# FactExplorer: Embedding-based visual analysis of data fact space

Roy G. Biv, Ed Grimley, *Member, IEEE*, and Martha Stewart

Fig. 1. In the Clouds: Vancouver from Cypress Mountain. Note that the teaser may not be wider than the abstract block.

**Abstract**—Automatically discovering interesting data patterns from multi-dimensional data has become a prevalent visual analysis approach because of it's effectiveness in conveying facts from data without requiring excessive user efforts. Existing authoring tools mostly filter out facts that perform poorly in evaluation, and only show the most distinctive ones. However, it can not conclusively determine the role of facts with poor rating by algorithm in aiding exploration, and their absence also leads to a lack of context for significant facts. To address these challenge, a multi-factor embedding approach, which can measure the similarity of data facts, is designed in this work to provide an overview of the whole facts and the context for each fact. A multi-storyline generation algorithm is also designed to organize fact sequences from multiple perspectives to support an exploratory narrative These ideas are implemented in a visual analysis system named, FactExplorer. Finally, we evaluate the proposed technique through two FactExplorer usage scenarios and a user study. Our evaluation shows that FactExplorer helps users easily and efficiently to analyze and explore the fact space.

**Index Terms**—Automatic design, data embedding, data story, human-computer interaction

## 1 Introduction

Currently, automatic insight delivery is emerging as a promising visual analysis approach in multi-dimensional data exploration. This technique can effectively reduce the effort of users in exploring data, through automatic insight discovery, well-designed layout and narrative structure, and engaging visual mapping. More and more researchers are attracted to create authoring tools for automated delivery of insights, such as datashot, data videos, and calliope. These tools liberate users from creation and allows them to focus on high-level exploration, that is friendly to ordinary audiences who are lack of domain-specific knowledge, and promotes the popularization of visual analysis tools. Tabular data, as a common information storage medium, is selected as the focus of this work. In the following discussion, we utilize data fact to represent the insight in tabular data.

Ranking the data facts according to certain evaluation metrics and selecting the top-k ones is a common fact delivery approach utilized in aforementioned works. However, this will make a large number of facts invisible. Although these facts have a poor performance in the evaluation, it is not conclusive that their role in aiding users understanding. Due to technical limitations, there are certain hard-to-find facts, as shown by the outermost dots in figure 2. As with the existing work, improving the fact detection ability is not our focus. This work focuses more on users' understanding of the fact space and whether they can effectively and smoothly locate the ones that meet their needs (dots in orange boxes in figure 2). Furthermore, another limitation of the aforementioned approach is that the contextual information of the top-k facts is also lacking, which is not beneficial for users to switch between different facts. In summary, the first challenge of this work is to support an overview of the whole fact space, as well as the context of each fact.

Data story, which assembles fact pieces into well-designed narrative structure to generate a meaningful sequence, is a more effective approach of delivering insights. Existing work has focused on researching how to construct high-quality narrative sequences. Such as, Calliope incorporates a logic-oriented Monte Carlo tree search algorithm to progressively generate data facts and organize them in a logical sequence. Storylines, which generated in previous work, generally have good performance in evaluation or rule constraints. However, such storylines usually have a single narrative perspective and a fixed mindset, and can not provide a flexible exploration. Users can only receive information in the mindset that perform well in algorithm evaluation, instead of independently exploring according to their own mindset. Exploratory narratives can make user exploration more interesting and reveal more

- Roy G. Biv is with Starbucks Research. E-mail: roy.g.biv@aol.com.
- Ed Grimley is with Grimley Widgets, Inc.. E-mail: ed.grimley@aol.com.
- Martha Stewart is with Martha Stewart Enterprises at Microsoft Research. E-mail: martha.stewart@marthastewart.com.

*Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx*

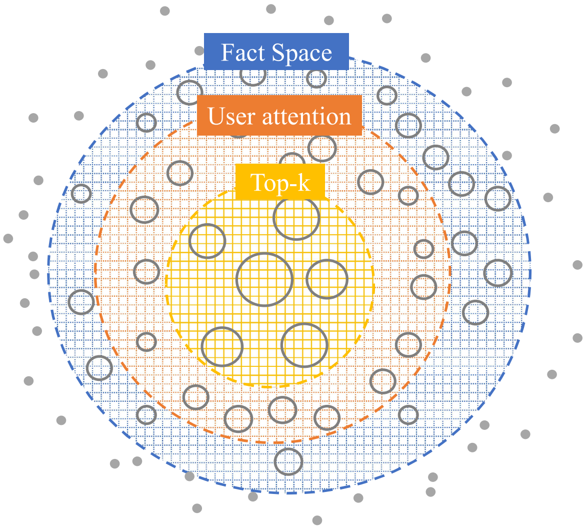

Fig. 2. The illustration of fact space. The dots in blue box are the facts automatically extracted, orange box includes the facts that may be of interest to users, and top-k facts are in yellow box

underlying patterns. Thence, the second challenge of this work is how to generate storylines from multiple perspectives, enhance the autonomy of exploration, and allow users to decide the direction of the narrative.

To address the above challenges, we propose a framework to assist in exploring the fact space, which consists of three parts: fact extraction, fact embedding, and storyline generation. In the fact embedding module, we introduce a two-factor embedding approach, which can measure the similarity of data facts, to embed facts into the fact space. More specifically, a novel variational encoder is employed to encode the visual style details of facts, and the logical relation of facts is embedded by computing the logical distance among facts based on the findings from previous work. Furthermore, we design a novel multi-storyline generation algorithm, which expands multi-perspective storylines from top-k facts respectively. A fact path search algorithm is also designed to seek a smooth and informative transition path from one fact to another. Finally, these ideas are implemented in a fact space exploration system, FactExplorer. Certain interactions are also provided in the system to support flexible editing of facts and fact storylines. The major contributions of this paper are as follows:

- A runnable system named FactExplorer is designed to automatically build a fact space from tabular data. Appropriate interaction components are also provided in the system to support flexible and efficient exploration.

- A two-factor (visual style and logical) fact embedding approach for embedding facts into the event space. A novel variational encoder model is utilized to embed visual styles, and logical embedding is achieved by computing the logical distance among facts, and finally aggregate these results to get fact embedding vector.

- An exploratory narrative approach is proposed to improve the flexibility and interest of exploration. More specifically, a multi-storyline generation algorithm is designed to generate multi-perspective storylines. A fact path search algorithm is also designed to seek a smooth and informative transition path from one fact to another.

## 2 Relate Work

### 2.1 Automatic Visualization

### 2.2 Modeling Visualization Similarity

### 2.3 Data-Driven Storytelling

## 3 The Design Of FactExplorer

In this section, we discuss the design goals and system pipeline of FactExplorer. The details of the system implementation will be described in the next section.

### 3.1 Design Goals

The main design goal of FactExplorer is to minimize users' efforts to explore and discover interesting patterns in the fact space. More specifically, we summarize the following four goals:

**G1 Ensure the high quality of data facts.** The system should ensure that the extracted facts can cover most of the original data space, ensure that each fact includes a rich amount of data, and ensure that the visual styles of the facts are diverse.

**G2 Support a semantic overview of fact space.** The system should model relationships between facts and provide an overview of the entire fact space to facilitate users to develop a holistic understanding.

**G3 Organize data facts into storylines.** The system should organize fragmented facts into informative, encoding related, and logically related storylines. In addition, system should also provide a transition path between specified facts to incorporate real-time insights from users.

**G4 Support convenient user interactions.** The system should provide certain flexible interactive components to support querying, filtering, and viewing specific facts, and to enable comprehensive editing of generated facts and storylines for users.

### 3.2 System Pipeline

To achieve these design goals, we utilize a **plane-line-point** hierarchy to present data facts[cite]. We first extract the facts (**point**) from the tabular data. These facts are then embedded into the fact space to revealed the global features of fact collection (**plane**). Finally, related facts are assembled into meaningful storylines (**line**). Figure 1 shows the pipline of the FactExplorer system, which consists of three core modules

**Fact extraction.** Common fact extraction methods are utilized to extract facts from tabular data[cite]. The procedures executed include: subspace slicing, fact enumeration, fact screening, and fact scoring[**G1**]. These processes are automatically completed by the system , user only need to select the data attributes and fact types that participate in the fact extraction[**G4**]. After this module, the underlying raw data is converted into an fact collection.

**Fact embedding.** In this module, facts are embedded into fact space to provide an overview of the fact collection. We adopt the method of multi-channel embedding, and factors closely related to the fact (visual encoding, logical, and deviation) will be embedded separately. Finally, these one-factor embeddings are aggregated into the final embedding by assigning different weights[**G2, G4**].

**Storyline generation.** In this module, we delineate the structure in the fact space by extracting storylines to bridge the connections among facts and fact clusters[**G3**]. Users can browse the trunks along the storyline, it can effectively promote users' cognition of the fact space and accelerate the exploration process. The details of storyline extraction will be discussed in the next section.

## 4 FactExplorer System

## 5 Casestudy

## 6 Userstudy

## 7 Discussion

## 8 Conslusion

In this paper, we introduce FactExplorer, a system designed to help users efficiently and conveniently explore and analyze fact space. In this system, entire facts are automatically extracted from tabular data.

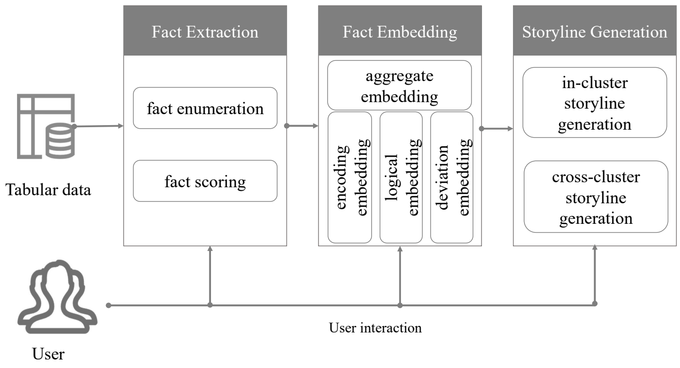

Fig. 3. The processing pipeline of FactExplorer. First, FactExplorer automatically extracts facts from the tabular data and scores these facts. Next, these facts will be embedded into fact space from three perspectives: visual encoding, logical, and deviation. Finally, storylines will be extracted after clustering the facts.

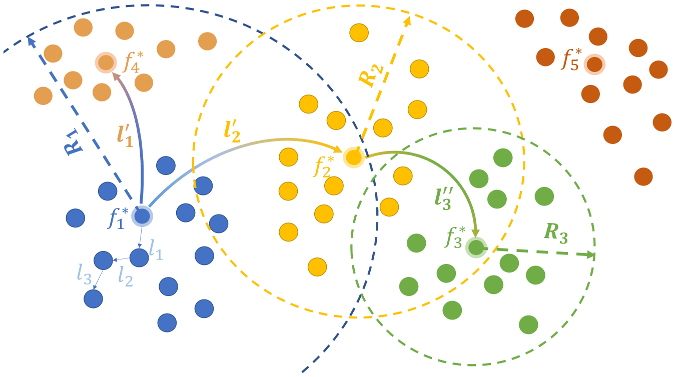

Fig. 4. The processing pipeline of FactExplorer. First, FactExplorer automatically extracts facts from the tabular data and scores these facts. Next, these facts will be embedded into fact space from three perspectives: visual encoding, logical, and deviation. Finally, storylines will be extracted after clustering the facts.

A two-factor (visual style and logic) fact embedding approach is introduced to embed facts into the fact space, which provides an overview of all facts and the context of each fact. A multi-storyline generation algorithm is also designed to generate multiple perspectives storylines. The whole facts are organized well to promote exploration and deepen the impression of the fact space on the user. Certain interactive components are also implemented to support users to flexibly edit facts and storylines. These techniques proposed in this work is evaluated through two case studies and a user study. Finally, we discuss several limitations of the current system, which will be addressed in the future.