Title	DataShot: Automatic Generation of Fact Sheets from Tabular Data
Authors	Wang, Yun; Sun, Zhida; Zhang, Haidong; Cui, Weiwei; Xu, Ke; Ma, Xiaojuan; Zhang, Dongmei
Source	IEEE Transactions on Visualization and Computer Graphics, v. 26, (1), January 2020, article number 8805442, p. 895-905
Version	Accepted Version
DOI	10.1109/TVCG.2019.2934398
Publisher	IEEE
Copyright	© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

DataShot: Automatic Generation of Fact Sheets from Tabular Data

Yun Wang*, Zhida Sun*, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang

Abstract— Fact sheets with vivid graphical design and intriguing statistical insights are prevalent for presenting raw data. They help audiences understand data-related facts effectively and make a deep impression. However, designing a fact sheet requires both data and design expertise and is a laborious and time-consuming process. One needs to not only understand the data in depth but also produce intricate graphical representations. To assist in the design process, we present DataShot which, to the best of our knowledge, is the first automated system that creates fact sheets automatically from tabular data. First, we conduct a qualitative analysis of 245 infographic examples to explore general infographic design space at both the sheet and element levels. We identify common infographic structures, sheet layouts, fact types, and visualization styles during the study. Based on these findings, we propose a fact sheet generation pipeline, consisting of fact extraction, fact composition, and presentation synthesis, for the auto-generation workflow. To validate our system, we present use cases with three real-world datasets. We conduct an in-lab user study to understand the usage of our system. Our evaluation results show that DataShot can efficiently generate satisfactory fact sheets to support further customization and data presentation.

Index Terms—Fact sheet, infographic, visualization, and automated design

1 Introduction

Fact sheets are a presentation of data, knowledge, and information in a format that emphasizes key points from a variety of perspectives in a concise way [59]. In a fact sheet, multiple data facts, which are numerical or statistical results derived from data [48], are composed together to tell a data story. Visualizations and infographic components are usually adopted to illustrate the data facts. Leveraging captivating visuals instead of a text-heavy explanation, fact sheets help ease the process of absorbing information by humans [6].

However, the creation of fact sheets is not easy. It requires two completely different types of expertise, namely, data analysis and graphical design. On one side, a fact sheet needs to be informative, interesting, thought-provoking, insightful, and reliable [22]. To compose a fact sheet, users need to explore data, find important data facts, and organize them into an interesting data story, which is demanding and burdensome for most users. Taking on this opportunity, commercial tools such as Power BI [2] and Google Sheets [1] have introduced functions to help users gain insights instantly by automatically recommending data facts. However, data facts generated by these tools have no logical connections, and users have no idea of the whole picture of the data. Therefore, they still need to examine the entire pool of recommendations to distill a meaningful story. On the other side, a good data fact sheet needs to be not only informative and interesting, but also aesthetically pleasing. To create a fact sheet, design tools, such as Adobe Illustrator, are common choices for professional designers. Recently, researchers have improved the design environments to ease the creation of flexible and expressive data-driven infographics, such as DDG [23], Data illustrator [28], Charticulator [40], and InfoNice [58]. While these tools have more or less simplified the process of conducting data-binding and editing graphical shapes, users still need to take a great amount of time to consider fact sheet content and design choices autonomously, and do data binding manually with trial-and-error.

In practice, data scientists and graphical designers usually work together, communicate closely, and design iteratively to compose visually appealing data fact sheets. They need to determine fact sheet topics, choose important data attributes, comprehensible visual representations, and adjust the overall visual effects, causing a significant amount of communication and trivial design issues. For general users without data and design expertise, such tasks can become practically impossible [16]. It is, therefore, unsurprising that most fact sheets are only created by data and design professionals.

The goal of this research is to significantly reduce the efforts involved in creating fact sheets and make fact sheet creation accessible to general users. We take a new approach to ease the fact sheet design process: to automatically generate fact sheets from tabular data. We choose tabular data because it is widely used and familiar to general users. There are two main challenges to overcome. The first challenge is to extract data facts from the data table and organize facts into meaningful topics. The generated data facts should be reliable and interesting. The topics extracted should be meaningful and understandable. The second challenge is to choose proper visualizations that can demonstrate the data facts. To transform data content into a page of expressive infographics, the fact sheet design should consider not only the visual element-level but also the sheet-level presentation.

To achieve this, we propose a fact sheet generation framework based on a formative survey on a collection of awarded infographics designs. The framework consists of three parts, namely, fact extraction, fact composition and visual synthesis, corresponding to the challenges above. Then, we implement a proof-of-concept system that automatically creates infographic fact sheets. Given a tabular dataset with multiple columns and rows, we first extract various data facts based on the statistic characteristics of the columns and rows with importance scores. Then, we organize the data facts into different topics and select the facts most related to the topics. After that, we visualize the data facts and add descriptions. We arrange the data facts of a topic into one page and unify the style of the fact sheet. Users can further customize the data fact sheet according to their needs.

Considering the huge design space of data fact sheets, the fact sheets generated cannot cover all the fact choices and visual representations. Instead, the generated fact sheets provide a stepping stone for users to explore, organize, and design data facts. Based on our design candidates, users can further customize the fact sheets to cater to their needs. The contributions of this paper are threefold:

- We investigate an award-winning infographic dataset to analyze the common design practices of data fact sheets.
- We describe the DataShot framework with novel techniques to organize data facts into topics and transform data facts into fact sheets. To validate our techniques, we implement a proof-of-concept system to automatically compose fact sheets from tabular data.
- We use real-world data to demonstrate the usage of our system and conduct an in-lab user study to reveal the potential benefits of DataShot.

2 Related Work

2.1 Data-Driven Storytelling

Data-driven storytelling has been discussed extensively in recent research publications. The research on data-driven storytelling is about how to communicate data effectively and give data a voice.

- *Y. Wang, H. Zhang, W. Cui, and D. Zhang are with Microsoft Research Asia. E-mails: {wangyun, haizhang, weiweicu, and dongmeiz}@microsoft.com*
- *Z. Sun, K. Xu, and X. Ma are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. Work done during Z. Sun and K. Xu's internship at Microsoft Research Asia. E-mails: {zhida.sun, kxuak}@connect.ust.hk and mxj@cse.ust.hk*
- **These authors contributed equally to this work.*

Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.
Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2019.2934398

Segel and Heer [46] identified seven genres of narrative visualization. Two later separate studies [26, 27] discuss and summarize the future research directions of data-driven storytelling. Recently, other researchers [32, 41, 49] have further studied the design space and theories around data-driven storytelling. As fact sheets can also be viewed as a type of data-driven storytelling, our survey on fact sheet design further extends this line of research.

The research of creating data-driven storytelling can be categorized into two branches. One branch focuses on authoring systems to ease the design process, with the assumption that users already have a deep understanding of data and what they want to present. These systems are usually designed for a specific genre of story to cater to users' needs. For example, DataClips [4] helps users to craft data-driven video easily; Ellipsis [44] and ChartAccent [38] enable dynamic annotations to support data storytelling; Timeline Storyteller [9] is a tool for expressive timeline narratives.

The other branch of data storytelling research focuses more on the automated systems and techniques to save users' efforts. These systems take data as input and generate organizations of the story components for users. Most of them target a specific domain or data format and generate stories accordingly. For example, NewsViews [18] is an automated pipeline to create annotated maps from news articles. Contextifier [20] annotates time series stock data with text extracted from stock news. Wang *et al.* [57] visualize the MOOC temporal changes with animated visualization to highlight different categories of key events within a time span. Bryan *et al.* [10] create auto-annotated visual summary images through an interactive exploration process for time varying data. Some researchers study the automated organization of data content. For example, Hullman *et al.* [21] study story composition into a linear narrative presentation flow. GraphScape further builds a graph model for sequencing based on visualization similarity [24].

To support data storytelling, DataShot organizes automatically generated data facts into topics around the data. From a large set of data fact candidates, we extract and rank the topics extracted from the table to help users easily get an overview of the dataset from different aspects.

2.2 Visualization Recommendation

Researchers have explored the generation of visualizations for years. Visualization generation systems usually aim to help users interactively explore data and provide insights. Mackinlay's APT [29] introduces a compositional algebra to enumerate visualization encoding. The research of automatic visualization recommendation has long used rules and heuristics to facilitate exploratory data analysis. *e.g.,* Tableau, Power BI and Polestar [36] (formerly Polaris [50]) recommend charts based on user specifications of the data fields. Users need to specify data fields by dragging data fields into the x- and y- axes.

More recently, researchers have started to recommend interesting visualizations based on the statistical properties of data and search more deeply into any possible data insights from the data table. For example, Voyager [60] and Voyager 2 [61] recommend visualization charts based on statistical and perceptual measures; DataSite proactively recommends data analysis results with a set of heuristic algorithms [13]; Foresight recommends data visualization ranked within different types of data insights for large scale data [14]; Tang *et al.* [52] and Vartak *et al.* [56] further recommended top-k insights with respect to an importance or interestingness measure; Srinivasan *et al.* explore how system-generated data facts can be illustrated with visualizations [48]. Industry systems such as Microsoft Power BI and Google Sheets [15] also recommend visual charts based on the data insights detected by the insight mining engines. These studies are formative for DataShot, and our research further supports users' understanding and presentation of data by a fact sheet generation workflow, including fact extraction, composition, and visualization.

2.3 Infographic Creation Tools

Infographics can effectively and engagingly convey data, knowledge, and insights. Although the forms and objectives of infographics diverge greatly, the common characteristic of infographics is the combination of data visualization and text description, embellished with icons and images. Mostly created by graphical designers, an artistic and embellished design can help data information spread quickly and widely. To demonstrate data-heavy information, visualizations are designed to support data presentation. To create these visualizations, designers adopt different methods. Programming toolkits (*e.g.,* D3) are widely used due to their flexibility. They give content creators a high degree of control with the help of the strong expressive power of programming languages. However, the flexibility of programming languages comes with a steep learning curve for most designers.

Modern design environments are developed to lower this barrier. Commercial products, such as Adobe Illustrator and Microsoft PowerPoint, are extensively used by professional and non-expert users. While they provides high flexibility, they lack effective data binding functions. Research on visualization design environments are developed to ease the process. For example, Lyra [45] used a drag-and-drop manner to enable the design of customized visualizations; iVisDesigner [39] incorporates similar configuration panels and menus to facilitate users specifying data bindings; Data Illustrator [28] helps designers define various combinations of shapes and their data bindings; meanwhile, Charticulator [40] focuses more on visualization layout. These systems support the visual configuration of data binding with basic graphical shapes. More recently, researchers have proposed design tools for more expressive and flexible infographic design by including images, icons, and hand-drawn shapes. For example, InfoNice (*a.k.a.* Infographic Designer) [58] supports expressive pictographs through a mark customization approach. Users can specify how icons, images, and texts are bound with data. DDG [23] and DataInk [62] enables data binding with hand-drawn shapes.

These systems enable infographic design through user specifications. However, they are not fully automated. Users need to have an idea about the design target. We simplify this whole process to a one-click experience. We support infographic-style visualization templates and trained a decision tree model to select the best visualization style and generate fact sheets.

3 Survey on Fact Sheet Design

To understand how designers make design choices when they create fact sheets, we conduct a formative study on a collection of award-winning examples before we explore the automatic generation method. In this section, we first describe in detail the collected fact sheet dataset; then we introduce our qualitative analysis method; and finally we present the findings, which we have derived from the analysis, to characterize the general patterns for fact sheet design in practice.

3.1 The Fact Sheet Dataset

As fact sheets are widely adopted in the real world, there is an abundance of design sources available on the Internet. To make the dataset practical and cover enough topics and presentations, we choose the *Kantar Information is Beautiful Awards* [22] as our data source and finalize a dataset with 298 infographic examples after retrieving the works under the *Infographic* genre during the period 2012-2018. We select this dataset because the works in it are of high quality and intensive coverage. Specifically, 1) These works have been awarded by a panel of experts [11] based on the evaluation criteria on topic interestingness, data accuracy, information usefulness, and presentation design [22]; 2) These works cover a wide variety of topics and domains. The examples in our dataset illuminate *Arts* (1.0%; 3), *Breaking News* (1.7%; 5), *Business, Finance and Marketing* (17.8%; 53), *People, Language and Identity* (12.4%; 37), *Politics, Global and Humanitarian* (30.5%; 91), *Science and Technology* (8.1%; 24), *Leisure, Games and Sport* (18.1%; 54), and *Nature* (10.4%; 31); 3) These works have diverse cultural backgrounds. Other than English (92.3%; 275), our dataset also includes infographics presented in other languages, such as German (2.3%; 7), Chinese (1.3%; 4), French (1.3%; 4), Spanish (1.0%; 3), Italian (1.0%; 3), Portuguese (.3%; 1), and Dutch (.3%; 1); 4) These works have been delivered across a broad spectrum of communication medias, including educational materials, news reports, and personal blogs, *etc.* We do not report the statistical results of media source because some works lack the source information while some have multiple publishing channels. The entire exemplary showcase can be accessed online[1].

We further remove 53 examples from the dataset, as they contain specially designed artistic images or glyphs, which do not meet the definition of fact sheets. As a result, we finally got 245 fact sheets (82.2% of the original dataset) that consist of common single visuals or composite visuals.

[1] *Kantar Information is Beautiful Awards*, accessed March, 2019: `https://www.informationisbeautifulawards.com/showcase?acategory=infographic&type=awards`

3.2 Qualitative Analysis

We conduct a qualitative analysis of the 245 examples to further understand the fact sheet designs from the following aspects:

- **Content structure**: How do designers structure the content?
- **Presentation layout**: What types of layouts are usually employed?
- **Visualization style**: How do designers choose the visualization styles to represent the data facts?
- **Fact type**: What kinds of data facts are usually covered?

The qualitative analysis is divided into two stages. In the first stage, our target is the entire fact sheet design. We examine the sheet-level designs of all fact sheet examples to learn the patterns on content structures and presentation layouts. In the second stage, we drill down into each component of the infographic or chart presented in the whole fact sheet and consider them as visual elements inside a fact sheet. We aim to investigate the element-level designs to identify commonly adopted visualization styles and fact types based on the detailed visual elements of the fact sheets. Two coders went through all the exemplars and conducted independent coding along the predefined scheme. The disagreements were resolved through discussion.

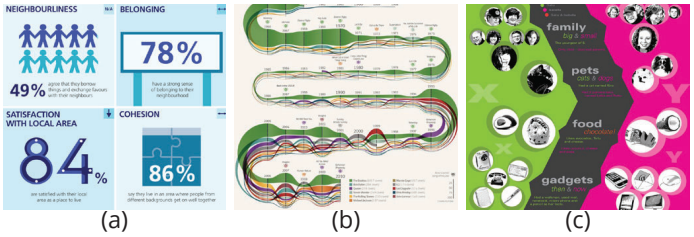
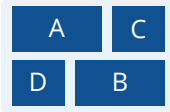


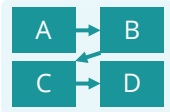
Figure 1. The sample fact sheets with three different content structures: (a) random facts [55]; (b) sequence [30]; and (c) multiple plots [42].

3.3 Sheet-Level Design

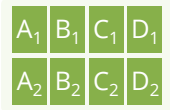
To explore the fact sheet design patterns related to content structures and presentation layouts, we borrow the definition of "panel" (*i.e.,* an independent module that can convey a full fact and/or complete story) from [5] and take it as the coding unit. We coded three content structures (Figure 1) of fact sheets:



Random Facts (52.3%; 128): Designers arrange the fact sheet elements in random order without specifying any logical relationship among all the elements. Swapping any two elements will not affect the clarity and easiness of understanding the fact sheet.



Sequence (25.7%; 63): Fact sheet elements are arranged in a sequential order. Such an order can be a specified time series, procedure, or narrative structure, *etc.* The positions of the fact sheet elements are fixed and cannot be changed.



Multiple Plots (22.0%; 54): Designers compare or contrast the information of different topics/subjects from the same aspect by a series of fact sheet elements. Elements can switch positions following certain rules (*e.g.,* along the measure) but not within the sub-groups.

To learn common presentation layouts applied in fact sheets, we adopt the same method in Bach *et al.* [5], where the authors have identified nine panel layouts (*i.e., Large panel*, *Annotated*, *Tiled*, *Grouped*, *Grid*, *Parallel*, *Network*, *Branched*, and *Linear*) based on a set of 59 paper-presented infographics. Since we focus on the sheet-level layouts in our analysis, we specifically remove the category of "*Grouped*", which is the hierarchical structure mixed within fact sheet elements. As a result, we coded eight layouts types for fact sheets. The results are shown in Table 1.

Figure 2 shows the distributions of different presentation layouts in each type of content structure. We identify that the "**Random facts**" structure and the "*Tiled*" layout are the most commonly adopted presentation in practise. Moreover, the "*Tiled*" layout in "**Random facts**" structure (Count = 76) are also applied more often than other presentation combinations. Therefore, we decide to start with these two configurations to develop our auto-generation method for fact sheets.

Table 1. Distribution of different presentation layouts

Layout	Count	Ratio	Layout	Count	Ratio
Tiled	**99**	**40.4%**	Parallel	70	28.6%
Linear	26	10.6%	Large panel	26	10.6%
Grid	10	4.1%	Annotated	7	2.9%
Network	5	2.0%	Branched	2	.8%

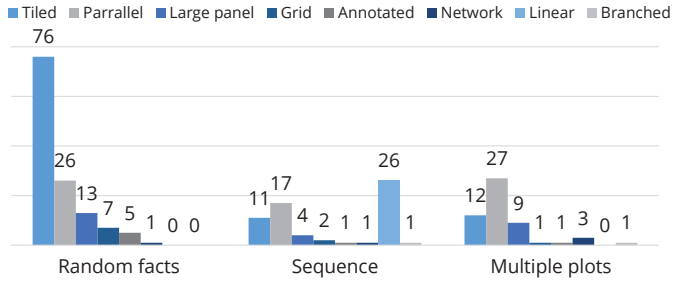


Figure 2. Distributions of different presentation layouts across content structures.

3.4 Element-Level Design

To investigate the element-level designs, we decompose each example fact sheet into atomic visual elements. The visual elements are usually infographics or visualizations. Each element corresponds to a complete fact and could not be further unpacked into any other lower level fact sheet components. We further extract 793 single fact sheet elements from the 245 examples. We code the basic visualization types (Table 2) appearing in the fact sheet examples following the taxonomy in [8].

Table 2. Distribution of different visual element types

Type	Count	Ratio	Type	Count	Ratio
Image	125	15.8%	Area	100	12.6%
Pictogram	94	11.9%	Pie/Donut	91	11.5%
Text/Number	86	10.8%	Bar	86	10.8%
Column	80	10.1%	Line	55	6.9%
Map	30	3.8%	Heatmap	15	1.9%
Treemap	12	1.5%	Sankey Diagram	10	1.3%
Scatter Plot	5	.6%	Venn Diagram	4	.5%

We also code commonly used fact types by referring to low-level analysis tasks [3] and facts taxonomy [12]. Based on the survey results, we summarize 11 categories of fact types that are commonly adopted in fact sheets with examples as follows.

Value (24.5%; 194): Retrieve the exact value of data attribute(s) under a set of specific criteria. Such facts answer the question of "what is/are the value(s) of $\{A, B...\}$ in the criteria of $\{X, Y...\}$". For example, "46 horses have won two out of tree Triple Crown Races".

Proportion (15.0%; 119): Measure the proportion of selected data attribute(s) within a specified set. Such facts answer the question of "what is the proportion of data attribute(s) $\{A, B...\}$ in a given set S". For example, "Protein takes 66% in the diet on Sunday".

Difference (14.4%; 114): Compare any two/more data attributes or compare the target object with previous values along with the time series. Such facts answer the question of "what is the difference between data attributes $\{A, B...\}$ within a given set S". For example, "There are more blocked beds in the Royal London Hospital compared with the UK average".

Distribution (11.5%; 91): Demonstrate the amount of value shared across the selected data attribute or show the breakdown of all data attributes. Such facts answer the question of "what is the summary/overall distribution over the data attribute(s) $\{A, B...\}$". For example, "The distribution of the unicorn companies is approximately normal over their age".

Trend (10.2%; 81): Present a general tendency over a time segment. Such facts answer the question of "what is the trend of the data attributes $\{A, B...\}$ over a period of time T". For example, "The budget for the Border Patrol Program has been rising from 1990 to 2013".

Rank (9.1%; 72): Sort the data attributes based on their values and show the breakdown of selected data attributes. Such facts answer the question of "what is the order of the selected data attribute(s) $\{A, B...\}$".

Table 3. The statistical results of visualization and fact types based on the collected fact sheets. "C" represents categorical data (includes geological information), "N" represents numerical data, and "T" represents temporal data. Note that when categorical fields are taken as measures, they are counted, listed, or encoded with colors. The numbers in the cells indicate the total amount of charts surveyed from our infographic dataset. Each row represents a certain type of fact.

Fact type	Analysis Task [3]	Dimension	Measure	Image	Area	Pictogram	Pie/Donut	Text/Number	Bar	Column	Line	Map	Heatmap	Treemap	Sankey Diagram	Scatter plot	Venn diagram
Value	Retrieve Value	C	-	59		4		5				3					
		-	N		3	8		16	1								
		C	C	10		3		1	1			2	1				
		C	N	11	15	12	5		4	7	1	4	1	1			
		C×C	C										2				
		C×C	N		3		1				1		1				
		T	C	5													
		T	N		1					1			1				
Categorization	Filter	C	-	3		2		1				1					
		C	C	2		1											
		T	C					1									
Aggregation	Compute Derived Value	-	N			6	2	14									
		C	N	4	2	7	2	6									
		T	N					1									
Extreme	Find Extremum	C	-	2				2									
		-	N					2									
		C	C		1								1				
		C	N		3	3	1	1	6	2							
		C×C	N			2											
Rank	Sort	C	N	5	17	3		3	24	15							
		T	N						1	3			1				
Proportion	-	-	N		2	6	15	12	2								
		C	N	6	4	10	47	1	7	2							
		T	N				1										
		T×C	N		1		2		1								
Distribution	Characterize Distribution	C	C	2				4		1	1	5					
		C	N	3	12	6	3		17	14	8	8	1				
		C×C	C										1				
		C×C	N		1				1	3							
Trend	-	T	C	4	1	1					2						
		T	N		8	1			3	17	37						
		T×C	N		1				2	1	3						
Difference	Compare	C	-	1													
		-	N		1		1	9									
		C	C	2	1					2							
		C	N	5	20	19	6	6	13	8		1					
		C×C	N		1		2		1								
		T	C	1				1									
		T	N		2		3		2	4	2						
Outlier	Find Anomalies	N	N													5	
Association	Correlate	C	C									6	1	2	10		3
		C	N										3	9			1
		C×C	N										1				

For example, "The top reason for consumers to engage in showrooming is '(the) price (is) better online'".

Aggregation (5.5%; 44): Calculate the descriptive statistical indicators (*e.g.,* average, sum, count, *etc.*) based on the data attributes. Such facts answer the question of "what is the value of the statistics function F over the data attribute(s) $\{A, B...\}$". For example, "The national average price for regular gas is $4.06 in July 2008".

Association (4.5%; 36): Identify the useful relationship between two data attributes or among multiple attributes. Such facts answer the question of "what is the correlation between/among data attributes $\{A, B...\}$ over a given set S". For example, "There is a negative correlation between the number of quality food and the distance between the vendor city and the eastern market".

Extreme (3.3%; 26): Find the extreme data cases along with the data attributes or within a certain range. Such facts answer the question of "what is/are the top/bottom N or -est value regarding attribute(s) $\{A, B...\}$". For example, "The character with the most epigrams in the collected dataset is Oscar Wilde himself, who has 12".

Categorization (1.4%; 11): Select the data attribute(s) that satisfy certain conditions. Such facts answer the question of "what is/are the data attribute(s) $\{A, B...\}$ which satisfy conditions $\{X, Y...\}$". For example, "Joshua and Samuel are two popular names for boys in 2004".

Outlier (.6%; 5): Explore the unexpected data attribute(s) or statistical outlier(s) from a given set. Such a fact answers the question of "what are the exceptional data attribute(s) $\{A, B...\}$ in a given set S". For example, "*Rocky Raccoon* has the most unique words given the other songs from the Beatles".

Following the fact type definitions, we further map the visualization styles to different data facts by referring to the input data types. We present the statistical results of visualization styles and fact types in Table 3. From the results we see that images are widely adopted among the majority fact types, and this finding aligns with the conclusion of [11]. Besides that, designers prefer to apply line charts to represent "**Trend**"s and pie/donut charts to show "**Proportion**"s. Interestingly, we find designers have creative ways of adopting visual charts. For example, they sometimes use length or area (*e.g.,* bar chart) to encode categorical data, showing ambiguous values or varying degrees (*e.g.,* "a lot" *vs.* "a little", "big" *vs.* "small"), different from common visualization guidelines. We exclude such cases when we design a fact-visual mapping model for our system.

4 DataShot

From our survey results (Section 3), we observed that the design space of fact sheets is huge. Therefore, as a first step, we decided to design DataShot by starting with the majority design options. According to Figure 2, we selected *random facts* for content structure and *tiled* for presentation layout to build our auto-generation method for data sheets. Overall, our proposed solution consists of three modules, *i.e.,* Facts Extraction, Fact Composition, and Visual Synthesis. In this section, we first discuss the design goals we proposed before implementing the proof-of-concept system; then we present the system pipeline; and finally we describe the details of each module.

4.1 Design Goals

In DataShot, our main goal is to minimize users' efforts for creating fact sheets from tabular data. More specifically, we conceive to achieve the following five goals:

G1 Ensure data facts' accuracy and reliability. The system should make sure that all the computed results are based on the original tabular data, and ensure the appropriateness of visual presentation.

G2 Support efficient data fact extraction. The system should offer data-driven insights quickly. To lower the barriers of processing tabular data, novices with less data analysis knowledge can obtain meaningful results without technical obstacles.

G3 Organize data facts into meaningful topics. The system should organize the computed data facts into related and meaningful topics. The topics should be able to cover the important data fact elements, and deepen users' understanding of the data.

G4 Aim for a succinct and expressive presentation. The visual design should be able to express data facts.The audience can learn the facts in a clear and understandable manner.

G5 Enable simple user interactions. To facilitate better understanding and presentation of the data, the system should enable users to explore and modify the fact sheet content based on their interests.

4.2 System Pipeline

Based on the aforementioned design goals, we follow the common practice of designing graphical representations [31, p. 18] and propose the DataShot system to automatically compose fact sheets from tabular data. To achieve this, we first extract common data facts from the data table based on different data subspaces. After that, we organize data facts into topics, choose the best topics, and select top-n facts from fact candidates. Then we visualize each data fact and compose the topics into a fact sheet. Figure 3 shows the fact sheet auto-generation pipeline, which consists of three core modules:

P1 Fact Extraction. The system first transforms the raw tabular data into data facts. During the preprocessing procedures, the system constructs data subspaces, enumerates fact types, and calculates fact scores based on original tabular data (**G1**). All these steps are conducted at the back-end module without extra input (**G2**).

P2 Fact Composition. After collecting all the computed results, this module runs topic extraction and ranking algorithms to select the recommended fact sheets. For each fact sheet, N top facts are selected from the data fact candidate pool. (**G3**).

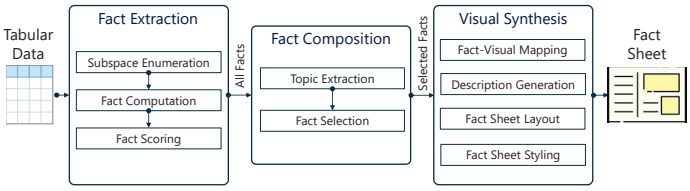


Figure 3. DataShot system overview and processing pipeline.

P3 Visual Synthesis. The visual representation module aims for presenting the final computed result to the end users. Through matching for suitable visualization style and computing the succinct layout, the system supports the presentation of extracted data facts in a fact sheet manner (**G4**). Considering the individual differences in interpreting fact sheet content, the system enables the end users to interactively revise the final presentation results (**G5**).

4.3 Fact Formulation

A tabular dataset with rows and columns can conceptually be modeled as multi-dimensional data. There are two types of usages for columns in a table: dimensions and measures. Dimensions are used to group or filter records, while measures are numerical columns on which certain aggregations (*e.g.,* SUM, AVG) can be performed. DataShot treats categorical or temporal values as dimensions and numerical values as measures. For the aggregation function COUNT, categorical columns are also treated as measures. In this section, we use a simplified *TabletSales* table as a running example (as shown in Table 4) in our descriptions.

Table 4. A sample tabular dataset of tablet sales.

Year	OS	Region	Country	Sales(M)	Units
2010	iOS	USA	United States	1.1	11231
2011	Android	Asia	India	1.4	10342
2011	Android	Asia	China	1.5	28221
...	...	...	...	...	...

In the context of multi-dimensional data analysis, a data fact can generally be used to represent a certain type of numerical property on a specific data subspace or multiple data subspaces. In DataShot, we model a data fact from five perspectives, namely *Fact Type*, *Fact Parameters*, *Fact Measures*, *Fact Subject*, and *Fact Score*. Accordingly, we formalize *data fact* as a 5-tuple structure:

$$fact := \{type, parameters, measure(s), subject, score\} \quad (1)$$

Fact Type. We implement various fact types, corresponding to the summarized fact types based on the survey results in section 3.4. The facts are calculated through the combination of different dimensions and measures in a data table. For example, the number of different values inside a dimension column is a *Value* fact. To extract facts like, "There are in total 5 brands", we enumerate the Brand column and count the total different values of this column.

Fact Parameter. For each fact type, we need a set of parameters to describe the characteristics of a fact. For example, for *Trend* fact, does it increase or decrease? For *Extreme* fact, is it a maximal or minimal number? We put all the specific description arguments in $fact_{parameter}$.

Fact Measure. A fact measure is a field that is treated as a dependent variable for this fact; its value is derived from a function of one or more dimensions. For example "The Sales of iOS increased over the year" is a *Trend* fact. This is extracted by taking *Sales* as Measure, *Year* as Dimension, and further input the data into the trend detection engine.

Fact Subject. *Fact Subject* scopes the content of a data fact. A subject reflects three important factors: context, breakdown(s), and focus. To compose interesting data fact sheets, we not only need to extract superficial data facts, which gives overviews of different data tables. It is also important to get deep insights from the data. For example, users may need to know "sales by year in the United States" besides the overall "sales by year". Therefore, we need to construct data subspaces for facts. A *subspace* is defined as a combination of data filters as follows (where {f[i] : v[i]} denotes a filter item with a particular value v[i] on dimension f[i]):

$$subspace := \{\{f[1] : v[1]\}, ..., \{f[n] : v[n]\}\} \quad (2)$$

For example, $\{\{OS : iOS\}, \{Country : China\}\}$ is a subspace corresponding to "device sales in China for iOS".

For a data fact, the context is the whole data subspace. For each data fact, it should have only one context. The breakdown(s) of a data subspace is used to divide the subspace into child groups in this subspace. The focus of the subspace is one or more groups in this subspace to highlight. A fact can have no focus.

$$fact_{subject} := \{context, breakdown(s), focus\} \quad (3)$$

For example, a subject $\{\{*\}, Country, \{\}\}$ denotes a value (e.g., "Sales") break down by "Country" in the whole data space, which can be used to show a distribution; a subject $\{\{Year : 2010\}, Country, \{\}\}$ denotes a value breakdown by "Country" in "2010"; a subject $\{\{Year : 2010\}, Country, \{Japan\}\}$ denotes the distribution of a value with "Japan" highlighted, which is usually used to show data facts like *Extreme*, *Outlier*, etc.

Fact Score. Clearly, different data facts are not equally attractive to users. For example, users in general might be more interested to see a sharp rise in "Sales by Years for iOS" than a Sales value for "iOS in the Year 2010, Asia, China". Therefore, we need to consider the importance of the data for further fact selection. We represent the importance of a fact as a score.

4.4 Fact Extraction

The fact extraction module mines the input data table to systematically extract a variety of facts from the data **(G1)**. All the extracted data facts are saved in a data store as the input for the Fact Composition module. More specifically, this module first enumerates the data subspaces by slicing and dicing the data table with different dimensions. Within each data subspace, it conducts corresponding computations on measures to derive different types of data facts. It also determines the importance of each data fact by assigning it an importance score.

4.4.1 Subspace Enumeration

In our implementation, we employ the BUC algorithm [19] to explore the data and enumerate the subspaces in a top-down order. The enumeration starts from the apex cuboid, and iteratively adds a new dimension to break down the current cuboid to generate a new one. We limit the depth of the lattice of the cuboid to be three as a data subspace filtered by more than three dimensions is usually uninteresting to users. However, as an open framework, DataShot allows users to configure the search depth.

4.4.2 Fact Computation

DataShot conducts computation to search for different types of facts during the process of enumerating data subspaces. For each subspace, *Value* facts and *Aggregation* facts are calculated for each measure on this subspace. Currently, DataShot only supports one measure in each calculation. However, the framework can be easily extended to multiple measures. When a subspace is broken down by a dimension, specific types of facts are computed accordingly based on the relationship of the parent-subspace and the child-subspaces. For example, *Rank* facts and *Proportion* facts are extracted based on the corresponding ranks and ratios of each child-subspace within the parent-subspace for each measure. *Extreme* facts and *Outlier* facts are generated for those child-subspaces with minimum/maximum values or abnormal values. Furthermore, if the break-down dimension is a temporal one, a *Trend* fact can be obtained if the values of child-subspaces exhibit an increasing or decreasing trend along the temporal dimension. The Fact Computation part is designed as an extensible framework in our implementation. Therefore, new computation components can easily be plugged-in to support new fact types in the future.

4.4.3 Fact Scoring

When a fact is extracted, DataShot assigns a score to determine the importance of the fact. Fact scoring is widely adopted in insight recommendation systems. Based on state-of-the-art research on automatic insight mining [17, 52, 56], we consider two factors when calculating fact scores: significance and impact. The fact score is a normalized value between 0 and 1. The fact scoring function can be further extended to cater for different requirements.

Significance. The significance score, $score_{significance}$, reflects the importance of a fact from the aspect of statistical properties. For example, a sharp increase corresponds to a high score for *Trend*. In our example, a sharp increase of Sales in India over time corresponds to a high score. A high proportion corresponds to a high score for *Proportion*. In our example, a high proportion of Sales for iOS corresponds to a high score. For *significance*, we consider the statistical significance and the fact types. When extracting a data fact, we can get the statistical significance of the fact. We score the data fact based on the significance of the facts and normalize them. Some data facts (*e.g., Value*, *Aggregation*) only derive values from the data. We simply assign them to zero.

Impact. The Impact score reflects how general the data fact sheet is. A fact is considered less important than the facts from a larger data subspace. For *impact*, we consider the focus and fact context. For example, we should treat "iOS" as a more impactful data value than "Android" if there are more records of "iOS" than "Android" in the data table, and thus "iOS" is a larger data subspace. Another example is we treat a data fact under "iOS and China" as a less impactful space than "iOS" because "iOS and China" filters less data records than "iOS". Therefore, when considering context, we define

$$score_{context} = \frac{number_{record}(subspace)}{number_{record}(whole\ space)}. \quad (4)$$

Similarly, we define the impact of a focus as

$$score_{focus} = \frac{number_{record}(focus)}{number_{record}(whole\ space)}. \quad (5)$$

The overall score of a data fact is defined as a weighted sum of the three factors. Based on experiments, the default values of the weights are 0.6, 0.2, 0.2, respectively.

$$score = \omega_s \cdot score_{significance} + \omega_f \cdot score_{focus} + \omega_c \cdot score_{context} \quad (6)$$

4.5 Fact Composition

Once we collect a large set of data facts from the data extraction phase, we have a large set of data fact candidates, the number of which depends on the number of rows and columns of the data table. DataShot further composes the data facts into fact sheets for users to understand the data table from different perspectives **(G3)**.

4.5.1 Topic Extraction

DataShot extracts topics from the data facts by examining the fact subjects. Specifically, we select all the facts with the same filters for context or focus. Formally, the topic of a fact sheet is defined as a collection of data facts $sheet_{topic}(t) = \{fact[1], ..., fact[d]\}$, where fact[i].subject.context = t or fact[i].subject.focus = t. A fact with "iOS" as the focus, or "iOS" as context will come under the same topic. For example, the fact $\{top(n), OS, Sales, \{*\}, \{OS : iOS\}, 0.6\}$ denotes "the sales of iOS ranked first among all different OSes"; the fact $\{top(n), Country, Sales, \{OS : iOS\}, \{Country : USA\}, 0.5\}$ denotes "USA ranked first in sales among all countries for iOS". These are two examples both related to iOS. They put iOS as the focus and context, respectively. From this example, we find that mixing these two kinds of data facts makes a topic more diverse. If we only select facts with a certain focus, all the facts in a topic will relate to this topic. Or, if we only select facts with a certain context, all the facts will come under the same data subspace.

After we obtain a list of fact sheet topics, we need to recommend top-k fact sheet topics which are the most interesting to the users. We rank the topics based on the average score of all the related data facts. Then we get a ranked list of the topics for users to further explore.

4.5.2 Fact Selection

To generate a fact sheet titled with *"N facts about topic(t)"*, we need to select the top-*n* data facts from the related facts. Apparently, there might be a lengthy list of facts especially when the data table is big. Naturally, we select the top-*n* data facts with the highest scores. However, in many cases, the top facts might be very similar semantically because they share attribute(s) that are very statistically significant, dominating the final scores. To avoid this consequence, we propose a density-based top-*n* algorithm to select *n* data facts, balancing fact diversity and fact importance.

The algorithm works as follows: We take three elements of each fact tuple, *i.e.,* $\{type, measure, subject\}$, as a vector. Then we calculate the distance between every two data facts, applying similarity distance functions for categorical values. To calculate the distance between data facts, a set of measures can be adopted, which has been thoroughly researched and evaluated [7]. For simplicity, we choose to measure the overlap of the fact tuples for similarity calculation. By default, the three elements are equally treated. After that, we start from the top of the data facts ranked by score. We iterate through data facts one by one. For each potential data fact, if it is ranked highest and allowed to be chosen, we exclude the nearest (most similar) data facts in this iteration. If several data facts are equally scored, the farthest one is chosen. At the end of each iteration, we release the excluded data facts. We continue iterating until reaching the end of the list or when the top-*n* facts have been collected.

The distance that determines the nearest data facts is a changeable parameter. The larger the distance, the more diverse the facts are. When the distance is set to 0, the results of the algorithm are equivalent to directly taking the top-ranked data facts according to the fact score. Moreover, the weights of the five tuples can also be changed, for example, a larger weight for the "type" of results in more diverse fact types.

4.6 Visual Synthesis

Here we describe how we design and arrange the visual representations for different kinds of data facts in DataShot **(G4)**, including fact-visual mapping, fact description generation, fact sheet layout, and fact sheet styling.

4.6.1 Fact-Visual Mapping

After surveying prior work, we did not find any algorithms or rules available for mapping data facts onto infographics in fact sheets. Therefore, based on the results collected in our formative study, we build a fact-visual mapping model to find an ideal style of visualization for the extracted data facts. The design options are built by referring to the award-winning infographic examples. Following the formative study result, we support all the listed visualizations in Table 2. For popular visualizations (such as pie/donut chart), DataShot offers multiple styles of visualizations to enrich the fact sheet presentation. Following the method adopted in [31, 43], we train a decision tree model with the 793 exemplary infographic elements, taking data fact types and data types as the input of the model, and retrieve the suitable infographic or visualization options as the output. When applied in the system, the model may return one single choice for a certain data fact. In other cases, a set of equal-weight options are recommended. To filter the optimum result out, we further adopt two design criteria to extract the output from the potential candidates. Inspired by [37], we specify the following two constraints when selecting visuals:

- **Inter-consistency** When the similar data facts (*i.e.,* with the same $fact_{type}$ and $fact_{score}$ and same type of $fact_{subject}$ and $fact_{measure}$) appear in one single sheet, DataShot adopts the same visualization to present different facts to keep unity and support a visual comparison.
- **Intra-diversity** When data facts with the same $fact_{type}$ have different types of $fact_{subject}$ or $fact_{measure}$, or even different $fact_{score}$, DataShot adopts different types of visuals to present data facts to show diversity.

Under these two constraints, we compose the fact sheet by referring to all the adopted elements' visualization styles over the fact sheet presentation. To further make the visualization expressive and appealing, we employ the visualization variants based on our survey. All different design options are stored in the design option pool and selected based on the aforementioned two constraints. For example, among all the facts of sporty car sales, the "**Trend**" facts of the brand BMW, Ford, and Volkswagen over 2007 to 2011 are in the same type of $fact_{subject}$ and $fact_{measure}$ and with the same fact type and ranking. Therefore, DataShot keeps employing the same visual (*i.e.,* line charts) under the inter-consistency rule (Figure 4 (B)); but in shark attack data, the "**Proportion**" facts on swimming events with different $fact_{score}$ or different type of $fact_{subject}$ or $fact_{measure}$ will be assigned with different visuals from the visual design pool (*i.e.,* the various design for pie charts and pictographs) by following the intra-diversity rule (Figure 4 (A)). We automatically insert images and icons to make visuals more

Figure 4. Fact sheets generated by DataShot from three data tables, namely, *SharkAttack*, *CarSales*, and *SummerOlympics*. (A) is about the shark attack events happened in swimming activity; (B-C) are the sales status of sports cars and the manufacturer BMW; (D) shows the conditions for winning gold medals in the Summer Olympics from 1896 to 2012.

vivid. For example, we adopt pictographs and replace the icons based on the $fact_{subject}$. We have collected an icon library covering a range of topics, and accomplished this in a key word matching way. Icon-matching based on semantic similarity can also be adopted to improve the matching results [33]. Users can further replace the icons with the ones they have in mind.

4.6.2 Fact Description Generation

In addition to presenting data facts, fact sheet elements also rely on short descriptions to enrich the content and make visualizations easy to interpret. Readers are used to looking for descriptive messages to digest information conveyed by the visual design, while such meaningful explanations are highly related to the fact types themselves [25]. Therefore, we adopt a template-based method to generate corresponding descriptions for each type of data fact. To be more specific, we construct templates including subjects, measurements, dimensions, and fact details with additional statistical indicators for each type of fact. When presenting the explanatory note, we highlight the subject part with an emphasized font to make the information eye-catching. For example, a text description "For the Category of SUV, the increase in Sales in 2011 compared with 2010." is composed from $fact.subject$ ($fact.context = SUV$, $fact.focus = \{2010, 2011\}$), $fact.measure = \{Sales\}$, $fact.type = \{Difference\}$.

4.6.3 Fact Sheet Layout

After collecting all the fact sheet elements, we seek to arrange those candidates into one single page. Since *tiled* is the most chosen layout (40.4%) for all the fact sheets (Table 1) and also the most chosen layout (59.4%) for "random facts" sheet from our survey (Figure 2), we choose to implement a tiled layout to arrange data facts in DataShot. We arrange the fact sheet elements based on a fluid grid system for page layout [54]. To simplify the problem, we assign the same height for each visual element and adjust the width according to visual type and fact content [51]. More complex layout algorithms can also be adopted and extended to improve layout variability in DataShot [34, 53].

4.6.4 Fact Sheet Styling

To refine the final presentation of the generated fact sheets, we worked with a professional designer from a local tech company to style the fact sheet design. Our generated fact sheets can easily be customized according to users' needs. DataShot supports three styles, including font, title, and embellishments, to diversify the fact sheet presentation (as shown in Figure 4). Moreover, seven different color schemes are supported to enrich the final design. Users can combine the three styles and seven themes to create their own fact sheets.

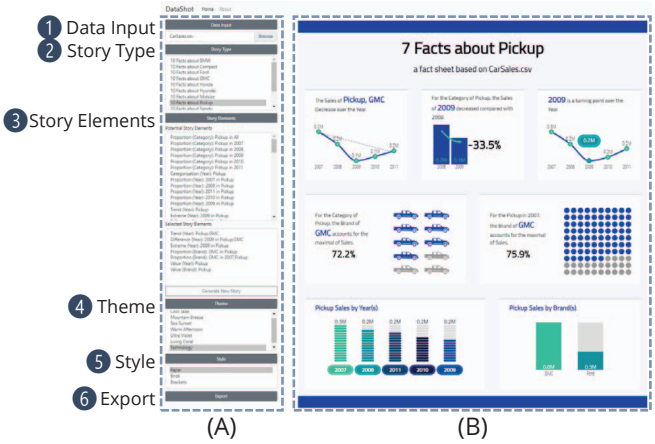


Figure 5. The interface of DataShot. (A) is the *control panel*; (B) is the fact sheet *presentation zone*. After clicking the item listed in "Data Source", corresponding data sheet will be presented in the presentation zone first. Users can further interact with the system to generate different fact sheets with different styles.

4.6.5 User Interface and Interaction

We describe how to interact with DataShot **(G5)**. The user interface of DataShot can be divided into two parts: the *control panel* and the *presentation zone*. By uploading a data table, the raw data will be shown in the presentation zone first (Figure 5 ①). At the same time, all the potential fact sheets with certain topics will also be listed in the "Story Type" field (Figure 5 ②). After clicking one topic, all the facts related to the topic will be shown as a fact sheet with correponding facts listed in the "Story Elements" field (Figure 5 ③). Users can further interact with the fact sheet elements by adding facts from the control panel or removing facts from the sheets. In the control panel, users can add data facts from the "Potential Story Elements" list or remove data facts from the "Selected Story Elements" list or directly from the fact sheet in the *presentation zone*. Users can also change "Theme" and "Style" to adjust presentation style, color theme, and font to further customize the fact sheet (Figure 5 ④⑤). Finally, users can export the fact sheets into PDF files by clicking the "Export" button (Figure 5 ⑥).

5 Evaluation

To evaluate DataShot, we first presented a set of use cases with real-world datasets to demonstrate the usefulness of DataShot. Further, we

carried out an in-lab user study to understand the usage and collect user feedback of DataShot.

5.1 Use Cases

To demonstrate the usefulness and expressiveness of DataShot, we show a set of fact sheets created from the datasets described in Table 5. We collect the datasets from the Internet and ensure that they cover different topics; The CarSales data is the sales records of eight different automobile manufacturers within five years (*i.e.,* 2007-2011). The data table consists of four columns (*i.e.,* brand, category, sales, and year); The SharkAttack data collects all the shark attack events across different countries and regions from 1554 to 2011. The data table consists of six columns (*i.e.,* activity type, #attacks, country, fatal, gender, and year); The SummerOlympics data lists the medals of the Summer Olympic games from 1896 to 2012. The data table consists of nine columns (*i.e.,* athletes' name, athletes' origin city, athletes' origin country, discipline, event, athletes' gender, medal type, sport type, and year).

Table 5. The details of datasets for use cases

Dataset	#Row	#Col	Data Type
CarSales	275	4	temporal, categorical, numerical
SharkAttack	4580	6	temporal, categorical, numerical
SummerOlympics	31165	9	temporal, categorical

We provide the final results generated by DataShot in Figure 4. There are many interesting insights, for example, Figure 4 (A) shows a fact sheet on the shark attacks with swimming activity. From the horizontal bar chart we can tell that swimming is the top activity for female who suffered shark attacks (Figure 4 ①). However, the swimming attacks which were fatal for females (25 times) were significantly lower when compared to males (305 times) in the column chart (Figure 4 ②). Figure 4 (B), (C) are generated based on the CarSales data. Figure 4 (B) presents the sales facts Sporty cars from all brands. We can easily tell that there are a total of three manufacturers who produce Sporty cars from the number (Figure 4 ③) and the sales trend decreases over the year in those line charts (Figure 4 ④). Figure 4 (C) gives the facts about the manufacturer BMW specifically. The fact sheet presents that the Compact category accounts for the majority of overall BMW sales from the donut chart (Figure 4 ⑤). While the sales of SUVs in 2011 increased compared with 2010 in the column chart (Figure 4 ⑥). Figure 4 (D) demonstrates nine interesting facts under the topic of gold medals based on the SummerOlympics data. From the fact sheet results we see that the total number of gold medals increased over the years, and the years in which the Olympics were not held for special reasons were not shown in the line chart (Figure 4 ⑦). In addition, the pictograph shows that gold medals for aquatics dominate women's sports (Figure 4 ⑧).

5.2 User Study

The overall goal of our evaluation was to investigate how users understand data facts and the corresponding visuals generated by the proposed pipeline in real cases. To this end, we conducted an in-lab user study with 10 users to assess the following aspects:

- whether DataShot can assist users in finding insightful facts;
- whether the visualisations generated by DataShot can aid users in understanding the data facts;
- whether the final presentation of DataShot can help users communicate their findings.

Since DataShot is the first system proposed to generate fact sheets directly from tabular data, there is no ideal tool or technique for us to compare with. Thus, we designed a user study to ask the participants to evaluate DataShot with the questionnaire of Likert scales first, then collect their feedback according to their experience with DataShot through post-study interviews.

5.2.1 Apparatus

We implemented DataShot on a Windows 10 operating system, running with Google Chrome and a monitor with a resolution of 1920×1200. Participants interacted with the DataShot system by an external mouse and keyboard. The study took place in a quiet office with only the investigators and participants involved.

5.2.2 Participants

We recruited 10 participants (aged between 22 to 30, $Mean_{age} = 25$, three females) through recruitment messages posted on social media platforms and word-of-mouth. Among all the participants, five of them have experience in designing infographics. In addition, three of the participants have a design background while the rest work in the computer science domain. All participants have the experience of using Microsoft Excel to analyze data and generate statistical graphs, and all the computer science background participants have experience in processing and presenting data with Python (*e.g.,* Pandas and Matplotlib). We did not financially compensate the participants for their efforts, and participation in this study was voluntary. Before the study, we confirmed with each participant to ensure none had seen any of the datasets used in this study before.

5.2.3 Study Procedure

The study started with a 10-minute tutorial introducing the DataShot's user interface and presenting a demo on how to generate fact sheets based on an example of tabular data. Then we asked the participants to freely explore the generated fact sheets and system, and create their own fact sheets with different tabular datasets. We encouraged them to ask any questions they had encountered during their interactions with DataShot. To observe user interactions and gain various subjective feedback, we did not set a fixed time limit. During the exploration, we adopted the think-aloud protocol to keep track of participants' intentions and their timely feedback. We collected participants' subjective feedback on DataShot with a 5-point Likert scale questionnaire derived from [23, 35]. The questionnaire is designed to assess DataShot from three aspects: (1) the content of the generated fact sheets; (2) the visualizations and designs of the fact sheets; (3) the overall user experience of the system. We also conducted a short interview with each participant and collected their feedback. The overall study lasted between 25-40 minutes.

5.2.4 Participant Feedback

All participants finished exploring the system while voicing their intentions of moving to their next goal during the study. We did not strictly time the whole exploration period for each participant. For the results of a 5-point Likert scale evaluation, where "1" means *strongly disagree* and "5" means *strongly agree*, participants generally assessed their experience of DataShot with a positive feedback, as shown in Table 6. Participants also provided promising feedback in the post-study questionnaire. Generally, all participants agree that DataShot is a useful and efficient tool for presenting tabular data with meaningful insights and easy to understand charts. They would like to use DataShot to help them understand the data with intriguing fact details.

Table 6. Overall ratings of DataShot on a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*).

Assessment	Measurement	Mean	SD
Content	Insightful	3.90	.57
	Comprehensive	3.90	.74
	Effective	4.60	.52
Visual	Easy to understand	4.70	.48
	Aesthetic	3.90	.74
	Expressive	3.70	.67
System	Useful	4.00	.67
	Usability	4.30	.67

Effective Data Fact Extraction Ease Data Exploration Participants responded positively to interacting with DataShot and appreciated the efficiency of the system for data exploration. Two participants commented that "*(DataShot is) so efficient with just one single click...*" (P8, female, 30) and "*(DataShot is) like a simultaneous data translator*" (P7, male, 22). One participant commented "*...(DataShot) saved a lot of time on exploring new data*" (P2, male, 28); "It saved me a lot of effort to find out all those significant differences" (P3, male, 22). Participants all agreed that they can benefit from the rapid and comprehensive data processing module, as the whole procedure significantly reduced excessive dependence on human resources and error-prone

processes. Unlike previous analyzing and mining tasks, users do not need to delve into detailed data relationships to find interesting facts. Furthermore, participants also reported that fact processing results are insightful enough to facilitate the understanding of the data, "*I used to get lost when I had a new data sheet...This tool makes it much easier to let me know where to get started*" (P2, male, 28), and help them find a promising direction for explaining the original tabular data.

Expressive Visual Design Improve Data Comprehension Participants also provided other positive comments on DataShot's visual design. They all agreed the visualization presented in DataShot is easy to understand. Participants commented regarding the visual design that "*...(DataShot) provides meaningful annotations for me to understand the data better*" (P1, male, 22), and "*the short description (of each fact sheet elements) helps me dive into the content of the data*" (P3, male, 22). At the same time, participants commented on the expressiveness of DataShot, "*The visual design is quite rich...I have plenty of choices to show the data from different aspects*" (P9, female, 29). Moreover, participants also commented on the attractiveness of the fact sheet design. "*I got a rough idea after I read the data, but DataShot presents the data in a more thought-provoking and eye-catching way*" (P9, female, 29). One participant (P7, male, 22) noted, "*I want to copy the charts into my presentation slides*". However, participants also pointed out that such an auto-generated process could diminish the diversity and decrease the creativity of the final output, which could hamper the expressiveness.

Enable Control Over Data With Rich Interactions We also learned lessons on implying necessary improvements to enhance the performance of DataShot. First, participants suggested improving the data sheets either from the back-end data processing module or through interactions with the end users. During the study, one participant (P2, male, 28) pointed out that the system should further process the data according to the semantic information of the data items. For example, we can assign different values to binary data item as "0/1", "false/true", or phrases with meaningful content, like "without/with...". Second, participants reported they would prefer more functional interactions albeit the current interface was simple and easy-to-use. The limited system interactivity affects users' understanding of the overall dataset. As one participant (P4, male, 22) commented, "*I think it would be more interesting if I could update the visualization styles...I can explore the other visualization types to gain insights from different aspects*". Another participant (P1, male, 22) mentioned, "*...the ideal way is that I could choose a data attribute, then I can further decide what information I would like to dig for in the following steps...The current list is not clear for me to tell the difference between those items...I feel that I have no freedom to explore the data*". Lastly, one participant (P8, female, 30) expressed her thoughts in that "*I prefer to edit the infographic facts sheet with more design functions...I have a lot of ideas but I cannot apply them to the design*". Since the focus of our research is the automatic generation of fact sheets, these concerns are beyond the scope of our paper. We further discuss these thoughtful comments in the next section.

6 Discussion

Benefits of Auto-Generated Fact Sheets. The creation of fact sheets is traditionally the joint work of data analysts and graphic designers. DataShot has made it possible for non-experts to create fact sheets easily. In our user study, our participants commended the convenience of generating fact sheets with just a click. In addition to the convenience, DataShot deeply integrates data exploration and data presentation, which allows people to process and consume data in a different paradigm. Traditionally, visual analytics systems usually follow an "overview first, zoom and filter, then details on demand" mantra [47]. Our study results suggest a possible new way of exploring data, "results first, assess and select, then refinements on demand". The fact sheets generated from DataShot organize data facts in an inviting and meaningful way and makes complex datasets more accessible to the users, inspiring and engaging them to explore data proactively. This indicates opportunities for the design of future data analytics workflow. Auto-generated data insights should be properly organized and presented to serve as a stepping stone for users' exploration and deeper understanding of data.

Opportunities for Infographic Authoring Tools. Obviously, it is very challenging, if not impossible, for a fully auto-generated fact sheet to meet users' needs perfectly. This indicates a new way of thinking for infographic and visualization authoring tools - presenting potential designs first and encouraging users to modify further. Users do not need to start from scratch, but instead, they can start from a draft. Although the current system of DataShot is limited in authoring interactions, our non-designer users are engaged with the presentation form and itch to create their own. This suggests an interesting avenue for the design process and interactions that supports further modification and re-creation. Previous design authoring systems take the assumption that every user has a clear design goal in mind, while this is often not true, especially for non-designers. To achieve this, additional considerations, such as novel editing operations and interactive design iterations, can also be put into the equation.

Limitations. There are several limitations in DataShot. First, we treat all data columns in a data table independently and do not consider semantic meanings among them. However, many datasets in the real world are highly domain-specific. For example, region and country have strong semantic relations or dependencies between themselves. Treating them as only categorical data impairs potential stories around the data. Second, the icons and visuals generated are largely from our predefined library, which is highly constrained with the library size. Our current system only supports the "tiled" layout style, which limits the flexibility and expressiveness of fact sheet design. As future work, we plan to integrate more fact sheet design choices to enrich our system. Third, the selection of visual types highly depends on our fact-visual model, which comes from the limited number of fact sheet examples in our survey. To determine the visualization choices, we base the dataset on the 793 infographics examples. For future work, we will include more data and labels from multiple sources to make the mapping more robust. Fourth, DataShot currently keeps the consistency of visualization types among similar data facts (e.g., same subject, measure). Further keeping the consistency of color encodings and axes among views may further improve the readability and interpretation of fact sheets [37]. We envision automatic consistency design algorithms to be studied to further improve the quality of the generated fact sheet. Fifth, the recommendation of facts depends on the scoring functions of our system. Recommending facts based on heuristic scoring functions is imperfect, and may introduce biases. It is a promising research direction to develop better fact evaluation metrics. More interactions can also be introduced to enable users to adjust parameters according to their needs. We believe our research opens the door to the future work of generating and authoring data fact sheets.

7 Conclusion and Future Work

In this paper, we have introduced a framework to automatically generate fact sheets from a tabular dataset. To understand how infographic fact sheets are designed in the wild, we have conducted an empirical study on 245 award-winning examples. To demonstrate its feasibility, we have implemented a proof-of-concept system that automatically generates fact sheets from tabular data for random facts. An automated pipeline has been proposed to enable the automatic design process. First, DataShot extracts a large number of interesting data facts from the data table. Then, DataShot organizes them into a ranked list of topics. After that, DataShot maps the data facts to visualizations based on a decision tree trained from 793 example elements. It then composes them into a one page fact sheet. Results from an in-lab study show that DataShot can effectively show the results and help users get a deeper understanding of the dataset. Participants are willing to customize the data sheet based on their understanding of the data and present the data with the fact sheets generated with DataShot.

As a proof-of-concept system, we have developed DataShot to support random facts, where all the facts around a topic are parallel. Our study shows benefits of a fact sheet generation system, which indicates potential research directions of future work. We plan to improve DataShot for other types of fact sheet structures and will explore more complex data storytelling logic and design of data fact sheets. At the same time, we will also explore more flexible ways of enabling users to plug in their own data fact and visual design preferences.

Acknowledgments

We would like to thank the anonymous reviewers and participants for their helpful suggestions and valuable feedback.

REFERENCES

[1] Google sheets. https://www.google.com/sheets/about/. (Accessed on 03/31/2019).

[2] Microsoft power bi. https://powerbi.microsoft.com/en-us/. (Accessed on 03/31/2019).

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117, Oct 2005. doi: 10.1109/INFVIS.2005.1532136

[4] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani. Authoring data-driven videos with dataclips. *IEEE transactions on visualization and computer graphics*, 23(1):501–510, 2017.

[5] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, and N. Henry Riche. Design patterns for data comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 38:1–38:12. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173612

[6] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2573–2582. ACM, 2010.

[7] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pp. 243–254. SIAM, 2008.

[8] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, Dec 2013. doi: 10.1109/TVCG.2013.234

[9] M. Brehmer, B. Lee, B. Bach, N. H. Riche, and T. Munzner. Timelines revisited: A design space and considerations for expressive storytelling. *IEEE transactions on visualization and computer graphics*, 23(9):2151–2164, 2017.

[10] C. Bryan, K. Ma, and J. Woodring. Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):511–520, Jan 2017. doi: 10.1109/TVCG.2016.2598876

[11] L. Byrne, D. Angus, and J. Wiles. Acquired codes of meaning in data visualization and infographics: Beyond perceptual primitives. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):509–518, Jan 2016. doi: 10.1109/TVCG.2015.2467321

[12] Y. Chen, J. Yang, and W. Ribarsky. Toward effective insight management in visual analytics systems. In *2009 IEEE Pacific Visualization Symposium*, pp. 49–56, April 2009. doi: 10.1109/PACIFICVIS.2009.4906837

[13] Z. Cui, S. K. Badam, M. A. Yalçin, and N. Elmqvist. Datasite: Proactive visual data exploration with computation of insight-based recommendations. *Information Visualization*, 18(2):251–267, 2019.

[14] c. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Recommending visual insights. *Proc. VLDB Endow.*, 10(12):1937–1940, Aug. 2017. doi: 10.14778/3137765.3137813

[15] K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan. Analyza: Exploring data with conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pp. 493–504. ACM, 2017.

[16] V. Dibia and Ç. Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *arXiv preprint arXiv:1804.03126*, 2018.

[17] J. Ding, S. Han, Y. Xu, H. Zhang, and D. Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *ACM SIGMOD/PODS International Conference on Management of Data*, June 2019.

[18] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos. Newsviews: an automated pipeline for creating custom geovisualizations for news. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 3005–3014. ACM, 2014.

[19] J. Han, M. Kamber, and D. Mining. Concepts and techniques. *Morgan Kaufmann*, 340:94104–3205, 2006.

[20] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2707–2716. ACM, 2013.

[21] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher, and E. Adar. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2406–2415, Dec 2013. doi: 10.1109/TVCG.2013.119

[22] Kantar. Information is beautiful awards. https://www.informationisbeautifulawards.com/. (Accessed on 03/31/2019).

[23] N. W. Kim, E. Schweickart, Z. Liu, M. Dontcheva, W. Li, J. Popovic, and H. Pfister. Data-driven guides: Supporting expressive design for information graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):491–500, Jan 2017. doi: 10.1109/TVCG.2016.2598620

[24] Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer. Graphscape: A model for automated reasoning about visualization similarity and sequencing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2628–2638. ACM, 2017.

[25] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 438. ACM, 2018.

[26] R. Kosara and J. Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013.

[27] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications*, 35(5):84–90, 2015.

[28] Z. Liu, J. Thompson, A. Wilson, M. Dontcheva, J. Delorey, S. Grigg, B. Kerr, and J. Stasko. Data illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 123:1–123:13. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173697

[29] J. D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5:110–141, 1986.

[30] M. Mauri. Cover mania-information is beautiful awards. https://www.informationisbeautifulawards.com/showcase/436-cover-mania. (Accessed on 03/31/2019).

[31] R. Mazza. *Introduction to information visualization*. Springer Science & Business Media, 2009.

[32] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. In *Computer Graphics Forum*, vol. 36, pp. 377–387. Wiley Online Library, 2017.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, eds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[34] P. O'Donovan, A. Agarwala, and A. Hertzmann. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 1221–1224. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702149

[35] H. L. O'Brien, P. Cairns, and M. Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28 – 39, 2018. doi: 10.1016/j.ijhcs.2018.01.004

[36] N. J. Pioch and J. O. Everett. Polestar: collaborative knowledge management and sensemaking tools for intelligence analysts. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 513–521. ACM, 2006.

[37] Z. Qu and J. Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):468–477, Jan 2018. doi: 10.1109/TVCG.2017.2744198

[38] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe. Chartaccent: Annotation for data-driven storytelling. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 230–239. IEEE, 2017.

[39] D. Ren, T. Höllerer, and X. Yuan. ivisdesigner: Expressive interactive design of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2092–2101, Dec 2014. doi: 10.1109/TVCG.2014.2346291

[40] D. Ren, B. Lee, and M. Brehmer. Charticulator: Interactive construction of bespoke chart layouts. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):789–799, Jan 2019. doi: 10.1109/TVCG.2018.2865158

[41] N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-driven storytelling*. CRC Press, 2018.

[42] G. Rodriguez. X and Y: a generational comparison — information is beautiful awards. https://www.informationisbeautifulawards.com/showcase/12-x-and-y-a-generational-comparison. (Accessed on 03/31/2019).

[43] B. Saket, A. Endert, and C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2829750

[44] A. Satyanarayan and J. Heer. Authoring narrative visualizations with ellipsis. In *Computer Graphics Forum*, vol. 33, pp. 361–370. Wiley Online Library, 2014.

[45] A. Satyanarayan and J. Heer. Lyra: An interactive visualization design environment. In *Computer Graphics Forum*, vol. 33, pp. 351–360. Wiley

Online Library, 2014.

[46] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[47] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *In IEEE Symposium on Visual Languages*, pp. 336–343, 1996.

[48] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics*, 25(1):672–681, 2018.

[49] C. D. Stolper, B. Lee, N. H. Riche, and J. Stasko. Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories. *Microsoft Research, Washington, USA*, 2016.

[50] C. Stolte and P. Hanrahan. Polaris: A system for query, analysis and visualization of multi-dimensional relational databases. In *INFOVIS*, 2000.

[51] Z. Sun, M. Sun, N. Cao, and X. Ma. Videoforest: Interactive visual summarization of video streams based on danmu data. In *SIGGRAPH ASIA 2016 Symposium on Visualization*, SA '16, pp. 10:1–10:8. ACM, New York, NY, USA, 2016. doi: 10.1145/3002151.3002159

[52] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1509–1524. ACM, 2017.

[53] K. Todi, D. Weir, and A. Oulasvirta. Sketchplore: Sketch and explore with a layout optimiser. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pp. 543–555. ACM, New York, NY, USA, 2016. doi: 10.1145/2901790.2901817

[54] B. Tondreau. *Layout Essentials: 100 Design Principles for Using Grids*. Design Essentials Series. Rockport Publishers, 2009.

[55] T. UK. TNS BMRB and Cabinet Office, Community Life infographic — information is beautiful awards. `https://www.informationisbeautifulawards.com/showcase/332-tns-bmrb-and-cabinet-office-community-life-infographic`. (Accessed on 03/31/2019).

[56] M. Vartak, S. Madden, A. Parameswaran, and N. Polyzotis. Seedb: automatically generating query visualizations. *Proceedings of the VLDB Endowment*, 7(13):1581–1584, 2014.

[57] Y. Wang, Z. Chen, Q. Li, X. Ma, Q. Luo, and H. Qu. Animated narrative visualization for video clickstream data. In *SIGGRAPH Asia 2016 Symposium on Visualization*, p. 11. ACM, 2016.

[58] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, and H. Qu. Infonice: Easy creation of information graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 335:1–335:12. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173909

[59] Wikipedia. Fact sheet. `https://en.wikipedia.org/wiki/Fact_sheet`, Feb 2019. [Online; accessed 29-March-2019].

[60] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.

[61] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2648–2659. ACM, 2017.

[62] H. Xia, N. Henry Riche, F. Chevalier, B. De Araujo, and D. Wigdor. Dataink: Direct and creative data-oriented drawing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 223. ACM, 2018.