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

Users

Treatment Group

Control Group

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 I: Exploratory data analysis

*Identify structure present in “unstructured” data*

- Frequency and co-occurrence analysis
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling

Part II: Predictive data analysis

*Make predictions using known structure in data*

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

Lecture 4: Visualizing high-dimensional data

George Chen
So Far...

Text doc #1
Feature vector #1 (histogram)

Text doc #2
Feature vector #2 (histogram)

We can visualize this histogram

How do we visualize all $n$ text doc’s at once if $n$ is large?

Text doc $n$
Feature vector $n$ (histogram)
Here’s another concrete example
Imagine we had hundreds of 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)!

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 the dimensionality of the data preferably to 1, 2, or 3.
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

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!

Most variability is along this green direction

Rotate!
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Most variability is along this green direction.
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Most variability is along this green direction.

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 $\rightarrow$ 2D as well (and just involves rotating, and not "flattening" the data)

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

Most variability is along this green direction
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