DATA8015 Math Foundation of Data Science

Review: Probability and Statistics

Probability: What is it good for?

- Language to express **uncertainty**

In AI/ML Context

- Quantify predictions

$[p(\text{lion}), p(\text{tiger})] = [0.98, 0.02]$

$[p(\text{lion}), p(\text{tiger})] = [0.01, 0.99]$

$[p(\text{lion}), p(\text{tiger})] = [0.43, 0.57]$

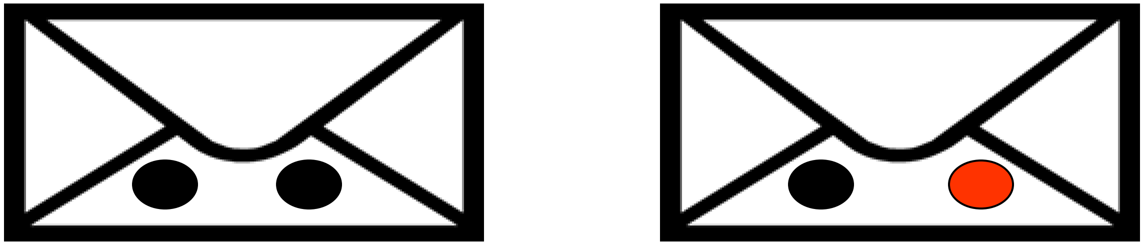
Model Data Generation

- Model complex distributions

StyleGAN2 (Kerras et al ’20)

Probabilistic Decision Making Example: Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth \$100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**



Statistical Learning Example: Flu Diagnosis Problem

- Evaluating probabilities:
  - Wake up with a sore throat
  - Do I have the flu?
- Logic approach: $S \rightarrow F$? Too strong
- **Inference**: estimate probability given evidence (records of 1000 persons)
  - Sore throat: 100 persons; With Flu: 10; Sore throat among flu sufferers: 9

Outline

- Basics: definitions, axioms, RVs, joint distributions
- Independence, conditional probability, chain rule
- Bayes' Rule and Inference

Basics: Outcomes & Events

- **Outcomes:** possible results of an **experiment**

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

- **Events:** subsets of outcomes we're interested in

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

- Always include $\emptyset, \Omega$

Basics: Probability Distribution

- We have outcomes and events.
- Assign **probabilities**: for each event $E, P(E) \in [0,1]$

Back to our example:

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

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

Basics: **Axioms**

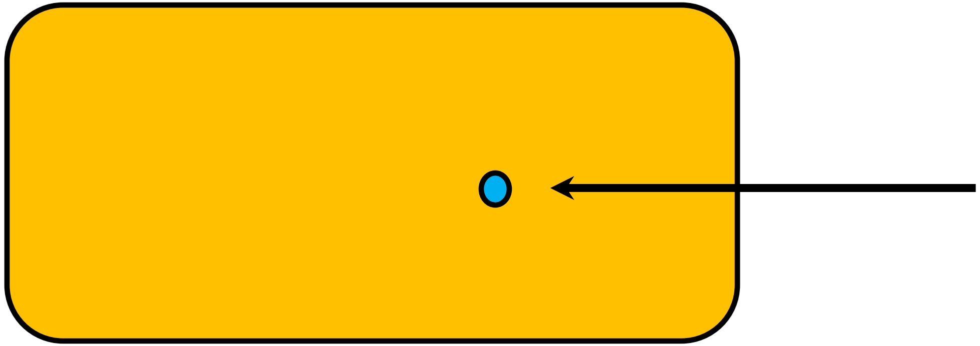
- Rules for probability:
  - For all events E, $P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

- Easy to derive other laws. Ex: non-disjoint events

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Visualizing the Axioms: I

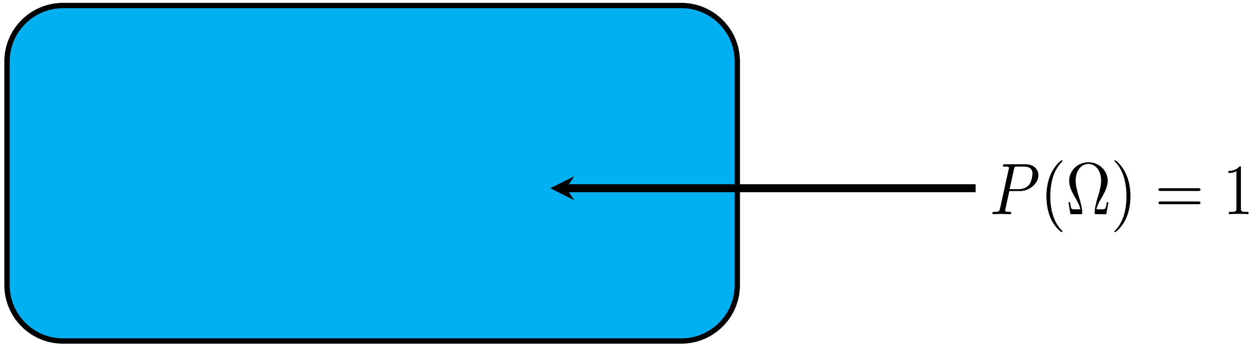
- Axiom 1: for all events E, $P(E) \geq 0$



$P(E) \geq 0$

Visualizing the Axioms: **II**

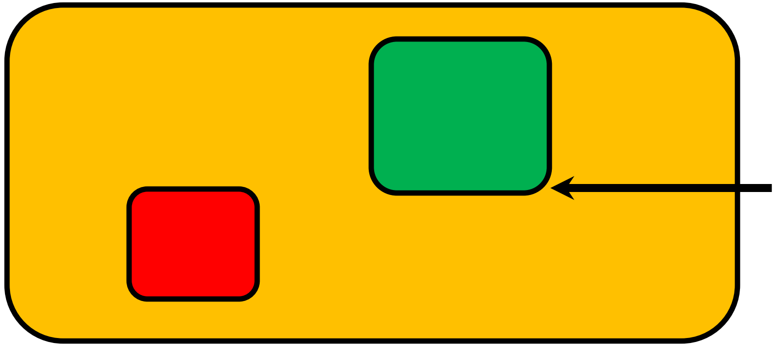
- Axiom 2: $P(\emptyset) = 0, P(\Omega) = 1$



Visualizing the Axioms: **III**

- Axiom 3: disjoint $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

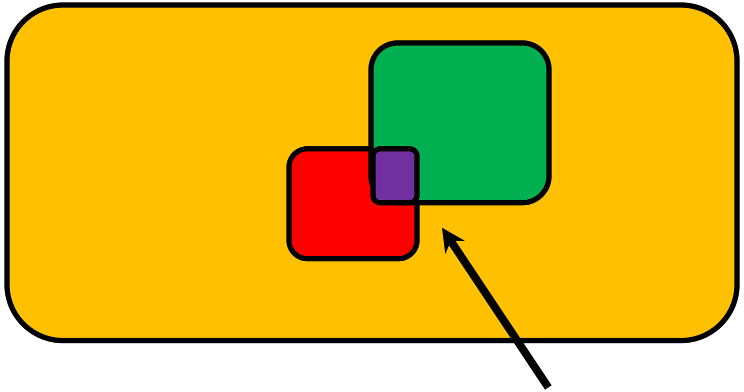
$\boxed{E_1 \textbf{ or } E_2}$



$$P({\color{red}E_1} \cup {\color{green}E_2}) = P({\color{red}E_1}) + P({\color{green}E_2})$$

Visualizing the Axioms

- Also, other laws:



E_1 **and** E_2

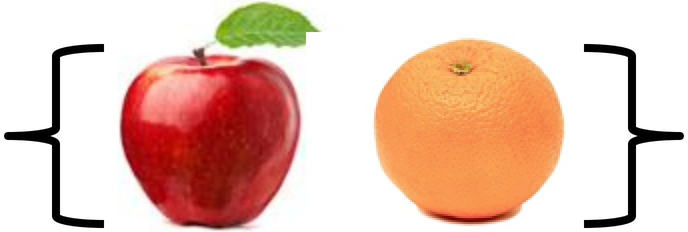
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Basics: **Random Variables**

- Intuitively: a number X that's random
- Mathematically: map random outcomes to real values

$$X : \Omega \to \mathbb{R}$$

- Why?
  - Previously, everything is a set.
  - Real values are easier to work with



Basics: **CDF & PDF**

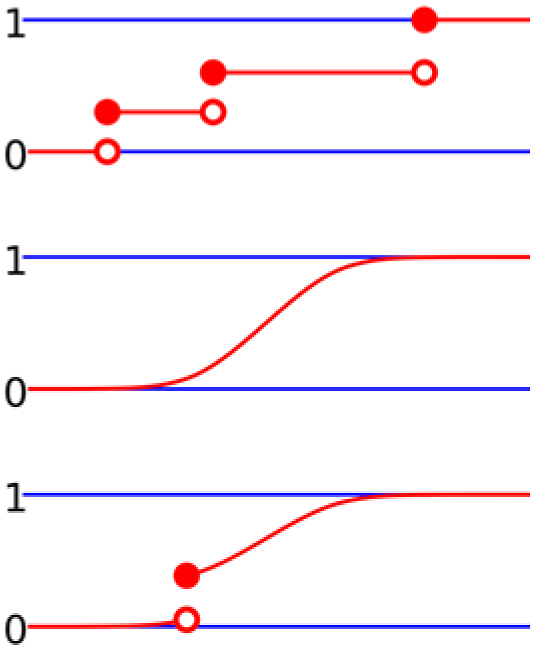
- Can still work with probabilities:

$$P(X = 3)$$

- Cumulative Distribution Func. (CDF)

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

- Density / mass function $p_X(x)$



Wikipedia CDF

Basics: **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"

Basics: **Joint Distributions**

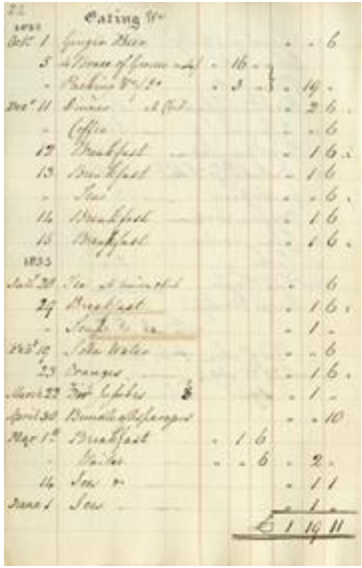
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty that correlate with each other

Basics: **Marginal** Probability

- Given a joint distribution $P(X = a, Y = b)$
  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the "marginal" distribution.



Probability Tables

- Write our distributions as tables

- # of entries? 4.
  - If we have n variables with k values, we get k^n entries
  - **Big!** For a 1080p screen, 12 bit color, size of table: $10^{7490589}$
  - No way of writing down all terms

Independence

- Independence between RVs:

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

- Why useful? Go from k^n entries in a table to $\sim kn$
- Expresses joint as **product** of marginals

- requires domain knowledge

Conditional Probability

- For when we know something (i.e. Y=b),

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

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$P(cold|white) = \frac{P(cold, white)}{P(white)} = \frac{120}{45 + 120} = 0.73$$

Conditional independence

– require domain knowledge

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

Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \ldots, A_n)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1)\ldots P(A_n|A_{n-1}, \ldots, A_1)$$

- Note: still big!
  - If some **conditional independence**, can factor!
  - Leads to **probabilistic graphical models**

Bayes' Rule

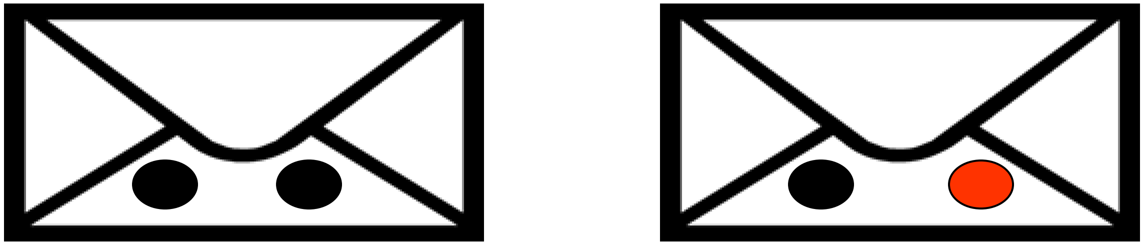
- Apply the conditional probability definition twice:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Note: fundamental rule in statistical learning
  - Leads to **Bayesian Inference**

Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth \$100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**



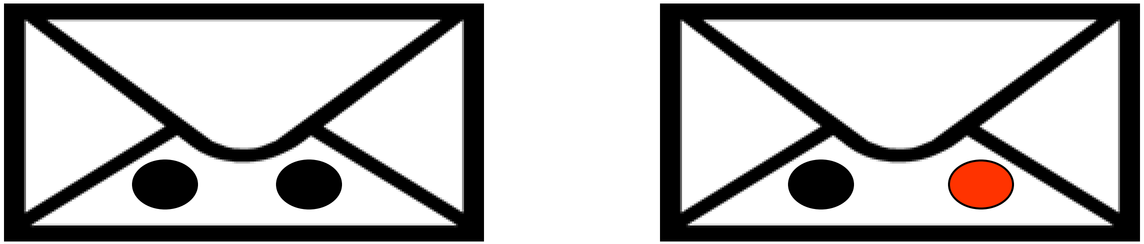
Two Envelopes Solution

- Let's solve it. $$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- Now plug in: $$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!



Flu Diagnosis Problem

- Evaluating probabilities:
  - Wake up with a sore throat
  - Do I have the flu?
- Logic approach: $S \to F$? Too strong
- **Inference**: estimate probability given evidence $P(F|S)$ (records of 1000 persons)
  - Sore throat: 100 persons; With Flu: 10; Sore throat among flu sufferers: 9

Flu Diagnosis Problem

- Want: $P(F|S)$
- **Bayes' Rule:** $P(F|S) = \frac{P(F,S)}{P(S)} = \frac{P(S|F)P(F)}{P(S)}$
- Estimate parts via data:
  - $P(S) = 0.1$ Sore throat rate
  - $P(F) = 0.01$ Flu rate
  - $P(S|F) = 0.9$ Sore throat rate among flu sufferers

So $P(F|S) = 0.09$

Reasoning With Conditional Distributions Using Bayes' Rule

- Interpretation $P(F|S) = 0.09$
  - Much higher chance of flu than normal rate (0.01).
  - Very different from $P(S|F) = 0.9$
    - 90% of folks with flu have a sore throat
    - But, only 9% of folks with a sore throat have flu

- Idea: **update** probabilities from

evidence

Bayesian **Inference**

- Fancy name for what we just did. Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- *H* is the hypothesis
- *E* is the evidence

Bayesian **Inference**

- Terminology:

$$P(H|E) = \frac{P(E|H)\color{red}{P(H)}}{P(E)} \longleftarrow \textbf{Prior}$$

- Prior: estimate of the probability **without** evidence

Bayesian Inference

- Terminology:

Likelihood

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Likelihood: probability of evidence **given a hypothesis**

Bayesian Inference

- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Posterior

- Posterior: probability of hypothesis **given evidence**.

Review: Bayesian Inference

- Conditional Probability & Bayes Rule:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Evidence *E*: what we can observe
- Hypothesis *H*: what we'd like to infer from evidence
  - Need to plug in prior, likelihood, etc.
- Usually do not know these probabilities. How to estimate?

Samples and Estimation

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

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

Samples and Estimation

- Estimate probability $P(H)$, $P(E)$, $P(E|H)$
- Estimate the mean $E[X]$
- Estimate parameters $P_\theta(X)$

- Example: Bernoulli with parameter *p* *(i.e., a weighted coin flip)*
  - $P(X = 1) = p, P(X = 0) = 1 - p$
  - Mean $E[X]$ is *p*

Examples: Sample Mean

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

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

  - That is, counting heads

Estimating Multinomial Parameters

- k-sized die (special case: k=2 coin)
- Face i has probability p_i, for $i=1 \ldots k$
- In n rolls, we observe face i showing up n_i times

$$\sum_{i=1}^{k} n_i = n$$

- Estimate $(p_{1,\ldots,} p_k)$ from this data $(n_{1,\ldots,} n_k)$

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p_1}, \ldots, \widehat{p_k})$

$$\widehat{p_i} = \frac{n_i}{n}$$

- Estimate using frequencies

Regularized Estimate

- Hyperparameter $\epsilon > 0$

$$\widehat{p_i} = \frac{n_i + \epsilon}{n + k\epsilon}$$

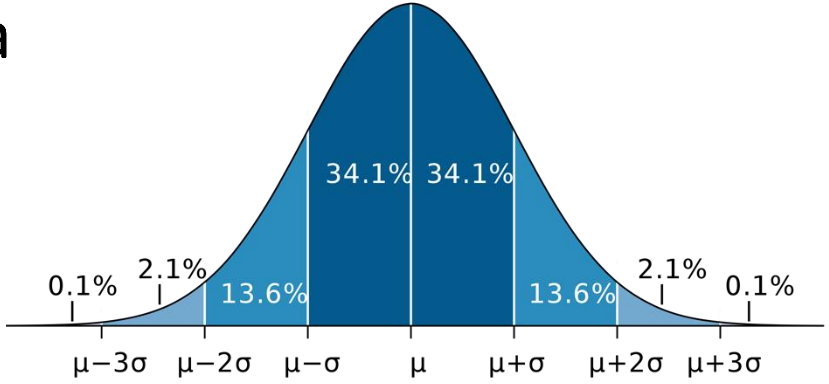
- Avoids zero when *n* is small
- Biased, but has smaller variance
- Equivalent to a specific Maximum A Posterori (MAP) estimate, or smoothing

Estimating 1D Gaussian Parameters

- Gaussian (aka Normal) distribution $N(\mu, \sigma^2)$
  - True mean μ, true variance σ^2
- Observe *n* data points from this distribution

$$x_1, \ldots, x_n$$

- Estimate μ, σ^2 from this data



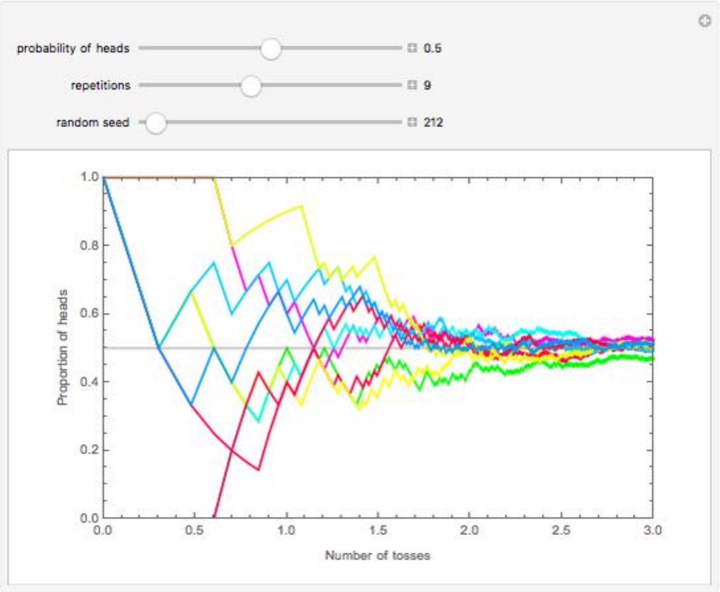
Wikipedia: Normal distribution

Estimating 1D Gaussian Parameters

- Mean estimate $\hat{\mu} = \dfrac{x_1 + \cdots + x_n}{n}$
- Variance estimates
  - Unbiased $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$
  - MLE $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$

Estimation Theory

- Is the sample mean a good estimate of the true mean?
  - Law of large numbers
  - Central limit theorems
  - Concentration



Wolfram Demo

Estimation Errors

- With finite samples, likely error in the estimate
- Mean squared error
  - $\mathrm{MSE}\left[\hat{\theta}\right] = \mathbb{E}\,[\left(\hat{\theta} - \theta\right)^2]$
- Bias / Variance Decomposition
  - $\mathrm{MSE}\left[\hat{\theta}\right] = \mathbb{E}\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right] + \left(\mathbb{E}\left[\hat{\theta}\right]\ - \theta\right)^2$

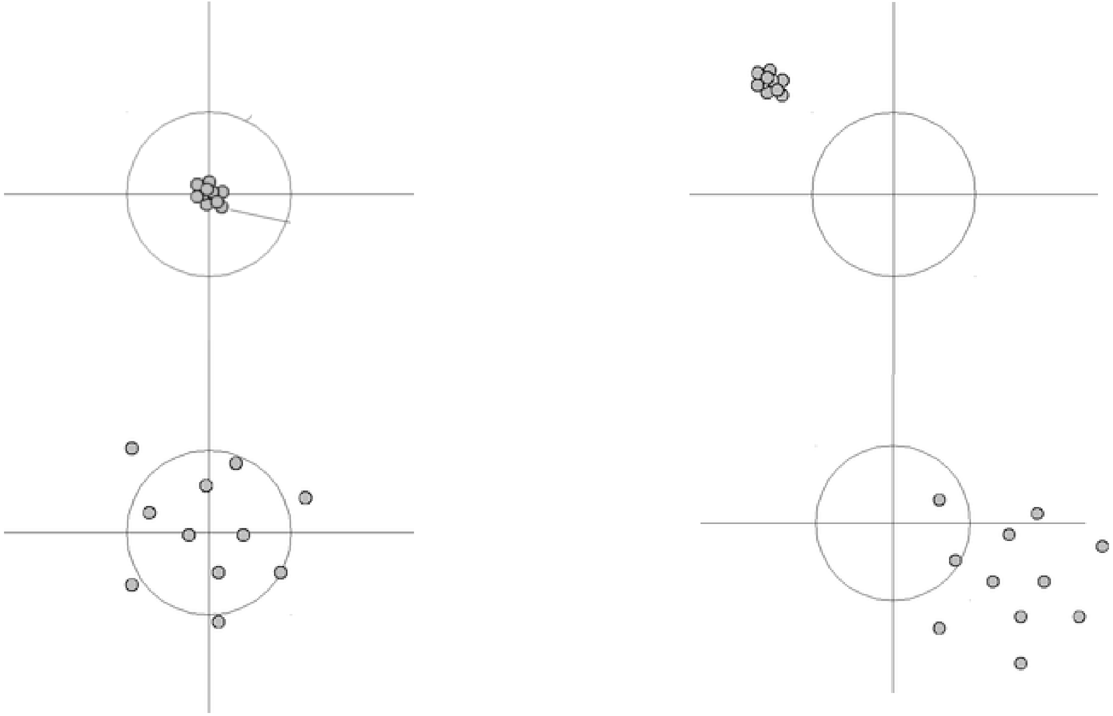
Variance Bias

Bias / Variance

Low Bias | High Bias

Low Variance

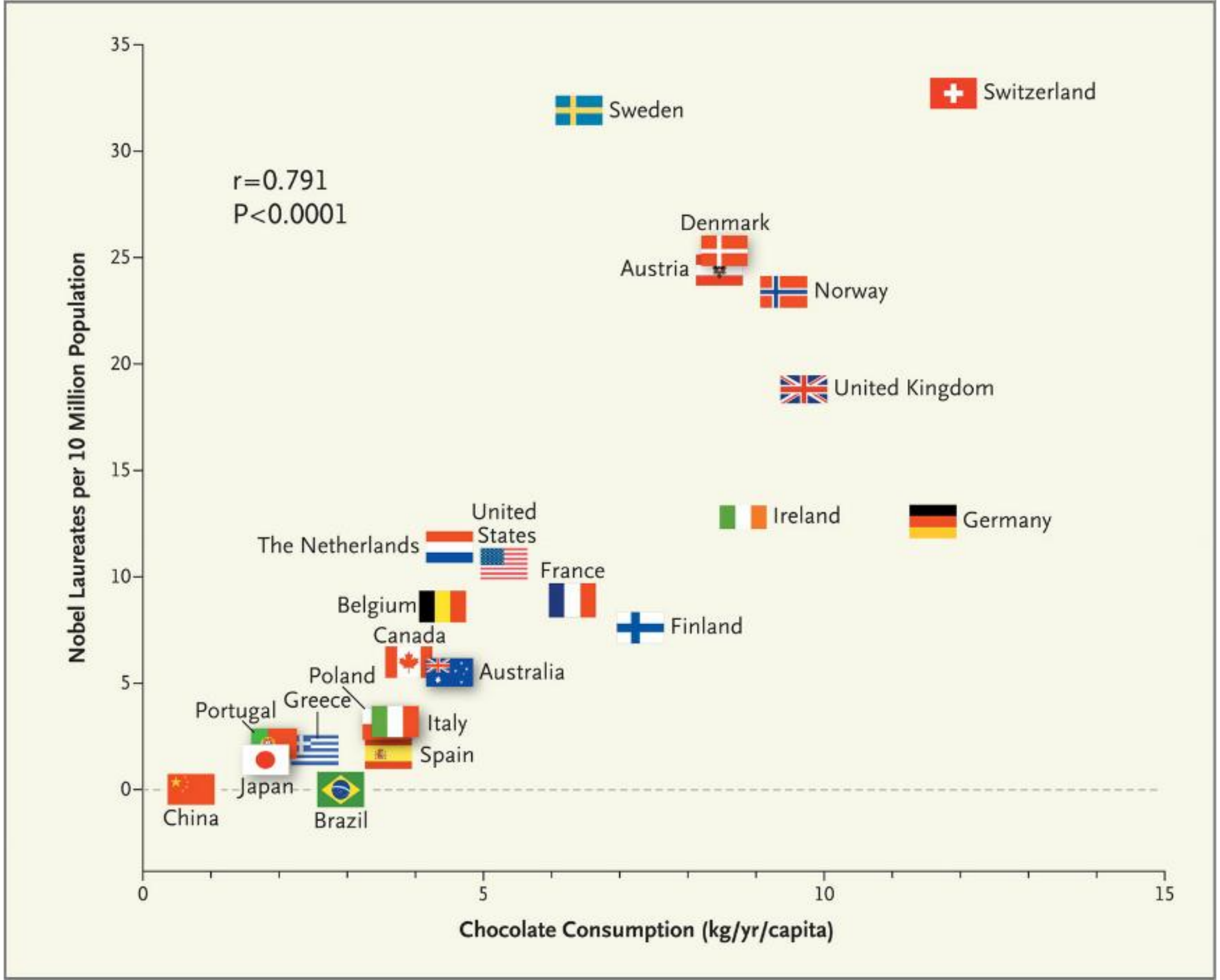
High Variance



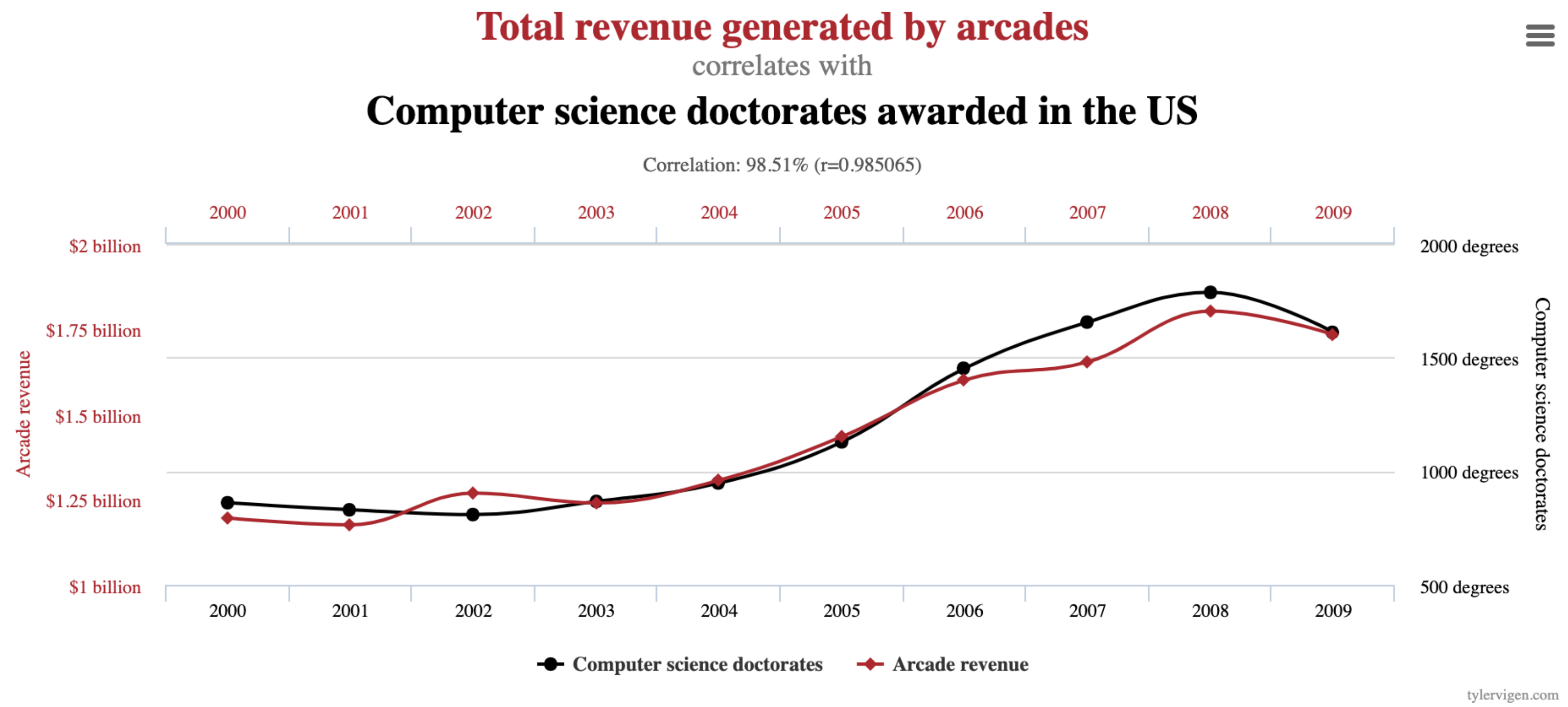
Wikipedia: Bias-variance tradeoff

Correlation vs. Causation

- Conditional probabilities only define correlation (aka association)
- $P(Y|X)$ “large” does not mean X causes Y
- Example: X=yellow finger, Y=lung cancer
- Common cause: smoking



https://www.nejm.org/doi/full/10.1056/NEJMon1211064



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