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

THE UNIVERSITY OF UTAH

# Visualization in Data Science:
# Challenges and Opportunities

visualization design lab

SCI
www.sci.utah.edu

# ME & ICG

**I’ve spent a lot of time at ICG – and it was great!**

**2003** First ICG Class with **Franz Leberl** and (then new) **Horst Bischof**; Teaching Assistant: **Tom Pock**

**2006** ICG **BS** Thesis w. **Horst Bischof** & **Martin Urschler**

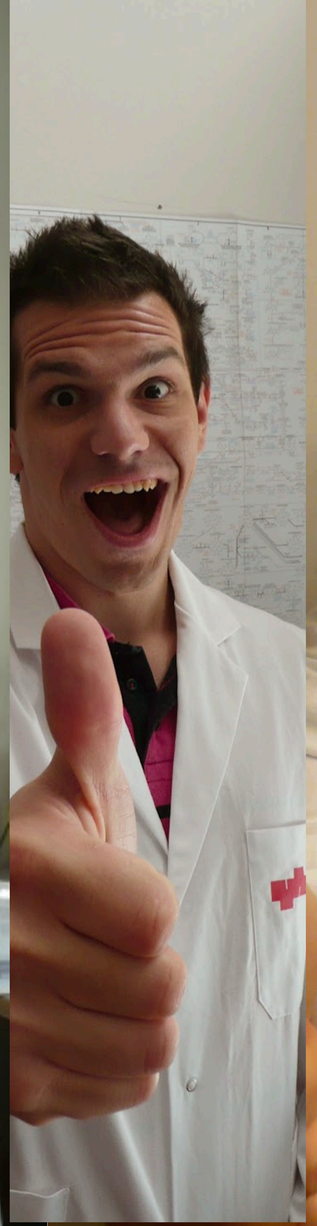
**2008** ICG **MS** Thesis w. **Dieter Schmalstieg**

**2012** ICG **PhD** Thesis w. **Dieter Schmalstieg**


KeyBank
KeyBank
ZIONS BANK
WELLS FARGO
WALKER CENTER
usbank
PLAZA


WESTERN CAPE
CAPE ADVENTURE
BEER MECH



# RESEARCH AREAS

## TECHNICAL CONTRIBUTIONS

**Novel Visualization Techniques**

**Visualization Process Innovations**

**Data Wrangling Methods**

**Accessibility**

## DOMAIN DRIVEN TECHNIQUES

**Tailored Methods and Systems for High Impact Science Problems**

## EMPIRICAL & THEORETICAL WORK

**Evaluation Methodology**

**Qualitative Studies**

**Design Spaces / Taxonomies**

# VIS IN DATA SCIENCE

## CHALLENGES & OPPORTUNITIES

1. SYSTEMS ARE HARD

2. INTERACTION IS POWERFUL BUT EPHEMERAL

3. DATA IS NOT THE TRUTH

# 1. SYSTEMS ARE HARD

# SYSTEMS ARE HARD

**Publishing software increases impact of your work**

**Spent a lot of time on building interactive visualization systems**

**Adoption is minimal. Why?**

CALEYDO

# SYSTEMS ARE HARD

Hard to build good UX in academic setting

- $$$
- Maintenance != publishing

Analysts don't want a new & complicated tool unless it is a significant improvement

Limited expressivity

- Simplicity – expressivity trade-off

# WHAT CAN WE DO?

1. The idea matters most

2. Work on reusable components

3. Meet users where they are

4. Commercialize

# Alexander Lex

FOLLOW

Associate Professor, SCI Institute, School of Computing, University of Utah
Verified email at sci.utah.edu - Homepage

Information Visualization  Visualization  Bioinformatics  Visual Analytics  Data Science

| TITLE | CITED BY | YEAR |
|---|---|---|
| UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties<br>JR Conway, A Lex, N Gehlenborg<br>Bioinformatics 33 (18), 2938-2940 | 1763 | 2017 |
| UpSet: Visualization of Intersecting Sets<br>A Lex, N Gehlenborg, H Strobelt, R Vuillemot, H Pfister<br>IEEE Transactions on Visualization and Computer Graphics 20 (12), 1983-1992 | 1390 | 2014 |
| LineUp: Visual Analysis of Multi-Attribute Rankings<br>S Gratzl, A Lex, N Gehlenborg, H Pfister, M Streit<br>IEEE Transactions on Visualization and Computer Graphics 19 (12), 2277-2286 | 344 | 2013 |

**Reusable component in common environment**

**Idea matters**

**Commercialize**

# Alexander Lex

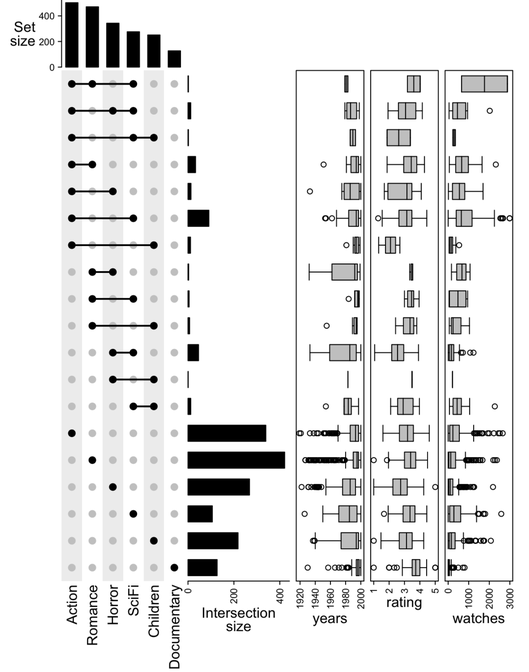
FOLLOW

Associate Professor, SCI Institute, School of Computing, University of Utah
Verified email at sci.utah.edu - Homepage

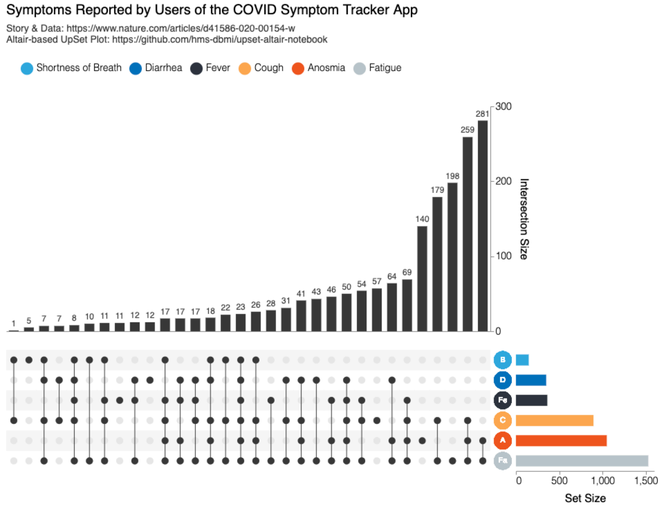
Information Visualization  Visualization  Bioinformatics  Visual Analytics  Data Science

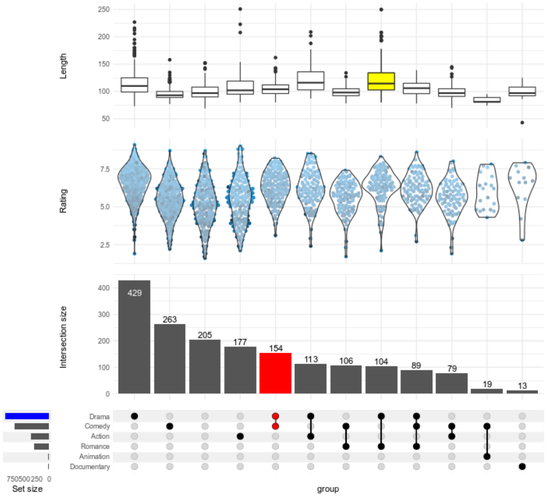
| TITLE | CITED BY | YEAR |
|---|---|---|
| UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties<br>JR Conway, A Lex, N Gehlenborg<br>Bioinformatics 33 (18), 2938-2940 | 1763 | 2017 |
| UpSet: Visualization of Intersecting Sets<br>A Lex, N Gehlenborg, H Strobelt, R Vuillemot, H Pfister<br>IEEE Transactions on Visualization and Computer Graphics 20 (12), 1983-1992 | 1390 | 2014 |
| LineUp: Visual Analysis of Multi-Attribute Rankings<br>S Gratzl, A Lex, N Gehlenborg, H Pfister, M Streit<br>IEEE Transactions on Visualization and Computer Graphics 19 (12), 2277-2286 | 344 | 2013 |

# Idea matters




# Basic UpSet Idea has been re-implemented in R, Python, Tableau, JavaScript, etc.




# Alexander Lex

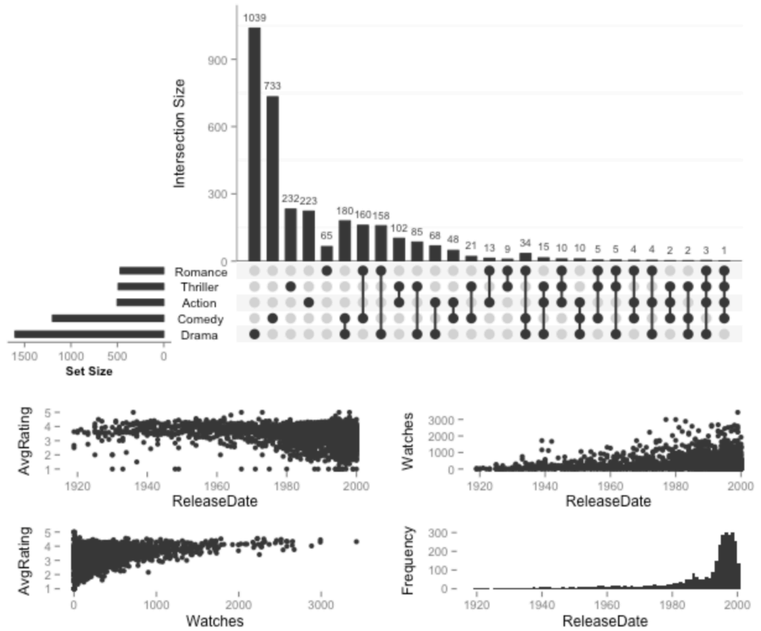
FOLLOW

Associate Professor, SCI Institute, School of Computing, University of Utah
Verified email at sci.utah.edu - Homepage

Information Visualization  Visualization  Bioinformatics  Visual Analytics  Data Science


| TITLE | CITED BY | YEAR |
|---|---|---|
| **UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties**<br>JR Conway, A Lex, N Gehlenborg<br>Bioinformatics 33 (18), 2938-2940 | 1763 | 2017 |
| **UpSet: Visualization of Intersecting Sets**<br>A Lex, N Gehlenborg, H Strobelt, R Vuillemot, H Pfister<br>IEEE Transactions on Visualization and Computer Graphics 20 (12), 1983-1992 | 1390 | 2014 |
| **LineUp: Visual Analysis of Multi-Attribute Rankings**<br>S Gratzl, A Lex, N Gehlenborg, H Pfister, M Streit<br>IEEE Transactions on Visualization and Computer Graphics 19 (12), 2277-2286 | 344 | |

## Reusable component in common environment




```
upset(movies,attribute.plots=list(gridrows=60,plots=list(list(plot=scatter_plot, x="ReleaseDat
list(plot=scatter_plot, x="ReleaseDate", y="Watches"),list(plot=scatter_plot, x="Watches", y="Avgra
list(plot=histogram, x="ReleaseDate")), ncols = 2))
```

## An R package that generates useful figures for inclusion in papers

# Alexander Lex

FOLLOW

Associate Professor, SCI Institute, School of Computing, University of Utah
Verified email at sci.utah.edu - Homepage

Information Visualization  Visualization  Bioinformatics  Visual Analytics  Data Science

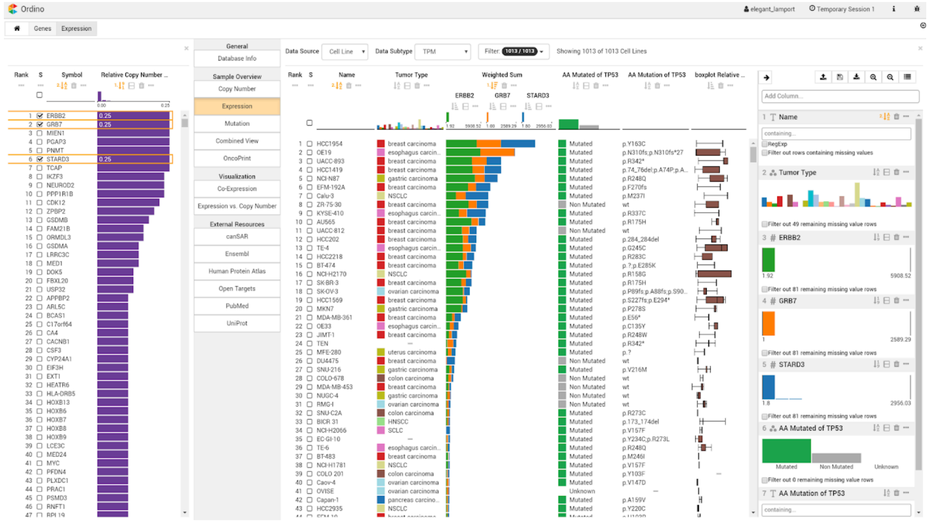
datavisyn

| TITLE | CITED BY | YEAR |
|---|---|---|
| UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties<br>JR Conway, A Lex, N Gehlenborg<br>Bioinformatics 33 (18), 2938-2940 | 1763 | 2017 |
| UpSet: Visualization of Intersecting Sets<br>A Lex, N Gehlenborg, H Strobelt, R Vuillemot, H Pfister<br>IEEE Transactions on Visualization and Computer Graphics 20 (12), 1983-1992 | 1390 | 2014 |
| LineUp: Visual Analysis of Multi-Attribute Rankings<br>S Gratzl, A Lex, N Gehlenborg, H Pfister, M Streit<br>IEEE Transactions on Visualization and Computer Graphics 19 (12), 2277-2286 | 344 | 2013 |

## Commercialize

**There is a need for complex visualization systems, but you need resources to make it useful.**

**datavisyn works with 5 of the 10 biggest pharmaceuticals**



# OPPORTUNITIES

Publish good ideas! [dough]

Publish small, maintainable software packages that fit into an ecosystem

- Reusability, flexibility, less maintenance burdens

Develop DSLs instead of UI systems (Vega, reVISit, …)

- UIs are just A LOT OF WORK
- DSLs can do things in well defined ways
- Tooling (including UIs) can follow later

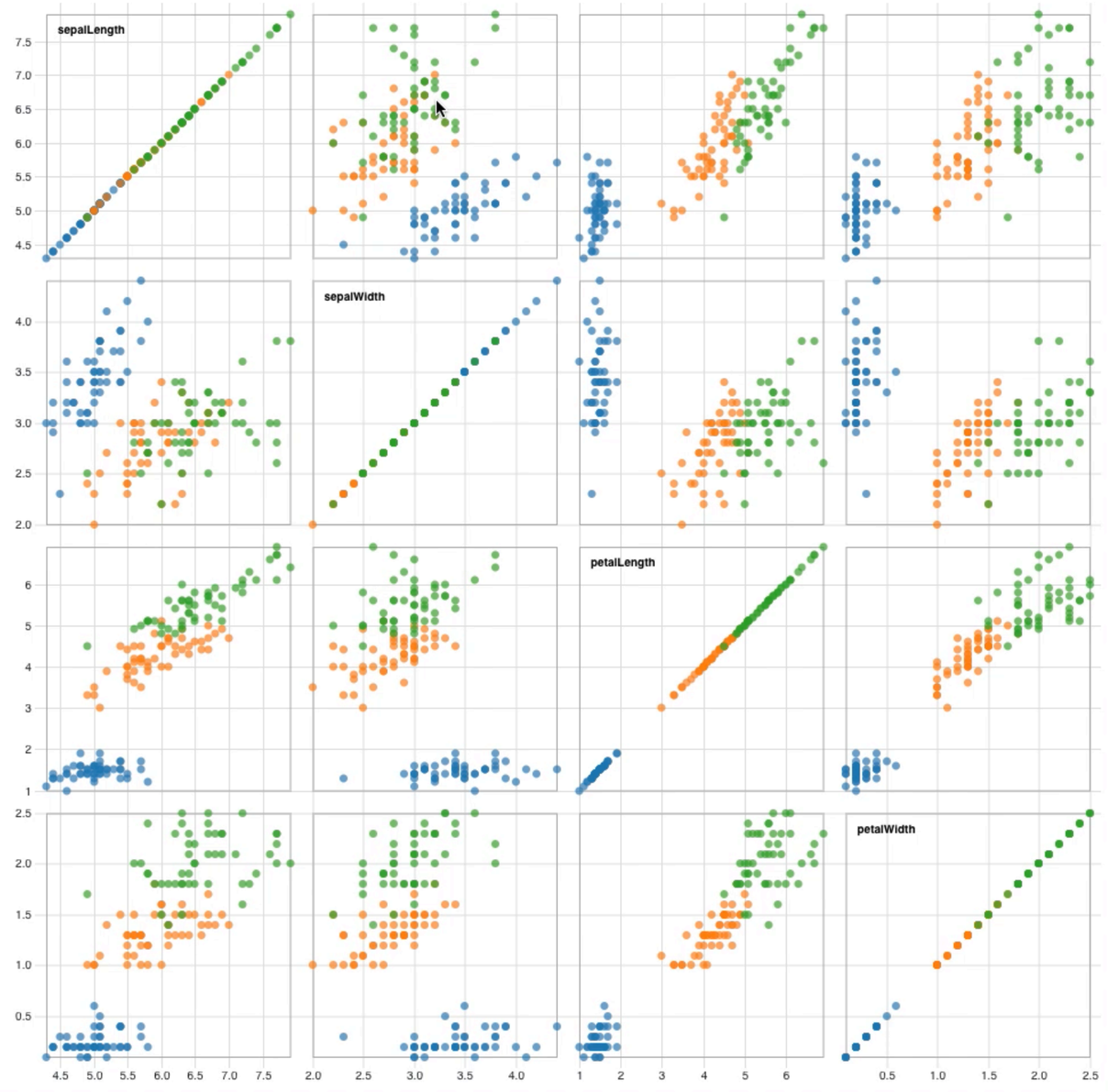
1. SYSTEMS ARE HARD

2. INTERACTION IS POWERFUL BUT EPHEMERAL

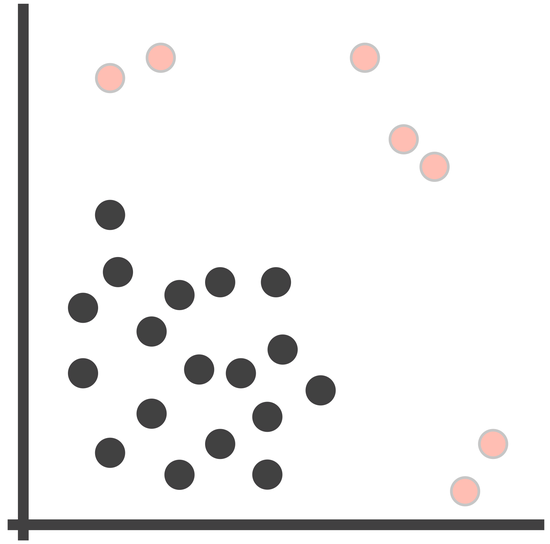
3. DATA IS NOT THE TRUTH

# 2. INTERACTION IS POWERFUL BUT EPHEMERAL

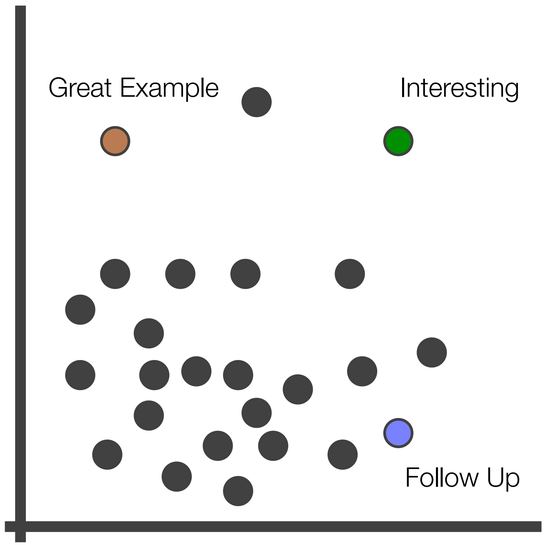
# WHAT ARE SELECTIONS?



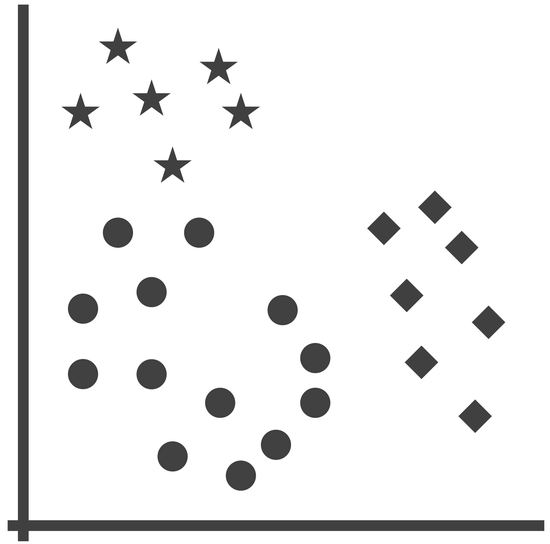
# SOME THINGS ARE BEST DONE THROUGH VISUALIZATION + INTERACTION



Filter

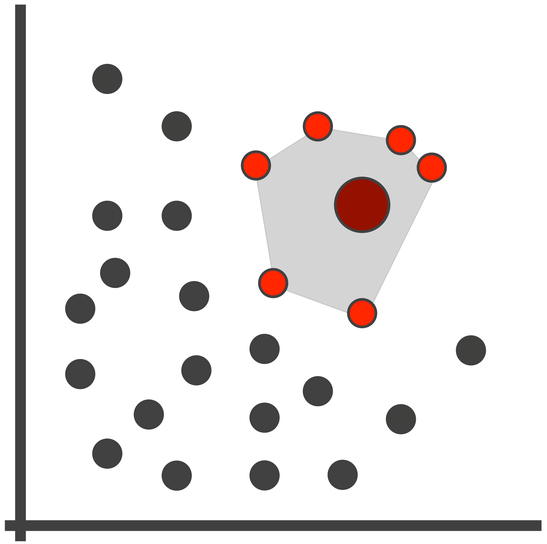



Label



Unassigned
Category A
Category B

Categorize



Aggregate

# EPHEMERAL INTERACTION

**BUT, unlike code interaction typically leaves no trace**

- **Not reproducible**
- **Not reusable**

**Interaction is also “siloed”**

- **Typically a dead end**

# INSPIRATION:

# LITERATE PROGRAMMING

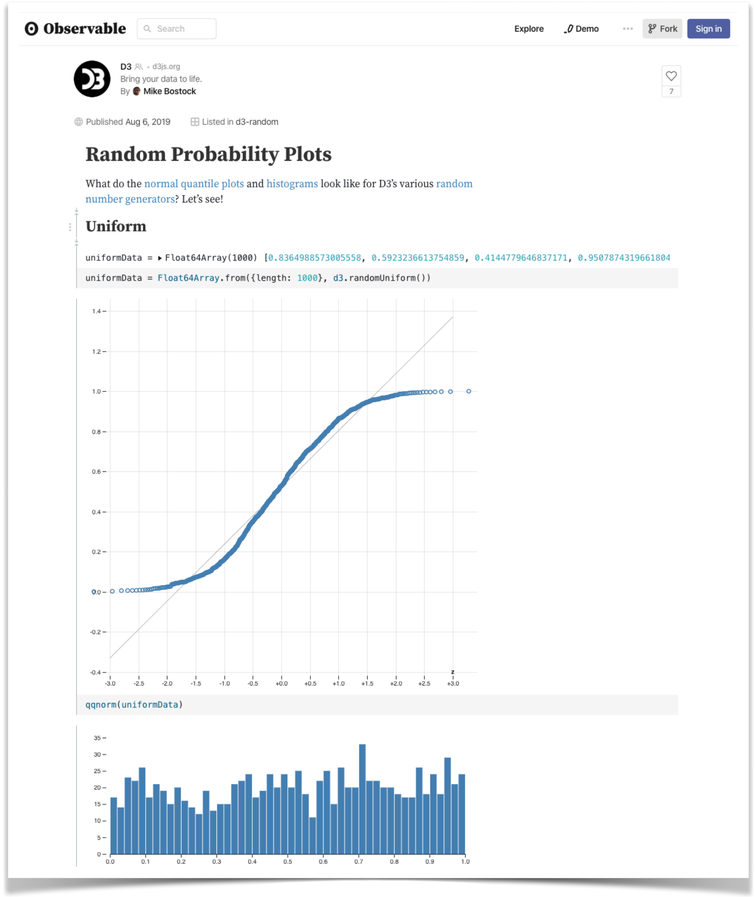
# LITERATE PROGRAMMING

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

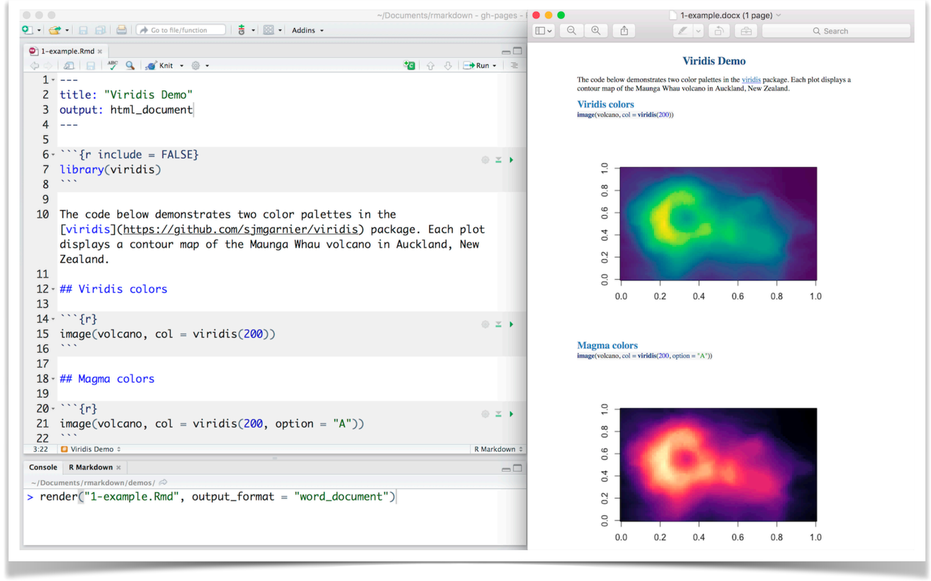
- Text
- Images / Visualizations
- Formulas
- Videos
- Links
- Code

[Donald E. Knuth, 1984]

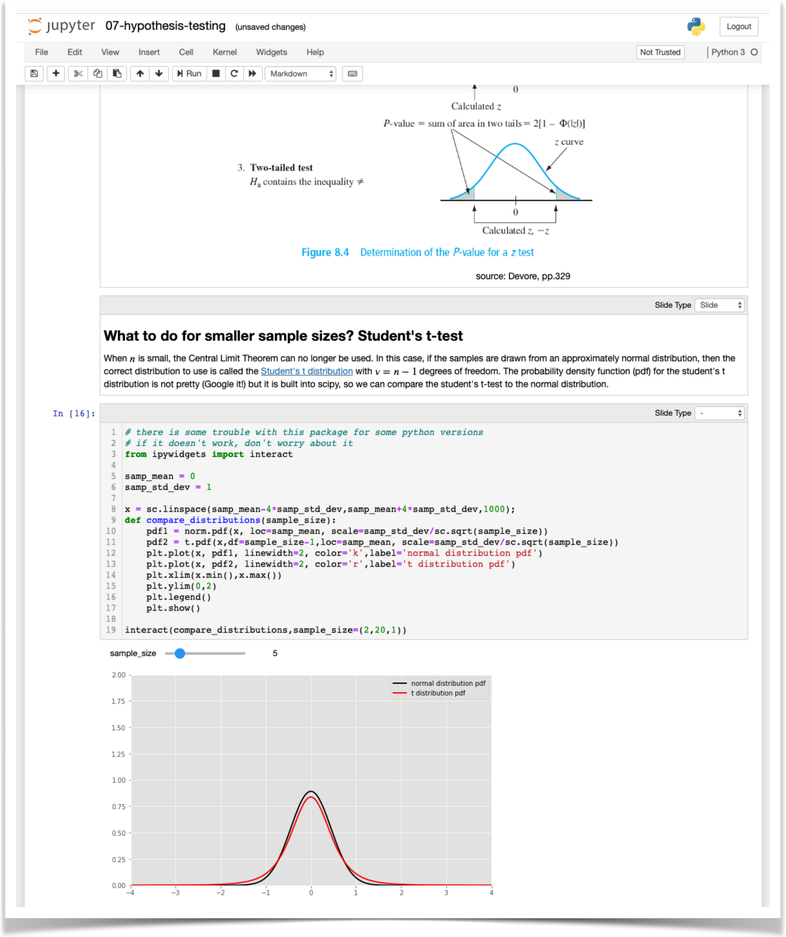
# LITERATE PROGRAMMING IN THE WILD




Observable




R Markdown




Jupyter Notebooks

THERE IS NO STRAIGHTFORWARD WAY TO DO

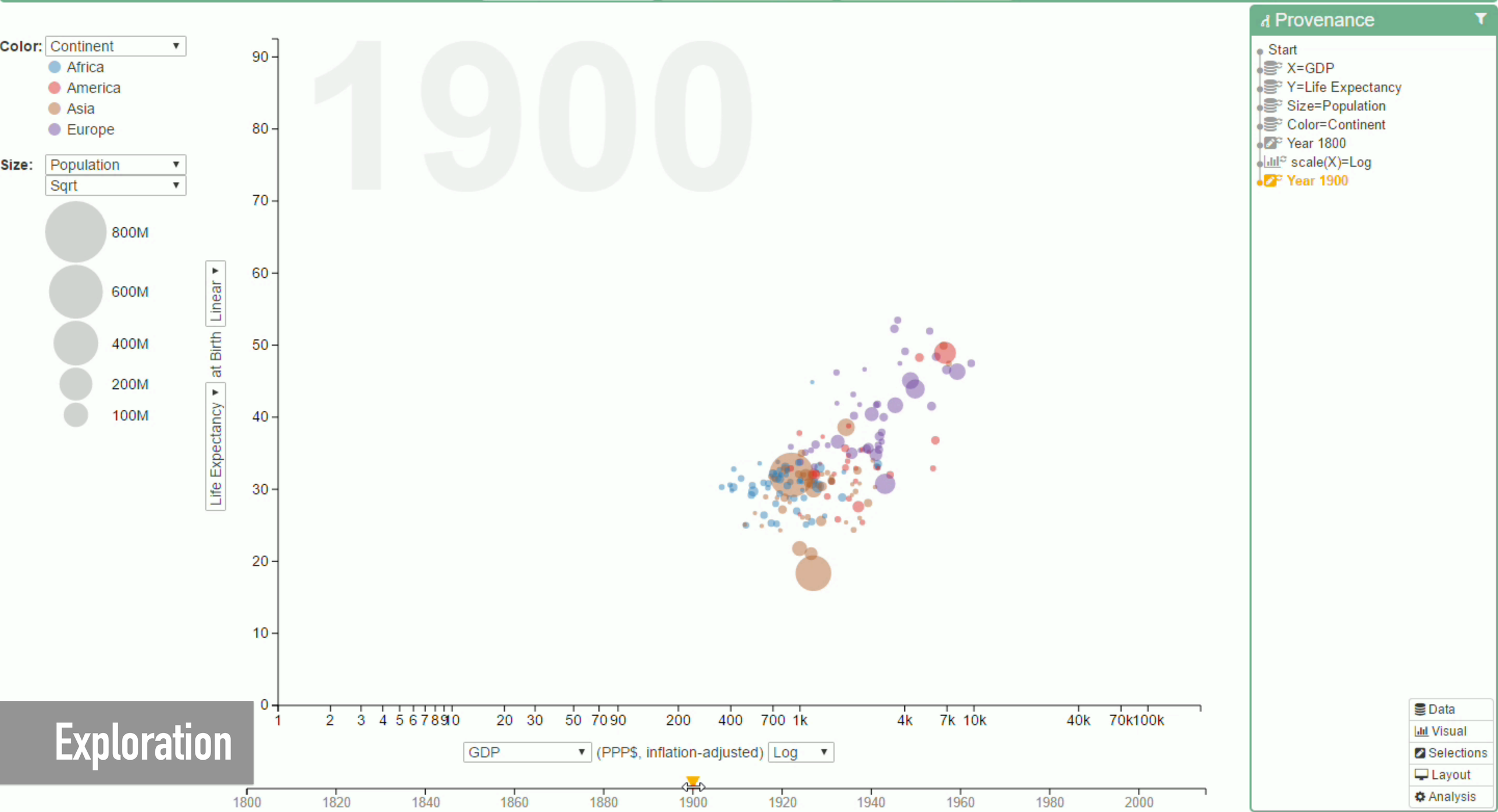
LITERATE DATA VISUALIZATION

# 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**

**Gratzl, Lex, Gehlenborg, Cosgrove, Streit; EuroVis 2016**

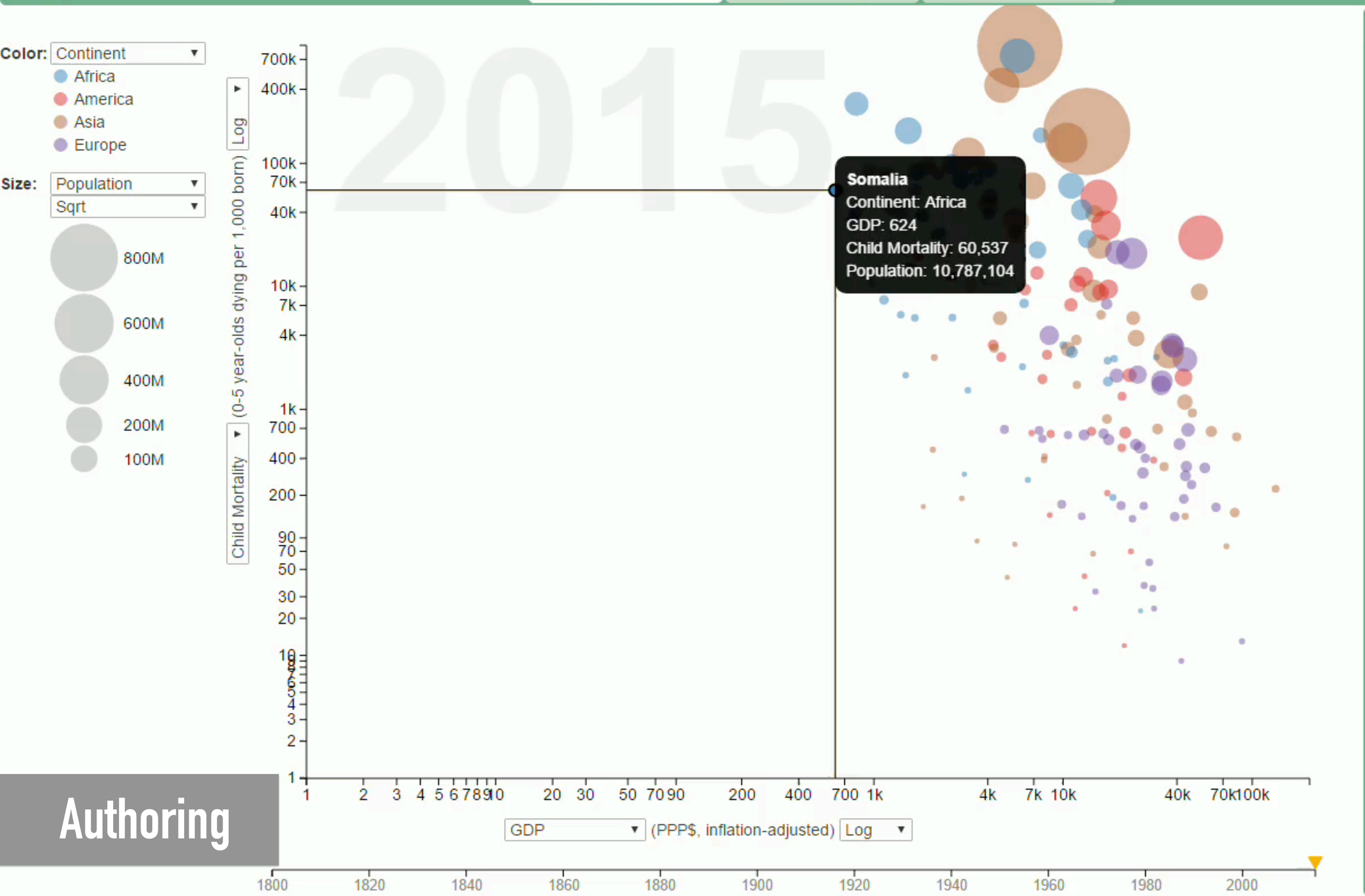



Provenance

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

Data
Visual
Selections
Layout
Analysis

# Exploration

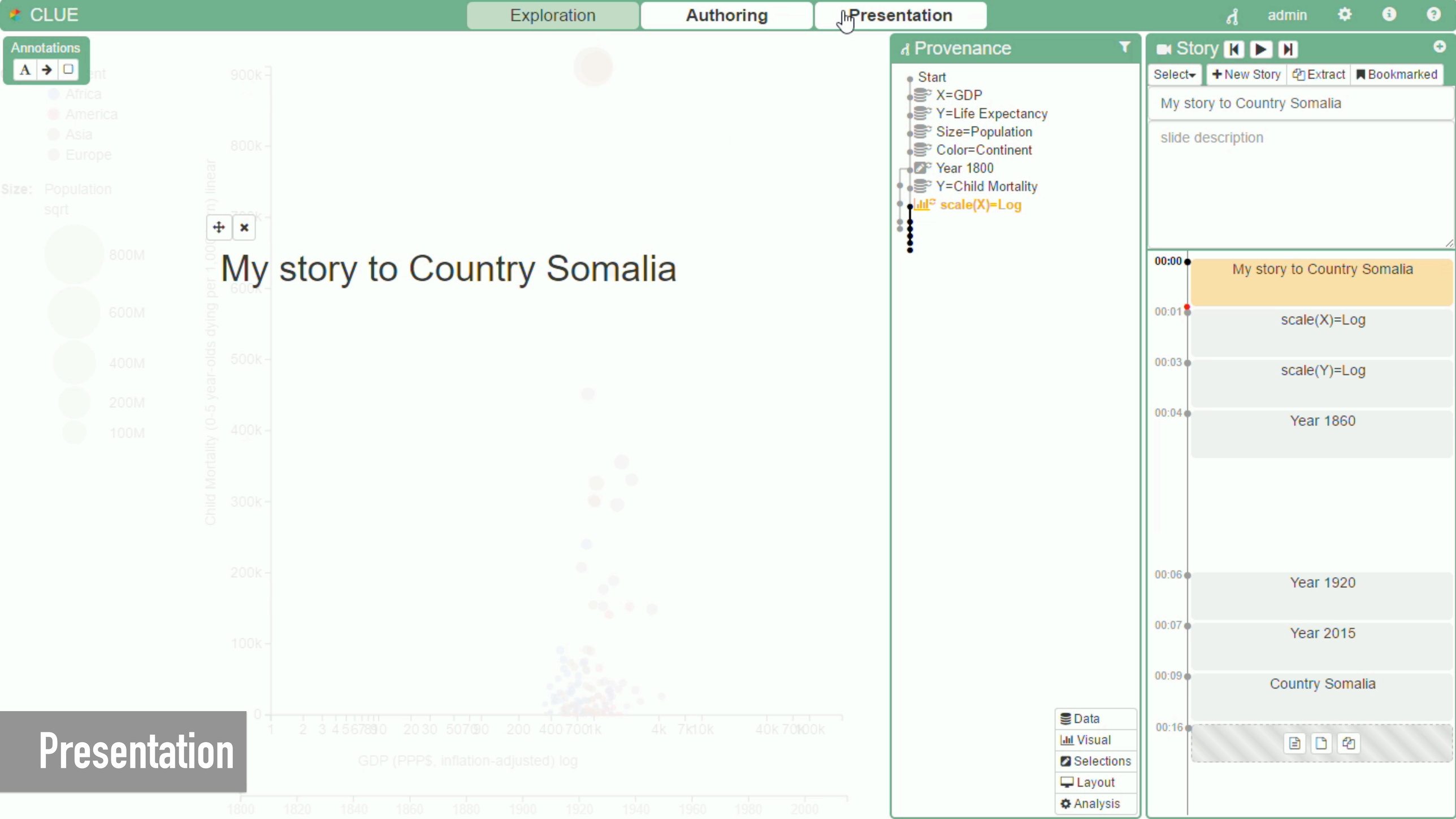



Provenance

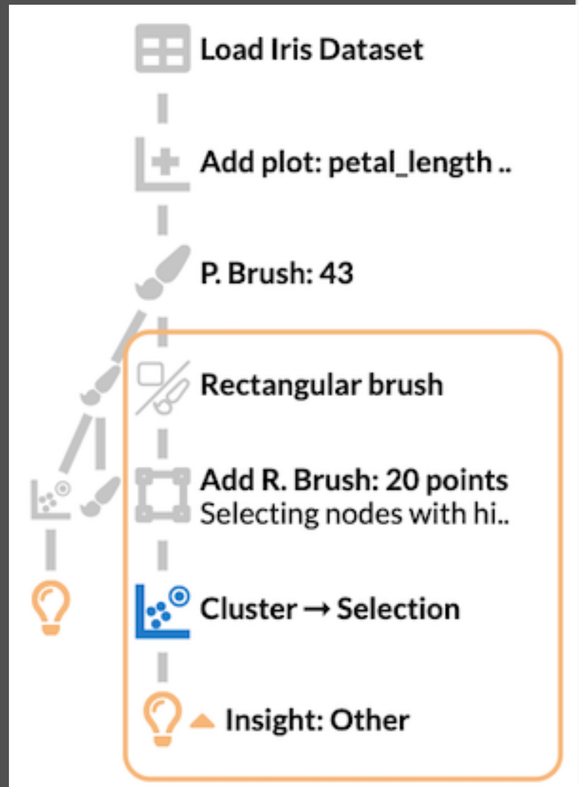
- Year 1800
- Y=Child Mortality
- scale(X)=Log
- scale(Y)=Log
- Year 1860
- Year 1920
- Year 2015
- Country Somalia

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

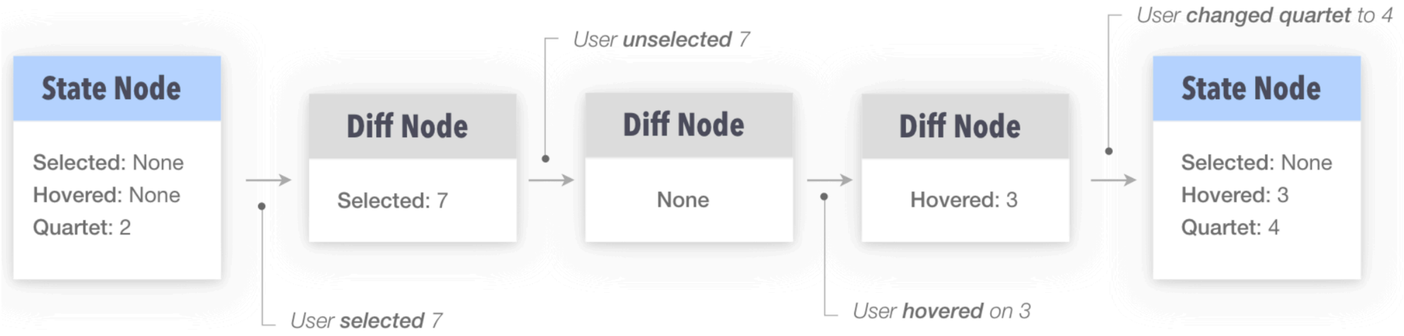
# TRRACK




## A web-based provenance library

## Easy to integrate in web apps

https://github.com/Trrack/trrackjs



# PROGRESS; BUT PROVENANCE IS "DUMB"

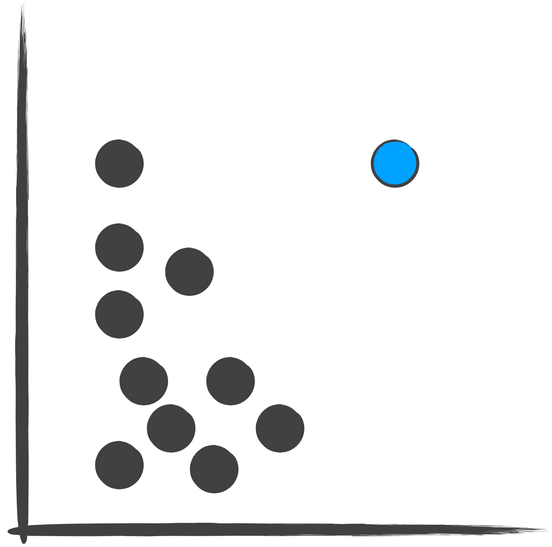
We solved the **WHAT,**
but not the **WHY**

**No progress towards reusability..**

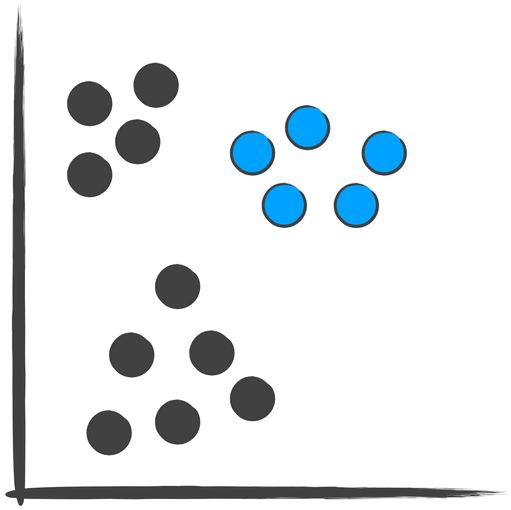
So, what else can we do?

# SEMANTIC SELECTIONS

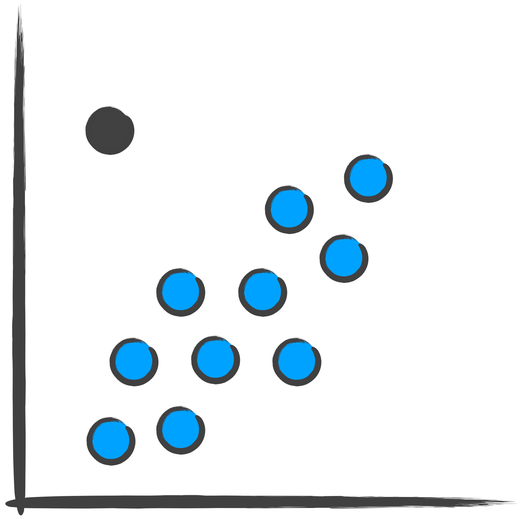
**Information Visualization 2021.** Kiran Gadhave, Jochen Görtler, Carolina Nobre, Oliver Deussen, Miriah Meyer, Jeff Phillips, Alexander Lex



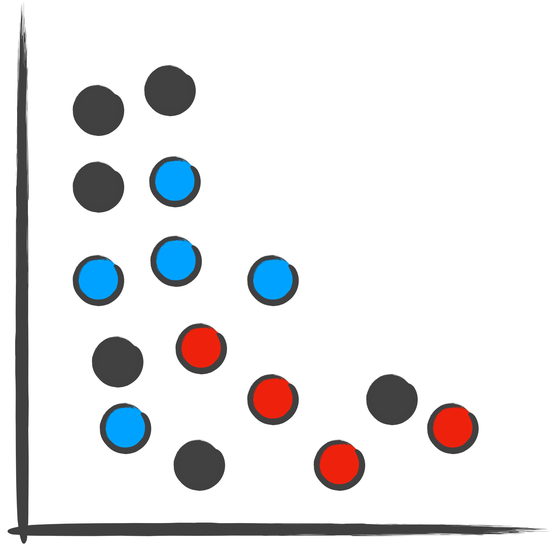
Outlier



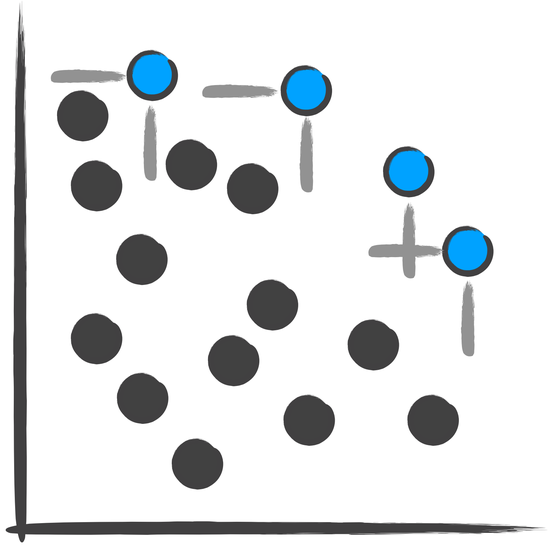
Clusters



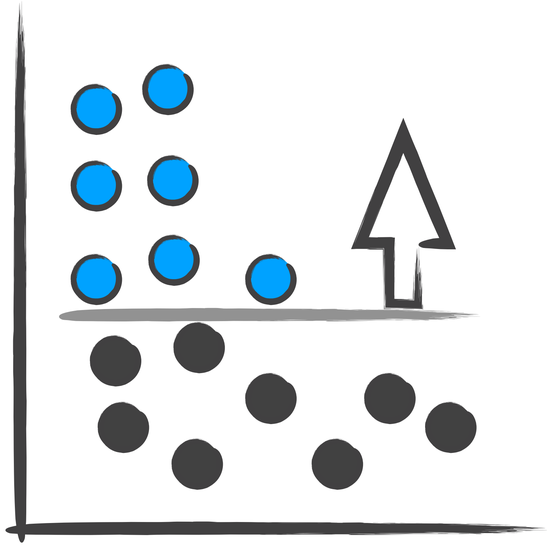
Correlation



Categories



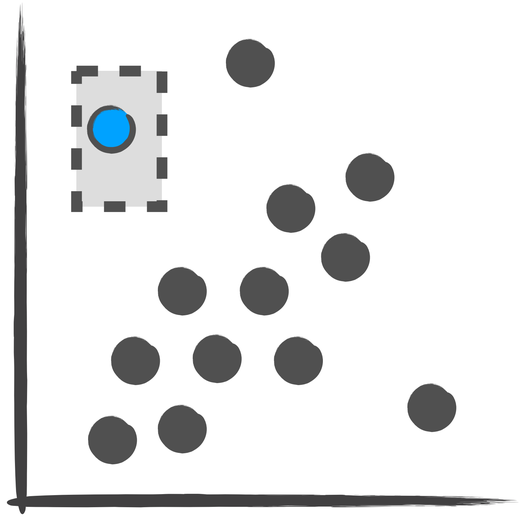
Multivariate Optimization



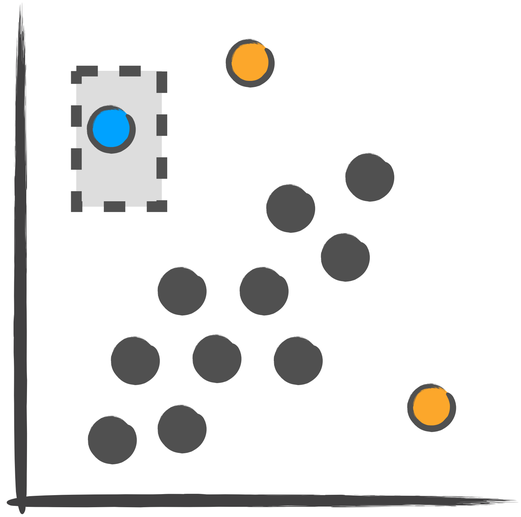
Ranges

# WHY DO WE CARE?

**Speed up** complex selections



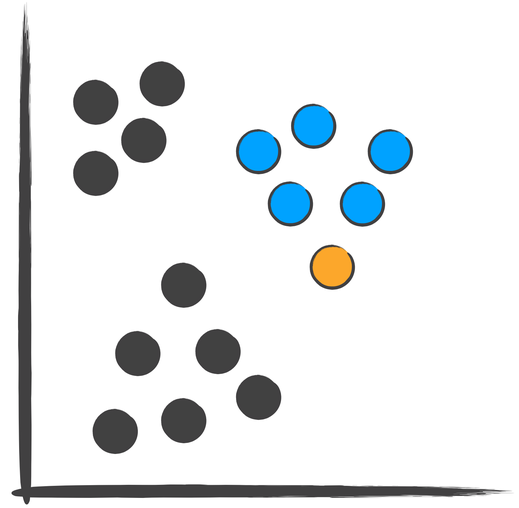
Selection



Outliers?

# WHY DO WE CARE?

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

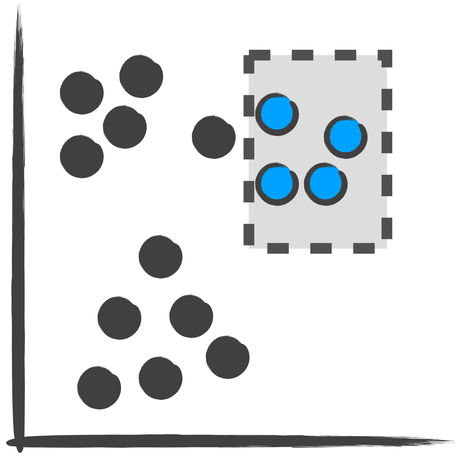


**Semantic Selection:**
Elements in K-Means cluster centered at [2, 3]

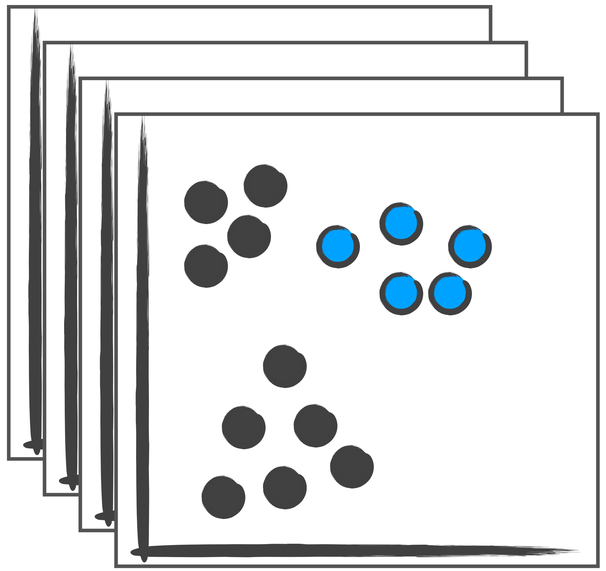
**Meaningful, higher level concept:**
**improves reproducibility**

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

# HOW DO WE INFER INTENT?



**Selection**



**Predictions**

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

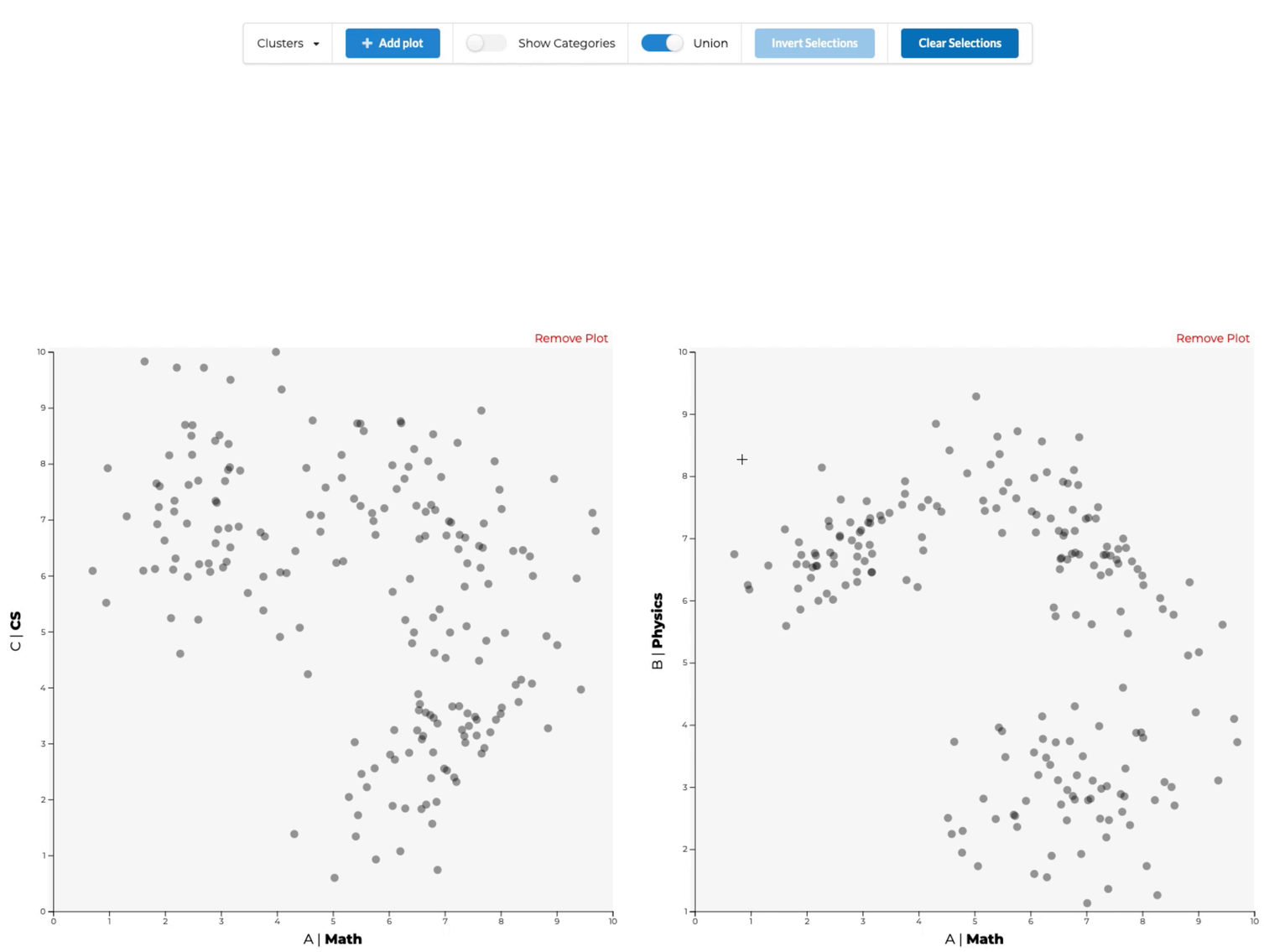
1. Range
2. Cluster
3. Outlier

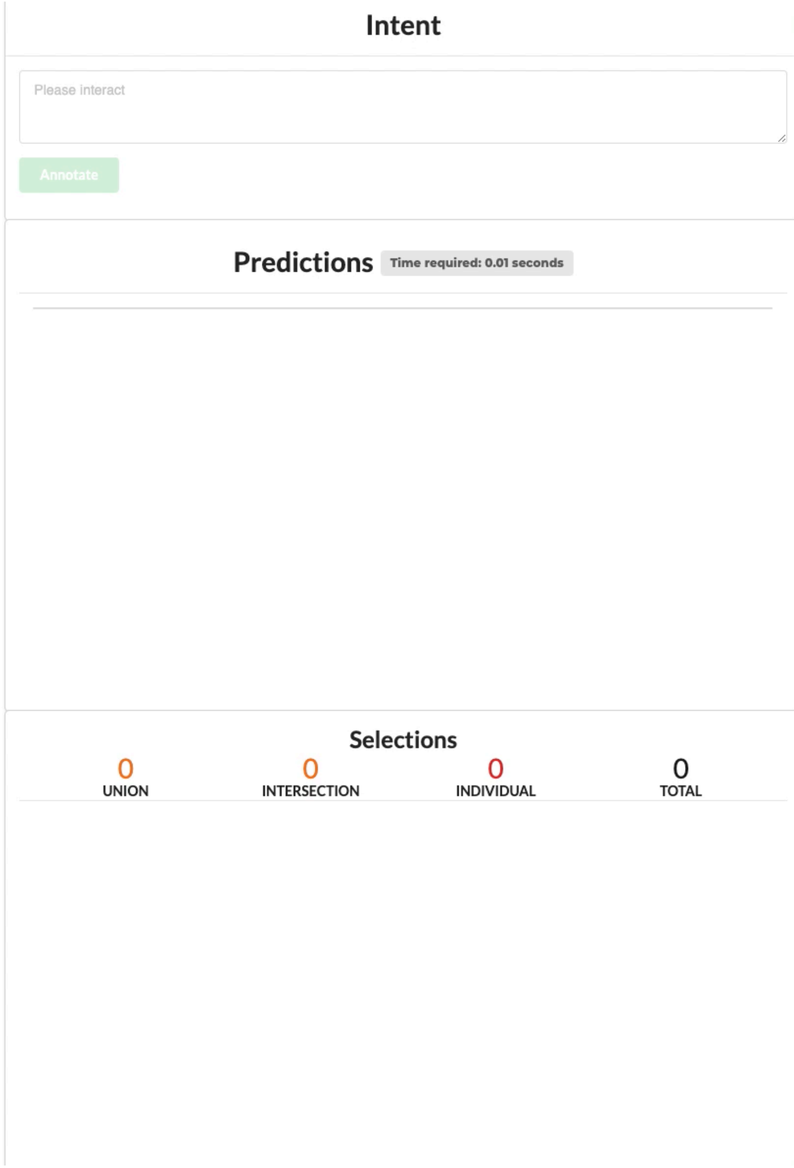
*I think this cluster…*

**Ranking**

Jaccard Distance
Naive Bayes
Classifier
Heuristic
Measures

**Confirming Intent & Annotation**




**Visualization and Selection**

**Annotation of Intent and Predictions**

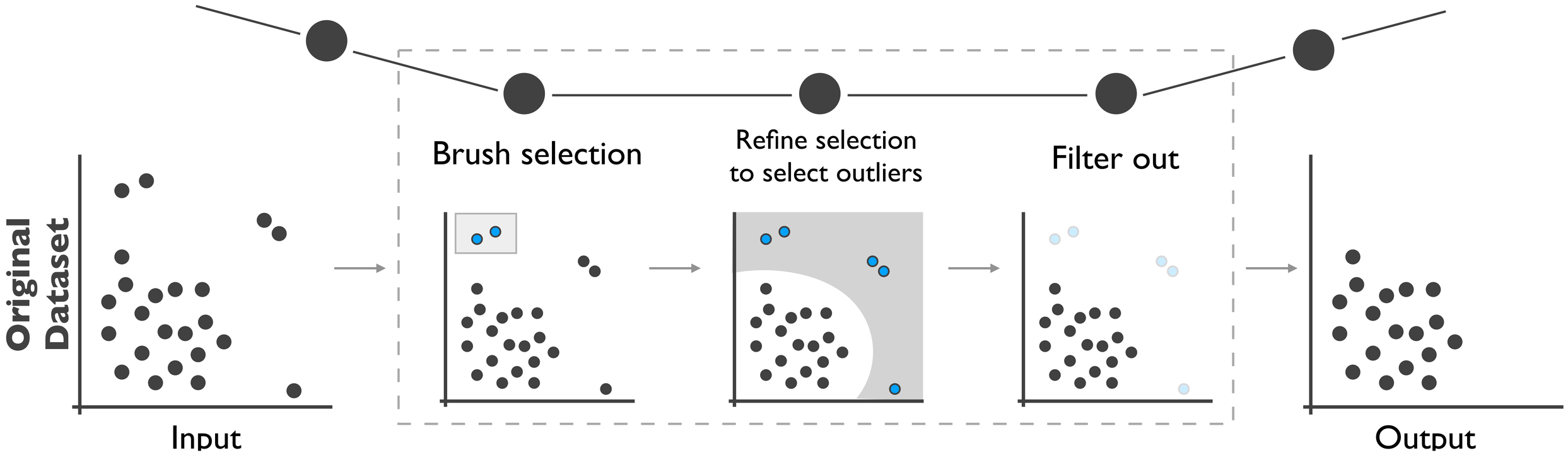
http://vdl.sci.utah.edu/predicting-intent/

# WORKFLOWS

**Based on semantic selections, we can create reusable workflows!**

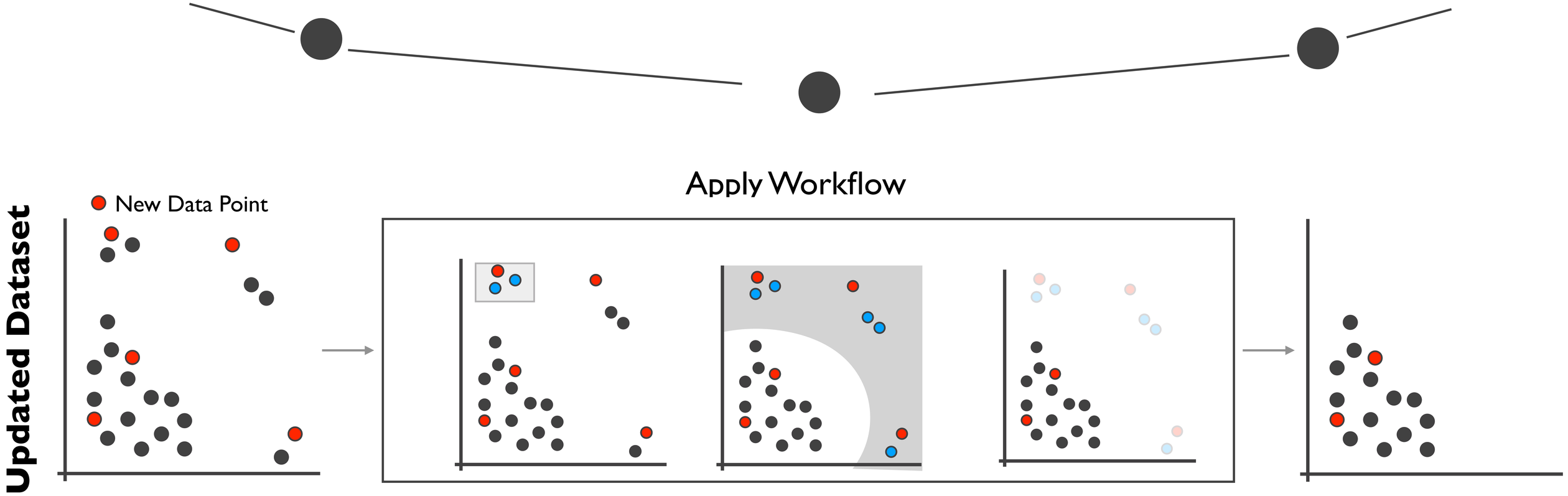
**EuroVis 2022.** Kiran Gadhave, Zach Cutler, Alexander Lex

# CAPTURING SEMANTICS OF WORKFLOWS

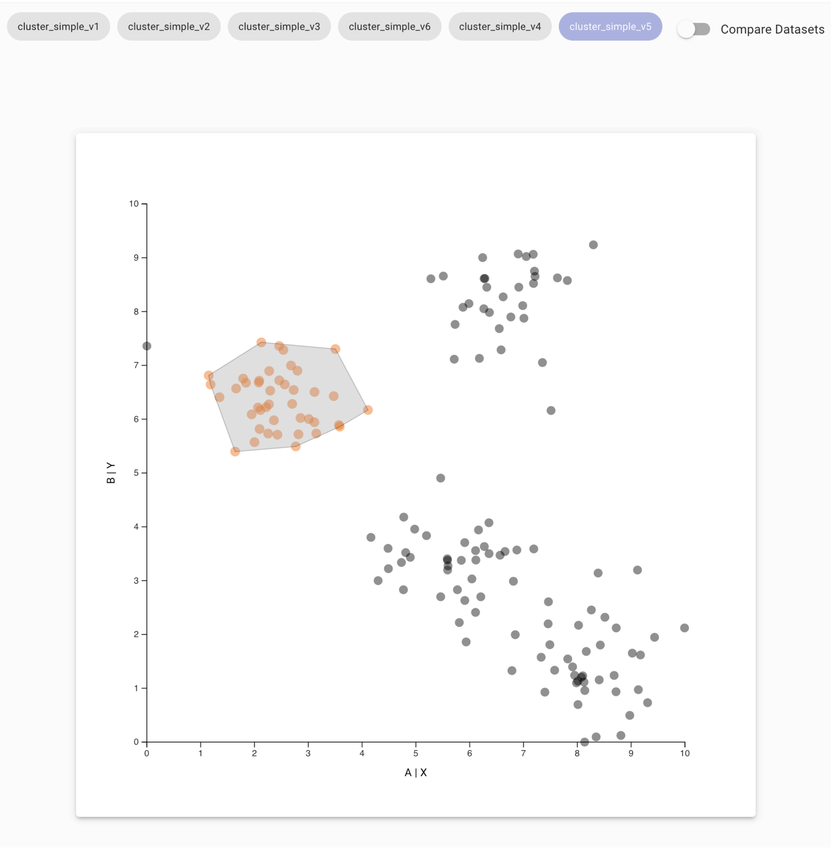



**Robust “Filter Outliers” Workflow**

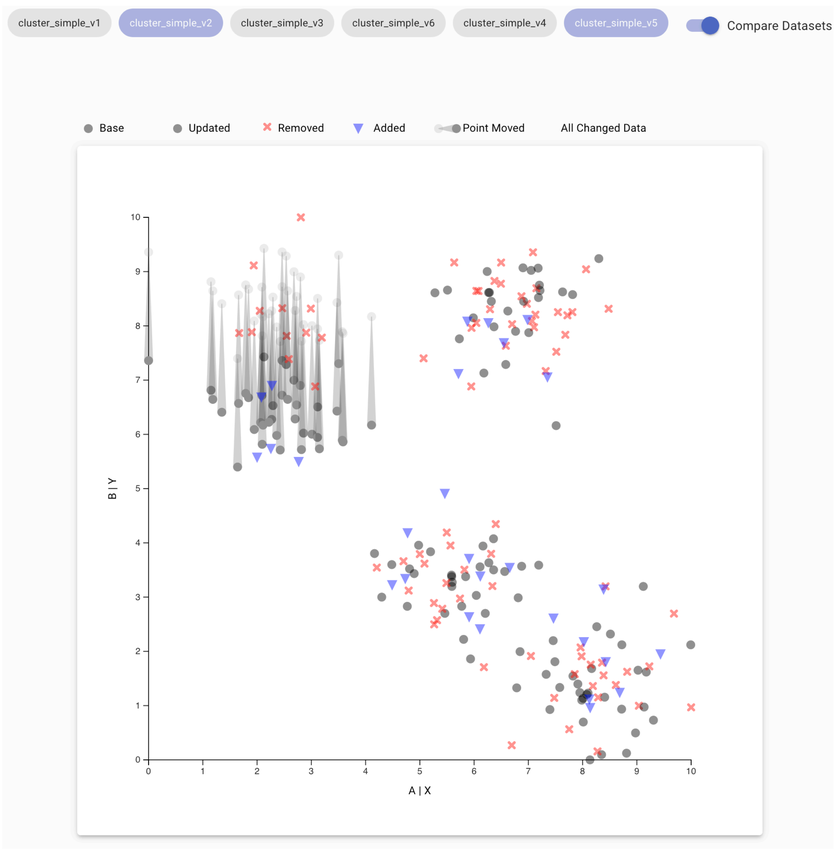
# REUSING WORKFLOWS ON UPDATED DATA



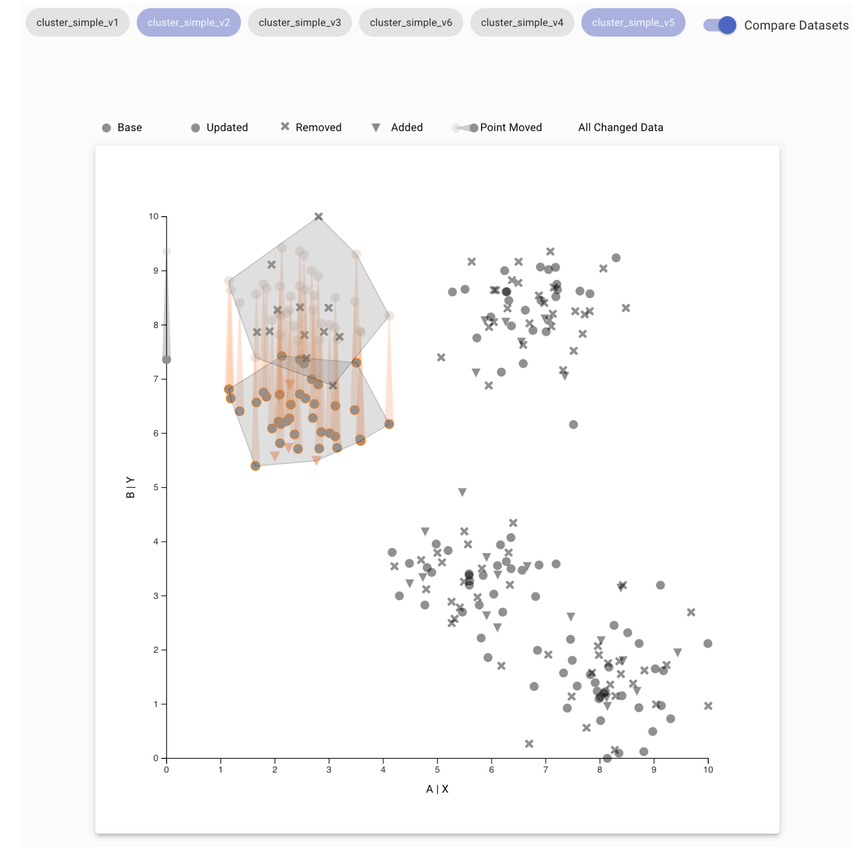
# REUSING SELECTIONS ON UPDATED DATASETS




Original Selection




Changed Dataset




Tracking A Selected Cluster

# NEXT: CAN WE ALL DO THIS WITH STANDARD PYTHON PLOTS?

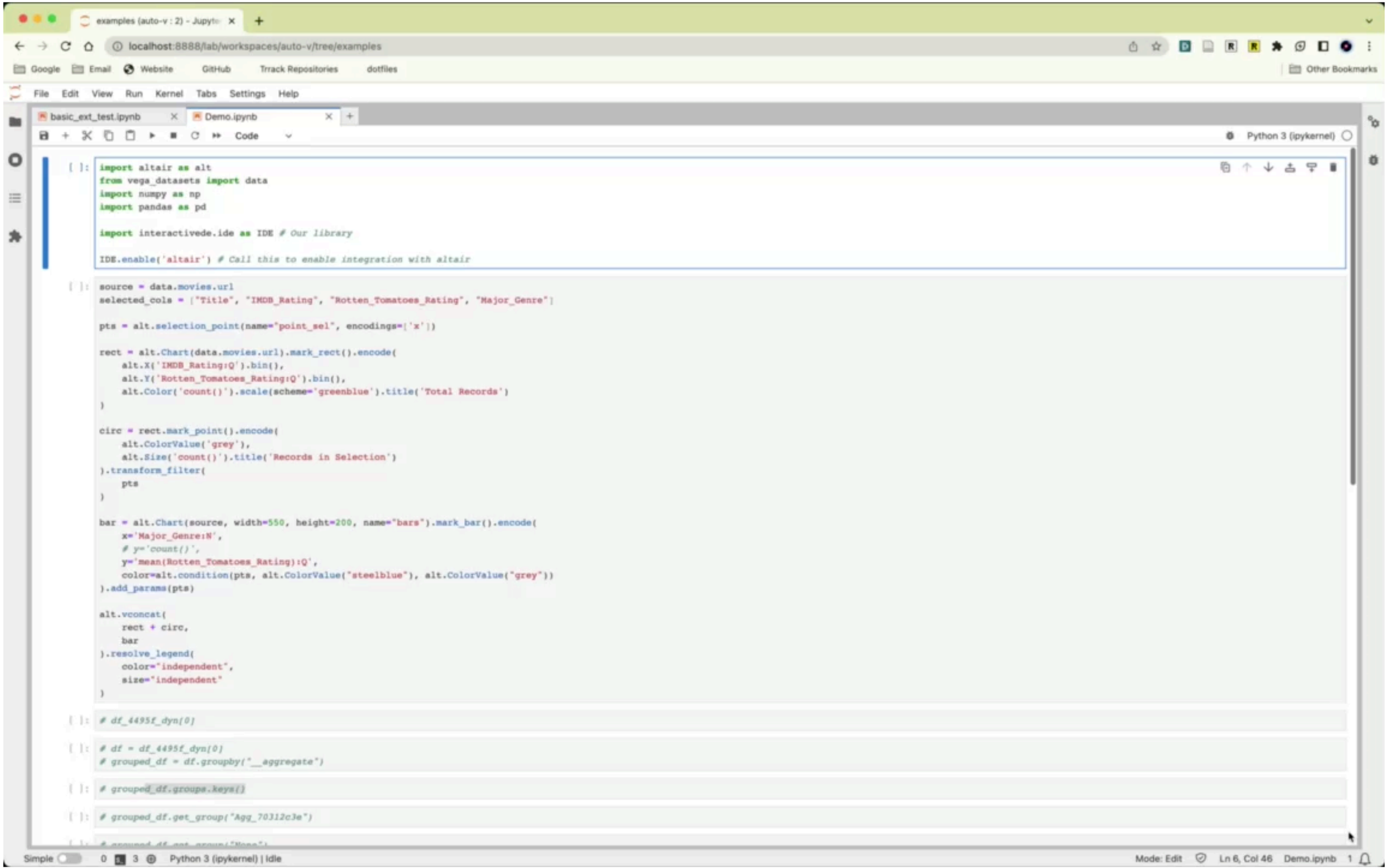
**Track interaction in native plots**

**Enable data wrangling operations (filter, label, aggregate, etc.)**

**Make steps permanent (like code)**

**Allow downstream use of modified data**

**Vega-Altair**


examples (auto-v : 2) - Jupyter
localhost:8888/lab/workspaces/auto-v/tree/examples
Google
Email
Website
GitHub
Trrack Repositories
dotfiles
Other Bookmarks
File Edit View Run Kernel Tabs Settings Help
basic_ext_test.ipynb
Demo.ipynb
Code
Python 3 (ipykernel)
[ ]:
import altair as alt
from vega_datasets import data
import numpy as np
import pandas as pd
import interactivede.ide as IDE # Our library
IDE.enable('altair') # Call this to enable integration with altair
[ ]:
source = data.movies.url
selected_cols = ["Title", "IMDB_Rating", "Rotten_Tomatoes_Rating", "Major_Genre"]
pts = alt.selection_point(name="point_sel", encodings=['x'])
rect = alt.Chart(data.movies.url).mark_rect().encode(
alt.X('IMDB_Rating:Q').bin(),
alt.Y('Rotten_Tomatoes_Rating:Q').bin(),
alt.Color('count()').scale(scheme='greenblue').title('Total Records')
)
circ = rect.mark_point().encode(
alt.ColorValue('grey'),
alt.Size('count()').title('Records in Selection')
).transform_filter(
pts
)
bar = alt.Chart(source, width=550, height=200, name="bars").mark_bar().encode(
x='Major_Genre:N',
# y='count()',
y='mean(Rotten_Tomatoes_Rating):Q',
color=alt.condition(pts, alt.ColorValue("steelblue"), alt.ColorValue("grey"))
).add_params(pts)
alt.vconcat(
rect + circ,
bar
).resolve_legend(
color="independent",
size="independent"
)
[ ]:
# df_4495f_dyn[0]
[ ]:
# df = df_4495f_dyn[0]
# grouped_df = df.groupby("__aggregate")
[ ]:
# grouped_df.groups.keys()
[ ]:
# grouped_df.get_group("Agg_70312c3e")
Simple
0 3 Python 3 (ipykernel) | Idle
Mode: Edit
Ln 6, Col 46
Demo.ipynb
1

# MAINTAINABLE SOFTWARE FITS INTO AN ECOSYSTEM

Fits perfectly in ecosystem of python data analysis

No installation burden

Interactive VIS is suddenly a first class operation in notebooks

1. SYSTEMS ARE HARD

2. INTERACTION IS POWERFUL BUT EPHEMERAL

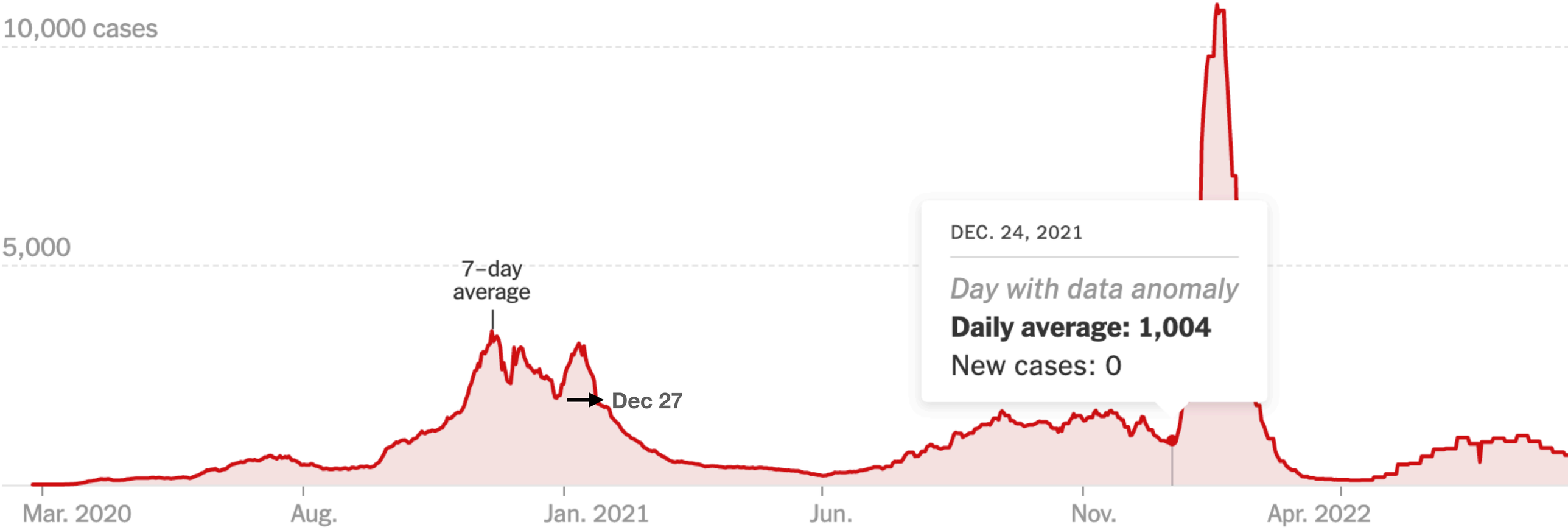
3. DATA IS NOT THE TRUTH

# 3. DATA IS NOT THE TRUTH

# Tracking Coronavirus in Utah: Latest Map and Case Count

Updated Aug. 11, 2022

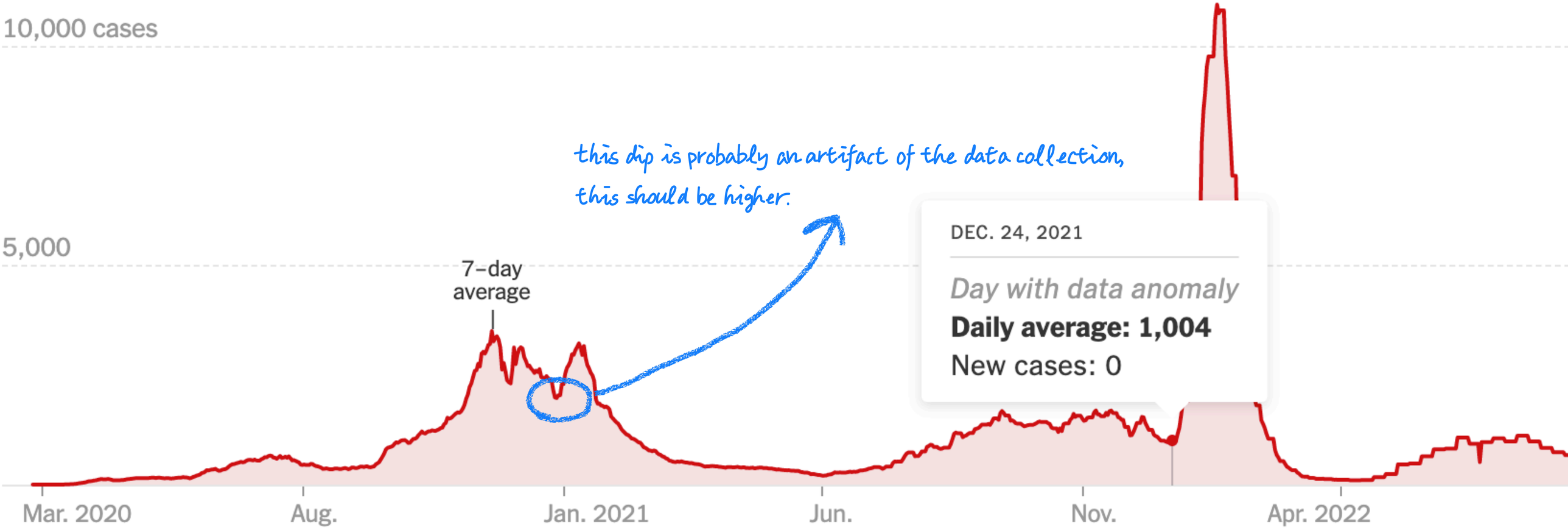
## New reported cases



# Tracking Coronavirus in Utah: Latest Map and Case Count

Updated Aug. 11, 2022

## New reported cases



Info Vis

*Article*

Information Visualization
2021, Vol. 20(2-3) 123–137

DOI: 10.1177/14738716211028565
journals.sagepub.com/home/ivi
SAGE

# Sanguine: Visual analysis for patient blood management

**Haihan Lin[1*], Ryan A Metcalf[2*], Jack Wilburn[1] and Alexander Lex[1]**

**Abstract**

Blood transfusion is a frequently performed medical procedure in surgical and nonsurgical contexts. Although it is often necessary or even life-saving, it has been identified as one of the most overused procedures in hospitals. Unnecessary transfusions not only waste resources but can also be detrimental to patient outcomes. Patient blood management (PBM) is the clinical practice of optimizing transfusions and associated outcomes. In this paper, we introduce Sanguine, a visual analysis tool for transfusion data and related patient medical records. Sanguine was designed with two user groups in mind: PBM experts who oversee blood management practices across an institution and clinicians performing transfusions. PBM experts use Sanguine to explore and analyze transfusion practices and their associated medical outcomes. They can compare individual surgeons, or compare outcomes or time periods, such as before and after an intervention regarding transfusion practices. PBM experts then curate and annotate views for communication with clinicians, with the goal of improving their transfusion practices. We validate the utility and effectiveness of Sanguine through case studies.

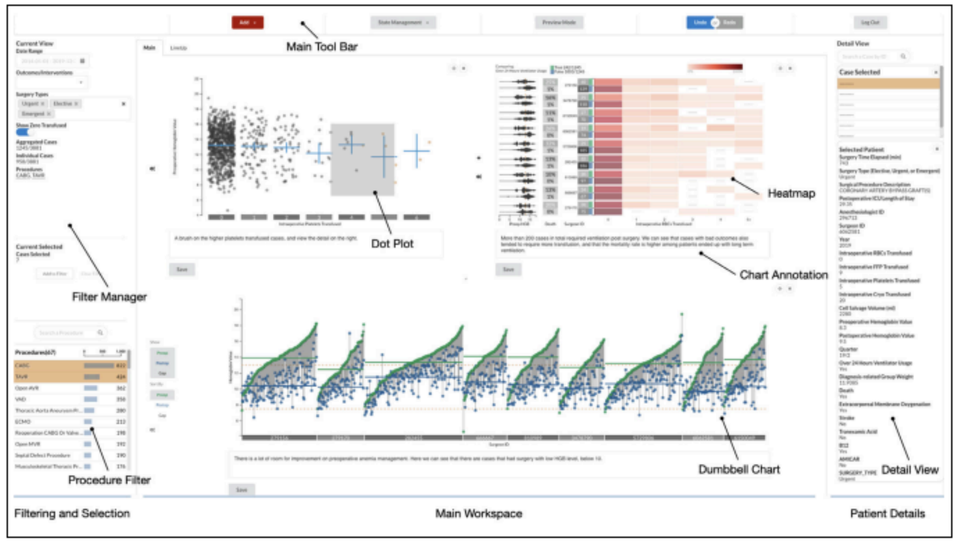
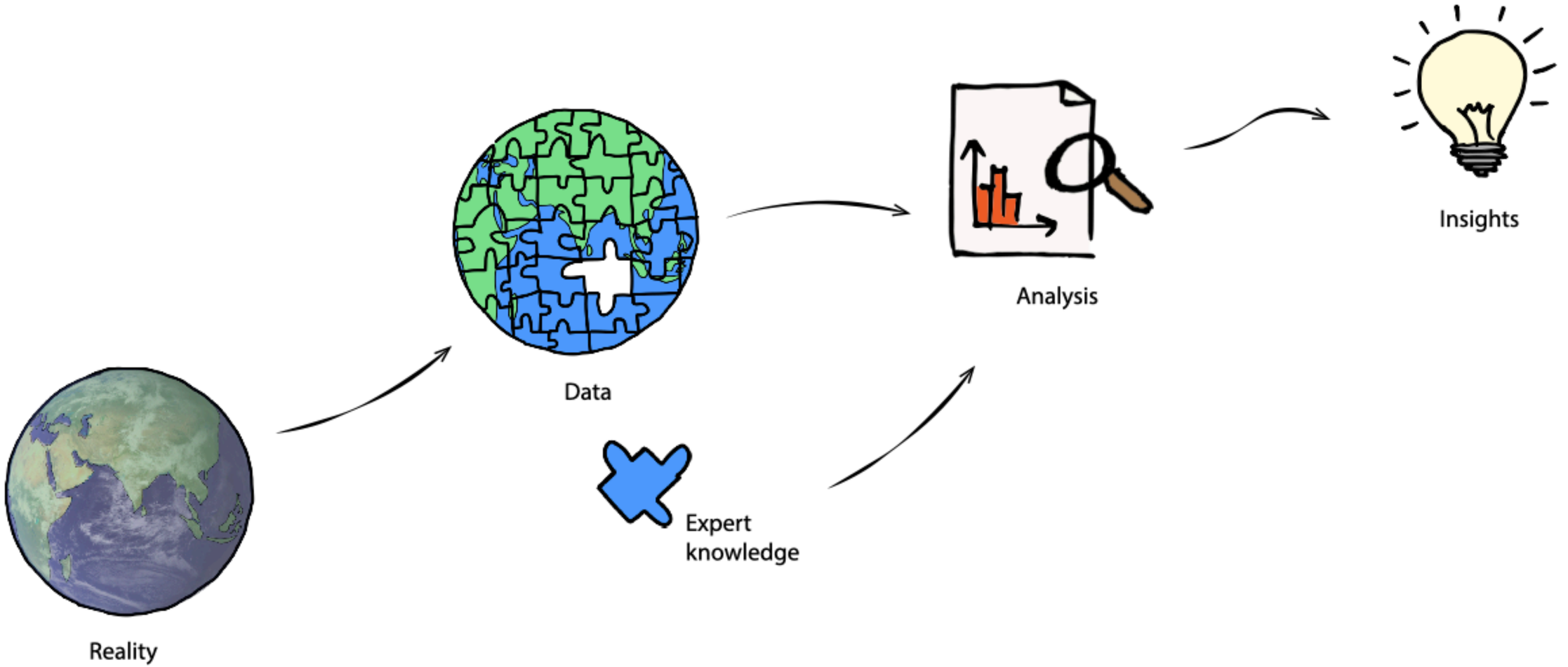



**Figure 1.** An overview of Sanguine visualizing patient blood management data with multiple views. The left panel is dedicated to managing filters and selections. The workspace in the center contains visualizations that can be flexibly arranged. A heatmap, a dot plot, and a dumbbell chart are shown. On the right, a patient-specific detail view shows attributes of a case.

The visualization showed many cases not using blood recycling. But to my knowledge, we almost always turn on the machine for it.

Likely, sometimes they were not recorded.


Reality
Data
Expert
knowledge
Analysis
Insights

# IMPLICIT KNOWLEDGE

Data is imperfect

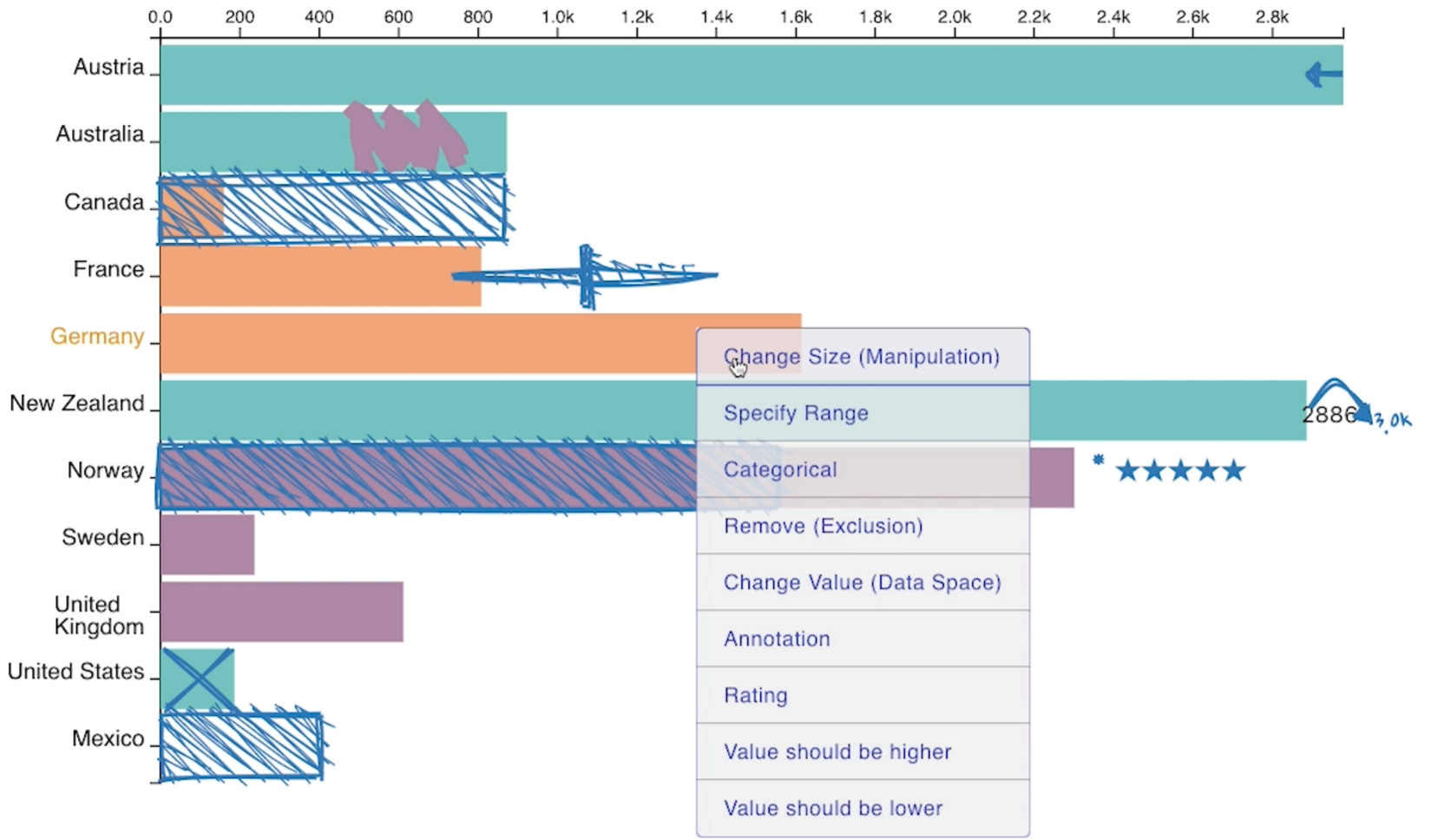
but experts know how and why

Challenges:

- Hardly documented (lab notebooks, methods sections)
- Not saliently available to others
- Bad tool support
- Limited metadata formats

A **DATA HUNCH** IS AN ANALYST'S KNOWLEDGE ABOUT HOW AND WHY THE DATA IS AN IMPERFECT AND PARTIAL REPRESENTATION OF THE PHENOMENA OF INTEREST.

Data Hunches | Add New Value (Inclusion) | Transform Data | Add Sketch | Show Category | Show Table | Select Data: COVID Cases | Data Hunch Demos: proposal talk




Add an annotation about the chart

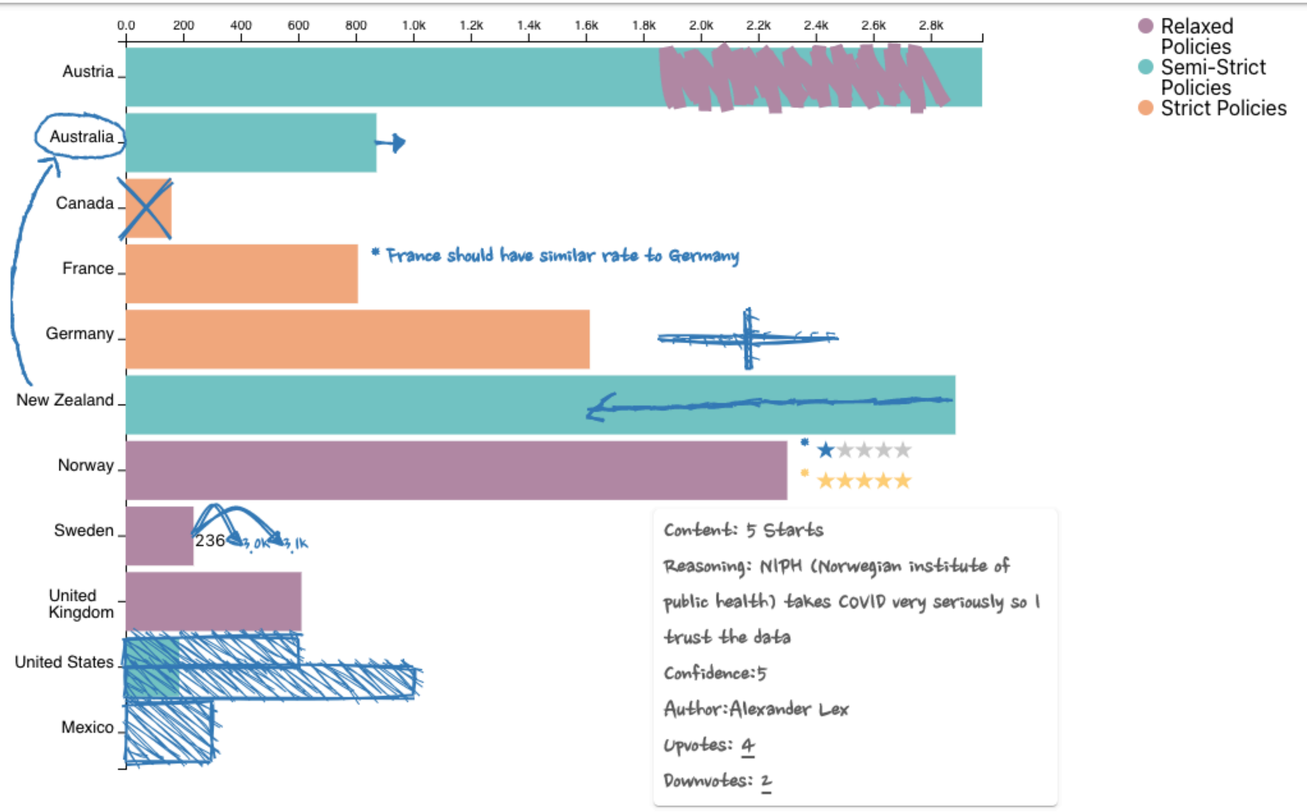
# COVID Cases

New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people, colored by stringency of country's response as of March 01, 2022. Strict Policies stringency indicates a stricter response. If policies vary at the subnational level, the result is shown as the response level of the strictest sub-region. Data shows Mar 01, 2022. Data Source: OurWorldInData

*The categorical labels do not help much

# RATING, COMMENTING, COLLABORATING

Data Hunches | Add New Value (Inclusion) | Transform Data | Add Sketch | Show Category | Hide Table | Select Data: COVID Cases | Data Hunch Demos: All type data hunches




Add an annotation about the chart

## COVID Cases

New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people, colored by stringency of country's response as of March 01, 2022. Strict Policies stringency indicates a stricter response. If policies vary at the subnational level, the result is shown as the response level of the strictest sub-region. Data shows Mar 01, 2022. Data Source: OurWorldInData

*The categorical labels do not help much

*sketch

| Type | Userna... | Label | Reasoning | Content |
|---|---|---|---|---|
| manipulati... | Haihan Lin | United St... | Many states do not ... | 600 |
| data space | Miriah Meyer | Sweden | Sweden does not t... | 3000 |
| data space | Alexander Lex | Sweden | Sweden does not t... | 3100 |
| categorical | Alexander Lex | Austria | Austria has similar ... | Relaxed |
| range | Derya Akbaba | Germany | Germany might hav... | 1856.25 |
| data space | Derya Akbaba | United St... | United States do no... | 1000 |
| rating | Miriah Meyer | Norway | I do not trust Norwa... | 1 |
| rating | Alexander Lex | Norway | NIPH (Norwegian i... | 5 |
| direction | Haihan Lin | Australia | I think there are mo... | higher |
| annotation | Alexander Lex | France | They are close to e... | France s |
| annotation | Miriah Meyer | all chart | This looks like a ve... | The cate |
| inclusion | Haihan Lin | Mexico | A conjecture about ... | 300 |
| exclusion | Haihan Lin | Canada | Canada has much ... | Canada |

Rows per page: 100 | 1-14 of 14

# CHALLENGES

## Biases and Trust

- Reinforcing preconceived ideas?
- Explain away “inconvenient” data points?

Need to provide reasoning and justifications

Intended for trusted teams

# IMPLICATIONS

Currently a stand-alone tool

Goals:

integrate with plotting library, such as Vega-Altair (see “systems are hard)

preserve hunches across data structures and analysis steps

Many other efforts needed to address the gap between data and truth

1. SYSTEMS ARE HARD

2. INTERACTION IS POWERFUL BUT EPHEMERAL

3. DATA IS NOT THE TRUTH

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


Thanks to: **Kiran Gadhave, Hainan Lin, Zach Cutler**, Devin Lange, Max Lisnic, 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