Data Visualization

for data scientists

Janu Verma

Data Scientist, Hike

http://jverma.github.io/

J.verma5@gmail.com

@januverma

Data Visualization

- Visual representation of data, usually to provide a quantitative and qualitative understanding.

- Data can be numeric, text, speech, images, relations, processes, concepts etc.

- Hard to make sense of data in CSV/dB/table. Dataviz helps in overview, aggregation, dissection, exploration, and understanding of data.

- A (good) picture is worth a thousand words! Humans are better at processing the visual information. (Cleveland). Can't detect patterns from numbers.

- Limited understanding of numbers, especially large numbers e.g. comparison of numbers - 12M vs 2B, astronomical distances and sizes e.g. size of earth and sun.

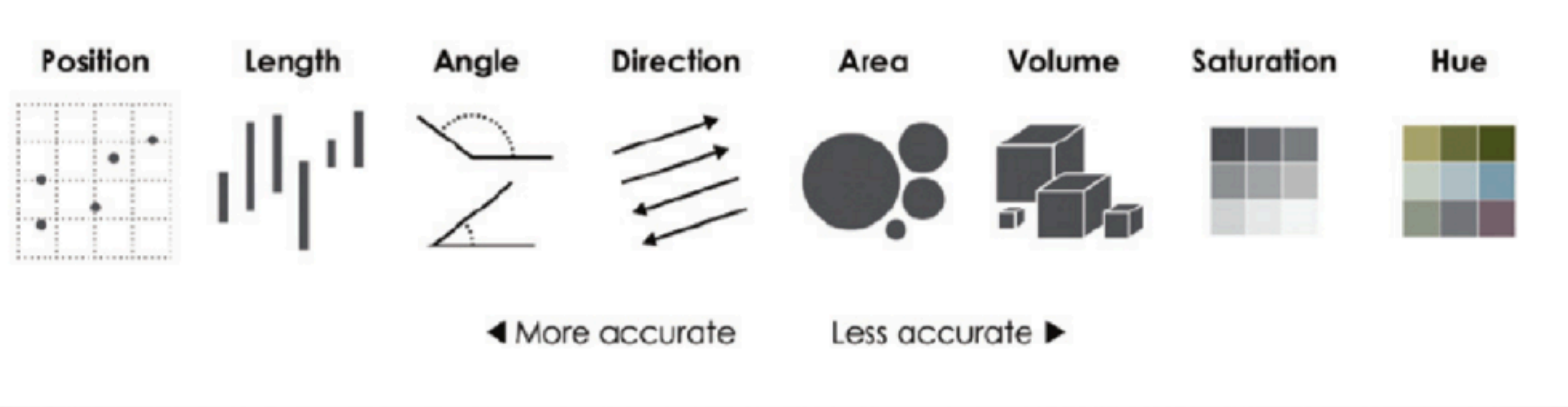
- Visual information stays in memory for longer, we are not good at remembering numbers.

- Data is often high dimensional.

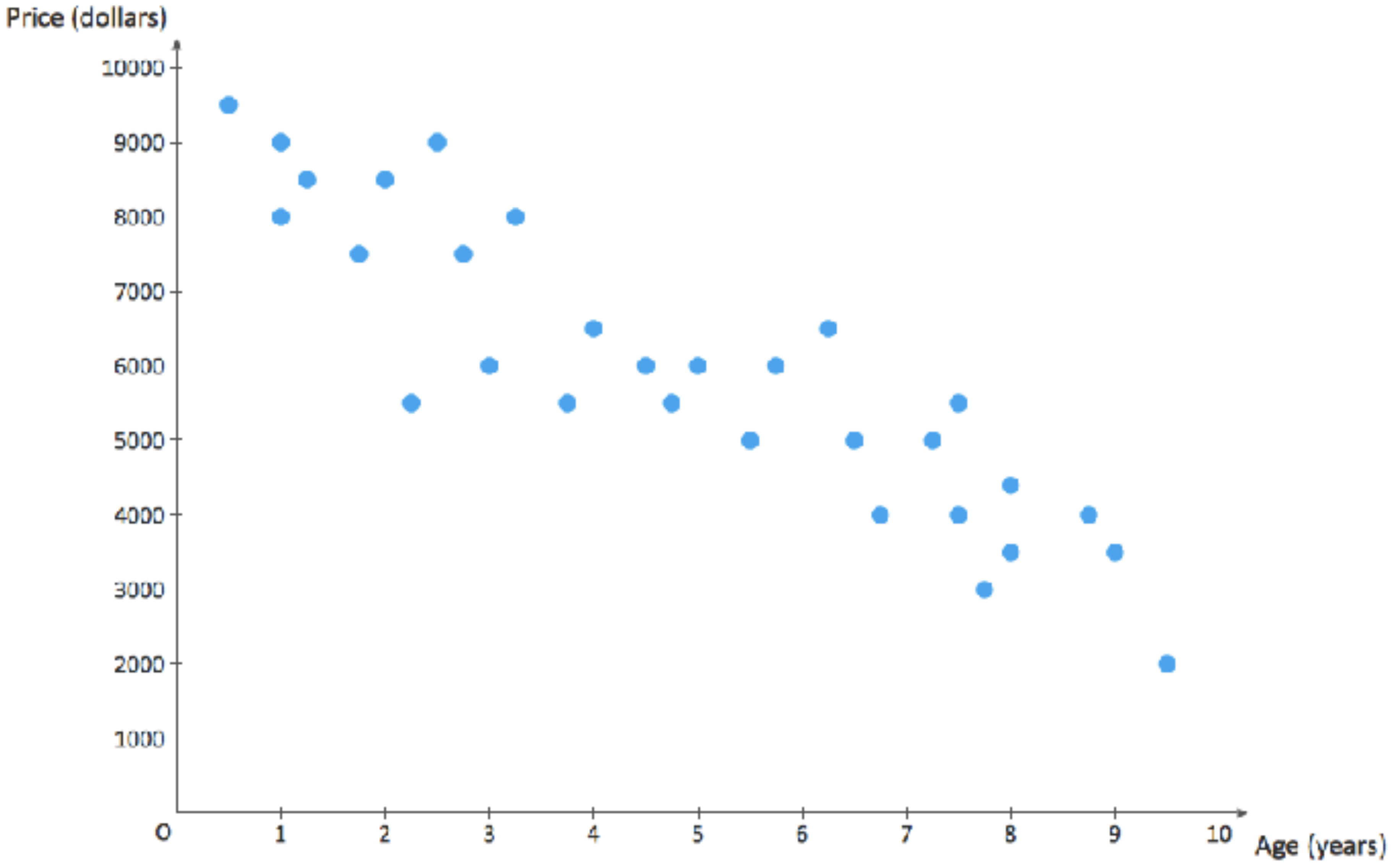
“Data Complex, Visualization Accessible”

Visual encoding for data points - the relationships/patterns/trends translate to graphical properties, which are easier to understand.

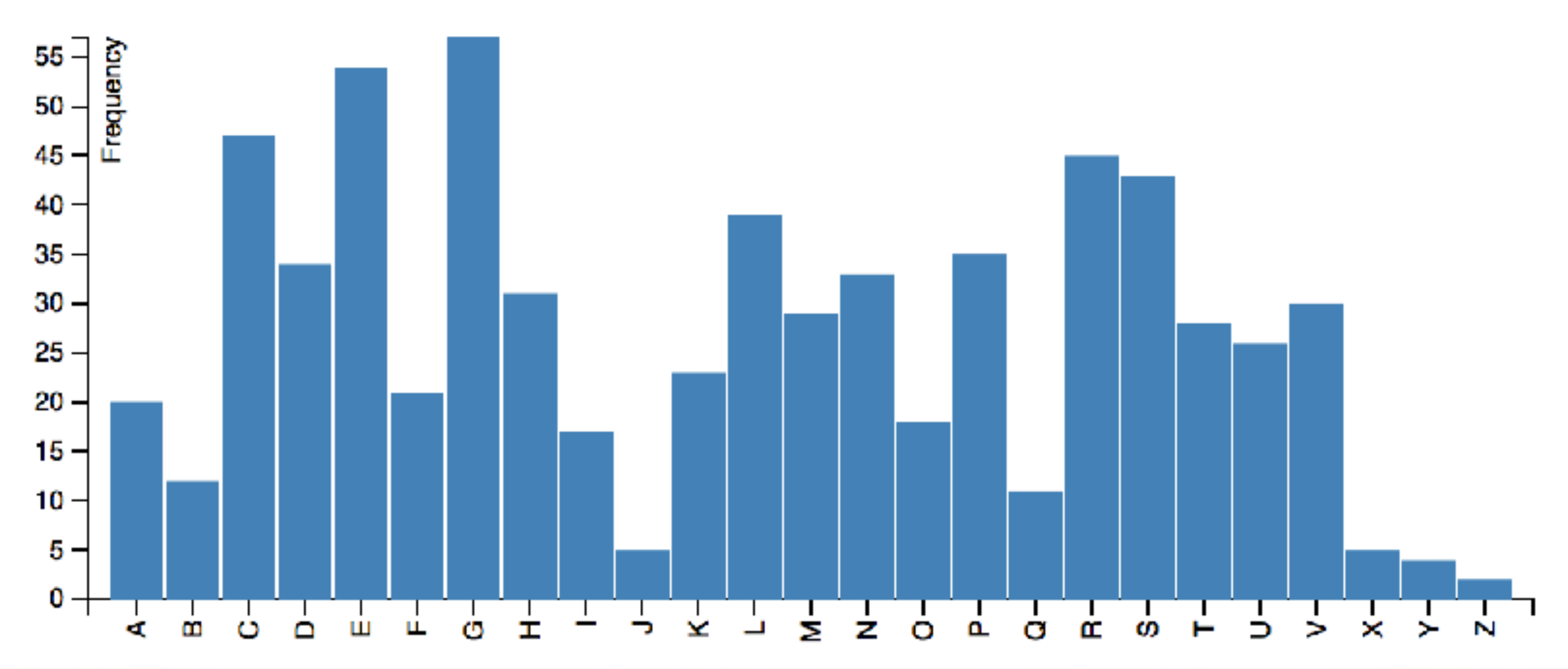
Visual Variables



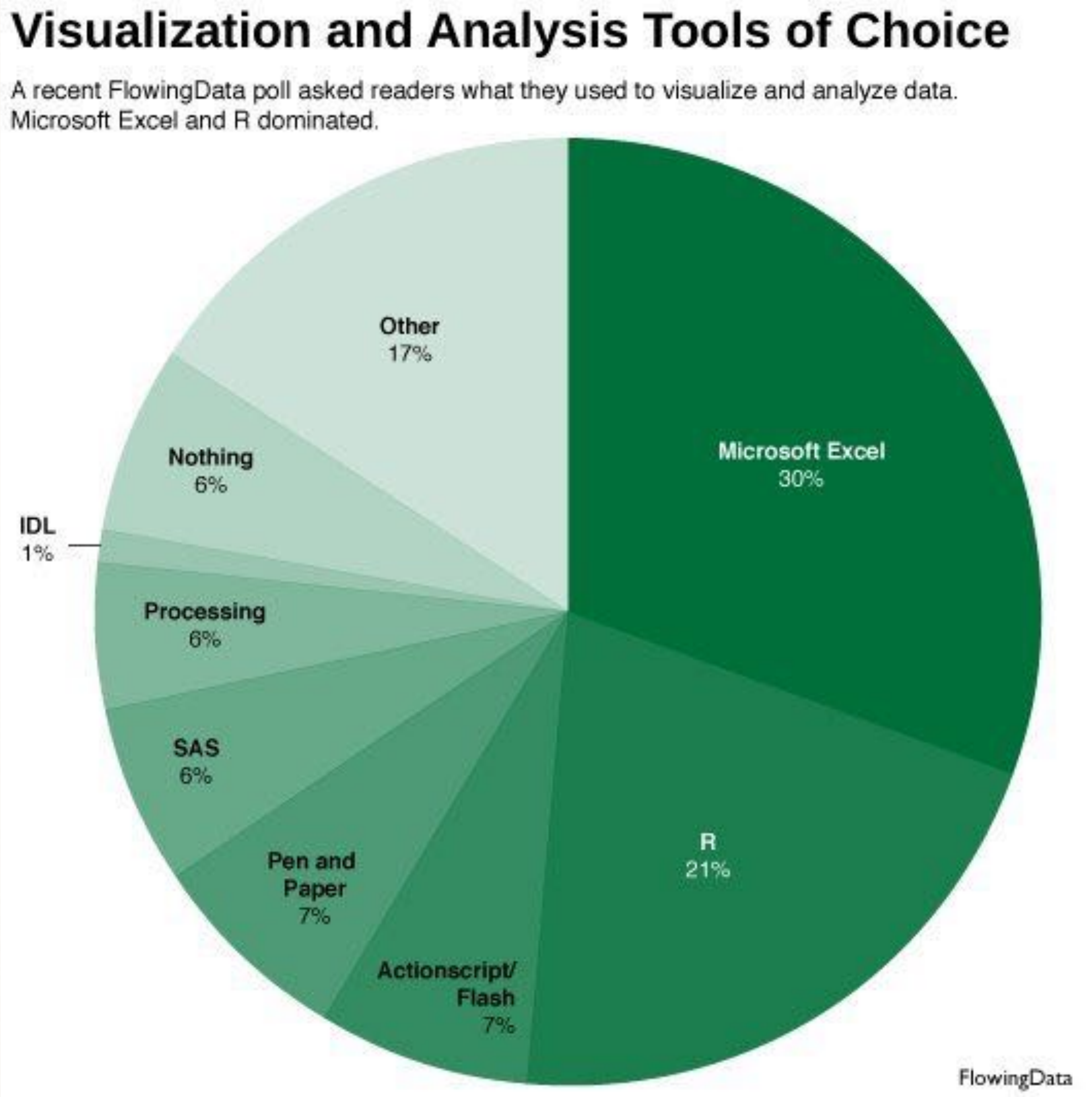
Scatterplot



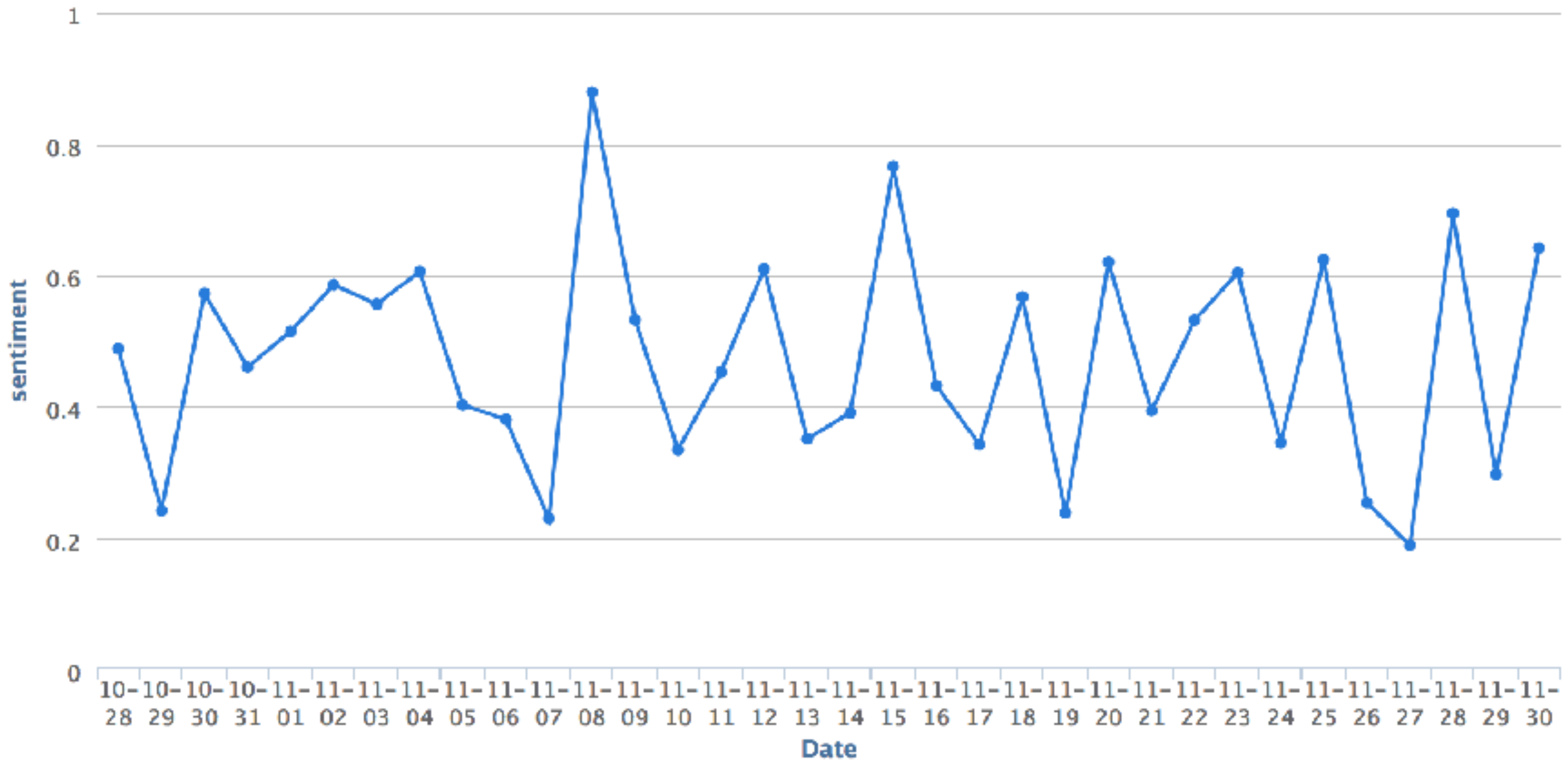
Barchart


Frequency
0
5
10
15
20
25
30
35
40
45
50
55
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
X
Y
Z

Pie Chart

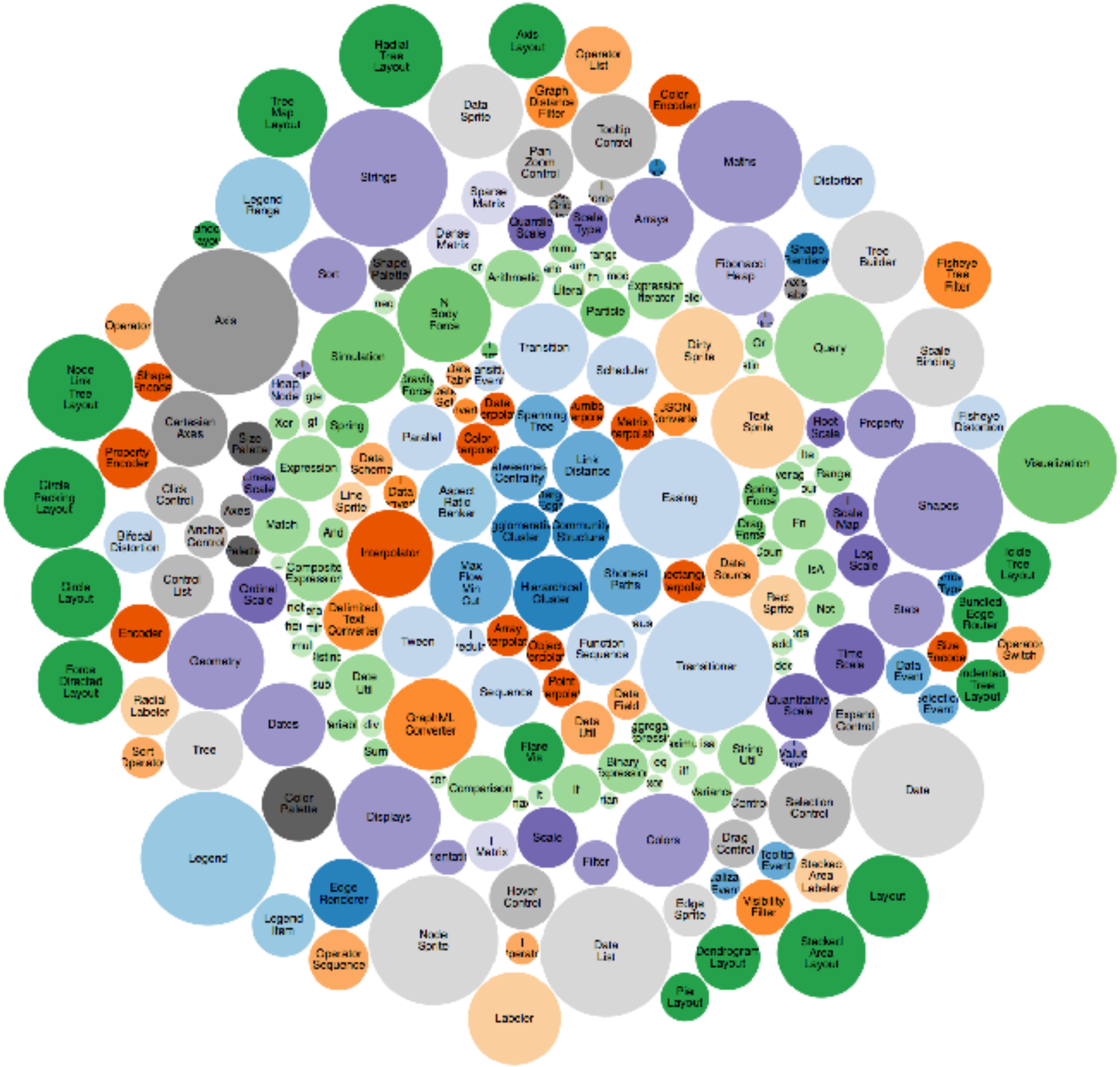

Visualization and Analysis Tools of Choice
A recent FlowingData poll asked readers what they used to visualize and analyze data. Microsoft Excel and R dominated.
Other 17%
Microsoft Excel 30%
Nothing 6%
IDL 1%
Processing 6%
SAS 6%
Pen and Paper 7%
Actionscript/ Flash 7%
R 21%
FlowingData

Line Chart

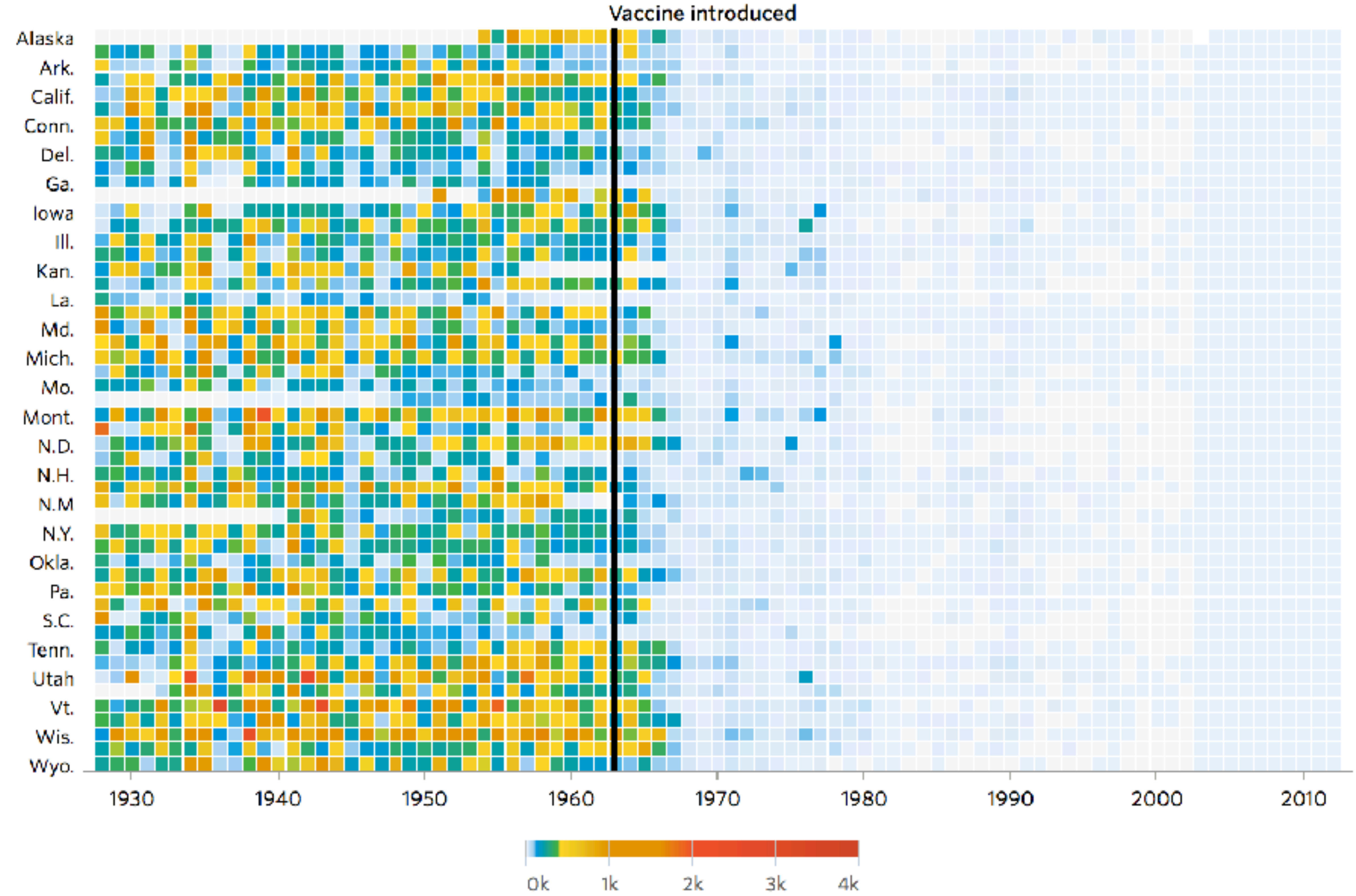


1
0.8
0.6
0.4
0.2
0
sentiment
10-28
10-29
10-30
10-31
11-01
11-02
11-03
11-04
11-05
11-06
11-07
11-08
11-09
11-10
11-11
11-12
11-13
11-14
11-15
11-16
11-17
11-18
11-19
11-20
11-21
11-22
11-23
11-24
11-25
11-26
11-27
11-28
11-29
11-30
Date

Bubble Chart



Measles


Vaccine introduced
Alaska
Ark.
Calif.
Conn.
Del.
Ga.
Iowa
Ill.
Kan.
La.
Md.
Mich.
Mo.
Mont.
N.D.
N.H.
N.M
N.Y.
Okla.
Pa.
S.C.
Tenn.
Utah
Vt.
Wis.
Wyo.
1930
1940
1950
1960
1970
1980
1990
2000
2010
0k
1k
2k
3k
4k

Dataviz

- Visual encoding for data points, the relationships/patterns/trends translate to graphical properties, which are easier to understand.

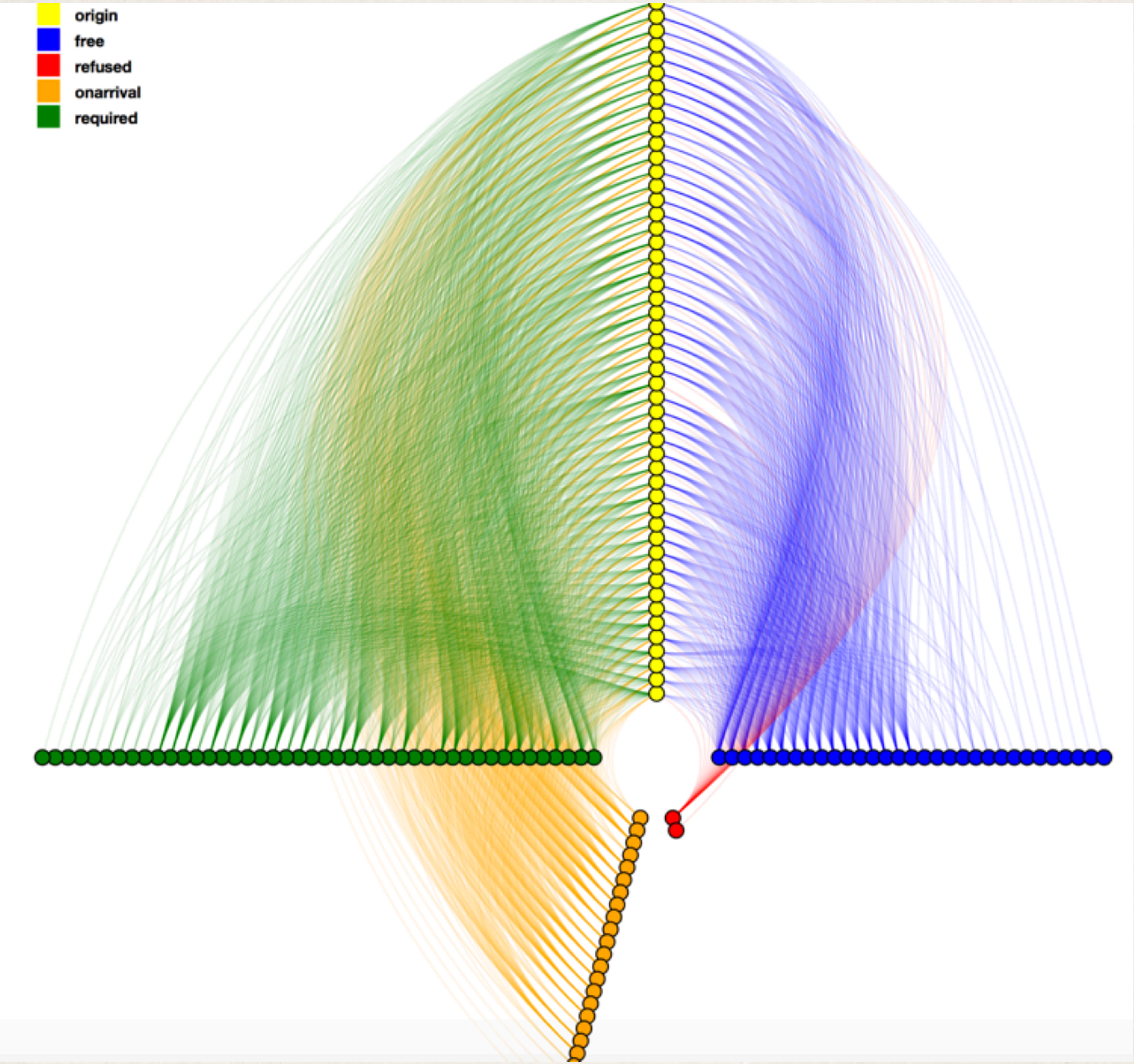
- Understanding may involve detection, measurement, and comparison, finding relationships, patterns, trends, groups, and anomalies and is enhanced via interactive techniques and providing the information from multiple views and with multiple techniques

- Often used for exploration and decision making. Exploratory and Explanatory dataviz. (Hadley Wickham)

- Coming up with good questions - vis tells you what, we can investigate further for how and why.

"The best data visualization should raise questions and inspire exploration, not just sum up information or try to tell us the answer."

– Jake Porway, Cofounder DataKind

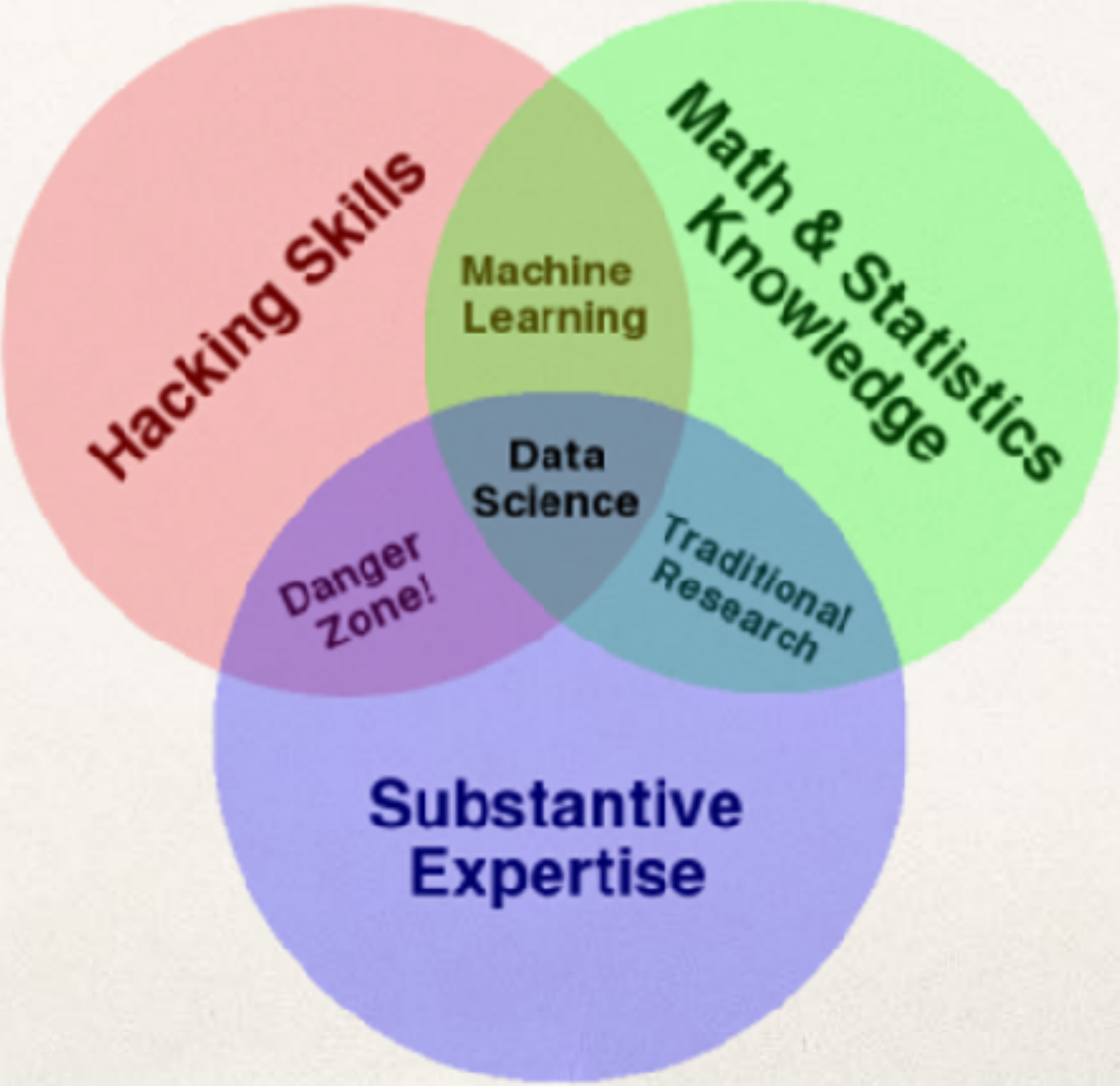


Visafree,

JV and Bryan Bischof

Data Science

Involves using automated methods to analyse massive amount of data and to extract knowledge from it.



Drew Conway

“Unlike any time before in our lives, we have access to vast amounts of free information. With the right tools, we can start to make sense of all this data to see patterns and trends that would otherwise be invisible to us. By transforming numbers to graphic shapes, we allow readers to understand the stories those numbers hide.”

Alberto Cairo

The Functional Art

Data Science

- Influences many areas:
  - Business
  - Finance
  - Advertising
  - Retail and manufacturing
  - Government and policy
  - Search Engines & internet
  - Healthcare and Insurance
  - Journalism
  - Biology and genomics
  - Astronomy and High Energy Physics

Data Science

- Formulating the question.

  "Finding the question is often more important than finding the answer" - John Tukey

- Find the answer.

  Analyze/model the data to arrive at the answer. ML/statistical analysis, mathematical modelling.

- Understanding the answer.

  "The numbers have no way of speaking for themselves. We speak for them" - Nate Silver

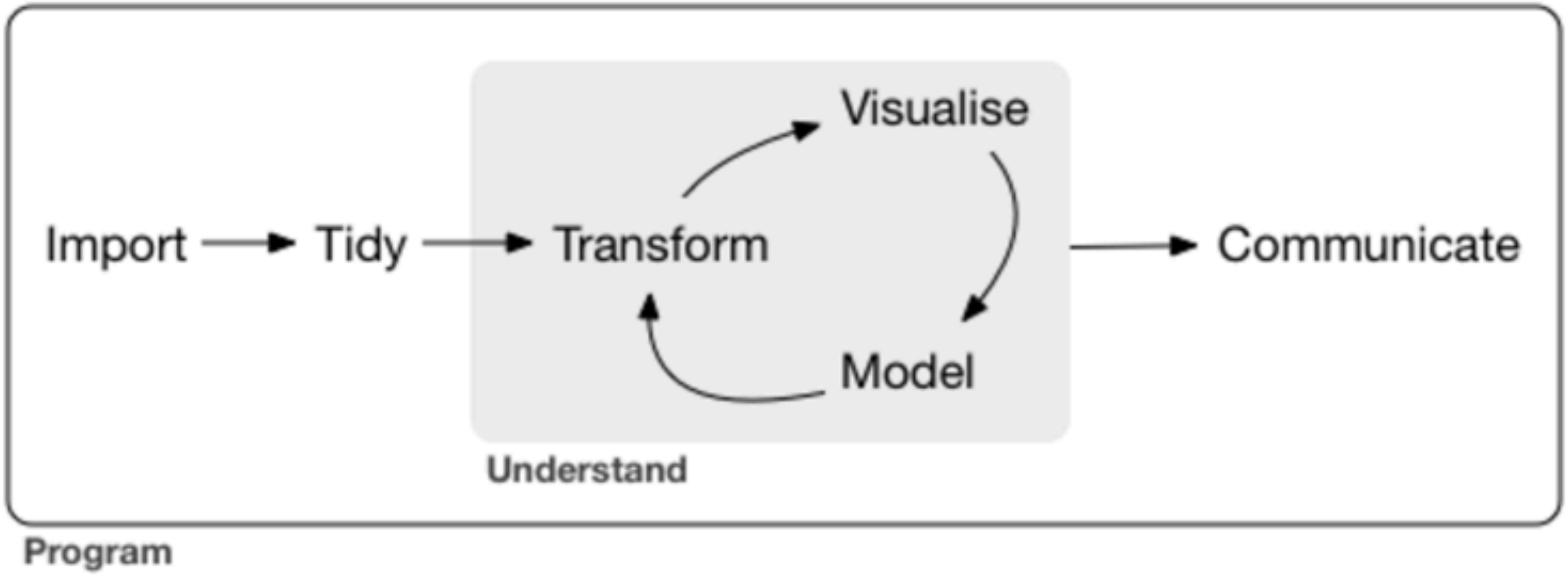
- Getting an hypothesis for the answer and derive insights from it.

  Data Science, at its core, is all about the reason for the outcome, and what can we do with this knowledge.

- Communicate the results.

  Dashboards/presentations

THE DATA ANALYSIS PROCESS



Source: Wickham and Grolemund, *R for Data Science*

DataViz in Data Science

- Visualizations are helpful (and often necessary) for efficient execution of each of the steps in a typical data science pipeline.
- More notably, visualization helps a great deal in
    - exploration and understanding of the data before we build a model
        - is this the right data?
        - is this the right model?
        - are these the right choices for parameters?
    - communication the results of the analysis/model.
        - what the model is doing?
        - why did it make a specific decisions?

Viz in Explorative Data Analysis

- ‘***Investigator***’ in crime series/movies. *Think Fox Mulder and Dana Scully in X-files*.

- Transformation, aggregation, summarization of data, often involves computation of statistical descriptors of the data e.g. mean, median, mode, variance, correlations etc.

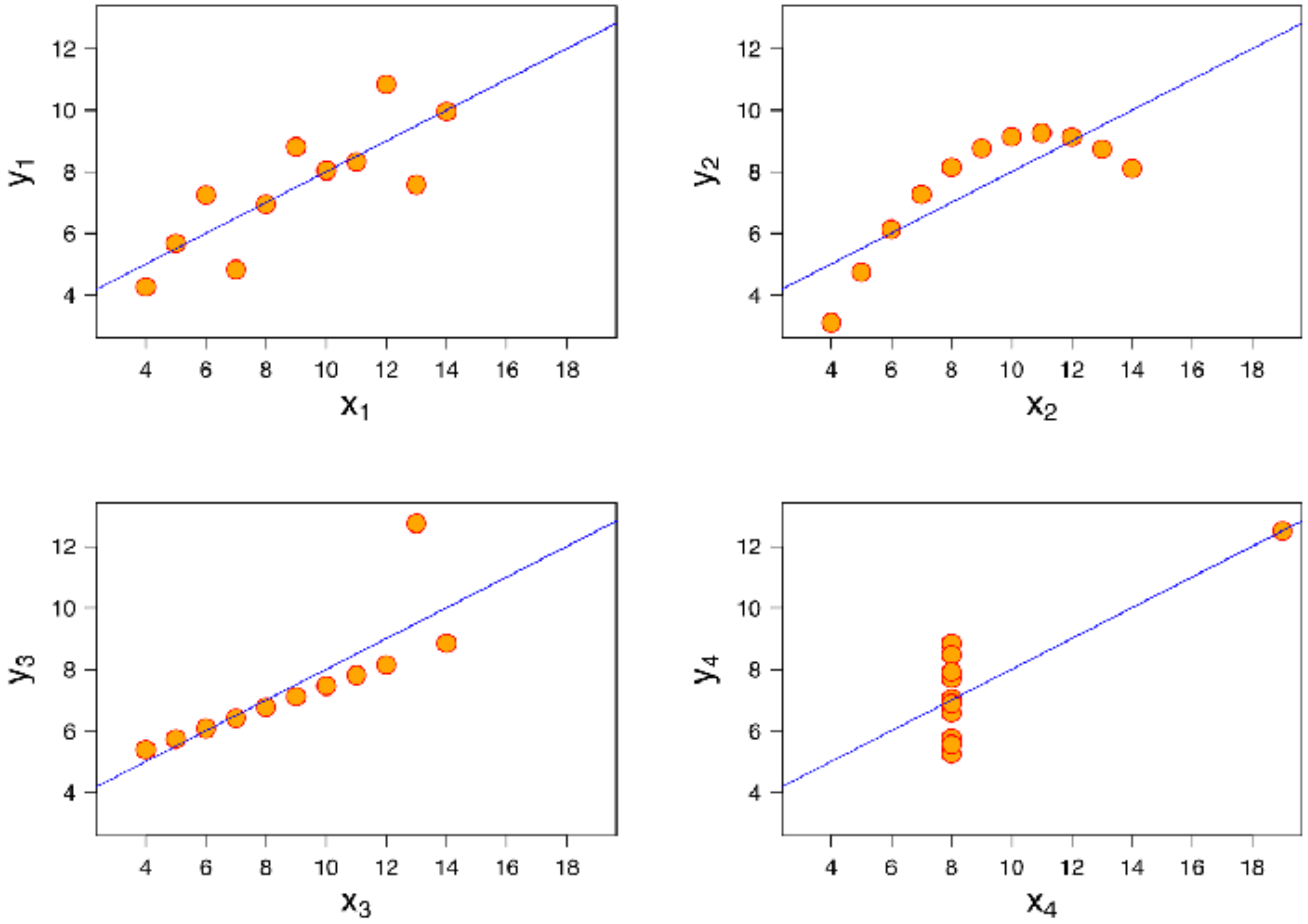
- Sometimes fitting functions e.g. line, splines etc.

- Clustering, low-dimensional projection, distribution of data.

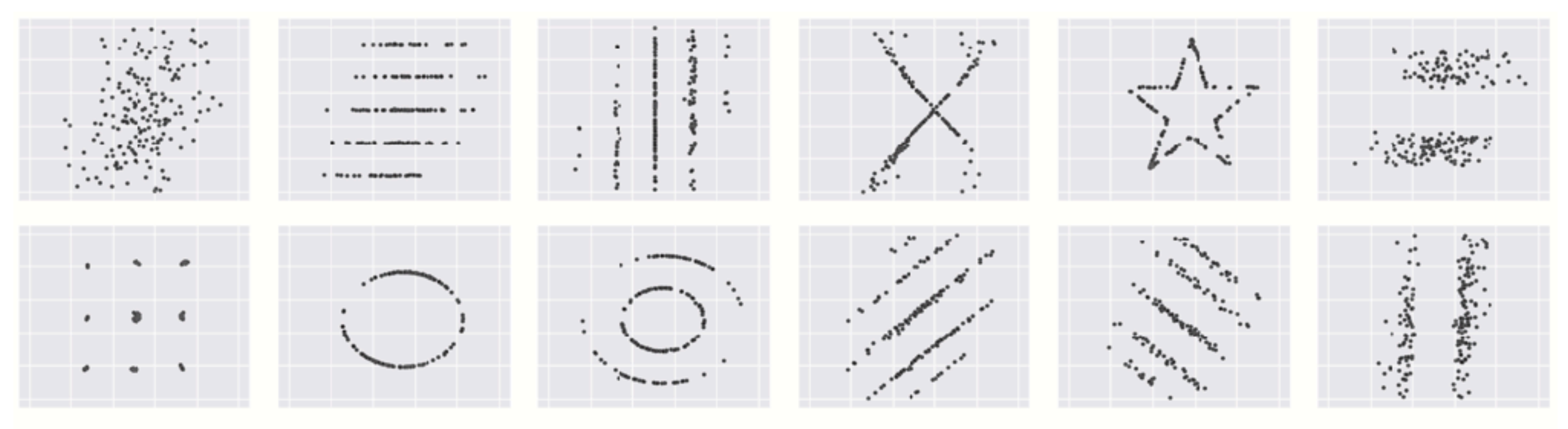
- Often, when exploring a dataset, you’ll want to use graphical representations of your data to help reveal insights/trends.

- Summary statistics can hide information.

ANSCOMBE'S QUARTET



Same stats, different graphs

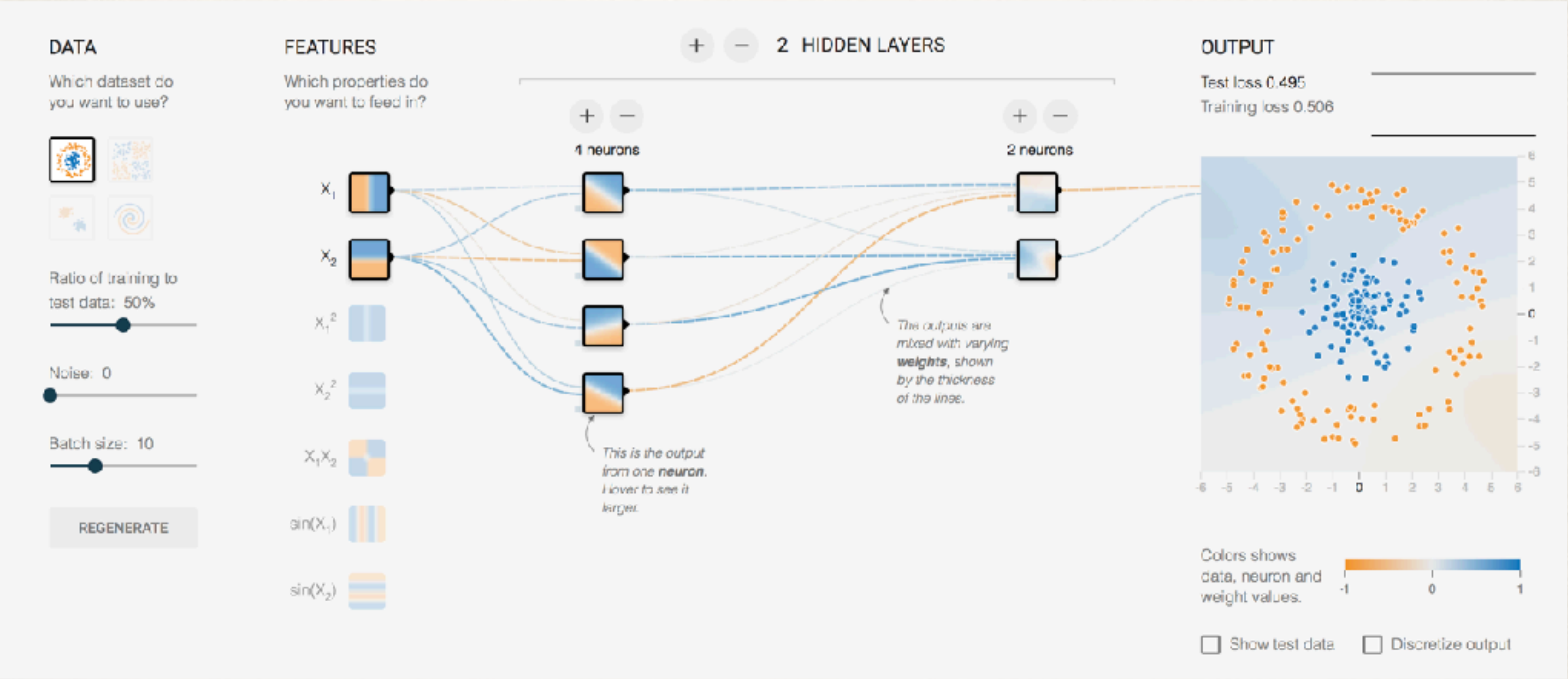


Same stats, different graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing,
Justin Matejka and George Fitzmaurice, CHI 2017

Dataviz in Machine Learning

- We live in exciting times to work at the intersection of ML and infovis. A lot of work has been done in this direction.
- The speaker is actively engaged in this topic.
- Intersection at two main points
  - Viz for improving ML model building process. Understanding of ML algorithms, params, features etc.
  - Viz for Interpretation of ML models.
  - Viz for better usability of ML-driven systems. (for less sophisticated)

Tensorflow Play



DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

REGENERATE

FEATURES

Which properties do you want to feed in?

2 HIDDEN LAYERS

4 neurons

2 neurons

This is the output from one ***neuron****. Hover to see it larger.*

The outputs are mixed with varying ***weights****, shown by the thickness of the lines.*

OUTPUT

Test loss 0.495

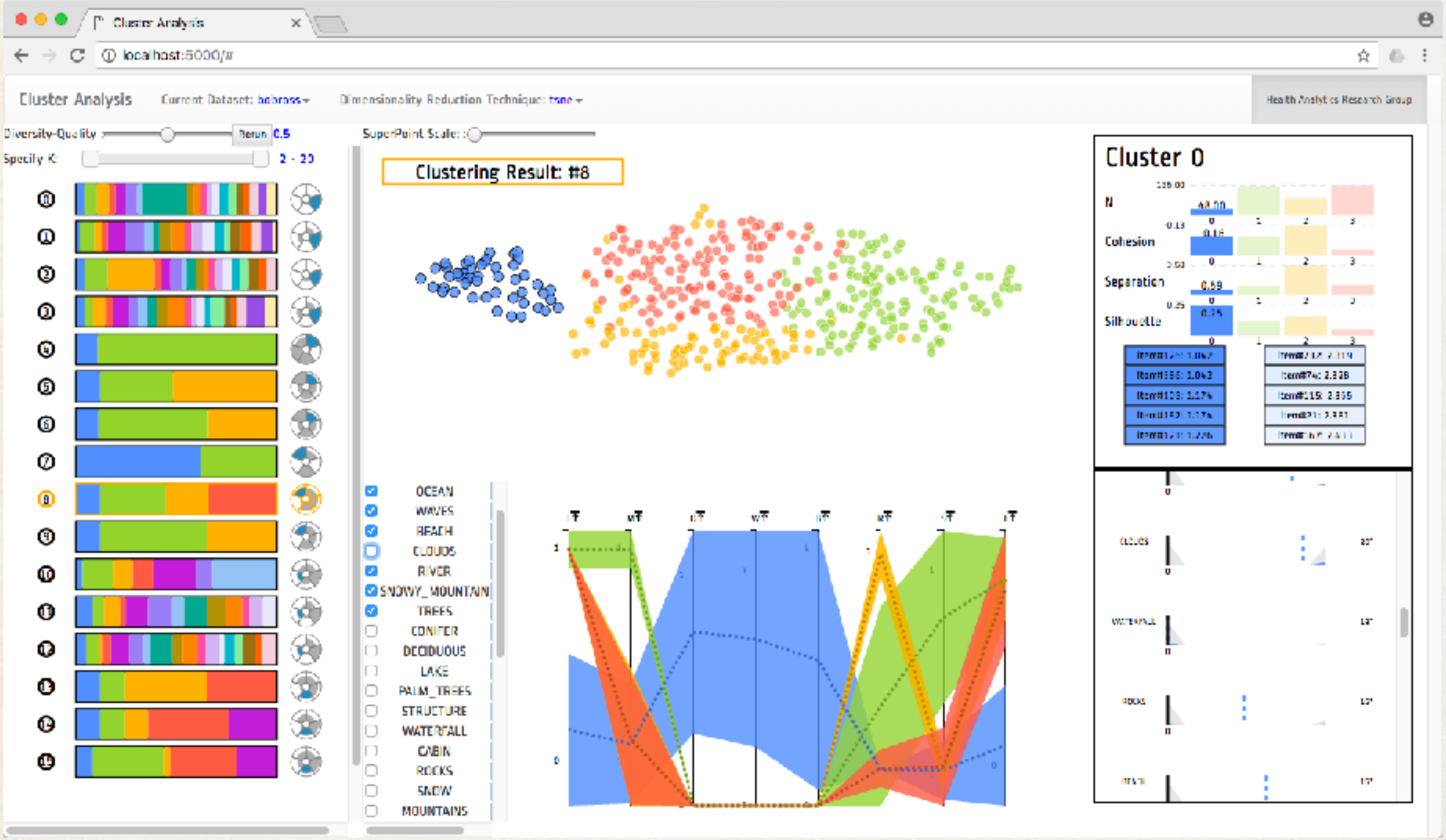
Training loss 0.506

Colors shows data, neuron and weight values.

□ Show test data

□ Discretize output

Clustervison



Cluster Analysis

Current Dataset: hdbross

Dimensionality Reduction Technique: tsne

Diversity-Quality: 0.5

Specify K: 2 - 20

SuperPoint Scale:

Clustering Result: #8

Prospector

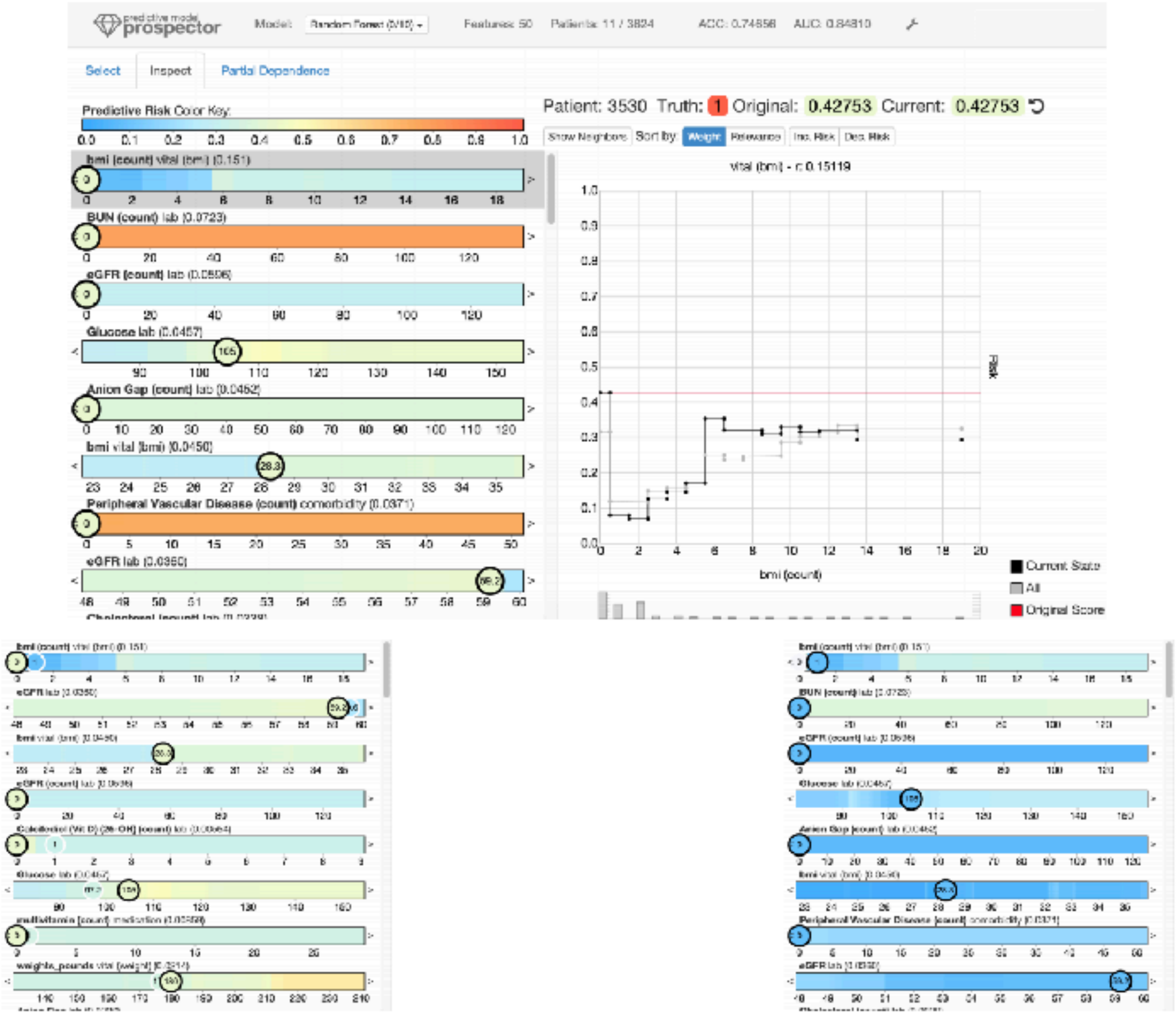

predictive model prospector
Model: Random Forest (V10)
Features: 50
Patients: 11 / 3824
ACC: 0.74656
AUC: 0.84810
Select
Inspect
Partial Dependence
Predictive Risk Color Key:
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
bmi (count) vital (bmi) (0.151)
BUN (count) lab (0.0723)
eGFR (count) lab (0.0606)
Glucose lab (0.0457)
Anion Gap (count) lab (0.0462)
bmi vital (bmi) (0.0450)
Peripheral Vascular Disease (count) comorbidity (0.0371)
eGFR lab (0.0360)
Cholesterol (count) lab (0.0239)
Patient: 3530 Truth: 1 Original: 0.42753 Current: 0.42753
Show Neighbors
Sort by:
Weight
Relevance
Inc. Risk
Dec. Risk
vital (bmi) - c 0.15119
Risk
bmi (count)
Current State
All
Original Score
Colecalciferol (Vit D) (25-OH) (count) lab (0.0064)
multivitamin (count) medication (0.0058)
weight_pounds vital (weight) (0.0214)

Figure 9. The user interface of *Prospector* is shown at the top. The bottom left shows suggestions on what changes (white circles) would decrease the predicted risk the most. The bottom right shows how the color plots change due to changing a value (namely changing the bmi value from 0 to 1). Fully white circles show the original value of the given patient.

Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow

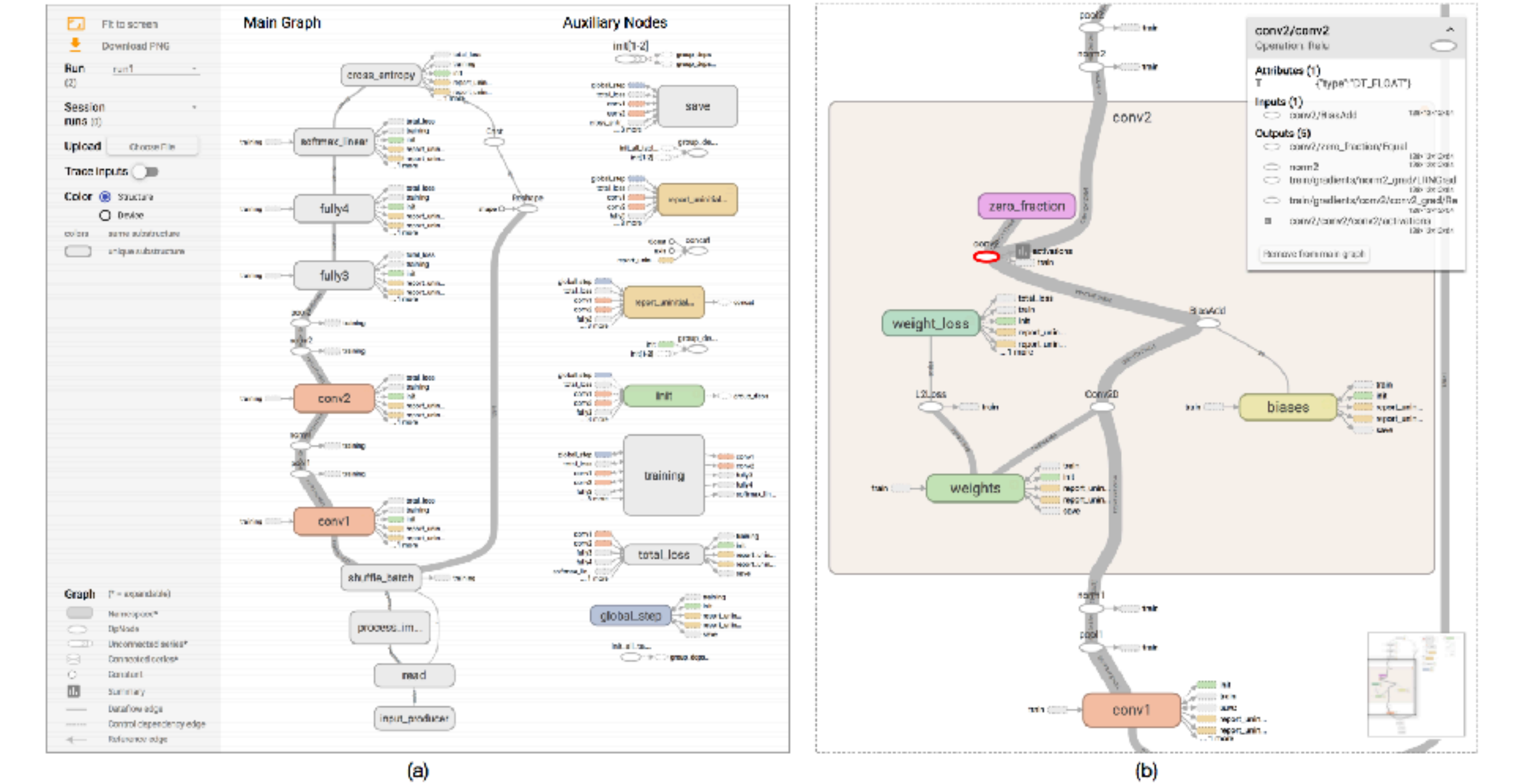


Fig. 1. The TensorFlow Graph Visualizer shows a convolutional network for classifying images (`tf_cifar`). (a) An overview displays a dataflow between groups of operations, with *auxiliary nodes* extracted to the side. (b) Expanding a group shows its nested structure.

Dataviz in Politics

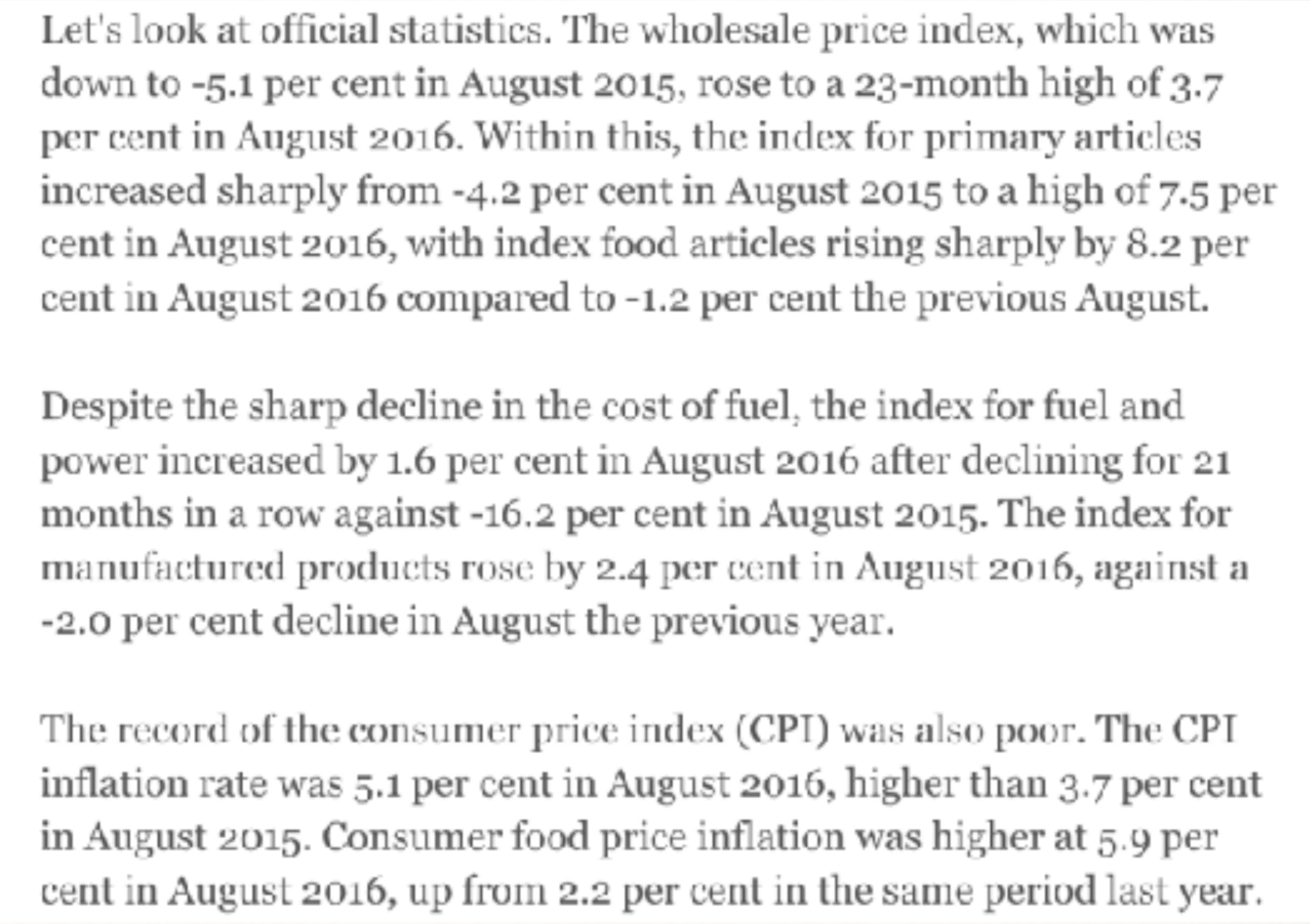
- One of the major applications of dataviz in current times. Data journalism and data storytelling are buzzing keywords. News outlets like NYT, WP, WSJ, NPR, The Guardian, FT, and even TOI, the Hindu etc. are working to channel full potential of viz in reporting and analysis.

- In addition to showing information visually, political studies/news benefits from viz in two major ways:

    - Viz for better communication of prediction results.

    - Viz for better predictive model building.

- In both the cases, the key contribution is to quickly iterate over"what if" scenarios based on changing data and model parameters.

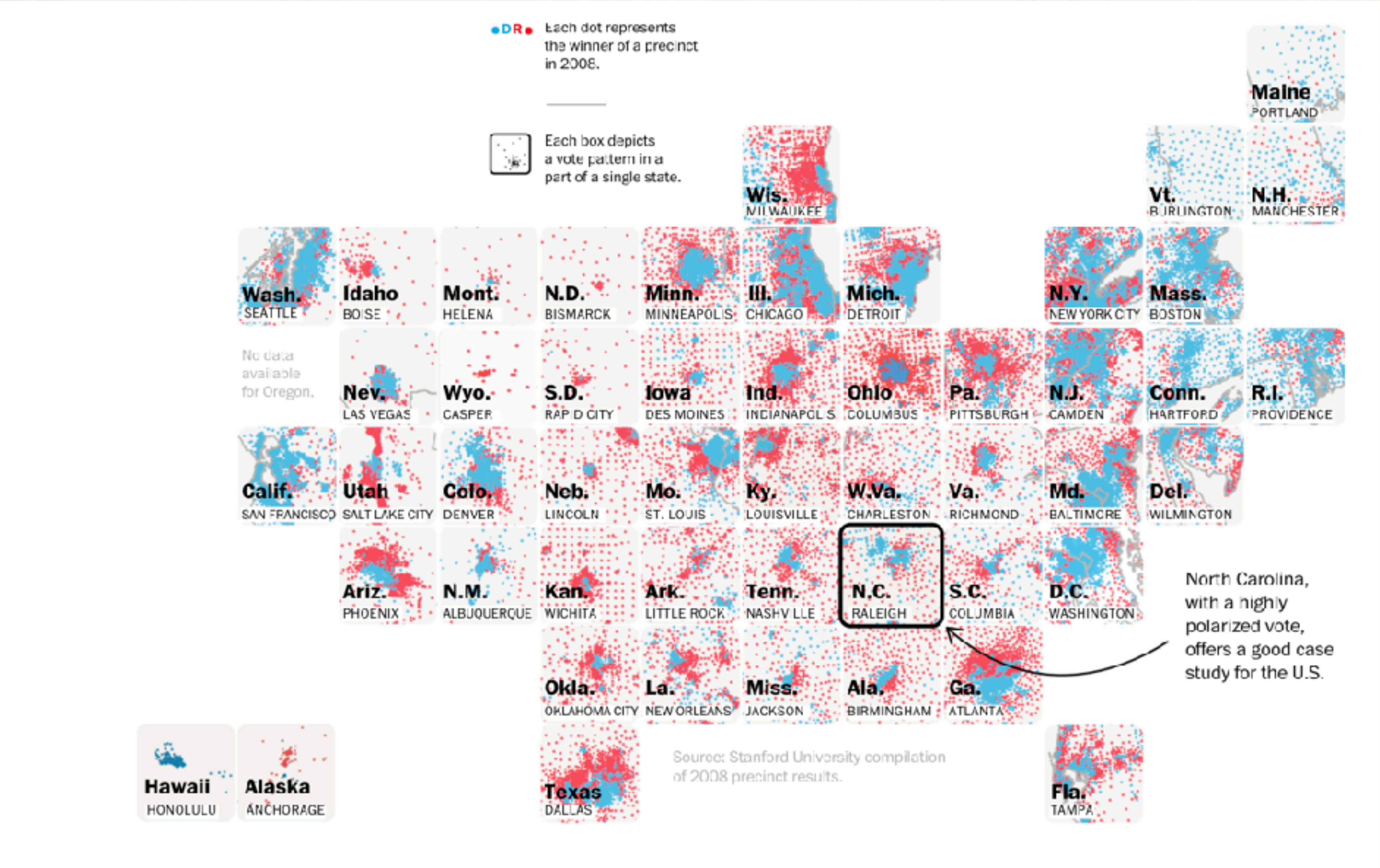


Let's look at official statistics. The wholesale price index, which was down to -5.1 per cent in August 2015, rose to a 23-month high of 3.7 per cent in August 2016. Within this, the index for primary articles increased sharply from -4.2 per cent in August 2015 to a high of 7.5 per cent in August 2016, with index food articles rising sharply by 8.2 per cent in August 2016 compared to -1.2 per cent the previous August.

Despite the sharp decline in the cost of fuel, the index for fuel and power increased by 1.6 per cent in August 2016 after declining for 21 months in a row against -16.2 per cent in August 2015. The index for manufactured products rose by 2.4 per cent in August 2016, against a -2.0 per cent decline in August the previous year.

The record of the consumer price index (CPI) was also poor. The CPI inflation rate was 5.1 per cent in August 2016, higher than 3.7 per cent in August 2015. Consumer food price inflation was higher at 5.9 per cent in August 2016, up from 2.2 per cent in the same period last year.

Political Polalization in America



Each dot represents the winner of a precinct in 2008.

Each box depicts a vote pattern in a part of a single state.

No data available for Oregon.

North Carolina, with a highly polarized vote, offers a good case study for the U.S.

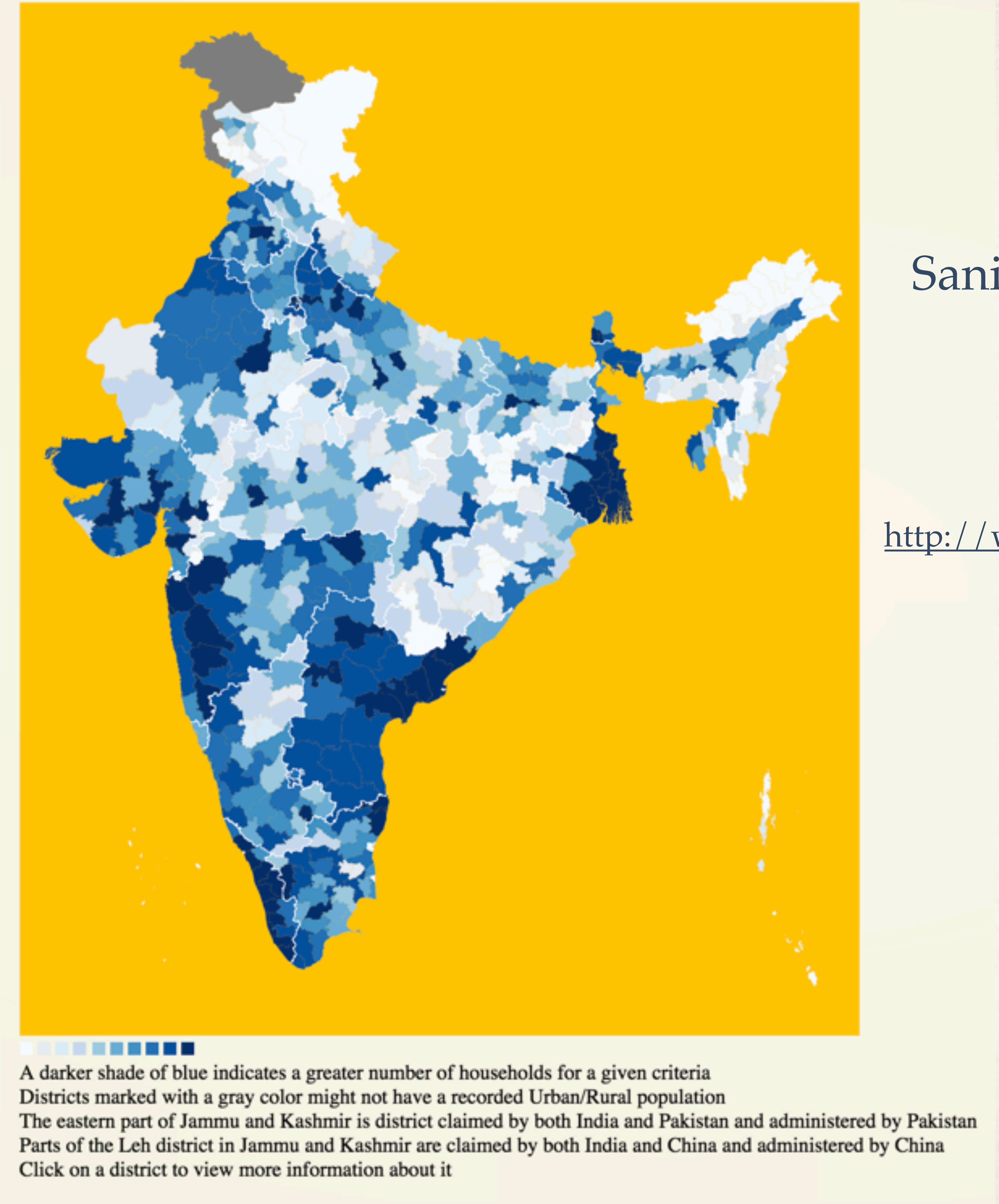
Source: Stanford University compilation of 2008 precinct results.

- *https://projects.fivethirtyeight.com/2016-election-forecast/*

Dataviz in Policy

- Data for social good is such an important theme today, and aptly so with so much of open data.
- Citizens needs to be educated about data to understand how law-makers are doing things. e.g. Refugee crisis. Check on governments/municipalities.
- Policy makers can (or can be persuaded to) make well informed decisions based on data. Best way to communicate the data is through a visualization. Accessible, and saves time!
- Data offices in White House, NY Mayors office.

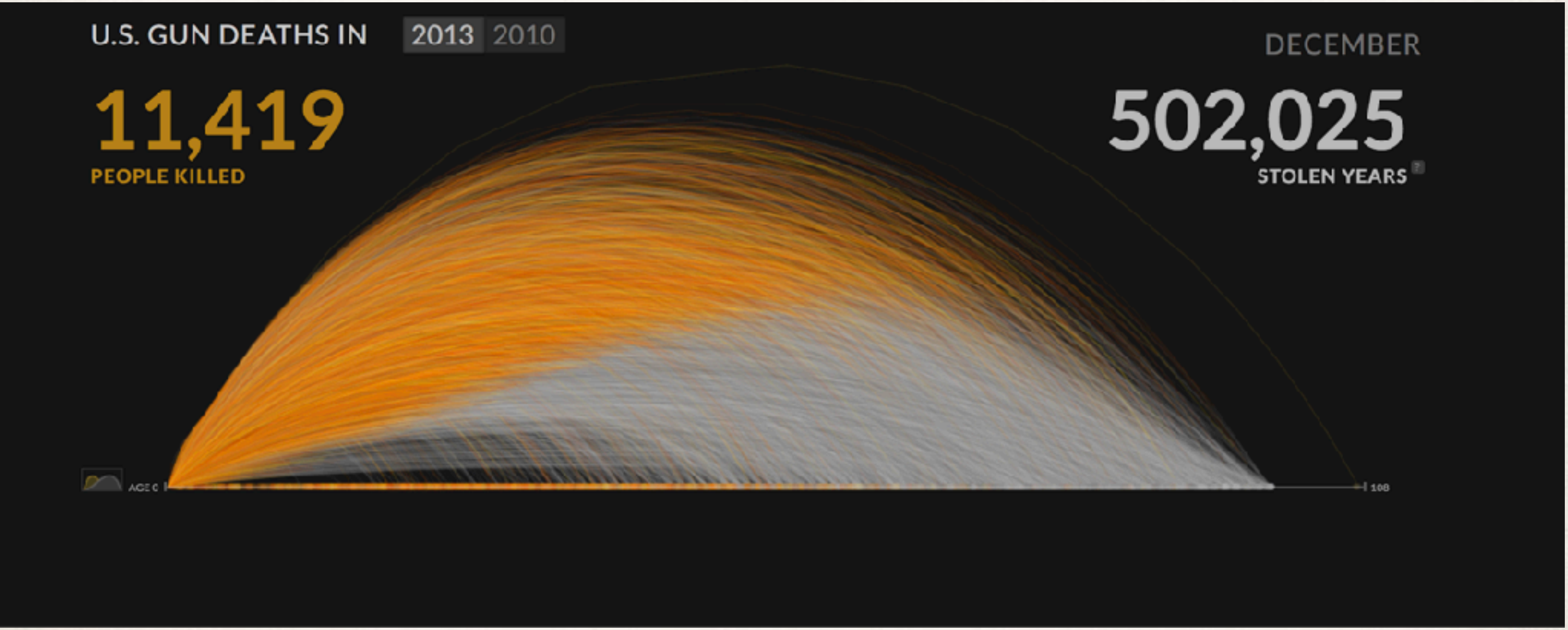
Sanitation conditions in India



A darker shade of blue indicates a greater number of households for a given criteria
Districts marked with a gray color might not have a recorded Urban/Rural population
The eastern part of Jammu and Kashmir is district claimed by both India and Pakistan and administered by Pakistan
Parts of the Leh district in Jammu and Kashmir are claimed by both India and China and administered by China
Click on a district to view more information about it

http://www.sigri.com/portrait_sanitation.html

Gun Violence in the US



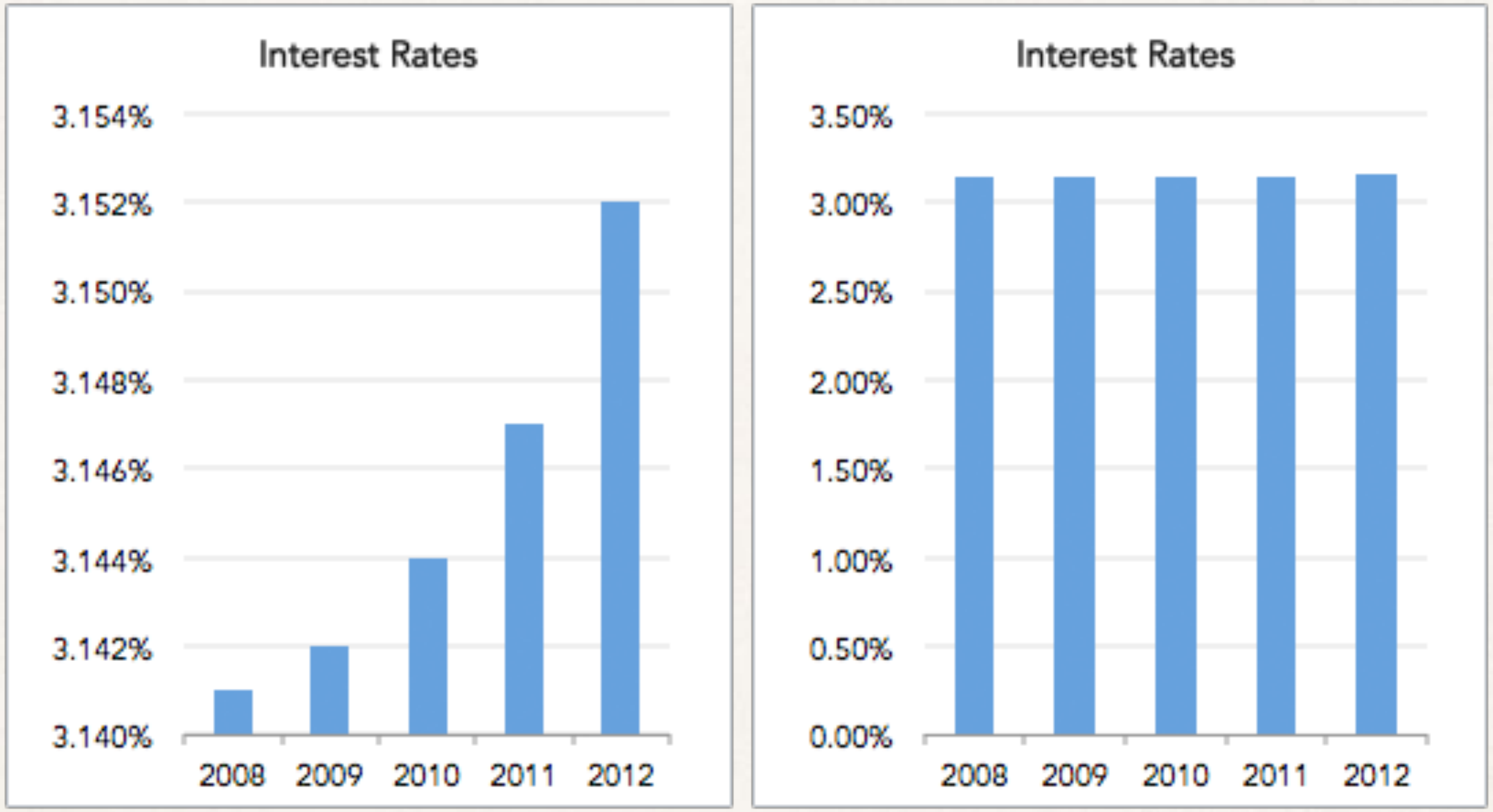
https://guns.periscopic.com/?year=2013

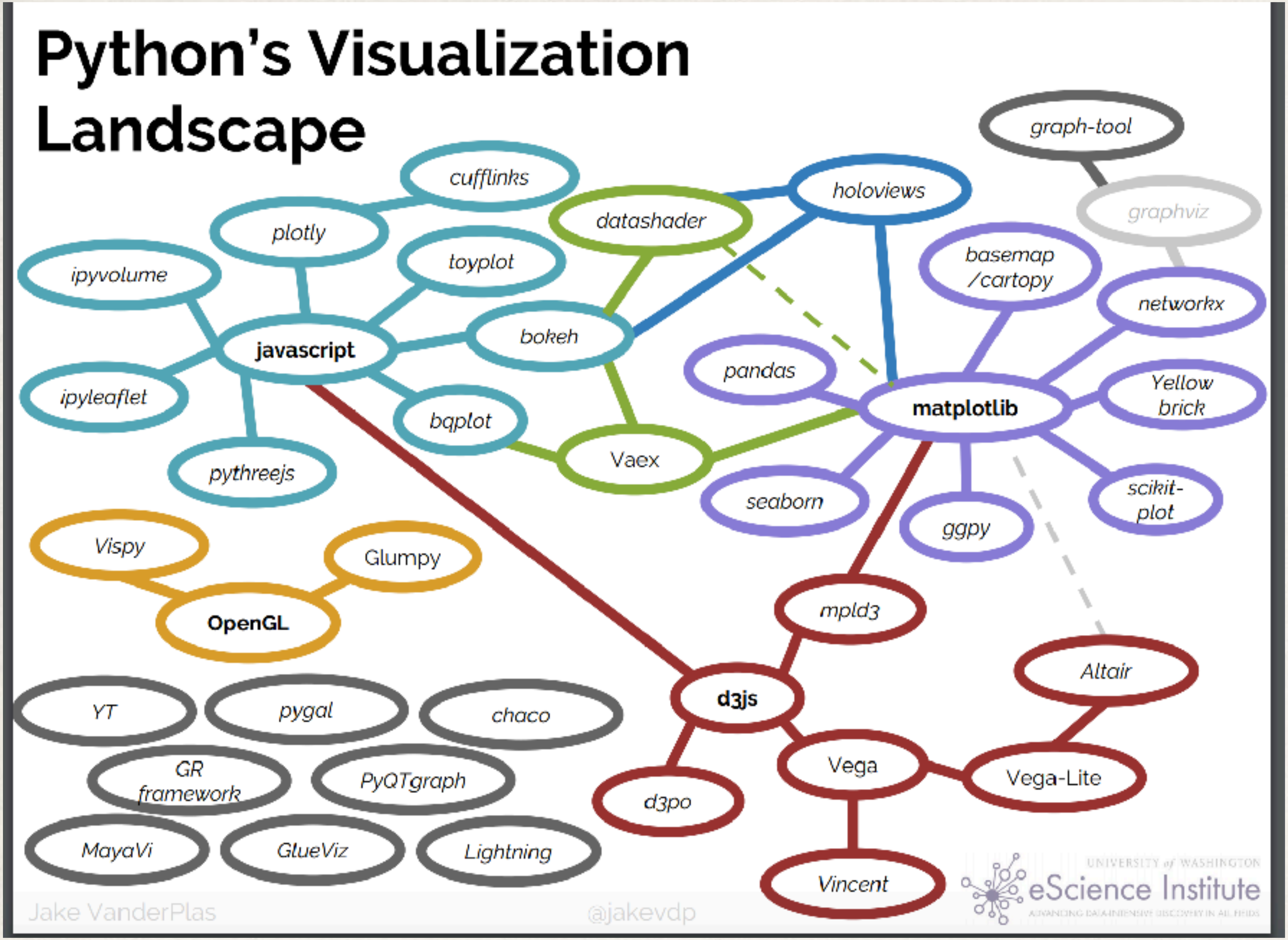
“data visualization without rigorous analysis is at best just rhetoric and, at worst, incredibly harmful.”

–Jake Porway, Founder DataKind

How not to lie with statistics

Same Data, Different Y-Axis





Python's Visualization Landscape

graph-tool
graphviz
cufflinks
holoviews
datashader
plotly
basemap /cartopy
toyplot
ipyvolume
networkx
javascript
bokeh
pandas
Yellow brick
matplotlib
ipyleaflet
bqplot
Vaex
pythreejs
seaborn
scikit-plot
ggpy
Vispy
Glumpy
OpenGL
mpld3
Altair
d3js
YT
pygal
chaco
GR framework
PyQTgraph
Vega
Vega-Lite
d3po
MayaVi
GlueViz
Lightning
Vincent

Jake VanderPlas

@jakevdp

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Jake VanderPlas, PyCon 2017

matplotlib

- Most common python visualization library.

- Used extensively in scientific research for its ability to programmatically generate a wide range of plots/charts. Very powerful and versatile.

- Created to bring the charting power of MATLAB to python.

- Lots of resources/tutorials/docs available, well tested and experimented with.

- Many rendering backends e.g. *pdf, png, jpeg, eps* (scientific publications), *svg* (web) etc.

- *matplotlib* tries to make easy things easy and hard things possible.

- Completely open source.

matplotlib

- Can be very verbose, lot of coding.
- Default styles are a bit outdated, not as sharp and pretty as modern libraries like *ggplot2* in R.
- No interaction
- Can be slow for large and complex data.

> *“**Design Goal**: Create new, high-level libraries, which are more expressive, less verbose, while retaining all the good parts of matplotlib.”*

– Jake VanderPlas, PyCon 2017

seaborn

- Statistical data visualization
- High-level language built on top of *matplotlib*.
- Great set of colour palette and stylistic features.
- Support for visualizing exploratory statistics : univariate and multivariate distributions, statistical analysis, comparisons, subsetting etc.
- Visualization of exploratory analysis: linear regression, clustering etc.
- Compatible with pandas dataframes.
- Visualization as the central part of exploring and understanding data, quick functions to build visuals. A very few lines of code is required to build complex visuals.

“The plotting functions operate on dataframes and arrays containing a whole dataset and internally perform the necessary aggregation and statistical model-fitting to produce informative plots.

If matplotlib “tries to make easy things easy and hard things possible”, seaborn tries to make a well-defined set of hard things easy too.”

– seaborn documentation

https://seaborn.pydata.org/introduction.html

DataViz for the web: Interactive visualization

- Data science and analytical workflows are becoming increasingly connected to the web. Today, most of the data science platforms are completely web-based.

- User is an active participant rather than a passive observer.

- Many libraries in python:
  - *ipywidgets:* a package for interactive mini-apps in the Jupyter Notebook.
  - *plotly:* produces d3.js charts using Python, and can convert Python charts as well.
  - *cufflinks:* binds pandas to plotly.

plotly

- Web-based platform for interactive visualization in python.

- Includes both offline components (Jupyter Notebook) and an online cloud/web GUI.

- *plotly* can also convert *matplotlib* charts into plotly interactive charts.