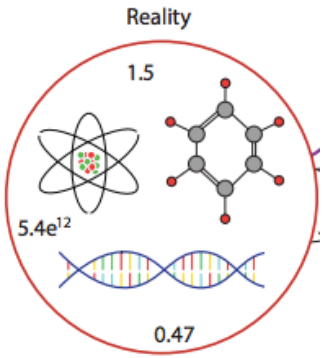
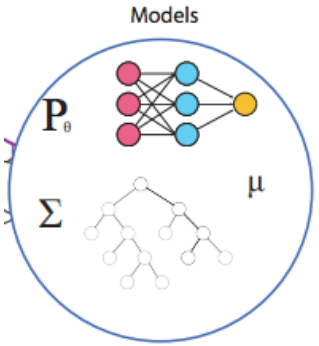


Future reality



Statistics 154, Spring 2019

Modern Statistical Prediction and Machine Learning

Lecture 3: Causality, experimental design, data quality

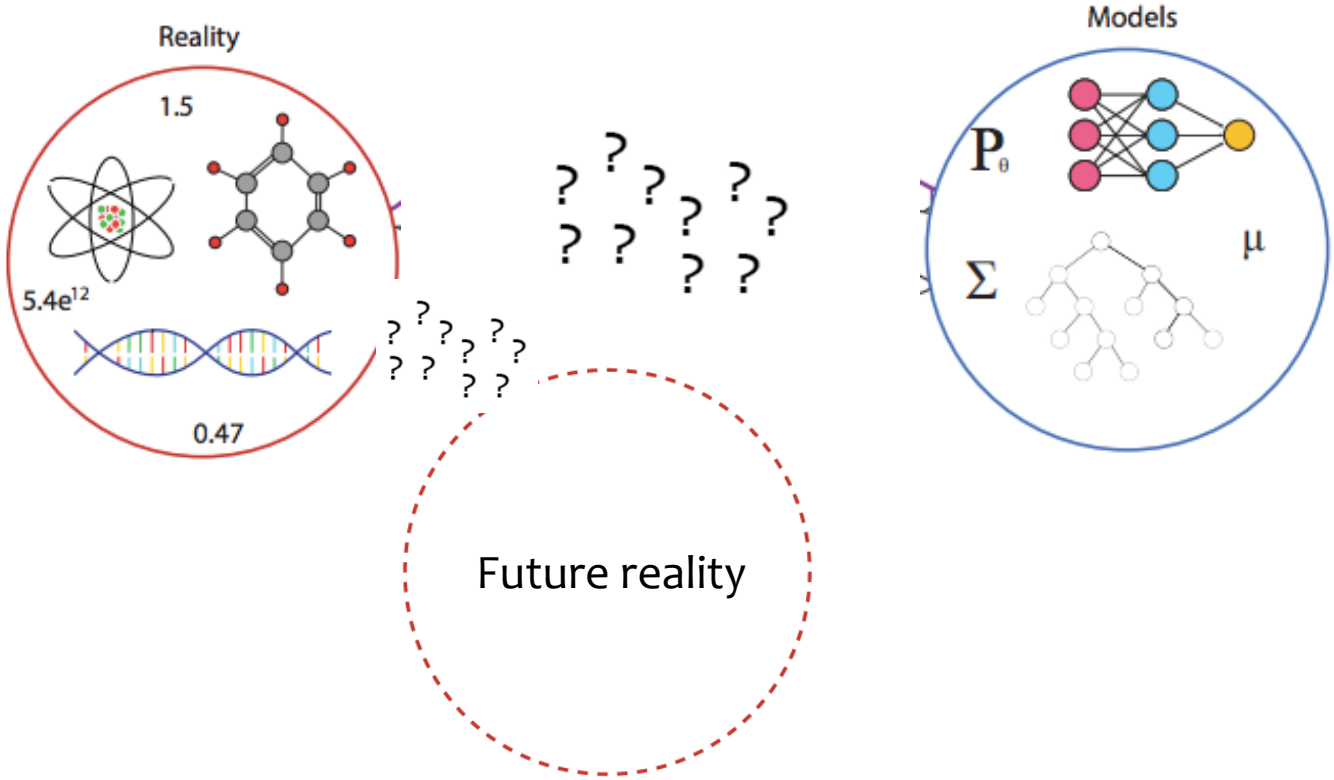
Instructor: Bin Yu
(binyu@berkeley.edu); office hours: Tu: 9:30-10:30 am; Wed: 9:00-10:00 am
office: 409 Evans

GSIs: Yuansi Chen (Mon: 10-12; 12-2); Raaz Dwivedi (Mon: 2-4; 4-6)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu
(office hours to be announced)

Special thanks to Rebecca Barter

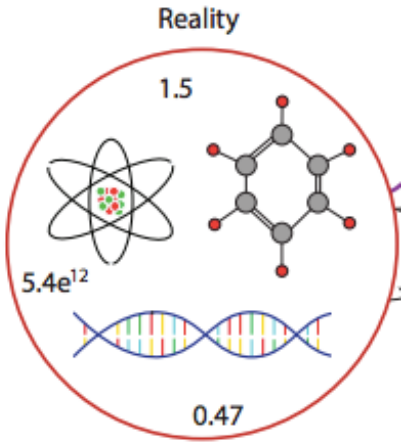
Recap: **Why are we here?**

To solve prediction problems in real world
by connecting the two solid circles below
in a justifiable way to say things about the third cycle

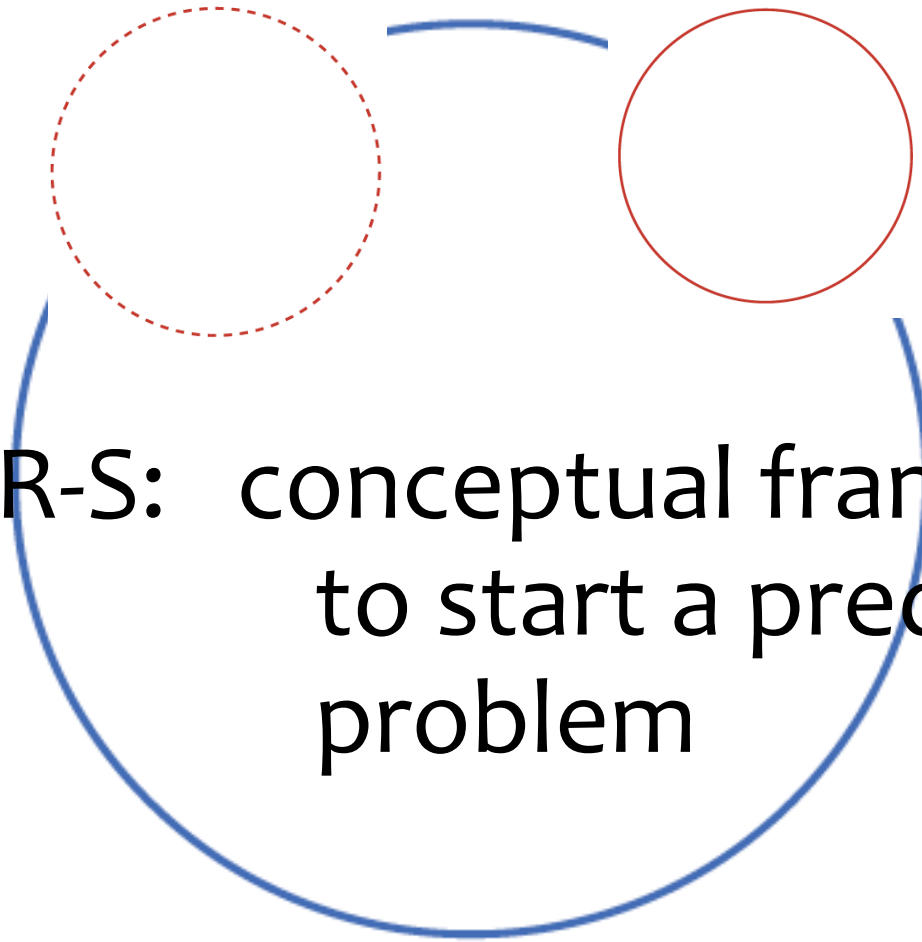


Recap: **What will you learn?**

Future reality



- Problem formulation
- Data collection, data cleaning
- EDA (exploratory data analysis, visualization)
- Unsupervised learning (e.g. PCA, clustering)
- Supervised learning (LS, regularized LS, Kernel regression, logistic regression, Support Vector Machines (SVMs), Nearest Neighbor (NN), Decision trees, Random Forests, Deep Learning)
- Data results, validation, conclusions

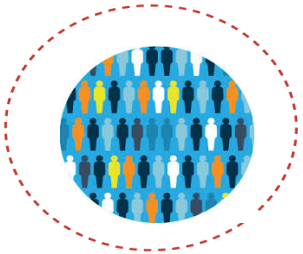


PQR-S: conceptual framing to start a prediction problem

or a check-list

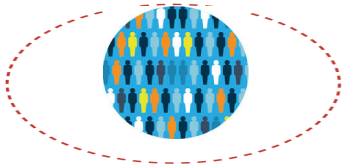
PQR-S helps you think straight!

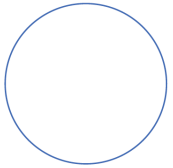
- **P**opulation



- **Q**uestion

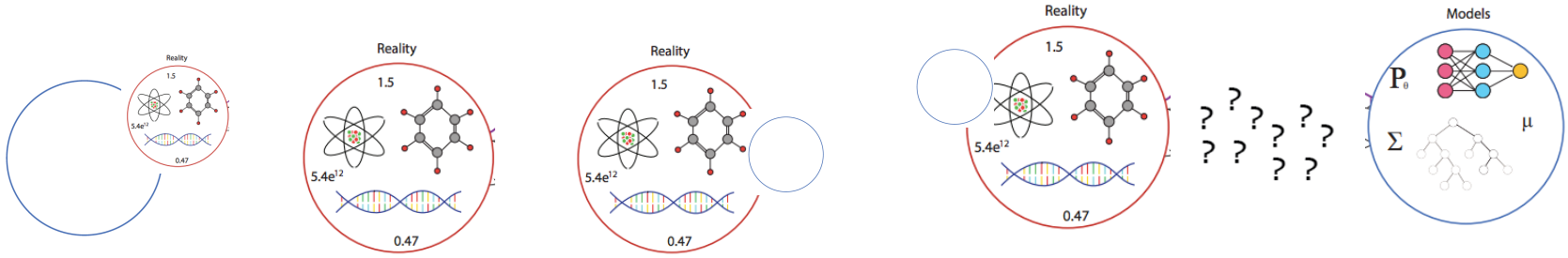
- **R**epresentative data collection (data neutral, fairness)

is  similar to  ?

- - (to be filled in throughout the course )

- **S**crutinize or validate data results

PCS=Predictability, computability and stability Stability (or robustness): a paramount issue in the data science life cycle



Domain Problem → Data collection/ storage → **Data cleaning** → Data visualization or EDA → **Unsupervised and supervised ML** → Post-hoc analysis/ EDA → **Model validation** → Interpretation data results → Get new data → Revised domain problem

http://www.odbms.org/2015/04/data-wisdom-for-data-science/

Experimental design

The science and subfield of statistics about how to collect data effectively…

R. A. Fisher (1890-1962)
Founding father of Modern Statistics
Geneticist

Predictability vs. causality

Association vs. causation

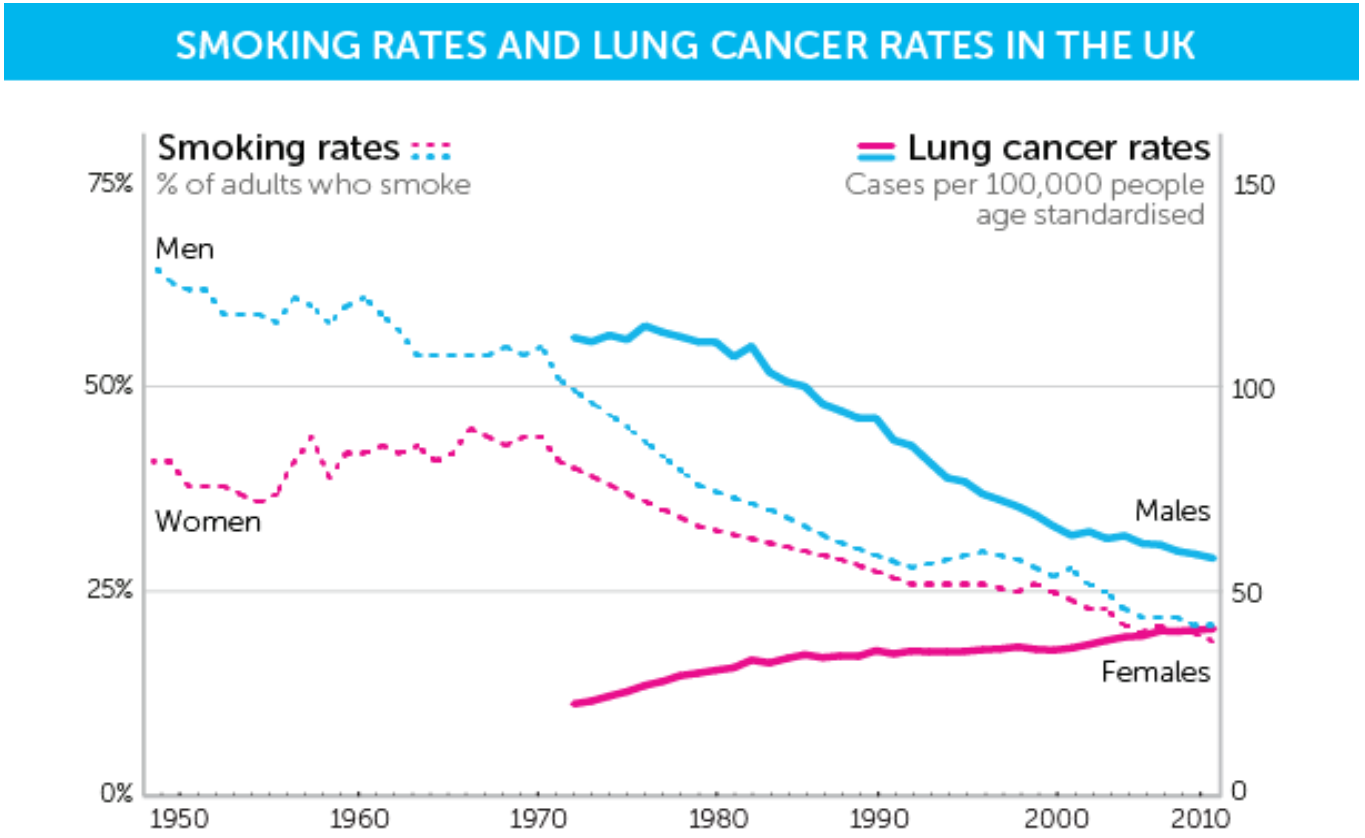
- It is a common practice that people jump from association to causation
- Believing in causation somehow gives a sense of control
- Marketing people use this confusion to their advantage
- Q: when did you do it last time?

Predictability vs. causality

- How to collect data to establish causality?
- Or what is the design principle to establish causality

Does smoking rate predict cancer rate?

UK Cancer Research



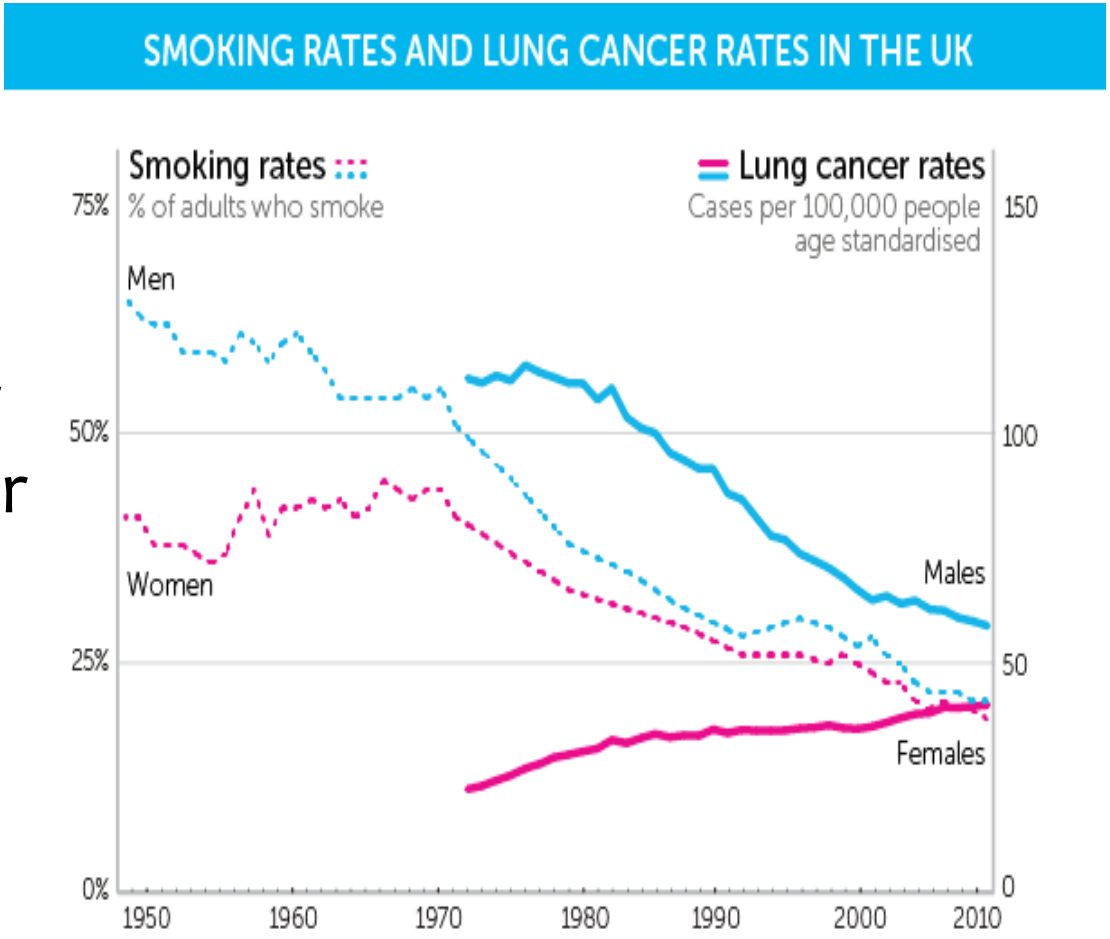
Does smoking cause lung cancer?

Population?

Thanks to R. Barter for help on some smoking-cancer slides

Does smoking cause cancer?

UK Cancer Research



Can we conclude that smoking causes cancer for men; but is good for women??

Possible explanations based on "colleague-sourcing"

- Lung cancer is not all the same
- Only 10-15% is closely associated to smoking
- Another type is not, but occurs more among women
- Women started smoking later than men and the cohort with peak smoking is still working its way through the population
- Women are more susceptible to tobacco toxins
- …

Thanks to P. Stark, P. Ding, J. Sekhon

Lessons

- Ecological correlation does not imply correlation at person level (ecological correlation = correlation between rates)

- Association (correlation) is not causation, although association is great for prediction.

- Confounding factors are always lurking in the back (confounding factor: a possible driver for both smoking and lung cancer, e.g. genetics)

The first solid epidemiological evidence, or observational study

BRITISH MEDICAL JOURNAL

LONDON SATURDAY NOVEMBER 10 1956

LUNG CANCER AND OTHER CAUSES OF DEATH IN RELATION TO SMOKING

A SECOND REPORT ON THE MORTALITY OF BRITISH DOCTORS

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

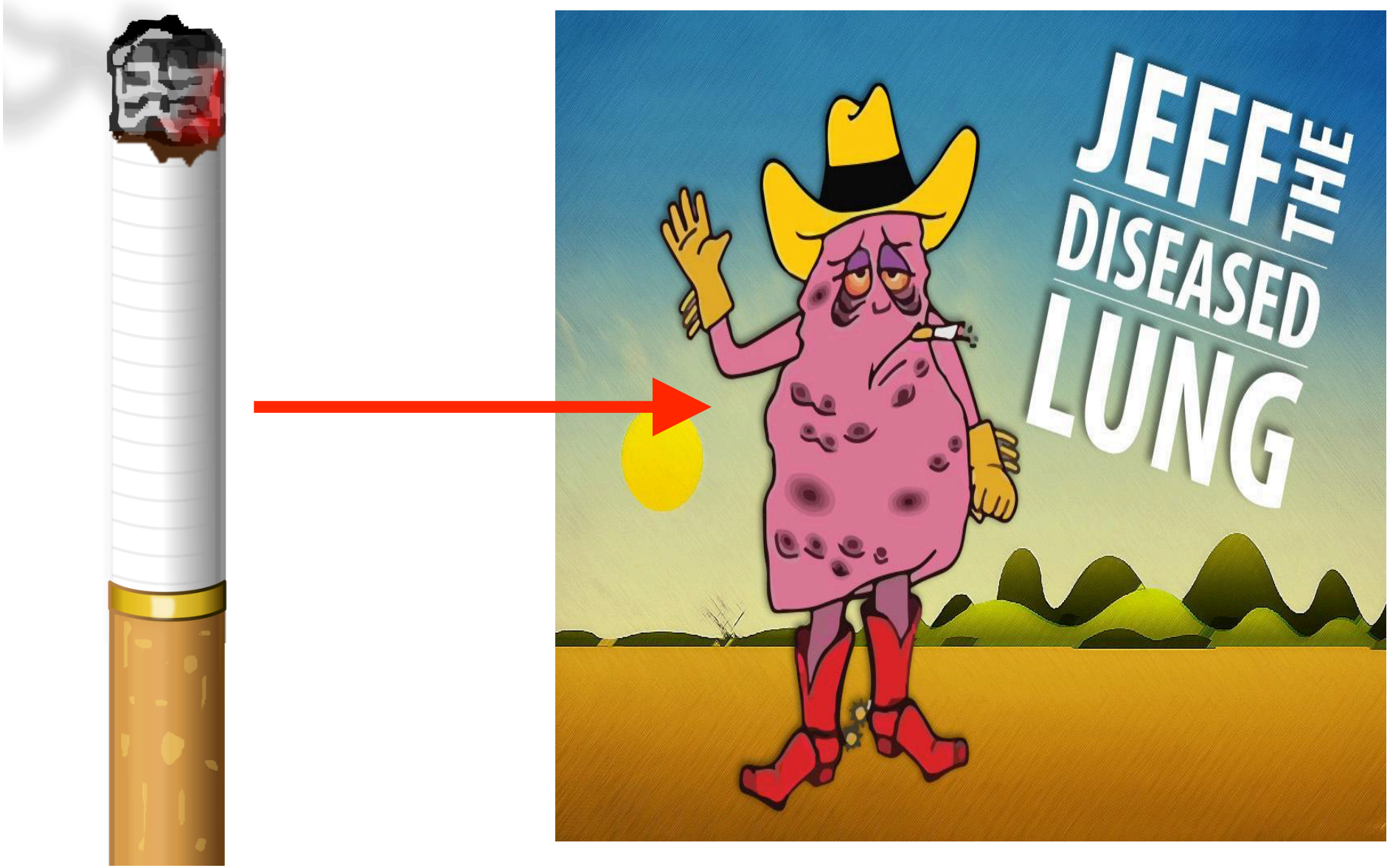
AND

A. BRADFORD HILL, C.B.E., F.R.S.

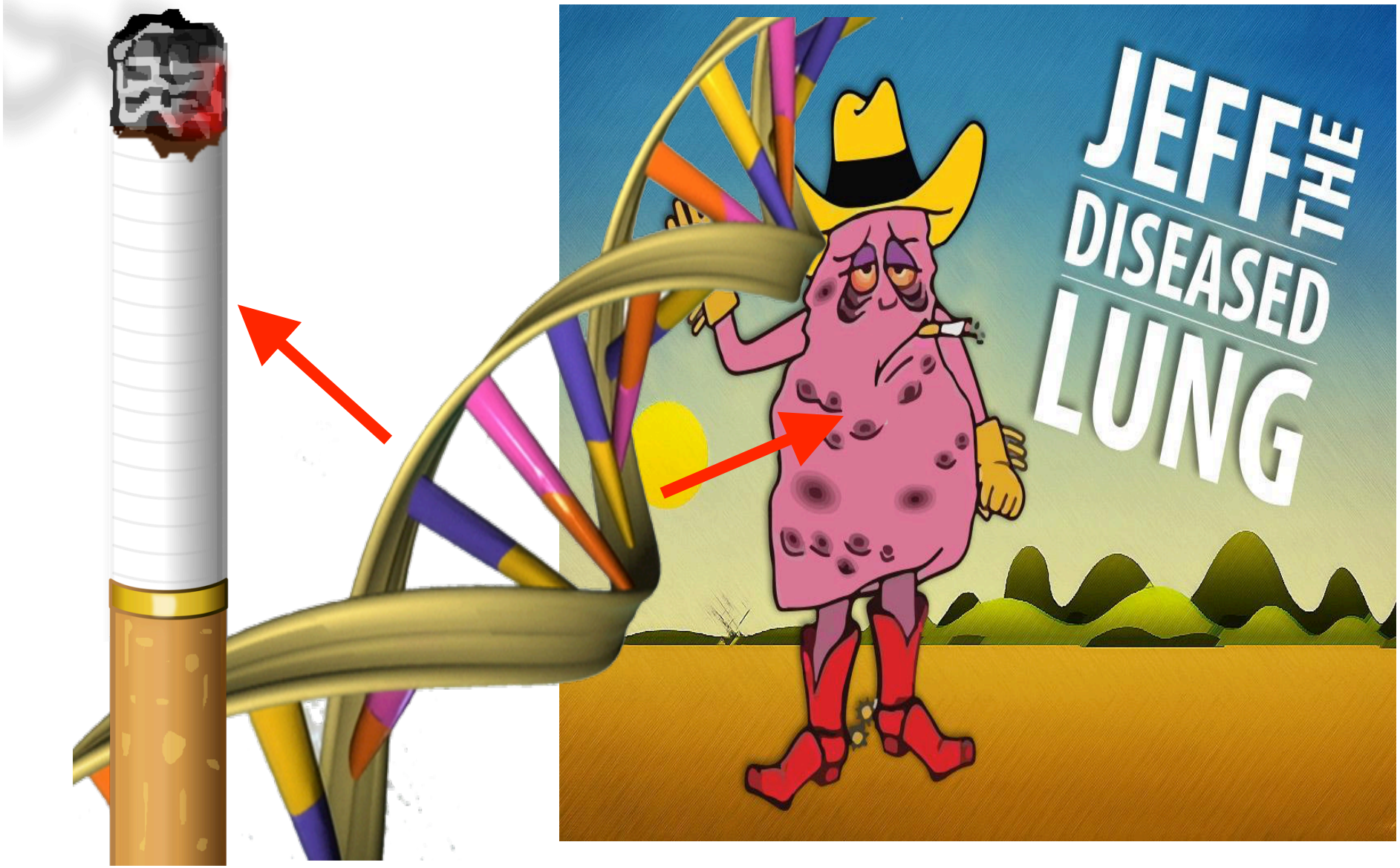
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

On October 31, 1951, we sent a simple questionary to all members of the medical profession in the United Kingdom. In addition to giving their name, address, and age, they were asked to classify themselves into one of three groups—namely, (*a*) whether they were, at that time, smokers of tobacco; (*b*) whether they had smoked but had given up; or (*c*) whether they had never smoked regularly (which we defined as having never smoked as much as one cigarette a day, or its equivalent in pipe tobacco or cigars, for as long as one year). All smokers previously have been a light smoker or may since then have given up smoking altogether; we shall have continued to count him, or her, as a heavy smoker. If there is a differential death rate with smoking, we must by such errors tend to inflate the mortality among the light smokers and to reduce the mortality among the heavy smokers. In other words, the gradients we present in this paper may be understatements but (apart from sampling errors due to the play of chance) cannot be overstatements.

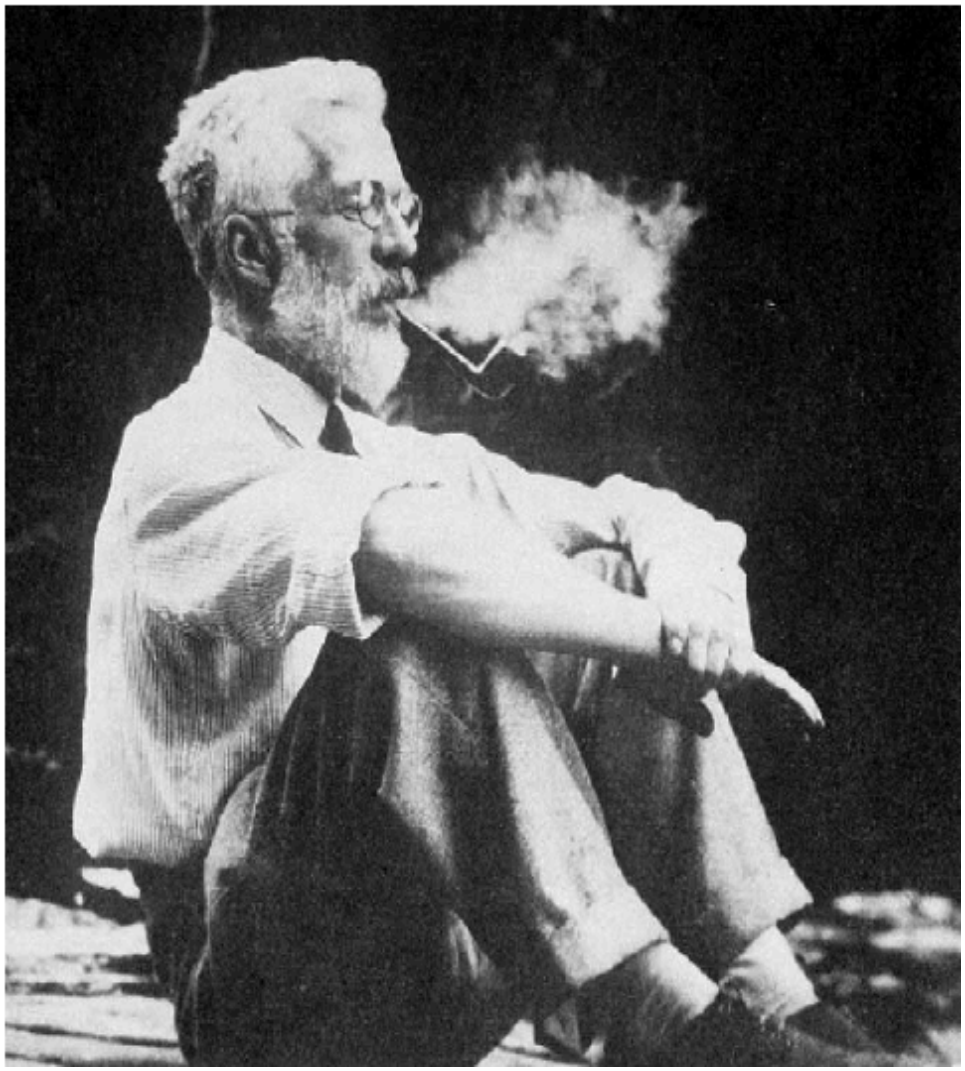
Cigarettes cause lung cancer!



Genetics cause both smoking and lung cancer? Or genetics could be a confounding factor

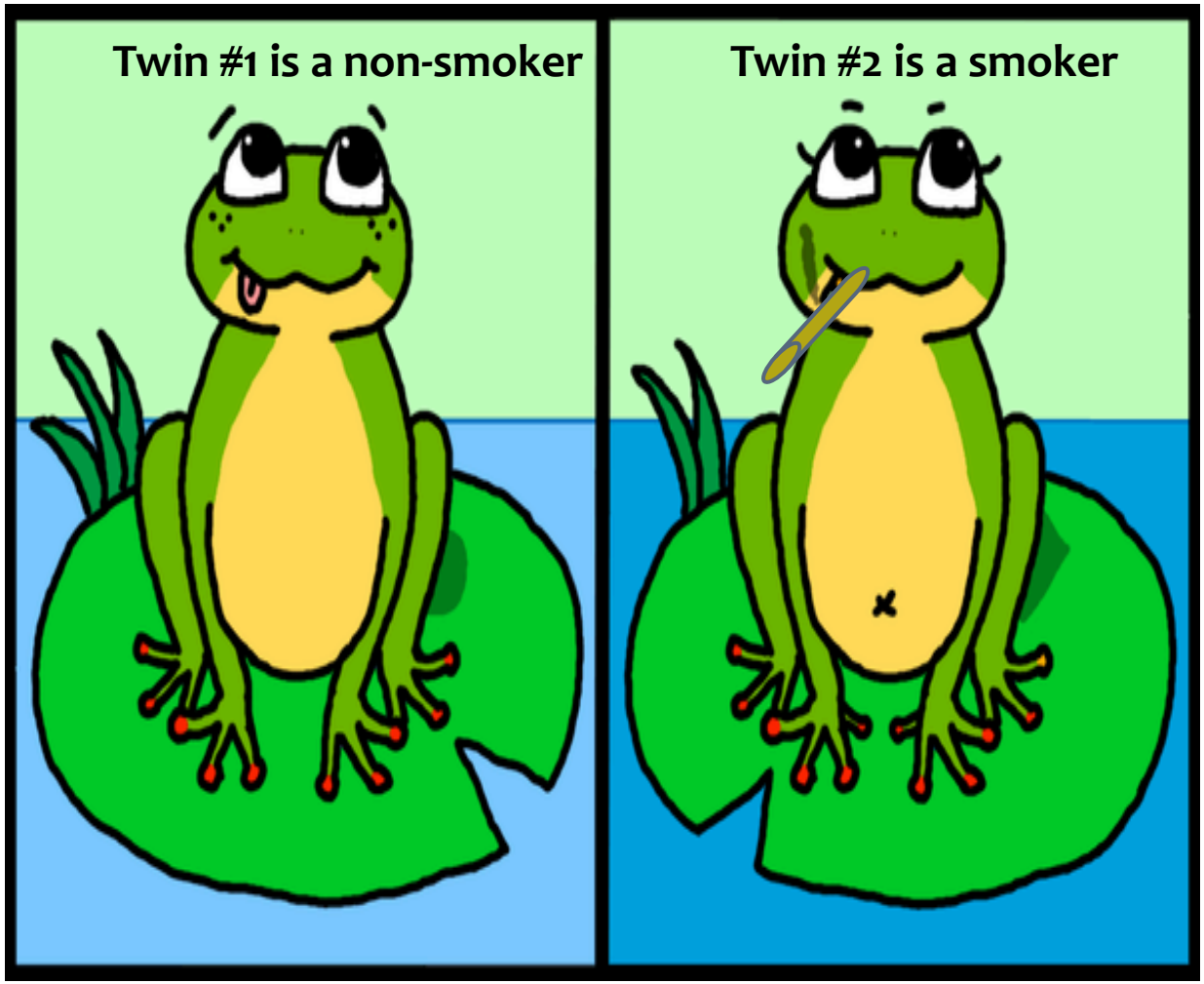


R. A. Fisher strongly believed in a common genetic cause!



17 February 1890 –
29 July 1962

Control for genetic differences: compare identical twins!



The evidence kept piling in from observational studies… Until finally in 1964

SMOKING *and* HEALTH

REPORT OF THE ADVISORY COMMITTEE
TO THE SURGEON GENERAL
OF THE PUBLIC HEALTH SERVICE

"Since 1939 there have been 29 retrospective studies of lung cancer alone which have varying degrees of completeness and validity."

"After appraising 16 independent studies carried on in five countries over a period of 18 years, this group concluded that there is a causal relationship between excessive smoking of cigarettes and lung cancer."

U.S DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

What is an ideal experiment to see whether smoking causes cancer?

- What is the population of interest?
- Are you going to ask some people to smoke? What is the difference from people choose to smoke?
- And then what?

Let us look at another problem: does BrainPlus IQ change humanity?

Anderson Cooper: Stephen Hawking Predicts, "This Pill Will Change Humanity" And It's What I Credit My $20 Million Net Worth To

Featured In: YAHOO! GQ MensHealth TIME People Aol.

Recently Hawking made some comments in an interview with Anderson Cooper about BrainPlus IQ that would become the biggest event in human history.

"This pill unlocks your brain power, allowing you to adventure further inside your own brain than ever before. This is the most groundbreaking BrainPlus IQ ever created, and we had to showcase it to the world!"

National Geographic
Limited Edition Cover Page

How do we translate "change humanity" into a measureable outcome?

How do we measure "adventure further inside your own brain"?

Does BrainPlus IQ change humanity?

- What is the population? All people who want to take BrainPlus IQ (many are unborn yet)…
- An easier question: is BrainPlus IQ better on average for people who have signed up for a study conducted by the company? Translate “humanity” into IQ measure…

Data collection by observing the population

- Administer BrainPlus IQ on Monday
- Administer Placebo on Tuesday

What assumption are we making?

If want to remove the assumption

- Administer BrainPlus IQ to half of the group and placebo to the other half

  How would you choose the half?

Gold Standard of Causal Evidence

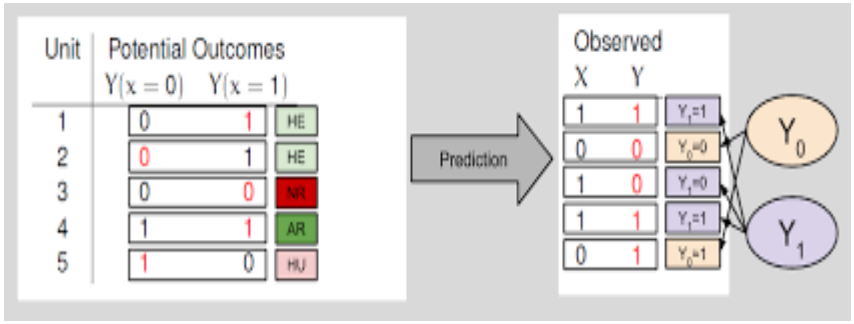
Randomization

which often relies on pseduo-random number generators (PRNGs) – deterministic processes that can be flawed …

It randomly assigns each subject to the treatment or control group – to take BrainPlus IQ or Placebo – or a random split.

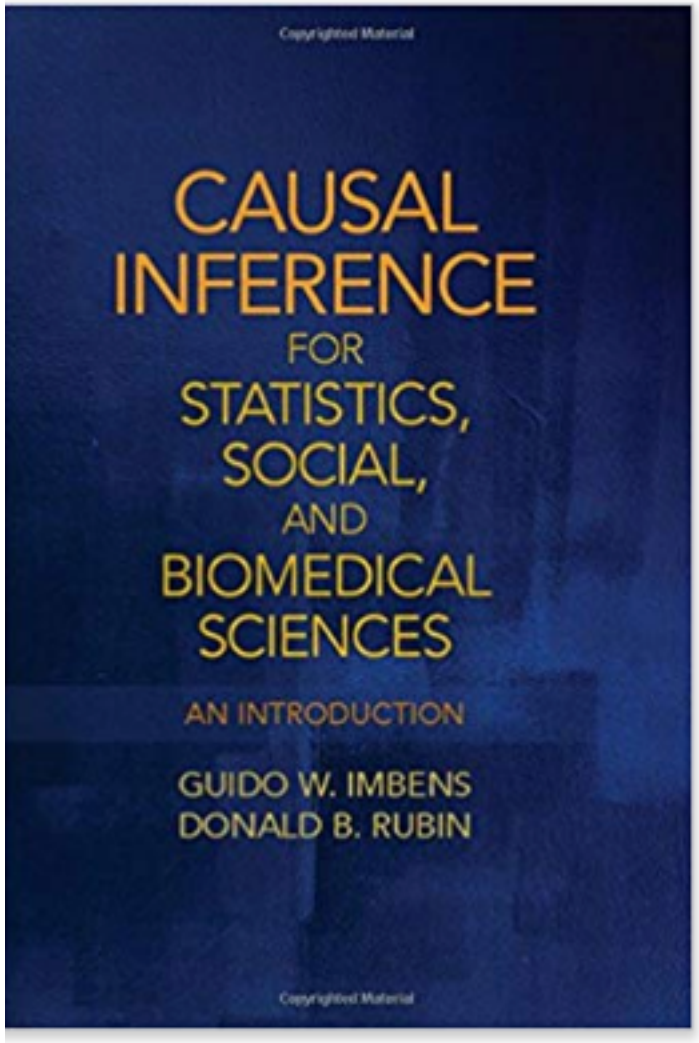
Neyman-Rubin model

- Population: potential outcomes under treatment and control for the subjects under study



Neyman was the founder of Berkeley Statistics;
Rubin is a Harvard Statistics Professor.

Book on causal inference by Imbens and Rubin



Three principles of experimental design

- Replication
- **Randomization**
- Blocking (reducing variability by using extra information – highly needed for the election situation)

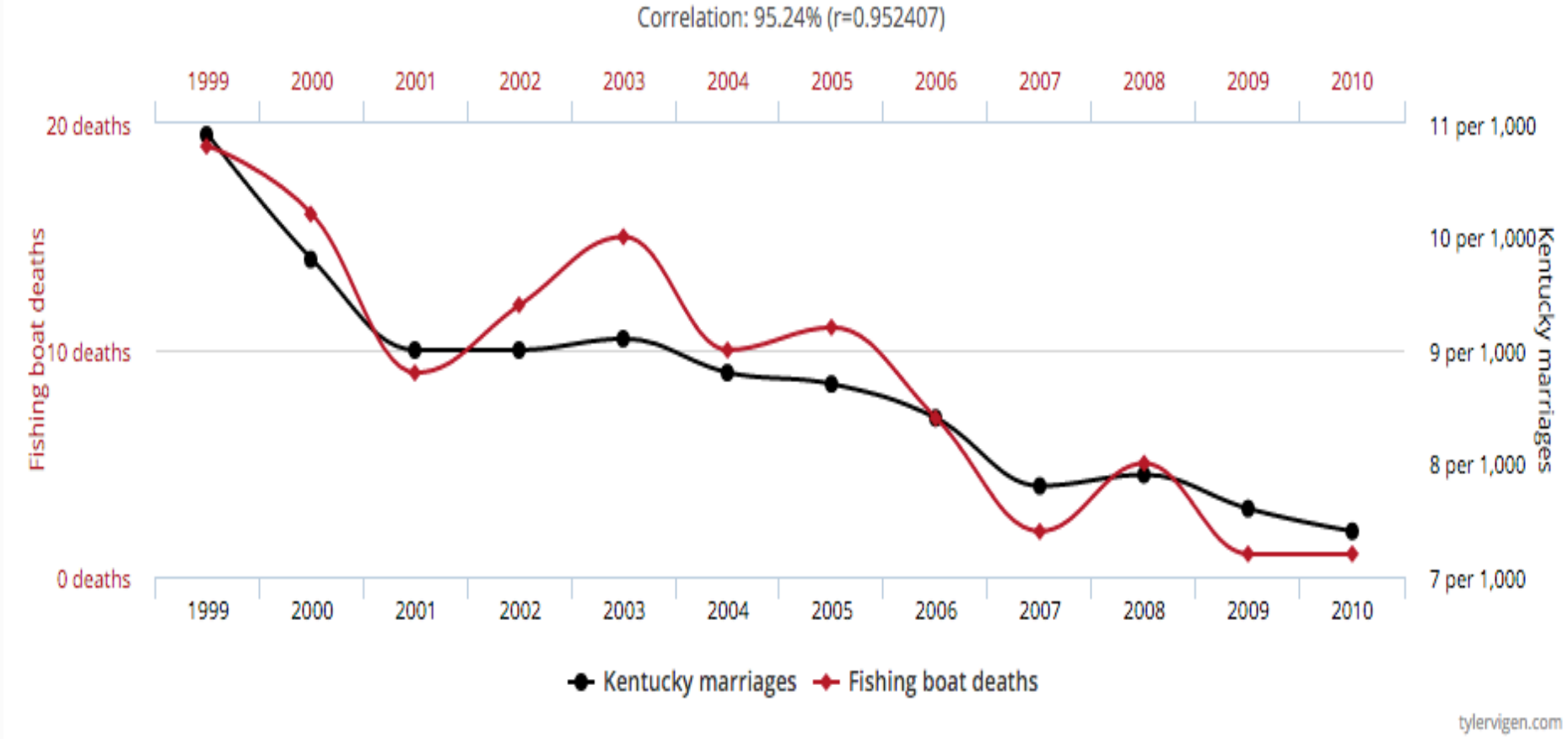
Summary: association is not causation

- Association is not causation
- Confounding factors often at play
- Randomization is the gold standard (randomized experiments)
- Many observational studies (not randomized experiments)

Last but not least:

a data-driven claim

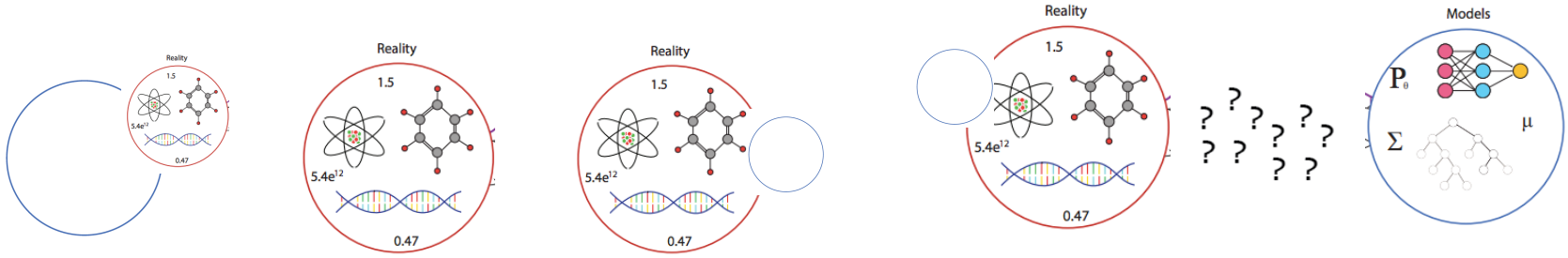
Marriage “causes” drowning...



Data sources: Centers for Disease Control & Prevention and National Vital Statistics Reports

Back to prediction…

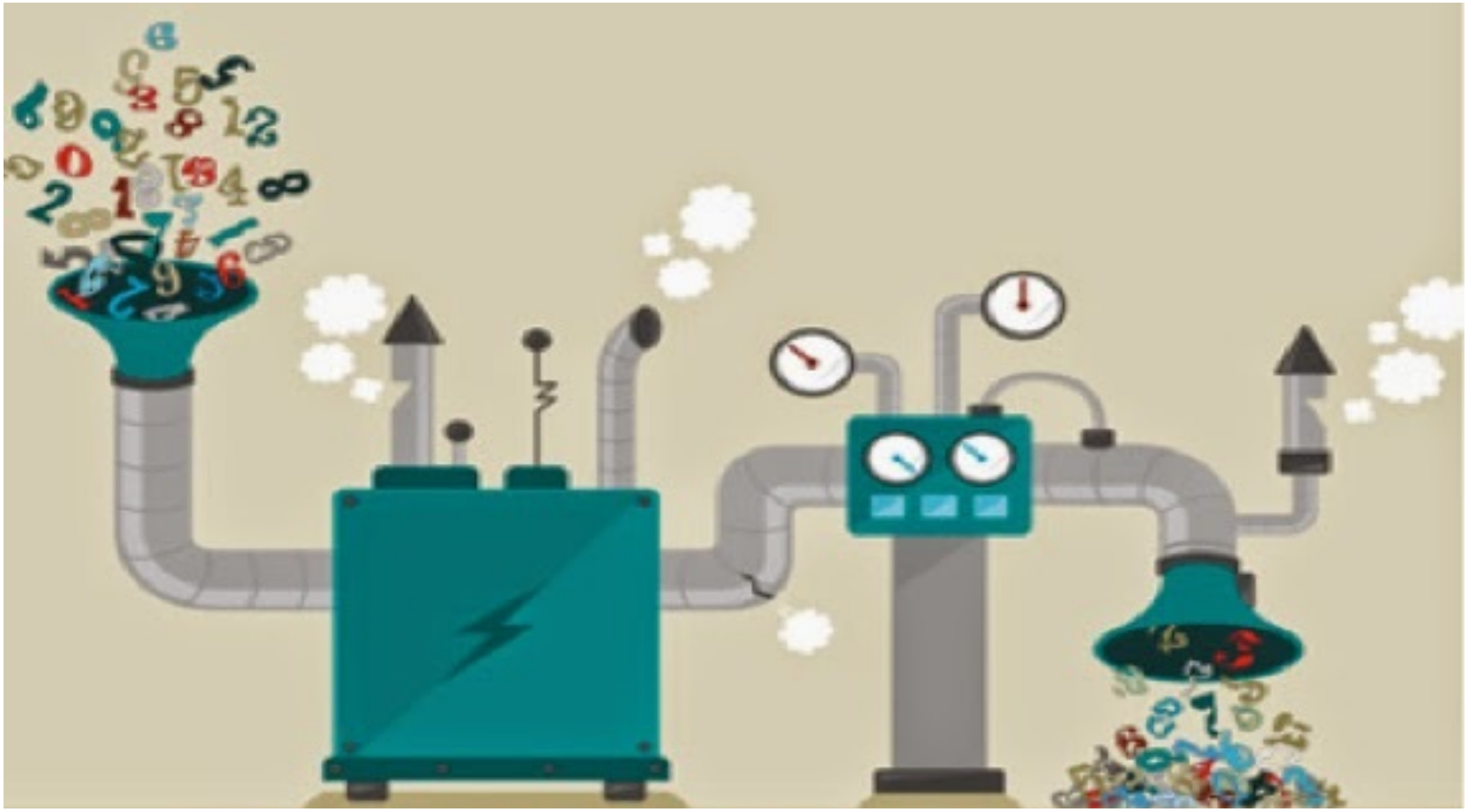
PCS=Predictability, computability and stability Stability (or robustness): a paramount issue in the data science life cycle



Domain Problem → Data collection/ storage → **Data cleaning** → Data visualization or EDA → **Unsupervised and supervised ML** → Post-hoc analysis/ EDA → **Model validation** → Interpretation data results → Get new data → Revised domain problem

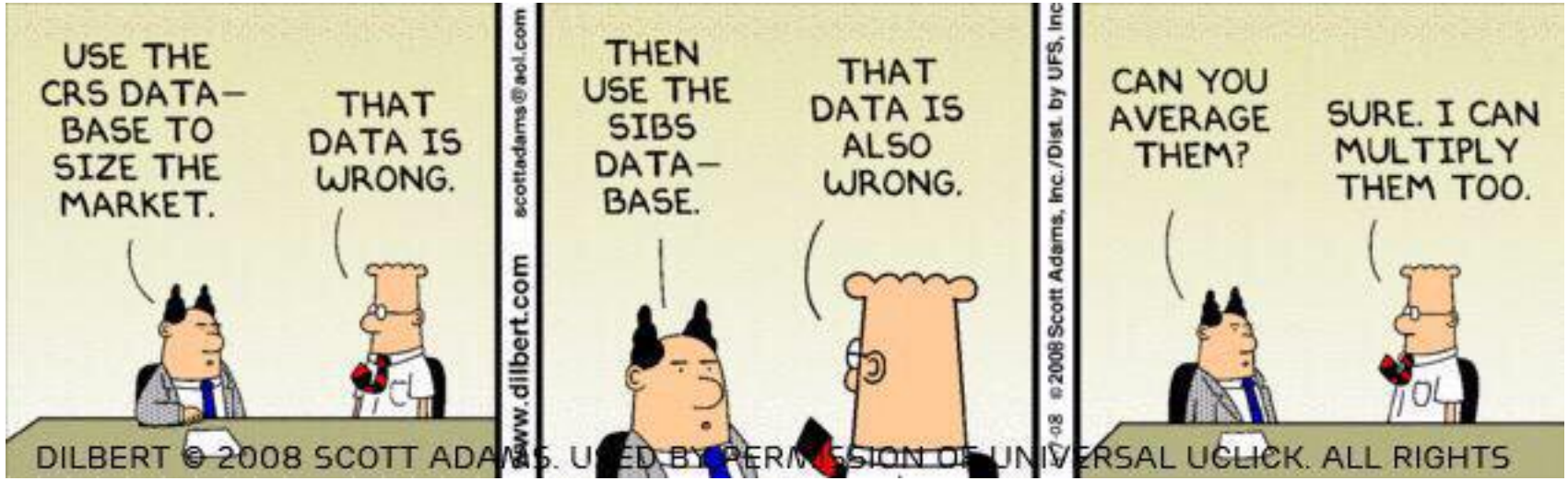
http://www.odbms.org/2015/04/data-wisdom-for-data-science/

Importance of data quality
garbage in, garbage out



advanced-outdoor.biz

No fancy algorithm can save bad data

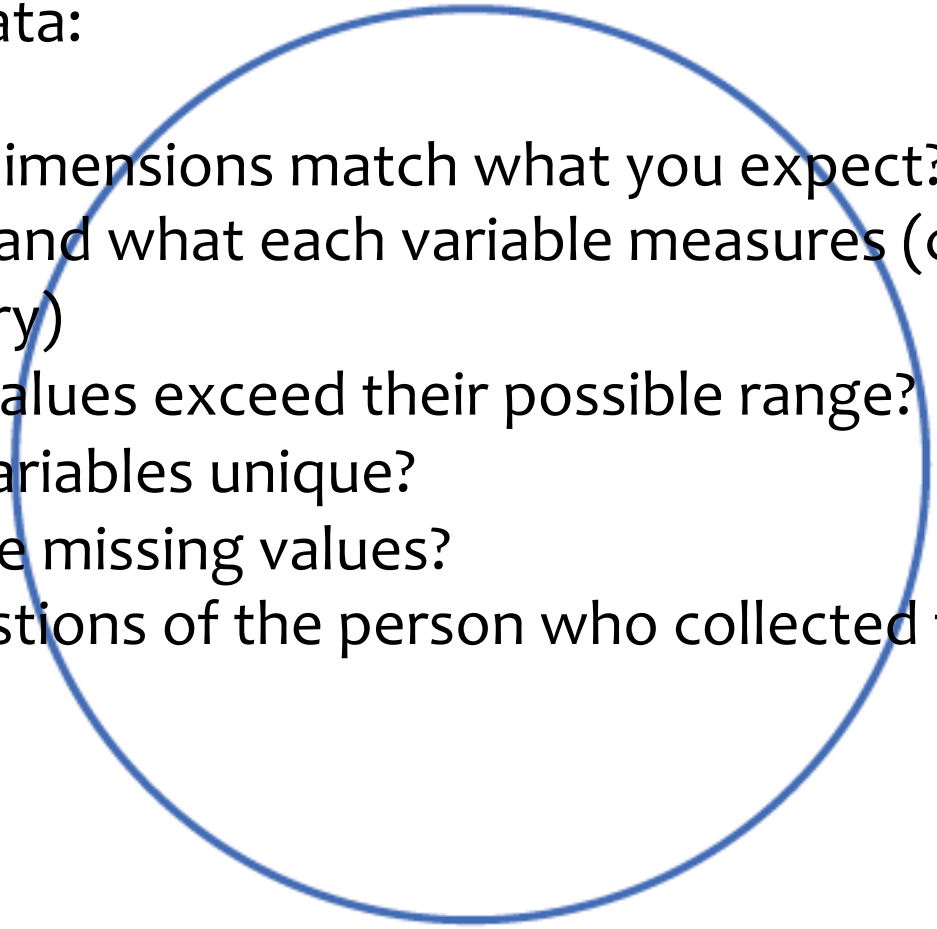


twitter.com

Data cleaning/pre-processing, data quality

Check the data:

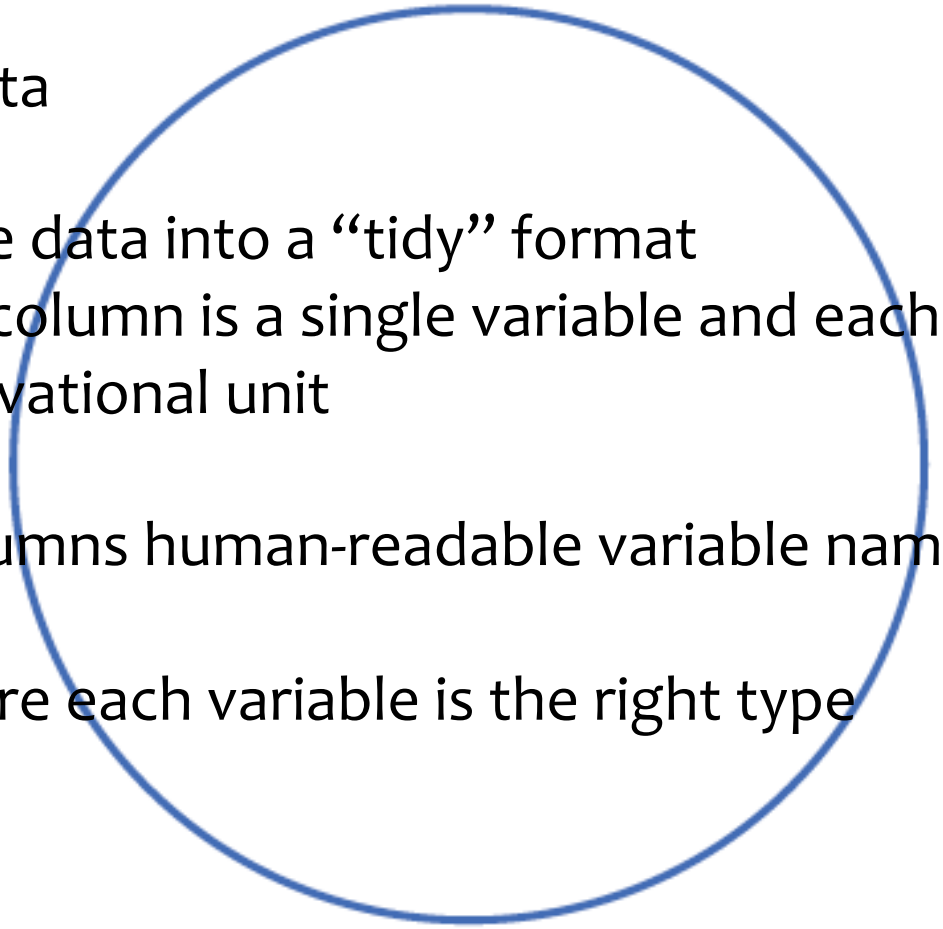
- Do the dimensions match what you expect?
- Understand what each variable measures (check data dictionary)
- Do the values exceed their possible range?
- Are ID variables unique?
- Are there missing values?
- Ask questions of the person who collected the data if you can!!!



Data cleaning/pre-processing, data quality

Clean the data

- Mold the data into a "tidy" format
  - Each column is a single variable and each row is an observational unit
- Give columns human-readable variable names
- Make sure each variable is the right type



Data cleaning/pre-processing, data quality

Ensure reproducibility

- Document and save everything you did to the data in R scripts (that can be run by another person to get to the exact same results)

- Keep a list of all decisions you made in the processing and analysis pipeline
  - E.g. how did you deal with missing values?
  - E.g. why did you remove data before July 2014?

UCD-UCB project team on prediction of surgery site infection from EMR

UCB Team

R. Barter
PhD student
Statistics

K. Kumbier
PhD Student
Statistics

B. Yu

UC Davis SSI Team Photos

Parul Dayal, MS, (PhD)
Graduate Student, Epidemiology
UC Davis

Prabhu Shankar, MD, MS
Clinical Informatician &
Assistant Professor
Health Informatics
School of Medicine, UC Davis

Hien H. Nguyen, MD, MAS
Health Sciences Clinical Professor
Division of Infectious Diseases
Department of Internal Medicine
UC Davis Health

Gavin Pereira,MD,FRCS(Eng)
Orthopaedic Surgeon,
Associate Professor,
Adult Reconstructive Service
UC Davis Medical Center

Zachary C (Zach) Lum, D.O.
Associate Physician
Orthopaedic Surgeon
Adult Reconstructive Service
UC Davis Medical Center

UC Davis project: predicting infection after surgery using EMR

- How would you begin?
- In what form do you expect your data?
- Are you going to run deep learning on this data? Why?
- Do want to go back to the UCD people after you have analyzed the data?

Thanks to Rebecca Barter for help with the slides

UC Davis project: predicting infection after surgery using EMR

30,000 surgeries between 2014 and 2018
<600 surgeries led to infections (<2%)

The data:

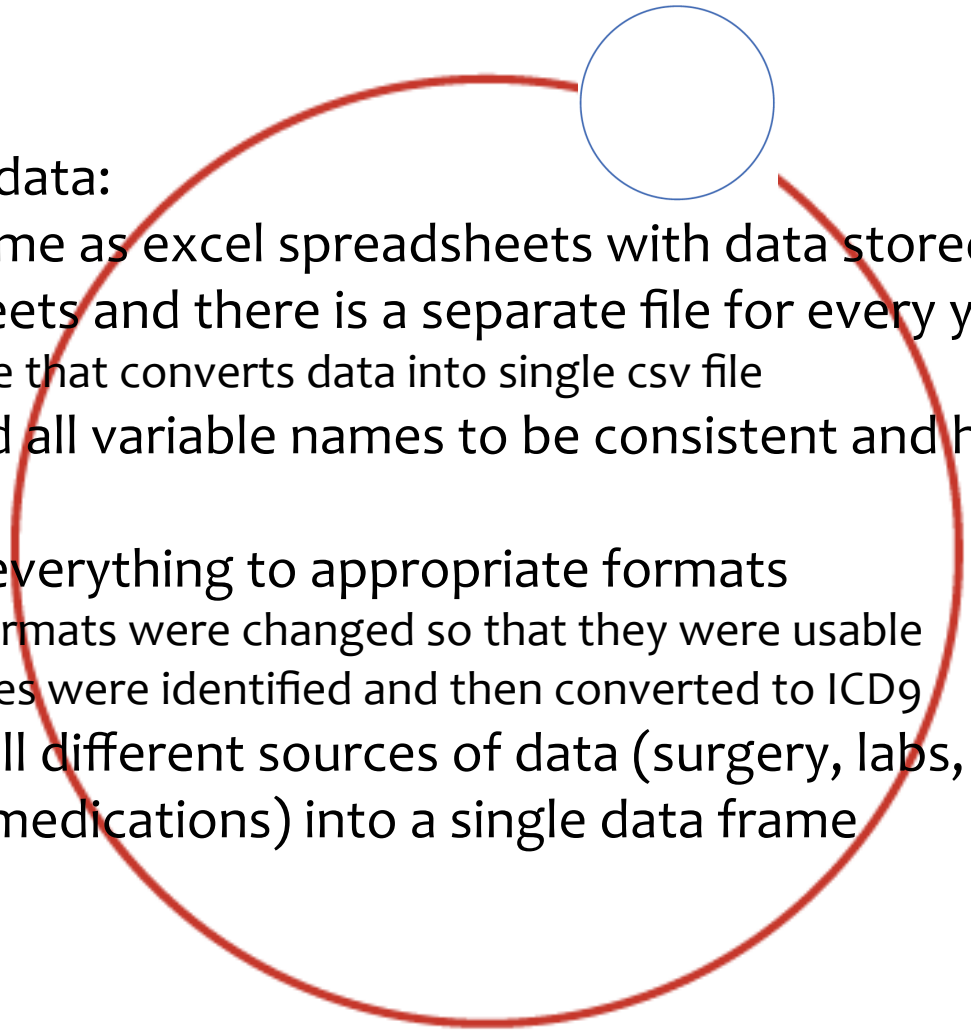
1. **Patient**: age, gender, height, weight, BMI, race, smoking status
2. **Surgery**: surgeon, date, procedure, anesthesia, wound closure type, emergency, outpatient, surgical risk, laparoscope, trauma, surgery time, department, transfusion
3. **Vitals**: BMI, temperature, weight, pulse
4. **Diagnoses**: all ICD9/ICD10 diagnoses codes
5. **Labs**: hematocrit, glucose, platelet count, sodium, red cell count, white blood cell count, hemoglobin, albumin, potassium, serum creatinine, …
6. **Medications**: all medications ordered

UC Davis project: predicting infection after surgery using EMR

Checking the data:

- The number of patients matched what was expected
- Reading the definitions of hundreds of variables led to many questions about what was being measured
- The date range didn't match what we expected (we had more years than the study period for some datasets)
- 20% of the surgery times were missing
  - Digging deeper we found that surgery times were not recorded prior to July 2014 due to changing to a new EMR system
- ICD9 and ICD10 diagnoses codes were all muddled together
- There were several supposedly unique identifiers that can match patient data to their lab data and diagnosis data etc, but not all IDs appear in all files

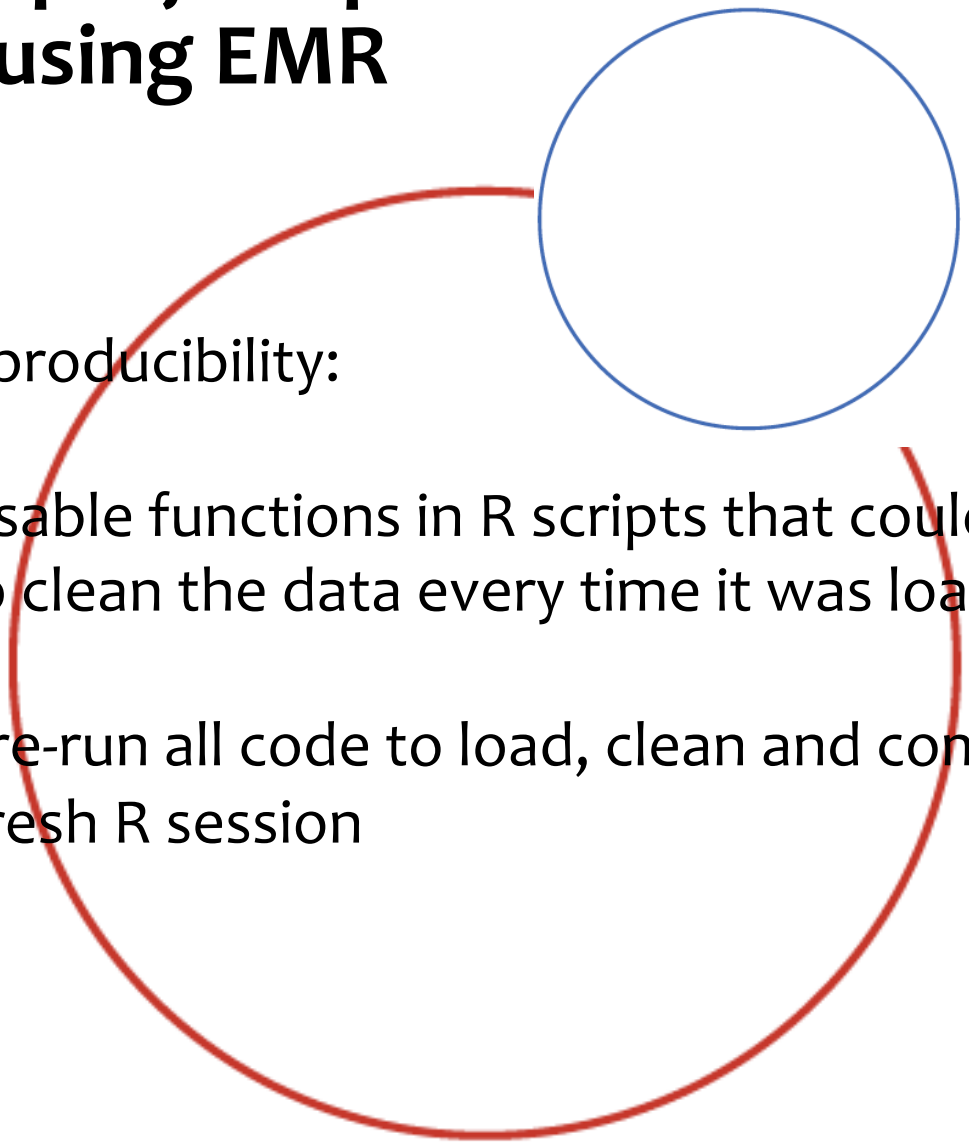
UC Davis project: predicting infection after surgery using EMR



Cleaning the data:

- The data came as excel spreadsheets with data stored across multiple sheets and there is a separate file for every year of data
  - Write code that converts data into single csv file
- We changed all variable names to be consistent and human-readable
- Converted everything to appropriate formats
  - All date formats were changed so that they were usable
  - ICD10 codes were identified and then converted to ICD9
- Combined all different sources of data (surgery, labs, vitals, diagnoses, medications) into a single data frame

UC Davis project: predicting infection after surgery using EMR



Ensuring reproducibility:

- Wrote reusable functions in R scripts that could be sourced to clean the data every time it was loaded in
- Regularly re-run all code to load, clean and combine the data in a fresh R session

UC Davis project: predicting infection after surgery using EMR

- “My questions would often lead to Parul and Prabhu literally collecting more data to answer them (e.g. they obtained data on each surgeon's age after we discussed whether or not simply using a surgeon id would be useful)

- Prabhu and Parul were in direct communication with the people who manage the EMRs at UC Davis as well as surgeons and infection experts within the hospital who had a lot of intuition about what kinds of things lead to infections and what is relevant and what is not.Prabhu set up a meeting where we all met in person for a 3-hour period and listened to each party's perspective, and we asked them all a lot of questions that we had, such as how long do they follow up with patients to see if they developed an infection (and what are the real world barriers to doing so)? what kind of things are we *not* capturing in the data that might be relevant for infections (they said things like the number of people in the surgery room, whether or not the patient takes care of their wound after surgery, etc)”

UC Davis project: predicting infection after surgery using EMR

- “One surgeon and an infection expert occasionally attend our fortnightly zoom meetings and clear up any misconceptions we have (and they get to learn about data and modeling too!)

- They have been very useful in making sure that we are developing a model that will actually be useful in practice!

- For reproducibility we use github, for data cleaning and exploration we use R, and the data lives in Box. For modeling, we will most likely be using python (since Python has more extensive deep learning capabilities) -- I'm working on this now.”

Live data vs. dead data

- Our project uses live data, which is defined as data that can be enriched by more data and human knowledge and with human domain experts interests

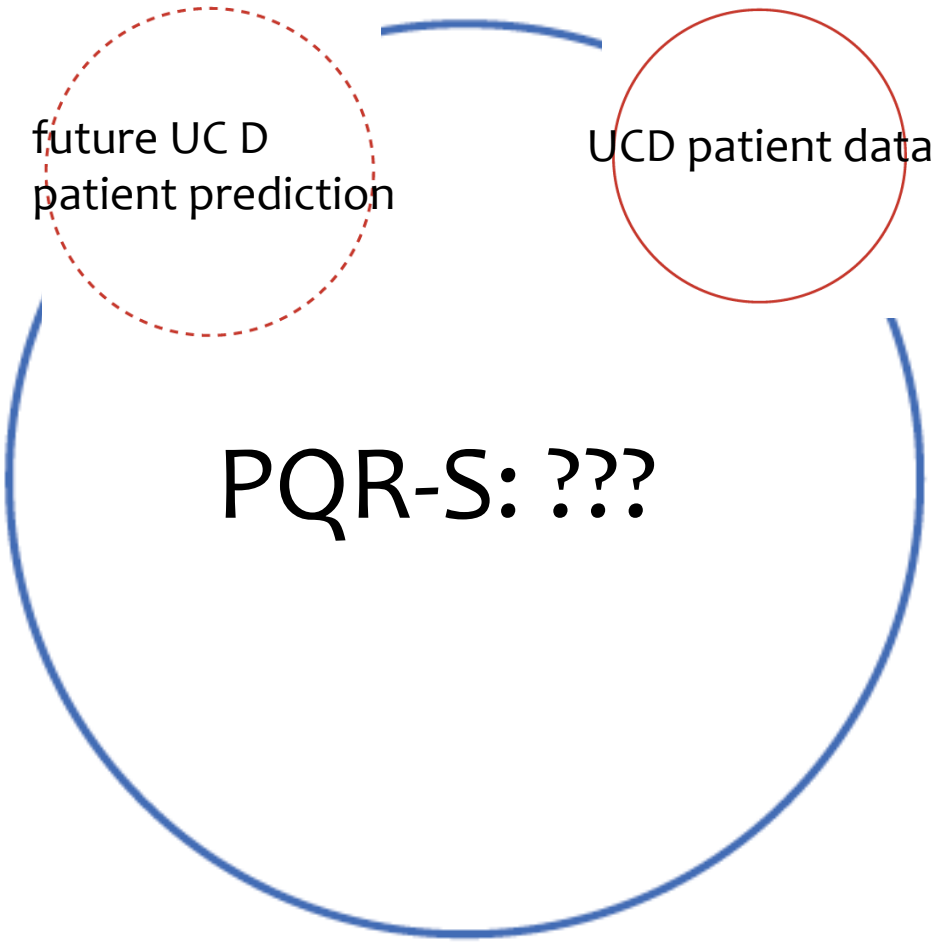
- Dead data: from the web with unknown data collection process and/or purpose, no interests from domain experts on its analysis

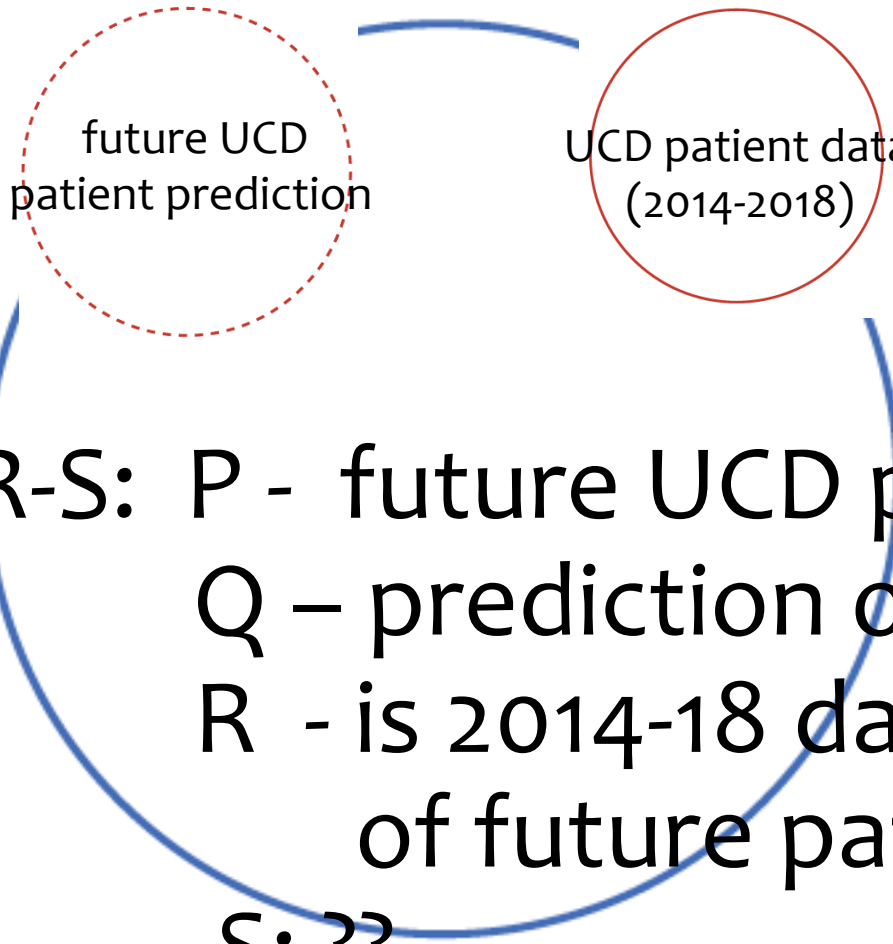
The above definitions were inspired by G. Gelman's distinction between "live problem" and mere "real data". He said in an email

"It's not enough for the data to be `real'; the data also should connect to some live question of interest."

Read his blog at

https://statmodeling.stat.columbia.edu/2009/07/23/that_modeling_f/


future UC D
patient prediction
UCD patient data
PQR-S: ???



PQR-S: P - future UCD patients (?)

Q – prediction of infection

R - is 2014-18 data rep.
of future patients?

S: ??

Reading assignments

- Review of pre-requisites
- Reading of James et al book chapter on cross validation