Data Visualization (CSCI 627/490)
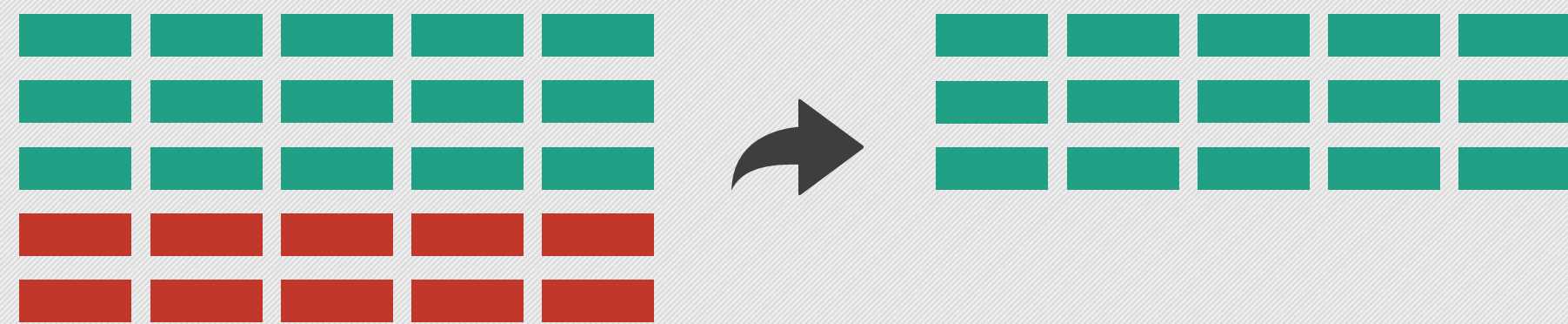
Aggregation & Focus+Context

Dr. David Koop
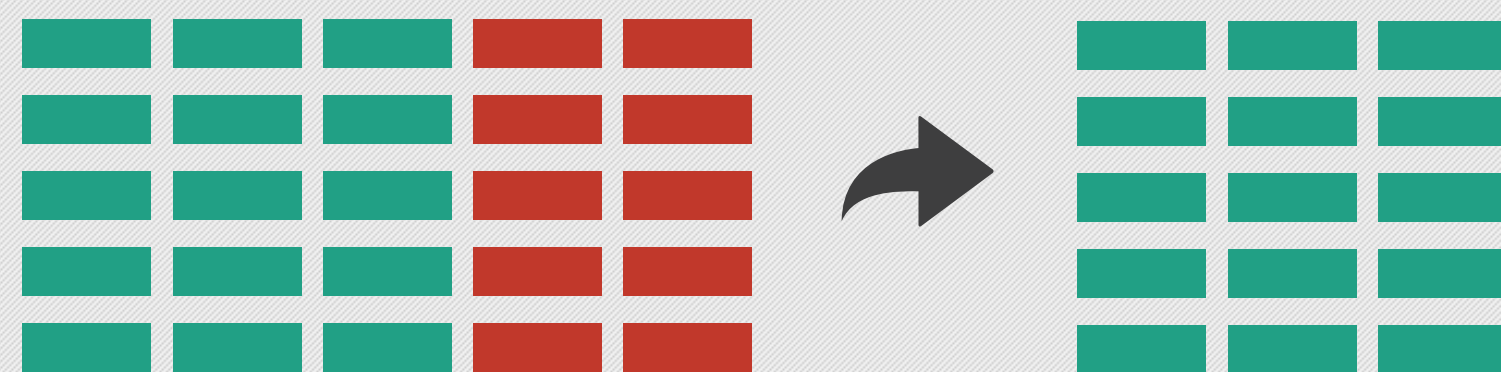
Northern Illinois University

Overview: Reducing Items & Attributes
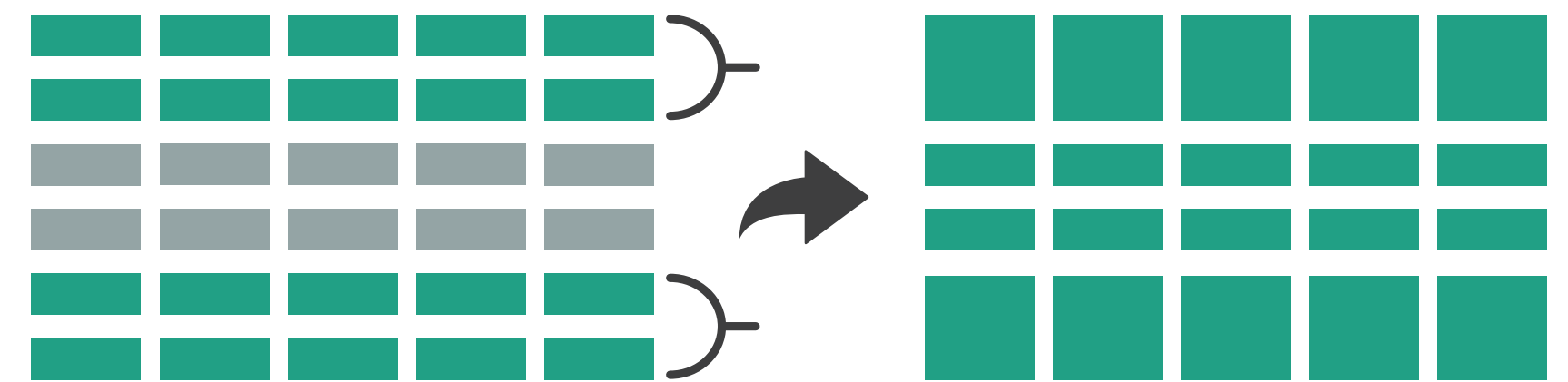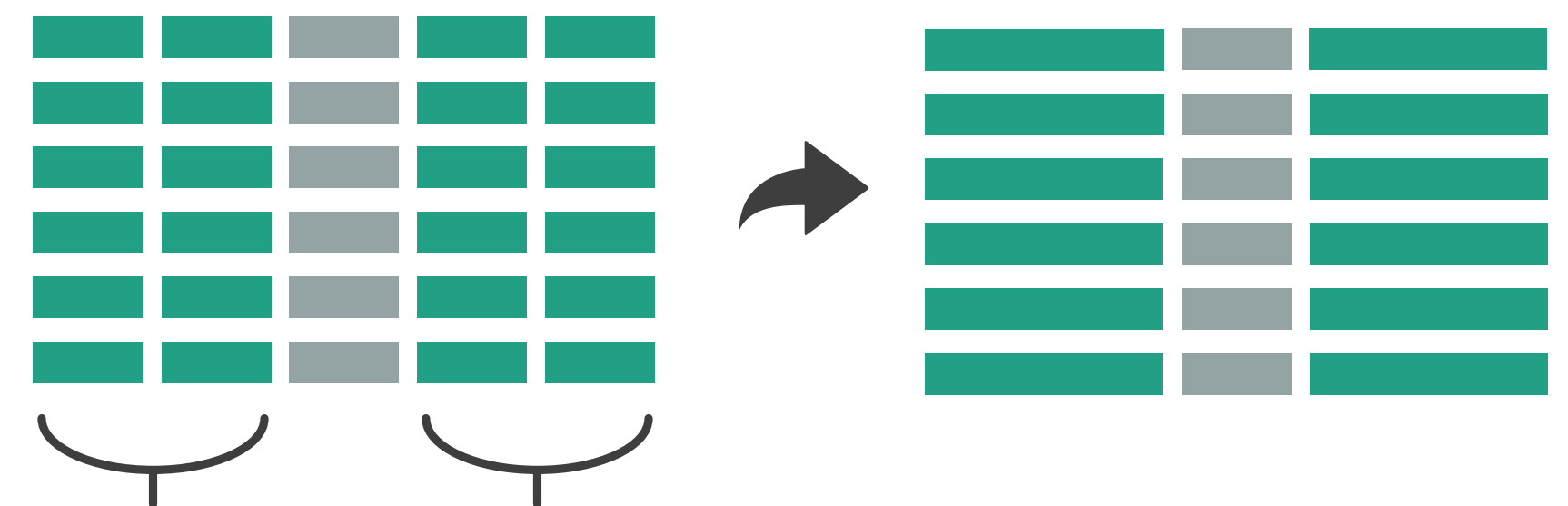
⊕ Filter

→ Items

→ Attributes

⊕ Aggregate

→ Items

→ Attributes

[Munzner (ill. Maguire), 2014]

Item Filtering on Maps

FIND A RESTAURANT   FIND A LOCATION

FILTER

All grades   All violations   All cuisines

Gracie's Cafe
Grade   **Grade pending**
Violation points   **27**
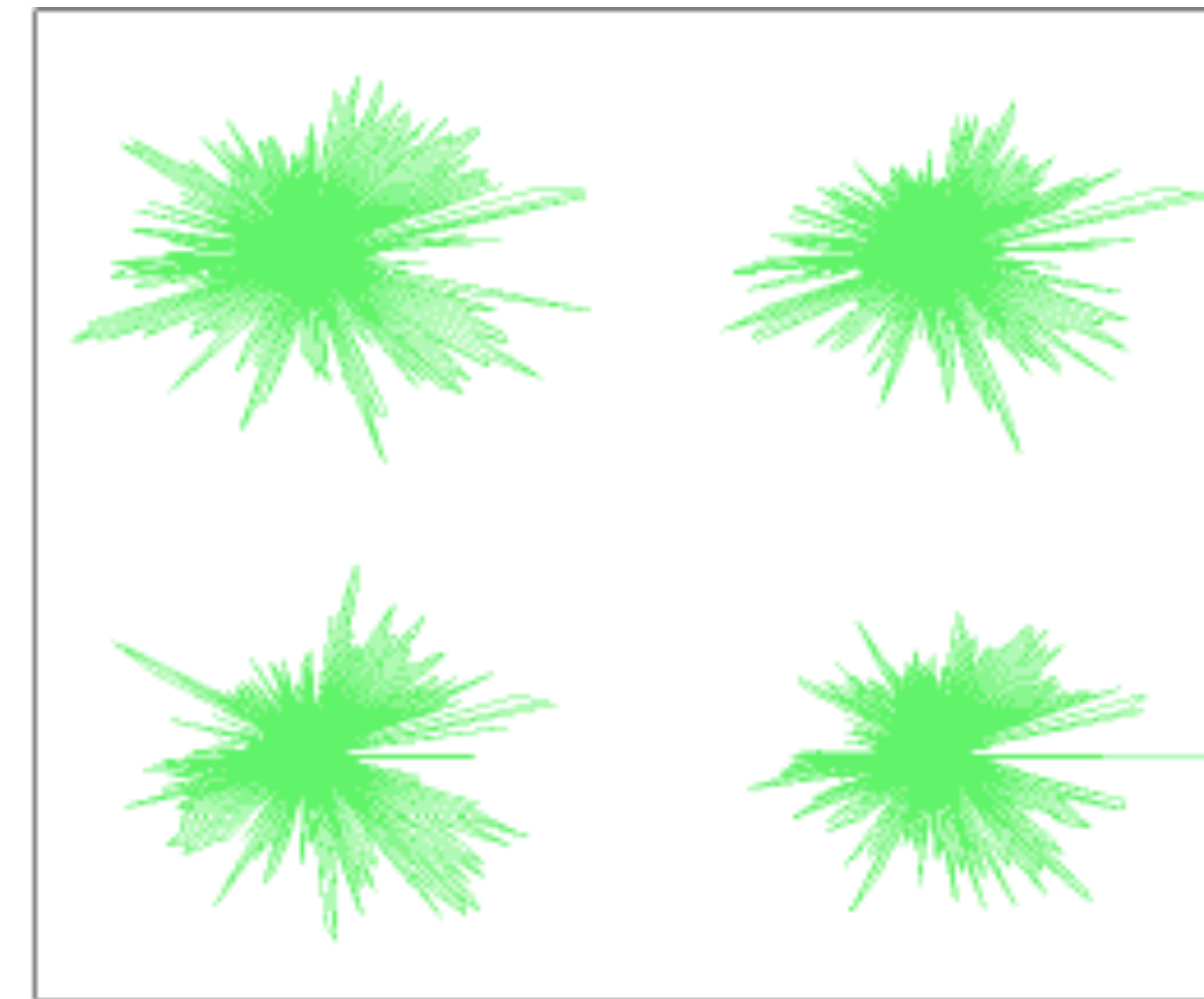*Click for details*
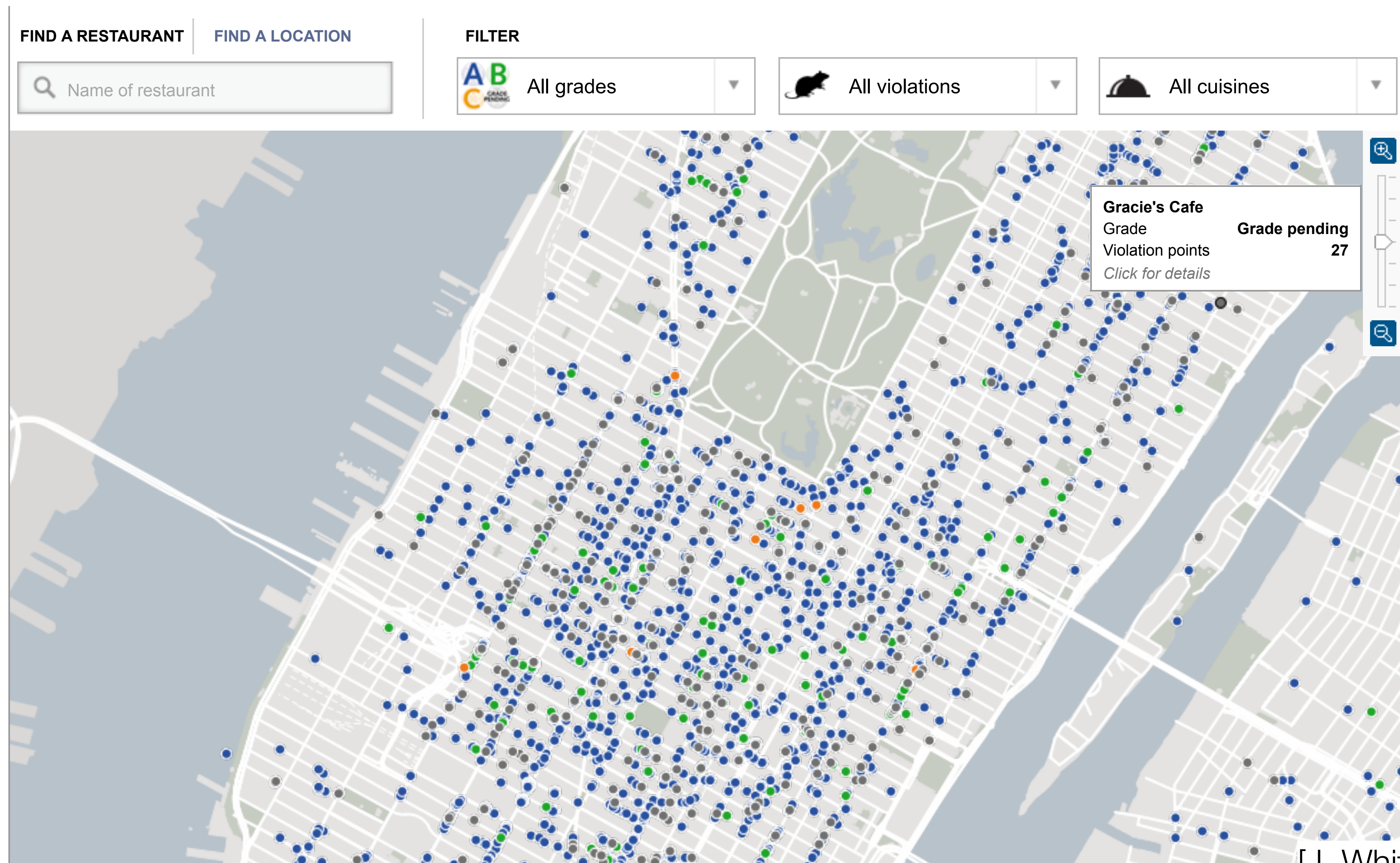
[J. White, New York Times]

Attribute Filtering on Star Plots
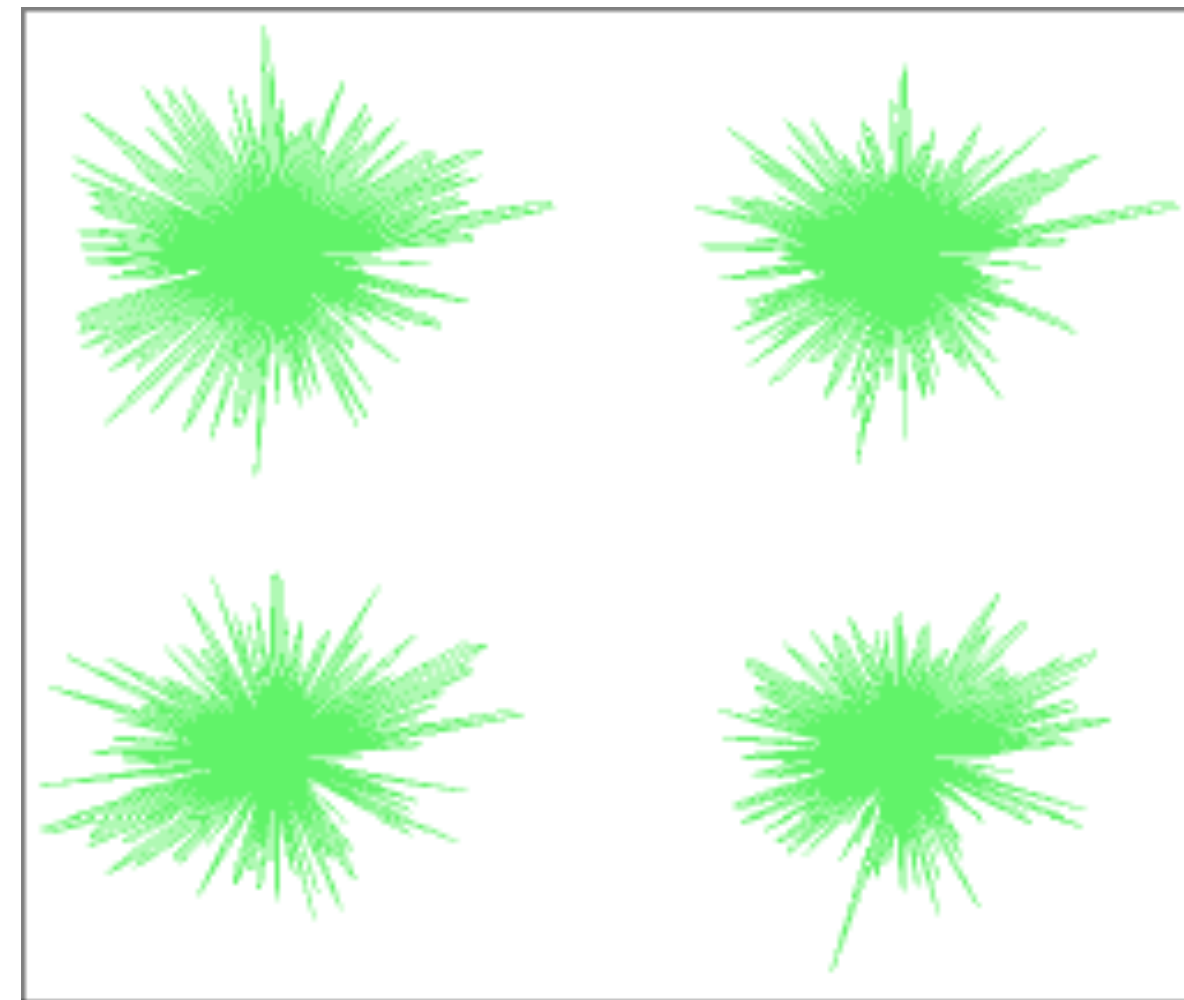
(a)

(b)
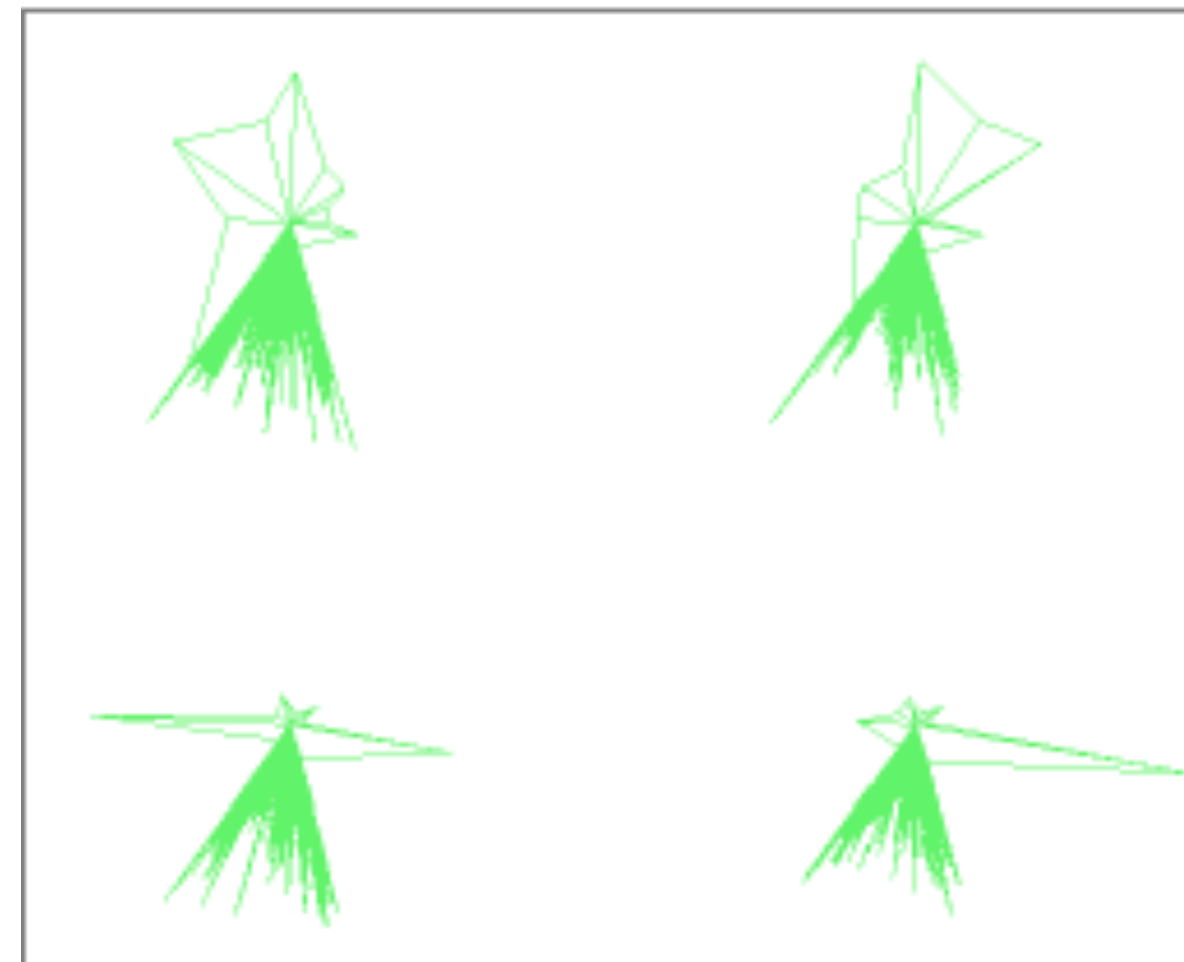
(c)
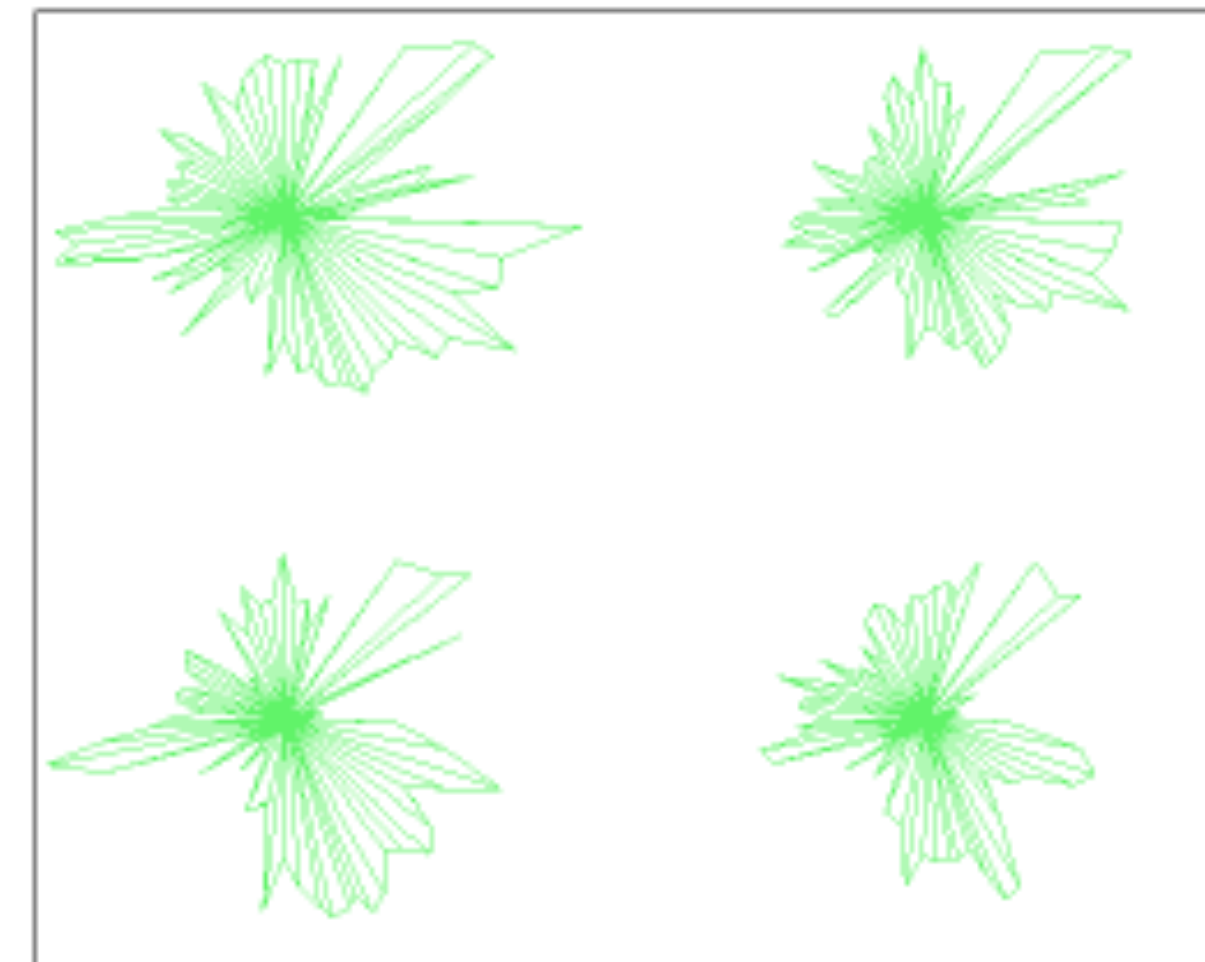
(d)

[Yang et al., 2003]

Project Design

- Feedback:
  - Data Manipulation?
  - Questions lead, not technique!
  - Be creative! (interaction too) https://xeno.graphics
- Work on turning your visualization ideas into designs
- Turn in:
  - Two Design Sketches (like sheets 2-4 from 5 Sheet Design)
  - One Bad Design Sketch (like sheets 2-4: here, justify why bad)
  - Progress on Implementation
- Due Today

Assignment 5

- Focus on Multiple Views and Interaction
- Soon…

Aggregation

Aggregation

- Usually involves **derived** attributes

- Examples: mean, median, mode, min, max, count, sum

- Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

| I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Aggregation

- Usually involves **derived** attributes
- Examples: mean, median, mode, min, max, count, sum
- Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

Mean of x	9
Variance of x	11
Mean of y	7.50
Variance of y	4.122
Correlation	0.816

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

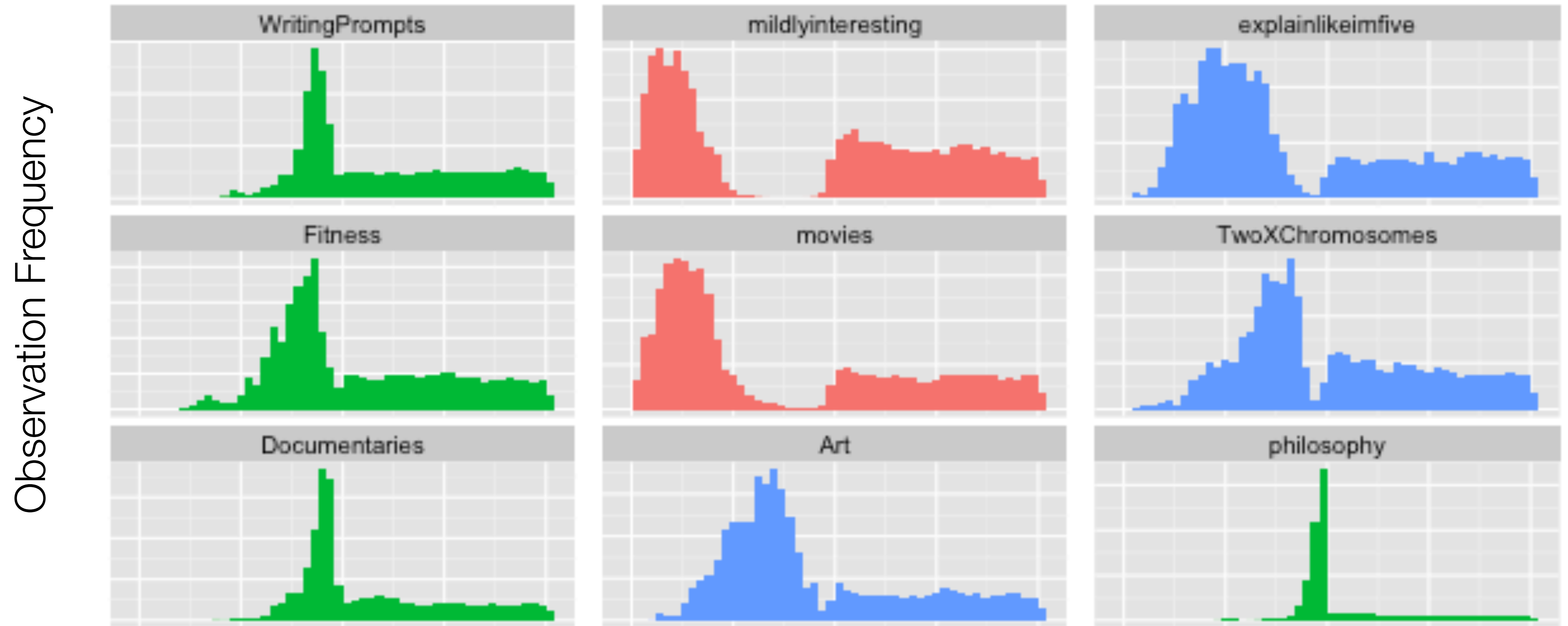# Anscombe's Quartet
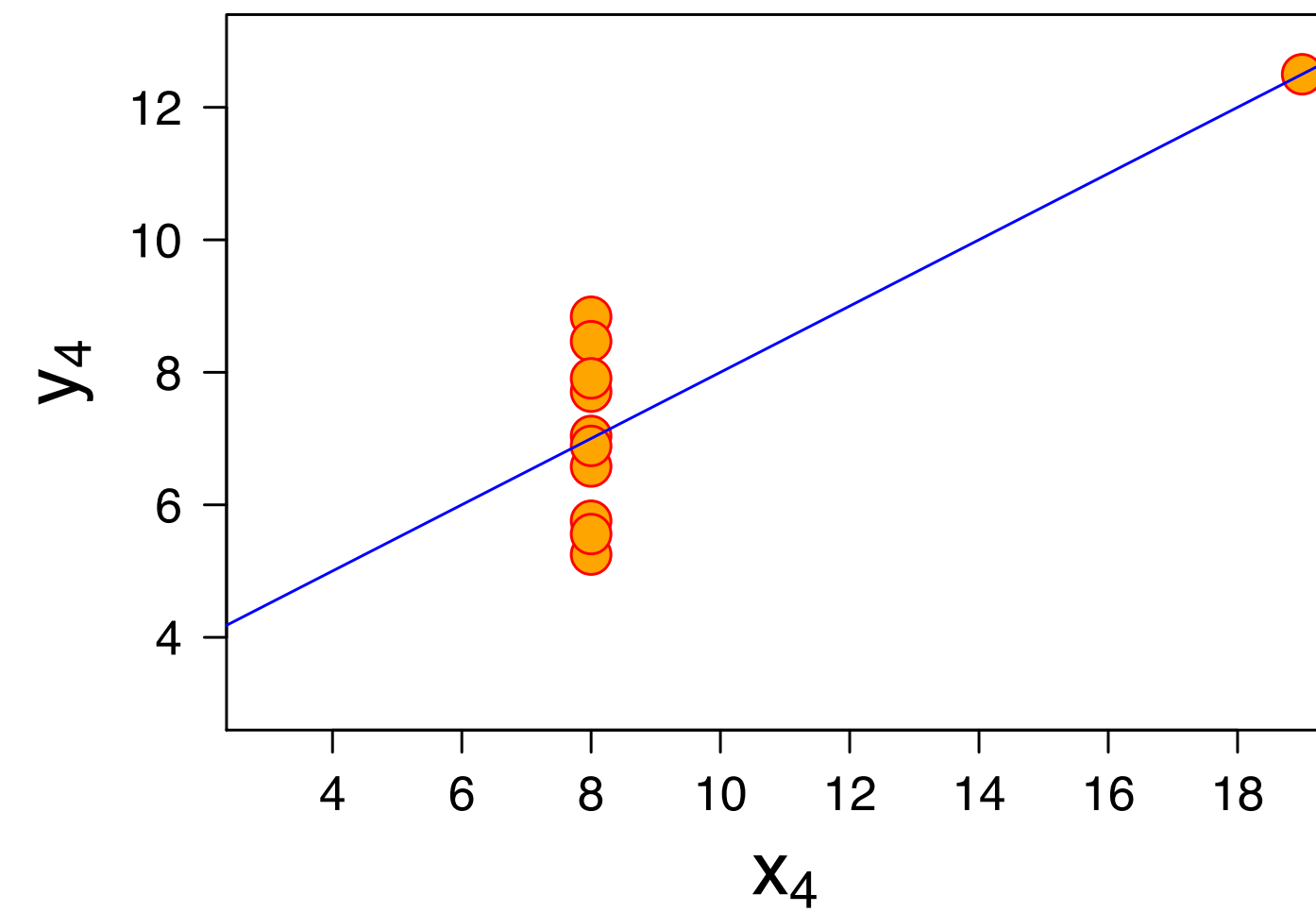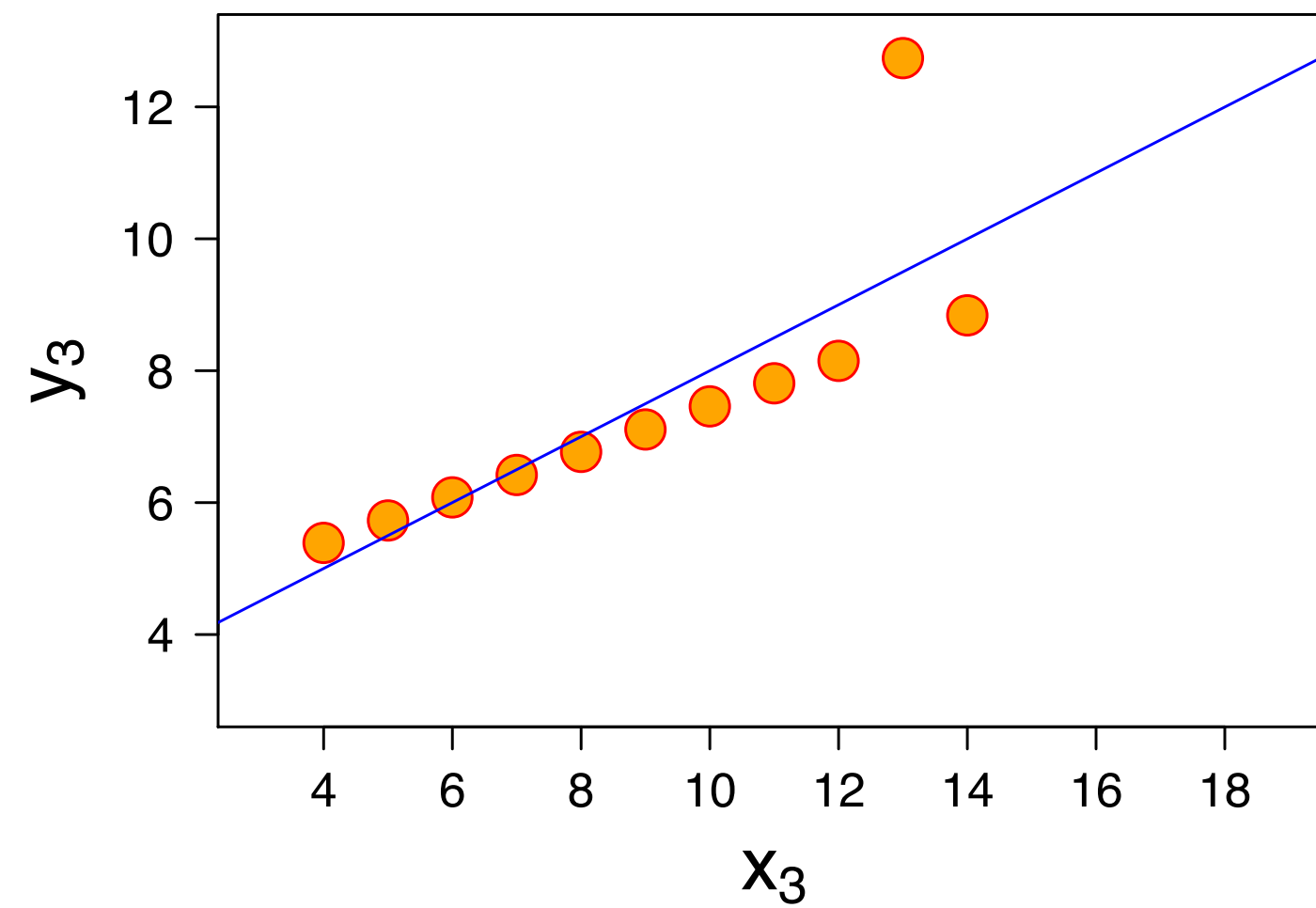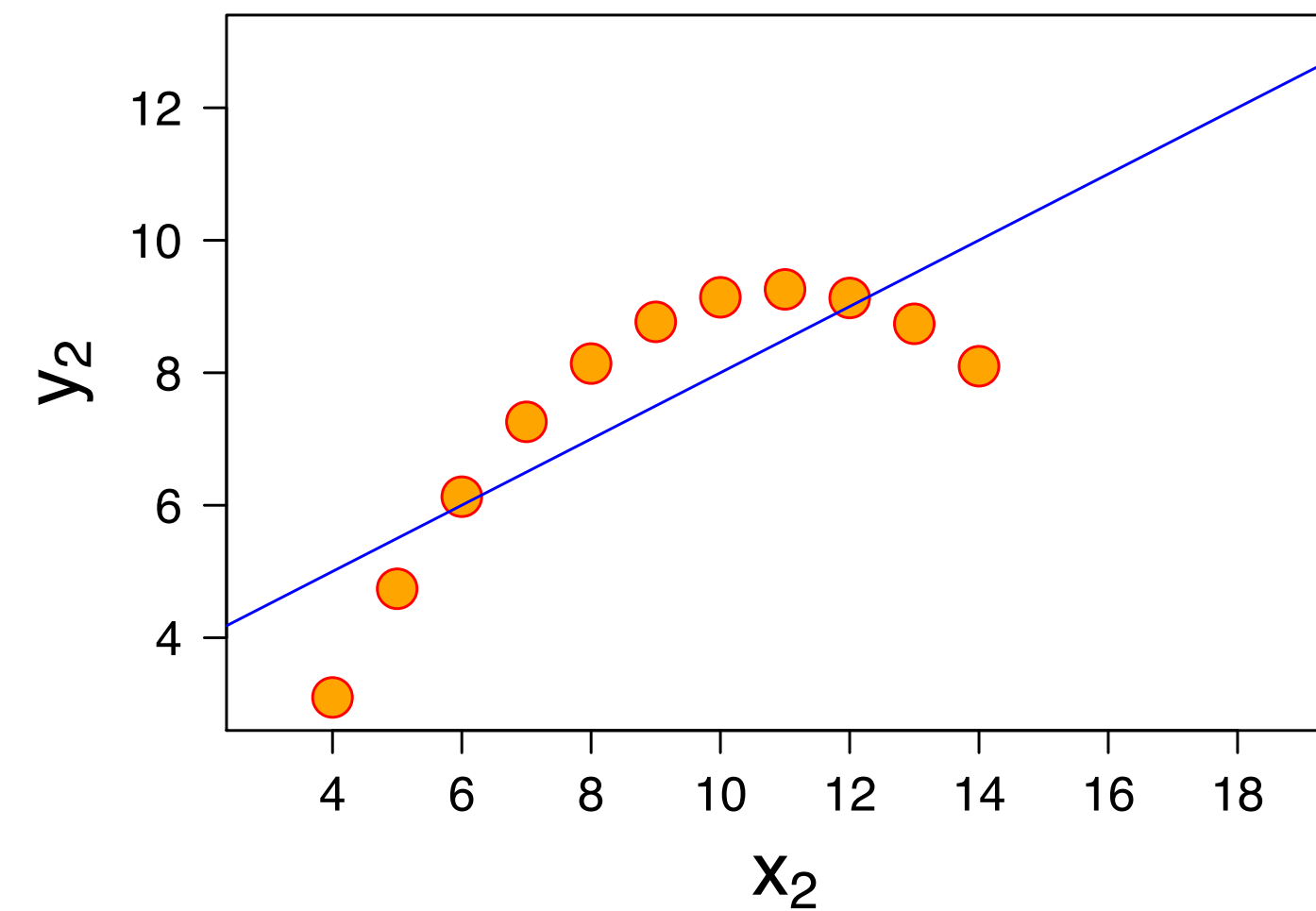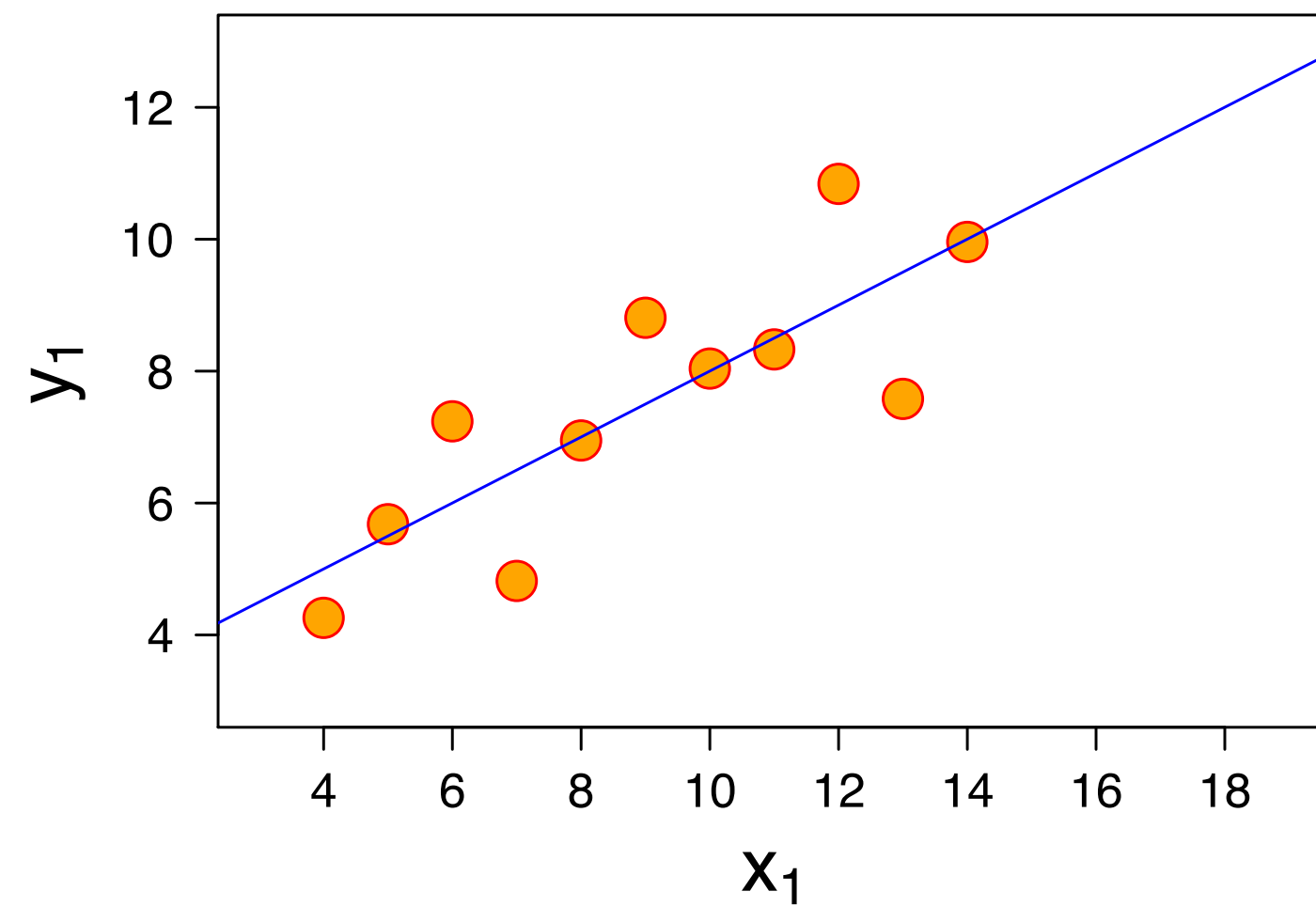
[F. J. Anscombe]

Aggregation: Histograms
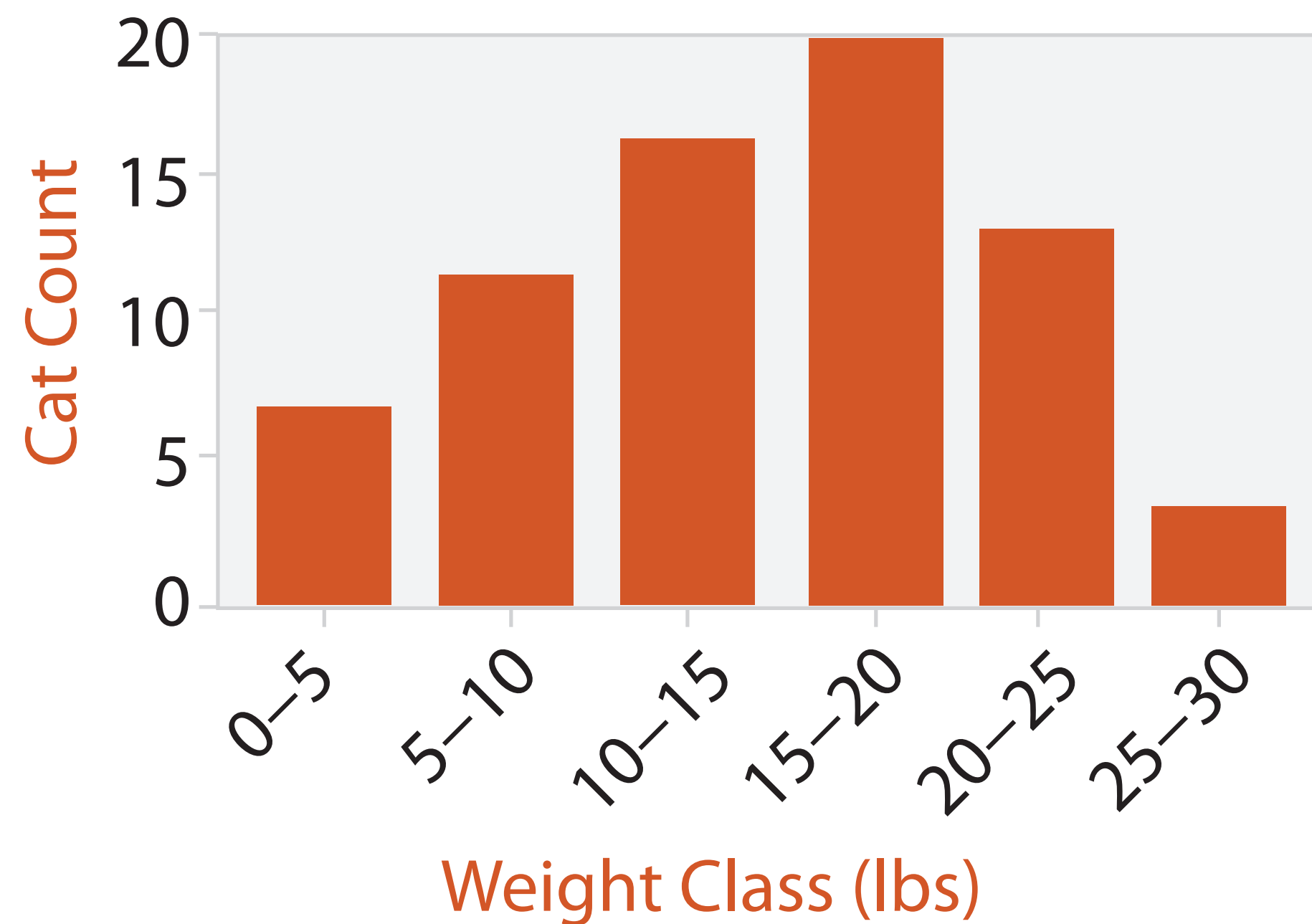
- Very similar to bar charts
- Often shown without space between (continuity)
- Choice of number of bins
  - Important!
  - Viewers may infer different trends based on the layout
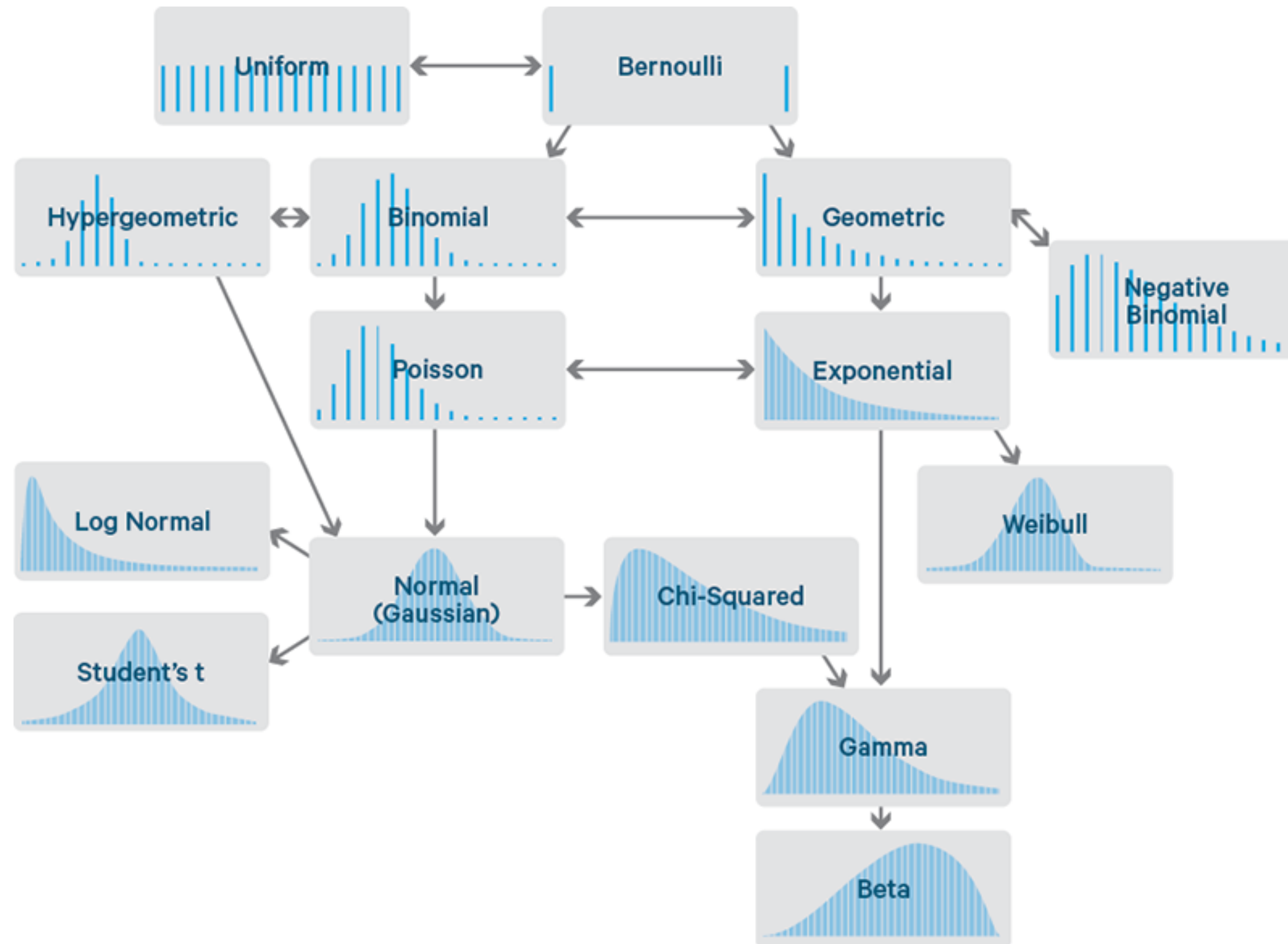
Aggregation: Histograms

Observation Frequency

Observed ranks of posts by subreddit

["The reddit Front Page is Not a Meritocracy", T. W. Schneider]

Common Distributions

Binning Scatterplots

- At some point, cannot see density

- Blobs on top of blobs

- 2D Histogram is a histogram in 2D encoded using color instead of height

- Each region is aggregated
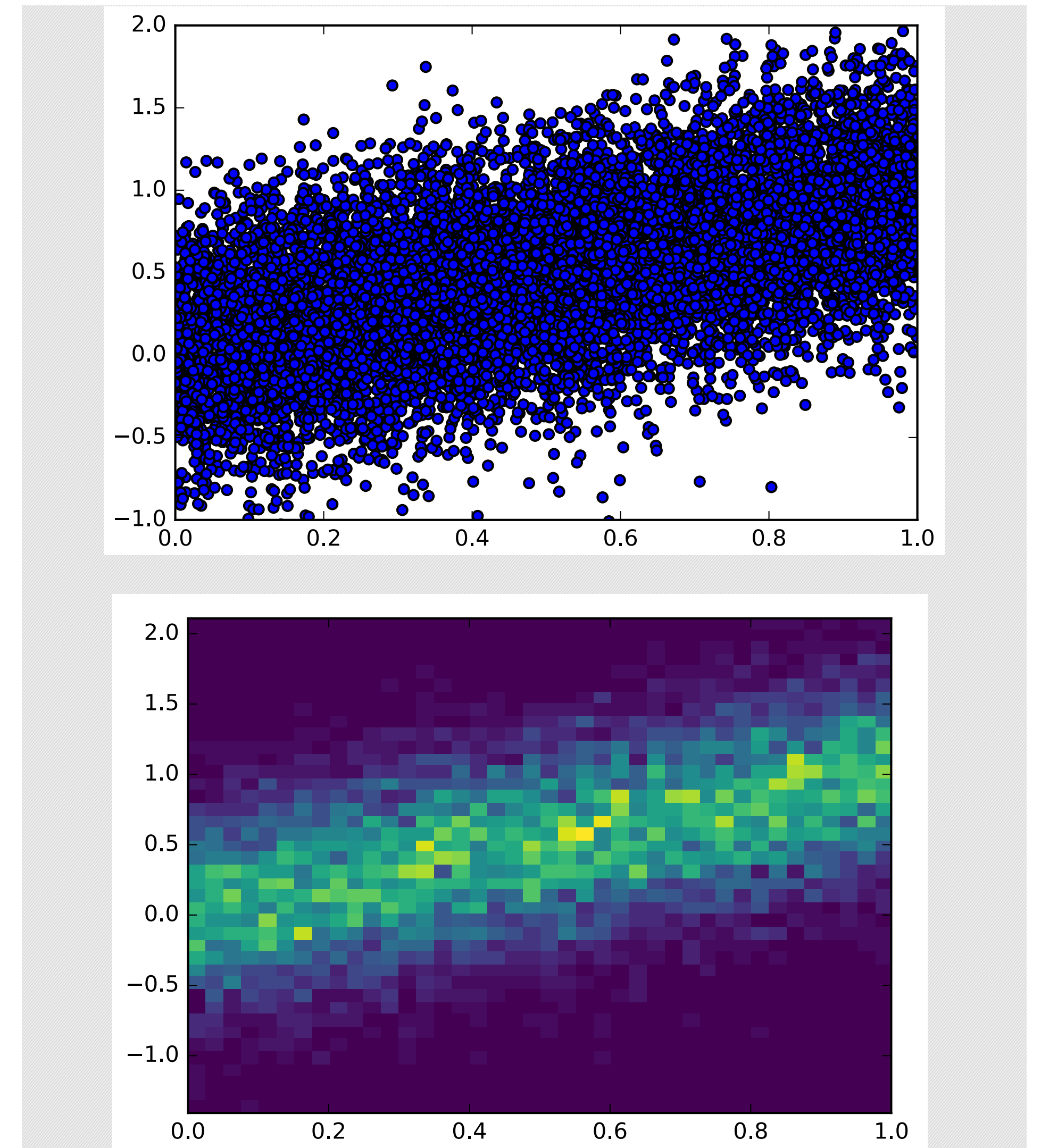
Binning

- Hexagonal bins are more circular

- Distance to the edge is not as variable

- More efficient aggregation around the center of the bin
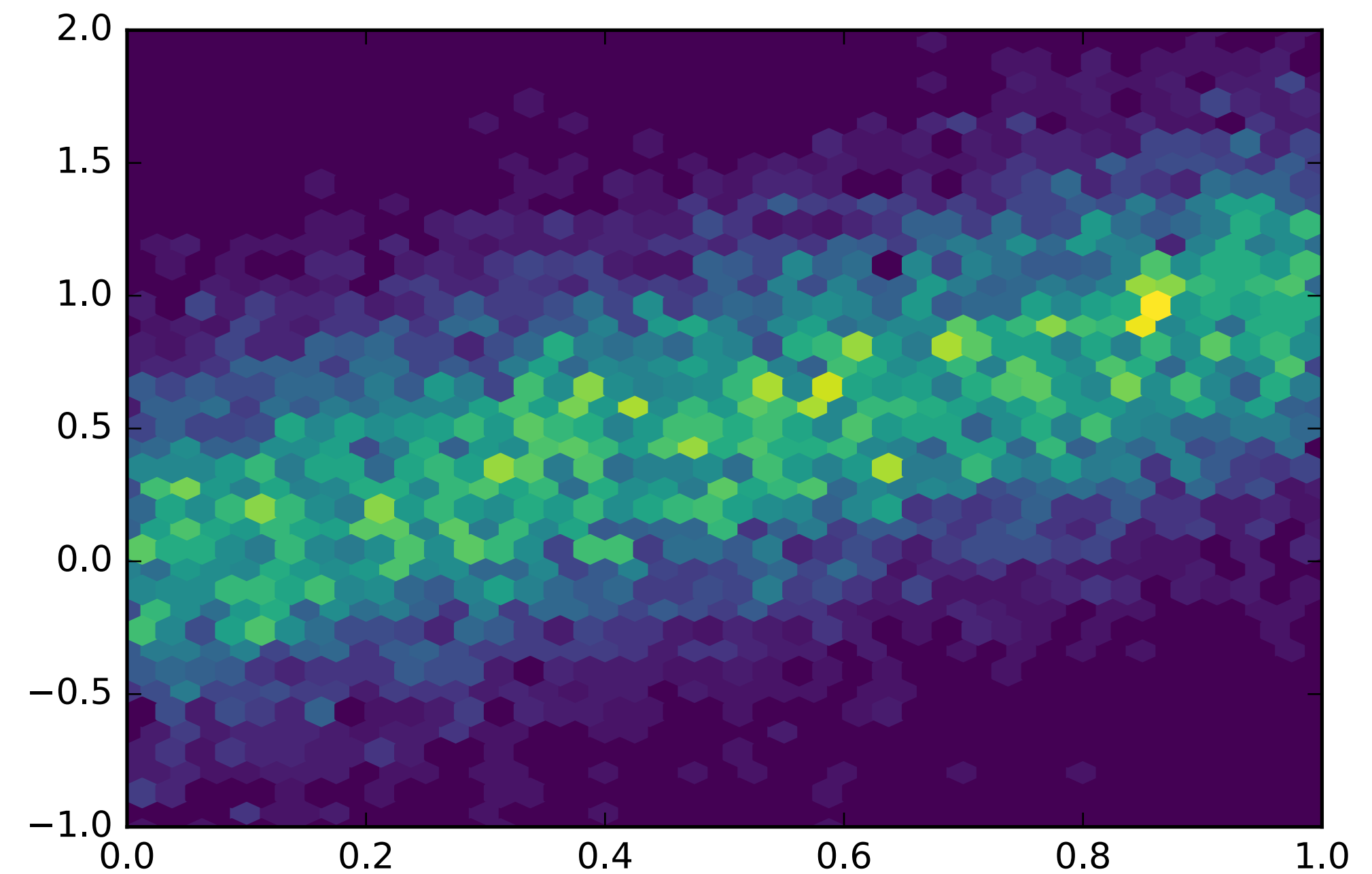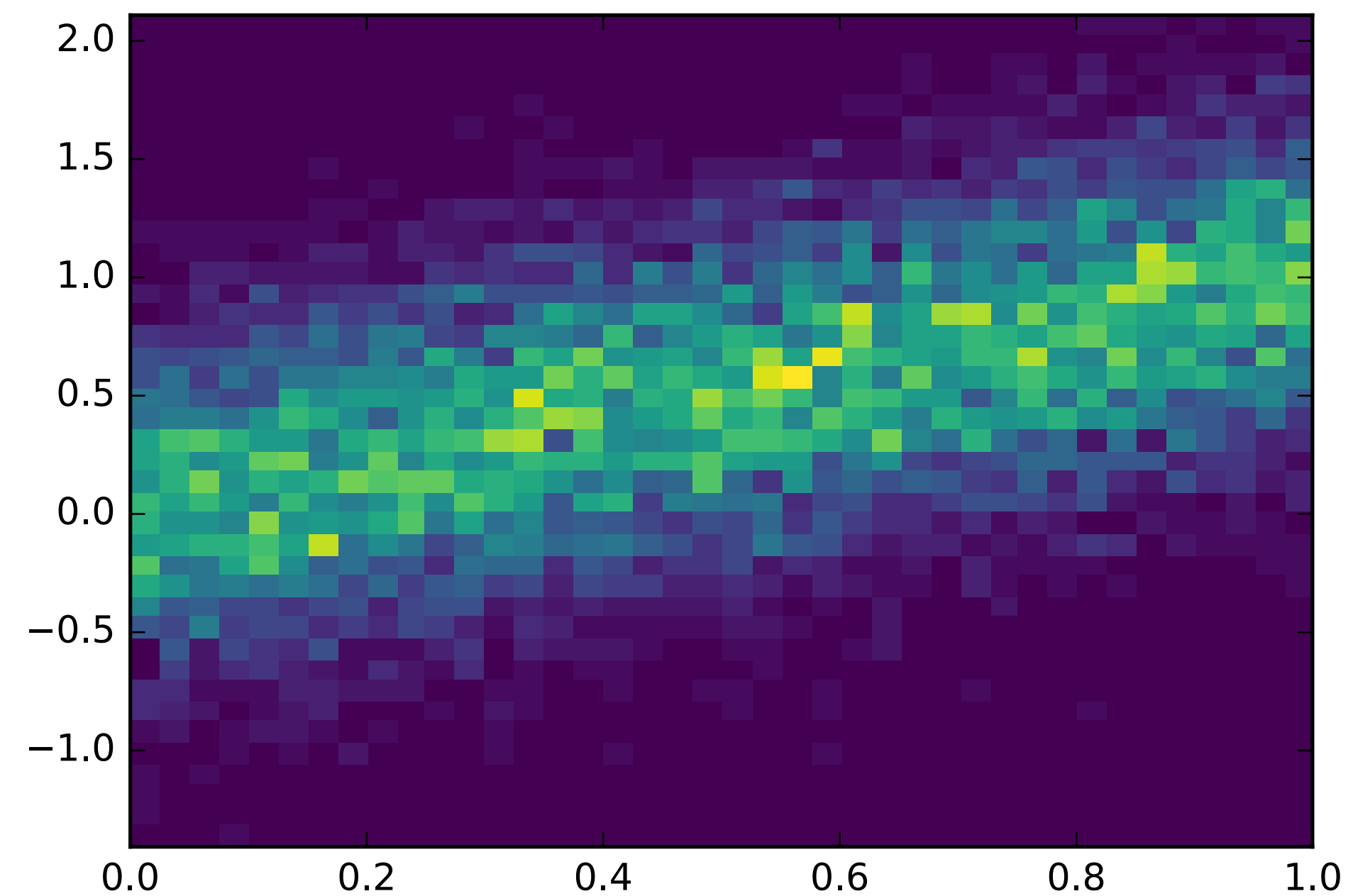
Spatial Aggregation

Modifiable Areal Unit Problem

- How you draw boundaries impacts the type of aggregation you get

- Similar to bins in histograms

- Gerrymandering

Pennsylvania-7
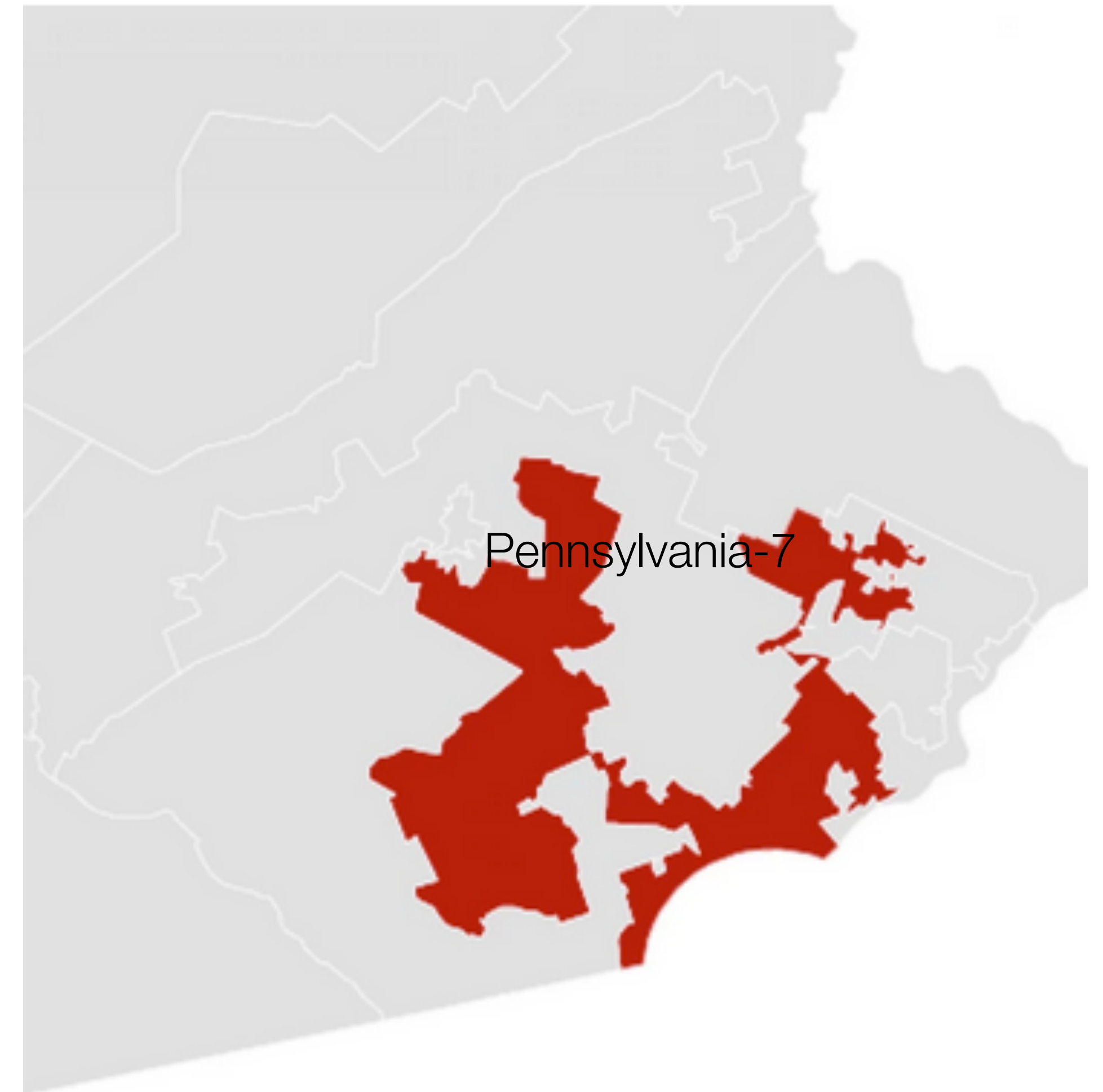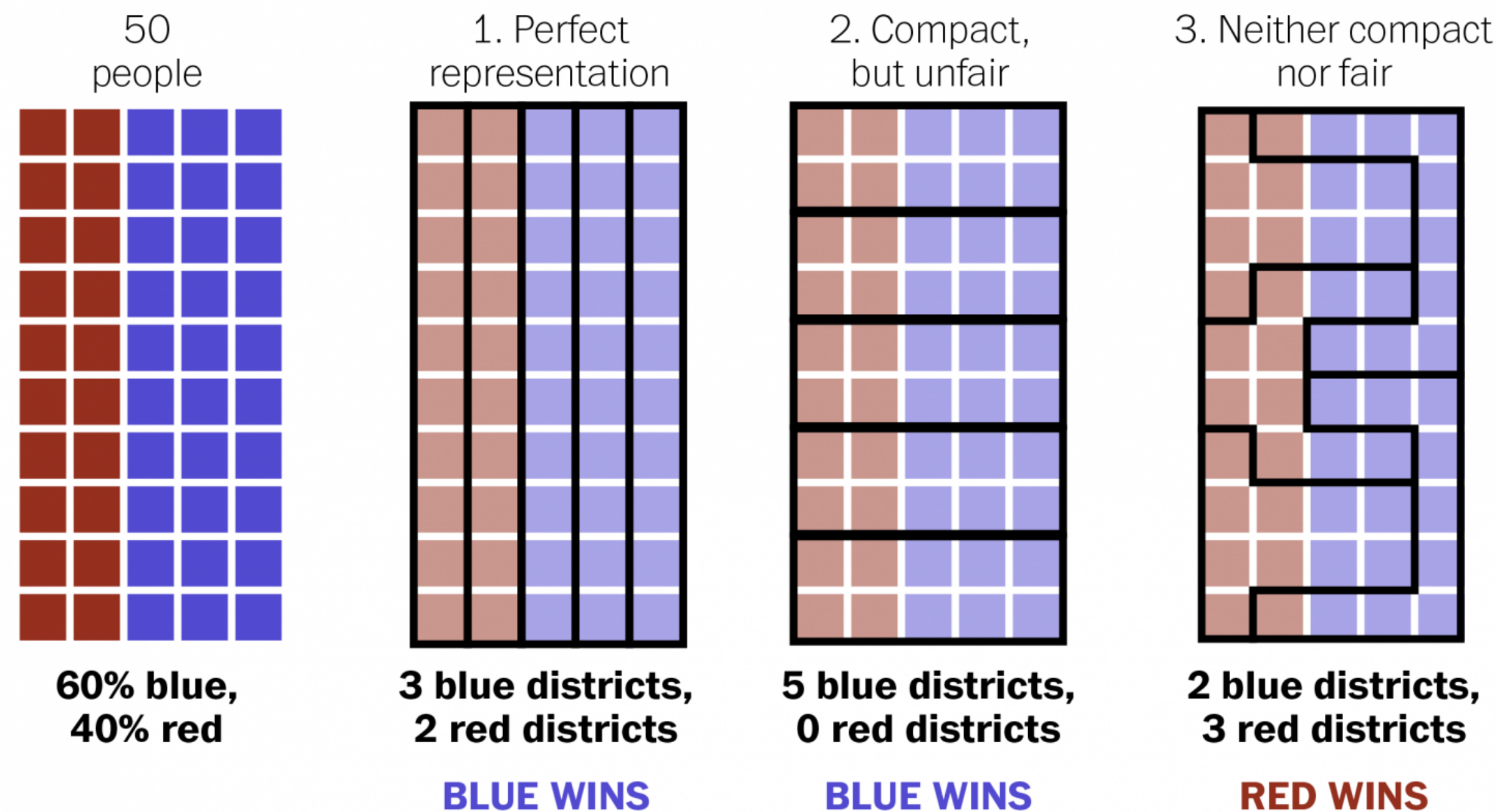
Drawing Different Maps: Compactness

Congressional districts drawn to be compact while trying to respect county borders

How often we'd expect a party to win each of the nation's 435 seats over the long term — not specifically the 2018 midterms — based on historical patterns since 2006

CHANCE OF BEING REPRESENTED BY EITHER PARTY

100% D ← → 100% R
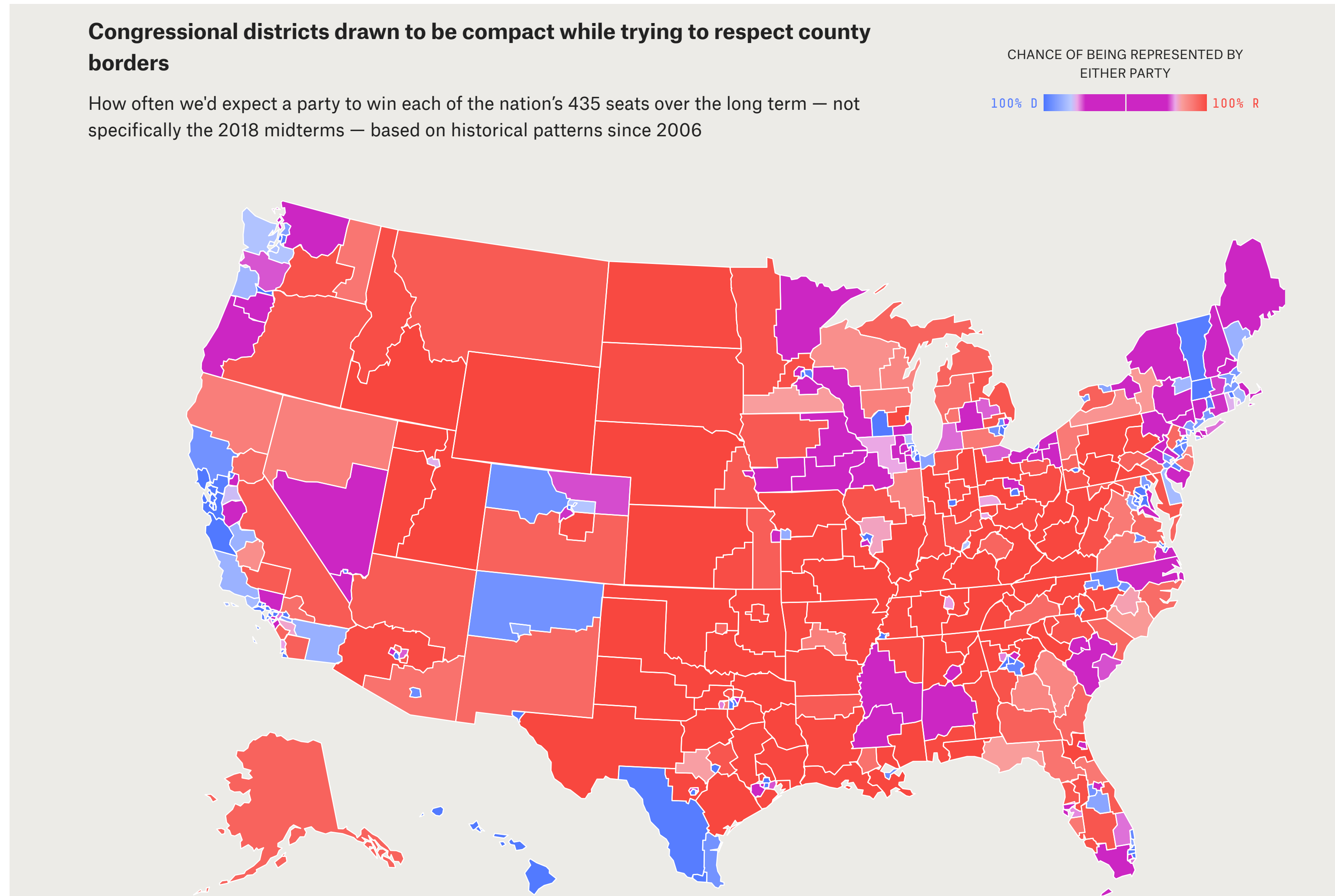
[A. Bycoffe et al., 538]

Drawing Different Maps

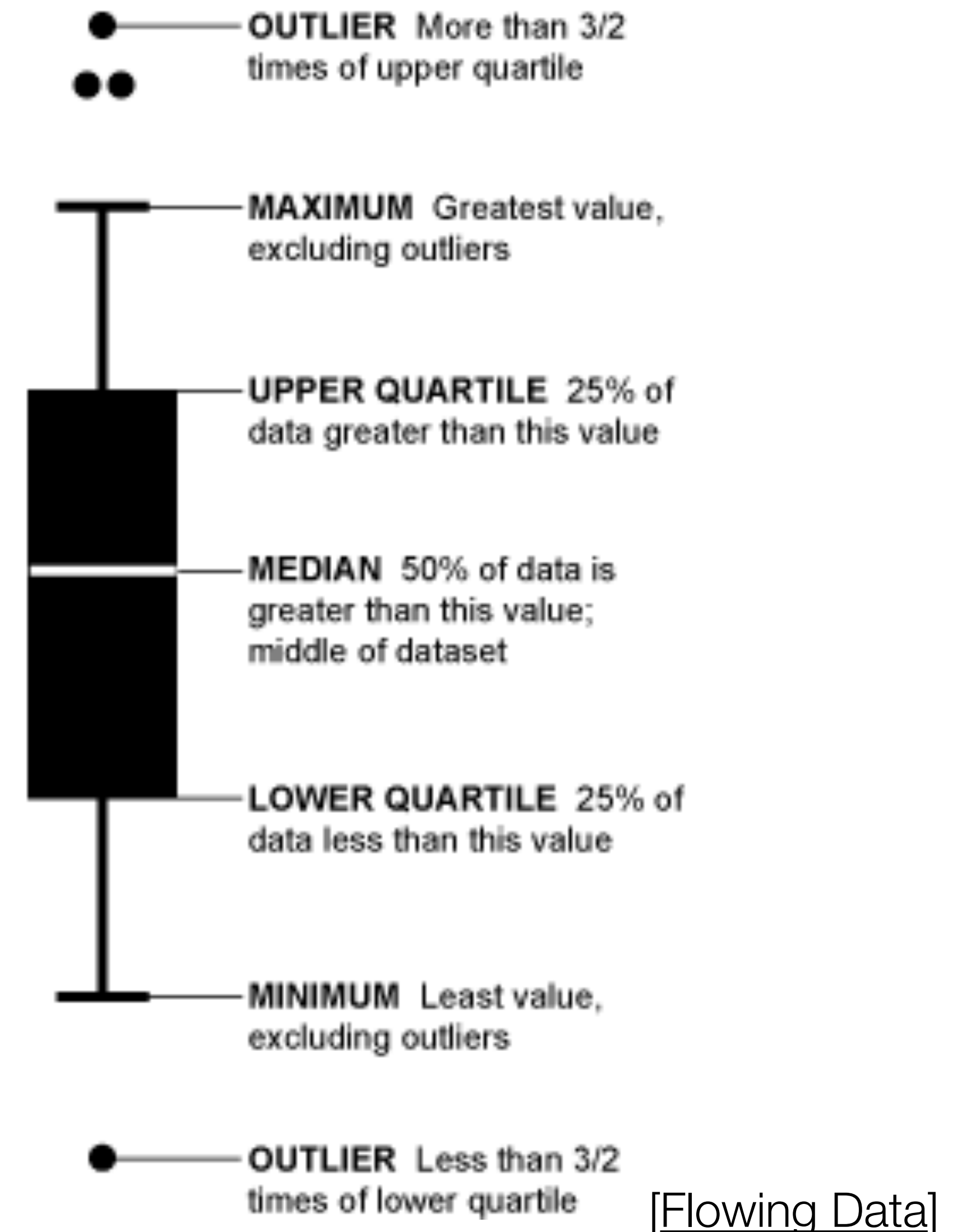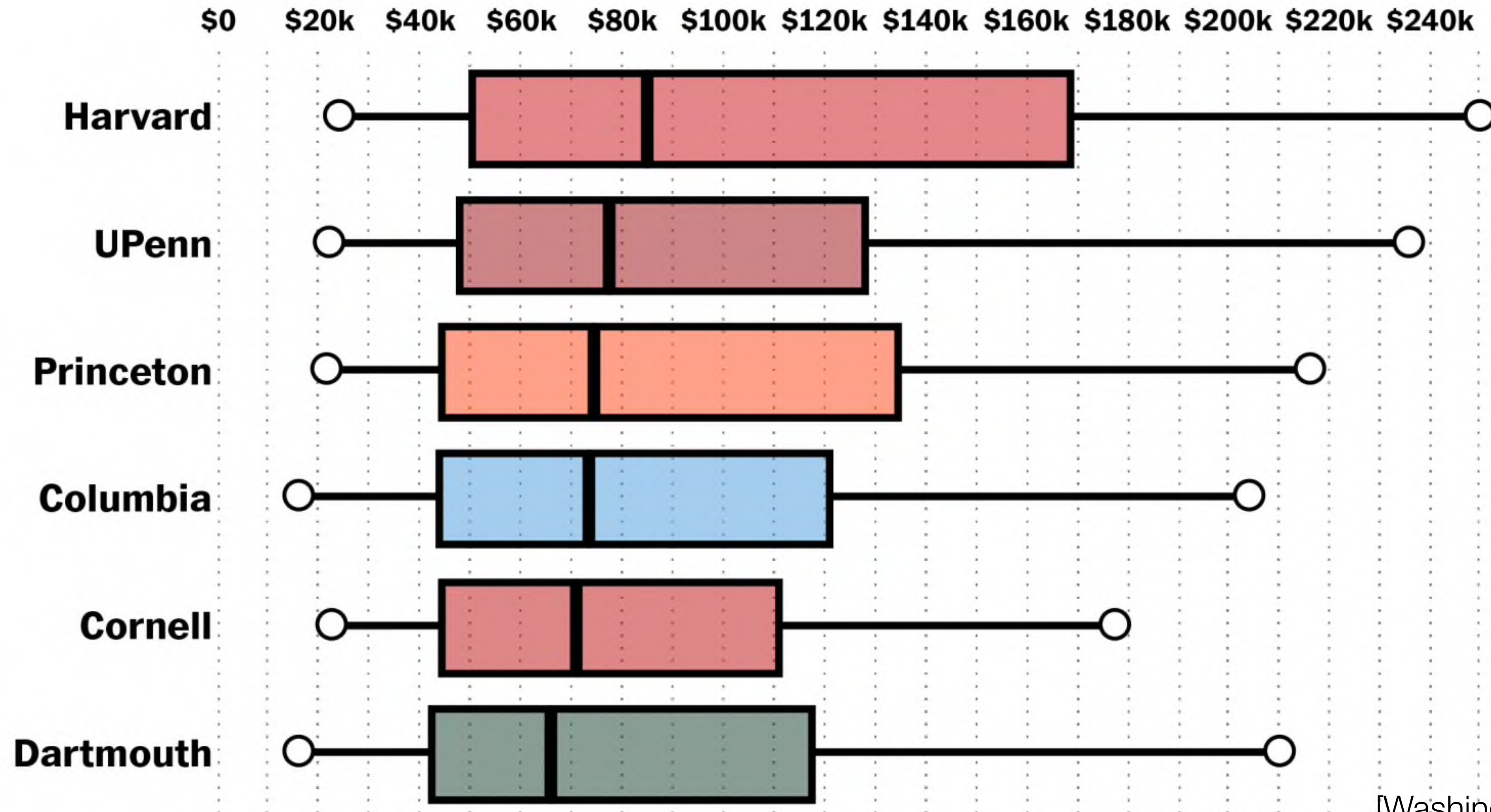| MAP	USUALLY DEM. DISTRICTS	HIGHLY COMPETITIVE	USUALLY REPUBLICAN	EXPECTED SEAT SPLIT DEM.	GOP
Democratic gerrymander	263	27	145	250.6	184.4
Proportionally partisan	174	82	179	214.0	221.0
Majority minority	169	82	184	209.8	225.2
Highly competitive	94	242	99	209.4	225.6
Compact (borders)	155	99	181	203.9	231.1
Compact (algorithmic)	151	104	180	202.8	232.2
Current	168	72	195	**200.6**	**234.4**
Republican gerrymander	139	21	275	171.3	263.7

[A. Bycoffe et al., 538]

Boxplots

- Show **distribution**

- Single value (e.g. mean, max, min, quartiles) doesn't convey everything

- Created by John Tukey

- Show **spread** and **skew** of data

- Best for **unimodal** data

- Variations like vase plot for multimodal data

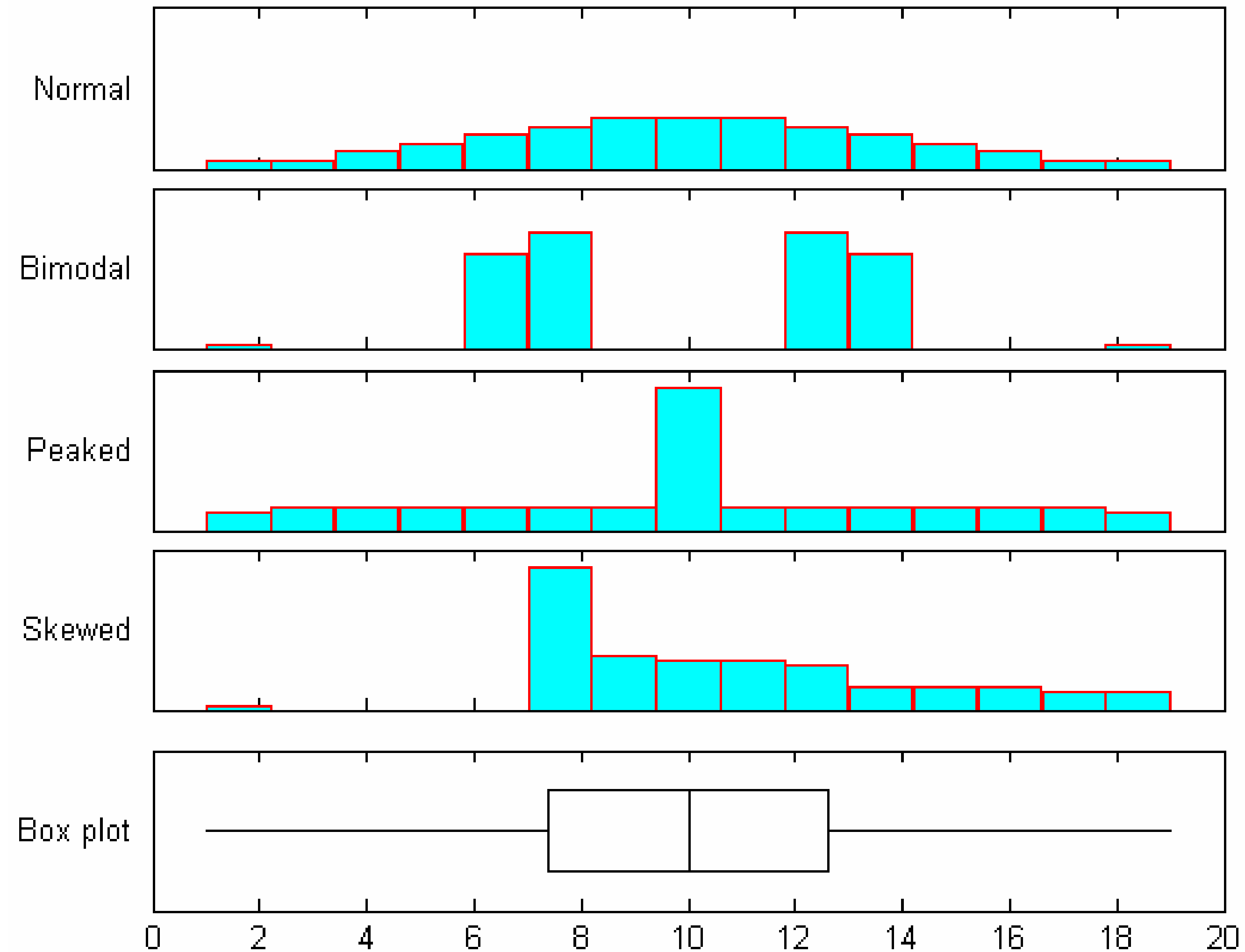- Aggregation here involves many different marks

OUTLIER More than 3/2 times of upper quartile

MAXIMUM Greatest value, excluding outliers

UPPER QUARTILE 25% of data greater than this value

MEDIAN 50% of data is greater than this value; middle of dataset

LOWER QUARTILE 25% of data less than this value

MINIMUM Least value, excluding outliers

OUTLIER Less than 3/2 times of lower quartile

[Flowing Data]
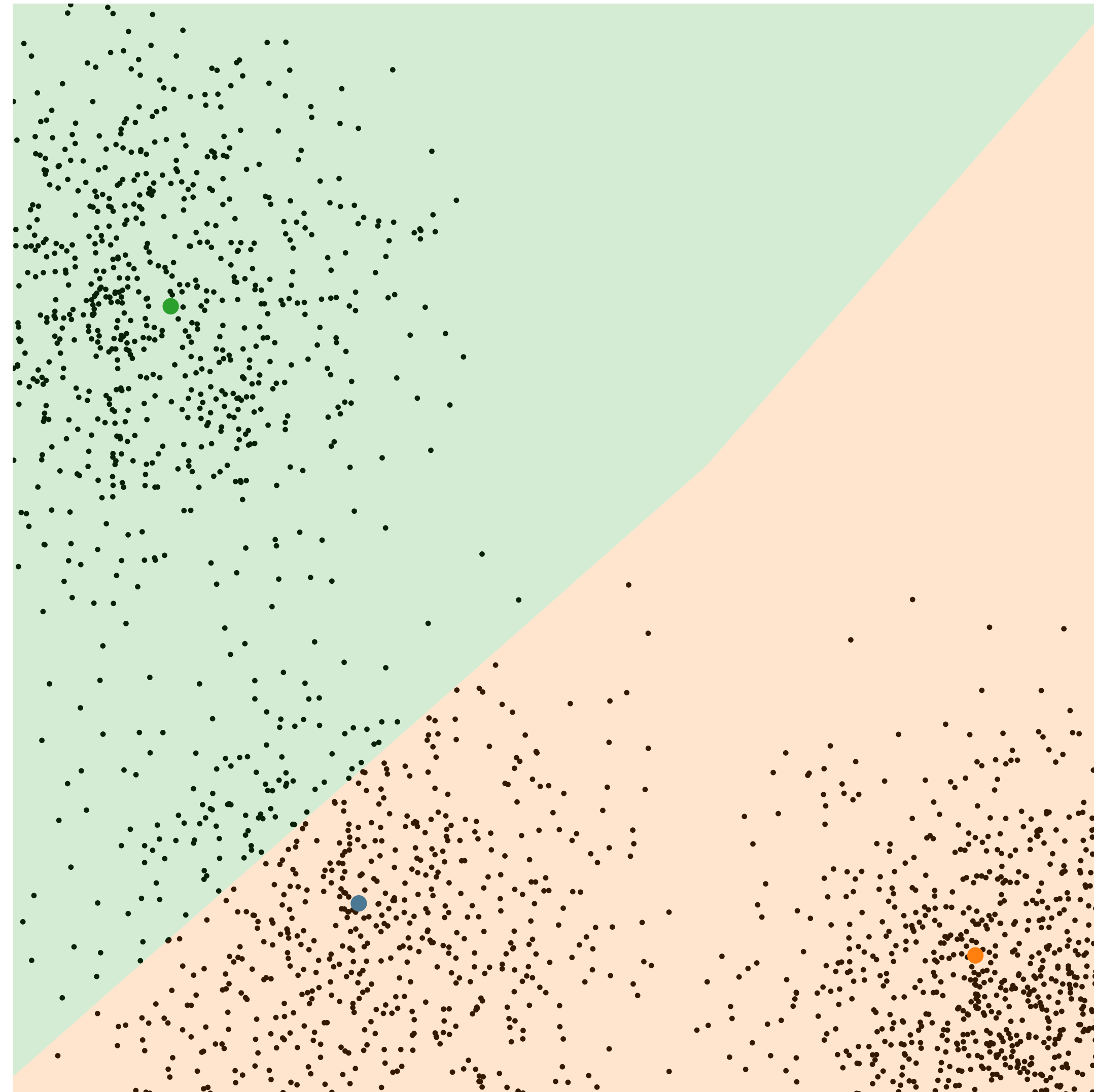
Aggregation: Boxplots

[Washington Post, 2015]

Four Distributions, Same Boxplot…

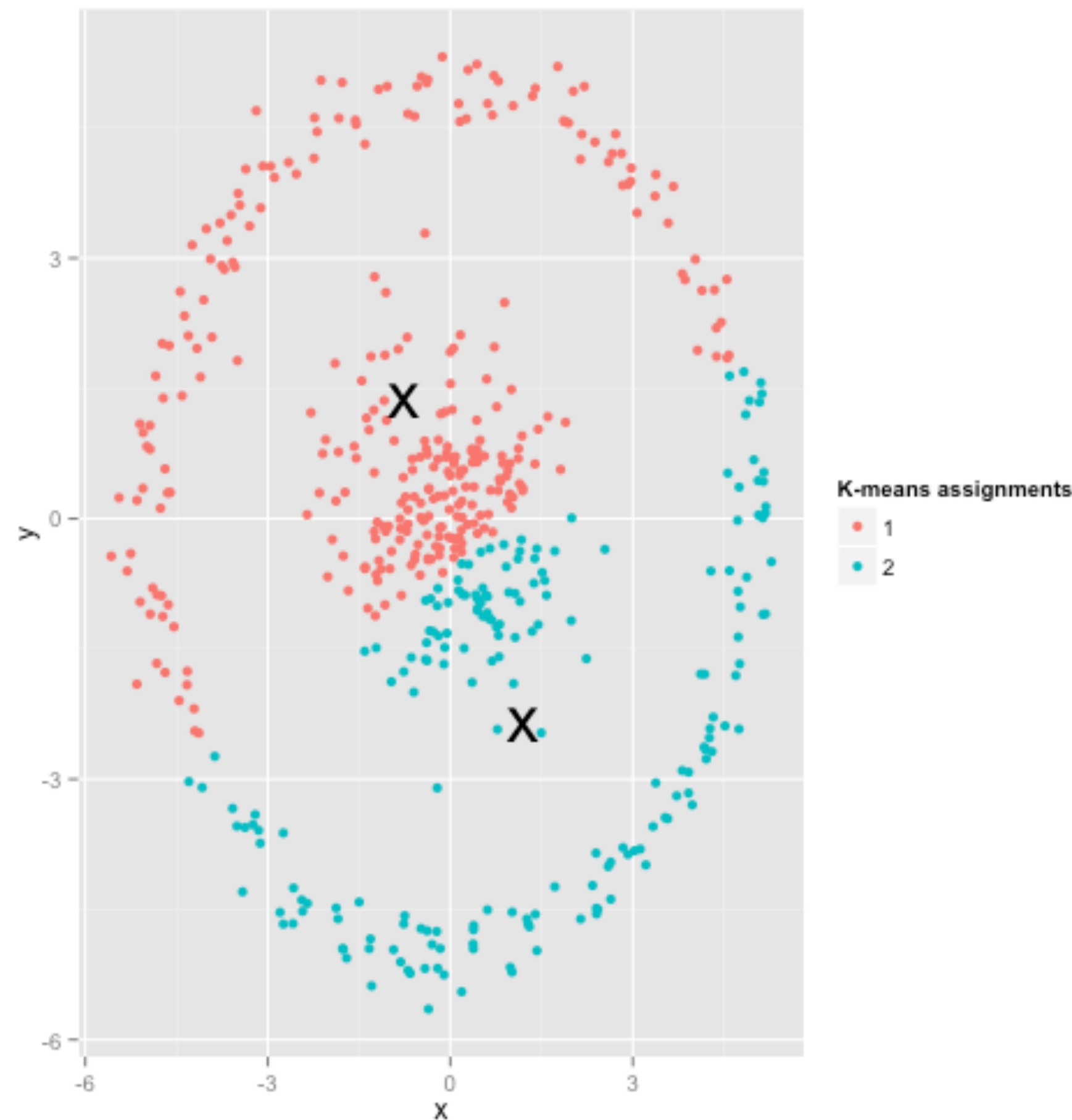[C. Choonpradub and D. McNeil, 2005]
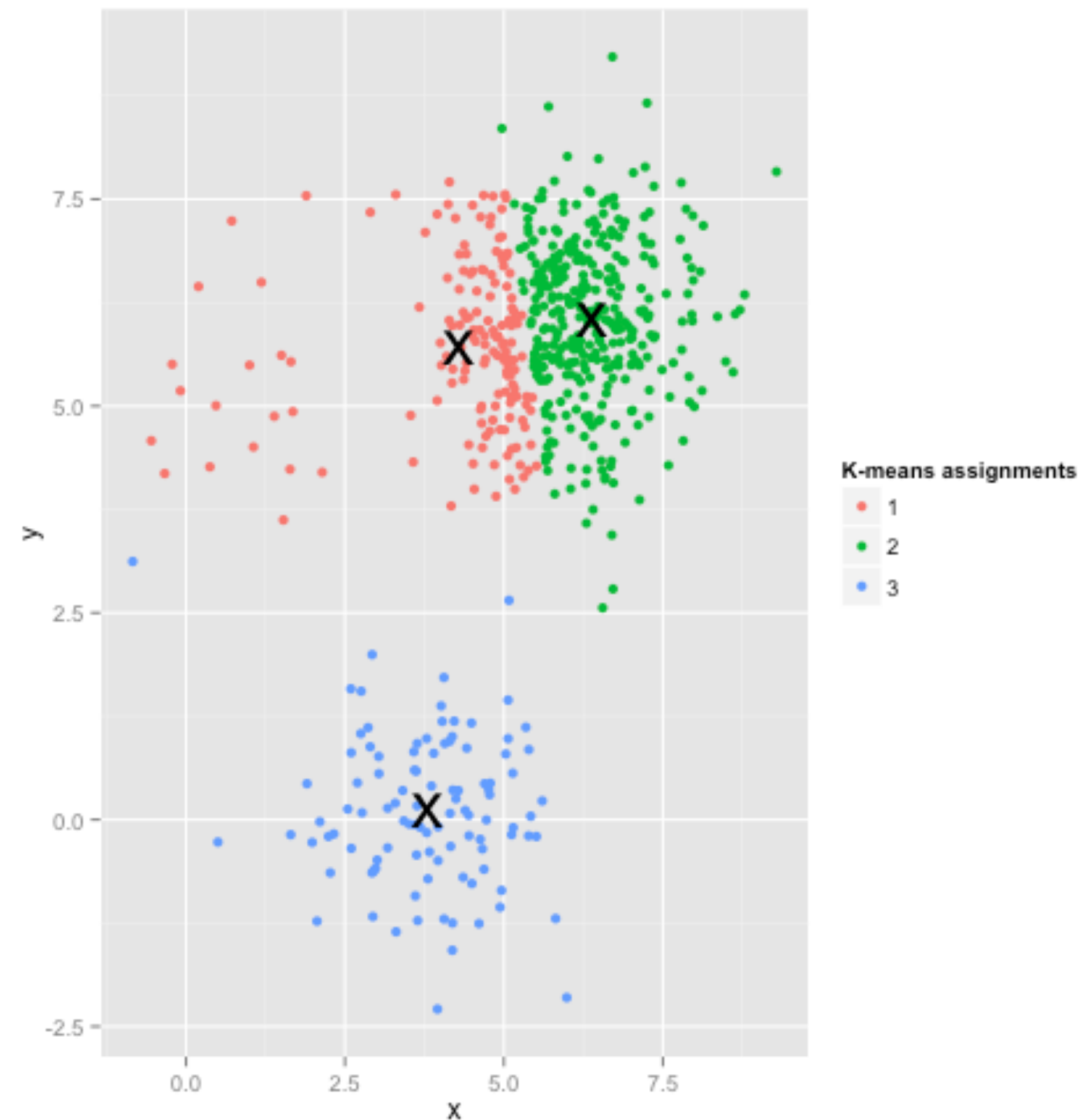
Attribute Aggregation

K-Means

Run
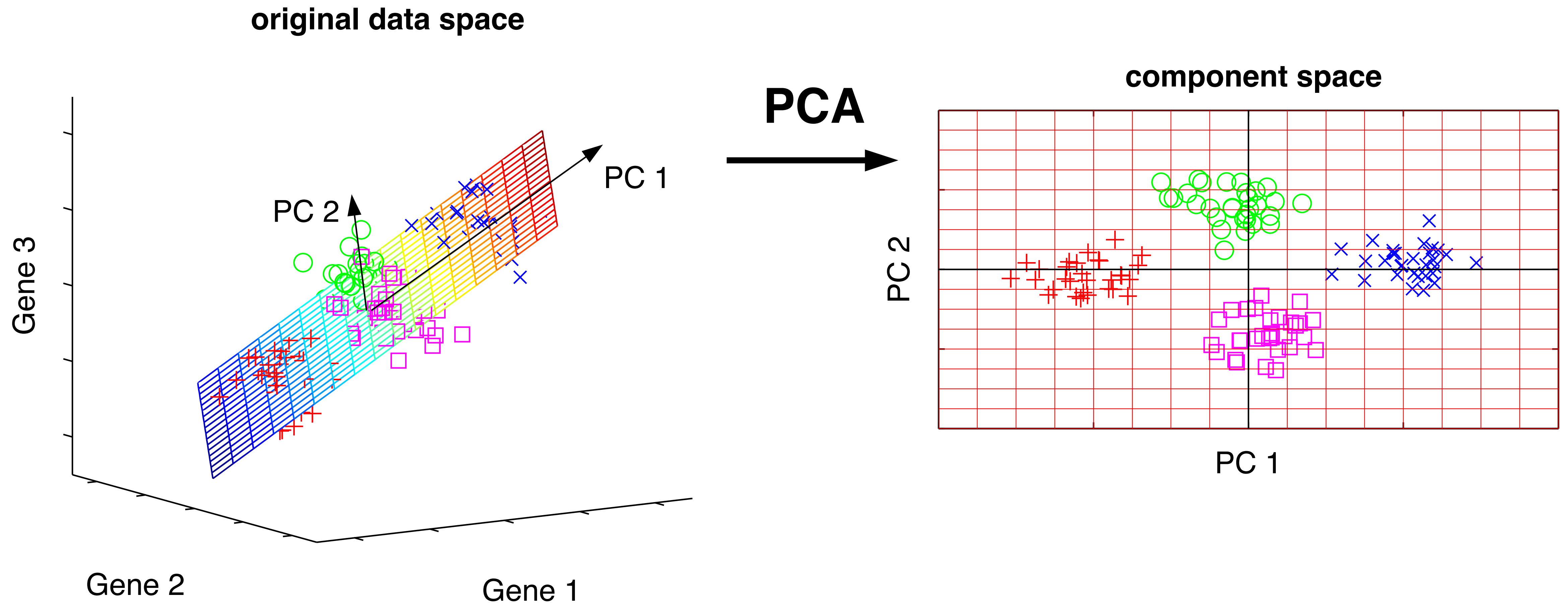
K-Means Issues

Shape

Number of Clusters

[D. Robinson, 2015]
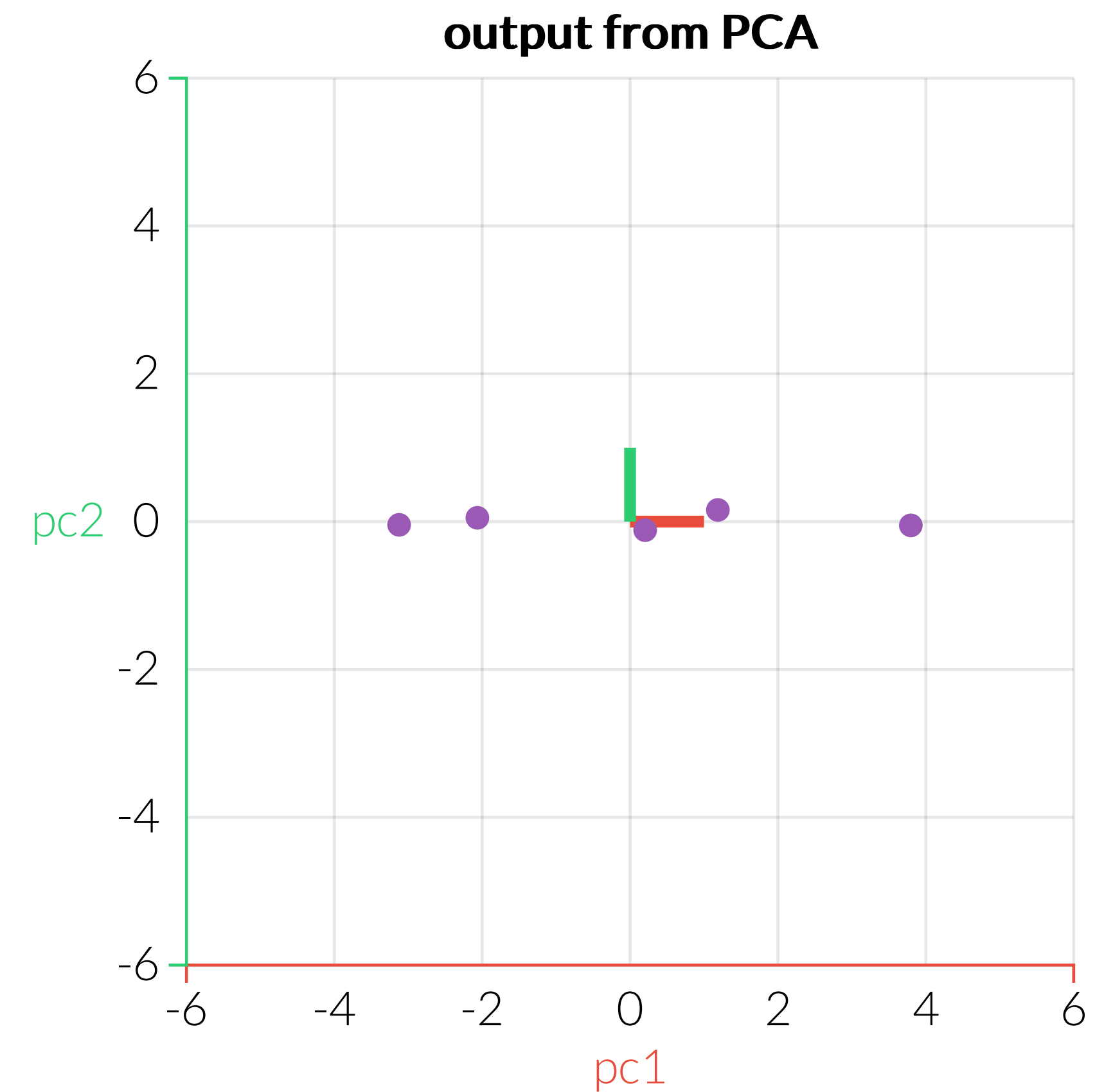
Dimensionality Reduction

- Attribute Aggregation: Use fewer attributes (dimensions) to represent items
- Combine attributes in a way that is more instructive than examining each individual attribute
- Example: Understanding the language in a collection of books
  - Count the occurrence of each non-common word in each book
  - Huge set of features (attributes), want to represent each with an aggregate feature (e.g. high use of "cowboy", lower use of "city") that allows clustering (e.g. "western")
  - Don't want to have to manually determine such rules
- Techniques: Principle Component Analysis, Multidimensional Scaling family of techniques

Northern Illinois University

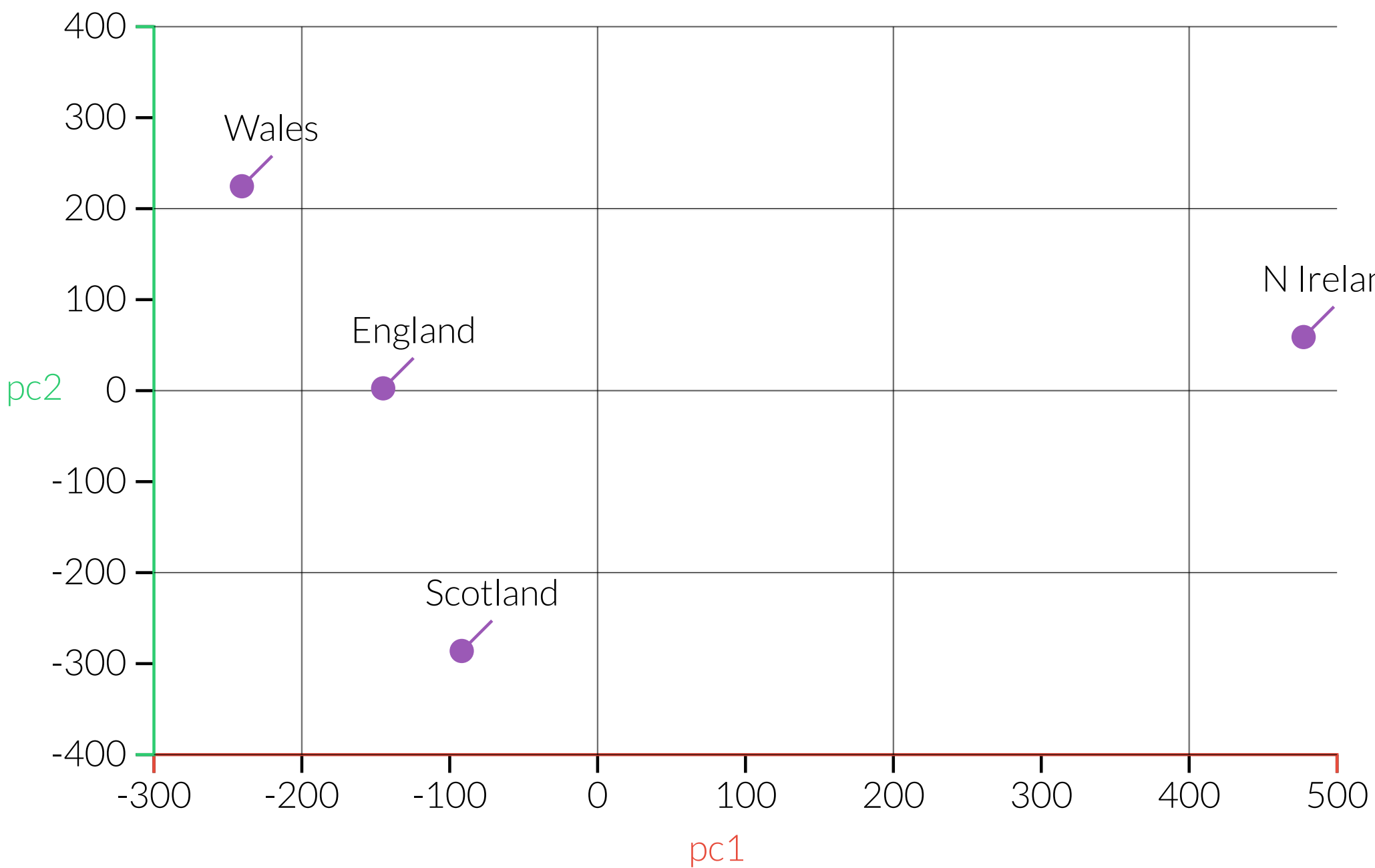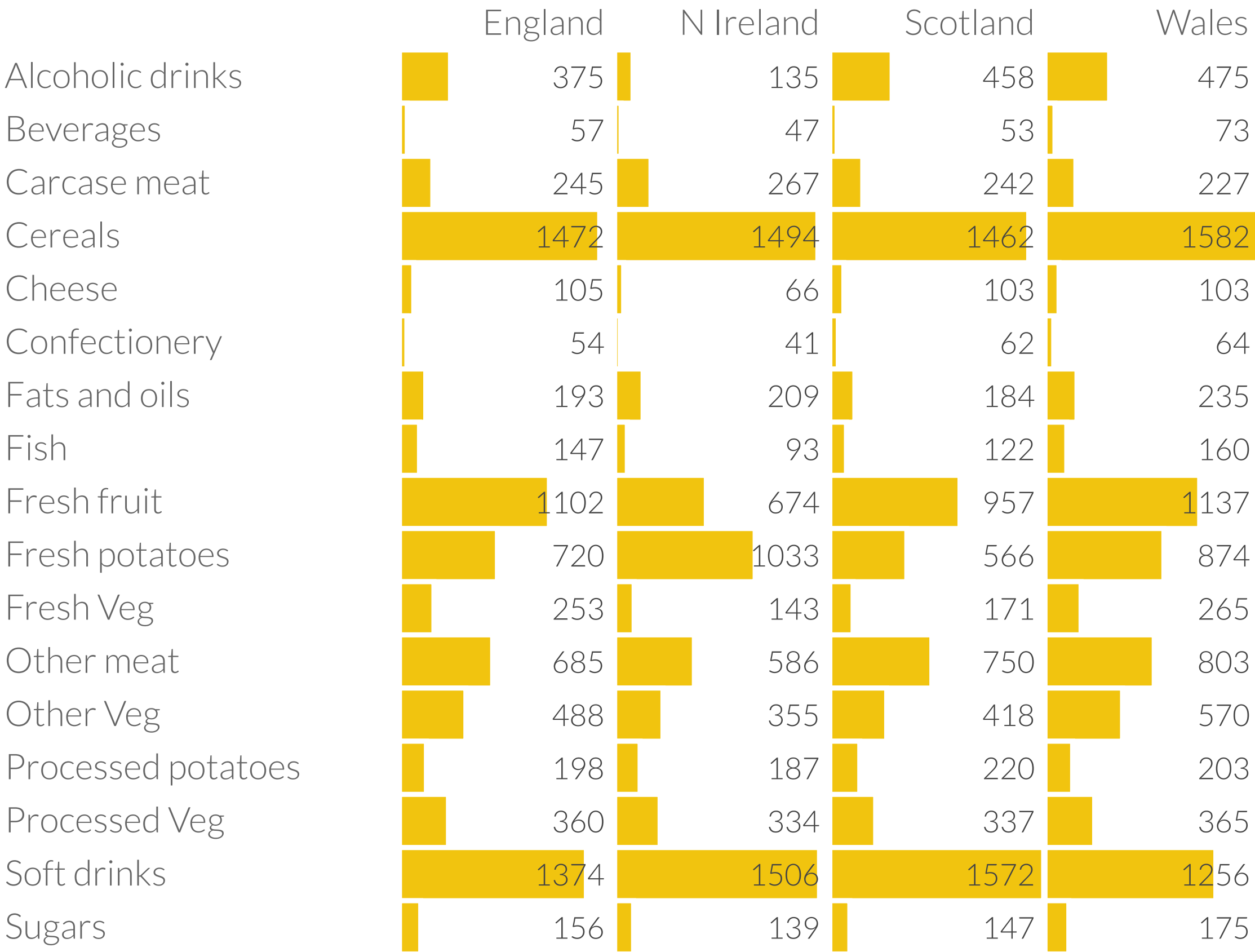# Principle Component Analysis (PCA)

original data space

component space

PCA

PCA

original data set

output from PCA

[Principle Component Analysis Explained, Explained Visually, V. Powell & L. Lehe, 2015]

17 dimensions to 2

	England		N Ireland		Scotland		Wales
Alcoholic drinks		375		135		458		475
Beverages		57		47		53		73
Carcase meat		245		267		242		227
Cereals		1472		1494		1462		1582
Cheese		105		66		103		103
Confectionery		54		41		62		64
Fats and oils		193		209		184		235
Fish		147		93		122		160
Fresh fruit		1102		674		957		1137
Fresh potatoes		720		1033		566		874
Fresh Veg		253		143		171		265
Other meat		685		586		750		803
Other Veg		488		355		418		570
Processed potatoes		198		187		220		203
Processed Veg		360		334		337		365
Soft drinks		1374		1506		1572		1256
Sugars		156		139		147		175

[Principle Component Analysis Explained, Explained Visually, V. Powell & L. Lehe, 2015]

Non-linear Dimensionality Reduction

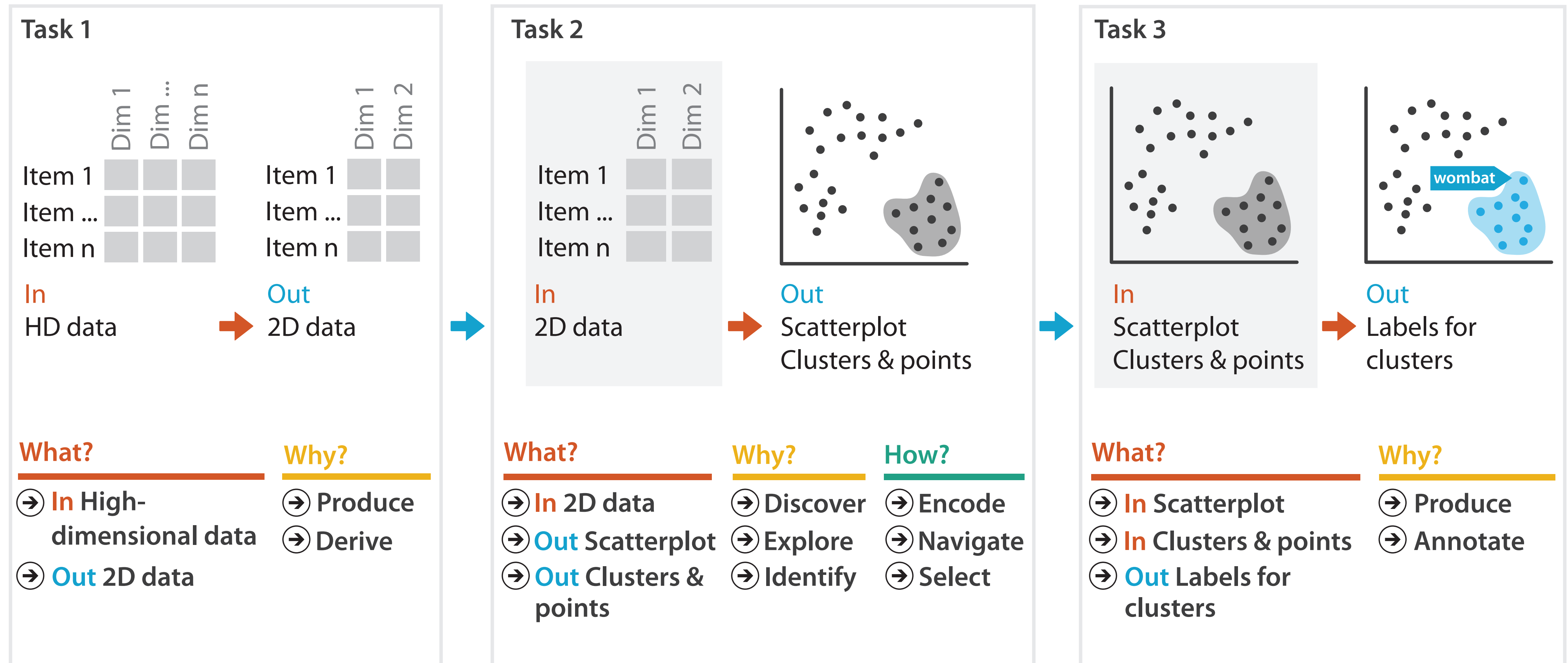$$\Phi_{gen} : \mathcal{Z} \rightarrow \mathcal{X}$$

$$\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$$

original data space $\mathcal{X}$

component space $\mathcal{Z}$
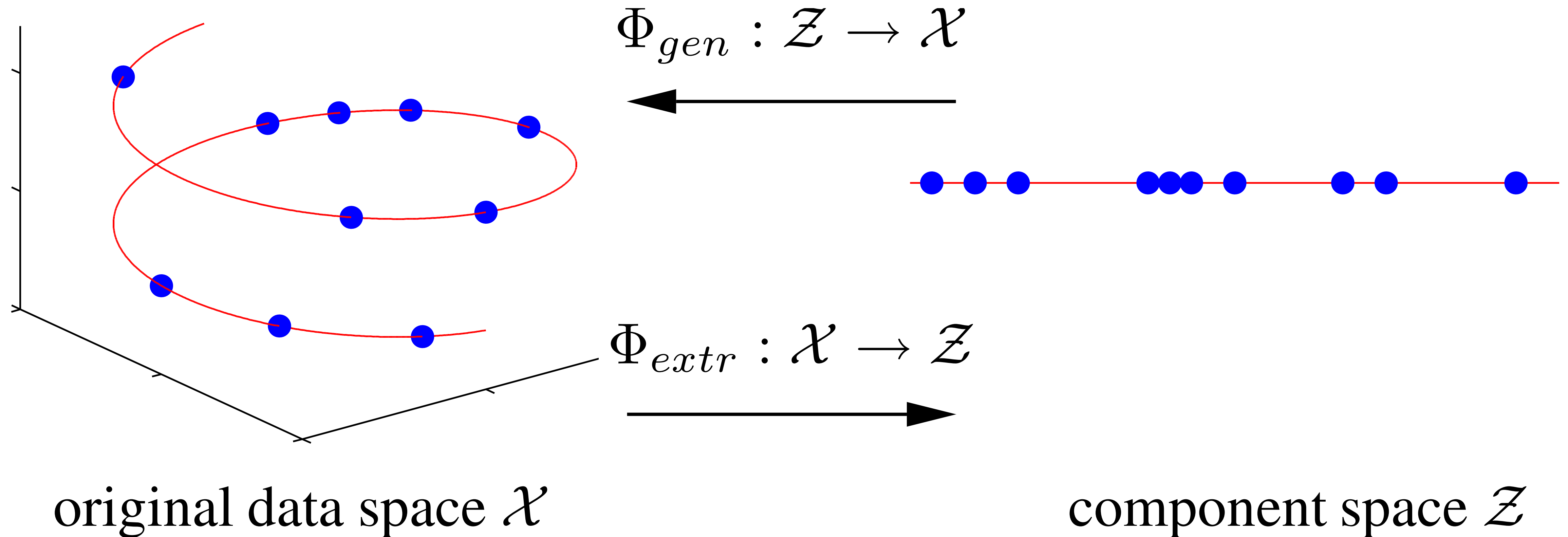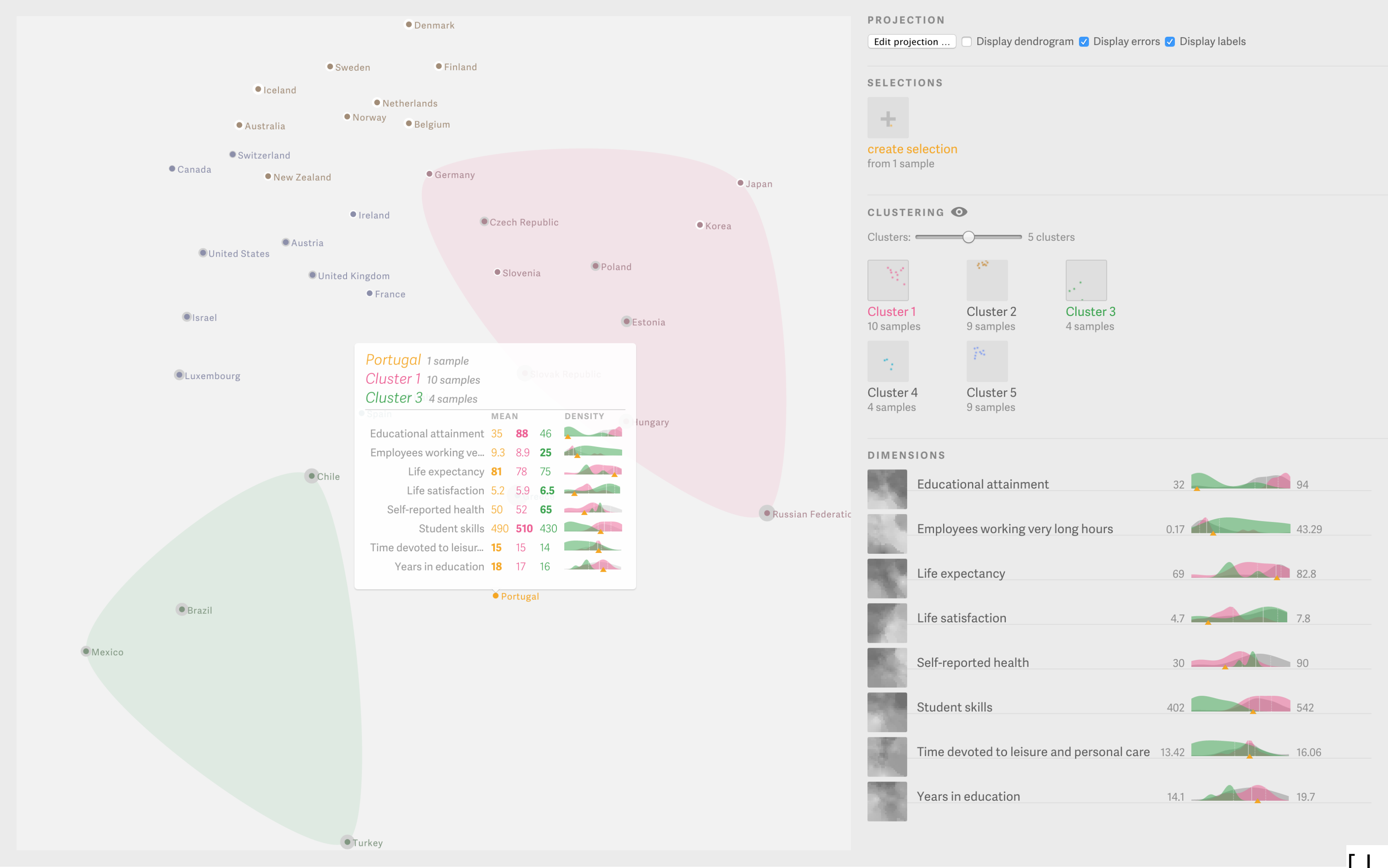
Dimensionality Reduction in Visualization

[Glimmer, Ingram et al., 2009]

Tasks in Understanding High-Dim. Data

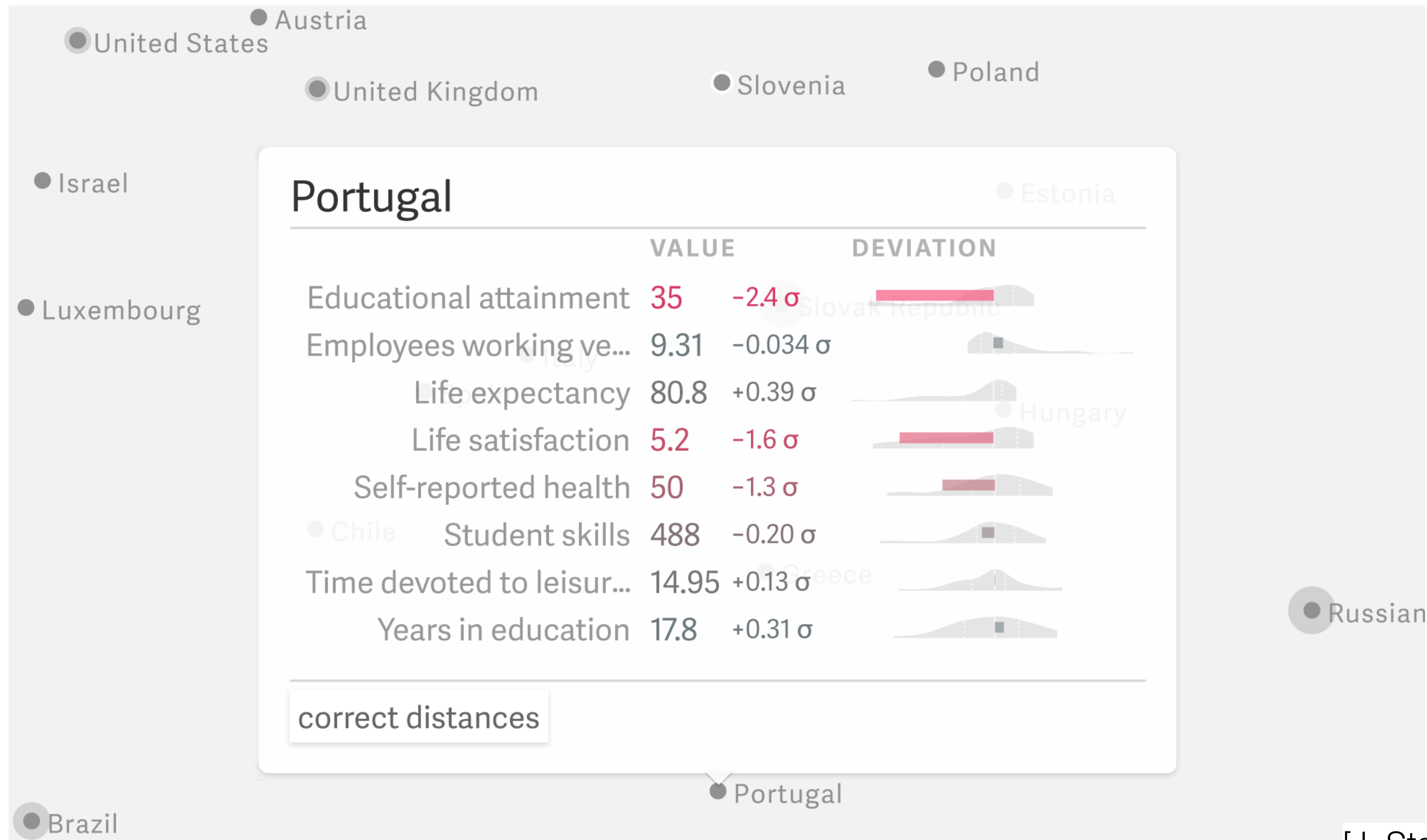[Munzner (ill. Maguire), 2014]

Probing Projections

[J. Stahnke et al., 2015]

Probing Projection Goals

- Examining the Projection

- Exploring the Data

- Design Goals:

  - Show and correct approximation errors

  - Allow for multi-level comparisons

  - Spatial orientation

  - Consistent design

- Allow **grouping** of samples

  - Selections

  - Classes

  - Clusters

[J. Stahnke et al., 2015]

Tooltips with statistics

Northern Illinois University
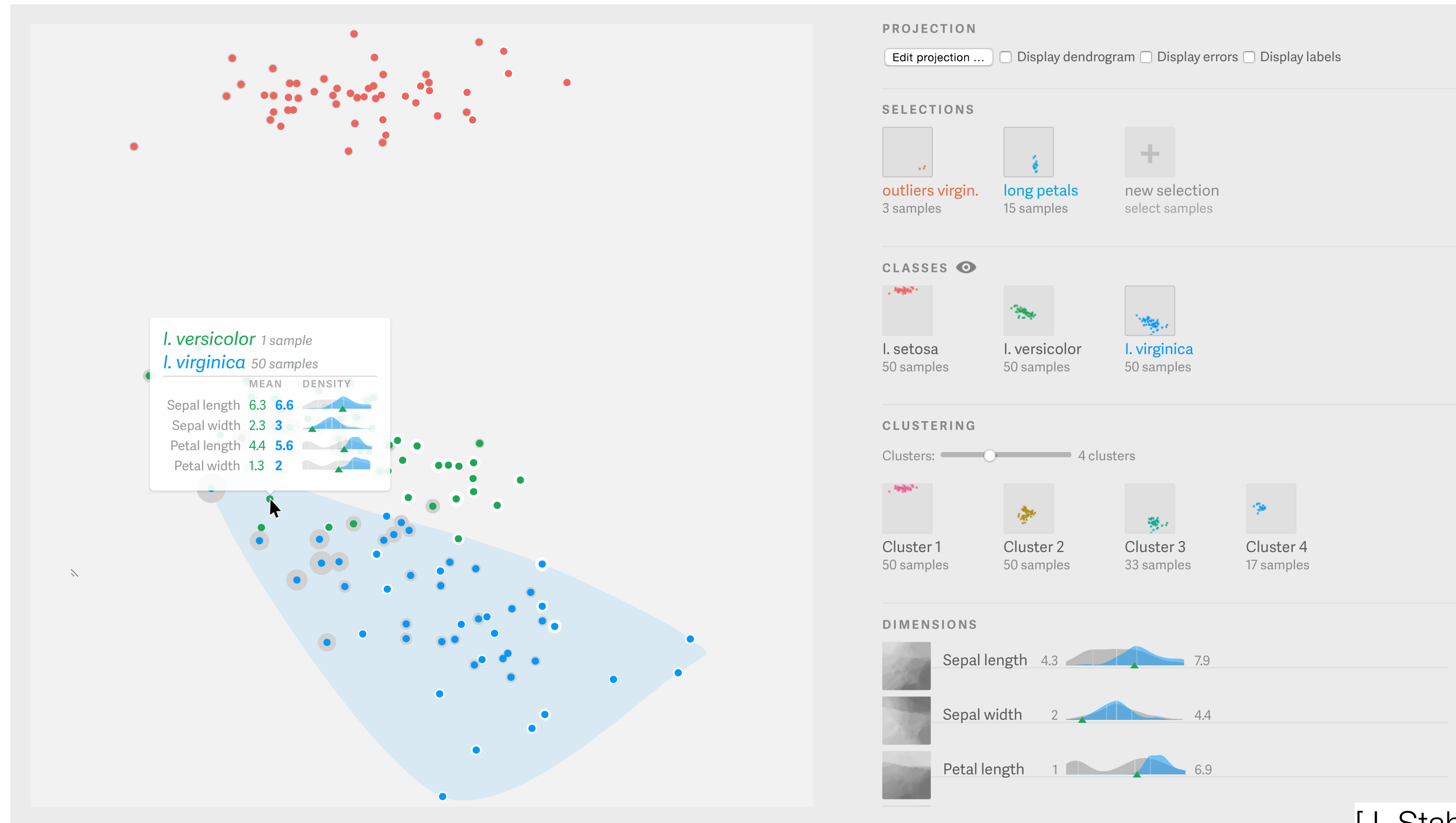
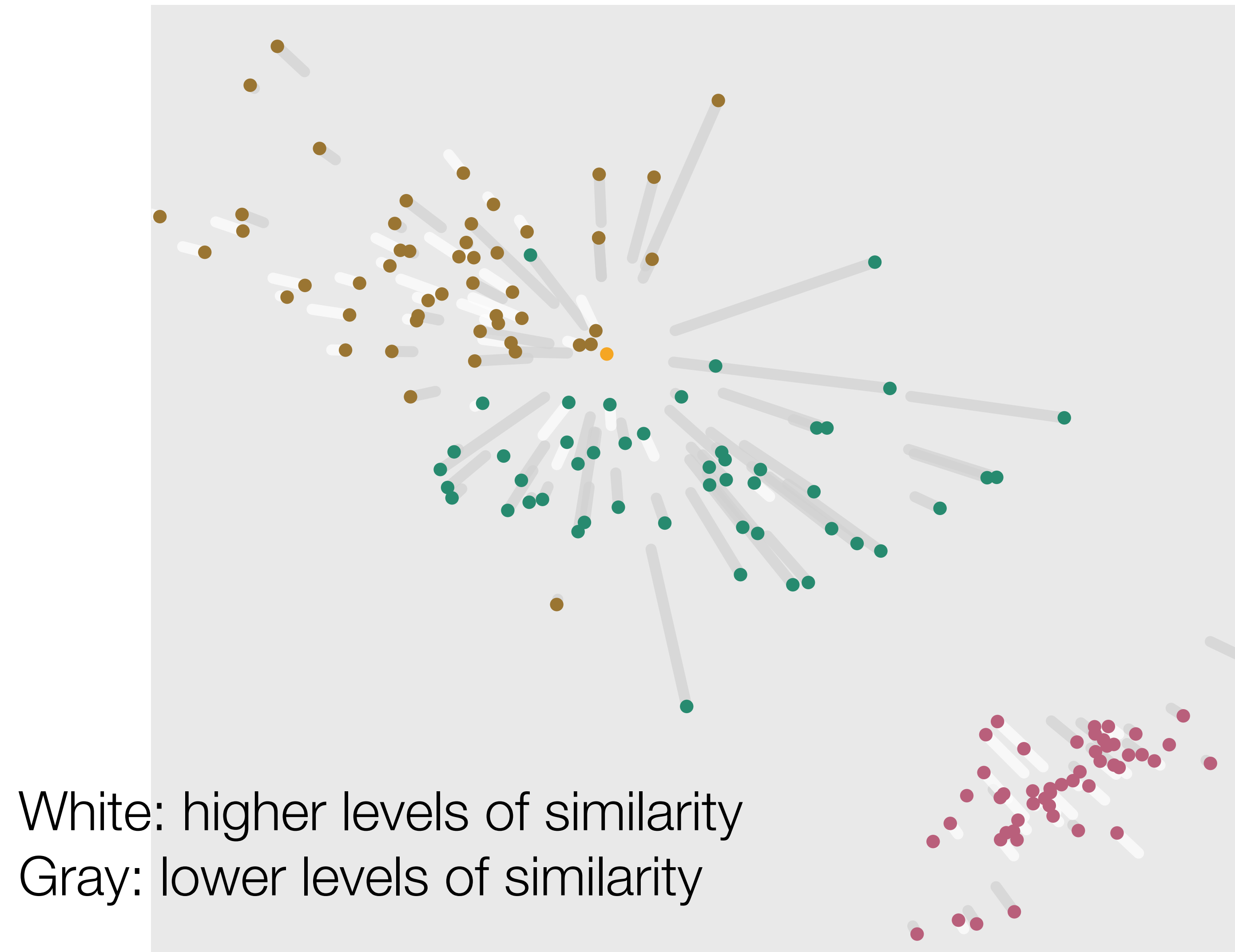# Comparing Two Groups

[J. Stahnke et al., 2015]

Heatmap from Dimension Hover

[J. Stahnke et al., 2015]

Showing Error via Sample-centric Halos

[J. Stahnke et al., 2015]

Showing Projection Errors

White: higher levels of similarity
Gray: lower levels of similarity

[J. Stahnke et al., 2015]

User Study & Results

- Types of Questions:
  - How would you try to characterize the type X?
  - In what way are X and Y different in their properties?
  - Are the projections of X and Y correct or do they deviate? How do you interpret this?
  - Can you discover which parts of the cluster combinations are A, B, and C?
- Discussion:
  - Learnability: need more effective mechanisms for grasping the concepts behind dimensionality reduction
  - Manipulation: What happens with results?
  - Large data: What about text corpora?

[J. Stahnke et al., 2015]

Focus+Context

Focus+Context

- Show everything at once but compress regions that are not the current focus
  - User shouldn't lose sight of the overall picture
  - May involve some aggregation in non-focused regions
  - "Nonliteral navigation" like semantic zooming
- Elision
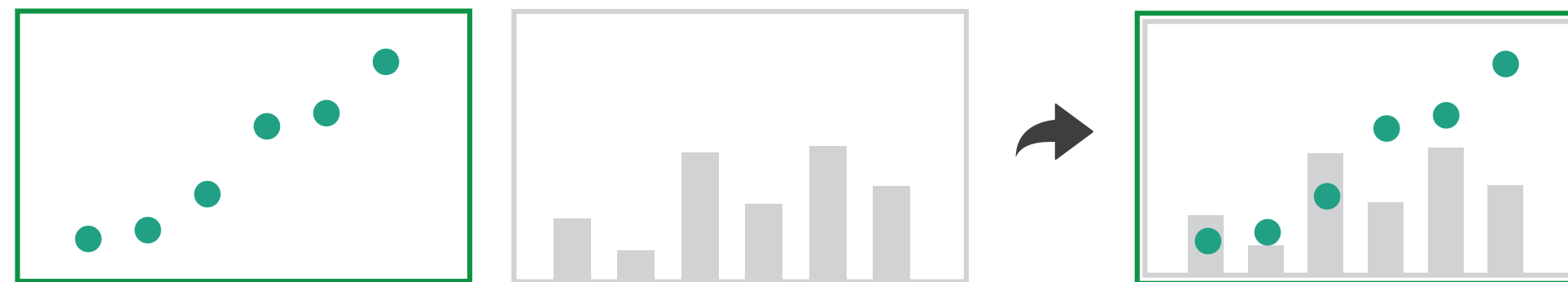- Superimposition: more directly tied than with layers
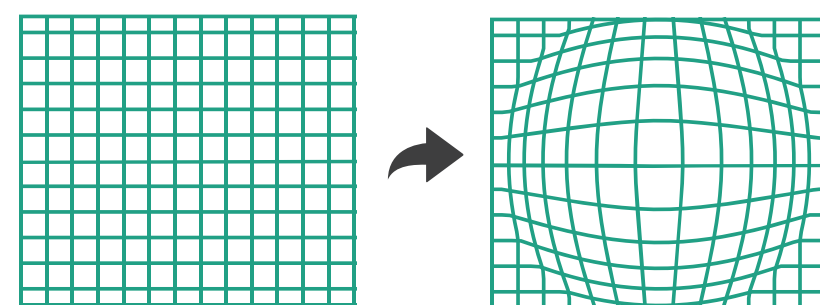- Distortion

Focus+Content Overview

Embed

➔ Elide Data

➔ Superimpose Layer

➔ Distort Geometry

Reduce

➔ Filter

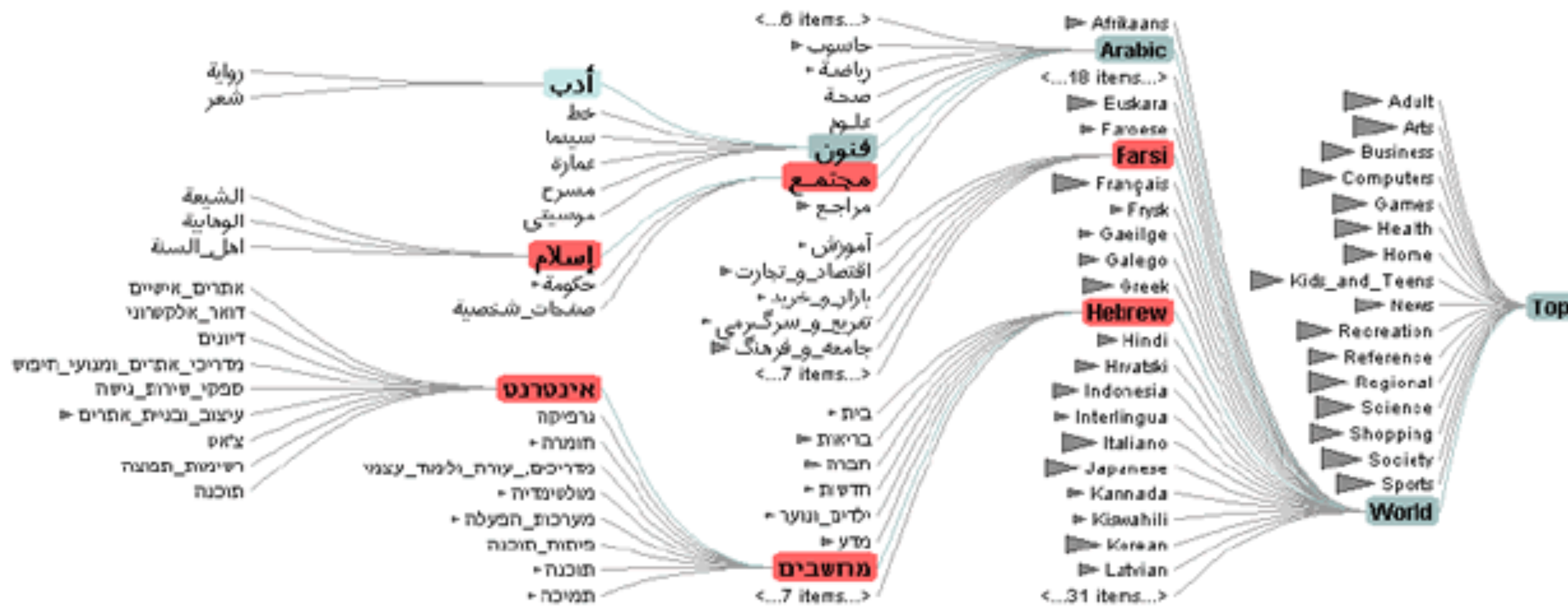➔ Aggregate

➔ Embed

[Munzner (ill. Maguire), 2014]

Elision

- There are a number of examples of elision including in text , DOITrees, …
- Includes both filtering and aggregation but goal is to give overall view of the data
- In visualization, usually correlated with focus regions

Degree of Interest Function

- DOI = I(x) - D(x,y)
  - I: interest function
  - D: distance (semantic or spatial)
  - x: location of item
  - y: current focus point (could be more than one)
- Interactive: y changes

Elision: DOITrees

- Example: 600,000 node tree
  - Multiple foci (from search results or via user selection)
  - Distance computed topologically (levels, not geometric)

[Heer and Card, 2004]