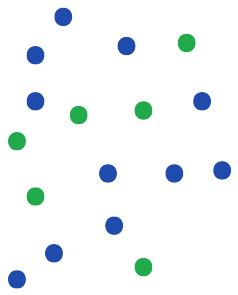


Data Exploration and Linked Views

Steven Braun

Data Analytics and Visualization Specialist

May 9, 2017

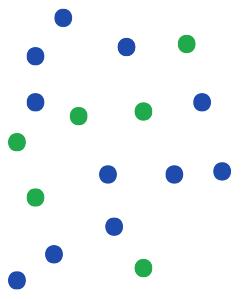


Data Collection

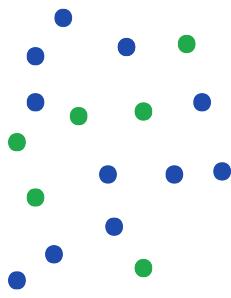
A lot can happen in the space between

A → B

Interpretation

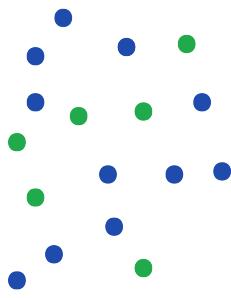


Data Collection

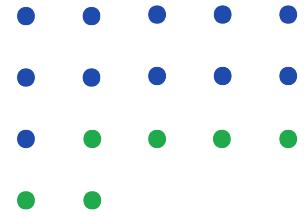


Data Collection

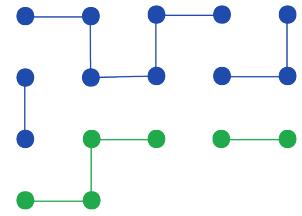
Structuring and Cleanup



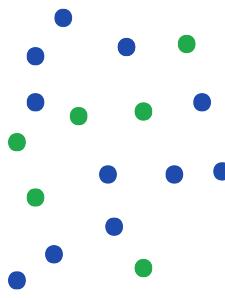
Data Collection



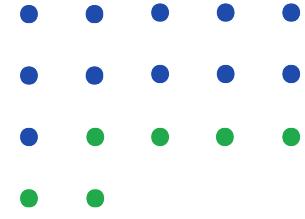
Structuring and Cleanup



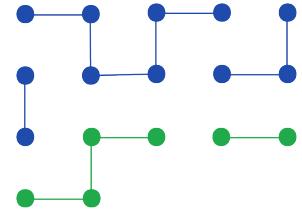
Analysis



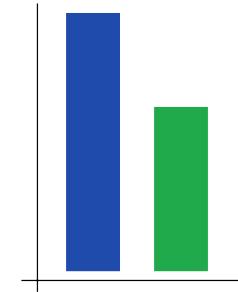
Data Collection



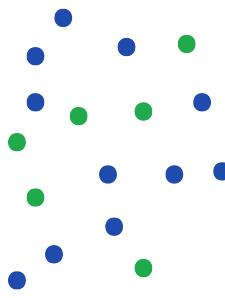
Structuring and Cleanup



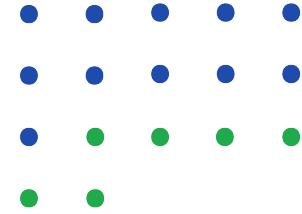
Analysis



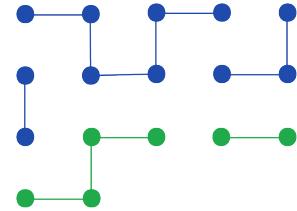
Visualization



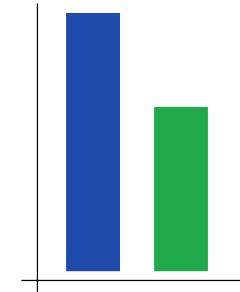
Data Collection



Structuring and Cleanup



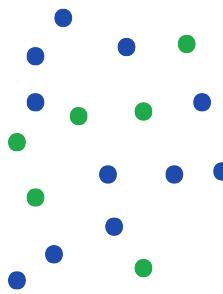
Analysis



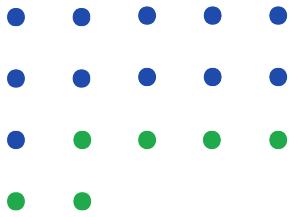
Visualization

A → B

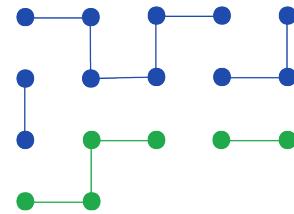
Interpretation



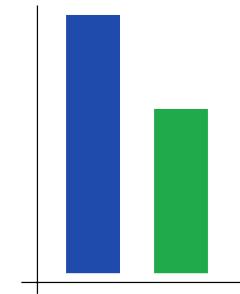
Data Collection



Structuring and Cleanup



Analysis



Visualization

A → B

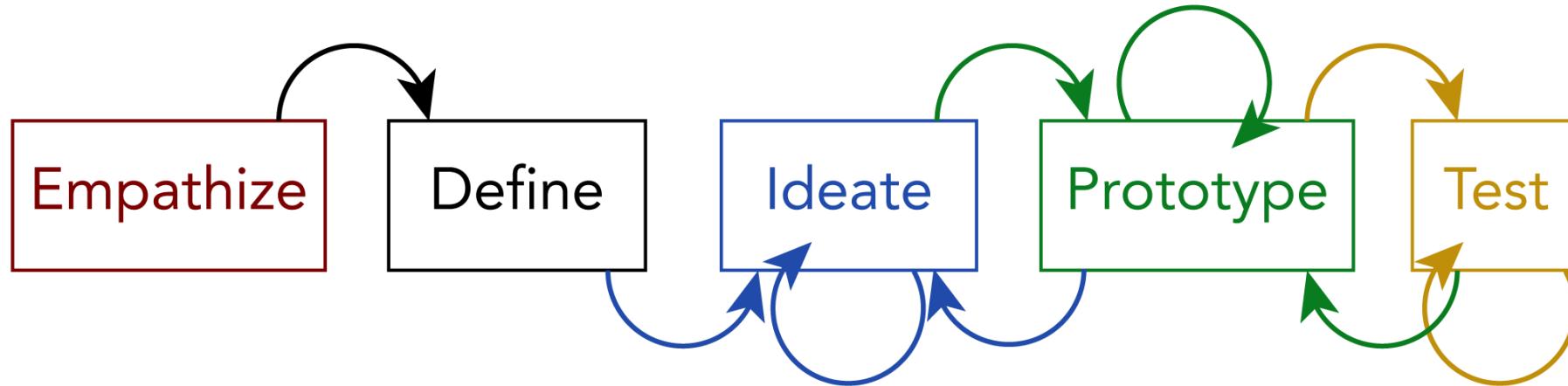
Interpretation

well-defined



ill-defined

DESIGN THINKING PROCESS



Stanford d.school

Why do we use data visualization
to explore our data?

Why do we use data visualization
to explore our data?

Visually displaying data makes it
easier to **find patterns**

Visualization helps make **complex data**
accessible and understandable

How?

By mapping visual and spatial cognition to
experiential knowledge

Data that seem statistically identical
may actually be graphically distinct

ANScombe's Quartet

I	
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

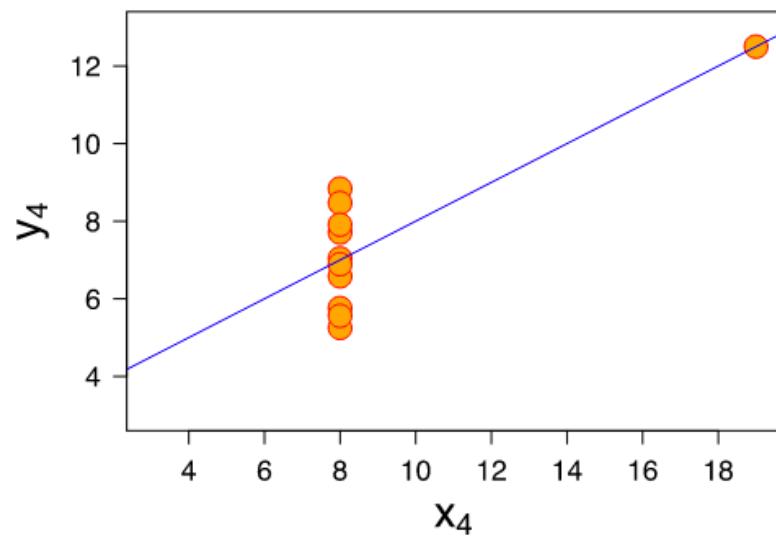
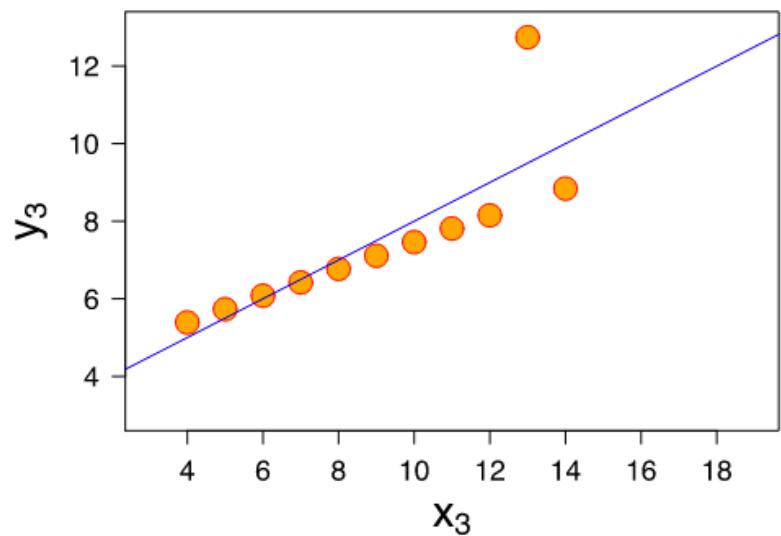
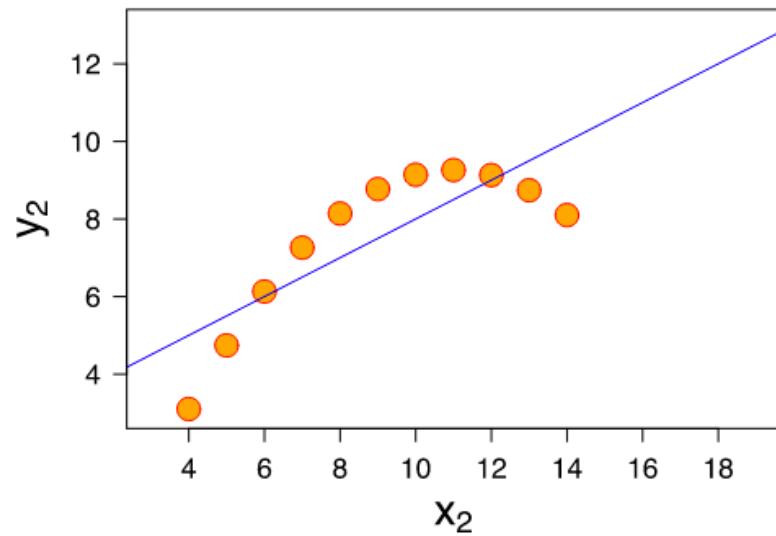
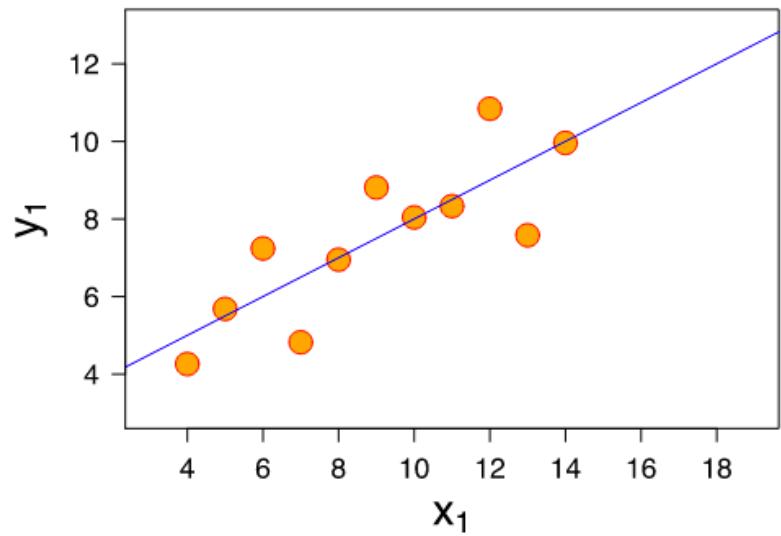
II	
x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

III	
x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

IV	
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

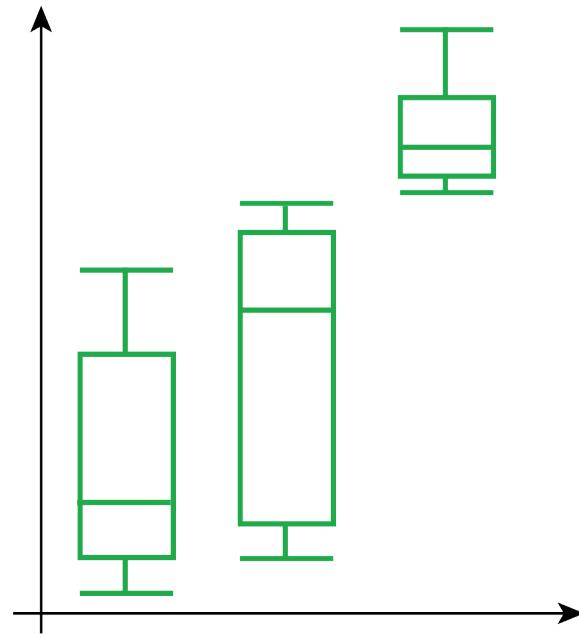
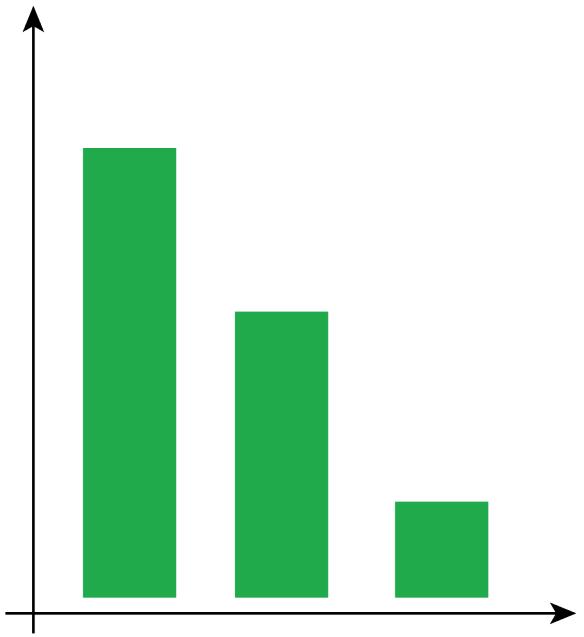
ANScombe's Quartet

Mean of x	9
Variance of x	11
Mean of y	7.5
Variance of y	4.125
Correlation between x and y	0.816
Linear regression	$y = 3.00 + 0.500x$



Lesson Learned

Be cautious of analyses that collapse distributions into summary representations – including charts



One way we can use visualization to explore data is through making charts and graphs **interactive**

DATA EXPLORATION TECHNIQUES

Zooming and panning

Highlighting

Brushing

Filtering (faceting)

Eliding

Details on demand

Animation

DATA EXPLORATION TECHNIQUES

Zooming and panning

Highlighting

Brushing

Filtering (faceting)

Eliding

Details on demand

Animation

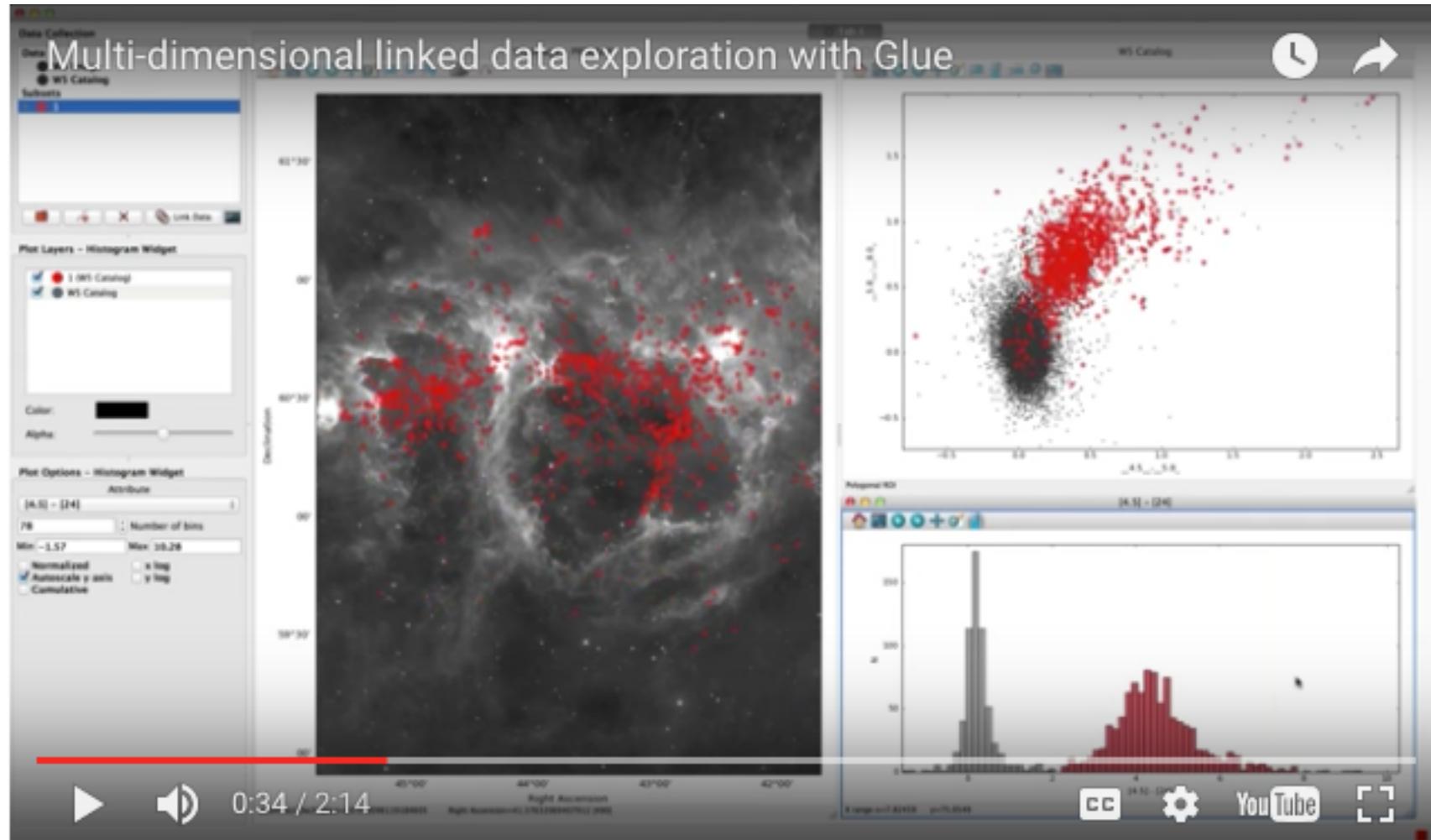


Linking

LINKED VIEWS

Interaction with data in one chart
(view) is mirrored or translated in
another chart

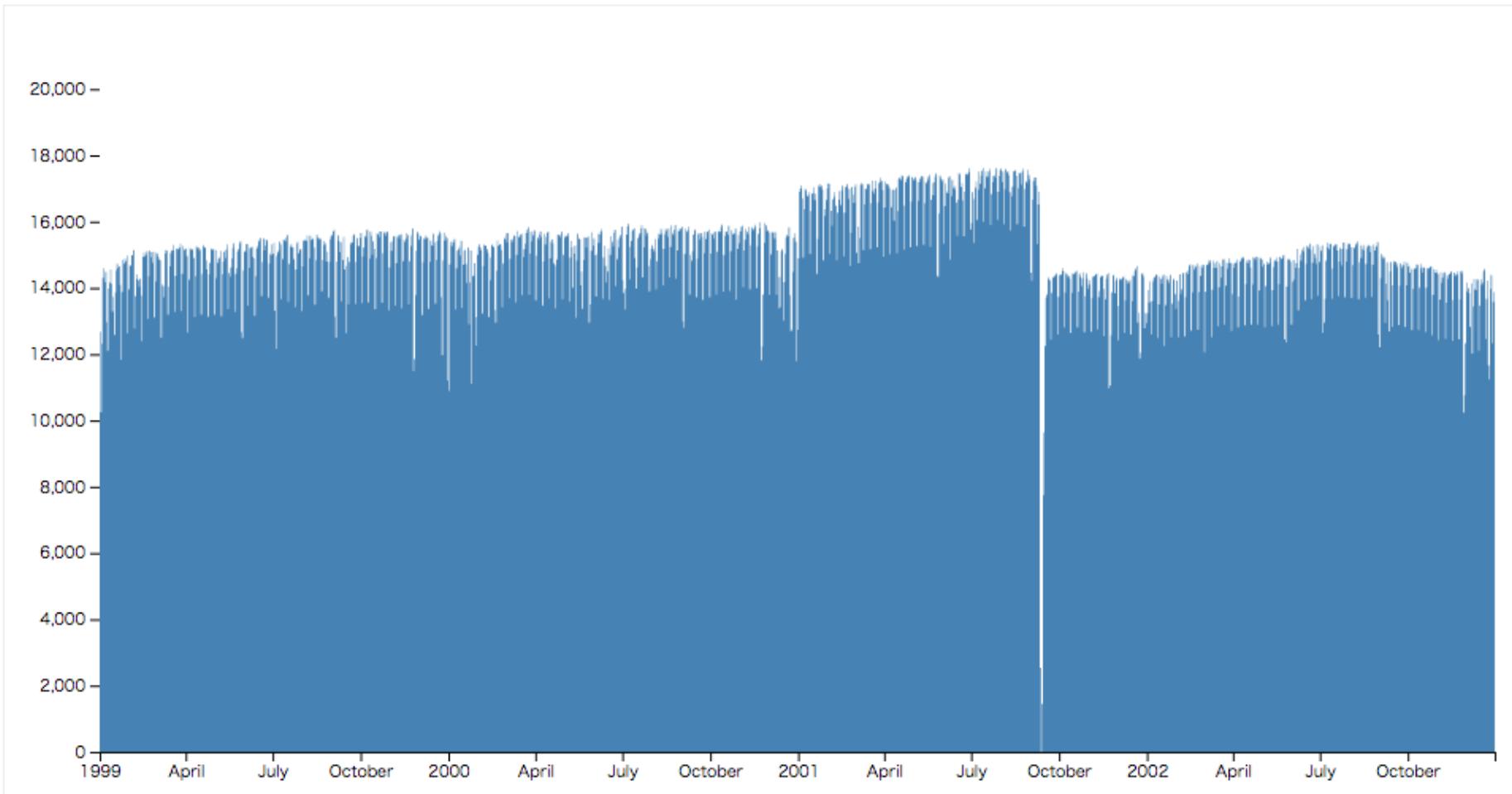
LINKED VIEWS



<https://www.youtube.com/watch?v=qO3RQiRjWA4>

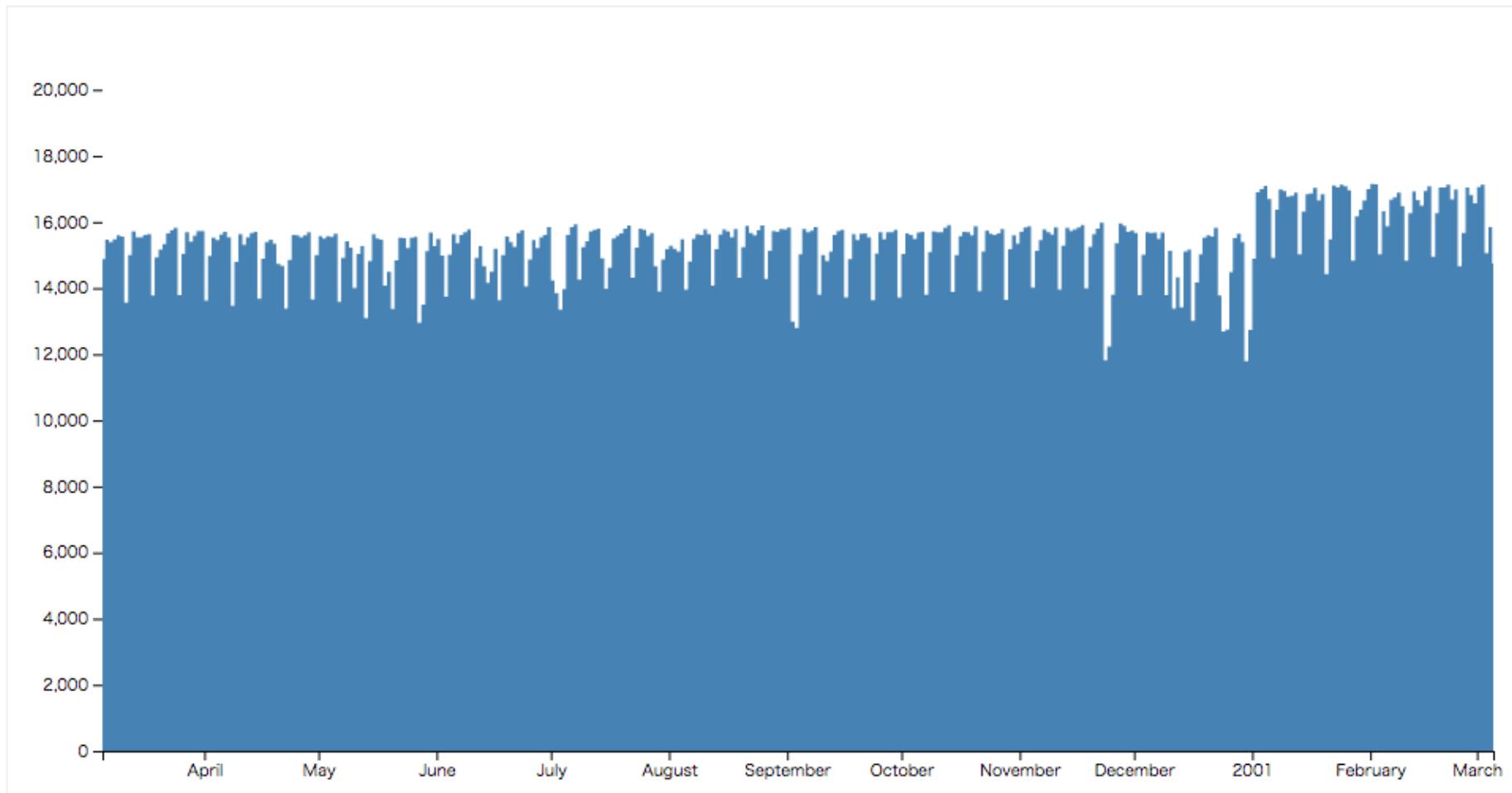
Interactivity enables
focus + context

ZOOMING AND PANNING



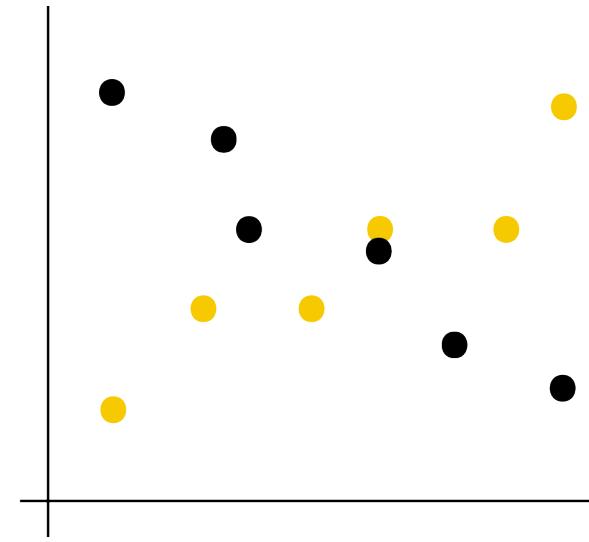
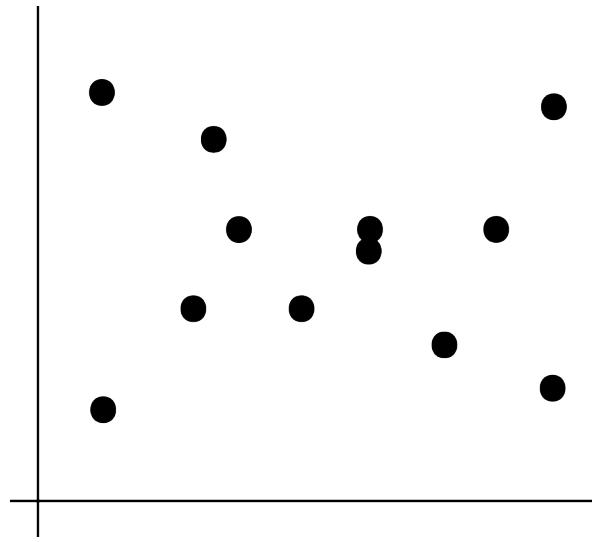
<https://bl.ocks.org/mbostock/4015254>

ZOOMING AND PANNING

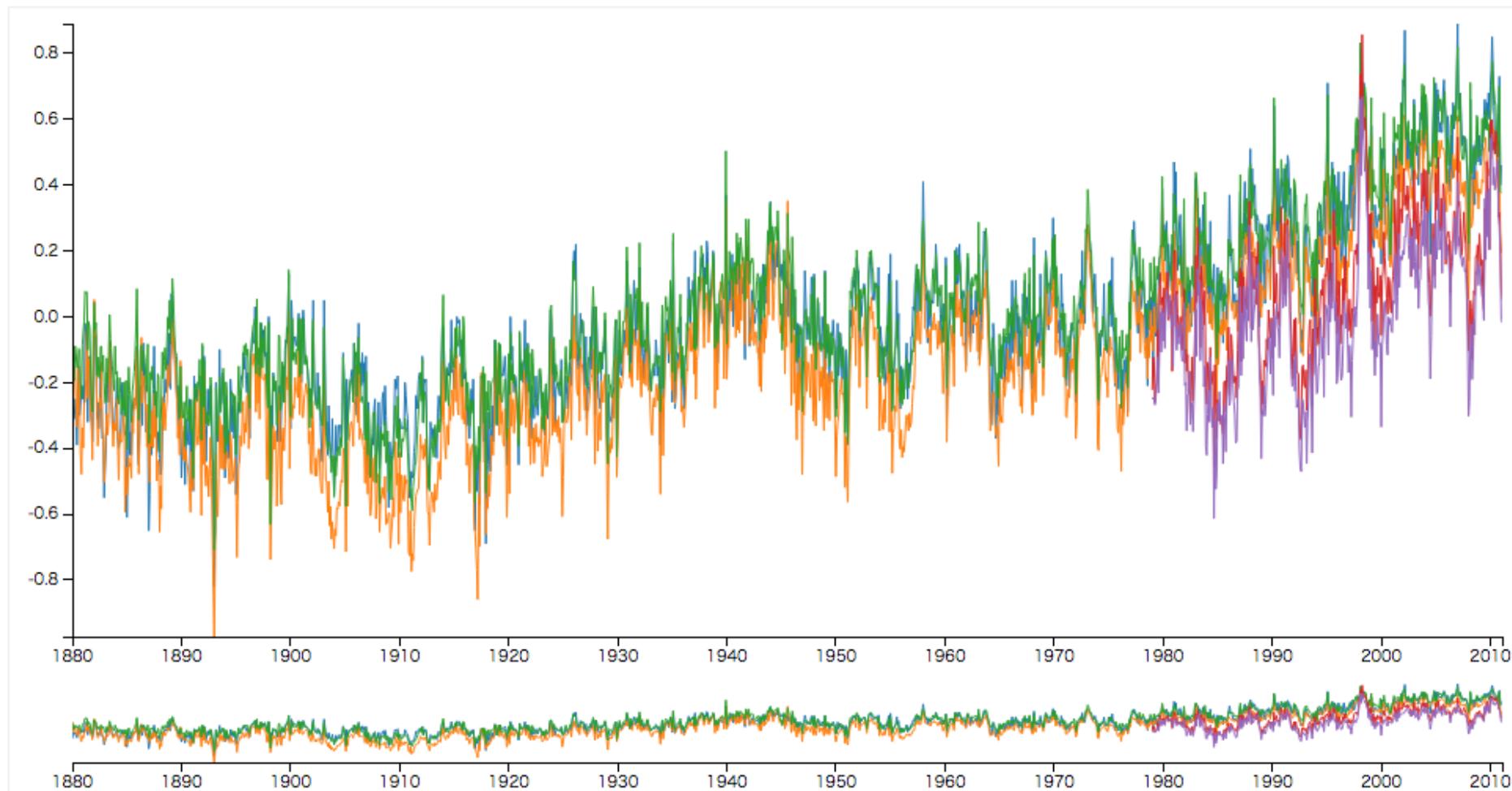


<https://bl.ocks.org/mbostock/4015254>

HIGHLIGHTING

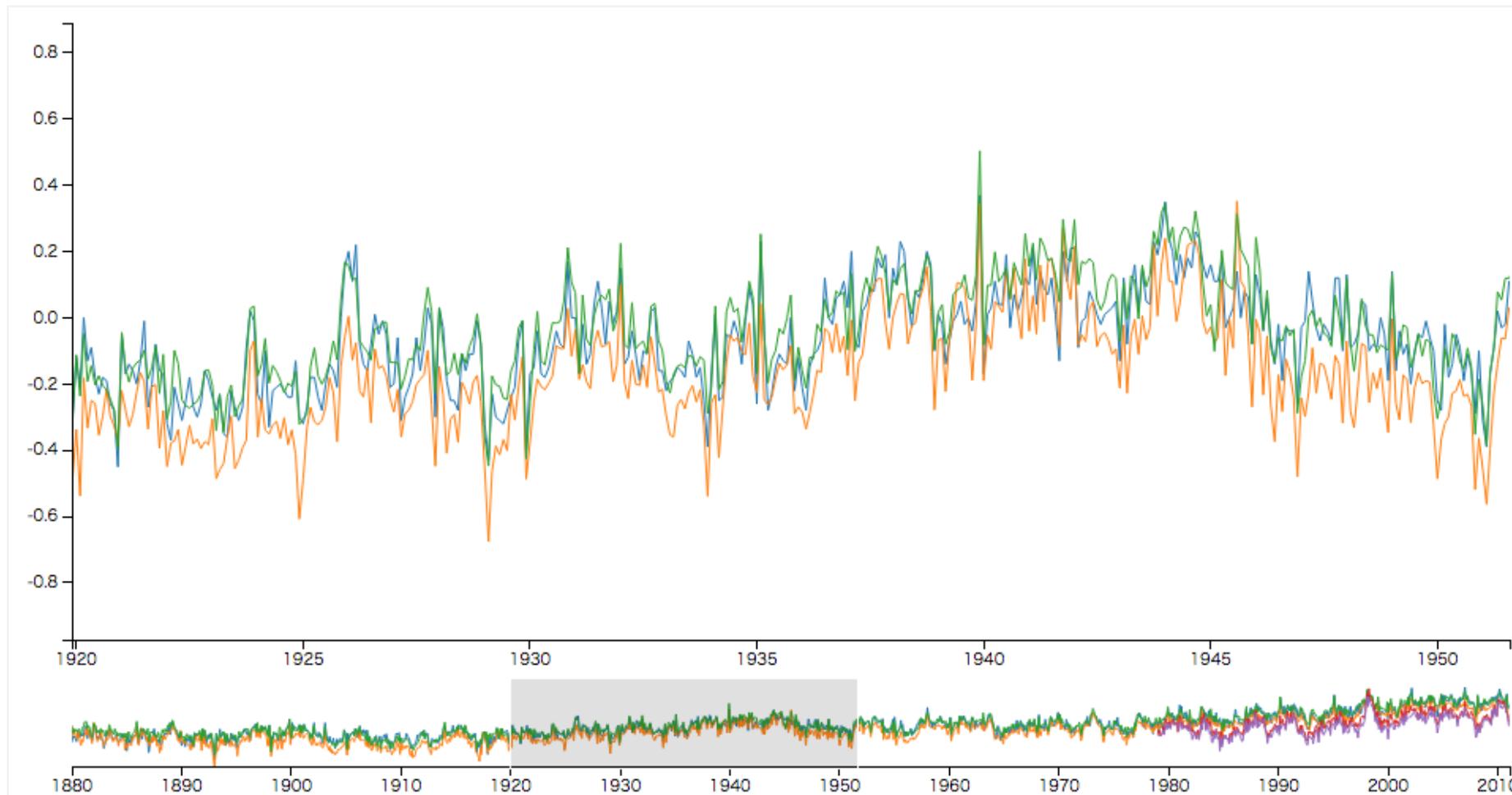


BRUSHING



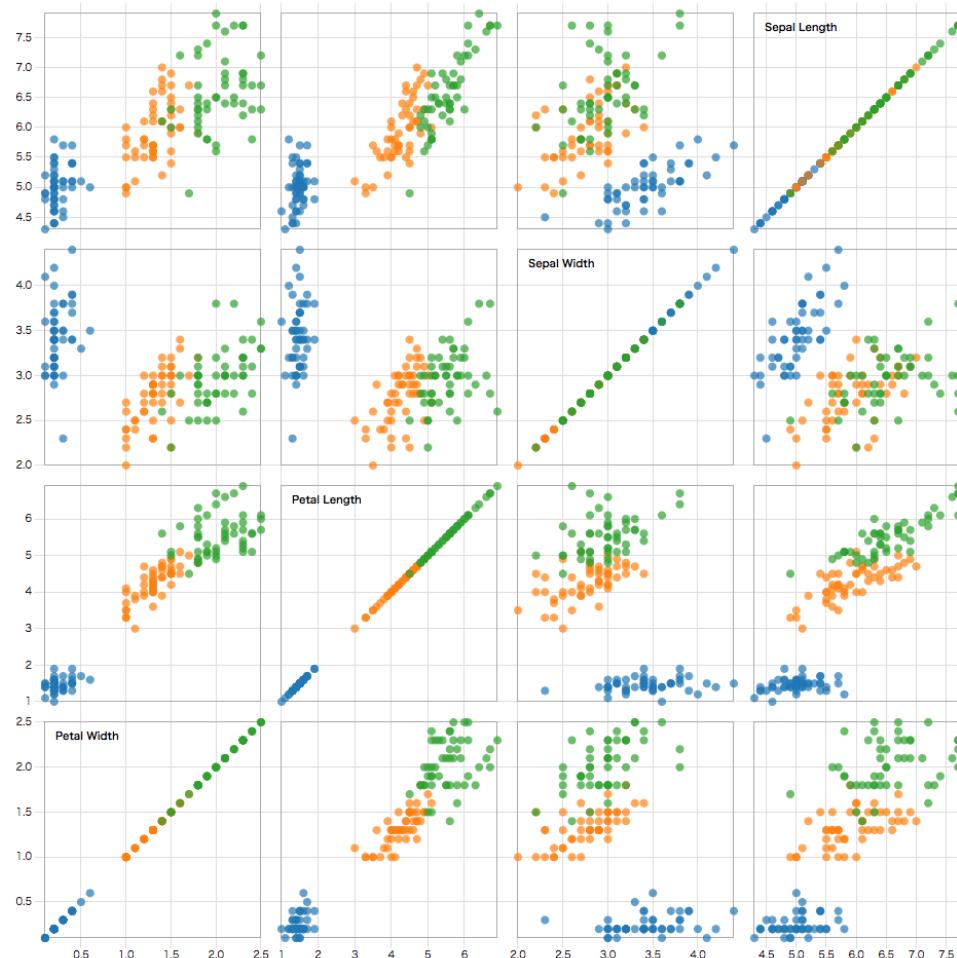
<http://blocks.org/natemiller/7dec148bb6aab897e561>

BRUSHING



<http://blocks.org/natemiller/7dec148bb6aab897e561>

BRUSHING



<https://bl.ocks.org/mbostock/4063663>

FILTERING AND FACETING

Side-by-side

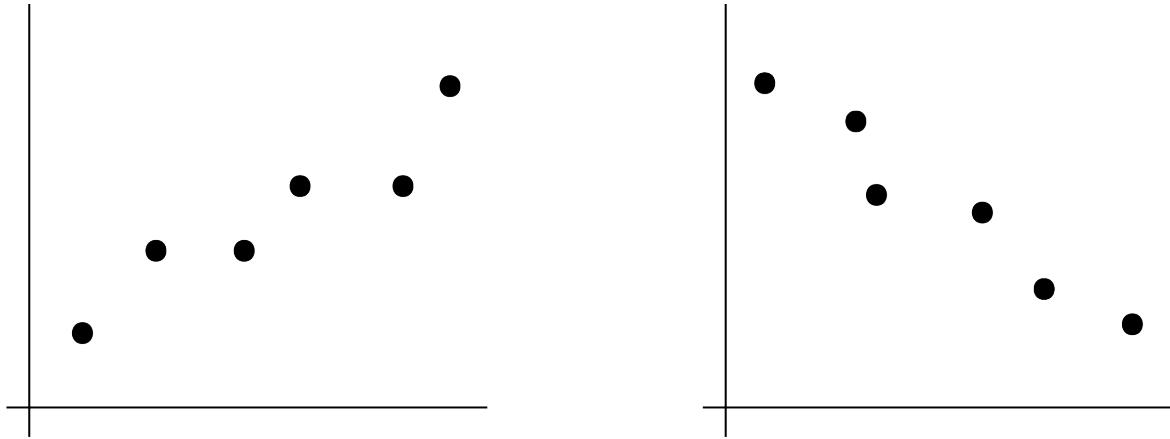
Juxtapose data classes using common axis

Layering

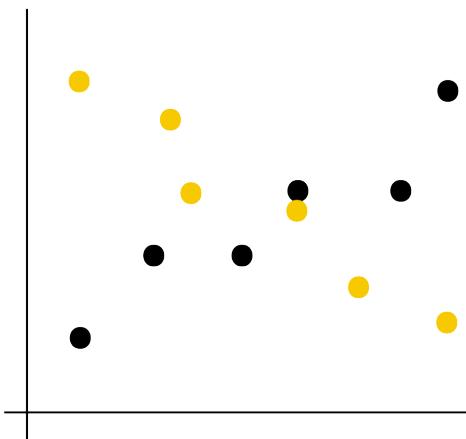
Superimpose data classes with different encodings

FILTERING AND FACETING

Side-by-side

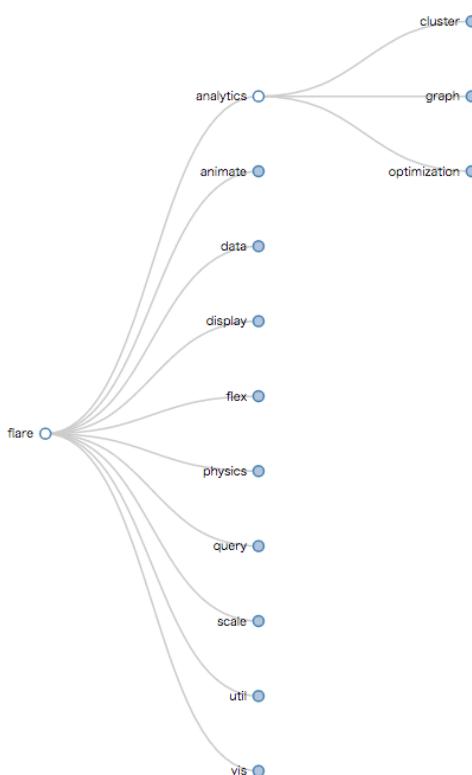


Layered



ELIDING

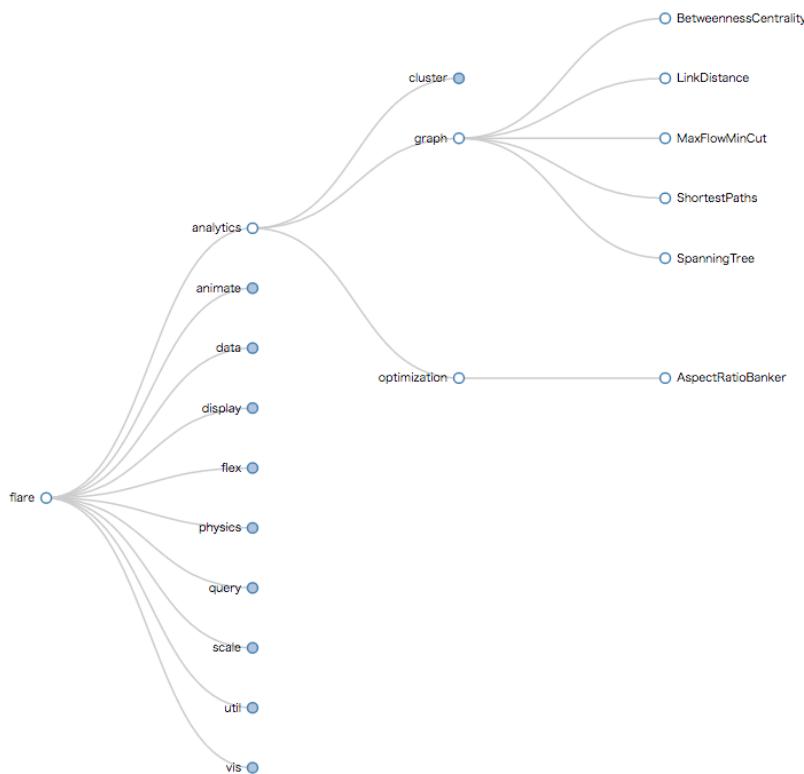
Focus on selected item to show detail while remaining data are shown in summary



<https://bl.ocks.org/mbostock/4339083>

ELIDING

Focus on selected item to show detail while remaining data are shown in summary



<https://bl.ocks.org/mbostock/4339083>

reductionist reading

vs

holistic reading

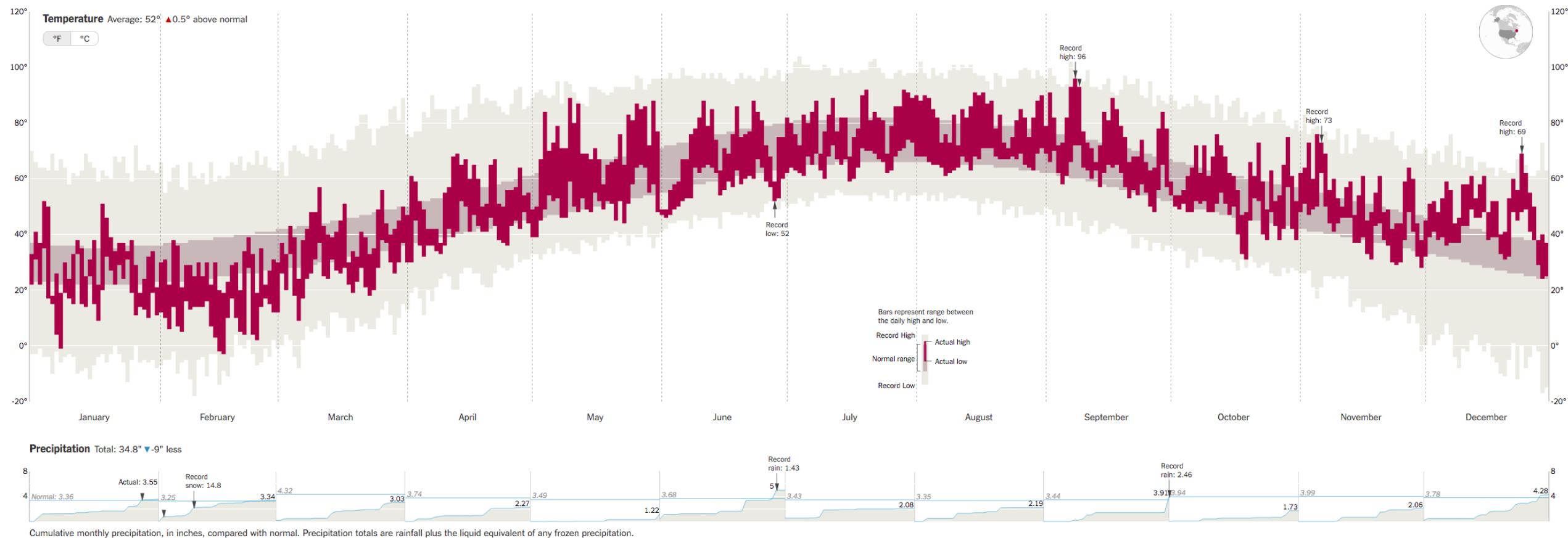
How Much Warmer Was Your City in 2015?

By K.K. REBECCA LAI FEB. 19, 2016

Scientists declared that 2015 was Earth's [hottest year on record](#). In a database of 3,116 cities provided by AccuWeather, about 90 percent of them were warmer than normal. Enter your city in the field below to see how much warmer it was last year.

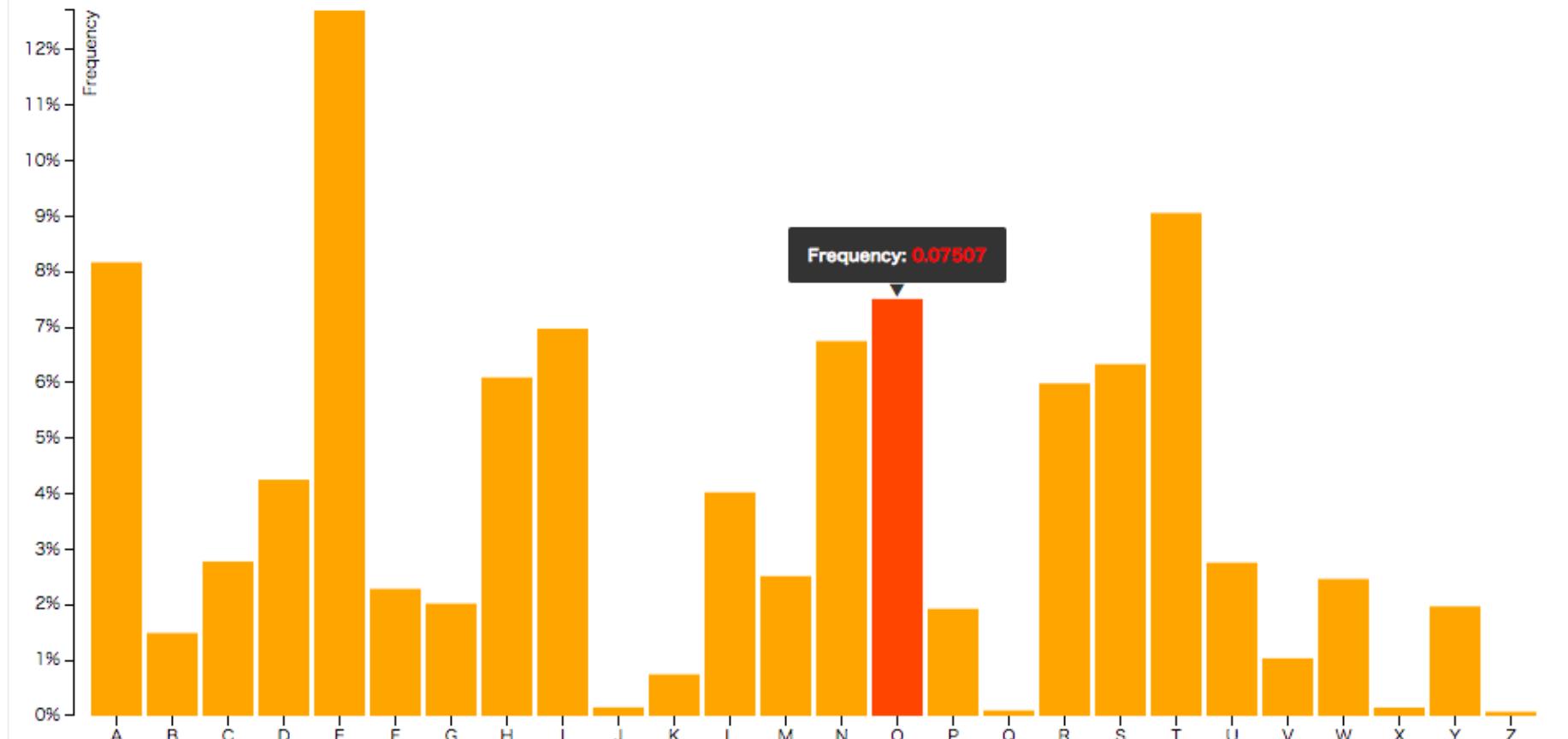
RELATED ARTICLE

Boston, Mass.



DETAILS ON DEMAND

Using d3-tip to add tooltips to a d3 bar chart.



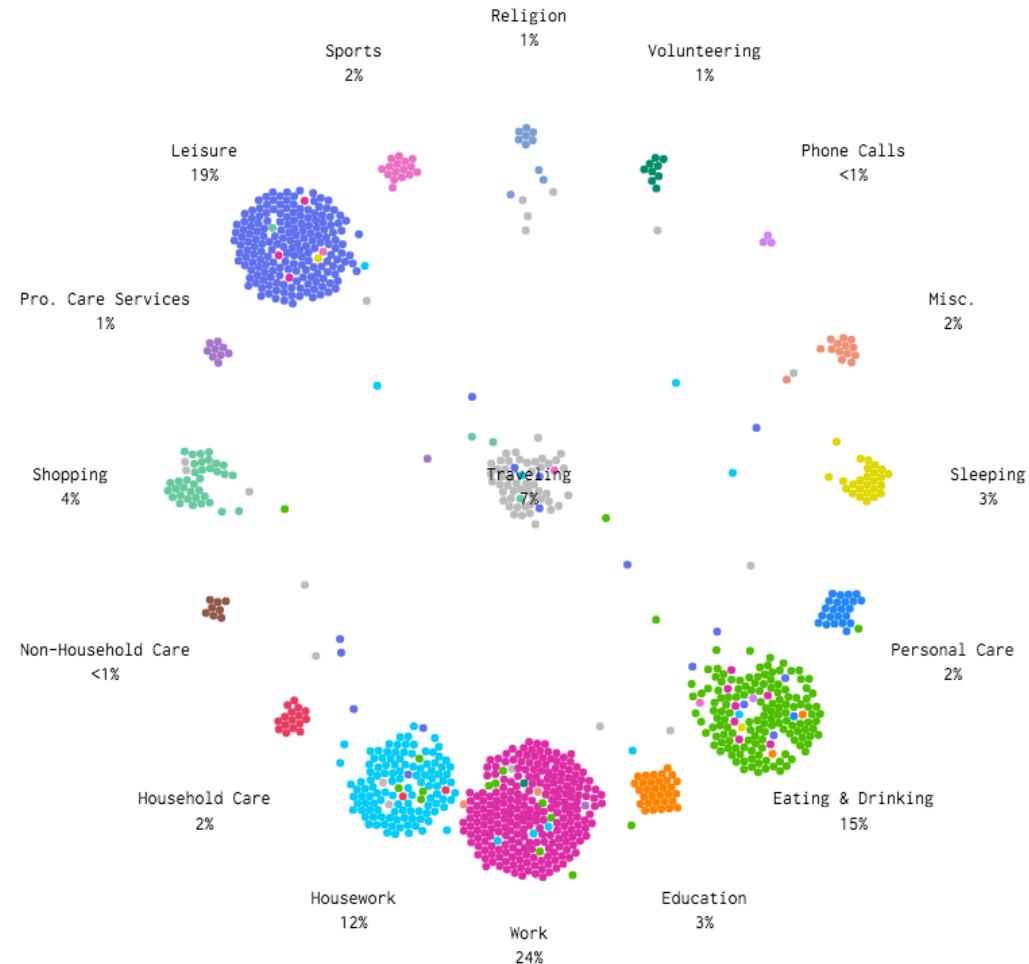
ANIMATION

12:16pm

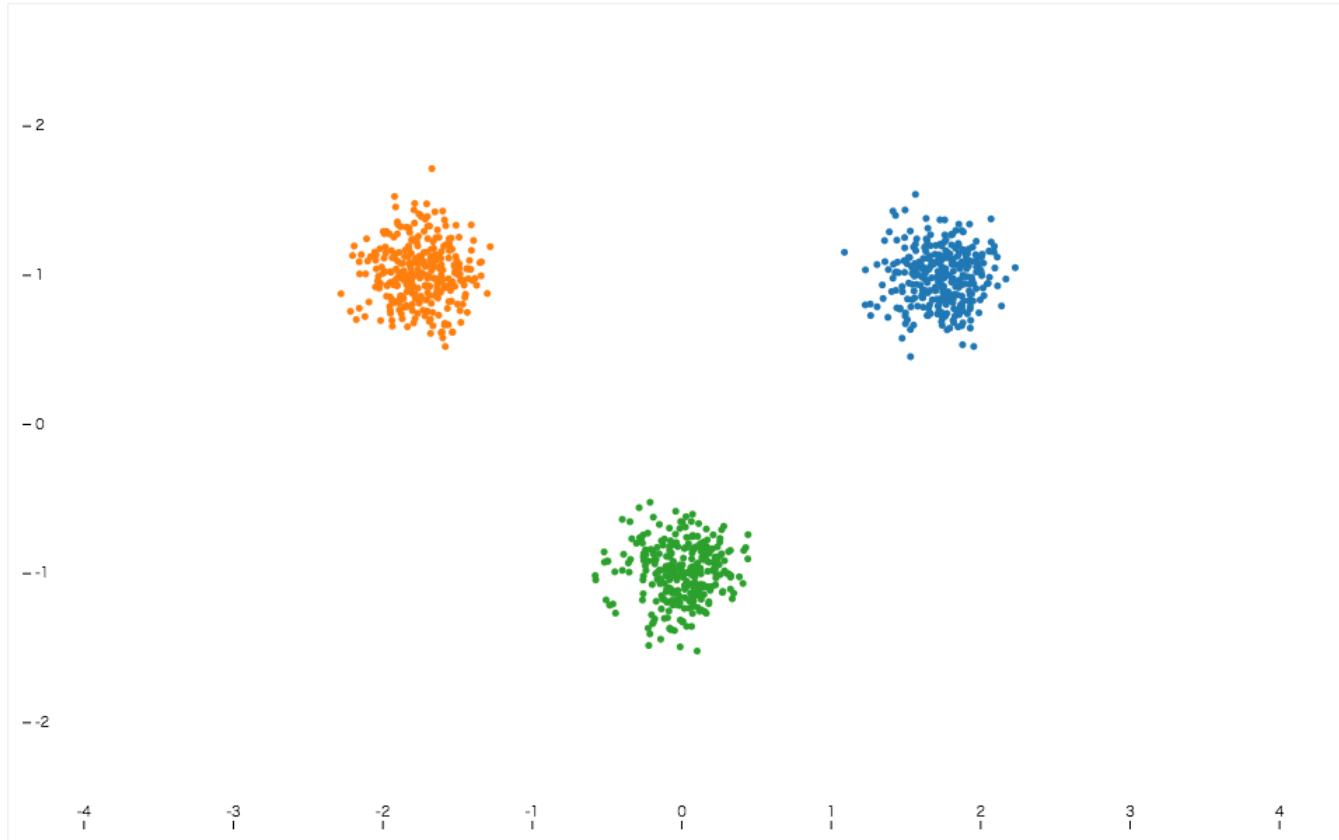
SLOW MEDIUM FAST

Lunch hour. Many go eat, but there's still activity throughout. You see a small shift at the end of the hour.

This is a simulation of 1,000 people's average day. It's based on 2014 data from the [American Time Use Survey](#), made way more accessible by the [ATUS Extract Builder](#).

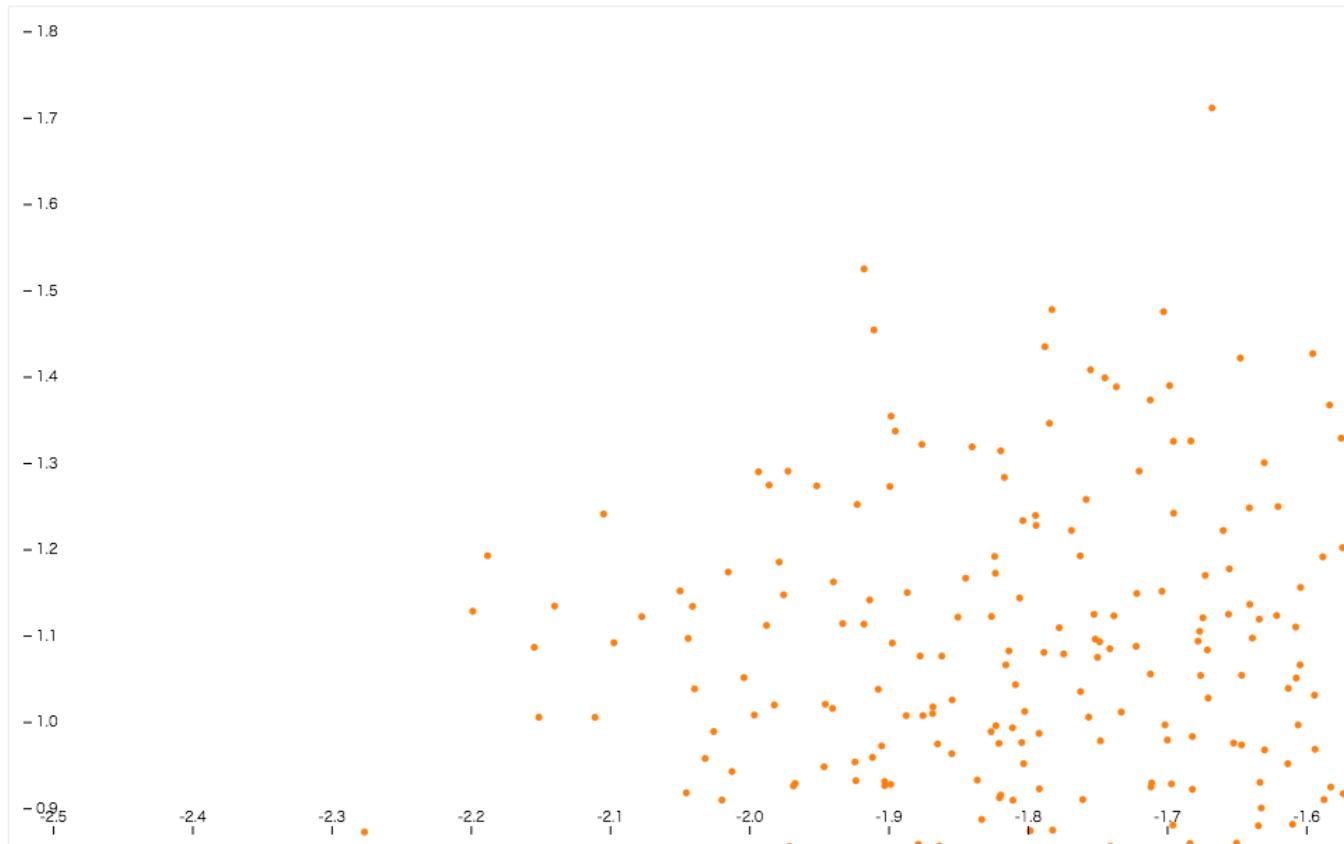


COMBINING INTERACTIVITY



<https://bl.ocks.org/mbostock/f48fcdb929a620ed97877e4678ab15e6>

COMBINING INTERACTIVITY



<https://bl.ocks.org/mbostock/f48fcdb929a620ed97877e4678ab15e6>

How can we embed “interactivity” and exploration into static visualizations?

STRATEGIES FOR STATIC VISUALIZATIONS

Small multiples

Callouts (selection/highlighting)

Filtering and faceting

Multiform encoding

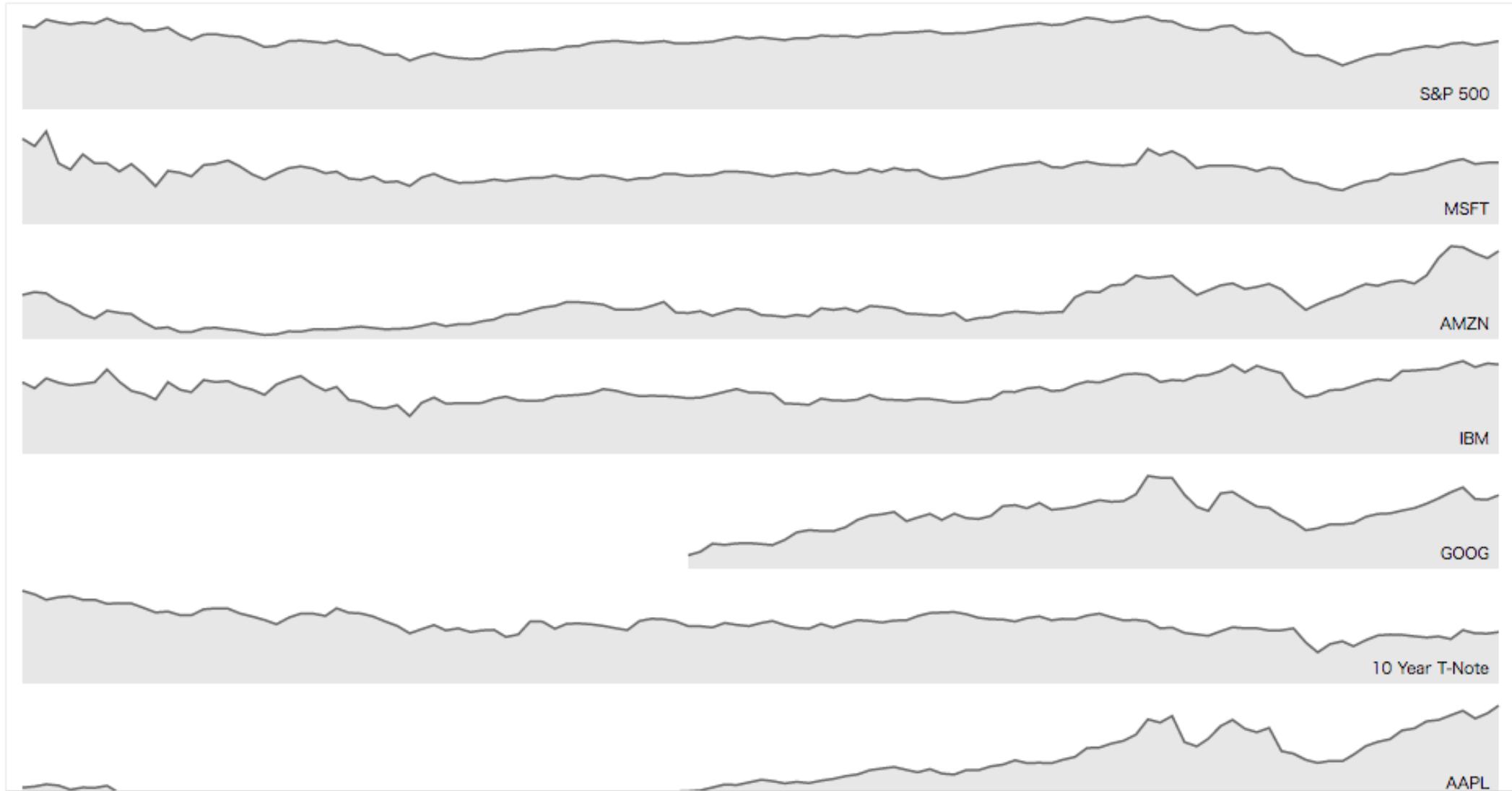
STRATEGIES FOR STATIC VISUALIZATIONS

Small multiples

Callouts (selection/highlighting)

Filtering and faceting

Multiform encoding



<https://bl.ocks.org/mbostock/1157787>

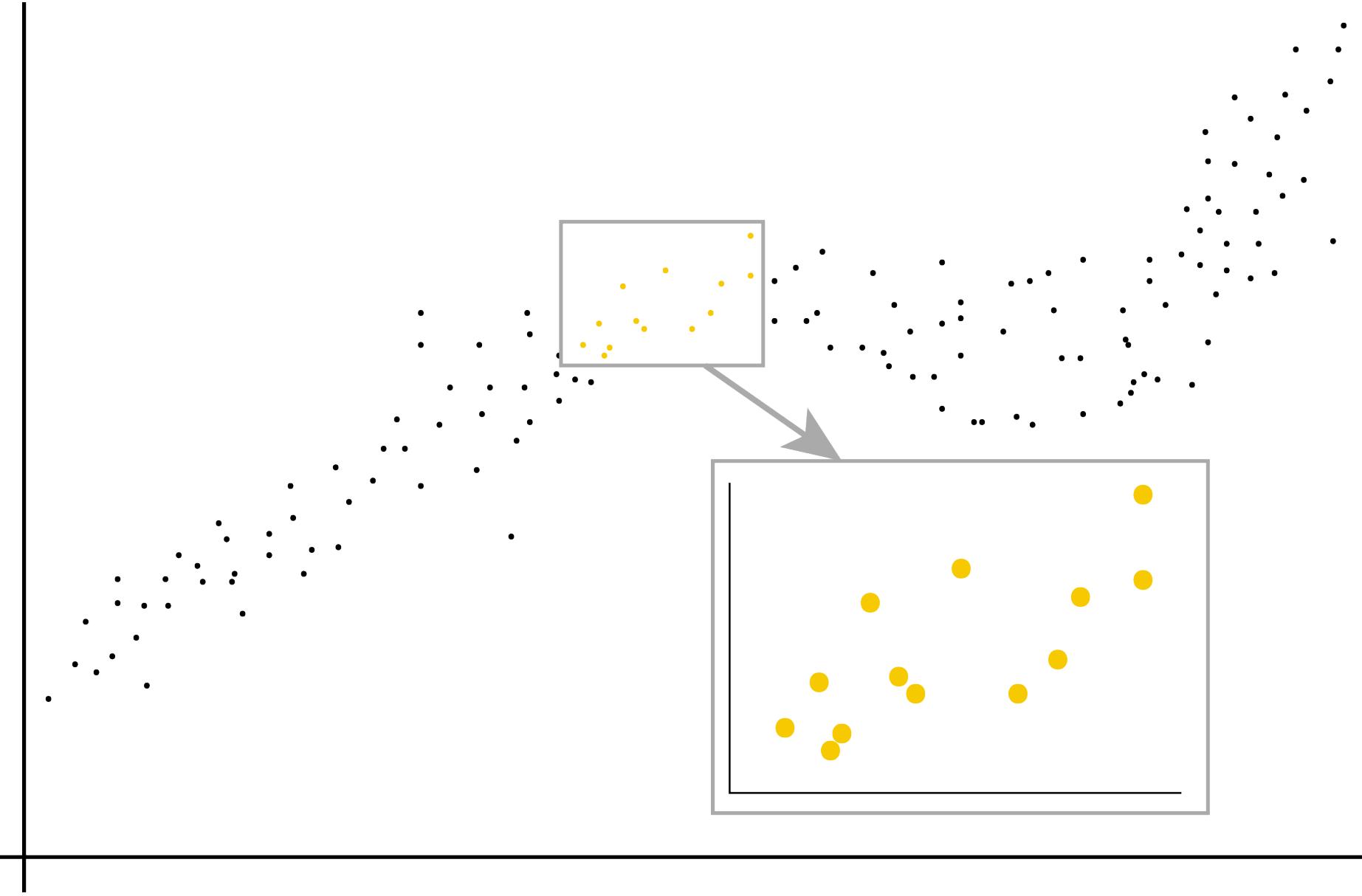
STRATEGIES FOR STATIC VISUALIZATIONS

Small multiples

Callouts (selection/highlighting)

Filtering and faceting

Multiform encoding



STRATEGIES FOR STATIC VISUALIZATIONS

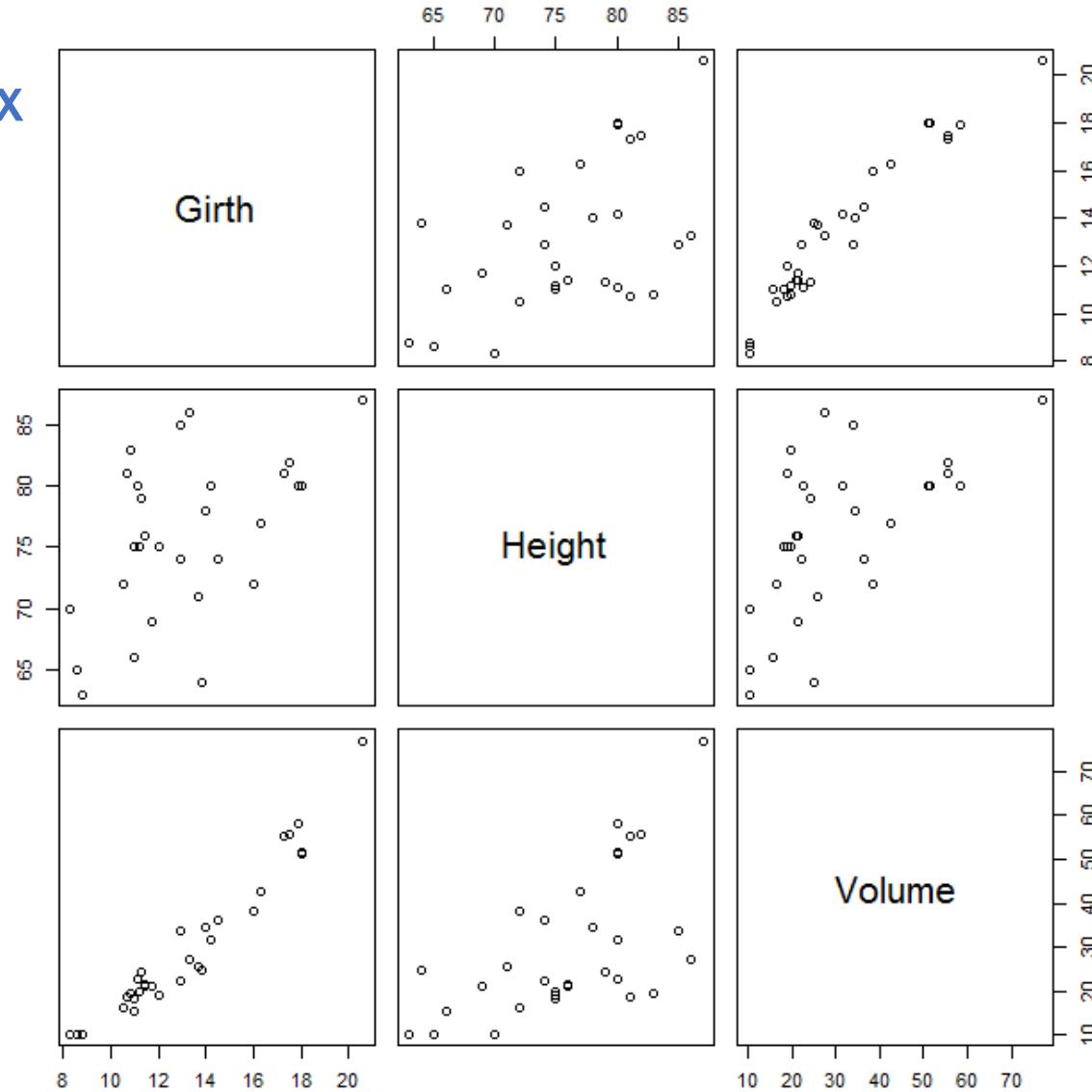
Small multiples

Callouts (selection/highlighting)

Filtering and faceting

Multiform encoding

Scatter plot matrix



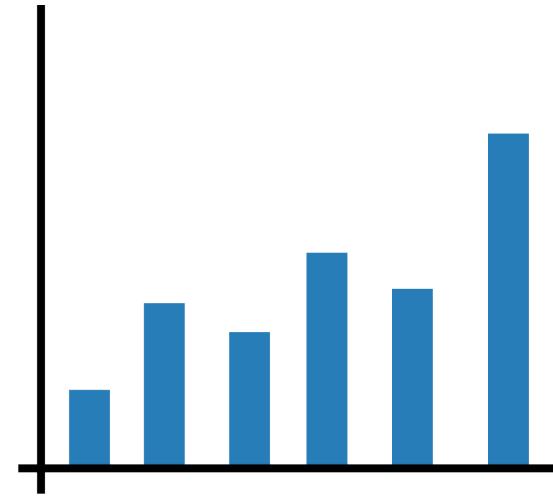
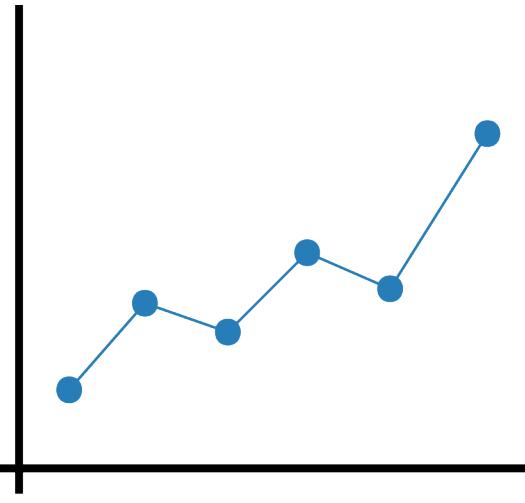
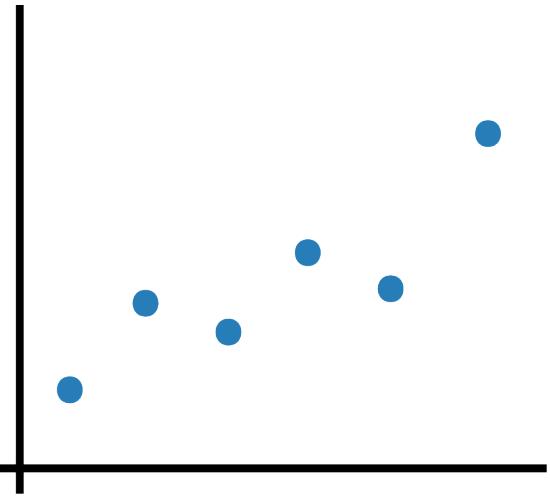
STRATEGIES FOR STATIC VISUALIZATIONS

Small multiples

Callouts (selection/highlighting)

Filtering and faceting

Multiform encoding



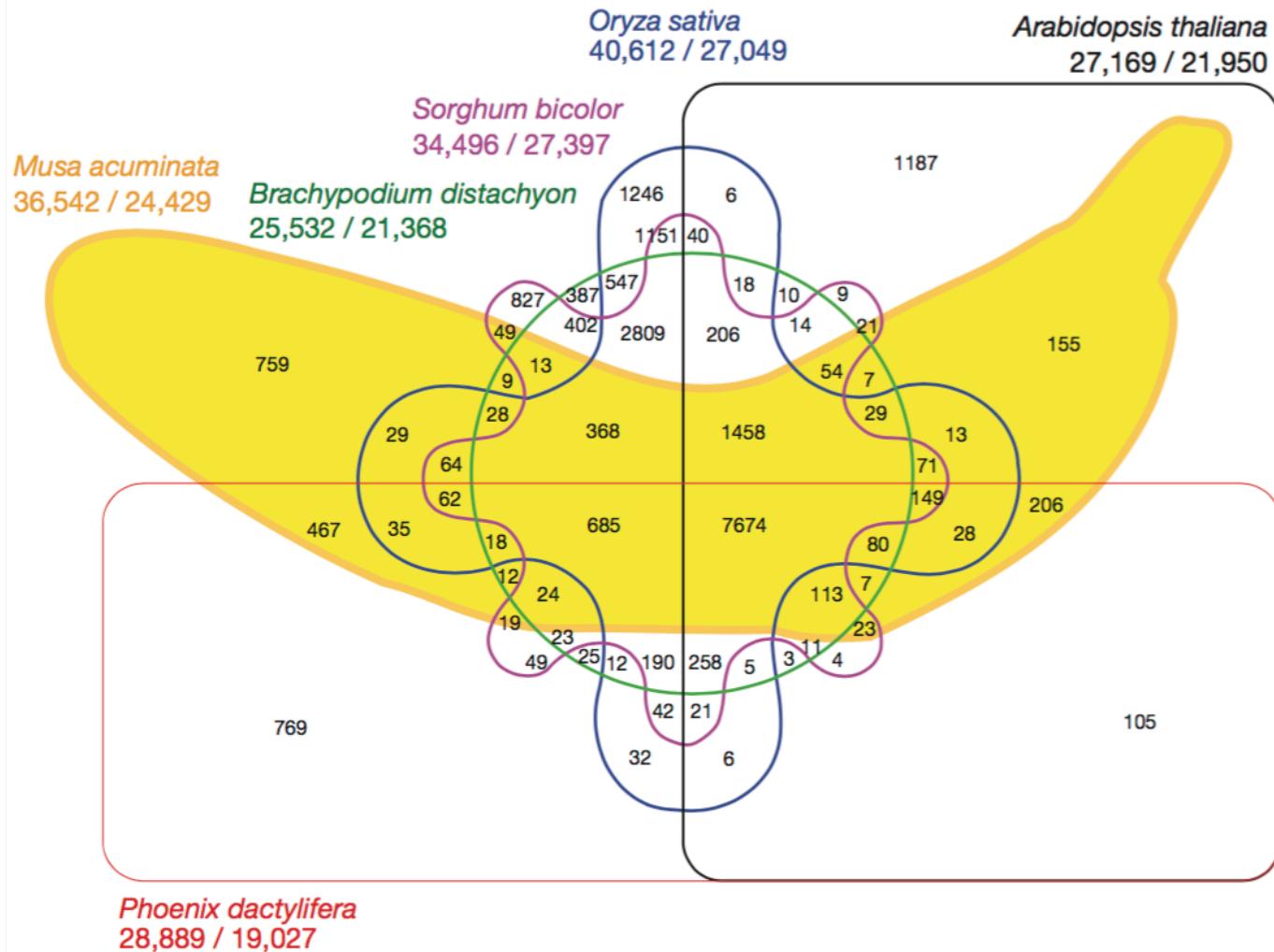
UpSetR

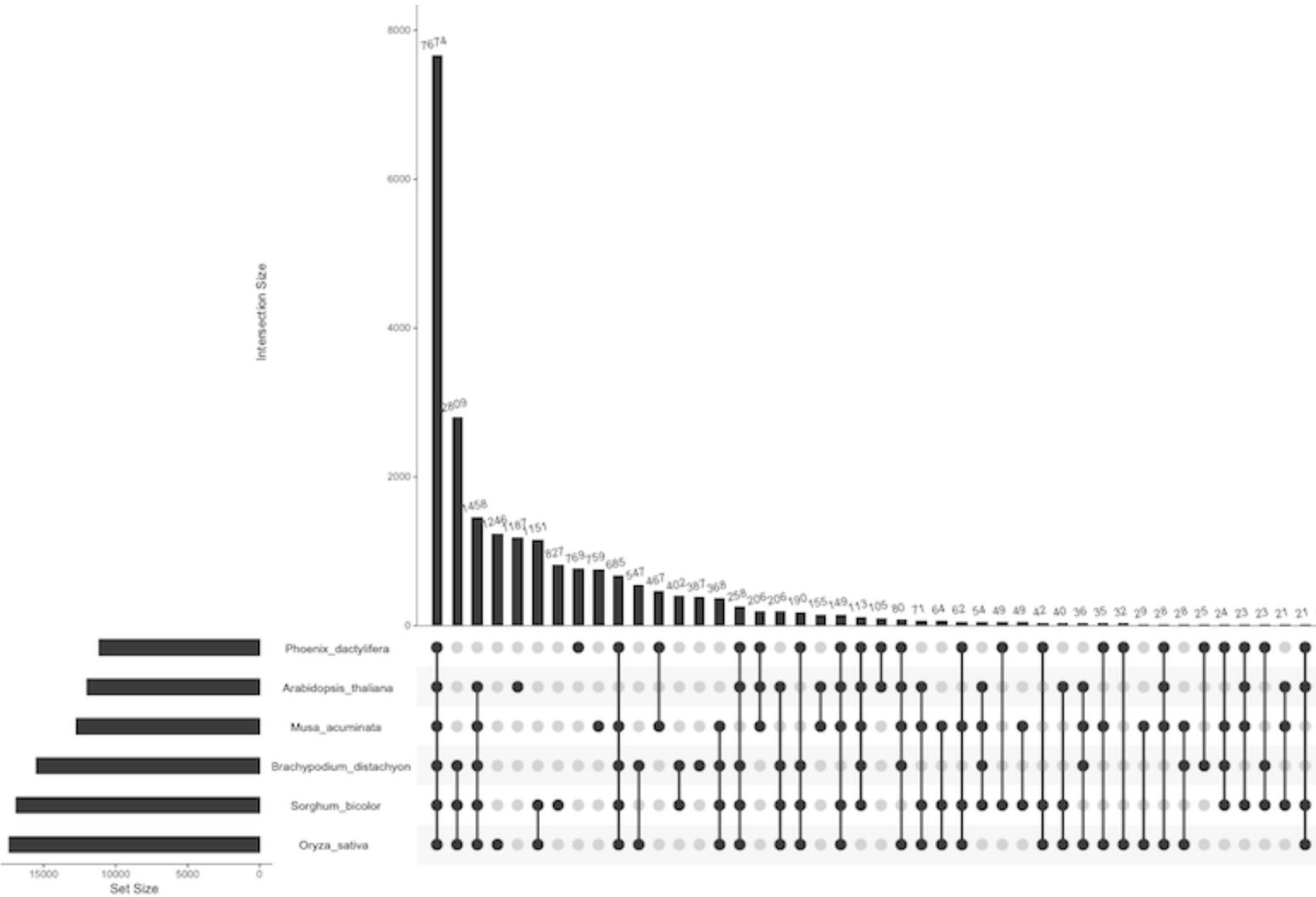
R package for quantitative analysis of sets,
intersections of sets, and aggregates of
intersections

<http://caleydo.org/tools/upset/>

<https://gehlenborglab.shinyapps.io/upsetr>



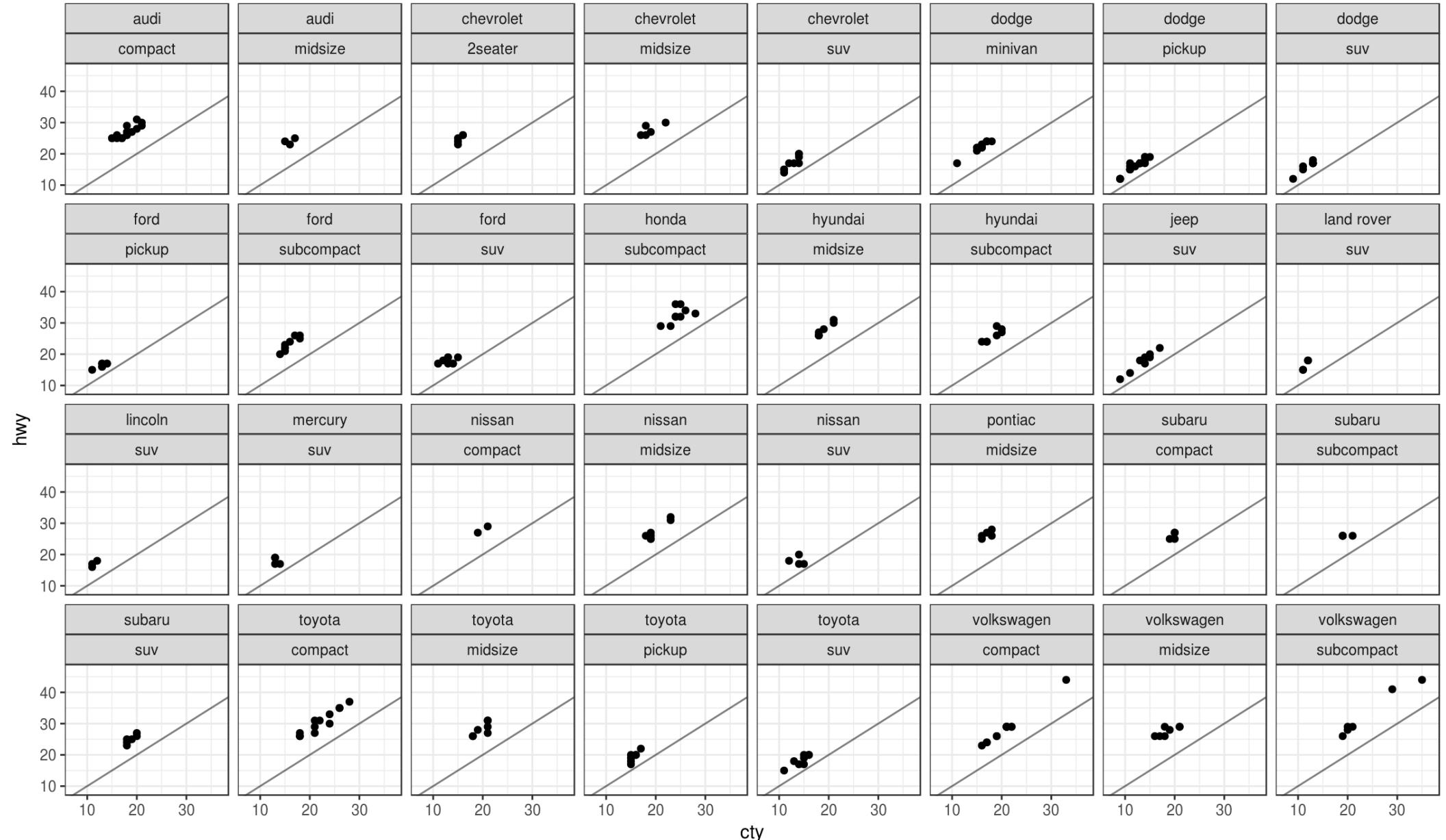


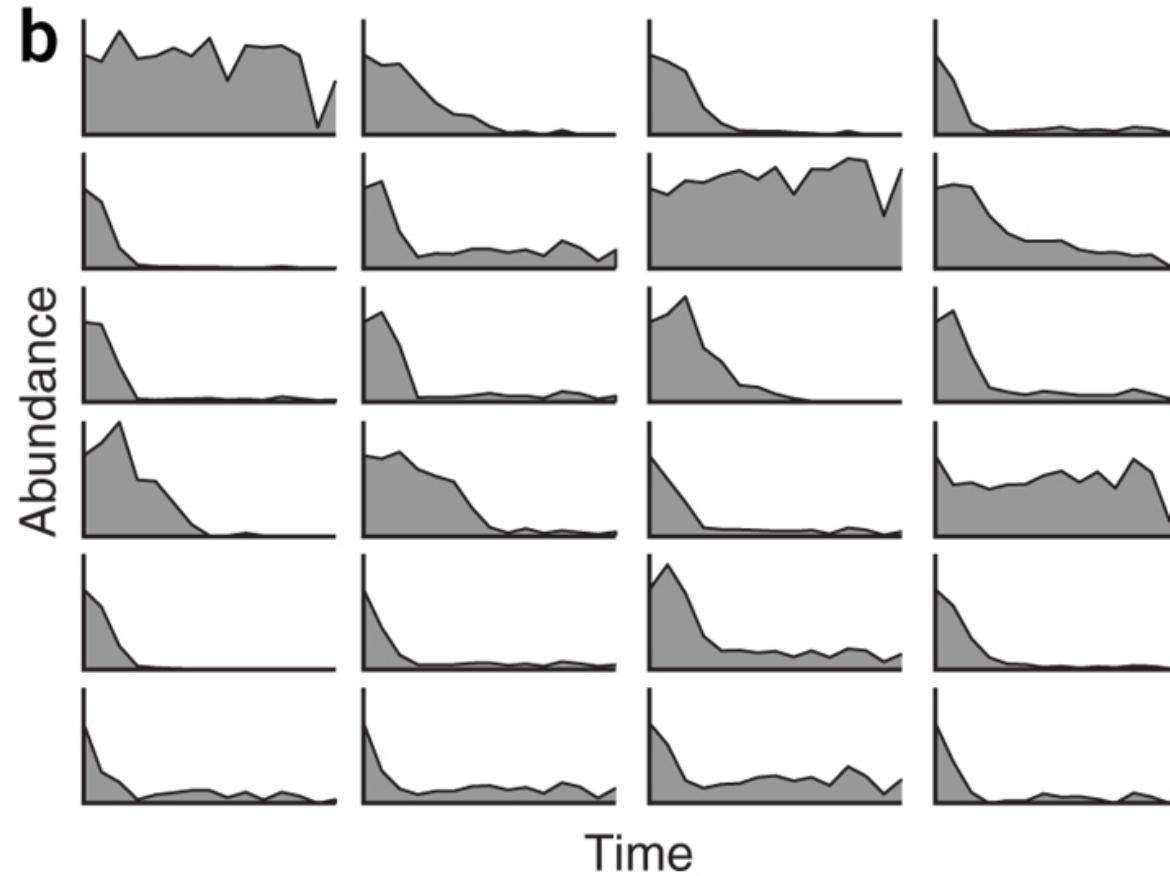


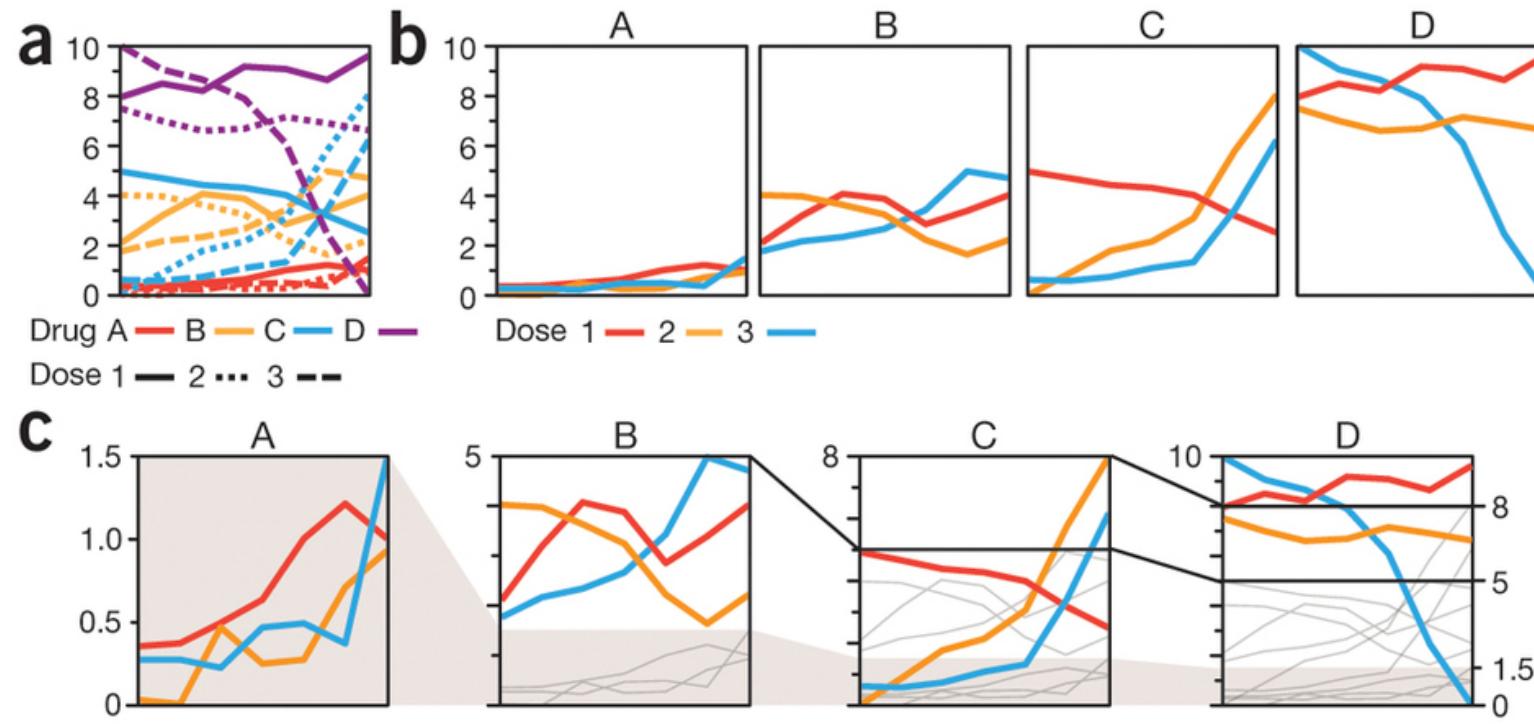
trelliscopejs

An R package for creating interactive
trellis displays of data

<https://hafen.github.io/trelliscopejs/>





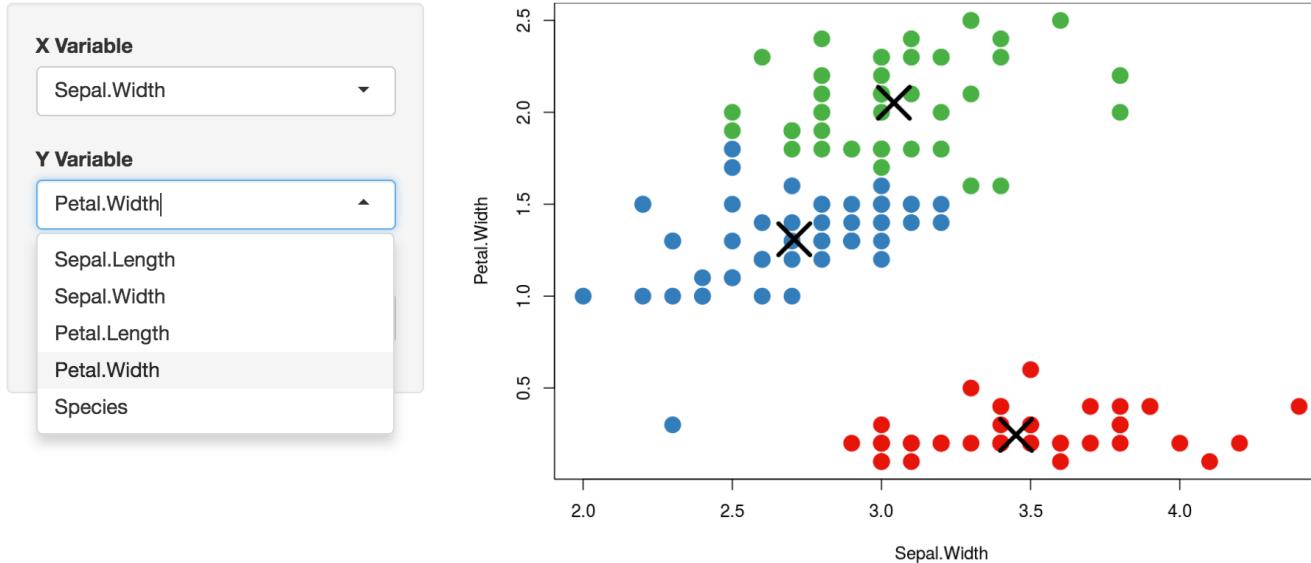


Shiny

An R package for creating interactive
web-based applications

<https://shiny.rstudio.com/>

Iris k-means clustering

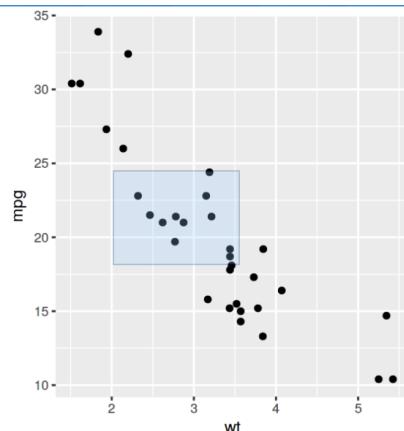


server.R

ui.R

↓ show below

```
function(input, output, session) {  
  
  # Combine the selected variables into a new data frame  
  selectedData <- reactive({  
    iris[, c(input$xcol, input$ycol)]  
  })  
  
  clusters <- reactive({  
    kmeans(selectedData(), input$clusters)  
  })  
  
  output$plot1 <- renderPlot({  
    palette(c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3",  
            "#FF7F00", "#FFFF33", "#A65628", "#F781BF",  
            "#999999"))  
  
    par(mar = c(5.1, 4.1, 0, 1))  
    plot(selectedData(),  
         col = clusters()$cluster,  
         pch = 20, cex = 3)  
    points(clusters()$centers, pch = 4, cex = 4, lwd = 4)  
  })  
}
```



Points near click

```
[1] mpg cyl disp hp wt am gear dist_
<0 rows> (or 0-length row.names)
```

Brushed points

	mpg	cyl	disp	hp	wt	am	gear
Mazda RX4	21.0	6	160.0	110	2.620	1	4
Mazda RX4 Wag	21.0	6	160.0	110	2.875	1	4
Datsun 710	22.8	4	108.0	93	2.320	1	4
Hornet 4 Drive	21.4	6	258.0	110	3.215	0	3
Hornet Sportabout	18.7	8	360.0	175	3.440	0	3
Merc 240D	24.4	4	146.7	62	3.190	0	4
Merc 230	22.8	4	140.8	95	3.150	0	4
Merc 280	19.2	6	167.6	123	3.440	0	4
Toyota Corona	21.5	4	120.1	97	2.465	0	3
Ferrari Dino	19.7	6	145.0	175	2.770	1	5
Volvo 142E	21.4	4	121.0	109	2.780	1	4

app.R

↓ show below

```
library(ggplot2)
library(Cairo) # For nicer ggplot2 output when deployed on Linux

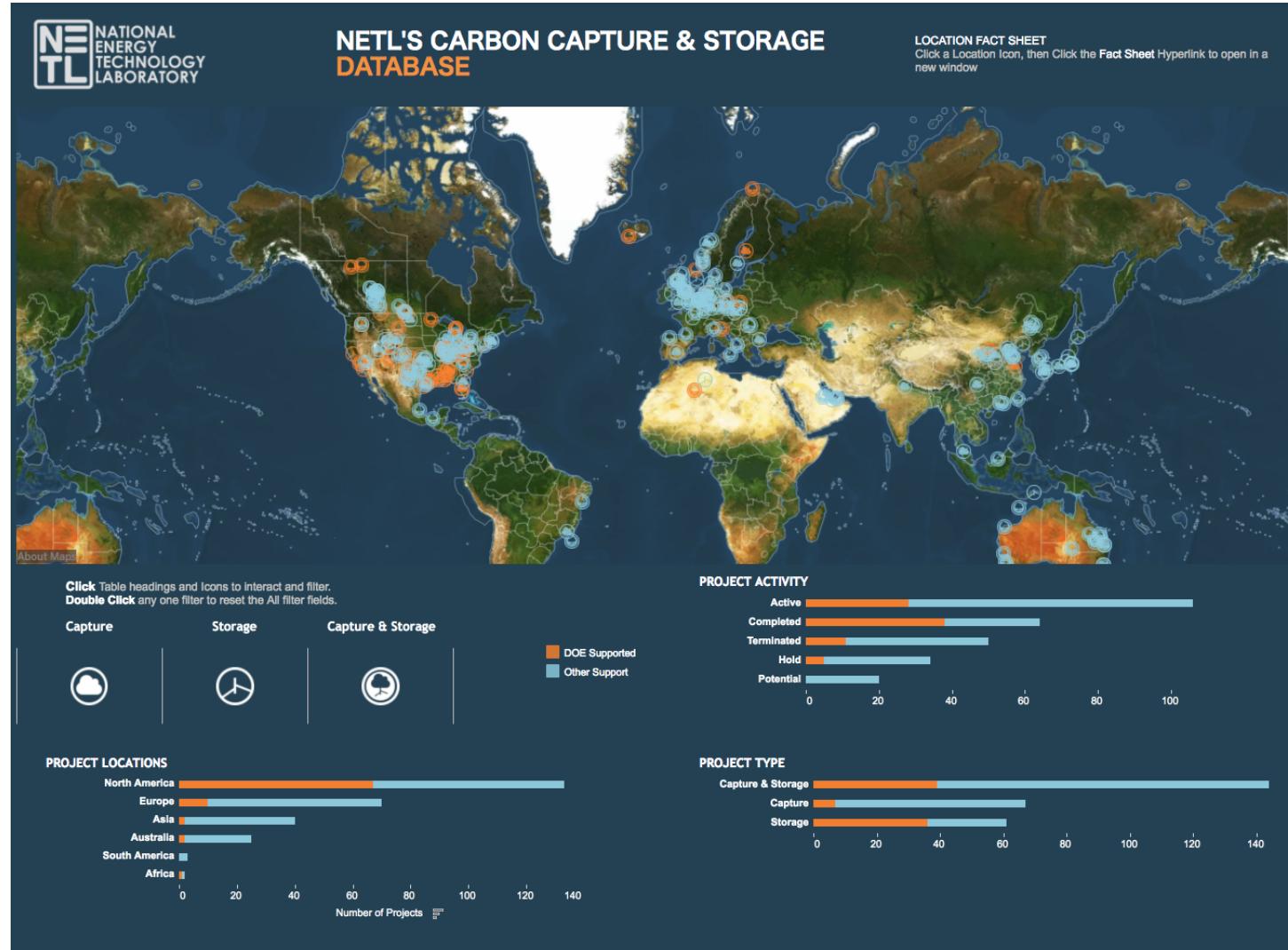
# We'll use a subset of the mtcars data set, with fewer columns
# so that it prints nicely
mtcars2 <- mtcars[, c("mpg", "cyl", "disp", "hp", "wt", "am", "gear")]

ui <- fluidPage(
  fluidRow(
    column(width = 4,
      plotOutput("plot1", height = 300,
        # Equivalent to: click = clickOpts(id = "plot_click")
        click = "plot1_click",
        brush = brushOpts(
          id = "plot1_brush"
        )
      )
    ),
    fluidRow(
      column(width = 6,
        h4("Points near click"),
        verbatimTextOutput("click_info")
      )
    )
  )
)
```

Tableau

Application for creating linked
visualizations and dashboards

<https://www.tableau.com/>



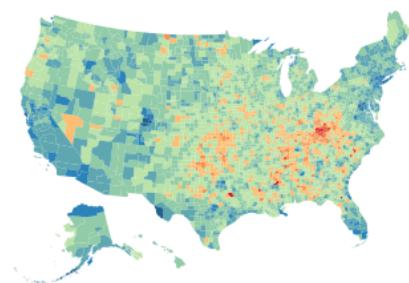
<https://public.tableau.com/en-us/s/gallery/carbon-capture-and-storage>

Percent Change in deaths per 100,000 from 1980 to 2014, by county

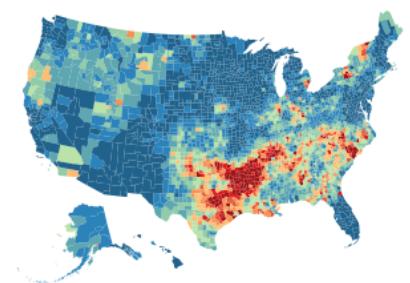


Hover over map for details by county

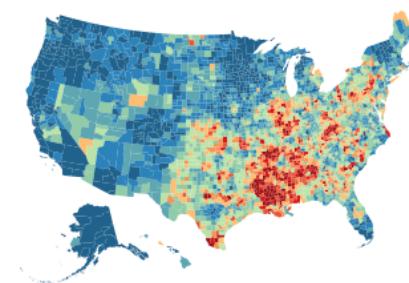
Cancer



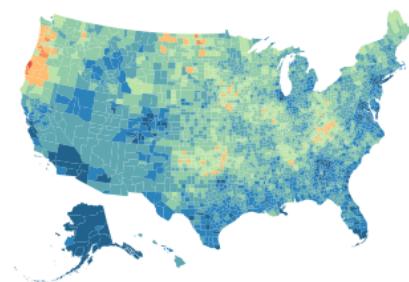
Nutritional deficiencies



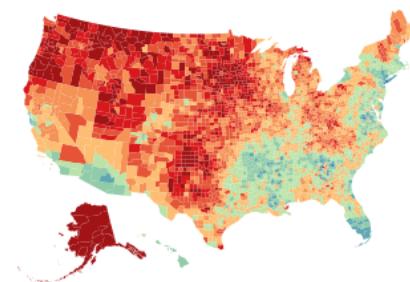
Common infectious diseases



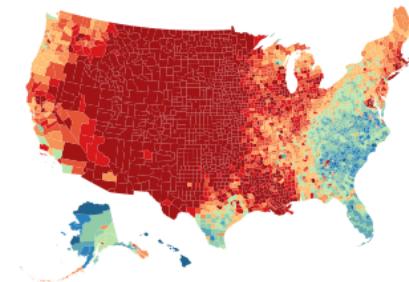
Digestive diseases



Musculoskeletal disorders



Neglected tropical diseases & malaria



It appears that the regions with a high growth in Nutritional Deficiencies related deaths, are predominately the areas with a decline in fatal Musculoskeletal Disorders. Lack of physical activity could be a leading factor to explain this phenomenon.

-50% +50%

<https://public.tableau.com/en-us/s/gallery/changing-diseases>

Data Exploration

In pairs, identify all the independent dimensions in the data set(s) provided.

Determine a narrative arc in the data to highlight, and using R markdown and ggplot2, create a series of visualizations with annotations that depict that narrative