Alexander Lex
@alexander_lex
http://alexander-lex.net

THE UNIVERSITY OF UTAH

# Literate Visualization: Making Visual Analysis Sessions Reproducible and Reusable

visualization design lab

SCI
www.sci.utah.edu

# Outline

- What is Data Science
  - Programming is essential
- What's the role of visualization in DS?
- Reproducibility, Communication, Reusability
  - Literate Programming is a way to achieve this
  - What's literate programming for visualization?
    - Provenance Tracking; extracting analysis stories.
    - Inferring Analyst's Intent

# DATA SCIENCE

# What is Data Science?

**A data scientist is a statistician who lives in San Fransisco.**

**Data science is statistics on a Mac.**

**A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.**

# What is Data Science?

**Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.** [Wikipedia]


Hadley Wickham

# Why is Programming So Important?

Expressivity and Generality!

It's hard (if not impossible) to build a higher-level system that anticipates any possible need.

But programming

is difficult,

not everyone can do it,

and it's slow.

Ideal World: Programming where necessary; Interactive tools where possible.

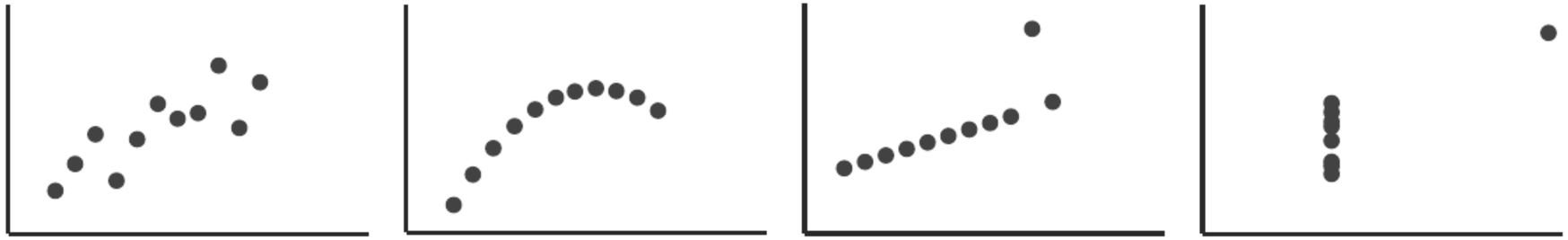
# THE ROLE OF VISUALIZATION IN DATA SCIENCE

# Can We Trust Statistics?

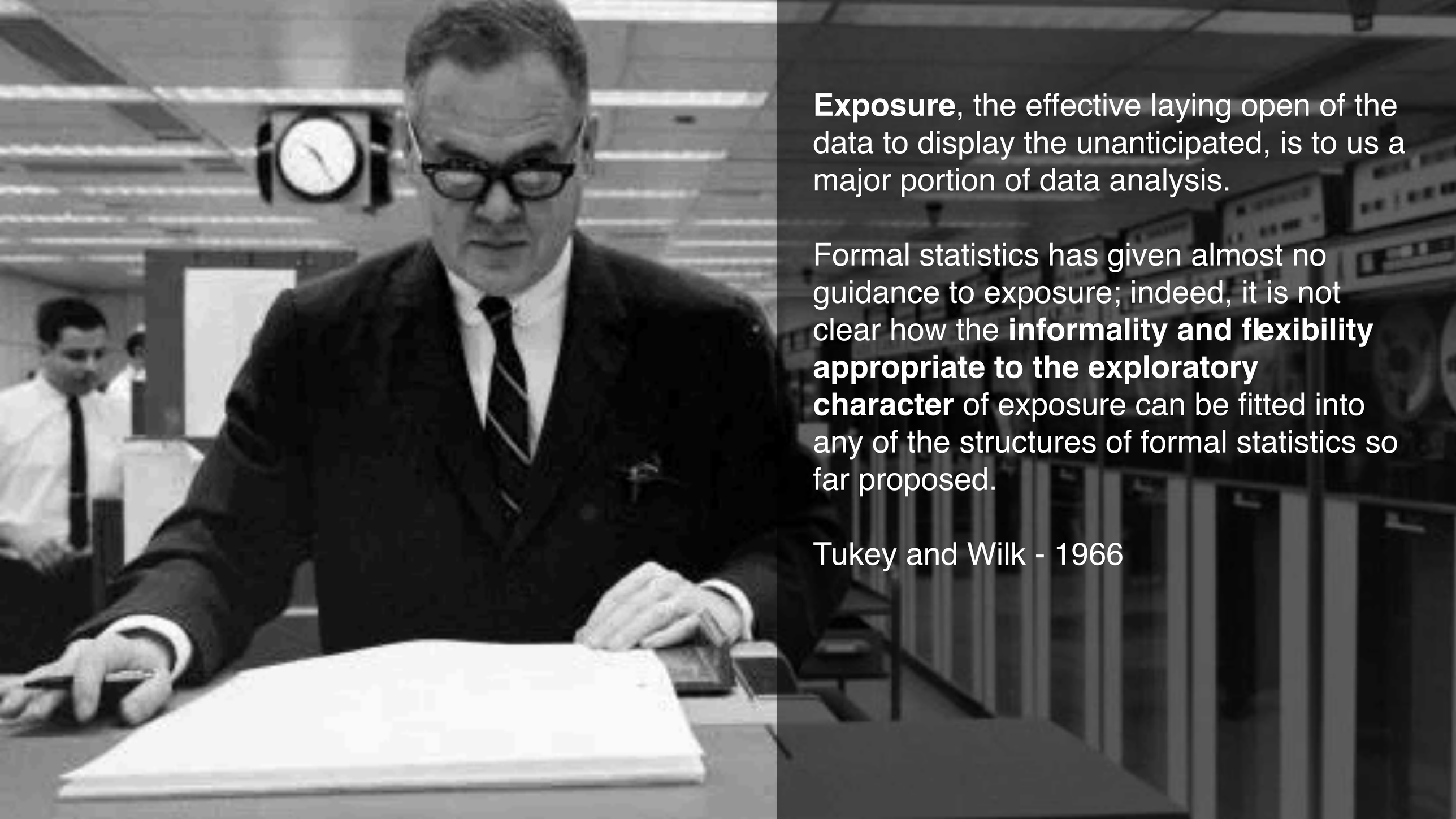
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.6 | | | | | 8 | 6.89 |

**Mean x: 9 y: 7.50**
**Variance x: 11 y: 4.122**
**Correlation x – y: 0.816**
**Linear regression: y = 3.00 + 0.500x**

# Anscombe's Quartett



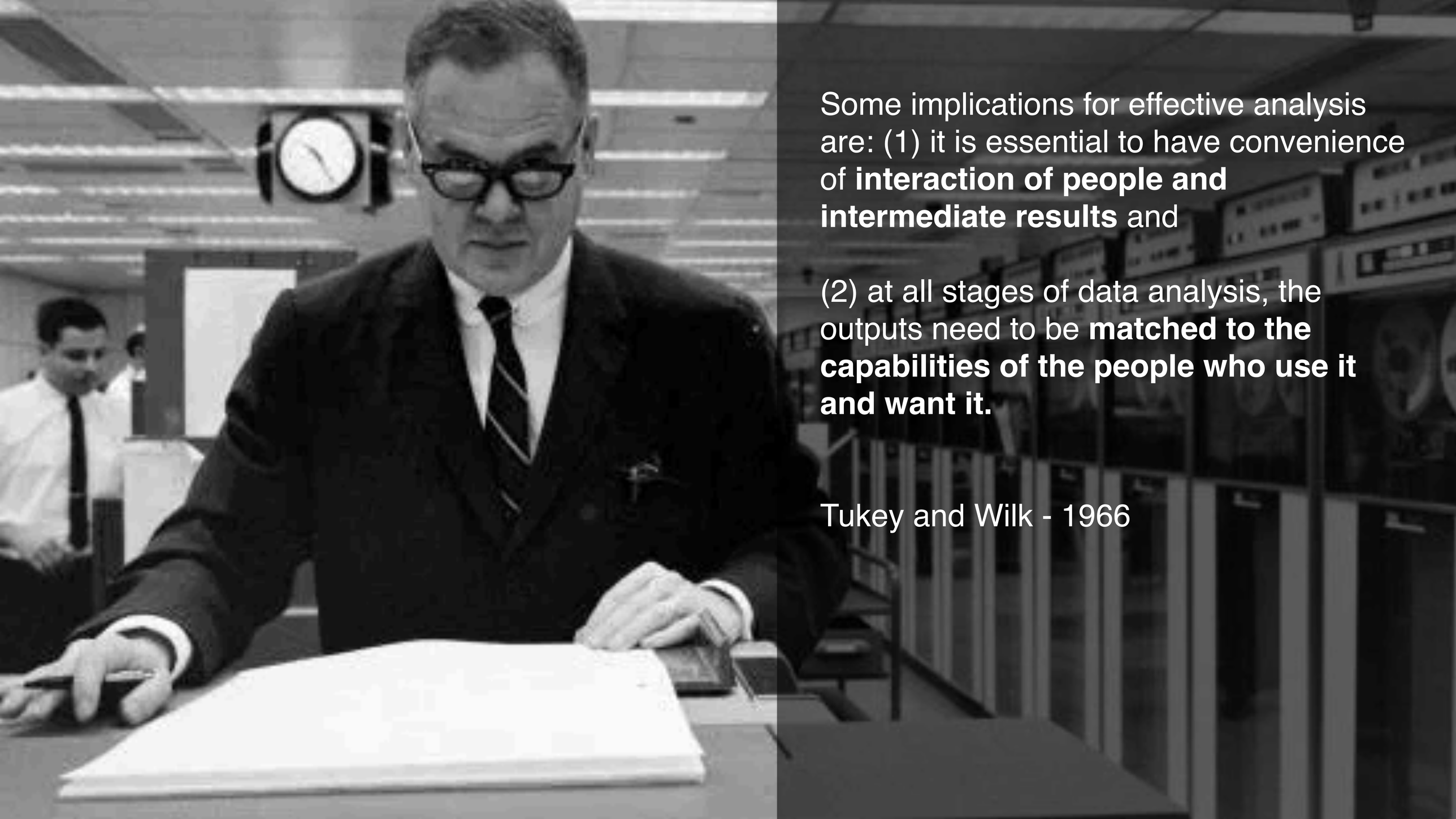
**Mean x: 9 y: 7.50**
**Variance x: 11 y: 4.122**
**Correlation x – y: 0.816**
**Linear regression: y = 3.00 + 0.500x**



**Exposure**, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis.

Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality and flexibility appropriate to the exploratory character** of exposure can be fitted into any of the structures of formal statistics so far proposed.

Tukey and Wilk - 1966

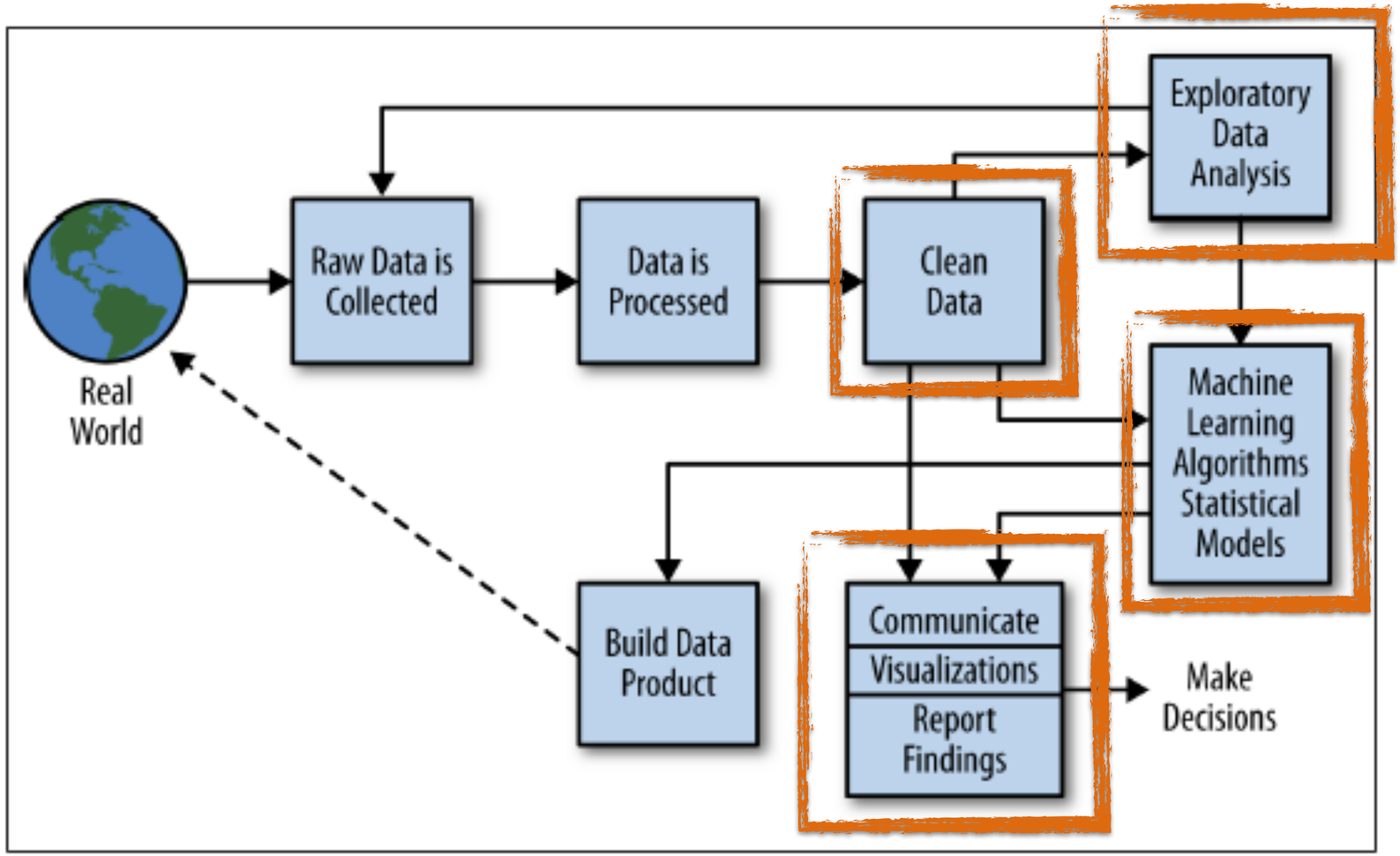


Some implications for effective analysis are: (1) it is essential to have convenience of **interaction of people and intermediate results** and

(2) at all stages of data analysis, the outputs need to be **matched to the capabilities of the people who use it and want it.**

Tukey and Wilk - 1966

# What's the Role of Visualization in Data Science?




Doing Data Science: Straight Talk from the Frontline,

by Cathy O'Neil and Rachel Schutt

# Vis as commonly used by Data Scientists

**Embedded, and code-generated static graphics**

**Great, but**

**(commonly) static**

**you have to write code…**

## Embedding ggplot into knitr document fails

We will generate a graphic in the following chunk.

```
library(ggplot2)

mtcars_ggplot <-
  ggplot(mtcars) +
  aes(x = wt, y = mpg) +
  geom_point() +
  stat_smooth()
```

In-line code is great for things like summary statistics. The mean miles per gallon for the `mtcars` data set is 20.090625.

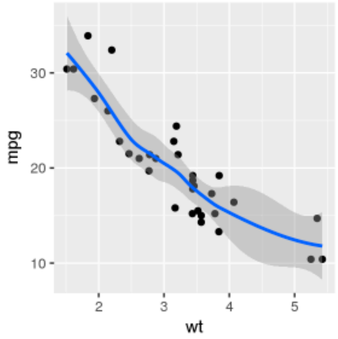
In-line code is not good for graphics.

`r mtcars_ggplot`

Instead, use a chunk. This will also allow you to control how the graphic is rendered in your final document.

```
mtcars_ggplot
```

```
## `geom_smooth()` using method = 'loess'
```



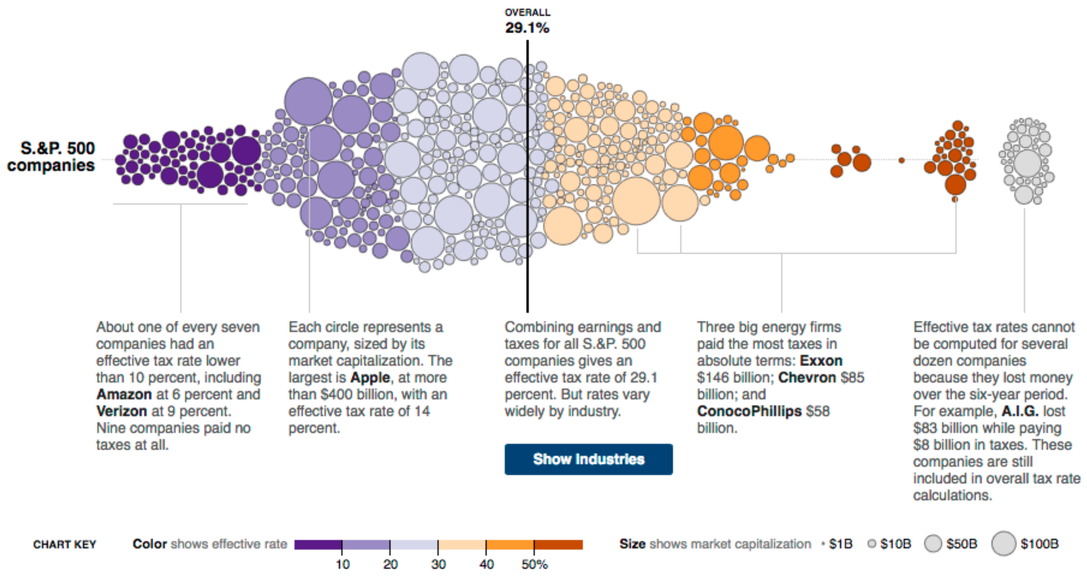
# Why do we need Interaction?

Explore data that is big / complex

**There is too much data**

**There are too many ways to show it**

**There are too many questions to ask**





http://www.nytimes.com/interactive/2013/05/25/sunday-review/corporate-taxes.html

# Visualization =

## Human Data Interaction

**REPRODUCIBILITY,**

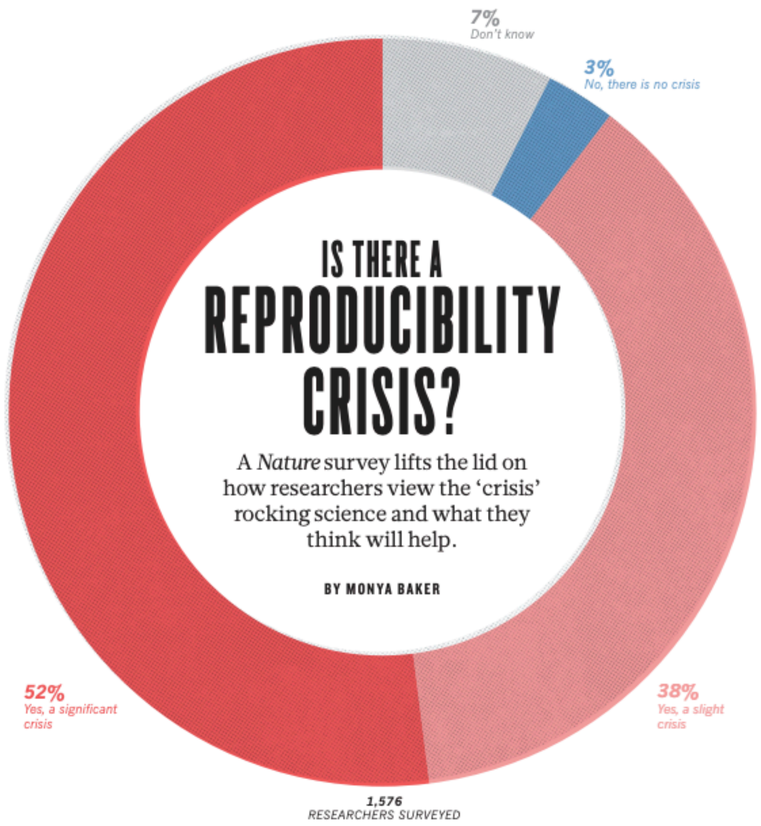
**COMMUNICATION,**

**REUSABILITY**

# Reproducibility Crisis in Science

**Psychology: 40% of experiments could be reproduced**

**Cancer Biology: 11%**




More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

Failing to reproduce results is a rite of passage, says Marcus Munafo, a biological psychologist at the University of Bristol, UK, who has a long-standing interest in scientific reproducibility. When he was a student, he says, "I tried to replicate what looked simple from the literature, and wasn't able to. Then I had a crisis of confidence, and then I learned that my experience wasn't uncommon."

https://www.nature.com/news/polopoly_fs/1.19970!/menu/main/topColumns/topLeftColumn/pdf/533452a.pdf
https://science.sciencemag.org/content/349/6251/aac4716
https://www.nature.com/articles/483531a

# Why is there a reproducibility crisis?

Perverse incentives (publish or perish)

Bias for “flashy” results

Biases in data analysis

Lack of data sharing

Lack of sharing the analysis process

…

# Just Sharing is Not Enough

**Data needs context/annotation**

**Analysis needs documentation and justification**

## Literate Programming

The past ten years have witnessed substantial improvements in programming methodology. This advance, carried out under the banner of "structured programming," has led to programs that are more reliable and easier to comprehend; yet the results are not entirely satisfactory.

My purpose in the present paper is to propose another motto that may be appropriate for the next decade, as we attempt to make further progress in the state of the art. I believe that **the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature.**

Donald E. Knuth, 1984

Hence, my title: "**Literate Programming.**"

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do,
**let us concentrate rather on explaining to human beings what we want a computer to do**.

Donald E. Knuth, 1984

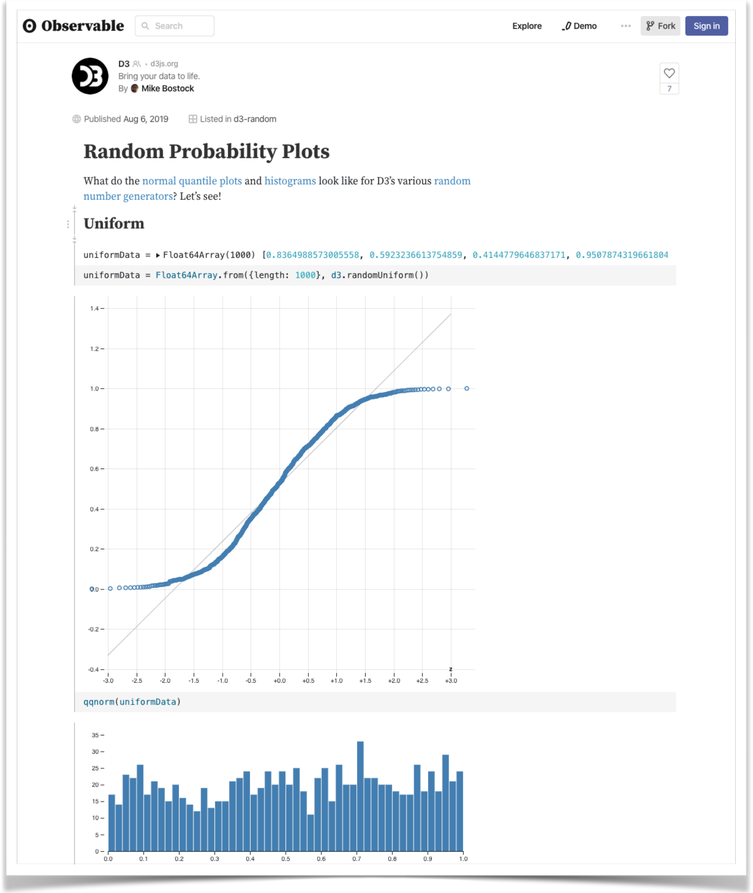
# Literate Programming

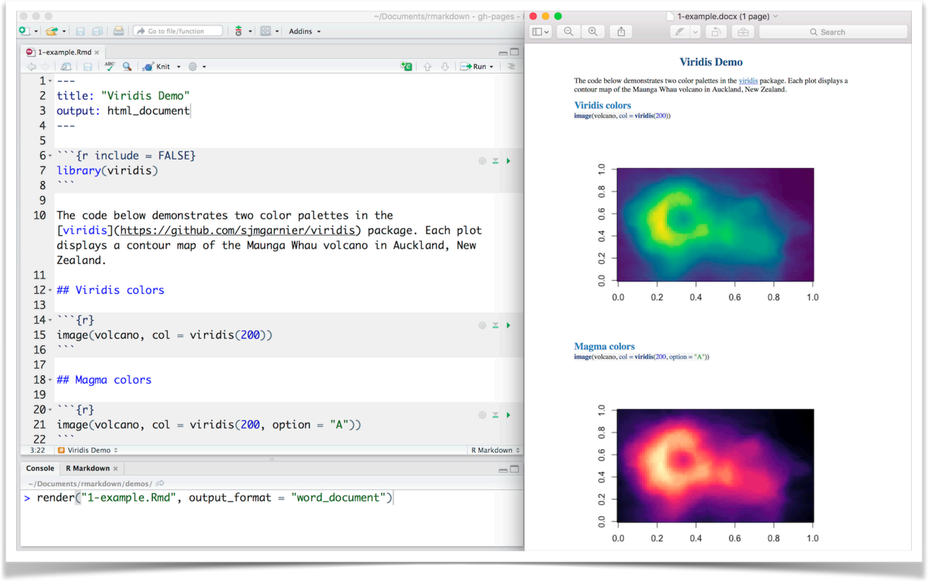
**Explain the why and how using any means necessary!**

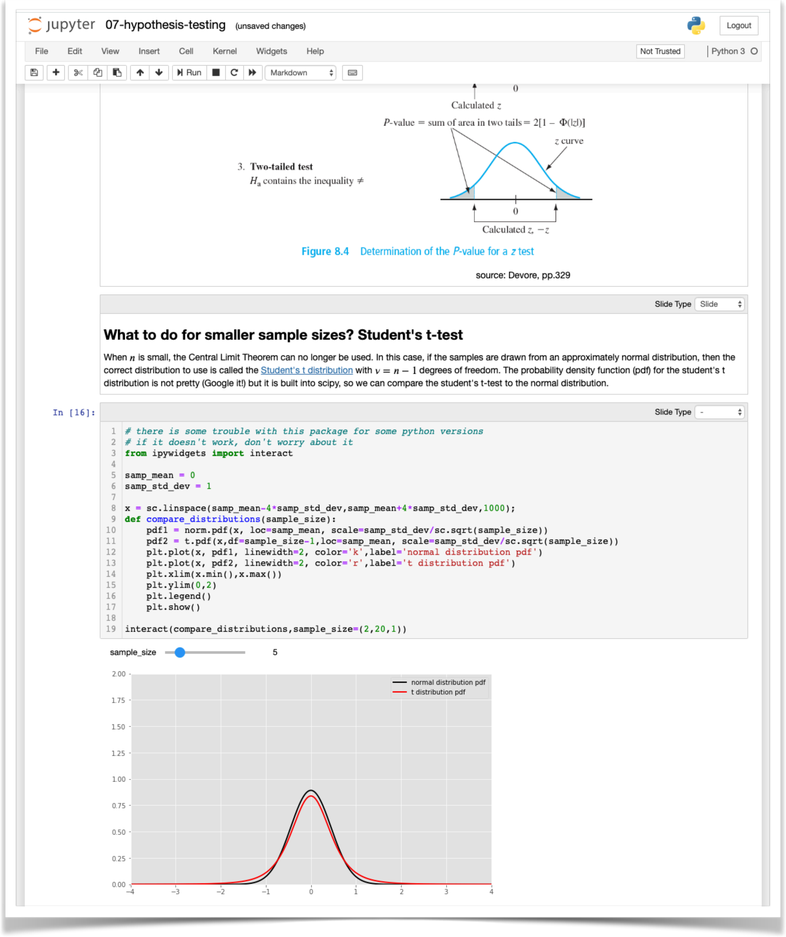
- **Text**
- **Images / Visualizations**
- **Formulas**
- **Videos**
- **Links**
- **Code**

**All work in concert to ensure reproducibility**

# Literate Programming in the Wild




**Observable** **R Markdown** **Jupyter Notebooks**

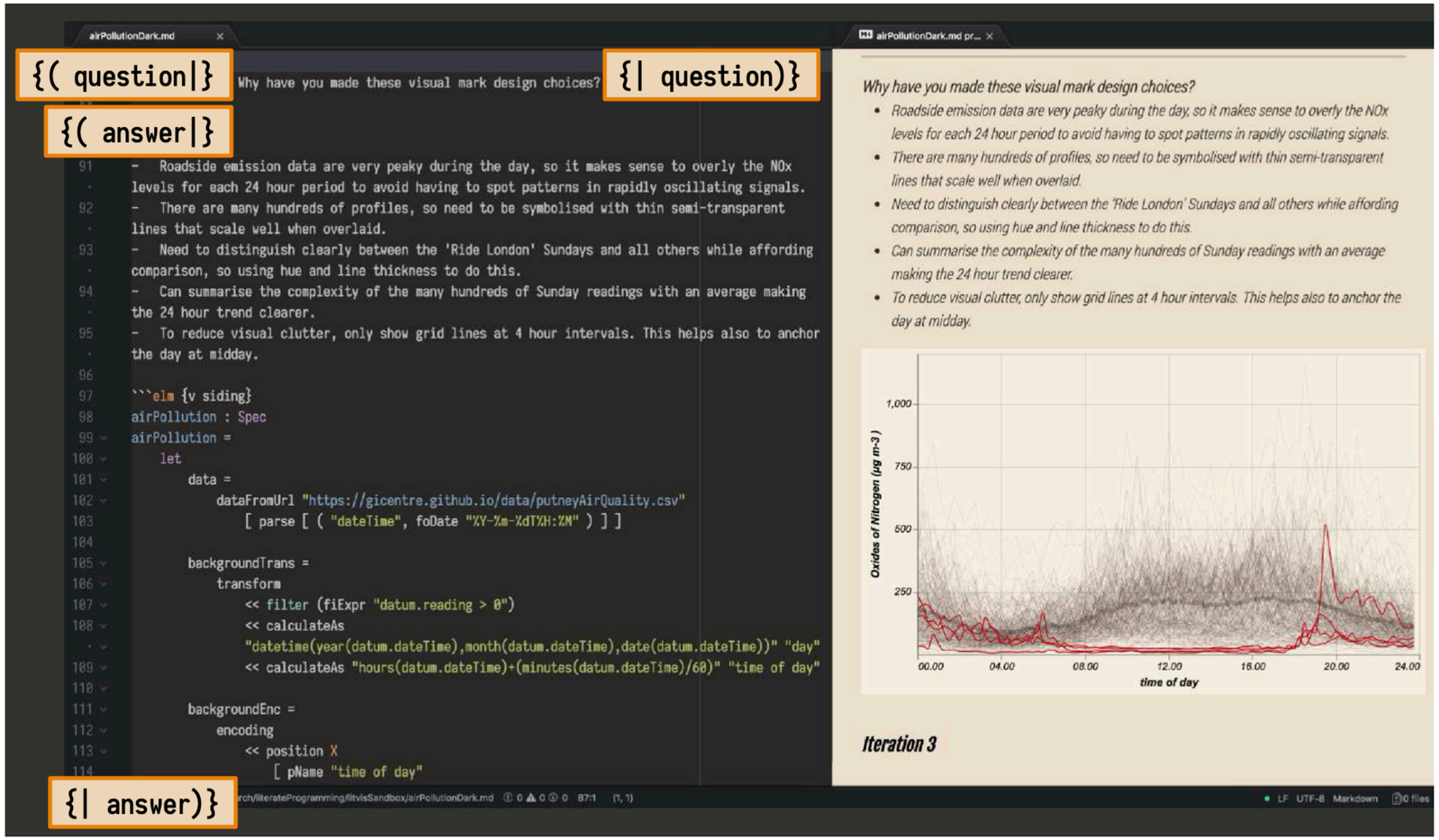
# Literate Visualization

**Idea: make the process of an interactive, visual analysis session well reasoned and documented.**

**Problem: no record of what was done (let alone why)**

# Aside: Literate Visualization Design

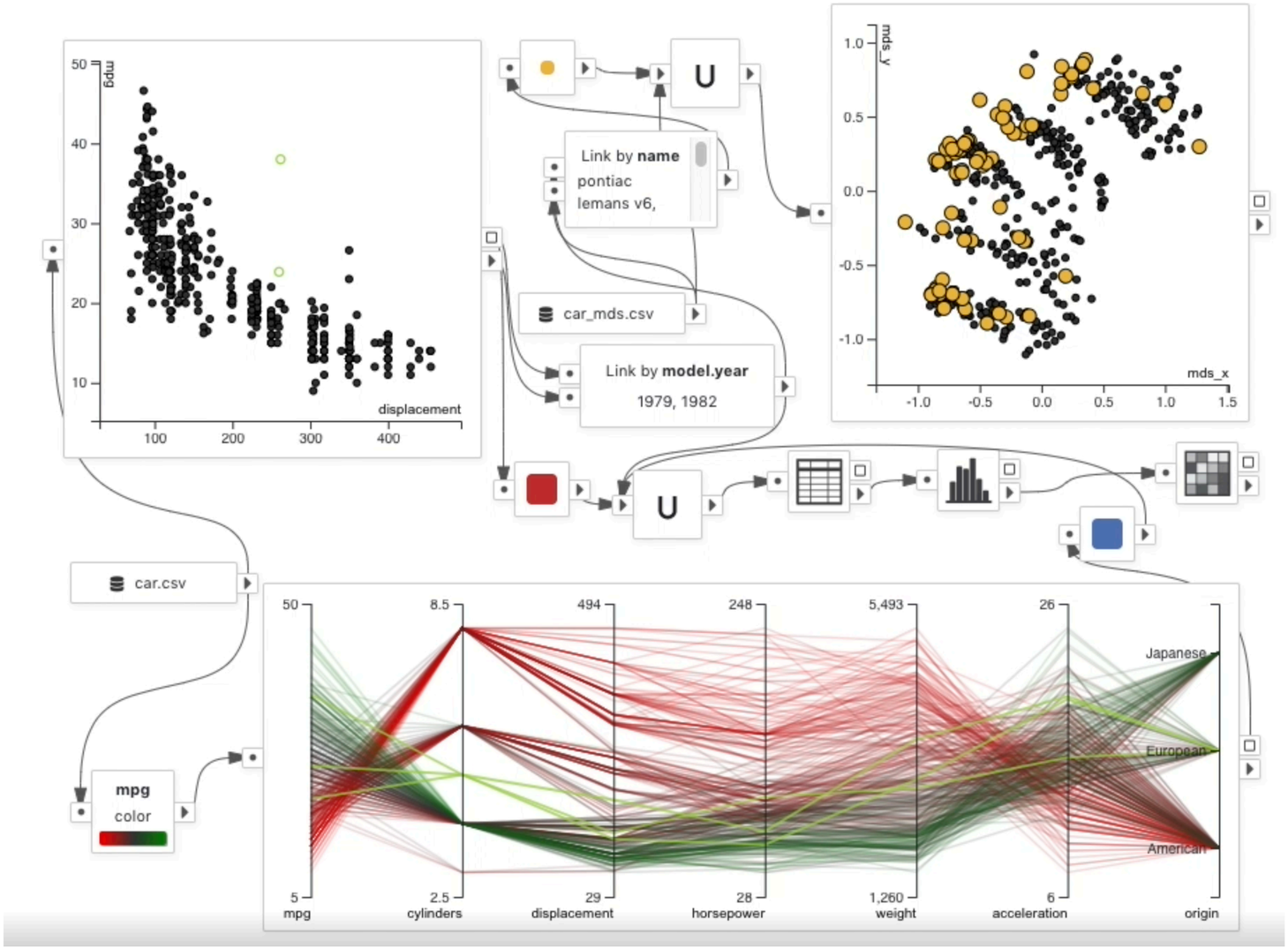
## Idea: Use Literate Programming to Justify Visualization Design Choices




Design Exposition with Literate Visualization, Wood, Kachkaev, Dykes, 2018

https://www.gicentre.net/litvis

# Pipeline-Based Visualization Specification




**Pro:**

**Explicit, similar to source code**

**Con:**

**Complex**

**Still not "Literate"**

Bowen Yu, and Claudio T. Silva, 2017

# Provenance Tracking

**Provenance (from the French provenir, 'to come from/ forth') is the chronology of the ownership, custody or location of a historical object.** [Wikipedia]

**In CS: a log, a record of everything that lead to a state**

# Provenance Uses

Recall

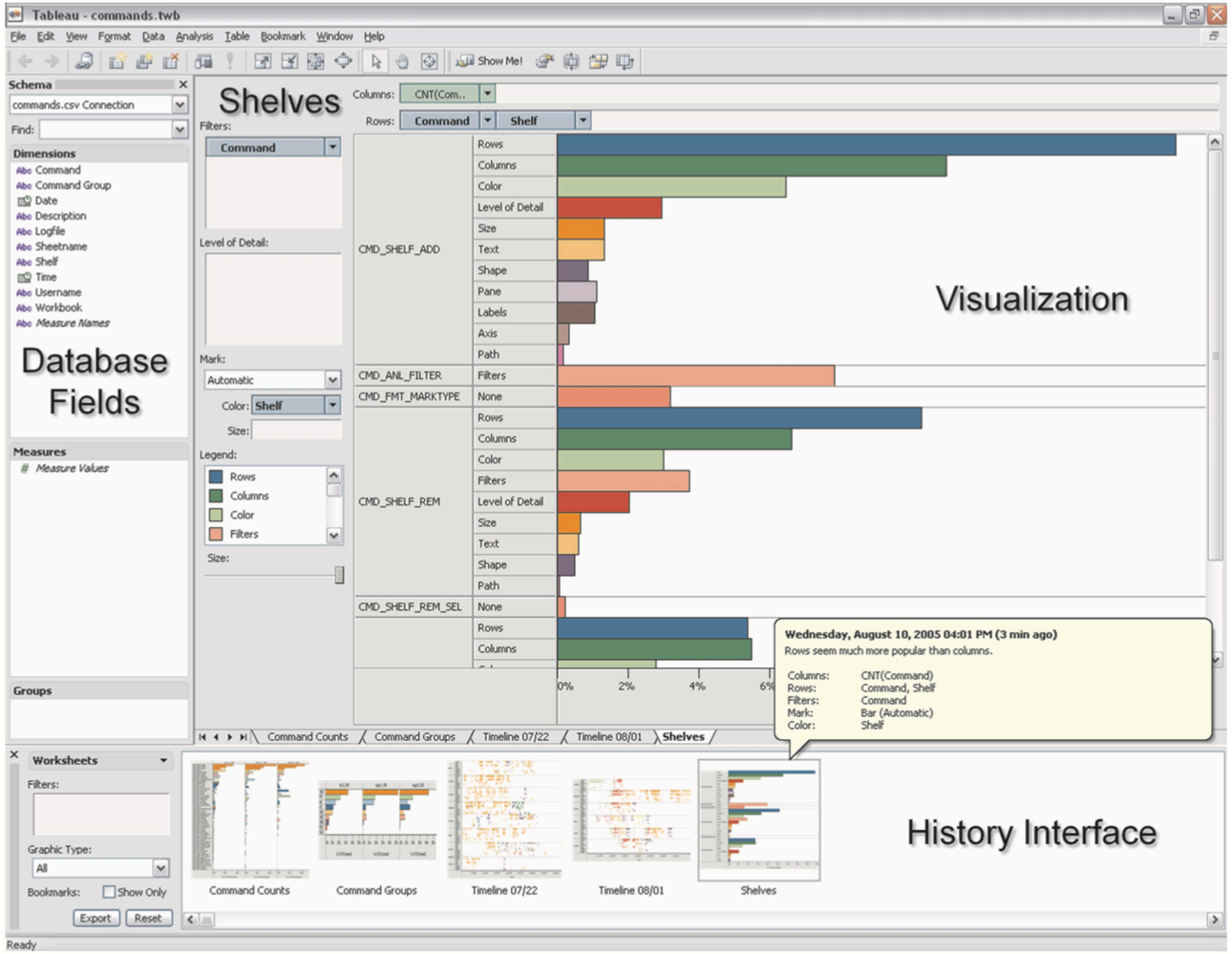
Replication

Action Recovery

Collaborative Comm.

Presentation

Meta-Analysis




Ragan et al., 2015

Heer et al., 2008

# How are “Literate Programs” Created?

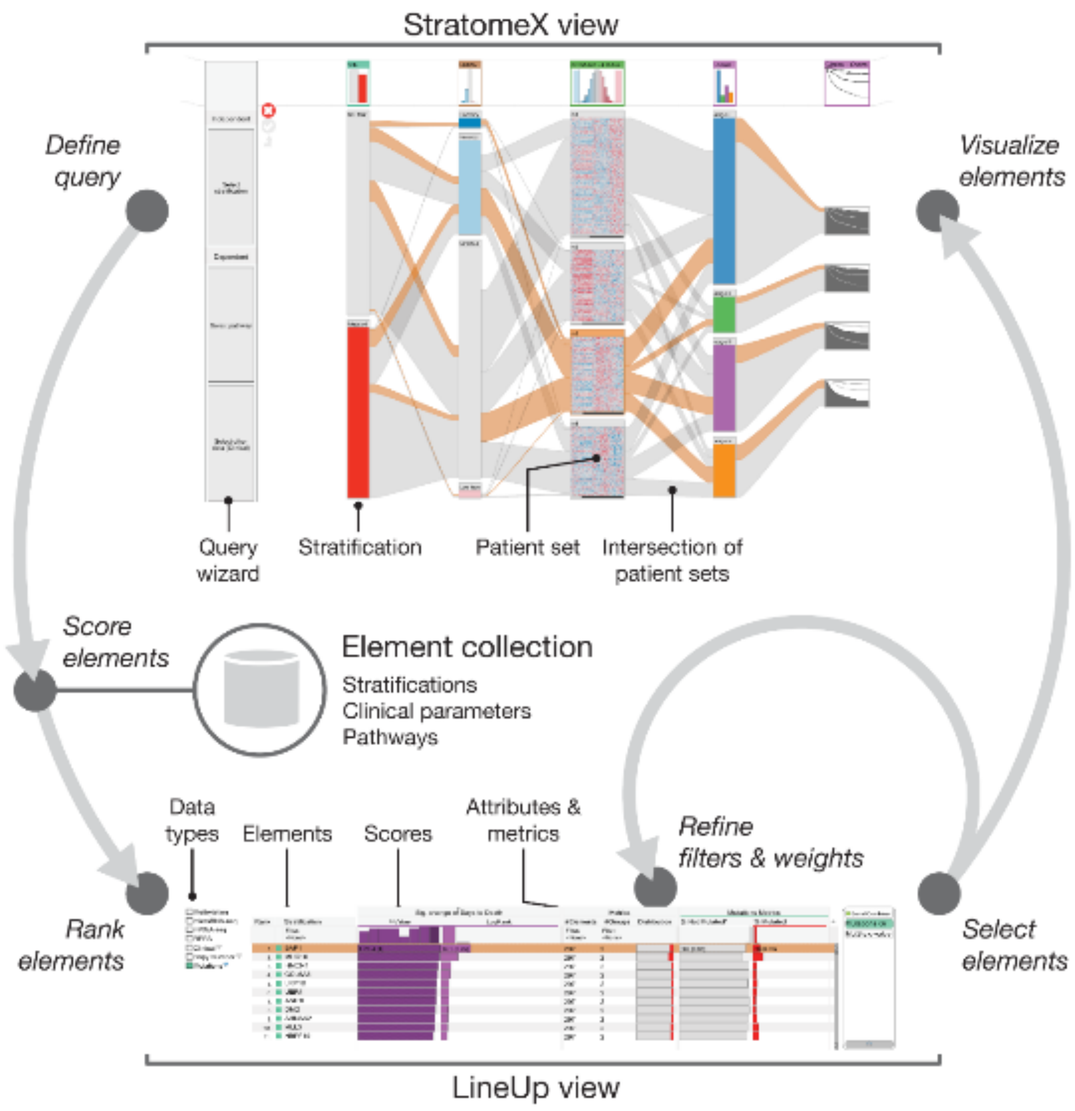
1. Write draft code that might work for test case
2. Edit code to generalize
3. Refactor for readability, good SE practices
4. Document
5. Provide context with images/links/equations/text

-> Complex, multi-stage process

A history (e.g., version control) isn’t very relevant.

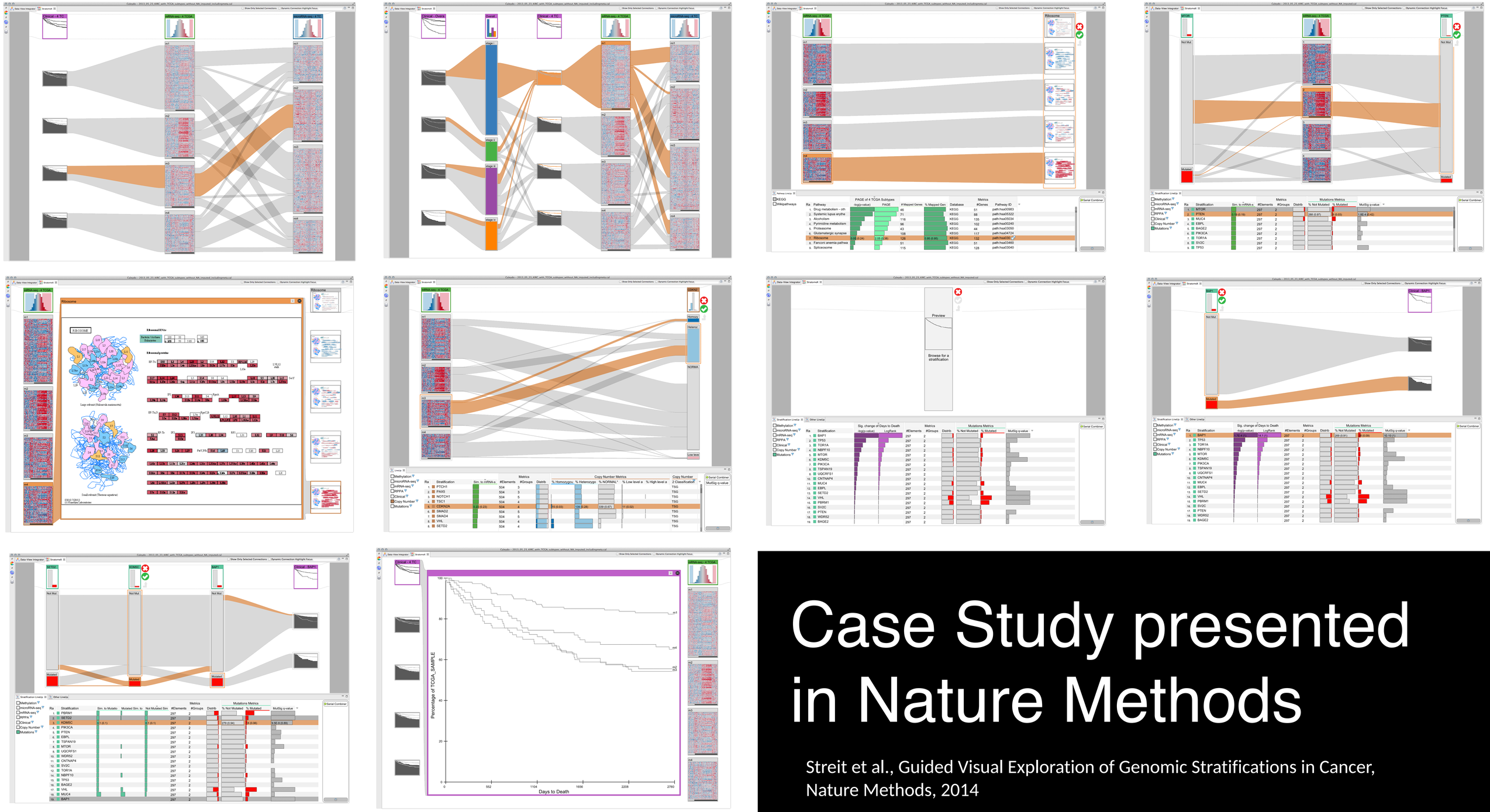
-> Visualization Provenance is necessary but insufficient

# How to do Literate Visualization?




Guided Visual Exploration of Genomic Stratifications in Cancer

Nature Methods 11, 9 (2014), 884–885

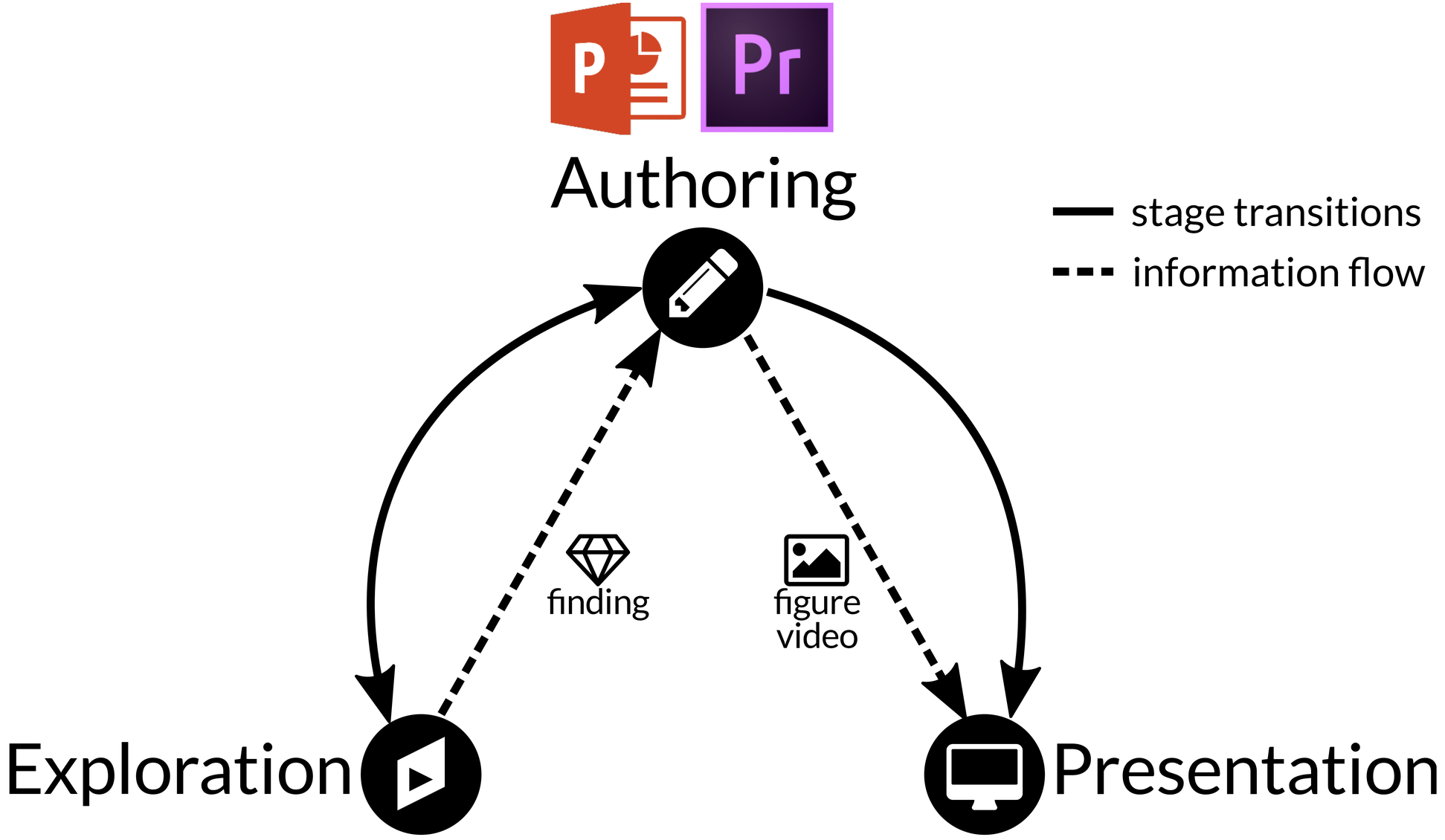


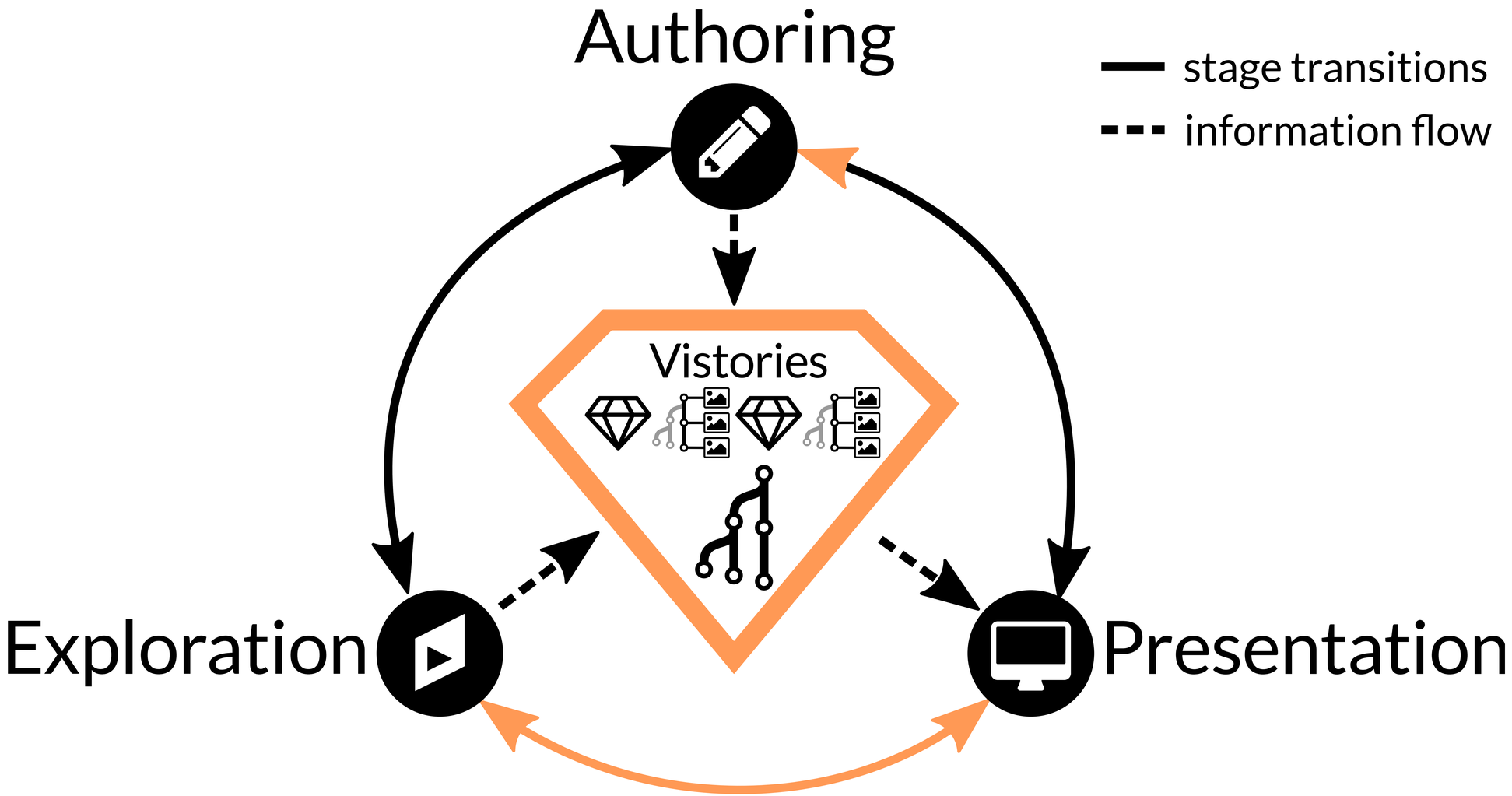
# Case Study presented in Nature Methods

Streit et al., Guided Visual Exploration of Genomic Stratifications in Cancer, Nature Methods, 2014

Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Nicola Cosgrove, Marc Streit (EuroVis 2016)

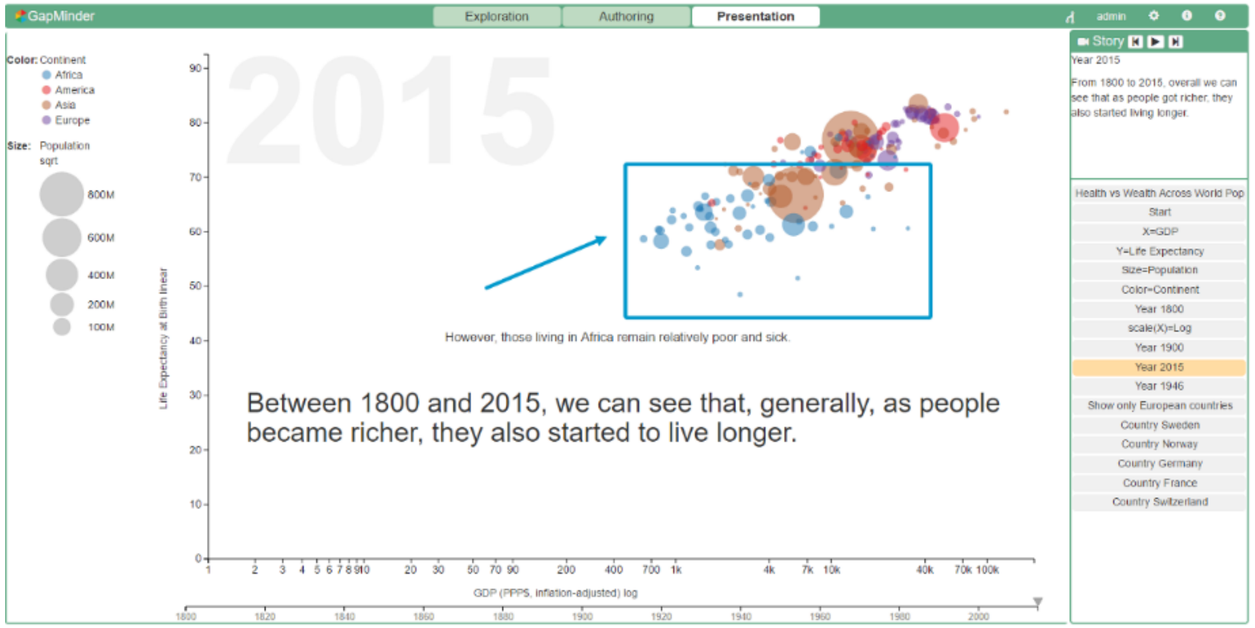
# VISTORIES
# FROM VISUAL EXPLORATION TO STORYTELLING AND BACK AGAIN


Authoring
stage transitions
information flow
finding
figure
video
Exploration
Presentation



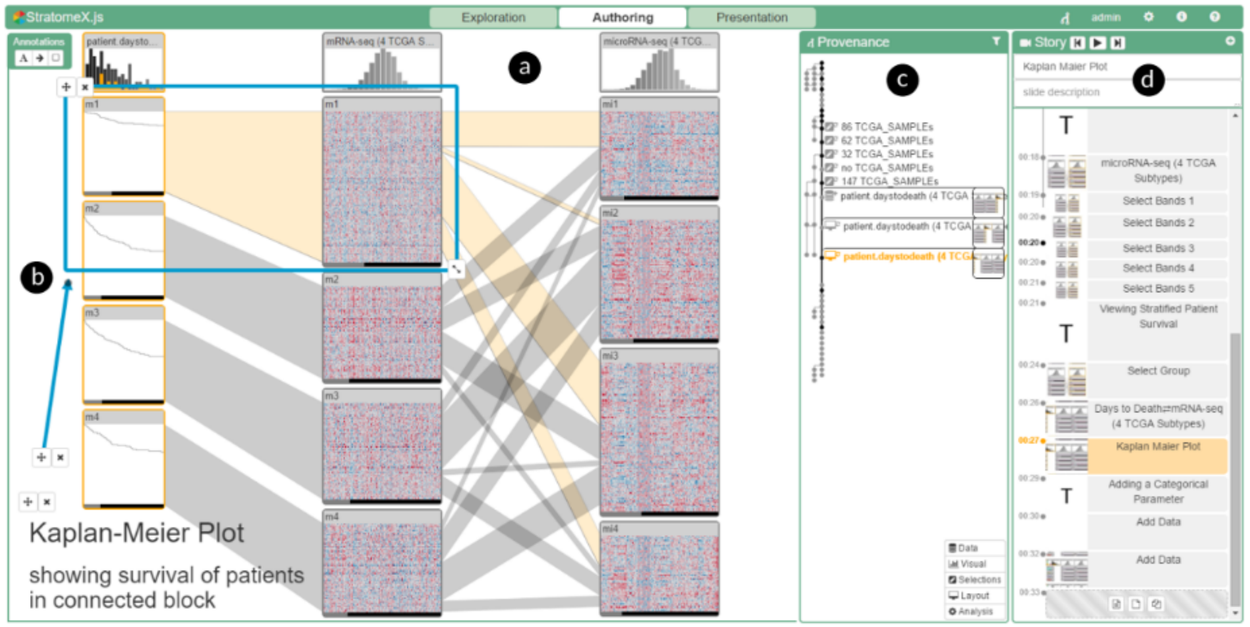

Gratzl et al., Computer Graphics Forum (EuroVis), 2016

# Examples




Gapminder

http://vistories.org/gapminder




StratomeX

http://vistories.org/stratomex

CLUE

Exploration | Authoring | Presentation

admin

Color: Continent
- Africa
- America
- Asia
- Europe

Size: Population
Sqrt

800M
600M
400M
200M
100M

1900

Life Expectancy at Birth Linear

90 80 70 60 50 40 30 20 10 0

1 2 3 4 5 6 7 8 9 10 20 30 50 70 90 200 400 700 1k 4k 7k 10k 40k 70k 100k

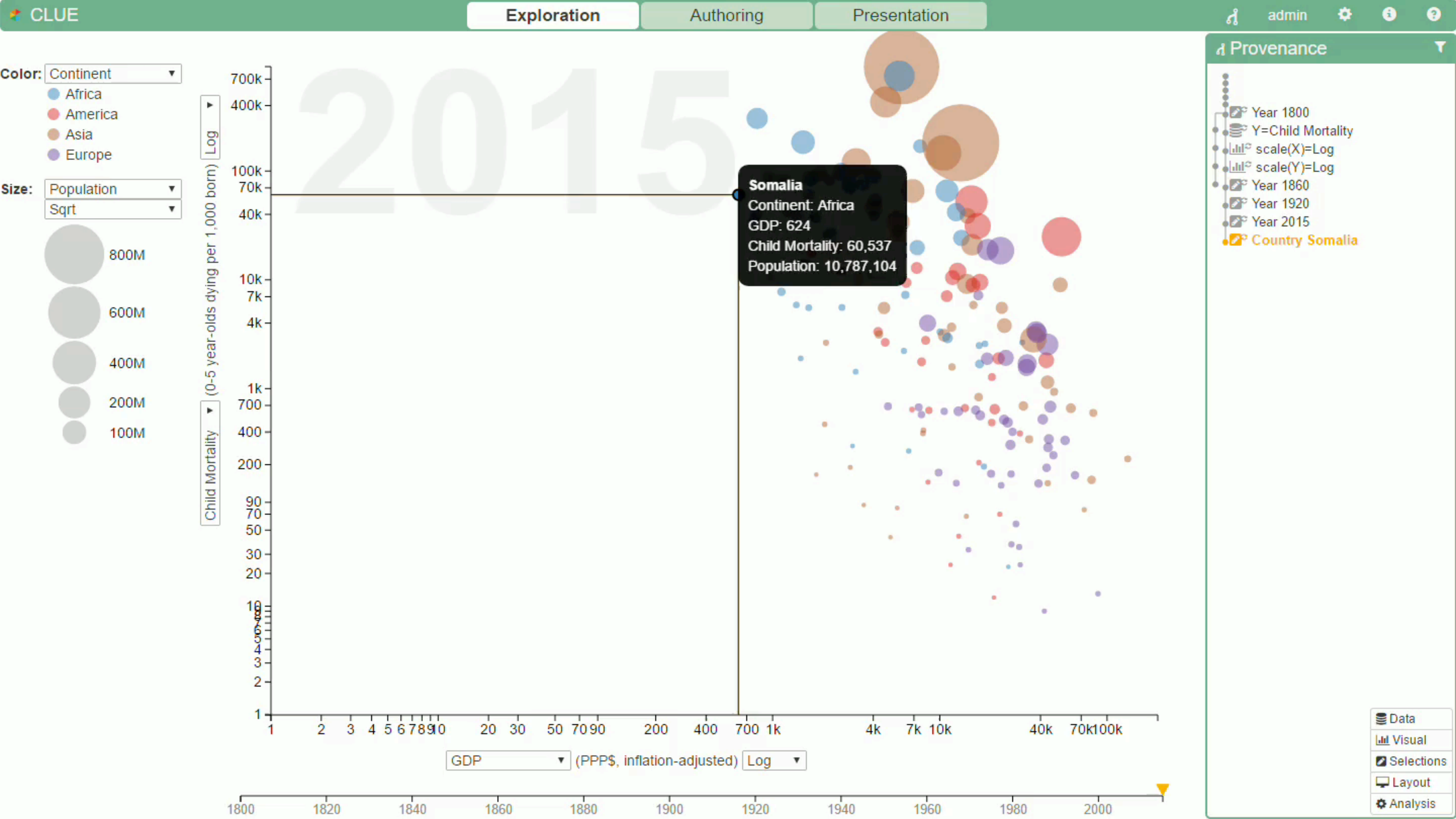
GDP (PPP$, inflation-adjusted) Log

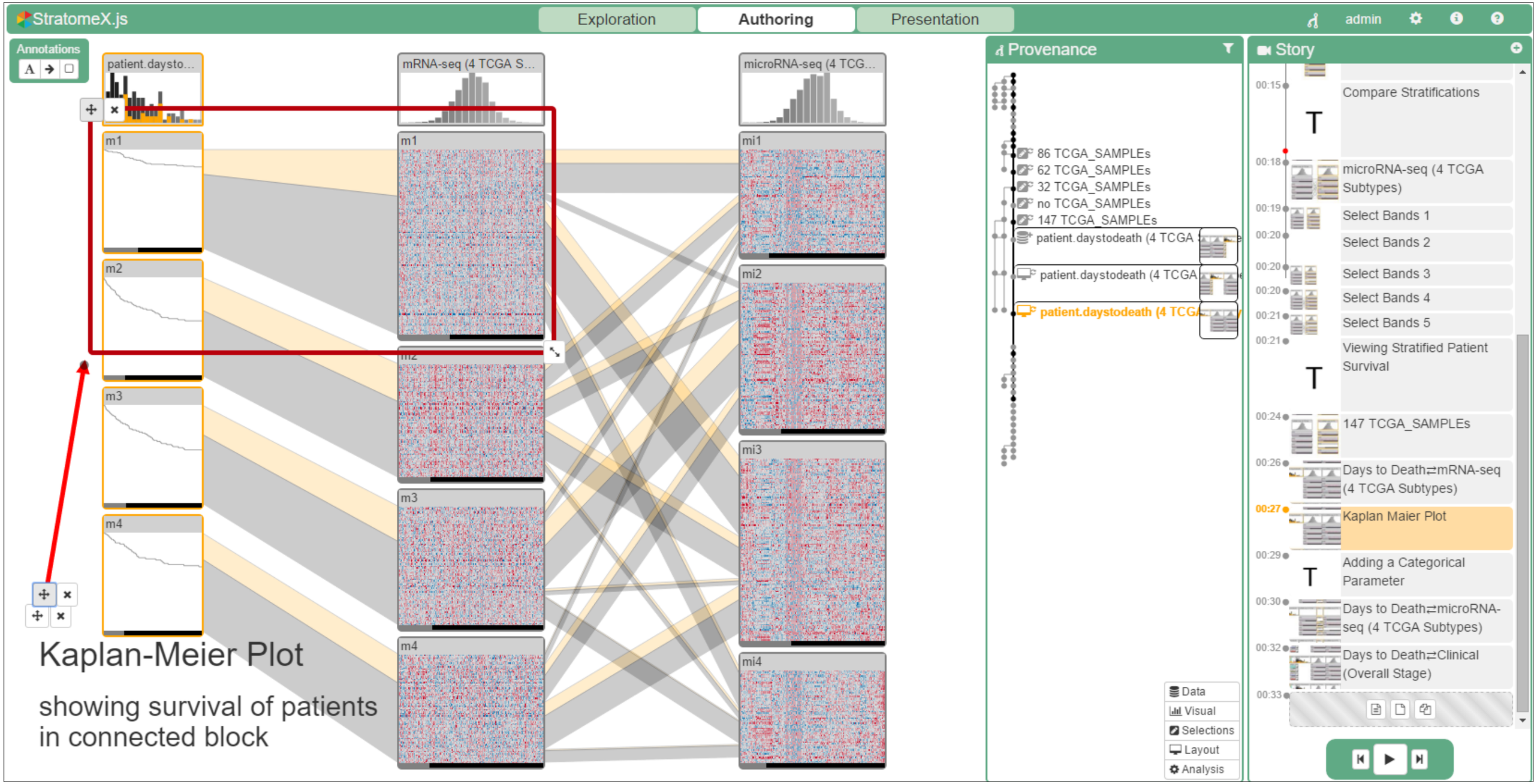
1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000

Provenance

- Start
- X=GDP
- Y=Life Expectancy
- Size=Population
- Color=Continent
- Year 1800
- scale(X)=Log
- Year 1900

Data
Visual
Selections
Layout
Analysis


CLUE
Exploration
Authoring
Presentation
admin
Provenance
Color: Continent
Africa
America
Asia
Europe
Size: Population
Sqrt
800M
600M
400M
200M
100M
2015
Log
Child Mortality (0-5 year-olds dying per 1,000 born)
Somalia
Continent: Africa
GDP: 624
Child Mortality: 60,537
Population: 10,787,104
GDP (PPP$, inflation-adjusted) Log
Year 1800
Y=Child Mortality
scale(X)=Log
scale(Y)=Log
Year 1860
Year 1920
Year 2015
Country Somalia
Data
Visual
Selections
Layout
Analysis




http://vistories.org/stratomex

# Challenges

**Privacy**

- Curated analysis stories vs full log

**Integration**

- Visualization tools are only one part of the workflow

**Scalability**

- Provenance Graph gets fast quickly

**Effort**

- It’s work to create an analysis story

**Re-Use**

- Unlike code, can’t easily apply a visual analysis to updated data

# Provenance is “dumb”

Record of what you did, but not why you did it.

Knowing why would help with scalability, effort, and re-use

So, can we make it easier to capture the why?

Kiran Gadhave, Jochen Görtler, Oliver Deussen,
Miriah Meyer, Jeff Phillips, Alexander Lex

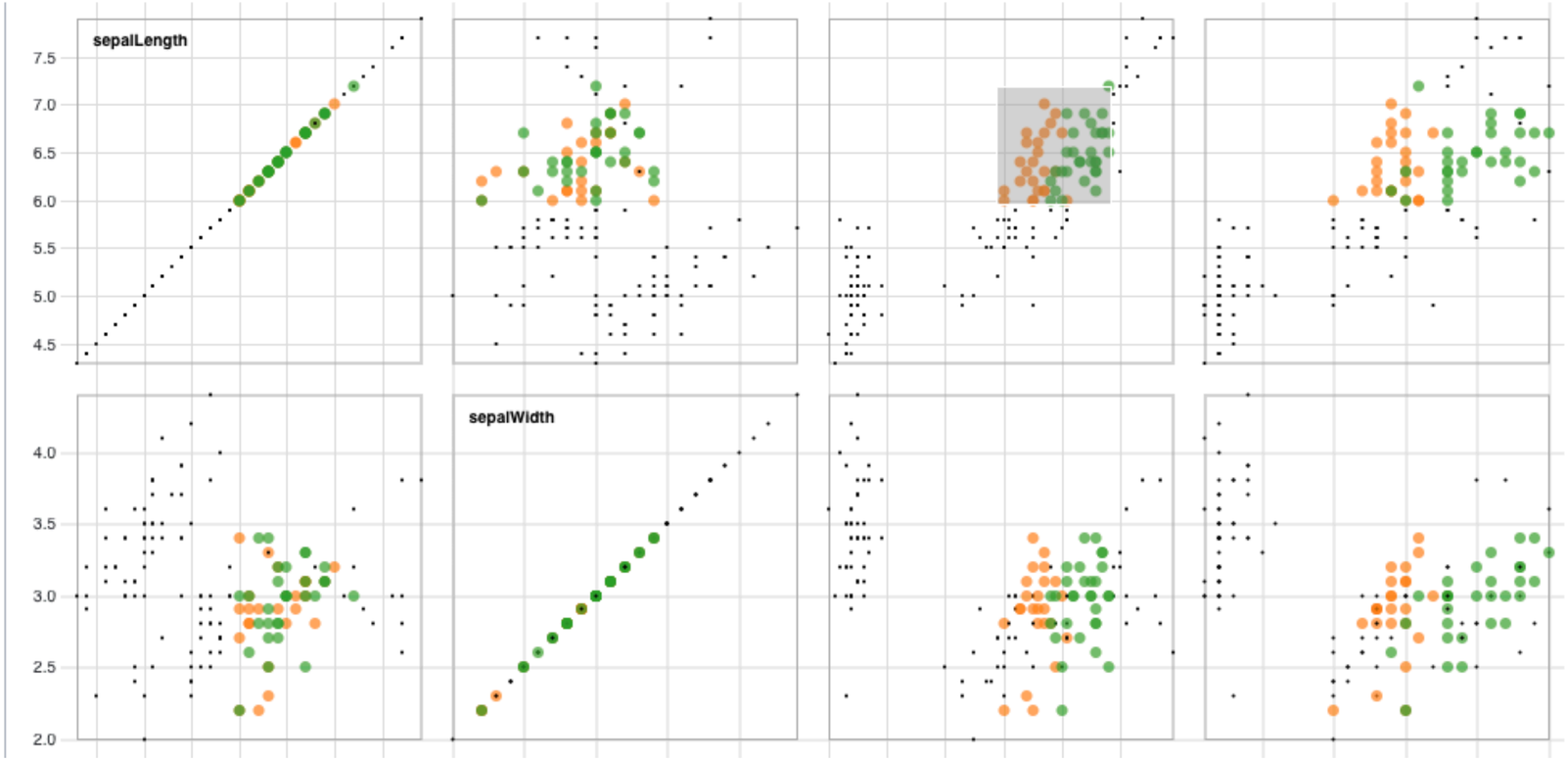
# CAPTURING USER INTENT WHEN BRUSHING IN SCATTERPLOTS

# What's Brushing?

Published Jun 15, 2019 Fork of Scatterplot Matrix · 4 Forks 1 File Listed in d3-brush

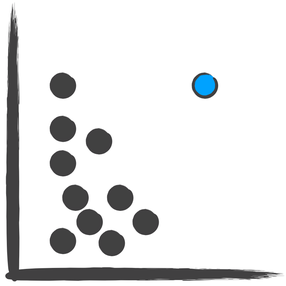
## Brushable Scatterplot Matrix

This scatterplot matrix allows brushing to select data points in one cell, and highlight them across all other cells.

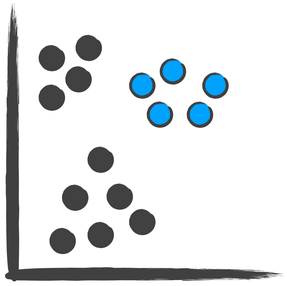


# What is Intent when Brushing?

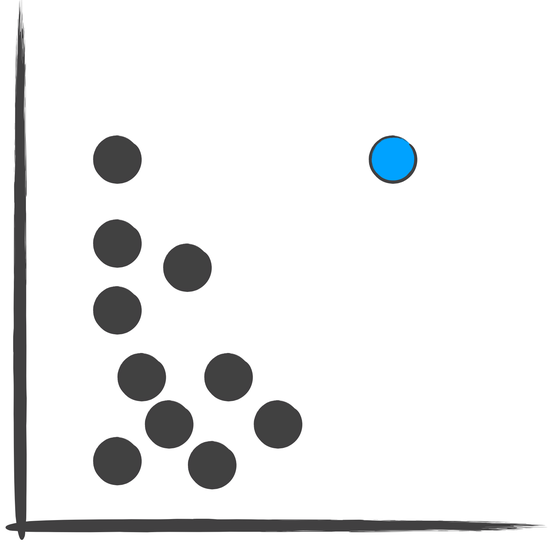
**Intent is the user's reason for performing a brush with a visualization.**



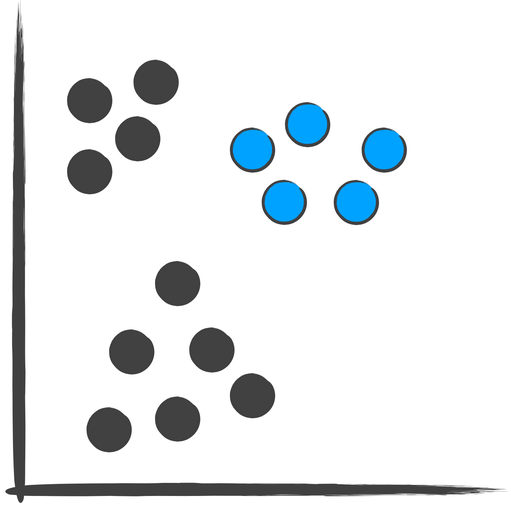
Outlier



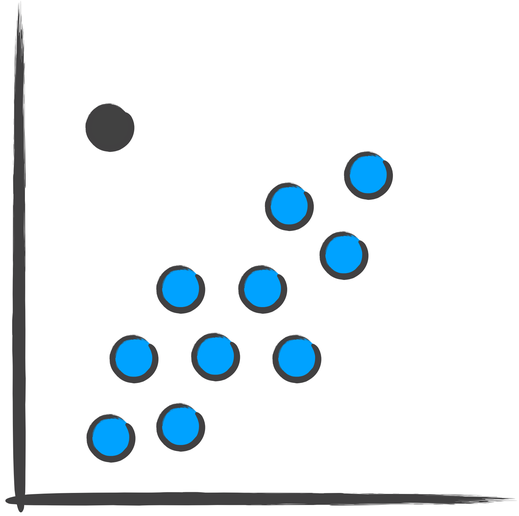
Clusters



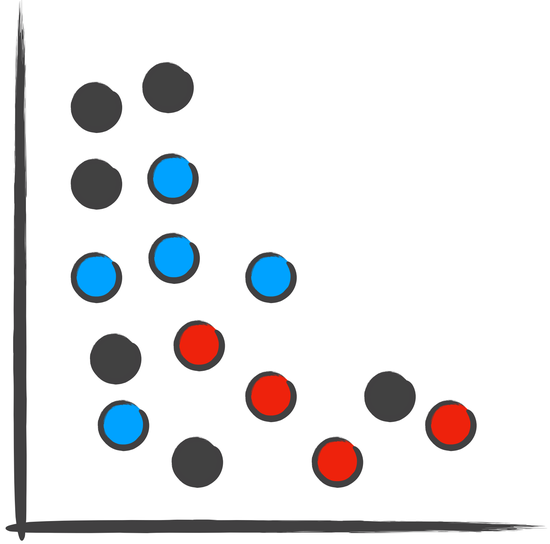
Outlier



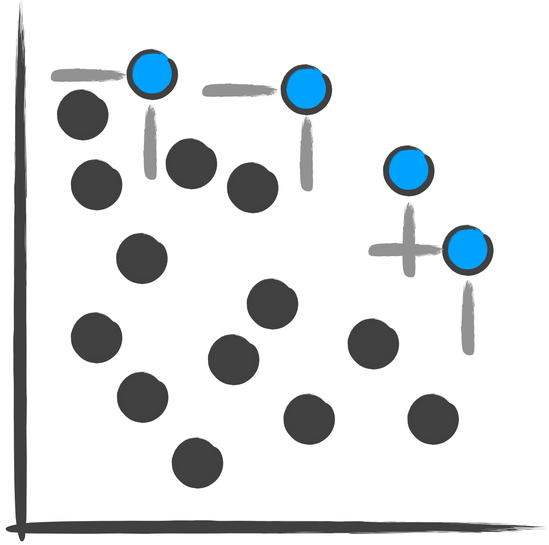
Clusters



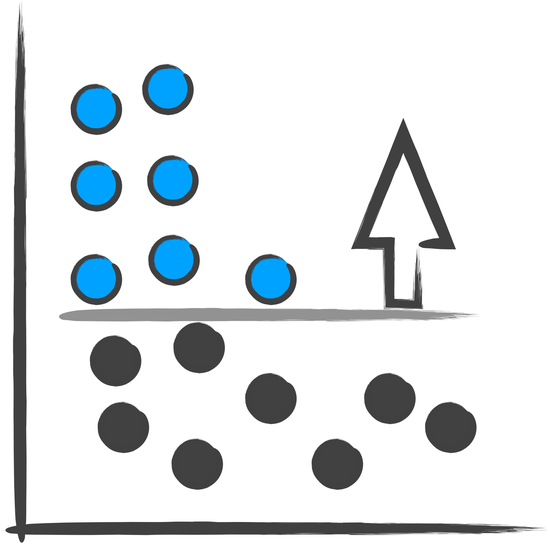
Correlation



Categories

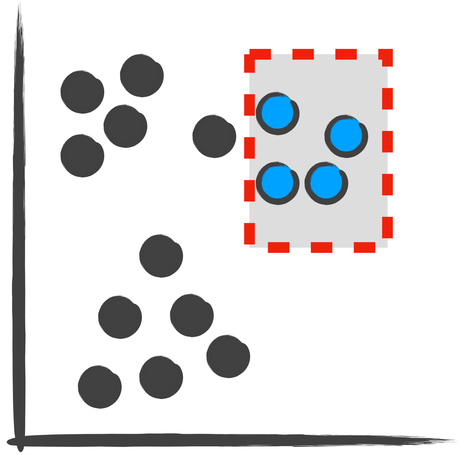


Multivariate Optimization



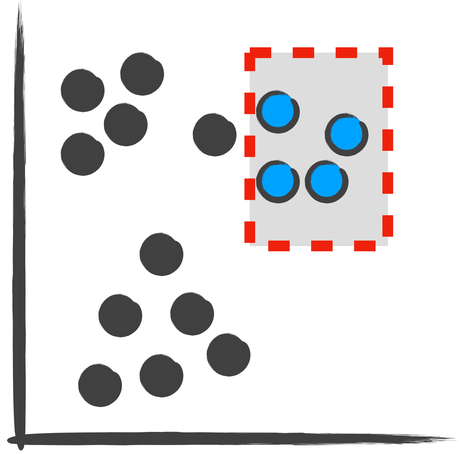
Ranges

# How do we infer Intent?

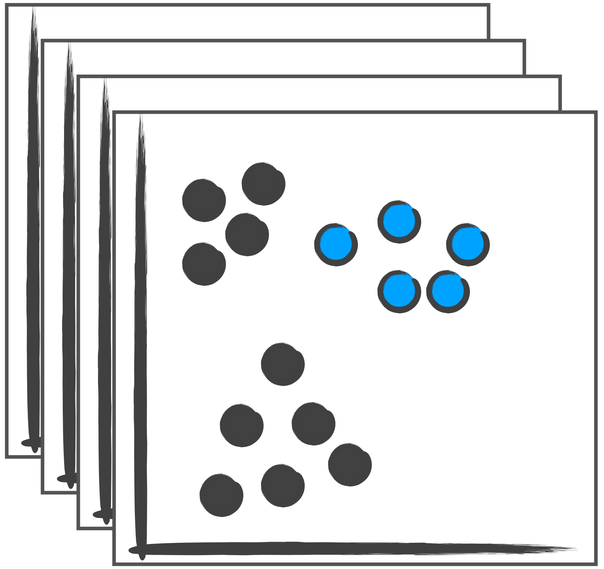


Selection

# How do we infer Intent?



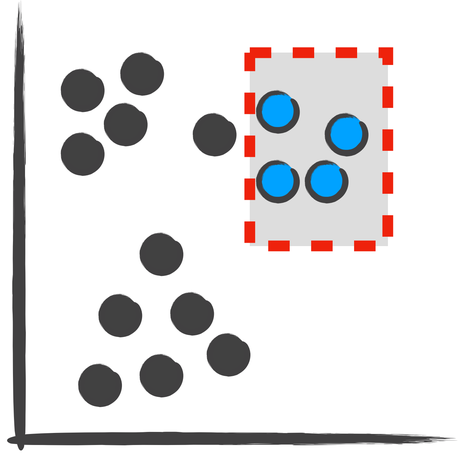
**Selection**



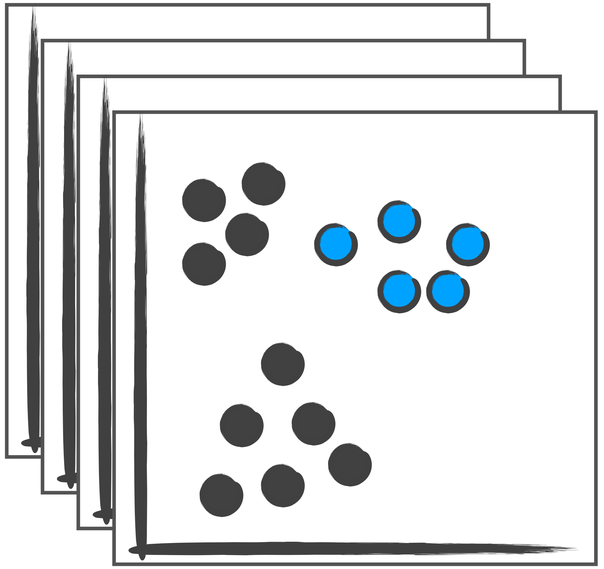
**Predictions**

K-Means
DBScan
Regression
Outlier Detection
Skyline
Decision Trees / Ranges
Categories

# How do we infer Intent?



**Selection**



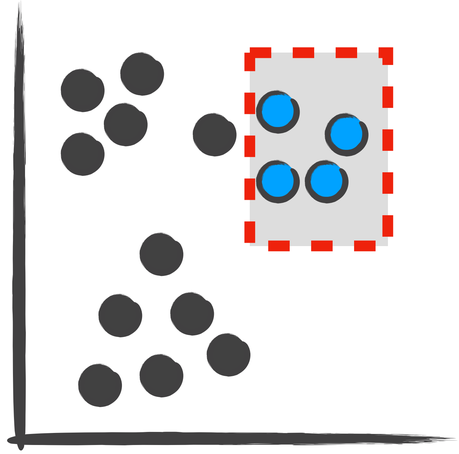
**Predictions**

K-Means
DBScan
Regression
Outlier Detection
Skyline
Decision Trees / Ranges
Categories

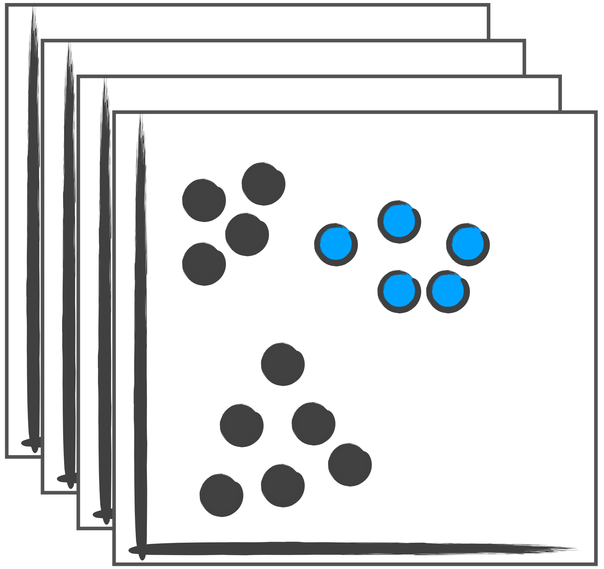
1. Range
2. Cluster
3. Outlier

**Ranking**

# How do we infer Intent?



**Selection**



**Predictions**

K-Means
DBScan
Regression
Outlier Detection
Skyline
Decision Trees / Ranges
Categories

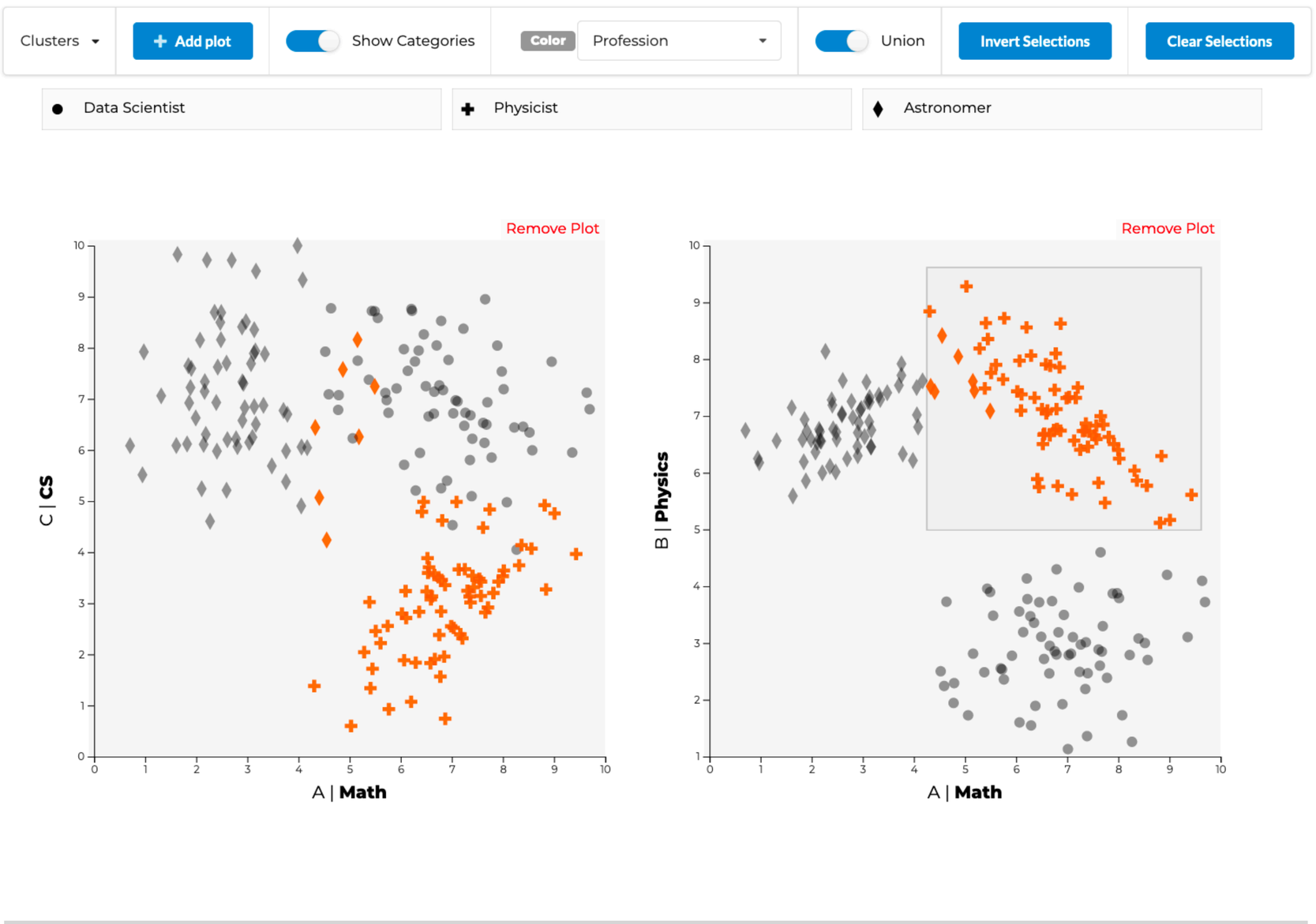
1. Range
2. Cluster
3. Outlier

**Ranking**

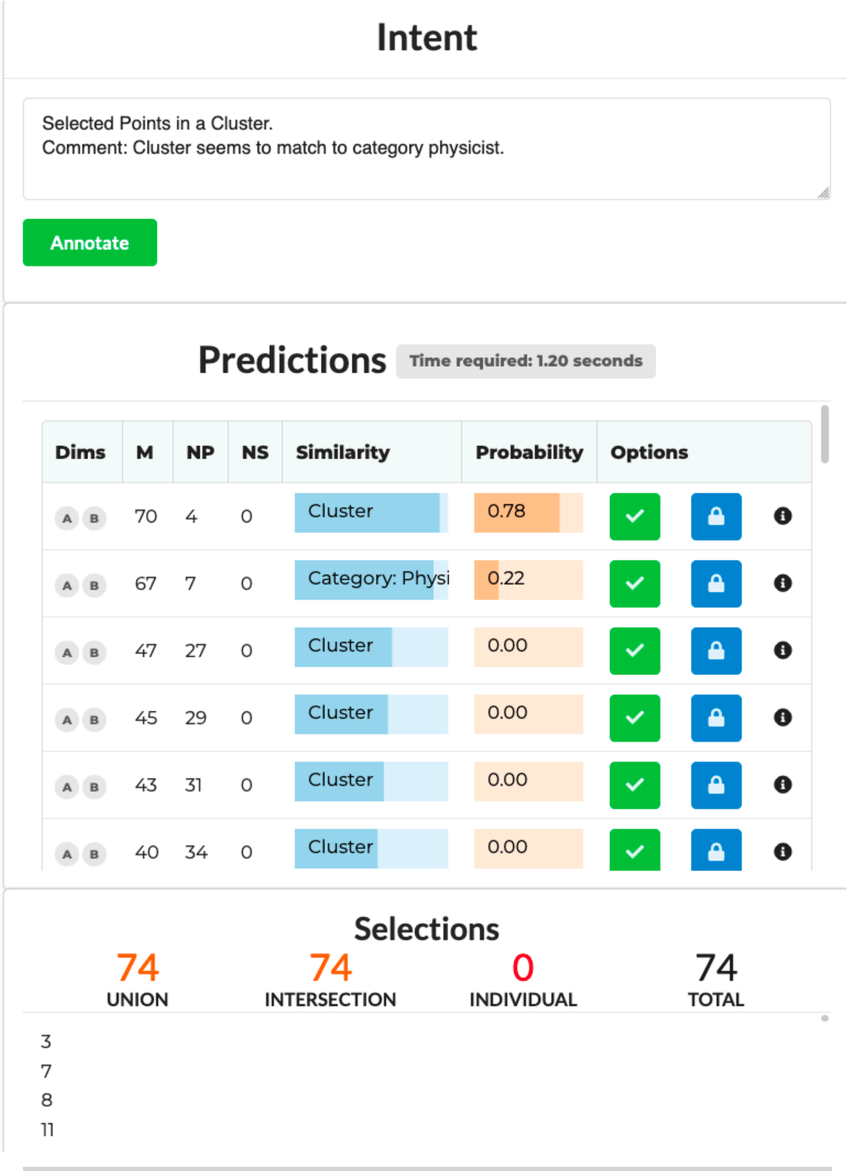
1. Range
2. Cluster
3. Outlier

*I think this cluster…*

**Confirming Intent & Annotation**




Visualization and Selection




Annotation of Intent and Predictions

**http://vdl.sci.utah.edu/predicting-intent/**

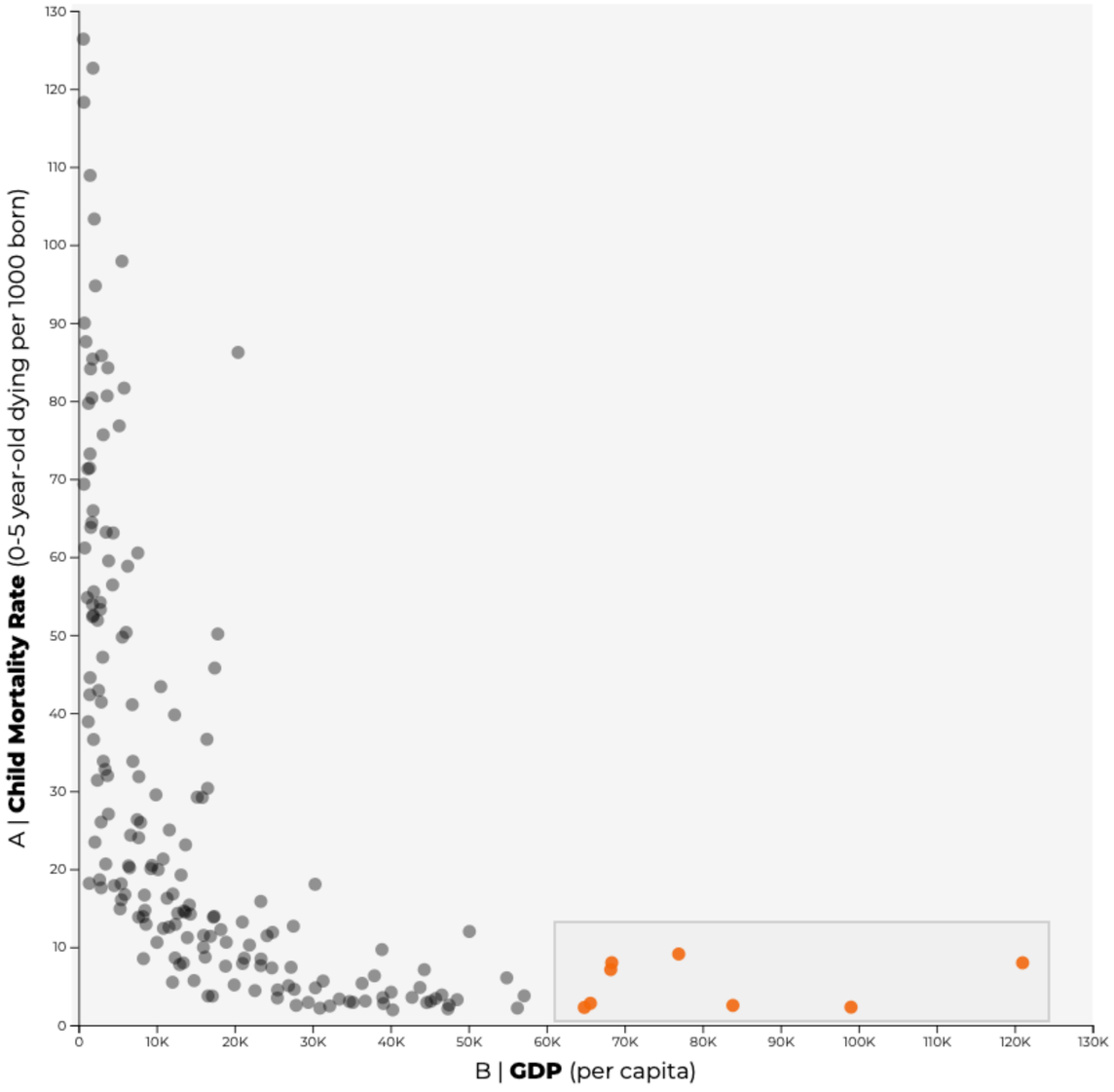
# Detail: Ranges

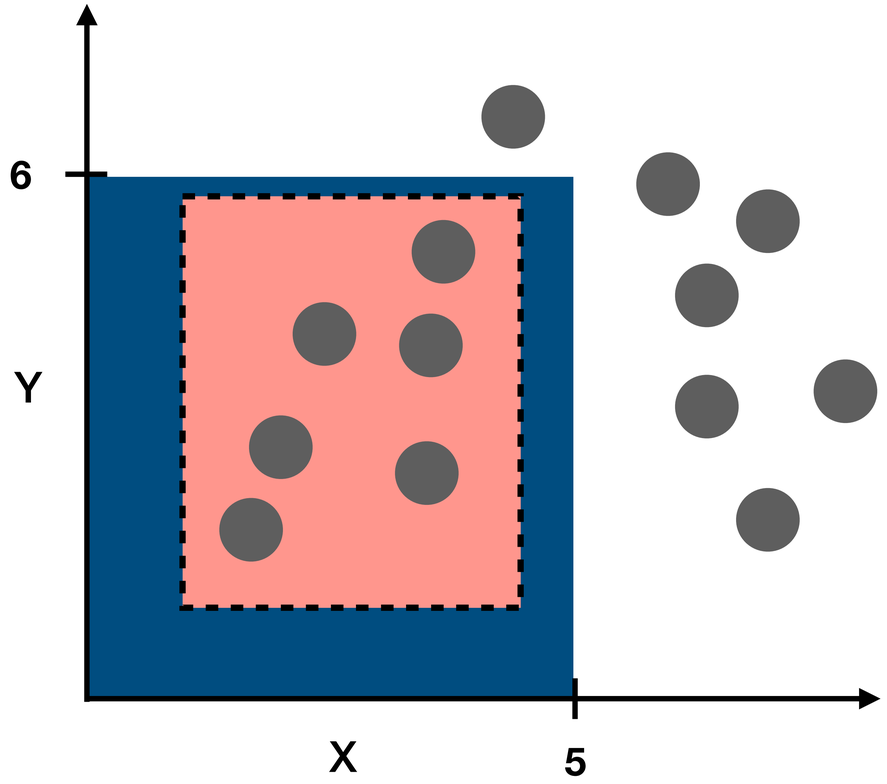
## What did I want to select?

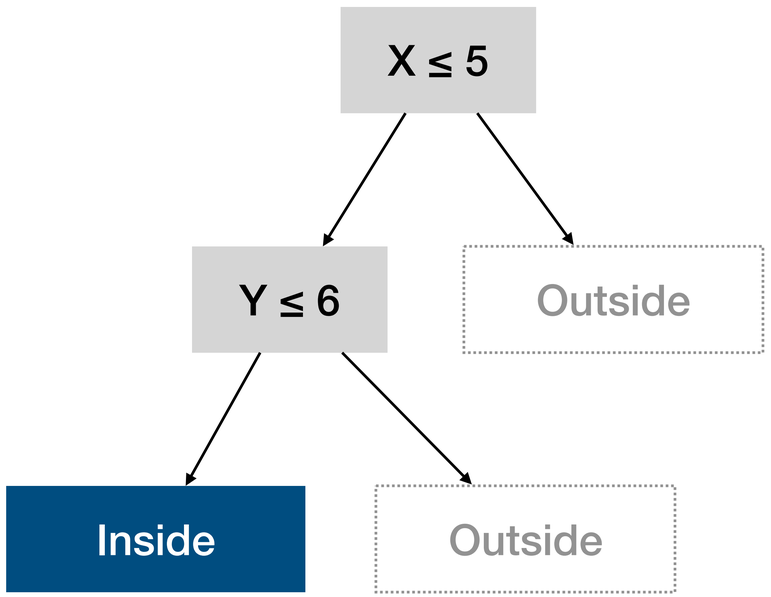
**GDP > 61,250**
**AND GDP < 111,250**
**AND Child Mortality Rate > 1.5**
**AND Child Mortality Rate < 13**

## Or

## GDP > 60k?

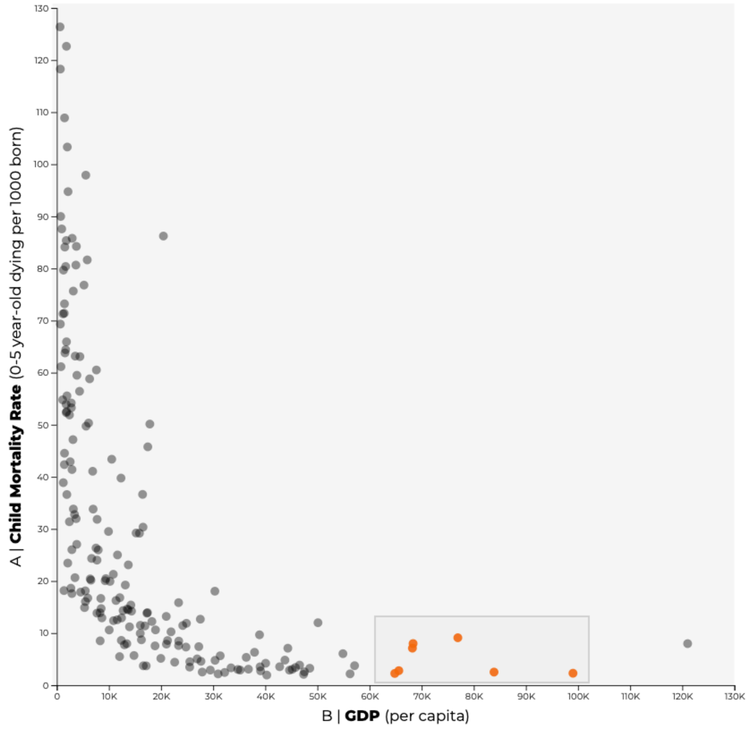


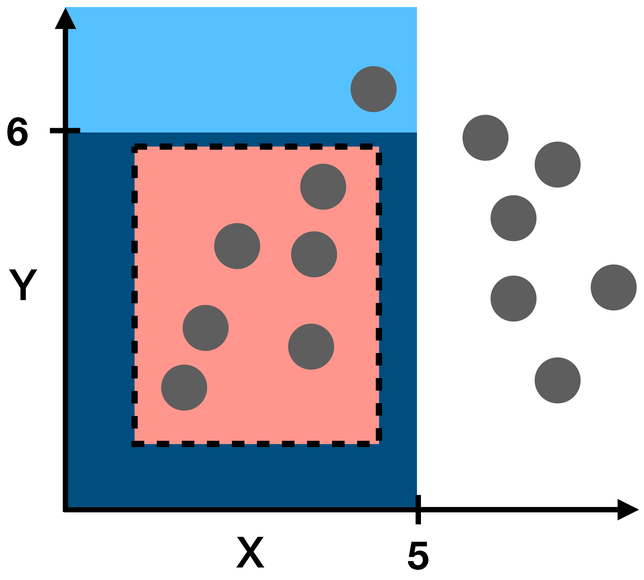

6
Y
X
5

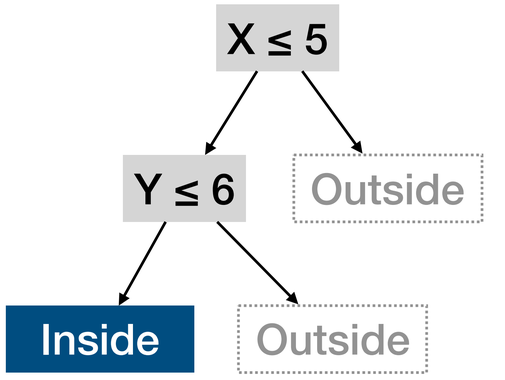

X ≤ 5
Y ≤ 6
Outside
Inside
Outside

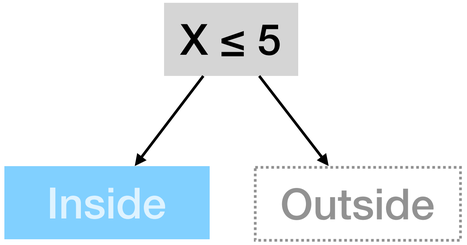
# Correcting Errors / Simplifying Selections

## Did I intent to select elements > 60k?



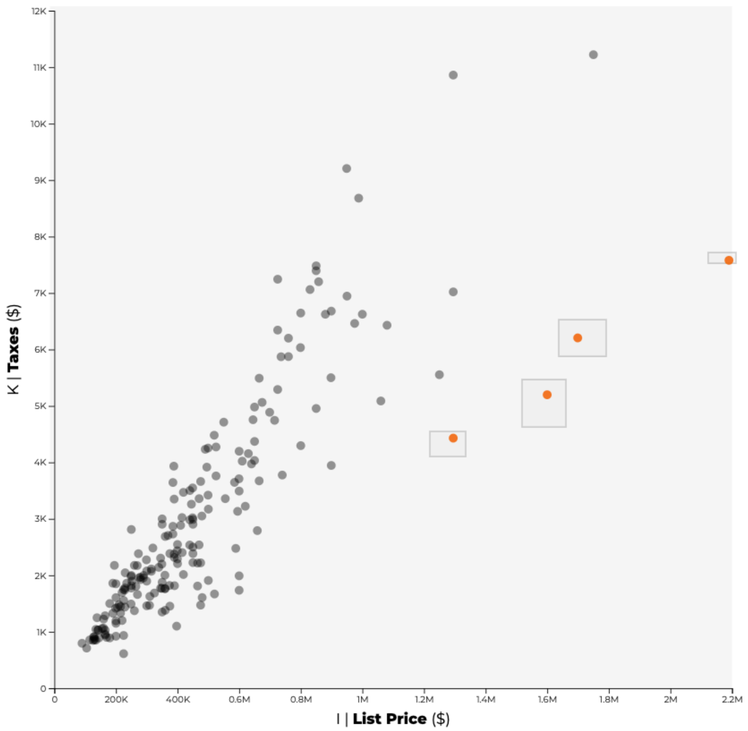



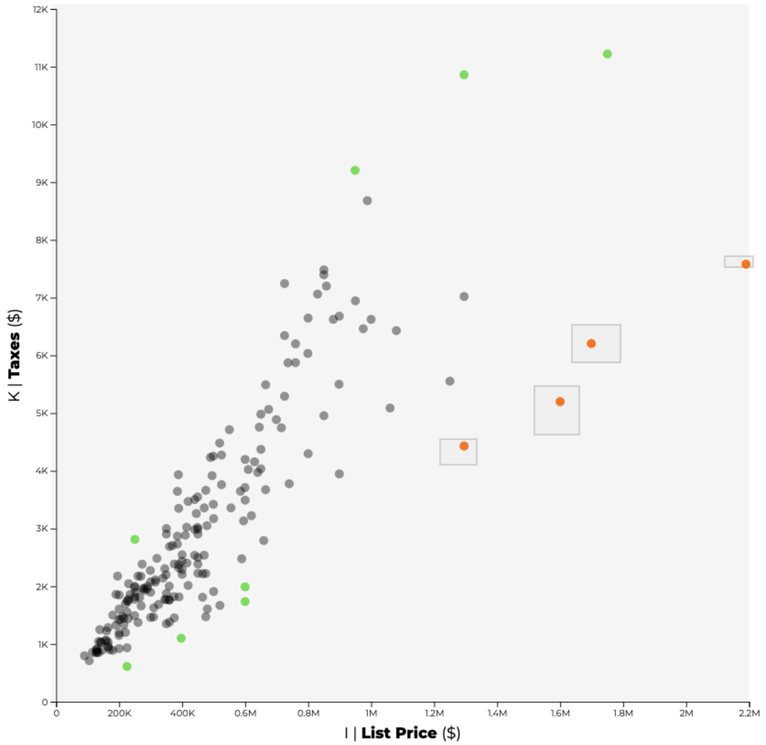



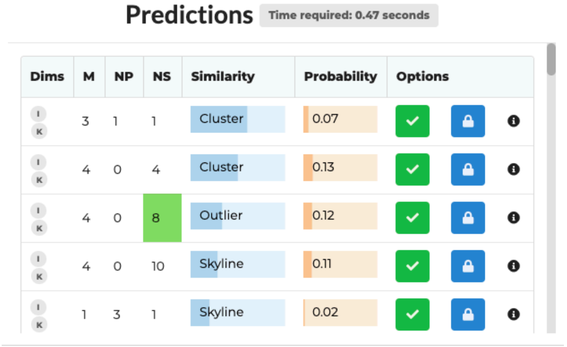
# Correcting Errors / Simplifying Selections

## Did I intent to select outliers?




## Turn prediction into selection.



# Implications

**Low effort to capture intent: reduce scalability problem**

**Understanding intent: potential for re-use**

**Correcting selection: speed up analysis**

# RECAP & OUTLOOK

# These are first steps…

**Reproducibility**

- Analysis stories and provenance make visualizations reproducible
- Annotating the Why makes them justifiable

**Communication**

- Curated stories make it easier for others (collaborators or “consumers”) to understand
- A future of “interactive figures”

**Re-Use**

- Curated, generalized visual analysis process can be applied to updated datasets

# Much is left to do!

Integration

Data Provenance

Review for Re-Use

Extend Intent Inference to other Vis Types and Interaction Types

python™

R

jupyter

+

Jupyter Notebook

Vistories

jupyter Taggle Last Checkpoint: 11/09/2017 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

In [5]:
```
import pandas as pd
```

In [6]:
```
%%javascript

window.asTaggle = (dataset, element) => {
  // element = JQuery output element magic by jupyter
  // increase output height
  element.parent('.output_area').css('height', '48em');
  element.css('position', 'relative');
  element.html('<main style="position: absolute; left: 0; right: 0; top: 0; bottom: 0"></main>');

  return new Promise((resolve) => {
    require(['./lib/TaggleJS_bundle.js'], function(taggle) {
      const parent = element.find('main')[0];

      const data = taggle.createLocalStorage(dataset, taggle.deriveColumnDescriptions(dataset));
      data.deriveDefault();
      instance = taggle.createTaggle(data, parent);
      instance.update();
      resolve(instance)
    });
  });
}
```

In [9]:
```
import json
from IPython.display import Javascript

Javascript("window.asTaggle({}, element);".format(iris.to_json(orient='records')))
```

Out[9]:

| Rank | sepal_length | sepal_width | petal_length | petal_width | species |
|---|---|---|---|---|---|
| 1 | | | | | virginica |
| 2 | | | | | virginica |
| 3 | | | | | virginica |
| 4 | | | | | virginica |
| 5 | | | | | virginica |
| 6 | | | | | virginica |
| 7 | | | | | virginica |
| 8 | | | | | virginica |
| 9 | | | | | virginica |

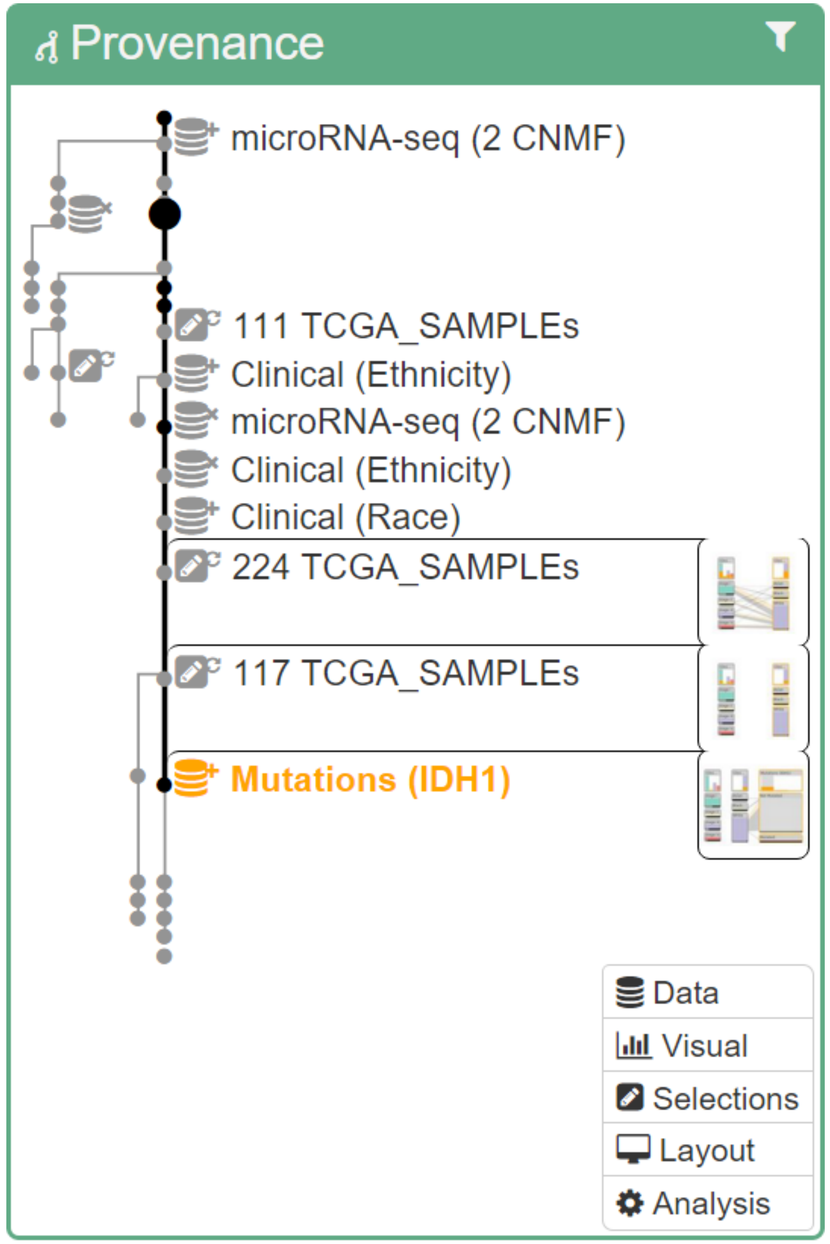
Overview

Add Column...

1 sepal_length

4.3 7.9

Filter out 0 missing value rows



**Alexander Lex**
@alexander_lex
http://alexander-lex.net

THE UNIVERSITY OF UTAH

**Thanks to: Kiran Gadhave, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Samuel Gratzl, Nils Gehlenborg, Nicola Cosgrove, Marc Streit**

**Funded by NSF, the Utah Genome Project, NIH, and DoD**

HEALTH
UNIVERSITY OF UTAH

NSF

NIH

visualization design lab

SCI
www.sci.utah.edu