Lecture 6
Filter + Aggregate

DSCI 532, Data Visualization II

January 21, 2019

Cydney Nielsen

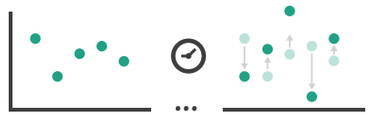
Senior Designer, Microsoft

Adjunct Professor, UBC Department of Computer Science

Dealing with complexity

Manipulate

- Change

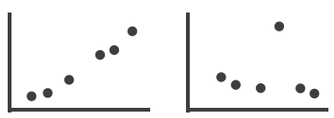


- Select

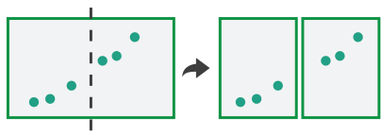
- Navigate

Facet

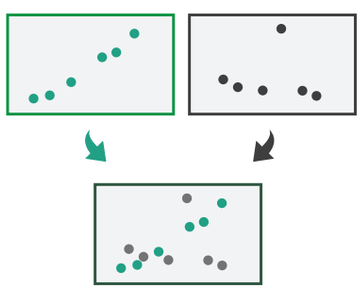
- Juxtapose



- Partition



- Superimpose

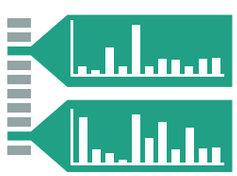


Reduce → *Derive*

- Filter

- Aggregate

Embed



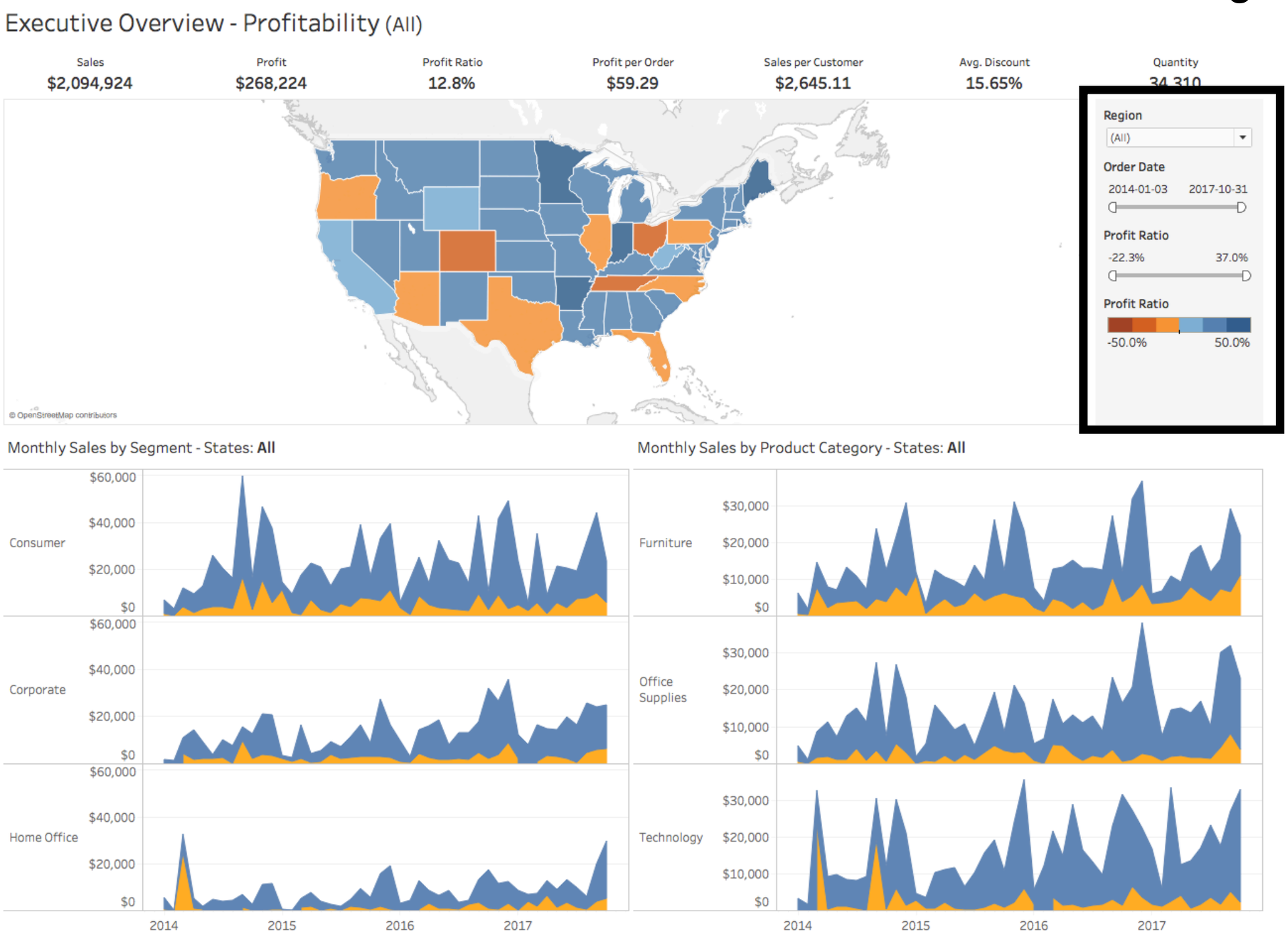
Visualization Analysis and Design
Tamara Munzner

Filter

reducing complexity and creating focus

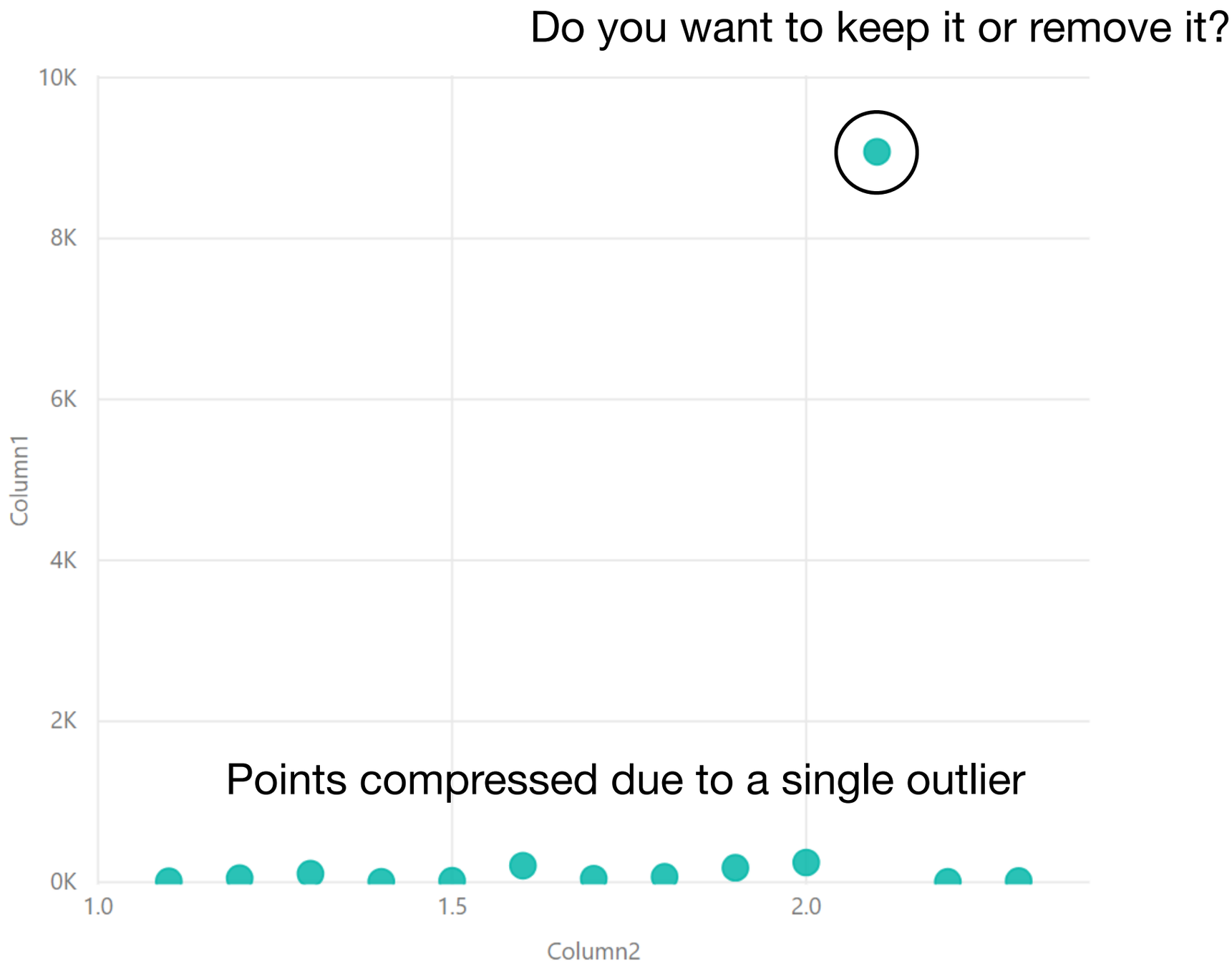
Dashboards

Filter widgets



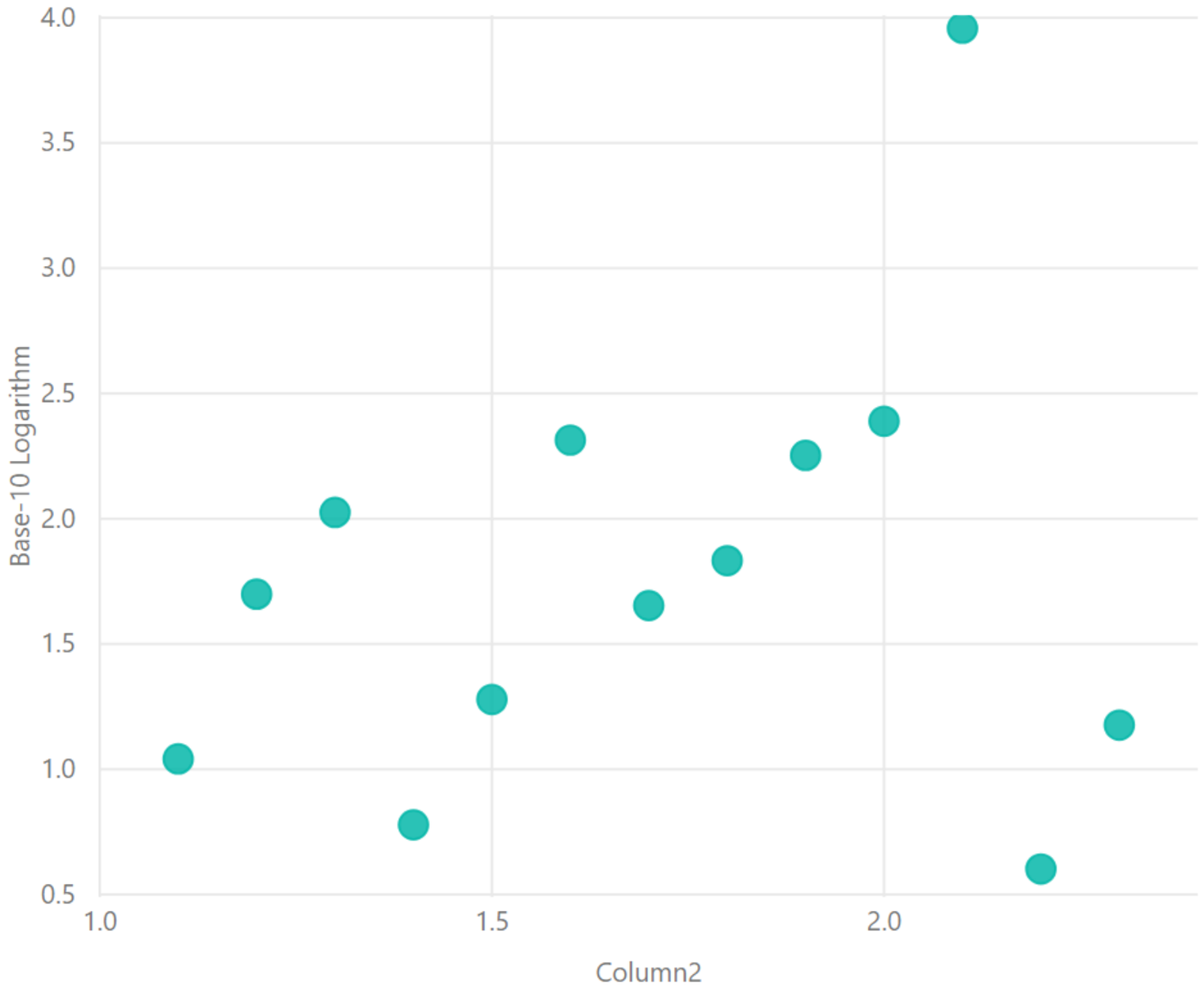
Tableau

Dealing with outliers



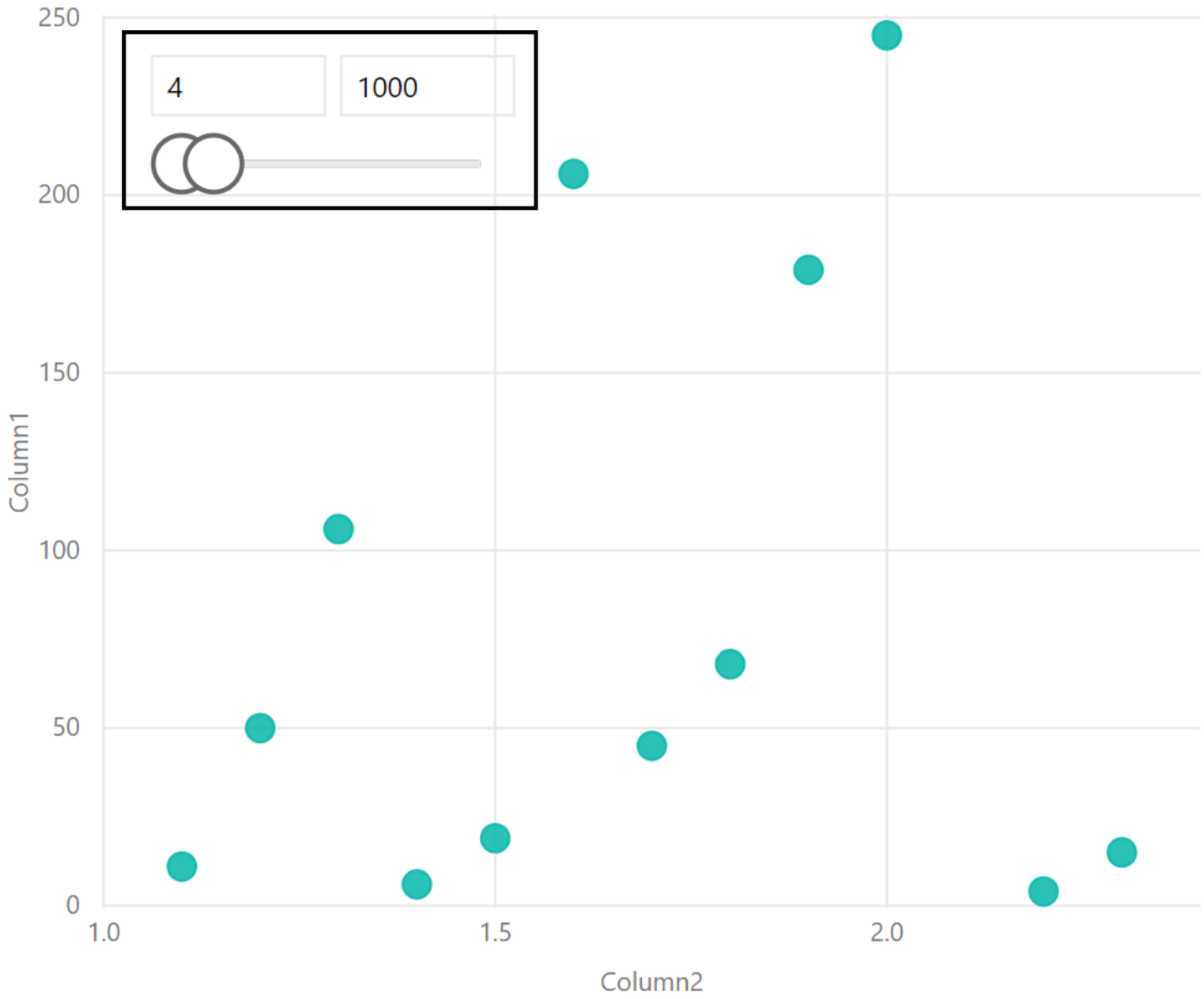
Dealing with outliers

log scale helps - *appropriate if you want to keep all points*



Dealing with outliers

interactive filtering is another solution - *appropriate when want to remove outliers*



Faceted search

Home / All / Camping and hiking / Packs and bags

CATEGORY

- All
  - Camping and hiking
    - **Packs and bags (92)**
      - Backpacking packs (56)
      - Daypacks (50)
      - Hydration packs and accessories (2)
      - Child carrier backpacks (1)

FEATURES

- [ ] Compression straps (83)
- [ ] Hydration compatible (73)
- [ ] Ice axe/pole attachment (70)
- [ ] Hipbelt pocket(s) (59)
- [ ] Ventilated back panel (44)

MORE

VOLUME

- [x] **30 to 44 L (68)**
- [x] **45 to 59 L (31)**
- [ ] 20 to 29 L (64)
- [ ] 10 to 19 L (46)
- [ ] 60 to 74 L (36)

MORE

HIKING AND CAMPING PACKS

92 items

30 to 44 L ⊗ 45 to 59 L ⊗

SAVE 30%

★★★☆☆ (2)

Gregory Targhee 45 Backpack - Unisex

$181.00 | ~~$259.00~~

☐ Compare 🛒 Add to cart

SAVE 20%

★★★★☆ (3)

Gregory Zulu 40 Backpack - Unisex

$167.95 | ~~$209.00~~

☐ Compare 🛒 Add to cart

SAVE 20%

★★★★½ (9)

Gregory Jade 28 Daypack - Women's

$127.95 | ~~$159.00~~

☐ Compare 🛒 Add to cart

SAVE 29%

★★★☆☆ (13)

SAVE 20%

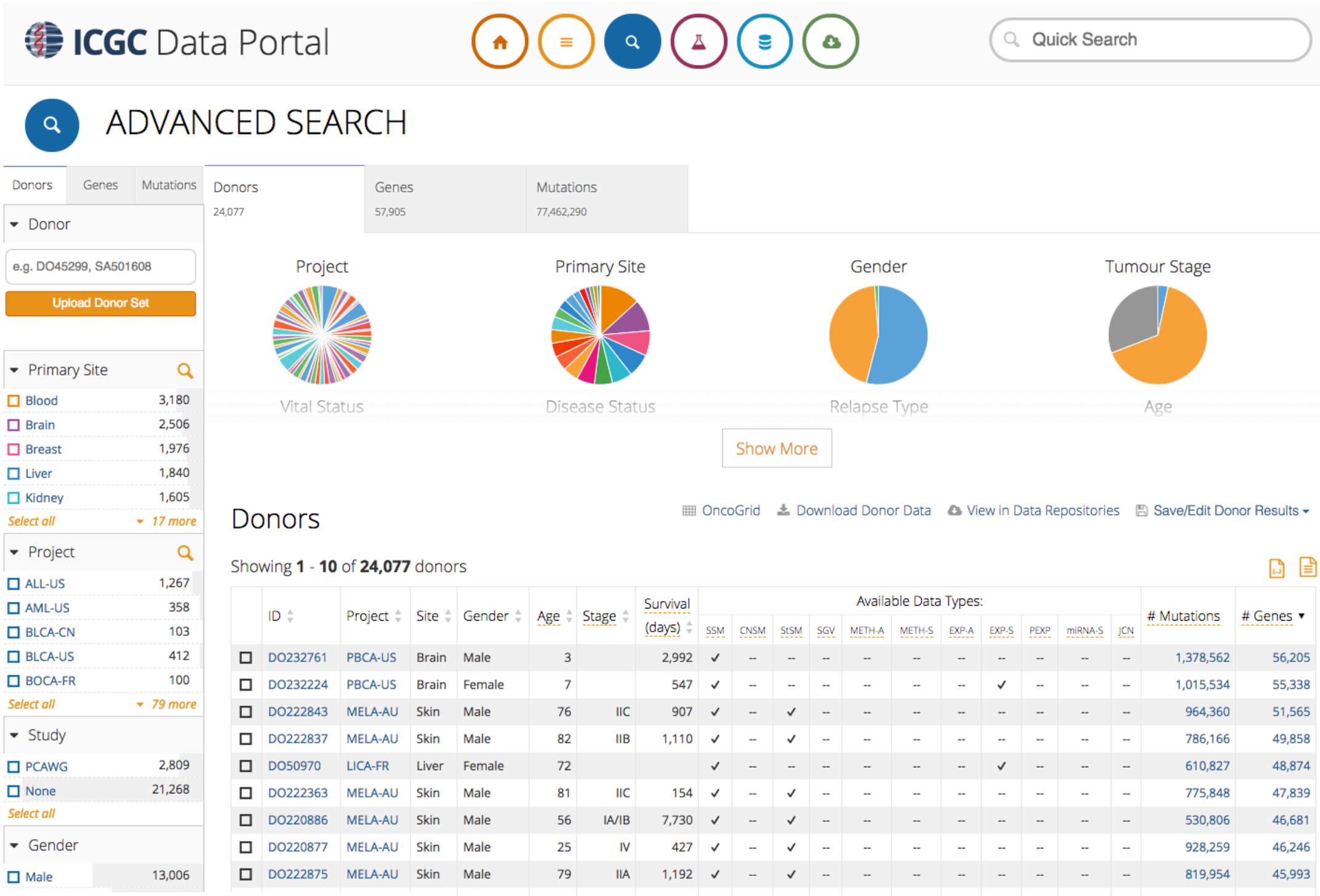
☆☆☆☆☆ (0)

Osprey Kestrel 48L Backpack - Men's

SAVE 20%

★★★★½ (4)

Gregory Zulu 55 Backpack - Unisex



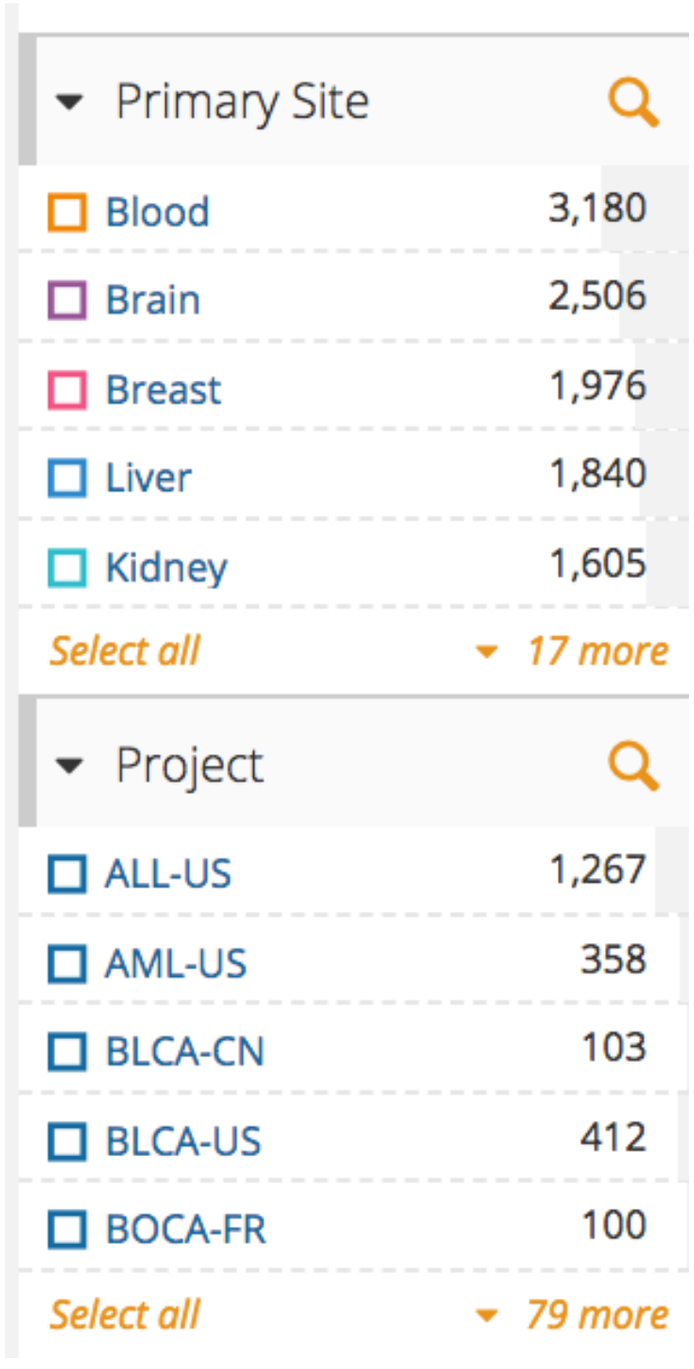
What works well?

What could be improved?

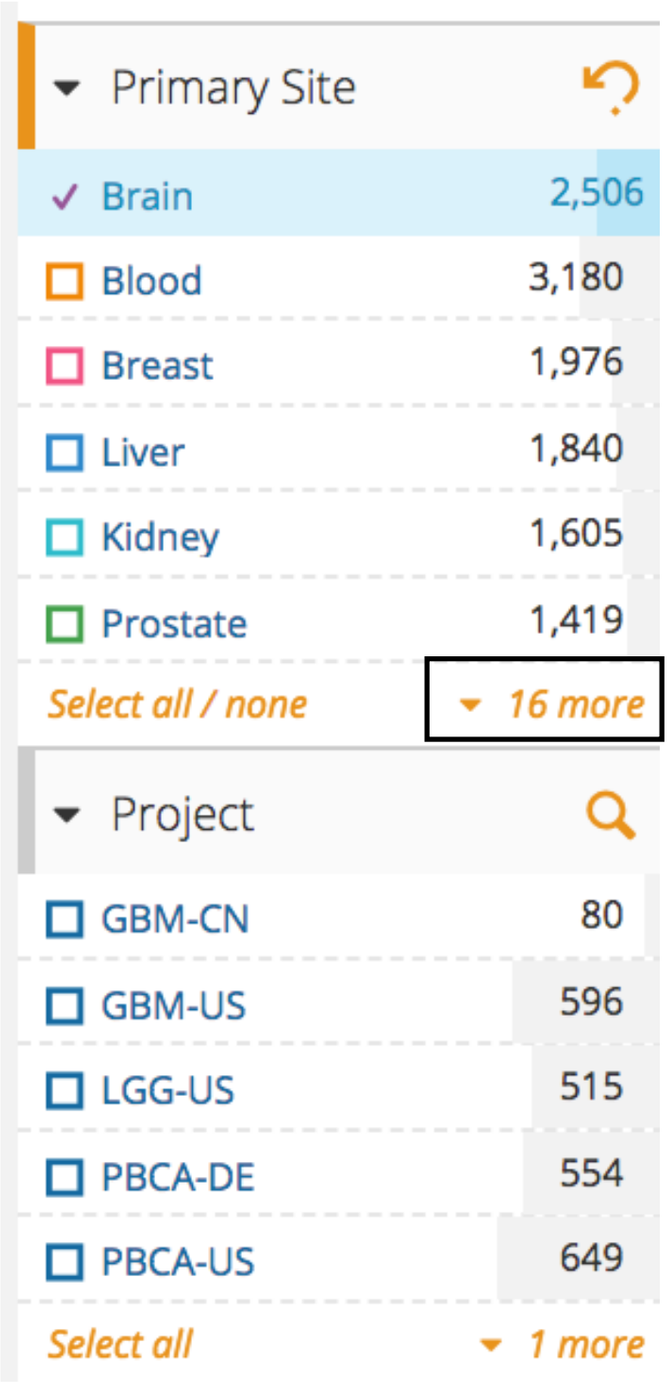
https://dcc.icgc.org/search

Dynamic filters

Before | No selections



After | **Brain** selected

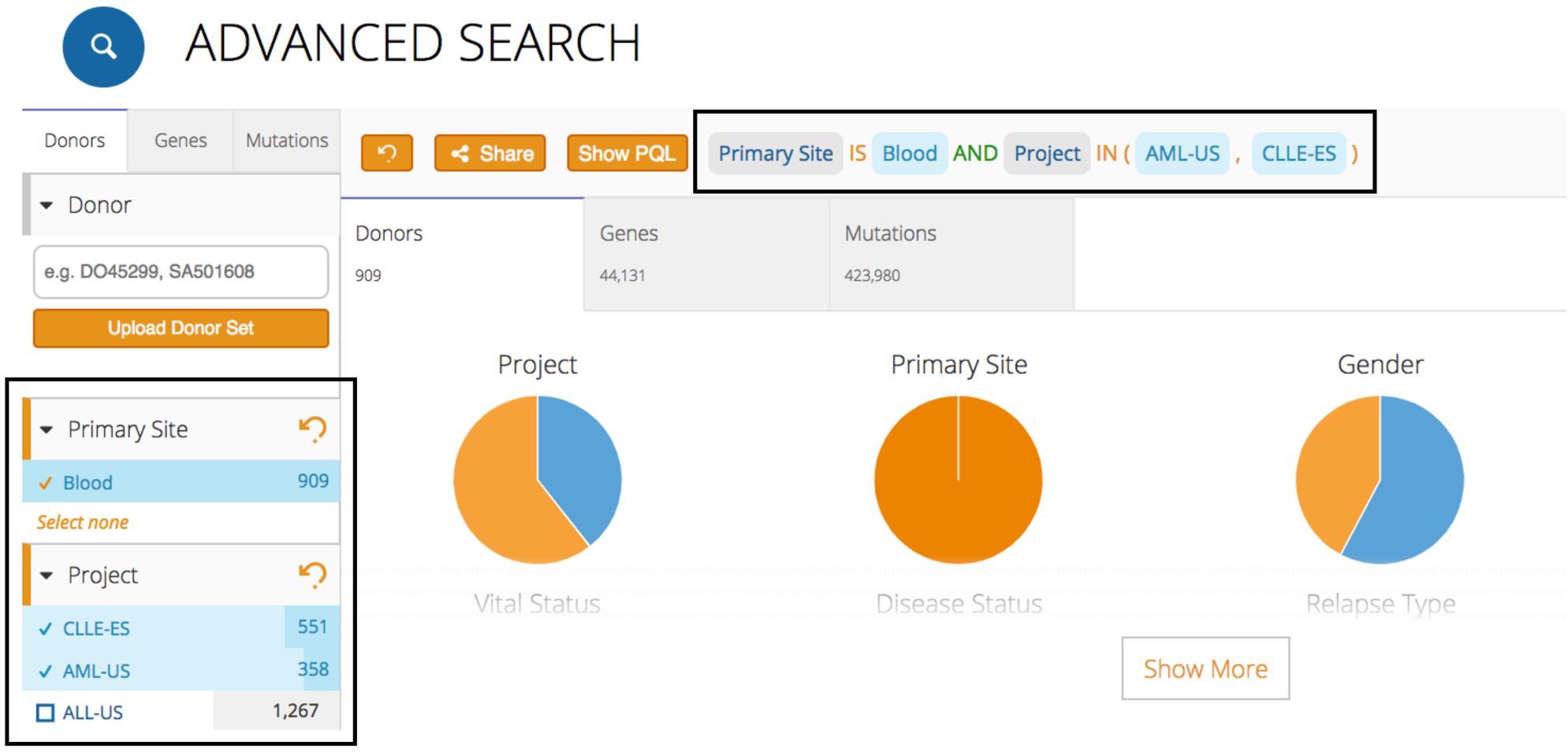


Also selection moves to the top

Truncate list

Project list changes based on **Primary Site**

Query is visible

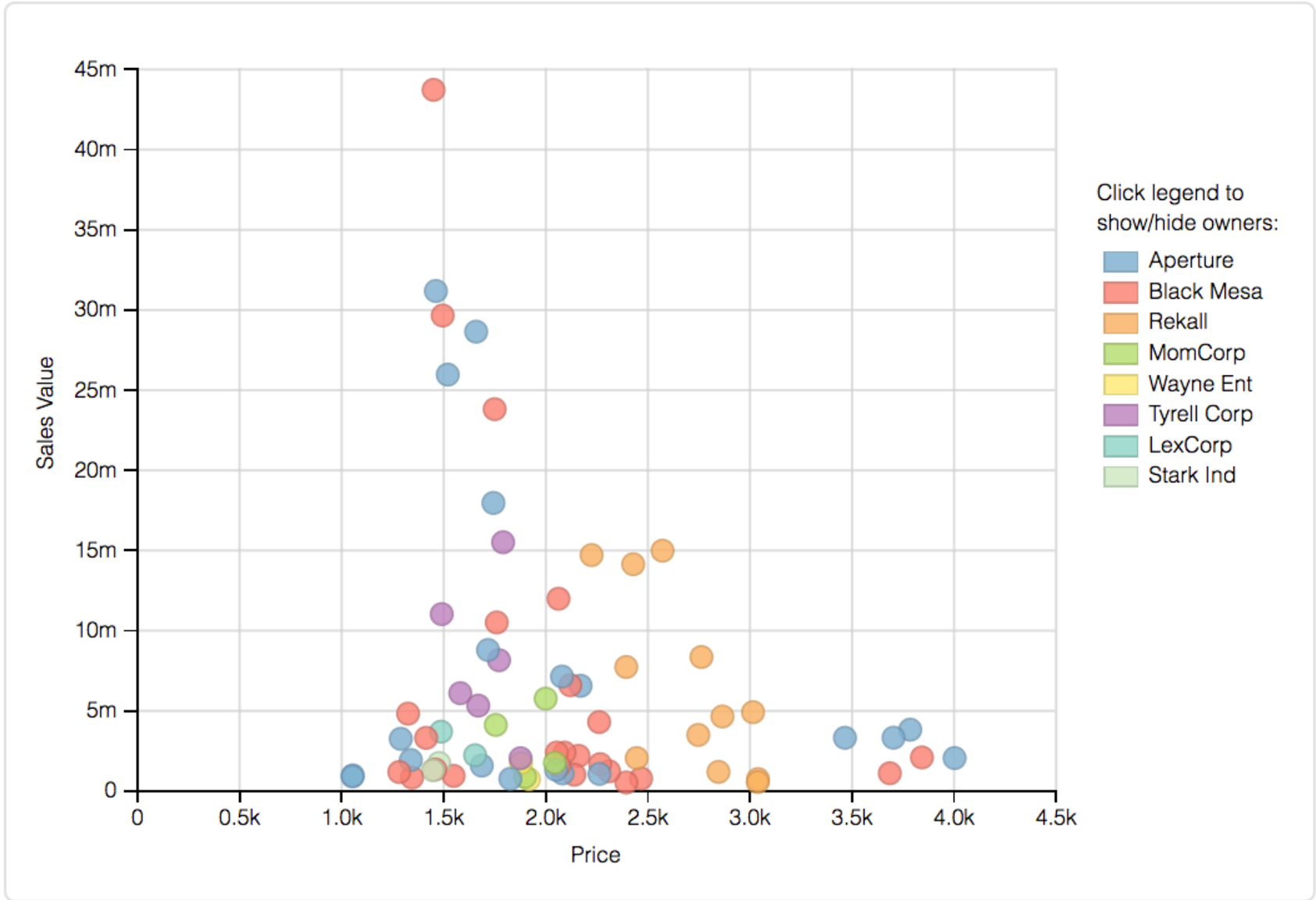


Query also captured in the page URL for easy sharing of the filter state

https://dcc.icgc.org/search

Interactive legend

Legend also functions like a filter widget

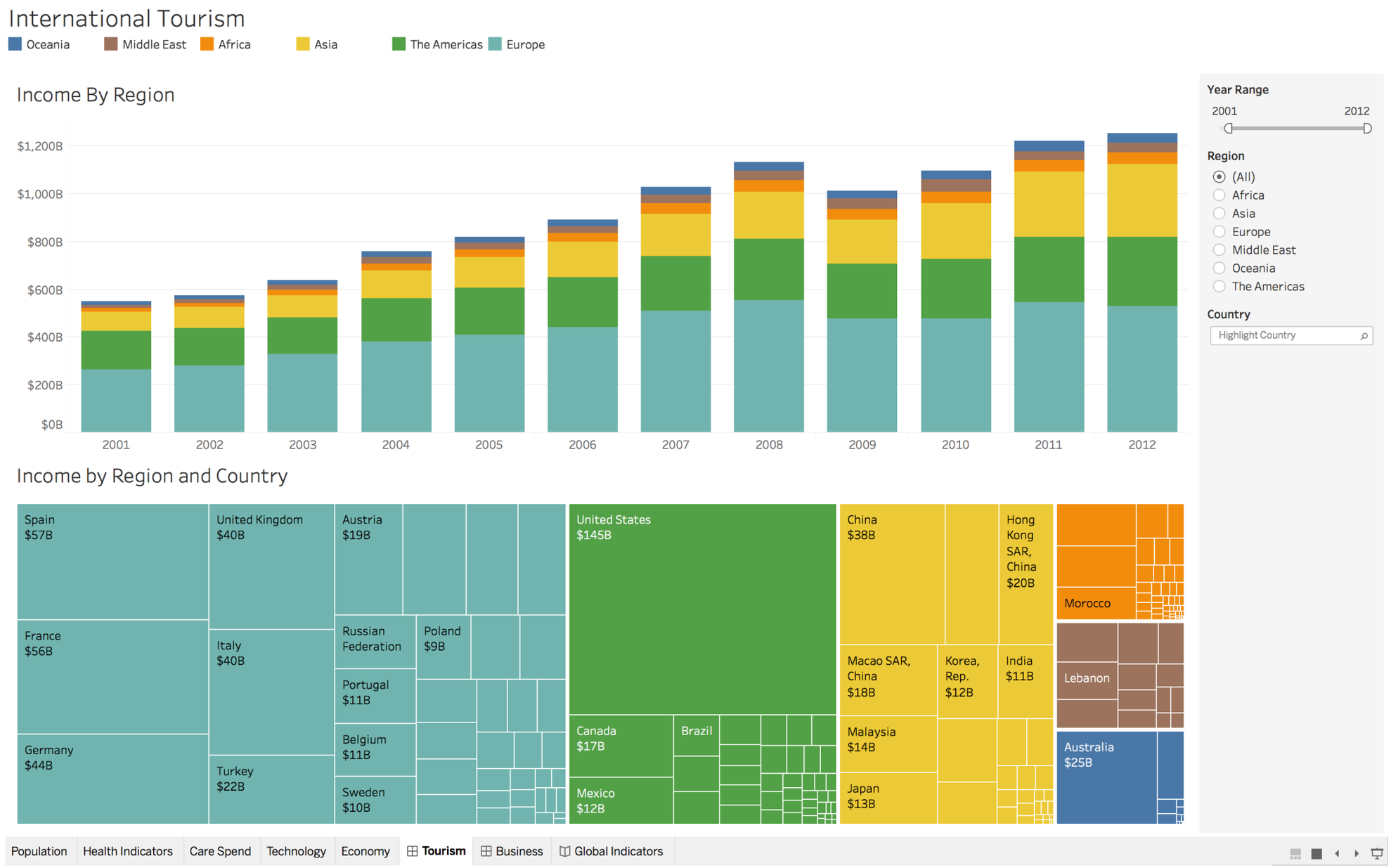


Can click on the legend to filter the data - efficient use of space

http://dimplejs.org/advanced_examples_viewer.html?id=advanced_interactive_legends

Cross filtering

Selection on plots themselves serve as filters to linked views

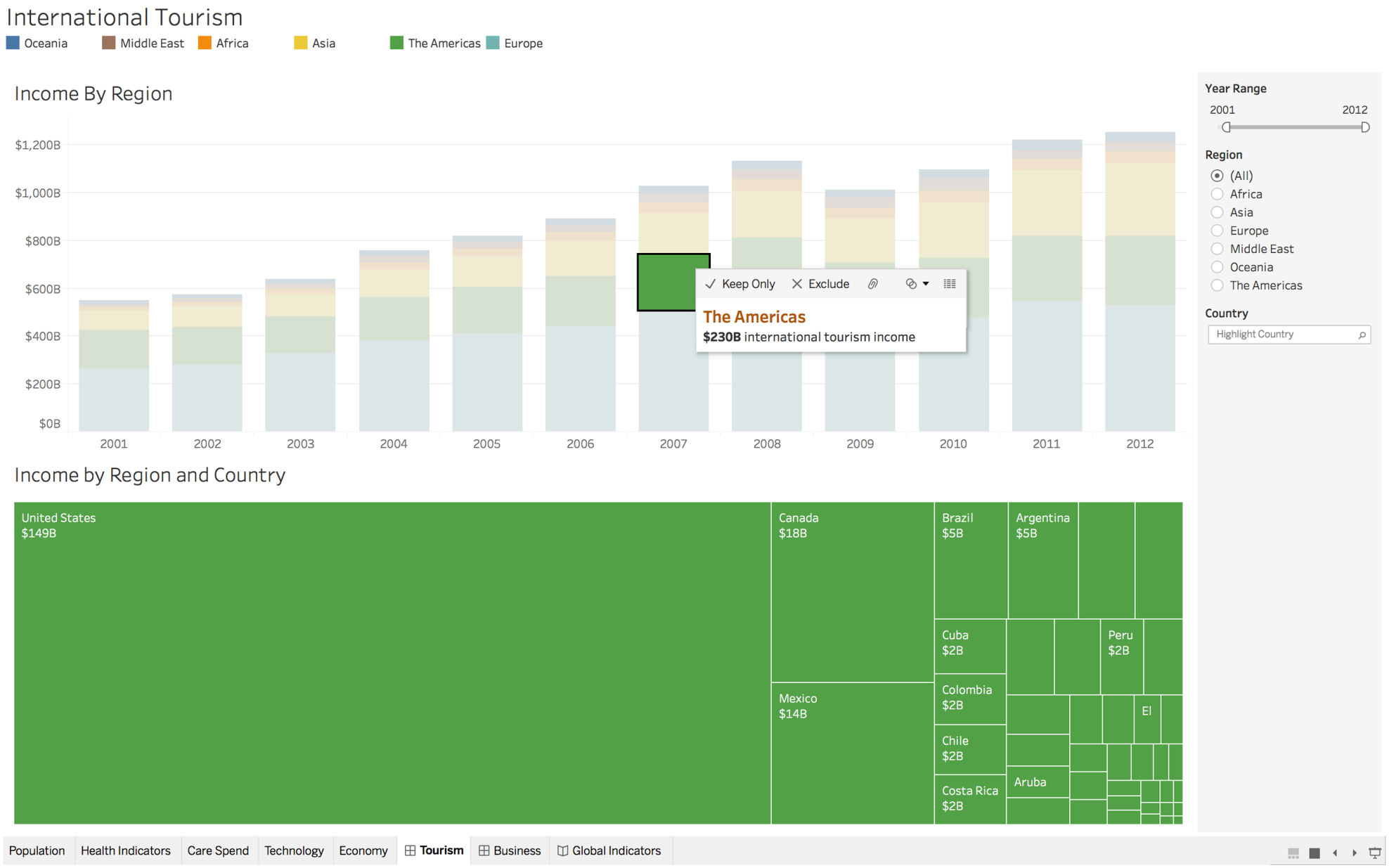


Tableau

World Indicators Sample Workbook

Cross filtering

Selection used to filter data in the linked views



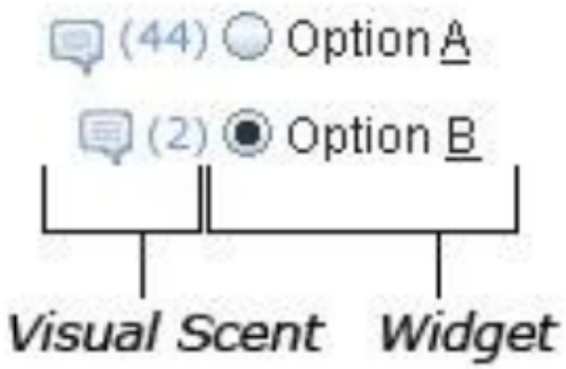
More space for selected values, but lose context

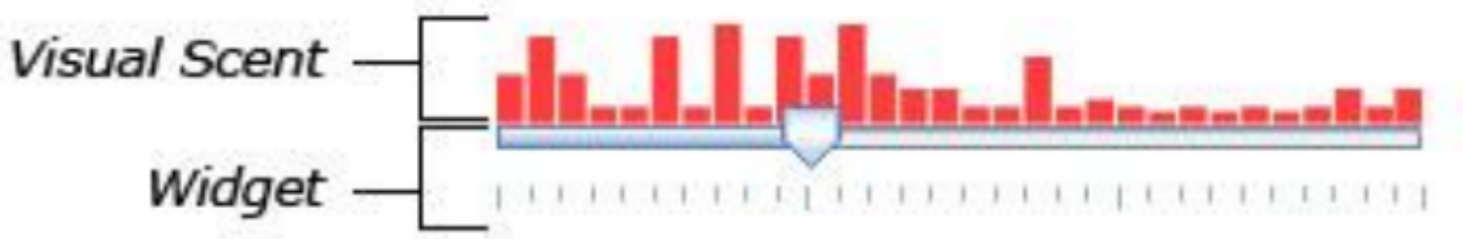
Tableau

World Indicators Sample Workbook

Scented widgets

Embedded visualizations that help you decide what to filter

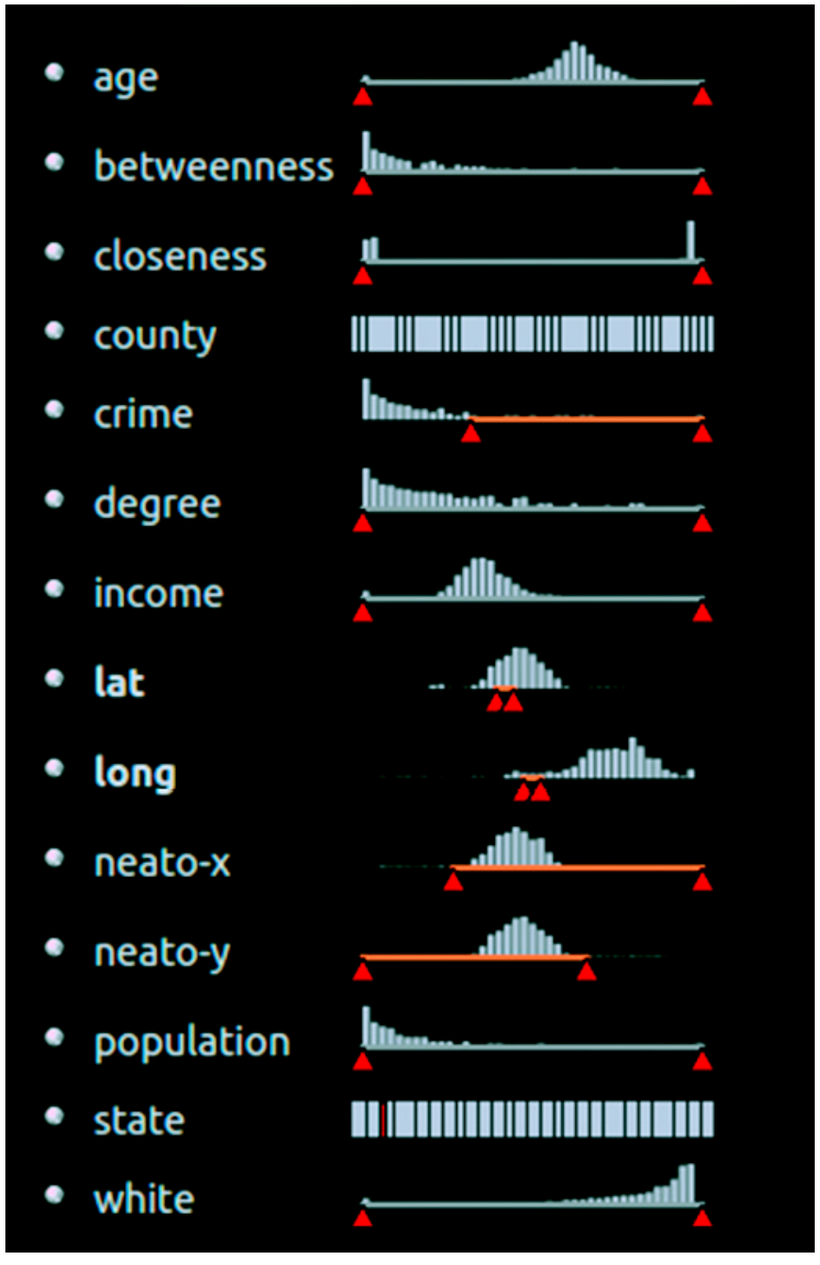




http://vis.stanford.edu/files/2007-ScentedWidgets-InfoVis.pdf

Scented widgets

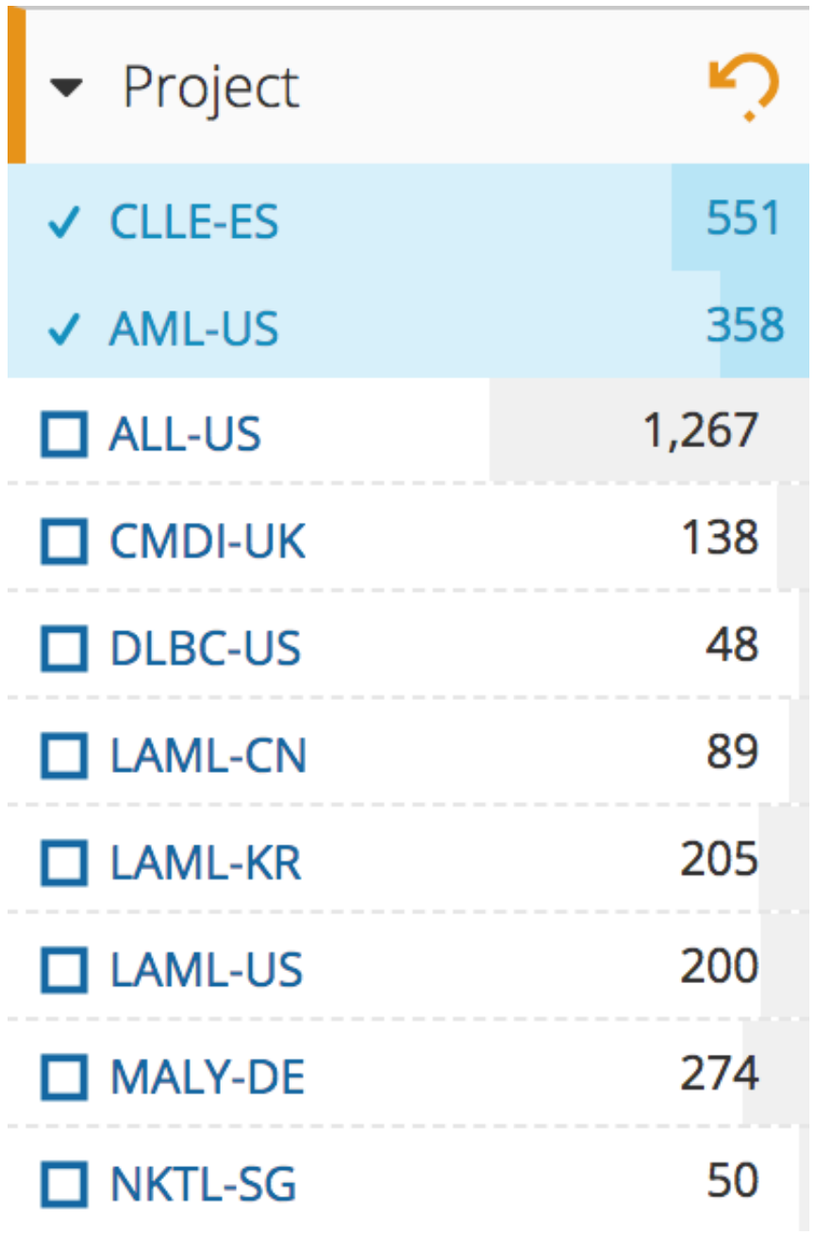
Embedded visualizations that help you decide what to filter



https://www.win.tue.nl/~selzen/paper/InfoVis2014.pdf

Scented widgets

Embedded visualizations that help you decide what to filter



Bar chart shows number of cases in each category

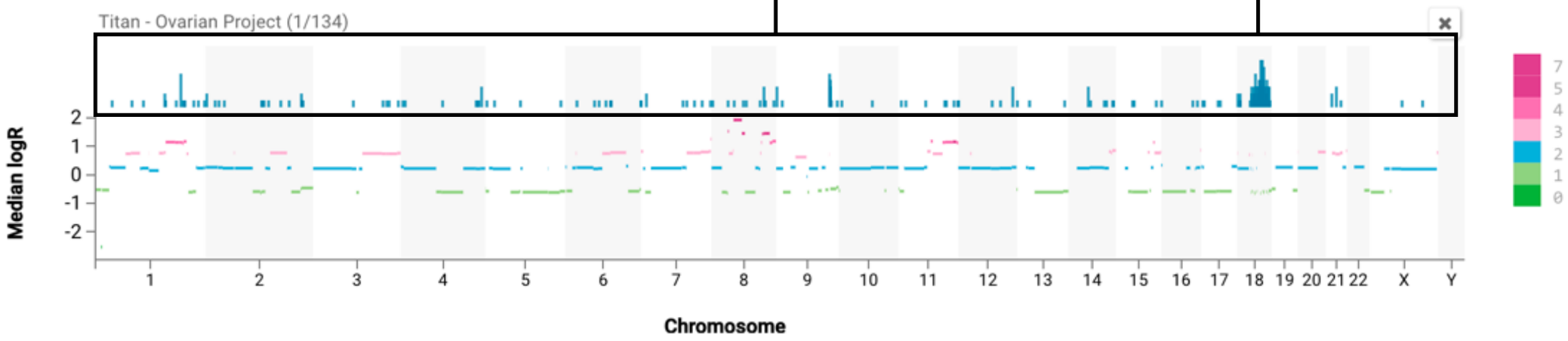
https://dcc.icgc.org/search

Out of sight, out of mind

A visual reminder of what you removed

As you zoom into this peak, small features become visible (peak tells you where to look)

Histogram of how many **features are too small** to be displayed in the current view



←——— Human genome is >3 billion base pairs long ———→

Want to display in <1000 pixels

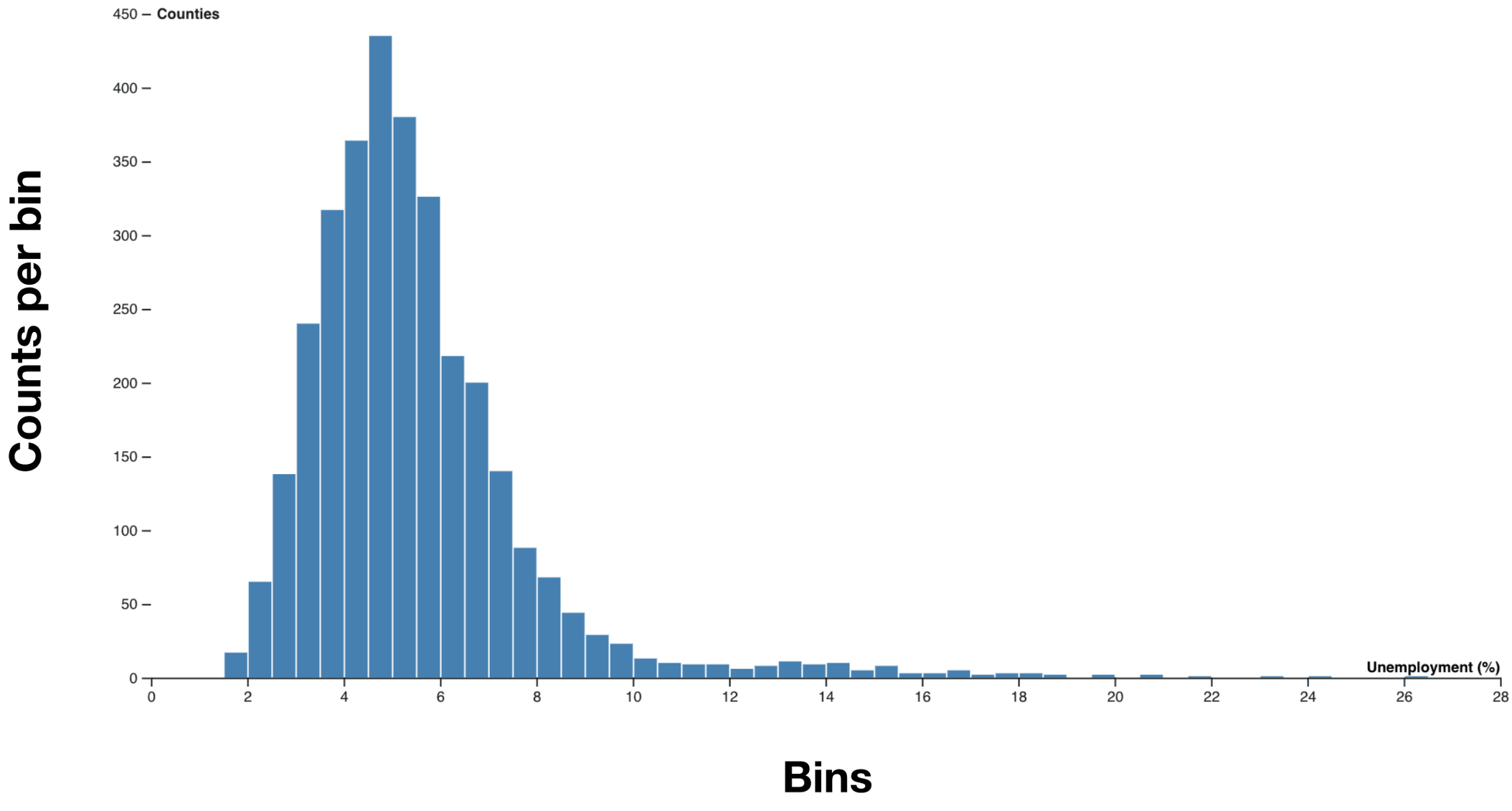
At that scale, small features won't be visible

Aggregation

Common aggregations

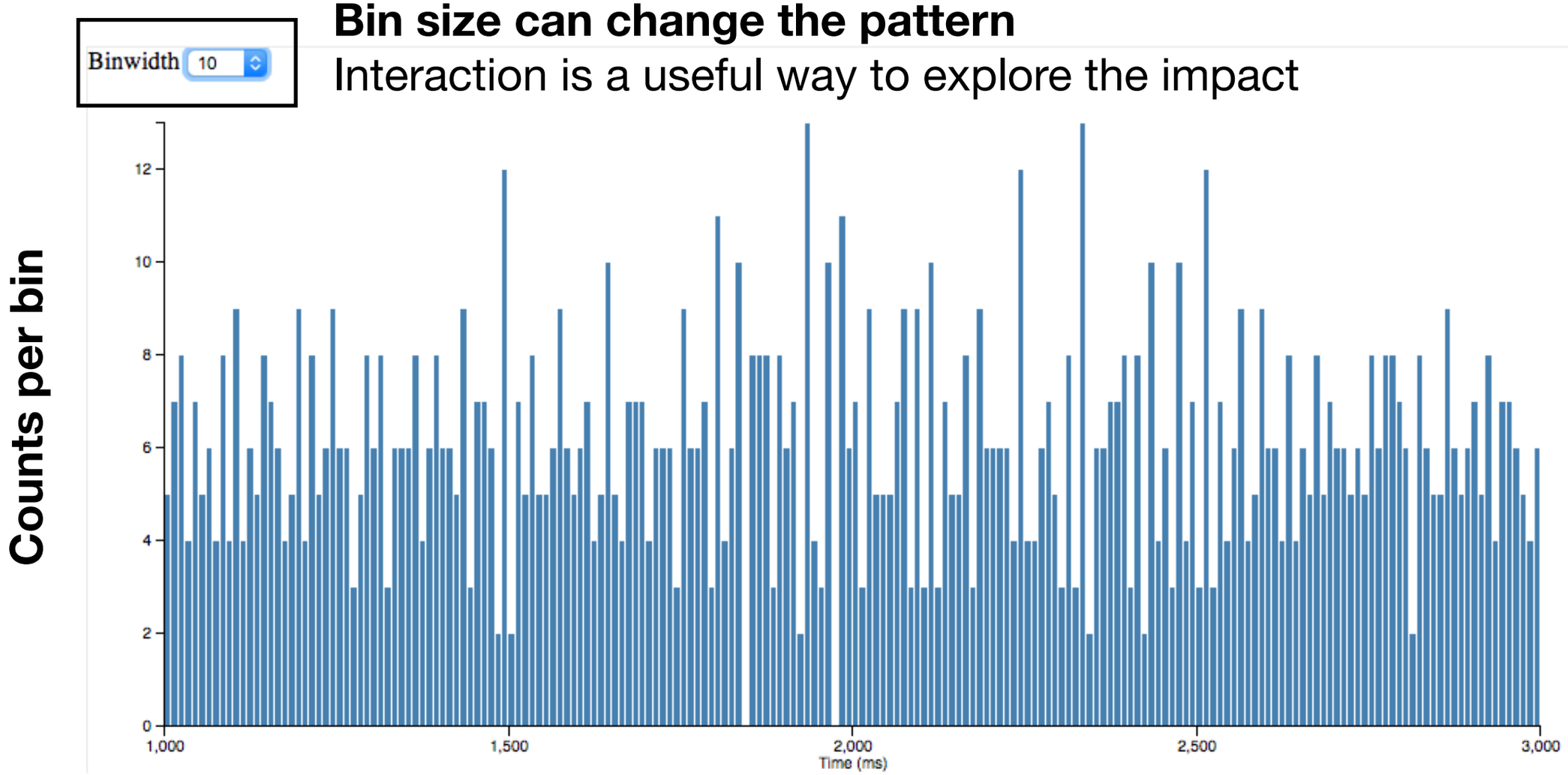
Histogram | Visualizing a distribution

Unemployment rate by county, August 2016. Source: Bureau of Labor Statistics.



https://beta.observablehq.com/@mbostock/d3-histogram

Histogram | Visualizing a distribution

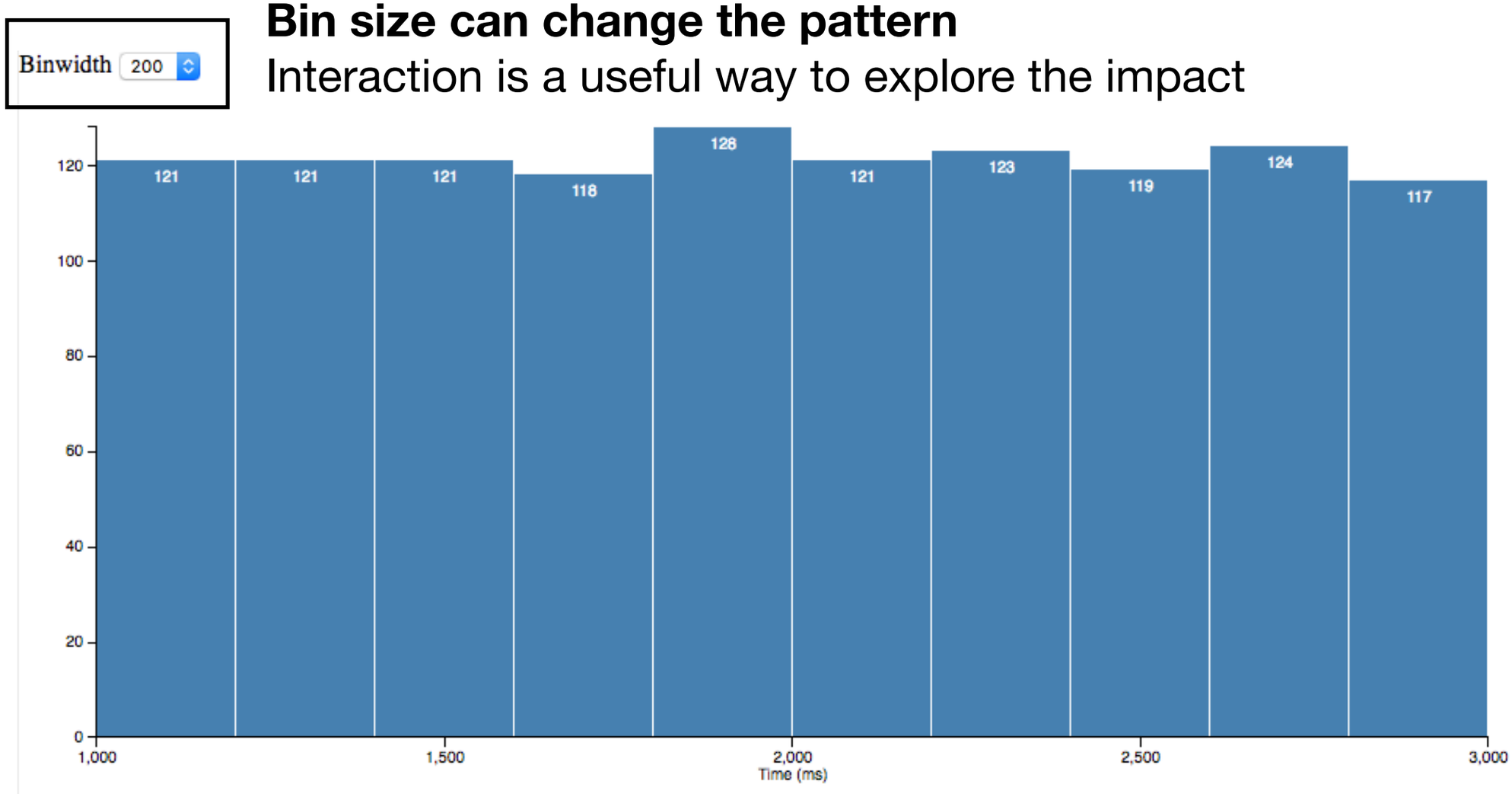


http://bl.ocks.org/babsey/9b428f5f20e538c15b9e

Histogram | Visualizing a distribution

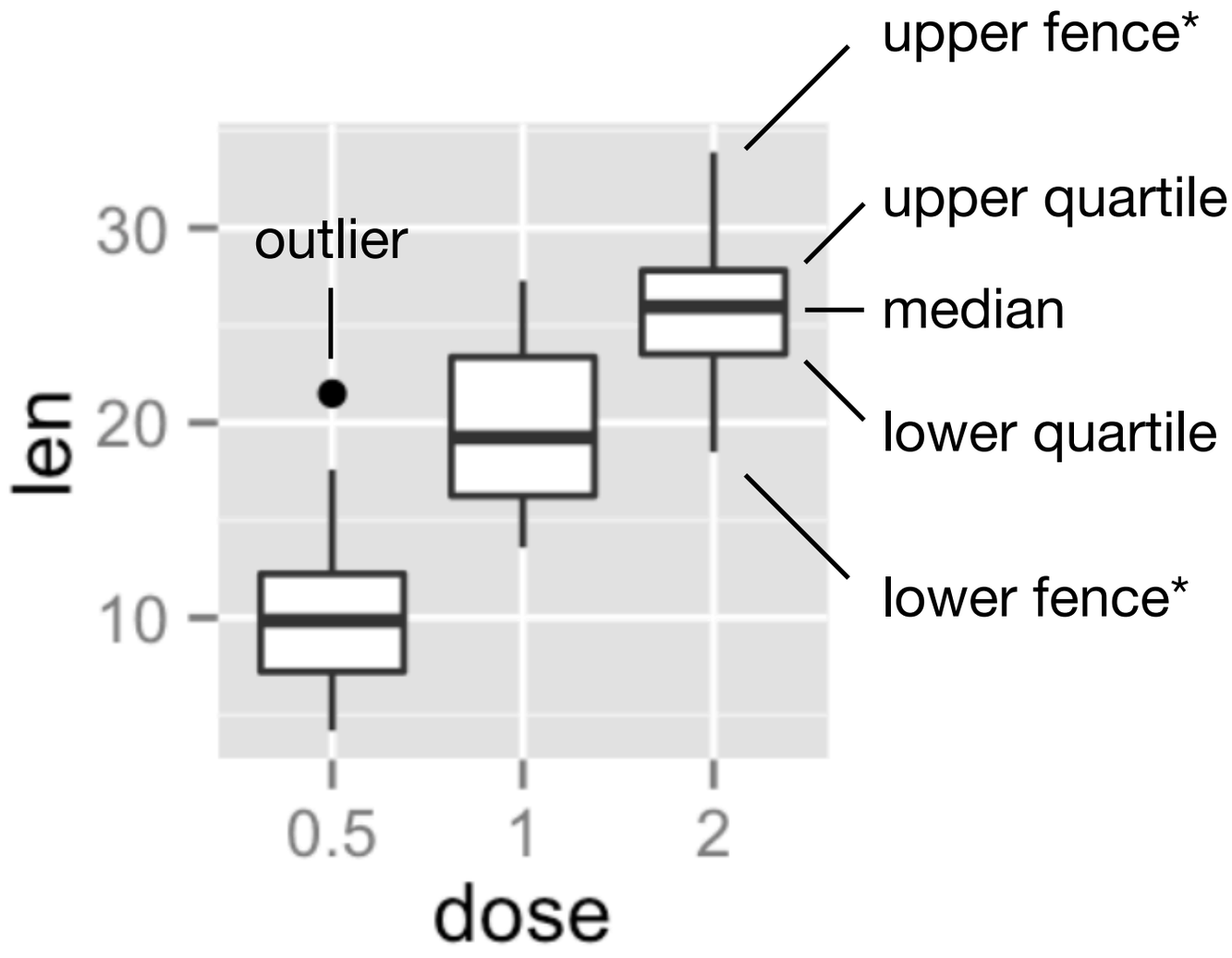
Bin size can change the pattern

Interaction is a useful way to explore the impact



http://bl.ocks.org/babsey/9b428f5f20e538c15b9e

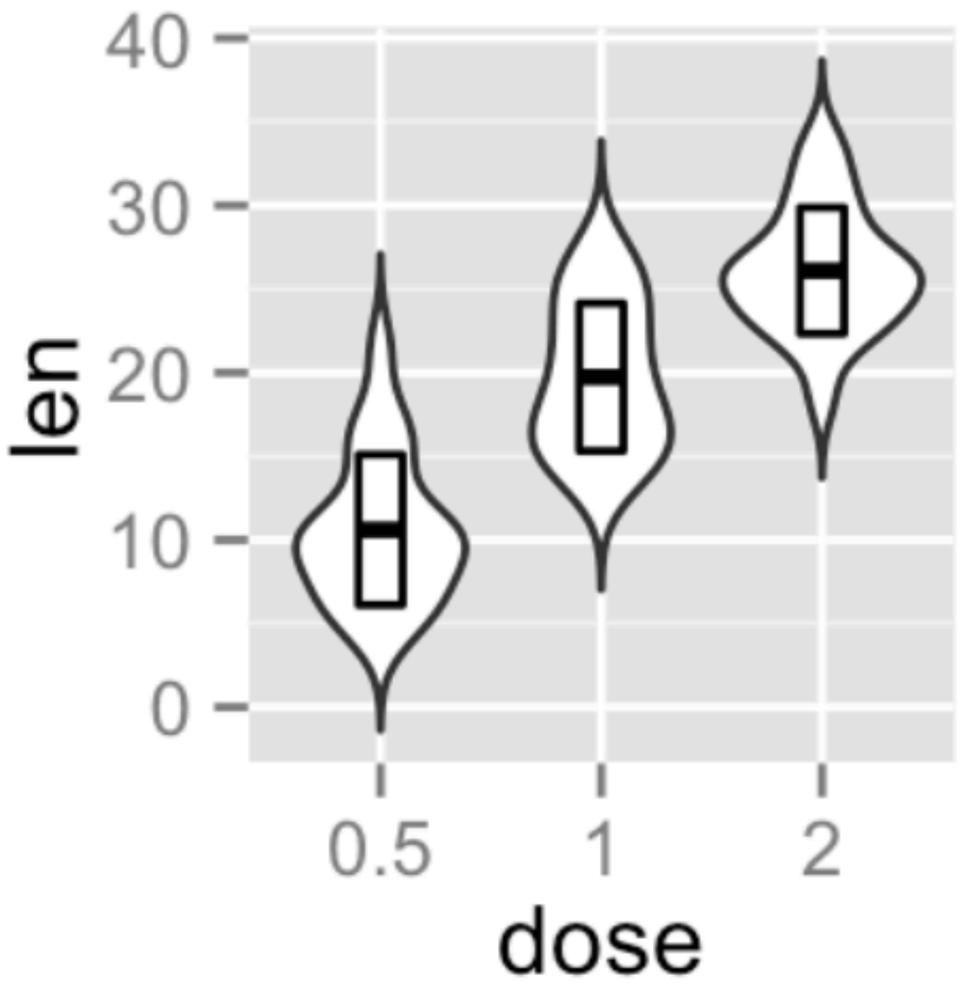
Box plots | Visualizing distributions



*__fences__ often extend from upper/lower quartile to the largest value no further than +/- 1.5 * inter-quartile range

http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-...

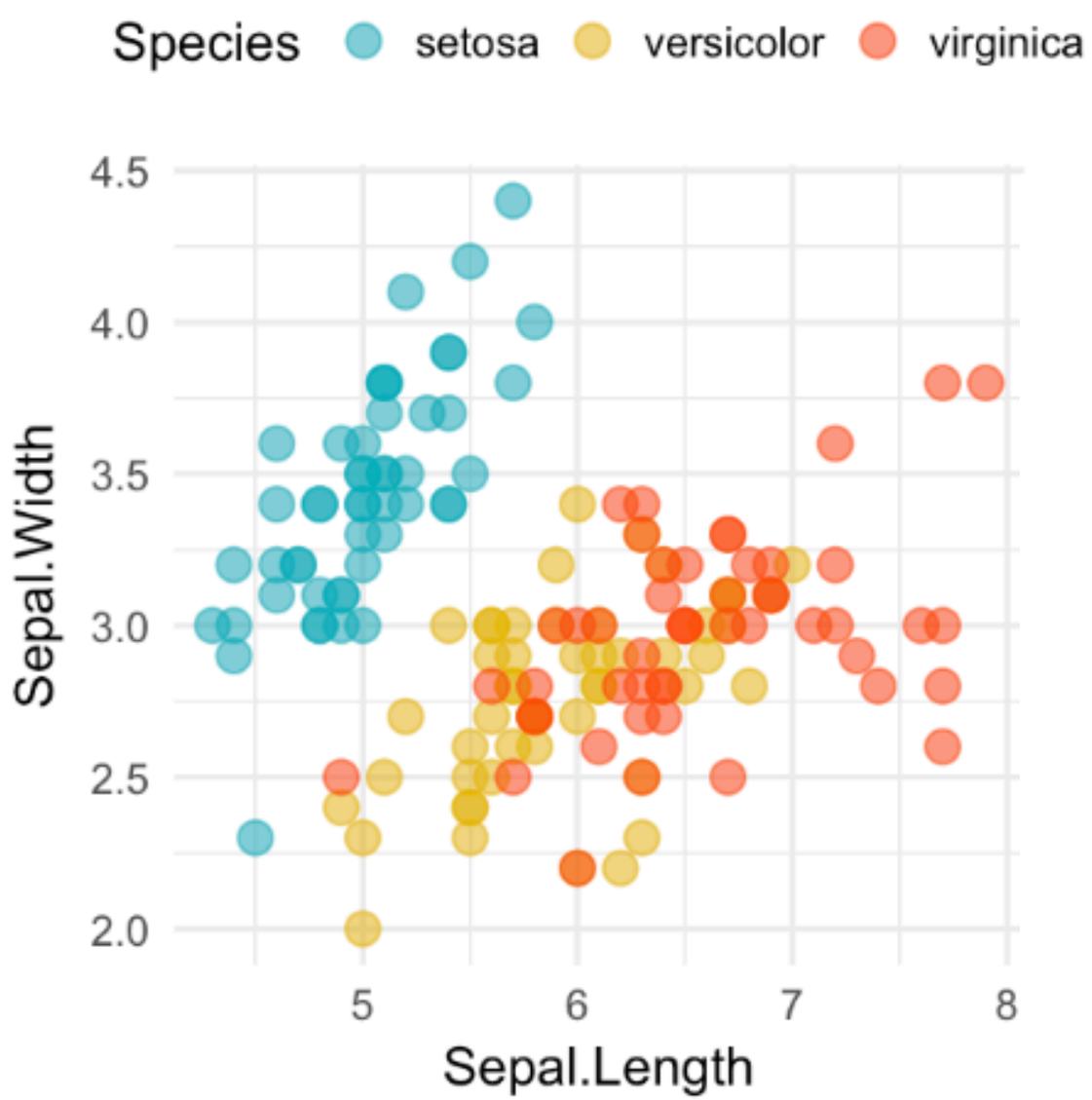
Violin plots | Visualizing distributions



Similar to a box plot but replaces the box with a mirrored density plot

http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-...

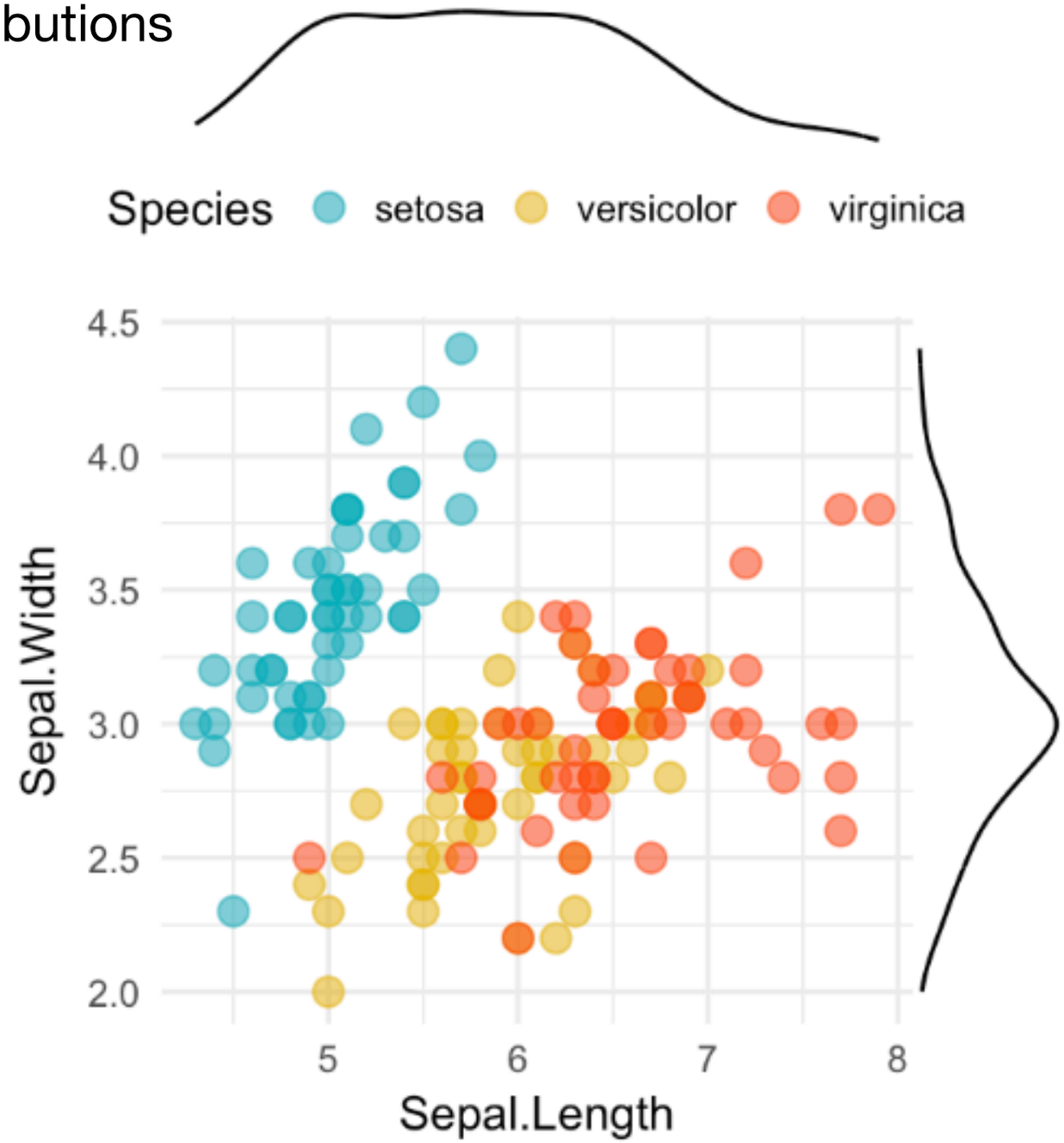
Scatter plots



http://www.sthda.com/english/articles/32-r-graphics-essentials/...

Scatter plots

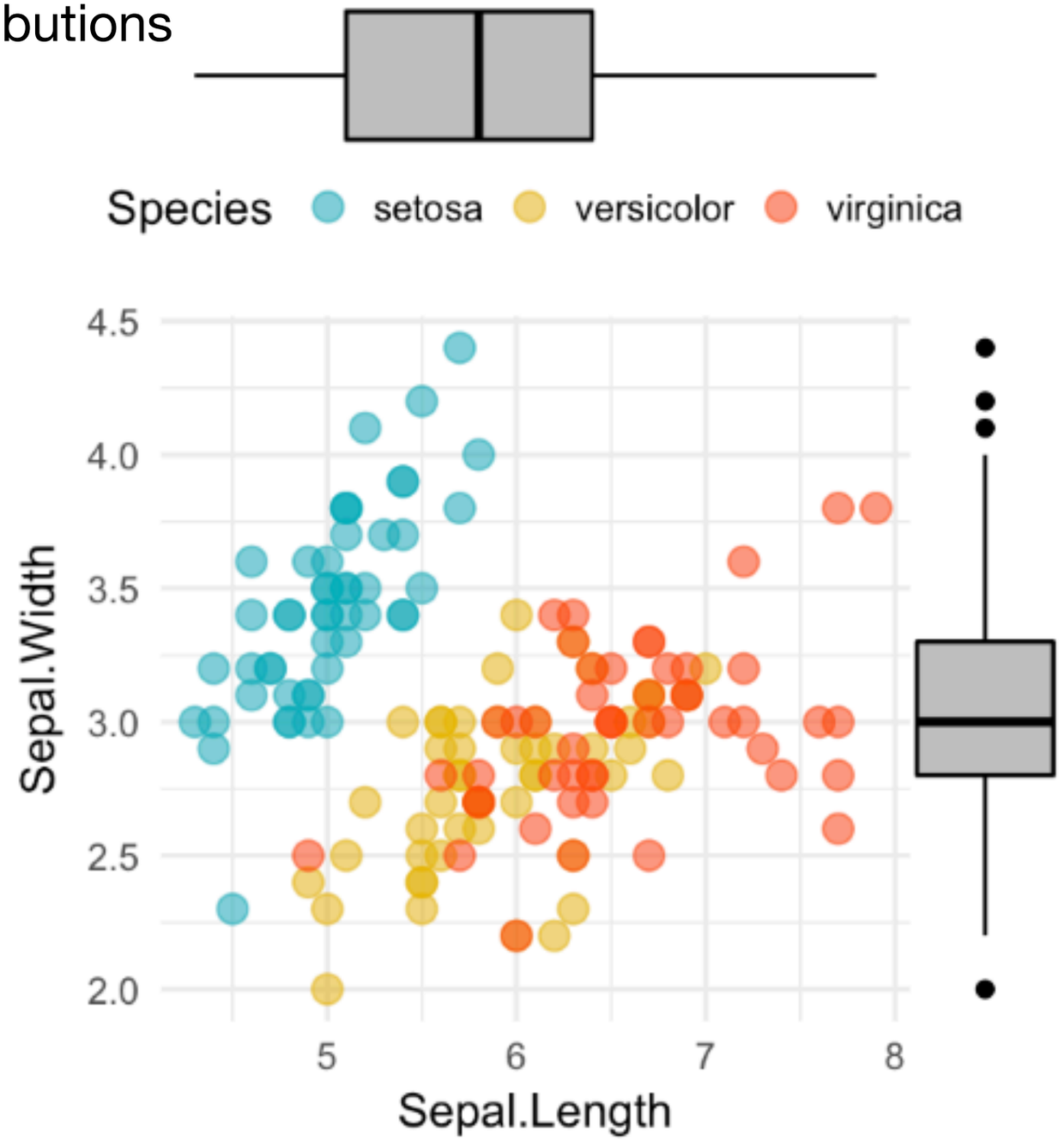
with marginal distributions



http://www.sthda.com/english/articles/32-r-graphics-essentials/...

Scatter plots

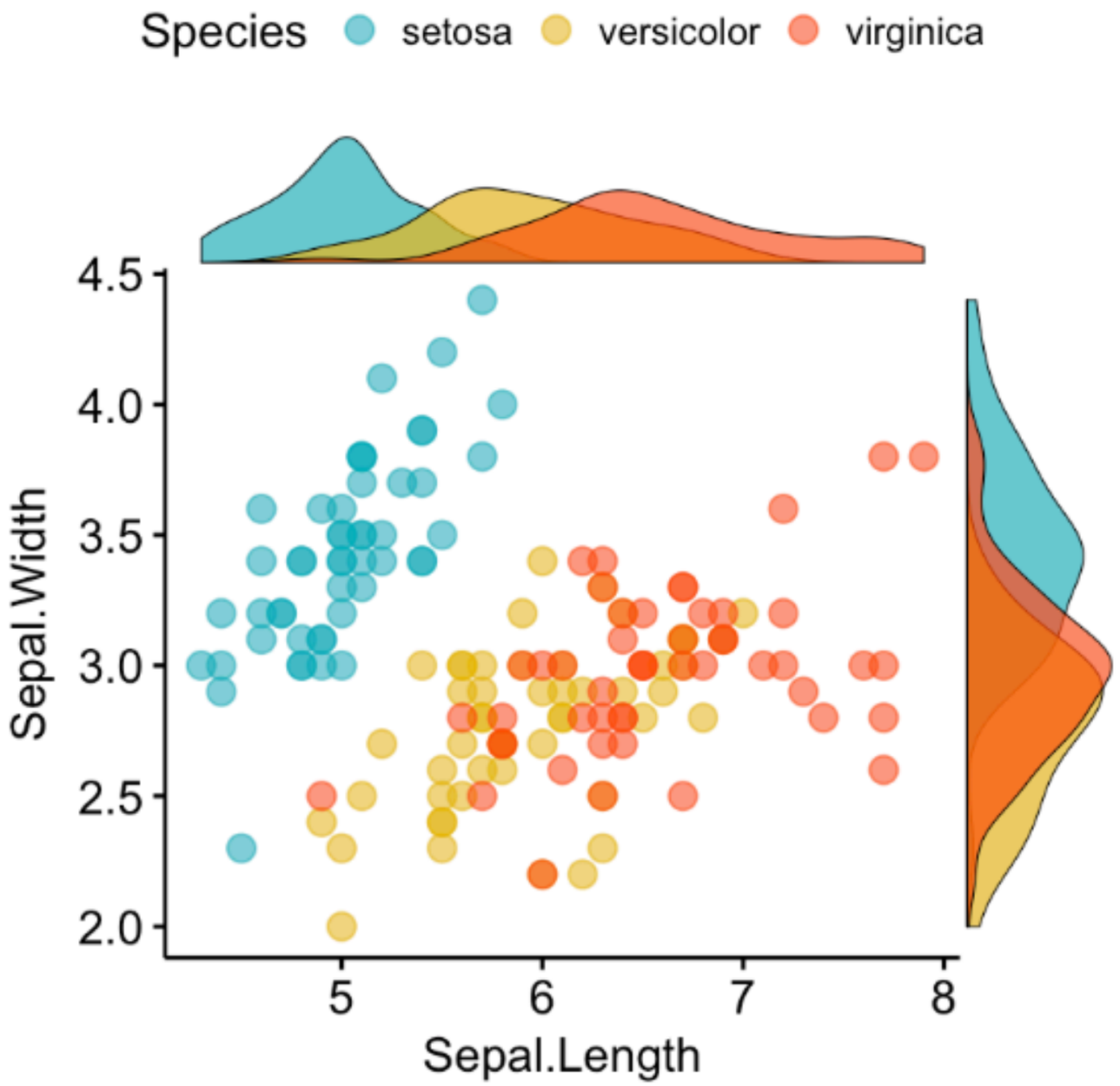
with marginal distributions



http://www.sthda.com/english/articles/32-r-graphics-essentials/...

Scatter plots

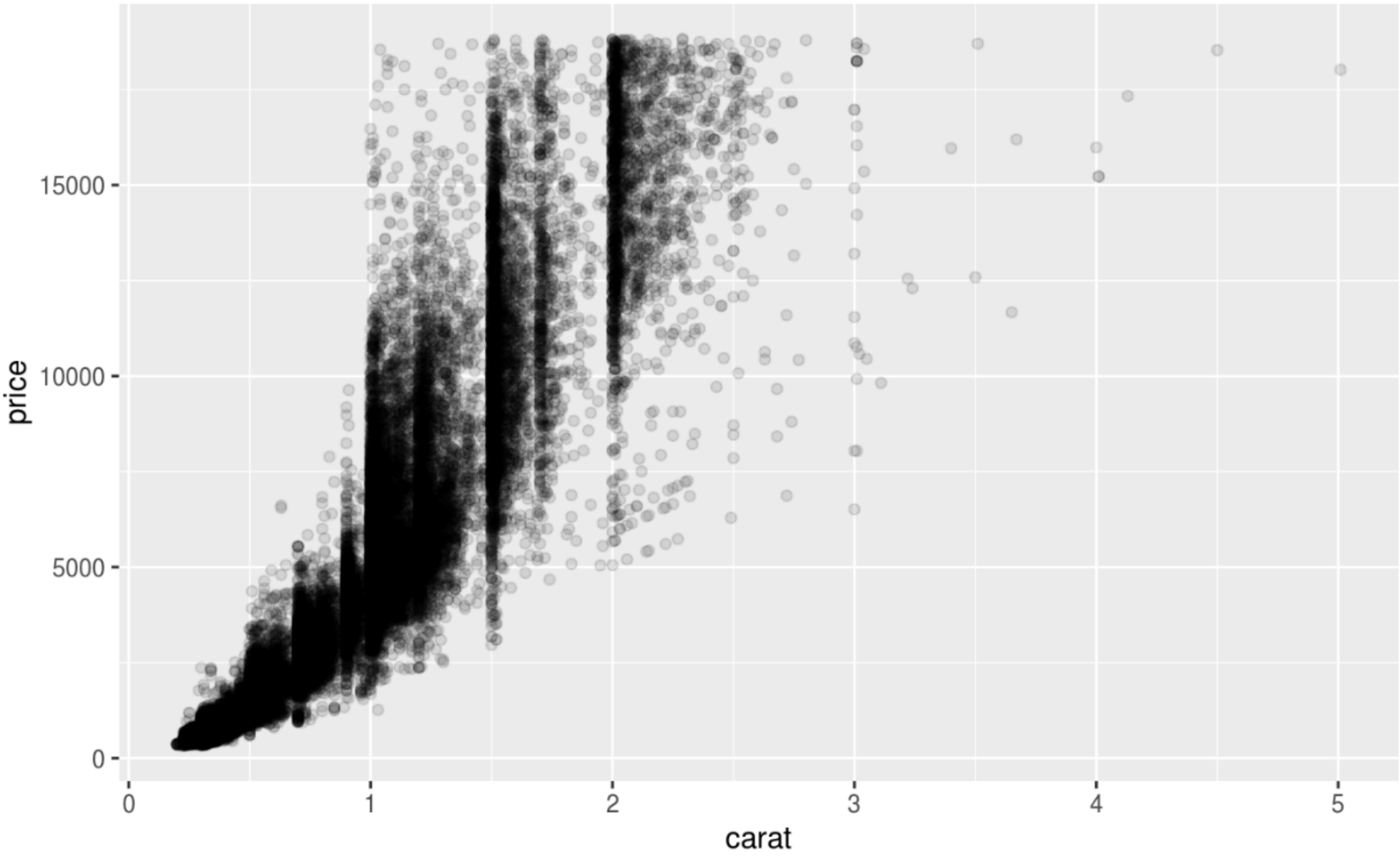
with marginal distributions



http://www.sthda.com/english/articles/32-r-graphics-essentials/...

What if there are too many points to see them all?

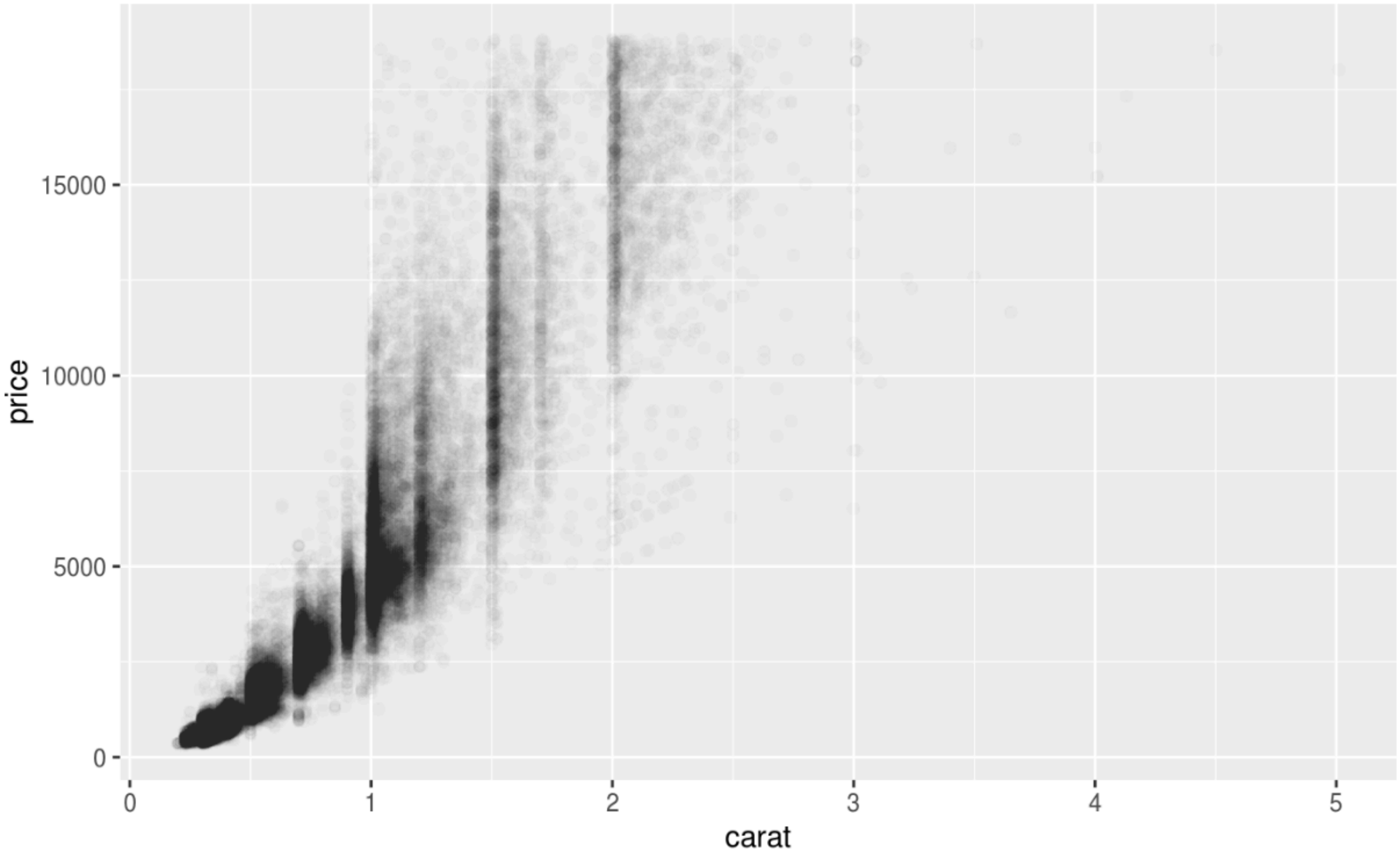
Overplotting



https://ggplot2.tidyverse.org/reference/geom_point.html

Overplotting

Changing the alpha value can help

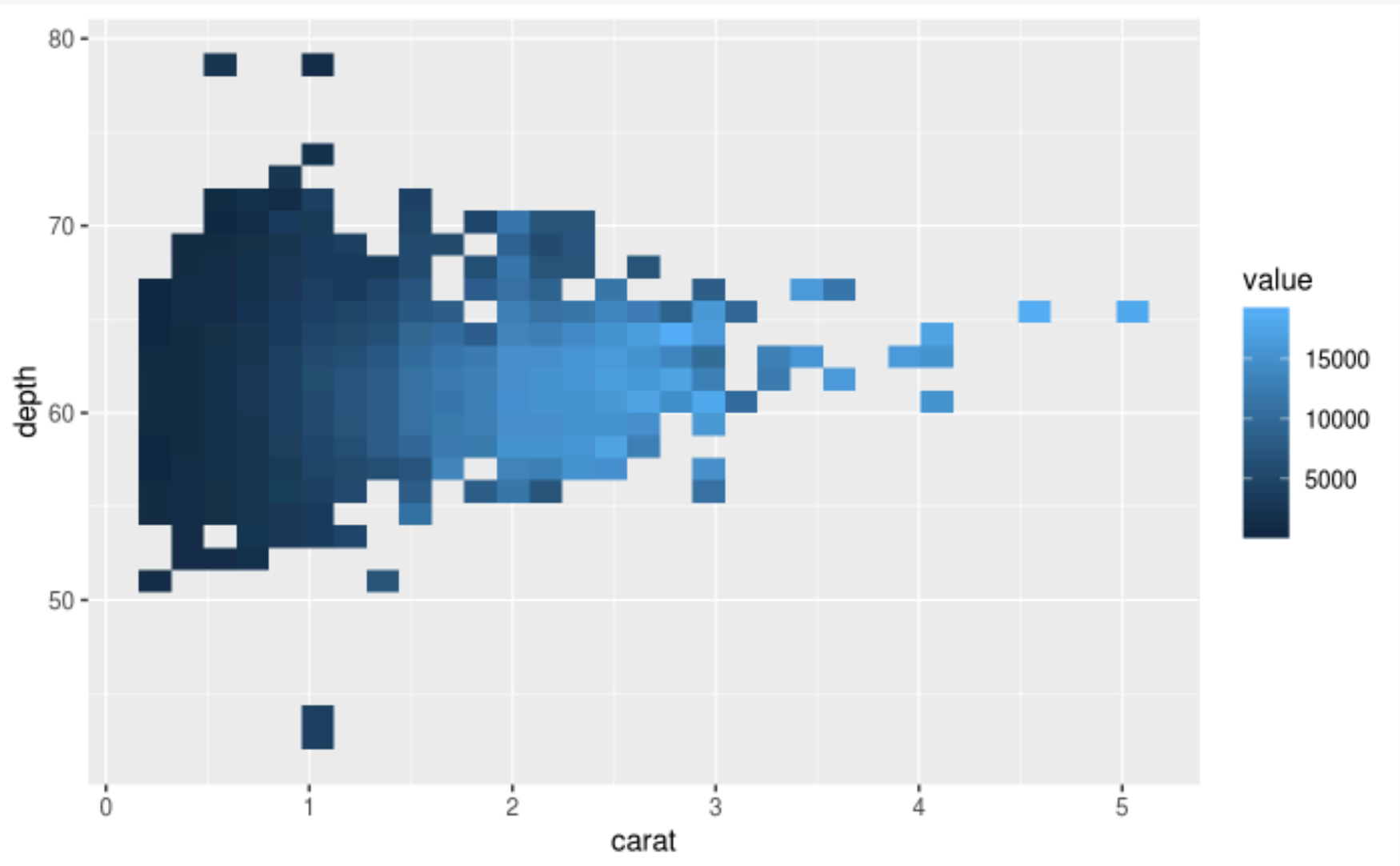


https://ggplot2.tidyverse.org/reference/geom_point.html

Binning | Rectangular binning

Great solution for large data sets

```
d <- ggplot(diamonds, aes(carat, depth, z = price))
d + stat_summary_2d()
```

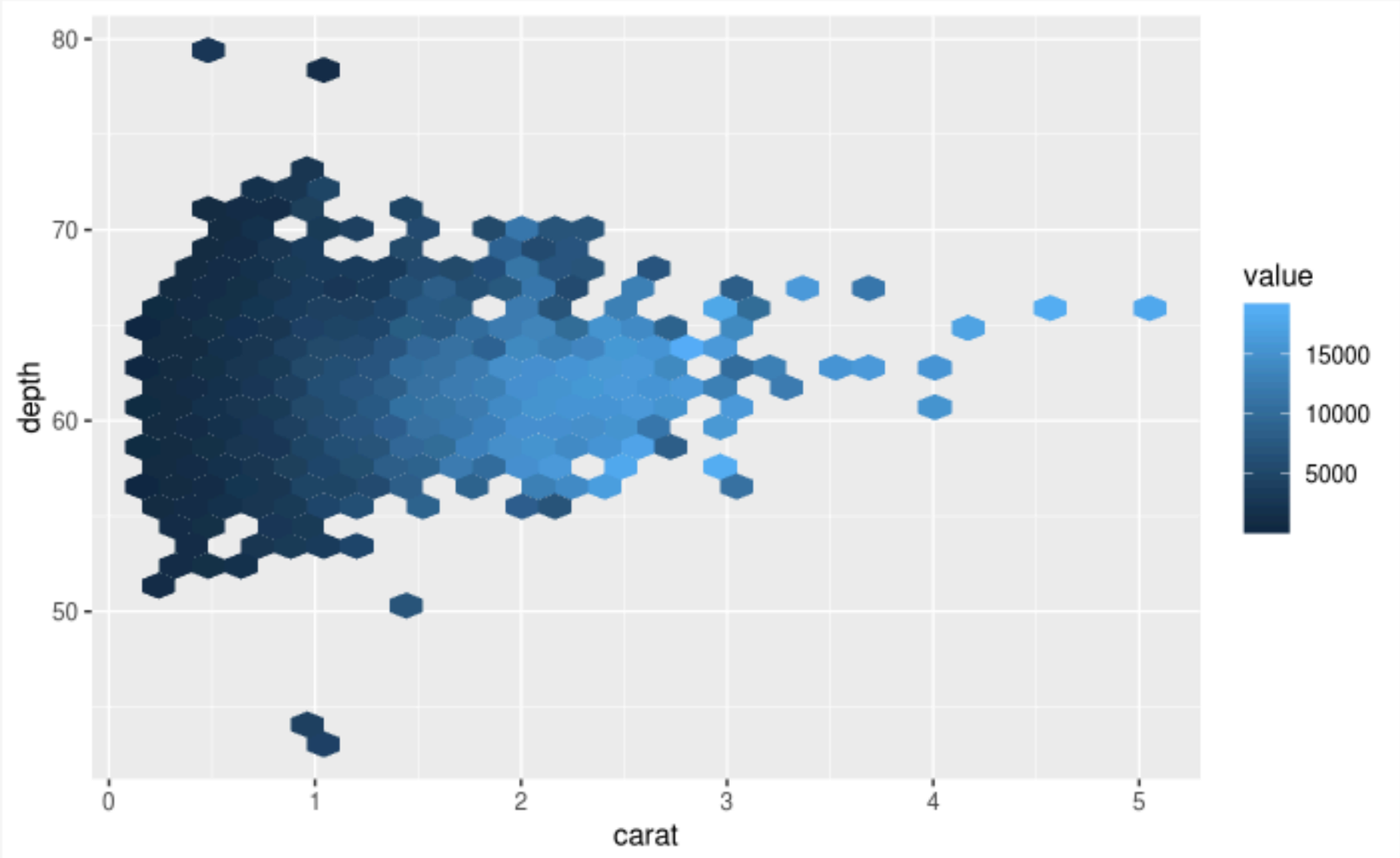


https://ggplot2.tidyverse.org/reference/stat_summary_2d.html

Binning | Hexagonal binning

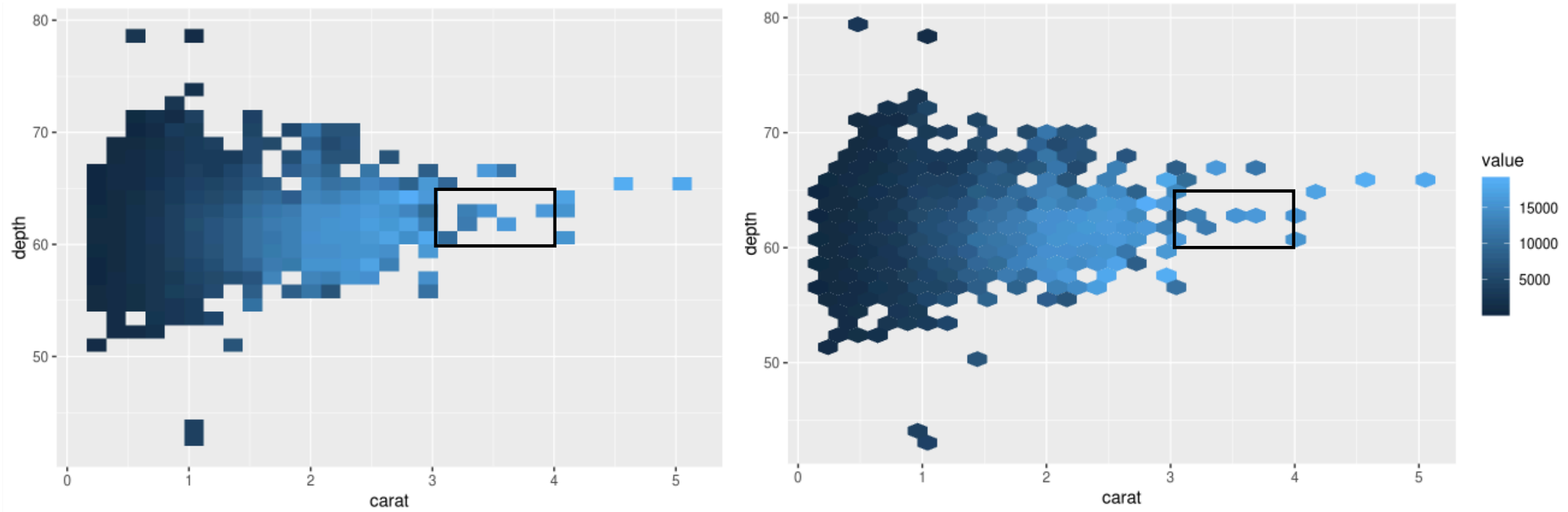
Great solution for large data sets

```
if (requireNamespace("hexbin")) {
d + stat_summary_hex()
}
```



https://ggplot2.tidyverse.org/reference/stat_summary_2d.html

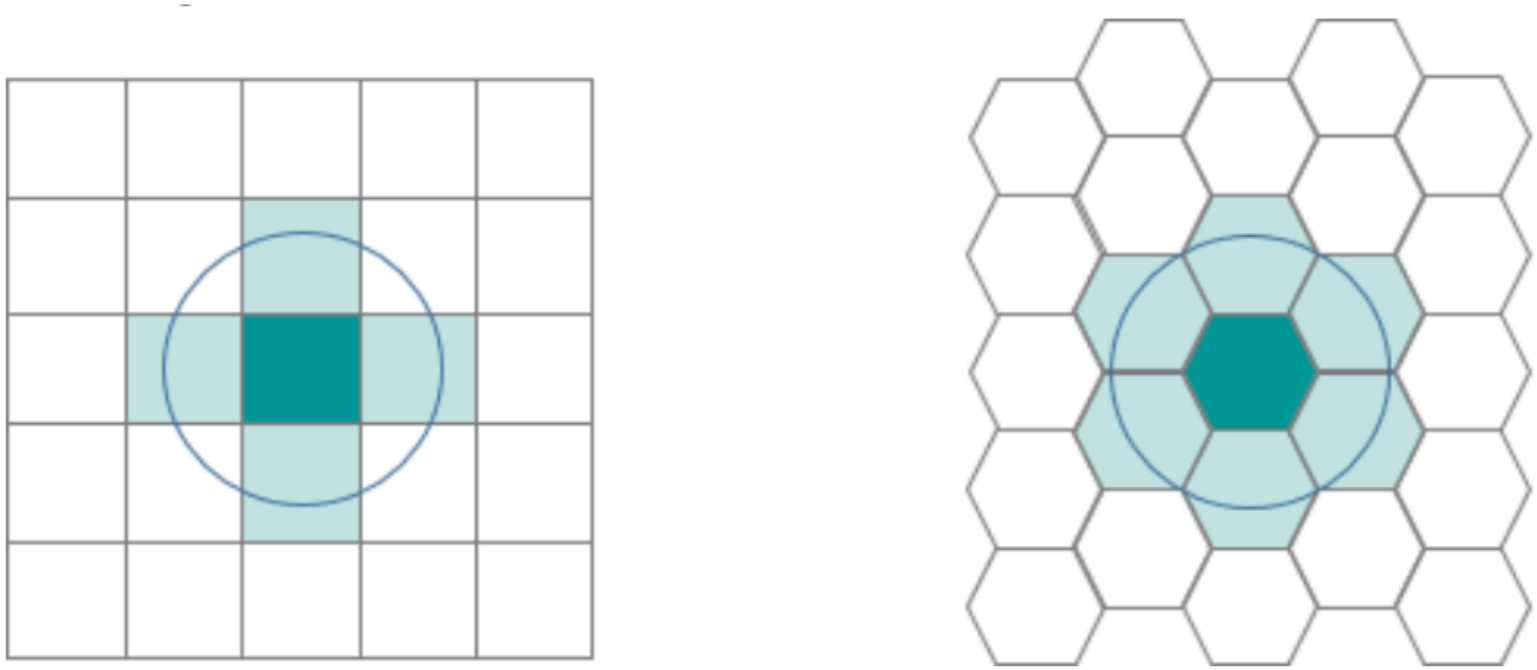
Rectangles versus Hexagons



- Rectangular binning can lead to artificially straight, parallel line patterns
- Circularity of a hexagon grid allows it to represent curved patterns in your data more naturally than square grids

http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-whyhexagons.htm

Rectangles versus Hexagons

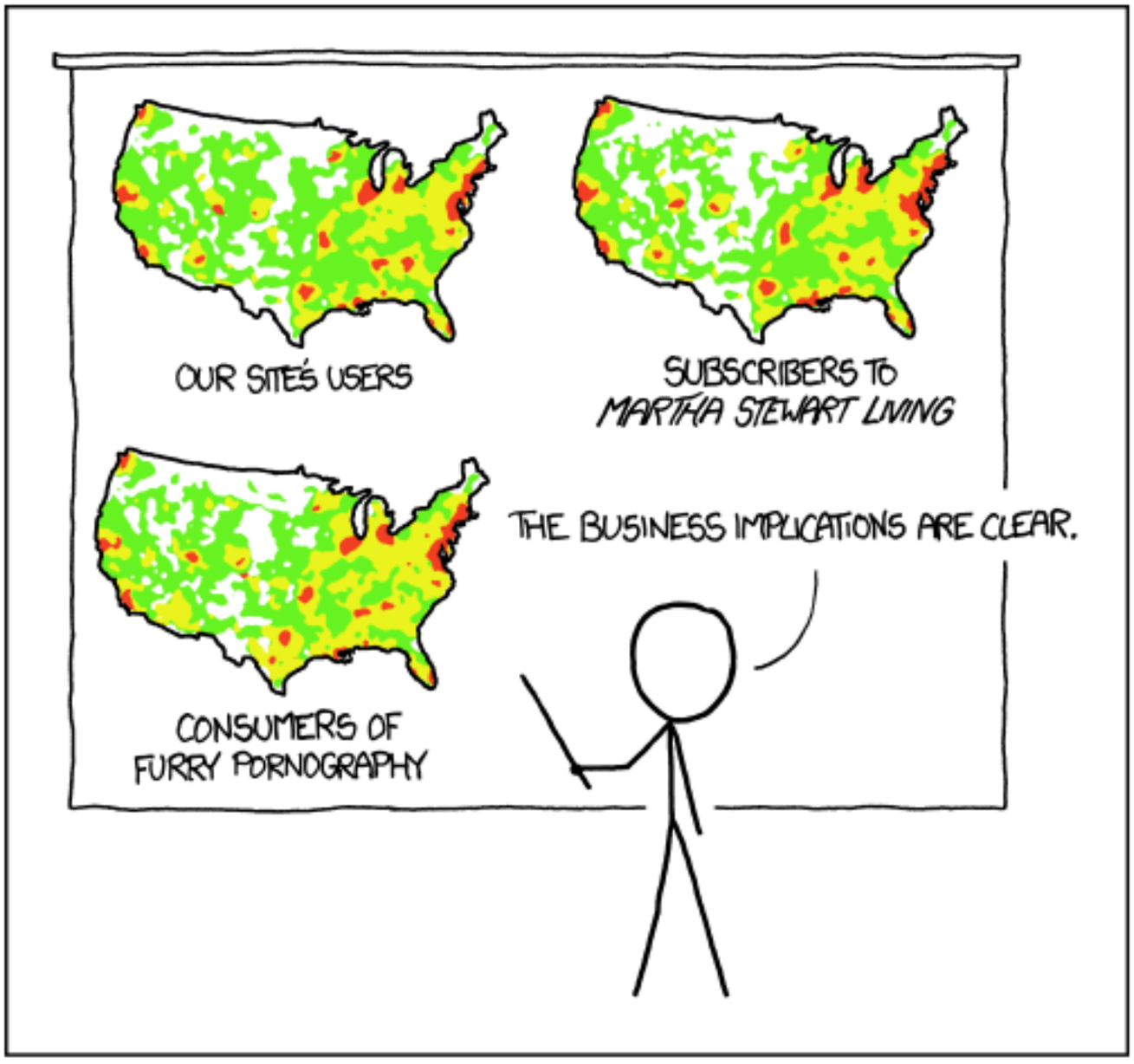


- Rectangular binning can lead to artificially straight, parallel line patterns
- Circularity of a hexagon grid allows it to represent curved patterns in your data more naturally than square grids
- Nice property that the distance between centroids is the same in all six directions with hexagons; this will impact calculations of number of neighbours within a given distance

http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-whyhexagons.htm

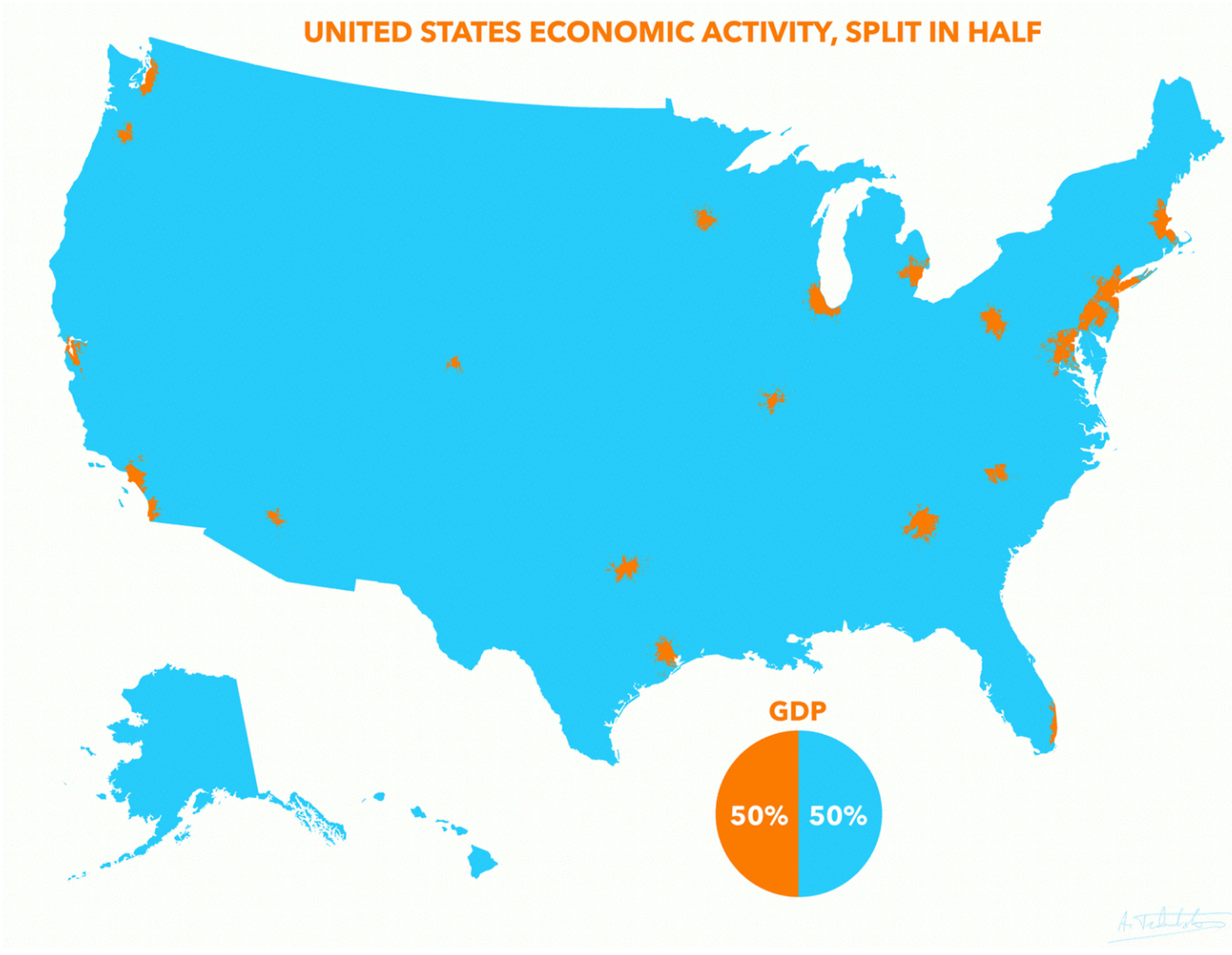
Aggregation on a geographic map

Choropleth maps



https://xkcd.com/1138/

The incredible map that shows that half of the U.S. population produces half of the GDP



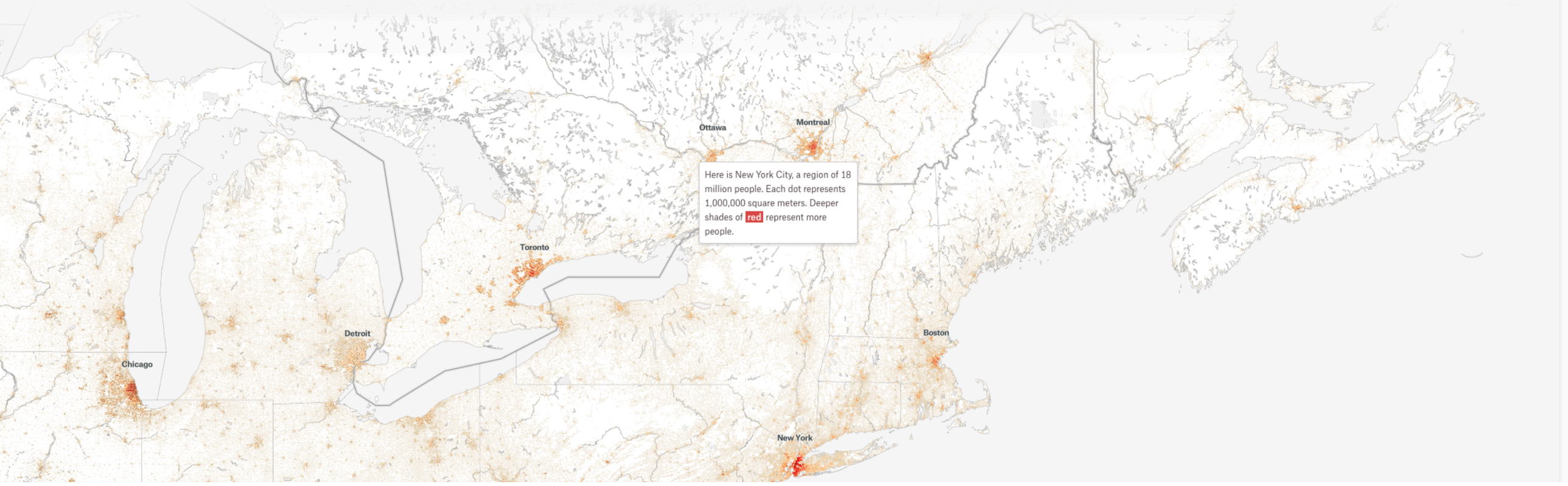
See discussion and links on Alberto Cairo's blog | the Functional Art
http://www.thefunctionalart.com/2014/02/the-incredible-gdp-map-that-shows-that.html

What does the world’s population distribution look like?

Population Mountains

By Matt Daniels

This is a story about how to perceive the population of cities.



https://pudding.cool/2018/12/3d-cities-story/

ThePudding

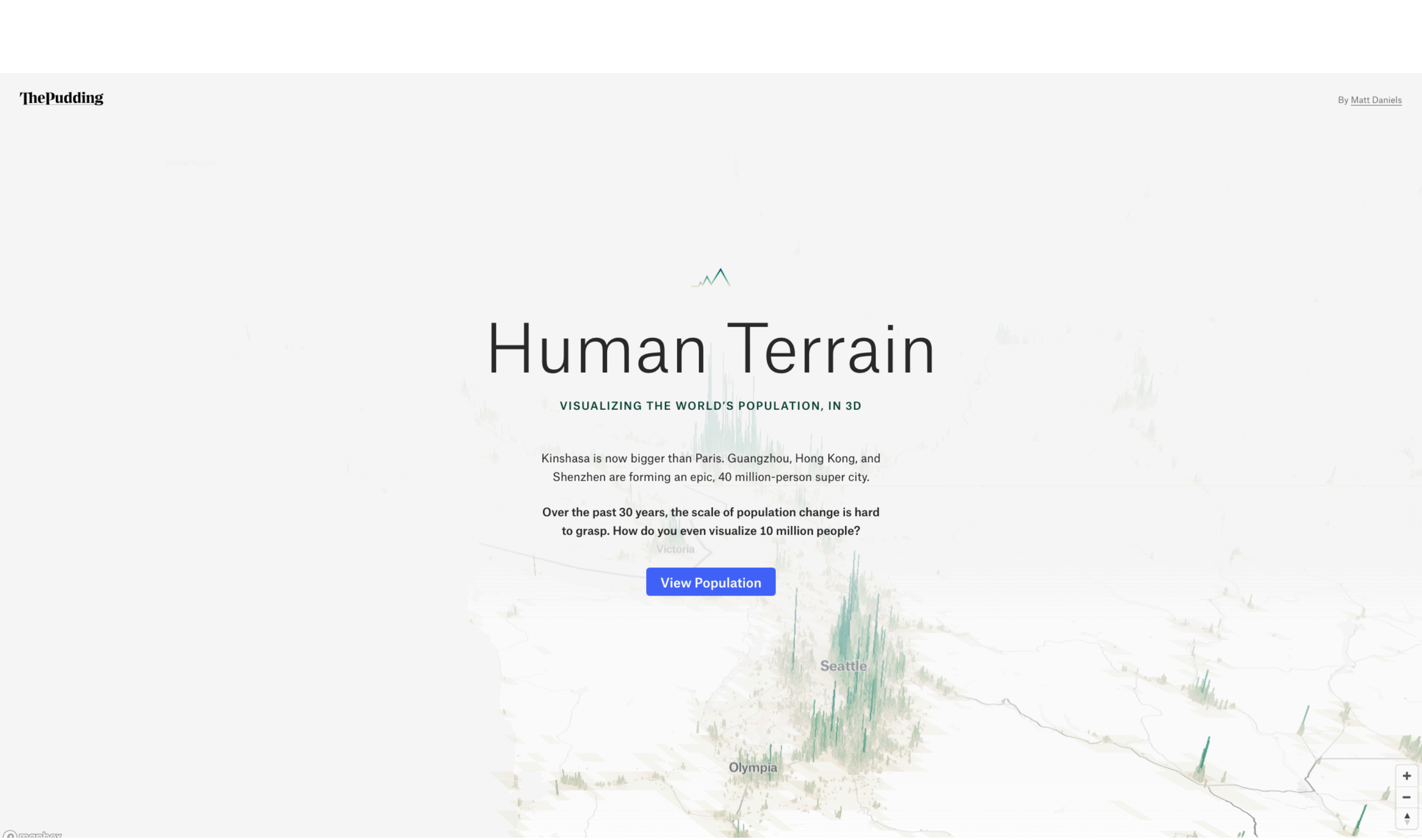
By Matt Daniels

Human Terrain

VISUALIZING THE WORLD'S POPULATION, IN 3D

Kinshasa is now bigger than Paris. Guangzhou, Hong Kong, and Shenzhen are forming an epic, 40 million-person super city.

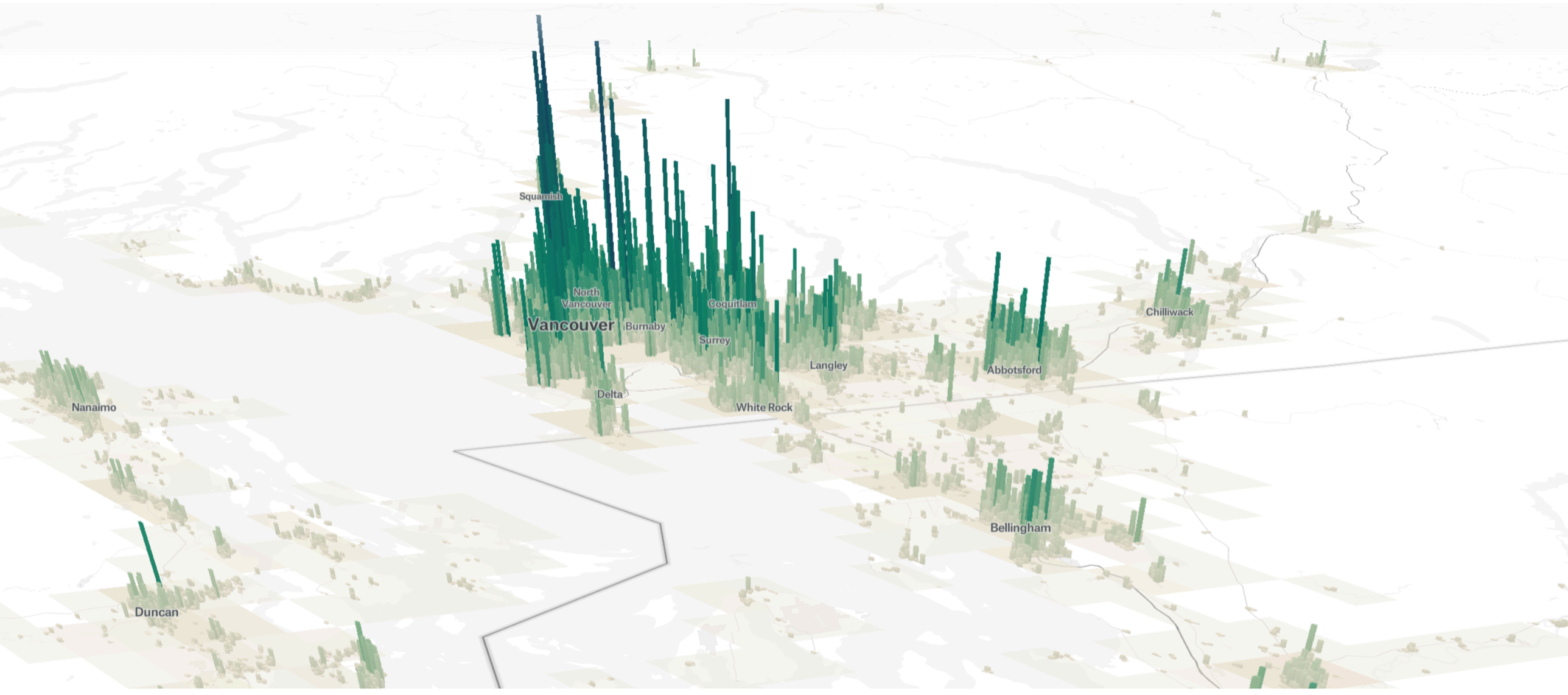
Over the past 30 years, the scale of population change is hard to grasp. How do you even visualize 10 million people?



What works well?

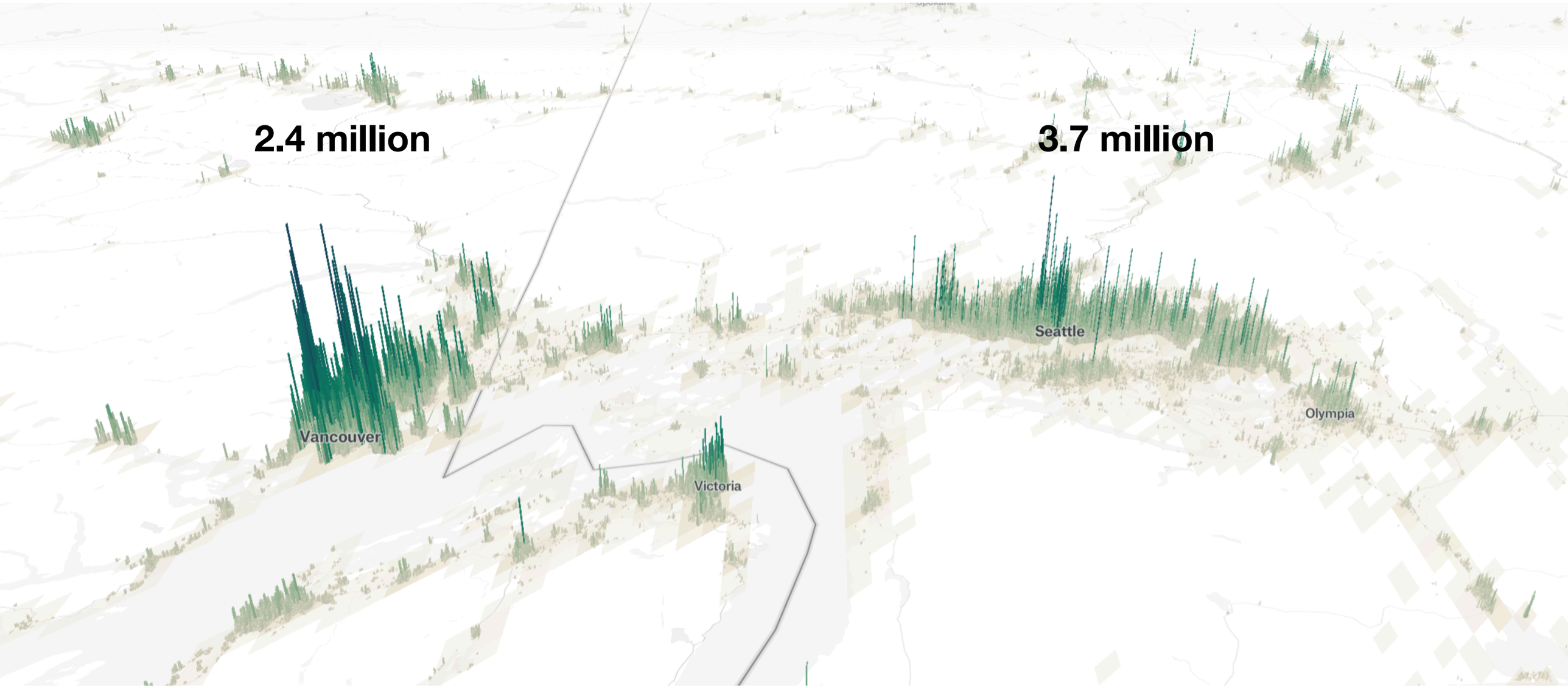
What could be improved?

https://pudding.cool/2018/10/city_3d/


Squamish
North Vancouver
Vancouver
Burnaby
Coquitlam
Surrey
Langley
Delta
White Rock
Abbotsford
Chilliwack
Nanaimo
Duncan
Bellingham

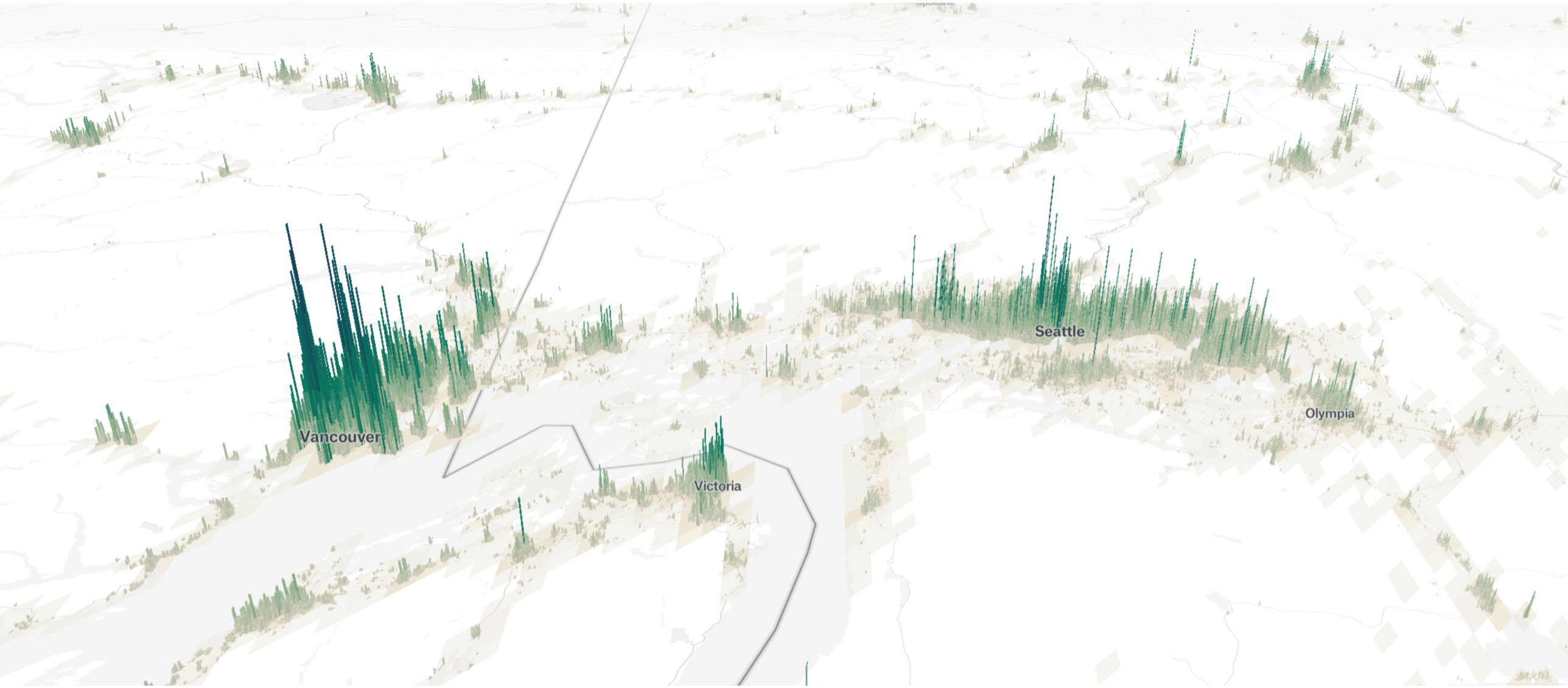
Occlusion problem

- tall bars block shorter ones and it’s hard to tell where the labels map



Which city is bigger?

- Colour + bar height emphasize tall bars (areas with dense population)



Other discussion points

- A scale would be helpful!

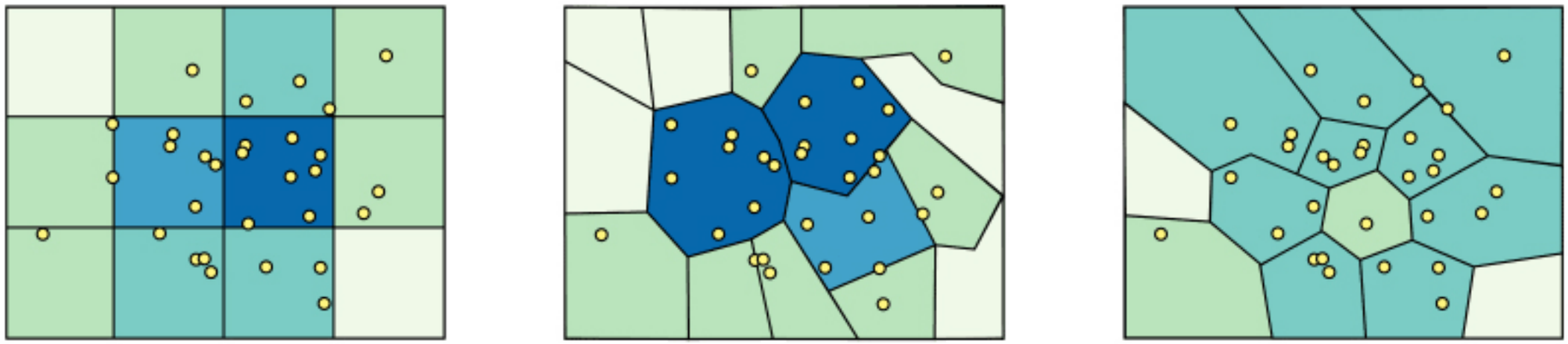
Zone effects | Where you draw the line matters

“…the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating.”

- Openshaw (1984)

Modifiable Areal Unit Problem
https://blog.cartographica.com/blog/2011/5/19/the-modifiable-areal-unit-problem-in-gis.html

Zone effects | Where you draw the line matters

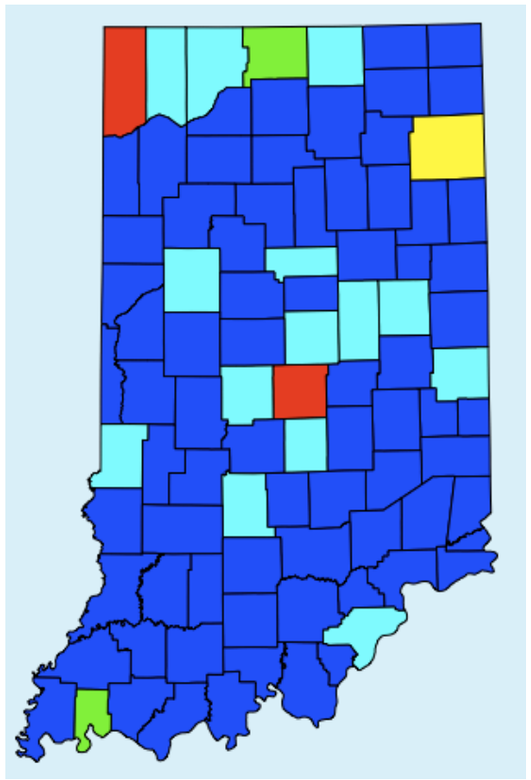


Different boundaries give different results

- gerrymandering (manipulating voting district boundaries) is one example

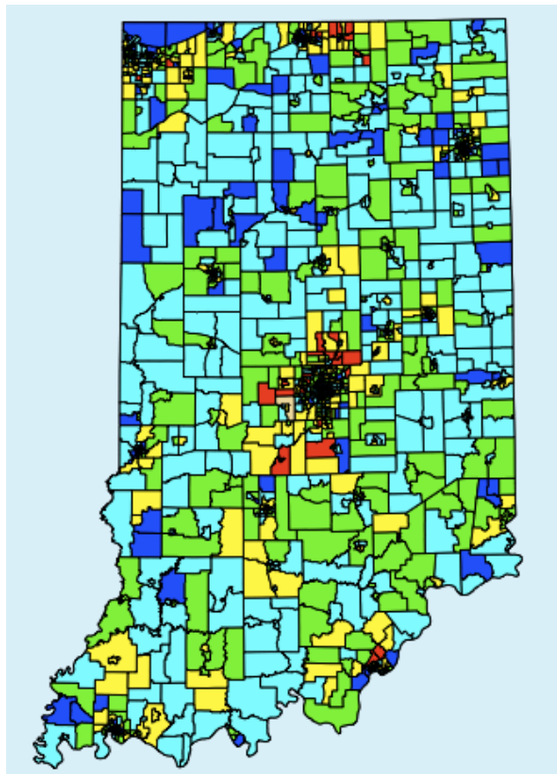
Scale effects | Bin size matters

Household counts across the state of Indiana



Counties

Larger bins, less variation



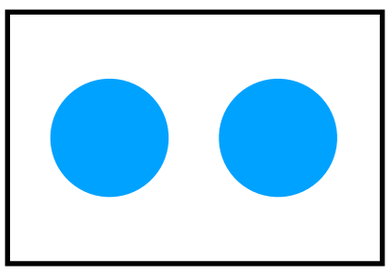
Census Tracts

Smaller bins, more variation

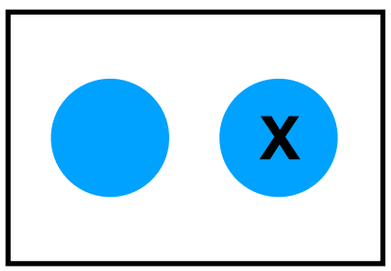
Modifiable Areal Unit Problem
https://blog.cartographica.com/blog/2011/5/19/the-modifiable-areal-unit-problem-in-gis.html

When populations are low, variation tends to be high

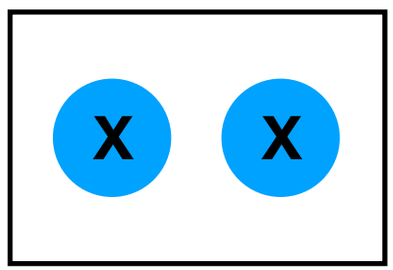
Imagine a population of 2 people



0% crime rate
Lowest in the country!



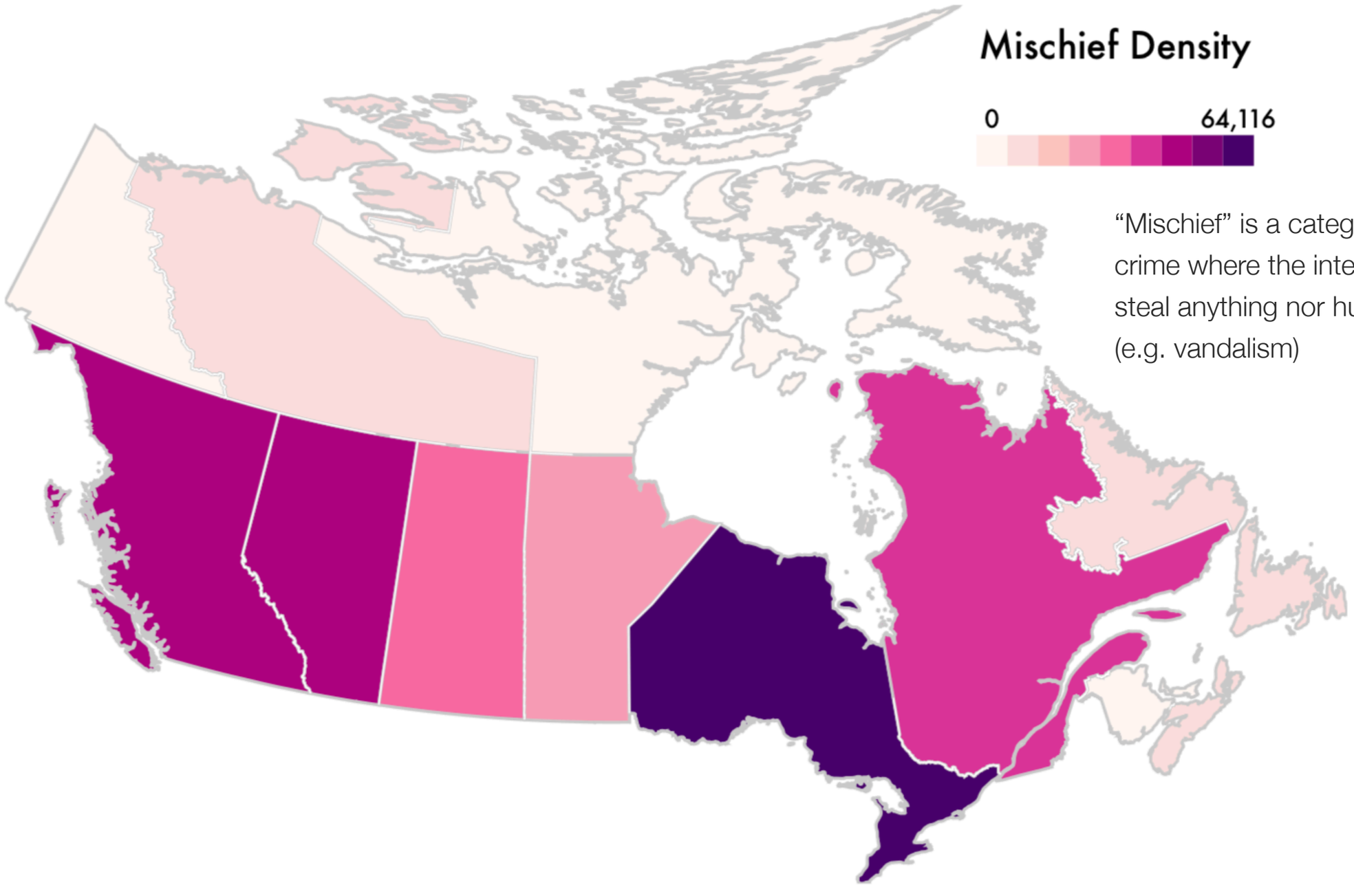
50% crime rate
Highest in the country!



100% crime rate
What kind of a place is this!

https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

Choropleth map | Event counts



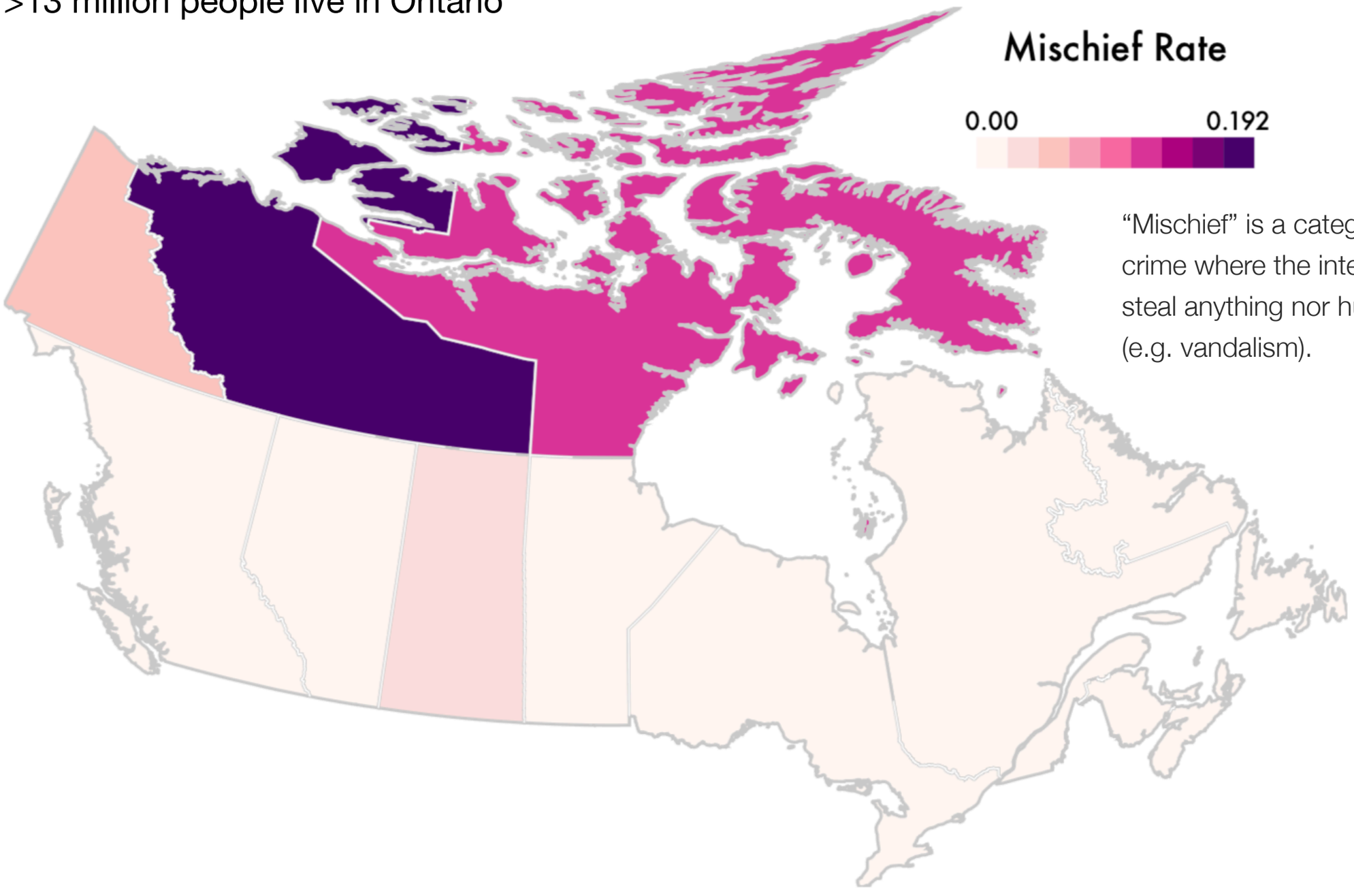
"Mischief" is a category of property crime where the intent is neither to steal anything nor hurt anybody (e.g. vandalism)

https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

Choropleth map | Per capita rate of crime

< 44,000 people live in the Northwest Territories

>13 million people live in Ontario

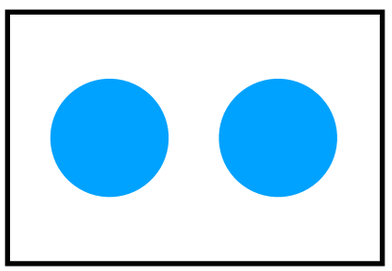


"Mischief" is a category of property crime where the intent is neither to steal anything nor hurt anybody (e.g. vandalism).

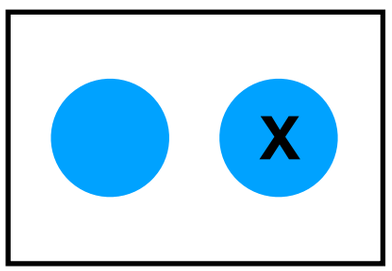
https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

When populations are low, variation tends to be high

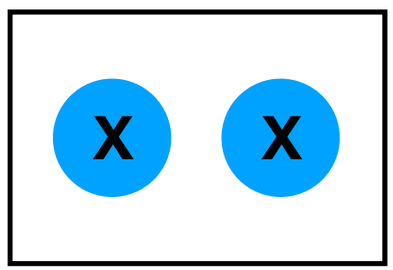
Imagine a population of 2 people



0% crime rate
Lowest in the country!



50% crime rate
Highest in the country!

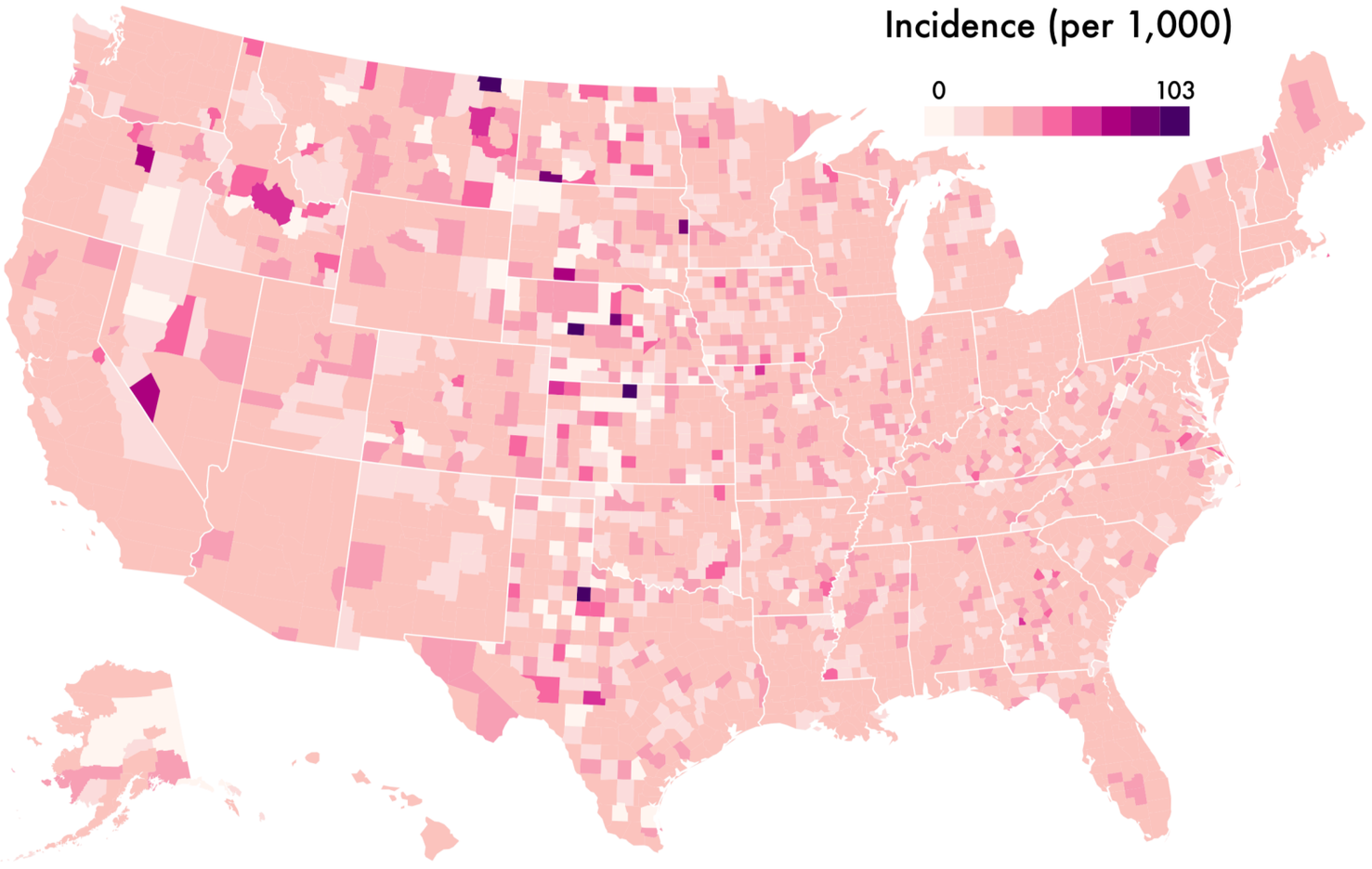


100% crime rate
What kind of a place is this!

None of these cases offer much evidence that this two-person province is really the safest or most dangerous place to live

https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

When populations are low, variation tends to be high



Random incidence

- Flip a coin for every person in the US

  (0.1% chance of success for each citizen, regardless of location)
- The apparent geographic patterns are an artifact of the high variance in counties with low population

https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

Surprise map

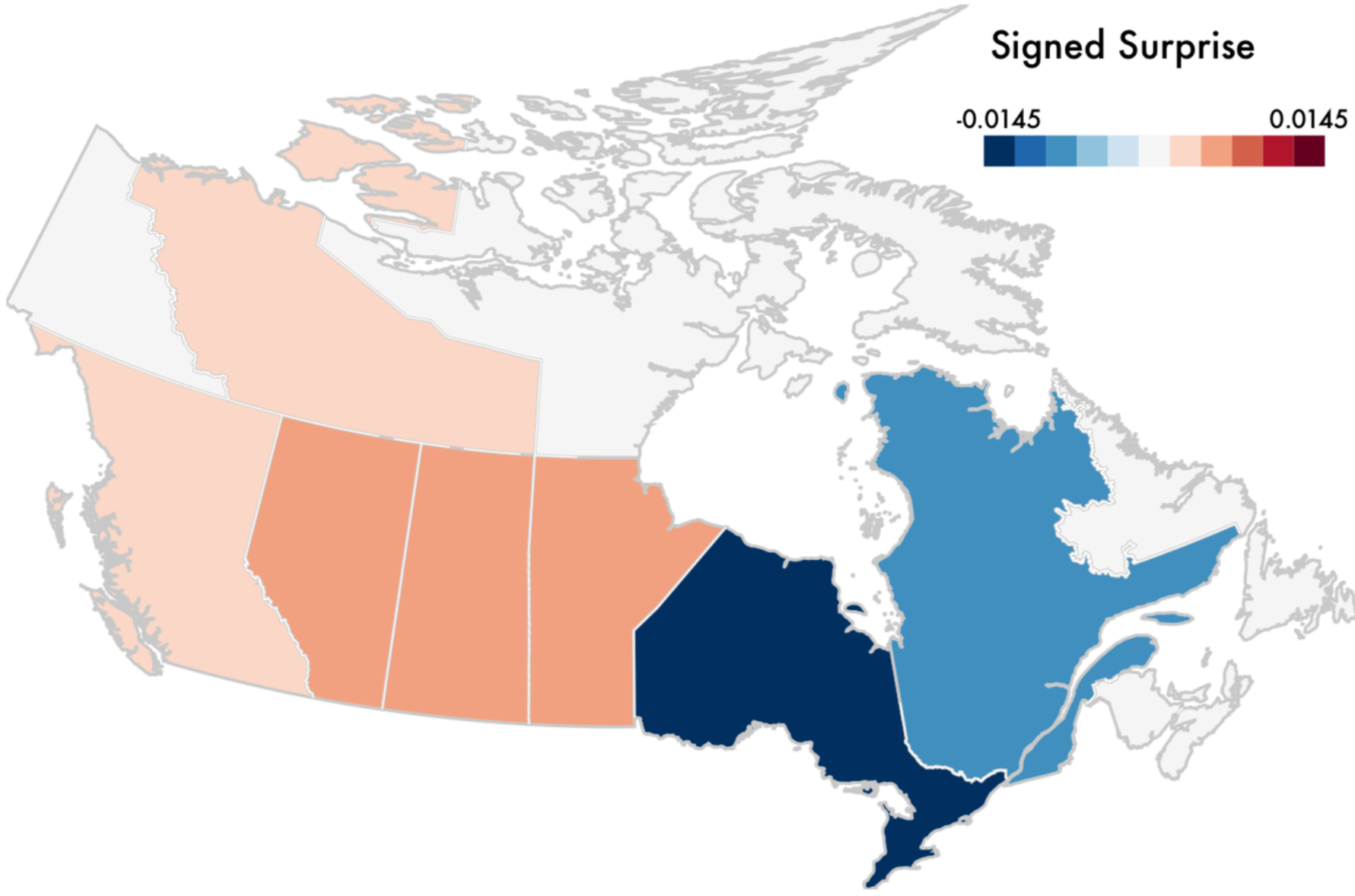
- **Let's assume two things**
  - If there is no big geographical differences in mischief, we'd expect each province to have the same per capita rate
  - If there is no big geographical difference in mischief, we'd expect variability to increase for smaller populations
- **Given these two models of how we expect the data to appear, we can measure deviations from these models**

https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

Surprise map

Bluish regions - where we have ***less*** mischief than we'd expect, given our models

Reddish regions - where we have ***more*** mischief than we'd expect, given our models

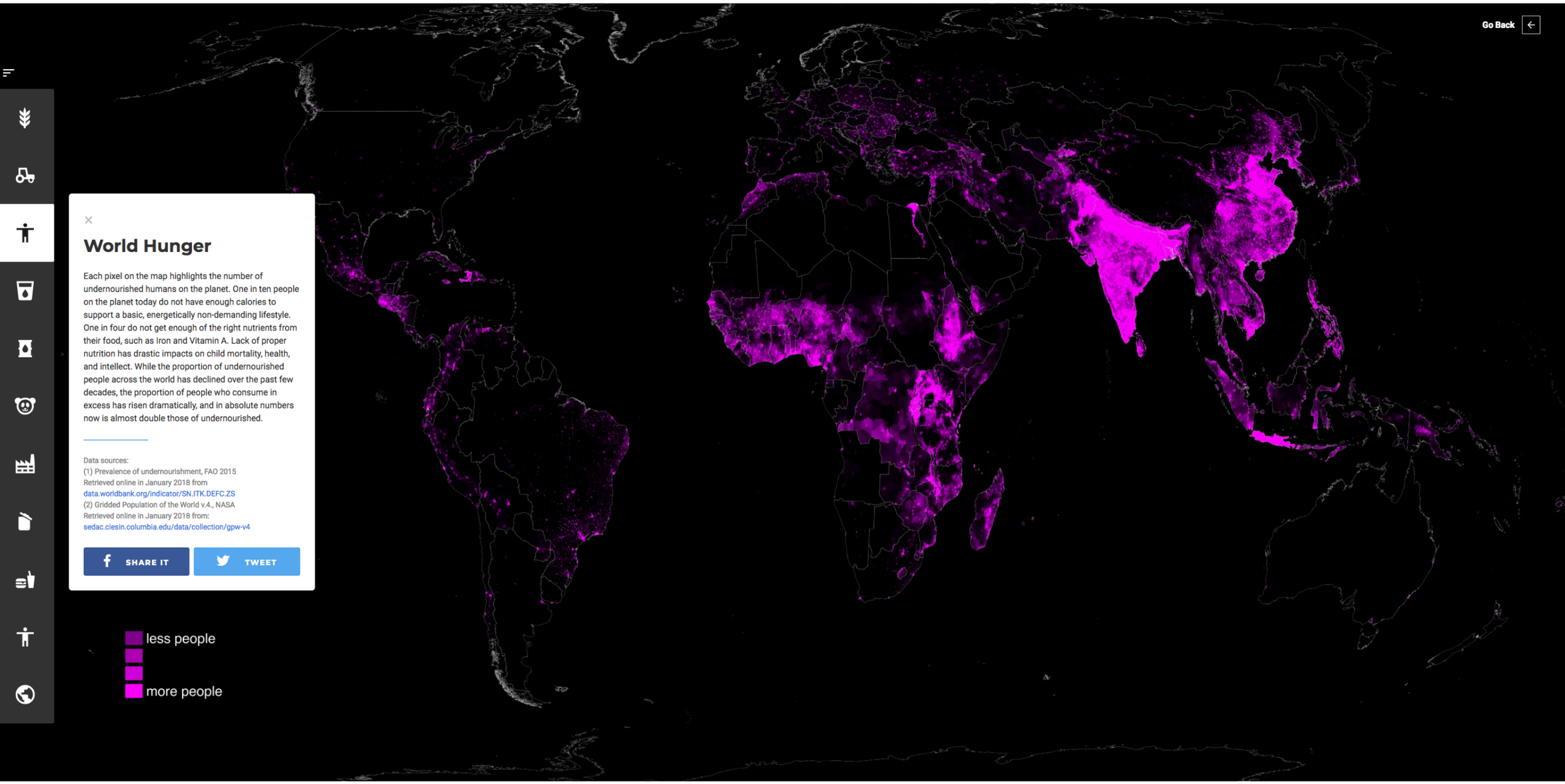


https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

Interested to read more…

Surprise maps

- **Paper** | http://idl.cs.washington.edu/papers/surprise-maps/
- **Code** | https://github.com/uwdata/bayesian-surprise



What works well?

What could be improved?

https://www.colours-of-food-security.com/exhibition/181/

Discussion points

- Beautiful and engaging visualization
  - Dramatic black background and bright colours
- Several maps just show population; could benefit from per capita normalization
- Difficult to see values on very small areas
- Do we need to show these data on a map?
  - Map helps spot differences between continents quickly (probably faster than reading labels on a bar chart)
  - Could complement with a small bar chart of countries with highest values (link to map); bars give equal visual weight to differently sized regions
- Could add zooming/panning interaction
- Consider visually distinguishing areas where no measure was taken (no data) and where measured data has a value of zero

Keypoints

Key points

- **Filtering**
  - Consider helping your user to make informed filtering decisions
    - Embed visuals in your widgets (scented widgets)
    - If appropriate, give visual cues as to how much data has been removed
- **Aggregation**
  - Binning is a great way to deal with large data sets (e.g. histograms, binned scatterplots)
  - Be aware that how you define your bins (shape and size) can change the resulting pattern
  - Avoid making population maps
    - Normalize per capita
    - Consider a surprise map approach

What I’ve been reading…

https://flowingdata.com