Previously on 95-865…

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{\boxed{P(A, B)}}{\boxed{P(A)\ P(B)}}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI measures (the log of) a ratio that says how far A and B are from being independent

There are *lots* of connections of information theory to prediction

Rough intuition:
Something surprising ↔ less predictable ↔ more bits to store

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\, P(A, B) - P(A)\, P(B)\,]^2}{P(A)\, P(B)}$$

Phi-square is between 0 and 1

0 ➔ pairs are all indep.

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\, P(A, B) - P(A)\, P(B)\,]^2}{P(A)\, P(B)}$$

Chi-square = N × Phi-square

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$N = 5750$

Sum comprises of 6 terms

Green, Green: $\dfrac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)} = 0.2216\ldots$

Green, White: $\dfrac{\left[\frac{200}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)} = 0.0496\ldots$

Green, Black: $= 0.0496\ldots$

$\vdots$

White, White: $= 0.1161\ldots$

White, Black: $= 0.0937\ldots$

Black, Black: $= 0.1161\ldots$

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Add these up to get:
Phi-square = 0.6470…

Interpretation: neighboring pixels not close to being indep.

Back to Earlier Example

	[redacted]	Tesla	Apple	[redacted]
Elon Musk	[redacted]	300	1	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Tim Cook	[redacted]	1	195	[redacted]

Often we know what kind of named entities are found

Example: Elon Musk and Tim Cook are people,
Tesla and Apple are companies

➔ can ask what people are related to what companies

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

PMI, phi-square, chi-square calculations done same way as before

Main things to calculate first:

P(Elon Musk, Tesla)	P(Elon Musk)
P(Elon Musk, Apple)	P(Tim Cook)
P(Tim Cook, Tesla)	P(Tesla)
P(Tim Cook, Apple)	P(Apple)

The math here is actually a bit easier to think about
because the rows and columns aren't indexing the same items

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

Divide by total

These are the joint probabilities!

Compute "marginals"

	Tesla	Apple	
Elon Musk	300/497	1/497	300/497+1/497
Tim Cook	1/497	195/497	1/497+195/497
	300/497+1/497	1/497+195/497	

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

P(Elon Musk)

P(Tim Cook)

P(Tesla)

P(Apple)

Not just for 2 by 2 tables (e.g., we could have many people, many companies)

Summary: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is "interesting"
- Find interesting relationships between pairs of items by looking at PMI
  - Intuition: "Interesting" co-occurring events should occur more frequently than if they were to co-occur independently
- In practice: some times it is helpful to generalize PMI and look instead at

$$\mathrm{PMI}_\rho(A, B) = \log_2 \frac{P(A, B)^\rho}{P(A)\, P(B)}$$

Tune parameter
$\rho > 0$

(we'll talk about parameter tuning later in the course)

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/rented/consumed more
  - Products & dates
- If you have a bunch of physical stores:
  anticipate *where* certain products are likely to be purchased/rented/consumed more
  - Products & locations
- If you're the police department:
  create "heat map" of where different criminal activity occurs
  - Crime reports & locations

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/
  re

  -

- If y
  an sed/
  re

  -

- If y
  cr curs

  - Crime reports & locations

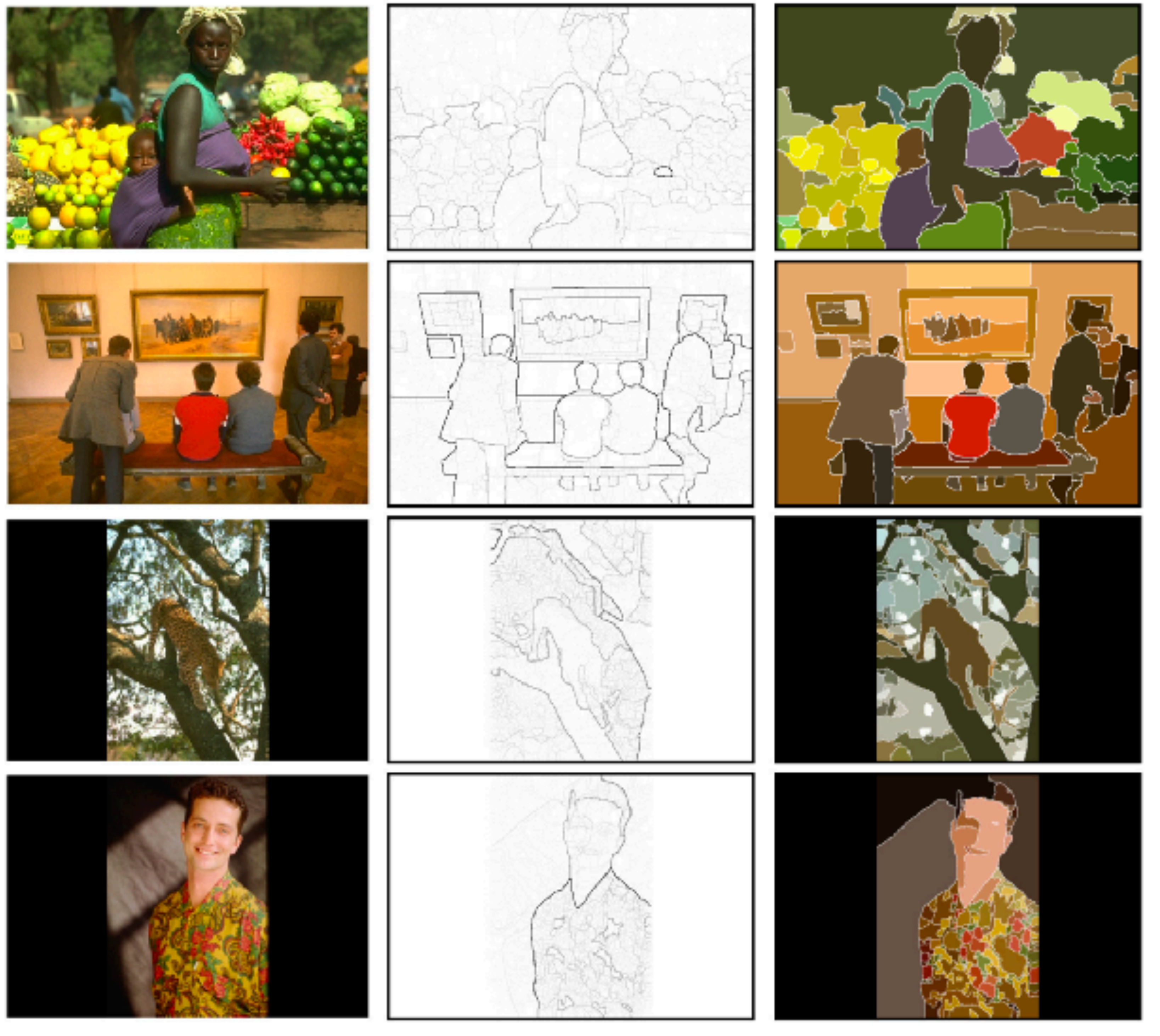
Examples of data to take advantage of:

- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:

- Scrapy
- Selenium (great with JavaScript-heavy pages)

Example Application of PMI: Image Segmentation



Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Crisp boundary detection using pointwise mutual information. ECCV 2014.

Example Application of PMI: Word Embeddings

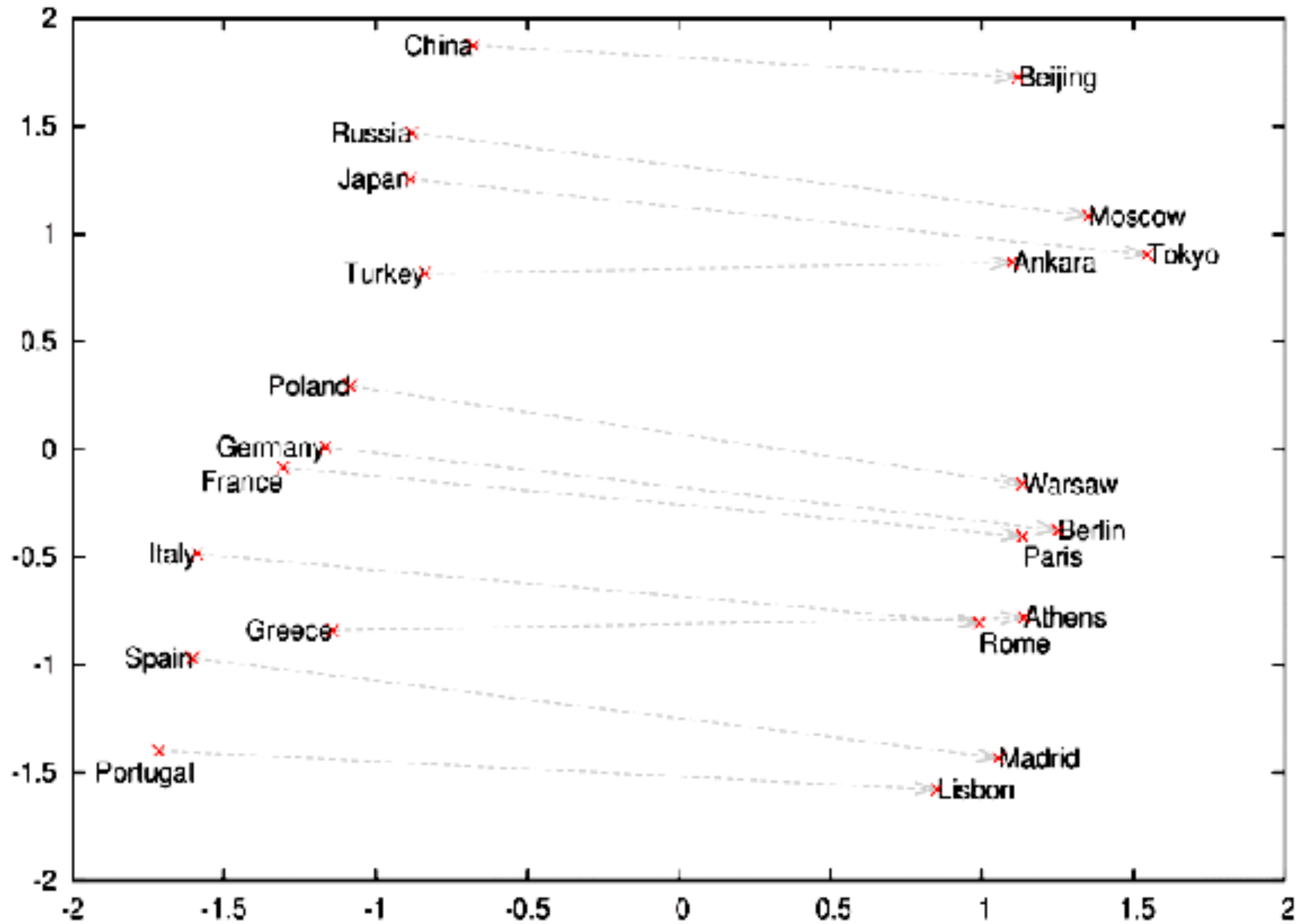
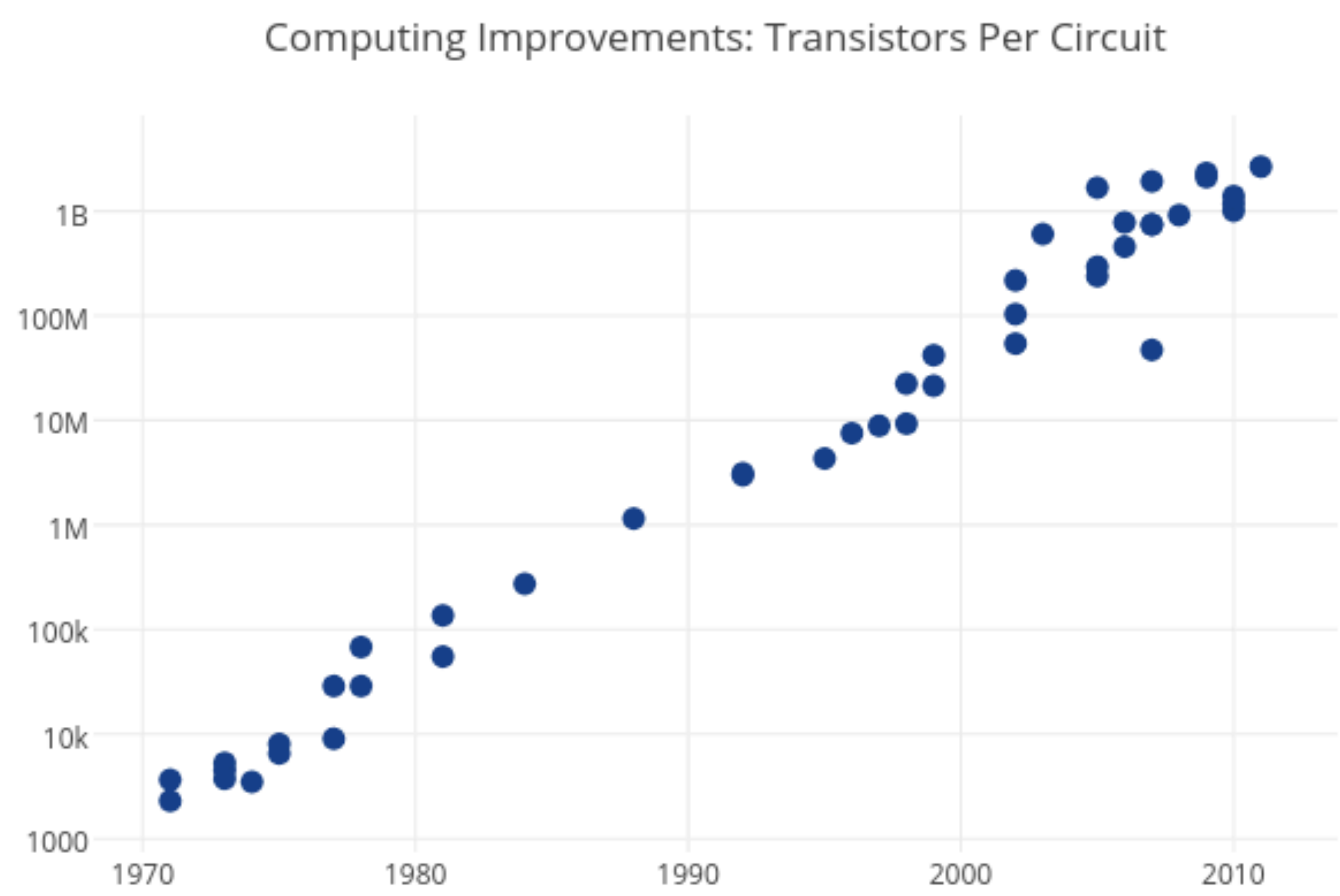


Image source: https://deeplearning4j.org/img/countries_capitals.png

Omer Levy and Yoav Goldberg. Neural word embeddings as implicit matrix factorization. NIPS 2014.

Continuous Measurements

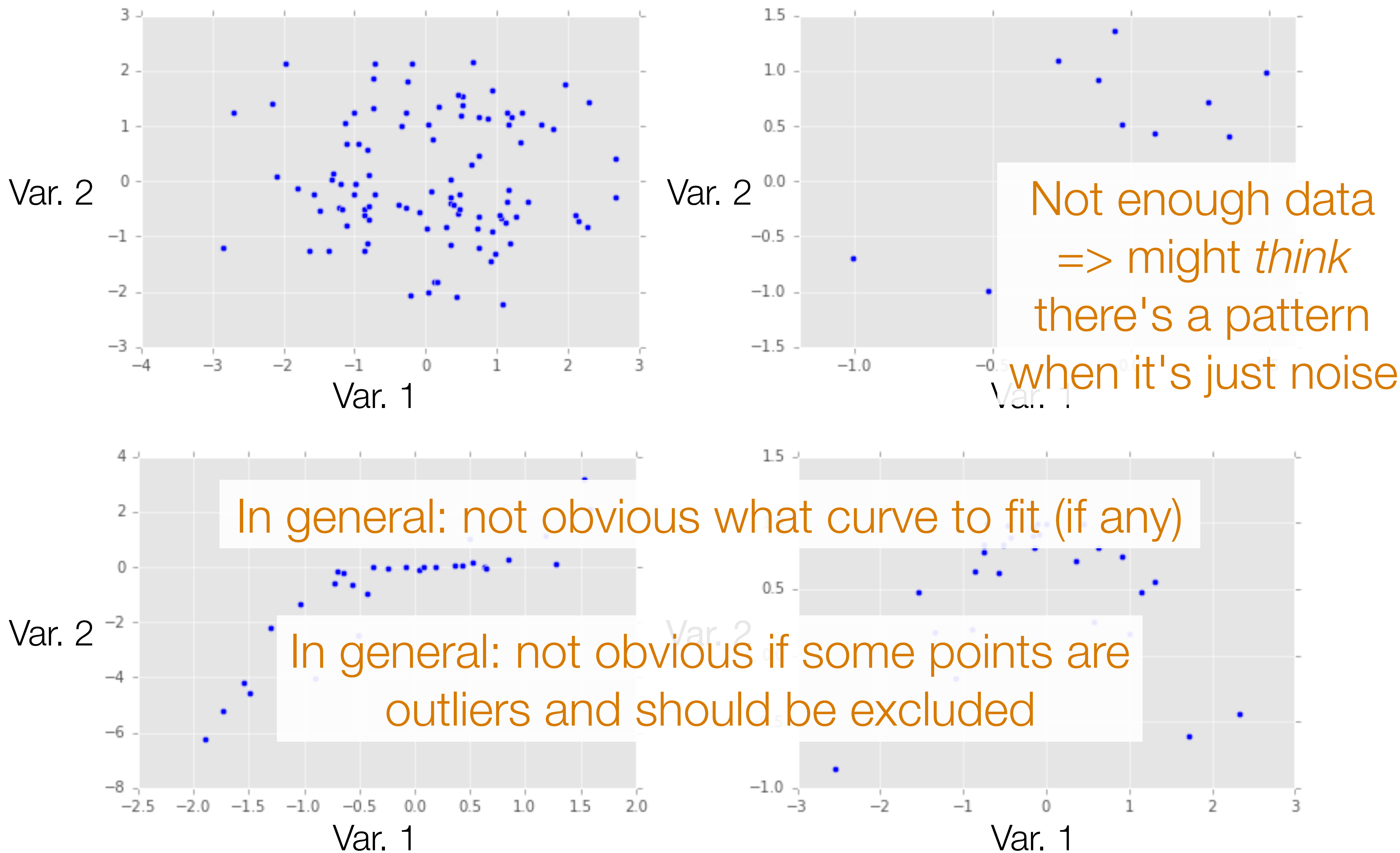
- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**



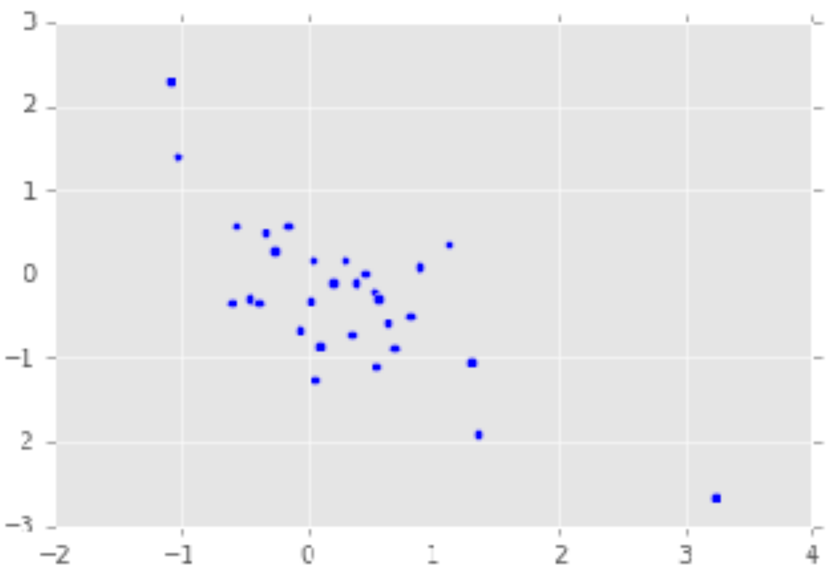
Of course, not all trends look like a line

Image source: https://plot.ly/~MattSundquist/5405.png

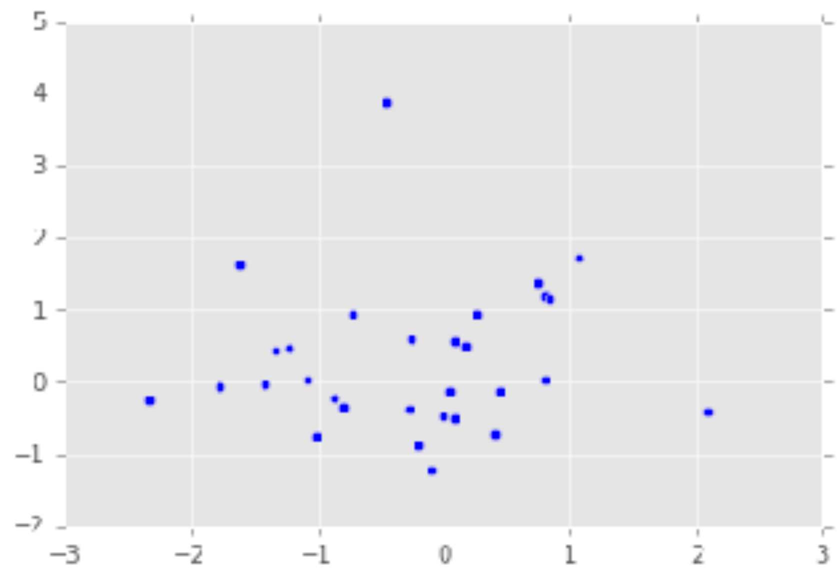
The Importance of Staring at Data



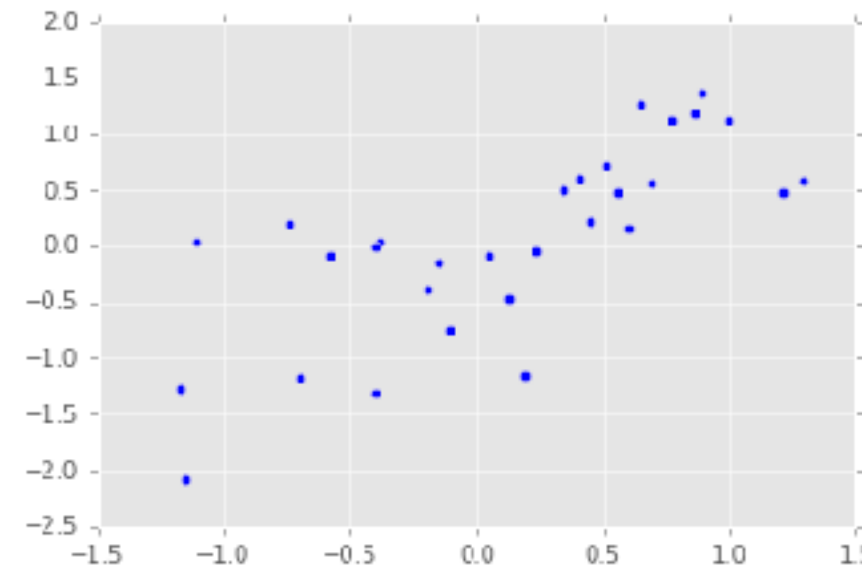
Correlation



Negatively correlated



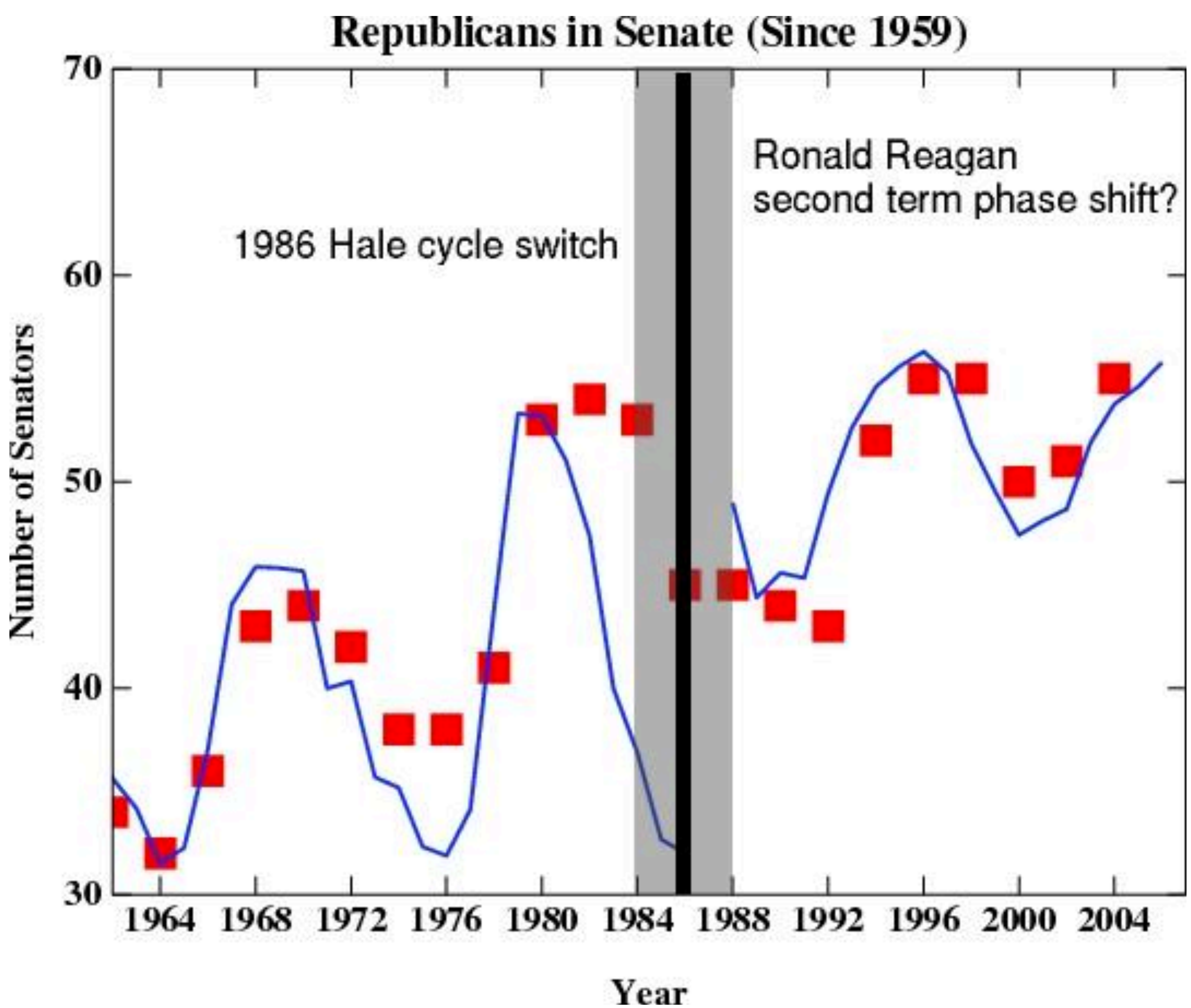
Not really correlated



Positively correlated

Beware: Just because two variables appear correlated doesn't mean that one can predict the other

Correlation ≠ Causation



Blue: Scaled sunspot number (inverted after Reagan's 2nd term)

Red: Number of Republican senators

Moreover, just because we find correlation in data doesn't mean it has predictive value!

Image source: http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/

Important: At this point in the course, we are finding *possible* relationships between two entities

We are *not* yet making statements about prediction (we'll see prediction later in the course)

We are *not* making statements about causality (beyond the scope of this course)

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Can we claim that coffee is a cause of lung cancer?

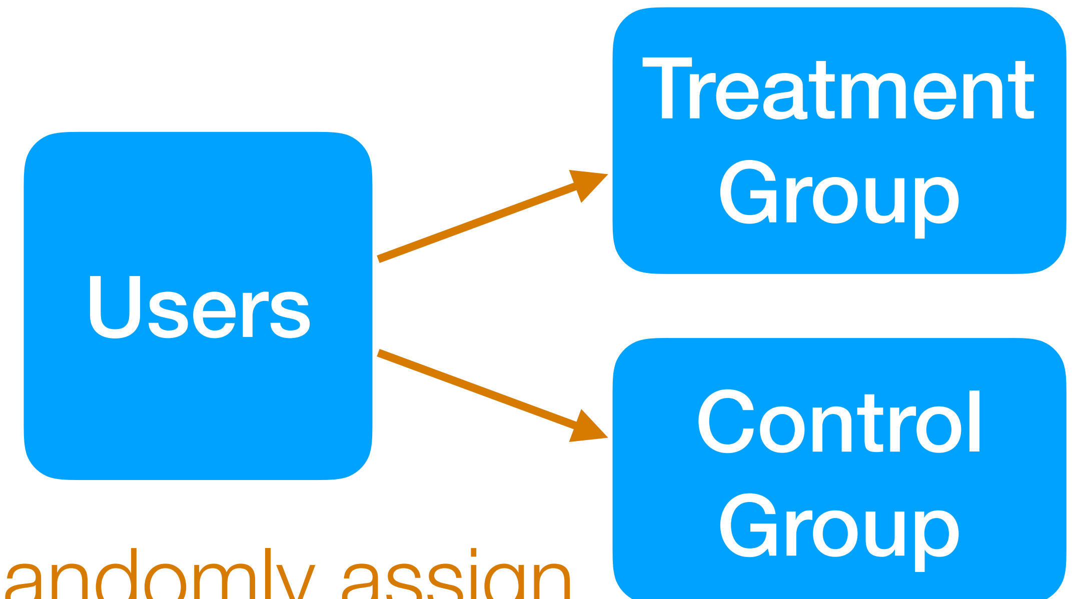
Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

To establish causality, groups getting different treatments need to appear similar so that the only difference is the treatment

Image source: George Chen

Establishing Causality

If you control data collection



Randomly assign

Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

Example: figure out webpage layout to maximize revenue (Amazon)

Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

In general: *not* obvious establishing what caused what

95-865 Outline

Part I: Exploratory data analysis

Identify structure present in "unstructured" data

- Frequency and co-occurrence analysis *Basic probability & statistics*
- Clustering
- Topic modeling (a special kind of clustering)

Part II: Predictive data analysis

Make predictions using structure found in Part I

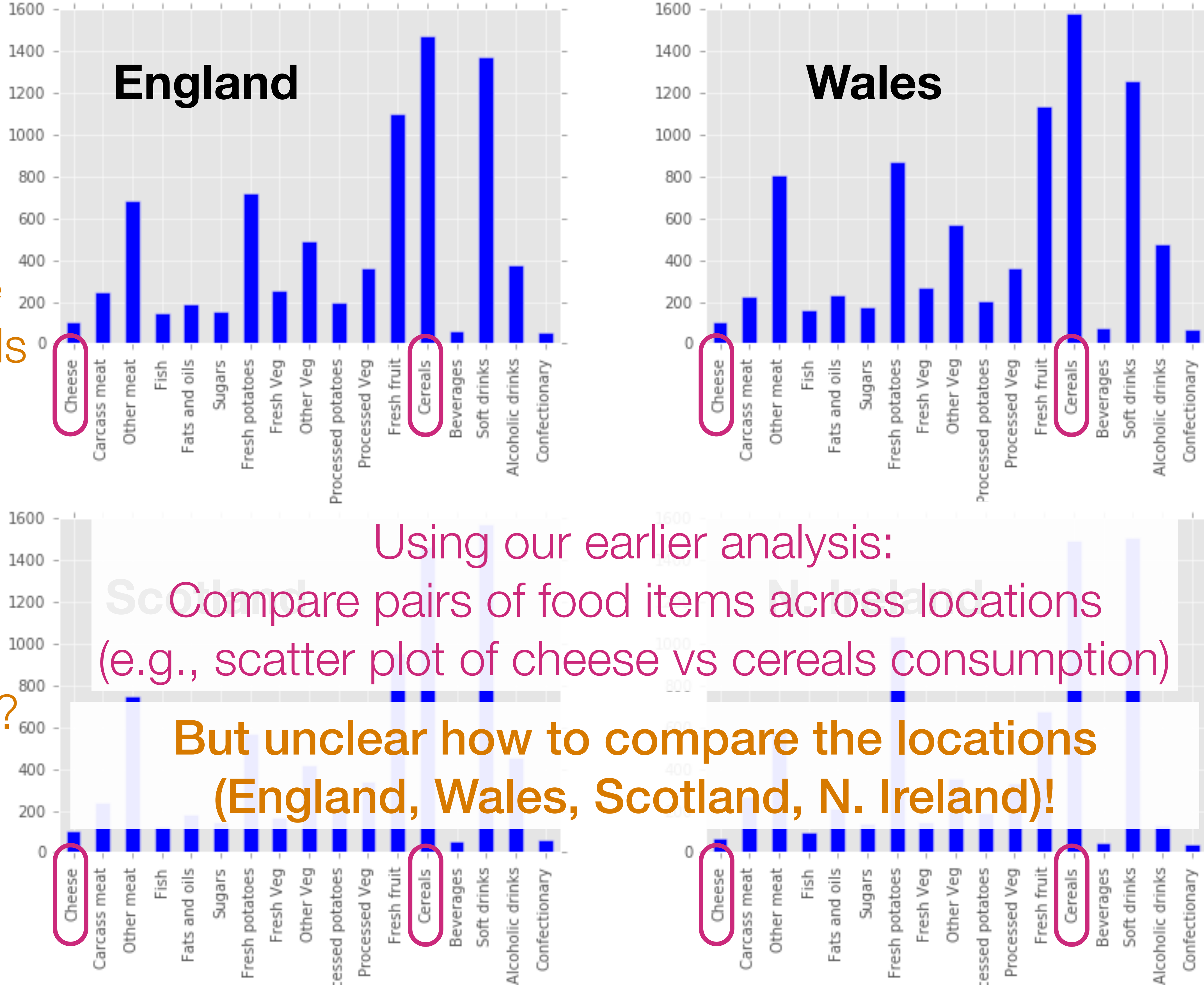
- Classical classification methods
- Neural nets and deep learning for analyzing images and text

Visualizing High-Dimensional Vectors

George Chen

The next two examples are drawn from:
http://setosa.io/ev/principal-component-analysis/

Visualizing High-Dimensional Vectors



Imagine we had hundreds of these

How to visualize these for comparison?

Using our earlier analysis:
Compare pairs of food items across locations
(e.g., scatter plot of cheese vs cereals consumption)

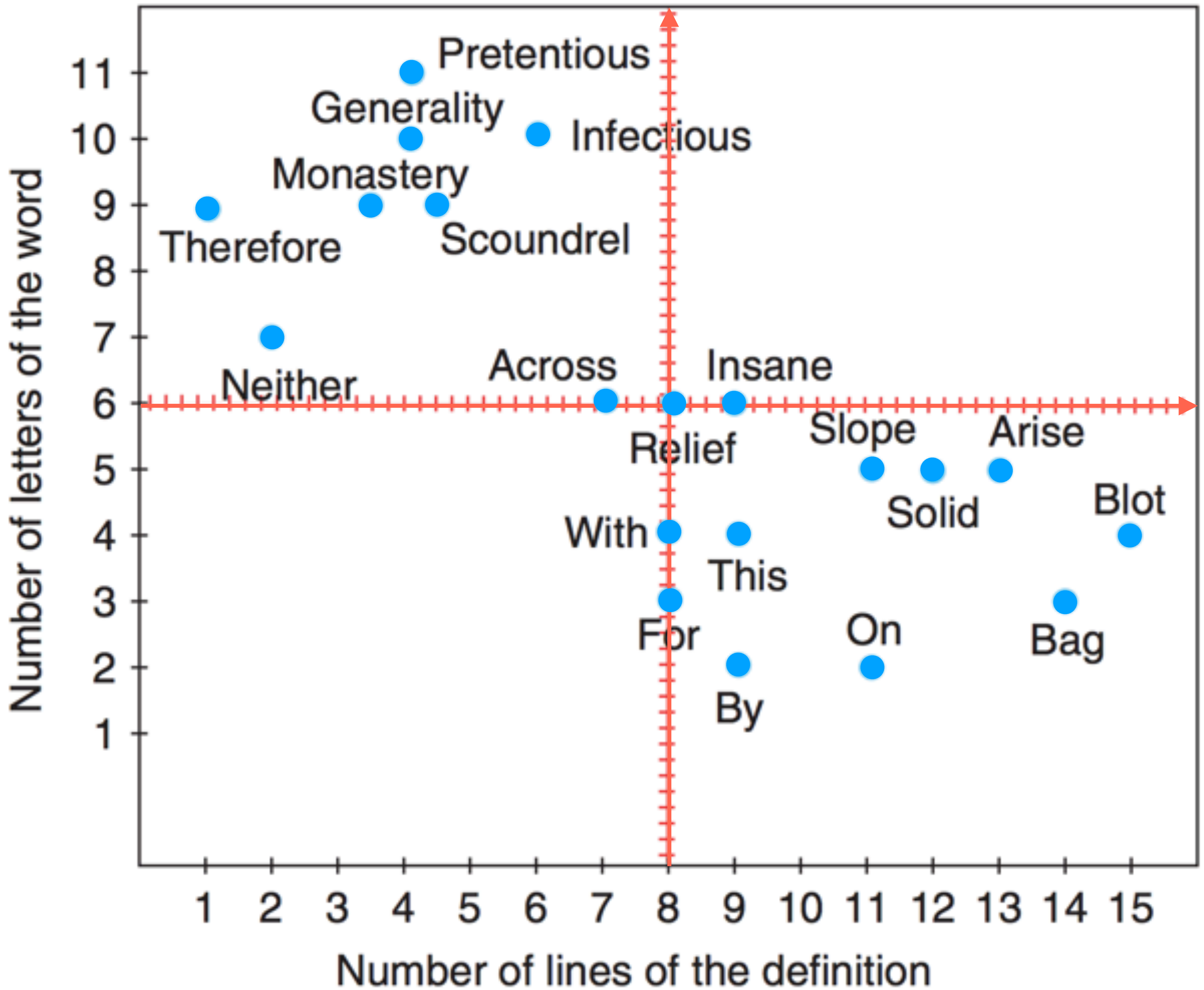
But unclear how to compare the locations (England, Wales, Scotland, N. Ireland)!

The issue is that as humans we can only really visualize up to 3 dimensions easily

Goal: Somehow reduce the dimensionality of the data preferably to 1, 2, or 3

Principal Component Analysis (PCA)

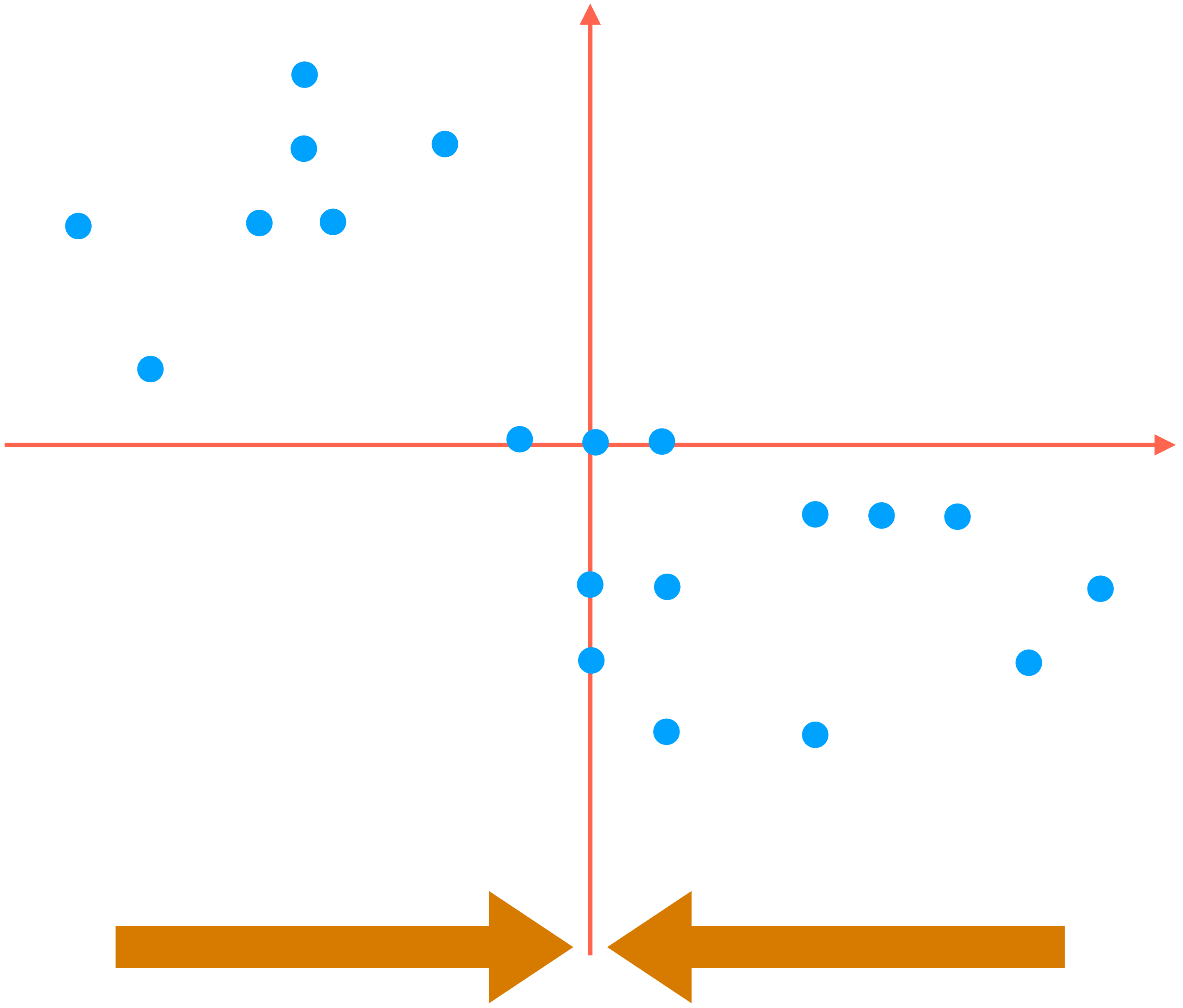
How to project 2D data down to 1D?



Hervé Abdi and Lynne J. Williams. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

Principal Component Analysis (PCA)

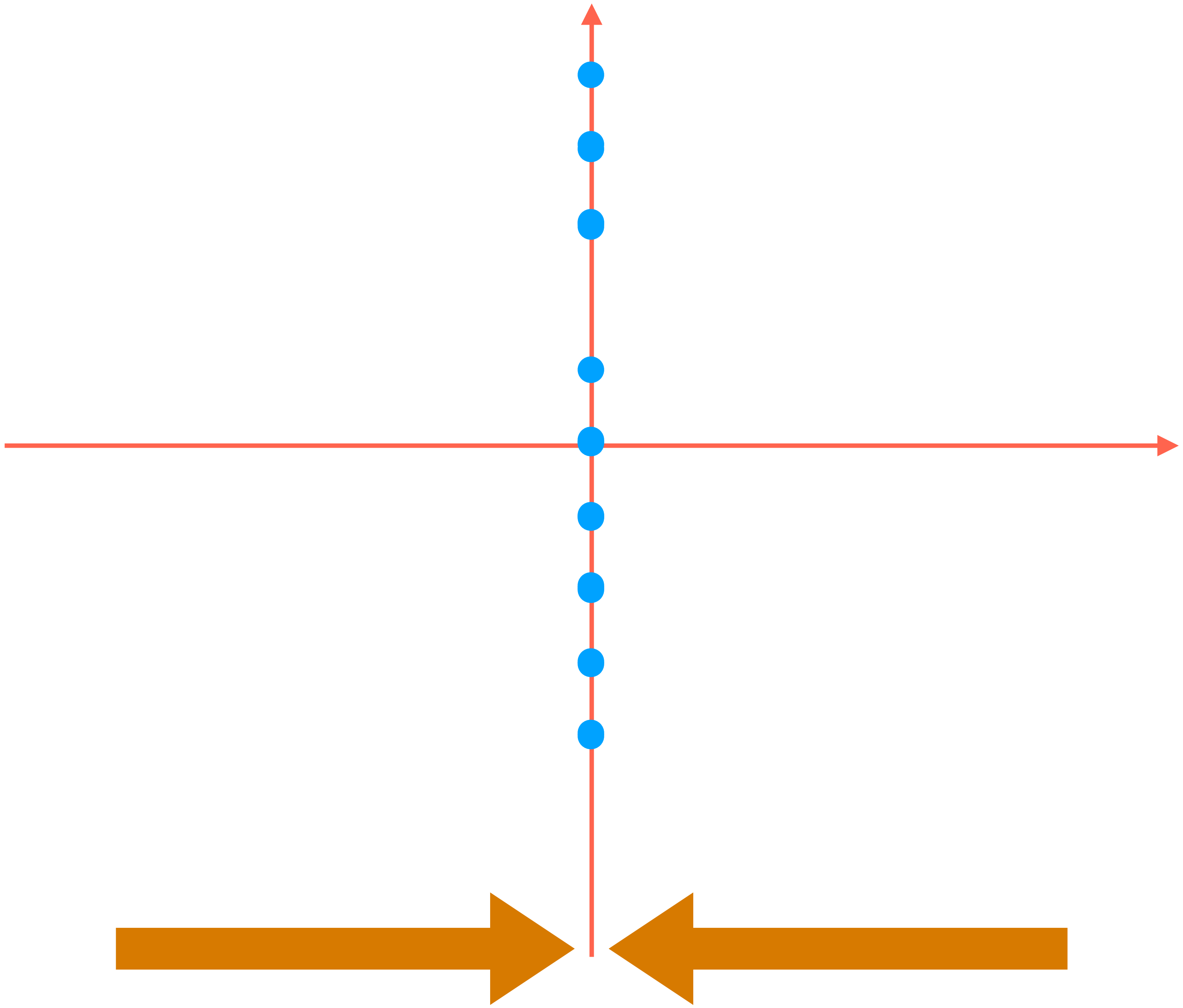
How to project 2D data down to 1D?



Simplest thing to try: flatten to one of the red axes

Principal Component Analysis (PCA)

How to project 2D data down to 1D?

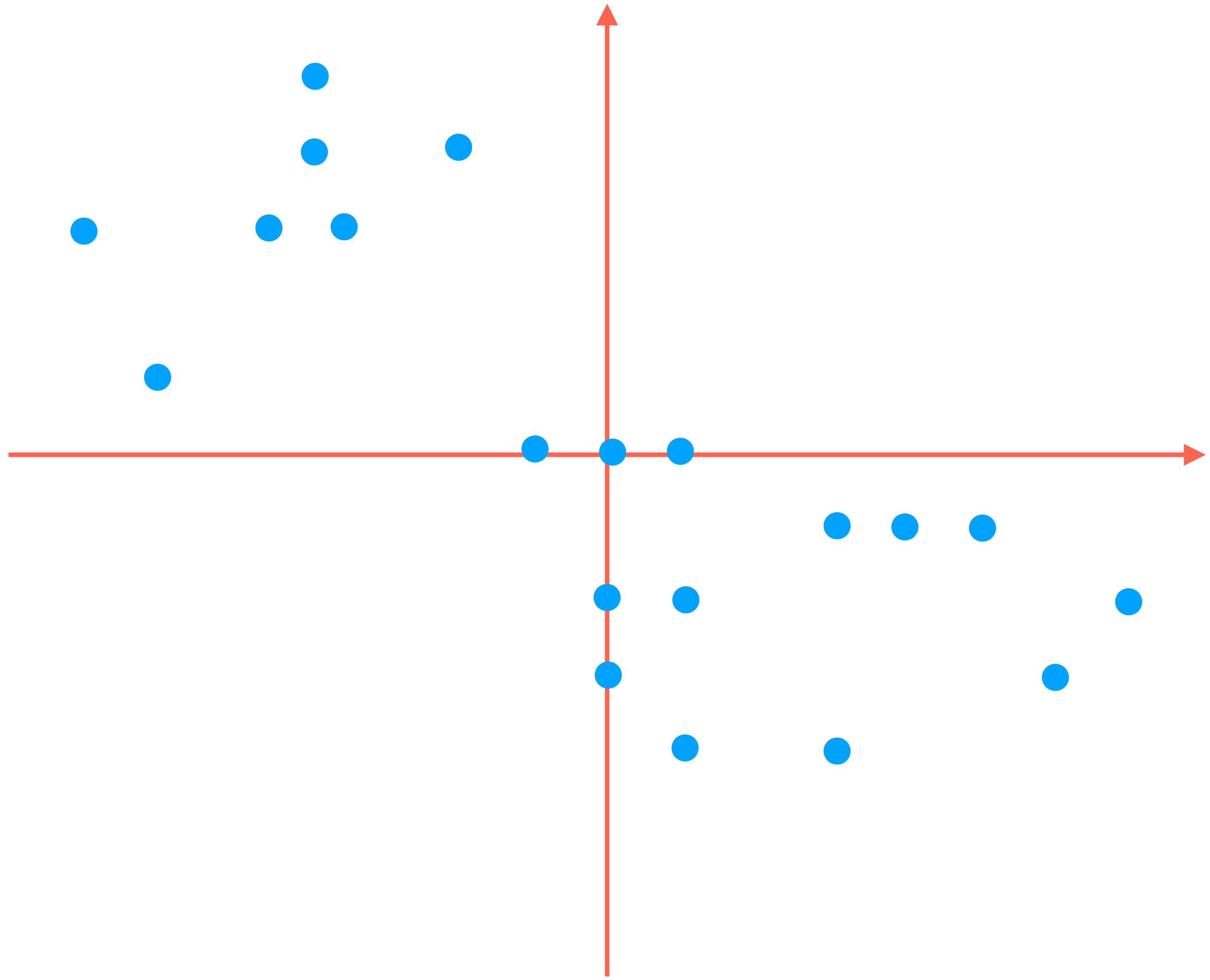


Simplest thing to try: flatten to one of the red axes

(We could of course flatten to the other red axis)

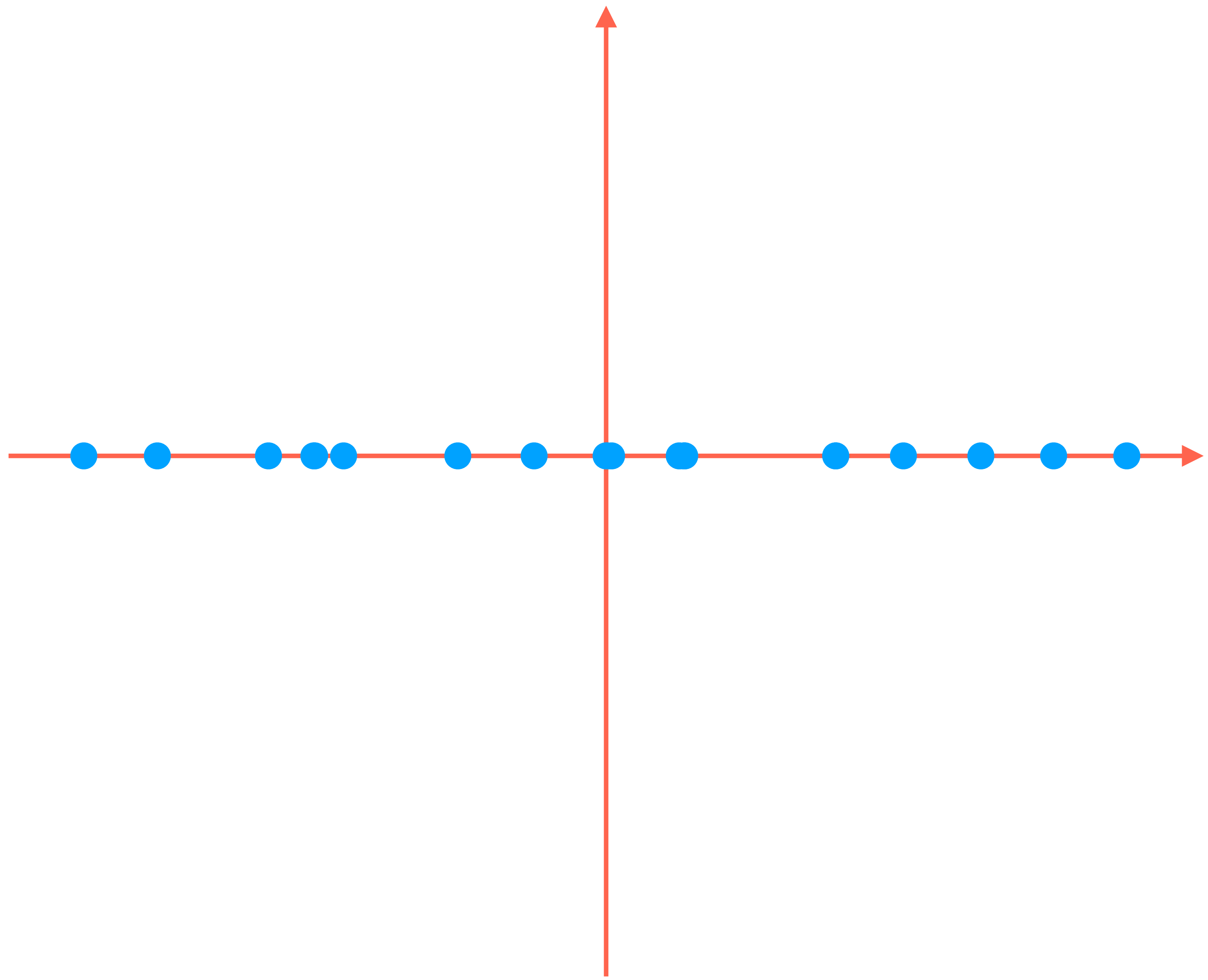
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



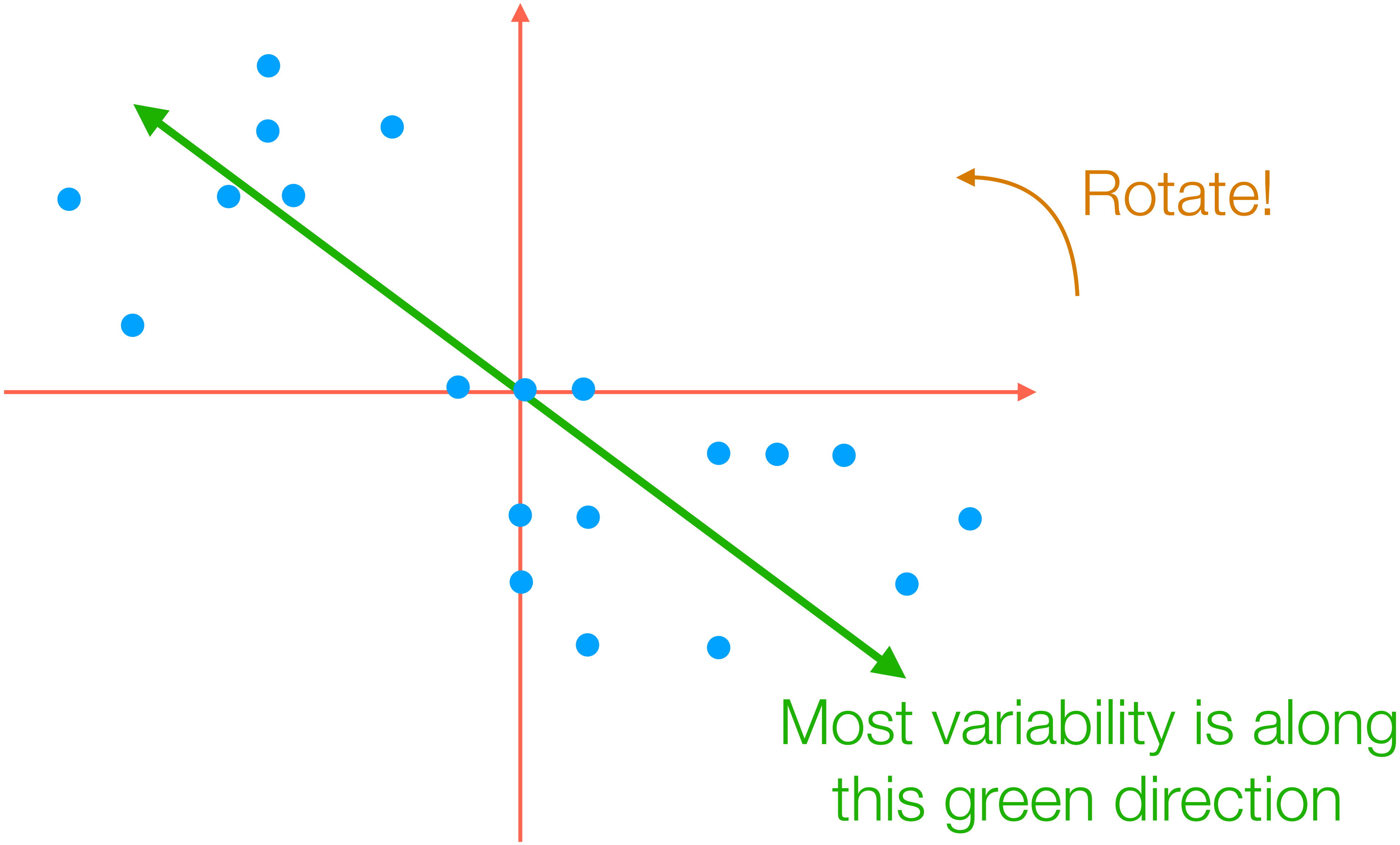
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

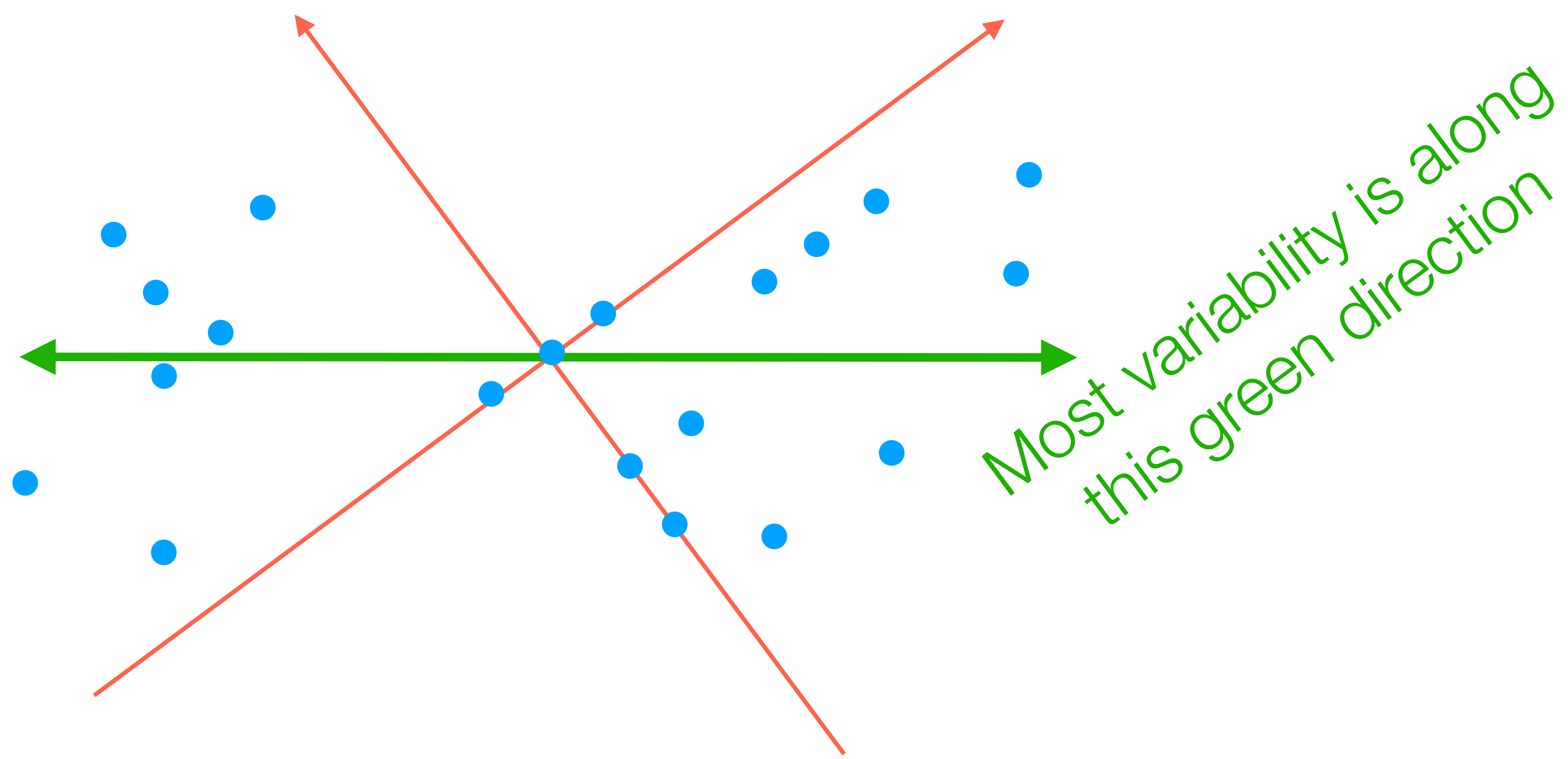
How to project 2D data down to 1D?



But notice that most of the variability in the data is *not* aligned with the red axes!

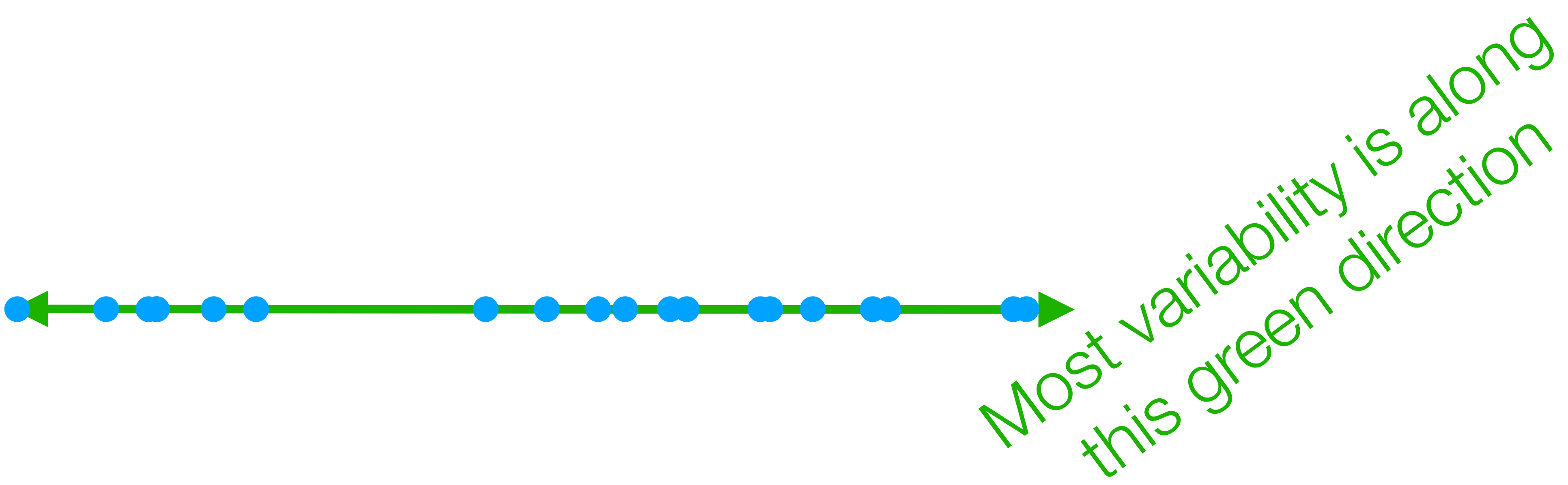
Principal Component Analysis (PCA)

How to project 2D data down to 1D?



Principal Component Analysis (PCA)

How to project 2D data down to 1D?

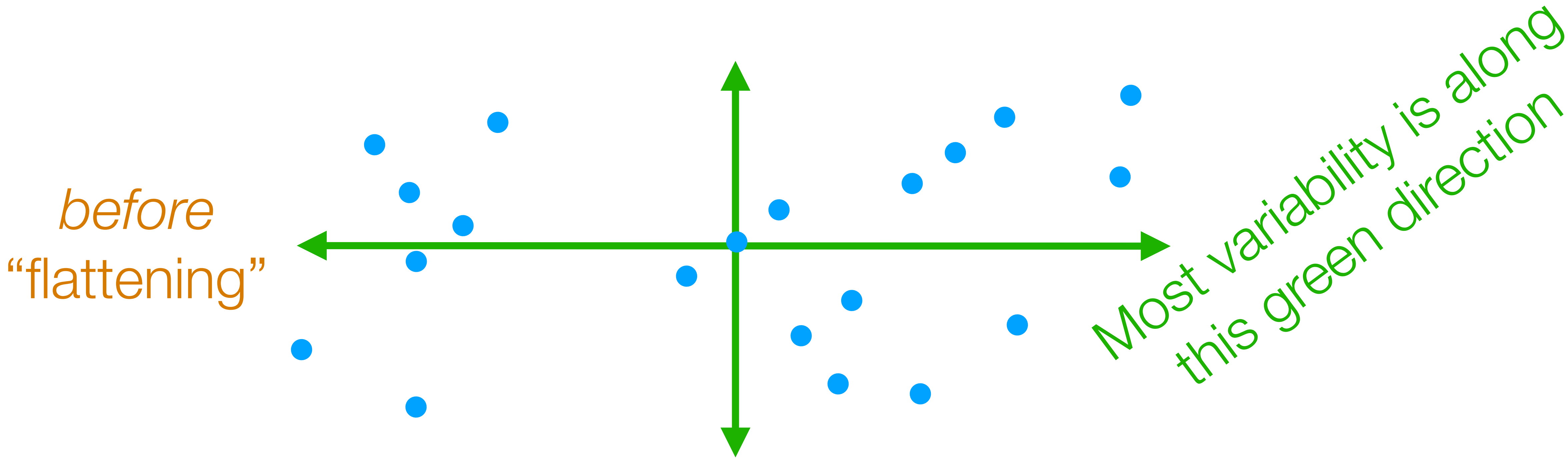


The idea of PCA actually works for 2D ➔ 2D as well
(and just involves rotating, and not "flattening" the data)

Principal Component Analysis (PCA)

~~How to project 2D data down to 1D?~~

How to rotate 2D data so 1st axis has most variance



The idea of PCA actually works for 2D ➔ 2D as well
(and just involves rotating, and not “flattening” the data)

2nd green axis chosen to be 90° (“orthogonal”) from first green axis

Principal Component Analysis (PCA)

- Finds top *k* orthogonal directions that explain the most variance in the data
  - 1st component: explains most variance along 1 dimension
  - 2nd component: explains most of remaining variance along next dimension that is orthogonal to 1st dimension
  - …

- “Flatten” data to the top *k* dimensions to get lower dimensional representation (if k < original dimension)

Principal Component Analysis (PCA)

3D example from:
http://setosa.io/ev/principal-component-analysis/

Principal Component Analysis (PCA)

Demo