Advanced Data Visualization

CS 6965

Spring 2018

Prof. Bei Wang Phillips

University of Utah

Lecture 28

Graph-based ML + VIS
Future directions

NV

Announcement

- Reminder of final project presentation time:
  - April 24 (Tuesday) 9:10 - 10:30 a.m
  - April 27 (Friday) 8:00 to 10:00 a.m.
- Final project report due time:
  - April 30 (Monday) 9:10 a.m.
- Survey
  - What materials would you like to see more/less of?

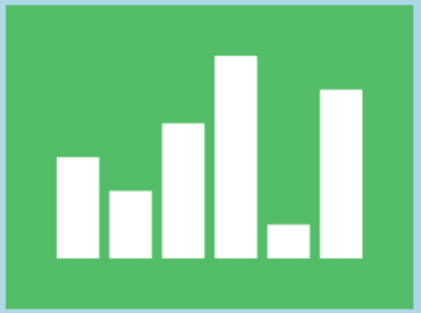
Graph-Based ML Plus Visualization

Visualizing Execution of Algorithm

VISUALGO.NET/EN

visualising data structures and algorithms through animation

https://visualgo.net/en

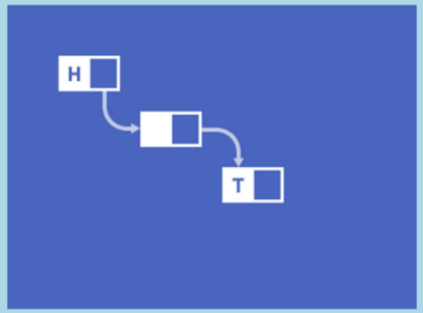


Sorting Training

array | algorithm | bubble | select | insert

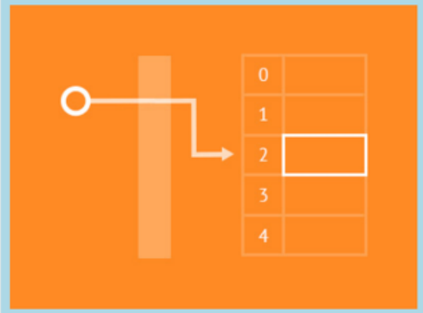
Bitmask Training

bit manipulation | set | cs3233 | array | list | ds



Linked List Training

stack | queue | doubly | deque | cs1020 | cs2020

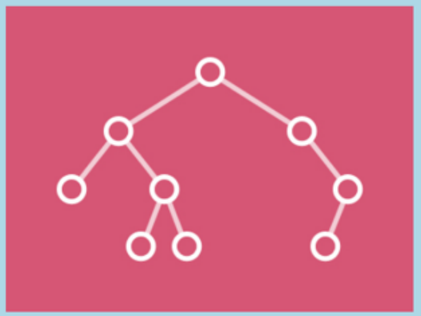


Hash Table Training

open addressing | linear | quadratic | probing

Binary Heap Training

priority queue | recursive | cs2010 | cs2020



Binary Search Tree Training

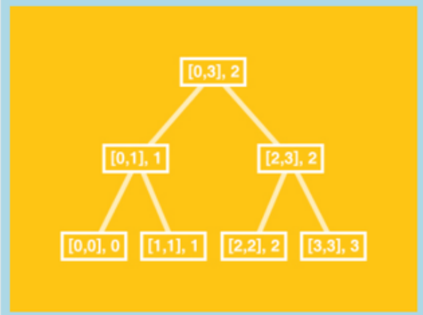
adelson velskii landis | set | table | avl | cs2010

Graph Structures Training

tree | complete | bipartite | dag | cs2010 | cs2020

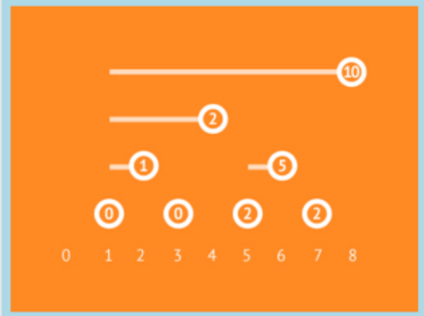
Union-Find DS Training

path compression | disjoint | set | data structure



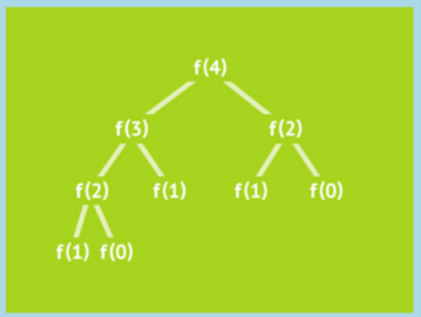
Segment Tree Training

dynamic | range | sum | min | max | cs3233



Fenwick Tree Training

binary indexed tree | bit | dynamic | fenwick | range



Recursion Tree/DAG Training

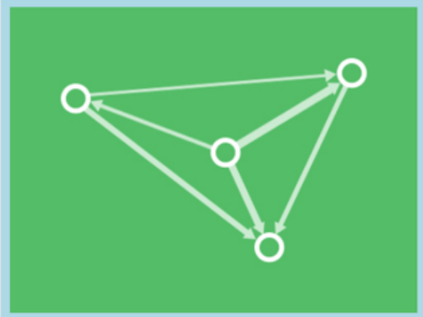
dynamic programming | dp | generic | cs1010

Graph Traversal Training

bfs | dfs | cs2010 | cs2020 | cs2040 | bipartite

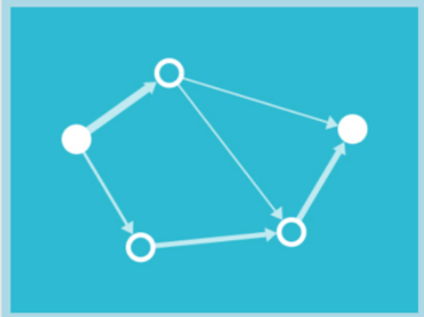
Min Spanning Tree Training

mst | prim | kruskal | graph | min | spanning



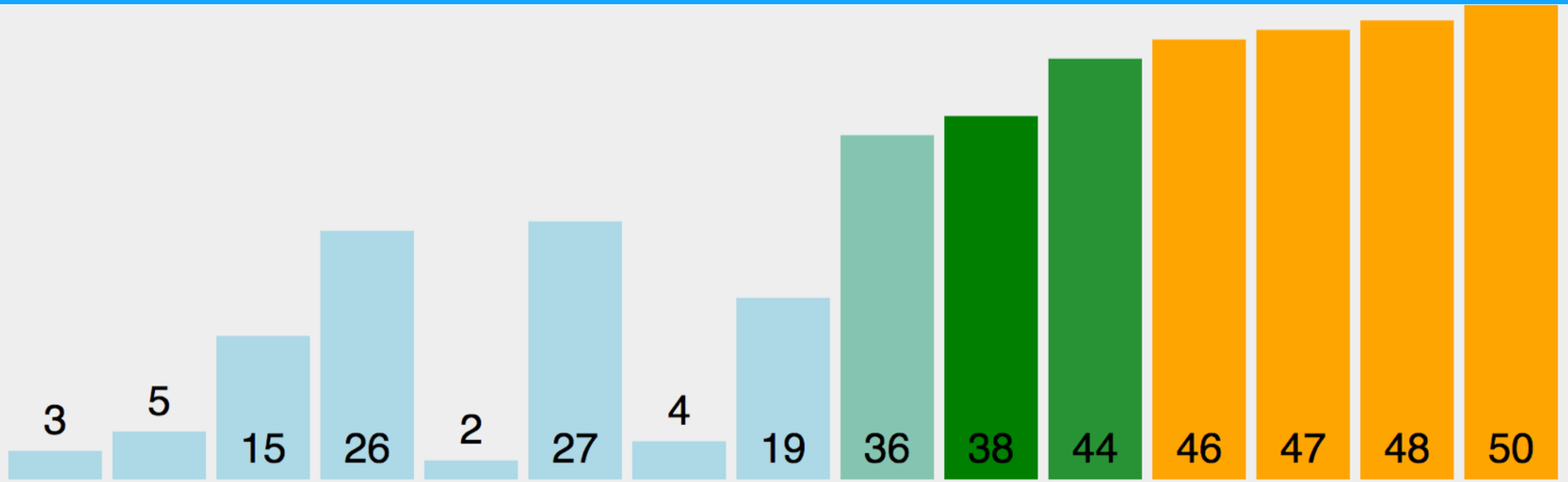
SS Shortest Paths Training

sssp | single-source | bfs | dijkstra | bellman ford



Network Flow

max flow | edmonds karp | min cut | dinic

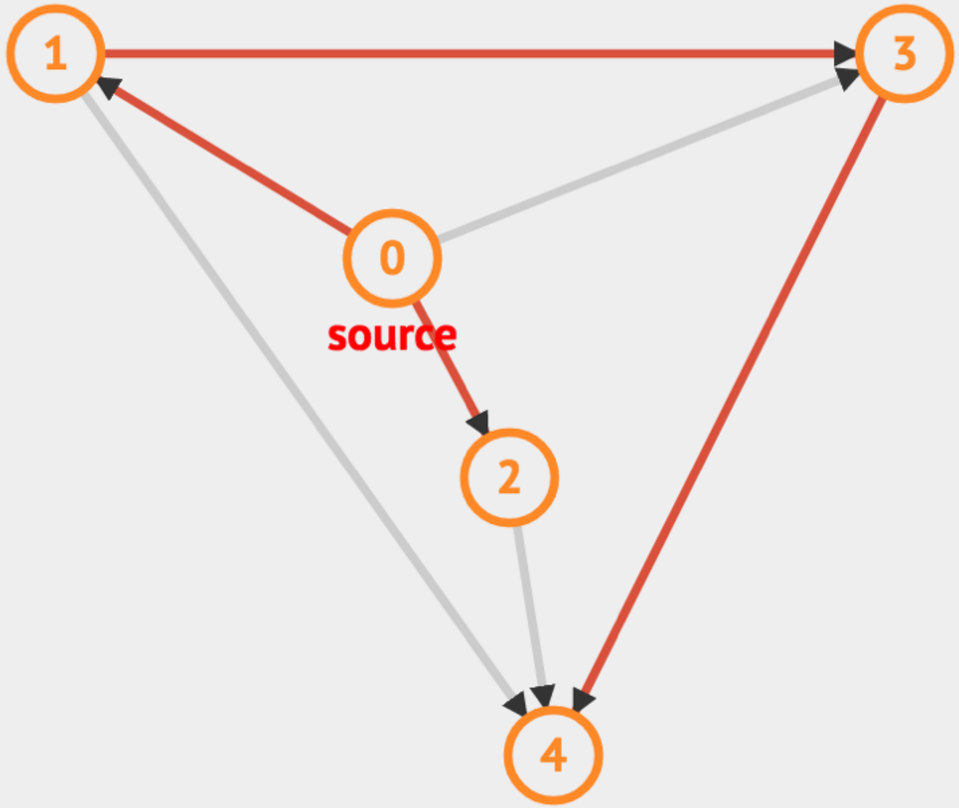


Bubble Sort

https://visualgo.net/en/sorting

Checking if 38 > 44 and swap them if that is true.
The current value of swapped = true.

```
do
  swapped = false
  for i = 1 to indexOfLastUnsortedElement-1
    if leftElement > rightElement
      swap(leftElement, rightElement)
      swapped = true
while swapped
```



https://visualgo.net/de/dfsbfs

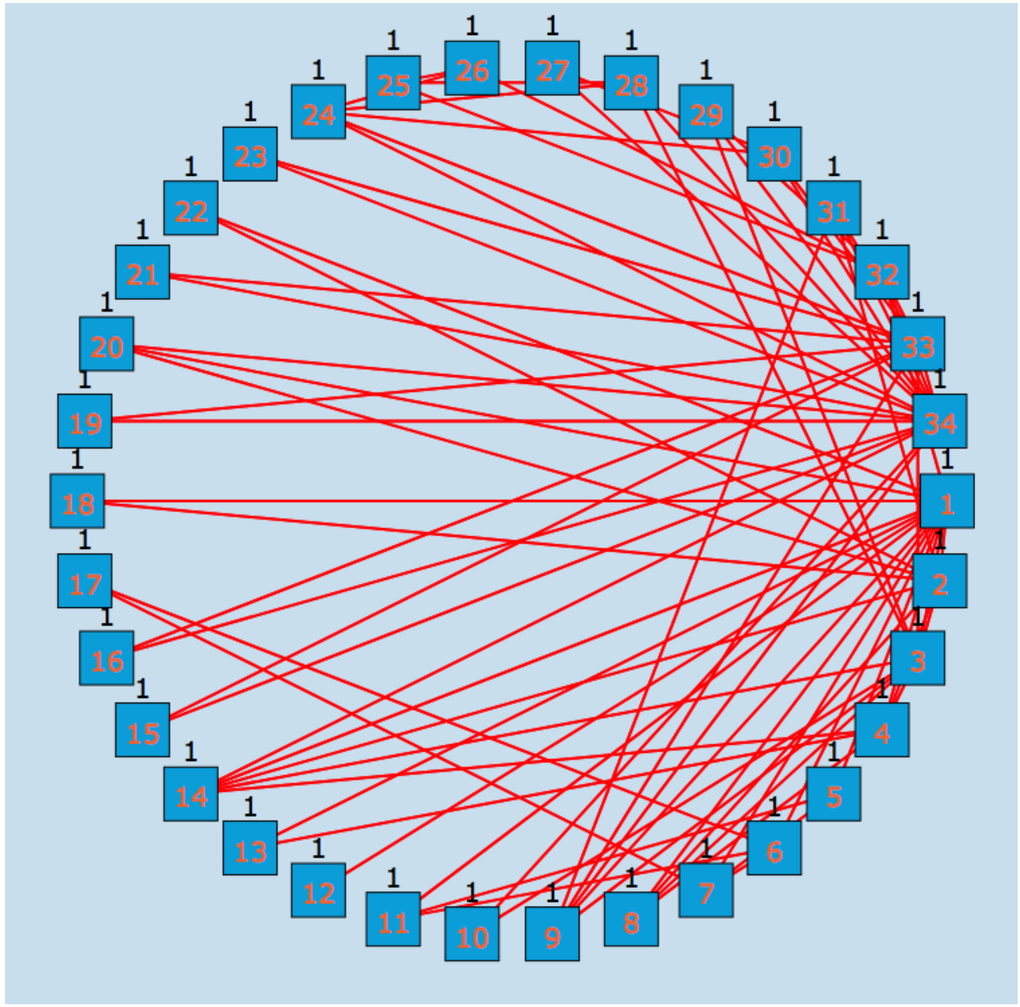
DFS (0)

DFS(0) is completed. Red/grey/blue edge is tree/cross/forward/back edge of the DFS spanning tree, respectively.

```
DFS (u)
for each neighbor v of u
  if v is unvisited, tree edge, DFS (v)
  else if v is explored, bidirectional / back edge
  else if is visited, forward / cross edge
// ch4_01_dfs.cpp / java, ch4, CP3
```

Visualizing LP?

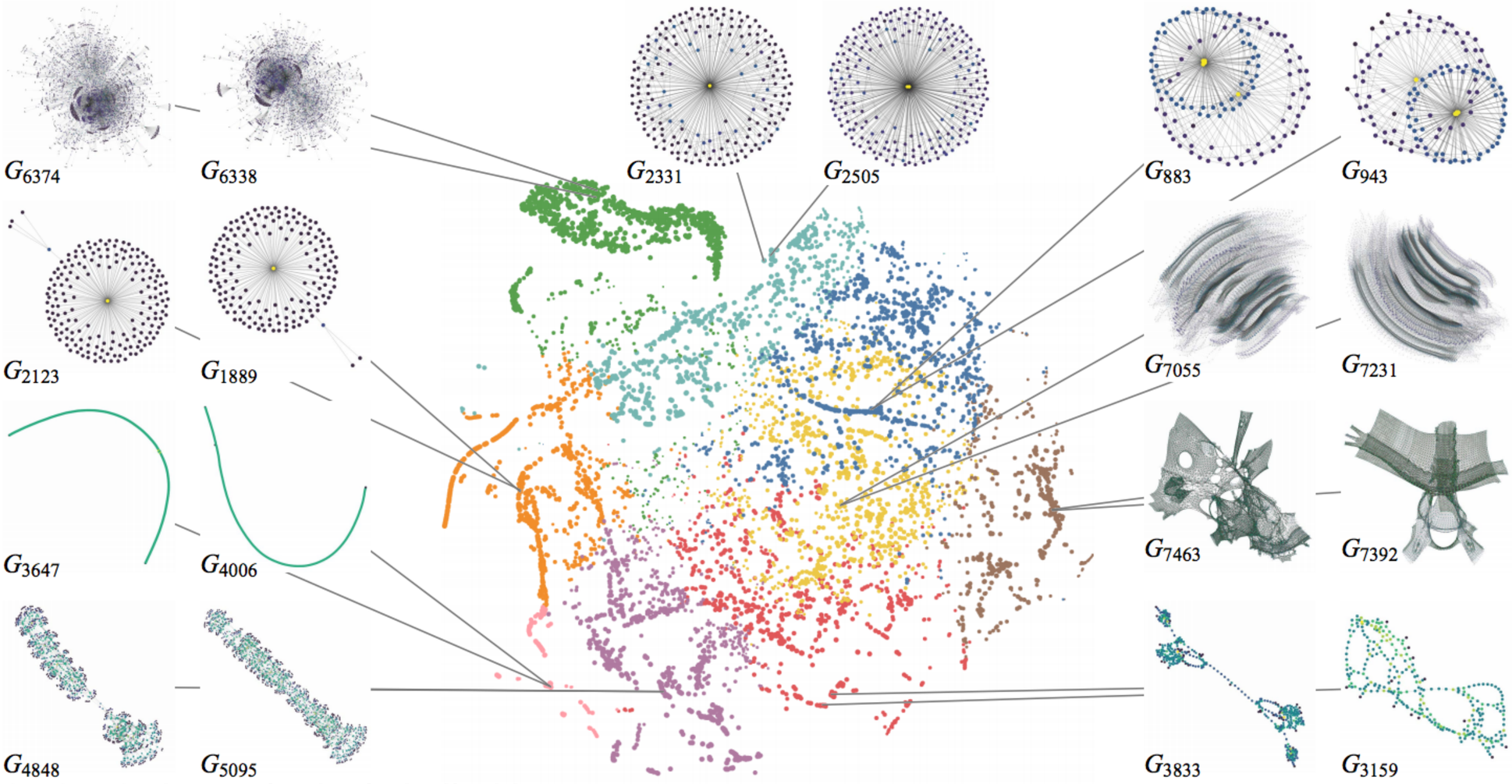
- Visualization of random walks on graphs?



http://www.opcoast.com/demos/label_propagation/index.html

ML approach to large graph visualization

Machine learning approach to large graph visualization



Cluster graphs by their topological similarities.

https://arxiv.org/abs/1710.04328

Graphlets

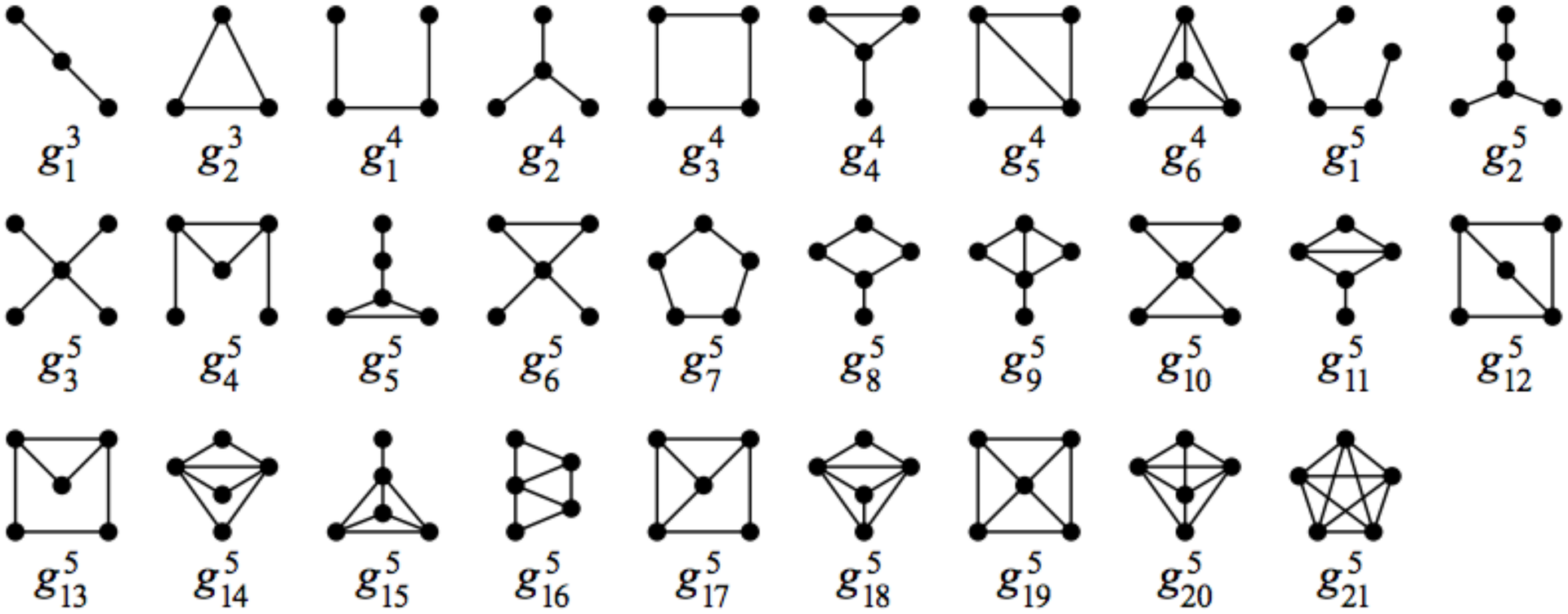
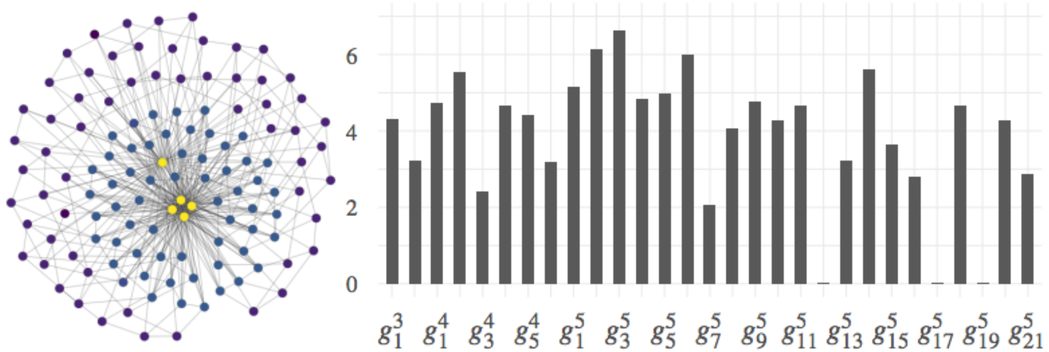


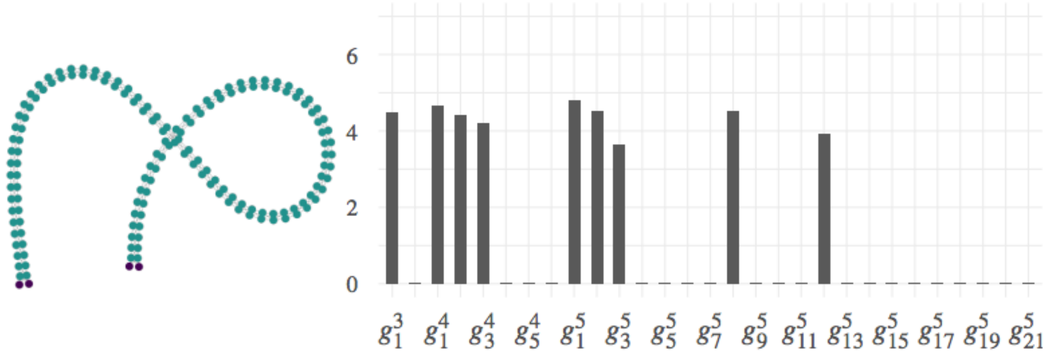
Fig. 2. All connected graphlets of 3, 4, or 5 vertices.

https://arxiv.org/abs/1710.04328

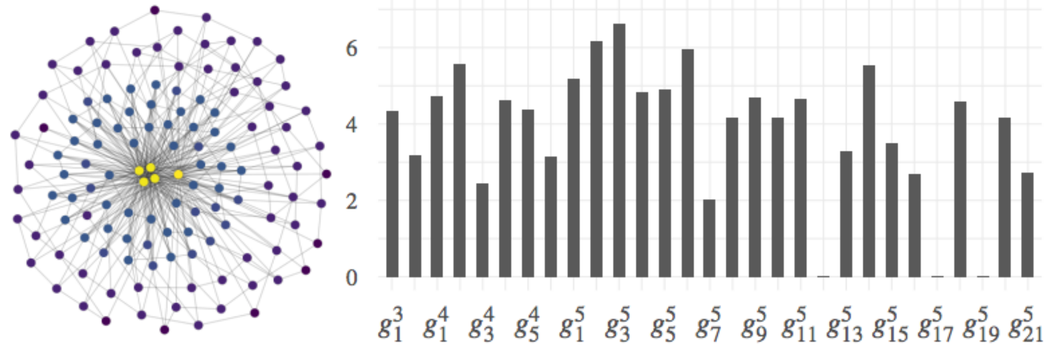
Graphlets Frequencies



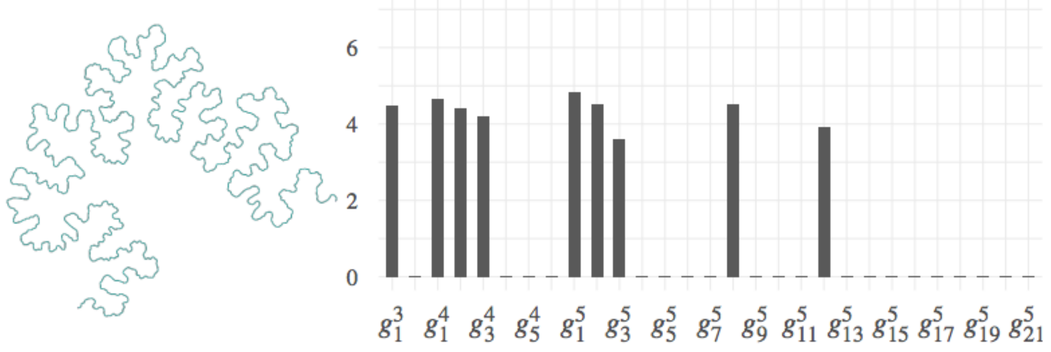
(a) G_{883} $|V| = 122$ $|E| = 472$



(b) G_{1788} $|V| = 158$ $|E| = 312$



(c) G_{943} $|V| = 124$ $|E| = 462$



(d) G_{7208} $|V| = 19,998$ $|E| = 39,992$

Fig. 3. Examples of graphlet frequencies. The x-axis represents connected graphlets of size $k \in \{3,4,5\}$ and the y-axis represents the weighted frequency of each graphlet. Four graphs are drawn with sfdp layouts [45]. If two graphs have similar graphlet frequencies, i.e., high topological similarity, they tend to have similar layout results (a and c). If not, the layout results look different (a and b). However, in rare instances, two graphs can have similar graphlet frequencies (b and d), but vary in graph size, which might lead to different looking layouts.

https://arxiv.org/abs/1710.04328

Topological similarities: kernels

Cosine similarity (COS): Most existing graphlet kernels use the dot product of two graphlet frequency vectors in Euclidean space, then normalize the kernel matrix. This is equivalent to the cosine similarity of two vectors, which is the L_2-normalized dot product of two vectors:

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \frac{\mathbf{x} \cdot \mathbf{x}'^{\mathsf{T}}}{\|\mathbf{x}\| \, \|\mathbf{x}'\|}$$

Gaussian radial basis function kernel (RBF): This kernel is popularly used in various kernelized machine learning techniques:

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2} \right)$$

where σ is a free parameter.

Laplacian kernel (LAPLACIAN): Laplacian kernel is a variant of RBF kernel:

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma} \right)$$

where $\|x - x'\|_1$ is the L_1 distance, or Manhattan distance, of the two vectors.

Treating graphlets frequencies as feature vectors. Define dot product of two graphlet frequency vectors.

https://arxiv.org/abs/1710.04328

Aesthetic criteria for graph VIS

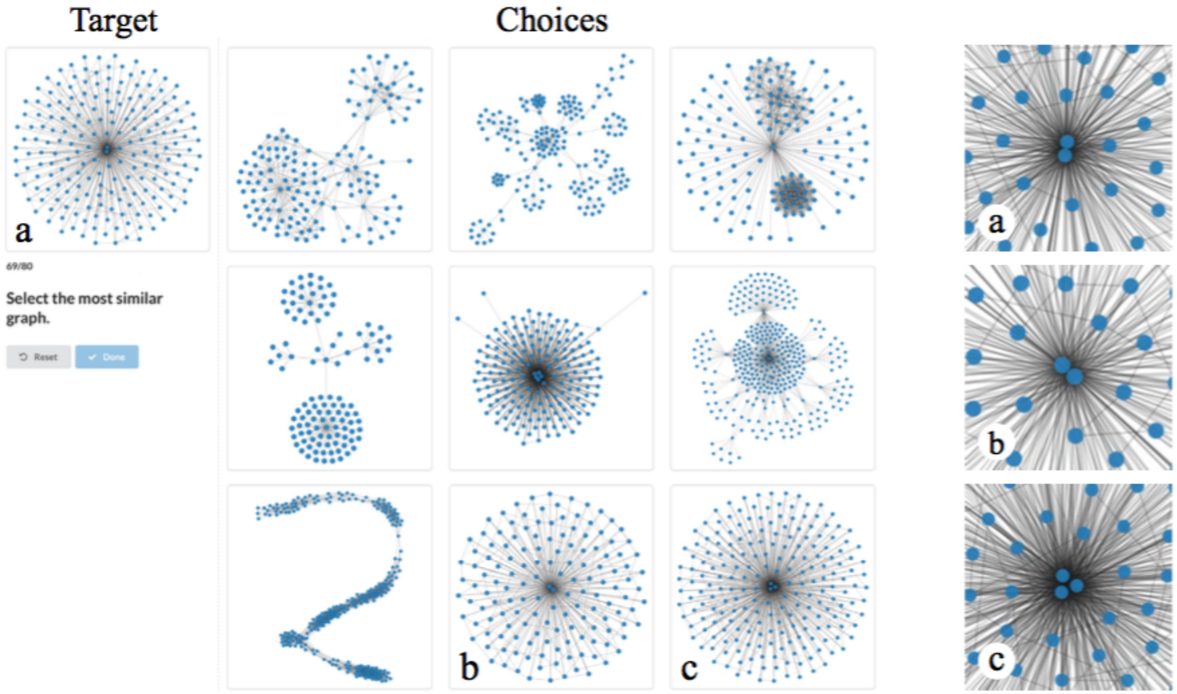
- Crosslessness: Minimizing the number of edge crossings
- Minimum angle metric: maximizing the minimum angle between incident edges on a vertex.
- Edge length variation: Uniform edge lengths.
- Shape-based metric: Mean Jaccard similarity between the input graph and the shape graph.

$$\mathrm{MJS}(G_1, G_2) = \frac{1}{|V|} \sum_{v \in V} \frac{|N_1(v) \cap N_2(v)|}{|N_1(v) \cup N_2(v)|}$$

https://arxiv.org/abs/1710.04328

What would a graph look like in this layout?

- Compare accuracy of ML predicted similar graph layout vs human chosen layout



https://arxiv.org/abs/1710.04328

Apply ML to Graph Drawing

Existing approaches

- Approaches that learn from human interaction: learn information from users based on their interactions to a graph drawing system.
- Approaches that are not based on human interaction: gather and evolve knowledge about how to draw a graph from the results of other automatic graph drawing algorithms or from the graph structure itself.

Rate by users

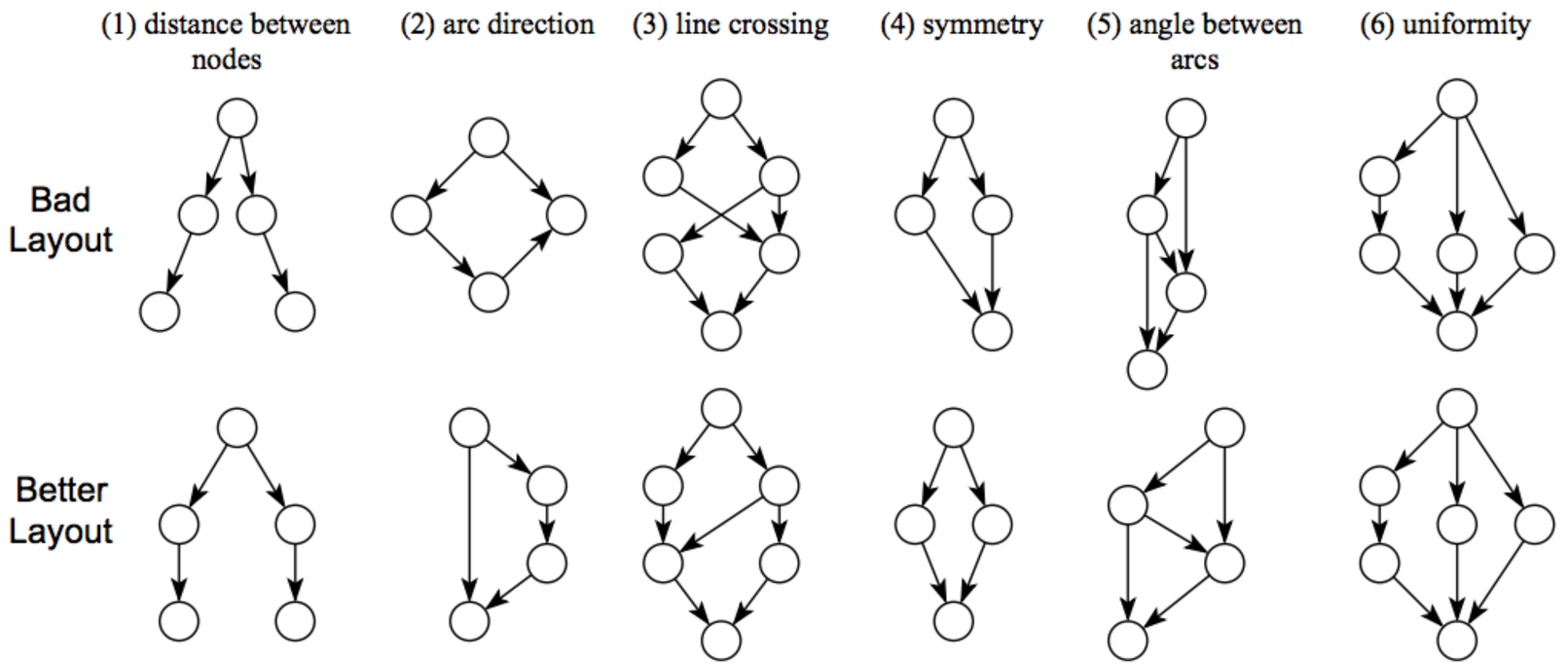


Figure 1: Constraints used in the layout of directed graphs.

https://www.cse.unr.edu/~sushil/class/gas/papers/masui94evolutionary.pdf

Learning optimal graph drawing for clustered graph

- Construct a handcrafted feature vector of a cluster from a number of graph measures: number of vertices, diameter, and maximum vertex degree.
- Find an optimal layout for each cluster.

https://dl.acm.org/citation.cfm?id=345354

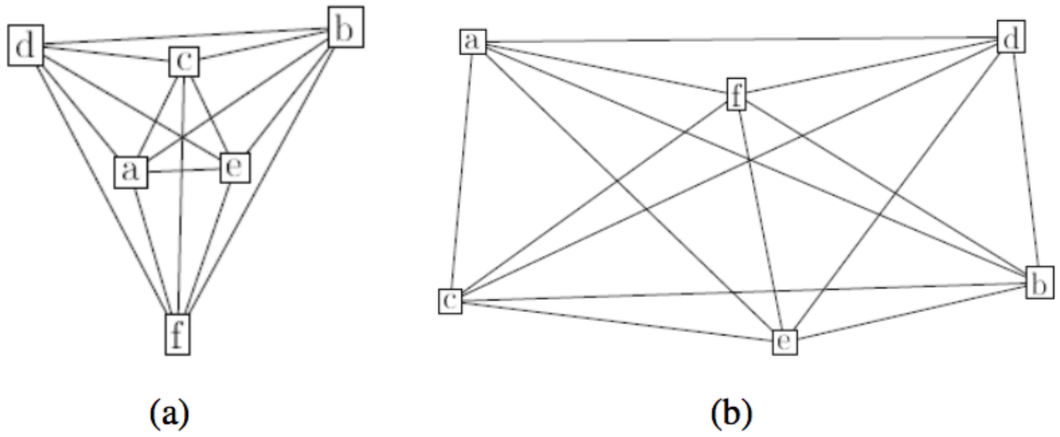
Using ML to improve layout quality

- Human in the loop evaluation.
- Using neural network algorithms to optimize a layout for certain aesthetic criteria.

Self-organizing graphs

B. Meyer: Self-organizing graphs - a neural network perspective of graph layout

- Not building a general neural network that learns aesthetic criteria or hints about how to draw a graph. Instead, it models the graph structure and the drawing problem as a network coupled with an energy system.



(a) (b)

Future directions?

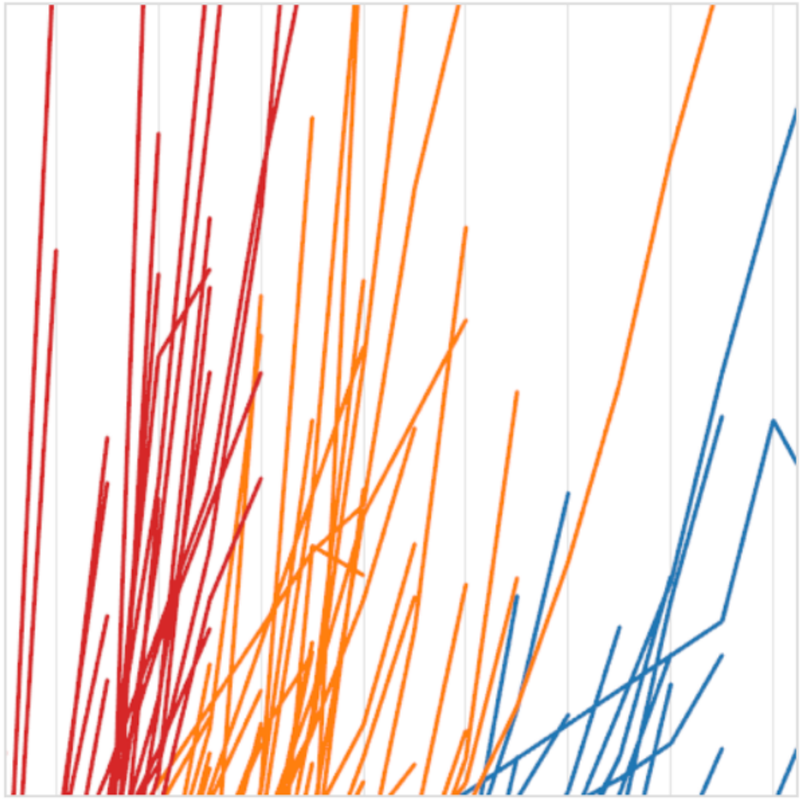
- Using neural network algorithms to optimize a layout for certain criteria
- Really learn how to draw a good graph

The Future of Data Visualization

Current & Next Cool Startups

CUSTOMER STORY

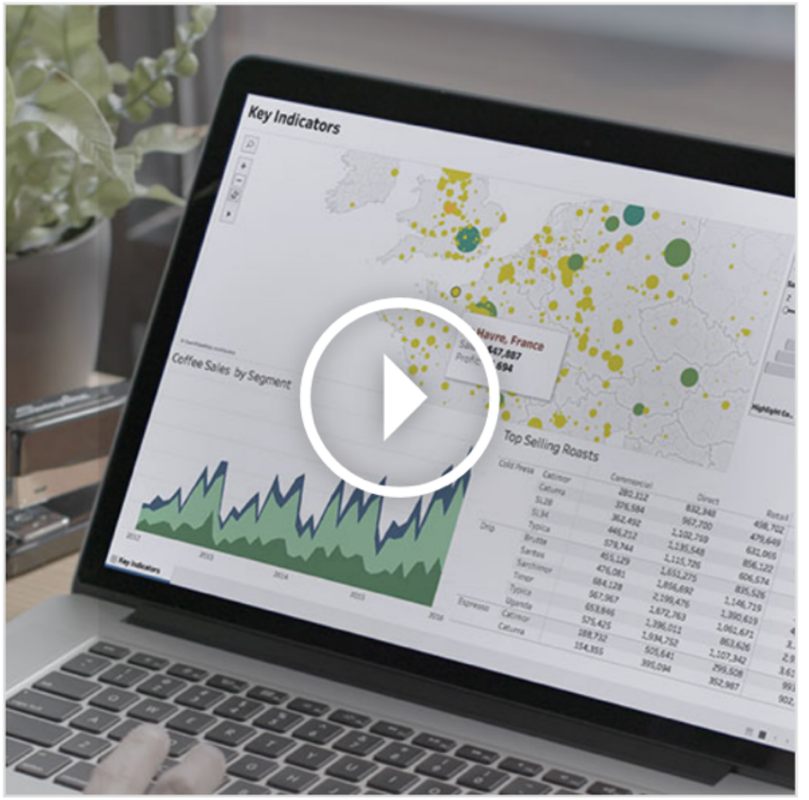
Wells Fargo wrangles data from over 70 million customers to redesign customer banking portal



VISUALIZATION

Test the myth of tech companies' 'rocket-ship' growth

CLICK TO INTERACT →



PRODUCT VIDEO

Create rich analyses and share your insights with colleagues in seconds

https://www.tableau.com/

OCT 23, 2017 @ 09:15 AM 16,848

The Little Black Book of Billionaire Secrets

Startup Raised $180M To Take On Tableau Software In $18B Analytics Market

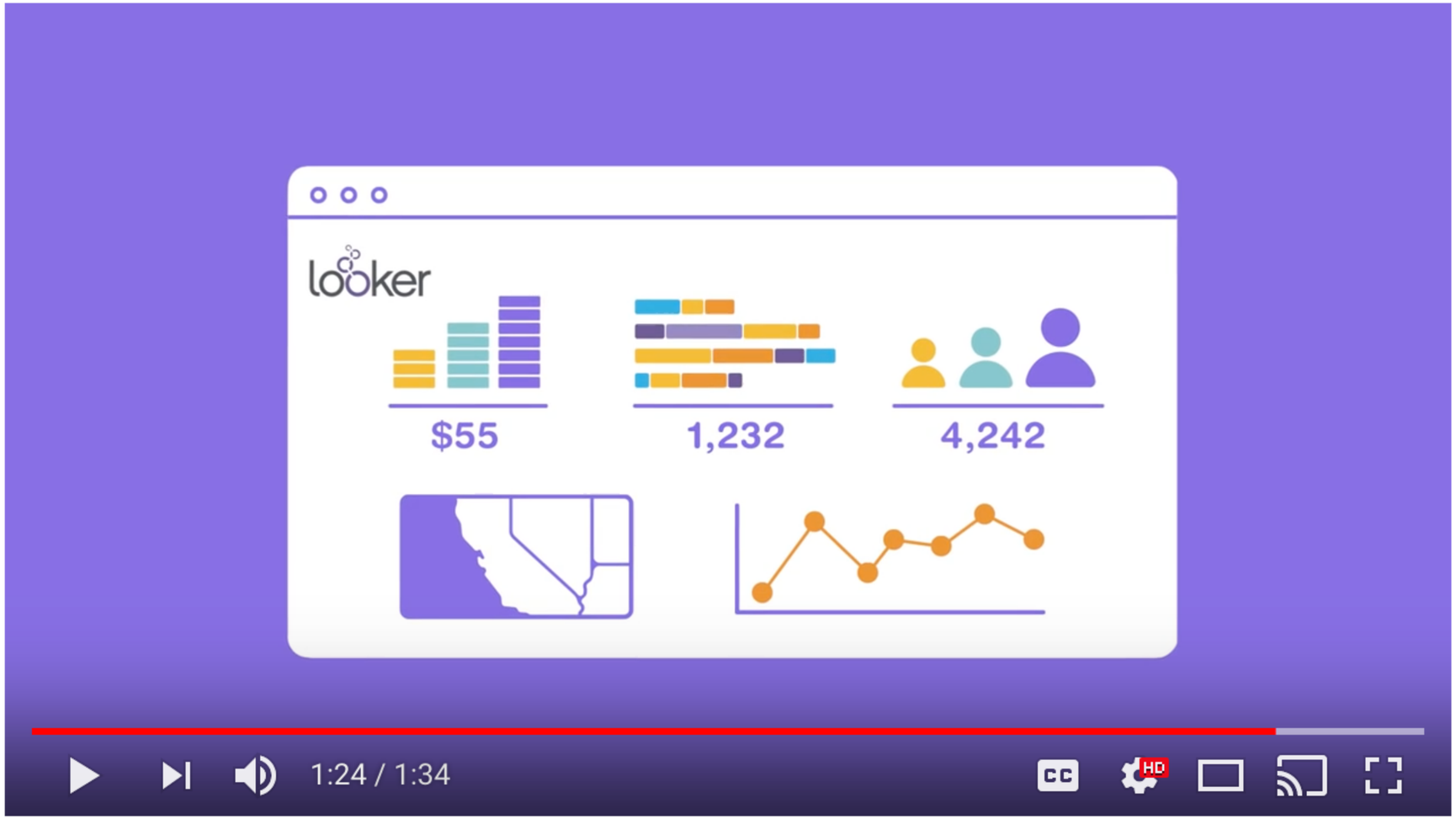
Peter Cohan, CONTRIBUTOR

FULL BIO

Opinions expressed by Forbes Contributors are their own.

Business intelligence and data analytics are popular buzzwords. But businesses face considerable obstacles to turning the buzz into better decisions. Now Looker, a Santa Cruz, Calif.-based supplier of analytics tools, is taking on publicly-traded Seattle-based Tableau Software in the $18 billion market for data analytics that includes business intelligence (BI) and data visualization, according to Gartner. Is Looker a threat to Tableau or is there room enough for many players in this industry?

https://www.forbes.com/sites/petercohan/2017/10/23/startup-raised-180-million-to-take-on-tableau-software-in-4-billion-data-visualization-market/#73162865ba2d



Looker Overview

https://www.youtube.com/watch?v=krXaBEi3f1s

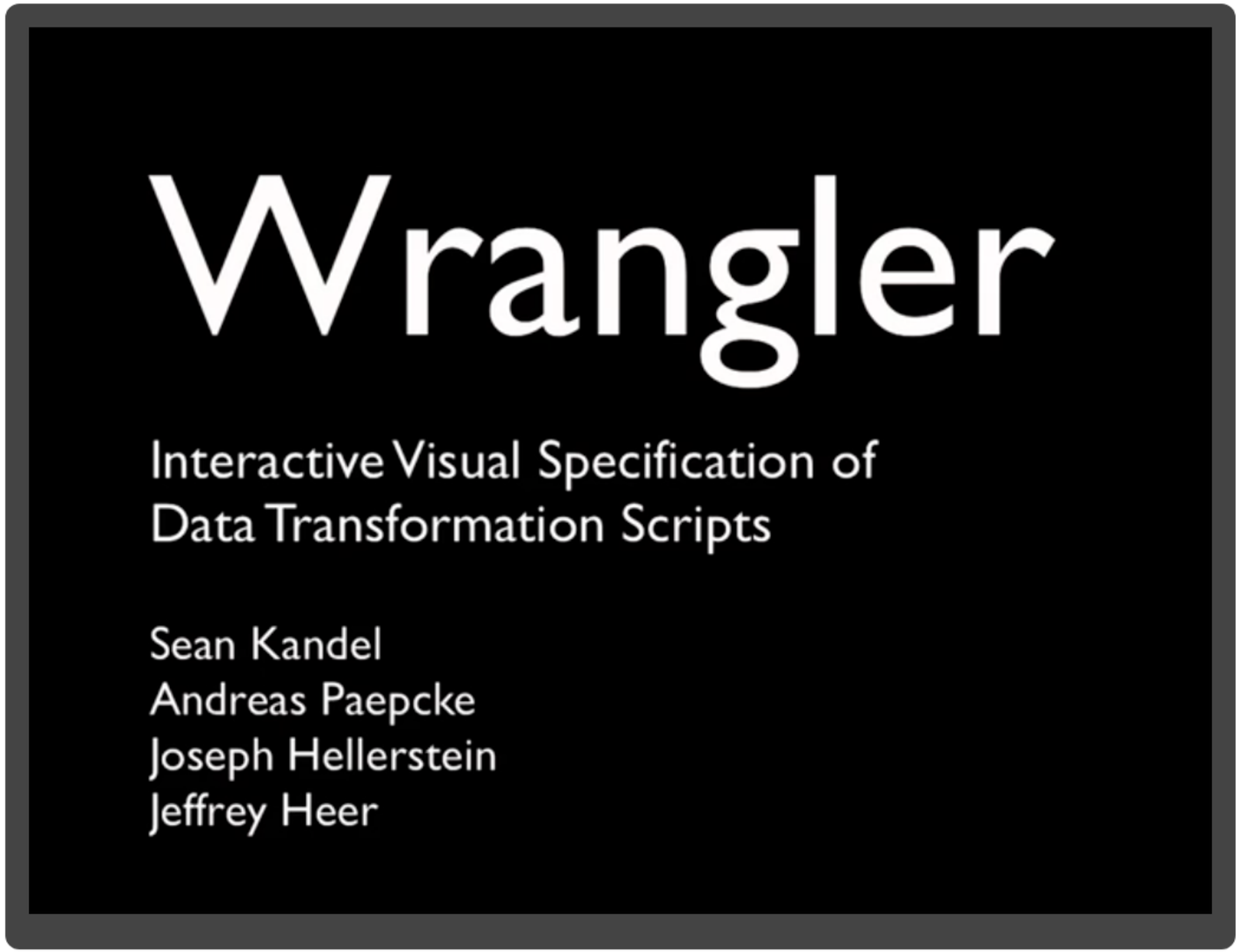
https://looker.com/company

Data Wrangler

- Accelerate data manipulation: spend less time fighting with your data and more time learning from it.
- Allow interactive transformation of messy, real-world data into the data tables analysis tools expect.

http://vis.stanford.edu/wrangler/

Data Wrangler



http://vis.stanford.edu/wrangler/

Trifacta

What We Do

Our focus is to create radical productivity for people who work with data.

Wrangling data is the most time-consuming and inefficient part of any data project – taking up over 80% of the time and resources. Trifacta enables anyone to more efficiently explore and prepare the diverse data of today by utilizing machine learning to provide a breakthrough user experience, workflow and architecture.

JOIN TOGETHER DISPARATE DATA SOURCES

ONBOARD EXTERNAL OR 3RD-PARTY INFORMATION

CLEAN RAW AND MESSY DATA

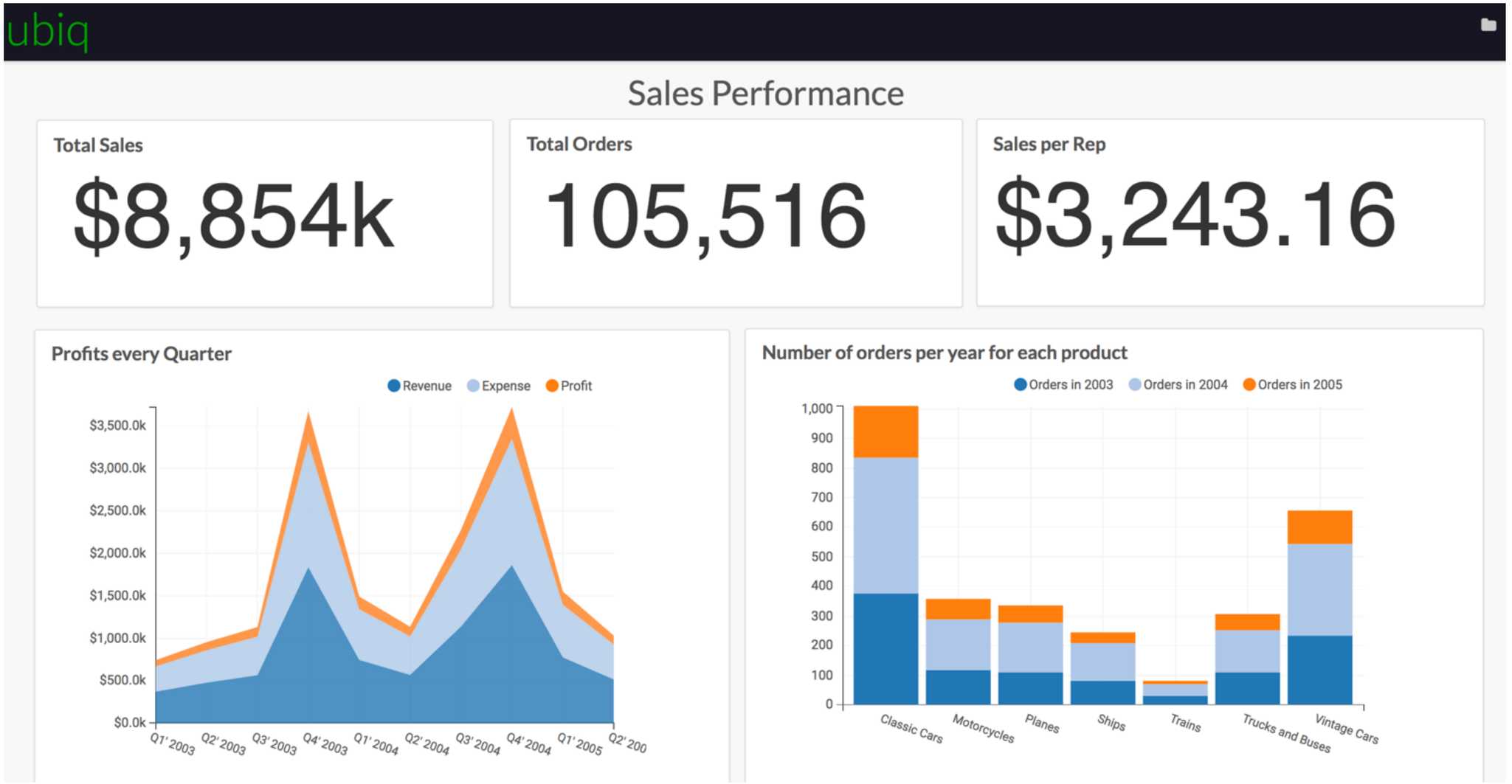
https://www.trifacta.com/

Trifacta

"The Future of Data Visualization" - Jeffrey Heer (Strata + Hadoop 2015)

https://www.youtube.com/watch?v=vc1bq0qIKoA

ubiq


ubiq
Sales Performance
Total Sales
$8,854k
Total Orders
105,516
Sales per Rep
$3,243.16
Profits every Quarter
Revenue
Expense
Profit
$3,500.0k
$3,000.0k
$2,500.0k
$2,000.0k
$1,500.0k
$1,000.0k
$500.0k
$0.0k
Q1' 2003
Q2' 2003
Q3' 2003
Q4' 2003
Q1' 2004
Q2' 2004
Q3' 2004
Q4' 2004
Q1' 2005
Q2' 200
Number of orders per year for each product
Orders in 2003
Orders in 2004
Orders in 2005
1,000
900
800
700
600
500
400
300
200
100
0
Classic Cars
Motorcycles
Planes
Ships
Trains
Trucks and Buses
Vintage Cars

ubiq

- Intuitive UI
- Real-time visualization
- Line, area, bar, scatter plot to gauges and maps
- Share visualizations…

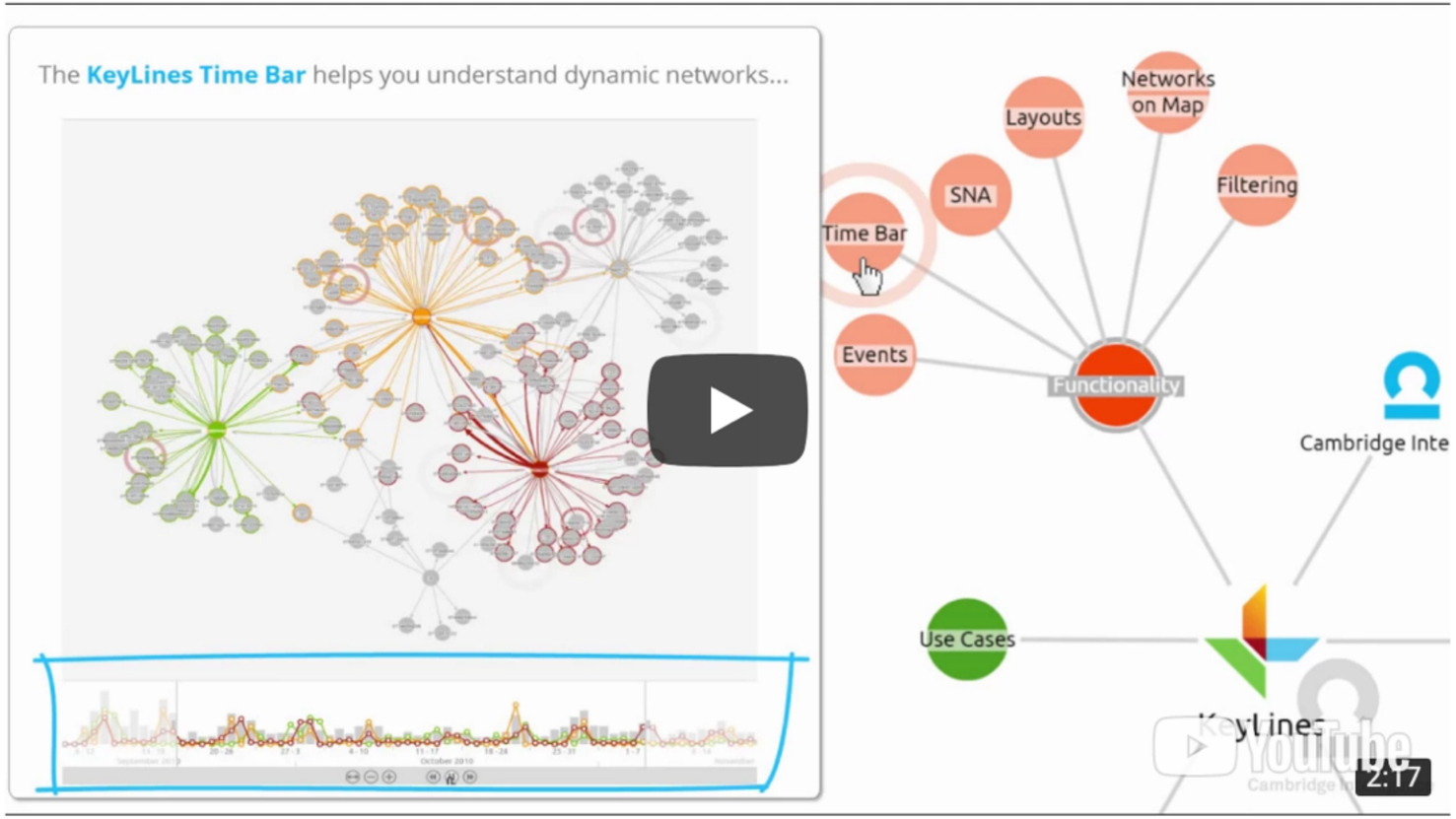
Cambridge Intelligence

Build products with game-changing interactive visualization that turns data into insight.

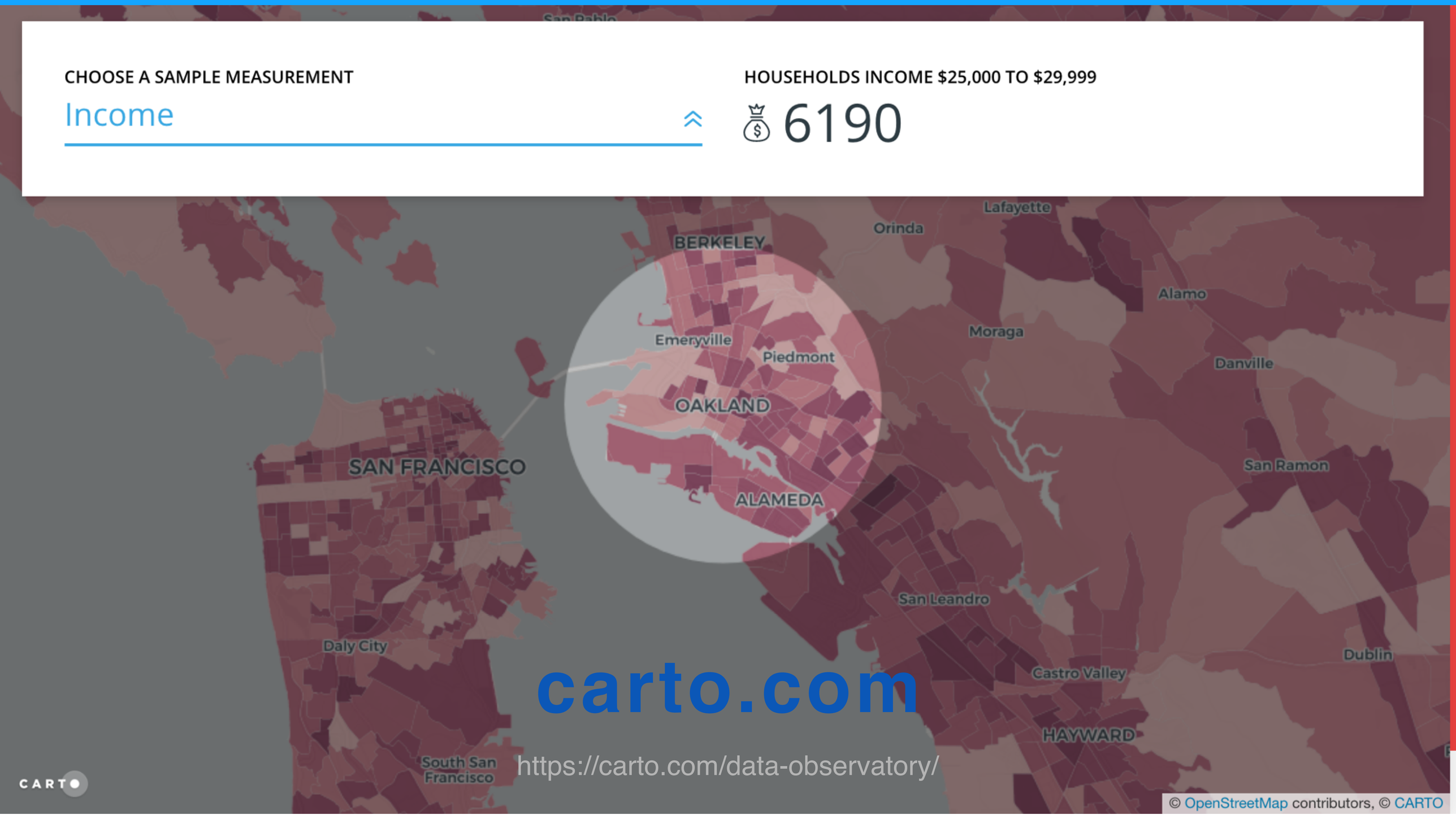
Connected data is all around us. It's in financial transactions, communications records, IT networks and beyond.

The best way to understand it is to visualize it.

Applications built with KeyLines offer new ways to join the dots in your data and uncover valuable buried insights.



https://cambridge-intelligence.com/keylines/


CHOOSE A SAMPLE MEASUREMENT
Income
HOUSEHOLDS INCOME $25,000 TO $29,999
6190
Lafayette
Orinda
BERKELEY
Alamo
Moraga
Emeryville
Piedmont
Danville
OAKLAND
SAN FRANCISCO
San Ramon
ALAMEDA
San Leandro
Daly City
Dublin
Castro Valley
HAYWARD
South San Francisco
carto.com
https://carto.com/data-observatory/
CARTO
© OpenStreetMap contributors, © CARTO

TOPICS:

Health Risks

RISK INDICATORS:

Opioid Overdose Death Rate (Age-Adjusted)

STATE

Opioid overdose death rate per 100,000 population (age-adjusted).

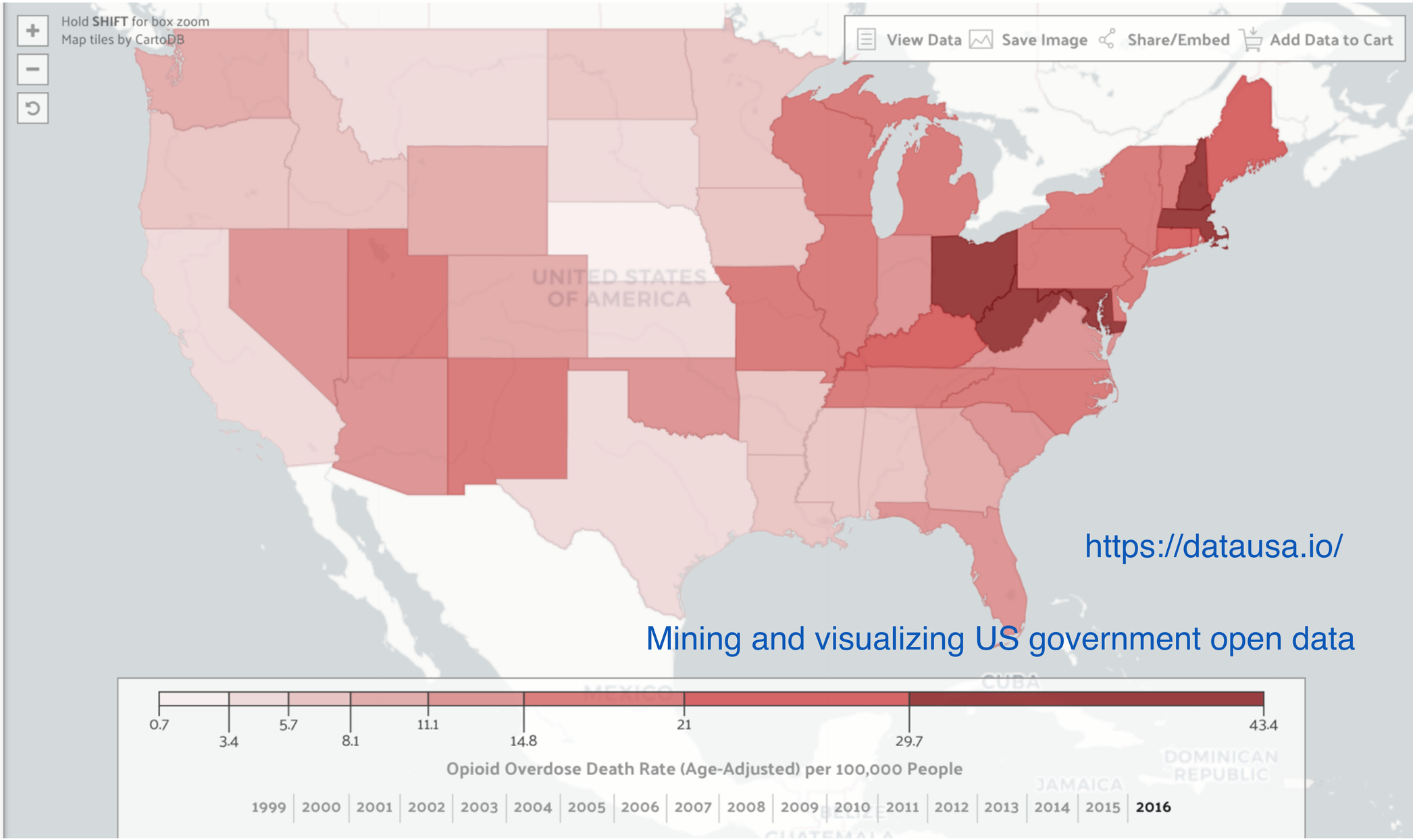
Top 10 Locations

1.	West Virginia	43.4
2.	New Hampshire	35.8
3.	Ohio	32.9
4.	District of Columbia	30
5.	Maryland	29.7
6.	Massachusetts	29.7
7.	Rhode Island	26.7
8.	Maine	25.2
9.	Connecticut	24.5
10.	Kentucky	23.6

Bottom 10 Locations

42.	Iowa	6.2

Dataset: Kaiser Family Foundation analysis of Centers for Disease Control and Prevention (CDC), National Center for Health Statistics



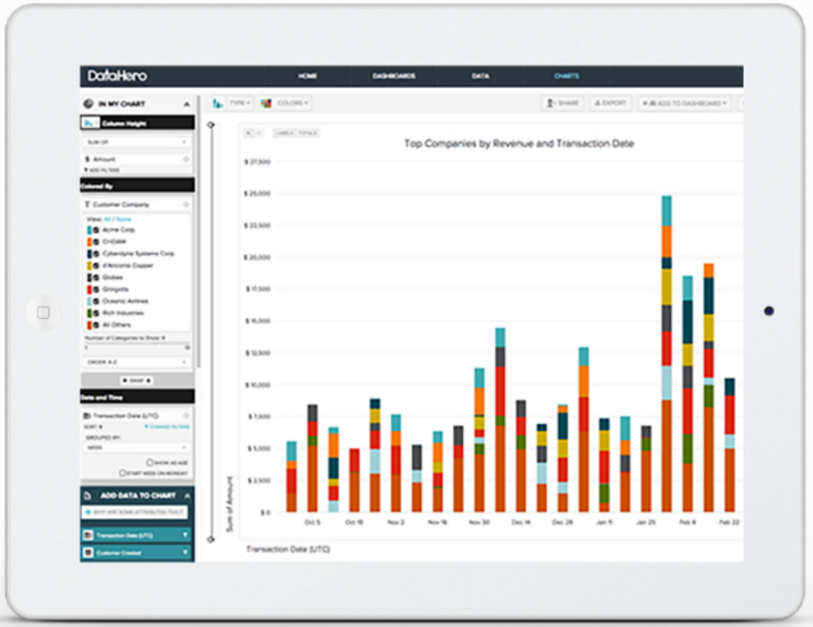
https://datausa.io/

Mining and visualizing US government open data

https://datausa.io/map/?level=state&key=opioid_overdose_deathrate_ageadjusted

Drag-and-Drop Chart Creation

Get suggested charts based on your specific data within DataHero or simply create your own customized charts from scratch. Segmenting, filtering, cohort analysis and more are at your fingertips. DataHero makes it easy to create beautiful data visualizations and data dashboards that you'll want to share with your team and clients.



Making simple charts for BI

https://datahero.com/

What are hot topics in advanced data visualization?

What are the next startups in data visualization?

http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

1. Design smart

Smart visualization design
Effective visual encoding

2. Better user interface

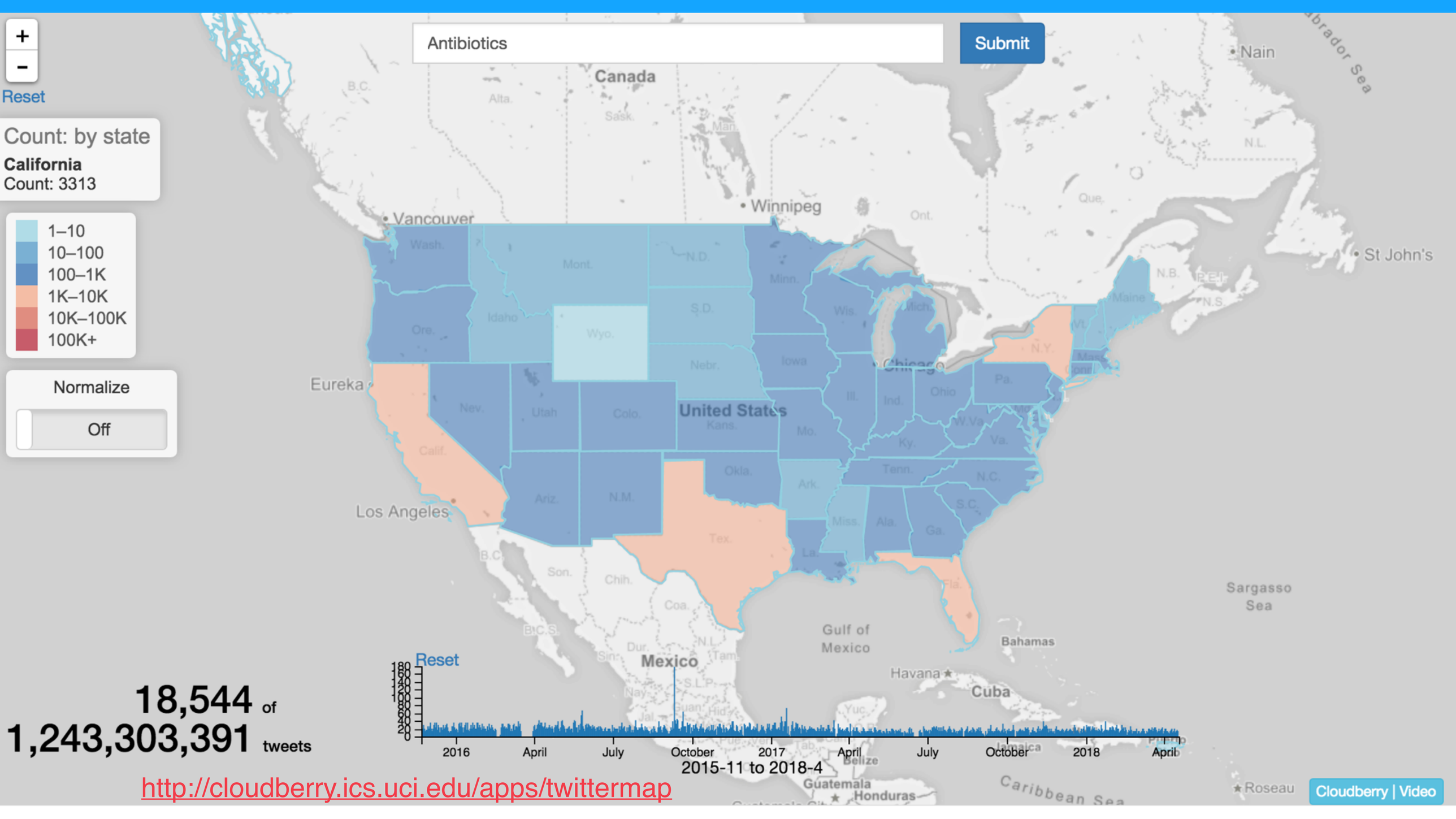
Recommend visualization views
New end user exploration users
No distractions
User-in-the-loop

3.Show data variation, not design variation

By Edward Tufty

4.Scalability

HPC, databases


+
-
Reset
Antibiotics
Submit
Count: by state
California
Count: 3313
1–10
10–100
100–1K
1K–10K
10K–100K
100K+
Normalize
Off
Canada
Vancouver
Winnipeg
Nain
St John's
Eureka
Los Angeles
United States
Mexico
Gulf of Mexico
Bahamas
Havana
Cuba
Sargasso Sea
Caribbean Sea
Guatemala
Honduras
Belize
Roseau
Reset
2016
April
July
October
2017
April
July
October
2018
April
2015-11 to 2018-4
18,544 of
1,243,303,391 tweets
http://cloudberry.ics.uci.edu/apps/twittermap
Cloudberry | Video

5.Visualization that support mobile devices

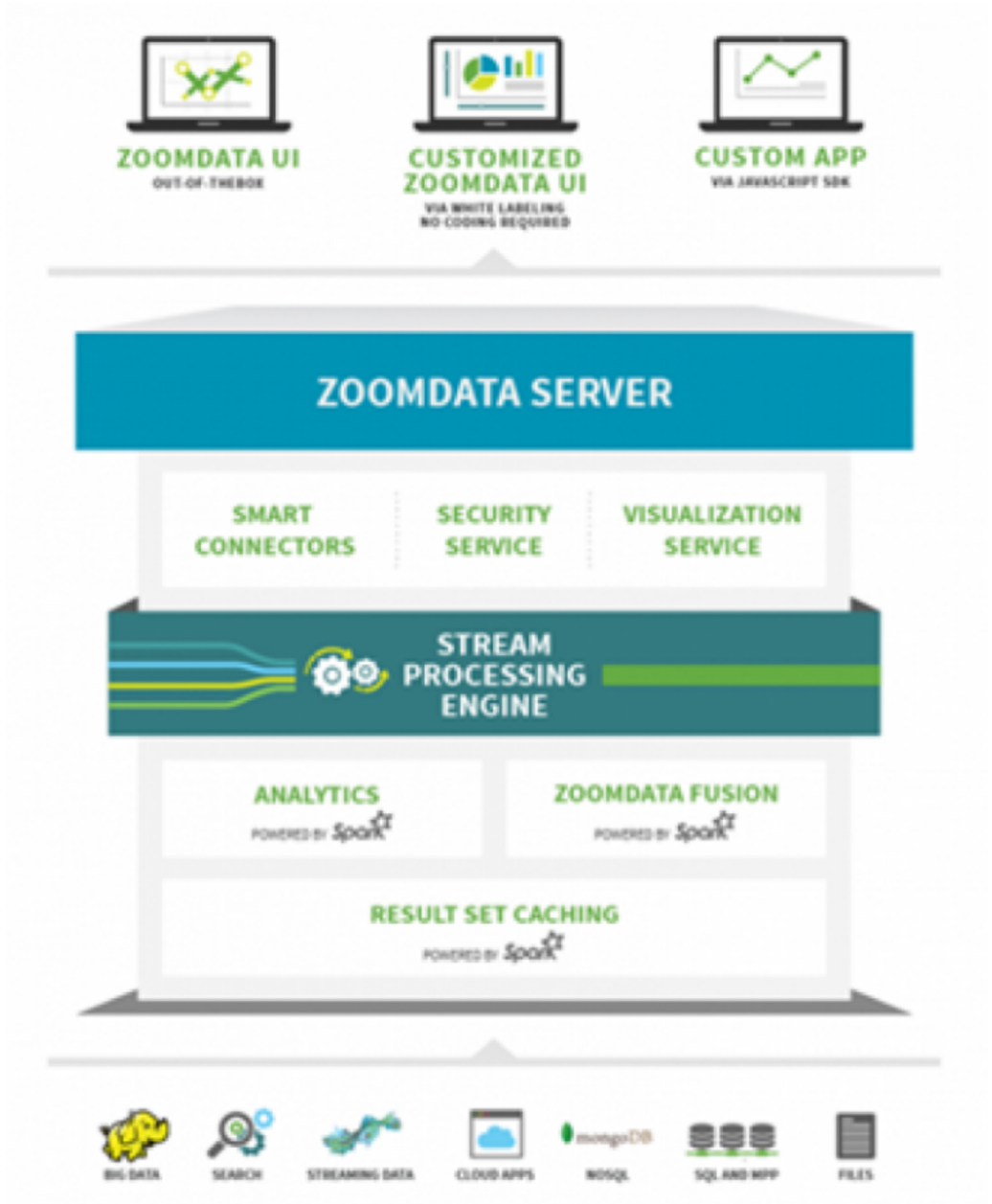
ZOOMDATA®

Master Class | Blog | Developers | Knowledge Base | Support | Contact

Products | Resources | Partners | Company | Free Trials | Demos

Home > Products > What Makes Zoomdata Unique? > Fastest Visual Analytics

https://www.zoomdata.com/



https://www.zoomdata.com/product/zoomdata-architecture/

6.Effective communication

Not just creative visual
Need substances
Edward Tufte: “Every single pixel should testify directly to content.”

http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

7. Integration with other disciplines

The Internet of Things

Tens of billions of devices will be connected to the Internet in the next decade. From smart appliances and wearables to automobile sensors and environmental monitors, the Internet of Things will provide unprecedented insight into what's happening around us. High-throughput, interconnected data streams will help us improve safety, drive operational efficiencies and better understand consumer demand.

In the words of Kevin Ashton, who first coined the term "the Internet of Things" in his seminal 2009 RFID Journal article, "The Internet of Things has the potential to change the world, just as the Internet did. Maybe even more so."

Network Theory

Network Theory builds on Graph Theory, which applies algorithms to understand and model pair-wise relationships between objects. Network Theory examines relationship symmetry, with the existence of asymmetric relationships providing grounds to predict the likely spread of information (social network analysis), dissect complex disorders (biological network analysis), find the shortest path between two points (network optimization) and identify target objects based on their behavior (link analysis).

Complexity Theory

Complexity Theory posits that many systems are characterized by complex, non-linear interactions that evolve dynamically and often unpredictably. Known as the "butterfly effect," small perturbations in one state ("here") can result in large repercussions in a seemingly unrelated state ("there"). According to Complexity Theory, it's impossible to predict with certainty a future state, but it is possible to understand the structure and potential states of complex systems.

Multidimensional Visualization

The adage "a picture is worth a thousand words" gained credence from our ability to process visuals more easily than text. Visualization has also been shown to improve learning and recall, and can portray complex concepts and relationships more easily than can text. Recent developments in computer graphics are making possible visualizations that enable the integration, manipulation and exploration of dynamic multidimensional data sets. Multidimensional visualizations allow users to not only examine data from new perspectives but also interact with it more effectively.

http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

8.Accurate and contextual vis

How data are connected?

http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

9.Facilitate decision making

http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

10. BI and medical domain

Comments
Suggestions?

Thanks!

Any questions?

You can find me at: **beiwang@sci.utah.edu**

CREDITS

Special thanks to all people who made and share these awesome resources for free:

- Presentation template designed by Slidesmash
- Photographs by unsplash.com and pexels.com
- Vector Icons by Matthew Skiles

Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:

http://www.1001fonts.com/oswald-font.html

https://www.fontsquirrel.com/fonts/open-sans

Colors used