Visualization in HCI

05-499/05-899 Section C

Carnegie Mellon University

Human-Computer Interaction Institute

Views + Filter & Aggregate

March 27, 2017

Final Project

Next Wednesday, April 5

Informal 10-minute group meeting in class.

Be prepared to describe your finalized dataset and show sketches or prototypes of (some) of your views.

Sign up for a timeslot on Slack!

Wednesday, April 19

“Project Milestone” Due

Submit draft of the current state of your process book via github

Refer to details at:

https://cmu-vis-course.github.io/2017/project/

PARTITIONING

action on the dataset that **separates the data into groups**

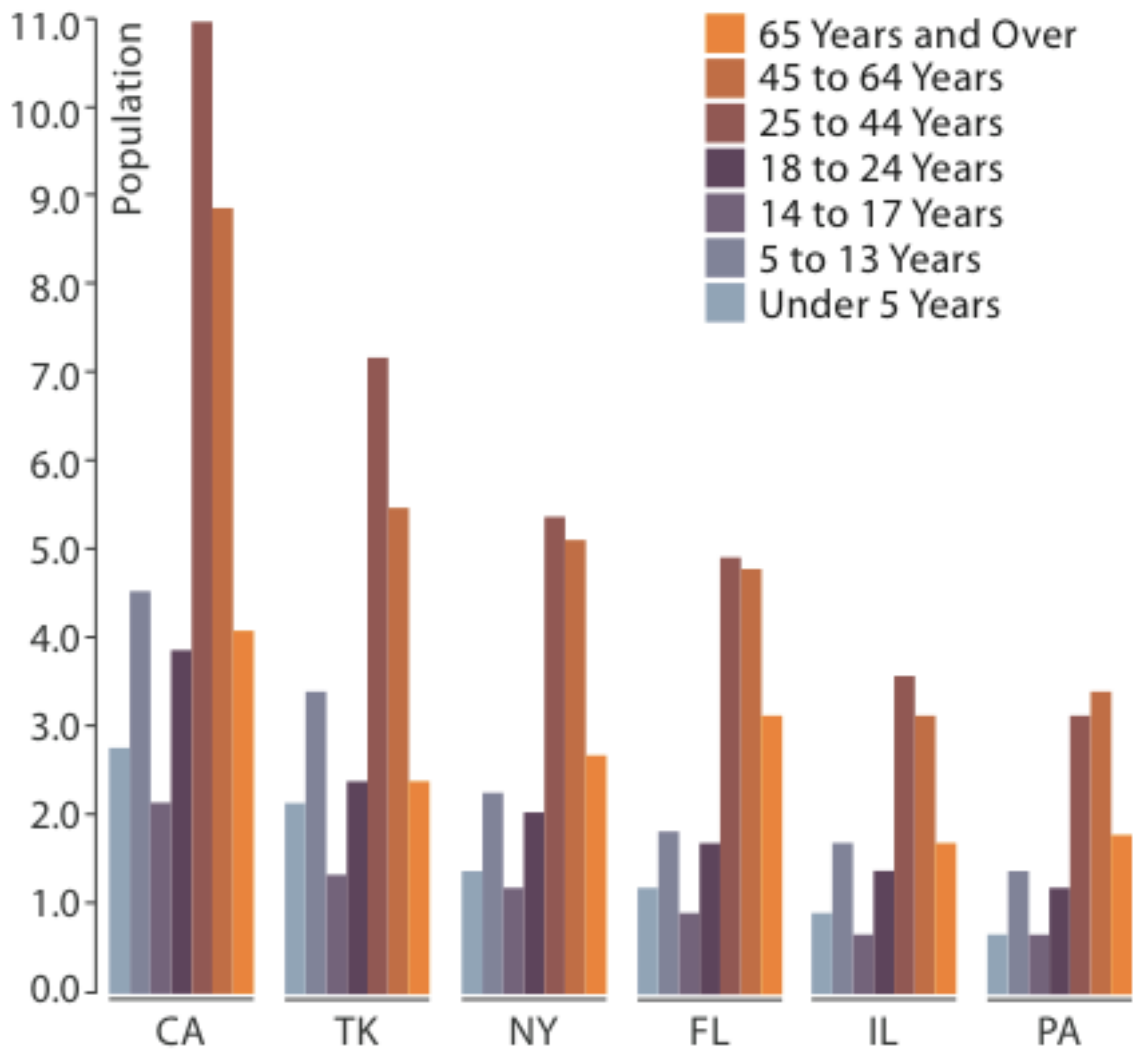
design choices

- how to divide data up between views, given a hierarchy of attributes
- how many splits, and order of splits
- how many views (usually data driven)

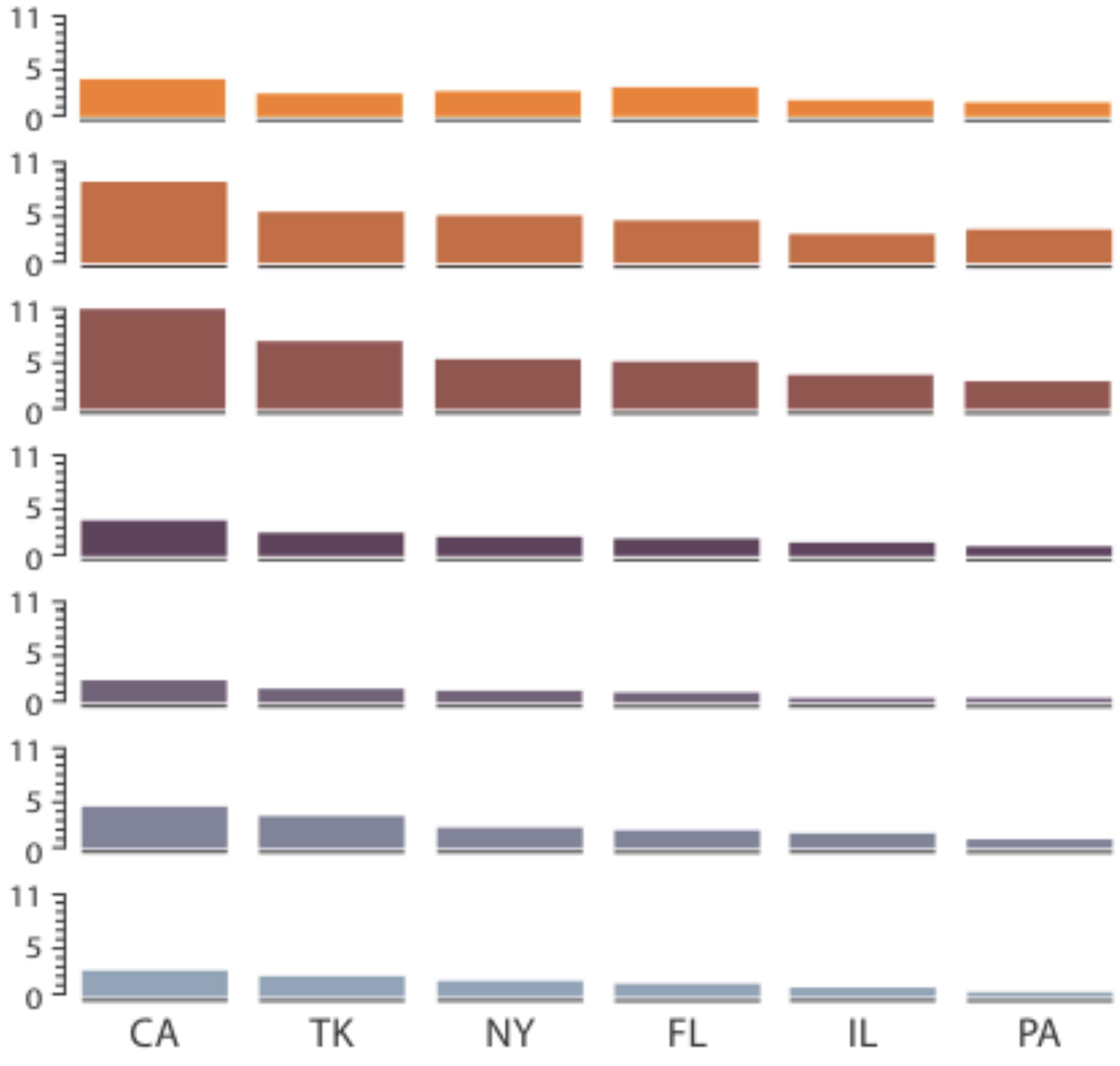
partition attribute(s)

- typically categorical

Partitioning



Partitioned by State



Partitioned by Age Group and State

Trellis Plots

panel variables

attributes encoded in individual views

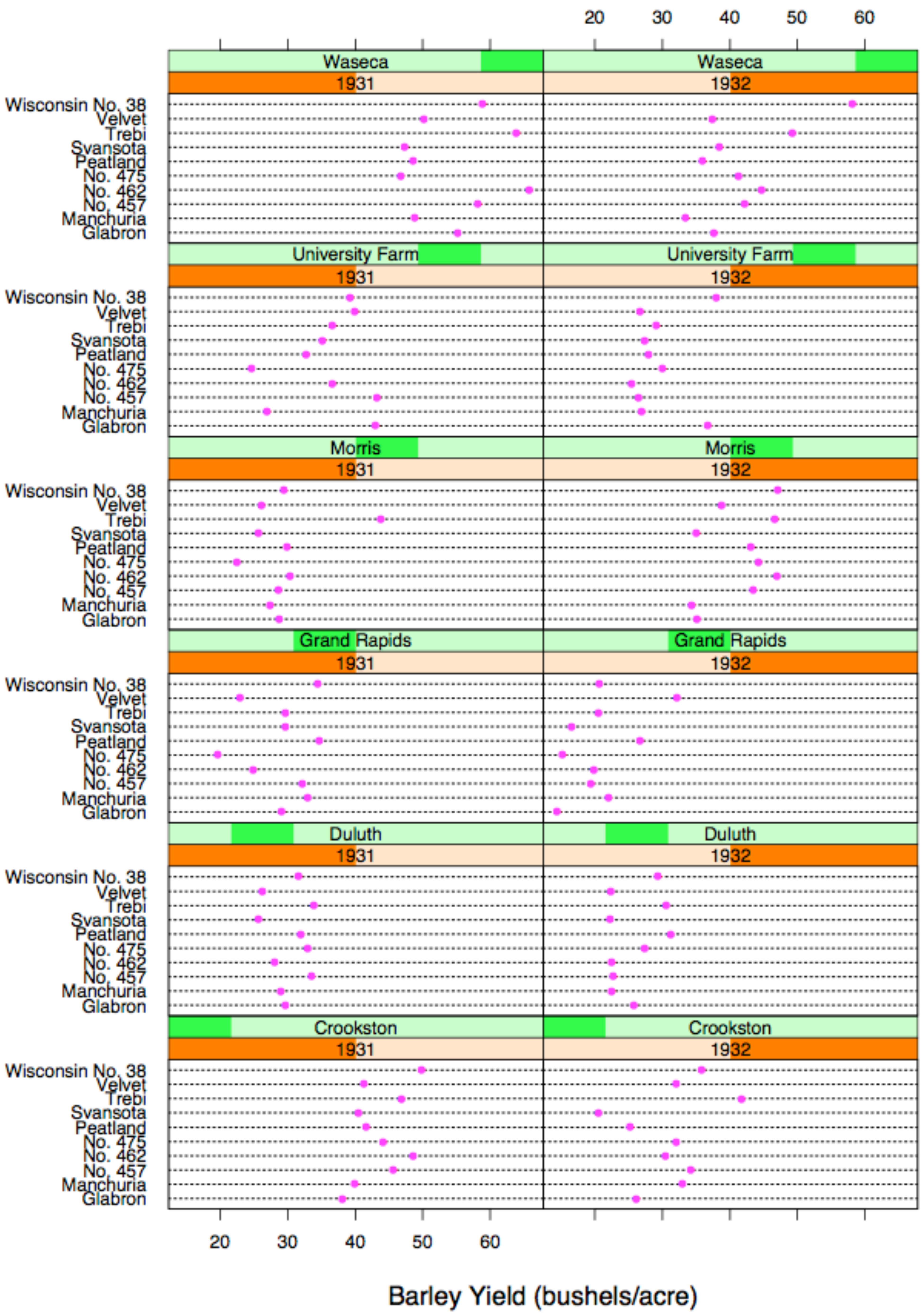
partitioning variables

partitioning attributes assigned to columns, rows, and pages

main-effects ordering

order partitioning variable levels/states based on derived data

support perception of trends and structure in data

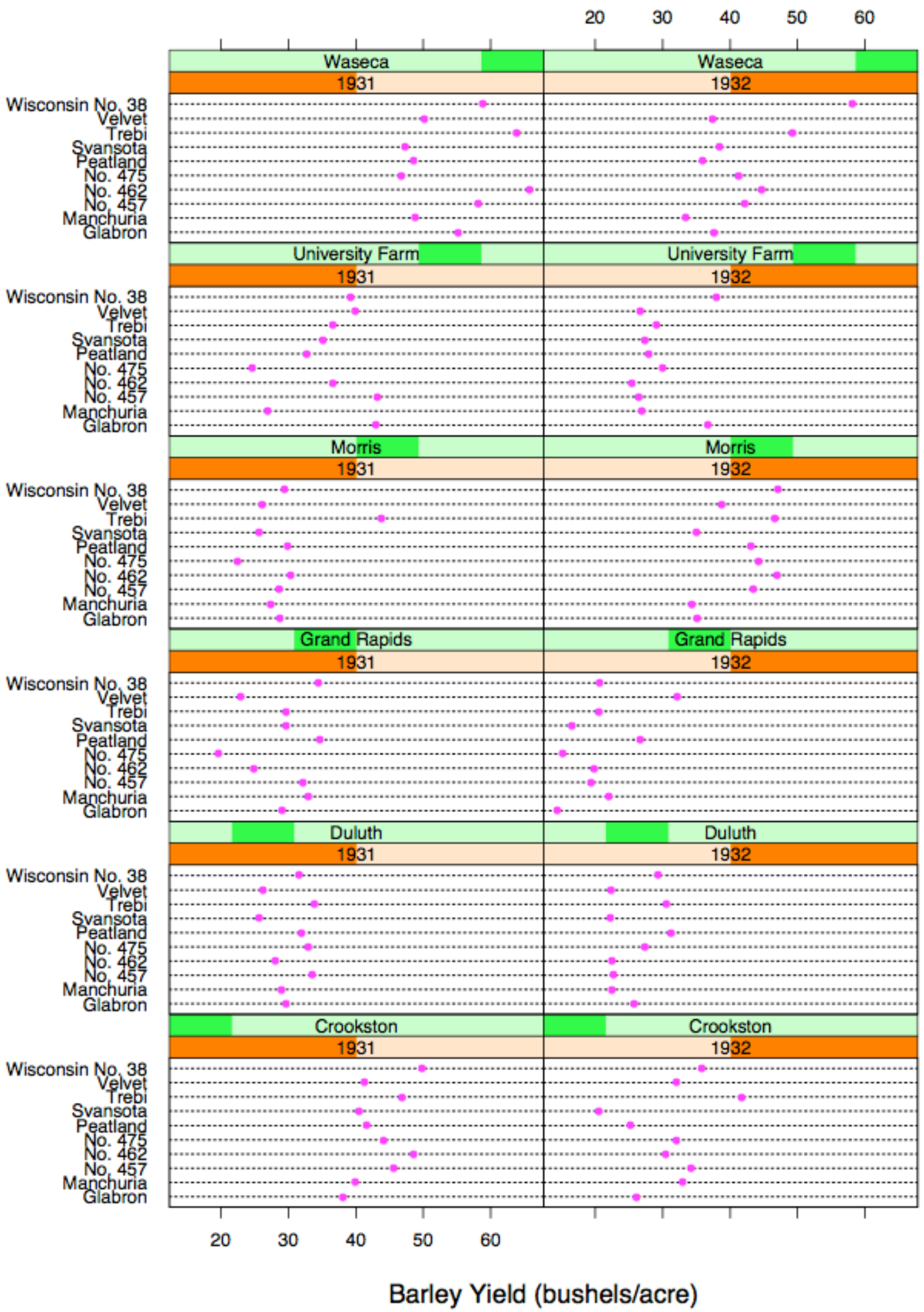


Data

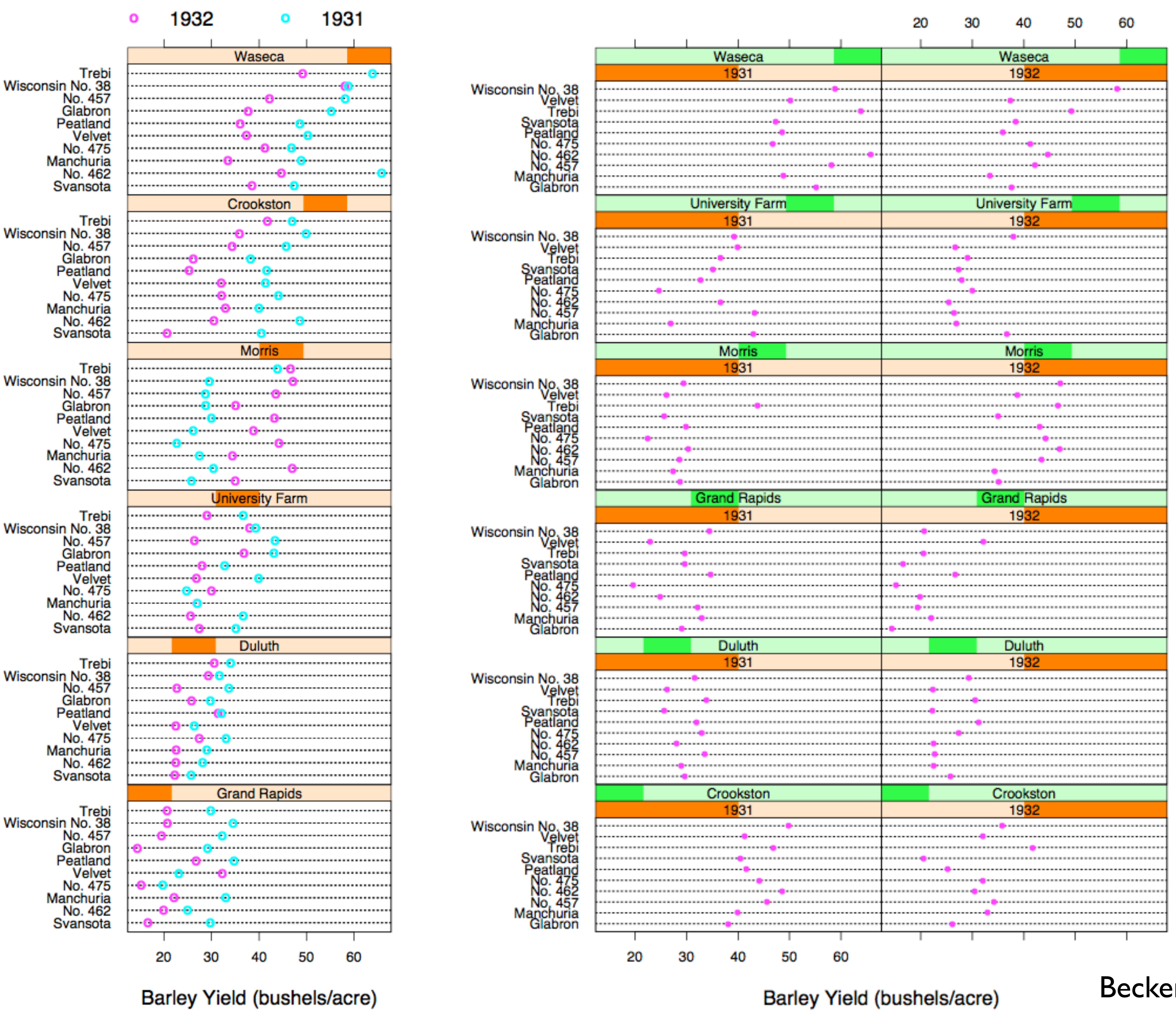
Barley Yields in two years across multiple farms for multiples barley strains

partitioning variables

- Columns partitioned by year
- Rows partitioned by farm



Becker 1996



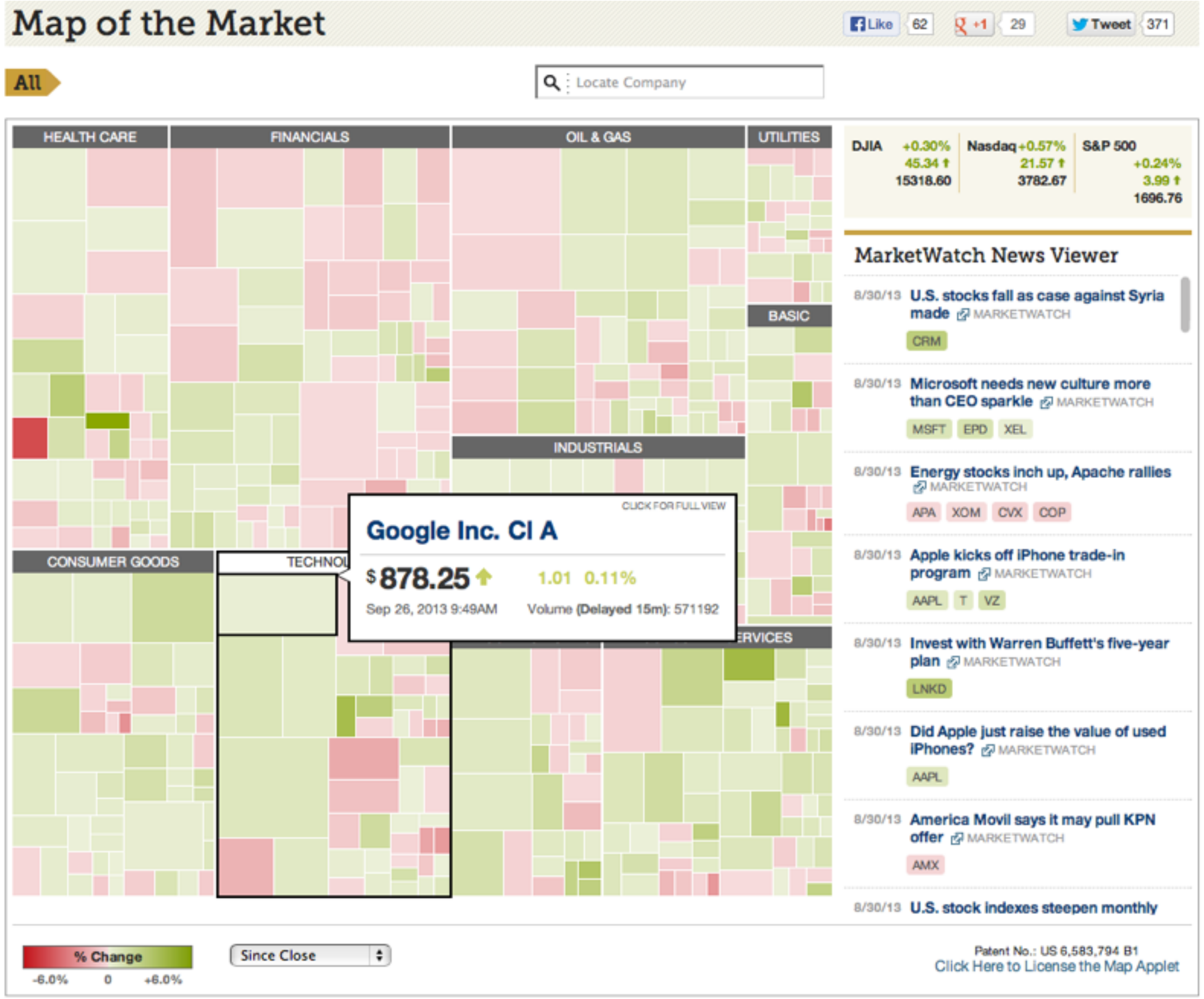
1932
1931
Waseca
Crookston
Morris
University Farm
Duluth
Grand Rapids
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Svansota
Peatland
No. 462
20
30
40
50
60
Barley Yield (bushels/acre)
Barley Yield (bushels/acre)

Becker 1996

Recursive Subdivision

partitioning: flexibly transform data attributes into a hierarchy

use treemaps as spacefilling rectangular layouts



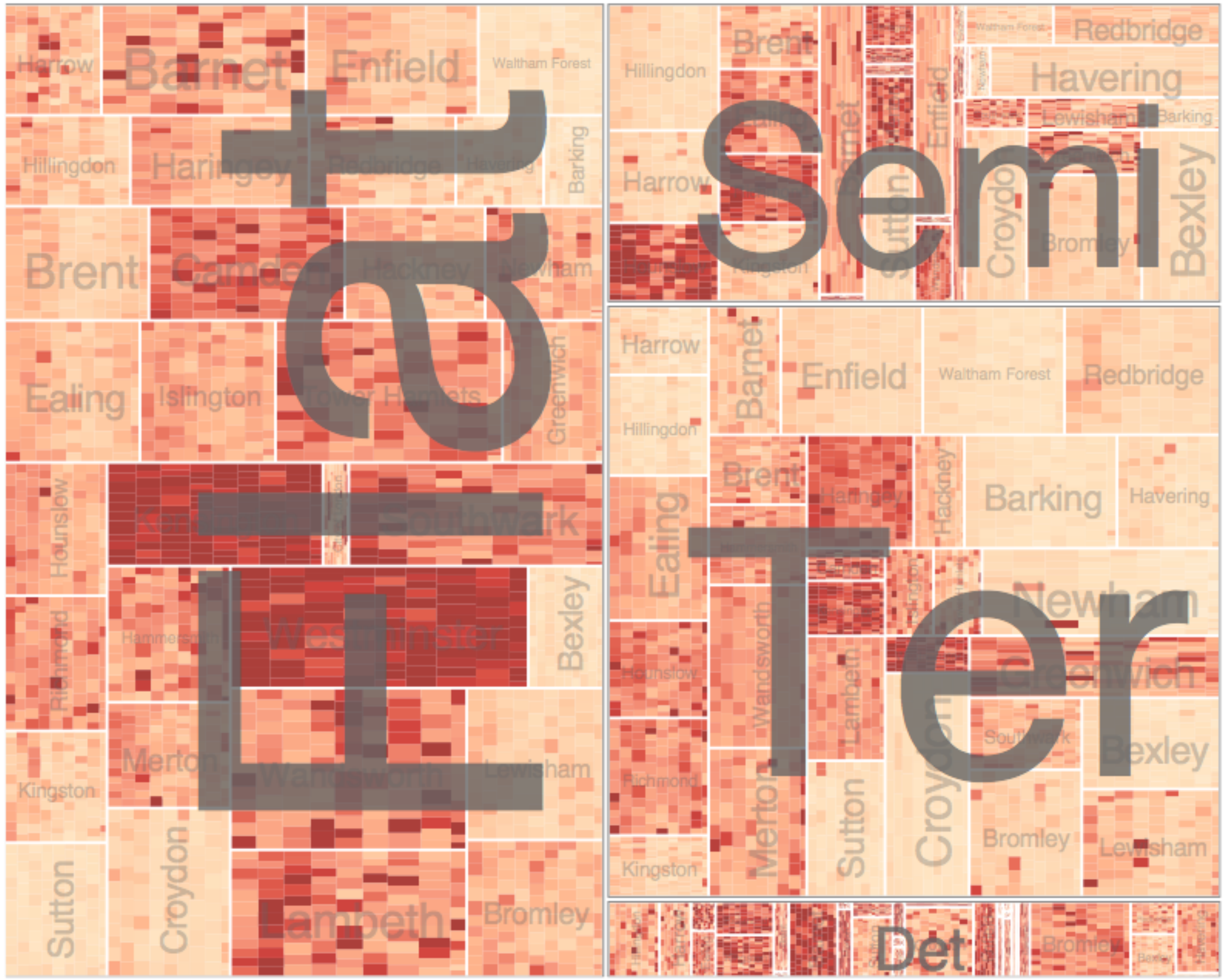
Treemap

HiVE example: London property

partitioning attributes
house type
neighborhood
sale time

encoding attributes
average price (color)
number of sales (size)

results
between neighborhoods, different housing distributions
within neighborhoods, similar prices



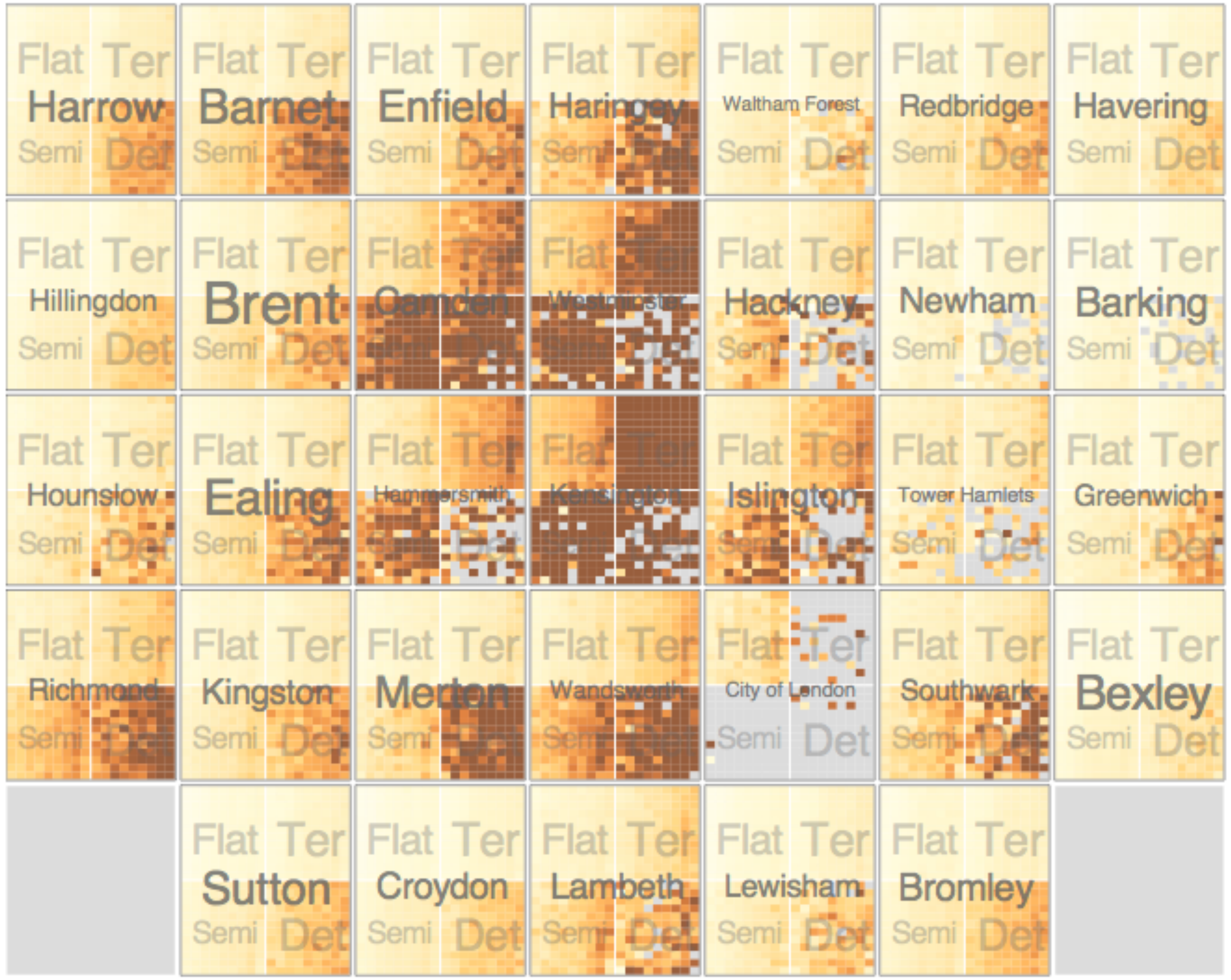
Slingsby 2009

HiVE example: London property

partitioning attributes
neighborhood location
neighborhood
house type
sale time (year)
sale time (month)

encoding attributes
average price (color)
n/a (size)

results
expensive neighborhoods
near center of city



Slingsby 2009

Configuring Hierarchical Layouts to Address Research Questions

CITY UNIVERSITY LONDON

Aidan Slingsby, Jason Dykes and Jo Wood
giCentre, Department of Information Science, City University London
http://www.gicentre.org/hierarchical_layouts/

CITY UNIVERSITY LONDON

LAYERING

combining multiple views on top of one another to form a composite view

rationale

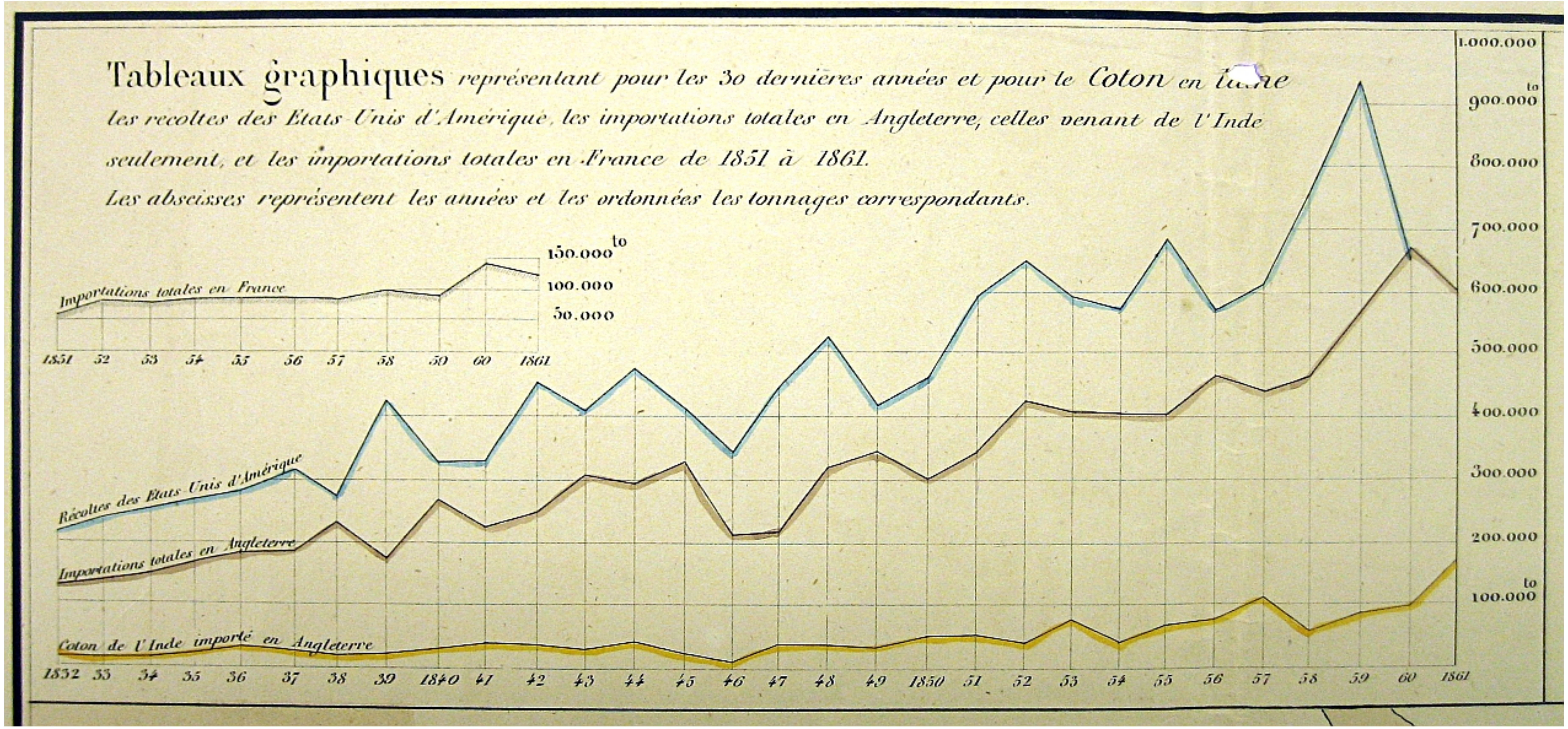
supports a larger, more detailed view than using multiple views

trade-off

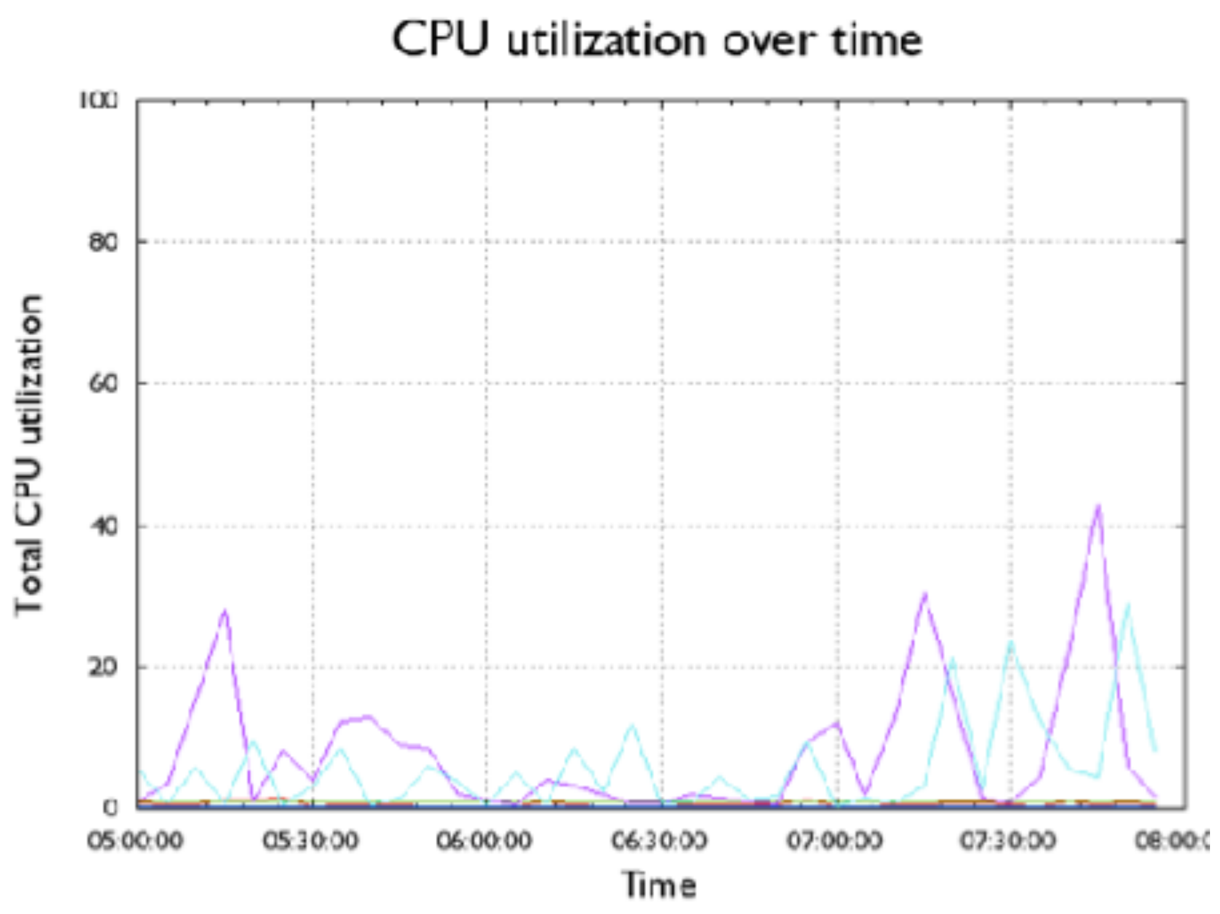
layering imposes constraints on visual encoding choice as well as number of layers that can be shown

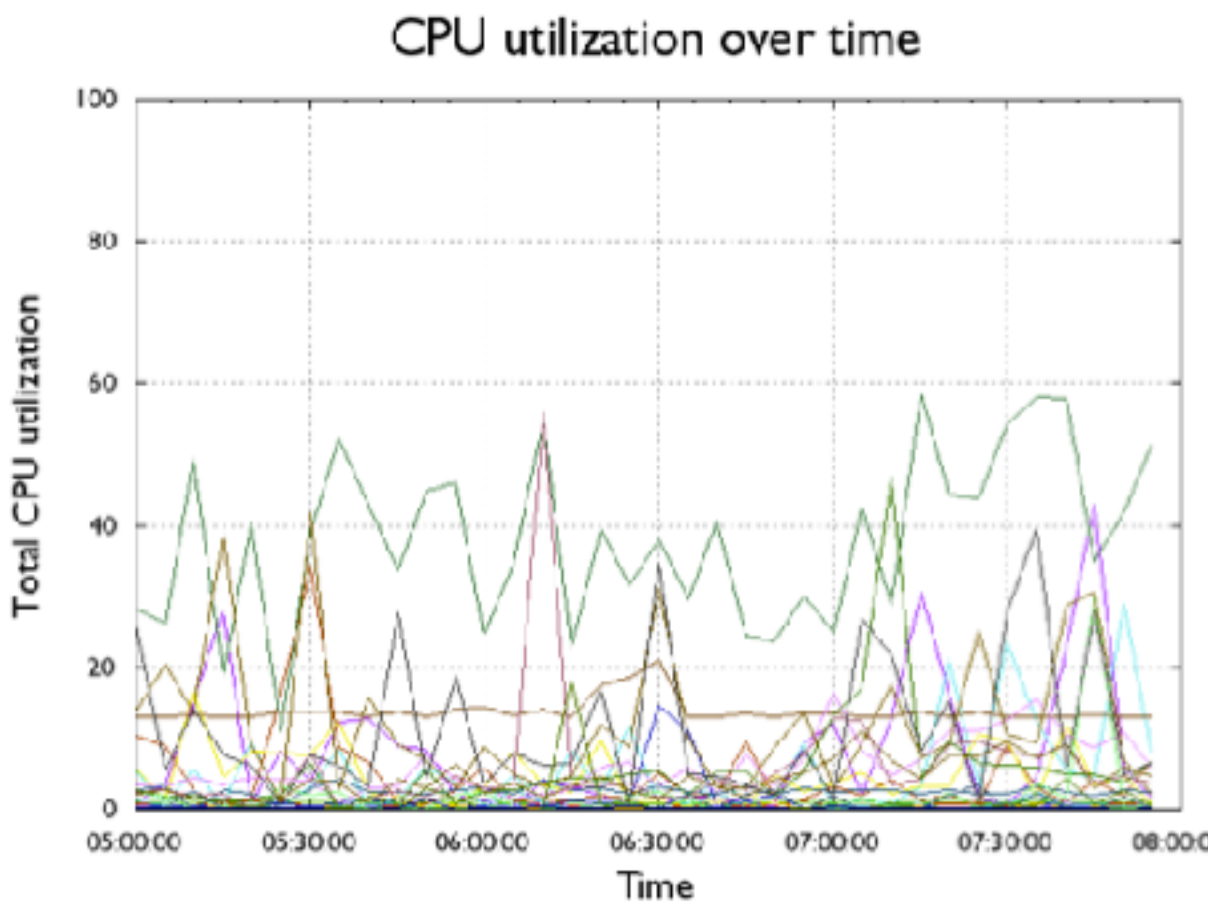
JOSEPH MINARD

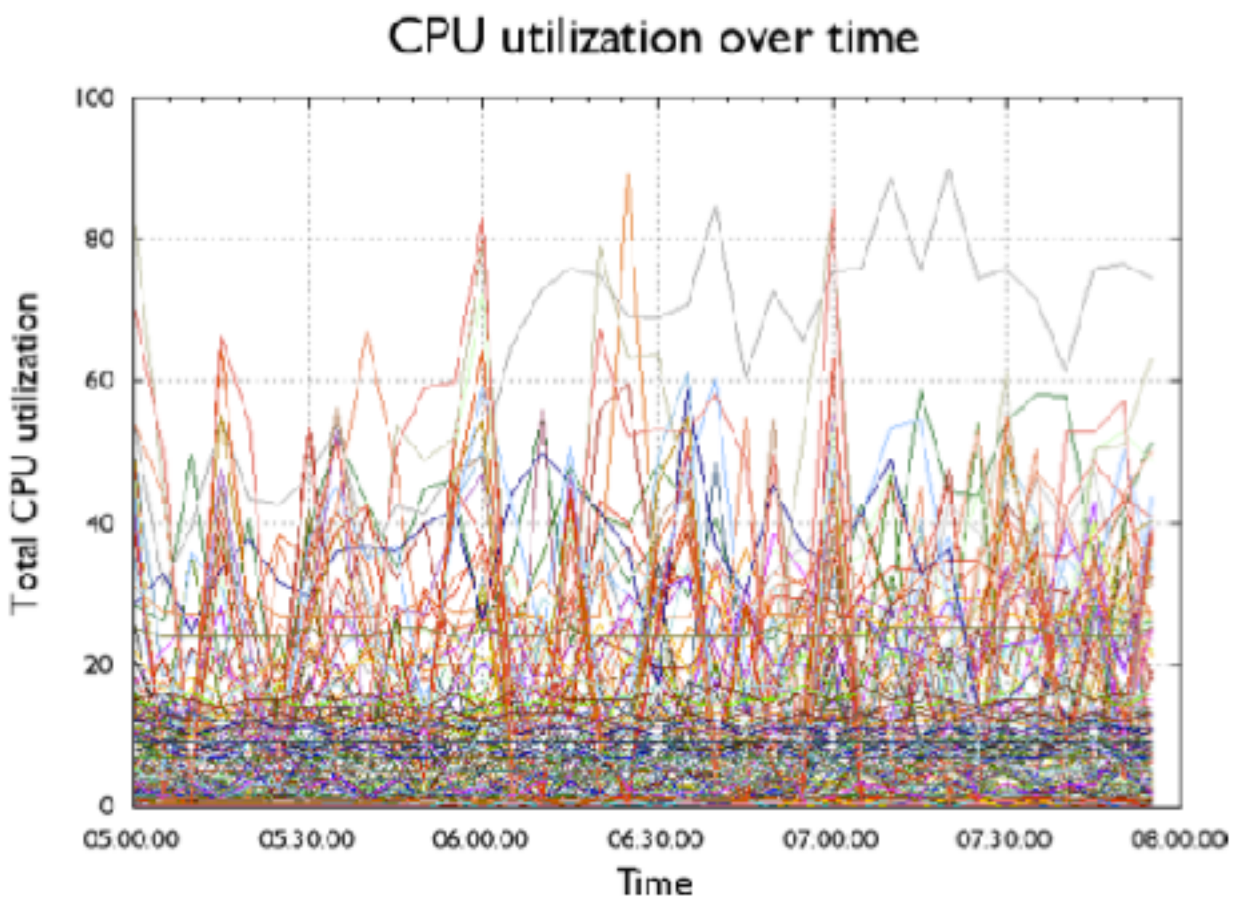
1781-1870



overlays


CPU utilization over time
Total CPU utilization
Time

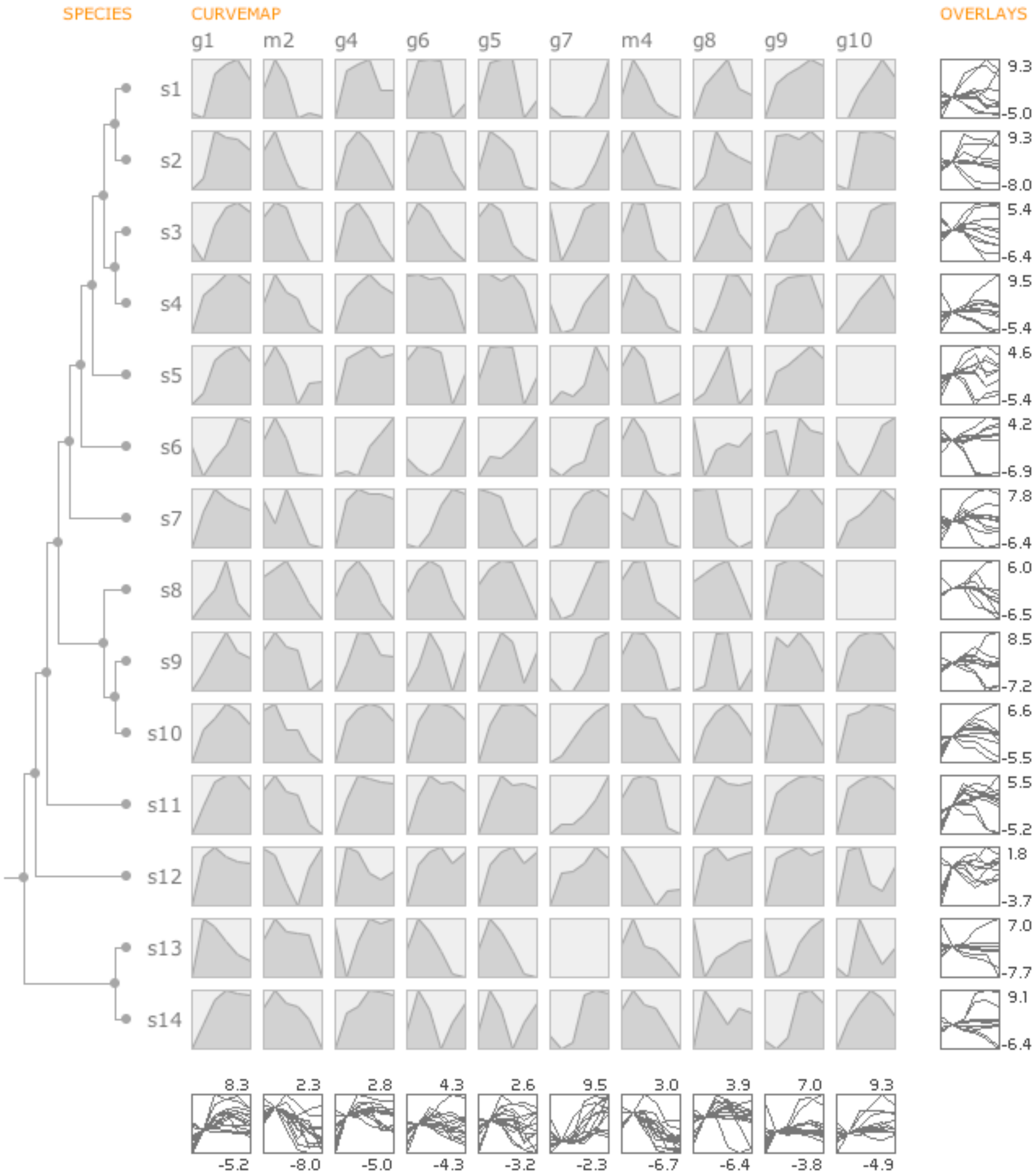

CPU utilization over time
Total CPU utilization
Time


CPU utilization over time
Total CPU utilization
Time

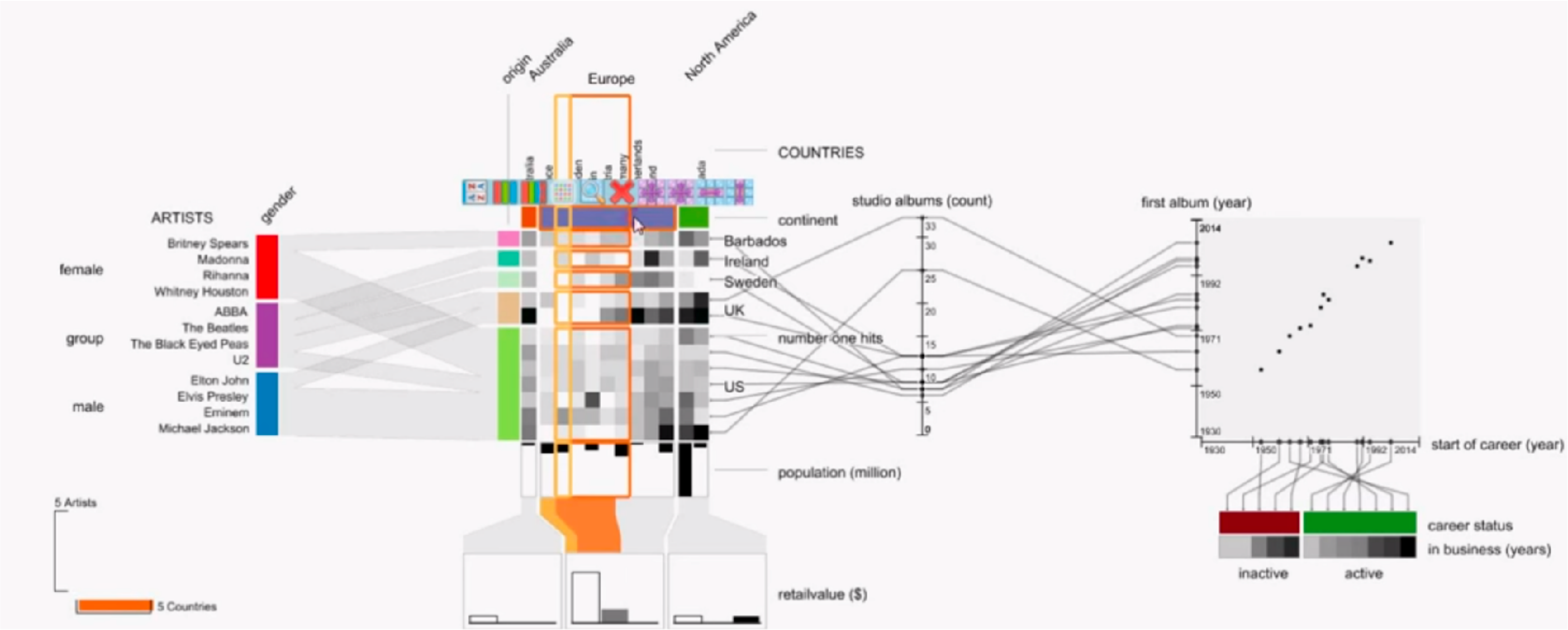
Combined

Partitioned + layered graph

Synchronized through highlighting



MCV to the Max



Filter & Aggregate

Reducing Items and Attributes

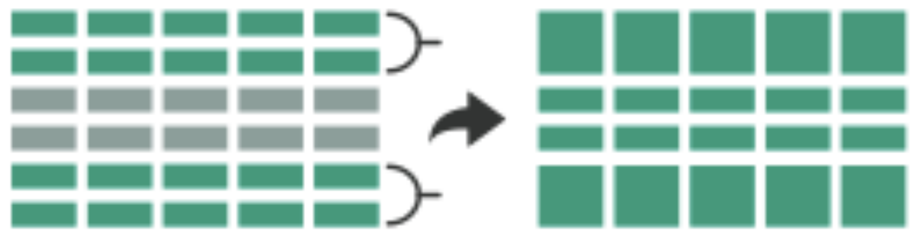
Filter

Items

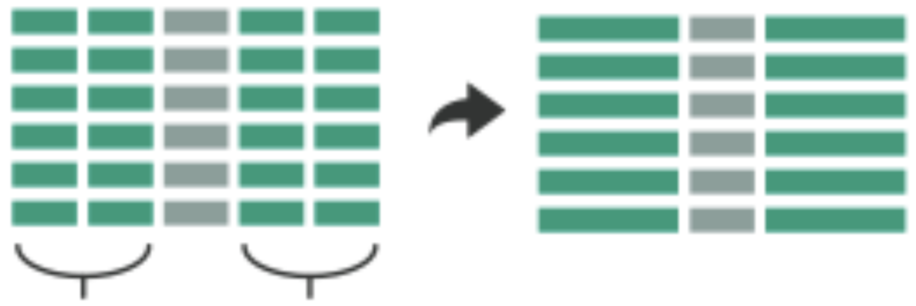
Attributes

Aggregate

Items



Attributes



Filter

Elements are eliminated

What drives filters?

Any possible function that partitions a dataset into two sets

- Bigger/smaller than x
- Noisy/insignificant

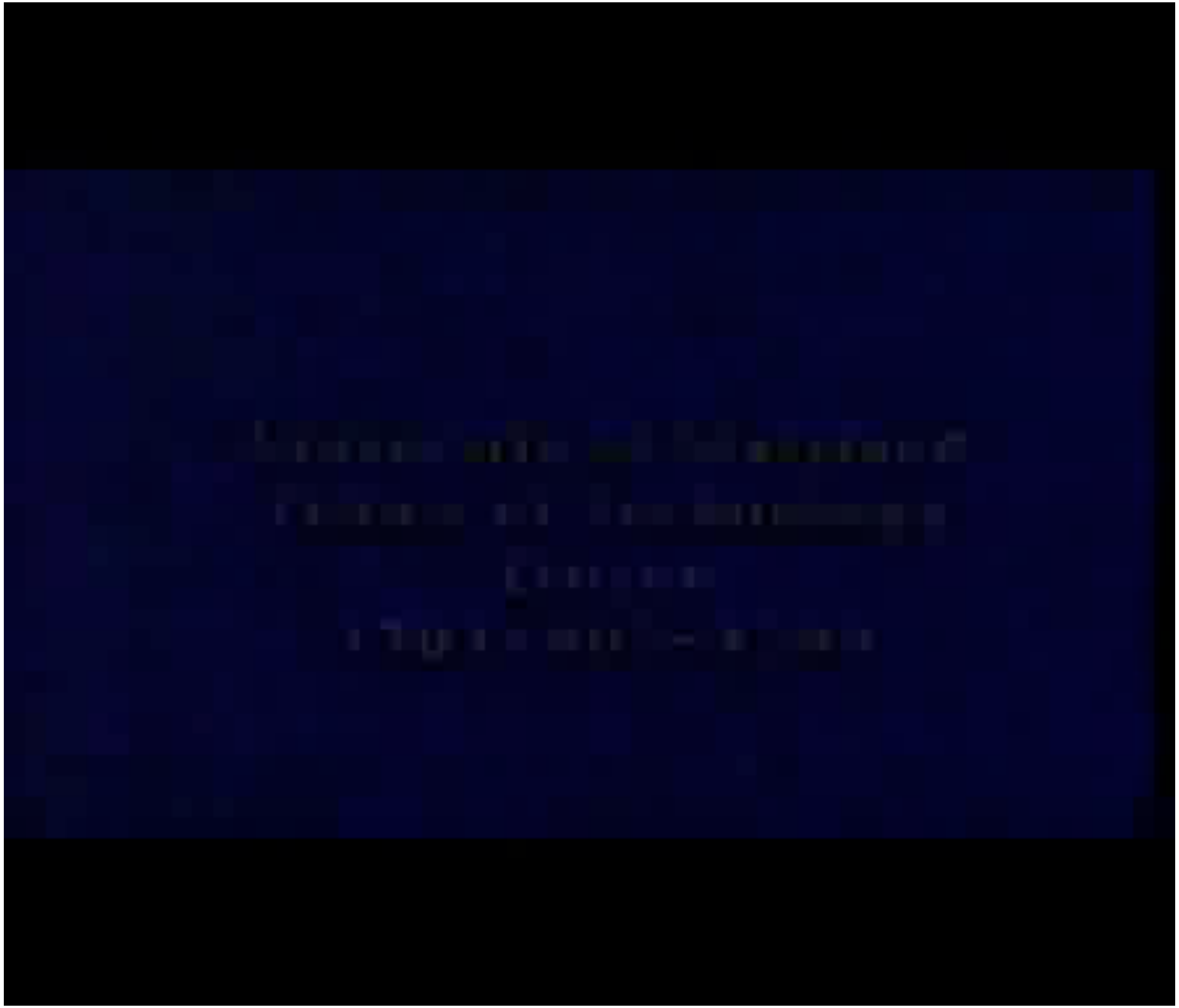
Dynamic Queries / Filters

coupling between encoding and interaction so that user can immediately see the results of an action

Queries: start with 0, add in elements

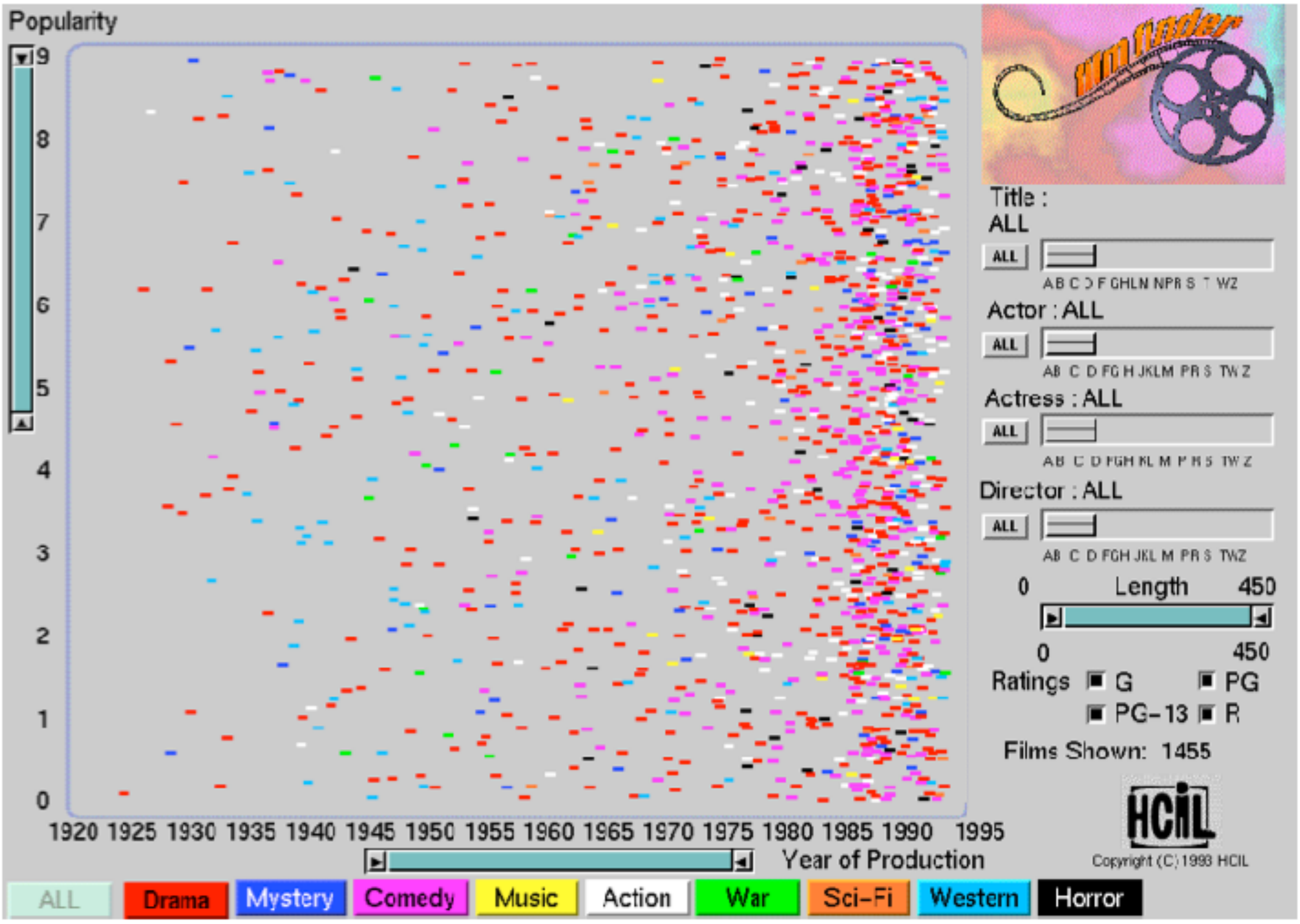
Filters: start with all, remove elements

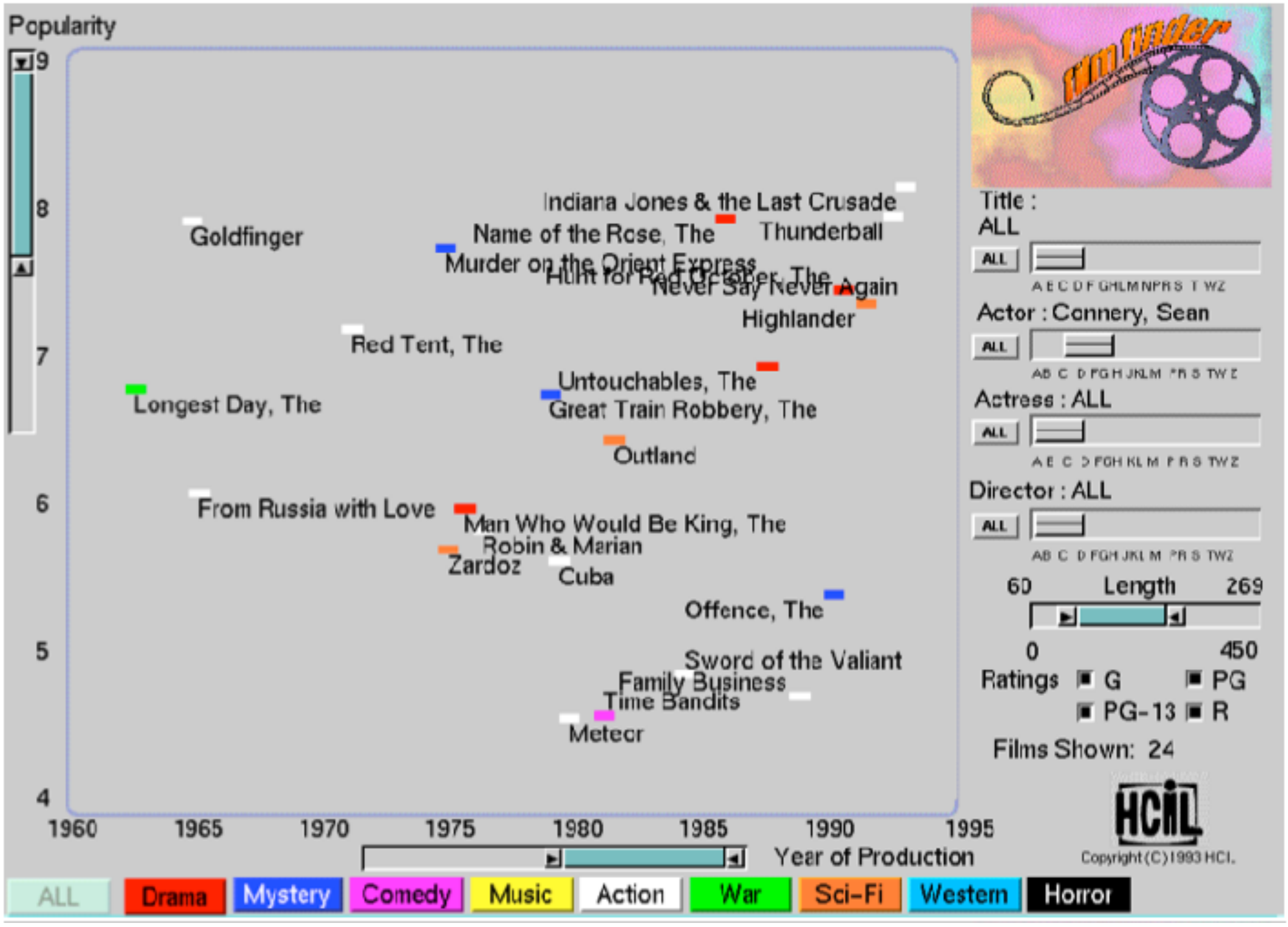
Approach depends on dataset size



Ahlberg 1994

ITEM FILTERING





Ahlberg 1994

FIND A RESTAURANT | FIND A LOCATION

Name of restaurant

FILTER

All grades

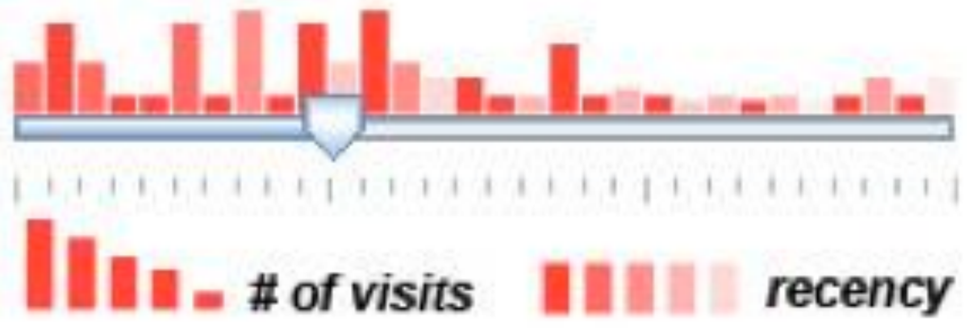
All violations

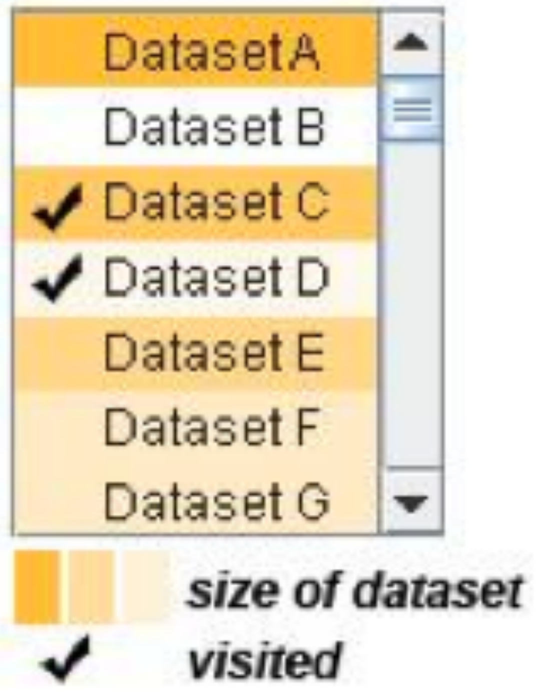
All cuisines

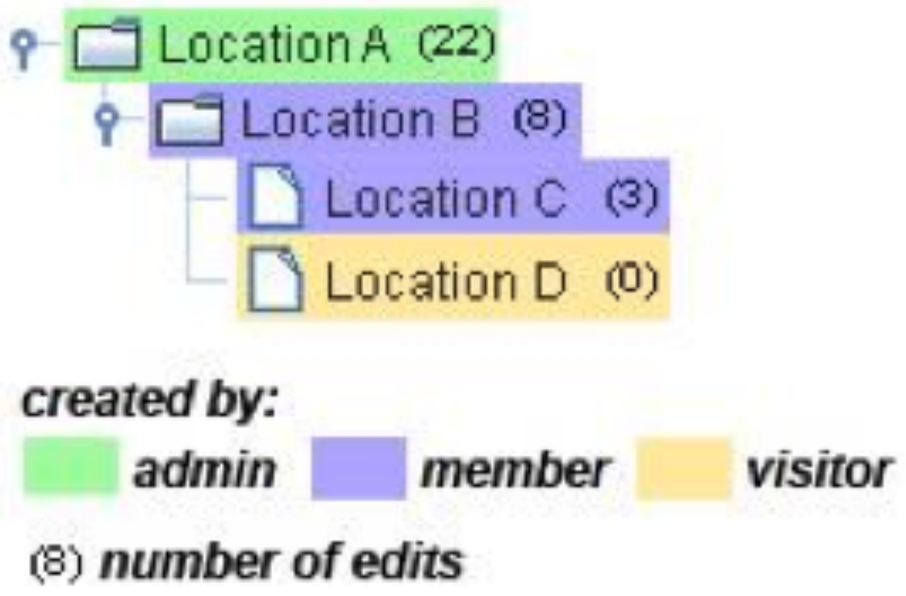
Scented Widgets

information scent: user's (imperfect) perception of data

GOAL: lower the cost of information foraging through better cues





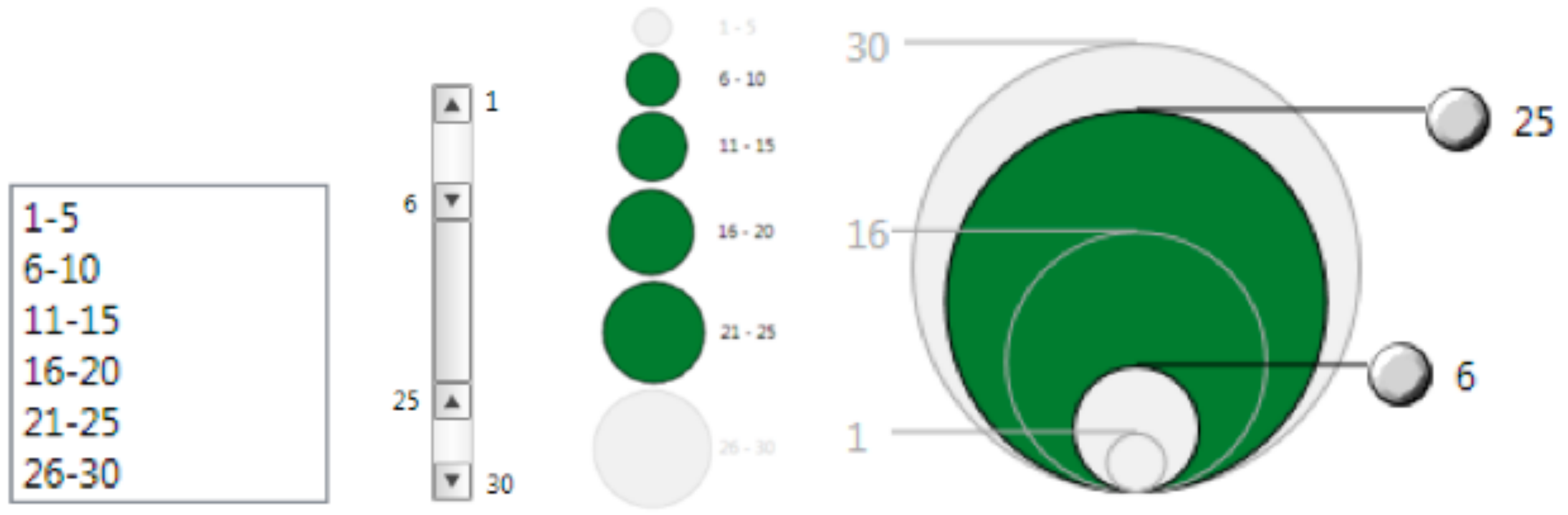


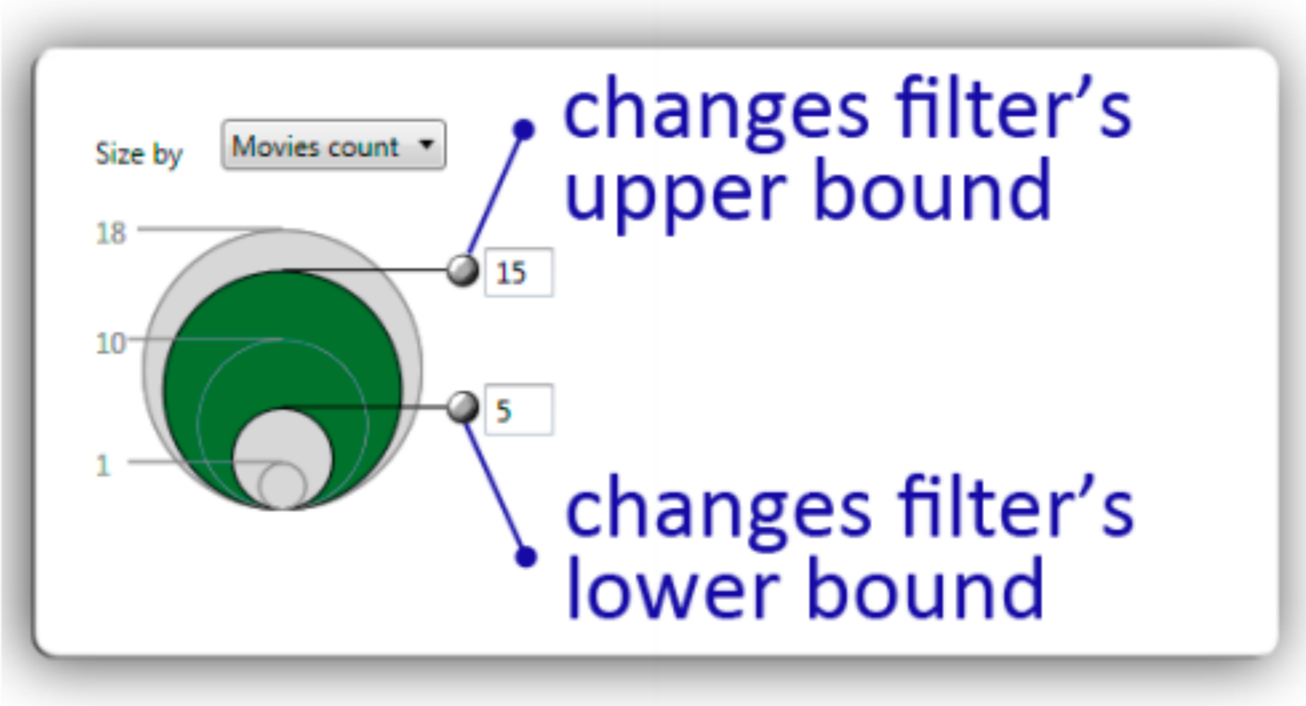
Willett 2007

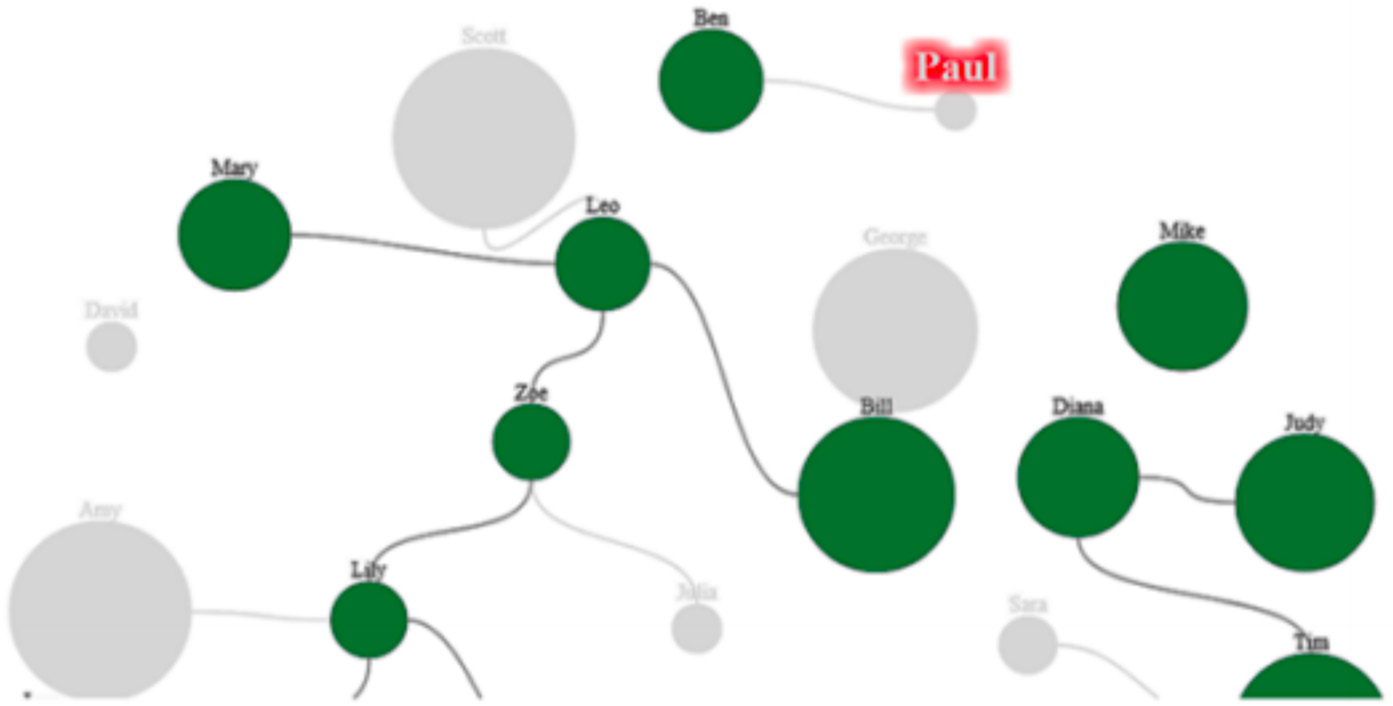
Interactive Legends

Controls combining the visual representation of static legends with interaction mechanisms of widgets

Define and control visual display together





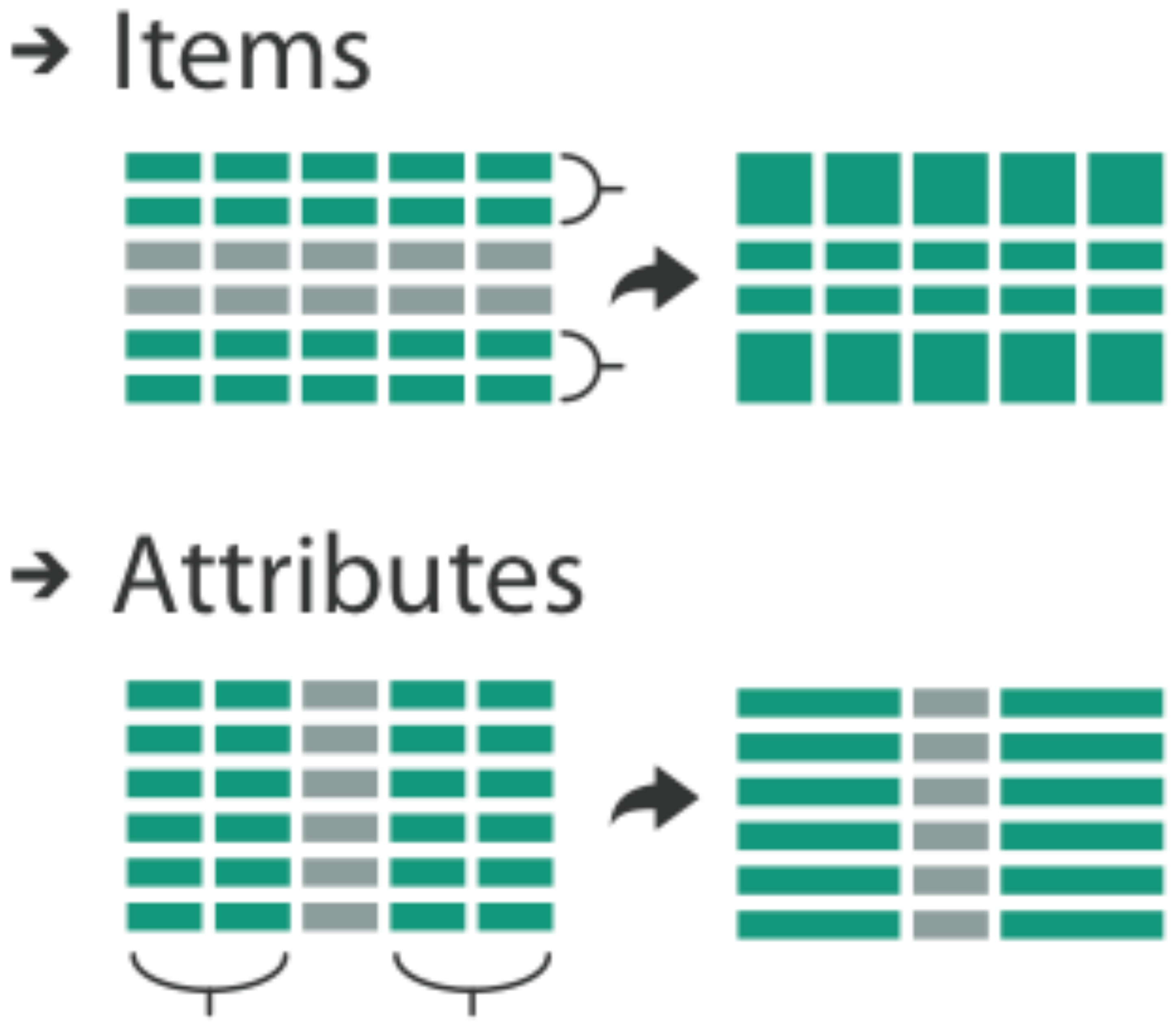


Riche 2010

Aggregation

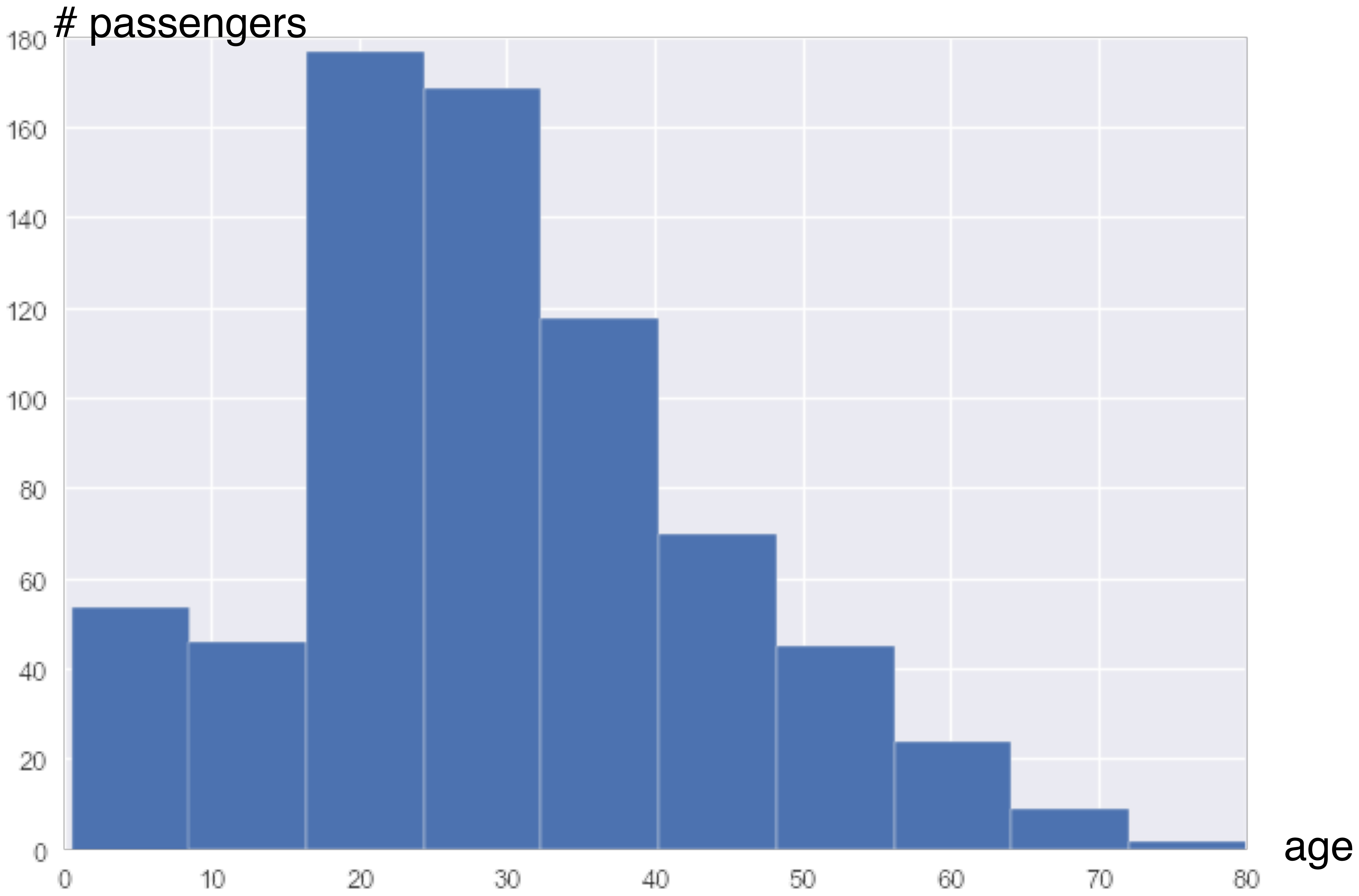
Aggregate

a group of elements is represented by a (typically smaller) number of derived elements



Item Aggregation

Histogram



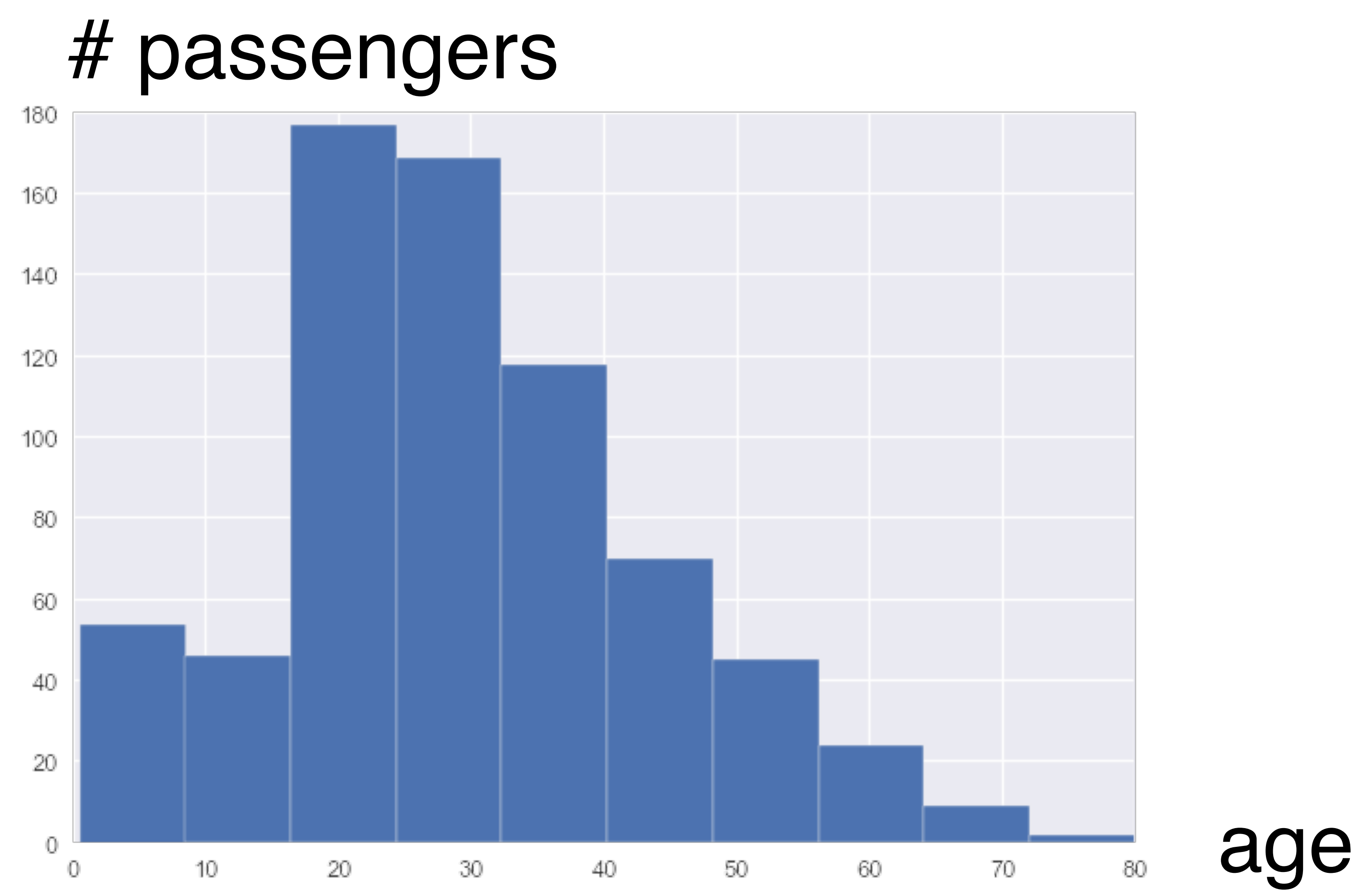
10 Bins

Histogram

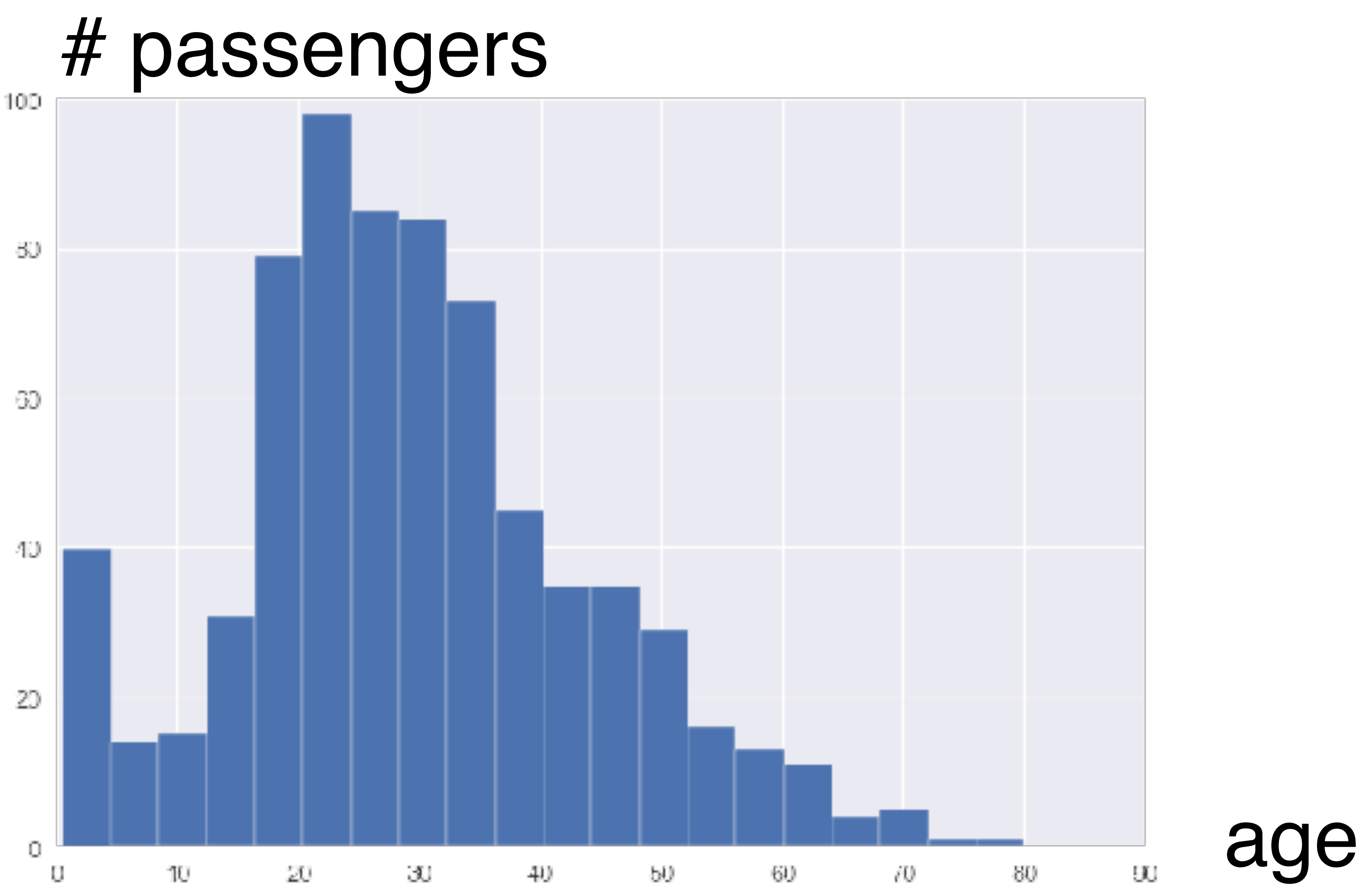
Good # bins hard to predict

make interactive!

rule of thumb: #bins = sqrt(n)

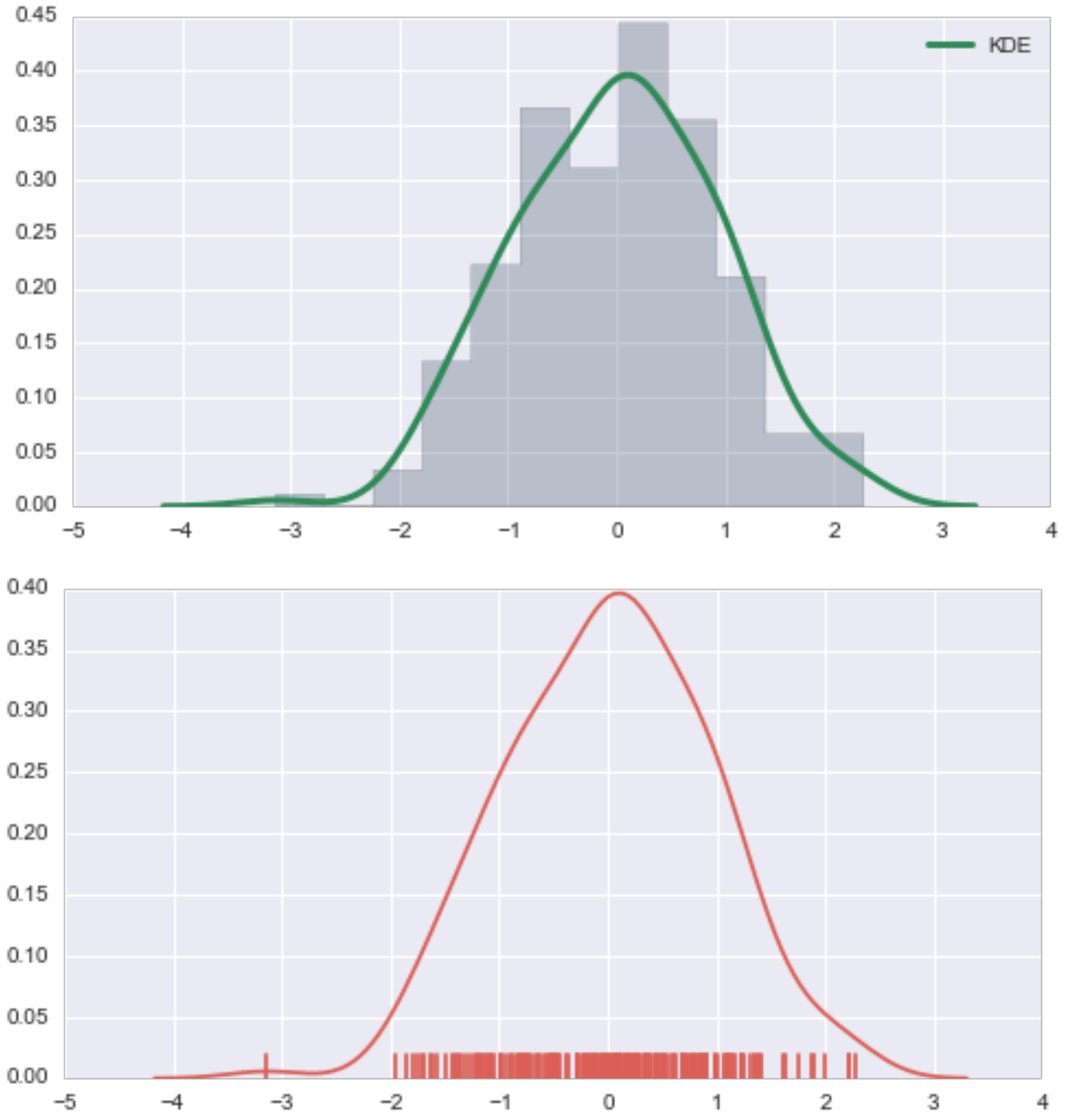


10 Bins



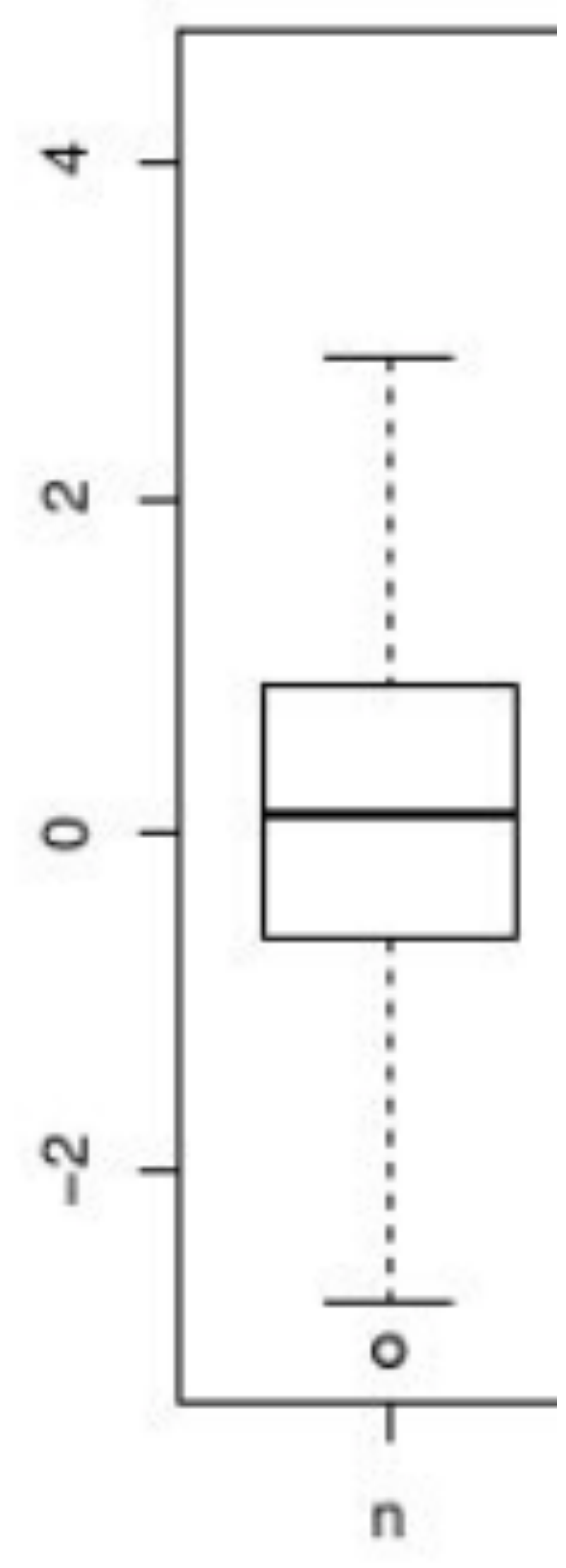
20 Bins

Density Plots

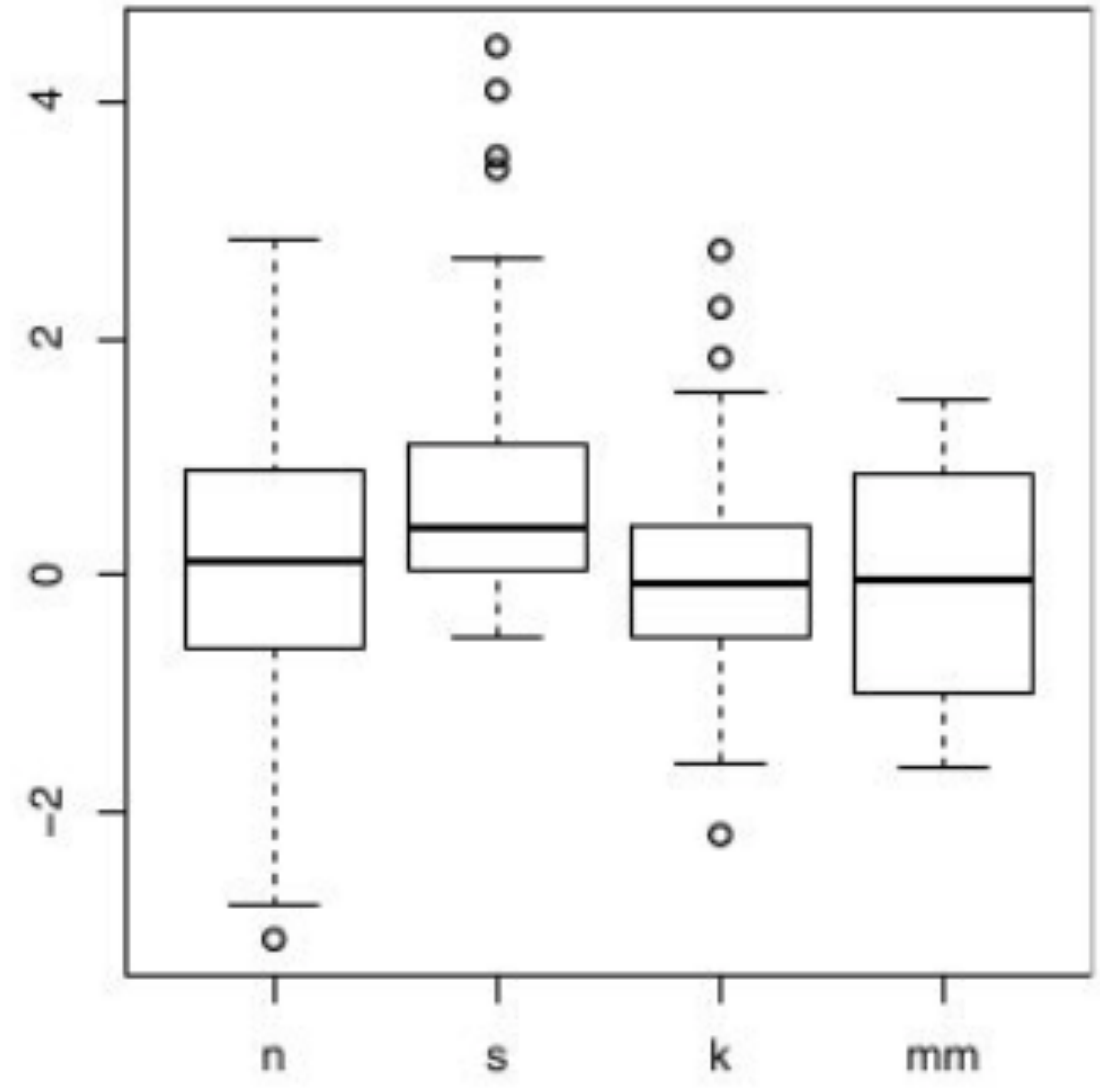


http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Box Plots
(aka Box and Whisker Plot)



Box Plots
(aka Box and Whisker Plot)



(a)

One Boxplot, Four Distributions

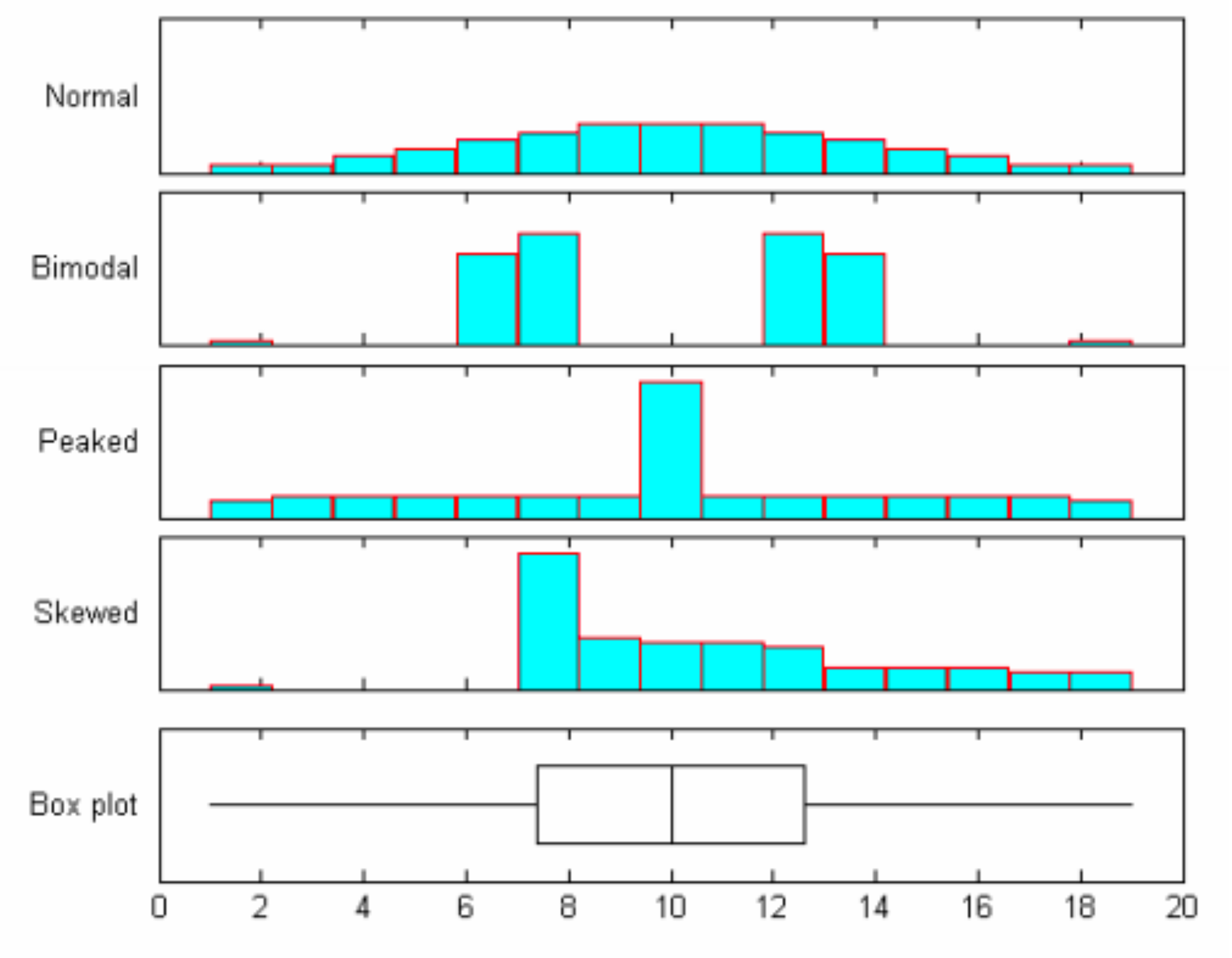
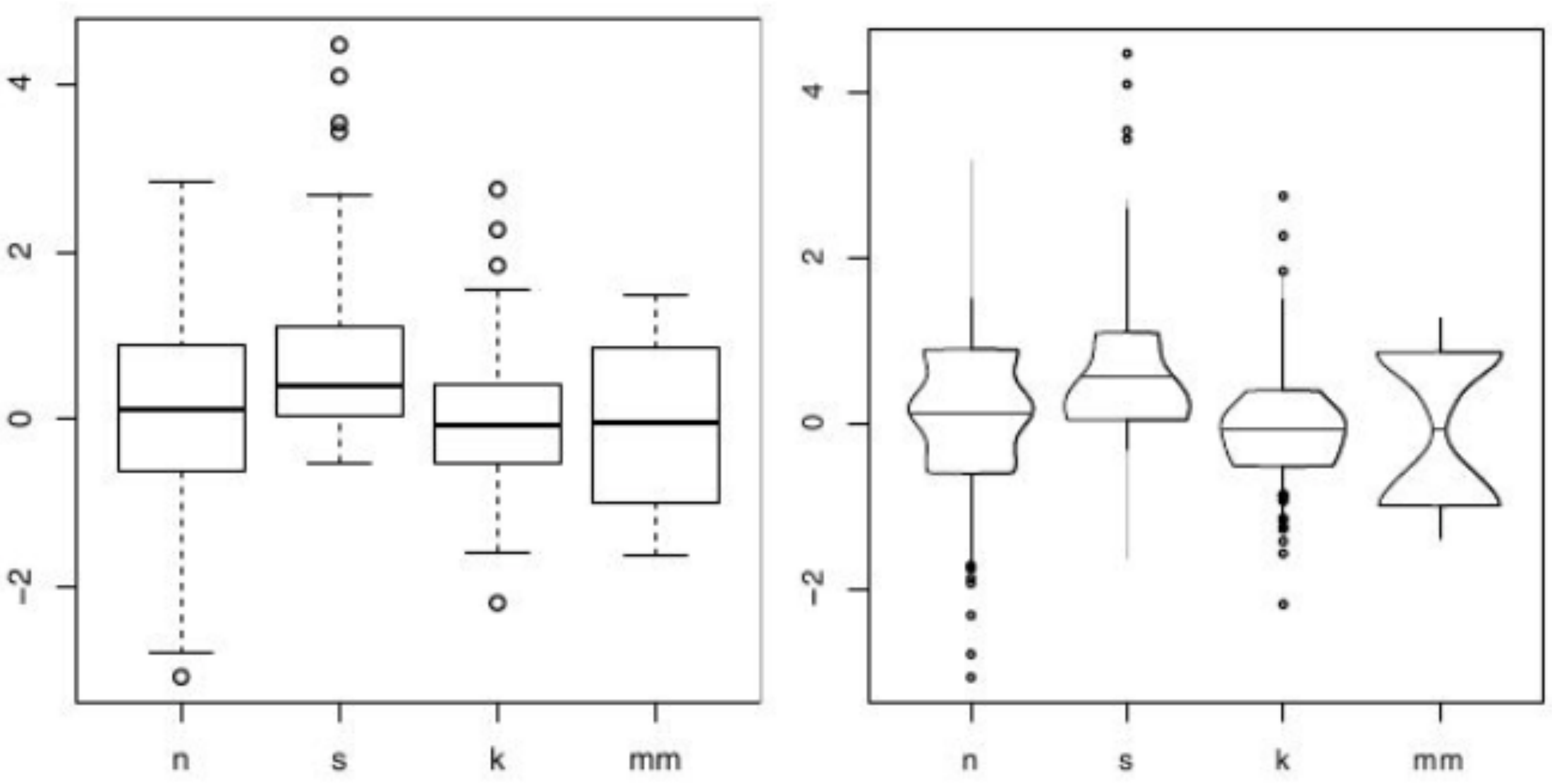


Figure 1: Histograms and box plot: four samples each of size 100

http://stat.mq.edu.au/wp-content/uploads/2014/05/Can_the_Box_Plot_be_Improved.pdf

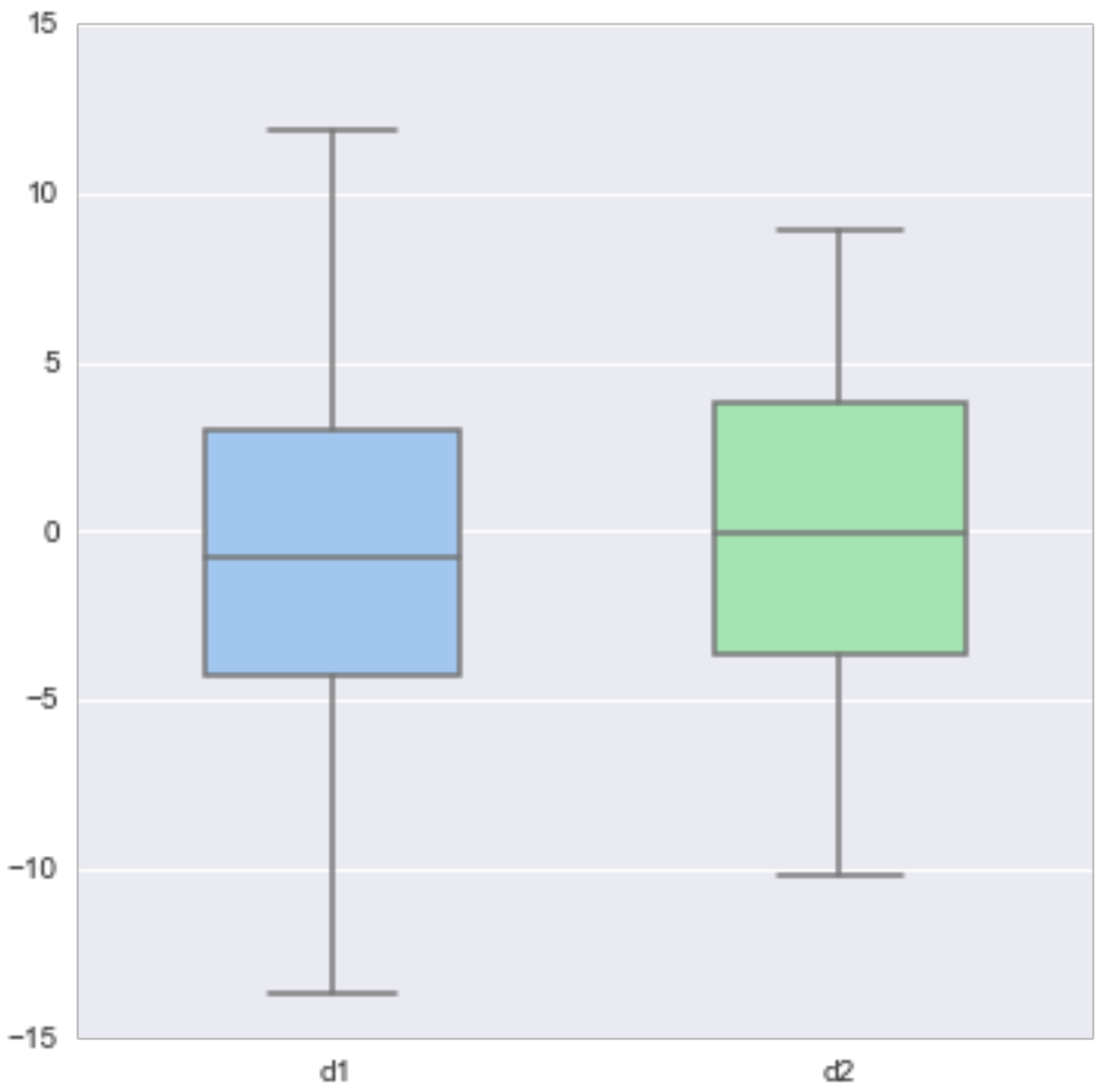
Box Plots
(aka Box and Whisker Plot)

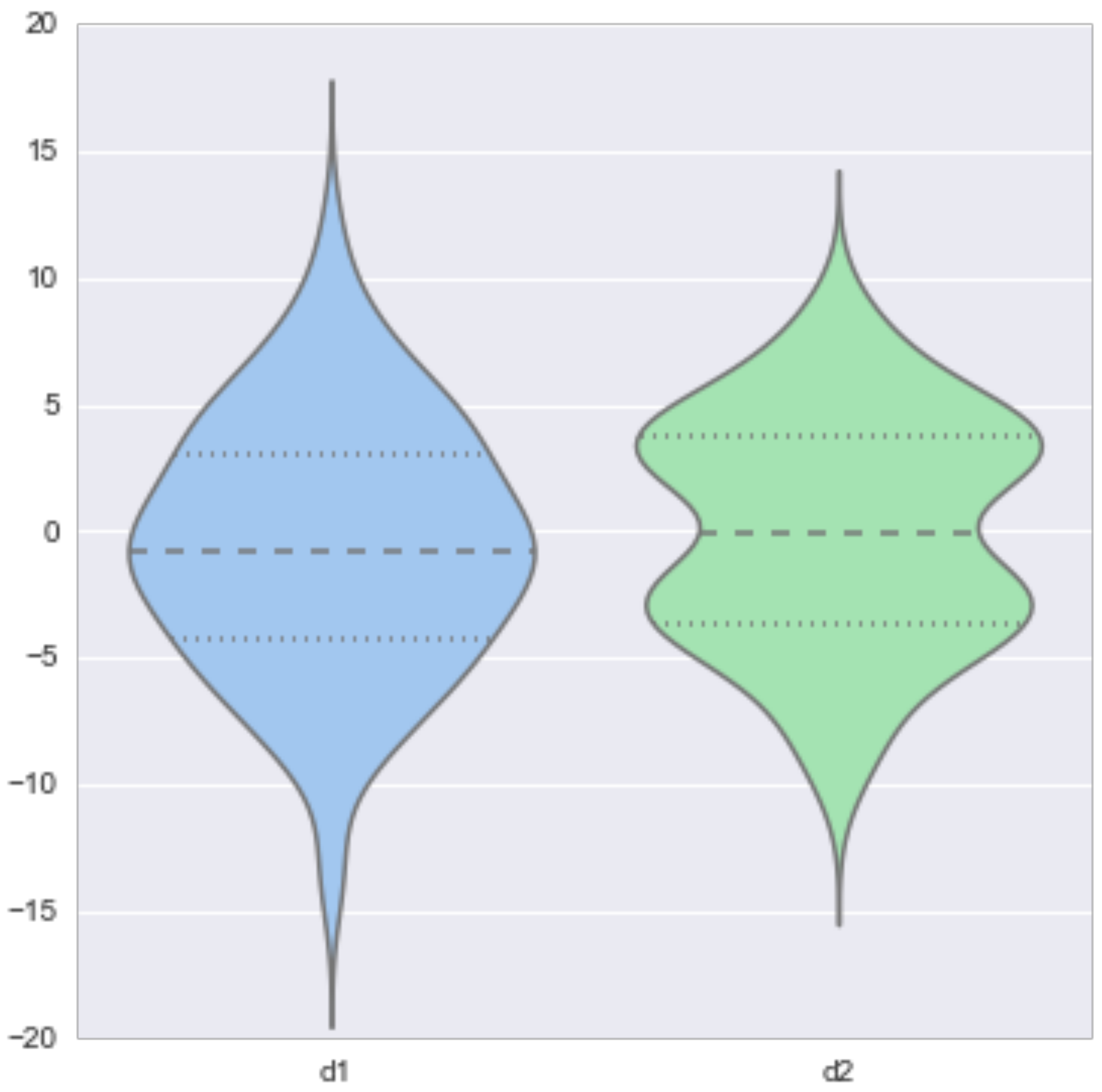


(a) (b)

Violin Plot

= Box Plot + Probability Density Function



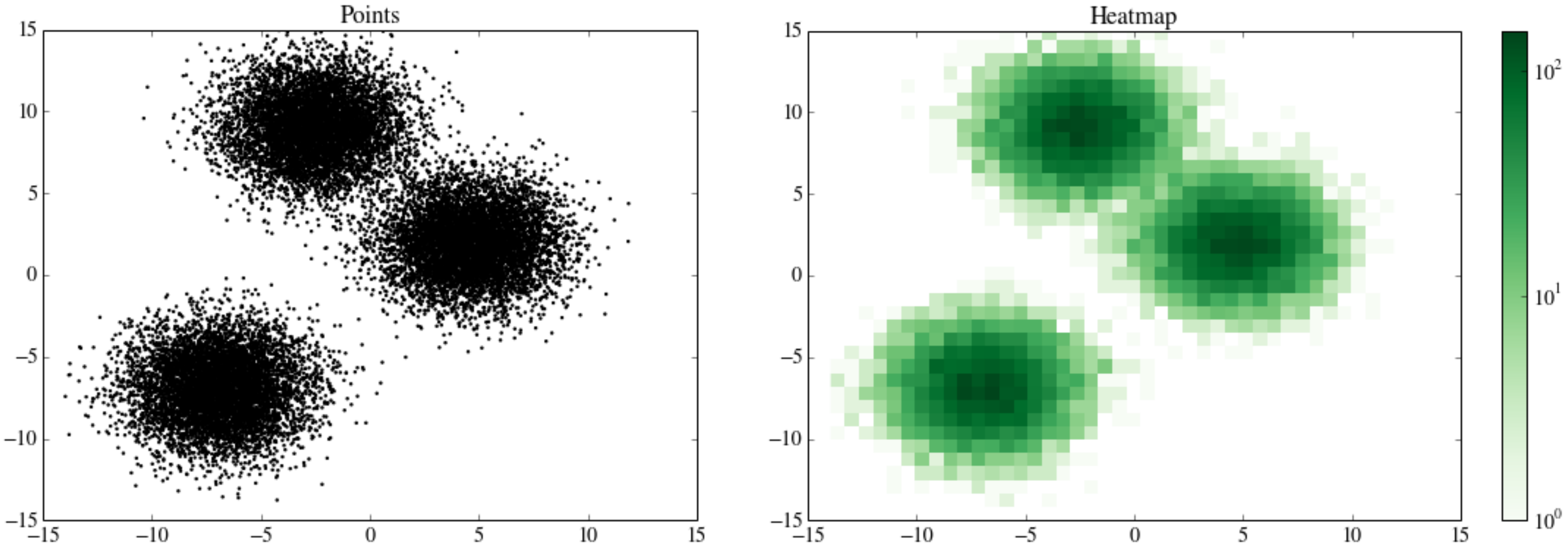


http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Heat Maps

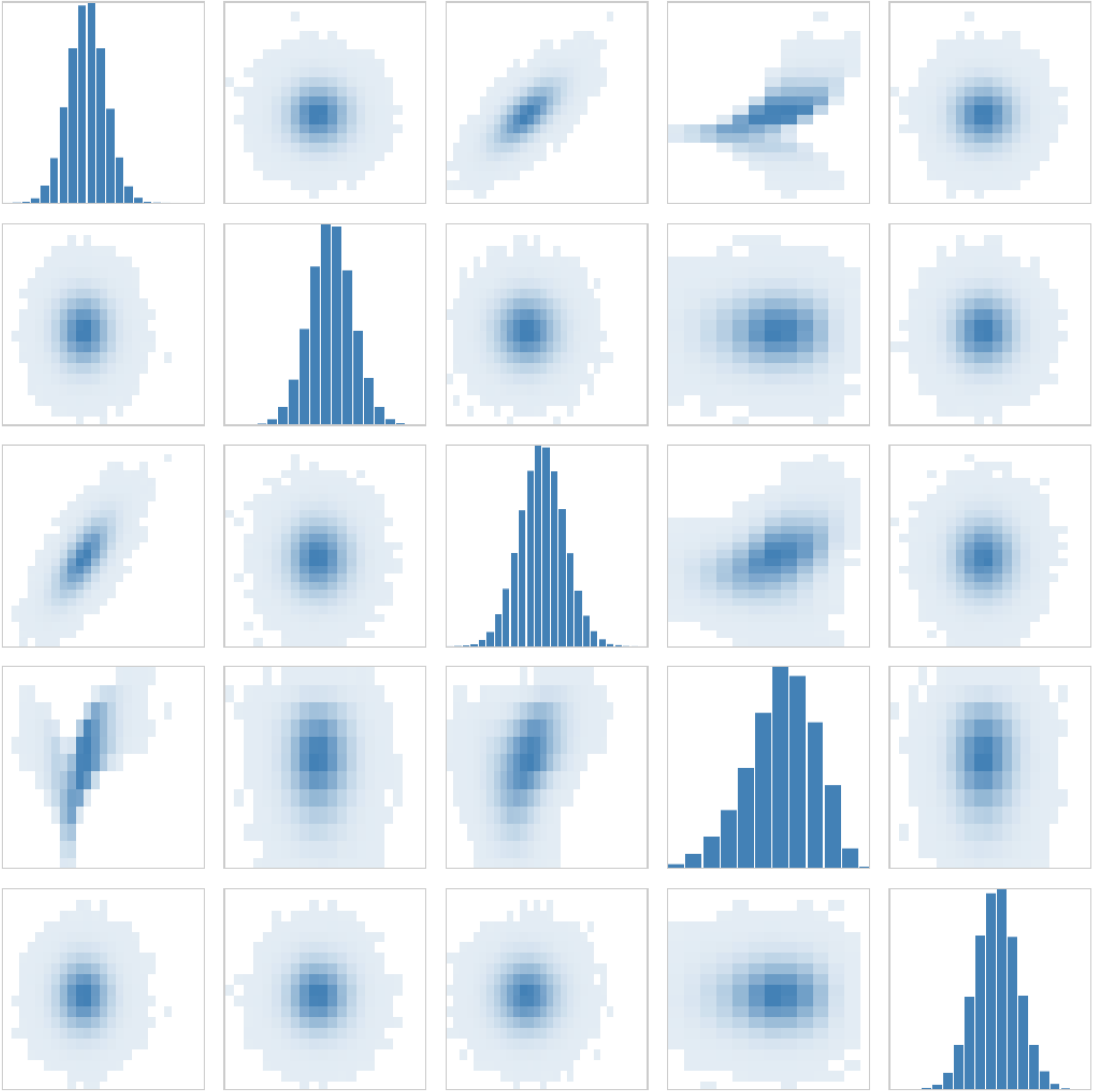
binning of scatterplots

instead of drawing every point, calculate grid and intensities



2D Density Plots

Interactive Binned Scatterplot Matrix Dimensions: 5 Bins: 20 Data Points: 100k

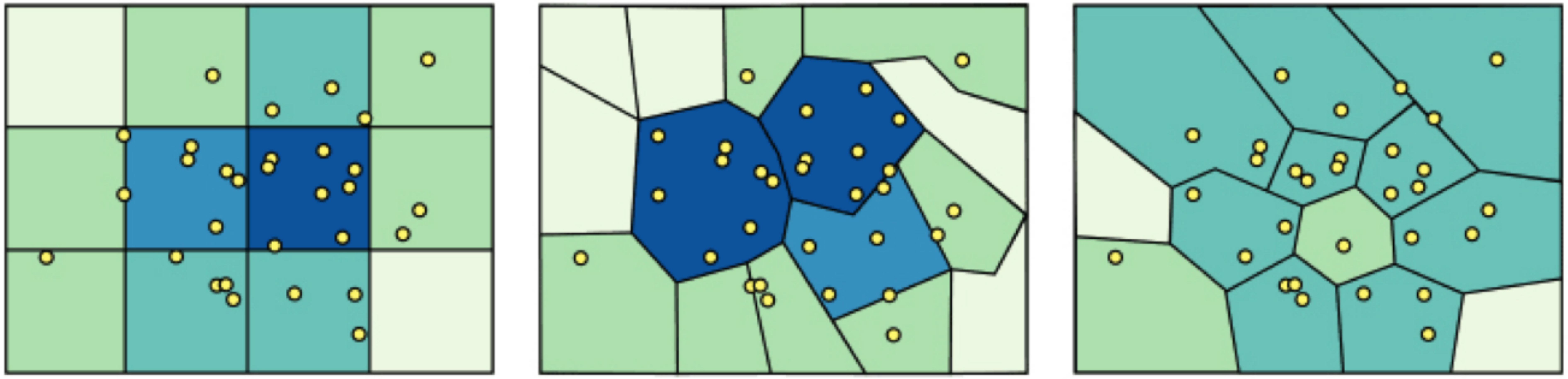


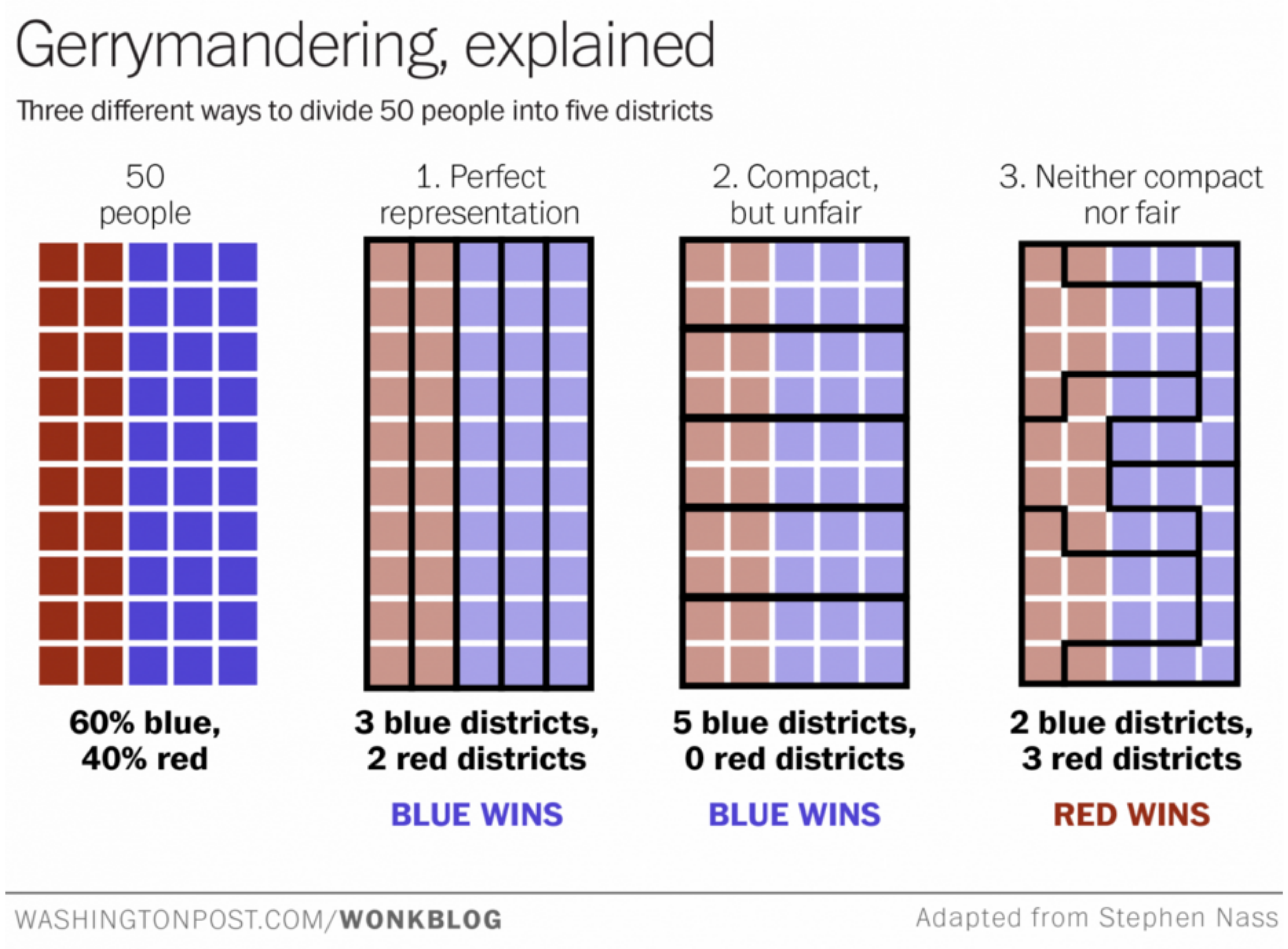
Powered by Datavore and D3.

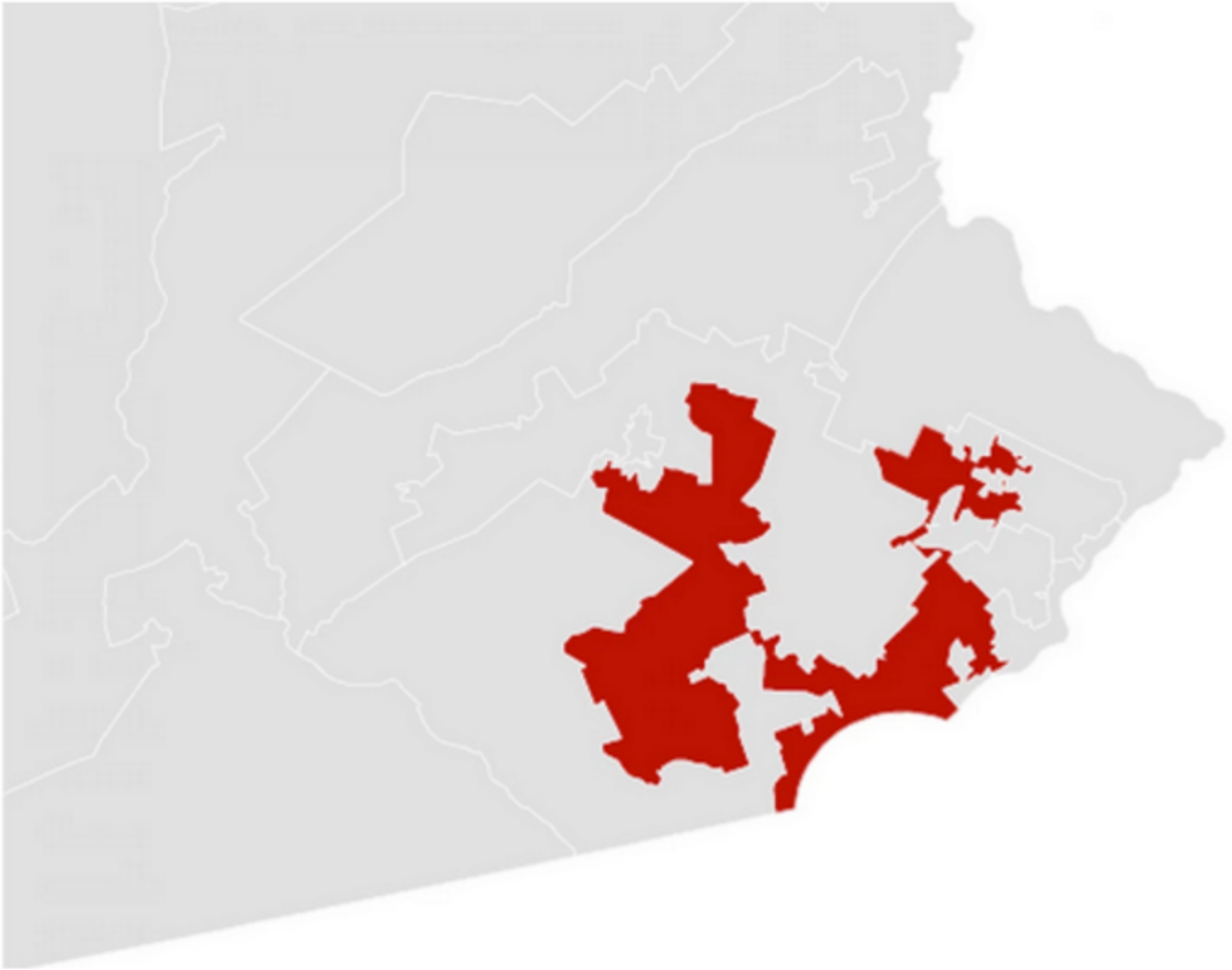
Spatial Aggregation

modifiable areal unit problem

in cartography, changing the boundaries of the regions used to analyze data
can yield dramatically different results



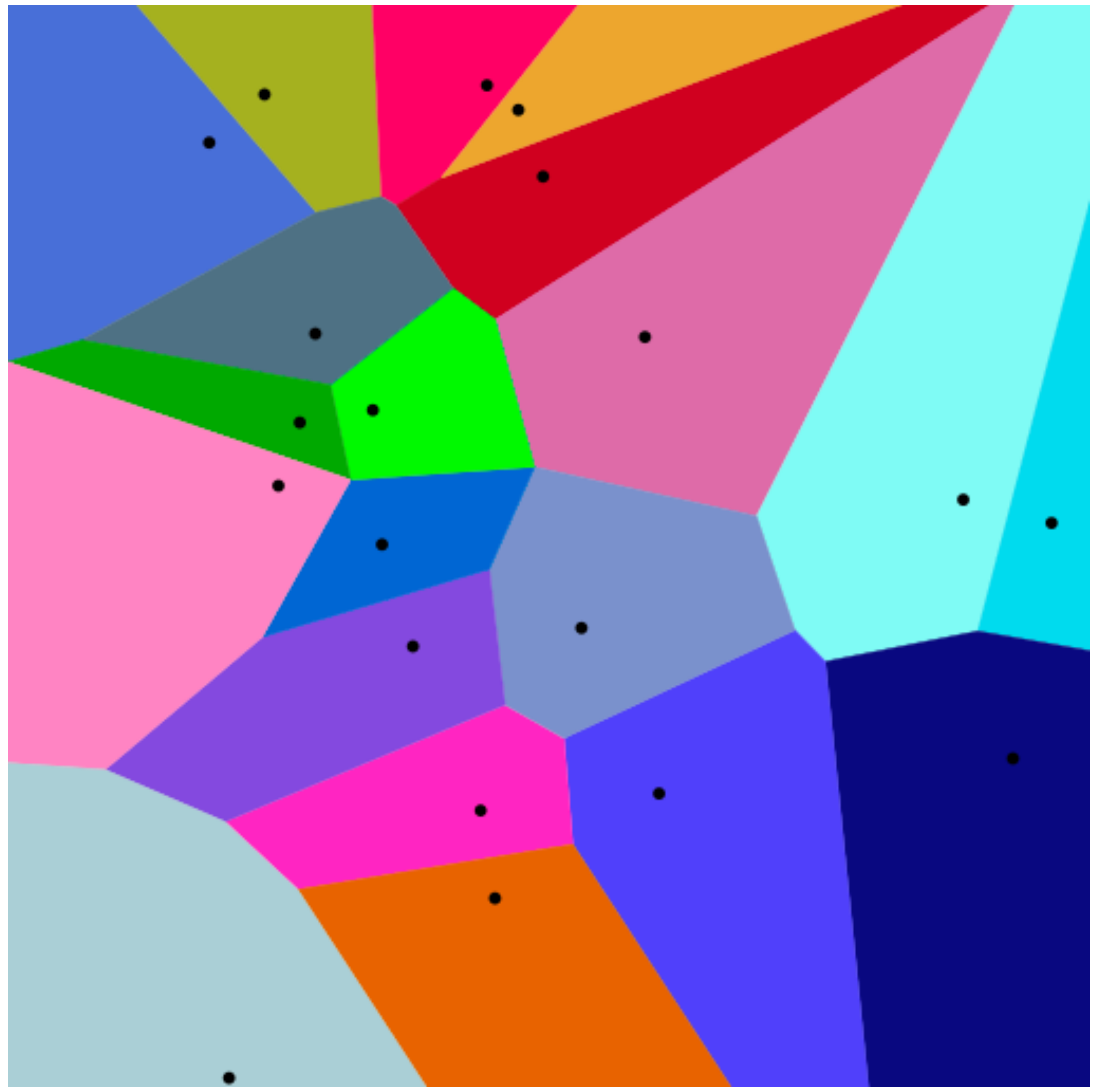




A real district in Pennsylvania
Democrats won 51% of the vote
but only 5 out of 18 house seats

Voronoi Diagrams

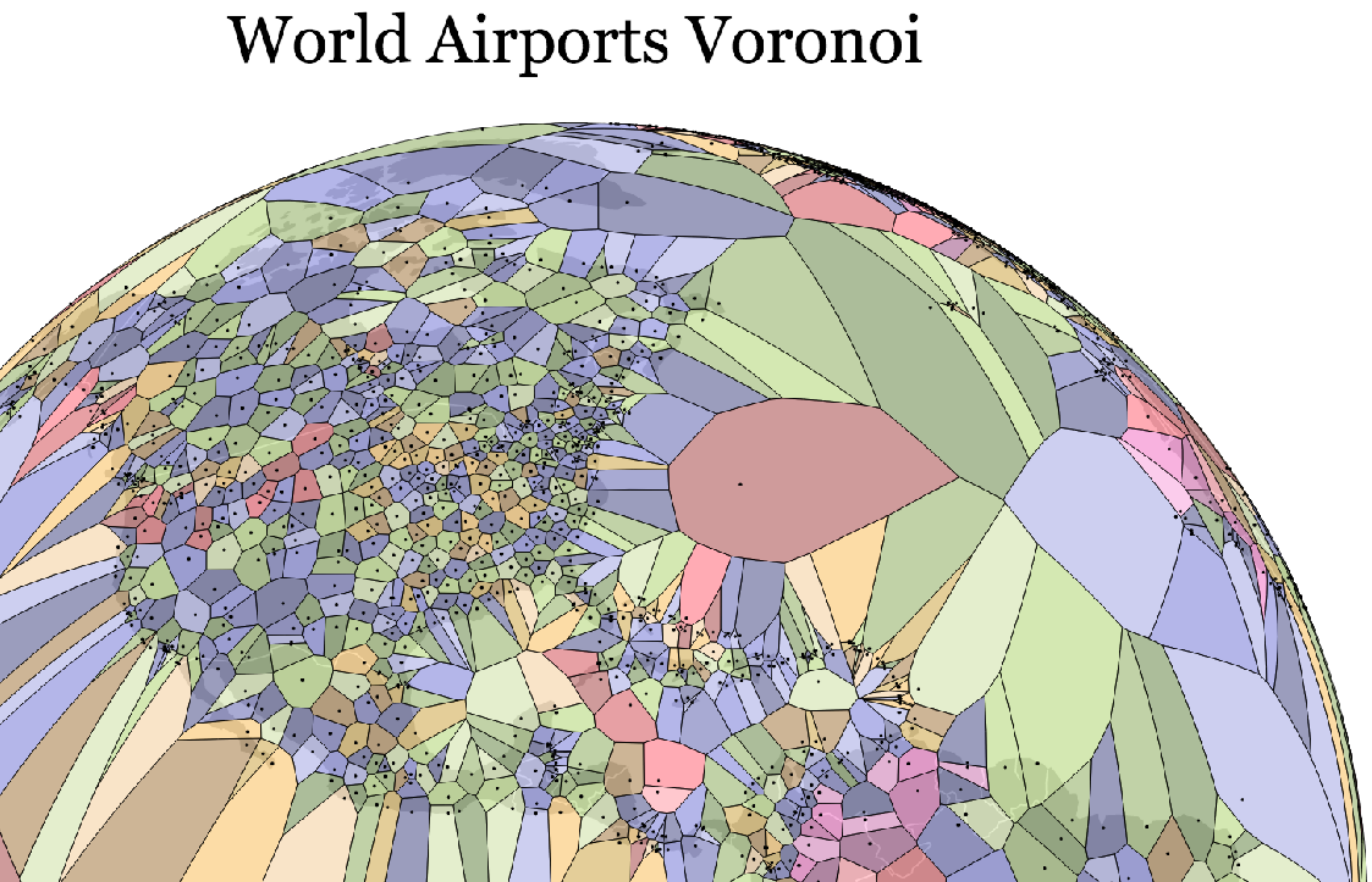
Given a set of locations, for which area is a location n closest?



D3 Voronoi Layout:

https://github.com/d3/d3-voronoi

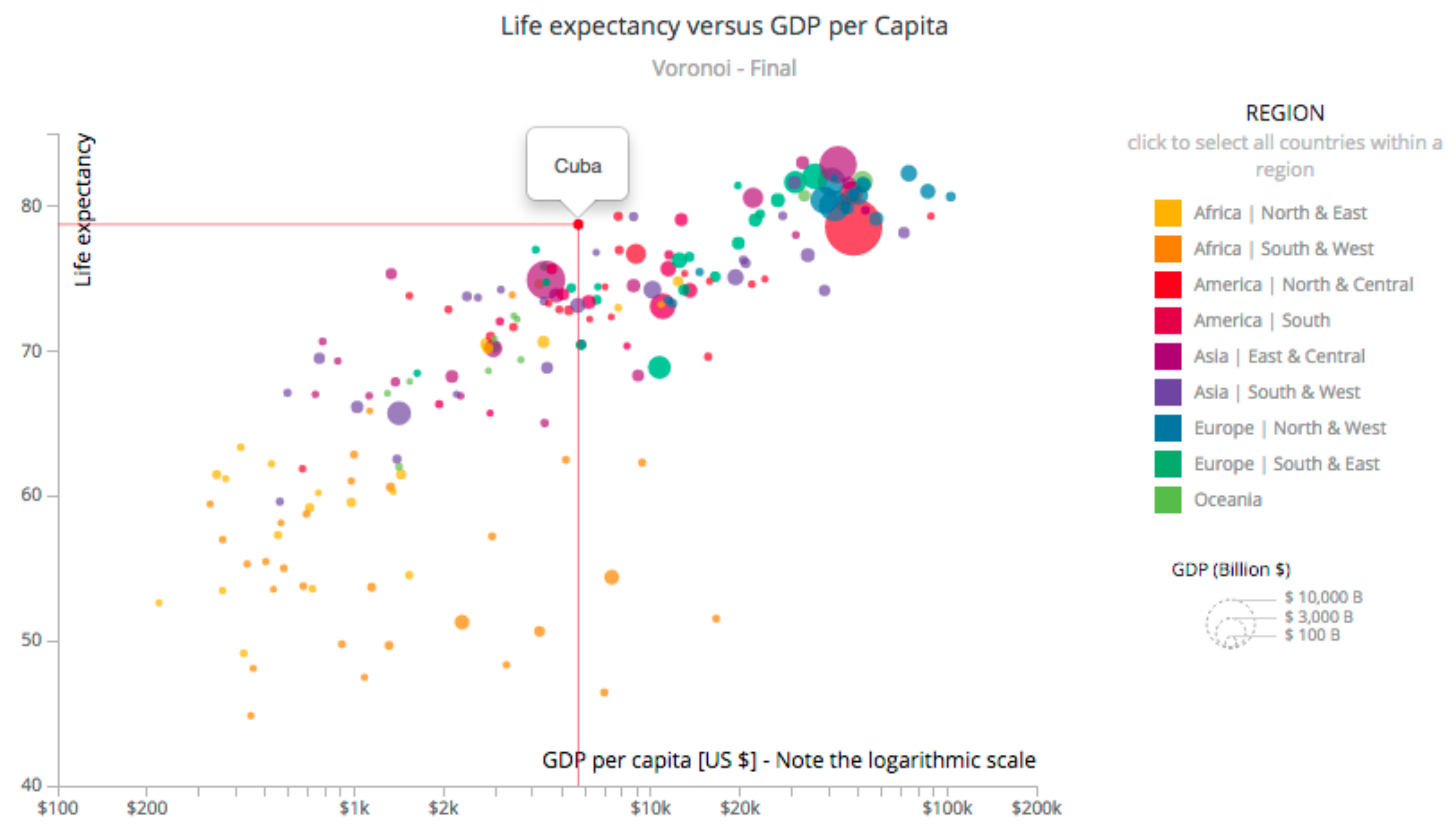
Voronoi Examples

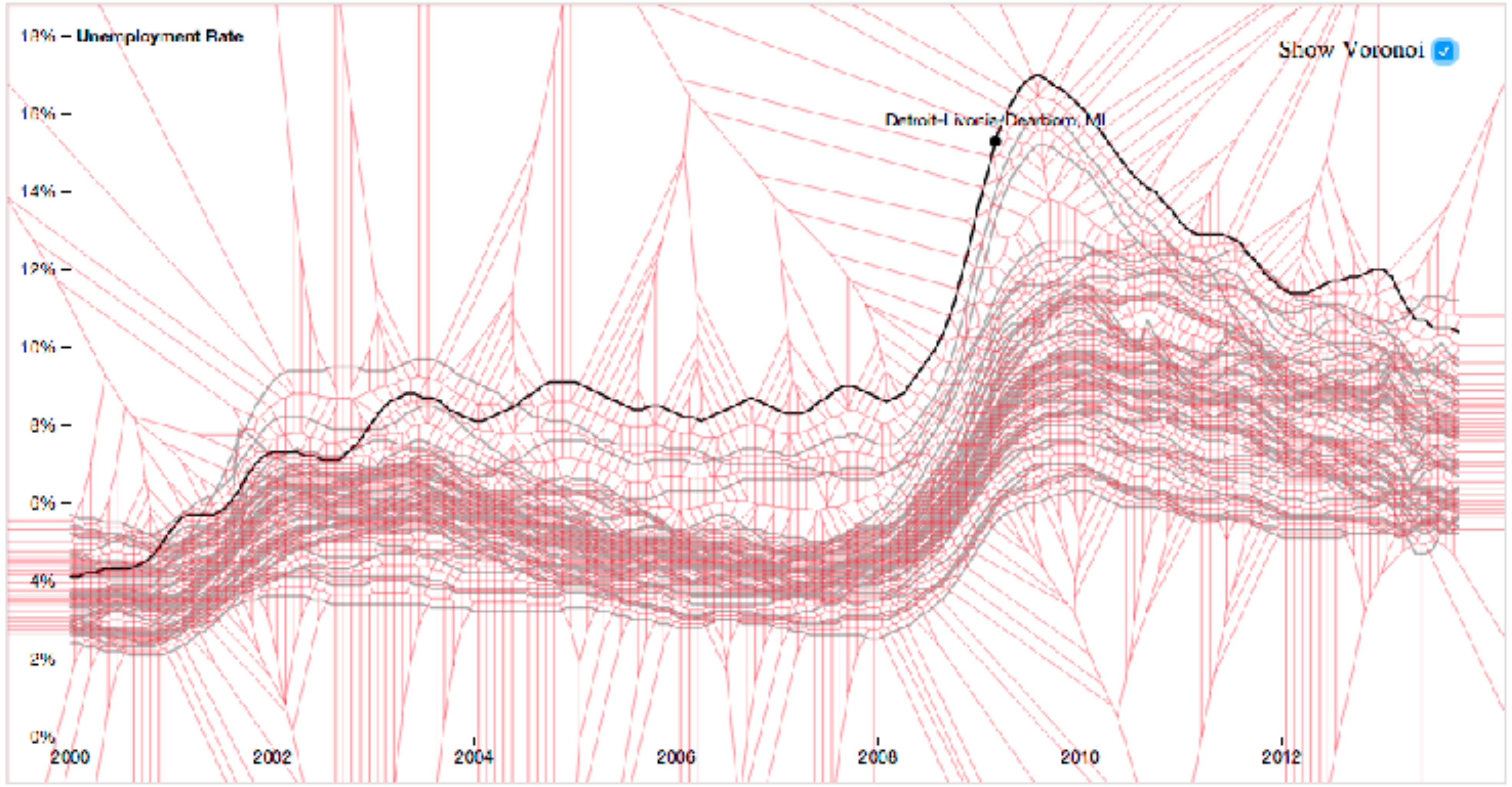


Voronoi for Interaction

Useful for interaction:
Increase size of target area to click/hover

Instead of clicking on point, hover in its region





https://github.com/d3/d3-voronoi/

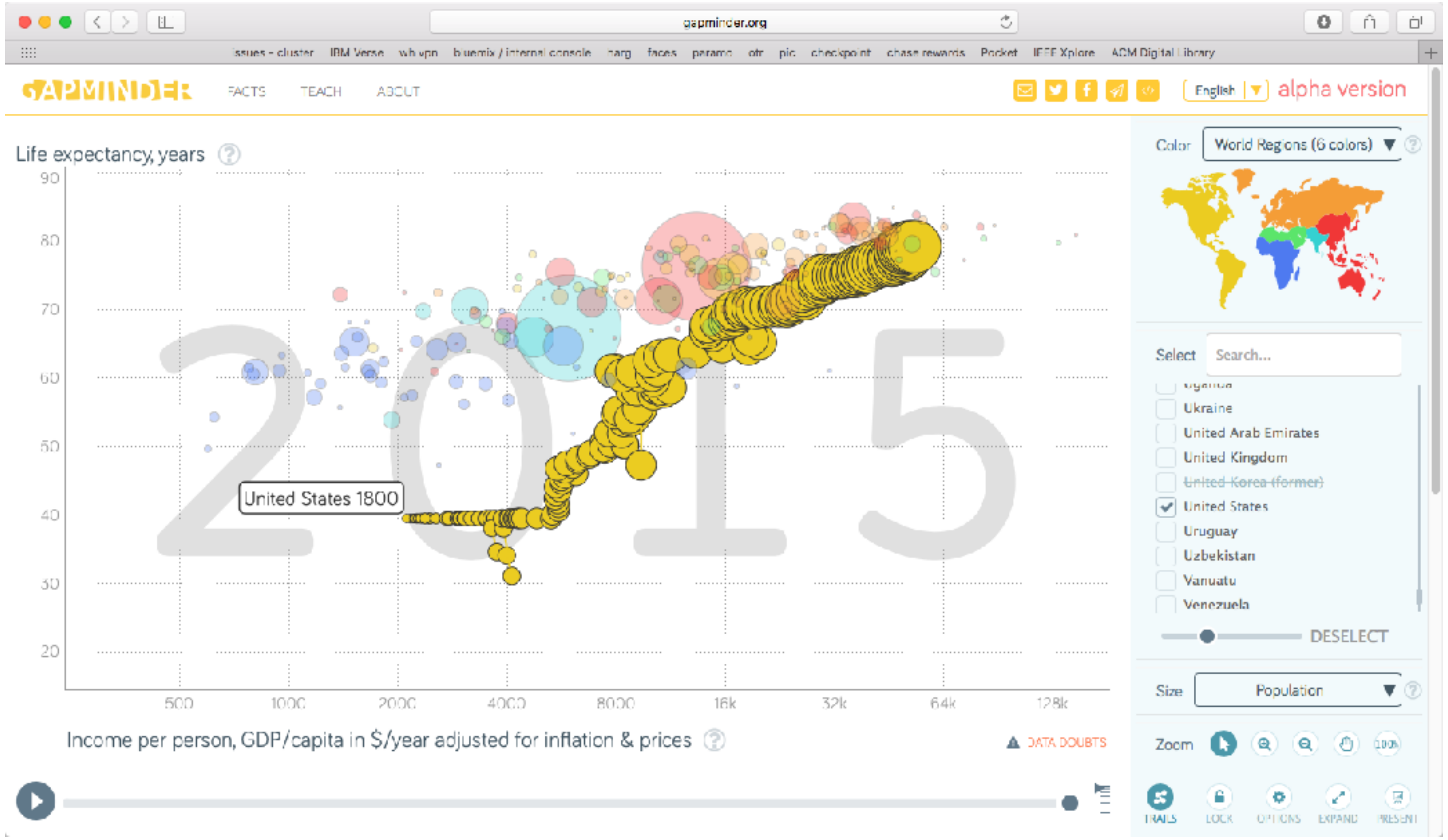
Design Critique

GapMinder - http://www.gapminder.org/tools

In breakout groups,

find an interesting story using this tool.

Change the axes and/or visual channels that demonstrated a new insight to you!



Attribute aggregation

1) group attributes and compute
a similarity score across the set

2) dimensionality reduction,
to preserve meaningful structure

Attribute aggregation

1) group attributes and compute a similarity score across the set

2) dimensionality reduction, to preserve meaningful structure

Clustering

Classification of items into “similar” bins

Based on similarity measures

- Euclidean distance, Pearson correlation, ...

Partitional Algorithms

- divide data into set of bins
- # bins either manually set (e.g., k-means) or automatically determined (e.g., affinity propagation)

Hierarchical Algorithms

- Produce “similarity tree” – dendrogram

Bi-Clustering

- Clusters dimensions & records

Fuzzy clustering

- allows occurrence of elements in multiples clusters

Clustering Applications

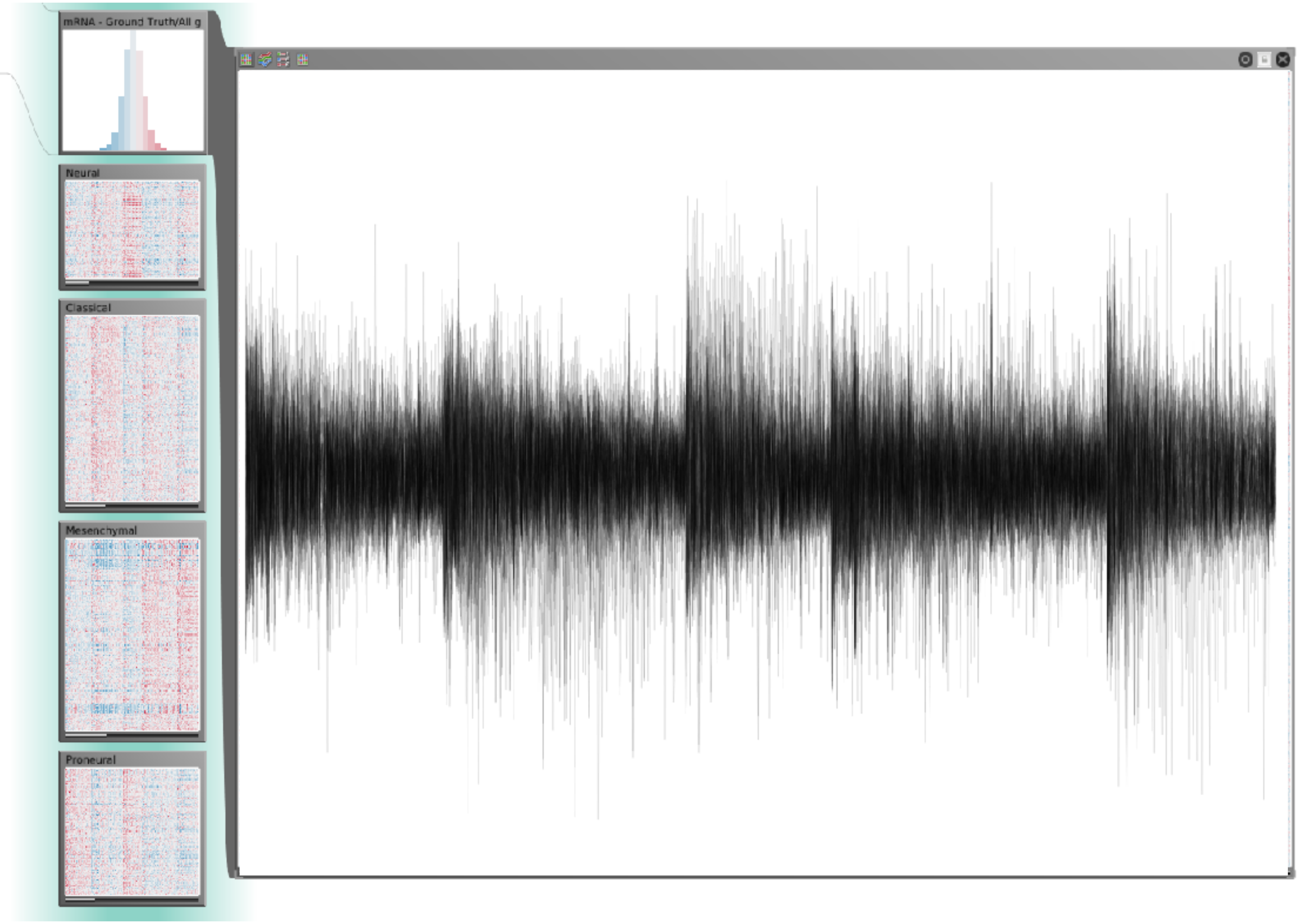
Clusters can be used to

- order (pixel based techniques)
- brush (geometric techniques)
- aggregate

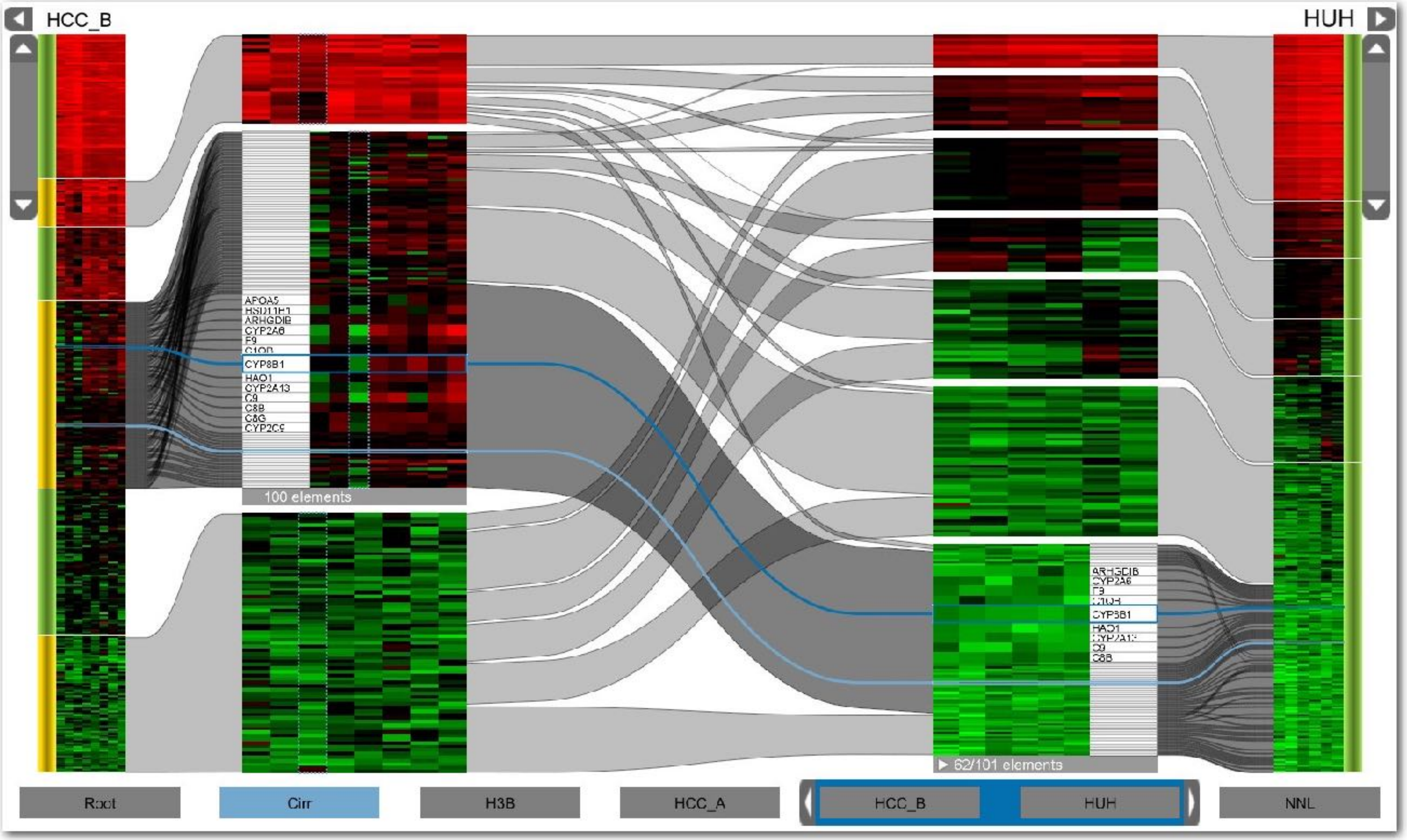
Aggregation

- cluster more homogeneous than whole dataset
- statistical measures, distributions, etc. more meaningful

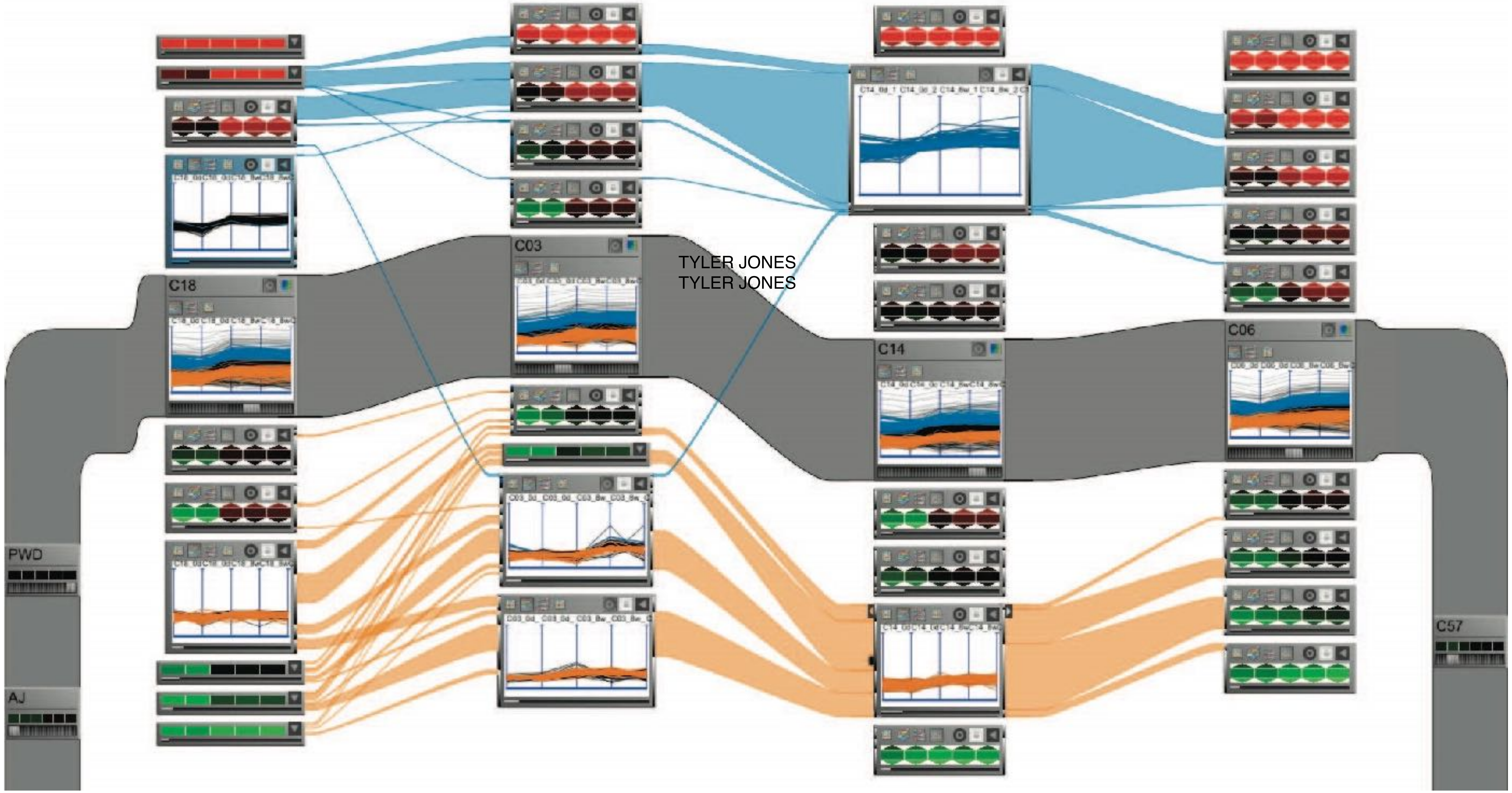
Clustered Heat Map



Cluster Comparison



Aggregation



Example: K-Means

Goal: Minimize aggregate intra-custer distance (*inertia*)

$$\underset{C}{argmin} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

total squared distance from point to center of its cluster

for euclidian distance: this is the variance

measure of how internally coherent clusters are

Lloyd's Algorithm

Input: set of records $x_1 \dots x_n$, and k (nr clusters)

Pick k starting points as centroids $c_1 \dots c_k$

While not converged:

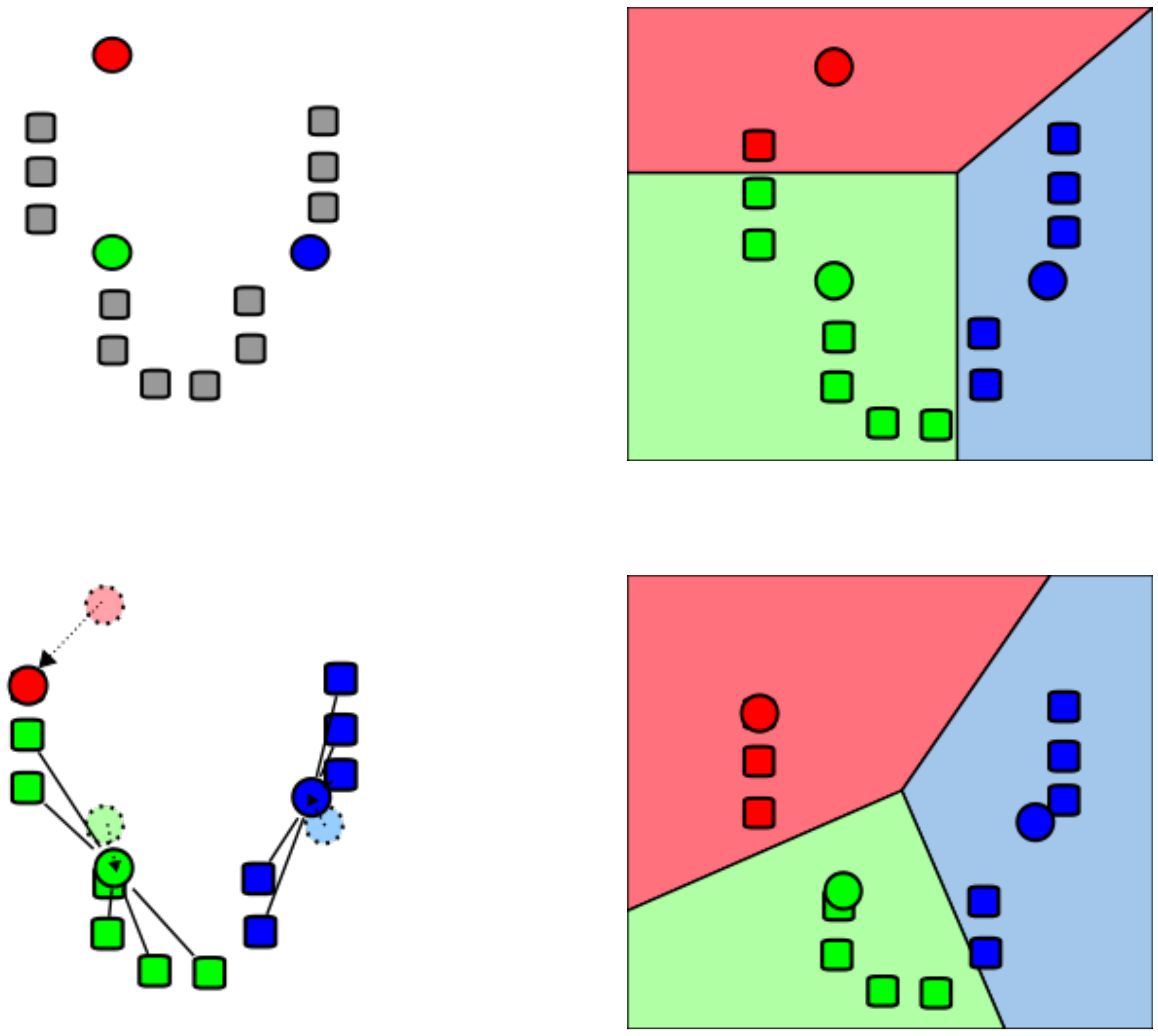
1. for each point x_i find closest centroid c_j
   - for every c_j calculate distance $D(x_i, c_j)$
   - assign x_i to cluster j defined by smallest distance
2. for each cluster j, compute a new centroid c_j
   by calculating the average of all x_i assigned to cluster j

Repeat until convergence, e.g.,

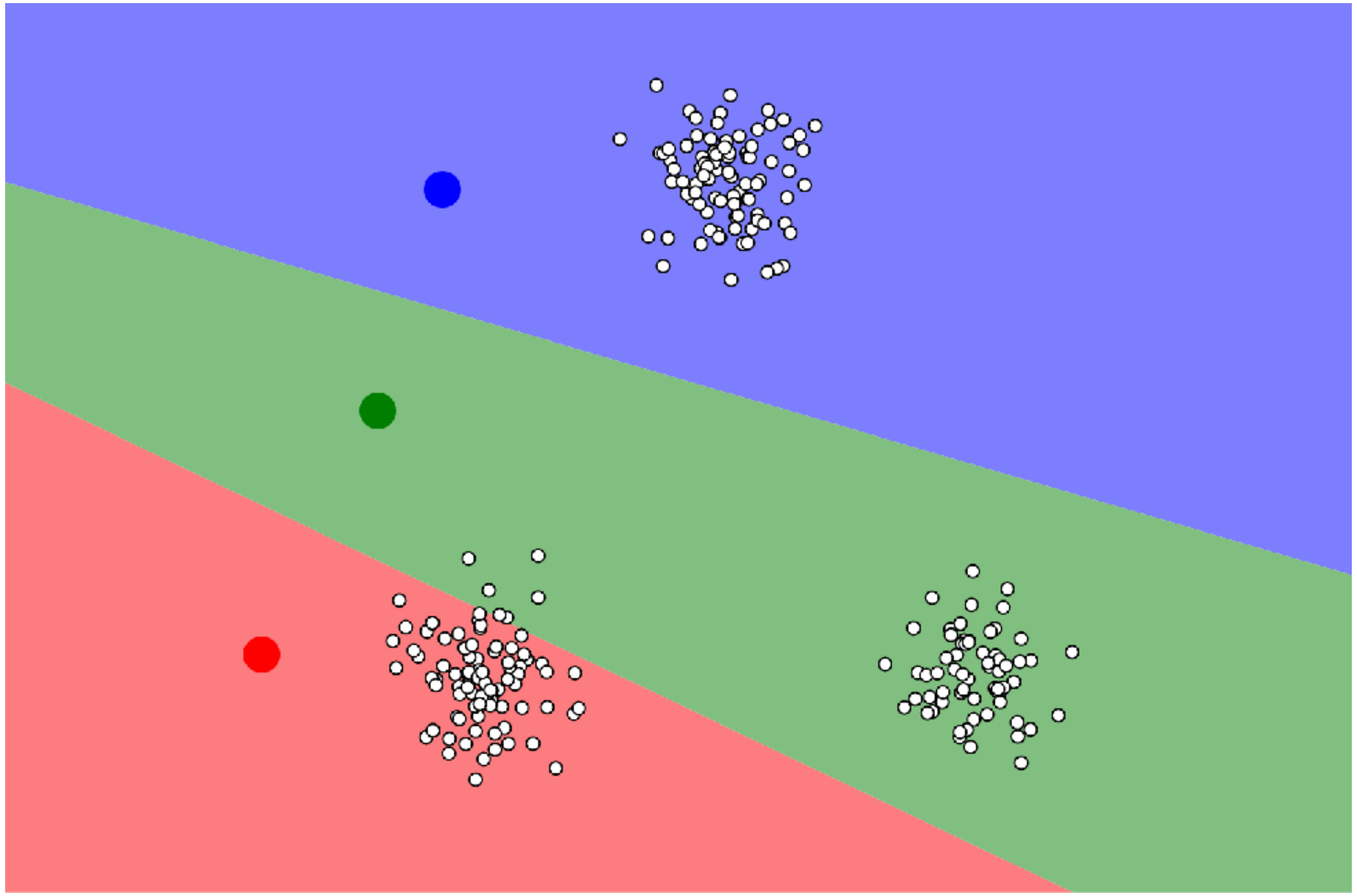
no point has changed cluster

distance between old and new centroid below threshold

number of max iterations reached

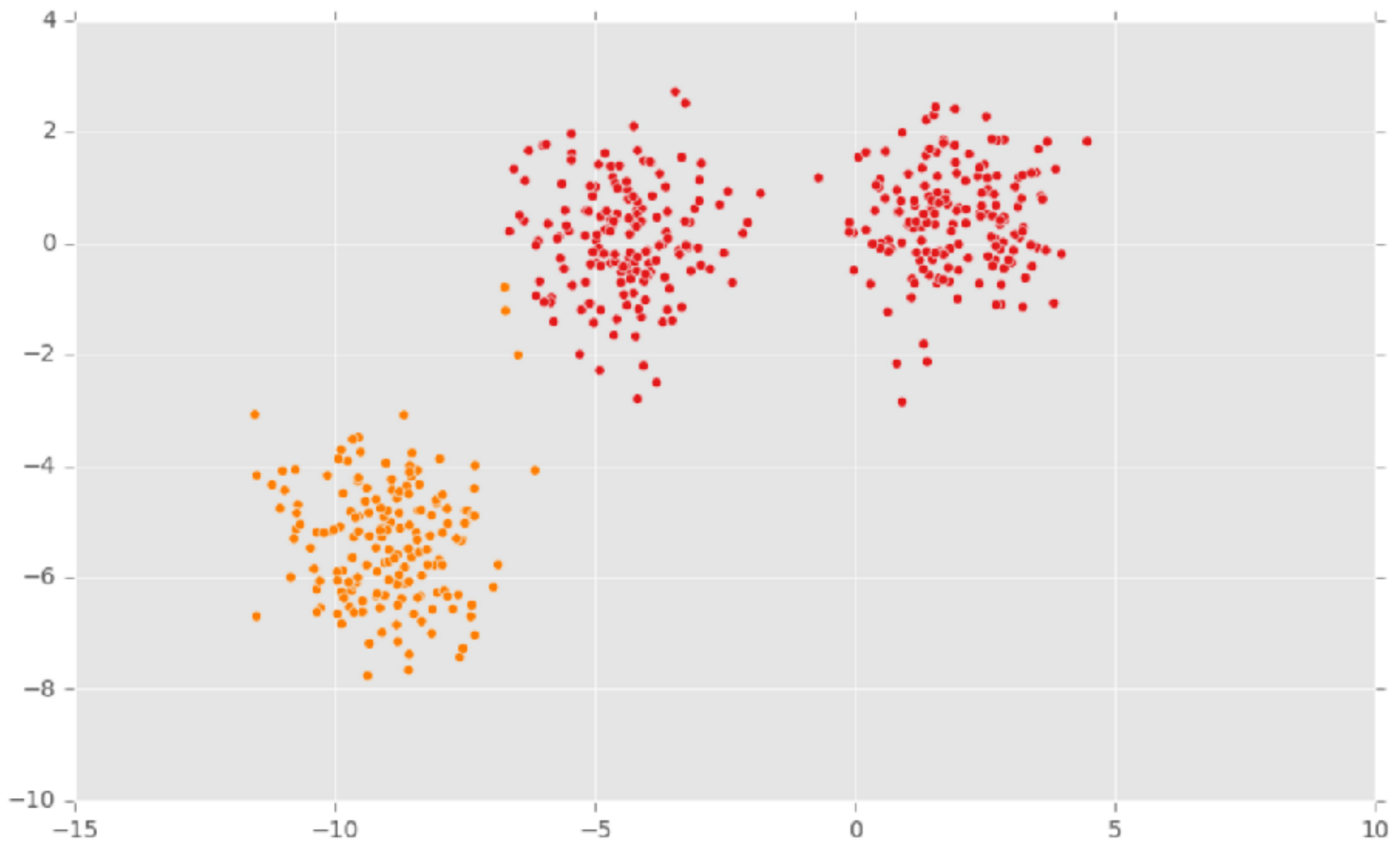


Illustrated



https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Choosing K



Properties

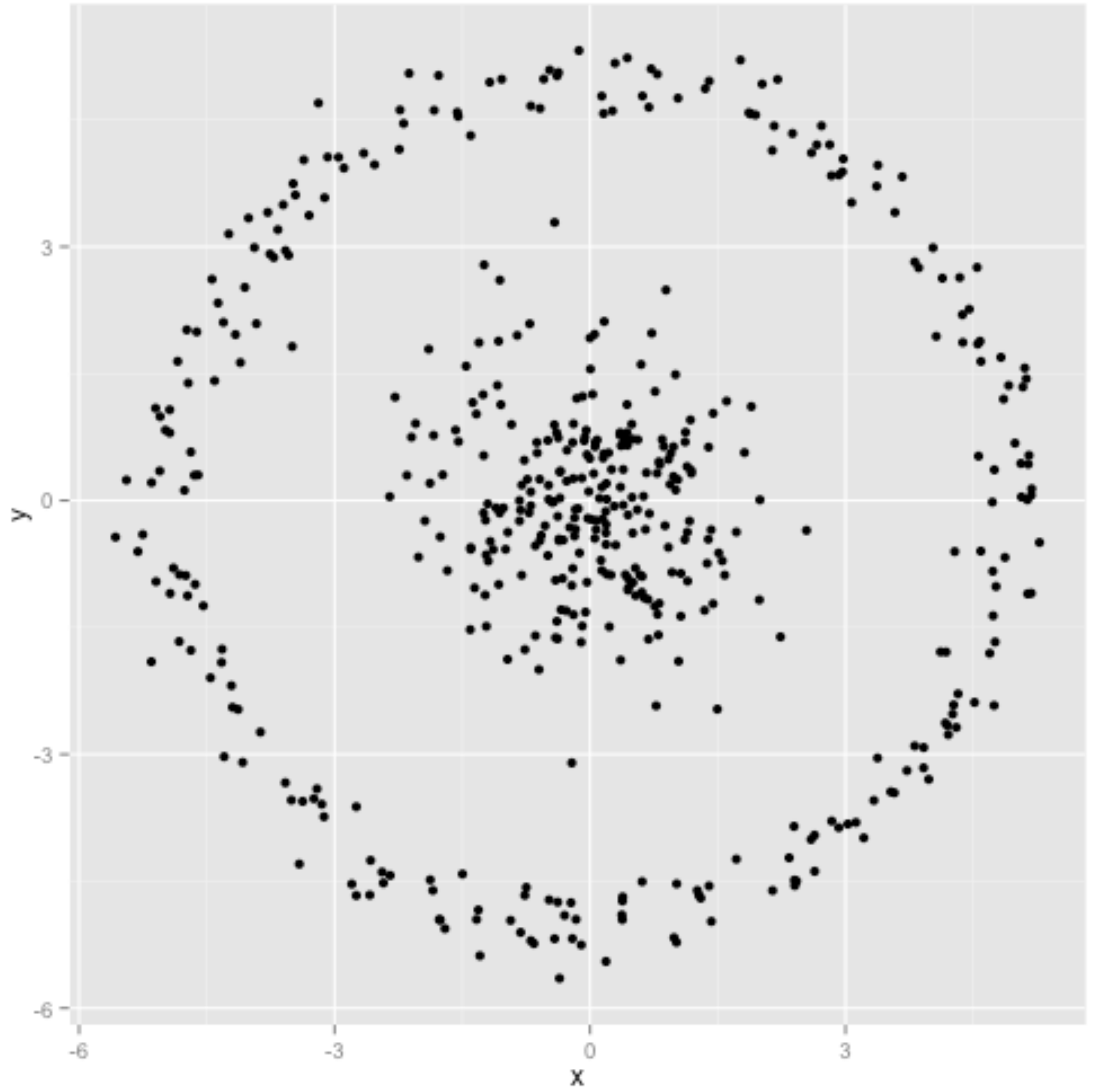
Lloyds algorithm doesn’t find a global optimum

Instead it finds a local optimum

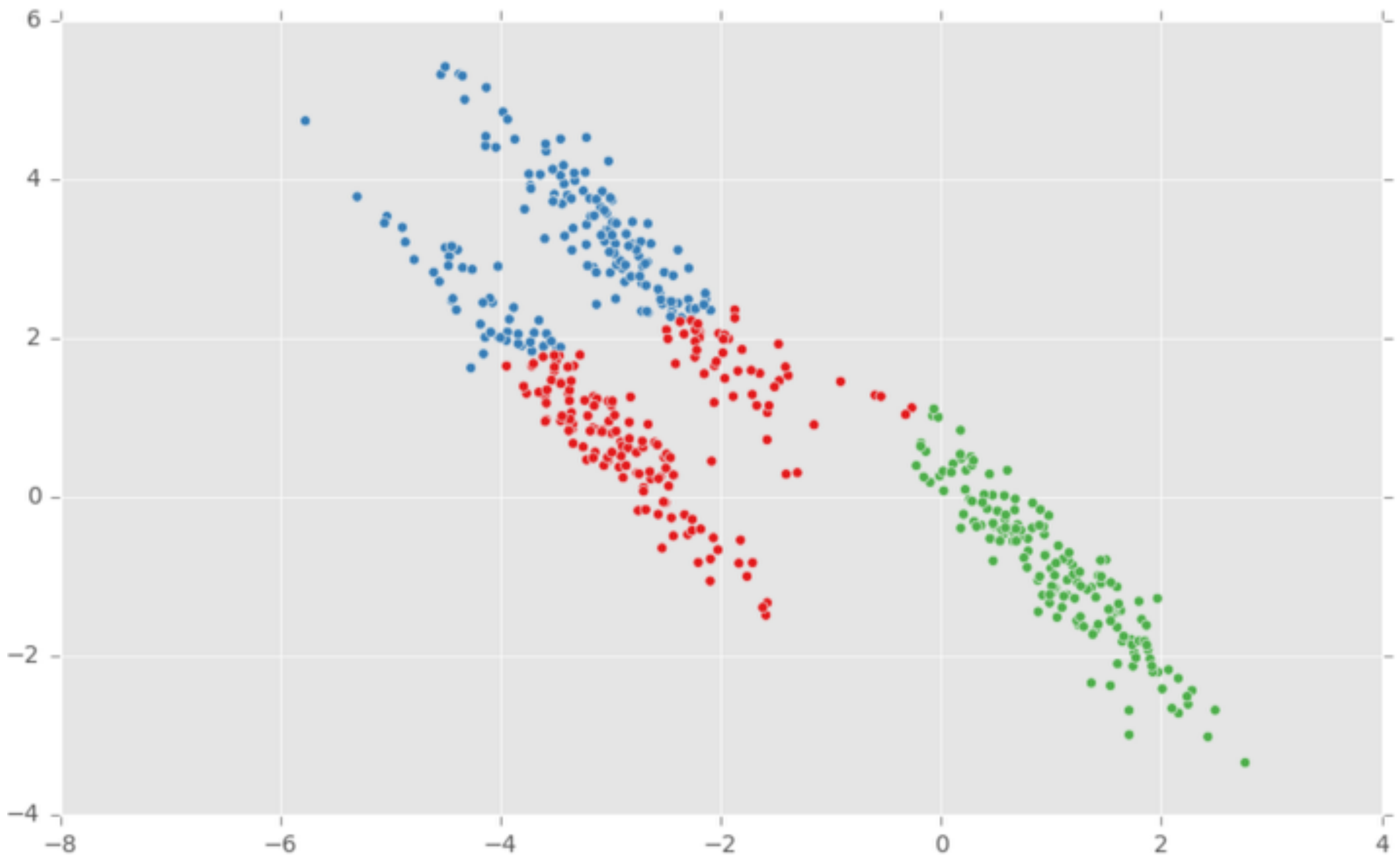
It is very fast:

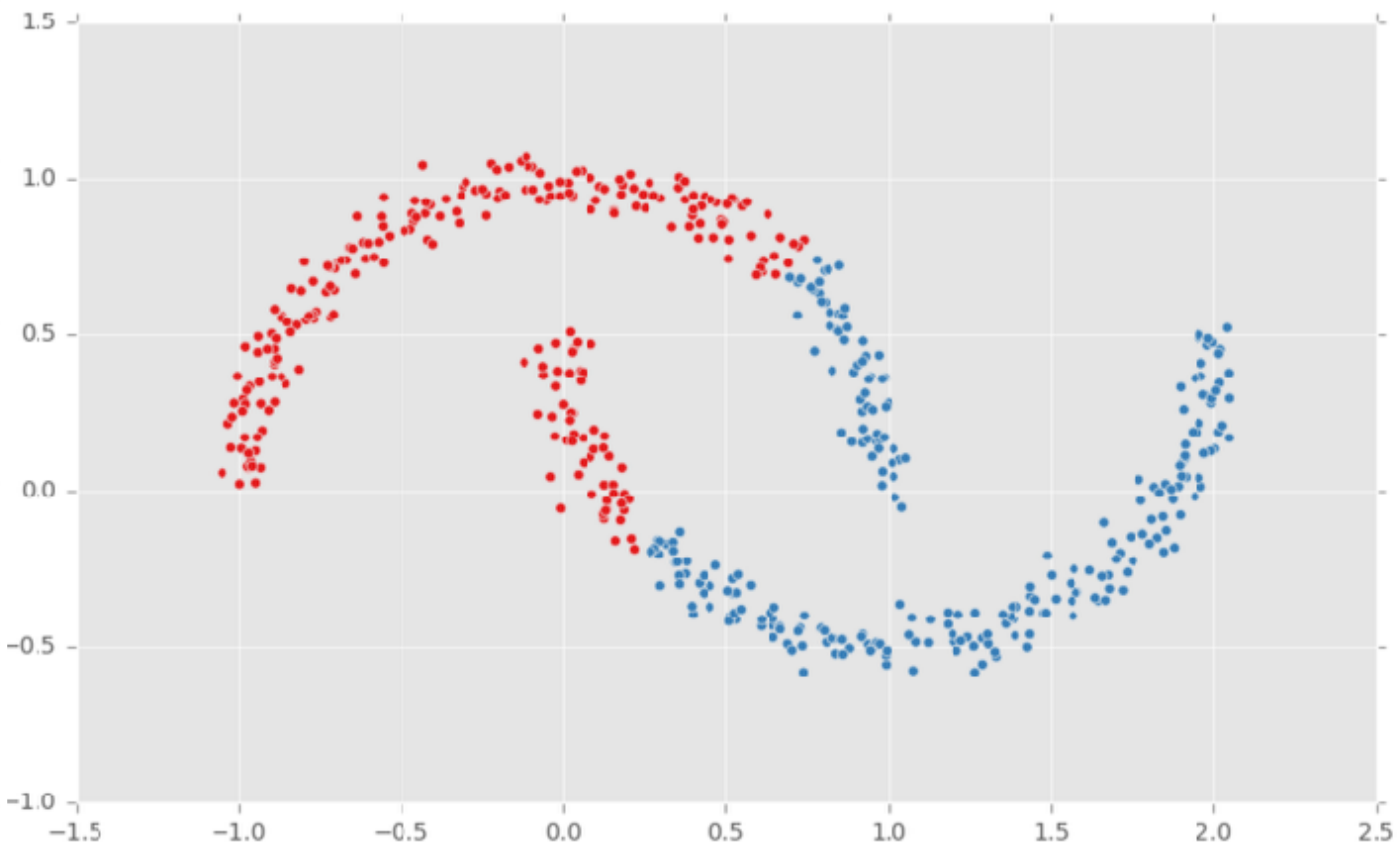
- common to run multiple times and pick the solution with the minimum inertia

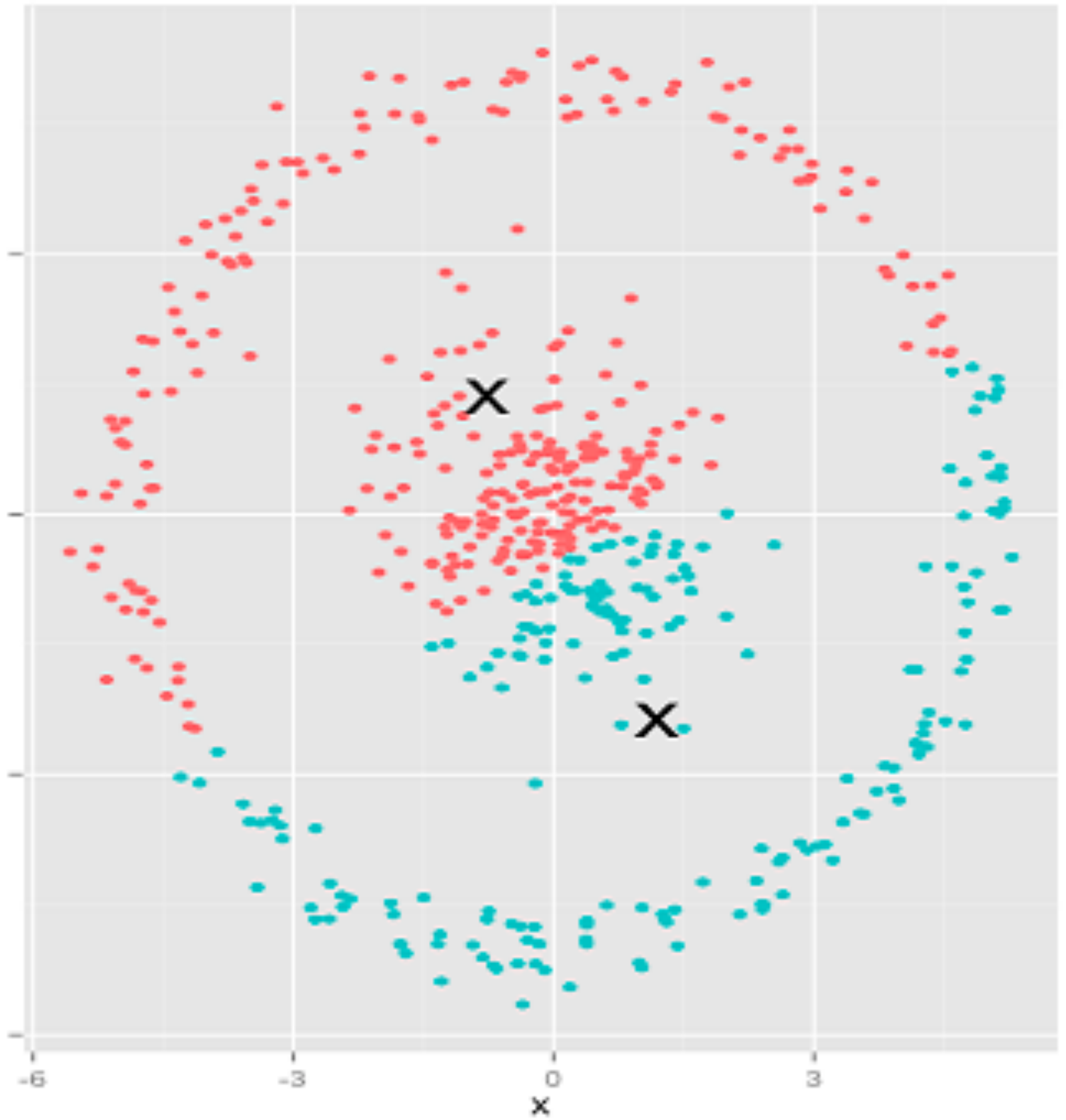
K-Means Properties



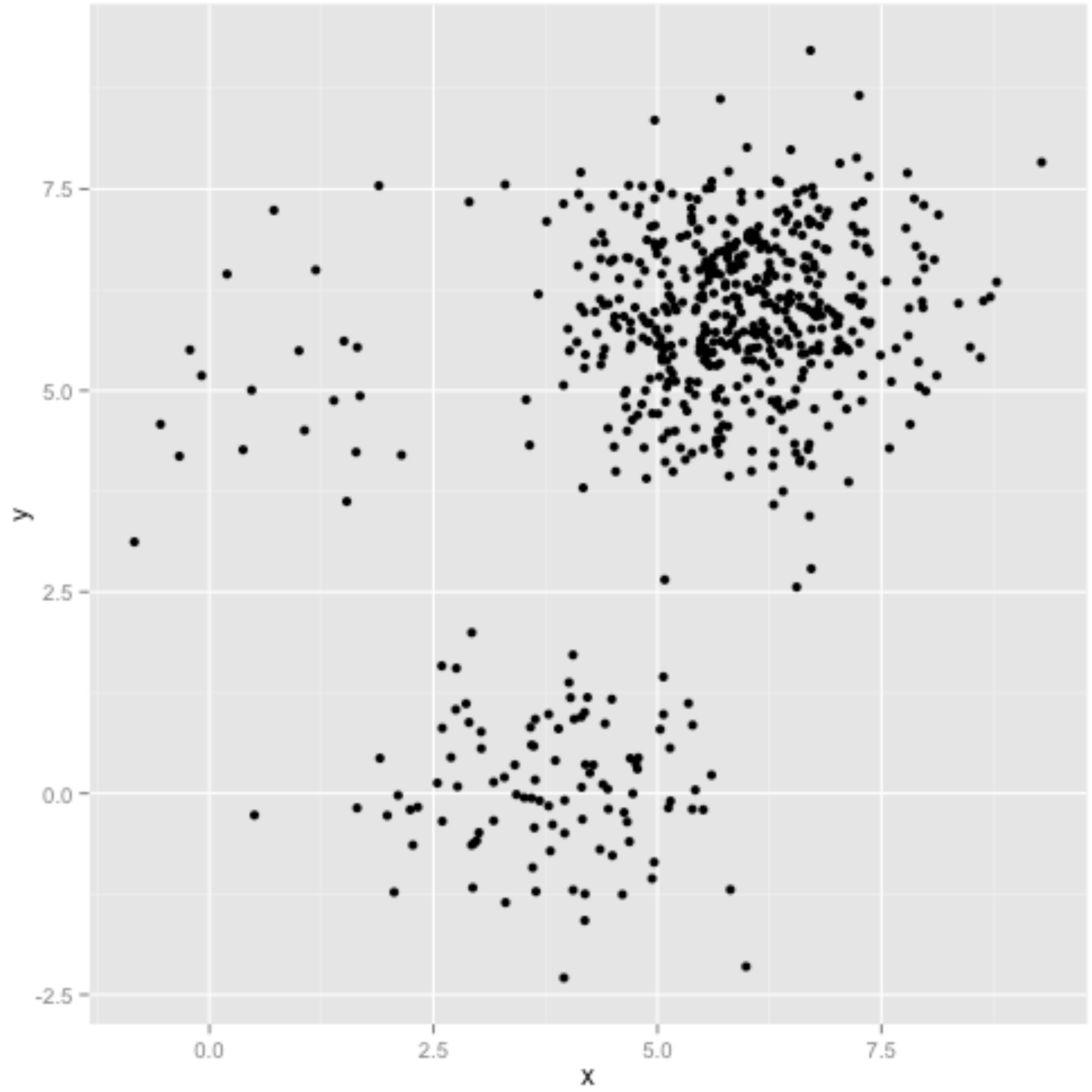
Assumptions about data: roughly “circular” clusters of equal size







K-Means Unequal Cluster Size

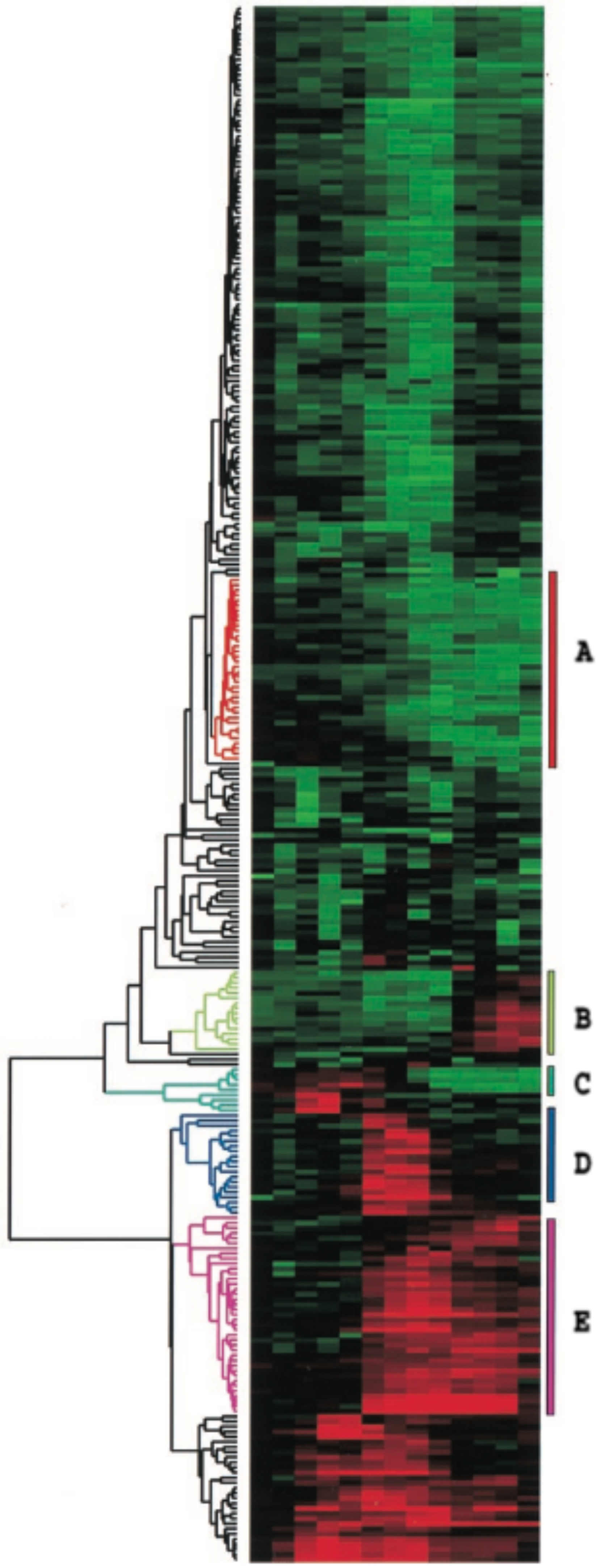


http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Hierarchical Clustering

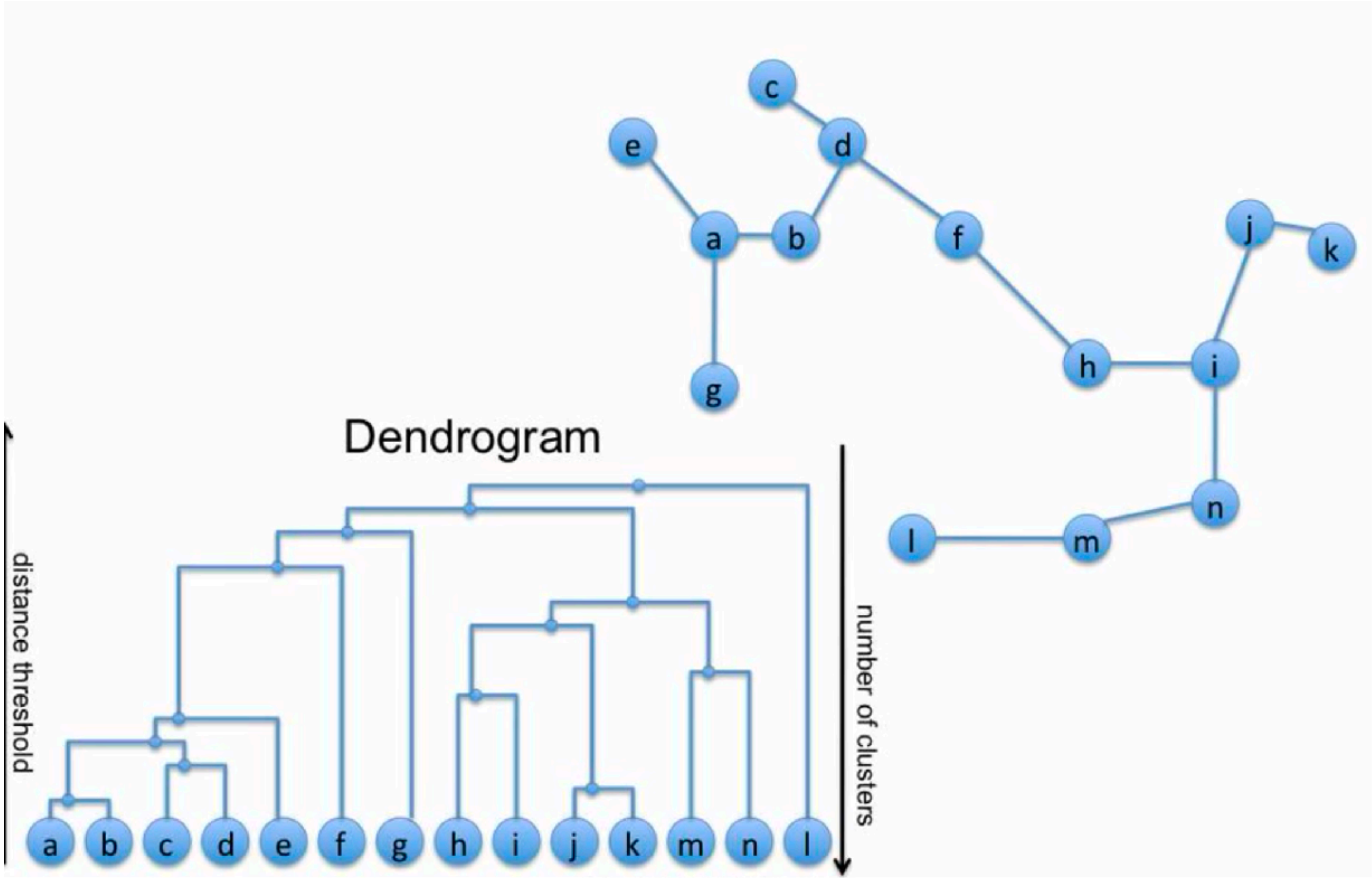
Two types:

- **agglomerative** clustering
  - start with each node as a cluster and merge
- **divisive** clustering
  - start with one cluster, and split



Agglomerative Clustering Idea

Agglomerative Clustering Idea



https://www.youtube.com/watch?v=XJ3194AmH40&list=PLBv09BD7ez_7qIbBhyQDr-LAKWUeycZtx&index=2

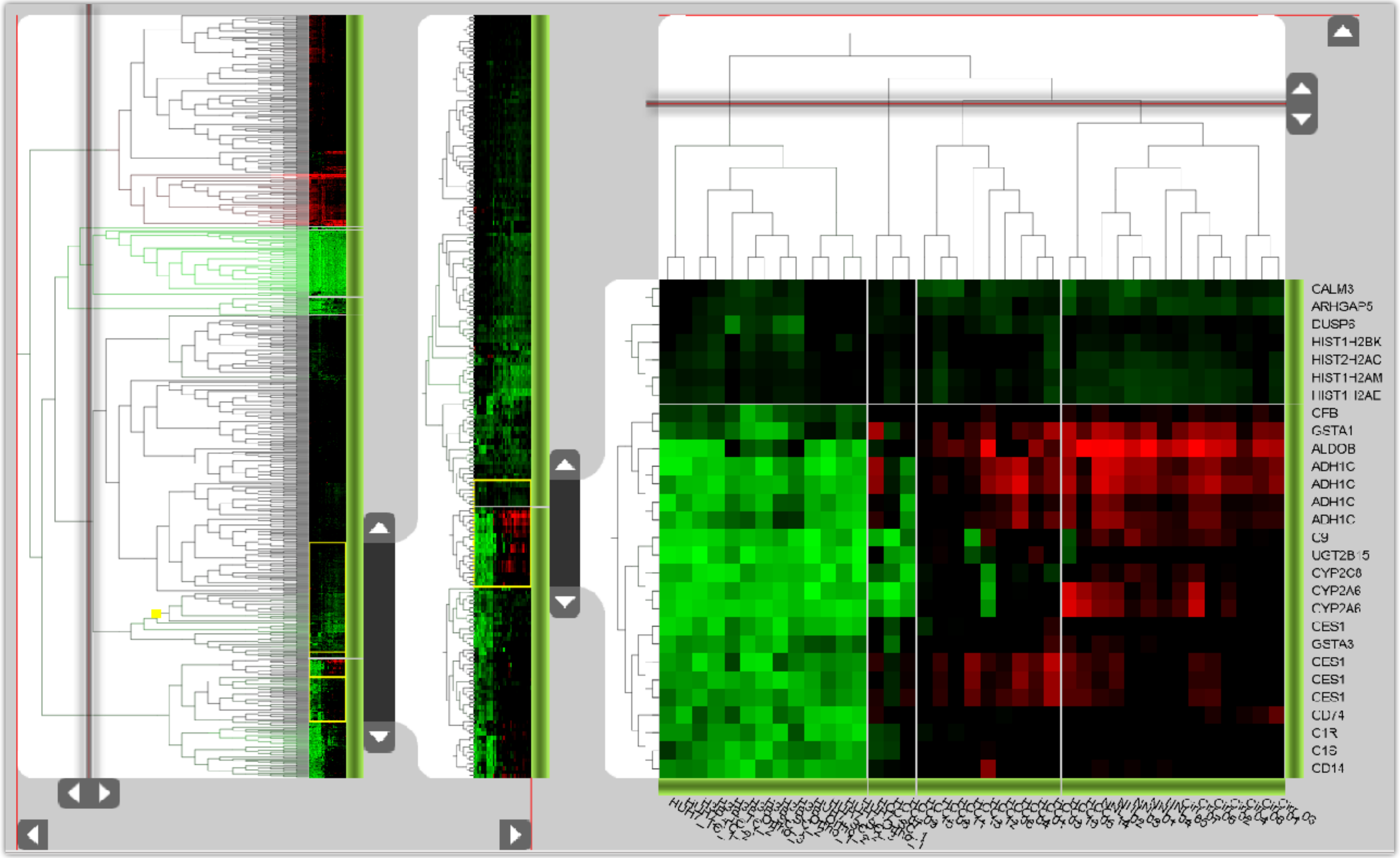
Linkage Criteria

How do you define similarity between two clusters to be merged (A and B)?

- use maximum linkage distance
- use minimum linkage distance
- use average linkage distance
- use centroid distance

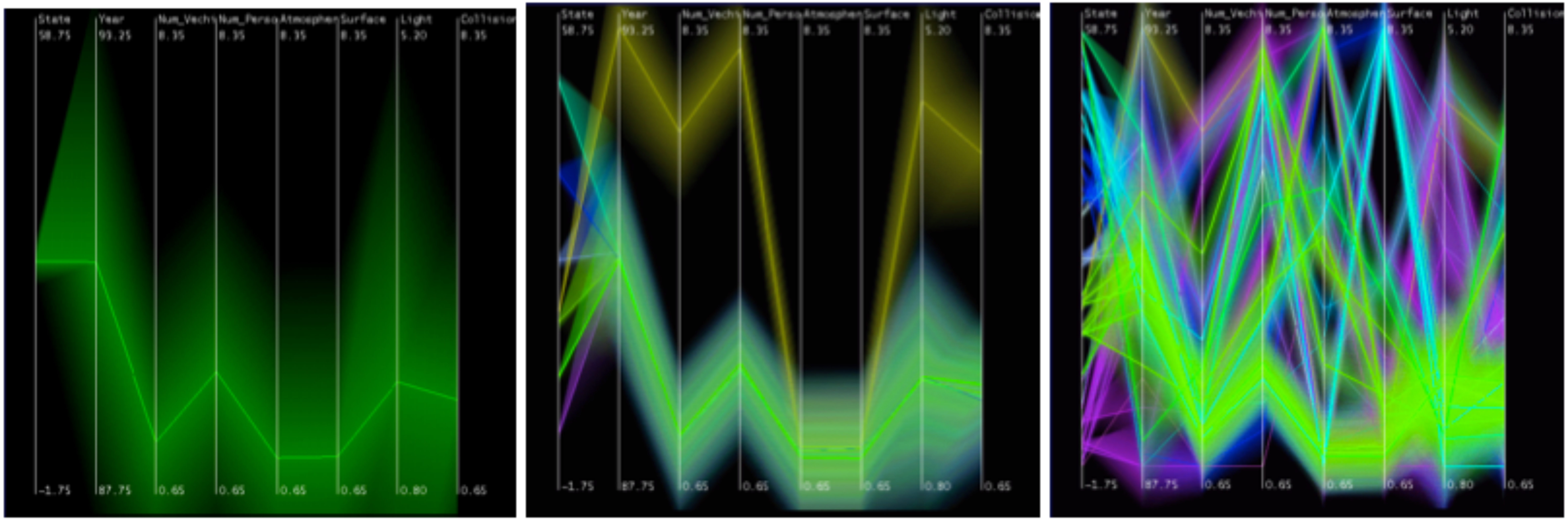
Names	Formula
Maximum or complete-linkage clustering	$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Minimum or single-linkage clustering	$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\lvert A\rvert \lvert B\rvert} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Centroid linkage clustering, or UPGMC	$\lVert c_s - c_t \rVert$ where c_s and c_t are the centroids of clusters *s* and *t*, respectively.

F+C Approach, with Dendrograms



[Lex, PacificVis 2010]

Hierarchical Parallel Coordinates



Fua 1999

Attribute aggregation

1) group attributes and compute
a similarity score across the set

**2) dimensionality reduction,
to preserve meaningful structure**

Dimensionality Reduction

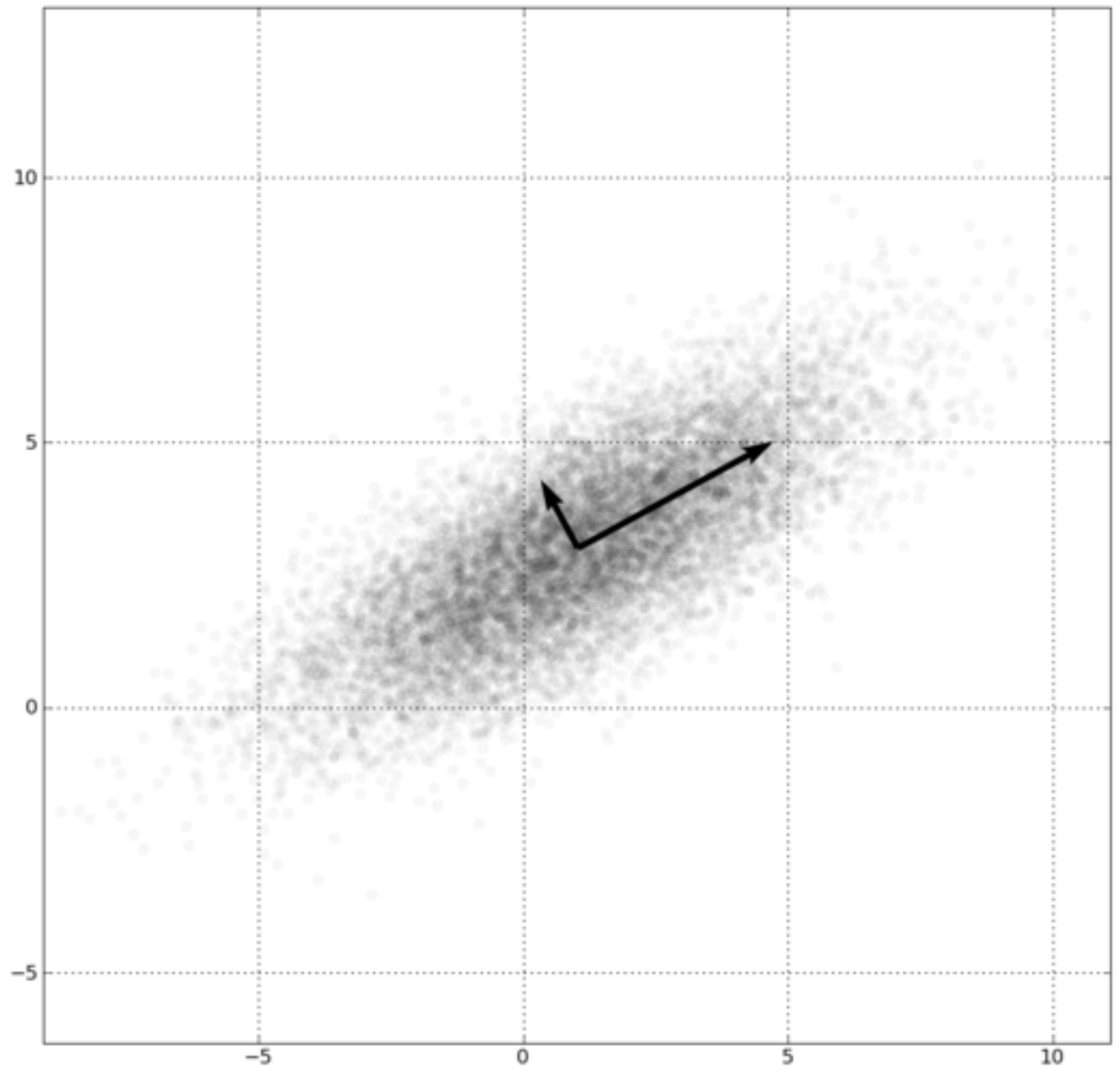
Reduce high dimensional to lower dimensional space

Preserve as much of variation as possible

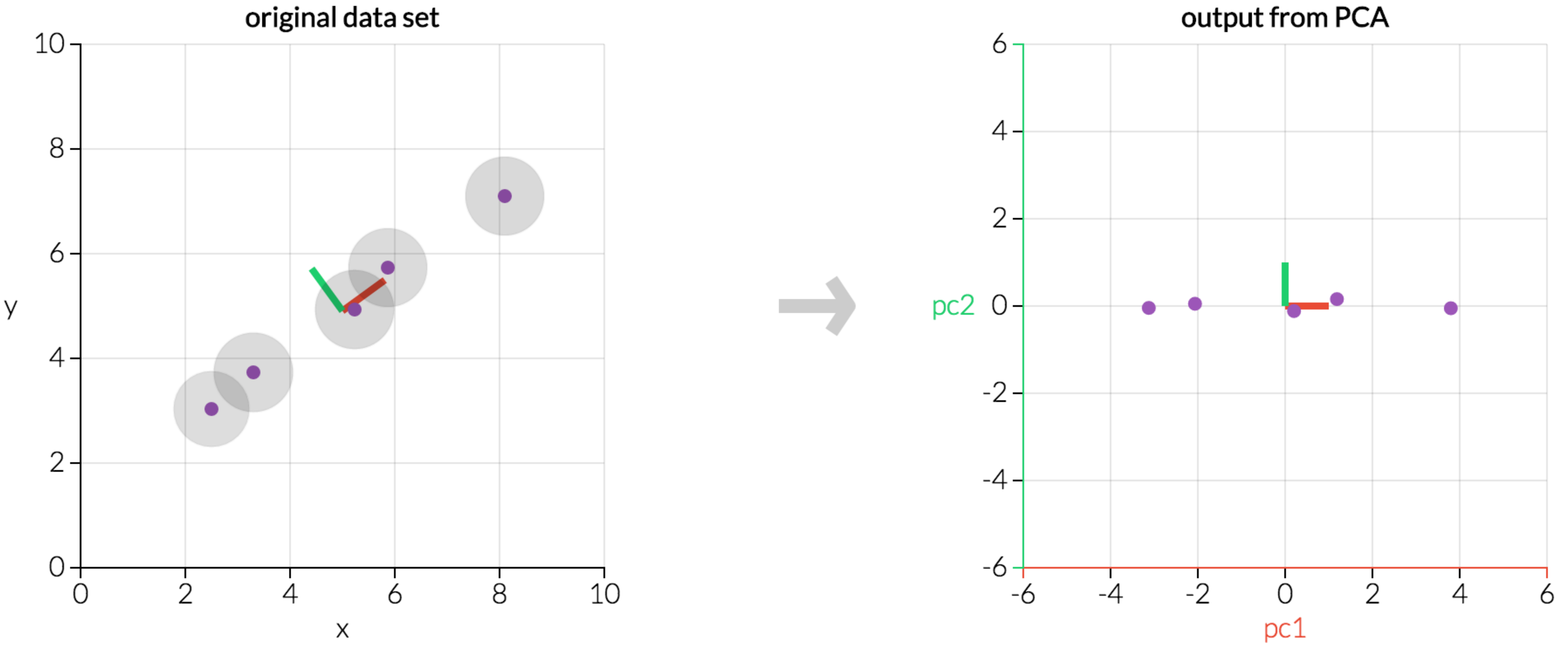
Plot lower dimensional space

Principal Component Analysis (PCA)

linear mapping, by order of variance



PCA



original data set
10
8
6
4
2
0
y
0
2
4
6
8
10
x
output from PCA
6
4
2
pc2
0
-2
-4
-6
-6
-4
-2
0
2
4
6
pc1

Multidimensional Scaling

Nonlinear, better suited for some DS

Multiple approaches

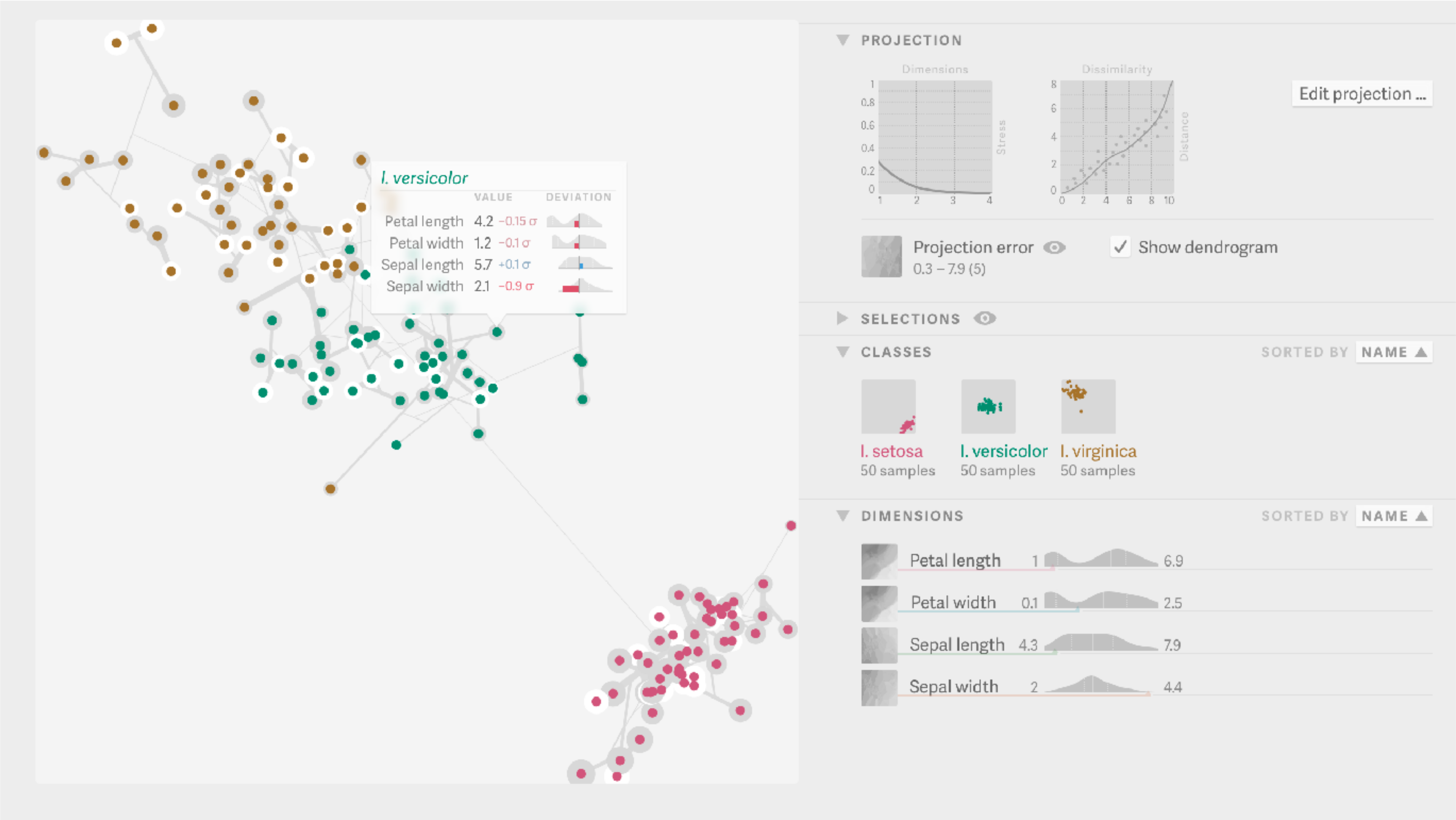
Works based on projecting a similarity matrix

- How do you compute similarity?
- How do you project the points?

Popular for text analysis

[Doerk 2011]

Probing Projections



http://julianstahnke.com/probing-projections/