

# **Spatio-temporal and Immersive Visual Analytics for Advanced Manufacturing**

Von der Graduate School of Excellence Advanced Manufacturing Engineering der Universität Stuttgart  
zur Erlangung der Würde eines  
Doktor-Ingenieurs (Dr.-Ing.)  
genehmigte Abhandlung

Vorgelegt von

**Dominik Alexander Herr**

aus Stuttgart

Hauptberichter: Prof. Dr. Thomas Ertl  
Mitberichter: Prof. Dr. Jörn Kohlhammer  
Tag der mündlichen Prüfung: 28. Mai 2019

Graduate School of Excellence Advanced Manufacturing Engineering  
der Universität Stuttgart

2019



---

# Acknowledgments

I would like to extend my sincerest thanks to my primary supervisor Thomas Ertl, who gave me the opportunity to become a PhD student at the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) and gave me guidance and support throughout this entire time. Further, I would like to thank my co-examiner Jörn Kohlhammer for reviewing my thesis and being part of my PhD defense committee. I also thank GSaME for the great time, funding, and valuable insights into the manufacturing and business domains, Professor Bernhard Mitschang for his guidance, and my colleagues and friends at GSaME, especially Jan Königsberger, Stefan Silcher, and Christian Weber.

I thank the co-authors who made this thesis possible: Michael Becher, Fabian Beck, Sören Brückmann, Qi Han, Robert Krüger, Kuno Kurzhals, Steffen Lohmann, Christoph Müller, Guido Reina, and Daniel Weiskopf. Further, my students Sebastian Grund, Jan Reinhardt, and Rafael Villanueva Ferrari provided great work that provided valuable insights for some of the papers in this thesis.

Special thanks to Michael Raschke, who convinced me to become a PhD student and Steffen Koch for the valuable discussions about my projects. I also want to extend my gratitude to Michael Wörner, who gave me guidance at the beginning of my studies and introduced me to Lena Wagner, whom I collaborated with for large parts of my time at GSaME and VIS. I thank my colleagues Qi Han, Robert Krüger, Sanae Mahtal, and Franziska Huth for bearing with me as their room mate and Kuno Kurzhals for collaborating with me with my later research projects and for his help with this dissertation's cover design.

In addition, I want to express my gratitude to all domain experts who contributed to my research with their advice and feedback, especially Lena Wagner, Abdullah Demir, Florian Schmoll, and Stefan Heitz. I thank Dominic Lutz from Gaugler & Lutz for his valuable feedback for some of my work and the companies AIT GmbH & Co. KG and iFakt GmbH for their collaboration.

Special thanks to the various after (or between) work groups that made the time as a PhD student even more fun: Tanja and Florian for the fun times at Maulwurf; Oli, Alex, Marcel, and Kuno as the core board game group; Gleb, Moataz, Kuno, and Valentin as the core Pokkez group.

Last but not least I want to thank my family, especially my parents, for their boundless supported throughout my entire time as a PhD student.



---

# Contents

<b>Acknowledgments</b>	iii
<b>Abstract</b>	ix
<b>Zusammenfassung</b>	xi
<b>1 Introduction</b>	1
1.1 Research Questions . . . . .	2
1.2 Structure and Contribution . . . . .	3
<b>2 Foundations</b>	7
2.1 Visual Analytics . . . . .	7
Concept . . . . .	8
Visualization . . . . .	9
Data Processing and Analysis . . . . .	10
Visual Analytics Application Example . . . . .	19
2.2 Advanced Manufacturing . . . . .	23
From Traditional to Advanced Manufacturing . . . . .	23
Product Lifecycle . . . . .	26
Overall Equipment Effectiveness . . . . .	26
Process Planning . . . . .	27
<b>3 Exploration of Concept Relationships</b>	29
3.1 Patents and their Classification System . . . . .	30
Patent Structure . . . . .	30
International Patent Classification System . . . . .	32
3.2 Related Work . . . . .	33
Dimension Reduction and Data Projection . . . . .	33
Word Cloud Visualization . . . . .	34
Hierarchical Aggregation and Visualization . . . . .	34
Patent and Scientific Literature Visualization . . . . .	36
3.3 Visual Exploration of Patent Relations using IPC Classes . . . . .	36
Patent Data Retrieval and Storage . . . . .	37
Data Preprocessing . . . . .	38

	IPC Cloud Visualizations . . . . .	39
	Example of Use . . . . .	43
	Discussion of Scalability . . . . .	44
3.4	Understanding Topic Relations Through Hierarchized Projection . . . . .	45
	Approach . . . . .	46
	Data Preprocessing . . . . .	46
	Recurring Steps of Analysis . . . . .	47
	Switch Between Levels of Detail . . . . .	50
	Explore Cluster Relationships . . . . .	53
	Use Cases . . . . .	56
	Evaluation . . . . .	58
	Task-based Results . . . . .	61
<b>4</b>	<b>Visual Analytics for Production Line Layout Planning</b>	<b>67</b>
4.1	Related Work . . . . .	68
	Layout Planning . . . . .	68
	Visual Support for Optimization Algorithms for Layout Planning . . . . .	69
	Virtual & Augmented Reality in Manufacturing . . . . .	70
4.2	Visual Analysis and Optimization of Worker Paths in U-Shaped Factory Layouts . . . . .	71
	Approach . . . . .	72
	Layout Inspection . . . . .	73
	Visual Optimization Guidance . . . . .	75
	Optimization with an Estimation of Distribution Algorithm . . . . .	76
	Application Scenario . . . . .	81
	Further Results . . . . .	83
4.3	Immersive Analysis of Production Line Simulations . . . . .	84
	Approach . . . . .	84
	Evaluation . . . . .	91
<b>5</b>	<b>Visual Event Analysis in Production Lines</b>	<b>95</b>
5.1	Production Domain Introduction . . . . .	96
	Factory Hierarchy and Event Data Structure . . . . .	96
	Specific Scenario of the Industry Partner . . . . .	97
5.2	Related Work . . . . .	98
	Visual Analysis for Event Relationships . . . . .	98
	Event Series Analysis . . . . .	100
	Visual Analytics in Manufacturing . . . . .	101
5.3	Visual Analysis for Spatio-Temporal Event Correlation in Production Lines . . . . .	102
	Approach . . . . .	103
	Evaluation . . . . .	112
	Findings . . . . .	112

Case Study . . . . .	114
Feedback Session . . . . .	117
5.4 Visual Analysis of Temporal Event Patterns . . . . .	119
Requirements . . . . .	120
Approach . . . . .	121
Selection and Optimization of Parameters . . . . .	126
Evaluation . . . . .	128
<b>6 Integration of Augmented Reality Monitoring and Visual Event Analysis</b>	<b>133</b>
6.1 Motivation . . . . .	133
6.2 Domain Problem Characterization . . . . .	134
6.3 Approach . . . . .	137
6.4 Evaluation . . . . .	146
<b>7 Conclusion and Outlook</b>	<b>149</b>
7.1 Summary of Contributions . . . . .	149
7.2 Discussion . . . . .	150
7.3 Outlook . . . . .	153
<b>Author's Work</b>	<b>157</b>
<b>Bibliography</b>	<b>159</b>





---

# Abstract

The increasing amount of digitally available information in the manufacturing domain is accompanied by a demand to use these data to increase the efficiency of a product's overall design, production, and maintenance steps. This idea, often understood as a part of *Industry 4.0*, requires the integration of information technologies into traditional manufacturing craftsmanship. Despite an increasing amount of automation in the production domain, human creativity is still essential when designing new products. Further, the cognitive ability of skilled workers to comprehend complex situations and solve issues by adapting solutions of similar problems makes them indispensable. Nowadays, customers demand highly customizable products. Therefore, modern factories need to be highly flexible regarding the lot size and adaptable regarding the produced goods, resulting in increasingly complex processes.

One of the major challenges in the manufacturing domain is to optimize the interplay of human expert knowledge and experience with data analysis algorithms. Human experts can quickly comprehend previously unknown patterns and transfer their knowledge and gained experience to solve new issues. Contrarily, data analysis algorithms can process tasks very efficiently at the cost of limited adaptability to handle new situations. Further, they usually lack a sense of semantics, which leads to a need to combine them with human world knowledge to assess the meaningfulness of such algorithms' results. The concept of *Visual Analytics* combines the advantages of the human's cognitive abilities and the processing power of computers. The data are visualized, allowing the users to understand and manipulate them interactively, while algorithms process the data according to the users' interaction.

In the manufacturing domain, a common way to describe the different states of a product from the idea throughout the realization until the product is disposed is the product lifecycle. This thesis presents approaches along the first three phases of the lifecycle: design, planning, and production. A challenge that all of the phases face is that it is necessary to be able to find, understand, and assess relations, for example between concepts, production line layouts, or events reported in a production line.

As all phases of the product lifecycle cover broad topics, this thesis focuses on supporting experts in understanding and comparing relations between important

aspects of the respective phases, such as concept relationships in the patent domain, as well as production line layouts, or relations of events reported in a production line. During the design phase, it is important to understand the relations of concepts, such as key concepts in patents. Hence, this thesis presents approaches that help domain experts to explore the relationship of such concepts visually. It first focuses on the support of analyzing patent relationships and then extends the presented approach to convey relations about arbitrary concepts, such as authors in scientific literature or keywords on websites. During the planning phase, it is important to discover and compare different possibilities to arrange production line components and additional stashes. In this field, the digitally available data is often insufficient to propose optimal layouts. Therefore, this thesis proposes approaches that help planning experts to design new layouts and optimize positions of machine tools and other components in existing production lines. In the production phase, supporting domain experts in understanding recurring issues and their relation is important to improve the overall efficiency of a production line. This thesis presents visual analytics approaches to help domain experts to understand the relation between events reported by machine tools and comprehend recurring error patterns that may indicate systematic issues during production.

Then, this thesis combines the insights and lessons learned from the previous approaches to propose a system that combines augmented reality with visual analysis to allow the monitoring and a situated analysis of machine events directly at the production line. The presented approach primarily focuses on the support of operators on the shop floor. At last, this thesis discusses a possible combination of the product lifecycle with knowledge generating models to communicate insights between the phases, e.g., to prevent issues that are caused from problematic design decisions in earlier phases. In summary, this thesis makes several fundamental contributions to advancing visual analytics techniques in the manufacturing domain by devising new interactive analysis techniques for concept and event relations and by combining them with augmented reality approaches enabling an immersive analysis to improve event handling during production.

---

# Zusammenfassung

Die zunehmende Menge an digital verfügbaren Informationen im Fertigungsbereich geht einher mit dem Bedarf, diese Daten zur Steigerung der Effizienz der gesamten Design-, Produktions- und Wartungsschritte eines Produkts zu nutzen. Diese Idee, die häufig als Teil von *Industrie 4.0* verstanden wird, erfordert die Integration von Informationstechnologien in das traditionelle Fertigungshandwerk. Trotz der zunehmenden Automatisierung in der Produktion ist die Kreativität des Menschen bei der Entwicklung neuer Produkte immer noch von entscheidender Bedeutung. Die kognitive Fähigkeit von Fachkräften, komplexe Situationen zu verstehen und Probleme durch Anpassung der Lösungen ähnlicher Probleme zu überwinden, macht sie zudem unverzichtbar. Heutzutage verlangen Kunden nach hochgradig anpassbaren Produkten, die in modernen Fabriken zu immer komplexeren Prozessen führen, da sie eine hohe Flexibilität in Bezug auf die Losgröße und Anpassungsfähigkeit an die produzierten Waren benötigen.

Eine der größten Herausforderungen im Fertigungsbereich ist die Optimierung des Zusammenspiels von Expertenwissen und Erfahrung mit Datenanalysealgorithmen. Experten können zuvor unbekannte Muster schnell nachvollziehen und ihr Wissen und ihre Erfahrungen nutzen, um Lösungen für neue Probleme zu erarbeiten. Im Gegensatz dazu können Datenanalysealgorithmen Aufgaben auf Kosten einer begrenzten Anpassungsfähigkeit an neue Situationen sehr effizient verarbeiten. Darüber hinaus fehlt ihnen in der Regel ein Sinn für Semantik, was dazu führt, dass sie mit menschlichem Weltwissen kombiniert werden müssen, um den Sinngehalt der Ergebnisse zu bemessen. Das Konzept der *visuellen Analytik* kombiniert die Vorteile der kognitiven Fähigkeiten des Menschen und der Rechenleistung von Computern. Dabei werden die Daten visualisiert und somit dem Benutzer ermöglicht, die Daten verstehen und damit zu interagieren, während Algorithmen die Daten entsprechend der Benutzerinteraktion verarbeiten.

Im Fertigungsbereich ist der Produktlebenszyklus eine gängige Methode, um die unterschiedlichen Zustände eines Produkts von der Idee über die Realisierung bis zur Entsorgung des Produkts zu beschreiben. Diese Arbeit stellt Ansätze entlang der ersten drei Phasen des Lebenszyklus vor: Design, Planung und Produktion. Eine Herausforderung, der alle Phasen gegenüberstehen, besteht darin, dass Zusammenhänge, z. B. zwischen Konzepten, Fertigungslinienlayouts

oder in einer Fertigungslinie gemeldeten Ereignissen, gefunden, verstanden und bewertet werden müssen.

Da alle Phasen des Produktlebenszyklus breite Themenbereiche abdecken, konzentriert sich diese Arbeit auf die Unterstützung von Experten beim Verständnis und Vergleich von Beziehungen zwischen wichtigen Aspekten der jeweiligen Phasen, wie z. B. Konzeptbeziehungen im Patentbereich, sowie Produktionslinienlayouts oder Beziehungen von Ereignissen, die in einer Produktionslinie gemeldet werden. Während der Entwurfsphase ist es wichtig, die Zusammenhänge von Konzepten zu verstehen, wie z. B. Schlüsselkonzepte in Patenten. Daher stellt diese Arbeit Ansätze vor, mit denen Domänenexperten die Beziehung solcher Konzepte visuell untersuchen können. Zunächst wird die Unterstützung bei der Analyse von Patentbeziehungen fokussiert. Anschließend wird der vorgestellte Ansatz erweitert, um Beziehungen über beliebige Konzepte zu vermitteln, wie z. B. Autoren in der wissenschaftlichen Literatur oder Schlüsselwörtern auf Webseiten. In der Planungsphase ist es wichtig, verschiedene Möglichkeiten zur Anordnung von Komponenten der Produktionslinie und zusätzlichen Lagern zu entdecken und zu vergleichen. In diesem Bereich reichen die digital verfügbaren Daten oft nicht aus, um optimale Layouts vorzuschlagen. Daher schlägt diese Arbeit Ansätze vor, die Planungsexperten helfen, neue Layouts zu entwerfen und die Positionen von Werkzeugmaschinen und anderen Komponenten in bestehenden Produktionslinien zu optimieren. In der Produktionsphase ist es wichtig, Domänenexperten dabei zu unterstützen, wiederkehrende Probleme und deren Beziehungen zu verstehen, um die Effizienz einer Produktionslinie zu verbessern. Diese Arbeit stellt visuelle Analytikansätze vor, die Domänenexperten helfen, die Zusammenhänge zwischen der von Werkzeugmaschinen gemeldeten Ereignisse zu verstehen und wiederkehrende Fehlermuster nachzuvollziehen, da diese auf systematische Probleme während der Produktion hinweisen können.

Schließlich kombiniert diese Arbeit die Erkenntnisse und Erfahrungen aus den bisherigen Ansätzen, um ein System vorzuschlagen, das Augmented Reality mit visueller Analytik kombiniert, um die Überwachung und eine situierte Analyse von Maschinenereignissen direkt an der Produktionslinie zu ermöglichen. Der vorgestellte Ansatz konzentriert sich in erster Linie auf die Unterstützung von Operatoren in der Fertigungshalle. Zuletzt wird in dieser Arbeit eine mögliche Kombination des Produktlebenszyklus mit wissenserzeugenden Modellen diskutiert, um Erkenntnisse zwischen den Phasen zu vermitteln, z. B. um Probleme zu vermeiden, die durch problematische Designentscheidungen in früheren Phasen verursacht werden. Zusammenfassend liefert diese Arbeit mehrere grundlegende Beiträge zur Weiterentwicklung visueller Analysetechniken im Fertigungsbereich, indem neue interaktive Analysetechniken für Konzept- und Ereignisbeziehungen entwickelt und mit Augmented-Reality-Ansätzen kombiniert werden, die eine immersive Analyse ermöglichen, um die Ereignisbehandlung während der Produktion zu verbessern.



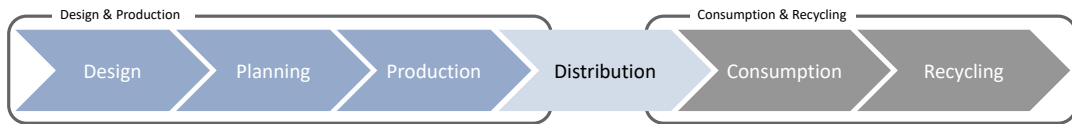


# Introduction

The increasing amount of digitization in the manufacturing domain regarding the quality and quantity of available data produces more information about everyday processes than ever before. Storing these data and making it accessible allows non-IT personnel to get information that helps with their tasks, e.g., getting real-time updates about the current issues of a production line or the quality of the produced goods. Despite the large amounts of data, most of the specialized knowledge and detailed insights of workers are still based on observations and experience gained during everyday tasks. Backing these personal insights with data and making knowledge available to colleagues and superiors is a challenge still persisting up to this day. This hampers the overall efficiency of the workers, as the knowledge has to be made by every employee separately.

Visualizing and allowing domain experts to analyze the available data enables them to support their observations with data and communicate their insights to colleagues or superiors. The combination of data analytics approaches to efficiently process the available data and visualization to present and interact with these data allows data analysis that would otherwise not be possible. The concept of combining data analytics and interactive visualization is called *visual analytics (VA)*.

Every product first needs to be designed and produced before it can be consumed and disposed of. This *product lifecycle* (see Figure 1.1a) is commonly used in the production domain and splits the lifetime of a product into several phases. Westkämper et al. [194] separate the phases into two groups: first, the product is designed and its production is planned before it is produced. Second, the finished product is shipped to consumers, where it is used and then disposed or recycled. This thesis focuses on the first part of the product lifecycle and proposes visual analytics approaches to support domain experts in understanding and comparing relations between important aspects during the design, planning, and production phase (see Figure 1.1b).



(a) General Product Lifecycle including categorization proposed by Westkämper et al. [194].



(b) First part of the Product Lifecycle, which this thesis focuses on.

**Figure 1.1:** This thesis presents approaches to support processes along the product lifecycle (a), specifically focusing on the first part (design and production) (b).

The remainder of this chapter first presents the research questions that this thesis aims to answer (Section 1.1). Then, the structure of this thesis is detailed and my contributions to the presented works are explicated (Section 1.2).

### 1.1 Research Questions

The overarching goal of this thesis is to improve the overall efficiency of processes along the product lifecycle. Thus, the central research question of this thesis is:

Overarching Research Question: **How can visual analytics be applied to support domain experts during the first part of the product lifecycle?**

Many optimizations in the production domain are conducted within the boundaries of the individual phases of the product lifecycle. The primary question is, how visual analytics can help with the tasks of these phases.

Answering this general question is a wide and challenging task. Therefore, this thesis breaks the overarching research question down to research questions that are focusing on the specific phases of the product lifecycle:



**Research Question 1: How can visualization help to understand the relation of relevant topics when designing a new product?**

Novel designs of products are often complex and may necessitate further inventions, such as new production processes. Having knowledge about similar topics or technologies in other fields may help to solve challenges and prevent legal issues due to patent infringements at a later point in time.

**Research Question 2: How can visual analytics support layout planning of factories and production lines?**

Planning and adapting factory or production line layouts is a challenging task that heavily relies on planning software and human experience. Therefore, integrating software and human knowledge through visual analytics has the potential to improve current processes.

**Research Question 3: How to support domain experts during the analysis of event data to understand issues in a production line?**

Keeping a production line operable is already a challenging task by itself. However, the exploratory visual analysis of event data reported by machinery to understand correlations and temporal patterns may help to understand production issues.

## 1.2 Structure and Contribution

This section outlines the remainder of this thesis and gives an overview of each chapter's content. I was first author of the majority of the publications that are presented in these chapters. The chapter overview also details with whom I collaborated during each contribution and which parts I contributed.

**Chapter 2 - Foundations** This chapter introduces the concepts and techniques that are relevant during the remainder of the thesis. The first part presents the general concept of visual analytics (VA) and the necessary fundamentals of visualization, data preprocessing, and analysis algorithms. The first part concludes with an application example for visual analytics through an approach that was developed in the context of the VAST challenge 2014. I contributed the data preprocessing views and helped with the analysis of the provided data

in this approach. In the second part, this chapter gives a brief summary of the development from traditional to advanced manufacturing and introduces important concepts from the production domain, of which the product lifecycle is of great importance, as the thesis structure follows along its different phases.

**Chapter 3 - Exploration of Concept Relationships** This chapter introduces approaches that help experts to understand topic relationships that may be of relevance during the design phase. It first focuses on the support of the analysis of patents based on their classification relationships. Then, this concept is extended to allow the analysis of the relations of concepts of documents in general, such as keywords on websites or in scientific literature. I was the first author of all papers in this chapter and collaborated with my colleagues Qi Han and Steffen Lohmann during the design and implementation process of the presented approaches. They provided parts of the data preprocessing and their experience in the fields of natural language processing and word cloud visualization.

**Chapter 4 - Visual Analytics for Production Line Layout Planning** This chapter presents approaches that support domain experts during the layout planning phase. Specifically, it first presents an extension of the desktop-based visual analysis approach to simulate modular production lines by my colleague Michael Wörner to be usable on an augmented reality device. This way, it is possible to compare different layout alternatives with each other on-site. The publications of this approach are based on the results of the master theses of Jan Reinhardt, who investigated how to transfer the aforementioned simulator to an augmented reality application, and Rafael Villanueva Ferrari, who complemented the simulator with a visualization to support the analysis of critical paths in the simulated production line. In addition to defining the research question, I extended the combined theses with a visual encoding to present the differences between two simulated layouts. My colleagues Guido Reina and Robert Krüger contributed their experience of simulation and geographic visualization to make the papers possible. The second half of the chapter focuses on the optimization of worker paths by rearranging movable components in a production line. The presented approach supports planners in choosing, which layout elements to optimize and where to move these parts to. This work is based on a prototype that was developed in the course of the bachelor thesis of Sebastian Grund, who also contributed the implementation of the pathfinding algorithm and parts of the graphical user interface of the final approach.

**Chapter 5 - Visual Event Analysis in Production Lines** The approaches presented in this chapter focus on supporting technical management staff to analyze the relation of events reported by machinery in a production line. Both

approaches were developed in cooperation with an industry partner, who explained the industrial relevance of such approaches and provided the used data. Further, the approaches were evaluated by domain experts of the industry partner. The first approach focuses on the analysis of event correlations and was developed together with my colleague Kuno Kurzhals who helped in designing the concept of the approach. The second approach helps experts to extract and analyze temporal patterns of the events' occurrence and was collaborative work with Fabian Beck, who helped to design the concept and provided his expertise in temporal data series visualization.

**Chapter 6 - Augmented Reality Monitoring and Visual Event Analysis** This chapter first combines the idea of the previously presented approach of using augmented reality to complement the visual analysis with the analysis of event data during production. Specifically, the event data from Chapter 5 is used as a basis for a system that allows operators to monitor and analyze data at the shop floor of a production line. This work was a collaboration with my co-authors Michael Becher, Kuno Kurzhals, Christoph Müller, Guido Reina, and Daniel Weiskopf. I contributed the overall approach of combining augmented reality and a tablet-based visual analysis, the analysis components, and the data preprocessing in this project. Further, this project was co-authored by Lena Wagner from Robert Bosch GmbH, who contributed her domain knowledge and the industry related questions.

**Chapter 7 - Conclusion & Outlook** The last chapter of this thesis first summarizes the contributions presented in this thesis. Then, the results of the approaches are discussed with regard to the stated research questions. At last, the thesis concludes with an outlook of open challenges, such as the visual analysis of the quality of the available data, collaborative analysis, and the combination of visual analysis and prediction to allow for predictive maintenance.



## Foundations

This thesis presents visual analytics approaches specifically designed for the advanced manufacturing domain. This chapter provides the necessary foundations by introducing the general concept of visual analytics (Section 2.1.1), an introduction to visualization (Section 2.1.2), and data processing and analysis algorithms that are used in the course of this thesis (Section 2.1.3). The introduction to visual analytics ends with a brief application example that shows how visual analytics can be applied to analyze previously unknown datasets in an artificial scenario (Section 2.1.4). In the second part of this chapter, the advanced manufacturing domain and concepts that this thesis makes use of, such as the product lifecycle, are introduced (Section 2.2).

This chapter is partly based on the following publication:

- R. Krüger, D. Herr, F. Haag and T. Ertl. “Inspector-Gadget: Integrating Data Preprocessing and Orchestration in the Visual Analysis Loop”. In: *Proceedings of the EuroVis Workshop on Visual Analytics*. EuroVA. The Eurographics Association, 2015 [9]

### 2.1 Visual Analytics

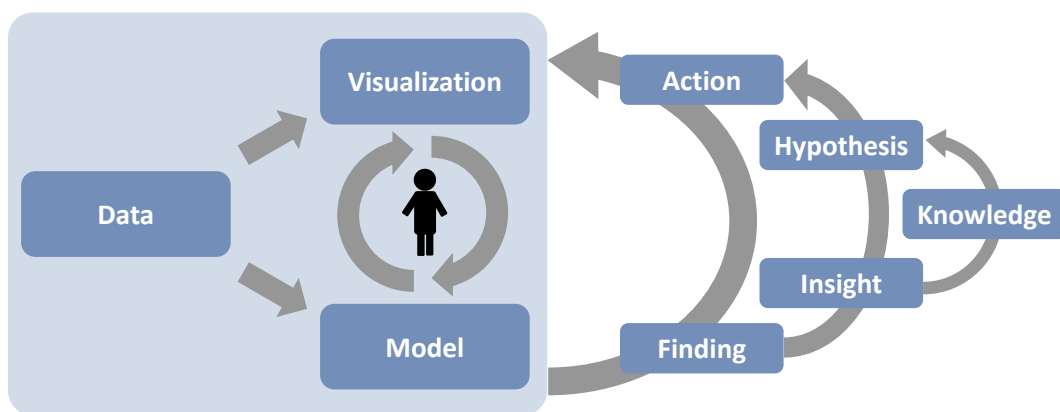
Historically, the field of visual analytics grew in importance when James J. Thomas and Kristin A. Cook published the book “Illuminating the Path - The Research and Development Agenda for Visual Analytics” as a response to the terrorist attacks on September 11, 2001, in the USA. They defined visual analytics (VA) as “the science of analytical reasoning facilitated by interactive visual interfaces” [181, p. 4]. Generally, VA combines visualization with data processing and analysis algorithms to make use of domain experts’ knowledge and experience and the computer’s ability to quickly process a large amount of

data. The following section first introduces visual analytics on a conceptual level. Afterwards, the core components of visualization and concrete data processing and analysis algorithms, which are used in this thesis, are presented. Finally, an application example for visual analytics shows how VA can be applied to help forensic analysts in getting insights about initially unknown datasets.

### 2.1.1 Concept

Visual analytics concepts can generally be viewed from two angles: On the one hand, VA intends to present workflow concepts that enable data processing that intertwines automatic data processing with human-readable visualizations, much like the definition by Thomas and Cook [181]. On the other hand, VA concepts intend to enable humans to extract insights from the data to aid them in their decision-making. Pirolli and Card [154] presented a prominent attempt for such a sensemaking process. They use the analogy of the shoebox system used by law enforcement and intelligence agencies to organize the available evidence during their investigations as a workflow to gain insights from the available data.

Sacha et al. [162] combine these two concepts to a knowledge generation model that merges both of the presented points of view in one model (see Figure 2.1). A computer processes the available data and generates data models and visualizations that can be inspected by human experts who want to analyze the data. The analysis can result in different stages of knowledge about the data. While the analysis is performed, experts can extract findings from the data. If the findings are already known or are of no relevance, the analysis continues. In

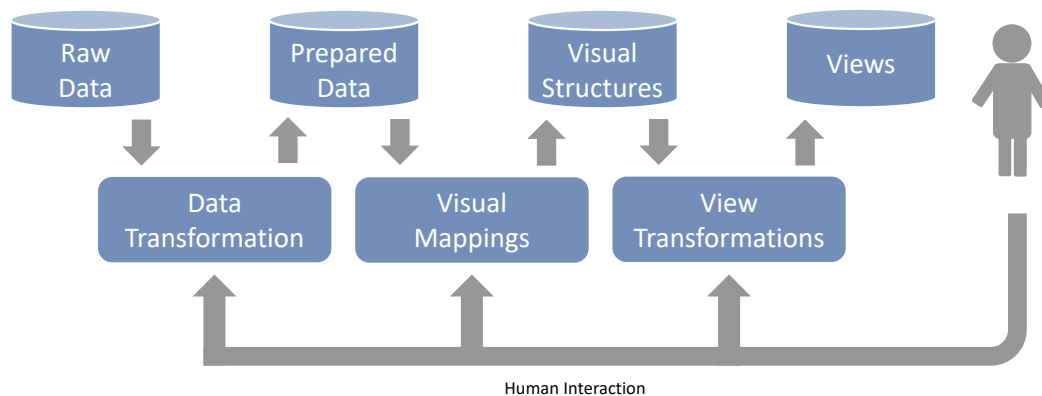


**Figure 2.1:** Example for a model to extract knowledge from data through the combination of visualizations and data analysis (cmp. Sacha et al. [162]). The analysis can result in different stages of understandings of the data: findings that are relevant become insights. If the hypotheses derived from the gained insights can be verified, they are seen as knowledge, otherwise the analysis continues.

case the findings were relevant, they become insights that need to be verified. This can be done by formulating a hypothesis about the insight and continuing the analysis with regard to the hypothesis. If the hypothesis is supported by the data, the insight becomes knowledge, that may also be transferable by the human analyst to other data sets or scenarios.

### 2.1.2 Visualization

Raw data are usually structured so that computers can efficiently process them, making them difficult to read and understand for humans. Card defined *visualization* as “the use of computer-supported, interactive, visual representations of data to amplify cognition” [42, p. 6]. It aims to transform data into a visual representation that is understandable for humans, which helps them to make sense of the data. As an extension to general visualization, *information visualization* is defined as the visualization of abstract data. In this context, abstract data describes data that does not contain spatial information that can be visualized directly. Card et al. defined an information visualization reference model [42, p. 17] that explains how raw data needs to be manipulated to transform it into human-readable views (see Figure 2.2). The model first transforms raw data to extract relevant data aspects that shall be presented to human users. Then, the prepared data are mapped to visual structures, which are arranged through view transformations to obtain the final views that are displayed. This workflow can be manipulated by the users through interaction at any stage where the data are manipulated.



**Figure 2.2:** The visualization reference model as proposed by Card et al. [42, p. 17]. The raw data is first transformed to create the data that shall be presented to the users. Then, the data is mapped to visual structures, which are then arranged through view transformations to be displayed in views.

### 2.1.3 Data Processing and Analysis

Nowadays, the amount of available data is often too large to process and analyze manually. Therefore, data processing and analysis algorithms are necessary to process, filter, and combine the data so that they can be presented adequately. The following presents algorithms and metrics used in the course of this thesis.

#### Similarity Metrics for High-Dimensional Data

Data is often complex and has a multitude of dimensions. An important step to get an understanding of the relation between the data points is to have a notion of which points are similar or different, and for what reason. There are multiple ways to measure similarity or dissimilarity of high-dimensional data points. The ones that are of relevance in this thesis are presented in the following.

**Euclidean distance** The arguably most intuitive way of measuring similarity is to calculate the Euclidean distance between points. For data with  $n$  dimensions, the distance of two points  $\vec{x}$  and  $\vec{y}$  is calculated as

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=0}^{n-1} (y_i - x_i)^2},$$

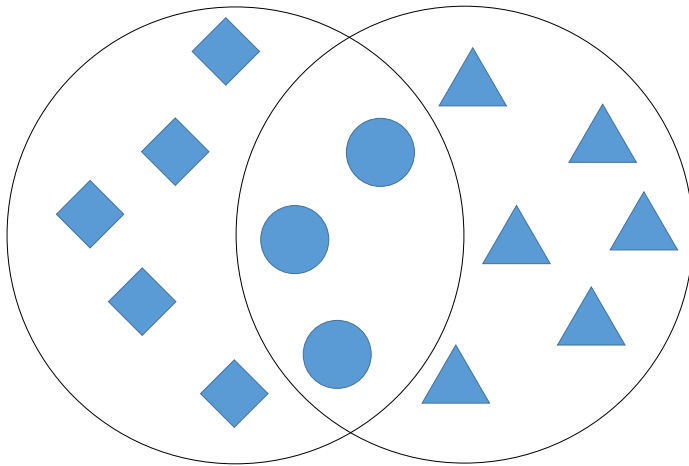
where  $x_i$  and  $y_i$  indicate the value of the point's  $i$ -th dimension. Although this measure is easy to understand, it has drawbacks. One is that the influence of each dimension on the results depends on the scale of its values. For example, assume a data point has two dimensions. The first dimension's values range from 0 and 100, whereas the second dimension's range from 0 to 10,000. When measuring the distance between two such data points, even large differences in the first dimension are outweighed by small changes in the second dimension.

**Cosine similarity** Unlike the Euclidean distance, the cosine distance is invariant to the ranges of a data point's dimensions. It considers every data point as a vector in a high-dimensional space and then calculates the angle between two points  $\vec{x}$  and  $\vec{y}$  using the cosine function:

$$sim(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \cos(\vec{x}, \vec{y})$$

As the cosine similarity implicitly discards the ranges of the individual dimensions, it solves the dimension range issue. However, both the cosine similarity and the Euclidean distance only work if the difference between the compared dimensions can be calculated. This is the case only for ratio scales. For all other data types, such as categorical data, these measures are not applicable.





**Figure 2.3:** Originally, the Jaccard index calculates the ratio of overlapping elements in two sets and their total number of distinct elements.

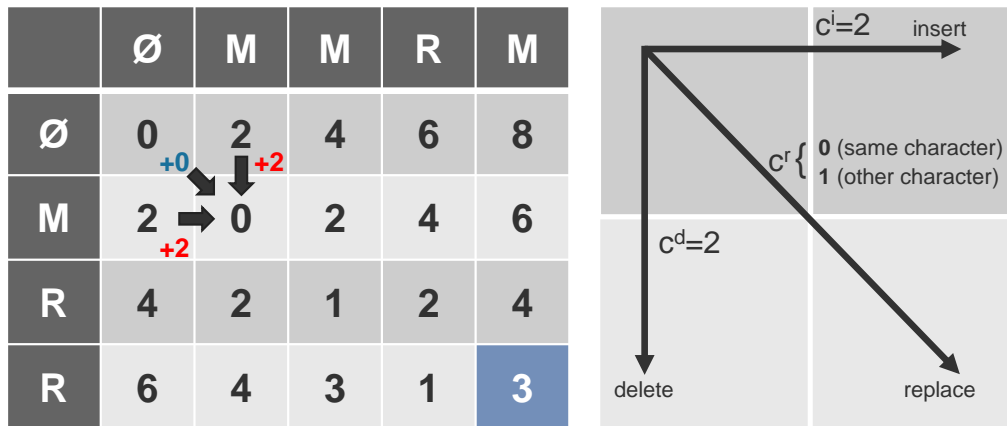
**Set overlap / Jaccard index** The Jaccard index takes a different approach compared to the previous measures. Originally, it was designed to measure the similarity of two given sets based on their overlap (see Figure 2.3). However, this approach can be adapted to compare high-dimensional data points with each other, by interpreting each dimension of the data point as an element in a set. To calculate the similarity of two data points, the overlap of both sets is calculated and then the value is divided by the total number of distinct elements in both data points [132, pg. 61]. The Jaccard index of two data points  $A$  and  $B$  can be defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where  $A = \{x_i | i \in [0..n - 1]; n = \# \text{ of dimensions}\}$ ;  $B$  is defined analogous to  $A$ . Although this measure only returns *exact* matches, it can cope with any data.

### String Comparison

When coping with text data, one piece of information that may be of interest is the (dis-)similarity of two given strings. An often used measure is the cost to transfer one string into another. The *Levenshtein distance* is an often used algorithm for this task. It transfers the strings' characters with the three operations *insert* ( $c^i$ ), *delete* ( $c^d$ ), and *replace* ( $c^r$ ). The cost is calculated using a table with  $(N + 1) \times (M + 1)$  cells, where  $N$  and  $M$  correspond to the size of the strings (see Figure 2.4, left). Both strings are extended with a leading empty character. The columns of the table correspond to the first string's characters, the rows to the second string's. Each cell  $c_{i,j}$  encodes the cost to transfer the first  $i$  characters of the first string to the first  $j$  characters of the second string. The cell  $(0,0)$ , which corresponds to the transformation of an empty string into an empty string, is initialized with no cost. The cost of all other cells is the smallest value of the



**Figure 2.4:** Example of the table that encodes the cost to transfer the string *MRR* to the string *MMRM*. The numbers beside the arrows indicate the transfer cost and the green cell represents the total transfer cost.

neighboring cell (top, left, top-left) plus their corresponding operation cost (see Figure 2.4, right):

$$C(i, j) = \min \begin{cases} C(i-1, j) + c^i \\ C(i, j-1) + c^d \\ C(i-1, j-1) + c^r \end{cases}$$

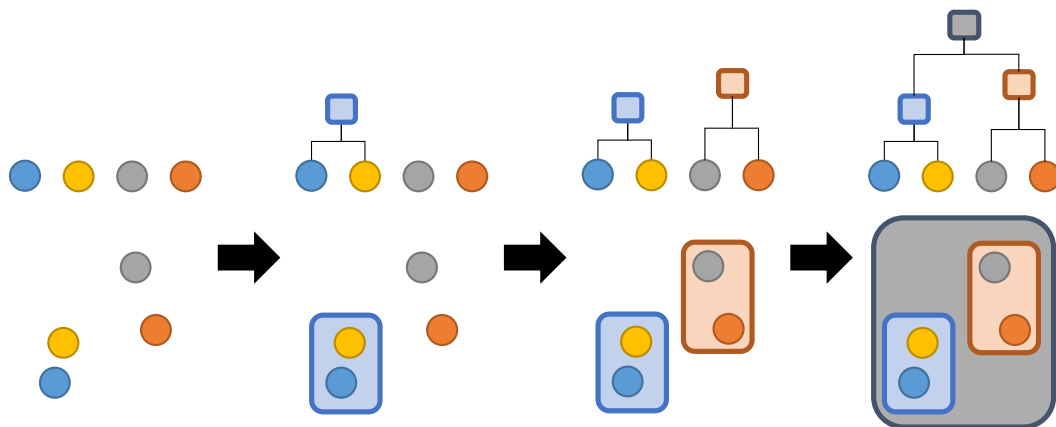
If the *replace* operation would replace a character with itself, the cost of the operation is reduced to zero. The costs for the operations are arbitrary, but common costs are either  $c^i = c^d = 2$  and  $c^r = 1$  or  $c^i = c^d = c^r = 1$ . The cost of the total string transformation is the value at the bottom right of the table.

After all values are calculated, it is possible to reconstruct the order of the necessary operations to transform the strings.

## Clustering

Representing similar data points as distinct groups is helpful to get a better understanding and overview of the data. In case the data is not labeled, such groups, also called clusters, can still be approximated in case a similarity metric between the data points can be defined. Two kinds of clustering algorithms can be distinguished: flat and structured. Flat clustering algorithms usually assign data points to one specific cluster and no relation between the clusters is assumed. Structured clustering algorithms introduce a tree-like structure, which enables clustering at varying levels of detail and therefore implicitly describe a relationship between the clusters.

The hierarchical agglomerative clustering algorithm (HAC) belongs to the structured clustering algorithms. Initially, every data point is assigned to a



**Figure 2.5:** The hierarchical agglomerative clustering algorithm first assigns every element to a separate cluster and then merges the clusters pairwise based on a previously specified linkage criterion.

separate cluster. Afterwards, the two most similar clusters are merged repeatedly until all data points are part of one cluster (see Figure 2.5). This procedure results in a binary tree of clusters where the leaves represent the individual data points and the nodes represent clusters. One advantage of this procedure is that the clustering granularity can be defined on an arbitrary level of detail. Aside from a pairwise similarity metric between the individual data points, it is necessary to define how the similarity between the clusters is calculated. There are various methods for defining the similarity of clusters:

**Single-linkage** defines the distance of two clusters as the least distance between any elements of the clusters. This linkage criterion is easy to understand, but its greedy approach makes its semantic meaning hard to understand [88, p. 525], as the elements are not compared to the elements that represent the core of the cluster, but to its outliers and elements at the border.

**Average-linkage** calculates the average similarity of each element in one cluster to all elements in the other cluster. The clusters' similarity is the average of the individual elements' average similarity. This linkage does not overemphasize outliers like single-linkage, but it can still be impacted by outliers and it is not possible to pick a cluster representative based on the metric.

**Medoid-linkage** first calculates the element with the least total distance to all other elements within the cluster (called *medoid*) as its representative and then defines the similarity between cluster as the similarity between the clusters' medoids. A medoid is the high-dimensional counterpart of the median in the one-dimensional space. This metric is a better way to describe a cluster's content compared to the elements used by single-linkage

for multiple reasons. For example, it implicitly compensates for a high variance and is more robust to outliers within the cluster [132, p. 392, 398].

One drawback of using medoid-linkage is the lack of monotonicity regarding the similarity of the merged clusters. Further, if two clusters are merged, the new cluster's medoid usually differs from the medoids of the merged clusters, requiring it to be recomputed as soon as the cluster changes.

Raschke et al. [10] use HAC to cluster study participants of an eye tracking study based on their scan paths. The cluster of each participant depends on the order in which they look at areas of interest and then represent each cluster by the participant that is in total the most similar to all other cluster members, which equals a medoid. These cluster representatives are then visualized and provide insights about common eye movement patterns during tasks, such as comparing bars in a bar chart and reading the highest value.

### **Dimension Reduction / Projection**

Projection techniques map data from a high- to a low-dimensional space. As this process usually includes information loss, most techniques focus on preserving structures that exist in the high-dimensional space also in the low-dimensional space. However, depending on the technique used to determine the mapping from the high- to the low-dimensional space, different characteristics of the data are preserved. For example, Principal Component Analysis (PCA) [152] uses the principal components of the high-dimensional data points to determine which dimensions vary the most between the data points and are therefore most promising to represent structure in the low-dimensional space. However, this way of separating data points does not represent the similarity of groups of data points in the low-dimensional space. Other techniques, such as Multidimensional Scaling (MDS) [114], interpret the high-dimensional distance between the data points as forces and the entire high-dimensional space as a mass-spring system. Depending on the function that maps the distance to force, such a system can build clusters of similar data points and push dissimilar points apart. However, like most approaches that are based on spatial differences in the high-dimensional space, it suffers from an ambiguousness that arises in sparse, high-dimensional spaces. This is one of the aspects of a phenomenon called *the curse of dimensionality*, which describes that in a sparse, high-dimensional space “the distance to the nearest data point approaches the distance to the farthest data point”, which makes the distance less meaningful [29].

Instead of using spatial distances between the data points, van der Maaten and Hinton's t-Distributed Stochastic Neighborhood Embedding (t-SNE) [187] uses pairwise probabilities that two data points are related in the high-dimensional space. If the data does not provide this information already, they define the

probability  $p_{i|j}$ , which describes the probability that a point  $x_i$  relates to a point  $x_j$  based on a Gaussian distribution. The probability that a point relates to itself is set to zero. One issue of the Gaussian model is that the needed variance depends on the sparsity of the analyzed data point's neighborhood. If the variance is low and the data density is sparse, then no neighbors will be included. On the contrary, a high variance with a dense data space will lead to a large neighborhood. To compensate for this, t-SNE uses a *perplexity*, which semantically corresponds to a point's neighborhood in the high-dimensional space that is invariant to the spatial distance of the points. However, because the neighborhood of the points  $x_i$  and  $x_j$  is different and due to the Gaussian model,  $p_{j|i} \neq p_{i|j}$ . This has several drawbacks during the dimension reduction, such as a high influence of outliers (see van der Maaten and Hinton [187] for details). To alleviate these drawbacks, the  $n \times n$  probability matrix is made symmetric:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

Once the matrix is set up, t-SNE iteratively minimizes the projection cost  $C$  of the cost function

$$\frac{\delta C}{\delta y_i} = 4 \cdot \sum_j \frac{(p_{ij} - q_{ij})(y_i - y_j)}{1 + \|y_i - y_j\|^2},$$

where  $y_i$  describes the low-dimensional representation of the data point that is currently mapped and  $y_j$  describes the other points. The cost function minimizes the Kullback-Leibler divergence [117] that uses a t-distribution to describe the pairwise distance between the high-dimensional relation  $p_{ij}$  and the corresponding low-dimensional relation  $q_{ij}$  of two data points.

### Pattern Detection and Extraction for Spatio-Temporal Data

When analyzing temporal data series, multiple aspects that could be of interest, for example, any long-term trends, seasonal behaviors, or outliers. Seasonal Trend Decomposition using Local Polynomial Regression (STL) [48] aims to explain time series by decomposing them into a trend, seasonal, and remainder component. A time series  $Y_v$  is defined as

$$Y_v = T_v + S_v + R_v,$$

where  $T_v$  is the trend series,  $S_v$  is the seasonal series, and  $R_v$  is the remainder that represents the difference between the trend and seasonal series to the actual data series. To extract these components, STL uses a nested loop approach. The inner loop extracts the seasonal and trend series, whereas multiple passes iteratively refine the results. The outer loop makes the resulting series more

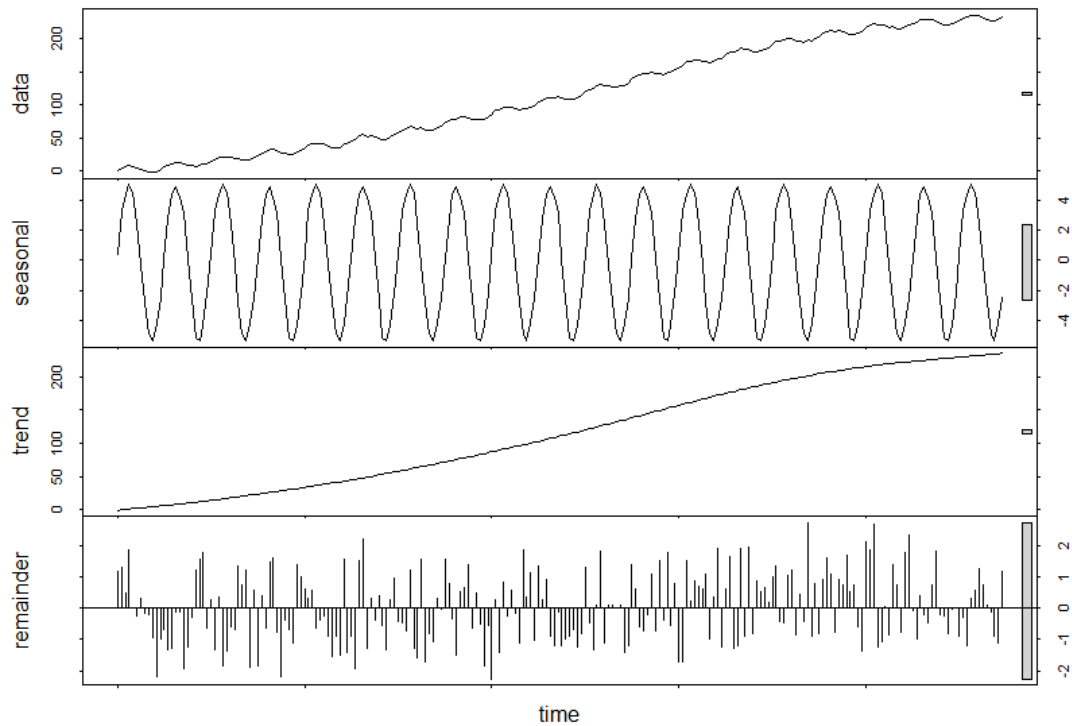
robust towards outliers by downweighting their impact during the extraction of the seasonal and trend series in future iterations of the inner loop. The following gives a brief overview of how the seasonal and trend components are extracted in the inner loop. The original work by Cleveland [48] provides more details for the inner loop and the mechanics of the outer loop. The inner loop comprises six steps to extract the trend and seasonal series:

- ① **Detrending:** The trend series is subtracted from the original series. This step can be skipped during the first pass, as there is not yet a trend series available. Alternatively, a predefined trend series can be provided.
- ② **Cycle-subseries smoothing:** The remaining data series is first split into  $n_p$  *cycle-subseries*, where  $n_p$  defines the length of the season that should be extracted. For example, if a series has a data point for each day and weekly patterns are of interest,  $n_p$  would be seven. After the cycle-subseries are extracted, each of the subseries is smoothed using a local polynomial regression (LOESS) smoother within a neighborhood of  $n_s$  elements. Low values for  $n_s$  result in seasons that may also include noise, whereas high values result in homogeneous seasons that may not represent changes over time. After the cycle-subseries smoothing, the subseries are recombined to a preliminary seasonal series  $C_v$ .
- ③ **Low-pass filtering of smoothed cycle-subseries:** A low-pass filter is applied to  $C_v$  to extract trend data that may unintentionally be included in the cycle-subseries.
- ④ **Detrending of smoothed cycle-subseries:** The previously extracted trend data is subtracted from the preliminary seasonal series, resulting in the seasonal series  $S_v$ .
- ⑤ **Deseasonalizing:** The extracted seasonal series is removed from the original time series.
- ⑥ **Trend smoothing:** The remaining deseasonalized series is smoothed using another low-pass filter. The resulting series is used as the trend series  $T_v$ .

Figure 2.6 shows an example in which an artificial time series is decomposed into an almost linear trend component, a cyclic seasonal component, and a uniformly distributed remainder (noise) component.

## Evolutionary Algorithms

Evolutionary algorithms (EA) are a class of optimization algorithms used to minimize a black-box objective function. Figure 2.7 illustrates the process EAs

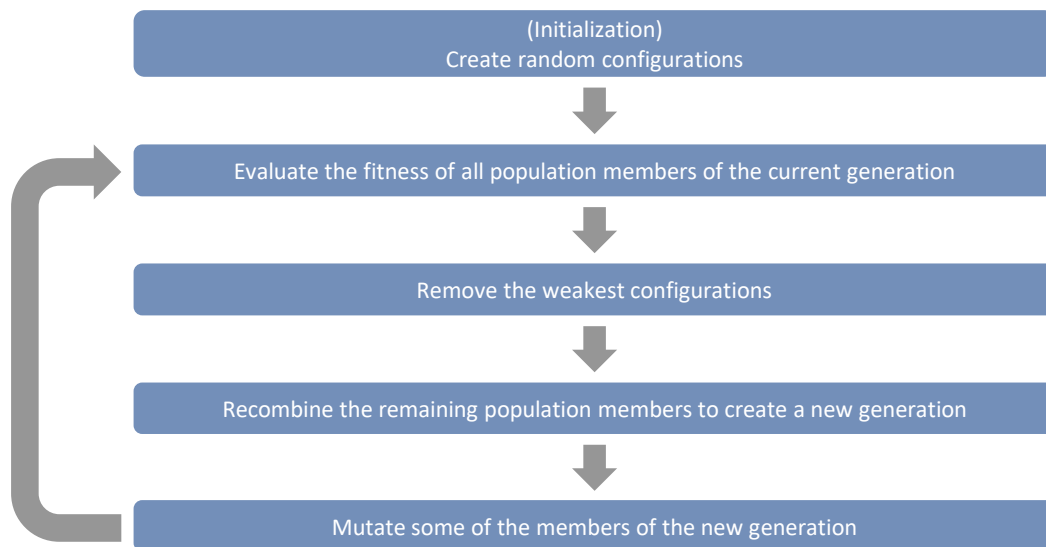


**Figure 2.6:** Example that decomposes an artificial time series into a trend, seasonal, and remainder component.

follow to imitate the natural evolution. First, a set of function input configurations is generated. The whole set is called a *population* and its configurations are called *members*. Every population belongs to a *generation*, which corresponds to the current iteration of the process. Upon their creation, the members' performance is evaluated using the black-box function. Based on the results some of the members are used to create a new generation of configuration by recombining the members' configuration values and mutating some of the results. This process repeats until a satisfactory result is achieved.

Algorithms that derive from evolutionary algorithms mostly differ in the way they pick the members and how they are recombined. Natural evolution recombines the genetic code of the parents' genomes to build a child's genome. The subclass closest to this behavior are called *genetic algorithms* and are the most widely known subclass of evolutionary algorithms. The following presents other subclasses that are used in the remainder of this thesis.

**Differential Evolution Algorithms** Differential evolution algorithms (DEA) behave mostly like genetic algorithms. However, differential algorithms do not take binary decisions when recombining the parents' attributes but interpolate between them. For example, assume two population members  $A$  and  $B$  comprising one scalar attribute.  $A$  has the attribute value 0.0 and  $B$  has the attribute value



**Figure 2.7:** Generic step-by-step procedure of an evolutionary algorithm.

1.0. In case of a genetic algorithm, the offspring's attribute could either be 0.0 or 1.0. In a differential evolution algorithm, the offspring's attribute could be any value between 0.0 and 1.0.

DEAs implicitly assume that the interpolated value between two well-performing values will also perform well. Unlike other optimization algorithms, like hill climbing, they do not require the function to be differentiable, which means DEAs can optimize a wider range of black-box functions.

**Estimation of Distribution Algorithms** Estimation of distribution algorithms (EDA) [89] differ from other evolutionary algorithms as they do not directly take a given generation's population members and recombine them. Instead, they evaluate a generation's members and build a high-dimensional probability space that describes for every dimension the likelihood that certain values should be picked again in future generations. There are multiple possibilities for designing this probability space, which all require a definition of how the dimensions of the high-dimensional space relate to each other. Typically, one of three different scenarios is assumed:

**Univariate dependency:** The parameters are assumed to be independent of each other. This means that every parameter in the high-dimensional space can be optimized separately, which is easy to model but often not truthful to realistic data.



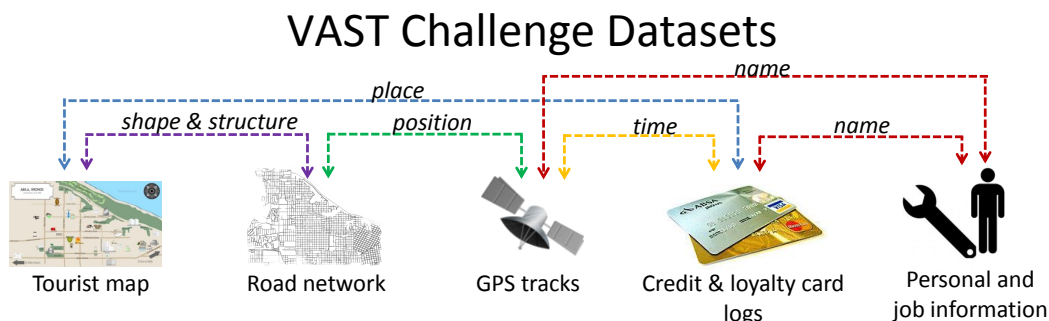
**Bivariate dependency:** It is also possible to assume that the parameters have a pairwise correlation. This often results in models that have a graph-like structure to describe the dependencies between the parameters.

**Multivariate dependency:** Although it is the most difficult to model, the assumption that all parameters are possibly related provides the most possibilities to describe the dimension's relation. One prominent way to model such high-dimensional dependencies are Bayesian Networks [151].

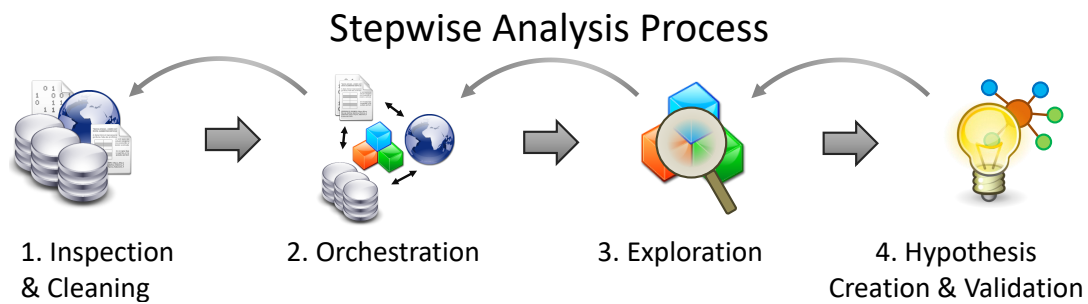
### 2.1.4 Visual Analytics Application Example

The following section shows how visual analytics can be used to derive insights from a previously unknown dataset through an iterative exploratory approach. The application example was designed in the context of the *VAST Challenge 2014* [190]. The scenario was set in the city *Abila* on the fictitious island *Kronos*, where a number of employees of the company *GASTech* were kidnapped. The task was to gather information about the whereabouts of these employees and who kidnapped them. The *VAST Challenge 2014* comprises of three mini-challenges that address different aspects of the kidnappings, such as e-mail traffic, movement and transaction logs, and social media data streams. The data provided in the second mini challenge, which this approach focuses on, comprises a tourist map, a road map, and historical data of the past two weeks about the inhabitants of *Abila*, such as credit and loyalty card logs and GPS data about the rental cars of *GASTech*'s employees (see Figure 2.8).

At the beginning, very little is known about the data. Aside from the usefulness and quality of the individual datasets, the synergies of the different data sources are unclear. To analyze the data, a visual analytics system called *InspectorGadget* was developed that helps to analyze the data iteratively. Analyzing the data comprises multiple steps, as shown in Figure 2.9.



**Figure 2.8:** Overview of the heterogeneous datasets provided in the *VAST Challenge 2014 – Mini Challenge 2* that had to be aligned. ©2015 The Eurographics Association

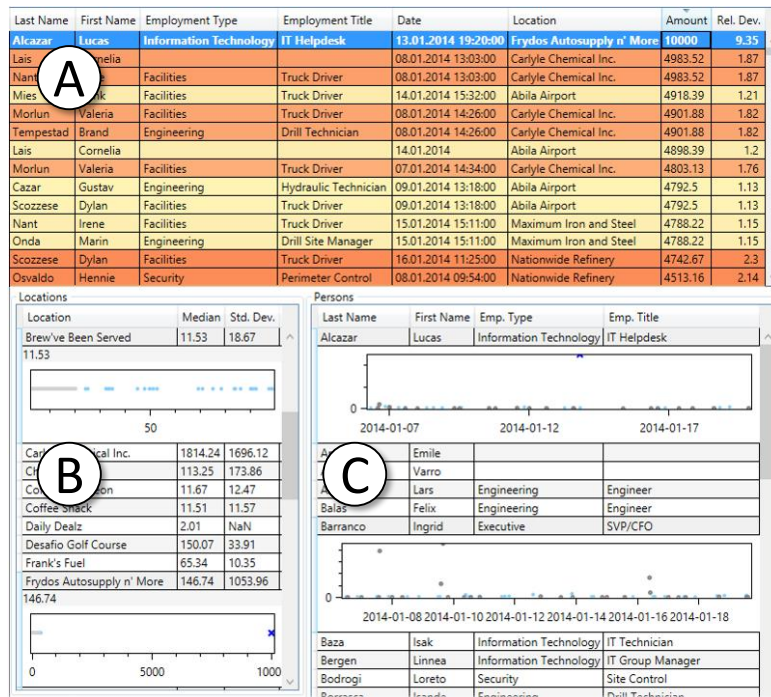


**Figure 2.9:** ① Inspect domain unspecific data characteristics/data cleaning & pre-filtering; ② Map the data sources in domain-specific views; ③ Explore details with various visualization; ④ Externalize findings, refine, validate. © 2015 The Eurographics Association

Initially, the data needs to be inspected and cleaned. The inspection usually needs to be performed manually, as a computer does not know what to do with the data at this point. During the inspection, analysts can already perform manual cleaning of the data or define automated processes for this task. Once the data is prepared, different data sources need to be orchestrated to make use of possible connections between the data, which may later provide additional insights. Afterwards, the data can be explored, findings can be extracted, and hypotheses can be created and validated.

In the context of the VAST challenge, the tourist map and the road map need to be aligned to a city map. Additionally, the GPS data of the rental cars must be converted to movement trajectories that describe the movement of the employees who rented the car. The trajectories must then be aligned with the registered city map. In *InspectorGadget*, the alignment between the map and the data is performed manually, whereas the trajectories are extracted automatically.

Further, the credit and loyalty card data need to be combined to gain further insights about the places where the cardholder was at specific points in time. *InspectorGadget* provides table views to analyze the logs (see Figure 2.10). The upper tabular view (see Figure 2.10 (A)) presents the combined loyalty and credit card data, as well as an additional metric that shows the deviation of the transaction from average payments made at the location of the transaction. The table rows have a color coding ranging from yellow to red that indicates the strength of the deviation, where red indicates a strong deviation from average transactions. This enables analysts to quickly see which transactions may be of interest, based on the transaction value. The second table view (see Figure 2.10 (B)) lists all locations, the median of the transactions, and the standard deviation. Analysts can expand any location to get further details about the distribution of the transaction amounts. The third table view (see Figure 2.10 (C)) presents the transactions by person. The detailed information shows the transaction amounts



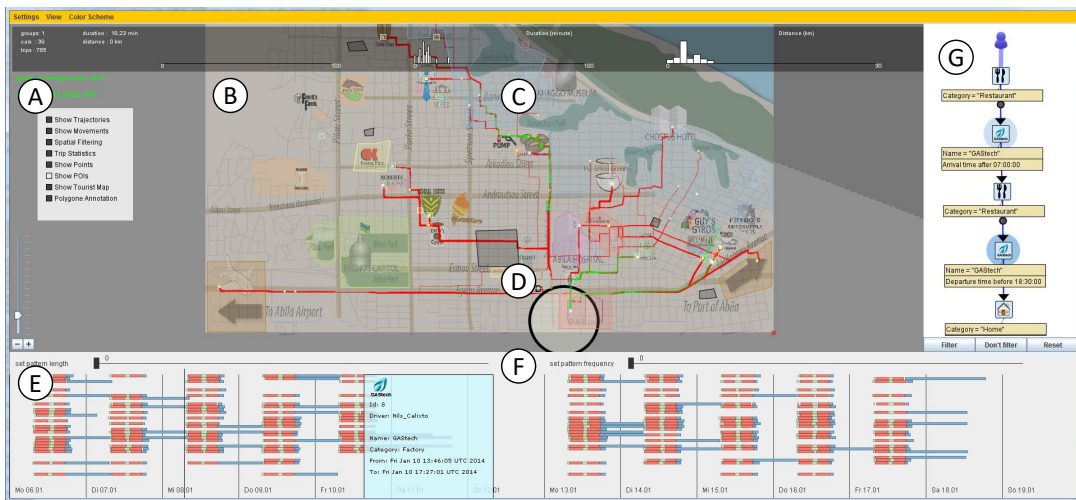
**Figure 2.10:** Inspection and Cleaning: A table (A) shows initially available data (here transactions). Further, detailed information about locations (B) and persons (C) can be inspected. Color indicates deviation from the average expense at a location (more intense red  $\Leftrightarrow$  higher deviation). © 2015 The Eurographics Association

of the selected person over time. All views are linked so that the selection of an entry, for example in the overview table (A), opens the detailed views of the other tables for the according entry and highlights the transaction.

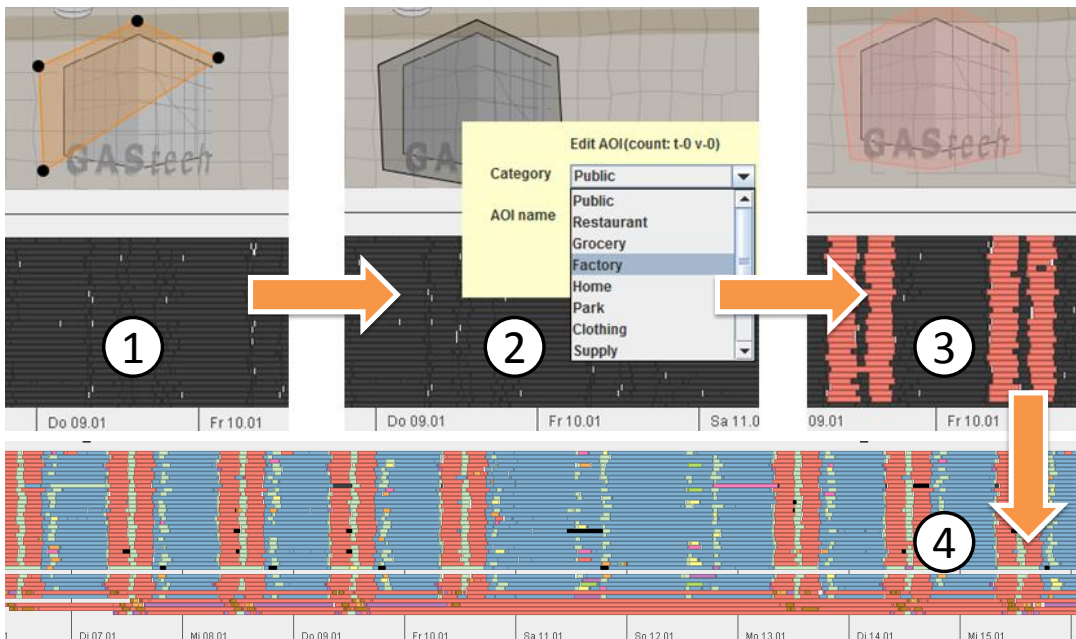
The table views can be used to extract first findings of unusual transactions. Suspicious transactions lead to hypotheses that either the corresponding person or location may be of interest for further investigations. These can be investigated in detail in the second component of *InspectorGadget*, which focuses on the analysis of the spatio-temporal aspects of the data (see Figure 2.11).

Before the analysis of the spatio-temporal data can begin, the data needs to be enriched with areas of interest (AOIs) that can be used to create movement profiles of the employees at *GASTech*. To create movement profiles, analysts can iteratively define AOIs, which are then synchronized with the GPS and transaction data to provide information about the whereabouts of the persons. An area of interest comprises its area, a textual description, and a category, for example, *office* or *home*. These categories are used in the sequence view to highlight daily patterns. Figure 2.12 shows the process of the definition of an AOI and how it affects the sequence view.

Once the movements have been enriched with AOI information, analysts can start the exploration process. The general inspection of the car movements



**Figure 2.11:** The analysis system (A) Map Overlays on/off; (B) Geographic Map View - animated movements (green) and trajectories (red); (C) Annotations - define and extract areas of interests; (D) Lenses - filter trajectories by origin/destination/way; (E) Sequence View - AOI movement sequences per employee; (F) N-Gram Sequence Filter - detects frequent and outlier patterns on a per user basis; (G) Pattern Filter - externalize knowledge, refine hypotheses and query the data. © 2015 The Eurographics Association



**Figure 2.12:** Annotation Process - (1) With a polygon tool an AOI can be created; (2) The AOI can be annotated with name and category, e.g., *GASTech*; (3) Movements are enriched and colored based on their destination (here *GASTech*); (4) All AOIs are annotated / all movements enriched. © 2015 The Eurographics Association

reveals that during night time nearly all cars are parked in the north-east city area—the employees’ homes. This behavior can be included in a pattern filter (see Figure 2.11 (G)). The results of the pattern are shown in the Sequence View (see Figure 2.11 (E)). Afterwards, a magic lens tool [112] can be used to further query the movements based on their origins and destinations while obtaining immediate feedback on the map (see Figure 2.11 (D)). For example, by inspecting trips from the city airport, one can see the arrival of the *GAStech* CEO a few days prior to the kidnapping. Further, the analysis of the movement behaviors during lunchtime reveals that two of the employees regularly drive to a hotel at the other end of Abila. This behavior is suspicious as it is unlikely that there are no other restaurants closer to *GAStech*. In addition, during the night, some of the cars are neither parked at *GAStech*, nor at any of the employees’ homes, which makes the cars’ owners suspicious. Also, the transaction shown in Figure 2.10 (A), which was conducted by Lucas Alcazar at Frydo’s Autosupply, is unusually high for an IT helpdesk employee. This could be a coincidence, but the fact that the shop (Frydos Autosupply n’ More) also never had such a high transaction makes this event suspicious.

The analysis of the data provides starting points for forensic analysts for further analyses, which of the suspicious behaviors may be connected to the kidnappings. These need to be combined with findings from the other mini-challenges to reveal the whereabouts of the missing employees and the culprit’s identity and motive.

## 2.2 Advanced Manufacturing

This section provides foundations regarding (advanced) manufacturing and principles and techniques used in this thesis. First, the historical background that lead to modern manufacturing principles, such as lean production and advanced manufacturing, is given (Section 2.2.1). Afterwards, the product lifecycle is introduced, which the remainder of this thesis is aligned to (Section 2.2.2). In addition, the overall equipment measure (Section 2.2.3) and basics of process planning (Section 2.2.4) are introduced.

### 2.2.1 From Traditional to Advanced Manufacturing

Manufacturing is generally understood as the transformation of physical goods into goods or services of higher value. The term production, although it is often used as a synonym to manufacturing, describes a more holistic concept that also comprises the design and sale of the produced goods. Originally, most goods were produced through manual labor (manufacturing originates from the Latin

term *manu factum*, which means *by hand*) or with the help of simple tools (e.g., hammer, saw).

Before the 18<sup>th</sup> century, the largest parts of the population worked in the agricultural sector to produce food for themselves or their landlord. The invention of the Roberts Loom, which was the first steam-driven power loom, in 1830 and the invention of various other machine tools allowed the automation of many repetitive tasks, such as spinning. This *first industrial revolution* allowed large improvements of the productivity in various industrial sectors, such as the textile, chemical, and metal casting industry and inventions such as the seed drill solved the constant food supply issues that were present up to that point.

The *second industrial revolution* started when electricity became usable in 1880 and electric devices were invented, such as the telephone in 1885. With the parallel extension of railroad networks, communication and travel speeds increased considerably. About the same time, Frederick Taylor introduced his so-called *scientific management*, which fundamentally changed how factories operated: Until 1880, most workers acquired their skills through practice and rules-of-thumb based on prior experience were commonly accepted standards. Taylor proclaimed that workflows should comprise methods that are based on scientifically studied processes and workers should be specifically selected and trained for the tasks that they need to perform. Together with the invention of assembly lines by Henry Ford in 1908, the efficiency and effectiveness of the overall production of factories rose and many goods also became affordable by persons with average earnings. In his autobiography, Ford wrote:

Therefore, in 1909 I announced one morning, without any previous warning, that in the future we were going to build only one model, [...] and that the chassis would be exactly the same for all cars, and I remarked: 'Any customer can have a car painted any colour that he wants so long as it is black.' [...] We were, almost overnight it seems, in great production. How did this come about? Simply through the application of an inevitable principle. By the application of intelligent directed power and machinery.

*Ford [72, p. 72–74]*

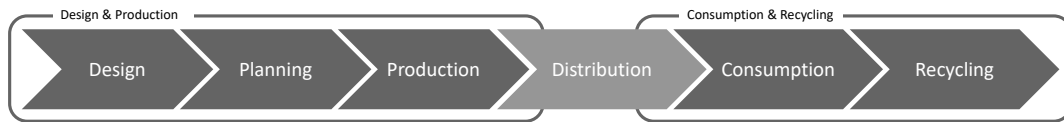
Ford's statement already indicates that the mass production of goods comes at the price of the product range that production lines were able to produce. In 1908, this was not a problem, as most people were not able to afford most goods produced through mass production before. However, in the years after the second world war, most economies grew and as a result the welfare of the population of many countries rose along with it. As many people were now able to afford former luxury goods, a need to distinguish oneself from the others

through customized products arose. As a result, factories needed to increase their flexibility regarding the products and product variants they were able to build. One of the most important principles that helped with this transition was the introduction of *lean production*. *Lean production* is derived from the Japanese manufacturing industry, specifically from Toyota's production and work principles. Generally, it endorses a production that minimizes storage of goods, be it intermediate or final products, while optimizing the workload of a production line. To minimize the storage of goods, supplemental products need to be supplied and processed just when they are needed. This increases the complexity of the overall production process, as this means that not only a company's own production needs to be optimized, but the supply chain also needs to be taken into account.

In parallel to the introduction of lean production, the invention of the first bipolar transistors in 1947 and its application in integrated circuits resulted in the invention of computers, mobile phones, and the internet. The transition of analog to digital technologies lasts until today and almost every aspect of the economy is affected. Examples for applications in the manufacturing domain are order acceptance, work scheduling, planning and designing of new products, and the introduction of computer numeric control (CNC) machine tools.

Nowadays, most IT infrastructures in the manufacturing domain follow a centralized approach, wherein all information is collected and distributed through a central system. For example, manufacturing execution systems (MES) collect data, such as the scheduling of production processes, the execution of production orders, or performance analyses of machines. These data can then be used to alert technicians of problems in a production line or they can be used to optimize the production. To represent the increasing integration of information technology in manufacturing the term *advanced manufacturing* was introduced. The Organisation for Economic Co-operation and Development (OECD) defines advanced manufacturing technology as "*computer-controlled or micro-electronics-based equipment used in the design, manufacture or handling of a product*" [146].

Recently, concepts like *Industrie 4.0* in Germany or the *Advanced Manufacturing Partnership* in the USA focus a tighter integration of information technologies with production systems. An important technology in this field is the concept of *cyber-physical systems*, which "comprise interacting digital, analog, physical, and human components engineered for function through integrated physics and logic" [143]. This technology supports the manufacturing domain in several aspects, such as autonomous purchasing or the deployment of smart machine tools, allowing a decentralized communication within a factory. The latter could lead to self-organizing machine tools that can ease production downtimes. For example, in case a machine breaks down unexpectedly, the production order could be rearranged to a limited degree to rebalance the workload of still functional machines without requiring human intervention.



**Figure 2.13:** The lifecycle of a product is often separated into five phases: initial design, planning & production, distribution, consumption, and recycling. These phases can be grouped into two parts: first, the product is being designed & produced and in the second part it is being consumed & recycled (cmp. Westkämper et al. [194, p. 153]).

### 2.2.2 Product Lifecycle

The product lifecycle is a concept commonly used in the production domain to describe the different stages of a product from its initial design until it is disposed of or recycled [173]. Usually, the lifecycle is divided into five phases:

- 1. Design & planning:** Initially, a new product is designed, either by improving other products in certain aspects or by designing an entirely new product. Afterwards, the way to produce the product is planned. Among other things, this includes defining the different process steps needed to create, manipulate, or assemble different components and, if necessary, to plan the layout of the production line that later produces the product.
- 2. Planning & production:** The production line is prepared and the production of the necessary components and the final products is performed.
- 3. Transportation:** The product is distributed to a consumer.
- 4. Consumption:** The product is being used, either as a component for another product or as a final product.
- 5. Recycling:** After the product either breaks or is not being used anymore, it is disposed or recycled.

According to Westkämper et al. [194, p. 153ff.], the cycle can be split into two parts: The design and production part and the consumption and disposal part. Further, They model design and planning as separate phases (see Figure 2.13). This thesis focuses exclusively on the first part.

### 2.2.3 Overall Equipment Effectiveness

To successfully run a factory, it is necessary to be aware, how *good* the performance currently is. One often used measure to define *good* is the overall equipment effectiveness (OEE) [141]. It takes into account how long a production line was operational, as well as its efficiency and effectiveness.



The OEE is defined as

$$OEE = Availability \cdot Efficiency \cdot Quality, \text{ with}$$

$$Availability = \frac{\text{run time} - (\text{planned and unplanned downtime})}{\text{run time}}$$

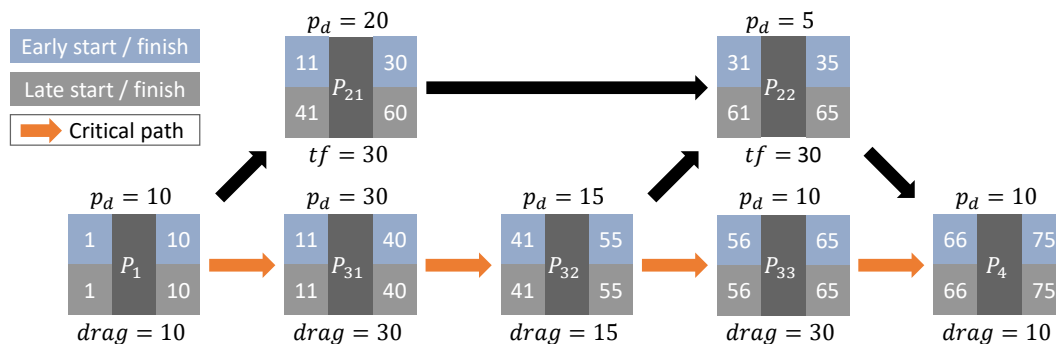
$$Efficiency = \frac{\text{produced parts}}{\text{max. producible parts}}$$

$$Quality = \frac{\text{produced parts} - \text{defective parts}}{\text{produced parts}}$$

It is important to note that this measure only accounts for the time when a production line is (or should be) running. It does not consider that a production line could be idle, e.g., if the day's workload is finished prematurely or if there is currently no need for the goods produced at a given production line.

### 2.2.4 Process Planning

In manufacturing, the output of products per hour is an important measure for the effectiveness of a production line (see Section 2.2.3). This throughput is defined by the time it takes a production line to complete all necessary processing steps. Usually, the production comprises of multiple steps with a given process duration  $p_d$ , which usually require one or more prior steps to be completed first. The following denotes the transport time between the steps with  $p_t$ . During the process planning, these dependencies are interpreted as a graph. The longest path in this graph from the start to the end point is called the *critical path* [71], which defines the theoretically possible throughput. Figure 2.14 presents an example of processes and their critical path along with other metrics that are presented in the following.



**Figure 2.14:** Example of a critical path analysis including drag and total float.

In addition to the critical path, other metrics are calculated for the individual process steps:

**Early start** ( $p_{es}$ ) is the earliest point in time at which a process can be started (considering possible dependencies on prior process steps).

**Early finish** ( $p_{ef}$ ) is the earliest time a process can finish.  $p_{ef} = p_{es} + p_d + p_t$

**Late start** ( $p_{ls}$ ) is the latest point in time this process can start to realize the planned throughput time.  $p_{ls} = p_{lf} - (p_d + p_t)$

**Late finish** ( $p_{lf}$ ) is the latest point in time at which a process can finish to meet the planned throughput time (considering prior process steps).

Using these metrics, three more important measures can be defined:

**Total float (tf)** is the time a process step can be delayed before it impacts the final deadline. On a critical path, the total float is zero.  $tf = p_{lf} - p_{ef}$

**Free float (ff)** is the time a process step can be delayed before it impacts the early start of its successor activities. This is different from the total float, as other required process steps for the successor activities may finish later.

**Drag** describes the time difference from the critical path to the next shortest path. In other words: if the processes on the critical path can be reduced by more time than the drag value, then they will not be on the critical path anymore.

## Exploration of Concept Relationships

At the beginning of the creation of any product is the design phase. Therein, the functionality and the look of the product are planned. Companies often have characteristic details that distinguish them from their competitors. There are easily perceivable characteristics, such as the chassis of a car. Other characteristics that may be too technical for end-users are typically communicated through key indicators that combine multiple processes to an attribute affecting the consumer. For example, the processes affecting the efficiency of a car motor are usually communicated through indicators such as the car's acceleration, its fuel consumption, or its engine power. Nowadays, having a technological advantage over competitors in any of these properties is often the key to a superior positioning on the market. Patents are a popular mechanism to protect the intellectual property of this advantageous knowledge. This applies to the design of the product itself, as well as the processes, machinery, and technologies needed to produce the product. Aside from protecting one's knowledge, it is important to be aware of possibly conflicting patents issued by competitors, as these may potentially halt the entire production of a product.

Ideally, a specialist is tasked with handling all matters regarding intellectual properties is aware of every patent that may apply to the currently planned and produced goods. However, the number of patents is continuously growing. According to the world intellectual property organization (WIPO), more than three million patents were issued in 2017 [195]. It is unfeasible to read all of these patents to assess if they may be applicable to one's own product range. Therefore, it is important to get a quick overview of the patent landscape and provide experts with means to filter the number of patents so that they can assess them individually. In addition, it is important to know, which persons are knowledgeable in a specific field or how technologies relate to each other, to be able to find patents that are possibly designed for another purpose but may be applicable nevertheless.

After the introduction of the general structure of patents and the international patent classification system (IPC) [99] (Section 3.1), this chapter presents related work regarding the visualization of patents and scientific literature, and visual analysis of documents, focusing on approaches that use dimension reduction (Section 3.2). Afterwards, two approaches are presented. The first, named *IPC Clouds*, uses the IPC taxonomy to provide an overview on how topics relate to each other (Section 3.3). The second approach extends the idea of *IPC Clouds* by replacing the IPC taxonomy with concepts, which are projected onto a concept map using a hierarchical clustering approach (Section 3.4). These concepts can be extracted from patents, but they may also be important persons from a domain or tags from websites.

This chapter is partly based on the following publications:


- D. Herr, Q. Han, S. Lohmann, S. Brüggemann and T. Ertl. “Visual Exploration of Patent Collections with IPC Clouds”. In: *Proceedings of the 1st International Workshop Patent Mining and Its Applications*. Vol. 1292. CEUR-WS. CEUR-WS.org, 2014 [3]
- D. Herr, Q. Han, S. Lohmann and T. Ertl. “Visual Clutter Reduction through Hierarchy-based Projection of High-dimensional Labeled Data”. In: *Proceedings of Graphics Interface*. CIPS / ACM, 2016, pp. 109–116 [4]
- D. Herr, Q. Han, S. Lohmann and T. Ertl. “Hierarchy-based projection of high-dimensional Labeled Data to Reduce Visual Clutter”. In: *Computers & Graphics* 62 (2017), pp. 28–40 [5]

## 3.1 Patents and their Classification System

The following section first gives a brief overview of the structure of a patent and the metadata it comprises (Section 3.1.1). One important information is the classes a patent belongs to. These classes are standardized by the international patent classification system (IPC). This classification system is presented in more detail, as parts of the approaches to analyze the patent landscape rely on the IPC taxonomy (Section 3.1.2).

### 3.1.1 Patent Structure

Patents are structured legal documents that comprise multiple metadata fields that can be used for analysis, partly independent of the patents’ actual content. Depending on the country, different patent offices may be able to publish a patent (e.g., the European Patent Office (EPO) or the German Patent and Trade Mark Office (DPMA)), but they all share large parts of the available metadata. Figure 3.1 shows an exemplary patent published by the EPO. Important properties of a patent are its date of publication (Fig. 3.1 Ⓐ), the title (Fig. 3.1 Ⓑ), and the

(19)		
		(11) <b>EP 2 222 950 B1</b>
(12)	<b>FASCICULE DE BREVET EUROPEEN</b>	
(45) Date de publication et mention de la délivrance du brevet: <b>10.08.2016 Bulletin 2016/32</b>	(51) Int Cl.: <b>F02N 11/08 (2006.01) F02N 15/06 (2006.01)</b>	(D)
(21) Numéro de dépôt: <b>08868886.6</b>	(86) Numéro de dépôt international: <b>PCT/EP2008/067893</b>	
(A) (22) Date de dépôt: <b>18.12.2008</b>	(87) Numéro de publication internationale: <b>WO 2009/083477 (09.07.2009 Gazette 2009/28)</b>	
<hr/>		
(B) (54) <b>PROCEDE DE COMMANDE POUR DEMARREUR D'UN MOTEUR A COMBUSTION ET SON APPLICATION</b>		
VERFAHREN ZUR STEUERUNG DER ZÜNDUNG EINES VERBRENNUNGSMOTORS UND ANWENDUNG DAVON		
METHOD FOR CONTROLLING THE STARTER OF A COMBUSTION ENGINE AND APPLICATION THEREOF		
<hr/>		
(C) (84) Etats contractants désignés: <b>AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MT NL NO PL PT RO SE SI SK TR</b>	(72) Inventeurs: • <b>COUETOUX, Herve</b> F-91470 Les Molières (FR) • <b>GUILLOZ, Nicolas</b> F-92140 Clamart (FR) • <b>SCHEFFGES, Olivier</b> F-92400 Courbevoie (FR)	(E)
(30) Priorité: <b>20.12.2007 FR 0708950</b>	(56) Documents cités: <b>EP-A- 1 041 277 WO-A-2007/101770</b> <b>DE-A1- 10 005 005 US-A1- 2004 017 086</b> <b>US-A1- 2007 084 429</b>	(F)
(43) Date de publication de la demande: <b>01.09.2010 Bulletin 2010/35</b>		
(73) Titulaire: <b>Renault S.A.S.</b> <b>92100 Boulogne-Billancourt (FR)</b>		

**Figure 3.1:** Example of a patent published by the European Patent Office [95]. A patent comprises various metadata information, such as its publication date (A), title (B), applicable countries (C), IPC classification (D), the inventor (E), and other related patents (F).

countries the patent applies to (Fig. 3.1 C). In addition to these data, all patents are assigned to one or more classes that describe the overall topic(s) of the patent (Fig. 3.1 D). This standardized classification system is called *international patent classification system (IPC)*. Further, the inventor (Fig. 3.1 E), and other patents related to the patent itself (Fig. 3.1 F) are provided, which implicitly contains information about related topics and other important patent authors, much like in scientific literature. Due to their legal relevance, these metadata are checked and cleaned before a patent is published.

### 3.1.2 International Patent Classification System

The International Patent Classification (IPC) [...] provides for a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain.

The IPC divides technology into eight sections with approximately 70,000 subdivisions. Each subdivision has a symbol consisting of Arabic numerals and letters of the Latin alphabet.

*World Intellectual Property Organization [200]*

As stated above, the IPC describes the topics to which a patent belongs to based on a multi-level hierarchy. It is updated every year to reflect shifts in technology. If a new technology field emerges or a subfield grows noticeably, this change will be reflected in the next IPC, either by splitting an already existing (sub-)class or by introducing a new class. A recent example is the introduction of the IPC class *G16 (information and communication technology [ICT] specially adapted for specific application fields)* in the IPC edition of 2018 to reflect the increasing importance for this technology field.

Figure 3.2 shows an example, what information is provided for the IPC class *F02N 11/08*. The first level of an IPC class is its *section*, denoted with a capital letter followed by a *class* described by two digits. Every class is split into *subclasses*, which are indicated by a capital letter after the class. To provide further details, a subclass is divided into *groups*, which are indicated by two numbers separated by an oblique stroke. The IPC differentiates between *main groups* and *subgroups*. The second number block of a main group is always *00*, whereas the number of subgroups can be any number greater than *00*. More details about the international patent classification is given in the IPC guide provided by the WIPO [201].

Section	F	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING ENGINES OR PUMPS
Class	02	COMBUSTION ENGINES; HOT-GAS OR COMBUSTION-PRODUCT ENGINE PLANTS
Subclass	N	STARTING OF COMBUSTION ENGINES; STARTING AIDS FOR SUCH ENGINES, NOT OTHERWISE PROVIDED FOR
Main group	11/00	Starting of engines by means of electric motors
Subgroup	11/08	Circuits specially adapted for starting engines

**Figure 3.2:** Exemplary breakdown for the IPC code *F02N 11/08*.

## 3.2 Related Work

Enabling experts to explore, which patents relate to each other or which documents share common topics is an important step while a new product is being designed. Showing these relations on a two-dimensional plane, wherein close elements indicate a stronger relationship, is a common approach for this task. To present this information, the topic relations first need to be placed on a two-dimensional plane, which is often done using data projection algorithms. The following section focuses on the visual exploration of concept relationships through their projection as labels on a two-dimensional plane. These concepts can be either metadata from documents, such as patents, scientific literature, or websites. Therefore, the following presents related work from the fields of dimension reduction and data projection, word cloud visualizations, the visualization of hierarchically structured data, and patent and scientific literature visualization in general.

### 3.2.1 Dimension Reduction and Data Projection

One common approach for visualizing the content of datasets, for example, document collections, is to take extracted features, such as keywords or concepts, and compare the data based on those features. The natural language processing community developed many ways to extract keywords from documents [86]. The relatedness of such features is usually computed pairwise. To visualize these high-dimensional relationships, the relations need to be transformed so that they can be visualized. Methods to solve this problem can be projection-based, for instance, by using Principal Component Analysis (PCA) [197], Multidimensional Scaling (MDS) [115], Least Squares Projection [149], or t-SNE (see Section 2.1.3). Van der Maaten et al. [188] survey a number of techniques to project high-dimensional data onto a low-dimensional space. In such approaches, data is usually visualized as data points that take up almost no space. This representation suffers from visual clutter when the data is not clearly separable and is usually not designed for a data representation with labels. Sacha et al. [163] reviewed how dimension reduction techniques are used and integrated into interactive visualization techniques.

Due to the high complexity of projection techniques, other potentially faster and more intuitive approaches, such as force-based layouts [74, 76] are often used when interactive visualizations are needed. However, García-Fernández et al. [77] conclude in their study that projection-based approaches are superior to force-based layouts when a complete and large dataset needs to be visualized.

Other approaches are based on neural networks, such as hierarchical self-organizing maps (SOM) [110]. If such approaches are applied to subdivided areas, as proposed by Suganthan [178] or Endo et al. [59], then they create a

visual structure seemingly based on a hierarchy. However, by generating a visual hierarchy this way, only elements contained in the same area are mapped relative to each other. The approach presented later in this chapter uses a hierarchy not only to visualize the relation between parents and their children but also between the siblings across the clusters.

## 3.2.2 Word Cloud Visualization

As the approaches presented in this chapter visualize concepts as labels, they are related to word cloud visualizations that show the most frequent words of a text as a weighted list in some specific spatial arrangement, such as a sequential, circular, or clustered layout [127]. Several variations and advancements of word clouds have been proposed in the past years. One line of research concerns the improvement of the layout of word clouds. For instance, Seifert et al. [167] developed algorithms for space-filling word clouds based on a set of heuristics, while related layout algorithms have also been presented in a number of other works, such as ManiWordle [109] and Rolled-out Wordles [177].

Some layout strategies consider word relationships and implement spatial arrangements where strongly related words are placed in close proximity, similar to the presented approaches. The layout strategies range from simple line-by-line approaches [87] to treemap-like layouts [103] and force-directed placements in combination with Venn diagrams [46]. Some works even apply projection techniques, such as the aforementioned MDS, to reflect the relatedness of words [150, 203], while others use topographical word landscapes [75].

There are also attempts to explicitly depict the relationships in word clouds, either by adding links between related words [174] or via interactive highlighting [93]. DocuBurst [50] uses a sunburst visualization related to word clouds to show a hierarchy of concepts extracted from text documents. Prefix Tag Clouds [39] make use of prefix trees to group different word forms, whereas the Word Cloud Explorer uses advanced NLP processing to link word forms and to support the visual analysis of text documents via interactive word clouds [93].

## 3.2.3 Hierarchical Aggregation and Visualization

The approach presented in Section 3.4 introduces a hierarchical structure to the data before projecting it on a plane. Various approaches exist to provide detailed information about hierarchy-based data, which are discussed in the following.

Dou et al. [54] generate topic models in which the users can interactively modify the created hierarchical structure. Afterwards, users can inspect the development of individual or groups of topics over time. In contrast to the approach presented in Section 3.4, this attempt uses the hierarchies to aggregate topics to analyze changes over time. Like most hierarchical visualization approaches,



Dou et al. assume the availability of a hierarchical structure and use predefined hierarchy levels to show information. The approach presented in Section 3.4 goes beyond that by enabling users to set the shown hierarchy levels by themselves.

Liu et al. [125] developed a system to build hierarchies based on topic graphs. They visualize the relations of extracted topics by using stacked trees in combination with force-based graph layouts. Unlike the later presented approach, they use hierarchies to distinguish between topics. The presented approach aims not only to provide relational information across clusters but also of their content when more information is shown.

Fried and Kobourov [73] presented a system, that maps the titles of papers in the DBLP database onto a two-dimensional landscape based on a hierarchy. Users can create a search profile that is used to highlight topics using a heat map visualization, focusing on temporal aspects of the data.

Wise et al. [196] proposed to show large document collections through a galaxy metaphor, in which every document is represented by a star. By doing so, the user gets a more intuitive understanding of the relations between documents. Similarly, SPIRE [182] and INSPIRE [198] use the same metaphor, but they combine it with a visual analytics approach to enable users to further analyze the data. The STREAMIT system [13] uses force-based layouts, clustering discovery techniques, and topic modeling to visualize document streams in real-time. The clustering is based on the graph layout and does not create a hierarchy to examine the document streams on a semantical level. An early version of the *Overview* system [35] projects documents onto a two-dimensional scatterplot while showing a hierarchical tree structure of the documents in another view. The system uses brushing and linking to connect those two visualizations. However, the selected elements of the hierarchy view are not represented in the number of data points shown in the scatterplot.

Stahnke et al. [172] propose an interaction technique to interpret arrangements and errors in dimensionality reduction. They do this by enabling the user to *probe* the projected data and show more detailed information about the data and the relevant projection information.

Only a few of these landscape-based approaches support hierarchical data and most that do assume the hierarchy to be given. Thom et al. [180] use hierarchical topic clustering combined with a treemap-based visualization to show Twitter data on different levels of detail. This level of detail can be interactively steered by the user during an analysis run, which enables the user to see more details about a specific topic.

The goal of most hierarchical visualization techniques is to show the structure of the hierarchy. As such, the elements contained in the various presented clusters need to be adequately visualized. Elmqvist and Fekete [58] propose guidelines how hierarchical data can be used to limit the amount of shown information, such as taking the most important element of a cluster as cluster representative.

### 3.2.4 Patent and Scientific Literature Visualization

The first approach presented in this chapter focuses on the visualization of the relations between IPC classes. The IPC space is rarely visualized in related work. As it is a hierarchy, it is often presented in some kind of tree view that the user can navigate to find IPC symbols of interest. Kutz uses a sequence of treemaps to visualize the evolution of the IPC system over time [118]. However, the treemaps are structured according to the manually designed IPC hierarchy without considering other IPC relations in the patent data, which may not represent their actual relations based on patents.

Another popular visualization technique in the patent domain are node-link diagrams. Other approaches use node-link diagrams to show relations between patents and priority documents [79, 108] or to graphically depict networks of applicants or inventors [176]. Node-link diagrams can be very useful to explore the patent space and to identify important clusters in the patent data.

Due to their similar structure, approaches that are designed for scientific literature can also be applied to patents. Federico et al. [69] reviewed visual analytics approaches to analyze scientific literature and patents. Heimerl et al. [90] present an approach to analyze important topics in scientific literature over time. This approach was extended by Han et al. [83] for a detailed analysis of important authors over time and they applied this approach to patents.

Heimerl et al. [91] project documents to a two-dimensional plane and use magic lenses to show important keywords in documents. When covering multiple documents, the most important keywords are ordered around the lens and the selection of keywords highlights their occurrences in other documents.

## 3.3 Visual Exploration of Patent Relations using IPC Classes

The IPC taxonomy introduced in Section 3.1.2 can be used as a reliable and easy way to access the data source to provide patent experts with an overview, which topics a patent covers. However, it is also possible to use IPC classes to find related patents that may not directly refer to each other. To do this, the similarity between IPC classes needs to be computed based on patents, not on the IPC hierarchy. Then, the similarity of the IPC classes needs to be communicated to experts so that they can extend their search from the relevant IPC classes they are aware of ones that they did not look into before. One approach to provide such insights is to show IPC classes on a map, where closer distance encodes the similarity between the IPC classes. To support this exploration, a visualization approach, named *IPC Clouds*, is introduced. The following details the exploration of the IPC landscape using *IPC Clouds*. First, the retrieval and storage (Section 3.3.1) and the preprocessing (Section 3.3.2) of the used patents is

presented). Then, two *IPC Cloud* views are introduced (Section 3.3.3). Finally, an example shows how the approach can be used (Section 3.3.4) and the scalability of the approach is discussed (Section 3.3.5).

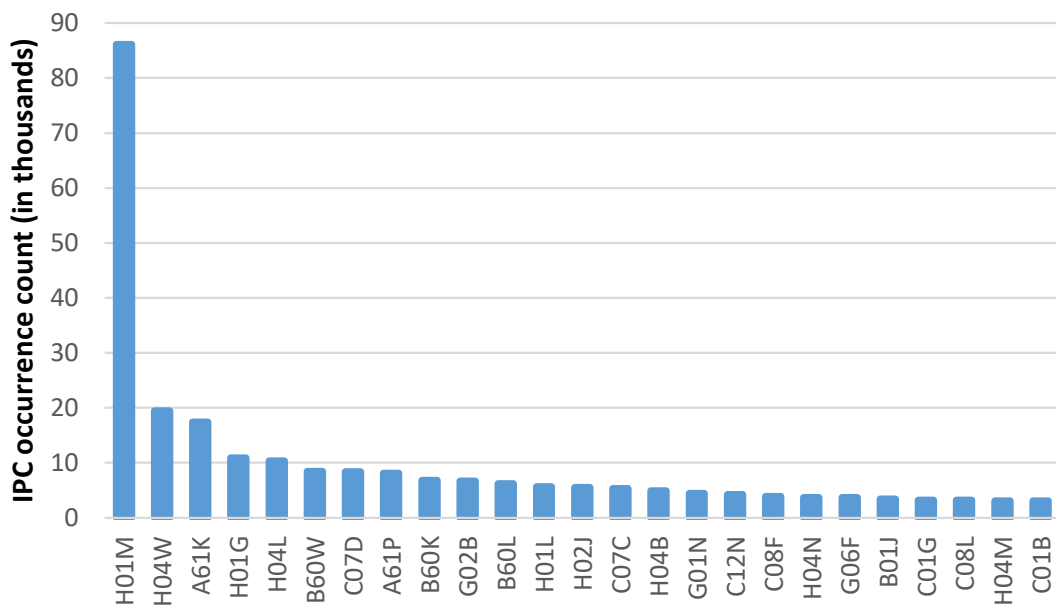
### 3.3.1 Patent Data Retrieval and Storage

Although patents are publicly accessible, they still need to be retrieved and stored so that they can be used for an efficient analysis. As patents contain structured information, such as the previously introduced metadata (see Section 3.1.1), and unstructured data, e.g., their description and included figures, the usage of relational databases is problematic. Therefore, the patents used in the following approaches are stored in the document-oriented NoSQL database Elasticsearch [57]. Elasticsearch is based on the information retrieval software Apache Lucene, and stores the data in the JavaScript Object Notation (JSON) format, which does not require a specific structure for its entries. One advantage of the unstructured data storage is that new information can easily be added to a subset of records without the need to update other records in the database or to use empty fields. Another useful characteristic of document-oriented databases is that they use indexers with a specific focus on an efficient retrieval of documents based on their text content.

Elasticsearch provides an interface to efficiently retrieve data through HTTP requests and exchanges them in JSON format. Moreover, the Lucene repository can be used to directly preprocess the data and perform computationally expensive tasks, such as the computation of the pairwise similarity of the IPC classes or the extraction of specific content from the patents.

For demonstration purposes, the used database comprises two repositories. The first contains a large number of patents of which only bibliographic information is available. The bibliographic information was taken from the PatStat database [62] of the European Patent Office. It includes the patent ID, title, abstract, applicant, inventor, filing and application dates, all IPC classifications, as well as citations for more than 70 million patents. In the following approaches, the classifications were limited to the IPC subclass level. They have an exponential distribution of their occurrence across all patents (see Figure 3.3). The second, smaller repository also contains the abstracts and description of the patent documents. These patents were retrieved from Espacenet [64], the European Patent Register [65], and the European Publication Server [61], using RESTful web services of the Open Patent Services [66].

The patent texts comprise the descriptions and claims for 88,000 arbitrarily chosen patents. All texts are indexed by Lucene and linked to the bibliographic information via their unique patent IDs. The PatStat data was transformed into the JSON structure of the Elasticsearch database using MongoDB [138].



**Figure 3.3:** The distribution of the IPC usage frequencies roughly follows a power law, as illustrated for the 25 most often used IPC symbols in the 88.000 patent records that were analyzed (in thousands).

### 3.3.2 Data Preprocessing

Before the *IPC Cloud* views can be generated, the patent data needs to be pre-processed. The preprocessing comprises two steps: First, the pairwise similarities between the IPC symbols are computed. Then, the IPC classes are projected onto a two-dimensional space based on the calculated similarities.

#### Computation of IPC Similarities

Similarities can be computed on different levels of the IPC hierarchy, i.e., on the class, subclass, group, or subgroup level. Discussions with patent experts showed that the IPC subclass level provides a good balance between the information that is contained in the IPC classes and the generality that provides an overview of other potentially relevant IPC classes. Consequently, *IPC Clouds* provides all information on the IPC subclass level. As explained in Section 3.1.2, the IPC subclass symbols comprise of four characters: a letter for the section, followed by a two-digit number for the class, and a letter for the subclass (e.g., *A01B*). In the IPC version IPC-2014.01, which is used in the processed patents, the IPC subclass level comprised 638 entries.

The patent data is converted into a vector space to compute the similarities between the IPC subclass symbols. Here, the 88,000 patents from the second repository of the database introduced in Section 3.3.1 were used. Each of the 615

subclass symbols contained in the dataset<sup>1</sup> is represented by one vector, with the patents as dimensions of the vector space. If the considered IPC subclass symbol is used to classify a patent, the corresponding dimension has a positive value; otherwise, it is zero. Then, the cosine similarity (see Section 2.1.3) of each pair of subclass symbols is computed to determine their relatedness in the patent data.

The cosine similarity is an efficient measure for sparse vectors, which is useful in this case, as each subclass symbol is associated with only a small fraction of the patents. This results in a small number of non-zero dimensions per vector compared to the total number of dimensions in the vector space, and hence in sparse vectors.

### Dimensionality Reduction of IPC Space

In the second step, the IPC subclass symbols are mapped onto a 2D plane that is used for the visualization of the classes. The objective of the two-dimensional representation is to place similar IPC classes close together, while dissimilar classes should be placed far apart. *IPC Clouds* uses t-SNE (see Section 2.1.3) to project the IPC subclass vectors to a two-dimensional space, with the previously calculated similarity matrix as the input data.

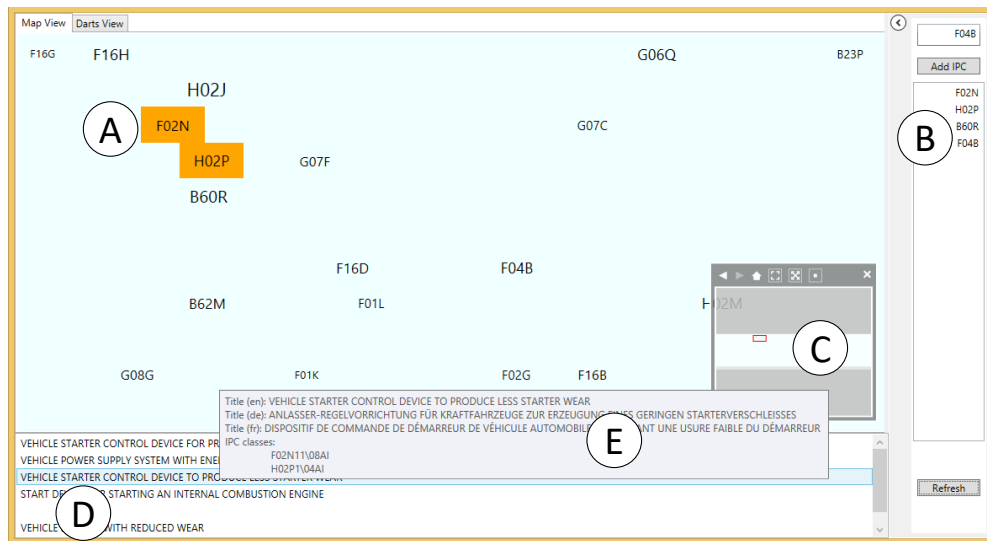
#### 3.3.3 IPC Cloud Visualizations

After the IPC subclasses are projected onto a two-dimensional space, the *IPC Cloud* views can be generated. The approach comprises two views that can be used for analysis, called *map view* and *darts view*, which is detailed in the following. While the *map view* provides a global overview on the IPC space, the *darts view* centers previously selected IPC symbols to support an easier visual identification of subclass symbols that are related to the selected ones. Both views follow the “visual information seeking mantra” [171] by first providing an overview, then allowing to zoom and filter the data, and finally showing details on demand.

##### Map View

The *map view* depicts the two-dimensional representation of the IPC space. As t-SNE maps the data to an arbitrary Cartesian coordinate system, the coordinates need to be normalized and rescaled first. By doing so, the mapping is transformed into a coordinate system appropriate for visualization, while the spatial distribution is retained. Scatterplots are a common way to visualize projected data because they are highly scalable regarding the number of data

<sup>1</sup> 23 of the 638 available subclass symbols were not used in the dataset.



**Figure 3.4:** The IPC subclass symbols are shown in the overview (A). Users can filter the shown subclasses through the filter component (B). The shown IPC symbols are limited to the subclasses added to the filter and IPC symbols that are used together with them. The minimap (C) indicates, which part of the IPC space is currently shown. The two highlighted IPC symbols have been selected by the user. The bottom part lists all patents that are associated with the selected IPC symbols (D). Further information about the patent, including all associated IPC symbols, can be displayed on demand (E).

points that can be visualized at once. Although clusters of points can be easily perceived this way, users need to interact with the visualization to understand the meaning of these clusters. Both *IPC Cloud* views show the IPC subclass symbols as labels so that users can quickly understand the meaning of the data points. The usage frequency of the IPC symbols is encoded in the font size of the labels. The font size uses a logarithmic scaling to counterbalance the exponential distribution of the IPC symbols. Using a non-linear scaling prevents often used IPC symbols from being overemphasized. When the coordinate system is being rescaled, it is important to consider the width and height of the labels, as the *map view* would otherwise be cluttered due to a possibly high number of overlapping labels. In case of the used dataset, a scaling factor of 25,000 resulted in a good overview and only few overlaps of the text labels.

After the layout has been computed, the IPC subclass symbols are placed at the determined positions on the screen, as shown in Figure 3.4 (A). In addition, users can remove the remaining overlaps. Keeping the relative distances of the labels roughly stable is important, as they reflect the relatedness of the IPC symbols. This disqualifies many algorithms for overlap removal that preserve the orthogonal ordering of the labels but not their relative distances [56]. The

*map view* uses the *push* variant of the Force-Scan Algorithm (FSA) [136] for this purpose, which preserves the general layout and, in particular, the relative distances of the nodes. The algorithm compares the label areas with each other and, if an overlap is detected, fixates the label that is further to the upper left and moves all other labels in the direction where the overlap is resolved the fastest. A common drawback of the push variant of FSA is the increased space needed by the labels. To compensate for this, the *map view* supports zooming and panning to navigate in the IPC space.

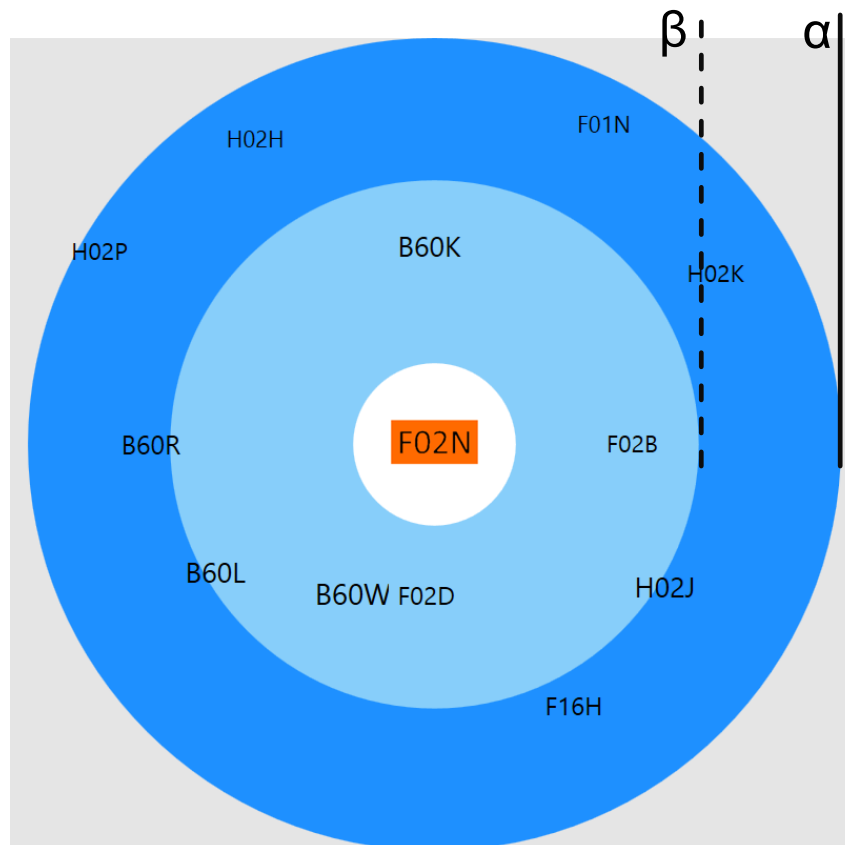
Panning and zooming are basic but important interaction techniques that enable the users to explore different parts of the *map view* in more detail. Furthermore, a minimap shows the whole IPC space and indicates which part of it is focused in the main view (Figure 3.4©). The minimap can also be used to change the focused area and to reset the zoom level.

Since experts are typically interested in specific IPC symbols, they can filter the *map view* to show only the filtered IPC subclasses and subclasses that are used together with them. This can be done by selecting any number of IPC symbols on the map, which adds them to a whitelist displayed on the right of the visualization (Figure 3.4®). As it can be hard to spot specific IPC symbols on the map, the IPC symbols can alternatively be entered in a search field (which has an autocomplete feature).

If users select an IPC symbol in the visualization, the titles of patents associated with that symbol are listed beneath the main view (Figure 3.4ⓓ). If several IPC symbols are selected, only titles of patents associated with all of the symbols are listed (i.e., they are connected by a logical conjunction operator). More details about a patent, such as the whole list of associated IPC symbols and its titles in German and French, are shown in a tooltip when hovering over the patent's title in the list.

### Darts View

The *darts view* provides another perspective on selected IPC symbols using the metaphor of a dartboard. In contrast to the *map view*, it does not provide a global overview on the IPC space but focuses on specific IPC symbols and their local context. IPC symbols selected in the *map view* or selected in the search field are placed in the center of the *darts view* (the bullseye) to represent what users are interested in. Related IPC symbols are concentrically arranged around the bullseye. The proximity of an subclass symbol to the center represents their relatedness to the selected symbols. IPC symbols close to the bullseye are strongly related, whereas symbols near the border have a weaker relation. This provides a more truthful visualization of the relation between a limited number of focused IPC labels and their neighborhood, which may be unclear on the *map view* due to the information loss introduced in the process of the dimension



**Figure 3.5:** Darts view showing one selected IPC symbol in the bullseye and related IPC symbols concentrically arranged around it indicating their relatedness.

reduction. Figure 3.5 shows an example where the IPC symbol *F02N* (*starting of combustion engines*) has been selected and is therefore placed on top of the bullseye. Labels, such as *B60K*, *B60W*, or *F02D* are most similar to this label, whereas *H02P* has only little relation to *F02N*. Overall, it appears that classes starting with *B60* (*vehicles in general*) are most related to the selected class.

The *darts view* requires the definition of two key parameters: 1) a maximum number of related IPC symbols ( $n$ ) shown in the visualization, and 2) a threshold ( $\alpha$ ) defining the minimum similarity value a related IPC symbol must have to be shown in the visualization. This ensures that only relevant subclass symbols are shown. Both parameters are interrelated and suitable values are dependent on the application context, such as the available screen space or the average font size of the labels. In most cases, showing between ten and 20 related labels led to a visualization that can be well perceived. An appropriate value for  $\alpha$  is more difficult to choose, as the similarity values are dependent on the considered patent data and IPC symbols. For the used patent data, a value between 0.5 and 0.7 led to good results in most cases. Figure 3.5 shows a darts view that was generated with an  $\alpha$  value of 0.6.



The *darts view* uses a static  $\alpha$  value instead of returning a static the number of related IPC symbols without considering the similarity, because the users may not recognize that the similarities of the classes may differ when analyzing different *darts views*. This could lead to misinterpretations of the data.

After the related IPC symbols have been determined based on the parameter values for  $n$  and  $\alpha$ , their positions on the dartboard are computed. Like the *map view*, the *darts view* makes use of the previously calculated 2D representation (see Subsection 3.3.2). When placing the related subclass labels, the *darts view* considers the direction vector towards the labels in the center of the *darts view* when determining the order of the IPC classes in the view. If multiple IPC symbols are selected, their labels are placed on top of each other in the bulls eye. The surrounding labels' angle is the average of the angles towards the individual labels. To minimize overlapping labels, the *darts view* does not preserve the actual direction of the IPC label but places the labels in an equiangular way so that they form a circle around the selected IPC symbol(s).

The distances of the IPC symbols in relation to the bullseye is also computed based on the pairwise similarity of the subclasses. The distance towards the center is modified using a logarithmic scale. This optimizes the space used by the view as it compensates for the distribution of the IPC subclasses. Finally, the IPC symbols are placed at the determined positions on the dartboard, while their font sizes indicate how often they are used in the patent data, analogous to the *map view*.

The *darts view* also supports the introduction of additional thresholds to indicate different levels of interest. For example, a threshold  $\beta = 0.75$  may indicate that subclasses with a similarity value  $\geq 0.75$  may be considered *important*, whereas subclasses with lower similarity may be *notable* (see Figure 3.5).

### 3.3.4 Example of Use

Let us assume the role of a company's patent expert who is tasked to file a patent for a new technique to start combustion engines. The IPC symbol *F02N* is ideally suited to classify our invention since it refers to the "starting of combustion engines" [195]. In the *map view*, we have already spotted said IPC symbol and noticed that the IPC symbol *H02P* is very close to it (as in Figure 3.4). It classifies patents that describe the "control or regulation of electric motors, generators, or dynamo-electric converters" [195]. Although this IPC subclass does not even share the section with our original IPC class it is clear that these two classes are related to each other, as electric and combustion engines share many technologies. Further, modern cars often have hybrid power trains that in fact comprise an electric and a combustion engine. For technologies relating to such composite engines, it is plausible to classify related patents with classes for combustion and electric engines. It seems to be a good idea to analyze the

patents related to electrical engine starters, because there may already be a patent that is in conflict with the invention we want to protect.

After switching to the *darts view*, we realize that there seem to be several other IPC symbols that are also strongly related to the IPC symbol we are interested in. In this case, these subclasses mostly relate to the IPC class *B60 (vehicles in general)*, which were not shown nearby our starting class in the *map view*. This insight leads us to further technologies and patents that might be of relevance and should be considered before filing our patent.

#### 3.3.5 Discussion of Scalability

Due to the massive number of patents that are digitally available nowadays, scalability is one of the main issues in any patent visualization approach. For the presented approach, the scalability of the data storage, the data preprocessing, and the data visualization regarding the number of patents and used IPC classes is most important. Table 3.1 summarizes the scalability of the components that are part of the presented approach. It indicates, how well the data storage, data preprocessing, and the views scale with an increasing number of patents and IPC symbols.

The scalability of the data storage is unproblematic, as Elasticsearch and Apache Lucene were designed to handle large amounts of text data. If new IPC symbols are added to the database, only the patent records classified by these symbols need to be updated, without updating any other patent records.

The data preprocessing also scales well with the number of patents regarding the retrieval of actual patents based on specific IPC subclass symbols. However, the introduction of new patents implicitly changes the similarity matrix, which leads to a changed result of the projected IPC space. Due to the large number of patents used for the calculation of the initial similarity matrix, the similarity of the IPC classes will not change considerably by the introduction of individual patents. Therefore, it is not necessary to run the projection each time. As the similarity matrix will become increasingly incorrect regarding its projection, the dimension reduction must eventually be repeated, which will likely change the resulting IPC space. If the IPC taxonomy changes, the projection must be performed again, regardless how many patents were added to the dataset, as the new classes are not part of the visualization and some of the already existing patents' classes may be updated with the newly introduced classes.

Both *IPC Cloud* views scale well with the number of patents, as these are not directly presented in the views, but in a separate detail view that necessitates a previous limitation of the interesting IPC classes. If the number of labels shown in the views increases, they are more likely to cause visual clutter due to overlapping labels. Further, it will become increasingly difficult to get a quick overview of the whole IPC landscape if more labels are shown. At last, although

Table 3.1: Scalability of the data storage, data preprocessing, and *IPC Cloud* views in relation to the number of patents and the number of IPC symbols.

	# of patents	# of IPC symbols
<b>Data storage</b>	+	+
<b>Data preprocessing</b>	Search: + Sim. accuracy: ±	-
<b><i>IPC Cloud</i> views</b>	+	-

patent experts usually know the meaning of the IPC subclasses that they often cope with, it may occur that they do not recognize some of the IPC classes.

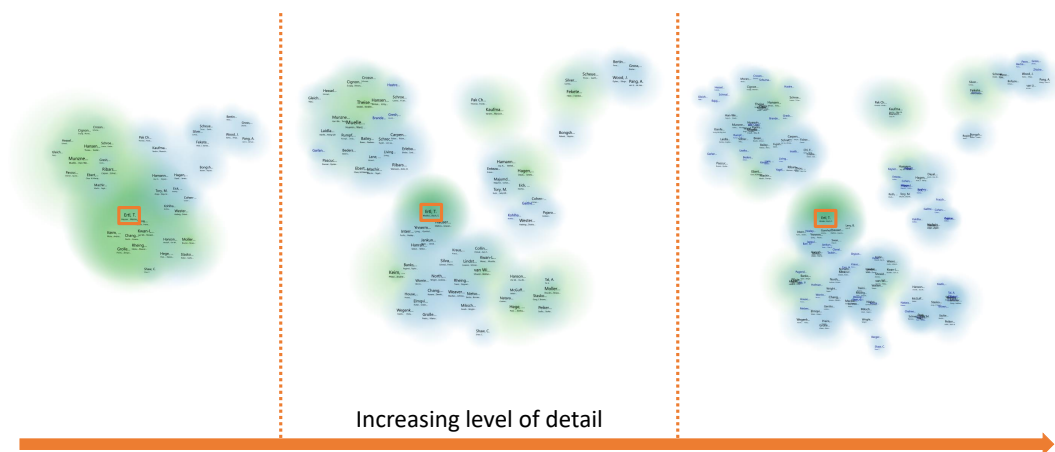
### 3.4 Understanding Topic Relations Through Hierarchized Projection

The approach presented in Section 3.3 allows patent experts to get a first overview of the patent landscape based on the subclass symbols defined by the IPC. However, the shortcomings of the approach regarding the scalability of the presented views causes high cognitive load for the users in case the labels need to be projected anew. Further, the analysis of a specific topic during the design phase of a product is not necessarily limited to patents. It is also important to be aware, which other persons or organizations are influential regarding specific topics and what other topics may also be of interest.

To alleviate these challenges, a new approach was developed based on the lessons learned in *IPC Clouds* (see Section 3.3). The following section presents this reworked approach that

- i) supports arbitrary document-based datasets that provide a similarity matrix comparable to the one presented in Section 3.3.2 (instead of focusing explicitly on the patent domain),
- ii) constructs a hierarchy that is based on the given similarity matrix, and
- iii) visualizes the created hierarchy, while preserving the labels' positions as much as possible.

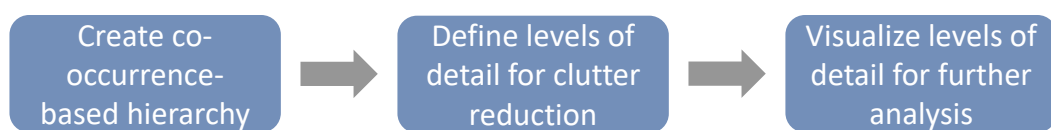
To prove the general applicability of the approach, it was evaluated with datasets from a question-and-answer website and with a dataset about the relations of researchers in the visualization community. Figure 3.6 presents, how users can explore authors from the visualization domain relate to each other based on co-authored papers.



**Figure 3.6:** The approach reduces visual clutter of projected labels by showing the user the relatedness of the labels on increasing levels of detail. Here, the cluster *Ertl, T.*, highlighted with a green background and an orange box, and its related clusters, marked in with a slightly lighter green, are shown in increasing levels of detail. © 2017 Elsevier

### 3.4.1 Approach

The main goal of this approach is to reduce visual clutter caused by the projection of labels onto a two-dimensional plane. This is done by introducing a hierarchy in the projection, which provides a smooth transition between overview and detail. At the same time, the approach aims to preserve and indicate the relationships between labels on different levels of the hierarchy. Those *levels of detail* are based on the hierarchy generated in a preprocessing step. Like before, the distribution of the labels on each level supports the users' intuition that related labels tend to be placed in close proximity. The approach consists of three steps (see Figure 3.7), which will be detailed in the following.



**Figure 3.7:** Workflow of the approach.

### 3.4.2 Data Preprocessing

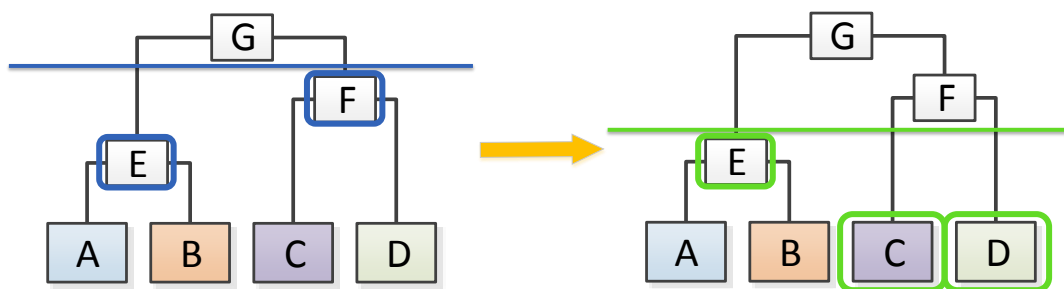
Depending on the measure used to create a hierarchy, this process may be expensive regarding computation time. As there is no need to recompute the hierarchy of a dataset unless the presentable labels (e.g., IPC subclass labels in *IPC Clouds*) change, the hierarchy is computed in a preprocessing step and

then stored for later use. This ensures a quick and smooth entry point into the analysis, as the information about the hierarchy is directly available.

**Step 1: Hierarchy creation.** Hierarchical Agglomerative Clustering (HAC) (see Section 2.1.3) is used to create the hierarchy. Often, HAC-based clustering approaches use single-linkage. However, it is unclear what merging clusters this way means on a semantic level [88, p. 525]. Therefore, medoid-linkage was used as the linkage criterion, wherein the similarity of two clusters is defined by the similarity of the clusters' medoids.

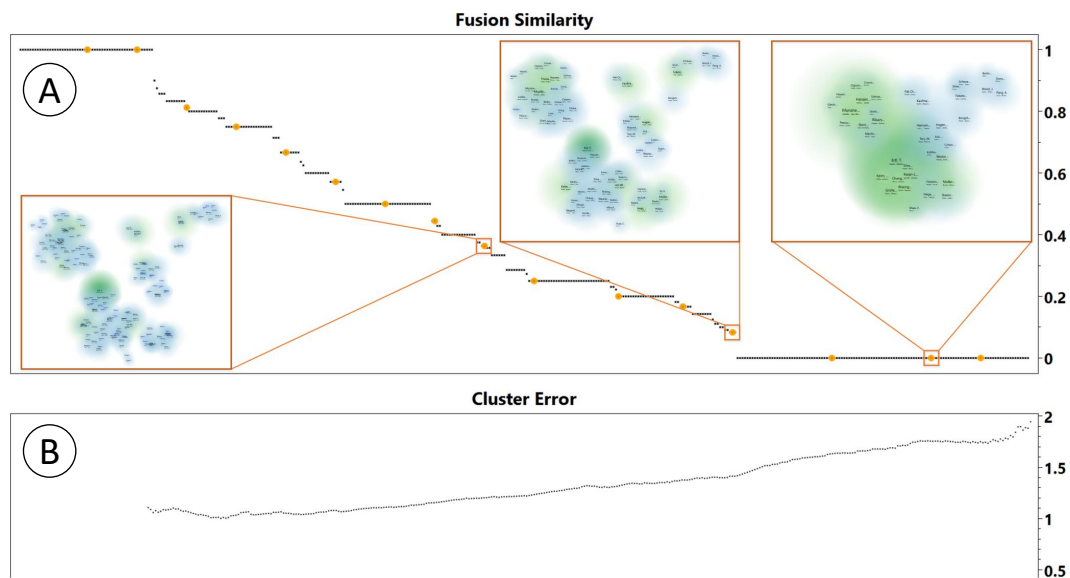
### 3.4.3 Recurring Steps of Analysis

The following steps are executed every time an analysis is performed. As users may be interested in defining different levels of detail for each run of an analysis, the visualized information may change and is therefore computed on demand.



**Figure 3.8:** The labels shown on a given level of detail are similar to a cut in the dendrogram of the hierarchy. The first nodes after the cut represent relevant clusters for that level of detail (they are highlighted by a border in the figure). The labels of these clusters are then used in the subsequent visualization step. © 2017 Elsevier

**Step 2: Setting the Levels of Detail.** The hierarchy construction through hierarchical agglomerative clustering results in a binary branching of the hierarchy. Therefore, a given level of depth of the hierarchy describes the number of shown clusters at the same time. An analysis of the hierarchy by stepping through it one level of depth at a time can be tedious as only one cluster splits apart and is therefore unfeasible. To reach the before mentioned goal, the user can step through multiple levels of depth at once. In this approach, the chosen levels of depth are referenced to as *levels of detail*. A level of detail corresponds to a cut at the level of depth within the hierarchy. Only the clusters directly below the cut are being shown to the user. An example illustrating this idea is shown in Figure 3.8.



**Figure 3.9:** Two plots are initially presented. The fusion similarity (A) indicates the similarity of the clusters that have been merged over the course of the hierarchization. The cluster error (B) shows the clusters' correctness using the Davies-Bouldin index (a lower value indicates better clusters). Also, initial cuts for the levels of detail are proposed to the user. For demonstration purposes, a reduced number of levels of detail are used and small depictions of the visualizations on three different levels were added. The same visualizations are shown in larger size in Figure 3.6. ©2017 Elsevier

Therein, the first cut is after cluster  $G$ , which means that the clusters  $E$  and  $F$  are the next clusters directly below the cut. In the second step, the cut is below  $F$ . Cluster  $E$  remains, but cluster  $F$ , which is now above the cut, is replaced by the clusters  $C$  and  $D$ .

Initially, an increase of 20 levels of depth per level of detail is set. This ensures that each level of detail shows a comprehensible amount of new information. Afterwards, the user may add, change or remove any number of levels of detail. To support the users in this task, two plots that contain information about the hierarchy are shown (see Figure 3.9).

Figure 3.9(A) shows the similarity of the merged clusters. Every point of the plot represents one merge step of the clustering algorithm (or going down one level of depth in the hierarchy). The similarity is expressed in the y-value of the points, whereas the clustering step is equal to the x-value of the points. A higher similarity value indicates that the clusters are much alike, whereas a low value indicates little overlap regarding the co-occurrence of clusters' medoids.

Figure 3.9(B) indicates the quality of the clustering at a given step. As an indication measure, a variation of the Davies-Bouldin index [52] is used. The Davies-Bouldin index is designed to have a low value when the distance between clusters is high and the distance within the clusters is low. This step assumes

that the similarity value  $s$  is normalized and calculates the distance  $d$  of two clusters  $C_i$  and  $C_j$  as  $d_{i,j} = 1 - s_{i,j}$ . The index was slightly adapted, since the approach uses medoid-linkage, whereas the Davies-Bouldin index uses centroid distances. The modified Davies-Bouldin index  $DBI_{mod}$  is calculated as

$$DBI_{mod} = \frac{1}{n} \cdot \sum_{i=0}^{n-1} \cdot \max_{i \neq j} \left( \frac{\sigma_i + \sigma_j}{d_m(C_i, C_j)} \right),$$

with  $i$  and  $j$  being the cluster indexes,  $\sigma_x$  representing the average distance between the elements within cluster  $x$  and  $d_m(C_i, C_j)$  being the distance between the medoids of the clusters  $C_i$  and  $C_j$ .

This way, the users get a visual insight into the clustering and can decide whether an adjustment of the levels of detail is necessary and useful. For instance, a sudden change of the similarity or of the Davies-Bouldin index could be a reason for an adjustment.

**Step 3: Visualizing the Clusters.** In the third step, the users can visually explore the hierarchy and the relationships of the shown clusters on increasing levels of detail. At first, the labels from the topmost level of detail are taken and projected onto a two-dimensional space, which is henceforth called *projection view*. Analogously to Section 3.3, the data is projected using the t-SNE projection technique. T-SNE is used as it is possible to apply it with or without an initial spatial mapping of the data.

For every level of detail, a separate distance matrix that is used by t-SNE is created. The cost to traverse the hierarchy tree between all shown clusters is calculated and the results are used as the distances between the clusters. Once the positions of the clusters have been determined, representative labels can be shown in the projection view. Details about the projection and positioning of the labels are presented in Section 3.4.4. Several key aspects of the data and its structure are encoded into the projection view:

**Representative of the cluster.** Every cluster is represented by a label with three text boxes in two rows. Following the recommendation of Elmqvist and Fekete [58], the elements of the cluster with the highest overall occurrence frequency are used as the cluster’s labels. The upper row shows the element with the highest overall occurrence frequency with its according font size (see below). As showing only one label to describe a term may be ambiguous or even misleading, the second and third most important elements are shown in the second row. Their font size is half of the font size of the most important element.

**Font size for importance.** As the font size is a prominent visual aspect, it is used to indicate the importance of a cluster, similar to word clouds. Since clusters usually consist of several elements, the accumulated occurrence frequency of the elements is mapped to the font size. The importance of the clusters on the

currently shown level of detail is represented by normalizing the accumulated frequencies of all shown clusters.

**Numbers of elements within a cluster.** To further indicate the distribution of the elements across the clusters, a colored radial background is added to every label. The size of the radius corresponds to the number of elements that are contained within the cluster. The color scale is a linear gradient that starts with a light blue in the center and fades out towards the outside. In case the cluster is selected or marked to be relevant, the light blue is replaced by a dark or light green (see Section 3.4.5).

**Cluster density.** If two or more clusters overlap, their colors add up, resulting in a heat map-like visualization. This helps users to get a better visual impression of the clusters' distribution. The radius of the clusters remains constant when the users zoom in or out of the projection view. Thus, smaller clusters are aggregated into bigger ones in a zoomed-out view.

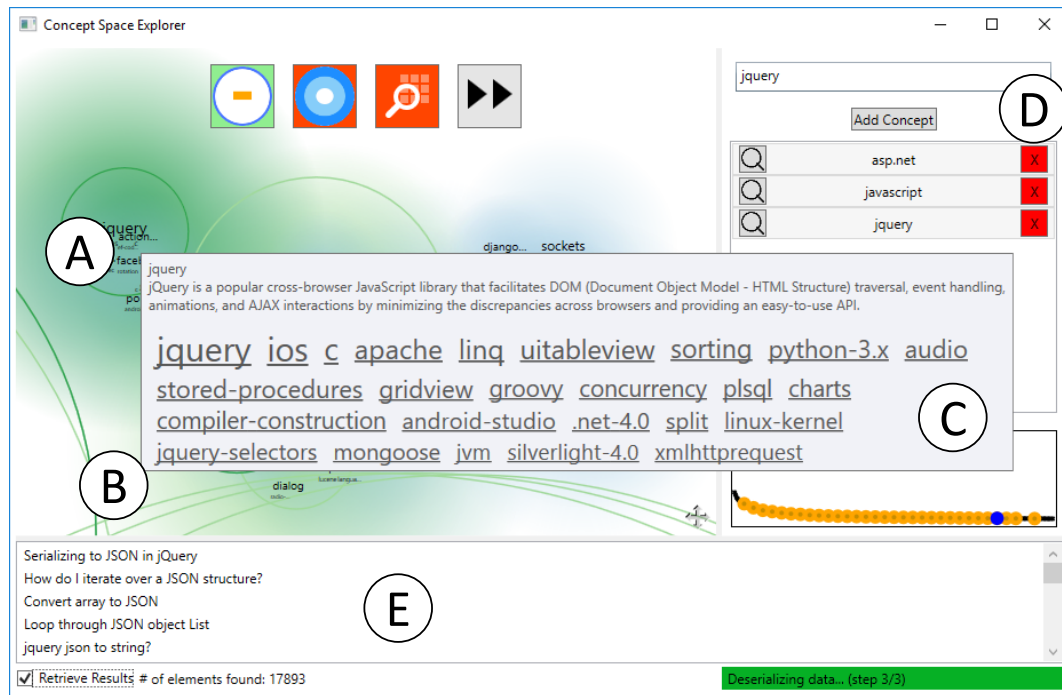
**Position of the labels.** As in *IPC Clouds*, the position of the labels follows the Gestalt principle that visual closeness correlates with similarity. More precisely, in t-SNE, the closeness of two elements indicates the likelihood that two elements are related to each other. This aspect is incorporated implicitly by using the similarity as the projection measure in the second step of the approach. To keep the cognitive load low, it is important to keep the positions of the already visualized labels as stable as possible when changing the level of detail. The positioning of the labels is being detailed in the following section.

### 3.4.4 Switch Between Levels of Detail

To view the data on different levels of detail, the users need to be able to switch between the chosen levels. In this approach, the new clusters and the changes of the map are presented through an animation. Whenever a visualization changes over time, it is important to minimize the cognitive load of a user. However, projection algorithms are not designed to support the iterative projection of a dataset with subsets of increasing levels of detail and do not make use of previously positioned clusters. Nevertheless, users should be supported in understanding the changes between levels of detail. It is necessary to visually inform the user about the relation of the newly available subclusters in the context of the already visualized clusters. The construction of the map when switching levels of detail should therefore minimize the movement of unchanged data. In this approach, new data points are first drawn at their respective parent cluster's position and then moved with an animation to their final destination.

At the first level of detail, there is no previous projection result that needs to be considered. Therefore, the projection is performed using the standard implementation of t-SNE. For subsequent projections, several parameters used





**Figure 3.10:** Screenshot of the prototype showing data from the question-and-answer website StackOverflow. The projection view (A) shows the projected level of detail's clusters. Some clusters are selected and highlighted with a dark green background. All selected and related clusters are indicated by a green background color and halos. Some clusters are not visible in the focused viewport, but the halos indicate their position (B). A tooltip (C) shows the content of the focused cluster *jquery*, including its name, description and a word cloud with the most frequent elements contained in that cluster. The selected labels are also shown and highlighted in the search component on the right (D). On the bottom (E), questions are listed that contain at least one element from every selected cluster. © 2017 Elsevier

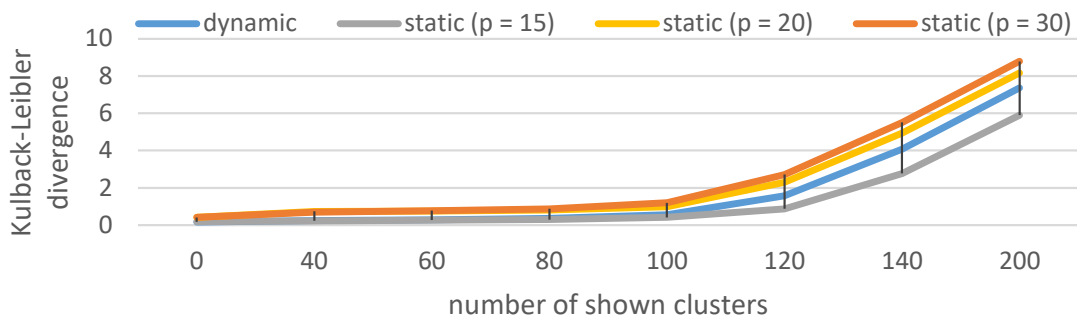
by t-SNE can be adapted to stabilize the subsequent projections, which are explained in detail in the following:

**1. Initial map.** Initially, t-SNE distributes all points that need to be projected with a Gaussian distribution and optimizes this state. However, it is also possible to initialize the map manually, for example, with the results of a previous projection step. To do this, all clusters that split into smaller clusters in the next level of detail are retrieved. Unchanged clusters keep their position and all newly generated clusters inherit the position from the cluster they originate from. The resulting map is used as the initial map of t-SNE, replacing the Gaussian distribution of the clusters, which reduces unnecessary changes in the positions of the clusters.

**2. Perplexity.** Further, the perplexity parameter used by t-SNE can be adapted to the visible clusters, as the perplexity can be interpreted as a factor

that controls the size of the neighborhood used during the mapping of each point (see Section 2.1.3). Typically, this is a constant value that must only be adapted, if the results are not satisfying. Van der Maaten and Hinton [187] recommend a value between five and 50 for this parameter. Usually, this value is robust and small differences do not heavily impact the algorithm’s results. However, in the case of a modified initial mapping, a too high or too low perplexity value can lead to visual artifacts. In case the perplexity is too high, the visualization will look like all clusters center around one point because every cluster tries to optimize with regard to all other clusters. If the perplexity is too low, the clusters will not move at all because they only consider themselves as relevant and therefore all newly added clusters overlap at their parent’s position. To keep the user from following a trial and error approach to find a proper value, the perplexity value is dynamically approximated depending on the number of shown clusters. The used perplexity function  $p(x) = 6.929 \cdot x^{0.252710}$  was designed to initially increase fast in order to ensure that clusters in the first few levels of detail already consider some of their neighboring clusters. The more clusters are shown, the lower the increase of the perplexity is (compared to the previous level of detail). The values are based on the results that were achieved with the datasets described in Section 3.4.6.

The function was compared to static perplexities on different levels of detail. The Kullback-Leibler divergence was used as a performance measure, as it is also the function minimized during the projection and the authors of the t-SNE algorithm propose to use it as the statistical quality measure. The results based on the dataset of use case 2 are shown in Figure 3.11. It becomes clear that the function-based perplexity performs slightly better than most static perplexities. Although the static perplexity of  $p = 15$  performs better regarding the projection’s divergence, a lot more overdraw of the labels can be seen. This issue likely amplifies in case the perplexity is lowered further.



**Figure 3.11:** Comparison of the used dynamic against static perplexities. ©2017 Elsevier

**3. Scaling of target projection space.** Once the clusters are projected, their positions are min-max normalized. To prevent a shift of all clusters due

to outliers, the clusters' geometric center is calculated. Afterwards, all clusters are normalized and the 10% of the clusters that are the farthest away from the center are discarded. Then, the target projection space is rescaled based on the number of visible clusters using a root function.

### 3.4.5 Explore Cluster Relationships

In case the users want to explore a given level of detail, there are two scenarios: either they already have an initial idea what they are looking for, or they want to get an overview of the chosen level of detail without any premises. In both cases, the first step is to decide on interesting clusters to inspect.

#### Search for Specific Clusters and Elements

The users can use a search box that provides an autocomplete feature, suggesting every partly matching label contained within the dataset. Once an element has been found, it can be added to the list of topics of interest below the search box (Figure 3.10 (D)). The selections made in the topic list are linked to the projection view (Figure 3.10 (A)).

Sometimes the searched element is not visible in the projection view because it is hidden within a cluster. In this case, the cluster that contains the element is selected. It is also possible to focus on the element or cluster that contains the element in the center of the view with the focus button on the left of the label and to remove them from the list with the button on the right.

#### Freely Explore the Projection View

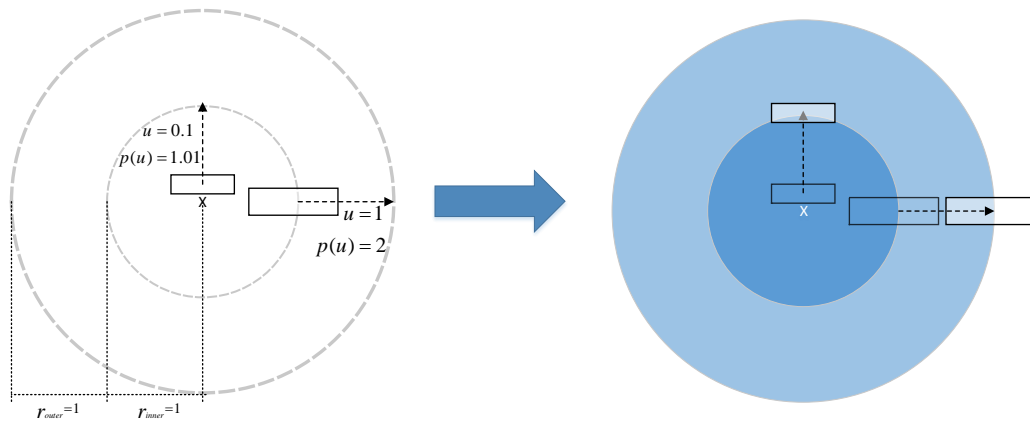
The projection view supports basic navigation interactions through zooming and panning to let the users freely explore the visualized level of detail. As described in Section 3.4.3, the size of the background color of the clusters is independent of the visual zoom. This way, the clusters are visually aggregated when the user zooms out, as the size of the cluster backgrounds increases compared to the labels of the clusters. This helps users to distinguish between areas with a higher cluster density and regions that are sparser.

The chance of an overlap of the shown labels increases with the number of shown clusters. In case the users want to inspect or select the content of overlapping clusters, they can use a lens that spatially distorts the labels below the lens. The lens consists of an inner and an outer segment. Only labels within the inner segment are moved, but they also be relocated to the outer segment of the lens. The two segments make it easier to see, which part of the lens is distorted and allows users to select clusters more easily. The spatial distortion effect of the lens is similar to the optical distortion effect of a fisheye lens.

However, instead of enlarging elements closer to the center, they are pushed away from it stronger than elements that are farther away. The push force of the lens is based on an Epanechnikov-Kernel [60], which is used in statistics for density estimation. The kernel was modified to limit the maximum from  $\frac{3}{4}$  to  $\frac{1}{2}$ . This kernel increases the lens' push force with a behavior similar to a Gaussian kernel, but the kernel's size is limited. The proportional distortion  $p_{lens}$  between the center of the lens and center of the cluster is calculated as

$$p_{lens}(u) = \underbrace{\left(1 - \frac{1}{2}(1 - u^2)\mathbf{1}_{|u|\leq 1}\right)}_{\text{Epanechnikov-Kernel}} \cdot \underbrace{\frac{r_{inner\ lens} + r_{outer\ lens}}{r_{inner\ lens}}}_{\text{scaling factor}}$$

wherein  $u$  is the original, unmodified proportion between the center of the lens and the center of the cluster compared to the inner radius of the lens. This way, the center of the lens is free of labels and the labels themselves are located around the center. The push effect with one dimension is shown in Figure 3.12.



**Figure 3.12:** Schematic depiction of the distortion behavior of the lens. © 2017 Elsevier

Once the users decide on one or more interesting clusters, the clusters can be selected in the *projection view*. The selected elements will then be added to the aforementioned topic list (Figure 3.10 ©).

### Navigation Aids in the Projection View

Once a set of clusters and/or elements is selected, the users may be interested in other clusters related to the selected ones. However, due to the information loss during the dimension reduction, this information may not be available directly. The approach distinguishes between two kinds of relatedness: *global* and *local*. The global relatedness is depicted by the spatial arrangement of the clusters, wherein closer clusters are more likely to be related than clusters that

are farther apart. The local relatedness is available for selected clusters and shows, which clusters are similar to the selected based on the high-dimensional data. To measure the local relatedness, the average similarity  $s_{avg}$  between the set of selected clusters and the cluster they are compared with is calculated:

$$s_{avg}(C_{other}) = \frac{1}{n} \sum_{i=0}^{n-1} \left( \frac{1}{m} \cdot \sum_{j=0}^{m-1} \frac{1}{o} \cdot \left( \sum_{k=0}^{o-1} sim(C_{i,j}, C_{other,k}) \right) \right),$$

where  $i$  is the index of the cluster within the cluster set,  $j$  is the index of the compared element within the cluster and  $k$  is the index of the element within the cluster that is compared with the set of selected clusters.

As the locally related clusters may lie spatially apart from the selected clusters, several visual cues help to indicate their positions.

**Halos.** Users are provided with *halos*, introduced by Baudisch and Rosenholtz [22], to find clusters that are related to the selected clusters. Halos indicate the position of the selected and other related clusters by drawing a circle around these clusters. In case the cluster is outside of the visible area, the radius of the halo is expanded, so it stays within the view. The curvature of the Halo fragment indicates the direction and distance of the corresponding clusters. An example of the visual indication provided by halos is shown in Figure 3.10 (B).

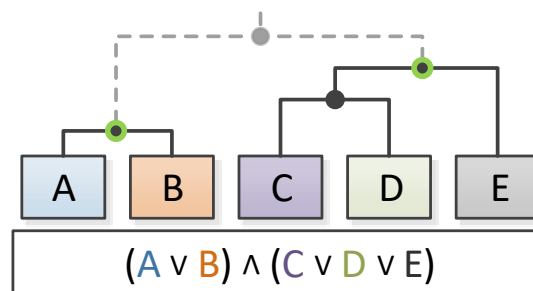
**Color coding.** To make selected, related, and other clusters distinguishable, the backgrounds and halos of selected clusters are colored with a noticeable dark shade of green. Analogously, the backgrounds and halos of related clusters are shaded with light green. The background of the other clusters is drawn in a light blue. Color coding was used instead of approaches such as isolines or glyphs because the color map indicates the uncertainty of the positions of the clusters and retains the different states of the clusters.

**Darts View.** Furthermore, users can quickly spot related clusters using the *Darts View* presented in Section 3.3.3.

Figure 3.5 and Figure 3.10 (A) depict how these aspects look like in the implemented prototype.

### Analyzing a Cluster's Content

Once the users found a relevant cluster, they may be interested in further information about it. They have several options to inspect its content. First, the users may request additional information about the cluster by looking at its tooltip. In case only the name of the elements is available, the tooltip shows the name of the representative label as well as a word cloud. The labels shown in the cloud belong to the elements contained in the cluster. They are ordered by



**Figure 3.13:** For each selected cluster (here indicated with a green circle), the leaf nodes are retrieved. Then, data is requested from the dataset, which contains at least one of the leaf nodes from every selected cluster. ©2017 Elsevier

their weight ( $\hat{=}$  occurrence frequency), which is also encoded into the font size of the words. If it is available, the textual description of the representative label is shown at the top. The word cloud contains the terms that occur most often in the descriptions of all elements. An exemplary depiction of the tooltip is shown in Figure 3.10©.

Second, the user can select one or more clusters. When doing so, the individual documents that contain at least one term from each selected cluster are retrieved. A schematic demonstration of such a request is depicted in Figure 3.13. Here, the two highlighted clusters were selected. As the elements *A* and *B* are contained within the first cluster, only one of them has to be part of the resulting document. By selecting an individual document, further details can be retrieved.

### 3.4.6 Use Cases

The following illustrates the applicability and usefulness of the approach with two scenarios using real-world data.

#### **Use Case 1: Analysis of StackOverflow**

In the first use case, we assume the role of the head of a newly founded department of a software company that used to develop server-sided software solutions. Our new department has to supplement a client-sided component to the software suite. The company already supplied us with some of its software developers, but we still need to recruit a programmer that is well versed with web technologies. As we only possess basic knowledge of this field, we need to take an explorative approach in order to understand, which aspects are important. To create a list of key skills for the profile of our recruit, we use the presented approach to explore the tags and relations of the question-and-answer website *StackOverflow*. Since StackOverflow contains data that comprises more than ten million questions, we

limit our inspection to tags that have been assigned to at least 5000 questions. The similarity between the tags is based on the Jaccard coefficient (see Section 2.1.3).

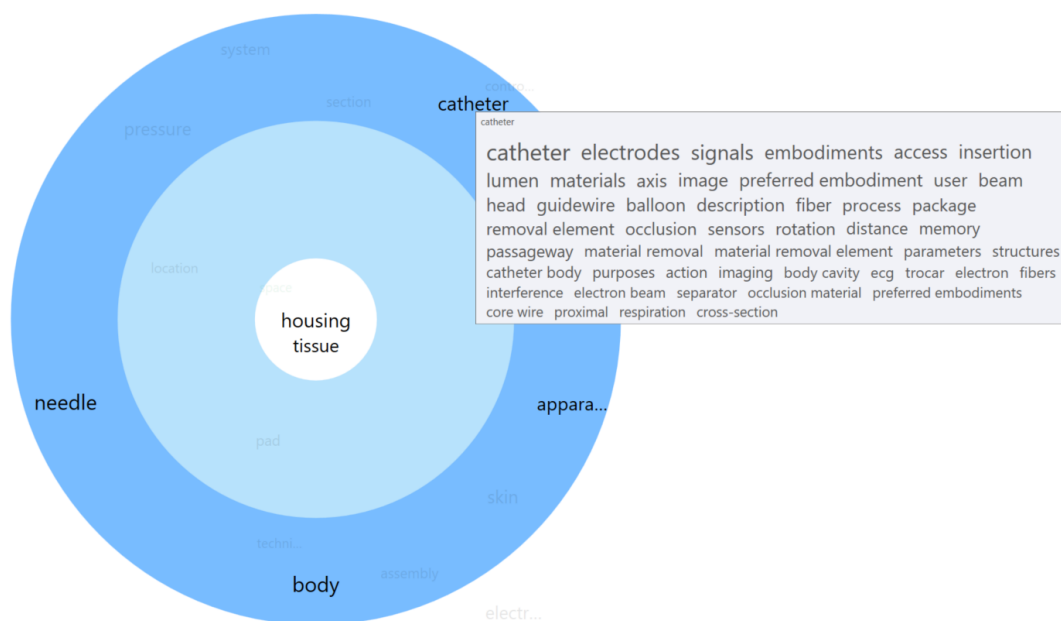
We decide to remove the first four proposed levels of detail because we want to see some detailed information at the top level already. We know that our company's server-sided software returns JSON formatted answers and JavaScript is a common solution for web clients, so we select the clusters containing *JSON* and *JavaScript*. To see if our interview candidates have some background knowledge, we note some of the most often viewed questions in StackOverflow that contain elements from the JSON and JavaScript clusters. After selecting the clusters, the tool marks the cluster *jQuery* to be relevant. Once reading the description of the tag in the tooltip and looking at its contents, we decide to add it to our profile by selecting it (Figure 3.10 ©). We write down questions such as 'How to iterate over a JSON structure?' for the interview of candidates for our team. At last, we freely explore the space around our selected clusters and notice the cluster *d3.js*. After looking at the description of the cluster description, we decide to add *d3.js* to our profile, as a developer with knowledge in web-based visualization may be useful in later software development stages. By now, we have a profile that we can use to search for a software developer that has knowledge in key technologies used in web development, which fits our companies existing technologies.

### Use Case 2: Patent Analysis

In the second use case, we take the point of view of an intellectual property analyst at a large company that produces and sells medical apparatuses. The company plans to develop a new product to apply an adhesive to open tissue during a surgery and wants to avoid conflicts with existing patents.

Our task is to limit the number of patents that have to be inspected individually to a reasonable number. First, we reduce the number of patents by searching for an IPC class, which specifies medical surgery or diagnosis. Also, the patents need to contain the keyword *glue* or *adhesive* in their description. This results in a patent set of about 300 patents, which are too many to analyze manually. Therefore, the patents are automatically processed and concepts are extracted using the workbench presented by Brüggmann et al. [37]. The similarity between concepts is calculated by measuring the co-occurrence of the concepts within sentences and comparing them using a cosine similarity (analogous to Section 3.3). Then, the concepts are loaded into the presented prototype to reduce the number of patents even further.

We could change the initially proposed levels of detail, but in this case, we leave the levels as proposed. Once the first level is projected, we search for the clusters that contain the concepts *adhesive* (which is hidden inside the cluster *housing*) and *tissue*. By doing so, we notice that the concept *catheter* is marked to be relevant to the selection. This becomes even clearer as we switch to the



**Figure 3.14:** Depiction of the concepts related to the clusters *adhesive* (contained in *housing*) and *tissue*. There are five concepts that are strongly related to those clusters. One is *catheter*, which makes sense, as a catheter may be used to apply the adhesive. © 2017 Elsevier

parts view, which is shown in Figure 3.14. As it makes sense to apply an adhesive with a catheter, we add it to the set of selected clusters.

When looking at the resulting patents, we notice that the results are not precise enough yet. Therefore, we add the cluster containing the concept *apparatus* to the selected concept list and increase the level of detail several times. By doing so, the selected clusters get more specific as dissimilar elements split apart, which results in a more specific patent request. By iteratively increasing the level of detail and inspecting the results, we are eventually left with 27 potentially relevant patents, which is a reasonable number for a manual analysis.

#### 3.4.7 Evaluation

A qualitative user study was conducted to evaluate the effectiveness and comprehensibility of the presented approach. The approach was compared with the unmodified projection results of t-SNE, i.e., without any introduction of a hierarchy of the data. The studied hypothesis was that the introduction of a hierarchy helps the users to get a faster overview and a better understanding of how the mapped clusters and individual labels are related.



#### Participants

The study was conducted with ten expert users between 31 and 42 years of age (mean age: 34.9 years). All of them had a university degree in either computer science or mathematics, which reflects the user group of academic professionals who may be required to use a visualization approach like the one presented. The participants were asked to specify their expertise in visualization and programming as well as their familiarity with dimensionality reduction techniques. They had on average 15.4 years (SD: 6.2) of programming experience and 7.3 years (SD: 3.5) of experience with visualization techniques. All but one of the participants had some knowledge about dimensionality reduction. The familiarity with this topic was given with 3.6 years (SD: 1.3) on average on a scale ranging from “no knowledge” (1) to “expert knowledge” (6).

#### Materials and Procedure

The study was conducted in a closed room, with one participant at a time. The prototype ran on a Lenovo W540 Laptop with an Intel Core i7 processor, 32 GB RAM, and an SSD hard drive. It was shown on a 24” monitor in full screen with a resolution of 1920x1080 pixels.

First, participants were asked to provide some demographic data in a questionnaire. They were asked for their gender, age, educational degree, current job position, and their research area of expertise. Additionally, they were asked about any previous programming and visualization experience in years and their self-assessed experience with dimension reduction. Subsequently, they were familiarized with the basic ideas of the approach. It included the motivation of dimensional reduction, the challenge of showing labels in a projected space (in contrast to points), and the introduction of a hierarchy to create an overview to detail the approach. Then, the implementation was briefly explained to them using the patent dataset presented in Section 3.4.6. After the explanations, the participants were invited to use the prototype themselves to get familiar with it.

Each participant was shown both the prototype implementing the hierarchical projection approach and a reference implementation that skipped the hierarchy and directly projected all of the labels at once with a static perplexity. For the remainder of this section, the latter will be referred to as *flat projection*. That way, the participants were able to compare between the two approaches. Participants were assigned to one of two groups in order to balance carry-over effects that may result from remembering tasks and/or answers. In addition, the order in which the participants used the two projection approaches was counterbalanced.

Two datasets were prepared that the participants were asked to analyze: the StackOverflow dataset presented in Section 3.4.6, and a dataset about IEEE

Table 3.2: A  $2 \times 2$  mixed study design with visualization type (flat vs. hierarchical) tested as independent variable within subjects and two conditions resulting from counterbalancing the order of the visualization types tested between subjects were used. ©2017 Elsevier

group	$vis_1$	$dataset_1$	$vis_2$	$dataset_2$
$G_1$	flat	StackOverflow	hier.	VisWeek papers
$G_2$	hier.	StackOverflow	flat	VisWeek papers

Visualization publications since 1990 [100]. In the latter case, the similarity of the authors of the publications was visualized based on co-authorship information derived from the dataset. In particular, the similarity was calculated as

$$sim(A_a, A_b) = max\left(\frac{\#pub(A_a \cap A_b)}{\#pub(A_a)}, \frac{\#pub(A_a \cap A_b)}{\#pub(A_b)}\right).$$

The maximum of the similarities of any pair of authors ( $A_a$  and  $A_b$ ) was taken to compensate for the effect that authors who wrote papers with many different authors automatically get a low similarity to all of them, even if they were from the same research group. This alleviates the similarity issue, but it does not solve it as two seasoned authors will still have a low similarity when they are directly compared to each other. Table 3.2 shows in what order the two groups saw the compared visualizations and what dataset was used.

In sum, a  $2 \times 2$  mixed design was used with visualization type (flat vs. hierarchical) tested as independent variable within subjects, and two conditions resulting from counterbalancing the order of the visualization types tested between subjects. Participants were randomly assigned to one of the two groups. The participants were asked to verbalize their thoughts as they were solving the tasks, and their statements were noted (*think aloud protocol*).

Finally, the participants had to complete a questionnaire to gather information on which parts of the visualization caused confusion and which elements were helpful for understanding the dataset. The participants were asked to rate a number of prepared statements about the approach and its implementation:

1. The introduction of the hierarchy levels was...
2. The animated transitions between the hierarchy levels were...
3. The navigation support through the halos was...
4. The color coding of the clusters and halos was...
5. The interactive lens to reduce overlappings was...
6. The implementation with the hierarchy-based projection was...

7. The implementation with the flat projection was...

A Likert scale ranging from “not helpful” (1) to “very helpful” (6) was used to rate the statements. In addition to the scale, the participants were also allowed to give the answer “I do not know” in case they had no opinion on the statement for any reason. Furthermore, the participants were asked to explain their reasoning for the picked answer orally.

#### Tasks

The participants were asked to answer five questions using the provided visualizations. The questions were slightly altered depending on the given dataset.

1. Which authors/tags are important, and where are they located in the currently shown hierarchy level?
2. Which groups of authors/tags are important and what characterizes them?
3. Which subgroups of authors/tags are contained within the previously named groups?
4. Where is the group with the author “Ertl, T.” / the tag “Java” located?
5. Which groups are similar to this group?

#### 3.4.8 Task-based Results

In both groups, the participants stated that the hierarchical view is better suited to solve the first task (important authors/tags) compared to the flat projection because the clusters reduce the amount of information initially shown to the user. Every participant was able to resolve the task with the hierarchical view by searching for the label with the biggest font size at the highest level of the hierarchy (which was 20 labels by default). In the flat projection, none of the participants were able to solve the question, as there were too many labels to find labels that are considerably bigger than others. Most of the participants stated that they would have to search through all of the clusters and compare their size, which they considered too much effort.

For the second task (important clusters), most of the participants stated that the results in the hierarchical view are the same as for the first task. Most participants explained that it looks like the cluster size and the most important authors/tags in a cluster often correlate. To characterize the content, the participants mostly used the tag cloud provided in the tooltip.

The flat view does not contain any easily distinguishable visual encoding for the cluster size (as all “clusters” contain one element). Thus, the participants

either tried to zoom out and searched for the areas where most of the background circles overlap or they estimated which area contained most elements by looking at the amount of shown text. Both of those actions took the participants longer than reading the encoding of the circle size in the hierarchical view.

Many participants implicitly answered the third task (subgroups of previously inspected groups) when they characterized the content of a cluster in the second task. When they actually read the task, most of them increased the level of detail in order to confirm their statements from before. In case of the flat projection view, most participants distinguished between groups and subgroups by looking at the visualization and visually cluster the labels depending on their distance.

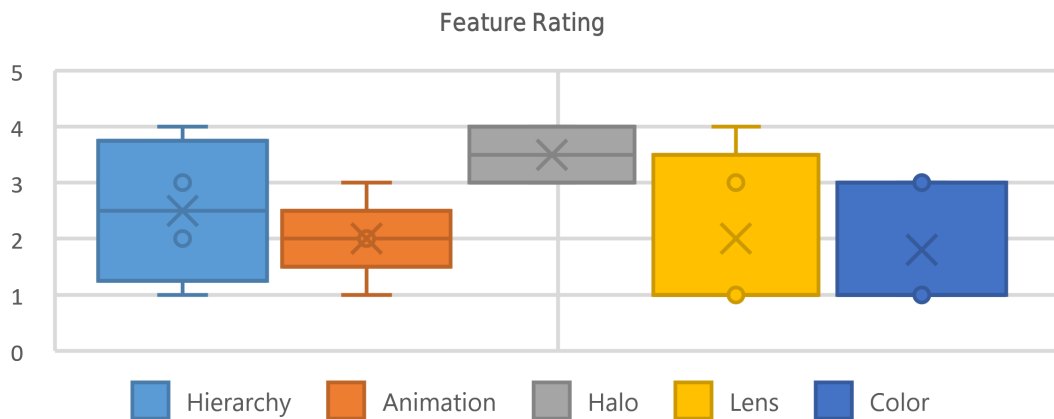
The approach of the participants to solve task 4 (find the cluster of a certain tag or author) was the same for both datasets and visualizations: In case they saw the tag or author before, they tried to find it again through zooming and panning. If they did not see the group before, some still tried to find it in the projection view for about ten to 30 seconds before using the search field and the focus button.

The last task asked the participants to find clusters similar to the cluster they had to locate in task 4. Instead of selecting the cluster to get similar clusters highlighted, most participants assumed that the clusters close to the viewed one are likely to be the most similar clusters. Only four participants selected the cluster to search for similar clusters. Two of those additionally used the halos to find the relevant clusters more quickly.

Overall, the hierarchical view outperformed the flat view in tasks that required to summarize the characteristics of the clusters and the relations of the clusters. One of the main reasons for this observation is that the participants had to find clusters visually in the flat view, as it does not contain any clusters on a logical level. This did not only take time, but it also decreased the accuracy of the answers, as the visualization may not be entirely correct due to the information loss in the process of the dimension reduction. Noticeably, some of the participants still assumed that the proximities of the clusters directly reflect their similarity. However, due to the information loss during the dimension reduction, this is not the case. Therefore, it can be concluded that the visualization needs to be extended with a visual cue to indicate the location of the relevant clusters compared to the individual clusters.

### **Further Results**

In addition to the oral feedback during the analyses of the datasets with the prototype and the reference implementation, some information about the usefulness of the features was collected through a questionnaire. Figure 3.15 summarizes the rating of the evaluated features. The participants were first asked about their opinion of the introduction of a hierarchy and the configurable step choice



**Figure 3.15:** Results about the usefulness of the prototype’s features that were evaluated in the questionnaire. The used Likert scale ranged from one for “strongly disagree” to six for “strongly agree”. The results suggest, that most features were well received, especially the introduction of a hierarchy. However, there is some disagreement about the usefulness of the color mapping. © 2017 Elsevier

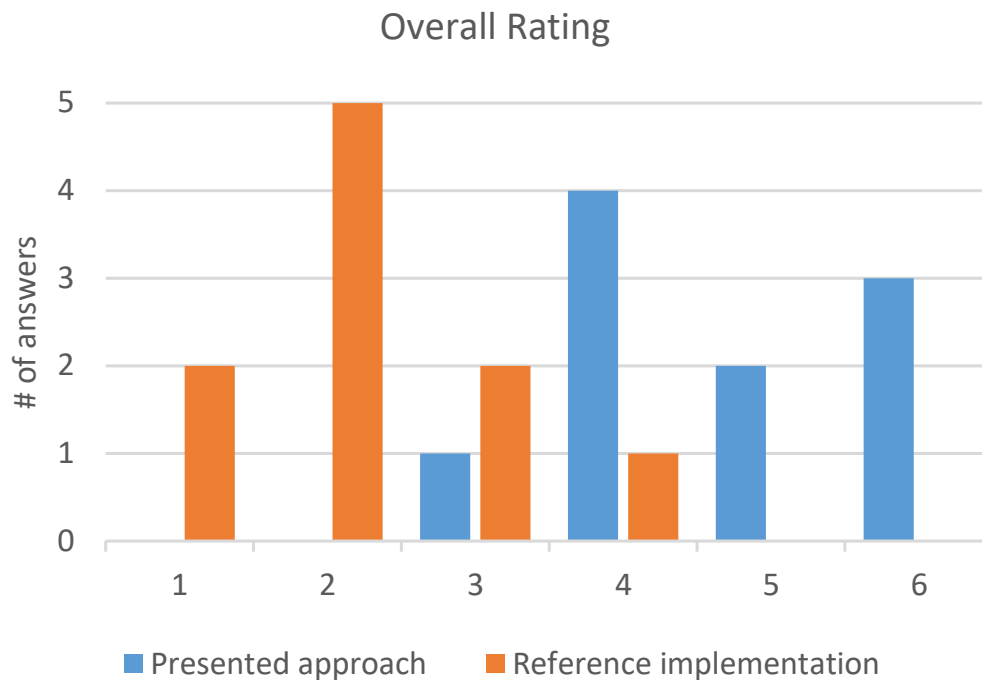
(statement 1). Nine of the ten participants rated the usefulness between four and six points on the Likert scale. Some further stated that the usefulness of the introduction of a hierarchy likely depends on the dataset. This aligns with observations during the study, where the participants had an easier time to make sense of the clusters in the VisWeek dataset compared to the StackOverflow dataset (which had a long tail distribution with regard to the clustering similarity).

The ratings about the usefulness of the animations when switching the level of detail (statement 2) were almost evenly distributed between the scores of two and six points. Most of the participants that gave a mediocre score stated that it is possible to track the changes of a specific cluster, but it is very hard to follow the changes of many clusters at once. Some proposed to give the users some kind of visual cue on where the clusters were located in the previous level of detail.

Three participants had no opinion on the usefulness of the halo feature as they did not use it. The others gave a score of either four or five points. Some of the participants stated that the halo function may be helpful in general, but it takes time to get used to its availability.

The ratings about the usefulness of the color coding were mixed. Four participants gave a score of two or three points and the others gave it between four and six. Some of the participants argued that the colors for selected and related clusters are too similar and therefore not as useful. Others said that the created heatmap effect is very helpful to distinguish areas with few from areas with many clusters and that it gave a good overview of the overall structure.

Eight of the ten participants gave the usefulness of the lens (statement 5) a score between four and six. However, independent of the rating of the lens, most



**Figure 3.16:** Overview of the overall ratings given by the study participants for the presented approach and the reference implementation with the flat projection. The ratings are on a Likert scale ranging from one (“not helpful”) to six (“very helpful”). Nine of the ten participants rated the implementation with the hierarchy between four and six. On the contrary, nine of the ten participants rate the implementation without a hierarchy between one and three. ©2017 Elsevier

participants proposed to change the behavior of the scaling effect to something that a) keeps the local structure intact and b) guarantees an overlap-free view at the same time, for example by showing the result of the distortion of the lens in a separate view.

At last, the participants were asked to rate the presented implementations with an overall score. The results are shown in Figure 3.16. The ratings suggest that the approach was generally well received, although there is still space for improvement. The reference implementation was not as well received, which is likely caused by the high amount of overdraw that was caused by the high number of shown labels.

The participants were also asked, if and where they could imagine to apply such an approach. Most of them saw an application scenario in their own field of expertise. For example, one participant who works in the field of eye tracking said that the approach can be used to identify groups of participants during studies which have a similar behavior. Another participant said, that the approach would be suitable to get an initial overview of text corpora.

Overall, the participants found the approach useful, especially for exploring datasets where the user is only familiar with the general topic that the dataset is about. Aside from the choice of colors, the heatmap-like visualization was well received and mentioned to be helpful in getting an overview of the projected structure of the datasets.

#### **Discussion**

When analyzing a new dataset, the users are confronted with a chicken and egg problem: on the one hand, the zoom levels are supposed to help users in understanding the possible relations within the dataset by only showing information on a very coarse level. On the other hand, without knowledge about what levels of detail may be interesting, it is difficult to decide how many levels of detail should be shown and what information granularity they should have. This approach addresses this dilemma by proposing the users a preset number of levels of detail that can be used as a starting point for the exploration. Further, the presentation of the error index of the clustering to support an easier manipulation of the levels. However, the shown information is of little help if the users are interested in which clusters are actually merged at a specific step. Even in case the users had that knowledge, it is still impossible to infer any further knowledge about the previous or the next cluster fusion step. Moreover, the users only get an abstract measure of the quality of the clustering, which is hard to comprehend. This problem may worsen when the binary tree of the clustering algorithm is simplified, for example, by using Bayesian Rose Trees [31]. In that case, it is hard for users to comprehend how many clusters will be shown in the next level of depth of the tree, as every step may imply multiple aggregation steps.





## Visual Analytics for Production Line Layout Planning

The next step after a product is designed is its production. However, in most cases, it is not possible to produce a new product with the exact layout setup and machinery that is currently available in a factory. Therefore, a production line needs to be modified to support the new product or a new production line needs to be designed. Generally, there are two scenarios when designing a factory's layout plan or individual production lines. In a *greenfield* scenario, the entire factory can be designed from the ground up. In a *brownfield* scenario, the production needs to be incorporated into already existing structures, ranging from already existing buildings down to already running production lines.

In addition to these planning constraints, it is important to consider what type of production is suitable for the goods that will be produced. For custom-made products, the most suitable setup often consists of individual manual work stations wherein workers process the product and then move it to the next station. It is typical for such setups that many process steps are performed by workers. In case some of the process steps can be automated, a u-shaped layout can be deployed to keep the walking distance between the stations short so that workers can operate multiple stations without bringing the overall production to a halt. For products with a limited variety that need to be produced in large numbers, linear production lines are an established option. An important factor for choosing the production setup is how adaptable the production needs to be. On the one hand, manual working stations are usually very adaptable, for example regarding their product range and the positioning of the individual stations. On the other hand, production lines can produce more effectively, but they support only a limited product range that needs to be known before the line is deployed and most of the parts in a production line cannot be repositioned.

The following approaches assume a brownfield scenario. These often face additional challenges, e.g., missing or partly incorrect information about the

measures of existing rooms, building supports, or the positioning of power or water outlets. Further, there are factors that are usually not considered by fully automatic algorithms, for example, if planners prefer to emphasize space around manual labor stations to make the workers feel more comfortable. The first of the following approaches (Section 4.2) focuses on a u-shaped layout setup wherein the positions of movable objects should be optimized regarding the paths workers have to travel between the stations they have to work on. The second approach (Section 4.3) assumes a completely modular production line setup and extends the simulation for advanced manufacturing (SAM) by Wörner [202] with an augmented reality component to allow an on-site optimization of existing production lines.

This chapter is partly based on the following publications:

- D. Herr, S. Grund and T. Ertl. “BlueCollar: Optimizing Worker Paths on Factory Shop Floors with Visual Analytics”. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019, pp. 1580–1589. URL: <http://hdl.handle.net/10125/59598> (visited on 2019-10-17) [2]
- D. Herr, J. Reinhardt, R. Krüger, G. Reina and T. Ertl. “Immersive Visual Analytics for Modular Factory Layout Planning”. In: *Proceedings of the Workshop on Immersive Analytics*. IEEE, 2017. URL: [https://groups.inf.ed.ac.uk/vishub/immersiveanalytics/papers/IA\\_2278-paper.pdf](https://groups.inf.ed.ac.uk/vishub/immersiveanalytics/papers/IA_2278-paper.pdf) (visited on 2019-10-17) [7]
- D. Herr, J. Reinhardt, G. Reina, R. Krüger, R. Villanueva Ferrari and T. Ertl. “Immersive Modular Factory Layout Planning using Augmented Reality”. In: *Procedia CIRP* 72 (2018), pp. 1112–1117 [8]

## 4.1 Related Work

This chapter focuses on supporting planning experts for production lines to create or improve the positions of machinery and movable objects to improve the overall production workflow. Therefore, the following section covers related work in the fields of layout planning and visual support for optimization algorithms for layout planning. Further, previous work in the field of augmented and virtual reality in manufacturing is presented.

### 4.1.1 Layout Planning

Planning an optimal and adaptable facility layout is an ongoing challenge that already exists for a long time [20]. The parameters to optimize are manifold and range from value stream analysis over incorporating external suppliers to optimizing worker paths. There are also approaches that take the assembly

process [82] into account and combine it with the layout planning process into a holistic approach [51].

Most visualization and visual analytics approaches that target the manufacturing domain focus on the optimization of simulations and production schedules. Rohrer [160] argues that visualization helps domain experts to get a better understanding of manufacturing simulations, for example, by visually representing the paths operators take between machines. Also, visualization enables an interactive communication of results between a simulation software and its users. For example, Wörner [202] presents a production simulation framework with an integrated visual analytics approach that visualizes bottlenecks in production lines with modular components. This system is introduced in more detail in Section 4.3.1, as this thesis complements the system with an augmented reality approach. Other examples help to find anomalies [130] in production line processes, as well as to gain general process insights [206].

In order to optimize production schedules, Klöpffer et al. [106] presented a system generating a set of possible production schedules that can be iteratively reduced based on aspects experts deem to be most important. *LiveGantt* [102] helps experts to explore Gantt charts of large concurrent schedules. Users interact with the schedule and get visual feedback about the effects of their changes.

Further, there is an increasing number of approaches that make use of genetic algorithms to optimize layouts based on *Key Performance Indicators* (KPIs) resulting from layout simulations [96, 49]. They optimize layouts without human intervention. However, automatic methods neglect that experts may be able to further improve a layout using their experience and that possibly unmodeled (or hard and costly to model) constraints cannot be considered.

Many approaches exist that apply optimization rules and algorithms to find a satisfying layout [55]. Often, layout optimizations and layout simulations are used together to improve the outcome of the layout planning phase [11]. Modern layout planning systems also make use of computer-aided design models combined with visualization to make use of human experts' domain knowledge in the planning process [147] or rely on automatic approaches [122].

#### 4.1.2 Visual Support for Optimization Algorithms for Layout Planning

Recently, deep neural networks become increasingly popular for classification or recommendation systems, although their internal mechanisms are often difficult or impossible to understand [124, 157]. Further, such approaches need a lot of training data (that may not be available) and they are unresponsive during the training, as their intermediate results often cannot be used.

Other optimization approaches, such as simulated annealing [105], use physical models to control the search space used to find a global optimum for a black-box function. They are easy to understand and provide intermediate results.

However, they are heavily reliant on their starting configuration and they are unable to represent multiple local optima with similar function values as the global optimum.

The first approach presented in this chapter uses an estimation of distribution algorithm (EDA) to provide recommendations. Unlike evolutionary algorithms, EDAs create a high-dimensional probability space (dependent on the number of parameters) to pick new population members. This high-dimensional probability space can be used to visualize intermediate results, which makes the picking strategy easier to understand compared to the recombination effects used by evolutionary algorithms. There are various alternatives to visualize such high-dimensional data, for example, scatterplot matrices [85], parallel coordinate plots [98], glyphs [28], or projection techniques [187], which can be presented using scatter plots or heat maps.

### 4.1.3 Virtual & Augmented Reality in Manufacturing

In the past years, several approaches that make use of virtual or augmented reality were proposed. Planning entirely virtual production layouts before setting them up in reality [142, 120] reduces costs and collaborative scenarios as presented by Menck et al. [134] enable the planning of factory layouts with other planners independent of their real-world location.

By incorporating digital data into the real world, augmented reality approaches are able to provide a wide range of information to users. This can be currently available machine data [165] or insights and practicability assessments of assembly processes [145, 144]. Further, it is possible to understand the consequences of changed production layouts or production processes [153, 53]. The improving quality and prices of VR/AR devices make these approaches increasingly relevant.

The idea of using AR headsets to support workers in a factory is not new [44]. Many of the early AR applications like the Studierstube [164] relied on custom and oftentimes inflexible video see-through hardware for augmenting the real world. Today, many ready-to-use devices are available, which prompted Palmarini et al. [148] to present a structured workflow for choosing the right AR technology for industrial maintenance tasks. The upcoming of untethered devices offering a reasonably stable out-of-the-box registration, most prominently the Microsoft HoloLens, led to an increase of AR research in various areas. These include surgery [116, 135], molecular graphics [80], and a variety of industrial applications. For instance, in factory planning scenarios, augmented and virtual reality support can help to deal with the complexity of modern production processes and reduce the planning costs [120, 53, 142, 153]. Training of new workers is also an area in which AR was used [123].

Wang et al. [192] surveyed augmented reality systems for product assemblies. Typically, such systems focus on guiding technicians through predefined steps, for example, a step-by-step explanation of how to service or repair a machine [70, 107, 12, 185]. Likewise, Henderson et al. [94] explore the applicability of augmented reality to repairing military vehicles. Making digital information available in the real world is another application field for AR. A study of Erkoyuncu et al. [63] showed that users of their AR system, which overlays workpieces with adaptive AR content, completed tasks nearly twice as fast as the control group that had to use instructions on paper. However, in a real production process with multiple machines working in parallel, e.g., an assembly line, these approaches address only the lowest-level step of the manufacturing process.

## 4.2 Visual Analysis and Optimization of Worker Paths in U-Shaped Factory Layouts

One important aspect for the efficiency of a factory is the productivity of the workers that operate and maintain machine tools. There are many possibilities for improving their productivity, such as enhancing the ergonomics at a workstation [68] and optimizing the work schedule [102]. The layout of a factory is tightly coupled to the work schedule, as their combination influence the paths that workers have to travel to their next work station. This becomes even more important in setups like u-shaped layouts, as they may require workers to operate multiple machines or to share special equipment with other workers [168].

Often, pathing problems for workers are hard to account for during the initial layout planning phase. Exemplary reasons are changes in a production line over time (e.g., machinery replacement), the addition of new production lines, or changing work schedules. There are numerous approaches to support the pathing of workers during production (e.g., visual cues on the floor or on the nearest machine terminal). It is also possible to optimize the positions of movable parts, such as shared tool caches. However, this optimization is challenging [140]. On the one hand, an automatic optimization is expensive to calculate due to the large number of possible solutions. Further, its results are prone to errors due to unmodeled constraints (e.g., availability of adequate power supply) or constraints only known to experts, such as understanding that workers avoid being close to loud machinery. On the other hand, manual optimization by experts is challenging, as many paths need to be considered during the optimization.

The following visual analysis approach uses an estimation of distribution algorithm (EDA) (see Section 2.1.3) to support layout planners in optimizing the locations of movable machinery and containers, such as tool caches. Planners are first provided with an overview of a layout's performance regarding the

pathing of workers. Afterwards, they are provided with visual cues that indicate, which components have high optimization potential. Upon the selection of a component, a heat map visualization visually highlights where the component could be repositioned to.

To evaluate the approach, a prototypical system called *BlueCollar* was implemented. An application scenario based on an experimental production line layout provided by a production optimization company is presented to demonstrate the applicability of the approach.

### 4.2.1 Approach

Targeting layout planning experts, *BlueCollar* supports planners to improve the efficiency of factory layouts by optimizing the paths workers have to take to complete a work schedule. The following first presents the requirements and the resulting workflow of the approach. Afterwards, the different components used in the approach and the underlying estimation of distribution algorithm (EDA), which provides the optimization recommendations, are detailed. Three system requirements for interactively optimizing factory layouts regarding the pathing of the workers were identified in previous informal expert interviews:

**Requirement 1: Current Performance**  $(R_1)$

Provide information about the current layout's performance regarding the taken paths of workers.

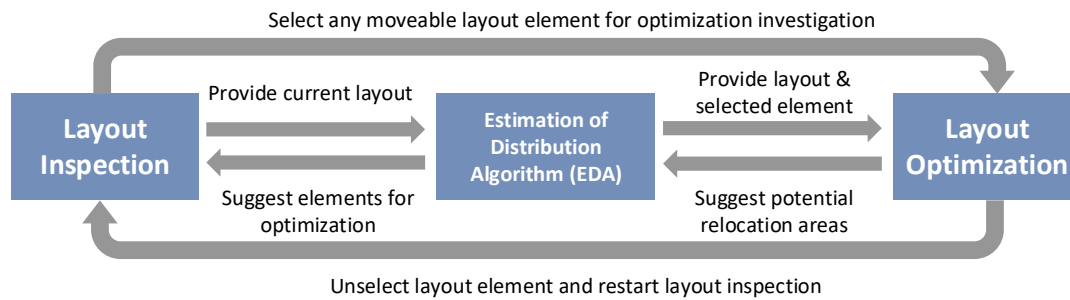
**Requirement 2: Overview Guidance**  $(R_2)$

Present visual feedback regarding the optimization potential of the layout elements.

**Requirement 3: Optimization Guidance**  $(R_3)$

Visualize suitable relocation areas for specific layout elements and provide information about the impact of the relocation on the layout's performance.

Based on these requirements, a visual analytics approach was developed that comprises two stages (see Figure 4.1). Both stages support experts to decide, how to continue the optimization through EDA-based recommendations. Initially, planning experts can inspect the layout and get a first overview of the positions of the layout elements (e.g., machine tools or tool caches). At this point,



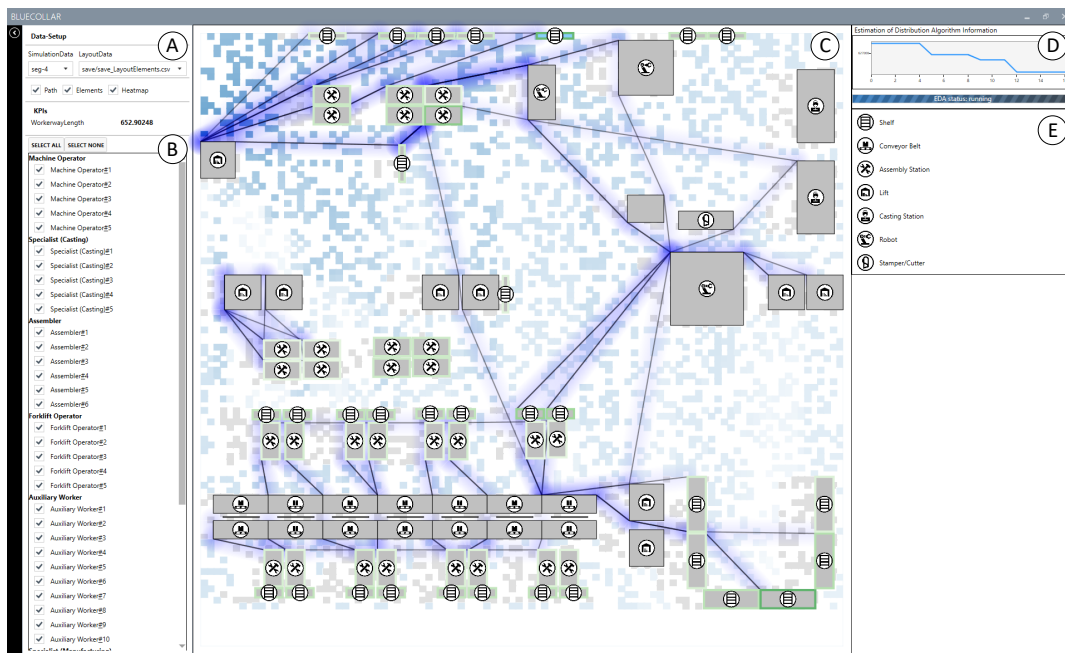
**Figure 4.1:** *BlueCollar* comprises two stages that use an estimation of distribution algorithm (EDA) to provide visual optimization recommendations. The layout inspection provides an overview of the layout and recommends layout elements that have high potential to improve the layout performance. The layout optimization recommends suitable areas to relocate specific elements to. Source: [2] under CC BY-NC-ND 4.0

*BlueCollar* provides information about the most likely taken paths of workers, the performance of the current layout regarding the pathing of the workers, and which elements have the highest optimization potential. Once experts decide, which element's position to optimize, *BlueCollar* visualizes suitable areas for the relocation of the selected layout element through a progressively updating heat map visualization. Based on these recommendations, experts can manually modify the layout and continue the optimization.

### 4.2.2 Layout Inspection

To optimize a layout, experts first need to get an overview of the status quo. *BlueCollar* provides the current layout and enables users to inspect the performance either for the entire workforce or by selecting individuals or groups of workers. The analysis starts with an already existing factory layout and a planned work schedule (see Figure 4.2 (A)). A work schedule describes all of the steps needed to complete the production of certain goods. As the approach targets the optimization of the paths of the workers, the work schedule was restricted to tasks that require workers to walk to other machinery.

**Layout Performance Overview.** To get an overview of the performance of the current layout, *BlueCollar* presents the current layout as a 2D plan view (see Figure 4.2 (C)). Every layout element is represented by its rectangular bounding box and an additional icon to represent its function. A legend provides detailed information about the meanings of each icon (see Figure 4.2 (E)). Further, the most likely taken paths of the workers (based on the shortest walking distance, which are calculated using the A\* pathfinding algorithm [84]) are indicated through semi-transparent polylines. It is possible that some paths are taken multiple times, by either a single worker or multiple workers. Therefore, the line



**Figure 4.2:** *BlueCollar* requires layout planning experts to load a layout and work schedule (A) and choose which workers' schedule data to include in the optimization (B). The layout of the factory (C) provides a first overview of the positions of the components and the taken worker paths. Further, it recommends suitable components for optimization through a color coding of their border's color and presents fitting relocation areas as a heat map. A line chart (D) shows the progress of the optimization algorithm. The elements in the layout view are annotated with icons, which are explained in a legend on the right (E). Source: [2] under CC BY-NC-ND 4.0

segments that were used multiple times are less transparent than segments only used once. This provides layout planners with an overview of the current layout's performance, which meets requirement (R<sub>1</sub>).

**Data Filter.** Layout planners can also view the performance of individual workers or specific groups of workers, e.g., the group of assemblers or machine operators. The ideally taken paths are shown as lines in the layout. They can be filtered by selecting them individually or based on their task in the side panel (see Figure 4.2 (B)).

To get a better overview of high traffic zones, *BlueCollar* emphasizes the paths with a semi-transparent blue color glow to encode the worker density in the layout view. This information can be used to get first insights about possible current or future bottlenecks, where workers may collide or have to take detours. In contrast to the path lines, which emphasize the taken paths, the heat map emphasizes areas with highly frequented crossings.



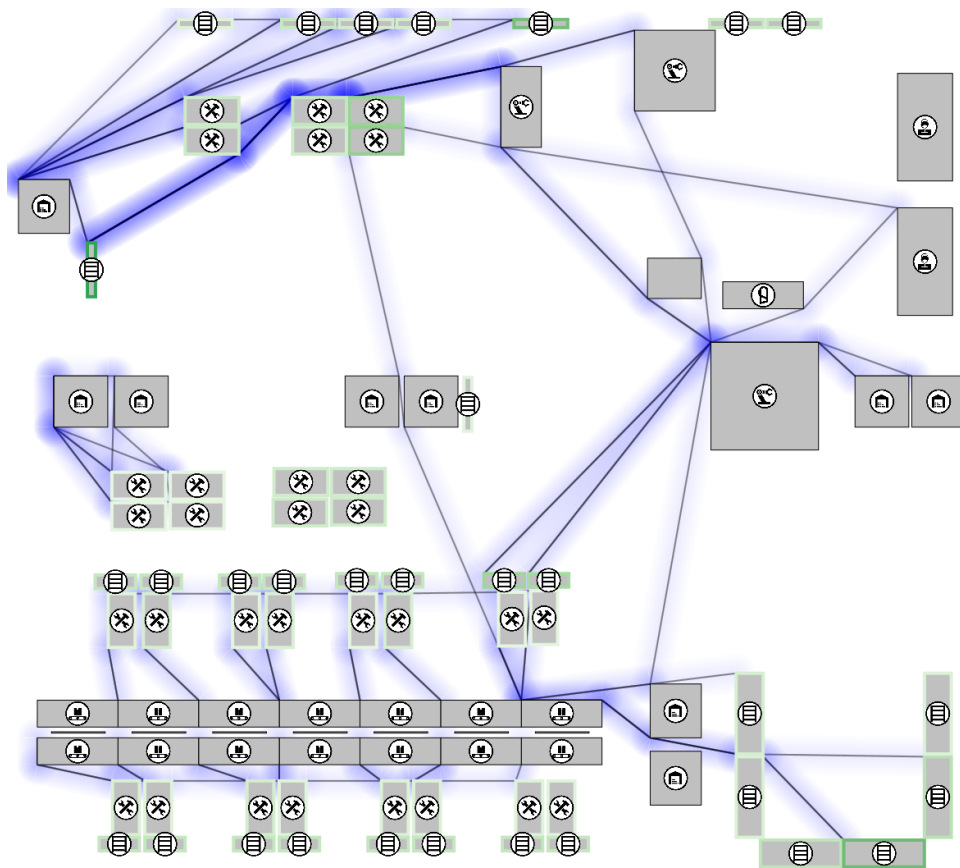
### 4.2.3 Visual Optimization Guidance

*BlueCollar* guides experts during the optimization process in two stages: it recommends elements that have high optimization potential and it suggests better positions for these elements. In the following, the visual guidance is presented in more detail.

**Layout element recommendation.** *BlueCollar* indicates the optimization potential of movable layout elements, such as machines or tool caches by color coding their borders. A light green border indicates very small changes and the more saturated the border becomes, the higher the layout element relocation potential is. The possibility of a worse performance can be discarded, as leaving the layout element at its current position is always a possibility and therefore marks the worst case for the optimization. To compensate for the non-uniform distribution of the improvement potential values of the individual component, the non-linear color saturation mapping introduced by Liu et al. [126] was used. The optimization potential is iteratively refined in the background. Figure 4.3 shows an example, where *BlueCollar* provides information of the optimization potential of the layout elements. Section 4.2.4 details how the optimization algorithm handles the evaluation of multiple layout elements. The layout element recommendation meets requirement  $(R_2)$ .

**Relocation recommendation.** Once planners select a specific layout element for optimization, the element is highlighted with a light blue background and the EDA will only optimize the position of the selected element. To give the experts an overview of the optimization progress, a line chart (see Figure 4.2 (D)) shows the best layout scores after each iteration. The x-axis shows the iteration and the y-axis the cost. A decreasing value indicates that a better relocation position was found.

Additionally, *BlueCollar* visualizes the optimization progress by mapping already available results onto a heat map visualization overlay on top of the currently viewed layout. This results in a progressively updating heat map, which gives planners an early impression of possible relocation areas. The color scheme of the heat map ranges from white (worst performance) to dark blue (best performance) and uses a min-max normalization of the available data. Due to the way EDA works, the surroundings of well-performing relocation areas are more likely to be sampled, which additionally emphasizes these regions. Figure 4.4 shows the suggested relocation areas for the selected casting station (D) based on a simplified work plan. The legend on the bottom right side of Figure 4.4 shows the heat map's color coding. The results show that the station should be relocated between the conveyor belt (A) and the robot station (C). As the heat map includes information about the relocation areas and their optimization potential, they meet analysis requirement  $(R_3)$ .



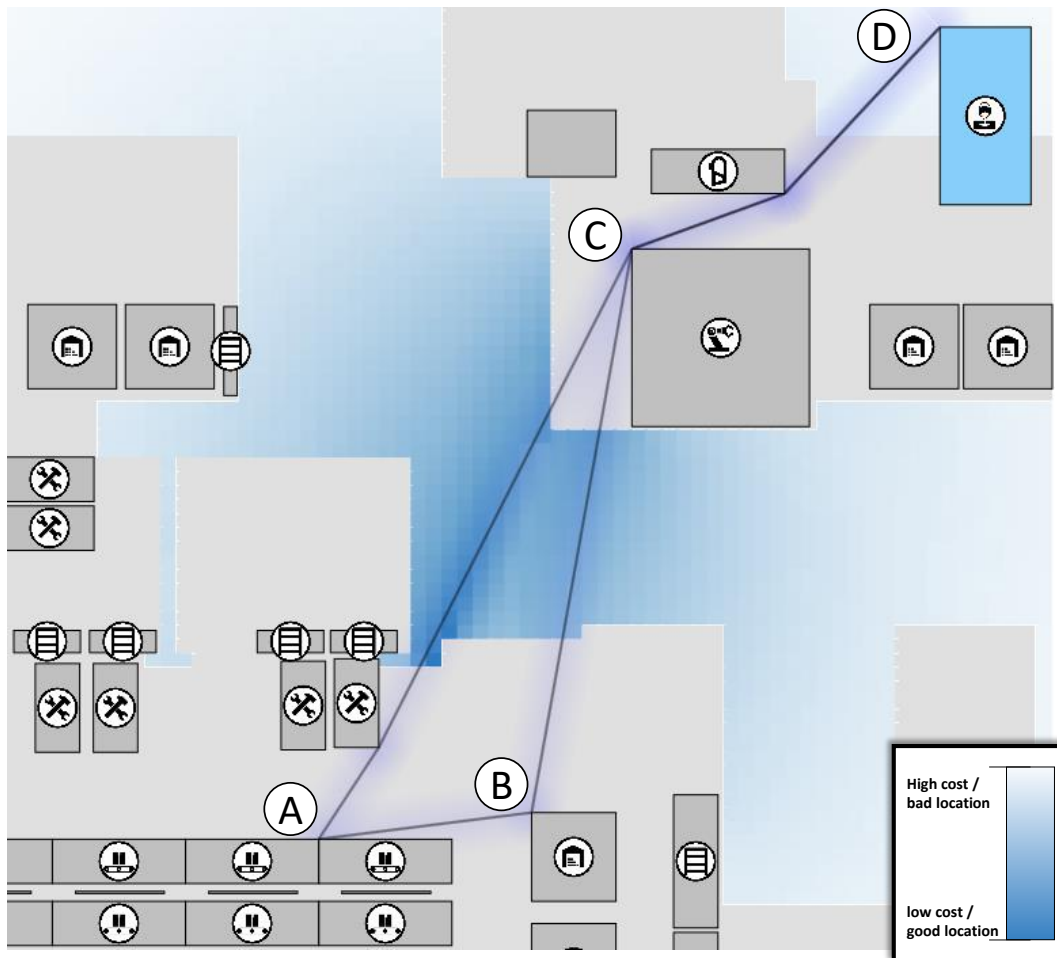
**Figure 4.3:** *BlueCollar* provides a visual indication of the optimization potential for movable components through the border of the elements. The color ranges from light green (low potential) to dark green (high potential). Source: [2] under CC BY-NC-ND 4.0

Planners can also relocate layout elements manually by dragging them within the layout view. If the dragged element is currently selected, the heat map keeps updating. Otherwise, the selection is switched to the dragged element, the heat map is cleared, and the EDA is restarted.

#### 4.2.4 Optimization with an Estimation of Distribution Algorithm

The suggestions for suitable layout elements for relocation and the recommendations for suitable relocation areas (see Section 4.2.3) are based on an estimation of distribution algorithm. EDAs are a class of optimization algorithms that aim to minimize the output of a given black-box function (see Section 2.1.3).

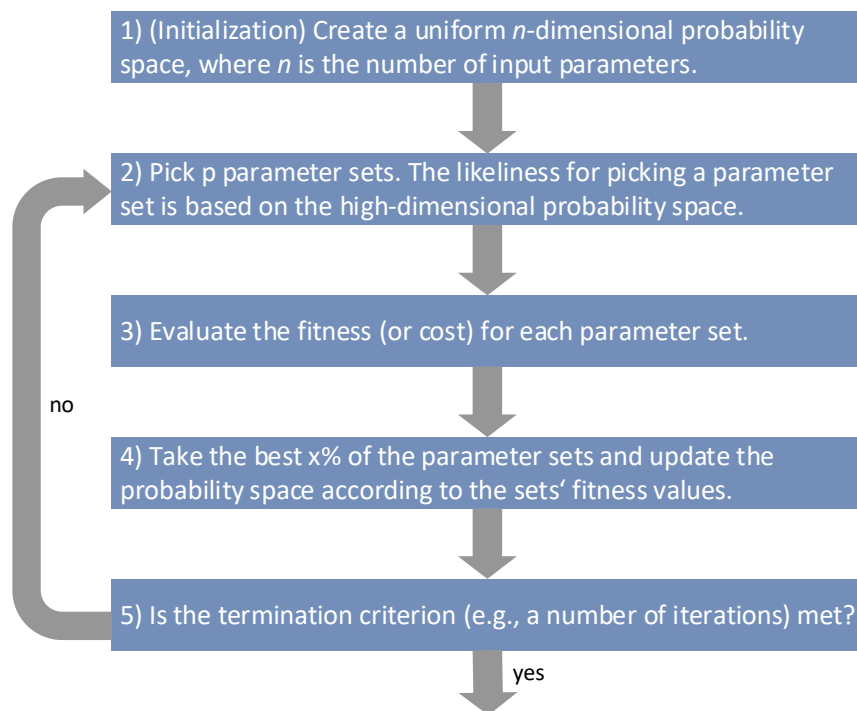
Figure 4.5 shows the general steps of an EDA. It initializes with a uniform probability distribution and then iteratively picks parameter sets for the black-



**Figure 4.4:** The worker has to walk in a circle ((A)→(B)→(C)→(D)→(A)). The legend on the right shows the heat map’s color coding. The heat map indicates that the selected casting station (D) should be relocated between the conveyor belt (A) and the robot station (C). Source: [2] under CC BY-NC-ND 4.0

box function based on the probability space. The picked sets are evaluated and the probability space updates depending on the evaluation results.

Generally, there are two options for the probability space update: i) rebuild the probability space depending on all previously evaluated input parameter sets (requires more memory); ii) use only the evaluated input parameter sets of the last iteration (requires more computing power, as more parameter sets need to be picked). Although the probability space may converge faster with the latter option, the probability space may change considerably in each iteration step, resulting in quickly changing heat maps, which increases the cognitive load on the users. Additionally, the evaluated input parameter sets can be used for the heat map visualization. Therefore, *BlueCollar* uses the first option.



**Figure 4.5:** Generally, an EDA comprises five steps [89]. After initializing its probability space with a uniform picking chance for all configurations, it iteratively picks parameter sets based on the probability space. Then, the probability space is iteratively updated by evaluating the picked parameter sets. This process repeats until a predefined termination criterion is met. Source: [2] under CC BY-NC-ND 4.0

*BlueCollar* uses the combined length of all paths that the workers have to take to complete a work plan as the cost function to measure a layout's performance. To efficiently calculate the workers' paths, it first maps the layout to a graph representation. The approach assumes that each layout element has a rectangular bounding box, each comprising four corner vertices. In an initialization step, all vertices are connected to each other (which results in a complete graph). After that, edges that intersect any layout element are removed. To calculate a worker's path, the corresponding work plan is split into individual tasks. Then, the optimal path is calculated to solve each task that requires the worker to change the location using the A\* pathfinding algorithm [84]. Afterward, *BlueCollar* reconstructs the total path of the worker's total path by concatenating all task-based paths. At last, the path lengths of each worker are summed up.

As generating the layout graph is the most expensive operation, the relocation operation was optimized to reuse the original graph and perform as few graph changes as possible. This is possible, as the EDA implementation of *BlueCollar* only optimizes the position of one layout element at a time. It is assumed that



**Figure 4.6:** Exemplary one-dimensional weighting stamp that influences all neighbors with distance  $\leq 2$ . Source: [2] under CC BY-NC-ND 4.0

this layout element is known at the start of the optimization and the complete graph of all connections between the elements is built without this element. Then, the element can be added at an arbitrary position. A quadtree is used to check if the newly added layout element collides with other layout elements or edges. In case the layout element collides with another element, the layout is assumed to be invalid and the layout’s cost value is increased accordingly. Setting the cost close to infinity would disproportionately decrease the picking probability of surrounding configurations (see *updating the probability model* below). Therefore, the cost value of an invalid layout position is set to a plausible value that would indicate a very inefficient layout:  $cost_{max} = s \cdot l_{layout}$ , where  $s$  is the total number of path segments and  $l_{layout}$  is the width of the layout.

If the layout element collides with connections, these are temporarily removed from the graph. Every affected connection is implicitly replaced by two edges: one from the starting vertex to the inserted element and one from the inserted element to the ending vertex. For further relocations of the element, the only needed adaptation is to remove the previously inserted layout element and its edges and inserting the deleted edges again before adding the layout element elsewhere. This skips the initialization step, which is the most expensive part of the procedure.

To improve the pathfinding’s scalability regarding the number of workers, *BlueCollar* uses a lookup table that stores all previously calculated movement tasks. This reduces the calculation time, as each path needs to be calculated only once.

After all parameter sets are evaluated, the probability model that decides, which parameter sets are likely to be picked in the next generation, needs to be updated. In addition to the assumption that the two position parameters have a dependency, it is assumed that the neighboring cells of the evaluated parameter sets have similar cost values. Consequentially, the cost of the neighboring cells is approximated based on the already evaluated parameter configurations. The influence of neighboring cells is modeled as a two-dimensional triangular function, where the center has the highest influence and the values are halved for every step towards the outer corner of the neighborhood (see Figure 4.6 for a 1D example). The size of the stamp depends on the dataset. In the dataset used in Section 4.2.5, a 100x100 cell grid is used for the heat map and the stamp size covers five cells in every direction.

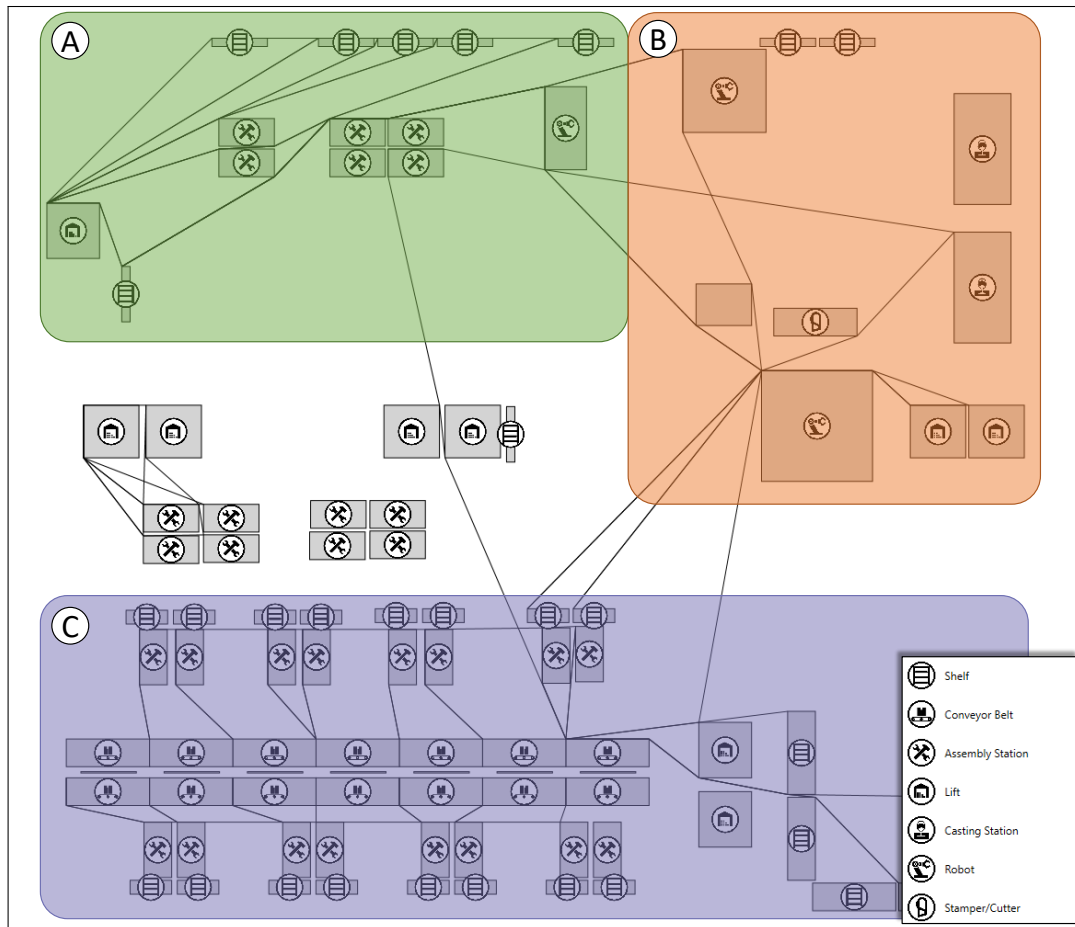
Without further modifications, the algorithm is likely to get stuck in local minima, as the initially picked parameter sets strongly influence the development of the probability space. To prevent this, a base probability of 10% for all elements to be picked is reserved (e.g., if there are four elements, each of the elements has a base probability of  $\frac{10\%}{4} = 2.5\%$ ).

The runtime of each iteration depends on the cost to pick and evaluate the population members and the cost to update the probability space. In the EDA implementation of *BlueCollar*, the runtime cost is

$$\begin{aligned} \text{runtime cost} &= p \cdot \overbrace{(\mathcal{O}(2 \cdot d + \text{runtime}_{eval}))}^{\text{pick \& evaluate}} + \overbrace{(\mathcal{O}(k^2))}^{\text{prob. space update}}, \\ &= p \cdot (\mathcal{O}(d + \text{runtime}_{eval}) + \mathcal{O}(k^2)), \end{aligned}$$

where  $p$  is the number of picked parameter sets,  $d$  is the size of the grid's dimensions and  $k$  is the size of the kernel. The cost of picking new elements mainly depends on the runtime complexity of the evaluation function, which depends on the number of necessary node expansions in  $A^*$ . In the implementation of *BlueCollar*, the cost of updating the probability space depends on the number of nodes that need to be updated. In the worst case, none of the picked population members were evaluated before and therefore, every element's neighborhood (which is  $k^2$ ) must be updated, leading to the runtime complexity given above.

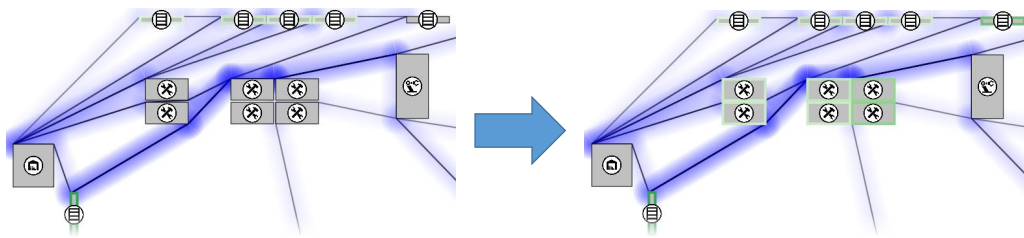
Until this point, only the position optimization of a previously known layout element was considered. To assist users in deciding, which element to optimize, *BlueCollar* initially provides the optimization potential for various layout elements at the same time. To provide this information, all movable layout elements are added to a queue. The element at the front of the queue is evaluated with one pass of the estimation of distribution algorithm. The best layout result is used as the respective optimization potential value. Afterwards, the calculated layout and the corresponding performance value is stored for that layout element, which is then added at the end of the queue again. If the layout element that is taken from the queue already contains calculated positions, these are used to initialize the probability space. The main challenge in this step is to balance the calculation time assigned to each layout element to calculate its optimization potential against the evaluation of as many layout elements in as little time as possible. This means that more time per layout element results in a better indication quality of individual layout elements, while less time means that more layout elements can provide an indication, but with a lower indication quality. Based on observations made while testing the approach, running one EDA iteration that generates  $\sqrt{n}$  parameter configurations, with  $n$  being the number of grid cells, seems to be a good trade-off between the reliability of the potential and its calculation speed.



**Figure 4.7:** The layout can be subdivided into three areas: The green area (A) comprises assembly stations and multiple shelves that are connected to a lift. The orange area (B) contains mainly robot stations. The purple area (C) contains a conveyor belt, which is connected to multiple assembly stations and a high rack storage. Source: [2] under CC BY-NC-ND 4.0

#### 4.2.5 Application Scenario

A software engineering company that offers planning and production simulation software provided the layout used in the following application scenario. The production layout was planned for a prototypical production line and focuses on manually operated assembly stations between which the workers have to move. The layout is 35.23 meters wide and 31.5 meters long. It can be subdivided into three areas (see Figure 4.7). The green area (A) comprises several shelves that are refilled and assembly stations that produce parts for a robot station. Further, one station gets supplies from a separate shelf. Area (B), highlighted in orange, contains three robot stations. The bottom-most station's goods are continuously delivered to two lifts. The purple area (C) is composed of two conveyor belts that transport goods from an (unmodeled) external supplier. The workers at



**Figure 4.8:** At the beginning, only optimization potentials for the shelves are available. Soon after, the potentials of the assembly stations become available, but the shelf near the lift still remains the station with the highest optimization potential. Source: [2] under CC BY-NC-ND 4.0

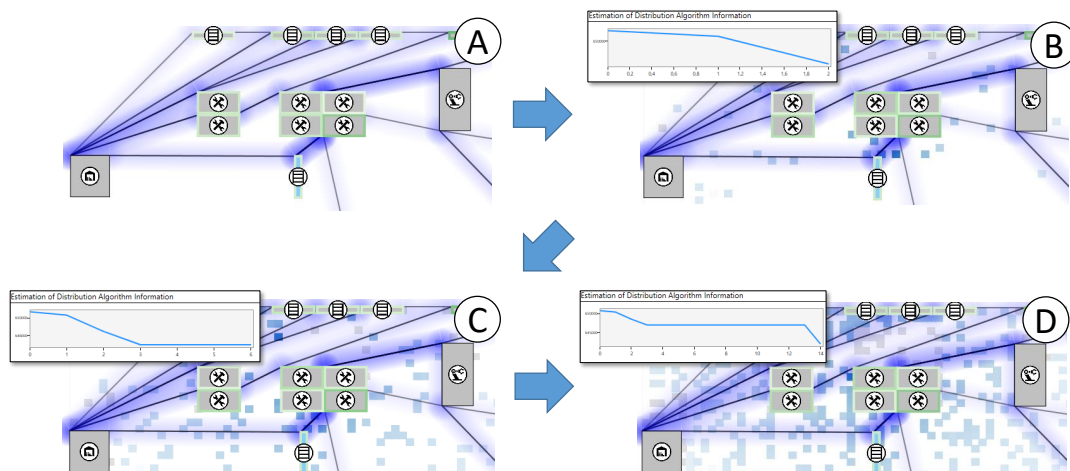
the assembly stations along the conveyor belts take the goods, process them, and put them back on the belt. At last, the parts are taken off the belt and get stored in a high rack storage area. Plausible work schedules that incorporate different tasks in several areas of the factory were added beforehand.

**Exploration of the optimization potential.** We assume the role of a layout planner and start the analysis with the layout view as presented in Figure 4.7. It seems obvious that the planned layout can be optimized, as there are many long paths to a limited number of stations. One of the areas that seems viable for optimization is area (A). The positions of the lift and the robot are fixed, but the shelves at the top, the assembly stations, and the shelf below the lift can be relocated. However, it is unclear, which layout element has the highest optimization potential. Therefore, we use the optimization suggestion that *BlueCollar* continuously extends and improves (see Figure 4.8). The system recommends optimizing the shelf near the lift. With this insight, we manually relocate the shelf towards the block of assembly stations (see Figure 4.9 (A)).

**Optimization of the layout.** To further optimize the shelf’s position, we decide to select it and optimize its position with the support of *BlueCollar*’s relocation heat map. As the analysis progresses, it becomes apparent that the most suitable position for relocation is in the gap between the two groups of assembly stations (see Figure 4.9 (B)–(D)). Although the best position is plausible, we decide to leave it below the assembly station. Based on total walking distance, this position may be worse (650m vs. 640m), but this placement prevents crowding of workers.

After the shelf’s relocation, we return to the element recommendation mode to validate the optimization potential of the shelf at the top right. Then, we select the shelf to get detailed relocation recommendations (see Figure 4.2). *BlueCollar* recommends to place the shelf directly above the lift (reducing the distance to 623m), but we opt to place it to the left of the top shelves to keep obstructions at a minimum, still reducing the path length to 632m.



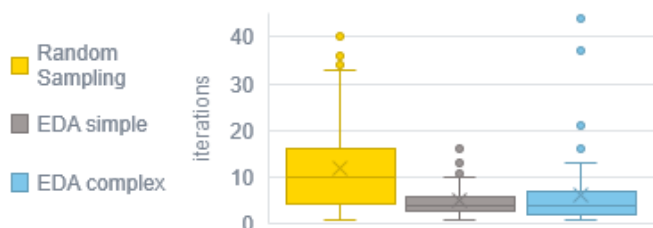


**Figure 4.9:** As the optimization continues, it becomes apparent that the most ideal relocation area for the shelf is between the two groups of assembly stations. Source: [2] under CC BY-NC-ND 4.0

#### 4.2.6 Further Results

BlueCollar’s EDA implementation converges quickly towards good results but has difficulties to find the best result (due to its probabilistic nature). To support this hypothesis, it was measured, how many iterations *BlueCollar* needs to find one of the ten best results for the relocation of a specific layout element. The ground truth for the grid cells was built in a separate pass. For this evaluation, the simple layout (see Section 4.2.3) and the complete plan (see Section 4.2.5) were used. The used population size per iteration was 100 and the measurements were repeated 100 times per layout. The results are shown in Figure 4.10 and will be annotated as  $\varnothing(\text{average}) \pm \sigma(\text{standard deviation})$  in the following.

For the simple plan, the EDA needed  $5.18 \pm 3.04$  iterations to find a top 10 value. For the complex plan, it needed  $6.12 \pm 6.48$  iterations for a top 10 value. The values were benchmarked against a random sampling, which resulted in  $12.07 \pm 10.40$  iterations for a top 10 pick. Overall, *BlueCollar* does not only find suitable values in fewer iterations than random sampling, but it also guides planners quicker towards areas of interest, as areas that perform better have a higher chance of being evaluated and therefore being highlighted.



**Figure 4.10:** EDA finds the best results in fewer iterations and with less variance than random sampling. An outlier of 61 iterations in during random sampling is not shown.

## 4.3 Immersive Analysis of Production Line Simulations

Nowadays, factories have to be flexible and adaptable to address the consumers' rapidly changing demands of highly customizable products. One way to achieve this is to deploy quickly re-arrangeable production line components.

Production line layout planning tools are usually designed as desktop applications. They require high cognitive effort to reconcile the final physical setup with a complex real-world environment starting from an abstract model (see Fig. 4.11). It is challenging to imagine how efficiently a production line layout can be used by workers after being built. Augmented Reality (AR) [30] can help to reduce the cognitive gap between a virtual scene and its mapping to the physical environment. Corresponding hardware, e. g., head-mounted displays, such as Microsoft HoloLens and Sony SmartEyeglass, superimpose a real-world scene with virtual items. The provided immersion, stereoscopy, and direct interaction enable planners to gain information directly at the location where it is needed.

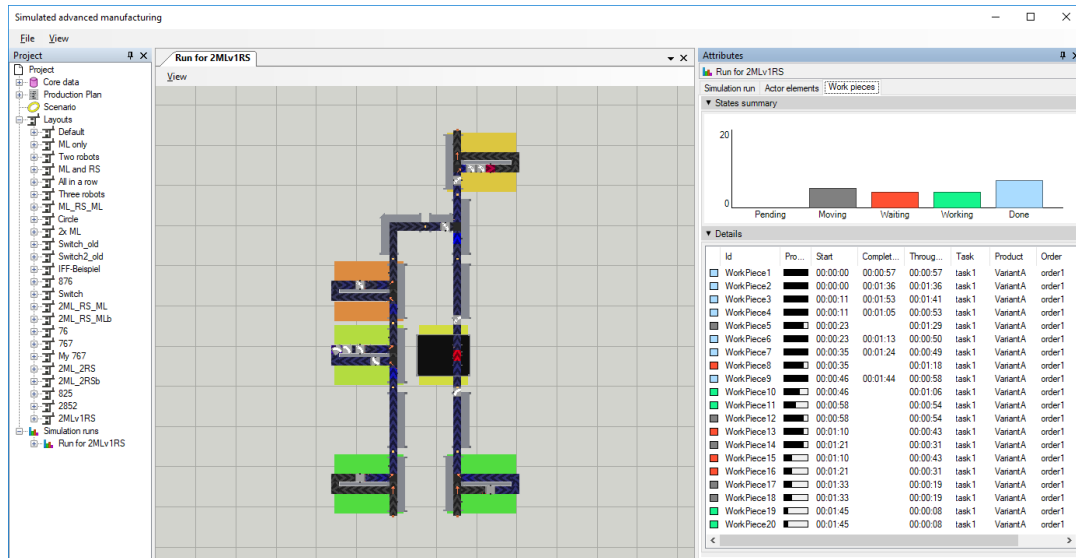
To address the gap between planning and implementation, the following contributes a methodical approach to extend the currently desktop-based layout planning process by an augmented reality component. The following approach is based on the work by Wörner [202, p. 27ff.] and employs the same application example. After its introduction, an approach to extend such desktop-based simulators through augmented reality is presented (Section 4.3.1). Afterwards, the approach is evaluated through an application example and a think aloud expert interview (Section 4.3.2).

### 4.3.1 Approach

This section first introduces the *Simulator for Advanced Manufacturing* (SAM) presented by Wörner [202, p. 34 ff.]. Then, the system is extended by an approach that allows to provide intermediate and final simulation data in an augmented reality environment. The (running) simulation is further enriched with additional information and highlights the differences between a loaded and automatically suggested layouts.

#### **SAM – Simulator for Advanced Manufacturing**

SAM runs on desktop computers and uses the iTRAME system [158], which demonstrates the applicability of modular production lines. iTRAME uses standardized connection modules such that its components can easily be rearranged in an arbitrary order to produce different products more easily. The simulator is able to simulate iTRAME production line layouts, which are composed of linear and corner conveyor belts, as well as lifts, robot stations, manual labor stations, automatic storage, vision stations, and switches. As it focuses on providing



**Figure 4.11:** A production line simulation run with SAM. It shows the layout's components, the workpieces, the actor load (background color), and the workpiece density (conveyor belt color). The right side shows the current status of workpieces and their processing progress. ©2018 Elsevier

information about the overall production line performance, it does not simulate the individual processes within the stations, e.g., the exact process of assembling two components. In SAM, users can manually design new and manipulate existing production line layouts. Further, it can automatically generate and propose new layouts using an evolutionary algorithm. The evolutionary algorithm (see Section 2.1.3) iteratively creates new layouts based on already stored or previously generated layouts by splitting the layouts at an arbitrary position and recombining those layout fragments. Then, the members are assessed based on their validity and Key Performance Indicators (KPIs), such as the number of used components, the required area of the layout, the running costs of the machines, the completion time of the current order, and the average load of the components. Based on their performance, the worst-performing members are discarded and some layouts are arbitrarily changed.

Experts can assess the performance of any layout by running a real-time simulation that provides information about the load of each station, the average workpiece density on the conveyor belt, and the status history of the workpieces. Figure 4.11 demonstrates a simulation run with SAM in which the left side creates the product with manual labor stations, whereas the right side uses a robot station to perform this task. The color coding of the stations backlog indicates that the right line has a better average load and lower workpiece backlog.

The process of the automatic layout generation is presented in a separate view. It shows the current simulation progress for the generated layouts and their

KPIs. It also shows the layouts' overall score, which is a weighted sum of KPIs that can be adjusted to the user's goals to simplify comparison. Furthermore, regardless of its overall score, information about the best-suited layout with respect to the KPIs are given. SAM visually indicates whenever it finds a better layout, which the users can then inspect in detail and edit manually, for example, to change the layout footprint. Each of the layouts is shown in a separate tab, so it is possible to open and quickly compare multiple layouts by hand. The manipulated results are also used by the evolutionary algorithm to find additional layouts with improved KPI.

Overall, SAM enables users to plan, simulate, and assess manually or automatically processed layouts. Most of the process works reasonably well as desktop input mechanisms, such as mouse and keyboard, are well-suited for the tasks and efficient. However, it is difficult to assess aspects related to real-world distances, spatial arrangement, paths that can be walked through, or work safety aspects based on a result presented in a 2D or 3D scene on a desktop client. Often, this issue is emphasized by out-of-date plans, such as for piping or electrics, or the general lack of accurate digital representation of the factory's status quo in brownfield scenarios. Hence, the following approach uses *augmented reality for simulated advanced manufacturing (ARSAM)* and was designed to complement the desktop-based layout planning and simulation approach, not to replace them.

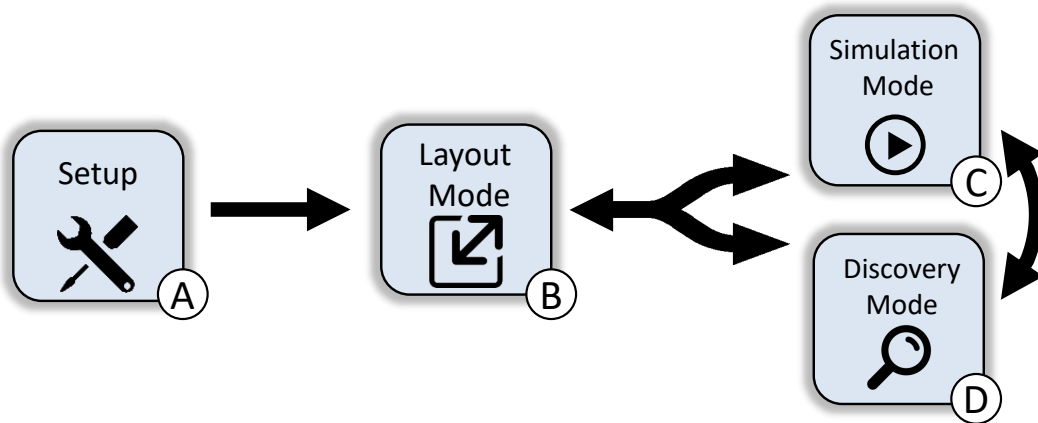
### Augmented Reality Extension

ARSAM's goal is to make use of augmented reality technology, allowing for an immersive layout planning and analysis process. It assumes that one or more layouts were already created on a desktop client that can be transferred to an augmented reality application. After creating the layouts, the users start with the initial setup phase in which they define the global coordinate system and load an initial layout (Figure 4.12 (A)). Subsequently, they choose whether to manipulate the loaded layout (Figure 4.12 (B)), simulate the current layout (Figure 4.12 (C)), or compare the layout to layout variants that are automatically created (Figure 4.12 (D)).

In the following, the individual workflow steps are detailed and the implementation is showcased. ARSAM's prototype is implemented as a HoloLens application. As recommended by Microsoft, the prototype was implemented using Microsoft's Mixed Reality Toolkit framework and Unity<sup>1</sup> as a development environment. HoloLens automatically handles gesture recognition, as well as the automatic registration of the real-world environment with the virtual scene. If necessary, users can manually adjust the alignment of the virtual layout with the real-world environment.

---

<sup>1</sup> <https://unity3d.com/partners/microsoft/mixed-reality>

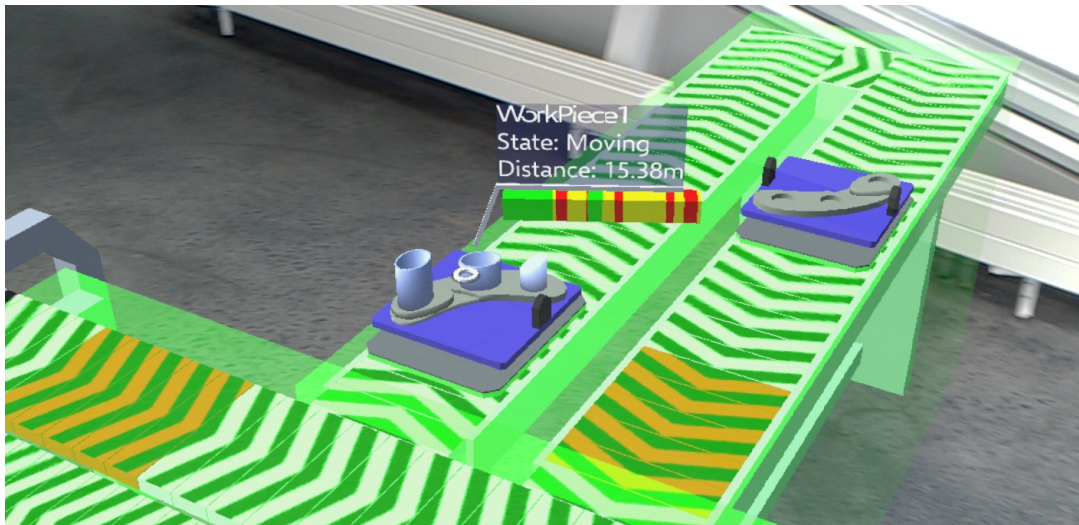


**Figure 4.12:** The layout planning workflow starts by setting the plane for the global coordinate system (A). Thereafter, the planners load and may alter a layout (B). The layout can then be used as a basis for simulations (simulation mode) (C). To explore other layout suggestions and optimizations, the planners can switch to discovery mode (D). They can switch arbitrarily between modes after setting the coordinate system. © 2018 Elsevier

**Initial Setup** Before the users can start the planning process, they need to set the plane for the global coordinate system. In ARSAM, this can be done semi-automatically, where the plane is aligned to the detected floor. Alternatively, the plane can be set entirely manually. If a plane was set before, its position can be reused. The position of the plane affects the positioning of the layouts when they are loaded. Then, the users can load the layouts that were either created using SAM or saved in previous sessions.

**Layout Mode** Users can view the current layout either in an adjustable model size that can fit on a desk and provides a good overview or in its real-world size that shows the virtual layout in a real-world context. The *layout mode* enables users to rearrange parts if they see potential to improve overall performance or satisfy yet-unmodeled constraints in the context of a real environment. Especially in the second mode, experts are able to perceive paths that are too narrow to walk through, the obstruction of safety-relevant inventory, or unmodeled workshop objects, such as supports and piping. However, the AR interaction is still more cumbersome than the desktop user interface when modeling complex layouts from scratch. Hence, if users want to compare an already deployed physical layout with alternatives, the physical layout is best modeled with the desktop application first and then transferred to its AR extension.

In ARSAM's prototype, users need to align the loaded layout manually to the physical counterpart, either by selecting and moving/rotating layout elements or by manipulating the global floor plane. Parts can be selected either individually or

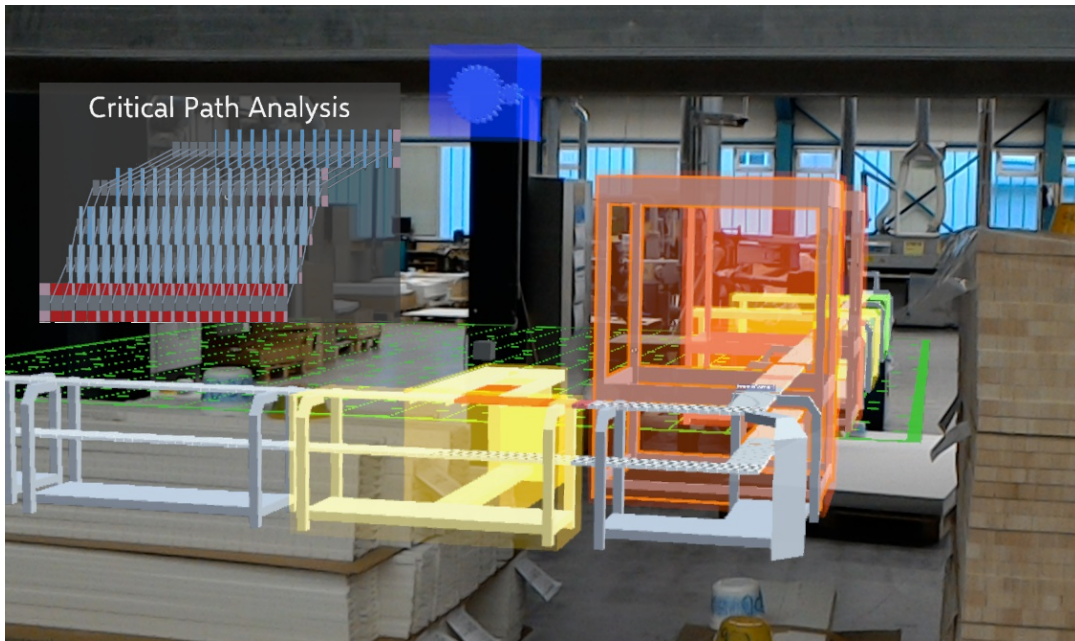


**Figure 4.13:** Screenshot of the detailed information of a specific workpiece. It contains the workpiece’s ID, its current state, its total moved distance from the start of the production line, and its status history. ©2018 Elsevier

as an entire group of connected components. During the manipulation, ARSAM presents a wireframe “ghost” model as a preview of the result alongside the original positioning. Users can thus perform minor layout changes in ARSAM to get a direct impression of their effects in a real environment.

**Simulation Mode** To analyze the current layout’s performance, users can switch to the *simulation mode* to run a simulation that shows how the layout performs during a production run. The provided information depends on the capabilities of the simulator. Analogous to SAM, ARSAM provides real-time information about the status of the work stations (working/idle), the location of the workpieces, and state of the workpieces (e.g., transport between stations, being processed, finished, see Figure 4.13). It also provides information about the load of individual work stations by color coding their bounding volume between red (for no load) to green (used permanently, see Figure 4.14). Similarly, the workpiece density of the conveyor belt segments is encoded in the segments’ color hue. ARSAM uses the same semantics as with the stations’ workload color coding, where red represents an undesired effect. The longer the backlog at a station, the more segments are colored red. Users can inspect any station’s status history for the simulation run through a tooltip, which is shown when the users directly look at them. It provides the current state, average load, and a continuously updating status bar that concisely indicates the load distribution of the station over time, analogous to the workpiece tooltip.

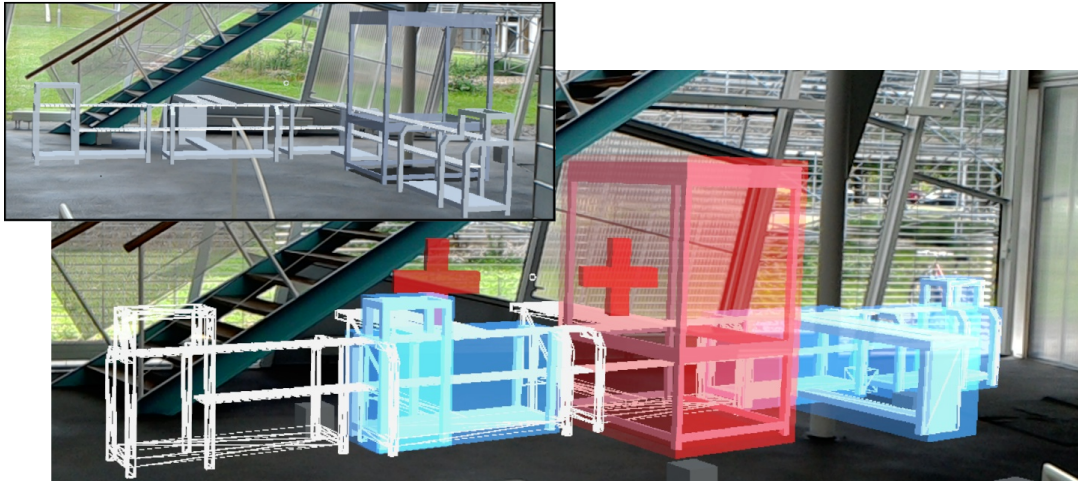
In addition to the visual encoding of the layout, the planning experts can optionally inspect bottlenecks that are calculated using a modified critical path



**Figure 4.14:** Screenshot of a simulation run taken from the users' view. It shows the layout components, their load (bounding volume color), the workpiece density (e.g., on the leftmost manual labor station), and the critical path view to find bottlenecks, where red denotes the critical station which should be optimized first. Image used with permission by Gaugler & Lutz oHG. ©2018 Elsevier

method (see Section 2.2.4). Its results, e.g., the critical path and the drag values, are visualized in the critical path view (see Figure 4.14). Therein, every row represents one station and the individual blocks represent the workpieces during the simulation. A red background marks the station on the critical path with the largest process time and a blue background represents the buffer (i. e., downtime) that a station has before it becomes critical itself. At this point, users are able to get an overview and detailed information about a specific layout's performance, find possible performance bottlenecks and use their expertise to assess possibly unmodeled layout issues, such as the spacing between the work stations.

**Discovery Mode** Analogous to *SAM*, the *discovery mode* uses an evolutionary algorithm, which automatically searches for new layouts that are better than the currently viewed layout regarding the KPIs explained in Section 4.3.1. The users are first presented a view that contains the currently discovered layouts sorted by their overall score. Any of the layouts can be selected for further analysis and comparison with the currently loaded layout. In contrast to *SAM*, where the layouts were inspected in separate tabs and compared in a summary view, *ARSAM* makes use of the augmented space to allow users to inspect both layouts at the same time for a situated analysis in relation to a reference layout. This



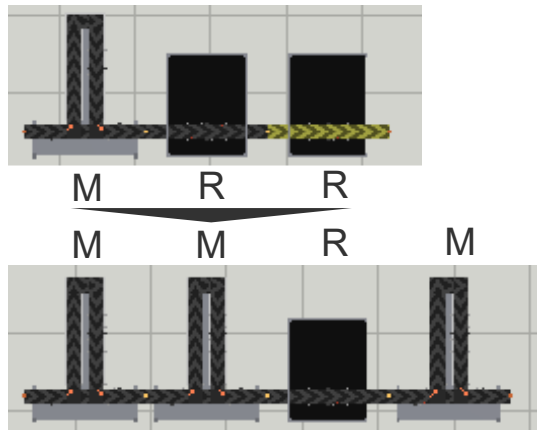
**Figure 4.15:** ARSAM enables users to inspect and edit modular factory layouts (top left). In addition, an evolutionary algorithm proposes new layouts and provides information about the needed changes to transform a given layout (solid layout) to the proposed layout (as wireframe preview). The door described in the scenario is just to the right of the visible field of view. ©2018 Elsevier

enables users to directly compare the differences of the layouts, see the needed changes to transform the loaded into the proposed layout, and inspect and edit the proposed layout in the *layout mode*. The comparison is especially useful if the originally loaded layout is also physically available, but it can also be used to compare two potential layout solutions in the real-world context.

ARSAM highlights the differences between the original and the proposed layout to further assist the users in comprehending the necessary effort (and incurred cost) to transform the original into the new layout. It encodes the needed changes visually into the components of the original and new layout's bounding volume. In case elements remain at their current position or if they need to be moved, their bounding volume is filled with a light blue. Components that are not used anymore are filled with red and components that need to be bought have a red '+' on top of their geometry (see Figure 4.15). All colors are semi-transparent to make sure that the users are still able to perceive the underlying components, regardless of whether they are physically present or virtually added. This additional visualization enables the users to assess if the possible performance increase outweighs the costs of buying new layout components or taking existing components offline.

To provide this information, ARSAM converts both layouts into a string representation where each character represents one component. It then compares the strings using a modified Levenshtein distance. Originally, the Levenshtein distance [121] transforms a string into another using three operations: *insert* or *delete* a character, and *replace* a character with another. However, in this context,





**Figure 4.16:** Depiction of the string encoding of two production lines.

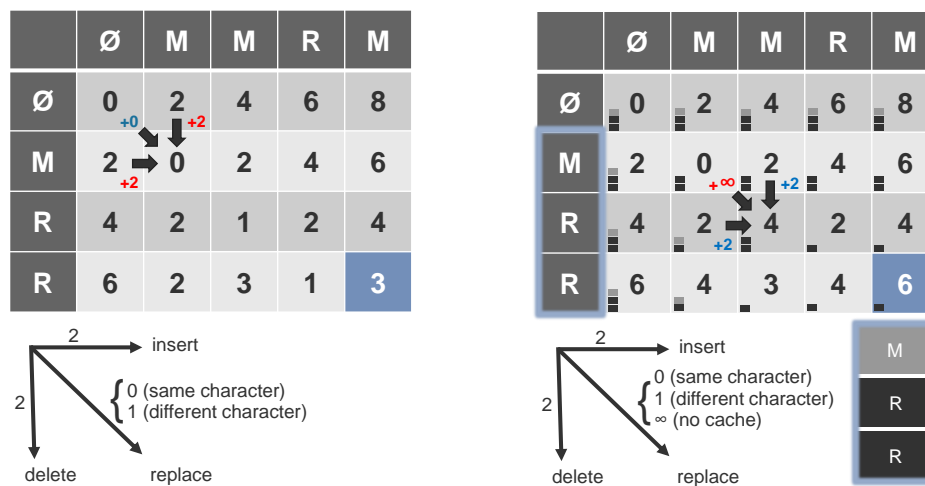
*replacing* a component is only reasonable if the original layout already contained the component, resulting in the *move* operation. To ensure this, a component cache is first built based on the original string. A *move* is only possible if the needed component is still available in the cache (from which it is removed during the process). After calculating the distance, the used operations are reconstructed based on the table built by the algorithm. The modified Levenshtein distance can be used as a KPI in the evolutionary algorithm, as well as providing a reconstruction of the performed transformation operations, which is used for the visual encoding described above. Figure 4.16 shows the encoding of a line that is transformed from one manual labor station and two robot stations to a line with two manual labor stations, followed by a robot station and another manual labor station. Figure 4.17 (A) shows the cost calculation using the traditional Levenshtein distance, Figure 4.17 (B) shows the modified Levenshtein distance.

### 4.3.2 Evaluation

The following evaluation was conducted with the ARSAM's prototypical implementation. First, a synthetic application scenario is showcased to demonstrate how ARSAM can help assess whether a simple production line can be deployed in a room with spatial constraints. Then, feedback about the applicability of the approach collected during an expert interview is presented.

#### Synthetic Application Scenario

In a synthetic application scenario, we assume the role of a layout planning expert that got the task to place a production line in a new hall. The new facility has some spatial restrictions, as the production line should be placed nearby the stairway, while it must not obstruct the door on the right wall. After setting the coordinate plane, we load the layout that was used in the old facility. It is



(a) Cost table built using the traditional Levenshtein distance. (b) Cost table built using the modified Levenshtein distance. The replace operation is not possible in the annotated transition as the cache contains no more manual labor stations.

**Figure 4.17:** Comparison of a traditional with the modified Levenshtein distance.

comprised of lifts at both ends and a manual labor station, a corner element and a robot station in between (see Figure 4.15, top-left).

As we inspect the real-world sized layout on-site, we notice that a major issue of the former layout in the new facility is its cornered structure, as the second half obstructs the door on the right. We solve this issue by first removing the angled conveyor belt and then add the rotated robot station and lift back to the layout. After an analysis of our new layout, we notice that the robot station’s load is not ideal.

Rebalancing the load of the layout’s components is not a trivial task, so we start the layout discovery mode to find a more suitable layout. After inspecting some generated layouts, we end up with a layout that replaces the robot station with two additional manual labor stations (see wireframe model in Figure 4.15, bottom right). By inspecting the layout preview, we notice a remaining issue of this layout: its manual labor stations are facing towards the glass front of the hall, so they may be difficult to reach. Therefore, we edit the proposed layout one more time and turn the manual labor stations by 180 degrees, which results in a well-performing layout that meets the spatial restrictions of our current location. We further notice that the ceiling of the robot station barely fits under the stairway without colliding with it. In this case, there is no need to further change the layout. However, without an on-site inspection this problem may have stayed unnoticed, as the height and geometry of the staircase are not modeled by the simulator.

### Think Aloud Expert Interview

An expert interview was conducted with the procurator of a medium-sized wood processing company to evaluate the applicability of the approach. Among others, the expert needs to decide when and how to rearrange the company's production layout to better meet the current product order situation, which by experience is at least done once per year. The evaluation lasted approximately two hours and was conducted at the interviewee's production facility. At the beginning, the prototype was demonstrated and the workflow was explained for approximately ten minutes. A model-sized scaling was used to present a small layout on a table. To give the expert a first impression of what he will see, the presenter's view was streamed onto a screen. Then, the expert was given the opportunity to test the prototype. Afterwards, the evaluation continued in one of the company's production buildings, where the expert tested the prototype with the real-world scaling on the shop floor. He was asked to speak out his thoughts, no matter if they are questions, remarks, or general thoughts.

After getting familiar with the prototype, he explained that, independent of the prototype, the HoloLens has too many limitations to be really usable in its current state. Aside from generally known drawbacks of the HoloLens, such as its weight and the limited display size (and thus the field of view), he also mentioned that it may be challenging to remember the most important voice commands used by the system. Also, the interaction is currently not robust enough with regard to imprecise selections, for example, to be usable by a wider audience without specific training. Afterwards, he explained that the prototype's simulation mode is not directly applicable to his specific application case, as the work stations of the company are not always connected through conveyor belts. However, he still explicated that the planning mode would help during the planning phase. In a brownfield scenario, he deemed it much easier to make plausibility checks, for example, if more space is needed around a work station than it may seem in the planning software. In a greenfield scenario, he even mentioned that the approach should go further and start literally on a field of grass to make it possible to plan the building size as well as infrastructure like wiring and piping accordingly. One of the most important advantages is that the prototype helps to bridge the cognitive gap between a 2D planning software and the perception of how the planned layout will look like in reality. This is of special importance, as building plans often do not match the actual building neither completely regarding information such as piping nor accurately, e.g., regarding measurements. Also, the possibility of pretending actual interaction at a work station shown in its real size would significantly reduce planning mistakes regarding freedom of movement and manual workpiece transport. Finally, the expert also mentioned that the difference view in the discovery mode is very useful when presenting changes to a manager who has to decide if the proposed

changes should be implemented. He also advised reducing user interaction further for that scenario, for example, such that only the viable layouts can be chosen and no alterations are possible.

## Visual Event Analysis in Production Lines

Once a production line is set up and operational, the main goal of a manufacturer is to optimize the line's productivity. The workflow and individual process optimization are out of scope of this thesis, as it primarily requires the expertise of production engineers. However, the overall equipment effectiveness (OEE) (see Section 2.2.3), an often used key performance indicator, also includes machine downtime and the ratio of okay vs. not okay parts to describe the productivity of a production line.

These factors can be measured with increasing precision through the availability of a multitude of sensors that provide a constant feed of multivariate data. These data are often stored in large databases for a later analysis that can be used to further improve the factory's efficiency and effectiveness.

One way to improve the OEE is to increase the availability of a production line, for example, by reducing its unplanned downtime. Sending an operator to every broken machine is a self-suggesting action to return a production line into an operating state. However, this does not necessarily prevent issues from reoccurring at a later point in time. Visual analytics approaches can help domain experts to get a better understanding of causes and patterns of events to reduce occurrences in the future.

This chapter first gives a brief overview of relevant domain aspects and the data used in the later presented approaches (Section 5.1). After discussing related work (Section 5.2), two visual analytics approaches in the domain of event analysis are presented. They were designed to support machine operators and technical managers in improving the availability and quality measures by assisting in the analysis of error events reported in a production line. The first approach helps experts to understand how these events correlate (Section 5.3). The second approach focuses on supporting domain experts in finding recurring events over an extended time period to identify and prevent the causes of such error events (Section 5.4).

This chapter is partly based on the following publication:

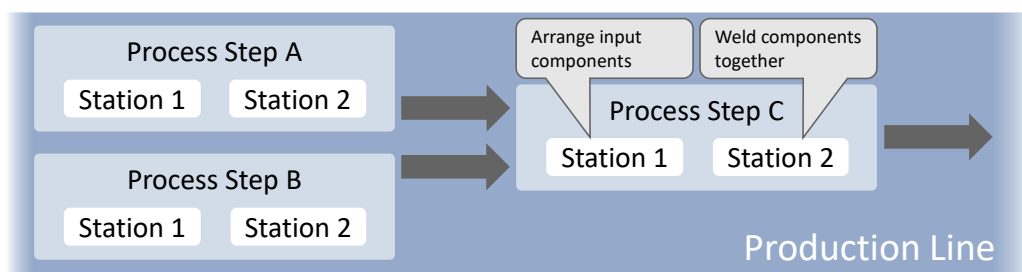
- D. Herr, F. Beck and T. Ertl. “Visual Analytics for Decomposing Temporal Event Series of Production Lines”. In: *Proceedings of the 22nd International Conference Information Visualisation. IV’18*. 2018, pp. 251–259 [1]
- D. Herr, K. Kurzahls and T. Ertl. “Visual Analysis for Spatio-temporal Event Correlation in Manufacturing”. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*. (To appear, accepted on 2019-08-19) [6]

## 5.1 Production Domain Introduction

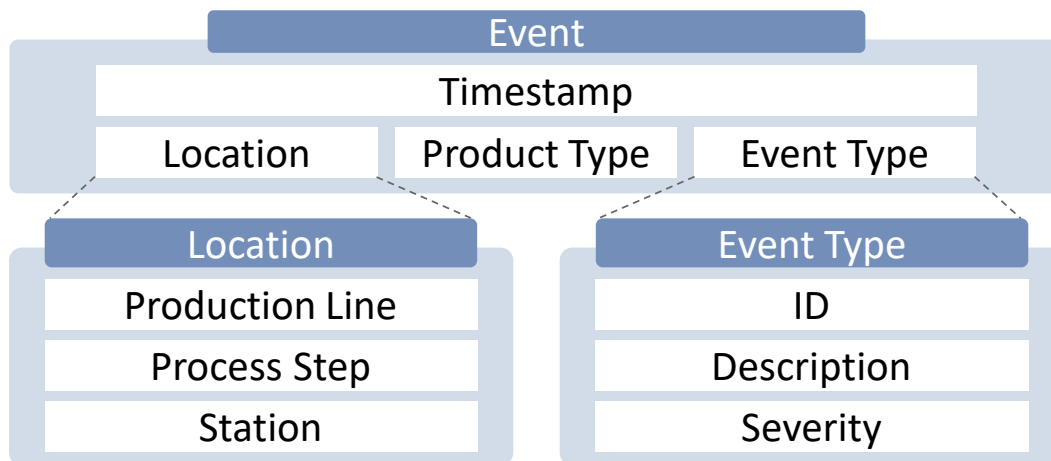
The following visual analytics approaches were developed in close collaboration with an industry partner. First, this chapter presents the general hierarchical structure of the studied production lines and the resulting specific characteristics of the data structure of the events reported by the machinery of the production lines (Section 5.1.1). Then, current measures to respond to events and to exchange observations and solutions among the staff on the factory’s shop floor and the production line that was studied in this chapter are introduced (Section 5.1.2).

### 5.1.1 Factory Hierarchy and Event Data Structure

Factories often produce a wide range of products. Typically, different types of goods are produced on different production lines. To optimize the workload of the machinery, similar products are often processed on the same line, where the machines may need to be adapted when switching between products. Any given *production line* comprises multiple *process steps* that manipulate products until they arrives at their final state, e.g., by combining two components by welding their connection points together. These steps may require multiple actions that are performed at separate *stations*. Typically, a station comprises one or more machine tools or human operators. In some cases, multiple stations perform long-lasting actions in parallel. Figure 5.1 shows an excerpt of a production line



**Figure 5.1:** A typical production line structure that comprises several process steps, each composed of several stations that perform the production tasks.



**Figure 5.2:** Data model of events used in the remainder of this chapter. Each event comprises a timestamp, the location of the reporting machine, the product type that was processed, and an event type.

that produces multiple components in parallel and then combines them into one product. In the given example, the last process step would comprise two stations: one for the component placement and another for welding parts together.

In case an error occurs, the affected station files a structured event report (see Figure 5.2). Each event instance has four components: a timestamp, the location of the event, the product type that was processed at the time of the event, and the type of the reported event. To distinguish events from their specific time-dependent instances, the term *event class* will be used in the following to describe general, time-independent events.

The date and time when the event occurred are provided in the *timestamp*. Details about the *location* of each event are structured according to the factory hierarchy presented before. The *event type* is detailed by an ID, a human-readable description, and a severity level (information, error, etc.). In the following, only event reports that have the severity level *error* are considered, as these have the highest impact on productivity. Event type information could be arbitrarily complex and range from a simple ID that may only identify an event’s severity (e.g., a low or high-risk factor) to complex information such as a precise event description (e.g., human-readable text that describes the event).

### 5.1.2 Specific Scenario of the Industry Partner

The collaborating industry partner produces small to mid-sized electric motors. Among others, the production lines record data of its stations’ process times, what product type is being produced, and events reported by the deployed machine tools. Currently, the data is used to monitor the processing times and,

if they deviate from their expected values, the cause is investigated at the shop floor. Further, currently persisting error events and information about material shortage are visualized on a display above the production line, a so-called Andon board [67]. In combination with indicator lights above the machines, the display helps operators on the shop floor to respond to malfunctioning machinery.

The data analysis is done on two levels: during a daily meeting, all technicians discuss recently occurring problems. The analysis of specific problems is usually conducted by looking at the process times and occasionally at the reported events during a specific time period, which is usually one day. Further, the head of the production line occasionally analyzes the process times of various stations. Such an analysis may include the data of the past hours up to a few weeks. If the head of the production line finds any anomalies, he consults the operators at the shop floor to get further details about the identified problem and assesses solution strategies.

The studied production line produces six types of motors with 84 variants overall. It comprises 19 process steps with 96 stations in total. The primary goal of the following visual analytics approaches is supporting domain experts tasked with the analysis of the production line, but also to support operators working on the shop floor of the factory.

## 5.2 Related Work

The approaches presented in this chapter focus on the visual analysis of spatio-temporal event data regarding their correlations and temporal occurrence patterns. Therefore, the following related work first summarizes approaches to visually analyze event relationships. Afterwards, previous work in the field of event series analysis is shown. Finally, an overview of approaches to support experts in the manufacturing domain through visual analysis is given.

### 5.2.1 Visual Analysis for Event Relationships

The need to analyze event relationships is not exclusive to the manufacturing domain. In the past, several visualization techniques and their combination were proposed that show the correlation of dimensions in high-dimensional or multivariate datasets. Zhang et al. [208] make use of scatterplots and parallel coordinate plots (PCPs) to visualize the relations between key features when using biomarkers to analyze their effect with different kinds of cancer. Wang and Mueller [191] first build multiple causal models, which they merge to find each correlation's credibility based on the number of causal models that contained the correlation. They show the result in a correlation map, as well as with a scatterplot and a parallel coordinate plot. Zhang et al. [209] also use PCPs and



a graph structure to visualize the relation of numerical, as well as categorical dimensions. They encode the correlation information in the graph's edge length as well as their color. Behrisch et al. [25] present an approach that allows to compare sets of matrices of varying size. Their technique can be used to get a better understanding of similarities across different data sets. To make the relations within a dataset comprehensible, Alsallakh et al. [14] use a star plot-like approach. Their *contingency wheel* provides information about the individual dimensions towards the outside of the plot, while they use the space at the center to provide information about the relation of the dimensions.

The prediction and handling of error events and attacks in networks are similar to the correlation of events in a production line, as it can also be interpreted as directed graphs in which packages (products) are transported. Qin and Lee [155, p. 95ff] assume a network monitoring scenario in which security administrators are hampered by the abundance of alerts reported within a system. To reduce the amount of shown information, they build a correlation graph that clusters alert patterns based on the starting node of the pattern. In another approach, Qin and Lee [156] take isolated attack alarms and correlate them to extract *attack plans*, which can be used for attack prediction. They correlate extracted scenarios to build causal networks that can be used to identify attack patterns. Xie et al. [204] use Bayesian networks to detect attacks with a focus on real-time detection and the analysis of the attacks' scope, severity level, possible consequences, and potential countermeasures.

Sedlmair et al. [166] analyze in-car communication network data to identify error and warning messages by first defining state machines for the messages and then visualizing the states' transitions over time. Further, the user can visually compare multiple state machines to find possible dependencies. Shi et al. [169] show network sensor data on a radial tree view that represents the network's routing logic to detect outliers that need to be investigated. They use a correlation graph to show the similarity of the sensor nodes, which can be used to find outliers in event series. Steiger et al. [175] use time-series data of sensors with known geo-positions and visualize the relationship of the sensors regarding their geographic position and the similarity of their time-series.

Alternatively, relations can be implicitly used to separate the data into groups, which can then be presented to users. Behrisch et al. [26] show samples of scatterplot views that visualize only two dimensions of the dataset at the same time. Users can label good and bad samples, which is used to build a decision tree that classifies the data and refines the proposed samples. Similarly, Heimerl et al. [92] train a support vector machine to classify documents by asking the users to select badly classified documents on a scatterplot. Xu et al. [205] propose a co-clustering approach for bi-partite graphs to find common attributes that allows users to derive insights. They present the results as a hybrid visualization that connects entity clusters, represented as treemaps or matrices,

through edges. The former approaches may have used temporal information during their data processing, but they did not expose it, although it may contain important further information. Beck et al. [24] give an overview of dynamic graph approaches, which often express the temporal information. Jäckle et al. [101] extend multidimensional scaling (MDS) to include temporal information, which provides insights about the correlation of dimensions over time.

If the events contain spatial information, spatio-temporal analysis approaches often include a map that shows the events' occurrence over time. As an example, *Voila* [40] enriches a city map with a heat map to provide information about abnormal patterns. Often, such approaches use additional views that provide additional information or enable users to analyze the events' trends or periodic patterns, e.g., as shown by Malik et al. [131].

Many of the presented approaches use graphs or similar structures to model their correlations without showing any temporal information. However, one goal of the approaches in this chapter is to preserve and visualize the temporal aspect of the correlating events. As dynamic graphs quickly get complicated to interpret, the presented approaches use other visualizations to represent the data.

## 5.2.2 Event Series Analysis

Analyzing temporal patterns of event series is also relevant outside the manufacturing domain. For instance, there exist event visualization tools focusing on security issues [97, 119], meteorological and oceanographic events [23], or historic events manifested in documents and media [15, 128, 111]. To perform analyses with time-series data, an approach to preprocess the data can help to clean the data beforehand. For example, Bernard et al. [27] present an approach to allow domain experts to interactively design preprocessing pipelines of time-series data. Like the second approach of this chapter, they visualize events over time. However, they do not allow for decomposing the time series interactively into trends and seasonal components.

There exist only a few other visual analytics approaches that build on such a decomposition of time series. Bögl et al. [34] provide interactive visual guidance for selecting appropriate parameters of *autoregressive integrated moving average (ARIMA)* and seasonal ARIMA models, which decompose a time series similarly to STL (see Section 2.1.3). They extended their work [32, 33] adding predictive analysis features to their approach, allowing for more insights. Chae et al. [45] apply STL on Twitter data to filter out any seasonal and trend effects to visualize unusual events. They assume that the outliers contained in the data without trend or seasonal series indicate special events that could be of interest. Maciejewski et al. [129] use STL to forecast hotspots of geo-located events. These works are rather interested in prediction and outlier detection than explaining the time series to the analyst by applying a decomposition. The presented approach

in Section 5.4 uses multiple decompositions to reflect different types of events through a single steerable model to decompose the event series.

Some information visualization techniques of time series particularly highlight seasonal patterns, for instance, spiral plots of the time axis [43, 193]. Alternatively, the time series can be split by season and plotted overlaid in 2D, juxtaposed in 3D, or encoded in color in a calendar grid (or any other 2D grid) [189]. This procedure scales even to large time series when color coding the values of the series in a pixel grid [104]. Showing the series in different resolutions with different seasonal lengths can reveal different patterns, which themselves can be juxtaposed [170]; it becomes difficult, however, to see trends and seasonal patterns because the time series is not explicitly decomposed. Cycle plots [159] use a form of dimensional stacking to compare data points from different seasons on a linear axis; this is limited to a single decomposition with a single season length at a time. However, this chapter's second approach (Section 5.4) allows to support the analysis of multiple decompositions with different season lengths.

To efficiently identify causal dependencies of events, the identification of event sequences is very important. Several approaches exist that extract and visualize sequential patterns (e.g., Guo et al. [81]), present the distribution of common sequences (such as Wongsuphasawat and Gotz [199]), or even allow to search for fuzzy sequences (for example Chen et al. [47]). As datasets may contain a large number of such sequences, Cappers and van Wijk [41] present an event querying system that allows for fuzzy searches that provide feedback, which event sequences match the searched pattern. Likewise, Krüger et al. [113] use a visual query language to highlight semantically annotated events to extract or confirm complex movement sequences. Instead of filtering the data on the model level, Monroe et al. [139] reduce the visual complexity to assist users in getting an overview of relevant data.

### 5.2.3 Visual Analytics in Manufacturing

Most visualization and visual analytics approaches that target the manufacturing domain focus on the optimization of simulations and production schedules. However, recently there is also an increasing number of approaches focusing on helping domain experts to monitor and analyze their production lines as well. Works regarding the simulation of production lines were already covered in Section 4.1. Therefore, the following primarily focuses on works regarding the production of goods and their scheduling.

Before the actual production of goods can begin, it is necessary to schedule at what point in time different products should be produced. To optimize these production schedules, Klöpper et al. [106] introduce a system that generates a set of possible production schedules that can be iteratively reduced based on aspects that experts deemed to be currently most important. *LiveGantt* [102]

helps experts to explore Gantt charts of large concurrent schedules. Users can interact with the schedule and get visual feedback about their changes' effects.

*ViDX* [206] analyzes a production line's performance based on product tracking data with the goal to better understand the effects of machine problems. It extends a Marey's graph [183, p. 31] to visualize products moving through a production line. Outliers are visually emphasized by aggregating products with similar process times. Further, the approach provides real-time tracking of a production line's performance. The visualization of individual products and their processing times improves the understanding of a line's performance and helps to analyze the effects of problems in a production line.

In contrast to *ViDX*, the first approach presented in this thesis (Section 5.3) focuses on the detection of systematic issues in a production line based on event reports filed by machine tools with the goal of finding recurring issues over an extended period of time.

### 5.3 Visual Analysis for Spatio-Temporal Event Correlation in Production Lines

Monitoring systems with live prediction of possible issues in the machinery are often a desirable goal. The development of systems that support predictive maintenance requires expert knowledge to help understand the complex relationship between different events. However, the knowledge of possibly existing correlations is based on the experience of domain experts and sometimes limited (e.g., regarding events that occur with some delay). Further, it requires specialized knowledge to decide if statistical correlations are also semantically plausible. Once this information is accessible, event sequences can be labeled to improve machine learning methods to detect correlations automatically. Visual analytics can support the reasoning for common event analysis tasks and communicates complex changes in events during the development and deployment of such monitoring systems. In addition, experts can also directly use gained insights to improve the productivity of a production line.

The following approach focuses on providing insights to domain experts about the correlation of errors with the goal to derive possible event causalities using the experts' experience and knowledge. Its requirements were analyzed in an iterative design process and design decisions for multiple coordinated views were derived based on the domain experts' analysis tasks. A prototype of the developed visual analytics approach was implemented (see Figure 5.3) to derive insights that can potentially increase the productivity of the analyzed assembly line. It fosters the interplay of event timelines, correlation plots, projections, and a spatial layout view, which supports hypothesis building and validation.



**Figure 5.3:** Experts can inspect the individual error event occurrences in the *Timeline View* (A). The *Location View* (B) provides information about the spatial distance between the selected stations. Stations that are similar regarding the patterns of reported events can be identified in the *Location Projection View* (C). To verify these correlations, the locations' correlation over time can be inspected in the *Correlation View* (D).

The combination of different data views is presented in a case study, which demonstrates how answers for typical domain-specific questions can be found. Afterwards, domain expert feedback of the approach is presented.

### 5.3.1 Approach

During the initial meetings with experts from the industry partner, it became clear that the domain experts, who are mostly technical engineers, prefer visualization approaches that are close to their current visualization expertise, which is mainly the interpretation of heatmaps, line plots, and raw tabular data. To provide an easier entry point for engineering experts and to address the different levels of complexity of the requirements, this approach combines views that provide views known by the experts, such as the layout of the production line, and scatterplots that show the correlation of the error classes, which are connected through brushing & linking.

**Data Model and Processing** The approach assumes a data structure as presented in Section 5.1. An example of such an event is:

Timestamp	Location
$\overbrace{04.12.2017\ 06:11:38; \text{Line}=1; \text{PID}=4553; \text{Station}=2;}$	
$\underbrace{\text{Description: Fehler bei Bewegung X125D X-ACHSE LINKS}}$	
Event Type	

To gain insights about interdependencies between event classes, an additional event correlation metric to describe the event classes' dependencies is required.

**Pairwise Event Correlation** Based on the experience of the industry collaboration partner who provided the data, two scenarios can result in an event relation:

- a process step has a problem and therefore all stations that are part of the process step report the same event, and
- stations that are part of different process steps report events. If the latter event(s) are caused by the former, they are most likely caused by a product that is being processed by both stations.

The first case is trivial, as checking if the events are part of the same process step is sufficient. The second case is more complicated, as it is necessary to know how long it takes for a product to be transported from station A (which *causes* the relation) and station B (which is *affected*).

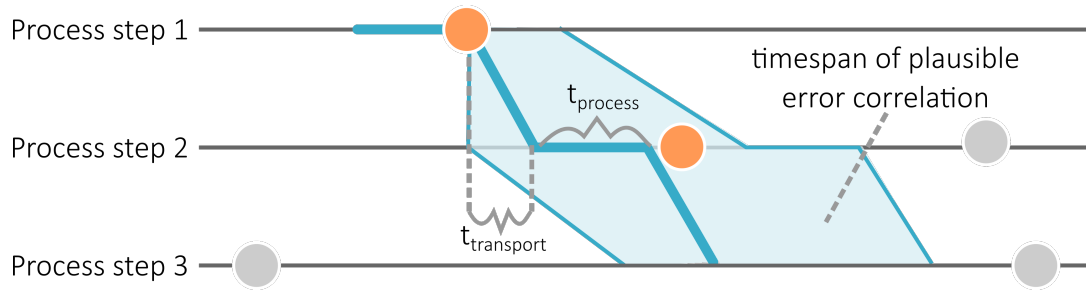
To reflect these scenarios, the correlation metric presented in the following assumes that correlating events are caused by the products that are transported between the stations. To derive a pairwise correlation between event classes, there needs to be an understanding of when two events are similar. This can be formalized as a function that takes the temporal and spatial difference of two given events into account.

As the products are transported on a conveyor belt that runs at a constant speed, the distance can also be defined through a *temporal distance* that describes how long it takes a product to travel between two given stations. Such a measure requires data specific to each production line, such as the transport time on a conveyor belt between stations and the process times of all stations the part or product passes through. In this thesis, the correlation of two event classes is defined as the quotient of the number of plausible co-occurring events by the sum of the individual event occurrences.

Formally, the pairwise correlation  $C(E_A, E_B)$  of two event classes  $E_A$  and  $E_B$  is defined through the correlation plausibility of the event classes' instances:

$$C(E_A, E_B) = \frac{\overbrace{\sum_{e_i \in E_A} \sum_{e_j \in E_B} IsPlausible(e_i, e_j)}^{\text{plausible co-occurrences of } E_A \text{ and } E_B}}{\underbrace{|E_A| + |E_B|}_{\text{occurrences of } E_A \text{ and } E_B}}, \quad (5.1)$$

where  $IsPlausible(e_i, e_j)$  defines if the event  $e_i$  possibly caused  $e_j$  based on their spatial and temporal distance:



**Figure 5.4:** Schematic description of the used fuzzy matching regarding the first reported event. The time frame is calculated based on the transport time between the stations and the processing time of the stations themselves and allows an offset of  $\pm 10\%$ . Events within an allowed time frame are colored in orange.

$$IsPlausible(e_i, e_j) = \begin{cases} 0 & e_j\text{'s station is located before } e_i\text{'s,} \\ 1 & \text{the start time of } e_j \text{ is within a reasonable,} \\ & \text{timeframe after } e_i \text{ was reported (see below)} \\ 0 & \text{else.} \end{cases} \quad (5.2)$$

The transport times  $t_{transport}$  between the stations and the passed stations' processing times  $t_{process}$  are summed up to decide, how long the transport of a product between two stations should take. In addition, an uncertainty factor allows to compensate for dynamic changes in the actual production process (e.g., unexpected delays). For the investigated dataset, a temporal deviation of up to  $\pm 10\%$  is allowed for the estimated time for transport and processing. Figure 5.4 demonstrates the described event matching, where colored events are considered as a match, whereas grayed out events are mismatches.

### Requirements and Design Decisions

In many application scenarios, the overarching goal for event analysis is the development of a reliable predictor that is capable of foreseeing issues before they happen, allowing for a faster response by personnel at the shop floor. A general understanding of the underlying data and the occurring chains of events that lead to critical issues is important to design and train such predictors. Hence, this visual analytics approach aims to provide insights about the relations between events and their spatial and temporal coherence.

The following requirement analysis is based on typical research questions for spatio-temporal data as proposed by Andrienko et al. [16]. As such, the requirements can be categorized according to the four categories *when*, *where*, *what*, and *relational* coherences. As the approach focuses on the extraction of possible event relationships and not their comparison, the category compare/relate is not

considered. To provide concrete examples, the following includes questions from the manufacturing domain that fit in this categorization and can be solved with the presented approach:

### Category 1: **When**

The temporal data dimension provides important information to answer multiple questions:

- ①  $Q_1$  When did an event (re-)occur?
- ②  $Q_2$  In which order did different events occur?

To answer these questions, a timeline representing the discrete occurrence of single events is one of the most common and therefore familiar visualizations to many people. As a single timeline with pictograms representing different event types is limited in terms of scalability, the events are distributed based on their location along the vertical axis. The industry partner's domain experts explained that it is important to investigate the temporal order of issues in the log files with respect to the questions ① & ②.

### Category 2: **Where**

The spatial context becomes important to identify specific locations that might be involved in a chain of events:

- ③  $Q_3$  Where did an event occur?
- ④  $Q_4$  What is the spatial relationship between events that occur together?

All questions related to the spatial context can be intuitively represented on a map or plan. In the given production line example, the shop floor's layout is provided containing the machines and the conveyor belts that the products are transported on. Such a visual representation makes it easier to understand the spatial distance between the locations that reported events, the stations' connectivity, and the *hierarchical structure* of the production line, as described in the Sections 5.1.1 and 5.3.1.



#### Category 3: **What**

Finding the details about a specific event that incorporates the information when and where it was reported:

- Ⓚ<sub>5</sub> What happened when an event occurred at a specific time?  
*The meaning of an event is not always included in the data. Usually, domain experts are required to answer this question.*

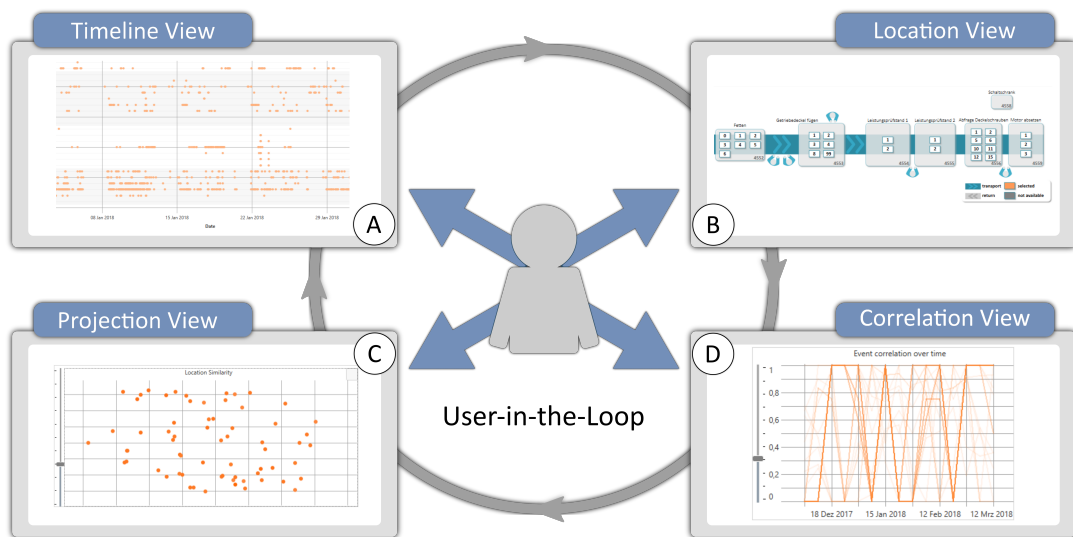
Combining the information provided by the timeline and the map helps to reconstruct what caused an event. In the given scenario, the event types are already annotated by the industry partner to include human-readable text (e.g., which component of a machine is not operational).

#### Category 4: **Relational**

Understanding how different events relate to each other is important to identify possible chains of events that lead to failures:

- Ⓚ<sub>6</sub> Which pattern does an event belong to?
- Ⓚ<sub>7</sub> Which events usually occur together?
- Ⓚ<sub>8</sub> Which co-occurrences are persistent and which are outliers?
- Ⓚ<sub>9</sub> Which locations are similar regarding reported events?

The most common approaches to visualize relations between data are matrix and graph visualizations. However, these approaches have limited capabilities when it comes to visualizing temporal changes. Time-to-space approaches often result in visualizations that require much space and time-to-time approaches are unsuitable for interaction due to their changing content. Other approaches, such as hybrid visualizations (e.g., combining graphs and matrices), were discarded due to the aforementioned requirement that the proposed visualizations should be similar to visualizations that the experts are familiar with. Such approaches are often intransparent regarding the way they aggregate the presented data. To take the temporal changes in the dataset into consideration, two views present temporal statistical measures and the overall relatedness of the data: (1) a projection view that indicates high correlations between event types or locations based on their events by spatial proximity and (2) a line plot that




**Figure 5.5:** Visual analysis approach with four linked views for spatio-temporal event analysis: (A) In the *Timeline View*, individual events can be investigated in full detail. (B) The *Location View* provides the spatial context to specific events. (C) The *Projection View* helps identify multiple co-occurrences. (D) The *Correlation View* displays pairwise event co-occurrences over time.


shows pairwise correlations between event types over time that has analogies to a parallel coordinates plot where the discrete time steps are the dimensions. In combination, the relational questions can be investigated with the proposed analytical approach. The case study (Section 5.3.4) exemplifies how different issues related to each other can be found.

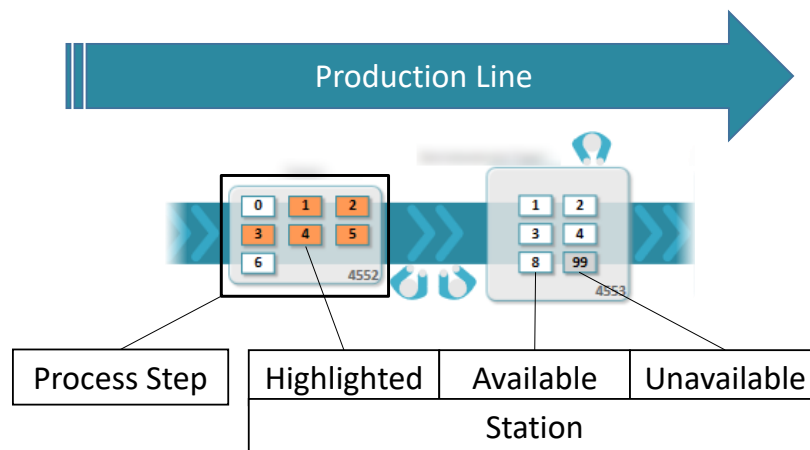
### Visual Analytics Approach

The approach consists of four linked views (see Figure 5.5) that provide an overview and detailed temporal (A), spatial (B), and relational (C & D) information about a dataset. In addition, experts can search and filter for specific event messages through a text search component that is only shown on demand, as it does not convey any additional information, unlike the other views. The general design concept aims for a combination of abstracted overviews for correlation analysis and detailed views for the temporal and spatial components of the data that represent the underlying data domain as close as possible.


(A)  **Timeline View** This view provides a detailed plot of individual events over time. Each row corresponds to a station in the production line. The rows are ordered first by the process step and then by the station number. An alternating background color helps to distinguish, which rows belong to the same process step. The horizontal axis provides temporal information. A tooltip

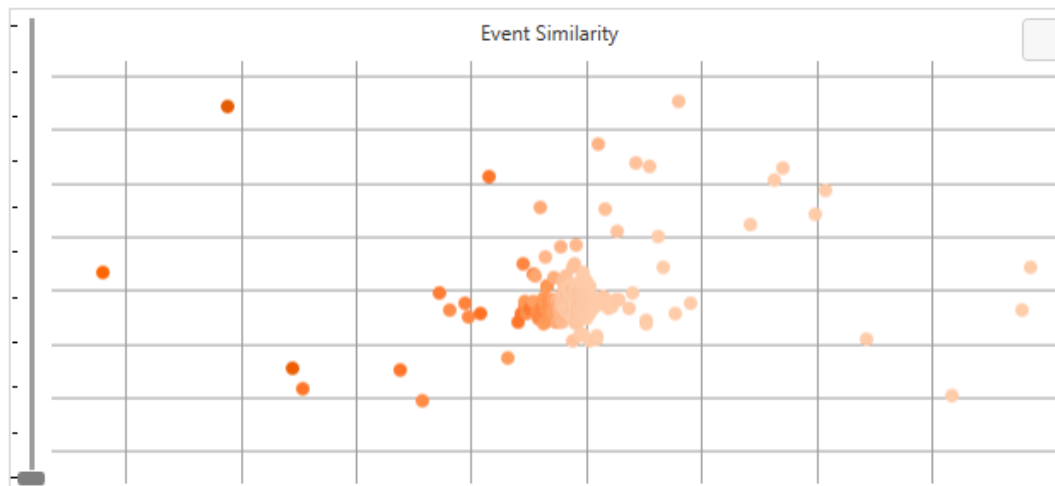
provides additional information about the individual events, e.g., their detailed description. By selecting an area on the timeline, the contained events' locations and event classes are highlighted in the other views.

Ⓑ  **Location View** Spatial context is important to relate events with the location of the process steps and stations where they occurred. Hence, a map with all stations that report events is available in the *Location View*. Stations that do not contain any events for the analyzed range of time are grayed out. Selected stations are highlighted in orange in the *Location View* as well as in the others. In case experts select events in other views, the stations where the selected events were reported from are highlighted. Figure 5.6 shows an enlarged excerpt of the *Location View* where multiple stations in the first process step were selected. This view helps to interpret correlations from the views Ⓒ (temporal correlation) and Ⓓ (projection of event/location similarity) and provides detailed information about the spatial domain of the data.




**Figure 5.6:** Hierarchical structure of a production line. Each line has several process steps that can be seen as tasks (e.g., drilling). The process steps contain stations that perform the same task in parallel to increase the processing speed. The stations have a visual indication if data at a station is highlighted (orange), available (white), or if no events were reported (gray).

Ⓒ  **Projection View** Since event classes can correlate although they are located at process steps that are far apart, a general overview of their overall correlation is necessary. This approach includes a view that projects the correlation matrix of the event classes or locations based on t-distributed stochastic neighborhood embedding (t-SNE) [186, 187] onto a two-dimensional plane. This way, the potential correlations are displayed through spatial proximity.



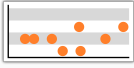


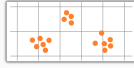
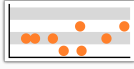


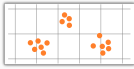
**Figure 5.7:** The quality of the points in the *Event Projection View* is encoded in each point’s color saturation. A light orange indicates a bad result, whereas darker orange indicates a good result.

Due to the data loss during the dimension reduction process, not all of the event classes can be placed correctly. To prevent users from assuming event type relations due to falsely placed event types, the placement quality of each event class is indicated in its corresponding point’s color (see Figure 5.7). The placement quality uses the quality measure proposed by Mokbel et al. [137]. Users can set a minimum projection quality threshold through a slider control to hide badly placed event classes.

**D**  **Correlation View** This view provides an abstracted overview of the pairwise event class correlations over time as a line chart. Each line corresponds to a specific pair of event classes. On the horizontal axis, the temporal dimension of the data is discretized by modifiable time intervals. The default interval is one week due to the production line’s production schedule. Within each interval, the correlation measure defined in Equation 5.1 is used. To prevent data loss, only the “source” event must start within the time interval, whereas the “affected” event can also be part of the following. As this view is prone to visual clutter, experts can use a slider to set a threshold to filter the shown event class pairs either based on their average or their maximum correlation. Selecting lines in the correlation view highlights those lines in orange, whereas other lines are grayed out. The affected event classes and their corresponding events are also highlighted in the other views.

**Combination of Views** Table 5.1 gives an overview of the task categories and which tasks can be answered through a combination of the presented views.

Table 5.1: Comparison matrix for the provided views and their capability to answer for the four question categories *when*, *where*, *what*, *relation*.

				
	<b>when</b> where <b>what</b> relation	<b>when</b> <b>where</b> <b>what</b> relation	<b>when</b> where <b>what</b> <b>relation</b>	<b>when</b> where <b>what</b> <b>relation</b>
	<b>when</b> <b>where</b> <b>what</b> relation	when <b>where</b> what relation	when <b>where</b> <b>what</b> <b>relation</b>	when <b>where</b> <b>what</b> <b>relation</b>
	<b>when</b> where <b>what</b> relation	when <b>where</b> <b>what</b> <b>relation</b>	when where <b>what</b> <b>relation</b>	when where <b>what</b> <b>relation</b>
	<b>when</b> where <b>what</b> <b>relation</b>	when <b>where</b> <b>what</b> <b>relation</b>	when where <b>what</b> <b>relation</b>	when where what <b>relation</b>

Examples on how to derive insights of the corresponding categories are provided in Section 5.3.4.

**System Architecture** The prototype (see Figure 5.3) is implemented with C# and .NET Framework 4.6. The data is stored in a Microsoft SQL Server database that contains the reported event data. Dapper<sup>1</sup> was used to map the data from the database to the prototype’s data model. The front-end is implemented as a Windows Presentation Foundation (WPF) desktop client. During an analysis run, data from a specified timespan is retrieved from the server and processed on the client. Some of the views, such as the *Correlation*, *Timeline* and *Projection Views* use the SciChart WPF Framework<sup>2</sup> to present the data.

<sup>1</sup> <https://github.com/StackExchange/Dapper>

<sup>2</sup> <https://www.scichart.com/>

### 5.3.2 Evaluation

The approach was evaluated with three domain experts from the industry partner in two stages:

1. First, the findings were gathered with the introduced approach, according to the questions  $Q_1$ – $Q_9$  from the catalog shown in Section 5.3.1. Section 5.3.3 presents these findings and Section 5.3.4 shows, how these they can be extracted.
2. Then, the three domain experts were interviewed in a feedback session. Along with the findings, they got a demonstration, how the presented findings can be derived with the approach. Following the demonstration, questionnaires were handed out for rating the findings, the importance of the individual questions, and the visualization views. Afterwards, individual findings and improvement suggestions were discussed with the experts.

The experts were asked to rate the insights, importance of the research questions, and the system's views in a qualitative user study. A Likert scale that ranged from 1 (not expected/not useful/impossible) to 7 (very expected/very useful/very easy) was used for the ratings. Further, the experts had the option to give no answer.

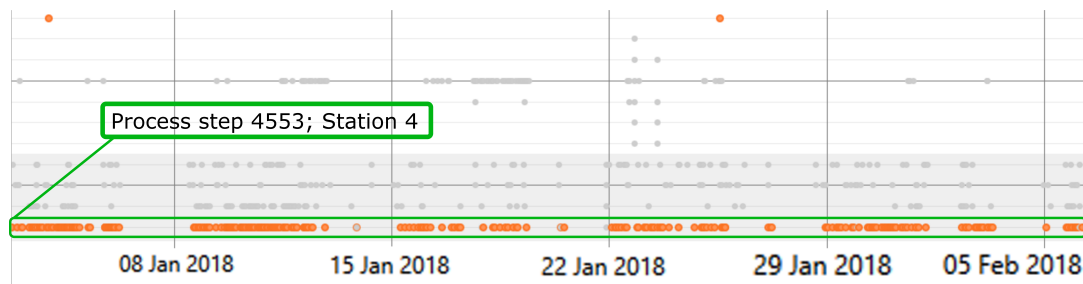
The first expert was only involved in the evaluation and did not know the system before, whereas the latter two experts were also involved during the development of the prototype. The same experts participated in all parts of the study. One expert is responsible for the production line where the analyzed events originated from. The second expert is a project leader tasked with the advancement of industry 4.0 concepts, for example, to make collected data available to workers in an understandable way. The third expert is a team leader and responsible for the implementation of the accessibility of data on the shop floor level in the factory.

### 5.3.3 Findings

Initially, nine findings were derived with the presented prototype. Table 5.2 presents these findings along with ratings from two of the three domain experts (see Section 5.3.2 for details). The second expert opted out of the insight rating, as she is not involved in the daily routine of the production line and cannot assess the plausibility of specific findings. Section 5.3.4 demonstrates, how these findings can be acquired.

Table 5.2: Nine findings presented to the domain experts. All findings were derived with respect to the related analysis questions. The experts rated the findings according to how useful they are to improve the manufacturing process (1 = not useful – 7 = very useful). Some of the findings were anonymized to protect the industry collaboration partner’s intellectual property.

Finding	Description	Related Question	Expert Ratings
<b>F1</b>	The error <i>Fehler bei Bewegung AAA</i> at process step 4553 occurs frequently (varies between 15 minutes and two hours).	Q <sub>1</sub> , Q <sub>2</sub>	5 / 4
<b>F2</b>	The error <i>Stoerung gesamt Taktachse / Roboter</i> at process step 4543 occurs regularly (often errors are reported within minutes up to an hour, sometimes there are gaps of several hours).	Q <sub>1</sub> , Q <sub>2</sub>	6 / 6
<b>F3</b>	Station 1 in process step 4552 rarely reports any errors, but if it does, then it reports <i>Fehler Kinematik 1 (siehe Intramotion)</i> several times in a short timespan (less than an hour)	Q <sub>1</sub> , Q <sub>2</sub>	5 / 5
<b>F4</b>	Process step 4546 and process step 4553 are spatially almost half of the production line’s length apart.	Q <sub>3</sub>	5 / 4
<b>F5</b>	If the error <i>Fehler Kinematik 1 (siehe Intramotion)</i> is reported, a technician can quickly look up what process step this error belongs to and where it is located.	Q <sub>3</sub>	7 / 6
<b>F6</b>	At process step 4552, the error <i>Fehler Kinematik X*</i> ( <i>siehe Intramotion</i> ) often occurs at most stations (1-5) at the same time. (X* depends on the station that reports the error.)	Q <sub>3</sub> , Q <sub>4</sub> , Q <sub>6</sub>	5 / 6
<b>F7</b>	Process step 4546, Station 2, and process step 4553, Station 1, have a cause-effect relationship regarding their reported errors (→ if something breaks at 4546, then there is a chance that something will break later at process step 4553).	Q <sub>7</sub>	5 / 6
<b>F8</b>	At process step 4549, Station 8, the errors <i>Fehler Stellglied BBB</i> and <i>Fehler Stellglied CCC</i> occur often and usually occur together.	Q <sub>8</sub>	4 / 6
<b>F9</b>	The reasons vary, but if there are problems at process step 4543, then there is also a chance that there are problems at process step 4547.	Q <sub>9</sub>	5 / 6



**Figure 5.8:** The *Timeline View* provides a quick overview of which event types occur rarely (highlighted events at the top) and frequently (highlighted events at the bottom). The highlighted events within process step 4553; Station 4 contain the error type *Fehler bei Bewegung AAA*.

### 5.3.4 Case Study

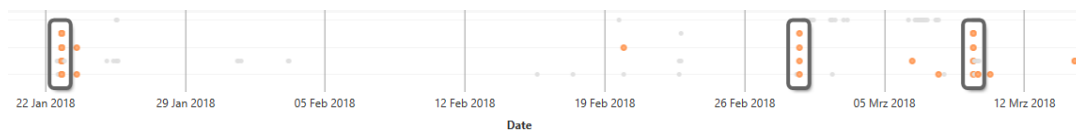
Following the general analysis questions introduced in Section 5.3.1, the following case study demonstrates, how the approach can be used to derive the findings presented in Section 5.3.3. The dataset analyzed for the evaluation comprised 20,872 error events reported over a timespan of four months. All events occurred in the same production line, which contains 19 process steps and 96 stations in total. As different parts of the event data are relevant to answer a question (e.g., temporal or spatial information), users may choose different views to enter the analysis.

#### When did errors occur? (F1–F3)

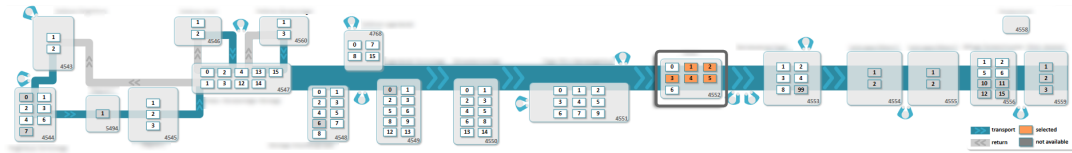


Some of the most relevant questions regarding the temporal aspect of the data are if there are events that occur very frequently or if they set in at a certain point in time and then stop again. Especially the first question can be efficiently solved with the *Timeline View*. For example, experts may search for a period of time where a station reports many events. When selecting this range, they have access to all events during the selected time at the selected station. This gives a quick overview if the reported events are all of the same class or if multiple event classes contribute to the high number of error reports. To inspect a specific event class, users can filter for all events of the type of interest (e.g., by using the text search) to inspect its occurrences in the *Timeline View*. In Figure 5.8, process step 4553; Station 4 reports a high number of events. After selecting the events during the second week of January, the selection output shows that 139 of the 143 events are the event type *Fehler bei Bewegung AAA*. By highlighting all errors of this event class, it becomes clear that this error is the most often occurring event type throughout the entire analyzed range of time (880 of 1069 events). This was reported as Finding *F1* to the interviewed experts (see Table 5.2).





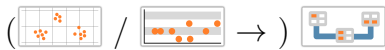
(a) *Timeline View* with a highlighted recurring pattern.



(b) *Location View* in which the stations of the selected pattern are highlighted.

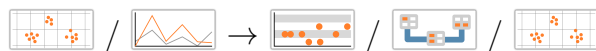
**Figure 5.9:** The *Location View* provides experts with a spatial context to selected events. The annotated pattern in the *Timeline View* 5.9a indicates that various events occur at the same time. Further, this pattern repeats at several points in time. The *Location View* 5.9b shows, that the stations, where the errors occurred, are all part of the same process step and where this process is located in the production line.

**Where are the stations that reported errors? (F4 & F5)**



The most intuitive way to solve questions related to space is to use the *Location View*. When combining it with other views, experts can either quickly locate the station that reported a specific error or find event types related to the location selected in the *Location View*. In more complex scenarios, the view can be used to provide a link between an abstract pattern of events. For example, the pattern highlighted in Figure 5.9a seems to recur over time. Through a selection of one pattern occurrence it becomes apparent that all events occur at the same process step, but at different stations (see Figure 5.9b). This insight is part of the Findings *F5 & F6* (see Table 5.2).

**Which errors or locations relate to each other? (F6–F9)**



Due to its complexity, finding relations between event classes is not as straightforward as the other questions. Usually, an interesting pattern or unexpected outlier leads to the need to inspect the relation between event classes or locations. Therefore, this approach is not designed with a fixed workflow, but it allows to start at any view to conduct the analysis.

One possibility is to start with the *Event Type Projection View* and select two or more event classes that are close to each other. To verify the event type’s relation, experts can check the *Location View* to quickly assess if the relationship is plausible or use the *Timeline View* to see the distribution of the highlighted



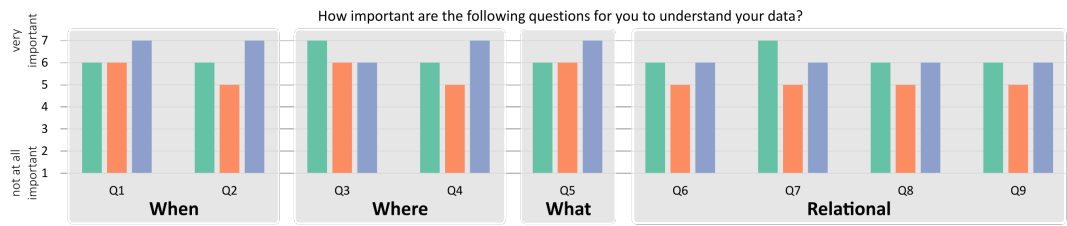
**Figure 5.10:** Through the selection of a clearly **separated group of events types**, it becomes apparent that there are error dependencies between the process steps **4547**, **4768** and **4549**. An expert can verify this hypothesis through the *Correlation View* and the *Timeline View*.

errors. Alternatively, the *Correlation View* allows to check if the correlation persisted during the analyzed period of time or if the correlation is temporally restricted. In Figure 5.10, a group of event classes that is clearly separated from the other classes (see other rectangle) was highlighted. The selected events were all parts of the process steps 4547, 4768, and 4549, which are highlighted in the *Location View*. The process steps are emphasized with blue, green, and purple borders respectively. In addition, the experts can use the *Correlation View* to verify, which specific error types have a high correlation and use the *Timeline View* to see, at what times the individual events occurred. Finding *F8* (see Table 5.2) was derived analogously to the presented example.

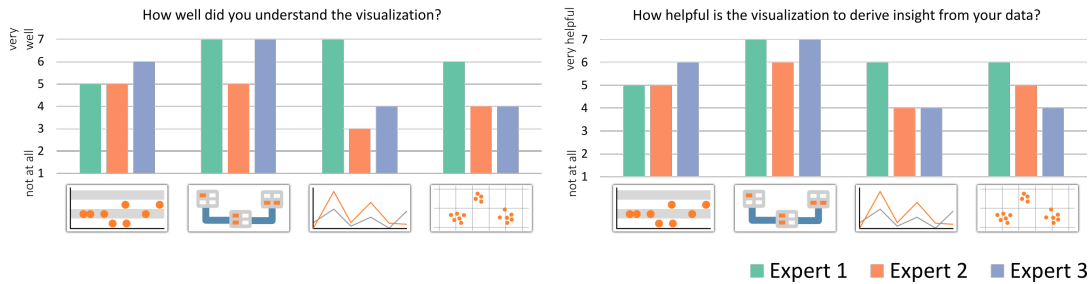
Another possibility to find relationships is to use the *Location Projection View* to analyze event correlations on an overview level. In Figure 5.3, a group of three stations was selected in the *Location Projection View*. The *Location View* shows that the stations are part of different process steps and the *Correlation View* indicates that the relationship is not caused by a single pair of event types, but that it is composed of multiple relations. Findings *F7* & *F9* (see Table 5.2) were both derived similarly to this example.

If the experts are interested in a specific group of error messages (e.g., errors that mention a specific sensor), they can use the text search component to highlight only errors or locations that contain the entered text. This is especially useful in combination with the *Correlation View*, as errors can be part of many correlation pairs and the filter helps to find these correlation pairs faster. Further, it enables users to broaden their exploration, for example, if they start with a very specific error, but then they want to find all errors that are similar to the error they found. The pattern shown in Figure 5.9a was discovered by starting with the insight from the example in the previous scenario (*Where are*

### 5.3 Visual Analysis for Spatio-Temporal Event Correlation in Production Lines



(a) Expert rating of the requirement questions presented in Section 5.3.1.



(b) Expert rating of the individual views and the overall system regarding their understandability (left) and usefulness (right).

**Figure 5.11:** Results of the qualitative evaluation with experts from the industry collaboration partner.

the stations that reported errors?), in which events were observed that occurred at the same time. Through the selection of these events, it becomes apparent that all descriptions are similar (*Fehler in Kinematik...*). Possible further steps are to use the text search to find similar events or to verify the similarity of the event classes in the *Event Projection View*.

#### 5.3.5 Feedback Session

For each finding (see Table 5.2), the experts were asked to state if the finding is plausible. In case it is, they were further asked to rate how expected the finding is, how useful it is to improve the manufacturing process, and how much effort is required to derive the insight with their current methods. As explained before, the second expert opted out of the insight rating. All of the findings were deemed plausible by the experts. None of the questions were rated with less than four points (borderline) and most questions scored at least five points on average. This means that all of the derived findings were deemed to be useful insights.

Before the system was introduced, the experts were asked to rate the requirement questions (see Figure 5.11a). All of the presented research questions got ratings higher than five, which indicates that the system meets the analysis requirements of the domain experts.

At last, the domain experts were asked to evaluate the system. They were asked to rate, how easy it is to understand the individual views, how much the views help to derive insight from the data, if the overall system would help them to gain insights, and if they can think of other areas in the company where such a system could be useful. The results of the questionnaire are shown in Figure 5.11b.

The *Timeline View* and the *Location View* with average scores of  $5.\bar{3}$  and  $6.\bar{3}$  respectively were better understood than the *Correlation View* and the *Projection View*, which scored 4 on average each. The score difference can be expected, as the *Timeline View* and the *Map View* are common visualizations that are easy to read, understand, and interpret. Further, they do not transform or aggregate the data. Generally, the scores for the first two views are similar, but expert 1 (head of the production line) gave much better scores for the *Correlation View* and the *Projection View* than the other experts (7 and 6 compared to  $2/3$  and  $3/3$ ). Compared to the understandability score, the experts rated the helpfulness of the views mostly equal or slightly higher than the respective views' understandability score.

When the experts were asked for oral feedback all of them gave similar explanations for their scoring: The *Timeline View* and the *Location View* are simple enough to be used by anyone who has a general understanding of the production line, including operators on the shop floor during their daily routine. This led them to give a high understandability and helpfulness score.

The other two views are more suitable for experts that are specialized in the analysis of the overall production line performance. The head of the production line is such a specialized user and gave the views very high scores (six or seven points). The other two experts explained that the more complex views may be powerful, but it requires training and time to get accustomed to these views. As the introduction to the views was very brief, they stated that they rated how well they currently understand the views. However, the experts explained that this score may improve, but this could not be answered without the before mentioned training and familiarization phase. As a consequence, they gave the views low helpfulness scores, as it was hard for them to estimate, how well insights can be derived from the views.

The remainder of this section presents additional feedback received during the presentation of the system and after the experts filled out the questionnaire. Generally, the approach was well received. In addition to its current functionality, the experts suggested to add more detailed data to the system, especially which products were processed by a station when it reports an error. This information is important for multiple reasons, for example, because different products may cause different error relations between stations.

During the presentation, the experts noticed a relation between two stations that are both part of processes that add components to the production line. The



**Figure 5.12:** The facet views (A)–(C) on the left provide a first overview of the event distributions. The temporal distribution of the filtered events and its outliers are presented as a stacked bar chart (D). Through brushing, analysts can get more information about the reported events during the selected timespan (E). The decomposition parameters for STL are set in the *Data Decomposition Control* (F) and results are shown as three line charts (G). The *Calendar Plot* (H) provides information about the distribution of outliers or event data. © 2018 IEEE

expert responsible for the production line explained that, although it cannot be shown with the currently available data, this may be a plausible finding if the supplier of the added components switches at this point in time. This finding led to the general consent of the domain experts that the incorporation of data from other departments would help especially during the reasoning step after building hypotheses using the presented approach. In addition, the experts suggested incorporating data about products that were taken out of the production line due to issues and link this information to the machine error reports.

## 5.4 Visual Analysis of Temporal Event Patterns through Event Series Decomposition

As presented in Section 5.3.1, one possibility to reduce the downtime of production lines is to understand, which events are caused by other previously reported events. This knowledge can be used to reduce the occurrence of the original cause of errors.

Another possibility to prevent errors is to know, whether they are recurring over an extended period of time. It is important to understand if any events have an underlying trend or if their time of reports follows a pattern.

The goal of the following approach is to support production experts in finding and comprehending long-term issues of production lines. To find recurring event patterns, the reported event logs are iteratively decomposed using *Seasonal Trend decomposition using locally weighted regression (loess)* (STL) (see Section 2.1.3). The system presents original and decomposed event series as time series plots and in a calendar view (see Figure 5.12).

The following first presents design requirements for such an approach, which were obtained in interviews with domain experts (Section 5.4.1). Then, the approach and its interaction support using a evolutionary algorithm is presented (Section 5.4.2 & Section 5.4.3). At last, the approach was evaluated through use cases and expert feedback 5.4.4.

### 5.4.1 Requirements

Several meetings with domain experts of the industry partner revealed what insights they hope to find in their data regarding recurring events. The following approach aims to extend the currently used short-term analysis by supporting the extraction of problems that occur regularly over an extended time period to find previously unnoticed event patterns. Such an approach needs to meet the following requirements:

**Requirement 1: Overview & Faceted Information**  $(R_1)$

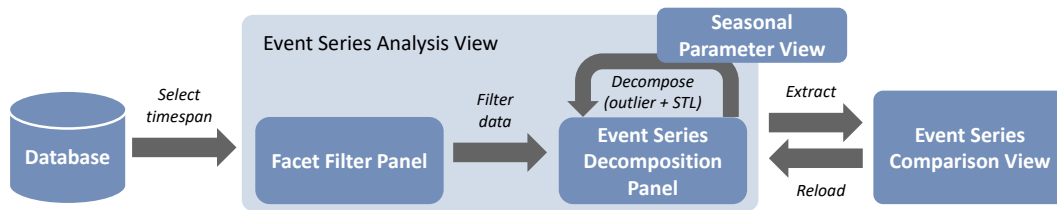
Present an overview of the available events with respect to the process step, event type, and product type. Further, provide interactive data filters and visualize the temporal distribution of events.

**Requirement 2: Pattern & Outlier Identification**  $(R_2)$

Help analysts to semi-automatically find seasonal patterns, trends, and outliers of the reported events.

**Requirement 3: Extract & Compare**  $(R_3)$

Facilitate analysts to extract analysis results, visually compare them, and interactively extend them.



**Figure 5.13:** At first, the users need to select, which timespan to analyze and what event type should be used. The data used in each analysis is set through facet views, which also provide an overview of the various data aspects through sparkline visualizations. After filtering the data, the users can inspect the aggregated event series in a line plot, which can be decomposed into a trend, seasonal and remainder component. In this view, analysts can iteratively extract outliers, trends, and seasonal patterns. ©2018 IEEE

## 5.4.2 Approach

Analogous to the requirements, the concept comprises three parts: first, analysts get an overview of the data and filter it according to their needs with facet views. Second, they analyze the filtered data regarding temporal aspects. Third, analysts can extract and compare findings (see Figure 5.13).

### Data Subset Configuration through Faceted Search

Initially, the analysts choose a time period. The following analysis is conducted in the *Detailed Analysis View* (Figure 5.12). The *Facet Filter Panel* (A)-(C) on the left uses faceted browsing [207] to filter the events based on their process step, product type, and event type. All facets have a common color scheme for event counts, which ranges from a light yellow via orange and red to black. In most cases, the facets also provide an overview of the distribution of the events along the production line’s process steps. The distribution is visualized as a sparkline visualization [184] that presents the production line’s process steps as bars, with details available via tooltips. The height of a bar encodes the number of events of the step normalized per row.

**Process steps.** The *Process Step Facet* (see Figure 5.12 (A)) lists all available steps in their order of occurrence in the production line. Each item provides the ID of the step, its description, and its event count. In addition, it shows a visualization that represents the relative share of events compared to the total number of events, similar to a Pareto chart [179], which is commonly used in the quality management domain. The event share is visualized as a bar, where the width of the bar represents the step’s share relative to the total event count. Further, a line indicates the cumulative event share of the current and all previous process steps. The experts of the industry partner explained that such a piece of information is an important aspect for prioritizing analysis tasks.

**Event types across process steps.** The data can also be filtered based on specific event types (see Figure 5.12 (B)). Each row corresponds to one type and comprises its ID, description, occurrence count (which is also the sorting criterion of the list), and the event distribution sparkline. This provides analysts with a quick overview of the event distribution along the production line.

**Product types across process steps.** Analysts can filter the data with respect to the products that were being produced when events occurred (see Figure 5.12 (C)). The *product type filter facet* is similar to the facet explained above, but instead of using event descriptions, the facet provides information about the distribution of the events depending on the produced good. Each product is represented through a unique product number. The experts from the industry partner explained that this number is readable by analysts who are familiar with the production line. This way, it is possible to quickly find similar event distributions of different products.

In case entries from multiple facets are selected, the data needs to meet at least one selection from each facet. Except for the temporal aspect, the faceted views meet requirement *R1 (Overview & faceted information)*.

### Temporal Analysis using Event Series Decomposition

The analysis to find temporal event patterns is conducted in the *Event Series Decomposition Panel* (see Figure 5.12 (D)). For the temporal analysis, the filtered data are aggregated by the hour in which they occurred. Initially, this panel consists of four plots.

All series in the *Event Series Decomposition Panel* have a common x-axis that represents the loaded time frame. The y-axes represent the number of events at a point in time. They adapt to the data shown in the plot and not to the global maximum to use as much space of the plot as possible.

The *event series plot*, shown at the top of the panel, is always available and shows the filtered data, as well as (optionally) outliers, in a stacked bar chart. At last, the panel provides three (initially empty) line plots that, once the analysts choose to decompose the series, provide the trend (green), seasonal (purple), and remainder (brown) components of the series. To improve the comparability between the seasonal and trend series, their y-axes have a shared min-max range. The trend, seasonal, and remainder series are described in more detail later in this section. The *event series plot* meets the temporal aspect of requirement *R1 (Overview & faceted information)*.

**Outlier configuration and extraction.** Analysts can optionally inspect and extract outliers from the event series plot. Outliers may indicate unexpected events that should be further investigated. In addition, the extraction of outliers before the event series' decomposition improves the results, as STL may miss some outliers and include them in the seasonal component.

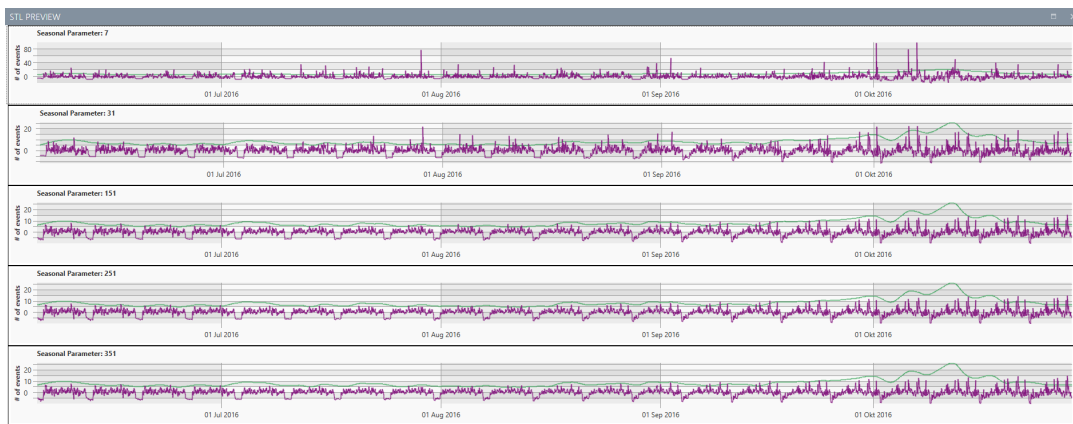


In case the outlier detection is enabled, parts of the *event series plot* are replaced with red bars proportional to the outlier part of the series. To decide, if and to what extent a point is an outlier, the standard score of the data point is used, which describes how many standard deviations a data point deviates from the mean value. A data point is identified as an outlier if its standard score is higher than  $x$ , which is four by default. However, the outlier threshold can be changed through a slider control, because it depends on the dataset and the analysts' notion of what an outlier is.

Removing the entire outlier would likely cause another outlier because no events during one hour are unlikely (except nothing is produced). Therefore, a compensated value for every outlier is calculated. First, the value between the previous and next data point is interpolated linearly and then  $x$  standard deviations are added to or subtracted from the interpolated value to move it towards the data point's actual value.

**Iterative Analysis of Trends and Recurring Behaviors.** After filtering the data and optionally extracting outliers, analysts can decompose the *event series plot* using STL, which decomposes an event series into a trend, seasonal, and remainder component (see Section 2.1.3). The trend represents long-time effects in a time series. The seasonal component represents the recurring effects during the series. These are of the most interest, because recurring events may indicate systematic problems in the production line. The visualization of the trend and seasonal component contribute towards requirement *R2 (Pattern & outlier identification)*. The remainder component represents the difference between the initial event series and the extracted trend and seasonal components. Thus, it consists of noise that is present in the data and non-extracted outliers. Further, it may contain non-extracted seasons that have a shorter length than the current season (longer seasons are extracted partly into the trend component).

Section 2.1.3 provides a brief introduction to STL, whereas the work by Cleveland et al. [48] provides further technical details. When STL is applied to the error series, the resulting trend, seasonal, and remainder components are shown in their respective plots in the *Event Series Decomposition Panel* (see Figure 5.12 (D) & (G)). Analysts need to provide three arguments to run STL: the time series represents the dataset used for the decomposition. In most cases, this is the entire series shown in the event series plot, but the users may manually select parts of the series (see *Inspection and Filtering of User-Selected Data* below). The seasonal period length defines the number of data points per season (e.g., 24 to analyze daily seasons). The strength of the seasonal smoother defines how strongly the seasonal component should be smoothed. A high value will lead to seasons with very low variations over time, while a low value allows high variation, which may also include noise. The strength of the smoother cannot be predetermined, as the analysts need to decide how much variation is allowed in the seasonal component. As the analysts are usually neither data

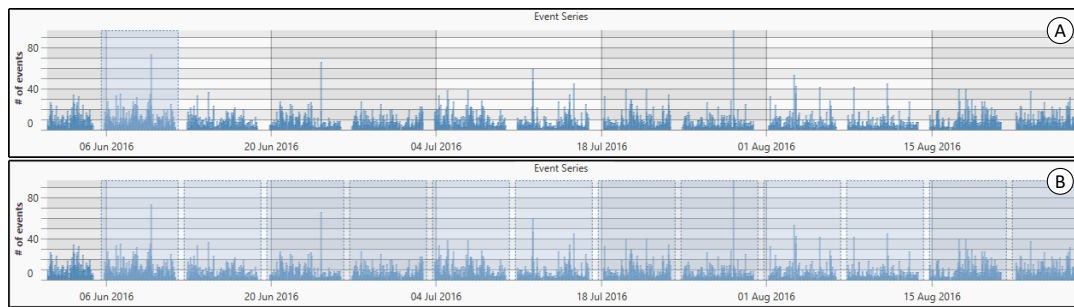


**Figure 5.14:** Exemplary depiction of a choice between various seasonal parameters used by STL. The first two options (with values of 7 and 51) still contain much noise, while the remaining three choices (with values of 151, 251, and 351) do not differ noticeably. © 2018 IEEE

scientists nor statistics experts, the other parameters, such as the window sizes used during the trend and seasonal series extraction, are not exposed. Instead, they are approximated automatically (see Cleveland et al. [48]).

A predefined set of seasonal period lengths is provided to detect daily, weekly, or monthly patterns, but the period length can also be set manually (see Section 5.4.3 for details). Once the analysts decided on the event series and the period length, they are provided with a preview of the seasonal and trend component for different values of the smoothing parameter. Figure 5.14 depicts an exemplary choice for several seasonal parameters. The first and partially the second choice still contain much noise, and the remaining options show similar results, so experts may choose the third parameter. The proposed values were determined empirically and deemed appropriate for the decomposition of event series in a production line. In case the decomposition yields a season or trend that the analysts deem interesting and plausible, the current analysis state can be stored in the *Event Series Comparison View* (see Section 5.4.2–*Event Series Comparison*). This contributes to requirement *R2 (Pattern & outlier identification)*.

**Calendar Plot.** A calendar below the event series plots (see Figure 5.12 (H)) provides more temporal context, as the production schedule usually repeats every week. The analysts can switch the data source through a combo box. If the data source is the event series, the analysts can further choose to show the regular data or the outliers. The same color scheme as in the *Facet Panel* is used (which ranges from a light yellow via orange and red to black). Initially, the calendar is grouped monthly and every entry represents one day, much like the calendar plot used by van Wijk and van Selow [189]. The analysts can change the time granularity of the calendar so that the calendar’s cells can represent hours, days,



**Figure 5.15:** The users can define their custom event series by first selecting a start and end point of the first season occurrence (A). Then, they define a gap (which can be zero) between the seasons' occurrences. To get a feedback of the selected pattern, all seasons that fit in the event series are also highlighted (B). © 2018 IEEE

months, etc. If the calendar cells have a coarser granularity than the plot views (days, months, years, etc.), the average number of outliers or events per hour are calculated for the color mapping.

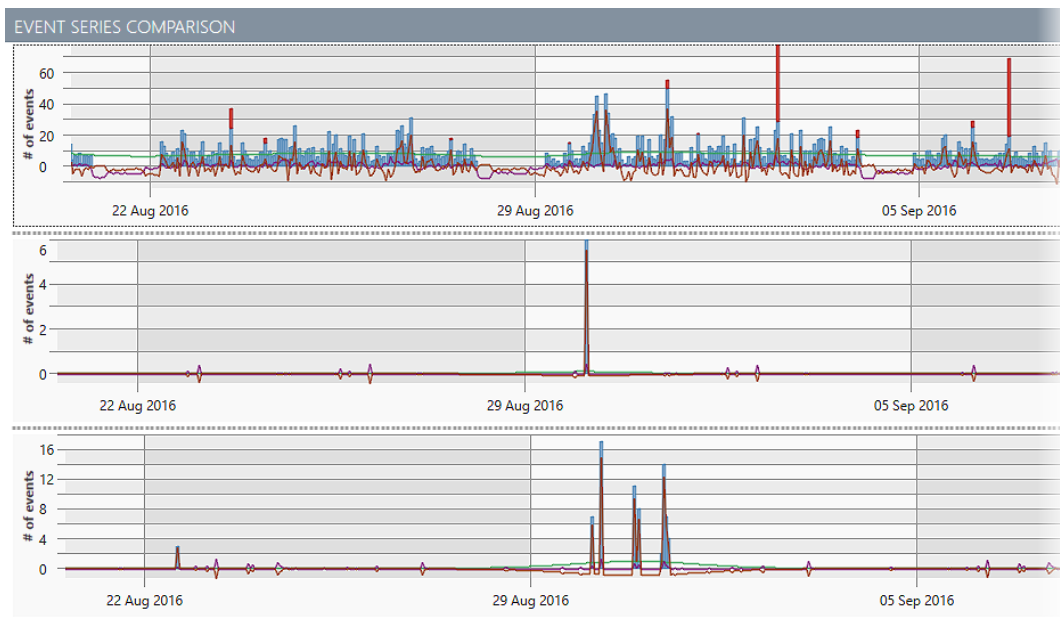
### Inspection and Filtering of User-Selected Data

In addition to the explicit filter facet, the analysts can retrieve information about event types by directly interacting with the event series plot. They can either inspect reported events at a certain time (e.g., one specific hour) or during a time interval (from 8 am to 8 pm). A specific point in time can be selected by right-clicking. An extended time frame is selected by brushing over the interesting part of the *event series plot*. The selection is highlighted with a light blue background (see Figure 5.15 (A)). Then, the selected event reports can be retrieved by right-clicking the highlighted area.

The retrieved event types are listed to the right of the event series (see Figure 5.12 (E)). The listed selection of the event types is coupled with the event type filter facet, so selecting an event type in one filter will also select it in the other. The context-dependent event type list provides a secondary filter option and therefore supports requirement *R1 (Overview & faceted information)*.

### Event Series Comparison

Analysts can store insights gained during the analysis process by storing the current filter settings and decomposition results in the *Event Series Comparison View* (see Figure 5.16). The series are visualized as a superimposed representation. To perceive the original data as well as its decomposed series, the stacked bar chart containing the event series and outliers is overlaid by the line charts of the decomposed event series. The analysis configuration can also be restored to the *Detailed Analysis View* to continue the analysis. Figure 5.16 presents an



**Figure 5.16:** The event series comparison view visualizes event series and (if extracted) their outliers as a stacked bar chart, which is overdrawn by line charts of the decomposed event series with a color coding analogous to the *Event Series Decomposition Panel*. This example shows the original event series as well as two subseries of different event types. © 2018 IEEE

exemplary comparison that contains the original data series, as well as two series filtered by two different event types. The outlier events occur at similar points in time, which may indicate that the events are related.

### 5.4.3 Selection and Optimization of Parameters

If the predefined STL seasonal periods are insufficient, analysts can define a custom pattern. First, they need to select the pattern's first occurrence analogous to Section 5.4.2—*Inspection and Filtering of User-Selected Data*. Further, analysts can indicate a timespan to skip during the decomposition by dragging a replica from the highlighted pattern to the next occurrence. The user-defined event series are highlighted across the *event series plot* to make the resulting series more comprehensible (see Figure 5.15 (B)).

If the event series comprises a high number of data points, analysts may have problems to set the exact values they desire. Therefore, an optimization approach based on a differential evolution algorithm (see Section 2.1.3) automatically improves the users' input parameters. It comprises the following six steps, where step (0) is executed just once and steps (1)–(5) repeat  $n$  times:

**0. Initialization.** An initial set of 30 parameter configurations is created.

The parameter values are randomly initialized with a uniform distribution

around the users' input values. The starting point can vary by up to  $\pm 24$  hours, while the period and gap length can vary by up to  $\pm 48$  hours.

- 1. Cost Evaluation.** To evaluate each parameter configuration, the event series is first decomposed based on the configuration's parameters and then its seasonal component ( $cs$ ) is extracted. To do so,  $cs$  is split into its seasonal subseries  $cs_i$ . Then, an average subseries  $cs_{avg}$  is built from the selected and the following two subseries, assuming that the searched pattern will be most visible close to the first pattern occurrence. Afterwards, the average variance ( $cs_{var}$ ) between the first three subseries  $cs_i$  and  $cs_{avg}$  is calculated. The cost value  $C$  was defined as:

$$C(cs) = \frac{cs_{var}}{\underbrace{(\max(cs) - \min(cs))^2}_{\text{Range of the series' values}}}$$

It should be noted that a lower cost value is better.

- 2. Keep Best Configurations (Elitism).** To save the best results, the top 10% of the configurations are transferred to the next generation.
- 3. Discard Unfit Configurations.** The lowest scoring 50% of the configurations are removed.
- 4. Recombination.** New parameter configurations are created by combining two still existing "parent" parameter configurations  $A$  and  $B$ . The new configurations' values  $v$  are based on a weighted average between the parents' values  $v_A$  and  $v_B$ . The parameter  $\alpha$  is randomly picked based on a uniform distribution:

$$v = \alpha \cdot v_A + (1 - \alpha) \cdot v_B, \text{ where } \alpha \in [0, 1]$$

- 5. Mutation.** At last, the new configurations' parameter values have a slight change ( $p = 0.1$ ) to be randomly changed within the variation explained in the initialization step. Values that are out of bounds (e.g., a start date earlier than the first data point) are assigned the outmost allowed value.

The user-defined selection is updated whenever a better result was found. Due to step (5), which broadens the pool of available parameter configurations, an evolutionary algorithm is more robust regarding local optima than simpler optimization approaches (such as hill climbing).

## Results

The optimization approach was not part of the pair analytics sessions presented in Section 5.4.4, as the domain experts had no experience with the presented approach and the evaluation of the general approach was prioritized during the

feedback session with the domain experts. In the following, an exemplary result achieved with the input optimization approach is presented and discussed.

As a proof of concept, it was tested, if the evolutionary algorithm is able to recommend a simple pattern if the input parameters are similar to a presumably good result. The unfiltered event series provides only one trivial insight, but it is clearly visible that the seasonal component reflects the weekends during which usually no events were reported. Figure 5.12 ④ & ⑤ show what the decomposition should look like. It was hypothesized, that the evolutionary algorithm recommends that the season and the season gap should add up to seven days and that the period length should be between five and seven days.

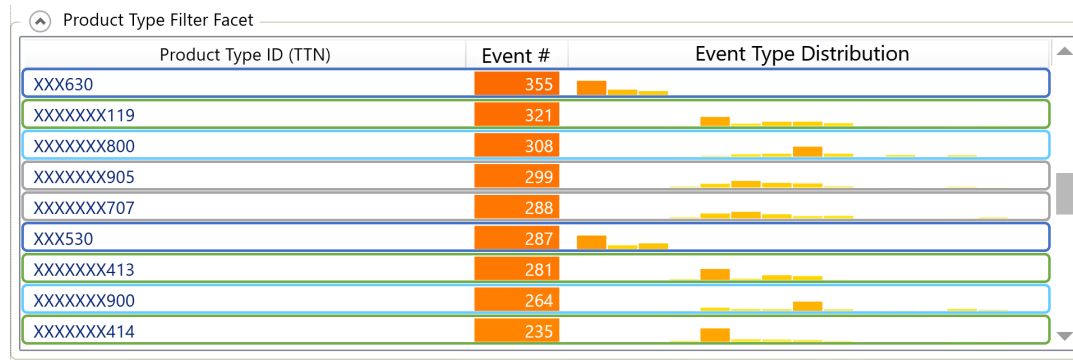
Multiple test runs were conducted and manually set the period length to be six days and the gap between the seasons to one day. This setup covers the assumed results, but also leaves space for other results outside of the assumed optimum (e.g., four days without a gap). The results show that a seven days period without a gap achieves the best result ( $C(cs) = 6.66 \cdot 10^{-5}$ ), whereas a period length of five or eight days without a gap resulted in the worst results ( $C(cs) = [1.71 \cdot 10^{-4}, 1.74 \cdot 10^{-4}]$ ). Further, a seasonal length of six days with one day gap resulted in suboptimal results ( $C(cs) = 1.18 \cdot 10^{-4}$ ). A possible explanation could be that the variation on Sundays is still a better trade-off than ignoring Sundays and only considering the variation of the remaining six days.

Overall, the optimization approach showed promising results, but there is room for improvement regarding the handling of the data during the gaps between the seasonal patterns.

#### 5.4.4 Evaluation

The approach was evaluated in two pair analytics sessions [17] with experts from the industry partner. In pair analytics, one or more domain experts work together with a visual analytics expert. The domain experts contribute their expertise and experience to focus the analysis on interesting data, while the visual analytics expert is operating the visual analysis tool.

The pair analytics sessions were prepared by searching for potentially interesting patterns, which were used as a starting point during the sessions. In the following, three use cases that were derived during the pair analytics sessions are showcased to demonstrate how the presented approach contributes to the analysis of event patterns in a production line. All of the use cases are based on approximately six months of event data. The evaluation was conducted with two domain experts in each session: the first expert was responsible for the data acquisition and to make the data available to workers on the shop floor of the factory. He was also involved in the development process of the approach. The second expert was the head of the studied production line and was consulted



**Figure 5.17:** The distribution of the events per product type exhibit several distribution patterns (highlighted in different colors). The domain experts already assumed such profiles. Further, they noticed that these profiles do not always correlate to the different groups of produced goods (e.g., electric motors with and without a power train extension). © 2018 IEEE

only for the evaluation of the approach. Each of the pair analytics sessions lasted approximately 60 minutes. Furthermore, general feedback is presented that was collected from the domain experts after the pair analytics sessions.

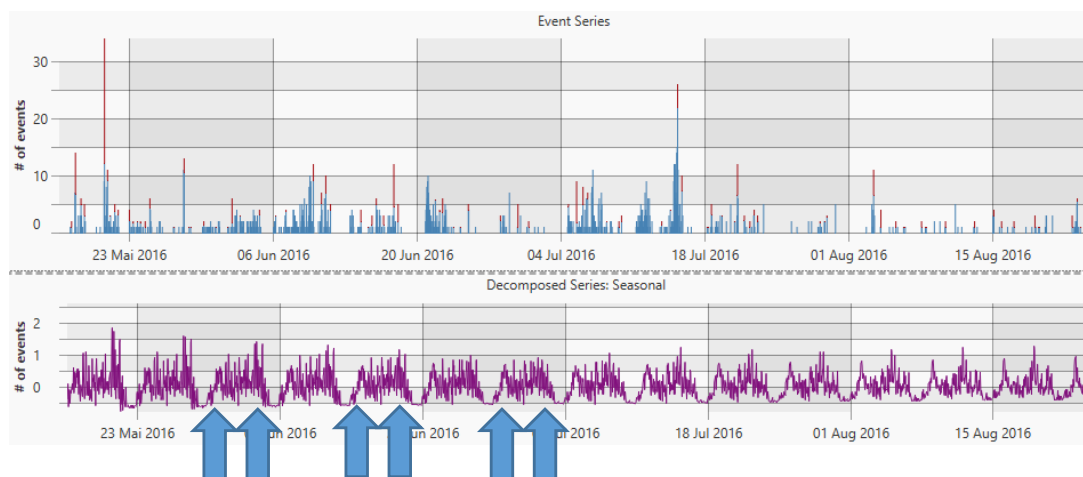
### Use Case 1: Event distribution along a production line

Before the first detailed analysis was started, one expert mentioned that the event distribution shown in the *event type facet* differs from what he expected. There are, generally speaking, four different types of electric motors produced in the production line. He expected that the four types would have different event distributions when compared to each other and that variations of the motor types would have similar events and distributions. However, the data only partly supports these assumptions. There are similarities in the event distributions, but the motor type is not always the same. Figure 5.17 shows an excerpt from the product type facet, wherein the different event distributions are highlighted. The similar distributions are partly explainable because even if the motor types are different, they still share parts of their production plan. However, some of the similarities were not explainable this way and further investigations in cooperation with workers on the shop floor level are required to assess other reasons.

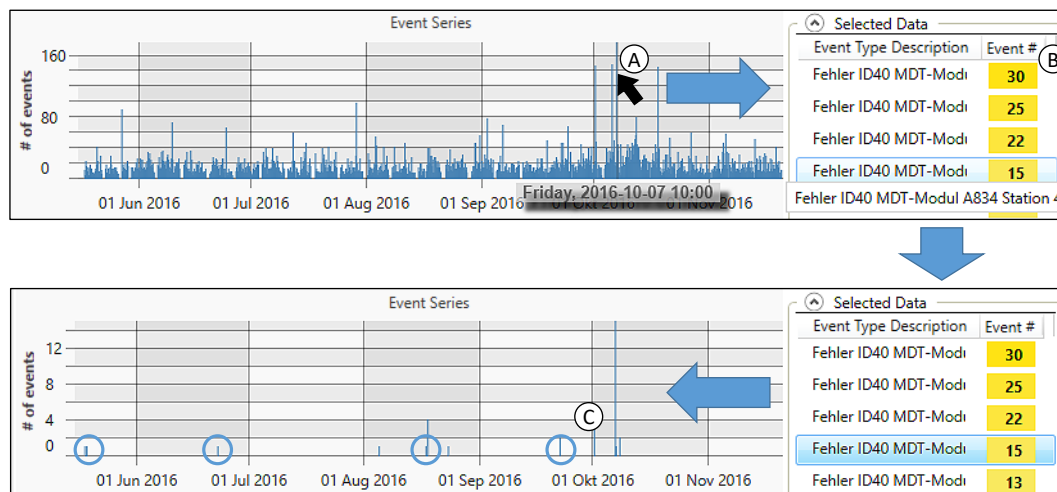
### Use Case 2: Analysis of Recurring Event Patterns

Next, the second domain expert was interested in the most frequently occurring event code. This event was at the *label checker station* that has to take the part off the workpiece carrier to process it. The series was decomposed with a period length of one week to check for event patterns, as the line's production schedule

## 5 Visual Event Analysis in Production Lines



(a) The data shown in the event series plot (top) is decomposed to search for weekly seasonal patterns. There are no events on Sundays (flat line before the arrow pairs) and the number of events is increasing at the beginning and end of the week. The arrows indicate at what times the overall number of events increases.



(b) The inspection of one of the event series' outliers reveals that the same component caused issues on various stations along the production line within an hour. When filtering for individual events (highlighted entry in the tooltip), it becomes apparent that the events repeat approximately every month (highlighted through circles).

**Figure 5.18:** Results of use case 2 (a) and use case 3 (b). © 2018 IEEE

usually repeats every week. Figure 5.18a shows the event series and the resulting seasonal plot. The seasonal pattern shows two characteristics: first, the pattern starts with a short period of time, where no events occurred. This is expectable, as the line usually does not produce anything during Sundays and therefore there cannot be any events during that time. However, the pattern disappears towards the end of the series, as the production line runs more often on Sundays.



Second, the number of reported events is remarkably higher towards the beginning and the end of the week compared to other days. The head of the production line hypothesized that this finding may be caused by the quality of the parts used during the production. He explained that the line usually uses parts from the main supplier, but towards the end of the week, these parts often run out. In that case, they switch to parts from another supplier, whose parts have a higher quality variance than the main supplier's. Although the parts' quality is mostly within the allowed margin of error, the stations that process these parts have a higher likelihood to encounter problems when processing these parts. This hypothesis was supported when the expert accessed the logistic department's inbound delivery list, which the presented tool cannot access. He further mentioned that they were aware of this issue before. However, they could not prove that this is a regularly recurring issue because their previous analysis methods did not allow them to transform the data as needed to support this hypothesis. Therefore, this finding is helpful because it can be used to argue for an improvement of the robustness of stations that need to process the mentioned supplied parts.

### Use Case 3: Analysis of Outliers for Pattern Analysis

The event series of the *Event Series Decomposition Panel* shows a remarkable outlier in the second half of the data (see Figure 5.18b (A)). The events at that point in time were extracted by selecting the peak. Almost all events were related to the *ID-40 module*, which is a sensor that reads the ID of the workpiece holders to provide tracking within the production line (Figure 5.18b (B)).

Usually this event indicates a broken sensor, but in this case, the event was reported from various stations at the same time. One expert explained that such an event distribution may indicate a problem with the bus system that connects the sensors to the IT infrastructure. Afterwards, some of the events were filtered one after another and the experts noted that the events repeat approximately every month (see Figure 5.18b (C)).

The head of the production line stated that this is an interesting and unexpected finding that they were not aware of before. Due to the found result, two measures were taken: First, the operators will be informed to watch such sensor events more closely. Second, the finding is forwarded to the responsible department, as it is not possible to fix it by improving the production line, but other lines in the factory may be affected by this problem as well. Some time after the analysis, the industry partner found out that parts of the identified issues were caused by an error in one of the machine tool's programs.

### General Expert Feedback

After the pair analytics sessions, the participating experts were invited to give feedback regarding the general approach and the different components in a semi-structured interview. They were asked about advantages and drawbacks of the currently available views and if they could imagine any enhancements that would help them with their work. They first remarked that the approach results in a powerful tool that should be used by an engineer in a supervising position, as shop floor operators have neither time nor necessary expertise to analyze such event patterns. They also mentioned that this is not a problem, as experts such as production line heads can use the approach and forward the gained insights to the operators at the show floor level.

The experts summarized that the *Event Series Comparison View* is especially useful to see what data was already analyzed in the past. They further explained that the iterative analysis approach on the overview and detailed analysis level is helpful. It enables experts to either pursue a specific event until its occurrence is completely understood (varying STL configurations) or to analyze the most important events separately (varying filter configurations). The overview of the analyses provided in both views is important because the analysis runs are often interrupted, e.g., because talking to specialists is required to solve an issue. The experts further inquired that the *Event Series Comparison View* should be extended to contain information on how the shown series were extracted, for example by showing the used filters.

The experts also found the *Facet Panel* useful. They mentioned that the product type facet is helpful to gain more insight about the events' distributions. The process step and event type facets are useful, but unlike the product part facet, they cannot be used for a free exploration of the data. Instead, the analysts must already have a specific analysis goal that requires a selective investigation.

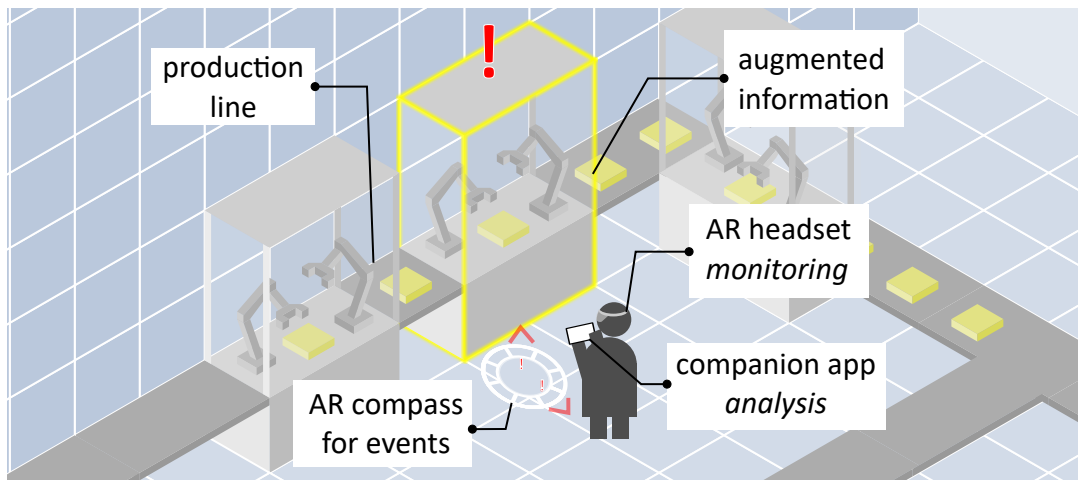
The experts stated that the extraction of outliers and seasonal trends are both of importance. An outlier can indicate special events that may require special attention. The seasonal series extraction allows to form hypotheses about systematic errors that are backed by the available data. Furthermore, the experts emphasized that engineers are usually not experts in statistical analysis and thus require an easy to use approach to decompose the temporal event series. They understood that the approach exposes only the required parameters towards the user. However, the results are still sometimes difficult to interpret, making more indications that help them to configure the analysis necessary. One of the experts proposed that the system could suggest decomposition configurations automatically. An expert would still have to evaluate, which suggestions are useful, without the need of any additional knowledge. At last, the experts rated the *Calendar Plot* to be useful, because it gives a different view on the time series and also provides information about other time granularities.

## Integration of Augmented Reality Monitoring and Visual Event Analysis

The augmented reality extension for the simulator for advanced manufacturing (ARSAM) presented in Section 4.3 and the corresponding related work (see Section 4.1) show that using AR to support on-site analyses has high potential to support domain experts. The following chapter uses the lessons learned in the previous chapters and proposes a holistic system that combines the advantages of situated analysis [133, p. 185 ff.] using augmented reality devices with the capabilities of classical desktop applications.

### 6.1 Motivation

In the past years, the tasks of operators on the shop floor shifted away from manual tasks, such as welding, towards monitoring of individual, automated process steps as well as the overall manufacturing process, only taking action in case of malfunctions or unexpected events. Through discussions with the experts of the industry partner, two important aspects for making use of the data provided by an automated production line were identified. First, live monitoring of sensor data is necessary in order to timely react to malfunctions of machinery and to remedy critical issues, which in turn is crucial to keep a production line as effective as possible. Real-time analysis is a demanding task that often requires efficient data filtering and meaningful abstraction, the latter of which can often be best provided to a human user by a visual representation. Second, a retrospective analysis of previously collected data and logged events helps to understand and improve current processes. Currently, the analysis of events and the propagation of recent observations is often limited to daily briefings, which are partly based on data, such as deviations from planned processing times or unusually high numbers of errors at specific machines. It is desirable to be able to update this analysis with the latest data continuously and, more



**Figure 6.1:** Illustration of the approach for the monitoring and analysis of event reports in production lines. Operators are notified of current issues via an augmented reality headset and can get details and an overview of the situation on a tablet.

importantly, to provide the results from such an analysis in real-time and on-site during production. This enables operators not only to react to emerging problems immediately but also to apply the accumulated knowledge of past events.

The following introduces an immersive analytics system that covers both a real-time situated approach for presenting data to an operator and the in-depth analysis of historical event data on a tablet device (see Figure 6.4). Immersive analytics [133] recently gained attention due to the availability of advanced and more affordable portable and wearable devices, such as head-mounted displays for virtual and augmented reality. These are crucial tools for developing effective and more compelling systems for immersively exploring large and complex data. The presented system contributes an immersive analytics approach for monitoring and analyzing automated manufacturing processes and combines the unique benefits of an AR headset with the larger screen and well-known input modalities of a tablet. It is designed as a linked-view system consisting of an analytic component (running on the tablet device) for advanced data analysis and a monitoring component (running on the augmented reality headset) for a situated inspection of identified issues. The approach was evaluated by deploying a prototype implementation on-site at a production line of the collaboration partner introduced before (see Section 5.1) and by gathering feedback from their production line operators.

## 6.2 Domain Problem Characterization

For this project, a production line for small electrical motors for cars was chosen. Such a motor consists of several parts, fabricated in different parts of the line. As

the production line itself contains more than sixty single machines, the interplay of all machines together is quite complex. Operators on the shop floor must manage to keep the production line running. They get information directly from display above the production line, a so-called Andon board [67]. On this display, the current Overall Equipment Effectiveness (OEE), which represents the current performance of the production line (see Section 2.2.3), and error events (see Section 5.1.1) are shown. Further, indicator lights above the machines indicate their status (e.g., missing parts or a problem with the machine). The daily challenge for the operators is to identify the importance of the error event and to know (mostly from their experience) where the machine is located on the shop floor. These disturbances cause downtime and have a direct impact on the OEE. Therefore, the main question for the operators is: *Which machine should I fix first to maintain the OEE and keep the line running?*

Currently, the Andon boards above the lines only show error codes but not the locations of the erroneous machine. Further, due to the layout of a complex production line, operators usually cannot see all machinery or their indicator lights from where they are. When they notice an issue, operators prioritize and complete tasks based on their experience or subjective judgment of urgency of specific failure events. The lack of objective guidance for prioritization of failure events is an important issue in current production lines. Thus, a solution supporting operators with their decisions and helping them to prioritize and locate machines is a valuable addition to existing indicators.

To better understand the domain problem, to establish relevant task classifications, and to identify requirements for the system's application, several meetings with the collaboration partners were held. During these meetings, the overall system design was discussed, iteratively improving upon the initial concept by gathering qualitative feedback. Meetings also frequently included preliminary test runs of the prototype implementation at the studied production line. Subsequently, four main categories for classifying the tasks were identified that have to be covered by a system to provide immersive monitoring and analytical capabilities for a manufacturing environment. Based on discussions with domain experts and relevant related work, these four categories are *monitoring*, *analysis*, *prediction*, and *maintenance*.

**Monitoring** The operators have to react to error events as soon as possible. These events cause the production line to halt and must be addressed immediately. As explained before, events (e.g., error messages from machines) are shown on a central display above the assembly line. Presenting such information on a wearable device, such as a head-mounted AR display, has several advantages: (1) The operators are notified within their field of view, independent from an external display. (2) Additional information can be displayed that would otherwise only be

visible at the respective machine. (3) Responsibilities for events can be centrally distributed by providing specific operators with customized information based on event priorities.

**Analysis** The retrospective analysis of the previous events plays an important role. Currently, the collaboration partner's experts discuss such events daily. An analytical approach should support (1) a situated analysis of previous events with longer time spans on demand, (2) visual analysis to identify correlations between events and whether a causality between them exists. A detailed analysis is less suited to interact with in AR due to the amount of presented data and necessary interaction with the data. Therefore, the analysis is externalized onto an additional application running on a mobile device that links selected data with the augmented view.

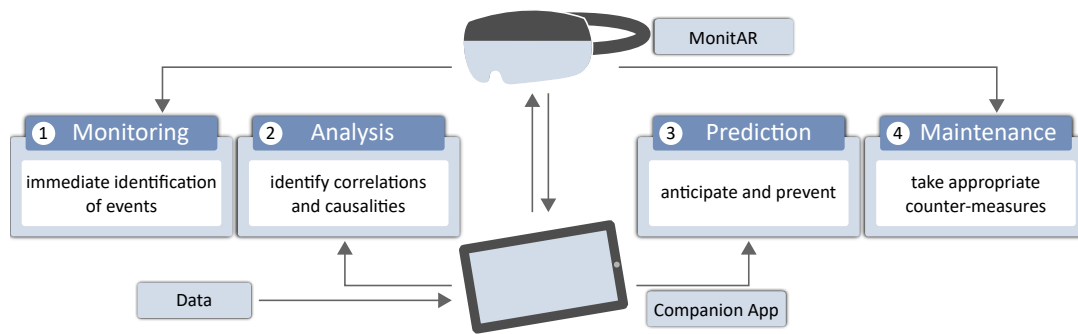
**Prediction** Anticipating events is tightly coupled with analysis. In the long run, the prediction of events should be automatic, but this goal requires prior knowledge about causalities to model an appropriate predictor. Hence, analysis and prediction should be conducted with a visual analysis approach and event-related predictions should be presented in AR in order to prevent possible issues.

**Maintenance** In case of an error event, the operator has to react appropriately. For example, in case of a machine failure, a repair procedure has to be initiated. An often suggested method is providing information supporting the repair process in an AR view. In most cases, operators need both hands for repair, so a full-size computer or tablet is impractical for information display and interaction.

The following approach focuses on the first two categories and how *predictions* could be integrated in future iterations. *Maintenance* is not extensively addressed, as this was the focus of numerous previous works and became (to some degree) a reference utilization of AR systems (see Section 4.1.3).

Furthermore, the domain experts stated some requirements particularly important for monitoring and analysis procedures for production lines:

**Effortless access and hands-free usage** The most important information needs to be provided concisely and in a way that does not require the operator to interact with the system to access the information. Specific tasks, mainly concerning the maintenance of defect machinery, require that the operator can use both hands. Hence, the system has to support interruptions of ongoing tasks and provide information hands-free.



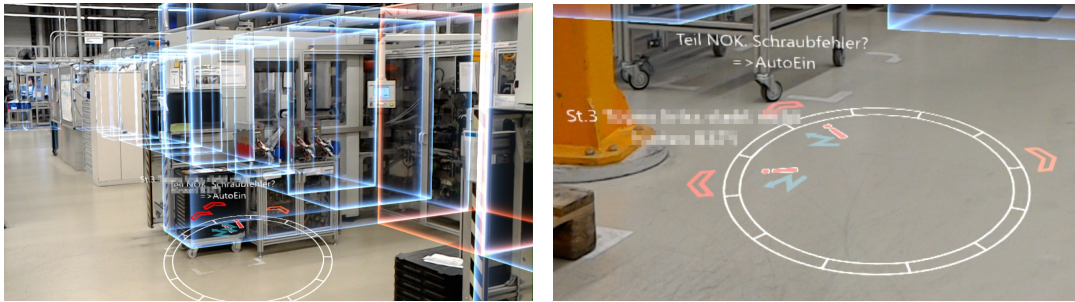
**Figure 6.2:** Immersive analytics scenario for manufacturing. Four main application categories for this approach were identified: (1) monitoring for immediate reactions, (2) analysis of recent events for correlation analysis, (3) prediction based on knowledge from the analysis, and (4) maintenance support. Augmented reality is mainly applied for monitoring and maintenance and a companion app on a mobile device is suitable for advanced analysis and prediction. Linking between the AR device and the companion app supports all scenarios.

**Analysis limitation to most relevant tasks** As their time is limited, the data provided to the operators must be limited to aspects relevant to their daily tasks. In the given scenario, the analysis needs to focus on how to support the decision, which malfunction to fix next, where this error is located, and to retrieve further details about the error.

## 6.3 Approach

The related work in Section 4.1.3 indicates that a situated analysis of issues in a production line may help to get a better understanding of events and improve the response time when they arise. Hence, supporting real-time visual monitoring is essential for an immersive analytics system. Furthermore, in scenarios exceeding the scope of simple reaction to events, an analytical component is required for providing the necessary context of recent and historical events to the on-site operators. This and the wide range of tasks and constraints listed in Section 6.2 suggest that no single device can appropriately cover all aspects of the desired system. On the one hand, a traditional device with a sufficiently large screen for presenting and browsing detailed information, such as a tablet computer, does not meet the requirements of hands-free usage. On the other hand, smaller, wearable devices like smartwatches or head-mounted displays can only provide a limited amount of information at a time and have a limited range of input interactions.

Thus, the purely informative and the analytical component of the presented system are separated and distributed among two physical devices combining their respective benefits. The prototype of the presented approach uses an



**Figure 6.3:** Left: The compass showing indicators for two active events and an additional third one for the currently selected process step. Right: Information about individual stations is always visible, even with real-world occlusions on the machinery.

AR headset, specifically the Microsoft HoloLens, as the hands-free wearable device. It presents the user with only the most important information that should be immediately available while making use of the spatial context to show information at fitting real-world locations. For more detailed information and in-depth analysis of the situation, the HoloLens application is complemented by a *companion application* (henceforth *companion app*) running on a tablet. When needed, this device offers additional information and provides sufficient space for complex visualizations and sophisticated interactions that are cumbersome or impossible on the HoloLens. If not needed, the companion device can simply be stowed away to free the users' hands. Both devices are tightly coupled and exchange information to provide the user with a comprehensive system that is adaptable to usage circumstances.

Figure 6.2 depicts the proposed design concept for immersive analytics of production lines. It covers all four task categories presented in Section 6.2: *monitoring*, *analysis*, *prediction*, and *maintenance*. Effortless and hands-free requirements are met by the augmented reality device, which is mainly used for monitoring and maintenance. The companion app handles incoming data and covers the requirements for analysis and prediction. By design, both devices share events through linked views that synchronize information according to the current task. The presented system focuses on the design of appropriate visualizations that cover monitoring and analysis tasks. Possibilities on how to include prediction and maintenance tasks are discussed in Section 7.3.

### Visual Monitoring

The main goal of visual monitoring is to provide users with specific information about important events. Therefore, event notifications comprise a prioritization and visual guidance to where these events occurred in the spatial context. These tasks are handled by the AR application for the HoloLens, named *Line MonitAR* (subsequently *MonitAR*). A number of design and hardware considerations

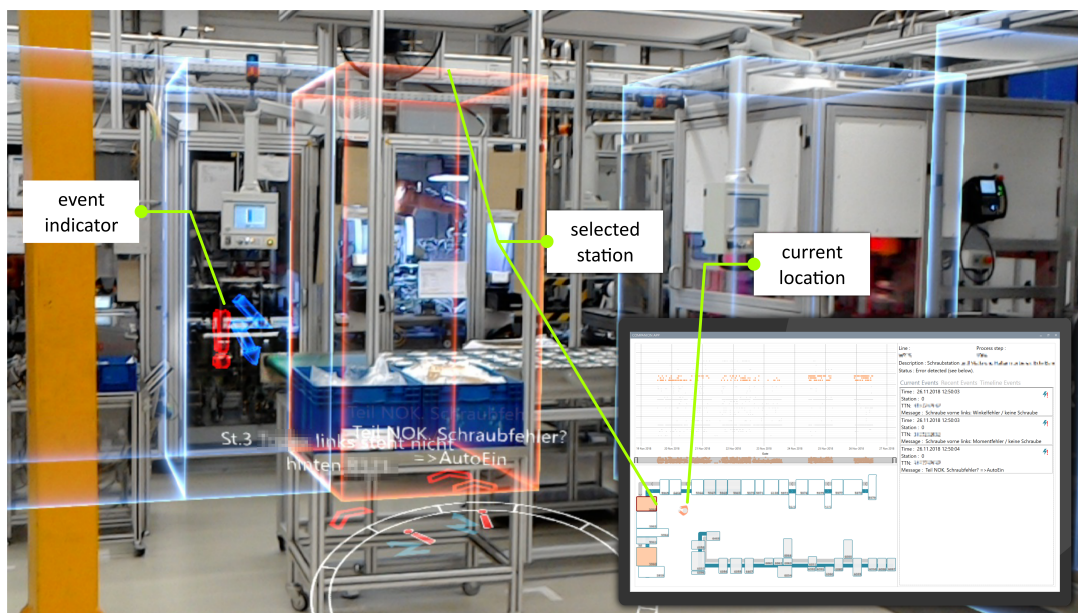


have to be made for a HoloLens application (or AR in general), primarily including the limited field of view and computing power, but also aspects such as display brightness. *MonitAR* only uses simple graphical elements, which keeps the necessary computational power low and improves the battery life of the HoloLens.

More importantly, the limited field of view demands to only display the most critical information within *MonitAR* to avoid a visual overload of the operators. For well-lit environments, e.g., the shop floor, display brightness and contrast between the virtual elements on the AR-screen and the real world becomes an issue. Overall, the visualization overlay provided by *MonitAR* prioritizes being functional and helpful. However, engaging the user with a visually appealing interface and graphical scene elements might be helpful for improving the acceptance of a new modality, as the application has to remain unobtrusive.

*MonitAR* comprises two components. First, each process step in the production line is presented as a 3D frame around its real-world counterpart (see Figure 6.3, left), which indicate the current status of the individual process steps. *MonitAR* intentionally does not perform any occlusion checks to also show process steps that may be hidden behind other machinery. This way, the status of all machines is visible when the operators look into their direction. Second, a continuously visible compass floats at hip height in front of the operators to provide directions towards currently erroneous machines (see Figure 6.3, right). The decision for a compass as primary navigational support element is the result of discussions with domain experts and has benefits compared to more sophisticated navigation aids: A compass is a well-known, intuitive concept that usually does not require any explanation or training. Furthermore, it can be assumed that operators who use such an AR system have basic knowledge of the production line layout. They primarily need to know *where* to go to deal with a specific event, rather than *how* to get there. More elaborate navigation overlays, such as plotting a route on the shop floor, seems therefore unnecessary at best and distracting at worst. The workflow proposed for *MonitAR* is split into four major phases: Alert, Steer, Survey, Instruct.

**Alert** During the alert phase, the virtual overlay is reduced to a minimum while waiting for events. This way, the operators are not distracted from normal activities. Furthermore, it is more likely that changes in the virtual scene caused by incoming events attract each operator's attention. Relevant events are shown as rotating 3D icons representing specific error classes in the frame of the respective step (see Figure 6.4). The approach distinguishes between events that are related to moving parts of the machine, time-related issues, such as process timeouts, and events that do not fit the other categories. The selection, which is synchronized with the companion app (see Section 6.3), is indicated by a visual



**Figure 6.4:** Augmented reality view of the production line. A 3D event icon indicates an error left to the selected process step (orange frame). All active error events and the selection are also indicated in the compass at the bottom. The inset on the right shows the corresponding view on the companion device. Note that the AR view has a finer resolution of process steps for this part of the production line than the server provides for the companion app.

highlight. Using similar icons, the compass indicates the direction of an event as discussed for the next phase.

**Steer** The second phase (steer) acts as a transition between the *monitoring* and *maintenance* tasks. Once the operators noticed an error, the system assists navigating to the location of the faulting machine. Selecting the problematic process step highlights it in the AR view, making it a natural navigation waypoint. The highlight is easily distinguishable from the other process steps, even if this step is occluded by other machines in the real world. Nevertheless, the operators might not directly face the affected process step. In that case, the aforementioned compass, which is always located in front of the users, indicates the direction of the selection and error events (see Figure 6.3). For each active event, an indicator element is shown on the compass, comprising an arrow pointing towards the event, an icon identifying the event category, and a textual description. Descriptions always face the users such that the text remains readable even for events behind the operators. If several events occur in a similar direction, the arrows intelligently stack up to avoid overlaps while taking up as little additional space as possible. Once the operators arrive at the indicated location, they can retrieve further details about the currently faulting machine through the companion app.

**Instruct** Once the problem has been identified, the instruct phase begins. Depending on how well-known the current issue is, operators can either explicitly retrieve instructions on how to deal with the issue on the tablet or get direct instructions overlaid on the affected machine parts in AR.

**Survey** Between the completion of maintenance tasks, operators can retrieve and analyze details about the issues they recently fixed on the companion app. This allows them to gain further insights, for example, about simultaneously occurring errors. The visual analysis capabilities of the companion app are detailed below.

The survey and in particular the instruct phase are both primarily associated with the *maintenance* task. Since machine maintenance in AR has already been studied in many publications (see Section 4.1.3), the last two phases were not explored in greater detail in the prototype application.

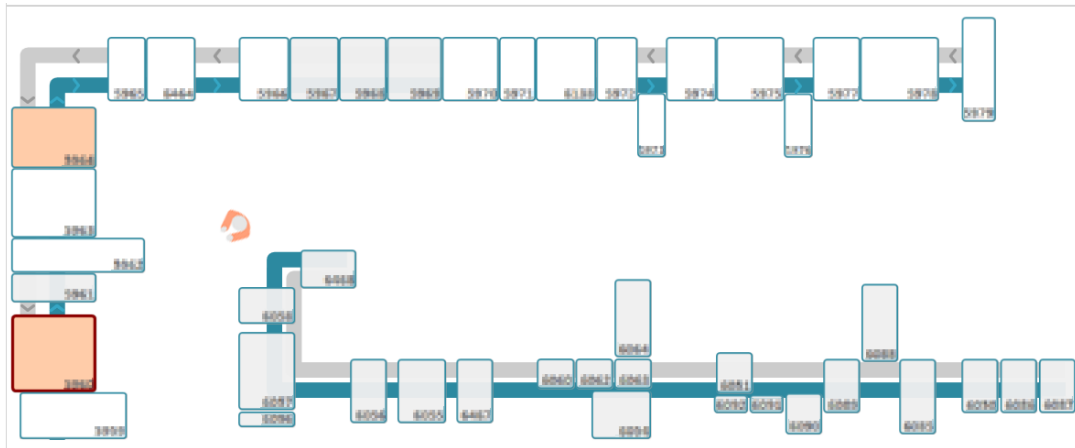
Due to their high importance, the currently relevant events are also shown in the companion app used for the visual analysis (see Section 6.3). This redundancy is necessary, as the operators have to be made aware of critical events regardless of which device they are currently focusing on. In contrast to the AR view, wherein the focus is to provide only the most critical information in the spatial context of the real world, the companion app provides a visual overview of the events across the entire production line. An example of the complete system in use is shown in Figure 6.4.

### Visual Analysis

Detailed analyses are performed on a mobile companion device offering enough space for the data and efficient interactions to explore them. Once an error has been identified, operators can retrieve contextual information about it, such as where the event occurred and which other errors exist at the process step. Understanding how often issues arose at the inspected process step in the past and which other events occurred under similar conditions may reveal further insights into dependencies between errors.

To support this process, the companion app comprises a *spatial layout view* of the production line, a *detail view* that provides information about the selected process step and events, and a *timeline view*. All components are linked within the companion app, and with the AR visualization on the HoloLens.

**Layout View** In addition to the information provided in AR, the companion app features an abstract map of the production line in the *Layout View*. The view is similar to the layout view presented in Section 5.3.1. It helps to provide

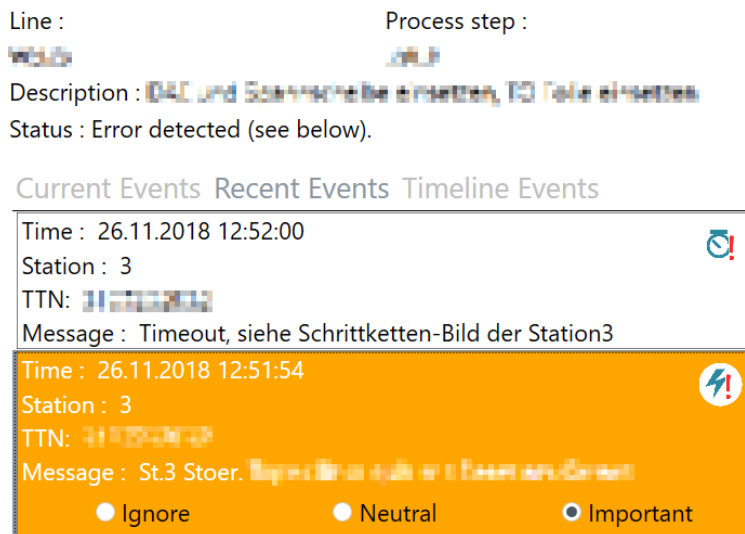


**Figure 6.5:** The *Layout View* presents an abstract portrayal of the production line. The operator’s position and view direction are indicated with a small orange icon. Erroneous process steps are filled with orange. In case no data is available, they are grayed out. The selected step has a red border.

information about the spatial arrangement of the production line’s process steps, to identify the location of current error events (see Figure 6.5), and to show the operator’s position. As the interaction precision on a mobile device is limited compared to interactions with mouse and keyboard, the granularity of the layout was restricted to the process steps (compared to the stations level in the desktop analysis presented in Section 5.3). Users can select any process step to obtain its status in the *Detail View* (see below). Due to the similarity of the displayed information, the *Layout View* can be seen as a mediator between the views on the companion device and the AR application.

**Detail View** The *Detail View* presents further information about the currently selected process step. Besides basic location information, such as the line, a process step identifier, a description, and the current status, it details current and historical error events for the process step. Each event comprises information about the time at which it was reported, the exact station in the process step that reported the issue, the product being produced, and a description of the event. Furthermore, analogous to *MonitAR*, an icon indicates the category the event is related to.

Currently active events listed in the *Current Events* tab help to get a quick overview of all issues that caused a failure at a specific process step. Aside from knowing where errors exist, the description of active events is one of the most important pieces of information needed before beginning the actual maintenance of the machinery. Recent issues with a machine are also valuable information, which can be looked up for the past hour in the *Recent Events* tab. Finally, the

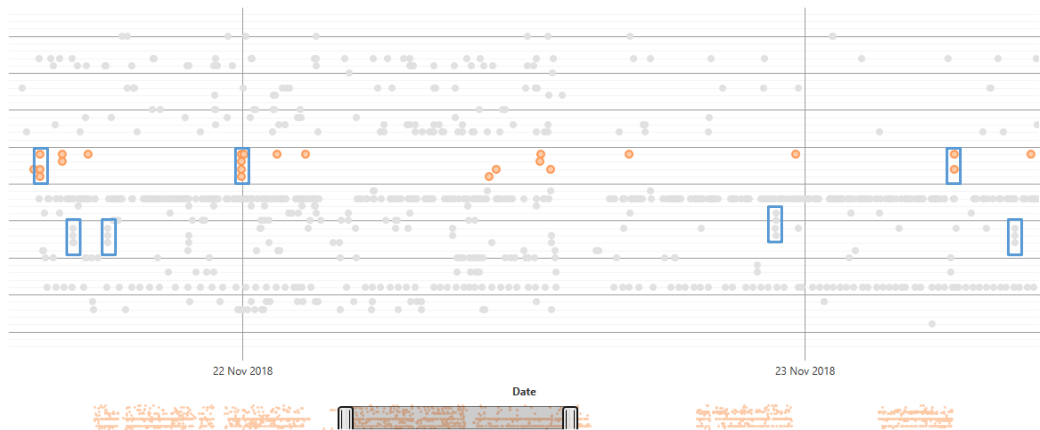


**Figure 6.6:** The *Detail View* presents detailed information about any selected process step. The top part shows information about the process step itself, whereas the bottom part can show current, recent, and filtered events.

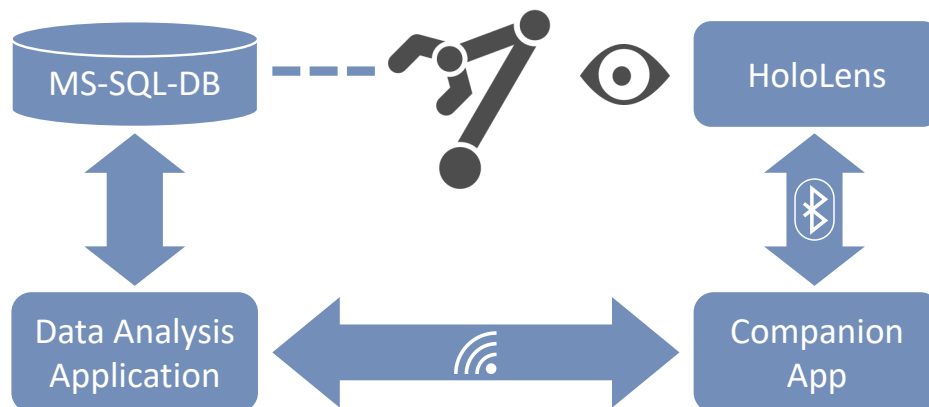
*Timeline Events* tab provides detailed information about all the events for the selected process step that are also visible in the *Timeline View* (see below).

For any event in these lists, the operators can change each event's importance between *ignore*, *neutral*, and *important*. This prioritization is applied to all events of the same event class, independently of the time when it occurred. *Ignored* events are hidden in the *Detail View* as well as the *Timeline View*, whereas *important* events are highlighted in the *Detail View* to be easily recognized.

**Timeline View** The *Timeline View* (see Figure 6.7) presents the events reported by the machines in the past seven days as a scatter plot. Except for an additional slider, the plot is entirely analogous to the timeline view in Section 5.3.1, except that it does not allow for any interactions. The x-axis represents time and the y-axis encodes the order in which machines process a workpiece in the production line. The top of the view represents the start of the production line, the bottom the end. Events of a selected process step (from AR or from the *Layout View*) are emphasized (see Figure 6.7). The overall structure of this view is similar to the Marey's graph used by Xu et al. [206], but this view only encodes special events, not regular processes. This way, experts can quickly recognize repeating patterns and verify assumptions of their understanding of the relations between events. For example, the *Timeline View* in Figure 6.7 reveals that some errors frequently occur simultaneously (emphasized with a blue border), which may indicate a relation between them. The additional slider that can be used to filter the timespan shown in the scatterplot. For easier navigation, the slider uses the



**Figure 6.7:** Similar to the approach in Section 5.3.1, the *Timeline View* presents the events of the past seven days as a scatterplot. The x-axis encodes time and the y-axis encodes the individual machines along the production line (from the start at the top to the end at the bottom). The selected process step's machine errors are highlighted in orange. The scrollbar enables operators to filter the data shown at the top, while still providing an overview of the available data. This view provides insights into which events often occur at similar times (emphasized in this figure with blue borders).



**Figure 6.8:** Machines report events to an MS-SQL server. This data is processed by a data analysis application that also provides the companion app with data. The companion app forwards live data to the HoloLens application and exchanges selection and position data via Bluetooth.

entire timeline in miniature as its background such that users always retain an overview of the available data.

### Implementation Details

The machines in a production line report the beginning and end of events to a Microsoft SQL database. The presented system combines three components to

process the data and present it to the operators. First, a *data analysis application* regularly checks if there are any new events. In case there are, it tries to match any still ongoing event with events that flag the end of an error. The matched events are stored in the database so that they can be retrieved as historical events to the companion app. For ongoing events, the application checks how long the event already persists. The industry partner proposed that operators should be notified about error events that last for more than five seconds. Any events that match this criterion are sent to the companion app, which forwards the information to the HoloLens via Bluetooth.

*MonitAR* is implemented on the HoloLens using C++ and Direct3D. It establishes a connection from the companion app to the HoloLens using Bluetooth RFCOMM. Compared to a connection via WiFi, the connection can be established automatically once the two devices are paired, thus providing a much better user experience for users without IT expertise. An additional advantage of Bluetooth is that it does not rely on a factory-wide coverage of WiFi access points. Although there will be no more live-event updates in case WiFi is temporarily lost, *MonitAR* and the companion app can still exchange position information and operators can inspect the available historical event information.

The virtual 3D model of the production line shown by *MonitAR* is aligned with its real-world counterpart using ArUco markers [78, 161] when the application is first started. While the HoloLens builds up its own spatial mapping of the surroundings, the operator will have to actively seek out markers placed at previously defined locations in the production line. Once *MonitAR* recognized at least three markers, the transformation between the predefined marker locations in the virtual scene and their real-world locations given in the HoloLens' coordinate system is estimated by a least squares fitting [18]. This transformation is then applied to the complete 3D model to align it with the real world. Once the real-world position and orientation of the virtual model have been successfully established, *spatial anchors*<sup>1</sup> are created at the locations where *MonitAR* spotted the ArUco markers. These spatial anchors allow the HoloLens to find a location based on visual features it has found in the environment. Once set up, *MonitAR* can restore the transform of the virtual model using the anchors without requiring the user to look at the markers again. Marker detection is explicitly activated and deactivated on demand in order to improve battery life by switching off the front camera and detection algorithm and also to avoid updating marker locations by mistake. The system makes use of voice commands to control such more complex actions, as relying solely on the air tap gesture input of the HoloLens might tire out and frustrate the operators quickly, especially if repeated input is required for an action. Furthermore, in situations that require hands-free use of the system, gesture input would not be feasible.

<sup>1</sup> <https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-anchors>

The companion app is developed in C# and uses the Windows Presentation Foundation (WPF). Besides presenting the user with detailed event information, which would be difficult on the HoloLens, the companion app relays information from the data analysis application to the HoloLens. The companion itself obtains its information from the data analysis application by means of a Windows Communication Foundation (WCF) web service. The data transferred to the HoloLens via Bluetooth is formatted in JSON, which is easy to produce and consume both in C# (companion app) and C++ (*MonitAR*).

## 6.4 Evaluation

The final prototype of the presented approach was presented to four experts from the industry partner. The following presents the results of their feedback and additional observations made during the presentation of the prototype.

### Expert User Feedback

To test the presented technique with domain experts, the prototype was deployed in the aforementioned production line of the industry partner. The group of domain experts consists of operators who are directly responsible for the handling the machines, and employees who are responsible for implementing Industry 4.0 concepts at the production line and keep track of the whole production process. In total, four experts participated in this first feedback round. Each expert tested the combination of both devices for approximately 10 to 15 minutes, observing the production process with live events. The *System Usability Scale* (SUS) [36] was applied for usability feedback, combined with free-text questions to identify potential issues. *SUS* was specifically designed to get feedback from domain experts in an industrial context who only have limited time between their usual tasks to rate a system and have little or no prior experience with usability studies.

The SUS is a questionnaire comprising ten questions that, in summary, provide a normalized score between 0–100 for the usability of a system. The experts rated the approach with a score of 87.5, 77.5, 82.5 and 87.5, respectively. Based on the rating categories used by Bangor et al. [21], all of these ratings correspond to good to excellent scores and are above a value of 68, which they found to be the average across 3,500 SUS results.

Overall, the experts stated in the free-text fields that they found the system was helpful to them. They appreciated the visual presentation, general ease of use, and the fast responsiveness to live error events. Navigational support by the compass was specifically identified as a helpful feature. Also, two participants commented that having no occlusion in the AR view is advantageous because it



can visually provide information about machines that are physically occluded by other machines in the real world. Furthermore, a more detailed 3D representation of individual machines, which should include the parts in the machine, was noted as a useful feature for the future. The operators from the production line agreed that the tablet companion device used in the current design was still too large to conveniently carry it with them at all times. An additional comment stated that the HoloLens is still too bulky for all-day usage.

### Further Results & Discussion

Feedback from domain experts and observations made during live testing of the system confirm that the design presents a feasible solution for monitoring and analyzing a production line. The feedback also provided valuable insights into the limitations of the approach and possible future extensions.

Operators at the production line immediately adopted the idea of using an immersive AR system after wearing the device and experiencing the AR view for the first time. They appreciated the real-time responsiveness of the prototype that displays error events within a few seconds, slightly faster than the Andon board that is already installed and used at the factory. While the limited comfort of wearing the HoloLens was commented upon, operators still successfully repaired a malfunction that required both operating the physical control board outside of the machine and manually readjusting parts inside the machine with limited available space, without being hindered by the headset.

With regard to user interaction and experience, all participants confirmed that the system design is easy to understand and use. One of the domain experts stated that it might be helpful to automatically select process steps of new events that were previously flagged as being important. On the one hand, this may emphasize the visual feedback of the event and therefore increase the reaction time. On the other hand, it may interrupt any current tasks, independent of their current progress, which means that the highlighted location may change multiple times or a currently maintained station may lose focus unexpectedly. It is currently unclear if one of the arguments outweighs the others and further investigations are needed to make a statement about the potential of such a process. Another question that was raised during the live test was how the issues of depth occlusions of event icons and overlapping text sprites could be solved. Currently, event icons that are located along a user's line of sight occlude one another, limiting the overview in some cases. Parts of this effect could be eased by moving the shown text to increase its overlap or by showing only event descriptions or prioritized events in case of an overlap.

The tablet companion device was less well-received, as the used 10" tablet was too large to easily carry around at all times without an obvious solution for quickly stowing it away. Nevertheless, the tablet is still essential as the current

level of input interactions available to most AR headsets is still limited and therefore an approach such as the presented one benefits from an additional physical input device.

While the presented system is specifically tailored to the production line scenario, the general concept of this immersive analytics system, built from an AR device for situated visualizations and a more traditional device for detailed on-site information and interaction, is applicable to other domains of similar composition. This broadly includes scenarios where critical information that benefits from being displayed in a spatial context needs to be presented while at the same time having quick access to a large amount of historical or related data is advantageous. For example, emergency and rescue services is a possible scenario for introducing immersive monitoring and analysis [38].

## Conclusion and Outlook

This thesis presented visual analytics approaches to support domain experts during the first phases of the product lifecycle, specifically the design, planning, and production phase of a product. The following first recapitulates the contributions of the different chapters of this thesis (Section 7.1). Afterwards, the results are discussed regarding the research questions stated in Section 1.1 of the thesis (Section 7.2). At last, an outlook of still open challenges and how visual analytics may help to solve them is presented (Section 7.3).

### 7.1 Summary of Contributions

The beginning of the thesis (Chapter 1) presented the overall research question “*How can visual analytics support the overall production quality along the first part of the product lifecycle?*”. As this research question is very broad, it was divided into three aspects that are part of the product lifecycle.

**Chapter 3** introduced approaches to help to understand the relationships of patents and other documents with keywords. The first approach focused on the analysis of the relations between patents by projecting classes of the international patent classification based on their co-usage in patents. Afterwards, this approach was extended to allow the visual analysis of general concepts that are used across documents (patents, websites, scientific literature).

**Chapter 4** presented visual analytics approaches that support planning experts to create or improve production line layouts. The first part supports domain experts in optimizing factory layouts regarding the paths workers have to take by suggesting which movable components in a factory should be relocated and where to locate these components to. In the second part, the desktop-based simulator for advanced manufacturing (SAM), which allows experts to simulate, compare, and discover new layouts in a modular production line setup, is extended with an augmented reality approach. It allows experts to simulate and discover new

layouts analogous to SAM and allows experts to compare the differences between two layouts.

**Chapter 5** focused on aiding technical management staff in understanding and analyzing the relation of errors reported by machinery in a production line. The first part focused on the spatio-temporal analysis of correlations between error classes. It combined the spatial information of a production line's layout with the temporal information when errors were reported and information about the event classes' correlations. The second part presented a visual analysis approach, in which the error event temporal distribution was decomposed into trend, seasonal, and remainder series to find recurring event patterns and outliers.

**Chapter 6** combined the approaches from the previous chapters by combining the analysis of events with an augmented reality application to allow operators on the shop floor to monitor and analyze events on-site. Four major tasks (monitor, analyze, predict, maintain) were identified that need to be considered during everyday production activities. This chapter first focused on the first two tasks and then proposes how to include the third task in the approach.

**This chapter** concludes this thesis with a discussion on how the research questions introduced in Chapter 1 were addressed in the other chapters. Finally, an outlook of still open challenges, such as the visual analysis of the quality of the available data, collaborative analysis, and the combination of visual analysis and prediction to allow for predictive maintenance, are presented.

## 7.2 Discussion

At the beginning of the thesis, the overarching research question “*How can visual analytics support the overall production quality along the first part of the product lifecycle?*” was stated as a motivation to show possible application fields of visual analytics in the production domain. The following first discusses the results of the presented approaches with regard to the research questions presented in Section 1.1. Then, the overall applicability of visual analytics for advanced manufacturing is discussed, as stated in the overarching research question, and a possible integration of the product lifecycle with the knowledge generation model presented in Section 2.1 is presented.

### Research Question 1

*How can visualization help to understand the relation of topics relevant to the product to be designed?*

There are several ways how the relationship of topics that are relevant when designing new products can be analyzed. The approaches presented in Chapter 3 focus on visualizing these relationships by analyzing the co-occurrences of the

topics across a large number of documents, such as patents, question-and-answer websites, or scientific literature. The results of the presented approaches indicate that the projection of topics based on their pairwise similarity can help experts to understand, which groups of topics exist. Further, it is important to allow experts to access further details about the documents and to separately present the most similar topics of a set of selected topics to avoid the intuition that projections always show the correct topics nearby on a map.

## Research Question 2

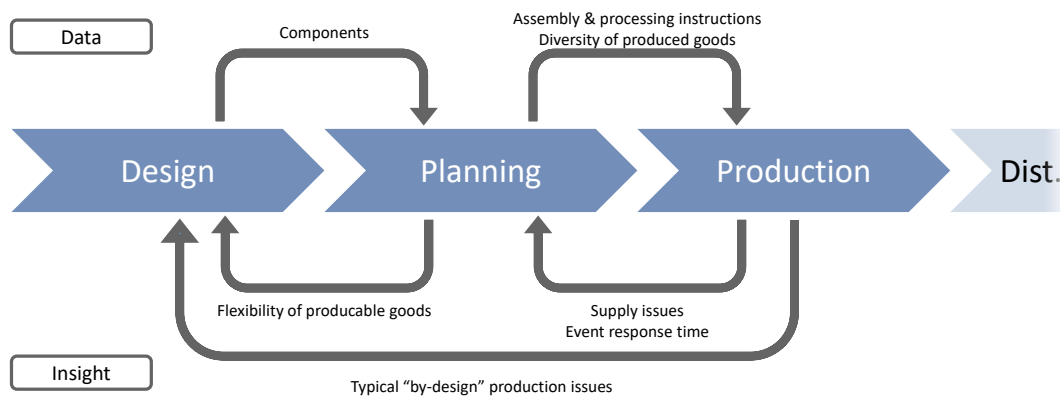
*How can visual analytics support layout planning of factories and production lines?*

Planning or optimizing existing factory or production line layouts is an important step towards running a production line efficiently. The evaluation of the approaches in Chapter 4 showed that visual analytics concepts can help domain experts in solving this task more efficiently. Using evolutionary algorithms to guide layout planners when optimizing layouts proved to be a promising approach. Further, domain experts confirmed that extending desktop-based visual analysis systems with an augmented reality application helps them in coping with their everyday tasks.

## Research Question 3

*How to support domain experts during the exploratory analysis of event data to understand issues in a production line?*

During production, it is important to assert that a production line runs without unplanned interruptions. To prevent errors from occurring, it is important to understand, if there are any temporal patterns in their occurrence and which errors have a cause-effect relationship. Such knowledge can help experts to understand what caused errors so that they can be prevented in the future. The evaluations of the approaches presented in Chapter 5 show that using the explorative nature of visual analytics can help to uncover such systematic issues and improve the processes that caused them. Further, Chapter 6 shows that the addition of situated analytics through wearable augmented reality devices further improves the responsiveness of operators on the shop floor.



**Figure 7.1:** By combining the knowledge generation model (see Section 2.1) with the first part of the product lifecycle, a holistic approach to transport knowledge between across the creation phases of a product can prevent issues when creating new products in the future.

## Overarching Research Question

*How can visual analytics support the overall production quality along the first part of the product lifecycle?*

Most related work and the approaches presented in this thesis handle challenges that arise during everyday tasks of any phase of the product lifecycle as isolated problems. The presented approaches show that visual analytics can successfully be applied in the production domain to assist experts with their tasks. However, the insights gained during any phase can also be used in prior process steps. This can prevent or ease problems that are caused by design, e.g., because of a problematic layout of the production line or because the design of the product makes some process steps more difficult than they have to be.

The combination of the knowledge generation model (see Section 2.1) and the product lifecycle (see Section 2.2.2) results in a holistic workflow that may be able to improve the production quality in the long run. The resulting workflow (see Figure 7.1) is similar to the idea of the Building Information Model (BIM) [19] from the architecture domain. Therein, all available information should be accessible during any stage of the creation of a building from its sketch until the finished construction. This way, later trades, such as piping, have access to the data at early stages of the design process and may be able to inform earlier trades about problematic designs that they may not be aware of. In the context of the production domain, the design phase should forward needed components or early drafts of the product that is currently designed so that the production planning experts can give feedback about the production complexity of the product. This way, issues, such as a possibly limited flexibility of the production line regarding the number of producible product variants, can be communicated early on. Further, the production experts can give feedback about expectable

issues during the production that may be caused by some design aspect of the product. In addition, the planning and production experts can exchange their experiences, for example, to provide precise assembly instructions or issues with the planned production line layout, e.g., regarding the response time in case of an issue with the machinery.

## 7.3 Outlook

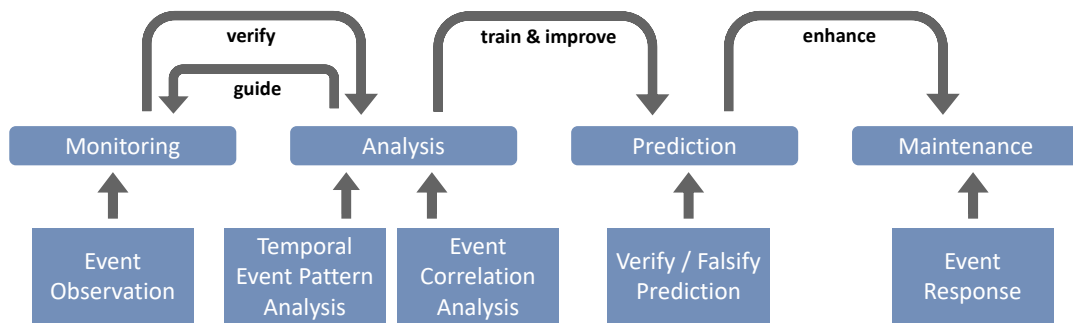
The approaches that are presented in this thesis show that visual analytics can help experts in the production domain in solving their tasks along the different phases of the product lifecycle. However, the development and evaluation of the approaches also showed that there are still open challenges in which visual analytics can help.

### Visual Data Quality Analysis

Understanding the quality of the available data is one of the major challenges in the production domain. Many producing companies are aware that collecting and analyzing data can potentially help them in some way. However, it is often unclear what kind of data may help during the analysis. As a result, an often followed practice is to collect and store all the data that is available. The quality of the data and what kind of insights they hope to find are often unclear in this process. The explorative nature of visual analytics can help to understand the quality and the analysis potential of the available data and possibly indicate, which data may still be missing for an effective analysis. Allowing experts to explore the general connection between data collected from different sources, as well as providing information about inconsistencies and incomplete or missing data helps to get an understanding of the available data. Further, the improved data quality helps during the analysis later on.

### Collaborative Immersive Data Analysis

The approaches in Section 4.3 and Chapter 6 show the potential of an on-site analysis of the data. Although the isolated approaches already help the domain experts, there is potential to improve the overall productivity by integrating multiple domain expert groups in the main application categories (see Figure 6.2). These groups, for example, technical managers and workers, could exchange their insights more efficiently. Figure 7.2 extends the categories with the tasks of the respective roles in a production line (bottom) and how they can improve the overall productivity by sharing their specialized knowledge (top). On the one hand, technical managers, such as the head of a production line, could



**Figure 7.2:** Complementing the approach from Section Chapter 6 with the lessons learned previously allows for a more efficient transfer of information and allows to provide and improve predictions about upcoming issues.

*provide insights* during their data analysis directly to the affected personnel. Those workers can *verify findings* and work together with their superiors to find and solve possible problems indicated by the findings. On the other hand, operators at the shop floor could *report observations* of unusual behavior directly to their superiors to prevent the loss of insights and also allows for a more timely response to such events. This would allow for a quicker response in case multiple machines or even multiple production lines are affected (which an operator might not be aware of).

Aside from becoming useful for technical management staff, the extension towards the analysis of correlations would also contribute towards the prediction component of the design concept of the system (see Figure 6.2). By including the analysis conducted by the technical management staff, the analysis of correlations could be used to *build and provide models* for the prediction of future errors. Incorporating the event correlation analysis into an overall analysis concept would have multiple advantages. At the same time, operators could *provide feedback* about the prediction quality, e.g., by rating the semantic plausibility of the provided correlations. Overall, this could lead to an *improved response time* to maintain broken machinery.

To prevent the operators from having to cope with analysis options that they have no time to use, the analysis component would need to have predefined role profiles, e.g., operators and technical management staff. Views such as the *Correlation View* presented in Section 5.3.1 could be further supported with extensions during the data projection, such as the approach presented in Section 3.4.

## Predictive Maintenance

As discussed in Section 7.3, the visual analysis and rating of event correlations can help to provide predictions of which errors may occur in the near future.



Visually conveying this information to workers can be seen as a first step towards predictive maintenance, as the correlation information can be used to indicate, *which stations* may break in the near future, *how likely* the breakdown is, and *how much time is left* before the error occurs. One drawback of using a correlation measure to predict events is its inability to check for the semantic plausibility of the prediction. Therefore, rating the semantic meaning of the predicted correlations helps to show only correlations that make sense. Visualizing the influence of the ratings given by the individual workers may further increase their motivation to contribute prediction ratings further and increase the acceptance of such a prediction system.

In summary, many open research questions provide room for further investigation on how to support the manufacturing domain through visual analytics. Although production is a long-established trade, an increasing amount of digitization and the introduction of concepts such as *Industry 4.0* force companies to find suitable approaches to deal with these data and benefit from them as much as possible. Visualization and Visual Analytics allow experts to investigate their data on different levels of detail, while still providing an easy entry point for specialists that did not have to deal with data analytics before.



---

## Author's Work

- [1] D. Herr, F. Beck and T. Ertl. “Visual Analytics for Decomposing Temporal Event Series of Production Lines”. In: *Proceedings of the 22nd International Conference Information Visualisation*. IV'18. 2018, pp. 251–259 (cit. on p. 96).
- [2] D. Herr, S. Grund and T. Ertl. “BlueCollar: Optimizing Worker Paths on Factory Shop Floors with Visual Analytics”. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019, pp. 1580–1589. URL: <http://hdl.handle.net/10125/59598> (visited on 2019-10-17) (cit. on pp. 68, 73, 74, 76–79, 81–83).
- [3] D. Herr, Q. Han, S. Lohmann, S. Brüggemann and T. Ertl. “Visual Exploration of Patent Collections with IPC Clouds”. In: *Proceedings of the 1st International Workshop Patent Mining and Its Applications*. Vol. 1292. CEUR-WS. CEUR-WS.org, 2014 (cit. on p. 30).
- [4] D. Herr, Q. Han, S. Lohmann and T. Ertl. “Visual Clutter Reduction through Hierarchy-based Projection of High-dimensional Labeled Data”. In: *Proceedings of Graphics Interface*. CIPS / ACM, 2016, pp. 109–116 (cit. on p. 30).
- [5] D. Herr, Q. Han, S. Lohmann and T. Ertl. “Hierarchy-based projection of high-dimensional Labeled Data to Reduce Visual Clutter”. In: *Computers & Graphics* 62 (2017), pp. 28–40 (cit. on p. 30).
- [6] D. Herr, K. Kurzhals and T. Ertl. “Visual Analysis for Spatio-temporal Event Correlation in Manufacturing”. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*. (To appear, accepted on 2019-08-19) (cit. on p. 96).
- [7] D. Herr, J. Reinhardt, R. Krüger, G. Reina and T. Ertl. “Immersive Visual Analytics for Modular Factory Layout Planning”. In: *Proceedings of the Workshop on Immersive Analytics*. IEEE, 2017. URL: [https://groups.inf.ed.ac.uk/vishub/immersiveanalytics/papers/IA\\_2278-paper.pdf](https://groups.inf.ed.ac.uk/vishub/immersiveanalytics/papers/IA_2278-paper.pdf) (visited on 2019-10-17) (cit. on p. 68).

- [8] D. Herr, J. Reinhardt, G. Reina, R. Krüger, R. Villanueva Ferrari and T. Ertl. “Immersive Modular Factory Layout Planning using Augmented Reality”. In: *Procedia CIRP* 72 (2018), pp. 1112–1117 (cit. on p. 68).
- [9] R. Krüger, D. Herr, F. Haag and T. Ertl. “Inspector-Gadget: Integrating Data Preprocessing and Orchestration in the Visual Analysis Loop”. In: *Proceedings of the EuroVis Workshop on Visual Analytics*. EuroVA. The Eurographics Association, 2015 (cit. on p. 7).
- [10] M. Raschke, D. Herr, T. Blascheck, T. Ertl, M. Burch, S. Willmann and M. Schrauf. “A visual approach for scan path comparison”. In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. Safety Harbor, Florida: ACM, 2014, pp. 135–142. ISBN: 978-1-4503-2751-0 (cit. on p. 14).

---

## Bibliography

- [11] E. E. Aleisa and L. Lin. “For effective facilities planning: Layout optimization then simulation, or vice versa?” In: *Proceedings of the 37th Conference on Winter Simulation*. WSC '05. Winter Simulation Conference, 2005, pp. 1381–1385. ISBN: 0-7803-9519-0 (cit. on p. 69).
- [12] M. Aleksy, M. Troost, F. Scheinhardt and G. T. Zank. “Utilizing HoloLens to Support Industrial Service Processes”. In: *Proceedings of IEEE 32nd International Conference on Advanced Information Networking and Applications*. AINA. 2018, pp. 143–148 (cit. on p. 71).
- [13] J. Alsakran, Y. Chen, D. Luo, Y. Zhao, J. Yang, W. Dou and S. Liu. “Real-Time Visualization of Streaming Text with a Force-Based Dynamic System”. In: *IEEE Computer Graphics and Applications* 32.1 (2012), pp. 34–45 (cit. on p. 35).
- [14] B. Alsallakh, W. Aigner, S. Miksch and M. E. Gröller. “Reinventing the contingency wheel: Scalable visual analytics of large categorical data”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2849–2858 (cit. on p. 99).
- [15] P. André, M. L. Wilson, A. Russell, D. A. Smith, A. Owens and M. C. Schraefel. “Continuum: Designing timelines for hierarchies, relationships and scale”. In: *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. UIST. ACM, 2007, pp. 101–110 (cit. on p. 100).
- [16] N. Andrienko, G. Andrienko and P. Gatalsky. “Exploratory spatio-temporal visualization: An analytical review”. In: *Journal of Visual Languages & Computing* 14.6 (2003), pp. 503–541 (cit. on p. 105).
- [17] R. Arias-Hernandez, L. T. Kaastra, T. M. Green and B. Fisher. “Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics”. In: *Proceedings of the 44th Hawaii International Conference on System Sciences*. Piscataway, N. J.: IEEE, 2011, pp. 1–10. ISBN: 978-1-4244-9618-1 (cit. on p. 128).

- [18] K. S. Arun, T. S. Huang and S. D. Blostein. “Least-Squares Fitting of Two 3-D Point Sets”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.5 (1987), pp. 698–700 (cit. on p. 145).
- [19] S. Azhar. “Building Information Modeling (BIM): Trends, Benefits, Risks, and Challenges for the AEC Industry”. In: *Leadership and Management in Engineering* 11.3 (2011), pp. 241–252. ISSN: 1532-6748 (cit. on p. 152).
- [20] J. Balakrishnan and C. H. Cheng. “Dynamic layout algorithms: A state-of-the-art survey”. In: *Omega* 26.4 (1998), pp. 507–521 (cit. on p. 68).
- [21] A. Bangor, P. Kortum and J. Miller. “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale”. In: *Journal of Usability Studies* 4.3 (2009), pp. 114–123 (cit. on p. 146).
- [22] P. Baudisch and R. Rosenholtz. “Halo: A Technique for Visualizing Off-screen Objects”. In: *Proceedings of the Conference on Human Factors in Computing Systems*. CHI. ACM, 2003, pp. 481–488 (cit. on p. 55).
- [23] K. Beard, H. Deese and N. R. Pettigrew. “A framework for visualization and exploration of events”. In: *Information Visualization* 7.2 (2008), pp. 133–151 (cit. on p. 100).
- [24] F. Beck, M. Burch, S. Diehl and D. Weiskopf. “A taxonomy and survey of dynamic graph visualization”. In: *Computer Graphics Forum* 36.1 (2017), pp. 133–159. ISSN: 1467-8659 (cit. on p. 100).
- [25] M. Behrisch, J. Davey, F. Fischer, O. Thonnard, T. Schreck, D. Keim and J. Kohlhammer. “Visual Analysis of Sets of Heterogeneous Matrices Using Projection-Based Distance Functions and Semantic Zoom”. In: *Computer Graphics Forum* 33 (2014), pp. 411–420. ISSN: 1467-8659 (cit. on p. 99).
- [26] M. Behrisch, F. Korkmaz, L. Shao and T. Schreck. “Feedback-driven interactive exploration of large multidimensional data supported by visual classifier”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2014, pp. 43–52 (cit. on p. 99).
- [27] J. Bernard, T. Ruppert, O. Goroll, T. May and J. Kohlhammer. “Visual-Interactive Preprocessing of Time Series Data”. In: *Proceedings of SIGRAD 2012*. Växjö, Sweden, 2012, pp. 39–48 (cit. on p. 100).
- [28] J. Bertin. *Semiology of graphics: Diagrams, networks, maps*. Madison, WI, USA: University of Wisconsin Press, 1983. ISBN: 0-299-09060-4 (cit. on p. 70).
- [29] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft. “When Is “Nearest Neighbor” Meaningful?” In: *Proceedings of the Conference on Database Theory*. Ed. by C. Beeri and P. Buneman. ICDT’99. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 217–235. ISBN: 978-3-540-49257-3 (cit. on p. 14).

- 
- [30] M. Billinghamurst, A. Clark and G. Lee. “A survey of augmented reality”. In: *Foundations and Trends in Human–Computer Interaction* 8.2-3 (2015), pp. 73–272 (cit. on p. 84).
- [31] C. Blundell, Y. W. Teh and K. A. Heller. “Bayesian Rose Trees”. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI, 2010, pp. 65–72 (cit. on p. 65).
- [32] M. Bögl, W. Aigner, P. Filzmoser, T. Gschwandtner, T. Lammarsch, S. Miksch and A. Rind. “Visual analytics methods to guide diagnostics for time series model predictions”. In: *Proceedings of the IEEE VIS Workshop on Visualization for Predictive Analytics*. VPA. 2014 (cit. on p. 100).
- [33] M. Bögl, W. Aigner, P. Filzmoser, T. Gschwandtner, T. Lammarsch, S. Miksch and A. Rind. “Integrating Predictions in Time Series Model Selection”. In: *Proceedings of the EuroVis Workshop on Visual Analytics*. EuroVA. The Eurographics Association, 2015, pp. 73–77 (cit. on p. 100).
- [34] M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch and A. Rind. “Visual analytics for model selection in time series analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2237–2246 (cit. on p. 100).
- [35] M. Brehmer, S. Ingram, J. Stray and T. Munzner. “Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2271–2280 (cit. on p. 35).
- [36] J. Brooke. “SUS: A retrospective”. In: *Journal of Usability Studies* 8.2 (2013), pp. 29–40 (cit. on p. 146).
- [37] S. Brüggmann, N. Bouayad-Agha, A. Burga, S. Carrascosa, A. Ciaramella, M. Ciaramella, J. Codina-Filba, E. Escorsa, A. Judea, S. Mille, A. Müller, H. Saggion, P. Ziering, H. Schütze and L. Wanner. “Towards content-oriented patent document processing: Intelligent patent analysis and summarization”. In: *World Patent Information* 40 (2015), pp. 30–42 (cit. on p. 57).
- [38] P. Brunetti, A. Croatti, A. Ricci and M. Viroli. “Smart Augmented Fields for Emergency Operations”. In: *Procedia Computer Science* 63 (2015), pp. 392–399. ISSN: 1877-0509 (cit. on p. 148).
- [39] M. Burch, S. Lohmann, D. Pompe and D. Weiskopf. “Prefix Tag Clouds”. In: *Proceedings of the IEEE International Conference on Information Visualisation*. IEEE, 2013, pp. 45–50 (cit. on p. 34).

- [40] N. Cao, C. Lin, Q. Zhu, Y.-R. Lin, X. Teng and X. Wen. “Voila: Visual Anomaly Detection and Monitoring with Streaming Spatiotemporal Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 23–33 (cit. on p. 100).
- [41] B. C. M. Cappers and J. J. van Wijk. “Exploring multivariate event sequences using rules, aggregations, and selections”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 532–541 (cit. on p. 101).
- [42] S. K. Card, J. D. Mackinlay and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. The Morgan Kaufmann series in interactive technologies. San Francisco, CA, USA and Great Britain: Morgan Kaufmann Publishers, 1999. ISBN: 1-55860-533-9 (cit. on p. 9).
- [43] J. V. Carlis and J. A. Konstan. “Interactive visualization of serial periodic data”. In: *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*. UIST. ACM, 1998, pp. 29–38 (cit. on p. 101).
- [44] T. P. Caudell and D. W. Mizell. “Augmented reality: An application of heads-up display technology to manual manufacturing processes”. In: *Proceedings of the 25th Hawaii International Conference on System Sciences*. Vol. ii. 1992, pp. 659–669 (cit. on p. 70).
- [45] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert and T. Ertl. “Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2012, pp. 143–152 (cit. on p. 100).
- [46] Y.-X. Chen, R. Santamaría, A. Butz and R. Therón. “TagClusters: Semantic Aggregation of Collaborative Tags Beyond TagClouds”. In: *Proceedings of the International Symposium on Smart Graphics*. Springer, 2009, pp. 56–67 (cit. on p. 34).
- [47] Y. Chen, P. Xu and L. Ren. “Sequence Synopsis: Optimize Visual Summary of Temporal Event Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 45–55 (cit. on p. 101).
- [48] R. B. Cleveland, W. S. Cleveland, J. E. McRae and I. Terpenning. “STL: A seasonal-trend decomposition procedure based on Loess”. In: *Journal of Official Statistics* 6.1 (1990), pp. 3–73 (cit. on pp. 15, 16, 123, 124).
- [49] M. Colledani, D. Gyulai, L. Monostori, M. Urgo, J. Unglert and F. van Houten. “Design and management of reconfigurable assembly lines in the automotive industry”. In: *CIRP Annals – Manufacturing Technology* 65.1 (2016), pp. 441–446 (cit. on p. 69).



- 
- [50] C. Collins, M. S. T. Carpendale and G. Penn. “DocuBurst: Visualizing Document Content using Language Structure”. In: *Computer Graphics Forum* 28.3 (2009), pp. 1039–1046. ISSN: 1467-8659 (cit. on p. 34).
- [51] L. Da Xu, C. Wang, Z. Bi and J. Yu. “AutoAssem: An automated assembly planning system for complex products”. In: *IEEE Transactions on Industrial Informatics* 8.3 (2012), pp. 669–678 (cit. on p. 69).
- [52] D. L. Davies and D. W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1.2 (1979), pp. 224–227 (cit. on p. 48).
- [53] F. Doil, W. Schreiber, T. Alt and C. Patron. “Augmented Reality for Manufacturing Planning”. In: *Proceedings of the Workshop on Virtual Environments. EGVE '03*. New York, NY, USA: ACM, 2003, pp. 71–76. ISBN: 1-58113-686-2 (cit. on p. 70).
- [54] W. Dou, L. Yu, X. Wang, Z. Ma and W. Ribarsky. “HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2002–2011 (cit. on p. 34).
- [55] A. Drira, H. Pierreval and S. Hajri-Gabouj. “Facility layout problems: A survey”. In: *Annual Reviews in Control* 31.2 (2007), pp. 255–267. ISSN: 1367-5788 (cit. on p. 69).
- [56] T. Dwyer, K. Marriott and P. J. Stuckey. “Fast Node Overlap Removal”. In: *Proceedings of the International Conference on Graph Drawing. GD'05*. Springer, 2006, pp. 153–164. ISBN: 3-540-31425-3 (cit. on p. 40).
- [57] *Elastic Search*. URL: <http://www.elasticsearch.org> (visited on 2019-10-17) (cit. on p. 37).
- [58] N. Elmqvist and J.-D. Fekete. “Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.3 (2010), pp. 439–454 (cit. on pp. 35, 49).
- [59] M. Endo, M. Ueno and T. Tanabe. “A Clustering Method Using Hierarchical Self-Organizing Maps”. In: *Journal of VLSI signal processing systems for signal, image and video technology* 32.1-2 (2002), pp. 105–118 (cit. on p. 33).
- [60] V. A. Epanechnikov. “Non-parametric estimation of a multivariate probability density”. In: *Theory of Probability & Its Applications* 14.1 (1969), pp. 153–158 (cit. on p. 54).
- [61] *EPO – European Publication Server*. URL: <https://data.epo.org/publication-server> (visited on 2019-10-17) (cit. on p. 37).

- [62] *EPO Worldwide Patent Statistical Database (PATSTAT)*. URL: <https://www.epo.org/searching-for-patents/business/patstat.html> (visited on 2019-10-17) (cit. on p. 37).
- [63] J. A. Erkoyuncu, I. F. del Amo, M. Dalle Mura, R. Roy and G. Dini. “Improving efficiency of industrial maintenance with context aware adaptive authoring in augmented reality”. In: *CIRP Annals – Manufacturing Technology* (2017) (cit. on p. 71).
- [64] European Patent Office. *Espacenet*. URL: <https://www.epo.org/searching-for-patents/technical/espacenet.html> (visited on 2019-10-17) (cit. on p. 37).
- [65] European Patent Office. *European Patent Register*. URL: <https://register.epo.org> (visited on 2019-10-17) (cit. on p. 37).
- [66] European Patent Office. *Open Patent Services (OPS)*. URL: <https://www.epo.org/searching-for-patents/data/web-services/ops.html> (visited on 2019-10-17) (cit. on p. 37).
- [67] R. J. Everett and A. S. Sohal. “Individual Involvement and Intervention in Quality Improvement Programmes: Using the Andon System”. In: *IJQRM* 8.2 (1991), pp. 21–34 (cit. on pp. 98, 135).
- [68] A.-C. Falck, R. Örtengren and D. Högberg. “The impact of poor assembly ergonomics on product quality: A cost-benefit analysis in car manufacturing”. In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 20.1 (2010), pp. 24–41. ISSN: 1520-6564 (cit. on p. 71).
- [69] P. Federico, F. Heimerl, S. Koch and S. Miksch. “A Survey on Visual Approaches for Analyzing Scientific Literature and Patents”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (2017), pp. 2179–2198 (cit. on p. 36).
- [70] S. Feiner, B. Macintyre and D. Seligmann. “Knowledge-based Augmented Reality”. In: *Communications of the ACM* 36.7 (1993), pp. 53–62. ISSN: 0001-0782 (cit. on p. 71).
- [71] J. W. Fondahl. *A Non-Computer Approach to the Critical Path Method for the Construction Industry*. Technical Report (Stanford University. Department of Civil Engineering). Stanford, CA, USA: Department of Civil Engineering, Stanford University, 1962 (cit. on p. 27).
- [72] H. Ford and S. Crowther. *My Life and Work*. Garden City, NY, USA: Doubleday, Page & Company, 1923 (cit. on p. 24).
- [73] D. Fried and S. G. Kobourov. “Maps of Computer Science”. In: *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2014, pp. 113–120 (cit. on p. 35).

- 
- [74] T. M. J. Fruchterman and E. M. Reingold. “Graph drawing by force-directed placement”. In: *Journal of Software: Practice & Experience* 21.11 (1991), pp. 1129–1164 (cit. on p. 33).
- [75] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada and H. Okuda. “Topigraphy: Visualization for large-scale tag clouds”. In: *Proceedings of the International Conference on World Wide Web*. 2008, pp. 1087–1088 (cit. on p. 34).
- [76] E. R. Gansner and S. C. North. “Improved Force-Directed Layouts”. In: *Proceedings of the International Symposium on Graph Drawing*. Springer, 1998, pp. 364–373 (cit. on p. 33).
- [77] F. J. García-Fernández, M. Verleysen, J. A. Lee and I. Díaz. “Stability Comparison of Dimensionality Reduction Techniques Attending to Data and Parameter Variations”. In: *Proceedings of the EuroVis Workshop Visual Analytics using Multidimensional Projections*. Ed. by M. Aupetit and L. van der Maaten. The Eurographics Association, 2013. ISBN: 978-3-905674-53-8 (cit. on p. 33).
- [78] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas and R. Medina-Carnicer. “Generation of fiducial marker dictionaries using mixed integer linear programming”. In: *Pattern Recognition* 51 (2016), pp. 481–491 (cit. on p. 145).
- [79] M. Giereth, S. Koch, M. Rotard and T. Ertl. “Web Based Visual Exploration of Patent Information”. In: *Proceedings of the IEEE International Conference on Information Visualization*. IV '07. IEEE, 2007, pp. 150–155. ISBN: 0-7695-2900-3 (cit. on p. 36).
- [80] A. Gillet, M. Sanner, D. Stoffler, D. Goodsell and A. Olson. “Augmented Reality with Tangible Auto-Fabricated Models for Molecular Biology Applications”. In: *Proceedings of IEEE VIS*. 2004, pp. 235–242. ISBN: 0-7803-8788-0 (cit. on p. 70).
- [81] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha and N. Cao. “EventThread: Visual Summarization and Stage Analysis of Event Sequence Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 56–65 (cit. on p. 101).
- [82] D. Gyulai, A. Pfeiffer, B. Kádár and L. Monostori. “Simulation-based production planning and execution control for reconfigurable assembly cells”. In: *Procedia CIRP* 57 (2016), pp. 445–450 (cit. on p. 69).
- [83] Q. Han, F. Heimerl, J. Codina-Filba, S. Lohmann, L. Wanner and T. Ertl. “Visual patent trend analysis for informed decision making in technology management”. In: *World Patent Information* 49 (2017), pp. 34–42 (cit. on p. 36).

- [84] P. E. Hart, N. J. Nilsson and B. Raphael. “A Formal Basis for the Heuristic Determination of Minimum Cost Paths”. In: *IEEE Transactions on Systems Science and Cybernetics* 4.2 (1968), pp. 100–107. ISSN: 0536-1567 (cit. on pp. 73, 78).
- [85] J. A. Hartigan. “Printer graphics for clustering”. In: *Journal of Statistical Computation and Simulation* 4.3 (1975), pp. 187–213 (cit. on p. 70).
- [86] K. S. Hasan and V. Ng. “Automatic keyphrase extraction: A survey of the state of the art”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2014, pp. 1262–1273 (cit. on p. 33).
- [87] Y. Hassan-Montero and V. Herrero-Solana. “Improving Tag-Clouds as Visual Information Retrieval Interfaces”. In: *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies*. 2006, pp. 25–28 (cit. on p. 34).
- [88] T. Hastie, R. Tibshirani, J. Friedman and J. Franklin, eds. *The elements of statistical learning: Data mining, inference and prediction*. Vol. 27. 2005 (cit. on pp. 13, 47).
- [89] M. Hauschild and M. Pelikan. “An introduction and survey of estimation of distribution algorithms”. In: *Swarm and Evolutionary Computation* 1.3 (2011), pp. 111–128 (cit. on pp. 18, 78).
- [90] F. Heimerl, Q. Han, S. Koch and T. Ertl. “CiteRivers: Visual Analytics of Citation Patterns”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 190–199 (cit. on p. 36).
- [91] F. Heimerl, M. John, Q. Han, S. Koch and T. Ertl. “DocuCompass: Effective exploration of document landscapes”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2016, pp. 11–20. ISBN: 978-1-5090-5661-3 (cit. on p. 36).
- [92] F. Heimerl, S. Koch, H. Bosch and T. Ertl. “Visual classifier training for text document retrieval”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2839–2848 (cit. on p. 99).
- [93] F. Heimerl, S. Lohmann, S. Lange and T. Ertl. “Word Cloud Explorer: Text Analytics Based on Word Clouds”. In: *Proceedings of the 47th Hawaii International Conference on System Sciences*. HICSS. IEEE, 2014, pp. 1833–1842 (cit. on p. 34).
- [94] S. Henderson and S. Feiner. “Exploring the benefits of augmented reality documentation for maintenance and repair”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.10 (2011), pp. 1355–1368 (cit. on p. 71).

- 
- [95] Herve Couetoux [FR], Nicolas Guilloz [FR] and Olivier Scheffges [FR]. “Method For Controlling the Starter of a Combustion Engine and Application Thereof”. EP2222950 (B1) (cit. on p. 31).
- [96] S. Huan, L. Aiping, L. Xuemei, X. Liyun and M. Giovanni. “A heuristic approach to solve an industrial scalability problem”. In: *Procedia CIRP* 63 (2017), pp. 21–26 (cit. on p. 69).
- [97] C. Humphries, N. Prigent, C. Bidan and F. Majorczyk. “ELVIS: Extensible log visualization”. In: *Proceedings of the Tenth Workshop on Visualization for Cyber Security*. VizSec. ACM, 2013, pp. 9–16 (cit. on p. 100).
- [98] A. Inselberg. “The plane with parallel coordinates”. In: *The Visual Computer* 1.2 (1985), pp. 69–91 (cit. on p. 70).
- [99] *International Patent Classification (IPC)*. URL: <https://www.wipo.int/classifications/ipc/en/> (visited on 2019-10-17) (cit. on p. 30).
- [100] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller and J. Stasko. *Visualization Publication Dataset*. 2015. URL: <http://vispubdata.org/> (visited on 2019-10-17) (cit. on p. 60).
- [101] D. Jäckle, F. Fischer, T. Schreck and D. A. Keim. “Temporal MDS plots for analysis of multivariate data”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 141–150 (cit. on p. 100).
- [102] J. Jo, J. Huh, J. Park, B. Kim and J. Seo. “LiveGantt: Interactively Visualizing a Large Manufacturing Schedule”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2329–2338 (cit. on pp. 69, 71, 101).
- [103] O. Kaser and D. Lemire. “Tag-Cloud Drawing: Algorithms for Cloud Visualization”. In: *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization*. 2007 (cit. on p. 34).
- [104] D. A. Keim, M. Ankerst and H.-P. Kriegel. “Recursive pattern: A technique for visualizing very large amounts of data”. In: *Proceedings of the 6th Conference on Visualization*. InfoVis. IEEE, 1995, pp. 279–286. ISBN: 0-8186-7187-4 (cit. on p. 101).
- [105] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. “Optimization by simulated annealing”. In: *science* 220.4598 (1983), pp. 671–680 (cit. on p. 69).
- [106] B. Klopfer, S. Honiden, J.-P. Pater and W. Dangelmaier. “Decision making in adaptive manufacturing systems: Multi-objective scheduling and user interface”. In: *Proceedings of the IEEE Symposium On Computational Intelligence In Control And Automation*. 2011, pp. 123–130 (cit. on pp. 69, 101).

- [107] C. Knöpfle, J. Weidenhausen, L. Chauvigne and I. Stock. “Template based authoring for AR based service scenarios”. In: *Proceedings of IEEE VR*. 2005, pp. 237–240 (cit. on p. 71).
- [108] S. Koch, H. Bosch, M. Giereth and T. Ertl. “Iterative integration of visual insights during patent search and analysis”. In: *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 203–210. ISBN: 978-1-4244-5283-5 (cit. on p. 36).
- [109] K. Koh, B. Lee, B. Kim and J. Seo. “ManiWordle: Providing Flexible Control over Wordle”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1190–1197 (cit. on p. 34).
- [110] T. Kohonen. “The self-organizing map”. In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480 (cit. on p. 33).
- [111] M. Krstajic, E. Bertini and D. A. Keim. “CloudLines: Compact display of event episodes in multiple time-series”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2432–2439 (cit. on p. 100).
- [112] R. Krüger, D. Thom, M. Wörner, H. Bosch and T. Ertl. “TrajectoryLenses—A Set-based Filtering and Exploration Technique for Long-term Trajectory Data”. In: *Computer Graphics Forum* 32.3 (2013), pp. 451–460. ISSN: 1467-8659 (cit. on p. 23).
- [113] R. Krüger, T. Tremel and D. Thom. “VESPA 2.0: Data-Driven Behavior Models for Visual Analytics of Movement Sequences”. In: *Proceedings of the International Symposium on Big Data Visual Analytics*. IEEE, 2017, pp. 1–8. ISBN: 978-1-5386-0781-7 (cit. on p. 101).
- [114] J. B. Kruskal. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”. In: *Psychometrika* 29.1 (1964), pp. 1–27. ISSN: 0033-3123 (cit. on p. 14).
- [115] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. SAGE Publications, 1978. ISBN: 9780803909403 (cit. on p. 33).
- [116] I. Kuhlemann, M. Kleemann, P. Jauer, A. Schweikard and F. Ernst. “Towards X-ray free endovascular interventions – using HoloLens for on-line holographic visualisation”. In: *Healthcare Technology Letters* 4.5 (2017), pp. 184–187. ISSN: 2053-3713 (cit. on p. 70).
- [117] S. Kullback and R. A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. ISSN: 0003-4851 (cit. on p. 15).

- 
- [118] D. O. Kutz. “Examining the Evolution and Distribution of Patent Classifications”. In: *Proceedings of the IEEE International Conference on Information Visualization. IV '04*. IEEE, 2004, pp. 983–988. ISBN: 0-7695-2177-0 (cit. on p. 36).
- [119] J. Landstorfer, I. Herrmann, J.-E. Stange, M. Dörk and R. Wettach. “Weaving a carpet from log entries: A network security visualization built with co-creation”. In: *Proceedings of the 2014 IEEE Conference on Visual Analytics Science and Technology. VAST*. IEEE, 2014, pp. 73–82 (cit. on p. 100).
- [120] K. I. Lee and S. D. Noh. “Virtual manufacturing system—a test-bed of engineering activities”. In: *CIRP Annals – Manufacturing Technology* 46.1 (1997), pp. 347–350 (cit. on p. 70).
- [121] V. I. Levenshtein. “Binary codes capable of correcting deletions, insertions and reversals”. In: *Soviet Physics Doklady*. Vol. 10. 1966, p. 707 (cit. on p. 90).
- [122] R. S. Liggett. “Automated facilities layout: Past, present and future”. In: *Automation in construction* 9.2 (2000), pp. 197–215. ISSN: 0926-5805 (cit. on p. 69).
- [123] C.-F. Liu and P.-Y. Chiang. “Smart Glasses Based Intelligent Trainer for Factory New Recruits”. In: *Proceedings of MobileHCI*. 2018, pp. 395–399. ISBN: 978-1-4503-5941-2 (cit. on p. 70).
- [124] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu and S. Liu. “Towards Better Analysis of Deep Convolutional Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 91–100 (cit. on p. 69).
- [125] S. Liu, X. Wang, J. Chen, J. Zhu and B. Guo. “TopicPanorama: A full picture of relevant topics”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology. VAST*. IEEE, 2014, pp. 183–192 (cit. on p. 35).
- [126] Z. Liu, B. Jiang and J. Heer. “imMens: Real-time Visual Querying of Big Data”. In: *Computer Graphics Forum* 32.3pt4 (2013), pp. 421–430. ISSN: 1467-8659 (cit. on p. 75).
- [127] S. Lohmann, J. Ziegler and L. Tetzlaff. “Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration”. In: *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction. INTERACT '09*. Berlin, Heidelberg: Springer, 2009, pp. 392–404. ISBN: 978-3-642-03654-5 (cit. on p. 34).

- [128] D. Luo, J. Yang, M. Krstajic, W. Ribarsky and D. A. Keim. “Eventriver: Visually exploring text collections with temporal references”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.1 (2012), pp. 93–105 (cit. on p. 100).
- [129] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland and D. S. Ebert. “Forecasting hotspots—A predictive analytics approach”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.4 (2011), pp. 440–453 (cit. on p. 100).
- [130] A. Maier, T. Tack and O. Niggemann. “Visual Anomaly Detection in Production Plants”. In: *International Conference on Informatics in Control, Automation and Robotics*. 2012, pp. 67–75 (cit. on p. 69).
- [131] A. Malik, R. Maciejewski, N. Elmquist, Y. Jang, D. S. Ebert and W. Huang. “A correlative analysis process in a visual analytics environment”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2012, pp. 33–42 (cit. on p. 100).
- [132] C. D. Manning, P. Raghavan and H. Schütze. *Introduction to Information Retrieval*. Vol. 1. New York, NY, USA: Cambridge University Press, 2008 (cit. on pp. 11, 14).
- [133] K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stürzlinger and B. H. Thomas. *Immersive Analytics*. Vol. 11190. SpringerLink : Bücher. Cham, Switzerland: Springer, 2018. ISBN: 978-3-030-01387-5 (cit. on pp. 133, 134).
- [134] N. Menck, X. Yang, C. Weidig, P. Winkes, C. Lauer, H. Hagen, B. Hamann and J. C. Aurich. “Collaborative factory planning in virtual reality”. In: *Procedia CIRP* 3 (2012), pp. 317–322 (cit. on p. 70).
- [135] A. Meola, F. Cutolo, M. Carbone, F. Cagnazzo, M. Ferrari and V. Ferrari. “Augmented reality in neurosurgery: A systematic review”. In: *Neurosurgical Review* 40.4 (2017), pp. 537–548. ISSN: 1437-2320 (cit. on p. 70).
- [136] K. Misue, P. Eades, W. Lai and K. Sugiyama. “Layout adjustment and the mental map”. In: *Journal of Visual Languages & Computing* 6.2 (1995), pp. 183–210 (cit. on p. 41).
- [137] B. Mokbel, W. Lueks, A. Gisbrecht and B. Hammer. “Visualizing the quality of dimensionality reduction”. In: *Neurocomputing* 112 (2013), pp. 109–123 (cit. on p. 110).
- [138] *MongoDB*. URL: <https://www.mongodb.com/> (visited on 2019-10-17) (cit. on p. 37).
- [139] M. Monroe, R. Lan, H. Lee, C. Plaisant and B. Shneiderman. “Temporal event sequence simplification”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2227–2236 (cit. on p. 101).



- [140] D. Mourtzis, M. Doukas and D. Bernidaki. “Simulation in manufacturing: Review and challenges”. In: *Procedia CIRP* 25 (2014), pp. 213–229 (cit. on p. 71).
- [141] P. Muchiri and L. Pintelon. “Performance measurement using overall equipment effectiveness (OEE): Literature review and practical application discussion”. In: *International Journal of Production Research* 46.13 (2008), pp. 3517–3535 (cit. on p. 26).
- [142] T. S. Mujber, T. Szecsi and M. S. J. Hashmi. “Virtual reality applications in manufacturing process simulation”. In: *Journal of materials processing technology* 155 (2004), pp. 1834–1838 (cit. on p. 70).
- [143] National Institute of Standards and Technology. *Cyber-Physical Systems*. URL: <https://www.nist.gov/el/cyber-physical-systems> (visited on 2019-10-17) (cit. on p. 25).
- [144] J. Novak-Marcincin, J. Barna, M. Janak, L. Novakova-Marcincinova and J. Torok. “Visualization of intelligent assembling process by augmented reality tools application”. In: *4th IEEE International Symposium on Logistics and Industrial Informatics*. LINDI. 2012, pp. 33–36 (cit. on p. 70).
- [145] S. K. Ong, Y. Pang and A. Y. Nee. “Augmented reality aided assembly design and planning”. In: *CIRP Annals – Manufacturing Technology* 56.1 (2007), pp. 49–52 (cit. on p. 70).
- [146] Organisation for Economic Co-operation and Development. *Advanced Manufacturing Technology*. 2001. URL: <https://stats.oecd.org/glossary/detail.asp?ID=52> (visited on 2019-10-17) (cit. on p. 25).
- [147] H. M. Osman, M. E. Georgy and M. E. Ibrahim. “A hybrid CAD-based construction site layout planning system using genetic algorithms”. In: *Automation in construction* 12.6 (2003), pp. 749–764. ISSN: 0926-5805 (cit. on p. 69).
- [148] R. Palmarini, J. A. Erkoyuncu and R. Roy. “An innovative process to select Augmented Reality (AR) technology for maintenance”. In: *Procedia CIRP* 59 (2017), pp. 23–28 (cit. on p. 70).
- [149] F. V. Paulovich, L. G. Nonato, R. Minghim and H. Levkowitz. “Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.3 (2008), pp. 564–575 (cit. on p. 33).
- [150] F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim and L. G. Nonato. “Semantic Wordification of Document Collections”. In: *Computer Graphics Forum* 31.3 (2012), pp. 1145–1153. ISSN: 1467-8659 (cit. on p. 34).

- [151] J. Pearl. “Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning”. In: *Proceedings of the 7th Conference of the Cognitive Science Society, 1985* (1985), pp. 329–334 (cit. on p. 19).
- [152] K. Pearson. “On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. ISSN: 1941-5982 (cit. on p. 14).
- [153] K. Pentenrieder, C. Bade, F. Doil and P. Meier. “Augmented Reality-based factory planning—an application tailored to industrial needs”. In: *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. 2007, pp. 31–42 (cit. on p. 70).
- [154] P. Pirolli and S. K. Card. “The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis”. In: *Proceedings of the International Conference on Intelligence Analysis*. Vol. 5. 2005, pp. 2–4 (cit. on p. 8).
- [155] X. Qin and W. Lee. “Statistical causality analysis of infosec alert data”. In: *Proceedings of the International Workshop on Recent Advances in Intrusion Detection*. 2003, pp. 73–93 (cit. on p. 99).
- [156] X. Qin and W. Lee. “Attack plan recognition and prediction using causal networks”. In: *20th Annual of the Computer Security Applications Conference*. 2004, pp. 370–379 (cit. on p. 99).
- [157] P. E. Rauber, S. G. Fadel, A. X. Falcao and A. C. Telea. “Visualizing the Hidden Activity of Artificial Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 101–110 (cit. on p. 69).
- [158] P. Riffelmacher, S. Kluge, R. Kreuzhage, V. Hummel and E. Westkämper. “Learning factory for the manufacturing industry: Digital learning shell and a physical model factory-iTRAME for production engineering and improvement”. In: *Proceedings of the 20th International Conference on Computer-Aided Production Engineering*. 2007, pp. 120–131 (cit. on p. 84).
- [159] N. B. Robbins. *Introduction to Cycle Plots*. 2008. URL: [http://mail.perceptualedge.com/articles/guests/intro\\_to\\_cycle\\_plots.pdf](http://mail.perceptualedge.com/articles/guests/intro_to_cycle_plots.pdf) (cit. on p. 101).
- [160] M. W. Rohrer. “Seeing is Believing: The Importance of Visualization in Manufacturing Simulation”. In: *Proceedings of the 32nd Conference on Winter Simulation*. WSC '00. San Diego, CA, USA: Society for Computer Simulation International, 2000, pp. 1211–1216. ISBN: 0-7803-6582-8 (cit. on p. 69).

- 
- [161] F. J. Romero-Ramirez, R. Muñoz-Salinas and R. Medina-Carnicer. “Speeded up detection of squared fiducial markers”. In: *Image and Vision Computing* 76 (2018), pp. 38–47 (cit. on p. 145).
- [162] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis and D. A. Keim. “Knowledge Generation Model for Visual Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1604–1613 (cit. on p. 8).
- [163] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. North and D. A. Keim. “Visual interaction with dimensionality reduction: A structured literature analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 241–250 (cit. on p. 33).
- [164] D. Schmalstieg, A. Fuhrmann, G. Hesina, Z. Szalavári, L. M. Encarnação, M. Gervautz and W. Purgathofer. “The Studierstube Augmented Reality Project”. In: *Presence: Teleoperators and Virtual Environments* 11.1 (2002), pp. 33–54 (cit. on p. 70).
- [165] M. Schmitt, G. Meixner, D. Gorecky, M. Seissler and M. Loskyll. “Mobile interaction technologies in the factory of the future”. In: *IFAC Proceedings Volumes* 46.15 (2013), pp. 536–542 (cit. on p. 70).
- [166] M. Sedlmair, P. Isenberg, D. Baur, M. Mauerer, C. Pigorsch and A. Butz. “Cardiogram: Visual analytics for automotive engineers”. In: *Proceedings of the Conference on Human Factors in Computing Systems*. CHI, 2011, pp. 1727–1736 (cit. on p. 99).
- [167] C. Seifert, B. Kump, W. Kienreich, G. Granitzer and M. Granitzer. “On the Beauty and Usability of Tag Clouds”. In: *Proceedings of the IEEE International Conference on Information Visualisation*. IEEE, 2008, pp. 17–25 (cit. on p. 34).
- [168] J. P. Shewchuk. “Worker allocation in lean U-shaped production lines”. In: *International Journal of Production Research* 46.13 (2008), pp. 3485–3502 (cit. on p. 71).
- [169] L. Shi, Q. Liao, Y. He, R. Li, A. Striegel and Z. Su. “SAVE: Sensor anomaly visualization engine”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2011, pp. 201–210 (cit. on p. 99).
- [170] M. H. Shimabukuro, E. F. Flores, M. C. F. de Oliveira and H. Levkowitz. “Coordinated views to assist exploration of spatio-temporal data: A case study”. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization*. IEEE, 2004, pp. 107–117 (cit. on p. 101).

- [171] B. Shneiderman. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”. In: *Proceedings of the IEEE VIS Symposium on Visual Languages*. VL '96. IEEE, 1996, pp. 336–343. ISBN: 0-8186-7508-X (cit. on p. 39).
- [172] J. Stahnke, M. Dörk, B. Müller and A. Thom. “Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 629–638 (cit. on p. 35).
- [173] J. Stark, ed. *Product Lifecycle Management*. Decision Engineering. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-17439-6 (cit. on p. 26).
- [174] M. Stefaner. “Visual Tools for the Socio-Semantic Web”. Master thesis. University of Applied Sciences Potsdam, 2007 (cit. on p. 34).
- [175] M. Steiger, J. Bernard, S. Mittelstädt, H. Lücke-Tieke, D. Keim, T. May and J. Kohlhammer. “Visual analysis of time-series similarities for anomaly detection in sensor networks”. In: *Computer Graphics Forum* 33 (2014), pp. 401–410. ISSN: 1467-8659 (cit. on p. 99).
- [176] C. Sternitzke, A. Bartkowski and R. Schramm. “Visualizing patent statistics by means of social network analysis tools”. In: *World Patent Information* 30.2 (2008), pp. 115–131 (cit. on p. 36).
- [177] H. Strobel, M. Spicker, A. Stoffel, D. A. Keim and O. Deussen. “Rolled-out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives”. In: *Computer Graphics Forum* 31.3 (2012), pp. 1135–1144. ISSN: 1467-8659 (cit. on p. 34).
- [178] P. N. Suganthan. “Hierarchical overlapped SOM’s for pattern classification”. In: *IEEE Transactions on Neural Networks* 10.1 (1999), pp. 193–196 (cit. on p. 33).
- [179] N. R. Tague. “Seven basic quality tools”. In: *The Quality Toolbox*. Milwaukee, Wisconsin: American Society for Quality (2004) (cit. on p. 121).
- [180] D. Thom and T. Ertl. “TreeQueST: A Treemap-Based Query Sandbox for Microdocument Retrieval”. In: *Proceedings of the 49th Hawaii International Conference on System Sciences*. 2015, pp. 1714–1723 (cit. on p. 35).
- [181] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, 2005. ISBN: 9780769523231 (cit. on pp. 7, 8).
- [182] J. J. Thomas, P. J. Cowley, O. Kuchar, L. T. Nowell, J. Thompson and P. C. Wong. “Discovering knowledge through visual analysis”. In: *Journal of Universal Computer Science* 7.6 (2001), pp. 517–529 (cit. on p. 35).

- 
- [183] E. R. Tufte. *The Visual Display of Quantitative Information*. 1st edition. Cheshire, CT, USA: Graphics Press, 1983. ISBN: 978-0961392147 (cit. on p. 102).
- [184] E. R. Tufte. *Beautiful Evidence*. 1st edition. Graphics Press, 2006. ISBN: 0961392177 (cit. on p. 121).
- [185] A. E. Uva, M. Gattullo, V. M. Manghisi, D. Spagnulo, G. L. Cascella and M. Fiorentino. “Evaluating the effectiveness of spatial augmented reality in smart manufacturing: A solution for manual working stations”. In: *The International Journal of Advanced Manufacturing Technology* 94.1 (2018), pp. 509–521. ISSN: 1433-3015 (cit. on p. 71).
- [186] L. van der Maaten. “Accelerating t-SNE using tree-based algorithms”. In: *Journal of Machine Learning Research* 15.1 (2014), pp. 3221–3245 (cit. on p. 109).
- [187] L. van der Maaten and G. Hinton. “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605 (cit. on pp. 14, 15, 52, 70, 109).
- [188] L. van der Maaten, E. Postma and J. van den Herik. “Dimensionality reduction: A comparative review”. In: *Journal of Machine Learning Research* 10.1-41 (2009), pp. 66–71 (cit. on p. 33).
- [189] J. J. van Wijk and E. R. van Selow. “Cluster and Calendar based Visualization of Time Series Data”. In: *Proceedings of the IEEE Symposium on Information Visualization*. 1999, pp. 4–9 (cit. on pp. 101, 124).
- [190] *VAST Challenge 2014: Mini-Challenge 2*. 2014. URL: <http://www.vacommunity.org/VAST+Challenge+2014%3A+Mini-Challenge+2> (visited on 2019-10-17) (cit. on p. 19).
- [191] J. Wang and K. Mueller. “Visual Causality Analysis Made Practical”. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. VAST. IEEE, 2017, pp. 151–161 (cit. on p. 98).
- [192] X. Wang, S. K. Ong and A. Y. C. Nee. “A comprehensive survey of augmented reality assembly research”. In: *Advances in Manufacturing* 4.1 (2016), pp. 1–22. ISSN: 2195-3597 (cit. on p. 71).
- [193] M. Weber, M. Alexa and W. Müller. “Visualizing time-series on spirals”. In: *Proceedings of the IEEE Symposium on Information Visualization 2001*. Vol. 1. InfoVis. IEEE, 2001, pp. 7–14 (cit. on p. 101).
- [194] E. Westkämper. *Digitale Produktion*. Berlin: Springer Vieweg, 2013. ISBN: 9783642202582 (cit. on pp. 1, 2, 26).
- [195] *WIPO – World Intellectual Property Organization*. URL: <http://www.wipo.int> (visited on 2019-10-17) (cit. on pp. 29, 43).

- [196] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur and V. Crow. “Visualizing the Non-visual: Spatial Analysis and Interaction with Information from Text Documents”. In: *Proceedings of the IEEE International Symposium on Information Visualization*. IEEE, 1995, pp. 51–58 (cit. on p. 35).
- [197] S. Wold, K. Esbensen and P. Geladi. “Principal component analysis”. In: *Chemometrics and Intelligent Laboratory Systems 2.1* (1987), pp. 37–52 (cit. on p. 33).
- [198] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner and J. J. Thomas. “IN-SPIRE InfoVis 2004 Contest Entry”. In: *Proceedings of the IEEE International Symposium on Information Visualization*. IEEE, 2004 (cit. on p. 35).
- [199] K. Wongsuphasawat and D. Gotz. “Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2659–2668 (cit. on p. 101).
- [200] World Intellectual Property Organization. *About the International Patent Classification*. URL: <http://www.wipo.int/classifications/ipc/en/preface.html> (visited on 2019-10-17) (cit. on p. 32).
- [201] World Intellectual Property Organization. *Guide to the IPC (Version 2018)*. URL: [http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide\\_ipc.pdf](http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf) (visited on 2019-10-17) (cit. on p. 32).
- [202] M. Wörner. “Visual analytics for production and transportation systems”. PhD thesis. Stuttgart: University of Stuttgart, 2014 (cit. on pp. 68, 69, 84).
- [203] Y. Wu, T. Provan, F. Wei, S. Liu and K.-L. Ma. “Semantic-Preserving Word Clouds by Seam Carving”. In: *Computer Graphics Forum* 30.3 (2011), pp. 741–750. ISSN: 1467-8659 (cit. on p. 34).
- [204] P. Xie, J. H. Li, X. Ou, P. Liu and R. Levy. “Using Bayesian networks for cyber security analysis”. In: *Proceedings of the International Conference on Dependable Systems and Networks*. 2010, pp. 211–220 (cit. on p. 99).
- [205] P. Xu, N. Cao, H. Qu and J. Stasko. “Interactive visual co-cluster analysis of bipartite graphs”. In: *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2016, pp. 32–39 (cit. on p. 99).
- [206] P. Xu, H. Mei, L. Ren and W. Chen. “ViDX: Visual diagnostics of assembly line performance in smart factories”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 291–300 (cit. on pp. 69, 102, 143).

- [207] K.-P. Yee, K. Swearingen, K. Li and M. Hearst. “Faceted metadata for image search and browsing”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI. ACM, 2003, pp. 401–408 (cit. on p. 121).
- [208] L. Zhang, S. Klimov and Y. Zhu. “CancerVis: An interactive exploratory tool for cancer biomarker analysis”. In: *Proceedings of the International Conference on Bioinformatics and Biomedicine*. 2015, pp. 785–792 (cit. on p. 98).
- [209] Z. Zhang, K. T. McDonnell, E. Zadok and K. Mueller. “Visual correlation analysis of numerical and categorical data on the correlation map”. In: *IEEE Transactions on Visualization and Computer Graphics* 21.2 (2015), pp. 289–303 (cit. on p. 98).