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

visualization

The purpose of computing is insight,

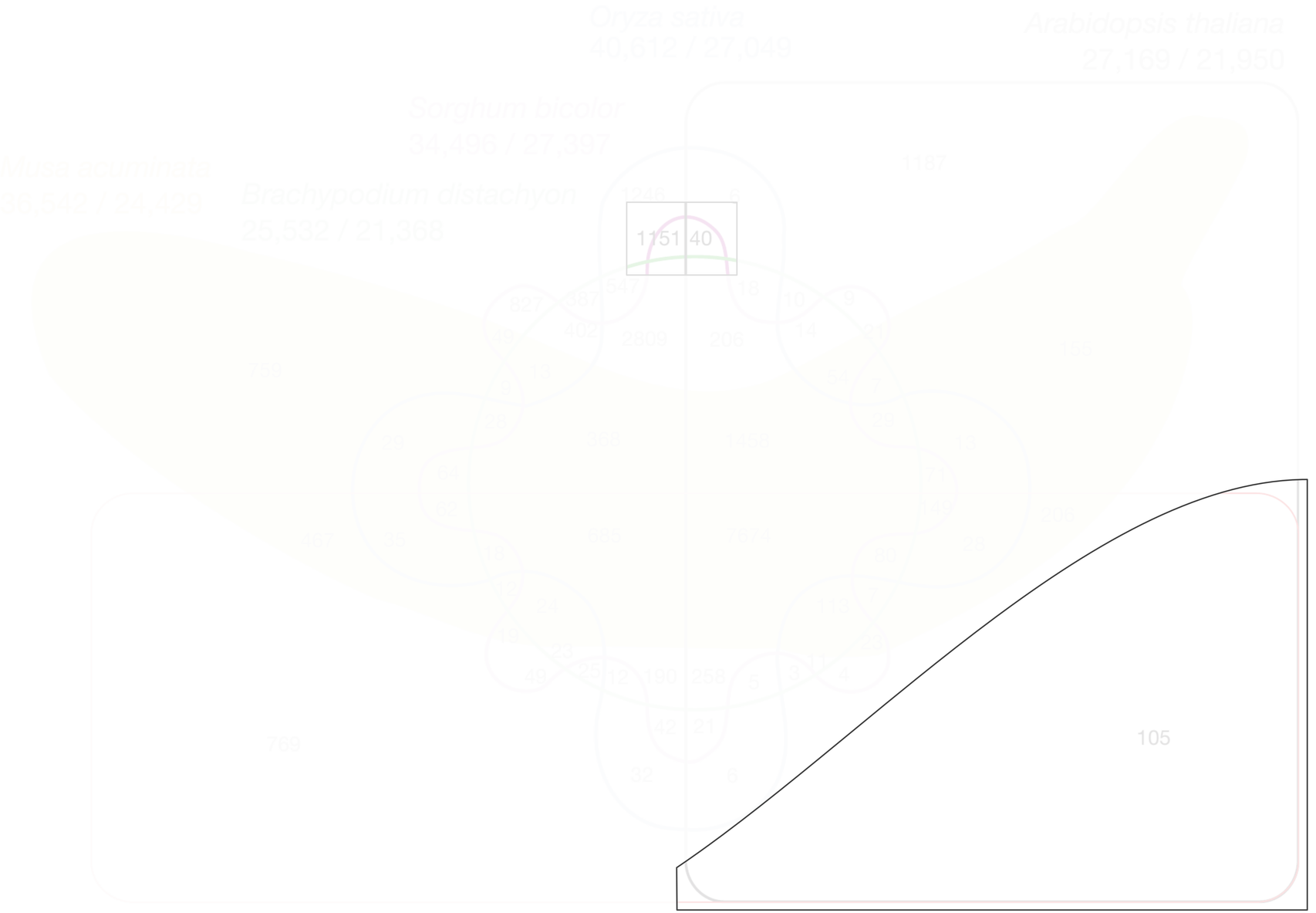
not numbers.

pictures

[Card, Mackinlay, Shneiderman]

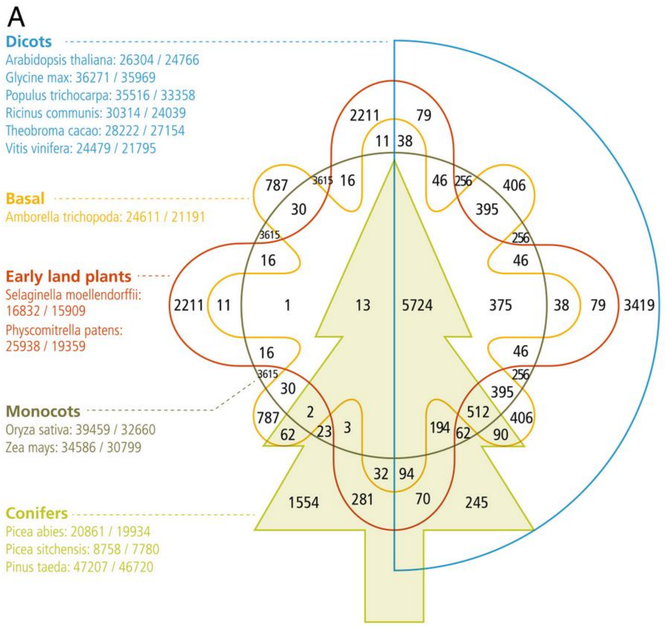
[Richard Wesley Hamming]

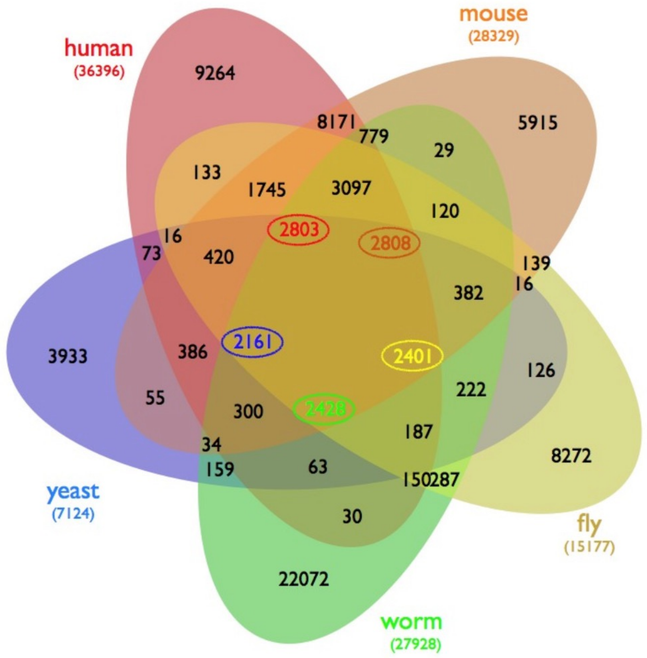
Banana	*M. acuminata*
Date	*P. dactylifera*
Cress	*Arabidopsis thaliana*
Rice	*Oryza sativa*
Sorghum	*Sorghum bicolor*
Brome	*Brachypodium distachyon*



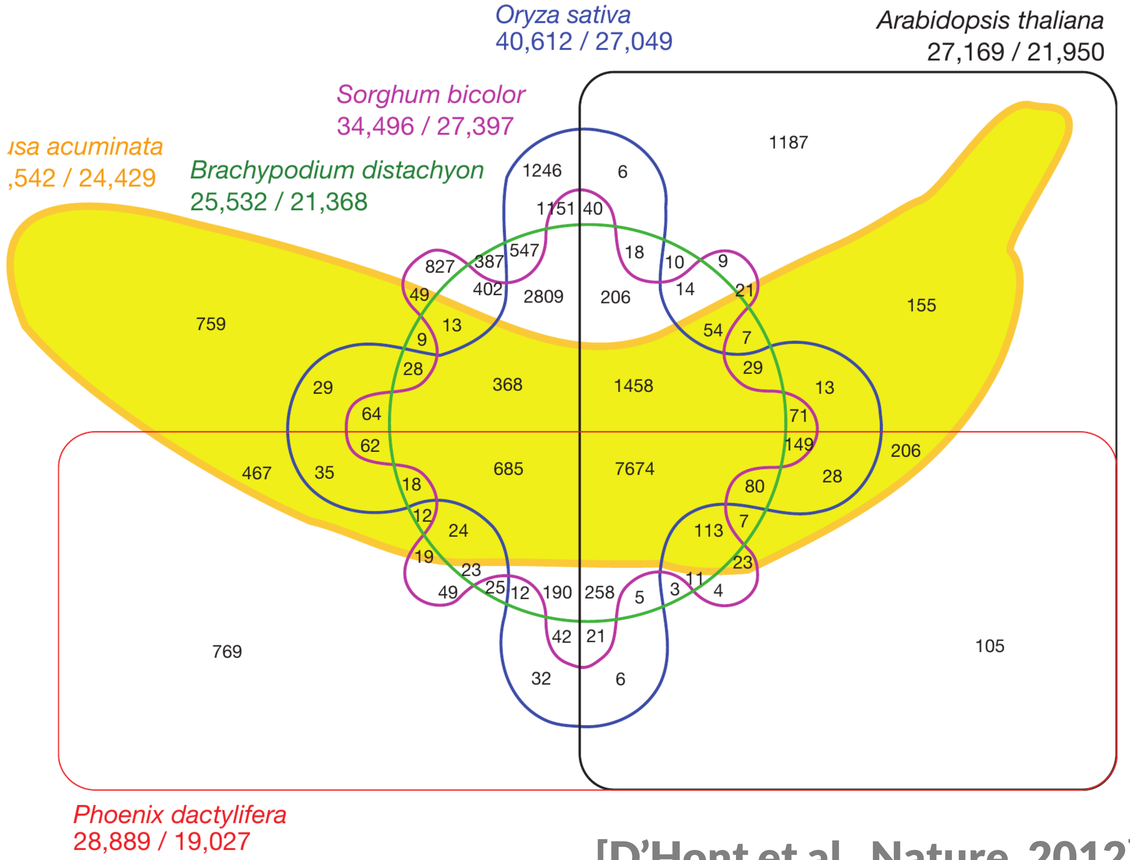
[D'Hont et al., Nature, 2012]

[Neale et al., BMC Genome Biology, 2014]

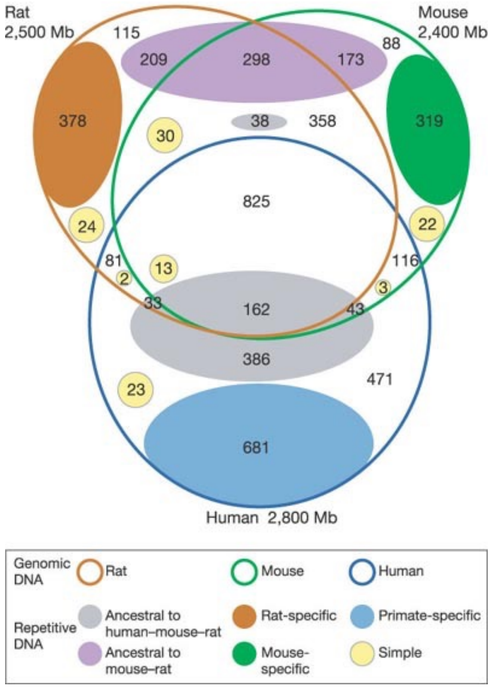




[Wiles et al., BMC Systems Biology]

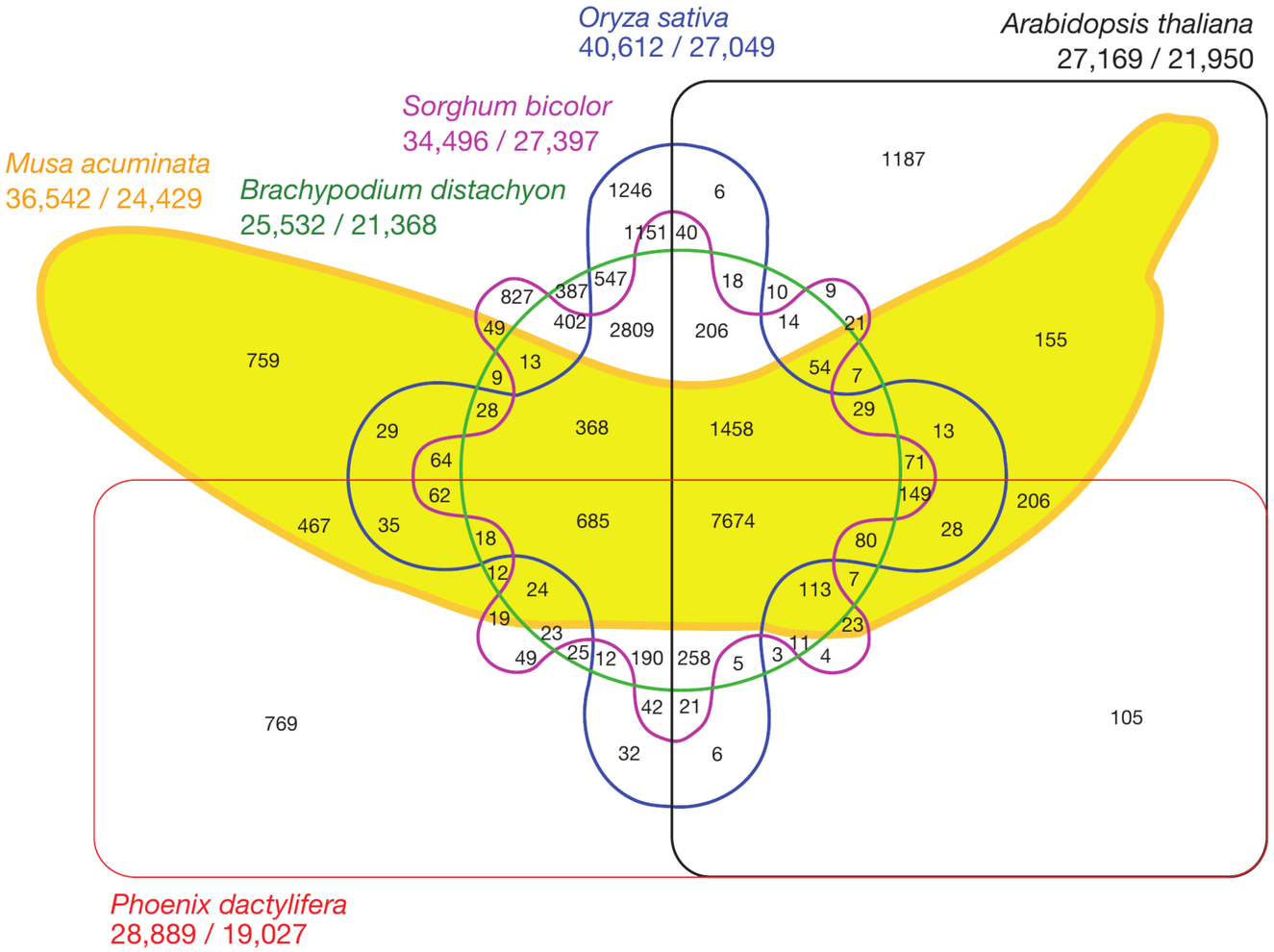


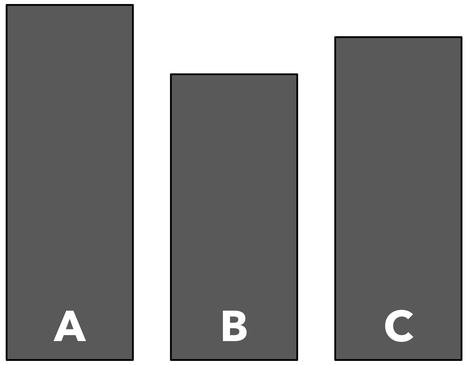
[D'Hont et al., Nature, 2012]

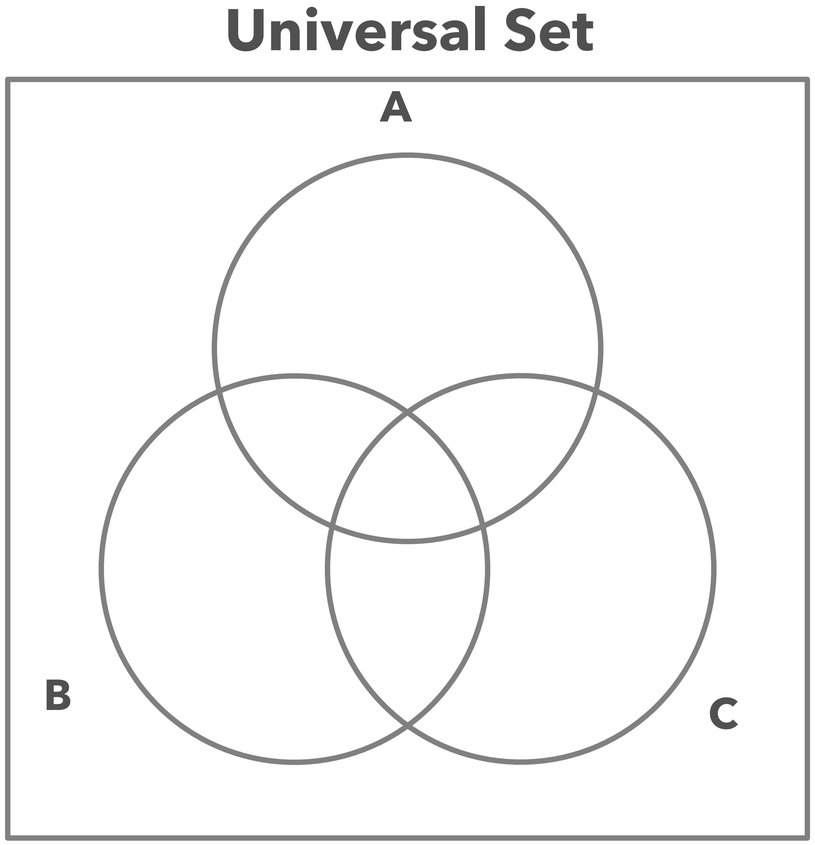


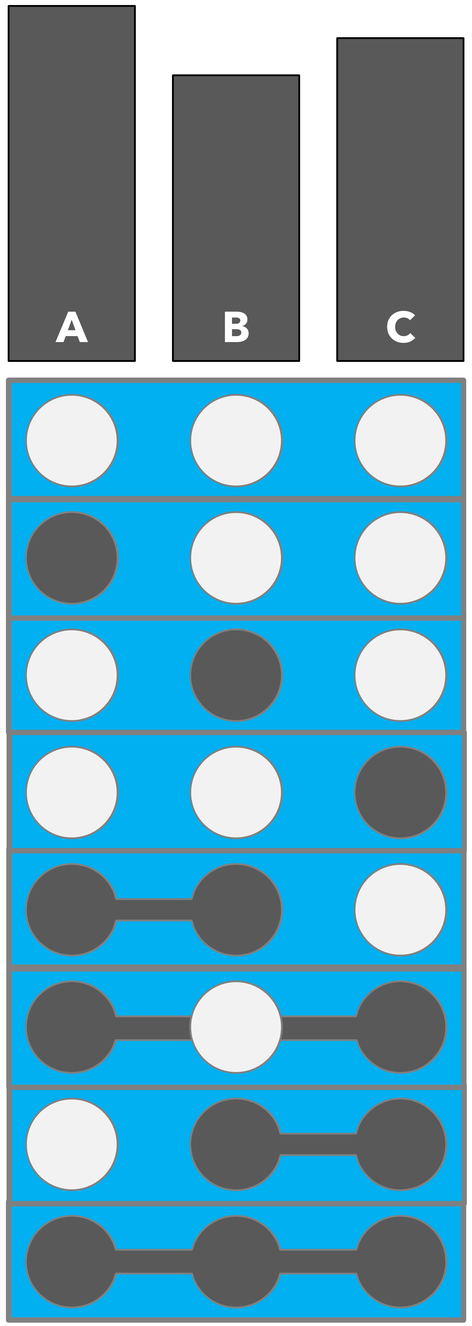
[Gibbs et al., Nature, 2004]

SO CAN WE DO BETTER?

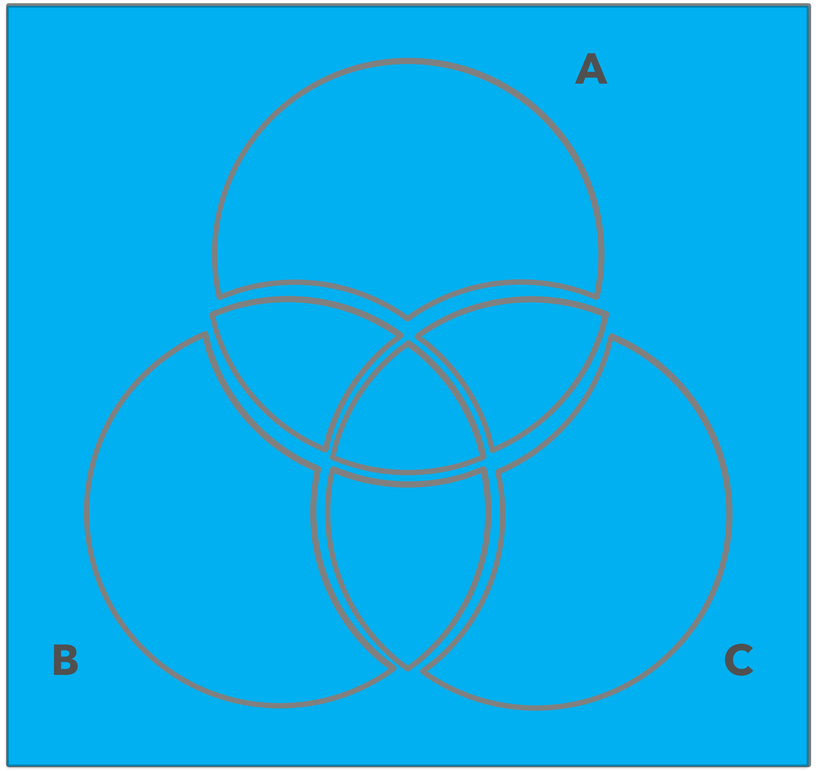


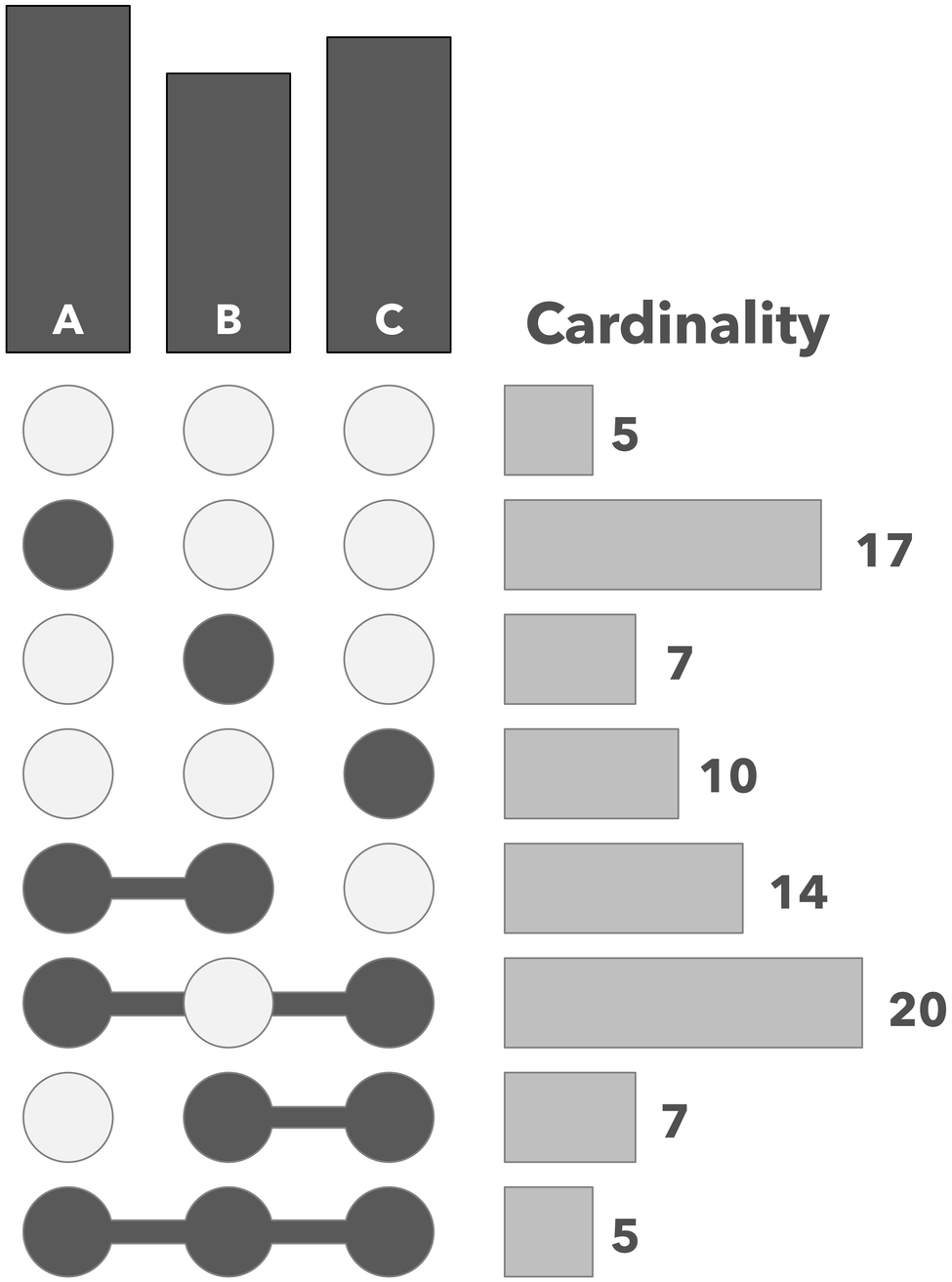

A
B
C

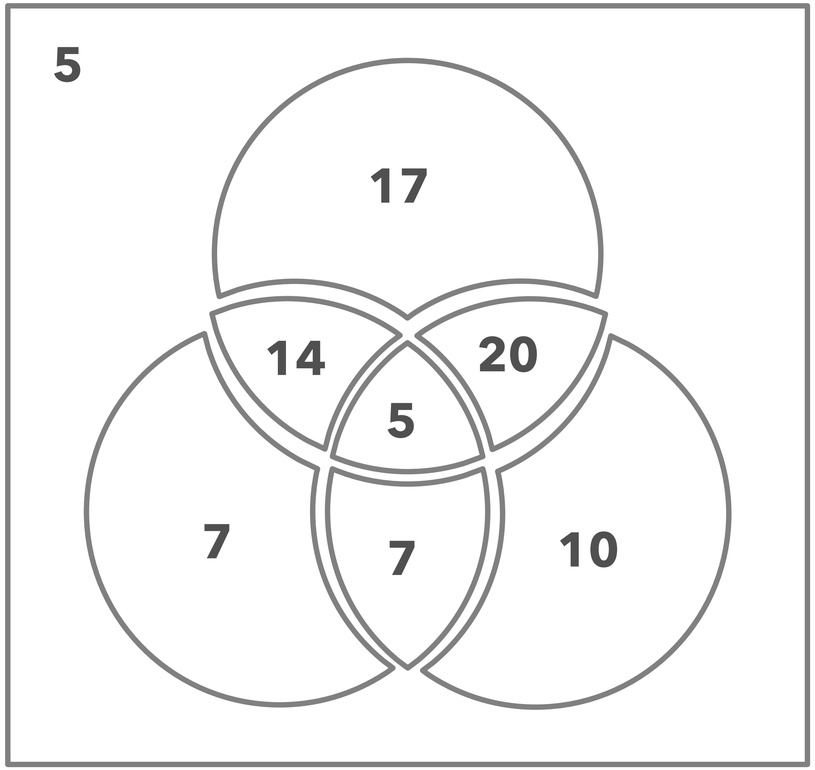

Universal Set
A
B
C



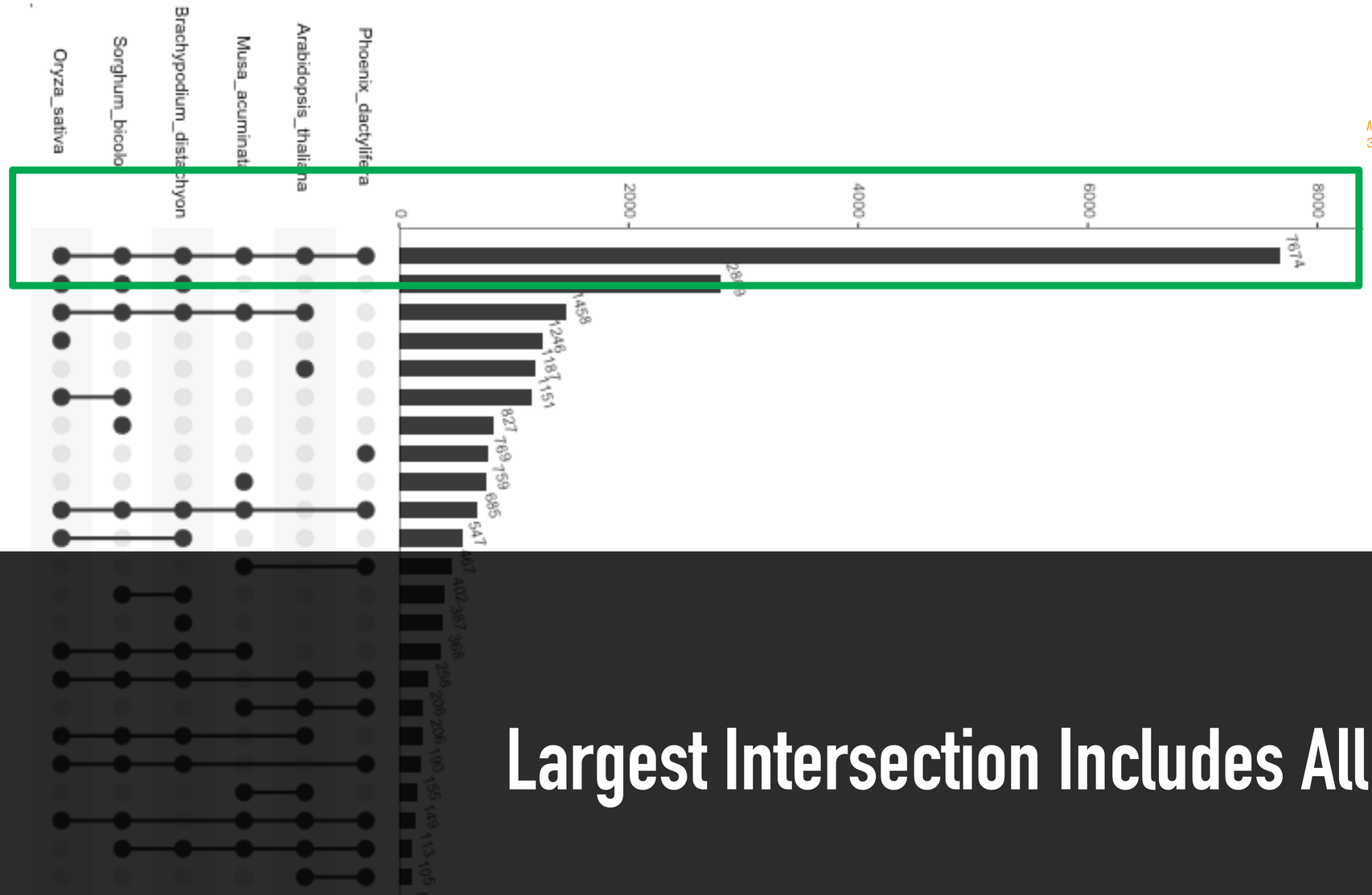
Universal Set

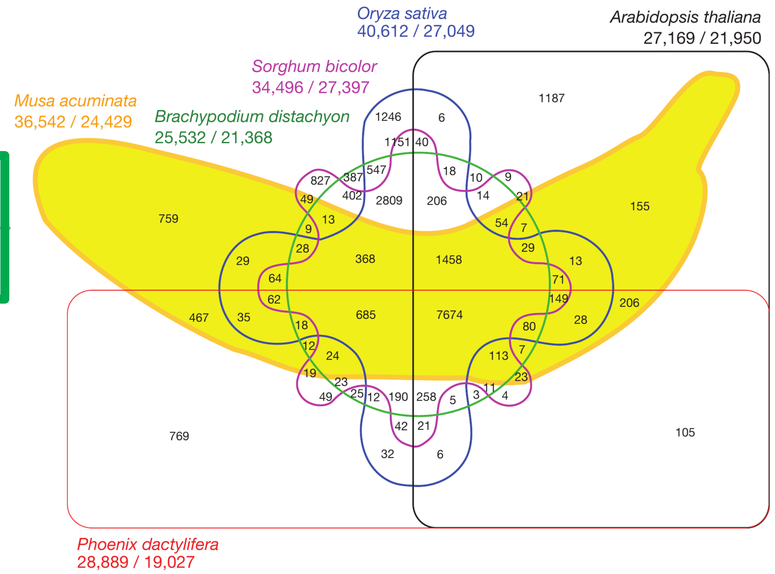



A
B
C
Cardinality
5
17
7
10
14
20
7
5


5
17
14
20
5
7
7
10

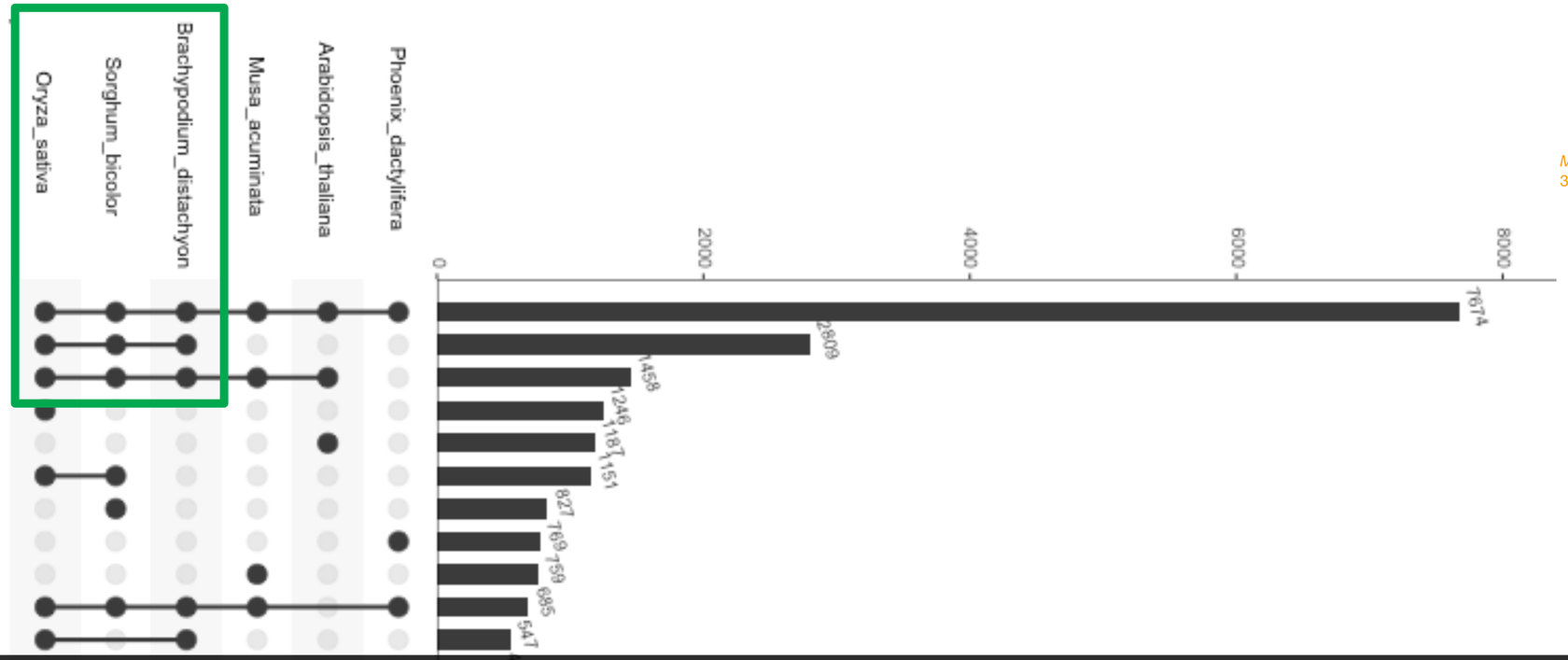
THE BANANA CHART REDESIGNED: UPSET

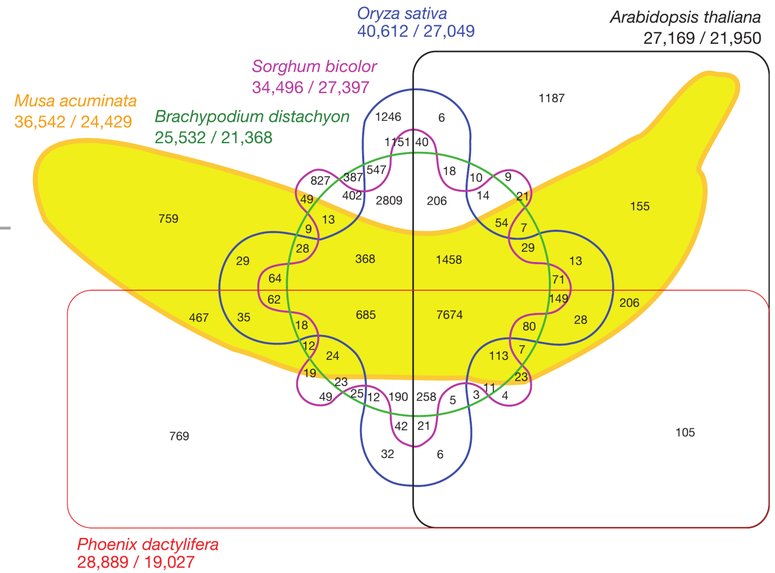




Largest Intersection Includes All Sets

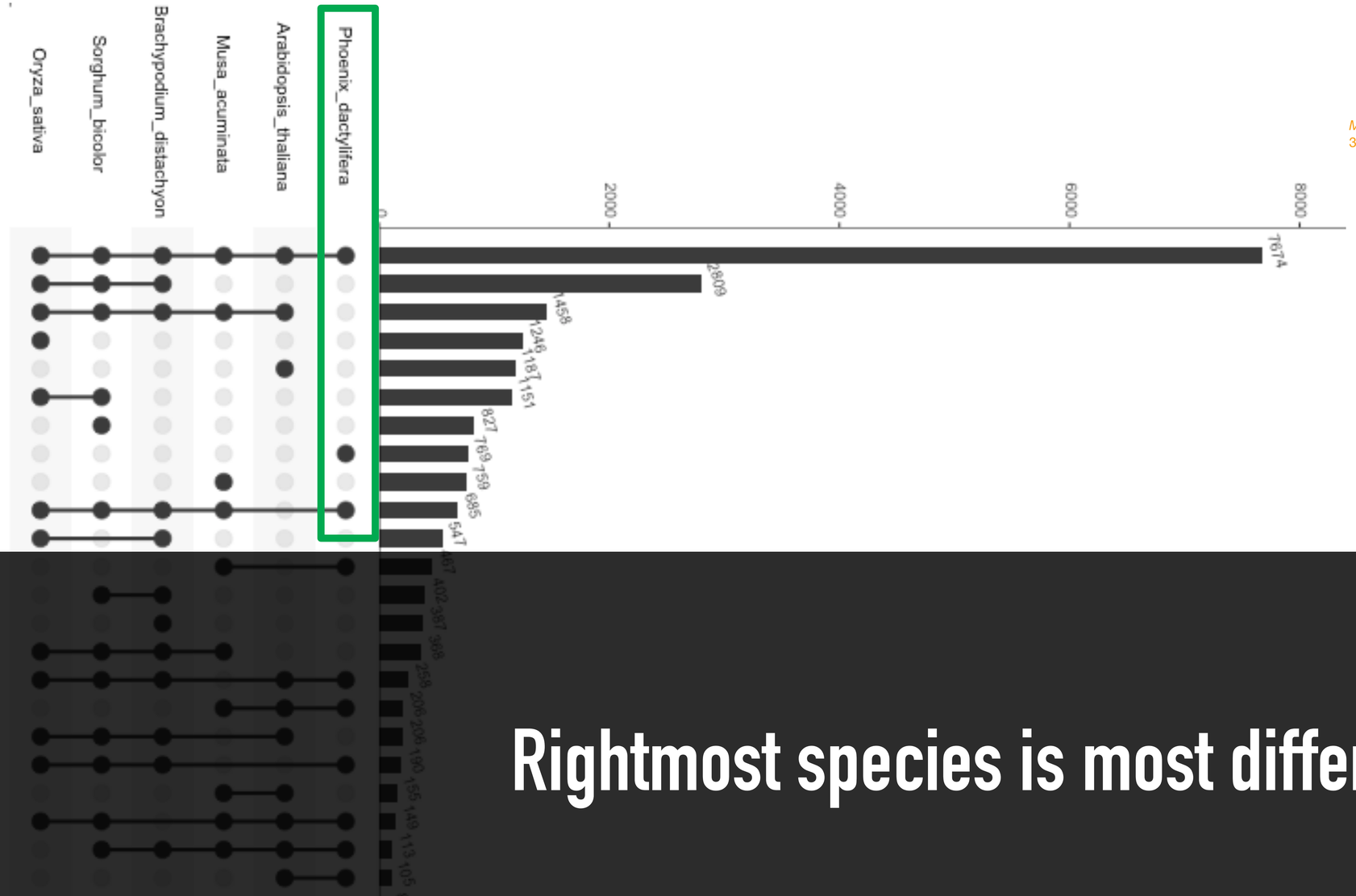
THE BANANA CHART REDESIGNED: UPSET

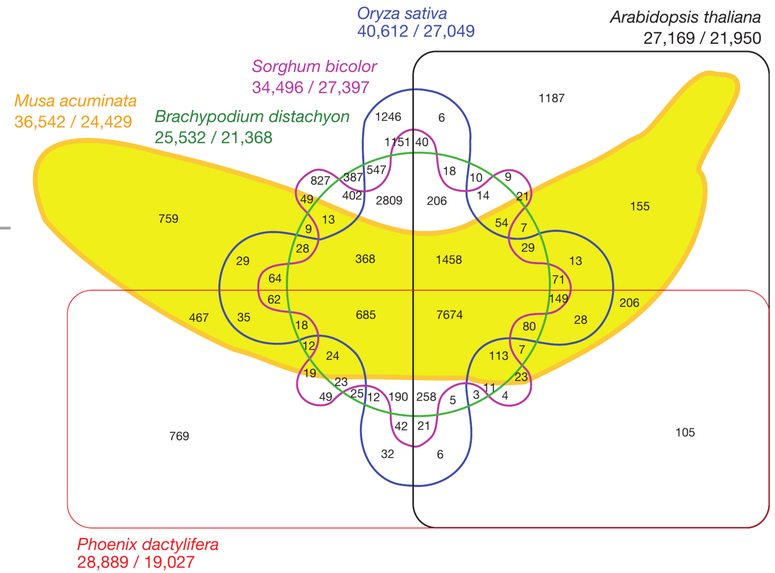




Three Leftmost Species Are Most Similar

THE BANANA CHART REDESIGNED: UPSET





Rightmost species is most different

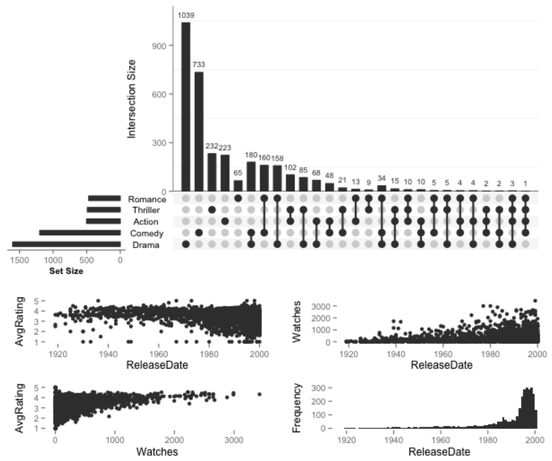
http://vcg.github.io/upset/

UPSET

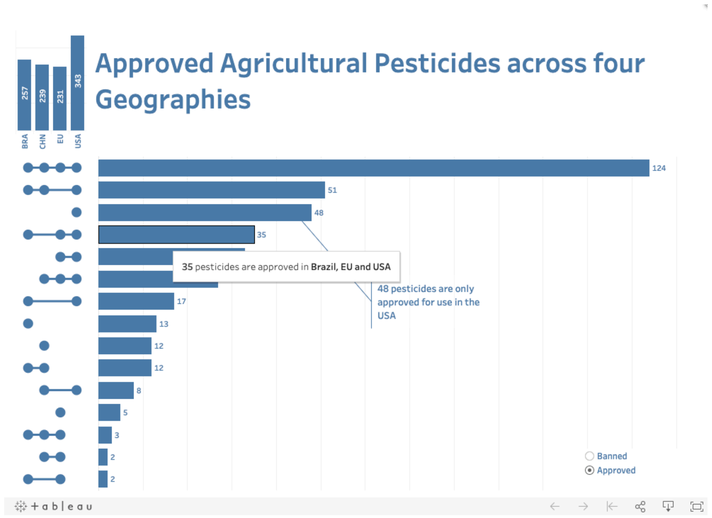
The canonical way to show set data with > 3 sets

Second-most cited VIS paper of the last decade

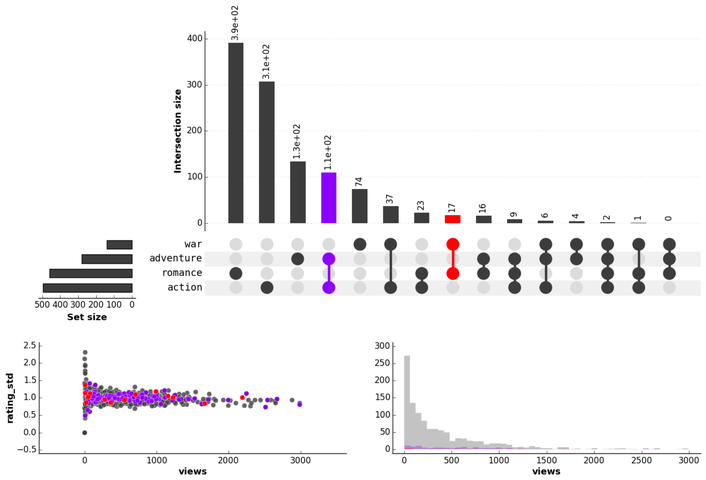
Multiple implementations in various languages



R



Tableau



Python

RESEARCH AREAS

TECHNICAL CONTRIBUTIONS

Novel Visualization Techniques

Visualization Process Innovations

Data Wrangling Methods

DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

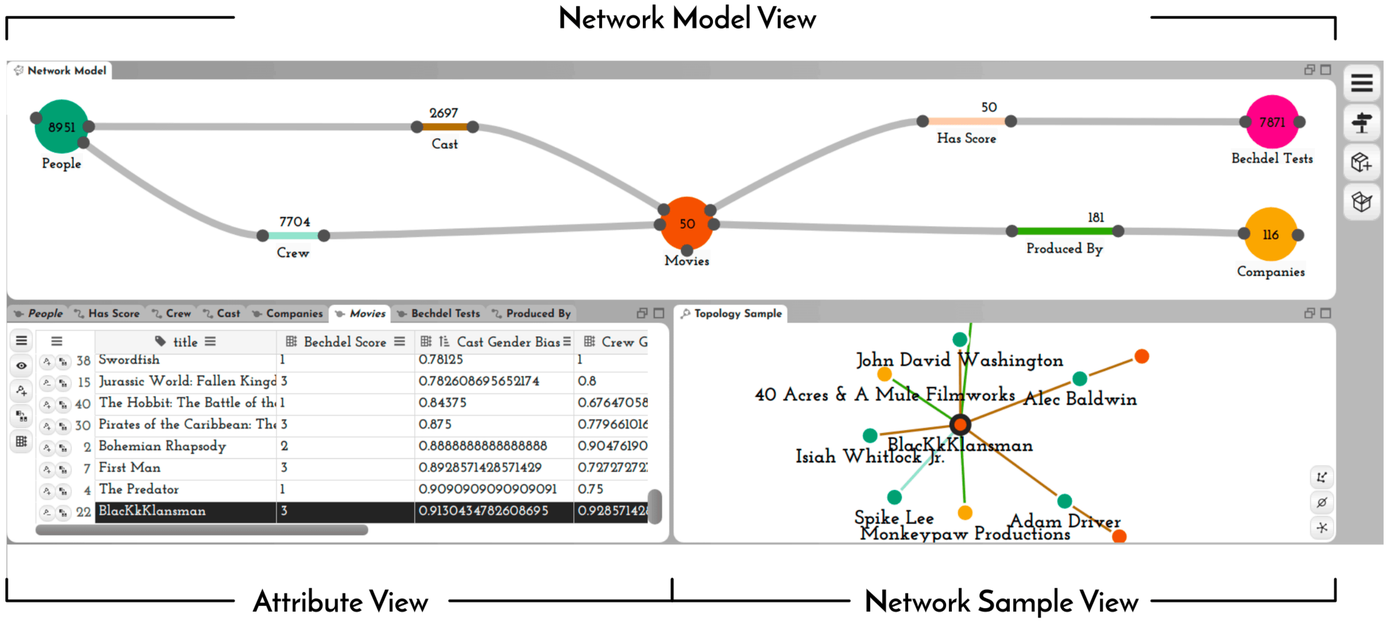
EMPIRICAL & THEORETICAL WORK

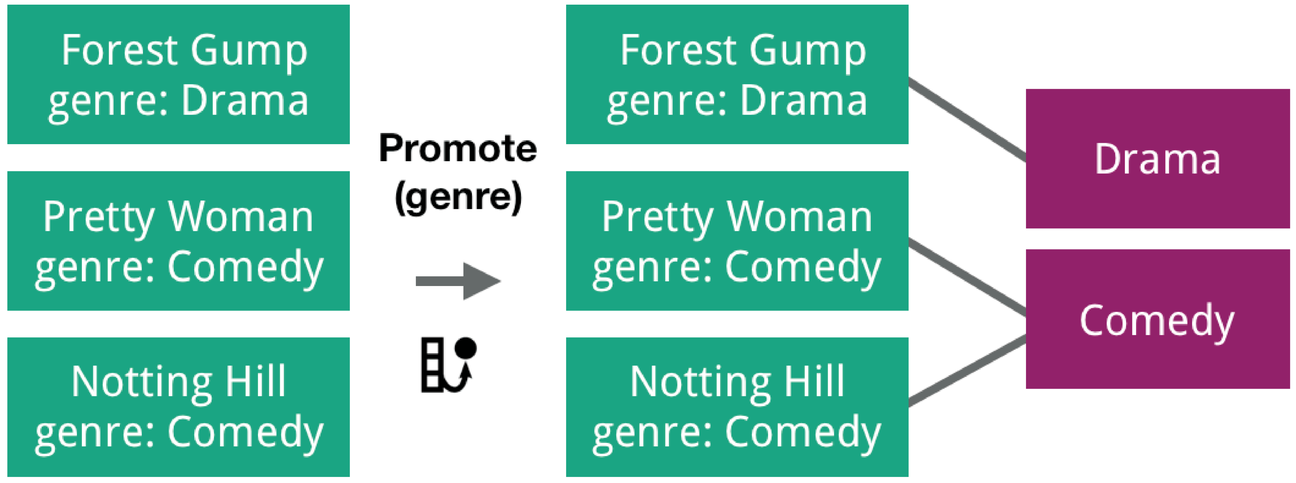
Evaluation Methodology

Design Spaces / Taxonomies

Reshaping Networks

Data Wrangling Methods





TECHNICAL CONTRIBUTIONS

Data Wrangling Methods

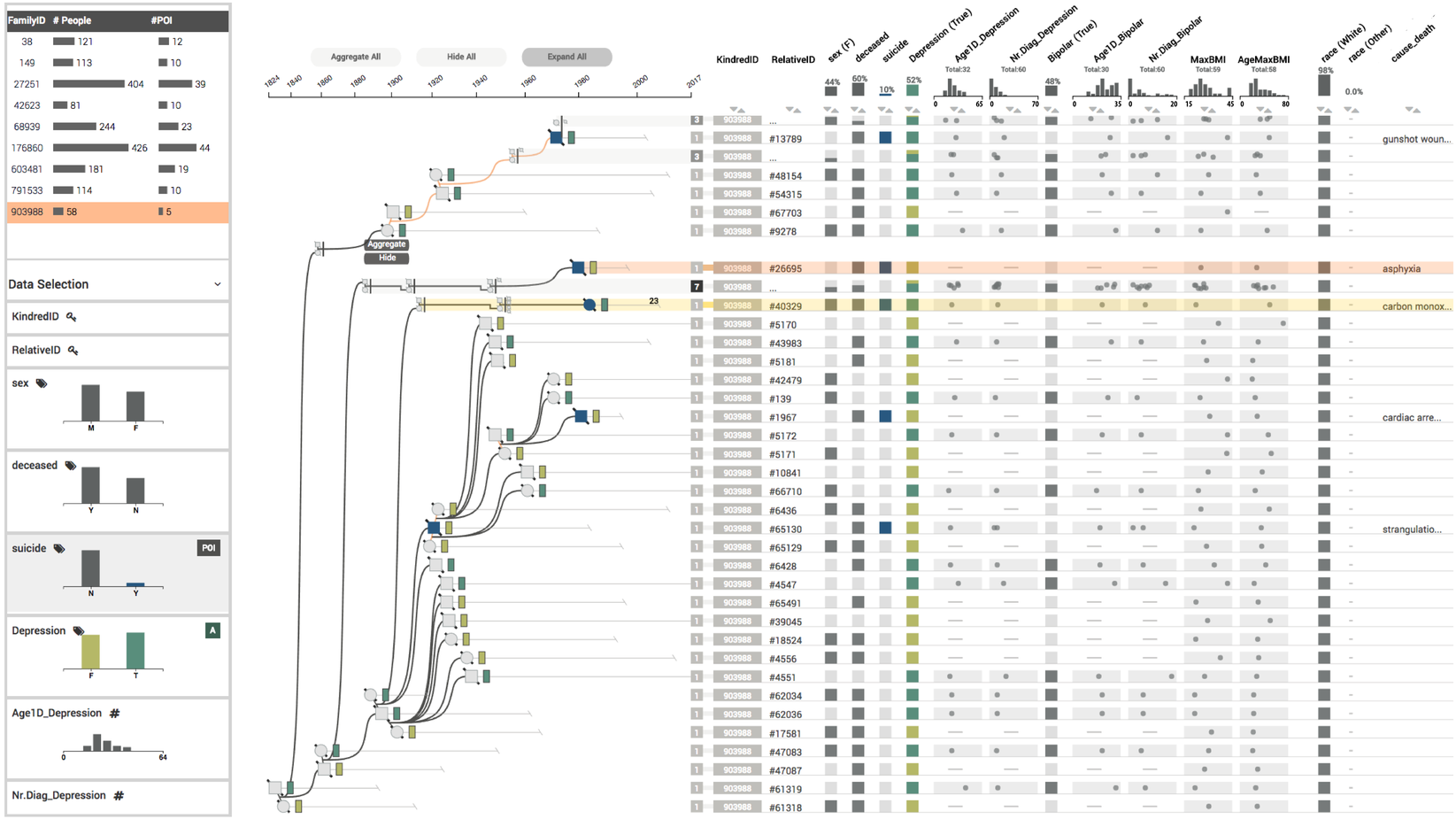
DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

DOMAIN DRIVEN TECHNIQUES

Genealogies for Clinical Data Analysis

Tailored Methods and Systems for High Impact Science Problems



DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

EMPIRICAL & THEORETICAL WORK

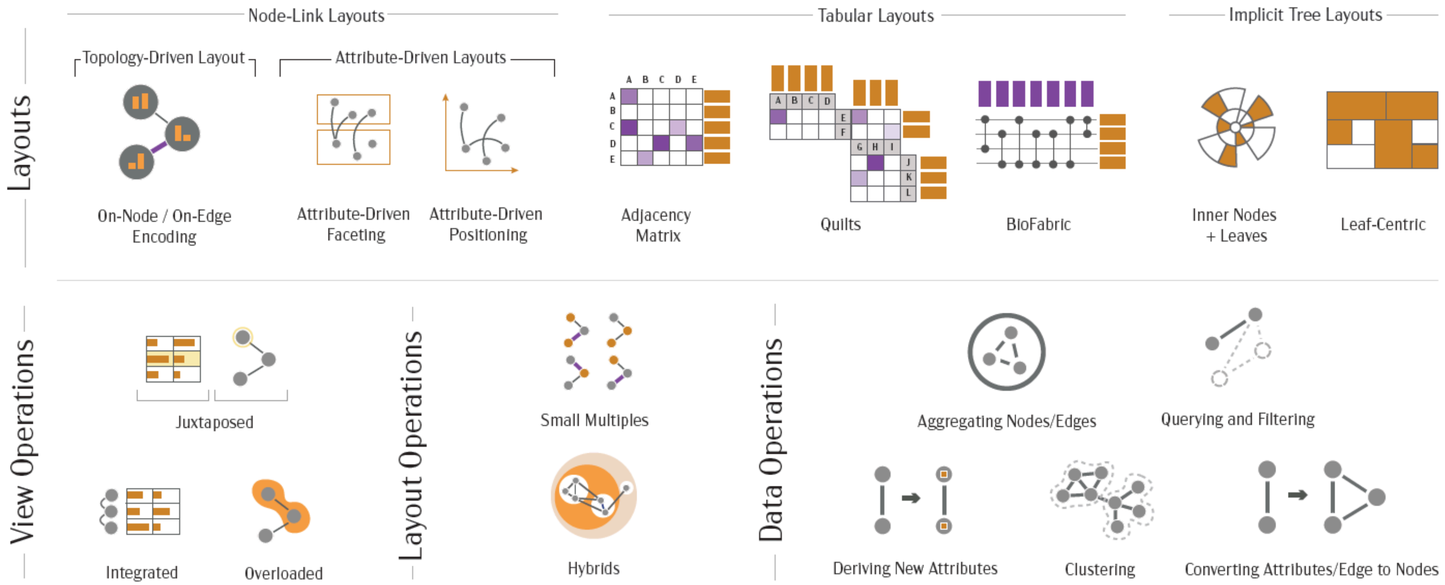
Evaluation Methodology

Design Spaces / Taxonomies

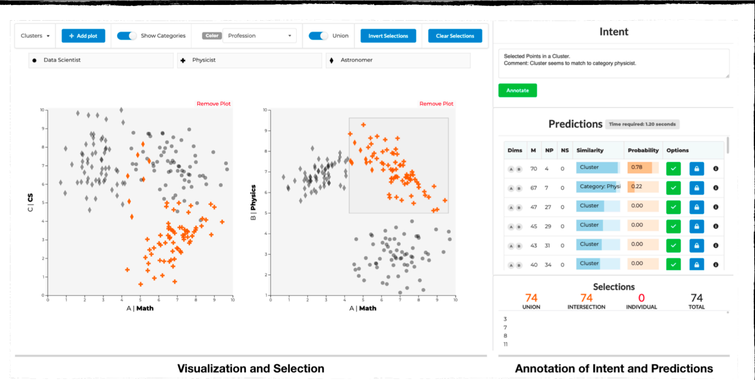
EMPIRICAL & THEORETICAL WORK

Evaluation Methodology

Design Spaces / Taxonomies



TECHNICAL CONTRIBUTIONS



LITERATE VISUAL DATA ANALYSIS:

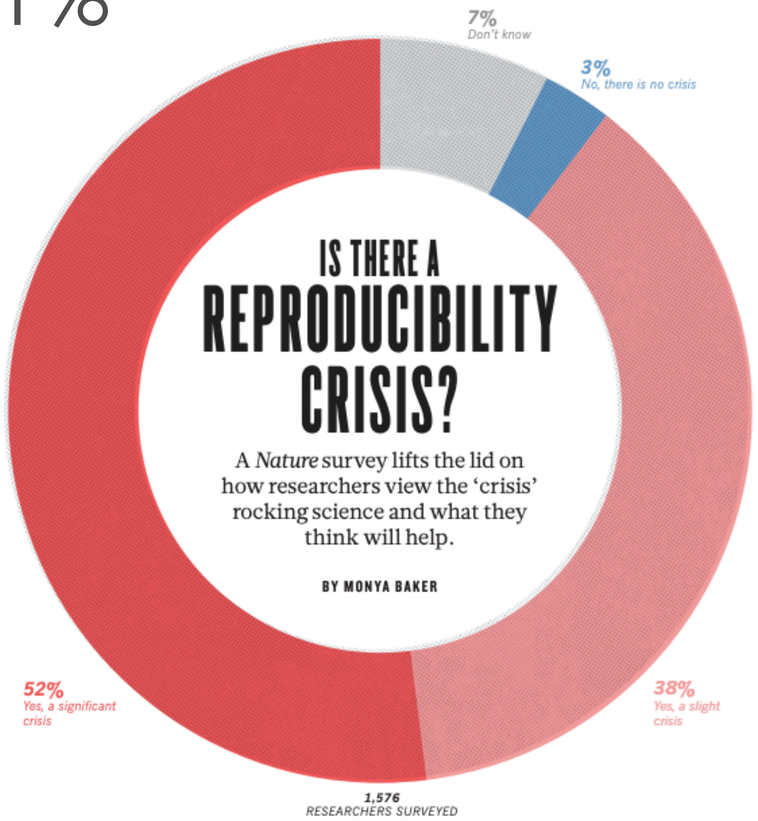
MAKING VISUAL ANALYSIS SESSIONS REPRODUCIBLE AND REUSABLE

REPRODUCIBILITY CRISIS IN SCIENCE

Reproducibility Rates

Psychology: 40%

Cancer Biology: 11%



[Baker, Penny, 2016]
[Open Science Collaboration, 2016]
[Begley, Ellis, 2012]

WHY IS THERE A REPRODUCIBILITY CRISIS?

Perverse incentives (publish or perish)

Bias for “flashy” results

Problems in data analysis

Lack of data sharing

Lack of sharing the analysis process

…

TACKLE PROBLEMS IN DATA ANALYSIS AND ENHANCE REPRODUCIBILITY: LITERATE PROGRAMMING

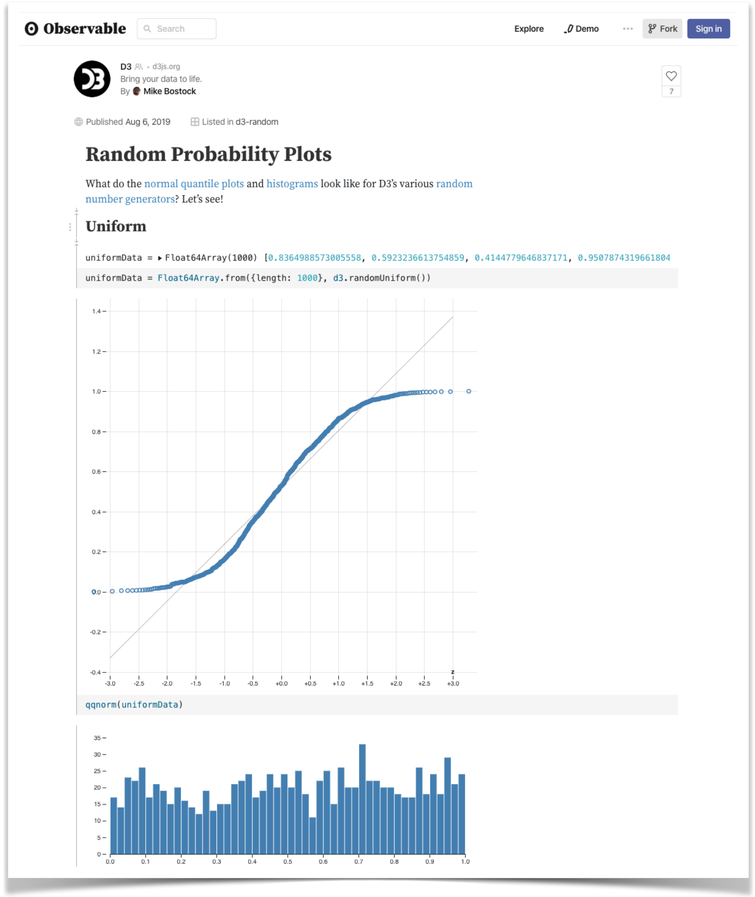
LITERATE PROGRAMMING

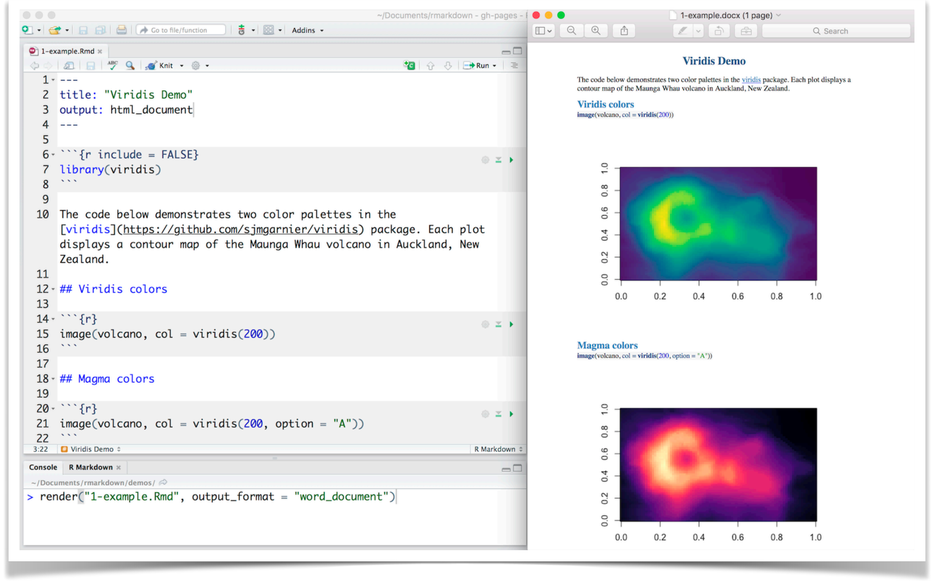
Explain the why and how using any means necessary!

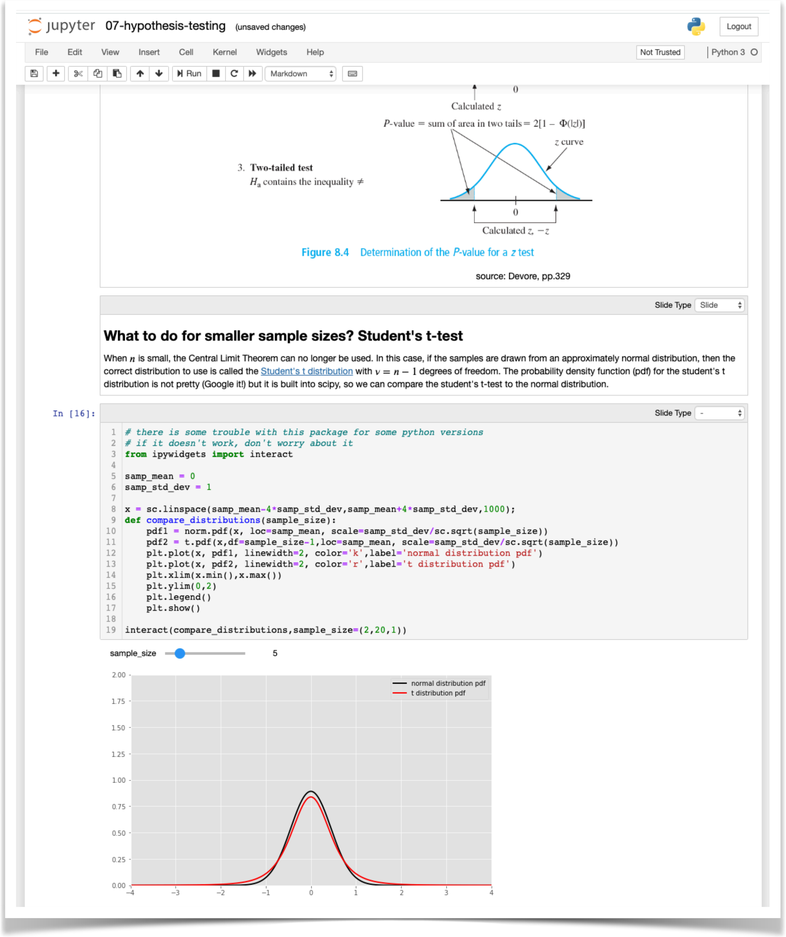
- Text
- Images / Visualizations
- Formulas
- Videos
- Links
- Code

[Donald E. Knuth, 1984]

LITERATE PROGRAMMING IN THE WILD







Observable

R Markdown

Jupyter Notebooks

CHARACTERISTICS OF A GOOD NOTEBOOK

Carefully curated, well narrated.

Clean code for readability

Illustrations, Formulas, Visualizations

Complex, multi-stage process

THERE IS NO STRAIGHTFORWARD WAY TO DO

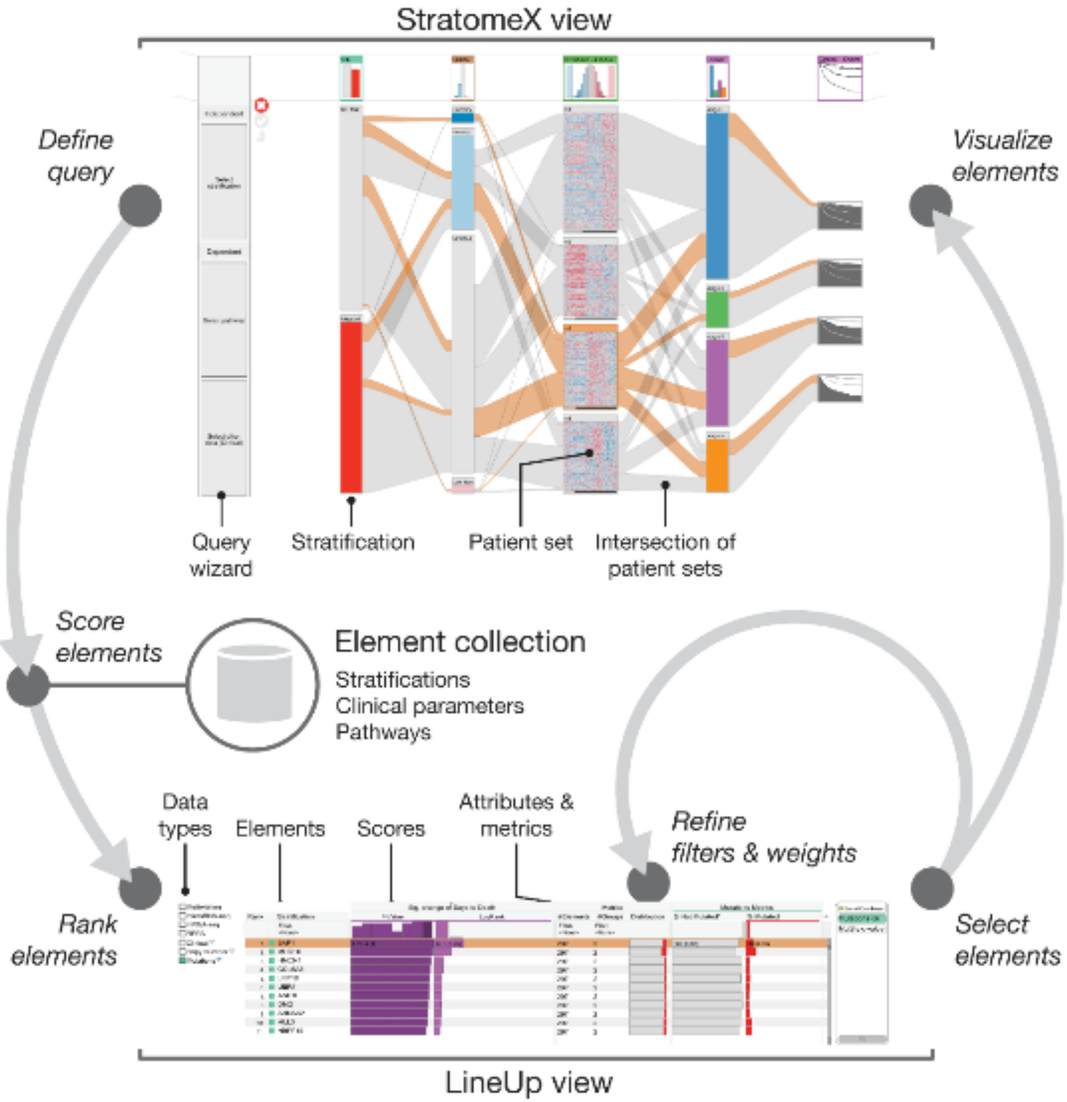
LITERATE DATA VISUALIZATION

LITERATE VISUAL DATA ANALYSIS

Current State:
no record of what was done,
let alone why

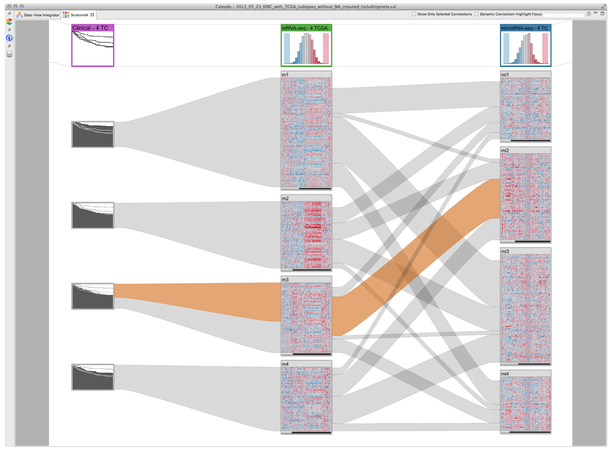
Idea: make the process of an interactive, visual analysis session well reasoned and documented

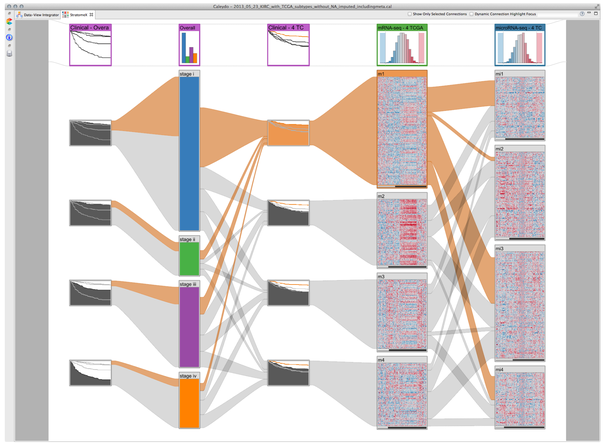
A MANUAL ATTEMPT AT LITERATE VISUALIZATION

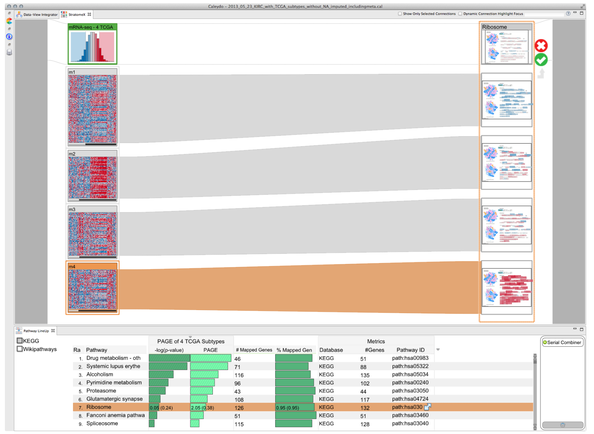


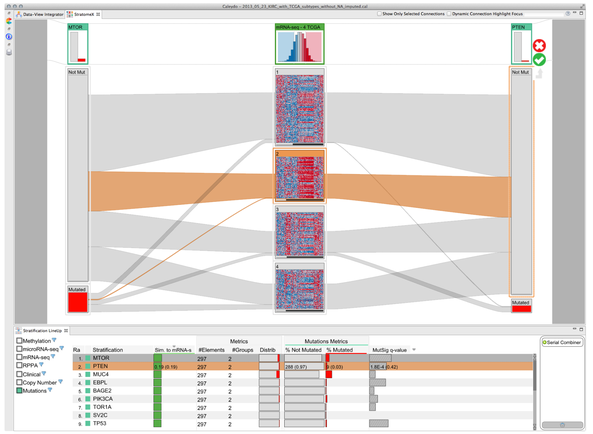
Guided Visual Exploration of Genomic Stratifications in Cancer

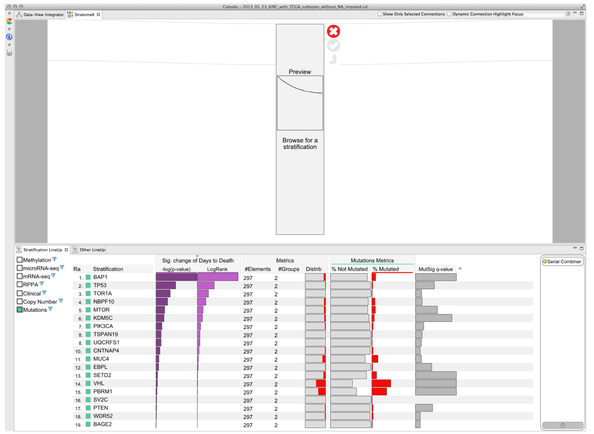
Nature Methods 11, 9 (2014), 884–885

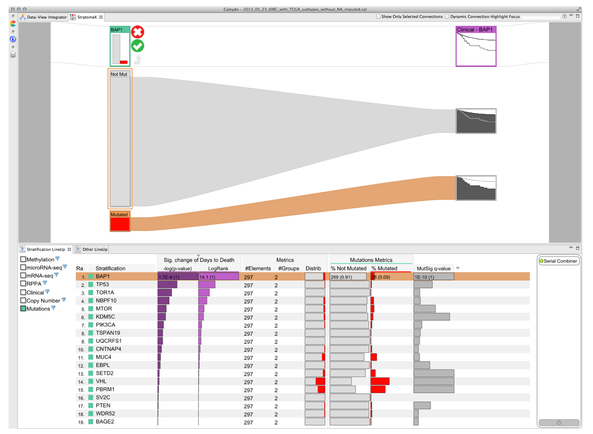


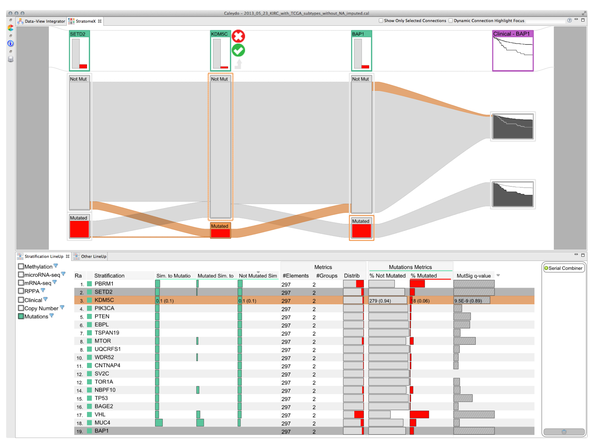


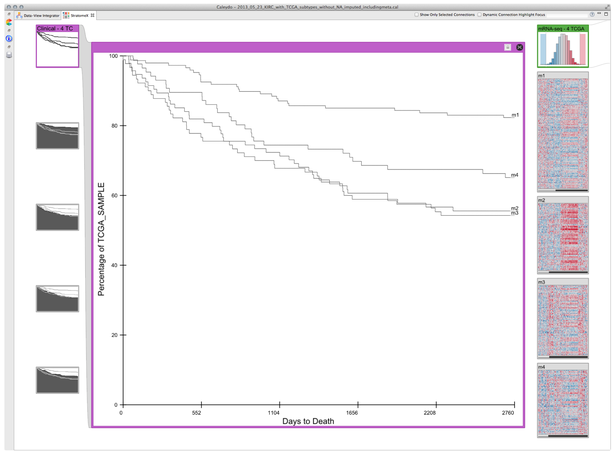












CASE STUDY

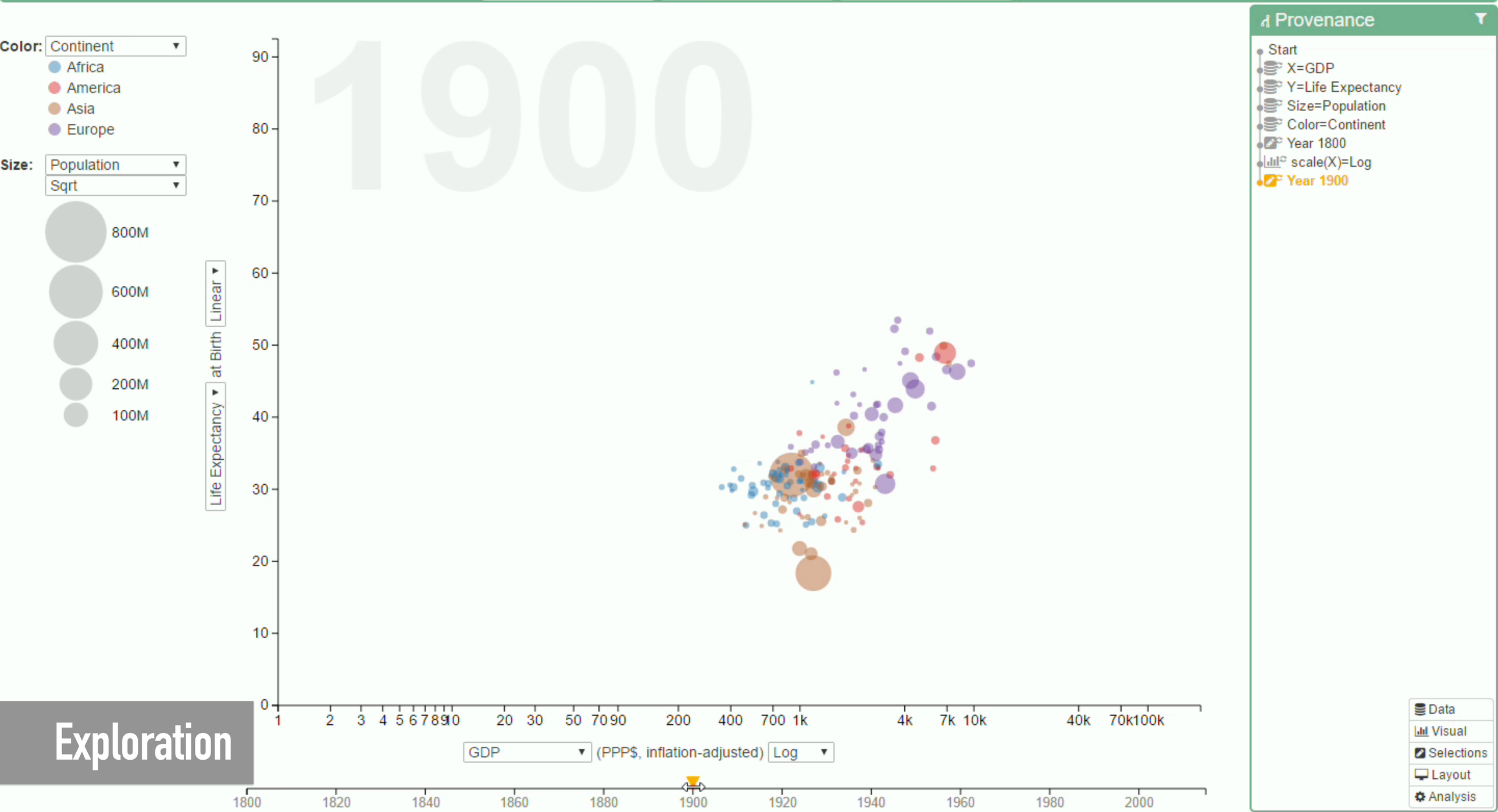
Streit et al., Guided Visual Exploration of Genomic Stratifications in Cancer, Nature Methods, 2014

IDEA: USE ANALYSIS SESSION PROVENANCE AS BASIS FOR A DATA STORY

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

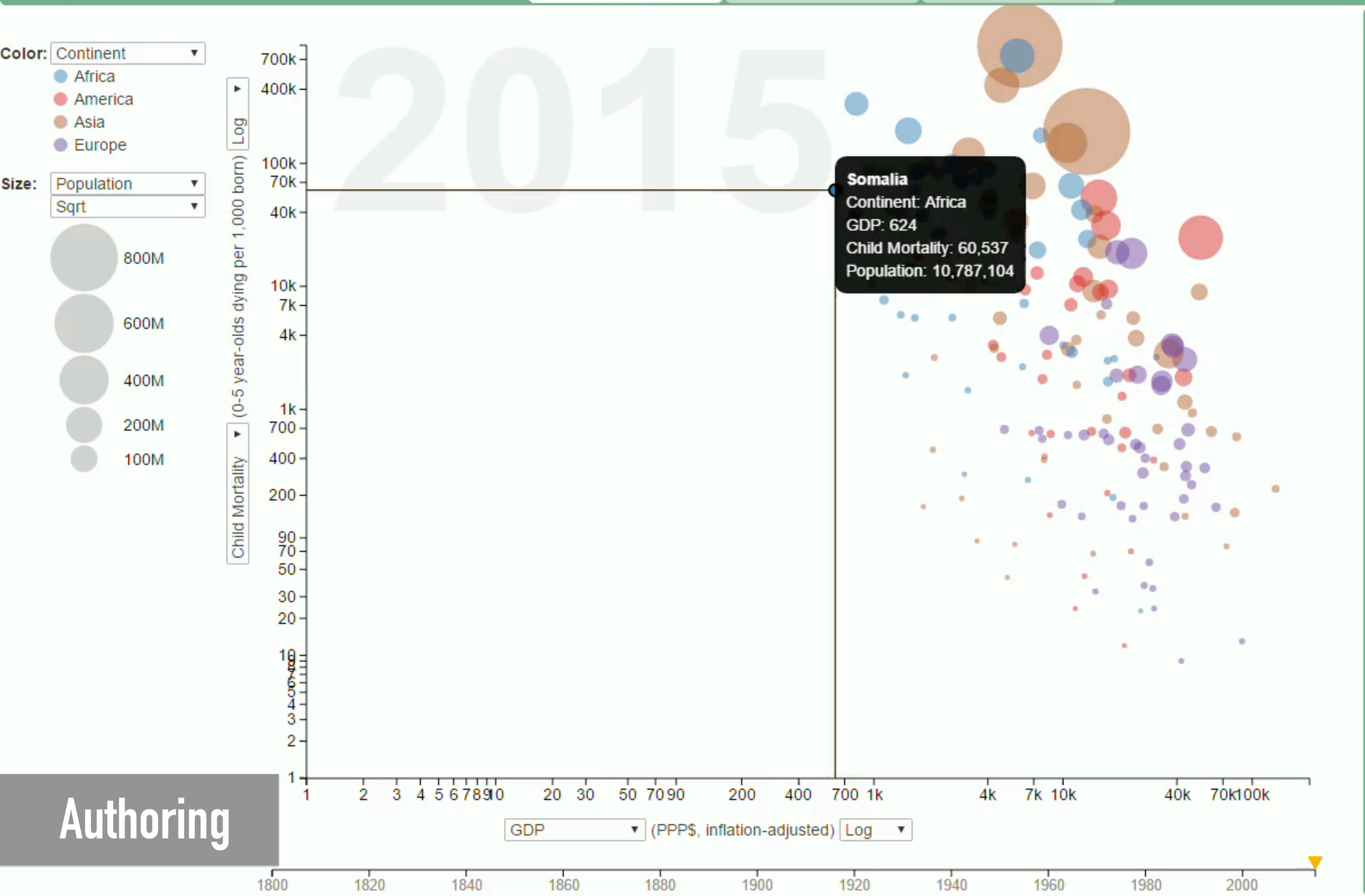


Provenance

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

Data
Visual
Selections
Layout
Analysis

Exploration

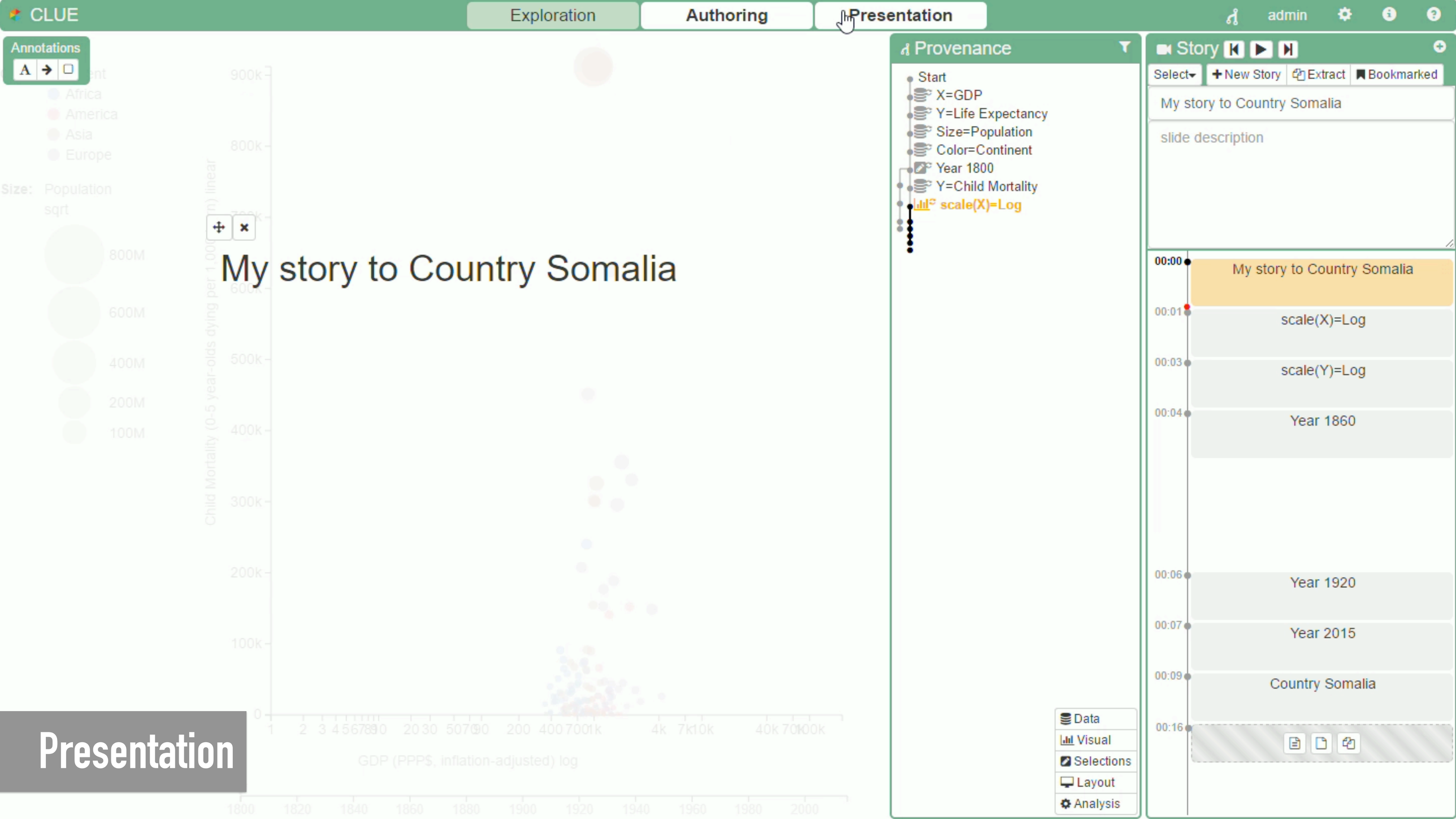


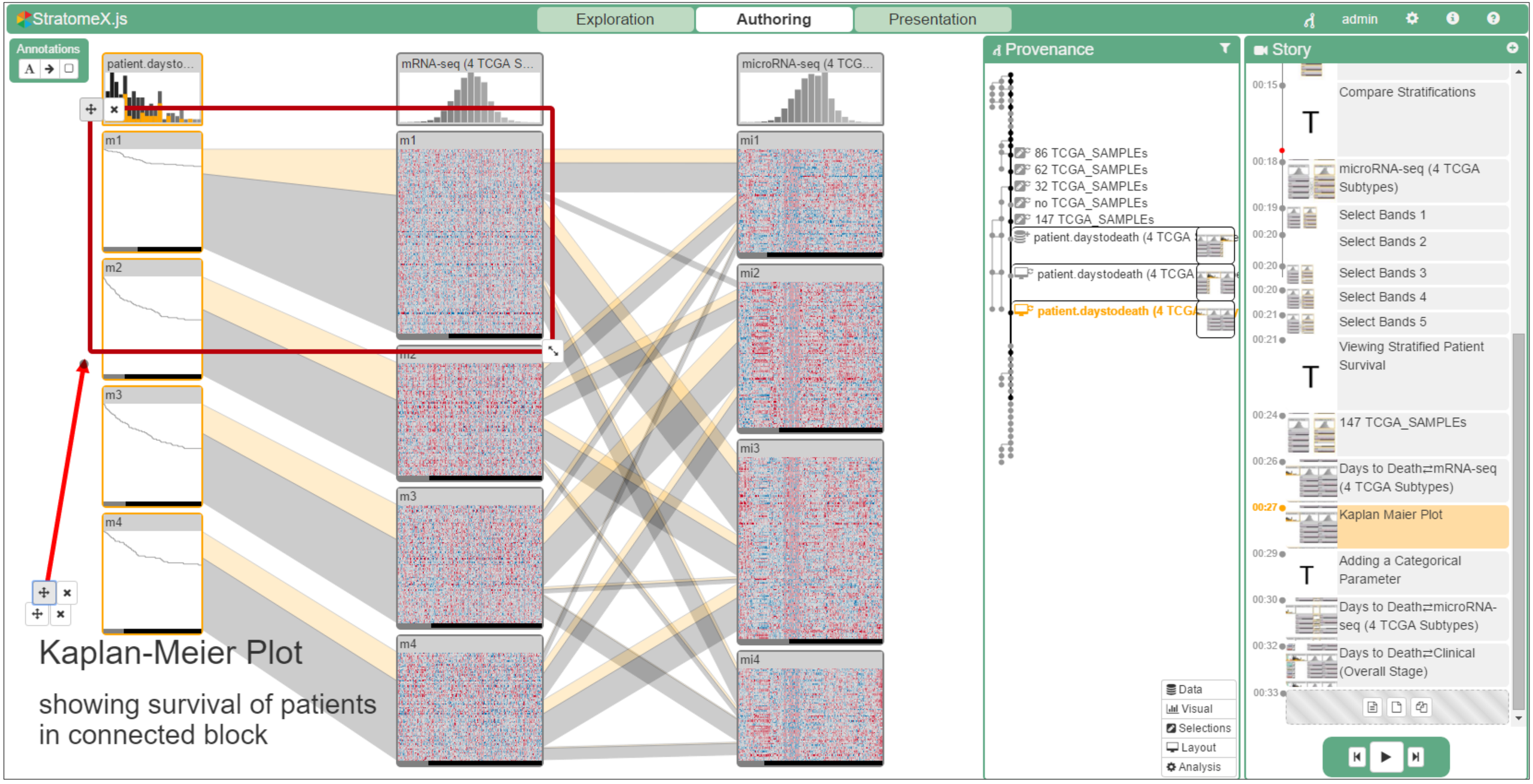
Provenance

- Year 1800
- Y=Child Mortality
- scale(X)=Log
- scale(Y)=Log
- Year 1860
- Year 1920
- Year 2015
- Country Somalia

Data
Visual
Selections
Layout
Analysis

Authoring


CLUE
Exploration
Authoring
Presentation
admin
Annotations
Provenance
Start
X=GDP
Y=Life Expectancy
Size=Population
Color=Continent
Year 1800
Y=Child Mortality
scale(X)=Log
Story
Select
New Story
Extract
Bookmarked
My story to Country Somalia
slide description
My story to Country Somalia
00:00
My story to Country Somalia
00:01
scale(X)=Log
00:03
scale(Y)=Log
00:04
Year 1860
00:06
Year 1920
00:07
Year 2015
00:09
Country Somalia
00:16
Data
Visual
Selections
Layout
Analysis
Africa
America
Asia
Europe
Size: Population
sqrt
800M
600M
400M
200M
100M
Child Mortality (0-5 year-olds dying per 1,000 born) linear
900k
800k
600k
500k
400k
300k
200k
100k
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
My story to Country Somalia
Presentation



http://vistories.org/

PROGRESS; BUT PROVENANCE IS "DUMB"

We solved the **WHAT,**
but not the **WHY**

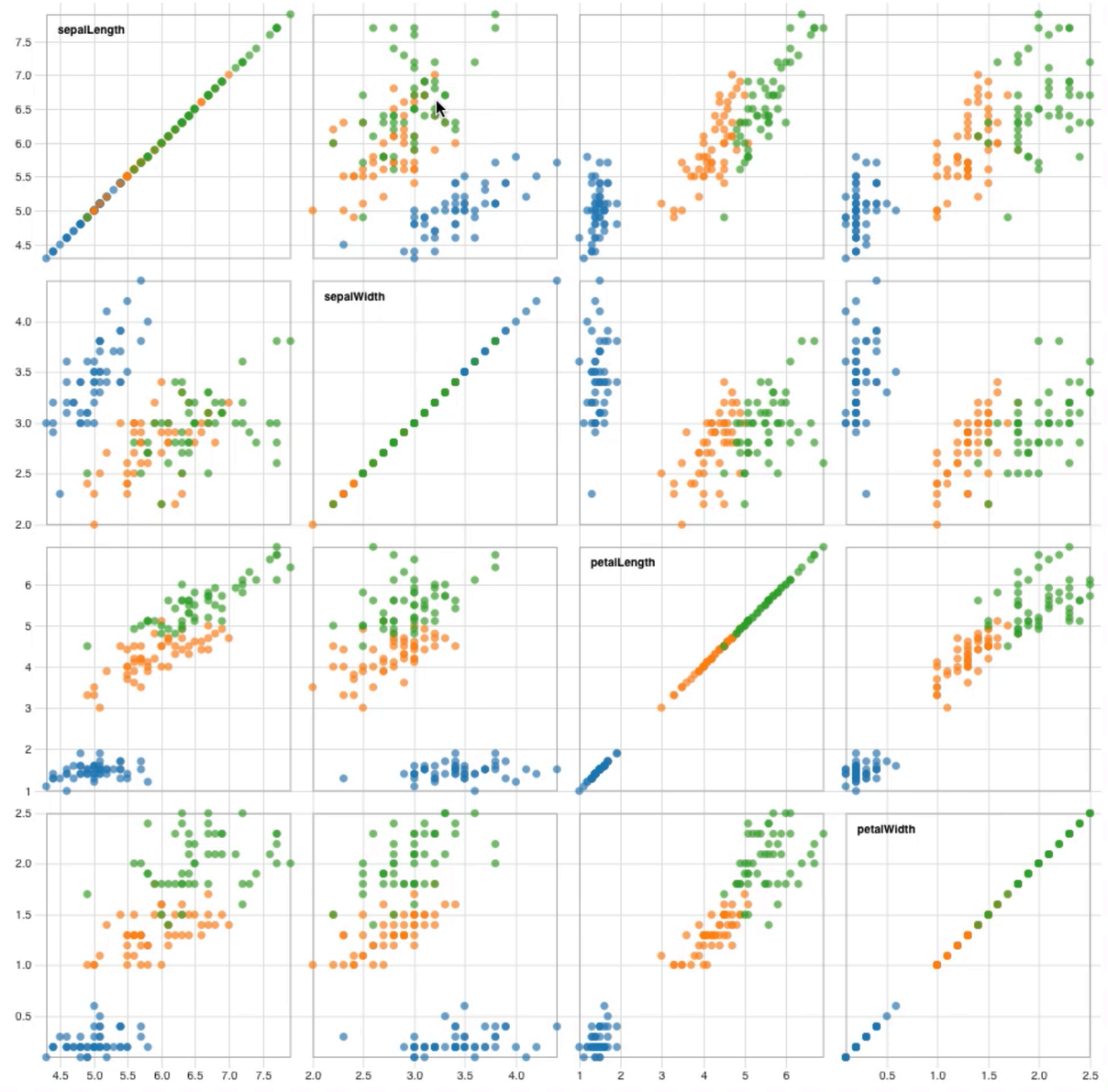
No progress towards reusability..

So, what else can we do?

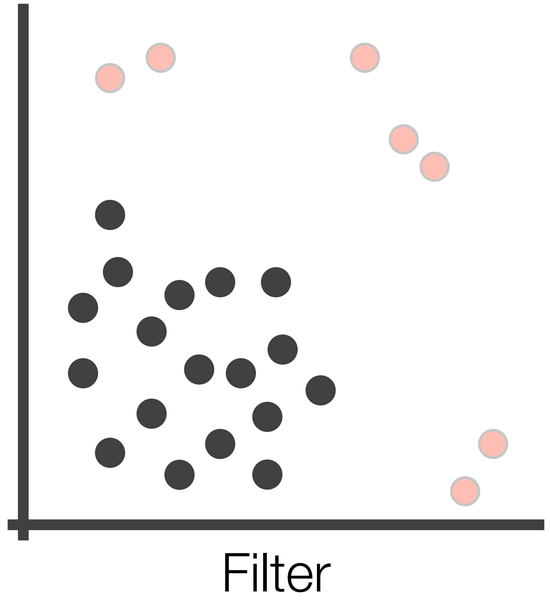
SEMANTIC SELECTIONS

Information Visualization 2021. Kiran Gadhave, Jochen Görtler, Carolina Nobre, Oliver Deussen, Miriah Meyer, Jeff Phillips, Alexander Lex

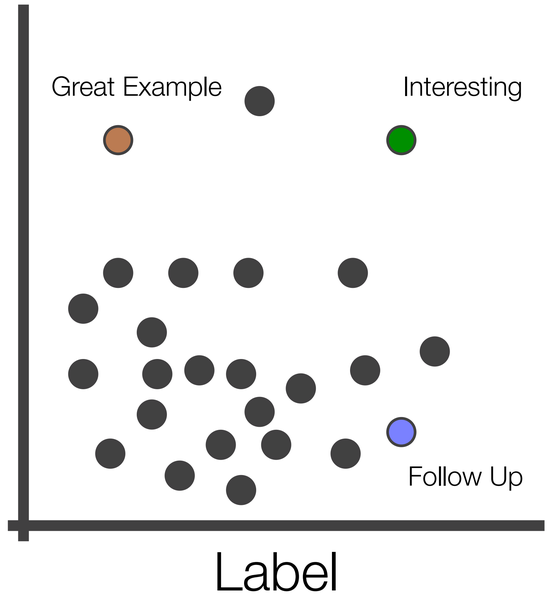
WHAT ARE SELECTIONS?



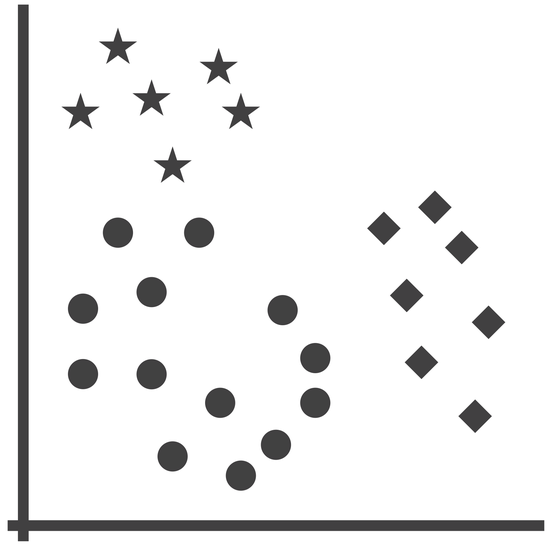
FROM SELECTIONS TO ADVANCED OPERATIONS

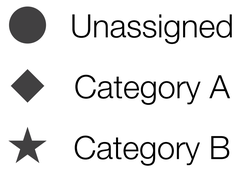


Filter

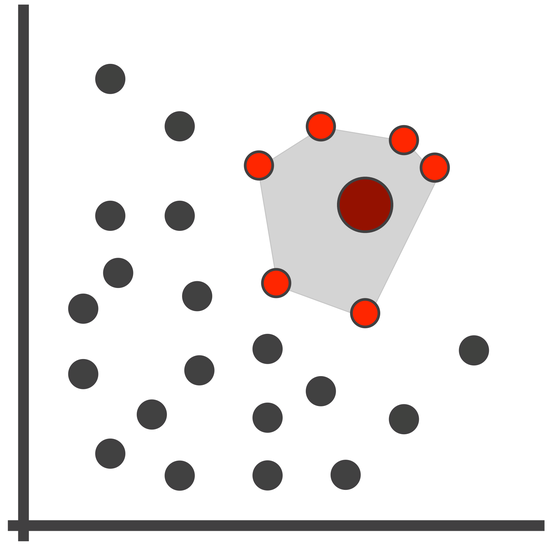


Label





Categorize



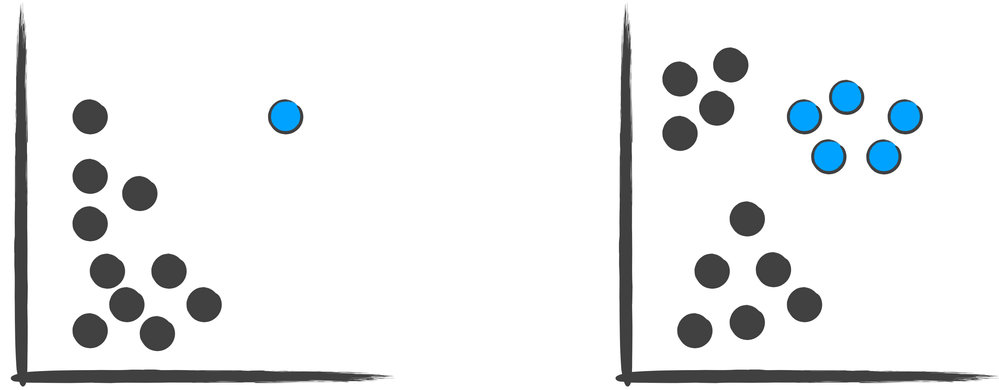
Aggregate

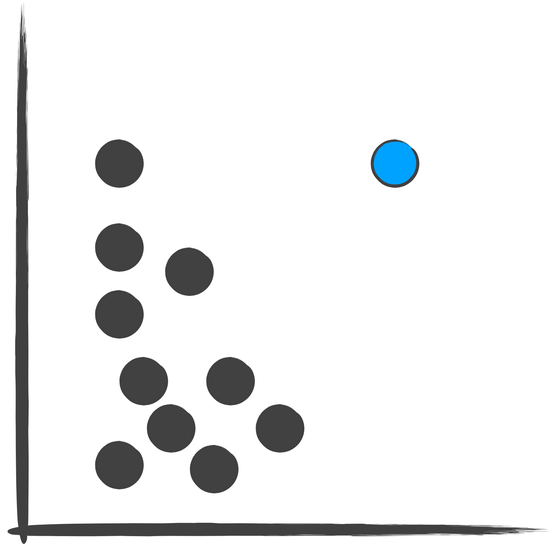
WHAT IS INTENT WHEN SELECTING?

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

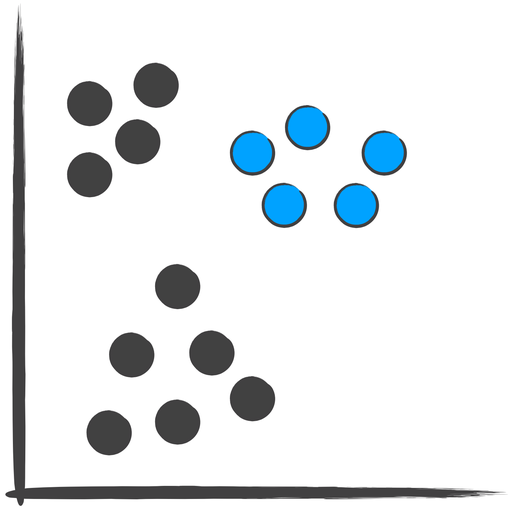
Domain Specific Intent: Capture through Annotation

Pattern-Based Intent: Capture Automatically

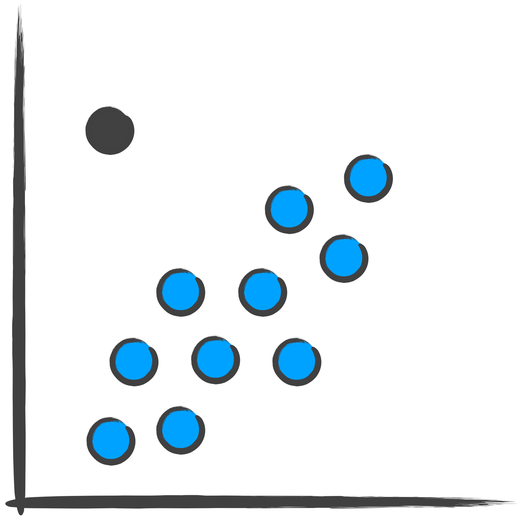




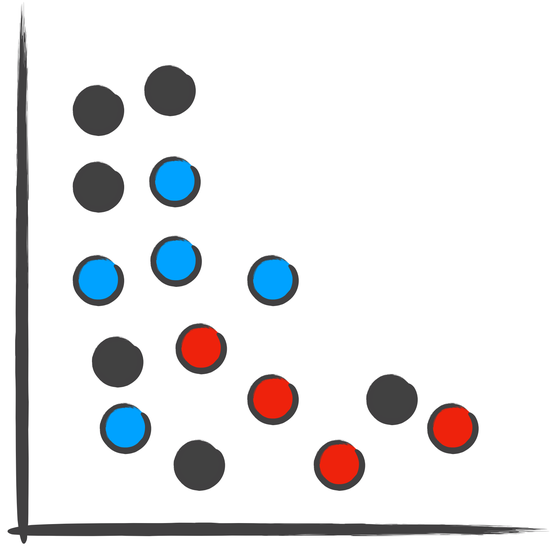
Outlier



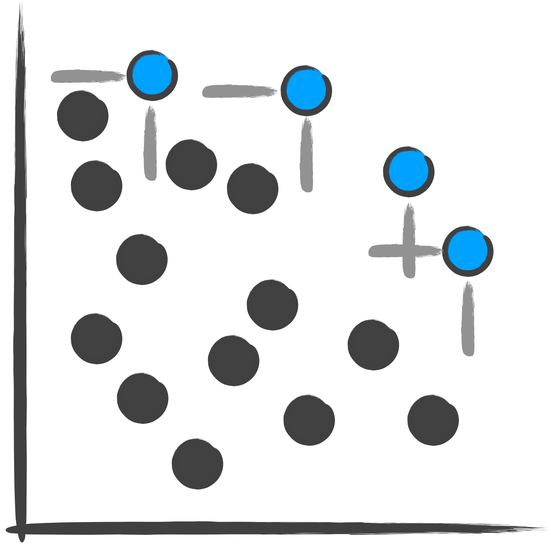
Clusters



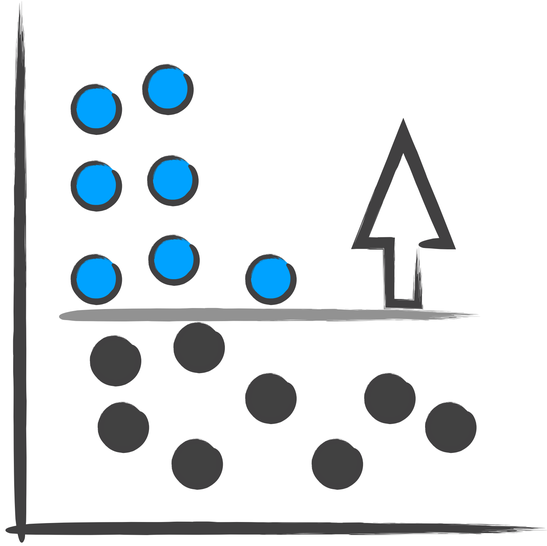
Correlation



Categories



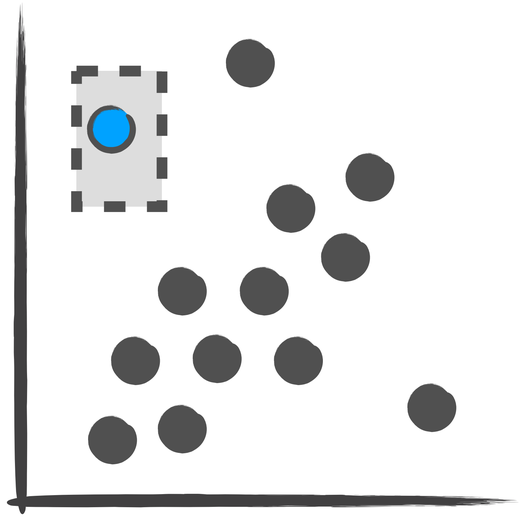
Multivariate Optimization



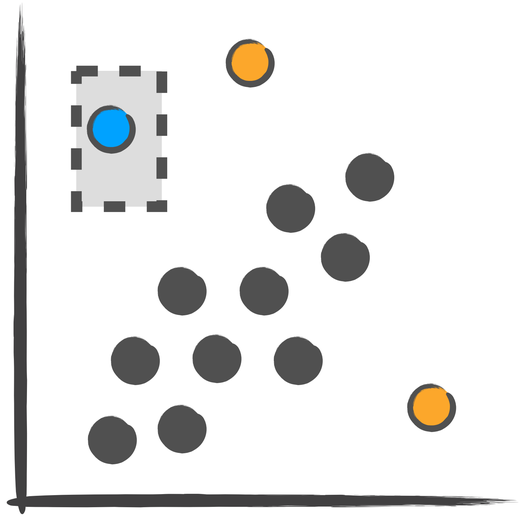
Ranges

WHY DO WE CARE?

Speed up complex selections



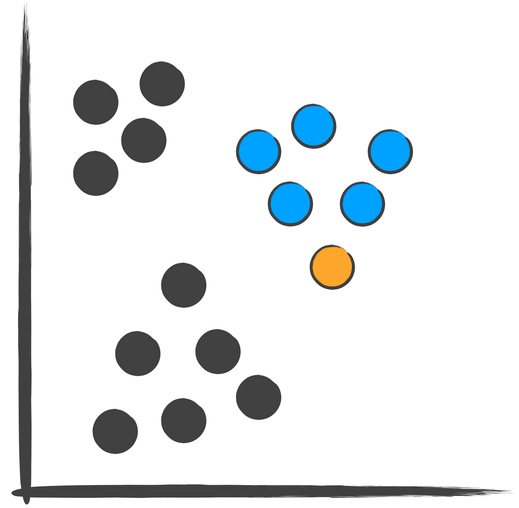
Selection



Outliers?

WHY DO WE CARE?

ID Based Selection:
Selected Elements: 7, 9, 13, 18, 22

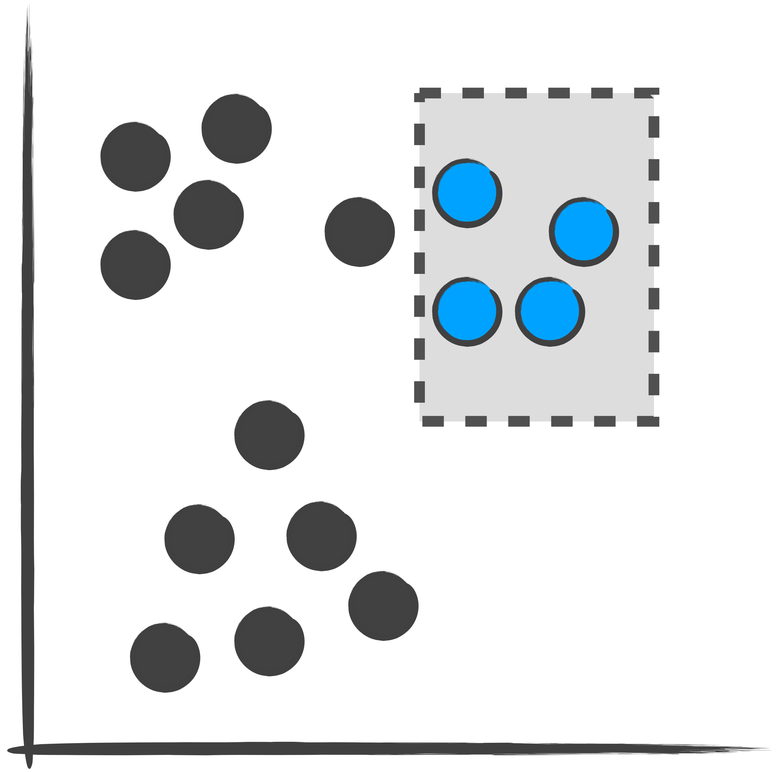


Semantic Selection:
Elements in K-Means cluster centered at [2, 3]

Meaningful, higher level concept:
improves reproducibility

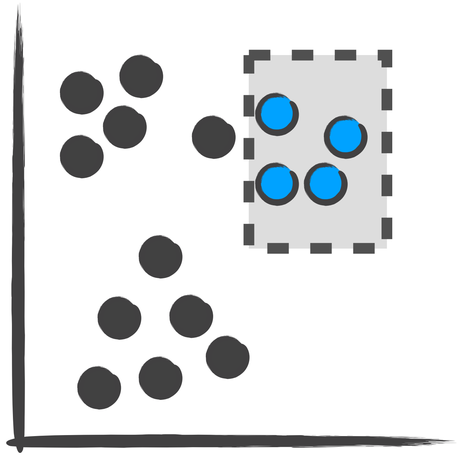
Robust to changes and updates in dataset:
enables re-usability

HOW DO WE INFER INTENT?

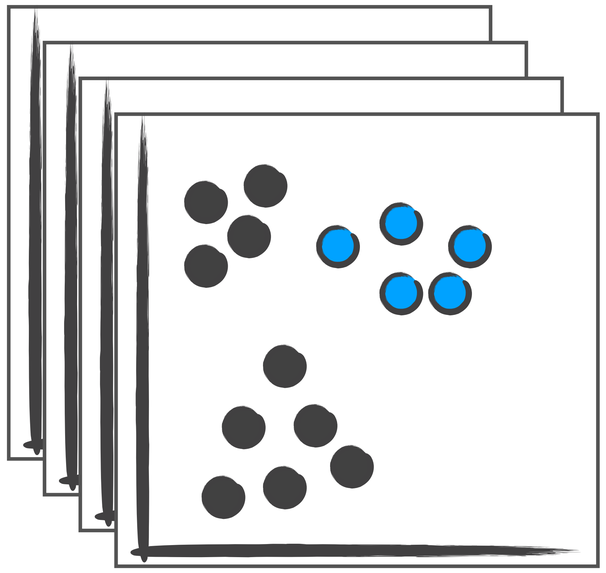


Selection

HOW DO WE INFER INTENT?



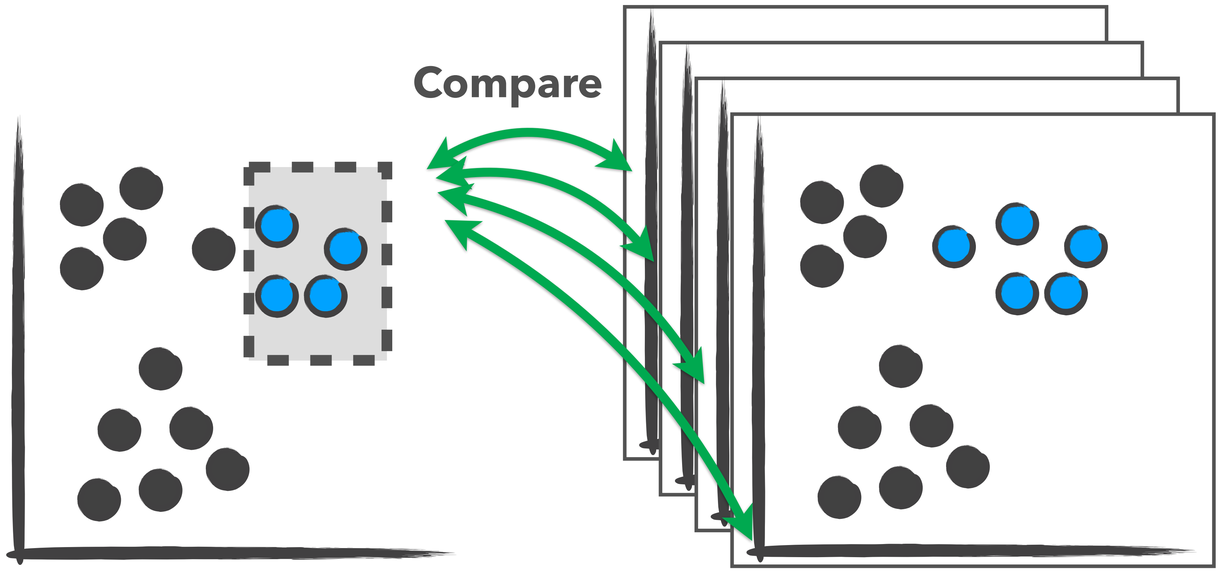
Selection



Predictions

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

HOW DO WE INFER INTENT?

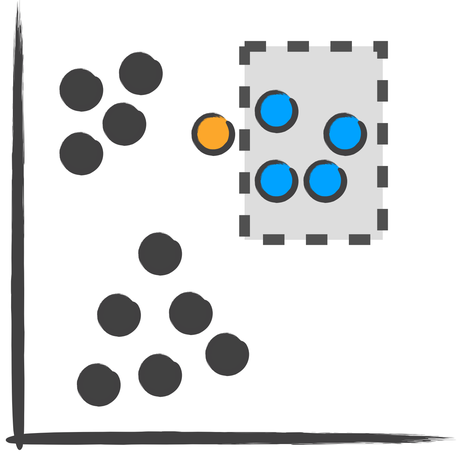


Selection

Predictions

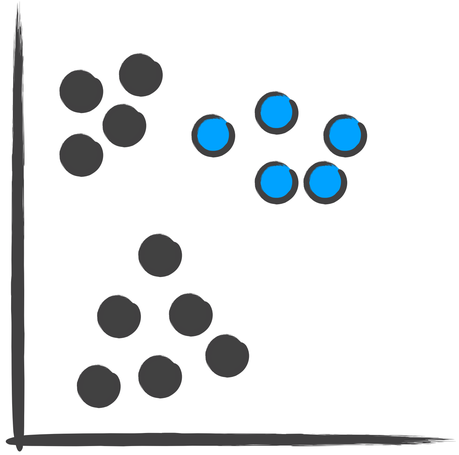
K-Means
DBScan
Regression
Outlier Detection
Skyline
Decision Trees / Ranges
Categories

1. Range
2. Cluster
3. Outlier



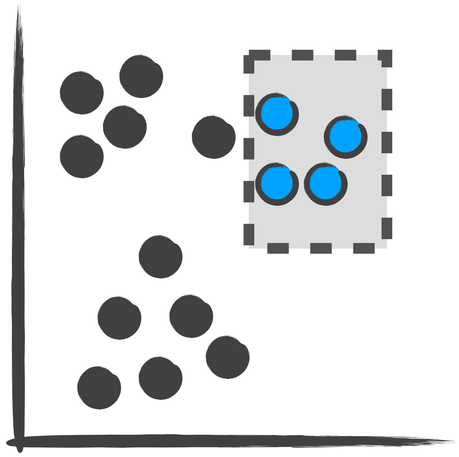
Ranking

Jaccard Distance
Naive Bayes
Classifier
Heuristic
Measures

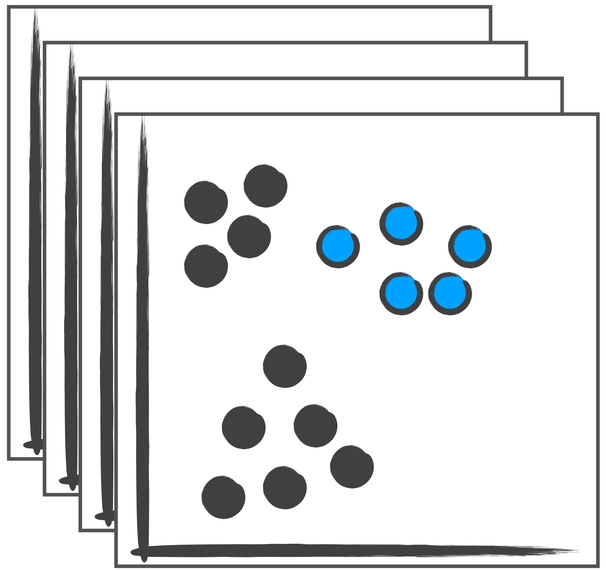


$$J(S,C) = \frac{|S \cap C|}{|S \cup C|}$$

HOW DO WE INFER INTENT?



Selection



Predictions

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

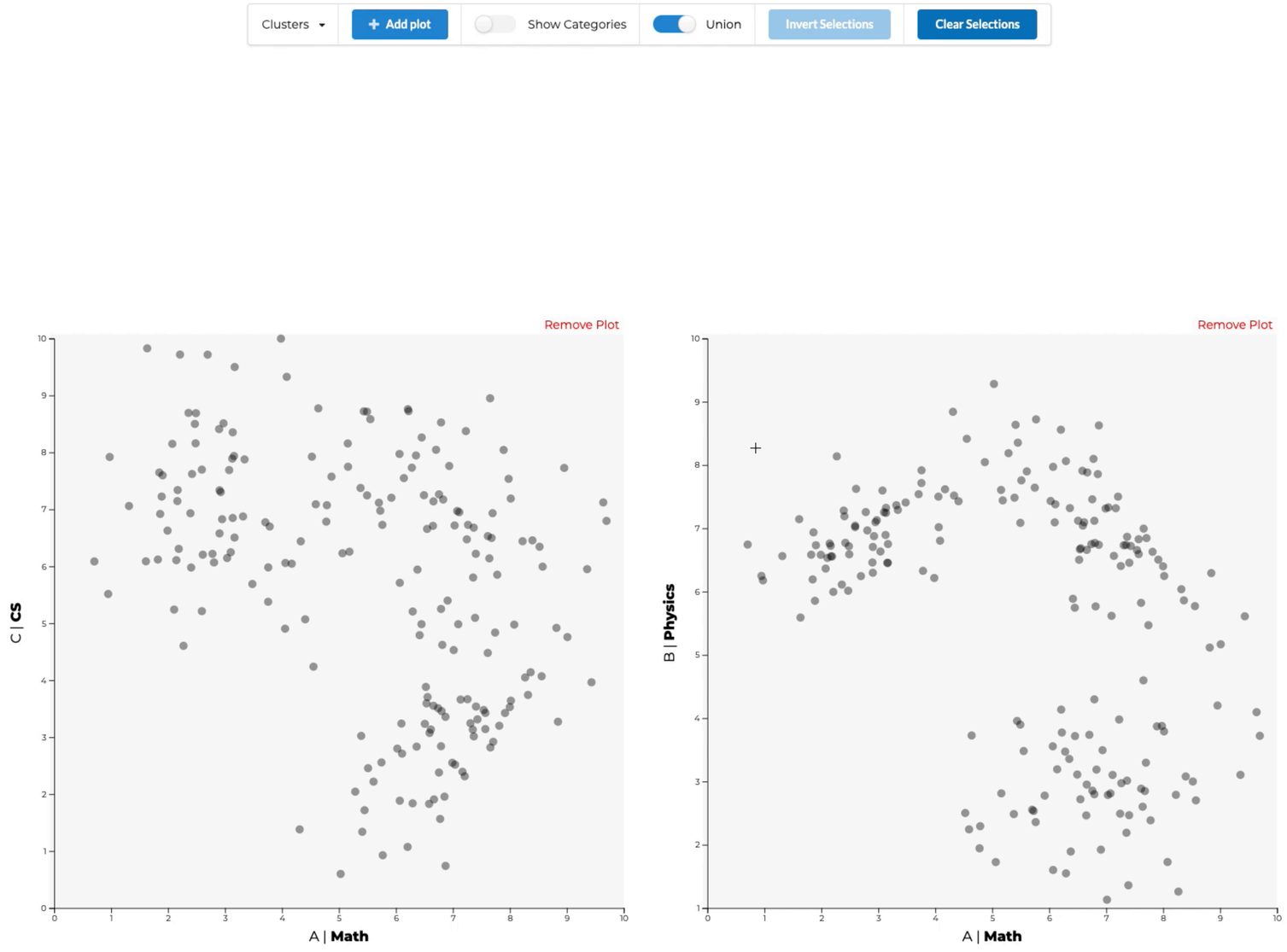
1. Range
2. Cluster
3. Outlier

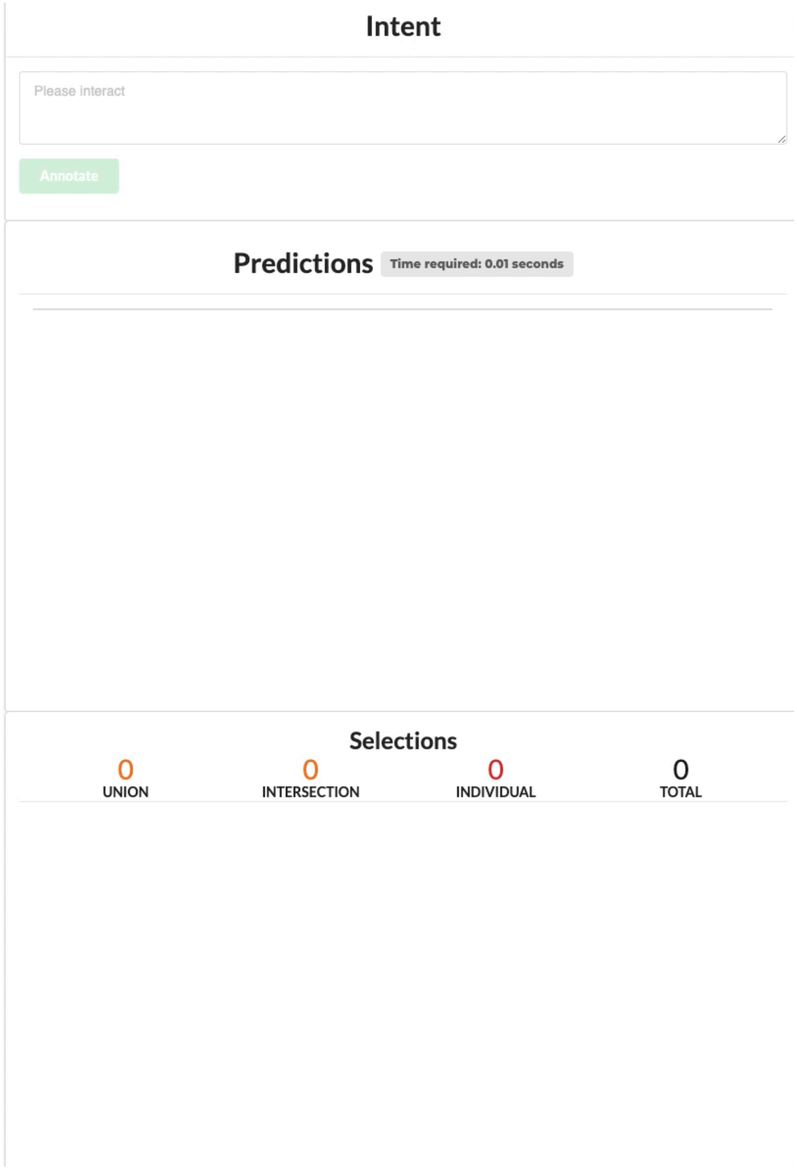
I think this cluster…

Ranking

Jaccard Distance
Naive Bayes
Classifier
Heuristic
Measures

Confirming Intent & Annotation





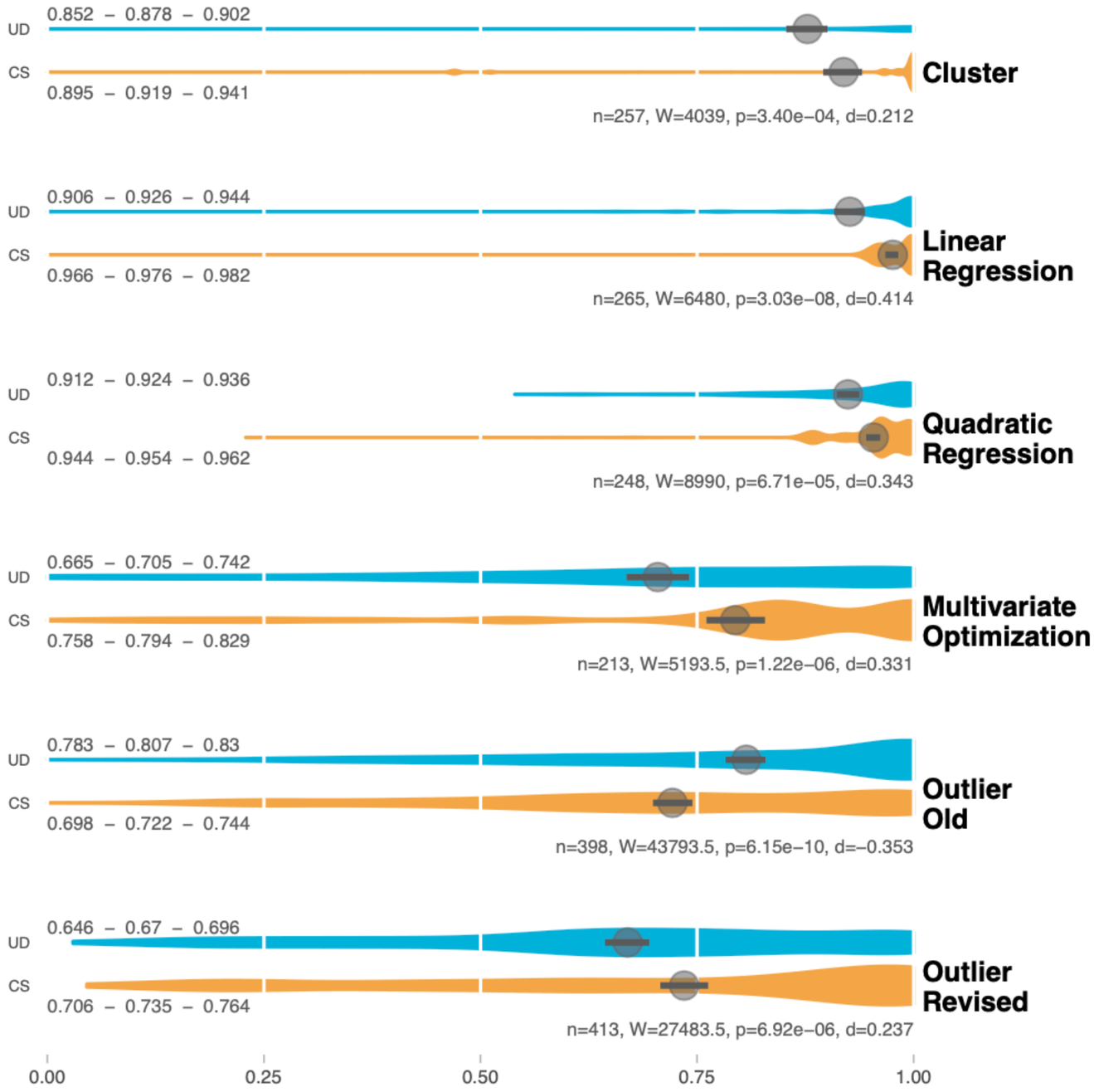
Visualization and Selection

Annotation of Intent and Predictions

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

VALIDATION

User Driven (UD)
Computer Supported (CD)



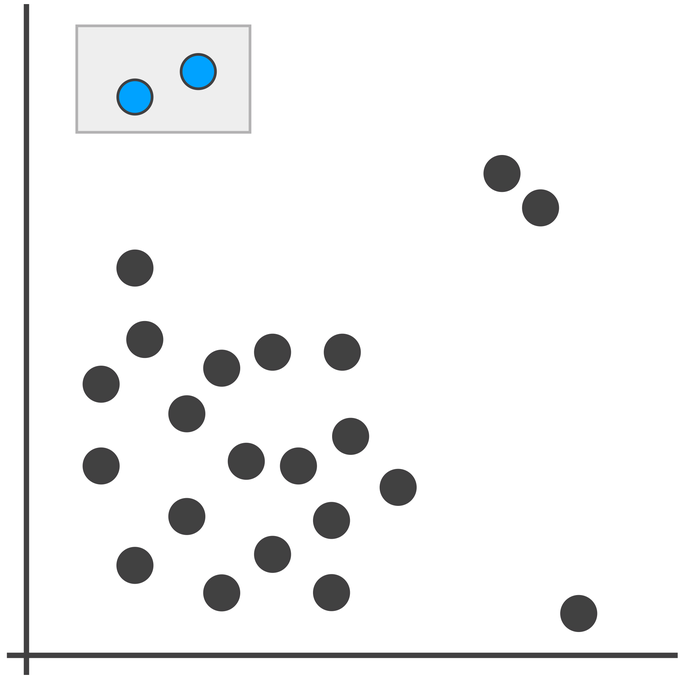
WORKFLOWS

Based on semantic selections, we can create reusable workflows!

Kiran Gadhave, Zach Cutler, Alexander Lex

Pros and Cons

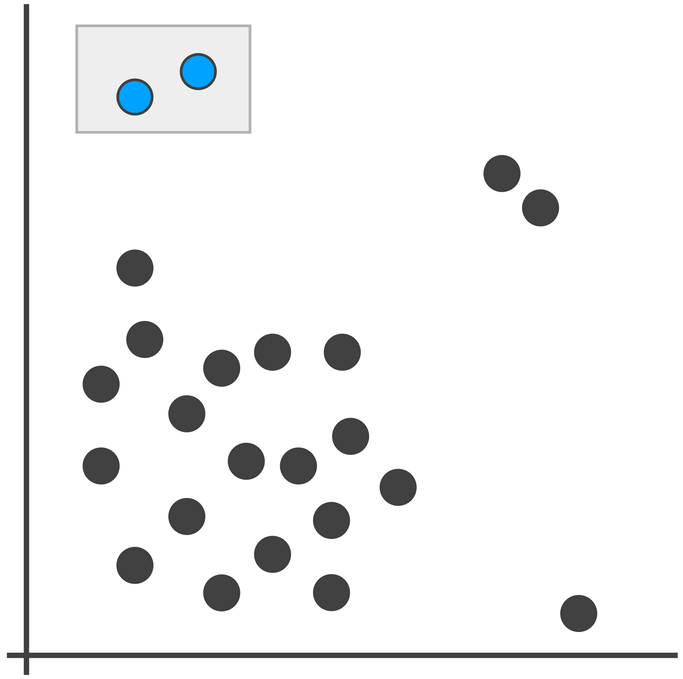
Interactive Visual Analysis



Computational Analysis

Pros and Cons

Interactive Visual Analysis

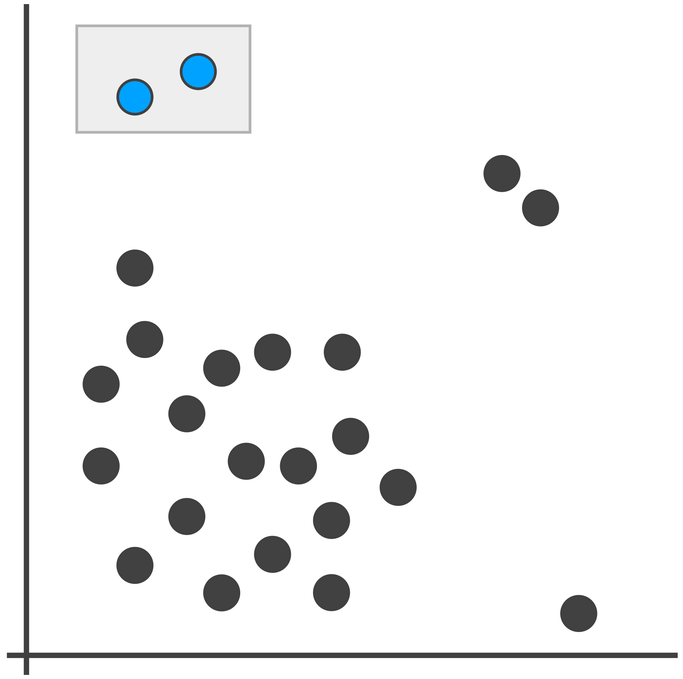


Intuitive and Fast
Uses Human Perceptual Capabilities

Computational Analysis

Pros and Cons

Interactive Visual Analysis

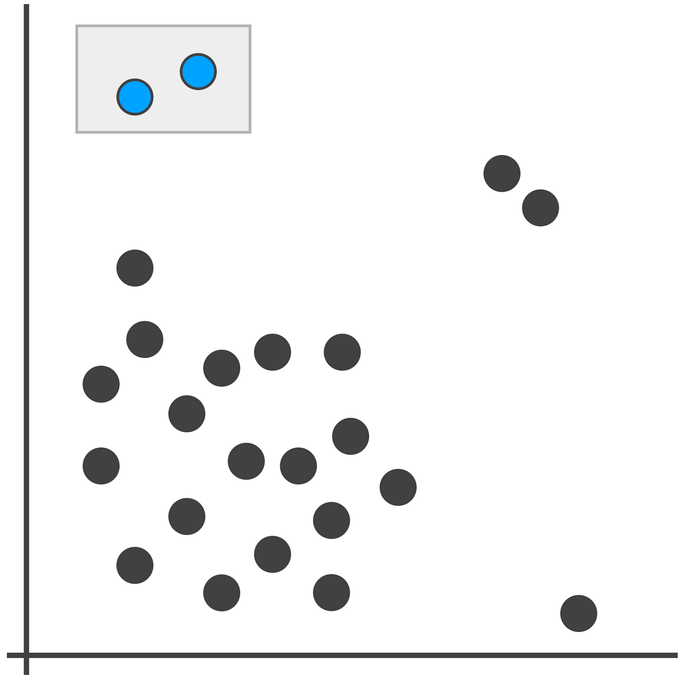


Need to redo the analysis when datasets update

Computational Analysis

Pros and Cons

Interactive Visual Analysis

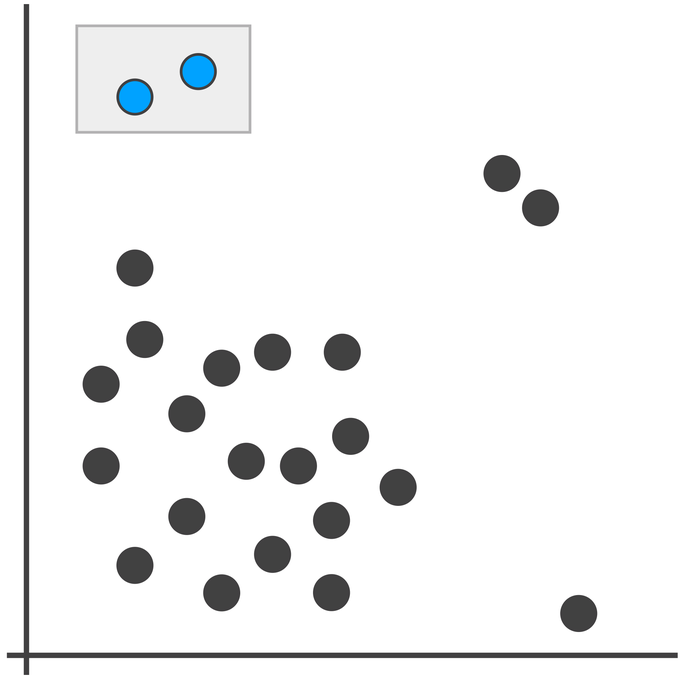


Computational Analysis

Flexible, powerful, reusable.

Pros and Cons

Interactive Visual Analysis



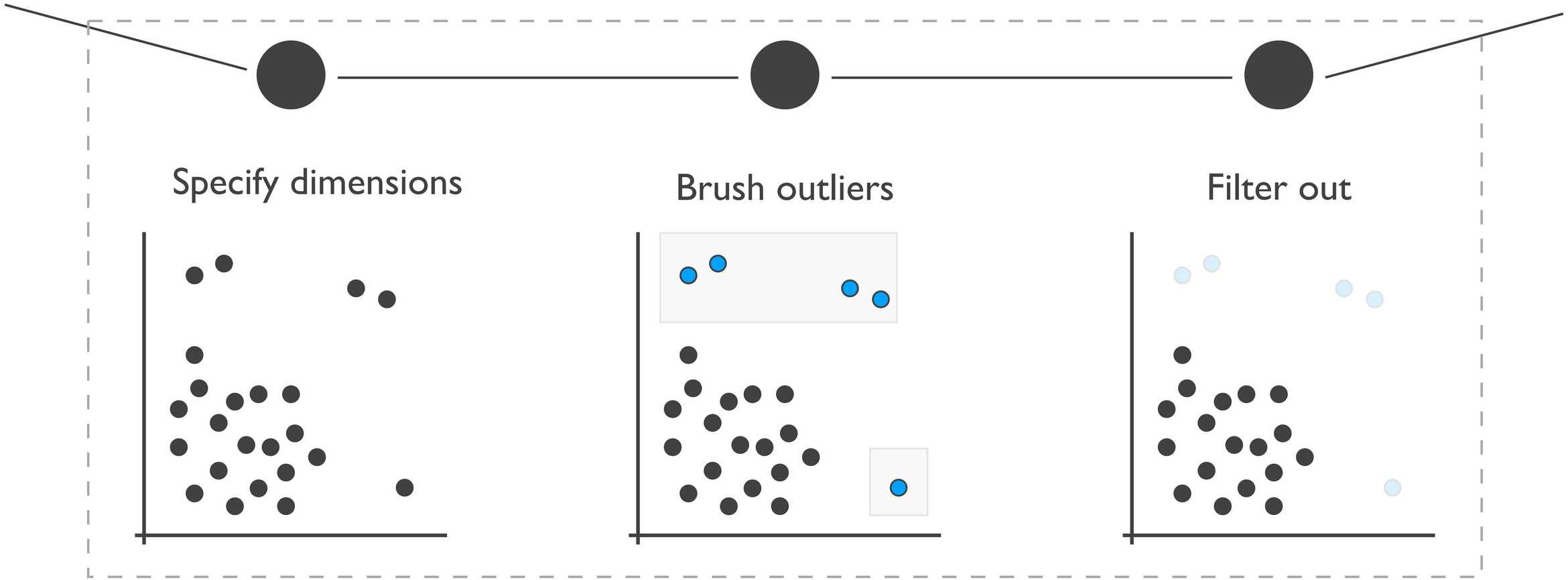
Computational Analysis

Time Consuming

Require analyst can and want to program

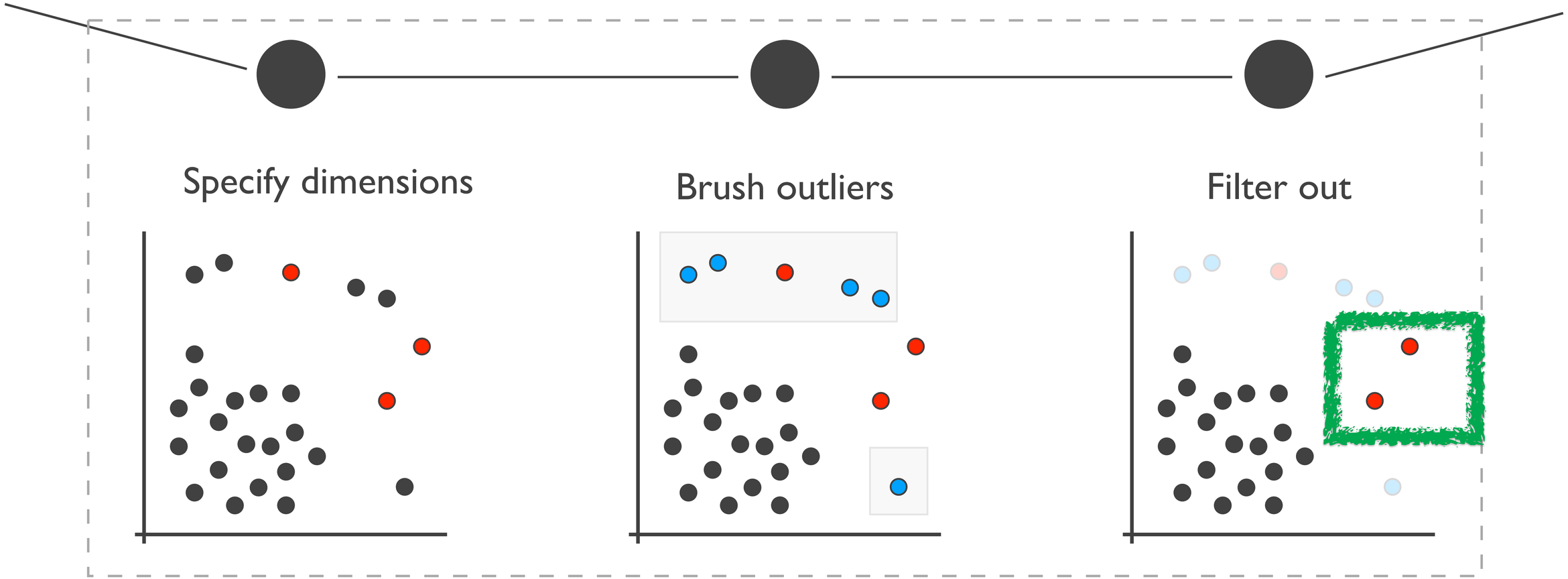
Difficult to see what's in the data

IDEA: CAPTURING WORKFLOWS

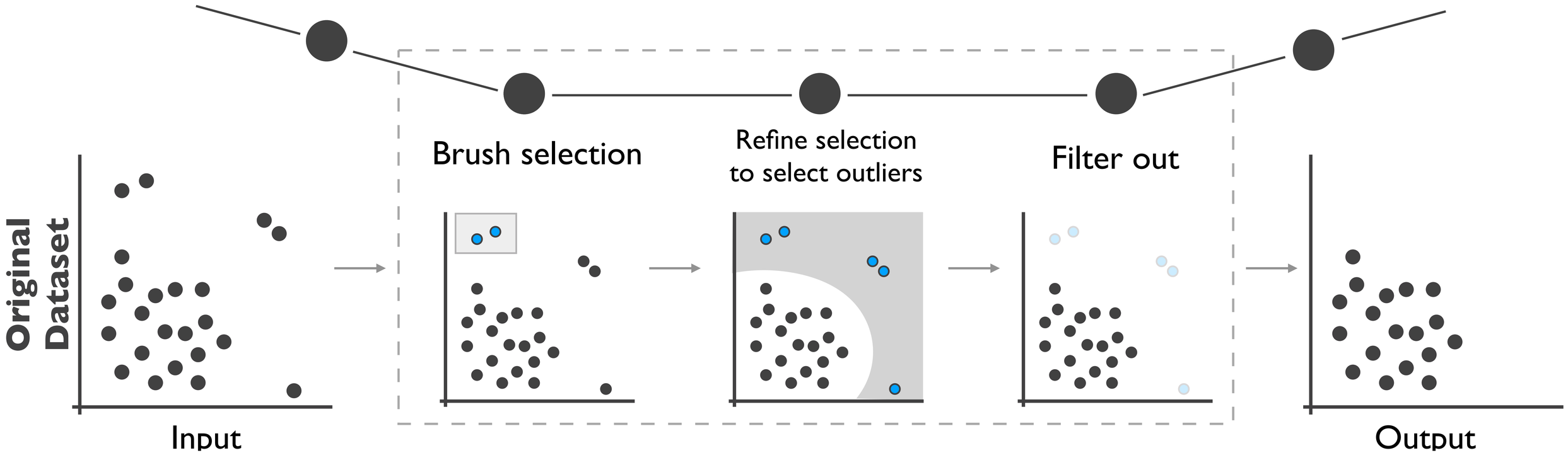


"Filter Outliers" Workflow

USING WORKFLOW ON UPDATED DATA

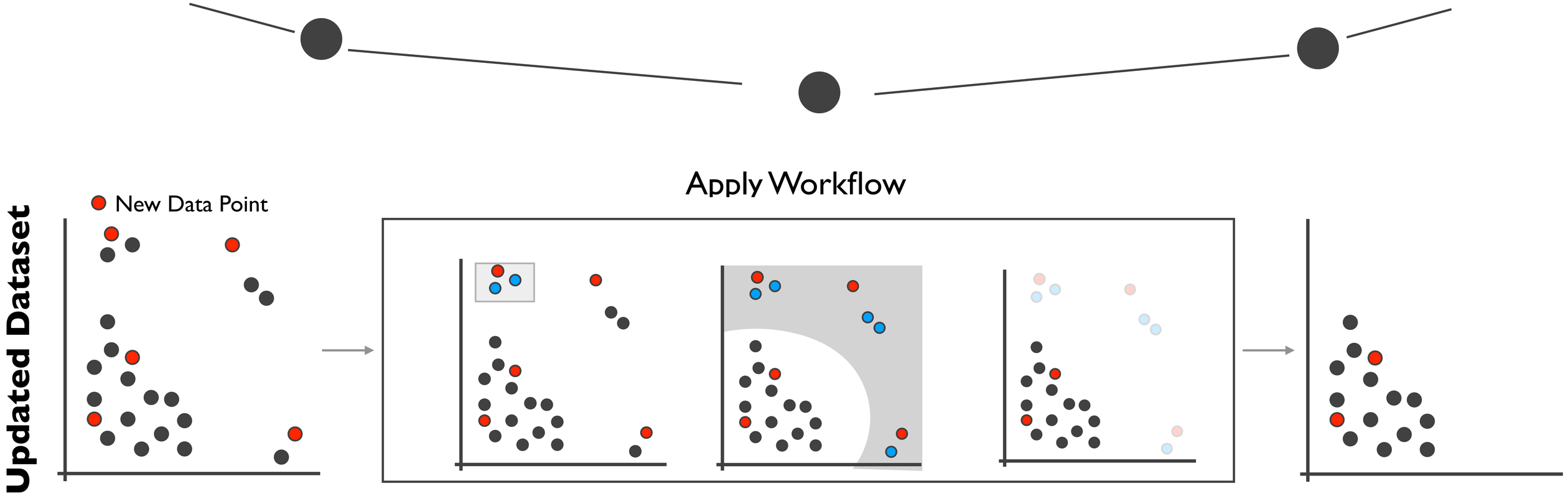


CAPTURING SEMANTICS OF WORKFLOWS

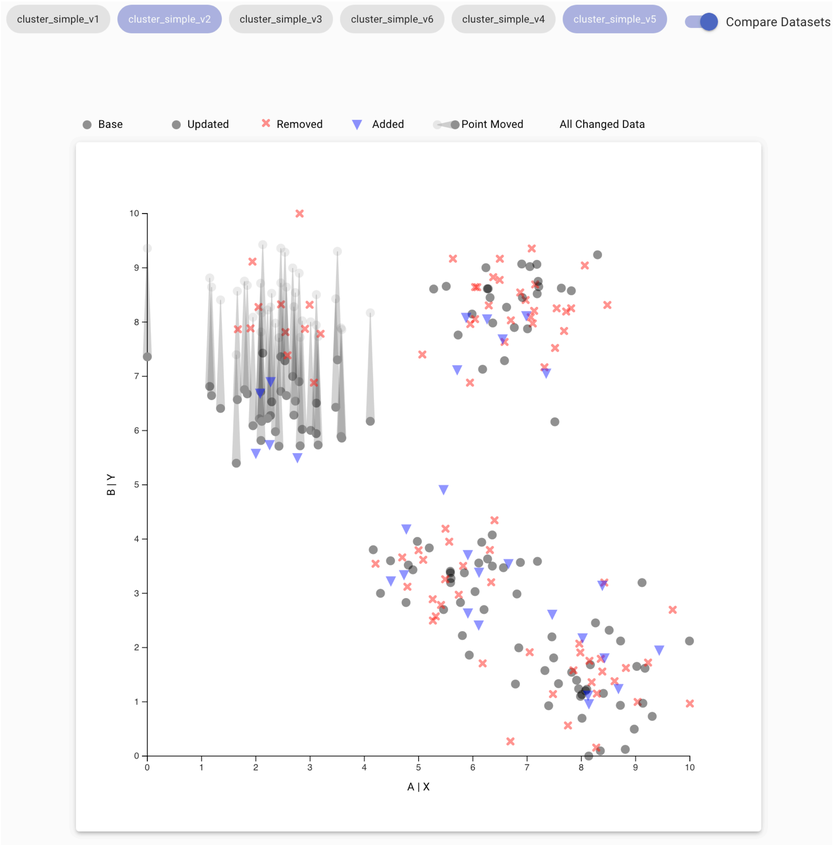


Robust “Filter Outliers” Workflow

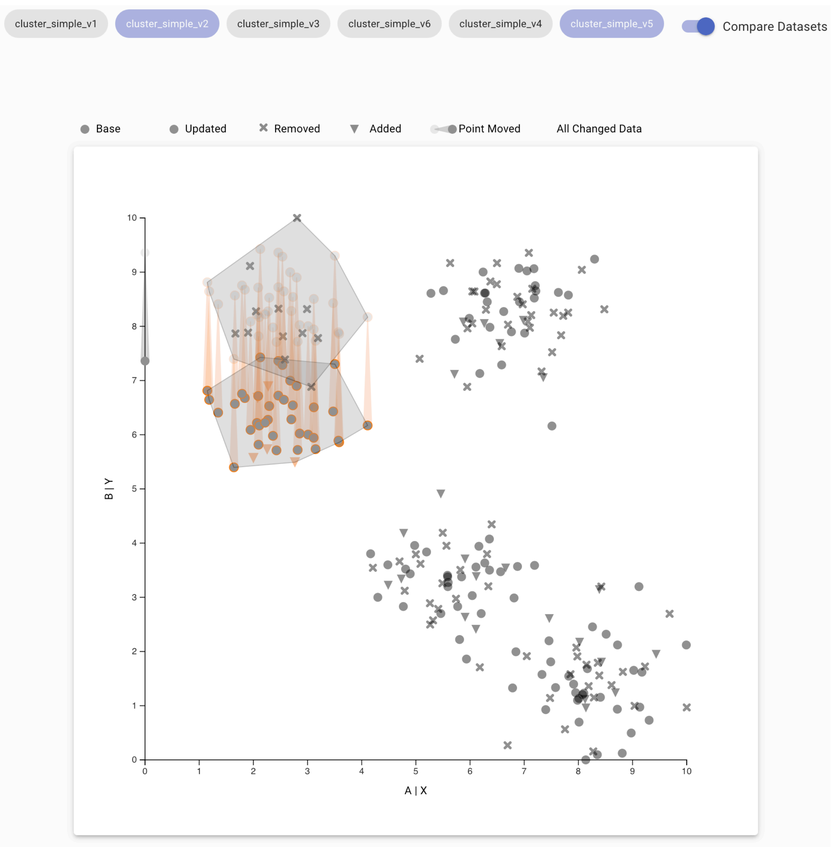
REUSING WORKFLOWS ON UPDATED DATA



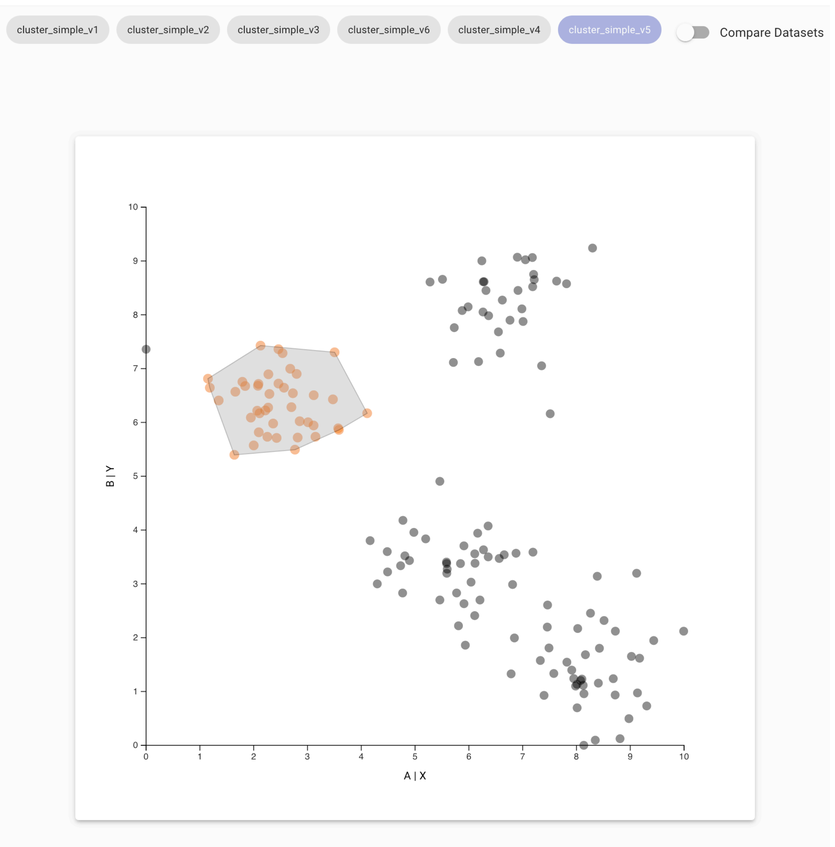
REUSING SELECTIONS ON UPDATED DATASETS



Changed Dataset

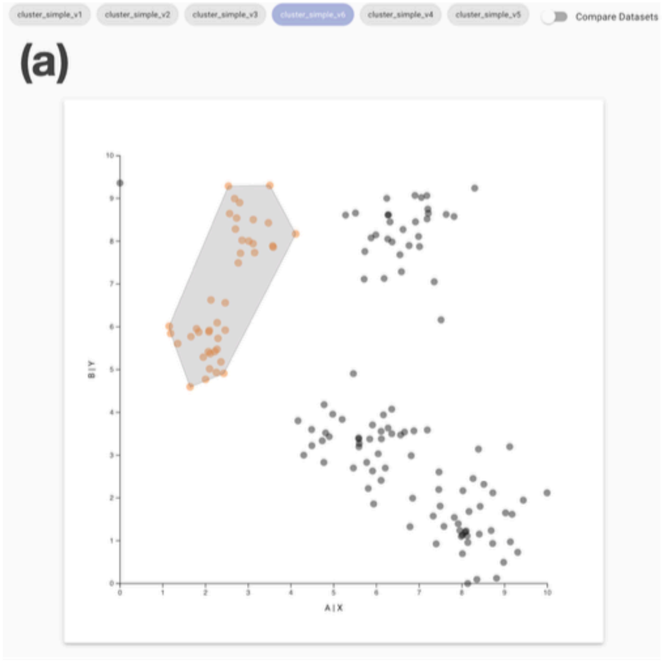


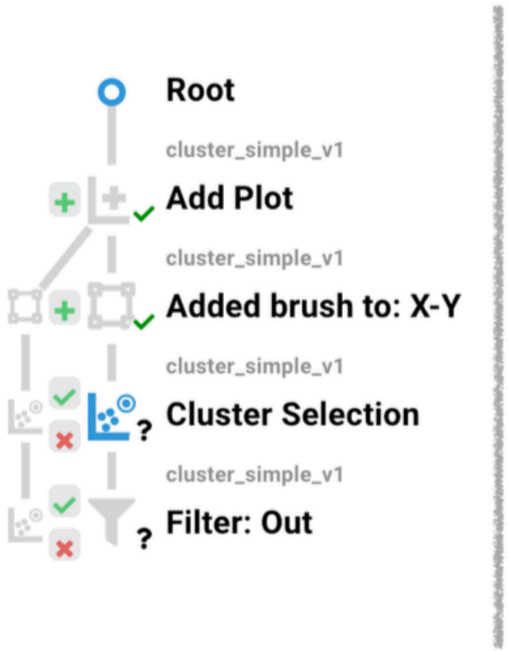
Tracking A Selected Cluster

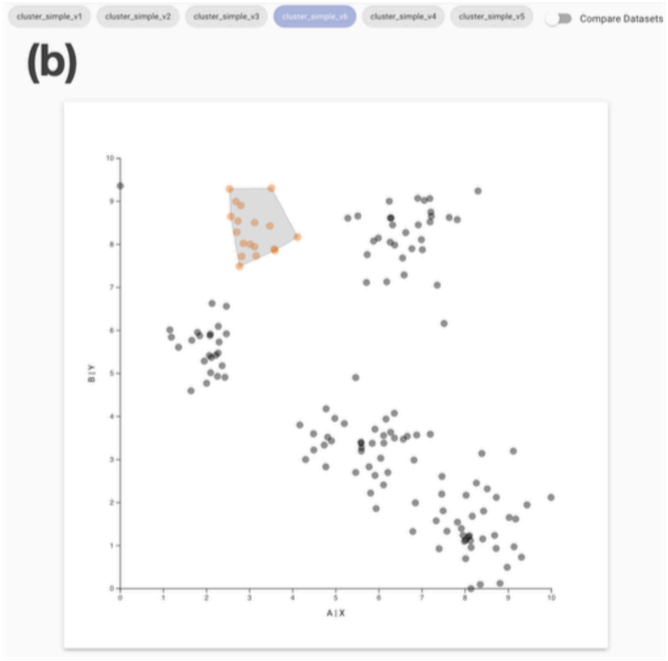


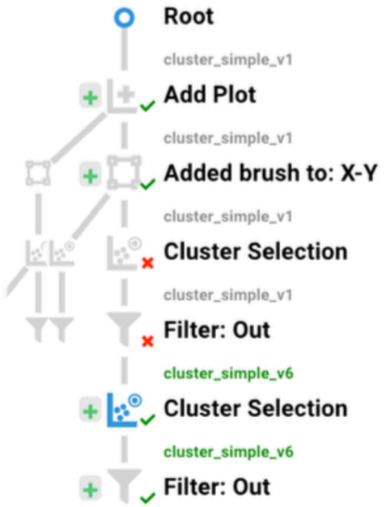
Selected Cluster on Changed Dataset

HUMAN REVIEWS

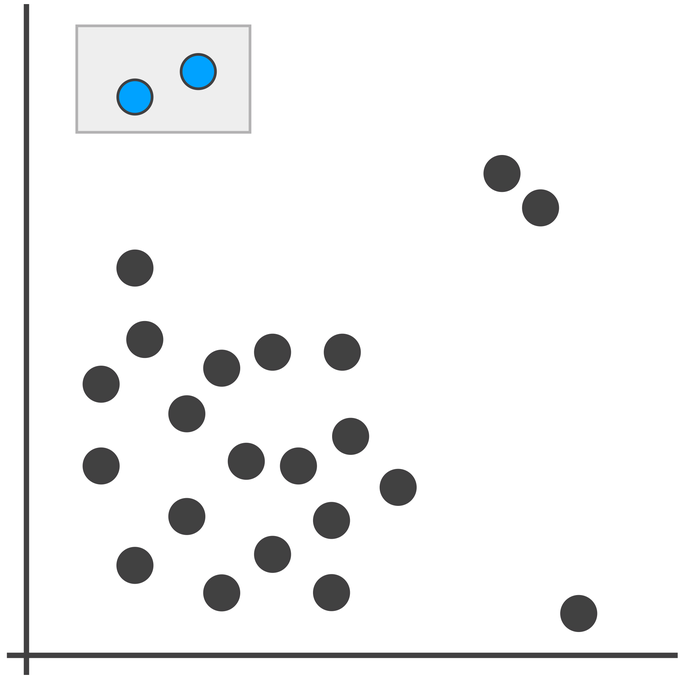








Interactive Visual Analysis



Workflow
Database

Computational Analysis

jupyter

USING WORKFLOW IN A COMPUTATIONAL NOTEBOOK

```
# Installing the reapply-workflows adds a module called backend
# This module exposes the Reapply class which initializes the library

from backend import Reapply
```

```
# Here we load the reapply_workflows library.
r = Reapply()

# We add a workflow from our workflow database.
workflow = r.load_workflow("workflow16178086816 20")
```

```
# Print the workflow name
print("Workflow: ", workflow.name, "\n")

# Description of the workflow and the operations in it
workflow.describe
```

```
Workflow:  Deleting Cluster

| Root
 +--| Add Plot
     +--| Added brush to: X-Y
         +--| Cluster Selection
             +--| Filter: Out
```

```
# Prints the reapply results for all interactions, along with review status.
```

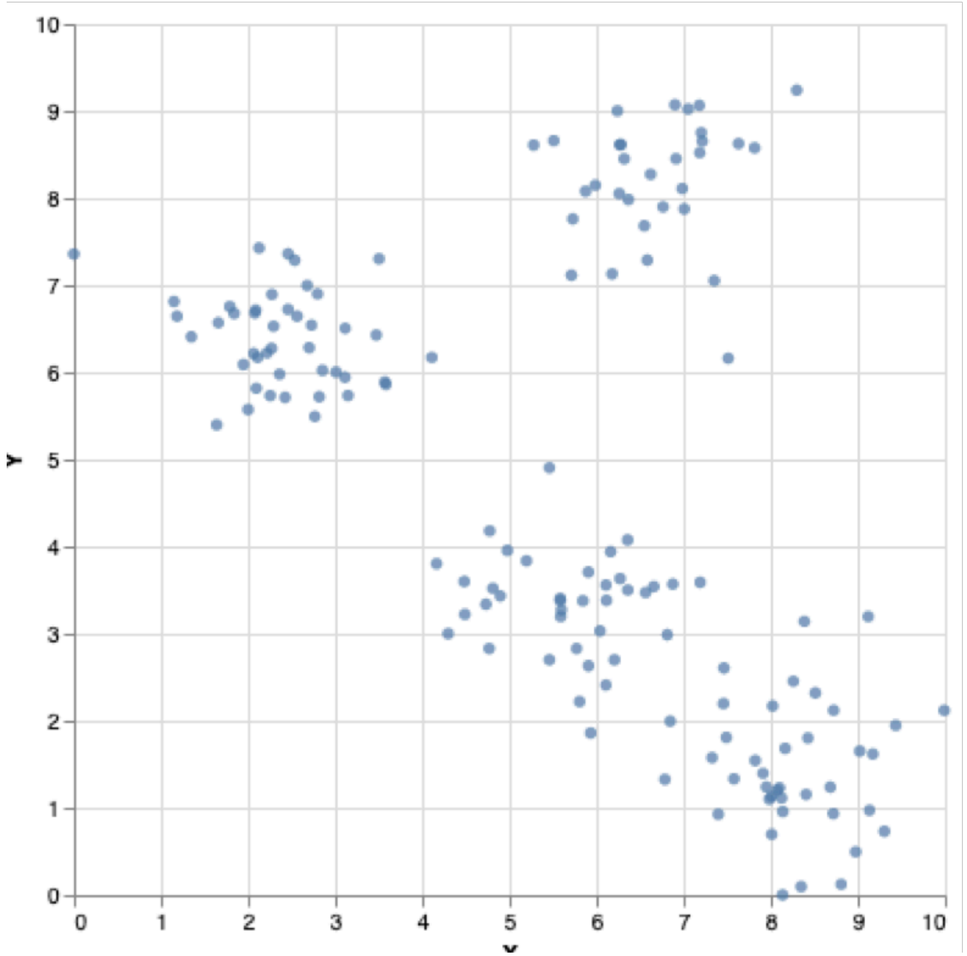
```
# Apply the workflow to target dataset.
# apply function requires the target dataset
# and the label column as arguments.
res = workflow.apply(target, "Label")

# Results is an array of datasets for each interaction
# we grab the final one.
result_dataset = res.results[-1]['data']
result_dataset
```

```
This workflow has not been reviewed for all interactions.
Please go to following url: https://reapply-workflows.gith
```

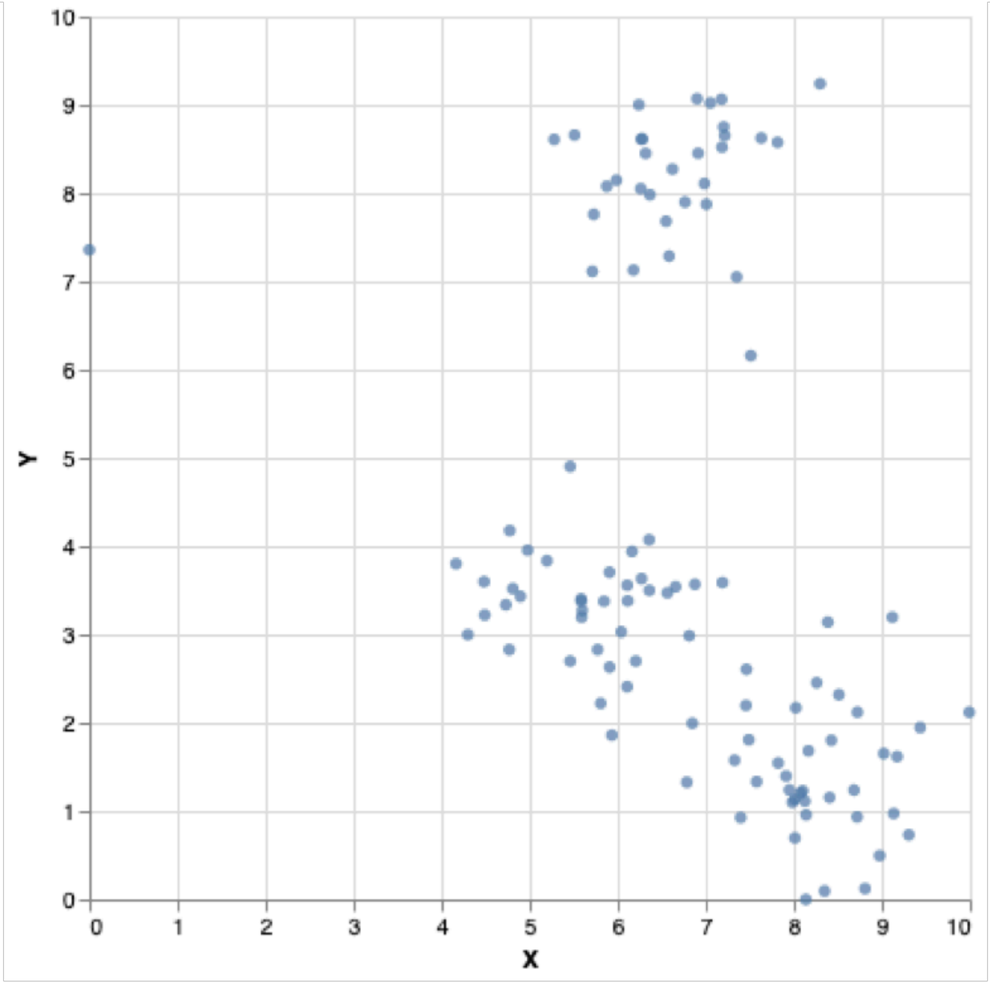
	Label	X	Y
3	P52	6.58351	7.28796
5	P171	4.77421	4.17980
8	P199	8.34966	0.09550
9	P183	8.42670	1.80299
10	P61	4.29760	2.99981
...	...	...	...
141	P138	7.35179	7.05215
142	P46	6.62171	8.27311

BEFORE AND AFTER

150 rows × 3 columns



108 rows × 3 columns



jan_2021 | june_2021 | dec_2020 | Compare

Graph | Bookmarks/Annotations

Brush Type

Transforms

FILTER

LABEL

AGGREGATE

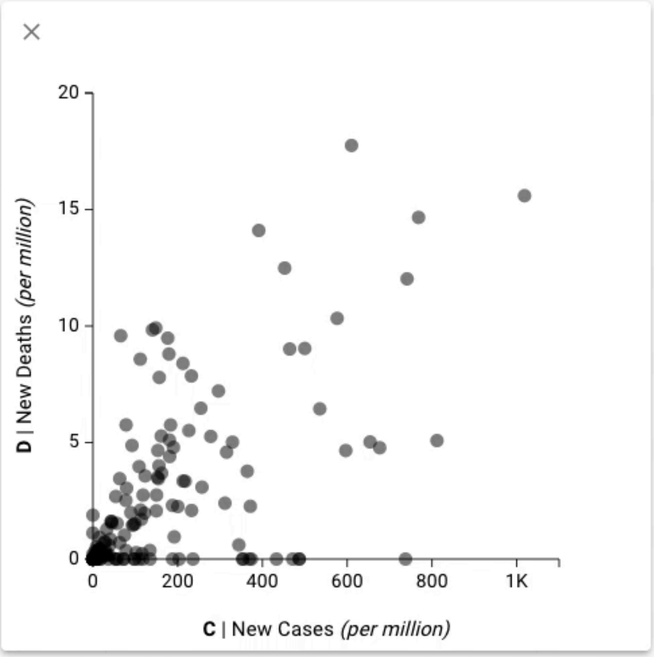
Categories

Enable category encoding

Showing categories

Case Category

- UNASSIGNED
- HIGH DEATHS - HIGH CASES
- LOW DEATH - HIGH CASES
- HIGH DEATHS - LOW CASES



NO VALID SELECTIONS

Undo | Redo

- Root
- jan_2021 Adding scatterplot f..
- jan_2021 Adding scatterplot f..
- jan_2021 Remove Scatterplot

COVID Case Study

TOWARDS LITERATE & REUSABLE VISUALIZATION

Semantic selections & annotations allow us to have higher-level information about events in an analysis process.

Analysts have the means to justify their choices.

Makes it possible to reuse interactive analysis processes on updated datasets.

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

Thanks to: **Kiran Gadhave, Zach Cutler**, Marc Streit, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Samuel Gratzl, Holger Stitz,, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister, and many others!

visualization
design lab