STUDENT FOOD SECURITY SURVEY

UC San Diego researchers are conducting a study to better understand food security among undergraduate students. The survey will take 10 minutes or less to complete and is voluntary.

Use this link or scan the QR code to take the survey by October 31st, 2024.

Questions? Please contact:
Dr. Richard Garfein, HWSPH
rgarfein@ucsd.edu
Elle Mari, CCH
emari@ucsd.edu

UC San Diego

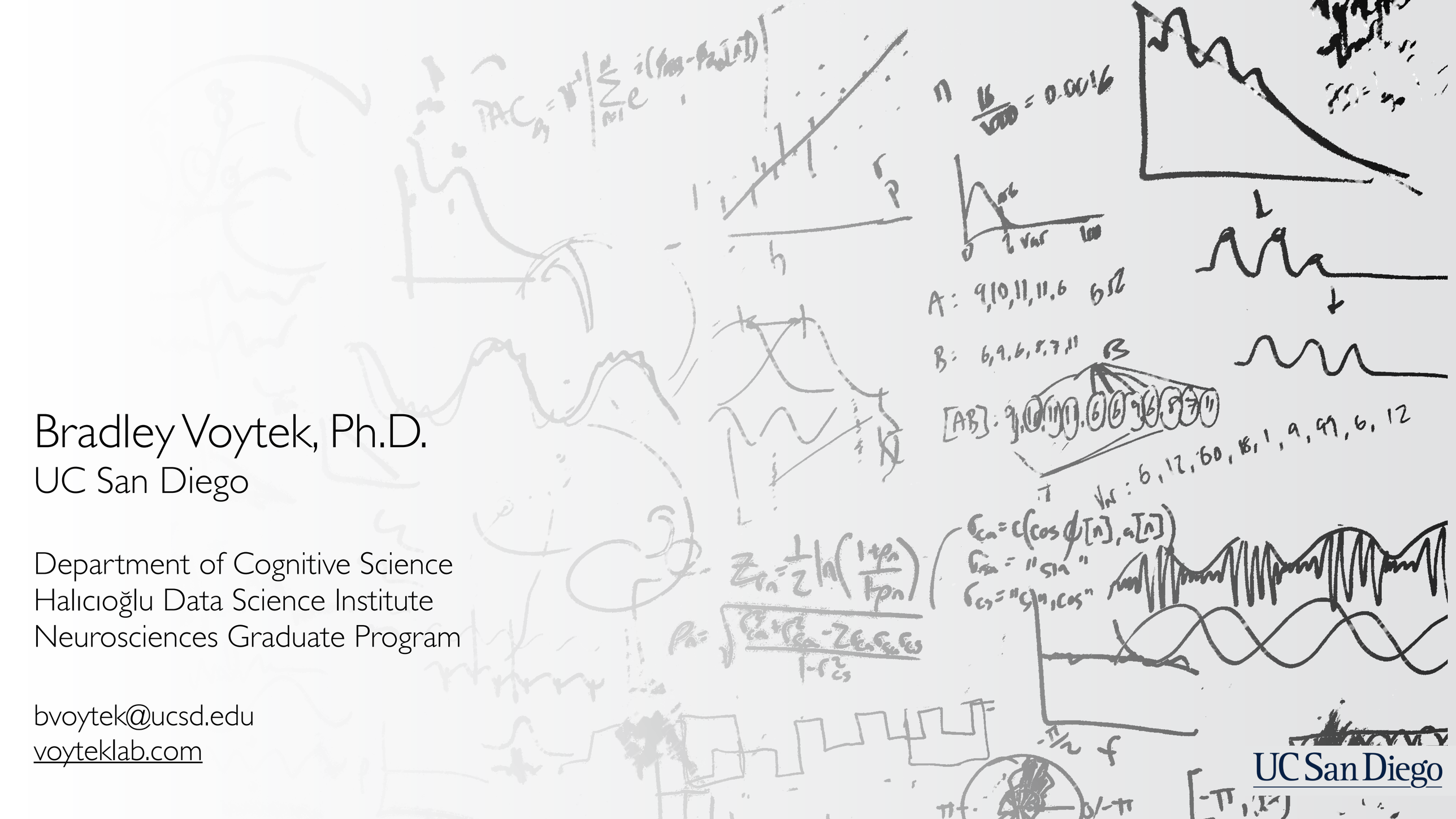
R3 is released!

Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

COGS 9
Introduction to Data Science

Statistical inference and ML

Statistical inference and ML

UC San Diego

Today's Learning Objective

What is machine learning, and what are its limitations?

Data Intuition

In today's pattern recognition class my professor talked about PCA, eigenvectors & eigenvalues.

I got the mathematics of it. If I'm asked to find eigenvalues etc. I'll do it correctly like a machine. But I didn't **understand** it. I didn't get the purpose of it. I didn't get the feel of it. I strongly believe in

> you do not really understand something unless you can explain it to your grandmother -- Albert Einstein

Well, I can't explain these concepts to a layman or grandma.

1. Why PCA, eigenvectors & eigenvalues? What was the *need* for these concepts?
2. How would you explain these to a layman?

Theory *vs* Practice

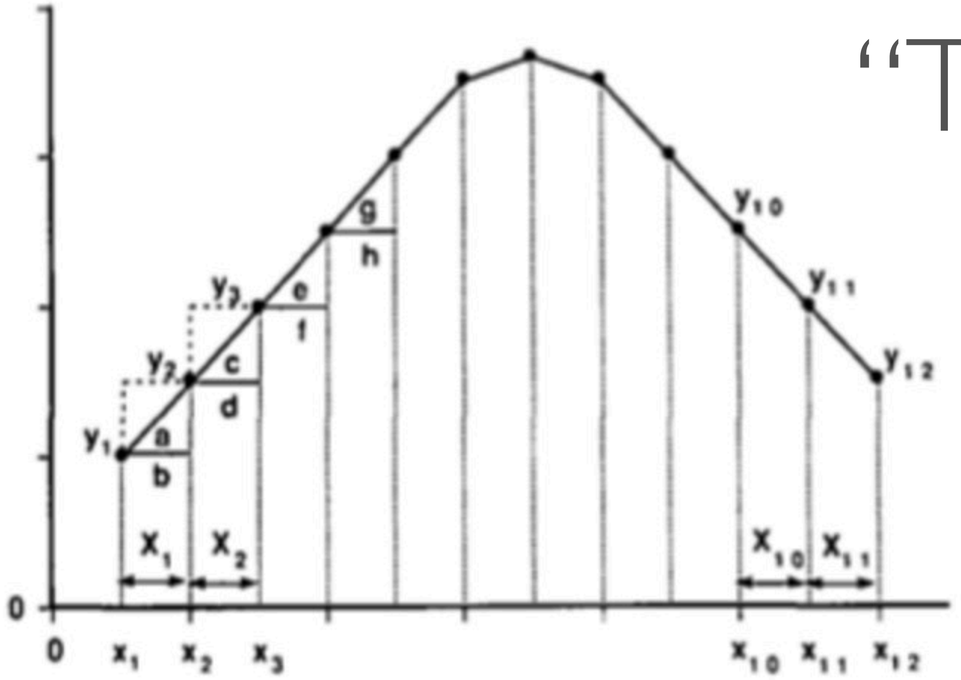


Figure 1—*Total area under the curve is the sum of individual areas of triangles a, c, e, and g and rectangles b, d, f, and h.*

"Tai's Model"

Integral

Theory *vs* Practice

"In Tai's Model, the total area under a curve is computed by dividing the area under the curve between two designated values on the X-axis (abscissas) into small segments (rectangles and triangles) whose areas can be accurately calculated from their respective geometrical formulas. The total sum of these individual areas thus represents the total area under the curve."

UC San Diego

Source: Tai, *Diabetes Care* 1994

What is machine learning?

ML is the science of getting computers to act without being explicitly programmed

"Machine learning is the science of getting computers to act without being explicitly programmed" - Andrew Ng, Stanford, ex-Google, formerly chief scientist at Baidu, founder Coursera

What does ML look like?

- Algorithms that can improve in performance with training or experience
- Typically by fitting large numbers of parameters
- Usually used in situations where it is difficult to define rules by hand
  - Face detection
  - Stock market prediction
  - Spam filters

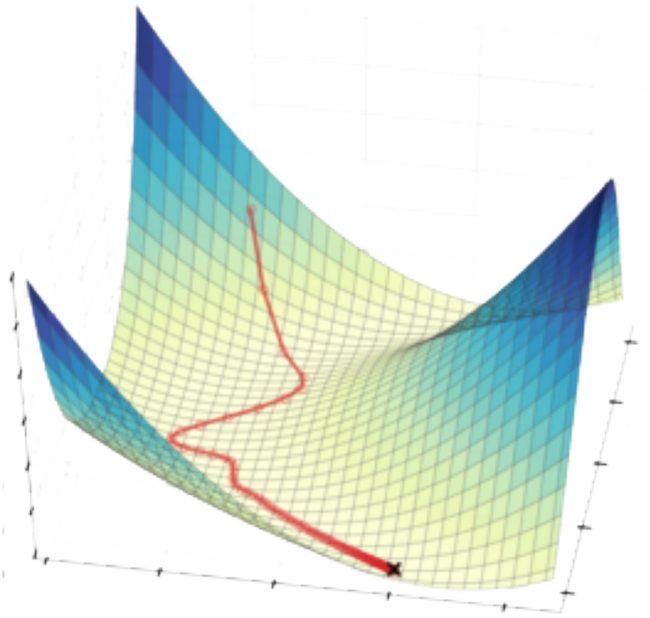
hard to define by hand

What is ML really?

What is ML really?

What is ML really?

Pick a set of parameter values that minimize some objective function

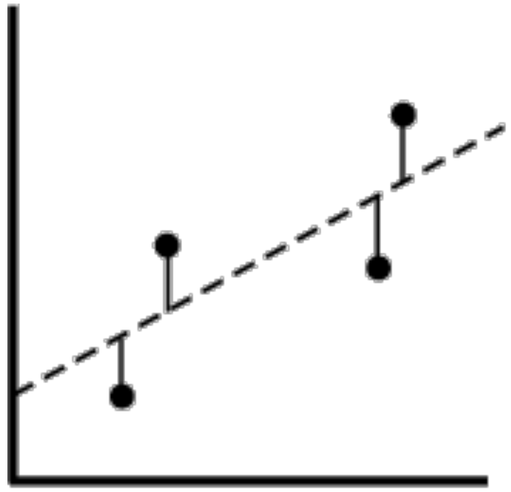


All machine learning is gradient descent

What is ML really?

Pick a set of parameter values that minimize some objective function

$$\text{argmin}_{a,b} \sum_{i=1:n} (y_i - (ax_i + b))^2$$



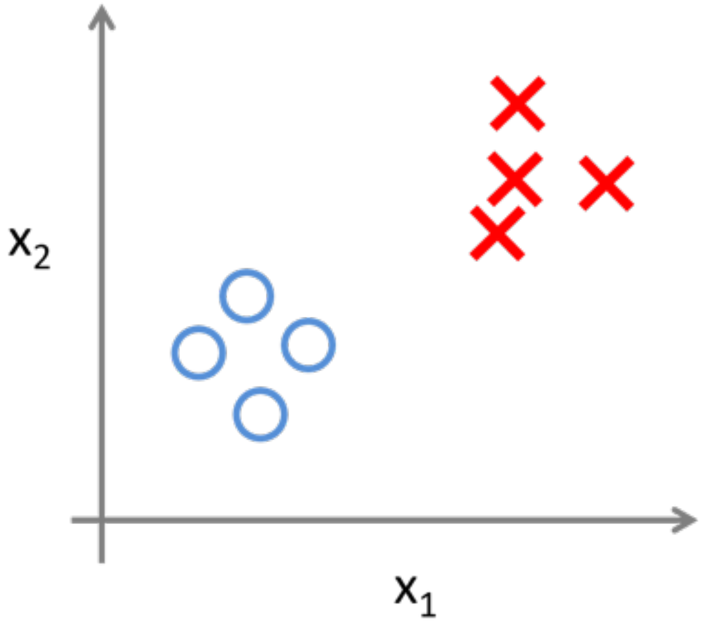
Two major modes of ML

Support vector machines use kernal trick

*Alot of machine learning is just fitting lines

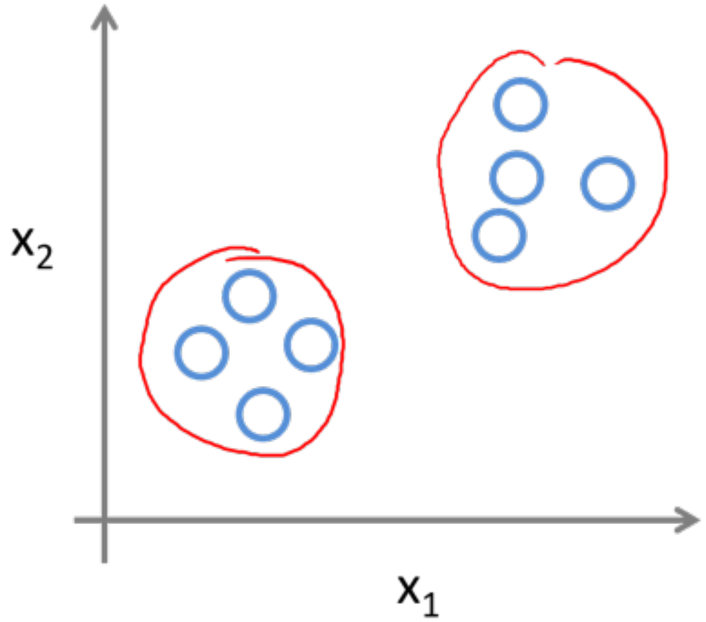
Supervised Learning

- regression



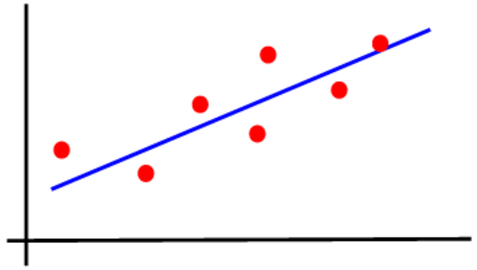
Unsupervised Learning

- you know know anything



Three canonical problems

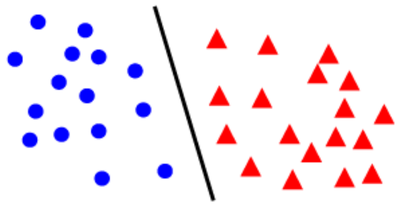
1. Regression - supervised
   - estimate parameters, e.g. of weight vs height

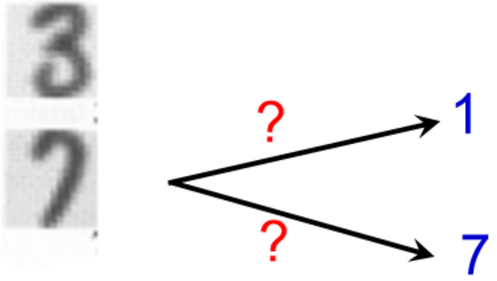


Three canonical problems

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height

2. Classification - supervised
   - estimate class, e.g. handwritten digit classification



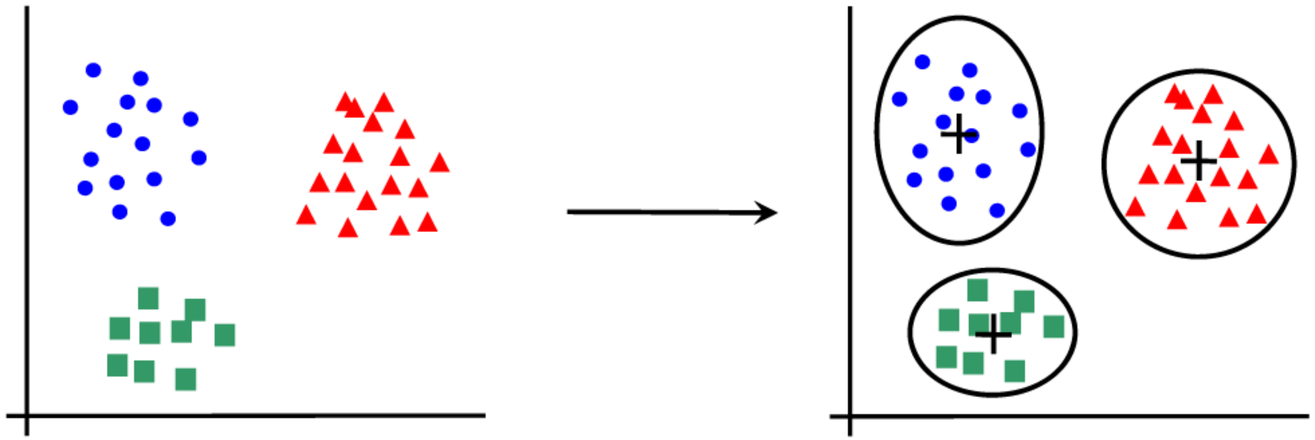


Three canonical problems

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height

2. Classification - supervised
   - estimate class, e.g. handwritten digit classification

3. Unsupervised learning – model the data
   - clustering



Regression is supervised

Classification is supervised

Clustering – unsupervised

Supervised classification demo!

Today's Learning Objective

What is Bayes' Theorem and why and how is it useful?

UC San Diego

Fermi Estimation

benfords law ? smaller numbers appear more frequently

Paint the Earth

Has humanity produced enough paint to cover the entire land area of the Earth?

—Josh (Bolton, MA)

Source: *XKCD*, What If?

Fermi Estimation

This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using a brush.

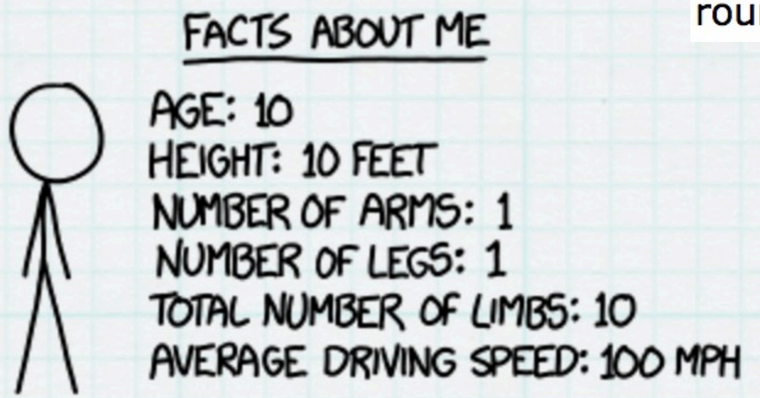
Source: *XKCD*, What If?

Fermi Estimation

Fermin Estimation The Thinking of the ways to Solve a problem

But first, let's think about different ways we might come up with a guess for what the answer will be. In this kind of thinking—often called **Fermi estimation**—all that matters is getting in the right ballpark; that is, the answer should have about the right number of digits. In Fermi estimation, you can round[1] all your answers to the nearest order of magnitude:

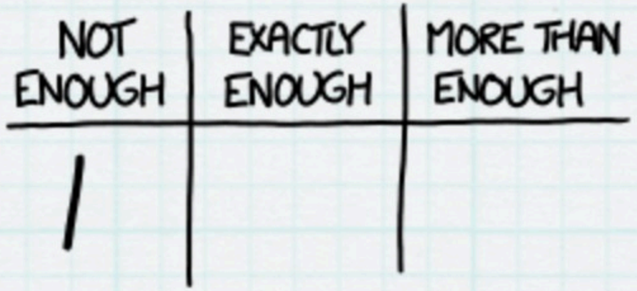
Using the formula $\text{Fermi}(x) = 10^{\text{round}(log_{10}x)}$, meaning that 3 rounds to 1 and 4 rounds to 10.



Source: *XKCD*, What If?

Fermi Estimation

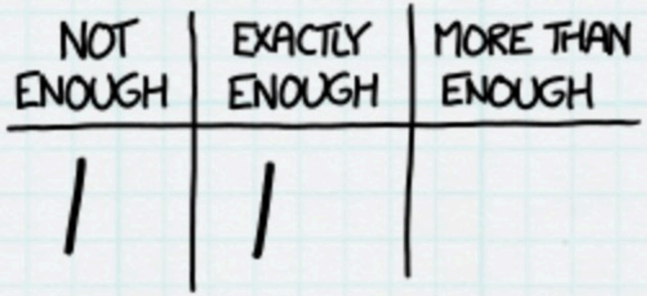
Let's suppose that, on average, everyone in the world is responsible for the existence of two rooms, and they're both painted. My living room has about 50 square meters of paintable area, and two of those would be 100 square meters. 7.15 billion people times 100 square meters per person is a little under a trillion square meters—an area smaller than Egypt.



Source: *XKCD*, What If?

Fermi Estimation

Let's make a wild guess that, on average, one person out of every thousand spends their working life painting things. If I assume it would take me three hours to paint the room I'm in, [2] and 100 billion people have ever lived, and each of them spent 30 years painting things for 8 hours a day, we come up with 150 trillion square meters ... just about exactly the land area of the Earth.



Source: *XKCD*, What If?

Fermi Estimation

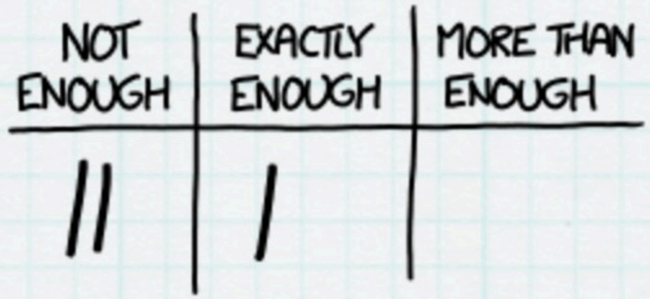
How much paint does it take to paint a house? I'm not enough of an adult to have any idea, so let's take another Fermi guess.

Based on my impressions from walking down the aisles, home improvement stores stock about as many light bulbs as cans of paint. A normal house might have about 20 light bulbs, so let's assume a house needs about 20 gallons of paint.[3] Sure, that sounds about right.

Source: *XKCD*, What If?

Fermi Estimation

The average US home costs about $200,000. Assuming each gallon of paint covers about 300 square feet, that's a square meter of paint per $300 of real estate. I vaguely remember that the world's real estate has a combined value of something like $100 trillion,[4] which suggests there's about 300 billion square meters of paint on the world's real estate. That's about one New Mexico.



Source: *XKCD*, What If?

Fermi Estimation

Of course, both of the building-related guesses could be overestimates (lots of buildings are not painted) or underestimates (lots of things that are not buildings[5] are painted) But from these wild Fermi estimates, my guess would be that there probably isn't enough paint to cover all the land.

So, how did Fermi do?

Source: *XKCD*, What If?

Fermi Estimation

According to the report **The State of the Global Coatings Industry**, the world produced 34 billion liters of paints and coatings in 2012.

There's a neat trick that can help us here. If some quantity—say, the world economy—has been growing for a while at an annual rate of **n**—say, 3% (0.03)—then the most recent year's share of the whole total so far is $1 - \frac{1}{1+n}$, and the whole total so far is the most recent year's amount times $1 + \frac{1}{n}$.

Source: *XKCD*, What If?

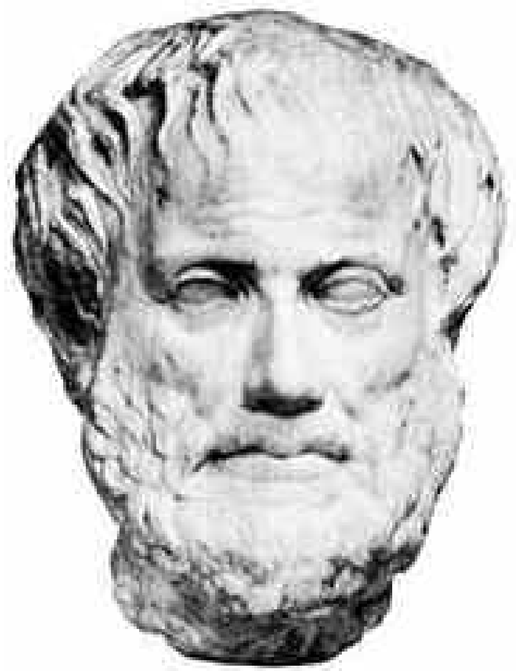
Fermi Estimation

If we assume paint production has, in recent decades, followed the economy and grown at about 3% per year, that means the total amount of paint produced equals the current yearly production times 34.[6] That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,[7] that's enough to cover 9 trillion square meters—about the area of the United States.

So the answer is no; there's not enough paint to cover the Earth's land, and—at this rate—probably won't be enough until the year 2100.

Source: *XKCD*, What If?

Logic



Aristotle
(384-322 BC)

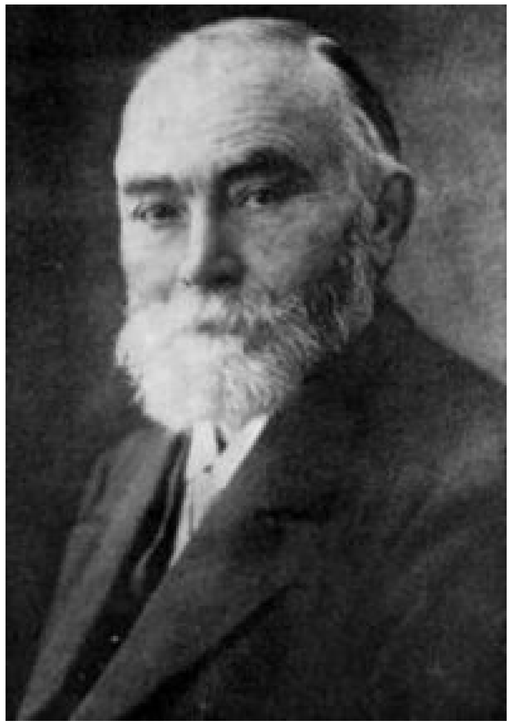
All *A*s are *B*s

All *B*s are *C*s

All *A*s are *C*s

Modern logic

George Boole
(1816-1854)



Gottlob Frege
(1848-1925)

$$\frac{P \rightarrow Q \quad P}{Q}$$

- A mathematically sound system of rules telling us how to infer the truth of one proposition from the truth of others.
- Given a set of facts, with symbols standing for things in the world, simple rules can lead us to new conclusions.

Inductive problems

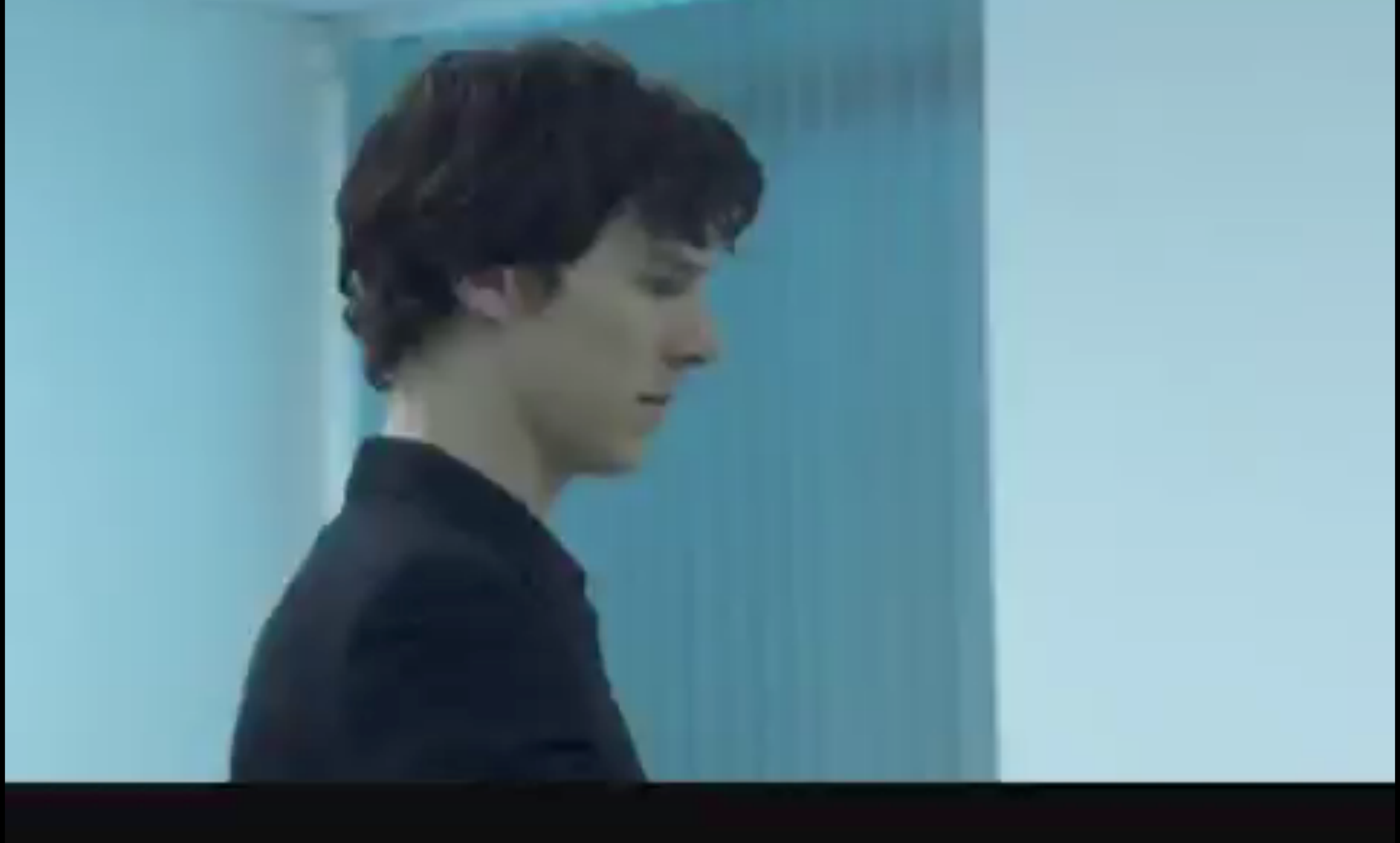
- Drawing conclusions that are not fully justified by the available data
  - (detective work)

unsupervised learning is inductive learning

- Much more challenging than deduction!

"In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment, and a very easy one, but people do not practice it much.

- Sir Arthur Conan Doyle



Bayes' Theorem

bayes' theorem prior knowledge with data

- Bayes' theorem tells us how to combine prior knowledge with data
  - a language for describing the constraints on human inductive inference

- Probabilistic approaches also tell us how to make decisions and interact with others

Bayes' Theorem

How rational agents should update their beliefs in the light of data

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Bayes' Theorem

- 1% of people have *fleeb* cancer (and therefore 99% do not).
- 80% of tests detect *fleeb* cancer when it is there (and therefore 20% miss it).
- 9.6% of tests detect *fleeb* cancer when it's not there (and therefore 90.4% correctly return a negative result).

Bayes' Theorem

	cancer (1%)	**no cancer (99%)**
test positive	80%	9.6%
test negative	20%	90.4%

- 1% of people have cancer

- If you *already have cancer*, you are in the first column.
  - There's an 80% chance you will test positive.
  - There's a 20% chance you will test negative.

- If you *don't have cancer*, you are in the second column.
  - There's a 9.6% chance you will test positive.
  - There's a 90.4% chance you will test negative.

Bayes' Theorem

$P(B|A) = \frac{P(B \cap A)}{P(A)}$

	cancer (1%)	no cancer (99%)
test positive	true positive 80%	false positive 9.6%
test negative	false negative 20%	true negative 90.4%

- Now suppose you get a positive test result.

- *What are the chances you have cancer?*
  - 80%?
  - 99%?
  - 10%?

Bayes' Theorem

	cancer (1%)	**no cancer (99%)**
test positive	true positive 80%	false positive 9.6%
test negative	false negative 20%	true negative 90.4%

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Bayes' Theorem

- Ok, we got a positive result. It means we're somewhere in the top row of our table. Let's not assume anything — it could be a true positive or a false positive.

p(cancer|positive) = (p(positive|cancer) * p(cancer)) / p(positive)

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

p(cancer|result) = (0.8 * 0.01) / p(result)

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

- p(result) tells us the chance of getting *any* positive result, whether it's a real positive in the cancer population (1%) or a false positive in the non-cancer population (99%). It's a bit like a weighted average, and helps us compare against the overall chance of a positive result.

p(cancer|result) = (0.8 * 0.01) / p(result)

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

- p(result): The chance of getting any type of positive result is the chance of a true positive plus the chance of a false positive.

p(cancer|result) = (0.8 * 0.01) / p(result)

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

- The chances of a false positive = chance you don't have *fleeb* cancer * chance test caught it anyway = 99% * 9.6% = 0.09504

p(cancer|result) = (0.8 * 0.01) / p(result)

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

- p(result) = true positive (0.008) plus false positive (0.09504) = 0.10304

p(cancer|result) = (0.8 * 0.01) / 0.10304

Bayes' Theorem

- The chances of a true positive = chance you have *fleeb* cancer * chance test caught it = 1% * 80% = 0.008

- p(result) = true positive (0.008) plus false positive (0.09504) = 0.10304

p(cancer|result) = 0.078 = 7.8%

Bayes' Theorem

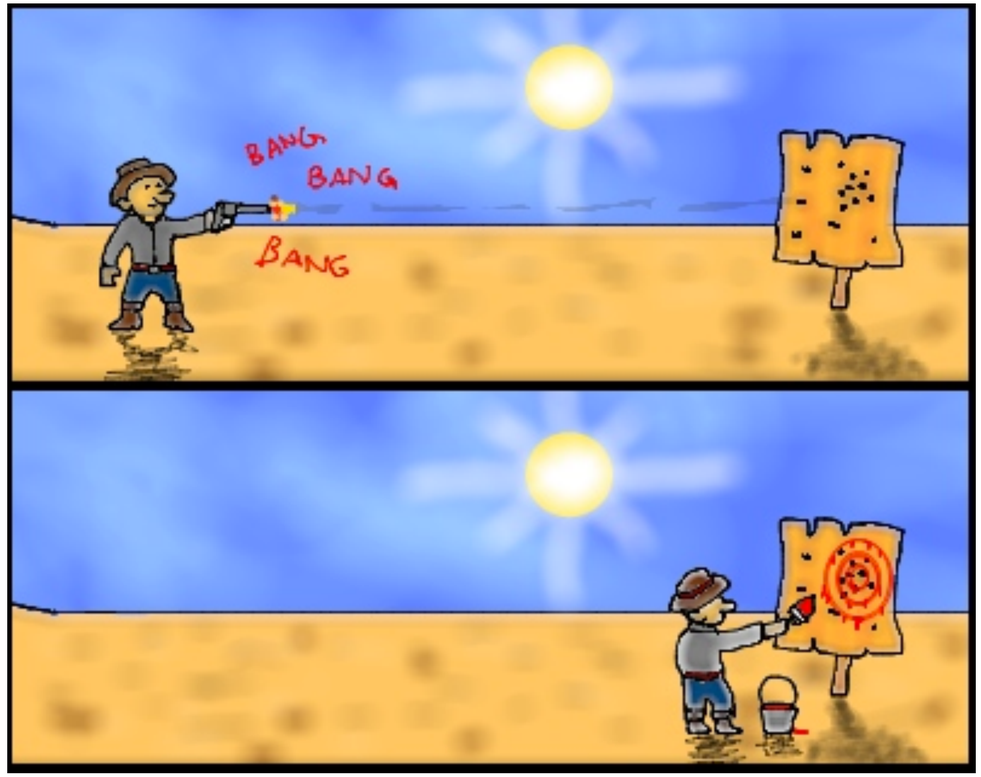
- A positive *fleeb* test only means you have a *7.8% chance of having fleeb cancer*.

- *Not* 80% (the supposed accuracy of the test).

- It might seem strange at first but it makes sense: the test gives a false positive 10% of the time, so there will be many false positives in any given population.

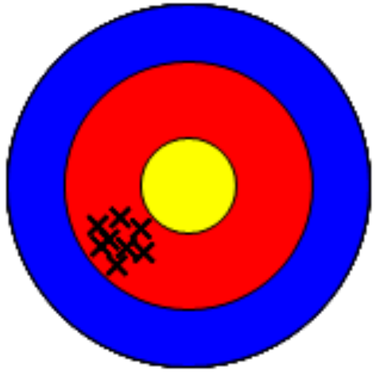
- There will be so many false positives, in fact, that most of the positive test results will be wrong.

Texas sharpshooter fallacy



The fallacy is characterized by a lack of a specific hypothesis prior to the gathering of data, and the formulation of a hypothesis only after data have already been gathered and examined.

Accuracy vs. Precision

	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)