



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Towards Making a Large Data Set Accessible for Multiple
Different Stakeholders“

verfasst von / submitted by

Dipl.-Ing. Raphael Sahann, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Doktor der Technischen Wissenschaften (Dr. techn.)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

A 786 880

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Dr.-Studium der technischen Wissenschaften
(Dissertationsgebiet: Informatik)

Betreut von / Supervisor:

Univ. Prof. Torsten Möller, PhD

Acknowledgments

First and foremost, I want to express my sincere gratitude to my supervisor Torsten Möller. You always pushed me to give my absolute best and challenged me to strive for more remarkable achievements. But you also knew to pull the brakes when the situation required it and respected my boundaries. I enjoyed working with you, and I am in awe every time I reflect on how much I learned from you.

Collaborating with Johanna Schmidt showed me that Aristotle's metaphysical quote on emergence applies to research projects as well. The result of our collaborations was indeed *greater than the sum of its parts*. Thank you for the patience and joy you brought to our projects!

I am grateful to Axel Sonntag and Claudia Plant for collaborating with me on different aspects of my research. They are both thorough researchers whose comprehensive advice and thoughtful support I appreciate.

My most heartfelt appreciation goes to Stephan Prechtel, who truly believed in my visions from our first meeting, and helped me with all his capabilities to achieve them. He started to support me on a professional level but grew to be one of my most valued sources of personal energy, perseverance, and a dear friend.

I want to thank my colleagues in research, Christoph Kralj, Christian Knoll, Bernhard Fröhler, Laura Koesten, Thomas Torsney-Weir, and Michael Sedlmair, for their constant encouragement and helpfulness. Your feedback, ideas, and our sometimes hour-long discussions helped me out of situations where I seemed stuck and lost.

Particular expressions of appreciation are dedicated to Martin Polaschek, who saw potential in my capabilities at a very early point of my studies. He initially recruited me for academia, and I am sure that my path would have looked different without the doors he opened for me.

I enjoyed working together with many different and uniquely great students. I hope they learned at least half as much from me as I learned from them. I want to specially mention Ivana Gajic, Julian Gruber, Christoph Pressler, and Katharina Wünsche, who actively engaged in projects with me and were great collaborators.

I am very grateful to Anne Marie Faisst and the whole, ever-changing team of the research group Visualization and Data Analysis. All of you made long hours in the office ever-so-slightly more bearable, which I mean as the highest possible praise! My spontaneous appearance as basically an outsider in the Coordination of Student Services team never felt like that. On the contrary, I was immediately a part of the welcoming team, which I grew very fond of despite my comparatively minor role. A special nod of appreciation goes to Carmen Fuchs, who was always there to be weird with me.

Finally, I want to thank the anonymous reviewers and editors for their helpful comments, which provided valuable assistance in improving my work.

Acknowledgments

I dedicate this work to my wife Sigrid, our breathtaking children, and my mother, Gudrun.

Baden, June 2021

Abstract

This dissertation is a compilation of publications and publication manuscripts that seek to improve the accessibility of large data sets for multiple different stakeholders. Therefore, it focuses on three essential aspects (i) data structures and abstraction, (ii) perception of data visualizations and their interactivity, and (iii) user experience. The thesis explores these facets using the extensive student data set from the University of Vienna.

The first publication defines the abstract concept of a *study path* which represents the reported mental model of how the progress in a curriculum is perceived. This concept is then used to calculate a distance metric, making it possible to numerically express the difference between study paths. We show that this metric can then be used for clustering and predicting study paths. This abstraction makes the data approachable and intuitive to use.

The second publication manuscript focuses on human perception for interpreting data distributions in histograms. It evaluates the error when judging distribution shapes in terms of the maximum number of shown bars in a histogram. We directly compare these findings to commonly used recommendations for choosing histogram binnings. Our work finds that notably fewer bins than common binning methods recommend achieve comparable perceptual errors when judging distributions but are easier to comprehend for the viewer.

The third publication presents a novel intuitive brushing technique for parallel coordinate plots. It uses established concepts for highlighting lines in parallel coordinate plots and makes them easily accessible by reducing the interaction to a simple click-and-drag mouse gesture. The intuitive usage approach of this brushing method reduces the mental load when interacting with parallel coordinate plots and, therefore, makes them easier to grasp.

Finally, the fourth publication manuscript describes the implementation and evaluation of a user interface for semester planning. This four-year user-centered design process shows the value of a well-integrated, easy-to-use user interface. The manuscript also abstracts the semester planning process into a broadly applicable abstract planning task that provides a guideline for future planning tools. It concludes with an analysis of different evaluation approaches for the design process and how a combination of different methods can benefit from another.

These four publications and manuscripts deal with different aspects, driven by different stakeholders that all need access to the same large data set. Each individual approach and any combination of presented approaches help make large quantities of data more accessible to expert and non-expert users equally.

Kurzfassung

Diese Dissertation ist eine Sammlung von Publikationen und Manuskripten welche die Zugänglichkeit großer Datensets für mehrere unterschiedliche Interessenträger zu verbessern. Um dies zu erreichen liegt der Fokus auf den drei wesentlichen Aspekten (i) Datenstrukturierung und Datenabstraktion, (ii) Wahrnehmung von Datenvisualisierungen und deren Interaktionsmöglichkeiten, und (iii) dem Nutzererlebnis. Um dies zu ergründen wird der umfassende Datensatz der Studentendaten der Universität Wien verwendet.

Die erste Publikation definiert das abstrakte Konzept des *Studienpfads*, welches dem mentalen Modell wie Fortschritt im Studium empfunden wird entspricht. Auf Basis dieses Modells wird eine Distanzmetrik berechnet, die den Unterschied zwischen zwei Studienpfaden numerisch ausdrücken kann. Dadurch können Studienpfade gruppiert und auch für Vorhersagen verwendet werden. Diese Abstraktion macht die Studiendaten den Interessenträgern leichter zugänglich und intuitiv nutzbar.

Die zweite Arbeit fokussiert sich auf die menschliche Perzeption bei der Interpretation von Datenverteilungen in Histogrammen. Sie evaluiert den perzeptuellen Fehler bei der Identifikation von Verteilungskurven im Verhältnis zur maximalen Anzahl der im Histogramm gezeigten Balken. Diese Observation wird direkt mit anderen in der Literatur häufig angewandten Methodiken zur Bestimmung der Klassenhäufigkeit verglichen. Unserer Arbeit zeigt, dass mit erheblich weniger Klassen bereits vergleichbar niedrige perzeptuelle Fehler erzielt werden können, und diese außerdem für den Betrachter leichter verständlich sind.

Die dritte Publikation stellt eine neuartige, intuitive Markiermethode für Linien in Parallelen Koordinaten Diagrammen vor. Dafür werden bekannte Konzepte der Markierung von Parallelen Koordinaten verwendet, welche mittels Reduktion der benötigten Maus-Aktionen zu einer einzelnen Klick- und Ziehinteraktion zusammengefasst werden. Die intuitive Verwendbarkeit der neuen Methode reduziert die mentale Anstrengung beim Interagieren mit Parallelen Koordinaten Diagrammen und macht diese dadurch verständlicher.

Das vierte und letzte Manuskript beschreibt die Implementation und Evaluation einer nutzerfreundlichen Bedienoberfläche für die Semesterplanung von Studierenden. Anhand einer vierjährigen Designstudie wird der Mehrwert, den ein gut integriertes, leicht zu verwendendes Interface bringt, beleuchtet. In dem Manuskript wird außerdem der Prozess der Semesterplanung soweit abstrahiert, dass dieser auf eine große Bandbreite allgemeiner Planungsprozesse angewendet werden kann. Dieser abstrakte Planprozess kann als Richtlinie für die Erstellung zukünftiger Planungstools fungieren. Abschließend beleuchtet eine Analyse unterschiedliche Evaluationsmethoden im Designprozess und zeigt, wie eine Kombination dieser Methoden deren Ergebnisse noch verbessern kann.

Kurzfassung

Diese vier Publikationen und Manuskripte behandeln unterschiedliche Aspekte des Zugriffs auf den gleichen darunterliegenden Datensatz aus der Perspektive mehrerer Interessenträger. Jede der gezeigten Methoden für sich, aber auch deren Kombination, helfen dabei große Datenmengen für Experten und Nicht-Experten gleichermaßen zugänglich zu machen.

Contents

Acknowledgments	i
Abstract	iii
Kurzfassung	v
1 Preamble	1
1.1 Abstraction and Structure	2
1.2 Perception and Interaction	3
1.3 User Experience	3
2 A Distance Metric for Sets of Events	5
2.1 Introduction	6
2.2 Related Work	6
2.3 A Distance Metric For Sets Of Events	7
2.4 Experiments	9
2.5 User Study	12
2.6 Discussion	13
2.7 Conclusion	15
3 Histogram binning revisited with a focus on human perception (accepted)	19
3.1 Introduction	20
3.2 Related Work	21
3.3 Quantitative User Study	22
3.4 Results	24
3.5 Impact and Discussion	26
3.6 Conclusion	28
4 Selective Angular Brushing of Parallel Coordinate Plots	31
4.1 Motivation	32
4.2 Related Work	32
4.3 Selection Design	33
4.4 Implementation	34
4.5 Results	35
4.6 Conclusion	35
5 Designing a Semester Planner for Students (manuscript)	37
5.1 Introduction	38

Contents

5.2	Related Work	39
5.3	Methodology	40
5.4	Planning Tool	41
5.5	Evaluation Results	45
5.6	Discussion	47
5.7	Conclusion	48
6	Discussion	51
6.1	Future Work	55
6.2	Conclusion	55
	Bibliography	57

1 Preamble

Universities already collect a substantial amount of data each semester. Even focussing on student data alone, disregarding all personnel, building, facility, and research data, a near-insurmountable data collection remains. The University of Vienna registered 15 faculties, 178 curricula, and 88,756 students in the most recent count from the winter semester of 2019 [oV19]. Each relevant action of each student is stored in a central database. Included are course and exam registrations, received grades, graduations, new study registrations, dropouts, curricula changes, and submitted theses. More than a million new data points are added to this database each year. These vast quantities of data collected and stored can easily hide essential factors or be overwhelming, even to expert users.

In order to complete their studies, students have to complete a set amount of credits, which are defined by the *European Credit Transfer System (ECTS)*. 180 ECTS are needed for a typical Bachelor's program, 120 ECTS for a Master's program. A positive grade in either a course, an exam, or a thesis is required to gain ECTS. Completing courses is, therefore, the most critical aspect of any curriculum, which is also the focal point of this work.

This *core* student data set is an event-based data set which consists of the following information:

- course/exam information (including ECTS)
- course/exam registrations
- course/exam grades
- personal student data
- instructors
- curricula texts
- date and time of each event

These data are the basis of a multitude of workflows all across the university. Some examples of stakeholders and their workflows that use data from this set include:

- the presidents' office
e.g., key performance indicators, global budget definition, issuing of new curricula, semester and year summaries, faculty evaluations

1 Preamble

- finance and controlling
e.g., calculating the budget for individual curricula, paying teachers, monitoring spent budget and evaluating its effectiveness
- administrative planning
e.g., planning of courses and exams, assigning teachers to courses, creating new, and closing unneeded courses
- instructors
e.g., planning individual courses, planning exams, booking rooms, grading students
- students
e.g., planning which courses to take next, printing their academic records, managing scholarships

Even though this list of tasks is nowhere near exhaustive, it shows that this data is both increasingly large and crucial to multiple different stakeholders. Each stakeholder and each task need different parts, aggregations, views, methods, and approaches to access those data properly. The raw data contains all information necessary to accomplish those tasks but is not helpful to most stakeholders since they typically are not data analysts.

My Master's thesis and the resulting publication *OCP - Operational Curricular Planning: A Visual Decision Support System for Planning Teaching Resources at Universities* [SM18] aims to help administrative planners to offer suitable courses for the upcoming semester. Comparative visualization of past and planned semesters and the accessibility of task-specific aggregated data support their decision-making process. Based on this knowledge, this thesis seeks to extend the topic of accessibility of large data sets by answering the following research questions:

- RQ1: How do data representations have to differ in order to accomplish different tasks from different stakeholders efficiently??
- RQ2: How can perception and interaction be used to make data visualizations more comprehensible and easier to use?
- RQ3: How does the process of creating a research prototype differ from the process of creating a ready-to-use-tool?

1.1 Abstraction and Structure

A crucial factor that makes raw data hard to understand for non-experts is the lack of clarity. Raw data is commonly stored in tabular form, often distributed across multiple databases. This very abstract representation does not fit with the experts' understanding of the data. For example, a workshop we held in early 2019, which included members of all stakeholders listed above, showed that the users have an explicit mental image when thinking about aspects of their work. Thus, they have difficulty making sense of existing data if it does not fit with that mental image.

In our work *A Distance Metric for Sets of Events*, shown in chapter 2, we synthesize *study paths* from the existing data, which represent the mental image of the surveyed stakeholders. To make these study paths more accessible, we present a distance metric that makes these structures comparable to each other. This metric can cluster similar study paths, and help experts understand the students' actual path to complete their studies. The use of this metric for predictive analysis also shows potential for administrative planning of semesters.

The planning tool we present in chapter 5 *Designing a Semester Planner for Students* combines different aspects of the data, which were initially available from different sources. This combination closely represents the task of planning as identified in an initial study of the planning process. Furthermore, familiarization with the abstracted data enhances its accessibility. We achieve this by showing the aggregated data in the commonly-known patterns of a shopping cart and a calendar. These patterns prove to be intuitively usable for users of the tool.

1.2 Perception and Interaction

While choosing a visual encoding that respects both the underlying data and task appropriately is essential, it is also necessary that the user correctly perceives what is shown. Working with student data and multiple stakeholders, we frequently encountered the need to use histograms as means of understanding data. The most common question we faced was how many bins to use in a specific situation. Many different methods for choosing the correct amount of bins in a histogram have been published in the past. They all focus on the characteristics of the underlying data. In our work on *histogram binning* (see chapter 3), we show that reducing the number of bins up to a certain amount results in equally few errors as high bin counts when judging distribution shapes. This user-centered approach suggests a perception-based amount of bins which makes histograms easier to understand for non-expert users.

The analysis of how students progress in their studies led us to use parallel coordinate plots. We noticed that this multi-dimensional data visualization method is not very approachable for novice users. Most insights generated by this technique rely on different interaction methods to sort, filter, or highlight parts of the chart. A multitude of highlighting techniques allow all sorts of individual selections of lines in parallel coordinate plots. In *Selective Angular Brushing of Parallel Coordinate Plots* shown in chapter 4, we renew an existing brushing method by making it more intuitive to use, therefore more user-friendly in the process. This highlighting method proved helpful when visualizing study paths in parallel coordinate plots.

1.3 User Experience

The user interface and, therefore, the user experience is vital for accessibility and acceptance of any visualization tool. A well-designed tool that provides a good user experience will motivate users to come back and use it. On the other hand, a tool that only shows

1 Preamble

the required data without considering the user experience makes users question if they want to continue using it each time they are frustrated with one of its parts.

Students at our faculty often struggle to finish their Bachelor's degree in the recommended number of semesters. A short, informal survey showed that a notable amount of lost time could be attributed to issues while planning which courses to take. Unawareness of additional options, general misconceptions, and missing information while establishing a semester plan are significant contributors to prolonged study times. To tackle this issue, we created a tool that helps students plan their next semester. Chapter 5 *Designing a Semester Planner for Students* describes the abstraction and creation process that ultimately tries to mimic the existing planning process closely. The tool assists the process, enhances it, and avoids compromising flexibility in planning. This extensive user-centered design process resulted in an intuitive user interface that combines functionality and data to create a positive user experience, and is currently used by more than 11,000 students every semester.

2 A Distance Metric for Sets of Events

Synopsis

© 2020 IEEE. Reprinted, with permission, from Raphael SAHANN, Claudia PLANT, and Torsten MÖLLER. "*A Distance Metric for Sets of Events.*", 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2020 [SPM20].

This paper was presented at the IEEE Conference on Data Science and Advanced Algorithms (DSAA) on October 9, 2020, nominally hosted in Sydney, but finally occurred as an online conference.

Raphael Sahann did all research and writing of the paper. Torsten Möller assisted the formalization of some of the formulas in Appendix A. Both Claudia Plant and Torsten Möller gave feedback on the writing and pointed Raphael towards helpful resources regarding the related work and the conceptualization of some experiments.

A Distance Metric for Sets of Events

Raphael Sahann
Faculty of Computer Science
University of Vienna
Vienna, Austria
raphael.sahann@univie.ac.at

Claudia Plant
Data Science @ Uni Vienna
Faculty of Computer Science
University of Vienna
Vienna, Austria
claudia.plant@univie.ac.at

Torsten Möller
Data Science @ Uni Vienna
Faculty of Computer Science
University of Vienna
Vienna, Austria
torsten.moeller@univie.ac.at

Abstract—In this work, we introduce a novel distance metric that describes the distance between sets of events, where events in the most common form are actions that happen at a given time. More generally, an event can be any object that is in an ordered relation to other objects. In our case, an event is a course taken by a student that happens during a specific semester. Calculating the distance uses the difference between the positional relations of all individual events in the set. For this, we do not use the absolute position of events but instead use the sum of differences of the relations *before*, *concurrent*, and *after* to express distance. We describe our metric algorithmically and evaluate it formally as well as exemplarily on an existing data set of student exams. We also show that the results of the metric are intuitive to interpret for humans by comparing them to the results of a user study that we ran.

This metric can be applied to a range of problems that rely on the positional relation of events by removing the dependency of timestamps for events and replacing them with a set of ordered identifiers. We show a specific application of the metric by tackling the problem of clustering and predicting study paths from university students.

Index Terms—distance metric, event, student data, user study, clustering

I. INTRODUCTION

The problem of clustering sets of events initially emerged after the introduction of a new curriculum at our faculty. The immediate questions that arose were whether the new curriculum works, how it compares to the old curriculum, and the difference in student performance between the two? Comparing student grades is one way to tackle some of this, but it does not show the actual picture. We wanted to understand what factors make students successful or let them fail and whether the new curriculum affects that. Hence we needed a way to cluster students into meaningful groups, not only based on their grades but instead focusing on the way they approach their studies.

By looking at courses and their semesters as sets of events that are time-dependent to each other, we needed a way to cluster them. Therefore, we introduce the notion of *study paths*, which represents all courses a student takes of the same curriculum as sets of events. The idea is that if we cluster similar *study paths*, we can easily find clusters of students who finish their studies faster than others, students who work while studying, or even students that are more likely to drop out.

The most straightforward approach would be to treat *study paths* as time series since we know the exact date and time of each exam taken. This approach does not capture the classes' actual nature since an exam usually happens on the last date of a class in the semester, but the actual class lasted the whole semester. It also poses the issue of date-inconsistencies between semesters, which is problematic when comparing two different paths. An exam that might have happened on a Monday in one semester could be on a Wednesday in the next year, and vice versa. Therefore comparing the time series of two students who started their studies one year apart might show differences, even though they took the same classes in the same respective semesters, just because the exam dates were different. We needed a possibility to specify that two events are *concurrent* — in this case, two courses happening in the same semester — and the structure of a time series could not provide that.

Therefore, we introduce our distance metric, which distinguishes between three positional relations, namely *before*, *concurrent*, and *after*. Using these relations, the metric abstracts from absolute times or specific time distances, and can detect distinct structural differences. Figure 1 shows the idea of how our metric calculates the distance, and section III describes the underlying method and all details of our method.

The remainder of the paper is organized as follows:

- 1) in section III we define a novel metric to measure the distance between sets of events and evaluate it
- 2) we apply our metric for clustering and prediction of study paths in section IV
- 3) we ran a user study to evaluate the performance and intuitiveness of our metric, shown in section V

Section II reviews the related work and section VI discusses different approaches we considered, the limitations of the presented metric and also highlights future work. Finally, section VII concludes the paper.

II. RELATED WORK

Our initial approach to tackle similarity of study paths was using a representation of our data set as time series. Fu [1], as well as Ding et al. [2], show a plethora of approaches and problems in data mining tackled using some form of time series representation. Dynamic Time Warping [3] is one of the most common approaches when dealing with time series,

2.3 A Distance Metric For Sets Of Events

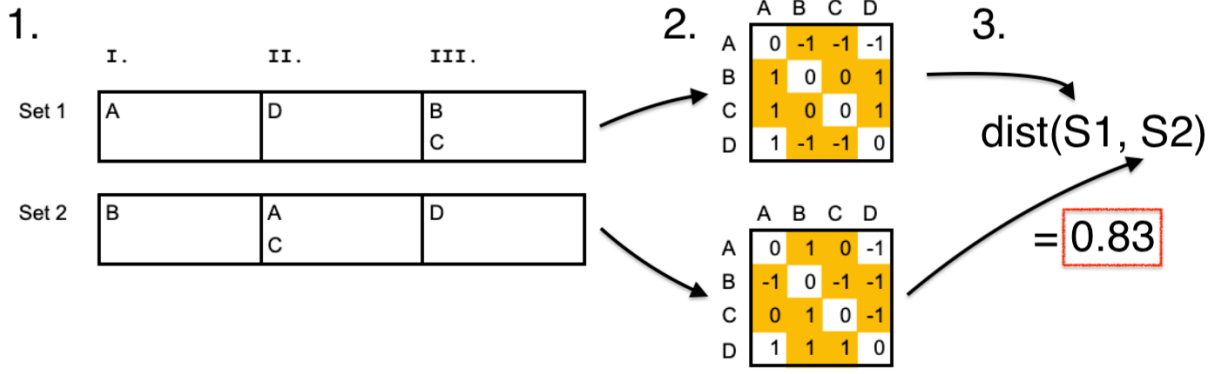


Fig. 1. Two sets of events to compare are shown in step 1. Each event consists of an object (A, B, C or D) and a time interval (I, II or III). Using the order information from the intervals a matrix of positional relations is constructed for each set (step 2). Finally, in step 3, the distance is calculated by evaluating the differences between the two matrices. This distance measure quantifies the distance between two sets of events. The details of this process are described in section III.

as it can handle different length inputs and is more robust than Euclidean distances [4], [5]. The problem we faced when working with time series was that they did not account for date changes over multiple semesters, namely a course that happened on one day might be scheduled on another day of the week the following semester. So two students who attended the same classes in different semesters were treated as dissimilar because the courses switched weekdays.

The works on sequential patterns mining [6] and temporal association rules [7]–[9] could further be used to gather data and dynamically determine the window sizes for time intervals in our metric. Our proposed metric could then be used to detect similarities and cluster results from these sets defined by temporal association rules.

In order to understand the structures better, we started to look for solutions with simplified versions of study paths using a shorthand string representation, which we introduce in section III-A. Since the string representation worked well for creating a mental model of a study path, we considered different string metrics to evaluate our study paths. Cohen et al. [10], [11] describe different approaches and compare them.

We also considered different graph-based metrics [12], but graphs always consist of single nodes and do not facilitate the inclusion of concurrency. In order to introduce concurrency into the graph representation, we created artificial bundling nodes before and after each set, which we linked to all nodes within that set. These nodes affected the size of the compared structure and skewed the results increasingly as the number of sets rose. We finally abandoned this approach because it was not a feasible solution, neither structurally nor computationally.

A. Metrics used for comparison

As a reference to compare our metric against after testing it, we used the following three well-known distances:

1) Energy distance [13], [14]

The Energy Distance is a statistical difference metric that

measures the distance between distributions of random vectors.

2) Earth-Mover's distance [15]

It is sometimes also known as the Wasserstein metric and is also a measure of the distance between two probability distributions, and is informally described as the cost of turning one pile of dirt over a set region into another pile, where cost is equal the amount of dirt, times the distance moved.

3) Damerau-Levenshtein distance [16], [17]

The Damerau-Levenshtein distance is a string metric measuring the edit distance between two strings. In addition to the traditional Levenshtein distance, it also includes the transposition of two adjacent characters as an available edit option.

We chose both the Energy distance and the Earth-Mover's distance because the courses in the data set associated with our initial problem also have a certain probability of appearing in a specific semester, which is individual for each student.

We also included the Damerau-Levenshtein distance because it matched our approach when dealing with string representations of study paths the closest. It is also a widely used standard in the field of text analysis.

When testing Dynamic Time Warping and Euclidean distance, we found that they produced worse results than the other considered metrics, which is why we left them out of the comparison.

III. A DISTANCE METRIC FOR SETS OF EVENTS

Figure 1 shows the core concept of our metric, which takes two sets of events, builds a matrix of relational differences for each using -1, 0, and 1 to express *before*, *concurrent* and *after*, and calculates the distance using the differences between two of these matrices.

In this section, we describe the method we developed in detail, its properties, as well as a shorthand notation for sets of events, which we use over the remainder of the paper to facilitate legibility and understandability.

2 A Distance Metric for Sets of Events

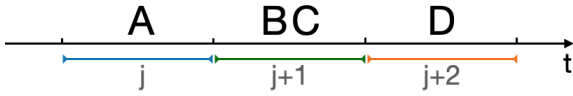


Fig. 2. A timeline view of set S_1 . Each interval $j, j+1, j+2$ on the timeline represents a time-interval during which all occurring events are considered concurrent.

A. Notation

We introduce a shorthand notation of sets of events that will be used over the remainder of the paper to facilitate the reading of examples. A capital letter (e.g., A, B) denotes an event in the set. Dashes separate objects belonging to time intervals, and the reading direction is from left to right, where the leftmost group of objects is the earliest interval, and the rightmost group is the latest. The placement of an item within a group does not matter, but we generally write objects in alphabetical order to improve legibility. Table I provides some examples.

TABLE I
EXAMPLES OF THE SET NOTATION USED IN THIS WORK

AB	a group of two events A and B which happen simultaneously (both in interval one)
A-B-C	three groups, each containing one event, A happens before B and C, B happens after A but before C and C happens after A and B
A-BC	two groups, group 1 containing only event A and group 2 containing B and C, where B and C happen simultaneously and both happen after A
A-CB	identical to the previous row

Using this notation, set 1 from figure 1 is written as A-D-BC and set 2 is B-AC-D.

B. Distance Metric

This metric builds on three relations between events: *before*, *concurrent*, and *after*. The idea of this metric is that events in the past can influence the outcome of a later event, but not vice versa. As an example: a class already taken by a student builds knowledge, which is then present in all future courses. The notion of concurrent in terms of courses applies to two courses taken in the same semester — these courses incrementally build knowledge over the whole timespan, from which the student can benefit. Therefore, in terms of education, relations are explained as follows: *before* denotes already acquired experience, *concurrent* is simultaneously acquiring knowledge, and *after* is still unknown.

Using these relations, we calculate a distance score between zero and one, where zero represents complete similarity, and one is entirely different.

We now illustrate calculating the score by the example of the distance between set S_1 A-BC-D (see fig. 2) and set S_2 D-BC-E (see fig. 3). The first step is to separate all events into those appearing in both sets — the overlap — and those contained in only one. As seen in figure 4, in this example, the overlap is B, C, D, and the non-overlapping members are A and E. Events in the non-overlapping set contribute one

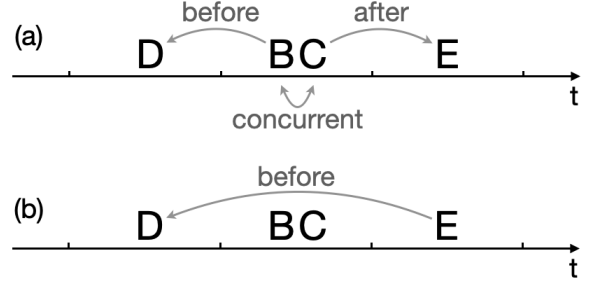


Fig. 3. A timeline view of set S_2 . The arrows represent the positional relations that are considered in our approach (a). The relations *before* and *after* do not only apply to neighbouring intervals (a), but also across multiple (b).

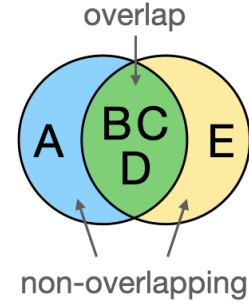


Fig. 4. The blue circle shows the events of set S_1 , the yellow circle shows set S_2 and in the green area in the middle the overlap of events that appear in both sets is shown.

to the overall sum because they only appear in one set, and therefore, the distance is maximal.

Applying the metric to sets of different sizes and, therefore, sequences of different lengths, works for all combinations, because we use this separation into overlapping and non-overlapping members. When comparing a short and a long sequence, many events will only be present in the long sequence. Each of these events adds the maximal distance of one to the overall sum, which is then weighted by the number of events in total. Since all events are weighted equally, overlapping events occur twice, while non-overlapping events only occur once. Thus the resulting distance is not disproportionately skewed by size differences.

Next, we create a matrix of relational positions for each set (see fig. 5). A row represents an event. The relations we use are -1 for *before*, 0 for *concurrent*, and 1 for *after*. The first row in figure 5, therefore, should be read as follows: B happens concurrently with B, B happens concurrently with C, and B happens before D. To reduce the computing load, we only look at the upper triangular matrices since the resulting matrices are skew-symmetric, and the diagonal is always zero. In the next step, we compare each position in the first matrix with the matching position from the second matrix and create a sum of all differences by adding one each time the two entries are not identical (highlighted positions in fig. 5). Finally, we normalize the resulting sum by dividing it by the number of entries compared.

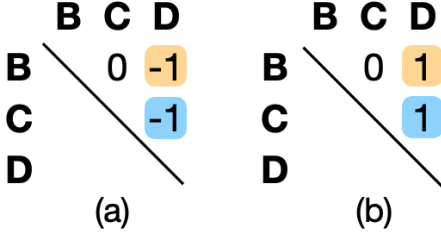


Fig. 5. Matrices of positional differences from set 1 A-BC-D (a) and set 2 D-BC-E (b). A and E are omitted because they are in the non-overlap. The highlighted entries are pairs of different entries in the two matrices, and each pair contributes one to the sum of positional differences.

The sum of positional distances, therefore, is

$$\text{pos_dist} = \frac{1}{n^2 - n} \sum_{i=1}^n \sum_{j=1}^n \text{sgn}(|a_{i,j} - b_{i,j}|) \quad (1)$$

where $a_{i,j}$ and $b_{i,j}$ are the positions in the matrices from set 1 and set 2, respectively, and n is the number of entries.

In this example, the first entry in the first row is identical, but the second entry in the first row and the entry in the second row are different. The resulting normalized sum is $2/3$.

The last step combines the results from the overlapping, and the non-overlapping parts by weighing them by the number of events they represent. The overlap represents three events from set 1 and three events from set 2, which results in a weight of $3 \times 2 = 6$. The non-overlap are two events, which gives a sum of $\frac{2}{3} \times 6 + 2 \times 1 = 6$. This sum normalized by the number of events in total (8) results in $\frac{6}{8} = \frac{3}{4}$ as the final distance between the set A-BC-D and D-BC-E.

Note that we normalized by the number of events in total, and not the sum of events in the union of the two sets. If we instead use the number of events in the union, we skew the results towards the events in the non-overlap. As an example the difference of A-B vs. A-X results in $\frac{1}{2}$ using all events, but $\frac{1}{3}$ when using only the union of the sets.

This calculation fulfills the axioms of *identity of indiscernibles*, *symmetry* and the *triangle inequality*, as well as the condition of *non-negativity*, which thus qualifies it as metric. We provide the full mathematical description of the metric and the proof of all axioms in appendix A.

IV. EXPERIMENTS

This section describes the data as well as the different tasks we tested using our metric.

Find the code and an anonymous version of the data used in this section in the supplementary GitHub archive <https://github.com/VDA-univie/set-of-events-distance-metric>. Instructions on how to generate the results and the images are in appendix section B, which can also be found in the GitHub archive.

A. Data – Exam Corpus

We based our approach on the corpus of exam grade data from the University of Vienna. The exam corpus is initially

an event-based data set where new entries are added whenever a student gets a grade. A row, therefore, contains information about an exam, the student who took it, the course to which the exam belongs, the date, the semester, and the grade.

We aggregate all exams for each student and curriculum and divide them into sets of individual semesters to use the data. The exam is happening on an individual date, but the course to which that exam belongs usually lasts the whole semester. Therefore, we specify an event to be the combination of the course and the semester the student took it. Since the semesters are in temporal relation to each other, we end up with a set consisting of courses ordered into time intervals (semesters). We call the resulting structure the *study path* of a student, a set of events that contains the information which courses this student took and in which order. A student can have multiple distinct *study paths* (e.g., Bachelors and Masters).

B. Clustering

TABLE II
COMPARISON OF CLUSTERING PERFORMANCE SCORES USING DBSCAN TO CLUSTER THE RESULTS OF THREE VARIATIONS OF THE TEST SET WITH 100 PATHS. THE FIRST SECTION WAS GENERATED BY REARRANGING COURSES WITH THE FOLLOWING PROBABILITY: 70% CHANCE THAT COURSES STAYED IN THE SAME SEMESTER, 20% CHANCE TO MOVE ONE SEMESTER UP OR DOWN AND 10% CHANCE TO MOVE TWO SEMESTERS UP OR DOWN. EM STANDS FOR EARTH MOVERS DISTANCE, AND DLEV IS THE DAMERAU-LEVENSHTEIN DISTANCE.

parameters	distance metric	normalized mutual info	cluster	noise
70-20-10	our metric	1	4	0
	Energy	0.73399	6	1
	EM	0.66943	9	4
	DLev	0.46348	10	53
50-30-20	our metric	1	4	0
	Energy	0.72334	7	1
	EM	0.56801	14	14
	DLev	0.07689	1	98
30-40-30	our metric	1	4	0
	Energy	0.70399	5	3
	EM	0.53881	14	19
	DLev	0.00004	0	100

To test the clustering robustness of our metric, we use study paths from the exam corpus data set. Since there is no ground truth available, we synthesize a test set from the existing data. We chose four study paths for this task, that start in the same semester, are from the same curriculum and are all finished successfully. The only difference between these paths is their grade point average (GPA), being 4.0, 3.0, 2.0, and 1.0. From each of these paths, we build variations by changing the position of the courses. In the first iteration, each course had a 50% chance to stay in the same semester, a 30% chance to move one semester randomly up or down, and a 20% chance to move two semesters up or down. Using this setup, we then generated 24 new paths from each study path, which all combined — including the original study paths — gave us a labeled test set of 100 paths. Figure 6 shows the resulting distance matrix using our distance metric for this

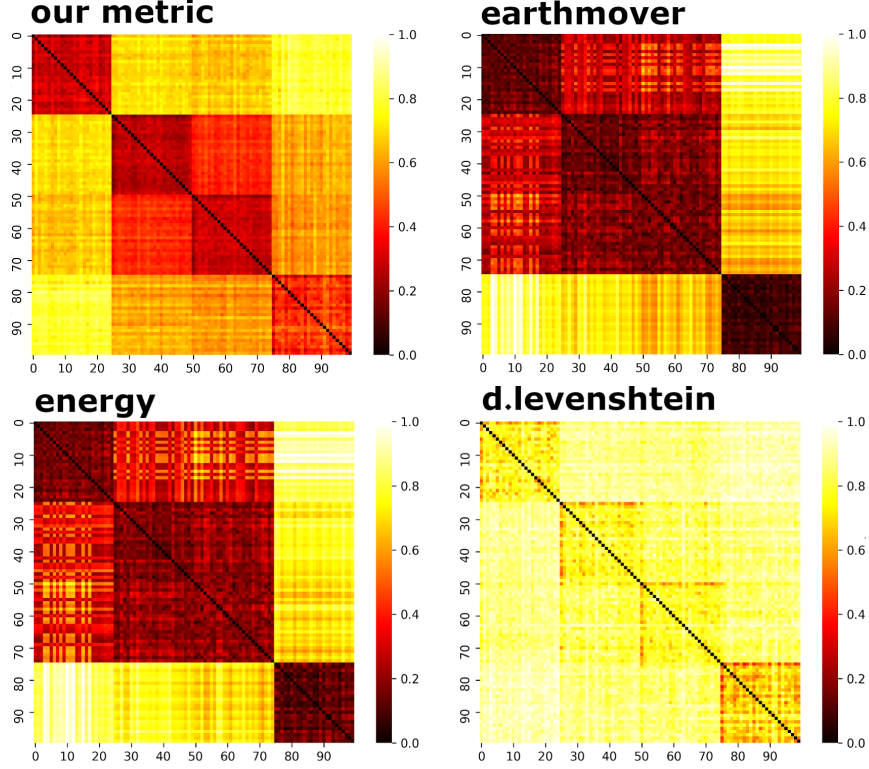


Fig. 6. The heatmaps of the resulting distance matrices are calculated using synthesized study paths, which stem from four actual study paths, using 50-30-20 variation probabilities. Our metric clearly shows the distinct clusters every 25 rows, which are the generated variations of each of the four input paths. Paths 2 and 3 are closer together than the rest, which is why both Earth Mover and Energy distance have trouble distinguishing those. We found that Damerau-Levenshtein works better for short paths, which explains the poor result since the input paths chosen for this experiment are all finished studies and contain around 60 events each.

test configuration. The first column contains the original path with a GPA of 4.0, and the next 24 entries are the variations we built from that path. Index 25 in the heatmap is the original path with GPA 3.0, followed by its variations, continuing in this fashion. We ran the same experiments with 1000 paths and got similar results, but the resulting heatmaps are not possible to visualize due to space limitations, which is why we omitted them here.

We chose the representation of the results as heatmap because it clearly shows the differences between the variants of the generated paths. It is especially noteworthy that the distances between the paths with a GPA of 4.0 and 1.0 are the highest, and the distances between the paths from 2.0 and 1.0 are relatively close to each other in comparison.

To evaluate the results, we compared them against difference calculations using the Energy distance [18], the Earth Movers distance [15] as well as the Damerau-Levenshtein distance [16], [17]. We then used k means as well as DBSCAN (epsilon: 0.8, min. points: 5) to find clusters in the resulting distances. Finally, we repeated the same setup using different probabilities of changes as well as a higher number of generated paths. Table II shows the resulting clustering performance measures for the above-described probabilities (50-30-20), as well as the probabilities 70-20-10 and 30-40-30,

each evaluated with 100 paths. The results from k means were omitted here and can be found in appendix section C since, in a real-world scenario, the number of clusters is usually unknown.

Table II shows that our proposed metric achieved optimal results for all three iterations of the test set, while the other distance measures could not reproduce the same results. The most problems for the other measures stem from the similarity between the paths with a GPA of 3.0 and 2.0. This closeness leads to intersections between the two. DBSCAN did not find any clusters using the Damerau-Levenshtein distance on the 50-30-20 test set, which is noteworthy since using k means, given four clusters, the Damerau-Levenshtein distance performed better than the Earth Movers and the Energy distance. Our metric got optimal results using k means clustering as well (see appendix C).

The appendix for this paper can be found at the supplementary GitHub archive <https://github.com/VDA-univie/set-of-events-distance-metric>.

C. Clustering real data

We also clustered real study paths using the four metrics and the same DBSCAN setup from before. For this, we do not have ground truth about the clusters, so we chose to show the resulting clusters as heatmaps in figure 7. The results from our

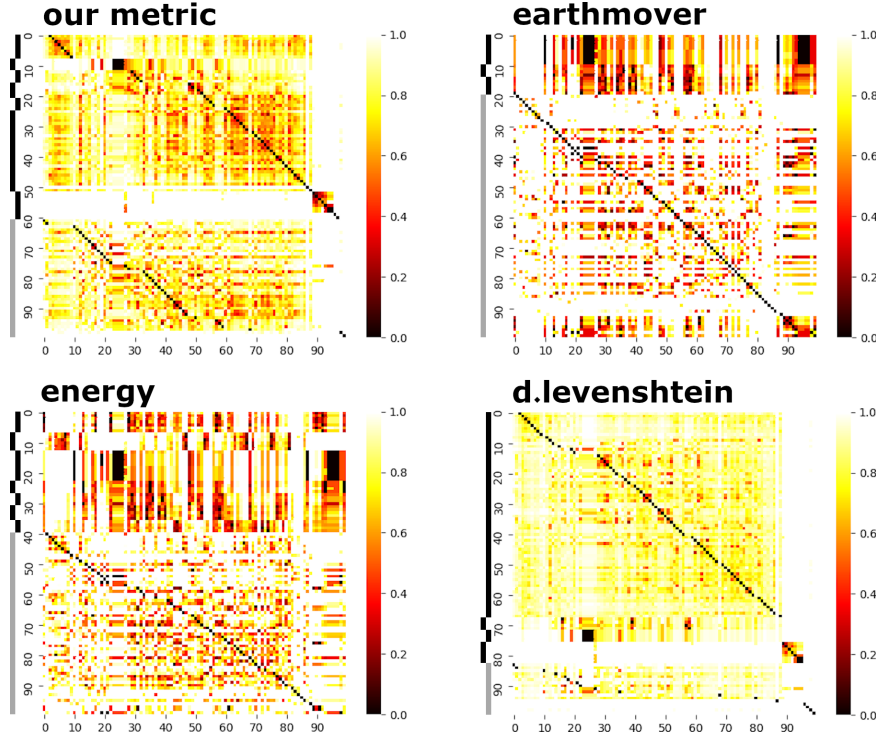


Fig. 7. The distance matrices from clustering a sample of 100 study paths with DBSCAN (epsilon 1.3, min. points 4) using the four different metrics. The cluster labels sort the rows. Black segmented bars to the left of the heatmaps indicate the rows belonging to the same cluster, and the grey bar shows the noise. Using our metric DBSCAN found 7 clusters and 39 rows were noise; Earth Movers finds 3 clusters, 81 rows of noise; Energy finds 7 clusters, 60 rows of noise; and Damerau-Levenshtein finds 4 clusters and 17 rows of noise.

metric show values spread across the spectrum, and clusters accordingly. It finds seven clusters and 37 rows of noise. The spread and size of the clusters fit with the expectations we had about the data, and we confirmed samples of them by manually comparing the clustered paths.

Earth Movers and Energy distance have many results in the extreme regions at the upper and lower end of the range and do not produce values evenly spread across the spectrum. The vast difference in individual distances makes it hard to cluster and, consequently, both those distance measures find more noise than clusters. Clustering with the Earth Movers distance results in three clusters and 81 rows of noise and the Energy distance finds seven clusters and 60 rows of noise.

Damerau-Levenshtein has almost all values in the range of 0.7 to 0.9. Therefore, it finds mainly one massive cluster containing all rows with values from that range, which does not represent the underlying data at all. Overall, it finds four clusters and 17 rows of noise.

We sampled results from all distance measures and compared the clustered paths, and our metric produced results similar to manual clustering, while most samples from the other metrics seemed more randomly clustered.

D. Prediction

We used the results of our metric to test a simple prediction of the number of semesters a student from the exam corpus is going to need until he or she finishes. For this task, we

sampled 1000 paths, then cut these paths after two, three, and four semesters. We then computed the distance matrices of 800 of these paths for every amount of cut off semesters. The remaining 200 paths — also cut accordingly — were then used to find the k nearest neighbors. Of these neighbors, we then averaged the actual number of semesters of the full path and used that as a prediction for the given path.

Table III shows the results for all semesters, using a k of 40 and the previously used distance metric. We tested other values for k between 5 and 50, and we finally chose a k of 40 because it had the least average error across all metrics and semesters. We also chose to omit predicting after the first semester, because the faculty requires a specific set of courses for the first semester. Therefore, students can only start choosing which courses to take in the second semester. The number of semesters of the paths used ranges from 1 to 18, with an average of 7 semesters.

The results in table III show that our metric starts at an average error of about 1/5 of a semester and steadily decreases with additional data from more semesters. Both Energy distance and Earth Movers distance are not well suited for this task, and their results varied massively across runs. The Damerau-Levenshtein (DLev) distance produced similar results to our metric for a low number of courses, most of them even better than the results from our metric, especially at lower k 's. However, as soon as the paths got longer, the

2 A Distance Metric for Sets of Events

TABLE III

THE AVERAGE ERROR OF PREDICTING THE NUMBER OF SEMESTERS OF STUDY PATHS FROM THE EXAM CORPUS USING THE AVERAGE AMOUNT OF THE 40 NEAREST NEIGHBORS. PREDICTIONS WERE MADE WITH 800 PATHS AS TRAINING DATA AND 200 PATHS FOR TESTING.

semester	our metric	Energy	EM	DLev
2	0.1954	0.2934	0.2479	0.1350
3	0.1821	0.5898	0.5642	0.1432
4	0.1382	0.3786	0.3429	0.2822

DLev distance produced increasingly worse results.

This straightforward application of our metric for prediction shows potential for even more uses, e.g., predicting the courses of the following semester for capacity planning at universities. The Future Work section (VI-E) talks about more uses for our metric in prediction.

E. Generic inputs

The exam corpus data set we used for testing has two issues — it is not easy to understand for someone who does not know our faculty, and it contains sensible data from our students, which makes it impossible for us to publish it. We tackled this problem by including a module that converts the string representation of paths — as described in section III-A — into the set of events structure with which the metric can work. This conversion module makes it easy to create individual paths that everybody can understand and use these to test the metric. It also allows us to create anonymous string representations of the actual study paths from our exam corpus data set, which we can provide publicly.

The string conversion module accepts arbitrary Unicode strings as input where each character is a single event and uses dashes for the separation of sets. Therefore, it is easy to apply our metric on all kinds of other external data sets by transforming them into this string representation.

V. USER STUDY

To assess how users would judge the difference between study paths, we conducted an online survey using generic study path representations and compared its results with the results from our metric. This section introduces a different shorthand notation, which we used in the user study, describes the study design, and discusses the results.

A. Study design

	1. Semester	2. Semester	3. Semester
Student 1	A	D	B C
Student 2	B	A C	D

Fig. 8. Example of a study path representation in the user study (Q5 in our survey)

To find out whether our metric represents our users' intuition, we designed a study that tested the users' understanding of the distance between study paths. For this, we conducted an online survey with 15 questions in total. Each question was asking the users to rate the similarity between two study paths. The study paths were laid out in a tabular fashion, shown in figure 8, where each cell represents a semester, and each course (capital letter) appears in a new row. We chose this layout over the representation shown in section III-A, because the pilot for this study clarified, that the textual notation introduced a bias from reading the paths like words and comparing them by their "sound." This pilot also showed that it was more intuitive for the users to judge similarities instead of differences, which is why we transformed our distance measure into a similarity measure for this study by calculating $(1 - \text{dist})$.

The questions asked were chosen as follows:

- one pairing of two opposite paths (A-B-C-D vs. D-C-B-A)
- one pairing of the same path twice (A-B-C-D vs. A-B-C-D)
- five more path pairings of paths with four items each, spread evenly across the similarity spectrum, chosen randomly
- three pairings with eight items each (handcrafted, such that the first path is easy to comprehend)
- three pairings with twelve items each (also handcrafted)
- two pairings with miss-matched events (e.g. A-B-C-D vs A-B-X-Y)

We gave the users a choice of five possible similarities for each question (0%, 25%, 50%, 75%, and 100%) from which they had to pick the closest one. At the beginning of the survey, a short textual task introduction was given, as well as two explained examples using pairings with three courses each for reference. The full survey with the introduction and all questions, as well as all answers, can be found in the supplementary GitHub archive <https://github.com/VDA-univie/set-of-events-distance-metric>.

B. Results

Thirty-four participants took part in that survey, all with a university background, to ensure they had a general understanding of the concept of a *study path*. The participants include undergrad, graduate, and postgraduate students of multiple faculties from different universities. We used 33 of the results and excluded one participant who judged both the identical and the opposite paths with 50% similarity.

Figure 9 shows the exact results for each question of the survey. *Q1* through *Q15* are the questions, the rows *0%* to *100%* show the number of times the participant chose an answer. *Avg similarity* is the average similarity that the users picked for a question, *calc similarity* is the similarity produced by our metric and *difference* is the difference between the two similarities. The total average difference between the calculated results and the results from the user study is 15.72%. The results show that the average difference of the short paths, with only four courses each (questions 1-7), have

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
0%	1	0	0	25	16	0	0	4	3	10	5	0	22	1	31
25%	3	0	0	4	9	17	1	3	13	9	9	14	6	4	0
50%	22	19	0	1	2	11	10	19	10	9	9	7	2	26	2
75%	7	11	0	3	6	5	22	7	7	4	10	12	3	2	0
100%	0	3	33	0	0	0	0	0	0	1	0	0	0	0	0
Avg similarity	0.5152	0.6288	1	0.1136	0.2348	0.4091	0.6591	0.4697	0.4091	0.3258	0.4318	0.4848	0.1439	0.4697	0.0303
Calc similarity	0.3333	0.8333	1	0	0.1666	0.6666	0.5	0.606	0.7142	0.4545	0.8484	0.8095	0.1428	0.5	0
Difference	-0.1819	0.2045	0	-0.1136	-0.0682	0.2575	-0.1591	0.1363	0.3051	0.1287	0.4166	0.3247	-0.0011	0.0303	-0.0303

Fig. 9. This table shows the individual results of the user study. Columns Q1 through Q15 are the questions from the survey, and the percentages are the possible answers for the questions. The values in these rows and the overlaid histogram show how often users picked that option, and the red lines indicate the calculated results from our metric. Bottom part: Avg similarity is the average answer given by the users, calc similarity is the similarity calculated by our metric, and the last row is the difference between the two.

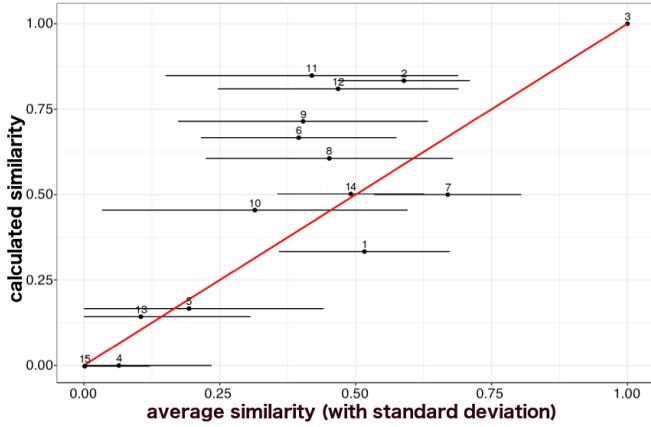


Fig. 10. The average answer of each question (including the standard deviation) compared to the calculated results from our metric — the red diagonal acts as a reference for full correlation.

a lower average difference to the calculated results at 11.62% compared to the difference of 21.88% between the longer paths (with eight and twelve courses – questions 8-13).

The comparison of the average answers to the calculated results in figure 10 shows the users generally underestimate the similarity between two paths, the only notable exclusions from that being question 1 (AB-CD vs. ABCD) and question 7 (D-AB-C vs. CD-AB). In question 1, false judgment appears to happen because moving two of the four courses creates the same paths. In terms of our metric, the matrix of positional relations is all zero for ABCD, and the matrix for AB-CD contains two zeroes and two ones in each row. The users miss that the diagonal does not contribute to the sum since the relation of an event to itself always stays zero, therefore, moving one event changes two relations, and only one stays identical. A similar misjudgment seems to happen for question 7 because moving C from the back to the front changes its relation to all three other events, thus changing three of the six positional relations.

Questions 3 (identical paths) and 4 (opposite paths) are almost identical between users and metric, but interestingly so are questions 5 (A-D-BC vs. B-AC-D) and 13 (A-BC-DEF-G vs DFG-AE-C-B). The latter two show a

much higher standard deviation, however. Question 5, question 10, and question 11 stand out since their standard deviation is above 25%, which is higher than the differences in possible answers. The fashion in which we crafted questions 10 (ABC-DEF-GHI-JKL vs. GHI-JKL-ABC-DEF — where we just swapped semesters 1, 2 and 3, 4 while leaving their contents identical) and 11 (ABC-DEF-GHI-JKL vs. AB-CD-EF-GH-IJ-KL — where we did not swap courses, but instead just changed the number of courses in each semester) shows that the participants put a different weight in the order of courses versus the semesters they were taken in. Users who put the weight in the order were generally closer to the calculated results in Q10 and Q11 because the metric also scores based on the relation of the courses. The other group of responses was that of users that put more weight in the respective semesters the courses were taken in, and, therefore, had a higher difference to the metric results.

The two questions with different events both score close to the calculated result, interestingly in Q15 (A-B-C-D vs. W-X-Y-Z) two users still answered that the similarity is 50%. This result could be because they judged the events and the structure of the paths separately, and the two paths are structurally identical — four sets with one event each.

The feedback from the users after taking the survey also showed that they had more issues judging the longer paths compared to the short ones. Most users reported coming up with specific criteria to estimate the similarity, which they could apply to the short examples easily, while they tended to use their intuition for longer paths.

Concluding the results of the user study show that the metric does not fully cover the intuition of all users we asked, but with an overall average difference of 15.72%, it comes reasonably close, and it even decreases for shorter paths to 11.62%. The average study path for which we apply this metric has 40+ different courses, which is a much more substantial amount than tested in this study. Therefore, the task of manually comparing sets of events becomes more error-prone, and thus a stable metric is very valuable.

VI. DISCUSSION

In this section, we discuss the results of our metric and talk about current limitations. We also give an overview of the

alternatives we explored and list some future projects to use and advance the metric.

A. Variations of the Levenshtein distance (string edit distance)

At first, we tried to use available metrics with slight adaptations to fit our problem's requirements. Since we abstracted the idea of finding similar study paths as a reordering problem, we first chose the Levenshtein distance [17]. We assigned each course a unique Unicode character and then concatenated these characters in their respective order. We tried different approaches with the resulting strings: every semester being a word, the full path being a word — with and without including 0 cost swaps within semesters — but all approaches failed.

While the initial test cases seemed promising, a straightforward test case could not be solved, namely the distance between ABC-D and D-ABC. It was maximal in all cases that we tried because every letter is shifted by one and thus compared to its neighbor. When every semester is a word ABC is compared to D and vice versa, so the distance is still maximal. The only version in which it did not turn out to be maximal was a multi-word scenario where we matched every word with the closest distance from the other path and tried summing up only the closest distances, thus comparing ABC to ABC and D to D which gave us a distance of 0. The actual result for this example from our metric is 0.5, since the positional relation between D and all other courses changes, but the remaining relation within courses stays unchanged.

B. Variations of the Graph Edit Distance

Our second approach was looking at the problem from the perspective of graphs. By defining the semesters as nodes and their temporal relation as directed edges, we tried appropriate versions of the graph edit distance [12]. The main issue we encountered with this was that nodes within graphs could not consist of multiple items. When trying to circumvent that by putting concurrent nodes in parallel, we ran into the problem that we would run into issues with circular connections between courses that introduced unwanted variations in the resulting distances. Another option was to introduce artificial separation nodes after each parallel set, but this made the resulting graphs proportionally longer and, therefore, skewed the results of long paths versus short ones.

The final issue with this method was the computing time, which grows exponentially based on the size of the graph, and since we want to compare multiple thousands of paths with lengths of up to one hundred courses, this approach was not a feasible option.

C. Variations of our own Metric

We settled for the distance as a weighted sum of positional differences early in the process, but the way the difference of time intervals was calculated varied. Initially, we used the difference between the actual positions and not only their relation to each other. E.g. the positional relation between A and C in A-B-C would be -2, since A is two intervals before

C. The result of this was that long paths had huge distances between them, which did not compare well against short paths.

The second version we tried has only one difference to the final metric, namely equation 11, did not have the signum function in the sum. This implies that a difference between before (-1) and after (1) contributed double the amount than a difference between before and concurrent (0) or concurrent and after, since $|a_{i,j} - b_{i,j}| = 2$ for $a = 1$ and $b = -1$, but it is only 1 for $a = 1$ and $b = 0$. The problem with this version is that when calculating the distance between a path with an even number of semesters and a path with an odd number of semesters, the maximal distance can never be one since at least one object would only contribute 1 to the total sum. E.g., the distance between the paths A-B-C and BC-A: the matrices of positional relations of the two are the following:

$$M_{A-B-C} = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \quad \text{and} \quad M_{BC-A} = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad (2)$$

The result of the sum of their differences using the signum function is 6, averaged by $1/n(n-1)$ the distance between the two paths is 1, since all courses happen in a different order. If the signum function is left out, the sum of differences in this example is 10, normalized by $1/2n(n-1)$ results in 0.833. In contrast, the distance between A-B-C and C-B-A, where also all courses happen in a different order, is 1 in both variants, therefore making the variant without the signum function less consistent.

Our final version of the metric includes the signum function, which ensures that every difference between positional relations contributes precisely 1 to the final sum, thus removing the problems mentioned above.

D. Limitations

A current limitation of our approach is that every event can only appear once, which can easily be tackled by uniquely identifying duplicate events (e.g., instead of having AA, labeling them A_1A_2). It might be a limitation on the one hand, but can also be used as a benefit when using the metric to predict multiple occurrences of the same event. For example, when dealing with *study paths*, we use this to determine if a student had multiple failed attempts of one course and in which semesters they were.

Another limitation is that events that do not occur in both compared sets just add one to the sum of differences and do not contribute to the structure of the path that is compared by the positional distance function. It works well for the exam corpus because the *knowledge* of a course that one student did not yet take is not available for other courses. In other data sets, which do not depend on *knowledge* of previous events, this fact might limit the usability of the metric.

E. Future work

We intend to use the results of the metric as a predictor for planning future semesters at the University of Vienna. Predicting could be done by finding the closest paths to every

currently active student and, based on these paths, computing probabilities for courses that this student is most likely to take in the next semester. We are also looking into using the results from our metric as weights for building a Markov-Chain of all courses in our faculty, which we can then use to predict future semesters. The metric can also be used to visually monitor current trends and the general state of a faculty by clustering study paths, thus making them easier to explore and evaluate.

Other application domains that seem applicable for our metric are the analysis of online purchases and incident management data. In online purchases, every individual purchase could act as time interval, the bought items as objects in those intervals, and then use the similarity to other customers' purchases for suggestions or targeted advertisements. Using the similarity of incidents in a similar fashion could automatically assign an incident to a group that is more likely to solve it, or prioritize incidents early that show signs of being unresolved for an extended period.

Finally, the medical domain, specifically the patient management and individual patient's treatment procedure in hospitals and other treatment facilities, looks to be a promising field. Treatments happen chronologically, and various events, such as administering medication, are happening during each treatment. The notion of concurrency can be set individually, for example, from different events happening during a single surgery, up to all treatments in a day, or even a week. In this case, the *knowledge* of previous events — as described in the Limitations subsection (VI-D) — would be the influence a previous medication or treatment has on the next ones. We are currently looking to find hospitals willing to cooperate with us.

VII. CONCLUSION

In this work, we presented a novel distance metric to calculate the difference between sets of events. Further, we proved that the mathematical conditions of metrics held for our approach and showed its application by using it for clustering and predicting a real-world data set. A user study further solidifies the validity of our approach by showing that the metric results represent the expectations of the users.

We provide a python implementation of the metric in the supplementary GitHub archive <https://github.com/VDA-univie/set-of-events-distance-metric>. It includes the data conversion code we used to build study paths from the exam corpus, an example workflow and the generator code we used to create the data for the robustness test. Since we cannot make the exam corpus publicly available, due to privacy protection policies, we provide the full exam corpus in an anonymized fashion using the string-representation of study paths. To further test the metric, the archive also includes a transformation script that can generate sets of events from string inputs based on the notation introduced in section III-A. The appendix of this work can also be found in the GitHub archive.

REFERENCES

- [1] T. chung Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164 – 181, 2011. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0952197610001727>
- [2] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," *Proc. VLDB Endow.*, vol. 1, no. 2, pp. 1542–1552, Aug. 2008. [Online]. Available: <http://dx-doi-org.uaccess.univie.ac.at/10.14778/1454159.1454226>
- [3] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS'94. <https://dl.acm.org/doi/abs/10.5555/3000850.3000887>: AAAI Press, 1994, p. 359–370.
- [4] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 285–289. [Online]. Available: <https://doi.org/10.1145/347090.347153>
- [5] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, mar 2005. [Online]. Available: <http://link.springer.com/10.1007/s10115-004-0154-9>
- [6] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*. <http://ieeexplore.ieee.org/document/380415/>: IEEE Comput. Soc. Press, 1995, pp. 3–14. [Online]. Available: <http://ieeexplore.ieee.org/document/380415/>
- [7] Y. Li, P. Ning, X. S. Wang, and S. Jajodia, "Discovering calendar-based temporal association rules," *Proceedings of the International Workshop on Temporal Representation and Reasoning*, vol. 44, pp. 111–118, 2001.
- [8] L. Wang, J. Meng, P. Xu, and K. Peng, "Mining temporal association rules with frequent itemsets tree," *Applied Soft Computing Journal*, vol. 62, pp. 817–829, 2018. [Online]. Available: <https://doi.org/10.1016/j.asoc.2017.09.013>
- [9] J. M. Ale and G. H. Rossi, "An approach to discovering temporal association rules," in *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1*, ser. SAC '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 294–300. [Online]. Available: <https://doi.org/10.1145/335603.335770>
- [10] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," *KDD Workshop on Data Cleaning and Object Consolidation*, vol. 3, pp. 73–78, 2003.
- [11] —, "A Comparison of String Distance Metrics for Name-Matching Tasks," in *Proceedings of the 2003 International Conference on Information Integration on the Web*, ser. IIWEB'03. Jakarta, Indonesia: AAAI Press, 2003, pp. 73–78.
- [12] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis and Applications*, vol. 13, no. 1, pp. 113–129, 2010.
- [13] G. J. Székely, "Potential and kinetic energy in statistics," *Lecture Notes, Budapest Institute*, 1989.
- [14] M. L. Rizzo and G. J. Székely, "Energy distance," *WIREs Comput. Stat.*, vol. 8, no. 1, p. 27–38, Jan. 2016.
- [15] Y. Rubner and C. Tomasi, *The Earth Mover's Distance*. Boston, MA: Springer US, 2001, pp. 13–28. [Online]. Available: https://doi.org/10.1007/978-1-4757-3343-3_2
- [16] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, p. 171–176, Mar. 1964. [Online]. Available: <https://doi.org/10.1145/363958.363994>
- [17] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.
- [18] G. J. Székely and M. L. Rizzo, "Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method," *Journal of classification*, vol. 22, no. 2, pp. 151–183, 2005.

2 A Distance Metric for Sets of Events

APPENDIX A PROPERTIES OF THE METRIC

TABLE IV
OVERVIEW OF THE SYMBOLS USED TO DESCRIBE THE DIFFERENT PARTS
OF OUR METRIC

S	set of events
e	event — combination of an object and an identifier
o	object (e.g. a course)
i	identifier which can be ordered (e.g. semester, can be ordered in time)
S^O	set of objects from set of events
$\text{dist}(S_A, S_B)$	metric distance function
δ	positional relation (before, concurrent, after)
M_S	matrix of positional relations in set S
w	contribution factor
$\text{pos_dist}(S_{AB}^*, S_{BA}^*)$	distance between the sets of identical events of the input

Given a set of objects O and an ordered set of identifiers I , we define an event e as $e \in O \times I$ and a set of events S as

$$S = \{e : e \in O \times I\}, \quad (3)$$

as well as its projection onto just the object space S^O as

$$S^O = \{o : \exists e = (o, i) \in S\}, \quad (4)$$

A set of events S is the basic representation of data used in our metric, its projection onto the object space is used to determine all objects that appear in both compared sets.

Given two sets S_A and S_B , the contribution factor w is the number of objects that occur in both S_A and S_B divided by the total number of objects in both sets.

$$w = \frac{|S_A^O| + |S_B^O| - (|S_A^O \setminus S_B^O| + |S_B^O \setminus S_A^O|)}{|S_A^O| + |S_B^O|} \quad (5)$$

Each object contributes equally, therefore the overlap of the two sets $|S_A \cap S_B|$ is counted twice, because each object appears twice in the overlap.

$$w = \frac{2(|S_A^O \cap S_B^O|)}{|S_A^O| + |S_B^O|} \quad (6)$$

The distance function of our metric consists of two parts, $w \times \text{pos_dist}(S_A^*, S_B^*)$ is the positional distance, which is used for the part of the sets, where the objects are identical. $(1 - w) \times 1$ describes the contribution of objects that only occur in one of the two sets, since the maximum distance possible in our function is one, these objects each contribute one to the distance.

$$\text{dist}(S_A, S_B) = w \times \text{pos_dist}(S_{AB}^*, S_{BA}^*) + (1 - w) \times 1 \quad (7)$$

S_{AB}^* and S_{BA}^* are the subsets of S_A and S_B , where the objects are identical.

$$S_{AB}^* = \{s = (o, i) \in S_A \text{ and } \exists i_B, \text{ s. t. } (o, i_B) \in S_B\} \quad (8)$$

We define the matrix M_S as the positional relations δ between all events in S . This matrix is computed for each set of events individually.

$$M_S = (\delta(e_i, e_j)) \quad : \quad \forall e_i, e_j \in S \quad (9)$$

The positional relation $\delta(e_1, e_2)$ indicates the relation of event e_1 to event e_2 using the order of the identifiers I . The function returns -1 if e_1 happens before e_2 , 0 if both are concurrent and 1 if e_1 happens after e_2 . It is defined as

$$\delta(e_1, e_2) := \begin{cases} -1 & i_1 < i_2 \\ 0 & i_1 = i_2 \\ 1 & i_1 > i_2 \end{cases} \quad (10)$$

where i_1 and i_2 are the positions given by the identifiers of the events e_1 and e_2 respectively.

Since the relation between two identical events is always zero, the diagonal entries of the matrix M_S are zero. Furthermore M_S is skew symmetric ($m_{i,j} = -m_{j,i}$), because if e_1 happens before e_2 , symmetry implies that e_2 happens after e_1 and vice versa.

The positional distance function sums up the differences in the matrices of positional relation of the subsets with identical objects of both input sets S_A and S_B .

Let M_A and M_B be the matrices of positional relations of the sets of events S_{AB}^* and S_{BA}^* , and $n = |S_{AB}^*| = |S_{BA}^*|$ be the number of events with common objects in one of the two sets, then the metric distance between S_{AB}^* and S_{BA}^* is defined by

$$\text{pos_dist}(S_{AB}^*, S_{BA}^*) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \text{sgn}(|a_{i,j} - b_{i,j}|) \quad (11)$$

where $a_{i,j} \in M_A$ and $b_{i,j} \in M_B$.

The sum of differences is averaged over all non-diagonal entries — of which there are $n(n-1)$ many — to ensure the result is between one and zero.

The signum function ensures that all differences contribute equally to the sum of differences. Further, this distance metric only makes sense if the position i of event e_i containing object o is identical in both matrices A and B .

A. Metric Properties

This subsection proves that our distance measure satisfies all metric properties.

Lemma 1. *Our distance measure $\text{dist}(S_A, S_B)$ is non-negative.*

Proof: $|a - b| \geq 0$ for any values of a and b , the number of rows/columns n in the matrix M_S is also always positive, and therefore the positional distance function is always greater or equal to zero.

$$\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \text{sgn}(|a_{i,j} - b_{i,j}|) \geq 0. \quad (12)$$

$$w \times \text{pos_dist}(S_{AB}^*, S_{BA}^*) + (1 - w) \times 1 \geq 0 \quad (13)$$

The contribution factor w (see eq. 6) describes a percentage of the input, thus it can only be between 0 and 1, therefore our metric is always greater or equal than zero. ■

Lemma 2. *If two sets of events are indiscernible, the result of our distance measure is zero.*

Proof: Two sets of events X and Y are indiscernible, if they contain the same objects, and those are in the same order. If this is the case their matrices of positional relations are identical ($M_X = M_Y$) and the sum of differences between two identical matrices is zero.

A distance of zero is only the case if X and Y are identical, if they differ in only one event the sum of differences is not zero anymore, which thus concludes that if $\text{dist}(X, Y) = 0$, then $X = Y$. ■

Lemma 3. *Our distance measure is symmetric.*

Proof: The positional distance function (equation 11) is the sum of differences between the positional relations of the events. One is added to the sum if the positional relations are different, zero otherwise. Changing the order is equivalent to changing $\text{sgn}(|x_{i,j} - y_{i,j}|)$ to $\text{sgn}(|y_{i,j} - x_{i,j}|)$. Since the sum is formed on the absolute value of differences, the sign is cancelled and the signum function guarantees that no other values than one and zero can be added. Therefore one is added to the sum whenever $x_{i,j}$ and $y_{i,j}$ are different independent of their order, which concludes $\text{dist}(X, Y) = \text{dist}(Y, X)$. ■

Lemma 4. *Our distance measure fulfills the triangle inequality, which means*

$$\text{dist}(X, Z) \leq \text{dist}(X, Y) + \text{dist}(Y, Z) \text{ for all } X, Y \text{ and } Z.$$

Proof: We are proving this inequality component-wise. The distance $\text{dist}(X, Z)$ is calculated by summing up the differences between all positional relations (entries) in M_X and M_Z . $\text{sgn}(|x_{i,j} - z_{i,j}|)$ can either take the value zero if $x_{i,j}$ and $z_{i,j}$ have the same positional relation, or one if their positional relation is different.

- **Case 1:** $\text{sgn}(|x_{i,j} - z_{i,j}|) = 0$
Zero is the lowest value that can appear, therefore $0 \leq \text{sgn}(|x_{i,j} - y_{i,j}|) + \text{sgn}(|y_{i,j} - z_{i,j}|)$.
- **Case 2:** $\text{sgn}(|x_{i,j} - z_{i,j}|) = 1$
Our proof is by contradiction. Let us assume that $\text{sgn}(|x_{i,j} - y_{i,j}|) + \text{sgn}(|y_{i,j} - z_{i,j}|) = 0$. The only case where this happens is, if $x_{i,j}$ and $y_{i,j}$, as well as $y_{i,j}$ and $z_{i,j}$, are concurrent events. Hence, by transitivity, $x_{i,j}$ and $z_{i,j}$ are concurrent as well. This means $\text{sgn}(|x_{i,j} - z_{i,j}|) = 0$ which is a contradiction.

By proving that the difference between individual entries of the positional relations fulfill the triangle inequality and since our distance measure is the sum of all individual differences, all of which are non-negative, it follows that it fulfills the triangle inequality. ■

APPENDIX B

INSTRUCTIONS FOR EXPERIMENT REPLICATION

All data, code as well as the instructions and answers to our user study can be found here: <https://github.com/VDA-univie/set-of-events-distance-metric>. The instructions are written for the bash shell, since it is available on all operating systems. To run the experiments described in this work, the following steps need to be done first to set up the environment.

- 1) download the 'code' directory to your machine
- 2) navigate to the 'code' directory in a terminal
- 3) create a virtual environment using python3
`python3 -m venv env`
- 4) activate the virtual environment
`source env/bin/activate`
- 5) install all required packages
`pip install -r pip-requirements`

Once completed the experiments can be run with the commands shown in table V.

TABLE V
COMMANDS USED TO REPLICATE THE DIFFERENT EXPERIMENTS FROM SECTION IV.

Clustering generated paths
<code>python path_clustering_generic.py</code>
Clustering of real paths
<code>python path_clustering.py</code>
Predicting path lengths
<code>python path_predict.py</code>

All parameters are set such that the results shown in this work are replicated. The parameters can be changed by editing the source code, where each parameter is named accordingly.

APPENDIX C

DETAILED CLUSTERING RESULTS

Tables VI, VII and VIII show more detailed results from clustering the generated study paths. The probabilities used for generating the paths have been varied, all other parameters stayed the same. DBSCAN was run with epsilon of 0.8 and the minimum number of points was 5.

2 A Distance Metric for Sets of Events

TABLE VI
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 70-20-10

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.740380813	4	0
Energy	kMeans	0.75	0.892601039	0.815110624	0.652883569	0.740611948	0.818199718	0.694392994	4	0
EM	kMeans	0.75	0.892601039	0.815110624	0.652883569	0.740611948	0.818199718	0.66375268	4	0
Dlev	kMeans	1	1	1	1	1	1	0.470027398	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.740380813	4	0
Energy	DBSCAN	0.792802453	0.679549349	0.731820194	0.601756278	0.658452905	0.733994817	0.633068367	6	1
EM	DBSCAN	0.77935697	0.575018842	0.661773399	0.520272708	0.535539742	0.669436287	0.457225376	9	4
DLev	DBSCAN	0.498300818	0.431093361	0.462267097	0.18400579	0.360822106	0.463480501	-0.142133249	10	53

TABLE VII
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 50-30-20

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.66528544	4	0
Energy	kMeans	0.75	0.869013399	0.805132372	0.632383295	0.740779426	0.807316573	0.619289976	4	0
EM	kMeans	0.75	0.869013399	0.805132372	0.632383295	0.740779426	0.807316573	0.572138792	4	0
Dlev	kMeans	1	1	1	1	1	1	0.213137912	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.66528544	4	0
Energy	DBSCAN	0.806200476	0.649006172	0.719113104	0.609812209	0.623622465	0.723345758	0.593211605	7	1
EM	DBSCAN	0.731543477	0.441038545	0.550304993	0.349707838	0.374413587	0.568013091	0.197628526	14	14
DLev	DBSCAN	0.020447872	0.289137359	0.038194612	0.000830441	0.007903456	0.076891117	0.092768369	1	98

TABLE VIII
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 30-40-30

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.613383675	4	0
Energy	kMeans	0.75	0.858419754	0.800555718	0.621994795	0.740884491	0.802380717	0.563085631	4	0
EM	kMeans	0.809563453	0.825070928	0.817243632	0.753666841	0.802826335	0.817280411	0.459402372	4	0
Dlev	kMeans	1	1	1	1	1	1	0.109391893	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.613383675	4	0
Energy	DBSCAN	0.764070217	0.64863478	0.701636248	0.563013699	0.629685354	0.703990424	0.490243353	5	3
EM	DBSCAN	0.714315308	0.406440431	0.518090805	0.247930083	0.338350653	0.538819656	0.148717497	14	19
DLev	DBSCAN	3.20E-16	1	6.41E-16	0	9.61E-16	4.44E-06	0	0	100

3 Histogram binning revisited with a focus on human perception (accepted)

Synopsis

The following chapter contains the contents of Raphael SAHANN, Torsten MÖLLER, and Johanna SCHMIDT "*Histogram binning revisited with a focus on human perception*". The version below was submitted to the IEEE VIS 2021 conference on March 31, 2021.

Most of the data creation, data analysis, and creation of the respective figures were done by Raphael Sahann. Johanna Schmidt and Raphael Sahann collaborated on everything else. Torsten Möller helped define the hypotheses, phrasing the user study questions, and gave feedback on the final text.

Note from the final edit of this thesis: A revised and shortened version of the paper in this chapter has been accepted to the IEEE VIS 2021 Short Paper track. It was decided to leave the previously submitted version of the full paper in the thesis, since its core contribution did not change, and it presents more information on the topic.

Histogram binning revisited with a focus on human perception

Submission ID 1682

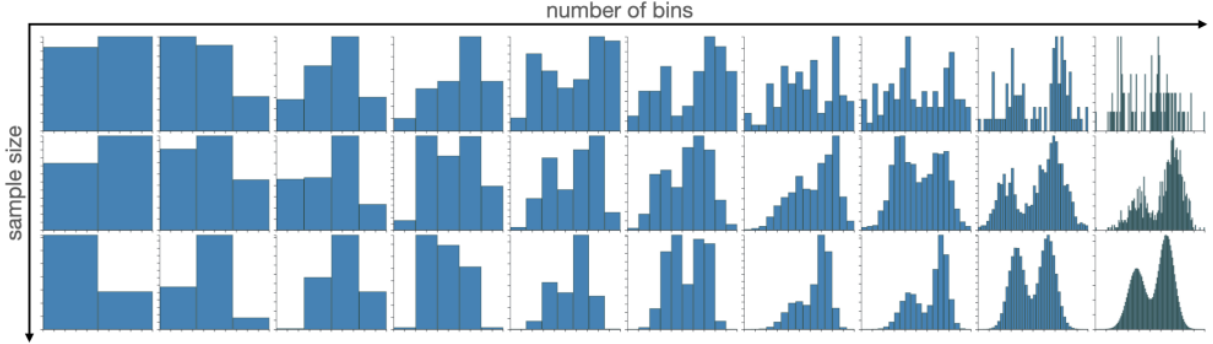


Fig. 1. Selection of histogram datasets used in our study. We evaluated how well human viewers can detect the underlying data distribution in a histogram when different sample sizes and bins are used. For this, we created datasets with a different number of samples (first row: few, last row: many) and a different number of bins (left column: 2, right column: 100). A bimodal distribution was used to create the datasets in this illustration.

Abstract—Many different approaches for selecting the right number of bins in a histogram have already been proposed. These models consider the underlying distribution of the data and the number of samples in the dataset. However, the number of bins suggested by mathematical approaches has not been evaluated with a focus on human perception yet. This paper presents a quantitative user study to evaluate how well users can visually perceive the underlying data distribution from a histogram representation. We used histograms with different sample and bin sizes and four different distributions (uniform, normal, bimodal, and gamma). The study results confirm that, in general, more bins correlate with fewer errors by the viewers. However, upon a certain number of bins, the error rate cannot be improved by adding more bins. By comparing our study results with the outcomes of existing mathematical models for histogram binning (e.g., Sturges’ formula, Scott’s normal reference rule, the Rice Rule, or Freedman–Diaconis’ choice), we can see that most of them overestimate the number of bins necessary to make the distribution visible to a human viewer. Our paper summarizes and discusses all outcomes and concludes with a suggestion for future work.

Index Terms—empirical studies in visualization; histogram binning; distributions; human-centered computing

1 INTRODUCTION

Having already been used in the 19th century [31], histograms can be denoted as one of the earliest types of data visualization techniques. Histograms are a well-known and prevalent visualization technique representing the distribution of univariate data by visualizing the tabulated frequency at certain intervals, represented as bars or *bins*. Bins are usually drawn as rectangles – several bins next to each other help human viewers to build a mental model of the data distribution. In this sense, histograms help estimate where values are concentrated and if outliers can be found in the data. The most important parameter visualization designers have to set when creating a histogram is the *number of bins*, sometimes called the *bin width*. (The greater the number of bins, the smaller the bin width.) With too few bins, the data cannot be accurately represented, and features might be obscured. With too many bins, random artifacts might be created in the visualization, which hinders the analysis of the underlying data’s true distribution.

The number of bins significantly influences how well we humans can interpret a histogram. Statisticians have developed several thumb rules to help researchers estimate the right number of bins when creating a histogram. For example, Sturge’s formula [42] defines how to split

the data into k bins based on the number of samples being available. Many other mathematical models try to match the information generated by binning with the data’s underlying distribution and to minimize the error between the two representations. Scott’s normal reference rule [37] measures the discrepancy between the bin representation and the data distribution by employing mean integrated squared error. The Freedman–Diaconis choice [16] is based on minimizing the difference between the area under the data distribution and the area under the probability distribution defined by the binning. The mathematical models, on the forefront being Sturge’s formula, Scott’s normal reference rule, the Rice Rule [43], and the Freedman–Diaconis choice, are nowadays in use in many visualization applications and libraries.

These mathematical rules use the number of samples as the main input for calculating the number of necessary bins in a histogram. They are fast and convenient binning estimations, but they also have drawbacks. There is an indication that the thumb rules’ strong assumptions about the data make them sub-optimal for non-normally distributed data. As an alternative, Knuth’s rule [25] or Bayesian Blocks [35] use fitness functions computed on the actual underlying distribution to choose an optimal binning, which is computationally more expensive but produces more accurate results for more complicated distributions. There is a lively discussion about which mathematical models describe the data best, and new models are continuously developed (see Section 2).

However, interestingly, the models and their suggested binnings have not been evaluated in perceptual user studies yet. The models in use today were statistically and mathematically evaluated. However, it is unclear how well the suggested numbers of bins match the human visual perception when analyzing a histogram.

In this paper, we therefore address the following research questions:

- ? Is it possible to define an optimal number for the number of bins for human viewers to be able to detect the data's underlying distribution in a histogram?
- ? What is the difference between the binnings suggested from perceptual experiments to the values suggested by mathematical models? Does the perception of a distribution in a histogram depend only on the binning, or also on the sample size of the data (as it is suggested by the mathematical models)?

To answer these research questions, we need to understand how well users can detect the underlying distribution of data in histograms. In general, this is in line with the current need in research to understand how viewers can construct and interpret data visualizations [7]. Laboratory and user studies are a legitimate approach to evaluating users' performance when interacting with data visualizations.

This paper presents a user study where we tested the viewers' abilities to detect the data's underlying distributions in histograms. We used datasets with four different distributions (*uniform*, *normal*, *gamma*, and *bimodal*), different sample sizes, and numbers of bins (see Figure 1) and asked participants to state which distributions they see in the different representations. We then compared the results of the user study to existing models for binning suggestions in histograms. The paper presents the following contributions:

- ! **Perceptually optimal number of bins.** The user study results indicate that a perceptually optimal number of bins can be set independently of the data's sample size. This is under the assumption that enough samples are available to represent the underlying distribution properly.
- ! **Comparison with mathematical models.** In comparison with the mathematical models, it can be seen that most of the models overestimate the number of bins that is necessary to represent the data's distribution for human viewers.

The paper is organized as follows: In Section 2 we discuss related work on histogram binning, perception studies, and evaluation in visualization. In Section 3, the user study setting is described. In Section 4 the results of the user study are described. In Section 5 we present our new suggestion for an optimal histogram binning and compare it with mathematical models. The paper is concluded in Section 6.

2 RELATED WORK

Our research is rooted in histogram binning, but methodologically draws on perception studies on histograms, and evaluation of visualizations in general.

2.1 Histogram binning

The number of bins significantly influences how well we can visually interpret a histogram. Many approaches have already been developed that assess the optimal number of bins for a given distribution. One of the earliest reported methods for constructing histograms was proposed by Sturges [42], who suggested to calculate the binning based on the data range. Although quite simple and easy to calculate, Sturges' formula does not ideally represent the underlying distribution if the sample size of the data is large, and if the samples are not normally distributed [38]. Alternatively, the so-called Rice's rule [43], or the approach by Doane [12] can be used, which also work for non-normal distributed data. Scott [37] proposed to use an error measure between the probability density represented by the histogram and the actual probability density of the underlying data. Freedman–Diaconis' choice [16] adapted Scott's normal to make it less sensitive against outliers. Also similar to Scott's normal, Shimazaki and Shinomoto [39] proposed a method based on minimizing an estimated cost function. These approaches are still quite popular and are often used in current visualization systems (e.g., Python, Matlab).

Some other methods have been, for example, proposed by Stone [40], who used a loss function minimization approach, or Rudemo [34], who

employ risk functions and cross-validation techniques. Wand [47] presented an extension to Scott's normal to have good large sample consistency properties. Hall [18] investigated the use of a different information criterion and, very recently, Knuth [25] proposed a method based on maximizing the posterior probability for a certain number of bins. Heinrich [21] researched how a selection of bins can best model the intuitive decision whether a histogram is uniform or not. Lolla and Hoberock [28] proposed to use Cumulative Distribution Functions (CDF) to assess the optimal number of bins. Birge and Rozenholc [2] proposed a statistical approach based on a penalty function which works especially for small sample sizes. He and Meeden [20] based their approach for selecting the number of bins on a loss function, reflecting the idea that smooth distributions need fewer bins than rough distributions.

All the solutions mentioned here depend on statistical models, mathematical approaches and optimization criteria. In this paper we compare the model-based binning approaches with the results we obtained in a quantitative user study.

2.2 Perception studies (histograms)

Visualizations are interpreted by using our human visual system, and several researchers already tried to better understand this process by conducting perception studies. When it comes to summary statistics, Lem et al. [27] and Kaplan et al. [24] noticed a general problem for students when trying to read and interpret aggregated information in histograms and box plots. Many of these misinterpretations are related to data mapping (e.g., how many variables are depicted in the graph). Dabos [11] concluded that students often have problems interpreting the variability of a variable in a histogram. Zubiaga and Mac Namee [48] assessed the data literacy of participants when interpreting a distribution of values with different charts, and could report more positive results on the literacy of histograms. Correll et al. [10] highlighted the importance of selecting the right number of bins for detecting missing values and outliers in a histogram. They also pointed out that more liberal rules of thumbs for selecting the number of bins (e.g., Freedman–Diaconis choice) should be preferred over the common Sturges' formula – especially when creating sanity checking histograms.

In general, it can be denoted that, although statistical and mathematical models for computing the optimal binning exist, the literacy of humans interpreting these visualizations seems to be decoupled from the statistical interpretation. According to Boels et al. [4], information reduction still seems to be an understudied topic in data visualization.

In this paper we present a quantitative evaluation of histogram literacy with different number of bins and sample sizes.

2.3 Evaluation of visualizations

User studies [26] offer a scientifically sound method to measure how people read visualizations [22], and to better understand a visualization's readability. In visualization research, however, evaluation approaches are often focused on evaluating the performance of a newly designed visualization in comparison to existing techniques [23], in many cases by collecting qualitative feedback [13]. This is in contrast to user studies where the purpose is to understand human perceptual or cognitive characteristics [45]. User study for understanding perceptual characteristics are targeted towards learning something new about human perception or cognition, which later on can be used to make informed decision for visualization design.

In this paper we decided to run a user study to understand human perception, also sometimes called user evaluation study [15]. We collected quantitative data (i.e., accuracy, confidence, and time it took to complete the task) to gain a basic understanding of the visual perception of data distribution in histograms.

This narrow task definition allowed us to evaluate the *design idea* [46] of the effect of varying binnings in a histogram. This setting is best described as a *judgement study*, where the study's purpose is to gather a person's response to a set of stimuli [9]. The setting is made irrelevant in judgement studies. According to the literature, judgement studies are a commonly used approach for perceptual studies and can provide considerable precision.

3 QUANTITATIVE USER STUDY

We conducted a user study to test how well users can detect the underlying distribution of a sample in a histogram. Histograms with different distributions, number of samples, and different number of bins have been used (see Section 3.1). We established a data generation mechanisms to create histograms with different binnings (see Section 3.2). The study was carried out as a web-based questionnaire (see Section 3.3). We also evaluated the study setting in a test run (see Section 3.4).

3.1 Hypotheses generation

Task definition: One of the first decisions we made when designing the study was to concentrate on one specific task users perform when analyzing histograms. Histograms as summary statistics provide the possibility to perform several tasks related to distribution analysis (e.g., identifying the mean and the median or comparing quartiles). One task related to distribution analysis is to identify the data’s underlying distribution, which has been classified as the task to “*describe and identify the shape and type of one distribution*” by Blumenschein et al. [3]. The identification of the underlying distribution is the task we evaluated in our study.

Distributions: Histograms are not restricted to special types of data distributions. In mathematics and statistics, hundreds of different density distributions can be found. To conduct a study on the shape and type of distributions, we decided that we have to reduce this high number of possible choices to a reasonable number that can be tested in a study. To get an overview of the distributions currently used in practice, we looked into literature targeted towards data scientists to learn more about the use of data distributions. Some examples are: *Doing Data Science* [36] lists 17 density functions that data scientists should be familiar with. The *Data Scientist’s Crib Sheet* [32] describes 15 density functions that are important and highlights their relationships. In the *KDnuggets* tutorials [41] five density functions are explained that data scientists should be aware of. Based on this literature research and based on our own experience when working with data, we decided to classify the available density distributions based on their main shape characteristics. We defined four main classes (see also Figure 2):

- *uniform*: uniform distributions
- *unimodal*: distributions with one peak, similar to a Gaussian kernel
- *bimodal*: distributions with two peaks
- *skewed*: distributions with one peak which are skewed to one side of the distribution

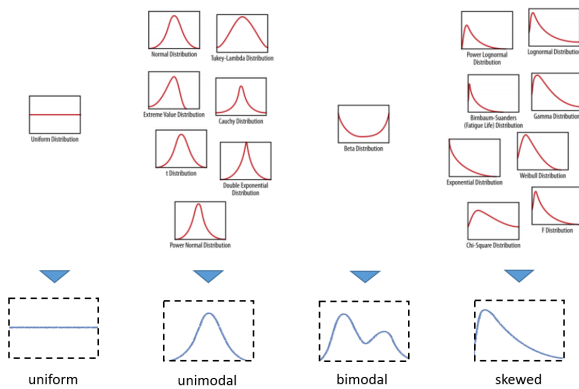


Fig. 2. Classification of density distributions. We decided to categorize distributions based on their main shapes (*uniform*, *unimodal*, *bimodal*, and *skewed*) and test these classifications in our study. This figure illustrates how the 17 density distributions described in *Doing Data Science* [36] fit into these classifications. No bimodal distribution was listed in this reference. Small images taken from [36].

This classification is also confirmed by Walker [29], who describes the most common shapes of distributions as *bell shaped*, *left skewed*, *right skewed*, *bimodal*, and *uniform*. In our study, we did not differentiate between left and right skewness but rather only considered distributions which are skewed to the left. After deciding on the four classes, we identified one mathematical density function representing each class best:

- For the class *uniform*, a uniform distribution fits best.
- For the class *unimodal*, we selected the normal density function to represent this class.
- For the class *bimodal* we joined two normal density functions to form a bimodal distribution with two peaks.
- For the class *skewed* we selected the gamma density function to represent this class.

Number of samples and bins: After selecting the distributions, we had to define the ranges for the number of samples and the number of bins. We applied a mathematical approach to calculate the four moments (mean, variance, skew, and kurtosis) for each of the distributions used to represent a class (uniform, normal, bimodal, and gamma). The combination of these four moments can uniquely identify a distribution’s shape.

For five different sample sizes (100, 1,000, 10,000, 100,000, and 1,000,000) we drew 1,000 times from the four distributions specified above and recorded the actual moments from each draw. We then created evenly spaced binnings in steps of 1 from 2 up to 100 bins for each draw. Using the bins’ centers as the outline of a new shape, we calculated its moments and compared them to the actual moments. Figure 3 shows the average error in percent across all draws up to 50 bins. The charts show that the error is large for small numbers of bins for all moments but reaches an almost constant rate above 10 bins. Adding more bins does not affect the error rate anymore.

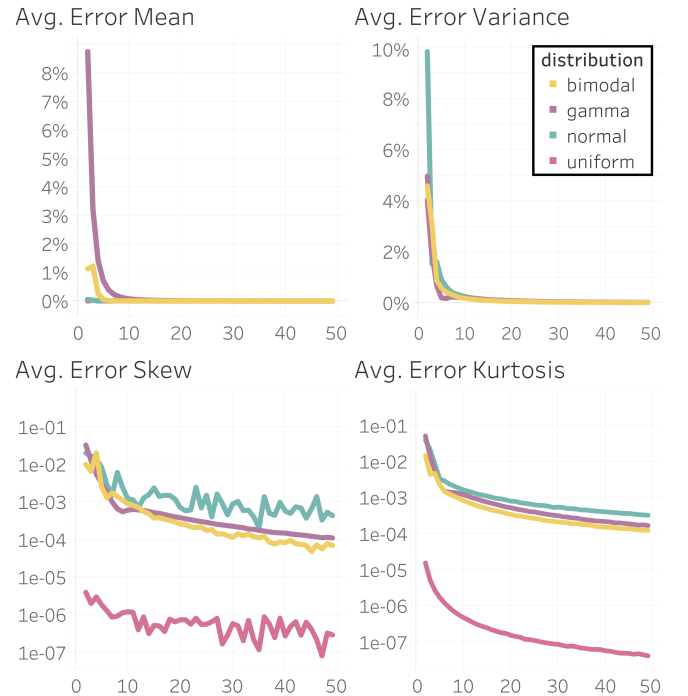


Fig. 3. Error made by binning. The charts show the average error in percent between the actual moments the moments of the different binning options. The *x*-axis shows the number of bins in all charts, the *y*-axis shows the error in percent. Please note that the *y*-axes of skew and kurtosis are logarithmic.

3.3 Quantitative User Study

We, therefore, decided that the range up to 10 bins would be of the greatest interest for our study. Since some moments (especially kurtosis) take a little longer to settle completely, we chose to include some values between 10 and 40 as well. We also included 100 bins as an upper boundary since it is equal to our study’s smallest sample size. Concluding from this investigation we decided on testing ten different bin counts (within the range [2–100]) and four sample sizes (within the range [100–1,000,000]). The sample size 100,000 was left out to reduce the overall number of combinations for the study—and also because the errors with 100,000 samples were almost identical to 10,000 and 1,000,000.

Hypotheses generation: The aim of the study was to test the effect of binning on the recognition of distributions in a histogram. To summarize, we decided to test

- four distributions (uniform, normal, bimodal, and gamma), with
- ten different bin counts (2, 3, 4, 5, 7, 10, 15, 20, 40, and 100), and
- four sample sizes (100, 1,000, 10,000, and 1,000,000).

This data collection and setting allows us to study the effect of binning under different sample sizes and with different distributions. We agreed upon testing the following hypotheses:

- **Hypothesis H1:** The number of bins influences how well humans can perceive the underlying data distribution in a histogram.
- **Hypothesis H2:** Upon a certain number of bins, adding new bins does not improve the perception of underlying data distributions in a histogram.

3.2 Data generation

Given the four distributions, ten bin counts, and four sample sizes, our complete test data consisted of 160 different parameter combinations. The data for the four different types of distributions was generated in the following way:

- *Uniform:* For generating datasets with a uniform distribution, we generated random integers between 0 and 10,000.
- *Normal:* For creating datasets with a normal distribution, we used a location-parameter of 0 and a scale of 1.
- *Bimodal:* For generating datasets with a bimodal distribution, we used two normal distributions. The first normal distribution was identical to the one mentioned previously, and the second normal distribution varied in scale between 0.5 and 1. We randomly placed the centers within 1.5 and 2 standard deviations (both distributions combined) apart. This placement ensured that the centers did not overlap and that the distributions also never entirely separated. Finally, we chose a random proportion between 0.3 and 0.7 to combine both distributions’ data points.
- *Gamma:* The shape and scale parameters for generating datasets with a gamma distribution were 2.

It has to be noted that we did not store and visualize the raw samples but computed the binning beforehand. We used Python NumPy [19] to generate the distributions. We picked random samples and split the data into bins for the histograms. To ensure reproducibility, we seeded each dataset individually. The final datasets containing the binning information were stored as JSON files.

3.3 Study design

We used the approach of a web-based questionnaire to be able to reach a large group of participants [33]. Participants could start the survey by accessing a website on one of our servers and then clicking a button to start the survey. Through the website, participants could inform themselves about the data protection regulations and could see contact details in case they wanted to ask further questions. A Cross-Site Request Forgery (CSRF) token was generated whenever a participant decided to start the survey. Since we then used only this token to identify the participant, the study was fully anonymized without any possibility to track the results back to the participants.

The study was implemented in Angular [17] version 10, and we used D3.js [8] to render the histograms. Since the bin information was stored instead of the raw data (see Section 3.2), it was not necessary to compute the binning online, which allowed us to render the histograms quickly. Participants did not encounter any delay when loading the questions. The Bootstrap [6] library was used for stylizing the buttons and control elements. With the built-in Typescript functionalities, we measured task completion time. The study results were sent via a POST command to the server and stored as a JSON file.

The study consisted of four parts:

1. Explanation

In the beginning, we gave a short explanation that the study will be about histograms, that different versions of histograms with a different number of bins will be shown, and that we ask the participants to judge the underlying data distribution.

2. Participant data

Afterwards, we collected statistical data about the participants, namely age group, residence, education, profession, and experience with visualizations.

3. Sanity check

As a next step, we integrated a sanity check to filter out careless participants [30]. On the next two pages, we showed a histogram to the participants where we also stated the sample size n . More specifically, a sample size of $n = 100,000$ was shown in the histogram visualization. On the first page we showed the histogram image and asked the participants

“What does n stand for in the above image?”.

Participants could select one out of five answers:

- *Sample size*
- *Number of bars*
- *Maximum number shown*
- *Statistical significance*
- *I don’t know.*

On the second page we showed the same histogram again and asked the question

“What is the sample size (= number of shown data points) in the above image?”

Again, participants could select one out of five answers:

- *1,000*
- *10,000*
- *100,000*
- *1,000,000*
- *I don’t know.*

Participants could only proceed with the questionnaire when they answered both questions of the sanity check.

4. Questionnaire

In the next step, we showed 20 histograms to every participant, one after the other. The histograms were randomly selected from the pool of datasets (see Section 3.2). The histogram plots had two axes with ticks, but we did not show any numbers or scales. Participants were asked to answer the question

“1. Choose the distribution which resembles the image above most closely”

by clicking on one of the icons below the histogram showing different possible distributions. Participants were also asked to state

“2. How confident are you about your answer?”

on a four-point Likert scale. Participants could only proceed with the next histogram if they answered both questions. An example of how the web-based implementation of such a histogram question looked like is shown in Figure 4.

5. Results

After judging 20 histograms, the answers of every participant were stored as a JSON file on the server. As a reward, participants had the chance to review their own performance during the study, where we showed them all 20 histograms again and depicted whether their answer was correct or not.

3.4 Pilot study

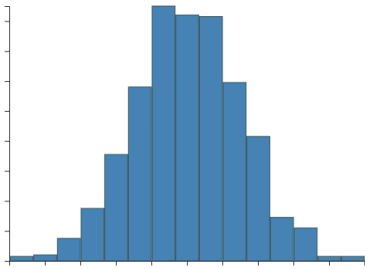
Before starting the larger user study, we started with a pilot study with 10 participants to evaluate our system. Several steps of the final study design (see Section 3.3) have been adapted after we received the comments and feedback from the pilot study. The main suggestions of the pilot participants, and our own suggestions for improvement were:

- To add an initial explanation at the beginning about the scope of the study,
- To revise the answers for *profession*,
- To add an *I don't know* option to both sanity check questions,
- To make options clickable as labels so that participants do not have to search for the radio button boxes,
- To not vertically align distribution icons and confidence boxes, since this caused misunderstanding through the alignment, and
- To add the possibility for the participants to review their own results.


The pilot study participants did not encounter difficulties in understanding the experiment nor the questions. They also did not experience any technical difficulties. There were no delays in loading and rendering the histograms. They also found that judging 20 histograms is within a acceptable time frame.

Question 9 of 23

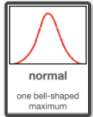
n = 1000



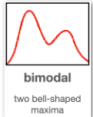
1. Choose the distribution which resembles the image above most closely:




uniform
even across
all values



normal
one bell-shaped
maximum



bimodal
two bell-shaped
maxima



gamma
one directionally
skewed maximum

2. How confident are you about your answer?

very unconfident

slightly unconfident

confident

very confident

Fig. 4. Study question. Participants were shown a histogram depicting the data's underlying distribution and asked to click the appropriate icon. Participants were also asked to state how confident they are about their answers. In this example *normal* and *confident* have been selected.

4 RESULTS

After the pilot phase, we distributed the user study to a larger group of possible participants. In total, 82 participants finished the user study within a 14 day time frame. We only counted complete submissions and did not record the dropouts. Based on the sanity check questions at the beginning (see Section 3.3), we had to exclude 10 data records from the evaluation, which led to a final number of 72 valid submissions. It took participants 44 minutes, on average, to complete the study. Participants could only proceed if they answered all questions, and only completed surveys were stored on the server, so we received 20 complete answers from all 72 participant.

4.1 Study participants

At the beginning of the study we collected statistical information about the participants, which is summarized in Figure 5. The majority of our participants were between 20 and 49 years old. One-third of the participants (33%) were bachelor's, master's, or PhD students. The other participants were working part-time (14%), full-time (23%), or, more specifically, in research and education (26%). Since we used our established networks to broadcast the study, most of our participants were inhabitants of the region A (*name removed for review*). The majority of our participants already had prior experience with data visualization. 27% of the participants had some experience in reading charts and plots in the media. 27% classified themselves as being experienced in reading data visualizations, and 41% stated that they are also creating data visualizations themselves. Only four participants stated that they do not have any experience with data visualization. Therefore, the focus of our study was on the literacy of histograms for people who already have experience with visual data representations.

4.2 Study results

In analyzing the results, we could identify significant differences in recognizing the underlying data distribution when comparing the results for different sample sizes and number of bins. The experience of participants with visualization did not have much impact, but influenced the confidence of the participants when answering the questions. More specifically, the evaluation of the quantitative results led to the following results:

Insight 1: Small sample sizes generally make it harder to detect the underlying data distribution, which can only slightly be mitigated by using a higher number of bins.

We could identify significant differences in recognizing the underlying data distribution when comparing the results for different sample sizes (see Figure 6, left). For datasets with 100 samples, 35.4% of the answers were wrong. With 1,000, 10,000, and 1,000,000 samples being available, the detection error rate could be halved to 16.1%, 18.3%, and 18.8%. A Mann–Whitney U test [14] resulted in p-values $p < 0.001$ when comparing the results for all sample sizes, which confirms the statistical significance of the results. Participants stated to be less confident when judging the distribution with a sample size of 100 (see Figure 6, right). A Mann–Whitney U test results in p-values $p < 0.001$ when comparing datasets with sample sizes of 100 with the results for other sample sizes. The amount of participants being *very confident* about their answers constantly increases with a rate of about 10% for larger samples sizes (100: 12.4%, 1,000: 22.6%, 10,000: 33.1%, and 1,000,000: 43.5%). It is, therefore, confirmed, that with a small number of samples, participants have troubles recognizing the underlying data distribution.

Insight 2: Beyond a certain number of bins the error rate stays constant and is not improved by adding more bins.

Due to the random selection of datasets for all participants, we could ensure that we received approximately the same number of answers for all possible parameters (*bins: percentage of answers* – 2: 9.02%, 3: 9.63%, 4: 10.79%, 5: 8.72%, 7: 10.49%, 10: 9.94%, 15: 9.57%, 20: 9.33%, 40: 11.89%, 100: 10.61%). Like the sample size, the number of bins affected participants' ability to recognize the underlying distribution correctly. More bins result in fewer errors being made by the participants (see Figure 7, left). This effect is different, depending on how many samples are available.

3.4 Results

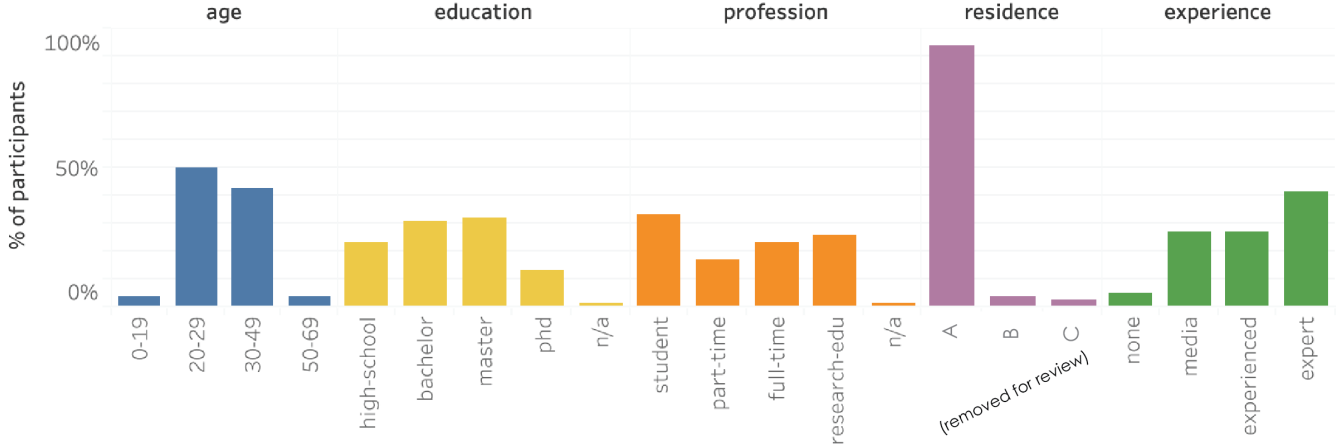


Fig. 5. Study participants. The participants were mainly in the age range from 20 to 49. The participants consisted of bachelor's, master's, PhD students, and people working part- and full-time, some of them in research and education. The majority of the participants were inhabitants of the region A (*removed for review*). Participants were, in general, experienced in using and reading visualizations.

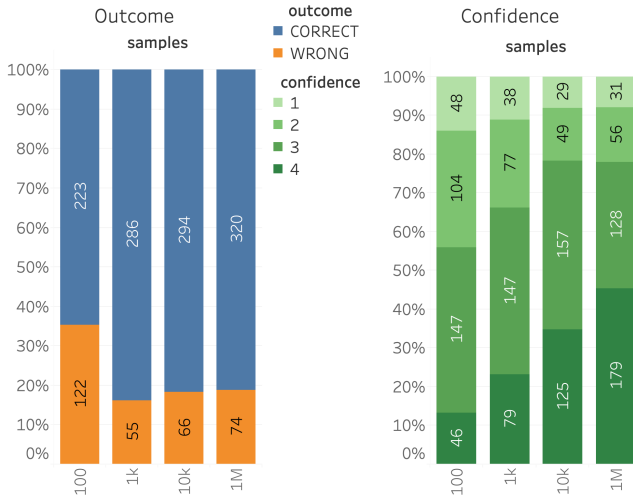


Fig. 6. Correct (blue) and wrong (orange) answers and confidence based on sample size. The percentage of wrong answers is especially high (35.4%) for a small number of samples (100). For a sample size of 100, participants were also rather unconfident in their answers.

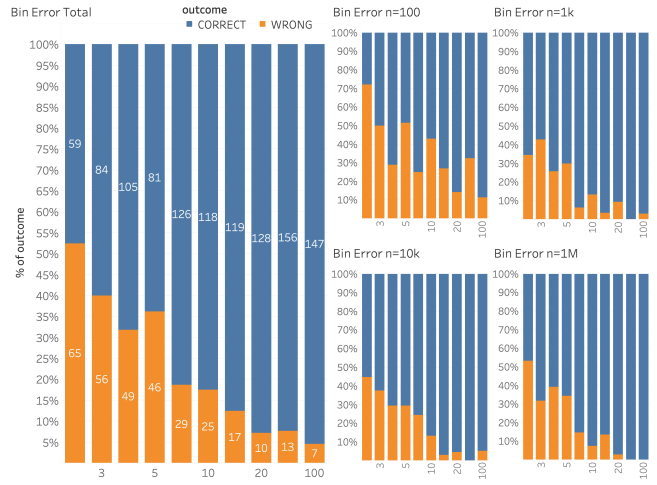


Fig. 7. Correct (blue) and wrong (orange) answers based on the number of bins and sample size. With only 100 samples, the data distribution recognition is generally challenging. For larger sample sizes, a larger number of bins increases the recognition of the correct distribution. However, beyond 20 bins, the detection rate does not increase significantly anymore.

The results in Figure 7 (right) show that in the case of 100 samples, the error rate stays rather high, also in cases where a higher number of bins was used. For other sample sizes, the error rate decreases in case a larger number of bins is used. For larger sample sizes, it can be seen that more bins do not improve the visual perception of the underlying data distribution. While the error rate is significantly better when comparing the bin size 2 to other parameters ($p < 0.001$), the difference between a larger number of bins is not significant any more (*bins/bins*: p -value – 15/20: $p = 0.072$, 20/40: $p = 0.442$, 40/100: $p = 0.121$).

Insight 3: For bimodal distributions the number of bins is more important to recognize them correctly.

The recognition of distributions under a different number of bins was not the same for all types of distributions. When looking more closely at the percentages of correct and wrong answers for every distribution in Figure 8, we can see that it is generally easier to detect gamma and uniform distributions. For normal distributions, only 2 bins are not enough to perceptually resemble the underlying distribution, but the detection was also quite successful with more bins. The detection of bimodal distributions is very strongly affected by the number of

bins. The bimodal distribution was the distribution that was most likely confused with another distribution by the participants. With a low number of bins, participants tended to confuse bimodal distributions with gamma or normal distributions (see Figure 9). The detection was easier the more bins were used. In the case of normal distributions, apart from 2 bins, the detection worked quite well. We can, therefore, conclude, that some distribution types are easier to detect than others.

Insight 4: Experience in reading data visualization had no impact on the error rate. More experience led to higher confidence when answering the questions.

Only minor, non-significant differences could be detected when analyzing the percentage of correct and wrong answers than the participants' stated experience with visualizations. We, however, have to note that the total count of participants without prior knowledge in our study was relatively small (4 out of 72). Significant differences between the participants' confidence when answering the questions and their stated experience could be identified. Participants with no or mediocre experience were generally less confident when answering the

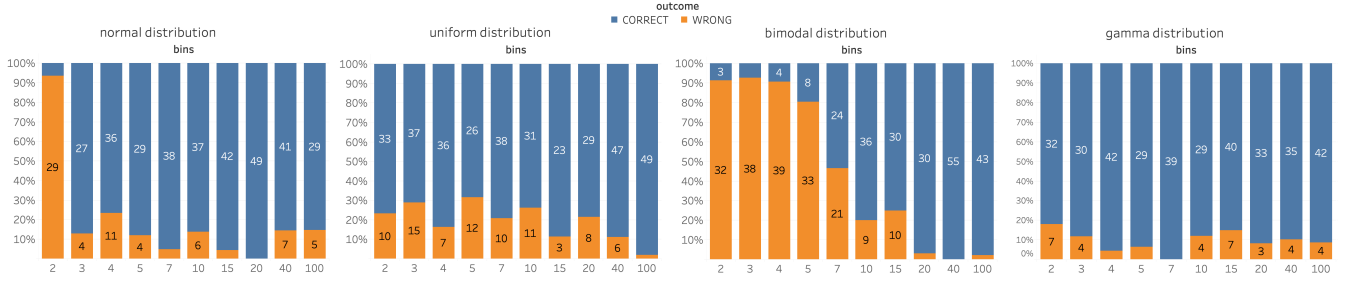


Fig. 8. Correct (blue) and wrong (orange) answers based on the number of bins and distribution. For the participants, it was generally easier to detect gamma and uniform distribution. The detection of normal distributions does not work in case only 2 bins are used. For the detection of bimodal distributions, a larger number of bins was needed.

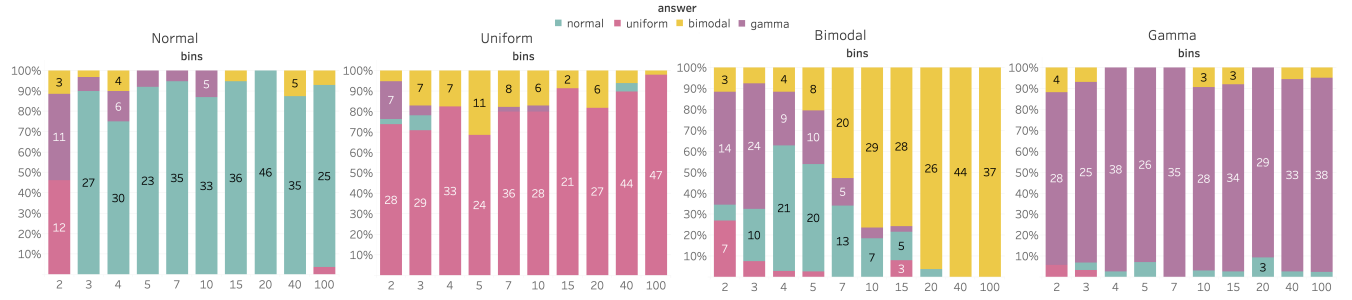


Fig. 9. Distribution confusions. Gamma and uniform distributions could be successfully detected for all numbers of bins. If 2 bins were used, normal distributions were confused with gamma or uniform distributions. Bimodal distributions were the most challenging ones to detect and were mostly confused with gamma and normal distributions for low bin counts.

questions than those who had extensive data visualization knowledge. Timings affected the error rate. If participants answered the questions very quickly, the answers were wrong in many cases. The timings for different bins and sample sizes were almost equal, but participants were a little bit faster in recognizing gamma and normal distributions.

4.3 Hypotheses testing

In Section 3.1 we defined two hypotheses to be tested in the study. We can summarize the results in the following way:

- **Hypothesis H1:** We can *partially confirm* that the visual perception of the underlying data distribution depends on the number of bins. At least with larger sample sizes (enough samples to resemble the underlying distribution), the recognition becomes better when using more bins. However, it seems that around 20 bins are enough for humans to detect the underlying data distribution.
- **Hypothesis H2:** We can *confirm* that upon a certain number of bins, adding new bins does not improve the perception of the underlying data distribution. After 20 bins, the error rate cannot be decreased significantly by adding additional bins.

5 IMPACT AND DISCUSSION

This paper described a web-based user study to evaluate how well users can detect underlying data distributions in a histogram. The main focus of the user study was on “describing and identifying the shape and type of one distribution” [3]. In this Section we will discuss the results and their impact on visualization design.

5.1 Binning with a focus on human perception

The study results confirm that adding additional bins beyond a certain number of bins does not further decrease the error rate for a human viewer. This means that the error rate is not fully correlated linearly with the number of bins. The difference of the error rate between 100 and 10 bins is still significant ($p < 0.001$), also when comparing 100 and 15 bins ($p = 0.007$), but not any more when comparing 100 and

Table 1. Mathematical models. In this table the number of bins as suggested by the mathematical models are listed, according to the number of samples in the dataset.

samples	Sturge's formula	Rice Rule	Scott's normal	Freedman-Diaconis
100	8	9	14	18
1,000	11	20	29	38
10,000	15	43	62	80
1,000,000	21	200	287	371

20 bins ($p = 0.163$). This tells us that from 20 bins on, the detection rate for distributions in a histogram reaches a stable state, even when more bins are added.

Therefore, we propose a new perception-based strategy for histogram binnings. A minimum of around 20 bins are needed for human viewers to detect the underlying data distribution. More bins can be added, but this does not improve the visual perception for human readers (as long as we have more samples than bins). It has to be noted that this binning suggestion works without considering the sample size of the data—we, however, have to assume that the number of samples in the data is enough to represent the underlying data distribution. We showed that binning does not improve visual perception for too small sample sizes.

To evaluate our rule of 20 bins, we compare it with other binning suggestions presented in the literature. We chose Sturges' formula, the Rice Rule, Scott's normal reference rule, and Freedman-Diaconis' choice since these rules are primarily used in many visualization applications and libraries. The results are displayed in Table 1.

Sturge's formula is closest to what has been identified as a minimum number of bins needed for humans to correctly interpret a histogram—although the number of bins for small sample sizes is lower. All other models overestimate the number of bins that are needed for sample sizes above 10,000. However, as stated above, adding more bins does not increase the error rate for human viewers. The information of a

possible minimum of around 20 bins can, nevertheless, be considered an important guideline for visualizations. On devices with limited display capacities (e.g., smartphones or tablets), where displaying more than 200 bins (as suggested by the Rice Rule, Scott’s normal and Freedman-Diaconis choice) might be problematic.

5.2 Sampling theory

After reviewing the study results and comparing them to the mathematical models, we also asked ourselves how this might be related to a sampling problem and how the results can probably also be explained by sampling theory. Binning can also be seen as a way to sample the original distribution. As known from sampling theory, it is impossible to reconstruct the original function with too few samples. We, therefore, wanted to evaluate our binning suggestions derived from the user study under this respect.

We transferred the four density distributions we used (uniform, normal, bimodal, and gamma) into the frequency domain. None of these is band-limited, hence, there is no concrete Nyquist frequency. As a baseline we, therefore, used a representation of each density distribution with 1,000,000 uniform samples. We then compared this baseline to representations of the same density distribution with less samples (5, 10, 15, 20, 40, and 100) – the samples in this case represent binnings. For a comparison we measured the deviation (i.e., error) between the baseline and the binned representation in the frequency domain.

The results can be seen in Figure 10. The charts show the frequency from 0 to π on the x-axis and the error (difference to the baseline) on the y-axis. In all graphs, the line with only five samples (i.e., bins) stands out to have the most significant error. From ten samples (i.e., bins) on, the error starts to converge to zero. 100 samples (i.e., bins) already ensure an error to the original representation very close to zero. It is important to note that the difference between five and ten samples (i.e., bins) is much more significant than the difference between 40 and 100. From a sampling theory perspective, adding more bins to a representation with only five bins reduces the error much faster than adding bins to a representation with, for example, 40 bins.

The comparison of the representations in the frequency domain is very similar to the results we got in our user study. Here, we also concluded that adding more bins from a threshold of about 20 bins does not further improve the user’s perception of the histogram’s underlying data distribution. It can also be seen that a number of bins of around 20 results in a small enough error also in the frequency domain.

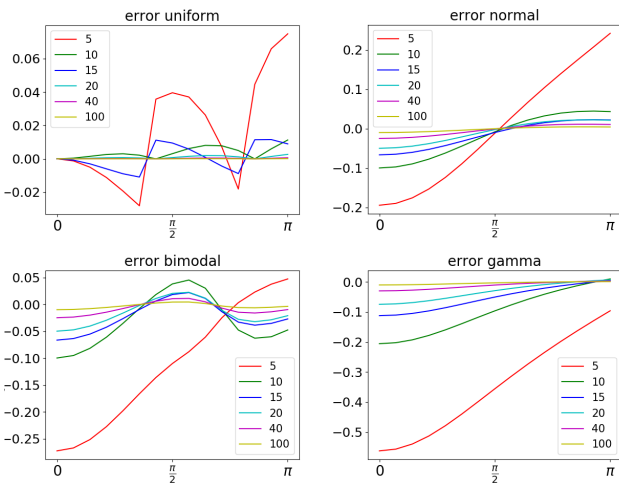


Fig. 10. Errors due to binning in the frequency domain. The charts show the errors between a baseline representation of the distribution with 1,000,000 samples and binned representations. The x-axis shows the frequency $[0-\pi]$ in all charts, and the y-axis shows the difference error to the baseline. If one representation for one frequency is below the original frequency, the error becomes negative.

5.3 Discussion

We made the following observations upon reflecting our results.

- Minimum of 20-40** Our new, perception-based, binning rule focuses on the range of 20 – 40 bins, since this is the number we tested in our study. Hence, the optimal number of bins, may be located within this interval. For a further study we would like to test different bin sizes within this interval, to get more detailed information on the error rate. Literature suggests that there exists a right answer to all questions [1], and we would, therefore, like to find out the right answer to histogram binnings in the future.
- Outliers** We did not specifically include outliers in our datasets. The reason for not considering outliers was to keep the set of parameters to be tested in the study within a reasonable complexity, as this was the first study on the depiction of data distributions in histograms for various bin sizes. It has to be noted, though, that the presence of outliers in the data should affect the choice of binning, as outliers may lead to binning artifacts [44].
- Parameters** Our study tested four different distributions: uniform, normal, bimodal, and gamma. These distributions were chosen since these are among the most commonly used distributions in data science [36], and since these four best represent the possible distribution shapes [29] that can be depicted in a histogram. We think it would be challenging to differentiate, for example, a normal distribution from a Student’s t distribution or a Chi-squared distribution from a gamma distribution, especially if only very few samples are available. We tested the number of bins between 2 and 100, as we wanted to see if positive results can be achieved with a low number of bins. Our datasets had sample sizes between 100 and 1,000,000, since we think that a histogram representation might be a useful visualization technique for distributions. For sample sizes smaller than 100, other visualization techniques (e.g., 1D scatter plots) might be more useful. We do not expect to get different results for sample sizes greater than 1,000,000.
- Experience** The majority of our study participants had mediocre to advanced knowledge in data visualization (54%), and the other major group considered themselves experts (41%). It was intended to study participants with prior knowledge in visualization and statistics to ensure that participants have basic knowledge about data analysis and data distributions. In terms of correctness of the answers, we did not find significant differences between the participants with mediocre to advanced knowledge and experts ($p = 0.079$), and not between the participants with and without any experience ($p = 0.198$). There was, however, a significant difference between participants with no experience and others in terms of confidence in their answers ($p < 0.001$). The group of participants with no experience in our study, however, was very small. Prior studies indicate that when looking at the general public, people have problems interpreting histograms and understanding data aggregation [4], which primarily comes from a lack of knowledge in data analysis and statistics. Therefore, we think that more studies on the literacy and interpretation of aggregated data would be important in the future (see also Börner et al. [7]), for a more general understanding of histogram literacy.
- Web-based setting** Through the web-based setting of our study, we were able to distribute the study to a larger group of participants. It is, though, not possible to enforce a controlled environment in a web-based study. Since a very neutral and easy-to-read representation of the histograms was chosen, we do not consider different monitor settings and lighting conditions as important parameters that might influence the results. The website was implemented in a responsive way; however, we encouraged participants to use a larger screen like a computer monitor. Distractions were always possible, and we also did not motivate the participants to finish the questions as fast as possible. This is also reflected by the

task completion times, ranging between 5s and 9m for answering one histogram question. We recorded these timings, but we have not given it so much importance due to the web-based setting.

- **Salient features** Since the study was web-based and, therefore, done in an uncontrolled setting, we did not collect any information on salient features in a histogram (e.g., via eye-tracking). This information would be beneficial to understand better how humans read and interpret histogram information. First attempts into this direction revealed that users mostly look at the axes and the average values in the plot, but not so much at the peaks [5], which is definitely of relevance when trying to depict a distribution. We also do not have evidence whether the participants used the sample size information n we gave for every plot.

6 CONCLUSION

In this paper, we presented a quantitative evaluation of different histogram binnings. We conducted a web-based user study with 82 (72 valid) participants, where we asked every participant to judge the underlying data distribution in 20 histograms. The study results show that participants had difficulties recognizing the distribution if only a few samples were available. These problems can also not be mitigated by using a higher number of bins. In case more samples are available, the recognition rate is increased by using more bins. However, with around 20 bins, the error rate becomes stable and does not improve by adding more bins. We conclude from this study that around 20 bins are sufficient for human viewers to detect the underlying data distribution, in case enough samples are provided to resemble the underlying distribution properly. 20 bins are less than what is suggested by the commonly used mathematical models for histogram binning. The mathematical models (e.g., Scott's normal reference rule, the Rice Rule, Freedman-Diaconis' choice) mostly overestimate the number of bins necessary for a correct perception for human viewers.

For future work, we would like to study the number of bins between 20 and 40 in more detail. We also consider including outliers in the data as the next step towards a new study. Repeating the experiment with more variations (e.g., scaled distributions) would also be very insightful. More participants with only minor experience in data visualization would be beneficial. In general, we consider further studies on histogram literacy a very valuable direction for future work.

ACKNOWLEDGMENTS

Removed for review

REFERENCES

- [1] D. Adams. *The Hitchhiker's Guide to the Galaxy*. Pan Books, 1979.
- [2] L. Birge and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10:24–45, 2006. doi: 10.1051/ps:2006001
- [3] M. Blumenschein, L. J. Debbeler, N. C. Lages, B. Renner, D. A. Keim, and M. El-Assady. v-plots: Designing Hybrid Charts for the Comparative Analysis of Data Distributions. *Computer Graphics Forum*, 39(3):565–577, 2020. doi: 10.1111/cgf.14002
- [4] L. Boels, A. Bakker, W. Van Dooren, and P. Drijvers. Conceptual difficulties when interpreting histograms: A review. *Educational Research Review*, 28:100291, 2019. doi: 10.1016/j.edurev.2019.100291
- [5] L. Boels, R. Ebbes, A. Bakker, W. Van Dooren, and P. Drijvers. Revealing Conceptual Difficulties when Interpreting Histograms: An Eye-Tracking Study. In *Proceedings of the 10th International Conference on Teaching Statistics*, ICOTS '10. International Statistics Institute, Kyoto, Japan, July 2018.
- [6] Bootstrap. Build fast, responsive sites with Bootstrap. <https://getbootstrap.com/>, 2020. last accessed [2020-11-28].
- [7] K. Börner, A. Bueckle, and M. Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, 116(6):1857–1864, 2019. doi: 10.1073/pnas.1807180116
- [8] M. Bostock. D3.js - Data-Driven Documents. <https://d3js.org/>, 2020. last accessed [2020-11-28].
- [9] S. Carpendale. *Evaluating Information Visualizations*, pp. 19–45. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-70956-5_2
- [10] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks Good To Me: Visualizations As Sanity Checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):830–839, 2019. doi: 10.1109/TVCG.2018.2864907
- [11] M. Dabos. A glimpse of two year college instructors' understanding of variation in histograms. In *Proceedings of the 9th international conference on teaching statistics*, ICOTS9, pp. 1–4. Flagstaff, AZ, USA, July 13–18 2014.
- [12] D. P. Doane. Aesthetic Frequency Classifications. *The American Statistician*, 30(4):181–183, 1976. doi: 10.2307/2683757
- [13] G. Ellis and A. Dix. An Explorative Analysis of User Evaluation Studies in Information Visualisation. In *Proceedings of the AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, pp. 1–7. Association for Computing Machinery, Venice, Italy, May 23 2006. doi: 10.1145/1168149.1168152
- [14] A. Field and G. Hole. *How to Design and Report Experiments*. SAGE Publications, 2003.
- [15] C. Forsell. A Guide to Scientific Evaluation in Information Visualization. In *Proceedings of the 14th International Conference Information Visualisation*, IV '10, pp. 162–169. London, UK, July 26–29 2010. doi: 10.1109/IV.2010.33
- [16] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981. doi: 10.1007/BF01025868
- [17] Google. Angular. <https://angular.io/>, 2020. last accessed [2020-11-28].
- [18] P. Hall. Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probability Theory and Related Fields*, 85(4):449–467, 1990. doi: 10.1007/BF01203164
- [19] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. G'érard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2
- [20] K. He and G. Meeden. Selecting the number of bins in a histogram: A decision theoretic approach. *Journal of Statistical Planning and Inference*, 61(1):49–59, 1997. doi: 10.1016/S0378-3758(96)00142-5
- [21] C. Heinrich. On the number of bins in a rank histogram. *Quarterly Journal of the Royal Meteorological Society*, 2020. doi: 10.1002/qj.3932
- [22] W. Huang. *Handbook of Human Centric Visualization*. Springer New York, 2014.
- [23] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2012. doi: 10.1109/TVCG.2013.126
- [24] J. J. Kaplan, J. G. Gabrosek, P. Curtiss, and C. Malone. Investigating Student Understanding of Histograms. *Journal of Statistics Education*, 22(2), 2014. doi: 10.1080/10691898.2014.11889701
- [25] K. H. Knuth. Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing*, 95:102581, 2019. doi: 10.1016/j.dsp.2019.102581
- [26] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. User Studies: Why, How, and When? *IEEE Computer Graphics and Applications*, 23(4):20–25, 2003. doi: 10.1109/MCG.2003.1210860
- [27] S. Lem, P. Onghena, L. Verschaffel, and W. Van Dooren. On the misinterpretation of histograms and box plots. *Educational Psychology*, 33(2):155–174, 2013. doi: 10.1080/01443410.2012.674006
- [28] S. V. G. Lolla and L. L. Hoberock. On Selecting the Number of Bins for a Histogram. In *Proceedings of the International Conference on Data Mining*, DMIN '11, 2011.
- [29] MathBootCamps. Common shapes of distributions. <https://www.mathbootcamps.com/common-shapes-of-distributions/>, 2017. last accessed [2020-11-29].
- [30] A. S. M. Niessen, R. R. Meijer, and J. N. Tendeiro. Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63:1–11, 2016. doi: 10.1016/j.jrp.2016.04.010
- [31] R. L. Nuzzo. Histograms: A Useful Data Analysis Visualization. *PM&R*, 11(3):309–312, 2019. doi: 10.1002/pmrj.12145

- [32] S. Owen. Common Probability Distributions: The Data Scientist's Crib Sheet. <https://medium.com/@srowen/common-probability-distributions-347e6b945ce4>, 2018. last accessed [2021-03-15].
- [33] U.-D. Reips. *The methodology of Internet-based experiments*, pp. 373–390. Oxford University Press, 2007. doi: 10.1093/oxfordhb/9780199561803.013.0024
- [34] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982. doi: 10.2307/4615859
- [35] J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang. Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations. *The Astrophysical Journal*, 764(2):167, 2013. doi: 10.1088/0004-637x/764/2/167
- [36] R. Schutt and C. O’Neil. *Doing Data Science*. O’Reilly Media, Inc., 2013.
- [37] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979. doi: 10.1093/biomet/66.3.605
- [38] D. W. Scott. *Sturges’ and Scott’s Rules*, pp. 1563–1566. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-04898-2_578
- [39] H. Shimazaki and S. Shinomoto. A method for selecting the bin size of a time histogram. *Neural computation*, 19(6):1503–1527, 2007. doi: 10.1162/neco.2007.19.6.1503
- [40] C. J. Stone. An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984. doi: 10.2307/2241002
- [41] L. Strika. 5 Probability Distributions Every Data Scientist Should Know (KDnuggets, Tutorials, Overviews). <https://www.kdnuggets.com/2019/07/5-probability-distributions-every-data-scientist-should-know.html>, 2019. last accessed [2021-03-17].
- [42] H. A. Sturges. The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926. doi: 10.1080/01621459.1926.10502161
- [43] G. R. Terrell and D. W. Scott. Oversmoothed Nonparametric Density Estimates. *Journal of the American Statistical Association*, 80(389):209–214, 1985. doi: 10.2307/2288074
- [44] E. Thoen. Binning Outliers in a Histogram. <https://edwinth.github.io/blog/outlier-bin/>, 2017. last accessed [2020-11-29].
- [45] M. Tory. *User Studies in Visualization: A Reflection on Methods*, pp. 411–426. Springer New York, 2014. doi: 10.1007/978-1-4614-7485-2_16
- [46] M. Tory and T. Möller. Human Factors In Visualization Research. *IEEE Transactions on Visualization and Computer Graphics*, 10(1):72–84, 2004. doi: 10.1109/TVCG.2004.1260759
- [47] M. P. Wand. Data-Based Choice of Histogram Bin Width. *The American Statistician*, 51(1):59–64, 1997. doi: 10.2307/2684697
- [48] A. Zubiaga and B. M. Namee. Graphical Perception of Value Distributions: An Evaluation of Non-Expert Viewers’ Data Literacy. *Journal of Community Informatics, Special Issue on Data Literacy*, 12(3), 2016.




4 Selective Angular Brushing of Parallel Coordinate Plots

Synopsis

The following chapter contains the contents of Raphael SAHANN, Ivana GAJIC, Torsten MÖLLER, and Johanna SCHMIDT *"Selective Angular Brushing of Parallel Coordinate Plots"*, presented at EuroVis (Short Papers) 2021 [SGMS21].

Ivana Gajic coded and evaluated the pilot study as part of her Bachelor's thesis. Johanna Schmidt and Raphael Sahann collaborated on the further evaluation and the coding of the released open-source version of the brushing method, including the range selection as an extension to large data sets. They also wrote all text and created all figures collaboratively. Torsten Möller helped with critical questions and feedback to the final text.

Selective Angular Brushing of Parallel Coordinate Plots

R. Sahann¹ , I. Gajic¹, T. Möller^{1,2} , J. Schmidt³ 

¹University of Vienna, Faculty of Computer Science, Austria

²Data Science @ Uni Vienna, Austria

³VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Austria

Abstract

Parallel coordinates are an established technique to visualize multivariate data. Since these graphs are generally hard to read, we need interaction techniques to judge them accurately. Adding to the existing brushing techniques used in parallel coordinate plots, we present a triangular selection that highlights lines with a single click-and-drag mouse motion. Our selection starts by clicking on an axis and dragging the mouse away to select different ranges of lines. The position of the mouse determines the angle and the scope of the selection. We refined the interaction by running and adapting our method in two small user studies and present the most intuitive version to use.

CCS Concepts

• **Human-centered computing** → **Visualization**; **Interaction techniques**;

1. Motivation

The visualization of multivariate data brings several challenges in terms of screen space and interaction. Parallel coordinates [Ins09] have become a well-known and increasingly used visualization technique for the visualization of multivariate data [JF16]. When using parallel coordinates in data analysis, interaction plays an important role [FL03] in enhancing analysis. Besides interacting with the axes (e.g., sorting), brushing is considered one of the most important interaction concepts in parallel coordinates plots. Brushing enables the users to select individual or multiple lines (i.e., data items) in a parallel coordinates plot. This is done by setting a *brush* (e.g., rectangle or lasso) and computing its intersection with the underlying data lines.

Multiple brushes [HW13] can be combined with logical operations (AND and OR) to further filter the data. For example, one could imagine defining two brushes—one for a specific range on axis x_1 and another for a specific range on axis x_2 . By combining them with a logical AND, the user will see all data defined within both ranges (i.e., all lines that “go through” both brushes). The definition of filter operations over multiple domains (i.e., axes) usually requires a sequence of different mouse interactions. In the previous example, users will have to define two brushes on two different axes, which require the same operation (click and drag) twice.

We present a novel method for selecting a subset of lines in a parallel coordinates plot. Our method allows users to specify the subset based on two axes but only one mouse operation. While a lasso selects all lines within the selected region, our techniques only selects lines in a particular region with a particular direction.

2. Related Work

Parallel coordinates are a well-studied technique [HW13], yet they are difficult to comprehend. Hence, several improvements have been suggested. For example, researchers proposed to apply illustrative representations [MM08] and focus+context techniques [NH06] to parallel coordinates to enhance perception. Density-based representations [HW09] and edge-bundling techniques [PBO*14] are proposed to deal with overplotting for large datasets. Effective interaction techniques, like selections, are crucial for users to deal with dense representations [Sii00].

In this paper, we specifically concentrate on brushing in parallel coordinates. Classical approaches include

- selecting values on an axis,
- range selection on axes, and
- brushing (e.g., using a line or lasso) in between axes.

In multi-dimensional settings, brushes can also be combined [MW95] to allow higher-order selections [War94, War97]. Multi-dimensional and higher-order brushing is an active research area. Here, researchers include additional guidance [RLS*19] to deal with high-density parallel coordinates [REB*16]. Another brushing technique is described by angular selection [HLD02] where users can define line subsets based on an angle at a specific point in the plot. Previous work suggests that interaction with parallel coordinates is intuitive and combining different methods for brushing can lead to increased performance for users [SR06]. In line with these findings, our new selection method can be seen as a combination of angular selection [HLD02] and multi-range selection [War97].

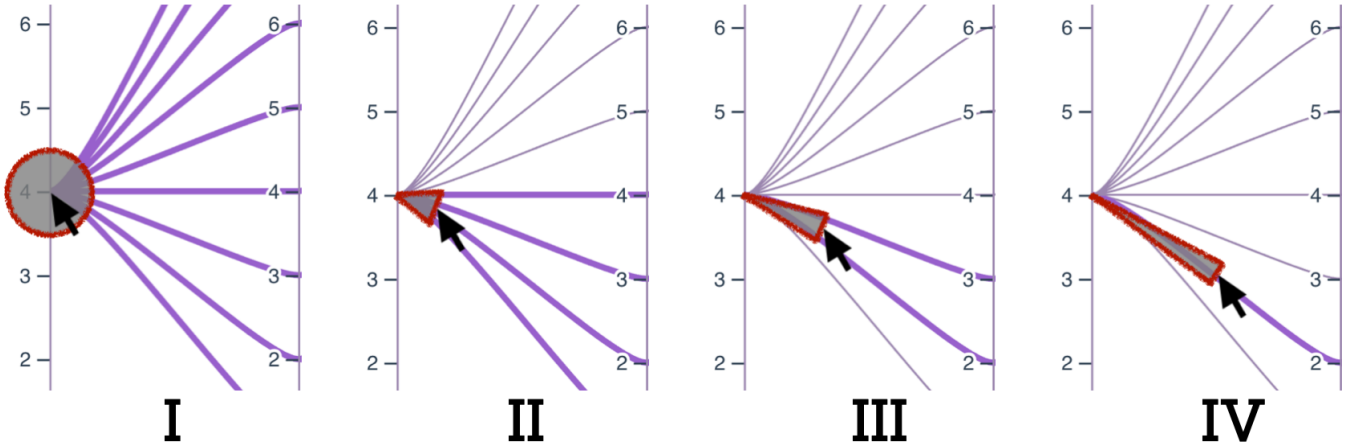


Figure 1: Selective Angular Brushing (basic concept). The selection method is shown in four steps—the red outlines have been added for clarity and are not part of the implementation. I) The interaction starts with a click on the axis, where all lines going through that point are highlighted. II) By dragging the mouse away from the axis, the circle selection changes to a triangle that highlights only lines beneath the triangle. III) Dragging the mouse even further narrows the region and in this way selects fewer lines. IV) Changing the angle is still possible anytime and also changes the selection.

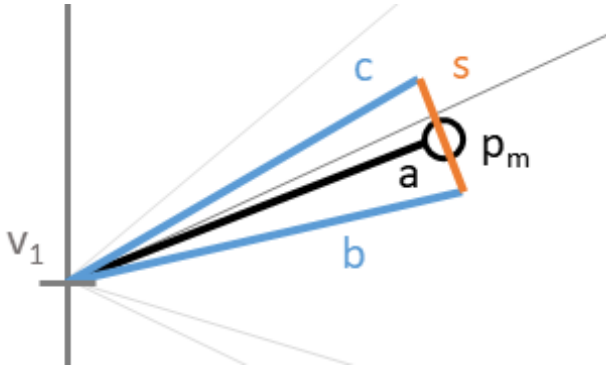


Figure 2: Basic concept. Starting from a value v_1 an imaginary line a can be drawn to the current mouse position p_m . Based on the angle between a and the axis, the lines b and c can be drawn. The endpoints of these lines define the line s , which is perpendicular to a and is located at the current mouse position.

3. Selection Design

Our novel method *Selective Angular Brushing* is based on two main concepts:

- Allow line selection based on the angle between two axes (similar to Hauser et al. [HLD02]).
- Enable users to do the selection with one single mouse operation.

The idea evolved out of working with data on student grades where the axes defined grades for specific lectures. In our case, users wanted to quickly verify whether students improved or deteriorated from one lecture to another. Basically, they wanted to verify “which lines move up from one specific point”, “which lines are parallel”, and “which lines go downwards”. Using the classical concept of multi-range selections, it was necessary to operate with

two ranges on two axes, which was tedious. Therefore, we came up with a new solution to answer these questions more quickly.

3.1. Basic Concept

Selective Angular Brushing is based on the mouse operation click-and-drag. The different steps of the selection process are outlined in Figure 1. Selective Angular Brushing starts with the user clicking on one axis in a parallel coordinates plot (step I). This initially selects all lines going through this point, shown by a circle centered on the current value. By dragging the mouse away in a certain direction, only the lines matching the current angle are selected (step II). The mouse marker, which was originally a circle, changes its shape to a triangle. Moving the mouse further away from the initial click shortens the triangle’s base, narrowing the selection (step III). An isosceles triangle with the apex at the initial click position and the center of the base at the mouse position shows the highlighted section. The mouse position can be changed at any time, which changes the angle to the original point and the triangle shape and, therefore, the selection (step IV). By now, releasing the mouse button applies the current selection, although other operations (e.g., filtering) may also be possible.

The main parameters for defining the selection are the distance and the angle between the axis and the current mouse position. Figure 2 shows the geometric concept. All data lines crossing the line s closest to the mouse position will be selected. In other words, it can be said that the triangle side s which is closest to the mouse pointer defines the selection interval. The whole selection area (consisting of the lines b , c , and s) is shown as an indicator. It should be noted, however, that the selection is not made by selecting all lines that somehow touch this area (compare step III in Figure 1).

The selection area is narrowed (i.e., the line s is shortened) the further the user moves the mouse away from the original click (compare step II and IV in Figure 1). When staying close to the

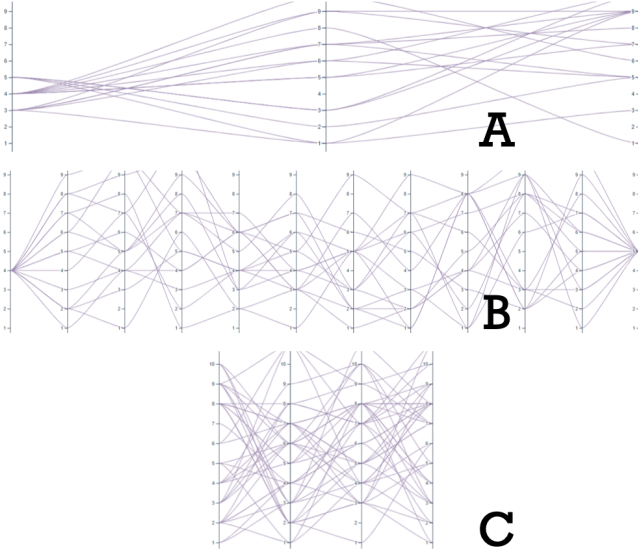


Figure 3: The three test datasets we used in the second evaluation, which were also used the first evaluation. A was previously called (viii) wide, B (ii) twelve axes, and C was referred to as (iv) 36 lines in the first evaluation. The focus of this evaluation was comparing the selection method with and without the triangular guide.

original click position, all lines going through this position are selected. When moving further away, only lines crossing the selection area stay selected. This means that more coarse selections are possible close to the original point, and more fine-grained selections are possible when moving further away. The change in the size of the selection area is up to the configuration done by the visualization designer and developer.

While this concept is easy to implement, we also wanted to evaluate whether Selective Angular Brushing is at the same time intuitive for users when interacting with parallel coordinates. We, therefore, conducted a user study on this (see Section 3.2). The results were promising and indicated that the concept is understandable even if the mouse markers (circle and triangle) are not shown.

3.2. Evaluation

Our evaluation of the intuitiveness of Selective Angular Brushing consisted of two steps. In our first evaluation we conducted ten interviews of users with at least some experience using parallel coordinate plots. Important to note, in this first iteration we did not include a visualization of the selection area, but rather just let the users use the click-and-drag mouse operation to highlight lines directly. We used nine different parallel coordinate plots with i. two axes, ii. twelve axes, iii. five lines, iv. 36 lines, a plot with three axes and 15 lines that was v. large, vi. small, vii. tight but high, viii. wide, and ix. a plot with actual student course data.

In individual digital interviews, which lasted approximately half an hour each, we tested whether selecting lines based on click-and-drag operations is intuitive to use. Participants were asked to share their screen. They received an initial explanation of the vi-

ualization and selection, and then interacted with the plots independently. No explicit tasks were given, so the testers were free to explore the interface and different data sets for up to 15 minutes. Afterward, we asked the users to rate the intuitiveness of using this technique in various parallel coordinate plots on a five-point Likert scale and if the plot's shape influenced their experience. Finally, we ended the interview with questions from the standard usability scale (SUS) [Bro13] test.

Our participants found the selection method generally intuitive to use, independent of size, shape, or number of lines (average 4.3 out of 5 points), but also reported, to some extent, that they would need the support of a technical person to be able to use the system (2.1 out of 5). Some users suggested that it would be more intuitive to use the system if some visual guidance would show the current selection. As a result, we adapted our implementation to include the triangular-shaped visual representation of selection area.

For the second evaluation, we recruited four participants, two of which already took part in the first round. Since the plots' size did not make a difference previously, we reduced the number of different charts to three, shown in Figure 3. Apart from the changes in charts the study setup remained identical to the first one. We also had the users compare the original version without visual guidance versus the selection area highlighting (see Figure 1). This evaluation scored slightly higher marks for usability when the sample included the visualization of the selection area. Interestingly, some users suggested that they liked using the visualization without the selection area being highlighted better—but only after they had a few minutes to use it with the visual guide to understanding how the mouse position is linked to the selected range.

We, therefore, conclude that Selective Angular Brushing is most intuitively used when the selection area is shown. In cases where the additional overlay might clutter the visualization, a short tutorial to familiarize the users with the selection method might be sufficient to use the selection without guidance.

3.3. Extension to axes range selections

When working with dense datasets, a lot of lines are shown in the plot, and then a simple click on an axis may not be enough for users to select the desired range of values. We, therefore, decided to extend Selective Angular Brushing to also work with range selectors [War97].

In this case, the selection with Selective Angular Brushing starts with a click on a previously set range selector placed on one of the plot axes. Again, a selection area is drawn when dragging the mouse away from the axes, and all lines following the same angle as the current selection are selected. Since we now start from a range selector, the selection area now is not a triangle any more, but rather a trapezoid. The area close to the mouse pointer becomes more narrow when moving away from the axis, so that more fine-grained selections are possible.

4. Implementation

We implemented two prototypes in web-based frameworks. To demonstrate the basic concept (as described in Section 3.1) we

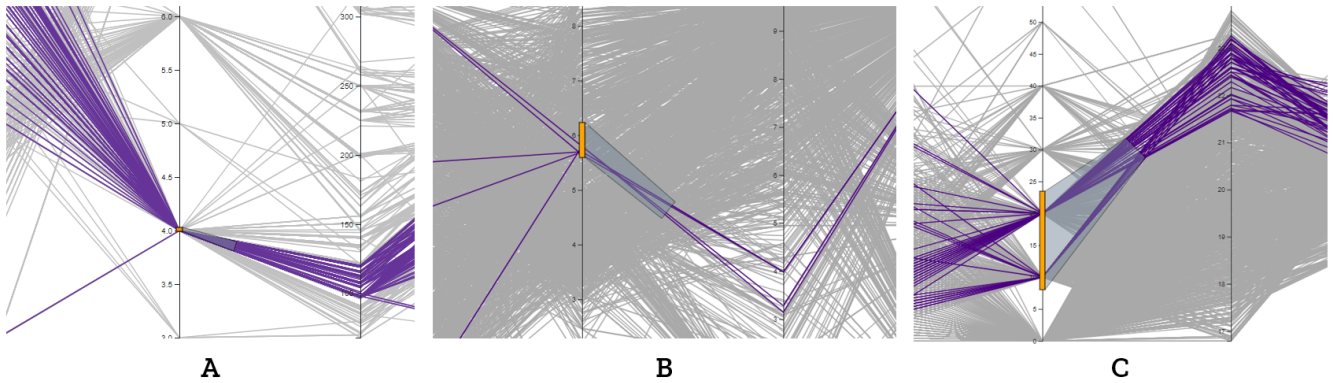


Figure 4: Selective Angular Brushing (range selectors). In case of dense data to be shown, range selection might be used instead of a single click on the axis. We could show that also in case of large datasets (A: 406 lines, B: 1,458 lines, C: 19,735 lines), Selective Angular Brushing provides an intuitive way to interact with the data.

used Vue.js [You21] and D3 [Bos21] version 6. This prototype was also used for the user evaluation (described in Section 3.2). We then implemented a second prototype to demonstrate the extension for range selections (described in Section 3.3) with D3 [Bos21] version 6 (for the interactive elements), WebGL [Khr21] (for the line rendering) and Angular [Goo21] version 11. The source code is available at <https://github.com/johanna-schmidt/selective-angular-brushing>.

5. Results

We tested Selective Angular Brushing with different datasets. The results can be seen in Figure 4. We used three different datasets to test our approach. In A a very small range selector is used, and therefore the selection is very similar to our initial point-based basic approach. The dataset used here contains specifications of cars from the 70s and 80s [Eva21] and consists of 406 lines. In B the range selector covers a larger area on the axis, which also affects the shape of the selection area. Even in a very dense data area, the data points with a specific angle can be nicely selected. In this case, we used data of the human freedom index [Kag21] which consisted of 1,458 lines. In C a large range selector is used, and again, with Selective Angular Brushing it is possible to extract the data items with specific characteristics based on the angle. Here we used a dataset consisting of different parameters for energy prediction [Kag18], which consisted of 19,735 lines. The results indicate that Selective Angular Brushing is a useful selection mechanism for both sparse and dense datasets.

Because of the simplicity of the interaction, there are a number of ideas on how Selective Angular Brushing can be extended and further evaluated:

- **Line rendering mechanisms** We would like to note that for our technique to work it is important that every polyline in the parallel coordinates plot touches its exact value at every axis. However, in other representations where, for example, smoothed curves [GK03] are used, it cannot be guaranteed that the curves touch the exact values. The initial click (basic concept) or a range selector might then not reach all intended lines.

- **Combinations** Selective Angular Brushing always starts at an axis, and could therefore easily be combined with a lasso brush—which can be applied whenever the click-and-drag begins in the space between two axes.
- **Scalability** We tested our approach with dense datasets (as seen in Figure 4). In case of many lines being drawn in the parallel coordinates plot, the performance of the technique depends on how fast the intersection of the selector and the lines can be calculated. In case of 19,735 lines, the calculation was still smooth in a JavaScript setting. In case larger datasets are used (e.g., more than 50,000 lines), parallel threads to compute the intersection, and faster drawing mechanisms (e.g., WebGL instead of SVG) are beneficial. The number of axes does not influence the performance, since only two axes are considered for calculating the selection.
- **Mobile devices** In the future we would like to explore the applicability of our approach on mobile devices with touch-and-drag functionalities.
- **Further evaluation** The user study we conducted was based on rather small and sparse datasets. In the future we would like to conduct further studies to evaluate the usefulness of our approach for the analysis of large datasets.
- **Other applications** Selective Angular Brushing works on line-based visualizations and could potentially also be useful for node-link diagrams and line charts.

6. Conclusion

This paper presents *Selective Angular Brushing*, a novel selection mechanisms for parallel coordinates based on a single click-and-drag mouse operation. Starting from a point or range on an axis, our technique allows to select all lines that follow a certain angle. This makes it possible to select all lines that “go up” or “go downwards” with one single operation, where previously multiple clicks (e.g., setting two ranges on two axes) were necessary. We conducted a user study to evaluate the intuitiveness of our approach. In the future, we would like to further study the applicability for mobile devices, and further evaluate the combination with other interaction and rendering techniques.

References

- [Bos21] BOSTOCK M.: D3.js - Data-Driven Documents . <https://d3js.org/>, 2021. [Accessed 2021-02-28]. 4
- [Bro13] BROOKE J.: Sus: A retrospective. *J. Usability Studies* 8, 2 (Feb. 2013), 29–40. 3
- [Eva21] EVANS C.: Parallel Coordinates. http://www.columbia.edu/~cme2126/datavisuals/bigdata_parallelcoordinates.html, 2021. [Accessed 2021-02-26]. 4
- [FL03] FERREIRA DE OLIVEIRA M., LEVKOWITZ H.: From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics* 9, 3 (2003), 378–394. doi:10.1109/TVCG.2003.1207445. 1
- [GK03] GRAHAM M., KENNEDY J.: Using curves to enhance parallel coordinate visualisations. In *Proceedings of the 7th International Conference on Information Visualization* (London, UK, July 16–18 2003), IV '03, pp. 10–16. doi:10.1109/IV.2003.1217950. 4
- [Goo21] GOOGLE I.: Angular - The modern web developer's platform. <https://angular.io/>, 2021. [Accessed 2021-02-28]. 4
- [HLD02] HAUSER H., LEDERMANN F., DOLEISCH H.: Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium on Information Visualization* (Boston, MA, USA, Oct. 28–29 2002), INFOVIS '02, pp. 127–130. doi:10.1109/INFVIS.2002.1173157. 1, 2
- [HW09] HEINRICH J., WEISKOPF D.: Continuous Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1531–1538. doi:10.1109/TVCG.2009.131. 1
- [HW13] HEINRICH J., WEISKOPF D.: State of the Art of Parallel Coordinates. In *STAR Proceedings of 34th Annual Conference of the European Association for Computer Graphics* (Girona, Spain, May 6–10 2013), Eurographics '13. doi:10.2312/conf/EG2013/stars/095–116. 1
- [Ins09] INSELBERG A.: *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer, 2009. 1
- [JF16] JOHANSSON J., FORSELL C.: Evaluation of parallel coordinates: Overview, categorization, and guidelines for future research. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 579–588. 1
- [Kag18] KAGGLE: Appliances Energy Prediction - Data driven prediction of energy use of appliances. <https://www.kaggle.com/loveall/appliances-energy-prediction>, 2018. [Accessed 2021-02-25]. 4
- [Kag21] KAGGLE: The Human Freedom Index - A global measurement of personal, civil, and economic freedom. <https://www.kaggle.com/gsutters/the-human-freedom-index>, 2021. [Accessed 2021-02-24]. 4
- [Khr21] KHRONOS M.: WebGL Overview. <https://www.khronos.org/webgl/>, 2021. [Accessed 2021-03-01]. 4
- [MM08] McDONNELL K., MUELLER K.: Illustrative parallel coordinates. *Computer Graphics Forum* 27, 3 (2008), 1031–1038. doi:10.1111/j.1467-8659.2008.01239.x. 1
- [MW95] MARTIN A., WARD M.: High dimensional brushing for interactive exploration of multivariate data. In *Short Paper Proceedings of the IEEE Visualization Conference* (Atlanta, GA, USA, Oct. 30–31 1995), VIS '95, pp. 271–278. doi:10.1109/visual.1995.485139. 1
- [NH06] NOVOTNÝ M., HAUSER H.: Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 893–900. doi:10.1109/TVCG.2006.170. 1
- [PBO*14] PALMAS G., BACHYNSKYI M., OULASVIRTA A., SEIDEL H., WEINKAUF T.: An Edge-Bundling Layout for Interactive Parallel Coordinates. In *Proceedings of the IEEE Pacific Visualization Symposium* (Yokohama, Japan, Mar. 4–7 2014), PACIFICVIS '14, pp. 57–64. doi:10.1109/PacificVis.2014.40. 1
- [REB*16] RAIDOU R., EISEMANN M., BREEUWER M., EISEMANN E., VILANOVA A.: Orientation-Enhanced Parallel Coordinate Plots. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 589–598. doi:10.1109/TVCG.2015.2467872. 1
- [RLS*19] ROBERTS R., LARAMEE R., SMITH G., BROOKES P., DCRUZE T.: Smart brushing for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 25, 3 (2019), 1575–1590. doi:10.1109/TVCG.2018.2808969. 1
- [Si00] SIIRTOLA H.: Direct manipulation of parallel coordinates. In *Proceedings of the Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, Apr. 1–6 2000), CHI '00, pp. 119–120. doi:10.1145/633292.633361. 1
- [SR06] SIIRTOLA H., RÄIHÄ K.: Interacting with parallel coordinates. *Interacting with Computers* 18, 6 (2006), 1278–1309. doi:10.1016/j.intcom.2006.03.006. 1
- [War94] WARD M.: XmdvTool: integrating multiple methods for visualizing multivariate data. In *Proceedings of the IEEE Conference on Visualization* (Washington, DC, USA, Oct. 17–21 1994), Visualization '94, pp. 326–333. doi:10.1109/VISUAL.1994.346302. 1
- [War97] WARD M.: Creating and Manipulating N-dimensional Brushes. In *Proceedings of the Joint Statistical Meeting* (Anaheim, CA, USA, Aug. 1997), ASA Proceedings, pp. 6–14. 1, 3
- [You21] YOU E.: Vue.js - The Progressive JavaScript Framework. <https://vuejs.org/>, 2021. [Accessed 2021-02-28]. 4

5 Designing a Semester Planner for Students (manuscript)

Synopsis

The following chapter contains the contents of Raphael SAHANN and Torsten MÖLLER *"Designing a Semester Planner for Students: Insights from Advancing a Concept into a Real-World Application"*. The presented version is a manuscript that will be submitted to the ACM SIGCHI 2022 conference on Computer-Human Interaction. The submission deadline is September 9, 2021.

Raphael Sahann did the design shown in this paper and the low-fidelity prototypes. Axel Sonntag and Torsten Möller helped to iterate on the initial design with helpful suggestions. The coding of the high-fidelity prototypes was done by Raphael Sahann together with Julian Gruber and Christoph Pressler. Finally, the real-world implementation of the tool was programmed by a SCRUM team from Cloudflight.io, for which Raphael acted as the *product owner*. Integration into the student portal at uspace.univie.ac.at was done by the department Coordination of Student Services at the University of Vienna. Raphael Sahann wrote all texts in the presented version. Finally, Torsten Möller gave feedback on the text.

Designing a Semester Planner for Students: Insights from Advancing a Concept into a Real-World Application

RAPHAEL SAHANN, Faculty of Computer Science, University of Vienna, Austria

TORSTEN MÖLLER, Faculty of Computer Science & Data Science @ Uni Vienna, University of Vienna, Austria

This paper reflects on a four-year design process of the creation of a semester planning tool available for more than ten thousand students. The project started as a simple research project based on findings of a behavioral sciences study but was transformed into a real-world implementation project after two years. First, we highlight design decisions and present the final design of the semester planning tool. Our qualitative and quantitative evaluations show users' acceptance of the tool in addition to positively affecting the course registration numbers. We compare the work in a classical research environment to collaborating with a SCRUM team of programmers and using methods from the SCRUM framework. Finally, we show how both research and SCRUM methods contributed to our work and give a recommendation on how combining both approaches might be beneficial for other research projects.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools*; • **Applied computing** → *Education*.

Additional Key Words and Phrases: planning, interface, SCRUM

ACM Reference Format:

Raphael Sahann and Torsten Möller. 2022. Designing a Semester Planner for Students: Insights from Advancing a Concept into a Real-World Application. In *CHI '22: ACM CHI Conference on Human Factors in Computing Systems, Apr 30–May 6, 2022, New Orleans, LA*. ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

A standard bachelor's degree at our university takes six semesters — three years — to complete. At least, that is how it is supposed to be. Data, however, show that only very few students finish their bachelor's degrees within six semesters.

There are no tuition fees in our country, so students do not have any immediate financial pressure to finish their studies in a timely fashion. Furthermore, there is no upper limit on the number of semesters to graduate, so students often work part-time while only studying part-time. Apart from a few mandatory prerequisites, students can freely choose which courses they want to take each semester. However, considerable freedom of choice while studying comes with the need for self-discipline and organizational skills. Providing an academic environment that minimizes completion time while maximizing completion rates is one of the key objectives of any higher education institution. A preceding study at our university looked into ways to enhance such an academic environment by testing various soft measures to promote study success in selected faculties. Internal documents of this yet-unpublished behavioral sciences study show that lack of guidance and experience in the semester planning process was an issue many students were experiencing. Based on these findings, we design and implement its most promising solution, a semester planning tool.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

Such a semester planner aggregates all information, opportunities, and constraints to enable students to plan their next semester.

In this work, we focus on three main research questions:

- (1) Can an online tool enhance the semester planning process?
- (2) Does such a planning tool affect student registration numbers?
- (3) Do the recommendations of the *Design Study Methodology* [30] and the *Multi-dimensional in-depth long-term case studies* [31] hold up to implementing production-ready tools using the *SCRUM Framework* [28]?

Thanks to the successful evaluation results of our initial research prototype, we were allowed to collaborate with a software development team, implement our prototype, and release it to the entire student body of the university. This enables us to reflect on the whole process from the conceptualization of the tool until its full release and gather experiences from several tens-of-thousands of interactions.

Finally, we found some notable differences between the suggested research process as shown in the literature and the actual SCRUM implementation process. We will elaborate on those differences at the end of this paper and share a selection of methods that benefited our work.

2 RELATED WORK

Planning and scheduling are organizational methods that have been around for centuries, are of ever-growing importance since the industrial revolution, and will continue to be relevant into the foreseeable future [4, 17]. Computer-aided planning tools are a logical consequence of the increasing amounts of data used while scheduling. Tory et al., [33], for example, use their tool to compare construction schedules, while Kokkalis et al. [15] show an automated approach that combines crowd wisdom and natural language processing to suggest plans that help users tackle high-level tasks. By using visual analysis approaches, Schneider and Aigner [27] integrate collaborative automated planning into existing business scenarios.

Likewise, some tools exist that help students in their university, career, and course planning efforts. For example, Tomy and Pardede [32] show a tool that helps students to choose the right subject based on their skills, while Li et al. use visualization alongside gamification elements to diversify linear curricula and therefore motivate students to learn more different skills. With EventAction, Du et al. [6] present a tool that analyzes past events from each student individually and acts as a recommender system for educational planning. Another option would be to automate the planning process, but Pass et al. [24] found that students do not like automatic assignments. Instead of choosing solely optimal courses based on their curriculum, they favor courses by their personal preferences, teachers they already know, and classes that their friends attend. The tools above have in common that they focus on the Anglo-Saxon system. While similar in large parts, the lack of majors and minors in our university system that make them less usable or not effective at all. Also, they do not focus on planning a specific semester – which is the main goal for our tool.

While designing our tool, we used standard practices that we commonly use for research projects. We accounted for the pitfalls highlighted by Sedlmair et al. [30] to guide our design study. For engaging and active user participation, we borrowed some techniques that Kerzner et al. [13] and Knoll et al. [14] describe in their work about running creative visualization-opportunities workshops.

In the second half of our project, we collaborated with a team of programmers from an external software company. This part of the project was structured based on the SCRUM framework [28]. Human-Centered Design methods are an integral component of the SCRUM design process and have already been shown to enhance the outcomes [2]. In

addition, low- and high-fidelity prototyping and usability evaluation are highly regarded and often used in SCRUM teams [11]. On the other hand, there are approaches to adapt the SCRUM framework for research projects. For example, Ota [23] shows the general applicability for managing research projects with SCRUM. Furthermore, an adaptation of the SCRUM framework for agile project management in distributed research projects is shown by Hidalgo [8], and Lima et al. [25] report using SCRUM in at least seven different research projects at their lab. When comparing research projects and SCRUM projects with a focus on user interfaces, it is clear that both follow the user-centered design cycle [16] and the idea of creating mental models to capture the user’s perspective [20] as closely as possible. One slight difference between the two approaches is apparent in the way they tend to handle evaluation.

Research evaluations tend to be well thought out and structured. Extensive work on the topic exists characterizing domains and research methods [21], the types of evaluation and when to use them [1], and different scenarios considering evaluation goals [9, 18]. In their work on multi-dimensional in-depth long-term case studies, Shneiderman and Plaisant [31] describe methods to cope with long-running projects, and Ivory and Hearst [10] show methods for automating the usability evaluation of user interfaces.

While classical research projects need this rigor because they seek to report quantifiable data, SCRUM projects are often tied to a tighter budget and need to optimize their evaluation process. Larusdottir et al. [19] report that feedback from small groups in short cycles is often more effective than large-scale evaluations for most SCRUM projects.

Sedlmair et al. [29] encountered some resistance when trying to apply scientific evaluation methods in a company setting and conclude that some methods need to be adapted. While extensive evaluations might lead to the best insights, it could be helpful to combine both approaches to get the most out of an agile research project, especially if it is constraint by funding or seeks to be relevant in a real-world scenario.

3 METHODOLOGY

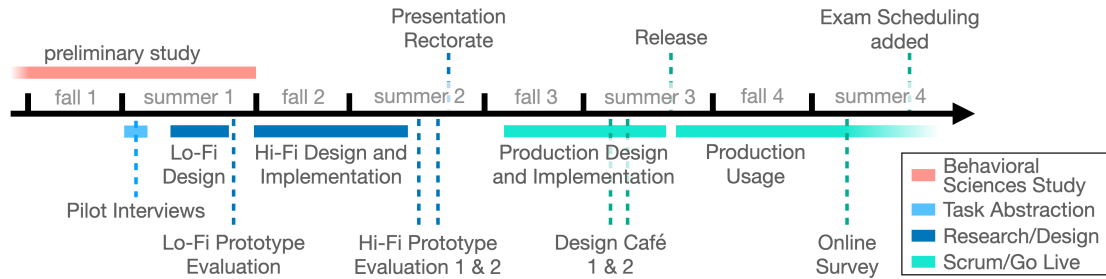


Fig. 1. A semester-based timeline of the whole four-year project with color-coding to differentiate the different parts of the project. It shows the duration of the different design, implementation and testing phases, and uses dashed lines to show the evaluations (bottom) and important events (top).

Fig. 1 shows a timeline of the full process, including all separate stages and milestones.

Our colleagues from the behavioral sciences ran a study that systematically tested the effect of various *soft measures* to promote study success. It included the measures of individual information, study groups with and without mentors, self-commitment, and a simple semester planning tool using LimeSurvey [7]. Eight faculties participated in that study, and even though the initial results showed no measurable effects with any treatment, a survey revealed that students who used the planning tool reported higher satisfaction with the university’s support than all others. The planning

tool let the students select possible courses and showed the number of credits already selected, contrasted with the recommended credits for one semester. Due to a lack of integration into university services, in order to use this tool, students needed to enter their status quo in terms of course credits manually. We, therefore, decided to build a much more user-friendly, integrated tool.

To maximize user acceptance and understand the planning process, we conducted a pilot study with three students from different faculties (three participants, one female, and two males; henceforth abbreviated with $f=1$, $m=2$). First, in a think-aloud session [5, 12], we observed how the students planned their semester. Then, combining the pilot's findings with our own planning experience, we abstracted this process' underlying tasks. Using these tasks, we designed low-fidelity prototypes of a tool that facilitated planning for the next semester. We evaluated the prototypes in guided interviews with seven students ($f=3$, $m=4$). The results helped us converge to one design, which we implemented as a high-fidelity prototype. Meanwhile, a workshop investigating use cases for student data in a separate project further solidified the need for a planning tool by independently coming up with an almost identical idea. In the first evaluation session of the high-fidelity prototype ($f=1$, $m=3$), which also used guided interviews, we found that our users could not identify with the mock student data provided in the prototype. Therefore, we conducted a second evaluation ($f=1$, $m=2$) for which we manually added accurate student data of the test users first.

After presenting our high-fidelity prototype to the president's office of our university, they decided to integrate the semester planning tool into the university website, initially as an extended test scenario, but ultimately for all students. Working with a SCRUM [28] team of three developers and one product owner, we designed and implemented our tool to be compatible with the university's homepage. While designing the production tool, we had to adapt the interface to fit the available data, adding a layer of structural dependencies that we were unaware of when designing our initial prototypes. Additionally, we had to work with a limited budget, such that some features did not make it into the final version of the tool. We gathered user feedback in two test cafés during the design and implementation process ($f=5$, $m=3$, and $f=4$, $m=2$). After releasing the planning tool on the university website, students from select faculties could use it for two semesters, and an online survey ($f=105$, $m=55$) concluded this test phase. The tool is still maintained, gets new features, and will successively become available for more students. At the time of writing the monitoring software counted 79,833 unique visits of the tool. We reflect on lessons learned from this process in sections 5 and 6.

All prototypes, interview guides, surveys, study designs, high-resolution images, and anonymized results are available in the supplemental materials.

4 PLANNING TOOL

The planning tool's general idea is to help students choose which courses to take in the following semester. The curriculum has a seemingly overwhelming amount of courses to offer, but not all are available for everyone. By reducing options to only feasible choices, the decision process is more straightforward and manageable. This section highlights the design process, elaborates on the planning process's abstract tasks, reflects on generating agency, and presents some crucial changes to the interface that happened during the project.

4.1 Task Abstraction

All collaborators on this work had their own experiences with how they planned their semesters during their studies, but we wanted to find out how current students — potential users of our tool — tackle this task. Thus we started the design process by inviting three students from different faculties ($m=2$, $f=1$) and asked them to plan their next semester while telling us about their thought process.

5 Designing a Semester Planner for Students (manuscript)

Designing a Semester Planner

CHI '22, Apr 30–May 6, 2022, New Orleans, LA

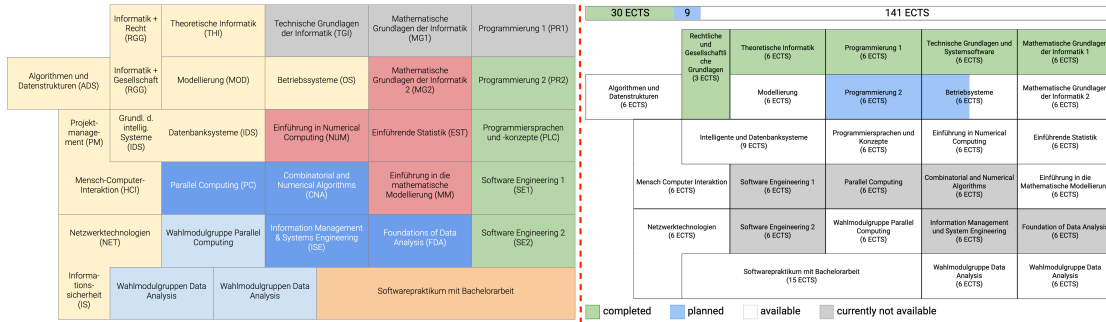


Fig. 2. The semester plan is a suggested course distribution, released by the faculty, where each row is one semester — in this case, for the bachelor's degree in data science. On the left: the semester plan as released by the faculty, on the right: our high-fidelity prototype.

One student's solution used a paper organizer, where a hand-drawn table of the faculty's semester plan (see fig. 2) was present. The student crossed out all courses that she already completed to evaluate which courses to take next. She then looked up those courses on the course directory website and copied all possible courses to a blank calendar page. Finally, she crossed out overlapping courses until the resulting schedule was overlap-free. Another student had a print-out of the semester plan and copied the courses he wanted to attend into his Google calendar, and the last participant used Microsoft Excel, where he also had a self-made version of the semester plan with different colored cells and used a separate sheet to find overlaps in courses he wanted to attend.

The semester plan shows courses structured into modules – groups of courses dealing with a similar topic. Therefore, students may need to complete multiple courses before finishing a module. Study participants used strategies like shading or glyphs to tackle partially finished modules.

Adding our own experience, we condensed the planning process into the following steps:

- (1) Get an overview of the current state by looking at available, as well as already completed courses
- (2) Choose courses that are interesting, useful, and possible to take in the following semester
- (3) Schedule the chosen courses as efficiently as possible, without overlaps
- (4) Create a timetable for the semester

By abstracting further, we can boil these tasks down to their core purposes: *Overview first, reflected choice, manage and adjust, summarize*. After some research, we found that many decision processes are structured that way [22, 33]. As an illustrative example, we apply it to travel planning:

- *Overview first*: Find applicable travel routes, e.g., bus, train, plane, and car
- *Reflected choice*: Comparing prices, and with the requirement to work on a laptop, choose traveling by train
- *Manage and adjust*: Select the optimal train by date, departure and arrival times, and intermediate stops
- *Summarize*: Print ticket and travel details

4.2 Design

We created the tool's interface in a multi-stage, user-centered design process (shown as Research/Design in fig. 1). With the abstract tasks as guidance, we structured the interface accordingly:

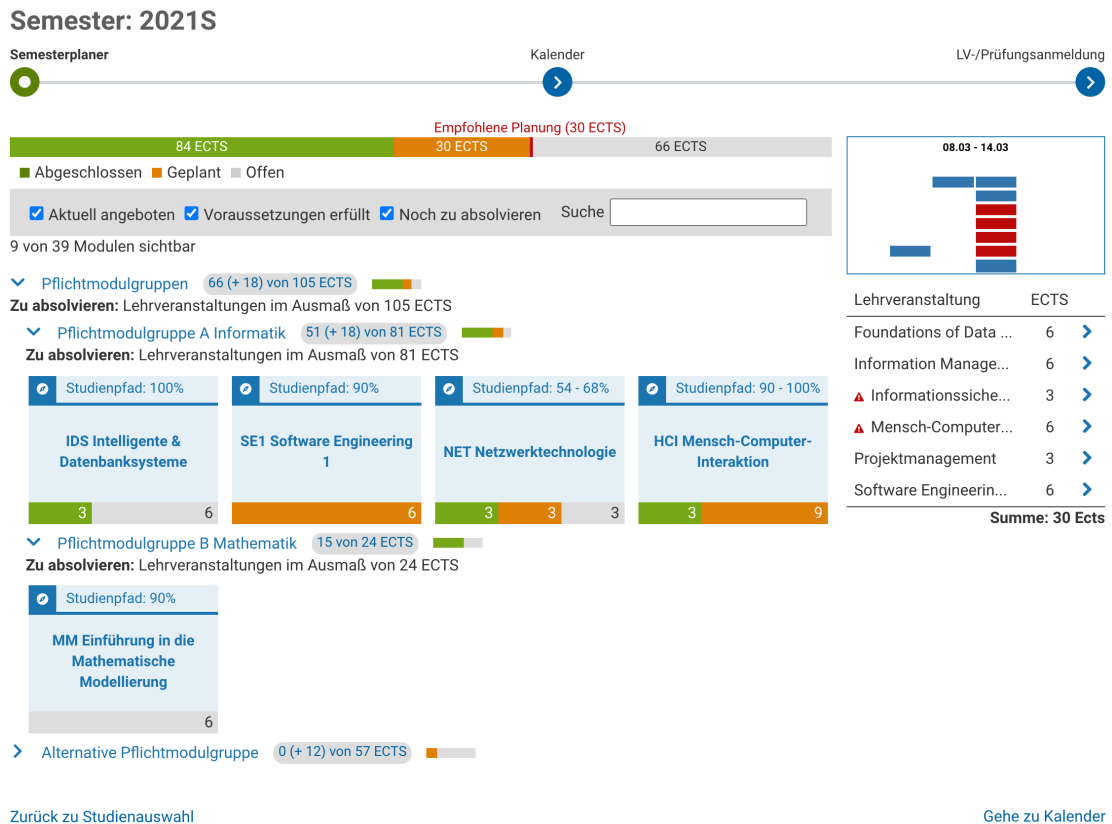


Fig. 3. The main view of our interface in its standard filtering shows all modules that are available to the current user. The three line-connected dots at the top show the current step of the planning process. A progress bar gives an overview of the already completed, currently planned, and open credits. Below is a filter and search bar to manually change the shown modules. Each blue rectangle represents one module, which can contain several courses. The status of these courses shown by individual progress bars in each module. Clicking on a module shows the contained courses. By selecting a course it gets added to the planning selection. The sidebar on the right shows a preview of the calendar view and a list of planned courses.

Overview first. The first view in our interfaces is an overview of the current possibilities the user has. Fig. 3 shows the overview page of our tool. In its initial state, it shows all available courses for the next semester. By modifying the filters, the user can also display unavailable and already completed courses.

Reflected choice. Clicking on a course shows its details, such as instructors, groups, times, and dates. From there, the user can add any number of courses to their planned courses on the right side of the interface and the calendar.

Manage and adjust. The calendar view (see fig. 4) highlights time overlaps of courses, making it easy to schedule them and detect time conflicts. We further enhanced this by including a semester overview at the bottom of this view that shows if an overlap exists for the whole semester.

Summarize. The last step in the interface summarizes all chosen courses, displays them with their registration links, and provides a download option for the resulting calendar file.

5 Designing a Semester Planner for Students (manuscript)

Designing a Semester Planner

CHI '22, Apr 30–May 6, 2022, New Orleans, LA

Semester: 2021S

Semesterplaner

Kalender

LV-/Prüfungsanmeldung

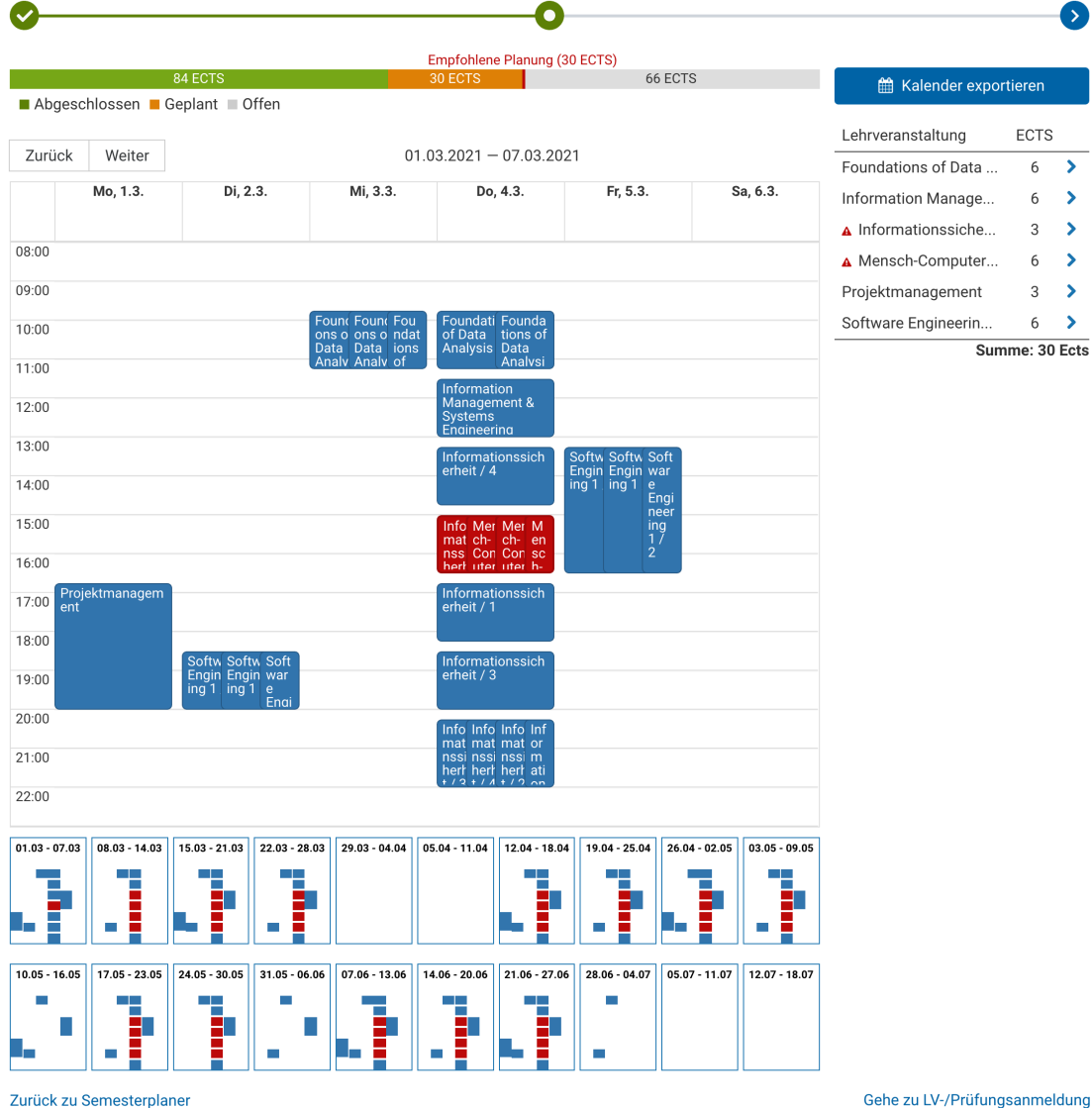


Fig. 4. We use a calendar to show occupied timeslots from planned courses. A red highlighting shows overlaps of different course times. Below the primary calendar are previews of all weeks in the semester, making it easy to distinguish between recurring and single overlaps. A calendar export button and the list of planned courses is located on the right side.

A small preview of the calendar is also visible in the top right corner of the overview (see fig. 3). It is set to show the current week during the semester and the next whole week of the semester during the holidays. While it does not show detailed information, it gives a good overview of the already planned courses, and the red highlight allows users

to detect overlaps without switching to the calendar view. Our low- and high-fidelity prototypes featured a similar preview of selected courses for the calendar view based on the semester plan, but it was replaced with a simple list of selected courses in the final version. This list looks similar to a shopping cart in an online shop, which is more familiar to users, and was more cost-efficient to implement.

4.3 Trust and Agency

We aimed at creating an approachable and enjoyable interface for students, which has more benefits for the users than a simple scheduling tool. A crucial part of creating agency in our interface — while focusing on the main task of planning — was to provide the option for our users to look back at what they had already done. By seeing the bigger picture and context, they could identify and confirm the integrity of the data they saw, thus making an informed choice. The semester plan (see fig. 2) provided an ideal way of showing the current status in our prototypes. We got very positive feedback about it in the low-, and high-fidelity prototype evaluations. The final version of the tool eventually did not include a semester plan because it turned out that only about half of all faculties in our university provide one. We just happened to test our prototypes with students of those faculties. We still included the overview of seeing the detailed study progress by adding an option to the main view to show already completed courses.

Another highly appreciated feature in decision-making was the inclusion of a semester overview at the bottom of the calendar view (see fig. 4). This feature allows our users to see if an overlap happens only once or repeatedly during the semester. If a single appointment is conflicting, a student might choose to ignore the collision and still schedule both courses, allowing decision-making based on the complete available information. Students seemed positively surprised that we included such a feature in the planning tool because they expected that planning to skip single courses would not be appreciated by the university.

5 EVALUATION RESULTS

This section discusses the results for each of the evaluations shown at the bottom of fig. 1. For our first evaluation, we created two different interfaces as clickable low-fidelity prototypes using Balsamiq [3]. Then, we invited students, including some of our pilot study participants, to individual guided interviews ($f=3$, $m=4$). First, participants had to complete a few simple tasks, after which we questioned them about their preferences, the ease of use, and their recommendations on how to improve. One of our designs used the layout of the semester plan (see fig. 2) to show all available data, but the users noted that most of the shown modules were irrelevant for planning. In the second design, we categorized the modules and only showed the selection relevant for planning, which allowed users to focus on planning. The progress bar from the first design occupies a central spot in the final design because it was the most requested feature from the first evaluation. Unfortunately, some of our testers got confused with the low-fidelity prototypes because they could only click on specific elements and the rest was not interactive.

Using this feedback, we implemented a high-fidelity prototype that combined the most-liked features and recommendations. The high-fidelity prototype also mitigated the previous confusion since all elements were clickable. To evaluate the resulting prototype, we invited more students, alongside some initial testers, to our lab again, where we conducted a similarly structured guided interview as in the previous evaluation. We noticed that the users ($f=1$, $m=3$) could not identify with the interface's data during the first day of interviews. They stated that they were unfamiliar with the shown student data from a random computer science student. Therefore, we postponed the second day of interviews and manually integrated our participants' actual study data into the interface. On the second day of interviews, the students ($f=1$, $m=2$) used the interface more intuitively since they recognized their curricular choices. Both days resulted

in positive feedback by our testers, and the usability rating from both days was similarly good. Notably, the answers to what the users would change about the tool varied between the days. Users from the first session mainly talked about a few general changes to the interface, while the users from the second session had a lot more ideas for features, which were primarily additions to the tool. We find this distinct difference in answers interesting since it could intentionally steer an evaluation in a general direction. The users pay closer attention to aesthetics when looking at mock data while they focus on the tool’s functionality when working with their actual data.

When designing the tool’s release version, we took advantage of the monthly scheduled test cafés hosted by our collaborators. These took place in a coffee shop on campus where students could walk in and participate without invitation. Usability experts guide them to perform usability tasks on prototypes and document their feedback. Participating students received a ten euro voucher of their choice as compensation. Due to the nature of the incremental implementation process and quick feedback cycles, we did not evaluate the whole tool but instead evaluated different parts of our tool individually. Also, the two test cafés mainly focused on usability. As one of the results, we added a progress indicator that indicates the user’s current step of the planning phase to the tool.

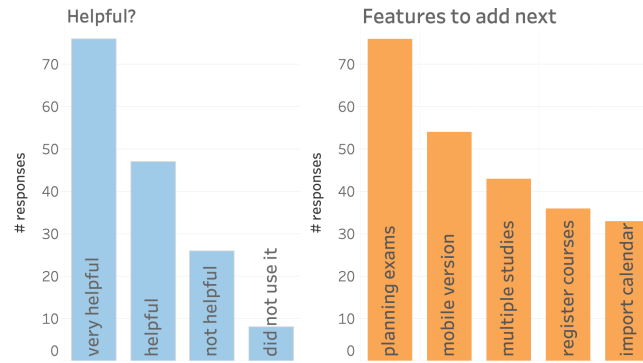


Fig. 5. These charts show the results from the online survey ($n=160$) we ran after our tool was released to students of five faculties. The left chart shows the rating how helpful our users found the tool, the right chart shows the features our users requested the most. Multiple answers were allowed for the second question.

We counted more than 11,000 unique users in around 35,000 unique sessions in the first two semesters after releasing the tool for our students. At this point, we conducted an online survey about the user’s experience. This survey ($f=105$, $m=55$) showed that 123 participants found the tool helpful or very helpful and the most valued features were the calendar view and the overview of available modules. However, analyzing the comment field of that survey, we found that the tool had a database issue for law students that resulted in incorrect courses being displayed, which we promptly fixed. Finally, we asked what feature the students would value most as the next addition to the tool. The most requested feature was to use the planning tool for scheduling exams as well, which we already added at the time of writing. Making the interface mobile-friendly ranked second, followed by planning multiple studies side-by-side, registering for courses directly from the planning interface, and lastly importing private calendars to see overlaps with personal events (see fig. 5).

Besides these direct user evaluations, we also evaluated the semester planner quantitatively using an experimental between-subjects design. The Faculty of History agreed to an A/B test scheme, where we randomly selected 50% of their 561 Bachelor’s students to gain access to the tool during the first semester after its release, while the other 50% needed to

plan without our tool. Data shows that 400 of the 561 students were active during the semester. We investigated whether students completed more credits than without our semester planner and found a statistically significant improvement. In particular, we found a causal treatment effect of the semester planner for history students of 1.45 credits, i.e., History students who had access to the semester planner completed 13% more credits on average than those who had not. This effect is particularly strong when contrasted with the results obtained from our colleagues in the behavioral sciences. When they evaluated the simple semester planner implemented in LimeSurvey, they found a 4% increase for tool-using students of the Faculty of History. The fact that the performance increase caused by the more sophisticated semester planner is more than three times as high as compared to the simple planner prominently demonstrates how vital a proper user-centered design, evaluation, and development process can be.

6 DISCUSSION

A benefit of working on a research project is near limitless room for exploration. At the start of the project, after narrowing down the task, we applied the *Five Design-Sheet Methodology* [26] that created three very different designs for the planning interface. Combining two options, we then designed two different versions of the interface for our first evaluation. We did not start to realize a high-fidelity prototype until this evaluation was concluded, and we knew on which parts to focus. Later, when we concluded the first day of our high-fidelity prototype evaluation, we had the flexibility to postpone the second day to include actual student data. These steps caused a delay in the process, which might be hard to justify in a business context and translate to additional costs.

A significant benefit of working in a budget-tied environment was prioritizing each part of our design. It challenged us to question every little aspect and focus on the primary task our tool wants to achieve. We gathered many ideas from the previous prototype evaluations, which were welcome additions to the functionality of our tool. When stripping down every aspect to focus on the planning process, we labeled a lot of added features as *nice-to-have*. That label implies that a feature can be added later but is unnecessary for the tool's initial version. This concept of a *Minimal Viable Product (MVP)* is crucial in the SCRUM framework. It ensures that the resulting product from the iterative implementation process can complete the required task and stays within the given budget. A practical but unexpected consequence for the user evaluation stems from the iterative implementation process. The SCRUM process creates *increments* in short cycles. An increment is a concisely defined part of the whole interface, which provides a concrete function and is already usable on its own. The nature of these increments makes it possible to evaluate individual pieces of the interface out of context to see if they also function independently. Evaluating single pieces allows for detailed feedback on their usability and whether they are intuitively usable. This evaluation is easily set up, provides quick results and short feedback cycles.

Both approaches can enhance each other since they generally aim at a similar target – creating user-friendly interfaces. The abstract task of *Overview first, reflected choice, manage and adjust, summarize* was extensively used when we decided which parts of the prototypes need to be prioritized for the MVP. These four steps could be achieved by displaying everything that has already been completed, aggregating all relevant information on available courses, a calendar that shows overlaps, and an export option. In the end, even the semester plan, which we deemed crucial for this tool up to this point, could be left out by simply adjusting the filters to show completed courses as well. On the other hand, the SCRUM team members stated that they wished MVP conceptions would always be this straightforward because finding and defining the core task is often an essential part of the MVP definition process. Therefore, both processes have their advantages and disadvantages but can be combined for even better results in either case.

We can, therefore, answer our research questions as follows:

- (1) Can an online tool enhance the semester planning process?

The evaluation of our prototypes and the online survey clearly show that our online tool simplifies and enhances the planning process.

- (2) Does such a planning tool affect student registration numbers?

A/B testing reveals that it affects registration numbers and even course completion numbers by up to 13%. However, we consciously omit to show the registration numbers because they can be erroneous due to early course drop-outs or students who discontinued their studies.

- (3) Do the recommendations of the *Design Study Methodology* [30] and the *Multi-dimensional in-depth long-term case studies* [31] hold up to implementing production-ready tools using the *SCRUM Framework* [28]?

Some parts of the real-world implementation process are well supported and documented in the literature. Additionally, both works cited above help maintain scientific rigor and prioritize proper evaluation, which is not the main focus from the business perspective. They are a welcome addition to the SCRUM framework in that regard. One thing they both lack is the focus on the core task. Both methods give valuable input on properly evaluating any task that could come up in different situations but do not guide the reader towards creating an MVP.

Breaking down our design into all its components showed us how few of them are necessary for the tool to be functional. Consequently, we encourage prioritizing and separating features between *MVP relevant* and *nice-to-have* for scientific projects as well. It challenges researchers to question the design, facilitates task abstraction, and considerably improves production feasibility. Even though we think that including the semester plan would, nevertheless, be a worthwhile addition, the simplified interface we implemented brings us closer to a joking remark from one participant of our pilot study. When asked how he imagined a planning tool, he initially said “*Just one button: plan now. I press it, everything gets sorted out automatically, and I only have to show up on time.*”

6.1 Future Work

The most requested feature of planning exams with our tool has already been added to the tool, and we are currently discussing an additional budget for a mobile-first version with the president’s office of our university. Meanwhile, the current version of the tool is continuously rolled out to more faculties, which we will continue to evaluate and use to compare to an eventually available mobile-first version of the tool.

In future projects, we want to test if our assumption that using actual user data vs. generic data in evaluations continues to produce differences in focus between aesthetics and functionality. A review of other planning tools would reveal insight into the general applicability of our new mantra of *Overview first, reflected choice, manage and adjust, summarize..* Finally, we want to explicitly compare and contrast research and SCRUM methods to understand better how they complement each other.

7 CONCLUSION

In this work, we summarized a four-year-long research project that resulted in a tool that helps students plan their semesters. We analyzed the semester planning process and abstracted it into a broadened version that can be applied to a plethora of general planning tasks, from construction work to travel planning – *Overview first, reflected choice, manage and adjust, summarize..* This abstraction can help future designers of planning support tools to identify crucial tasks and focus on the core purpose of planning.

Using this abstraction, we created a semester planning tool using an iterative, user-centered design approach. We evaluated the tool during multiple stages of development and reported the findings of each evaluation. In the second stage of our development, we collaborated with an external SCRUM team and implemented the final version for real-world use. In order to stay within our budget, we had to focus on the essential parts of our tool to create a minimal viable product that was still capable of achieving all four main tasks. This prioritization process presented a different view on the research project, which finally resulted in a released tool with 79,833 unique visits thus far. It also enabled a streamlined and focused user evaluation, and a measurable increase of 13% completed credits from students using our tool compared to a control group.

Another remarkable result from our evaluations was the difference in the feedback we got when evaluating the high-fidelity prototype with and without actual user data. Our users commented on the usability in both cases, but the feedback from showing mock data revolved around the tool's aesthetics, while the focus was mainly on functionality when showing actual user data. Studying this effect in more detail could show whether this is a generalizable aspect for evaluation designs or just a coincidence in our evaluation.

Finally, we want to encourage other research projects to prioritize all aspects of their design, already starting at an early stage. Focusing on the main tasks and examining each part whether it is necessary to achieve that task can later help to evaluate a project and increase its chances of adoption in a real-world scenario. In addition, the SCRUM practice of evaluating parts of the design out of context, solely based on their usability, helps generate additional insights in the process.

REFERENCES

- [1] Keith Andrews. 2008. Evaluation Comes in Many Guises. *In AVI Workshop on BEyond time and errors (BELIV)* (2008), 8–10.
- [2] Carmelo Ardito, Maria Teresa Baldassarre, Danilo Caivano, and Rosa Lanzilotti. 2017. Integrating a SCRUM-Based Process with Human Centred Design: An Experience from an Action Research Study. *Proceedings - 2017 IEEE/ACM 5th International Workshop on Conducting Empirical Studies in Industry, CESI 2017* (2017), 2–8. <https://doi.org/10.1109/CESI.2017.7>
- [3] LLC Balsamiq Studios. 2021. *Balsamiq Wireframes - Industry Standard Low Fidelity Wireframing Software*. Balsamiq Studios, LLC. <https://balsamiq.com/wireframes/>
- [4] B.J. Cox. 1990. Planning the software industrial revolution. *IEEE Software* 7, 6 (1990), 25–33. <https://doi.org/10.1109/52.60587>
- [5] B Davey. 1983. Think Aloud: Modeling the Cognitive Processes of Reading Comprehension. *Journal of Reading* 27, 1 (1983), 44–47. <http://www.jstor.org/stable/10.2307/40029295>
- [6] Fan Du, Catherine Plaisant, Neil Spring, Kenyon Crowley, and Ben Shneiderman. 2019. EventAction: A Visual Analytics Approach to Explainable Recommendation for Event Sequences. *ACM Transactions on Interactive Intelligent Systems* 9, 4 (dec 2019), 1–31. <https://doi.org/10.1145/3301402>
- [7] LimeSurvey GmbH. 2021. *Universities - LimeSurvey - Easy online survey tool*. LimeSurvey GmbH. <https://www.limesurvey.org/en/solutions/universities>
- [8] Enric Senabre Hidalgo. 2019. Adapting the SCRUM framework for agile project management in science: case study of a distributed research initiative. *Heliyon* 5, 3 (2019), e01447. <https://doi.org/10.1016/j.heliyon.2019.e01447>
- [9] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. 2013. A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2818–2827. <https://doi.org/10.1109/TVCG.2013.126>
- [10] Melody Y. Ivory and Marti A. Hearst. 2001. The state of the art in automating usability evaluation of user interfaces. *Comput. Surveys* 33, 4 (2001), 470–516. <https://doi.org/10.1145/503112.503114>
- [11] Yuan Jia, Marta Kristin Larusdottir, and Åsa Cajander. 2012. The usage of usability techniques in scrum projects. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7623 LNCS (2012), 331–341. https://doi.org/10.1007/978-3-642-34347-6_25
- [12] Christopher J Johnstone, Nicole A Bottsford-Miller, and Sandra J Thompson. 2006. Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners (NCEO Technical Report). *National Center on Educational Outcomes* (2006), 1–25. <http://hdl.handle.net/11299/174019%0Ahttp://www.nceo.info>
- [13] Ethan Kerzner, Sarah Goodwin, Jason Dykes, Sara Jones, and Miriah Meyer. 2018. A Framework for Creative Visualization-Opportunities Workshops. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 748–758. <https://doi.org/10.1109/TVCG.2018.2865241> arXiv:1808.02502

5 Designing a Semester Planner for Students (manuscript)

- [14] Christian Knoll, Asil Çetin, Torsten Möller, and Miriah Meyer. 2020. Extending Recommendations for Creative Visualization-Opportunities Workshops. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*. 81–88. <https://doi.org/10.1109/BELIV51497.2020.00017>
- [15] Nicolas Kokkalis, Johannes Huebner, Steven Diamond, Dominic Becker, Michael Chang, Moontae Lee, Florian Schulze, Thomas Koehn, and Scott R. Klemmer. 2012. Automatically providing action plans helps people complete tasks. *AAAI Workshop - Technical Report WS-12-08*, 5 (2012), 116–117.
- [16] Olga Kulyk, Robert Kosara, Jaime Urquiza, and Ingo Wassink. 2007. Human-centered aspects. In *Human-centered visualization environments*. Springer, 13–75.
- [17] Kaushik Kumar, Divya Zindani, and J Paulo Davim. 2019. Process Planning in Era 4.0. In *Industry 4.0*. Springer, 19–26.
- [18] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. <https://doi.org/10.1109/TVCG.2011.279>
- [19] Marta Lárusdóttir, Ása Cajander, and Jan Gulliksen. 2014. Informal feedback rather than performance measurements - User-centred evaluation in SCRUM projects. *Behaviour and Information Technology* 33, 11 (2014), 1118–1135. <https://doi.org/10.1080/0144929X.2013.857430>
- [20] Zhicheng Liu and John Stasko. 2010. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 999–1008. <https://doi.org/10.1109/TVCG.2010.177>
- [21] JOSEPH E. MCGRATH. 1995. METHODOLOGY MATTERS: DOING RESEARCH IN THE BEHAVIORAL and SOCIAL SCIENCES. In *Readings in Human-Computer Interaction*, RONALD M. BAECKER, JONATHAN GRUDIN, WILLIAM A.S. BUXTON, and SAUL GREENBERG (Eds.). Morgan Kaufmann, 152–169. <https://doi.org/10.1016/B978-0-08-051574-8.50019-4>
- [22] Michael J. Mortenson, Neil F. Doherty, and Stewart Robinson. 2015. Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research* 241, 3 (2015), 583–595. <https://doi.org/10.1016/j.ejor.2014.08.029>
- [23] Martin Ota. 2010. SCRUM in research. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6240 LNCS (2010), 109–116. https://doi.org/10.1007/978-3-642-16066-0_18
- [24] Michael W Pass, Sanjay S Mehta, and Gurinderjit B Mehta. 2012. Course selection: Student preferences for instructor practices. *Academy of Educational Leadership Journal* 16, 1 (2012), 31.
- [25] Igor Ribeiro Lima, T de Castro Freire, and Heitor Augustus Xavier Costa. 2012. Adapting and Using SCRUM in a Software Research and Development Laboratory. *Revista de Sistemas de Informação da FSMA* 9, - (2012), 16–23. http://www.fsma.edu.br/si/edicao9/FSMA_SI_2012_1_Principal_2_en.pdf
- [26] Jonathan C. Roberts, Chris Headleand, and Panagiotis D. Ritsos. 2016. Sketching Designs Using the Five Design-Sheet Methodology. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 419–428. <https://doi.org/10.1109/TVCG.2015.2467271>
- [27] Thomas Schneider and Wolfgang Aigner. 2011. A-Plan: Integrating Interactive Visualization with Automated Planning for Cooperative Resource Scheduling. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (Graz, Austria) (i-KNOW '11)*. Association for Computing Machinery, New York, NY, USA, Article 44, 8 pages. <https://doi.org/10.1145/2024288.2024341>
- [28] Ken Schwaber and Jeff Sutherland. 2017. The SCRUM Guide: The Definitive The Rules of the Game. *Scrum.Org and ScrumInc* November (2017), 19. <https://doi.org/10.1053/j.jrn.2009.08.012> arXiv:arXiv:1011.1669v3
- [29] Michael Sedlmair, Petra Isenberg, Dominikus Baur, and Andreas Butz. 2011. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization* 10, 3 (2011), 248–266. <https://doi.org/10.1177/1473871611413099>
- [30] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (dec 2012), 2431–2440. <https://doi.org/10.1109/TVCG.2012.213>
- [31] Ben Shneiderman and Catherine Plaisant. 2006. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. *Proceedings of BELIV'06: BEyond time and errors - novel EvalUation methods for Information Visualization. A workshop of the AVI 2006 International Working Conference* (2006). <https://doi.org/10.1145/1168149.1168158>
- [32] Sarath Tomy and Eric Pardede. 2018. *Course map: A career-driven course planning tool*. Springer International Publishing. 185–198 pages. https://doi.org/10.1007/978-3-319-95165-2_13
- [33] Melanie Tory, Sheryl Staub-French, Dandan Huang, Yu Ling Chang, Colin Swindells, and Rachel Pottinger. 2013. Comparative visualization of construction schedules. *Automation in Construction* 29 (2013), 68–82. <https://doi.org/10.1016/j.autcon.2012.08.004>

6 Discussion

This chapter discusses how each of the four presented publication manuscripts contributes to the three main research questions. Further, it summarizes the methodology used to come to these results and closes with open questions for future research.

RQ1: How do data representations have to differ in order to efficiently accomplish different tasks from different stakeholders?

Both works on the *distance metric* and the *semester planner* deal with the identical problem that the data on course completions in its full extent is too much to grasp in its entirety. While they address perspectives and tasks from different stakeholders, they both present ways to augment the data that make it easier to understand and work with it.

The representation of a *study path* in our *distance metric* is a linear representation of courses along a *timeline*. This visual abstraction into concretely stepped timelines that allows concurrencies of events is a close representation of the mental model the interviewed stakeholders use in their daily work. To take advantage of this structure, we stripped all unnecessary details from the data, and our metric allows our users to sort, filter, cluster, and browse all study paths.

For our *semester planner*, instead of reducing the data to the minimal amount of needed features, we enrich it with meta-information that is relevant for planning. Therefore, we gather additional data about courses from the curricula texts, the available course dates and times, information about prerequisites, and registration deadlines. This data aggregation for each course combines all the information that the interviewed students reported using for their planning. Proper structure in the user interface and pre-filtering to only show planning-relevant objects help not overwhelm our tool's users.

The main component of both representations is that it fits with the mental model of the stakeholders. In the case of the *distance metric*, we tried to represent the data as timelines first. We encountered the problem that timelines do not have an easy way to contain concurrent events, and the length of semesters is not equal. In Austria, the summer break is longer than the winter break, which results in differences when calculating distances between timelines. Therefore, we created the representation we call *sets of events* that deals with both of these problems. To summarize, a suitable data representation needs to capture the mental model of the stakeholders entirely to help them understand their data. It is, therefore, crucial to fully understand the task, requirements, and environment of the stakeholder to create an adequate representation of their mental model.

RQ2: How can perception and interaction be used to make data visualizations more comprehensible and easier to use?

Shape recognition and shape abstraction are well suited for human perception since human vision can quickly recognize and differentiate a variety of shapes [MKO⁺02]. Interestingly, the commonly used methods for estimating the number of bins in a histogram did not consider human perception. They solely looked at the characteristics of the underlying data, resulting in a massive number of bins for large data sets. Therefore we reverse the approach and primarily focus on perception in our work on *histogram binning*. Comparing the results from our study to the recommended amount of bins, we conclude that shape recognition does not benefit from showing more bins.

A similar observation can be reported with our work on *brushing of parallel coordinate plots*. Parallel coordinate plots are a powerful multi-dimensional data visualization method, but they are inherently difficult to understand even for expert users. Interaction, in the realm of this work the highlighting of lines, makes understanding the depicted results easier. To further facilitate the process, we reduce the number of mouse interactions needed to a single click-and-drag gesture that is intuitive to use. While this does not reduce the overall difficulty of these plots, it reduces mental load and motivates users to explore the depicted data.

Data visualizations are a generally powerful tool for making data accessible to human viewers. Not considering the users' perception limits the value of visualizations, making users reluctant to deal with them. Therefore, human perception should be considered when designing visualizations, especially for non-expert users. Throughout many interviews conducted in this thesis, the users remarked that they felt less overwhelmed by a visualization if less information was visible at once. For example, showing fewer, already filtered courses in the planning interface or showing fewer bars in a histogram made the users more likely to continue using a visualization. This tradeoff of losing expressiveness through simplification to not overwhelm the user can make the visualization more useful. This confirms the conventional wisdom that the hard part of any user interface design is to figure out what to leave out.

RQ3: How does the process of creating a research prototype differ from the process of creating a ready-to-use-tool?

Dealing with a large data set in a user interface often translates to reducing the amount of data shown. In our work on *designing a semester planner*, we explored different strategies that both reduce the amount of data shown not to overwhelm the user while providing additional information whenever necessary. We found three principles that helped provide trust and agency for our users when using the planning tool.

1. Do not artificially limit options for interaction.
A well-received user interface contains the option to interact with the available data freely. Even if they help achieve the underlying task, obvious restrictions overrule the user's choice, thus making the user experience feel forced and unnatural.
2. Make deductions and calculations transparent.
If the user experience is guided or even limited by algorithmic decisions, e.g., calculating a threshold, try to show how this calculation was done. Also, show ways the user can influence the algorithm, if possible.

3. Show all related data and its context.

Large amounts of data can quickly become overwhelming when no context is present. Showing the data's origin, historical data, or other defining factors make it easier for the user to grasp its meaning. While it is generally considered good practice to hide such data to reduce possible information overflow, it is equally well-received if the data can become available through an additional interaction when needed.

The overall methodical approach of this thesis to make large data sets accessible for multiple stakeholders is entirely focused on being user-centered. All presented works use this paradigm as a foundation for significant decisions in their processes. Evaluation plays an almost equally significant role in the used methodology. User-centered design and evaluation are intertwined to ensure good results. Initial analysis of the current situation allows us to understand the situation and the state-of-the-art of how tasks or problems are currently solved. Workshops, pilot studies, and, in the case of the highlighting of lines in parallel coordinate plots, our personal experience led us to real-world problems when tackling the work with large data sets. Building on these problems, we abstracted the core task behind them and formulated hypotheses and research questions on its basis. This abstraction process ensures that a problem is not solely domain-specific such that our findings that base on it can also be applied to other domains.

After the abstract task is defined, an appropriate methodology gets chosen. In both *Chapter 4: Selective Angular Brushing of Parallel Coordinate Plots* and *Chapter 5: Designing a Semester Planner for Students*, we applied an iterative design process which consisted of low-fidelity prototyping, evaluating the low-fidelity prototypes, high-fidelity prototyping, and finally evaluating the high-fidelity prototypes. We went one step further for the semester planner by fully implementing and releasing it for students of the University of Vienna. In *Chapter 2: A Distance Metric for Sets of Events* we take the abstracted concept of the study path, define it mathematically, and create a distance metric based on this mathematical representation. We evaluate the metric by mathematically proving the metric properties and additionally test if the results from our metric match students' expectations when applied to study paths. This added evaluation step ensures that our stakeholders understand and confirm our results. Finally, in *Chapter 3: Histogram binning revisited with a focus on human perception* we formulate hypotheses based on the abstract task that can be answered in an interactive online survey. We use these hypotheses to tailor data, charts, and questions accordingly.

The differences above show that no one-fits-all solution is possible when designing systems that allow multiple stakeholders to access a large data set. Even when looking at only data and stakeholders from the University of Vienna, the tasks are too varied, and the expertise of stakeholders and their problems differ vastly.

Therefore, the findings of this thesis can be summarized to a general workflow to ensure accessibility for relevant parties when working on large data sets:

1. **Investigate actual situation**

Do not treat problems as *given* just because they were presented as such. Take

some time to investigate and personally understand the current situation, the roles different stakeholders play, and record tasks as they happen.

2. Find the abstract task

After defining the concrete problem or task, try to abstract it, reduce it to its core parts. Using this abstraction, make sure it still fits previous observations and then apply it to other domains. Some tasks might already be solved in domains that do not seem obvious at first glance.

3. Choose the appropriate method(s)

Can the problem be solved by data abstraction and data structure, perception and interaction, a user interface, or a combination of the three concepts?

4. User-centered design

The two key components while designing/implementing are to keep the users in the loop and always keep the abstract task in mind. When adding functionality, make sure it is necessary to achieve the core task and that the users can use it.

5. Constant feedback

Designing prototypes and quick feedback cycles ensure that eventual errors get spotted immediately and have no great impact on the project. Each cycle should take between two and four weeks and produce a result that users can test. Sometimes it is better to focus on a single element that can be tested instead of creating a full view that does not sufficiently work.

6. Evaluate the abstract task

Evaluate if the final product fulfills all requirements of the abstract task. Also, make sure that it fits into the user workflow, can be used intuitively, does not obstruct but rather enhances the workflow, and does not increase the overall amount of work. These conditions increase the chance of adoption.

This workflow was independently derived by summarizing findings from the presented work. However, it fully concurs with the *Nested Model* presented by Tamara Munzner [Mun09]. Therefore, this thesis gives concrete recommendations on implementing user-centered data accessibility and validates the applicability of the *Nested Model*.

Limitations

All research conducted for this thesis took place in Austria, most of it at the University of Vienna. This locality introduces a noteworthy bias. While the author assumes general applicability of the presented results, they might differ for different university settings or cultures.

For example, the concept of a *study path* might not be transferrable to other universities or other university systems. The distance metric for sets of events, on the other hand, is generally applicable.

A locality bias also affects our work on the brushing of lines in parallel coordinate plots. Almost all participants in our study were from Austria, and a substantial amount were

either students or from academia. A broader study setting could eliminate that bias and also focus on the measurable benefits of that method.

Our work on the perception of data distributions in histograms focuses only on the perceptual consequences for detecting general shapes in distributions. Other aspects like outlier detection or the judgment of minima, maxima, or other concrete values were not tested.

The interface for the semester planning process possibly contains a bias towards students using a desktop computer for planning since no mobile interface for the tool exists. Users may have an increased focus on planning and longer attention when using a desktop machine compared to a mobile device. Also, the reported evaluations from the SCRUM context in this work were all done with the same team. Even though this team follows the SCRUM guidelines closely, the individual interpretation of such guidelines always plays a role in their implementation. Additionally, the SCRUM framework is constantly evolving, which might influence the evaluations as well.

6.1 Future Work

Using the presented framework in more projects, on other data sets with diverse stakeholders, preferably in different locations or countries, will help verify and extend its recommendations. In addition, a literature study that tries to code findings from related projects could further extend this framework's scope and yield additional suggestions.

Currently, the field of highest interest to the author regards the perception of visualizations. While working on the perception of histograms, additional questions that need more research emerged. For instance, if other tasks that judge distributions in histograms, e.g., outlier detection, finding minima or maxima, work with the same amount of bins we recommend for shape detection or need different bin counts. The position and amount of labels in histograms can also play a role in their perception, especially for determining bin ranges. Similar perceptual studies could also enhance other visualizations apart from histograms.

6.2 Conclusion

This thesis summarizes several concrete concepts for dealing with the accessibility of large data sets for multiple stakeholders. It presented four different publications and manuscripts that contribute guidelines for

1. data representation and structure,
2. perception and interaction, and
3. user interface and user experience.

Many different evaluation methods were used throughout all presented works, and their results further contribute to solidifying the proposed guidelines. This thesis aims to help future projects with a similar scope by combining all concepts and recommendations into

6 *Discussion*

a simple six-step checklist. Following this workflow should ensure a solid foundation for multi-user accessibility of large data sets.

Bibliography

- [MKO⁺02] Scott O. Murray, Daniel Kersten, Bruno A. Olshausen, Paul Schrater, and David L. Woods. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23):15164–15169, 2002.
- [Mun09] Tamara Munzner. A Nested Process Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.
- [oV19] University of Vienna. Verzeichnis der Wissensbilanz-Kennzahlen 2019. https://www.univie.ac.at/fileadmin/user_upload/startseite/Fotos/Publikationen/LB_2019_Kennzahlen.pdf, 2019. [Online; accessed 06-June-2021].
- [SGMS21] Raphael Sahann, Ivana Gajic, Torsten Möller, and Johanna Schmidt. Selective Angular Brushing of Parallel Coordinate Plots. In Marco Agus, Christoph Garth, and Andreas Kerren, editors, *EuroVis 2021 - Short Papers*. The Eurographics Association, 2021.
- [SM18] Raphael Sahann and Torsten Möller. OCP - Operational Curricular Planning: A Visual Decision Support System for Planning Teaching Resources at Universities. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pages 1–9, 2018.
- [SPM20] Raphael Sahann, Claudia Plant, and Torsten Möller. A Distance Metric for Sets of Events. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 506–515. IEEE, 2020.

