PURDUE
UNIVERSITY

CS37300:
Data Mining and Machine Learning

Linear Algebra + Hypothesis Testing

Profs. Tianyi Zhang and Rajiv Khanna

1 Sep 2023

Indiana Center for Database Systems ™

PURDUE UNIVERSITY

Likelihood function

- Assume we have *n* **independent** samples $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$
- Define the dataset $D = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$
- The likelihood function represents the probability of the dataset *D* as a function of the model parameters

$$L(D;\theta) = P(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n;\theta) = \prod_{i=1}^{n} P(\underline{x}_i;\theta)$$

by independence

Likelihood function

- The likelihood function represents the probability of the dataset *D* as a function of the model parameters

- Gives **relative probability of data given a parameter**
- We can compare two values θ and θ' by comparing their likelihoods
- We say that θ is better for explaining the dataset *D* than θ' if

$$L(D;\theta) > L(D;\theta')$$

Maximum likelihood estimation (MLE)

- Most widely used method of parameter estimation
- **Intuition:** a θ with higher likelihood explains better the data
- "Learn" the best parameters θ that maximizes likelihood:

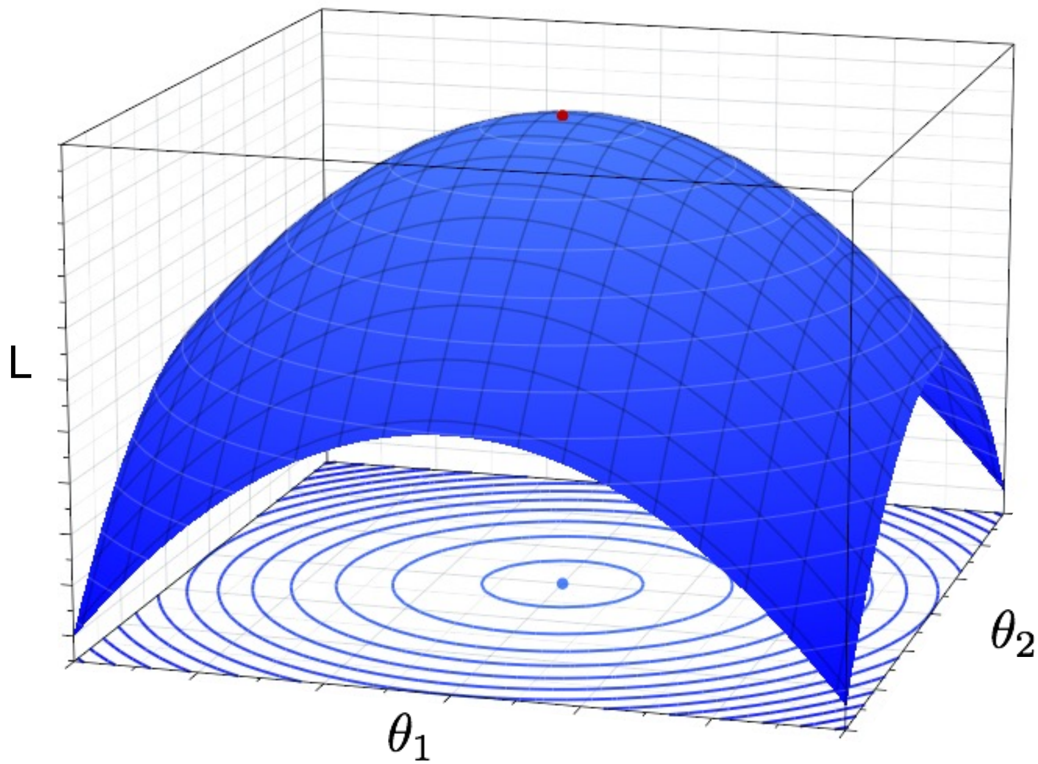
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; L(D;\theta)$$

- Often easier to work with log-likelihood:

$$l(D;\theta) = \log L(D;\theta) = \log \prod_{i=1}^{n} P(\underline{x}_i;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; l(D;\theta)$$

Likelihood surface



If the log-likelihood surface is concave, we can often determine the parameters that maximize the function analytically

PURDUE UNIVERSITY

Maximum Likelihood Estimation (MLE) for Bernoulli

- For a Bernoulli r.v. $x_i \in \{0,1\}$ $\theta = p$, $P(x_i;\theta) = p^{x_i}(1-p)^{1-x_i}$

$$\log P(x_i;\theta) = x_i \log p + (1-x_i)\log(1-p)$$

- Clearly:
- The **log-likelihood function** is:

$$l(D;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$
$$= \sum_{i=1}^{n} \left(x_i \log p + (1-x_i)\log(1-p) \right)$$
$$= \left(\sum_{i=1}^{n} x_i \right) \log p + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p)$$

- Recall that the **MLE** is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(D;\theta)$$

Maximum Likelihood Estimation (MLE) for Bernoulli

- We can maximize $l(D;\theta)$ by taking derivative equal to zero:

$$\frac{\partial l(D;\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \quad \text{then} \quad \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Linear algebra review

Vectors

- A vector is a matrix with several rows and one column

$$a = \begin{bmatrix} 5 \\ 7 \\ 1 \\ 4 \end{bmatrix} = (5,7,1,4)^T$$

- Notation: $a \in \mathbb{R}^m$

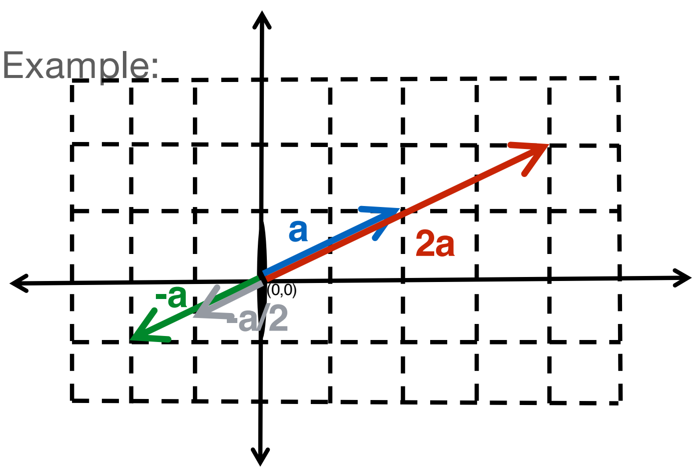
Vector: multiplication by scalar

- A scalar c is a real value
- Multiply/divide all entries of vector a by the scalar c

$$(ca)_i = ca_i$$

$$(a / c)_i = a_i / c$$

- Example:



$$a = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad 2a = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$$-a = \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \quad -a/2 = \begin{bmatrix} -1 \\ -0.5 \end{bmatrix}$$

Vector: addition and subtraction

- a and b have the same number of rows

$$a = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 7 \\ 3 \end{bmatrix}$$

$$(a+b)_i = a_i + b_i$$

- Add corresponding entries in a and b

$$a + b = \begin{bmatrix} 4 \\ 9 \\ 7 \end{bmatrix}$$

$$(a-b)_i = a_i - b_i$$

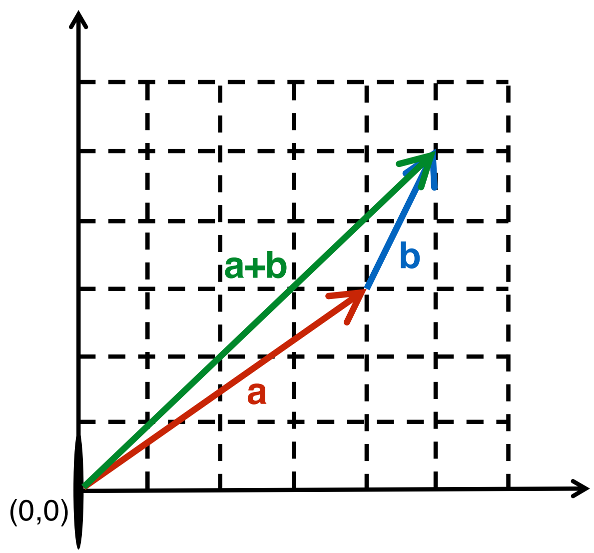
- Subtract corresponding entries in a and b

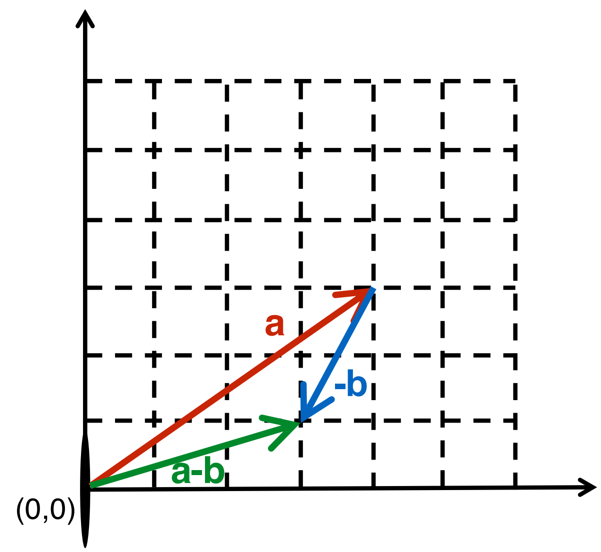
$$a - b = \begin{bmatrix} 2 \\ -5 \\ 1 \end{bmatrix}$$

Vector: addition and subtraction

- Geometrically…

$$a = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad a + b = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad a - b = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$





Vector: inner product

- Defined as

$$a \bullet b = a^T b = \sum_{k=1}^{m} a_k b_k \quad a \in \mathbb{R}^m \quad b \in \mathbb{R}^m$$

- Analog of scalar multiplication in many ways
- a and b have the same number of rows: ,

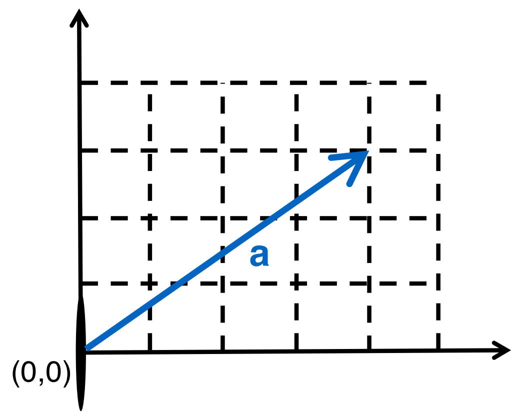
$$a = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -7 \\ 3 \end{bmatrix}$$

$$a \bullet b = 3 \times 1 + 2 \times (-7) + 4 \times 3 = 1$$

Vector: Euclidean norm

- The norm of $a \in \mathbb{R}^m$ is $\|a\| = \sqrt{a \bullet a} = \sqrt{a_1^2 + a_2^2 + \ldots + a_m^2}$

- Example



$$a = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad \|a\| = \sqrt{3^2 + 4^2} = 5$$

- Distance between two vectors a and b is $\|a - b\|$

Matrices

- A matrix is a 2-D array of numbers:

Row

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}.$$

Column

- Example notation for type and shape:

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}$$

Matrix: addition and subtraction

- A and B have the same number of rows and columns

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 0 \\ 0 & 4 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 5 & 1 & 0 \\ 5 & 7 & 2 \\ -5 & 3 & 1 \end{bmatrix}$$

$$(A+B)_{i,j} = A_{i,j} + B_{i,j}$$

- Add corresponding entries in A and B

$$A + B = \begin{bmatrix} 7 & 4 & 1 \\ 6 & 9 & 2 \\ -5 & 7 & 6 \end{bmatrix}$$

2+5

5+1

$$(A-B)_{i,j} = A_{i,j} - B_{i,j}$$

- Subtract corresponding entries in A and B

$$A - B = \begin{bmatrix} -3 & 2 & 1 \\ -4 & -5 & -2 \\ 5 & 1 & 4 \end{bmatrix}$$

2-5

5-1

Matrix: multiplication

- Number of columns of A = number of rows of B

$$(AB)_{i,j} = \sum_k A_{i,k} B_{k,j}$$

- Example:

$$A = \begin{bmatrix} 3 & 1 & -2 & 4 \\ -2 & 4 & 2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 2 & 1 \\ 4 & 5 & -3 \\ 2 & 3 & 2 \\ -1 & 2 & -4 \end{bmatrix}$$

$$3 \times 2 + 1 \times 5 - 2 \times 3 + 4 \times 2 = 13$$

$$AB = \begin{bmatrix} 5 & 13 & -20 \\ 14 & 22 & -10 \end{bmatrix}$$

Matrix: multiplication by scalar

- A scalar c is a real value
- Multiply/divide all entries of matrix A by the scalar c

$$(cA)_{i,j} = cA_{i,j}$$

$$(A/c)_{i,j} = A_{i,j}/c$$

- Example:

$$A = \begin{bmatrix} 4 & 5 \\ 0 & -2 \\ 3 & 6 \end{bmatrix}, \quad 3A = \begin{bmatrix} 12 & 15 \\ 0 & -6 \\ 9 & 18 \end{bmatrix}, \quad A/2 = \begin{bmatrix} 2 & 2.5 \\ 0 & -1 \\ 1.5 & 3 \end{bmatrix}$$

Matrix: transpose

- Rows become columns, columns become rows

$$(A^T)_{i,j} = A_{j,i}$$

- Example:

$$A = \begin{bmatrix} 3 & 1 & -2 & 4 \\ -2 & 4 & 2 & 0 \end{bmatrix}, \quad A^T = \begin{bmatrix} 3 & -2 \\ 1 & 4 \\ -2 & 2 \\ 4 & 0 \end{bmatrix}$$

- Multiplication property: $(AB)^T = B^T A^T$

- If $A = A^T$ then A is called **symmetric**

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 3 & -2 & 0 \\ 5 & 0 & 4 \end{bmatrix}$$

Identity matrix and Inverse

- **Identity** matrix has 1s in the diagonals and 0s everywhere else

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- For any vector x, we have $Ix = x$
- Matrix **inverse**: $A^{-1}A = I$
- A matrix cannot be inverted if:
  - More rows than columns, or more cols than rows
  - Redundant rows/columns (linear dependence)

Identity matrix and Inverse

- Example

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 4 & 1 \\ -2 & 1 & 7 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} -27 & 19 & 5 \\ 16 & -11 & -3 \\ -10 & 7 & 2 \end{bmatrix}$$

- Several languages provide functions/methods for computing the inverse
  (*We will not go into these details.*)

Functions and gradients

- We can define a function $f(x)$ of a vector $x \in \mathbb{R}^m$
- The **gradient** has the derivatives with respect to each entry:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \\ \partial f / \partial x_m \end{bmatrix} \in \mathbb{R}^m$$

- Example:
  $f(x) = 5e^{x_2} + x_3 e^{x_1},$

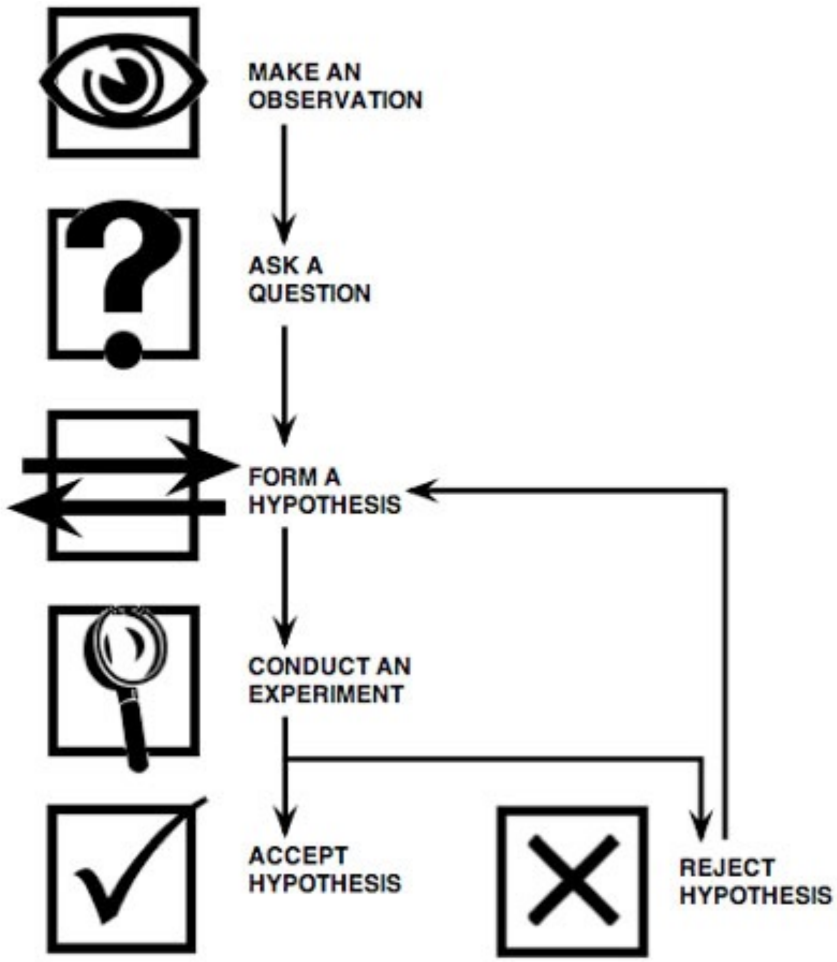
$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \partial f / \partial x_3 \end{bmatrix} = \begin{bmatrix} x_3 e^{x_1} \\ 5e^{x_2} \\ e^{x_1} \end{bmatrix}$$

Common gradients

- $\frac{\partial x^\top b}{\partial x} = b$
- $\frac{\partial x^\top A x}{\partial x} = (A + A^\top)x$. If A is symmetric ?
- Standard rules:
  - Differentiating under a summation
  - Chain rule: $\frac{\partial f(x^\top b)}{\partial x} = \frac{\partial f(z)}{\partial z} b$
- Check dimensions!

Hypothesis testing

Hypothesis Testing



PURDUE
UNIVERSITY

What is a hypothesis?

- **Hypotheses** are tentative statements of the expected relationships between two or more variables
  - **Inductive** hypotheses are formed through inductively reasoning from many specific observations to tentative explanations (*bottom-up*)
  - **Deductive** hypotheses are formed through deductively reasoning implications of theory (*top-down*)
- Reasons for using hypotheses
  - Provides a useful framework for organizing and summarizing results and conclusions
  - Provides focus and directs research investigation

Types of hypotheses

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe the relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

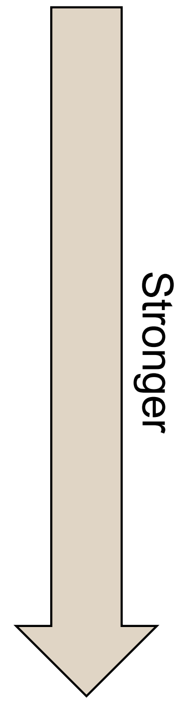
- **Non-directional**: an differential outcome is anticipated but the specific nature of it is not known (e.g., the tuning parameter will affect algorithm performance)
- **Directional**: a specific outcome is anticipated (e.g., the use of pruning will increase accuracy of models compared to no pruning)

Descriptive Hypothesis

Non-Directional Relational Hypothesis

Directional Relational Hypothesis

Directional Causal Hypothesis



From claims to testable hypotheses

- Israel COVID cases breakdown Aug 17, 2021

Age	Total	Vax %		Severe Cases		Score function
Conditional	Individuals (approx.)	Not Vax	Fully Vax	Not Vax per 100k	Fully Vax per 100k	Conditional Efficacy
[12,15]	650,000	62.1%	29.9%	0.3	0.0	100.0%
[16,19]	600,000	21.9%	73.5%	1.6	0.0	100.0%
[20,29]	1,200,000	20.5%	76.2%	1.5	0.0	100.0%
[30,39]	1,050,000	16.2%	80.9%	6.2	0.2	96.8%
[40,49]	900,000	13.2%	84.4%	16.5	1.0	94.2%
[50,59]	750,000	10.0%	88.0%	40.2	2.9	93.2%
[60,69]	550,000	8.8%	89.8%	76.6	8.7	89.8%
[70,79]	350,000	4.2%	94.6%	190.1	19.8	90.6%
[80,89]	120,000	5.6%	92.6%	252.3	47.9	84.0%
90+	50,000	6.1%	90.5%	510.9	38.6	93.0%

From claims to testable hypotheses

Claim: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals

Building the hypothesis:

- **Step 1:** Express data as random variables (joinly). E.g.:
  - A age
  - SV severe vax per 100k
  - SU severe unvax per 100k
  - $\mathrm{Y} = \mathrm{SU}/(\mathrm{SU} + SV)$ observed vaccine efficacy

From claims to testable hypotheses

Claim: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals

Building the hypothesis:

- **Step 2:** Restate claim as a hypothesis about the relationship between the random variables, e.g.,
  - Hypothesis: $E[Y|A > 90] > E[Y|80 \leq A \leq 89]$
- **Step 3:** Determine type of hypothesis (and consider whether you can make it stronger)

From claims to testable hypotheses

- **Claim**: Vaccine efficacy is higher for 90+ y/o individuals than for [80,89] y/o individuals
- **Types of hypotheses**:
  - *Descriptive*: Efficacy values vary (i.e., Y varies).
  - *Non-directional relational*: Y varies based on age (i.e., A and Y are associated)
  - *Directional-relational*: A > 90 folks have higher efficacy (i.e., A > 90 is associated with smaller Y)
  - *Causal-relational*: A > 90 folks have higher vaccine efficacy because these are monitored more closely in nursing homes and get medical interventions earlier before they get too sick

Using Data to Test Hypotheses

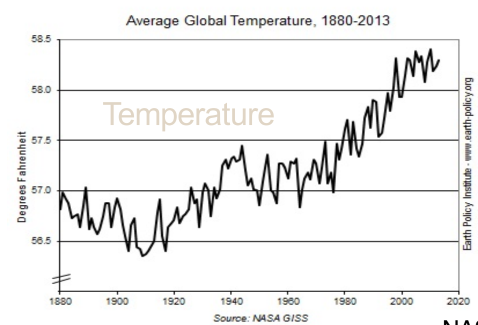
Is Aspirin effective in reducing cancer risk?

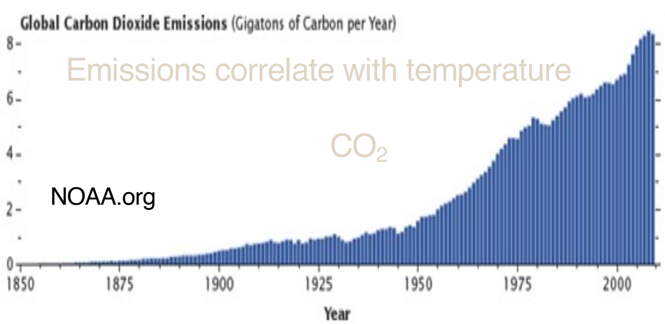
- Here, we are looking at causal effects…
- Data
  - A person represented by random variable $X \in$ {{Age}, {Sick, Not Sick}, …}
    - Recruit people: x_{john} , x_{mary} , x_{eve} , x_{adam} , …
    - Medicine to take: $T \in$ {aspirin, placebo}
- Hypothesis
  - Force ½ (randomly chosen) of the people to take aspirin : $Y|X, \mathbf{do}(T = \text{aspirin}) \in \{1 \text{ - Cancer in 1yr}, 0 \text{ - No Cancer in 1yr}\}$
    - Here, the **do**() notation means forcing T to be something (intervention)
  - Force remaining ½ to NOT take aspirin: $Y|X, do(T=\text{placebo})$
    - Hypothesis: $E[Y|X, do(T=\text{aspirin})] < E[Y|X, do(T=\text{placebo})]$

Directional
Causal Hypothesis

Example:

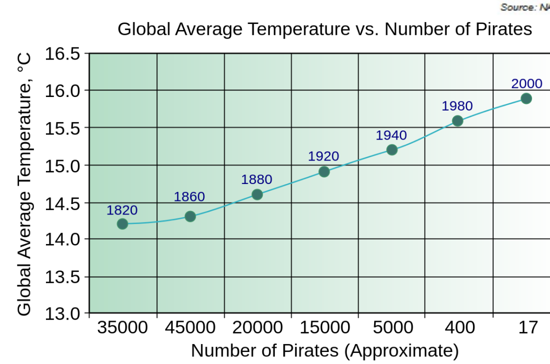
- CLAIM 1: The temperature of the planet is rising and the increase is **due** to human activities such as fossil fuel use and deforestation.
- Which kind of data could support such claim?





Not enough

Why?



Directional Relational Hypothesis

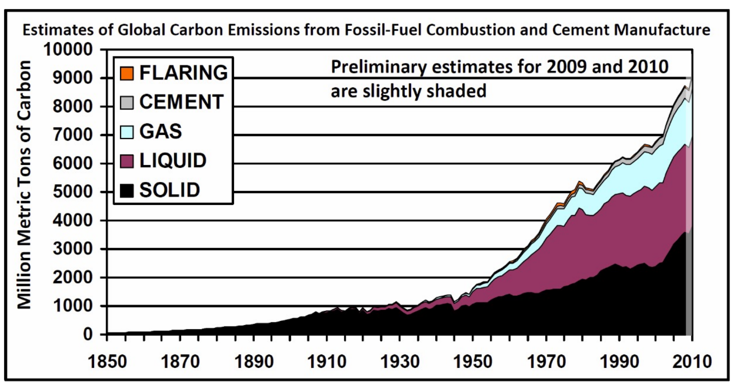
- CLAIM 2: The temperature of the planet is rising with increased human activity

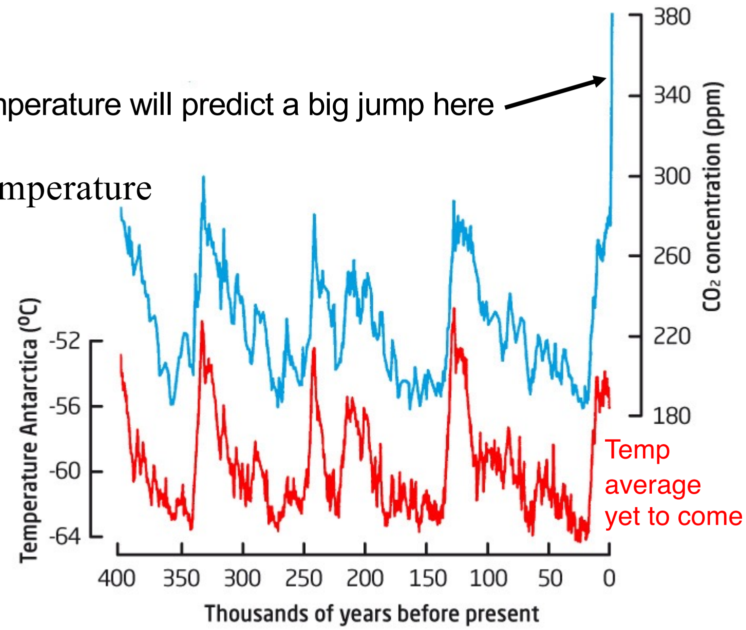
Model trained on CO2 data to predict Antartica's temperature will predict a big jump here

$f(x_{CO2})$ = Temperature

CO$_2$ jump related to human activity

$g(x_{fossil\text{-}fuel})$ = Atmospheric CO_2 – Natural CO_2





Directional Relational Hypothesis

Causal Claims without Experiments are Difficult

Curiosity

- CLAIM 1: The temperature of the planet is rising and the increase is **DUE** to human activities such as fossil fuel use.
  - How would you test it?
  - How it is tested:
    - Climate models (we know how climate works)
      - We know how much energy the sun outputs
      - We know how much energy the planet radiates back into space
      - We know where the energy goes inside the planet
      - https://earthobservatory.nasa.gov/features/EnergyBalance
    - Historic *natural experiment* events:
      "Coal-burning in Siberia after volcanic eruption led to climate change 250 million years ago"

Directional
Causal Hypothesis

PURDUE UNIVERSITY

Hypothesis Must Consider Observation Biases

Curiosity

- Your experience with buses at peak hours:
  - Bus at 99% capacity at peak hours
  - You wait on average 17 minutes and 9 seconds for it to arrive
- Purdue's transportation admin:
  - *buses at peak hour are at 60% capacity*
  - *average bus inter-arrival time is 10 minutes*

?