Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

Jan. 29, 2020

1.2: Thinking Like a Scientist

ANNOUNCEMENTS

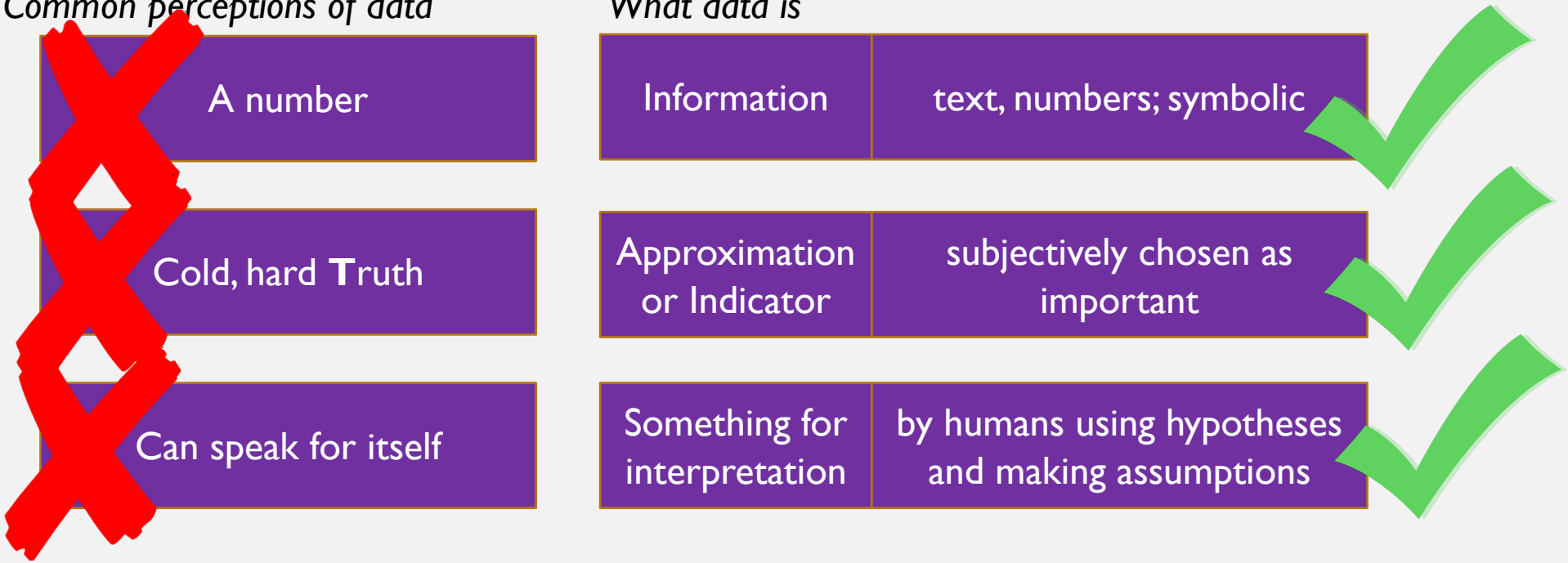
1. Data Science for Everyone vs. Intro. to Data Science (Prof. Policastro)
2. Homework 1 out Monday, Feb. 3; due Tuesday, Feb. 18
    - Due Tuesday (Monday, Feb. 17, is a holiday)
3. Lab 0 out Wednesday, Feb. 5; due Wednesday, Feb. 12
    - Content: practice formatting and submitting assignments for this semester
    - That's it! But this is *just* how important it is
    - Goal of labs: Practice.
    - Design: You can complete it in section!
4. Lecture & section materials are on Classes
    - And the syllabus, but you already knew (and read) that

Outline

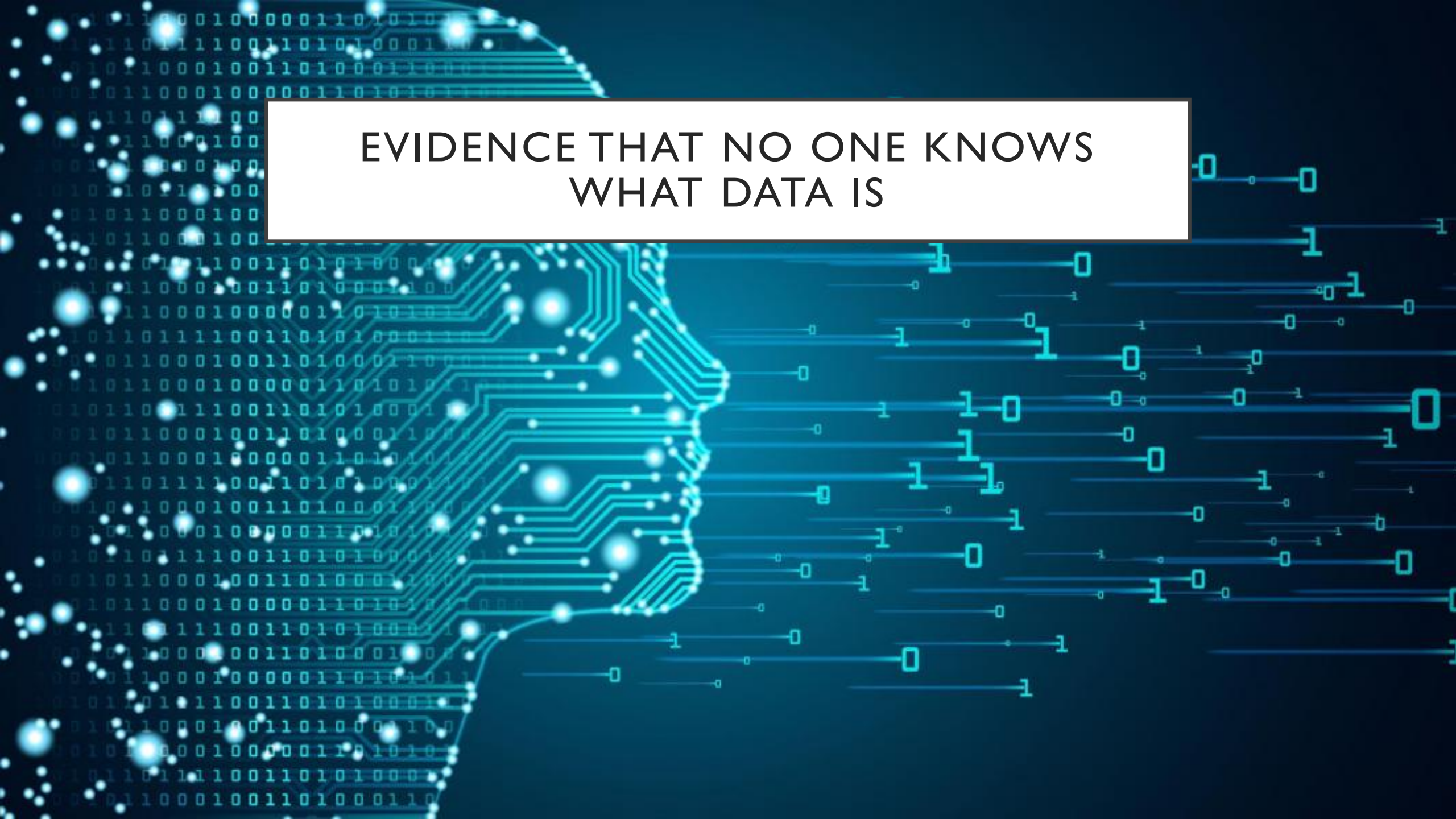
1. Data
2. Science
3. Thinking like a scientist

DATA ISN'T WHAT MANY THINK IT IS

Common perceptions of data	*What data is*	
A number	Information	text, numbers; symbolic
Cold, hard **T**ruth	Approximation or Indicator	subjectively chosen as important
Can speak for itself	Something for interpretation	by humans using hypotheses and making assumptions




EVIDENCE THAT NO ONE KNOWS
WHAT DATA IS


EVIDENCE THAT NO ONE KNOWS
WHAT DATA IS

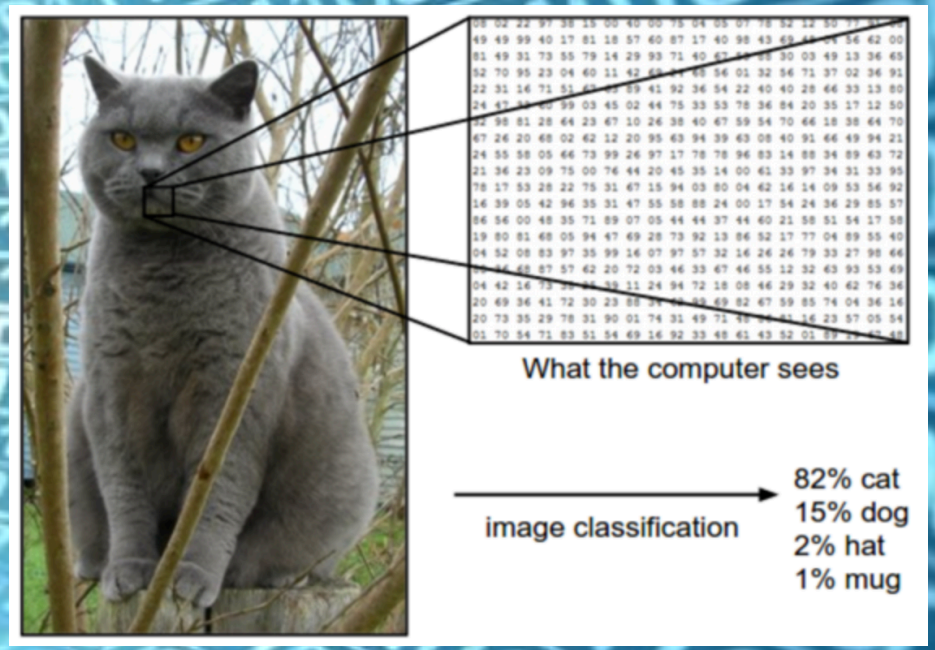
DATA =

Information

text, numbers; symbolic

Number of petals	Name	Color
8	lotus	pink
34	sunflower	yellow
5	rose	red

Chapters
ONE PLAYING PILGRIMS "Christmas won't be Christmas wit ...
TWO A MERRY CHRISTMAS Jo was the first to wake in the ...
THREE THE LAURENCE BOY "Jo! Jo! Where are you?" crie ...
FOUR BURDENS "Oh, dear, how hard it does seem to take ...
FIVE BEING NEIGHBORLY "What in the world are you going ...
SIX BETH FINDS THE PALACE BEAUTIFUL The big house did ...
ALLEY OF HUMILIATION "That boy is a perfe ...
TS APOLLYON "Girls, where are you going?" ...
ES TO VANITY FAIR "I do think it was the mo ...
C. AND P.O. As spring came on, a new set of ...



DATA IS NOT TRUTH

Forbes

Data Is Not The Same As Truth: Interpretation In The Big Data Era

One of the most common misconceptions of the "big data" world is that from data comes irrefutable truth. Yet, any given piece of data records only a small fragment of our existence and the same piece of data can often support multiple conclusions depending on how it is interpreted. What does this mean for some of the major trends of the data world?

Read here and here (reminder: links in lectures slides are optional).

The New York Times

Why Big Data Is Not Truth

The word "data" connotes fixed numbers inside hard grids of information, and as a result, it is easily mistaken for fact. But including bad product introductions and wars, we have many examples of bad data causing big mistakes.

Big Data raises bigger issues. The term suggests assembling many facts to create greater, previously unseen truths. It suggests the certainty of math.

Measurement?

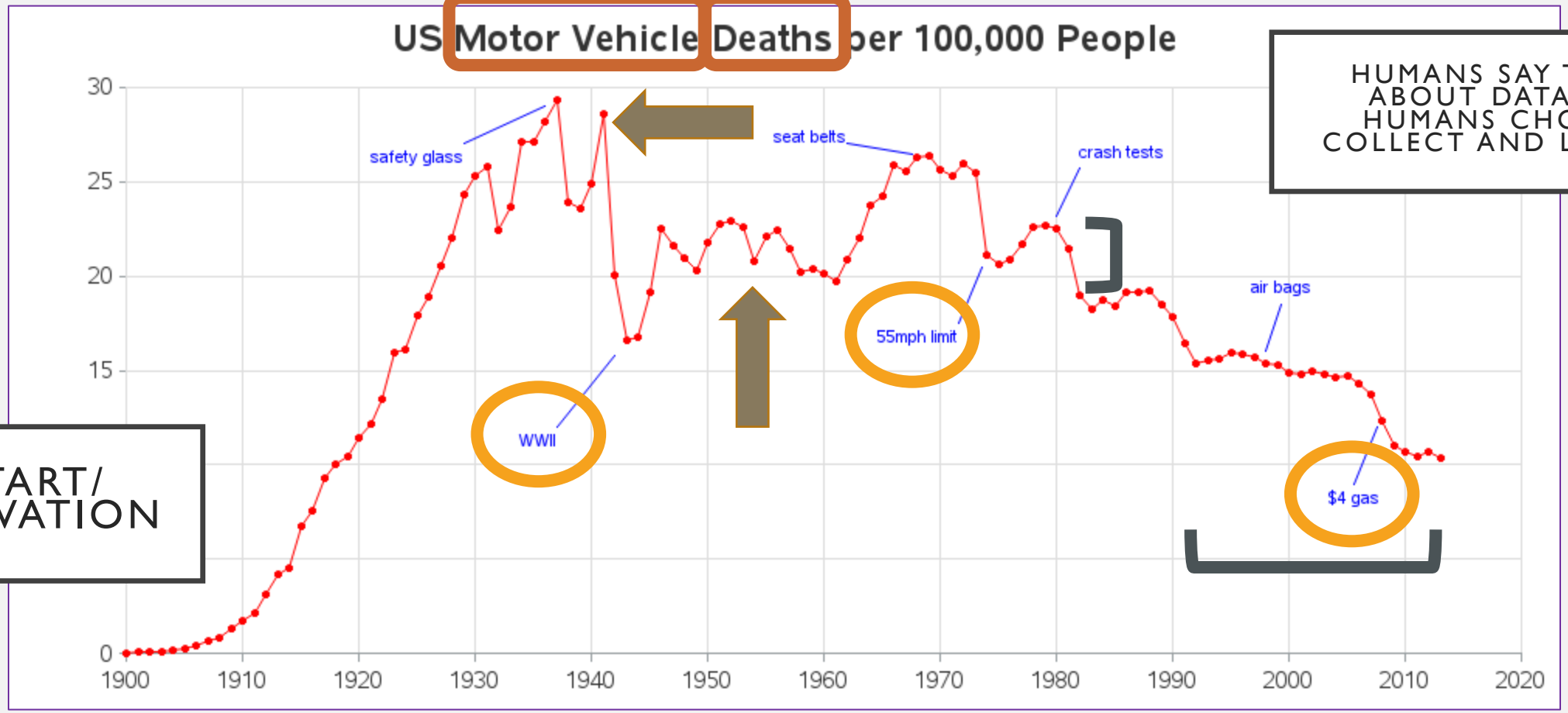
What about points not labeled?

Are any of these changes **meaningful**?

What policies should we adopt to reduce fatalities?

DATA DOESN'T SAY ANYTHING!

(on its own)



HUMANS SAY THINGS ABOUT DATA THAT HUMANS CHOSE TO COLLECT AND LOOK AT!

A START/ OBSERVATION

WHA DOES E DAT SAY?

Outline

1. Data
2. Science
3. Thinking like a scientist

SCIENCE ISN'T WHAT MANY THINK IT IS

Common perceptions of science	*What science is*
About certainty & truth	About doubt & skepticism; always tentative
About proving things	About disproving things
About getting answers and being correct	About asking questions; humility that we're always wrong; a study is a simplified snapshot

EXAMPLE: WHAT IS THE WEALTHIEST COUNTRY?

Measures:

- GDP
- GDP per capita
- Distribution of GDP
- Which currency? Current USD?
- Purchasing power parity
- The number of millionaires or billionaires
- The number of people below the poverty line (chose
- Health or well-being outcomes?
- Combined: GDP/Capita weighted by the Gini index

How is this measured?

This?

- Private consumption + gross investment + government spending + (exports-imports)
- Nominal (don't adjust for inflation); Real (adjust for inflation)
- Doesn't account for activities that are bad for the country long-term
  - Net positives:
    - Oil spills
    - Terrorism
    - Overfishing
    - Plane crashes

EXAMPLE: WHY ARE SOME COUNTRIES WEALTHIER THAN OTHERS?

- Stable government
- Natural resources
- Peace
- Friendly neighboring countries
- Low crime
- High investment
- Low inflation
- Property rights
- Rule of law
- Religion, culture, or values
- Luck

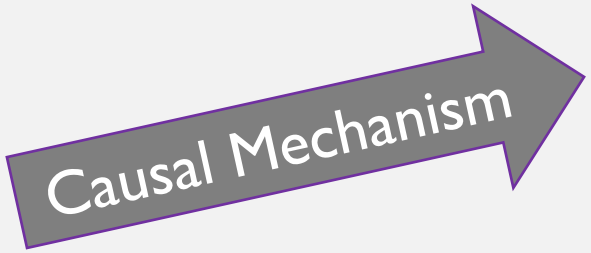
How is this measured?

This?

EXAMPLE: WHY ARE SOME COUNTRIES WEALTHIER THAN OTHERS?

Scientific method
Pre-step: Be interested in something

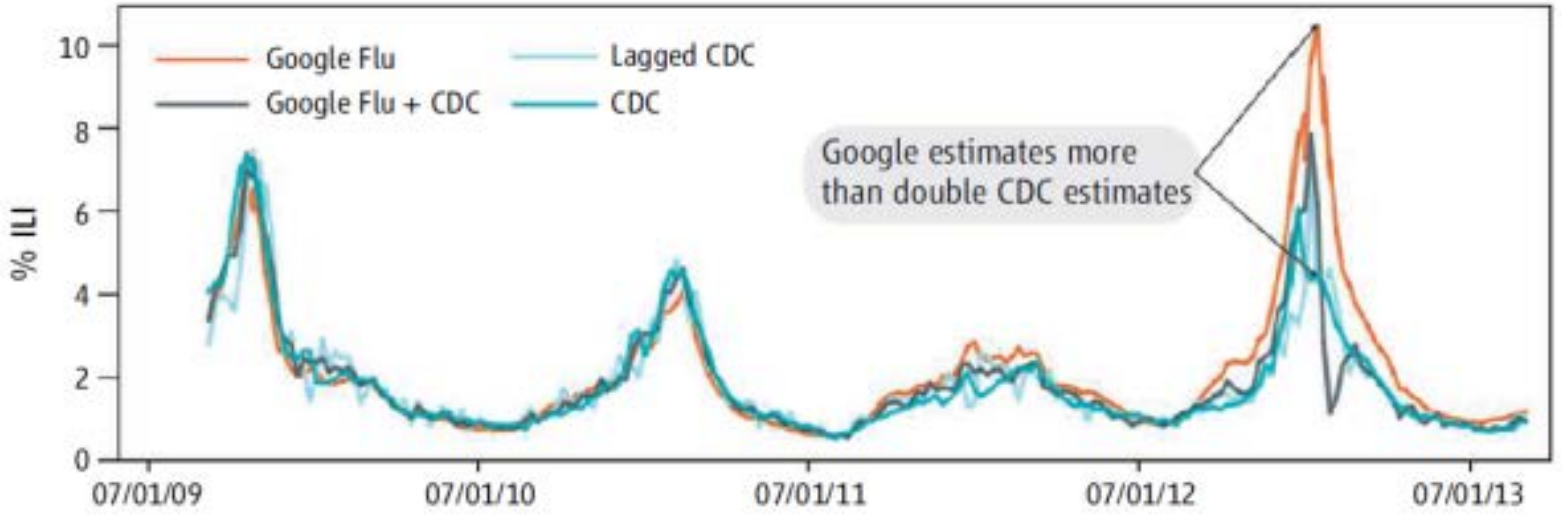
Observation	Question	Theory	Hypotheses	Test	Update theory
Some countries are wealthier than others	Why?	Political stability causes countries to be wealthier	H1: Countries that are wealthy are likely to be stable; H0	Design the test	Maybe, but endogeneity
		Because it allows for long-term investment	H2: Countries that are stable have more investment; H0	Generate or collect, organize, and clean data Conduct the test	Under what conditions? Necessary? Sufficient? Neither?



SCIENCE

EXAMPLE: GOOGLE FLU TRENDS

Google gathers data on search terms (including symptoms), and uses that to predict who has flu:



But what might happen if a state declared a 'flu' emergency?

Outline

1. Data
2. Science
3. Thinking like a scientist

THINKING LIKE A SCIENTIST

Generally

Using the scientific method to better understand likely causal relationships

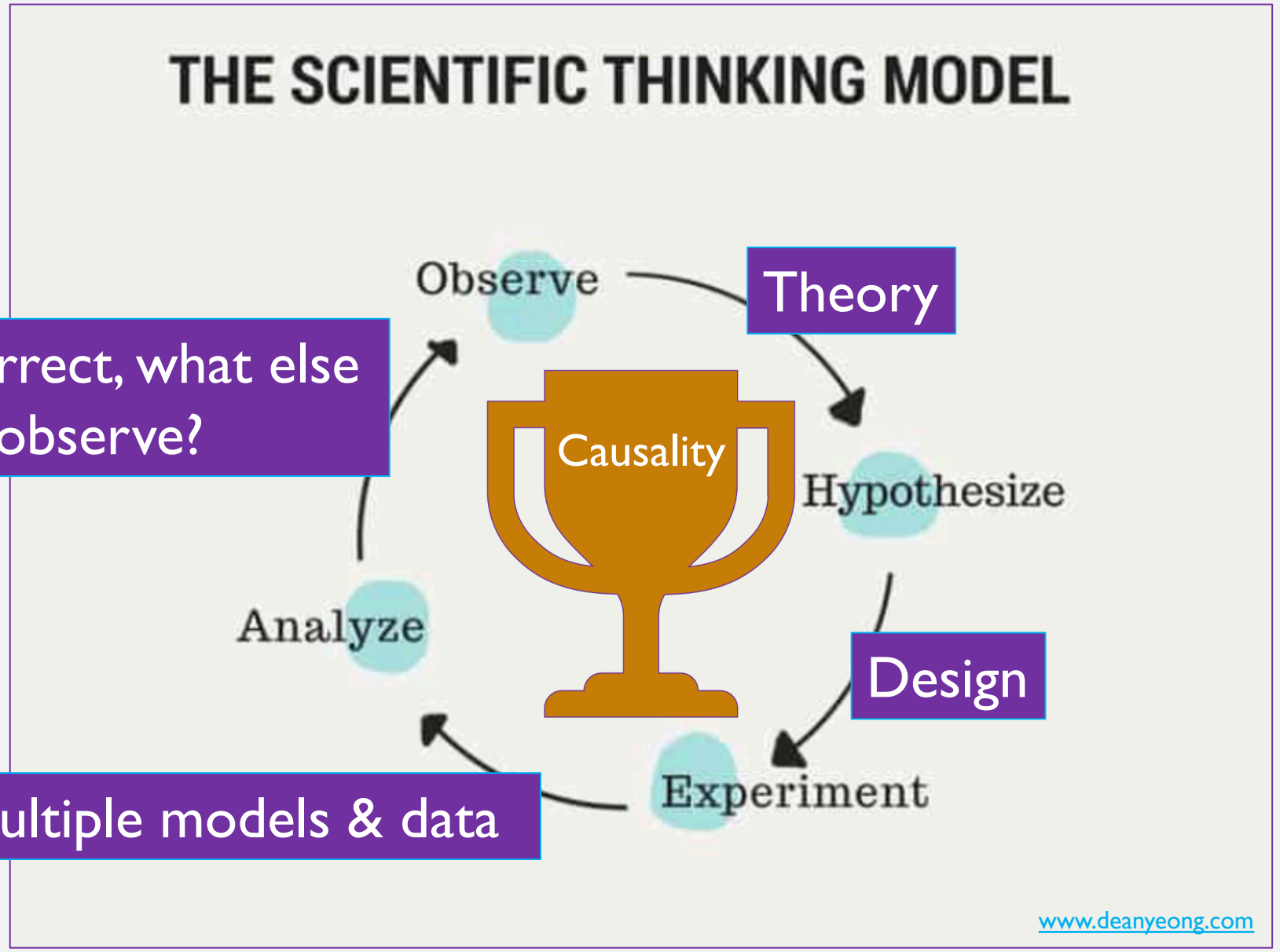
Bonus traits

Persistence in trying to disprove your own and others' findings

Humility about biases & challenges to inference

Skepticism (at best) of anyone claiming to have proven a theory

Simplification of the scientific method




Holy grail
An elusive object or goal that is sought after for its great significance
For science
Causality

Why do some people get cancer and not others?	What are the causes of cancer?
What makes two countries more likely to go to war?	What are the causes of war?
Why might a person experience depression or anxiety?	What are the causes of happiness?
How can we eradicate malaria?	What are the causes of infection?
What makes one football team better than another?	What are the causes of success?
How can we predict which startup companies will succeed?	What are the causes of success?

THINKING LIKE A SCIENTIST

Using the scientific method to better understand causal relationships

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test
6. Update theory

DATA

7. Repeat as desired

Check out a video of the story here.

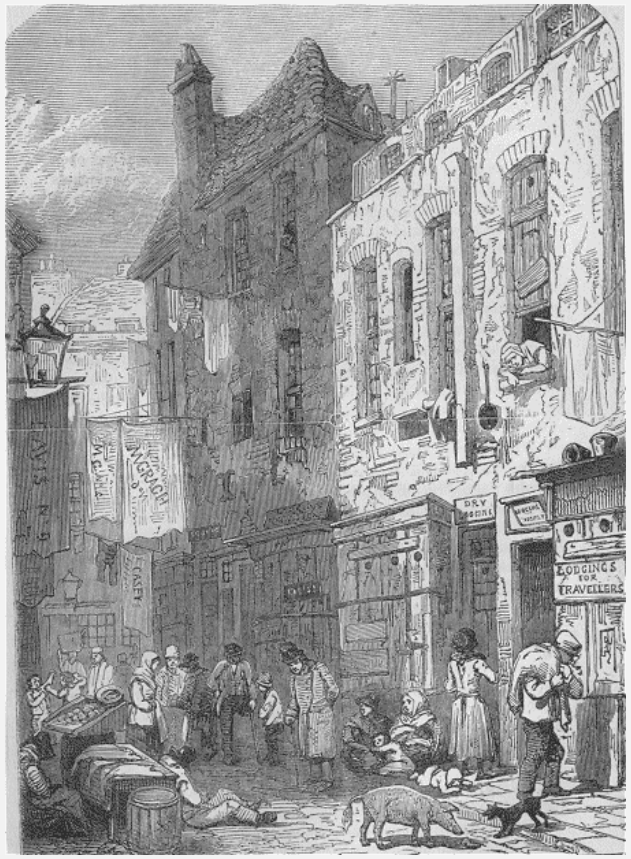
(Complement not substitute for the reading!)

EXAMPLE: READING FOR LECTURE 2.1

Ch. 2 John Snow & the Broad Street Pump

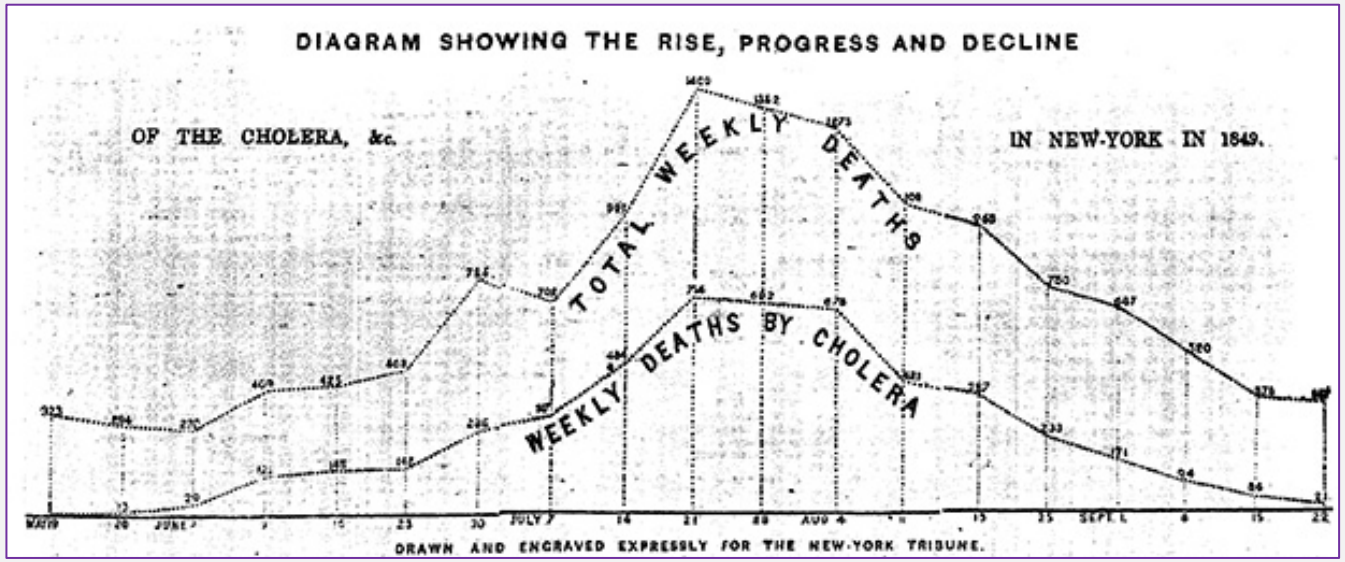
(a very good scientist)

JOHN SNOW & BROAD STREET PUMP



- London in the 1850s (but common in many cities)
- Waves of cholera, killing tens of thousands of people each
- John Snow: Doctor treating patients. Massive outbreak in summer 1854.

WHAT SHOULD WE DO?!



NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit !.

Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

SLEEP AND CLOTHE WARM !

DO NOT SLEEP OR SIT IN A DRAUGHT OF AIR.

Avoid getting Wet !

Attend immediately to all disorders of the Bowels.

TAKE NO MEDICINE WITHOUT ADVICE.

Medicine and Medical Advice can be had by the poor, at all hours of the day and night, by applying at the Station House in each Ward.

CALEB S. WOODHULL, *Mayor.*

JAMES KELLY, *Chairman of Sanatory Committee.*

OBSERVATIONS: WHAT DOES SNOW KNOW?

- Immediately ***deadly*** – you die within days of contracting
- Patterns of death:
  - Often people *within* one house would *all* die
  - But their *neighbors* weren't infected
- Symptoms:
  - Digestive problems

THINKING LIKE A SCIENTIST

SCIENTIFIC METHOD

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test
6. Update theory

DATA

7. Repeat as desired

Outline

1. Data
2. Science
3. Thinking like a scientist