

Paul Rosen

paul.rosen@utah.edu
@paulrosenphd
<https://cspaul.com>



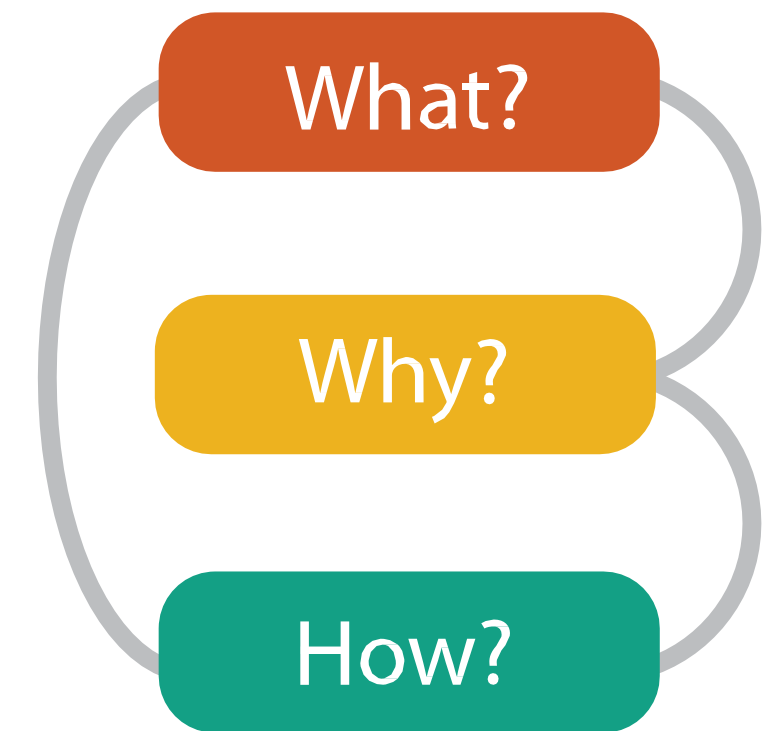
Visualization for Data Science

DS-4630 / CS-5630

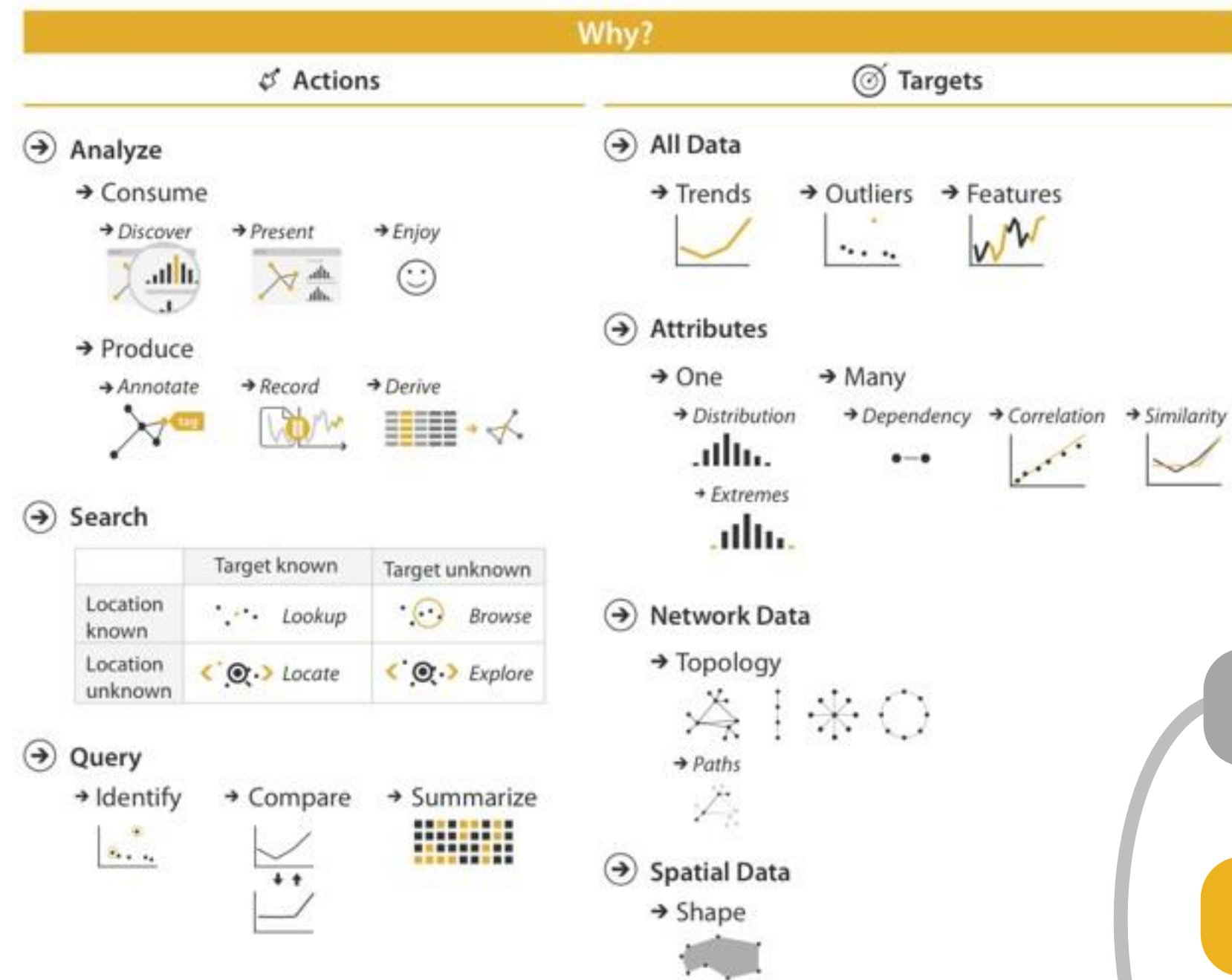
TASKS & INTERACTION

High-level Task Taxonomy: what, why, and how

- **what** is shown?
- **why** is the user looking at it?
- **how** is it shown?
 - abstract vocabulary avoids domain-specific terms
 - what-why-how analysis framework as scaffold to think systematically about design space



task abstraction



What?

Why?

How?

{action, *target*} pairs

- discover distribution
- compare trends
- locate outliers
- browse topology

→ Analyze

{action, target}

→ Search

→ Query

{action, target}

→ Analyze

→ Consume

→ Discover



→ Present



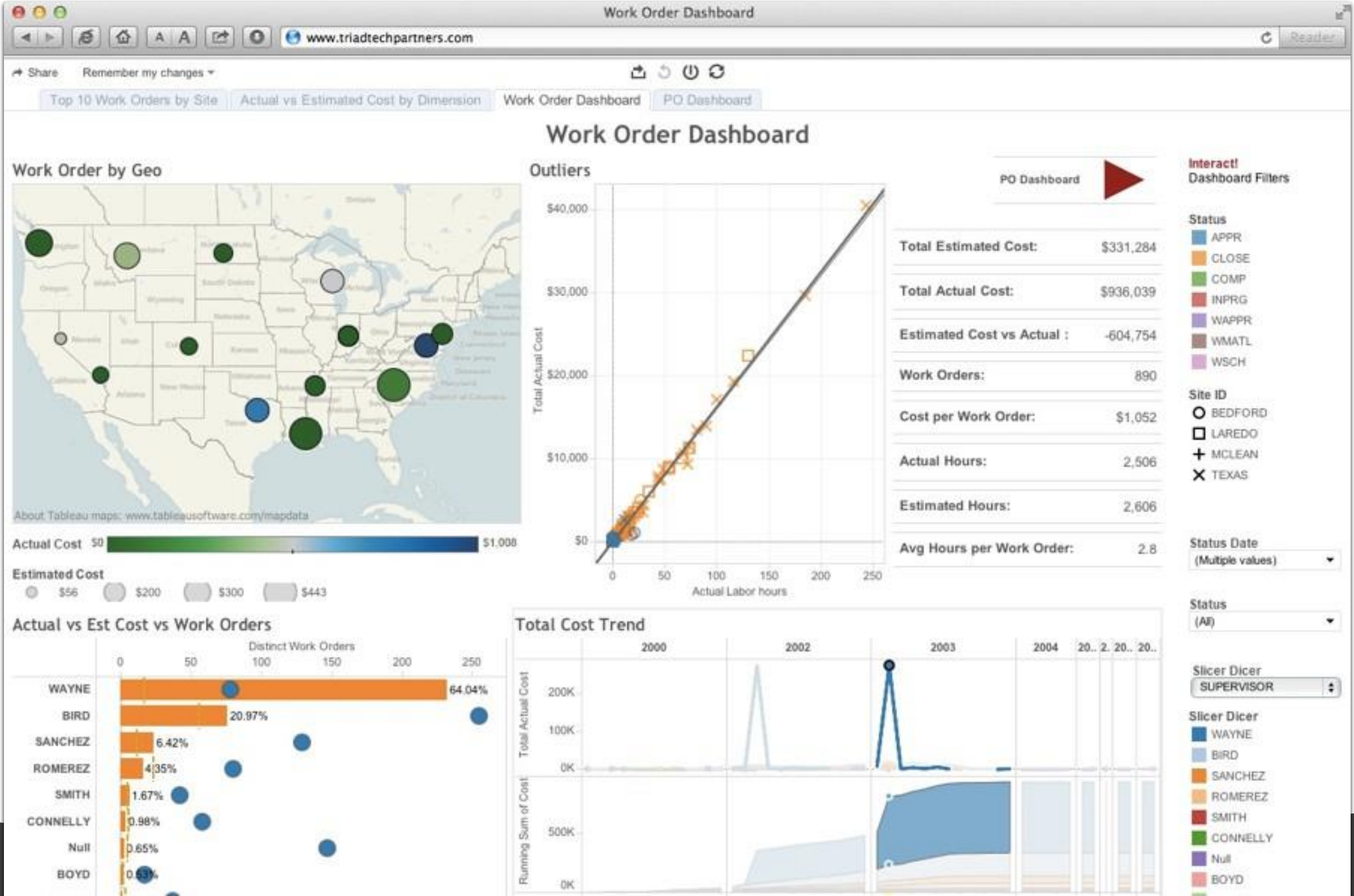
→ Enjoy



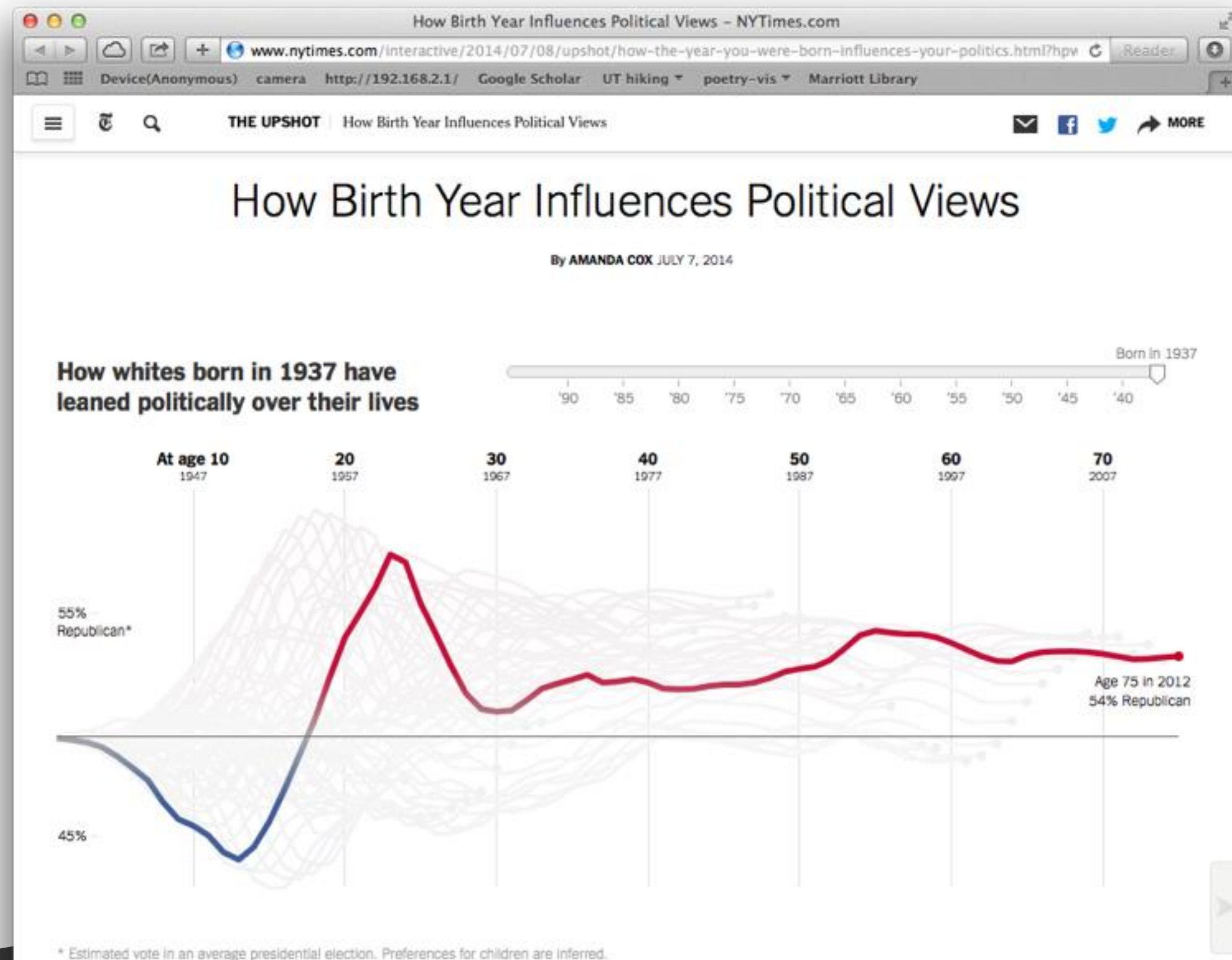
→ Search

→ Query

discover

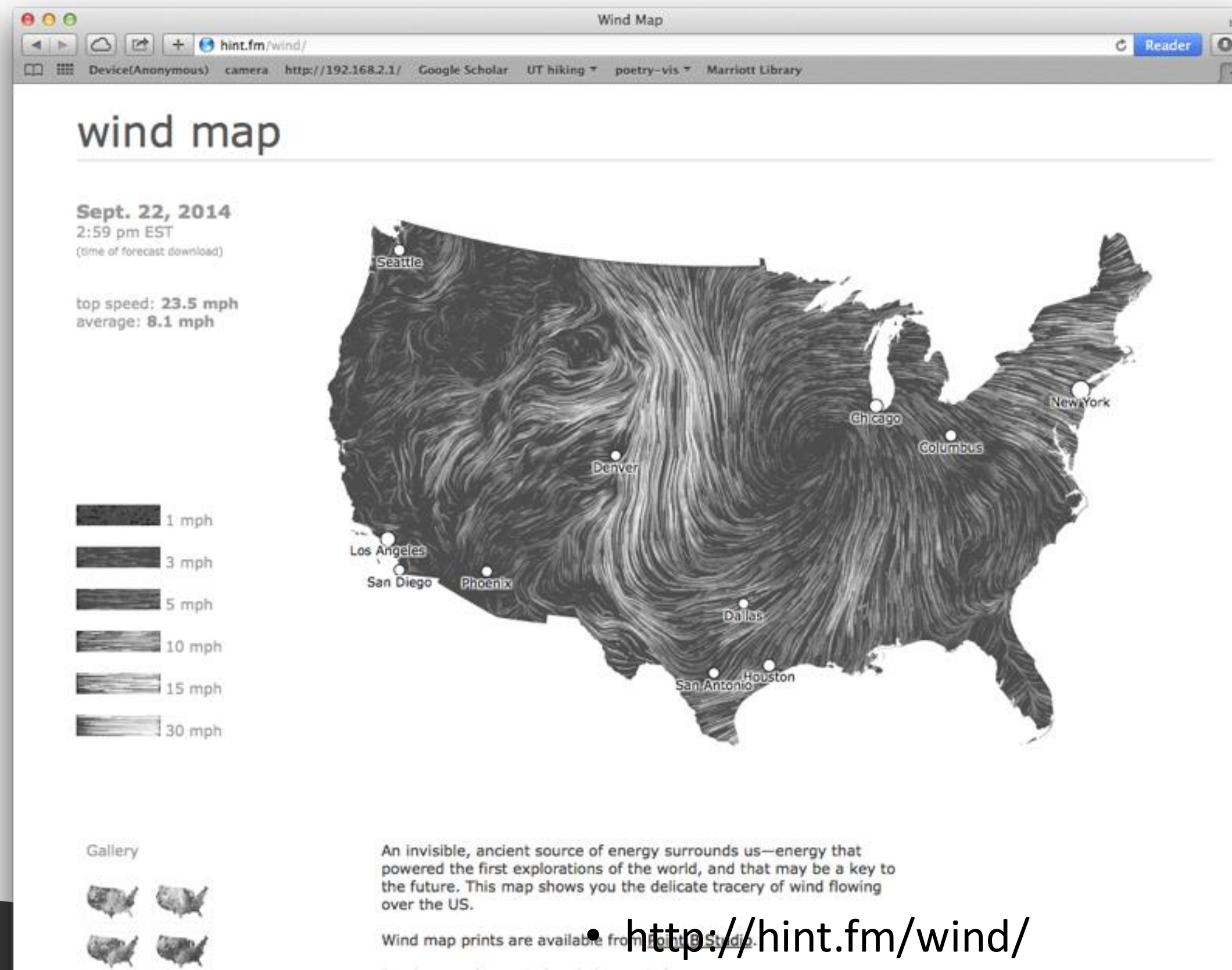


present



- <https://www.nytimes.com/interactive/2014/07/08/upshot/how-the-year-you-were-born-influences-your-politics.html>

enjoy



{action, target}

→ Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



→ Produce

→ Annotate



→ Record



→ Derive



→ Search

→ Query

{action, target}

→ Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



→ Produce

→ Annotate



→ Record



→ Derive



→ Search

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

→ Query

{action, target}

→ Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



→ Produce

→ Annotate



→ Record



→ Derive

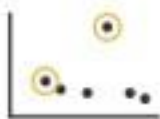


→ Search

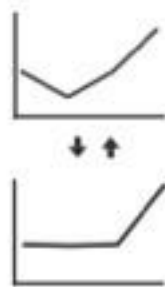
	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

→ Query

→ Identify



→ Compare



→ Summarize



{action, target}

→ All Data

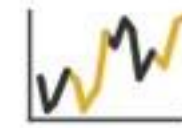
→ Trends



→ Outliers



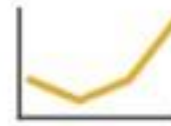
→ Features



{action, target}

→ All Data

→ Trends



→ Outliers



→ Features



→ Attributes

→ One

→ Distribution



→ Extremes



→ Many

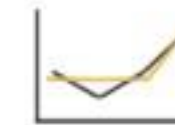
→ Dependency



→ Correlation



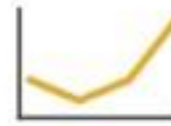
→ Similarity



{action, target}

→ All Data

→ Trends



→ Outliers



→ Features



→ Attributes

→ One

→ Distribution



→ Extremes



→ Many

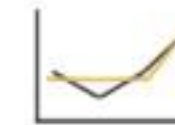
→ Dependency



→ Correlation



→ Similarity



→ Network Data

→ Topology



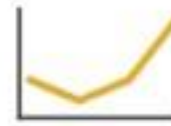
→ Paths



{action, target}

→ All Data

→ Trends



→ Outliers



→ Features



→ Attributes

→ One

→ Distribution



→ Extremes



→ Many

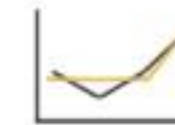
→ Dependency



→ Correlation



→ Similarity



→ Network Data

→ Topology



→ Paths



→ Spatial Data

→ Shape



Low-level task taxonomy

Low-Level Components of Analytic Activity in Information Visualization

Robert Amar, James Eagan, and John Stasko

College of Computing/GVU Center
Georgia Institute of Technology

ABSTRACT

Existing system-level taxonomies of visualization tasks are geared more towards the design of particular representations than the facilitation of user analytic activity. We present a set of ten low-level analysis tasks that largely capture people's activities while employing information visualization tools for understanding data. To help develop these tasks, we collected nearly 200 sample questions from students about how they would analyze five particular data sets from different domains. The questions, while not being totally comprehensive, illustrated the sheer variety of analytic questions typically posed by users when employing information visualization systems. We hope that the presented set of tasks is useful for information visualization system designers as a kind of common substrate to discuss the relative analytic capabilities of the systems. Further, the tasks may provide a form of checklist for system designers.

CR Categories and Subject Descriptors: H.5.0 [Information Interfaces and Presentation]: General; J.0 [Computer Applications]: General

Additional Keywords: Analytic activity, taxonomy, knowledge discovery, design, evaluation.

With the aim of generating an actionable means for supporting analytic activity, we wish to rethink some of the lower-level task taxonomies that focus on a generated presentation as an end result. In general, information visualization can benefit from understanding the tasks that users accomplish while doing actual analytic activity. Such understanding achieves two goals: first, it aids designers in creating novel presentations that amplify users' analytic abilities; second, it provides a common vocabulary for evaluating the abilities and affordances of information visualization systems with respect to user tasks.

We argue that a stronger focus on user tasks and analytic activities in information visualization is necessary as current tools do not seem to support analytic activity consistently. A 2004 study by Saraiya and North found that insights generated from tools used to visualize gene expression data were not generally valuable according to domain experts [11]. Systems such as IN-SPIRE [7] support analytic activities within the domain of document search but may not generalize across domains. Current tools may not even support representational activity very well; consider, for example, the Kobsa study showing only 68-75% accuracy on relatively simple tasks during commercial tool evaluation [8].

1 INTRODUCTION

1.2 The Nature of Analytic Activity

User analysis questions and tasks as part of analytic activity

Tasks

1. Retrieve Value

General Description: Given a set of specific cases, find attributes of those cases.

Pro Forma Abstract: What are the values of attributes {X, Y, Z, ...} in the data cases {A, B, C, ...}?

Examples:

- What is the mileage per gallon of the Audi TT?
- How long is the movie Gone with the Wind?

2. Filter

General Description: Given some concrete conditions on attribute values, find data cases satisfying those conditions.

Pro Forma Abstract: Which data cases satisfy conditions {A, B, C...}?

Examples:

- What Kellogg's cereals have high fiber?
- What comedies have won awards?
- Which funds underperformed the SP-500?

Tasks

3. Compute Derived Value

General Description: Given a set of data cases, compute an aggregate numeric representation of those data cases.

Pro Forma Abstract: What is the value of aggregation function F over a given set S of data cases?

Examples:

- What is the average calorie content of Post cereals?
- What is the gross income of all stores combined?
- How many manufacturers of cars are there?

4. Find Extremum

General Description: Find data cases possessing an extreme value of an attribute over its range within the data set.

Pro Forma Abstract: What are the top/bottom N data cases with respect to attribute A ?

Examples:

- What is the car with the highest MPG?
- What director/film has won the most awards?
- What Robin Williams film has the most recent release date?

Tasks

5. Sort

General Description: Given a set of data cases, rank them according to some ordinal metric.

Pro Forma Abstract: What is the sorted order of a set S of data cases according to their value of attribute A ?

Examples:

- Order the cars by weight.
- Rank the cereals by calories.

6. Determine Range

General Description: Given a set of data cases and an attribute of interest, find the span of values within the set.

Pro Forma Abstract: What is the range of values of attribute A in a set S of data cases?

Examples:

- What is the range of film lengths?
- What is the range of car horsepower?
- What actresses are in the data set?

Tasks

7. Characterize Distribution

General Description: Given a set of data cases and a quantitative attribute of interest, characterize the distribution of that attribute's values over the set.

Pro Forma Abstract: What is the distribution of values of attribute A in a set S of data cases?

Examples:

- What is the distribution of carbohydrates in cereals?
- What is the age distribution of shoppers?

8. Find Anomalies

General Description: Identify any anomalies within a given set of data cases with respect to a given relationship or expectation, e.g. statistical outliers.

Pro Forma Abstract: Which data cases in a set S of data cases have unexpected/exceptional values?

Examples:

- Are there exceptions to the relationship between horsepower and acceleration?
- Are there any outliers in protein?

Tasks

9. Cluster

General Description: Given a set of data cases, find clusters of similar attribute values.

Pro Forma Abstract: Which data cases in a set S of data cases are similar in value for attributes $\{X, Y, Z, \dots\}$?

Examples:

- Are there groups of cereals w/ similar fat/calories/sugar?
- Is there a cluster of typical film lengths?

10. Correlate

General Description: Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.

Pro Forma Abstract: What is the correlation between attributes X and Y over a given set S of data cases?

Examples:

- Is there a correlation between carbohydrates and fat?
- Is there a correlation between country of origin and MPG?
- Do different genders have a preferred payment method?
- Is there a trend of increasing film length over the years?

Tasks Not Considered

5.1 Compound Tasks

Considering the set of tasks in the taxonomy to be analytic “primitives” allows us to examine some questions that do not cleanly fit into one category but rather appear to be compositions of primitive tasks. For instance, the task “Sort the cereal manufacturers by average fat content” involves a Compute Derived Value (average fat) primitive followed by a Sort primitive.

5.2.2 Higher-level Questions

We have found that the proposed ten tasks cover the vast majority of the corpus of analytic questions we studied. Some questions, however, imply tasks not explicitly covered by our task set, but instead they can be thought of as guiding higher-level exploration in the data set. For example:

- “Do any variables correlate with fat?”
- “How do mutual funds get rated?”
- “Are there car aspects that Toyota has concentrated on?”

5.2.1 Low-level Mathematical and Cognitive Actions

In constructing the taxonomy, we abstracted away as low-level, and thus beyond the scope of the present work, some basic mathematical and cognitive operations, such as determining that a data case mathematically satisfies filtering criteria or conditions and computing aggregate values from a mathematical perspective. In particular, we explicitly acknowledge the existence of a low-level mathematical comparison operation, one in which a value is evaluated for being less than, greater than, or equal to another value or values.

This leads to the notion of questions whose overall goal is too “low-level” for our analytic task taxonomy. For instance, the following questions involve the aforementioned mathematical comparison operation:

- “Which cereal has more sugar, Cheerios or Special K?”
- “Compare the average MPG of American and Japanese cars.”

5.2.3 Uncertain Criteria

Other questions in the corpus contained uncertain criteria, for example:

- “Do cereals (X, Y, Z...) sound tasty?”
- “What are the characteristics of the most valued customers?”
- “Are there any particular funds that are better than others?”

Another style of question common in the set involves a comparison operation that is much higher in level and more abstract than the fundamental mathematical comparison operation discussed earlier in the section. For instance, consider the questions:

- “What other cereals are most similar to Trix?”
- “How does the Toyota RAV4 compare to the Honda CRV?”
- “Compare the distributions of values for sugar and fat in the cereals.”

Using Interaction

- Change Over Time
- Rearranging
- Selection & Highlighting
- Linking
- Filtering
- Navigation

CHANGE OVER TIME

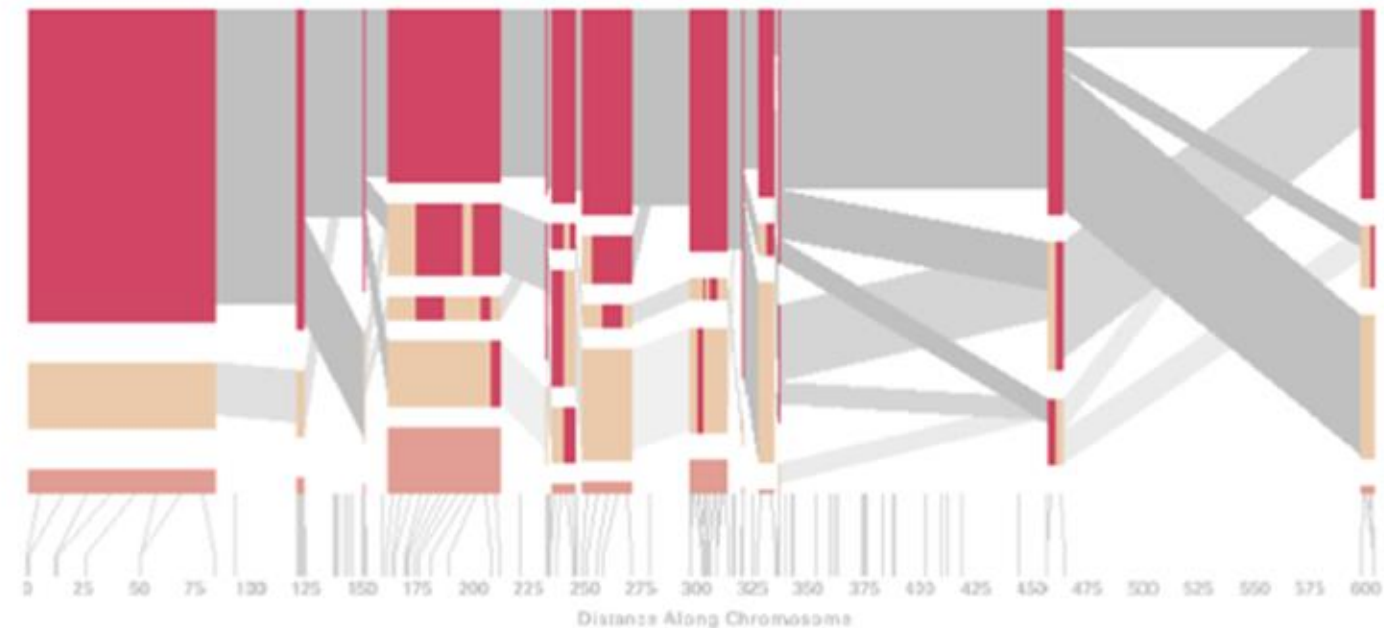
animated transitions

<< BEN FRY

isometricblocks

When comparing the genome of two different people, you'll see single letter changes (called SNPs, pronounced "snips") every few thousand letters. An interesting feature of SNPs is that their ordering has distinct patterns, where sets of consecutive changes are most often found together. There are many methods for looking at this data, so this piece combines several of them into a single visual display. The project is described in greater detail in my [dissertation](#), starting in chapter four.

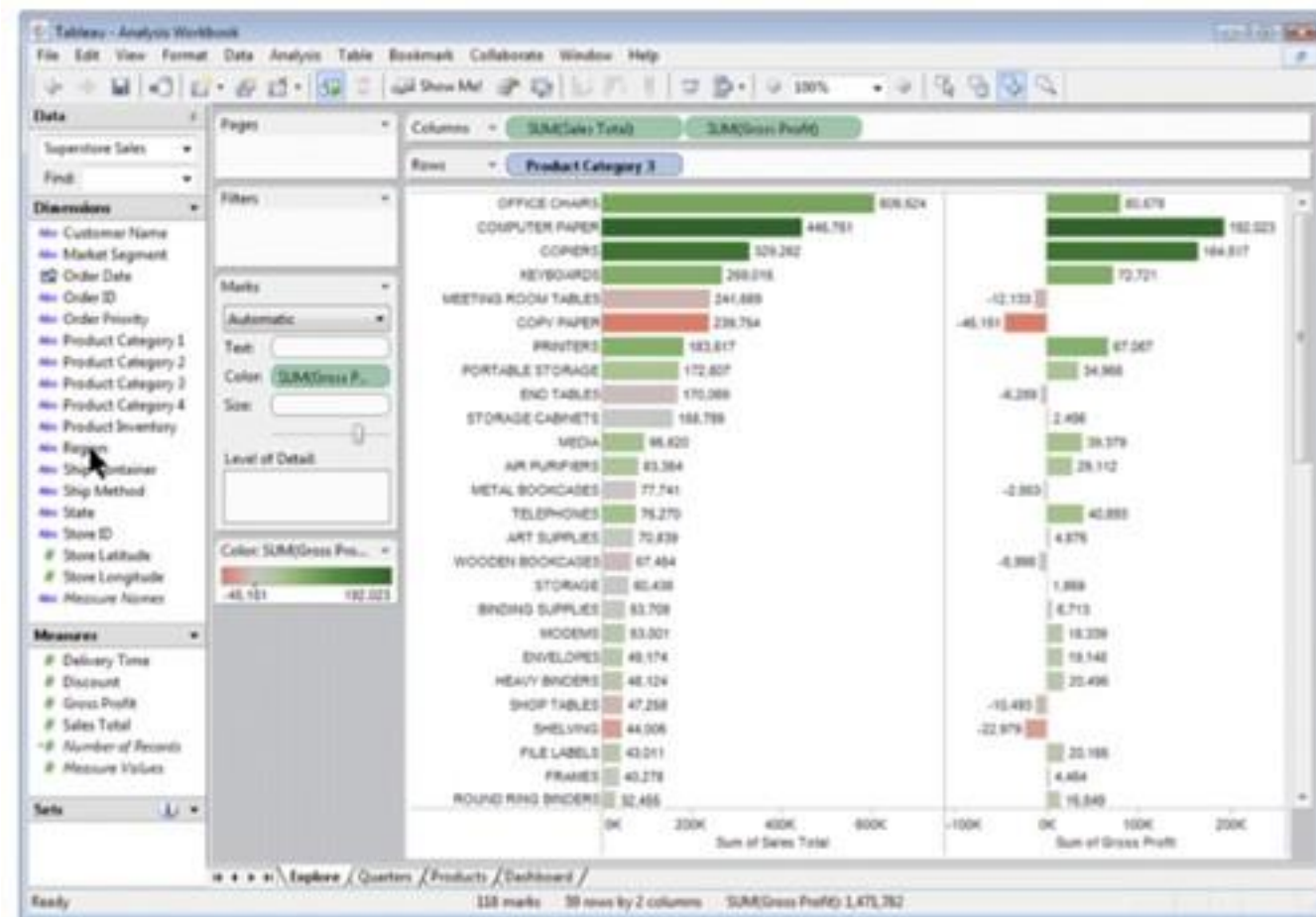
View ☒ 2D ☐ 2D Even Spacing ☐ 2D Quantitative ☐ 3D ☐ 3D with LD Units ☐ LD Units from above



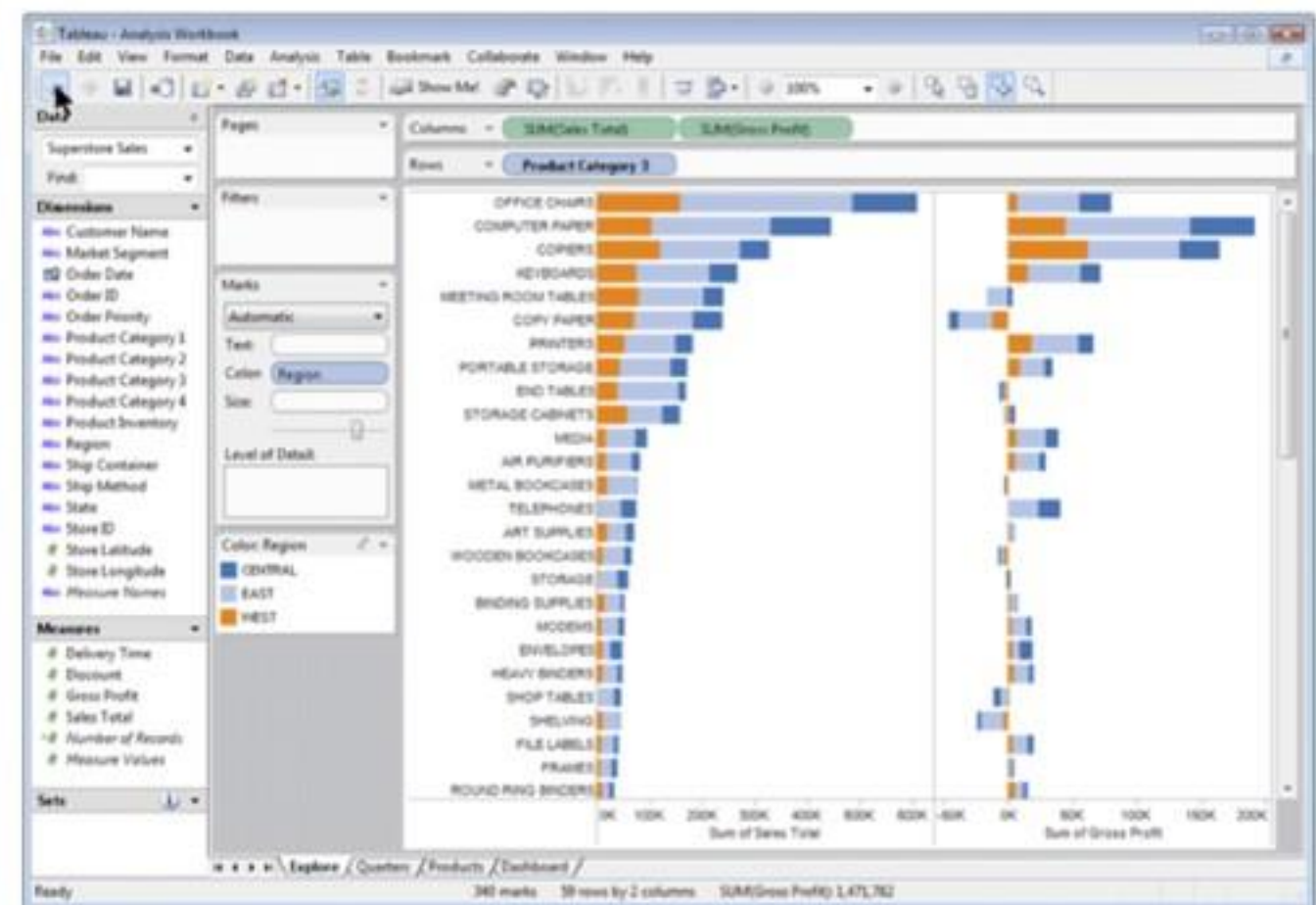
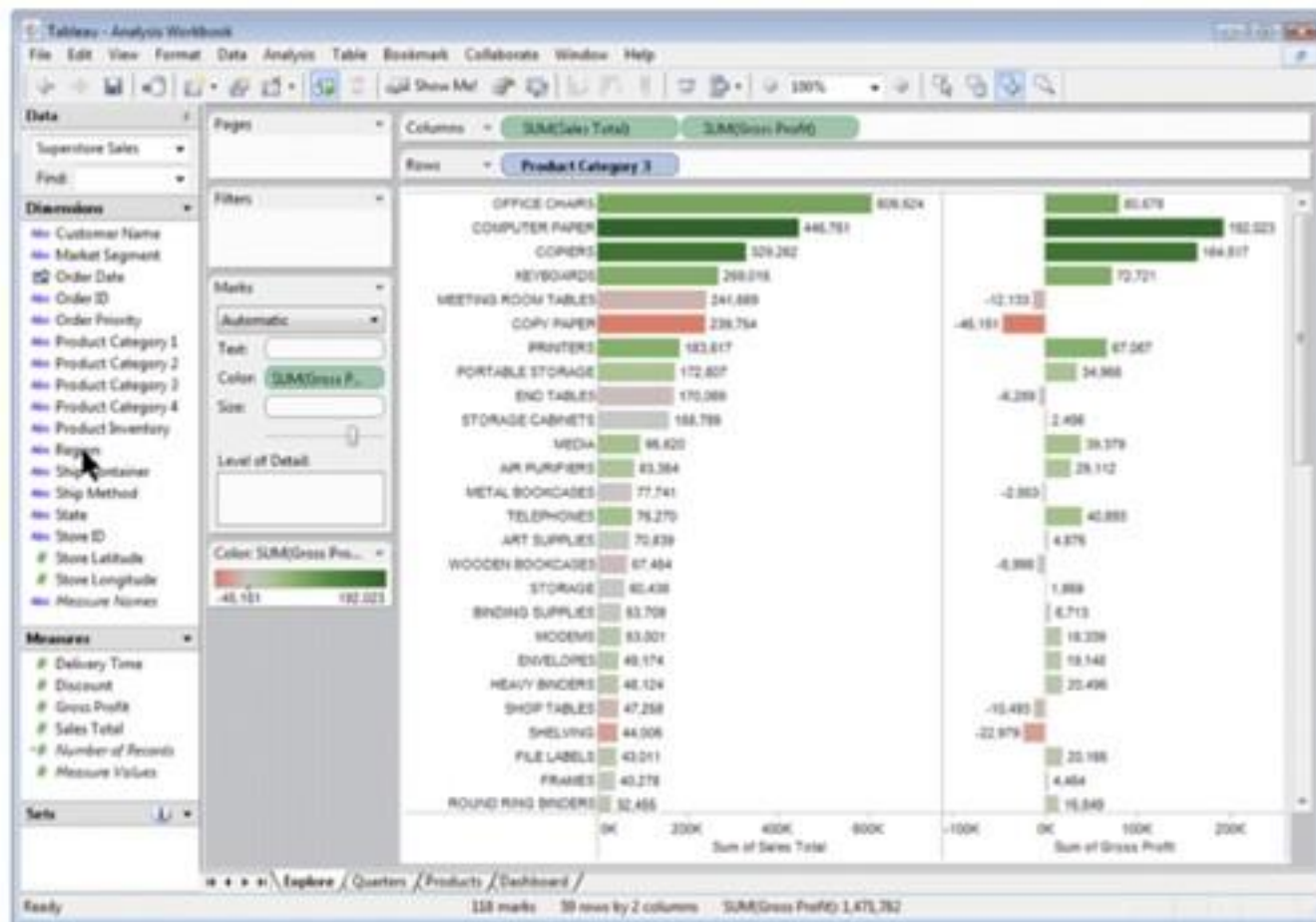
Cut High CI ☐ 98
Cut Low CI ☐ 70
Rec High CI ☐ 90
Min Strong LD ☐ 95

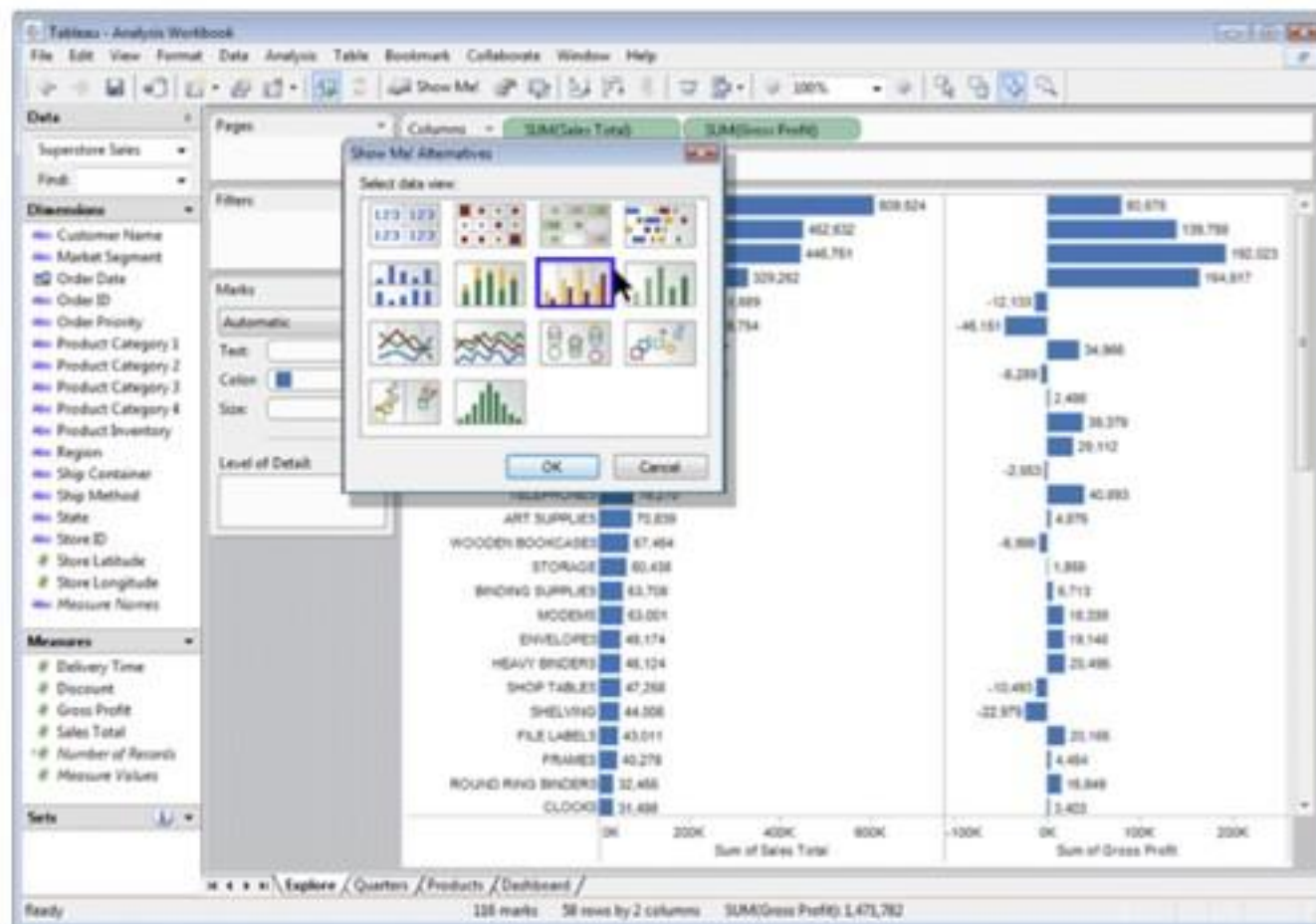
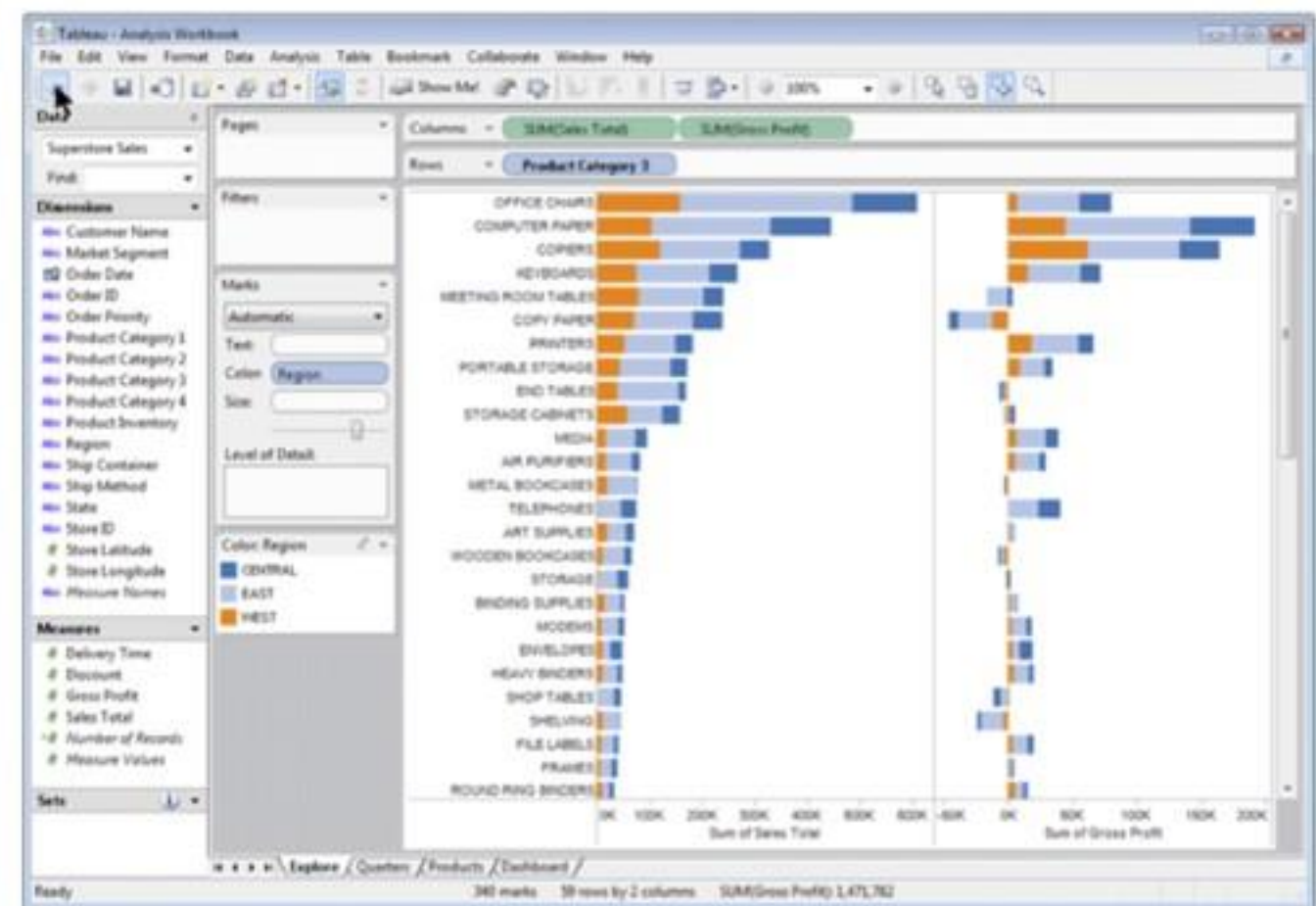
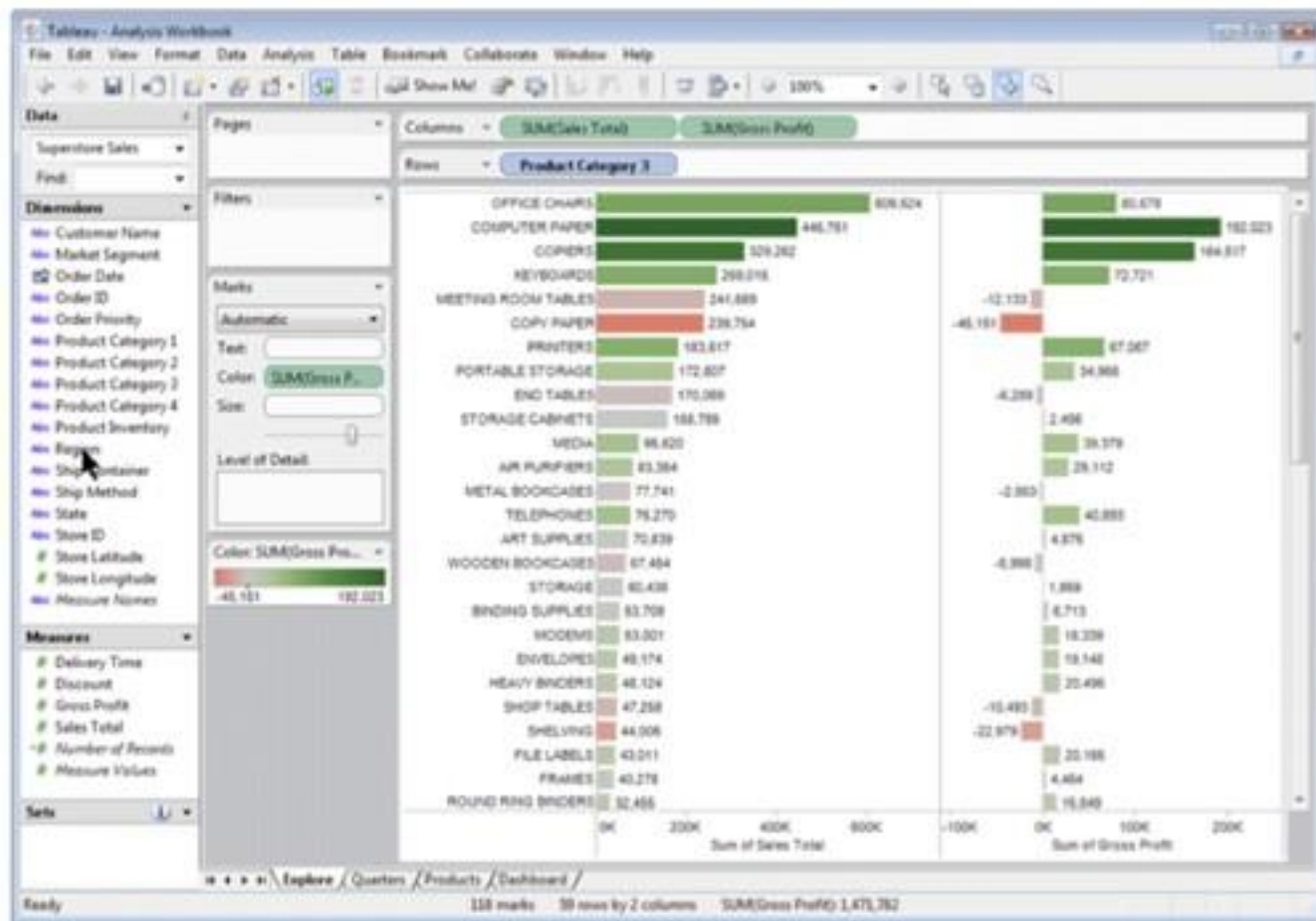
- <http://benfry.com/isometricblocks/>

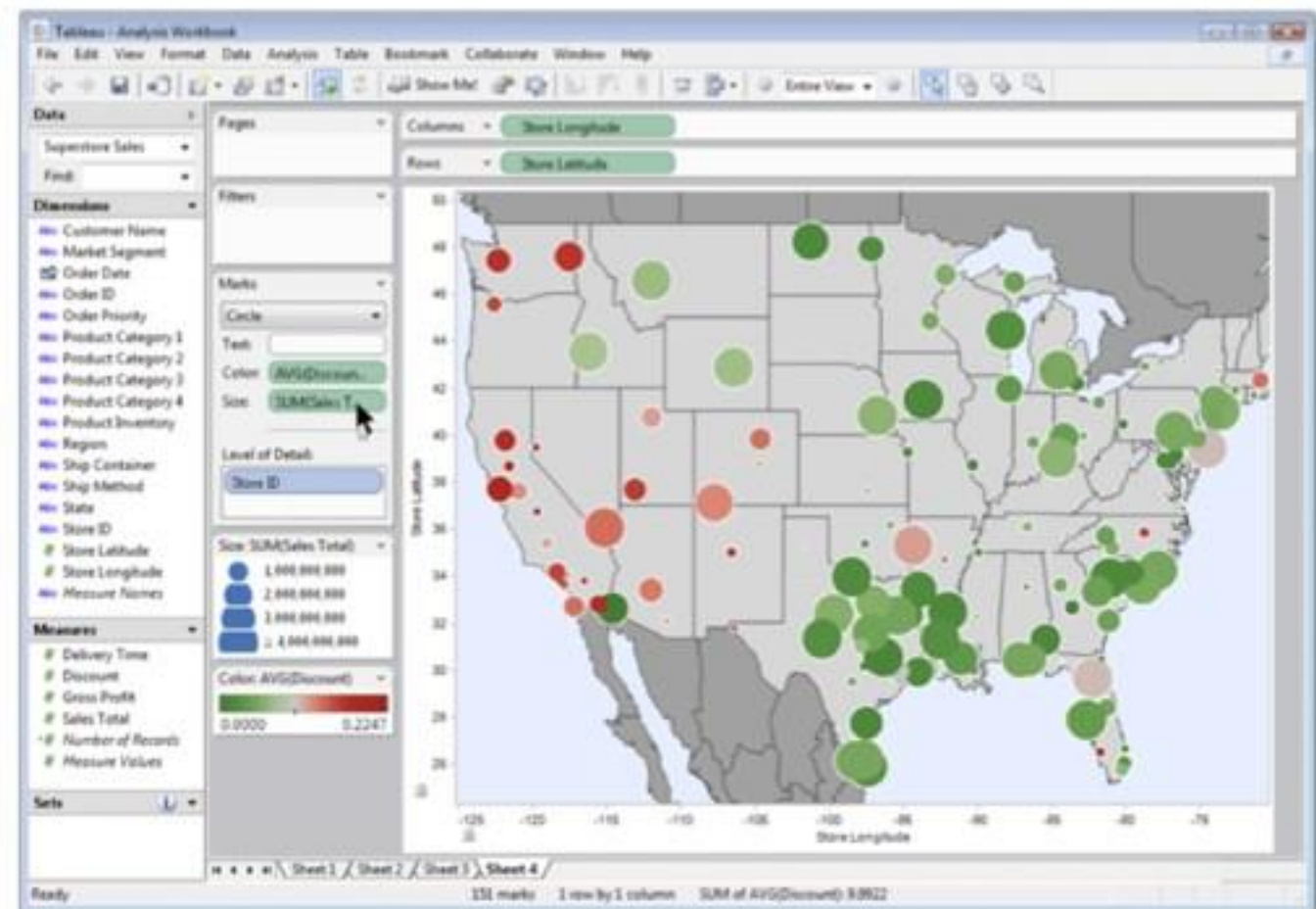
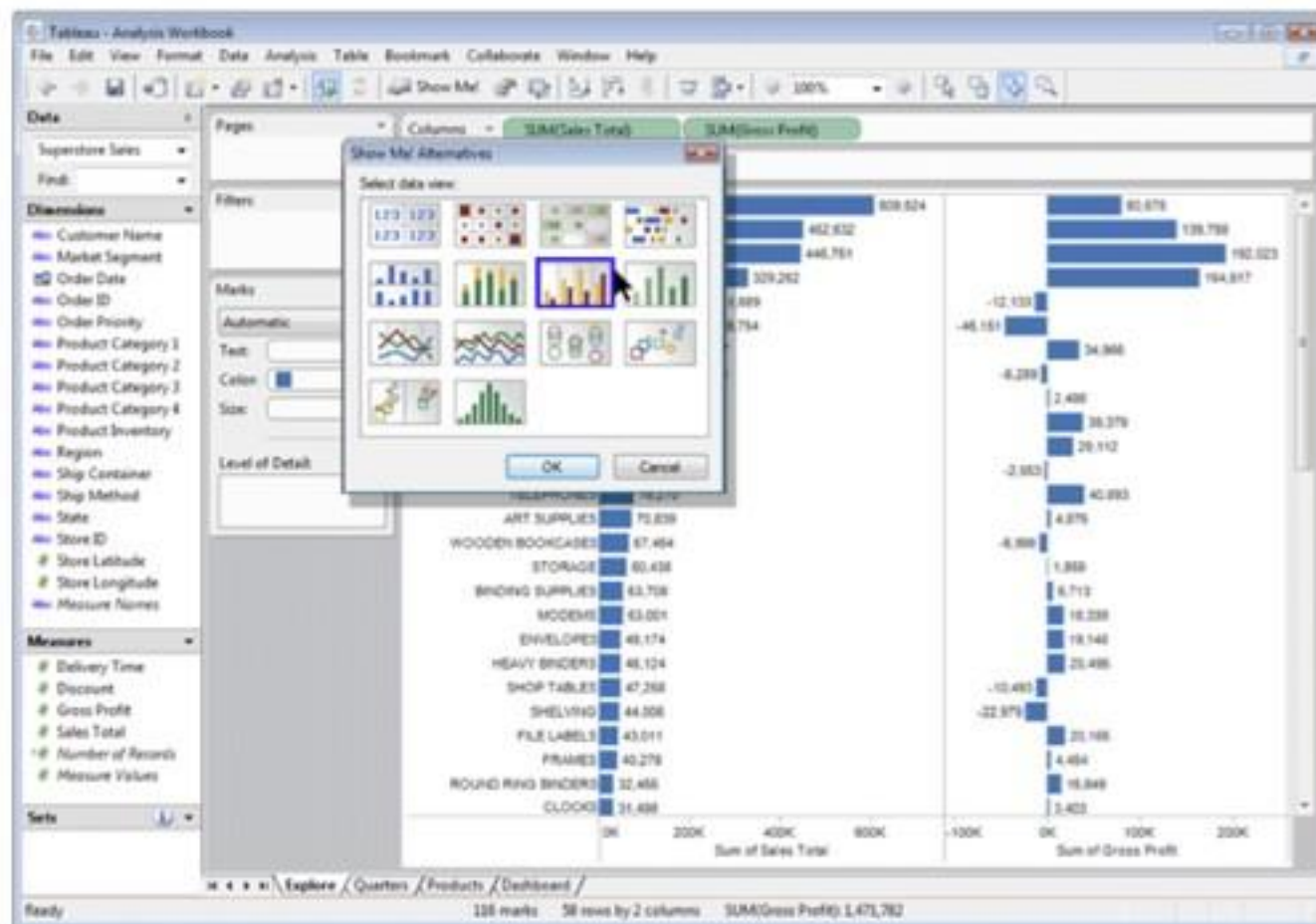
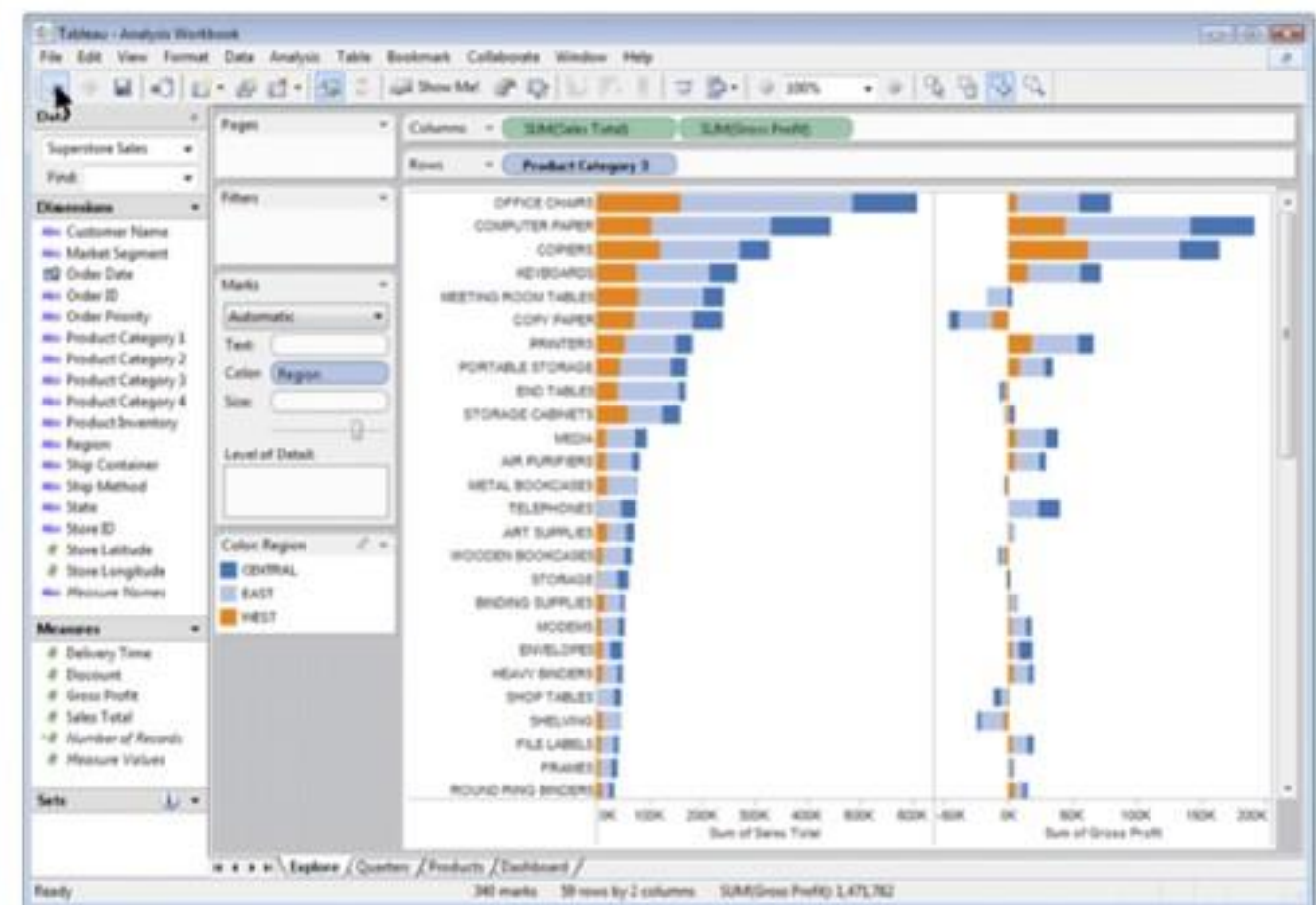
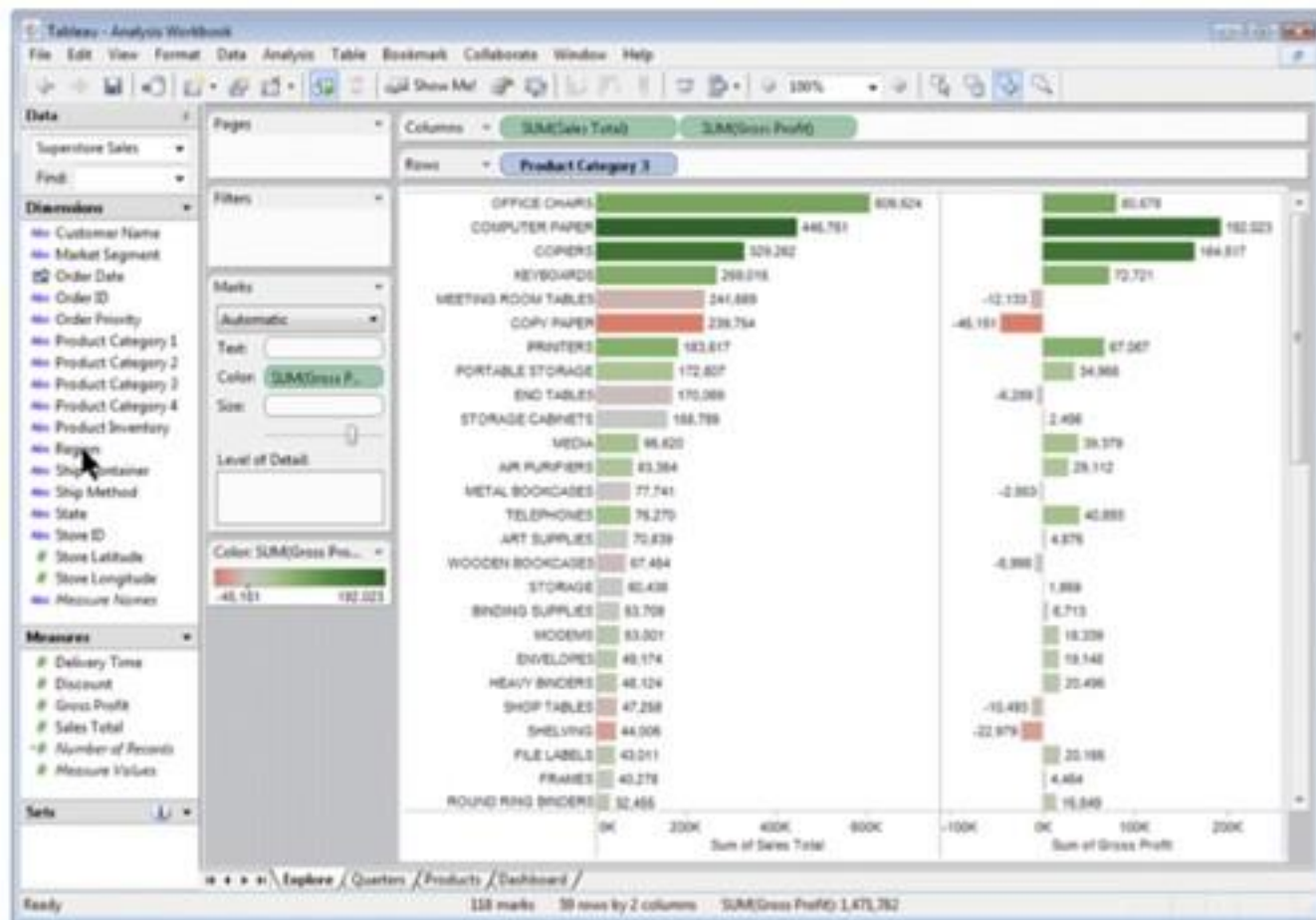
Rearranging



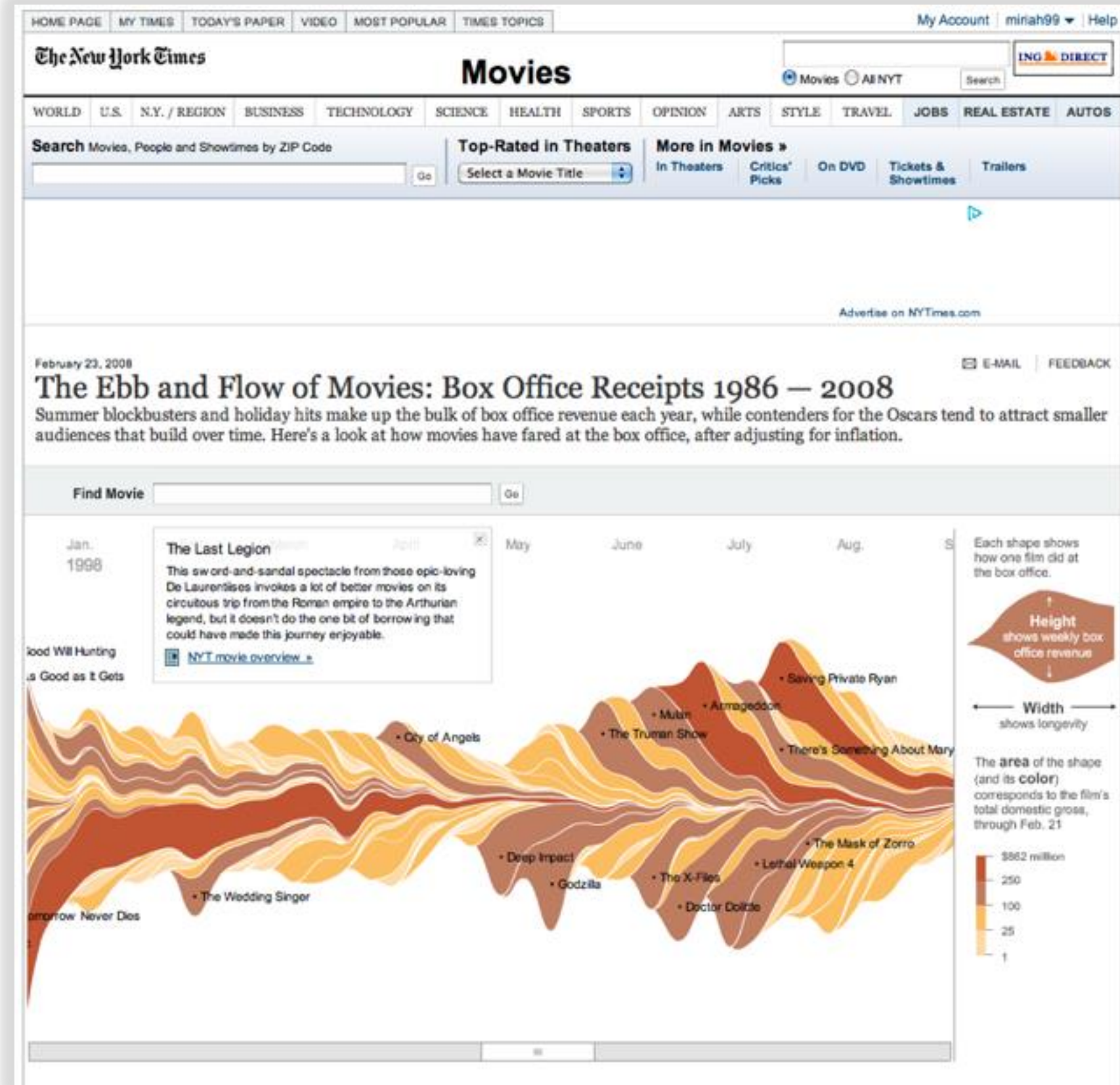
change encoding





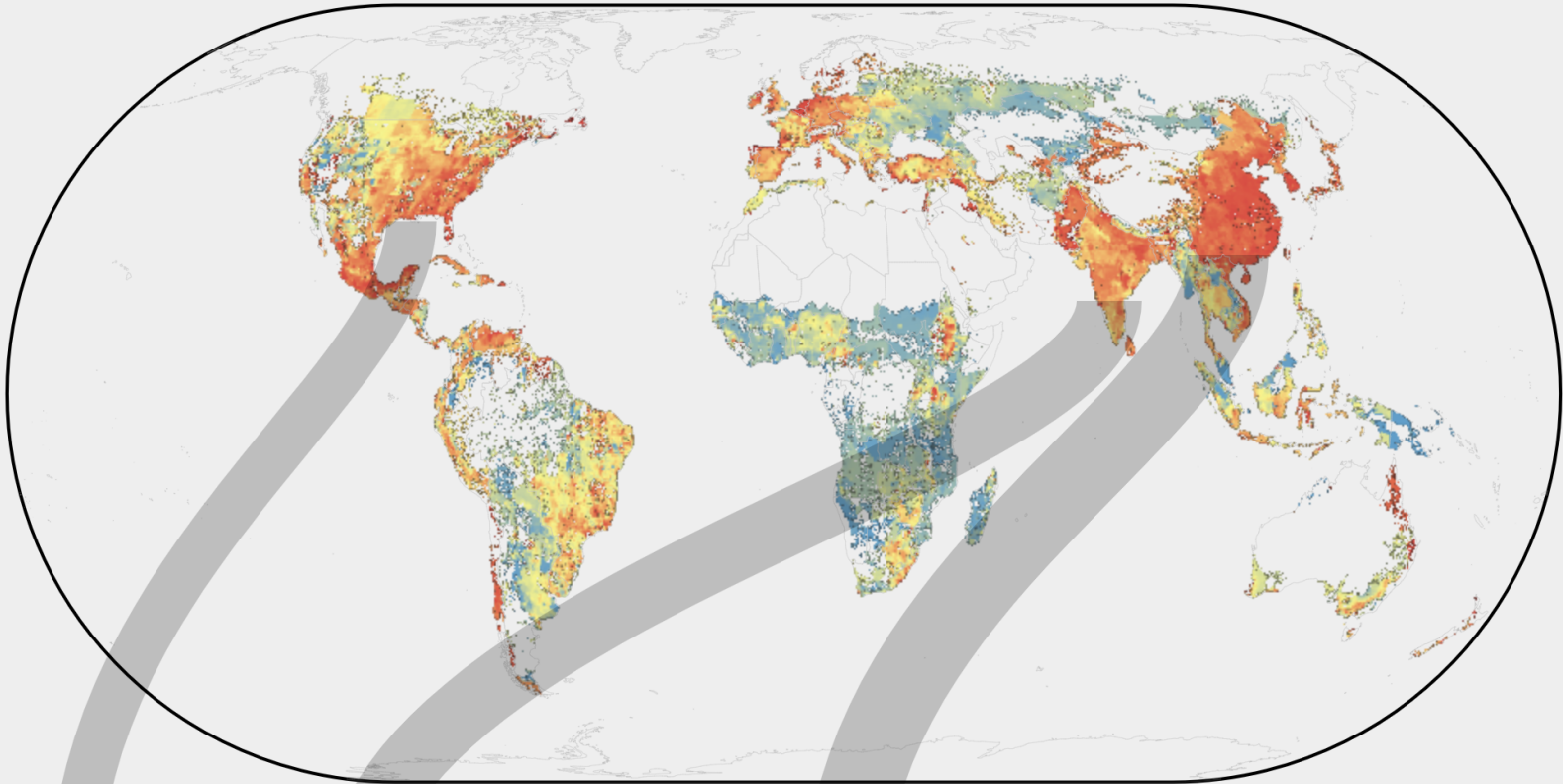


SELECTION & HIGHLIGHTING



Linking

Global Excess Nitrogen



Canada

Excess Nitrogen: 737,678.4 TONNS
Global Percentage: 1.49%

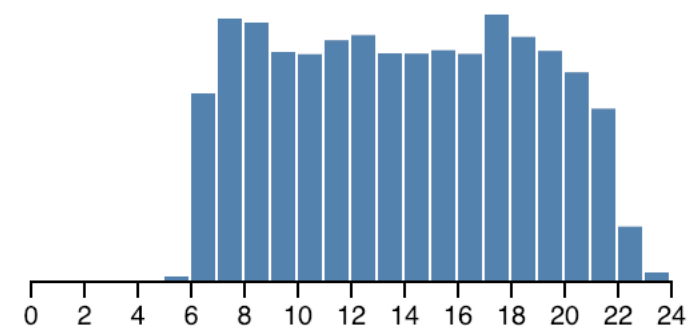
Developed by [Shipeng Sun](#)
[Global Landscapes Initiative](#)
[Institute on the Environment](#)
University of Minnesota
Now at [Environmental Studies](#)
Univeristy of Illinois Springfield

[Source Code](#)

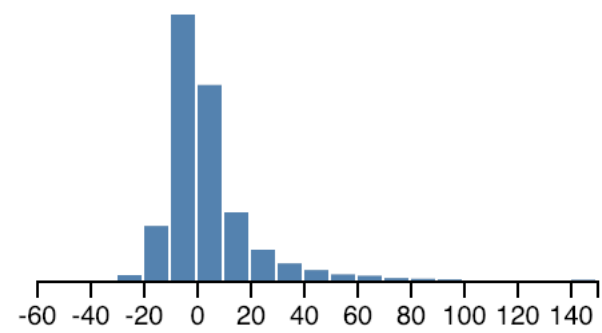
United States	India	China	Iran	Greece	Nepal	Argentina	Italy	Egypt	Vietnam	Mexico	Bangladesh
			Brazil	Bolivia	Cuba	South Africa	Iran	Thailand	Canada		
			Russia	Hungary	Ukraine	Morocco	Nigeria	Australia			
			Poland	Syria	Democratic Republic of Congo	Philippines		France	Turkey	Brazil	
			Finland	Sweden	Czech Republic	United Kingdom	Poland				
			Chile	Denmark	Netherlands	Sudan		Germany	Indonesia	Pakistan	
			Peru	Belarus	Uzbekistan	Japan	Spain				
			Australia	Iraq	Colombia	Ethiopia	Korea				
			Canada	Ghana	Honduras	Sri Lanka	Venezuela				

Filtering

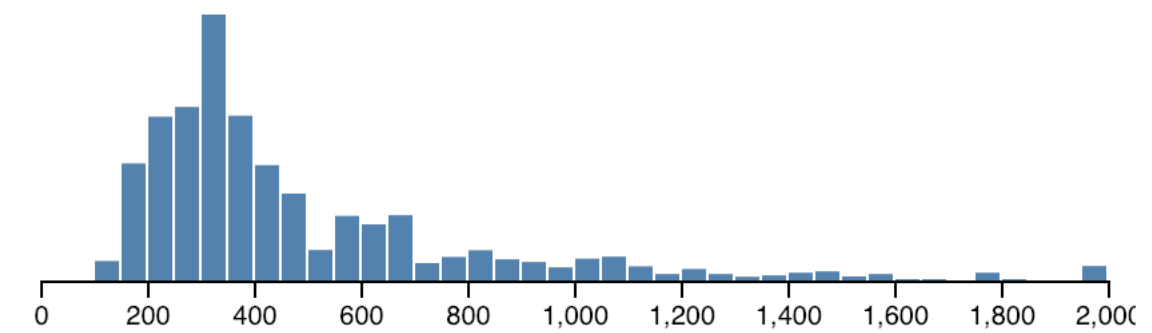
Time of Day



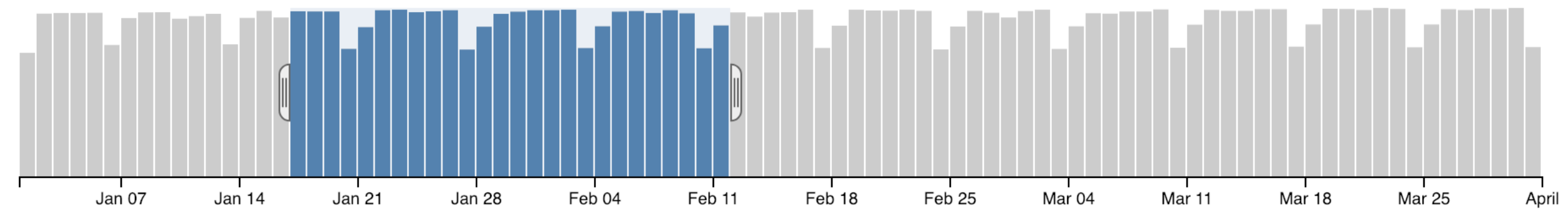
Arrival Delay (min.)



Distance (mi.)

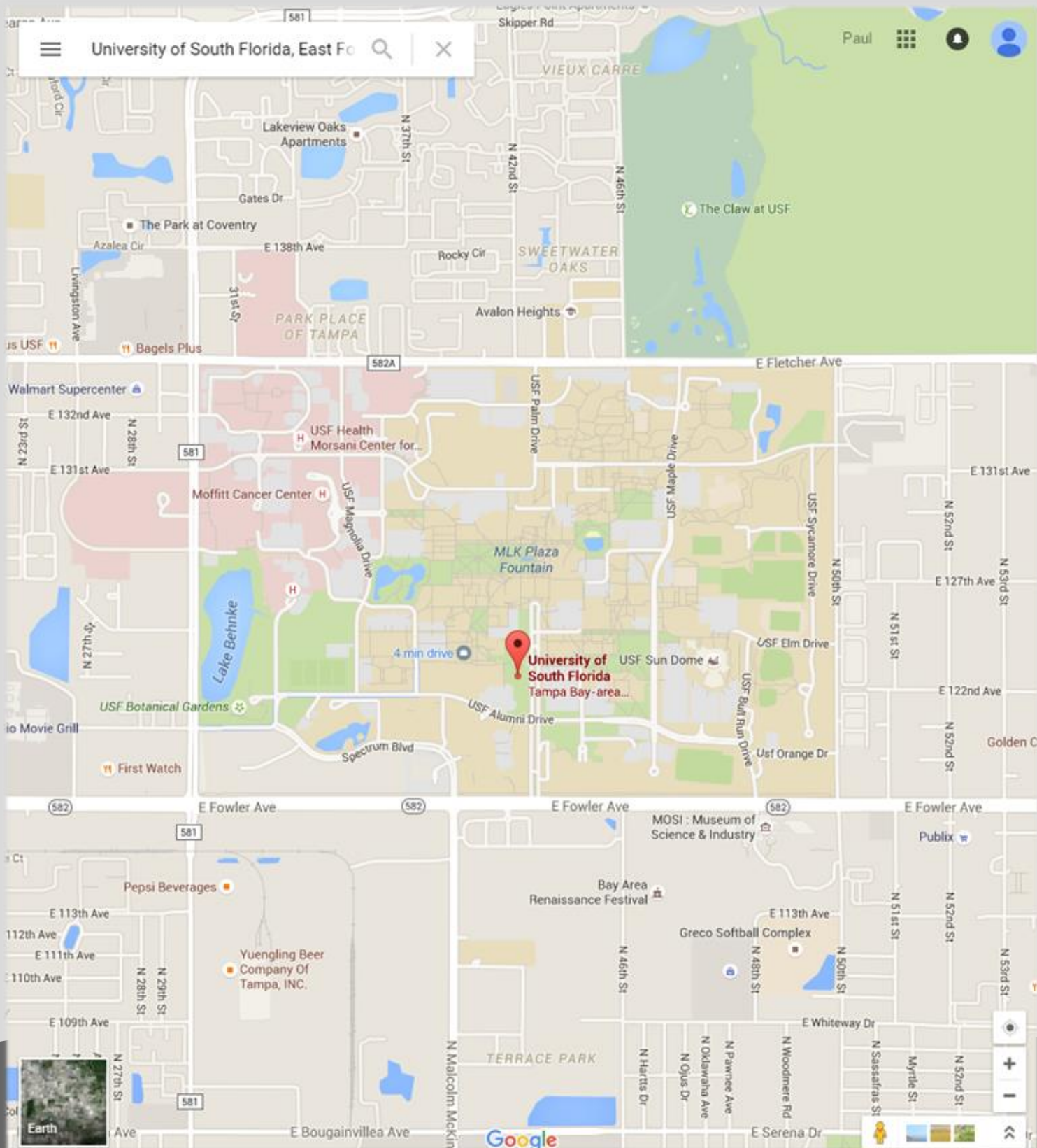


Date [reset](#)

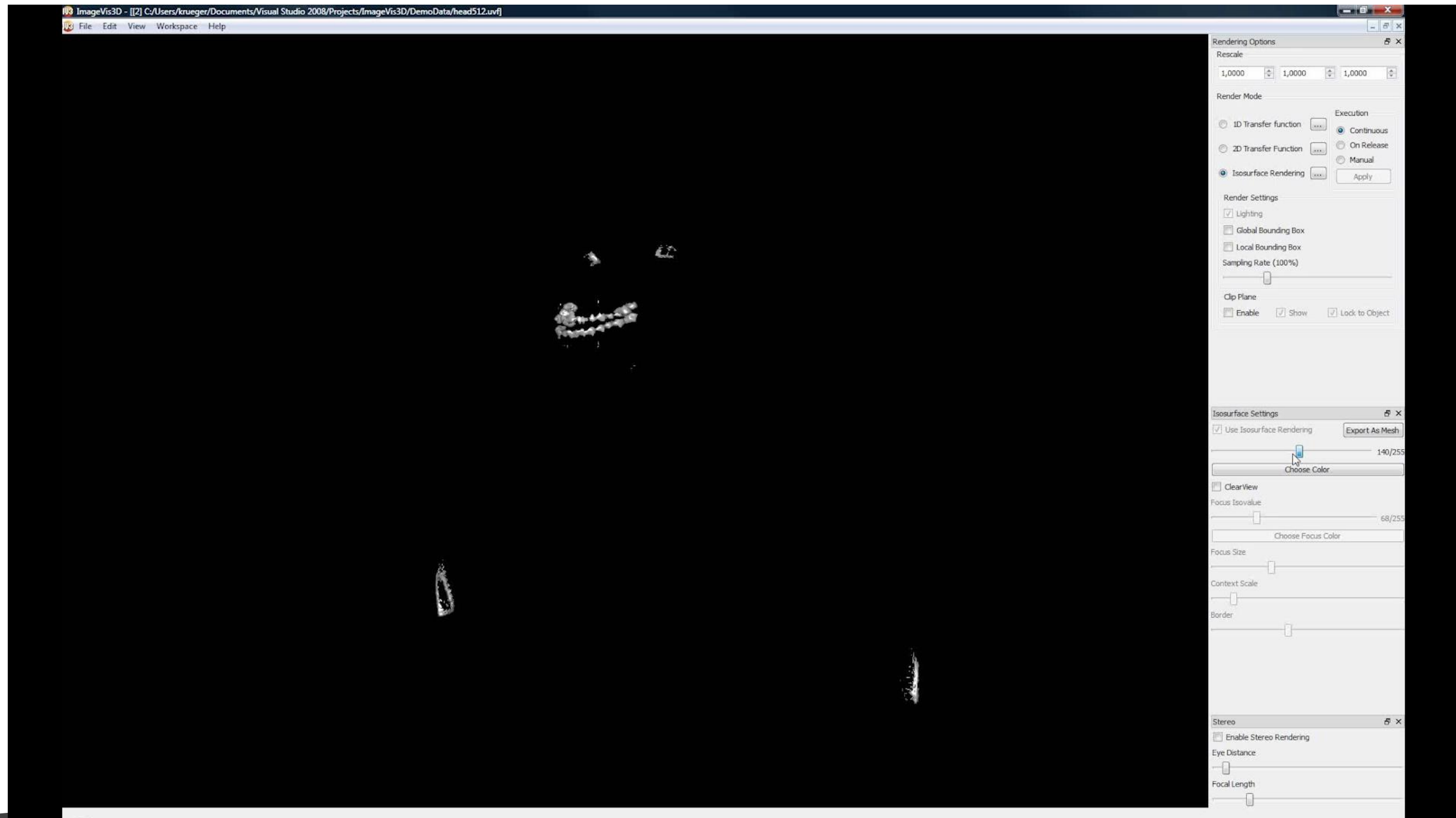


NAVIGATION

pan (and translate)

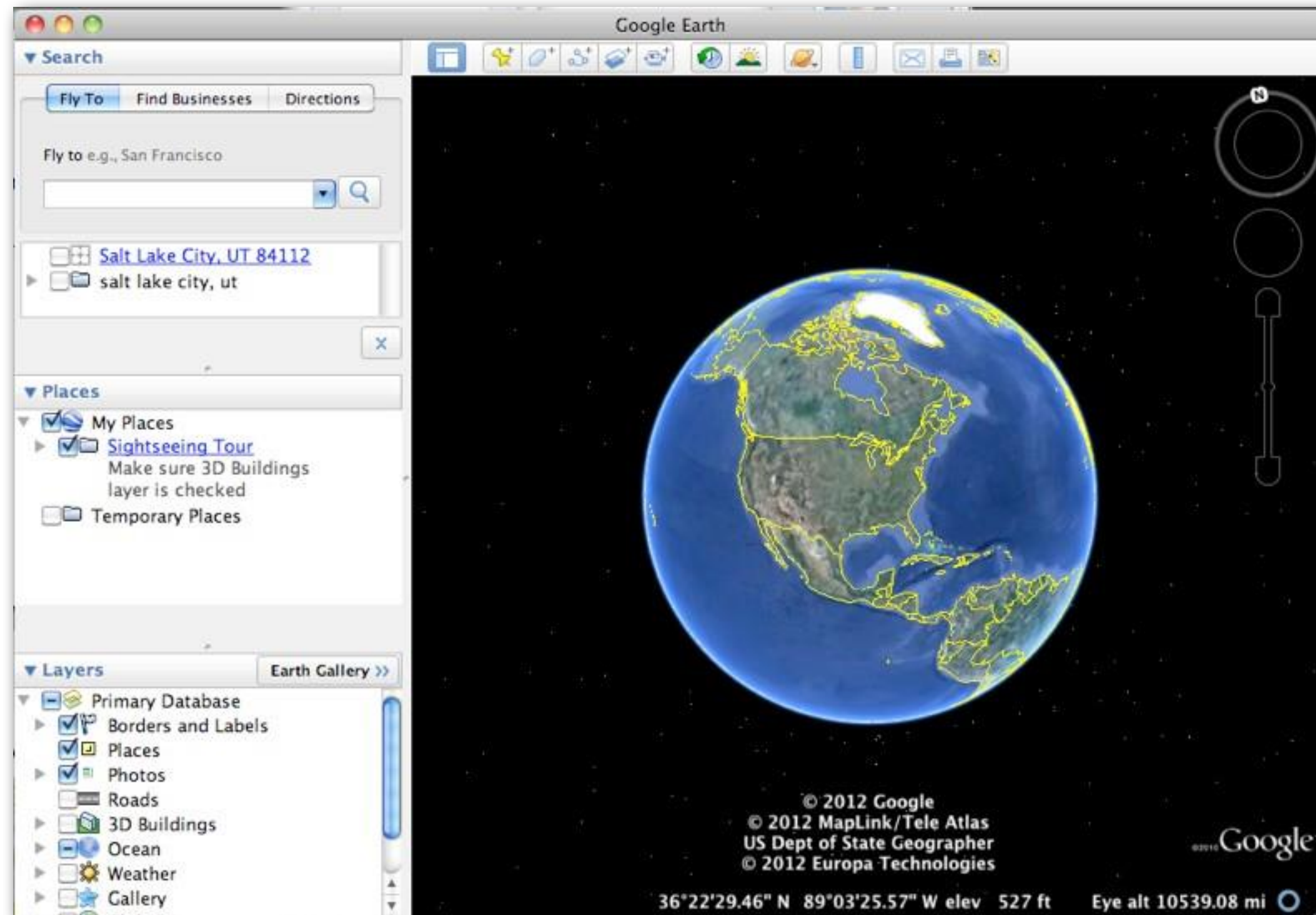


rotate



GEOMETRIC vs SEMANTIC ZOOMING

geometric



semantic



semantic

**LiveRAC: Interactive Visual Exploration of
System Management Time-Series Data**

