Week 1: Introduction, Software, and Data Considerations

MATH-517 Statistical Computation and Visualization

Linda Mhalla

September 13, 2024

Classes

Lectures

- Teacher: Linda Mhalla
- Time: Friday 10:15-12:00
- Place: GC D0 386

Exercises

- Teacher: Charles Dufour
- Time: Friday 13:15-15:00
- Place: CM 1 221

Statistical **Computation** and Visualization

Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill.

– Press et al., *Numerical Recipes*

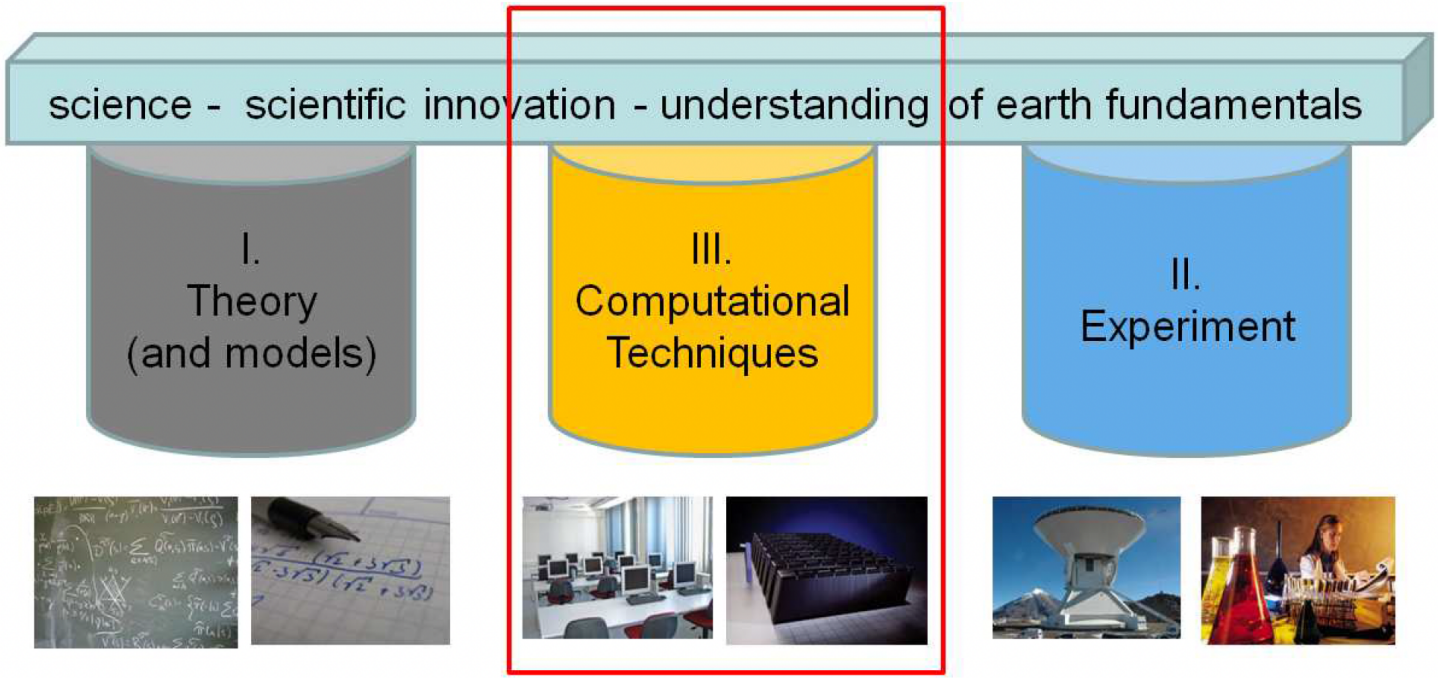
Statistical **Computation** and Visualization

> *Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill.*
>
> – Press et al., *Numerical Recipes*

Apart from Monte Carlo (MC), we will cover (re)sampling methods such as

- cross-validation
- bootstrap
- jackknife
- Bayesian MC extensions

The Three Pillars of Science



Statistical Computation and **Visualization**

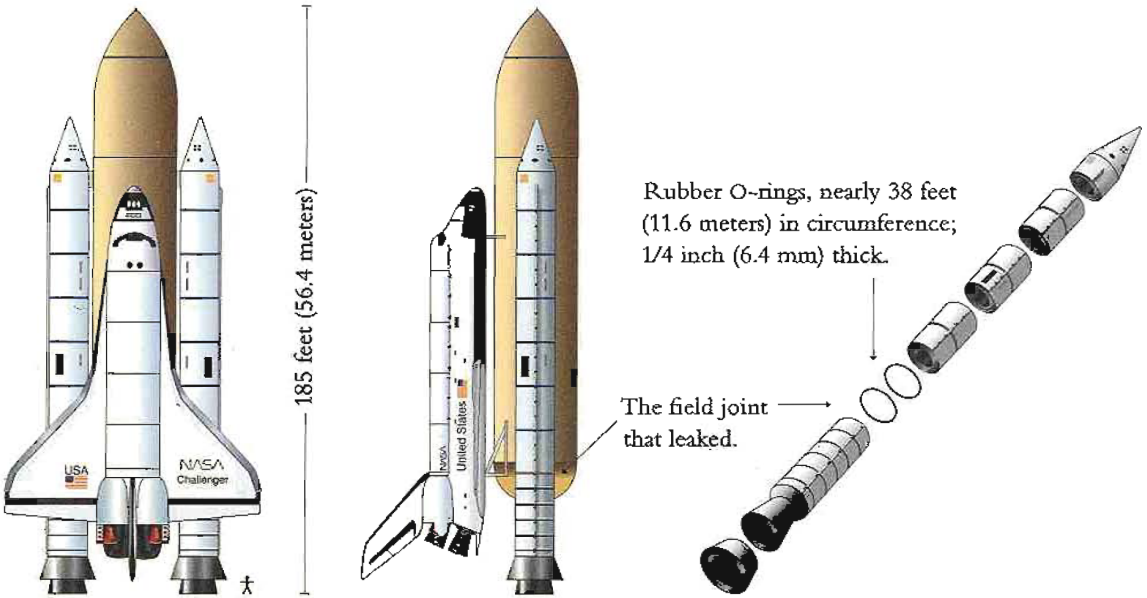
The scientific aim of any visualization is to allow the reader to understand data and extract information

- intuitively,
- efficiently, and
- accurately

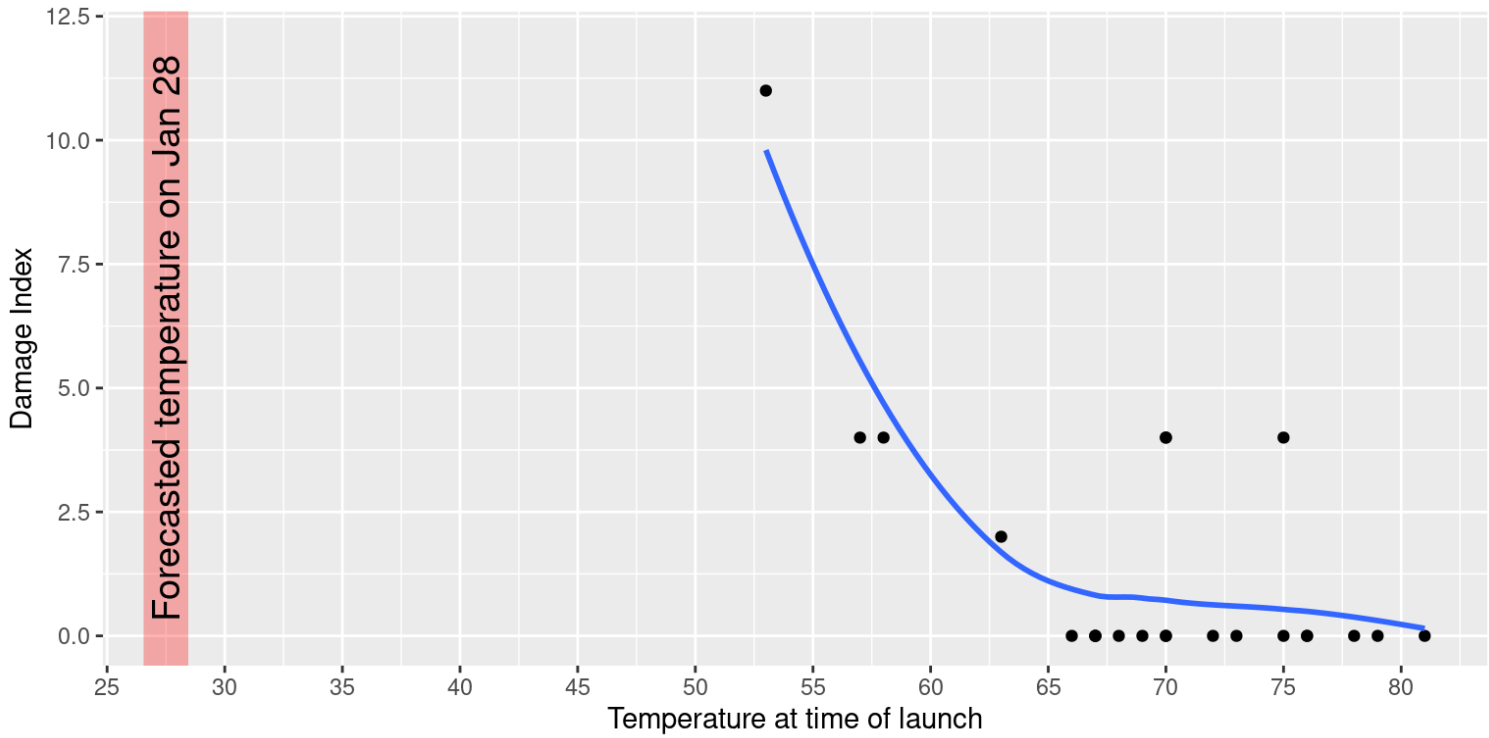
A successful data visualization will

- grab attention,
- improve access to information, and
- summarize content

Statistical Computation and **Visualization**



Statistical Computation and **Visualization**



NASA managers excluded the flights where no failures happened

Bad Visualization

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

Course Content (Chronologically)

- Soft Start
  - R and other software
  - reproducibility and ethics
  - data wrangling and visualization
- Course Core
  - kernel density estimation
  - non-parametric regression
  - cross-validation
  - EM algorithm
  - Monte Carlo (MC)
  - bootstrap
- Bayesian Inference
  - basic thinking
  - Markov Chain Monte Carlo (MCMC)
- Tree-based methods for classification

Polls

- Have you ever written a for-loop and if-else statement?
- Have you ever worked with R?
- Have you ever worked with Julia, Python, Matlab, etc.?
- Have you taken a class dedicated to linear models?
  - prerequisite
- Can you define the p-value?

Polls

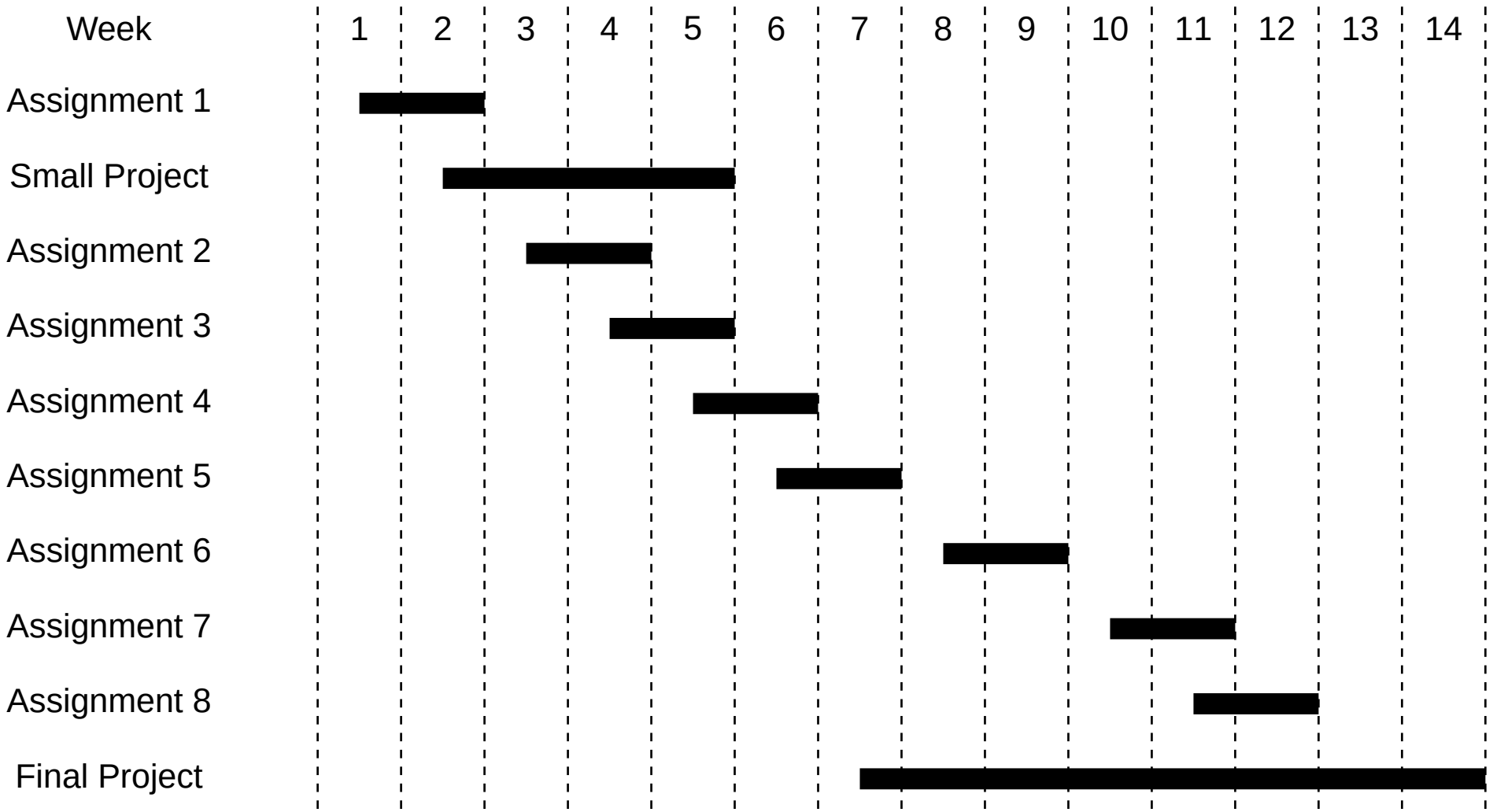
- Have you ever written a for-loop and if-else statement?
- Have you ever worked with R?
- Have you ever worked with Julia, Python, Matlab, etc.?
- Have you taken a class dedicated to linear models?
  - prerequisite
- Can you define the p-value?

- What are your expectations from the course?

Course Requirements

- assignments
  - 40 % of the grade (8 assignments of 5 % each)
  - to be solved during the exercise classes
  - graded on a three-point scale
  - collaboration (and questions) encouraged, but individual submissions required (avoid perfect copies . . .)
- data exploration – small project
  - 20 % of the grade
  - if chosen data set is too simple, can be composed of multiple data sets
  - in groups of 2-3 students
- project: data exploration+analysis *or* simulation study
  - 40 % of the grade
  - in groups of 2-3 students

Expected Progress



Course Requirements

- 1 assignment = 5 % of the grade = 0.25 on the 1-6 grade scale
  - missing all assignments ⇒ final grade 4.0 at best!
- R, Markdown and GitHub for the assignments and projects will be needed
  - submissions are made through GitHub Classroom (see dedicated tutorial on the course website)
  - this is not a programming course, learn by doing!
- 2 hours of lecture per week
  - going through the course content
- 2 hours of exercises per week
  - working on exercises, assignments, and projects
  - keeping up with the lecture

active participation = success in this course

Questions and feedback are always appreciated.

Evaluation starts right away!

Section 1

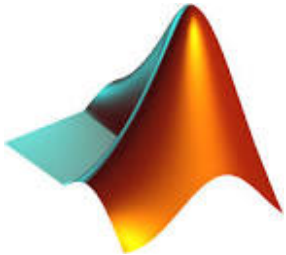
Software

Business Software

- all commercial
  - it has pros and cons
  - all (claim to) offer free academic versions
- popular in different fields
  - SAS: biomedicine, clinical research, etc.
  - SPSS: psychology, social sciences, etc.
  - STATA: econometrics, finance, etc.

Academic Software

Python

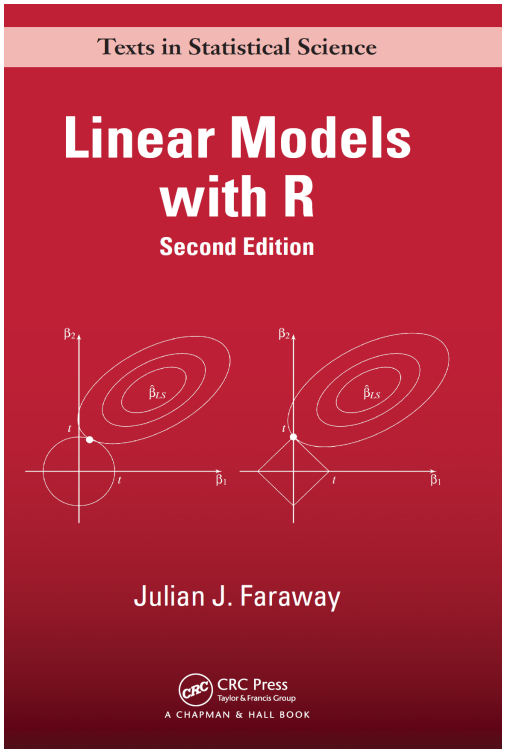


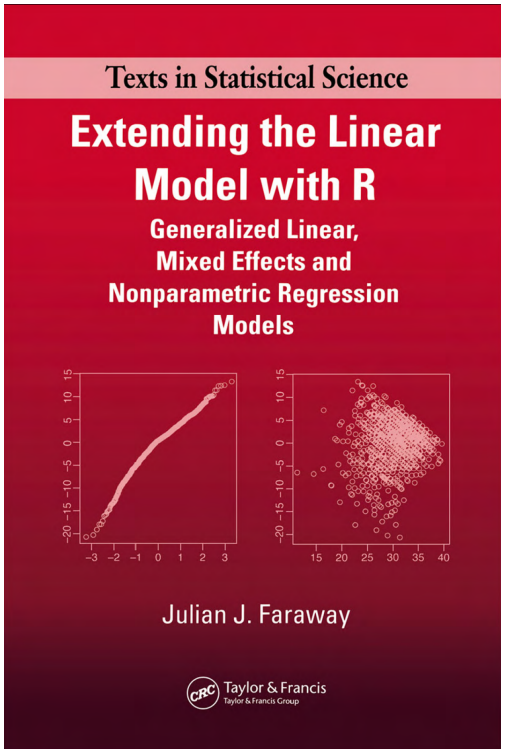
Matlab

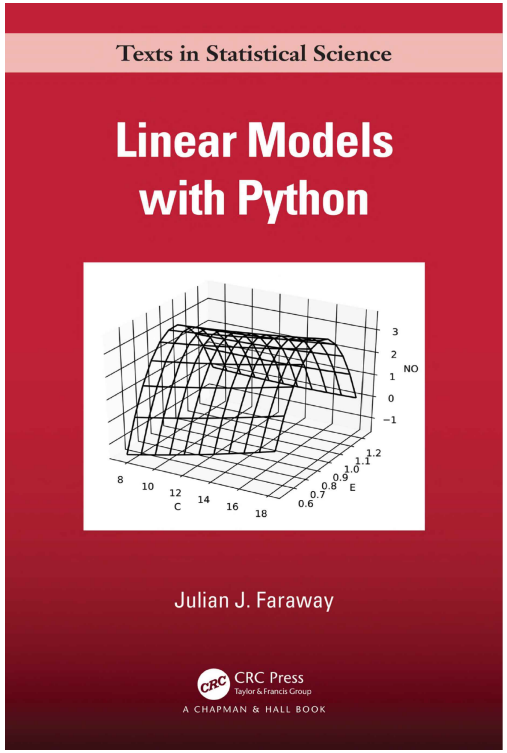
- all well documented, easy to use, with lots of examples and extensive community support
- each has its strengths and weaknesses, none is perfect
- we will mostly use R but Python or Julia are encouraged too!
- software packages are our **tools**, not skills!

free open source	free open source	paid (accessible) closed source
Statistics	Machine Learning	Numerical Math
Data Science		Optimization

Statistics done in R!







2004 – 1st Edition
2015 – 2nd Edition

2006

2020

Beware of **power without wisdom**!

References

JJ Faraway (2015) Linear Models with R (2nd Edition)

JJ Faraway (2020) Linear Models with Python

Poldrack (2019) Statistical Thinking for the 21st Century (online)

Tufte (1997) Visual Explanations

H Wickham et al. (2023) R for Data Science (online)

W McKinney (2022) Python for Data Analysis (online)

B Yu and R Barter (2024) Veridical Data Science online

Section 2

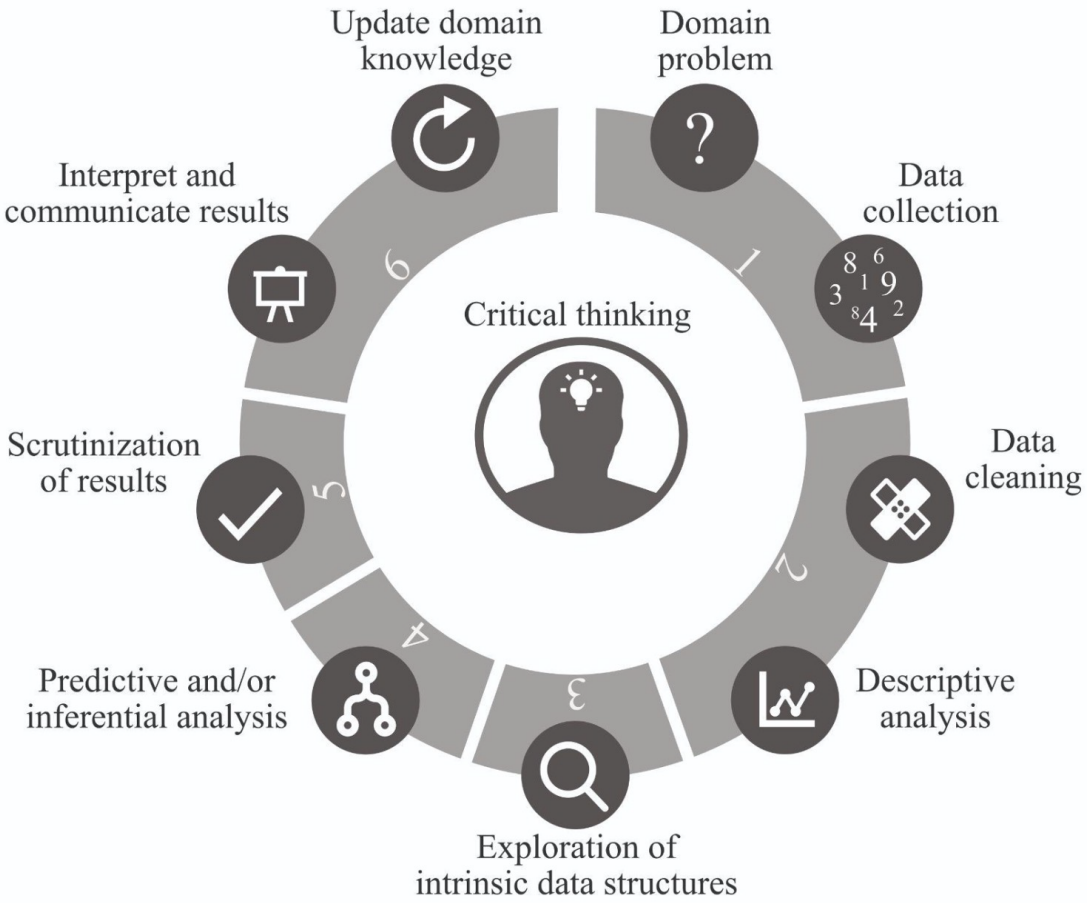
Data Considerations

Job of a Statistician

Understand a complex world by describing it in relatively simple terms that capture essential aspects of its structure, and provide us some idea of how uncertain we are about that knowledge

- think about uncertainty and bias (anticipate and reduce it)
- build models emulating nature
  - inference about the models leads to conclusions about nature – but what if the model is a poor representation of nature?
- provide *interpretable* models allowing for rational conclusions
  - prediction vs. information extraction
  - all models are wrong $\Rightarrow$ critical model validation
- estimate variation ($\Rightarrow$ confidence intervals, significance)
- draw conclusions from data
- traditional role: statisticians invited to analyze existing data
  - e.g., does the existing data contain the desired information?
- modern role: collaborative step-by-step
  - from acquisition of data to presentation of results
  - interdisciplinary communication

Cycle of (Data-driven) Science



credit: Bin Yu, Rebecca Barter (2024)

Reference book: Veridical Data Science (2024)

Cycle of (Data-driven) Science

- Stage 1: Problem formulation and data collection (part of modern role of statistician collaborating with domain experts)
- Stage 2: Data cleaning/wrangling (spot issues with data, preprocess to fit algo requirements) and **EDA** (through data viz)
- Stage 3: Uncovering intrinsic data structures (unsupervised learning s.t. dimensionality reduction and clustering to uncover complex patterns). Not covered
- **Stage 4**:
  - 4.1: Inferential analysis (role of statistician), e.g., parameter estimation and uncertainty quantification. Main focus of the course
  - 4.2: Predictive analysis (focus of supervised machine learning algorithms), e.g., classification and regression trees. Computational aspects will be briefly covered towards the end of the semester
- Stage 5: Evaluation of results (do they make sense?)
- Stage 6: **Communication of results (e.g., through data viz and sharing code)** and updating domain knowledge

Domains of Application

- actuarial science
- biostatistics (medicine, pharma, genetics, etc.)
- business
- chemometrics
- econometrics
- epidemiology
- finance
- geostatistics
- machine learning and AI
- official statistics (demography, surveys, etc.)
- psychology
- quality control
- reliability
- physics
- signal processing
- ...

Section 3

Data Considerations: Reproducibility

Statistics in Science

An overwhelming portion of contemporary scientific conclusions is based on the concept of *statistical significance*:

- there is a hypothesis (e.g. drug A is better than drug B)
- data is collected (e.g., patients are split, some are given drug A, others drug B, and some relevant response Y is collected)
- null hypothesis is formed (e.g., "effect A on Y = effect B on Y")
  - this usually requires a model
- if p-value $< 5\ \%$, conclusion is reached

Problems:

- was there really a hypothesis at the beginning?
- how exactly were data collected?
- is the model good? (confounders?)
- the "if" above: Cargo-cult statistics
  "the mechanical application of methods without understanding their assumptions, limitations, or interpretation will surely reduce scientific replicability"

Defining Reproducibility

- a type of stability assessment (e.g., to the data cleaning, to the code written, to the person who conducted the analysis) and/or,
- a predictability assessment (re-evaluation involves showing that the results reemerge using new data)

Questionable Practices I

- from a chapter titled "Writing the Empirical Journal Article" a popular career guide in psychology:

  "There are two possible articles you can write: (1) the article you planned to write when you designed your study or (2) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (2). [...] If you see dim traces of interesting patterns, try to reorganize [...] Go on a fishing expediction for something, anything..."

- the author suggests to rewrite their theory based on the facts, rather than using the theory to make predictions and then test them
- this is called *p*-hacking (strategies for rendering non-significant hypothesis testing results significant) and should be avoided!
- this includes many strategies such as selective reporting of (in)dependent variables or selective data trimming, all of which could increase the actual false positive rate!

Questionable Practices II

- Carney, Cuddy & Yap (2010)
  - Power posing . . . has positive effects on your mind and hormones
- Cuddy (2012) Your body language may shape who you are, TED talk
  - 2nd most viewed TED talk of all time
- 2015 first reproducibility issues
- 2016 Carney withdraws her name
  - she no longer believes in the effect, because

We ran subjects in chunks and checked the effect along the way. It was something like 25 subjects run, then 10, then 7, then 5. Back then, this did not seem like p-hacking. It seemed like saving money.

- this is called peeking (at p-values) and should be avoided! (or admitted and corrected for)

Example of Peeking

```
peeking <- function(a,b=10){
  x <- rnorm(25)
  Tstat <- mean(x)/sd(x)*sqrt(length(x))
  if(abs(Tstat) > qt(0.975,length(x)-1)){
    return(Tstat)
  }else{
    x <- append(x, rnorm(b))
    Tstat <- mean(x)/sd(x)*sqrt(length(x))
    return(Tstat)
  }
}
set.seed(517)
Tstats <- sapply(1:10000,peeking)
mean(I(abs(Tstats) > qnorm(0.975)))
```

```
## [1] 0.0851
```

The severity depends on the number of peeks the researcher takes at the data and on the number of observations added between peeks

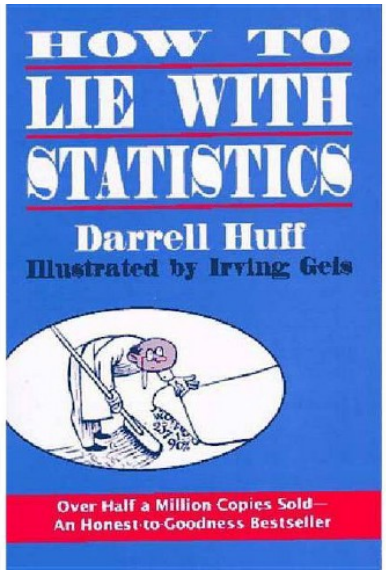
Reproducibility Crisis

- Based on 100 papers published in psychology, a meta-analysis "The Reproducibility Project" finds that only 37 out of the 97 significant results, were significant – "Estimating the Reproducibility of psychological science". *Science* (2015)
- similar issues in other fields, e.g., biology, chemistry, economics, social sciences, or nutrition
  - we are talking even about *Nature* and *Science* publications not being reproducible!
- it is *not* the fault of *p*-values

More on the reproducibility crisis in science can be found here and here

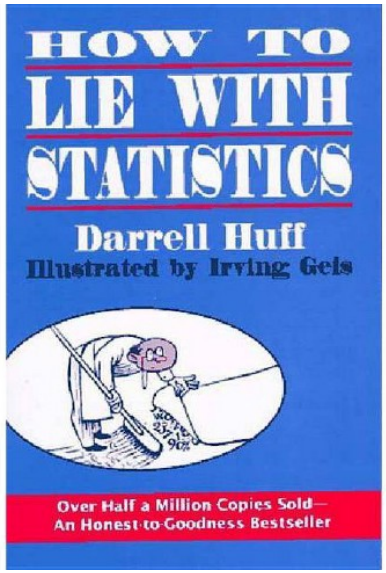
More on Shady Practices

- Review of *p*-hacking strategies: code and Shiny App provided here!
- A classic book discussing data literacy



More on Shady Practices

- Review of *p*-hacking strategies: code and Shiny App provided here!
- A classic book discussing data literacy



"Ironically, written by a journalist with little knowledge of statistics who later accepted thousands of dollars from cigarette companies and told a congressional hearing in 1965 that inferences in the Surgeon General's report on the dangers of smoking were fallacious."
– Andrew Gelman

Reproducible Practices

- sharing data (and code) using version control such as GitHub for example
- documenting data collection/cleaning and analyses
  - in particular any judgement calls (mostly data cleaning and modeling choices, but sometimes also tuning parameter selection, etc.)
  - avoid doing manual modifications to data files
- encapsulation or rerun code from start on fresh sessions
- pre-registration (e.g., at ClinicalTrials.gov)
- publishing negative results

Reproducible Data Analysis

Levels of Research Reproducibility

Trying things out in the R Console pane, saving tables and figures to named files.

Writing code in an RScript or Rmarkdown file, generating saved tables, figures, and manuscripts that can be re-run if needed.

Using Projects and {here} , sharing code, documenting with README and comments, doing code review,
and sharing code publicly on GitHub

All of the above, plus public sharing of code and data, and preserving your local package environment with {renv}

All of the above plus encapsulating the entire computing environment (R, packages, code) in a Docker image.

Section 4

Data Considerations: Ethics

Ethical Guidelines for Statistical Practice

- Professional integrity and accountability
  - expose yourself to (self-)criticism
- Integrity of data and methods
  - aim for reproducibility
- Responsibilities to stakeholders
- Responsibilities to research subjects
  - research on living beings must be supervised
  - privacy for human subjects
- Multidisciplinary teams
  - profession-specific ethical guidelines
- Responsibilities to the statistical profession, mentoring, etc.
  - the career guide above fails big here

See the ASA guideline for details

Assignment 1 [5 %]

"**Assignment**" = **mandatory**, counts towards the final grade [5 %]

Go to Assignment 1 for details.

"**Exercise**" = does not count towards the grade

Go to Exercise 1 for details.