Carnegie Mellon University
Heinz College

# Unstructured Data Analytics for Policy

## Lecture 3: Co-occurrence analysis (cont'd), visualizing high-dimensional data

George Chen

HW1 is now due

Tuesday March 29, 11:59pm

# (Flashback) Use PMI as a Numerical Score to Rank *Specific* Person/Company Pairs

$$\mathrm{PMI}(A, B) = \log \frac{P(A, B)}{P(A)P(B)}$$

- More positive value means a specific pair appears much more likely than if they were independent
- More negative value means a specific pair appears much less likely than if they were independent
- In practice: need to be careful with named entities that extremely rarely occur
- Sometimes people consider only pairs with positive PMI values to be interesting (called *positive PMI or PPMI*)

# How We Use PMI

| | | |
|---|---|---|
| Elon Musk, Alphabet | | PMI(Elon Musk, Alphabet) |
| Elon Musk, AMD | | PMI(Elon Musk, AMD) |
| Elon Musk, Tesla | | PMI(Elon Musk, Tesla) |
| Sundar Pichai, Alphabet | | PMI(Sundar Pichai, Alphabet) |
| Sundar Pichai, AMD | → Compute PMI's | PMI(Sundar Pichai, AMD) → Sort biggest to smallest |
| Sundar Pichai, Tesla | | PMI(Sundar Pichai, Tesla) |
| Lisa Su, Alphabet | | PMI(Lisa Su, Alphabet) |
| Lisa Su, AMD | | PMI(Lisa Su, AMD) |
| Lisa Su, Tesla | | PMI(Lisa Su, Tesla) |

What about figuring out if
people (as a whole)/companies (as a whole)
is an “interesting” relationship?

# Goal

people, companies

people, products

people, locations

people, dates

companies, products

companies, locations

companies, dates

products, locations

products, dates

locations, dates

rank these pairs from "most interesting" to "least interesting"

For analysis: might want to focus on most interesting pairs

Need a numerical score for "interesting"-ness

**PMI doesn't work here!**

# Score for People/Companies Pair

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

In this slide:
A = person, B = company

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

Phi-squared is between 0 and min(#rows, #cols)-1

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

Chi-squared $= N \times$ Phi-square

where $N$ = sum of all co-occurrence counts

0 ➔ pairs are all indep.

Cramér's V = Sqrt( Phi-squared / [min(#rows, #cols)-1] )

Cramér's V is always between 0 and 1

**To rank different category pairs, we can use Cramér's V**

Why not use phi-squared or chi-squared instead?

# How We Use Cramér's V

| | | | |
|---|---|---|---|
| people, companies | | Cramér's V(people, companies) | |
| people, products | | Cramér's V(people, products) | |
| people, locations | | Cramér's V(people, locations) | |
| people, dates | | Cramér's V(people, dates) | |
| companies, products | → Compute Cramér's V | Cramér's V(companies, products) | → Sort biggest to smallest |
| companies, locations | | Cramér's V(companies, locations) | |
| companies, dates | | Cramér's V(companies, dates) | |
| products, locations | | Cramér's V(products, locations) | |
| products, dates | | Cramér's V(products, dates) | |
| locations, dates | | Cramér's V(locations, dates) | |

# Recap

- Rank specific person/specific company pairs: can use PMI score
  - Other score functions exist, such as **Jaccard index**:

$$\frac{1}{\frac{P(A)}{P(A,B)} + \frac{P(B)}{P(A,B)} - 1}$$

- Rank category pairs (e.g., people/companies, people/locations): can use Cramér's V score
  - Phi-squared/chi-squared/Cramér's V are closely related to each other and you can convert between them
  - If different category pairs have co-occurrence tables of the same size, then we can also rank using phi-squared
  - If different category pairs have co-occurrence tables of the same size & number of co-occurrences, then we can also rank using chi-squared
  - Other scores are possible (such as **mutual information** — this is different from *pointwise* mutual information)

# Co-Occurrence Analysis

Demo

# 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
  re
  - 
- If y
  cre
  - Crime reports & locations

Examples of data to take advantage of:

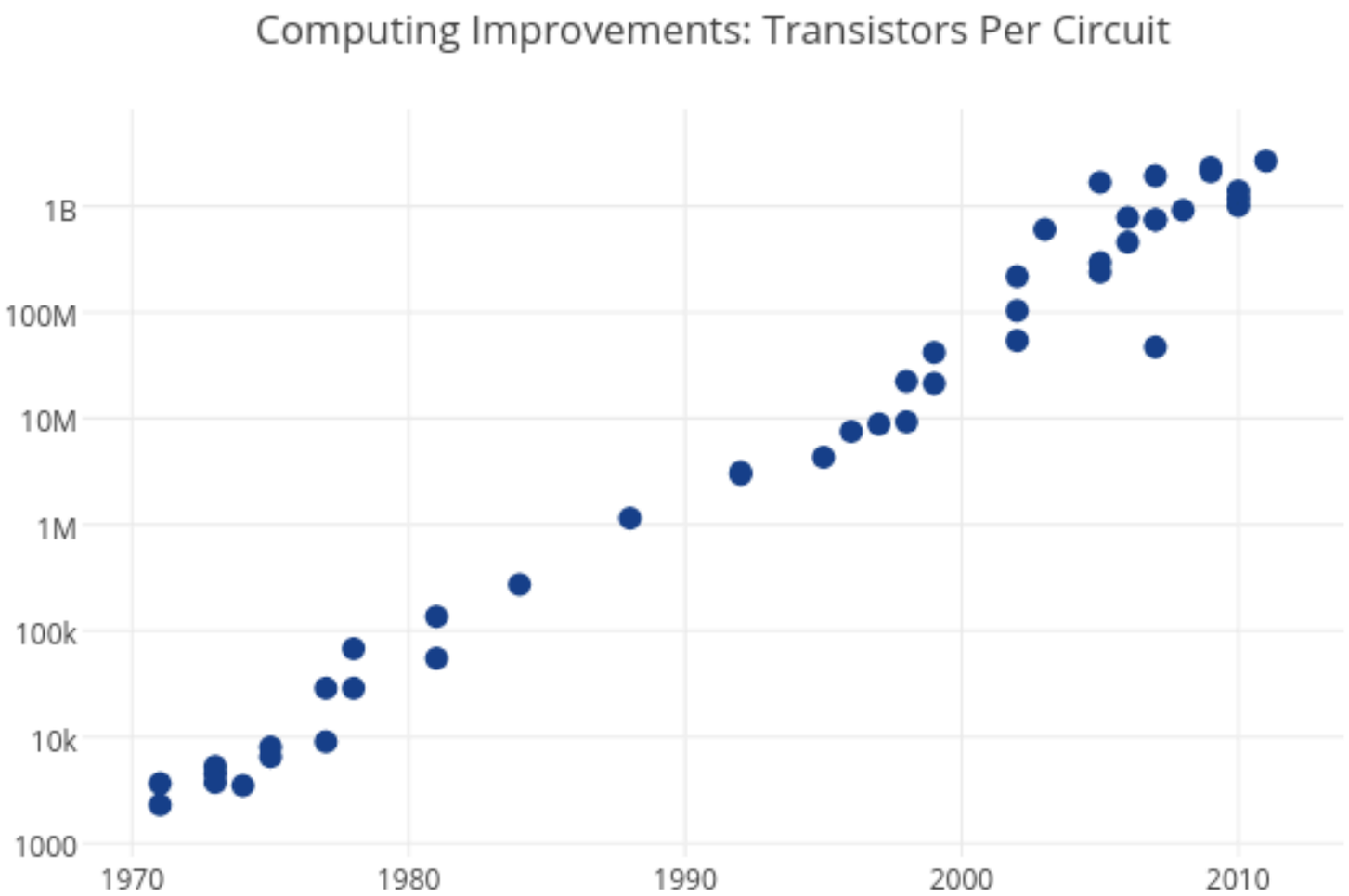
- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:

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

# Continuous Measurements

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

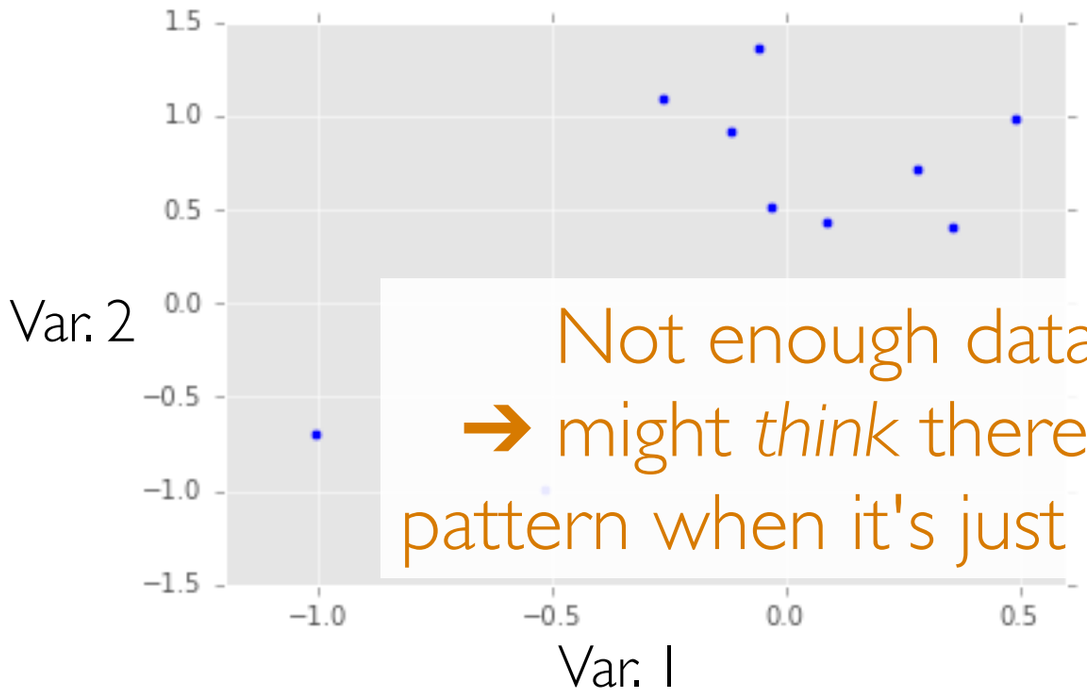



Of course, not all trends look like a line

(so don't just do linear regression!)

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

# The Importance of Staring at Data

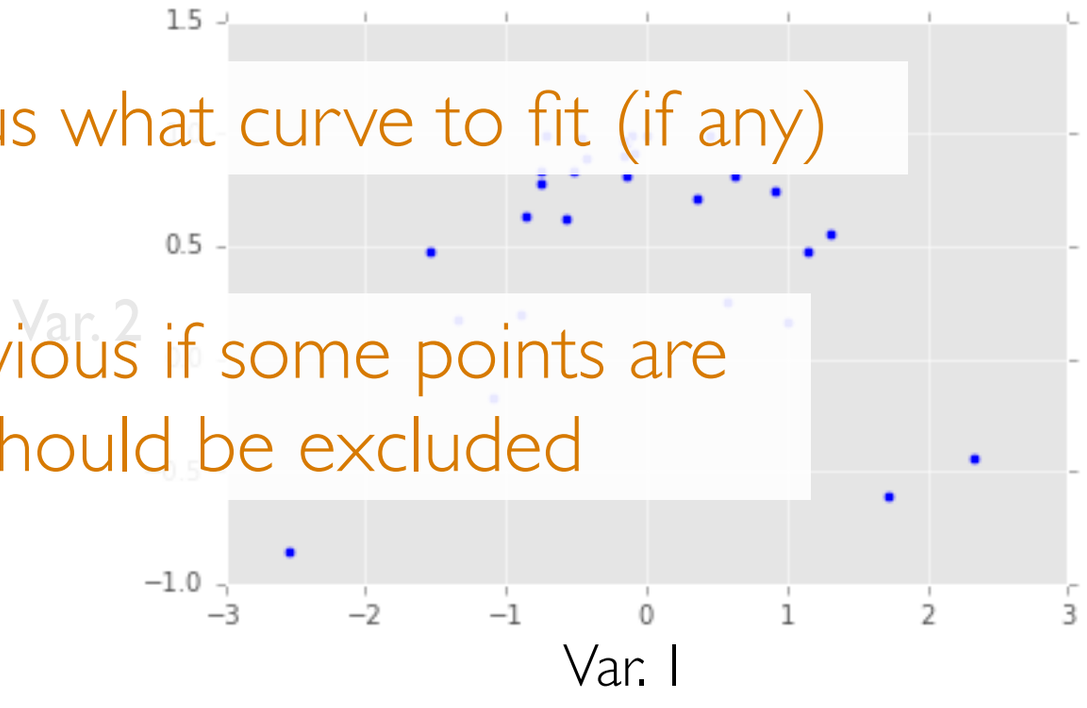




Not enough data
➔ might *think* there's a pattern when it's just noise

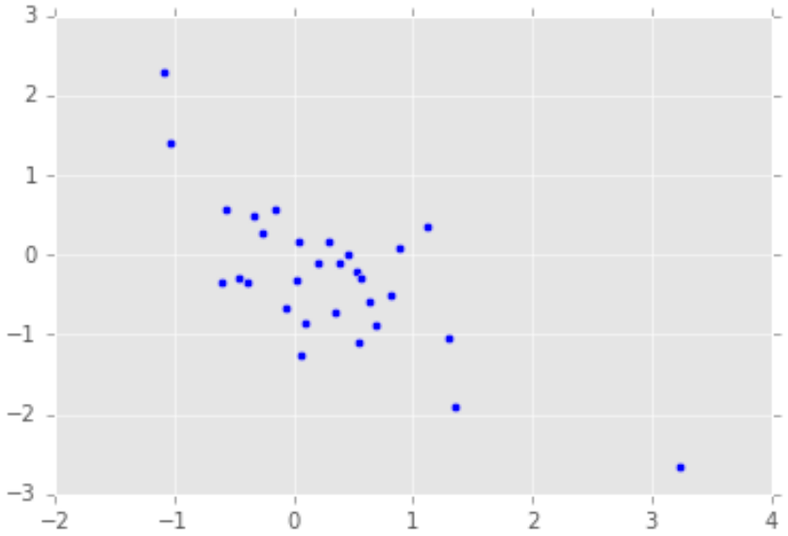
In general: not obvious what curve to fit (if any)

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

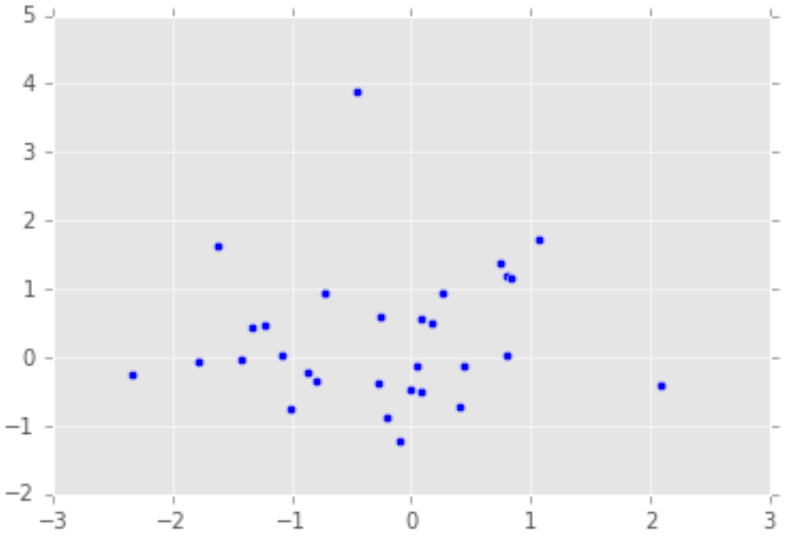


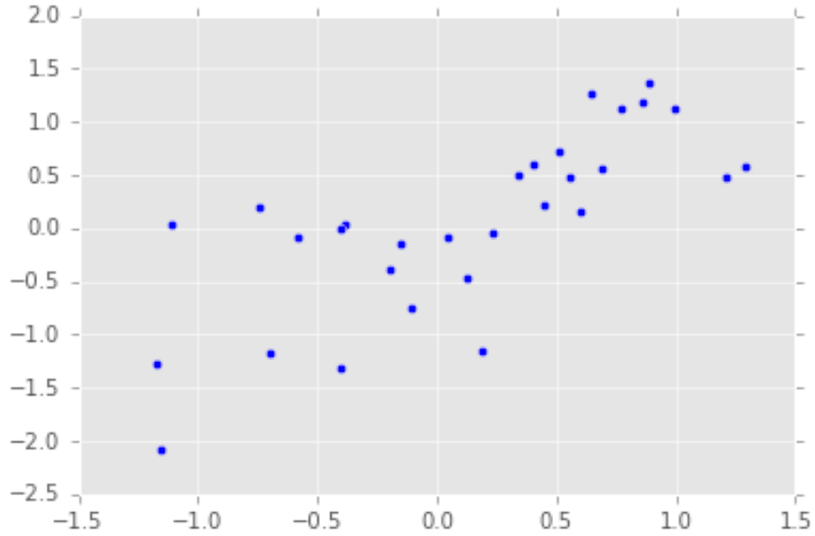
# Correlation



Negatively correlated



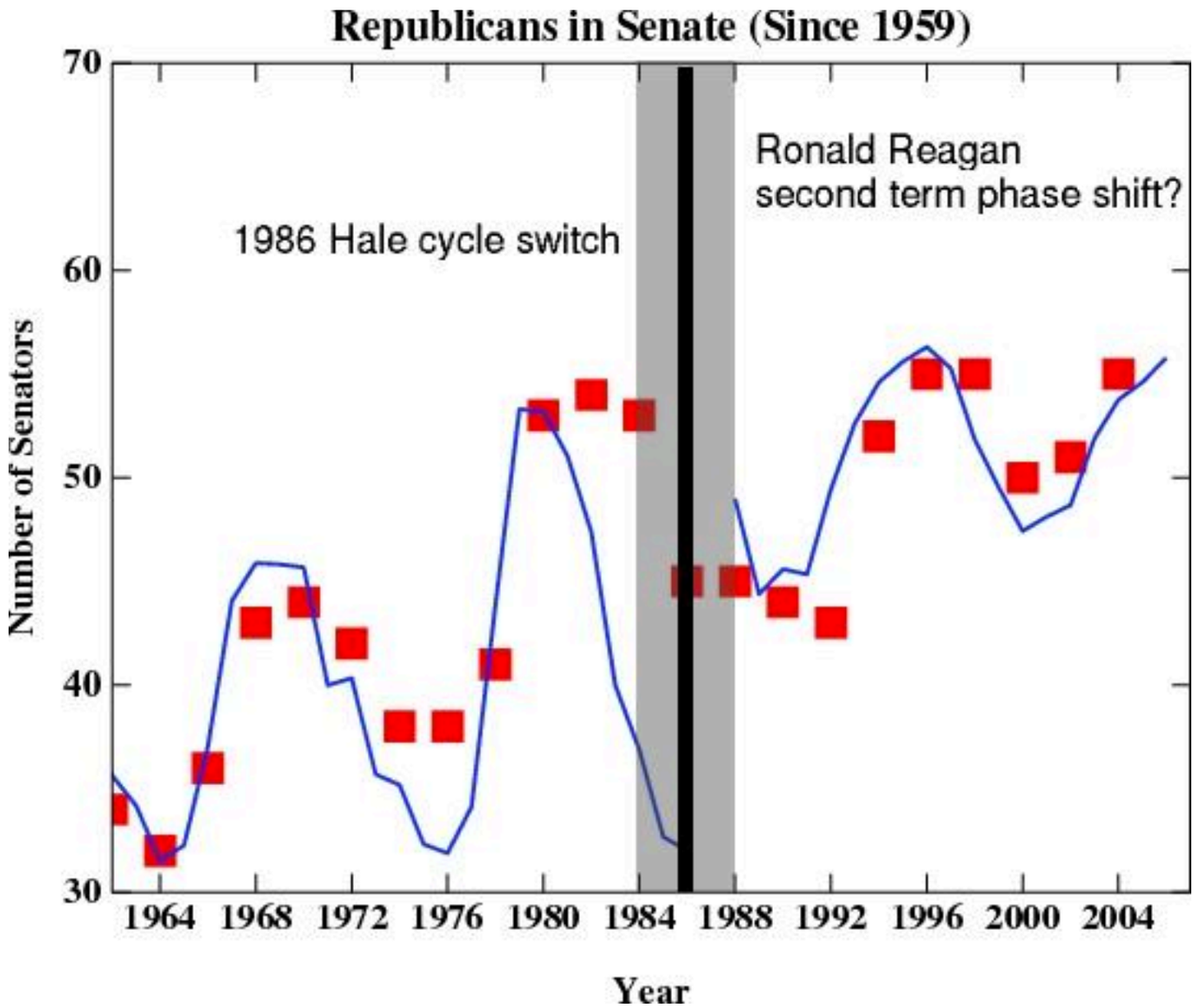
Not really correlated



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**

These are just *candidate* relationships that might be interesting

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)**

# A Recurring Theme: "Design Choices"

- Should I lowercase? Should I lemmatize? How do I count co-occurrences (at the sentence level? paragraph level? document level?), … *lots of design choices!*
  - When you do data analysis for a company/organization, often there is an **infinite number of design choices**
  - There usually will not be someone that tells you what is the "correct" way to choose all of these design choices
  - *You* have to make these decisions!
- **If you're not sure about what to use, try multiple options and see for yourself how the output changes and whether this affects conclusions that are drawn from the analysis!**
  - It's good for you to figure out which design choices lead to significant changes and which do not

# Course Outline

Part I: Exploratory data analysis

*Identify structure present in "unstructured" data*

- Frequency and co-occurrence analysis — *Basic probability & statistics*
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling

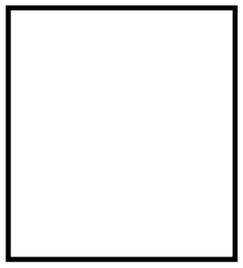
Part II: Predictive data analysis

*Make predictions using known structure in data*

- Basic concepts and how to assess quality of prediction models
- Neural nets and deep learning for analyzing images and text

# Visualizing High-Dimensional Data

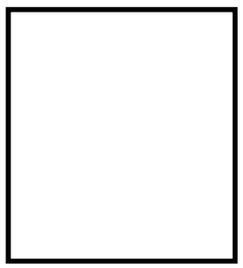
# So Far…



Text doc #1

Feature vector #1
(histogram)

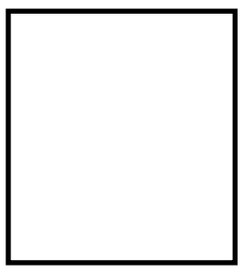
We can visualize this histogram



Text doc #2

Feature vector #2
(histogram)

How do we visualize all ***n*** text docs at once if ***n*** is large?

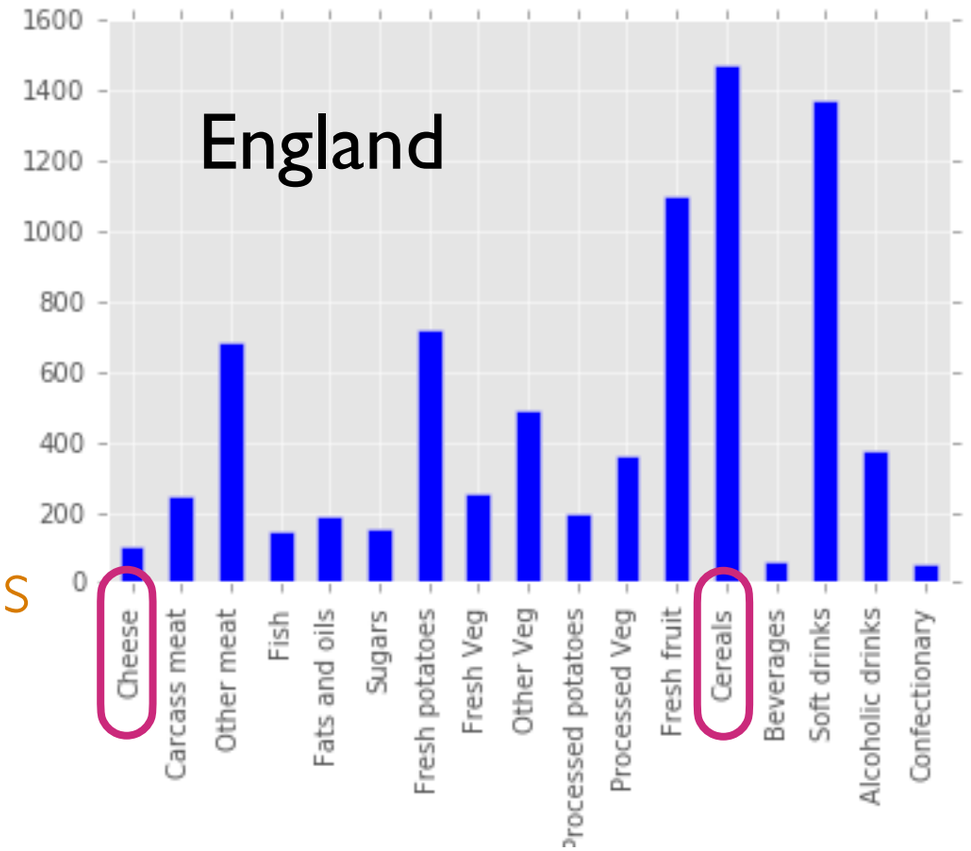


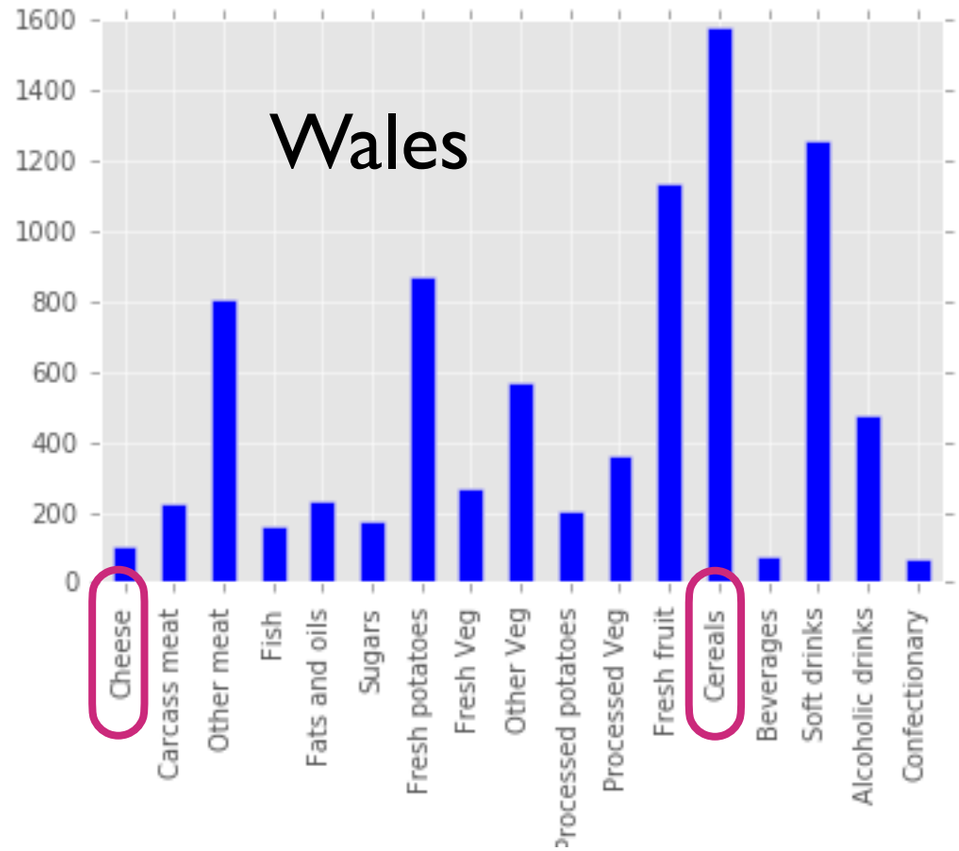
Text doc #***n***

Feature vector #***n***
(histogram)

Here's another concrete example

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 visualize the constituent countries for comparison accounting for all 17 features at once

*Source: http://setosa.io/ev/principal-component-analysis/*

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

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

We will begin with the most famous dimensionality reduction method: principal component analysis (PCA)