Explorable Visual Analytics

Knowledge Discovery in Large and High–Dimensional Data

Saman Amirpour Amraii
CREATE Lab, Robotics Institute
Carnegie Mellon University
and
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA
samirpou@cs.cmu.edu

Michael Lewis
School of Information Sciences
and
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA
ml@sis.pitt.edu

Randy Sargent
CREATE Lab, Robotics Institute
Carnegie Mellon University
Pittsburgh, PA
randy.sargent@cs.cmu.edu

Illah Nourbakhsh
CREATE Lab, Robotics Institute
Carnegie Mellon University
Pittsburgh, PA
illah@cs.cmu.edu

ABSTRACT

Visual analytic tools are invaluable in the process of knowledge discovery. They let us explore datasets intuitively using our eyes. Yet their reliance on human cognitive abilities forces them to be highly interactive. The interactive nature of visual analytic systems is facing new challenges with the emergence of big data. Massive data sizes are pushing against the boundaries of current visualization capabilities. Also the emergence of complex datasets is asking for new ways of navigation in the high–dimensional space. EVA (Explorable Visual Analytics) is an in-progress work for developing a web–based tool for visual exploration of large and complex datasets. EVA tries to handle large data sizes through utilizing local GPU resources and a novel client/server architecture. It also provides an easy navigation mechanism for exploring high–dimensional data. This paper presents our experiments in knowledge discovery with EVA, using US Census employment dataset as our testbed. We hope our experiences result in designing guidelines and techniques for the future visual analytic tools of the big data era.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: General; H.4 [**Information Systems Applications**]: General; H.1.2 [**User/Machine Systems**]: Human Information Processing

Keywords

visual analytics, data exploration, visualization, dimension reduction, data mining

1. SENSEMAKING AND BIG DATA

A data explosion is happening, promising invaluable opportunities in scientific and technological progress, yet this vast potential relies not only on our ability to collect and access this data, but also on us being able to understand it. Despite this fact, it seems that knowledge discovery from raw data has not still reached its full power. We are producing much more data than we can explore leading to massive amounts of untapped data waiting for future discoveries. But what makes knowledge discovery hard?

There are in general two major approaches to do knowledge discovery: either we use mathematical methods (e.g. machine learning) or we use human judgment by directly looking at data (e.g. visual analytics). Mathematical methods are profoundly powerful tools yet they still rely on human intuition for the following reasons. First, mathematical methods are a collection of tools. Finding the right tool, using the right models, tuning its parameters and feeding the right feature space into it are often done by human experts. Second, mathematical methods are not context-aware. It is this extra knowledge that usually leads human experts to find the right features or ask the right questions. Third, mathematical methods are not good at providing explanations. A famous example is a Neural Network which is great at finding patterns but does not provide any explanation for how does it find it. And last but not least, mathematical methods are best practiced by mathematicians and computer scientists while most data experts are from other fields, not proficient enough in using these tools on their own data. These facts force us to keep the human in the knowledge discovery loop. Therefore the important question to answer becomes how do people make sense of the data?

Jerome Bruner [10] argued that children posses three modes of representation, (1) interactive, (2) visual and (3) symbolic, and they use these modes to understand a new object or system. In other words we act, we see, and we ask to make sense of something new. For example, upon encountering a new object, the child uses her hands to play with the object, looks at it to find out what happens when she touches it and in the more abstract level she may even ask a question to acquire new sources of knowledge. This process is then repeated until the child amasses enough knowledge about the object until she can build a reliable mental model

representing it (Figure 1). It can be argued that even scientists upon facing a new system, be it a simple object or a complex dataset, go through the same process in order to build a mental model of it. This multi-modal exploration of data is an essential step in building the right intuition and plays a significant role in choosing and applying the right rigorous methods in the following steps. For example in a classification problem using machine learning tools, data scientists usually first draw the raw data and do some basic interactions with the data (e.g. scaling). This step provides the initial guidance which then translates into choosing the right model/machine learning tool. This process of building a mental model of the data is called sensemaking. It is only after acquiring this intuition that we can apply our mathematical tools in their full power and extract meaning and knowledge out of the raw data. It is worth mentioning that the model presented in Figure 1 has a hidden assumption: the feedback we see from interacting with an object should be almost instantaneous. If we devise a new theory and test it on the object/data but receive our answer after several hours, we will not be able to effectively build a mental model as we lose our train of thought after only a few seconds. Therefore query latency can have a major impact on the sensemaking process.

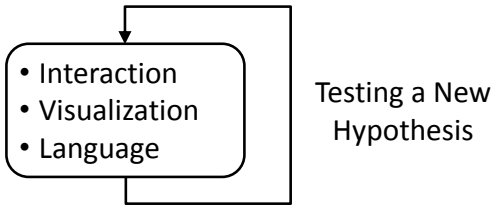


Figure 1: Multi-Modal Exploration: how people understand an object or a system.

Up until now, this sensemaking process has been done intuitively, usually through conventional visualization techniques (e.g. plotting). But the emergence of vast and high dimensional datasets is raising challenging issues not addressable by our current data analytic approaches. For example, current datasets are getting so large that asking even the simplest questions from them may take hours or days of computation. Even after accessing the data, usual visualization techniques may not work due to issues like overplotting. Furthermore, it is not even possible to fully visualize datasets that have hundreds or thousands of dimensions. Another issue is the lack of hypotheses for analyzing the data. Due to decreasing trend of storage prices, we are acquiring and storing an ever increasing amount of data without knowing which portions of that might be useful in a future analysis. Facing with these datasets, even finding the right questions becomes a part of data exploration process.

EVA (Explorable Visual Analytics) is an effort to seek for design guidelines and analytic tools which are capable of visualizing, exploring and analyzing large and complex datasets. Our hope is to promote a set of practices which lead to faster and easier data driven knowledge discovery. To achieve this goal, EVA attempts to facilitate hypothesis generation and query refinement through a series of consecutive multi-modal exploration loops. We also seek new computational techniques which can scale appropriately with the data size and complexity.

Section 2 gives some examples of how researchers are approaching large and complex datasets and what are the challenges they are faced with. In Section 3 we introduce EVA and give an example of using EVA for knowledge discovery on real data. Section 4 discusses some of the lessons we have learned so far in exploratory data analysis and suggests some possible approaches that might expand our ability to do knowledge discovery in large and high–dimensional data.

2. NEW APPROACHES IN VISUAL ANALYTICS

2.1 Knowledge Discovery, Visualization, and Big Data

The process of knowledge discovery is a fundamental aspect of science in general. A rich model for describing this process is presented in [22]. The authors argue that scientists navigate in a four dimensional space in order to extract meaning from their observations. The first dimension in this paradigm is called *data representation*. This is where an abstract representation of data is being formed from a set of features. The second dimension is *hypothesis space*. Here, the scientist generates new assumptions on the possible causal relationships. Then she moves to the third dimension of *experimental space* in order to test those hypotheses. It should be noted that the experiments themselves live in an experimental framework that defines the boundaries of valid experiments and expectable outcomes. Therefore the fourth dimension is *experimental paradigm space* where the scientist can choose a completely different class of experiments for her task. In visual analytics tools, a knowledge discovery process can be modeled based on the first three dimensions. A specific visualization is an example of the data representation space. The ability to interact with the data is happening in the experimental space. Finally, the visualization/exploration choices form a series of decisions in the hypothesis space. By using visual space for doing data representation, we have a tangible and more direct connection to the actual data. By forming a visual query, we actually form a hypothesis in our mind and when we do a visual search, we are experimenting with data in order to confirm or reject our hypothesis. This process has been explained in literature in various ways. Fry [13] presents this process in seven steps. First, we should acquire the data. Then we have to parse it and make it machine readable. This is then followed by filtering in which we select a subset of data that is relevant to our work. We then mine for useful information which usually means some sort of mathematical transformation. The results are then represented in an initial visualization. Then comes the refinement and finally interaction steps in which we explore the visualization and improve it by redoing the previous steps until we extract or discover the desired knowledge. A more general perspective on knowledge discovery is pursued in the field of visual analytics [17, 18]. The goal of visual analytics is to illuminate the way people understand data and then turn it into an algorithmic discipline which benefits from both the power of automated processing techniques and the capabilities of humans in discerning and analyzing visual patterns.

Visualization research has been successful in turning raw

data into meaningful visual presentations yet the general perspective of the field does not differentiate between small and simple datasets with large and complex ones. This paradigm is changing as visualization experts face with unforeseen challenges unique to the big data era. For example, as the size of the dataset grows, the responsiveness of traditional visualization systems drops until it is no longer interactive. In addition to the scalability issue, visualizing and understanding complex datasets with hundreds of features is very challenging. These issues have opened new lines of research which often try to change the underlying visualization approach in order to overcome these limitations.

Fekete [11] provides a nice summary of the challenges faced by current visual analytics tools and the paradigm shifts required to overcome these issues. He argues that as data sizes are getting larger, query latency is posing a serious problem. If the system does not provide an answer to a particular question within a few seconds, the analyst may forget her question and not benefit from the answer. He suggests that by shifting from conventional *accurate but slow* analytic tools toward *inaccurate but fast* paradigms, we can overcome the query latency issue when we are dealing with large and complex datasets. It is interesting to see this mindset has started to gain momentum for example in [12] where authors use partial but fast querying techniques to analyze very large databases. Fekete argues that another issue in current analytic systems is the lack of feedback and steering. When a user sends a query, the system starts producing a report. This process cannot be interrupted by the user. She should wait until a query–response "episode" is finished and then start asking a new question. We need to be able to steer the system toward our desired answer as it is analyzing the data. For example, we should be able to play with the parameters of our question or navigate through the data space and ask for finer and more accurate answers for a subspace of a large dataset. Interactivity is another important aspect of visualization systems. For example when a user tries to rotate a 3D object, the operation should happen instantaneously, usually within a 100 ms. This poses a great technical challenge in front of current visualization tools which their frame rate usually drops considerably fast even with modest data sizes of tens of thousands [8]. To summarize these issues we should expect new visual analytic tools which provide responsive multi-modal exploration mechanisms to support sensemaking, provide novel steering abilities to navigate large and high dimensional data, focus on small query latency even in cost of inaccurate answers, and provide non–episodic interactions where a user can modify her query while it is being processed.

2.2 Dealing with Size: Screen–Aware Tools

While data sizes are growing without any foreseeable limit, our cognitive abilities are fairly limited. We probably can only perceive a few million features or even less [11]. As it is us who are the actual bottlenecks in understanding visualizations of large data, a new class of solutions are emerging which focus on the output instead of input. These screen–aware (or output–sensitive [7]) tools use various data abstractions to reduce the size of presented information and avoid analyzing portions of data that are out of the scope of screen. They then use interactive and exploratory mechanisms to help the user navigate through the visualization and understand the data better. These tools are based on the assumption that we do not care for fine details in a big data visualization. A data analyst who looks at a visualization of millions of points is often only interested in the general shape of the visualization; the exact location of a single pixel is usually not important to her. On the other hand, she would prefer to be to able to interact (e.g. zoom, pan) with this visualization in a fluid manner in order to form a better mental model of the overall characteristics of the visualization.

One class of screen–aware solutions are called on-demand processing [7]. They only draw those things that would be visible. For example if the visual representation of a data point is smaller than a pixel or outside of the scope of screen, there is no need to process it. One of the most common techniques used in this class is semantic zoom [15]. In contrast to geometric zoom which redraws all pixels upon zooming, the semantic zoom provides more detail when zooming in and hides some of the detail when zooming out. This can result in tremendous conservation in processing and communication load and therefore it has been used extensively in visualization systems, such as online maps, etc. Semantic zooming is usually used in conjunction with multi–resolution data structures. The basic idea of a multi–resolution data structure is to pre–compute the visible data for each zoom level. As an example, this technique has been used in GigaPan and TimeMachine [21] to present massive high resolution images and videos in an interactive setup which allows zooming on any desirable part of the video while keeping the communication and processing under a manageable limit. Another example of multi-resolution data structures is presented in [19]. Here, the data structure is more complicated and has many dimensions but the fundamental idea is to aggregate over different features and pre–compute these values for several desirable zoom levels. This can then be used to interactively visualize multi–dimensional datasets with over billions of data points.

Another class of data abstraction solutions go further than only showing visible things and instead focus on only showing the important things. In one popular set of techniques, it is the computer/algorithm itself that decides on what is considered important. These techniques are usually pursued as clustering, sampling, aggregation, filtering, ... where the algorithm either combines several data points or selects a smaller subset of them and only processes those smaller representations. An excellent example in this class is presented in [12]. Here, when the user sends a query to visualize some aspects of the data, the algorithm will randomly select a small sample of data points and then visualizes only those points. It also presents some confidence intervals around each visualized object in order to help the analyst in understanding the error range of the incomplete visualization. With more time, the system grabs more data points and increases the accuracy of its visualization (also decreasing the confidence intervals). This system provides a very promising approach to visualization of large datasets by using both aggregation (in the form of queries) and sampling while in the same time it provides an inaccurate but responsive experience.

While using computer algorithms in choosing the important aspects of data results in highly scalable visualization systems, it is not obvious whether the algorithms will always choose the correct abstraction. This is why another class of solutions insert the user in the loop and ask her

to provide feedback on what is important and what should be visualized. The most common type of these techniques is query–based visualization [7]. Here, the user creates a query or search term and reduces the amount of data to a smaller subset which is then used for the final visualization. For example, Beyer et al. [6] present a query–based system for visualizing neurons in a terabyte scale dataset. The user selects regions and neurons of interest and then the system presents neighboring neurons and their relationship in an interactive setting. Another technique used for finding the correct abstraction is steering. Here, the user guides the visualization system in a two–way mechanism — the system provides an initial visualization and then the user refines it by steering the system toward her regions of interest and then the process repeats. An excellent example in this area is presented in [24] where the system uses a dimension reduction algorithm to present a large and high–dimensional dataset but instead of keeping users as passive observers, it actively engages them: the system gradually shows more points in the projected visual space while the user can steer the system towards her desirable regions. This allows the system to only focus on projecting data points in that region, therefore avoiding unnecessary calculations.

2.3 Dealing with Complexity: Human–Assisted Navigation

High–dimensional datasets are inherently hard to visualize (think of a 4-D cube) yet current big data trend is not only expanding in data size, but also in data complexity. Most high–dimensional visualization systems focus on some sort of dimension reduction. One class of these techniques are human–assisted methods which benefit from human feedback in their dimension reduction process [23]. These methods are often heavily interactive as it seams interacting with a visualization can somehow compensate for our inability to perceive high–dimensional space. Human–assisted dimension reduction usually starts with a projection algorithm that has some parametric values. The role of the human is to fiddle with these parameters until the final projection is more suited to her needs. This approach adds an extra layer of sophistication to the visualization system and extends its capabilities in generating meaningful projections of the complex data. It also has the added benefit of engaging the operator in the visualization process. This can both increase the awareness of the analyst plus through her feedbacks, the system can save valuable computational resources. One of the early examples of human–assisted methods in visualizing high–dimensional datasets is Grand Tour [23]. In a Grand Tour, the analyst can choose any arbitrary nonorthogonal projection of the data. This can reveal features that may remain hidden in the conventional orthogonal projections used in some other approaches such as parallel coordinate plots. Another early example of human–assisted methods is presented in [20]. This system has been used to visualize documents in a multi–dimensional setting. Each dimension is represented as a point in the visualization plane and documents would attract/repel to these points based on their similarity to each dimension. Also, by moving these feature points, the user can see how each document reacts. This helps in clustering documents into similar groups in their complex environment.

Steering is one of the recent techniques in human–assisted approaches. Williams and Munzner [24] introduce a navigation mechanism in which the operator steers the system toward the desired subspace of the original dataset. The projection algorithm is then focused on this area, avoiding unnecessary computations on the rest of the dataset. Also, by actively engaging the user in the process of complexity reduction, the operator builds a better mental model of the data. Ingram et al. [16] provide a different mechanism for engaging the user. Here, the system provides a collection of different dimension reduction algorithms and provides tools for tuning their parameters. The analyst can combine these algorithms together until she finds a desirable low–dimensional representation of the data. This is especially beneficial when the user is not an expert in machine learning and dimension reduction techniques. The authors also extensively use the idea of navigation and landmarks. Different levels of global and local navigation improve the exploration ability of the visualization tool while landmarks help the user to find interesting projections of the data. In a similar fashion, Gratzl et al. [14] introduce a tool for exploring rank–based data. Here, the projection algorithm is a simple weighted linear combination of dimensions, but the user has much more power on selecting each weight and the overall combination rules. The tool is also highly interactive, making it easy to create new hypotheses and then testing them through a simple drag and drop process.

2.4 Next Steps in Visual Analytics

The solutions discussed here are reshaping the conventional visualization paradigm. They put priority over speed and responsiveness even if it results in reduced accuracy, presenting a subset of data or presenting an abstract and compressed version of it. These solutions are also often screen–aware, which means their computational complexity is usually dependent on the screen size rather that data size. This makes them great candidates for emerging visual analytic tools that are capable of scaling with growing data sizes. Future visual analytic systems should also offer non–episodic interaction with the data. In this type of interaction, the user can constantly fiddle with the parameters of the query while the system instantly demonstrates new visualizations. This means that when the system receives a new input from the user, it would not wait until it completes the previous data analysis action. Instead, it adjusts its results to the new query. This interactive query building is essential in forming and improving our hypotheses about the data and as Fisher et al. [12] show in a case study, this can be highly beneficial for data analysts. Non–episodic interaction can be useful because in knowledge discovery we often need to ask many questions and perform multiple iterations on our hypotheses before we can form the right questions. A data analyst seldom asks only one question. She should form many assumptions and refine those assumptions through consecutive visualizations until she can find the answers she is looking for. The ability to change query parameters on the fly should be accompanied by fast response times from the system. Our memory is very limited, specially when dealing with vast quantities of visual information. Short query latency and intuitive navigation mechanism can help us go back and forth between several visualizations and look at the data from multiple perspectives, therefore increasing our chance for finding meaningful patterns.

3. EVA PROTOTYPE

Explorable Visual Analytics (EVA [5]) is a visualization system prototype. It has been developed to address the challenges arising in dealing with large and complex datasets. The main philosophy behind designing EVA is to improve hypothesis generation, both in quality and quantity. EVA tries to provide easy to use and intuitive navigation mechanisms. Through them, the user can easily navigate in a large space of data objects. It also helps the analyst to look at the multi–dimensional data from multiple perspectives, hence giving her a better chance for finding interesting phenomena in the data. In general, the interactive nature of EVA is critical in sense making and creating a mental model of the data. Also, EVA is designed to be responsive as it is beneficial to minimize the time between generating a question and testing it. There is an important period between when an analyst forms a question in her mind until she can see the relevant visualization to test that hypothesis. If it takes too long (e.g. even more than 10 seconds), the analyst may lose her train of thought. This is mainly due to our limited working memory. EVA minimizes this delay period and therefore lets the analyst to instantaneously test her new ideas. This is in turn helpful in generating more questions. In conclusion, EVA provides a simple navigation mechanism for studying a large and complex dataset through visual inquiries. It also has short processing time in order to avoid any delay between receiving a query from the user and visualizing it. EVA is also designed to provide a high resolution visualization, as richness of details is an important factor in doing knowledge discovery. All of these aspects helps the user to start with a relatively small set of assumptions, test them, generate new questions, refine them, and gradually build a better model of the data, which then results in finding new and meaningful patterns.

Based on the knowledge discovery framework presented in Section 2.1, EVA is composed of three major conceptual sections. In the *data representation* section, EVA provides a 5 dimensional visual space consisting of spatial coordinates (X, Y, Z), color and visibility period (named as Time). Each data point can be assigned to an instance of this visual space. In the *hypothesis space*, the user can use a simple one-to-one mapping function from data space to visual space. It is also possible to scale data values to better fit them in the visual space. In the *experimental space*, EVA provides various tools for interacting with and manipulating the visualization in order to do a visual search and find interesting patterns. These mechanisms include tools and techniques such as zoom, pan, rotation, choosing color palette, scaling, camera features, external visual aids such as Google Maps and also some textual helpers such as an information panel.

EVA is a web–based tool developed at CMU's CREATE Lab [1]. It is a part of Explorables [2] collaborative which consists of various projects aiming at interactive visual representations of large datasets. EVA is accessible from http://eva.cmucreatelab.org. It is written in JavaScript and HTML. It uses a selection of color palettes presented in Color Brewer [9]. It also uses the WebGL–based Three.js [4] library for its graphical engine. Choosing web–based technologies has been helpful in sharing EVA with other experts and incorporating their suggestions during development phase. EVA fully utilizes the GPU and RAM in order to visualize large datasets without sacrificing its response time. Currently, it can handle data sizes of up to a few million points consisting of tens of dimensions. Figure 2 shows a screenshot of EVA in a browser.

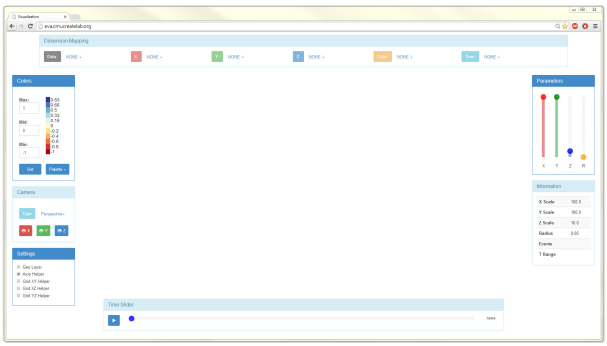


Figure 2: EVA's main screen.

Choosing the right dataset for EVA has been based on several factors. First, we wanted a dataset large enough to be beyond the processing capacity of usual visualization tools, yet not too large to complicate the development of our first prototype. As current tools are usually limited to visualizing a few tens of thousands of objects, we chose a limit around a few millions of points for our dataset. The second factor in choosing a dataset is its complexity. A dataset with a few dimensions (say 4) can be visualized completely using spatial dimensions and color. On the other hand, manually selecting and navigating through hundreds or more dimensions is tedious and very complicated. Therefore, we limited the datasets dimension cardinality to tens of dimensions. It is also important to chose a meaningful dataset acquired from real world measurements. This can lead to relevant and useful knowledge discovery. Also, the analyst can benefit from her expertise in the contextual information accompanying that dataset. Finally, the data should have some meaningful representation in the spatial space, otherwise a purely visual exploration may not be as beneficial.

Based on these characteristics, we chose United States Census Longitudinal Employer–Household Dynamics (LEHD [3]) dataset. This dataset provides information on employers and employees across country. This information includes categories such as employees earning, age, ethnicity, education level, etc[1]. It is aggregated over census blocks which are small geographical regions usually equivalent to a city block. Also, the data is produced yearly, therefore providing enough details both on the spatial and temporal levels. This dataset is being used by a wide span of scientists and analysts from economists to urban researchers. As such, it can be used with a rich set of contextual knowledge from various fields and therefore it can be a good candidate for doing meaningful knowledge discoveries. Currently, the visualization tools dedicated to LEHD are limited and they often work on aggregations of the original data, hence they do not visualize it with fine details. The LEHD dataset in its entirety is very large (more than a 100GB). Therefore we have limited our work to the state of Pennsylvania[2]. This

[1]Details of LEHD data structure is available at http://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.0.pdf

[2]Particularly to the residence–based workforce information subsection of LEHD. For Pennsylvania, this dataset is around 300MB.

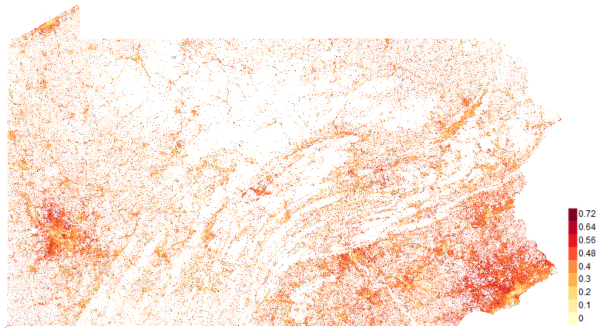


Figure 3: Earnings more than $3333 per month for Pennsylvania. Each dot represents the center point of the corresponding census block. Red areas show regions with a higher percentage of residents in high–end income range. The color palette on the right shows the minimum percentage of employees with the aforementioned income level in each census block.

subsection of LEHD has around 2.8M data entries and 44 dimensions. Next, we will go through some examples of using EVA on LEHD for understanding the data better and then doing discoveries as we interact with the data.

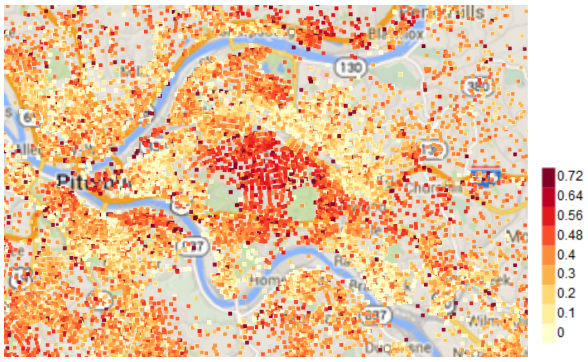


Figure 4: Earnings more than $3333 per month for Pittsburgh.

The first example is a simple visualization of income (Figure 3). Each dot represents one instance of data. The longitude dimension of each data instance is assigned to the visual dimension X (the horizontal orientation of the figure). The latitude dimension is assigned to the visual dimension Y (the vertical orientation of the figure)[3]. The visual dimension of color represents the ratio between the number of jobs with an income of $3333 or more per month with the total number of jobs. Therefore a pixel with bright red color shows a relatively wealthy neighborhood while a pixel with yellow color shows a poorer area. In general, there are 2.8 million pixels in the visualization. From this visual representation, it is easy to locate the major population poles of the state, such as Philadelphia on the bottom right corner or Pittsburgh on the left side. It is also possible to distinguish the major geological features of the area such as the distinctive Appalachian Mountains in the middle of the map. The other important observation is the non–uniform distribution of wealth throughout the state. Most of high–end income earning neighborhoods are concentrated in the suburbs of Philadelphia and Pittsburgh while the regions in the middle are usually less populated and often have a lower amount of income. Figure 4 shows a zoomed in version of Figure 3, focusing on Pittsburgh. This picture also includes a Google Map helper in the background. This layer can be helpful in distinguishing the exact location of each census block. Based on this map, the main wealthy neighborhoods are seen in the middle of the picture, where the University of Pittsburgh and Carnegie Mellon University are located.

[3]The latitude and longitude measures represent the central location of the corresponding census block.

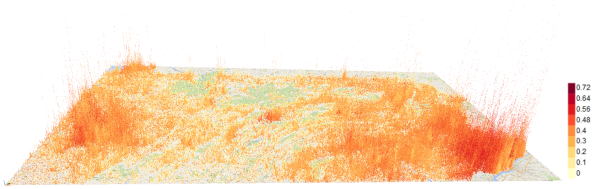
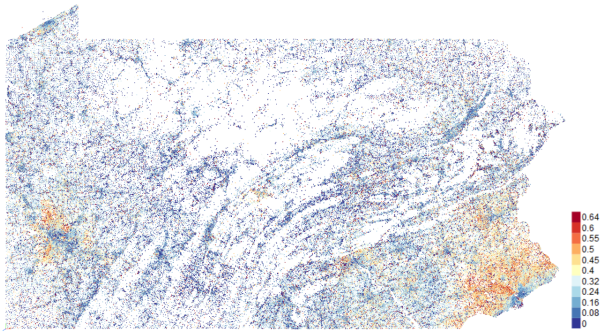


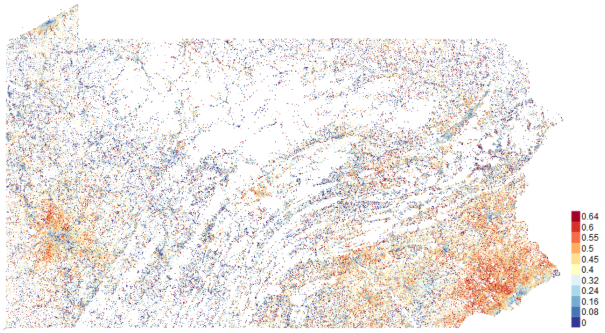
Figure 5: Earnings more than $3333 per month (as color) combined with total number of jobs (as elevation).

In Figure 5, we have utilized all the 3 spatial dimensions. Here, besides assigning longitude and latitude to X and Y, we have assigned total number of jobs in each location to dimension Z. By rotating the visualization, the user can look at the high–income levels (as color) and total number of jobs (as elevation) at the same time. Through this representation, it is again easy to find the major population hubs. Also, it is more evident that there is a more complex relationship between income level and number of jobs. For example by looking at Philadelphia at the bottom right corner, we can see areas of high income (red) and low income (yellow) with almost the same number of jobs adjacent to each other. Another interesting example is State College, home of Pennsylvania State University, located at the center of the map. This small city has a relatively low number of jobs, but the color of those jobs shows a high–income region, representative of its higher education employment sector. It should be noted that most of the visual objects in a point cloud are obscuring each other, therefore it is essential to have interactive capabilities. Through rotation, zooming and panning, the user has a much better chance of understanding the general outline of the visual space.

The last visual dimension available in EVA is Time. By assigning a data dimension to time, we can create an animation and control it through the bottom slider. Figure 6 shows the high–end income range percentages over a course of 10 years. As it is evident from comparing Figure 6(a) to Figure 6(b), the percentage of people with higher incomes is increasing over the decade. This can be due to the inflation in income or a real increase in the overall earnings. The time slider plays an important role in revealing this pattern as the user should go back and forth in time multiple times to better perceive the gradual change in earnings. Again, the interactive nature of visualization is vital in the knowledge discovery step. The same data is represented in a different view in Figure 7. Here, instead of the usual assignment of years to Time dimension, we have assigned it to Z. This results in a series of planes dissecting the data accord-



(a) 2002



(b) 2011

Figure 6: Earnings more than $3333 per month in years 2002 and 2011.

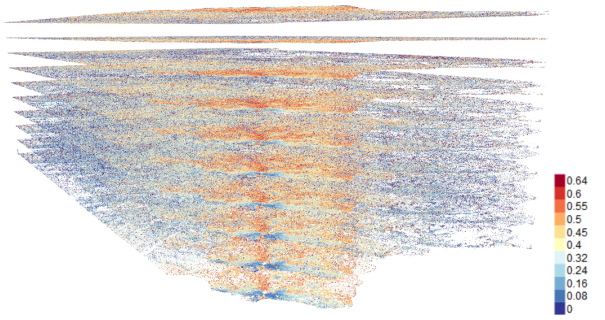
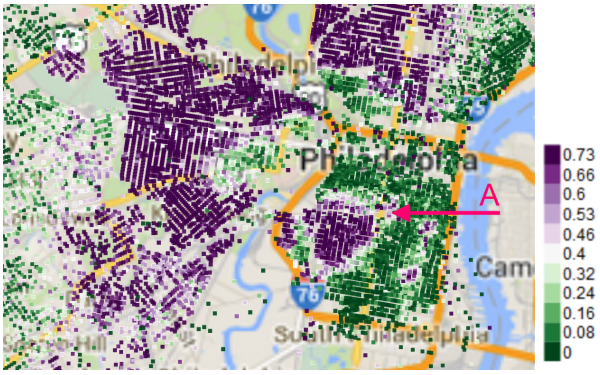


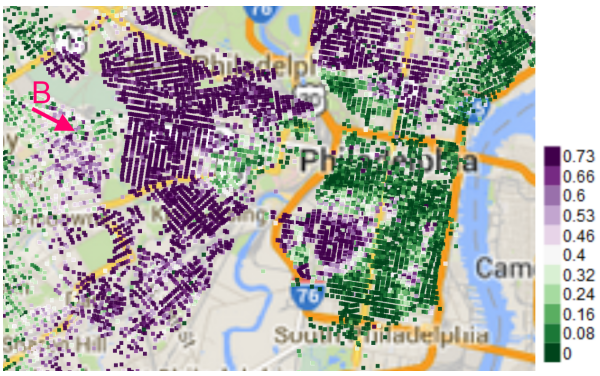
Figure 7: Earnings more than $3333 per month. The year dimension from the data is assigned to the Z dimension in the visual space.

ing to their year. This is useful for looking at the general trend. For example, the region in the front of the picture in Figure 7 is Philadelphia. We can see the lower layer (corresponding to year 2002) has more blue dots (corresponding to poor neighborhoods). As we go up in the layers we are going forward in time and we can see the shrinking of blue regions and the growth of higher–income neighborhoods.

Figure 8 looks at the distribution of races in the city of Philadelphia over the course of three years (from 2009 to 2011). The green regions represent neighborhoods with a majority of Whites while purple regions show neighborhoods with a majority of workforce from African American community. The first observation is the segregation between these two communities. Neighborhoods are mostly dominated by



(a) 2009



(b) 2011

Figure 8: Distribution of employees based on their race. Purple areas represent neighborhoods with a majority of African American workforce while the green areas represent neighborhoods with a majority of Whites. (a) shows this distribution in year 2009 and (b) is for year 2011.

only one race while in between there are some small border neighborhoods that accommodate a more balanced mixture of both races. The other observation is the relatively fast shifts in the population proportions of some border neighborhoods within a course of a few years. For example, the region marked as **A** in Figure 8(a) shows an area that is mostly composed of African Americans in 2009. But as we go forward in time to year 2011 (Figure 8(b)), this area becomes a more mixed race neighborhood. The opposite phenomena is happening in region **B** where it is changing from a mixed community to a more single–race neighborhood. During some informal discussion with a Philadelphia resident, he hypothesized that this population shift may be related to a new wave of African immigrants settling in the west side of the city.

The next example shows an accidental discovery. Here, the exploration was not driven by a hypothesis. Instead, it was the exploratory nature of the tool that led to an unexpected visualization. This later resulted in formation of new hypotheses. When working with geolocated data such as LEHD, it is common to visualize the data on a map. Figure 9 shows a visualization of LEHD data in an effort to view it outside of a geo–spatial representation. Here, each dot corresponds to one census block (i.e. neighborhood) on

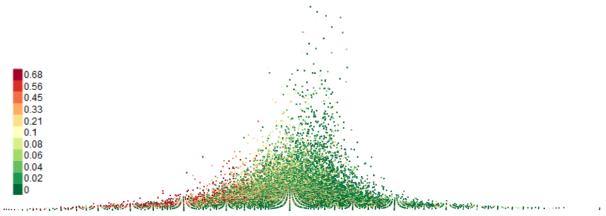


Figure 9: The relationship between race, gender, and total number of jobs. The dots on the right–hand side represent neighborhoods where a majority of workforce are men. The dots on the left–hand side are areas where the majority of working people are women. The elevation shows the relative total number of jobs. The color shows the percentage of African Americans in that neighborhood (red shows higher percentage of African Americans in that census block).

the map. The number of jobs for males has been assigned to the X dimension and the number of jobs for females has been assigned to Y dimension. Furthermore, the total number of jobs in each neighborhood has been assigned to the Z dimension. Viewing the final visualization from a perpendicular angel, we come up with Figure 9 where a dot on the right–hand side represents a neighborhood with a higher percentage of workforce being male, while a dot on the left–hand side shows a region with a higher percentage of females in the workforce. The elevation shows the total number of jobs. As it can be expected, most of the neighborhoods are located in the middle, with an almost 50–50 percent distribution of jobs between men and women. But the unexpected feature of this visualization is the one–sided distribution of red dots. Here, we have assigned number of jobs for African Americans to the Color dimension. Therefore the red dots show neighborhoods with a majority of workforce from African American community. Seeing that most of these dots are on the female side of the graph we can hypothesize that either there is a high unemployment rate among African American men or that they are working in areas with a majority of workforce from other races, hence their presence is not visible. In either case, the exploratory nature of EVA plus the ability of going through many visualizations in a short amount of time was crucial in creating this visualization and therefore forming new hypotheses about the nature of the data. It can be imagined that even randomly going through several different projections of the data can reveal some interesting patterns that are not evident in the first place, due to the lack of initial hypotheses in the mind of the analyst.

4. DISCUSSION

We can summarize EVA's contributions in three aspects: high resolution, explorability, responsiveness. High resolution is the ability of EVA to show as many data points as possible on a screen without aggregating them into overall summaries. The aggregation technique is used in many tools to improve their ability in working with larger datasets, but it also reduces the clarity of final picture and hides the fine details of the data. Knowledge discovery can be very dependable on the amount of detail a user can see. In the explorability aspect, EVA provides usual interactive techniques (e.g. zoom, pan, etc.) plus easy navigation between multiple projections of data through its dimension assignment tool. Our initial experiments showed that the ability of viewing data from multiple perspectives is crucial in understanding the data and finding the "wow" moments where the analyst observes some unexpected pattern. These moments usually lead to deeper investigations, new hypothesis generation, and sometimes to new discoveries. Finally, the responsiveness aspect of EVA fully utilizes its other features. Knowledge discovery is a memory intensive process. The analyst should form a series of assumptions and questions in her mind, and then create a series of visualizations, looking at one characteristic of the data in each step. It is important to remember all of these steps and their possible interpretations. If there is a long waiting period between each two step, the user can easily forget her previous observations and hence the general knowledge discovery process will be interrupted. EVA is designed from the ground up to address this issue by fully utilizing local computing resources available in order to make fast and smooth transitions from one visualization to the other. This is a fundamental feature in data exploration, specially when data size and complexity grows.

It is worth noting that EVA should be used in conjunction with a statistical tool. The main purpose of EVA is to facilitate hypothesis generation. It will also show visual representations of the data so the analyst can perform an initial test for each hypothesis, but coming up with a final accurate and reliable answer is the job of a statistical tool. Another important note about EVA is the role of experts in shaping it. From its inception, EVA has benefited from many experts. The choice of data, its visual characteristics (such as color palettes used), ... has been formed through many joint sessions with analysts from various backgrounds. Their realtime feedback while working with their own data on EVA has also been tremendously helpful in recognizing EVA's capabilities as well as its limitations. This collaboration would remain an ongoing part of EVA during the future expansions.

We are going to expand EVA in two major aspects: scaling and navigation in the action space. Currently, EVA downloads the full dataset into the local memory. In this way it can fully utilize clients local resources such as GPU and RAM. But this approach is limited to moderate data sizes of a few million points. Larger datasets take a long amount of time to download and they often cannot be fitted to local memory. Therefore, in the future EVA should support a client/server architecture which actively limits data transmission based on the screen resolution and user needs. This screen–aware method would not be accurate and complete, but can be scaled for large datasets. Another addition to EVA is a history function. When users explore a dataset they generate many different visualizations and sometimes they need to compare several views together in order to form a better mental model. A history function can help them navigate in their action space. This can also augment users working memory and improve the quality of their knowledge discovery.

5. ACKNOWLEDGMENTS

This project has been funded by Google.

6. REFERENCES

[1] CREATE Lab. http://www.cmucreatelab.org/, 2014.
[2] CREATE's Explorables. http://explorables.cmucreatelab.org/, 2014.
[3] Longitudinal Employer-Household Dynamics. http://lehd.ces.census.gov/, 2014.
[4] three.js, JavaScript 3D Library. http://threejs.org/, 2014.
[5] S. Amirpour Amraii. Explorable Visual Analytics. http://eva.cmucreatelab.org/, 2014.
[6] J. Beyer, A. Al-Awami, N. Kasthuri, J. W. Lichtman, H. Pfister, and M. Hadwiger. ConnectomeExplorer: query-guided visual analysis of large volumetric neuroscience data. *IEEE transactions on visualization and computer graphics*, 19(12):2868–2877, Dec. 2013.
[7] J. Beyer, M. Hadwiger, and H. Pfister. A survey of GPU-based large-scale volume visualization. In *Eurographics Conference on Visualization (EuroVis)*, page to appear, Swansea, UK, 2014.
[8] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.
[9] C. Brewer and M. Harrower. Color brewer 2.0. http://www.colorbrewer2.org, 2012.
[10] J. S. Bruner. The course of cognitive growth. *American Psychologist*, 19(1):1–15, 1964.
[11] J.-D. Fekete. Visual analytics infrastructures: From data management to exploration. *Computer*, 46(7):22–29, July 2013.
[12] D. Fisher, I. Popov, S. Drucker, and m. schraefel. Trust me, i'm partially right: Incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1673–1682, New York, NY, USA, 2012. ACM.
[13] B. J. Fry. *Computational information design*. Thesis, Massachusetts Institute of Technology, 2004. Thesis (Ph. D.)–Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2004.
[14] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. LineUp: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, Dec. 2013.
[15] I. Herman, G. Melancon, and M. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, Jan. 2000.
[16] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 3–10, Oct. 2010.
[17] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.
[18] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann, 2010.
[19] Z. Liu, B. Jiang, and J. Heer. imMens: Real-time visual querying of big data. *Computer Graphics Forum*, 32(3pt4):421–430, June 2013.
[20] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: The vibe system. *Information Processing & Management*, 29(1):69–81, Jan. 1993.
[21] R. Sargent, C. Bartley, P. Dille, J. Keller, I. Nourbakhsh, and R. LeGrand. Timelapse GigaPan: Capturing, sharing, and exploring timelapse gigapixel imagery. *Fine International Conference on Gigapixel Imaging for Science*, Nov. 2010.
[22] C. D. Schunn and D. Klahr. A 4-space model of scientific discovery. In *Proceedings of the seventeenth annual conference of the Cognitive Science Society*, page 106–111, 1995.
[23] M. Theus. High-dimensional data visualization. In *Handbook of Data Visualization*, Springer Handbooks Comp.Statistics, pages 151–178. Springer Berlin Heidelberg, Jan. 2008.
[24] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, pages 57–64, 2004.