Data Visualization (CSCI 627/490)

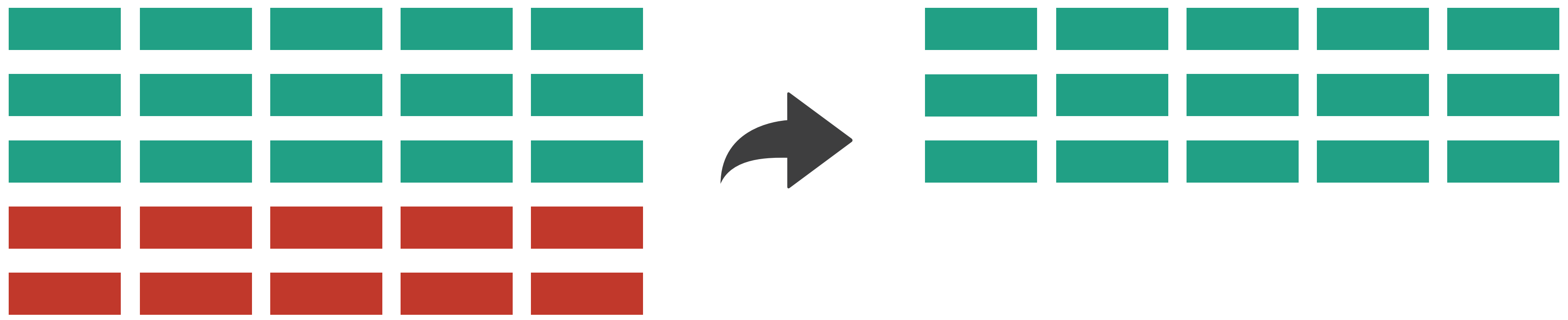
Aggregation & Focus+Context

Dr. David Koop

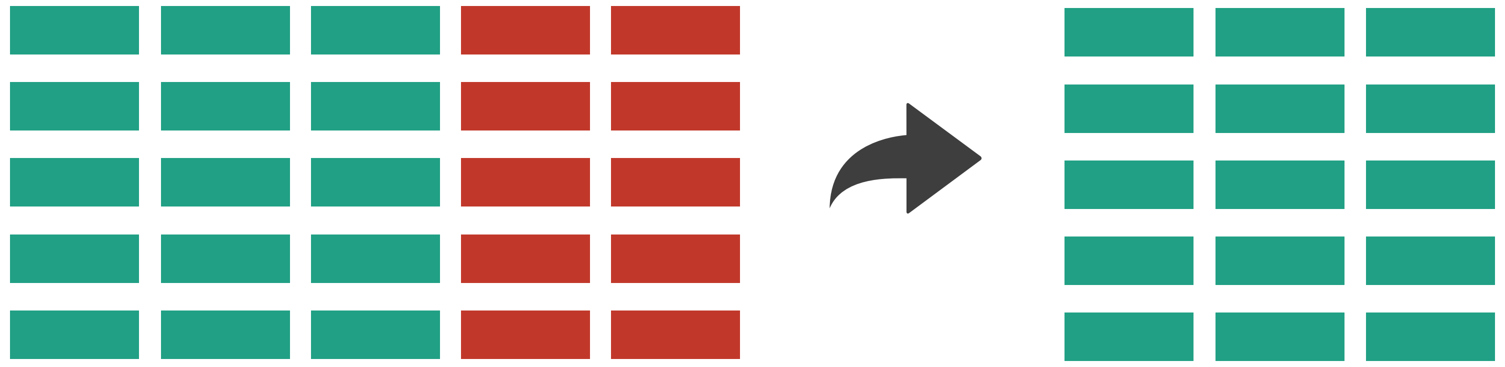
Overview: Reducing Items & Attributes

➔ Filter

→ Items

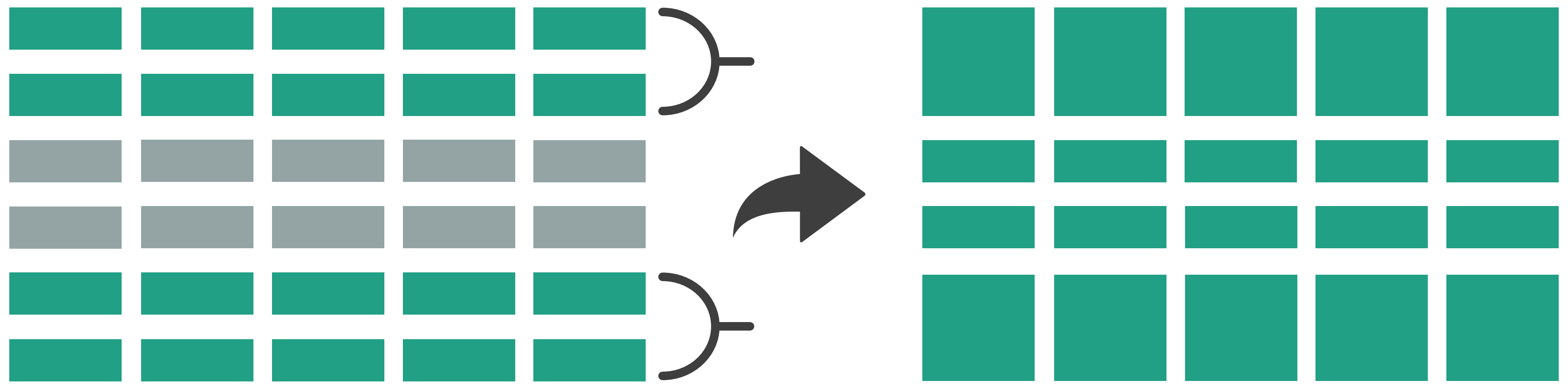


→ Attributes

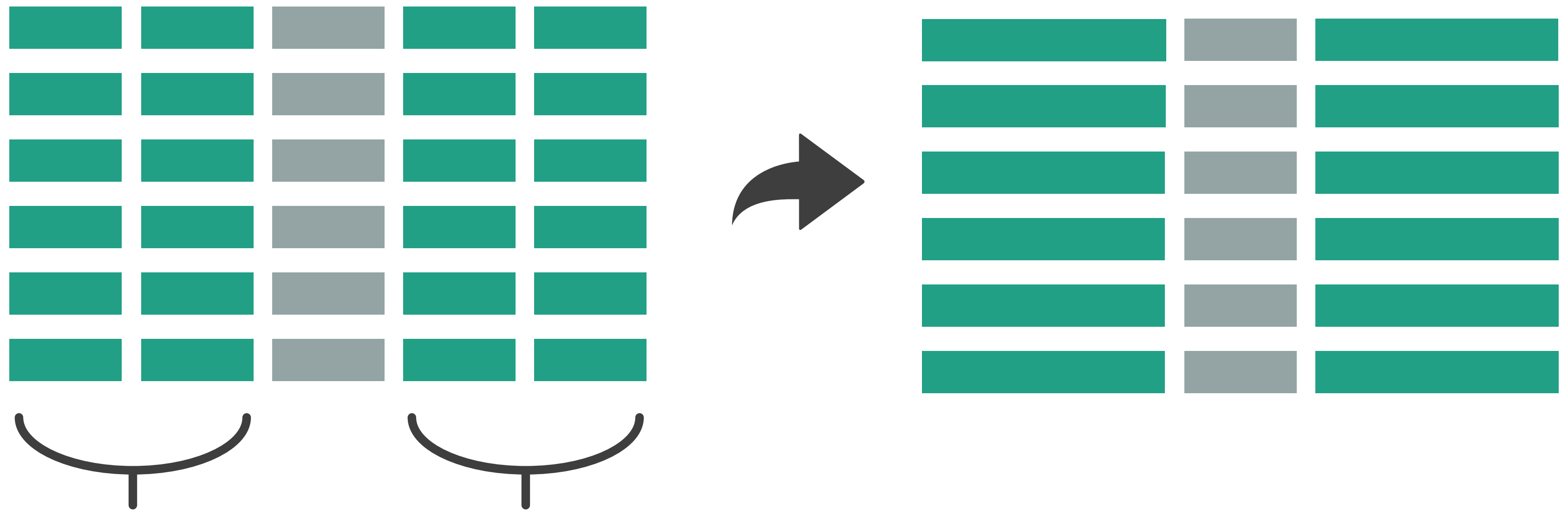


➔ Aggregate

→ Items

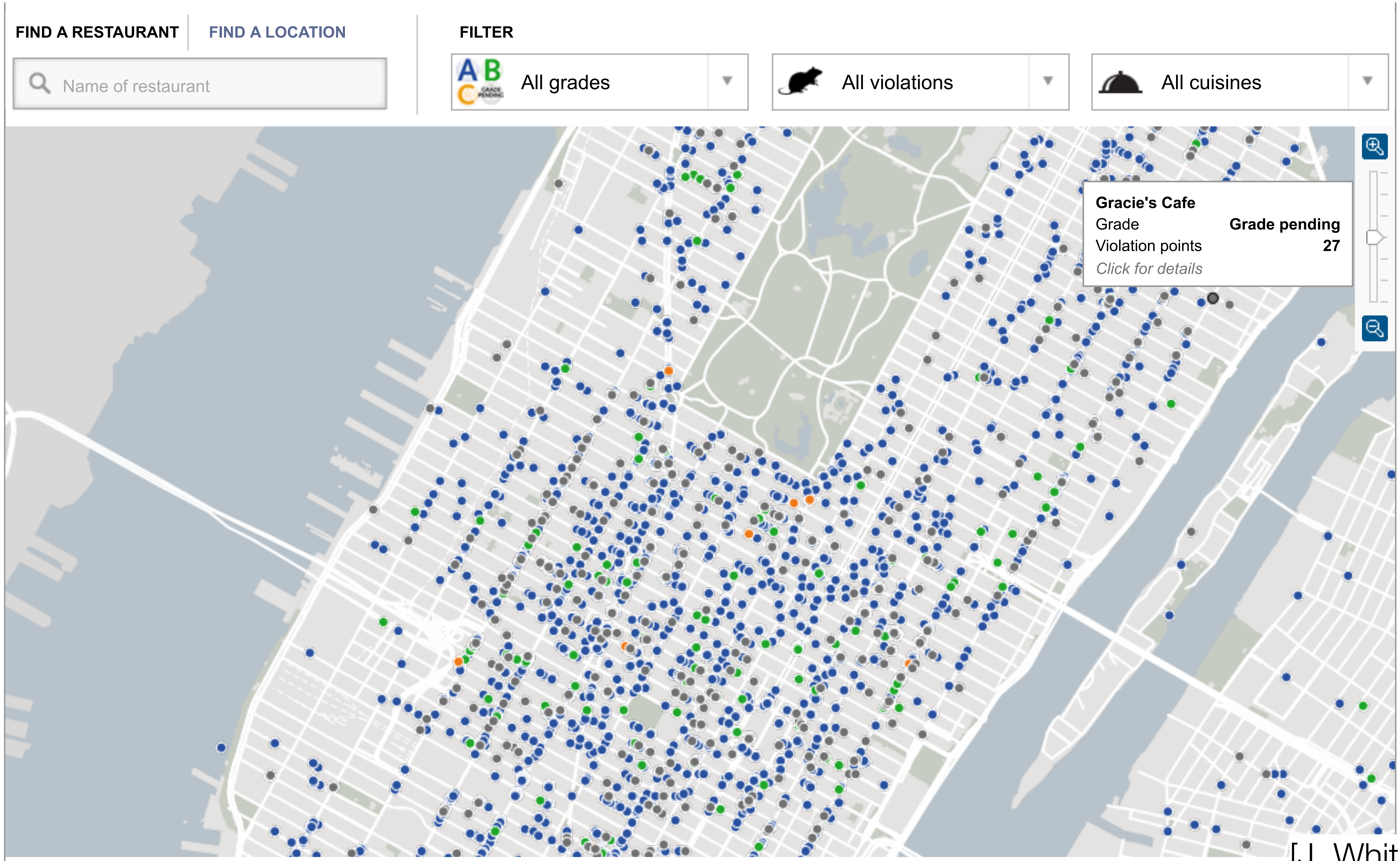


→ Attributes



[Munzner (ill. Maguire), 2014]

Item Filtering on Maps



FIND A RESTAURANT

FIND A LOCATION

Name of restaurant

FILTER

All grades

All violations

All cuisines

Gracie's Cafe

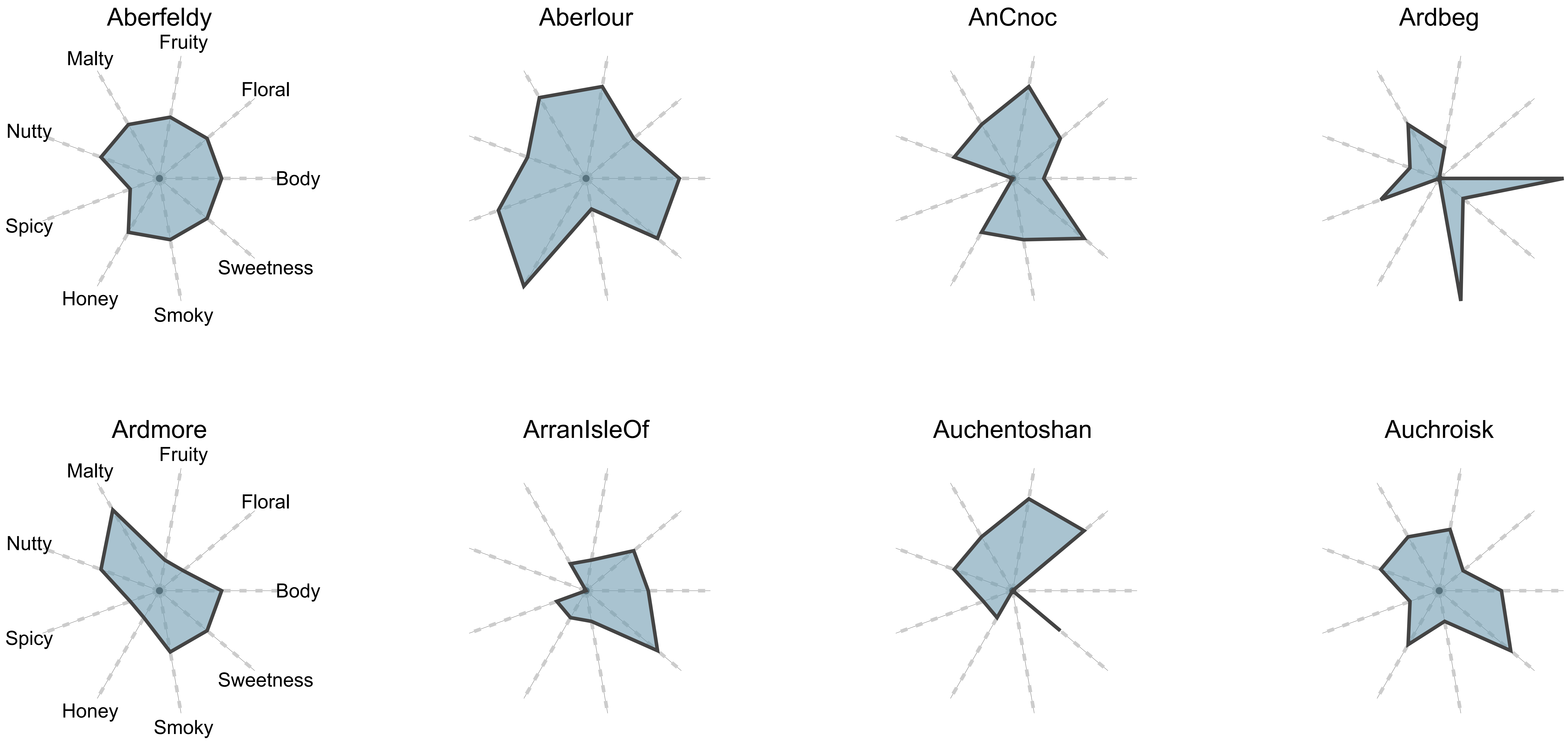
Grade **Grade pending**

Violation points **27**

Click for details

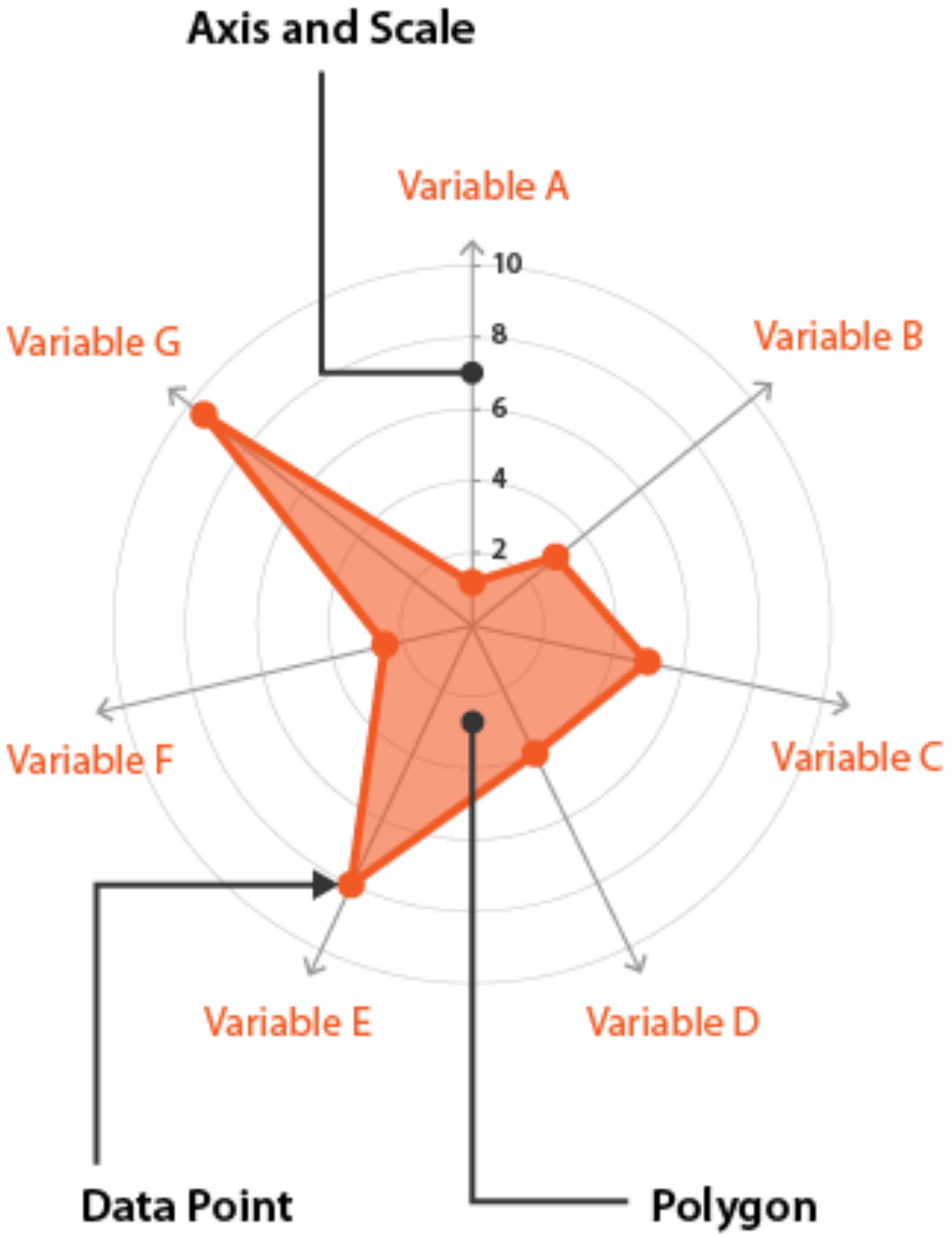
[J. White, New York Times]

Star Plots (aka Radar Charts)



[K. Schaul]

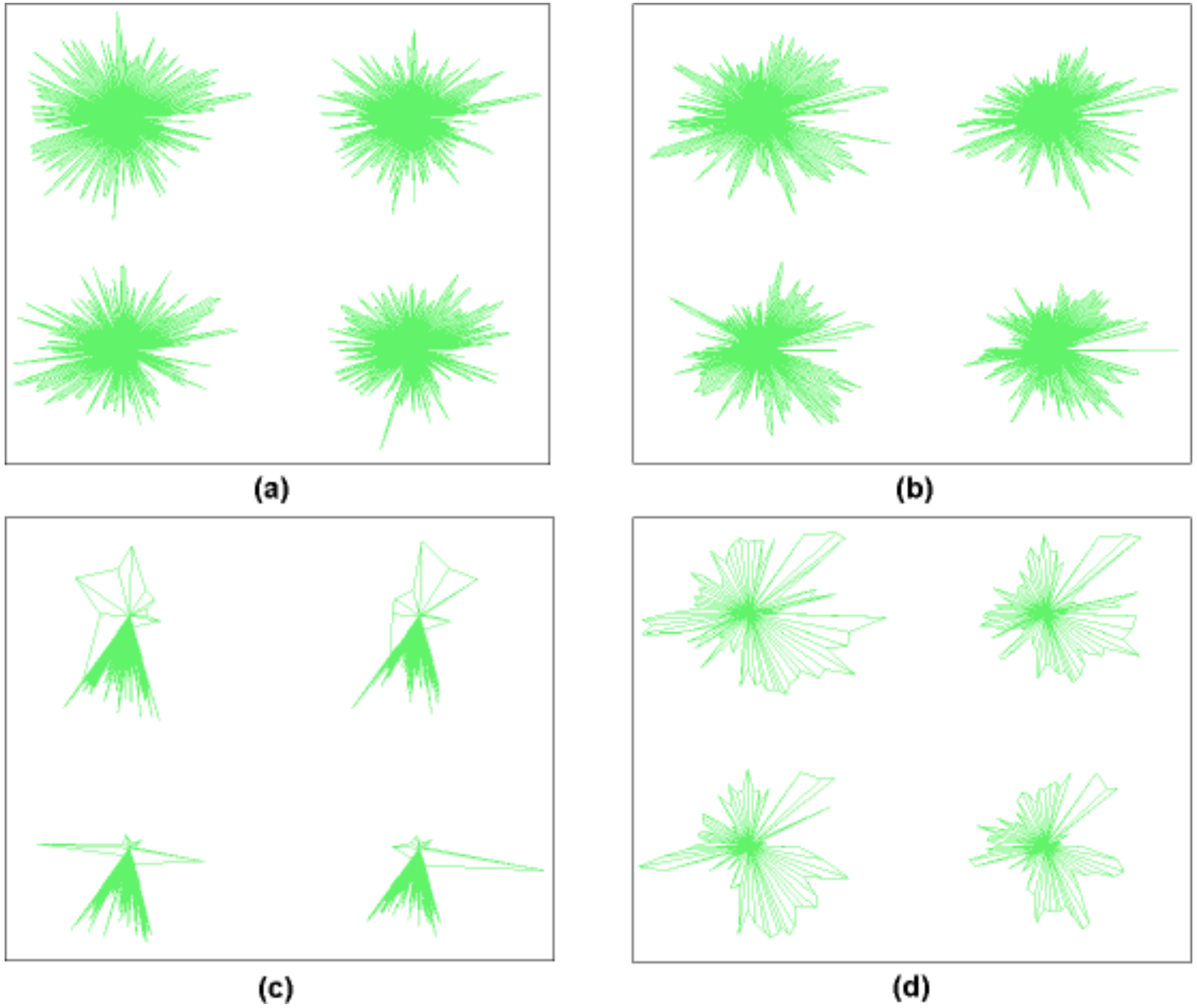
Star Plot / Radar Chart



- Use:
  - Compare variables
  - Similarities/differences of items
  - Locate outliers
- Considerations:
  - Order of axes
  - Too many axes cause problems

[S. Ribecca]

Attribute Filtering on Star Plots



(a)

(b)

(c)

(d)

[Yang et al., 2003]

Attribute Filtering

- How to choose which attributes should be filtered?
  - User selection?
  - Statistics: similarity measures, attributes with low variance are not as interesting when comparing items
- Can be combined with item filtering

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 Friday

Assignment 5

- Focus on Multiple Views and Interaction
- Soon…

Monday

- I am at a workshop so **no in-person lecture**
- Video lecture
- Assignment 5 will have been released

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

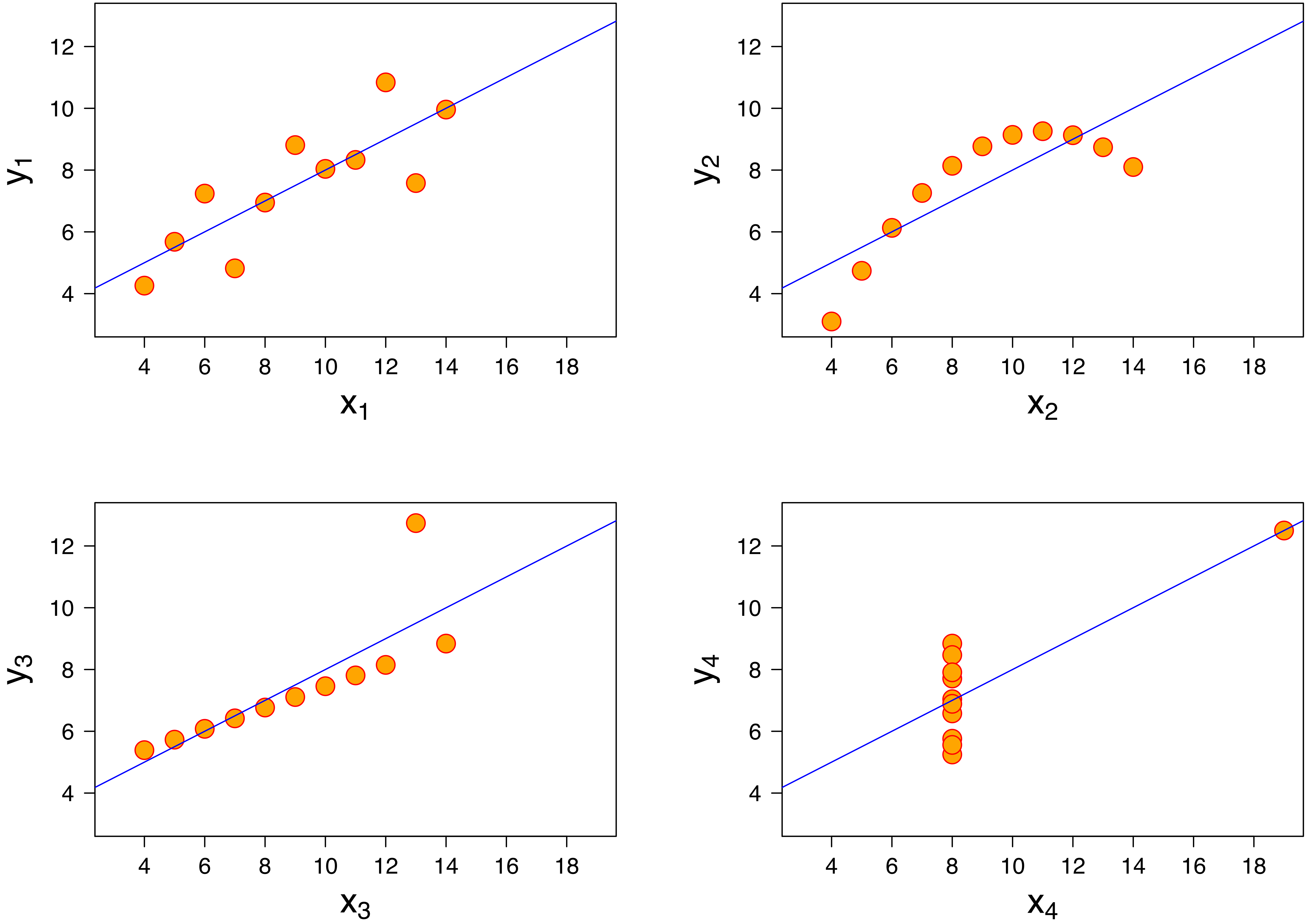
Aggregation

- 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

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

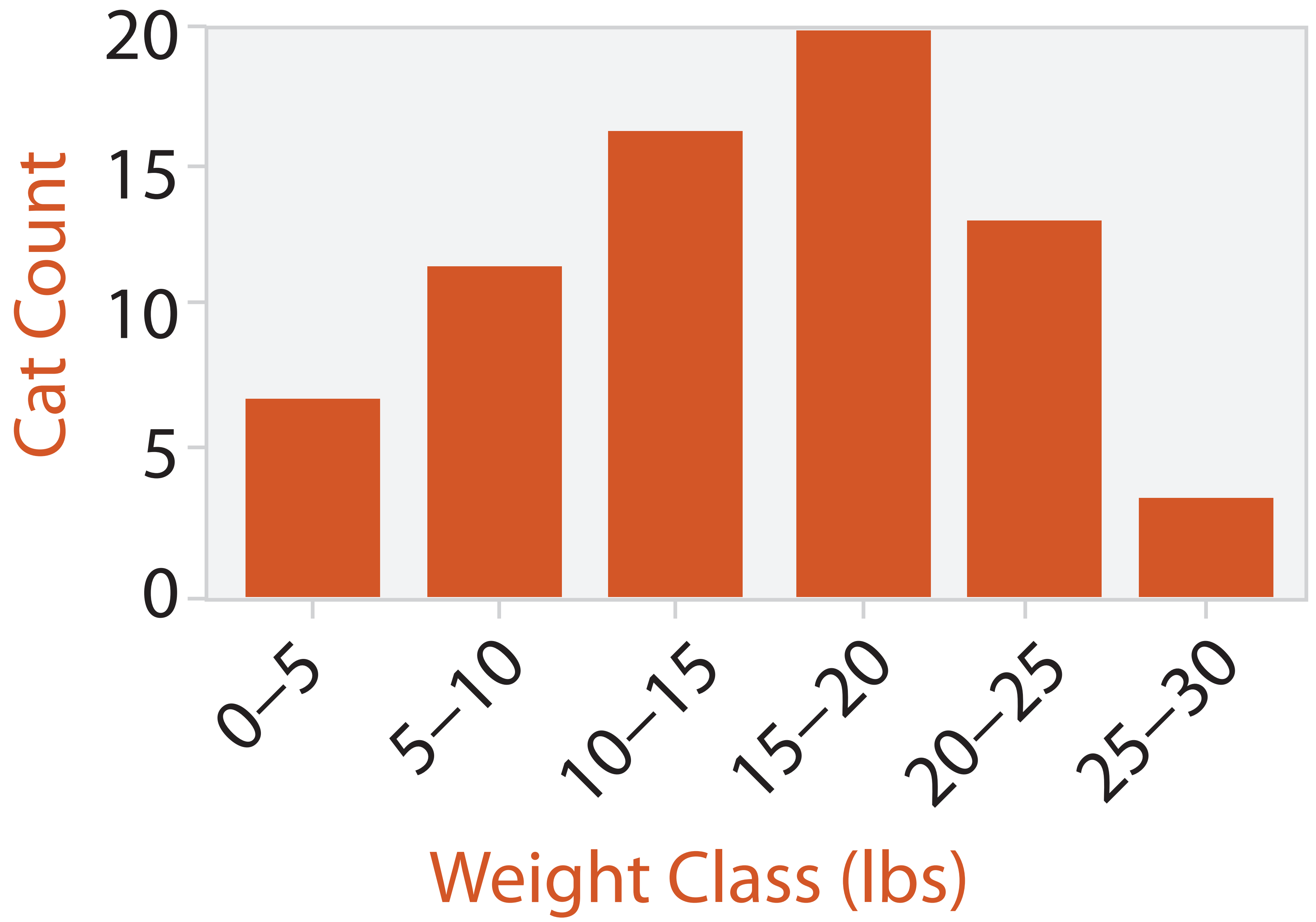
Mean of x	9
Variance of x	11
Mean of y	7.50
Variance of y	4.122
Correlation	0.816

Anscombe's Quartet



[F. J. Anscombe]

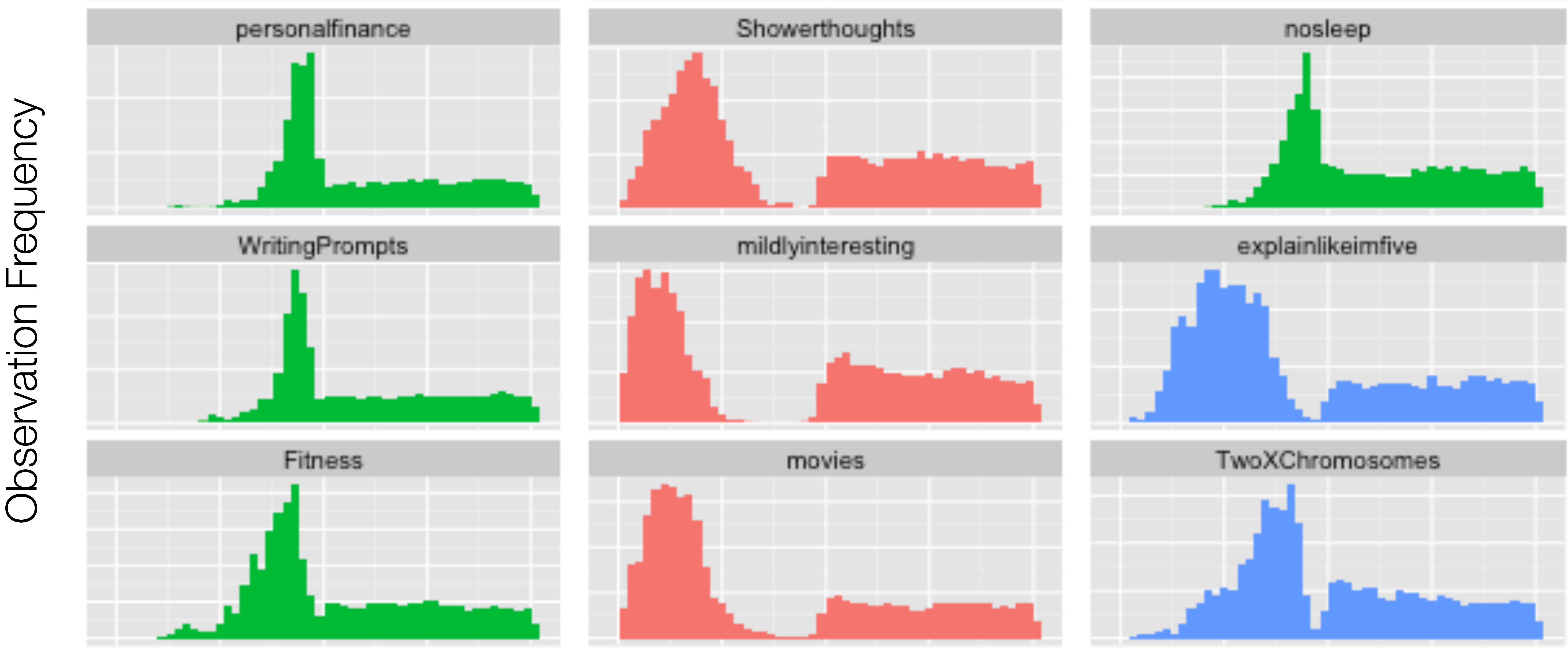
Aggregation: Histograms



- 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

[Munzner (ill. Maguire), 2014]

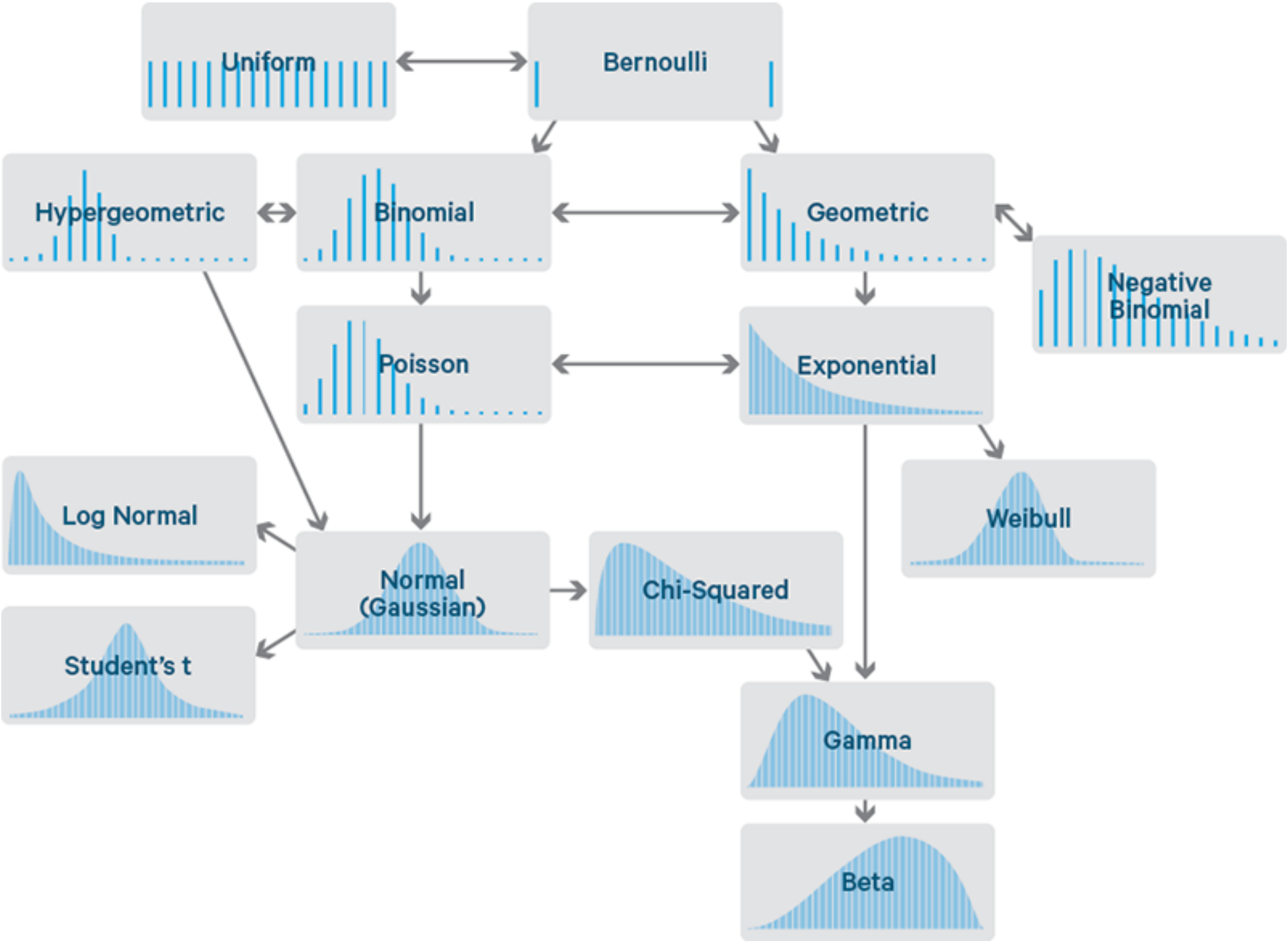
Aggregation: Histograms



Observed ranks of posts by subreddit

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

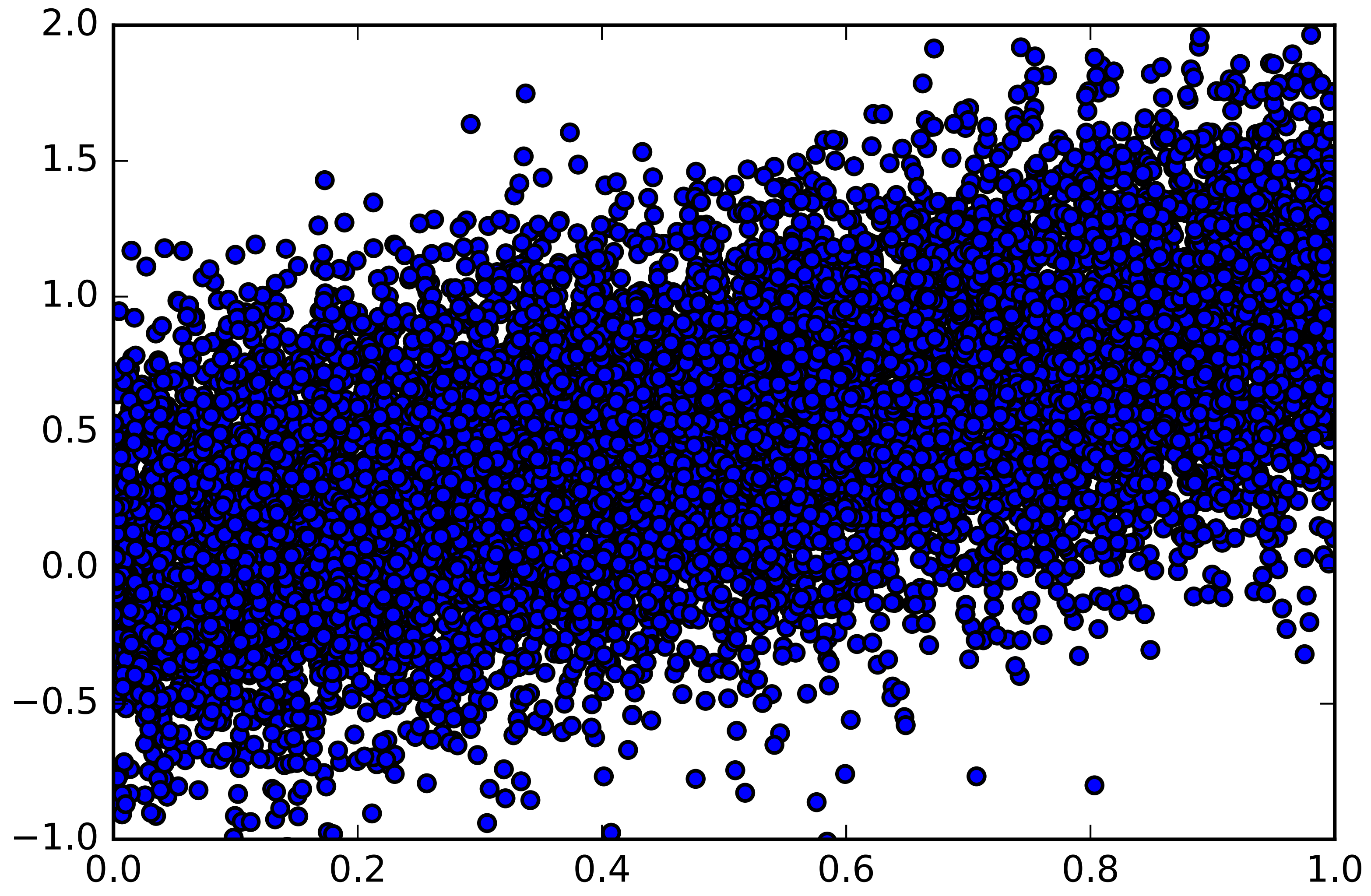
Common Distributions

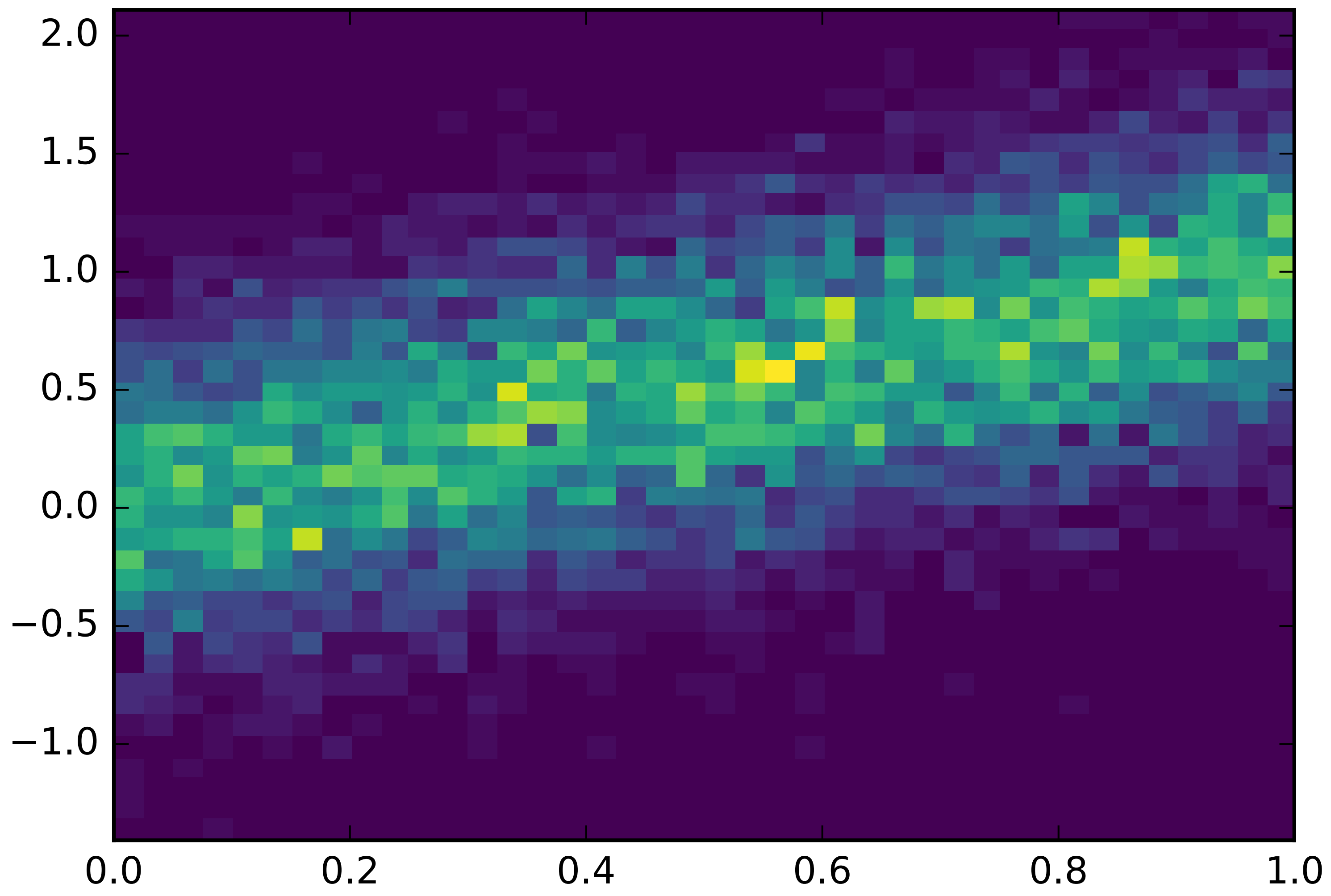


[Cloudera]

Binning Scatterplots

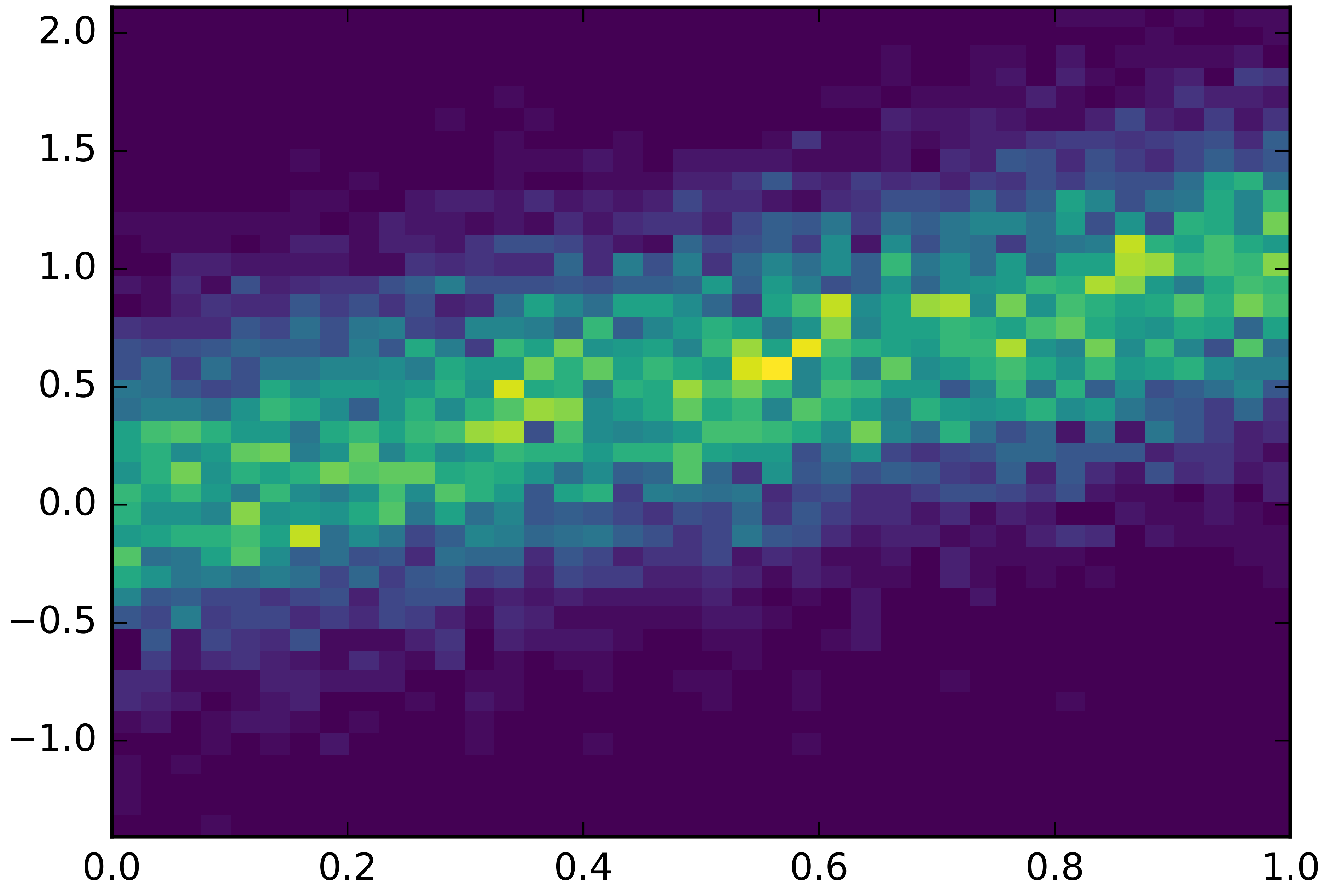
- At some point, cannot see density
- Blobs on top of blobs
- 2D Histogram is a histogram in 2D encoded using color instead of height
- Each region is aggregated

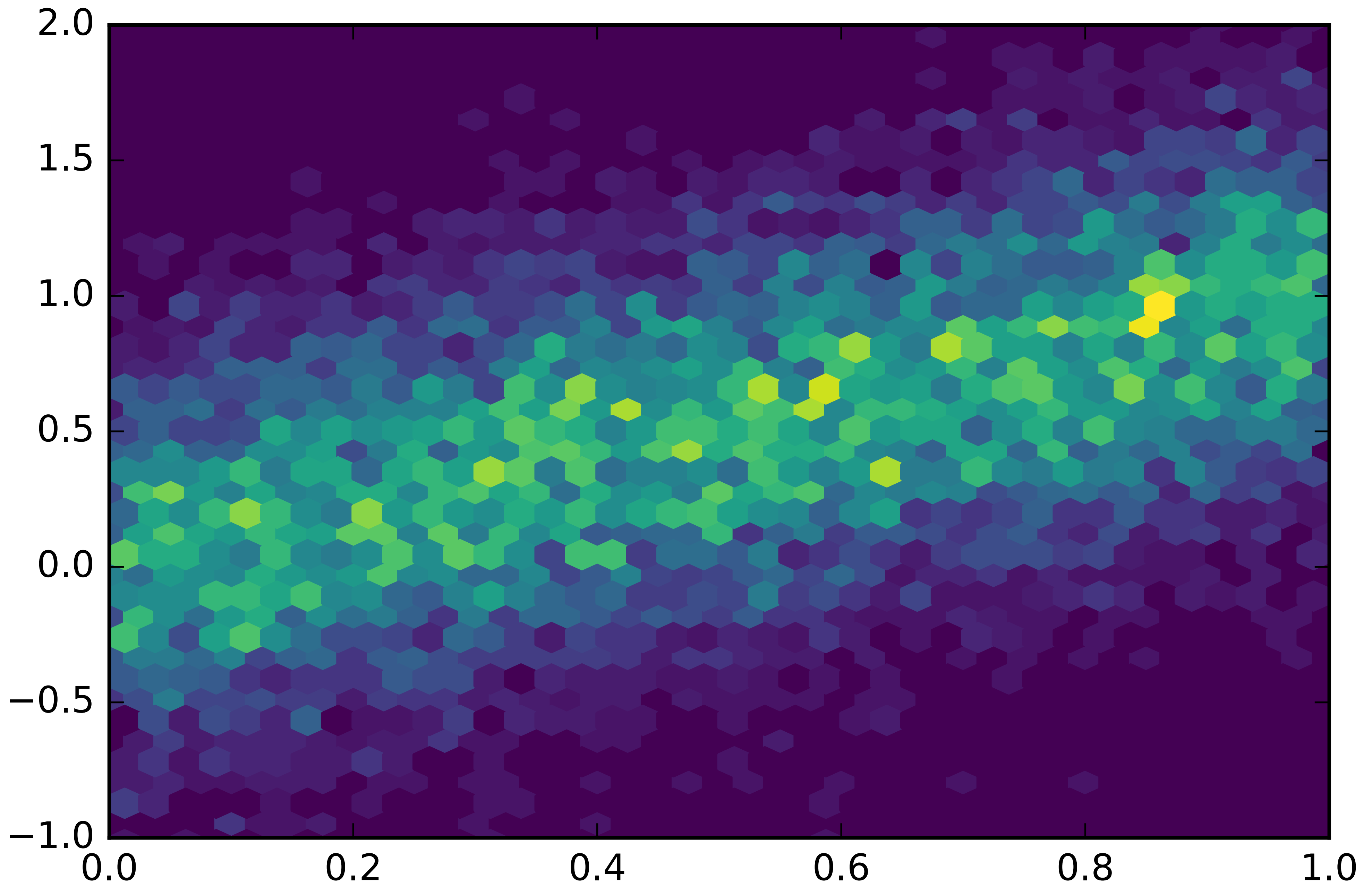




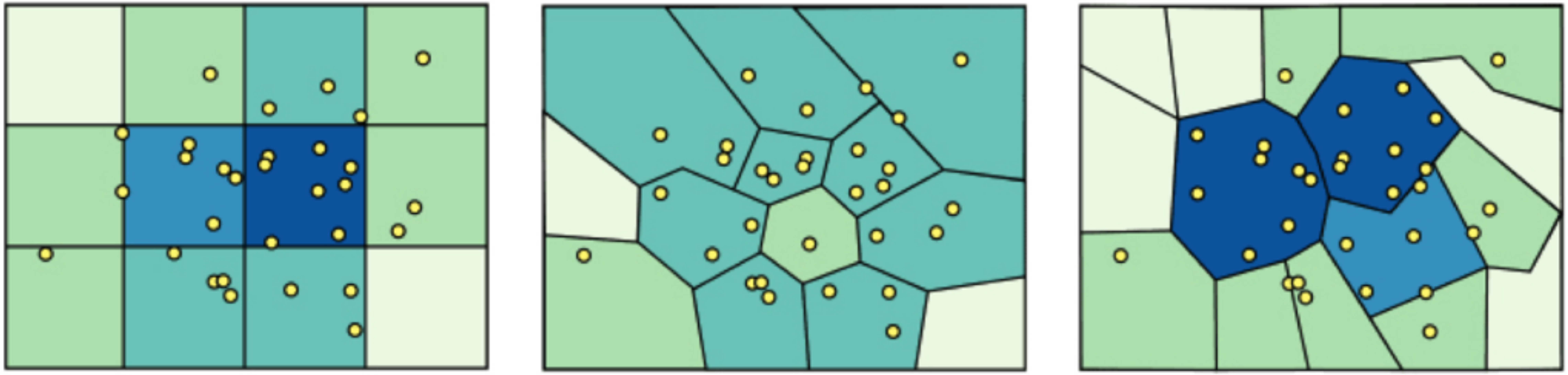
Binning

- Hexagonal bins are more circular
- Distance to the edge is not as variable
- More efficient aggregation around the center of the bin





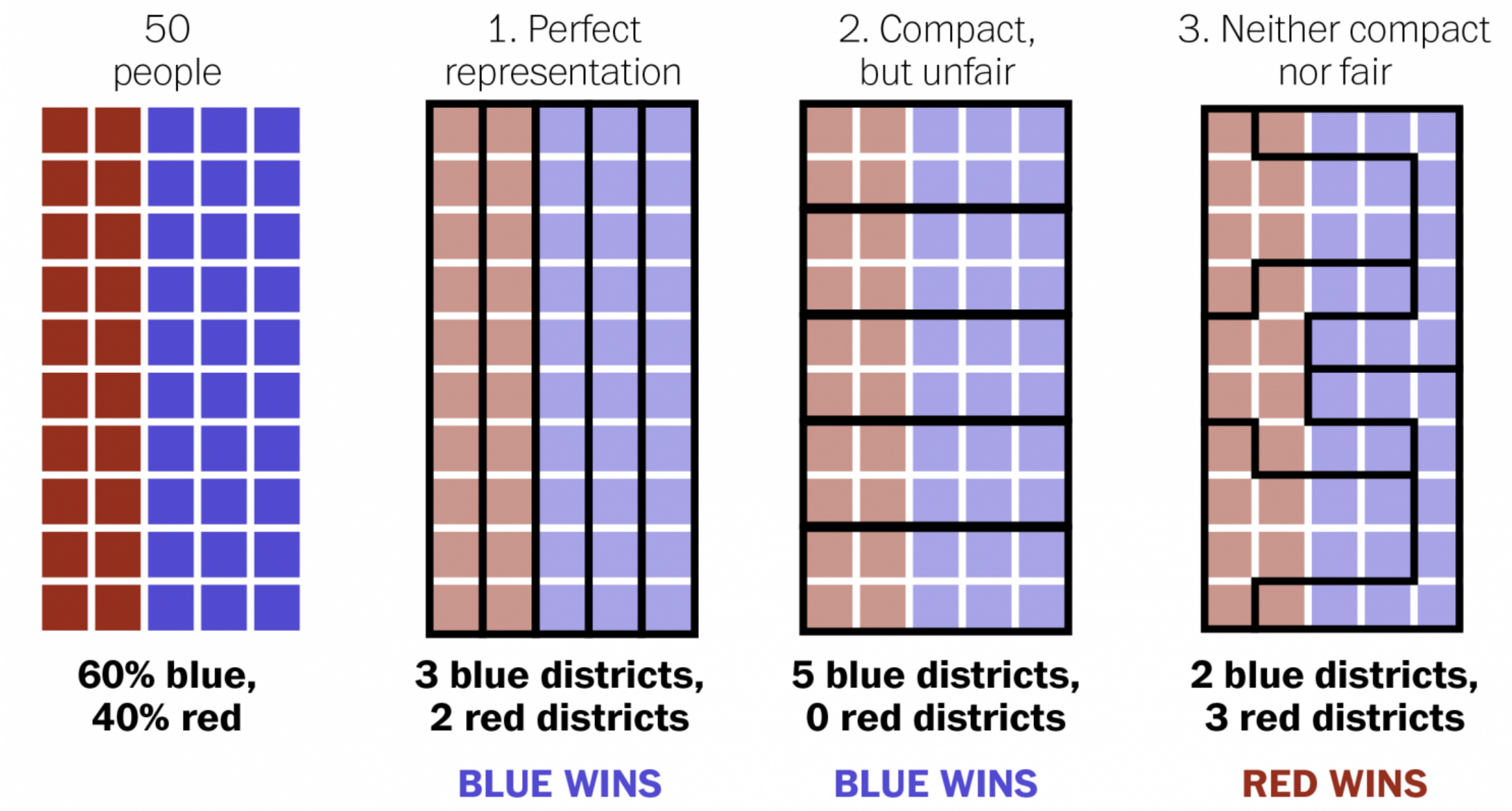
Spatial Aggregation

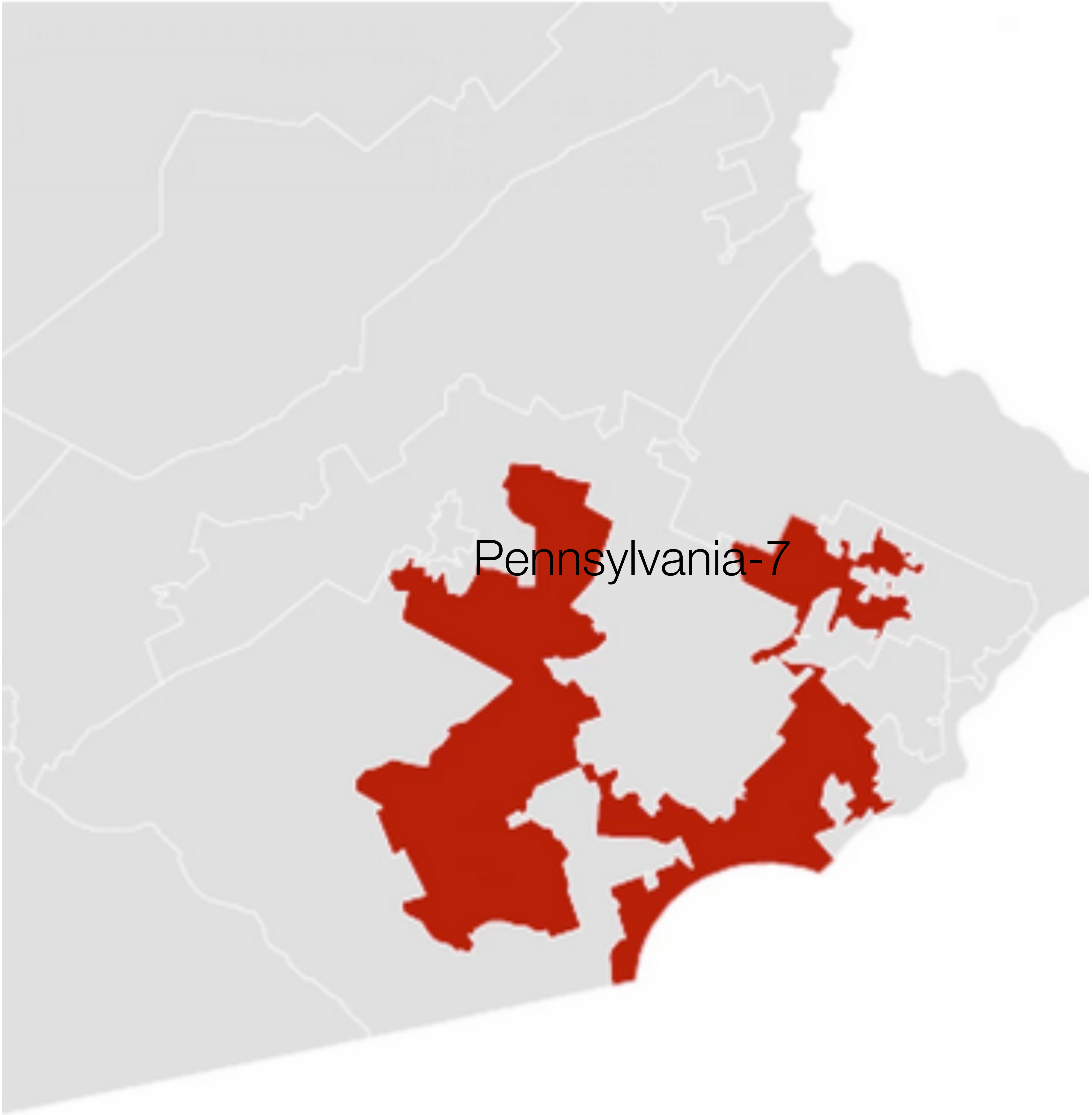


[Penn State, GEOG 486]

Modifiable Areal Unit Problem

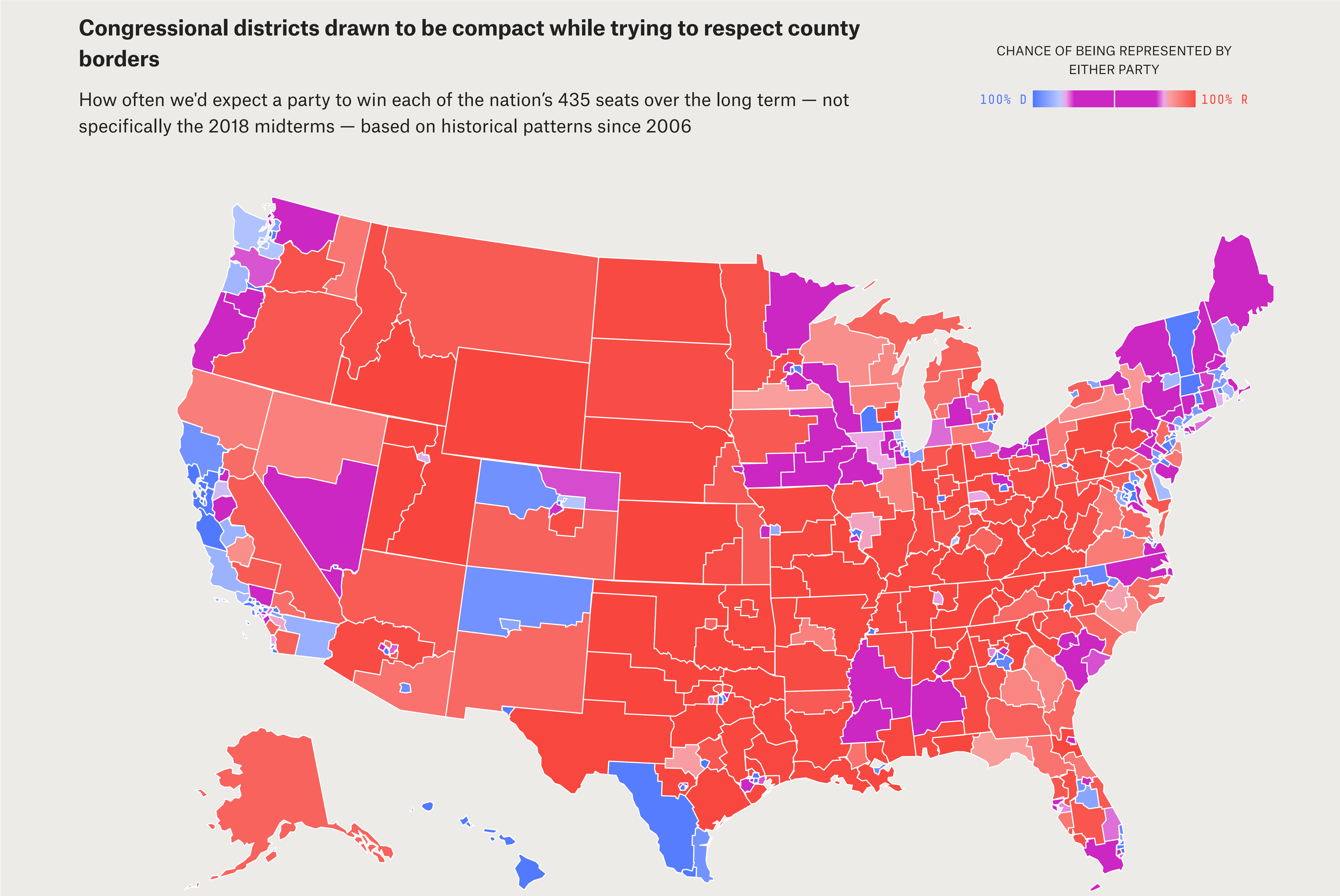
- How you draw boundaries impacts the type of aggregation you get
- Similar to bins in histograms
- Gerrymandering


50 people
1. Perfect representation
2. Compact, but unfair
3. Neither compact nor fair
60% blue, 40% red
3 blue districts, 2 red districts
BLUE WINS
5 blue districts, 0 red districts
BLUE WINS
2 blue districts, 3 red districts
RED WINS



[Wonkblog, Washington Post, Adapted from S. Nass]

Drawing Different Maps: Compactness

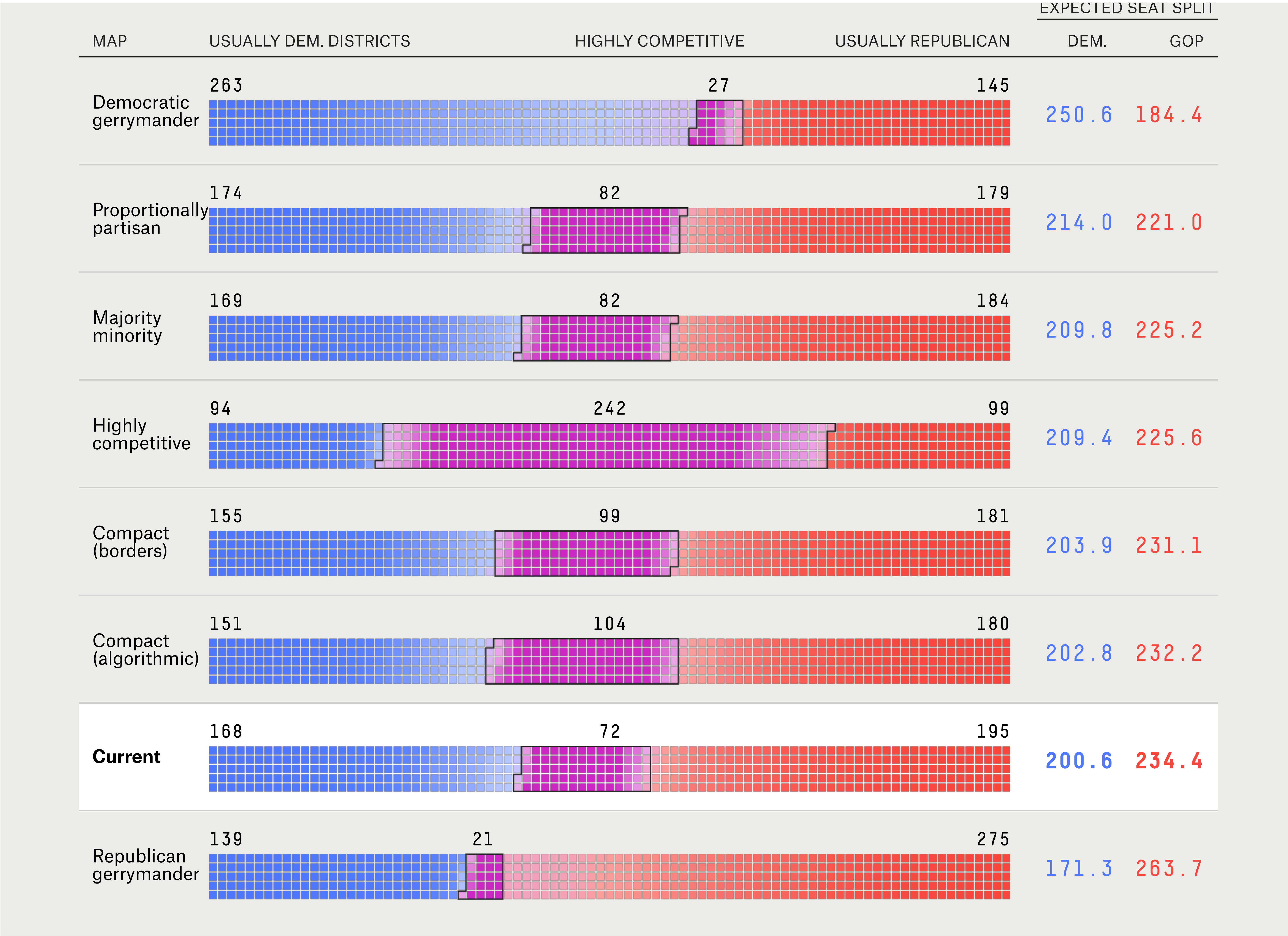


Congressional districts drawn to be compact while trying to respect county borders

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

[A. Bycoffe et al., 538]

Drawing Different Maps

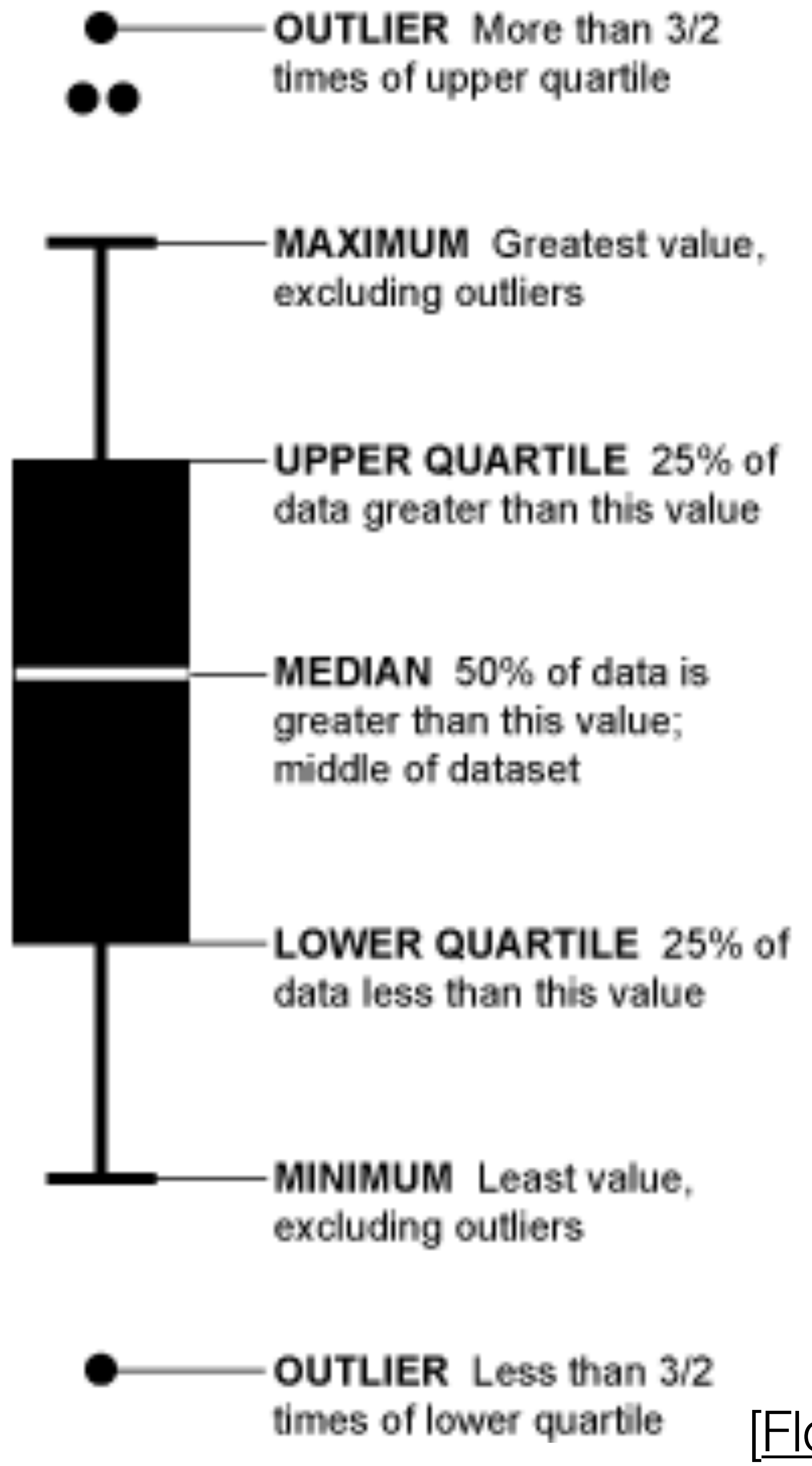


				EXPECTED SEAT SPLIT	
MAP	USUALLY DEM. DISTRICTS	HIGHLY COMPETITIVE	USUALLY REPUBLICAN	DEM.	GOP
Democratic gerrymander	263	27	145	250.6	184.4
Proportionally partisan	174	82	179	214.0	221.0
Majority minority	169	82	184	209.8	225.2
Highly competitive	94	242	99	209.4	225.6
Compact (borders)	155	99	181	203.9	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

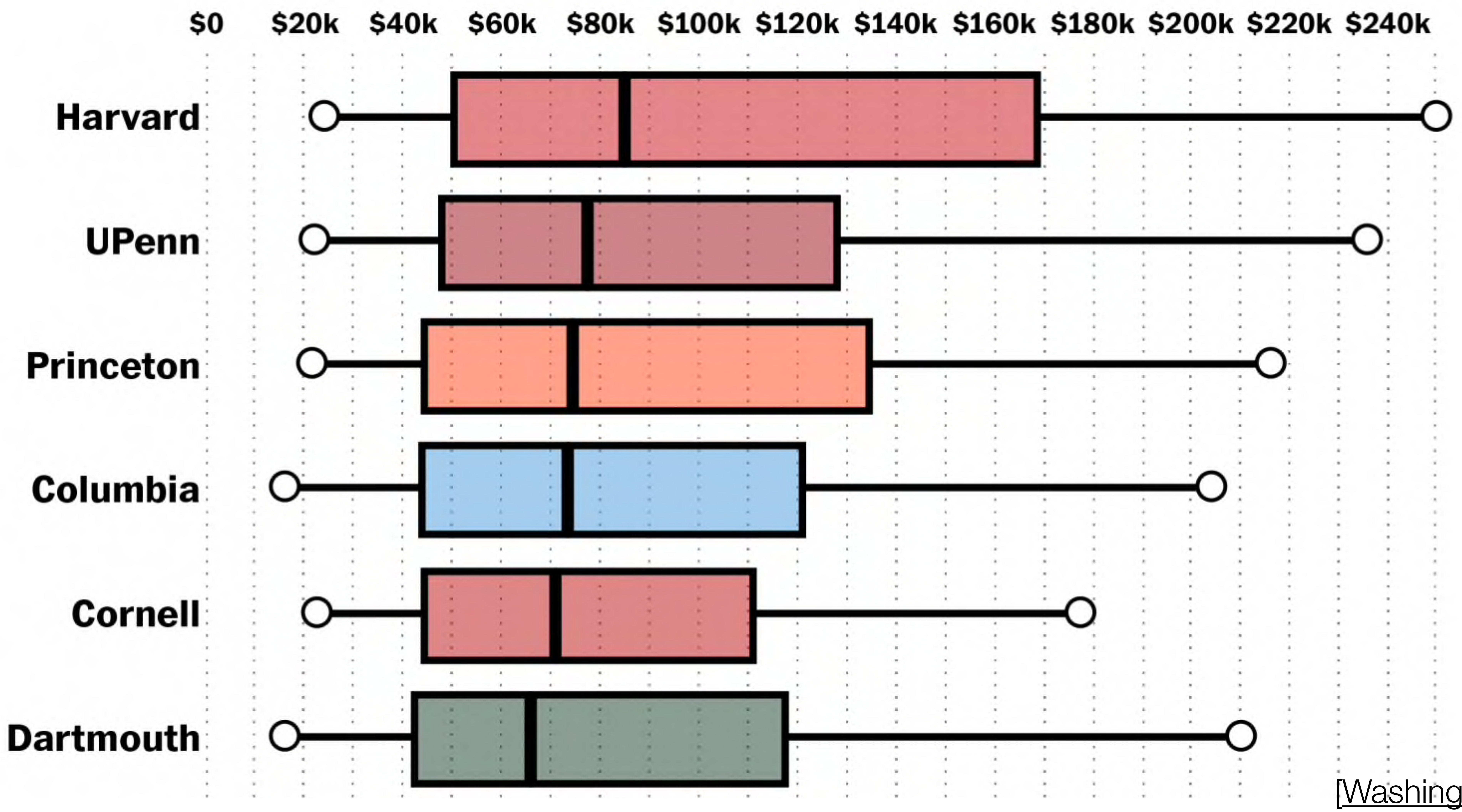
- Show **distribution**
- Single value (e.g. mean, max, min, quartiles) doesn't convey everything
- Created by John Tukey
- Show **spread** and **skew** of data
- Best for **unimodal** data
- Variations like vase plot for multimodal data
- Aggregation here involves many different marks



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

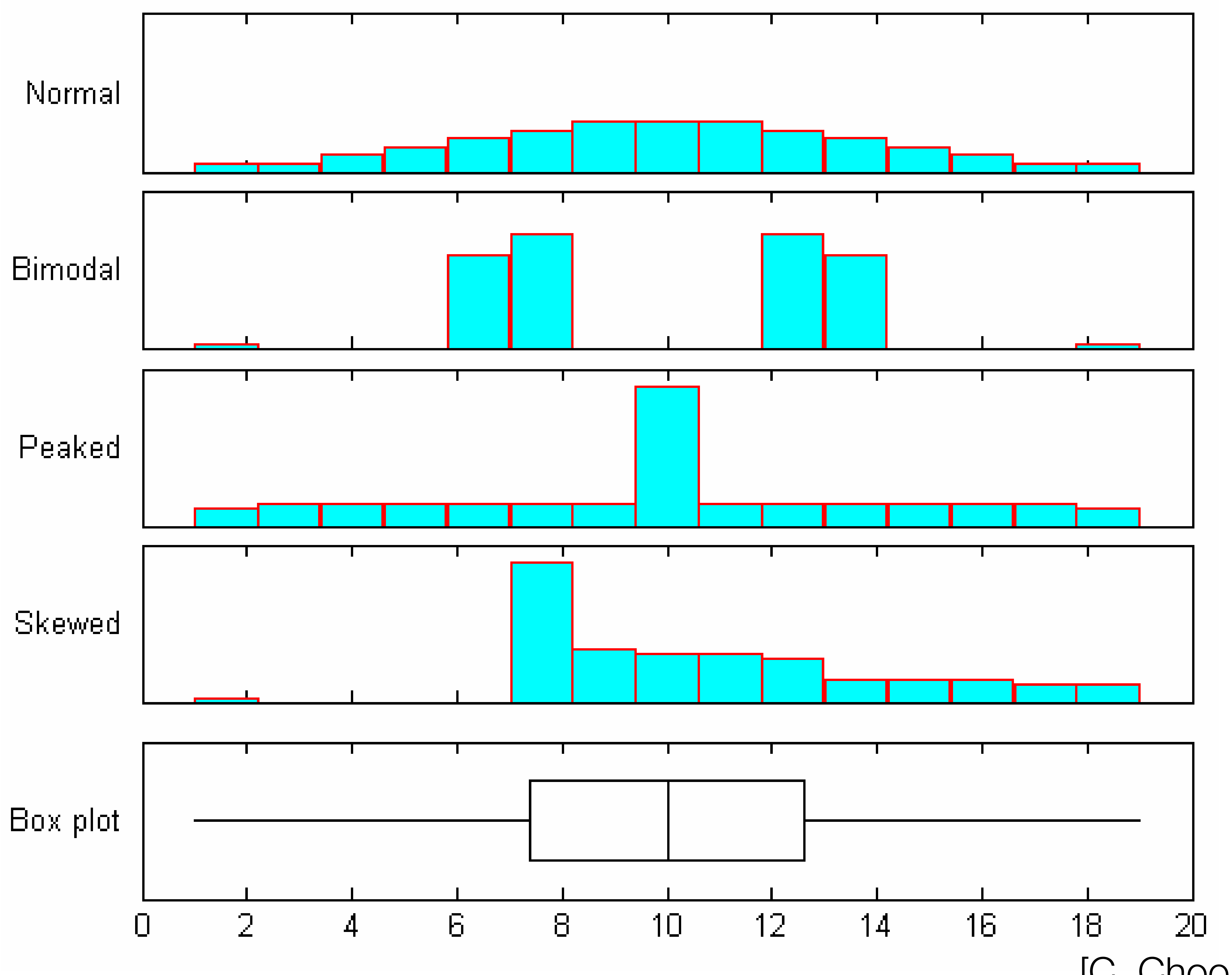
[Flowing Data]

Aggregation: Boxplots



[Washington Post, 2015]

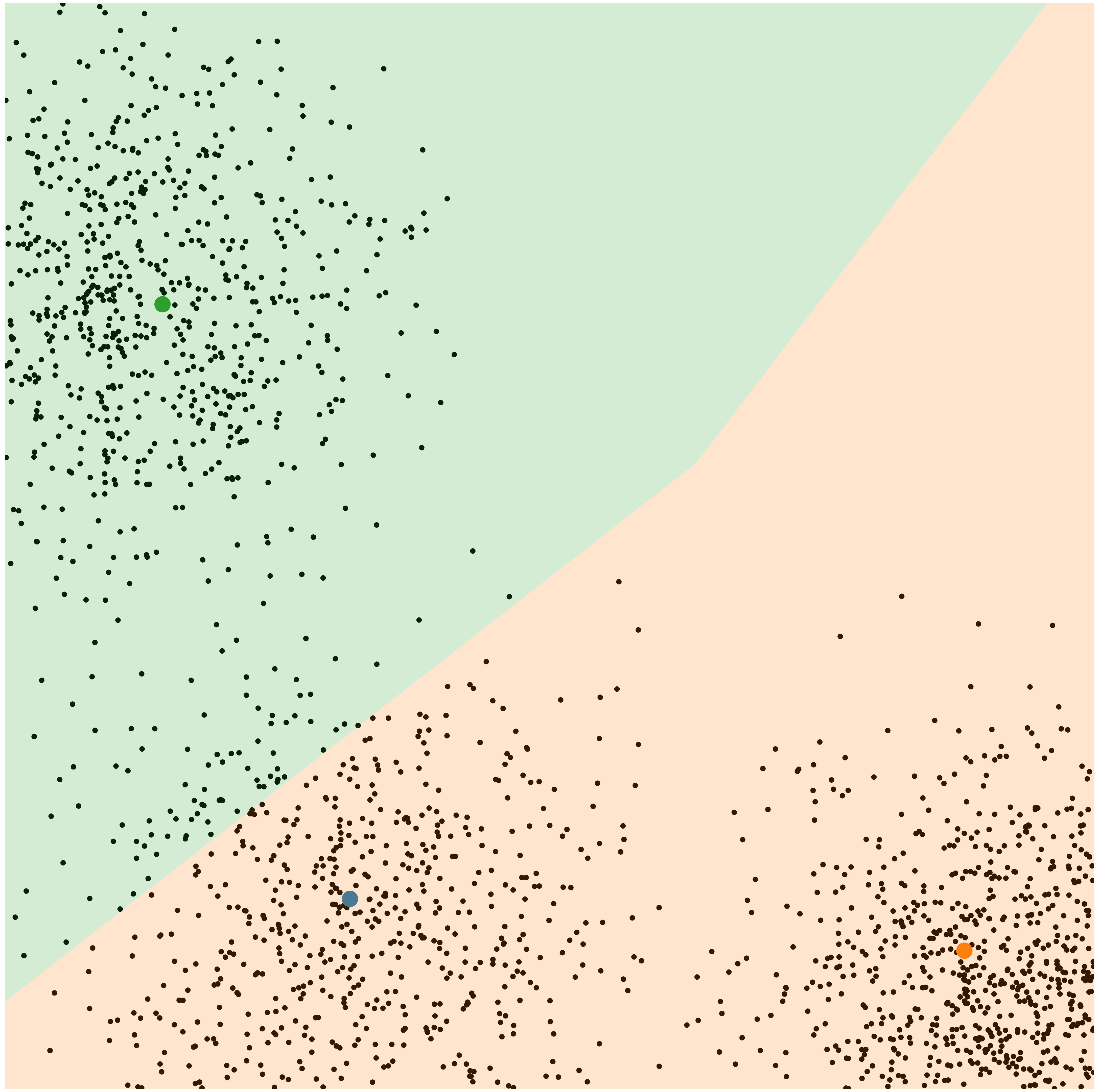
Four Distributions, Same Boxplot…



[C. Choonpradub and D. McNeil, 2005]

Attribute Aggregation

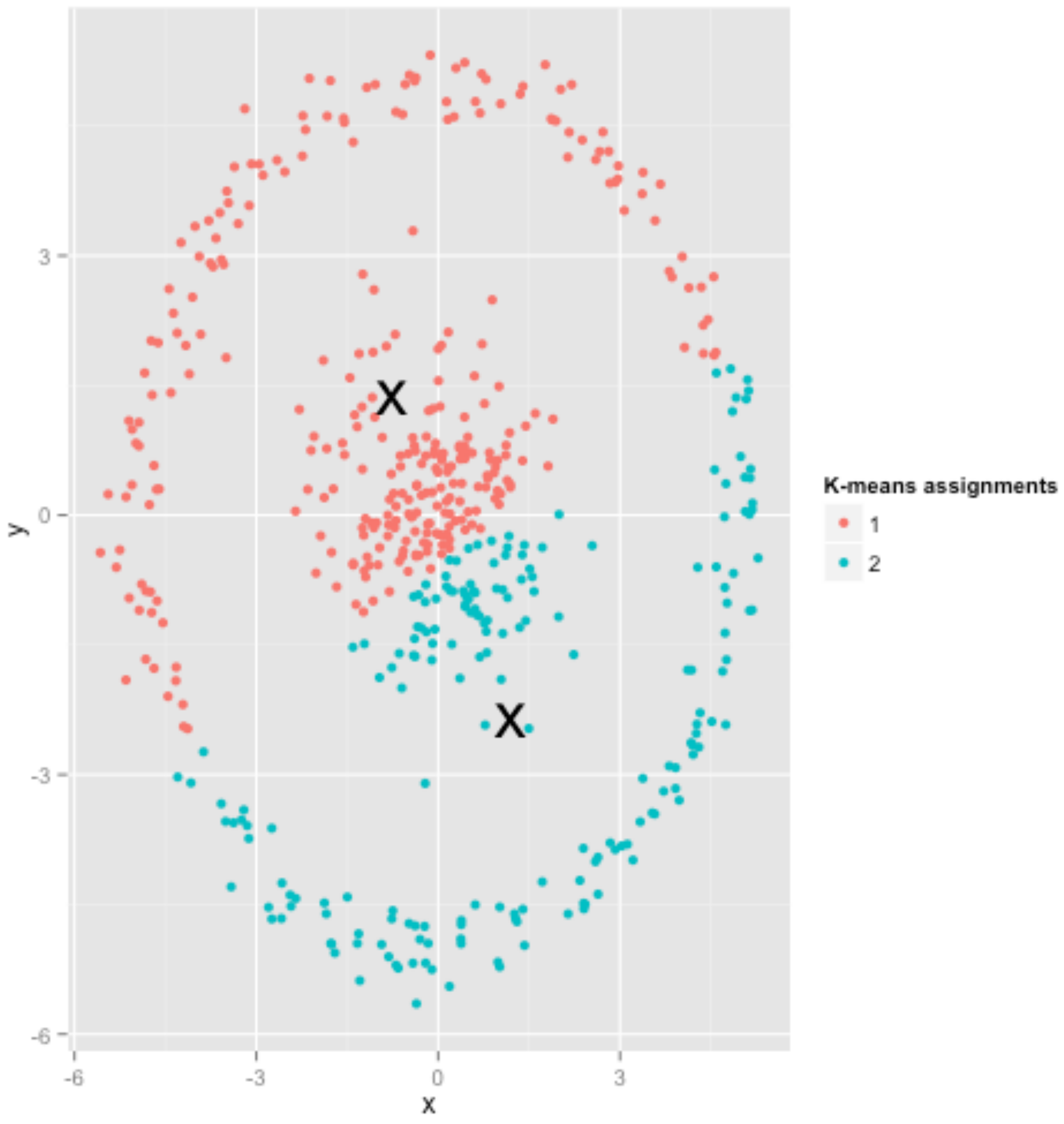
K-Means



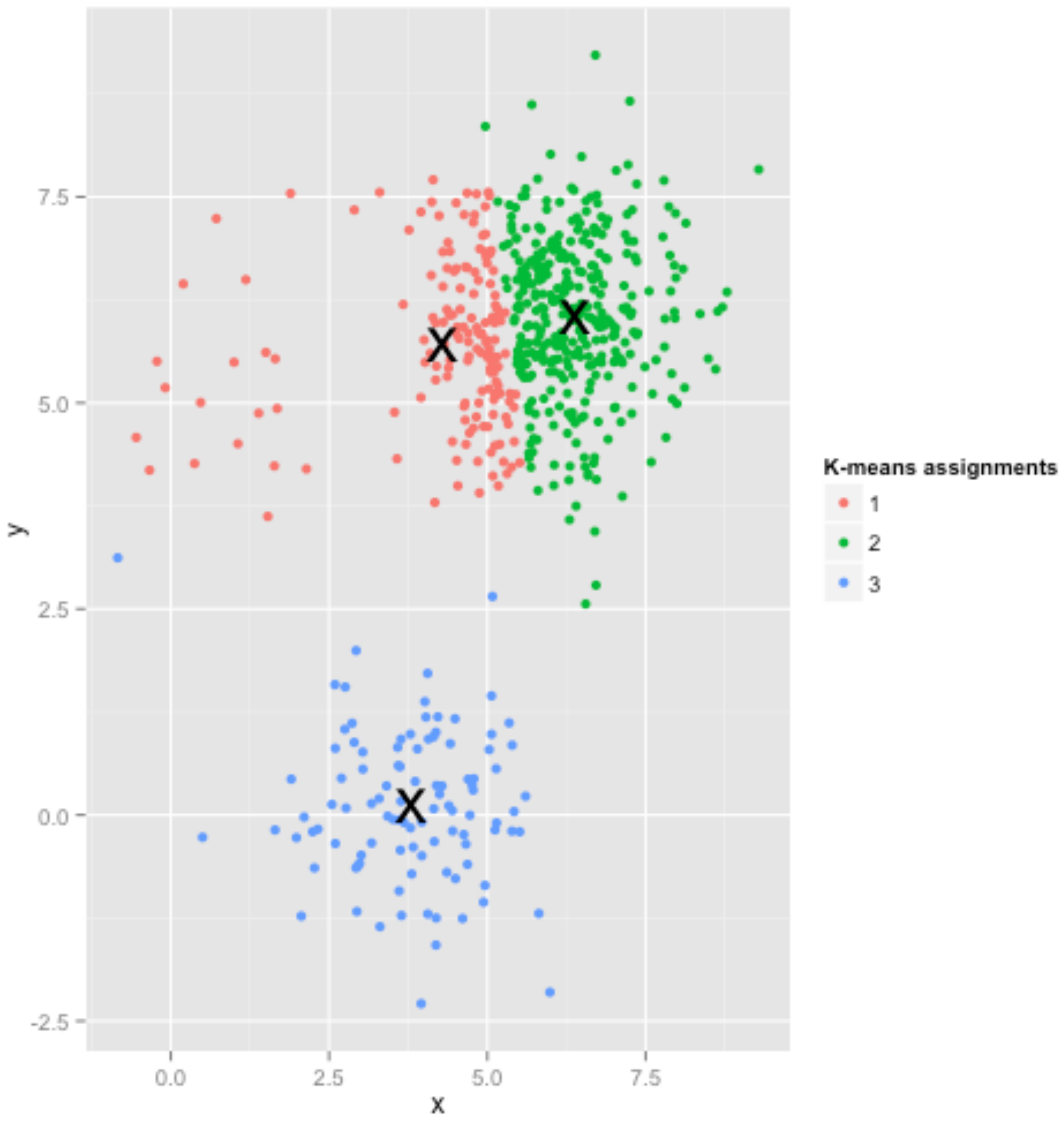
Run

[C. Polis, 2014]

K-Means Issues



Shape



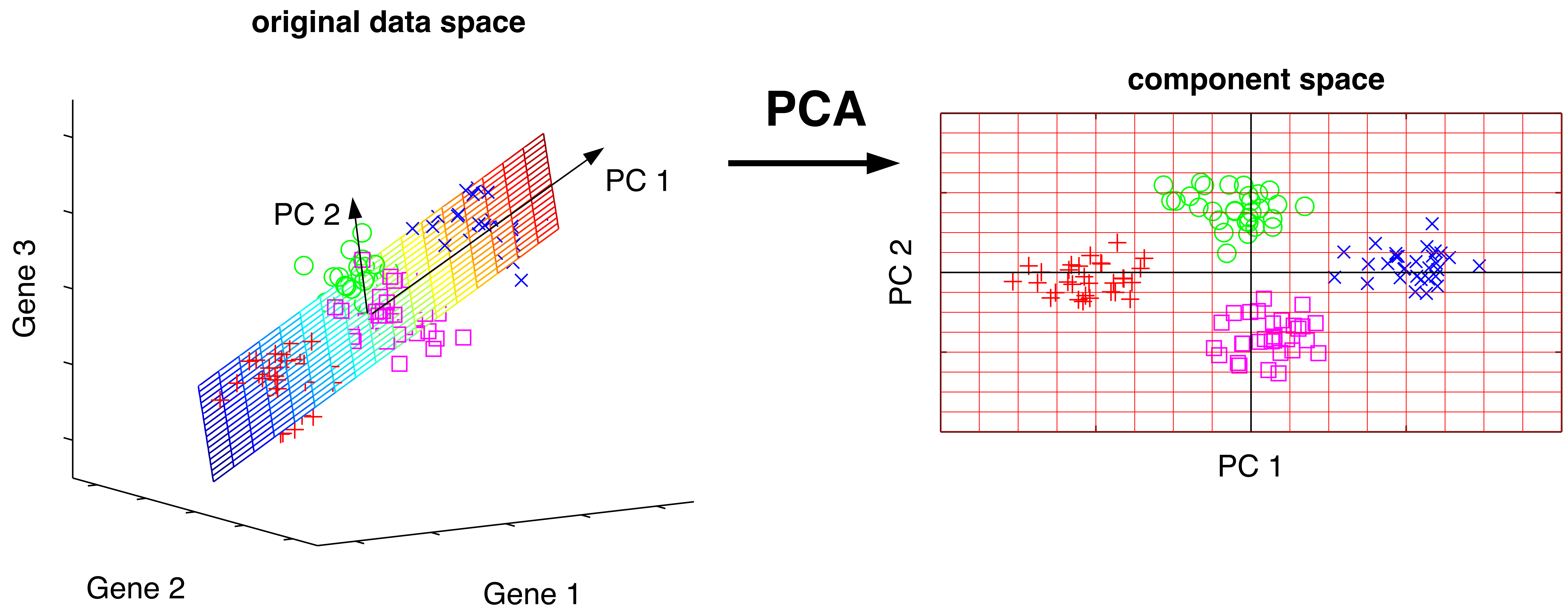
Number of Clusters

[D. Robinson, 2015]

Dimensionality Reduction

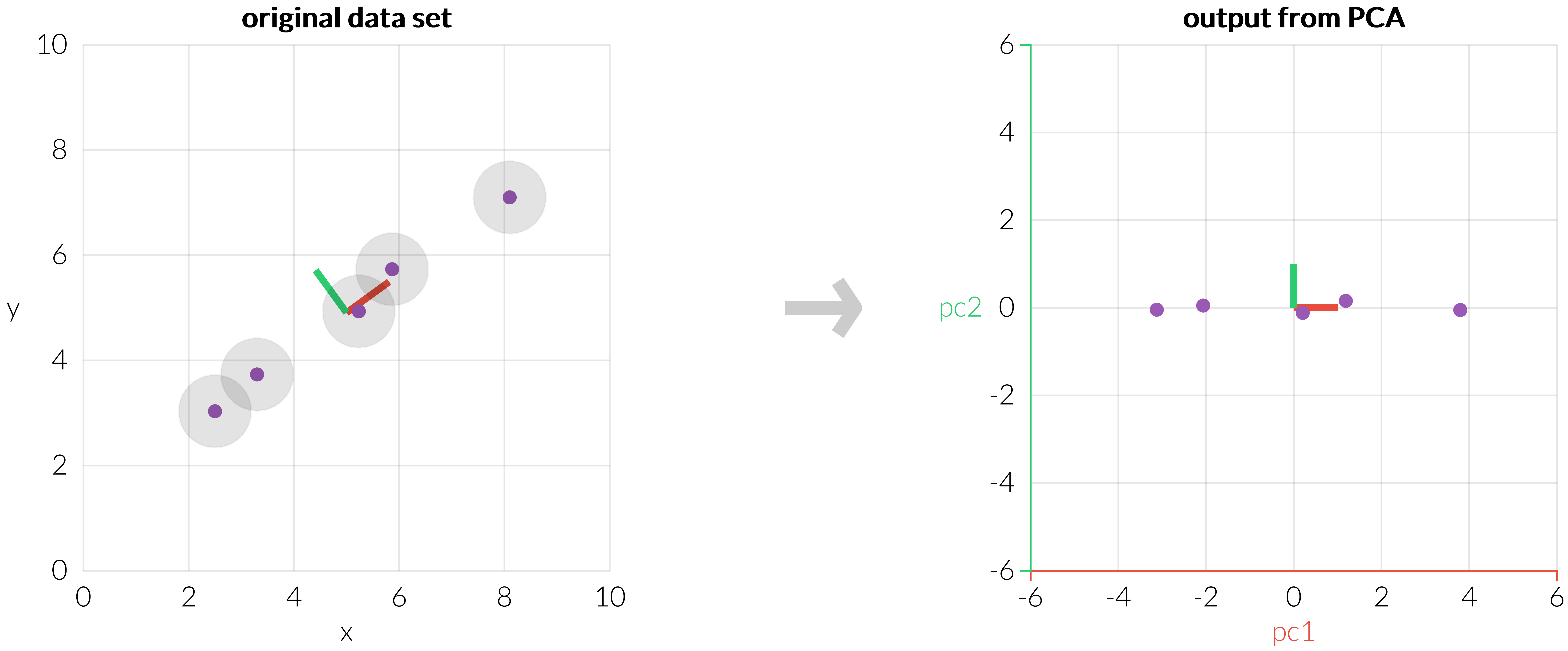
- Attribute Aggregation: Use fewer attributes (dimensions) to represent items
- Combine attributes in a way that is more instructive than examining each individual attribute
- Example: Understanding the language in a collection of books
  - Count the occurrence of each non-common word in each book
  - 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

Principle Component Analysis (PCA)



[M. Scholz, CC-BY-SA 2.0]

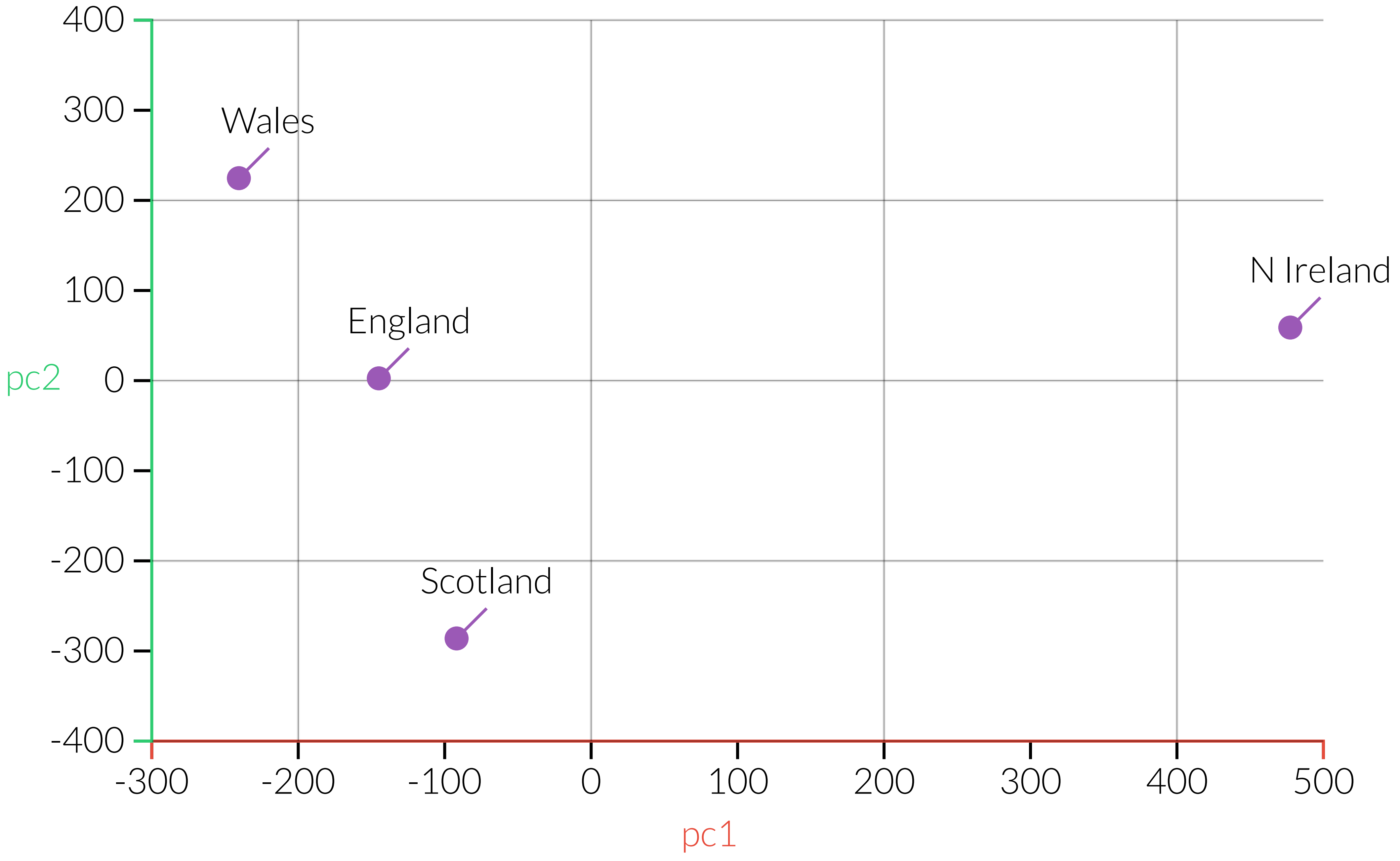
PCA



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

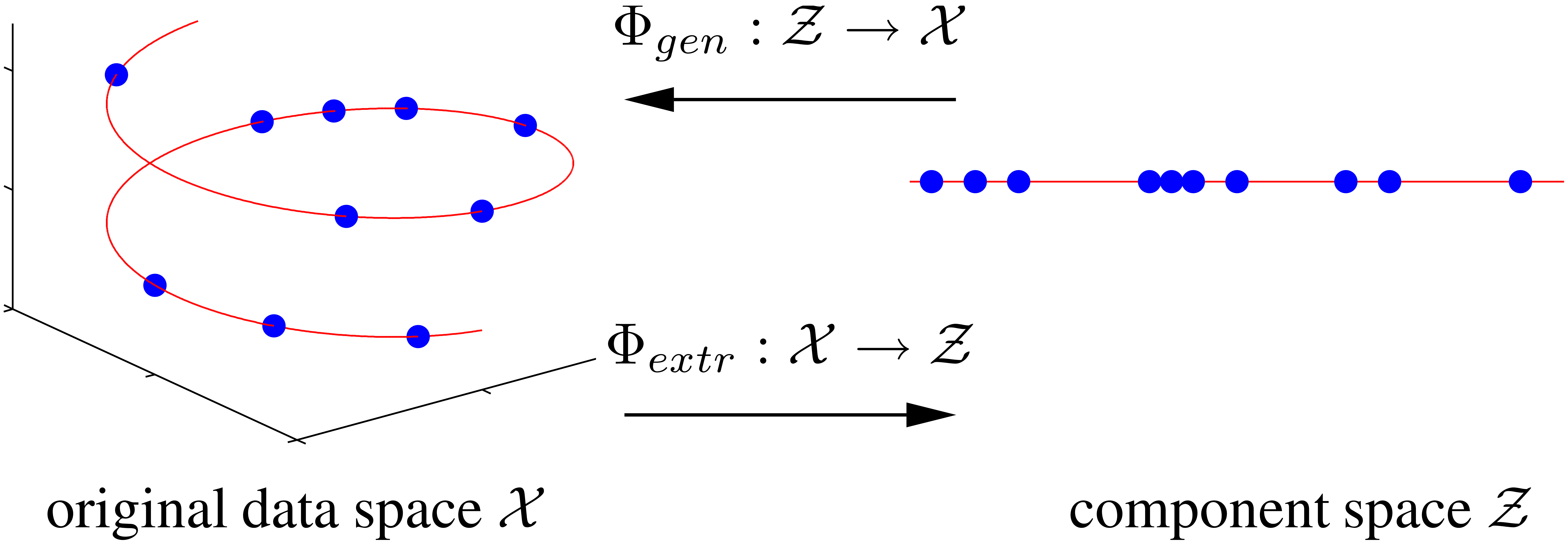
17 dimensions to 2

	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



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

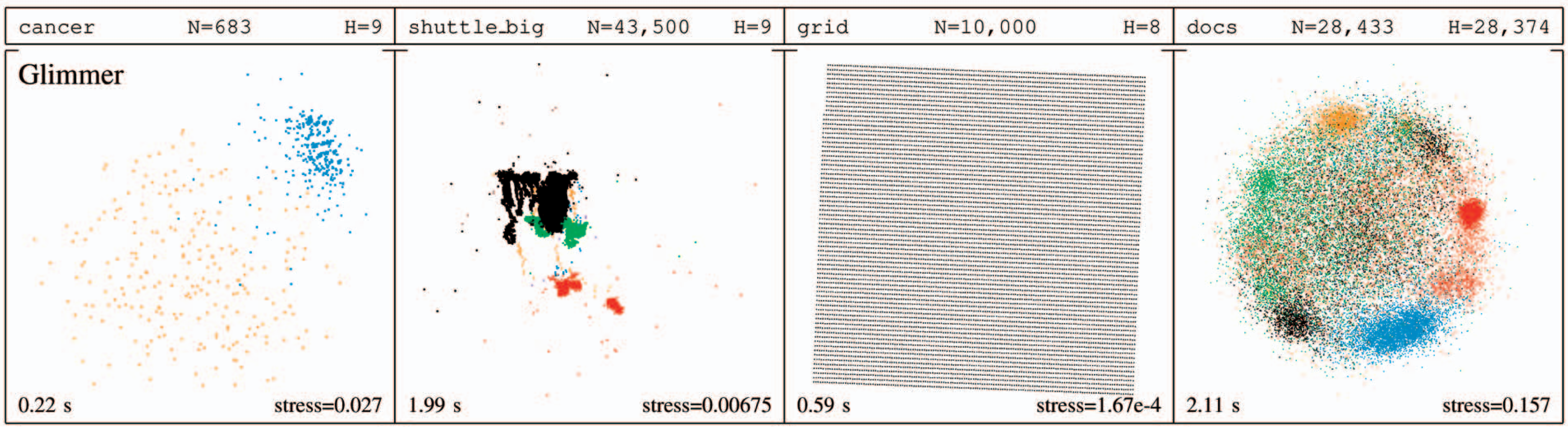
$\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$

original data space $\mathcal{X}$

component space $\mathcal{Z}$

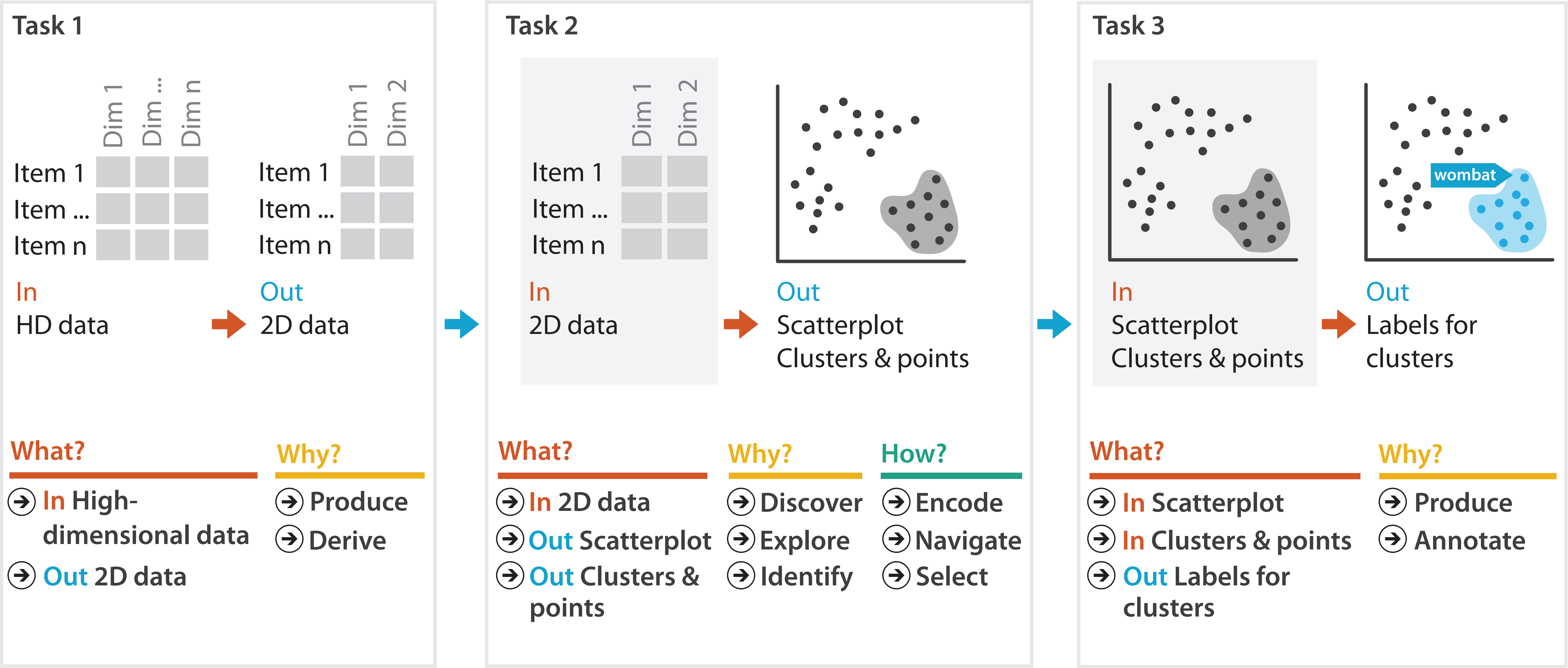
[M. Scholz, CC-BY-SA 2.0]

Dimensionality Reduction in Visualization



[Glimmer, Ingram et al., 2009]

Tasks in Understanding High-Dim. Data



Task 1

In
HD data

Out
2D data

What?
- In High-dimensional data
- Out 2D data

Why?
- Produce
- Derive

Task 2

In
2D data

Out
Scatterplot
Clusters & points

What?
- In 2D data
- Out Scatterplot
- Out Clusters & points

Why?
- Discover
- Explore
- Identify

How?
- Encode
- Navigate
- Select

Task 3

In
Scatterplot
Clusters & points

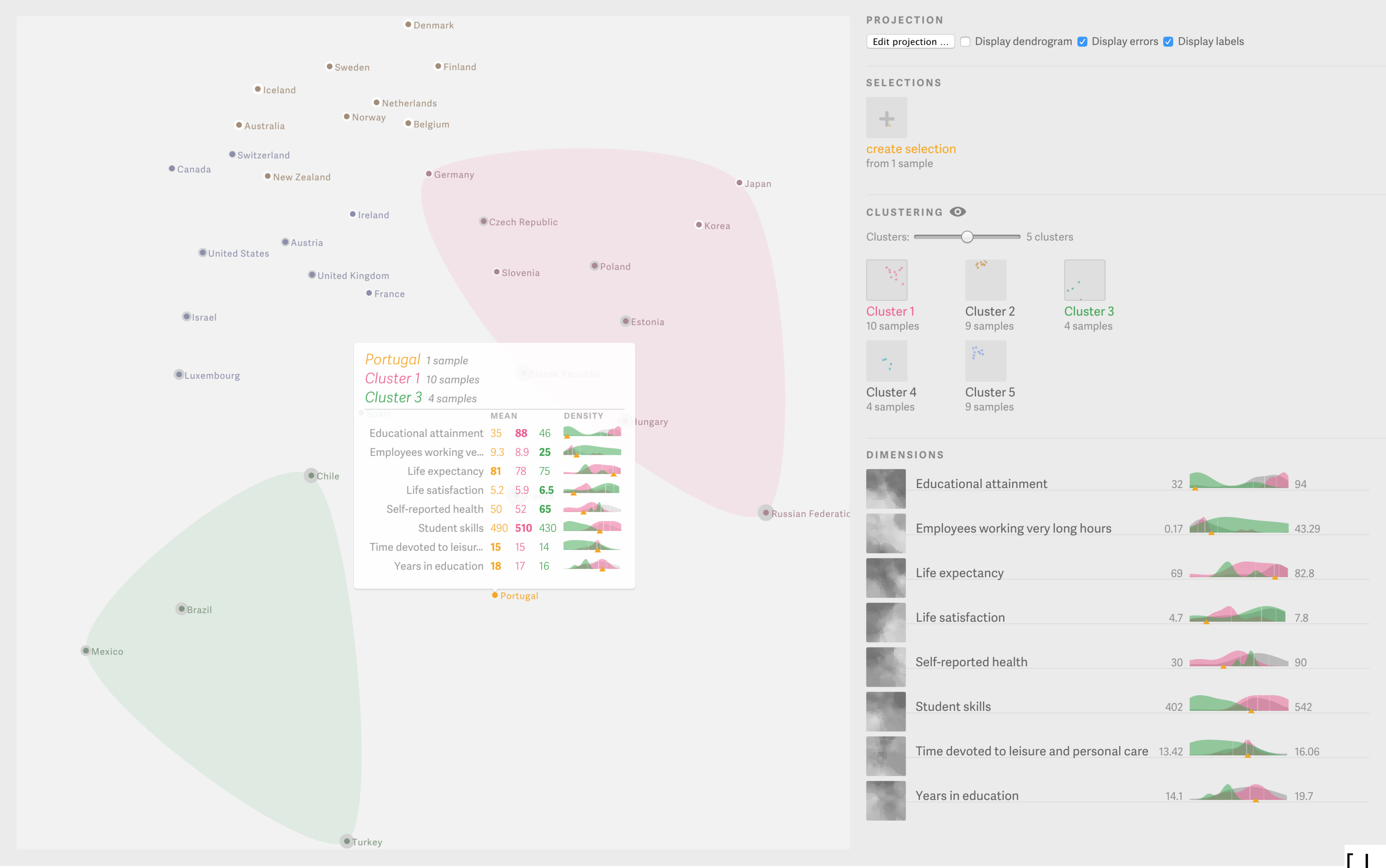
Out
Labels for clusters

What?
- In Scatterplot
- In Clusters & points
- Out Labels for clusters

Why?
- Produce
- Annotate

[Munzner (ill. Maguire), 2014]

Probing Projections



Portugal *1 sample*
Cluster 1 *10 samples*
Cluster 3 *4 samples*

	MEAN			DENSITY
Educational attainment	35	**88**	46	
Employees working ve...	9.3	8.9	**25**	
Life expectancy	**81**	78	75	
Life satisfaction	5.2	5.9	**6.5**	
Self-reported health	50	52	**65**	
Student skills	490	**510**	430	
Time devoted to leisur...	**15**	15	14	
Years in education	**18**	17	16	

PROJECTION

Edit projection ... □ Display dendrogram ☑ Display errors ☑ Display labels

SELECTIONS

create selection
from 1 sample

CLUSTERING

Clusters: 5 clusters

Cluster 1
10 samples

Cluster 2
9 samples

Cluster 3
4 samples

Cluster 4
4 samples

Cluster 5
9 samples

DIMENSIONS

Educational attainment	32	94
Employees working very long hours	0.17	43.29
Life expectancy	69	82.8
Life satisfaction	4.7	7.8
Self-reported health	30	90
Student skills	402	542
Time devoted to leisure and personal care	13.42	16.06
Years in education	14.1	19.7

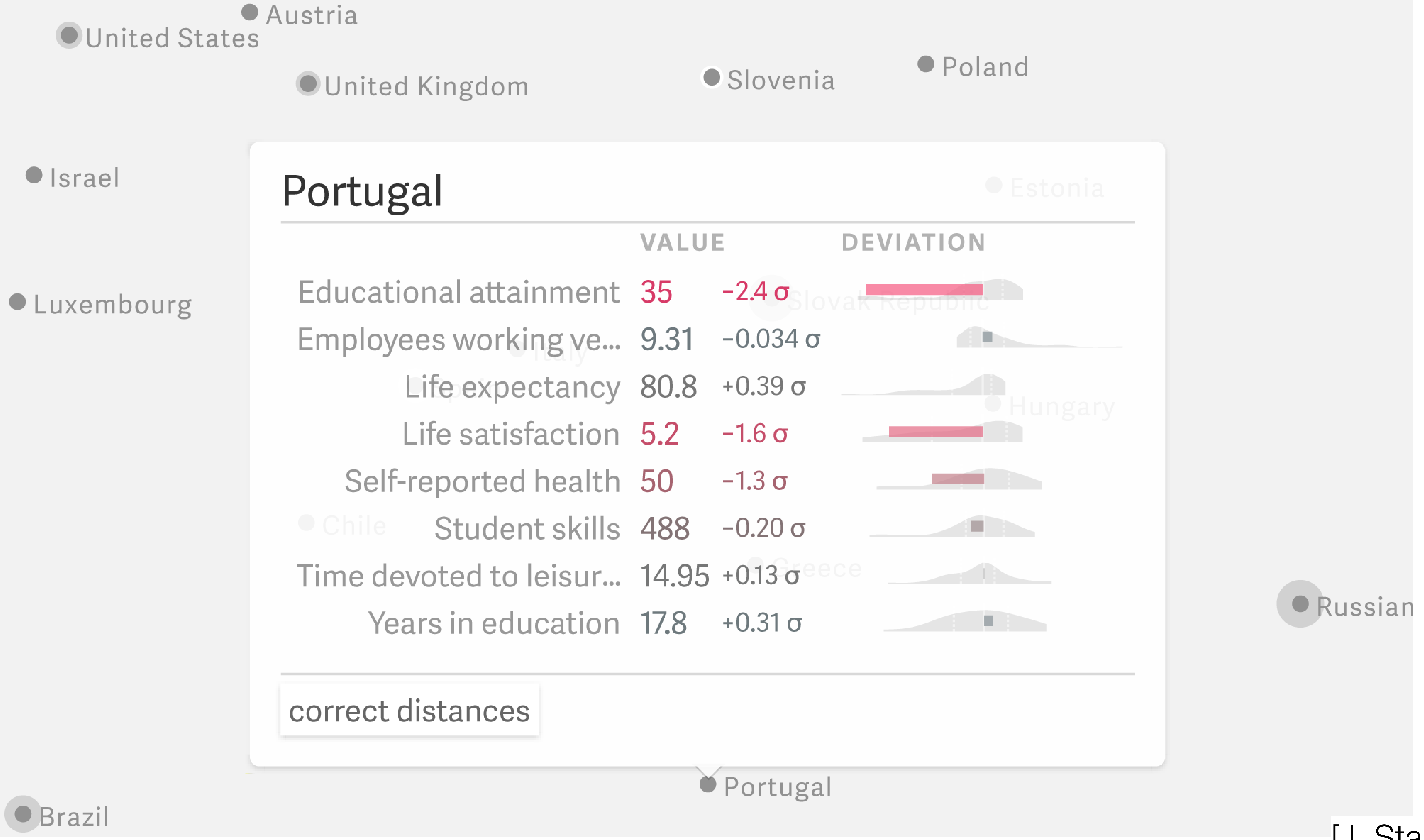
[J. Stahnke et al., 2015]

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

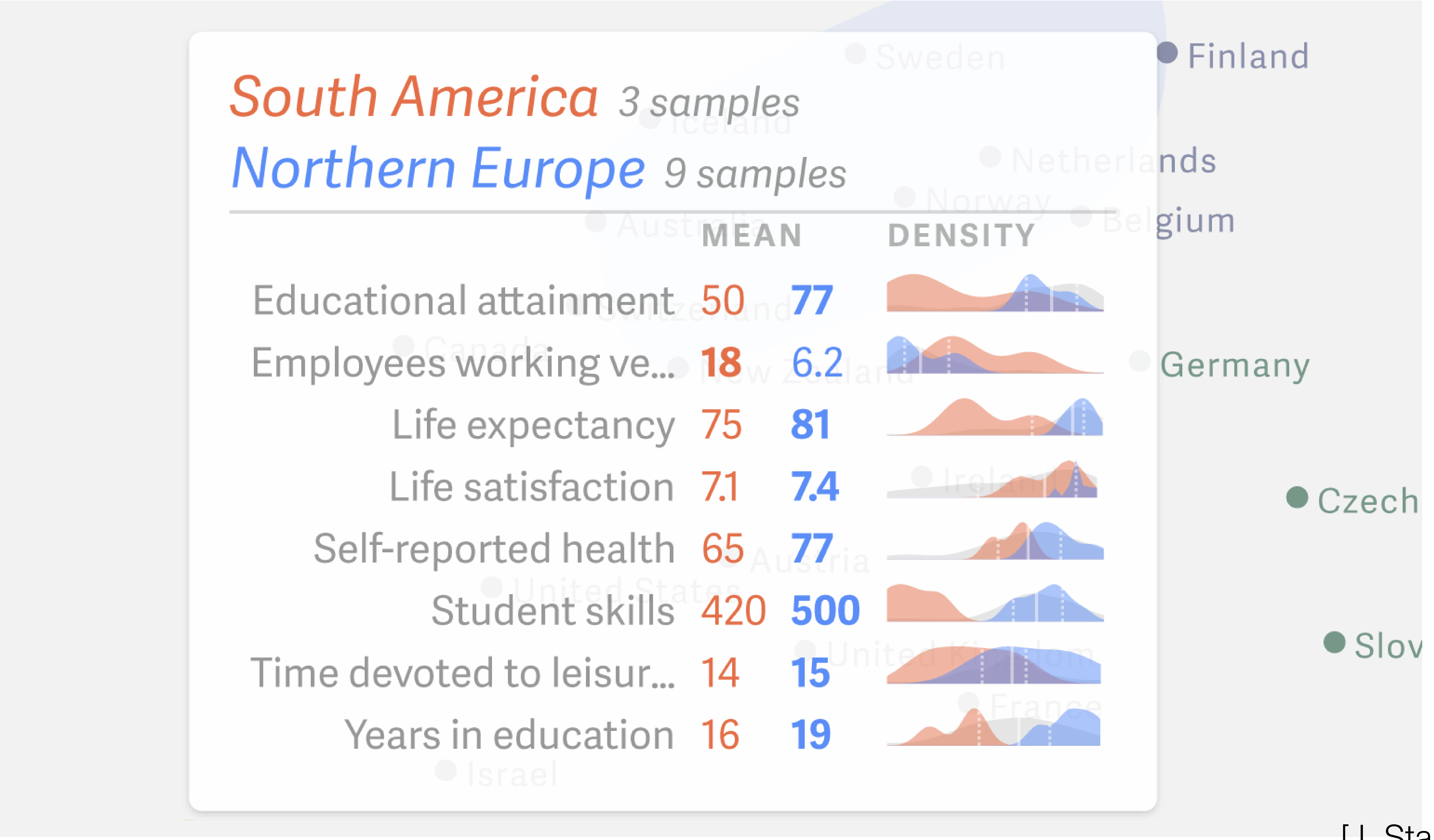


Portugal

	VALUE	DEVIATION
Educational attainment	35	−2.4 σ
Employees working ve…	9.31	−0.034 σ
Life expectancy	80.8	+0.39 σ
Life satisfaction	5.2	−1.6 σ
Self-reported health	50	−1.3 σ
Student skills	488	−0.20 σ
Time devoted to leisur…	14.95	+0.13 σ
Years in education	17.8	+0.31 σ

[J. Stahnke et al., 2015]

Comparing Two Groups



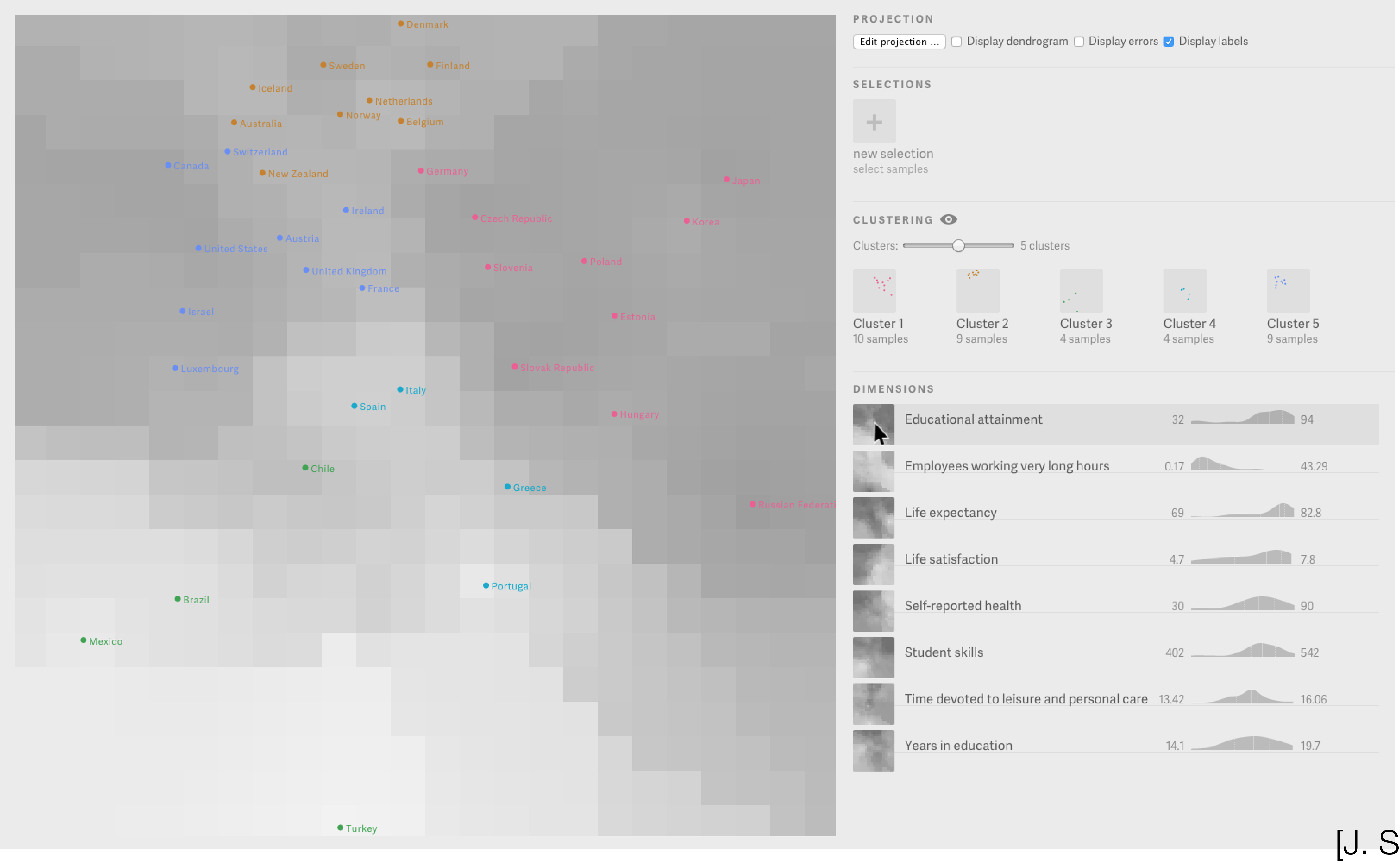
South America *3 samples*

Northern Europe *9 samples*

	MEAN		DENSITY
Educational attainment	50	77	
Employees working ve...	18	6.2	
Life expectancy	75	81	
Life satisfaction	7.1	7.4	
Self-reported health	65	77	
Student skills	420	500	
Time devoted to leisur...	14	15	
Years in education	16	19	

[J. Stahnke et al., 2015]

Heatmap from Dimension Hover



PROJECTION

Edit projection ... □ Display dendrogram □ Display errors ☑ Display labels

SELECTIONS

+ new selection
select samples

CLUSTERING

Clusters: 5 clusters

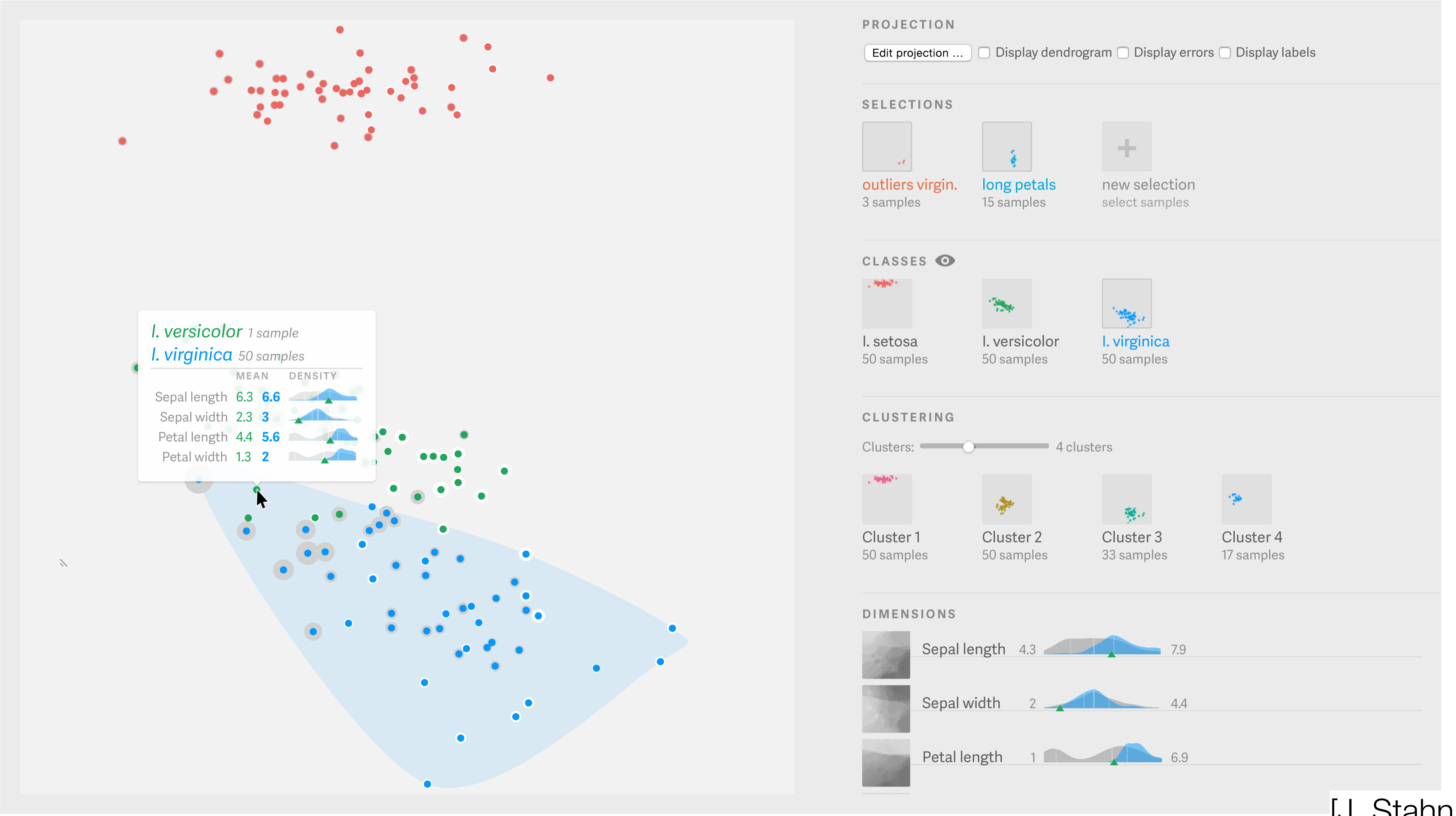
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
10 samples	9 samples	4 samples	4 samples	9 samples

DIMENSIONS

Dimension	Min	Max
Educational attainment	32	94
Employees working very long hours	0.17	43.29
Life expectancy	69	82.8
Life satisfaction	4.7	7.8
Self-reported health	30	90
Student skills	402	542
Time devoted to leisure and personal care	13.42	16.06
Years in education	14.1	19.7

[J. Stahnke et al., 2015]

Showing Error via Sample-centric Halos



I. versicolor 1 sample
I. virginica 50 samples

	MEAN		DENSITY
Sepal length	6.3	6.6	
Sepal width	2.3	3	
Petal length	4.4	5.6	
Petal width	1.3	2	

PROJECTION

Edit projection ... □ Display dendrogram □ Display errors □ Display labels

SELECTIONS

outliers virgin. 3 samples
long petals 15 samples
new selection select samples

CLASSES

I. setosa 50 samples
I. versicolor 50 samples
I. virginica 50 samples

CLUSTERING

Clusters: 4 clusters

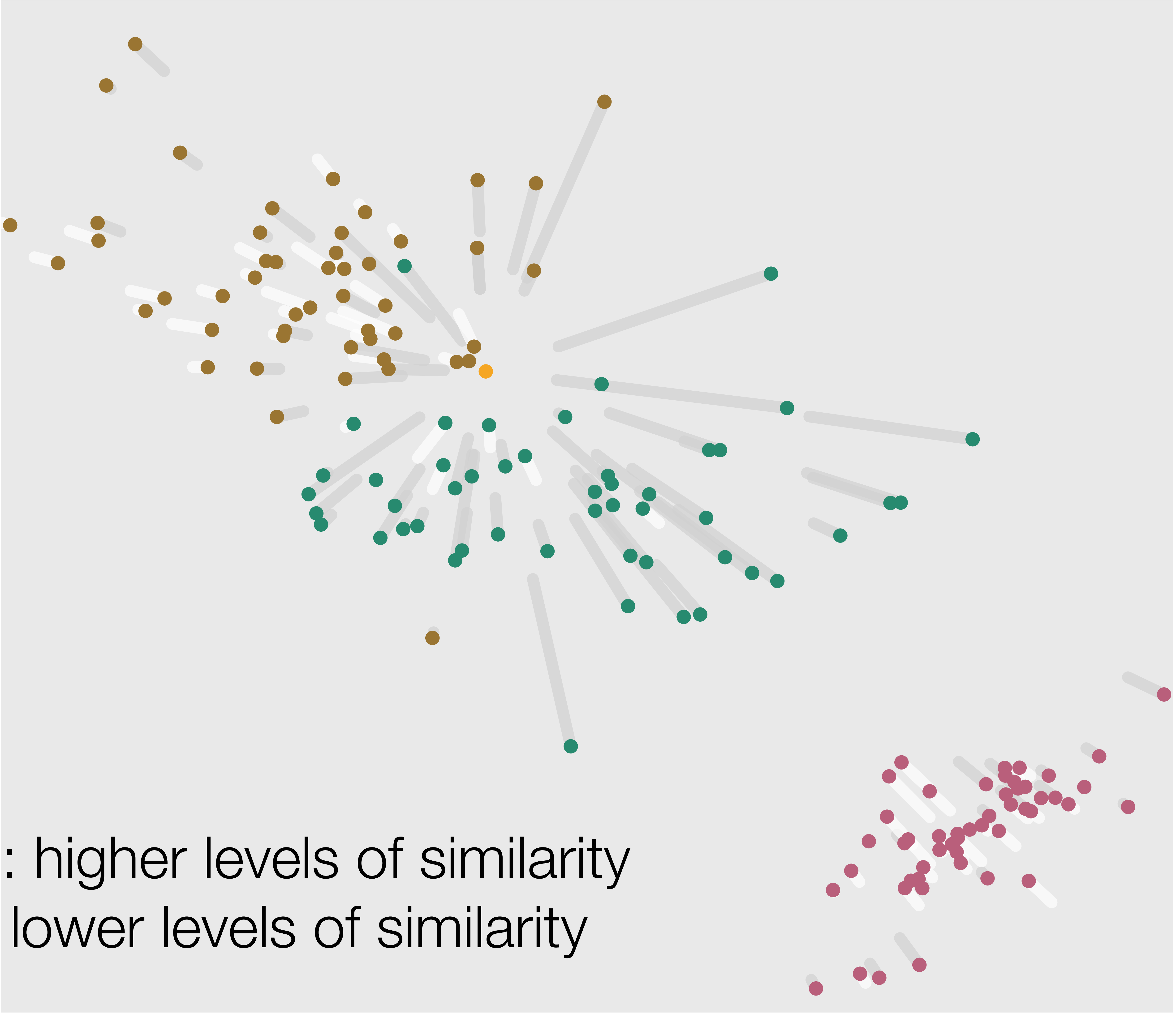
Cluster 1 50 samples
Cluster 2 50 samples
Cluster 3 33 samples
Cluster 4 17 samples

DIMENSIONS

Sepal length	4.3	7.9
Sepal width	2	4.4
Petal length	1	6.9

[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]