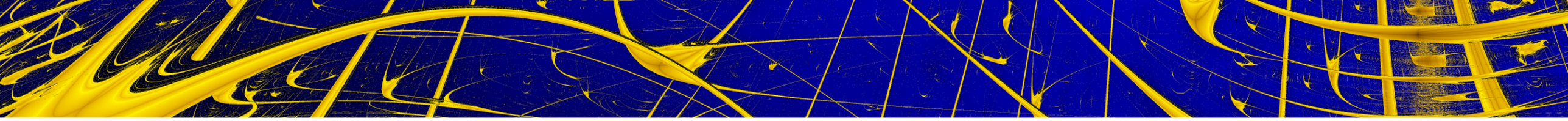


Introduction to Data Science

ESC 403

Lecture 2

Prof. Dr. Robert Feldmann
Department of Astrophysics
robert.feldmann@uzh.ch

Logistics

Course website: OLAT

lms.uzh.ch olat.uzh.ch

- Wiki page
- Forum
- Lecture slides, Exercise sheets
- Dropbox to upload exercises & project proposals

Group Project

- find group members (e.g., OLAT Forum)
- think about a project
- write a proposal (see OLAT Wiki for guidance)
- Proposal deadline **March 28**
- Project deadline May 17

Exercises

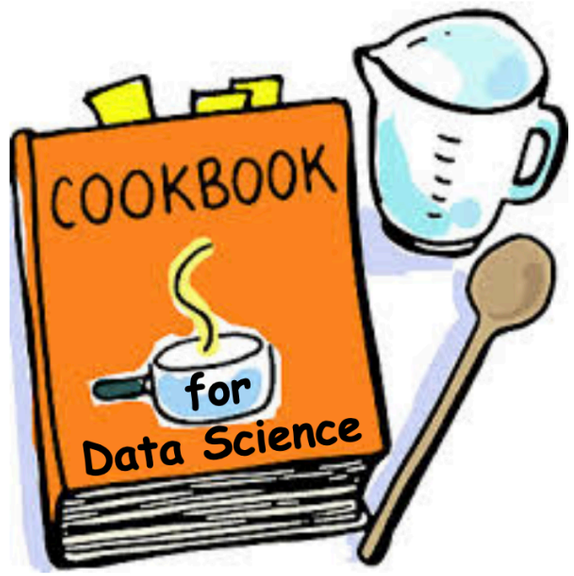
- handed out today, return Tuesday next week

Exam

- written exam, pen & paper style
- June 7 (Fri), 10 am — 11.30 am, room: Y15-G-40

Today's lecture

- Cook-book for a data scientist: what to do? which order? what to avoid?
  - What types of data are out there?
  - What is data wrangling?
  - What is the difference between inferential and predictive modeling?
  - How to deal with questions of causality?
  - What are common pitfalls?



References for todays lecture:

- Critical Questions for Big Data by Danah Boyd & Kate Crawford
- The Elements of Data Analytic Style by Jeff Leek
- An Introduction to Statistical Learning by Gareth James et al.
- The Elements of Statistical Learning by Trevor Hastie et al.

Step 1: Define your data analytic question first

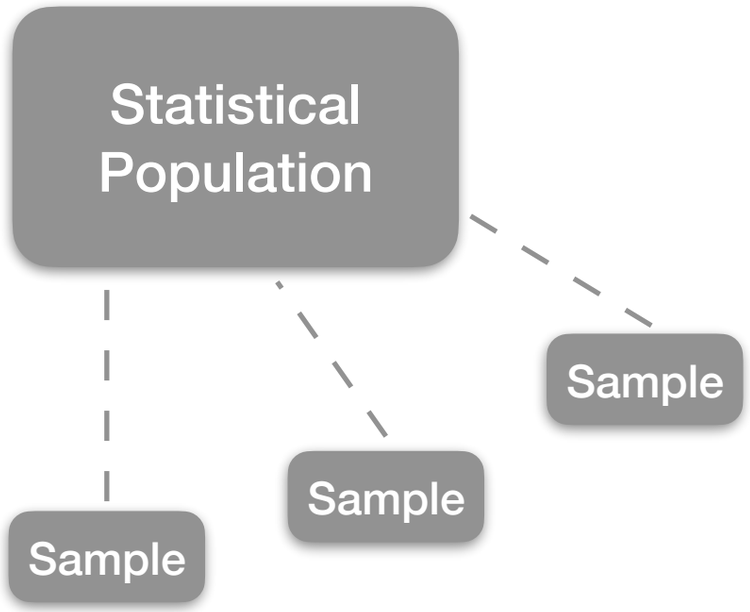
"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."

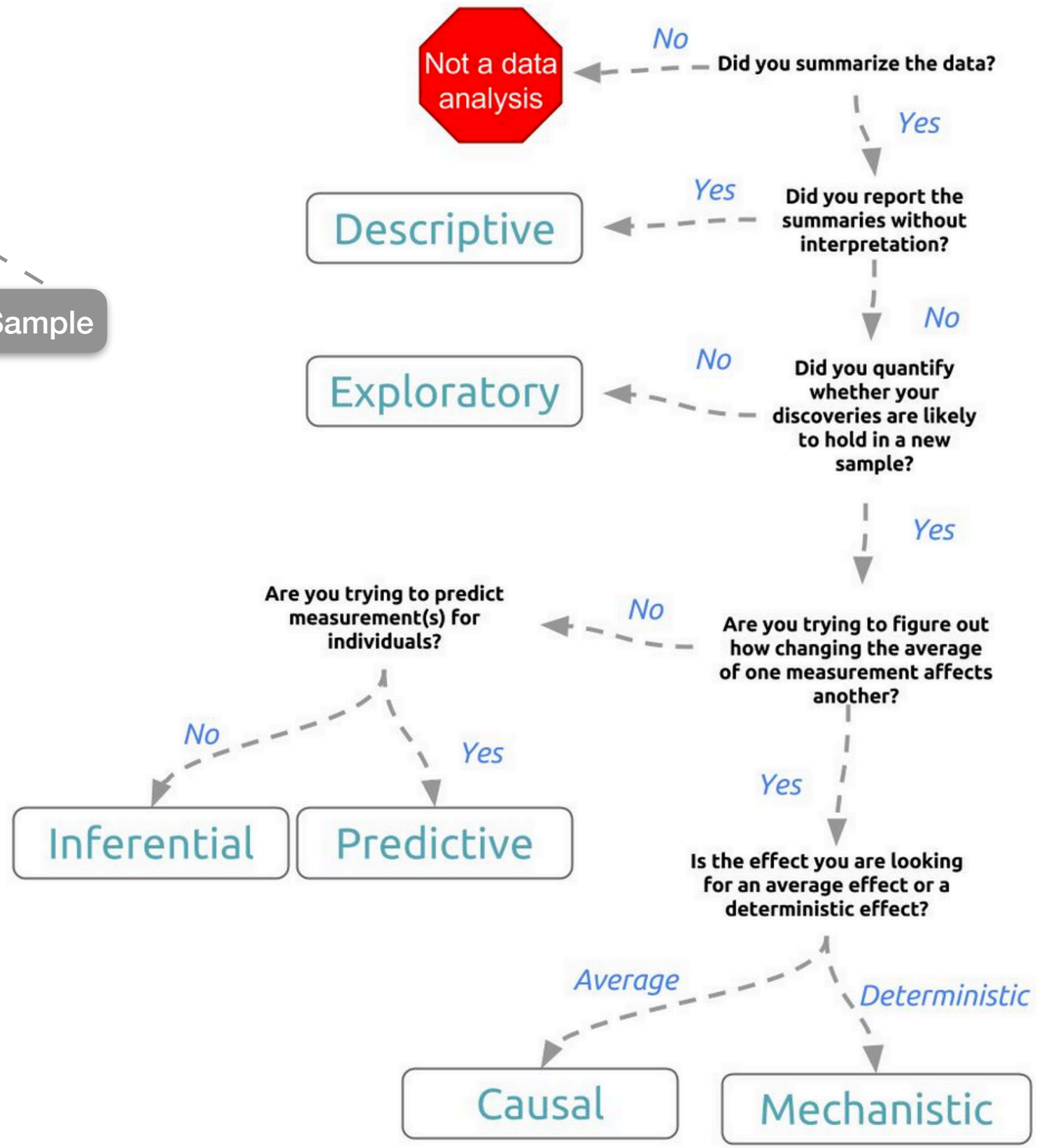
John Tukey

The question has to match the data (or vice versa)!

Understand which kind of question you ask!

Classification of Data Science Problems





Example

The safety of long-term use of bisphosphonate to treat Osteoporosis

Osteoporosis is a disease characterized by low bone mass, structural deterioration of bone tissue and an increase in the risk of fracture of the hip, spine, and wrist.

Bisphosphonates are a drug class widely used to treat and prevent osteoporotic related bone fractures.

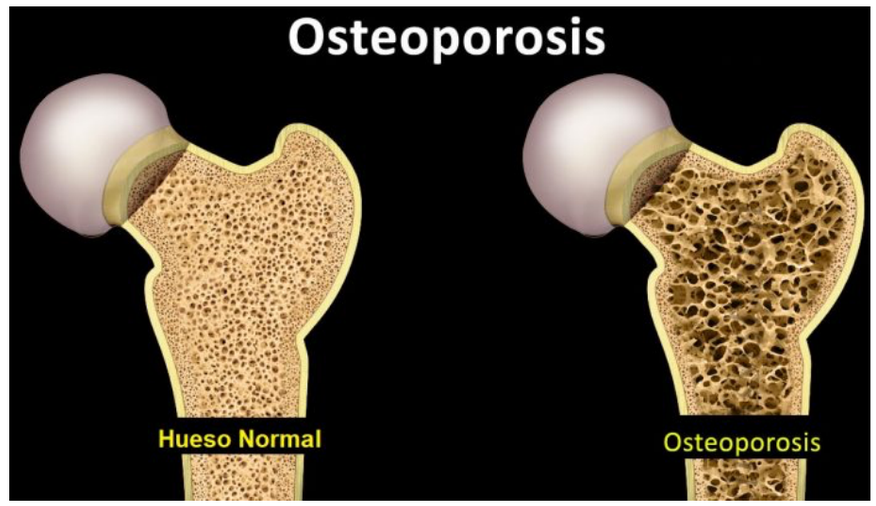
While this class of drugs has been proven effective in reducing the fracture risk, the safety of long-term bisphosphonate therapy has been questioned as adverse events continue to be reported and published.

Clinically serious but rare safety concerns with long-term use of bisphosphonates include femoral fractures, osteonecrosis of the jaw, and esophageal cancer.

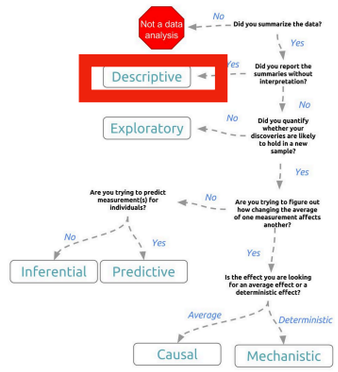
What question(s) should we ask?

Question determines other considerations

- Study design (e.g., randomization)
- Measure of safety => which endpoints?
- Strength of the effect => what sample size?
- Selection of participants?



Descriptive Statistics



Descriptive statistics is the term given to the analysis of data that helps **describe, show or summarize data in a meaningful way.**

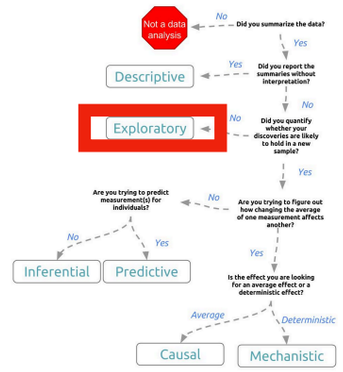
Descriptive statistics only concern with the data at hand. They usually do not allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made.

Example

Given a sample of 100 people:

- How many people in this sample take bisphosphonates?
- What is the average age of the people in this sample who take it?
- How many fractures do people in this sample have?
- How many cases of side effects are reported by people in this sample? etc

Exploratory Data Analysis (EDA)



*"Exploratory data analysis **isolates patterns and features** of the data and reveals these forcefully to the analyst."*

*"Exploratory data analysis . . . **does not need probability, significance or confidence**."*

*"Exploratory data analysis is actively incisive rather than passively descriptive, **with real emphasis on the discovery of the unexpected**."*

"If we need a short suggestion of what exploratory data analysis is, I would suggest that

1. *it is an attitude, AND*
2. *a flexibility, AND*
3. *some graph paper (or transparencies, or both)."*

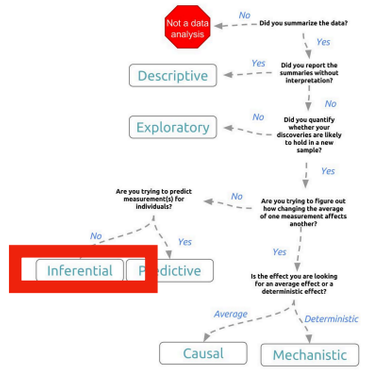
from David R. Brillinger 2011: Exploratory Data Analysis

Summarizing and visualizing data before performing formal modeling

Example

- check whether people in the sample that take bisphosphonates have higher frequency of femoral fractures or esophageal cancer than people in the sample who do not.
- look for other correlations in the data

Inferential Statistics



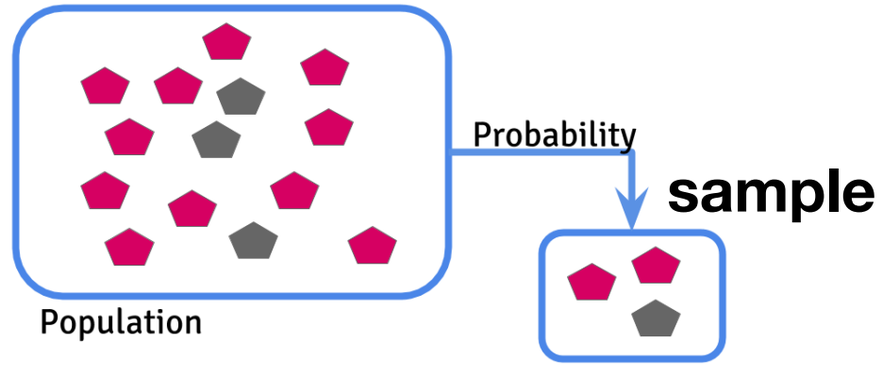
Infer the properties of a large data set ('population') from that of subset.

Inferential statistics are techniques that allow us to use these samples to **make generalizations about the population** from which the samples were drawn.

It is, therefore, important that the sample accurately represents the population. The process of achieving this is called **sampling**. Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population.

The main methods of inferential statistics are:

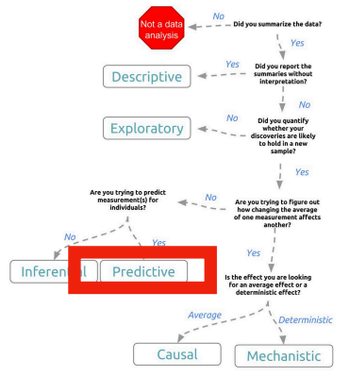
- the estimation of parameter(s) and
- testing of statistical hypothesis.



Example

- Estimate the average age of bisphosphonate users in the larger population
- Assess statistically whether correlations found in the data likely hold in the larger population
- Answer whether bisphosphonates use is correlated with higher incidence of, e.g., femoral fractures among the larger population

Prediction



Predict 'outcome' properties of data points from their 'feature' properties.

$$Y = f(X) + \epsilon$$ ϵ is random error (zero mean)

- given set of data points $\{(X_i, Y_i)\}_i$
- want F (estimation of f) such that $F(X)=\hat{Y} \approx Y$ for new data point (X,Y)
- often F treated as black box

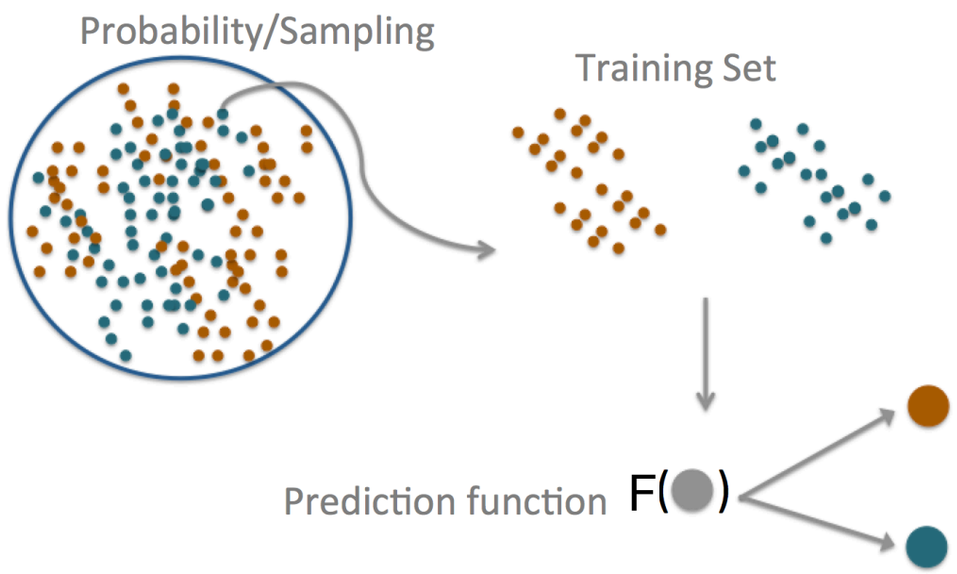
How good is the approximation F?

What is $E[(Y - \hat{Y})^2]$ (optimally ~0)

Residual: $Y - \hat{Y}$

What is bias $E[Y - \hat{Y}]$? (= 0 if unbiased prediction)

How do we get F?

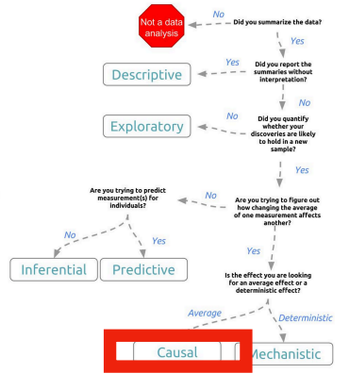


- take training sample from population
- train algorithm ➜ F (e.g., machine learning)
- validate F on other sample(s)

Example

- predict the number fractures of a patient based on age, sex, bisphosphonate use etc.

Causal Inference



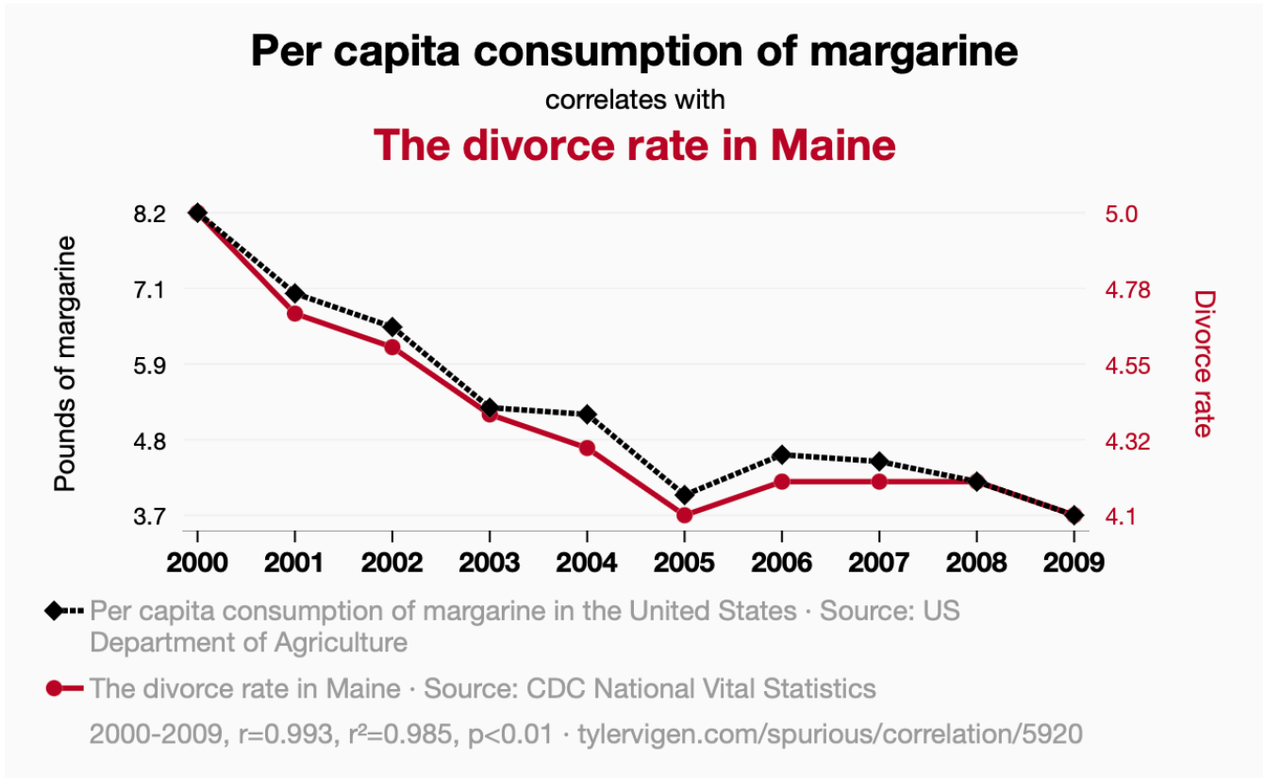
Want to know whether feature X *causes* outcome Y (in an average sense)

"*The world is richer in associations than meanings, and it is the part of wisdom to differentiate the two.*"

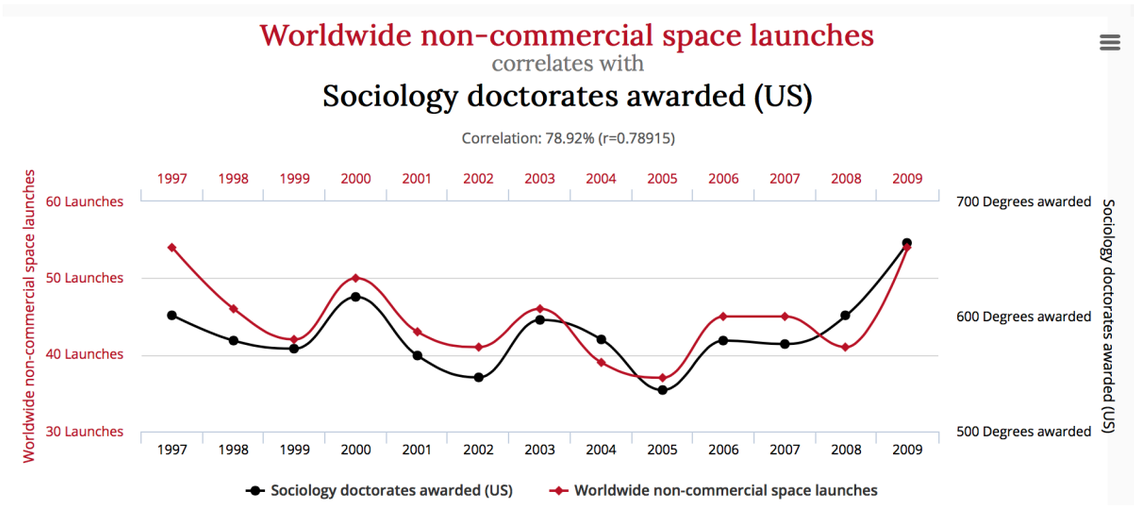
John Barth, novelist

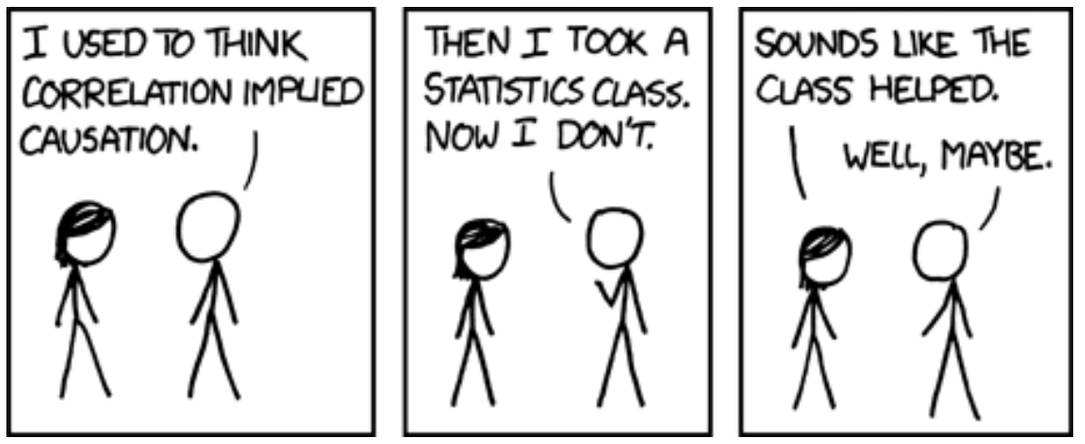
Complicated because:

Correlation does not imply causation!



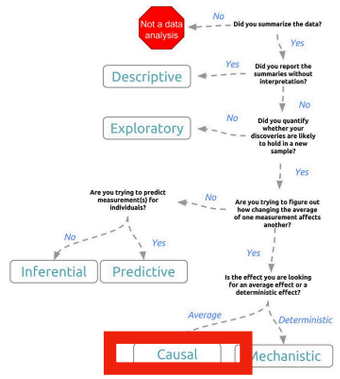
http://tylervigen.com





https://xkcd.com/552

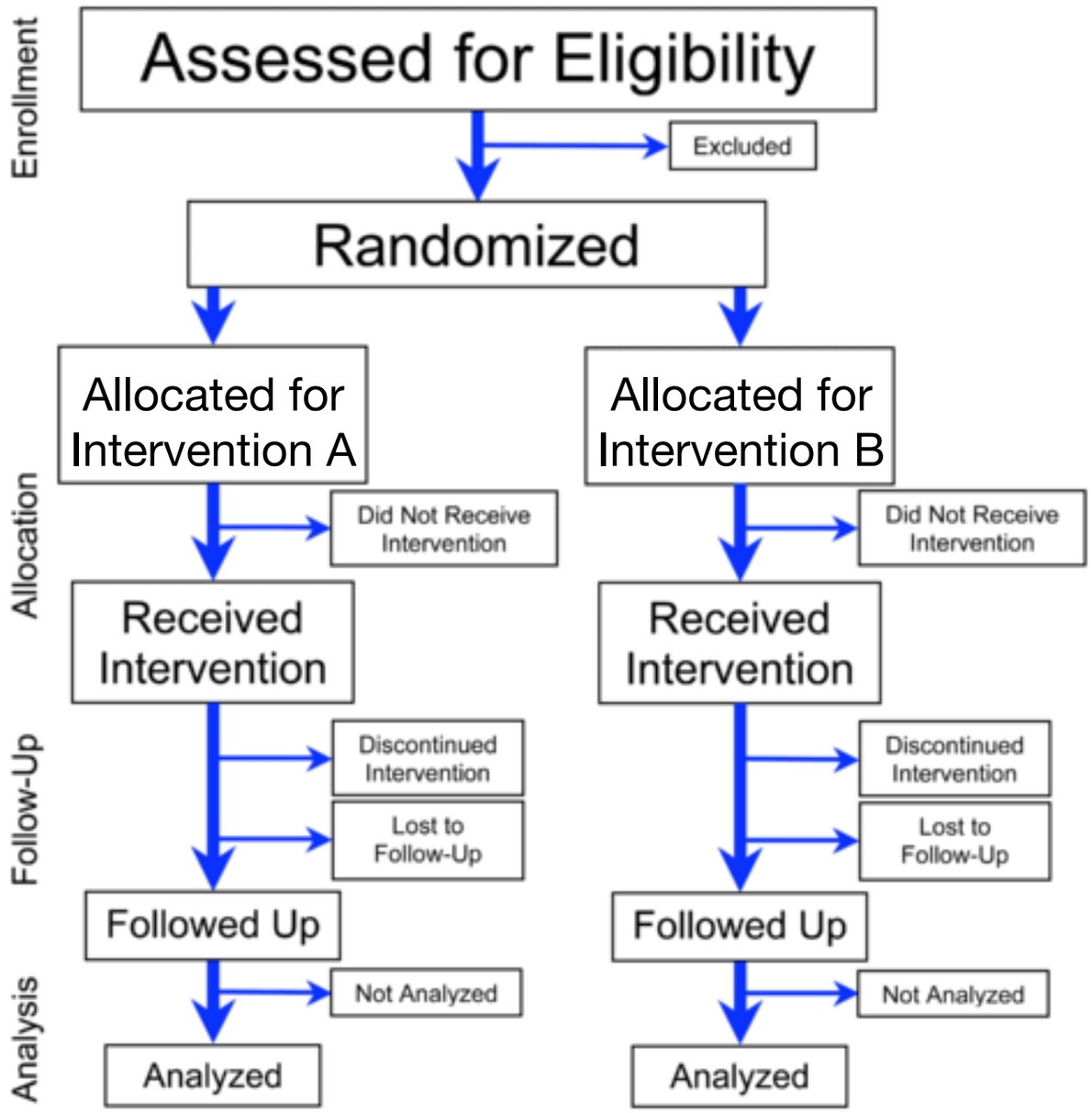
Causal Inference



- Typically only possible under well-controlled circumstances
- Study design critical!
- No 100% proof possible because influencing factor could have been missed

Randomized Control Trial:

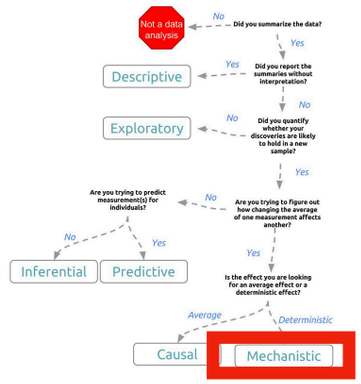
- Two (~identical) groups:
  one treatment, one placebo
- Eliminates selection bias and confounding in treatment assignment
- compare outcome for the two groups and assess differences statistically



Example

- Does long-term bisphosphonate use **cause** a higher risk of femoral fracture or esophageal cancer?

Mechanistic Analysis

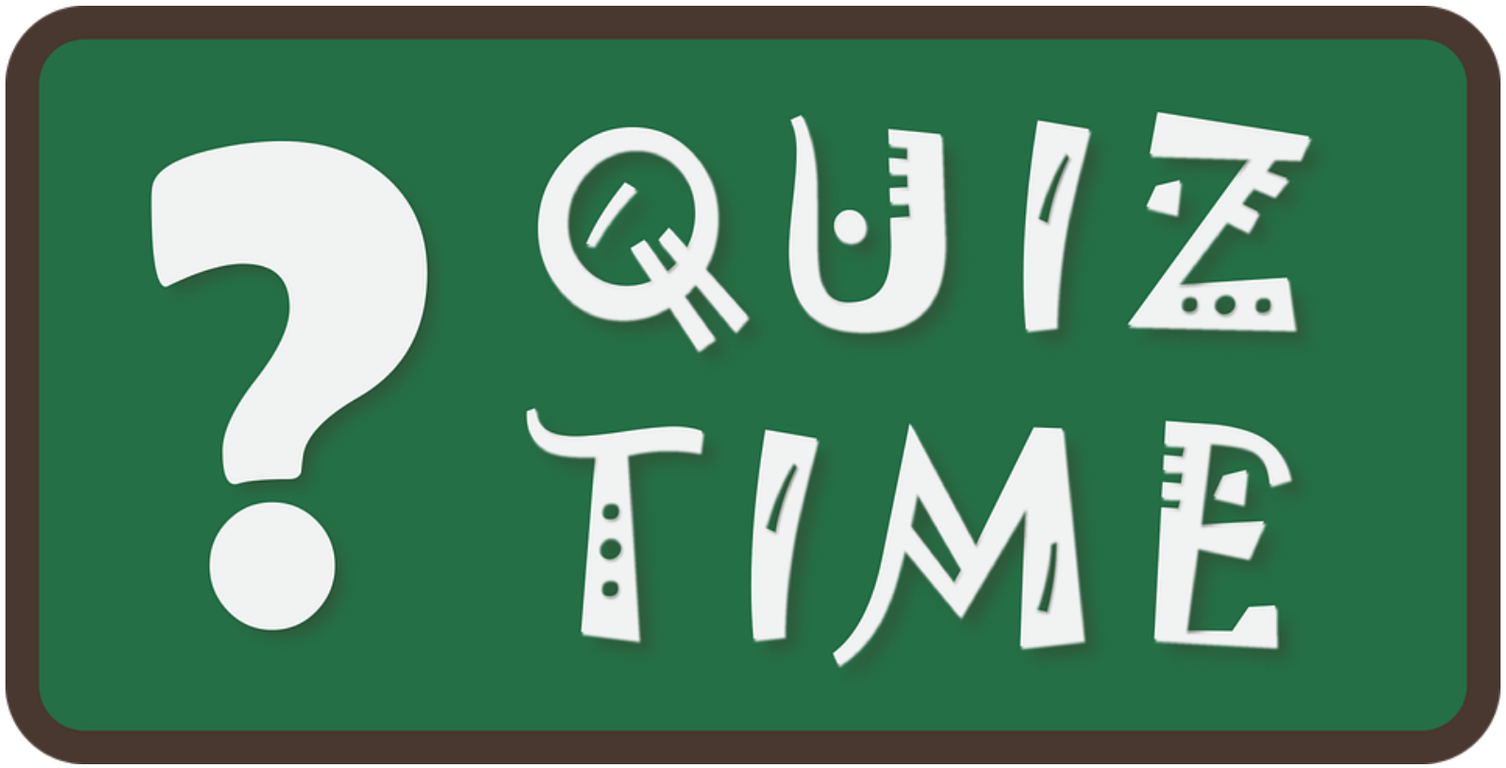


We want to show that changing on parameter always leads to a specific, deterministic behavior in another.

- Requires perfectly controlled experiments
- Desired in engineering, physics, chemistry, computational modeling etc
- Almost never possible in social sciences, health science etc.

Example

- study the biological effect of bisphosphonates
- How does it cause esophageal cancer?


QUIZ
TIME

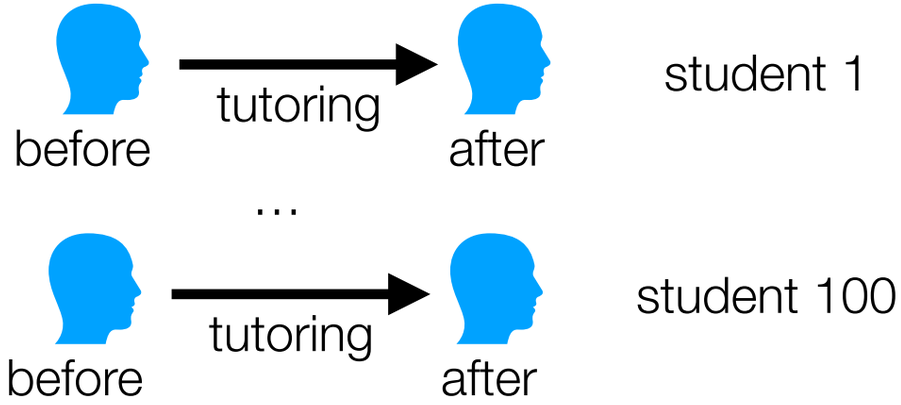
Does tutoring significantly increases student performance?

Study design:

- 8 week tutoring component, 100 students
- assume student performance is normally distributed
- measure their performance before the tutoring starts and after

Test:

- paired t-test to evaluate whether any difference seen is statistically significant



Result:

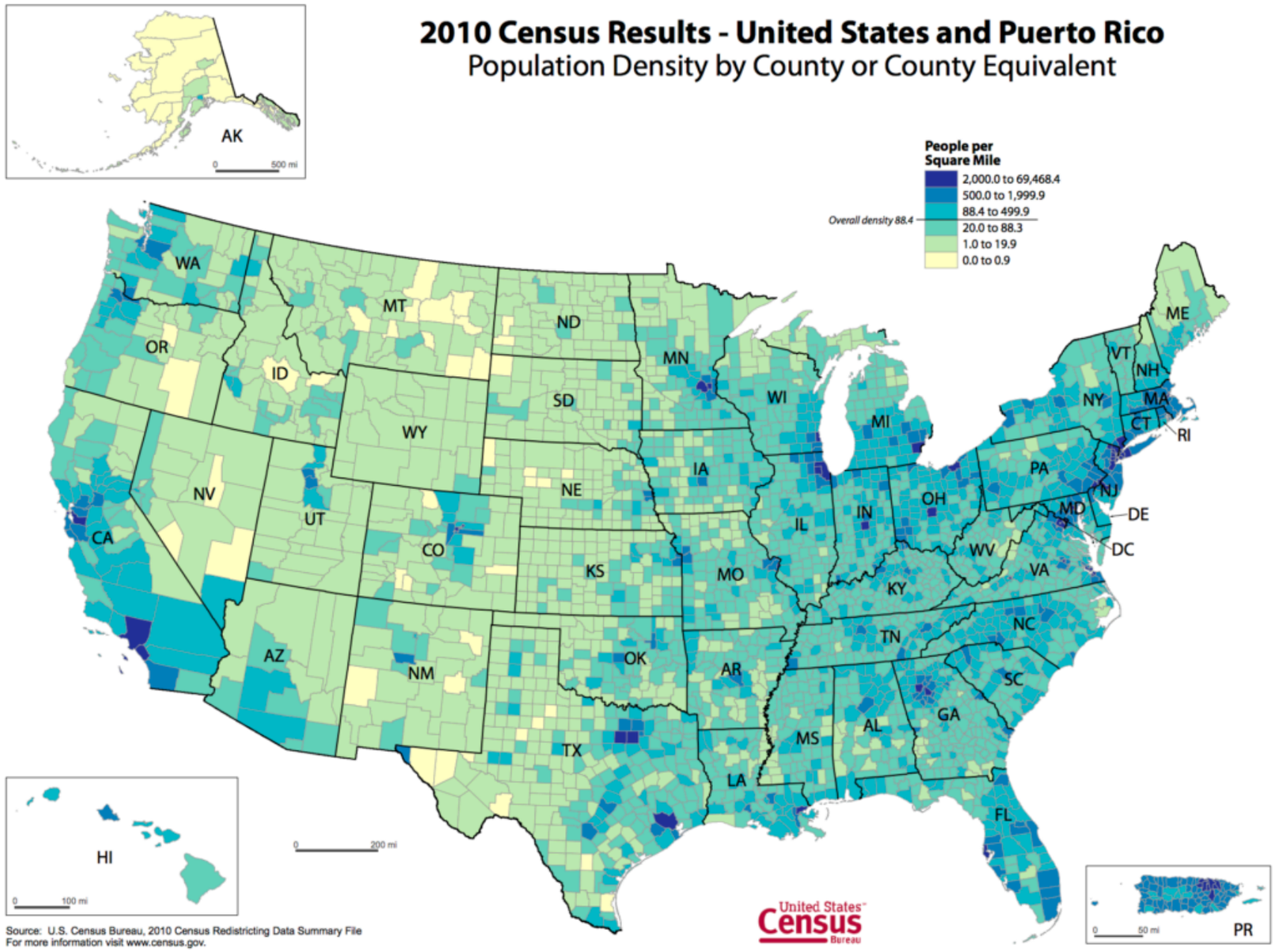
- if significant: tutoring is correlated with increased student performance

Other examples:

- 60% of the voters in the state favor proposition A, with a margin of plus or minus three percentage points.
- Brand A pain medicine brings noticeable relief significantly faster than brand B medicine.
- The 95% confidence interval for the population mean age is 35 to 45.

US Census

- ask ~every resident
- collect various data: age, sex, location, job, income, ethnicity etc.
- interpretation/use left to politics



Which team is favored of winning a Football tournament?

FiveThirtyEight

http://fivethirtyeight.com

Probabilities computed based on:

- domestic team strength: using e.g., goals scored, shots
- league strength: using e.g., league market value
- combine to get overall team strength
- calculate win/loss/draw probabilities for future matches
- simulate the season thousands of time to estimate winning probability

BUNDESLIGA
2017-18 season

Winner: Bayern Munich
Dortmund & S04: qualified for CL

TEAM	SPI RATING	QUALIFY FOR UCL	WIN LEAGUE
Bayern Munich 60 pts	91.8	>99%	>99%
Dortmund 40 pts	80.8	84%	<1%
Schalke 04 40 pts	75.3	68%	<1%

See all for Bundesliga

CHAMPIONS LEAGUE
2017-18 season

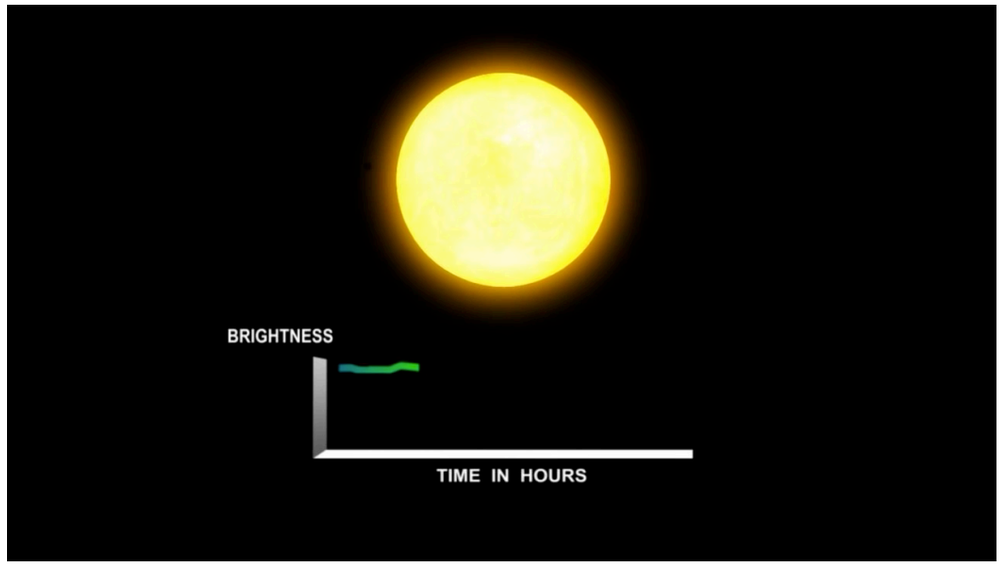
Winner: Real Madrid
Barcelona: lost Quarterfinals
Bayern Munich: lost Semifinals
Man. City: lost Quarterfinals

TEAM	SPI RATING	MAKE SEMIS	WIN FINAL
Barcelona	94.7	64%	24%
Bayern Munich	91.8	61%	17%
Man. City	91.2	60%	16%

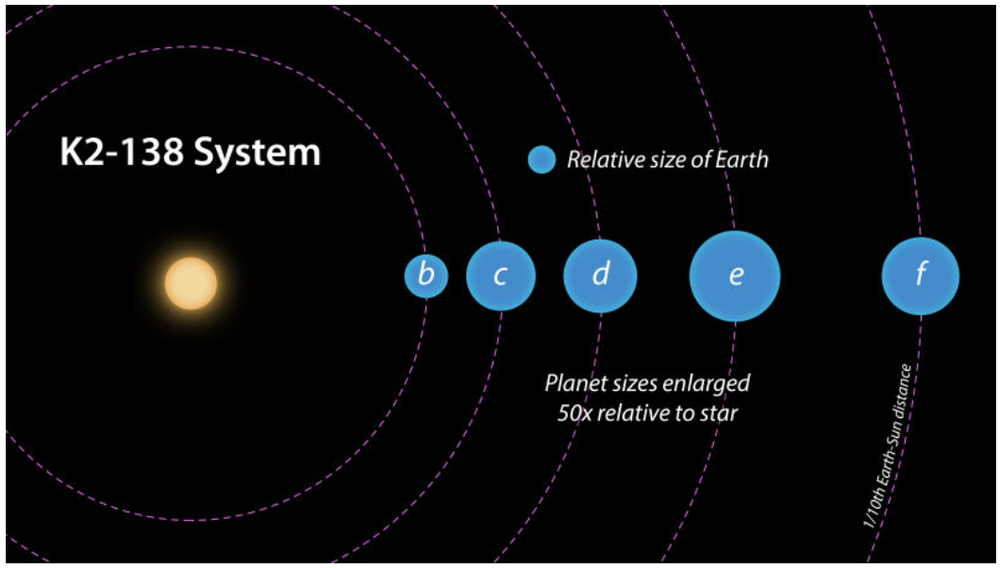
See all for Champions League

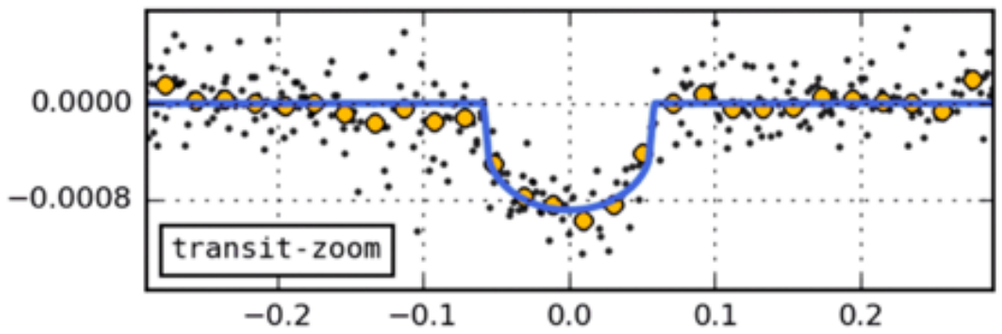
Citizen Scientists detecting planets with Kepler

- 1000's of citizen scientists sifting through Kepler data
- transit method => temporary reduction in stellar flux
- vote whether light curve contains evidence of a planet



- identified 4 planet candidates
- candidates later confirmed
- at least one more planet identified





https://www.zooniverse.org/projects/ianc2/exoplanet-explorers

Landing rockets

- Landing a rocket is rather tricky
- Main concern is the amount of fuel necessary
- many variables: speed, altitude, angle of attack etc.
- E.g., How does changing the angle of attack affect the fuel required for the landing; At what time should the legs be deployed?

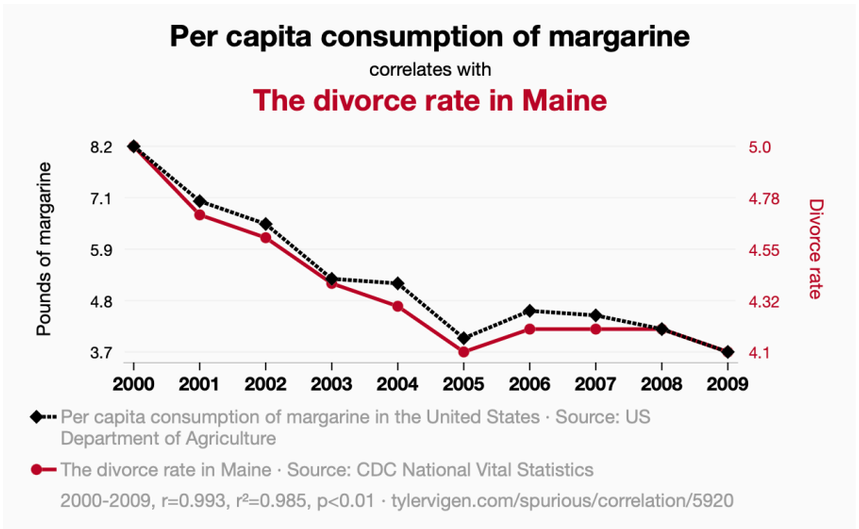
Should the government fund job training programs?

- Existing programs seem to help unemployed & underemployed to find better jobs
- Does the training really help or are there confounders?
- Maybe only motivated people go to job training?

Common mistakes

- Interpreting an inferential analysis as causal
  - avoid words: causing, affects, …
    unless it's causal inference!
  - can lead to **causation creep**:
    interpretation of results that suggests causation

Remember?



Researchers followed 397 children from pregnancy through their first year of life, and found that those living with dogs developed 31 percent fewer respiratory tract symptoms or infections, 44 percent fewer ear infections and received 29 percent fewer antibiotic prescriptions.

The researchers acknowledged that they couldn't account for all such factors [other than living with dogs that can explain the finding], and noted that they found a correlation, not a cause-and-effect relationship.

Headline *For healthier kids, get a cat or dog, study suggests.*

http://junkcharts.typepad.com/numbersruleyourworld/2012/07/the-causation-creep.html

Common mistakes

- Be careful with data used in EDA for subsequent analyses (Overfitting/Backtesting)
    - Do not use the same data for both model building and testing!
- Interpreting an exploratory analysis as inferential (Data Dredging/p-hacking)
    - Do not fit large number of models to a sample and try to infer about the general population

    “If you torture the data enough, nature will always confess.”

    Ronald Coase, economist

    - ‘p-hacking’

      The Inside Story Of How An Ivy League Food Scientist Turned Shoddy Data Into Viral Studies
- Interpreting a descriptive analysis as predictive/inferential
    - Avoid anything but descriptive if sample size is tiny!

Step 2: Tidy the data

The data set should contain at least the following

1. the raw data
2. a tidy data set ready for analysis
3. a code book describing each variable and its values in the tidy data set
4. an explicit description how the tidy data set is created from the raw data (e.g., a python script)

Step 2: Tidy the data

- Raw data
  - could be binary format, unformated Excel file, JSON data, hand-written log-book, etc.
  - No manipulation of the data, no removal, not summarized!
  - Raw is relative, but try to get rawest data possible
- Tidy data
  - each variable you measure in one column
  - each different observation in one row
  - one table for each “kind” of variable
  - have ID to link multiple tables if necessary
  - use descriptive names (e.g., AgeAtDiagnosis, not ADx)
  - CSV or tab-delimited text files work great

Step 2: Tidy the data

- The code book
  - info about the variables and their units
  - info about any summary choices you have made
  - info about the study design you used
  - info about how the data was collected

- An explicit description of creating the tidy data set
  - important for reproducibility
  - other people should be able to recreate tidy data set from raw data
  - simplest: provide the python/R/Shell script used to create the tidy data set
  - the version of the software products you used etc.

Common mistakes

- Combining multiple variables into a single column
- Merging unrelated data into a single huge file (better use ID & sep. tables)
- Instruction list not complete, missing software version, parameters

Step 3: Checking the data

Data wrangling / Data Munging

- the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics
- the goal is to assure quality and useful data
- Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.
- basically **always required**

Decide/Check data type, look for coding errors

• Continuous	age
• Ordinal	User satisfaction
• Categorical, Dichotomous	Ethnicity, Sex, Handedness
• Missing	‘NA’ or ‘.’
• Censored (i.e., only range known)	age < 18

Missing data and censored data are treated differently in statistical tests!

Check the units make sense

Look for label switching & logical inconsistencies
(e.g., one person’s height changes between tables)

Common mistakes

- Leaping to statistical modeling without proper data checking
- Wrong data coding (e.g., treat categorial variables as numerical)
- Just look at summaries, not sufficient plots
- Fail to look for outliers and missing values

Now we are ready for step 4: EDA and statistical analyses

Exploratory Data Analysis

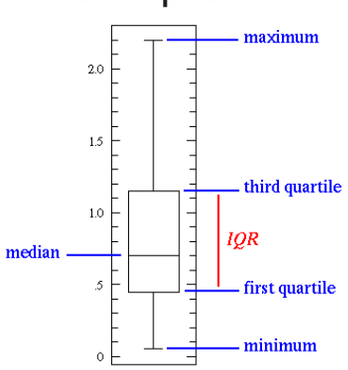
Summarizing and visualizing data before performing formal modeling

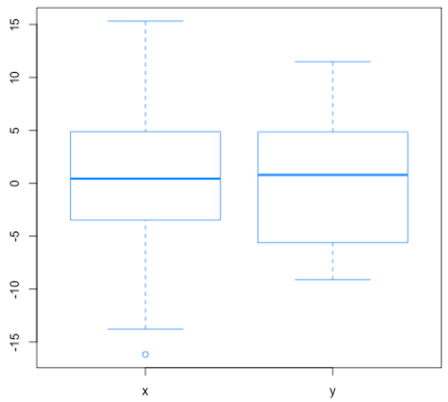
- to understand properties of the data
- to inspect qualitative features
- to discover new patterns or associations

Interactive analysis is best

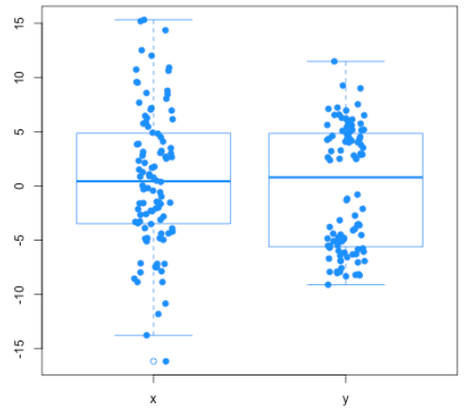
- if you have huge data sets ➜ analyze random (sub-)sample
- plot as much of the data as you can

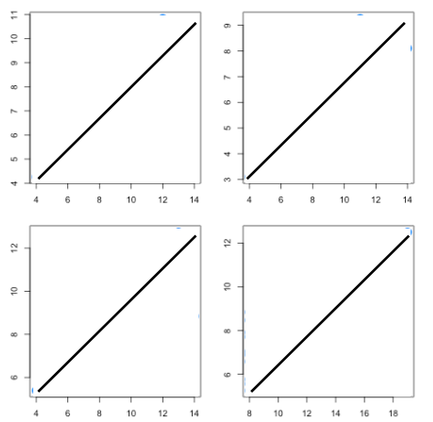
box plot



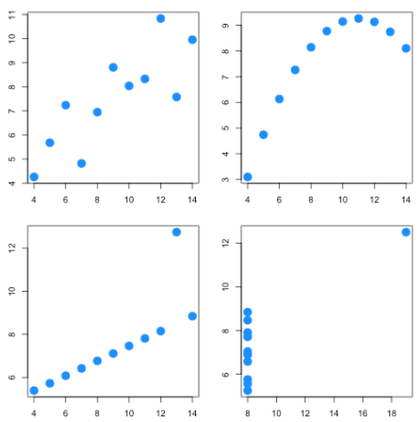


vs





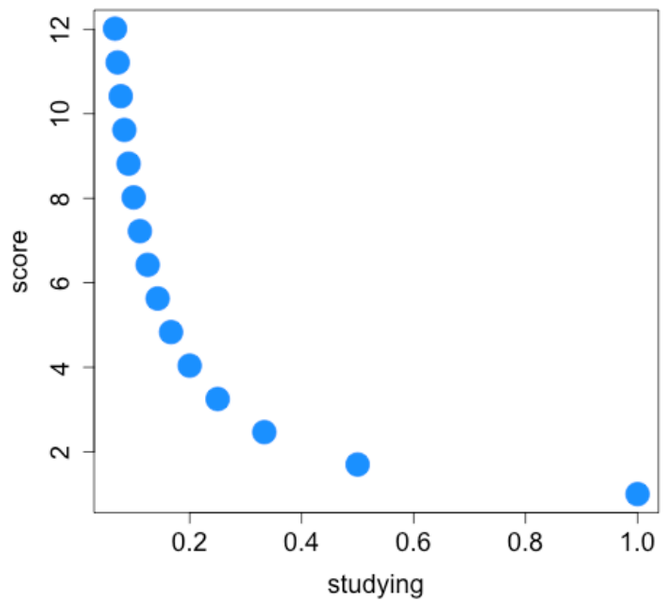
vs

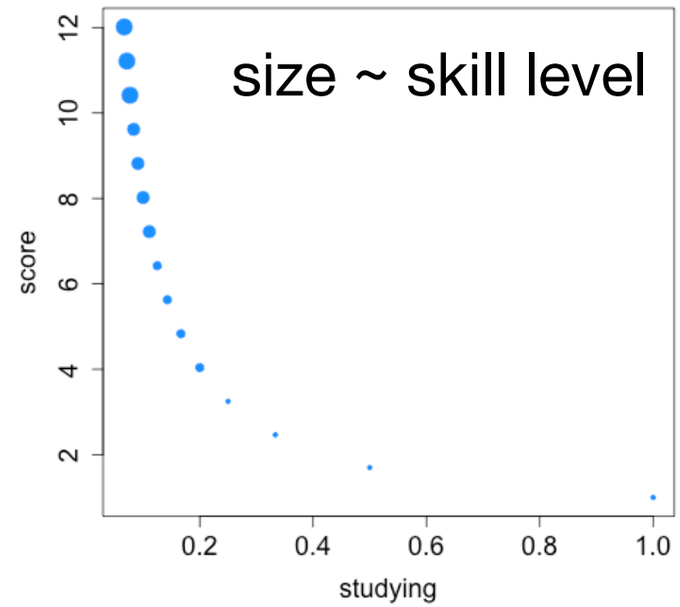


Make plots, tables etc quickly (pretty plots for published papers, not for EDA)

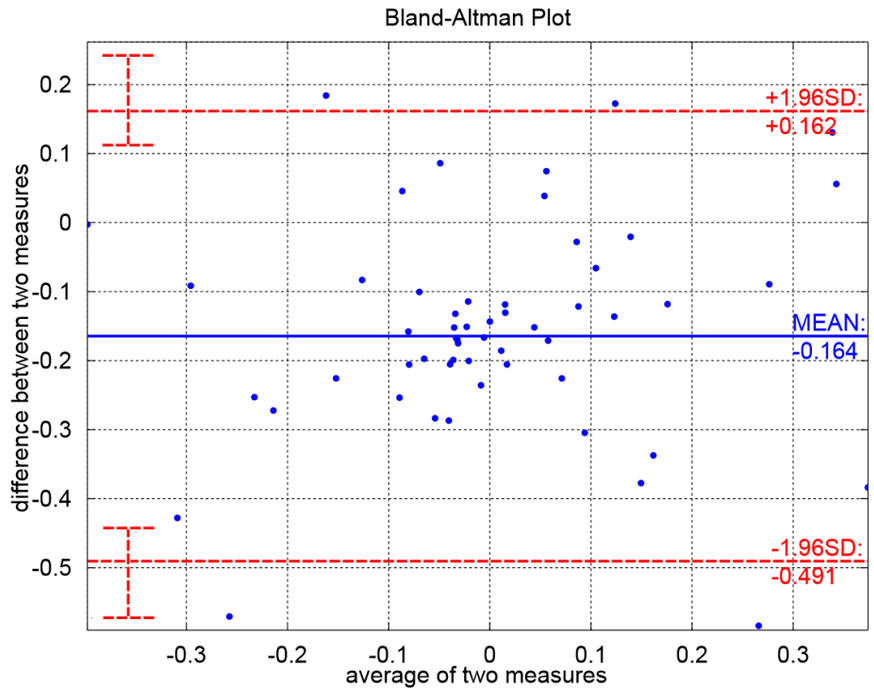
Exploratory Data Analysis

- Use colors, sizes etc to check for confounding





- If same quantity in multiple plots, use same axis limits
- Log-transform the positive data
- To compare the difference between two methods/instruments etc, consider a plot of averages vs differences



Common mistakes

- Not spending sufficient time and jumping right to statistical analysis
- Wasting time on making plots pretty (Speed is priority!!)
- Finding patterns but not exploring their origin (e.g., confounders?)
- Failing to look at patterns of missing values and their impact on conclusions (leading to bias)

Statistical modeling and inference

- Aim to perform exploratory and confirmatory analysis on separate data sets
    - Split data into two random(!) subsamples
    - perform exploratory analysis on first and confirm using second
    - 70% (exploratory) + 30% (test) are typical

- Be clear about the definition of the sample, population from which the sample is drawn, and the individual data sets

- Identify if/why sample may not be an unbiased representation of population one is interested in

- Identify potential confounders

Example: Literacy and shoe size are correlated
Why?

- Check for outliers & distribution of missing data

- Check that statistical estimates are plausible (sign and magnitude)

Statistical modeling and inference

- Be careful about very small or very large samples
  - small samples: measures of uncertainty are highly uncertain themselves
  - large samples: stat. uncertainty should be tiny, errors dominated by biases

- Know what your real sample size is
  - data may be correlated or from same source etc.

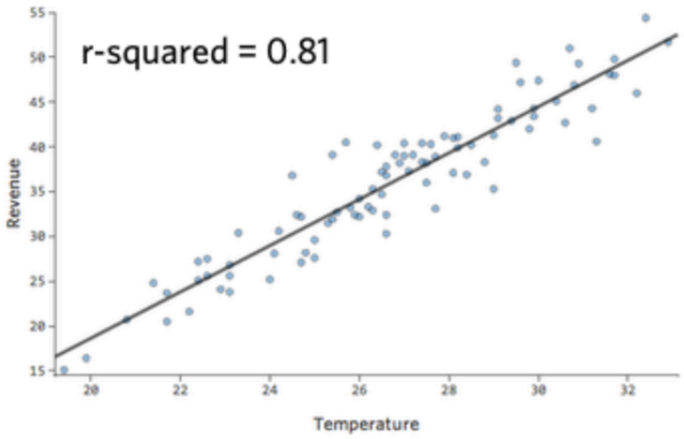
- When testing multiple theories: CORRECT FOR IT
  - single theory tested: typically check $p < \alpha = 0.05$
  - but if multiple tests: cannot test each and look at their p-values
  - one of the simplest methods: Bonferroni correction
    - if testing m theories, check for each whether $p < \alpha/m$
    - reduces the risk of incorrectly rejection the null hypothesis (type I error)
    - increases the probability of incorrectly accepting the null hypothesis (type II error)

Common mistakes

- Inference without exploration
- Not checking whether model fit describes the data well

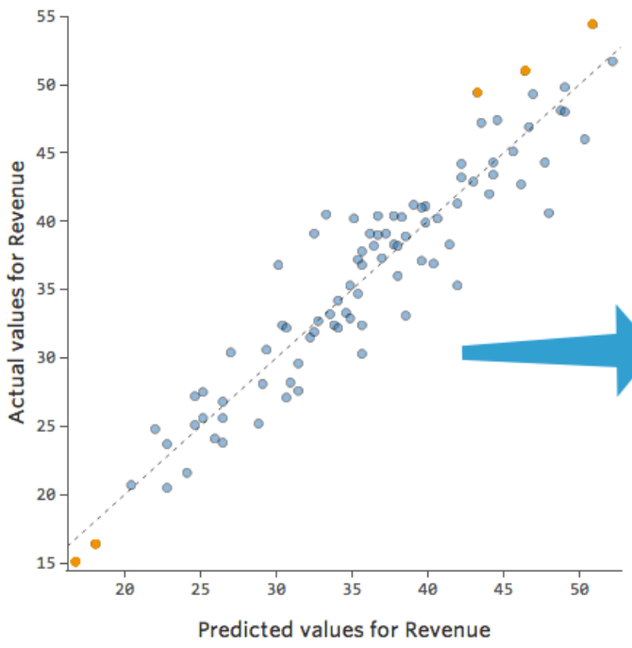
https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Lemonade stand data set

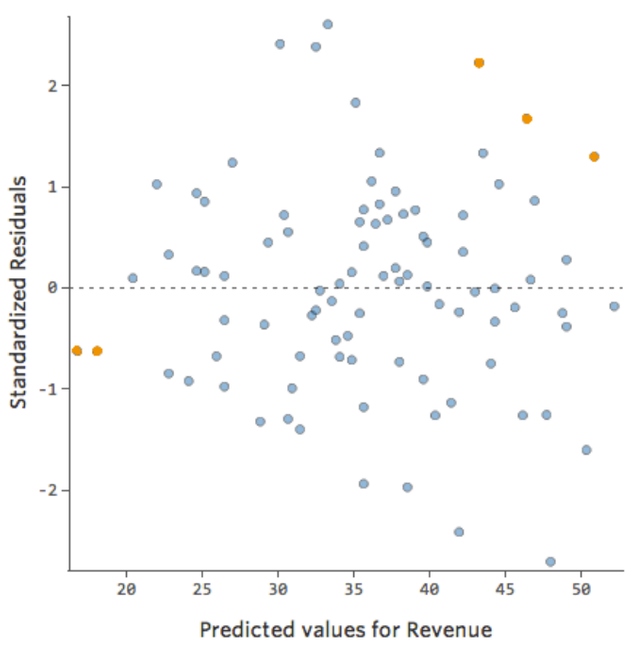


Revenue ~ Temperature

Predicted vs actual value



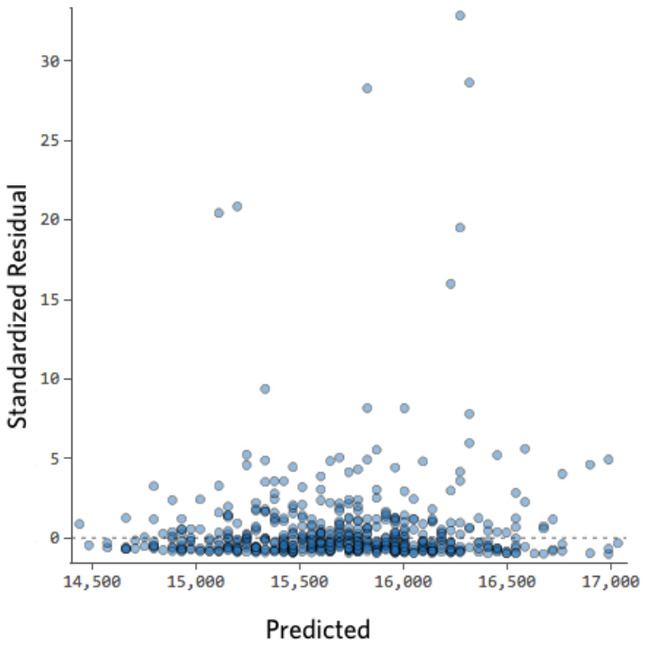
Predicted value vs residual



looks fine!

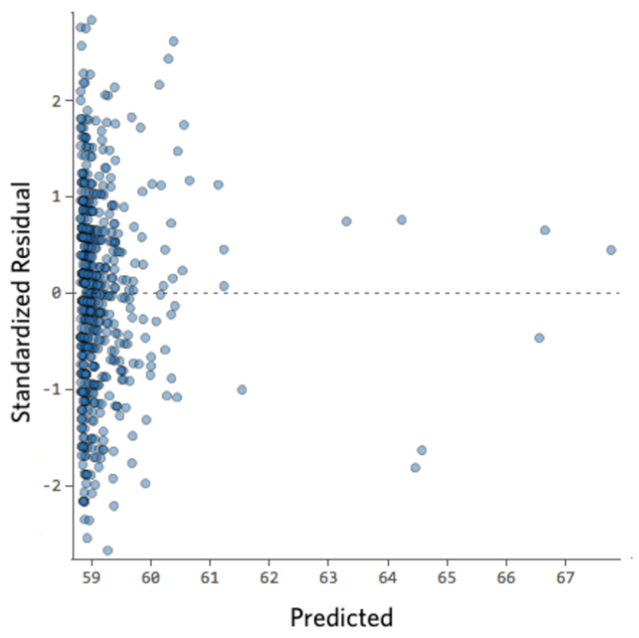
Common mistakes

bad!



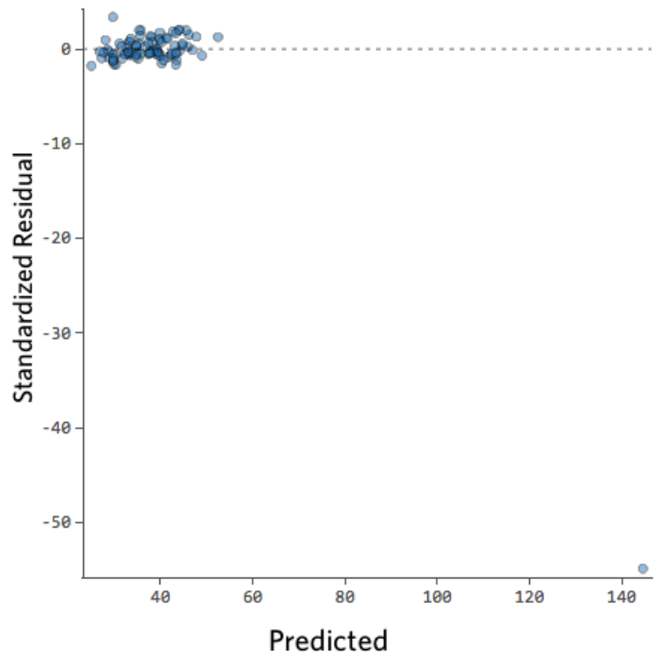
- unbalanced y-data

bad!



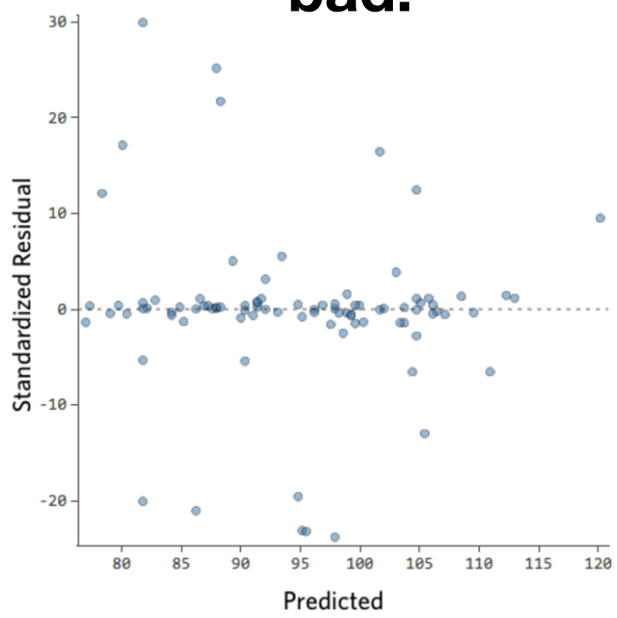
- unbalanced x-data

bad!



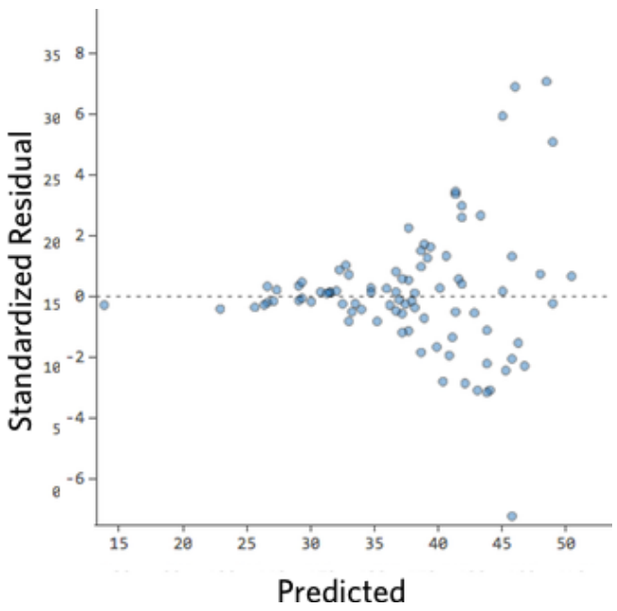
- outliers

bad!



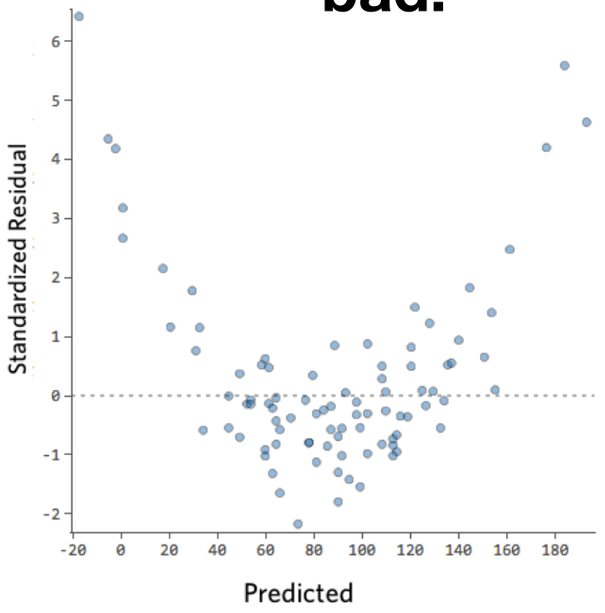
- missing predictors in model

bad!



- heteroscedasticity

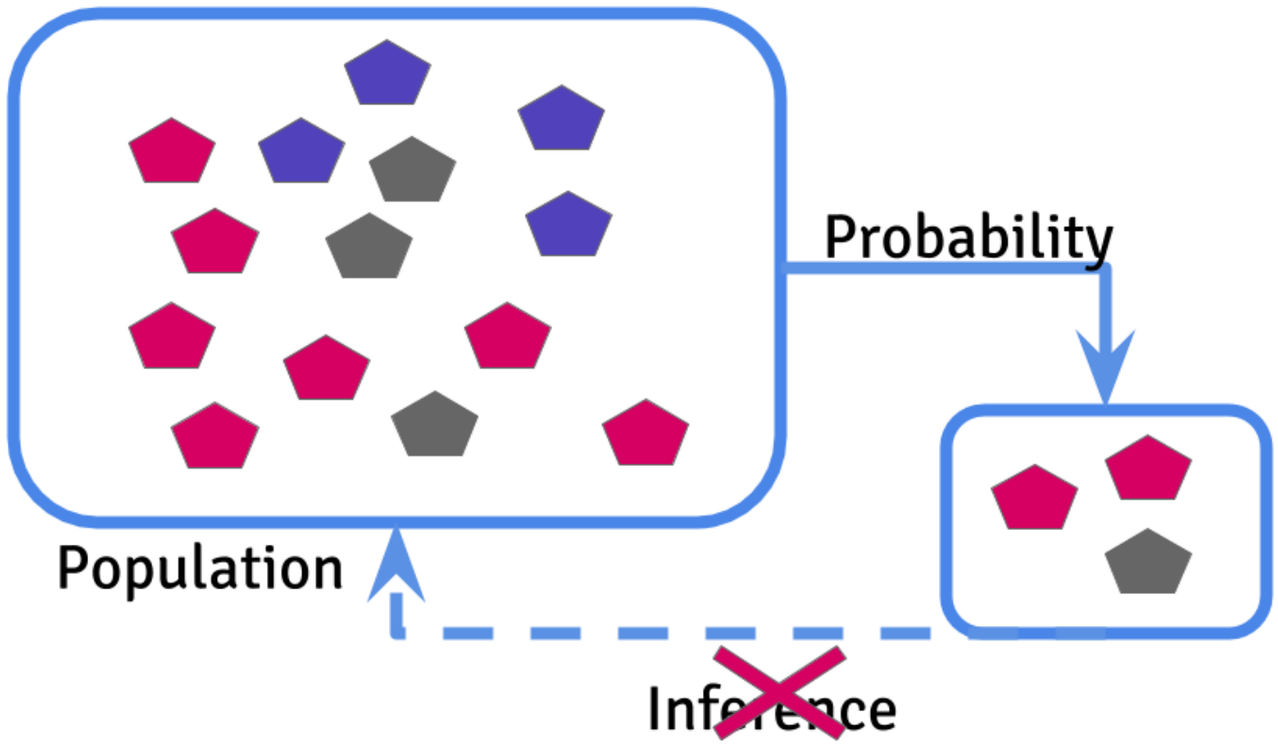
bad!



- non-linear

Common mistakes

- Drawing conclusions about the wrong population (selection effects)

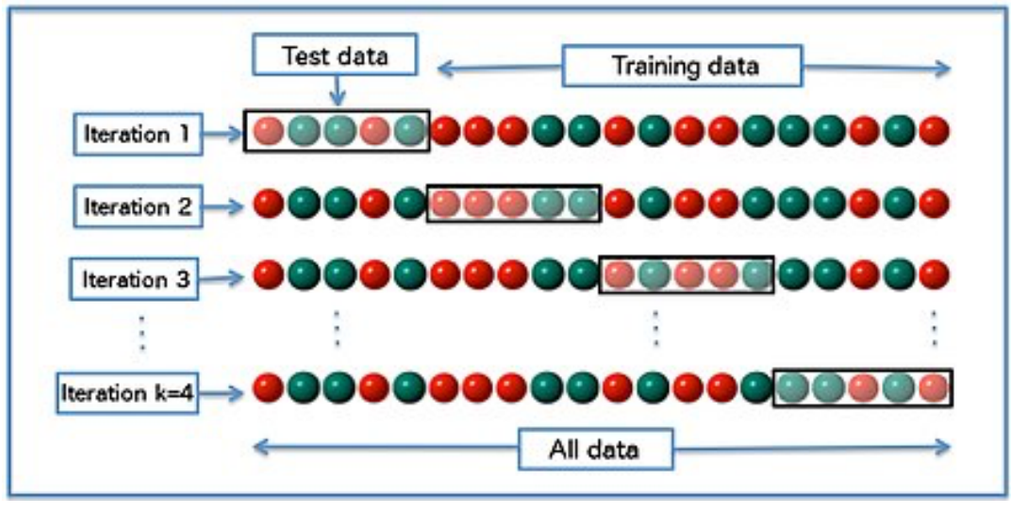


- Not addressing uncertainty, not providing confidence intervals

Prediction

- First thing: split sample randomly(!) into training set (~70%) and test set (~30%)
- put test set aside and use it at the very end once and only once to estimate true error rate of your algorithm
- to estimate true error of algorithm on smaller data sets: use cross-validation
  - split data set into two (randomly); train models on first, evaluate on second
  - repeat with different splits
  - average prediction errors from each split

example: 4-fold cross-validation



- more (better) data is typically more useful than better algorithm: "the unreasonable effectiveness of data"

Prediction

- If goal is prediction accuracy: average many prediction models together
  - reduce variability and does not introduce additional bias

- There is always a trade-off between

Accuracy vs
- Interpretability, Simplicity
- Speed
- Scalability

https://en.wikipedia.org/wiki/Netflix_Prize

- Predict user ratings for films based on previous ratings (grades=1..5)
- Training set: {<user, movie, date of grade, grade>}
- Qualification set: {<user, movie, date of grade>} (grades known to jury)

- Winning team received $1,000,000 prize in 2009 but their algorithm was never implemented (e.g., too complex, change in netflix operations)

https://www.techdirt.com/articles/20120409/03412518422/why-netflix-never-implemented-algorithm-that-won-netflix-1-million-challenge.shtml

Summary

Descriptive: Example US census

- collect resident type, age, location, etc.
- interpretation/use left to politics

Exploratory:

- search for trends, correlations to generate ideas or hypothesis
- to be tested rigorously later

Predictive:

- build (close) approximation of relationship
- predict individual outcomes on new data set

Inferential:

- Make statistical inferences about the data
- Is pattern real or by chance (stat. tests)?
- pattern holds in other data sets?

Causal:

- Learn about causal relationships in the data
- How does changing one measurement affect the outcome?

Mechanistic:

- Learn about deterministic relationships between individual data
- Goal: Explain what happens based on underlying (often deterministic) processes