User Study: Datasets and Visualization Sequences

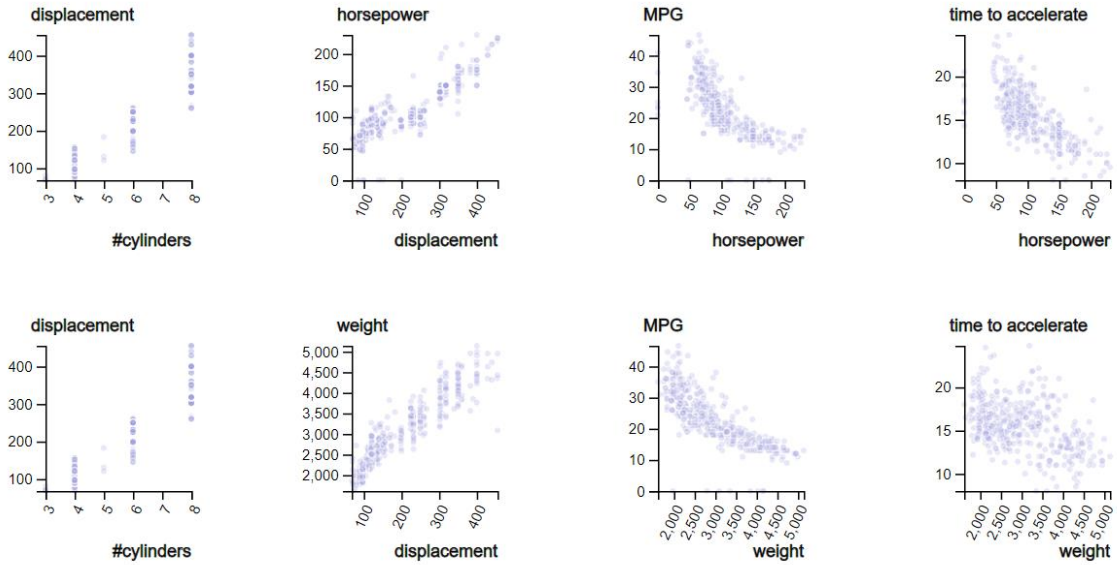
Data:

We use four test datasets - cars, college, pm10, and sales - for our study. For each dataset we generate random, spurious, causal detective, and causal exhaustive sequences. We show the different sequences for each dataset below. We also show the causal network for each dataset.

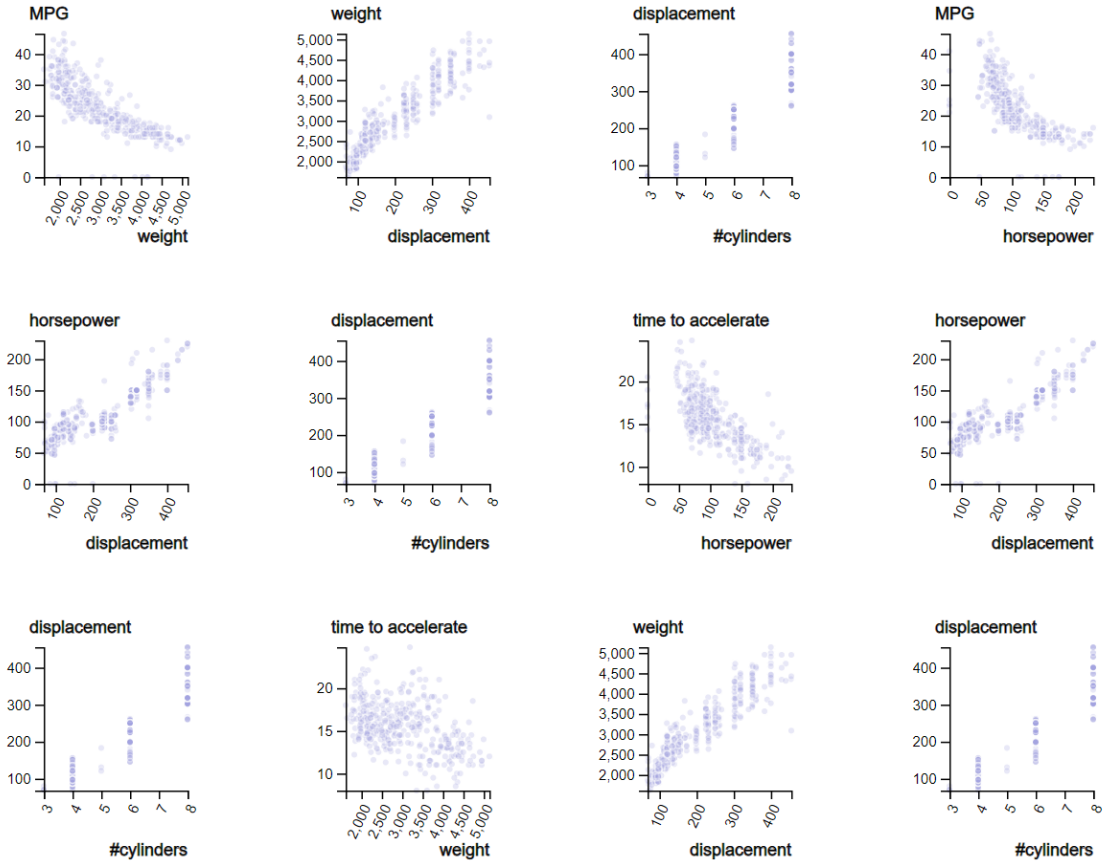
Cars:

Causal Network:

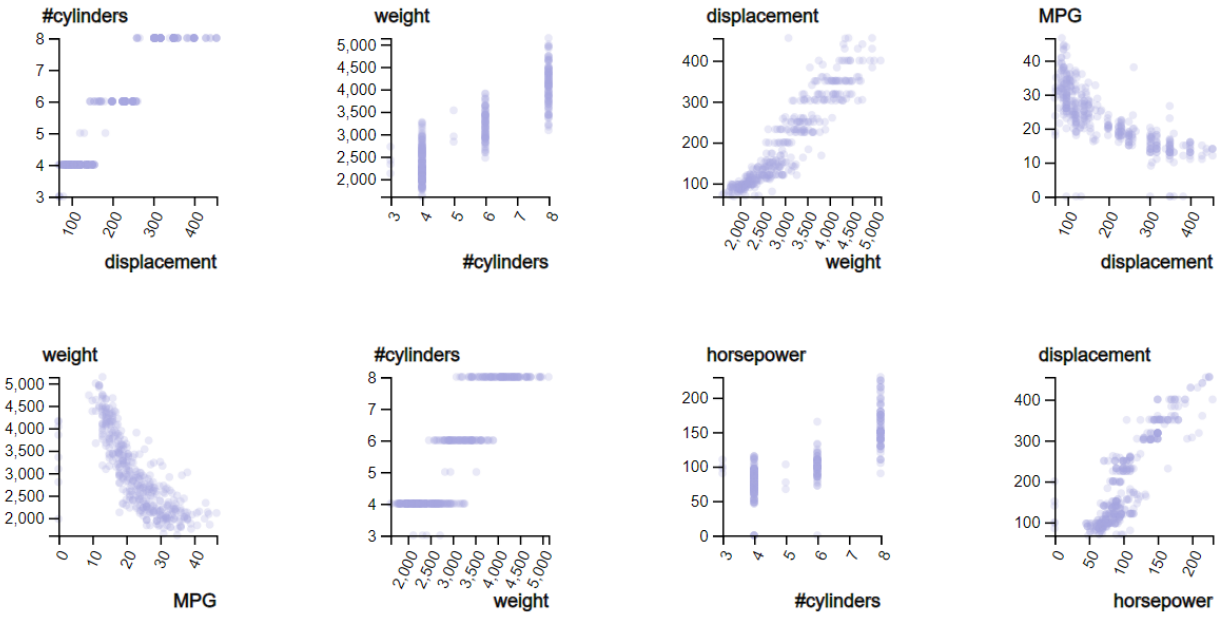
Exhaustive Scatterplot Sequence:



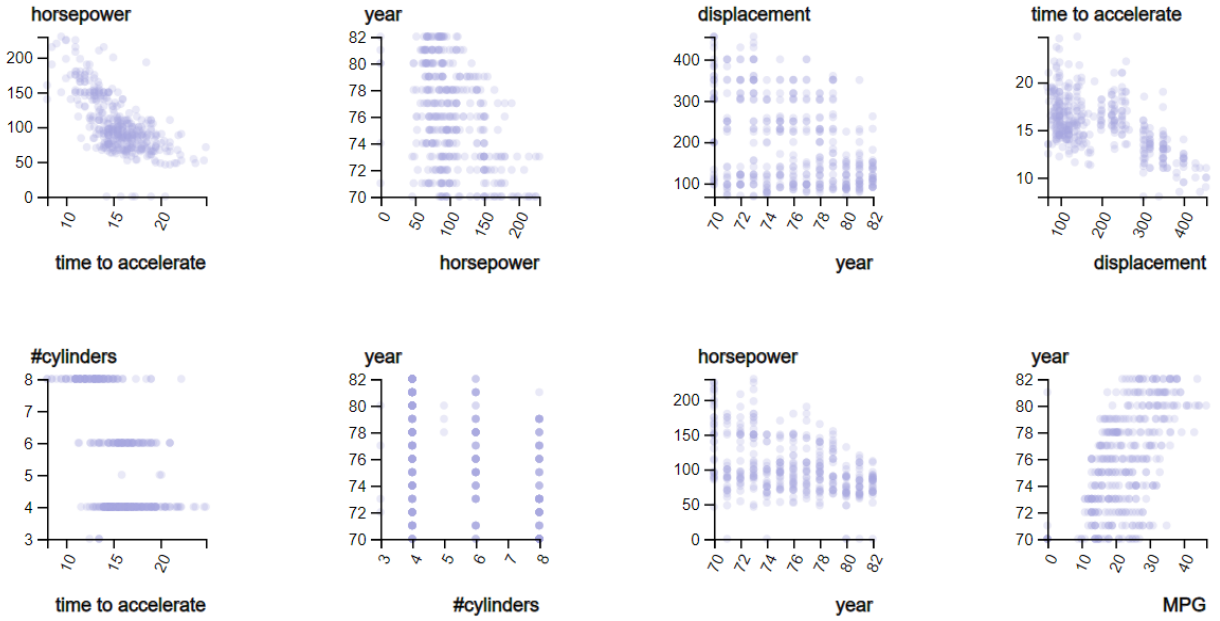
Detective Scatterplot Sequence:



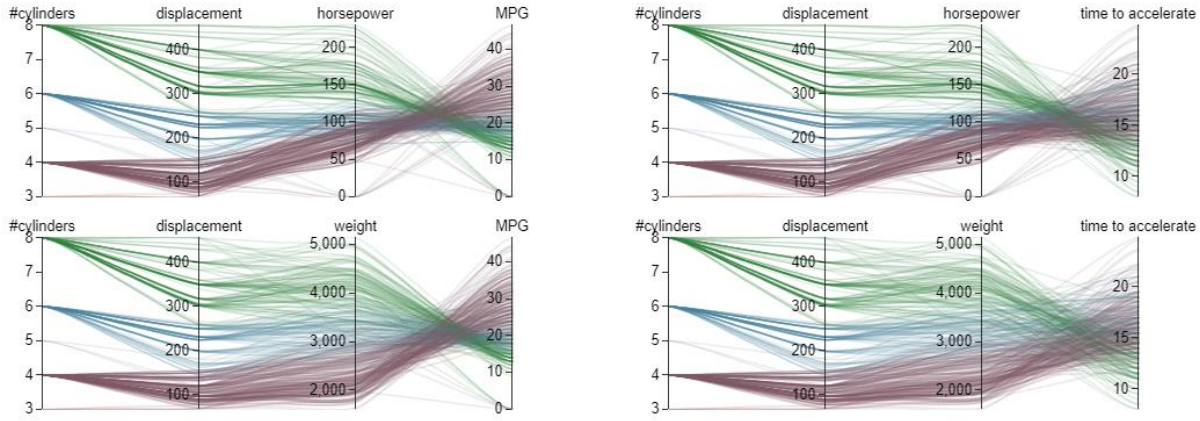
Spurious Scatterplot Sequence:



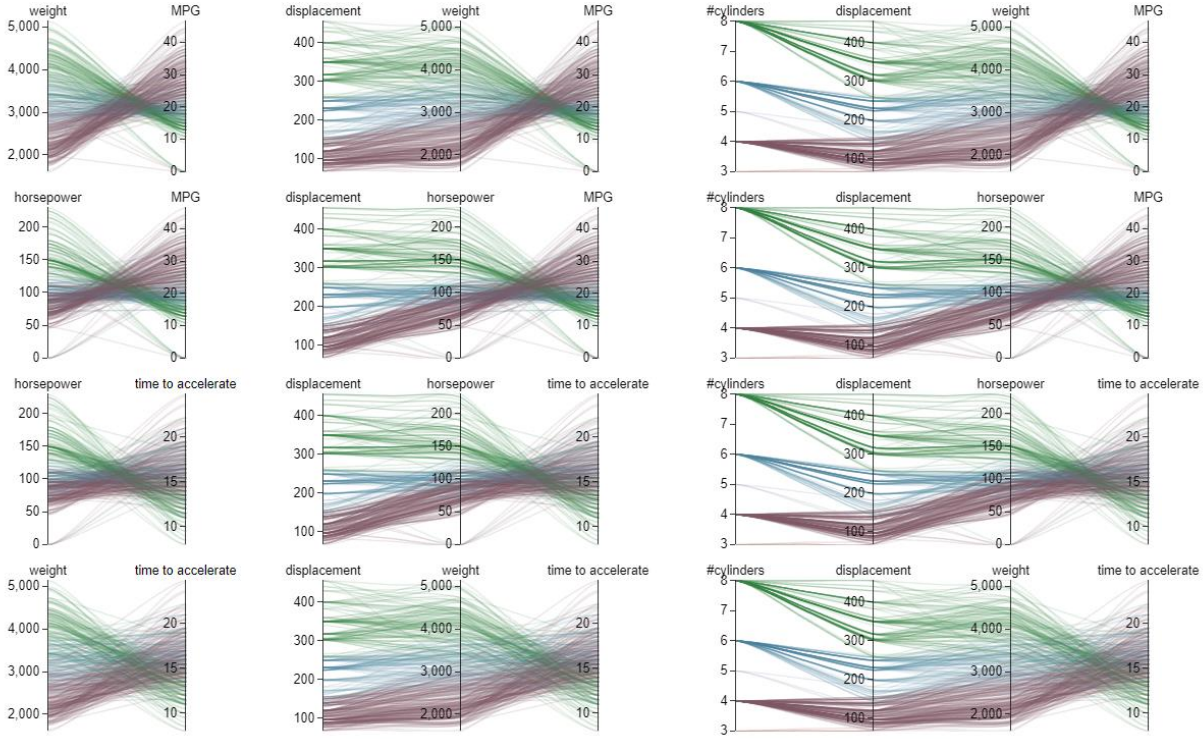
Random Scatterplot Sequence:



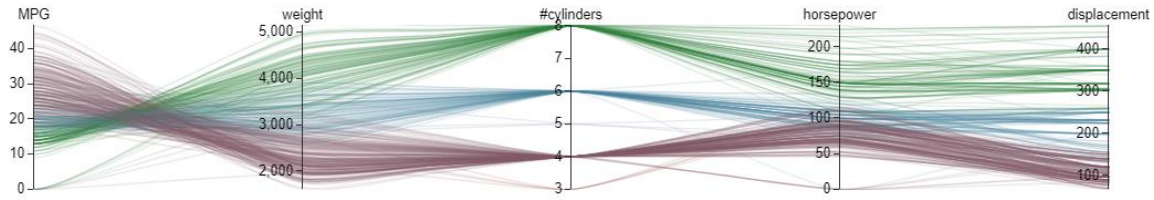
Exhaustive Parallel Coordinate Sequence:



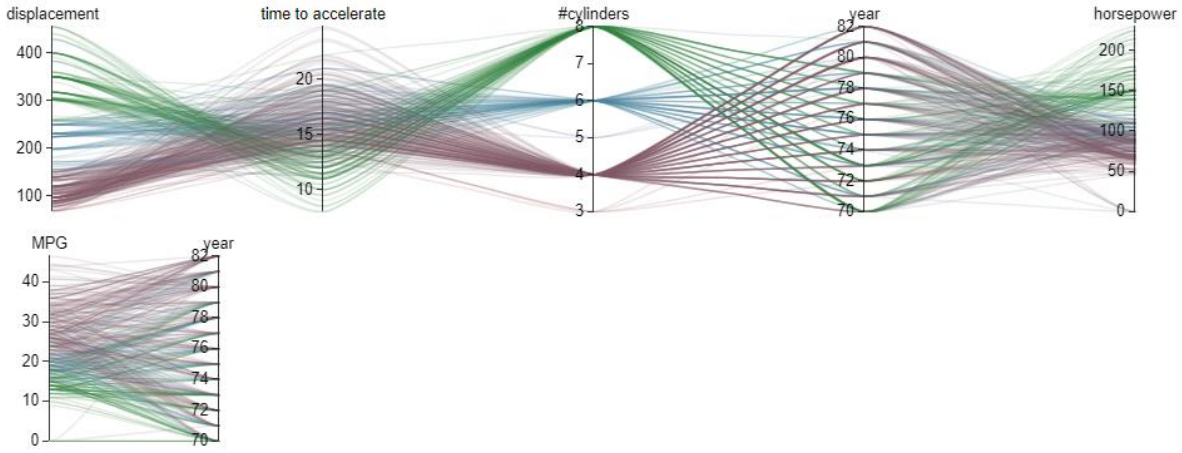
Detective Parallel Coordinate Sequence:



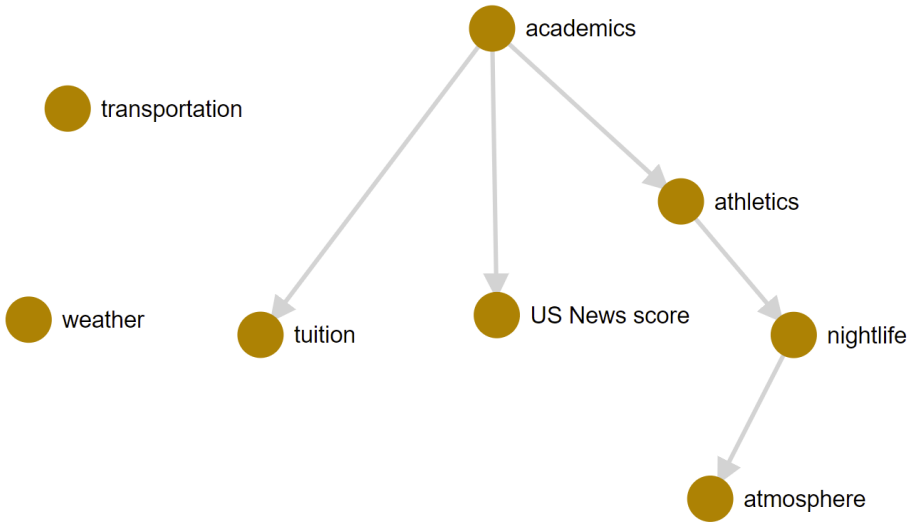
Spurious Parallel Coordinate Sequence:



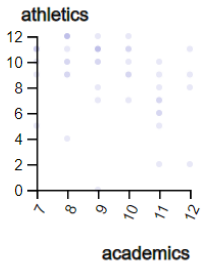
Random Parallel Coordinate Sequence:

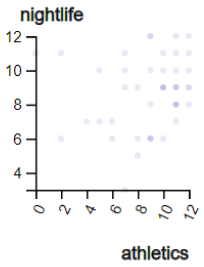


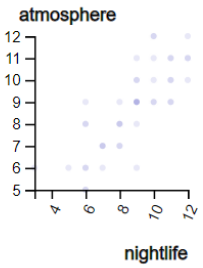
College:

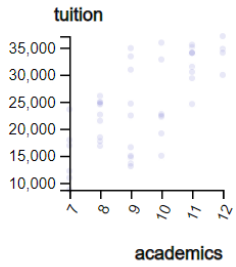


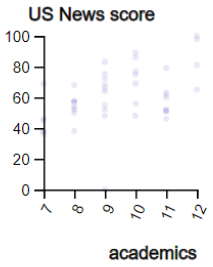
Exhaustive Scatterplot Sequence:



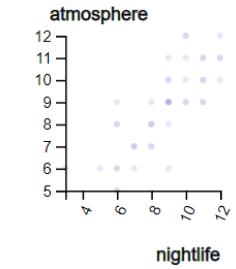


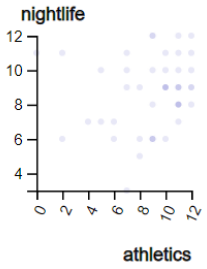


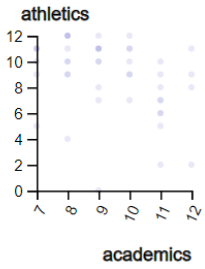


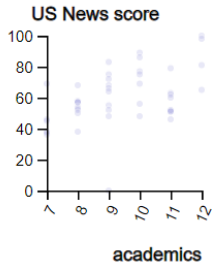


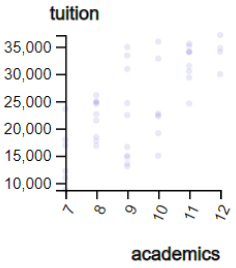
Detective Scatterplot Sequence:



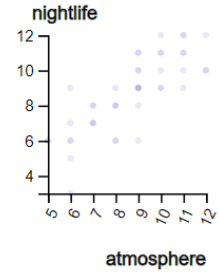


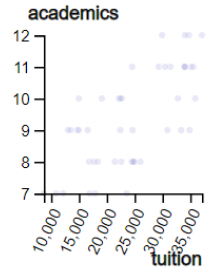


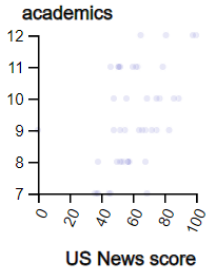


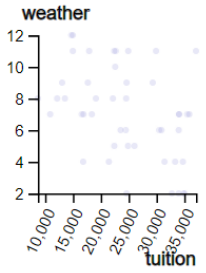


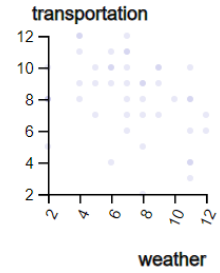
Spurious Scatterplot Sequence:



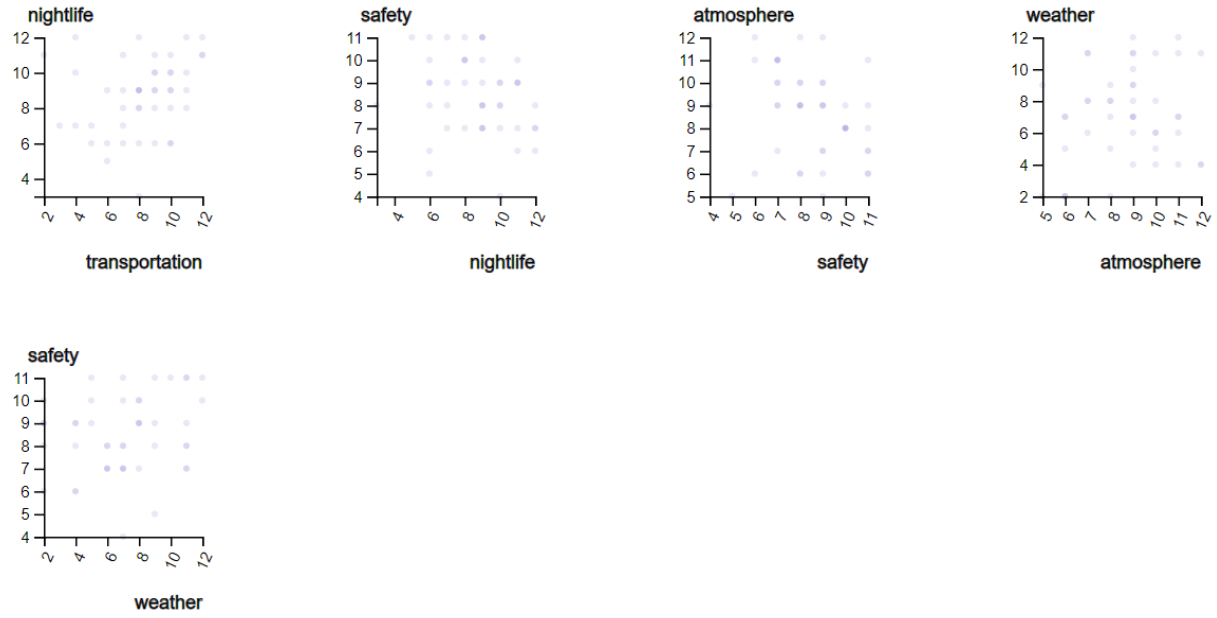




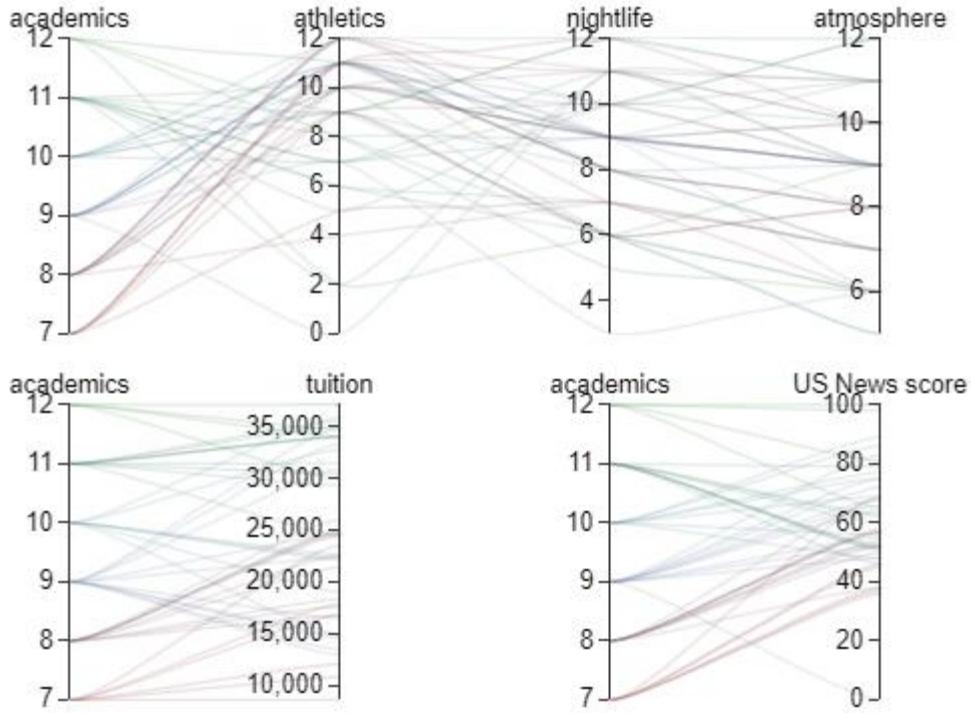




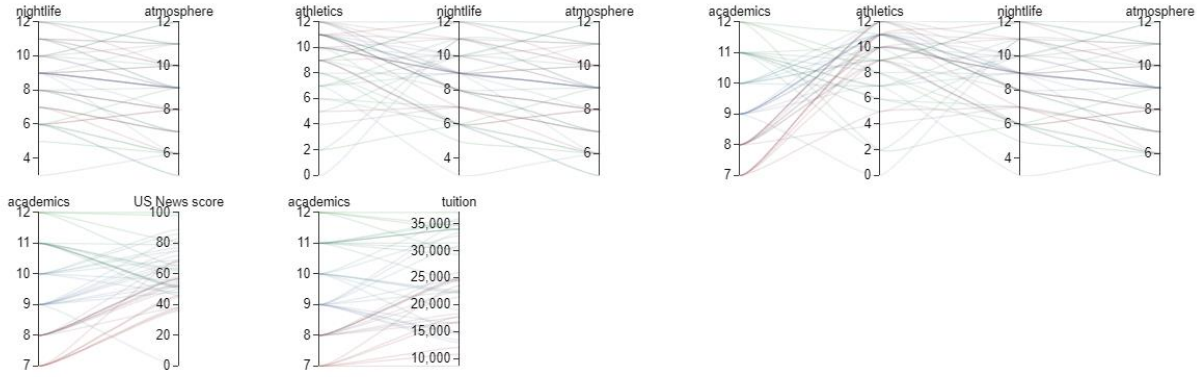
Random Scatterplot Sequence:



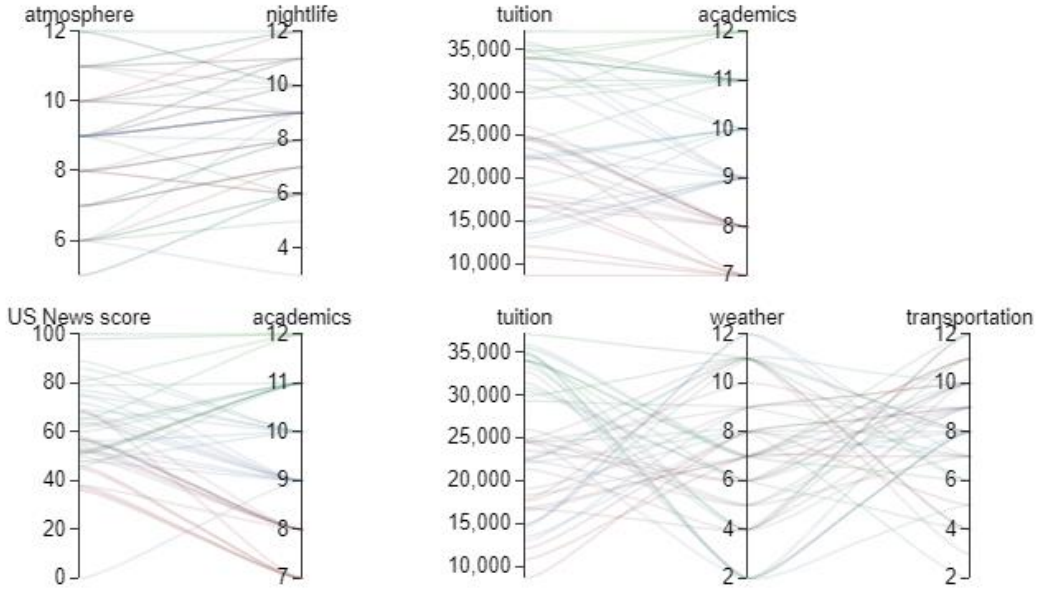
Exhaustive Parallel Coordinate Sequence:



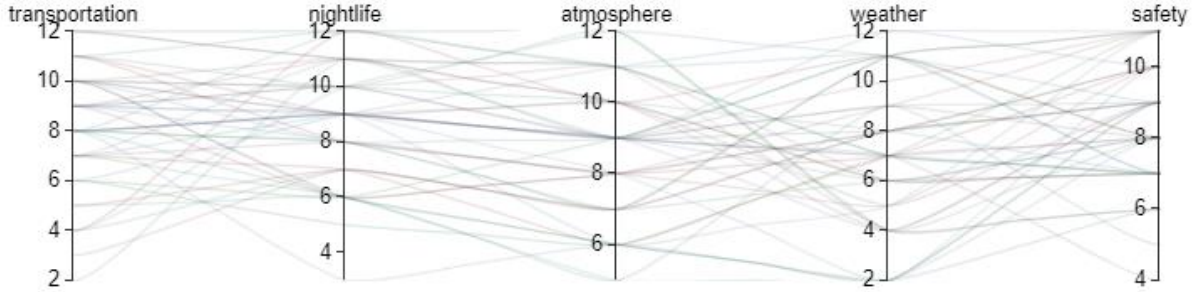
Detective Parallel Coordinate Sequence:



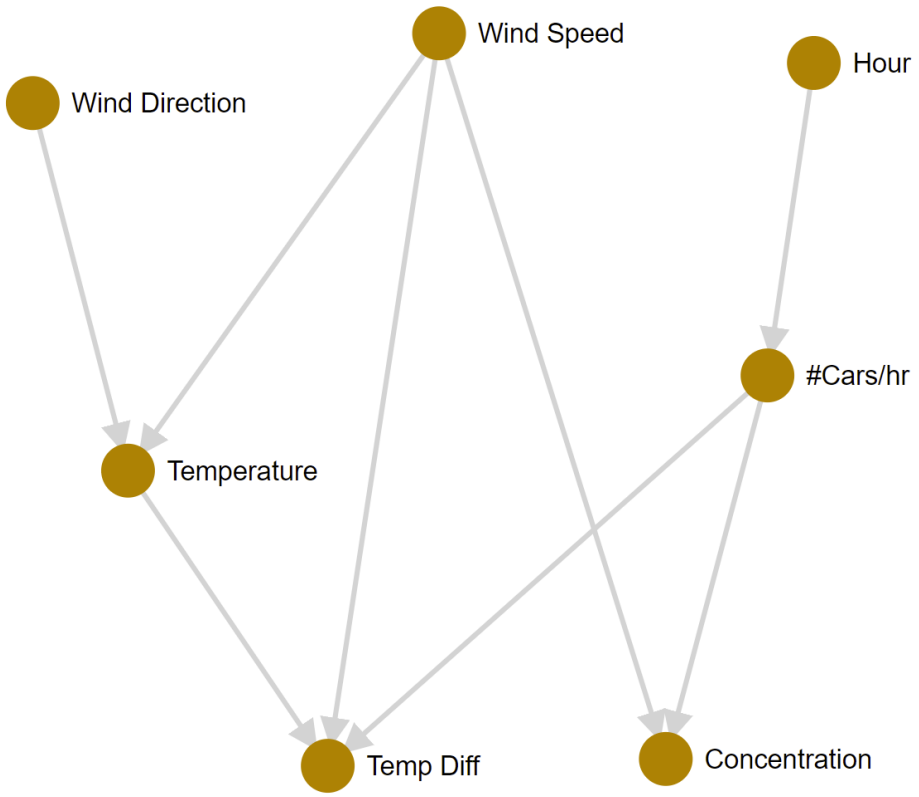
Spurious Parallel Coordinate Sequence:



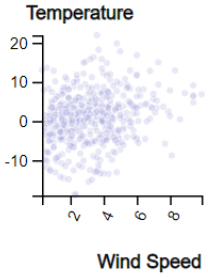
Random Parallel Coordinate Sequence:

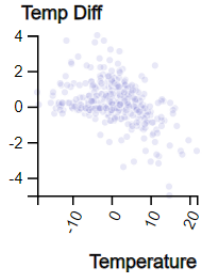


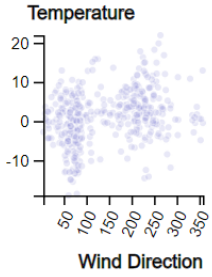
PM10:

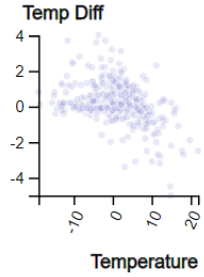

Wind Speed
Hour
Wind Direction
#Cars/hr
Temperature
Temp Diff
Concentration

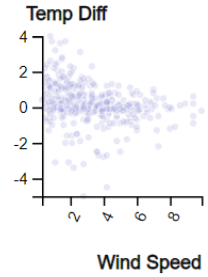
Exhaustive Scatterplot Sequence:

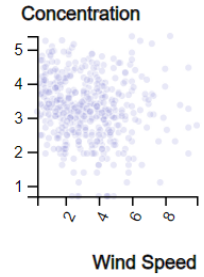


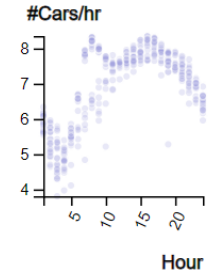


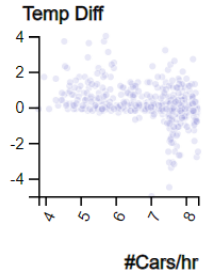


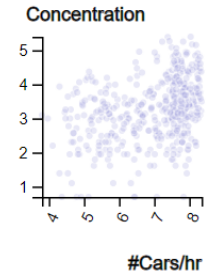




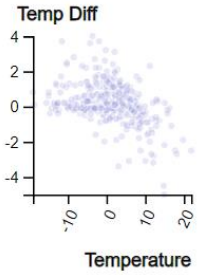


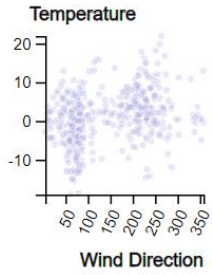


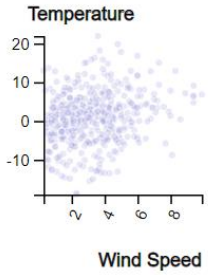


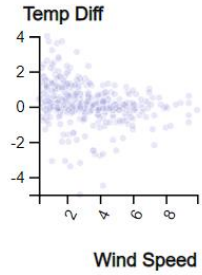


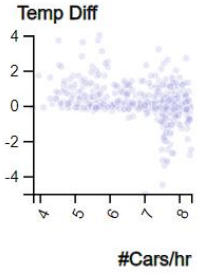
Detective Scatterplot Sequence:

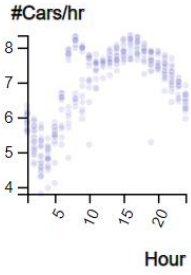


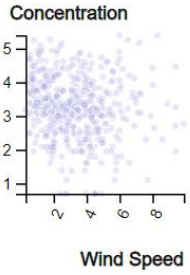


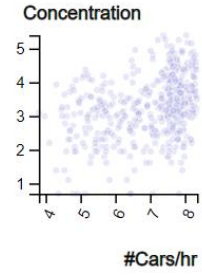


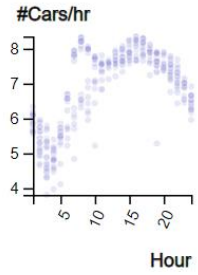




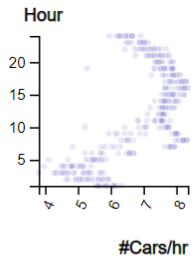


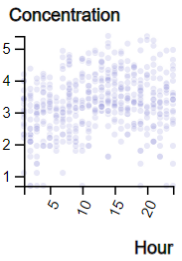


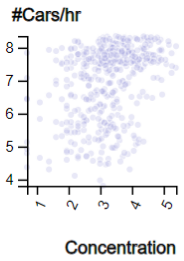


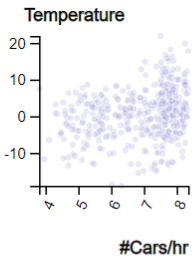


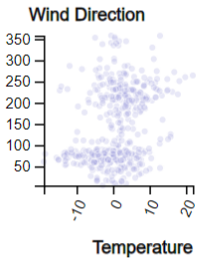
Spurious Scatterplot Sequence:

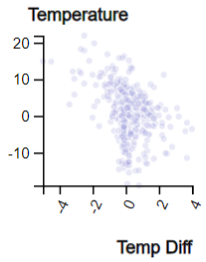


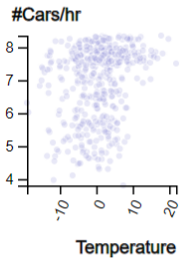


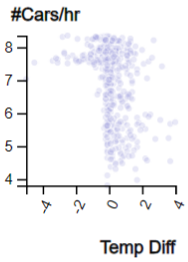


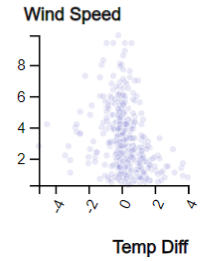




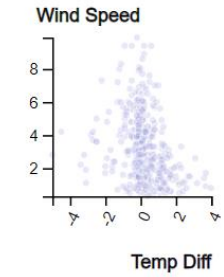


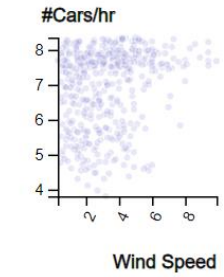


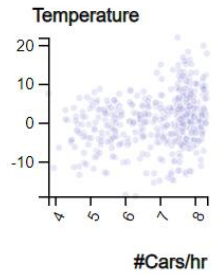


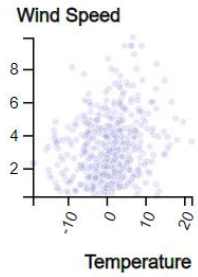


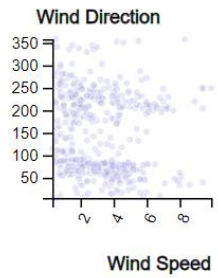
Random Scatterplot Sequence:

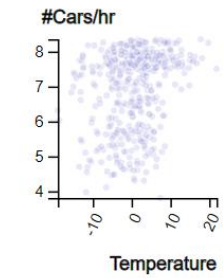


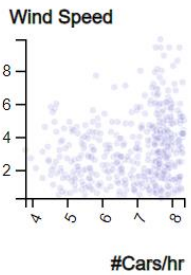




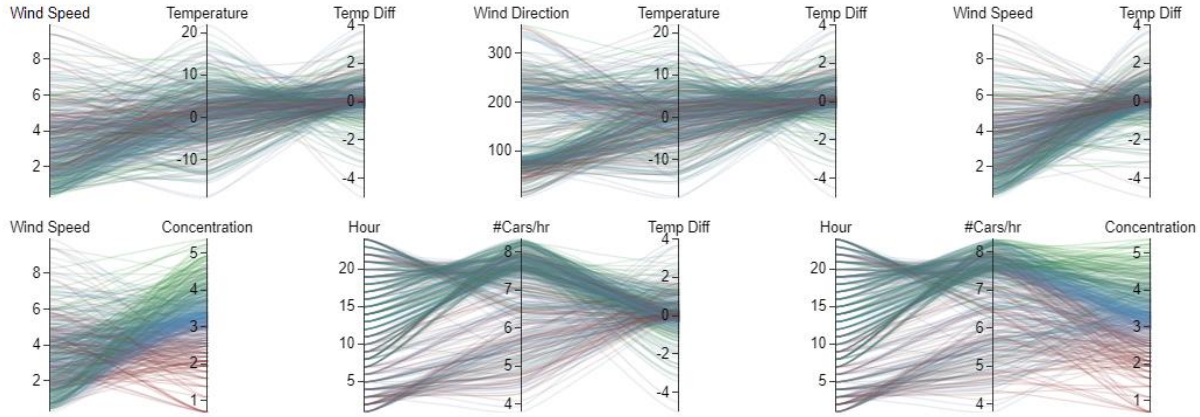




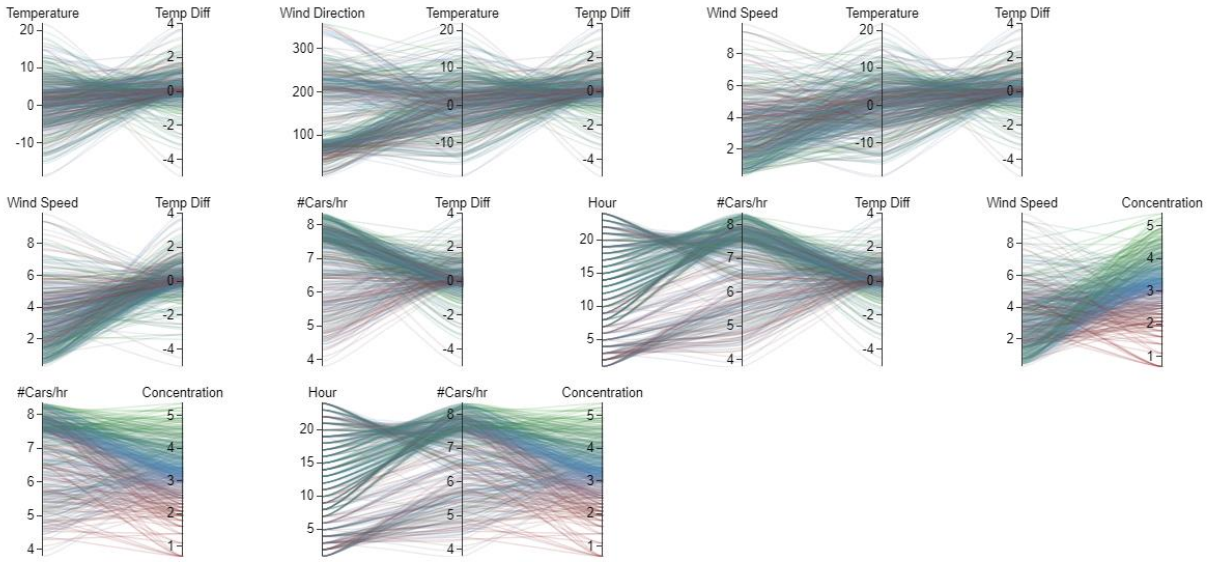




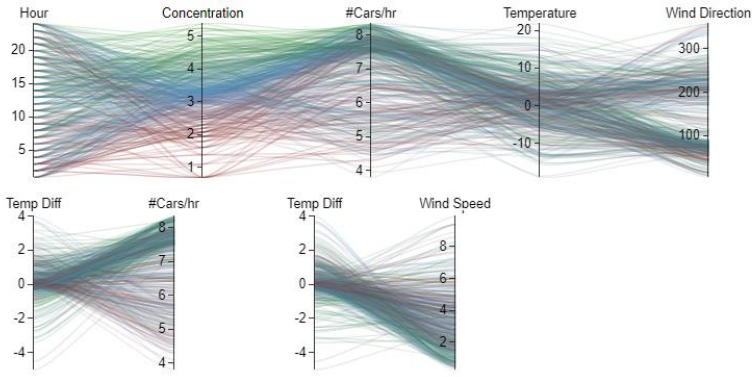
Exhaustive Parallel Coordinate Sequence:

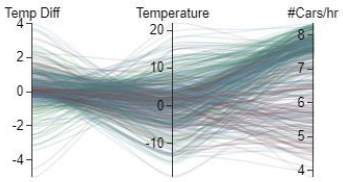


Detective Parallel Coordinate Sequence:

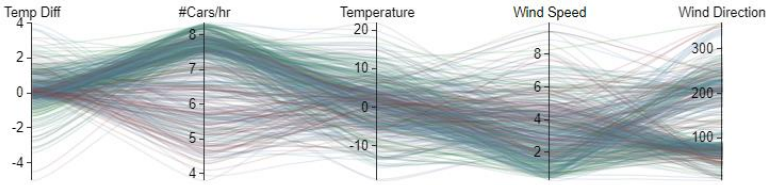


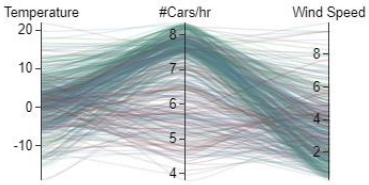
Spurious Parallel Coordinate Sequence:



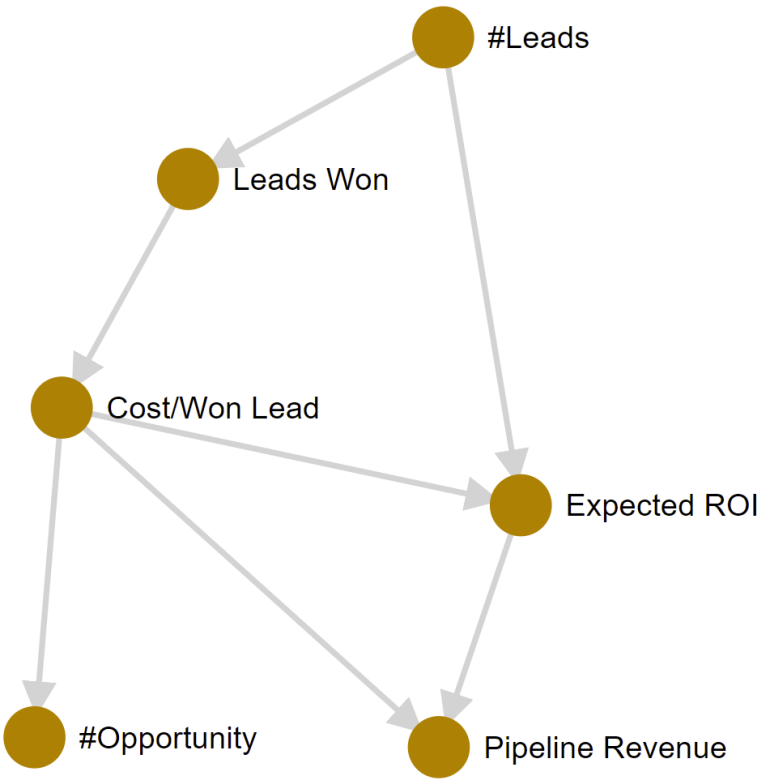


Random Parallel Coordinate Sequence:

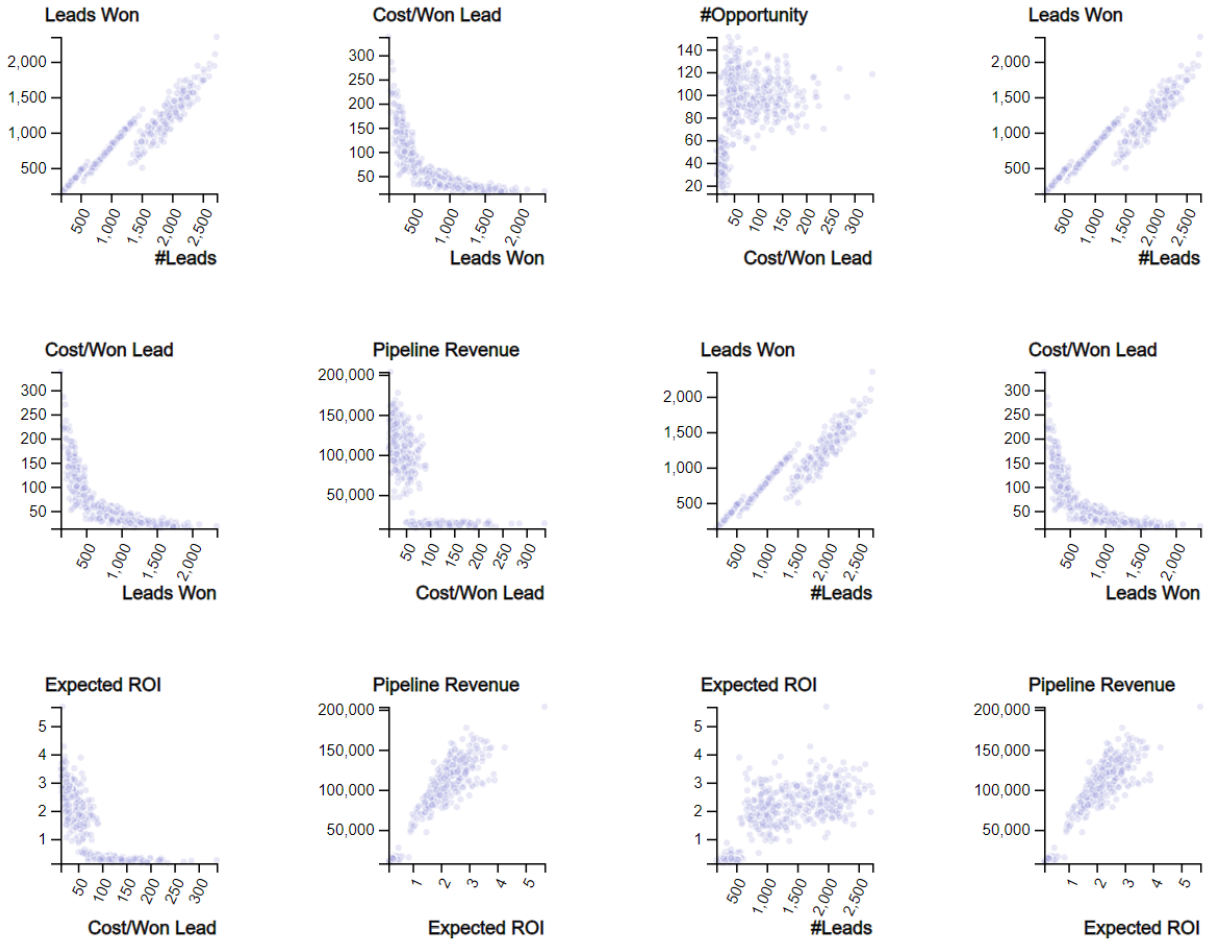




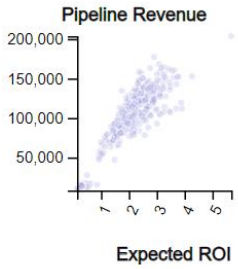
Sales:

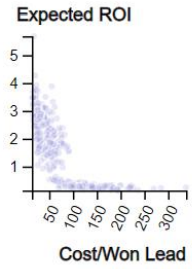

#Leads
Leads Won
Cost/Won Lead
Expected ROI
#Opportunity
Pipeline Revenue

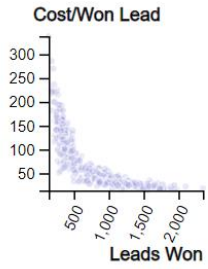
Exhaustive Scatterplot Sequence:

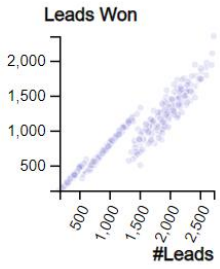


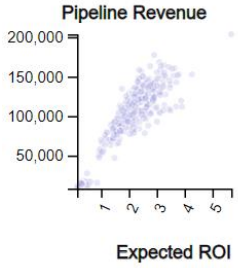
Detective Scatterplot Sequence:

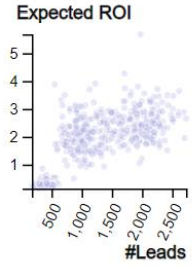


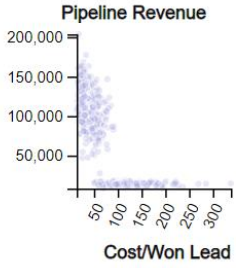


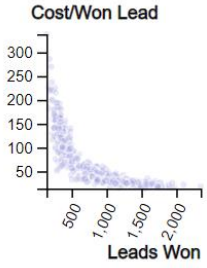


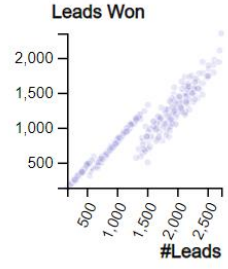


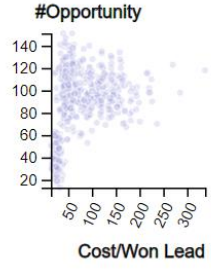


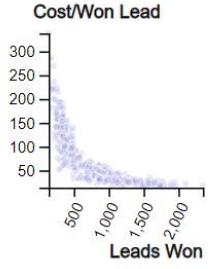


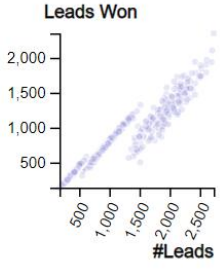




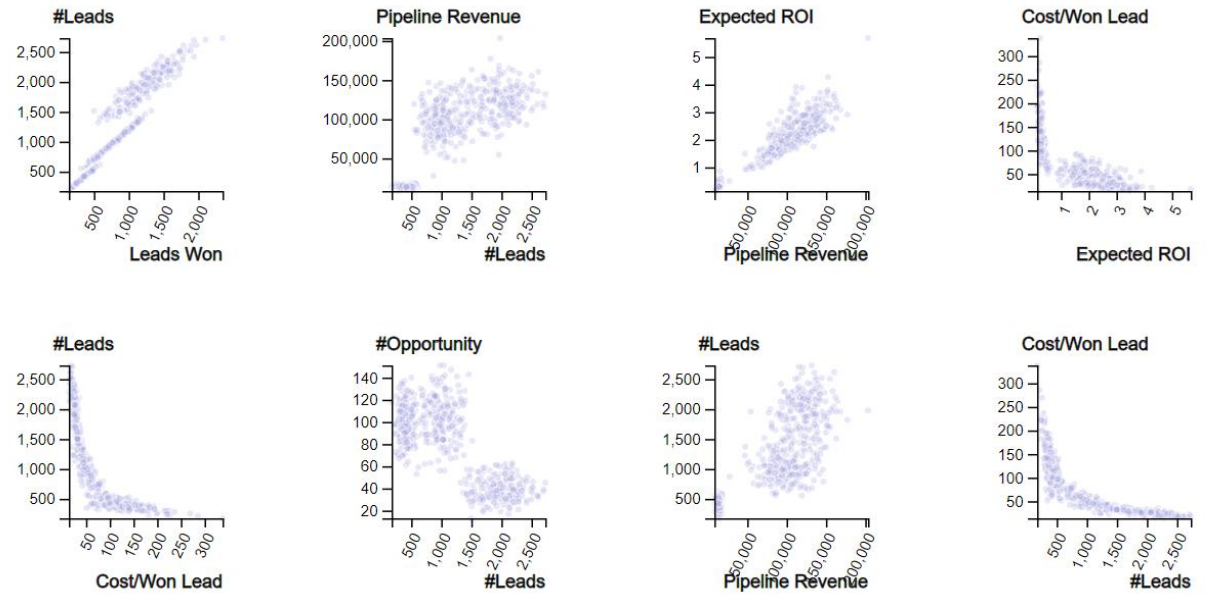




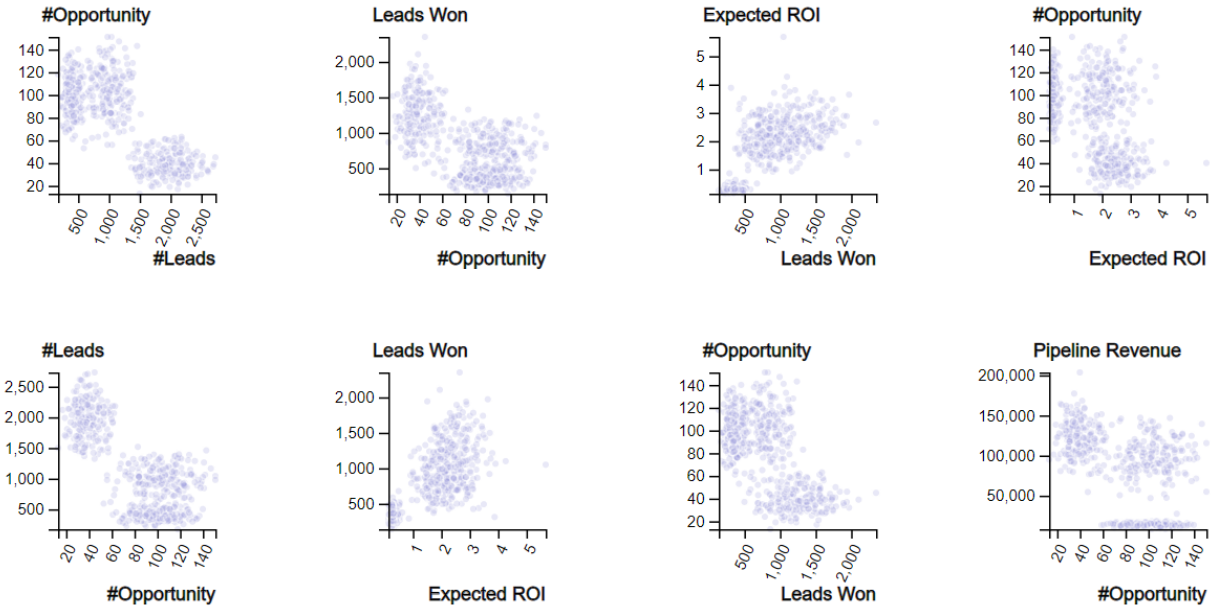




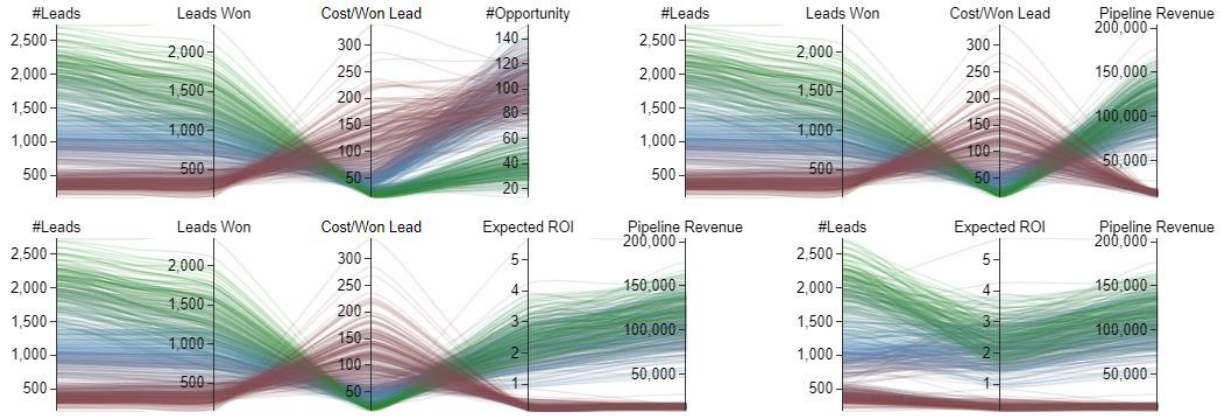
Spurious Scatterplot Sequence:



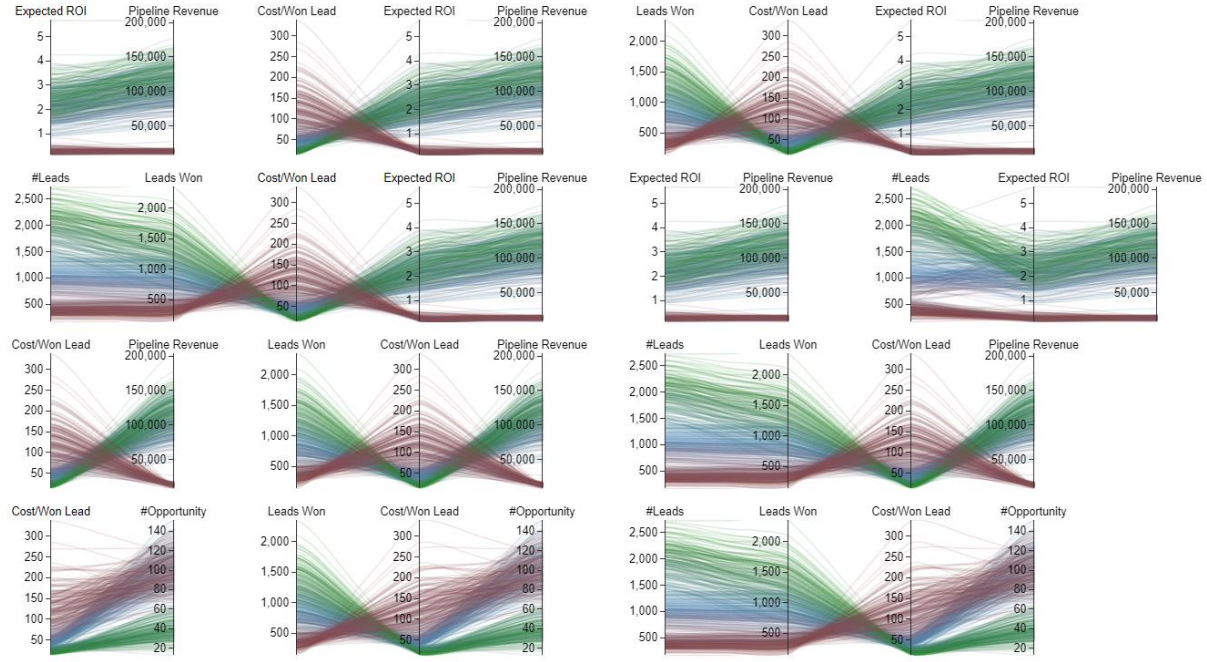
Random Scatterplot Sequence:



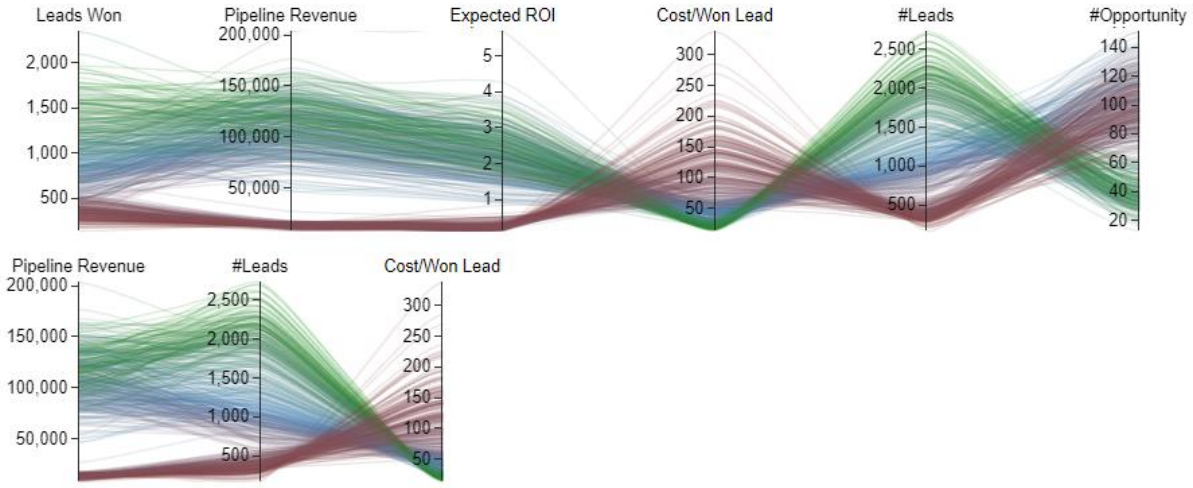
Exhaustive Parallel Coordinate Sequence:



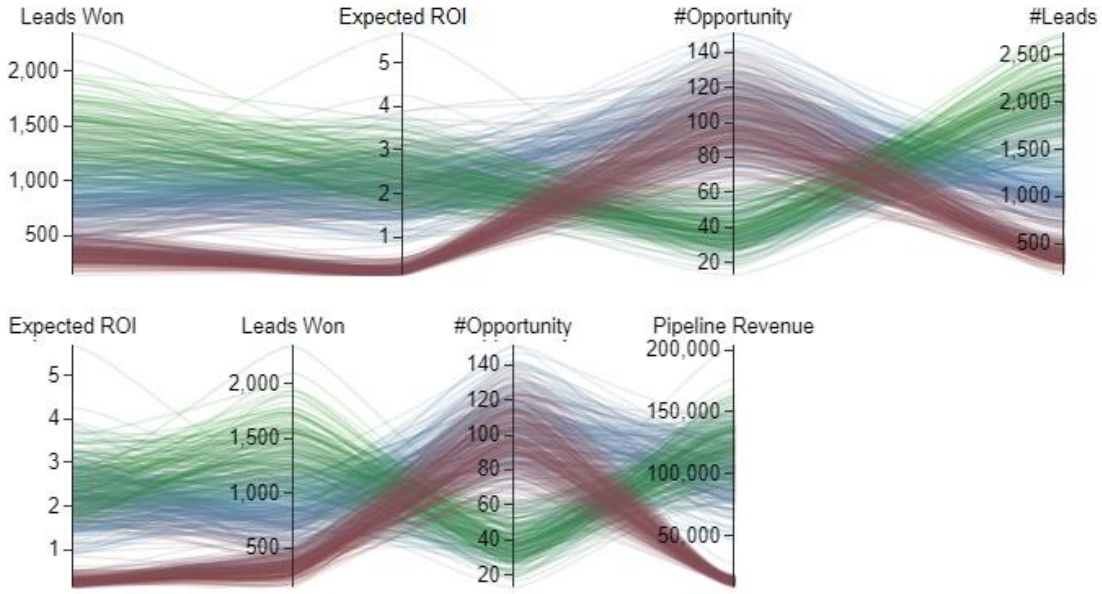
Detective Parallel Coordinate Sequence:



Spurious Parallel Coordinate Sequence:

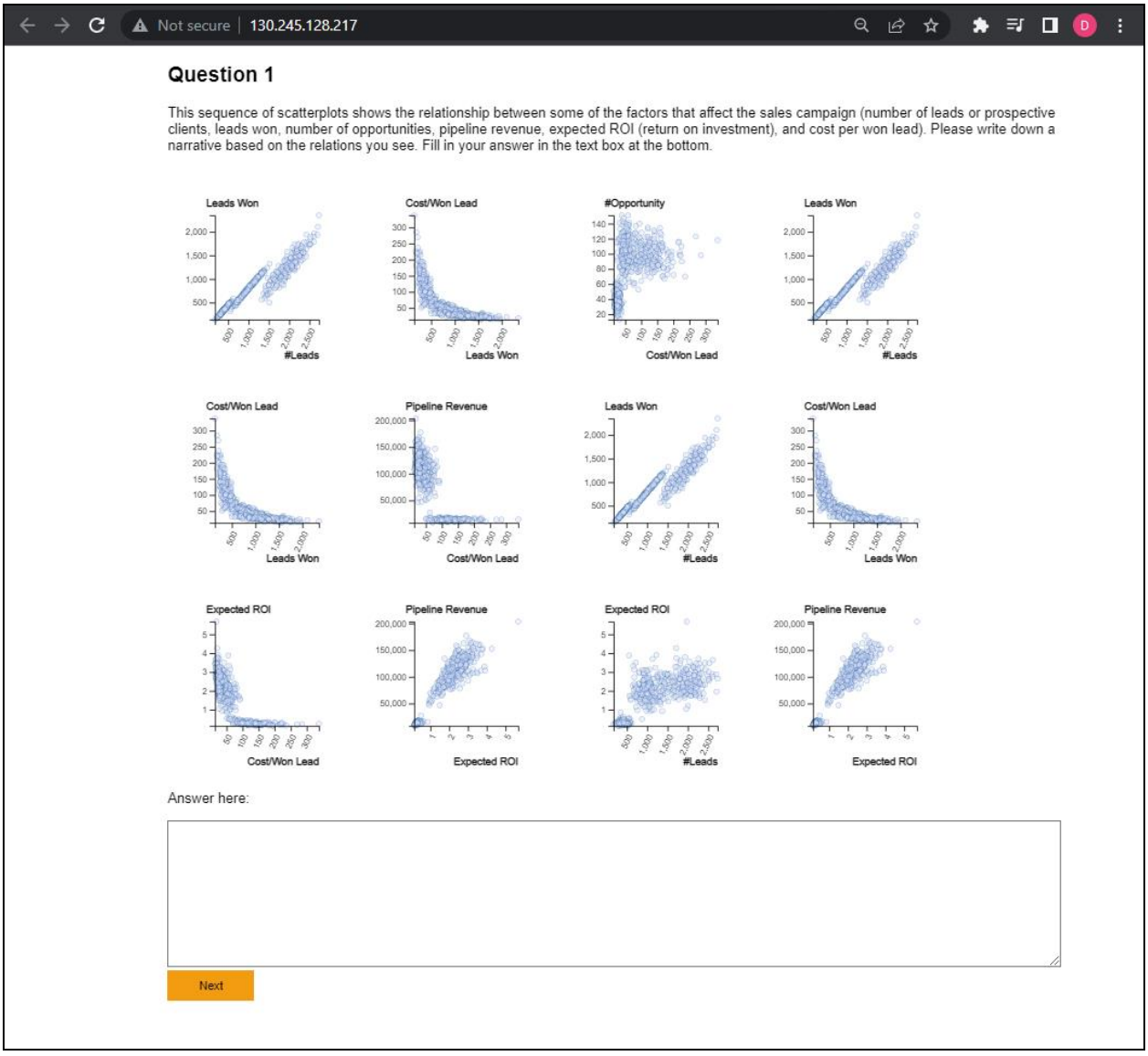


Random Parallel Coordinate Sequence:



Example Question

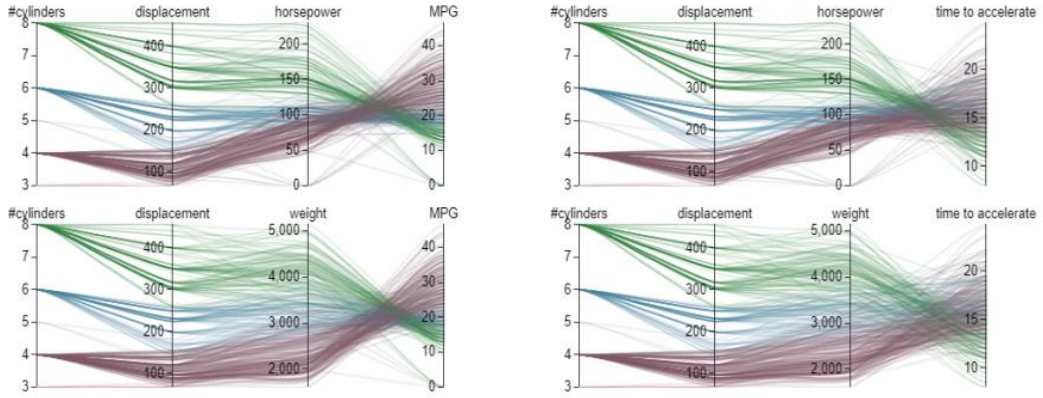
Here we show a complete example of a question a participant received with a scatterplot and a parallel coordinate sequence:



Not secure | 130.245.128.217

Question 1

This sequence of parallel coordinate plots shows the relationship between some of the attributes of cars (MPG or mileage, number of cylinders, displacement, horsepower, weight, time to accelerate, and the year of manufacture). Please write down a narrative based on the relations you see. Fill in your answer in the text box at the bottom.



Answer here:

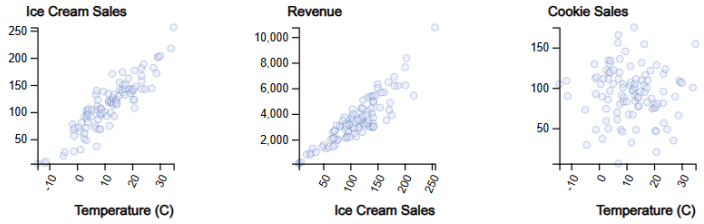
Next

Scatterplot Tutorial and Sample Question

Tutorial

The charts shown below are scatterplots. Here we use an example of data that shows the sales of ice cream, hot chocolate, and cookies at different outdoor temperatures. Each chart shows a pair of numerical data attributes - temperature & ice cream sales, revenue & ice cream sales, and temperature & cookies sales. Each attribute in the pair is assigned an x or y axis. The attribute values along an axis increase from left to right for the x-axis and bottom to top for the y-axis. Each point on the chart tells us the number sales of a product at a particular temperature.

So looking at these charts, what do we learn about how the attributes are related? Read on further below the graphs.



Looking at the scatterplots we can see that:
- As the temperature increases the sale of ice cream increases
- As the sale of ice cream increases the revenue increases
- We do not seem to see any strong relation between cookie sales and temperature.
- Finally we can conclude that as the temperature goes up more ice creams are sold which in turn leads to an increase in the revenue.

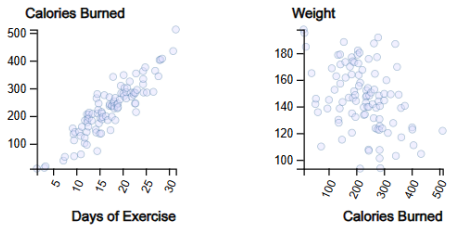
Next we will test if you understand the task. Click next to proceed.

Next

Practice Question

Let's see if you got the concept. This sequence of scatterplots shows the relationship between some of the factors that affect a persons weight (Calories Burned, Weight, Days of Excercise). From the options below please check all relevant options.

So looking at these charts, what do we learn about how the attributes are related? Read on further below the graphs.



- ☐ More calories are burnt if a person exercises for fewer days.
- ☐ More calories are burnt by lighter people.
- ☐ People with higher weight tend to exercise for more days.
- ☐ Exercising for more days burns more calories.
- ☐ Weight increases when more calories are burnt.
- ☐ Calories burnt has no real effect on weight.
- ☐ People exercising for more days tend to be lighter.

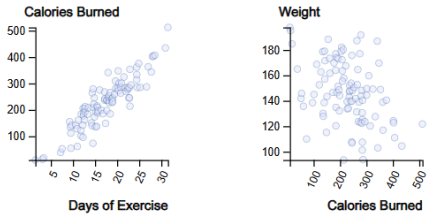
Click next to see the answers.

Next

Practice Question

Let's see if you got the concept. This sequence of scatterplots shows the relationship between some of the factors that affect a persons weight (Calories Burned, Weight, Days of Excercise). From the options below please check all relevant options.

So looking at these charts, what do we learn about how the attributes are related? Read on further below the graphs.



- [ ] ~~More calories are burnt if a person excercises for fewer days.~~
- [ ] More calories are burnt by lighter people.
- [ ] ~~People with higher weight tend to excercise for more days.~~
- [x] Exercising for more days burns more calories.
- [ ] ~~Weight increases when more calories are burnt.~~
- [ ] ~~Calories burnt has no real effect on weight.~~
- [x] People exercising for more days tend to be lighter.

For the questions that follow you will be required to observe sets of scatterplots showing data from different datasets and write down a narrative based on the relations you see. You will not receive a list of relations as in this practice question. Click next to proceed to the questions.

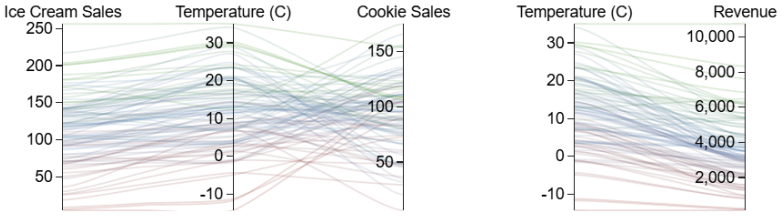
Next

Parallel Coordinates Tutorial and Sample Question

Tutorial

The charts shown below are two parallel coordinate plots. Here we use an example of data that shows the sales of ice cream, hot chocolate, and cookies at different outdoor temperatures. The first plot shows three numerical data attributes - ice cream sales, temperature, & cookies sales. The second plot shows two numerical data attributes - temperature & revenue. Each attribute is assigned a vertical axis and these axes are placed parallel to each other at equal distances. The attribute values along each axis increase from bottom to top. Each data point is represented as a line that intersects parallel axes where the intersection locations indicate the data point's value for that attribute.

So looking at these charts, what do we learn about how the attributes are related? Read on further below the graphs.



Looking at the parallel coordinate plots we can see that:
- As the temperature increases the sale of ice cream increases
- As the temperature decreases the sale of cookies increases (criss-crossing horizontal lines)
- As the temperature increases the revenue increases
- Finally we can conclude that ice cream sales bring in more revenue than cookie sales as the revenue is higher at higher temperatures which in turn is associated with higher ice cream sales.

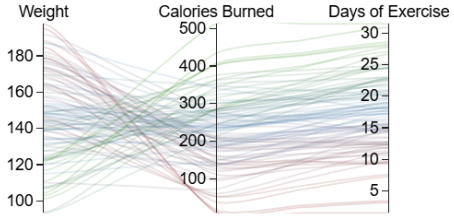
Next we will test if you understand the task. Click next to proceed.

Next

Practice Question

Let's see if you got the concept. This parallel coordinate plot shows the relationship between some of the factors that affect a persons weight (Calories Burned, Weight, Days of Excercise). From the options below please check all relevant options.

So looking at this charts, what do we learn about how the attributes are related? Read on further below the graphs.



- [ ] More calories are burnt if a person exercises for fewer days.
- [ ] More calories are burnt by lighter people.
- [ ] People with higher weight tend to exercise for more days.
- [ ] Exercising for more days burns more calories.
- [ ] Weight increases when more calories are burnt.
- [ ] Calories burnt has no real effect on weight.
- [ ] People exercising for more days tend to be lighter.

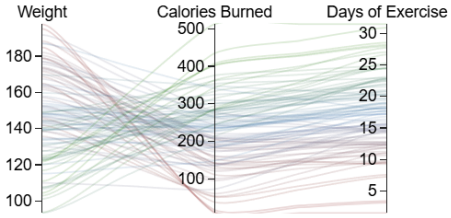
Click next to see the answers.

Next

Practice Question

Let's see if you got the concept. This parallel coordinate plot shows the relationship between some of the factors that affect a persons weight (Calories Burned, Weight, Days of Excercise). From the options below please check all relevant options.

So looking at this charts, what do we learn about how the attributes are related? Read on further below the graphs.



- [ ] ~~More calories are burnt if a person exercises for fewer days.~~
- [ ] More calories are burnt by lighter people.
- [ ] ~~People with higher weight tend to exercise for more days.~~
- [x] Exercising for more days burns more calories.
- [ ] ~~Weight increases when more calories are burnt.~~
- [ ] ~~Calories burnt has no real effect on weight.~~
- [x] People exercising for more days tend to be lighter.

For the questions that follow you will be required to observe sets of scatterplots showing data from different datasets and write down a narrative based on the relations you see. You will not receive a list of relations as in this practice question. Click next to proceed to the questions.

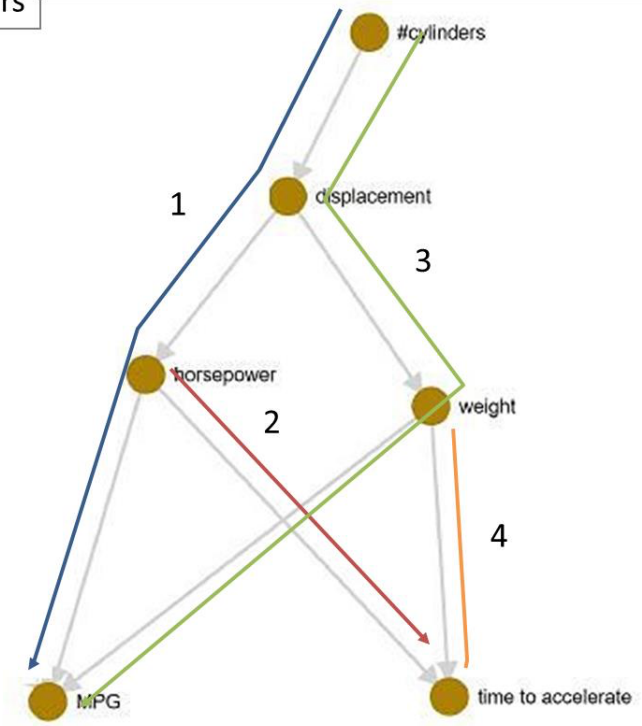
Next

Chain Decomposition

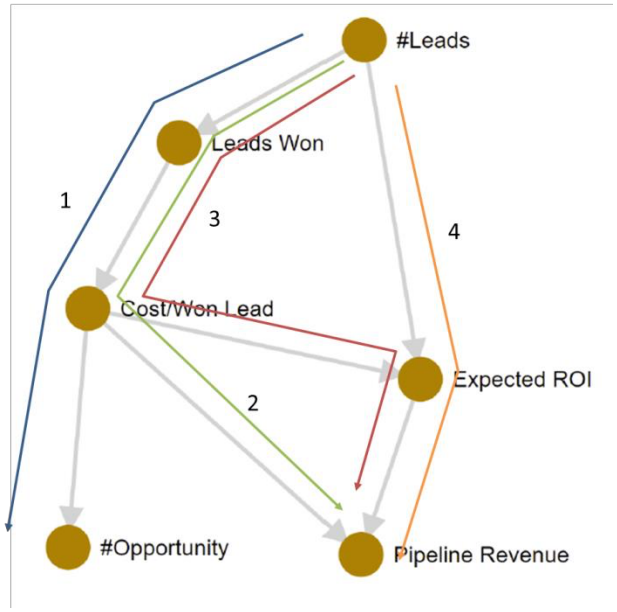
In the figures below we decompose the scatterplot sequences for each dataset used in the study into individual subsequences, each due to a partial or complete causal chain in the graph (indicated by the number next to each sub-chain).

- The Sales dataset uses the full flashback chains at branches (middle column in Fig. 2 main paper).
- The PM10 and Cars datasets use the shorter leaf chains at branches (right column of Fig. 2 main paper).
- The College dataset has no branches.

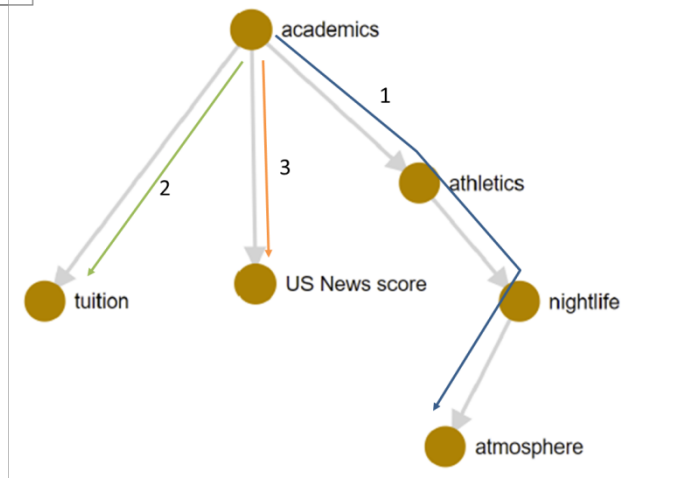
Cars



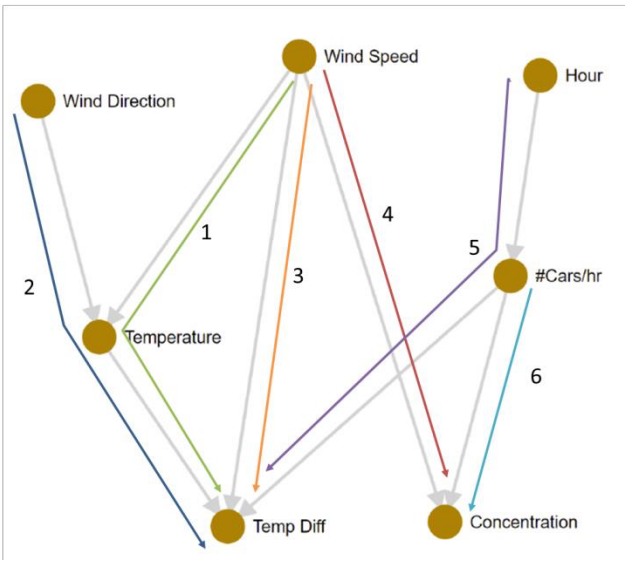
Sales



College



PM10

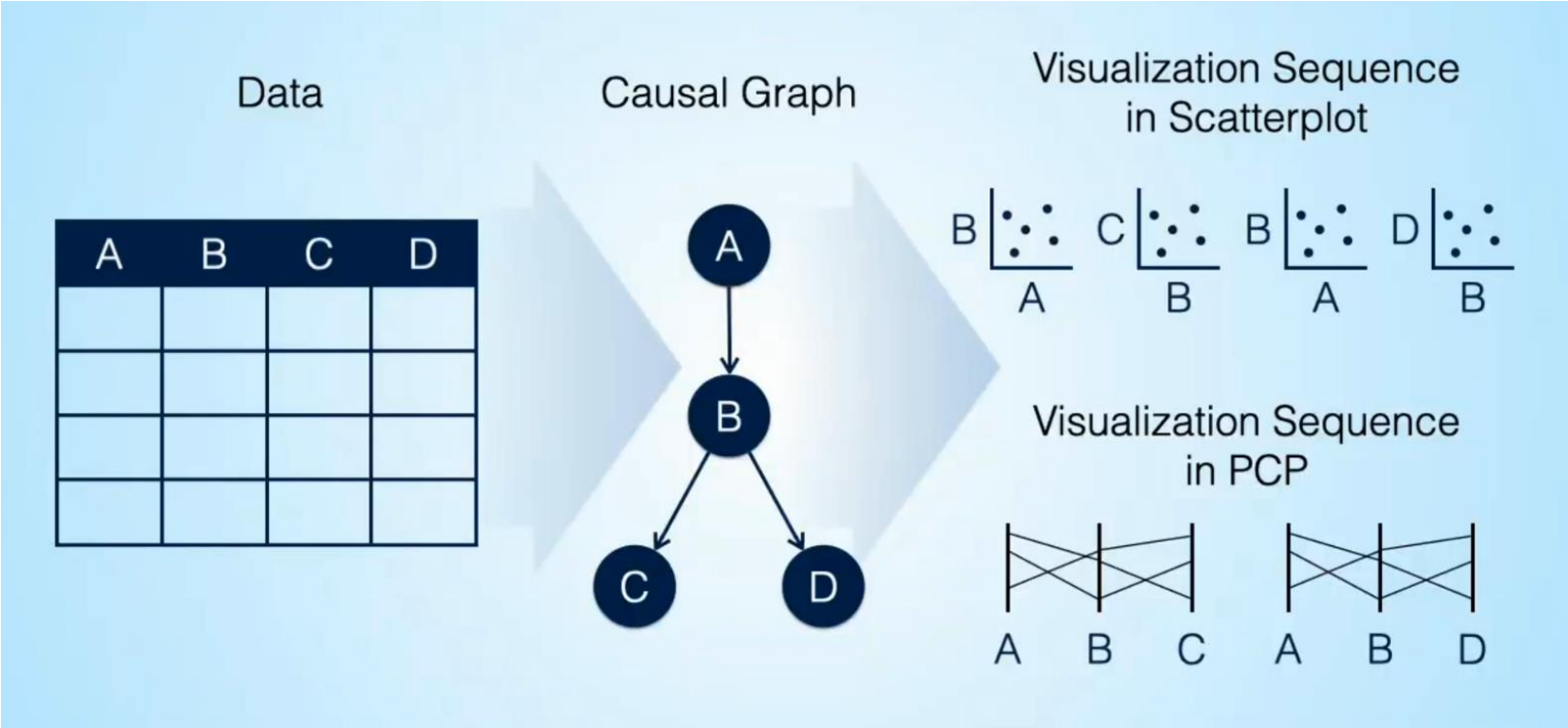


Generating Coherent Visualization Sequences for Multivariate Data by Causal Graph Traversal

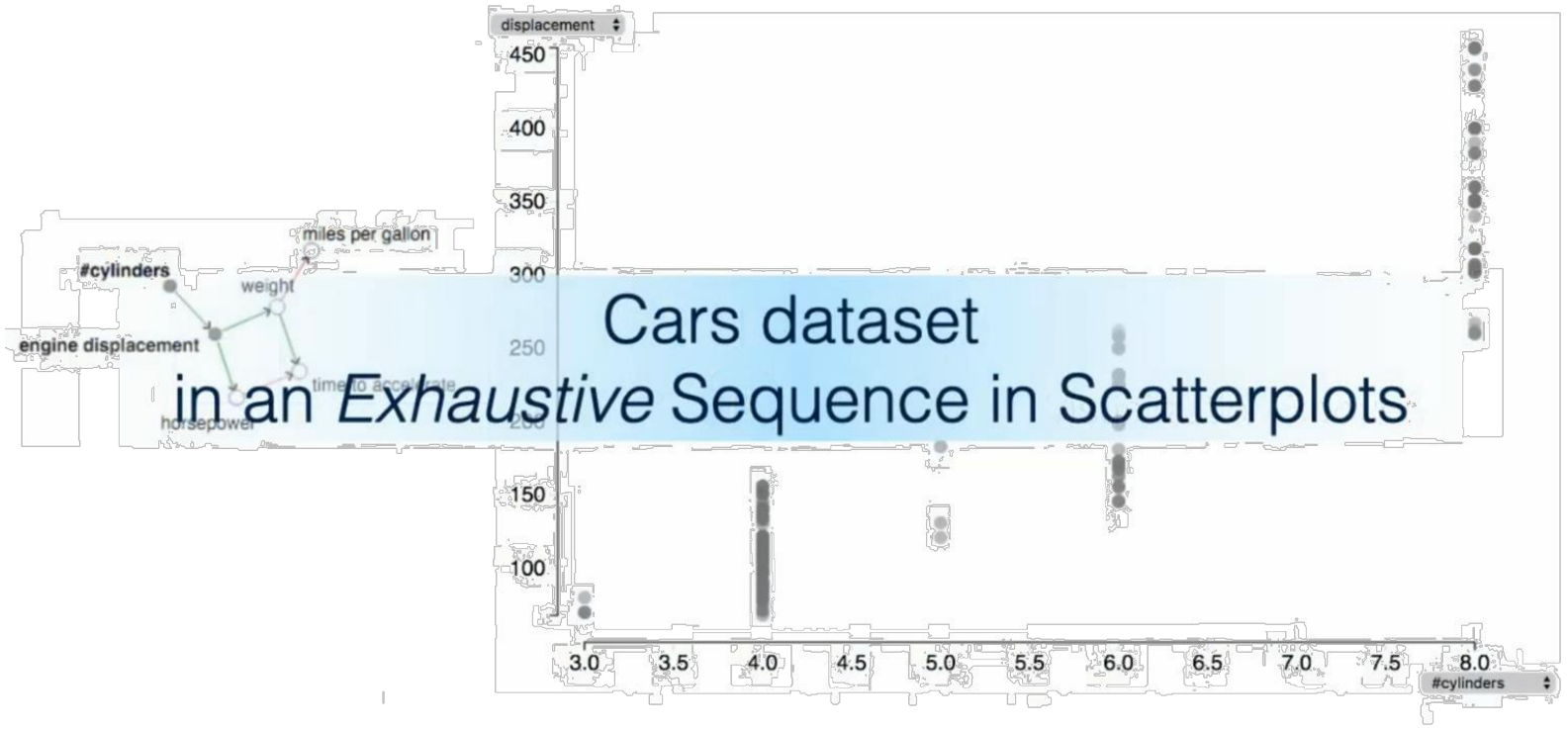
Puripant Ruchikachorn, Darius Coelho, Jun Wang, Kristina Striegnitz*, and Klaus Mueller

Stony Brook University, *Union College

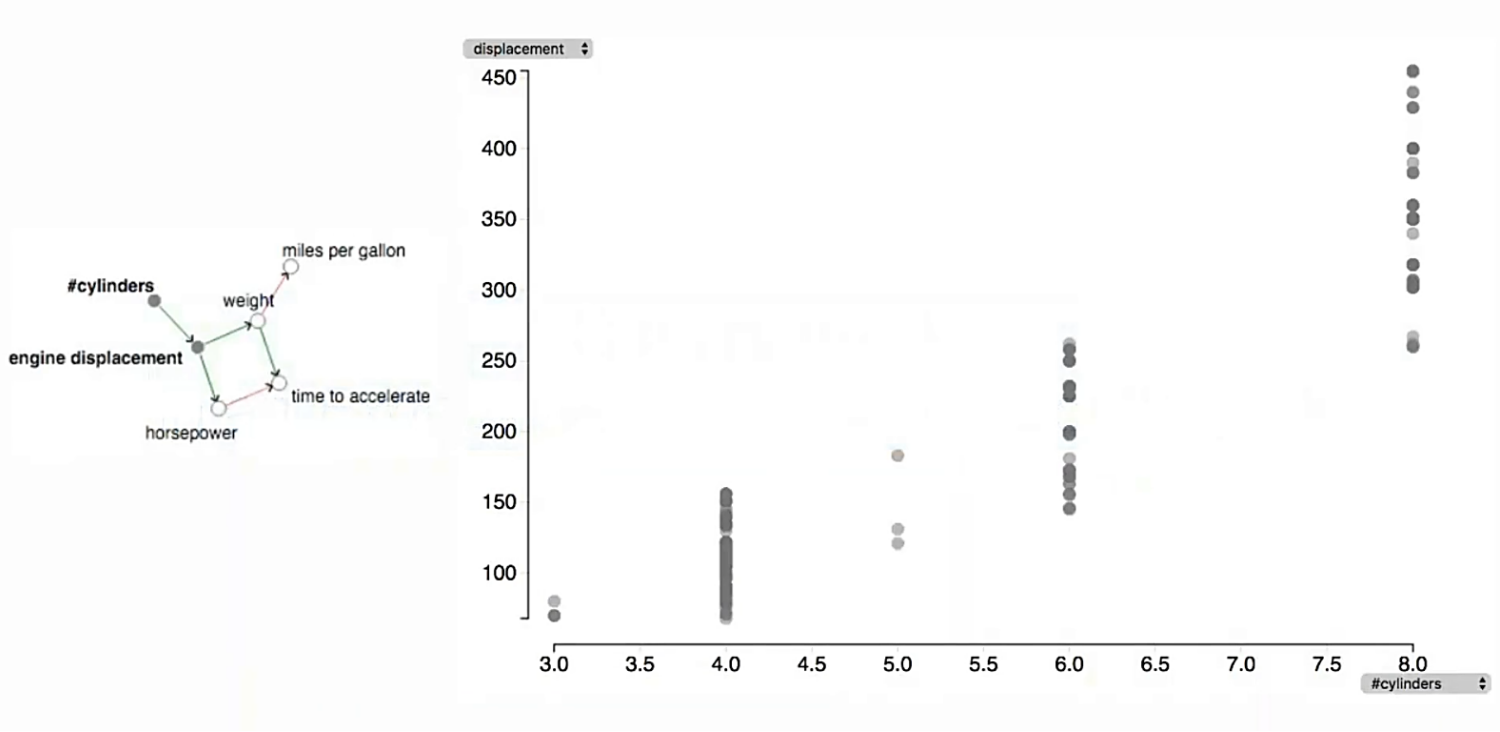
Illustrated Video Script



Our method automatically generates visualization sequences derived from an underlying causal graph, designed to support a coherent narrative of the data. To assess their effectiveness, we conducted both quantitative and qualitative studies, which validated the usefulness of these sequences for guiding interpretation. In the following sections, we present four datasets, each visualized through a distinct narrative sequence, inspired by different graph traversal strategies and storytelling genres. For each example, the corresponding causal graph is displayed on the left to illustrate the logic behind the chosen sequence. While this graph would typically remain hidden in real-world applications, we include it here for transparency and explanatory purposes.

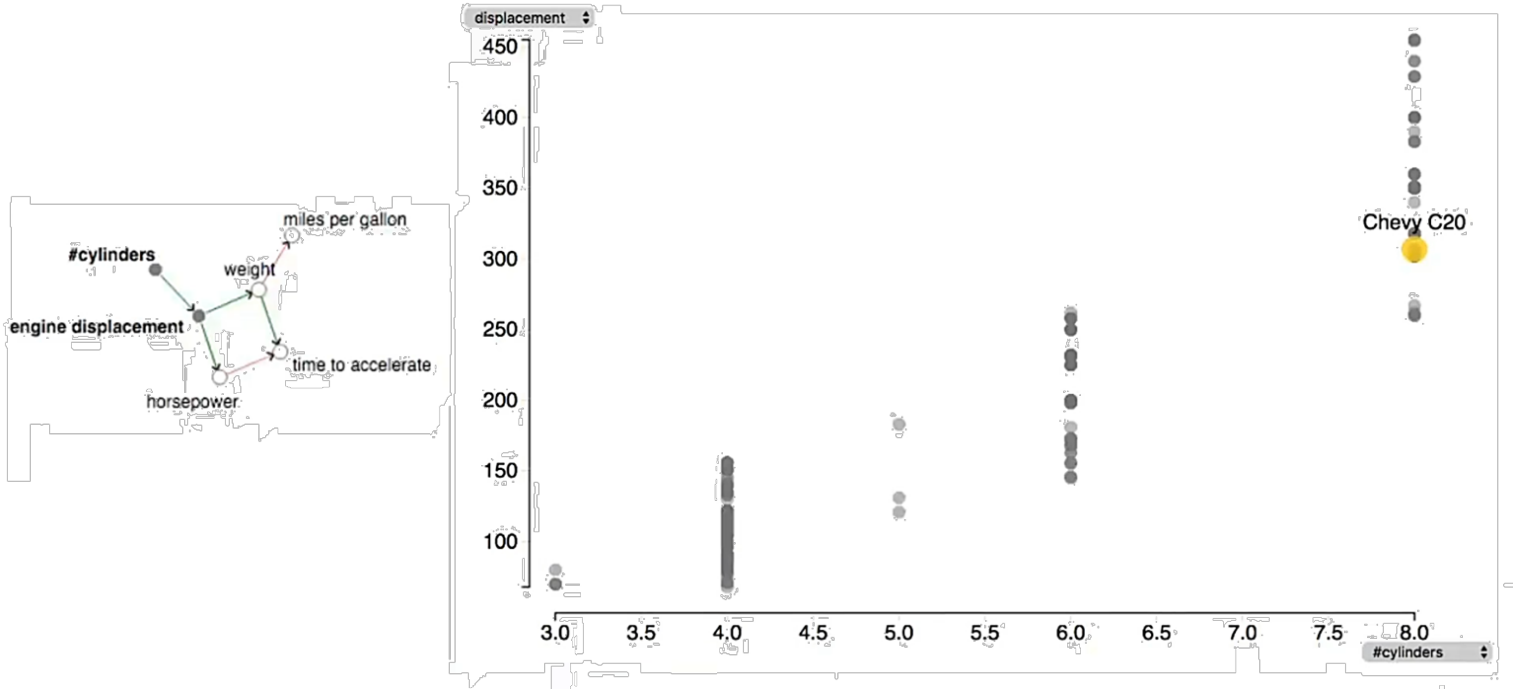


Our first example presents an **exhaustive sequence** of all causal relationships identified in a dataset of 1980s cars using a depth-first search algorithm. The sequence is available for visualization in both scatterplot and parallel coordinates formats.

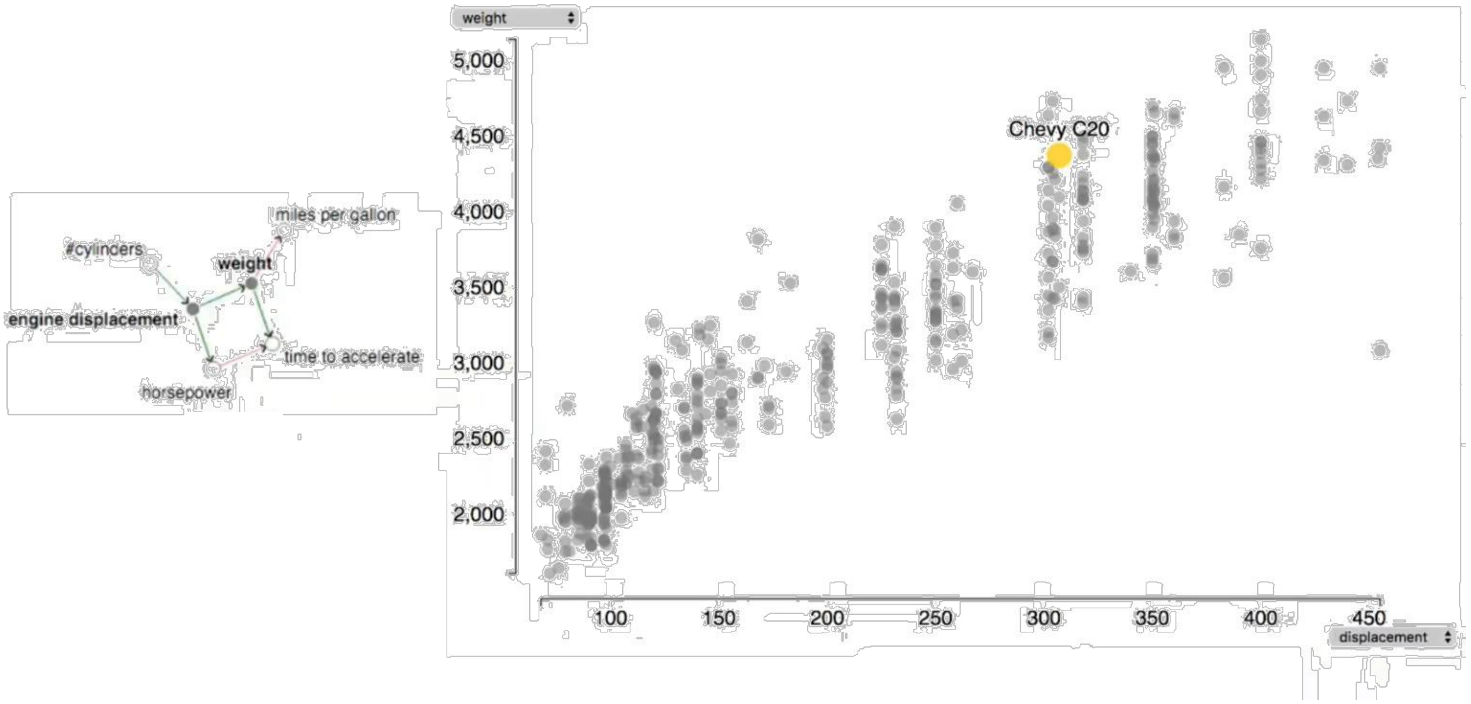


Shown here is part of a sequence of scatterplots for this dataset, specifically illustrating the causal relationship between the number of cylinders and engine displacement volume. The underlying causal network is displayed on the left, with the corresponding causal edge highlighted in the top-left corner.

(continued on the next page)

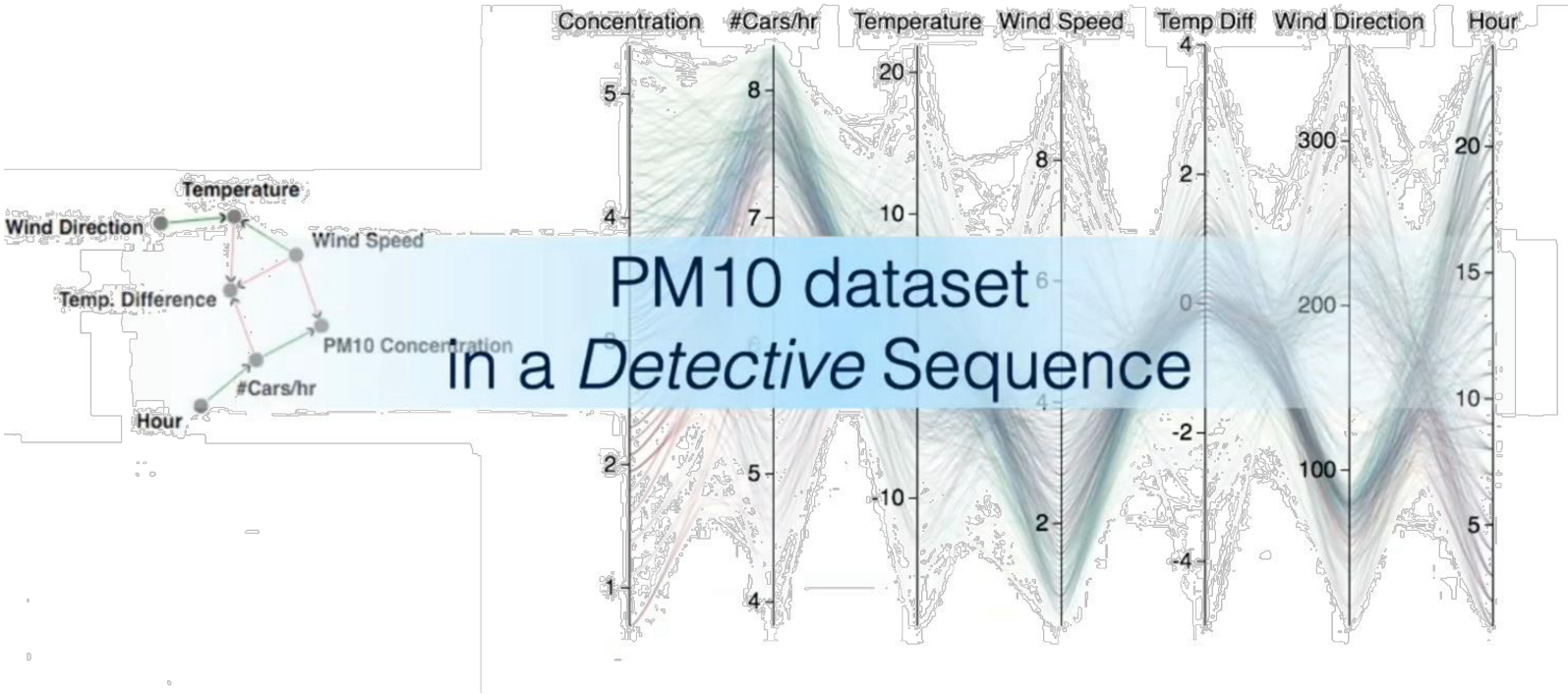


A sample data point—the Chevy C20—is highlighted in yellow to exemplify this relationship and it will serve as a running example throughout the sequence. In the scatterplot, the x-axis shows the cause—number of cylinders—and the y-axis shows the effect—displacement volume. We clearly observe the causal pattern encoded in the network: while there are variations in displacement among cars with the same number of cylinders, adding cylinders to an engine generally increases its displacement.



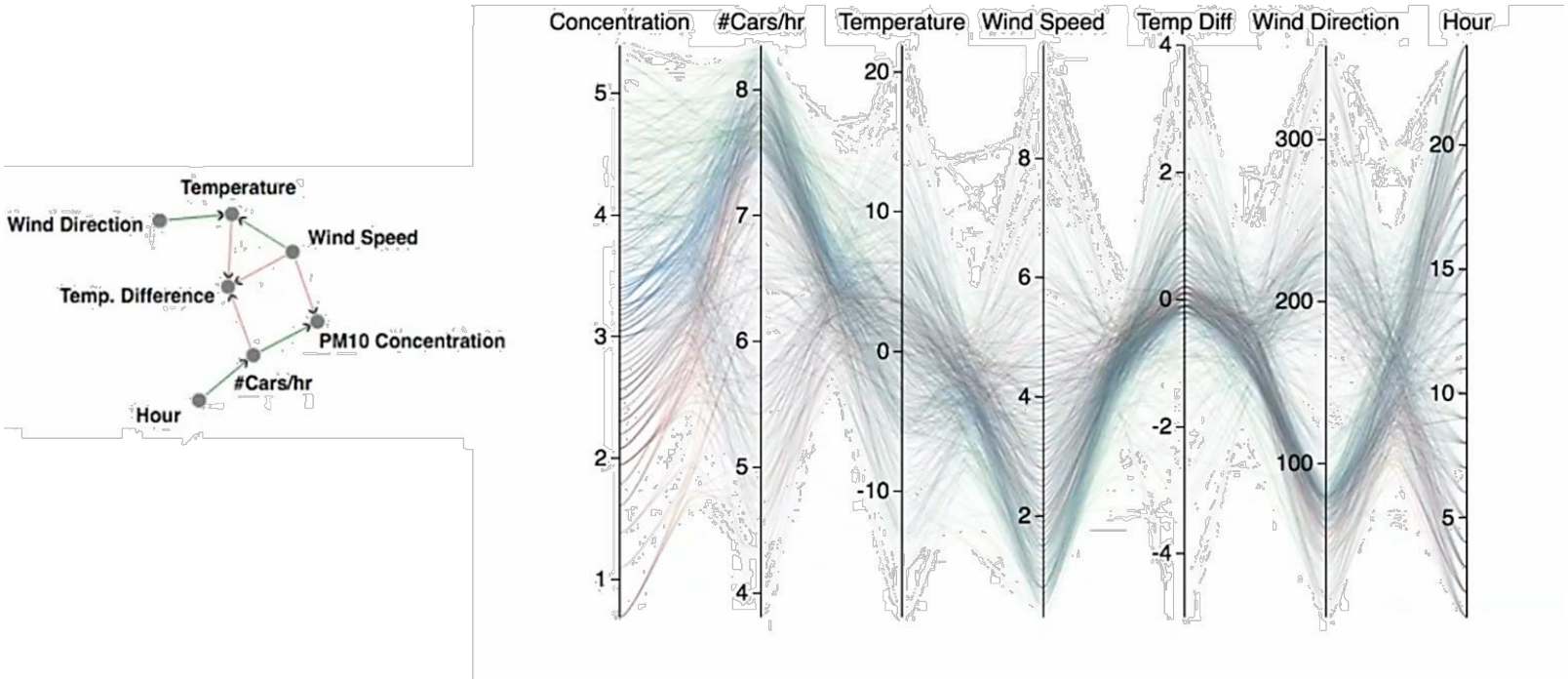
The sequence then continues to the next causal relationship, transitioning to the directed edge from engine displacement to car weight. Both are numerical variables, and here too we observe a clear pattern: larger engine displacement appears to cause higher weight. This reflects the underlying causal mechanism—cars with larger engines in the 1980s were typically built heavier, possibly to accommodate or complement the larger powertrain.

There are many more causal relations in this dataset and the exhaustive sequence will traverse through all of them.



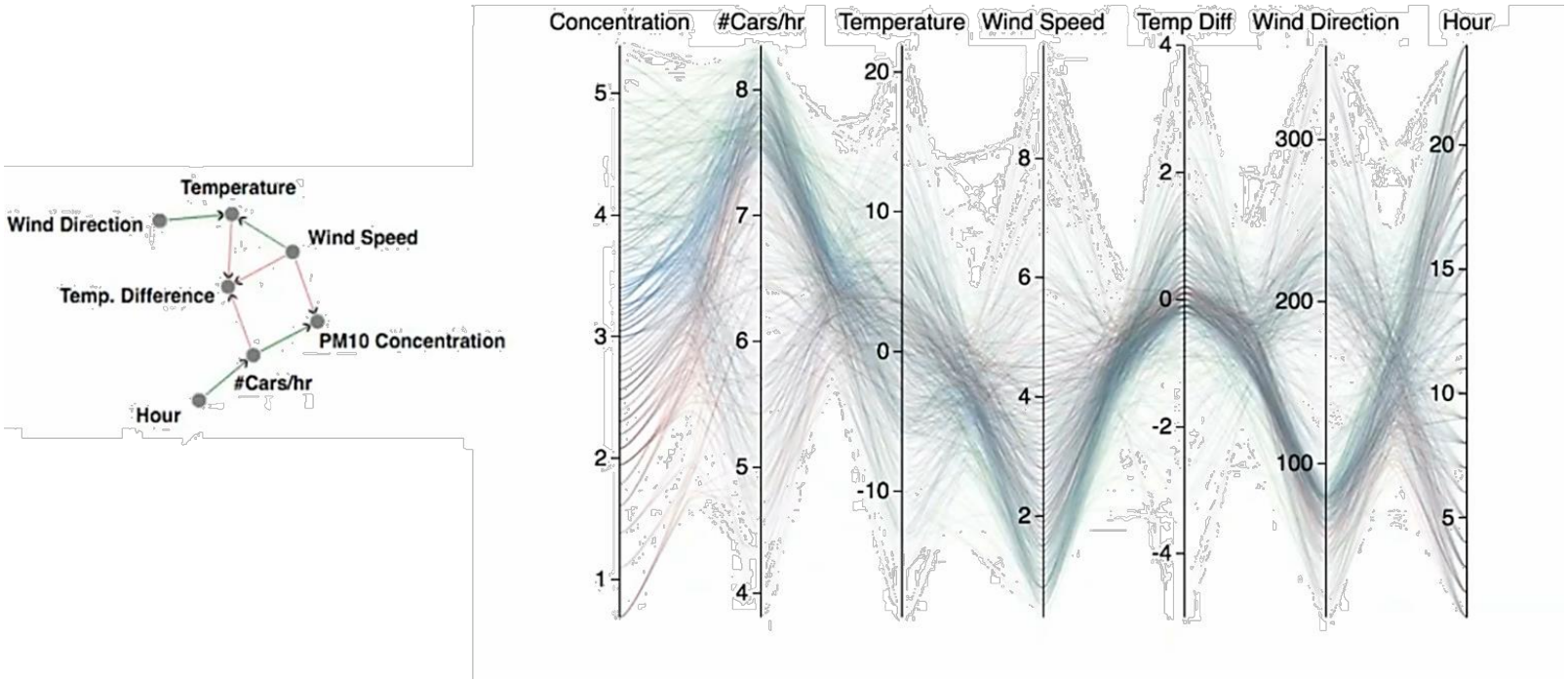
Next, we explore the **detective narrative sequence**. While similar to an exhaustive sequence, it reverses the typical cause-effect order. This approach mimics how detectives work—starting from an observed effect and tracing backward to potential causes. The goal is to make the sequence more engaging while filtering out unrelated causes to focus on the most plausible explanations.

We use the PM10 dataset as an illustrative example to investigate the possible causes of elevated concentrations of this air pollutant. Understanding these contributing factors is important, as high levels of particulate matter pose serious health risks.



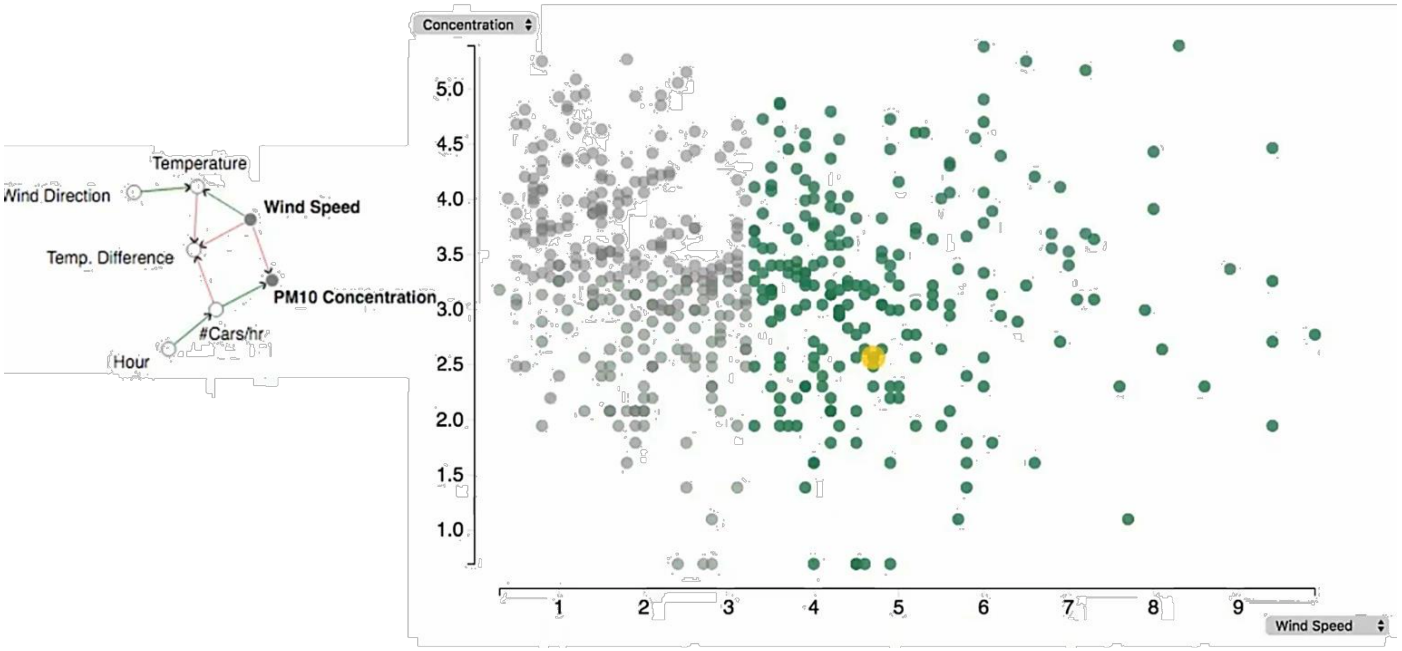
Causal analysis has found—see the network on the left—that elevated PM10 concentration can be directly caused by an increasing number of cars per hour, indicated by an incoming green edge, or it can be caused by decreasing wind speed, indicated by an incoming red edge. We use a mix of scatterplots and parallel coordinate plots to visualize these two competing causes and assess which one may be more influential.

(continue on next page)



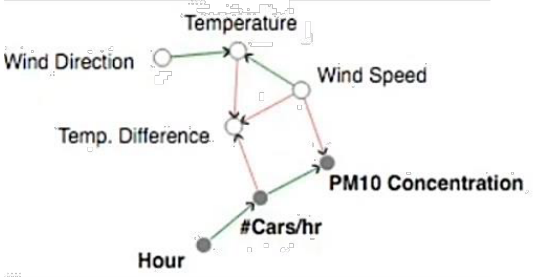
An advantage of Parallel Coordinates Plots (PCP) is that they can visualize multiple variables at once. In that way—in contrast to scatterplots—PCPs can show an entire causal chain in one visualization. Each curved line, also called polyline, is one multivariate environmental observation. To minimize abrupt visual changes, our algorithm sorts all chains by length and then swaps them to reduce local edit distances between adjacent paths. The polylines are also bundled and colored for better readability.

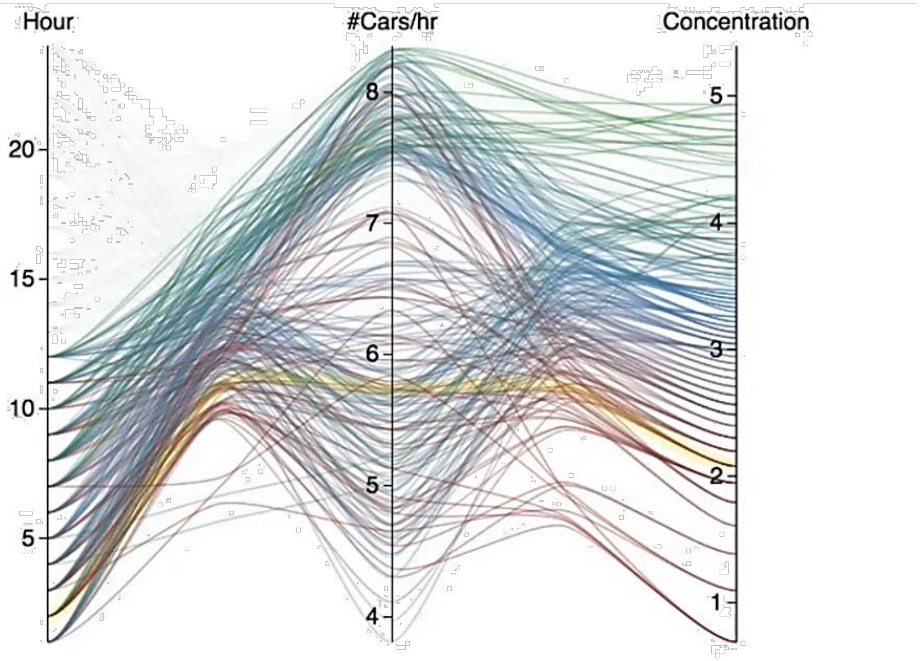
The next pages show the detective sequence of this dataset.



First, the algorithm shows the relation between wind speed and PM10 concentration, using a scatterplot since only two variables are involved This scatterplot of environmental observations reveals a negative trend, but the correlation is not strong—the observations are widely scattered about the trend.

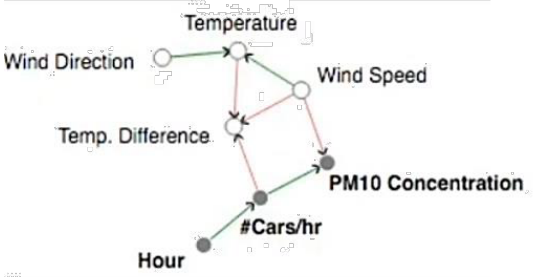
The coloring in this plot indicates whether samples meet (green) the goal polarity, i.e., low PM15 Concentration. Note that these colors differ from those used in the PCPs.

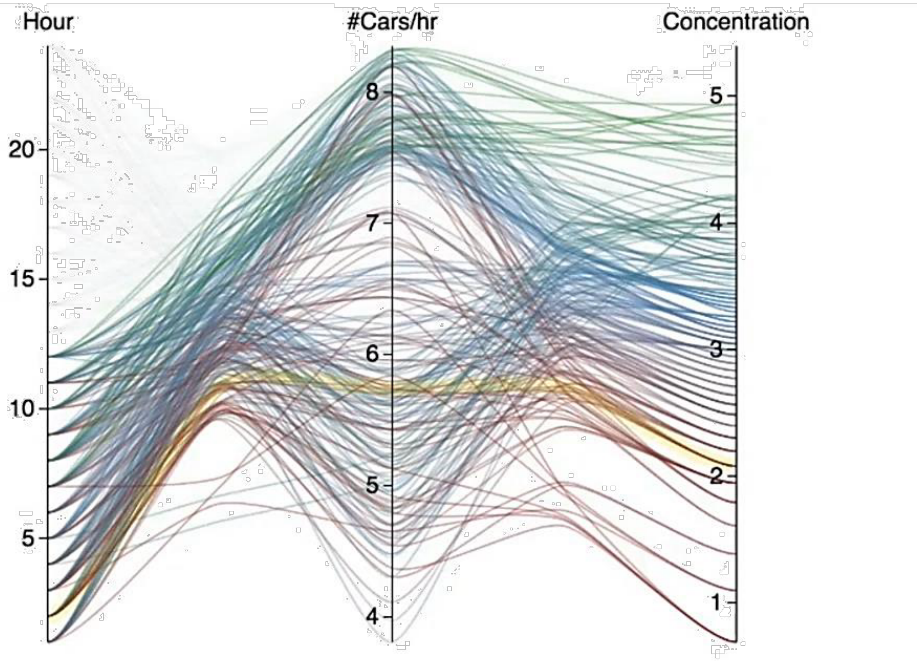




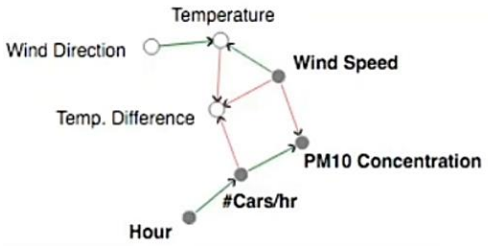
Since no other variable can affect wind speed, we now turn to the other causal path. This path forms a two-link causal chain, so the sequencing algorithm selects parallel coordinates to visualize it. The chain begins with the number of cars per hour, which itself can be influenced by the hour of the day. In other words, the hour of the day indirectly affects PM10 concentration through its impact on traffic volume.

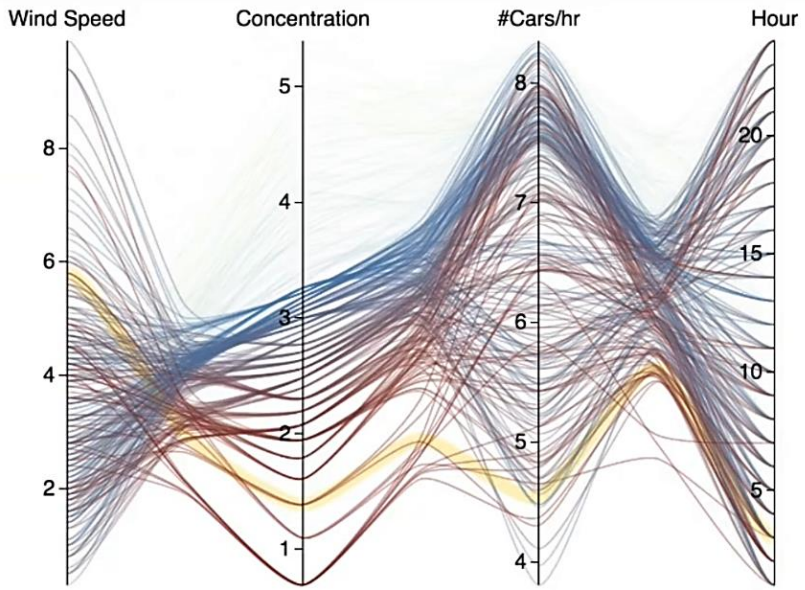
(continued on the next page)





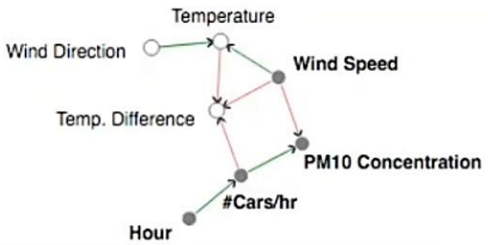
We observe a clear causal pattern between the number of cars and high PM10 concentration: the green lines and the upper bundles of blue lines are generally aligned. However, for lower pollution levels, the trend is less clear—many lines cross, indicating more variability. On the other hand, the hour of the day is a strong predictor of the number of cars per hour, though the distribution appears to be bimodal.

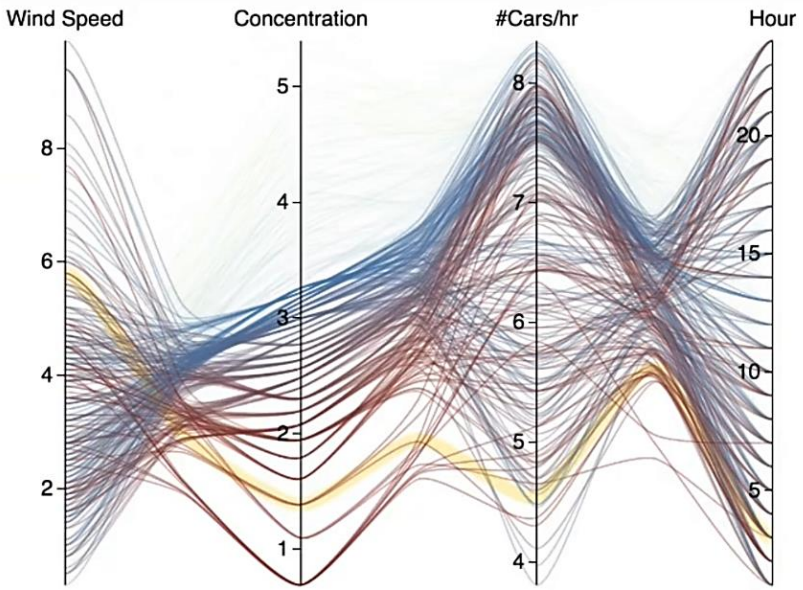




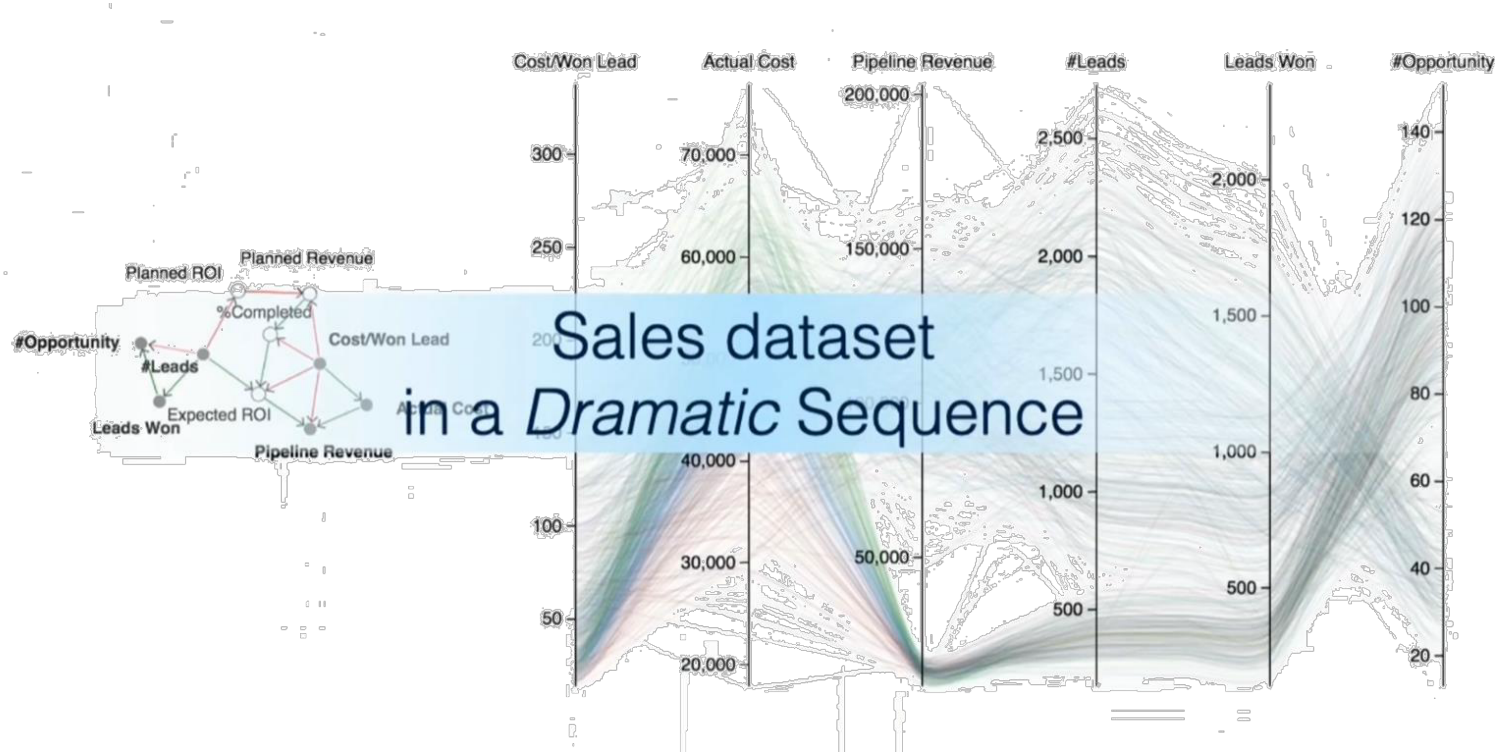
The narrative sequence concludes by presenting all causal relationships that affect our variable of interest—PM10 concentration—once again using parallel coordinates. As discussed, this format is particularly well-suited for depicting complex causal chains in a story-like, sequential fashion. The outcome—or effect variable—PM10 concentration is placed at the center, with the two colliding causal chains branching to the left and right. This layout highlights the multiple, potentially competing influences on pollution levels.

(continued on the next page)

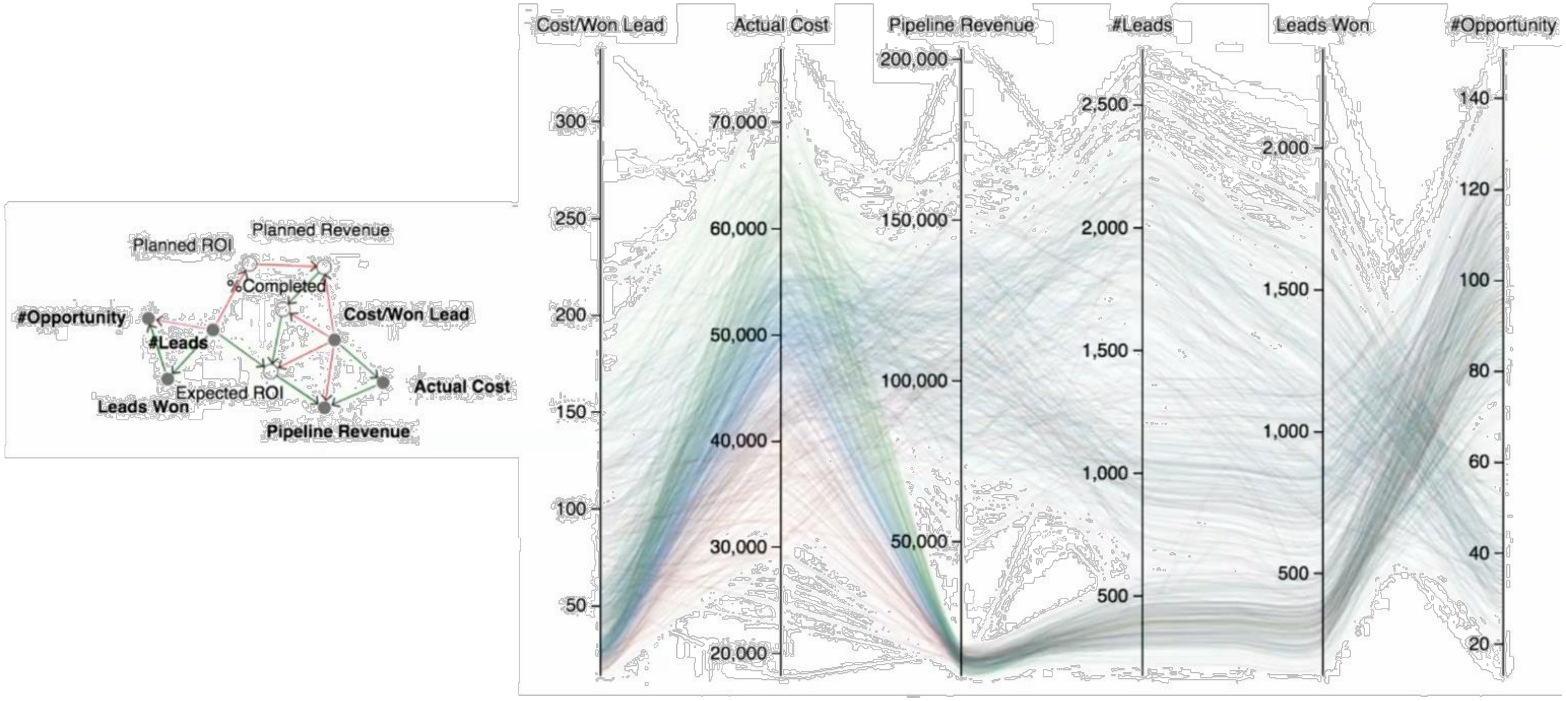




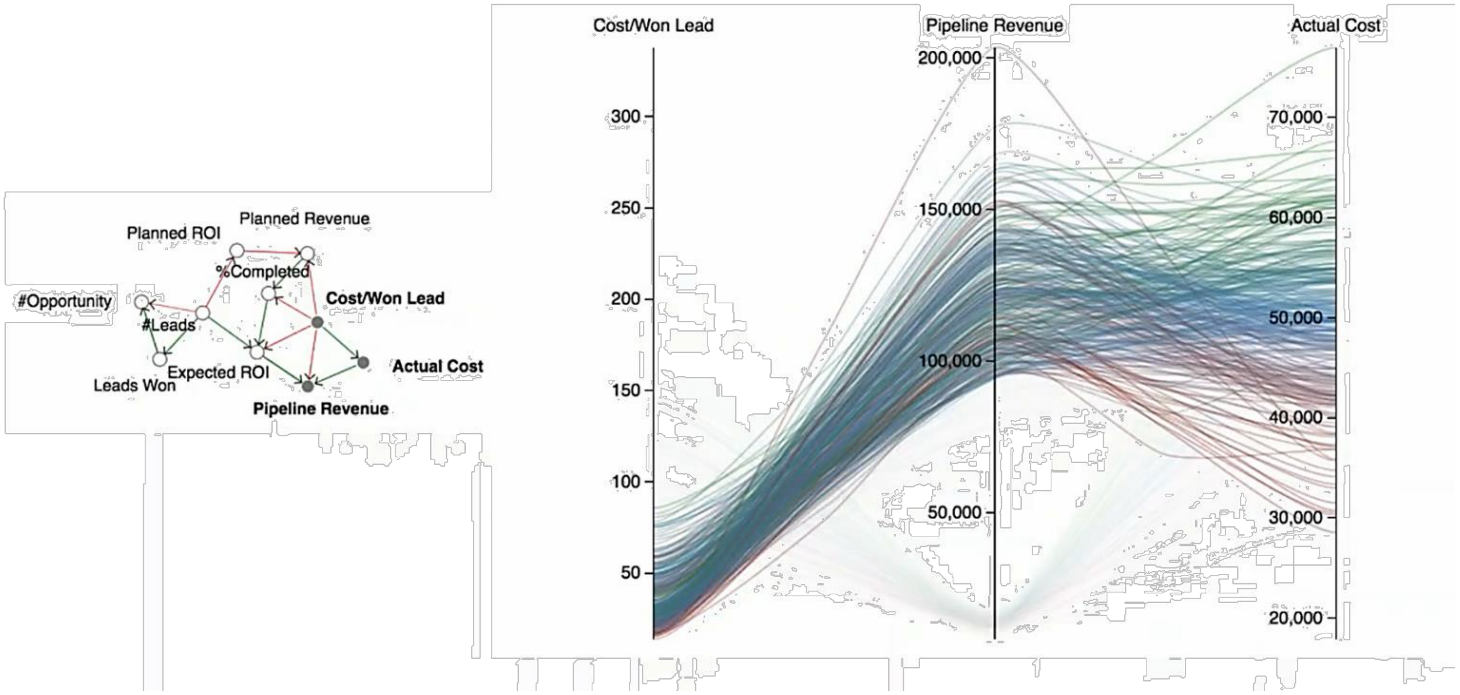
Now, the analyst—the detective—is tasked with weighing the relative impact of each path, tracing back through the two causal sequences. The yellow line highlights a specific observation, inviting closer inspection of how particular values along each causal chain may have contributed to the observed pollution level. This kind of analysis not only deepens the analyst's understanding of the data but can also guide targeted interventions—for example, by clarifying whether reducing traffic or addressing atmospheric conditions would be more effective in lowering pollution in a given scenario.



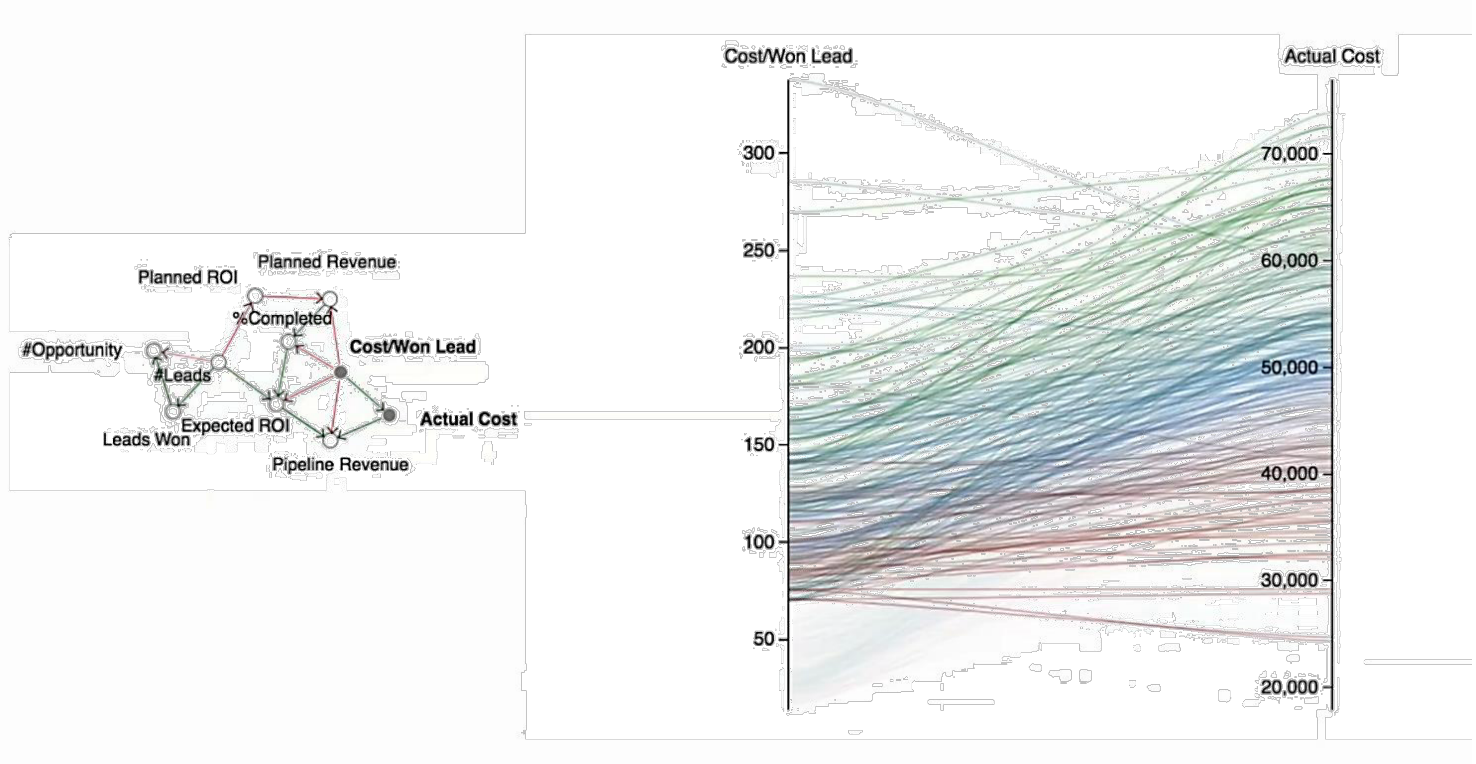
Next, we explore the **dramatic sequence**, which borrows the idea of conflict from the five-act dramatic structure. Here, the notion of conflict is mapped to a causal collider in the graph—where two or more causes converge on a common effect.



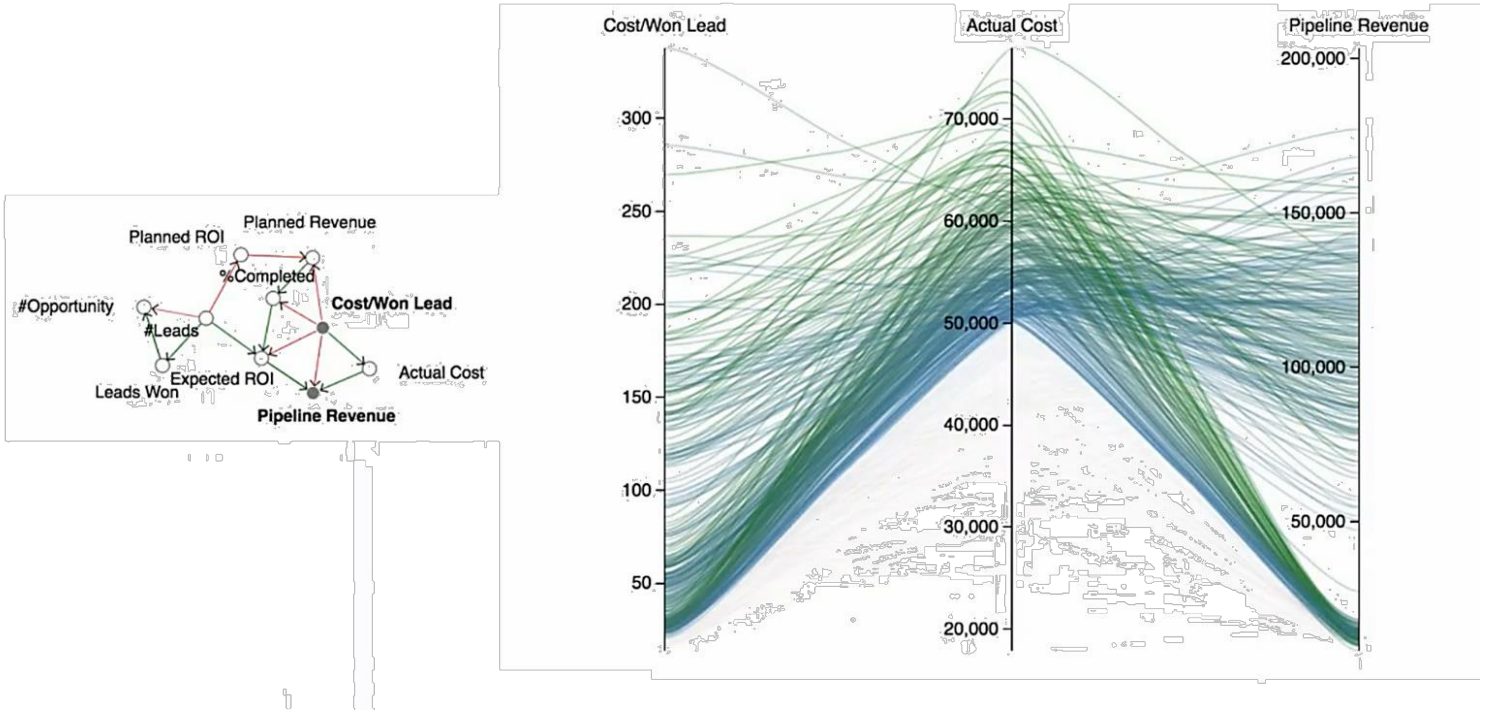
The sequence is built around this point of tension, inviting the viewer to consider how competing influences interact to shape an outcome. We use the Sales campaign dataset to illustrate this narrative style. In the corresponding causal graph, two conflicts are highlighted as triangular motifs.



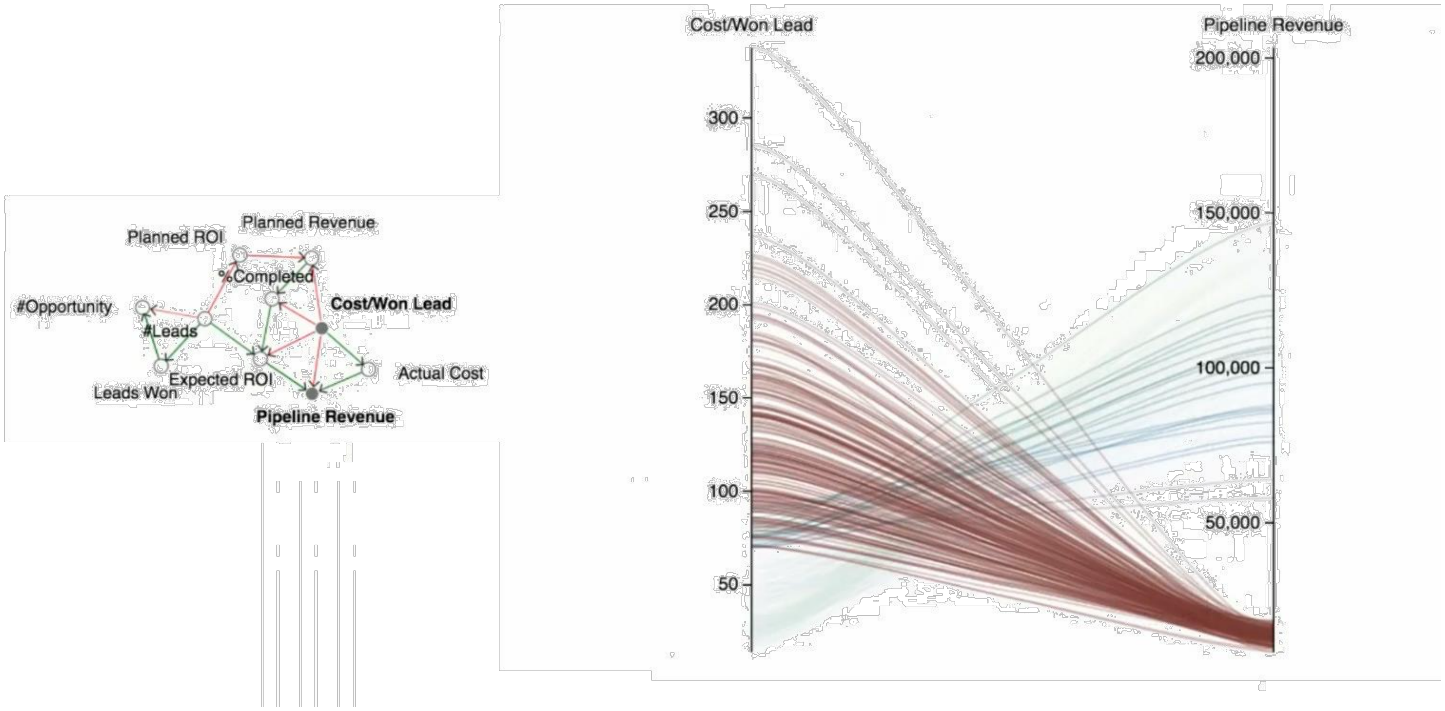
One such conflict involves three variables: pipeline revenue, actual cost, and cost per won lead, located in the lower right of the graph. In the parallel coordinate plot, these three variables are aligned with pipeline revenue—the effect variable and focal point of the conflict—placed at the center. The other two variables, cost per won lead and actual cost, represent competing causal influences. Their conflicting relationships with pipeline revenue are visually expressed through crossing lines and diverging trends, prompting the viewer to reason through the tradeoffs and the underlying causal structure. While pipeline revenue is positively influenced by actual cost and negatively by cost per won lead, it's important to note that cost per won lead also affects actual cost.



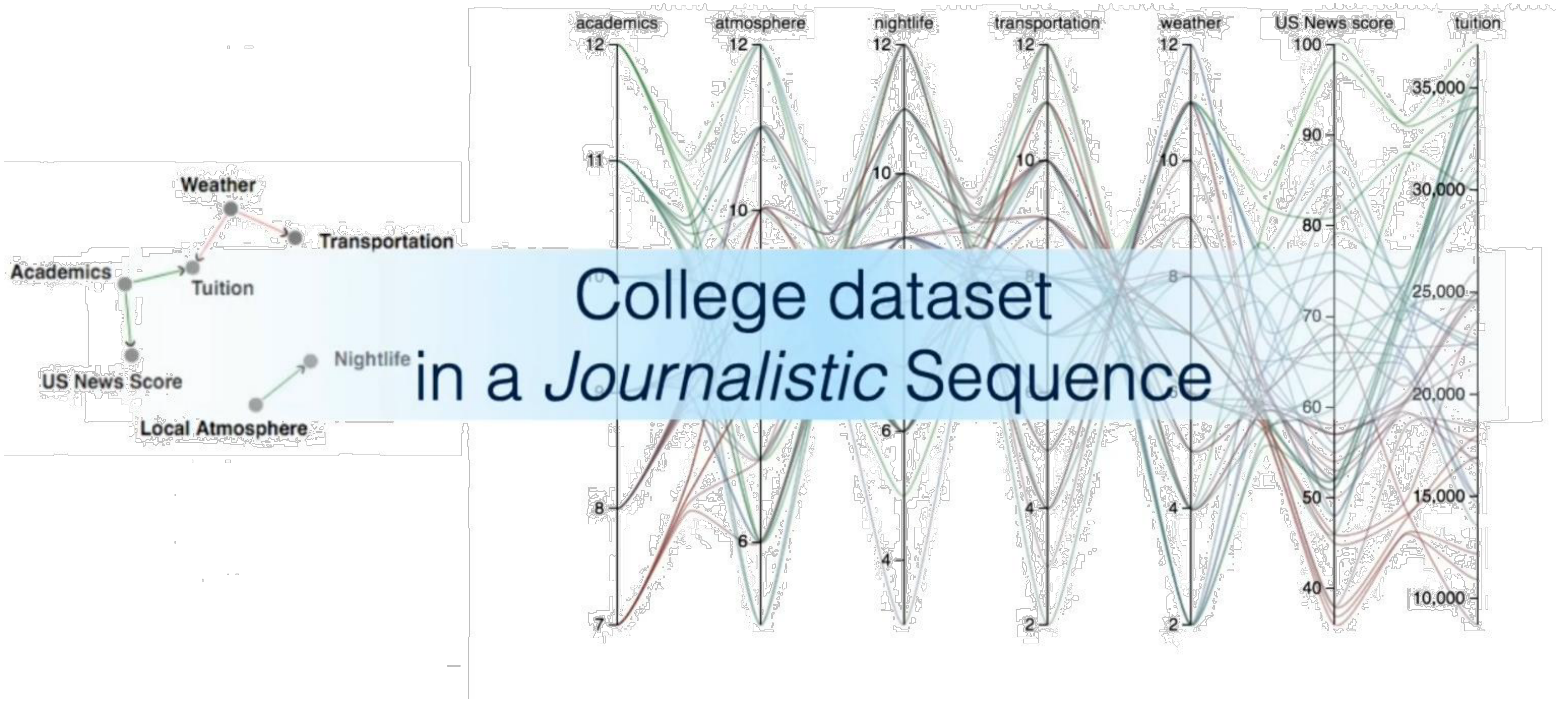
The next plot in the narrative aims to unpack this relationship by focusing on the interplay between these two conflict variables. In the parallel coordinate plot, a clear causal link from cost per won lead to actual cost is visible, as evidenced by the consistent pattern of lines and minimal crossing—suggesting a strong and direct dependency.



The next plot in the narrative reintroduces pipeline revenue—the focal point of the conflict—into the parallel coordinate display. This extended view reveals two distinct processes at play. In the first process, we observe low cost per won lead, followed by low actual cost, but also resulting in low pipeline revenue. In the second process, cost per won lead and actual cost remain similarly low, yet pipeline revenue ranges from medium to high. This divergence suggests that additional, unobserved factors may be driving more favorable outcomes, even when cost-related variables are held in check. These patterns highlight the complexity of the underlying causal dynamics and point to potential confounders or threshold effects that warrant further investigation.

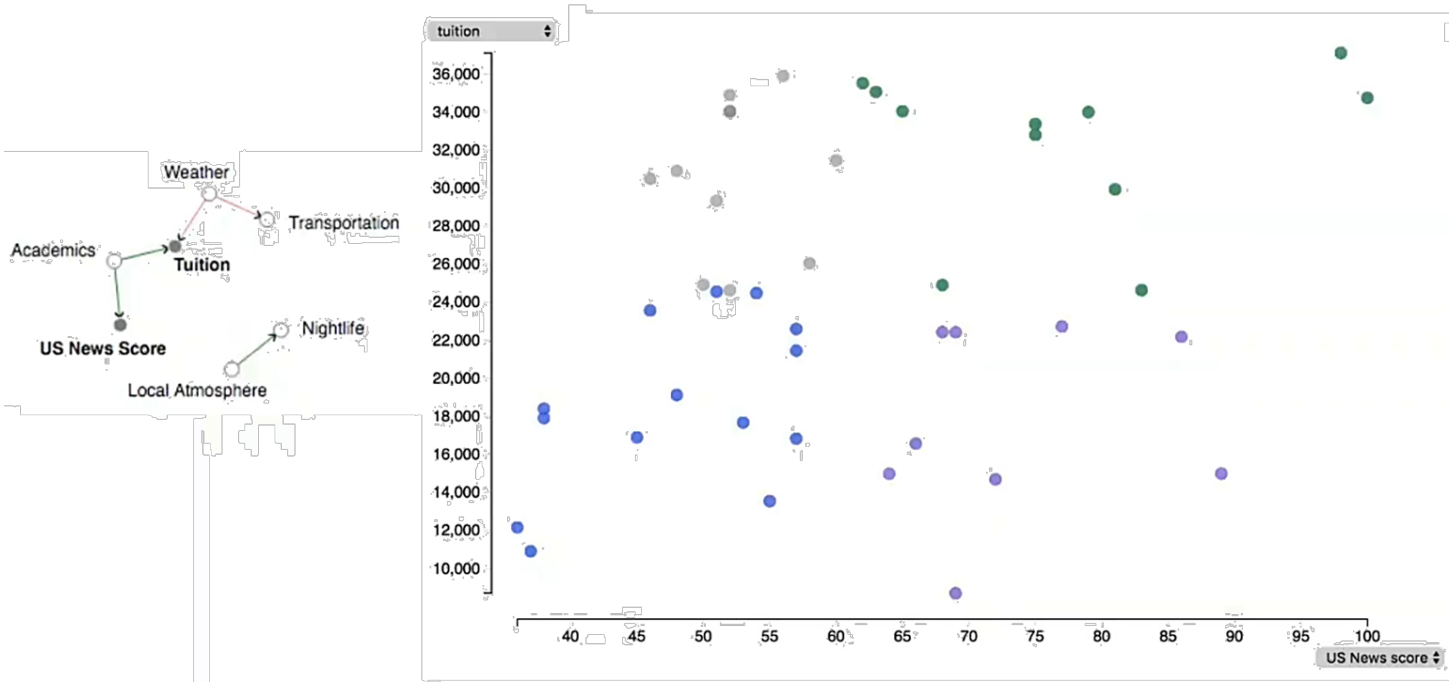


The final plot in the narrative isolates the relationship between cost per won lead and pipeline revenue, confirming the presence of an intermediate cost range where pipeline revenue exhibits diverging trends. This reinforces the earlier observation that similar cost levels can lead to markedly different outcomes—suggesting the influence of additional moderating factors or hidden variables that should prompt further investigation.



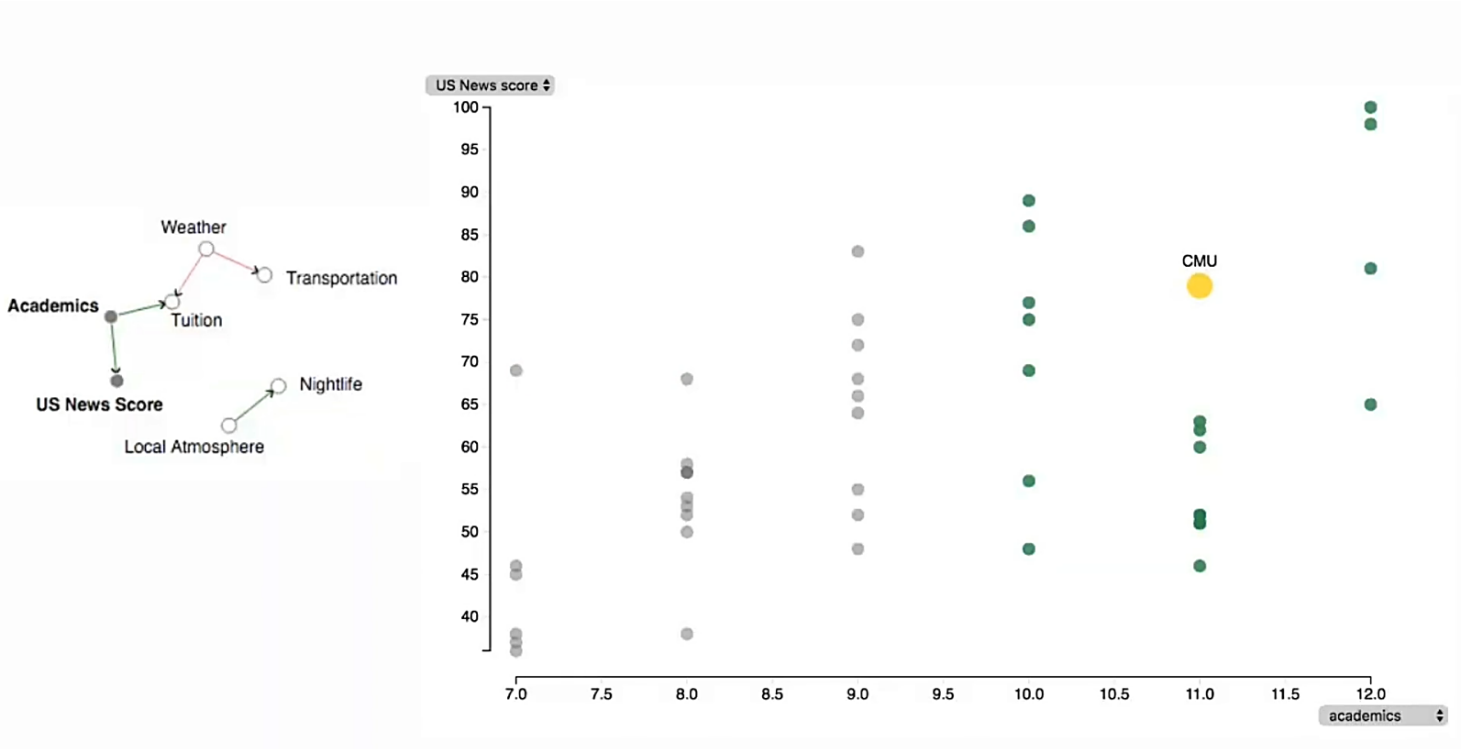
Last, we explore the **journalistic narrative**. A journalistic sequence presents the causal conflict as a trade-off between different parties which might be incompatible, similar to how a journalist would write a balanced story.

We use the College dataset for illustration. Ideally, a student would seek a school that combines a high U.S. News score with low tuition—maximizing quality while minimizing cost. However, such combinations are rare, and tradeoffs are often necessary. In many cases, improving one desirable attribute (e.g., ranking) comes at the expense of another (e.g., affordability), highlighting the tension between competing priorities in college selection.

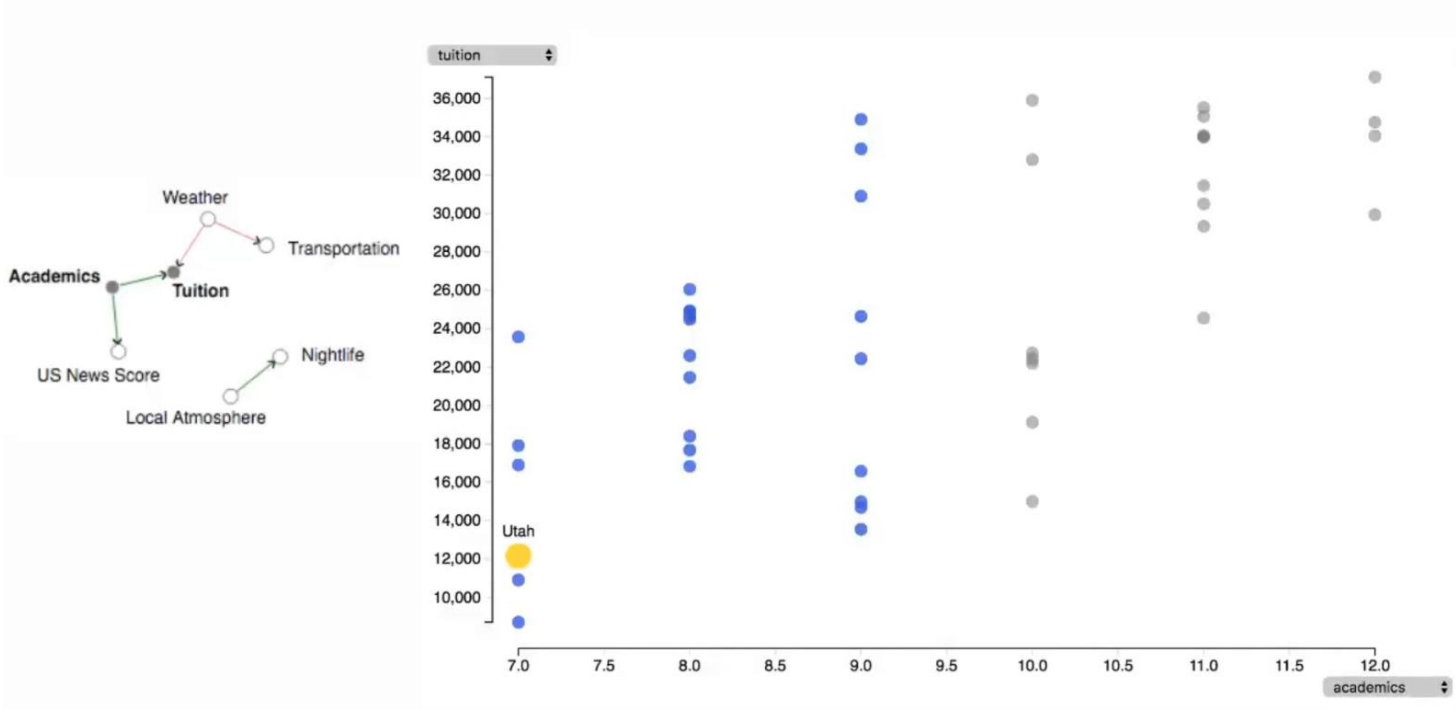


The first plot in this sequence visualizes the tradeoff, using a scatterplot to map U.S. News score against tuition. Interestingly, the relationship between these two variables is only modestly correlated, suggesting that the tradeoff is not absolute—and that there may be opportunities to identify schools offering both strong academic reputation and reasonable cost. This insight is reinforced by the underlying causal graph, which shows that U.S. News score and tuition are not connected by a direct causal link. Instead, their association appears to be mediated through other variables.

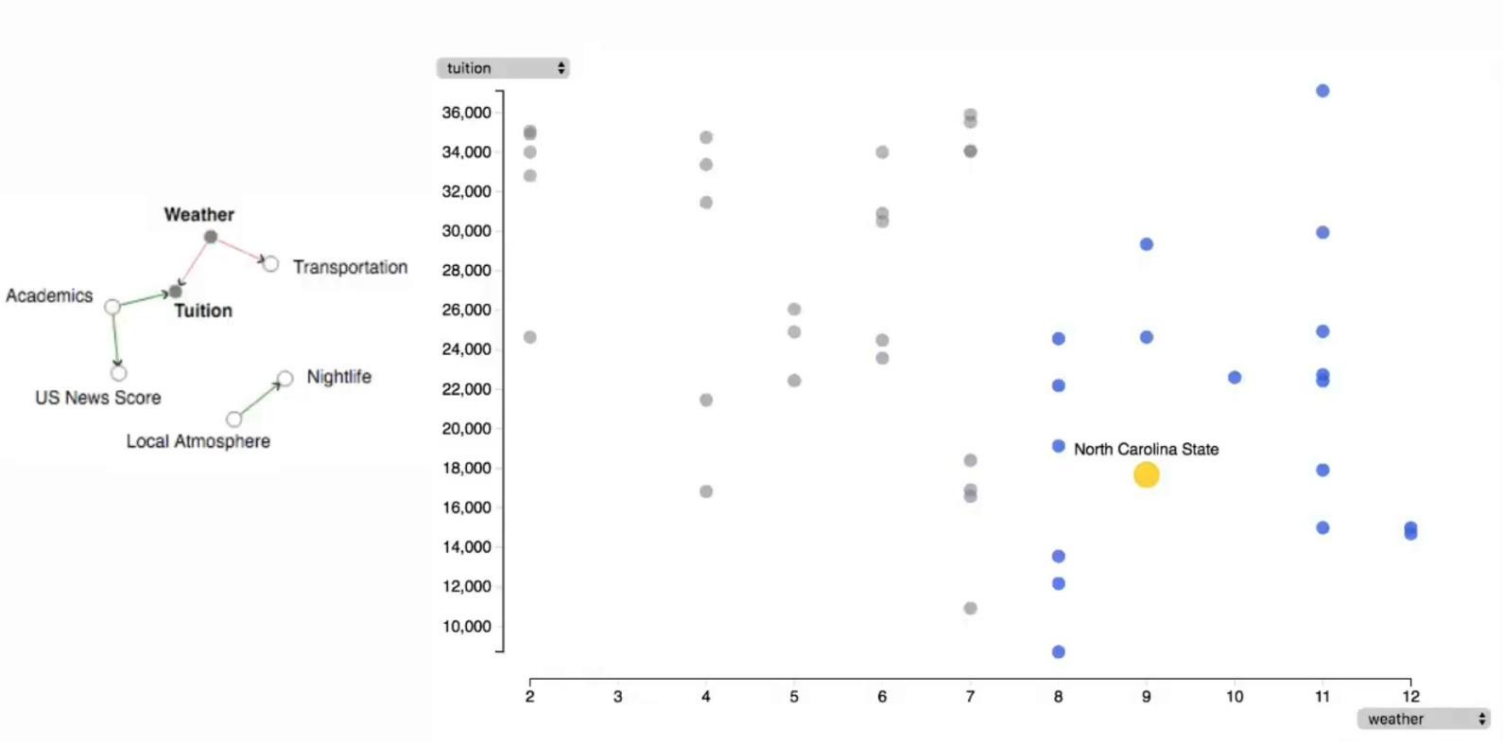
The remaining sequence of plots investigates these mediating factors and the underlying drivers of the tradeoff U.S. News score vs tuition cost in more detail.



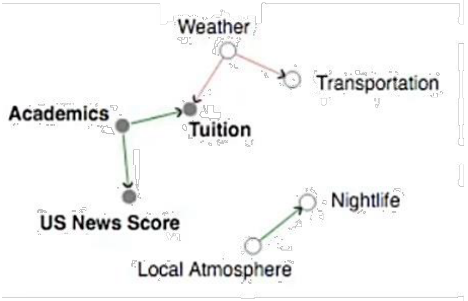
The first of these plots examines the relationship between U.S. News score and Academics score, revealing that while there is a positive trend, the relationship is moderate, with considerable variation at each level of academic quality. This suggests that other factors likely contribute to a school's overall ranking. To illustrate the connection concretely, Carnegie Mellon University is highlighted (in yellow) as a representative case—standing out with both a high Academics score and a correspondingly high U.S. News score.

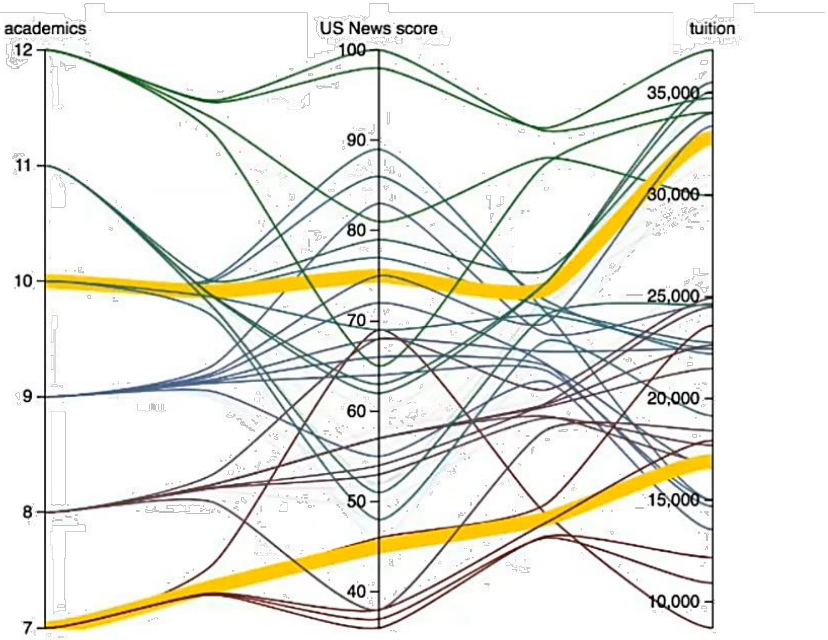


In the following plot, we examine how the Academics score also influences tuition, indicating that stronger academic programs often come at a higher financial cost. However, the plot also reveals considerable variation in tuition levels for a given academic quality, suggesting that academic strength is not the sole driver of cost. The University of Utah is highlighted as an illustrative example—demonstrating a relatively modest Academics score paired with correspondingly low tuition.



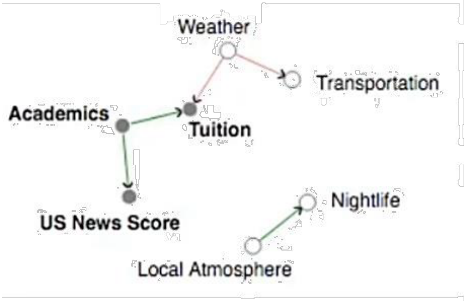
Another variable that affects tuition—though only in a moderate way—is the Weather score. While not a primary driver, it appears to have some influence on cost, possibly reflecting regional or lifestyle-related factors. North Carolina State University serves as an example in this context, exhibiting a high Weather score alongside low tuition, suggesting that favorable climate conditions do not necessarily come at a financial premium.

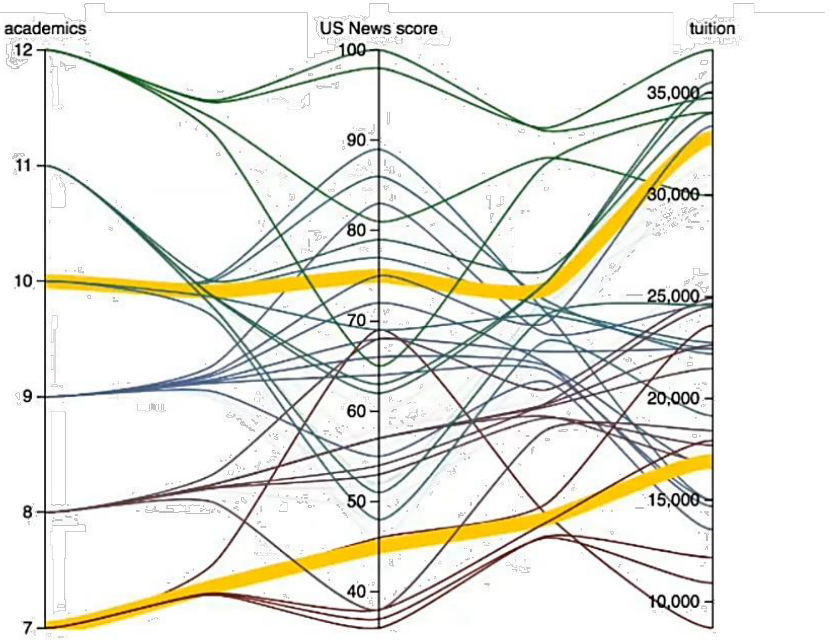




The final parallel coordinate plot brings the journalistic narrative to a close by revisiting the two key tradeoff variables: U.S. News score and tuition. It also shows the two incompatible choices, Carnegie Mellon University and University of Utah as yellow polylines.

(continued on the next page)





This summary view reinforces a central theme uncovered throughout the sequence—namely, that Academics score generally drives both U.S. News score and tuition upward. However, the plot also reveals considerable fluctuations along both outcome variables, highlighting the influence of additional factors, such as Weather score and other attributes discussed earlier. These variations underscore the complexity of the tradeoff and suggest that nuanced, multi-variable reasoning is essential when evaluating college options.

More detail on algorithms, methodology, use cases, and validation can be found in the paper

Generating Coherent Visualization Sequences for Multivariate Data by Causal Graph Traversal

authored by Puripant Ruchikachorn, Darius Coelho, Jun Wang, Kristina Striegnitz*, and Klaus Mueller

Stony Brook University, *Union College

Generating Coherent Visualization Sequences for Multivariate Data by Causal Graph Traversal

Puripant Ruchikachorn, *Member, IEEE*, Darius Coelho, Jun Wang, *Member, IEEE*, Kristina Striegnitz, and Klaus Mueller, *Fellow, IEEE*

Abstract—Multivariate data contain an abundance of information and many techniques have been proposed to allow humans to navigate this information in an ordered fashion. For this work, we focus on methods that seek to convey multivariate data as a collection of bivariate scatterplots or parallel coordinates plots. Presenting multivariate data in this way requires a regime that determines in what order the bivariate scatterplots are presented or in what order the parallel coordinate axes are arranged. We refer to this order as a visualization sequence. Common techniques utilize standard statistical metrics like correlation, similarity or consistency. We expand on the family of statistical metrics by incorporating the rigidity of causal relationships. To capture these relationships, we first derive a causal graph from the data and then allow users to select from several semantic traversal schemes to derive the respective chart sequence. We tested the sequences with a crowd-sourced user study and a user interview to confirm that the causality-informed visualization sequences help viewers to better grasp the relationships that exist in the data, as opposed to sequences derived from correlations or randomization alone.

Index Terms—Causality, Causal Graph, Visualization Sequence, Multivariate Visualization, Parallel Coordinates

1 Introduction

Multivariate data analysis can enable researchers and practitioners to uncover meaningful insights, make informed decisions, and optimize processes. Examples include analyzing healthcare data to identify associations between patient demographics, medical history, and treatment outcomes; investigating consumer behavior data to uncover connections between customer demographics, product preferences, and marketing campaigns; and exploring environmental data to understand the impact of climate change on temperature, precipitation, and species diversity.

A prime goal in multivariate data analysis is to uncover the intricate relationships that exist among the variables, conveying a comprehensive understanding of the complex system that underlies the data. An attractive method for this purpose is to learn a causal model from the data. Causal models can effectively elucidate relationships and interdependencies that exist among the variables and are easily visualized as node-link diagrams. While these models can reveal more semantics than simple summary statistics, they do not allow the analyst to detect irregularities, outliers, and unusual data patterns. Conversely, visualization capitalizes on the unsurpassed power of the human visual system to quickly recognize these types of irregularities.

There is a large arsenal of visualization methods, at a wide gamut of complexity. A bivariate scatterplot is the simplest visualization that can give a viewer insight into data relations, but among two variables only. A Parallel Coordinates Plot (PCP) [1] draws each data point as a line across several parallel data axes and as such can visualize more than two variables. Other techniques for multivariate data employ data embeddings, such as MDS [2], t-SNE [3], and UMAP [4], to name the most prominent, but their projections primarily focus on the depiction of clusters and neighborhood relations.

The limited screen space restricts extending bivariate scatterplots to include more than two variables. Therefore, to show more relationships, one might create a sequence of scatterplots but there are numerous combinations in which a set of scatterplots can be ordered. For example, with 6 variables, we get 15 unique scatterplots and $15!$ sequence permutations, excluding duplicates. Likewise, there are numerous ways in which 6 axes can be arranged in a PCP, amounting to $6!$ possible axes sequences.

Bridging the gap between causal inference and data visualization, we introduce a novel approach that leverages a causal network to create a coherent sequence of bivariate scatterplots or PCP axes. The order in which the plots are presented ensures that each visualization builds upon the previous one, progressively unveiling the underlying causal structure with visual evidence. Viewers follow the sequence from start to finish, thereby grasping the intricate relationships in multivariate data. As a causal graph can be traversed in many ways, multiple causal sequences can be created, each uniquely narrating the same data.

Our paper is organized as follows. Section 2 lists some related work. Sections 3 and 4 present our narrative graph traversal schemes and prototype to create and edit the resulting sequences. Section 5 presents quantitative and qualitative validations. Section 6 and 7 discuss the result, limitations of our work, and some future work.

- *Puripant Ruchikachorn, Darius Coelho, Jun Wang, and Klaus Mueller are or were with the Visual Analytics and Imaging Laboratory, Department of Computer Science, Stony Brook University, Stony Brook, NY. Email: {pruchikachor, dcoelho, junwang7, mueller}@cs.stonybrook.edu.*
- *Puripant Ruchikachorn is now with the Chulalongkorn Business School, Chulalongkorn University, Bangkok, Thailand. Email: puripant@cbs.chula.ac.th.*
- *Kristina Striegnitz is with the Computer Science Department, Union College, Schenectady, NY. Email: striegnk@union.edu.*