CS 540 Introduction to Artificial Intelligence
Statistics & Math Review

Josiah Hanna
University of Wisconsin-Madison

September 21, 2021

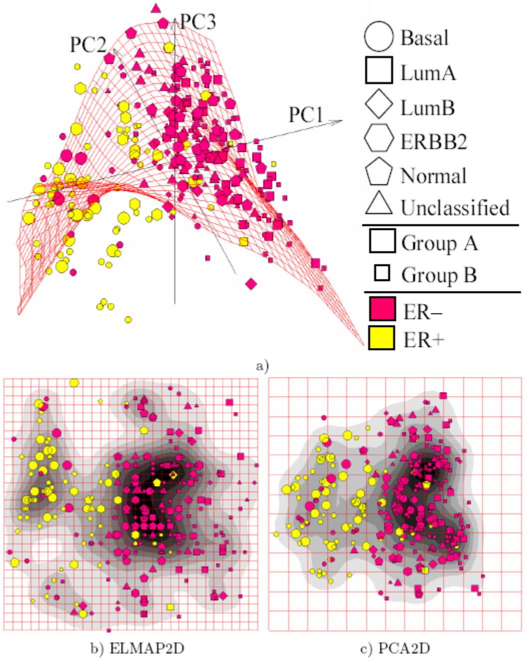
Announcements

- **Homeworks**:
  - HW2 released after class; due next Tuesday
- Class roadmap:

Date	Topic	Reading materials	Assignments
Thursday, Sept 9	Welcome and Course Overview	Slides	
Tuesday, Sept 14	Probability	Slides	HW 1 Released
Thursday, Sept 16	Linear Algebra and PCA	Slides	
Tuesday, Sept 21	Statistics and Math Review		HW 1 Due, HW 2 Released
Thursday, Sept 23	Introduction to Logic		
Everything below here is tentative and subject to change.			
Tuesday, Sept 28	Natural Language Processing		HW 2 Due, HW 3 Released
Thursday, Sept 30	Machine Learning: Introduction		

Outline

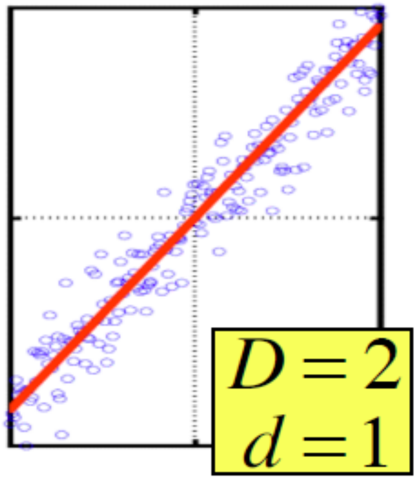
- Finish last lecture: **PCA**
- Review of probability
- Statistics: sampling & estimation

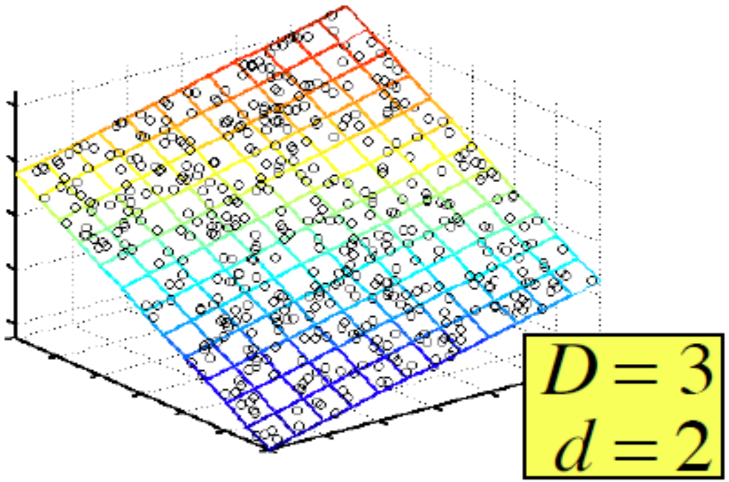


Wikipedia

Principal Components Analysis (**PCA**)

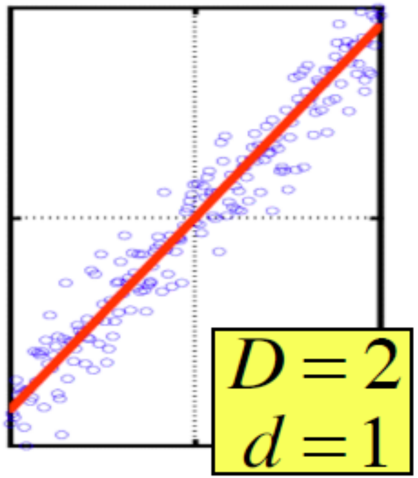
- A type of dimensionality reduction approach
  - For when data is **approximately lower dimensional**

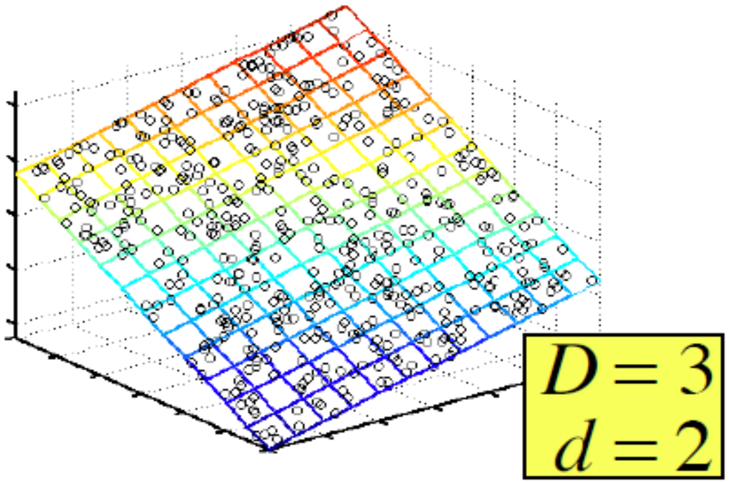




Principal Components Analysis (**PCA**)

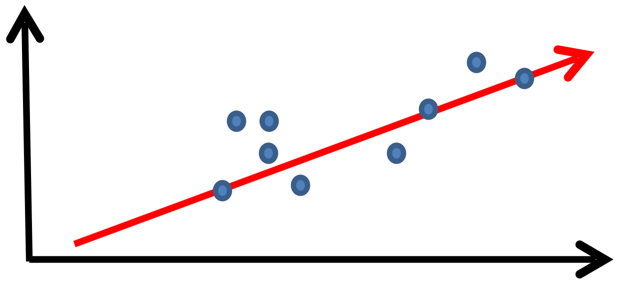
- Goal: find **axes** of a subspace
  - Will project to this subspace; want to preserve data





Principal Components Analysis (**PCA**)

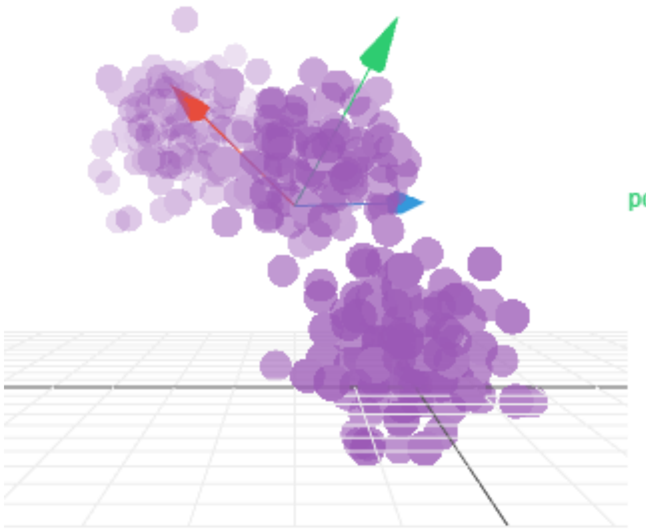
- From 2D to 1D:
  - Find a $v_1 \in \mathbb{R}^d$ so that we maximize "variability"
  - IE,



  - New representations are along this vector (1D!)

Principal Components Analysis (**PCA**)

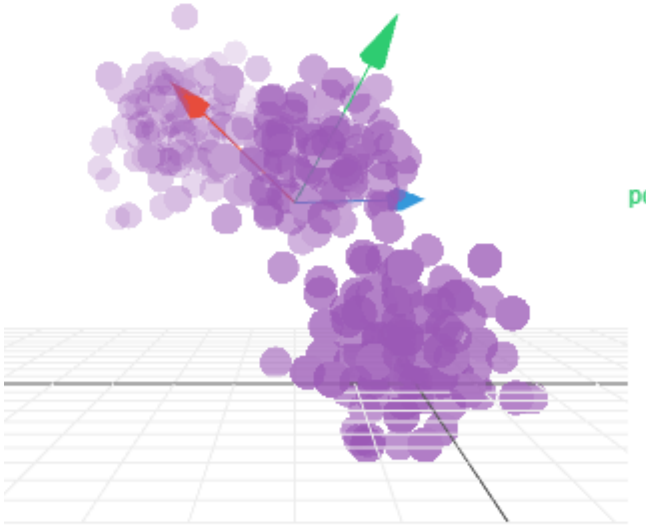
- From *d* dimensions to *r* dimensions:
  - Sequentially get $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$
  - Orthogonal!
  - Still minimize the projection error
    - Equivalent to **"maximizing variability"**
  - The vectors are the **principal components**



Victor Powell

PCA Setup

- **Inputs**
  - Data: $x_1, x_2, \ldots, x_n, \;\; x_i \in \mathbb{R}^d$
  - Can arrange into $X \in \mathbb{R}^{n \times d}$
  - **Centered**! $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$
- **Outputs**
  - Principal components $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$
  - Orthogonal!



Victor Powell

PCA Goals

- Want directions/components (unit vectors) so that
  - Projecting data maximizes variance
  - What's projection?

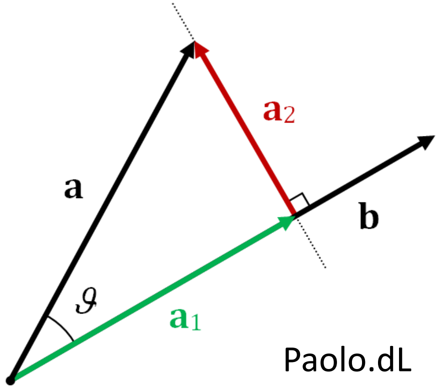
$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

  - To project *a* onto unit vector *b*,

$$\langle a, b \rangle b$$

Direction

Length



PCA Goals

- Want directions/components (unit vectors) so that
  - Projecting data maximizes variance
  - What's projection?

$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

- Do this **recursively**
  - Get orthogonal directions $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$

PCA First Step

- First component,

$$v_1 = \arg\max_{\|v\|=1} \sum_{i=1}^{n} \langle v, x_i \rangle^2$$

- Same as getting

$$v_1 = \arg\max_{\|v\|=1} \|Xv\|^2$$

PCA Recursion

- Once we have *k-1* components, next?

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X v_i v_i^T$$

Deflation

- Then do the same thing

$$v_k = \arg\max_{||v||=1} ||\widehat{X}_k v||^2$$

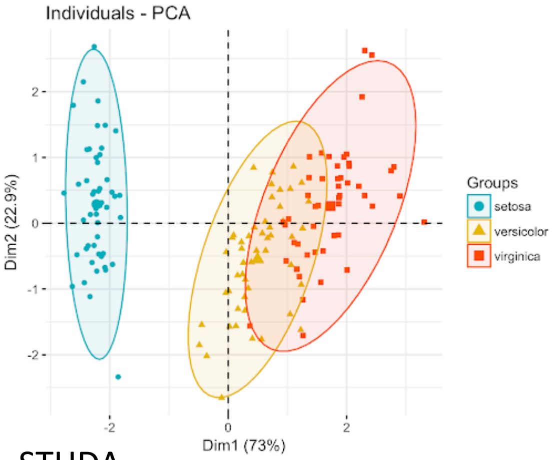
PCA Interpretations

- The v's are eigenvectors of X^TX **(Gram matrix)**
  - Show via Rayleigh quotient
- X^TX (proportional to) sample covariance matrix
  - When data is 0 mean!
  - I.e., PCA is eigendecomposition of sample covariance

- Nested subspaces *span(v1), span(v1,v2),…,*

Lots of Variations

- PCA, Kernel PCA, ICA, CCA
  - Unsupervised techniques to extract structure from high dimensional dataset
- Uses:
  - **Visualization**
  - Efficiency
  - Noise removal
  - Downstream machine learning use



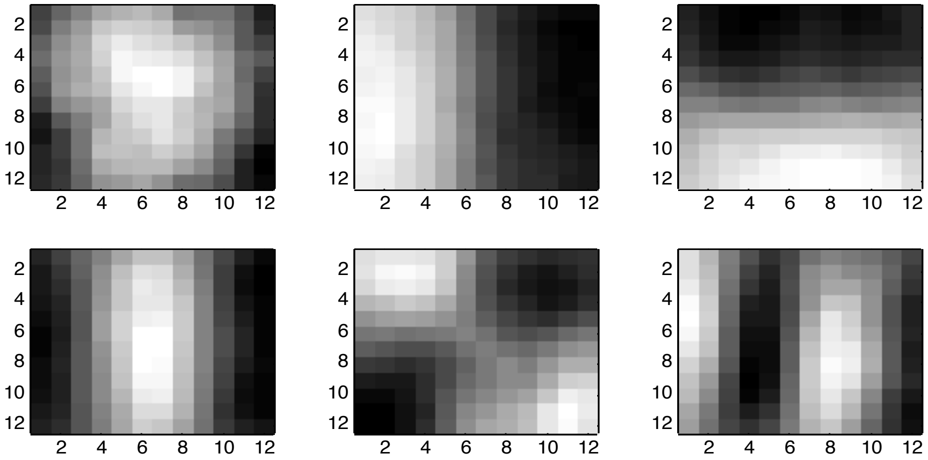
STHDA

Application: Image Compression

- Start with image; divide into 12x12 patches
  - I.E., 144-D vector
  - **Original image:**

Application: Image Compression

- 6 most important components (as an image)



Application: Image Compression

- Project to 6D,

Compressed

Original

Break & Quiz

Q 1.1: What is the projection of $[1\ 2]^T$ onto $[0\ 1]^T$?

- A. $[1\ 2]^T$
- B. $[-1\ 1]^T$
- C. $[0\ 0]^T$
- D. $[0\ 2]^T$

Break & Quiz

Q 1.1: What is the projection of $[1\ 2]^T$ onto $[0\ 1]^T$?

- A. $[1\ 2]^T$
- B. $[-1\ 1]^T$
- C. $[0\ 0]^T$
- **D. $[0\ 2]^T$**

Break & Quiz

Q 1.2: We wish to run PCA on 10-dimensional data in order to produce r-dimensional representations. Which is the most accurate?

- A. $r \leq 3$
- B. $r < 10$
- C. $r \leq 10$
- D. $r \leq 20$

Break & Quiz

Q 1.2: We wish to run PCA on 10-dimensional data in order to produce *r*-dimensional representations. Which is the most accurate?

- A. $r \leq 3$
- B. $r < 10$
- **C. $r \leq 10$**
- D. $r \leq 20$

Probability Review: Outcomes & Events

- Outcomes: possible results of an **experiment**
- **Events:** subsets of outcomes we're interested in

Ex: $\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$

$$\mathcal{F} = \underbrace{\{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \Omega\}}_{\text{events}}$$

Review: Probability Distribution

- We have outcomes and events.
- Now assign probabilities $\text{For } E \in \mathcal{F},\ P(E) \in [0, 1]$

Back to our example:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}}_{\text{events}}$$

$$P(\{1, 3, 5\}) = 0.2, P(\{2, 4, 6\}) = 0.8$$

Review: Random Variables

- Map outcomes to real values $X : \Omega \to \mathbb{R}$
- Can still work with probabilities:

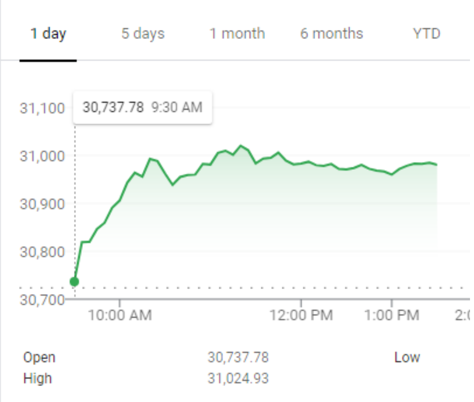
$$P(X = 3) := P(\{\omega : X(\omega) = 3\})$$

- Cumulative Distribution Function (CDF)

$$F_X(x) := P(X \leq x)$$

Review: Expectation & Variance

- Another advantage of RVs are ``summaries''
- Expectation: $E[X] = \sum_a a \times P(x = a)$
  - The "average"
- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of spread
- Higher moments: other parametrizations



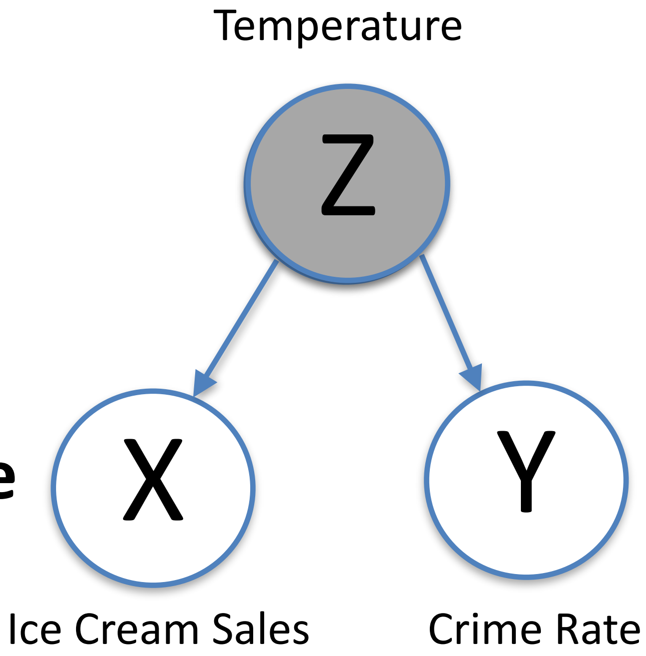
Review: Conditional Probability

- For when we know something,

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

- Leads to **conditional independence**

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$



Review: Bayesian Inference

- Conditional Prob. & Bayes:

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1, \ldots, E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- Has more evidence.
  - Likelihood is hard---but **conditional independence assumption**

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H)\cdots, P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

Review: Classification

- Expression

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H) \cdots , P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- *H*: some class we'd like to infer from evidence
  - Estimate prior *P(H)*
  - Estimate $P(E_i|H)$ from data!
  - How?

Samples and Estimation

- Usually, we don't know the distribution (P(X))
  - Instead, we see a bunch of samples

$$x_1, x_2, \ldots, x_n$$

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability *P(X)*
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

Samples and Estimation

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability *P(X)*
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

- Example: Bernoulli with parameter *p*
  - Mean $E[X]$ is p

Samples and Estimation

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability *P(X)*
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

- Example: Neural network
  - Model of $P_\theta(X)$ with connection weights as parameters

Examples: Sample Mean

- Bernoulli with parameter *p*
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - No different from counting heads

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $E[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $E[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $E[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $E[X^2]$

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

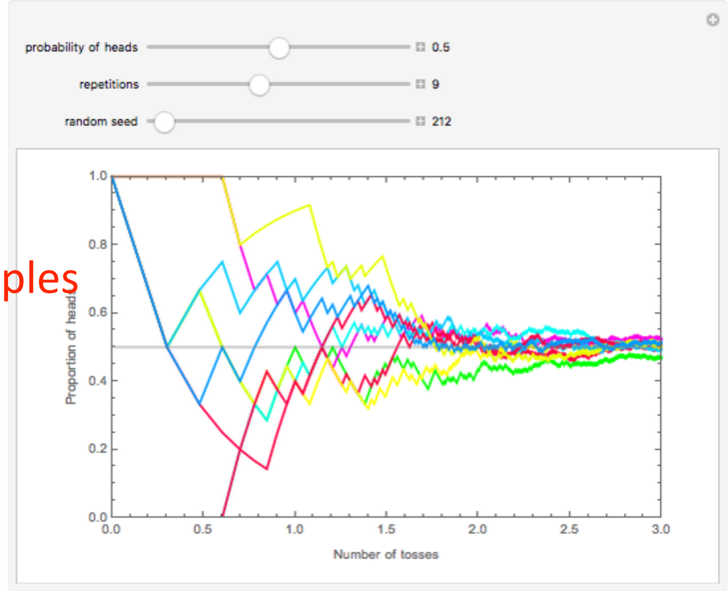
Estimation Theory

- How do we know that the sample mean is a good estimate of the true mean?
  - Concentration inequalities

$$P(|\mathbb{E}[X] - \hat{\mathbb{E}}[X]| \geq t) \leq \exp(-2nt^2)$$

True Mean
Estimate
Samples

  - Law of large numbers
  - Central limit theorems, etc.



Wolfram Demo

Estimation Error

- With finite samples, likely error in estimate.
- Mean Squared Error

$$MSE[\hat{\theta}] = E[(\underset{\text{Estimate}}{\hat{\theta}} - \underset{\text{True Value}}{\theta})^2]$$

- Bias / Variance Decomposition

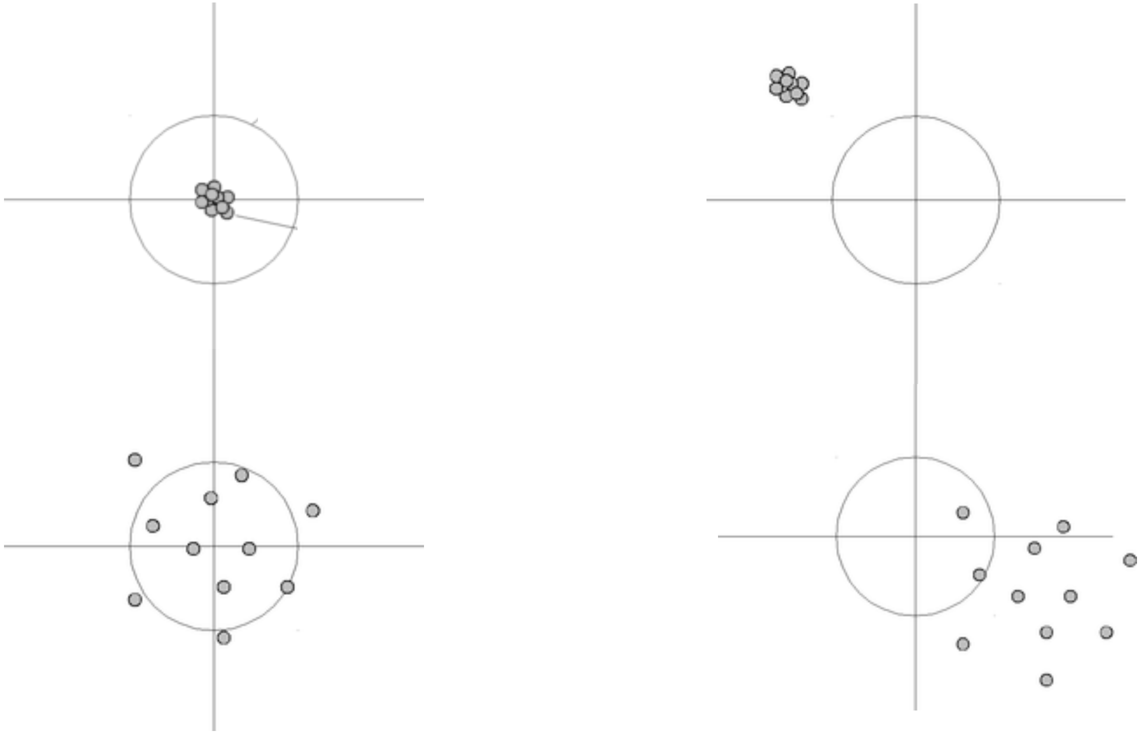
$$MSE[\hat{\theta}] = \underbrace{E[(\hat{\theta} - E[\hat{\theta}])^2]}_{\text{Variance}} + \underbrace{(E[\hat{\theta}] - \theta)^2}_{\text{Bias Squared}}$$

Bias / Variance

Low Bias High Bias

Low Variance

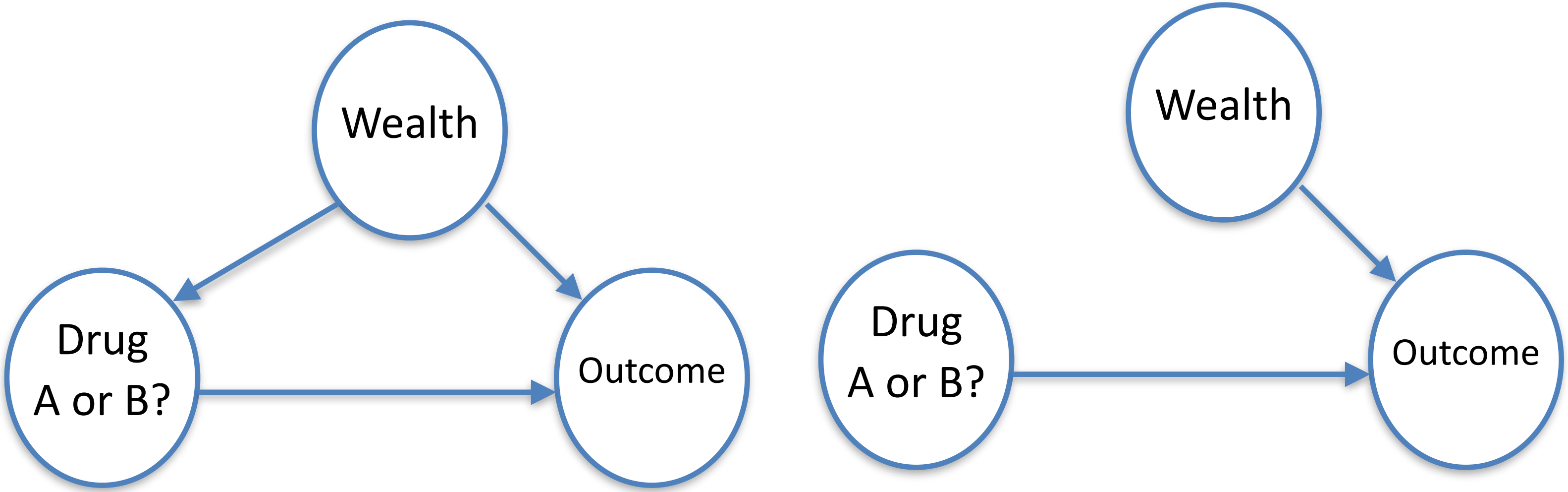
High Variance

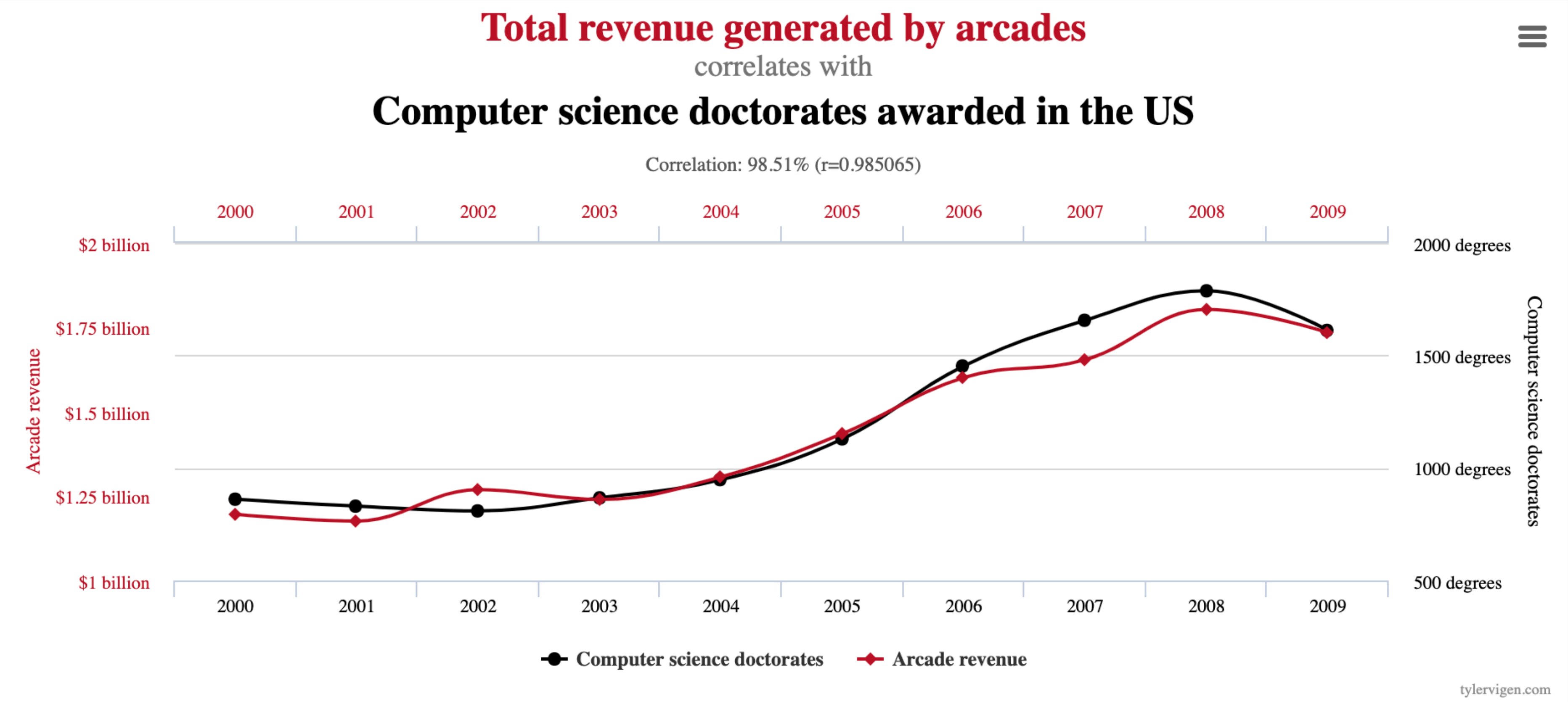


Wikipedia: Bias-variance tradeoff

Association vs Causation

- Conditional distributions give associational relationships
  - $P(Y|X)$ is not necessarily the causal effect of X on Y





Data sources: U.S. Census Bureau and National Science Foundation