Visual Analytics of Human Mobility Behavior

Von der Fakultät Informatik, Elektrotechnik und
Informationstechnik der Universität Stuttgart
zur Erlangung der Würde eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

Vorgelegt von

Robert Lutz Krüger

aus Aalen

Hauptberichter:	Prof. Dr. Thomas Ertl
Mitberichter:	Prof. Dr. Ross Maciejewski
	Prof. Dr. Johannes Maucher
Tag der mündlichen Prüfung:	19. Juni 2017

Institut für Visualisierung und Interaktive Systeme
der Universität Stuttgart

2017

Contents

List of Figures

List of Abbreviations and Acronyms

AOI	Area of Interest
API	Application Programming Interface
DBSCAN	Density-Based Scan Clustering
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IEEE	Institute of Electrical and Electronics Engineers
KDD	Knowledge Discovery in Databases
KDE	Kernel Density Estimation
MCV	Multiple Coordinated Views
MOBY	The MOBilitY Analysis System
NoSQL	Not only SQL
OD	Origin-Destination
OWL	Web Ontology Language
POI	Point of Interest
RDF	Resource Description Framework
RFID	Radio-Frequency Identification
SQL	Structured Query Language
VA	Visual Analytics
VAST	The IEEE Conference on Visual Analytics Science and Technology
VESPa	Visual Query Language for Event Sequence Patterns

Acknowledgment

This thesis would not have been possible without the support of my supervisors, colleagues, friends, and family. At first, I want to thank my doctoral advisor professor Thomas Ertl for giving me the opportunity to work in this successful research group, for giving me sophisticated advice when I needed it, but also for giving me enough freedom to be creative and to follow my research interests. I further want to thank my co-advisors professor Johannes Maucher and professor Ross Maciejewski for their advises, interest in my work, and their external report. I want to thank the University of Stuttgart and the Stuttgart Media University for choosing me as doctoral student for the PhD program "Digital Media" and granting me a three-year scholarship. I am thankful to the Institute for Visualization and Interactive Systems (VIS) and all people involved in research for the EU project CIMPLEX, the BMBF project VASA, and the DFG priority program VGI. These projects allowed me to work with other researchers in the field of visual analytics and to submit and present my research at international conferences. I am further grateful for my exceptional colleagues, for all the exchange of ideas and constructive criticism, for the joint work on my projects, and the long evenings, nights, and weekends we worked together. Special thanks go to Dennis Thom, Harald Bosch, and Steffen Koch who greatly integrated me from the very beginning. I further want to thank Florian Haag, Florian Heimerl, Kuno Kurzhals, Michael Woerner, Dominik Herr, Qi Han, Lars Lischke, Guodao Sun, Yafeng Lu, and Michael Steptoe for working with me on all the interesting projects and being co-authors on many publications. In addition, thanks go to Martin Falk for providing the LaTeX templates for this thesis. Lastly, I want to thank my partner Maike, my parents Erika and Roland, and friends who always believed in me, supported my decision and, most important, for giving me the necessary degree of freedom and understanding to work on this intense project. –It was a great journey.

Abstract

Human mobility plays an important role in many domains of today's society, such as security, logistics, transportation, urban planning, and geo-marketing. Both, government and industry thus have great interest in understanding mobility patterns and their driving social, economical, and environmental causes and effects. While stakeholders had to rely on manual traffic surveys for a long time, improvements in tracking technology made analyses based on large digital datasets possible. Recently, the omnipresence of mobile devices significantly increased the amounts of collected movement and context data. People are willing to reveal their position, but also further personal details such as visited places, observations, events, news, and sentiments in exchange for personalized services and social networking. This opens up new possibilities for many domains where a semantic mobility understanding is required but also raises major challenges.

To reveal a holistic picture, heterogeneous datasets of different services with different resolution and format have to be fused and analyzed. However, social sensing data is vast, has varying scale, is unevenly distributed, and constantly updated. Especially content from social media services is often inconsistent, unreliable, and incomplete, which requires special treatment. Fully automatic mapping approaches are not trustworthy as they do not take into account these uncertainties. At the same time, manual approaches become insufficient with large amounts of data. Even when data is perfectly aligned, analysts cannot purely rely on existing techniques. Answering questions about reasons for movement requires a broader perspective that takes into account environmental and social context, the driving forces for human mobility behavior.

Visual analytics is an emerging research field to tackle such challenges. It creates added value by combining the processing power and accuracy of machines with human capabilities to perceive information visually. Automatic means are used to fuse and aggregate data and to detect hidden patterns therein. Interactive visualizations allow to explore and query the data and to steer the automatic processes with domain knowledge. This increases trust in data, models, and results, which is especially important when critical decisions need to be made. The strengths of visual analytics have been shown to be particularly advantageous when problems and goals are underspecified and exploratory means are needed to discover yet unknown patterns.

This thesis presents novel visual analytics approaches to derive meaning and reasons behind movement, by taking into account the aforementioned characteristics. The approaches are aligned in a holistic process model covering all steps from data retrieval, enrichment, exploration, and verification to externalization of gained knowledge for various fields of application such as electric mobility, event management, and law enforcement. It is shown how data from social media can not only be used to retrieve up-to-date movement information, but also to enrich movement trajectories from other sources with structured and unstructured information about places, events, transactions, and other observations. Through highly interactive visual interfaces analysts can bring in domain knowledge to deal with uncertainties during data fusion and to steer the subsequent semantic analysis. Exploratory and confirmatory analysis techniques are presented to create hypotheses, refine them, and find support in the data. Analysts can discover routines and abnormal behavior with assistance of automatic pattern detection methods to cope with the vast amounts of data. Spatial drill-down is supported by a set-based focus+context technique, while a more abstract visual query language allows to explicitly formulate, extract, and query for movement patterns. The approaches are applied in different scenarios and are integrated in a visual analytics system. Evaluation with experts and novice users, case studies, and comparisons to ground truth data reveal the need and effectiveness of the contributions.

Overall, the thesis contributes a visual analytics process for human mobility behavior with novel semantic analysis approaches, ranging from global movements of many to local activities of a few people, for a wide range of application domains.

Zusammenfassung

Mobilität spielt eine wichtige Rolle in vielen Bereichen der heutigen Gesellschaft, wie zum Beispiel beim Thema Sicherheit, in der Logistik, dem Transportwesen, in der Städteplanung und im Geo-Marketing. Sowohl die öffentliche Verwaltung als auch die Industrie haben deshalb großes Interesse daran, Mobilitätsmuster und die sozialen, ökonomischen und umweltbedingten Ursachen und Wirkungen zu verstehen. Nachdem man lange Zeit auf manuelle Verkehrsumfragen angewiesen war, ermöglichten Fortschritte in der Tracking-Technologie Analysen auf Basis von großen digitalen Datensätzen. Durch die Allgegenwärtigkeit mobiler Geräte nahm die Anzahl von aufgezeichneten Bewegungs- und Kontextdaten stark zu. Im Gegenzug für personalisierte Dienste und soziale Vernetzung sind viele Menschen bereit, neben ihrem Standort weitere persönliche Details wie besuchte Orte, Beobachtungen, Ereignisse, Nachrichten und Stimmungen preiszugeben. Dies eröffnet neue Möglichkeiten in Anwendungsbereichen in denen ein semantisches Verständnis von Mobilität notwendig ist, aber birgt auch viele Herausforderungen.

Um ein ganzheitliches Bild zu erhalten, müssen heterogene Datensätze aus verschiedenen Diensten und mit unterschiedlichen Auflösungen und Formaten zusammengeführt und analysiert werden. Allerdings handelt es sich dabei um immense und sich ständig aktualisierende Datenmengen mit unterschiedlichen Maßstäben und ungleichen Verteilungen. Insbesondere Inhalte aus sozialen Medien sind häufig inkonsistent, nicht vertrauenswürdig und lückenhaft, was einen speziellen Umgang erfordert. Maschinelle Mapping-Verfahren beziehen diese Unsicherheiten jedoch meist nicht mit ein und liefern deshalb keine verlässlichen Resultate. Gleichzeitig werden manuelle Verfahren ineffizient, wenn sie auf sehr große Datensätze angewendet werden. Auch bei perfekt fusionierten Datensätzen können Analysten nicht immer auf bestehende Techniken zurückgreifen. Fragen nach Gründen für Bewegung erfordern eine breitere Sichtweise, die umweltbedingte und soziale Kontexte als treibende Kräfte für Mobilitätsverhalten mit einbezieht.

Visual Analytics ist ein aufstrebendes Forschungsfeld, um derartige Herausforderungen zu bewältigen. Ein Mehrwert entsteht insbesondere durch die Kombination von Rechenleistung und maschineller Genauigkeit mit den menschlichen Fähigkeiten, Informationen visuell aufzunehmen. Automatische Methoden werden verwendet, um die Daten zu aggregieren und zusammenzufügen und um versteckte Muster darin zu erkennen. Interaktive Visualisierungen erleichtern die Abfrage und Exploration der Daten und erlauben, die automatischen Prozesse mit Hilfe von Fachwissen zu steuern. Dies erhöht das Vertrauen in Daten, Modelle und Resultate und ist besonders wichtig, wenn kritische Entscheidungen getroffen werden müssen. Die Stärken von Visual Analytics kommen besonders zur Geltung, wenn Probleme und Ziele unterspezifiziert sind und explorative Methoden benötigt werden, um bisher unbekannte Muster zu entdecken.

Diese Arbeit stellt, unter Einbezug der zuvor benannten Besonderheiten, neue Visual Analytics Methoden vor, um Bewegungszusammenhänge abzuleiten. Die Methoden sind in ein ganzheitliches Prozessmodell integriert, das alle Schritte von der Datenabfrage, ihrer Anreicherung, Exploration und Überprüfung bis hin zur Externalisierung des erlangten Wissens beinhaltet. Zu den Anwendungsfeldern gehören unter anderem Elektromobilität, Event Management und Verbrechensaufklärung. Soziale Medien werden nicht nur genutzt, um sehr große Mengen an aktuellen Bewegungsdaten abzurufen, sondern auch, um Bewegungsdaten mit strukturierten und unstrukturierten Informationen über Orte, Ereignisse, Transaktionen und anderen Beobachtungen der Nutzer anzureichern. Hochinteraktive visuelle Methoden erlauben den Analysten, während der Fusion der Daten und der Analyse von semantischen Zusammenhängen Fachwissen einzubringen.

Um Hypothesen zu generieren, diese zu verfeinern und zu überprüfen, ob die gewonnenen Vermutungen durch Vorkommnisse in den Daten gestützt sind, werden verschiedene explorative und bestätigende Analysetechniken vorgestellt. Der Einbezug von automatischen Mustererkennungsverfahren ermöglicht es Analysten auch in sehr großen Datensätzen regelmäßige Abläufe sowie abnormales Verhalten zu entdecken. Räumliches Filtern der Daten wird durch eine auf Mengen basierende Fokus+Kontext Technik unterstützt, die aus mehreren interaktiven Lupen besteht. Eine abstraktere visuelle Abfragesprache ermöglicht zudem ein explizites Formulieren, Überprüfen und Extrahieren von Bewegungsmustern. Die Methoden werden in verschiedenen Szenarien angewendet und sind in ein Visual Analytics System integriert. Evaluierungen mit Experten und unerfahreneren Anwendern, Feldvergleiche und verschiedene Fallstudien verdeutlichen die Notwendigkeit, Nützlichkeit und Effektivität der Forschungsbeiträge.

Zusammengefasst beschreibt diese Arbeit einen generalisierten Visual Analytics Prozess für menschliches Mobilitätsverhalten. Dieser beinhaltet neue Methoden, die eine semantische Analyse in verschiedenen Anwendungsdomänen ermöglichen. Die Methoden eignen sich sowohl für Untersuchung globaler Bewegungsmuster als auch für die Analyse lokaler Aktivitäten.

CHAPTER 1

Introduction

The whole world is moving. Around 90,000 flights are scheduled each day, the number of cars surpassed one billion, and we walk around 100,000 kilometers in our lives by foot. Human movement covers important information and can be seen as a footprint of our society and our role in it, of our interests and needs. It is therefore not astonishing that both, government and industry share great interest in getting access to this information to solve the increasing amount of today's challenges. On a global scale, decision makers are overwhelmed with large transit and migration streams, while they have to maintain security and public health. Growing demands for transportation of people and goods damage the environment, raise logistic problems, and cause traffic jams on an already over-allocated road network. On a local scale, increasing market competition forces companies to get a deeper understanding of customer mobility behavior and to optimize their offers, accordingly.

While the industry monitors their fleets regarding position and performance, movement tracking technology has long reached the private sector, where smartphones and other wearable devices pervade our society. Their omnipresence allows to record precise digital traces. In addition to location, people share observations, places, activities, events, meetings, and sentiments in the web, in exchange of social networking and personalized information. Often the data is publicly accessible through social media. Prominent examples comprise microblog services such as Twitter and Sina Weibo, social location services like Foursquare, services from social networks such as Facebook and Google+, but also services from content sharing platforms, among them YouTube, Instagram, and Flickr[1]. This opens up new possibilities,

[1] Twitter: www.twitter.com; Sina Weibo: www.weibo.com; Foursquare: www.foursquare.com; Facebook: www.facebook.com; Google+: plus.google.com; YouTube: www.youtube.com; Instagram: www.instagram.com; Flick: www.flickr.com; last accessed: March 21, 2017

especially for applications where not only technical measures but a more holistic view that highlights semantics, dependencies, causes, and effects is important to understand the bigger picture and reasons behind human mobility.

While today's technology is capable to record and store increasing amounts of information [Gantz and Reinsel, 2012, Hilbert and López, 2011], the analysis of these large data sources to discover insights and gain new knowledge raises big challenges. In theory, fully automated approaches can handle the *big data problem* but reality is often different. Researchers have categorized and named the problems the *3Vs of Big Data* [Zikopoulos et al., 2011]. They comprise *volume*, *velocity*, and *variety*. Volume refers to the large amounts of data that raise not only computational but also analytical challenges regarding working processes and cognitive load limits. Velocity points to the increasing amounts of streaming data. Variety stands for data heterogeneity that hinders the creation of generic processes and tools as well as data fusion. Chen et al. [2014] add *veracity* as the fourth big data problem, referring to the uncertainty, inconsistency, and unreliability of many of today's data sources.

Because of these challenges many companies and institutions that have recorded large amounts of movement data cannot make use of it or are not even sure which problems the data is capable to solve. What is in the data first needs to be found out and analysis goals are underspecified, accordingly. The interdisciplinary research field of visual analytics (VA) [Thomas and Cook, 2005a, Keim et al., 2008] can help to tackle such challenges. It creates added value by combining the processing power and accuracy of machines with the human capabilities to perceive information visually. Automatic means are used to fuse and aggregate data, and to detect hidden patterns therein. Interactive visualization allows to explore and query the data, and to steer the automatic processes with domain knowledge. This increases trust in data, models, and results, which is especially important when critical decisions need to be made. The strengths of visual analytics have been shown to be particularly advantageous when problems and goals are underspecified and exploratory means are needed to discover yet unknown patterns.

1.1 Problem Statement

As for many domains, the four Vs of big data also account for the analysis of human mobility. However, due to the nature of movement data and the goals that afford a holistic mobility understanding, special challenges arise.

According to Laube [2014], traces that moving objects leave in space-time connect multiple data points. Movement data is thus more difficult to handle than simple point-based data. It can contain complex hierarchical structures of overlapping trajectories with diverse shapes, and directions. They may have varying spatio-temporal scale and resolution, and are unequally distributed. Relations between trajectories are more complex than relations of single data points, making data analysis even more challenging. Similarly, for an effective visual representation and understanding, one has to find means to overcome clutter and overplotting.

Another difficulty lies in the semantic gap, which is, according to Laube [2014], one of the main remaining challenges of movement analysis. Courgeau [1988, 2005] argues that human mobility is largely based on the social space in which it occurs. According to him, the study on mobility *"involves identifying the changes occurring over time in the relations between an individual or social group and space"*. It is the combination of movement and context that allows to derive such sophisticated insights. A post-analysis with recorded data, however, only reveals parts of the puzzle, because context information is often missing. Only by recomposing these parts a holistic overview and situational understanding can be recovered. As previously mentioned, with web 2.0 technologies, and especially social media, this data becomes available. However, if and how these sources can be utilized for human mobility analysis needs to be evaluated.

No matter where the context comes from, further challenges lie in the data fusion and joint analysis. Movement and context data are recorded, stored, and made available with different tracking techniques and services, which creates not only a patchwork of heterogeneous datasets with different timespans, resolutions, and formats, but also introduces many uncertainties for the mapping and interpretation of results. User-generated data is often inconsistent, and its content can rapidly change, which makes variety and veracity key issues. Large datasets and constant changes make a manual alignment and annotation inefficient but the heterogeneity and inconsistency render an automatic mapping ineffective. While visual analytics is promising to solve these challenges, most existing approaches concentrate on already well prepared data, offering no sufficient means that assist the analyst during preprocessing. In addition, the analysis of enriched movement data opens new challenges regarding exploration, querying, and knowledge extraction.

Lastly, the analysis of human mobility behavior has a wide field of applications, ranging from transportation and security to geo-marketing. On the one hand, different domain specifics have to be taken into account. On the other hand, researchers aim to find repetitive processes to built frameworks and best-practices for a large variety of applications. This leads to a trade-off between generalization of analysis approaches and processes, and applications that address domain specific requirements.

1.2 Research Questions

Based on the identified particularities the following research question arises:

Question 1: *Can visual analytics and the integration of social sensing data increase the semantic understanding on human mobility behavior?*

With respect to this research question, three following secondary questions emerge:

Question 2: *Can social media data serve as a source for movement and context data, and how can it be processed and represented with respect to its volume, variety, and veracity?*

Question 3: *Can a stronger integration of the human in the analysis loop through a closer and more interactive coupling of manual and automated techniques help to cope with these challenges?*

Question 4: *How can the approaches for human mobility analysis be applied to solve specific challenges in different application domains, and can a generic analysis processes be identified?*

1.3 Contribution

This thesis contributes approaches for a semantic analysis of human mobility for a variety of application domains. All approaches are integrated into *MOBY* (MOBilitY Analyzer), a prototypical visual analytics system, which was developed as part of this thesis for testing and evaluation. The system realizes a generic process model for a semantic analysis of human mobility behavior. The model is based on foundations from the fields of visual analytics, human cognition, and logical inference theory and abstracts best-practices from various application scenarios of electric mobility, event monitoring and planning, as well as law enforcement.

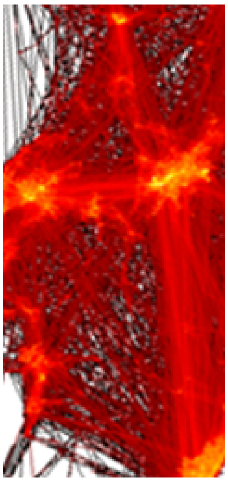
The integration of social media data plays a main part in the analysis process. The data is used for both, gaining access to massive and up-to-date movement profiles, but also enriching movement data with context. While so called *mashmaps* are commonly known to overlay different data in a stacked visualization, the approaches in this thesis additionally allow for a computational integration and explicit visual encodings of coherences. Volume, variety, and veracity issues during data retrieval and mapping are handled by a synergistic combination of automated and manual analysis means. An insightful handling of uncertainties resulting from the enrichment increases the analyst's awareness regarding these issues and allows to question, review, and improve data alignment at any stage of the analysis.

Many existing approaches propose visual analytics means for movement data. When carefully examined, however, they mostly propose an automatic processing and include visual means for result exploration only. By contrast, this thesis contributes means for a closer integration of the human into the analysis loop. This is achieved by highly interactive interfaces and a tight interplay of exploration and verification. Dynamic visual filters and concepts for faster computation allow to explore data with immediate visual feedback. Found cues can be inspected, and patterns can be extracted, refined, and directly executed as queries to verify if certain hypotheses match the data. This allows for a quantification of results, e.g., how often, where, when, and under which conditions certain routines or anomalies appear, and who is involved. The close interplay of hypotheses creation and verification enables a fluent narrowing from various heterogeneous datasets to compact patterns. Visual interfaces guide the analyst to saliences by highlighting routines and abnormalities in human mobility behavior. The analysis is supported by automated approaches from the field of unsupervised learning such as clustering and outlier detection.

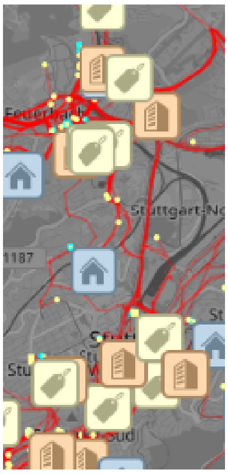
All design decisions, models, and approaches are evaluated with a variety of techniques in different domains. Groups of experts and novice users carried out comprehensive and analytical tasks, and answered questionnaires. An evaluation of low-level interactions using mouse-logging and eye-tracking was carried out to verify design decisions. Ground truth-based evaluations were applied and showed that a closer integration of human experts helps to cope with uncertainties to improve analysis results. In addition, multiple case studies demonstrate the applicability of the approaches. In summary, the evaluation results reveal the need for and the effectiveness of the contributions and provide ideas for further improvements. Through the integration of context information from a variety of social sensing data sources, semantic patterns that are not visible in the raw movement traces can be uncovered.

1.4 Thesis Structure

The thesis is structured as follows. Chapter 2 presents foundations and models in the field of visual analytics and introduces its terminology. It clarifies the technical interplay of its components but also gives a cognitive and task-oriented view. The second part of Chapter 2 introduces visual movement analysis, common datatypes, structures, and techniques. It then discusses the importance of context and the role of social media, before it combines the aforementioned to a new model for visual analytics of human mobility behavior. The steps of this generic pipeline, ranging from data retrieval via exploration and verification to the extraction of found patterns, are then subsequently discussed by presenting exemplary approaches in the Chapters 3–6. Because VA is an interdisciplinary field that benefits from the interplay of its components, there are often fluent transitions between two analysis steps. The approaches thus do not strictly focus on only one step. They are presented in order of their main focus.

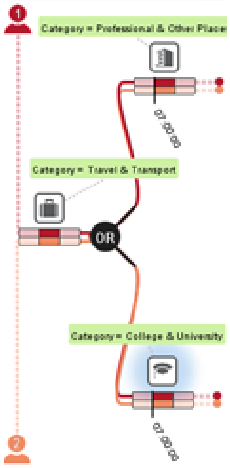


Chapter 3 addresses data retrieval, visual representation, and validation [Krüger et al., 2016b]. It discusses how social media, in particular geo-located microblog messages, can be used as an alternative source for movement data. Trajectories are reconstructed, and visual means to gain overview on different degrees of abstraction are provided, which cope with uncertainties resulting from low temporal resolution. Movements can be filtered based on inferred means of transportation, with respect to space, and to time. To evaluate the usefulness of this data source, i.e., if it covers real-world patterns, it is compared to ordinary datasets such as flight schedules on a global level and taxi data on a local level. Exemplary studies on large-scale events show the potential of this new data source.

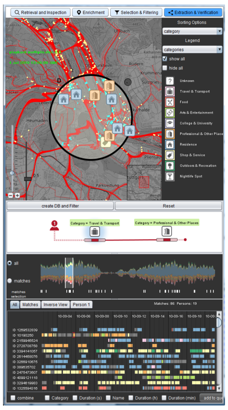


Chapter 4 presents means for manual [Krüger et al., 2012b, 2015b] and semi-automated enrichment [Krüger et al., 2014a,b] of movement data with context information on places, observations, and sentiments mainly from social media. The approaches allow to cope with heterogeneous data and large volumes. Services in focus are Foursquare and Twitter but data from transactions and mobile communication is also used. With this additional context, movements can be described as a sequence of events. Scalable geographic and temporal visualizations are proposed to explore and adjust these contextual mappings. While the chapter mainly focuses on the enrichment of e-mobility data, the case studies present an additional scenario in the domain of law enforcement.

To understand movement data in time and space methods to drill-down are essential. Many movement visualizations suffer from clutter, making it impossible to derive patterns. In Chapter 5 an interactive exploratory filtering method named *TrajectoryLenses* is described that allows for set-based filter combinations and immediate visual feedback [Krüger et al., 2013b]. Being introduced to work on raw trajectory data, two extensions for semantically enriched movements are provided. Close couplings with temporal views enable a spatio-temporal exploration. The applicability of the technique is evaluated with case studies on electric scooter and Twitter data. Recordings of eye and mouse interactions allow to verify the design decisions, and feedback from domain experts of a large energy provider and visualization experts emphasizes its usefulness.



To assist the exploration process of semantic patterns, Chapter 6 introduces automatic detection and extraction approaches from the field of unsupervised machine learning. They focus on the order and transitions of events. Clustering, outlier detection, and other sequence mining techniques guide the analyst to the routines and anomalies [Krüger et al., 2012b, 2015a, Steptoe et al., 2015] in the movement data. In addition, a pattern-based query language named *VESPa* gives a more expressive but compact representation of these patterns [Haag et al., 2016]. The visual language can be used to compose, query, refine, and extract them, and allows a closer interplay of exploration and verification tasks.



After all approaches are presented and their applicability is evaluated in various application domains, Chapter 7 shows their integration into a visual analytics system named MOBY (MOBilitY Analyzer). It enables a holistic analysis from data retrieval to the externalization of found patterns by implementing the process model (Chapter 2). Architecture and interfaces that allow to use the system with different data are presented. The section then reviews the process and its approaches regarding generic aspects, visual, computational, and interaction scalability. Lastly, Chapter 8 summarizes the contributions, before giving an outlook on open challenges.

Further Approaches
In the course of this thesis additional approaches were developed that are in the broader context of this work and referenced in the following chapters. They comprise means to encourage people in data collection [Tobien et al., 2016] and to estimate locations of social media messages when they are missing [Thom et al., 2014]. This thesis also discusses ideas from a visual social media analytics system [Bosch et al., 2013], results from a long-term study [Thom et al., 2015, 2016], and application scenarios [Bosch et al., 2015]. Furthermore, an approach for social media data orchestration for visual predictive analytics of box office success was developed [Lu et al., 2014, Krüger et al., 2013a], which influenced the enrichment approaches in this thesis. Lastly, the visual filtering technique was adopted to and evaluated with eye tracking trajectory data [Krüger et al., 2016a].

This thesis is partly based on the following publications:

R. Krüger, H. Bosch, S. Koch, C. Müller, G. Reina, D. Thom, and T. Ertl. HIVE-BEAT - A highly interactive visualization environment for broad-scale exploratory analysis and tracing: VAST 2012 Mini Challenge 1 award: Honorable mention for comprehensive visualization suite. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 277–278. IEEE Computer Society, 2012a

R. Krüger, S. Lohmann, D. Thom, H. Bosch, and T. Ertl. Using social media content in the visual analysis of movement data. In *2nd Workshop on Interactive Visual Text Analytics, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 1–3. IEEE Computer Society, 2012b

R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl. TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data. *Computer Graphics Forum*, 32(3):451–460, 2013b

R. Krüger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 193–200. IEEE Computer Society, 2014a

R. Krüger, D. Thom, and T. Ertl. Semantic enrichment of movement behavior with Foursquare—a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 21(8):903–915, 2014b

R. Krüger, F. Heimerl, Q. Han, K. Kurzhals, S. Koch, and T. Ertl. Visual analysis of visitor behavior for indoor event management. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1148–1157. IEEE Computer Society, 2015a

R. Krüger, D. Herr, F. Haag, and T. Ertl. Inspector Gadget: Integrating data preprocessing and orchestration in the visual analysis loop. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 7–12. The Eurographics Association, 2015b

M. Steptoe, R. Krüger, Y. Zhang, X. Liang, R. Garcia, S. Kadambi, W. Luo, T. Ertl, and R. Maciejewski. VAST Challenge 2015: Grand Challenge - award for outstanding comprehensive submission. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 119–120. IEEE Computer Society, 2015

R. Krüger, G. Sun, F. Beck, R. Liang, and T. Ertl. TravelDiff: Visual comparison analytics for massive movement patterns derived from Twitter. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 176–183. IEEE Computer Society, 2016b

F. Haag, R. Krüger, and T. Ertl. VESPa: A pattern-based visual query language for event sequences. In *11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (IVAPP)*, volume 7, pages 48–59, 2016

T. Tremel. Advancement of a query visualization for event sequences. Supervisor: R. Krüger, F. Haag, T. Ertl. Bachelor Thesis. Universität Stuttgart, 2016

Additional approaches were developed that are in the broader context of this work.

R. Krüger, H. Bosch, D. Thom, E. Puettmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl. Prolix - Visual prediction analysis for box office success. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, 2013a

H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013

D. Thom, H. Bosch, R. Krüger, and T. Ertl. Using large scale aggregated knowledge for social media location discovery. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1464–1473. IEEE Computer Society, 2014

Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski. Integrating predictive analytics and social media. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 193–202. IEEE Computer Society, 2014

H. Bosch, R. Krüger, and D. Thom. Data-driven exploration of real-time geospatial text streams. In *Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 203–207. Springer International Publishing, 2015

D. Thom, R. Krüger, T. Ertl, U. Bechstedt, A. Platz, J. Zisgen, and B. Volland. Can Twitter really save your life? A case study of visual social media analytics for situation awareness. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 183–190. IEEE Computer Society, 2015

D. Thom, R. Krüger, and T. Ertl. Can Twitter save lives? A broad-scale study on visual social media analytics for public safety. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1816–1829, 2016

P. Tobien, L. Lischke, M. Hirsch, R. Krüger, P. Lukowicz, and A. Schmidt. Engaging people to participate in data collection. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 209–212. ACM, 2016

R. Krüger, S. Koch, and T. Ertl. SaccadeLenses: Interactive exploratory filtering of eye tracking trajectories. In *2nd Workshop on Eye Tracking and Visualization (ETVIS), co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, volume 2, pages 31–34. IEEE Computer Society, 2016a

Chapter 2

From Foundations to Applications

"The purpose of computing is insights not numbers." is a quote by Richard Hamming who was involved in early advances of information science and telecommunication [Hamming, 1973]. An effective way to let humans derive insights from large amounts of data is visualization, as stated by Chang et al. [2009]. It takes into account the human capability to quickly extract patterns visually. Researchers assume that humans perceive more than 80 percent of the information through the visual channel [Kaplan, 2015]. Miller et al. [2008] found out that the gathering, encoding, integration, and processing of this information involves about 30 percent of the human brain. According to Ware [2012] humans *"acquire more information through vision than through all other senses combined"*. Nevertheless, often datasets are so large and complex that viewers would be overwhelmed with information. The number of data items by far exceeds not only the number of pixels on screen to represent them, but also the cognitive load to perceive and derive insights. Hence, clever means are needed to abstract from details, to give overview, to filter the data, and to highlight the important. Shneiderman [1996] coped with similar challenges in multiple analysis tasks with a design principle, he named *Information Seeking Mantra*: *"Overview first, zoom and filter, details on demand"*. With growing volumes and increasing variety, velocity, and veracity of data that is collected nowadays [Zikopoulos et al., 2011], this becomes all the more important but also increasingly challenging. To deal with massive and high dimensional datasets, automatic analysis methods have evolved that benefit from the fast processing power and high accuracy of machines. These approaches range from, e.g., data reduction, pattern detection, classification, and optimization to

automatic reasoning. For some problems a fully automatic solution is applicable. For other problems automatic means cannot replace the human mind, for example, when domain knowledge is needed, tasks are not clearly specified, or critical decisions have to be made. Decision makers need to understand and trust the meaning of analysis results as they are responsible for the effects of their actions. Similarly, Hamming [1973] stated that if an analyst *"[...] does not understand what is being done, he is very unlikely to derive much value from the computation. The bare numbers he can see but their real meaning may be buried in the computation"*.

The interdisciplinary research field of visual analytics couples algorithms and models with visual interfaces to support both, the human capability to derive information visually and the computational power of machines to rapidly calculate, aggregate, and extract patterns in the data. This opens the blackbox of automatic approaches and does not only increase the analyst's awareness of computational meanings but also allows to interactively steer the computation. This chapter introduces visual analytics by presenting different definitions and models (see Section 2.1). At first, a historical and technical understanding of visual analytics and its components and tools is given, before the focus is shifted to underlying cognitive aspects that influence the analysis process. Lastly, a task oriented perspective of information foraging and sensemaking is discussed.

While visual analytics can be applied to many application fields, the second part of this chapter narrows down the focus to the analysis of human mobility behavior, the topic of this thesis (Section 2.3). Types of movement data sources and data structures that play a certain role in the remainder of this thesis are presented, before the state-of-the-art for visualization and visual analytics of movement data is revisited. The chapter further introduces the importance of social aspects that drive human mobility behavior, as well as context information to enable the recovery of a bigger situational picture (Section 2.4). Social media is introduced as a vast resource for human movement and context data. The section then bridges the gap back to visual analytics and presents a process model for the analysis of human mobility behavior that takes into account this additional context. It is also outlined which of the remaining chapters address which steps of this process. The chapter finishes with Section 2.5, which presents four different application domains, stakeholders, and data sources that played a key role in the development and evaluation of the approaches of this work.

2.1 Visual Analytics

In 2004, after major terrorism attacks against the United States of America, the Department of Homeland Security (DHS) initialized research activities to counter future acts of terrorism. Being aware of the advantages of visualization to perceive and understand complex information, the National Visualization and Analytics

Center (NVAC) was chartered (among others) to research new approaches in the field of visual information analysis. Together with the industry, they set up a research agenda to reduce vulnerability by a set of measures, e.g., to prevent and protect the population, increase situational awareness, but also to respond to and recover from acts of terrorism. A more detailed agenda was published in a technical report called "Illuminating the path: The research and development agenda for visual analytics". In this work, Thomas and Cook [2005a] defined visual analytics as *"the science of analytical reasoning facilitated by interactive visual interfaces"*. Here, *analytical reasoning* is an abstract definition of methods by which users obtain deep insights. Since then the research definition and approaches spread out to other disciplines besides counter terrorism, law enforcement, and situational awareness. Internet security, healthcare, logistics, and digital humanities are some of a wide range of new areas where visual analytics techniques are applied.

While the technical report by Thomas and Cook [2005a] was one of the first publications where the expression *visual analytics* was explicitly named, visual analytics ideas have been developed for many years by researchers from different fields like databases, human computer interaction, and visualization. Most similar to today's definition of visual analytics is *visual data mining* which was first mentioned in the early 2000s by Keim [2002]. He stated that through visualization and interaction the user can be directly involved in the analysis process. Similarly, Shneiderman [2002] pointed out that the strength of statistical algorithms to automatically detect patterns could be combined with the strength of visualization to communicate patterns in a way that is easy to process for humans. Later advances in this area are also explicitly named visual analytics, although the definition by Keim et al. [2008], which was given later in 2008, slightly differs from Thomas and Cook [2005a]. It explicitly mentions the integration of automated analysis approaches:

"Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets."

Keim's definition of visual analytics is also represented in a process model [Keim et al., 2008, 2010] (see Figure 2.1). Here, automated analysis techniques are depicted as a step from the raw data to a model, i.e., an abstract representation of the data. This can be achieved by various means ranging from simple partition and aggregation techniques to more complex methods from the fields of data mining and machine learning. In the following, the main components of visual analytics are discussed: data visualization, model building/data mining, and interaction.

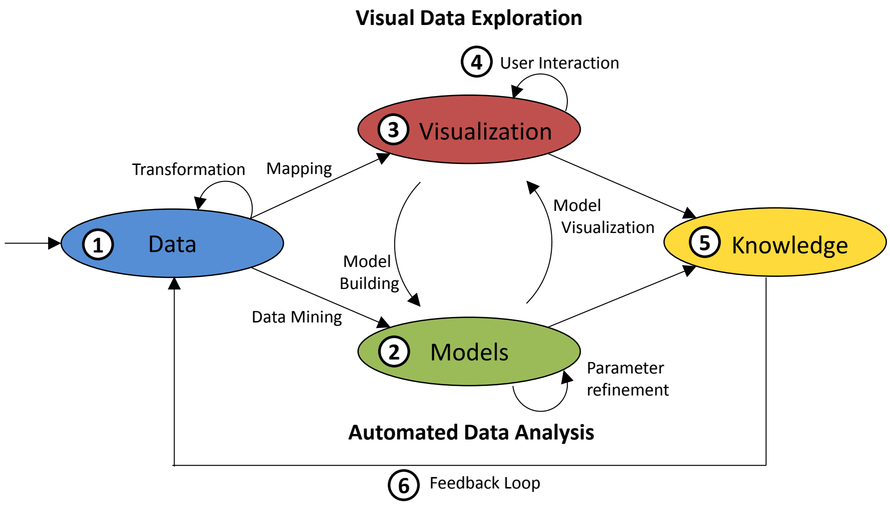


Figure 2.1 — The visual analytics model according to Keim et al. [2008] gives a technical perspective. Automated analysis approaches that abstract data (1) to models (2) are coupled with visualization (3) through interaction (4) to derive knowledge (5). This is done in an iterative manner (6).

2.1.1 Data Visualization

Visualization addresses the human perceptual capabilities to identify and interpret patterns visually (Figure 2.1, step 1 → 3). Card et al. [1999] define it as *"the use of computer-supported, interactive, visual representations of data to amplify cognition."* with the purpose of *"discovery, decision making, and explanation"*. The computation of visualizations follows a process by mapping data to visual primitives and rendering the data to images. This is also depicted in the visual analytics model. The reference model for information visualization, proposed by Card et al. [1999], gives a more precise definition (see Figure 2.2, top) that allows for interactive manipulation at different stages of the pipeline.

Depending on the data and visualization purpose, different mapping and rendering approaches are applied. Kerren et al. [2008] split the field of visualization into scientific visualization, which addresses data with given spatial information, and information visualization, which comprises data where the spatial representation is chosen. Similarly, Card et al. [1999] define information visualization to use abstract data that is not physical based. Geographic visualization can be located in between these two fields, since the data has both abstract and spatial components. Geographic visualization often uses abstraction and mixes spatial with abstract representation such as symbols and networks. MacEachren et al. [2004] stated that geographic visualization draws on *"[..] cartographic and geographic traditions, integrating their perspectives on representation and analysis of geospatial information with more recent developments in scientific and information visualization"*.

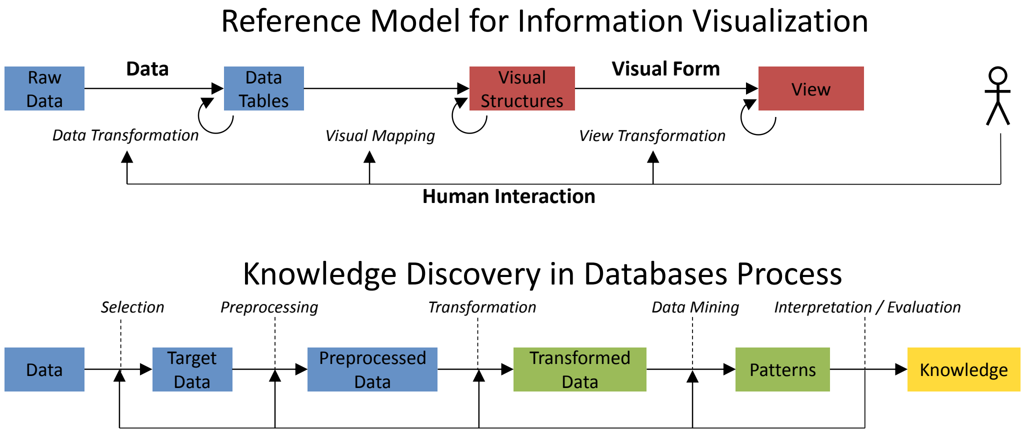


Figure 2.2 — Top) Reference model for information visualization according to Card et al. [1999]. Bottom) The KDD process according to Fayyad et al. [1996]. Both models share similar data processing and transformation steps (in blue).

2.1.2 Data Mining

Fayyad et al. [1996] propose an iterative process to automatically extract patterns from large and complex datasets, named the KDD (Knowledge Discovery in Databases) process (Figure 2.1, step 1 → 2). It ranges from target data selection, preprocessing, transformation, and data mining to knowledge (see Figure 2.2, bottom). After preprocessing, the transformation step includes feature selection and extraction methods, that serve as input for the *data mining* step where patterns are extracted from the data. The expression originates from mineral mining where, e.g., gold nuggets are carved out from stones and gravel. Data mining comprises different algorithms from the fields of machine learning that can be categorized in supervised and unsupervised approaches. Supervised approaches first train a regression or classification model based on known target values (e.g., classes). With the learned dependencies they then predict missing values or categorize class memberships. In many cases, labeled data where the outcome for an input is already known, is available. Unsupervised approaches do not have a supervised phase and are applied when target values are unknown. They learn groups, associations, and correlations based on any pre-defined metric. Clustering and outlier detection belong to the field of unsupervised machine learning. In this thesis unsupervised approaches are applied to support the data fusion and exploration process. They automatically aggregate and separate data based on characteristics and distributions and are helpful to abstract the relevant, to give overview, and to highlight routines and anomalies.

The KDD process has similarities with the information visualization process in data retrieval, preprocessing, and transformation (marked in blue), while the final steps differ. The KDD process automatically extracts patterns but does not explicitly define means to present them to humans. The visualization process does not extract any patterns but computes representations to let humans perceive them.

2.1.3 Interaction

Interaction (Figure 2.1, step 4) plays a crucial role in visual analytics and its components. Through interactive interfaces the user can navigate through, select, and manipulate data, models, and visualization. The visualization model (Figure 2.2) explicitly depicts the analyst who can transform data and map it to visual primitives such as position, shape, and color, and transform views. In the KDD process there is no explicit integration of an analyst but the arrows back to previous steps refer to any kind of adaption, either based on automatic optimization or based on parameters chosen by a user [Fayyad et al., 1996]. In visual analytics, both, data mining and visualization are combined through human-computer interaction (see Figure 2.1, user interaction). This allows for a close integration of the human analyst in all analysis steps, commonly named *human in the loop* [Endert et al., 2014] and leads not only to a deeper understanding of the analytic methods, but also allows for an integration of analysts' domain knowledge. An expert on human mobility, for example, might interactively adjust parameters of a clustering based on known distance and speed limits for flights or excludes routes with high risks from the list of possible connections in a transportation network. Similarly, a crime investigator could annotate abnormal movement paths from GPS-equipped devices based on eyewitnesses' statements to narrow down the set of possible culprits and train a classifier for automatic support when coping with larger datasets.

Especially for an exploratory data analysis where the problem may not be fully specified and discovery of yet unknown patterns is in focus, interactive means to steer the analysis and to get immediate visual feedback are important. Deep insights are usually not gained by looking at a static image only. Different viewing angles and perspectives on the data, different levels of abstraction and detail need to be investigated. Over the years, information visualization and visual analytics researchers have established best practices on interaction techniques, although there are multiple opinions on how they can be categorized (e.g., by Yi et al. [2007]). Shneiderman [1996] initially defined seven interactive tasks for data analysis. Four of them are included in the information seeking mantra: to gain *overview* of the entire dataset, to *zoom* in on items of interest, to *filter* out uninteresting items and select groups, and to get *details-on-demand*. The remaining three tasks are to *relate* items, to keep a *history* of actions to support, replay, and refine, and finally to *extract* sub-collections and query parameters.

Different approaches to realize these abstract ideas have been proposed with respect to the application domain, users, and devices. Cockburn et al. [2009] summarize approaches to get overview+detail, to zoom, and to focus on details without losing context. For them, the term *overview+detail* refers to *"a spatial separation to partition contextual and detailed information"*. A well-known example of this concept is a file explorer, where in one view the folder and hierarchy is shown and in the other detailed information of files for the selected folder are displayed.

Zooming, by contrast, is a practice that maps space to time and is a *"temporal separation between views"* [Cockburn et al., 2009]. Smooth animations can improve orientation because they allow viewers to keep their mental map for a short time, but on the contrary, do not provide overview and detail at the same time. Zooming has a strong analogy to real-world navigation where objects geometrically occur larger (geometric zooming), and more details can be seen (semantic zooming), if objects are closer. Common examples of visualizations that use zooming are interactive geographic maps. By contrast, *focus+context* techniques provide means to see details and keep the overview at the same time, within the same view. A common example are magic lenses as *movable display regions*, introduced by Bier et al. [1993], that show additional details underneath. They are often combined with *fisheye* distortion techniques [Furnas, 1986]. These techniques stretch parts of the visualization in a way that more details can be shown and compress it in areas further away from the focus. At last, Cockburn et al. [2009] name *cue-based* techniques that do not serve as an overview but just give orientation and hints by pointing, e.g., to an area of interest. Window sliders and halos can be categorized as cue-based approaches.

Among techniques for data filtering, *brushing and linking* became a state-of-the-art concept in visual analytics. Especially, when data is complex and multi-dimensional it can be difficult to understand dependencies between data instances (data cases) and features (also named data dimensions or attributes). Brushing and linking can coordinate multiple views that present different aspects of the data. When the user selects (brushes) parts of the data in one view, the same parts are emphasized in the other views (linking), making the connection among, e.g., features very clear. Roberts [2007] reviewed the state-of-the-art on multiple coordinated views (MCV) and North and Shneiderman [2000] evaluated that MCV improve user performance by 30-80% over detail-only and uncoordinated interfaces. For geographic analysis MCV usually comprise a geographic map for spatial exploration, a timeline for temporal overview and navigation, as well as multiple abstract views on further data attributes.

2.2 Knowledge Generation

The definition of visual analytics by Keim et al. [2008] is a rather technical perspective consisting of human interaction and computational steps for the generation of models and visualizations. It also includes a cognitive step from models and visualizations to knowledge (Figure 2.1, 5), which is, however, not further described. To get a deeper understanding of the fundamentals how analysts can derive conclusions one can investigate Peirce's [1974] theory of inference and how to move from premises to conclusion. A process model that takes into account real-world practices is given by Pirolli and Card [2005].

2.2.1 Inference Theory

The inference model consists of three types of reasoning: deduction, induction, and abduction that *"form a process flow for the construction of knowledge"* [Pike et al., 2009]. Deductive reasoning allows to reach a conclusion by applying general rules. An observation can be explained with preknowledge. This allows to conclude further dependencies and predictions. It is also referred to as *top-down process* that runs from theory to observation. With correct premises, deduction delivers correct results. Induction, on the other hand, infers such a rule by collecting observations that provide common evidence about the truth of a conclusion. Induction runs from observations to theory and allows to *"synthesize individual statements into general laws"* [Magnani, 2005]. Thirdly, abduction is commonly referred to as *inference to the best explanation*. This more creative process is controversially discussed. Many researchers and philosophers say that humans do not strictly follow an inductive or deductive process and abduction is the common way of how humans explore the world. The idea is that one tries to find the best explanation for a set of observations by fitting it to the most explanatory hypothesis available. In summary, Peirce [1974] states: *"Deduction proves that something must be; Induction shows that something actually is operative; Abduction merely suggests that something may be."*

Yu [1994] set Peirce's inference theory in relation to exploratory data analysis and discusses abduction as the goal, where the analyst discovers new dependencies and forms a general hypothesis. Existing hypotheses allow, e.g., to classify further data cases, accordingly (deduction). If and how many cases in the data really lead to/ align with that pattern can then be verified by example (induction). Yu [1994] defines induction to be necessary for the empirical substantiation that adds quantitative details to the conceptual knowledge.

2.2.2 The Sensemaking Process

Based on Pierce's foundations, different theories on *sensemaking* [Dervin, 1983], a process that combines action and cognition to gain insights and construct meanings, were created. Some researchers take a more abstract and cognitive perspective, while others have a narrow focus and try to extract repetitive processes from real-world practices of information retrieval and analysis. The latter aligns with the suggestion by Endert et al. [2014] that visual analytics should recognize and integrate human working processes into the system. A widely used definition of sensemaking in the visual analytics research community is given by Russell et al. [1993] and Pirolli and Card [2005]. They define sensemaking as the processes of *"searching for a representation and encoding data in that representation to answer task-specific questions"*. Their perspectives are based on an empirical study of real-world working processes in the domain of intelligence analysis, the root of visual analytics (see Section 2.1). It consists of a foraging and sensemaking loop (see Figure 2.3).

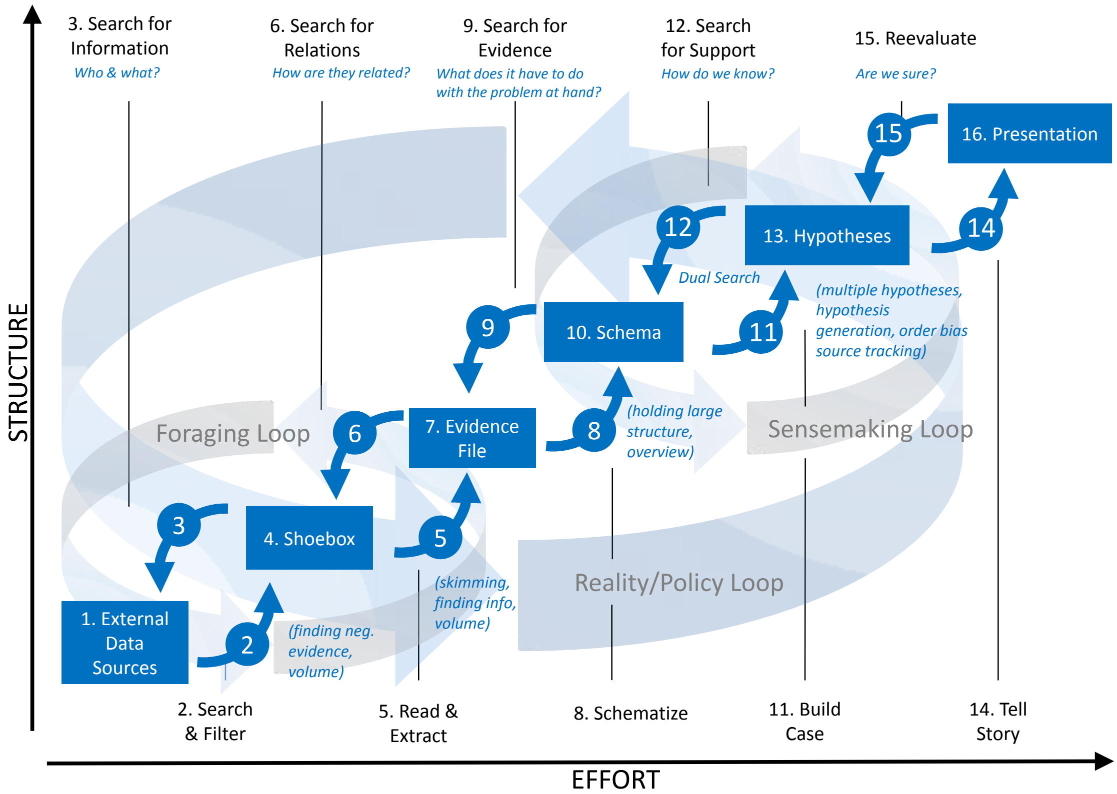


Figure 2.3 — The sensemaking process according to Pirolli and Card [2005].

The foraging loop *"involves processes aimed at seeking information, searching and filtering it, and reading and extracting information"* [Pirolli, 2007]. Data is retrieved and explored and stepwise reduced to a so called evidence file (see Figure 2.3, steps 1-7), a more narrow and precise description of the data that can be used to draw further conclusions. Pirolli and Card [1999] and Pirolli [2007] discuss a trade-off between recall and precision in the foraging process. Based on observations by Patterson et al. [2001], broader retrieval and exploration leads to higher recall but lower precision, while enrichment and fusion with other data sources and further restrictions narrow the possibilities and matches. This leads to higher precision but lower recall in the *exploitation* step, the generation of inferences and finding of patterns. Foraging aims to create a good starting point for the sensemaking loop *"that involves iterative development of a mental model (a conceptualization) that best fits the evidence."* [Pirolli, 2007]. From the evidence file (see Figure 2.3, 7) a schema can be generated (step 8). The schema holds a structural overview of the retrieved dependencies. This may be supported by tools or simply be in the mind of the analyst. The following steps (11-16) evolve around hypotheses generation, verification, and knowledge externalization to form a story and present the retrieved patterns.

There is a smooth transition between information foraging and sensemaking, and different starting points for the analysis are possible. [Pirolli and Card, 2005] introduce a bottom-up and a top-down perspective on how information processing takes place. The bottom-up process is data-driven and runs from data to theory. The analyst explores the data and derives relevant aspects from it, regarding the analysis goal. These aspects are then represented in a schematic way and can be used to form hypotheses. By contrast, the top-down process runs from theory to data. The analyst has a hypothesis in mind, expresses it, and can, e.g., query and search for matching or mismatching cases in the data. Klein et al. [2006] argue that in practice the cognitive process does not strictly follow this waterfall model, i.e., has no starting and ending point. Even data has to be collected or generated and hypotheses already exist from past experiences. For the sensemaking process by Pirolli and Card [2005] this means that there is a fluent and iterative transition of top-down and bottom-up approaches, which can be aligned with abductive reasoning (see Section 2.2.1), i.e., what is observed, what hypothesis exist, and which of them infers to the best explanation. The analyst may need to refine the hypothesis to narrow or widen the result sets and consider new perspectives, which may require further data exploration. While Klein et al. [2006] give a more generic view on human sensemaking, the model by Pirolli and Card [2005] is derived from a real-world process (intelligence analysis) and can be aligned with concrete computer-aided analytic actions, such as retrieval, querying, exploration, and reporting. A working process with concrete analytical steps to derive and extract patterns is especially important for decision makers that are in charge of their actions and need to verify decision foundations with care.

In summary, there are different orthogonal aspects to consider: the technical interplay of visual analytics components, the working process, and a cognitive process, making it difficult to create a combined model using boxes and arrows. Recently, Sacha et al. [2014] contributed their viewpoint with a model that integrates two of these aspects. They extended Keim's visual analytics model [Keim et al., 2008, 2010] by adding details about the knowledge component and combine technical and cognitive aspects on a abstract level. Specifics in different application domains, however, call for specialized models that take into account the domain's working processes and tasks.

2.3 Analysis of Movement Data

This thesis applies visual analytics on movement data to understand human mobility behavior. The analysis of human mobility is a wide and constantly growing field of research, relevant to many application domains. These range from sociological studies, over political studies, urban, traffic, and transportation planning, as well as security related topics, to visitor and customer analysis in geo-marketing and sales. The analysis is made possible by today's tracking technology that became an ubiquitous element in our lives. Sensors are integrated not only in various means of transportation from aircrafts to cars, but also in wearable devices like smart phones and watches. This section introduces possibilities to acquire positional data, discusses resulting specifics, and presents a taxonomy of movement data structures. It then shifts the focus to the state-of-the-art in visual movement analysis and presents spatio-temporal and event-based perspectives on movement.

2.3.1 Data Acquisition

There exist many possibilities to acquire movement data. The approaches presented in this thesis are based on three data source categories: sensor readings, schedule data, and artificial data generated by simulations (see Figure 2.4).

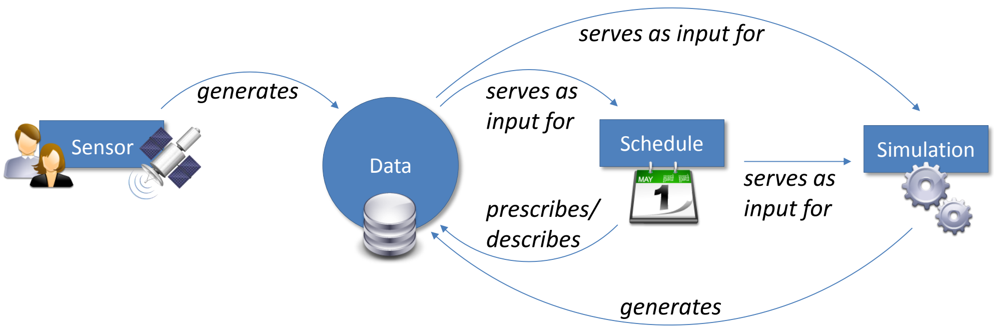


Figure 2.4 — Data can be gathered from technical or human sensors, observing a situation, from schedules that prescribe and describe a routine, or from simulations.

Sensor data is the preferred source for most analysis tasks, since the data covers recordings from real-world actions. Sensors are defined as objects responding to *"some type of input from the physical environment"*[1]. A prominent means to track location are GPS sensors. The Global Positioning System is currently the most commonly used space-based global navigation satellite system, followed by GLONASS, the Russian global navigation satellite system, and GALILEO, the European equivalent. Non-satellite-based tracking often utilizes mobile communication networks such as GSM (Global System for Mobile Communications). It is based on multilateration of radio signals and signal strength between cells. With different advances, GSM has a precision of up to 50 meters, while GPS accuracy is more precise (between 2 to 4 meters in 95 percent of cases), depending on atmospheric conditions and the accuracy of the GPS receiver[2]. Both are limited to outdoor usage because of signal scattering and attenuation. For indoor tracking WiFi (Wireless Fidelity)-based location detection became the state-of-the-art. Having different access points the location guessing works similar to the GSM networks and can further be enhanced with statistical tools like particle filters,

[1] http://whatis.techtarget.com/definition/sensor, last accessed: January 15, 2017
[2] Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report, http://www.nstb.tc.faa.gov/reports, last accessed: January 15, 2017

e.g., Sequential Monte Carlo Localization (SMCL). A more simplistic approach is stationary tracking. Examples are toll systems that track the number and the identity of surpassing cars, and RFID (Radio-Frequency Identification) readers that track a product along an assembly line or on its delivery chain. When only the amounts of moving objects are tracked but the objects' identities are not recorded, movement paths cannot be reconstructed. The data then simply contains temporal distributions of movement at specific places. Recently, people increasingly make use of the previously discussed positional tracking technology, e.g., in smartphones, and share their position along with personal information through social media. This will be further discussed in Section 2.4.2.

Secondly, schedules can be utilized to derive information about movement routines. An advantage of schedule data is that it is often available even when the moving object is not tracked or the data is not accessible for analysis but the underlying repetitive procedure is known. Schedules prescribe temporal and spatial conditions of a procedure. As they are often built on real-world observations, rules, and physical conditions, schedules can also be seen as descriptive models. People are used to schedules about transportation means such as buses, trains, and flights, and create their own plans of working, shopping, and living procedures. In most cases, schedules only define start or end position and time of an action. Data reliability depends on whether the schedule is up-to-date and whether prescriptions can be technically fulfilled. Another drawback is that schedule data covers planned routines only. It does not contain any outlying action such as delays.

Another possibility is to rely on simulations. According to Banks [1998], a simulation is *"the imitation of the operation of a real-world process or system over time"*. When real-world processes cannot be observed, simulations can help to understand the options, dependencies, and result spaces of a system [Wörner, 2014]. In logistics and traffic planning simulation is thus a main tool for the evaluation of new traffic networks. The more knowledge about real-world processes is available, the more accurate the simulation model can be configured. Many visual analytics approaches have been developed based on simulation data, because real data is often not publicly available. Simulated data is furthermore used for evaluation purposes. Patterns are simulated on purpose and analysts can then evaluate the sufficiency of novel analysis approaches by comparing analysis results to this simulated ground truth. A well-known forum for simulated datasets is the IEEE Vast Challenge workshop, colocated to the IEEE Visualization Conference (VIS). In the annual challenge participants have to uncover patterns and make predictions using various artificial datasets. Topics addressed in the challenges are derived from real-world problems. Often, stakeholders from industry are involved. They benefit from novel ideas of the community. Over the years a large repository has been developed[3] that offers datasets of all challenges since 2003 [Scholtz et al., 2012].

[3] http://hcil2.cs.umd.edu/newvarepository/benchmarks.php, last accessed: March 23, 2017

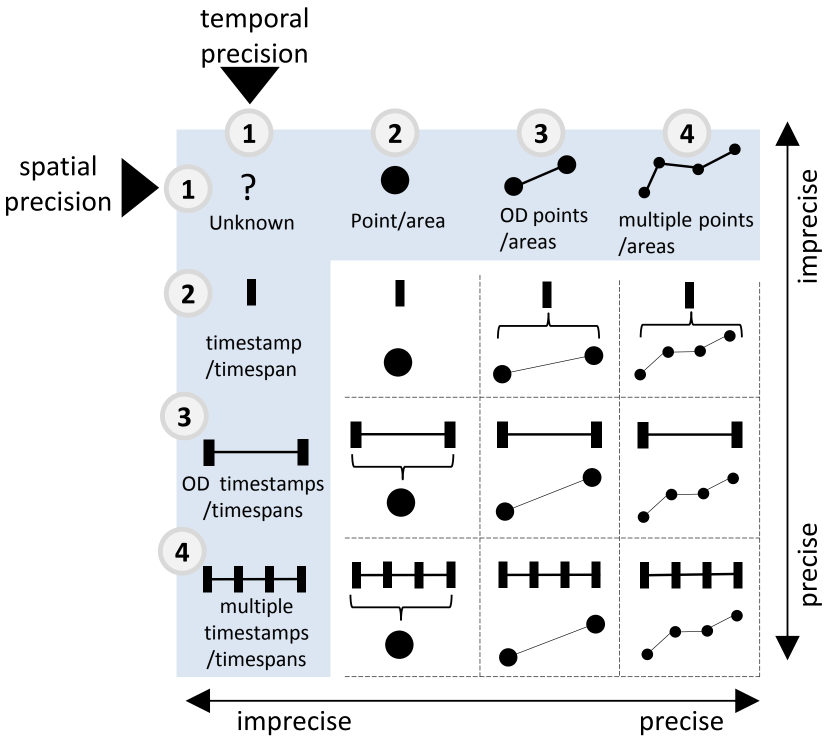


Figure 2.5 — Movement data can have different spatial and temporal resolutions, ranging from no information given (1), over a single recording (2), and origin-destination information (OD) (3) to multiple recordings (4), allowing to reconstruct the movement's path. Spatial information can be vague, depicting an area or a precise location. Temporal information can depict a timespan or a timestamp.

2.3.2 Data Structures

Although movement is continuous, movement recording technology usually samples it into a set of spatio-temporal measurements. With the recorded measurements of length n, one can rebuilt the movement's trajectory tr that describes its path through space as a temporally ordered sequence. This is expressed in Equation 2.1 where p_i is the spatial location of a measurement at time t_i with $i = 1, \ldots, n$.

$$tr = \big((p_1, t_1), (p_2, t_2), \ldots, (p_n, t_n)\big) \text{ with } t_1 < t_2 < \ldots < t_n \tag{2.1}$$

Movement datasets usually contain a set of moving objects O, where each object $o_j \in O, 1 \leq j \leq w, w = \|O\|$ holds a set of trajectories $TR = \{tr_1, \ldots, tr_m\}$. The temporal and spatial resolution of the trajectory data mostly depend on the tracking and storing technology that allows to record movement at different frequency.

Figure 2.5 categorizes recorded movement into precision classes. Besides the spatio-temporal resolution (1-4), the precision of a single record can vary. It can, e.g., roughly describe an area, further labeled as (a), or can have precise location information (b). The same applies to temporal information that either points to a timespan (a) such as a day, month, or year, or can precisely express a point in time (b). An example for low spatial and temporal resolutions are stationary scans that count passing objects. It is recorded, e.g., how many concert visitors enter a stadium at that day (see Figure 2.5, $(2, 2)$). Slightly more information is covered in so called OD (origin-destination) data $(3, 3)$, which is also used in this thesis. For example, in Section 2.5.1, OD data is reconstructed from geo-located microblog messages with precise location information but uncertain temporal data $(3a, 3b)$. It is unknown when the user has originally arrived or left a location. By contrast, Section 2.5.3 presents an OD dataset with low spatial resolution (rooms instead of precise locations are linked) but comparatively high temporal resolution $(3b, 3a)$. Category $(4, 4)$ represents trajectories with a high measurement frequency. Trips from electric scooters (Section 2.5.2) were recorded with a frequency of three seconds and precise location $(4b, 4b)$. Other combinations are less likely but occur.

Graphs

Besides data structures that describe movements of individual objects, movement can be spatially aggregated to reduce computation and storage costs, to correct signal dithering and remove outliers, and to get a better overview. Discrete spatial aggregation can be based on some existing boundaries, e.g., political or infrastructure boundaries or can be computed based on the data distribution itself, e.g., by clustering. For traffic analysis, aggregation can also be based on a road network. A full spatial and temporal aggregation of all trajectories in a dataset results in a static graph $G = (V, E)$ consisting of vertices V (also named nodes) and edges E. Each edge $e = (u, v)$ with $u, v \in V$ can encode directions and can contain a weight $w(e)$ that holds the travel volumes between nodes for the full time period. By varying the granularity of the temporal aggregation one can also keep certain temporal changes on an, e.g., hourly, daily, or weekly basis. Movement between areas can then be described as a dynamic (time-varying) graph $G := (G_1, ..., G_T)$ where each graph $G_t = (V_t, E_t), 1 \leq t \leq T$ holds a spatial aggregation of movements for the associated timespan. Depending on the analysis purpose, one can aggregate all trajectories to a single graph, construct multiple graphs, one for each object, or preserve each trajectory. Even in the latter case, spatial aggregation can significantly reduce data sizes, since trajectories can be expressed by referencing to nodes and edges of the graph structure.

Fields

The spatial characteristics of a large movement dataset can also be described as continuous density fields, which preserves the natural distribution of the data, and generates a smooth surface. The idea is that the recorded data is only a random sample of the population, but the population at a point x can be estimated from this sample using a probability density function (see Equation 2.2). x_i represents an instance of a set of n data points with $1 \leq i \leq n$ and h is a scale factor for the kernel $\mathscr{K}$.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathscr{K}\left(\frac{x - x_i}{h}\right) \tag{2.2}$$

The choice of the kernel function, e.g., Gaussian or Epanechikov kernel, depends on the assumptions regarding the data distribution. Maciejewski et al. [2010] integrates a kernel density estimation (KDE) for urban and rural population in a geovisual analytics approach. For trajectory data usually a line kernel is used [Lampe and Hauser, 2011] or supersampling between measurements is applied. Integrating over a two-dimensional area results in a density field. Willems et al. [2009, 2013] also apply KDE to visualize trajectory data. In addition to spatial attributes, they include time in their estimation of the density field.

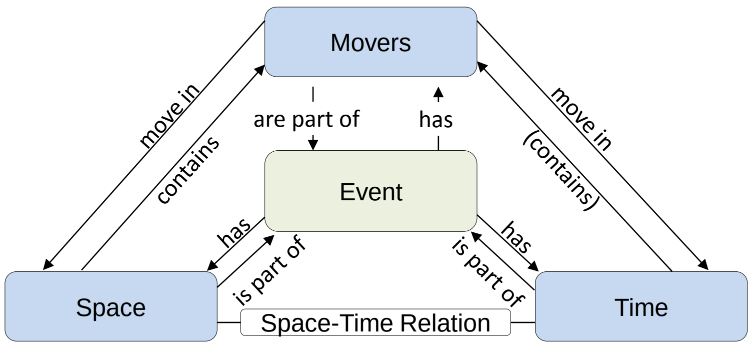
If not density but direction and speed are of interest, movement data can be represented as a vector field that encodes the direction and velocity. Similar to the density field the average movement direction and velocity for each position in a 2D space is calculated. Different approaches exist to transform trajectory data of individual moving objects into a vector field. Ferreira et al. [2013] suggest a grid-based approach. Each trajectory is sampled along a triangular grid and contributes to the average direction and velocity of the underlying cell. Subsequently, an interpolation function is used to form a continuous flow. Kim et al. [2017] represent and analyze flow maps of non-directional spatio-temporal data unaccompanied by trajectory information using Newton's law of gravitation.

2.3.3 Visual Analysis Perspectives

As introduced in Section 2.1, visual analytics combines the advantages of computational analytical means with the human capabilities of pattern recognition through visualization. Depending on the data characteristics, the application domain, and analysis goals, different perspectives on the data are of interest and different aspects can be highlighted through visualization. Andrienko and Andrienko [2013] contribute new taxonomies, ideas, and techniques for visual analytics of movement. They discuss four main perspectives on movement data: a spatial and a temporal perspective, a perspective on events, and a perspective on movers. The relationship is depicted in Figure 2.6. In the following, a brief overview is given.

Spatial Perspective

Movement data has a spatial dimension, and visualizations map geographic attributes usually to space, resulting in maps. Nowadays, maps are well established in our society and became the primary way to represent, analyze, and communicate geographic data. One of the first known maps is dated back to 6,500 BC.



◄ **Figure 2.6** — An altered version of the *Four perspectives of the phenomena of movement* according to Andrienko and Andrienko [2013]. The visual analysis can focus on space, time, movers, and events.

A neolithic cave painting shows a volcanic eruption of the city of Çatalhöyük in Turkey [Schmitt et al., 2014]. A prominent example for geographic analysis is the cholera map by John Snow, an English epidemiologist [Snow, 1855]. He visualized the locations of cholera related death in London. The spatial correlation to water pump locations revealed that the water source was contaminated, and indeed the death rate decreased after removing the pump handle. One benefit of maps is that they intuitively preserve the relative spatial distances, scales, or angles of the geographic data and hereby highlight the spatial autocorrelation. This was also described by Tobler [1970]: *"Everything is related to everything else, but near things are more related than distant things"*. It aligns with Newton's universal law of gravitation, measuring the attraction of objects based on their mass and distance. This also accounts for geographic phenomena, e.g., the larger and closer cities are, the more connected they usually are, including traffic and trade.

Geographic movement visualizations usually draw movement data on satellite imagery or on more abstract representations of road networks. This allows to discover strongly or barely connected regions, similar routes, and travel directions. Static visualization ranges from a detailed representation of single trajectories by showing paths or glyphs [Netzel et al., 2014] to more abstract representations, allowing to focus on patterns of varying granularity. Node-link diagrams that visualize aggregated movement (see Section 2.3.2) give a clearer image about strongly connected regions [Andrienko and Andrienko, 2008]. Depending on the layout, level of aggregation, and connectedness of the graph, however, visual clutter remains a major issue. Holten [2006] and Holten and Van Wijk [2009] thus propose clutter reduction techniques which bundle nearby edges that follow a similar direction. Andrienko and Andrienko [2008] present another approach that handles distant paths as multiple transitions between neighboring nodes. A smoother visual impression can be provided by accumulating trajectories to density fields (see Section 2.3.2) using *edge-splatting*. Resulting values are mapped to color [Lampe and Hauser, 2011, Willems et al., 2013]. To visually analyze other information, such as average movement direction or speed, flow/vector fields can be computed. Travel directions are typically encoded with glyphs [Scheepens et al., 2014].

Temporal Perspective

Trajectory visualizations on maps implicitly include temporal information as the tracked positions are usually connected in temporal order, even when no timeframe and no timestamps are given. Other visualizations focus explicitly on time to answer questions about the temporal movement distribution, trends, seasons, and peaks. Different means can be applied, depending on the temporal resolution of the data and the analysis goals. Aigner et al. [2011] present a taxonomy and visualization examples for time-oriented data. Different timelines show the amounts of moving objects either for a certain point in time (linear timeline) or regarding repetitive behavior (cyclic timeline showing, e.g., an average day). Static timelines map time to space and help to perceive a temporal overview of different features of movement data. By contrast, mapping time to time results in an animation. While animation does not give overview and suffers from inattentional and change blindness [Jensen et al., 2011], it allows to review situations as one would observe them in realtime.

Spatio-Temporal Perspective

Often the relation of time and space is important, for example, to detect gatherings and relations between moving objects. To understand spatio-temporal relationships, one can map the temporal aspect to visual primitives such as shape, size, and color or provide an animation on a map. A common practice is to employ multiple coordinated views (MCV) [Roberts, 2007] that show spatial and temporal dimensions independently and coordinate these views via brushing and linking. Wang et al. [2013] present a MCV system for the analysis of traffic jams based on a taxi fleet. A third way to visually analyze spatio-temporal correlation are space-time cubes [MacEachren, 1995], firstly mentioned by Hägerstraand [1970]. Here, time is mapped to the third axis on a three-dimensional coordinate system. While space-time cubes have drawbacks such as perspective distortion and occlusion, they explicitly highlight the spatio-temporal distribution of the data. Kapler and Wright [2005] and Eccles et al. [2007] apply space-time cubes for the analysis of trajectory data in law enforcement tasks, mainly to detect events.

Event-based Perspective

In addition to position and time, trajectories can contain further attributes such as sensor reading from the movement object (mover) or any happening within the object's context. Both, a strong constancy as well as significant change of attribute values of one or multiple moving objects over time is named an event [Andrienko and Andrienko, 2013]. For vehicle data, an event can be any significant change in the control unit of a motor. An event can include a single person paying a bill or entering an office. It can also be a meeting of several people at a stadium or at a restaurant (social event). Geographic movement visualizations can be enhanced with information about events, encoded with color or glyphs at the corresponding position. As mentioned in the previous subsection, space-time cubes highlight spatio-temporal aspects of events. They, for example, implicitly represent meetings of moving objects by crossing trajectories which can be explicitly highlighted [Kapler and Wright, 2005, Eccles et al., 2007]. However, for some analysis goals geographic aspects are less important and the visualizations are more sufficient if they focus on the visual encoding of events, movers, and time. Node-link visualizations can be applied if transitions between places are of interest. Independent from the geographic space, they can represent Markov chains with transition probabilities between, e.g., visited places [Heuer et al., 2012, Andrienko et al., 2013b]. Streamgraphs [Havre et al., 2000] and Sankey diagrams (see [Tufte and Graves-Morris, 1983], [Cui et al., 2011], and [Byron and Wattenberg, 2008]) can show temporal transitions between events [von Landesberger et al., 2012].

Mover-based Perspective

The analysis can also focus on movers and their behavior. Questions comprise whether multiple moving objects are *flocking* around an area [Andrienko et al., 2013a], whether movers share similar features and paths, and if common and outlying behavior patterns can be identified. Analysts can utilize spatial, temporal, and abstract visualizations that are accompanied with various capabilities to filter for single movers or to extract groups of similar spatial, temporal, or event-based behavior. For the latter, scarf plots (e.g., [Richardson and Dale, 2005]) show individual sequences of events by drawing an item for each event in consecutive order, usually along the x-axis, for each object (mover) on the y-axis.

2.4 Semantic Analysis of Mobility Behavior

The word *semantic* has its origins in ancient Greek[4], where *sēmantikós* means *derived from sign* and refers to the *study of meaning*. These signs or *signifiers*, from which meaning can be derived, can comprise words, symbols, and phrases, and their denotation in a particular context. When movement data is recorded these contexts usually get lost. The installed sensors record position and time, but usually do not take into account other environmental and social context. However, behind every movement there is a purpose, be it the way to work, home, to the family, to friends, to a cultural event, or to the doctor. Context data that encodes this semantic meaning with additional signs may exist, but may not be linked and stored in the same database. To enrich movement data with context one can make use of the spatial

[4] www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057:entry=shmantiko/s, last accessed: March 22, 2017

and temporal autocorrelation of geographic data and related empirical observed gravity laws (see Section 2.3.3), which allows to infer dependencies of happenings with a nearby spatial and temporal position. Unfortunately, these assumptions simplify the reality where many forces are involved. The simplifications introduce many uncertainties that need to be dealt with by an insightful handling of any enriched movement data. Humans tend to belief what they see. For this reason, visualizations need to communicate uncertainties and raise analysts' awareness. This can be done with the use of color, with legends, with distortion such as blurring, and by giving the user the possibility to interactively adjust mapping thresholds concerning distances and probabilities. Depending on the uniqueness, the reliability, and the resolution of movement and context data sources, there will, however, always be a risk regarding wrong assumptions of dependencies. Semantic enrichment can thus be seen as *best effort*, which is often better than getting no insights at all. Recently, more ground truth data became available, which will allow to test and improve the enrichment significantly.

2.4.1 Definition of Context

Without having been defined explicitly, the word *context* has already often been used in this work. Schmidt et al. [1999] define context as *"describing the surrounding facts that add meaning"*. One important parameter is location, which is often used to create context-aware applications. Location based services rely on location information. This comprises location-based navigation, location-based billing, and location-based communication and dating. But according to Schmidt et al. [1999] *"there is more context than location"*. They distinguish context in human factors and physical environment. Human factors comprise information about the user, the social environment and task. User information, for example, can include knowledge of habits, emotional state, biophysiological conditions, etc. The physical environment consists of conditions, infrastructure, and location. All factors can be further subdivided into more precise context types. Figure 2.7 shows the context feature space. In the later part of their work, Schmidt et al. [1999] suggest to extract *cues* of sensor readings. A cue is defined as an abstraction from physical and logical sensors. A cue for a temperature sensor can be *hot*. Different cues build the context on a more abstract level. For example, *hot summer day* consists of at least temperature and season. Other examples for context are *in the office* or *jogging*.

2.4.2 Data from Social Media

Context data for the enrichment can be found through many different channels. With billions of messages, pictures, and videos as well as structured meta data, social media services are one of the largest sources for context data. Social media data is produced by individuals. Besides the core data, people share places, events, meetings, their observations, their opinions and moods. In a sense, social media users are social sensors that observe and interpret their environment and publish it to

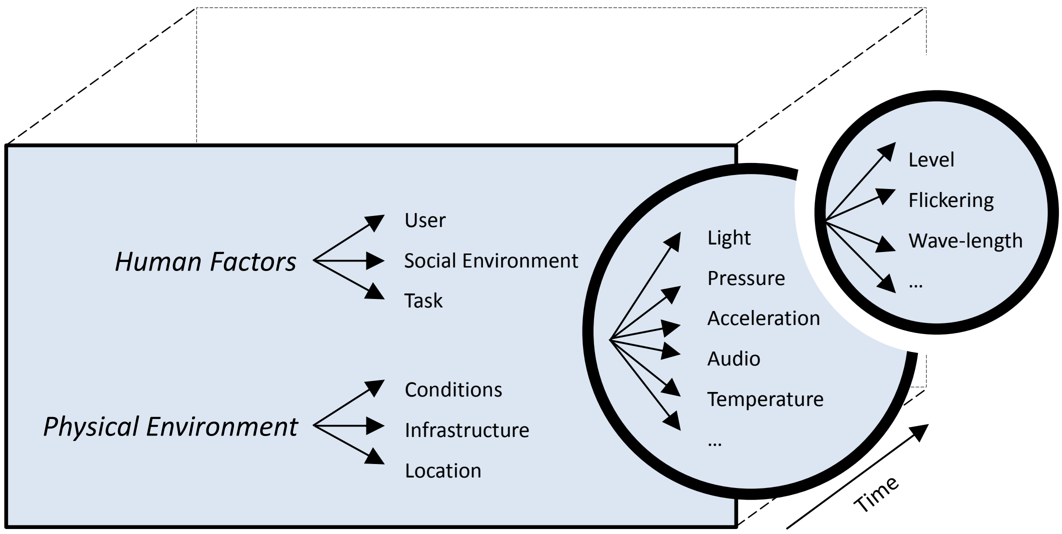


Figure 2.7 — The taxonomy by Schmidt et al. [1999] distinguishes context in a hierarchical manner, comprising human factors and physical environment.

the web. When information is shared via social media, it usually comprises content in different media types (e.g., text, images, videos), metadata such as information about the user, and a timestamp when this information is shared. Often, the data also contains location information. Location data can be of different abstraction level and ranges from precise information, e.g., latitude and longitude position, to less precise information, for example, the city or country a person is in. Many user interfaces of social media applications have location information as *opt-out* setting, which means that many users share their location information detected by their devices, even if they did not explicitly configure it this way.

Social media data can be used for two tasks. Firstly, arbitrary movement data can be enriched with context from social media nearby. For example, one can derive semantic information of a convention, a sports event, a demonstration, an epidemic, or a crime that takes place around the area and timespan of the movement sensor readings. For traffic data within a city, social sensors can give information of the causes of movement. For example, the cause of many people driving to the city's concert hall can be a famous musician giving a show. Many events appear periodically, e.g., in daily commute patterns, which can be explained by knowing the places, sentiments, and moods involved travelers share. Secondly, when the location information of a user is combined and ordered in time it is possible to reconstruct the user's movement. The reconstruction of movement from social media data is recently getting more and more attention, and researchers have started to understand its value. The data is up-to-date, often freely available, and covers worldwide mobility behavior of different scale and from different means of transportation.

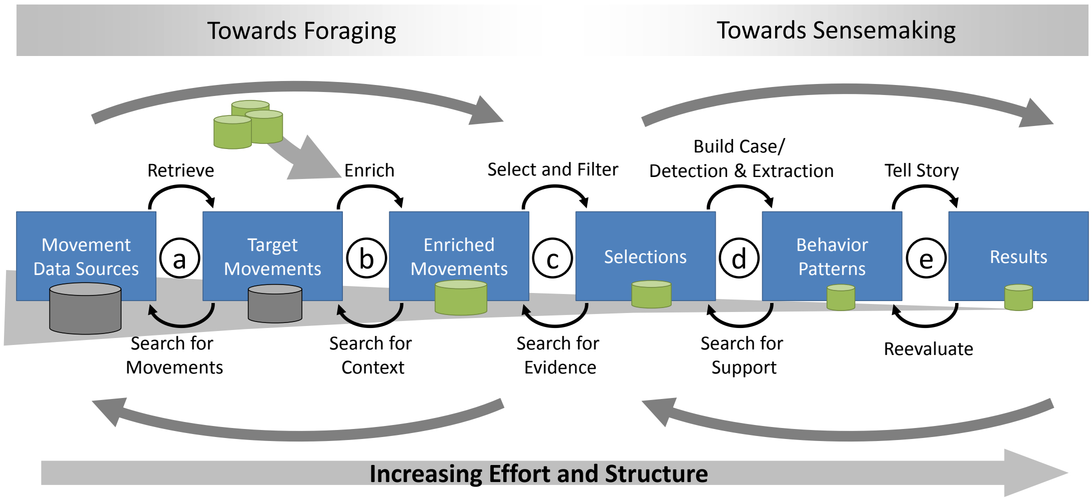


Figure 2.8 — The process model for semantic visual mobility analysis comprises a foraging and a sensemaking part, which flow seamlessly into one another.

2.4.3 The Analysis Process - Model and Instantiation

During this thesis, different approaches and systems were created for different datasets, application domains, and tasks (introduced in Section 2.5). Most of them basically followed a similar process pattern, which can be abstracted to a general model for a semantic analysis of human behavior (see Figure 2.8). The model is an adaption from Pirolli and Card's [2005] sensemaking process (see Section 2.2.2) that places the working process in the center. Boxes depict the datasets and its stepwise narrowing from a wide range of information with high recall to a narrow view of highly relevant findings. The process depicts both, exploratory and confirmatory analysis. Exploratory tasks and the creation of hypotheses run from data to theory (left to right) by stepwise combining and extracting the relevant. Retrieval and verification tasks, e.g., explicit querying, to find support for a hypothesis in the data, run from theory to data (right to left). The visual analytics model [Keim et al., 2008] is orthogonal to this working process. In each step of the analysis, data and models are closely coupled by highly interactive visual interfaces. Those build the bridge between cognitive and computational tasks. The analyst can steer the model building through these interfaces. At the same time, the analyst perceives information through the visualization.

As a first step, movement data needs to be acquired and loaded (see Figure 2.8, a). While in some cases the analyst starts with a suitable dataset at hand, there are cases where the right pieces need to be retrieved from a larger data source. This includes preprocessing steps like data cleansing from jitter and outliers, the handling of missing values, but also trip separation and categorization. Having the data at hand the analyst can inspect it and may refine the conditions to widen or narrow the amount of data. To cope with the vast data, aggregation approaches can be applied. The enrichment step (b) is an optional preceding step. Semantic enrichment fuses the movement data with additional context information from other data sources (see Section 2.4.1) and allows to get answers regarding movement reasons. This step significantly affects further analysis means and results. The analyst can thus come back to this stage at any time of the analysis, for example, to widen or narrow a set of rules for data fusion to control precision and recall. Especially, steps (a) and (b) can be seen as classical foraging steps where as much information as possible is collected and combined. After the enrichment process the analyst can search and filter the data (c) to explore spatial, temporal, and more abstract dependencies involving the semantic context. This bridges the gap between information foraging and sensemaking. During the exploration, patterns are found and hypotheses are created. From the explored subsets one can detect and extract these patterns that hold a compact description of the findings (d). Step (d) refers to the so called *dual search* and can be seen as a loop of hypotheses creation and verification. More precisely, analysts either explore and extract found semantic mobility patterns or they explicitly search for specific mobility patterns to verify if already existing hypotheses are supported by data cases. In both cases, the semantic enrichment leads to a different representation of the data where, in addition to geometric aspects of trajectories, sequences of happenings (events) are in focus. This affords new methods for exploration and verification. Lastly, the gained knowledge should be externalized and stored (e) to archive it and communicate it to other people in authority.

Based on practical approaches that are designed along this pipeline, the following chapters put focus on specific steps and highlight their individual importance. Figure 2.9 gives an overview of these approaches and their order in the remainder of this thesis. This aligns with the presented process model for semantic visual mobility analysis. In the next chapter (Chapter 3), movement data retrieval (see Figure 2.8, a) as well as the visual representation and inspection/validation of massive movement data are presented. More specifically, it is investigated how massive movement data can be reconstructed and visually explored using data from social media services. In Chapter 4 different approaches to enrich movement data with context (b), mostly from social media services, and to enable a semantic analysis are presented. Different scenarios ranging from electric mobility analysis to law enforcement are shown. For example, electric scooter data is enriched with point of interest information to find out more about their usage. After the enrichment, the data can be explored on a semantic level (c). Chapter 5 focuses on interactive exploration approaches by presenting a highly responsive filtering technique to drill down from overview to specific areas and places. This will be demonstrated on a large e-mobility fleet of electric scooters (see Section 2.5.2), but also on the movements extracted from Twitter. Step (d) addresses pattern detection and pattern-based querying (Chapter 6). Previously presented visualizations to understand temporal

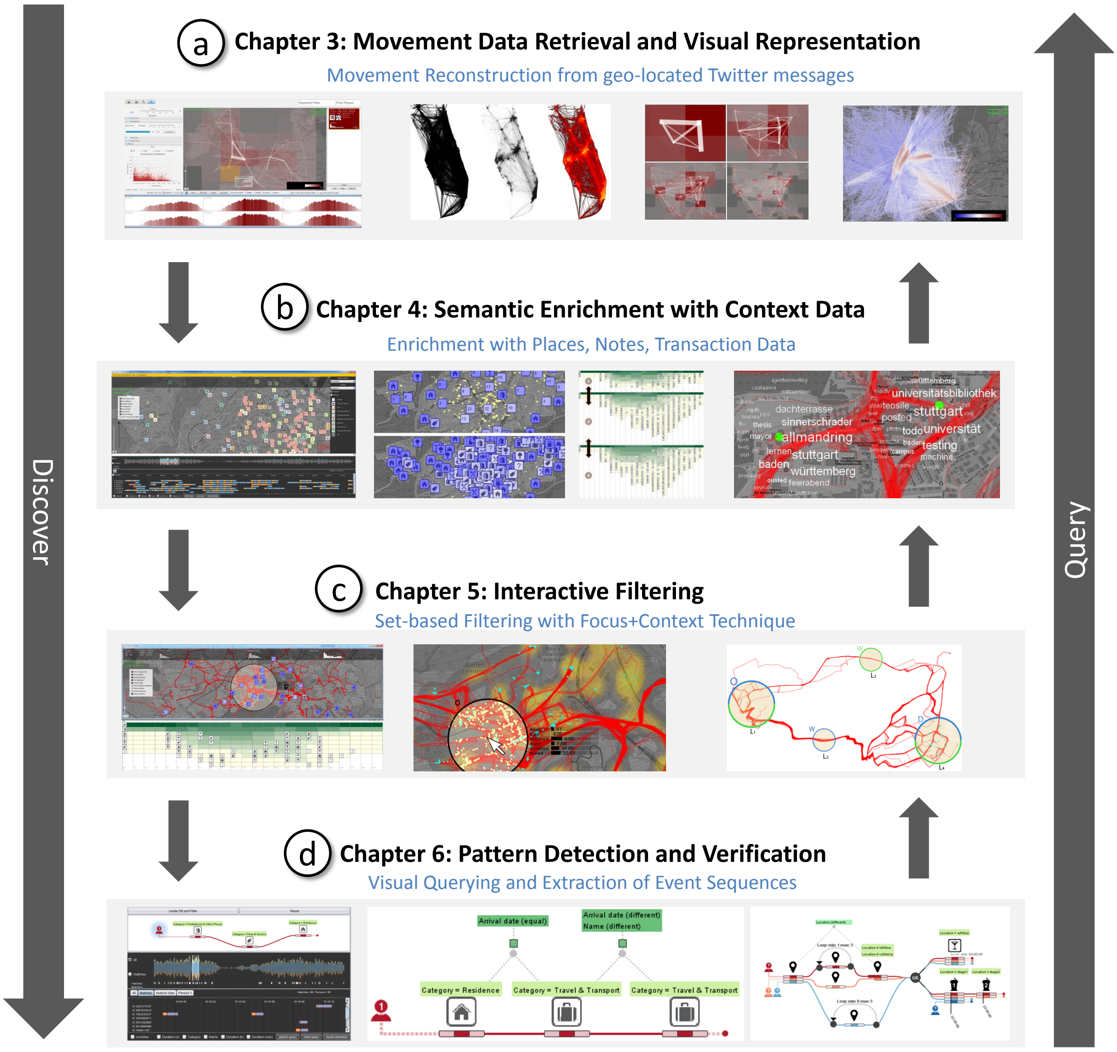


Figure 2.9 — Overview of visual analytics approaches of this thesis according to the process model for semantic human mobility analysis (Figure 2.8).

patterns, such as routines and anomalies, are enhanced with automatic approaches such as clustering, outlier detection, and distance-based sorting. A pattern-based visual language enables not only a compact description of extracted patterns for reporting and communication but also allows refinement and querying for patterns to verify hypotheses (e). Both, pattern detection and pattern-based querying are closely coupled to enable an iterative loop of data exploration and verification, as proposed by Sacha et al. [2014].

2.5 Application Scenarios

Previous sections motivated the needs for approaches and techniques that enable a semantic analysis of human mobility. Of equal importance are the application cases that afford and drive the approaches. Mobility analysis has a variety of applications (see Figure 2.10). In urban planning, for example, movement data helps to understand city dynamics. Knowledge about people's daily commute can be important for the construction of new shopping, living, and business areas. This also accounts for traffic planning, where mobility analysis can reveal bottlenecks in complex networks. Maintaining these networks is essential, because public and private transportation rely on an operative traffic system. To improve transportation and logistics, stakeholders can optimize schedules and coordination for public services such as buses and trains as well as private fleets, and make their systems more efficient and effective. However, the performance of single vehicles is of interest, too. To stay competitive, the automotive industry tests car performance and exhaustion in field studies and analyzes customer satisfaction by investigating mobility profiles of drivers. This bridges the gap to geo-marketing and demographics. By knowing which customers to expect at which locations, companies can place advertisement and plan new branches. Demographic patterns about travelers, e.g., age, nationality, and level of education, are also of interest in politics, where decision makers have to deal with migration and resulting challenges for security as well as for health care and health prevention. Here, movement data serves as a basis for epidemic simulations. With knowledge about the connectivity of suburbs, cities, and countries one can estimate the spreading of disease and other processes.

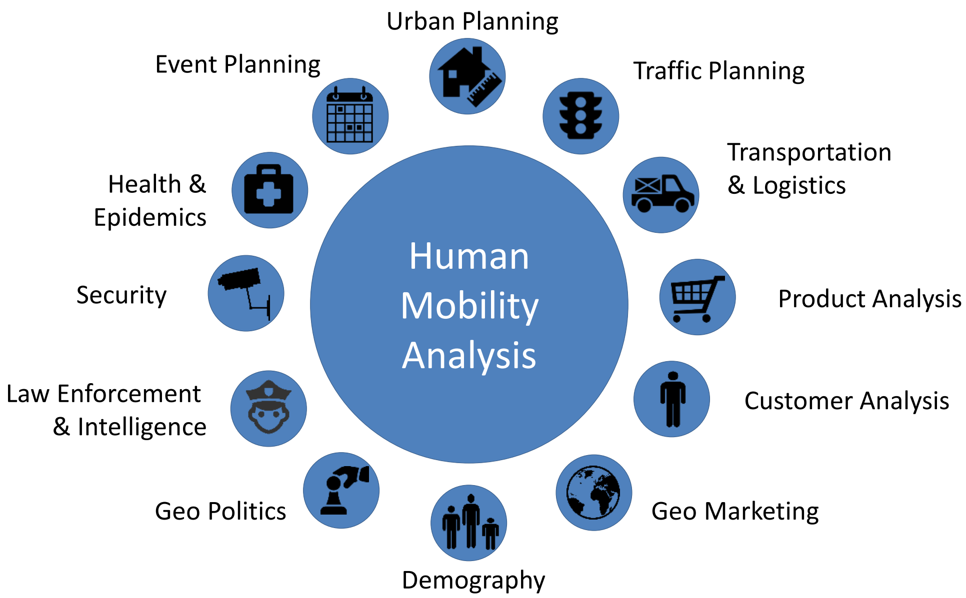


Figure 2.10 — Human mobility analysis has many application areas ranging from urban planning, via product and customer analysis, to law enforcement.

The scenarios presented in this thesis address only a subset of this wide range of application areas. In the remainder of this thesis the following scenarios give motivation and serve as exemplary cases to evaluate the proposed means.

2.5.1 Situation Awareness through Social Media

Situation awareness plays an important role in domains where stakeholders are in charge of making decisions and have to react to certain circumstances. This, for example, accounts for event management, crisis response, and the handling of global travel and migration. To improve decision making, an up-to-date information basis is essential. Unfortunately, information about mobility patterns is often monetized by large companies, hard to retrieve, and distributed in heterogeneous formats, creating a patchwork of data.

Twitter Data: As part of this thesis, social media data is evaluated as additional data source for movement and context data to improve situation awareness. Microblog platforms such as Twitter and Sina Weibo allow users to contribute short-term messages to a global audience. These messages contain information about people's activities and observations, about relations to others, and sentiments. Twitter is one of the largest microblog services. From about 500 mio. messages daily, approximately 1-2 % contain precise geo-location information. By connecting the locations of users, large trajectory datasets can be extracted, that cover up-to-date movement patterns of varying scale around the globe. In combination with the message content it can reveal social gatherings and other events that cause and effect temporal changes in human mobility. Besides urban planning, the vast movement data could also improve the understanding of migration streams as well as catastrophes such as terrorism, mass panics, and disease spreading. Different challenges arise due to the data specifics such as its low spatio-temporal resolution and resulting uncertainty, large volumes, and uneven spatial distribution.

2.5.2 Acceptance Evaluation of Electric Mobility

To stay competitive companies constantly search for and evaluate new markets. With global warming, increasing environmental awareness, decreasing resources, and funding by the government, the field of electric mobility gets more and more attractive as a new business model for various companies from the field of vehicle manufacturing and energy consumption. Germany plans to have around one million electric vehicles till 2020 and to become a leader in this new field of technology. Besides environmental goals, however, companies have to evaluate if business models are successful before they invest on a grand scale.

Electric Scooter Data: In the summer of 2010, a major German energy provider started a field study with 500 electric scooters in Stuttgart, Germany [EnBW AG, 2010a], creating the largest electric fleet in the country [EnBW AG, 2010b] at that time. The goal of the study was to assess if and how electric mobility could find its way into everyday life. The study participants received their scooters for one year of free usage under the condition that they agreed to their trips being tracked. The recording began in July 2010 and ended in August 2011 with measurements taken approximately every 30 seconds during the rides. Each of the measurements includes the scooter ID, GPS position, time of measurement, speed, battery charge level, and trip ID. In addition, a dataset with recorded battery chargings as well as a dataset with trip comments was given. The latter was created by the participants using a web-interface to comment on their trips afterwards. All three datasets are connected via the scooter and trip IDs. By summer 2011, a huge dataset had been collected, containing about 8 million measurements adding up to 155,000 trips with nearly a million kilometers traveled. As can be seen in Figure 2.11 besides charging behavior, data about charging stations is available as well. An expert interview revealed that basic statistics were created from the data but left many questions towards the original goal unanswered. A main question was if e-mobility can find a way into everyday life and if e-mobility can be seen as a promising future business case. Hence, the experts found it highly relevant to learn about what kind of activities people used their scooters for, and what they felt about it. They were interested to find out when scooters were used regarding time of day, week, and seasons, in which areas they were accepted, and what the specific advantages and disadvantages of using them as the preferred means of transportation were.

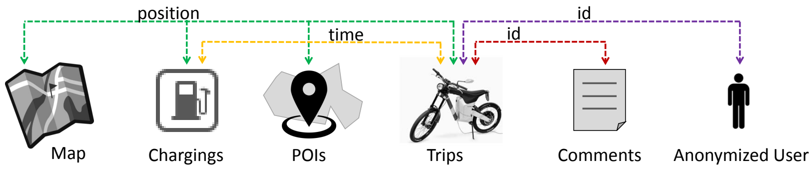


Figure 2.11 — Data readings of the electric scooter data. Points of interest (POI) data was derived in a semantic enrichment process using web 2.0 services.

The application scenario of electric mobility was one of the main motivations and driving factors to create the approaches presented in this thesis. Hence, in many of the following chapters, the scenario is present. It serves as a motivation but is also used to evaluate the contributed approaches in case and user studies.

2.5.3 Analysis of Visitor Behavior at Indoor Events

Anticipating the needs of customers and visitors by tracing and predicting their movements can be of great value, not only for transportation business and car industry, but also for stores and planning events, such as exhibitions, festivals, and conferences. With knowledge about visitor streams and room occupancy, managers and promoters can optimize the layout of show rooms, time schedules, and marketing strategies. For large malls and supermarkets, customer movement information reveals which shops and shelves are visited frequently and helps to understand why customers prefer or avoid certain areas. Detecting typical groups of people with movement, for example, can help shop managers to react to the needs of these groups by placing products they are presumably interested in, in the same area. For events, it is also beneficial to know about visitor movements and correspondingly deduced interests for prevention of crowded places and evacuation planning. With the rising popularity of mobile devices, people are increasingly using location based services and are also willing to reveal mobility details in exchange for personalized information. Indoors, for example, customers welcome free WiFi or navigation services and may accept advertisements or agree to position tracking.

re:publica 2013 Data: Big effort for indoor tracking was made at the *re:publica*[5] conference in 2013. The *re:publica* is a large annual event on Web 2.0 innovations with up to 10,000 visitors. Hot topics presented and discussed here deal with social media platforms and habits, but also with business ideas and the digital society in general. In 2013, the conference was held in Berlin, Germany, on two floors of a former train station. All conference rooms were equipped with WiFi access points and all attendees had free wireless internet access. In return, visitors signed an agreement that allowed their mobile devices to be tracked and the results to be used for analysis and research purposes. This way, a large dataset was created, containing information of over 6,700 visitors' movements. A floor plan gives information about the rooms on two different levels. In addition, a time schedule of the conference contains information on the talks: in which room they were held in, the session topic, and the speakers. Being a social media conference, the speakers are present on the web, and additional information about the popularity of the persons and the presentation topic, etc. could be derived. To analyze the data, different pieces of the puzzle have to be brought together (see Figure 2.12).

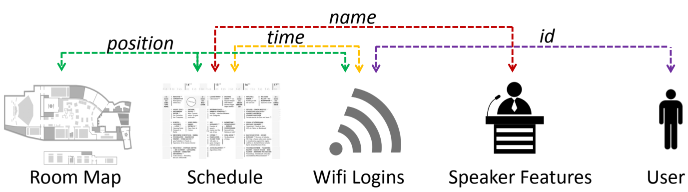


Figure 2.12 — Different data readings of the re:publica 2013 data.

2.5.4 Crime Detection for Law Enforcement

Visual Analytics has its roots in intelligence analysis and law enforcement. For crime investigation detectives need sophisticated tools that allow to collect and fuse all evidences and give a holistic view on the situation. Spatio-temporal correlations, daily routines, and outlying behavior such as meetings of suspects can lead to important insights. Pirolli and Card [2005] have generalized the schemata in intelligence analysis, as discussed in Section 2.2.2. The VAST Challenge is an annual competition in the visual analytics research community and colocated to the IEEE

[5] https://13.re-publica.de, last accessed: March 23, 2017

VIS conference, the main forum for visualization and visual analytics. Every year, a synthetic dataset is created. In the massive, highdimensional, and heterogeneous datasets different hidden patterns give cues about what has happened. In the year 2014 and 2015 the task was to investigate crime.

The Kronos Incident–The Vast Challenge 2014: In 2014, the challenge dealt with the fictitious island country of Kronos[6]. During a celebration of employees of a company named GAStech different people had gone missing. Among different questions to answer, the main task was to understand the daily routines of the employees as well as abnormal behavior that gives hints about the disappearance. For the analysis different datasets were given. A tourist map contained a rough imaginary description about points of interest such as the company's factory, home locations, parks, groceries, ports, and restaurants. A road network covered the main connections of the city. In addition, the GAStech employees working titles and GPS traces from rented cars were given as well as credit card and loyalty card payments (see Figure 2.13). One major difficulty was to align the data and deal with the spatio-temporal resolutions, inaccuracies, and uncertainties.

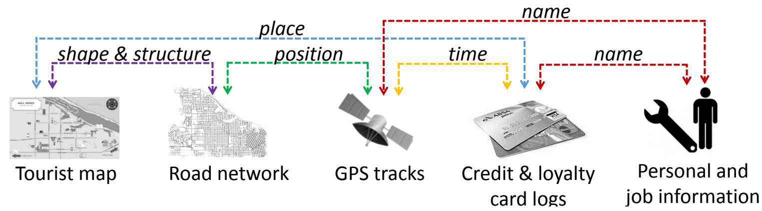


Figure 2.13 — Different heterogeneous datasets of the VAST Challenge 2014.

DinoFun World–The Vast Challenge 2015: The VAST Challenge 2015[7] dealt with law enforcement on a criminal activity happening in a theme park. In addition, the task was to understand the parks' main routines and visitor group behavior. Similar to the 2014 challenge, many different heterogeneous datasets, such as the visitor traces, mobile phone connections, a park map, and check-in information to attractions, were given (see Figure 2.14). The data contained a three day timespan of around 11,000 visitors. A backstory included additional information such as the visit of a soccer star and the disappearance of valuables.

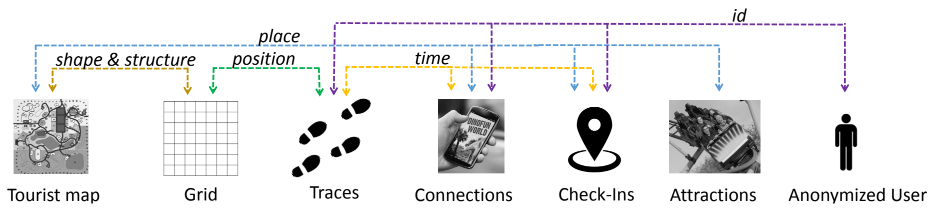


Figure 2.14 — Different heterogeneous datasets of the VAST Challenge 2015.

[6] www.vacommunity.org/VAST+Challenge+2014, last accessed: May 4, 2017
[7] www.vacommunity.org/VAST+Challenge+2015, last accessed: May 4, 2017

CHAPTER 3

Movement Data Retrieval and Visual Representation

Mobility tracking has become a commodity. Massive datasets are recorded that cover human mobility behavior of different scales. When global movement is of interest, analysts usually derive data from flight schedules. On a more local scale, cell phone and public transportation data provide valuable information of urban movement such as daily commute [Ratti et al., 2006]. Unfortunately, these datasets are often incomplete and of limited temporal and geographical resolution and scale. They can be expensive and hard to retrieve because they are monetized by large companies or cover privacy-sensitive data. With the rise of smartphones and social media, humans have become sensors themselves, contributing geo-located data points every day. On microblog platforms, many users send public messages with location information that can be used to retrieve a trajectory estimating the users' movement. Twitter users, for example, produce more than 200 million geo-located messages a month [Thom et al., 2014]. Based on this data, one can reconstruct movements of different scale, from global air traffic to local ground transportation. The data is available for large timespans and often free to record and analyze.

Recently, researchers have started to interpret movements reconstructed from microblog platforms [Blanford et al., 2015, Chae et al., 2015, Chen et al., 2014, 2016, Fuchs et al., 2013, Sakaki et al., 2010, Senaratne et al., 2014, von Landesberger et al., 2016] (see Section 3.1.1). While these approaches already reconstruct and classify trajectories, there are still many open research questions and challenges: Are the reconstructed trajectories reflecting real world patterns, comparable to arbitrary movement data? What similarities and differences exist when comparing

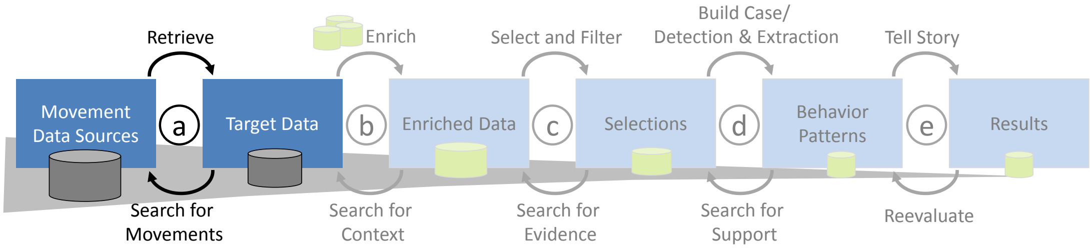


Figure 3.1 — This chapter mainly addresses visual data retrieval (first two highlighted stages) but also presents means for a visual representation of raw trajectory data.

such heterogeneous datasets and how can they be made visible? How to deal with varying spatial and temporal granularities, distributions, and volumes?

This chapter presents visual means to address these questions by taking into account the data specifics (Section 3.2). A major part lies in the visual data retrieval, but it also presents means for visual representation and inspection (see Figure 3.1). At first, data is retrieved and preprocessed to reconstruct massive amounts of movements (Section 3.3). Subsequently, different approaches to visualize the raw movement data are proposed, comprising alpha blending, edge splatting, and animation. Moreover, a hierarchical aggregation method is presented that allows to get an overview and investigate different levels of detail (Section 3.4). Comparative means to evaluate the data with respect to arbitrary movement data sources are discussed in Section 3.5. The applicability of the approaches is presented in Section 3.6. At first, an exemplary evaluation reveals how the reconstructed movements reflect human mobility patterns by comparing the movements to city commute and to flight schedules (Section 3.6.1). Thereafter, two case studies show the applicability to explore events (Section 3.6.2). They focus on the pilgrimage to Mecca in September 2015, where hotspots and travel directions are identified, and on a large scale sports event, the Super Bowl 2013 in New Orleans.

The approach can be considered as a step towards a better understanding of movement patterns retrieved from microblog messages and as an enabler for a comparative analysis (see Section 3.5) of such data.

Parts of this chapter have been previously published in:

R. Krüger, G. Sun, F. Beck, R. Liang, and T. Ertl. TravelDiff: Visual comparison analytics for massive movement patterns derived from Twitter. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 176–183. IEEE Computer Society, 2016b

3.1 Background

The background of this chapter is two-fold. In the first part, related work for movement reconstruction from microblog messages, its visualization, and its analysis is presented. Subsequently, visual means for comparative analysis are discussed.

3.1.1 Movements from Microblog Data

While the visual analysis of textual content from social media is a well-developed research field, the integration of geographic information recently gained more and more attention. *Senseplace2* [MacEachren et al., 2011], *ScatterBlogs2* [Bosch et al., 2013], and approaches by Fuchs et al. [2013] allow for a location-based analysis of Twitter messages. Others even trace the geographic path and evolution of topics over time: Sakaki et al. [2010] extract earthquake related microblog messages and apply a Kalman filter to reconstruct an earthquake's trajectory. Similarly, Senaratne et al. [2014] look for hotspots of a certain topic and reconstruct a trajectory of a musician's concert tour across the USA.

Besides content analysis there is new research that investigates human movement derived from geo-located microblog data. Gabrielli et al. [2014] present an analysis of movements from Twitter. They focus on the enrichment and leave details about the reconstruction open. They cope with much smaller volumes and do not integrate the human into the analysis loop but use visualization for result presentation only. Chae et al. [2015] reconstruct movements of Twitter users and map them to the road network to overcome data irregularities. They further present an expert-driven interactive analysis to identify outliers. Interactive filtering of movements derived from Twitter is also shown by Chen et al. [2014]. In their work, origin–destination (OD) patterns can be found with a *focus+context* technique. Chen et al. [2016] also developed a full-fledged system for the analysis of movements from Sina Weibo, a Chinese microblogging platform. They address major challenges such as large data volumes, irregular sampling, data sparsity, and resulting uncertainty. A Gaussian-mixture model is used to categorize movements to different means of transportation. The approach presented in this chapter focuses more on data comparison but also comprises a categorization method to classify means of transportation. By contrast to the method from Chen et al. [2014], it is rule-based and allows for an interactive trade-off between precision and recall. With a more regional and application-specific focus, Blanford et al. [2015] analyze cross-border movement in Africa derived from Twitter and build graph models to aggregate the movements. *Flow Sampler* by Chua et al. [2015] also adapts a graph-based method and infers pathways from geo-located microblog data to explore movement flow. Recently, von Landesberger et al. [2016] contributed another approach to analyze temporal changes. While they also present a comparison graph, they do not provide a hierarchical structure that allows for a smooth transition between overview and detail. They focus on Twitter analysis only and do not compare the patterns to other data.

3.1.2 Comparison of Trajectories and Graphs

In their taxonomy of visual comparison, Gleicher et al. [2011] discern three types of comparison methods: *juxtaposition*, *overlay*, and *explicit encoding*. These general approaches are applicable to many visualization problems but have different advantages and limitations. For a visual comparison of trajectories, one could juxtapose small multiples, each representing a different version of the dataset [Kehrer et al., 2013]. Also, different groups of trajectories can be overlaid while discerning them by coloring [Andrienko and Andrienko, 2012] or 3D stacking [Andrienko and Andrienko, 2012, Tominski et al., 2012]. Overlay or juxtaposition approaches work well to reveal coarse differences. However, overlays get difficult to read if too many visual entities overlap and, in juxtaposed images, only large differences can be retrieved easily, while identifying small changes becomes a *spot-the-difference game*. Since the comparison is mostly about rather subtle differences between variants of similar movement networks, explicit encoding is suitable: differences and commonalities are explicitly highlighted in order to guide the viewer to interesting patterns.

To gain an overview of massive trajectory data, graphs provide the necessary level of abstraction by aggregating trajectories to transitions [Andrienko and Andrienko, 2012]. Various approaches for visual graph comparison have already been proposed [Andrews et al., 2009, von Landesberger et al., 2011]. Also related are dynamic graph visualization techniques, which compare changing graphs over time [von Landesberger et al., 2011]. Among the comparison approaches, some employ explicit encodings: In node-link diagrams, differences are usually encoded with colors [Archambault, 2009, Archambault et al., 2011, Barsky et al., Schipper et al., 2009]. Also, adjacency matrix representations show differences between two graphs using color [Beck and Diehl, 2013] or glyphs [Ma et al., 2015]. In matrix representations, however, the geographic information, which is important for this analysis scenario, is not preserved. In an interactive comparison approach, Beck et al. [2011] compute union, intersection, and difference only on demand and thereby let the user control what should be visualized.

The approach presented in this chapter applies similar interactive means but in a hierarchical manner, comparable to Abello and van Ham [2004], giving the user even more control over the merging and filtering process of trajectories. This allows both, the visual analysis of a single group of trajectories as well as the visual comparison of two groups of trajectories in a *diff* view.

3.2 Data Specifics and Design Rationales

Geo-located Twitter data and its reconstructed movements have certain peculiarities that require special treatment. Four specifics (S1-S4) and resulting requirements are identified that go beyond common movement analysis practices:

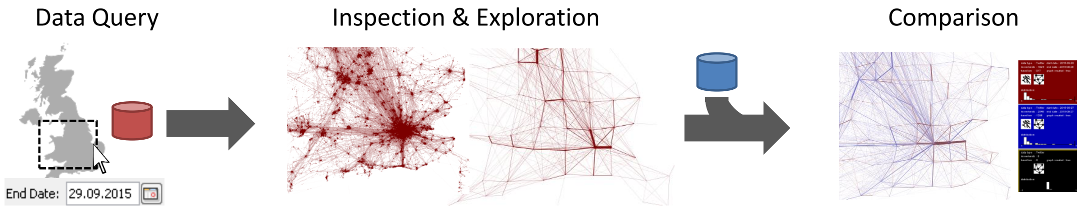


Figure 3.2 — The abstract analysis process consists of an exploration part and a comparison part. Colors indicate an exploratory analysis for two different datasets, that precedes the comparative analysis.

S1 **Implicit and unlabeled data.** Movements need to be reconstructed from geo-located microblog messages. This requires several preprocessing steps, including bot removal and categorization of means of transportation.

S2 **Massive and unstructured.** From about seven million tweets a day, the reconstructed movement data is massive but often only consists of origin and destination points. Hence, movements do not follow an underlying network (e.g. roads). This can result in overplotting and cluttering. Clear visualizations that preserve the density distribution are needed.

S3 **Varying scale and resolution.** Analysts are interested in movements of different scale, from commutes to intercontinental flights. However, varying by region, data resolution is low. Detailed visualization of uncertain movement can be misleading. Adaptive data aggregation helps to solve the challenge.

S4 **Uneven distribution and unknown quantities.** Different data has different characteristics, such as resolution, distribution, and structure. Moreover, it is difficult to compare quantities derived from microblogs to real world quantities such as number of travelers. Normalization and comparison methods that handle different data distribution and outliers are needed.

With respect to these data specifics and resulting requirements, the approaches presented in the following orchestrate visual means to explore and compare large movement data of different transportation means with multiple coordinated views (MCV). Using comparison visualizations, the data can be contrasted to other datasets or analyzed regarding seasonal changes and events in the microblog movements. All views are connected with brushing and linking. North and Shneiderman [2000] showed in a user study that multiple coordinated views can improve user performance in data analysis tasks by 30–80% over detail-only and uncoordinated interfaces for most tasks. Before the design rationales are explained (Sections 3.3 to 3.5) based on the identified requirements, this section gives a brief overview of the analysis process, illustrated in Figure 3.2.

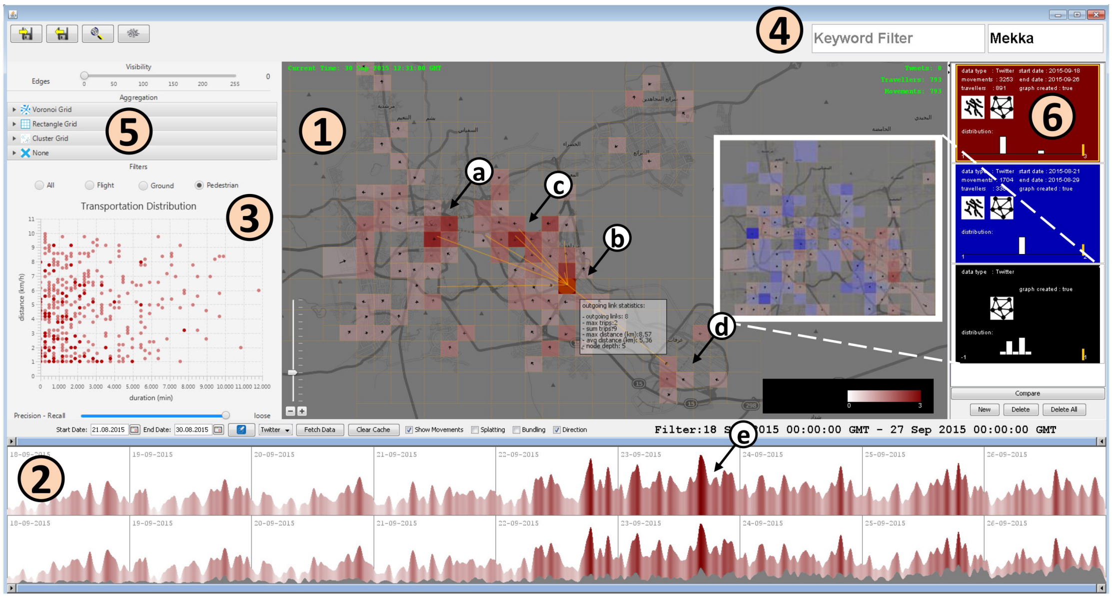


Figure 3.3 — The analysis components for movement data retrieval, exploration, and comparison: (1) map; (2) timeline; (3) velocity distribution & filtering; (4) searchboxes; (5) grid creation panel; (6) comparison interface. The map shows distribution and direction of pilgrim movements in Mecca, Sep. 2015. Patterns (labeled a–e) are described in a case study (Section 3.6.2).

The analyst starts the exploration process by selecting a region and time of interest. The user interface that comprises multiple linked views (see Figure 3.3) therefore provides a polygon tool and a calendar widget. After data querying and loading, the movements can be explored. The analyst may first take a look at the map, located in the center (see Figure 3.3, 1), exploring spatial movement distribution and flow of messages. The map provides both detail visualizations for investigating the raw data and layers aggregating data for overview and advanced filtering. The timeline view (2) allows for temporal navigation and analysis of time-dependent patterns. Depending on the analysis goals, the movements might be filtered by means of transportation; a scatterplot (3) shows the velocity distribution for the results. The approach supports keyword filters and a geographic search to quickly jump to a specific location (4). Besides exploring the raw movement data, the grid configuration panel allows to aggregate the data to a hierarchical graph model (5). For data comparison, a second dataset can be loaded, e.g., Twitter, flight, or taxi data. The interface to the right (6) allows switching between the two datasets and a comparison visualization. Comparisons are either done on a raw data level or in the aggregated model.

3.3 Data Retrieval and Preprocessing

To reconstruct movements from Twitter messages (S1), data from the official Twitter webservice is recorded and stored in a local database. The data is then retrieved locally and reconstructed to trajectories that need to be filtered regarding bot messages. Finally, the data can be categorized by estimating the means of transportation. In the following, these steps are described in detail.

3.3.1 Storage and Retrieval

Twitter is the largest microblog platform in the web and has around 313 million active users a month that send more than 500 million messages a day. The company allows access to their data via two different APIs, the Twitter REST API and the Twitter Streaming API. The first allows to query the database for messages containing keywords, about certain users, and connections like friendships and followers in the network. The latter gives access to live data, i.e., messages that are currently posted. Some of these messages contain not only information about the user, text, and timestamp but also detailed spatial information (latitude, longitude) where the message was written. To collect the incoming stream a local service was built that stores geo-located messages in a Lucene[1] database. The NoSQL database allows for fast queries using spatial, temporal, and text-based index structures. This data serves as the basis for the reconstruction of movement trajectories.

3.3.2 Extraction and Reconstruction

Once the data is retrieved from the local database the messages can be used to reconstruct the users' movements. This is done by user-based and temporal ordering. Afterwards, different bot-removal strategies are applied to filter out spam and content from non-human users.

Trajectory Composition Around 1–2% of the messages from Twitter contain detailed geo-location consisting of latitude and longitude information. With respect to overall volumes of around 500 million tweets a day[2], it still sums up to around 200 million geo-located messages per month. Many users share messages frequently, i.e., several times a day or a week. From their location and temporal information, movement trajectories can be derived. A description of a trajectory is given in Section 2.3.2. In this specific case p_i is the geo-location of a message at time t_i. There is a varying number of messages that cannot be connected to any movement. The number of reconstructed trajectories highly depends on the selected area (region and size) and timeframe. On average, the algorithm is able to connect about 30% of the loaded messages to trajectories when the timeframe is at least a day.

[1] https://lucene.apache.org/core/, last accessed: February 21, 2017
[2] https://about.twitter.com/company/, last accessed: December 12, 2017

Bot Detection and Filtering Twitter contains a large number of computer-generated (bot) messages. While not all bot messages can be removed, their amount can be significantly reduced by applying different filters. Content-based filters are widely used for bot removal in microblog data. This approach makes use of basic techniques, based on the observations by Lee et al. [2010]. It includes a dictionary of keywords, such as *weather report* and *special offer*, defining a black-list to filter out news and ads. Also, user accounts with high word and sentence repetition are excluded. In the future, more elaborated techniques can be integrated [Chu et al., 2012, Cao et al., 2016]. Besides content, certain spatio-temporal patterns relate to bots. One can assume that human users tweet at maximum with an average daily frequency of two messages per hour. Higher frequencies (>48 tweets a day) usually indicate bots (∼3–5%). Also, even fast passenger aircrafts do not exceed a speed of 1,500 km/h. If the derived movement speed of subsequent tweets exceeds this threshold multiple times, the corresponding user is excluded, because it is presumably a bot.

3.3.3 Movement Categorization

Often, there are days between two consecutive messages of a user and it is unknown where the user has been in between or if the user has arrived much earlier at the second location than the message was sent. Thus, classification of the used means of transportation is a non-trivial problem. To infer the most likely means, different categorization rules can be applied. Because the data sparsity and resulting uncertainty is too high for a finer-grained classification, the here presented approach splits the movements into three main categories: pedestrian, ground transportation, and flight. Still, it is challenging to define a clear boundary among different transportation categories. A solution is to apply a set of heuristics. Humans tend to walk at about 5 km/h [Browning et al., 2006], up to 9 km/h [Minetti, 2000]. Similarly, for ground transportation a distance threshold can be set to 300 km and above 5 km, the upper pedestrian walking distance. A fast passenger aircraft does not exceed a speed of 1,500 km/h and the maximum distance of a flight cannot exceed 20,000 km, as this is half the circumference of the earth and not possible without refuel. Exemplary results of the categorization can be seen in Figure 3.4. However, depending on the region and timespan to be analyzed, this strict categorization can lead to a small data sample that is susceptible to outliers. Hence, interactive means are provided to widen filter settings and let the analyst choose a trade-off between precision (more reliable categorization) and recall (higher data volumes). Figure 3.5 shows the interactive means to control precision and recall with immediate feedback.

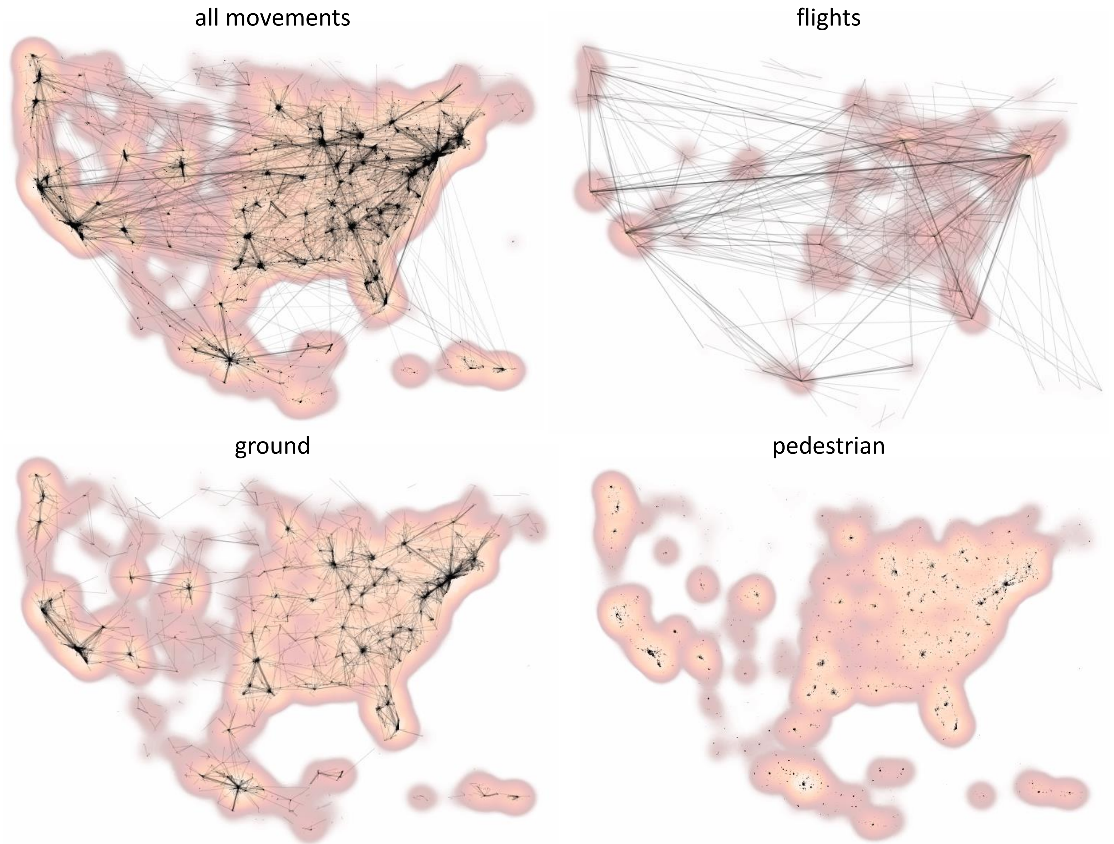


Figure 3.4 — The outcome of the categorization on distance and speed for two days Twitter data from North and Central America (overall 200,000 trajectories). The lines show reconstructed trajectories over larger distances from flights, medium distances for ground transportation and shorter distances for pedestrian movement. The heatmap in the background highlights regions with dense tweet coverage.

3.4 Exploratory Visualization

With respect to the data specifics, defined in Section 3.2, four approaches to visualize the reconstructed movements are proposed that fit different application cases. At first, animation is introduced to replay the movements and get an awareness of the situation happening in a selected timespan and region. To overcome occlusion and to address varying data distribution, two density-based visualization methods are presented. Thereafter, an aggregation technique that allows for abstraction and stepwise visual transition from overview to detail is shown. Closely coupled with a timeline component and different filters, all approaches can be interactively steered by the user.

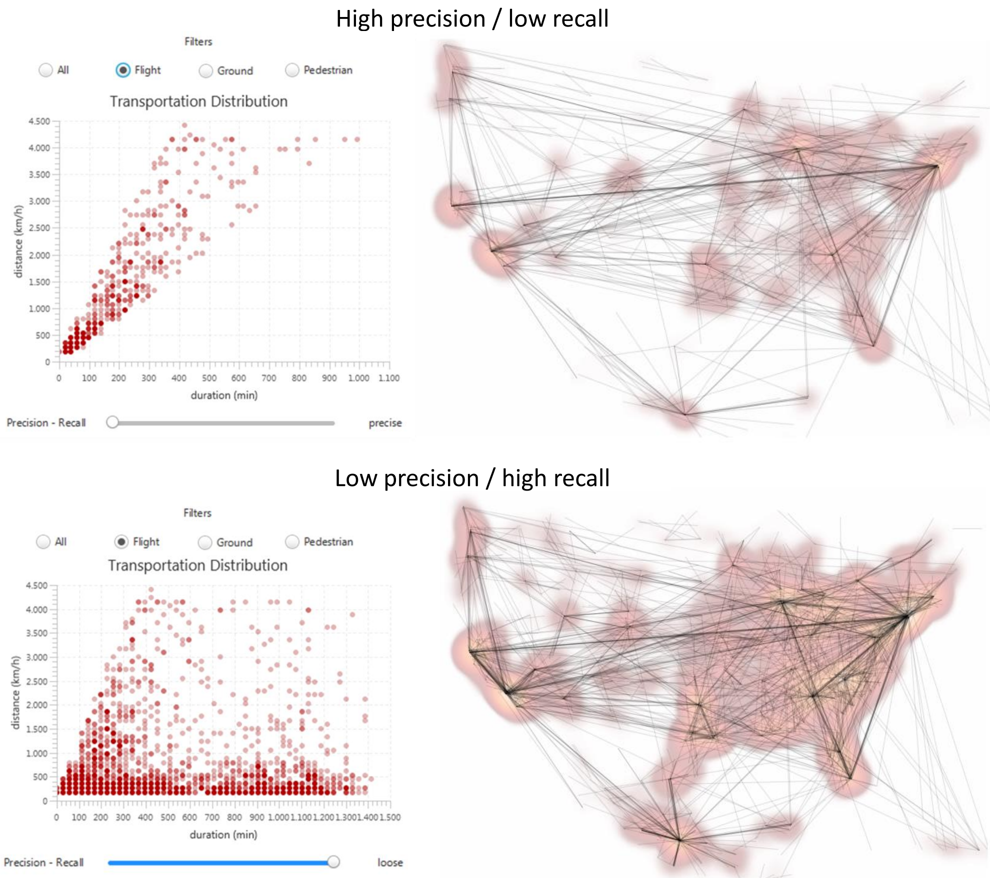


Figure 3.5 — The recall for any categorization can be interactively increased by widening the restrictions on speed and distances. This decreases precision. The trade-off is controllable with a slider. Top) 516 movements categorized as flights; Bottom) 1,813 movements categorized as flights.

3.4.1 Animation

Animation can be seen as a direct replay of movement. Card et al. [1999] state that *"animation means movement"* and *"is desired for the visualization of dynamic, complex processes"*. They also advice that *"animation should be part of any standard visualization tool kit"*. On the other hand, many researchers stated that animations violate the so called *apprehension principle* [Tversky et al., 2002] that says that visualizations should be accurately perceived and appropriately conceived. Animations are often too complex and suffer from inattention blindness and change blindness issues. How should this divergent discussion be interpreted? As for most things, it depends on the task. Tversky et al. [2002] discuss animation to be more effective than static graphics in situations where analysts do not want to convey complex systems, but for example for real time reorientation in time and space. If they are used judiciously, they can be effective to communicate movement directions and

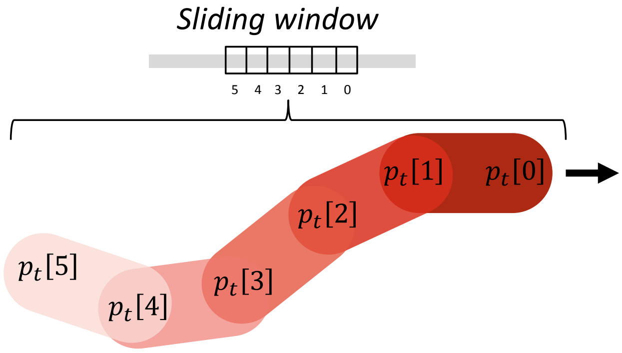
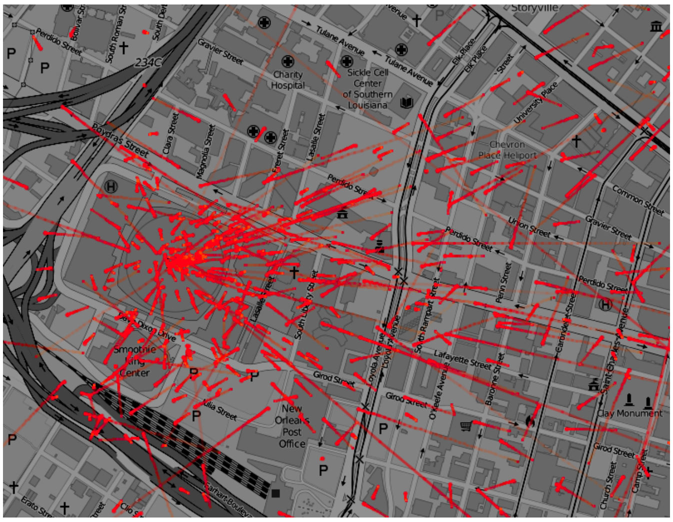


Figure 3.6 — Animation is achieved by interpolation. The current position $p_t[0]$ is visualized with full opacity. Earlier positions are kept in a sliding window for n iterations (here 6) and are drawn with decreasing opacity.

paths. Scheepens et al. [2016] have shown that animation can be supportive in movement analysis to understand main traffic flows. They use a particle system to show the spatial extent and direction of trajectories and show that this provides strong clues with little clutter.

Similarly, in this work, animation is applied to get orientation of a situation in the data as a direct replay of what has taken place. This can help to get a first glance of main movement between countries but also to (re)view micro situations of an event. Because the reconstructed trajectories do not follow an underlying network (S2), clutter and overlap hinder the understanding of movement directions and flow. While animation does not give a holistic overview, it shows movement flow in a clear way. The approach uses linear interpolation between two known measurements $m_{i-1} = (p_{i-1}, t_{i-1})$ and $m_i = (p_i, t_i)$. The animation redraws n times a second. To highlight the movement direction and path (where does an object come from and where does it go to) the position can be calculated with a sliding window and visualized with increasing transparency over time, as sketched in Figure 3.6. In the manner of a video player the analyst can steer the playback speed and playback direction (forward/reverse) via user interface. Figure 3.7 shows a snapshot of an animation of reconstructed movements from Twitter data in New Orleans during the Super Bowl in 2013. One can easily identify the main movement direction from the inner city to the Mercedes-Benz Superdome (left).



◄ **Figure 3.7** — Snapshot of an animation of reconstructed trajectories from microblog messages during the Super Bowl 2013 in New Orleans. Movement directions clearly become visible.

3.4.2 Density-based Approaches

A straight-forward way to give an overview on the reconstructed movements (S2) would be to simply draw trajectories on the map. Unfortunately, this results in clutter and overplotting (see Figure 3.8, left). A more scalable approach to visualize the movements is alpha blending (center). By decreasing the opacity of the trajectories, visual clutter can be reduced and trajectories with high density are clearly conveyed. However, alpha blending introduces new problems. Firstly, the quantity of trajectories to be drawn largely depends on the selected area and timespan. Users need to adjust the opacity when drawing trajectories under different circumstances to avoid overplotting in dense areas. This issue may reduce the efficiency of the users' visual exploration. Secondly, encoding flexibility is limited. Alpha blending results in a linear density mapping. Hence, it is often not possible to perceive both, small and large density differences within a single visualization.

To overcome these problems, edge splatting is supported in the analysis component (see Figure 3.8, right). Similar to the work by Scheepens et al. [2011], the trajectories are transformed to a density distribution of pixels in a 2D matrix with the same pixel size and ratio as in the current viewport. More specifically, the algorithm iterates over all trajectories and determines the intersecting pixels. Each pixel finally holds an amount of intersecting trajectories and one can easily obtain the minimum and maximum density in the distribution. Finally, the density of each pixel is mapped to color. This is done with a non-linear mapping to approximate perceptual linearity [Liu et al., 2013] and to reduce sensitivity to outliers. In the case study in Section 3.15, edge splatting is applied to two large datasets (see Figure 3.15). The calculated density values are mapped to different color schemes ranging from white (low density) to red or blue (high density).

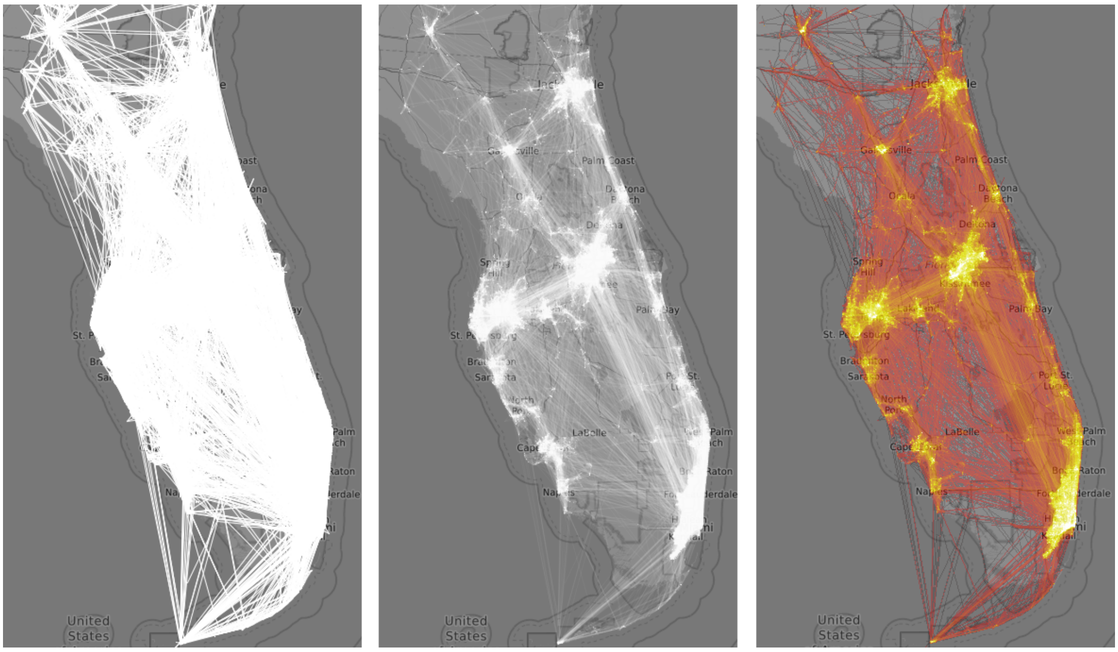


Figure 3.8 — Drawing of large movement data. From left: full opacity, alpha-blending, edge splatting with normalization and color mapping from red (low density) to yellow (high density).

3.4.3 Graph-based Approaches

For the exploration of movement volumes, edge splatting is an adequate technique. In addition to this, a graph-based method was developed that especially addresses the specifics of Twitter data (S2, S3). Firstly, the exact path between two geo-located messages is unknown. The more time is spent between two locations, the less certain it is that the user took a straight path (and uniform speed) to reach the destination. With animation or edge splatting, however, the direct path between two consecutive positions is visualized. Secondly, microblog data volumes vary strongly, depending on timespan and selected region. Peaks in popular locations skew the data distribution, and the large number of trajectories from or to these peaks can occlude underlying patterns. While edge splatting mitigates such clutter and occlusion, it is still difficult to understand traffic flows, especially when many trajectories with different angles are displayed.

Aggregation of movement data can help to overcome the aforementioned issues and is often achieved by graph computation. Most graphs, however, have the drawback of a fixed, pre-computed level of aggregation. On contrast, the approach presented here creates a hierarchical graph structure that allows for a stepwise data aggregation and abstraction from detail to overview. It hierarchically subdivides the geographic regions based on their data volumes. For this recursive subdivision, different structures could be applied. While political boundaries (e.g., countries,

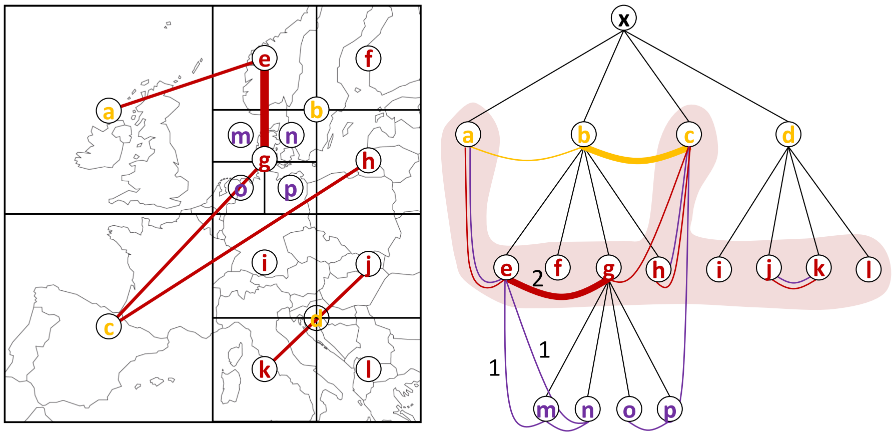


Figure 3.9 — The hierarchical graph model and weight-based link propagation. Original links in purple (leaf level) are propagated to upper levels (red, yellow), allowing for more abstract/aggregated views. The map shows links at the second (red) level.

states) are a common way to bin geographic data, the current implementation relies on an adaptive rectangular grid (quadtree). The advantage of this method is that it bins the data depending on the natural distribution rather than creating hard cuttings on political borders. However, as country-based statistics are often of interest, other shapes should be supported in the future.

To create the grid, the algorithm uses all locations of Twitter messages for the selected region and time. The reason why it does not only take into account the movement points is that all messages better reflect the Twitter population and sparsity problems in regions with low message density are reduced. At first, the analyst defines a cut threshold. This means that each cell is split when the cut value is exceeded, leading to large cells in areas with low amounts of messages and to a finer grid structure (smaller cells) in regions where many messages are posted. This does often but not necessarily correspond to the population. Figure 3.9 illustrates the spatial properties of a grid on the left and its data structure on the right side. The original links on leaf level are shown in purple. A cutting threshold of one leads to the highest detail, because each data point is represented as a single cell on the leaf level. By contrast, looking at a higher level of the hierarchy provides more overview. Depending on the underlying data and task, one can create flat or deep hierarchies with stronger or weaker aggregation per cell.

Next, the movement data is used to compute directed links between the grid cells on leaf level. Each link holds the number of movements between two cells as weight. To achieve a smooth transition from detail to overview, a link propagation algorithm is applied. It sums up the links' weights to a higher tree level. Once calculated, the tree holds information for all levels and does not need to be computed again. The

Figure 3.10 — Node-link visualization of the hierarchically aggregated graph. From top left to bottom center: Decreasing levels of aggregation using the computed hierarchy. The thickness of edges encodes the number of movements from cell to cell. Each cell is colored based on incoming or outgoing movements. Statistics can be derived through interactions (bottom right).

algorithm starts to propagate the links from the leaves to the first aggregation level (see Figure 3.9, red links). Thereby, links disappear that point to direct siblings (here m–n and o–p), and others get stronger because their weight sums up (e–g). Likewise, links for the next aggregation level are illustrated in yellow. Depending on the selected level, the respective nodes and their links are visualized in the map. Figure 3.9 illustrates the red link level. This also includes leaves with links that are shallower than the current selected level. Note that the illustration uses only a few links, while the quadtree can be large, holding thousands of nodes and links that have different directions. Figure 3.10 shows the results for a large dataset, covering three days in the USA. The analyst can step by step increase or decrease the level of detail and render the graph on different hierarchy levels. Grid cells represent the nodes of the graph. They are colored according to the number of incoming or outgoing link weights, which leads to a grid-based heatmap encoding aggregated movement volumes. To take the spatial area of each cell into account, the number of links is normalized with respect to the cell size. Similarly, Figure 3.14 shows results for mircoblog data in the USA and eastern Asia as part of a case study.

Besides links, the gridded heatmaps can be overlayed with arrows that indicate movement direction by showing the average angles of outgoing or incoming links per cell (see Figure 3.11). Increasing standard deviation of movement directions is encoded by increasing arrow transparency. While the link-based visualization shows main connections, the arrows have other benefits: they are free of clutter and reduce misinterpretation of routes travelled caused by sparse data resolution.

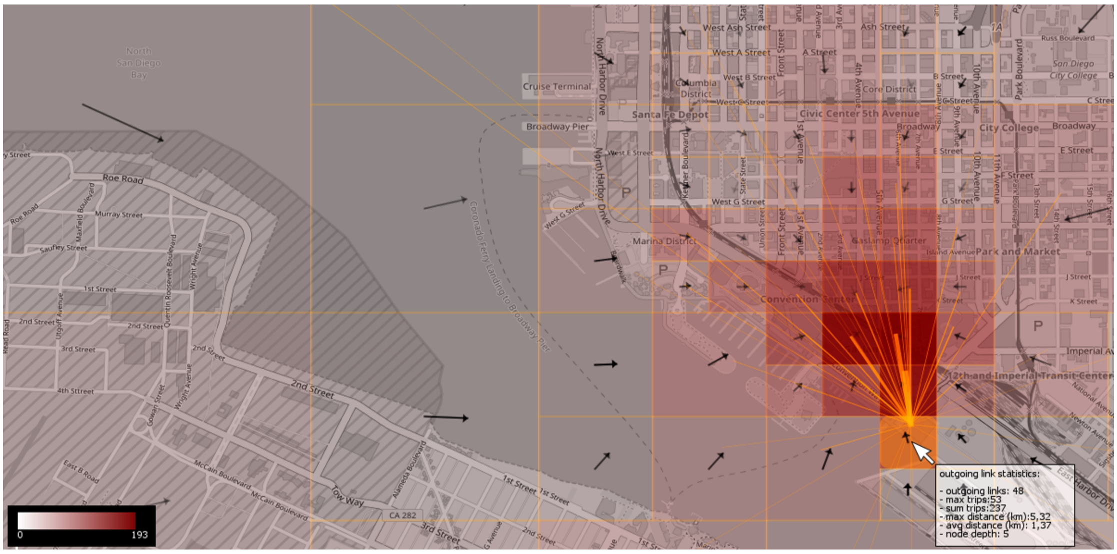


Figure 3.11 — The grid structure is colored according to the number of outgoing movements (here: San Diego Comic-Convention 2013). Arrows indicate average movement direction. By hovering a cell, a tooltip gives additional information and highlights outgoing movements (in orange).

3.5 Movement Comparison

With the techniques used to construct and visualize raw movements as well as aggregated ones, one can compare patterns on different levels of detail (S4). The statistics of the datasets loaded appear in the comparison interface to the right of the map view (see Figure 3.3). To help the analyst to distinguish between both datasets, the first dataset loaded is colored in red, the second one in blue.

The comparison of movement data is challenging. When comparing different datasets (e.g., Twitter and flight schedule data or Twitter data for different time periods), the data first needs to be normalized. Directly comparing the datasets would lead to severe bias, because their density may not be at the same level of magnitude. A min–max normalization transforms the data into the range [0.0, 1.0]. Compared to the flight schedule and taxi route data, the spatial distribution of the movements extracted from Twitter can have large peaks that would dominate the normalization and may lead to undesirable and incommensurable results. Often, the rest of the data distribution is hardly visible. To overcome this issue, a cube root function is applied that preserves differences in lower density range while it compresses the peaks to an acceptable level.

A histogram first gives an overview on the data density distribution. Depending on the visualization in use, it shows the movement distribution in the edge splatting matrix or of link weights in the hierarchical graph structure. If further adoption

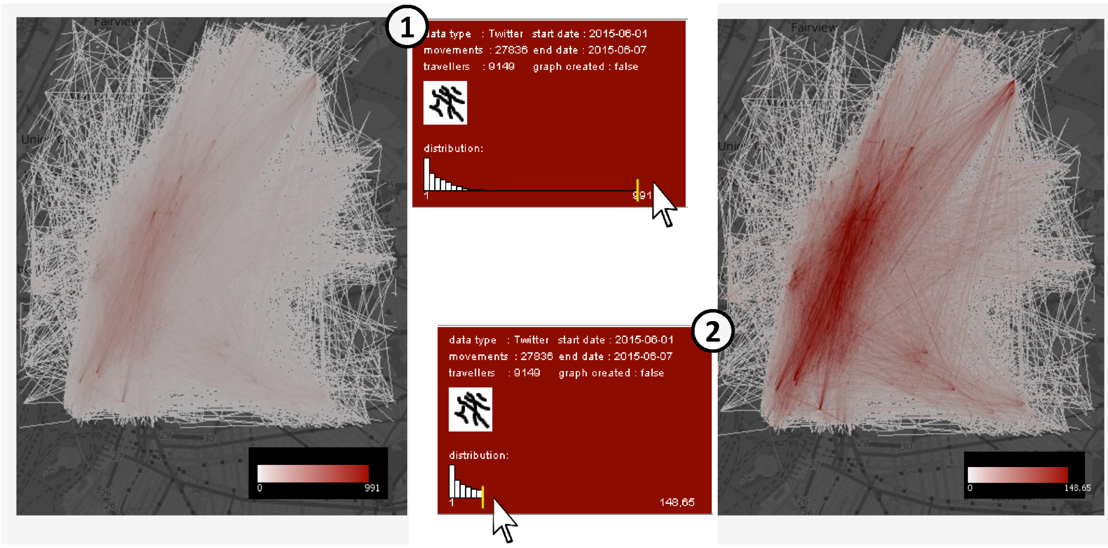


Figure 3.12 — (1) The density distribution of one week Twitter data in Manhattan (see Section 3.6.1) is dominated by large peaks. (2) They can be compressed interactively, shifting the maximum to a lower value. Besides the histogram, overall movement count, mean μ, and standard deviation σ per pixel give further information about the distribution.

is needed, the analyst can interactively shift the maxima to a lower range (see Figure 3.12). All values to the right of the marker will be downgraded to the value at the position of the marker (new maximum). Thus, a highly unbalanced distribution will be transformed to be more flat, and users become aware of the transformation. Together, all adjustments lead to the following equation with x' for the pixel value in case of splatting/link value for the graph and $x_{max_{ad}}$ for the adjusted maximum.

$$x' = \alpha + \left(min(1, \frac{x - x_{min}}{x_{max_{ad}} - x_{min}}) \right)^{\beta} * (1 - \alpha)$$

α (presetting: 0.05) is a threshold that preserves visibility for areas with low density) and β (0.33) introduces non-linearity [Liu et al., 2013] that controls the degree of data smoothing.

After the normalization, users can run a comparison analysis using either edge splatting (detail level) or gridded heatmaps (with either links or arrow glyphs), which rely on the hierarchical graph structure. In edge splatting mode, each pixel value of one image is simply subtracted from the pixel value of the other image. Thus, the pixel value in the result image is in the range $[-1.0,\ 1.0]$. The example in Figure 3.15 shows edge splatting results for two datasets and their comparison. The graph-based comparison is also done by subtraction. A precondition is that both graphs have the same layout, i.e., same quadtree structure with same depth and same number of nodes, otherwise the comparison will fail to find the right links to compare. By employing an *auto-filling* strategy one can obtain the same hierarchical grid structure for different datasets: users first load a dataset and create a grid. The resulting heatmap can be seen in Figure 3.13, left. When a second dataset is loaded for comparison, the same hierarchical grid structure is constructed but filled with the movement data from the second dataset.

Afterwards, the comparison of the two hierarchical graphs is performed by subtracting their link weights. E.g., a value of 0.5 subtracted with 0.8 leads to a slightly negative value of -0.3. Lastly, a divergent color scheme with colors, ranging from blue over white (similar) to red is applied. This can either be visualized by coloring cells (Figure 3.13, right) or edges (see Figure 3.14, bottom left).

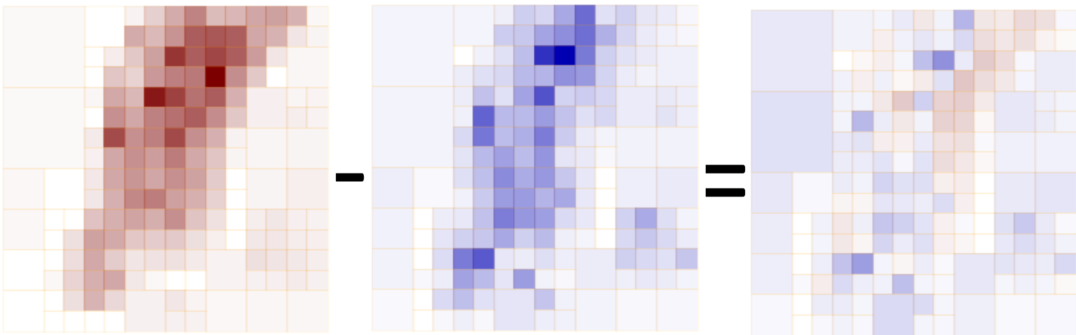


Figure 3.13 — For the first dataset loaded (here, taxis in New York) a grid structure is created (left). The same grid is filled with the second dataset (Twitter data, center). The subtraction is shown on the right.

3.6 Evaluation

To evaluate the applicability of the approaches different studies were carried out that are presented in the following. At first, the suitability of reconstructed movements from Twitter as an alternative movement data source (see Section 3.6.1) is evaluated. Then, two case studies show how the approaches can be applied to understand large scale events (Section 3.6.2).

3.6.1 Comparison to Other Data Sources

For an exemplary evaluation of the extracted microblog movements, a comparison to more commonly used data sources is carried out. Similarities and differences are investigated. On a global scale, movements extracted from Twitter are compared to flight schedule data. On a local level, they are compared to Taxi data, which reflects daily commute patterns.

Global Flight Schedules

Flight schedule data is often used to investigate global mobility patterns. It is thus interesting how movements extracted from Twitter conform to real world transportation on a global level. For this study schedule data from the OpenFlight project[3] is used. It was collected in August 2014 and includes 8,107 airports and 67,663 air routes. For each route, the data contains origin and destination. The connectivity of this dataset is very strong as many airports allow flying to many destinations. To reduce clutter, the analyst makes use of the hierarchical aggregation approach, which creates an adaptive grid structure (described in Section 3.4.3) and aggregates the flights into a hierarchical graph. The upper left image of Figure 3.14 shows results for the USA for a selected aggregation depth. The level of aggregation can be changed with a slider. Movement strength is encoded by the width of the links connecting different cells. Cell color encodes the accumulated weight of outgoing links. As the second dataset a sample of one month of Twitter data from the USA is used, also collected in August 2014. To get a comparable graph for the Twitter movements and filter the outcome to flights only, the same grid structure (see Section 3.4.3) is used, leading to 2,003 movements (Figure 3.14, upper right). One can see that visually similar connected links and movement volumes (relatively similar widths) exist in both results (flights and tweets). Large airports around Los Angeles (1) and New York (3) are slightly busier in the Twitter data (upper right), as indicated by the width of the links of each corresponding grid cell. The color of each cell is also comparable for both datasets, indicating the relative number of scheduled flight departures and outgoing Twitter movements is equivalent in most areas. However, around the area of Atlanta (2) the connectivity to other areas is stronger and the relative number of flights is higher as in the microblog data (Atlanta airport is the world's busiest airport). The comparison approach was also applied to eastern Asia. The lower images encode both datasets in a single view using a divergent color scheme for differences of outgoing links (left) and cells (gridded heatmap, right). For the aggregation level shown in the image, each cell contains on average 13.2 outgoing movements for the flight data and 3.1 for the Twitter data. The flight schedule data is more prominent in China (4), were Twitter is less popular (more people use Sina Weibo). Clear peaks in the Twitter data become visible in Indonesia, Malaysia (5), and Japan (6). Hovering over Malaysia highlights 10% of all flight related movements in the Twitter data, while in the flight schedule data, the relative amount is significantly lower (2%). Overall, this case study shows a high similarity for flight connections in the USA. In Asia, similar connections could also be identified, but their relative travel volumes show significant differences.

[3] http://openflights.org, last accessed: February 22, 2017

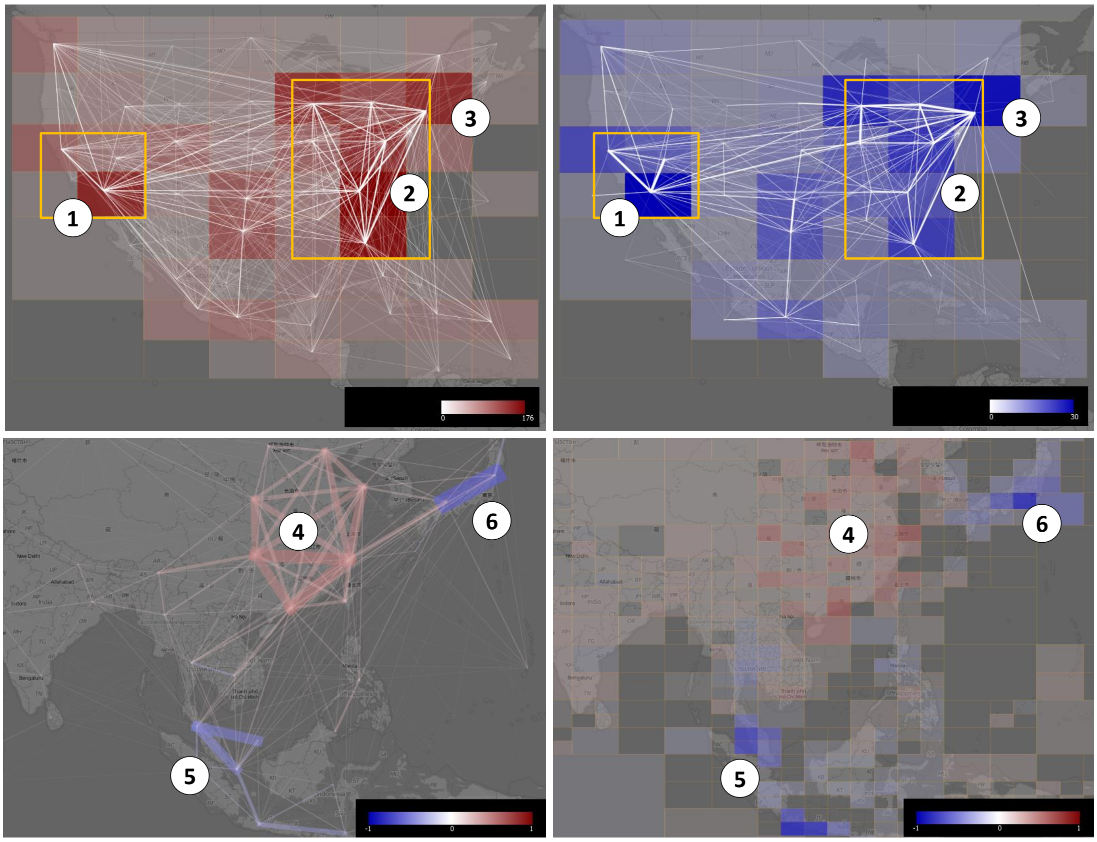


Figure 3.14 — The upper two images show two graphs on the same hierarchy level: flight data (left) and Twitter data (right). Flight data is more prominent around Atlanta (2), Twitter data in California and New York (1, 3). The lower images show that the data from flight schedules is more prominent in the Chinese area (4). Twitter data has its peaks in Indonesia and Japan (5, 6). The lower left image shows connection differences while the lower right heatmap highlights different volumes of outgoing movement.

Taxi Traffic in New York

Transportation data derived from public buses, taxis [Ferreira et al., 2013, Wang et al., 2013], and bike sharing systems [Wood et al., 2011] is often used to better understand urban mobility patterns. To determine if microblog data could be used instead, one week of Twitter movements in New York was compared with one week of taxi data. In New York, Twitter volumes are extremely dense and one can also choose from different traffic data sets that are available for the public. The dataset used in this study is from the NYC Taxi & Limousine Commission and was recorded in June 2015. A 5% sample of the first week in June was chosen, leading to 140,000 taxi trips. From Twitter, data of the same timeframe and spatial area was chosen, leading to 50,000 trajectories.

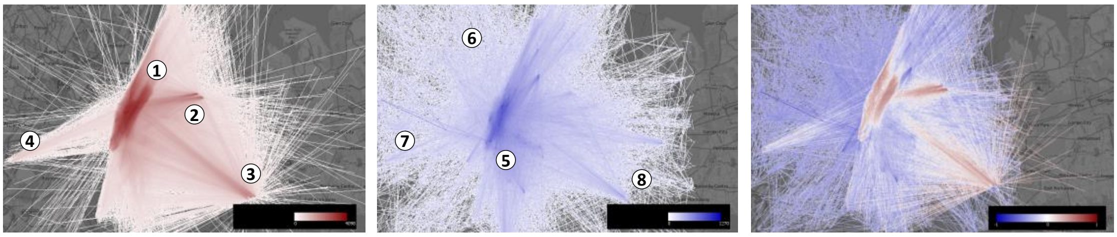


Figure 3.15 — The taxi data (left) shows different peaks at New York's airports (2, 3, 4) and in the inner city (1). Twitter data (center) has higher peaks and is not that smoothly distributed, but also gives information about movement in the north-western areas (6). The right image shows differences (red, blue) and some similarities (white), for example, the routes to/ from Newark airport (cf. 4, 7).

In Figure 3.15 one can see the individual edge splattings for both datasets and their diversity splat. The visualization on the left shows the taxi dataset that has certain peaks (red) at the airports Newark (4), John F. Kennedy (3), and La Guardia (2) that surround the New York area. The derived statistics $\mu = 27.1, \sigma = 73.6$ (density per pixel) align with the visual impression: in the center of Manhattan many taxis are in use while in remote areas the density of the data is mostly low, leading to a high standard deviation. Manhattan locates main attractions and businesses. The movements derived from Twitter (center) share some of these peaks. However, the color transitions are not as smooth. One can hypothesize that this is because most microblog movements show tourist routes to main attractions such as the Empire State Building and the City Hall (5). The JFK (8) and the Newark (7) airports also stand out. Compared to the taxi dataset, the number of Twitter movements is less but they also cover the north-western area. This leads to a lower mean and standard deviation ($\mu = 11.18$, $\sigma = 24.47$). By clicking on the comparison view, both edge splatting images are subtracted with a min–max normalization, and a cubic power function is applied (see Section 3.5). The result is shown in the right image. The aforementioned peaks clearly stand out. The routes to the Newark airport are quite similar, as is the area around JFK. While the taxi data is densest in the inner city, information about the western regions is only available from the movements reconstructed from Twitter.

3.6.2 Case Studies

To evaluate if the reconstructed movements also properly reflect large-scale events and seasonal changes, different events were investigated regarding travel volumes, where people come from, and the temporal distribution of the event.

Disaster Prevention: Pilgrimage to Mecca

This case study compares different time periods of trajectories derived from microblog data in the area of Mecca, Saudi Arabia. Mecca is the central pilgrimage destination for Islam. During the *Hajj* (the pilgrimage), more than two million people travel to Mecca to circumambulate *Kabaa*. In September 2015, more than 700 of them died in a mass panic[4] near the *Jamaraat bridge*. If changes in movement volumes and directions derived from Twitter can be identified it could help to gain insights and possibly make the journey more secure. The analyst starts the investigations by loading the data for the 2015 *Hajj* (see Figure 3.3). Therefore, she slightly extends the timeframe from 18th to 26th, to also cover arrivals and departures. Next, the filter is adapted to consider pedestrians only, which reduces the dataset from 3,253 movements drastically to 194 movements. Hence, she accepts to lose precision to increase recall and widens the filter restrictions (see Section 3.3.3). This leads to an amount of 793 movements for the selected area and time period. The analyst has a first look using alpha blending and edge splatting. However, the data is not dense enough to produce a smoothly distributed image, and she makes use of the hierarchical aggregation means (see Section 3.4.3) to reduce clutter from diverse movement directions. After interactively choosing a suitable level of detail using the hierarchy slider the grid-based heatmap shows clear patterns that are highlighted in Figure 3.3. Label (a) shows a peak at *Kabaa*, the cuboid building at the center of *Masjid al-Haram*, the sacred mosque. Label (b) points at the area of *Muzdalifah*, another stop of the pilgrimage, just like (d), which highlights the area around *Mount Arafat*. Arrows encode average travel directions. One can see that many arrows point from *Muzdalifah* to (c) *Jamaraat Brige*, a large pedestrian bridge build to carry hundred thousands of pilgrims. In this area, the catastrophe happened between the 23th and 24th of September, in 2015. This is also indicated by the temporal peak at the timeline (e). To compare this extraordinary situation to an usual time period, the analyst loads Twitter data from August, as a second dataset to make it comparable to the *Hajj* data. While the August dataset shows a peak around *Kabaa* only, the *Hajj* data clearly has its density peaks also at other pilgrimage places (see Figure 3.3, white box).

Visitor Analysis: Super Bowl 2013

The Super Bowl is the annual championship game of the American Football League. With 114.4 million viewers in the United States and approximately one billion viewers worldwide it is the most watched television broadcast [Craig, 2016]. The large-scale event has to be planned and supervised carefully as the host city of the event has to fulfill different requirements. According to an article of the magazine Sports Illustrated [Rose, 2014] the stadium needs to have a minimum size of 1,000,000

[4] Hajj Stampede Near Mecca Leaves Over 700 Dead, New York Times, September 2015, http://web.archive.org/web/20160905193042/http://www.nytimes.com/2015/09/25/world/middleeast/mecca-stampede.html, last accessed: March 23, 2017

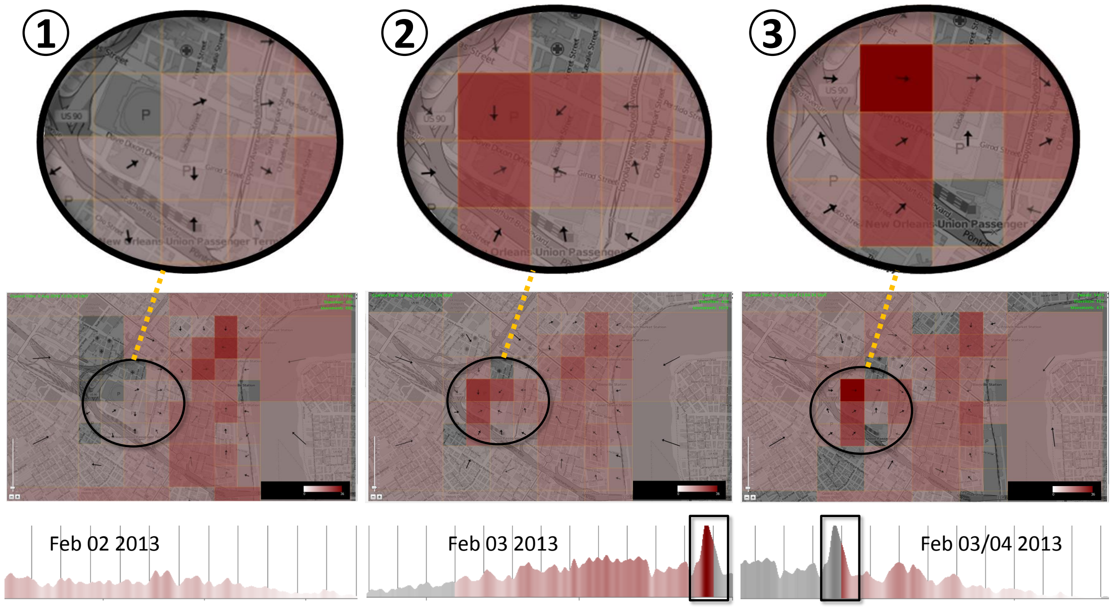


Figure 3.16 — The Super Bowl 2013 in New Orleans. Clear changes in mobility flow become visible. (1) One day before the game: normal movement focusing around the city; (2) hours before the game: people travel towards the stadium area; (3) after the game: people leave the Super Bowl stadium area towards the city.

square feet, a large entertainment area in walking distance, and requires hotels in the neighborhood to host at least 35 % of the stadium's capacities. On February 3, 2013, the Super Bowl XLVII (No. 47) was held in New Orleans, Louisiana, in the Mercedes-Benz Superdome. The city therefore invested more than one billion dollars in infrastructure. Planning and construction had started years before the event took place. Such preparations hold many challenges for urban and traffic planners, but also for the persons in charge of security.

The proposed approaches for movement reconstruction and exploration from Twitter allow to investigate the main mobility flows before, during, and after the event. This can not only help to monitor the situation during the happening. It can also be used to derive more generic insights useful for future planning. To start the analysis, the analyst retrieves Twitter data from the Lucene repository by interactively defining the area and timespan of interest. In this case, the area includes the city of New Orleans and surrounding regions. The timespan is set to start in the beginning of February 2013 and ends including February 5, two days after the event. After the data is fully loaded, the analyst can activate alpha blending or edge splatting to get a first impression of the raw data volumes and quality. From 33,392 loaded tweets 20,684 could be included in the trajectory reconstruction leading to 18,202 movement *edges* (trajectory parts) of 2,482 users. The analyst can also make use

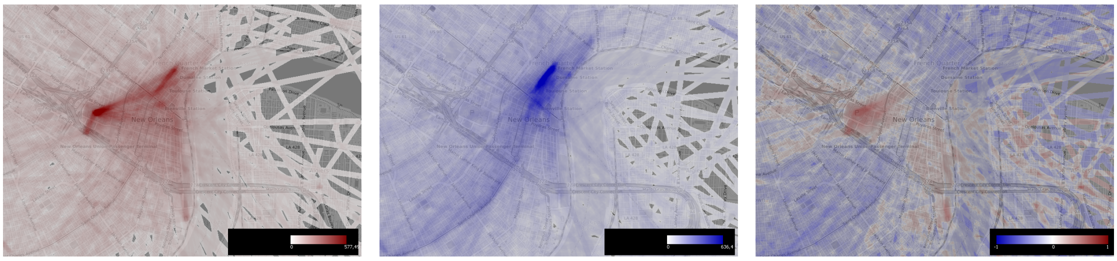


Figure 3.17 — Density-based splatting for the Super Bowl days (February 3 and 4, left image), an arbitrary week (center) and an explicit comparison by normalized subtraction of the distributions (right). The Superdome area clearly stands out.

of the animation capabilities. Figure 3.7 shows a snapshot before the event where people arriving at the stadium mostly come from the city center.

A clearer understanding of traffic flows can be gained after graph aggregation by looking at the average travel directions per cell (node) of the hierarchical grid. The analyst can therefore choose an aggregation resolution that fits the goals of the current task. In Figure 3.16, the area around the Mercedes-Benz Superdome is visualized. One can clearly see that the arrows point towards different directions on the 2nd of February, one day before the game took place. The main movement densities are further north-east, towards the Business district and the French quarter area (Figure 3.16, 1). On February 3, by contrast, there is a clear shift towards the stadium and a strong increase of movement, especially before and during the game (2). Of course, there are more messages sent during the game allowing to get more information about the movements. This can be seen in the timeline (2). After the game, the arrows start to point at various directions again, mostly towards the city, indicating that people leave the event towards this area (3). At last, the analyst contrasts the data by loading tweets for the following week, not including a large scale event. For the comparison she uses the pixel-based approach (subtraction of splattings), because the data is smooth and dense enough to produce a clear image. One can see the main differences in the views (see Figure 3.17).

The case studies revealed patterns that increase situational awareness, and help planning and monitoring large scale events in the future, especially when other data is not available. While there are biases in the data, e.g., users share an increasing amount of messages when an event is happening, the results still contain real world movement that may not be known without the application of the presented approaches. Further application areas, be it for urban and traffic planning, customer analysis, or security may arise in the near future.

Chapter 4

Semantic Enrichment with Context Data

Raw movement readings usually consist of spatial and temporal information. Spatio-temporal processing and analysis of such data leads to insights about *where* and *when* movement is happening. To answer questions about the *how* and *why*, however, more context information is necessary. Eyewitnesses are usually aware of further details about places, weather and climate, sentiments, actions, and interactions between moving objects. However, this context is often not directly recorded with the same sensors and stored in the same database. Hence, it is missing for a post-analysis, making it impossible to fully understand the situation, i.e., to derive all patterns and dependencies like cause and effect. For example, if a car stops at a specific point in the afternoon, there can be several reasons for the driver's behavior, and the recorded GPS positions are often not enough to come up with an explanation. Movement data needs to be enriched with further information to support the data analyst in the sensemaking process. These enriched tracks are referred to as semantic trajectories [Parent et al., 2013].

The following sections present different enrichment approaches. Figure 4.1 shows the corresponding parts in the visual mobility analysis pipeline, introduced in Section 2.4.3. The first contribution (Section 4.3.1) is a human-centered annotation approach with points and areas of interest, which builds on the analyst's background knowledge and the human ability to combine different pieces of a puzzle. An urban planner, for example, is aware of land use, housing, shopping, and industrial districts and can use this information in the annotation process accordingly. Manual annotation, however, reaches its limits when data is large and context is uncertain. In these cases, computational accuracy and efficiency

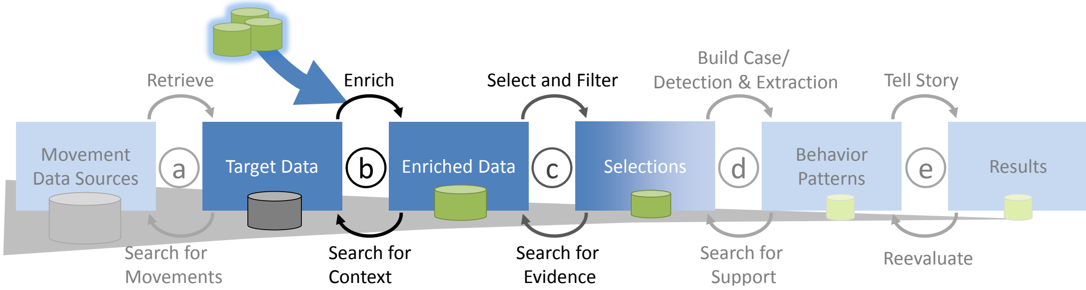


Figure 4.1 — This chapter mainly addresses semi-automated approaches for a semantic enrichment of movement data with contextual information, the visual representation, and interactive adoption.

is needed. Evaluation results (see Section 4.1.2) show, however, that automatic enrichment approaches are far away from generating trustworthy outcomes. To address this issue, the second contribution is a semi-automatic decision model for point of interest (POI) assignment (Section 4.3.2). POIs are derived from the social media service Foursquare. It enables an insightful handling of geo-spatial uncertainties resulting from the enrichment by integrating the human in the loop. The analyst can review results and make adjustments accordingly. Thirdly, a method to enrich data with textual information from a microblog service containing news, events, and sentiments is proposed (Section 4.3.3). To explore the enriched movements novel visualization techniques are needed that highlight the contextual data. Especially the temporal order of happenings is of interest as it allows to identify behavior routines. Section 4.4 presents two scalable visualizations. Finally, the enrichment with Foursquare POIs is evaluated using datasets with known ground truth (Section 4.5.1), and two case studies apply the approaches in the domain of electric mobility an law enforcement (Section 4.5.2).

Parts of this chapter have been previously published in:

R. Krüger, S. Lohmann, D. Thom, H. Bosch, and T. Ertl. Using social media content in the visual analysis of movement data. In *2nd Workshop on Interactive Visual Text Analytics, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 1–3. IEEE Computer Society, 2012b

R. Krüger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 193–200. IEEE Computer Society, 2014a

R. Krüger, D. Thom, and T. Ertl. Semantic enrichment of movement behavior with Foursquare—a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 21(8):903–915, 2014b

R. Krüger, D. Herr, F. Haag, and T. Ertl. Inspector Gadget: Integrating data preprocessing and orchestration in the visual analysis loop. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 7–12. The Eurographics Association, 2015b

4.1 Background

Related work on semantic movement analysis can be categorized in three areas. Firstly, researchers have proposed data structures for a semantic modeling of movement. These structures define semantic attributes and their relationships, either by ontologies or by entity-relation models that give a more technical, application centered view. This area is not within the focus of this thesis, but semantic modeling and its advantages and drawbacks will be briefly discussed. The second research area focuses on the semantic enrichment of movements, i.e., how multiple attribute values can be added to the data. Thirdly, it is of interest how the enriched data can be visually represented and analyzed. The latter two areas are often not clearly distinguishable. Many papers present approaches for enrichment and analysis, and especially in visual analytics a close coupling of these steps is reasonable.

4.1.1 Semantic Data Models

Semantic data structures became popular with the rise of semantic web technologies and an *"explicit specification of a conceptualization"* called *ontologies* that allow for sharing and reusage of knowledge among software entities [Gruber, 1995]. The word ontology originates from the philosophy of the nature of being, starting in ancient Greek history. Among others, the philosophers Aristoteles and Parmenides created a reasoned perspective on the structure of reality [Hacking, 2002]. Wolff [1736] was among the first authors who introduced the word ontology as describing structures, their properties, and relationships to a broader audience in his work *Ontologia*. Recently, it gained increasing practical usage in the semantic web, where ontologies are used to define semantic structures. A standard model for data interchange on the web is RDF (Resource Description Framework)[1]. So called RDF-tripels, consisting of subjects, predicates, and objects allow to model real world dependencies. The Web Ontology Language (OWL)[2] builds on RDF.

The technology has been also applied to the domain of mobility analysis to model attributes and relations among movement and moving objects. Yan et al. [2008] describe a framework for a *"semantically-oriented structuring, modelling and querying of trajectory data"* based on OWL. They rely on ontologies that address both, domain independent and application dependent semantic and geometric facets. A similar approach is presented by Hu et al. [2013]. Also, Klien and Lutz [2005] propose an ontology-based retrieval and discovery to support semantic annotation. Spaccapietra et al. [2008] define a *trajectory design pattern* that specifies similar structures with a more technical, UML-like description.

At a first glance, semantic web standards have advantages, because they increase compatibility and allow to make use of existing infrastructures, components, and

[1] https://www.w3.org/RDF, last accessed: March 24, 2017
[2] https://www.w3.org/OWL, last accessed: March 24, 2017

data. In some fields, semantic web structures are successfully used for knowledge management and allow to query and combine resources in the internet, similar to a data warehouse. However, the idea of semantic web and its implementation and resulting applicability differ. Firstly, regarding the three main parts of the semantic web, namely data models, services, and intelligent applications, McCool [2005, 2006] discovers several issues. This includes an often unnecessary complexity of the description languages, poor user participation, and an increasing but insufficient number of services and applications. In practice, web interfaces that offer OWL-based structures and retrieval on a complete and reliable basis are still rare. Those that exist do often not contain up-to-date information, are incomplete, and do not provide interfaces for high data throughput. Flexible combination of data affords strong normalization that does not necessarily fit big-data analysis requirements such as real-time streaming and fast retrieval of massive and high-dimensional datasets. Common web services from YouTube, Twitter, and Facebook barely support OWL-conform queries nor do they deliver data in RDF syntax. Similarly, data mining algorithms and libraries mostly work on the raw data or simple table-like formats. Secondly, in the domain of mobility analysis different semantic models have been proposed [Klien and Lutz, 2005], but there is no standardized semantic model for movement data, yet. Thirdly, semantic web technology barely addresses uncertainty and relies on fully automatic reasoning. By contrast, the approaches presented in this chapter aim to integrate these uncertainties to let the analyst get a *raw* picture and the ability to change and ponder on the enrichment.

As a last remark to this topic, this thesis has its goals and contributions not on the usage and improvements on semantic web technology but on novel approaches for the interactive visual analysis of human mobility behavior. With increasing processing power, lightweight data formats like JSON-LD[3], and new services as offered by Google's Knowledge Graph Search[4], semantic web technology may become important for the domain in the near future.

4.1.2 Semantic Enrichment

A summary on the field of semantic enrichment and analysis of movement is given by Parent et al. [2013]. They present concepts for (1) the construction of trajectories from plain movement tracks, (2) enrichment to enable desired interpretations, and (3) the application of data mining to extract semantic behavior patterns.

Regarding the first two topics, Alvares et al. [2007] propose a framework to transform trajectories in stops and movement first, and then to add context information. Guc et al. [2008] discuss web services as a means to derive context and to integrate the human in this process, but their implementation relies on manual annotation only. Yan et al. [2011] contribute *SeMiTri*, a multilayered framework for semantic

[3] http://json-ld.org, last accessed: April 26, 2017
[4] https://developers.google.com/knowledge-graph, last accessed: April 26, 2017

annotation of trajectories. It assigns POI categories based on previous states using Hidden Markov Models. Similarly, Spinsanti et al. [2010] propose an algorithm for POI enrichment. They compute probabilities by distances and stop duration. They discuss hierarchical POI aggregations, but do not argue where the structures and data come from. Statistics and uncertainty about the POIs are not taken into account. Slightly after the publication of parts of this thesis, Coffey and Pozdnoukhov [2013] published work where bike routes are enriched with context from Twitter messages. They apply topic modeling in order to find repetitive temporal patterns, but also analyze causal relationships between message topics and bike usage. However, non of the above mentioned approaches are validated.

Ground truth data for validation of semantic assignments was collected by Kochan et al. [2005] and Cools et al. [2009] from the Transportation Research Institute (IMOB), Hasselt University. They carried out a study where users harvested and annotated the data. Unfortunately, the data is not available for the public but some groups got access. Reumers et al. [2013] built a classification tree to make categorical assumptions, such as home or shopping, which purely relies on stop durations. They achieve mixed accuracies for the test data, ranging from 84% to 13% depending on the category. Furletti et al. [2013] also use the ground truth data to evaluate their fully automated semantic enrichment. It is based on POI web services and relies on temporal and spatial rules. They retrieve the semantic categories and their average duration times from the evaluation data itself, which may lead to biases and overfitting. Depending on the categories, they achieve good and average results (accuracy for food 83%, leisure 49%) as well as some poor results (education 3%, daily shopping 0%). Ideas for improvements are given by Andrienko and Andrienko [2013]. They, e.g., suggest to interpret meanings of personal POIs based on cyclic temporal patterns of visit times.

Different means were proposed to visualize semantically rich movement data. Andrienko et al. [2013b] focus on privacy-preserving movement analysis. They present patterns in abstract space by visualizing flows between activity types. While they enrich raw movement with context from web services to understand sematics, other research focuses directly on semantically rich movement data. Preoţiuc-Pietro and Cohn [2013], for example, analyze check-in behavior of Foursquare users. They use matrix visualizations to investigate transitions between POI categories and show temporal distributions in streamgraphs. Fujisaka et al. [2010] visually analyze movements of users based on geo-located Tweets and apply clustering on frequent POIs. Kling and Pozdnoukhov [2012] investigate semantic movement patterns of both Foursquare and Twitter users. They utilize streamgraphs, geographic heatmaps, and word clouds. Lastly, Zeng et al. [2017] build on approaches presented in this thesis and show a visual analytics system for enrichment and exploration of movement data from Foursquare with a focus on time-varying patterns.

While there is more and more research on semantic enrichment of trajectories, the evaluation results demonstrate the difficulties encountered here. The accuracy highly depends on the recorded data, the underlying urban areas, and the categories. Analysts have to deal with variety and veracity issues, which hamper an automatic alignment of data [Raman and Hellerstein, 2001]. This is known as the semantic integration problem [Bergamaschi et al., 2001, Doan and Halevy, 2005]. Visual analytics enables the integration of human expertise and decisions to value the outcome. In comparison to related work, approaches in this thesis combine enrichment with uncertainty-aware visual analysis in an interactive and semi-automatic manner. Rather than focusing on single trajectories, they rely on an aggregation of destinations to overcome data scattering and inaccuracies.

4.2 Semantic Data Model

Human mobility behavior is often described as a sequence of events. For example, people start the day by moving from home to work. After work, they go shopping at a nearby supermarket and finally, they drive back home. Andrienko and Andrienko [2013] discuss that a trajectory consists of many composited episodes, named events. An event is demarcated based on constancy or significant changes (as stated by Laube et al. [2007]) of the aforementioned attribute values such as *moving episode* or *behavioral episode*. The analysis of the behavioral episodes extends the focus from a geometric perspective of movement in an Euclidean homogeneous space to a semantic perspective where a movement can be described as a sequence of happenings. Hence, for the visualization of semantically enriched movement data, different approaches are needed that highlight events, transitions, and repetitions instead of geometric aspects only. Figure 4.2 illustrates this. On top, a *raw* movement trajectory is shown that consists of a set of measurements containing spatial and temporal values. After an enrichment with context information the same trajectory can be described as a sequence of events, e.g., daily routines of a person.

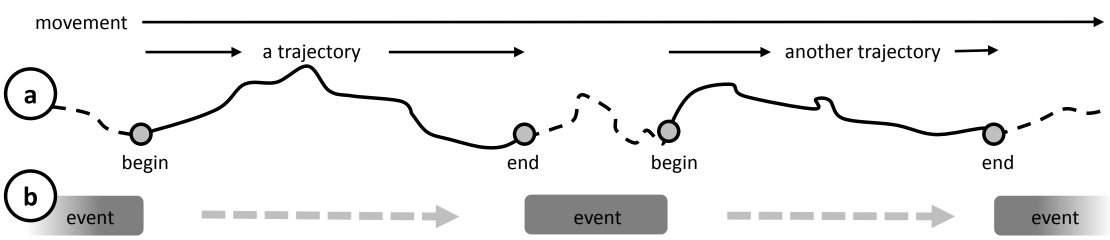


Figure 4.2 — Based on an illustration by Parent et al. [2013]: (a) Human movement can be described by a set of trajectories (focus on the movement itself). (b) In between movements, a person follows different activities taking place at a specific time and place (event). Depending on the application scenario, the movement itself (a) or the sequence of events taking place (b) is of interest.

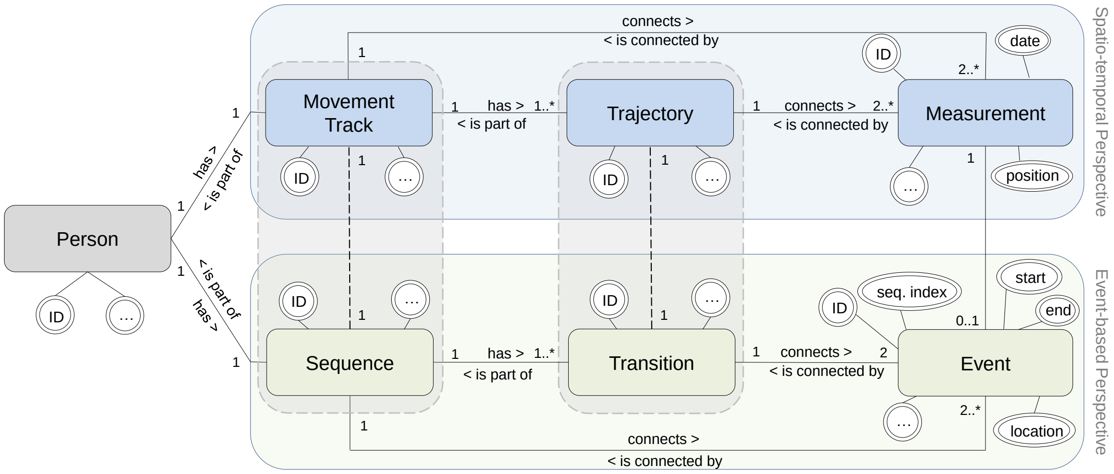


Figure 4.3 — A simplified version of the data model. It gives a spatio-temporal perspective (blue), representing moving episodes and an event-based perspective, representing behavioral episodes (green). Besides spatio-temporal attributes it allows to add environmental and social context.

To support the enrichment in the previously introduced manner, the data model has to allow the integration of additional context information. Figure 4.3 gives a simplified description of the main objects, attributes, and relations. Each movement object (a person using any means of transportation) has a movement track in geographic space. This track is split up in trajectories—an episode of the movement describing a trip from one place to another. A trajectory is composed of measurements, ordered in time. Measurements can contain more or less precise location information such as latitude and longitude and temporal information, e.g., a timespan or timestamps (see Section 2.3.2). Besides spatio-temporal information (blue entities) a moving person holds a sequence of events (green entities). An event comprises a location where it takes place, a start and end date, as well as an index that describes its position in the sequence. An event at least involves one person, but events can also be shared among multiple persons. In the model, however, this would be defined as separate events. On demand, a connection can be made via temporal and spatial comparisons or further attributes, such as name. A sequence defines the temporal order of events a person is involved in as well as further details about the transitions between them. Respectively, a measurement can be connected to an event if it describes the start or end of a trajectory. In the programmatic realization, a movement track and a sequence are represented as a single class, implementing two interfaces for a geographic and an event-based perspective. The same applies to trajectories and transitions.

The description here shows a minimal set of attributes. The empty attributes can be seen as wild cards for further information. In case of the e-mobility data (Section 2.5.2), a measurement comprises speed and battery drainage and a trajectory can contain comments from drivers annotating their trips via web interface.

Similarly, an event can have a name and a multitude of environmental and social conditions. In most cases, however, semantic context is missing, making enrichment necessary to derive semantic meaning.

4.3 Enrichment Approaches

In the following, different methods for a semantic enrichment of movement data are presented. At first, interactive means for manual data annotation with points and area of interests (POIs/AOIs) and other context attributes (Section 4.3.1) are proposed. To enable the enrichment for larger datasets and to allow for a fusion that goes beyond the analyst's knowledge, a semi-automated approach is introduced, which can be reviewed and adjusted by the analyst (Section 4.3.2). Besides POIs, a method allows to enrich data with textual information from a microblog service containing news, events, and sentiment (Section 4.3.3).

4.3.1 Manual Annotation and Alignment

The enrichment of geographical data with context information is not new. Geographers have used manual annotation for decades. Geographic information systems such as ArcGIS[5] and QGIS[6] offer interactive maps where overlays can stack any additional data to spot correlations [Klien, 2008]. This comprises information about landuse, business, and housing areas, as well as road networks. The here presented approach similarly allows analysts to manually annotate places and regions on an interactive map by making use of their expert knowledge or by deriving the information visually from another layer. The novelty is that the annotations are subsequently used to enrich movement data.

To annotate the map with points of interest, only a click to a certain region is needed. Different attribute values, such as category (e.g. *work* or *private*) and name can then be assigned. If larger areas are of interest, such as blocks, suburbs, or cities, the analyst can make use of a polygon selection tool. Therefore, the information in the underlying map serves as a visual template. If more assistance is needed, custom maps can be loaded and geographically aligned. When the annotation is completed, an algorithm checks for all movement trajectories that end in the defined area and enriches them with the given annotations. Figure 4.4 gives an example. At first, a university area is marked as area of interest (AOI) with an interactive polygon tool using a geographic map as underlying context (1). This area is then labeled (2). The color of the AOI is assigned based on the labeled category. As soon as the annotation is completed, the enrichment algorithm assigns the context to all trajectories ending in the university campus. After the enrichment, the movements can be expressed as a sequence of visited AOIs, or more generically

[5] www.arcgis.com, last accessed: February 26, 2017
[6] www.qgis.org, last accessed: February 26, 2017

Figure 4.4 — Annotation: (1) With a polygon tool an AOI is created. (2) It is annotated with category and name, here *College & University: Campus Stuttgart-Vaihingen*. Movements are enriched and colored based on their destination.

as a sequence $s := e_1 \to e_2 \to e_3 \to, ..., \to e_n$ of events e, where each event consists of a number of attributes such as location, start and stop date, involved objects, etc. (see Section 4.2). In a timeline that shows stop periods (x-axis) of all moving objects (y-axis) the matchings are colored accordingly (Figure 4.4, red colored bars). By annotating more and more regions on the map, additional *stop-events* colorize the timeline, making the event sequences more and more complete. In Section 4.4 different event sequence visualizations are described in detail. Lastly, the annotated areas can be stored or loaded for re-usage as a set of polygons and meta-information. Any creation, altering, or loading of stored annotations leads to a recalculation of matching trajectories and their annotation.

4.3.2 Semi-Automated Enrichment with POIs

When data volumes are large, manual annotation reaches its limits due to the workload and time needed for an analyst to manually align single cases. Klien and Lutz [2005] argue that manual semantic annotation is *difficult, time consuming, and expensive*. Another argument they raise is that data providers are often not willing and not capable to perform annotation tasks. To allow for a semi-automated process the approach presented in the following relies on the social location service Foursquare[7], where more and more people provide detailed geolocated information

[7] www.foursquare.com, last accessed: 2 May, 2017

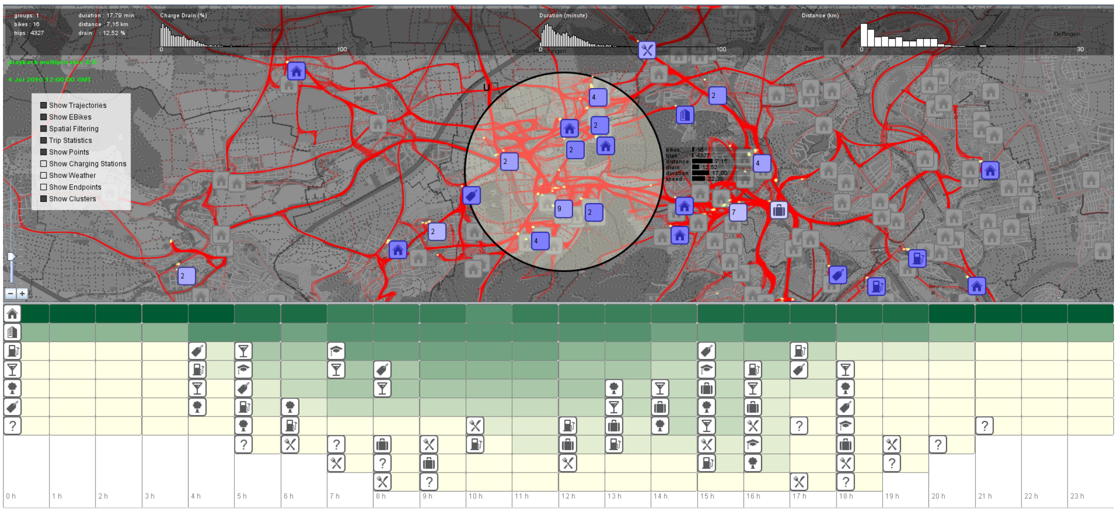


Figure 4.5 — Two linked views give a geographic and a cyclic temporal perspective on trajectories and visited places (symbols). They allow the analyst to adjust the semantic enrichment and analyze results. A lens (introduced in Chapter 5) is used to interactively selects a subset in the map that is displayed in the temporal view.

of several POIs, such as restaurants, bars, business locations, universities, sport parks, and public transport stations. Mainly used for '*What is here*' user-centered navigation applications, the web services can also be utilized to enrich geospatial data with context information, as summarized in Section 4.1. While most existing approaches rely on a fully automated decision model, evaluation results show that they are far away from generating reliable and trustworthy outcomes. Tracking is often not exact, and more inaccuracies can occur during broadcasting and logging. Thus, data can become noisy. Moreover, there is a high number of places not stored in POI databases, and thus many map locations cannot be resolved. By contrast, in city areas, the POI density is high, making it often not clear which POIs were visited. While mobile phones stick with the user, vehicles have to be parked somewhere and are often not moved all the way to the actual destination. All the above makes a *human in the loop* [Endert et al., 2014] who explores, analyzes, or even revises enrichment results still necessary. This is provided by a visual analytics interface (see Figure 4.5) that combines a geographic and a temporal view via brushing and linking. In the following, preprocessing and semantic enrichment steps are presented, before the visualizations are discussed in more detail.

Preprocessing

Preprocessing is important for the visual analysis of movement data, but it often receives little attention. However, the more ambiguities, inconsistencies, and inaccuracies in space and time can be eliminated, the more reliable the data is and the more accurate the outcome of a (semi-)automated enrichment will be. The techniques applied here are similar to Zheng et al. [2009] and use two threshold parameters: temporal distance and geographic distance. In case trips are not already separated, large geographical or temporal distances can indicate the end of a trip (Figure 4.6 a). While extracting trips and destination points, the measurements are smoothed to filter out outliers (b). Furthermore, for correct trip splitting, the approach differentiates between slow motion in a constant direction and dithering (c), resulting from noisy signals.

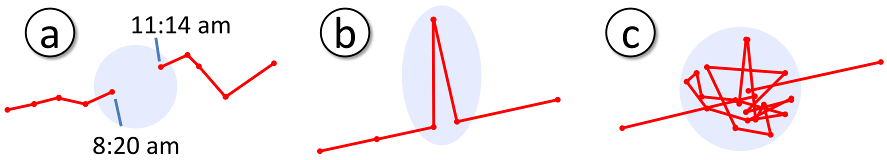


Figure 4.6 — Examples where measurement preprocessing is necessary: (a) Temporal and geographical gaps / missing data; (b) Signal outliers; (c) Signal dithering

A first glance at the processed data can be taken by visualizing it on a geographic map, as shown in Figure 4.7. One can see that the destinations of the trips build dense clusters over time, surrounded by noise. Such clusters may result from drivers that frequently return to the same location or indicate an event where multiple people go to. Noise can occur due to several reasons: Firstly, there are destinations only visited a few times. Thus, no clear cluster appears. Secondly, there are measuring inaccuracies due to GPS signal interference, deficient receiver performance, data logging, or broadcasting issues. When datasets are large, comprising many moving objects over a long time period, and behavior routines are of main interest, one can easily overcome this issue by focusing on destination clusters only and accepting a certain degree of information loss.

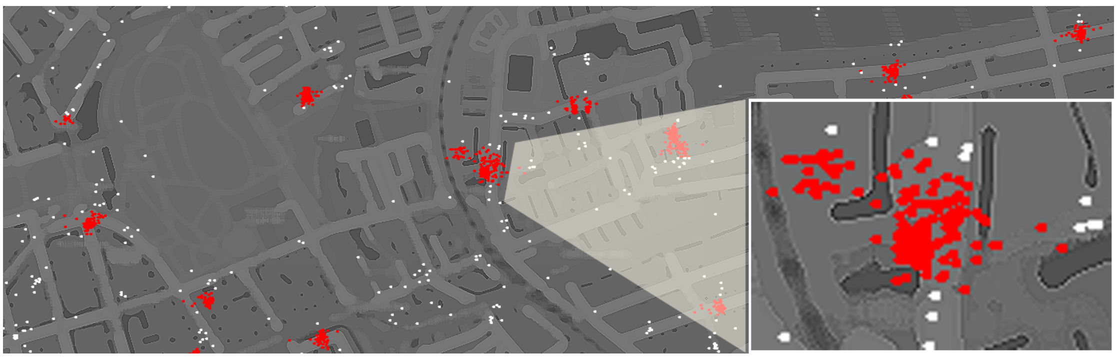


Figure 4.7 — Popular destinations of the large electric scooter dataset (see Section 2.5.2) that was used as a basis for the development of the following methods. The destinations build dense and clearly separated clusters (red) surrounded by some noise (white). Map details are blurred due to privacy reasons.

Criteria	Google	Facebook	Foursquare
Number of POIs	50,000,000	missing	30,000,000
Requests/min	1,000	3,600	5,000
Requests/day	100,000	1,000,000	1,200,000
Caching	forbidden	lacking info	30 days
Structure	fair	fair	good

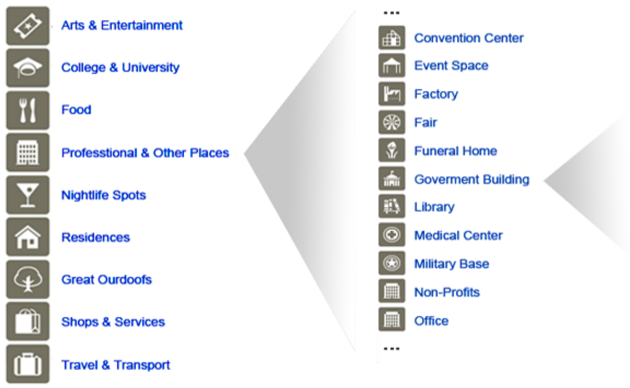
◄ **Figure 4.8** — API limits and facts (taken from [Constine, 2011], the Facebook API documentation, and public forum discussions).

For the automated detection of frequently visited places DBSCAN [Ester et al., 1996], a density-based clustering technique, is applied. Advantages using DBSCAN compared to partition-based approaches like k-means are that (1) there is no need for a-priori knowledge about the number of clusters, (2) DBSCAN implicitly filters out noise, and (3) it is able to detect clusters of varying size (area, num. instances) and shape (e.g., longish). For the clustering, a geographic distance function is used. In addition, the outcome depends on two thresholds: ϵ, which defines a maximum distance for range queries, and $minPts$, defining a minimal number of neighbors within ϵ distance for the cluster to be created or extended. Both parameters can be interactively steered by the user and depend on each other. Choosing a small $minPts$ and ϵ, for example, will result in many small clusters, considering even places where users only stopped a few times. Increasing $minPts$ usually reduces the number of clusters found, because only dense areas are considered. Increasing ϵ lowers the density threshold, as there is a higher chance to find $minPts$ within ϵ distance. This may connect single clusters to larger ones, covering a wider area. The configuration depends on the application and data. As an exemplary result on the scooter data, DBSCAN detects 1,215 clusters containing 105,808 of the 150,000 total trip endpoints. Hence, about two of three endpoints have often been visited by a single user or have at least been visited once by several users.

Hierarchical POI Decision Model

To query up-to-date context data for the detected destinations, the approach employs Foursquare, a location-based social networking service with dense information about POIs in city areas. Data from Foursquare is often quite recent as the data is regularly maintained by users, who can always add, delete, or update POIs (in Foursquare called *venues*). Furthermore, it delivers additional details such as the number of users that visited a venue and the number of individual check-ins, which can help to determine the recent prominence of a location. While there are other providers offering similar services, e.g., Facebook[8], the Foursquare API has less query limitations and still a high POI density (see Figure 4.8)[9]. Low query restrictions were an important factor in the decision, as they allow to compute the data enrichment online (at runtime), enabling more interactive requests. POIs

[8] developers.facebook.com/docs/reference/ads-api/api-rate-limiting, as constituted 2013
[9] Statistics derived in 2014



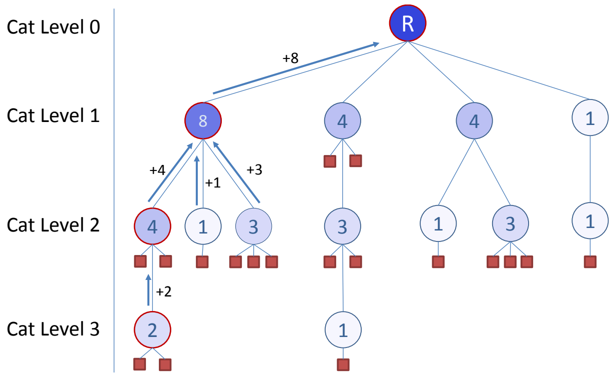
▶ **Figure 4.9** — Foursquare category tree: each level summarizes its child nodes in parent categories. Leaf nodes are highly detailed venue types, such as Post Office.

from Foursquare are categorized into three hierarchical levels. Each level aggregates its child categories to more comprehensive ones. The highest level contains nine overview categories (see Figure 4.9). Every POI belongs to one of these categories. For example *Ann's Blue Jeans Shop* belongs to the subsubcategory *Boutique*, subcategory *Clothing Store* and category *Shops and Services*.

POI Assignment Foursquare offers a search function which returns nearby POIs and their corresponding categories for a given geographic position and a defined query radius *rad*. This search function is used to add semantic information for all detected destination points in the movement data, but only in few cases the result is unique. Often more than one POI is found, and they often belong to different categories. Thus, it is uncertain which POI was actually visited. However, if the majority of found POIs for a cluster are restaurants, there is a good chance that the area is a food district, leading to the assumption that the person went there to have a meal. Foursquare returns category information in text notation appended to a POI result list. As this data format is hard to handle, a result tree is built, based on the category hierarchy with level 0 being the root element holding all found POIs (see Figure 4.10). The category depth depends on the category type and on how detailed the POI was described when added to Foursquare by the community. For some POIs there is only a main category available, while for others there is a finer hierarchy. Having calculated the result tree, the most likely category for the query at level l is determined by comparing POI counts of siblings (i.e., categories at the same level l having the same parent category). Depending on the current detail level, the certainty c, $0 \leq c \leq 1$ ($c = 1$ is certain) of the determined category cat at level l can be calculated as a ratio of the number of POIs found in this category and the overall number of POIs found for the destination.

$$c_{cat_l} = count(cat_l)/count(cat_0) \tag{4.1}$$

Thus, with higher level of detail categories usually get more and more uncertain, since only a low amount of found POIs belong to exactly that category, e.g. *Australian Restaurant*. With lower detail, by contrast, it becomes more certain because most POIs belong to a main category, e.g., *Food*.



◀ **Figure 4.10** — Example of a result tree with root element R, holding all found categories (nodes) with their POIs (red rectangles). POIs are appended at different levels, depending on how detailed they are described. The primary category on each level is based on the highest count (in red).

Relying on the number of POIs, however, is a naive approach. Considering the food district again, there could be other POIs, less frequently visited but much more important, like a large company with thousands of employees. Also, many restaurants may be within the geographic query radius, but at the very end of it, while a clothing store might be only a few meters away from the destination. As for each POI the distance to the query point is known, one could simply enrich the trajectory with the nearest neighbor to its destination. However, a way to make the category decision more robust (not depending on a single finding) is to calculate the distance average per category and rank categories accordingly. To penalize POIs far away even more, a Gaussian kernel $g(x)$ with $\mu = 0$ and $\sigma = rad/3$ is used, leading to $c = 1$ for POIs at the query point and $c = 0$ for POIs being rad meters away. Then, the average Gaussian value of all POIs within the category is calculated as follows.

$$avgGauss(cat_l) = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \qquad g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2}$$

An advantage of Foursquare is the additional check-in and user information for each POI. Besides the distance, this information can also be used to decide which POIs and corresponding categories are most certain to be visited. Regarding the discussed features again, certainty can be computed as follows:

$$cDist_{cat_l} = \frac{avgGauss(cat_l)}{avgGauss(cat_0)} \tag{4}$$

$$cCheckins_{cat_l} = \frac{numCheckins(cat_l)}{numCheckins(cat_0)} \tag{5}$$

$$cUsers_{cat_l} = \frac{numUsers(cat_l)}{numUsers(cat_0)} \tag{6}$$

Here, $cDist$ is a category's certainty score based on the distance, $cCheckins$ is based on the number of Foursquare check-ins and $cUsers$ is based on the number of Foursquare users logged in. To take all the information into account a mixed model is built, using an average of all features.

$$cM_{cat_l} = \frac{\alpha \cdot cDist_{cat_l} + \beta \cdot cCheckins_{cat_l} + \gamma \cdot cUsers_{cat_l}}{\alpha + \beta + \gamma} \qquad (7)$$

This equation can also be used as linear regression model, with α, β, and γ being factors which could be evaluated with a supervised learning algorithm (e.g. Support Vector Machine), if there is enough representative data with ground truth (known destination POIs) available. For now, each of these variables is considered to be equally weighted, leaving the choice to change these weights to the analysts.

Visual Inspection

As discussed in Section 4.1.2 the reliability of the enrichment results highly varies. Uncertainty visualization thus becomes an important topic. An overview of existing techniques is given by Griethe and Schumann [2006]. Also, MacEachren et al. [2005] and Correa et al. [2009] review progress in geographic uncertainty research and show frameworks suggesting different visual methods for uncertainty representation depending on data properties such as type and quality.

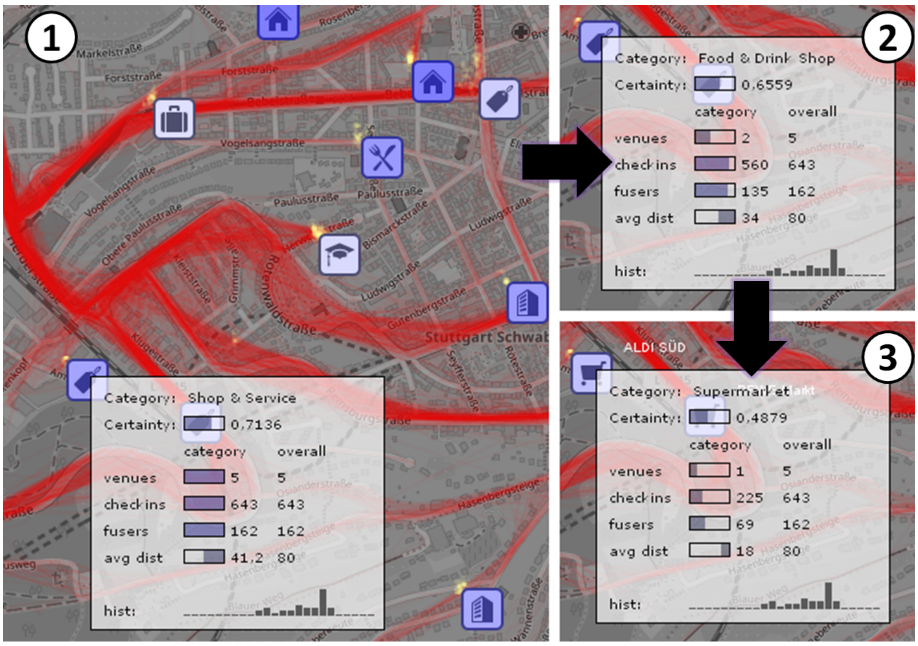
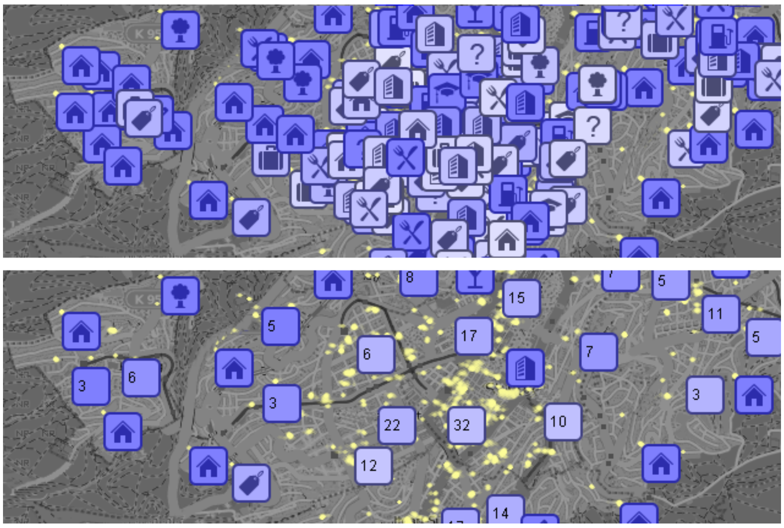


Figure 4.11 — Found POIs in the city area. The analyst can inspect the most likely POI information: (1) main category; (2) subcategory; (3) subsubcategory with POI name. With a higher level of detail certainty decreases. The histogram shows the daily visit distribution. The supermarket is mainly visited in the evening.

To visualize POI categories, the Foursquare icon set is used and a color mapping for certainty values is applied (see Section 4.3.2), as suggested for attribute accuracy by MacEachren et al. [2005]. The color gradient ranges from blue ($c = 1$) to white ($c = 0$), as can be seen in Figure 4.11. Simply visualizing every single POI in high detail (420 different categories), however, hampers fluent analysis and renders fast situation assessment impossible. To overcome this issue, the approach presented here makes use of two semantic aggregation methods. Firstly, information overload can be reduced by showing the main categories first (Figure 4.9, left), which can then be interactively explored by changing the current category level to see more or less details. By placing the mouse over an icon, a tooltip provides further statistics (Figure 4.11). Secondly, clutter resulting from overlapping items (see

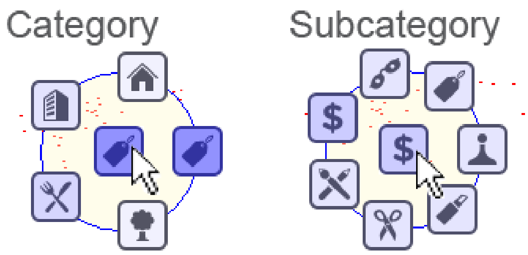


◄ **Figure 4.12** — Map snippet showing destination clusters and found POIs at category level 1. Top: Ungrouped icons may give better hints but clutter; Bottom: Aggregated icons avoid clutter.

Figure 4.12, top) can be eliminated using aggregation (bottom). Here, a mean-shift clustering algorithm is applied. At the beginning, each POI is in an individual cluster. Randomly, a cluster p is selected as starting point. Subsequently, other clusters are visited. If a cluster q overlaps with p it is added to p. An overlap is defined by a minimal threshold of the pixel distance between p and q. Afterwards, the cluster center is set to the mean position of its instances. Similarly, the approach accumulates the certainties of included POIs and computes an average. The result can be seen in Figure 4.12, bottom. The algorithm was chosen based on experiments on different datasets. It achieves overlap-free results after 5-10 iterations and can thus be interactively recalculated for different zoom levels.

Interactive Adjustment

With the presented means analysts can interactively explore the results of the automated enrichment approach that copes with large movement datasets. However, even when a clever decision model is applied, the accuracy depends on the underlying data source and POI service. If data is incomplete, imprecise, or even



▶ **Figure 4.13** — Analysts can use their expert knowledge and change preselected categories on each level of detail.

inaccurate, the model cannot always deliver the right solutions. For advanced control, a slider can be used to define a certainty threshold (similar to the approach in Section 3.3.3). This becomes extremely valuable for large datasets with dense information, i.e., in inner city areas, where the large number of POIs leads to more uncertainty. Depending on the task, the threshold can be increased to filter for only those POIs with an equal or higher certainty value. This leads to a lower recall (amount of POIs found) but higher precision. In addition, experts often have domain knowledge about context or understand coherences by manually aligning data. For example, they may recognize that a cluster has been formed on the large parking area of a warehouse store. Because the area also happens to house a small hot-dog stand it is falsely assigned based on its spatial proximity. The approach thus allows to change the icon and category on the current category level. In Figure 4.13 *Shops and Services* is pre-selected, since the certainty is the highest for all categories in that level. Considering advanced knowledge about a certain area, this category can easily be changed, e.g., to *Restaurant*. The same procedure can be applied on each category level. For example, the analyst stays with the *Shop and Services* category but on a lower level selects the hairdresser category instead of the pre-selected bank. Furthermore, the analyst can delete POIs she does not agree with.

4.3.3 Enrichment with Textual Content

POIs can give hints at movement purposes but often additional information is needed to understand a situation. Such information is available through social media, where people share their feelings, news, and observations. Unfortunately, augmenting movement data with microblogging content is challenging due to the large volume and heterogeneity of publicly available microposts. They must usually be presented in highly aggregated form to be useful for visual analysis. Straightforward means to visually depict aggregated microposts are provided by word clouds. They proved useful in summarizing text content by displaying the most frequently used terms as a weighted list [Gottron, 2009, Lohmann et al., 2009].

This visualization technique can also be used to enrich locations of interest with terms extracted from microposts. Figure 4.14 shows word clouds for two locations of interest on the map. The microblogging content was retrieved from a large database of Twitter posts, created during research by Thom et al. [2012]. The

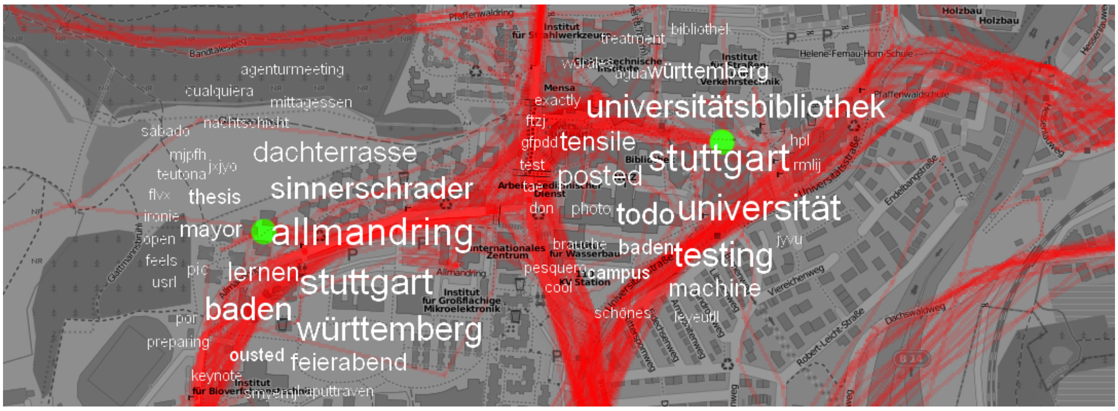


Figure 4.14 — Movement data is semantically enriched with microblog messages that give information about happenings and opinions during the same timespan and spatial area. This allows a deeper understanding for movement reasons. Frequent terms are extracted and visualized as word clouds.

database contains most of the publicly available geo-located messages, which are nearly 200 million per month (see Chapter 3). These posts are typically sent from mobile devices (e.g. smartphones). Roughly half of them contain a geotag (e.g. *Stuttgart*), while the other half comes with precise latitude/longitude coordinates. Only the latter is considered for the creation of the word clouds, as the former would be too imprecise to be plotted on a map with this level of detail. Furthermore, only microposts sent within a chosen timeframe and within a certain radius of the selected locations are considered. The decision of what a good radius is depends on the characteristics of the location and the micropost density in that region. For a detailed map of a well-populated area one would usually choose a relatively small radius in order to not let the context information interfere with that of a nearby location. The considered time frame for the microblog data would typically cover the timespan of the movements to enrich, but it could also be longer or shorter, depending on the interest of the analyst and the density of the available microposts. The selected microposts are processed with techniques commonly used in word cloud generation, including tokenization, stop word removal, and conversion to lowercase [Dork et al., 2010, Gottron, 2009]. That is, the microposts are split into individual terms, the terms are converted to lowercase, and stop words are removed. Then each term is weighted based on how often it occurs.

Finally, the visualization shows most frequently used terms in a circular word cloud. The frequency is encoded with text size, supporting the quick identification of the most popular terms [Lohmann et al., 2009, Thom et al., 2012]. As illustrated in Figure 4.14, the additional context information helps to better understand the movement data and related user behavior. For instance, the word cloud on the right-hand site depicts the German term for university library (*universitätsbibliothek*) in a large font size. This indicates that the library could be a reason why people parked their bikes at this location. As there exists also a POI for the library on the map, the additional context information would not have been necessary in this case. However, it supports the aforementioned assumption. In case of the left location of interest, by contrast, only few information is provided by the map alone. It shows neither names of buildings nor other information that helps with the interpretation. Here, the terms extracted from microblogging and visualized in the word cloud provide at least some hints. For instance, the terms *thesis* and *learning* (*lernen*) indicate that the buildings around the selected location belong to the university. One interpretation would be that they are student hostels, which is further supported by terms like *lunch* (*mittagessen*) or *roof terrace* (*dachterrasse*).

The tag map can also be used to overlay other textual data related to the movement data on a map. Some trips of the scooter data, for example, contain comments from the drivers, the VAST Challenge data has credit card billing information including location names, and the re-publica dataset contains information about presentation topics (see Section 2.5 for a description of the datasets).

4.4 Event Sequence Visualization

As discussed in Section 4.2, semantic enrichment adds another perspective on human movement data. With knowledge about behavioral episodes, movement can be described as a sequence of events (see Figure 4.2). To further verify the enrichment and analyze the data from a temporal perspective, two visualization approaches are proposed that highlight the temporal order and interplay of the happenings.

4.4.1 Cyclic Event Visualization

Timelines help to get an overview about the temporal distribution and varying frequencies of a dataset. For the enriched movement data with points and areas of interest, a temporal overview can reveal daily routines of human mobility such as private and business related behavior. A straightforward visualization of these sequences, however, reveals two scalability difficulties on vertical and horizontal axes. Firstly, sequences can be of different length. While it is possible to visualize somebody's sequence of visited places for a day, it can get difficult to do so for, e.g., a year whilst preserving both, overview and detail. Secondly, the dataset can contain not only a single person's behavior but the visit sequences of thousands. If mobility routines of a city or a suburb are of interest, the number of individual sequences would by far exceed the available screen space to visualize them at once. This leads to the conclusion that aggregation in both dimensions (for the movement objects and in time) is needed to preserve visual scalability and to allow for an overview on behavior patterns.

Depending on the analysis goals, different temporal visualization approaches are sufficient. For the electric scooter data (see Section 2.5.2) that spans over a long time range, a main goal was to cope with long time periods and to investigate mobility behavior of different temporal granularities. The developed visualization provides a cyclic view, based on a temporal aggregation for different time spans and an object-based aggregation by showing average frequencies of visited places instead of individual visits (see Section 4.3.2 for the enrichment with POIs). If the analyst is interested in daily behavior, the visualization subdivides the day into 24 sections. For each hour of the day, the sum for each category is calculated. The most frequently visited POI in the movement data is displayed on top, while less frequent ones are stacked below in descending order of frequency. The analyst can also switch to other cyclic aggregations to derive weekly (see Figure 4.15) or monthly patterns. By this means, the analyst can always identify the location

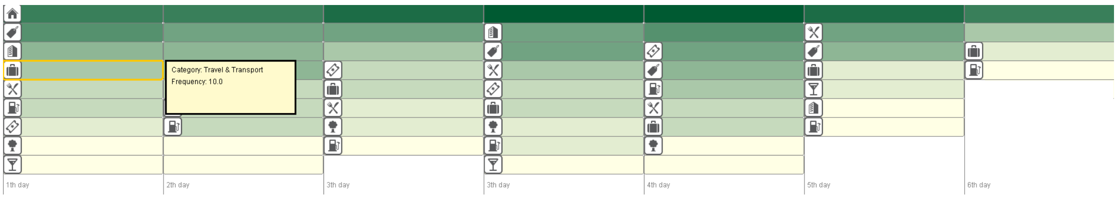


Figure 4.15 — Cyclic behavior patterns of electric scooter usage in Stuttgart, Germany. Every columns shows average daily behavior. For the filtered data, there are less visited places found on Saturday and Sunday (last columns).

most likely visited during the time span represented in a column, e.g., an hour of the day. However, the possibility that the driver has taken another means of transportation cannot be excluded. In order to remove periods where no movement occurred for a long time, the temporal aggregation can be restricted to POIs that have been visited for a duration not exceeding an upper threshold. For the electric scooter data, this threshold was set to 12 hours. In addition to the vertical order the frequency is mapped to a color scheme ranging from green (most frequent) to white (least frequent), which highlights the frequency distribution of the data. The visualization has many similarities to common streamgraph and Sankey diagram variations (see [Tufte and Graves-Morris, 1983], [Cui et al., 2011], [Byron and Wattenberg, 2008]), where the categories are shown as continuous stacked sorted ribbons. Although it might be easier to follow the temporal changes for a given category in these visualizations, experiments showed that they also lead to heavy clutter due to frequently crossing ribbons in the case of several changes in the category order. To understand mobility behavior it is more important to analyze the probability order of POIs visited at given time of the day than the temporal changes that a POI experiences. Therefore, the approach cuts the ribbon for each segment and indicates the category by a corresponding symbol at this point.

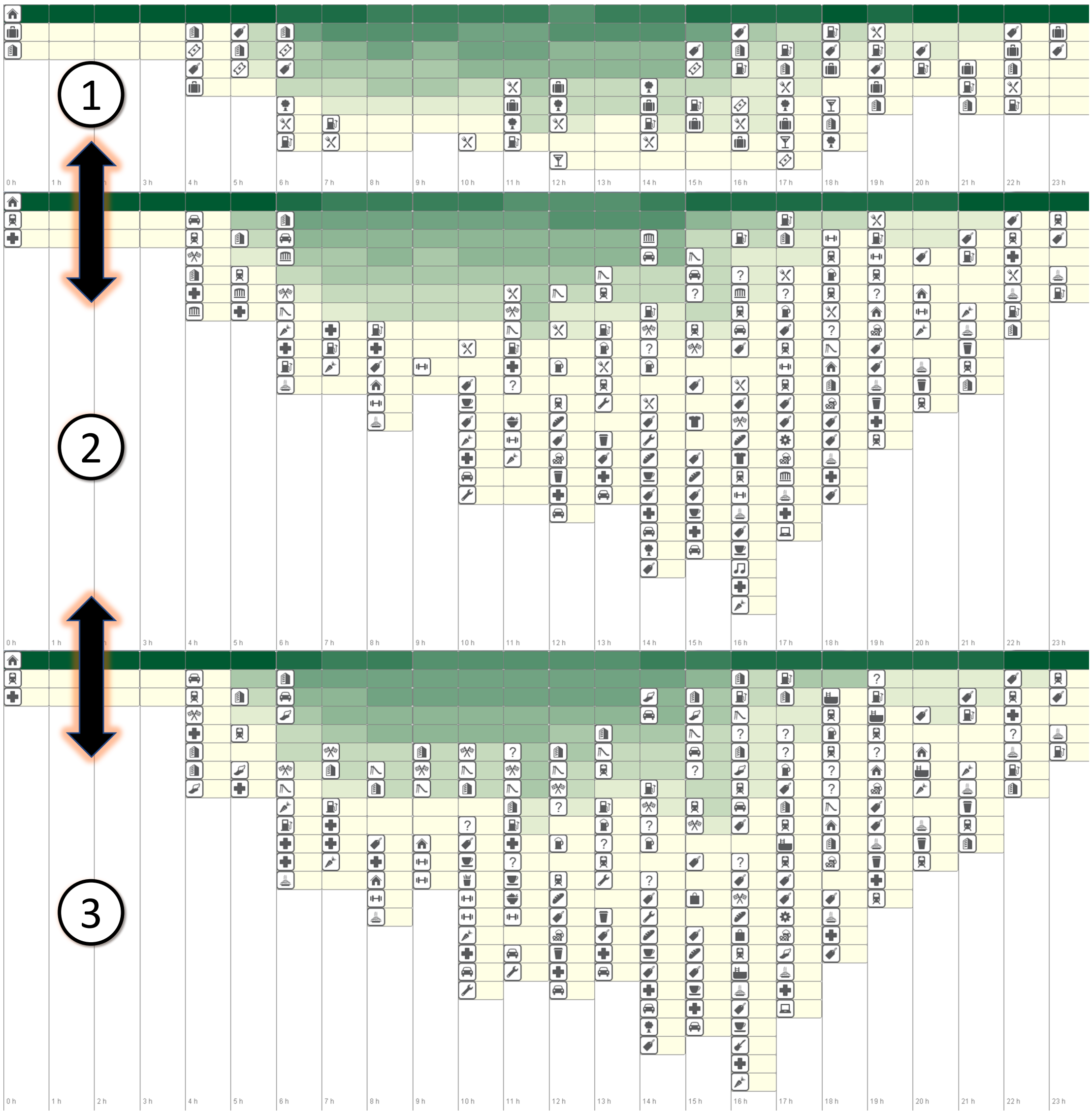


Figure 4.16 — The temporal component shows frequent daily patterns on different category aggregation levels. (1) The top level gives overview of the main categories like residence, profession, shopping, transportation, and food. Lower aggregation levels (2, 3) show more specific subcategories. A tooltip can be placed on the items to get more details and highlight matches in the map.

As pointed out in Section 4.3.2 context can have varying detail, and depending on the task analysts want to investigate rough patterns or drill down to specifics. Therefore, our approach implements a degree-of-interest function that visualizes the hierarchical data. Depending on the current level of detail, the POIs are aggregated according to respective categories (see Figure 4.16). Thus, the temporal visualization unfolds on higher detail levels (i.e., shows more distinct destinations) and folds up when lower detail levels like the primary categories are selected. The map view and the temporal view are interactively linked (see Figure 4.5). By investigating the destinations in the temporal view, details of the aggregated POIs (category type, number, frequency) are shown, and all corresponding destinations are consistently highlighted in the map, revealing their geographic position. Changes in the map view (e.g., selecting a more appropriate POI for a given destination) are also directly reflected in the temporal view.

4.4.2 Linear Event Visualization

While cyclic aggregation allows for a periodic overview of, e.g., average weekly or monthly behavior, it provides only limited capabilities to identify unexpected seasonality, trends, and outliers. Seasonal characteristics, for example, need to be known in advance to define the aggregation level for the cyclic wrapping. Also, differences on individuals or groups cannot be identified due to the object-based aggregation. Depending on the task it can thus be necessary to preserve details. To overcome the discussed scalability issues (Section 4.4.1) overview and detail is provided in separated but linked views [Cockburn et al., 2009].

A streamgraph visualization, e.g., ThemeRiver [Havre et al., 2000], gives overview by showing average frequencies of visited POI categories over time in stacked ribbons. To compute the visualization, the timespan of the movement data is separated into similar discrete timeslots (see Figure 4.17). Subsequently, an algorithm iterates over each sequence in the movement data and incrementally increases the count for a POI's category in the matching timeslot. The values are stored in a two-dimensional array. Figure 4.17 illustrates the counting process.

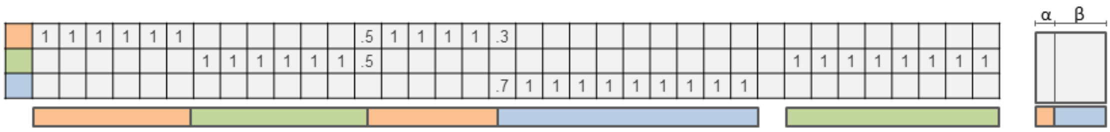


Figure 4.17 — Illustration of the event counting process to create a streamgraph. At the bottom, an exemplary event sequence is shown. In this case the events can be categorized in three groups (orange, green, and blue). A two dimensional array stores the amount of categories visited per timesplot. If an event ends within a timeslot only the proportional amount is increased (see α, β).

The visualization is rendered by drawing a polygon for each category (see Figure 4.18). To offset the polygons from the baseline, the y-values are shifted based on the previously rendered category (stacking). Figure 4.18 displays a streamgraph that shows the temporal distribution of POIs visited for the electric scooter data (introduced in Section 2.5.2). The colors represent the POI categories, resulting from the semi-automated enrichment process (see Section 4.3.2). As can be seen

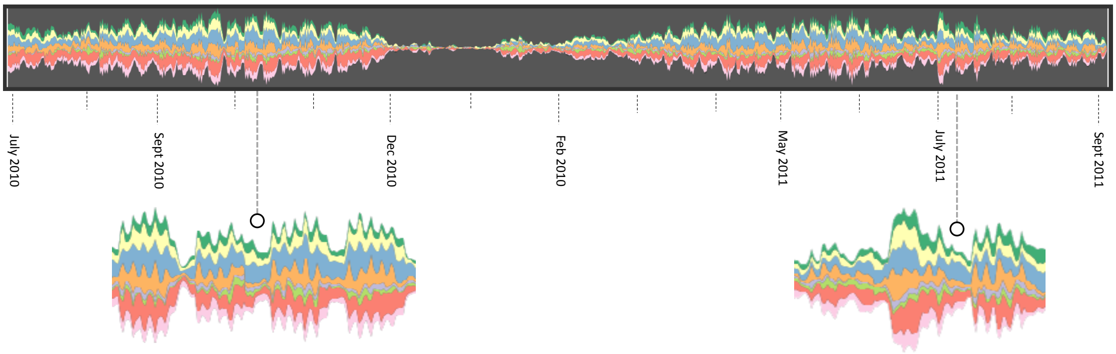


Figure 4.18 — A streamgraph gives a temporal overview on average frequencies of POI categories visited (here electric scooter data over 1 year). The scooters are less used during winter season (center). The bulging shows frequent usage during the week and less usage on weekends, especially on Sundays. Red color represents the food category (e.g., restaurants) and is still prominent on the weekends. The orange color shows business places, almost absent on weekends.

in Figure 4.18, the scooter study started in July 2010. During winter months (December till February) the scooters were employed poorly, probably due to the uncomfortable weather conditions. Then, in early 2011, the employment increased. There are several peaks, e.g. in May 2011 and especially at the end of June/ early July, in both cases after holidays ended. Taking a closer look at the patterns one can easily identify the five peaks (weekdays) that are clearly distinguishable from the weekends. While home locations and food places are still frequented on Saturday and Sunday, professional (working) places are visited less. More details are presented in the case study (see Section 4.5.2).

The overview (streamgraph) allows to select any part of the whole timespan for closer investigation. A detailed visualization will then show event sequences (sequences of visited places) for all moving objects. Figure 4.19 shows the electric scooter data with a selection in the streamgraph of May 2011 (a). A scrollbar (c) allows to browse through every single object's sequence (b), while the overview is maintained in the streamgraph. For the scooter data some usage patterns contain clear home-work-shopping-home activities while others are constantly employed for leisure or work only. The reasons are that some of the scooters were given to companies, while other owners live far outside of the city and do not employ them for work, etc. On average, however, one can still see daily patterns, as pointed out in Figure 4.18. Because many scooters are not used for weeks or even months, only the first 24 hours are integrated, because afterwards one cannot assume that the user is still at the same location (see empty spaces in Figure 4.19, b). The threshold to define the maximal timespan can be interactively adapted to the dataset. To take a closer look at certain point of interest categories, the analyst can activate a category filter panel (Figure 4.19, c). Here, single categories can be selected and deselected.

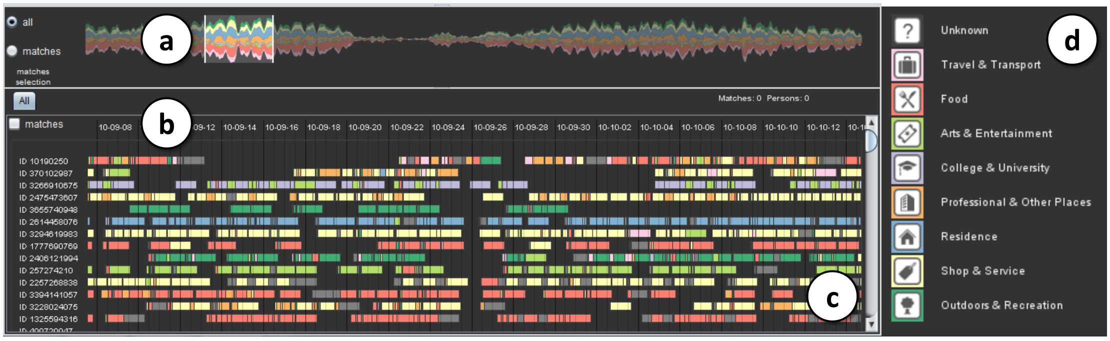


Figure 4.19 — Overview and detail: The streamgraph (a) provides overview of the category distribution over time. The selected area can be investigated in detail in the linear event sequence visualization (b). The categories of the visited POIs are colored according to the legend (c) that can also be used to select and deselect categories. A slider (d) allows to browse all sequences (here more than 500).

Accordingly, the streamgraph overview then only includes the selected categories and normalizes their volumes to the available space. Respectively, the detailed event sequence visualization greys out events (visits) of deselected categories.

4.5 Evaluation

To evaluate the approaches various case studies with different datasets and analysis goals are presented. For the semi-automatic POI enrichment approach a validation based on two different ground truth datasets is shown (Section 4.5.1), before different case studies in the application areas of customer analysis and law enforcement show the applicability of the approach (Section 4.5.2).

4.5.1 Ground Truth Validation

The two ground truth datasets cover two vacation trips, to Florida in winter/spring 2013 and to California and Washington in the autumn of 2012. Both datasets were recorded with a GPS tracker. While the Florida dataset has a tracking frequency of around 10 seconds, the California data is more precise, with a frequency of one to two seconds. Both datasets were not automatically separated into trips and had to undergo a proper preprocessing (see Section 4.3.2). Furthermore, in contrast to the Florida dataset, which was recorded in the car only, the California data was recorded while driving, walking, and even on the plane, meaning the tracking device was sometimes carried by the user and sometimes it was left somewhere. Both are interesting challenges for the POI enrichment, meaning the destination does not necessarily have to be the nearest POI. Thus, comparing POIs within a certain radius becomes even more important. First, the data was enriched with POI and category information from Foursquare. For the Florida dataset the query

► **Figure 4.20** — Results for model validation. All variants are compared to the ground truth. The nearest neighbor approach shows poor results on the holiday dataset. The number of users and the mixed model and manual assignment improve results significantly.

Model	California Data			Florida Data		
	Matches per Level (in %)			Matches per Level (in %)		
	L 1	**L 2**	**L 3**	**L 1**	**L 2**	**L 3**
Nearest	29	14	11	37	17	16
Amount	62	46	39	44	29	29
Distance	31	23	23	39	29	29
Check-ins	82	76	76	75	68	66
Users	85	82	77	83	78	76
Mixed	88	86	80	93	80	80
Manual	95	92	80	95	90	82

returns 46, and for the California dataset 117 visited POIs. Then, the system was presented to the persons who had recorded the trips. They reviewed their holidays using the replay mode proposed in Section 3.4.1, where they could follow their animated trip day by day, hour by hour in different playback speeds to validate the automated POI assignments. If the assigned POI category was wrong, the POI was relabeled using the approach shown in Section 4.3.2. They could also delete POIs if they could not remember or they were wrongly assigned. When the correct POI information was not available, they made notes.

After the ground truth information was gathered, different enrichment models (see Section 4.3.2) were evaluated. Figure 4.20 shows percentage values of correctly detected POI categories. Most interesting was that simply taking the nearest POI (most naive approach) led to bad results, even when aggregated (level 1). This is probably due to a large number of POIs, which are, however, very uncertain to be visited. Taking the number of POIs into account led to an acceptable result in the aggregation mode, but still a poor result in the most detailed mode (single POIs). Additionally taking the distances into account did not improve the results. However, in combination with the number of check-ins and number of users, the mixed model performed best, closely followed by the model based on number of users only. The correlation between wrongly detected POIs and the certainty value was also inspected. For example, with a certainty threshold of 40 percent on the mixed model the percentage of right detected POIs increased to 90 percent on the lowest level. There are few cases where the distance model and the nearest neighbor approach worked better. This occurred, for example, in the San Diego Zoo (see Figure 4.21, right) and in other areas where POIs are very dense. The last row of Figure 4.20 shows results which were further improved by a human. Here, another user without any ground truth knowledge was chosen to review the automated results and changed categories and POIs. For example, the Kennedy Space Center was detected as parking area, since the parking spot was simply too big and the entrance to far away from the parked car. Other refinements include hotel stays, which were wrongly detected as shopping and business centers, which did not suit the holiday context. Of course, the quality of manual corrections highly depends on

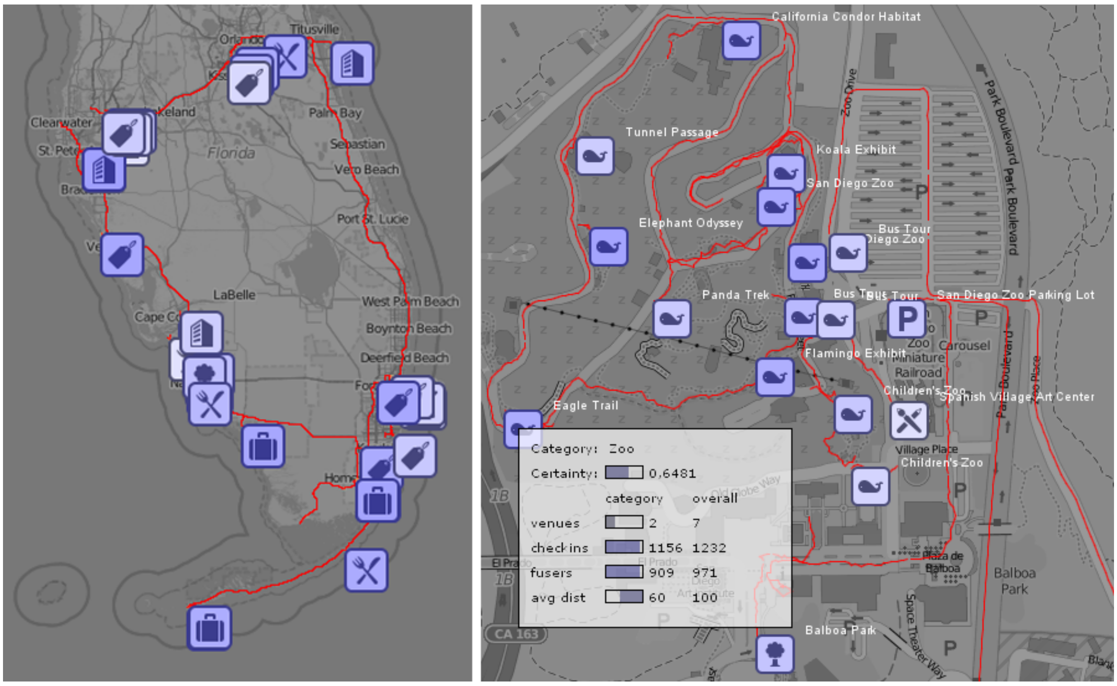


Figure 4.21 — Some results after the POI enrichment using the mixed decision model: Left) Florida trip on aggregation level 1 (overview); Right) A trip from the California dataset to the San Diego Zoo on level 3 (detail).

the analyst's domain knowledge. Overall, the results were very promising. However, for data in inner cities where POIs density is higher, the performance is expected to be worse. It remains to be validated in the future.

4.5.2 Case Studies

The applicability of the presented enrichment approaches was tested in different domains and for different datasets. The following section demonstrates two of them: The enrichment of electric scooter trips in Stuttgart, Germany (see Section 2.5.2) as well as the enrichment of rental car movements for law enforcement analysis as part of the VAST Challenge 2014 (see Section 2.5.4).

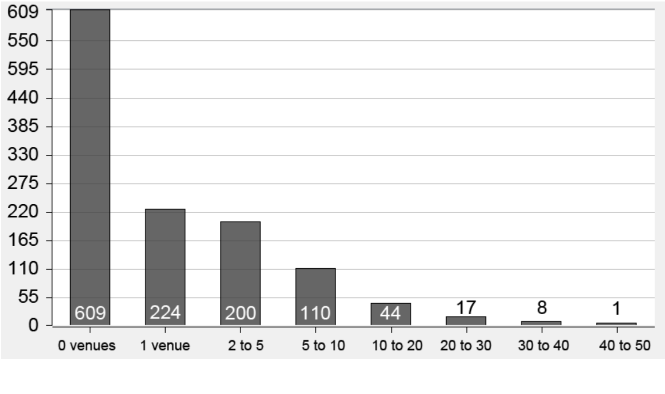
POI Enrichment for Electric Mobility

A main goal of the study on the electric scooter data (introduced in Section 2.5.2) was to investigate whether the scooters were accepted as a daily means of transportation and in a broader sense if e-mobility was able to find its way into every day life. This comprises to investigate for what purposes the products (scooters) were applied. To get a first glance about the spatial and temporal dimensions and characteristics of the dataset one can utilize different visual means that were introduced in Chapter 3. Overlays that show animation or a density image of the trips on a map already give a rough idea about movement purposes. To allow for an explicit visual encoding of semantics, however, a computational integration of movement and context is needed. To cope with the large data sizes, the semi-automated enrichment approach (Section 4.3.2) was applied. First, the stop points of the data were computed. From these locations, the clustering algorithm separated frequently visited places from noise. For the found clusters, Foursquare was then used to derive a number of nearby locations. When querying POI services there are several parameters to influence the search results. The geographic query radius for a given latitude/longitude pair is of importance. For the e-mobility dataset, a 50 meter radius around the cluster centers was chosen to be an appropriate range, as users often park their vehicles not directly at a POI but in short walking distance. Querying Foursquare with this setting returned POI (in Foursquare called venues) information for 607 clusters (see Figure 4.22). While most clusters received multiple possible venues, some clusters did not get a single venue assigned. These clusters are either private destinations (e.g. home, residence of a friend), charging stations, or anything unknown.

Home Location: Every scooter is assigned a home location based on the destination that is most frequently visited by the user. However, some scooters were used for testing purposes and are more frequently parked at various special locations (e.g. repair and assembly shops). Therefore, cleaning measures were used to remove vehicles that are missing a clearly identifiable home location.

Charging Station: A separate dataset contains exact locations of power charging stations in the Stuttgart area. Each cluster within a distance $\leq$ 50 meters to such a station is handled as charging cluster.

Overall, 74% of the found clusters were enriched with POI information. From 1,213 clusters, 266 were identified as home locations, 27 as electric scooter charging stations, and 604 clusters could be connected to venues via Foursquare. For some of these latter clusters up to 42 possible venues were retrieved in urban areas and often just one identifiable venue for clusters in more remote areas. Figure 4.22 illustrates the distribution of venues per cluster.



▶ **Figure 4.22** — Found venues per cluster, using a 50 meter query radius. For 604 of the 1,213 clusters venues were found.

The result of the enrichment can be visually inspected and adjusted (as presented in Section 4.3). This way, the analyst can get an overview. She may first investigate the distribution of categories using the category coloring before she switches to the uncertainty coloring. It can be seen that many of the enriched POIs are rather uncertain due to the high number of them in the inner city area or the wide distance to POIs in the outer areas. The analyst thus can make use of the precision-recall slider and adjust the threshold to at least 50 percent certainty. By taking a closer look on some locations the analyst can further correct POI assignments with domain knowledge. For example, in the north-east area of Stuttgart, one destination cluster is assigned an entertainment category. As this seems rather unlikely, the analyst activates the textual content layer (see Section 4.3.3) but chooses the scooter drivers' comments as textual data source instead of the Twitter messages. This reveals that there seems to be a repair shop. With this information she can reassign the POI. In a similar manner, she investigates different areas and fixes issues, especially for the most frequently visited destination clusters. A cross check of the temporal distribution of the data can be of value (see Figure 4.16). This way, data quality for further semantic analysis of behavior routines (see Chapters 5 and 6) is significantly increased.

Data Fusion for Law Enforcement

Another application field for semantic enrichment is law enforcement, where many different clues have to be brought together to reconstruct the situation of a criminal happening. In the VAST Challenge 2014 (see Section 2.5.4) a fictitious scenario takes place in the city *Abila* on the island of *Kronos*, where employees of a company named *GAStech* were kidnapped. Challenge participants had to provide digital forensic tools to analyze the situation and detect daily routines but also suspicious behavior.

Choosing suitable analytic tools is often a chicken-and-egg problem. On the one hand, without knowing the major data structure and properties, it is hard to find suitable visual representations, as demonstrated by Pretorius and van Wijk [2009]. On the other hand, without suitable representations it is hard to get a first impression of the data. To show plain and unfiltered but structured data, tables are generally an appropriate instrument [Rao and Card, 1994, Spenke et al., 1996]. Similarly, for an interactive exploration of the challenge data, the first view on the datasets is taken with a visually enhanced table, as proposed by Tweedie et al. [1994]. Even without knowledge about the data, one can calculate basic statistics such as mean and deviation. They are represented using visual primitives, which depend on the value scales (e.g. nominal, metric) of the feature (see Figure 4.23). The overview table (Figure 4.23, left) contains information about all employees, their transactions, and according POIs. Additional tables contain locations where transactions occurred and employees that performed the transactions. The tool further supports cleaning and filtering by various characteristics of the meta data

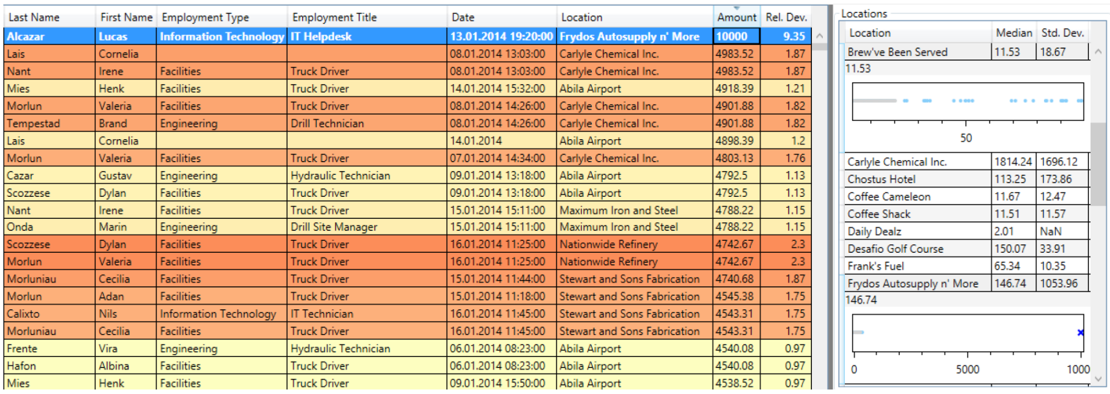


Figure 4.23 — Before the data is visually analyzed, analysts can get a first glance using an interactive table, enhanced with small multiple views that show distributions of selected attributes. In this case, a very high transaction of $ 10,000 at a car supply shop (highlighted in blue) is investigated.

in an easy and fast manner. For example, one can filter by means of deviating values, and by transactional mapping (credit and loyalty card data).

After getting an overview of the available data, its features, and its quality level (granularity, completeness, errors) the analyst can interactively use task-tailored visual tools for the enrichment. In the challenge 2014, POI information was partly provided with a comic-like image of a tourist map, which contains only imprecise location information. However, when combined with road and coordinate information, it can become very useful. As a first action, the analyst loads the image as a semi-transparent overlay and uses translation and scaling to adjust the datasets (see Figure 4.24, 1). That way, POI information is shown at the correct position.

As can be seen, the map contains different areas of interest, such as the company GAStech, public facilities like hospitals, restaurants, and cafes, but also housing and recreation areas. To get a semantic understanding of the movements, e.g., when and why employees visited which location and met others, the analyst makes use of the manual annotation approach, introduced in Section 4.3.1. She derives the relevant information from the tourist map and creates polygons that cover these areas, accordingly (see Figure 4.24, 2&3). For some areas, however, the tourist map does not give precise information. Not all POIs are covered or the exact location is not clearly marked. Unfortunately, the Foursquare service cannot deliver additional POIs in this fictitious scenario. Instead, likely POIs can be automatically extracted from provided transaction data. The credit card transactions contain location, username, and timestamps. By mapping the timestamp to the user's trajectory the algorithm identifies the most likely location. This also accounts for the loyalty card data. While the temporal resolution of loyalty card billings is lower, only including the date, one can still identify a likely location in some cases, e.g., when the rest

Figure 4.24 — The tourist map is loaded as map overlay and can be aligned to the underlying structure using translation & scaling (1). Subsequently, annotation is used to add areas of interest (2, 3). Accordingly, the linear event sequence visualization (4) reveals frequent behavior such as working, eating, and homecoming.

of the stops are already known. Figure 4.24, 2) shows a manual annotation of a park area, while next to this area automatically detected AOIs can be seen (small colored rectangles). Lastly, the movements are enriched with the AOI information. When an area of interest is extracted or manually defined, the corresponding event in the timeline is immediately colored by AOI category. This leads to a gradual reveal of the semantic routines and abnormal behavior. Figure 4.24, 4) shows the enriched movement data. Frequent blocks show repetitive behavior. It can be seen that the employees go to work on a daily basis (red blocks), except for the weekends. They usually grab a coffee on their way to work and back, visit restaurants for lunch (small cyan blocks) and drive home after work (larger blue blocks). Further analysis on this data is presented in Chapter 6.

Overall, the orchestration gives a new semantic view on the data. Especially a temporal view on the events (AOI visits per employee) can reveal routine and outlying behavior (Figure 4.24, 4). Means for a closer investigation of these patterns and for the extraction of findings will be presented in the next chapters.

CHAPTER 5

Interactive Filtering

In most data analysis scenarios, it is of advantage to begin data exploration with an overview visualization [Shneiderman, 1996]. In Chapter 3, different visual means to gain an overview were presented, including graph-based aggregation and edge splatting. At a certain point in the analysis process, however, details are needed to understand individual situations. For trajectory data, such detailed representation can lead to severe issues. In contrast to point-based data, trajectories already tend to overlap in small datasets and cause enormous clutter. When movement datasets cover thousands or millions of trajectories, it gets impossible to follow individual paths and extract patterns of frequent or outlying movement behavior. Additional context such as comments and points of interest even aggravates the situation. This is why sophisticated filtering techniques are needed that exclude specific attributes and ranges from an area of interest and allow to focus on the relevant. An ad-hoc definition of such filter criteria can be difficult, because analytical tasks are often underspecified and specific hypotheses and questions may not exist. In many cases it first needs to be explored what is in the data and what the data is capable of. Figure 5.1 highlights the selection and filtering step in the visual mobility analysis pipeline (Section 2.4.3), following on data retrieval and semantic enrichment.

A situation where exploratory filtering became necessary was given in the analysis of an electric mobility dataset (see Section 2.5.2). The driving force behind this study was a large energy provider who wanted to discover new business areas and was highly interested if electric mobility could find its way into everyday life. A deeper look into the energy provider's interests revealed multiple but rather vague questions concerning areas and times of scooter usage, the purpose of trips and visited places, the battery performances in different situations, and the road profiles taken. These

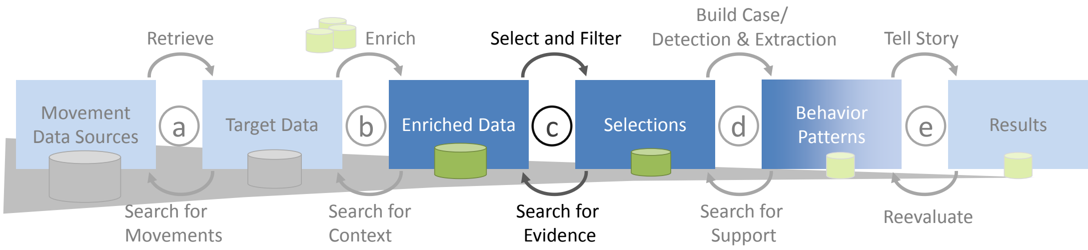


Figure 5.1 — This chapter addresses interactive exploration though set-based filtering and selection.

challenges led to the research and development of TrajectoryLenses (see Figure 5.2), a highly interactive exploration technique to gain insights from massive trajectory data. It extends the exploration lens metaphor to support complex, set-based filter expressions and the analysis of long time periods. Different spatial filter criteria are supported: The analyst can filter trajectories based on their origin, destination, or traversed area. To answer more complex questions, multiple lenses can be combined in set-operations making the approach flexible for many analysis scenarios. The user can comfortably handle these combinations in lens groups and does not need to specify or deal with the underlying mathematical operations.

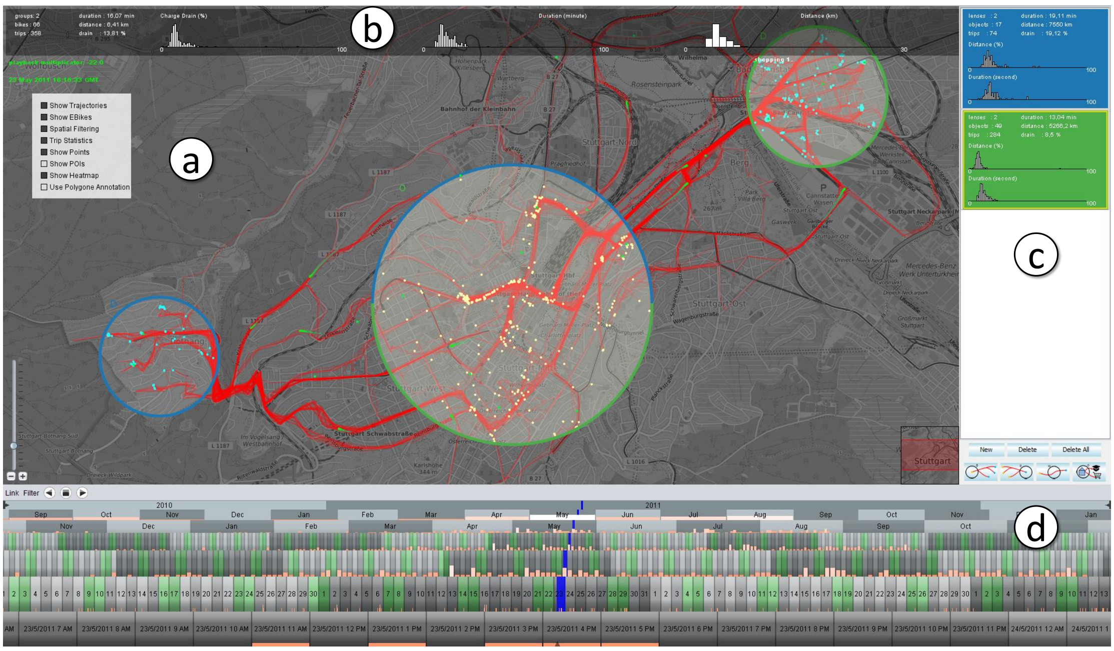


Figure 5.2 — The TrajectoryLenses interface: (a) map: interactive map with trajectories meeting the lens filter criteria; (b) result panel: details of the final filter result set (union of all groups); (c) group panel: create, select, and delete groups, group details; (d) timeline: hierarchical time slider and range selection.

In the following, related work in the field of exploratory filtering is discussed (Section 5.1). Thereafter, Section 5.2 describes the lens-based technique, the effectiveness of combined set operations, and the interaction with a hierarchical timeline to add time-based filtering capabilities for both, short and long time periods. Section 5.3 extends the approach by integrating the context resulting from the sematic enrichment (see Section 4). This comprises the integration of context in textual representation and points of interest. In Section 5.4, TrajectoryLenses is evaluated. Two case studies, expert feedback from the energy company, and a user study based on eye-tracking to verify design choices were conducted. While the studies primarily focus on the exploration of the electric mobility data, the technique was also applied to explore the reconstructed movements from Twitter, the VAST Challenge 2014 data, as well as taxi data from Hangzhou, China. Some of the gained insights are presented in the following sections.

Parts of this chapter have been previously published in:

R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl. TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data. *Computer Graphics Forum*, 32(3):451–460, 2013b

R. Krüger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 193–200. IEEE Computer Society, 2014a

R. Krüger, D. Thom, and T. Ertl. Semantic enrichment of movement behavior with Foursquare—a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 21(8):903–915, 2014b

5.1 Background

Thomas and Kielman [2009] state that transportation is one of the domains where rapidly expanding data volumes and complex analysis tasks create a need for visual analytics approaches for future developments. This thesis presents one such approach, which combines (1) lens-based and time-based focus+context methods for (2) visual filter and query construction in order to analyze large-scale trajectory data. To keep orientation during data exploration, the technique also comprises means for attention guiding (3). Hence, the related work to this approach is threefold.

5.1.1 Focus+Context Techniques

In many applications, the context of a focus region defined by the user or an algorithm is of great importance to the interpretation of the presented information. This has been illustrated by Furnas [1986] and Furnas [2006] for 'fisheye'-distorted views. A review of *focus+context* techniques, which try to maintain this context, was given by Cockburn et al. [2009]. An interaction technique often employed is the lens metaphor that changes the display between a focus and a context region by either showing additional information or hiding/distorting portions of the displayed information. A prominent example is *Magic Lenses* [Bier et al., 1993, 1997], describing lenses as a generic see-trough interface that lies between the application and the cursor. This concept has also been extended beyond the 2.5-dimensional space of regular user interfaces to tangible real-world objects [Spindler and Dachselt, 2010, Looser et al., 2007].

Lens interaction in TrajectoryLenses is exclusively used for the spatial domain of the data, and there is additional work on lenses more closely related to this. Graph edges are often drawn similarly to trajectories and Wong et al. [2003] present the *EdgeLens* to overcome occlusion problems in dense graphs by bending edges within the scope of the lens. For trajectories, however, the position of each element of a path is important and can not be moved independently from its accompanying context. The possibility of TrajectoryLenses to use multiple lenses in combination is shared with an approach by Panagiotidis et al. [2011]. Their work focuses on allowing drill-down interactions on large sets of densely packed or bundled edges in order to select specific sub-groups of edges. Butkiewicz et al. [2008] as well as *Semantic Lenses* [Rotard et al., 2007] display additional information about data in the focus area in an adjacent panel. TrajectoryLenses employs a similar method to show information in the vicinity of the lens or adjacent color-matching panels.

Regarding the temporal domain, TrajectoryLenses utilizes another focus+context technique, dubbed the *SmoothScroll* widget [Wörner and Ertl, 2011], which can be seen as the generalization of *Dual Scale Charts* [Isenberg et al., 2011] to n scales, all sharing a common focus point. For the trajectory analysis, a range selection similar to *ChronoLenses* [Zhao et al., 2011] was added to the SmoothScroll widget. In contrast to ChronoLenses, the selection does not change scaling information, but functions as a filter for subsequent spatial interactions.

5.1.2 Visual Filter Construction

TrajectoryLenses allows the exploration of a trajectory dataset and the simultaneous definition of spatio-temporal filters by providing immediate visual feedback on the result set of the current filter as has been suggested by Shneiderman's *Dynamic Queries* [Shneiderman, 1994]. Each lens defines a spatial filter, and several lenses can work in Boolean logic combinations. The *InfoCrystal* [Spoerri, 1993] is a visualization of all possible combinations of filters inspired by the Venn diagram, but this can become quite complex when involving more than 3-4 filters. The *Filter/Flow* representation [Young and Shneiderman, 1993] of Boolean queries focuses on user-defined combinations of filter terms instead of showing all possibilities and illustrates the size of the resulting set using the thickness of the line between the filter nodes. This helps to quickly judge the selectivity of filter terms, as has been demonstrated by Haag et al. [2014] for different RDF datasets. A similar approach is followed by the selection management component of the *PatViz* system [Koch et al., 2011, Bosch et al., 2009]. It is depicting the query in a graph structure. The filter definition is done in separate views that are specifically designed for displaying one or two attributes of the data elements. Similarly, the approach presented in this chapter defines the filter criterion directly in the spatial domain of the trajectories. *DataMeadow* [Elmqvist et al., 2008b] also employs a graph display, but at each node the arriving data is depicted as a star glyph, thereby providing context information for the filter definition. However, this is not a feasible way to meaningfully select ranges in the spatial domain.

FromDaDy [Hurter et al., 2009] is a trajectory analysis tool offering brushing-based set selection and set recombination. Its selection mechanism is closely related to the 'query sculpting' of *Rolling the dice* [Elmqvist et al., 2008a], where analysts can brush data elements in scatter plots of two freely exchangeable dimensions and then recombine the selected elements using conjunction or disjunction. However, once combined, the original sets are no longer available for interaction, which especially hinders the quick exploration of the data. In TrajectoryLenses, individual lenses of a combination can still be picked up and modified independently. The conjunction of filters in TrajectoryLenses is similar to the *Timebox widget* [Hochheiser and Shneiderman, 2004], but the explicit grouping of lenses additionally allows the disjunction of filters. Furthermore, by integrating group specific and aggregated data displays, TrajectoryLenses improves the exploration and comparison of alternative trajectories. In the work by Doleisch et al. [2003], users can brush fuzzy rectangular regions in 2D scatter plots, combine them with Boolean logic in a tree viewer, and show the result in a separate spatial view. The approach presented in this chapter simplifies the combination to only two levels: groups and lenses. Recently, Lu et al. [2015] proposed an approach that builds on TrajectoryLenses. They apply map matching to aggregate trajectories to a graph structure, and use combinations of lenses to filter routes in the graph. The map matching allows to cope with even larger data volumes, and to show diverse routes next to each other.

5.1.3 Visual Attention Guiding

Attention plays an important role in human perception. Koch and Ullman [1987] hypothesize that pre-attentive feature detection mechanisms operate in parallel and create a so called saliency map that contains the main features of a visual scene in an abstract way. These detection mechanisms are, among others, sensitive to color.

Itti et al. [1998] discuss similar theories, e.g., that visual saliency is computed in a pre-attentive manner across the visual field with 25 to 50 ms per item. Biederman's [1987] influential theory of scene recognition also states that a schema representation of a scene is activated within a few hundred milliseconds of scene representation. This has been studied on real world scenes. Yarbus [1967] revealed that people concentrate their fixation on the relevant aspects of a scene. [Chun, 2000] discusses different visual cues that draw our attention to those aspects and argues that visual context guides visual attention to fixate the most important with foveal vision. Moreover, he argues that *"contextual information provides useful constraints on the range of possible objects that can be expected to occur within that context."* [Chun, 2000]. Ceneda et al. [2017] present a taxonomy that categorizes the degree of guidance for visual analytic approaches into orienting, directing, and prescribing. TrajectoryLenses provides visual context, e.g., in form of heatmaps that aggregate underlying data to areas of similar density. These heatmaps do not only give overview on the data distribution but also guide the analyst's attention to areas of interest.

5.2 Interactive Visual Filtering

TrajectoryLenses extends the exploration lens metaphor to support complex set-based filter expressions and the analysis of long time periods, both adjusted to trajectory data. To filter and analyze trajectories, three different types of lens filters for different criteria (origin, destination, and waypoint) are introduced. By filtering the trajectories with these lenses, the analyst can access aggregated information like the number of filtered trajectories, the lengths, the time range, and domain dependent attributes. A possible analysis question could be which trips start or end in a certain area. These interactive filtering and exploration techniques are discussed in Section 5.2.1. However, there may be more complex questions with more criteria to be met. To address this demand, the lenses can be grouped with set operations to create combined geospatial queries. The mathematical implications are hidden behind the user interface, so that the analyst does not need to formulate complex logical expressions directly, but can simply arrange the lenses in groups. Section 5.2.3 will focus on these techniques. As trajectories describe a path through space as a function of time there is also a need for time-dependent filtering. Section 5.2.4 addresses this by combining the geospatial filtering technique with a hierarchical time visualization for selecting and browsing time ranges. It simplifies the navigation through long time ranges while retaining the possibility of making an accurate short-time selection.

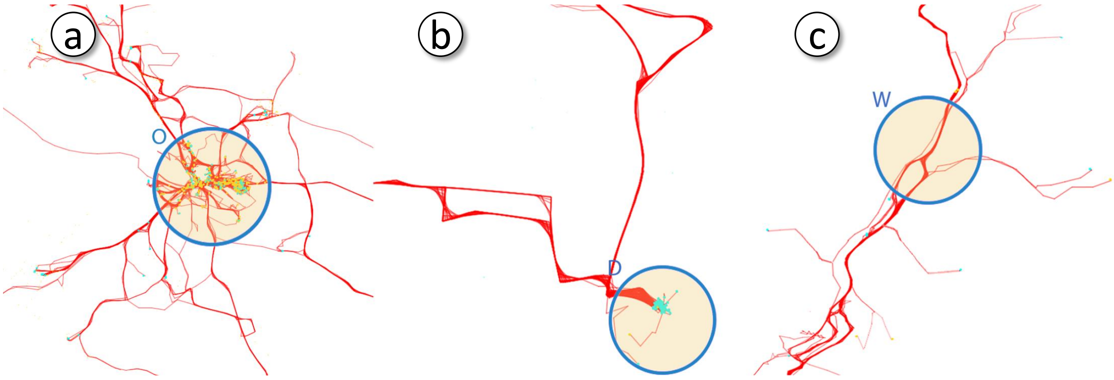


Figure 5.3 — Types of lenses with different filter criteria: (a) origin lens; (b) destination lens; (c) waypoint lens

5.2.1 Lens-based Filters

For the investigation of trajectory data, three different types of lenses are suggested (see Figure 5.3). The type of a lens determines the filter criterion for the data. Origin lenses filter for trajectories having their starting point in the lens area, destination lenses for trajectories ending in the lens area, and waypoint lenses filter all trajectories which somehow intersect the lens area, i.e., have at least one data point within the lens area. The user can set this filter criterion when spawning a new lens to the spatial visualization. For filtering by origin or destination points, each trajectory needs to be accessed once to retrieve its first or last element. This scales well for the 155,000 origin and destination points in the electric scooter dataset. The filtering for the waypoint criteria is more cost intensive, as all elements need to be checked. To fasten the filtering process for our large dataset of 8,190,507 elements, a quad tree is initialized with the spatial attributes of all trajectory elements to reduce the search effort. Hence, the filtering does not hinder interactivity.

For all lenses a circular shape is used, which is probably the most intuitive lens shape, showing all trajectories within a given distance from the point of interest. Furthermore, the tracking positions at the beginning of a trip have a random GPS error, which results in spatial Gaussian position variations. Therefore, a shape with a uniform surrounding area is helpful not to omit important tracks. This choice was also validated by expert feedback (Section 5.4.2). Depending on their needs, users can adjust the size of the lens, making it smaller or larger. They can move the lens over the underlying trajectories to explore the data and generate hypotheses. The implemented visualization is highly interactive, updating all displayed trajectories while the lens is moved over the data. Interactivity is still given when a filter set of 50,000 trajectories is drawn, which only occurs when the analyst positions a lens to include a whole inner city area. When an analyst found an area of interest, the lens can be dropped and stays on this position showing the filtered data until it is deleted or picked up again to change its size or position.

5.2.2 Focus+Context

Showing only data meeting the lens filter criteria, however, would lead to a lack of overview information, making it difficult to explore the data: Analysts would have to move the lens around without information on where they might find what they are looking for. Therefore, there is a need to show context information, depending on the lens filter criteria. This context includes all relevant elements that satisfy all filter constraints except for the one represented by the lens being moved. For an origin lens, this relevant context consists of the origin points, for a destination lens it consists of the destination points, and for the waypoint lens, the context are the entire trajectories. To improve the visual attention guiding density maps for the context points are computed. This allows to easily perceive areas where, e.g., many trips end or start. The approach uses slight distortion referring to the highlighting of covered origin or destination locations and the associated trips. As the map itself is fully interactive and zoomable, the technique refrains from magnifying the lens content because this would either occlude or distort the adjacent context.

TrajectoryLenses uses color coding to clearly distinguish between origin points (yellow) and destination points (cyan). Figure 5.4 shows an origin lens filtering the electric scooter data. All trajectories matching the current filter criterion are shown in red. The yellow heatmap provides a spatial distribution of trip origins, guiding the analyst to dense areas. Corresponding stop points appear in cyan. This dataset will be explored in more detail in the case study in Section 5.4.

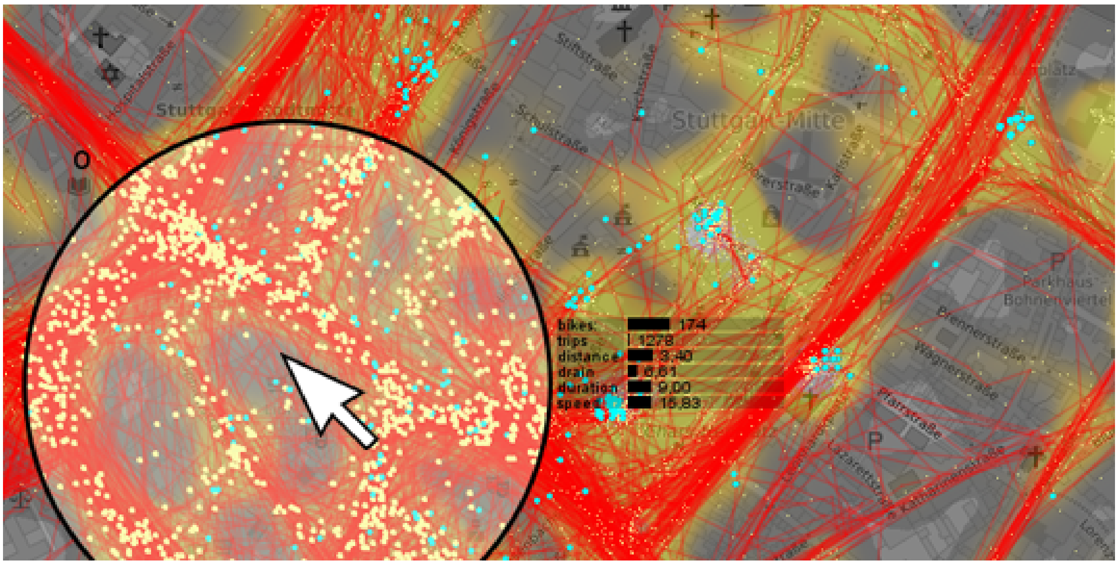


Figure 5.4 — The origin lens with context information. Trajectory origins are visualized as yellow heatmap, details are shown as yellow dots. The ones within the lens area are emphasized, and their corresponding destinations are shown (cyan dots). Different statistics are shown nearby the lens.

As analysts may not only be interested in the trajectories, but also in further statistics, e.g., distances, duration, and domain specific data, this information needs to be shown. For an exploratory analysis, it is important to focus on the lens. If context information was only shown in other views far away from the lens, the analyst could miss important information while switching views back and forth. Only a certain amount of information can be presented near the lens, because it covers the underlying data and there may be limited space and time to interpret it during the exploration. Therefore, the most important values are arranged as an interactive bar chart next to the lens (Figure 5.4), so that the values can be interpreted quickly. Further details (i.e., distributions, etc.) are shown in separated views, i.e., group panels, introduced in Section 5.2.3.

5.2.3 Set-based Filter Combinations

There may be questions more demanding than those presented in the previous section. The analyst may want to filter for trajectories originating in a certain area. Additionally, these trajectories might have to traverse certain areas before ending at their destinations. Regarding a dataset of traffic movements, there could be different driving patterns, and the analyst might be interested in only those from an origin O driving to a destination D or vice versa. To solve such tasks, the filtering technique arranges lenses in groups. Each group G_i, $i = 1, \dots, n$ combines its lenses by an intersection $G_i = L_1 \cap L_2 \cap L_3 \cap \dots L_m$ with each lens L_j, $j = 1, \dots, m$ containing a set of trajectories: $L_j = \{tr_1, \dots, tr_k\}$ with $k = |L_j|$. To also facilitate optional filtering (i.e., several valid origins, destinations or waypoints), the final result of our technique uses the union of the groups $R = G_1 \cup G_2 \cup G_3 \cup \dots G_n$. In conclusion, trajectories in the result set must meet the following filter condition: $g_1 \vee g_2 \vee g_3 \vee \dots g_k$ with $g_i = (tr \in G_i)$. As each logical formula can be expressed by a disjunctive normal form, one can formulate all logical expressions simply by arranging lenses in groups and groups in the final result set. This requires only two additional premises: The possibility of a lens being a member of multiple groups and the possibility to negate the meaning of a lens. Now assuming the analyst wants to filter for vehicle routes traversing either one of two areas to get from their origin to a destination, the trajectories have to meet this condition:

$$(l_1 \wedge l_2 \wedge l_4) \vee (l_1 \wedge l_3 \wedge l_4) \text{ with } l_j = (tr \in L_j) \tag{5.1}$$

L_1 is the trajectory set of the origin lens, L_2 and L_3 are trajectory sets for the alternative waypoint lenses, and L_4 is the set for the destination lens. As these terms can become complex to define and thus unreasonable for users without a mathematical background, the approach does not expect analysts to deal with set operations directly. Instead, they can create this query via the user interface as follows:

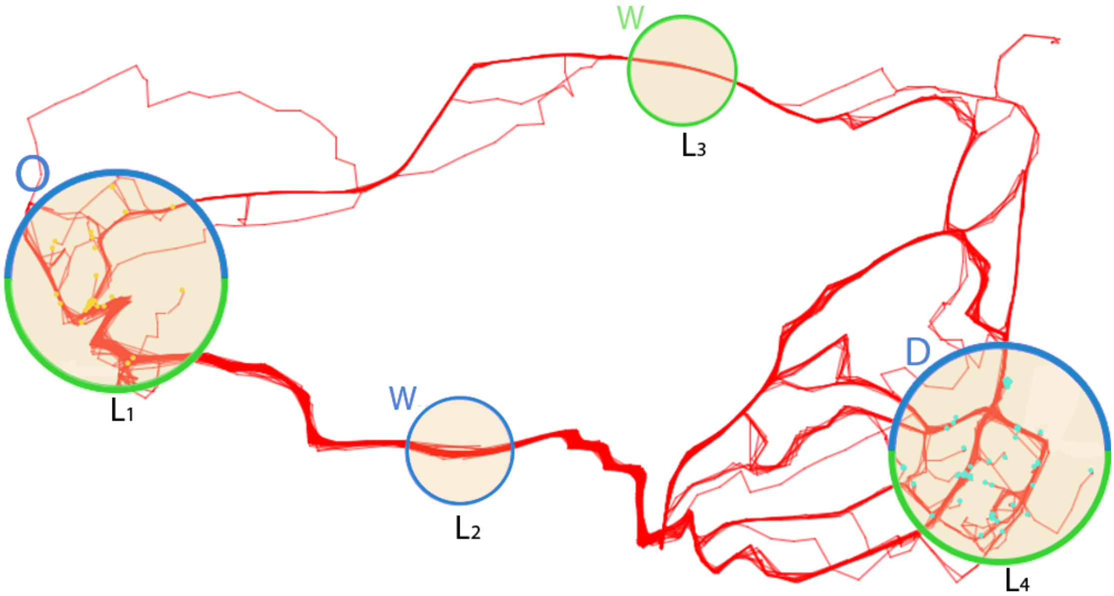


Figure 5.5 — Two filter groups (colored blue and green) filtering trajectories with a certain origin and destination area, traversing different waypoints.

1. Create an origin lens L_1 which implicitly defines a group G_1, adjusting and moving it to the origin point of interest.
2. Add a destination lens L_4 to G_1 containing destination points of interest.
3. Add a waypoint lens L_2 at a point of interest the trajectories must traverse on their way to L_1.
4. Create a second group G_2 and add L_1 and L_4 to it.
5. Create a waypoint lens L_3 in G_2.

Figure 5.5 shows the final filter setup and the trajectories meeting the criteria. The user interface supports the construction of such filter setups. The group panel (Figure 5.2, c) shows the generated groups. In addition to the bar chart mentioned in Section 5.2.2, the group panel shows more details about the filtered data of each group (i.e., mean values, correlations, and distributions). Here, the user can also spawn new groups and activate the group to which new lenses are added or lenses are to be removed from (see Figure 5.6). The lens border is always divided and colored by the amount of belonging groups. The upper picture (a) shows a lens in the blue group. Thus, it can be added to the active green group ('+' sign appears). The lower picture (b) shows the lens added to the green group. Thus, it can only be removed from the active green group ('-' sign appears). The information for the result set (union of all groups) is shown in the result panel (Figure 5.2, b). For the electric scooter data this information includes the average trip duration, battery drain, traveled distance, and the number of included scooters and trips.

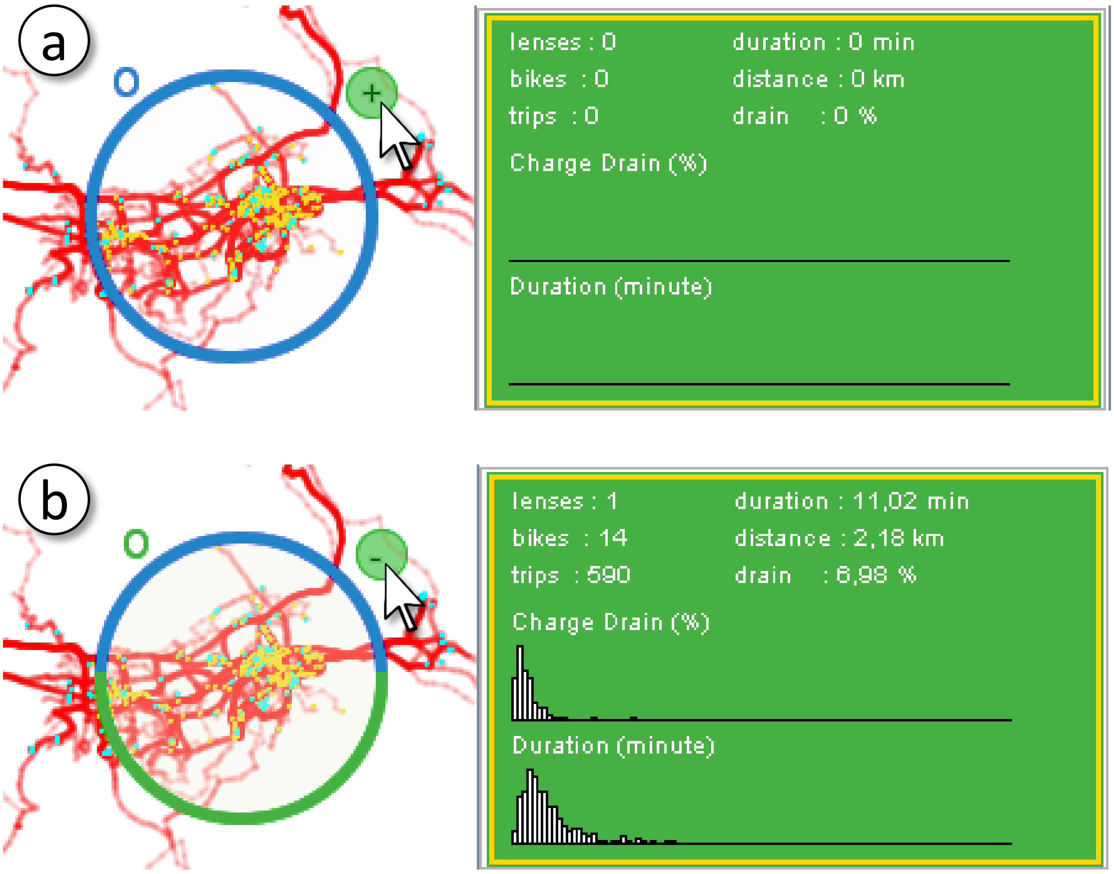


Figure 5.6 — Lenses can be added (a) and removed (b) from active groups via the '+' and '-' buttons, which appear on a dropped lens while hovering over the lens area.

5.2.4 Space-Time Exploration

For efficient temporal navigation TrajectoryLenses adapts the interaction control presented by Wörner and Ertl [2011] as the primary timeline view. This control extends variable-scale navigation [Ahlberg and Shneiderman, 1994, Appert and Fekete, 2006] by presenting the data range at multiple scales simultaneously. Our timeline is separated into a number of layers showing different scales and aggregation levels (Figure 5.7). The top layer displays years, the bottom layer hours. The intermediate layers use a gradual scale interpolation between those extremes and different levels of aggregation (trips per day or per hour). The most detailed layer displays a marker at the center of the control that indicates the current time position (just after half past 3 PM on 7 June 2011 in Figure 5.7). All other layers display a blue highlight that illustrates the time range that can be seen on the most detailed layer. Days marked in green indicate weekends.

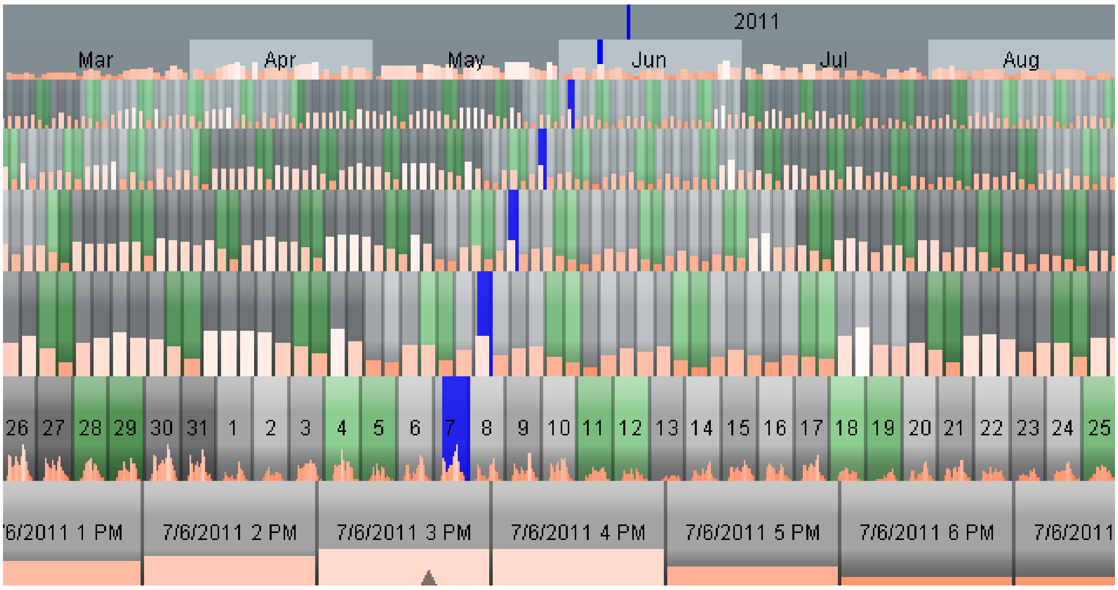


Figure 5.7 — A close-up view of the timeline showing layers at various scales and aggregation levels.

Because the timeline can be scrolled using any layer, a user can easily move through the data at various granularities and can even change the scrolling granularity by moving the cursor onto another layer while scrolling. While the top layer remains fixed, the intermediate layers move along with the current time position in accordance with their scale. The visual impression is that of a single, continuous display with a gradual progression from a finer to a coarser scale as opposed to several distinct bands showing different parts of the data, which scroll at different speeds. This establishes an intuitive, visual sense of context even when scrolling through the detail layer quickly. Clicking on any point on the timeline will set the focus point to the corresponding time, which allows for coarse long-range navigation, fine close-range navigation, and several intermediary granularities without having to switch between modes or views.

In addition to the various scale marks indicating years, months, days, and hours, the timeline displays a color-coded histogram of the trajectories selected by the lenses. Every item on a timeline layer shows a bar that represents the number of trajectories active during that hour, day, or month. There is also an option to subdivide timeline items: The second layer from the bottom is divided into days but shows the hourly distribution of trajectories over the corresponding day.

The functionality of the original control is extended by adding the possibility to define a time range selection. The analyst can drag the control display's start and end markers to change the start and end of the time range selection. The markers are drawn on every layer, which means that regardless of the current time position, they will always be visible on the topmost layer, and that the user can influence the granularity of the selection adjustment by dragging the marker on a coarser or finer layer. The time range selection will exclude trajectories outside the range from the other views.

5.3 Semantic Context Integration

In Chapter 4, various means to enrich mobility data with context information and different visualizations to gain an overview and alter the enrichment are presented. To enable a semantic exploration of human mobility behavior patterns, TrajectoryLenses has to be extended, accordingly. The inclusion of these semantic attributes in the exploratory filtering process, however, requires special treatment, because they contain more complex information that cannot simply be represented by numbers in the statistic panel. In the following, a tag cloud extension of TrajectoryLenses for textual data as well as a POI extension are presented. To reveal spatio-temporal dependencies, TrajectoryLenses is then coupled with the cyclic and linear event sequence visualizations introduced in Section 4.4.

5.3.1 Tag Cloud Extension

Text is the primary way of our society to communicate and archive thoughts. Tausczik and Pennebaker [2010] state that *"the words we use in daily life reflect who we are and the social relationships we are in"*. Text efficiently encodes information about opinions and sentiments, places and relations to other people. In Section 4.3.3, textual context enrichment was introduced as an effective means to derive further insights about movement reasons. Sometimes the movement data itself contains textual description about the purpose of a trip. If this is not the case, relevant context can be derived from geolocated news, event databases, or social media.

To better understand the filter sets resulting from the introduced origin, destination, and waypoint lenses, textual context is added to the result set visualization. Besides showing the filtered start and stop points and corresponding routes the most frequent keywords in the filtered data are visualized. The computation is simple. At first, a bag-of-words model is created by splitting all textual information into single words (terms) and counting their occurrences. Then, the most prominent terms are placed around the lens in a spiral pattern, starting with the most frequent word. This technique is adopted from Luboschik et al. [2008], and parts of the implementation are adopted from work by Thom et al. [2012]. Figure 5.8 shows mobility data derived from social media. The trajectories are reconstructed from geo-located Twitter messages (see Chapter 3). Thus, each trajectory point holds a short message of 140 characters. In this example, the dataset covers the Champions League final between Bayern Munich and Borussia Dortmund in 2013. The destination lens is placed around the Wembley Stadium area in London, where the game took place.

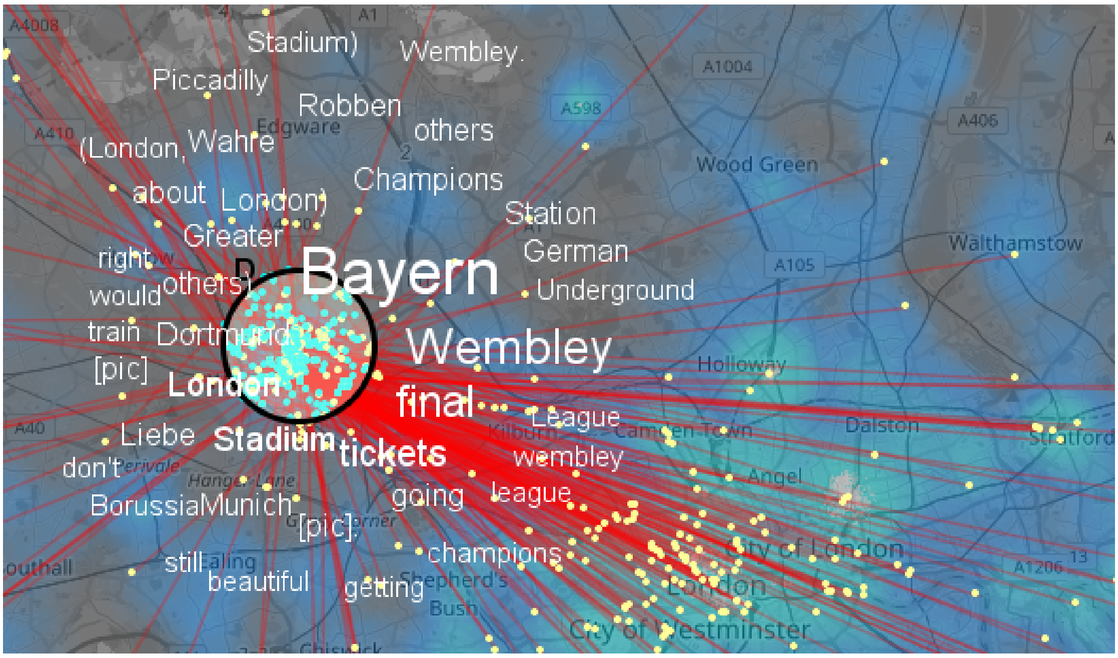


Figure 5.8 — The lens filters trajectories, reconstructed from geo-located Twitter messages, that end in the area around the Wembley stadium, where the Champions League final took place in 2013. The textual information helps to understand the movement purpose. The word cloud shows words in the Twitter messages used frequently, including the name of the event, teams and goal getters.

The visualization gives a joint view on shared messages of the Twitter users, as well as their trajectory. In this case, they come mainly from the inner city area. By zooming out one can also spot smaller clusters around Munich and Dortmund, where the teams and probably many fans come from.

5.3.2 Point of Interest Extension

For filtering the movement data based on the visited points of interest, the exploration lenses approach is extended to show POI symbols in the filtered area. This only acquires minor extensions for the existing lens types. In addition to start and stop points, the points of interest (see Section 4.3.2) that fulfill the lens criteria are visualized. This way, the analyst gets an understanding of movement purposes, such as shopping, working, housing, etc.

Besides the existing filters that were slightly adjusted, a new filter type was implemented. The *user lens* is motivated by the goal to understand human mobility behavior of residents of a particular area. For example, the stakeholders of the electric scooter project wanted the analysis to reveal information about the usage behavior of people living in specific suburbs (see Section 2.5.2). Instead of a heatmap that shows the context and guides the analyst to dense areas, the movement object's central location is shown (here: home location of the user), which was computed with a rule-based algorithm that clusters frequent stop points with certain constraints like stays over a long time, night times, etc. (see Section 4.5.2). Figure 5.9 shows a *user lens* filtering for all trips of users that have their home location within the lens area. This way, area dependent behavior can be investigated.

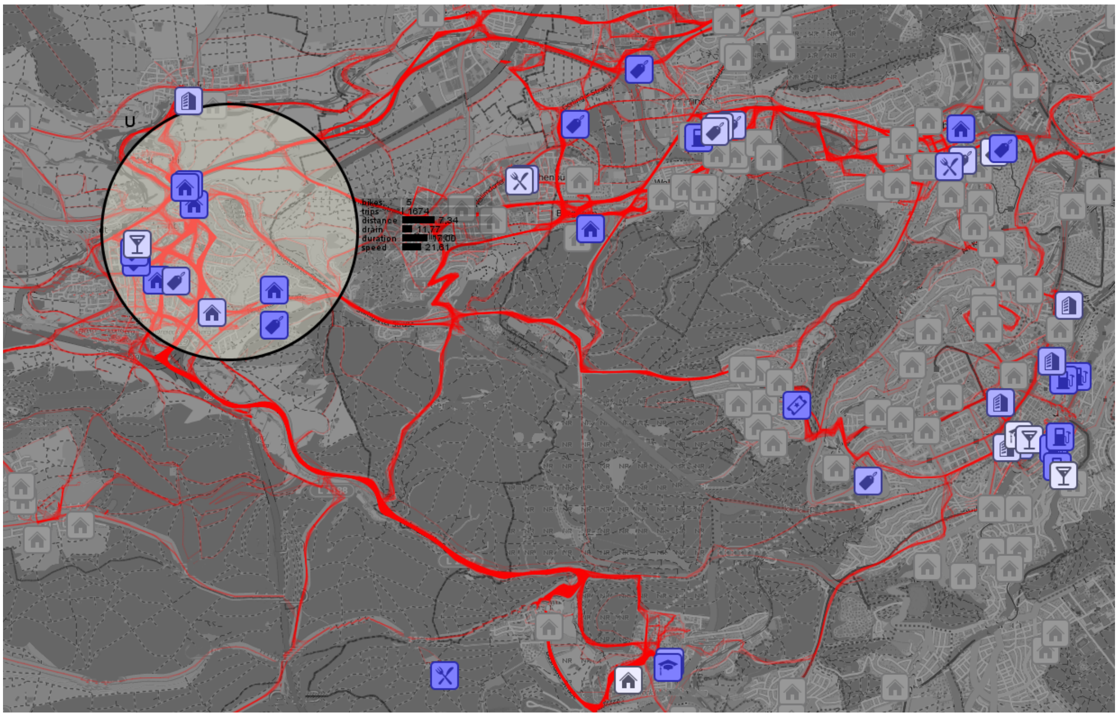
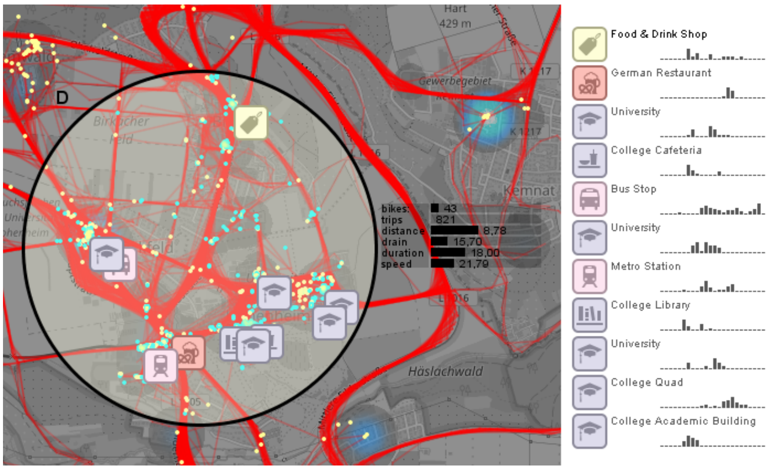


Figure 5.9 — Map snippet. All users living in the filtered suburb are selected. All their trips are shown in red. Icons show frequently visited POIs, colored according to their certainty.

5.3.3 Space-Time Exploration

To analyze spatio-temporal dependencies, the lens extensions can be combined with the different temporal visualization introduced in Section 4.4. Dependent on the analysis goal and the data characteristics, the analyst may choose the linear event sequence view to investigate temporal routines and outliers or select the cyclic visualization to investigate daily, weekly, or monthly temporal distributions. If details about individual POIs are in focus and the filtered data is relatively small, the POI histogram panel may be the right choice (see Figure 5.10). In all cases, the timeline is connected to the map view via brushing and linking. The visualization is fluently refreshed according to the current spatial filtering. Also, the analyst can select any event in the timeline to brush (hover over) or filter (click). The map



◄ **Figure 5.10** — The histogram view shows single daily visit distributions for each filtered POI. Through highlighting operations the analyst can identify spatio-temporal patterns. POI coloring is based on category.

view then only shows the selected event and connected trajectories, e.g., from or to the event. Figure 4.5 shows a filtered subset of the electric scooter data using the destination lens. The cyclic event sequence visualization highlights the temporal distribution of the visited places, accordingly.

5.4 Evaluation

Different evaluations of the developed techniques were carried out. Two case studies demonstrate how TrajectoryLenses can help to solve challenges in the field of mobility analysis. An implementation of the approach was also shown to three domain experts working in the research department of the energy provider. They gave oral feedback as well as ratings in a questionnaire. For a bottom up evaluation of how the focus+context technique works in detail, and to verify if the right design decisions were made, eye movements and device-based interactions of five visualization experts were tracked and analyzed.

5.4.1 Case Studies

The two case studies both address domain specific questions on the electric scooter dataset (see Section 2.5.2 for an introduction to the application scenario). Three of the most important questions that the energy provider specifically had were 1) for which purposes the people use their scooters, 2) whether they use them intensively for their daily commuting needs, and 3) whether there are any specific caveats or benefits resulting from choosing the electric scooters as a means of transportation. Questions 1 and 2 are addressed in the first case study, revealing visited places and temporal usage patterns. The second case study is centered around the extended capabilities resulting from the grouping technique and the analysis of battery drainage behavior, relating to question 3.

Consumer acceptance

Since the Stuttgart region has a rather undulating landscape, and the power of the scooters is limited to 2 kilowatts, long-term consumer acceptance, particularly of users living in more remote areas, was one of the most relevant issues for the energy provider. To address these questions, the analyst first gains a general overview using the edge-splatting approach presented in Section 3.4.2. Subsequently, she begins to examine usage behavior of participants that employ their scooters for daily trips to work. When spawning a destination lens, all trip-endpoints are accumulated in a heatmap that overlays the map. This way, one can now easily recognize clusters, which indicate important destination areas for the participants. By additionally activating the POI layer, the results from the semantic enrichment (see Section 4.3.2) indicate the most likely visited places. For example, a large cluster with assigned business location appears on the parking areas of most larger employers in the Stuttgart downtown area. The analyst thus places the destination lens over one of these clusters, which highlights all trips to this destination. Usually, some of the

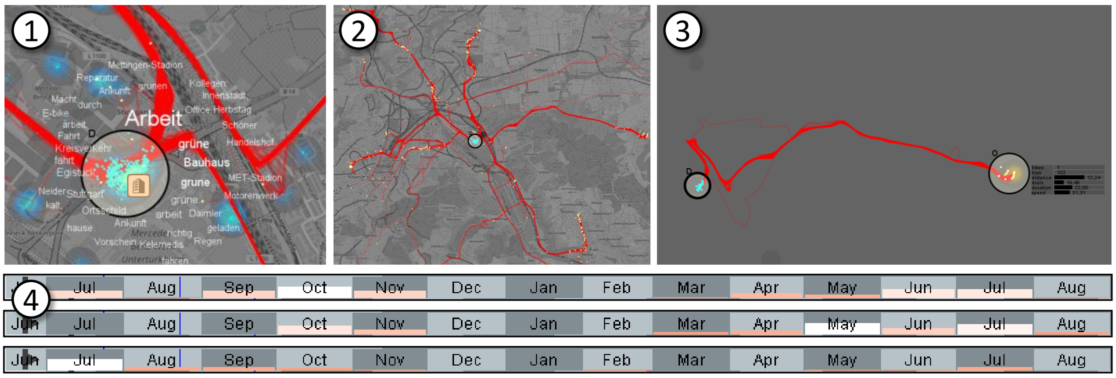


Figure 5.11 — Analysis of daily usage behavior: First, a cluster on the parking area of the company is selected, using the destination lens (1,2). The origin lens is then applied to examine individual participants (3). The temporal histogram gives information on increases and decreases in usage behaviors (4). (In the right image the map has been hidden in order to protect the privacy of the involved participants.)

trips repeatedly occur on the same routes, which is indicated by thicker red lines, and they are thus presumably generated by employees that visit the site on a regular basis. To investigate one of the larger companies (see Figure 5.11, 1), the analyst activates the tag cloud extension. This reveals frequent annotation terms such as *work*, the name of a car manufacturing company, *engine works*, and *Bauhaus*, a hardware store nearby. The analyst sees approximately five supposed employees that frequently drive to this area (see Figure 5.11, 2) and can now investigate their individual usage behavior. For this purpose, she spawns an origin lens and sees the available starting points of the trips having the destination lens area as endpoint (see Figure 5.11, 3). The clearly visible clusters will usually correspond to the participants' home locations, which is also indicated by the assigned POIs. Subsequently, she now moves the lens over one of these clusters and sees the movement behavior of the respective scooter driver. Using the trip histograms in the timeline view, she can investigate whether the intensity of usage increases or decreases over time (4). In the investigated examples, most of the users have an overall continuous usage behavior, which is only suspended during the winter months. However, there is also one user with a clear decrease in usage intensity in the second summer of the observation period. Due to the limited capabilities of the scooters, there are very few road trips or intercity rides.

◄ **Figure 5.12** — A large cluster is inspected by a destination lens with activated tag cloud that displays the most frequent terms in the trip annotations. It reveals that this area locates a large repair station for the scooters. The words are German and translate to, e.g., *customer service*, *repair*, *brake pads*, and *spokes*.

In addition, the analyst now also investigates other usage domains for the scooters by removing the destination lens and looking at the other trips of the employees. This way, she learns that the scooters are used for shopping, e.g., trips to a local hardware store, regular visits to other residential areas, and round trips for testing purposes. By activating the tag cloud feature, the analyst can gain further insights about often visited areas. The textual content shown in Figure 5.12 is gained from the users' trip annotations (see Section 2.5.2). In this case, the cluster filtered by the lens locates a repair shop for the scooters. Comments on the trips to this location thus reveal that people took their scooters there for *inspection* and *customer service*.

To compare consumer acceptance in different city areas and to investigate the range of uses that people have for their scooters the analyst makes use of the user lens (see Section 5.3.2) and places it over a specific suburb. As the lens is adjustable in size, it can be enlarged or shrinked until it covers the area to be analyzed. In this case, the analyst selects an area which is about 20 kilometers outside of the city. Besides some technical statistics for the current selection (e.g., average distance, number of trips, and battery drainage) the analyst also sees the tracks of the users and the corresponding POIs. In Figure 5.9 it can be seen that the main destinations are characterized in the categories *Professionals*, *Shops and Services*, and *University*. This reveals that these scooters were probably employed for daily transportation means, like work, shopping, and education. The thickness and intensity of the highlighted routes also gives the analyst a general idea of how frequently certain POIs were visited. The certainty coloring of the icons (introduced in Section 4.3.2) reveals that it is quite sure the selected users employ their scooters to head for the university. Also, there are office places detected. However, this information is less certain due to other POIs found nearby. For more details, symbols can be examined by mouse-hover to show a tooltip with more information, and the level of detail can be changed, as described in Section 4.4.1. To investigate the discovered locations, the analyst can now check other POIs that were found for a destination by clicking on the icon. This way she discovers a restaurant next to the offices, but it is less frequently visited by Foursquare users and thus also less likely to be a frequent destination for the scooter users.

To investigate the periodic frequencies and temporal behavior she continues her analysis using the temporal view. It shows that for the selected scooters the home location covers most of the time, indicating that the scooters are not used every day (see Figure 5.13). In addition, one can see that university and business areas

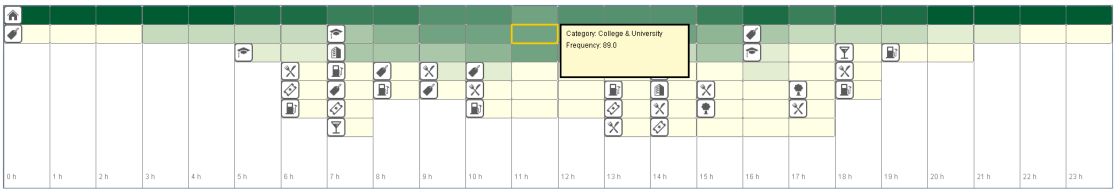


Figure 5.13 — Temporal component showing frequent temporal patterns per hour on a daily basis. It can be seen that at night times, the selected scooters are usually at home, while they are used to get to work and education at daytimes. In the early evening shopping becomes more frequent.

were primarily visited during the office hours. In the afternoon, other POIs are also visited frequently, such as the category *Outdoors*. People obviously use their scooter to go shopping and for dinner as well. By hovering over the POI bars in the temporal view a tooltip provides more detail about the POIs, e.g., how often they were visited at this time of the day during the year. The analyst switches to a different template, showing only days during weekends. This reveals that shopping is a frequent activity for which the scooters were used during leisure times. Referring to the question 1 and 2 in the beginning of Section 5.4.1, the case study exemplary showed that scooters are indeed used for daily commuting but also for various other daily activities.

Battery drainage

Another important question for the electric vehicle scenario were the possible shortcomings resulting from the small battery in combination with the low-performance engines of the devices. The correlation between the altitude profile of a trip and the battery consumption of the vehicle were analyzed in multiple cases. For this kind of investigation, one can try to find destinations that can be reached via known alternative routes and compare the sets of trips for each alternative. In this case, the task focused on the analysis of different ways from Stuttgart Obertürkheim to the suburb Vaihingen.

The analyst begins the analysis by spanning an origin lens over Obertürkheim (see Figure 5.14) and placing a destination lens over the Vaihingen area (outside of the image). It can be seen that the scooter drivers often reach their destination using routes through *Rohracker* or *Frauenkopf*. At the same time, one can observe that the scooter drivers also often take a route through *Sillenbuch*. The analyst now spawns three waypoint lenses for each route to investigate their differences. However, as long as all lenses are in the same group, all trips will be filtered out,

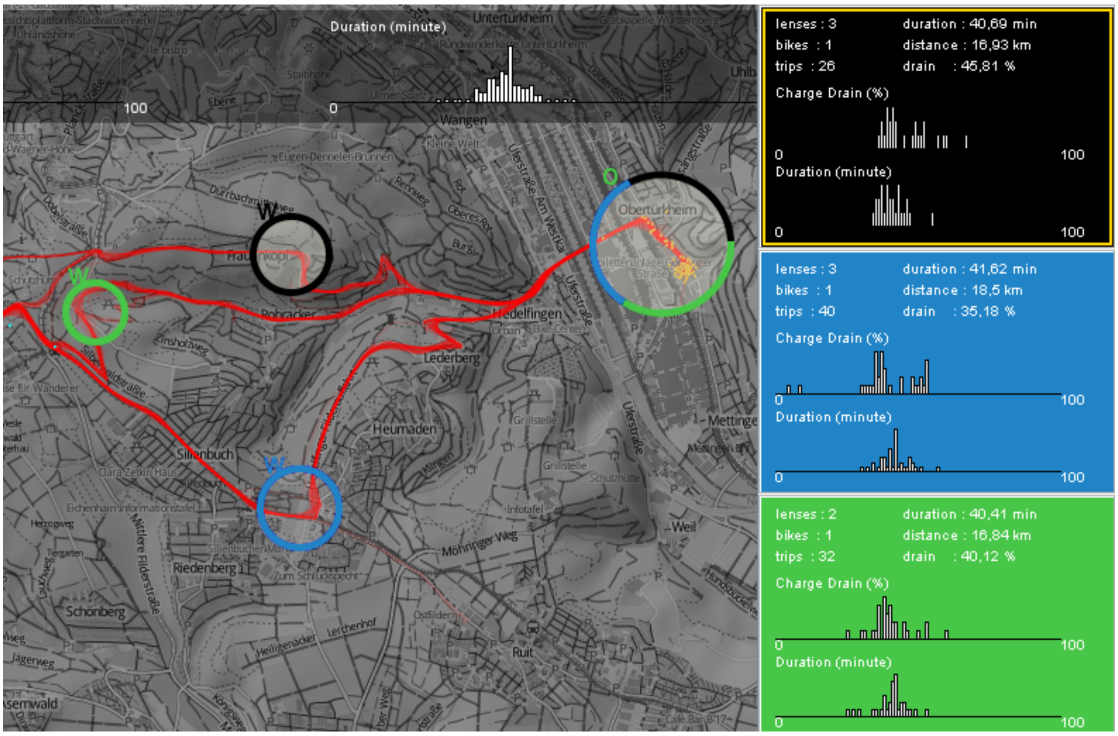


Figure 5.14 — Using three lens groups the analyst can investigate the battery drainage behavior based on three alternative routes from Obertürkheim to Vaihingen (outside of the image). Darker map areas show hilly terrain.

since the lenses would now only allow trips originating in Obertürkheim, ending in Vaihingen, and moving over *all three* waypoints. Therefore, she creates a new group for each alternative to include both the origin and destination lenses but *just one* of the waypoint lenses to create three logical disjunctions of the three alternatives. By looking at the three group info panels, she can now investigate the aggregated attributes of the alternatives (see Figure 5.14). At this point, the analyst observes that Rohracker (green group) and Frauenkopf routes (black) are only 17 km long on average, while the route via Sillenbuch (blue) has a length of 18.5 km. When comparing the average battery drainage between the alternatives, however, one can observe that although the distance of the Rohracker and Frauenkopf alternatives is shorter than the Sillenbuch route, the battery drainage is much higher (+5% and +10%) than the drainage using the Sillenbuch route. This is probably due to the steep slopes on the routes of Rohracker and Frauenkopf, which can be seen when switching to a topological map view (Figure 5.14). Based on these observations and the investigation of other similar case studies, one can form the hypotheses that steep slopes have a dramatic impact on battery drainage and should be avoided in favor of longer routes if energy preservation is desired. Even then, the high battery drainage limits people's mobility in many cases.

5.4.2 Expert Feedback

For further evaluation, the approach was presented to three employees of the energy provider working in the management of electric mobility research and projects. They first got a brief introduction and then had time to investigate the approach including parts of the application scenarios from Section 5.4.1. In the third part of the meeting they answered a four page evaluation survey. Overall they argued that the system supports the analysis of the relevant questions, and they were highly interested in using the approach also for other mobility field studies and fleet management. Figure 5.15 shows their ratings regarding the usefulness/relevance of different analysis goals, such as intended scooter usage, statistics about speed, battery drainage, etc. One can see that the experts were especially interested to find out about the intended usage as well as temporal and spatial patterns (usage

I estimate the usefulness of the following aspects as high:	fully agree				disagree
Intended usage (what for, e.g., work, shopping, leisure)	2	1			
Usage area (suburbs, routes, living areas)	2		1		
Speed statistics		2			
Route profile (street types, length, etc.)		2			
Battery drainage	1		1		
Charging (based on users, stations, capacity utilization)	1	1			
Temporal patterns (time of year, day, etc.)	2				
Correlation to weather (temperature, rain fall)	1		1		
Correlation to other events (traffic, cultural events, etc.)		1	1		

Figure 5.15 — Evaluation results regarding the usefulness of different analysis tasks. Especially intended usage and spatio-temporal patterns are of interest. One of the three experts rated only the first two features.

area). They mentioned that they mainly work with spreadsheets and only use simple visualizations like bar charts and graphs and thus really liked the idea of presenting the data in an interactive map view. Area related usage could help them to define future application scenarios, which are still not fully set. Respectively, in the related free-form field one expert wrote that the scooters might be used in metropolitan areas with 50.000 or more inhabitants, while another expert expected the usage more in rural areas, where users have more possibilities to charge their scooters at home.

The following is intuitive:	fully agree				disagree
Application in general	1	1	1		
Usage of lenses	2	1			
Grouping/Combination of lenses	1	2			
Temporal navigation and selection	1	1	1		

◄ **Figure 5.16** — The three experts rated the intuitiveness of the application, lens-based filtering and combinations, and temporal navigation and selection.

Furthermore, questions about the intuitiveness and usefulness of the approach were asked. According to the questionnaire, they deemed the approach highly intuitive, especially the spatial filtering (see Figure 5.16). The combination of lenses was rated as intuitive. The intuitiveness of the temporal navigation received neutral ratings, which might indicate a necessity for familiarization. Aligned with their rating of a spatial representation, they rated the implemented map representation as highly useful, as well as the geographic selection with the lens tool. Overall, they agreed on all features asked to be useful (see Figure 5.17). Regarding the shape of the lens opinions diverged. One expert argued that other shapes may be helpful in addition such as ellipses, rectangles or even free form shapes, while another expert stated that the circular shape is satisfactory. Lastly, one of the experts had additional ideas how the system could be useful including traffic analysis, and traffic planning. He considered it helpful for manufacturers of electric vehicles and charging infrastructure operators.

For future work, one of the experts mentioned interest in integrating weather data, points of interest, and event data to improve the semantic understanding.

I rate the usefulness of the following aspects as high:	fully agree				disagree
Representation of the filter results (histograms, etc.)	1	2			
Data representation on a map	3				
Temporal navigation and selection in timeline	2	1			
Representation of geographic selection (lenses)	3				
Representation of lens groups (and/or)	1	2			

◄ **Figure 5.17** — All experts agreed on the usefulness of features offered by TrajectoryLenses, especially regarding spatial representation and filtering.

5.4.3 User Study

To evaluate design decisions, the perceptual and interaction process was investigated during a hands-on study. For this purpose, a minimal setup of TrajectoryLenses was created, which consists of the geographic map view, a filter lens, as well as the statistic panel on the right side (see Figure 5.2). Eye movements and mouse interactions of five visualization experts (four male, one female) were recorded and analyzed in two three-minutes sessions of system usage. The first task was to use the destination lens to explore the scooter trips. In the second session, participants had to use the origin lens with the same data. The logging framework tracked all interactions and visual outputs. In addition, the users were asked to think aloud whilst using the lenses. One of the study conductors took notes of what participants were saying, and the screen was recorded while participants interacted with the system. Before each study, session participants had to sign a consent form, and give general information about gender, age, grade, job, but also visual impairments. In addition, all participants passed the Snellen visual acuity and Ishihara color test. After recording mouse movement and eye movements with a Tobi T60XL device, different preprocessing steps were applied to cluster the gaze and mouse points into fixations [Salvucci and Goldberg, 2000].

Interaction Log Results

Overlaying the stimulus with the eye and mouse movements revealed that the main focus of the fixations is clearly on the map, especially on dense areas. This indicates that the heatmap draws attention as planned. Almost all participants also investigated the group panel (see Figure 5.2, c). Some participants looked at it intensely, others just had a quick look at it, and even less inspected the panel with the mouse. The distance chart shown in Figure 5.18 allows to analyze correlations between eye and mouse movements. The *zigzag* patterns indicate that the distance between eye and mouse fixations increases when the eyes move to the next interesting region on the stimulus and decreases when the user moves the mouse cursor to the region accordingly. The small multiples of the spatial view (top) show a case where the distance decreases (2 to 6). A region was spotted, and the mouse is moved there, accordingly. The temporal chart below encodes distances over time, while the red bar links to the highlighted trajectory in the spatial view. The temporal view also allows to compare individual participants. This way, it was found that participants had the highest distances at the beginning of the study. Furthermore, three participants had high peaks of large distances towards the end of the study. The remaining two participants had several higher peaks throughout the study, indicating that they followed their eyes more often with the mouse.

Think Aloud Results

The videos and the think aloud protocol gave feedback on the usability and usage of TrajectoryLenses. During the free exploration task all participants investigated the

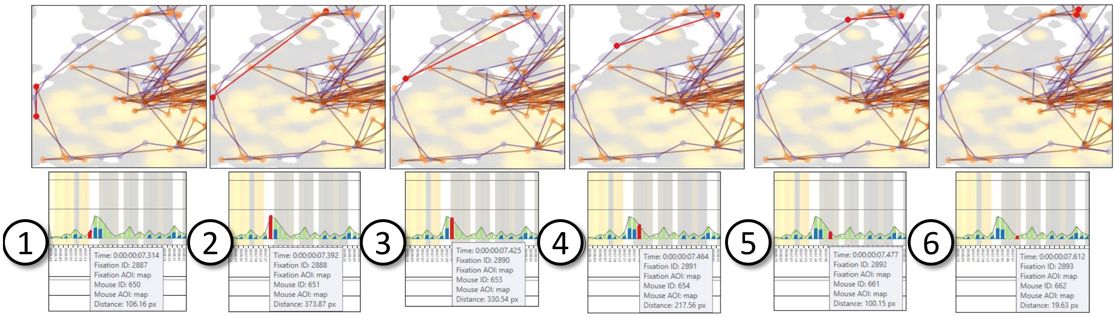


Figure 5.18 — A sequence of eye and mouse movement. Top) The red line marks the distance of eye and mouse pointer position at a certain timestamp of the analysis. One can see that the eyes explore the surrounding region first, making the distance larger, before the mouse pointer follows. Bottom) The plots (1–6) show the distances over time. The background is colored according to the data density within the participant's current eye focus. At dense areas, the distance is smaller.

inner city of Stuttgart and found out that this is the densest area for trips to start and end. Two participants investigated the main station area, probably because they were attracted by the high density in this area shown in the heatmap. One stated that he expected this area to be highly frequented because of the centrality of the location. The nearby cluster at the car manufacturing company (see Case Study in Section 5.4.1) was also investigated. One participant used the context provided by the map to reveal that 18 scooters were parked at the Zoo. Other places investigated comprise a beer garden, the bike repair shop (see Section 5.4.1), but also city areas like Feuerbach and Wangen where the participants increased the size of the lens, accordingly. One participant found it unsatisfactory to not getting more details about places, which highlights the importance of context. In the minimal setting for the user study the semantic enrichment results were not included. Indeed, all participants were busy finding out where people went to, although it was a free exploration task without any given analysis goals. One participant pointed out that the origin heatmap should be more sensitive, showing also density variations in less dense areas. This issue could be addressed by slightly adjusting the color mapping (Section 3.5). Besides investigating origins and destinations, trip duration and speed were of interest. Looking at the statistics, one participant found out that the speed limit is around 17 km/h on average and another mentioned that it is very useful to have the statistics nearby.

Regarding the usability, there was mixed feedback. Some participants found the lenses very interactive and intuitive. Others needed a little more time to get familiar and looked at the description handed out or asked repetitive questions about how to drop the lens, make it smaller or larger, etc. However, after a minute, everybody felt comfortable, which lead to the assumption that more features such as lens combinations and context integration would be easy to handle after a short training period. This has to be evaluated in the future.

Chapter 6

Pattern Detection and Verification

During the exploration process analysts select relevant subsets of the data, collect insights, and build a schema in their mind of what has happened. Regarding movement data, this could include daily routines of travel, meetings, as well as outlying behavior. At a certain point, however, free data exploration reaches its limitations. Firstly, datasets might be very large and complex, so that patterns in the data could be easily overseen. With only a few cases explored, hypotheses might be vague, and do not allow for a precise explanation. Thus, for the analysis of large and vast data, automated methods for pattern detection are needed. Secondly, exploratory means have limited capabilities to verify and quantify the support for hypotheses that already exist or are gained during exploration. Yu [1994] argues that *"exploratory data analysis performs the function of a model builder for confirmatory analysis"*. Confirmatory means become especially important when critical decisions need to be made, as those are often connected to public safety and enormous amounts of money. Decision makers are responsible for their suggestions and actions. They need to ensure that there are no patterns overseen, need to know where and when a pattern exactly appears, and that resulting hypotheses are plausible, i.e., there is enough support in the data. In many situations detailed reports are needed to present analysis results to supervisors in charge. Accordingly, this chapter contributes means for (1) pattern detection to assist data exploration, (2) the extraction and representation of patterns for storytelling and reporting, and (3) refinement and pattern-based querying to verify existing hypotheses. Figure 6.1 highlights the steps in the analysis pipeline (see Section 2.4.3).

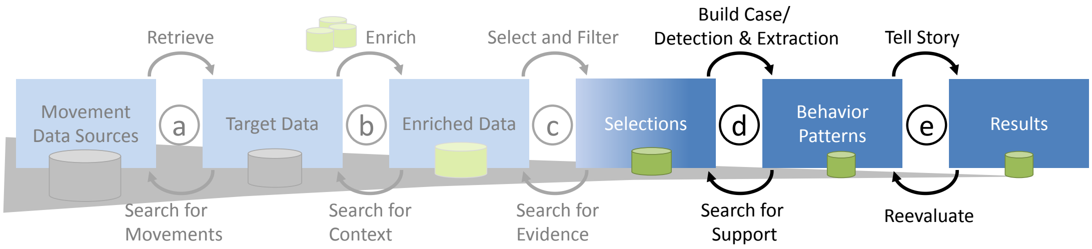


Figure 6.1 — The integration of automated approaches allows to guide the analyst to frequent patterns. For further extraction, verification, and refinement, a pattern-based visual query language is proposed. It gives a clear representation of findings and allows to communicate results in a compact way.

The chapter begins with related work in sequential pattern mining and querying (Section 6.1). Thereafter, novel visual analytics means to detect routines and anomalies in event sequences and their interactive steering, and visualization are presented (Section 6.2). Found patterns can be extracted and visualized with VESPa, a Visual language for Event Sequence Patterns that introduces a more compact and informative representation (Section 6.3). VESPa also enables confirmatory analysis. Existing hypotheses can be formulated as possible patterns, and executed to query for support in the data (Section 6.4). Exploration and verification can be coupled in an iterative loop, allowing for a stepwise refinement of results. Lastly, the approach is evaluated in two case studies and two user studies (Section 6.6).

Parts of this chapter have been previously published in:

R. Krüger, H. Bosch, S. Koch, C. Müller, G. Reina, D. Thom, and T. Ertl. HIVE-BEAT - A highly interactive visualization environment for broad-scale exploratory analysis and tracing: VAST 2012 Mini Challenge 1 award: Honorable mention for comprehensive visualization suite. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 277–278. IEEE Computer Society, 2012a

M. Steptoe, R. Krüger, Y. Zhang, X. Liang, R. Garcia, S. Kadambi, W. Luo, T. Ertl, and R. Maciejewski. VAST Challenge 2015: Grand Challenge - award for outstanding comprehensive submission. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 119–120. IEEE Computer Society, 2015

R. Krüger, D. Herr, F. Haag, and T. Ertl. Inspector Gadget: Integrating data preprocessing and orchestration in the visual analysis loop. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 7–12. The Eurographics Association, 2015b continued...

R. Krüger, F. Heimerl, Q. Han, K. Kurzhals, S. Koch, and T. Ertl. Visual analysis of visitor behavior for indoor event management. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1148–1157. IEEE Computer Society, 2015a

F. Haag, R. Krüger, and T. Ertl. VESPa: A pattern-based visual query language for event sequences. In *11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (IVAPP)*, volume 7, pages 48–59, 2016

T. Tremel. Advancement of a query visualization for event sequences. Supervisor: R. Krüger, F. Haag, T. Ertl. Bachelor Thesis. Universität Stuttgart, 2016

6.1 Background

The background given for this chapter is two-fold. At first, a brief review on sequential pattern analysis is given. Secondly, related work in the area of visual languages for pattern-based querying with a focus on event sequences is discussed. Also relevant to the approaches presented in the following is work on modeling and visualization of event sequences. This has been discussed in Chapter 4.

6.1.1 Sequential Pattern Analysis

In Section 4.2, movement data was enriched with contextual data, allowing to focus on the order of events, e.g., visited places and meetings, over time rather than on the movement (the trajectory) itself. Such an event sequence $s := e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow, ..., \rightarrow e_n$ consists of ordered nominal items (events) e_i $(1 \leq i \leq n)$. To understand routines and abnormal behavior therein mining algorithms from the field of unsupervised learning can be applied. They can be categorized into finding associative patterns among features and patterns among data instances.

The expression *sequential pattern mining* was first introduced by Agrawal and Srikant [1995]. They proposed a family of so called *Apriori* algorithms to find frequent sub-sequences in sequences of transactions, e.g., from shopping behavior of customers. It extends association-rule mining [Agrawal et al., 1993] by taking the order of transactions into account. Later, partly the same authors published advances named *GSP* [Srikant and Agrawal, 1996], *Spade* [Zaki, 2001], and *PrefixSpan* [Han et al., 2001]. In the context of this thesis the term sequential pattern mining is used in a broader sense and additionally comprises instance-based methods like sorting, clustering, and the detection of outliers in sequential nominal data. While sorting allows to find similar sequences to a selected sequence, clustering groups sequences based on similarity. Outlier detection allows to find sequences that show significant differences compared to others, e.g., do not fit in these groups.

Clustering can be categorized into partition-based and non-partition-based approaches. Former approaches cluster each data instance, while latter approaches declare instances as noise, if they do not fulfill the requirements. Another perspective categorizes clustering into connectivity-based, distribution-based, and centroid-based methods. Connectivity-based approaches connect instances based on a threshold and the amount of neighbors. *DBSCAN* [Ester et al., 1996] is a connectivity-based approach and was used in Section 4.3.2 to aggregate stop points. Distribution-based approaches assume an underlying distribution function to calculate cluster memberships. Centroid-based approaches assign instances to a cluster based on the distance to its center while updating the center iteratively. Besides the clustering method, the outcome highly depends on the chosen distance metric. Cohen et al. [2003] survey string metrics for matching names and records. A widely used metric to sort and cluster nominal sequential data is the *Levenshtein* distance [Levenshtein, 1966]. It assigns a cost function for any edit operation needed to transform a string into another. A drawback of edit-based distances is their high complexity. Especially in visual analytics a trade-off between effectiveness and efficiency has to be considered. If distances have to be computed in real-time one may consider a metric that allows for faster computation.

There exist several visual analytics approaches to detect patterns in nominal sequential data. In 2016, the workshop co-located with the IEEE VIS conference, named *The Event Event: Temporal and Sequential Event Analysis*, was held to address the area more precisely. Common tasks identified are to (a) simplify sequences, (b) find subsequences, (c) understand longitudinal changes, and (d) compare sequences [Ruddle et al., 2016]. Several works present concrete approaches to utilize mining approaches with visual analytics. Liu et al. [2016] propose a three-stage framework consisting of pattern mining (what types of patterns to extract), pattern pruning (what patterns to show), and interactive visualization to control algorithms and present patterns. Perer and Wang [2014] and Perer et al. [2015] apply *SPAM* [Ayres et al., 2002], an algorithm to mine frequent sequences. Other approaches make use of sequence clustering to reduce the visual complexity [Wongsuphasawat et al., 2012]. Gotz [2016] goes one step further by arguing that many of the state of the art approaches are not sufficient for real world event sequence data, as they do not allow for soft pattern discovery. Similarly, one objective presented by Perer et al. [2016] shifted the discussion of visual event sequence mining from theory to real-world event sequences. They argue that on the one hand the vast amounts of data require the integration of automated approaches, but the veracity of real-world event logs, such as the diversity of event types, noisiness, the lack of strict orderlines bring many challenges that afford manual exploration.

Other prominent examples on visual event sequence analysis are *LifeLine* [Plaisant et al., 1996] and *EventFlow* [Monroe et al., 2013a]. They do not apply unsupervised learning but allow for user-driven sequence simplifications based on event merging, category merging, and insertion.

6.1.2 Pattern-based Querying

A variety of approaches for visually expressing generic filter queries have been proposed in the past [Shneiderman, 1994, Seifert, 2011, Soylu et al., 2013, Russell et al., 2008]. There are specialized query visualizations for spatial data, some using symbolic representations of geographical relationships [Morris et al., 2004, Wu et al., 2013]. Others work directly on concrete maps [Kumar et al., 2013]. The symbolic representations lend themselves to use cases that work with logical locations.

For selectively finding particular pieces of information in spatio-temporal data, the spatial and the temporal dimensions are split up in some concepts [Boyandin et al., 2011, Certo et al., 2013, Krüger et al., 2013b]. Also, a graphical notation that expresses exact relative relationships between locations or areas and timespans has been proposed [Bonhomme et al., 1999]. Another visual approach displays the query restrictions in a space-time-cube [D'Ulizia et al., 2012]. Some of these concepts can be further abstracted to replace the geographical association with restrictions based on logical locations [Certo et al., 2013], and time-related works provide further ideas on how to visually represent certain features required for temporal queries [Monroe et al., 2013b].

So far, only few approaches specifically deal with the visual specification of queries for event sequences, which could be used for finding particular event patterns and validating hypotheses. Those that do so only visualize the event sequence without emphasizing the query structure and its restrictions in a visual way other than aligning sequential states horizontally or vertically [Plaisant et al., 1996, Gotz and Stavropoulos, 2014, Wongsuphasawat et al., 2012]. Only few concepts allow the analyst to visually specify pattern-based queries for event sequences as timelines [Fails et al., 2006] or as nodes with temporal constraints [Dionisio and Cárdenas, 1996], rather than just displaying results in a visual way [Plaisant et al., 1996]. Still, relations between distinct objects are usually not supported [Zgraggen et al., 2015]. If they are, relations are expressed by matching colors in otherwise visually disjoint elements [Jin and Szekely, 2009].

As opposed to the aforementioned work, VESPa, the approach presented here, focuses on the graphical representation of event sequence patterns of multiple persons. Instead of just recognizing similar sequences [Plaisant et al., 1996], VESPa also copes with actual overlaps that allow for an interaction of persons. Moreover, one can visually express the overlaps of event sequences by actual connections rather than just by matching properties such as node colors [Jin and Szekely, 2009] to make the connection explicit.

6.2 Detection of Routines and Anomalies

Finding patterns such as routines and anomalies in vast amounts of movement data is a challenging task. An overview can give analysts a general orientation of the data distribution. However, patterns are often hidden in the data, are superimposed by others, e.g., covert deeper in a hierarchical structure, or comprise multiple dimensions. They can thus be easily overseen, even in visualizations. At this point automatic approaches can help to highlight such structures. Computational methods are faster and, if the task is clearly specified, often more thorough. They can assist domain experts to fulfill their analysis tasks. The field of unsupervised machine learning has yielded many different approaches to find patterns in sequential data (see Section 6.1.1). From this catalog some suitable approaches for specific problems were picked and are presented in the following. At first, sorting approaches based on distance metric for sequential data are applied. Then, clustering and outlier detection is used to highlight human behavior routines and anomalies.

6.2.1 Similarity Sorting

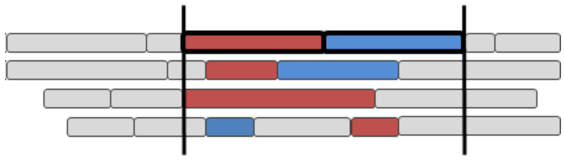
Sorting the data based on different distance measures introduces a systematic that helps to browse the data, detect similarities, and differences. The linear event sequence view (introduced in Section 4.4.2) thus provides different sorting options in a dropdown menu that allow analysts to either sort the data based on global features such as device ids and category counts or based on event order. A suitable distance metric for sequence comparison based on event order is the Levenshtein distance [Damerau, 1964, Levenshtein, 1966]. It defines the distance between two sequences as the number of insert, delete, and replace operations necessary to transform one sequence into the other. The Levenshtein distance between two strings a, b is given by $\mathrm{l}_{a,b}(|a|, |b|)$ (see Equation 6.1).

$$\mathrm{l}_{a,b}(i,j) = \begin{cases} \text{if } (\min(i,j) == 0): \ \max(i,j), \\ \text{else: } \min \begin{cases} \mathrm{l}_{a,b}(i-1,j)+1, \\ \mathrm{l}_{a,b}(i,j-1)+1, \\ \mathrm{l}_{a,b}(i-1,j-1) \\ \qquad +[a_i \neq b_j] \end{cases} \end{cases} \tag{6.1}$$

The indicator function $[a_i \neq b_j]$ equals zero if $a_i == b_j$, and 1 otherwise. In the following, the metric is applied to compute a matrix containing similarities for each pair of sequences. Because the Levenshtein distance is computationally expensive, this matrix is calculated offline for later use in the visualization.

When sequences get long, the Levenshtein distance may not be the best choice. Firstly, analysts are often not interested in sequences that are similar as a whole.

Secondly, event sequences over, e.g., a year are mostly very heterogeneous, and even if there are homogeneous parts, other parts outrule these local similarities. In such cases subsequences can be compared. This, however, cannot be pre-computed, as for a single sequence of n events there already exist $\frac{n(n+1)}{2}$ possibilities to select a subsequence of consecutive events for the sorting. For an online calculation, however, the Levenshtein distance still has a high complexity of $O(m \cdot n)$. To overcome this issue, the cosine distance is used, which can be computed interactively. Firstly, feature vectors of the selected subsequences are calculated based on the percental temporal category distribution of the selected event category within the selected time range. For the subsequence depicted in Figure 6.2 (marked in back) the feature



▶ **Figure 6.2** — Sketch of an ad-hoc sorting based on a selected subsequence (top). Subsequences within the timeframe are sorted based on their distance in descending order.

vector would be $v = (0.5, 0.5, 0.0)^T$ referring to the percental distribution of the colors red, blue, and gray. The similarity between the selected sequence, a reference vector r, and any other sequence in vectorform v in an n-dimensional vector space with n categories is defined as follows:

$$sim_{r,v} = \frac{\sum_{i=1}^{n}(v_i \cdot r_i)}{\sqrt{\sum_{j=1}^{n}(v_i)^2} \cdot \sqrt{\sum_{j=1}^{n}(r_1)^2}} \tag{6.2}$$

All vectors to be compared are normalized to length 1. This simplifies the formula to $\sum_{i=1}^{n}(v_i \cdot r_i)$. The analyst can for example brush a repetitive *home-work-home* pattern. This will immediately sort all sequences based on subsequence similarity for the same timeframe. The result can be seen in Figure 6.3.

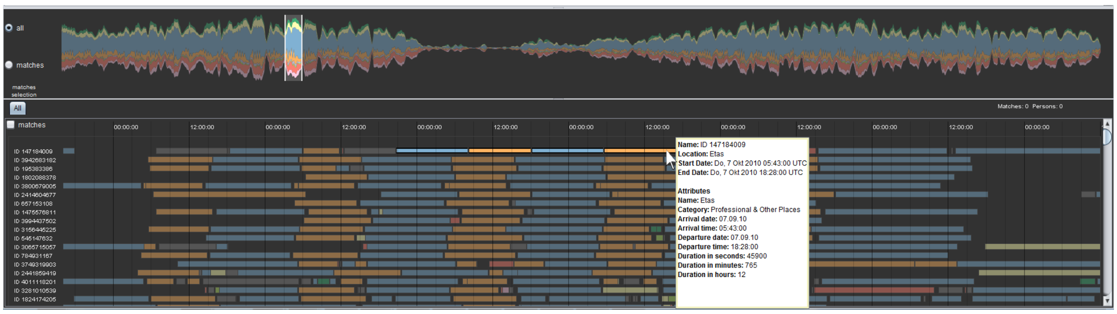


Figure 6.3 — Subsequence sorting on the enriched electric scooter data. The highlighted subsequence is selected by the analyst. The system sorts the sequences of all moving objects according to their subsequence similarity. In this case, the search by example is set to a *home-work-home-work* pattern.

6.2.2 Clustering

Sorting based on event-order affords that an object of interest has already been found as a reference. Clustering, in contrast, allows to detect groups of similar behavior without any preknowledge of the containing patterns. The aggregation can aid exploration, especially when datasets are large and it is not possible to manually browse the whole set. There exist different clustering techniques that can be combined with different distance measures and configured with different parameter settings, resulting in different outcome. Regarding the input data for the clustering, one can either use the sequence values over time directly or derive features that describe the event sequence more abstractly. The first option can help to extract groups of similar temporal behavior, e.g., groups that travel together while the second option may reveal groups of similar interest.

Partition-based Clustering

k-means clustering [Lloyd, 1982] is a centroid-based partitioning clustering method (see Section 6.1.1) and can be used to group event sequences into a predefined number of groups with similar features. k-means clustering has different advantages. Firstly, the computation is fast. It has a complexity of $O(n \cdot k \cdot i)$ with n being the number of instances, k the number of clusters and i the number of iterations the cluster centers are updated. Secondly, it is easy to understand and to steer by the analyst, because the outcome strongly depends on the parameter k, defining the number of clusters. Disadvantages are that the approach is not capable of finding clusters of longish shapes, and a suitable number of clusters that matches the natural data distribution is often not known in advance. Hence, the analyst either needs pre-knowledge or can adjust k in an iterative manner. Especially when data is very large and fast visual feedback is the dominant criteria, k-means clustering can be a good choice. In the visualization in Figure 6.4 a k-means clustering was computed. 192 million computer log sequences with three varying attributes over time are clustered [Krüger et al., 2012a]. Besides k the analyst can choose different measures, e.g., Euclidean distance. She can decide which data attributes to take into account and choose whether to sort by reference or group the sequences using clustering.

◄ **Figure 6.4** — The matrix shows sequences of 192 million computer log events, clustered with k-means. The cluster in the viewport shows machines with high connection counts at a certain timespan and healthy policy status.

Hierarchical Clustering

For the analysis of large datasets a hierarchical clustering can be of advantage. It constitutes a good basis for an interactive analysis of the dataset, because it allows analysts to derive an overview first and to stepwise increase the level of detail. The hierarchy is computed either by recursively grouping instances (bottom up) or by separating the data step by step based on different criteria (top down). A disadvantage is that hierarchical approaches have higher complexity and take much more time for computation and handling. To overcome this issue, the clustering structure can be precomputed, stored, and loaded in a short amount of time.

To find sequences of similar visitor behavior in the data recorded at the re:publica conference 2013 (see Section 2.5.3 for details) *spectral clustering* was applied in hierarchical manner. Spectral clustering [Ng et al., 2001] has its name from the spectrum (eigenvalues) of the similarity matrix, that are used to partition the data. One advantage of spectral clustering in comparison with other methods is that it is not biased towards any shape of the clusters in the instance space and is therefore able to produce more natural clusters. At first, the sequences of attended talks are processed and split into chunks of similar timespan. Next, a similarity matrix is computed based on the Levenshtein metric (see Section 6.2.1). The matrix is then transformed into a nearest neighbor graph. In this graph, each node represents one visitor pattern and is connected to its closest neighbors. Each edge connecting two neighboring event sequences is weighted by their respective similarity. The clustering algorithm then takes this neighbor graph as input and partitions it according to a specific criterion. A generally good objective for these graph partitions is that the within-cluster similarity should be high and the inter-cluster similarity should be low. This is achieved by recursively identifying good cuts within the neighbor graph, i.e., partitions of the graph that cut edges whose sum of edge weights is as low as possible. A second objective is that the resulting clusters should be equal in size in order to obtain a well-balanced tree.

Different criteria and heuristics exist that define a good cut according to these and other objectives. One popular criterion that incorporates balanced partitions is the normalized Cheeger cut [Cheeger, 1970]. It is defined as in Equation 6.3.

$$NCC(C,\overline{C}) = \frac{cut(C,\overline{C})}{min(vol(C), vol(\overline{C}))} \tag{6.3}$$

C represents one of the two tentative partitions of the subset at a given iteration of the algorithm, while $\overline{C}$ represents its counterpart. The function $cut(C,\overline{C})$, as defined in Equation 6.4 (left) represents the sum of the weights of the edges that would have to be disconnected in order to separate C and $\overline{C}$. To counterbalance this objective of the smallest cut, $vol(C)$ and $vol(\overline{C})$ are introduced. They represent the sum of the weight of all edges within C and $\overline{C}$, respectively (Eq. 6.4, right).

$$cut(C, \overline{C}) = \sum_{i \in C, j \in \overline{C}} w_{ij} \qquad\qquad vol(C) = \frac{1}{2} \sum_{i,j \in C} w_{ij} \tag{6.4}$$

Finding an optimal solution for the best-balanced graph cuts according to Equation 6.3, and thus recursively partitioning all visitor patterns, is a problem that can be solved efficiently using standard eigenvalue methods [Von Luxburg, 2007]. This problem can be reduced to the identification of the second smallest eigenvalue and corresponding eigenvector in the *Laplacian* matrix that is constructed based on the neighborhood similarity matrix.

$$L = D - W \tag{6.5}$$

This graph Laplacian L is defined according to Equation 6.5, where W represents the neighborhood weight matrix and D is a diagonal matrix containing the degrees of all nodes in the neighborhood graph on its diagonal [Von Luxburg, 2007]. The eigenvector of the Laplacian matrix L that corresponds to the second smallest eigenvalue can be interpreted as the optimal partitioning of the neighborhood graph into two subgraphs. The implementation provided by Bühler and Hein [2009] was used to efficiently create a hierarchical cluster tree of the visitor sequence dataset with 200 leaf clusters. Figure 6.5 shows the interactive visualization to explore hierarchically clustered sequences of attended talks for re:publica conference visitors (see Section 2.5.3 for details on the scenario). The user can interactively navigate through the cluster hierarchy using a range slider. The size of each cluster is indicated by the blue bars (1). Each cluster is visualized by a representative. This sequence shows the most frequent events on average. The analyst can click on a representative to unfold the visualization and inspect single instances of the cluster (2). By hovering over a sequence a tooltip displays further information, such as the category and topic of the attended talk, room number, and speaker (3).

Similarly, to find groups that visit amusement park attractions together in the 2015 VAST Challenge data (see Section 2.5.4 for more information) a hierarchical clustering with Levenshtein distance was used. By contrast to the re:publica scenario,

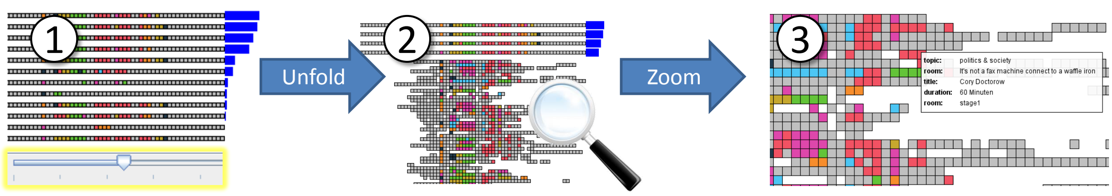


Figure 6.5 — Interactive hierarchical spectral clustering of the re:publica 2013 data. (1) The slider allows to navigate through the cluster hierarchy. (2) Clusters can be viewed by a representative, or in detail. (3) The analyst can zoom in and out and inspect single events in detail.

smaller groups of 2 to 50 people were of interest. Thus, an agglomerative (bottom-up) approach was applied (see Figure 6.6). Hovering over the event sequences links to the map view and highlights trajectories and visited locations.

6.2.3 Outlier Detection

Outlier detection can highlight abnormal mobility behavior patterns. Depending on the analysis goals different approaches can be applied. A simple approach is to show the instances with the highest distance from the mean value. More local approaches find outliers that are not member of any cluster (noise). Breunig et al. [2000] goes one step further by taking into account the local density in the neighborhood of the outlier. Similar to clustering, there is a trade-off between complexity and effectiveness. To keep response times as low as needed for an interactive analysis of large datasets, one can either apply algorithms with cheaper computation or use pre-computation which results in a loss of flexibility. Similarly, to find the most outlying sequences in the VAST Challenge 2015 (see Section 2.5.4 for a detailed description) three distance matrices were computed based on check-ins by attraction name, by category, and by region. In this case, the outlier detection just needs to sum up distances for each matrix row and return n rows with the highest sum in descending order. n can be interactively adjusted. Figure 6.6 shows the visual integration of the hierarchical agglomerative clustering (left) and outlier detection (right). It allows to find visitor groups that travel the park together and to find visitors behaving abnormally and may be connected to the crime (see Section 2.5.4).

Figure 6.6 — Interactive clustering and outlier detection in attraction visit sequences of amusement park visitors. Left: (1) define the hierarchy cutting level; (2) inspect cluster representatives; (3) inspect cluster members and geographic aspects; Right: (1) define outlier threshold; (2) largest outliers; (3) inspect details.

6.3 Pattern-based Visual Language

Detected patterns can be highlighted, e.g., by coloring, framing, resizing, etc. For further analysis and especially for presentation, however, it is of advantage to extract them and show them in a clearer, compact, and explicit way. In addition, analysts may look for variations, or have other specific questions to be answered, verified, and quantified. VESPa is a Visual language for Event Sequence Patterns. It can be used to express, to refine, confirm, or disproof hypotheses and was originally developed to solve the VAST Challenge 2014 (see Section 2.5.4). The challenge sets up a fictitious law enforcement task to search for routines and culprits with outlying behavior. In the following, VESPa 2.0 is presented, which includes lessons learned from a first user evaluation and enhances the original language (VESPa 1.0 [Haag et al., 2016]) with further possibilities and a more intuitive and interactive usage.

6.3.1 Language Components

In Section 4.3.1 an event was defined as a spatio-temporal happening that at least involves one person. The linear event visualization (see Section 4.4) represents event sequences per objects as individual rows. However, events can also involve multiple moving objects, which means that multiple objects share an event. This can be formulated as an acyclic directed graph $DAG = (V, E)$ with $E \subseteq V \times V$ where each node has a corresponding outgoing edge for each incoming edge (except the first and last node of the sequence), as depicted in Figure 6.7. The visual representation of VESPa fulfills a temporal order, having nodes and directed edges sorted from left to right.

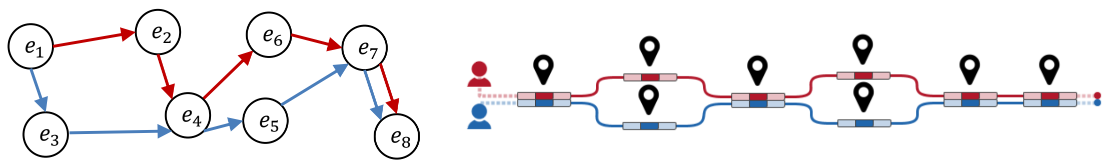
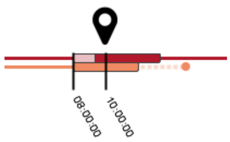


Figure 6.7 — An illustration of an event sequence as directed acyclic graph (left). The two event sequences have some events in common. (right) Same event sequence in an ordered sequential representation of VESPa with two actors as placeholders.

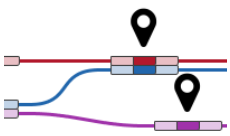
VESPa offers a set of components to define such an event sequence pattern, comprising actors, paths (representing the sequence), events, and timelines. In addition, the visual language allows to further specify events with attributes, to add loops and alternative branches, etc. In the following, all components are shortly described.

Actor
Every sequence, be it visited POIs or other events, is owned by an actor. This is depicted in the beginning of a sequence. In VESPa an actor is a placeholder for any possible moving person (or multiple). VESPa represents each path belonging to an object exactly with one actor. Respectively, paths with different actors cannot be represented by the same person.

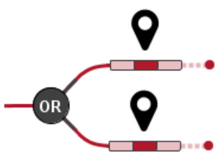
Event
Events are the core element of the query language. An event describes a spatio-temporal happening. The minimal set of attributes are place and time information. In addition, there can be further attributes. An event is shown as a placemarker with a timeline if no further restrictions are set.



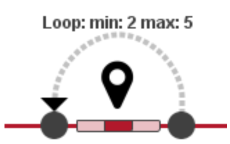
Timeline
A timeline is part of an event. Its color adapts to the actor involved. If the event is shared with multiple actors, it has multiple timelines. Conversely, if an event has only one incoming edge and timeline, this means that only the connected actor is involved in the event. The timeline can express restrictions, including start, end, and durations.



Edge
An edge connects actors with events and defines the temporal order of them. It also clearly connects the event sequence with the actors by color. The length and direction of edges do not have a meaning (can be freely layouted). Edges between two events do not necessarily result in a location change, as a new event can also differ in time only.



Alternatives
Alternatives define a logical exclusive OR condition. This means sequences that match the defined pattern include persons that follow either the one or the other defined path.



Loop
A loop allows multiple iterations of the included patterns. As depicted in the image, a condition with a minimal and maximal number of iterations can be defined.

VESPa 2.0 components can be orchestrated to define simple to complex patterns. A minimal pattern at least involves four components: an actor, a path, an event, and an event timeline (see Figure 6.8, 1). It is defined in temporal order, from the leftmost event, being the first event in the pattern, to the rightmost, the last event. Additional components allow to define loops and branches (XOR) (Figure 6.8, 2).

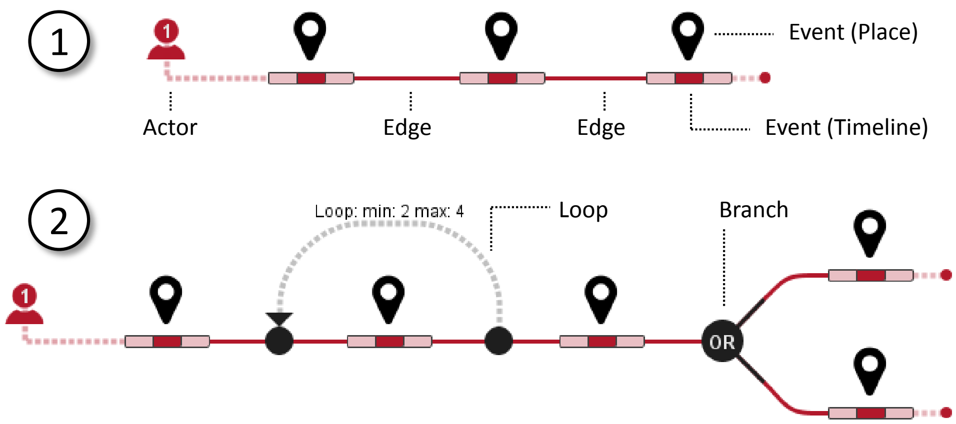


Figure 6.8 — Composition of VESPa components. (1) A minimal set allows to express simple patterns consisting of at least an actor, edges, and events. (2) Further components allow to formulate patterns with loops and alternatives.

6.3.2 Restrictions

Components can have restrictions on space, time, and other data dependent attributes. This can be used to express, e.g., the following situation:

A person starts at home, goes to work, and meets with another person for lunch at 3 pm, then goes shopping with two other persons that come from the university in Vaihingen, while the second person goes to another place not further specified.

This textual notation describes a relatively simple sequence of events, yet the text is already quite long and possibly confusing. Situations of the same or a higher complexity can be of interest, for instance, in geo-marketing, when investigating consecutive shopping behavior, or to understand city dynamics for urban planning. Similarly, in law enforcement investigators try to find out whether people meet after having visited certain locations that are connected to a crime. VESPa allows for a visual notation of such patterns which was evaluated as intuitive and more easy to define and understand (see Evaluation, Section 6.6). In Figure 6.9, the aforementioned textual notation is expressed using VESPa. Each actor is a placeholder for persons in the data, e.g., scooter drivers, and an event represents a visit of a POI of one or multiple actors at a certain time or timespan. Different constraints are added, concerning event category and name as well as time.

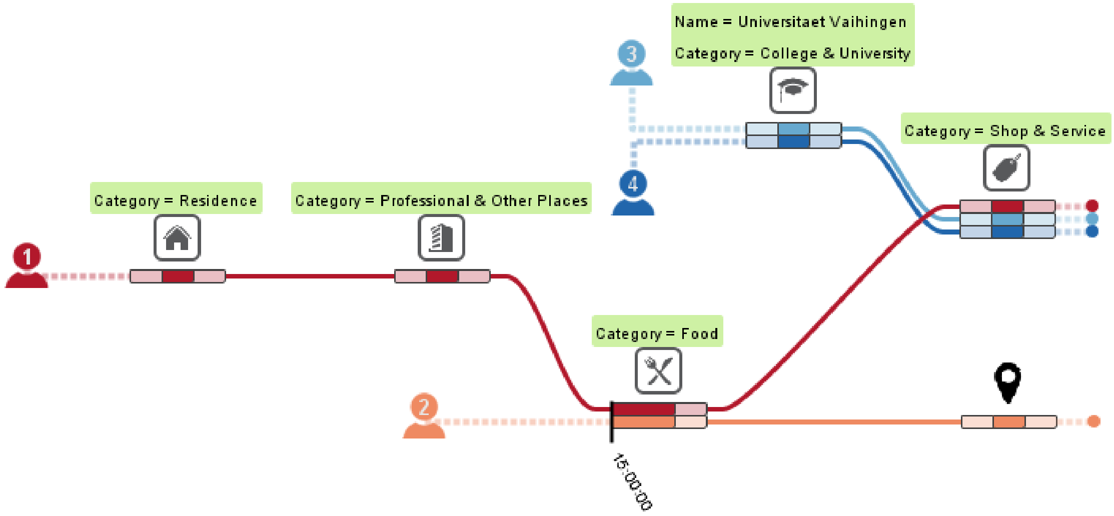


Figure 6.9 — The visual notation expresses the same situation as the textual description at the beginning of Section 6.3.2. It comprises different restrictions regarding event name, category, and starting time.

Attribute Restrictions

An event can have different attributes that further specify it. Regarding the scooter dataset, an event contains a POI that has a name, for example, *Universitaet Vaihingen* and a category such as *College and University*. Each attribute further restricts the matching set of situations described by the pattern. For the visual representation of attributes, different symbols can be added, making the expression more intuitive to read. In Figure 6.9, the Foursquare icon set is used, while in Section 6.6.1 the icons map to locations at an indoor conference. In addition, attributes can also be added to actors, for example characterizing the actor to be employed at a certain company or living in a certain region.

Temporal Restriction

An event is specified as a spatio-temporal happening. Hence, the temporal dimension is important. A user study of the first version of VESPa [Haag et al., 2016] revealed different misinterpreted aspects. Many participants took an event mainly as a POI and did not understand its temporal dependencies through the visualization. Hence, VESPa 2.0 addresses these issues by (1) explicitly showing a small timeline for each event and (2) allowing a sequential order of events only. Another feature of the timeline is that analysts can restrict the starting and ending time as well as the duration of the involvement of actors to an event in a compact way. In addition, the timeline allows to handle uncertainties of pre-post relations, as described by Allen [1983], by defining an earliest and latest start and end time (see Figure 6.10, 1), similar to *PlanningLines* by Aigner et al. [2005]. This way,

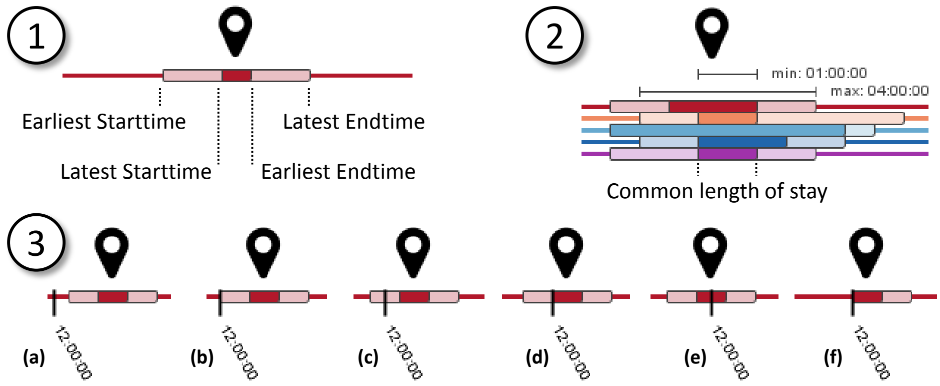
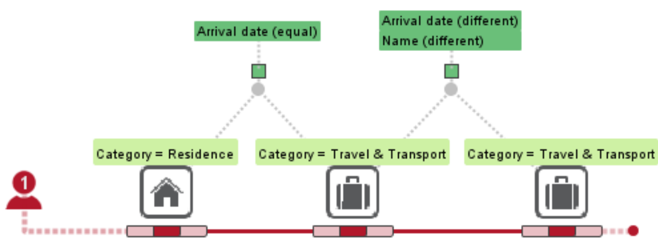


Figure 6.10 — The timeline component of an event allows to define (1) earliest and latest start and end times; (2) minimal and maximal common length of stay; (3) different concrete restrictions on start and end times: a) after 12 pm; (b) 12 pm at earliest; (c) before/ after 12 pm; (d) 12 pm at latest; (e) before 12 pm; (f) 12 pm.

temporal relations can be defined. For example, they can express if persons represented by an actor arrive or leave earlier than others as well as the minimal and maximal timespan in common for multiple involved actors (Figure 6.10, 2). Lastly, different temporal restrictions can be added to different positions on the timeline (Figure 6.10, a-f). Besides daytime these restrictions can also define concrete dates.

Comparisons

Sometimes it is necessary to explicitly define if events share or differ in different attributes. Therefore, VESPa offers comparison edges that can be added to any node. Figure 6.11 gives an example. Here, the first and second event share the value for the restriction *arrival date* but the second and third event differ, not only in the arrival data but also in the value for attribute *name*. For example, a traveler could have left the home location and arrives at a train station at the same day. She then takes the train to another city and arrives at its train station on the next day. Again, the available comparison possibilities depend on the restrictions of the dataset, resulting from data recording and semantic enrichment.



◀ **Figure 6.11** — Comparison edges allow to define shared attributes (here arrival date) and differences (again arrival date and name) between events.

6.4 Pattern-based Querying

Besides expressing a situation with VESPa, it can be used to verify if, where, when, and how often a pattern appears in the data. Patterns can be interactively composed and executed as query to search a database for matching event sequences. The analyst can manually compose a query based on background knowledge and assumptions (see Section 6.4.1). Subsequently, results are evaluated (Section 6.4.2) and shown in the linear event visualization (see Section 6.4.3) that is coupled with VESPa through brushing and linking. The combination also allows to extract interesting patterns from the visualization by brushing actions. The close combination of exploratory analysis (generate hypotheses) and confirmatory analysis (verify if the hypotheses are supported by the data) realizes an iterative substantiation of findings (Section 6.5). In the end, VESPa allows for a compact description of a patterns about routines and anomalies and can also be used for reporting.

6.4.1 Query Construction

To create a query the analyst can position the cursor at any point in the VESPa panel and spawn new query elements such as actors, events, and restrictions. Usually, an analyst would start by defining an actor (see Figure 6.12,1) and maybe add restrictions that further define this role, such as employment type, gender, age, etc. depending on the dataset. Events can then be freely created at any position in the panel (2). To define an event sequence the analyst can click on the starting point next to the actor and drag it along the events (3). A magnet function checks whether the cursor position containing the dragged node exceeds a certain threshold and if so pins the Bezier curve to the corresponding event. This way sequences can be created with one drag & drop operation (see Figure 6.12), which was assessed to be intuitive during user evaluation (see Section 6.6). The analyst can then define further restrictions for actors and events using different context menus (4).

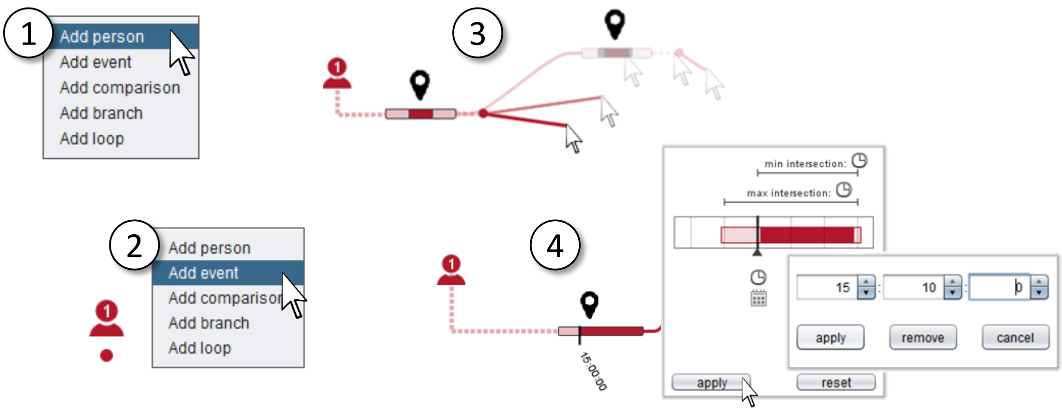
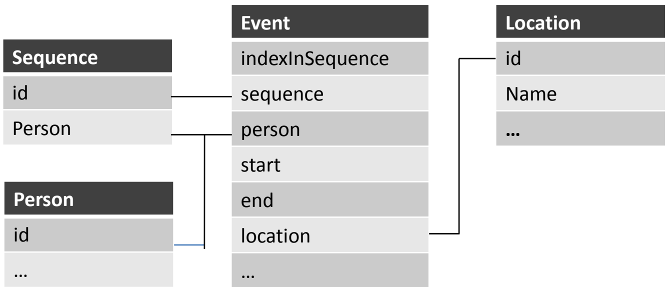


Figure 6.12 — VESPa queries can be defined step by step using simplistic and intuitive interaction approaches.

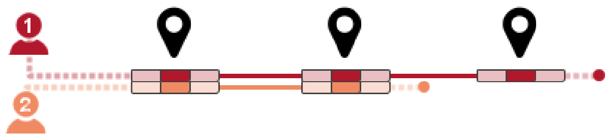
6.4.2 Query Interpretation

VESPa is developed as a standalone component that brings its own data storage, an SQL lite database. In the future, a faster backend, such as a graph database, could improve query response times, but the research so far concentrated on the visual language rather than on fast computation. The database schema is depicted in Figure 6.13. The normalization enables flexible querying, allowing any combination of event sequences and persons possible but also comes with higher computational costs. To integrate VESPa in an analytics system, the movement data model (see Figure 4.3, for example) has to match VESPa's database schema. Several interfaces have to be implemented, comprising the entities persons, events, and sequences connecting events. Thereafter, the data can be stored in the SQL database.



◄ **Figure 6.13** — The database schema normalizes the event sequence with respect to person, sequence, event, location, and others.

While the analyst adds and manipulates a VESPa query, all relevant visual query components are stored in a *MapQuery* object (Java-based data model of the query). When the query is executed, the VESPa interpreter transforms these constructs into an SQL statement, executes it, and filters the system's data model according to the query results. In Figure 6.14, a rather simple query graph is shown. The corresponding SQL statement is depicted in Listing 6.1. The example again highlights the simplicity of the visual representation. In comparison the SQL statement is not intuitive and affords expert knowledge in SQL, that cannot be expected for experts in movement data analysis, be it from the domain of law enforcement, consumer analysis, city planning, or any other task. In the *FROM* part of the query, the tables for persons and events are selected. All events that involve a certain actor have to be part of the same sequence, and their temporal indices have to match the visually defined order (lines 7-12, 14-15, 27-28, and 35). In addition, the query ensures that for every event that involves two persons, the persons were at the same place with a certain temporal overlap (making it the same event, line 16-17, and 29-30). Lastly, line 18-25 and 31-33 define that different events cannot temporally overlap.



◄ **Figure 6.14** — Example of a query graph that is translated into an SQL statement (see Listing 6.1).

Listing 6.1 — SQL statement equivalent to the graphical representation of the event sequence pattern in Figure 6.14, based on the schema in Figure 6.13.

```
SS
SELECT  A1.id, E1_1.sequence AS "Sequence1", E1_1.indexInSequence,
 E2_1.indexInSequence, E3_1.indexInSequence, A2.id,
 E1_2.sequence AS "Sequence2", E1_2.indexInSequence, E2_2.indexInSequence
FROM    Person A1, Person A2, Event E1_1,
 Event E1_2, Event E2_1, Event E2_2, Event E3_1
WHERE   (A1.id <> A2.id)
        AND (E1_1.sequence = E2_1.sequence)
        AND (E2_1.indexInSequence - E1_1.indexInSequence BETWEEN 0 AND 1)
        AND (E1_2.sequence = E2_2.sequence)
        AND (E2_2.indexInSequence - E1_2.indexInSequence BETWEEN 0 AND 1)
        AND (E2_1.sequence = E3_1.sequence)
        AND (E3_1.indexInSequence - E2_1.indexInSequence BETWEEN 0 AND 1)

        AND (E1_1.person = E1.id)
        AND (E1_2.person = E2.id)
        AND (E1_1.location = E1_2.location)
        AND (E1_1.start < E1_2.end AND E1_2.start < E1_1.end)
        AND (NOT ((E1_1.location = E2_1.location) AND (E1_1.start < S2_1.end)
        AND (E2_1.start < E1_1.end) AND (E1_1.start < E2_2.end)
        AND (E2_2.start < E1_1.end) AND (E1_2.start < E2_1.end)
        AND (E2_1.start < E1_2.end) AND (E1_2.start < E2_2.end)
        AND (E2_2.start < E1_2.end) ))
        AND (NOT ((E1_1.location = E3_1.location) AND (E1_1.start < E3_1.end)
        AND (E3_1.start < E1_1.end)
        AND (E1_2.start < E3_1.end) AND (E3_1.start < E1_2.end) ))

        AND (E2_1.person = A1.id)
        AND (E2_2.person = A2.id)
        AND (E2_1.location = E2_2.location)
        AND (E2_1.start < E2_2.end AND E2_2.start < E2_1.end)
        AND (NOT ((E2_1.location = E3_1.location) AND (E2_1.start < E3_1.end)
        AND (E3_1.start < E2_1.end) AND (E2_2.start < E3_1.end)
        AND (E3_1.start < E2_2.end) ))

        AND (E3_1.person = A1.id)
```

Being already bulky for this simple example, the SQL queries get more complex when more actors and events, and additional language components such as loops, branches, and restrictions are added. When alternatives are included, the query is decomposed at first. Subsequently, an SQL statement for each branch is executed separately, and the result sets are joined afterwards. The interpretation of loops follows a similar approach. For each loop iteration an SQL statement is created, e.g., a query containing a loop of three iterations results in three separate SQL statements, one containing the event one time, one containing the event twice, and one containing the event three times. Both cases are shown in Figure 6.15.

In Section 6.3.2, different types of restrictions were introduced. Attribute-based restrictions can simply be added to the SQL statement after the *WHERE* clause. This is implemented with different classes that match the internal data types of the programming language to SQL types such as string, long, and date. For comparative restrictions an equation can be added in a similar way, e.g., defining

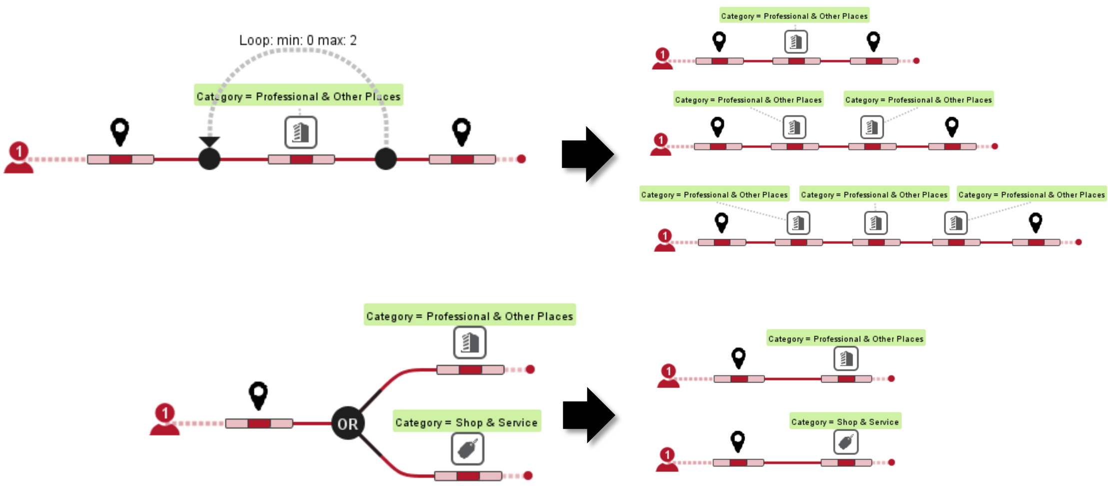


Figure 6.15 — Loops (top) and branches (bottom) are decomposed before they are translated into SQL queries.

that two attribute values have to be equal or different. For temporal restrictions different cases have to be considered. Firstly, there is a pairwise relation defining a before-and-after restriction. Secondly, there are time and date restrictions defining exact, earliest, and latest starting and stop points. Thirdly, there are intersection restrictions defining the minimal and maximal time of an event that actors have in common. In all cases, the execution of the SQL query returns a result table containing all cases that match the query pattern. Event sequences that do not match are excluded. Others might contain the query pattern even multiple times. The data table then needs to be transformed back to the event sequence data model. In case of loops and branches the result sets are joined.

6.4.3 Result Exploration

The query results can be explored in space and time. The map view shows all trajectories that are involved in the event sequence pattern defined in VESPa. It further shows POIs, or any other event attributes with spatial features. Similarly, the temporal distribution of the result set can be explored in, e.g., the linear event sequence view (see Section 4.4.2). This is depicted in Figure 6.16. Different options surrounding the timeline allow to set the focus on all sequences, to highlight matching subsequences but show the rest as context, or show matchings only. If different actors are involved in the VESPa query, the analyst can browse the results by match for each actor's pattern. The streamgraph gives overview of the distribution. It can be set to show all data, to show only categories included in the results, to show only event sequences included in the results, etc.

In Figure 6.16, the electric scooter data is queried for situations where a person goes to an event of the category *Professional & Other Places*, which is probably

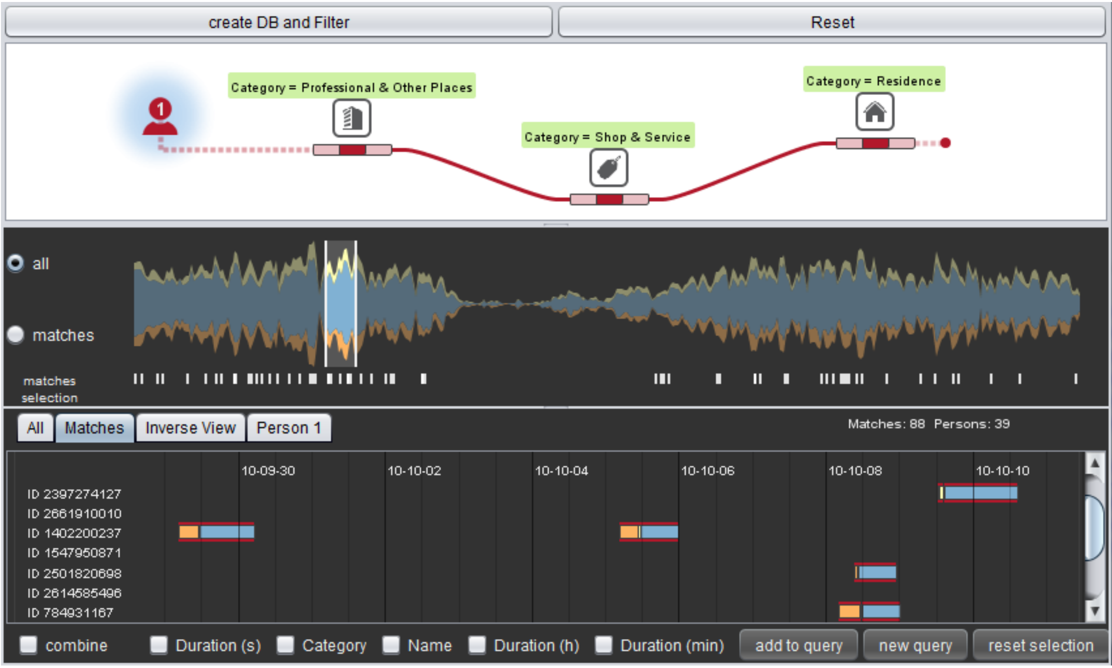


Figure 6.16 — VESPa combined with the linear event sequence visualization (Section 4.4.2) where results can be explored. In this case, the streamgraph shows categories included in the query and marks matching subsequences in white. The detailed view below shows the found subsequences, but can also be configured to show the rest as context, or the inverse result containing all parts that don't match, e.g., a common routine.

for work, then goes to a shopping place, before she or he arrives at home. The streamgraph marks the temporal areas where matches are found, allowing the analyst to navigate to these areas for detailed exploration. In this case, a timespan is selected in the streamgraph, and the detailed panel below shows the matches. It reveals different electric scooter users with the given patterns. The timespan at home is the longest (blue), the working period the second longest(orange). The shopping period is very short, about 10-20 minutes on average, only visible through further zooming.

6.5 Closing the Loop

When interesting subsequences are found in the data whilst exploring, it can be cumbersome to explicitly express them with VESPa, either for presentation and reporting or to query for more occurrences. The event sequence view thus allows to mark these parts directly in the visualization and deduce the VESPa pattern automatically. Attribute restrictions included can be selected in advance. Which restrictions are taken into account can be selected in the visual interface. By clicking on the *new query* button (see, e.g., Figure 6.17), a VESPa pattern is generated that contains the event sequence and restrictions marked by the analyst. It is also possible to mark sequences from different actors that may overlap, i.e., have events in common. In this case, an additional actor is added to the VESPa pattern, allowing to express meetings or differences. Thereafter, the pattern can be refined by adding more restrictions and events to the query or generalize it by removing restrictions, accordingly. Finally, the pattern can be executed as a query. This may reveal other parts in the data that also follow the defined routine. An example of this capability is given in the law enforcement case study in Section 6.6.1.

The combination of exploration and querying allows for an iterative generation, refinement, and verification of hypotheses. Figure 6.18 summarizes the iterative loop. The analyst might start by exploring event sequences and may gain a hypothesis or has already a hypothesis in mind from previous knowledge (step a). She then either extracts a found pattern or expresses another hypothesis in VESPa notation (b). The query can be refined (c) and executed (d). The results are shown in the event sequence visualization, leading to (a) again. The analyst can cycle this loop in an iterative manner. When she has finished this process, the generated pattern serves as a compact representation of the findings. Due to the simplified representation, VESPa can be used to communicate result (e), even to non domain experts without deeper knowledge of spatio-temporal visualization and analysis. This was evaluated in two consecutive user studies that are discussed in the next section.

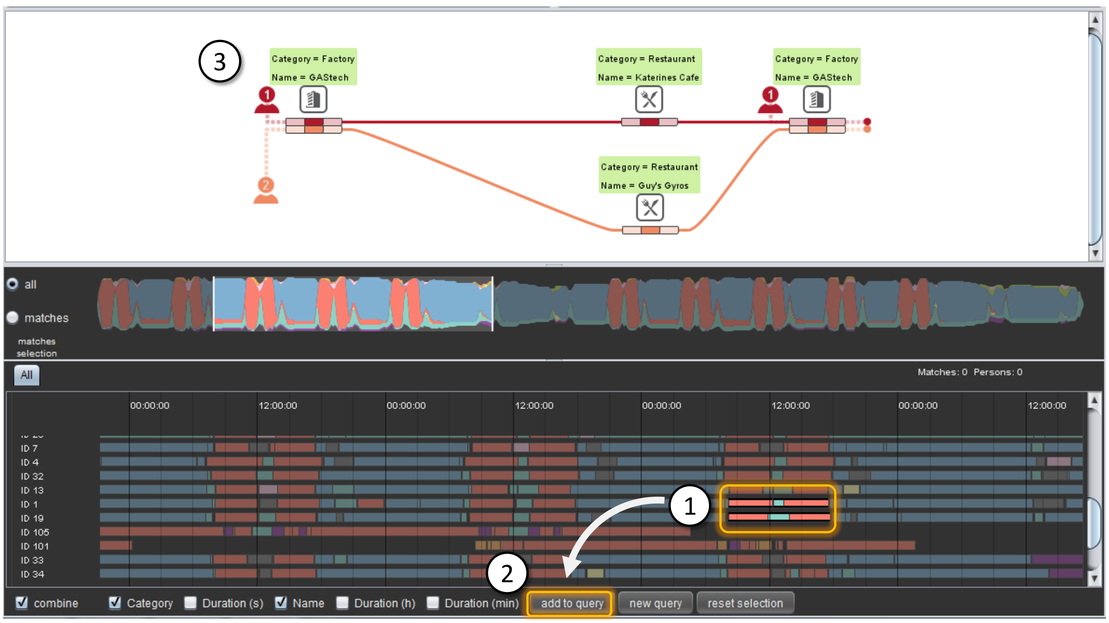


Figure 6.17 — Patterns can be extracted from the event sequence visualization to get a more compact and independent representation through VESPa. (1) mark pattern; (2) set attributes and extract; (3) pattern is represented through VESPa.

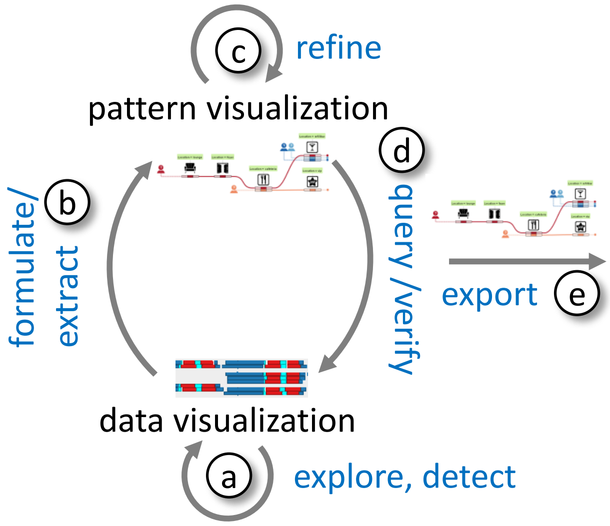


Figure 6.18 — Integration of VESPa in the analysis loop. (a) Data is explored, patterns detected; (b) Patterns are extracted or explicitly formulated; (c) Patterns are refined; (d) The query is executed, leading to (a) (result) exploration; (e) Final patterns are extracted in reports.

6.6 Evaluation

The proposed approach was evaluated based on case and user studies. The case studies demonstrate the applicability of the interactive visual query approach on real world datasets and problems as well as on simulated data with ground truth. The user studies tested the comprehensiveness and the learnability to construct visual queries with VESPa.

6.6.1 Case Studies

VESPa provides a small programming interface, making it easily usable on different datasets. While the examples of the scooter data were presented to introduce the approach, the first study focuses on an indoor dataset containing movement patterns and presentation attendances of visitors from a large conference. The second study deals with a simulated law enforcement dataset containing movement patterns of suspects. It was created for the VAST Challenge 2014 and has the advantage that ground truth data is available. This allows to evaluate approaches by comparing the number and accuracy of patterns found. In both datasets raw movement patterns were enriched with semantic information containing points of interest, information about persons, event schedules, and billing data.

Indoor Visitor Behavior

The re:publica is a large conference on web 2.0, social media, and the digital society. In 2013, free wifi was offered to visitors that agreed to positional tracking on a *per room* resolution, i.e., by connections to access points. Using the time schedule of the conference and additional information about speakers from the web, the visitors' movements (temporal sequence of access point connections) were semantically enriched. This resulted in sequences of stays and talk attendances, which can be used to analyze routines and outlying behavior to understand peoples' preferences and topic interests, to reveal bottlenecks, and to improve the organization of future events. More information about the dataset is given in Section 2.5.3.

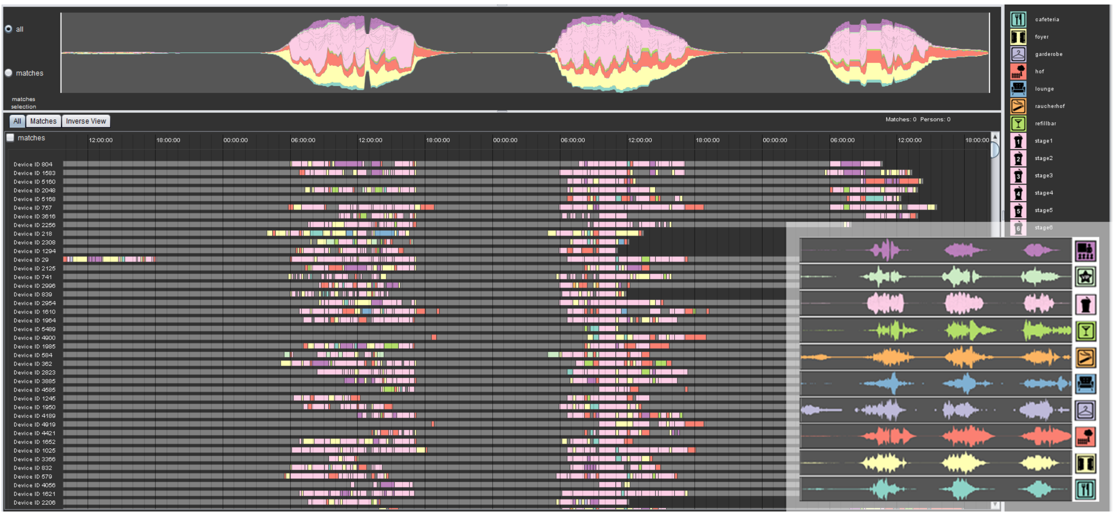
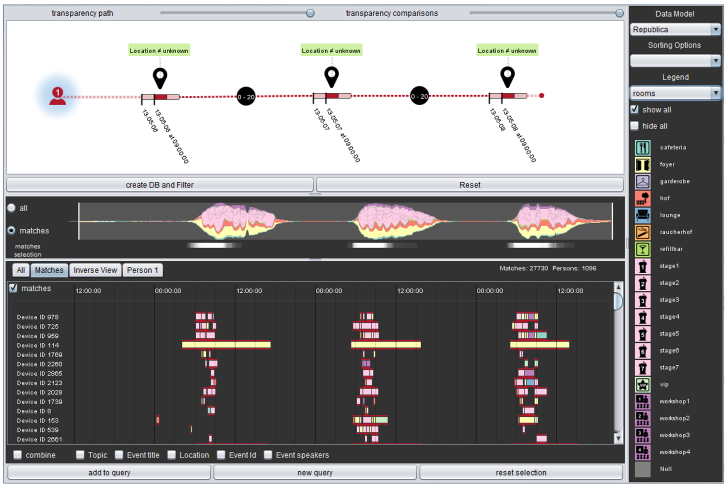


Figure 6.19 — re:publica visitor sequences, colored by room type. The streamgraph gives an overview. Users can also select single categories (as overlayed in the lower right), sort sequences, and cluster them.

After the data is loaded, the analyst starts to explore the visited rooms and topics over time in the streamgraph visualization (see Figure 6.19). A continuous morning increase and evening decrease of visitors on the three days is clearly visible. When sequences are colored by room, one can see that the stages 1-7 cover most of the visitors together with the foyer that has several peaks during breaks. Similarly, coloring based on talk categories reveals different peaks during the day, strongly correlating with the schedule. By activating and deactivating different categories in the streamgraph, the analyst detects activity in the conference area before it was officially opened. These people are most likely workers who prepared the rooms. Other sequences show constant overnight activity. The analyst selects one by example and sorts others based on similar length. This reveals that all these devices are distributed in different rooms, most likely workstations. The analyst chooses to investigate single rooms independently. The VIP room shows activity before talks start, probably because of speaker preparations. By contrast, rooms for leisure activities such as the bar, the smoking area, and the lounge are still busy after the last talks. As expected, the restaurant is busy during lunch time.



▶ **Figure 6.20** — VESPa is utilized to query for visitors that frequently attend the conference at latest 9 AM on all three days. 1,096 of 6,700 visitors match the query pattern.

To learn about visitors with similar interests and behavior the analyst makes use of the hierarchical clustering to subdivide visitors based on similar temporal talk attendances. By slightly increasing the cluster cutting threshold, the analyst examines that some guests frequently visited sessions, while other guests accessed WiFi only in the lounge areas. Many devices logged in only for one day of the conference. One cluster contains devices of guests that attended the conference on the last day only. They mainly visited talks on politics and society, for a long time. Overall, only small groups of about 2-4 people have identical talk attendances. Many visitors share smaller subsequences of, e.g., four to five talks and then split up.

The exploration led to the impression that not many visitors come to the conference on time. To verify if the hypothesis aligns with the data the analyst creates a visual query with VESPa (see Figure 6.20) and searches for visitors that arrive at 9 AM at latest on all three days. Only 1,096 of 6,700 visitors match this pattern.

Law Enforcement

As described in Section 4.5.2, the trajectory data of the VAST Challenge 2014 was enriched with context such as areas of interest and transactions (see Figure 4.24). Different views were presented to explore the semantics in space and time. By investigating the employees' mobility behavior one can see that repetitive patterns clearly stand out (see Figure 6.21). Here, the blue coloring represents trip destinations in the residence area, and the red coloring represents trips to offices, mainly to GAStech, the employees' company. The streamgraph shows that this routine repeats during the weeks with breaks on the weekends (Figure 6.21, 1). On their way to work and during lunch break the employees visit cafes and restaurants.

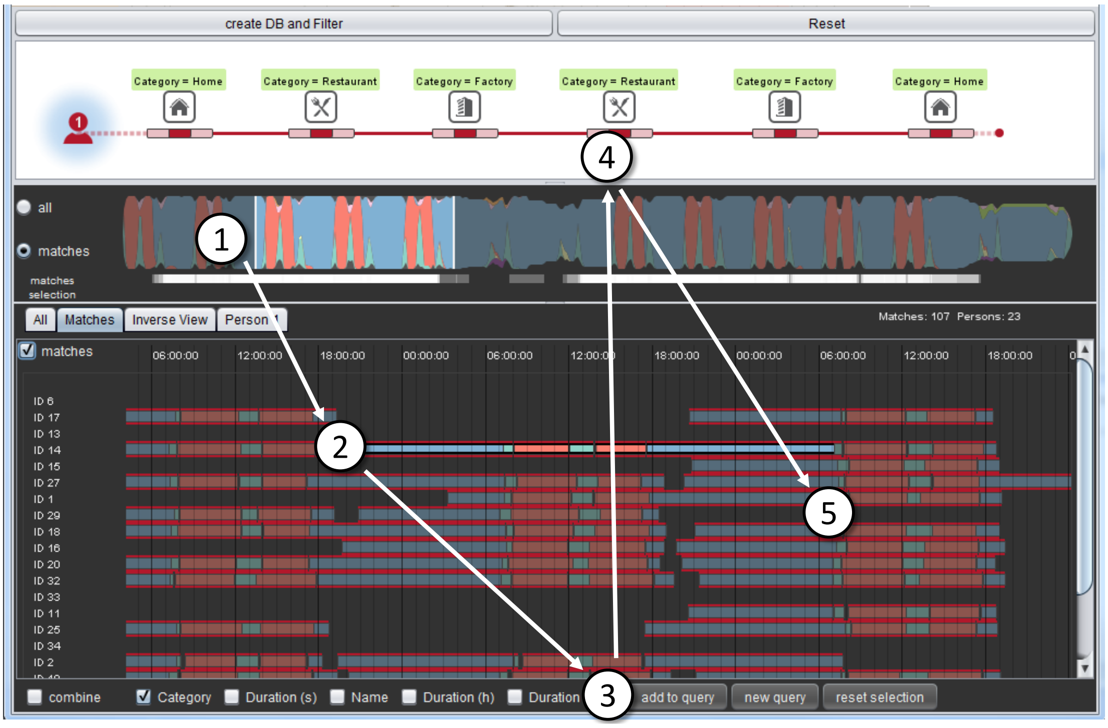
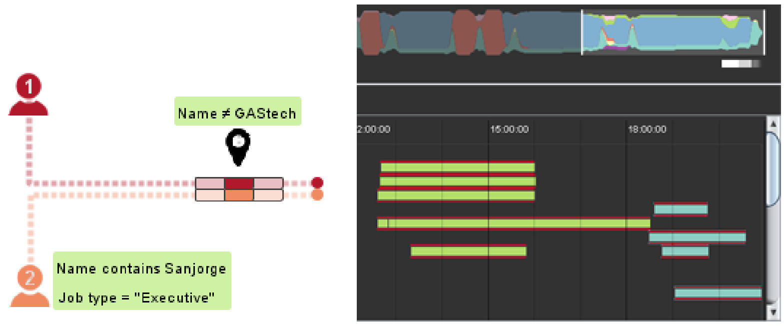


Figure 6.21 — The analyst drills down from overview (1) to detail and marks a pattern in the sequence timeline (2). The pattern is extracted (3) for a compact representation with VESPa (4). The execution of the pattern-based query reveals other matches (5).

For a closer investigation, the analyst brushes (Figure 6.21, 2) one of these sequences (home-restaurant-office-restaurant-office-home) as an example and uses the capability of VESPa to extract (3) and represent the pattern in a compact form (4). By executing the pattern, the query returns all matches in the dataset and highlights them, accordingly (5). Subsequently, the analyst can choose to see them in context with all other sequences, see matches only, or see inverted results, e.g., all outlying behavior that does not match the pattern. This reveals that some of the working routines show small differences and are thus not matched.

To investigate further, the analyst can now either loosen or strengthen the restrictions. VESPa's filter restrictions are adjustable to the use case. In this case, they included *name*, *category*, *employment title* and *type*, *time* and *date* of *arrival* and *departure*, and the amount of *payments*, if any. To increase the recall of the query she can remove restrictions and allow different branches that, for example, let the query return sequences of homecomers as well as sequences of employees that prefer to first meet at various bars. By contrast, additional restrictions can define the working day to start at 7 AM at the earliest and to end no later than 6:30 PM with a lunch break in between, according to the usual business habits of GAStech. This reveals that some people work outside of business hours, do not take lunch breaks, or do not appear at work at all, but go shopping. Other cases not matching the routines comprise trucks that commute between airport, harbor, and factories. In addition, suspicious expenses with credit and loyalty cards and especially a large payment of 10,000 dollars at a car supply shop stand out.

In her further analysis, the analyst wants to investigate the behavior of the GAStech CEO (chief executive officer) whose behavior also significantly differs from that of other employees. To investigate his activities, she constructs a filter pattern as follows: First, she creates an event, which she does, however, not specify any further in space and time. She then sets the actor for this event to be the CEO. Lastly, she adds another actor that is not restricted any further, either. By doing so, she



▶ **Figure 6.22** — The query for employees that meet with the CEO named Sanjorge reveals that he plays golf with five of them before having dinner (green and turquoise bars).

formulates a query that filters for events during which the executive meets any other person at any place but GAStech (see Figure 6.22). As a result, she gets three main events. Besides official meetings at the company GAstech on Friday, the CEO also meets three executives at the golf course. Furthermore, he is meeting other employees for lunch on Saturday and for dinner on Sunday. While this result does not fully confirm any involvement in a crime, it allows her to further substantiate her suspicion: The CEO is involved in unofficial activities, which he prefers to discuss outside of work. These discussions might be related to the kidnapping.

6.6.2 User Studies

VESPa was developed and evaluated in two iterations. Unfortunately, it was not possible to acquire domain experts in the fields of mobility analysis (see Section 2.5). For the first version of VESPa, five visualization experts were asked to solve comprehensive and composition tasks [Haag et al., 2016]. This information was used to improve VESPa to the stage that is presented in this chapter. A second user study followed a similar procedure and is presented in the following. This time, the participants were computer science students that had a solid background in information technology but were neither experts on human mobility nor in visualization and visual languages, allowing to especially evaluate how intuitive the approach is.

The study was carried out in a small office on a desktop computer with a 24 inch monitor with a resolution of 1920x1200 pixels. The study consisted of four parts: an introduction to the topic and to the query language VESPa, questions about the comprehension of VESPa patterns (interpretation), composition tasks (formulate patterns with VESPa), and a final discussion revealing the correct and incorrect answers. The introduction was given in written words to assure that all participants have the same preconditions. They were allowed to use the document during the whole study. It contained a short introduction into the topic and components of the language as well as different examples about how components can be orchestrated.

Comprehension Study

The comprehension task was carried out as think aloud study. Different VESPa patterns were given, and the participants should describe their meaning by word of mouth. The first two tasks of the study were designed to familiarize the participants with the possibilities to formulate VESPa queries in words. Therefore, two possible answers to choose from were already given. For the following tasks, however, they had to describe the meaning in their own words. To ensure comparability a list contained significant characteristics of the queries that enabled the rating of the answers given by the participants after the study. If they did not give satisfying answers they were explicitly asked about the aspects they left out. The comprehensive study concluded with a set of multiple-choice questions that focused on details of the query language.

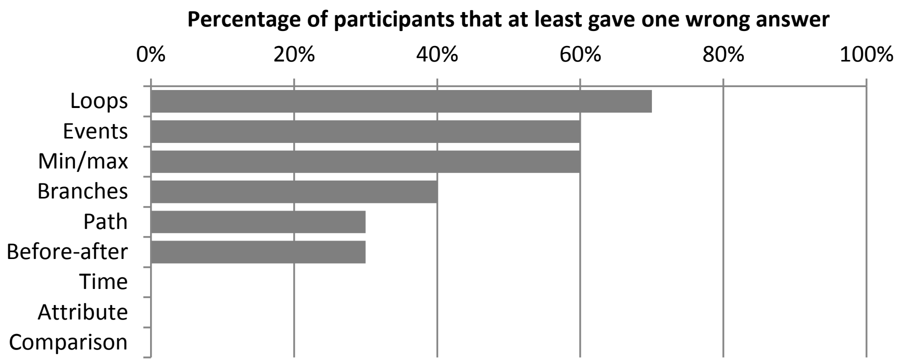


Figure 6.23 — Evaluation results: Comprehension problems of VESPa components. Regarding loops, 70 percent of the participants had comprehension problems. Timestamps, attributes, and comparisons were always understood correctly.

For the first two tasks nearly all participants gave the correct answers (one chose the wrong textual description). However, they were designed to introduce and were not considered for the evaluation. The rest was evaluated according to the predefined list of important aspects. The results are shown in Figure 6.23 and Figure 6.24. Many participants gave wrong answers regarding the number of loop iterations.

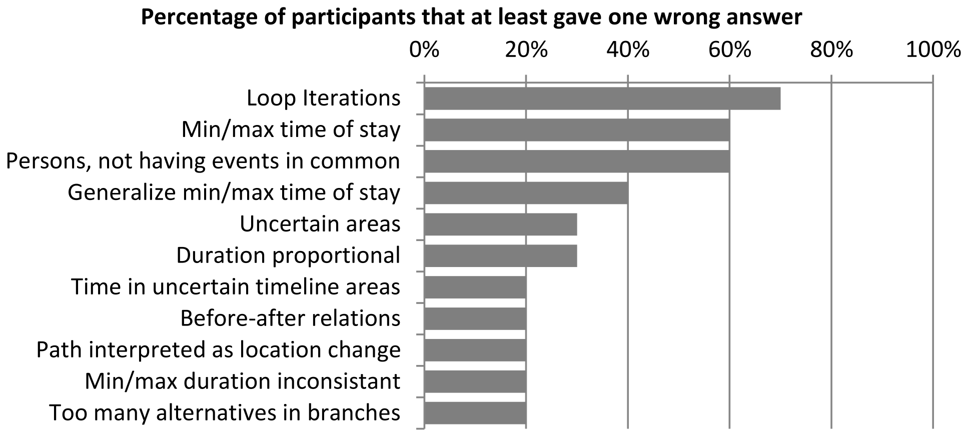


Figure 6.24 — Evaluation results: Most frequent wrong answers. The percentages shows the number of participants that at least gave one wrong answer.

They assumed that the loop includes the first iteration. Instead, the min variable was designed to start with the first additional iteration. During the final discussion, however, most participants agreed on the intended version and admitted that it would fit the visualization better. Besides these errors the participants understood the loop component very well, for example that they can be arranged freely around any event. Regarding the branches, two participants did not understand their purpose and correspondingly answered all parts containing branches wrong. In addition, two other participants assumed that the branches referred to a split in the movement of the actors, meaning they would both follow different sequence. The timeline was designed to highlight the temporal component of an event, based on the results of the first user study (VESPa 1.0 [Haag et al., 2016]), where participants often did not understand that an event is defined by both, space and time. Hence, even a second meeting at the same place with the same actors would lead to a different event. The timeline component significantly improved the understanding, but the minimal and maximal length of stay was still not always interpreted correctly. The same applies for the individual time of stays and for the before-after relations. The interpretation also took the participants much time, depending on the size and number of restrictions. For smaller expressions the participants needed less time and gave more correct answers.

Similar to the interpretation of VESPa 1.0 many participants had issues to understand the fact that persons without events in common cannot meet. Only one participant falsely added meaning to the length of the edges, while all others understood that no value was encoded here. Regarding the edges, another problem occurred. Two participants interpreted them as a movement phase and argued that the persons physically left a location before they came back to the same POI (the example included two events at the same location, directly after each other).

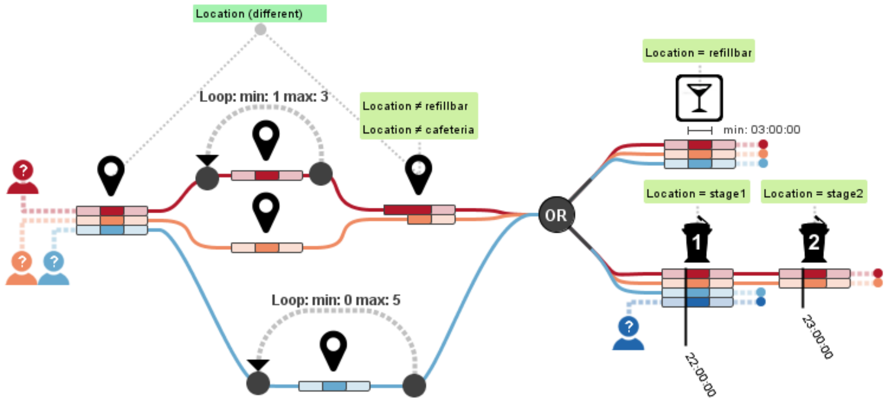


Figure 6.25 — The most complex query of the comprehensive tasks contained four actors, additional components such as loops and alternatives, and restrictions regarding locations and time.

Overall, the visual query language was understood well. Significant errors were only made by a few participants, mostly on more complex patterns, as shown in Figure 6.25. Also, they occurred in very special cases with many restrictions. Simpler examples were understood well. For the comparison, attribute restrictions, and temporal details (e.g., hour of day and date) no mistakes occurred. There was also no case where participants had no clue about a pattern. They always had an answer that was at least close to the intended happenings.

Composition Study

The purpose of the second part of the study was to evaluate how the participants can express textual notation using VESPa. To create expressions with VESPa the participants were introduced to the implemented query creation components and had time to get familiar with them by interactively experimenting. There was also an introductory exercise, similar to the one in the comprehension task. The graph was given on a printed paper and the participants had to reconstruct the graph with the graphical user interface. While there were no answers given regarding the semantic meaning and syntax of the language, the participants were allowed to ask questions regarding the interactions, e.g., how they can visually create components and choose restrictions from context menus. The following tasks consisted of an easy, an intermediate, and a difficult question.

Overall, all participants composed satisfying visual queries for the given tasks, even for difficult ones (see Figure 6.26). For all tasks the central statements were understood and composed correctly. Results of the first tasks were mostly right and even for the second task the participants made only small mistakes concern-

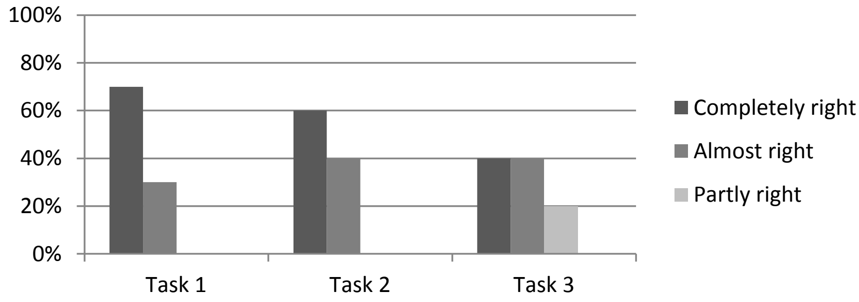


Figure 6.26 — Relative distribution of correct, mostly, and partly correct composition tasks. No one answered any task completely wrong.

ing the timeline restrictions. For the third task there were more problems. Four participants gave a correct answer, two made small mistakes, and two gave wrong answers regarding multiple components. Most mistakes were made for the loop iterations (80% of the participants made a mistake at some point), similar to the comprehensive task, and again for the temporal restrictions (60%).

Final Questionnaire and Conclusion

Finally, the participants got a questionnaire containing general questions about the understandability and usefulness of VESPa. One question asked the participants whether they would apply VESPa, if they were in a situation to query similar event sequences, which most participants confirmed. For example, they argued that it is very intuitive and not difficult to use, and that it would not be possible to describe such graph structures with words only. Some argued that it all depends on the length of the event sequence pattern and that they would probably use it, if the queries were complex. Only one participant stated to use arbitrary methods and would prefer to type in words. 9 of 10 participants stated that they understood the concept and also understood their mistakes.

Compared to VESPa 1.0 the new version achieved an improvement in the understandability of events as spatio-temporal happenings. Compared to the first evaluation less participants interpreted events as a place only. Regarding the temporal restrictions more mistakes were made. However, one can assume that for real-world cases the complexity of needed restrictions would be much lower, and queries would not be as complex as they were in the study where the purpose was mainly to evaluate all language components and arrangement possibilities. In the smaller tasks participants got the restrictions mostly right. Lastly, the participants were able to understand and apply VESPa after a short amount of time. VESPa can thus be seen as an intuitive or at least easy to learn domain specific visual query language.

Chapter 7

MOBY - The Mobility Analysis System

In Chapter 2, a process model for visual analysis of human mobility behavior was introduced (see Figure 2.8). The presented pipeline suggests both, a bottom up process (from data to theory) and a top down process (from theory to data) for exploratory and confirmatory analysis in an iterative, and flexible manner. In the subsequent chapters (Chapter 3 to 6), each step of this pipeline was exemplified by novel visual analytics approaches. While each approach can be applied in isolation, the best results are obtained when they are integrated into a common system[1] that realizes the multitude of loops in the process model. This allows users to go back and forth in the analysis process to widen or narrow information retrieval, to loosen or tighten rules for the enrichment, and to perform further exploration, querying, and pattern extraction. During the research on this topic, a system named MOBY (MOBilitY Analyzer) was developed (see Figure 7.1). MOBY's analysis workflow realizes best practices identified through a step by step integration and exemplary evaluation on multiple application scenarios with different datasets. In addition, MOBY reduces repetitive development efforts on data persistence, event handling, and layouting of facets and views that are needed to build novel approaches on top of the available solutions. It provides an architecture that allows for an easier adoption to future datasets and application scenarios. MOBY's architecture, user interface, and workflow are presented in Section 7.1. Therafter, Section 7.2 reviews possibilities and measures taken to maintain scalability for large and complex data.

[1] from ancient Greek *sýstēma* – whole compounded of several parts, www.perseus.tufts.edu, last accessed: March 22, 2017

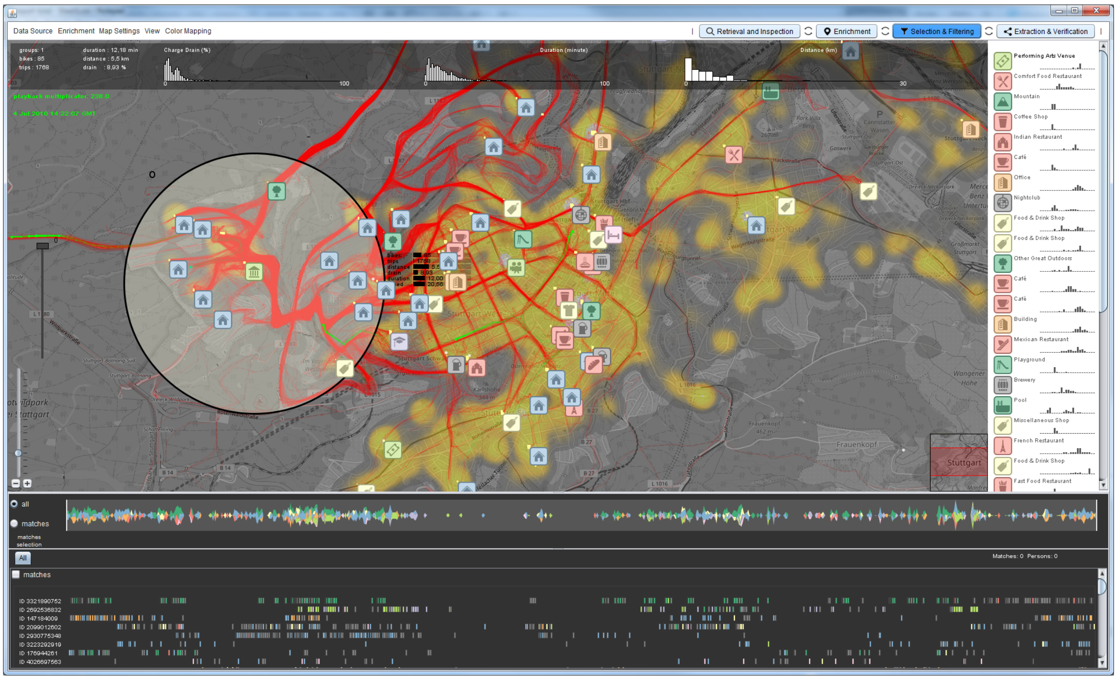


Figure 7.1 — MOBY offers coordinated views to visually analyze movement data. The setup shown in this image employs an origin lens to filter and explore trajectories and context in a geographic map. The timeline shows the linear order of POIs visited by users, accordingly. The side panel gives information about the temporal daily visit distribution per POI.

7.1 Integration

The integration of the previously presented approaches takes place on three levels. On the lowest level, MOBY's software architecture realizes a flexible technical integration. It makes use of the common characteristics of movement data (see Section 2.3) that allow for a generic data model and design of components. However, there will always be data and application specifics. MOBY's architecture takes this into account with different means to plug-in additional components, such as data readers, web-service connectors for enrichment, and visualizations. On a second level, MOBY allows the analyst to combine a variety of spatial, temporal, and other views and interfaces to select and filter the movement data. They are connected by brushing and linking. On a third level, MOBY consecutively leads the analyst through the analysis pipeline, while keeping the flexibility to step back and forth in iterative loops. This is realized by so called *facets* that provide a preselection of combined visualizations for different steps of the analysis process. More details on the technical realization of these concepts are presented in the following.

7.1.1 Architecture

MOBY is implemented in Java 8[2]. The system was mainly built to develop and test new research. Therefore, the architecture follows a modular concept that gives high flexibility on customization for different domains, data, and tasks. Each component encapsulates a certain task and offers a thin interface for an event-based communication with other components. Figure 7.2 gives an overview.

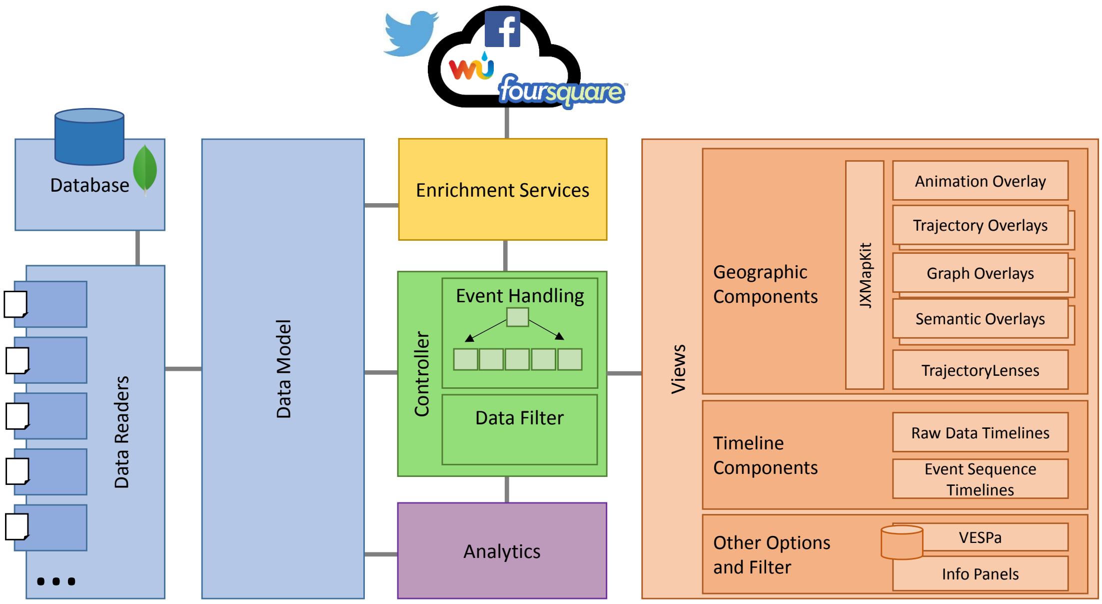


Figure 7.2 — The system has a layered architecture, realizing the model-view-controller pattern. The controller is the core element of the system that initializes and manages the interplay of the components.

MOBY's architecture follows the model-view-controller design pattern. A plug-in capability allows to write new movement data readers. Data can be loaded and transformed to match the structure of MOBY's inner data model (see Figure 7.2, blue components). A simplified version of this model was shown in Section 4.2. Data can also be persisted in NoSQL databases (e.g., MongoDB[3] and Apache Lucene[4]), which speed up data imports by restructuring, and spatial and temporal indexing. Whether database capability is used or not depends on the file-reader implementation. Currently, around ten different readers import data from taxi fleets, rental cars, electric scooters, pedestrians, geo-located microblog messages, and individual holiday trips. Independently of the data persistence, many parts of the data are loaded into main memory once a dataset is requested by the analyst. Here, spatial index structures organize origin and destination points of trajectories

2 www.java.com/download/, last accessed: March 22, 2017
3 www.mongodb.com, last accessed: March 22, 2017
4 lucene.apache.org, last accessed: April 20, 2017

in quad-trees. This guarantees immediate visual feedback to dynamic querying and ad-hoc filtering, as needed for, e.g., TrajectoryLenses (see Chapter 5).

To display the movements, the data model is accessible by all views (orange components). These comprise a geographic map with several overlays as well as timelines, filters, and statistics views. They are presented in more detail in Section 7.1.2. The controller (see Figure 7.2, green component) coordinates global data selection and filtering. Its event handling functionality implements the observer design pattern [Gamma, 1995], giving high flexibility to orchestrate visualizations depending on the task. Multi-threading is used to speed up interactive brushing and linking. If a selection is carried out through user interaction with the visual interface, the view utilizes the event handling functionality to forward an event to the data filter that provides different geographical, temporal, and abstract filtering operations. After the results are evaluated, the event handling notifies all subscribed views that re-render their visualizations to match the filtered data. Depending on the event type, the subscribers also include enrichment services and analytics components. The enrichment services (yellow component) provide means to enhance movement data with additional context, mainly from social media. The analytics component (highlighted in purple) contains algorithms for data processing and mining. Here, movement data is aggregated to graph structures, sorted, and clustered, and statistics are calculated.

7.1.2 User Interface Components

MOBY's user interface consists of different components that can roughly be categorized into spatial views, temporal views, and filter/query views. They were developed based on the UI libraries Java Swing and JavaFX[5].

For geographic visualization the library JXMapViewer[6] was used. It offers a two-dimensional geographic map and various transformation and interaction functions to calculate pixel positions from geolocation and vice versa, depending on the current zoom and viewport. The satellite and aerial imagery is retrieved from map servers such as OpenStreetMap[7] that provide a variety of map tiles to emphasize different aspects, comprising roads, buildings, landuse, or topography. In most views the tiles only show the most important information in different gray values, so that the geographic visualizations of MOBY overlaying these maps do not interfere with the map context. Different layers allow for different perspectives (Figure 7.3, a-f). For example, the animation overlay allows to replay movement data at different speed (a). An overview of a selected time period can be retrieved with alpha blending and edge splatting (b), or spatial aggregation based on different grids. This aggregation can either be visualized with node-link diagrams (c) or arrow

[5] https://docs.oracle.com/javase/8/javase-clienttechnologies.htm, last accessed: Feb. 10, 2017
[6] https://github.com/msteiger/jxmapviewer2, last accessed: February 9, 2017
[7] https://www.openstreetmap.org, last accessed: February 9, 2017

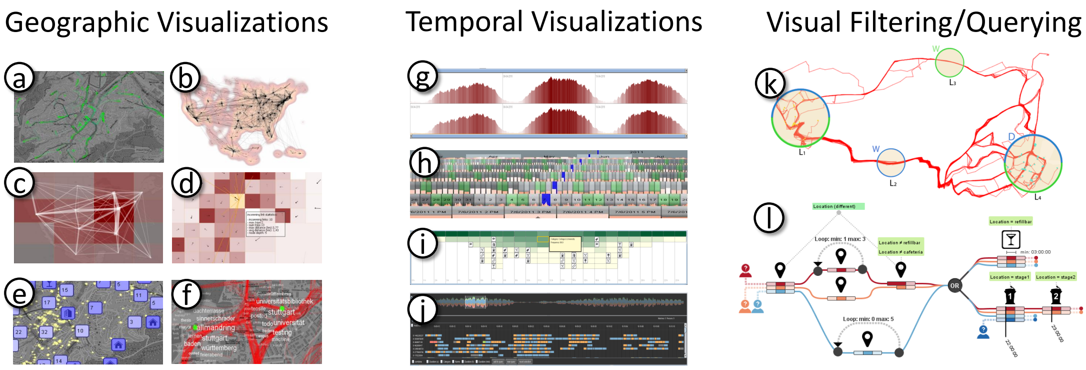


Figure 7.3 — The system offers different spatial (a-f) and temporal (g-j) visualizations that allow for navigation, selection, and filtering. More elaborated filtering/querying is supported by TrajectoryLenses and VESPa (k, l).

glyphs/flow maps (d). Enriched movements can be geographically investigated with symbol overlays (e) or word clouds (f). The overlays can be stacked on top of each other, e.g., to inspect the trajectories as a whole, while enhancing them with semantic information and animated moving objects that reveal main movement flows and direction for a given timeframe.

Furthermore, there exist four temporal views. For the aggregated trajectory numbers, a two-fold timeline provides overview+detail, navigation, and temporal filtering (g). In addition, a hierarchical view on the temporal dimension is provided, especially useful for long time ranges (h). For semantically enriched data, two event-sequence visualizations allow to investigate the happenings in linear or cyclic temporal order for current filter aggregations (i, j) or for individual POIs.

Multiple query and filter tools offer means to retrieve data and define restrictions in space, time, and other dimensions. They also allow to control the degree of data aggregation. Two more elaborated filter approaches are given with TrajectoryLenses (k) and VESPa (l). TrajectoryLenses mainly provides means to explore unaggregated trajectories in space, while VESPa focuses on explicit queries, not necessarily bound to a specific geographic location. Lastly, menu components allow to load and export data, to choose which views and filters are shown in the interface, and to decide which services are used to add contextual data.

7.1.3 Workflow

Chapters 3 to 6 presented each step of the analysis chain in detail but isolation. MOBY realizes a smooth interplay of these approaches, following the semantic mobility analysis process, presented in Section 2.4.3. While analysts can freely configure which views should be shown, the system also proposes so-called *facets*,

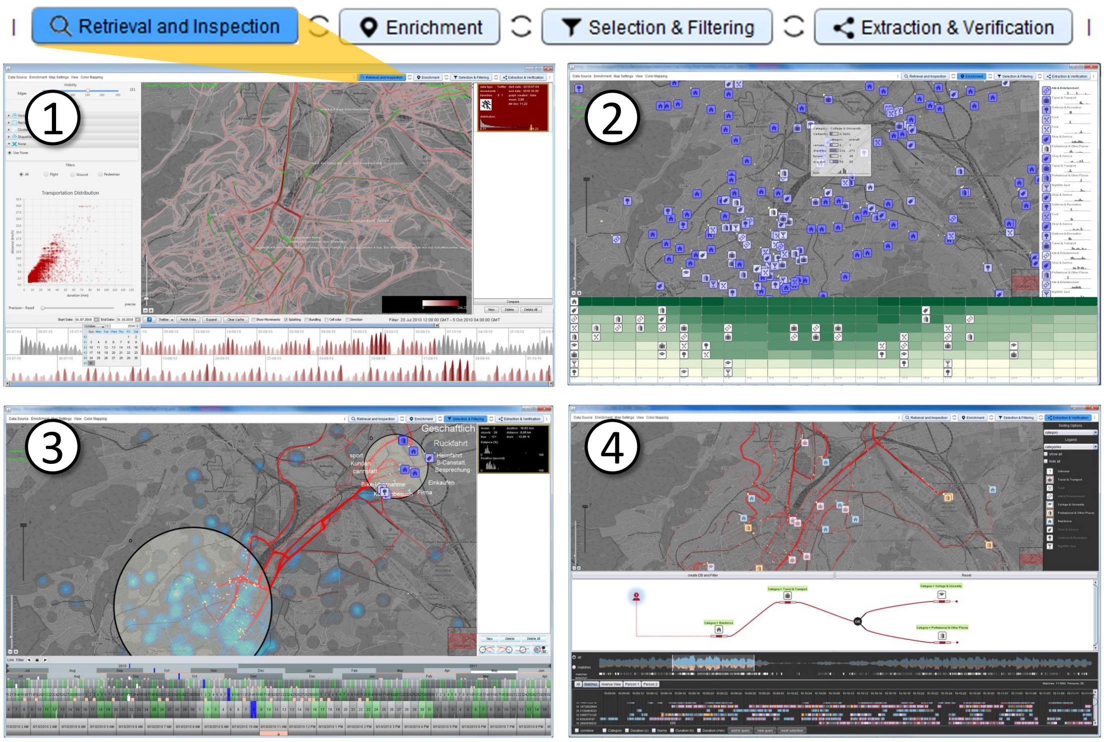


Figure 7.4 — The system offers different facets. They pre-orchestrate visualizations for different steps of the analysis: (1) retrieval & inspection; (2) enrichment; (3) selection & filtering; (4) extraction & verification.

i.e., pre-defined arrangements that support different stages of the analysis. As can be seen in Figure 7.4, the different steps of the pipeline (introduced in Section 2.4.3) are shown in the menu bar. For each step, a facet orchestrates the most suitable visualizations to support the analyst during retrieval and inspection, enrichment, selection & filtering, and pattern extraction.

Retrieval and Inspection

The analyst can start the analysis using the first facet (see Figure 7.4, 1). She can choose a dataset of interest under different spatial and temporal restrictions. To facilitate the selection, MOBY adopts a calendar widget and an area selection [Bosch et al., 2013] in the geographic view. Thereafter, the data is fetched. Different data readers allow to connect to different data sources. Data is either retrieved from the file system or database or comes from a remote service (see Figure 7.2, blue components). To gain an overview of main movement flows, the analyst can aggregate the movements based on different structures and data distributions (e.g. adaptive grid, Voronoi tessellation, arbitrary shapes from clusters). She can activate different visualizations (see Figure 7.2, view components) such as alpha blending and edge splatting to see finer grained structures of the trajectories. She can also explore the data in the animation overlay. Multiple datasets can be loaded and held in the data model, allowing the user to juxtapose visualizations or use means for an explicit comparison.

Enrichment

To get a semantic understanding of the movement data the analyst can switch to the *enrichment* facet (2) and utilize one of the implemented capabilities available in a dropdown selection in MOBY's main menu. Current implementations allow for an enrichment in a manual or in a semi-automated manner using context data from different webservices the analyst can choose from. This includes enrichment with POIs (based on Foursquare) or with a geo-referenced textual data source, such as Twitter (see Chapter 4). The enrichment can be carried out on individual trajectories or in an aggregated manner. The latter relies on a density-based clustering of stop points that can be carried out on demand or can be pre-calculated. Once the context data is received, it is persisted locally, allowing for a faster retrieval for resumed utilization.

Selection & Filtering

By switching to the *selection & filter* facet (3), TrajectoryLenses is activated. The analyst can drill down to subsets of the data and investigate user-based or region-based mobility profiles, route choice behavior, etc. This can either be done on raw or enriched movement data. The analyst can select a lens (origin, destination, intersection, user-based) of choice (see Section 5.2). A combination of multiple lenses creates Boolean expressions that are evaluated in the filter component (see Figure 7.2, green area), accordingly. The lenses allow not only for an exploration but also to query for specific cases. A limitation is that their application is bound to specific geographic areas.

Extraction & Verification

To formulate, detect, and extract more general semantic patterns, the analyst can switch to the *extraction & verification* facet (4). It consists of temporal event visualization, automatic pattern detection, and visual querying. The pattern detection is closely coupled with the visual representation and gives a wide range of options to find routines and outliers (see Section 6.2). Its computation takes place in the analytics component (see Figure 7.2, purple area). VESPa allows to query for sequences of happenings and meetings, etc. If no geographic restrictions are specified, these queries are independent from spatial limitations. The analyst can formulate patterns or retrieve them from the data. She can iteratively refine them, search for support, explore results, and finally extract a compact representation of findings. A future task would be to generate reports on the applied analysis process and findings.

7.2 Scalability

Thomas and Cook [2005b] motivated the importance for scalable analytics with the growing amounts of data. According to an IDC survey (International Data Corporation), the world-wide amount of generated data is expected to exceed 40 zettabytes till 2020 [Gantz and Reinsel, 2012]. The ability to cope with the growing amounts and complexity of data is commonly discussed under the term *scalability*. Different authors have categorized scalability aspects in the area of visual data analysis. Eick and Karr [2002] discuss factors that affect visual scalability: human perception, monitor (display) resolution, visual metaphors, interactivity, data structures and algorithms, and computational infrastructures. From these factors, Thomas and Cook [2005b] deduce a multitude of scalability types: software (computational) scalability, visual scalability, display scalability, information scalability, and human scalability. Although it is difficult to define a set of disjunct scalability aspects, this categorization gives a good compromise. Computational scalability addresses the data processing, which concerns algorithms and hardware. Visual scalability is referred to as the capability to visually represent large and complex data by suitable visual mappings and metaphors. Display scalability addresses the display resolution and size. Both have influence on perception. Information (interaction) scalability is discussed as a measure how well the approaches allow to extract relevant information, and how well they scale with the rate of change of information. It can also measure how well the approaches adapt to the audience, which might be experts or novice users. Lastly, human scalability addresses the number of people that can be involved in the analysis and which level of collaboration is supported.

MOBY allows to analyze movement data of different scale and different quality for different application domains and tasks. To achieve this, the main challenges regarding scalability were identified to be computational scalability, visual/display scalability, and information scalability. Accordingly, this section reviews the aspects in more detail, discusses how MOBY achieves scalability, and argues what can be improved in the future.

7.2.1 Computational Scalability

Computational scalability comprises the upscaling of computational data processing to larger datasets regarding the structure, the number of data instances, and the number of dimensions. This is of great importance for human mobility analysis because datasets can contain millions of trajectories. Especially exploratory approaches rely on dynamic queries with immediate visual feedback [Shneiderman, 1996], which requires solutions for fast data processing. A technique that works for individual traces of one or more persons may not be suitable to process traces of a large fleet. To keep computation applicable for larger datasets several measures can be taken (see Figure 7.5). MOBY implements a subset of them.

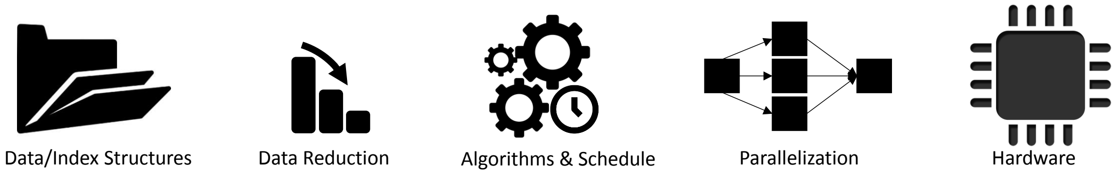


Figure 7.5 — Computational scalability can be achieved through suitable data and index structures, data reduction, by choosing scalable algorithms and scheduling, by process parallelization, and by improvements in hardware.

Data and Index Structures

The efficiency of data processing depends on how the data is stored and accessed. Relational databases usually normalize data into distinct tables that can be joined on demand. This gives high flexibility and reduces redundancy but can cause higher costs for accessing the needed data. By contrast, denormalized data tables can make querying fast but provide less flexibility. MOBY handles this trade-off with a combination of both. For fast data retrieval MOBY relies on the NoSQL databases MongoDB[8] and Apache Lucene[9] that offer a variety of indexes for geospatial queries. Once a subset of the data is loaded into the system (into main memory), a relational model is created (see Section 4.2) that allows for flexible filtering. To maintain high interactivity and immediate visual feedback, MOBY indexes trajectory measurements with quadtree structures. MOBY's implementation of DBSCAN (see Section 4.3.2), for example, utilizes them to speed up range and nearest neighbor queries. Similarly, the lens filters (see Section 5.2) use index structures for fast range queries, allowing to immediately select the subset of trajectories in focus. Future improvements could include advanced index structures such as R-trees [Guttman, 1984] and high dimensional index structures [Bentley, 1975] for fast access of trajectories as a whole. Suitable means for high-dimensional indexing of moving object trajectories have been summarized by Pfoser et al. [2000].

Data Reduction

Depending on the degree of detail that is necessary for the analysis, data can be reduced, so that it takes less processing time to display and analyze the data. Liu et al. [2013] categorize data reduction methods into filter and sampling, binned aggregation, model-based aggregation, and hybrid reduction. Firstly, filtering and sampling allow to select subsets of the data. This can be achieved by sorting out data items not fitting a spatial, temporal, or other attribute range as well as by random subsampling. The latter may results in loosing important details but still allows to extract main routines of the data. MOBY's lenses provide means for ad-hoc filtering based on spatial restrictions. To define multiple spatial criteria, multiple lenses can be combined, which further reduces the filtered subsets. Filtering can

[8] www.mongodb.com, last accessed: March 22, 2017
[9] lucene.apache.org, last accessed: April 20, 2017

also be applied in the temporal domain as well as in other domains (e.g., on context types). Secondly, binned aggregation groups data items based on adjacent intervals. MOBY's hierarchical aggregation relies on binning. Movement data is aggregated based on the underlying grid structure, resulting in a weighted and directed graph. Lastly, model-based aggregation describes the data with mathematical features or statistic summaries [Liu et al., 2013]. MOBY applies a variety of model-based reduction methods, ranging from simple statistics to more complex unsupervised machine learning approaches such as hierarchical clustering.

Algorithms and Scheduling

For the analytic components there is always a trade-off between accuracy and complexity. The challenge is to balance runtime and storage in a way so that it best fits the analysis tasks. Often, algorithms with lower complexity come at the price of losing accuracy but allow for more interactivity in the system. On the other hand, more elaborated algorithms have higher complexity but are more accurate and allow to solve the analysis tasks more precisely. In this thesis, methods range from simpler and faster approaches such as binning and k-means clustering to more elaborate means such as DBSCAN and spectral clustering. In addition, MOBY makes use of precomputation. This does not reduce computational costs, but shifts the processing to a period where it is less time critical because no interactive analysis is taking place. Interaction, visualization, and analysis tasks are sped up by precalculating distance matrices for sorting, clustering, and outlier detection. For the hierarchical clustering (see Section 6.2) not only the distance matrix but also the clustering itself is precomputed. This allows immediate visual feedback when the analyst changes the level of detail. A drawback is that clustering has to be rerun when the parameter configuration is changed.

Parallelization

Depending on the task, data processing can be parallelized. MOBY allows to speed up data imports by processing data from different files in parallel. To guarantee a fluent system handling, it separates data processing from data visualization and interfaces using threading. To ensure a fast delivery of events to all components, the event management is multi-threaded as well. The degree of parallelization depends on the environment where MOBY is executed. MOBY optimizes the number of threads based on the underlying hardware by detecting the number of cores available. However, the number of components that make use of parallel processing in MOBY is still limited. To scale the capabilities for larger datasets, a future challenge will be to add further parallel processing, e.g., for data mining components. One option is to externalize larger tasks to computing clusters or execute them on GPUs using parallel computing architectures and interfaces such as CUDA[10] or OpenCL[11]. A common method to run algorithms in parallel on multiple machines is *map-reduce* [Dean and Ghemawat, 2008], which first splits up the data into multiple chunks, processes these chunks on multiple machines and then joins the results. This is especially helpful to deal with data velocity, e.g., large data streams where aggregates have to be computed and processed in realtime.

7.2.2 Visual Scalability

Does it visually scale? is a common question from visualization researchers when they rate a visualization's usefulness. There are different limitations that hinder an unlimited displaying of data (see Figure 7.6). First, there is limited space available. Even when each data item or data dimension is displayed as single pixel, at a certain extent the data is larger than the number of pixels available. Second, humans have a limited capability of processing and understanding information. Showing details exceeding these limitations causes information overload, which hinders the analysis more than it helps. Data mapping and rendering that leads to a clear visualization for a few data items does not necessarily lead to a clear representation of a billion data items. As MOBY deals with large amounts of movement data, visual scalability was one of the main requirements for most visual components.

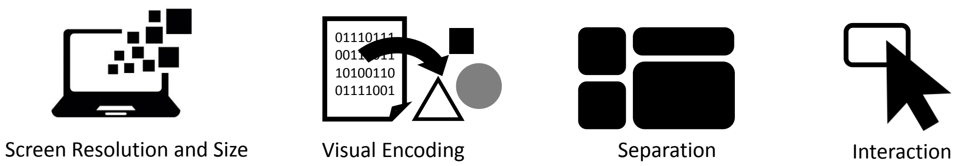


Figure 7.6 — Different means affecting visual scalability are screen resolution and size, mappings to visual primitives to build metaphors, separation in space and time, and the degree of interaction.

Screen Resolution and Size

Screen resolution and size are two important features that affect visual scalability. With higher resolution more pixels can be drawn and more information can be encoded on screen. Larger displays occupy a larger field of view for a similar viewing distance. This allows for an increased immersion, to identify more details, and to carry out tasks in collaboration. Screen resolution and display size depend on each other insofar as there are physical limitations to pixel sizes that a) viewers

[10] https://developer.nvidia.com/cuda-zone, last accessed: February 14, 2017
[11] https://www.khronos.org/opencl, last accessed: February 14, 2017

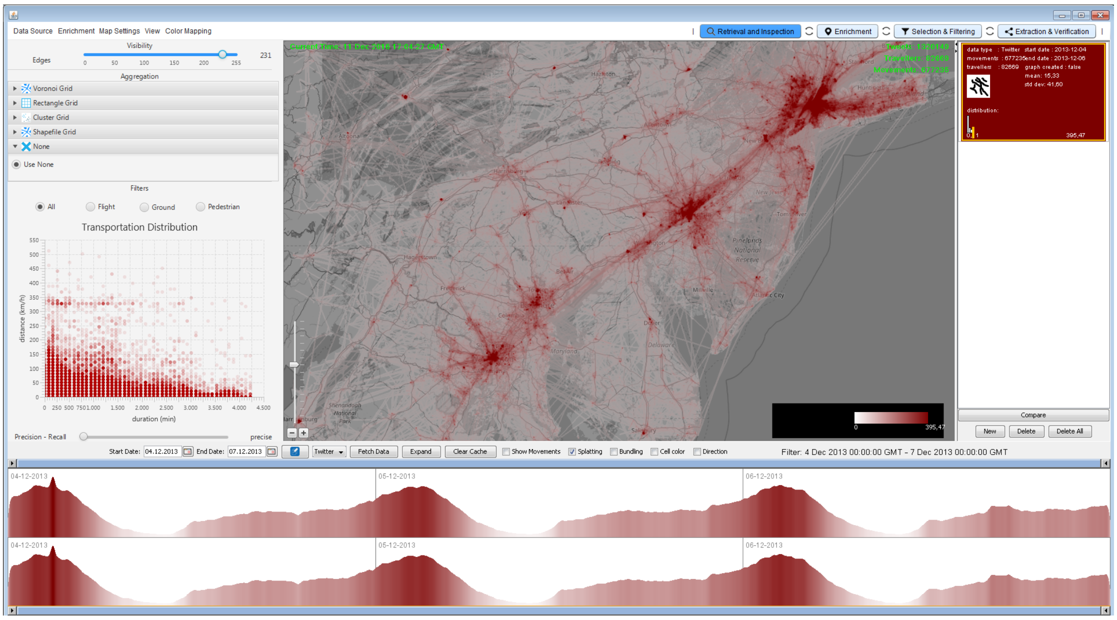


Figure 7.7 — Three days reconstructed movements from Twitter are drawn in realtime (677,235 trajectories) using hardware accelerated edge splatting.

can perceive and b) hardware is able to display. MOBY is developed to run on a desktop environment with full HD resolution. However, running MOBY with higher resolution gives advantages in identifying details and maintaining an overview within the same view, even when a larger area needs to be displayed.

Visual Encoding and Separation

A decisive factor influencing visual scalability is the visual encoding. Mackinlay [1986] proposed a taxonomy for suitable visual encoding with respect to data types. But also on a higher level, the chosen visual metaphor can influence visual scalability, as discussed by Eick and Karr [2002]. Liu et al. [2013] argue that in addition to computational data reduction, suitable visual encoding allows to improve scalability by visual aggregation, displacement, and distortion.

MOBY comprises a variety of visual encodings and visual aggregations. Trajectories can be shown as path using position, length, and angle or can be represented with glyphs (position, angle), or animation (position over time), depending on the amounts, resolution, and distribution of the data. Visual aggregation is supported by edge splatting that sums up density values locally for each pixel in the current viewport and maps them to color. This results in a clutter-free visualization, especially useful to understand large movement datasets (see Figure 7.7). Secondly, while MOBY's graph-based approaches rely on computational aggregation of movements based on areas of varying sizes, the clarity of the visualization also depends on the visual encoding. Node-link diagrams draw edges with varying thickness and transparency, highlighting important areas. This scales to a certain number of nodes and edges but networks can be very dense and globally connected. In such situations, the flow map visualization (see Section 3.4.3) maintains a clutter-free representation of the network by showing only average directions on a chosen aggregation level. Thirdly, MOBY visually aggregates data in the temporal domain, as presented in Section 4.4. Layered navigation and cyclic representation let the visualization scale horizontally, and aggregation by category (e.g., by event type) allows scalability on the vertical axis. The most compact representation in MOBY is given with VESPa (see Section 6.3) making the approach scalable to large amounts of data. Patterns such as the one shown in Figure 6.9 represent complex stories on small space, no matter if the pattern applies to one or a billion movers.

Lastly, spatial and temporal separation can be seen as a special case of visual encoding. Spatial separation in multiple coordinated views (MCV) has been studied and categorized by Roberts [2007] and proved to be useful by North and Shneiderman [2000]. MOBY allows to arrange a multitude of visualizations to MCV showing spatial, temporal, and abstract perspectives on the data for different levels of aggregation. Temporal separation is given through animation and through interactions like zooming and panning.

Interaction

Interaction is an important means to maintain visual scalability. When there is too much information to show on a static visualization, interaction allows to analyze different subsets of the data on demand by navigation, selection, and data manipulation. Accordingly, MOBY's coordinated views are interactively connected through brushing and linking. Geometric zooming supports fluent geographic and temporal navigation through temporal separation, either within single views (map view) or multiple views (e.g., layered timeline views). Beside changing the geographic scale by zooming on the map, MOBY utilizes semantic zooming. Examples are the hierarchical graph exploration (see Section 3.4.3) as well as the hierarchical POI category browsing (see Section 4.3.2). Accordingly, layered navigation realizes a stepwise overview+detail separation in the timeline. This allows analysts to inspect larger time ranges, for example on a yearly basis, as well as small time ranges, e.g., on an hourly basis (see Figure 7.3, g, h, j). Lastly, interactive filtering allows to focus on subsets of the data that can then be displayed in higher detail without irrelevant aspects. This reduces clutter and overplotting. The TrajectoryLenses approach (Figure 7.3, k) allows for a very dynamic drill down by providing immediate visual feedback.

7.2.3 Information Scalability

One aspect Thomas and Cook [2005b] discuss under the term *information scalability* is the upscaling of the user interaction to larger datasets and streams. Even when hardware and software architecture are designed to scale with the large datasets, there can be issues regarding the interactive data handling. This section discusses different concepts to maintain information scalability: guidance, abstraction, and generalization (see Figure 7.8).

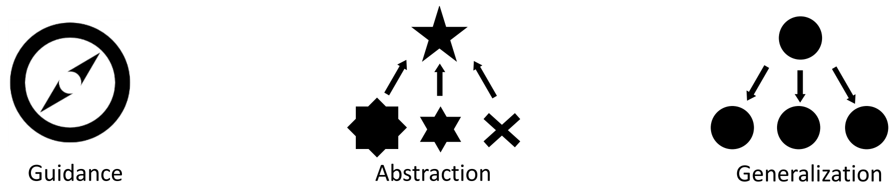


Figure 7.8 — Information scalability comprises multiple aspects. Among others, the upscaling of user interactions to handle larger datasets can be supported by guidance, abstraction, and generalization.

Guidance

When data is large and complex, navigation can get challenging because the analyst can easily oversee patterns or get lost completely. To make navigation scalable, guidance is a sufficient means. Ceneda et al. [2017] characterize different degrees of guidance ranging from giving orientation via giving direction to prescribing.

MOBY mainly provides orientation and direction for data exploration. To aid navigation of the lens filters (see Section 5.2) a heatmap guides the analyst to dense areas where many trips start, end, or events of a certain category take place. In Figure 7.9 a destination lens is spawned. The bluish heatmap shows the density of trips' destinations, accordingly. The analyst can first explore these areas with the eye, decide on which area she wants to focus, and then move the lens over to this area ($1 \rightarrow 2$). Subsequently, the filtered subset can be narrowed by additional lenses. In the lower images, the analyst adds an origin lens. Again, the heatmap (here yellow for origin points) highlights dense areas of the filtered subset. Thereafter, the analyst can move the lens to explore the paths that start in this area and end in the previously defined region ($3 \rightarrow 4$).

Abstraction

An abstract visualization often allows for a more compact representation than it would be possible with a visualization that represents, e.g., geographic or temporal dimension in detail. TrajectoryLenses can filter millions of trajectories. Filters are, however, always bound on a certain geographic area. The pattern-based query language VESPa (see Section 6.3), by contrast, is expressed in abstract space,

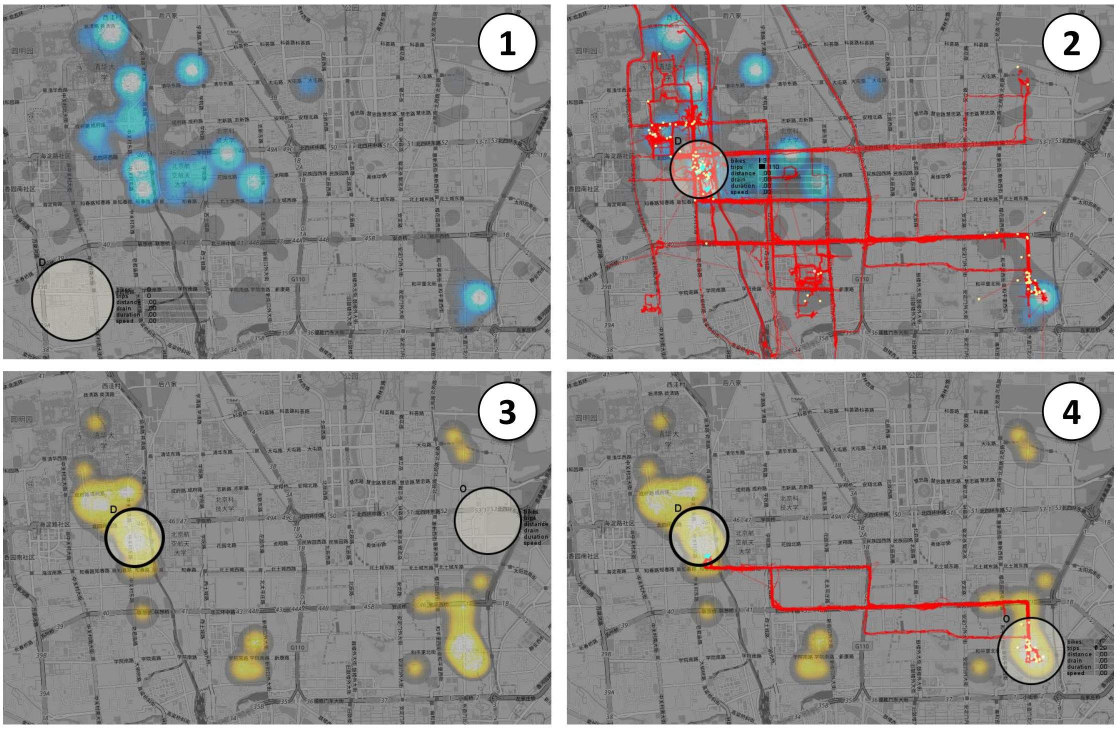
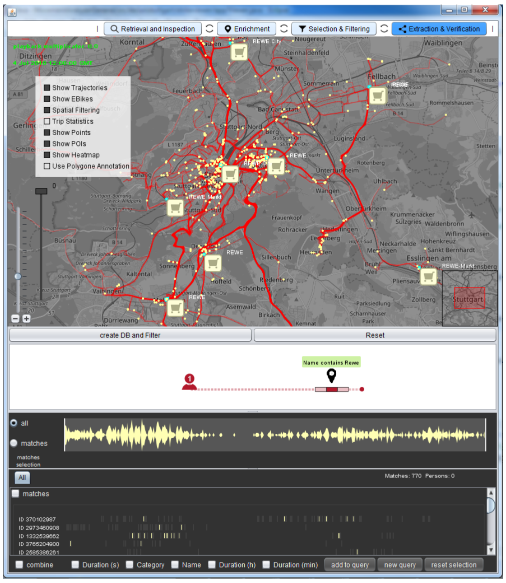


Figure 7.9 — Heatmaps guide the analyst: 1) The bluish heatmap shows areas where many trips end. 2) The lens is moved to one of the areas. 3) An origin lens is added. The yellow heatmap guides to starting points. 4) The lens is moved to such an area.

allowing to match all cases by defining a single pattern. Figure 7.10 gives an example. A simplistic VESPa query is used to filter a multitude of shopping locations of a certain brand. As presented in Section 6.5 this works also in opposite direction. By taking examples from the data one can abstract them to a common compact pattern.

Generalization

Data selection and manipulation (adding, deleting, or changing data values) is a rather simple task for a small number of data items, but can be challenging for large datasets. In Section 4.3.2, for example, relabeling of single POIs was proposed. As shown in Section 4.5.1, the investigation of single instances with domain knowledge increases accuracy, but only scales up to a few hundred items. One way to make data manipulation more scalable is reducing detail. In Section 4.3.1 area-based labeling was introduced. It labels not a single but all trajectories that end in the labeled area, allowing to define regions on different levels of detail, e.g., for suburbs, cities, or countries. This, however, comes at the cost of losing information. Another sufficient means to make selection and manipulation scalable is generalization. It proceeds from a premise about a sample to a conclusion about the whole. Applied



◀ **Figure 7.10** — VESPa is applied to filter for trips to a certain grocery brand. The map view shows the resulting set of trips and POIs.

to visual analytics approaches, analysts can, e.g., interact with a subset of the data and treat the remainder in a similar vein. MOBY makes use of this concept by letting users define or select an example with VESPa. The system then finds similar cases, accordingly. For more complicated cases, future research could focus on the integration of supervised learning, as done with *ScatterBlogs2* [Bosch et al., 2013]. Classifiers can be interactively created, trained, and applied to categorize geo-located microblog data, depending on the content. These classifiers can then be applied to filter the massive data streams according to, e.g., floods, riots, and epidemics to support situational awareness and crisis response. Integrated in MOBY, supervised learning could improve semantic enrichment with respect to large data sets, especially for repetitive tasks.

In summary, MOBY addresses computational, visual, and information scalability by offering a large amount of different views and analytical possibilities. This allows analysts to cope with large and complex data sources. Depending on the tasks, different approaches can be selected that highly aggregate the data and give an abstract view or provide detailed information on single instances. As the decisions which approaches to use in which situation can be difficult, MOBY guides users step by step through the analysis process, which ranges from data retrieval to the extraction of behavior patterns. In this process, analysts still maintain the necessary degree of freedom to step back and forth for iterative refinement and to freely orchestrate different perspectives and tools on the data.

Chapter 8

Conclusion and Outlook

At the beginning of this thesis, the need for an increased semantic understanding of human mobility behavior was motivated with the omnipresence of movement in our world, the potential it comprises for government and industries to solve existing and upcoming challenges, and the increasing technical possibilities to record, store, and make sense of the large and complex data. Nevertheless, much context is lost during data acquisition, making it challenging to infer meaning from recorded movements. Social sensing data, especially social media, was identified as a suitable source to bridge this semantic gap with additional data, be it areas and places, opinions, or eyewitness reports. For a semantic analysis of human mobility behavior, different challenges regarding data *volume*, *variety*, and *veracity* were identified that bring automatic computations as well as human perceptual and analytical capabilities to the brink of feasibility. Visual analytics was introduced to cope with these challenges by closely combining both, the strength of humans to perceive information visually and interpret patterns with domain knowledge and the accuracy and speed of computers. These motivations and opportunities draw the main research question (see Section 1.2), whether VA and the integration of multiple data sources gained through social sensing can increase the semantic understanding of human mobility patterns (question 1). A closer investigation of the challenges revealed secondary questions regarding the role of social media (question 2), the integration of human experts (question 3), and the application to different domains (question 4).

8.1 Contributions and Limitations

The research contributions in this thesis are multi-faceted. While each presented approach brings its own advances, a major contribution lies in the combination of the individual steps to a generic mobility analysis process (addressing research question 4). This process allows for a stepwise enrichment and narrowing from raw movement data to specific behavior patterns, describing routines and anomalies. The process was created and applied to solve specific challenges in application domains comprising electric mobility, law enforcement, indoor event management, and social media driven situation awareness. The generalization and application to multiple domains was enabled through an integration of all approaches in a visual analytics system named MOBY. The developed approaches provide a close integration of the human in the analysis loop through highly interactive interfaces and immediate visual feedback (addressing question 3). Especially, the interactive coupling of exploration and verification approaches enables a more natural analysis that aligns well with the basic modes of inference [Peirce, 1974].

Taking a closer look on the individual approaches, this thesis presented methods for human mobility analysis that make use of the increasing potential of social media data (question 2). Visual analytics methods were developed to reconstruct large amounts of movement data from geo-located messages. They allow to control the degree of precision and recall during retrieval and to explore and compare datasets on different levels of detail. Evaluations showed that the data contains valuable mobility patterns on a global and local level. Furthermore, social media data was used to derive missing context information. Semi-automated enrichment methods were proposed to enhance movement data with information about visited places, but also with textual information, comprising people's observations, opinions, and moods. The semantic enrichment offers new possibilities to understand reasons behind movement, for example why and under which conditions people move, with whom they meet, and which consequences may arise. To extract such information, analysts can navigate, filter, and drill down in the large and heterogeneous datasets. This is made possible by a powerful lens-based filter technique and a more abstract visual query language. The former allows for interactive exploration with immediate visual feedback. The latter provides means to extract, refine, and query for mobility behavior patterns. Automated techniques support the analyst coping with large datasets through sorting, grouping, and outlier detection. Different evaluations and a categorization of scalability aspects revealed potentials, strengths, and weaknesses of the approaches' realization and pointed out future challenges.

Overall, the results show that visual analytics and the integration of multiple data sources, derived through social sensing, can greatly increase the semantic understanding of human mobility behavior (question 1). Nevertheless, the success of VA approaches strongly depends on the analyst's knowledge and capabilities and on the available data. What is not in the data cannot be discovered. This is especially an issue when context information is missing or cannot be mapped accurately due to different resolutions and inaccuracies. While it is often better to gain a vague idea of what has happened than knowing nothing at all, there always remains a chance of making wrong decision due to confidence in false analysis results. Uncertainties that are not communicated to the analyst can easily lead to misinterpretations. Especially in social media, information is often imprecise, false, or missing, leading to uncertain results. This accounts for the reconstruction of movements from geo-located messages as well as for the enrichment with additional context. To deal with these challenges, the contributions of this thesis integrate the analyst in the sensemaking process. This allows to bring in domain knowledge and judge data quality, and increases the awareness of uncertainties. In the future, the integration of further data sources could provide additional insights and evidences.

8.2 Open Challenges

The contributed mobility analysis approaches allow to solve an increasing amount of tasks but there are still major challenges. These can roughly be categorized into big data challenges, new technological challenges, application specific challenges, and lastly privacy-related challenges.

Regarding big data challenges, this thesis presented approaches that cope with variety and veracity issues of large and heterogeneous datasets but there is a need for a more standardized process. Unfortunately, semantic web technology is still more an idea than applicable to the vast real world data, not yet fully accepted and supported, and does often not scale from a computational point of view. A more light-weighted description language, supported and fostered by the economy, and a better mapping of uncertain dependencies is needed. Another big data challenge arises from the increasing speed with which data becomes available. The application cases presented in this thesis were not time-critical. However, real-time processing will play a bigger role in the future. To cope with data velocity and the growing amounts of increasingly complex data, a closer integration of automated approaches will be needed to support users in the analysis process. Algorithms should be made as controllable and understandable as possible in terms of steering and its effect on the output. Besides uncovering behavior patterns and correlations in mobility data, future visual analytics approaches could support the discovery of causes and effects by making use of recent advances in causal machine learning [Hoyer et al., 2008, Mooij et al., 2016]. To improve and evaluate the accuracy of such methods more benchmark/ground truth datasets need to be established.

Another open challenge concerns the integration and evaluation of new technologies. Classical analysis, carried out on desktop computers is currently still the most comfortable and preferred way. However, the visual analytics community has started to make us of new technology such as virtual and augmented reality, that allows for increased immersion, more intuitive interaction, and a stereoscopic view that may put 3D information visualization back on the agenda. Likewise, power walls give means for collaborative analyses and offer a wide field of view that brings computational challenges in terms of interactive rendering. Other possibilities arise from the ubiquity of smartphones and wearable devices with increasing display resolution and computing power. They enable a more casual analysis for individuals right where the movement data is recorded, allowing them to understand, question, and improve own mobility habits.

With new technology and increasing digitalization new application fields arise. While generic solutions are worth striving for, visual analytics solutions are often most valuable if they address the needs and specifics of the application domain. For this reason, an adaptive process to new application areas will always exist. For the evaluation of domain specific solutions, a close integration of stakeholders is essential. Long term studies of visual mobility analytics approaches are still rare, but they are needed for improvement and dissemination of the research in practice.

Lastly, a field barely addressed in this thesis is privacy. The increasing amounts of personal information raise concerns on the protection of people's private life. Especially the enrichment with context data shows that a semantic picture can be recovered even from anonymized movement traces. While in many fields this may be a small issue, there are application areas where new privacy preserving means are strongly needed to protect citizens from companies or governments that want to gather as much information as possible. There does not necessarily have to be a trade-off between preserving privacy and gaining insights. For most application cases, aggregated profiles are expedient, because they represent larger groups of interest. Analytical solutions that can produce useful information but also safeguard people's lives are needed [Laube, 2014], because social sensing and large-scale semantic analysis of human mobility data has just begun.

Bibliography

J. Abello and F. van Ham. Matrix Zoom: A visual interface to semi-external graphs. In *IEEE Symposium on Information Visualization (InfoVis)*, pages 183–190, 2004. [page 42]

R. Agrawal and R. Srikant. Mining sequential patterns. In *International Conference on Data Engineering (ICDE)*, pages 3–14, 1995. [page 119]

R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM Sigmod Record*, volume 22, pages 207–216. ACM, 1993. [page 119]

C. Ahlberg and B. Shneiderman. The alphaslider: a compact and rapid selector. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 365–371, 1994. [page 103]

W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. PlanningLines: novel glyphs for representing temporal uncertainties and their evaluation. In *International Conference on Information Visualisation (IV)*, pages 457–463, 2005. [page 131]

W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer, 2011. [page 27]

J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983. [page 131]

L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In *ACM International Symposium on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 22:1–22:8. ACM, 2007. [page 66]

K. Andrews, M. Wohlfahrt, and G. Wurzinger. Visual graph comparison. In *Conference on Information Visualisation*, pages 62–67. IEEE Computer Society, 2009. [page 42]

G. Andrienko and N. Andrienko. Spatio-temporal aggregation for visual analysis of movements. In *Symposium on Visual Analytics Science and Technology (VAST)*, pages 51–58. IEEE, 2008. [page 26]

G. Andrienko and N. Andrienko. *Visual Analytics of Movement*. Springer-Verlag Berlin, 2013. [pages 25, 26, 27, 67, and 68]

N. Andrienko and G. Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, 12(1):3–24, 2012. [page 42]

N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi. Space transformation for understanding group movement. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2169–2178, 2013a. [page 28]

N. Andrienko, G. Andrienko, and G. Fuchs. Towards privacy-preserving semantic mobility analysis. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 19–23. Eurographics Association, 2013b. [pages 28 and 67]

C. Appert and J. D. Fekete. OrthoZoom scroller: 1D multi-scale navigation. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 21–30, 2006. [page 103]

D. Archambault. Structural differences between two graphs through hierarchies. In *Graphics Interface (GI)*, pages 87–94. Canadian Information Processing Society, 2009. [page 42]

D. Archambault, H. Purchase, and B. Pinaud. Difference map readability for dynamic graphs. In *Graph Drawing*, pages 50–61. Springer, 2011. [page 42]

J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 429–435. ACM, 2002. [page 120]

J. Banks. *Handbook of simulation: principles, methodology, advances, applications, and practice*. John Wiley & Sons, 1998. [page 22]

A. Barsky, T. Munzner, J. Gardy, and R. Kincaid. Cerebral: Visualizing multiple experimental conditions on a graph with biological context. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1253–1260. [page 42]

F. Beck and S. Diehl. Visual comparison of software architectures. *Information Visualization*, 12(2):178–199, 2013. [page 42]

F. Beck, R. Petkov, and S. Diehl. Visually exploring multi-dimensional code couplings. In *International Workshop on Visualizing Software for Understanding & Analysis*, pages 1–8. IEEE Computer Society, 2011. [page 42]

J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975. [page 157]

S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. Semantic integration of heterogeneous information sources. *Data & Knowledge Engineering*, 36(3):215–249, 2001. [page 68]

I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. [page 98]

E. Bier, M. Stone, and K. Pier. Enhanced illustration using magic lens filters. *IEEE Computer Graphics and Applications*, 17(6):62 –70, 1997. [page 96]

E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: The see-through interface. In *Proc 20th Annual Conf Computer Graphics and Interactive Techniques*, pages 73–80. ACM, 1993. [pages 17 and 96]

J. I. Blanford, Z. Huang, A. Savelyev, and A. M. MacEachren. Geo-located tweets. enhancing mobility maps and capturing cross-border movement. *PloS one*, 10(6): e0129202, 2015. [pages 39 and 41]

C. Bonhomme, C. Trépied, M.-A. Aufaure, and R. Laurini. A visual language for querying spatio-temporal databases. In *ACM International Symposium on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 34–39. ACM, 1999. [page 121]

H. Bosch, J. Heinrich, C. Müller, B. Höferlin, G. Reina, M. Höferlin, M. Wörner, and S. Koch. Innovative filtering techniques and customized analytic tools. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 269–270, 2009. [page 97]

H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12): 2022–2031, 2013. [pages 7, 41, 154, and 164]

H. Bosch, R. Krüger, and D. Thom. Data-driven exploration of real-time geospatial text streams. In *Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 203–207. Springer International Publishing, 2015. [page 7]

I. Boyandin, E. Bertini, P. Bak, and D. Lalanne. Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum*, 30(3): 971–980, 2011. [page 121]

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pages 93–104. ACM, 2000. [page 127]

R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, 100 (2):390–398, 2006. [page 46]

T. Bühler and M. Hein. Spectral clustering based on the graph p-laplacian. In *Annual International Conference on Machine Learning (ICML)*, pages 81–88. ACM, 2009. [page 126]

T. Butkiewicz, W. Dou, Z. Wartell, W. Ribarsky, and R. Chang. Multi-focused geospatial analysis using probes. *IEEE Transactions Visualization and Computer Graphics*, 14 (6):1165–1172, 2008. [page 96]

L. Byron and M. Wattenberg. Stacked graphs–geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008. [pages 28 and 82]

N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):280–289, 2016. [page 46]

S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. [pages 14, 15, and 48]

D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H. J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2017. [pages 98 and 162]

L. Certo, T. Galvão, and J. Borges. Time Automaton: A visual mechanism for temporal querying. *Journal of Visual Languages and Computing*, 24(1):24–36, 2013. [page 121]

J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert. Trajectory-based visual analytics for anomalous human movement analysis using social media. In *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015. [pages 39 and 41]

R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17, 2009. [page 11]

J. Cheeger. A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*, 625:195–199, 1970. [page 125]

S. Chen, C. Guo, X. Yuan, J. Zhang, and X. L. Zhang. MovementFinder: Visual analytics of origin-destination patterns from geo-tagged social media. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 239–240. IEEE Computer Society, 2014. [pages 2, 39, and 41]

S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):270–279, 2016. [pages 39 and 41]

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012. [page 46]

A. Chua, E. Marcheggiani, L. Servillo, and A. V. Moere. FlowSampler: Visual analysis of urban flows in geolocated social media data. In *Social Informatics*, pages 5–17. Springer, 2015. [page 41]

M. M. Chun. Contextual cueing of visual attention. *Trends in cognitive sciences*, 4(5): 170–178, 2000. [page 98]

A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):2:1–2:31, 2009. [pages 16, 17, 84, and 96]

C. Coffey and A. Pozdnoukhov. Temporal decomposition and semantic enrichment of mobility flows. In *ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 34–43. ACM, 2013. [page 67]

W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In *KDD workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003. [page 120]

J. Constine. What are the pros and cons of each "places" api?, 2011. http://www.quora.com/What-are-the-pros-and-cons-of-each-Places-API. [page 74]

M. Cools, E. Moons, T. Bellemans, D. Janssens, and G. Wets. Surveying activity-travel behavior in flanders: assessing the impact of the survey design. *BIVEC-GIBET Transport Research Day 2009, Part II*, 2009. [page 67]

C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Symposium on Visual Analytics Science and Technology (VAST)*, pages 51–58. IEEE, 2009. [page 77]

D. Courgeau. Méthodes de mesure de la mobilité spatiale: migrations internes, mobilité temporaire et navettes. *Population (french edition)*, 43(4/5):877–880, 1988. [page 3]

D. Courgeau. Mobility and spatial heterogeneity. In J. Fagerberg, D. C. Mowery, and R. R. Nelson, editors, *Demography: Analysis and Synthesis*, chapter 22, pages 279–292. Academic press, 2005. [page 3]

J. Craig. For NFL, New Orleans has always been a ball. HLNtv.com, February 2016. edition.cnn.com/article/2013/01/28/nfl-super-bowl-super-gras-favorite-city-new-orleans. [page 60]

W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421, 2011. [pages 28 and 82]

F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964. [page 122]

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. [page 159]

B. Dervin. *An overview of sense-making research: Concepts, methods, and results to date.* The Author, 1983. [page 18]

J. D. Dionisio and A. F. Cárdenas. MQuery: A visual query language for multimedia, timeline and simulation data. *Journal of Visual Languages and Computing*, 7(4):377 – 401, 1996. [page 121]

A. Doan and A. Y. Halevy. Semantic-integration research in the database community. *AI Magazine*, 26(1):83–94, 2005. [page 68]

H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Symposium on Visualization (VisSym)*, pages 239–248. Eurographics Association, 2003. [page 97]

M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1129–1138, 2010. [page 80]

A. D'Ulizia, F. Ferri, and P. Grifoni. Moving GeoPQL: A pictorial language towards spatio-temporal queries. *GeoInformatica*, 16(2):357–389, 2012. [page 121]

R. Eccles, T. Kapler, R. Harper, and W. Wright. Stories in GeoTime. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 19 –26, 2007. [pages 27 and 28]

S. G. Eick and A. F. Karr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, 2002. [pages 156 and 160]

N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions Visualization and Computer Graphics*, 14(6):1539 –1148, 2008a. [page 97]

N. Elmqvist, J. Stasko, and P. Tsigas. Datameadow: A visual canvas for analysis of large-scale multivariate data. *Information Visualization*, 7(1):18–33, 2008b. [page 97]

EnBW AG. Mission E-Mobilitaet - Elektronauten in Stuttgart, 2010a. URL `https://www.enbw.com/elektronauten/`. [page 35]

EnBW AG. President of the e-mobility federal association in Berlin: EnBW starts Germany's largest electric fleet, 2010b. URL `http://www.enbw.com/content/en/press/press_releases/2010/02/pm_20100219_cu_si/`. [page 35]

A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435, 2014. [pages 16, 18, and 72]

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, volume 96, pages 226–231. AAAI Press, 1996. [pages 74 and 120]

J. Fails, A. Karlson, L. Shahamat, and B. Shneiderman. A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. pages 167–174. IEEE Computer Society, Oct 2006. [page 121]

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996. [pages 15 and 16]

N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013. [pages 25 and 58]

G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, and H. Stange. Tracing the german centennial flood in the stream of tweets: first lessons learned. In *ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 31–38. ACM, 2013. [pages 39 and 41]

T. Fujisaka, R. Lee, and K. Sumiya. Discovery of user behavior patterns from geo-tagged micro-blogs. In *Conference on Uniquitous Information Management and Communication (IMCOM)*, page 36. ACM, 2010. [page 67]

B. Furletti, P. Cintia, C. Renso, and L. Spinsanti. Inferring human activities from GPS tracks. In *ACM SIGKDD International Workshop on Urban Computing*, pages 5–12. ACM, 2013. [page 67]

G. W. Furnas. Generalized fisheye views. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 16–23. ACM, 1986. [pages 17 and 96]

G. W. Furnas. A fisheye follow-up: further reflections on focus + context. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 999–1008. ACM, 2006. [page 96]

L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in Sensor Networks (CitiSens)*, pages 26–35. Springer, 2014. [page 41]

E. Gamma. *Design patterns: elements of reusable object-oriented software.* Pearson Education India, 1995. [page 152]

J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, pages 1–16, 2012. [pages 2 and 156]

M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011. [page 42]

T. Gottron. Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. In *International Conference on Theory and Practice of Digital Libraries*, pages 94–105. Springer, 2009. [pages 79 and 80]

D. Gotz. Soft patterns: Moving beyond explicit sequential patterns during visual analysis of longitudinal event datasets. 2016. [page 120]

D. Gotz and H. Stavropoulos. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1783–1792, 2014. [page 121]

H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *Conference on Simulation and Visualisation (SimVis)*, pages 143–156, 2006. [page 77]

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928, 1995. [page 65]

B. Guc, M. May, Y. Saygin, and C. Körner. Semantic annotation of GPS trajectories. In *11th AGILE International Conference on Geographic Information Science*, 2008. [page 66]

A. Guttman. *R-trees: a dynamic index structure for spatial searching*, volume 14. ACM, 1984. [page 157]

F. Haag, S. Lohmann, S. Bold, and T. Ertl. Visual SPARQL querying based on extended filter/flow graphs. In *International Working Conference on Advanced Visual Interfaces*, pages 305–312. ACM, 2014. [page 97]

F. Haag, R. Krüger, and T. Ertl. VESPa: A pattern-based visual query language for event sequences. In *11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (IVAPP)*, volume 7, pages 48–59, 2016. [pages 7, 128, 131, 143, and 145]

I. Hacking. Historical ontology. In *In the Scope of Logic, Methodology and Philosophy of Science*, pages 583–600. Springer, 2002. [page 65]

T. Hägerstraand. What about people in regional science? *Papers in Regional Science*, 24(1):7–24, 1970. [page 27]

R. Hamming. *Numerical methods for scientists and engineers.* McGraw-Hill Inc., 1 edition, 1973. [pages 11 and 12]

J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *International Conference on Data Engineering (ICDE)*, pages 215–224, 2001. [page 119]

S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization (InfoVis)*, pages 115–123. IEEE, 2000. [pages 28 and 84]

B. Heuer, H. Roßnagel, J. Zibuschka, and J. Maucher. Empirical analysis of passenger trajectories within an urban transport hub. In *International Conference on Information Systems for Crisis Response and Management (ISCRAM*, volume 9, 2012. [page 28]

M. Hilbert and P. López. The world's technological capacity to store, communicate, and compute information. *science*, 332(6025):60–65, 2011. [page 2]

H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004. [page 97]

D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006. [page 26]

D. Holten and J. J. Van Wijk. Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, 28(3):983–990, 2009. [page 26]

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *International Conference on Neural Information Processing Systems (NIPS)*, NIPS'08, pages 689–696, USA, 2008. Curran Associates Inc. [page 167]

Y. Hu, K. Janowicz, D. Carral, S. Scheider, W. Kuhn, G. Berg-Cross, P. Hitzler, M. Dean, and D. Kolas. A geo-ontology design pattern for semantic trajectories. In *International Conference on Spatial Information Theory (COSIT)*, pages 438–456. Springer, 2013. [page 65]

C. Hurter, B. Tissoires, and S. Conversy. FromDaDy: Spreading aircraft trajectories across views to support iterative queries. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1017–1024, 2009. [page 97]

P. Isenberg, A. Bezerianos, P. Dragicevic, and J.-D. Fekete. A study on dual-scale data charts. *IEEE Transactions Visualization and Computer Graphics*, 17(12):2469 –2478, 2011. [page 96]

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254–1259, 1998. [page 98]

M. S. Jensen, R. Yao, W. N. Street, and D. J. Simons. Change blindness and inattentional blindness. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):529–546, 2011. [page 27]

J. Jin and P. Szekely. QueryMarvel: A visual query language for temporal patterns using comic strips. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 207–214, 2009. [page 121]

M. Kaplan. *The Secrets in Their Eyes: Transforming the Lives of People with Cognitive, Emotional, Learning, Or Movement Disorders Or Autism by Changing the Visual Software of the Brain*. Jessica Kingsley Publishers, 2015. [page 11]

T. Kapler and W. Wright. GeoTime information visualization. *Information Visualization*, 4(2):136–146, 2005. [pages 27 and 28]

J. Kehrer, H. Piringer, W. Berger, and M. E. Gröller. A model for structure-based comparison of many categories in small-multiple displays. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2287–2296, 2013. [page 42]

D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual Analytics: Definition, Process, and Challenges*, volume 4950 of *Lecture Notes in Computer Science*, pages 154–175. Springer Berlin Heidelberg, 2008. [pages 2, 13, 14, 17, 20, and 31]

D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. [page 13]

D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010. [pages 13 and 20]

A. Kerren, J. Stasko, J.-D. Fekete, and C. North. *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950. Springer, 2008. [page 14]

S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. Ebert. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017. [page 25]

G. Klein, B. Moon, and R. R. Hoffman. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5):88–92, 2006. [page 20]

E. Klien and M. Lutz. The role of spatial relations in automating the semantic annotation of geodata. In *International Conference on Spatial Information Theory (COSIT)*, pages 133–148. Springer, 2005. [pages 65, 66, and 71]

E. M. Klien. Semantic annotation of geographic information. Disseration. Universität Münster, 2008. [page 70]

F. Kling and A. Pozdnoukhov. When a city tells a story: urban topic analysis. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 482–485. ACM, 2012. [page 67]

C. Koch and S. Ullman. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, pages 115–141. Springer Netherlands, 1987. [page 97]

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *IEEE Transactions Visualization and Computer Graphics*, 17(5):557–569, 2011. [page 97]

B. Kochan, D. Janssens, T. Bellemans, and G. Wets. Collecting activity-travel diary data by means of a hand-held computer-assisted data collection tool. In *Euro Working Group on Transportation (EWTG)*, pages 13–16. Citeseer, 2005. [page 67]

R. Krüger, H. Bosch, S. Koch, C. Müller, G. Reina, D. Thom, and T. Ertl. HIVEBEAT - A highly interactive visualization environment for broad-scale exploratory analysis and tracing: VAST 2012 Mini Challenge 1 award: Honorable mention for comprehensive visualization suite. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 277–278. IEEE Computer Society, 2012a. [page 124]

R. Krüger, S. Lohmann, D. Thom, H. Bosch, and T. Ertl. Using social media content in the visual analysis of movement data. In *2nd Workshop on Interactive Visual Text Analytics, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 1–3. IEEE Computer Society, 2012b. [pages 6 and 7]

R. Krüger, H. Bosch, D. Thom, E. Puettmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl. Prolix - Visual prediction analysis for box office success. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, 2013a. [page 7]

R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl. TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data. *Computer Graphics Forum*, 32(3):451–460, 2013b. [pages 6 and 121]

R. Krüger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 193–200. IEEE Computer Society, 2014a. [page 6]

R. Krüger, D. Thom, and T. Ertl. Semantic enrichment of movement behavior with Foursquare—a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 21(8):903–915, 2014b. [page 6]

R. Krüger, F. Heimerl, Q. Han, K. Kurzhals, S. Koch, and T. Ertl. Visual analysis of visitor behavior for indoor event management. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1148–1157. IEEE Computer Society, 2015a. [page 7]

R. Krüger, D. Herr, F. Haag, and T. Ertl. Inspector Gadget: Integrating data preprocessing and orchestration in the visual analysis loop. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 7–12. The Eurographics Association, 2015b. [page 6]

R. Krüger, S. Koch, and T. Ertl. SaccadeLenses: Interactive exploratory filtering of eye tracking trajectories. In *2nd Workshop on Eye Tracking and Visualization (ETVIS), co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, volume 2, pages 31–34. IEEE Computer Society, 2016a. [page 7]

R. Krüger, G. Sun, F. Beck, R. Liang, and T. Ertl. TravelDiff: Visual comparison analytics for massive movement patterns derived from Twitter. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 176–183. IEEE Computer Society, 2016b. [page 6]

C. Kumar, W. Heuten, and S. Boll. Geographical queries beyond conventional boundaries: Regional search and exploration. In *Workshop on Geographic Information Retrieval*, pages 84–85. ACM, 2013. [page 121]

O. D. Lampe and H. Hauser. Interactive visualization of streaming data with kernel density estimation. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 171–178, 2011. [pages 25 and 26]

P. Laube. *Computational movement analysis*. Springer, 2014. [pages 2, 3, and 168]

P. Laube, T. Dennis, P. Forer, and M. Walker. Movement beyond the snapshot–dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, 31(5): 481–501, 2007. [page 68]

K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 435–442. ACM, 2010. [page 46]

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966. [pages 120 and 122]

Z. Liu, B. Jiang, and J. Heer. immens: Real-time visual querying of big data. *Computer Graphics Forum*, 32(3):421–430, 2013. [pages 50, 55, 157, 158, and 160]

Z. Liu, H. Dev, M. Dontcheva, and M. Hoffman. Mining, pruning and visualizing frequent patterns for temporal event sequence analysis. In *The Event Event: Temporal and Sequential Event Analysis Workshop, colocated with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2016. [page 120]

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. [page 124]

S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Conference on Human-Computer Interaction (IFIP)*, pages 392–404. Springer, 2009. [pages 79 and 80]

J. Looser, R. Grasset, and M. Billinghurst. A 3d flexible and tangible magic lens in augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1–4. IEEE Computer Society, 2007. [page 96]

M. Lu, C. Lai, T. Ye, J. Liang, and X. Yuan. Visual analysis of route choice behaviour based on GPS trajectories. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–204, 2015. [page 97]

Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski. Integrating predictive analytics and social media. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 193–202. IEEE Computer Society, 2014. [page 7]

M. Luboschik, H. Schumann, and H. Cords. Particle-based labeling: Fast point-feature labeling without obscuring other visual features. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1237–1244, 2008. [page 105]

C. Ma, R. V. Kenyon, A. G. Forbes, T. Berger-Wolf, B. J. Slater, and D. A. Llano. Visualizing dynamic brain networks using an animated dual-representation. In *Eurographics Conference on Visualization*, pages 73–77. The Eurographics Association, 2015. [page 42]

A. M. MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 1995. [page 27]

A. M. MacEachren, M. Gahegan, W. Pike, I. Brewer, G. Cai, E. Lengerich, and F. Hardisty. Geovisualization for knowledge construction and decision support. *IEEE Computer Graphics and Applications*, 24(1):13–17, 2004. [page 14]

A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005. [pages 77 and 78]

A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2: Geotwitter analytics support for situational awareness. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 181–190. IEEE Computer Society, 2011. [page 41]

R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010. [page 25]

J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transaction on Graphics*, 5(2):110–141, 1986. [page 160]

L. Magnani. An abductive theory of scientific reasoning. *Semiotica*, 2005(153-1/4): 261–286, 2005. [page 18]

R. McCool. Rethinking the semantic web. part 1. *IEEE Internet Computing*, 9(6):88, 86–87, 2005. [page 66]

R. McCool. Rethinking the semantic web. part 2. *IEEE Internet Computing*, 10(1): 93–96, 2006. [page 66]

M. Miller, L. V. Kibiuk, D. Stuart, et al. *Brain Facts: a primer on the brain and nervous system*. The Society For Neuroscience, 2008. [page 11]

A. Minetti. The three modes of terrestrial locomotion. *Biomechanics and biology of movement*, pages 67–78, 2000. [page 46]

M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12): 2227–2236, 2013a. [page 120]

M. Monroe, R. Lan, J. Morales del Olmo, B. Shneiderman, C. Plaisant, and J. Millstein. The challenges of specifying intervals and absences in temporal queries: A graphical language approach. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 2349–2358. ACM, 2013b. [page 121]

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. [page 167]

A. Morris, A. Abdelmoty, B. El-Geresy, and C. Jones. A filter flow visual querying language and interface for spatial databases. *GeoInformatica*, 8(2):107–141, 2004. [page 121]

R. Netzel, M. Burch, and D. Weiskopf. Comparative eye tracking study on node-link visualizations of trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2221–2230, 2014. [page 26]

A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In *International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS)*, volume 14, pages 849–856, 2001. [page 125]

C. North and B. Shneiderman. Snap-together visualization: can users construct and operate coordinated visualizations? *International Journal of Human-Computer Studies*, 53(5):715–739, 2000. [pages 17, 43, and 161]

A. Panagiotidis, H. Bosch, S. Koch, and T. Ertl. Edgeanalyzer: Exploratory analysis through advanced edge interaction. In *Hawaii International Conference on System Sciences (HICSS)*, volume 44, pages 1–10. IEEE Computer Society, 2011. [page 96]

C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45 (4):42:1–42:32, 2013. [pages 63, 66, and 68]

E. S. Patterson, E. M. Roth, and D. D. Woods. Predicting vulnerabilities in computer-supported inferential analysis under data overload. *Cognition, Technology & Work*, 3 (4):224–237, 2001. [page 19]

C. S. Peirce. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press, 1974. [pages 17, 18, and 166]

A. Perer and F. Wang. Frequence: interactive mining and visualization of temporal frequent event sequences. In *International Conference on Intelligent User Interfaces (IUI)*, pages 153–162. ACM, 2014. [page 120]

A. Perer, F. Wang, and J. Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of biomedical informatics*, 56:369–378, 2015. [page 120]

A. Perer, B. C. Kwon, and J. Verma. The critical role of data mining for analyzing real-world event sequences. In *The Event Event: Temporal and Sequential Event Analysis Workshop, colocated with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2016. [page 120]

D. Pfoser, C. S. Jensen, Y. Theodoridis, et al. Novel approaches to the indexing of moving object trajectories. In *International Conference on Very Large Data Bases (VLDB)*, pages 395–406, 2000. [page 157]

W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009. [page 18]

P. Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007. [page 19]

P. Pirolli and S. Card. Information foraging. *Psychological review*, 106(4):643, 1999. [page 19]

P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *International Conference on Intelligence Analysis*, volume 5, pages 2–4. MITRE, 2005. [pages 17, 18, 19, 20, 31, and 37]

C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. LifeLines: Visualizing personal histories. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 221–227. ACM, 1996. [pages 120 and 121]

D. Preoţiuc-Pietro and T. Cohn. Mining user behaviours: a study of check-in patterns in location based social networks. In *Annual ACM Web Science Conference*, pages 306–315. ACM, 2013. [page 67]

A. J. Pretorius and J. J. van Wijk. What does the user want to see? What do the data want to be? *Information Visualization*, 8(3):153–166, 2009. [page 90]

V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *International Conference on Very Large Data Bases (VLDB)*, volume 1, pages 381–390. Morgan Kaufmann, 2001. [page 68]

R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 318–322. ACM, 1994. [page 90]

C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727, 2006. [page 39]

S. Reumers, F. Liu, D. Janssens, M. Cools, and G. Wets. Semantic annotation of GPS traces: Activity type inference. In *Annual Meeting of the Transportation Research Board*. Transportation Research Board of the National Academies, 2013. [page 67]

D. C. Richardson and R. Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29(6):1045–1060, 2005. [page 28]

J. C. Roberts. State of the art: Coordinated multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pages 61–71, July 2007. [pages 17, 27, and 161]

B. Rose. Nfl's lengthy list of requirements for super bowl host city leaked. Sports Illustrated, June 2014. [page 60]

M. Rotard, M. Giereth, and T. Ertl. Semantic lenses: Seamless augmentation of web pages with context information from implicit queries. *Computers & Graphics*, 31(3): 361 – 369, 2007. [page 96]

R. A. Ruddle, J. Bernard, T. May, H. Lücke-Tieke, and J. Kohlhammer. Methods and a research agenda for the evaluation of event sequence visualization techniques. In *IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis.* Leeds, 2016. [page 120]

A. Russell, P. Smart, D. Braines, and N. Shadbolt. NITELIGHT: A graphical tool for semantic query construction. In *International Conference on Posters and Demonstrations*, volume 543 of *CEUR-WS*, pages 110–111, 2008. [page 121]

D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Conference on Human Factors in Computing Systems (CHI)*, pages 269–276. ACM, 1993. [page 18]

D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. [pages 20 and 33]

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *International Conference on World Wide Web*, pages 851–860. ACM, 2010. [pages 39 and 41]

D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Symposium on Eye Tracking Research and Applications (ETRA)*, pages 71–78. ACM, 2000. [page 115]

R. Scheepens, N. Willems, H. van de Wetering, and J. J. Van Wijk. Interactive visualization of multivariate trajectory data with density maps. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 147–154. IEEE, 2011. [page 50]

R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 17–24. IEEE, 2014. [page 27]

R. Scheepens, C. Hurter, H. Van De Wetering, and J. J. Van Wijk. Visualization, selection, and analysis of traffic flows. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):379–388, 2016. [page 49]

A. Schipper, H. Fuhrmann, and R. von Hanxleden. Visual comparison of graphical models. In *IEEE International Conference on Engineering of Complex Computer Systems (ICECCS)*, pages 335–340. IEEE Computer Society, 2009. [page 42]

A. Schmidt, M. Beigl, and H.-W. Gellersen. There is more to context than location. *Computers & Graphics*, 23(6):893–901, 1999. [pages 29 and 30]

A. K. Schmitt, M. Danišík, E. Aydar, E. Şen, İ. Ulusoy, and O. M. Lovera. Identifying the volcanic eruption depicted in a neolithic painting at çatalhöyük, Central Anatolia, Turkey. *PloS one*, 9(1):e84711, 2014. [page 26]

J. Scholtz, M. A. Whiting, C. Plaisant, and G. Grinstein. A reflection on seven years of the VAST Challenge. In *BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, pages 13:1–13:8. ACM, 2012. [page 22]

I. Seifert. A pool of queries: Interactive multidimensional query visualization for information seeking in digital libraries. *Information Visualization*, 10(2):97–106, 2011. [page 121]

H. Senaratne, A. Bröring, T. Schreck, and D. Lehle. Moving on Twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories. In *ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 23–30. ACM, 2014. [pages 39 and 41]

B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6): 70–77, 1994. [pages 96 and 121]

B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages and Human-Centric Compucting (VL/HCC)*, pages 336–343. IEEE Computer Society, 1996. [pages 11, 16, 93, and 156]

B. Shneiderman. Inventing discovery tools: combining information visualization with data mining1. *Information visualization*, 1(1):5–12, 2002. [page 13]

J. Snow. *On the mode of communication of cholera.* John Churchill, 1855. [page 26]

A. Soylu, M. Giese, E. Jimenez-Ruiz, E. Kharlamov, D. Zheleznyakov, and I. Horrocks. OptiqueVQS: Towards an ontology-based visual query system for big data. In *International Conference on Management of Digital EcoSystems (MEDES)*, pages 119–126. ACM, 2013. [page 121]

S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008. [page 65]

M. Spenke, C. Beilken, and T. Berlage. FOCUS: The interactive table for product comparison and selection. In *Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 41–50. ACM, 1996. [page 90]

M. Spindler and R. Dachselt. Exploring information spaces by using tangible magic lenses in a tabletop environment. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pages 4771–4776. ACM, 2010. [page 96]

L. Spinsanti, F. Celli, and C. Renso. Where you stop is who you are: understanding people's activities by places visited. In *Workshop on Behaviour Monitoring and Interpretation (BMI)*, volume 5, 2010. [page 67]

A. Spoerri. InfoCrystal: A visual tool for information retrieval & management. In *International Conference Information and Knowledge Management (CIKM)*, pages 11–20. ACM, 1993. [page 96]

R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *International Conference on Extending Database Technology (EDBT)*, pages 1–17. Springer, 1996. [page 119]

M. Steptoe, R. Krüger, Y. Zhang, X. Liang, R. Garcia, S. Kadambi, W. Luo, T. Ertl, and R. Maciejewski. VAST Challenge 2015: Grand Challenge - award for outstanding comprehensive submission. In *VAST Challenge Workshop, co-located with the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 119–120. IEEE Computer Society, 2015. [page 7]

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1):24–54, 2010. [page 105]

D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 41–48, 2012. [pages 79, 80, and 105]

D. Thom, H. Bosch, R. Krüger, and T. Ertl. Using large scale aggregated knowledge for social media location discovery. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1464–1473. IEEE Computer Society, 2014. [pages 7 and 39]

D. Thom, R. Krüger, T. Ertl, U. Bechstedt, A. Platz, J. Zisgen, and B. Volland. Can Twitter really save your life? A case study of visual social media analytics for situation awareness. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 183–190. IEEE Computer Society, 2015. [page 7]

D. Thom, R. Krüger, and T. Ertl. Can Twitter save lives? A broad-scale study on visual social media analytics for public safety. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1816–1829, 2016. [page 7]

J. Thomas and J. Kielman. Challenges for visual analytics. *Information Visualization*, 8 (4):309–314, 2009. [page 95]

J. J. Thomas and K. A. Cook. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Laboratory (PNNL), 2005a. [pages 2 and 13]

J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005b. [pages 156 and 162]

P. Tobien, L. Lischke, M. Hirsch, R. Krüger, P. Lukowicz, and A. Schmidt. Engaging people to participate in data collection. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 209–212. ACM, 2016. [page 7]

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970. [page 26]

C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE Transactions Visualization and Computer Graphics*, 18(12):2565 –2574, 2012. [page 42]

T. Tremel. Advancement of a query visualization for event sequences. Supervisor: R. Krüger, F. Haag, T. Ertl. Bachelor Thesis. Universität Stuttgart, 2016.

E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, 1983. [pages 28 and 82]

B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002. [page 48]

L. Tweedie, B. Spence, D. Williams, and R. Bhogal. The attribute explorer. In *ACM Conference Companion on Human Factors in Computing Systems CHI)*, pages 435–436. ACM, 1994. [page 90]

T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011. [page 42]

T. von Landesberger, S. Bremm, N. Andrienko, G. Andrienko, and M. Tekušová. Visual analytics methods for categoric spatio-temporal data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192. IEEE, 2012. [page 28]

T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. MobilityGraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):11–20, 2016. [pages 39 and 41]

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007. [page 126]

Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering. Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159–2168, 2013. [pages 27 and 58]

C. Ware. *Information visualization: perception for design*. Elsevier, 2012. [page 11]

N. Willems, H. Van De Wetering, and J. J. Van Wijk. Visualization of vessel movements. In *Computer Graphics Forum*, volume 28, pages 959–966. Wiley Online Library, 2009. [page 25]

N. Willems, R. Scheepens, H. van de Wetering, and J. J. van Wijk. Visualization of vessel traffic. In *Situation Awareness with Systems of Systems*, pages 73–87. Springer, 2013. [pages 25 and 26]

C. Wolff. *Philosophia prima: sive ontologia*. prostat in Officina libraria Rengeriana, 1736. [page 65]

N. Wong, S. Carpendale, and S. Greenberg. Edgelens: an interactive method for managing edge congestion in graphs. In *IEEE Symp Information Visualization(InfoVis)*, pages 51–58, 2003. [page 96]

K. Wongsuphasawat, C. Plaisant, M. Taieb-Maimon, and B. Shneiderman. Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting with Computers*, 24(2):55–68, 2012. [pages 120 and 121]

J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of london's bicycle-hire scheme. *Cartographica*, 46(4):239–251, 2011. [page 58]

M. Wörner. *Visual analytics for production and transportation systems*. PhD thesis, Universität Stuttgart, Insitut für Visualisierung und Interaktive Systeme, 2014. [page 22]

M. Wörner and T. Ertl. Multi-layer distorted 1d navigation. In *International Conference of Information Visualization Theory and Applications (IVAPP)*, volume 2011, pages 198–203, 2011. [pages 96 and 103]

S. Wu, S. Otmane, G. Moreau, and M. Servières. Design of a visual query language for geographic information system on a touch screen. In *Human-Computer Interaction. Interaction Modalities and Techniques*, volume 8007 of *LNCS*, pages 530–539. Springer, 2013. [page 121]

Z. Yan, J. Macedo, C. Parent, and S. Spaccapietra. Hu2013. *Transactions in GIS*, 12 (s1):75–91, 2008. [page 65]

Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semitri: A framework for semantic annotation of heterogeneous trajectories. In *International Conference on Extending Database Technology*, EDBT/ICDT '11, pages 259–270. ACM, 2011. [page 66]

A. Yarbus. *Eye movements and vision. 1967.* Springer US, 1967. [page 98]

J. S. Yi, Y. a. Kang, and J. Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, Nov 2007. [page 16]

D. Young and B. Shneiderman. A graphical filter/flow representation of boolean queries: A prototype implementation and evaluation. *Journal of the American Society of Information Science*, 44(6):327–339, 1993. [page 97]

C. H. Yu. Abduction? deduction? induction? is there a logic of exploratory data analysis?. In *Annual Meeting of the American Educational Research Association*. ERIC, 1994. [pages 18 and 117]

M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2):31–60, 2001. [page 119]

W. Zeng, C. W. Fu, S. M. Arisona, S. Schubiger, R. Burkhard, and K. L. Ma. Visualizing the relationship between human mobility and points of interest. *IEEE Transactions on Intelligent Transportation Systems*, PP(99):1–14, 2017. [page 67]

E. Zgraggen, S. M. Drucker, D. Fisher, and R. DeLine. (s|qu)eries: Visual regular expressions for querying and exploring event sequences. In *Conference on Human Factors in Computing Systems (CHI)*, pages 2683–2692. ACM, 2015. [page 121]

J. Zhao, F. Chevalier, E. Pietriga, and R. Balakrishnan. Exploratory analysis of time-series with chronolenses. *IEEE Transactions Visualization and Computer Graphics*, 17(12):2422–2431, dec. 2011. [page 96]

Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *International Conference on World Wide Web*, pages 791–800. ACM, 2009. [page 72]

P. Zikopoulos, C. Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data.* McGraw-Hill Osborne Media, 2011. [pages 2 and 11]