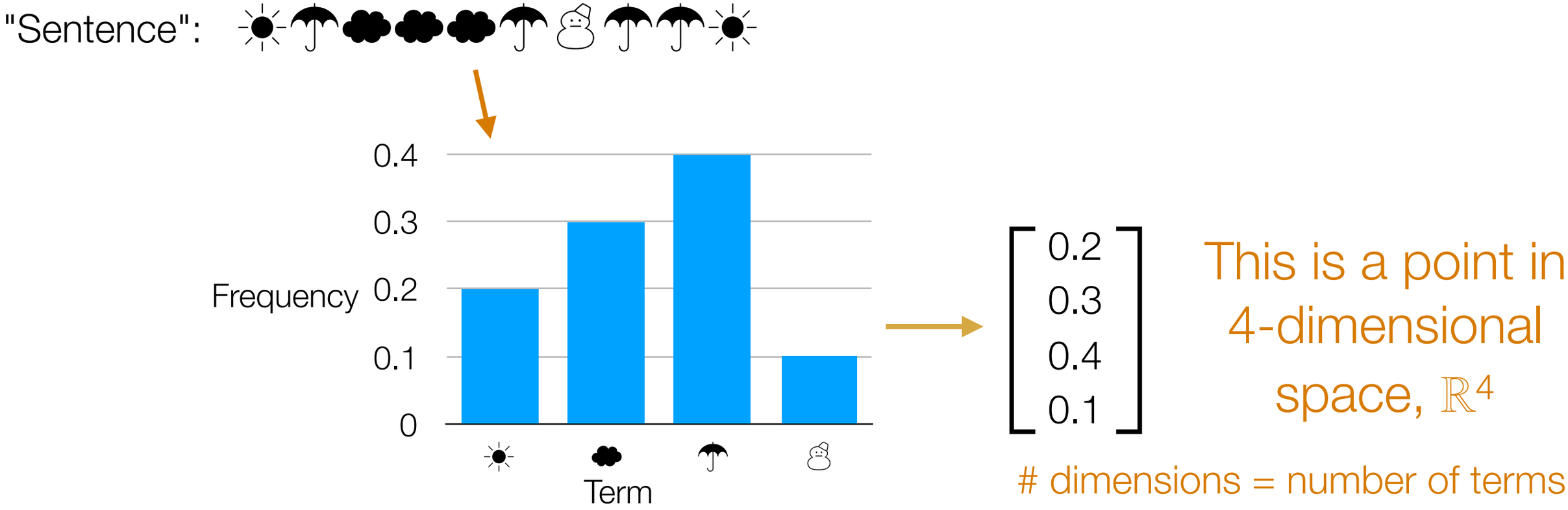
Recap: Basic Text Analysis

- Represent text in terms of “features”
  (e.g., how often each word/phrase appears, whether it’s a named entity, etc)
  - Can repeat this for different documents:
    represent each document as a “feature vector”



In general (not just text): first represent data as feature vectors

Example: Representing an Image

0: black
1: white

0

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

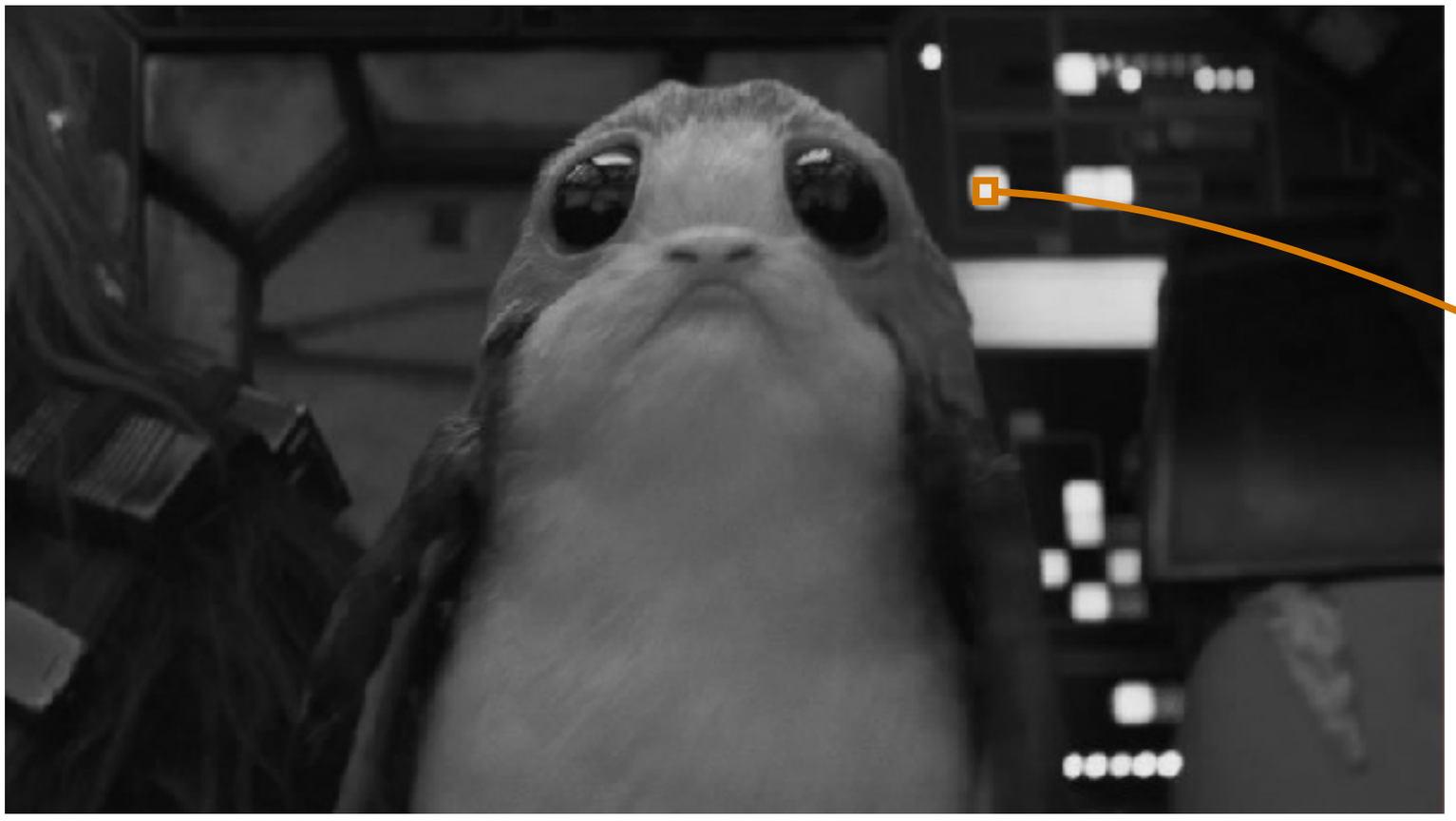
0: black
1: white

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image



0: black
1: white

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

0: black
1: white

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \\ 0.3 \end{bmatrix}$$

Go row by row and look at pixel values

dimensions = image width × image height

Very high dimensional!

Image source: starwars.com

Back to Text

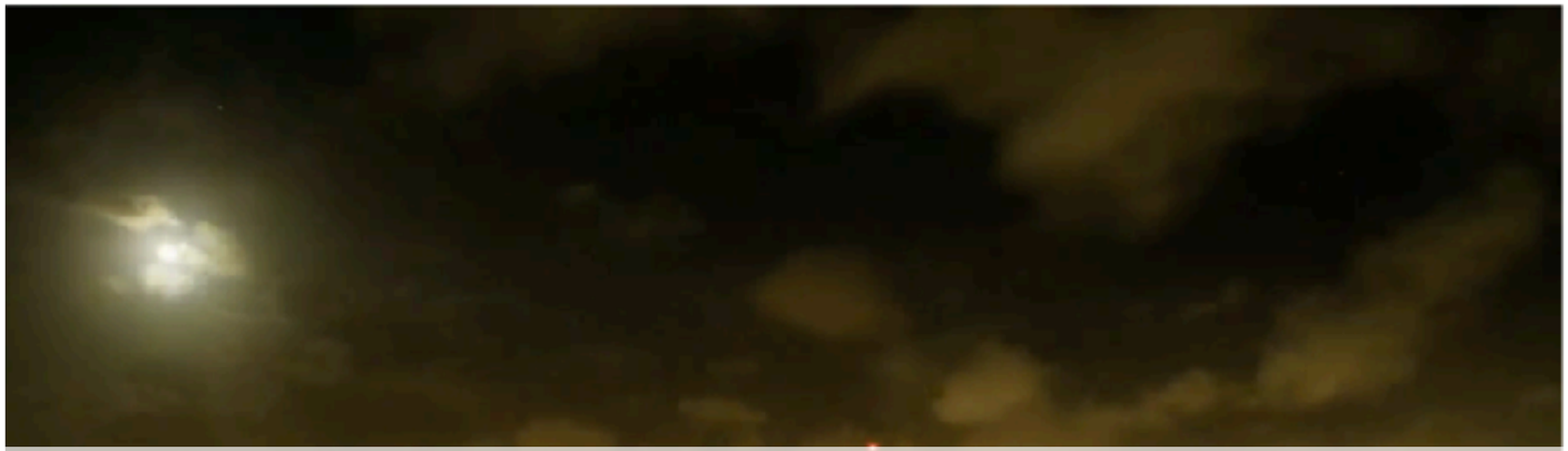
Unigram bag of words model is already quite powerful:

- Enough to learn topics
  (each text doc: raw word counts without stopwords)
- Enough to learn a simple detector for email spam

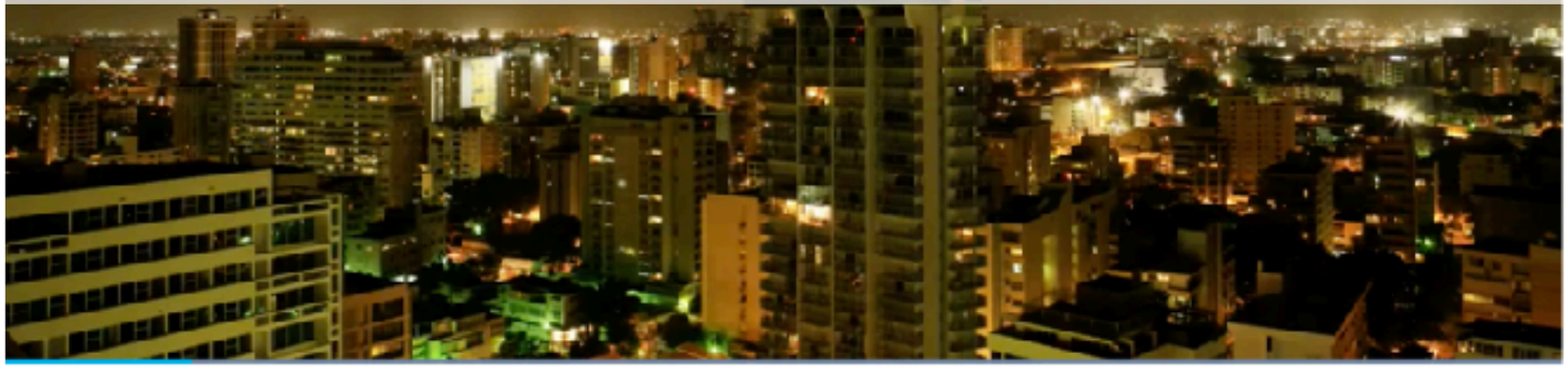
These are HW2 problems

Finding Possibly Related Entities

Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico



How to automatically figure out Elon Musk and Tesla are related?



The solar batteries have reportedly been spotted in San Juan's airport.

By **John Patrick Pullen** October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	5	1
Tesla	300			195
Apple	1	5		
Tim Cook	4	1	195	

Large values => possible related items

What does it mean for a named entity to co-occur with itself?

Example: could count # articles in which word appears ≥ 2 times

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Bottom Line

- There are many ways to count co-occurrences
- You should think about what makes the most sense/is reasonable for the problem you're looking at

We aim to find *interesting* relationships by looking at co-occurrences

Black and white frequently co-occur, but is this relationship interesting?

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

How I'm counting: For each pixel, look at neighboring 4 pixels and compare their values (1 of "green green", "green white", "green black", "white white", "white black", "black black")

Image source: awf.org/sites/default/files/media/gallery/wildlife/Plains%20Zebra/Z-Billy_Dodson_3.jpg?itok=rzMdZ7LM

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

Probability of drawing “White, Black”?

350/5750

Probability of drawing a card that has “White” on it?

(200+2000+350)/5750

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

Place into bag

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

Probability of drawing "White, Black"?

350/5750

Probability of drawing a card that has "White" on it?

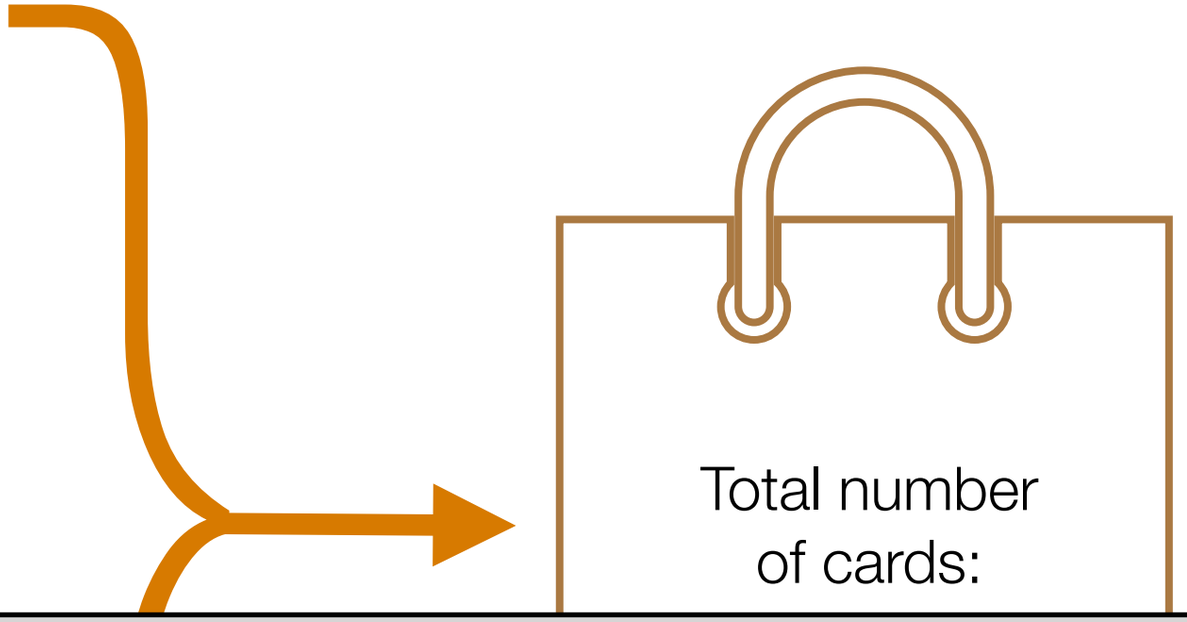
(200+2000+350)/5750

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White



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

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

Measuring Association: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

PMI can be positive or negative

Higher PMI ➔ more "interesting"

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI(Green, White)} = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \text{-1.63... bits}$$

$$\text{PMI(Green, Black)} = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \text{-1.63... bits}$$

-1.63 > -1.69

$$\text{PMI(White, Black)} = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = \text{-1.69... bits}$$

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

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

What is PMI Measuring?

Probability of A and B co-occurring

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

Probability of just A occurring

Probability of just B occurring

If A and B were “independent”

➔ probability of A and B co-occurring would be P(A)P(B)

What is PMI Measuring?

Probability of A and B co-occurring

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

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

if equal to 1
➔ A, B are indep.

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.

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

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

$$\text{Chi-square} = N \times \text{Phi-square}$$

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

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

N = 5750

$$\text{Chi-square} = N \times \text{Phi-square}$$

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

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

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

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

Green, Black: $= 0.0496\ldots$

White, White: $= 0.1161\ldots$

$\vdots$

White, Black: $= 0.0937\ldots$

Black, Black: $= 0.1161\ldots$

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

Interpretation: neighboring pixels not close to being indep.

Back to Earlier Example

		Tesla	Apple	
Elon Musk		300	1	
Tim Cook		1	195	

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!

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

Compute "marginals"

300/497+1/497

1/497+195/497

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

300/497+1/497

1/497+195/497

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)

Recap: Co-Occurrences

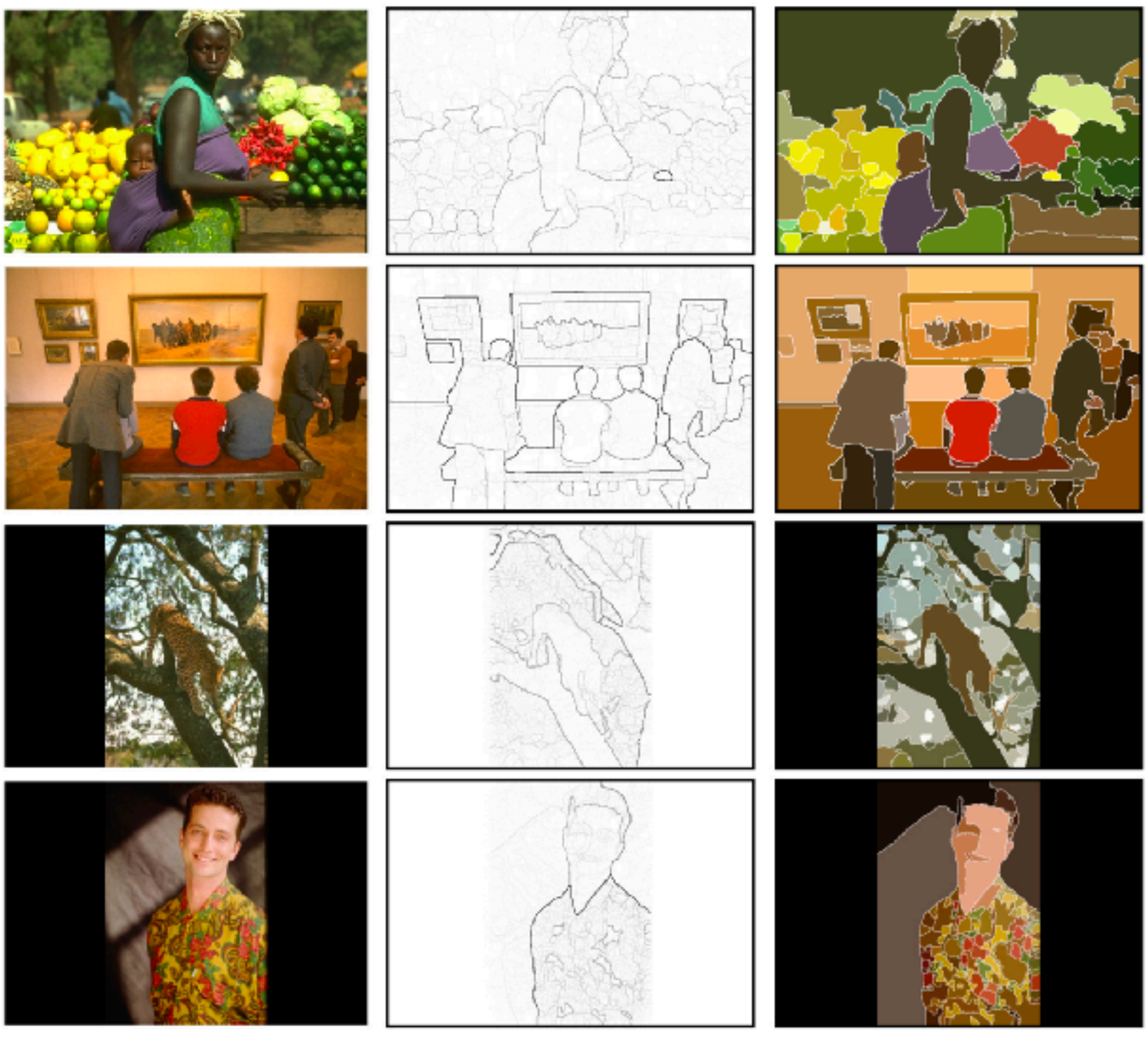
- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”
- Find interesting relationships between pairs of items by looking at PMI
  - Intuition: “Interesting” co-occurring events should occur more frequently than if they were to co-occur independently
- In practice: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

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

Tune parameter $\rho > 0$

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

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

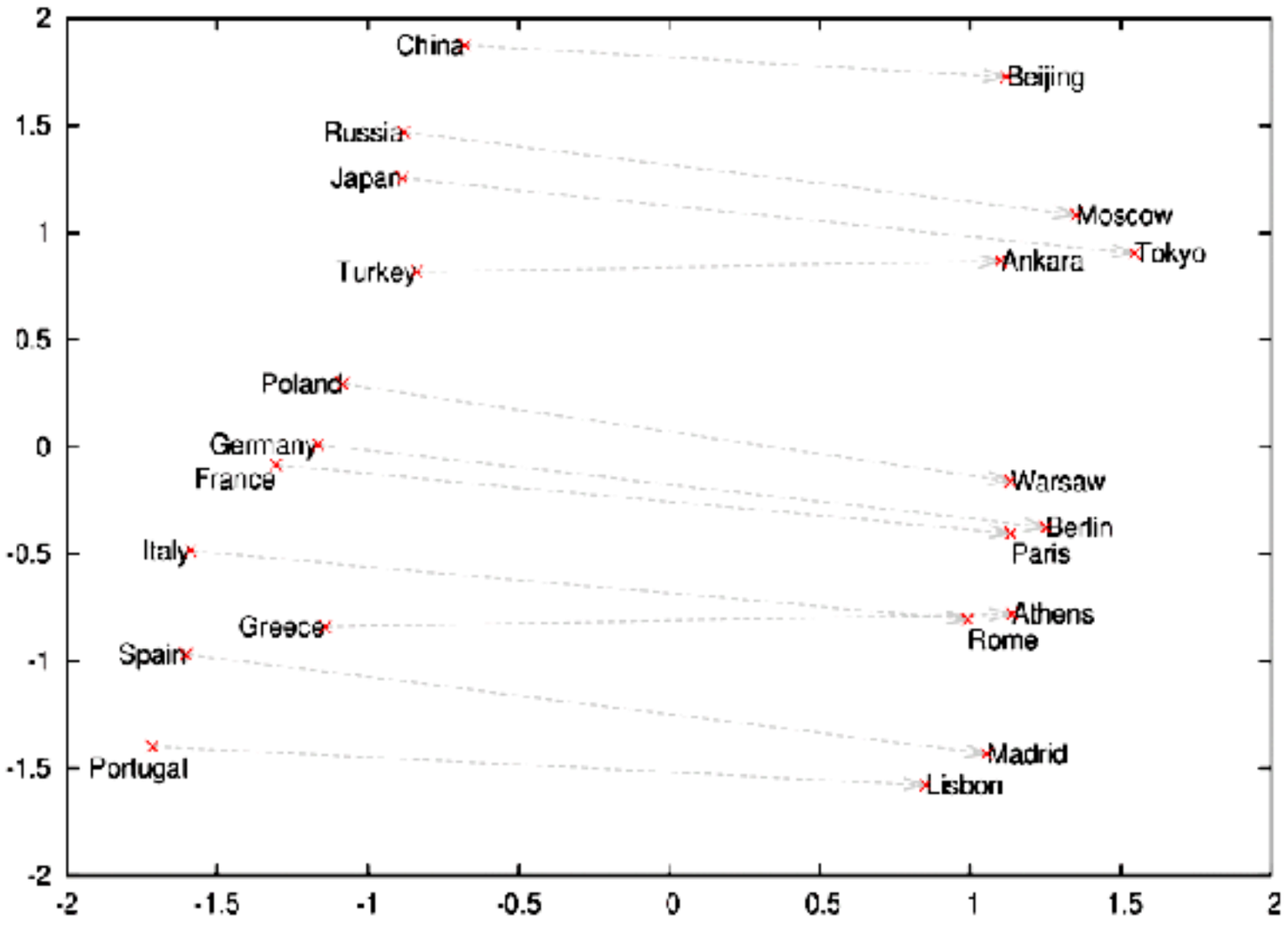
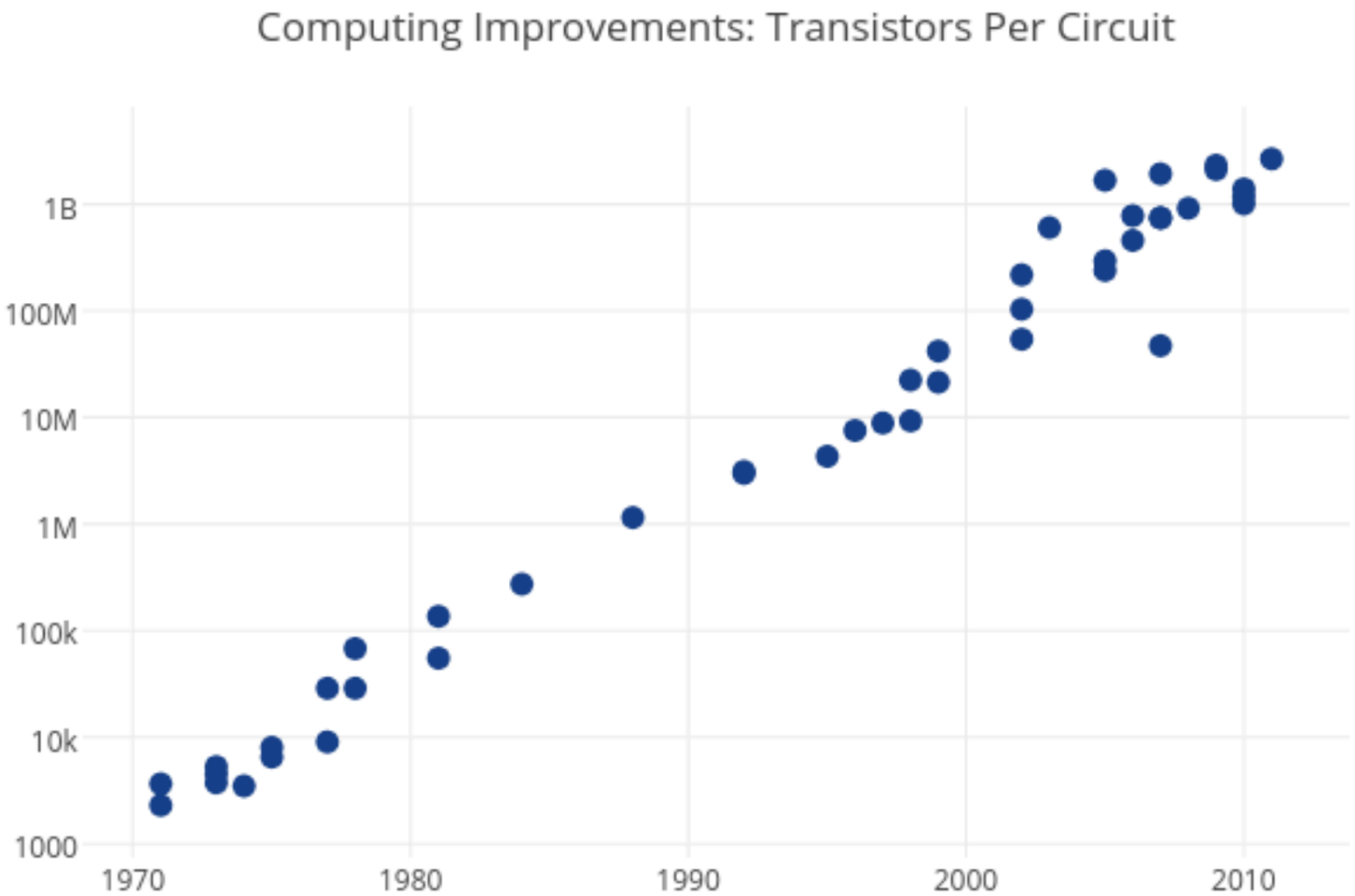


Image source: https://deeplearning4j.org/img/countries_capitals.png

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

Continuous Measurements

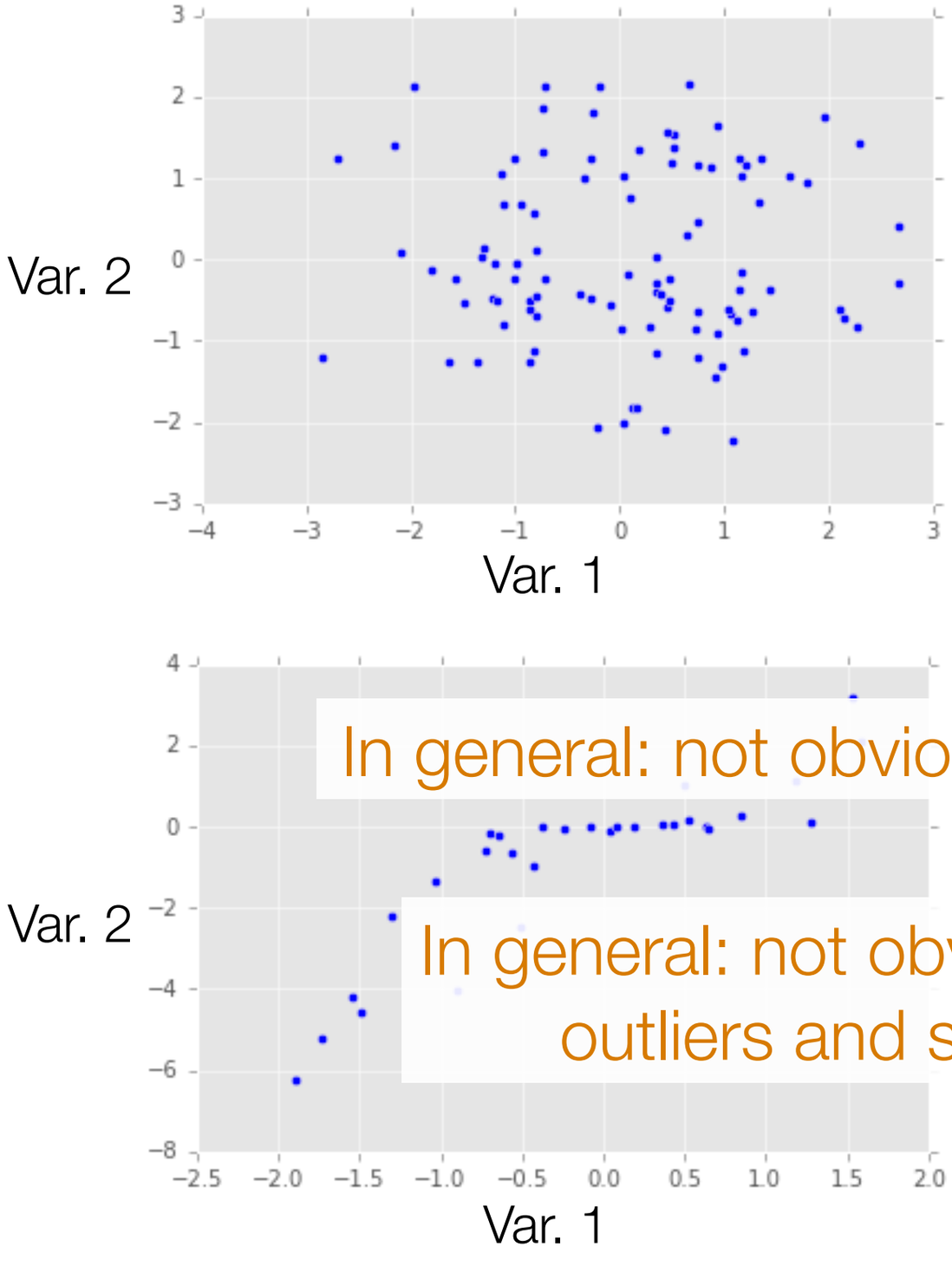
- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**

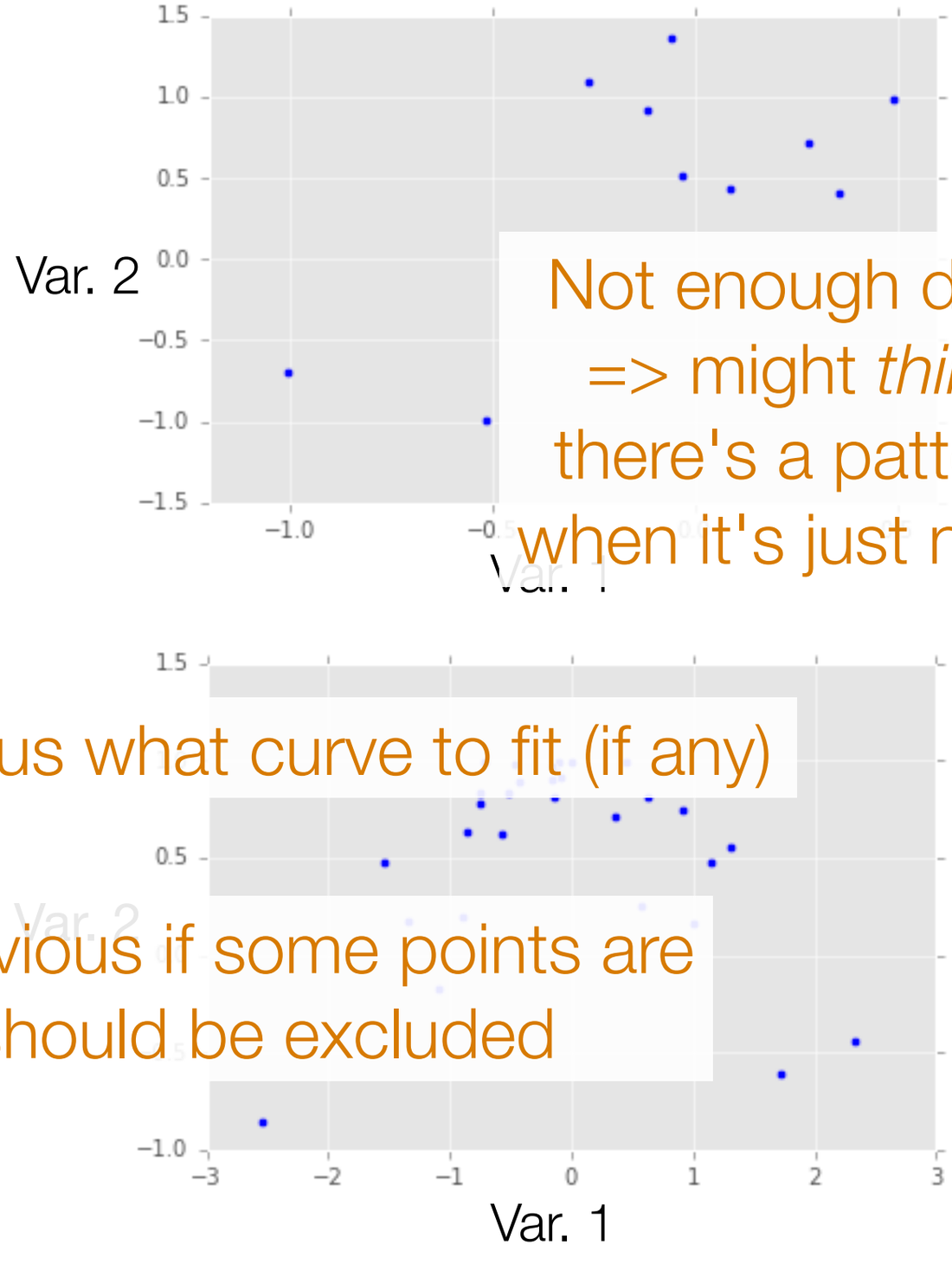


Of course, not all trends look like a line

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

The Importance of Staring at Data



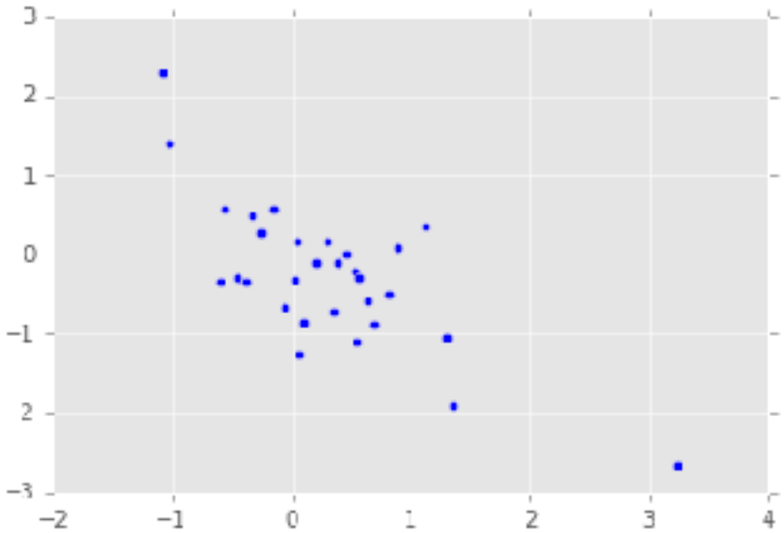


Not enough data
=> might *think*
there's a pattern
when it's just noise

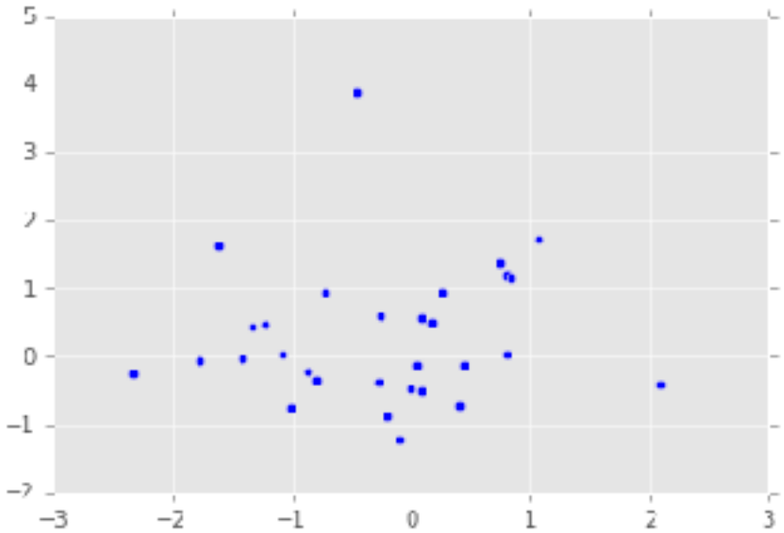
In general: not obvious what curve to fit (if any)

In general: not obvious if some points are outliers and should be excluded

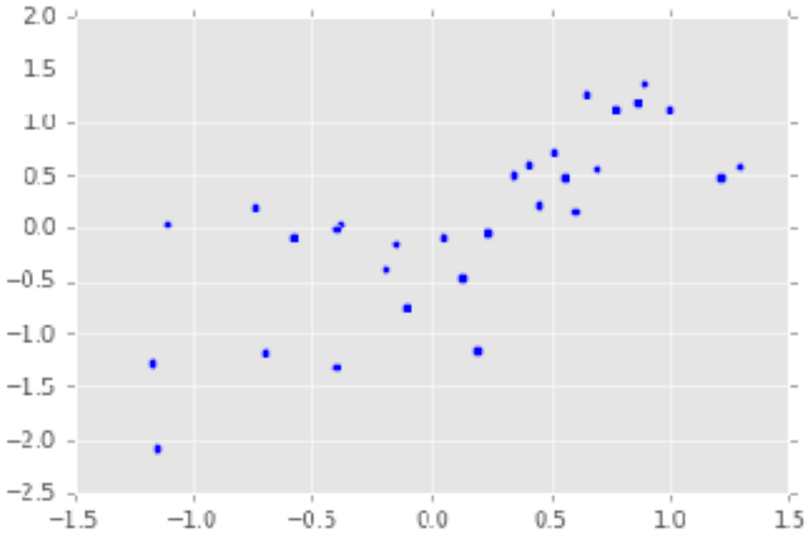
Correlation



Negatively correlated



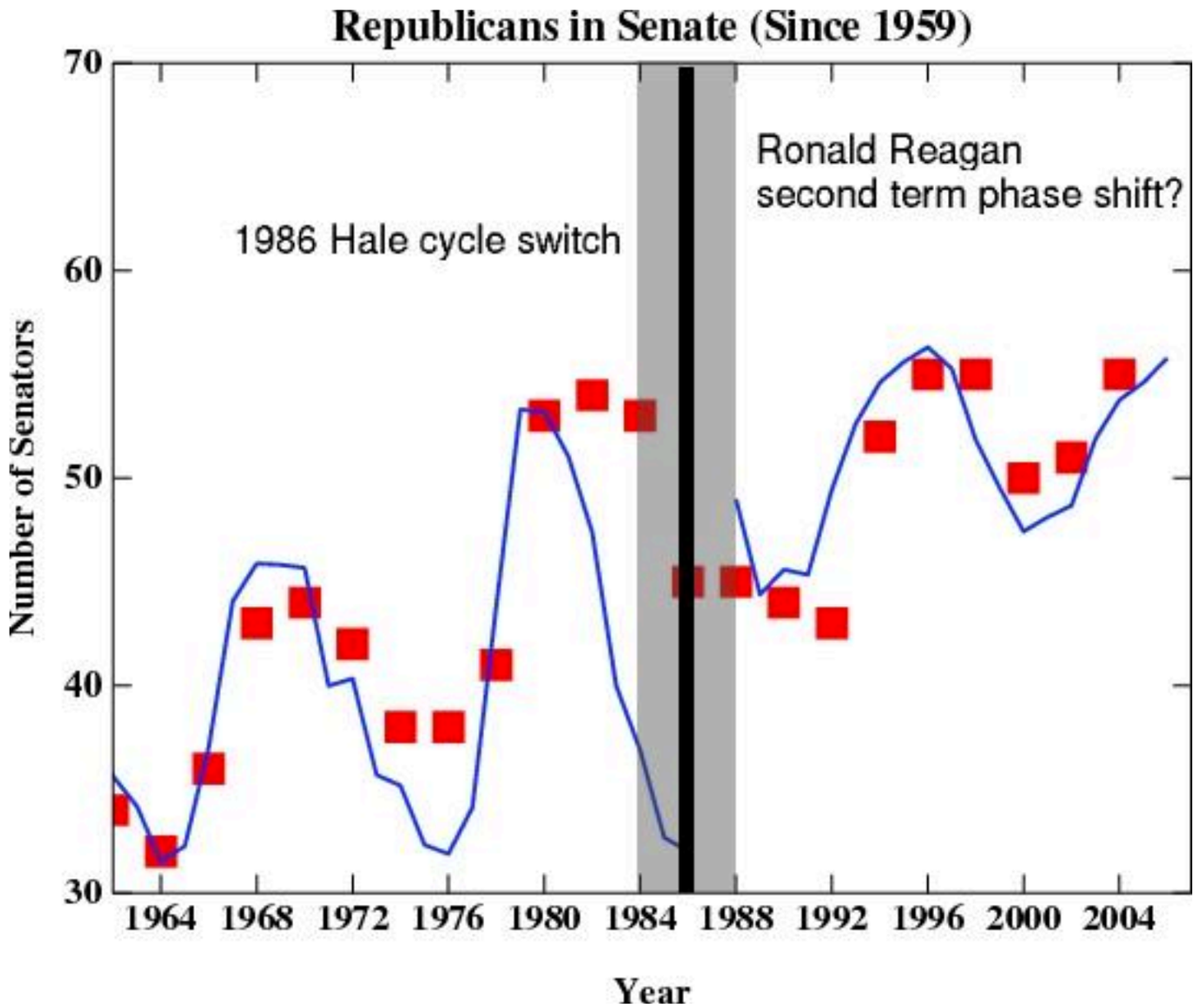
Not really correlated



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?

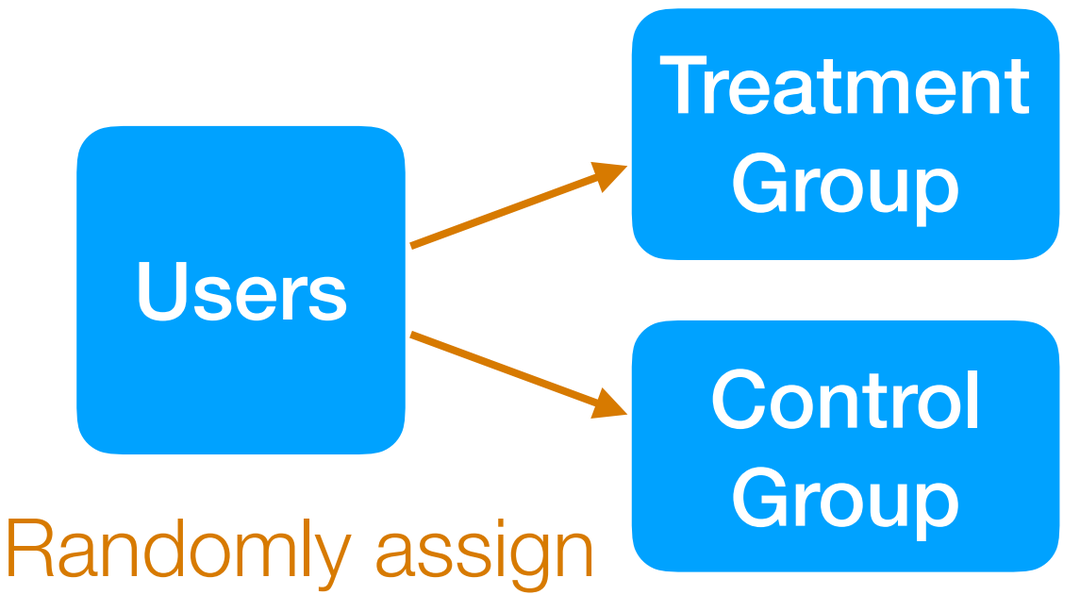
Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

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

Course Outline

Part 1: Identify structure present in “unstructured” data

Exploratory data analysis

- Frequency and co-occurrences *Basic probability & statistics*

- Clustering

Unsupervised learning

- Topic modeling (special kind of clustering)

Part 2: Make predictions using structure found in part 1

Predictive data analysis *Supervised learning*

- Introduction to classification

- Adaptive nearest neighbor methods

- Deep learning models for classification