Bachelorarbeit am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC)

Result-driven Interactive Visual Support of Parameter Selection for Dimensionality Reduction

Lilli Joppien

Betreuer: Dr.-Ing. C. Kinkeldey

Erstgutachterin: Prof. Dr. C. Müller-Birn

Zweitgutachter: Prof. Dr. L. Prechelt

Berlin, den 2. September 2020

Abstract

People without a technical background or knowledge of machine learning (ML) technology (non-technical experts) have become a major target group of ML-applications. Nonetheless, ML-systems still rarely support them in their informed use. This thesis investigates a new visual interface to explore new ways towards an informed use. The prototypical interface was built upon an existing ML-application in the research project IKON, in which dimensionality reduction (DR) is applied. DR is a widely used tool for the interpretation of high-dimensional ML-results. Due to the effects of information loss caused by reduced dimensionality, it can produce artifacts and transform the same high dimensional result into 2D representations that may vary a lot. Different criteria may be important depending on the task the DR visualization is used for. In the use case IKON the similarity of research projects is based on the embedding of their project abstracts into a multidimensional space. This embedding is produced by an ML-pipeline and is reduced into a two-dimensional scatter plot computed with the dimensionality reduction algorithm t-SNE. To explore the design requirements of the interface to be developed in this thesis, I conducted semi-structured interviews with HCI (human-computer interaction) experts and with help of existing research I uncovered the drawbacks of parameter value selection with classical interaction elements, such as numerical sliders, for non-technical experts.
The thesis is divided into two main tasks: The development of an interface on the one and the development of sorting measures for DR visualizations on the other hand. The goal of sorting visualized results is to enable comparing and evaluating the results against each other. For this, several metrics were explored and tested on 4 data sets and compared with a mathematical as well as a visual measure. I chose two metrics, which were deemed most suitable for the tasks: the first is a parameter intrinsic to the t-SNE algorithm. The second metric is obtained through a secondary, higher-level dimensionality reduction of the result space, again with t-SNE. The interface aims to enable parameter selection with a result-based visualization approach. I embedded the sorting of results into the interface prototype in a grid-like small multiples visualization, which I developed in a user-centered design process. Evaluation of the prototypical interface and the ordering measures was done through a pilot study, which provided evidence supporting the hypothesis that a result-based approach has the potential to support non-technical experts with parameter selection for different tasks, but also revealed new questions and possible applications.

Zusammenfassung

Da die Zielgruppe der Menschen ohne technische Ausbildung oder Expertise (nachfolgend nichttechnische Experten genannt) immer noch zu selten Fokus der Forschung über Anwendbarkeit von Machine Learning (ML) Systemen ist, wurde in dieser Arbeit ein neues visuelles Interface untersucht, welches informierte Nutzung der ML-Ergebnisse für sie ermöglichen soll. Es baut auf einer prototypischen Anwendung aus dem Forschungsprojekt IKON auf, in der die Ähnlichkeit zwischen Forschungsprojekten, berechnet durch einen ML-Algorithmus, in einem zweidimensionalen Streudiagramm visualisiert wird. Um diese Visualisierung zu erstellen, wird die Einbettung der Forschungsprojekte anhand ihrer Projektbeschreibungen in einen ML-generierten, multidimensionalen Raum auf zwei Dimensionen reduziert. Da durch die technischen Parameter der in dieser Applikation gewählten Dimensionalitätsreduktion mit dem Algorithmus *t-SNE* unerwünschte Artifakte auftreten können und es für diese Anwendung nicht *eine richtige* Lösung gibt, ist eine individuelle Anpassung der Parameter von Vorteil. Durch semi-strukturierte Interviews mit nichttechnischen Anwendern wurden Anforderungen für Parameterwahl untersucht und aus der Forschung ging hervor, dass klassische Interaktionstechniken, wie z. B. einfache numerische Schieberegler, für die Zielgruppe nicht hilfreich sind. Das Interface versucht deshalb, das Auswählen von Parametern durch eine ergebnisbasierte Visualisierung für Menschen ohne ML-Erfahrung zu vereinfachen. Damit das Vergleichen und Bewerten zwischen den einzelnen Ergebnissen und im Kontext unterstützt werden kann, wurde dazu eine Sortierung der Streudiagramme angestrebt, die menschlicher Ähnlichkeitswahrnehmung möglichst nahe kommen soll.
Zu diesem Zweck wurden mehrere Metriken, unter anderem Scagnostics, was schon in früheren Studien untersucht worden war, an 4 verschiedenen Datensätzen exploriert und mit einem mathematischen und visuellen Maß verglichen. Das beste Resultat erzielte eine Sortierung durch zwei verschiedene Metriken, von denen eine ein schon vorhandener Parameter des t-SNE Algorithmus ist, während die andere durch erneute Anwendung von dieser Dimensionalitätsreduktion auf einer Meta-Ebene produziert wurde. Diese Sortierung wurde in ein prototypisches Interface eingebettet, welches in einem benutzerzentrierten Prozess entworfen wurde. Zur Überprüfung des Prototypen inklusive der Sortierung wurde eine Pilotstudie durchgeführt, welche das Potential der ergebnisbasierten Darstellung für die Interaktion nicht-technischer Experten mit ML-Ergebnissen aufdeckt , andererseits aber auch neue Fragestellungen und Anwendungsmöglichkeiten anregte.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Goal . . . 2
- 1.3 Method . . . 3

2 Background **7**
- 2.1 Related Work . . . 7
- 2.2 Algorithms . . . 11
  - 2.2.1 Scagnostics . . . 11
  - 2.2.2 T-SNE . . . 13
  - 2.2.3 Other Ordering Measures . . . 14
- 2.3 Use Case Project IKON . . . 15

3 Prototype Design **19**
- 3.1 Design Requirements from Interviews with IKON Researchers . 19
- 3.2 LoFi-Prototype . . . 23
- 3.3 Ordering Measures and Sampling . . . 26
  - 3.3.1 Benchmark Data Sets . . . 26
  - 3.3.2 Implementation . . . 27
  - 3.3.3 Euclidean Distance Comparison . . . 31
  - 3.3.4 Sampling Method . . . 34
  - 3.3.5 Performance . . . 35
- 3.4 HiFi-Prototype . . . 36
  - 3.4.1 Design Decisions . . . 36
  - 3.4.2 Implementation . . . 37

4 Evaluation - User Study **41**
- 4.1 Technical Interface . . . 41
- 4.2 Study Setup . . . 42
- 4.3 Tasks . . . 43
- 4.4 Results . . . 45
- 4.5 Feedback Questionnaire Results . . . 47
- 4.6 Discussion . . . 49

5 Conclusion and Outlook **51**

References **54**

6 Appendix **61**
- 6.1 Pilot Study Forms . . . 61
- 6.2 Benchmark Data Sets Graphics . . . 63

6.3 Euclidean Distance To Neighbours 68

List of Figures

1.1 Color Picker 2
1.2 Random ordering of DR results 3
1.3 Research design 5

2.1 Viscoder 9
2.2 Generative Graph Layout 10
2.3 Scagnostics 12
2.4 Official IKON.projektor prototype 17

3.1 Taxonomy 21
3.2 Click Dummy Screenshots 23
3.3 Exploration of Ordering Measures 30
3.4 Visualizations Euclidean Distance 32
3.5 Final Interface 38
3.6 Class Diagram Interface 38
3.7 File Structure Interface 39
3.8 Color Blindness 40

4.1 Technical Interface 42

6.1 Consent Form 61
6.2 Feedback Questionnaire 62
6.3 Perplexity Ordering 64
6.4 Learning Rate Ordering 64
6.5 Silhouette Ordering 65
6.6 Outlying Ordering 65
6.7 Stability Ordering 66
6.8 Spearman Ordering 66
6.9 t-SNE Ordering 67
6.10 t-SNE and Perplexity Ordering 67
6.11 T-SNE Measure 69
6.12 Smallest Distances Measure 69
6.13 Mean Jaccard Measure 70
6.14 Stability Measure 70
6.15 Silhouette Measure 71
6.16 Learning Rate Measure 71
6.17 Clumpiness Measure (Scagnostics) 72
6.18 Sparsity Measure (Scagnostics) 72
6.19 $r_{spearman}$ Measure (Scagnostics) 73
6.20 Outliers Measure (Scagnostics) 73

List of Tables

3.1 Chosen t-SNE Parameters of the 3 participants 20
3.2 Typical tasks and their related factors in the taxonomy from Figure 3.1 22
3.3 Euclidean similarity ranking, a value around 0.52 means no better than random 33

4.1 Self-reported demographic data of the participants 43

1 Introduction

1.1 Motivation

As machine learning applications become ubiquitous in many fields, non-technical experts need to be able to adapt their output to their tasks and needs without professional knowledge about their implementation. There is evidence that users tend to not critically reflect on an outcome if they perceive it as the only result of an algorithm [SHW17]. Not only is a risk that the power and certainty of an algorithm can hence be overestimated, a disadvantageous effect is also that the application as a whole can be dismissed if the representation does not fit the viewers perspective or needs.

Specifically, selection of appropriate algorithmic parameters, that can hugely influence a result, remains a challenge. Interactive parameter selection with sliders or radio buttons can mislead laypeople to causally relate the numerical values of parameters to the machine learning (ML) output, but the actual influence of those values is usually much more complex [KNJ+20]. Adding to that, manipulating parameters to get a desired result takes a considerable amount of time, even for experts. This implies that state-of-the-art technical interfaces for parameter selection lack support for a person without ML background knowledge to find the result, which helps them best in solving their tasks. Furthermore these technical interfaces do not incite a non-technical, more subjective and open-ended exploration of possible results.

The task to be explored in this thesis is choosing the preferred dimensionality reduction (DR) parameters for an ML model, which is an open challenge in this field especially with non-technical stakeholders as a target-group [KNJ+20, SBIM12]. DR algorithms are prevailing tools for the visualization of multi-dimensional data, essentially mapping it onto a 2D (or 3D) scatter plot. Findings in a scatter plot, visual cues, such as outliers or clusters, can represent patterns in the underlying data.

An analogy to how this task is solved is selecting the best geographical map projection for a specific use. If you want to navigate a boat, the *Mercator* [SVU89] representation might be chosen, as it is conformal, that is, it preserves angles, but at the same time, heavily distorts shapes represented on the map. If you want to compare countries' sizes on the other hand, the *Winkel-Tripel* [SVU89] projection might be more suitable, as it minimizes distortion of areas (as well as direction and distance). There is no *one-fits-all* solution, trade-offs have to be made, although the input data, in this case describing the shape of the Earth in 3D, is always the same. Furthermore, there is no objectively correct way of choosing a projection, as not everyone uses the same strategies to solve a task. In our case, being 'better' at selecting a DR vi-

sualization means that humans are more successful at finding a result that contains the information they need to solve their specific tasks. Importantly, a non-technical expert who does not view at least a small selection of results in the case of dimensionality reduction, is not aware that compromises between different criteria *can* even be made to find the preferred outcome.

Visual ordering of different results may help the viewer to think of criteria by which to select one of them as a suitable outcome. Another way to look at the problem is therefore the analogy of choosing a color in a color picker as can be seen in Figure 1.1, an interface that is common in graphical tools. Instead of randomly trying different RGB-values until you find the color that fits your imagination, you can already get an overview of the entire, ordered color space in the interface and select the desired color, without directly choosing technical values.

The thesis aims to explore new visual interfaces for result-driven selection of parameters for dimensionality reduction in an ML application developed in the project IKON [Ben], i.e., interfaces that let non-technical experts choose parameters retro-actively based on outcome. The ordering of possible outcomes that is shown should support a more systematic selection of an embedding that preserves specific information needed for a certain task [SZS+16]. Tasks could be, for example, examining a single element or getting a rough overview over the space.

Figure 1.1: Screenshot of a typical color picker (taken from [Com18])

For the visual interface I investigate the approach of *small multiples* by finding a diverse representation of the latent space (of DR results) through a subset of instances. *Small multiples* visualizations are a popular way of showing changes, e.g., over time [BBL12], or over the scope of alternatives as is the case here. In the following I outline the specific goals and tasks of this thesis, which were motivated by the preceding paragraphs.

1.2 Goal

The task of this thesis is to investigate how a result-based visual interface in the form of small multiples might help non-technical experts to select suitable

parameters for dimensionality reduction. The goal is to make the diversity of equally valid DR outputs comprehensible through a variety of results so that an individual can make an informed decision. This small multiples visualization should be ordered by measures, that not only help inform about the latent space but also enhance the ability to quickly find a desired layout. The outcome is a prototypical interface, which is built into the prototype of the project IKON, the use case for this thesis detailed in Section 2.3.

Two main questions are deduced:

1. Which visual similarity measures help non-technical experts (especially in the specific scenario described later) to
   - grasp the entire space of possible DR outcomes?
   - successfully find the most useful visual representation of the information needed for their task?
2. How can a small multiples view of such ordered visualizations be designed and embedded into a visual interface that further enhances and simplifies the result-based selection and exploration of DR parameters?

The necessity of ordering DR outcomes, if you want to visualize them via a small multiples approach becomes clear, when looking at a random ordering as visible in Figure 1.2

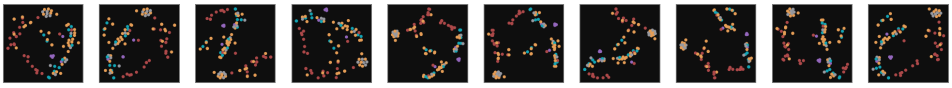


Figure 1.2: 10 results of reducing the IKON data set (described in Section 2.3) to 2D with different parameter values. They are randomly ordered and comparison between them is therefore difficult.

As the goal is to make interactive DR parameter selection accessible to non-ML-experts, the assumption is that their ability to understand the technical parameters and tune them appropriately is limited and therefore more intuitive orderings need to be found.

1.3 Method

This thesis follows an approach that consists of two main parts:

1. The generation of one or more visual measures for the ordering of DR outcomes that support finding the outcome best fit to solve specific tasks and get an overview over the result space.
2. The development of an interface with interaction and navigation techniques, which further help non-technical experts to choose their preferred visualization.

In iterative rounds, those two sides slowly converge to meet in the final prototype. The user centered design [Fou] approach, as pictured in Figure 1.3, has two parallel cycles of iteration. One of them concerned with the user interface, the other with the more technical task of ordering the small multiples visualization in this interface. Due to the extent of this thesis, formative evaluation is limited to informal cognitive walkthroughs on my own or with the help of a few HCI researchers and only in the summative evaluation a limited pilot study is conducted. The following steps are a more detailed description of the process:

1. Conduct interviews with IKON researchers to understand the requirements of a DR outcome suitable to their tasks. These interviews can be found in Section 3.1.

2. Develop ordering measures:
   a) Analyze the 2D embedding of research projects based on the topic extraction pipeline from project IKON, which Section 2.3 summarizes, for a better understanding of the visual and inner characteristics of the DR results.
   b) Analyze existing literature to compile an array of visual ordering measures applicable to a small multiples technique for DR results (summarized in Section 2.1).
   c) Explore and evaluate potential candidates on benchmark data sets and the IKON data set (described in Section 3.3).
   d) Find a representation of the complete latent space spanned by the parameters of t-SNE by sampling the result space in a useful way (to researchers at Museum für Naturkunde Berlin) that is supported by the selected measures.

3. Design and implement an interface prototype:
   a) Find tasks, goals, and perspectives tied with selecting an appropriate DR result in general and at IKON. This process builds on prior qualitative research in the project (e.g., [BMBK19]) and the interviews (Section 3.1, Section 3.2).
   b) Iteratively refine a prototypical visual interface with User-Centered-Design methods in cooperation with the researchers at IKON who have extensive knowledge on the needs of the specific user group. For this, similar solutions are taken into account [KM20, CHAS18], Section 3.4 describes the outcome.
   c) Implement a technical interface (e.g., with a slider for each technical parameter) to compare it to the developed small multiples interfaces.

4. Conduct a pilot study to explore the potential of the interface for parameter selection. This is done in an exploratory user study with two

non-technical researchers and one technical researcher at the HCC Lab, comparing the developed interface to the state-of-the-art solution as described in Chapter 4

5. Discuss the results and evaluate the approach of result-driven DR parameter selection interfaces in Chapter 5.

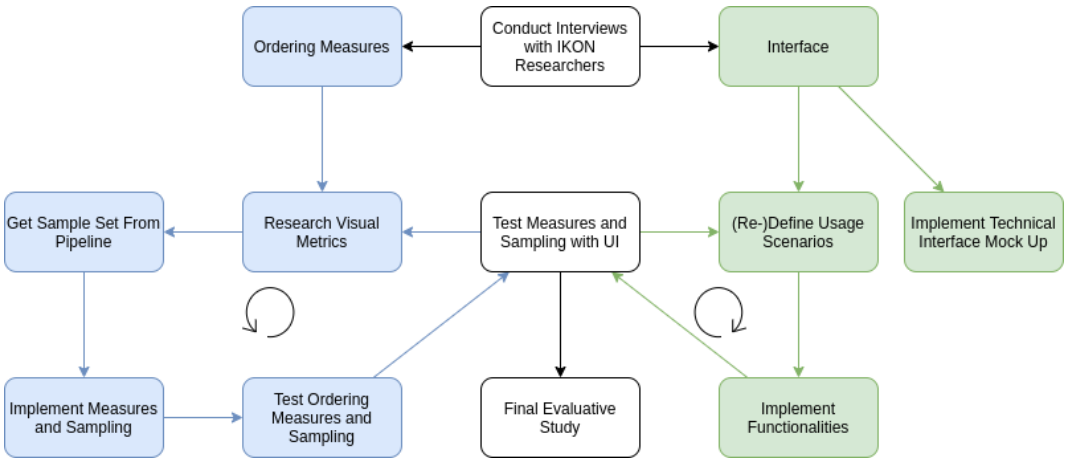


Figure 1.3: Iterative research design

2 Background

2.1 Related Work

The field of visual interaction with ML algorithms, specifically dimensionality reduction and the human perception of scatter plots and DR results has been active in recent years [PKF+16, DW14, ASA+19, TBB+10, AASB19]. Non-technical stakeholders are still rarely target group of new approaches although visual interfaces, where humans are brought together with DR results, have been introduced in some examples – the ones presented here (ScatterNet, VisCoDeR, Graph Layout) have an approach similar to the one of this thesis.

The application serving as a use case of this thesis (project IKON) uses t-distributed stochastic neighbour embedding (t-SNE) as a dimensionality reduction algorithm for the visualization of a thematic distribution of a set of research projects computed through a natural language processing pipeline. An array of work about DR techniques and their (usually more technical) evaluations, which help to find visual ordering criteria of DR visualizations, that may support people without formal ML education, is therefore introduced.

Evaluation of Dimensionality Reduction Results

One of the authors introducing t-SNE [vdMH08] conducted a user study showing that the quality of DR outputs is not trivial to measure and explain for non-experts [LvdMdS12]. An article about how to use it effectively also determined common understandings and misconceptions of t-SNE [WVJ16]. It shows that although it is often the best choice for the visualization of ML outputs, the selection of fitting parameter values is difficult for t-SNE. This motivates the idea that parameter selection in its generic technical form is not suitable for people without a formal ML education.

Dang and Wilkinson [DW14] investigate how scatter plots can be clustered based on their *Scagnostics*, a set of graph-theoretic metrics for scatter plots, resulting in an abstraction of highly dimensional data sets to a few, leading scatter plot examples. Although this approach can be successful to find interesting correlations in a diverse and large array of 2D plots, it does not help to find more fine grained differences between DR visualizations and is not based on human perception.

Different groups of researchers created a base knowledge on the human perception of visual similarity of scatter plots and DR visualizations [PKF+16, ASA+19, AASB19]. Their findings are taken into account for the design as well as the evaluation of this prototype. In their study Pandey et al. show, that *Scagnostics* does not coincide largely with human similarity perception [PKF+16]. It must be noted, that the scatter plots they are looking at, are mainly pairwise visualizations of dimensions from multi-dimensional data sets

and can therefore be ordered by, e.g., how striated or how monotonic they are, which is not usually the case with DR results. But this thesis still attains a similar result for Scagnostics as well as other computational metrics described in Section 3.3, when trying to order t-SNE visualizations. The authors find that the criteria *density, orientation, spread, regularity, grouping* and *edges* are important to human similarity-perception of scatter plots. Clusters grouped by humans had especially high relatedness of *density*, *edges* (edges being the prevalent pattern of a plot, such as a diagonal line) and *regularity*, which is of less importance here, as DR algorithms usually avoid creating regular or grid-like patterns.

The taxonomy in Figure 3.1 for clustering visualizations Sedlmair et al. propose [STMT12] has substantial overlap with the visual criteria for the selection of a DR result that were stated in interviews (as described in Section 3.1). It can serve as a secondary resource for a structured approach of verification of ordering measures.

Similar Interface Approaches

Other groups of researchers have developed a variety of different interfaces or systems that have similar approaches or applications to this thesis. A selection is described below.

ScatterNet

ScatterNet builds on the previous research of Pandey et al. [PKF+16] and feeds human-labeled triples of (dis-)similar scatter plots to a neural network. The results outdo the similarity ordering generated by Scagnostics and HOG (Histogram of Oriented Gradients, a feature descriptor in computer vision). Their subject matter is the same set of scatter plots Pandey et al. used, so the similarity that they found is more broad than that of different t-SNE results of the same data. For example, a sparse linear correlation can be distinguished from a dense l-shaped inverse correlation by the neural network. Another distinction is that the measures in my thesis are targeted towards smaller and clustered data, where the difference in clustering of just a few points can be of interest. Theoretically, a neural network could also solve the task of ordering these, but in this thesis a simpler solution appeared to suffice.

VisCoDeR

The metric that is later selected for this thesis, which is ordering an array of t-SNE plots by reducing them to one dimension again with t-SNE, was inspired from the interface Cutura et al. proposed with VisCoDeR [CHAS18]. The goal of their interface is, to "leverage comparative visualization to support learning and analyzing different dimensionality reduction (DR) methods" (page 1). Other than this thesis' prototype, it targets "junior data scientists" and "DR designers" and aims to inform them of the inner workings and outcomes of multiple distinct DR algorithms. Next to other functionalities it features a

2D higher-level t-SNE visualization over the space of results of different DR algorithms and parameter settings. This can facilitate an understanding of similarity between those results for people acquainted with clustering visualizations – for a less technical audience it potentially rather obscures and complicates their tasks. Nevertheless, as will be shown in Section 3.3, t-SNE can produce a visual ordering related to human perception and a small multiples method can bypass the additional complexity and make it accessible for non-technical experts.

In Figure 2.1 the VisCoDeR interface visualizes a set of t-SNE results for the IKON data set with different DR methods. A pronounced pattern is visible especially in the t-SNE result that could be explained by the rotation of clusters around the center of plots.

Figure 2.1: Data set of this thesis (a set of t-SNE visualizations with different parameters) fed to the VisCoDeR user interface, which visualizes it through different DR algorithms [CHAS18] (Screenshot from VisCoDeR application at [Cut17])

A Deep Generative Model for Graph Layout

A result-based approach with a similar goal as this thesis, is the visual interface for generating desirable graph layouts by Kwon and Ma [KM20]. The authors created an "interface to generate a layout they [the users] want, without either

blindly tweaking parameters of layout methods or requiring expert knowledge of layout methods" (page 1). Therefore different graph layouts are interpolated and made accessible in a 2D sample grid. Their understanding is that there is no *best* layout of a graph and a trade-off between different aesthetic criteria needs to be made. The result is a generative approach with a *WYSIWYG* interface, that enables users "to effortlessly generate a desired layout of the input graph"(page 9). Although the domain of graph layouts is different and continuous morphing from one layout to another with the help of a neural network is the main focus of their work, their problem and outcome are similar to the goal of this thesis: a result-based abstraction from technical parameters for non-technical stakeholders.

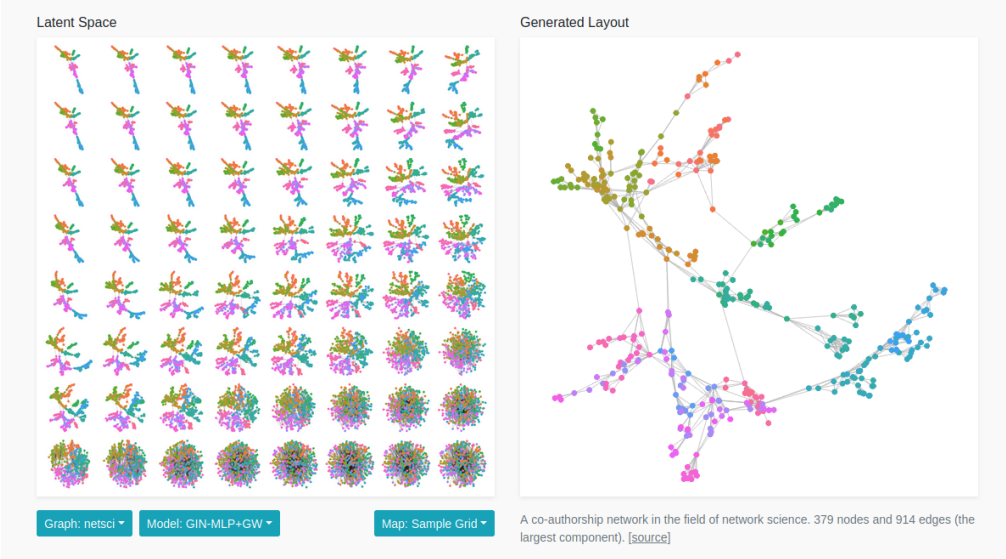


Figure 2.2: User interface for graph layout selection, screenshot taken from [KM20]

Visualization and Interaction Techniques

An overview on literature that covers visual interaction with DR summarizes existing methods [SZS+16]. Sacha et al. specifically state that they "did not identify a mature, ready-to-use system [...] for result-driven parameter tuning of DR algorithms" (page 4). Because of the mentioned difficulty with parameter selection in t-SNE, it is an especially promising algorithm for such a result-driven approach. In interviews Sedlmair et al. [SBIM12] analyze gaps and mismatches between research for DR interaction and real world applications. According to this analysis, current algorithms and tools contain "difficulties understanding and trusting the visual layout of reduced data sets..." and "...selecting, which algorithms to use and realizing when one has reached a stopping condition" (page 9).

As outlined in the introduction, a small multiples interface design appears especially suitable to non-technical stakeholders needs and is investigated. This

technique is deemed more effective as a tool for viewing and comparing variations or changes over time than e.g. animation, because comparison is instantly possible without waiting time or interaction [BBL12].

For the comparison, interface designs, such as a overview+detail or focus+context visualization, animation or zooming can be considered [CKB09]. In this case I chose an overview+detail approach similar to the graph layout interface (Figure 2.2), where overview and detail can be seen at the same time. This and other decisions are attended to in Section 3.2.

Most of the related approaches described in this chapter will be attended to again in the following chapters, especially when describing the definition of design requirements and the prototyping process.

2.2 Algorithms

Since I decided to order the small multiples visualization of DR results, which is the main element of the prototype, in a way that supports human perception, I tested different computational metrics for their applicability. Most of those metrics have been applied in similar situations before, their main workings are shortly described in the following.

2.2.1 Scagnostics

As the two-dimensional ordering of research projects computed with t-SNE is essentially a scatter plot, quality metrics for scatter plots can be investigated as a visual ordering measure for a small multiple of those DR results. *Scagnostics* originally proposed by Tukey and Tukey [TT85] for the analysis of scatter plot matrices have been extended by Wilkinson et al. with more graph-theoretic metrics [WAG05].

In contrast to the original goal of Scagnostics to statistically analyze multidimensional data, in this thesis (analogous to [PKF+16]) the aim is to order scatter plots based on a visual (and possibly semantic) similarity that can only be verified by the perception of humans with domain knowledge. I examined Scagnostics to verify what Pandey et al. already concluded in their user study: that human similarity perception and Scagnostics hardly correlate with each other. My approach also resulted in it not being useful for the IKON use case, although some similarity ordering can be seen with some of the measures.

Scagnostics is based on a representation of the scatter plot points in a Delaunay triangulation, convex hull, and a minimum spanning tree (illustrated in Figure 2.3). The measures are categorized in outlier, density, and shape measures. A detailed description including derivation and complexity on all measures can be found in [WAG05], the measures that seemed worth investigating are summarized here. As the shape of a t-SNE result is usually not informative for the clustering and in first experiments they did not show visible effects, the shape measures *convexity*, *skinniness* and *stringiness* were not included so the convex hull computation was not necessary.

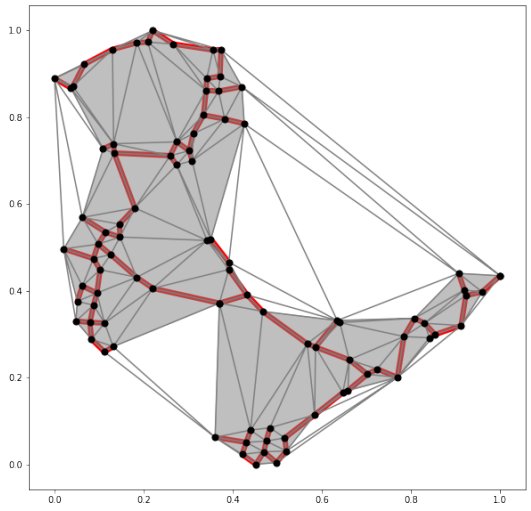


Figure 2.3: Scagnostics' graph theoretic approach: delaunay triangulation (thin lines), α - shape/ convex hull (grey area), minimum spanning tree/MST (red lines)

In the following the measures, which are explored in this thesis, are described:

Outliers: Calculate the 25% quantile and 75% quantile of edge lengths in the MST. Long edges are edges that meet following criterion:
long edges = $quantile_{75} + 1{,}5(quantile_{75} - quantile_{25})$

The outlier factor is calculated like: outliers = $\frac{sum\ of\ long\ edge\ lengths}{sum\ of\ all\ edge\ lengths}$

Skewed: Skewedness is a measure for the distribution of edge lengths, which approximates the relative density of points:
skewed = $\frac{quantile_{90} - quantile_{50}}{quantile_{90} - quantile_{10}}$

Clumpy: For each edge e_j remove it from the MST and then find the longest edge in the smaller of the two now disconnected components. The clumpiness factor is defined as:
clumpy = $max_j(1 - \frac{max(edgelength(k)\ \ for\ k\ in\ smaller\ component)}{edgelength(j)})$

Sparse: Sparsity is represented by the amount of long edge lengths:
sparse = $min(1, quantile_{90})$

Monotonic: Monotonicity is not usually a measure that can be applied to t-SNE visualizations, since a visible monotonicity is an artifact in t-SNE dimensionality reduction. Nonetheless a possible similarity ordering based on monotonicity was tested. Scagnostics is using the squared Spearman's rank correlation coefficient for this purpose.
monotonic = $r^2_{spearman}$

2.2.2 T-SNE

Dimensionality Reduction with t-distributed stochastic neighbour embedding (t-SNE) plays a central role in two parts of this thesis:

- To produce a diverse array of visualizations of the IKON data set by systematically changing t-SNE's parameters.
- To compute a visual similarity measure over the entire set of results.

Therefore it is crucial to not only understand the effect different parameters have on the outcome but also the metric (Kullback-Leibler divergence) applied for the arrangement of points.

T-distributed SNE is a variation to SNE, which Hinton and Roweis introduced in 2003 [HR03], and t-SNE was itself introduced by Hinton and van der Maaten [vdMH08]. It is an established method for dimensionality reduction in practice as it handles the crowding problem originating from the *curse of dimensionality* [BCC57] especially well and also preserves local structures. An important factor why, as mentioned earlier, t-SNE is a good candidate for a result-based approach is that, in opposite to other DR methods, such as PCA, it can have multiple local minima and is therefore non-deterministic.

T-SNE creates a Gaussian probability distribution, which defines relationships between the points in high-dimensional space, and uses a t-distribution (rather than a Gaussian like in SNE) to recreate the high-dimensional distribution in low-dimensional space. A gradient descent minimizes a single Kullback-Leibler (KL) divergence between a joint probability distribution in the high-dimensional and the low-dimensional space. KL divergence is a measure of how much information is lost with the projection to the low-dimensional embedding:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The *perplexity* parameter $Perp(P_i)$ gives an estimate on the amount of neighbours an element has in the multi-dimensional data set. Larger perplexity values concentrate on the global and small values on the local structure of an embedding. It specifies the variance of each points Gaussian distribution using its Shannon entropy $H(P_i)$:

$$Perp(P_i) = 2^H(P_i)$$

$$(H(P_i) = -\sum_j p_{i|j} \log_2 p_{j|i})$$

The other parameter varied in this thesis is the learning rate, which is a factor influencing the step size between each iteration of the gradient descent. When the learning rate is too low, a very dense cluster with few outliers far away from it might occur, if it is too high, this might result in a cloud of approximately equidistant points.

2.2.3 Other Ordering Measures

In addition to Scagnostics and before considering t-SNE as a possible measure, some other metrics can be compared:

Smallest Distances

A measure I implemented before finding Scagnostics is fairly similar in its effect to an inverse of the Scagnostics parameters for density. It produces an estimate on how densely packed the local clusters are:
Sort all euclidean distances between points in ascending order. Sum up the n smallest distances and divide by n, n being the amount of points in the visualization, resulting in an average over the n smallest distances:

$$\frac{\sum_{i=1}^{n}(distances[i])}{n}$$

Mean Jaccard Similarity

Shape-based metrics have successfully been applied to evaluate or order graph layouts as described in [EHKN15]. The authors of this article proposed an array of proximity graphs, which can also be used for scatter plots although they do not naturally have edges. One of those is a minimum spanning tree, which is calculated for each t-SNE result in this thesis and compared to one arbitrarily chosen result with the *Mean Jaccard Similarity* as done in their method.

Silhouette Score

A silhouette score compares the distance of samples to their cluster center and to the nearest cluster center [Rou87]. It gives an estimation on how well clusters are separated. The IKON data set is a real-world data set and hence well-separated clusters can not be expected in it. The clustering here is only defined by the first subject area a research project has been tagged with, e.g., "Life Sciences". This does not mean, that it cannot be similar to projects with other main subject areas. In other test data sets the silhouette score performed well in defining a visible similarity.

Stability Measure

In an attempt to stabilize the rotation of the points to the center of the plot in the small multiple grid I arbitrarily selected a point i and calculated its *rotation* on the unit circle with:

$$\text{stability} = \arctan(\frac{x_i}{y_i})$$

All metrics described here are explored and tested in Section 3.3, but first the use case, which this thesis deals with, is described in more detail.

2.3 Use Case Project IKON

This thesis is based on previous work in the project IKON on a real application scenario with non-technical experts. To comprehend the decisions as well as limitations of my work, it is therefore important to understand what the project is related to. IKON investigates potential for knowledge transfer through mapping research projects at the German Natural History Museum Berlin (MfN). Since over 300 people work at this institution, it becomes a challenge to promote their exchange of internal knowledge [Ben]. The project includes the development of a visualization, that helps the researchers to find new possibilities for knowledge transfer by seeing how projects are ordered thematically. For this, a natural language processing pipeline has been introduced, which computes the similarity of research projects through their project-abstracts [Kor19]. The result is visualized in 2D through a dimensionality reduction with t-SNE, in which semantically similar projects ought to be close to each other. The *uncertainty landscape* layered behind the project layout, conveys with a greyscale the degree of confidence of the algorithm for the location of each research project in the 2D representation [KKB19].

In project IKON, interviews and a workshop [BMBK19, BKH+20] with MfN researchers have created insights on their preconceptions of the way research is structured at the museum. Usability tests have been conducted with the latest prototype at the time. Participants seemed to rarely question the locations of the research projects in the visualization, expressing the semantic similarity between them. A further insight from usability tests was that, as a consequence of their preconceptions of, for example, the hierarchical structure at the museum, they are set in their *explanation strategies* [Hub15]. For the *uncertainty landscape* and the DR visualization they "only consider a specific subset of all possible types of explanations as valid."(page 4) [Kor19]. This leads to this thesis' hypothesis that only seeing one possible visualization of the projects does not support their understanding of the diversity of perspectives. Furthermore, if an array of visualizations is provided, this could help draw researchers attention to the variety of connections between as well as the diversity of research projects.

Pipeline and Data Set

I integrated the prototypical interface developed in this thesis into the existing prototype of project IKON. The machine learning application it is built on is a natural-language-processing (NLP) pipeline. The pipeline steps are:

1. Loading and cleaning of the training data and the IKON data set
2. Document embedding with HDP (Hierarchical Dirichlet Process) [TBJB03]
3. Dimensionality Reduction into a 2D scatter plot with t-SNE

As a training data-set the pipeline receives the abstracts of about 114,000 projects funded by the DFG (German Research Foundation) [(DF18]. The

pipeline then embeds the set of at this point 92 publicly released projects at the MfN into this multidimensional space. The museums projects and detailed information and linking are gathered from the internal wiki VIA [MfNB18].

The original pipeline is slightly adjusted to obtain a diverse array of visualizations through systematical changing of parameters of t-SNE namely the learning rate and perplexity.

Existing Prototype

The current front-end of the IKON project, *IKON.projektor*, is built using the React framework [Rea20] and some of the visualizations are implemented with Data-Driven Documents [Bos20]. The code can be found on `https://github.com/FUB-HCC/IKON-projektor` [WJEO]. The final visual prototype of this thesis also uses JavaScript with React to ease its incorporation into the project.

IKON.projektor incorporates three views, putting the data about the research projects at MfN into different contexts. *ZEIT* and *RAUM* dealing with temporal and spatial information about the research projects are not touched for this prototypical approach, while the *WISSEN* view showing the embedding with the NLP-pipeline described above, is subject of the small multiples selection interface to be built. It shows research projects as colorful dots in the middle, connected with arcs to an outer circle representing knowledge transfer activities and infrastructures used in the respective projects. With a sidebar menu the projects can be filtered depending on their research area or timeframe. The *uncertainty landscape* can optionally be layered behind the project dots. The official IKON.projektor attempts to solve the problem that a suboptimal embedding could be computed that, e.g., has overlapping project points, indirectly with the simplification of the embedding to a hexagonal grid. With the possibility of selecting ones preferred visualization through the small multiples approach investigated in this thesis, this transformation becomes obsolete.

Project IKON's results are necessary to help define requirements of this thesis in the following chapter. Their description also helps to understand the integration of the final prototype and its evaluation.

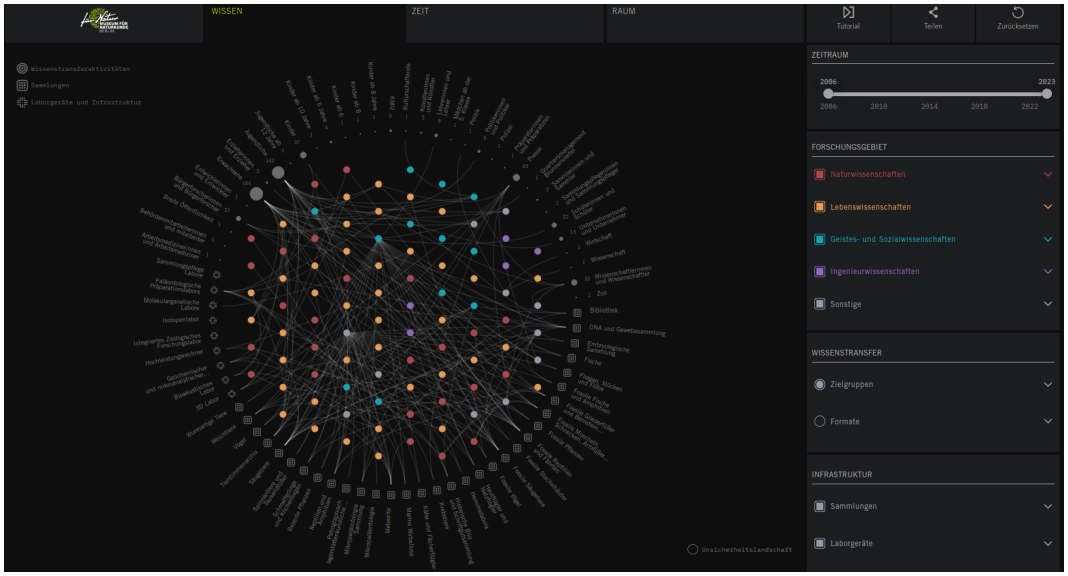


Figure 2.4: *WISSEN* view of the official IKON.projektor prototype with hexagonal grid [WJEO]

3 Prototype Design

The description of the design process for the interface developed in this thesis starts with a requirements definition, which comprises interviews with IKON researchers about the interface used at the moment and their feedback on click dummies I made in the first prototyping iteration. This is followed by a visual exploration and testing of the different ordering metrics I had already described in the algorithms chapter (Section 2.2) to find one that fits those requirements. Afterwards the implementation of the interface into the *IKON.projektor* with the integrated metrics is described.

3.1 Design Requirements from Interviews with IKON Researchers

To gain more insights on the scenario and requirements in project IKON, especially on the reasoning for choosing a DR visualization, semi-structured interviews with researchers of project IKON were conducted. The goal was to help define visual criteria, which support researchers at MfN with the selection of a result fitting their interests. As mentioned before, the results of these interviews are overlapping with the taxonomy Sedlmair et al. created for clustering visualization [STMT12].

Three researchers, who are part of the IKON project, one of them a t-SNE expert and two non-technical experts, were asked to select a visualization with the current technical interface, that consists of two sliders for perplexity and learning rate and the visual result, and reflect on their selection. In the semi-structured interview their personal preference as well as their estimate of the potential user groups interests was asked for. The interviews were conducted in German and the main findings are translated and summarized here.

Summary of Interviews

1. **Change the parameters until you find a visualization you would chose.**

   The parameter values each participant chose as well as the amount of trials and time the selection took them is seen in Table 3.1. While person 1 systematically tried out extrema of both parameters, person 2 and 3 had a more random approach and also voiced this during the interaction.

2. **How did you approach the manipulation and selection of the parameters?**

Person	1 (technical)	2 (non)	3 (non)
Perplexity	5	42	7
Learning Rate	10	72	3
# Trials	9	45	53
Duration	1:30 min	10 min	15 min

Table 3.1: Chosen t-SNE Parameters of the 3 participants

Like presumed the technical expert had a more structured approach in changing the parameters, as he understood their influence, whereas the other two were changing the parameters at random, although both were aware and noted that "small perplexity values make smaller clusters". They partly had misleading assumptions about the influence of the parameters, e.g., "The learning rate needs to be high for better clustering" (person 2). Person 3 said, they would not try to understand the technical parameters but instead "see the task as a game". The difference in approach is also reflected in the amount of trials and the time.

3. **Which parameters or which combinations of technical parameters do you think are most important?**

   The technical expert explained why perplexity is the most influential parameter to him while person 2 guessed that the learning rate had a higher effect and perplexity was random and person 3 guessed perplexity to be most important but did not have a reason.

4. **Which visual (or thematic) criteria did you apply to judge a visualization outcome?**

   Person 1: cluster separation, notable patterns, thematic connection between grouped elements, as suspected by someone without domain knowledge
   Person 2: cluster separation, distribution over the space, size of the clusters, difference in sizes of clusters, overlapping
   Person 3: interesting patterns, distribution of the subject areas, cluster separation, thematic composition of clusters

5. **How does the resulting visualization compare or relate to your interests?**

   Person 1: "interests are curiosity and wanting to explore [...] the chosen plot reflects my understanding of the underlying data."
   Person 2: "looks beautiful, patterns are visible [...] I am not an MfN-employee and can not make guesses about thematic correctness."
   Person 3: "It is an interesting pattern: one more broad cluster, small clusters around it seem to be expert fields [...] points are overlapping to much, but I would like to have the same with different perplexity values."

6. **How does the resulting visualization compare or relate to the information you have about the interests of researchers at MfN?**

Person 1: “MfN-researchers are less interested in global structure. [...] It is important that there is no overlap between points.”
Person 2: “MfN-researchers have diverse tasks and goals, some are only interested in thematic correctness of their own field, some want the hierarchical structure of the museum to be visible (e.g., subject areas well separated).”
Person 3: “MfN-researchers are diverse, some are more open and miss a broader perspective currently, others are mainly interested in their field or the structuring of subject areas.”

Deduced Tasks and Criteria

The interviews were conducted to refine, which criteria and tasks are related to selecting a t-SNE outcome. The gained information can be roughly separated into general tasks, that are comparable to Sedlmaier's taxonomy [STMT12] on the one hand and knowledge about the preferences of the target group in the IKON use case on the other. Factors that were also mentioned by the participants are marked in Figure 3.1 depicting the taxonomy.

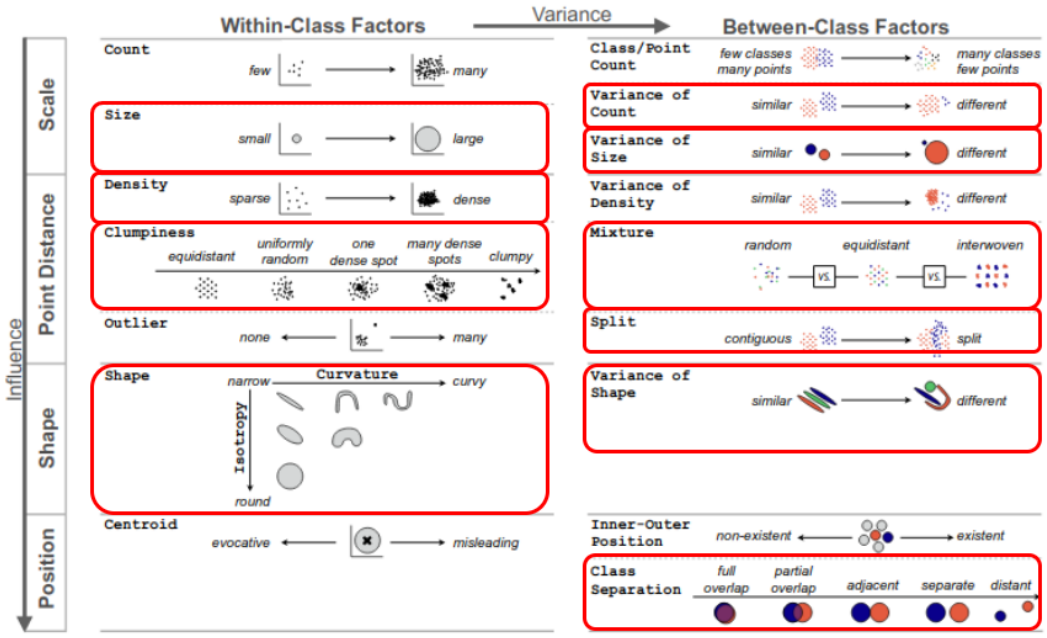


Figure 3.1: A taxonomy of visual cluster separation factors taken from Sedlmair et al.'s article [STMT12]. Factors also mentioned in the interviews are marked in red

There are some differences to the taxonomy that are based on the specific data set: Point count is not mentioned, because it stays unaltered in this use case. Classes are not directly inferred from the data, only a coloring through subject areas or the spacial clustering serve as cues, but statements about the perceived clusters are related to the *Between-Class Factors*. Factors that are intrinsic to the IKON data set are the structuring of subject areas and the

No	Task	Taxonomy Factors
1	**Clustering** identifying clusters	density, clumpiness, mixture, split
2	**Cluster sizes** comparing cluster sizes	size, density, variance of count, size
3	**Patterns** finding shapes and patterns	shape, variance of shape, mixture, split
4	**Details** exploring single points or neighbourhoods	density, clumpiness, class separation
5	**Subject areas** separating subject areas (classes)	class separation, mixture, split

Table 3.2: Typical tasks and their related factors in the taxonomy from Figure 3.1

thematic distribution of research projects, which can only be roughly assumed by people without appropriate domain knowledge through, e.g., the titles of projects.

In Table 3.2, the tasks are matched with their corresponding factor in the taxonomy. This list of tasks will be used throughout the thesis to validate design decisions. The overall goal mentioned in the introduction, to support exploring and getting an overview over the result space is concerned more with the ensemble of visualizations and navigation in the existing prototype. Nonetheless the factors play a role in facilitating an overview – the 5 tasks are implicit sub-tasks of this exploration. The finding of Pandey et al.'s study [PKF+16], that *density* and *edges* are closest related to human perception, is reflected here, as those were also often mentioned by the interviewees and are part of the *clustering* and *patterns* tasks.

Additional Findings

A parallel in all three interviews was, that the interviewees complained about the time it took to find a good visualization. Every participant stated at some point during the interaction that they had found a representation of which they liked some parts, but would change other parts. For example, person 3 found a clustering they liked, but points were overlapping, which they would have liked to change while preserving the global structure. Person 3 also noted, that the technical complexity and time involved with this interface is a hurdle that would prevent non-technical stakeholders to engage in a more exploratory and playful interaction. Person 1 mentioned that it seemed impossible with this interface to balance between criteria; as an example he used the trade-off between well-separated clusters and overlapping points.

These findings together with the defined tasks speak for the need of a more approachable interface for example with a small multiples grid, where an array of different results is ordered by similarity.

3.2 LoFi-Prototype

Before the actual implementation, hypotheses about the design were tested with simple click dummies made with diagrams.net [Tea]. All click dummies were bundled in one HTML document that was sent to two HCI researchers. In a video-meeting with them, the design hypothesis and interaction techniques of these first prototypes were discussed. The final design of the HiFi-prototype is based on the results of this discussion.

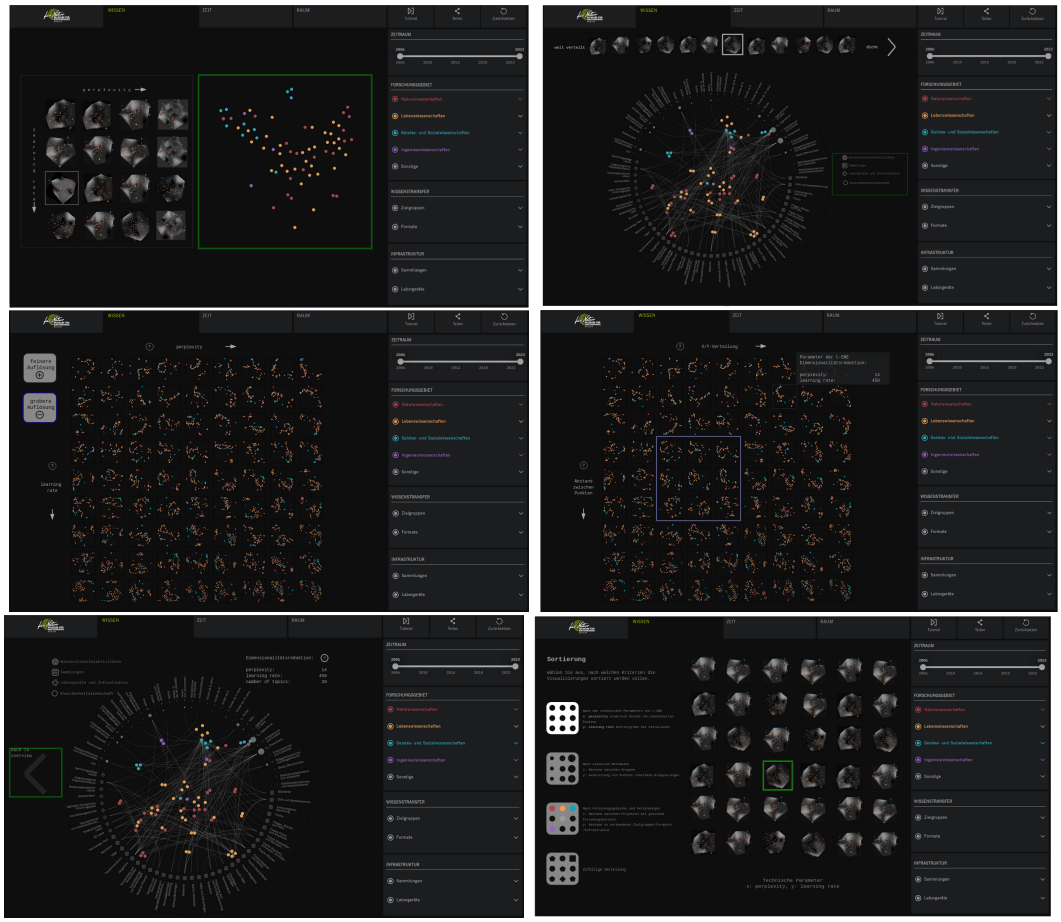


Figure 3.2: Screenshots of different click dummies to test design hypotheses. Clickable elements are marked with green or blue frames. Top-left **1** to bottom-right **6**

In the following a selection of defining usage hypotheses are described. The implementations can be seen in Figure 3.2. Pages of the click dummies, which did not add new findings in the discussion were left out of this description.

1. **Side by Side** (a 4x4 grid of visualizations on one and a single one on the other side of the screen) The amount of possibilities might not be as important, but the user wants to view and compare the details of each option. They want to convince themselves, that one layout groups certain projects better than another (*clustering*-,*details*-task). Two plots

are shown next to each other when selecting one item from the small multiples view on the left side. Clicking on it again returns to the overview.

2. **Ribbon** (A line of different plots over the standard *WISSEN* view) The differences in the DR output might not be the most important element of the interface to the non-technical experts. Only one measure is necessary for some comparison. Therefore a small selection and one-dimensional ordering is enough to find a desired result. It is more interesting for the users to see the projects in context with the labels and the rest of the page (*details*-task).

3. **Grid zooming** (Buttons for zooming in and out of grid) As it is not clear, which grid size is appropriate for the tasks, the viewer is able to change the grid size according to their needs. A coarse view might help getting an overview, while a finer view might enable them to find exactly their desired outcome (e.g., with *details*-task). Here an interaction technique close to zooming on a map is tested, where more detail appears between the former grid rows and columns when zooming.

4. **Grid focus+context**(clicking on corners of grid, enlarges them) Another possible interaction for adjusting grid size is a technique closer to a focus+context interface. The largest grid is also the one spanning the most of the result space and therefore supporting a coarse overview, while clippings show more detailed differences and might support the *details*-task. Hypothetically a user first looks at the preview-like small visualizations and enlarges a corner that might include plots that seem promising to them.

5. **In/Out interaction** As the selection interface can not possibly show all the same elements that the *WISSEN* view (the starting view of the normal IKON interface) shows, they need to be two separate views. Therefore there needs to be a form of interaction to get from one view to the other. The hypothesis here is, that the *WISSEN* view symbolizes a detail view in an overview+detail [CKB09] ensemble.

6. **Criteria Selection** Hypothetically the non-technical experts want to explore the latent space with more than one grid ordering (for example, orderings supporting each of the defined tasks separately) to understand potential criteria for the selection of a candidate. This click dummy already makes a suggestion for some criteria (such as separation of subject areas), the applicability of which might be dependent on the personal strategy. It has a more high-level approach as the sorting criteria are not predefined.

Expert Feedback

Together with the two experts I evaluated the hypotheses and they added insightful critique about other parts of the click dummies:

1. The researchers generally liked the side-by-side technique, although a comparison between two single plots did not prove effective. They also voiced that the side bar for filtering was not necessary in this view and would just take focus away from the main interaction. To layer the uncertainty landscape under every tile obscured and over-complicated the grid view. Therefore the uncertainty is excluded from the selection view in the final prototype.

2. A *ribbon* of different plots would convey that the space of possible results was simply linear, which it is not, and reduce the comparison options and was therefore discarded as a possible visualization.

3.+ 4. The researchers found the interactions for changing the grid size to be too complex. They proposed a simple interaction with a slider to be enough; they also considered to predefine the grid size to reduce complexity. The researchers liked the tooltips shown in click dummy 4 and proposed to have more tooltips and explanations also describing, e.g., the ordering measures and more descriptive labels for the axes.

5. In the discussion the idea that the *WISSEN* view would be framed like the detail view of this prototype was rejected. Instead, it should be up to the user to decide if they even want to open the parameter selection view or just work with the initially selected ordering. I therefore later decide on a simpler navigation.

6. Selecting criteria for ordering is not a task appropriate for the knowledge and interests of non-technical stakeholders according to the IKON researchers. It can not be assumed that the interest in this interface is big enough to enforce the additional step.

The feedback from the experts about the LoFi-prototypes summarized here, serve as a basis for the implementation of the HiFi-prototype.

3.3 Ordering Measures and Sampling

From the design requirements defined in Section 3.1 it becomes clear that an ordering of the small multiples visualization by criteria, which support the mentioned tasks, would greatly enhance the exploratory and comparative characteristic of the interface.
Therefore I evaluate a set of 10 metrics based on their support for the tasks.

3.3.1 Benchmark Data Sets

To assess the effect any of the measures have on the ordering of t-SNE results, 3 benchmark data sets, borrowed from VisCoDeR, were included next to the IKON data set. The sets were sourced from the VisCoDeR website as CSVs [Cut17]. Each element in the sets has a name, a class, and features, that in this case are the dimensions for the t-SNE reduction.

IKON

The 92 projects in the IKON data set are classified through the 4 subject areas they belong to, which is not necessarily a strict classification visible in underlying data. The multi-dimensional embedding of the research project abstracts uses 12 dimensions in this example.

IRIS

The well-known test set IRIS consists of petal and sepal length and width for 150 different iris flowers, which can be divided into three sub-species, so it has 4 dimensions and 3 classes.

MNIST

MNIST is a popular test set consisting of pixel representations of hand-drawn digits from 0 to 9. To prevent a cluttered visualization and to make the results comparable to the IKON plots, 20 instances for each digit were periodically drawn from the entire set, resulting in a set of 200 pixel representations. The digits 0 to 9 are the classes in this set, the 784 pixels the dimensions.

SPOTIFY

The SPOTIFY data set is a real-world data set consisting of a sample of 103 songs with the genre *metal* from the Spotify database. Each entry has 7 dimensions, such as *acousticness* or *popularity* that were computed by Spotify. The sample is divided into 8 sub-genres of metal, which are the classes, resulting in 7 dimensions and 8 classes.

3.3.2 Implementation

The t-SNE algorithm described in Section 2.2 is applied on a high level as a measure that proves as effective to order t-SNE results themselves again. For this, a set of t-SNE results with different parameters, which are stored as lists of coordinates of the projects, are used as an input for a second t-SNE embedding and reduced to 1D. Every point coordinates x and y values hereby represent a dimension (so $0_x, 0_y, 1_x, 1_y, \cdot 91_x, 91_y$) resulting in, e.g., 184 dimensions for each visualization in the IKON set of 92 points. The Python library Scikit [PVG+11] has a t-SNE implementation, which is used as follows:

```
from sklearn.manifold import TSNE
# representation as 184xN matrix
X = np.array([list(sum( [( i.projects[j][0],i.projects[j][1] )
    for j in range(len(i.projects))], ())) for i in self.dumps
   ])
# t-SNE embedding
X_embedded = TSNE(
    n_components=1,
    perplexity=30,
    learning_rate=100
).fit_transform(X)
# assigning the value to each result
for i in range(len(self.dumps)):
    self.dumps[i].tsne_measure = X_embedded[i][0]
```

Listing 3.1: Calculation of high-level t-SNE measure

The entire process and the computation of all the metrics were implemented in Python in a Jupyter Notebook (`https://jupyter.org`) and later extracted into a simple Python program, both found at `https://github.com/lillijo/result-based-dr/tree/master/computations`. The class *EmbeddingResult*, in which most measures are calculated, represents indivdual t-SNE results. The high-level t-SNE measure is computed with the *ResultCollection* class, which contains all results in a list of *EmbeddingResult* instances. This class also produces the dump with a sample of 160 results, which is later on fed to the prototypical interface as a JSON file.

Exploration of Measures

In the exploration of all metrics, for each of the benchmark data sets t-SNE is applied 160 times, each time altering the perplexity and learning rate. After a close inspection of results with different t-SNE parameters a range of *3 to 18* for perplexity and *10 to 100* for the learning rate in steps of 10, proved as sensible. Results outside of these ranges were either dense balls, with few extreme outliers or equidistant points distributed all over the drawing space, without visible patterns. This resulted from experiments for the IKON data but is not necessarily the same for the other sets. To make the visualizations comparable the same ranges are used for every data set nonetheless.

The 160 results are sorted by the respective measure. A sample of 7 results in roughly equal intervals illustrates, which ordering effect the respective measure

has. As an exception the *stability* and *t-SNE* measures are illustrated by taking 7 results that are direct neighbours of the sorting, as the effect can better be seen by looking at neighbours. The described effects for all measures become visible in illustrations in the appendix in Section 6.2.

The effects observed with some measures are described and discussed here – measures that do not produce reliable ordering effects different to the described ones are summarized below.

1. **Perplexity:** Small,dense clusters on the left side progress into a more homogeneous distribution on the right. While the visualizations on the left side support the *clustering*- and partly the *subject areas*-task, the other sides plots are more useful for task the *patterns*- and *details*-tasks. As the differences are quite stark and are representative of the result space of perplexity, this parameter as a measure also enables an overview over the result space. It does not do so satisfyingly though, as points in neighbouring plots are not similarly positioned and difficult to compare.

2. **Silhouette score:** For IKON and SPOTIFY silhouette values are low in general and variation in the well-separation of clusters is hardly recognizable. For IRIS and MNIST the variation in separation is visible, as the underlying data has more clearly separable classes. This measure might be attractive for such data but not for most real-world data sets, such as IKON and SPOTIFY. The assumption that the *subject areas*-task would be easier through this sorting therefore was not verified at least for the use case at hand.

3. **Outliers (Scagnostics):** The outliers measure of Scagnostics produces a similar result to t-SNE's perplexity, although it is inversed: low values indicate a homogeneous distribution, high values dense small clusters with overlapping points. The same is observed for the other density metrics of Scagnostics: skewedness, clumpiness, and sparsity. Perplexity, while producing similar results, has advantages as an ordering measure described later.

4. **Monotonicity (Spearman coefficient, Scagnostics):** Results are sorted by their x/y correlation, meaning top-left to bottom-right distributions are on the left and bottom-left to top-right distributions on the right side. The correlation here has nothing to do with the actual data sets and is an artifact of the t-SNE computation. Although the axis distribution could be interpreted as a shape or pattern, I believe that this measure is not helping the *patterns*-task as it is misleading.

5. **Stability:** A random point is selected and DR results where this point is at a similar position are neighbours in the visualization. For IKON, MNIST, and SPOTIFY the rough location of this randomly selected point can be made out through comparison of the results. For example, in the IKON sample it is part of the prominent round cluster on the bottom, which is at a similar position for every plot. As a well selected

point can not be guaranteed, this measure is not applicable for ordering. Although it could be used to roughly stabilize a sample (and therefore be beneficial for the *patterns*-task) by selecting only samples with similar values for the visualization – hence the name 'stability'.

6. **T-SNE ordering:** For IKON, MNIST, and SPOTIFY a similarity between neighbours is visible and morphs from one side to the other. As the changes happen gradually, this ordering aids in almost all tasks, but especially the *patterns*-task, to a great extent, because it is easier to compare neighbours. For IRIS some similarity ordering is also recognizable but neighbours are not as similar.

7. **Combination of t-SNE ordering and perplexity:** In this combination the high-level t-SNE sorts the entire sample. Then each rows order is set up by their perplexity. A progression from low to high perplexity is visible, while the positions of points stay similar, simplifying comparisons. This combination of measures produces especially good results for IKON but in the other test sets and even artificially generated random sets it had convincing results too.

The *learning rate* as a measure is not a viable solution as it does not result in any human-perceivable ordering in any of the test sets – nonetheless a plot can be seen in Figure 6.4. Sorting by the *mean jaccard* measure, is highly volatile, as it depends on the selected Pivot element. Therefore I decided not to include a plot, as a human-perceivable ordering rarely occurs and is not reproducible.

With the *smallest distances* measure, the ordering is almost identical to that of perplexity. This can be attributed to dense clusters in low perplexity plots and therefore a low average of the smallest distances. It was therefore also not necessary to illustrate its effects again.

The Scagnostics measures *outliers*, *skewedness*, *clumpiness*, and *sparsity* all produced such similar orderings, that only the *outliers* illustration is shown and discussed in more detail. This could be due to the fact that they are all based on long versus short edges in the minimum spanning tree of a visualization. In a plot with low perplexity, there tend to be small, but dense clusters, which are far apart from each other, resulting in a high value for all those measures, whereas a high perplexity results in a homogeneous distribution, meaning that distances are smaller and do not vary widely returning a low *outliers* etc. value.

Conclusions from Exploration

Through the exploration of possible measures and the information gained from the interviews, *perplexity* is chosen as one of the two measures necessary for ordering a grid. The summarized arguments speaking for it are:

- It produces an ordering that is perceivable by humans reliably and helps them to compare neighbours.

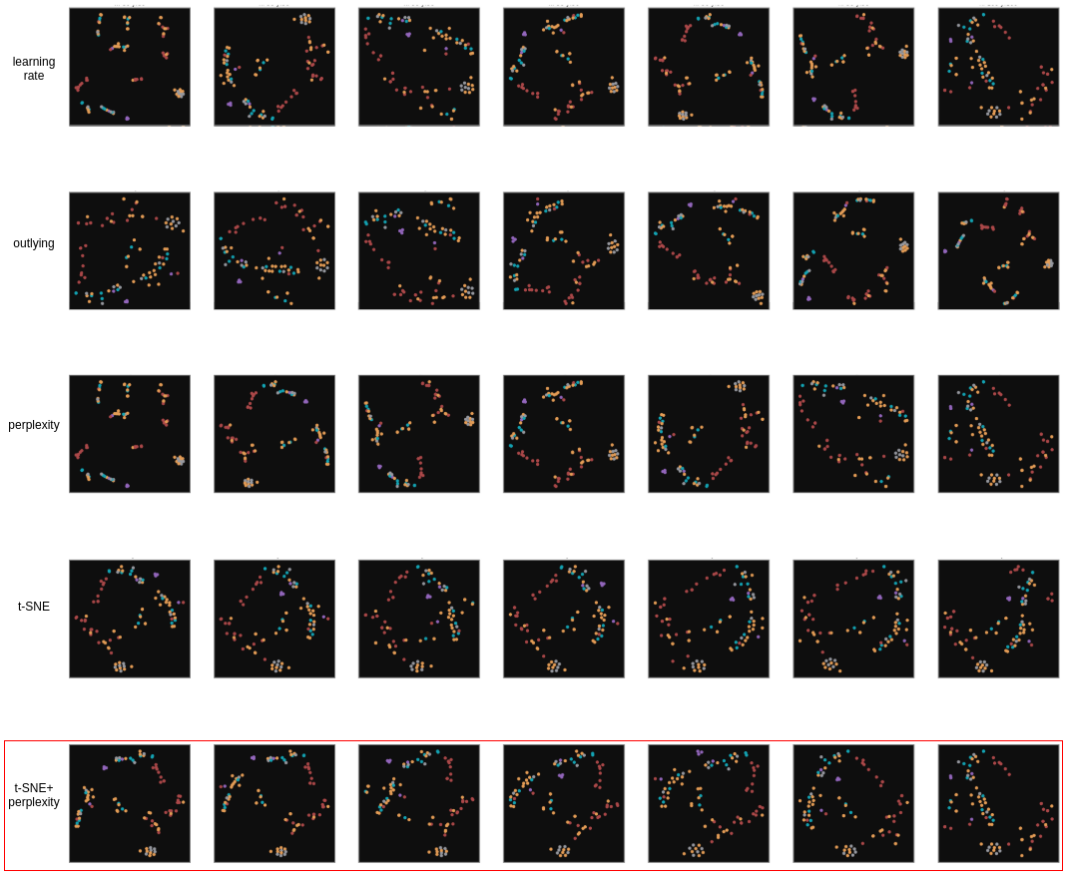


Figure 3.3: Ordering by learning rate, by Scagnostics outliers, by perplexity, by t-SNE measure and by t-SNE measure combined with perplexity for the IKON data set. The comparisons for the other sets and measures can be found in the appendix (Section 6.2)

- Many tasks defined in the previous section are supported through the perplexity ordering, as it reflects many of the factors of Sedlmair et al.'s taxonomy, namely cluster size, density, clumpiness, and variance of count.
- Perplexity is a direct technical parameter of t-SNE, so applying it does not increase complexity and possibly fosters an intuition for the inner workings of the algorithm.
- Five other measures produced very similar results to perplexity, conveying the assumption, that the ordering by perplexity is mathematically explicable.

The insights gained from this visual exploration led to the hypothesis, that the high-level t-SNE embedding as a second measure would be most helpful for the comparison by humans, as it creates more similar neighbourhoods than,

e.g., Scagnostics metrics. This way, a distinct change over the perplexity can be made visible, while other characteristics of the tiles change slowly and gradually over the entire space. To further investigate this hypothesis a second visual and computational exploration was undertaken only with the IKON set, which examines the similarity between neighbours.

3.3.3 Euclidean Distance Comparison

The goal of this neighbourhood visualizations is, to test how similar neighbours are in the grid, based on the average of euclidean distances of all points positions in one visualization to their position in the neighbouring visualization, when sorting by the different measures (Figure 3.4).

In the experiment the average euclidean distance between each tile and its neighbouring tiles in the grid (with shared edges) is calculated only once. This can simply be done without normalization, because all the results are scaled to 1x1 unit squares. It is enough to only calculate the distance to the right and the bottom neighbour of a tile ($i_{x|y}$) if they exist:

$$neighbour_{right} = \frac{\sum_{i=0}^{n} dist(i_{x+1|y}, i_{x|y})}{n}$$
$$neighbour_{bottom} = \frac{\sum_{i=0}^{n} dist(i_{x|y+1}, i_{x|y})}{n}$$

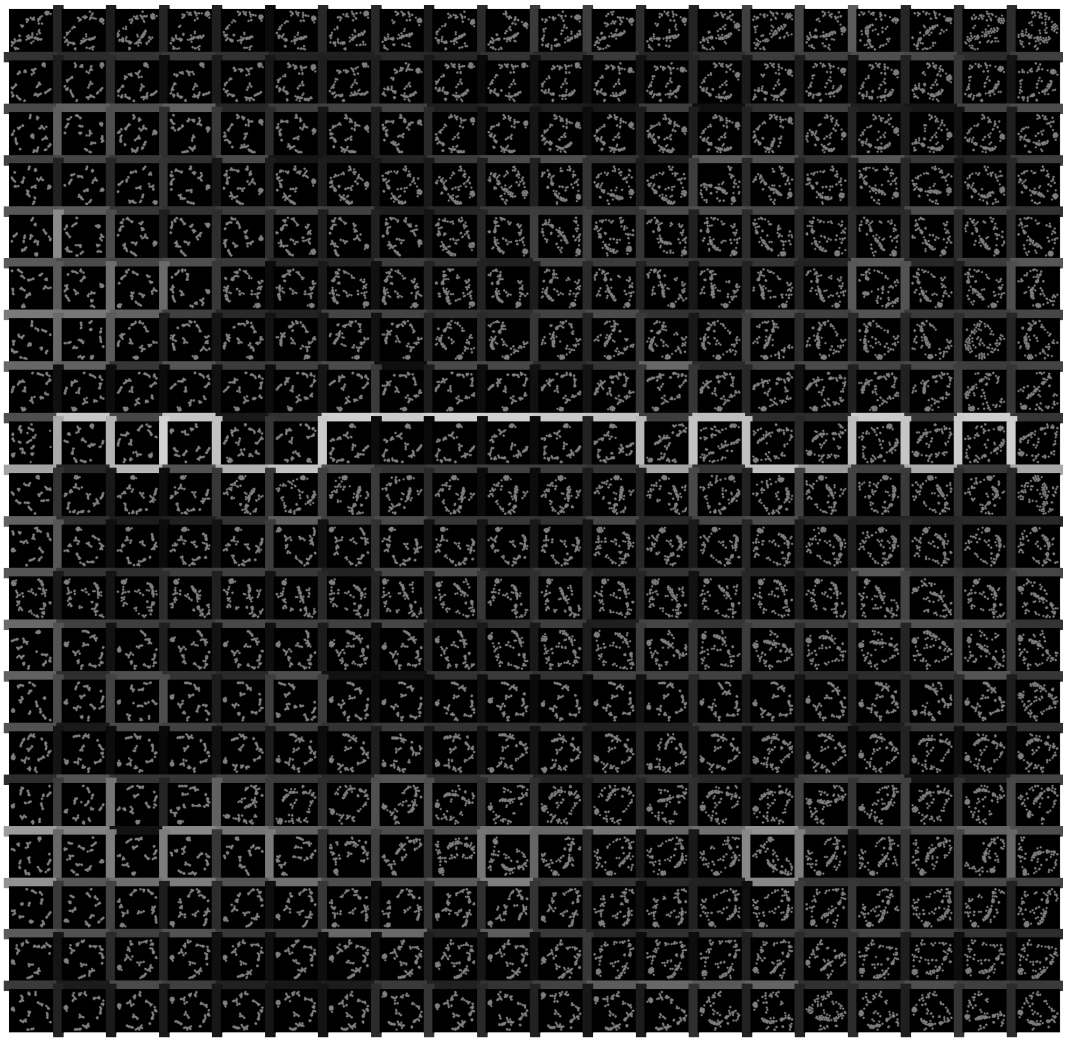
(x is the index of the column, y of the row, n the number of projects, here 92)

The sorting by perplexity on the x-axis is fixed in this exploration of measures, as perplexity fulfills the requirements discussed in the previous section and in Section 3.1. The goal here is, to have tiles with visibly different perplexities but similar other characteristics in each row. The 400 first tiles (20x20) are selected from a sorting of 640 computed results (generated with the method from the previous section) with the respective measure. Perplexity then sorts each row of 20. The plots represent each neighbourhood by coloring the belonging edge – the darker an edge is, the more similar are the two neighbours.

When comparing the results, it is apparent that the t-SNE measure creates the ordering with the best local neighbourhoods (see Figure 3.4). The visualizations of the euclidean similarities of neighbouring tiles for all mentioned metrics are depicted in the appendix in Section 6.3.

A metric to verify this visual comparison is the average distance between neighbours. Any value over the average euclidean distance of two random points in the unit square, so anything over $\frac{2+\sqrt{2}+5\log(1+\sqrt{2})}{15} \approx 0.5214\ldots$ [hd], implies, that the sorting is not better than a random sorting, when it comes to euclidean distance. While most of the measures proposed here produced a sorting under this limit, they are too close to it to be considered significant. The only exception to this are the *stability* and the *t-SNE* measure, which is also visible in the illustration. As mentioned earlier, the stability measure relies on the position of one randomly selected point, which explains, why the position of the cluster this point lies in, is similar for neighbouring tiles but

Figure 3.4: Visualization of similarity between neighbouring tiles based on the euclidean distance of the individual points inside the visualizations. Dark edges mean strong, light edges weak similarity. Here, the tiles are sorted through the perplexity and t-SNE measure, the other measures can be seen at Section 6.3



the stability measure is still worse than t-SNE, because it only takes one point (and indirectly its cluster) into account.

Other arguments speaking for the application of a high-level t-SNE embedding in the IKON scenario, leading to its selection, are:

+ Instead of trying to direct the viewers focus to hardly human-perceivable, metrical criteria, the t-SNE measure enables an impartial, perceivable sorting. This is especially helpful for non-technical experts as their experience with mathematical measures is limited and their decision-strategy is less technical and therefore also less predictable.

+ the t-SNE ordering is supporting especially the *patterns*-task, because

1	t-SNE	0.223
2	stability	0.379
3	learning rate	0.491
4	monotonicity	0.495
5	outliers	0.512
6	mean jaccard	0.515
7	sparsity	0.518
8	smallest dist.	0.52
9	silhouette	0.52
10	clumpiness	0.535

Table 3.3: Euclidean similarity ranking, a value around 0.52 means no better than random

shapes and patterns change gradually over the sample space and therefore become more emphasized. Recurring clusters become more visible (helping the *clustering*-task), as they only slightly mutate and are easily trackable.

\+ Through the neighbourhood of similar plots where, e.g., only the location or rotation of clusters varies strongly, a sense of (un-)certainty over the individual t-SNE results can be fostered. For example, a cluster that is visible in every tile of the grid and only slightly changes position in opposite to single points that completely change their position and clustering in every visualization.

Although the t-SNE measure shows the best result in this test, it also has some downsides:

– A white horizontal line (explained by less similar neighbours) dividing the grid in the middle is visible in the example in Figure 3.4. This is due to the fact, that t-SNE does not produce a linear ordering, but instead tries to find groupings. The edges between clusters (or rows) are therefore not necessarily smooth.

– As t-SNE is non-deterministic, the result differs slightly with every computation and can potentially also produce worse euclidean similarity values.

– A differentiation between plots that show good class separation (here based on subject area coloring) and ones that do not (*subject areas*-task), is not made explicit by this measure and neither perplexity. But the silhouette score aiming for this, does not produce satisfactory results either, at least in the IKON set. I assume this to be due to the nature of the underlying data, where a projects subject area is not decisive of its similarity to others.

– The measure is not revealing additional information other than (euclidean) similarity. In other words, it is not clear, which characteristics

plots have, e.g., in the top row, only that they are similar to each other. However, since the perplexity does produce a distinctive ordering in every row, I assume this to be of less importance.

The proposition that the t-SNE measure creates an ordering similar to human similarity-perception is based on the experiment and my own human perception – to verify it a user study similar to [PKF+16] is needed. As this studies negative findings for Scagnostics metrics coincide with the results presented here, it would be insightful to see how the t-SNE ordering performs with their approach.

3.3.4 Sampling Method

To select a comprehensible amount of plots for the *small* multiples visualization, a sampling method is necessary. A good sampling should span the latent space diversely and hereby provide an overview over the result space. Since in the interface the size of the grid is adjustable, a challenge is, to find a method, that produces good results for smaller and bigger sample sizes. For the purposes of this thesis a simple sampling seemed sufficient, though more sophisticated, e.g., non-linear approaches may ultimately yield more diverse yet complete results.

The 3 tested variations of the sampling each have advantages and disadvantages, when selecting a subset of n sorted results.

1. **Naive sampling:** n results are randomly chosen from the 640 results. The random method as it was to be expected often produces incomplete samplings, e.g., with almost all selected plots having a low perplexity.

2. **Periodic sampling:** the 640 results are sorted by the perplexity measure and n samples are selected in equal intervals. The advantage of this method is that, regardless of the grid size, the selection is always diverse in perplexity, the downside is that especially for smaller sizes, the (t-SNE) difference between neighbours can be big.

3. **Stabilized sampling:** the results are sorted by the stability measure, then the first n are selected, having a more similar stability measure, meaning that one element stays at a roughly stable position. This might ease comparison, because neighbours have the same rotation, but specifically for smaller grid sizes might limit diversity too much, as plots, where one element is at the same position tend to also have other elements distributed similarly.

I selected the periodic sampling, as it guarantees diversity in perplexity, the factor most relating to the needs mentioned in the interviews and the taxonomy. Although, for small grid sizes the effect of the t-SNE sorting might become less noticeable, as neighbours are too different, the perplexity remains a visibly changing factor. In addition, one could also argue that comparing fewer samples is less complex and therefore slow, gradual mutation is needed

less. If large grid sizes are selected on the opposite, the t-SNE (and perplexity) ordering is obvious with the sampling having smaller effects, as almost every result is included anyway.

It must be noted, that there is a trade-off between diversity and comparability when sampling the result space, which in this case was decided in favor of diversity.

3.3.5 Performance

The once-off computation of 640 results from the HDP/t-SNE pipeline required about 30 minutes on an Intel Core i5 2.30GHz x 4 with 8GB of RAM, although for the final visualization not all 640 results are needed. Computing all measures for those embeddings, which includes the calculation of the high-level t-SNE embedding and the Scagnostics measures, takes about 5 minutes each time. The majority of this runtime is due to to the graph-theoretic measures, where a minimum spanning tree has to be constructed for each visualization and computations on this graph quadratic to the project count are made for, e.g., the *clumpiness* measure. If the high-level t-SNE embedding of the 640 results is computed alone, this takes less than 10 seconds.

In this section 10 measures were tested for the applicability in the ordering of a small multiples grid of DR visualizations. The t-SNE parameter *perplexity* and a higher-level t-SNE computation over the result-space were selected as the most appropriate measures for solving the tasks defined in the requirements section. A simple sampling method was selected to dynamically create a subset of the results with the required size. I integrated these algorithms into the HiFi-prototype described in the following section.

3.4 HiFi-Prototype

This chapter is concerned with the decisions resulting from former prototyping iterations and describes the implementation of the HiFi-prototype. Since I had previously co-developed the *IKON.projektor* [WJEO], the visualization prototype of the project IKON, I was familiar with its implementation, architecture and styling. The decision to base the prototype of my thesis on it was therefore not only due to the potential incorporation into the official prototype but also because of my experience working with it.

The repository for my prototype can be found at `https://github.com/lillijo/result-based-dr`. For trying it out, the non-public IKON data about the research projects is required as well as the pre-computed DR results.

The design of the high-fidelity interface prototype is backed by the insights from the discussion of the LoFi-prototype as described in Section 3.2. The defined tasks are also taken into account for each decision made. After the implementation of the findings from the expert feedback, I refined details of the design with further feedback from my supervisor. Another influence into some decisions are the selected ordering measures and sampling method, on which the interface design itself also had an effect. The most important design decisions are:

3.4.1 Design Decisions

- From the feedback gathered with the LoFi-prototypes (mainly click dummies 1 and 2) and also inspired by the interface of Kwon and Ma mentioned (Figure 2.2), I selected the side-by-side overview+detail view as the most suitable basic layout for the interface. This was further motivated by the ordering measures, as they only make sense when applied simultaneously to a 2D grid. The spatial separation (opposed to, e.g., temporal in an animation) in a overview+detail technique also maximises the focus on one element while still enabling its comparison with the others, hereby improving support for all tasks.

- Based on the discussion of click dummies 5 and 6 for the navigation from the *WISSEN* view of the IKON prototype to my selection interface a simple button was chosen. In a later round an icon depicting a stylized version of the selection grid increased visibility and intuitive understanding.

- It is still possible to change the grid size but with a simplified interaction through a small slider on the bottom. This is because a fixed grid size might be poorly chosen and not optimal for every task. While a large amount might facilitate exploration, smaller grid sizes are better for seeing finer patterns and single points (*patterns*- and *details*-task).

- As the experts positively received the tooltips in previous iterations, every tile in the grid reveals its perplexity, learning rate and the t-SNE

measure used for the ordering on hover. This can support the understanding of the ordering. In the detail view the project dots show their title on hover just like in the standard IKON prototype. For MfN researchers the title should be enough to identify a project or at least its research topic. A person without domain knowledge would probably at least be able to make an assumption about the projects subject area.

- The axes also received tooltips, which explain them shortly. As a standard it seemed obligatory to add axes and a description to a (more or less scientific) data visualization.
- For navigation from the selection interface back to the *WISSEN* view, first a simple button saying 'Auswählen' (ger. for select) had been implemented. After an HCI researcher noted that a user might change their mind and want to keep the old selection, a 'Abbrechen' (ger. for abort) button was added too.
- A transition of 1 second between visualizations, when a new visualization is selected, is incorporated into the prototype, inspired by a bachelor thesis based on the same use case, that investigates the application of animated transitions. According to Camara's pilot study, transitions with a period between 0.7 and 3 seconds help with the understanding of change and identification of single project points [Cam19].
- The filtering interaction from the *WISSEN* view is transferred to the small multiples view, as it supports the *subject areas*-task.

3.4.2 Implementation

The code for the grid view as well as the technical interface are incorporated into the prototype as pages in the form of React components. Many decisions about the implementation were more passive, as the focus is on fully incorporating the interface into the existing project. For the purpose of this prototype it is sufficient to precompute all results and save them in JSON files, which initiate the state of the app.

Mainly, the components *SelectionGrid* and *TechnicalUI* were added to the project. The new *OverviewButton* component with buttons for navigating to both new views is a sub-view to the *ClusterMap*, which implements the standard *WISSEN* view from the IKON prototype. In the *Reducer* (the state controlling component of the React app) an initial ordering (indicated by indices in the array of orderings) and actions for changing the ordering as well as the grid size replaces the static ordering. The *selectedState* is a tuple of two integers, which represents the ordering selected at the moment and the previously selected ordering, to allow the user to abort their selection process. Figure 3.6 shows a class diagram mostly omitting components that had already been part of the IKON prototype and do not interact with the new components. The class diagram also abstracts or simplifies some methods and attributes as the actual data flow happens over a global state.

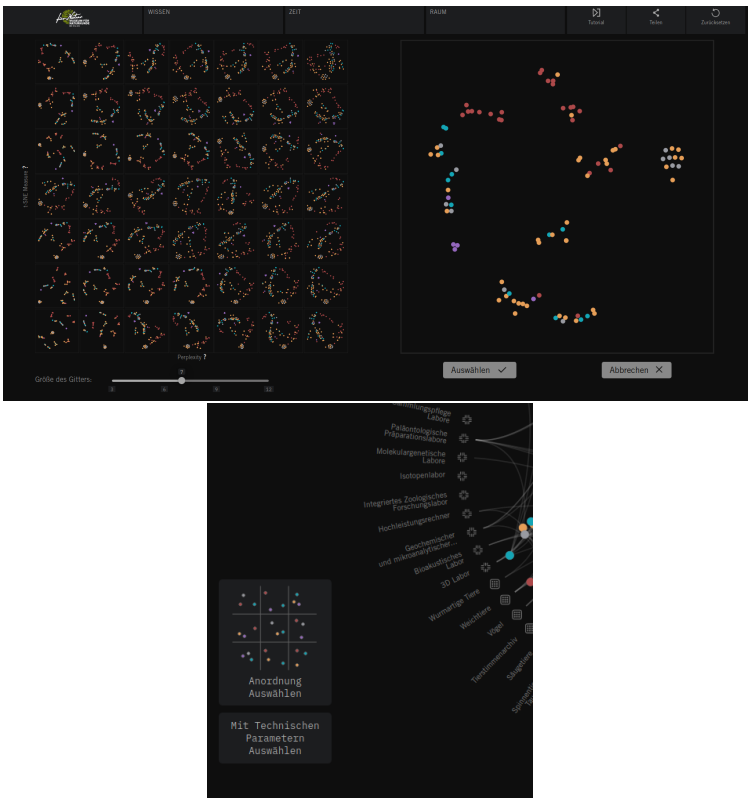


Figure 3.5: Integration into IKON.projektor, final interface on the left, navigation buttons with stylized icon on the right

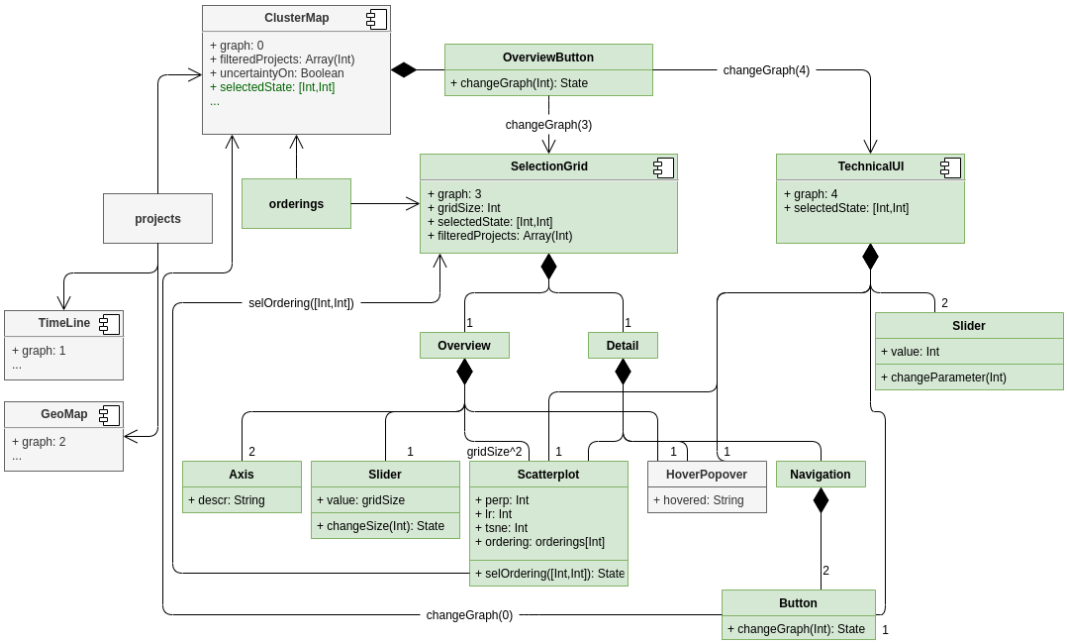


Figure 3.6: Class diagram, green are elements implemented for this thesis, grey are main components from existing implementation

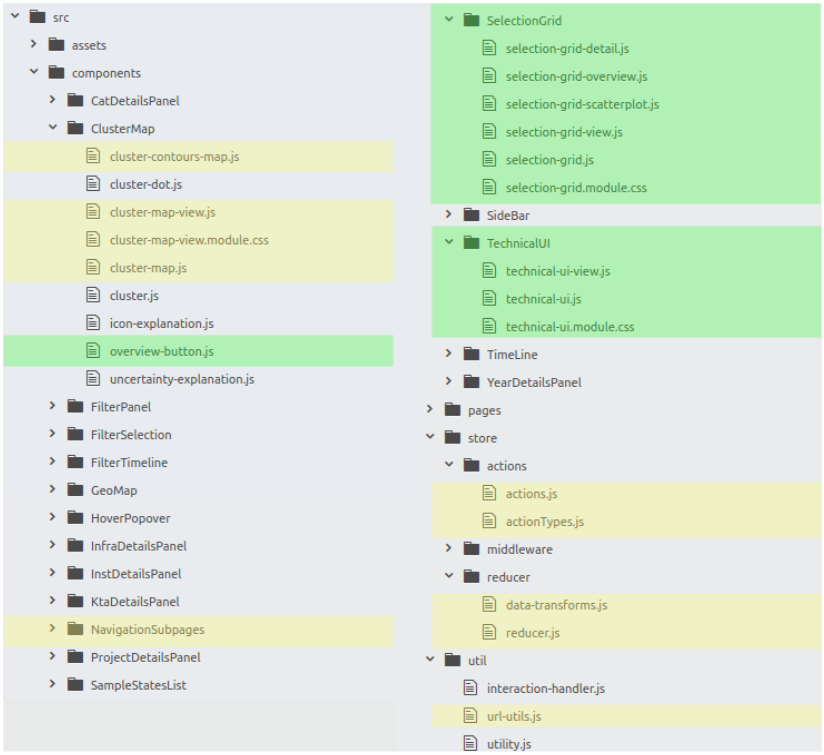


Figure 3.7: File structure, green are elements implemented for this thesis, yellow are changed files, grey are unchanged files from existing implementation

Color Deficiency

To test, how a color deficiency would influence the usability of the prototype, I used an online tool that simulates color blindness [MW]. The simulation for the most common, red/green deficiency is seen in Figure 3.8. None of the filters seemed to influence the usability of the interface greatly, as the coloring of the single points is still visible and also, selection of a visualization is still possible without the colors. The only task affected by color deficiency is the *subject areas*-task. All other elements do not use colors for coding. As will be shown in the evaluation in Chapter 4, the coloring can become somewhat unusable when the grid size is too large, resulting in very small points.

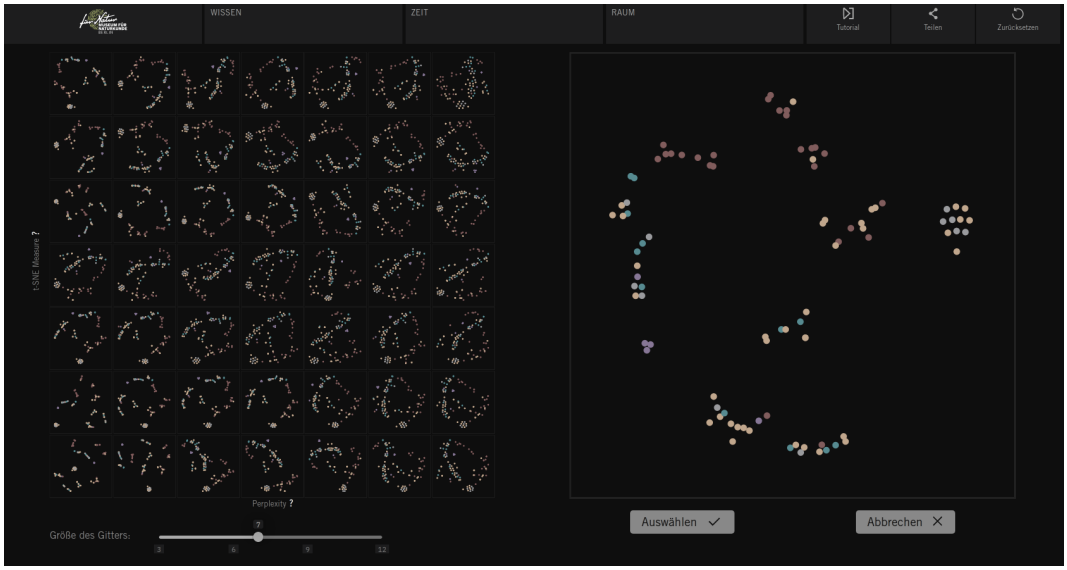


Figure 3.8: Simulation of red/green deficiency

4 Evaluation - User Study

The goal of the user study conducted for this thesis is a summative evaluation to gain insights into how people interact with the small multiples interface in comparison to a more technical interface with parameter sliders. Due to the limited extent with only a small sample of three participants, a bigger study should be conducted in the future for more certain results. As I had gained experience with the underlying t-SNE algorithm, the data as well as with selecting appropriate results, I needed a less biased opinion on the interface and the ordering. The tasks and questions I proposed to the participants are motivated by the general tasks I have defined through the interviews and Sedlmair's taxonomy. I also test design hypotheses I materialized through the feedback rounds with the HiFi-prototype. Questions that I hoped to get first answers on included but are not limited to:

1. Which interface (small multiples or technical) is preferred for the tasks formulated here?
2. Is the small multiples interface a useful enrichment for non-technical experts, how do they interact with it?
3. What are disadvantages of the small multiples interface?
4. Is an adjustable grid size helpful?
5. How should the ordering (with t-SNE and perplexity) be explained and is it noticeable and useful for non-experts?
6. How is knowledge or ignorance of the underlying algorithm influencing the interaction?

4.1 Technical Interface

A mock-up, which represents the current status of the selection interface in the IKON project (*technical interface* in the following), is integrated into the prototypical interface for the pilot study. The components of this interface are equivalent to what IKON researchers had previously used to select a t-SNE embedding. The incorporation enables direct comparison between the previous and the new interface and gives the participants the choice over which interface to use. It consists out of two sliders for perplexity and learning rate and one plot showing the t-SNE result. As this is only a prototype the computation is mocked with an array of 160 pre-computed results with different perplexity and learning rate values that show after an artificial delay of 1.5 seconds. For integrity this mock-up interface has the same styling as the rest of the prototype.

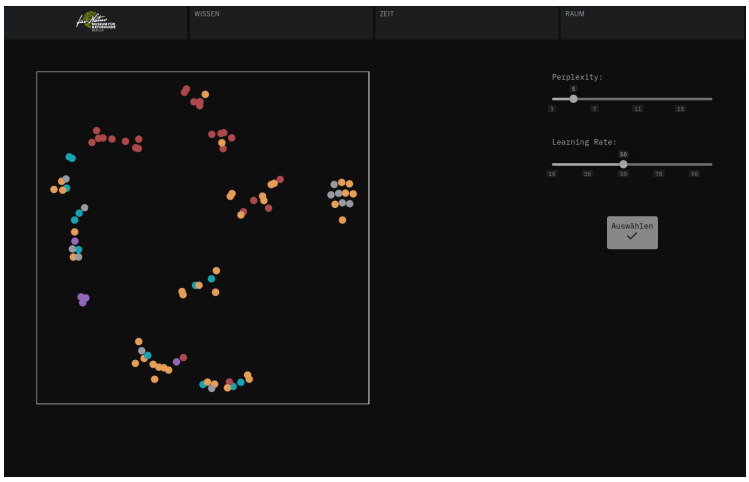


Figure 4.1: Technical interface mock-up with numerical slider for the parameters perplexity and learning rate used for comparison in the pilot study

4.2 Study Setup

For the pilot study three researchers (different to the ones from the semi-structured interviews described earlier) working in the field of Human-Computer Interaction were presented with the prototypical interface on a monitor along with tasks. Each candidate gave consent to recording the sound and taking notes on their interaction with the prototype (consent form template see Section 6.1). They were asked to *think-aloud*, meaning they should say everything coming to their mind out loud and explain every action (as described in [SBS94]). The tasks were read to them and they could also see them on a second monitor or on paper. After the tasks each participant received a feedback questionnaire (see Figure 6.2), also including demographic questions (such as expertise, age etc.), as for such a small sample size this information can not be omitted. The study took about one hour each, the first two interviews were conducted in an office that was kept as neutral as possible, while the third unfortunately had to be done in a coffee shop and on a smaller monitor. Those differences are taken into account when comparing and discussing the results. All interviews were recorded mostly in German (interviewees sometimes switched to English shortly) and insightful passages were translated by me and are presented here. In short, I varied the task of selecting a visualization result by asking for different use cases and characteristics of the visualization to be selected.

Participants

Since all participants in this study are working in the field of human-computer interaction and have a basic knowledge about the IKON projects goals and premises, the explanation given about the interface and project was kept rather short.

The first participant represents a *technical expert* as he has worked with plenty of machine learning algorithms, knows about dimensionality reduction and is acquainted with t-SNE. He was invited for the study to compare how expertise influences interaction with the prototype. The second participant is a social scientist and is maintaining a non-technical perspective. He has conducted user studies himself and knows common mismatches between non-technical stakeholders and (more technical) interfaces not only from his personal experience. Participant three has a masters degree in computer science, is concerned with a different field of research and never heard about the t-SNE algorithm before. Although she is per definition a technical expert, she has never interacted with the algorithms at hand before and therefore has a similar starting point as MfN researchers might have.

Person	1	2	3
T-SNE/ DR experience	knows t-SNE	no experience	no experience
Use of visualizations	daily	about once a week	about once a month
General field of work	Research assistant in HCC	Social Science, Science and Technology studies, HCC researcher	Research assistant in HCC
Degree	Master, Computer Science	Doctor, Social Science	Master, Computer Science
Age	28	39	>29

Table 4.1: Self-reported demographic data of the participants

4.3 Tasks

In the following the tasks and questions are presented, including the introduction texts:

Training

The project IKON is finding new ways to visualize natural language processing (NLP) results. The data set you will engage with today is a collection of research projects that have been conducted at the Museum für Naturkunde Berlin (MfN). With help of a data set of over 100'000 other research project abstracts, the collection has been embedded into a multidimensional space based on their abstracts. An array of 2D representations of this space has been computed with the dimensionality reduction algorithm 't-SNE'.

One output of this representation is currently shown in the IKON prototype in front of you. The dots in the middle represent research projects that are

colored based on their main subject area. Those research projects are linked to collections or laboratory devices (infrastructure), if they have been used in the project. A project is also linked to target groups or formats if a knowledge transfer activity was conducted, for example a workshop (format) for school-children (target group) was carried out on its topic.

Imagine the following scenarios and, while solving the tasks ahead, try to voice your thoughts, actions and impressions aloud about everything you do. This may feel odd at first, but its essential so that we learn from this test. Remember that we are not testing you: you are testing the prototype.

1. *Click on and hover over different projects, target groups, and infrastructures. Use filters. Explore the interface.*

2. *Find two ways you could change the ordering of the research project dots.*

3. *Change the grid size in the grid view.*

Try to solve the following tasks first with the technical parameter selection and then with the grid interface.

ROUND I

1a *Select a visualization where the points are evenly spread over the space. Ignore the coloring of the projects for this task.*

1b *Select a visualization where you can see clearly separated clusters. Ignore the coloring of the projects for this task.*

1c *Select a visualization where subject areas (i.e., the colors of the projects) are closest together.*

ROUND II (MfN Researcher)

You are working as head of a research group at the Museum für Naturkunde Berlin (MfN). You have been sent this new visualization of the research at the museum by one of your colleagues. You are interested in collaborating with other researchers and research groups at the museum. You are hoping that this visualization might help you understand the similarities between research projects and possibly find local or global connections that could inspire new collaboration. Your main research interest is paleontology and geology (a sub field of natural sciences). You know that some of your colleagues at the museum are often thinking more about the organisational structure of the museum rather than actual contextual similarities of research projects. This latter structure is visible in this prototype, as projects are colored according to their main subject area.

2a *Select a visualization that you would like to use to show other people how interdisciplinary the research at MfN is, that is, a visualization which has strong overlap between different subject areas.*

2b *Select a visualization that groups research projects together that may be part of your research area (paleontology and geology) or similar to it.*

ROUND III Feedback Questionnaire (see Section 4.5)

4.4 Results

Each participant's interaction with the interface and their statements for each task are summarized below. A word-by-word transcript of the sessions was excluded from this thesis, because everything not mentioned here was either already mentioned by another participant or not contributing to the critique of the interface in a new way.

1A and 1B

The tasks 1a and 1b, which relate to the *clustering- cluster sizes-* and partly *details*-task of the design requirements defined in Section 3.1, were easy and quick to solve for all participants with the small multiples interface by shortly scanning the according side of the grid (left for 1a, right for 1b). Person 1, who has experience with the t-SNE algorithm, was also able to quickly solve them with the technical interface as they understood the influence perplexity has on the result. For task 1a in the technical interface, person 2 chose the first visualization they got after changing the parameter perplexity once, seeming uninterested in reiterating but not pleased with the result. For task 1b they tried to find a result in the technical interface for a long time by randomly moving the parameter sliders. Unsuccessfully they remarked that "at the latest now I would need to understand how the parameters relate to the result, I am clueless and it is annoying now." Person 3 refused to use the technical interface completely saying "I can not even do it with that one", but was successful with the small multiples interface, commenting "Ah here (on the right side) it looks evenly spread, I can see that quickly in this interface".

1C

Task 1c relates to the *subject areas*-task of the design requirements and gave the participants some difficulties. However, from their statements it seems that this difficulty was more due to the nature of the single visualizations - their interactions for this tasks still revealed many hints about the usage of the interface. Person 1, the only one not refusing to use the technical interface at this stage, tried to find a relationship between the parameters and the result, e.g., "I think high learning rate will maybe help" but finally gave up saying "The only approach I could think of now, is to try every parameter combination, but that would be ridiculous".

After inspection of the grid interface every participant voiced the sentiment that the visualizations would actually be too similar in this tasks criterion (implying the indirect use of the *clustering-*,*cluster sizes-*, and *patterns*-tasks). For example, person 1 commented "When interacting I notice, that they are all very similar in this regard and it is probably just a matter of personal taste". And person 3 said "I have a hard time selecting, all visualizations are kind of mediocre. If there was at least like 3 that are a bit better." Person 3 turned to the axes explanations for help but appeared to not understand them. Overall the participants understood the task and with the small multiples visualization successfully realized that no optimal result exists for it.

2A

Task 2a was initially somewhat confusing to the participants. They were more interested to see how *their* research topic (paleontology and geology) looked in the main *WISSEN* view of the prototype and at first filtered out the other subject areas (corresponding to the *details-* and *subject areas*-tasks). To further their understanding of the scenario, I told to them to select a visualization that an MfN researcher would, e.g., post to social media. In the end, their approaches to the selection were quite similar (A wider spread visualization with mixed colors), but person 2 and 3 seemed to lack interest for this. Person 1 summarized their selection with: "I would select one that looks nice so maybe it should not have too much overlap and I want all red dots to be on one side." (*clustering-*,*cluster sizes-*, *patterns*-tasks) Person 2 first refused to use a selection interface and said: "I do not not care about the ordering for this task, it would be more interesting to put it into perspective, e.g., select a collection and show how many different projects relate to it. I would use the [IKON] interface itself and its 3 views to show how interdisciplinary we are.". After I asked them, which one they would select anyway: "If I have to select one... I want it to be evenly spread so that each one can be easily clicked and maybe that they are spread over the space so when I use one of the links like collections the whole circle is filled with links." (*patterns-*,*subject areas*-tasks)
Person 3 appeared unsatisfied with their selection and was critical of the entire interaction: "I do not have any criteria to say, which one is better or worse, because i do not understand them. I am not a fan of 2D representations, they give you a wrong image, I do not really care about the ordering, it does not make sense for this task."

2B

Task 2b corresponds to the *details-* and *subject areas*-tasks of the design requirements and was solved by every participant with the help of filters. Each person initially filtered all areas except their own area out and just looked at the projects for this. As there was no perfect grouping available they used different strategies for the selection. Person 1 searched for an ordering were the titles of projects from other research areas that were close to paleontology/geology projects seemed to fit thematically. They then used the small

multiple view to see, if this spatial connection prevailed in other plots. After person 2 had filtered out the other research areas, they were invested with the result and said that as an MfN researcher they would now be very proud of the amount of projects in their field. Only when I reminded them to select an ordering they selected a visualization where "my projects are closely clustered and there are a few outliers to show that we are not closed minded and work with other people as well". Person 3 was undecided between a visualization from the left side were "most are close but some are really far away" or one from the right were "they are all in one corner but not as close to each other". They also articulated that they were overwhelmed with the large selection to chose from: "Having more choices makes it even more difficult to select one and I can not see the colors when the points are too small". This problem might have been due to the smaller monitor, as the other two participants did not voice similar complaints.

4.5 Feedback Questionnaire Results

After solving the tasks and exploring the interface I asked the participants a few questions to get a more differentiated image of their use of the interface and personal opinions:

1. **What do you think about the grid interface compared to the technical interface, was it helpful for the tasks? What differences did you find in using the grid interface vs. the technical interface?**

   The overall sentiment is hugely in favor of the small multiples grid interface I created over the technical interface. However, none of the participants seemed to see a great demand for this kind of interaction for the MfN-researchers.

   Person 1 found the small multiples interface more useful because they could "detect patterns that are invariable over the space of results" (*patterns*-task) and technical parameter selection less useful "because parameters do not seem to have a meaningful effect". They said this, although they were the only one able to solve tasks 1a and 1b easily with the technical interface, as they were able to use at least the perplexity parameter.

   Person 2 thought "the grid interface was super useful and pleasant, it was also really fun to use, like, with a more exploratory approach. The other one jumped around too much, it is annoying without context knowledge.". They did however also argue that "the MfN researchers are probably not so interested in this ordering thing, when they actually want to do something with the interface." From their interaction with the prototype and their mimics it seemed, that although they liked playing around with the grid, they did not really estimate it useful for the more subjective tasks 2a and 2b.

   Person 3 had similar sentiments to person 1 and 2 and commended the possibility for "visual comparison" and was excited about the diversity

in the grid view, but also said "I do not think it is great for decision making. There are too many pictures to select from without knowing criteria". They personally disliked "spatial visualizations because they do not explain why points are, e.g., close or far away from each other." They had also said in the training round: "Do I even want to change the ordering? Okay now I can select one that I find *pretty*? I do not really understand the motivation."

2. **Which grid size did you find the most useful? Was changing the grid size for different tasks helpful?**

   All participants deemed the adjustable grid size as a positive element for selecting "more coarsely ore more fine" (person1). Person 2 and 3 found size 10 to be too cluttered and both set a maximum of 8x8 tiles.

   Person 1 also noted, that "it is annoying that the one you selected can disappear if you change the grid size, it needs a visual pointer". This flaw is something I had suspected before, possible solutions would be animating the size change or inserting visual markers in the place where the selected tile "disappears".

   The remark from **Person 2**: "Until now I did not think about what it really does and just kind of used it intuitively to *learn to see the data*, but how does the sizing really work?" showed that although it is intuitively usable, more speaks in favor of an animation or similar technique to show what is changing.

3. **What do you think about the explanation given about the axes? Did you need to know about the meaning of the axes to solve your tasks? (follow-up: What do you think of the ordering of the tiles?)**

   The description of the axes seemed confusing to the participants, but they also all said they did not need or use the axes for their tasks (person 1: "Axes are not really needed, the explanation of the y-axis is just confusing."). The ordering seemed to be used by all intuitively, although its extraction was not clear to them.

   Person 1 did not like "that most of them are basically the same just rotated". And suggested only showing visualizations in one rotation. Person 3 also criticized this. This problem could be partially solved by the alternative sampling method *stabilized sampling* I described in Section 3.3.4.

   Similar to my assumption, **Person 2** argued: "I think it is like a scientific minimal standard to have axes descriptions, but actually I do not care about them and to be honest, I think not every layperson would notice this ordering thing so much."

   Person 3 did not understand the axes and their ordering but said "I still somewhat noticed the ordering, for example that neighbours are

quite similar and the close or coarse groups thing, which I now learnt is the perplexity, but i thought it was more diagonally."

4. **What else did you like or dislike about the interface?**

   Participant 1 and 2 liked the layout and design of the grid interface (1: "I am not a designer, but I like the looks", 2: "I enjoy the design it looks professional"), while person 3 remarked "The colors are difficult to differentiate for me - As a glasses wearer, I am not a fan of dark mode interfaces". Again, this could be because of the smaller monitor and inferior lighting in the coffee shop. All three found the animated transitions when selecting a new ordering to be enhancing their understanding.

   Person 1 mentioned a few possible improvements: "I am not sure but maybe I would also like to see the labels and uncertainty and so on in the grid view.", "I would like to click on one point, that is then highlighted in all tiles to better compare the individual positions." and "It annoys me, that I can not highlight a sub-category separately.". While the first remark is something my advisor and I decided against, as we found it to be too cluttering, the second and third are valuable for future revision.

   Person 2 had difficulties finding the buttons for selecting orderings (with the *grid* vs. *technical* interface) in the training phase. They commented this here with: "For me, everything is *technical* so the naming of the buttons is not sensible. The left-bottom corner is also the last place I would look. From my research I learnt that it is really important for users to understand the scales or buttons if they want to use an interface effectively." Those issues can be resolved be placing the buttons differently (e.g., upper right corner) and finding more descriptive names.

   Person 3, with observing "Here, perplexity 13 is directly under one with perplexity 3, that makes no sense, that is confusing." found that the ordering on the axes is not linear, which speaks against the decision for designating axes in general. A solution that clearly shows, that the grid is based only on a 2-dimensional sorting of the samples might resolve this.

4.6 Discussion

For most of the initial questions a clear answer emerged. Furthermore, many remarks even about details of the interaction were made that could be applied in a future iteration of this prototype.

1. **Which interface (small multiples or technical) is preferred for the tasks?** All participants much preferred the small multiples view to the technical interface. The refusal of person 2 and 3 to use the technical interface is a clear indication for this. Especially person 1 and 2 also voiced their enjoyment interacting with it in a playful manner. They also praised the "exploratory approach" (person 2) and the possibility to

"detect patterns that are invariable over the space of results" (person 1, (*patterns*-task)), which were the facets I was most eager to enhance with this interface.

2. **Is the small multiples interface a useful enrichment for non-technical experts, how do they interact with it?** In contrast to their mostly positive experience using it, the participants all questioned the usage scenario to an extent, they seemed to disagree that MfN researchers were in demand for such an interface. The solving of the more objective tasks (1a and 1b) was clearly simplified by the interface, while the more subjective tasks were eased but remained difficult. Nonetheless, the comments the participants made here, indicate that they gained a better understanding of the possible result space through the grid interface, e.g., "most of them are basically the same just rotated" (person 1).

3. **What are disadvantages of the small multiples interface?** In comparison to the technical interface no disadvantages were mentioned. However, every participant saw some flaws in the interface, which they commented on as described previously in the feedback questionnaire results. These remarks not only help explain some difficulties they had with the tasks but also are potential improvements that can be made in future revisions.

4. **Is an adjustable grid size helpful?** The participants generally received the adjustable grid size as helpful, the expression participant 2 used, that it is good to "learn to see the data" matches with my observation of how the participants used it.

5. **How should the ordering (with t-SNE and perplexity) be explained and is it noticeable and useful for non-experts?** The axes and tooltips used to explain the ordering were received as unsatisfactory. As mentioned in the feedback, a better strategy for explaining the sorting (possibly omitting a linear metaphor) needs to be found, although the participants also noted that they did not need any explanations to solve the tasks. From my observations I conclude that the failed attempt to understand the more technical explanation confused the participants more than just making intuitive assumptions.

6. **How is knowledge or ignorance of the underlying algorithm influencing the interaction?** Comparing the first person who knew t-SNE to the other two participants revealed that knowledge about the parameters could be bypassed with the interface allowing them to make decisions similar to the ones the first person made by, e.g., using the parameter perplexity. The second and third participant seemed to be less interested in using the interface though, which could be due to their stated lack of understanding the reasoning behind embedding projects based on machine learning results.

5 Conclusion and Outlook

This thesis investigated how a small multiples visualization of dimensionality reduction results would help non-technical experts in the specific use-case IKON to select outcomes supporting their tasks.
I conducted semi-structured interviews to define the requirements and tasks for the interface, iteratively refined a LoFi- and then a HiFi-prototype and visually and computationally explored 10 different metrics for their applicability in ordering a small multiples visualization of DR results.

A small exploratory study provided evidence that the resulting small multiples interface (which I called 'grid interface' in the user study) was more useful to the participants than the currently used slider interface ('technical interface'). The findings suggest that the interface facilitates an exploratory and playful approach and invites the users to search and find patterns in the visualizations. As mentioned in the discussion of the pilot study, some design flaws and missing or mismatching interactions impaired the usability slightly, but overall the participants reported enjoyment and a positive experience. All participants of the study voiced general doubt about whether the application would be useful for the target group of the use case. I speculate that only individuals who already enjoy exploring visualizations will have the motivation to use it. Nonetheless this application is also still useful for other groups than the described target group. For example, researchers, who previously were in charge of selecting appropriate DR visualizations with technical parameters and are also non-technical experts, benefit from this interface.

In the exploration and experiments with ordering metrics the t-SNE measure combined with perplexity outdid the graph-theoretic measures and appeared to most effectively help with the selection of a result, which supports the tasks defined in the beginning. Through the ordering with perplexity, the *clustering*-, *cluster sizes*- and *details*-tasks became easier, while the t-SNE ordering supported especially the *patterns*-task but also somewhat the other tasks. The *subject areas*-task was only indirectly supported through the ordering, although this is presumably also due to the nature of the result space, which does not include visualizations with strongly separated subject areas. I assume this, because the participants of the study appeared to understand the ordering and the task well, but no ordering existed, which performed better than others in this regard. After all, what a small multiples visualization does, is not manipulating the data to fit the users expectations but present an equally valid set of visualizations (which sometimes just are not satisfying a specific need). All in all the participants of the user study understood the ordering and used it intuitively. When asked about it, some had difficulties describing and conceptualising the ordering. This could however also speak for it, as an ordering that is close to human perception might be more difficult to

describe for them.

The sampling method chosen here is simple as this was sufficient for the desired interface. The 'periodic' sampling guarantees for including very low and high values of perplexity. The user study showed that the 'periodic' sampling method led to some confusion, as tiles that are parallel on one axis do not necessarily have the same values. Another down-side mentioned was that many plots were almost identical and only their rotation was different. The 'stability' sampling might perform better if the effect of excluding similar but rotated plots is desirable as observed here. A more sophisticated maybe non-linear sampling method could also provide more even and less repetitive results.

Due to the framing of this project as a bachelor thesis some considerations would need to be tackled in the future:

- Because of the limited nature of the pilot study, findings can not be generalized. A follow-up study with more participants and more diverse application scenarios would be necessary. The agreement of the selected measure with human similarity perception also needs to be tested in a study, similar to the one described in [PKF+16].
- Measures (perplexity and t-SNE measure) are only tested on data sets with similar element counts as the IKON set to enable comparison, it does not necessarily scale to sets with more elements. A dimensionality reduction of the point coordinates with, e.g., PCA as a step before the high-level t-SNE might ensure good results for larger point counts. A bigger range for the t-SNE parameters in the background would produce a more diverse set of results and increase the chance that a suitable parameter setting can be found in more general cases.
- In comparison to other machine learning algorithms the t-SNE computation may not be the most optimal. Approaches, such as [DW14], where human-perception data was taken into account when creating a neuronal network could possibly reveal results closer to this data.
- The developed interface is bound to the IKON scenario and prototype. Data delivery to the front-end is only mocked with pre-computed results for simplicity and computation time concerns. It would be desirable to generalize the application for better testing and to include more use-cases. Most of the usability flaws found in the user study have obvious solutions that should also be implemented in a future revision of the interface.

The outcomes of this thesis imply that the result-based small multiple technique is preferable to classical parameter tuning techniques, such as numerical sliders, for the selection of technical parameters for the target group of non-technical experts. A similarity ordering can not only help them to search for their desired visualization more effectively but also inspire them to find stable

patterns over the result space and define their own criteria for the selection. This knowledge could be used in future work, where non-technical experts are confronted with the non-trivial task of selecting appropriate DR-parameters and possibly also other machine learning parameters.

References

[AASB19] Mostafa Abbas, Michael Aupetit, Michael Sedlmair, and Halima Bensmail. ClustMe : A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Computer Graphics Forum*, 38:225–236, 06 2019.

[ASA+19] Michaël Aupetit, Michael Sedlmair, Mostafa M. Abbas, Abdelkader Baggag, and Halima Bensmail. Toward perception-based evaluation of clustering techniques for visual analytics. In *2019 IEEE Visualization Conference (VIS)*, pages 141–145, 2019.

[BBL12] Ilya Boyandin, Enrico Bertini, and Denis Lalanne. A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. In *Computer Graphics Forum*, volume 31, 6 2012.

[BCC57] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

[Ben] Jesse J. Benjamin. Project abstract IKON. `https://www.mi.fu-berlin.de/en/inf/groups/hcc/research/projects/ikon/index.html`. Accessed: 2020-01-30.

[BKH+20] Jesse J. Benjamin, Christoph Kinkeldey, Eva-Maria Herbst, Tim Korjakow, and Claudia Müller-Birn. Participatory interpretability: Developing a design research method for machine learning interpretability with non-technical stakeholders. *Manuscript in preparation*, 2020.

[BMBK19] Jesse J. Benjamin, Claudia Müller-Birn, and Christoph Kinkeldey. Understanding knowledge transfer activities at a research institution through semi-structured interviews. *Technical Report*, 08 2019.

[Bos20] Mike Bostock. D3.js. `https://d3js.org/`, 2020.

[Cam19] Anja Susanne Camara. Animierte transitionen für die visualisierung von veränderungen in clustering-ergebnissen. `https://www.mi.fu-berlin.de/en/inf/groups/hcc/theses/finished/2019-Theses/animierte_transitionen_fuer_die_visualisierung_von_veraenderungen_in_clustering-ergebnissen/index.html`, 2019.

[CHAS18] Rene Cutura, Stefan Holzer, Michaël Aupetit, and Michael Sedlmair. Viscoder: A tool for visually comparing dimensionality reduction algorithms. In *ESANN*, 2018.

[CKB09] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.*, 41(1), January 2009.

[Com18] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[Cut17] Rene Cutura. VisCoDeR. `https://renecutura.eu/viscoder/`, 2017. [Online; accessed 2-July-2020].

[(DF18] Deutsche Forschungsgemeinschaft (DFG). GEPRIS. `https://gepris.dfg.de/gepris/OCTOPUS`, 2018. [Online; accessed 3-August-2020].

[DW14] Tuan Nhon Dang and Leland Wilkinson. ScagExplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium*, pages 73–80, 2014.

[EHKN15] Peter Eades, Seok-Hee Hong, Karsten Klein, and An Nguyen. Shape-based quality metrics for large graph visualization. In *Graph Drawing and Network Visualization*, pages 502–514, Cham, 2015. Springer International Publishing.

[Fou] Interaction Design Foundation. Definition user centered design. `https://www.interaction-design.org/literature/topics/user-centered-design`. Accessed: 2020-04-30.

[hd] Jack Aurizio (https://math.stackexchange.com/users/44121/jack daurizio). Average distance between two random points in a square. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/1254154 (version: 2020-06-12).

[HR03] Geoffrey E. Hinton and Sam T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 857–864. MIT Press, 2003.

[Hub15] Christoph Hubig. *Die Kunst des Möglichen I.* transcript Verlag, Bielefeld, Germany, 2015.

[KKB19] Christoph Kinkeldey, Tim Korjakow, and Jesse J. Benjamin. Towards Supporting Interpretability of Clustering Results with Uncertainty Visualization. In *EuroVis Workshop on Trustworthy Visualization (TrustVis)*. The Eurographics Association, 2019.

[KM20] Oh-Hyun Kwon and Kwan-Liu Ma. A deep generative model for graph layout. *IEEE Transactions on Visualization and Computer Graphics*, 26, 01 2020.

[KNJ+20] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists. In *CHI 2020*, April 2020.

[Kor19] Tim Korjakow. Developing interpretability techniques for unsupervised topic modeling. *Bachelor Thesis*, 08 2019.

[LvdMdS12] Joshua Lewis, Laurens van der Maaten, and Virginia de Sa. A behavioral investigation of dimensionality reduction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 01 2012.

[MfNB18] IKON-Team Museum für Naturkunde Berlin. VIA Wiki. `https://via.museumfuernaturkunde.berlin/wiki`, 2018. [Online; accessed 3-August-2020].

[MW] Human-Computer Interaction Resource Network Matthew Wickline. Color blindness simulator. `https://www.color-blindness.com/coblis-color-blindness-simulator/`. Accessed: 2020-07-31.

[PKF+16] Anshul Vikram Pandey, Josua Krause, Cristian Felix, Jeremy Boy, and Enrico Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3659–3669, New York, NY, USA, 2016. Association for Computing Machinery.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Rea20] React.js. A JavaScript library for building user interfaces. `https://reactjs.org/`, 2020.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[SBIM12] Michael Sedlmair, Matthew Brehmer, Sue Ingram, and Tamara Munzner. Dimensionality reduction in the wild : Gaps and guidance. 2012.

[SBS94] Maarten Someren, Yvonne Barnard, and Jacobijn Sandberg. *The Think Aloud Method - A Practical Guide to Modelling Cognitive-Processes*. 01 1994.

[SHW17] Aaron Springer, Victoria Hollis, and Steve Whittaker. Dice in the black box: User experiences with an inscrutable algorithm. *AAAI Spring Symposium Series*, 2017.

[STMT12] Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012.

[SVU89] J.P. Snyder, P.M. Voxland, and Geological Survey (U.S.). *An Album of Map Projections*. Number no. 1453 in An Album of Map Projections. U.S. Government Printing Office, 1989.

[SZS+16] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen North, and Daniel Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23, 01 2016.

[TBB+10] Andrada Tatu, Peter Bak, Enrico Bertini, Daniel Keim, and Jörn Schneidewind. Visual quality metrics and human perception: An initial study on 2d projections of large multidimensional data. *Working Conference on Advanced Visual Interfaces : AVI 10 ; Roma, Italy*, 01 2010.

[TBJB03] Yee Whye Teh, Matthew J. Beal, Michael I. Jordan, and David M. Blei. Hierarchical dirichlet processes, 2003.

[Tea] Diagrams Net Team. diagram desktop software diagrams.net formerly known as draw.io. `https://www.diagrams.net/`. Accessed: 2020-07-20.

[TT85] John W. Tukey and Paul A. Tukey. *Computer graphics and exploratory data analysis: An introduction*. 1985.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

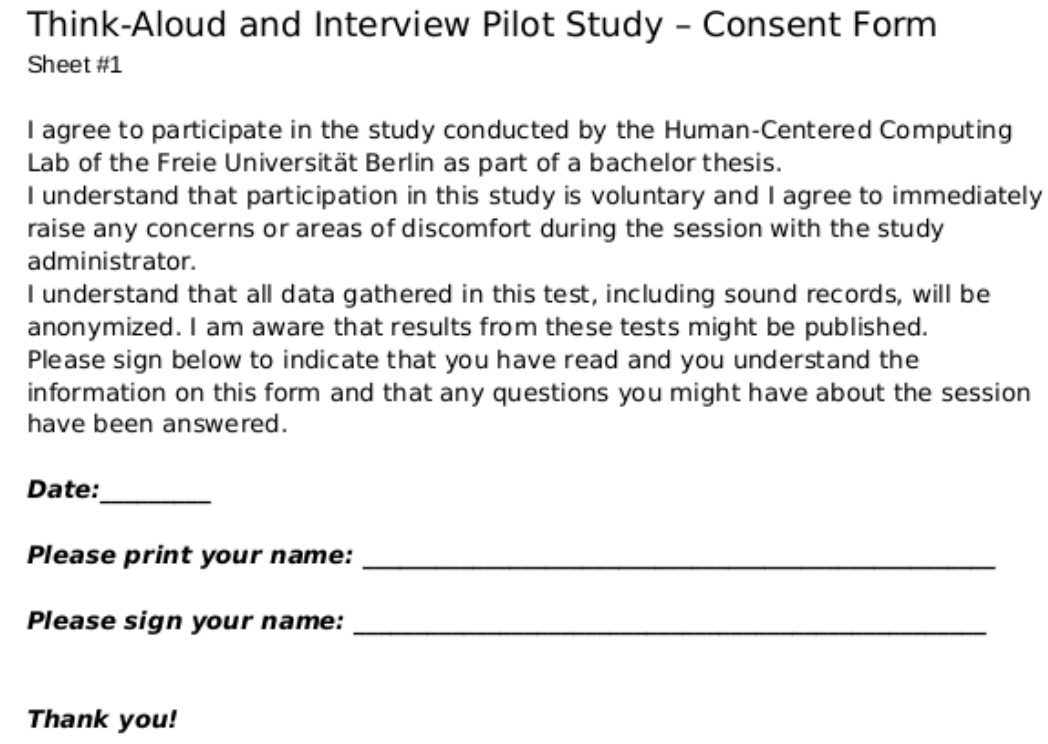
[WAG05] Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164, 2005.

[WJEO] Aaron Winter, Lilli Joppien, Max Eissler, and Jonas Oppenlaender. Front-end github-repository of project IKON. `https://github.com/FUB-HCC/IKON-projektor`. Accessed: 2020-04-23.

[WVJ16] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1, 10 2016.

6 Appendix

6.1 Pilot Study Forms



Think-Aloud and Interview Pilot Study – Consent Form
Sheet #1

I agree to participate in the study conducted by the Human-Centered Computing Lab of the Freie Universität Berlin as part of a bachelor thesis.
I understand that participation in this study is voluntary and I agree to immediately raise any concerns or areas of discomfort during the session with the study administrator.
I understand that all data gathered in this test, including sound records, will be anonymized. I am aware that results from these tests might be published.
Please sign below to indicate that you have read and you understand the information on this form and that any questions you might have about the session have been answered.

Date:_____

Please print your name: _____

Please sign your name: _____

Thank you!

Figure 6.1: This consent form was signed by every participant before the study. It is partly based on Jesse J. Benjamins formulations.

Figure 6.2: Questionnaire for Feedback and Demographic Data after the Test

Questionnaire

FEEDBACK:

After exploring the interface:
what do you think about the grid ordering of the results compared to the technical interface?

Was the grid interface helpful for the tasks? What differences did you find in using the grid interface vs. the technical interface?

Which grid size did you find the most useful? Was changing the grid size for different tasks helpful?

What do you think about the explanation given about the axes? Did you need to know about the meaning of the axes to solve your tasks?

What else did you like or dislike about the interface?

SELF-REPORTING

Do you have any expertise in dimensionality reduction (DR)? If so, how much?

None ◯ Used it before ◯ I know some DR algorithms ◯ I know t-SNE ◯

How often do you use visualizations in your work?

Never ◯ Once a Year ◯ Once a Month ◯ Once a Week ◯ Daily ◯

What is your general field of work?

What is your degree?

How old are you?

6.2 Benchmark Data Sets Graphics

Figure 6.3: A set of t-SNE results for each data set ordered by their perplexity

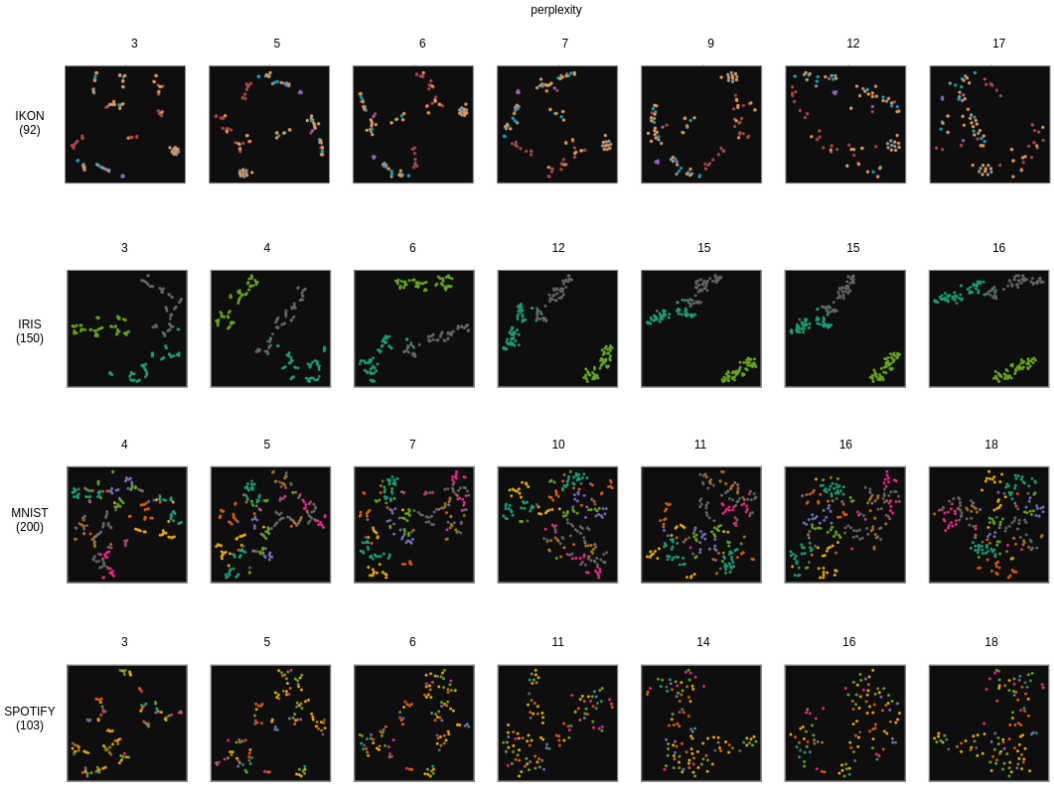


Figure 6.4: A set of t-SNE results for each data set ordered by their learning rate

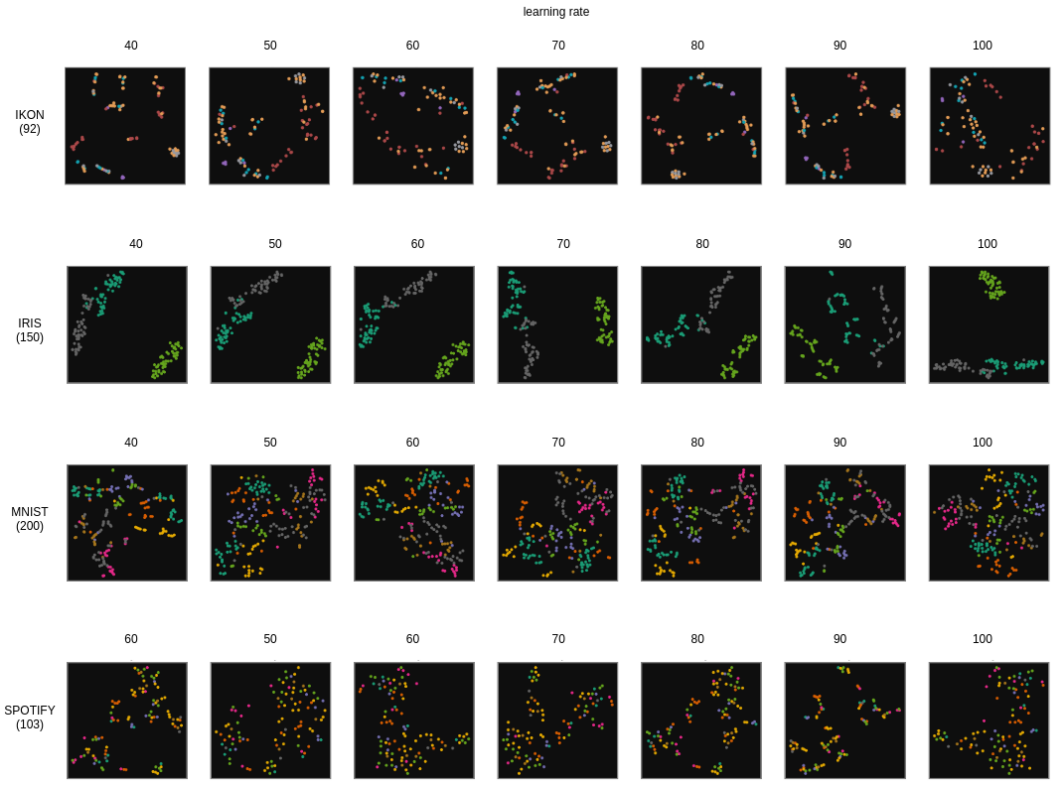


Figure 6.5: A set of t-SNE results for each data set ordered by their 'silhouette' score

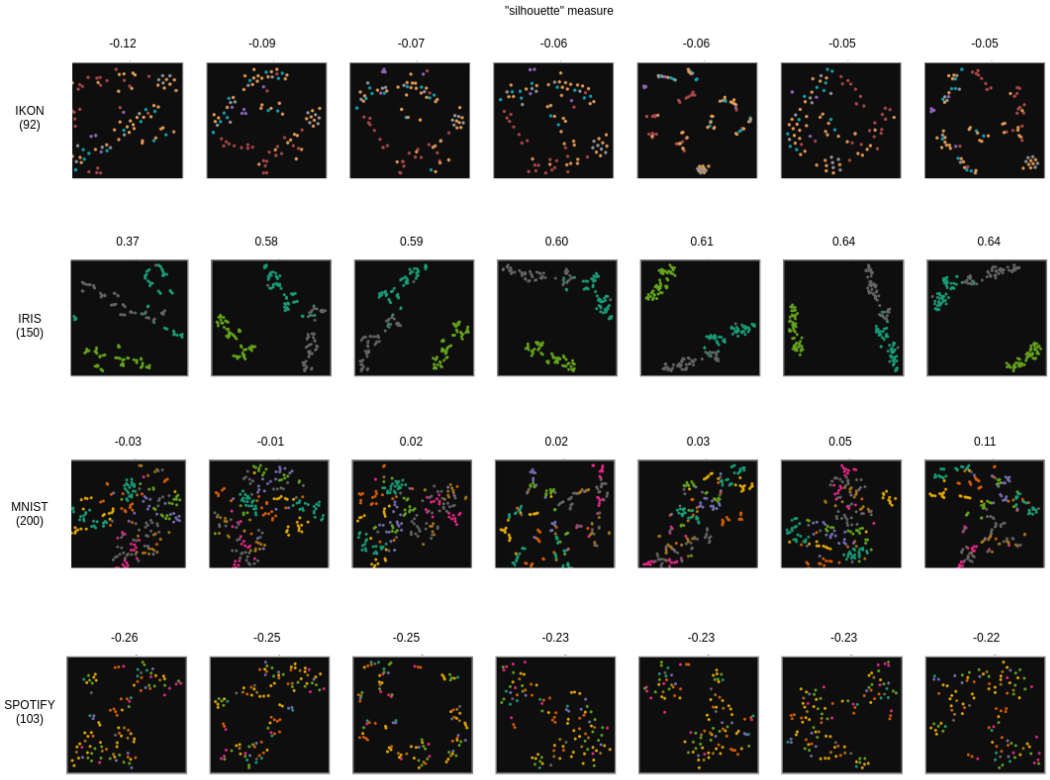


Figure 6.6: A set of t-SNE results for each data set ordered by the Scagnostics 'outlying' value

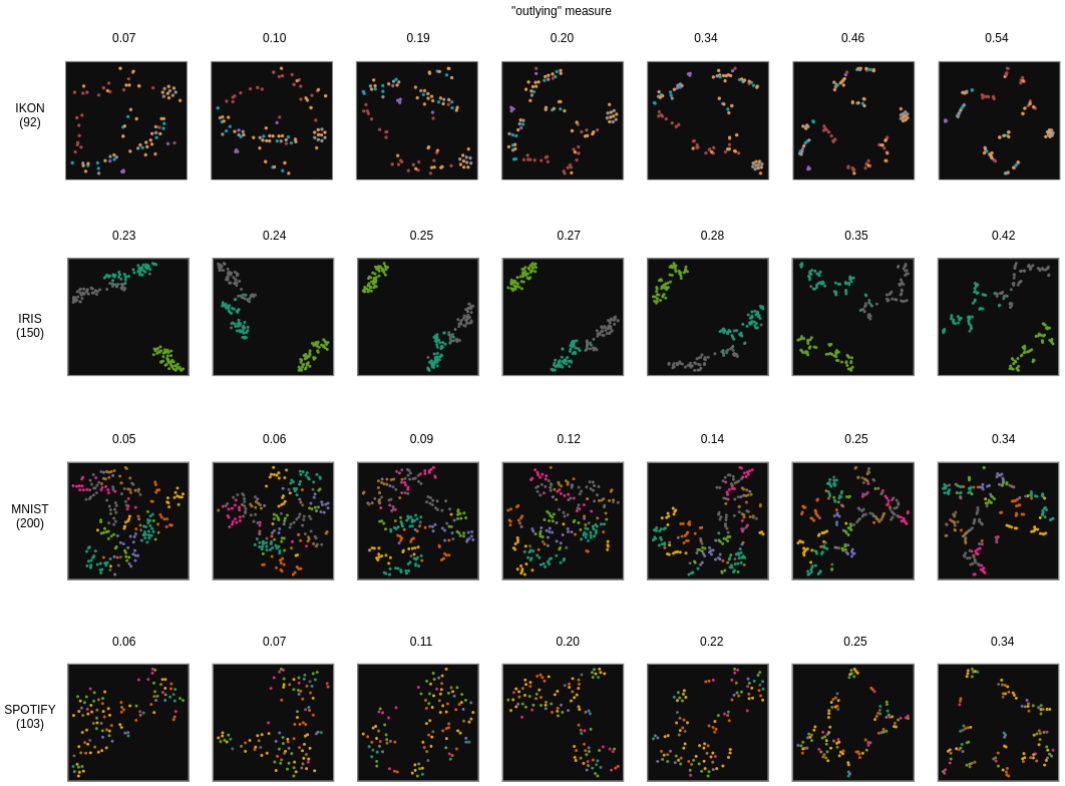


Figure 6.7: A set of t-SNE results for each data set ordered by the Stability value (the point position of one random point)

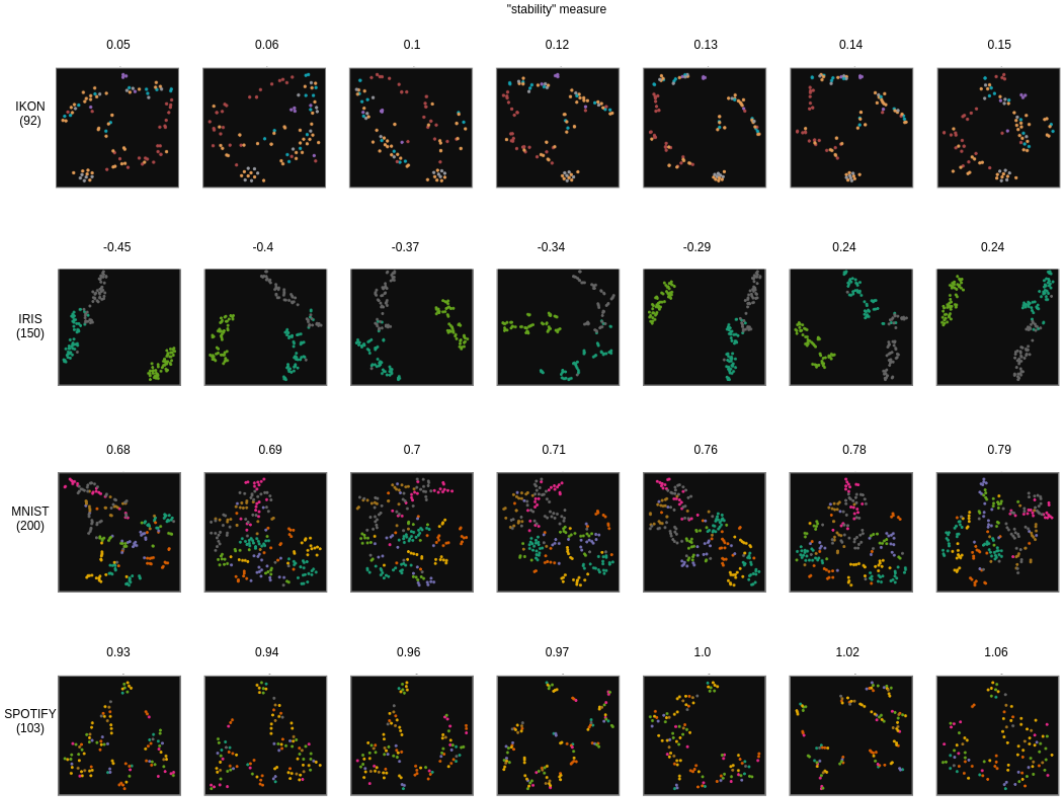


Figure 6.8: A set of t-SNE results for each data set ordered by the Scagnostics value for Spearman's rank correlation coefficient

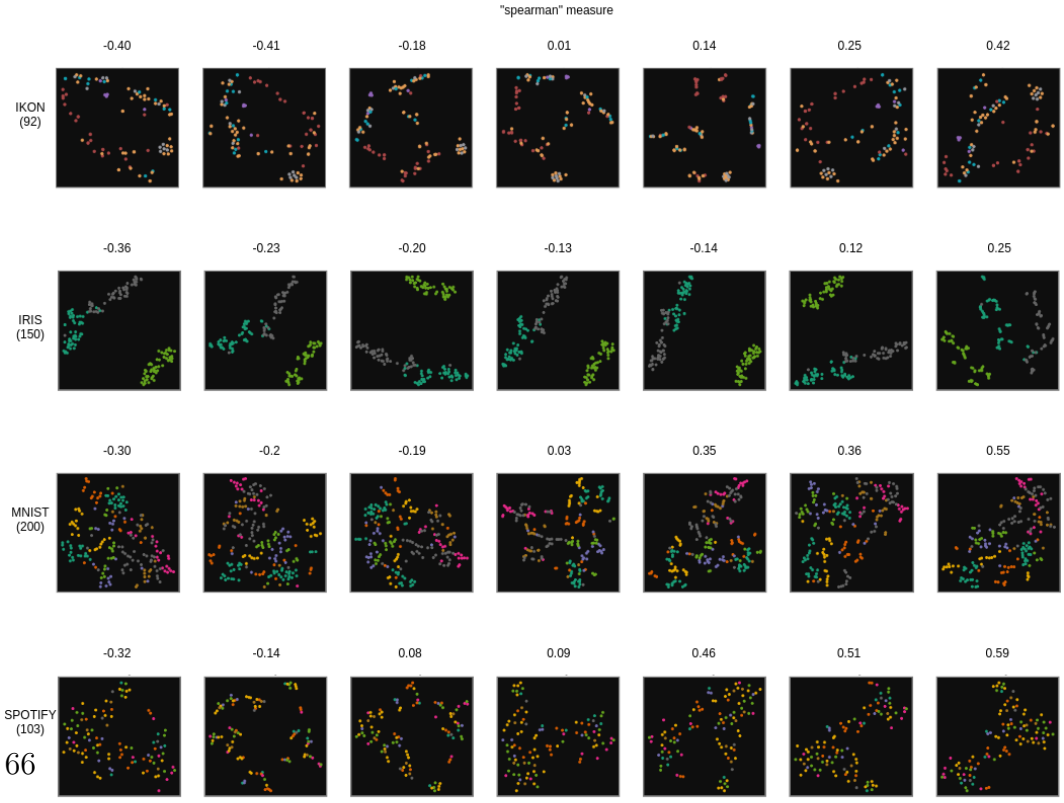


Figure 6.9: A set of t-SNE results for each data set with a similar value in their reduction to 1D through t-SNE

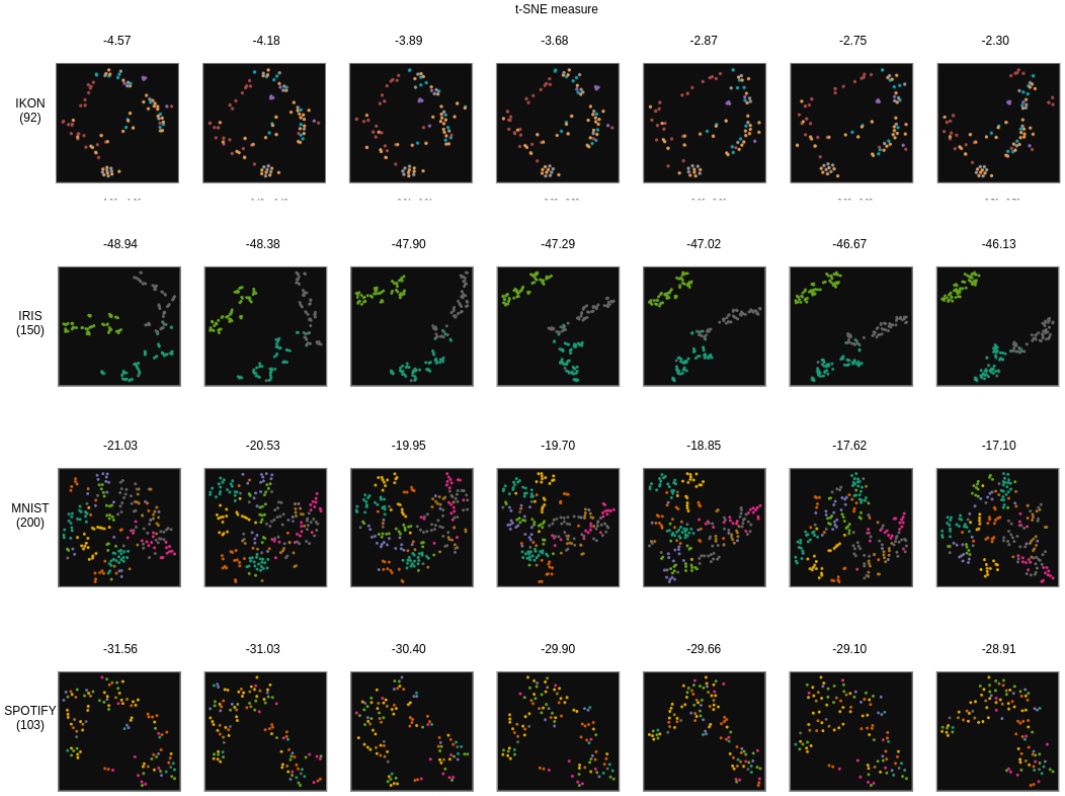
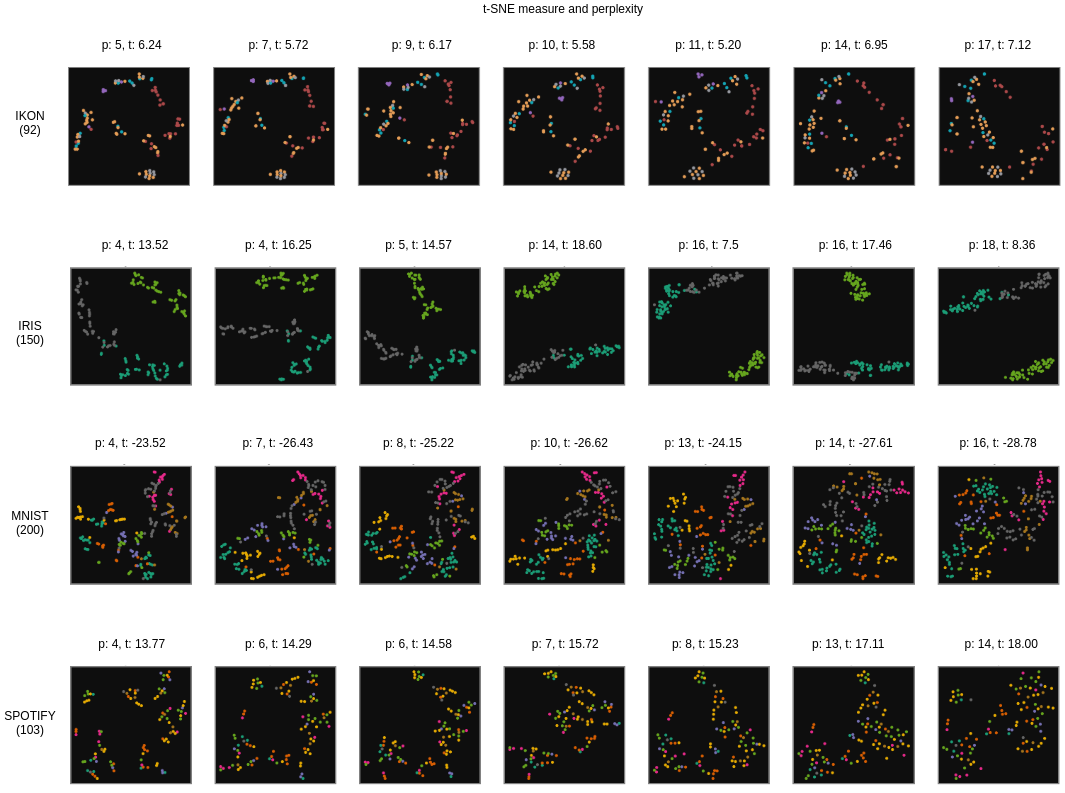


Figure 6.10: A set of t-SNE results for each data set with a similar value in their reduction to 1D through t-SNE ordered by their perplexity



6.3 Euclidean Distance To Neighbours

Figure 6.11: T-SNE Measure

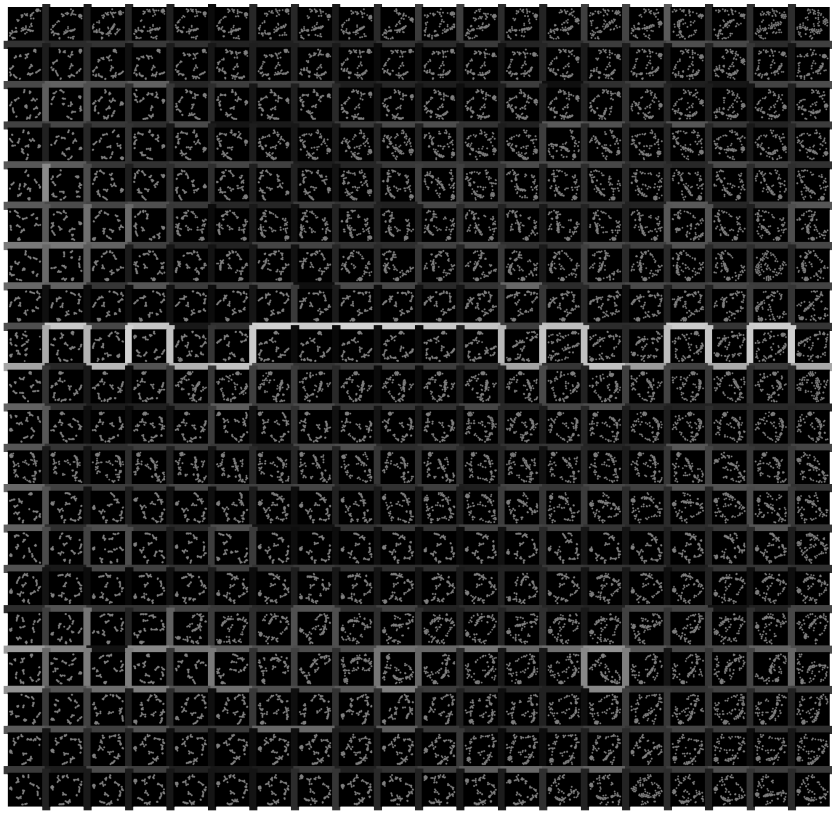


Figure 6.12: Smallest Distances Measure

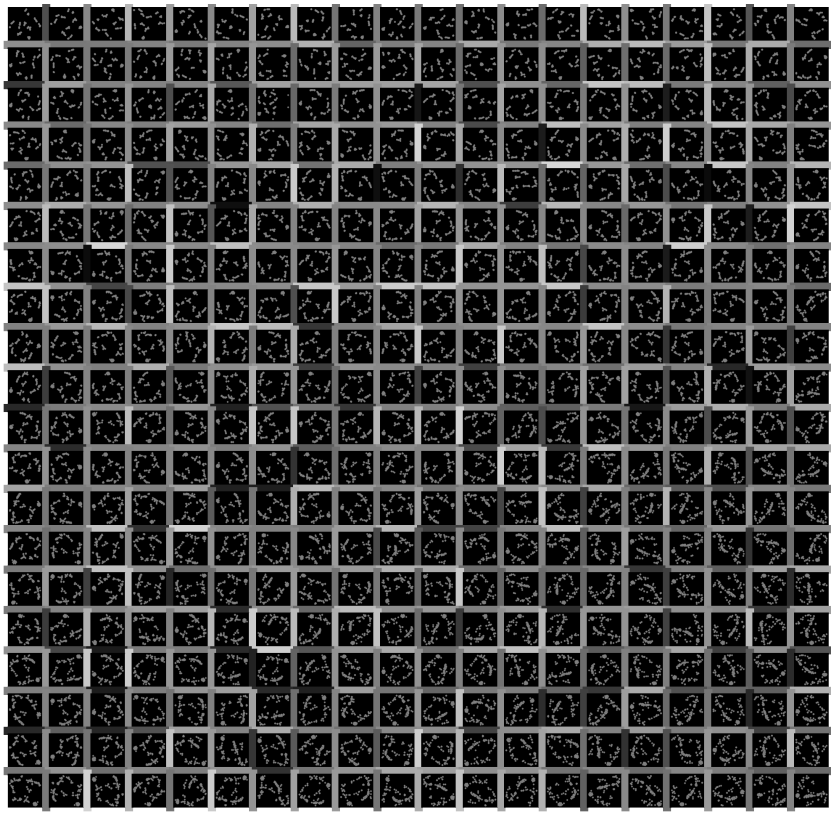


Figure 6.13: Mean Jaccard Measure

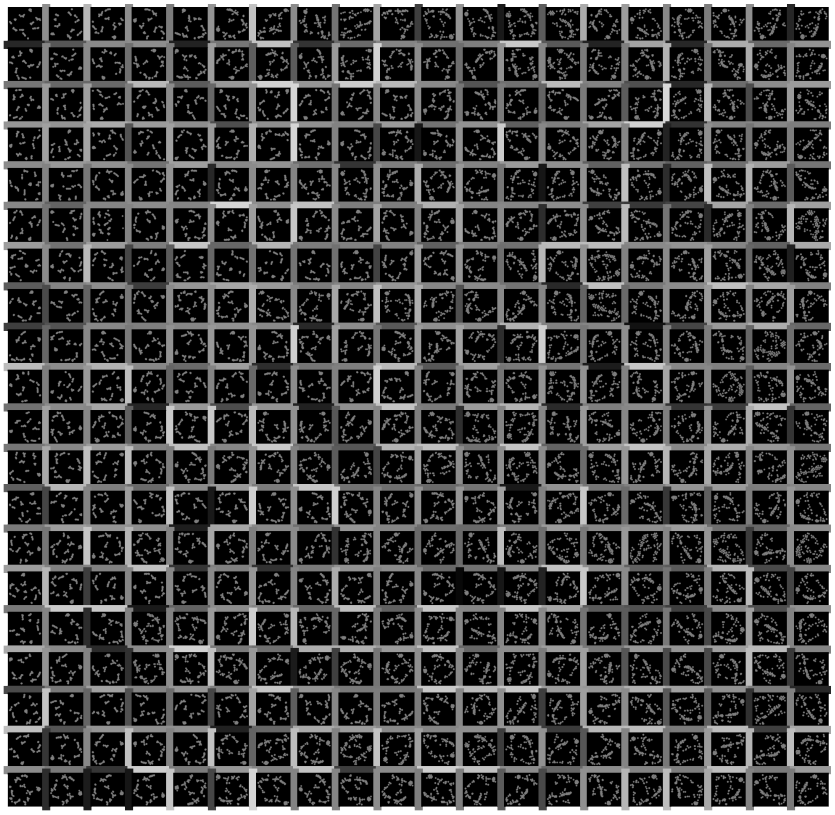


Figure 6.14: Stability Measure

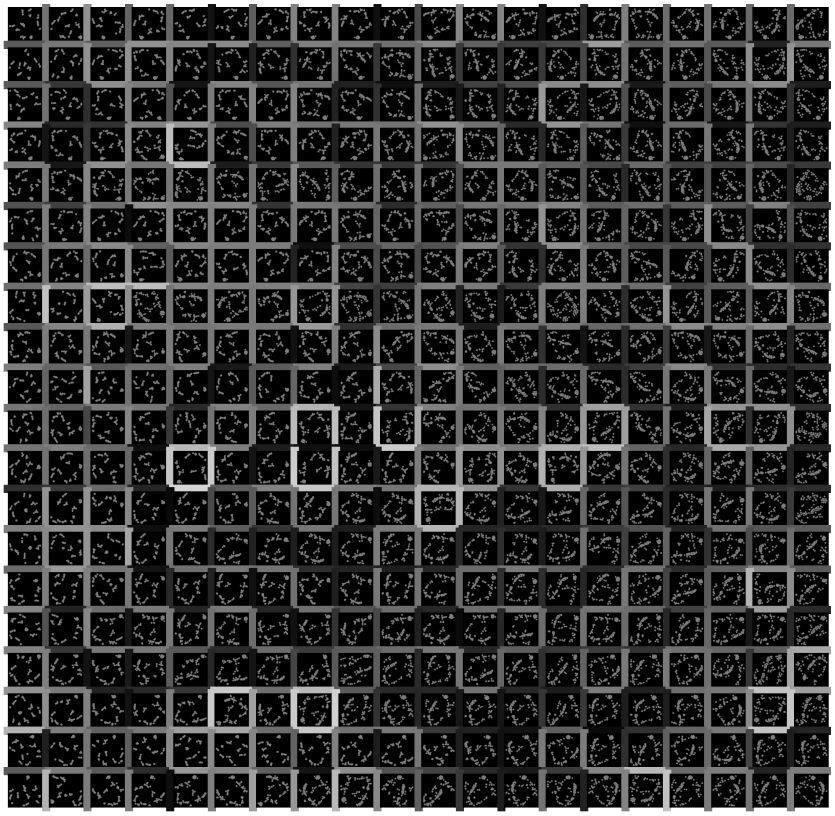


Figure 6.15: Silhouette Measure

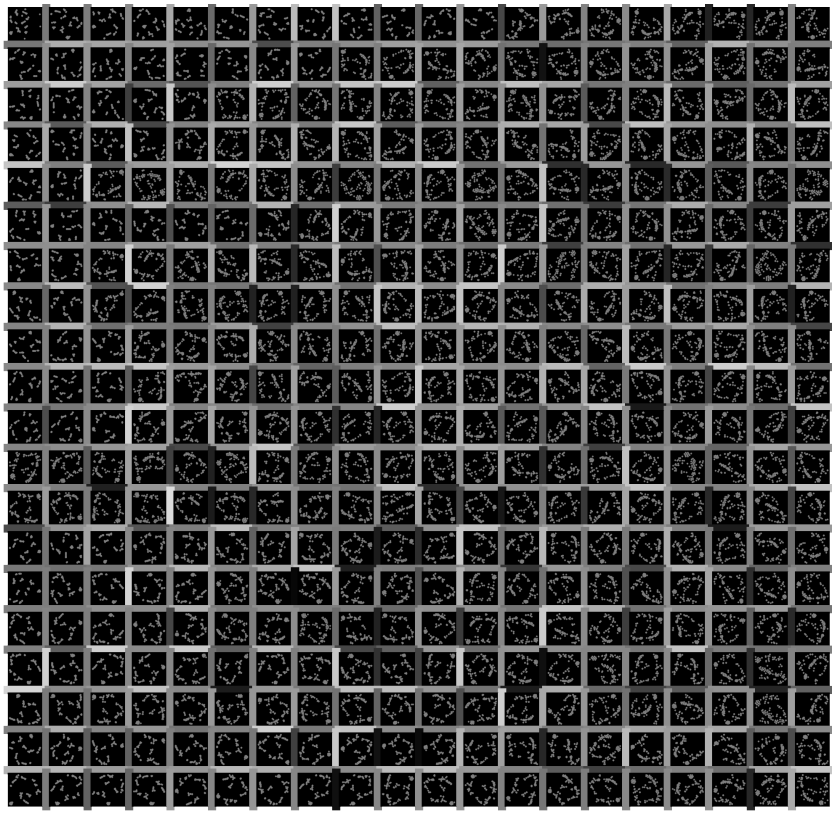


Figure 6.16: Learning Rate Measure

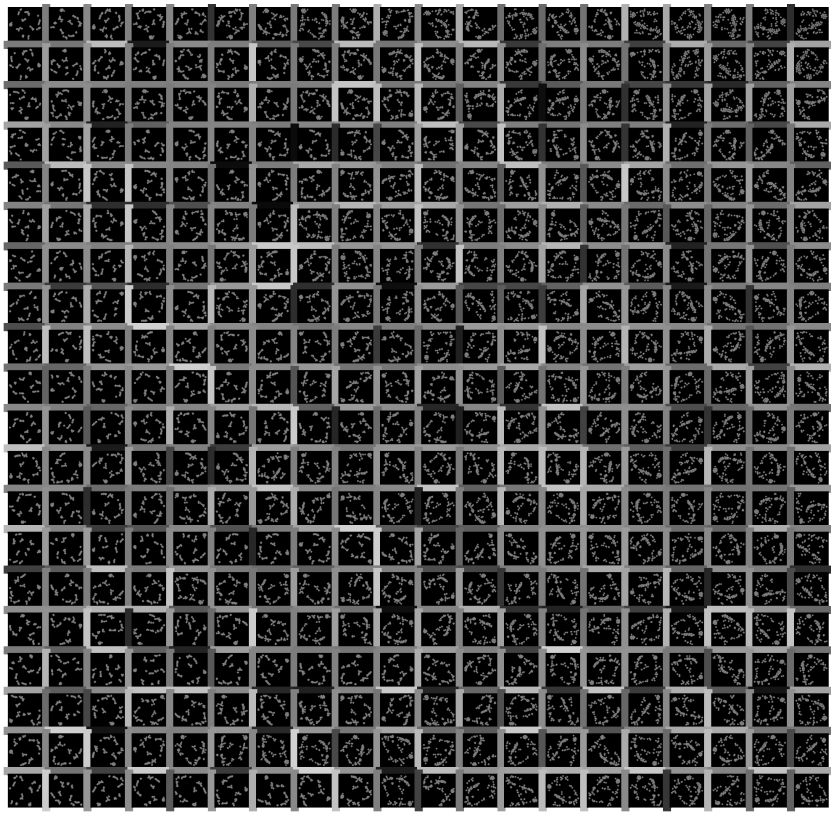


Figure 6.17: Clumpiness Measure (Scagnostics)

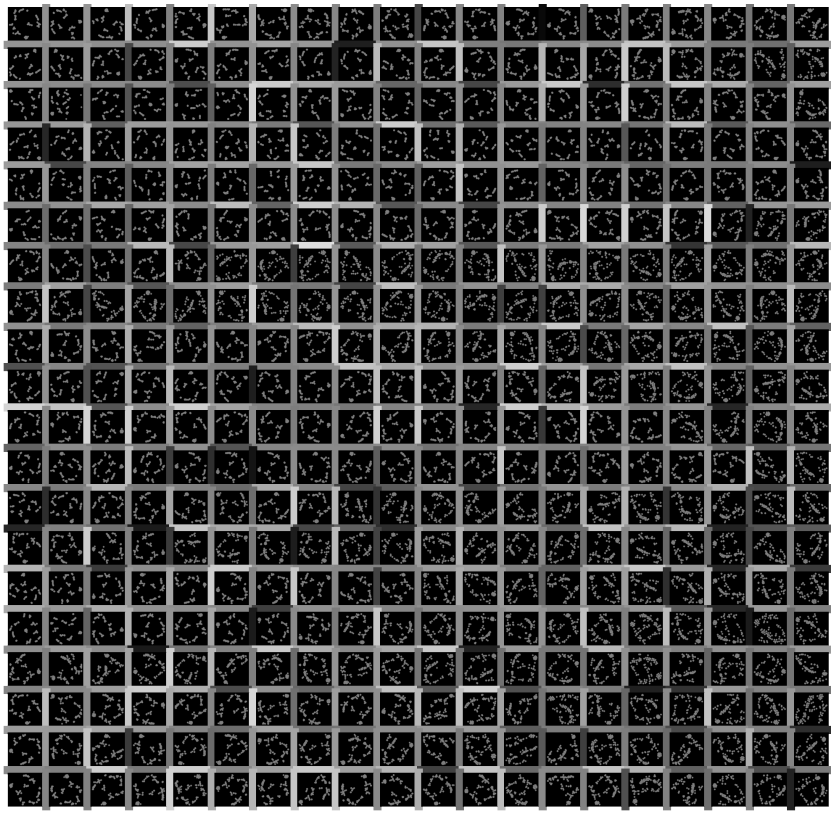


Figure 6.18: Sparsity Measure (Scagnostics)

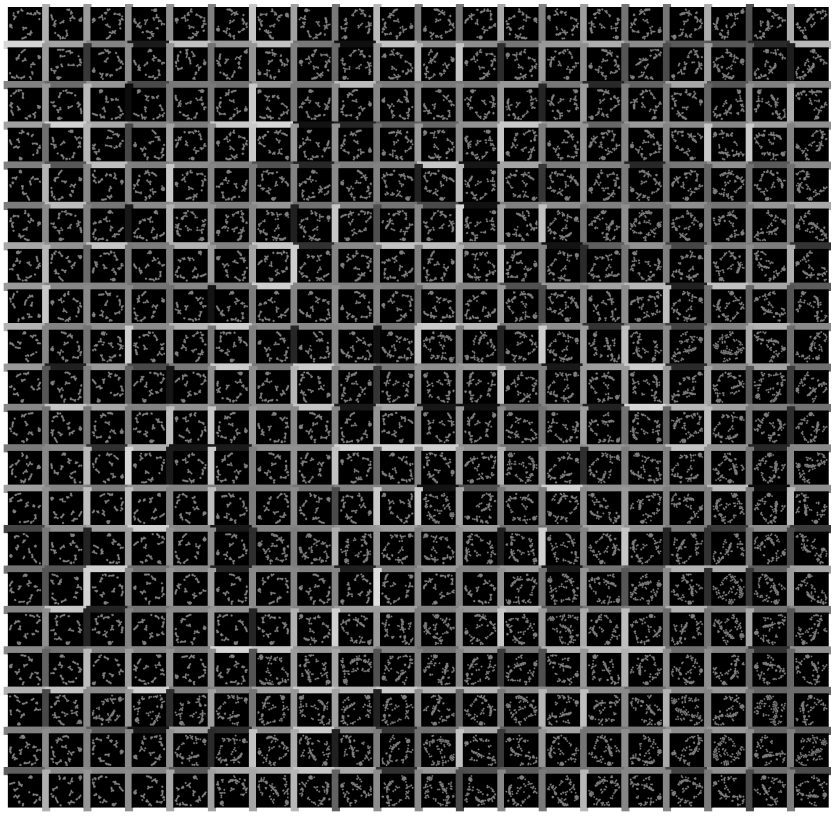


Figure 6.19: $r_{spearman}$ Measure (Scagnostics)

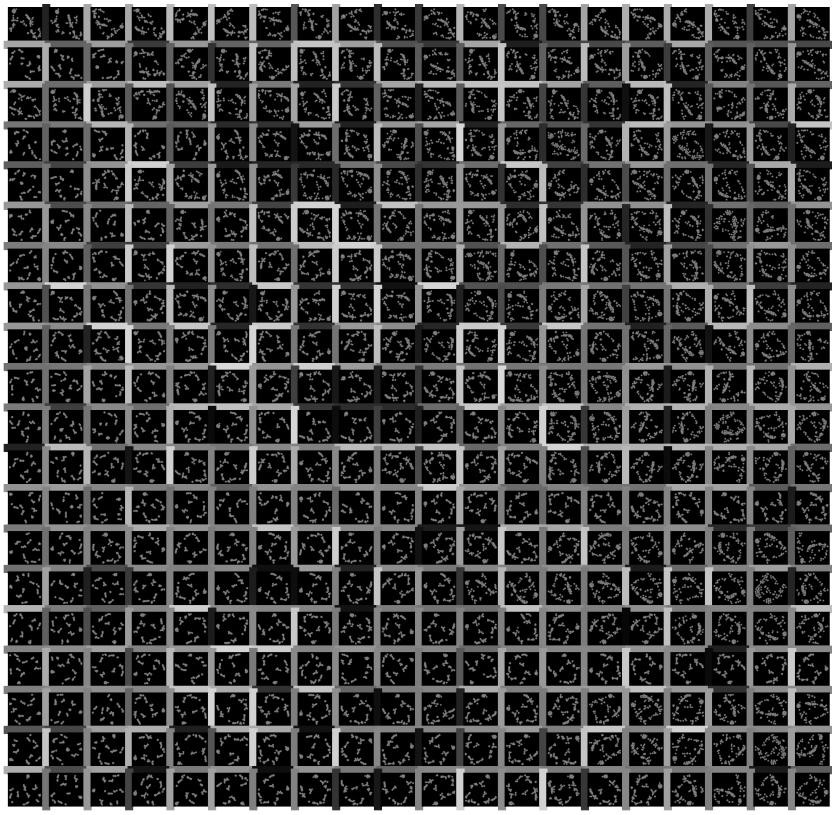


Figure 6.20: Outliers Measure (Scagnostics)