CMSE 890-310

Gaps, Missteps, & Errors in Statistical Data Analysis

Arjun Krishnan

arjun@msu.edu | @compbiologist | thekrishnanlab.org

Day 01

Welcome, Overview, Getting started

- Welcome, overview
- Introductions
- Scope & topics
- Website & communication
- Course activities
- What's due next week?
- Wrap-up

Introductions

- Please call me 'Arjun'.
- **arjun**@msu.edu | the**krishnan**lab.org | @**comp**biologist
- Assistant Professor
  - Dept. Computational Mathematics, Science, and Engineering
  - Dept. Biochemistry and Molecular Biology
- Research Interests: Computational genomics, Biomedical data science, Biological networks, Natural language analysis, Data integration, Machine learning

Breakout intros!!

Introductions

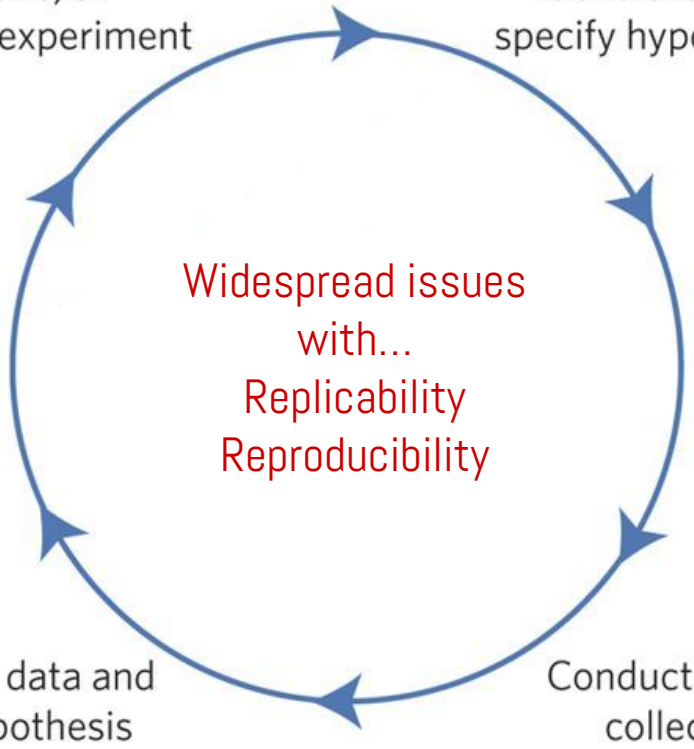
Introduce yourself to your fellow learners in the breakout room!

- I strongly recommend turning your video on so that you can all see each other :)

While you're in your breakout rooms, also introduce yourself to everyone in this class on the #welcome channel on Slack with:

- Name:
- Preferred pronoun:
- Three words/phrases to describe you/your-interests:
- Research/interests in emojis:

What's this course about?



Visualization challenges
Publication bias
Lack of holistic analysis

Failure to control for bias

P-hacking
Ignoring multiple testing
Ignoring base rates

Low statistical power
Not considering confounders

No exploration
Model abuse
P-hacking

Poor quality control
Not recording all metadata

https://www.nature.com/articles/s41562-016-0021

What's this course about?

Questionable requests that biostatisticians commonly receive:

- Altering some data to support hypothesis
- Interpreting findings on basis of expectation
- Not reporting missing data
- Ignoring violations of assumptions

[These requests are reported more frequently by younger statisticians.]

Trainees…

- Pressured by a PI or collaborator to produce "positive" data
- Pressure to publish influences the way they report data.

Ann Intern Med. 2018;169(8):554-558
Clinical Cancer Res. 2018;2(14)

What's this course about?

This is an advanced short (1-credit) course designed to:

- Discuss common misunderstandings & typical errors in the practice of statistical data analysis.
- Provide a mental toolkit for critical thinking and enquiry of analytical methods and results.

Prerequisites

We will assume:
1) Familiarity with basic statistics & probability
2) Ability to do basic data wrangling, analysis, & visualization using R or Python.

What's this course about?

The first principle is that you must not fool yourself, and you are the easiest person to fool.

– Richard Feynman

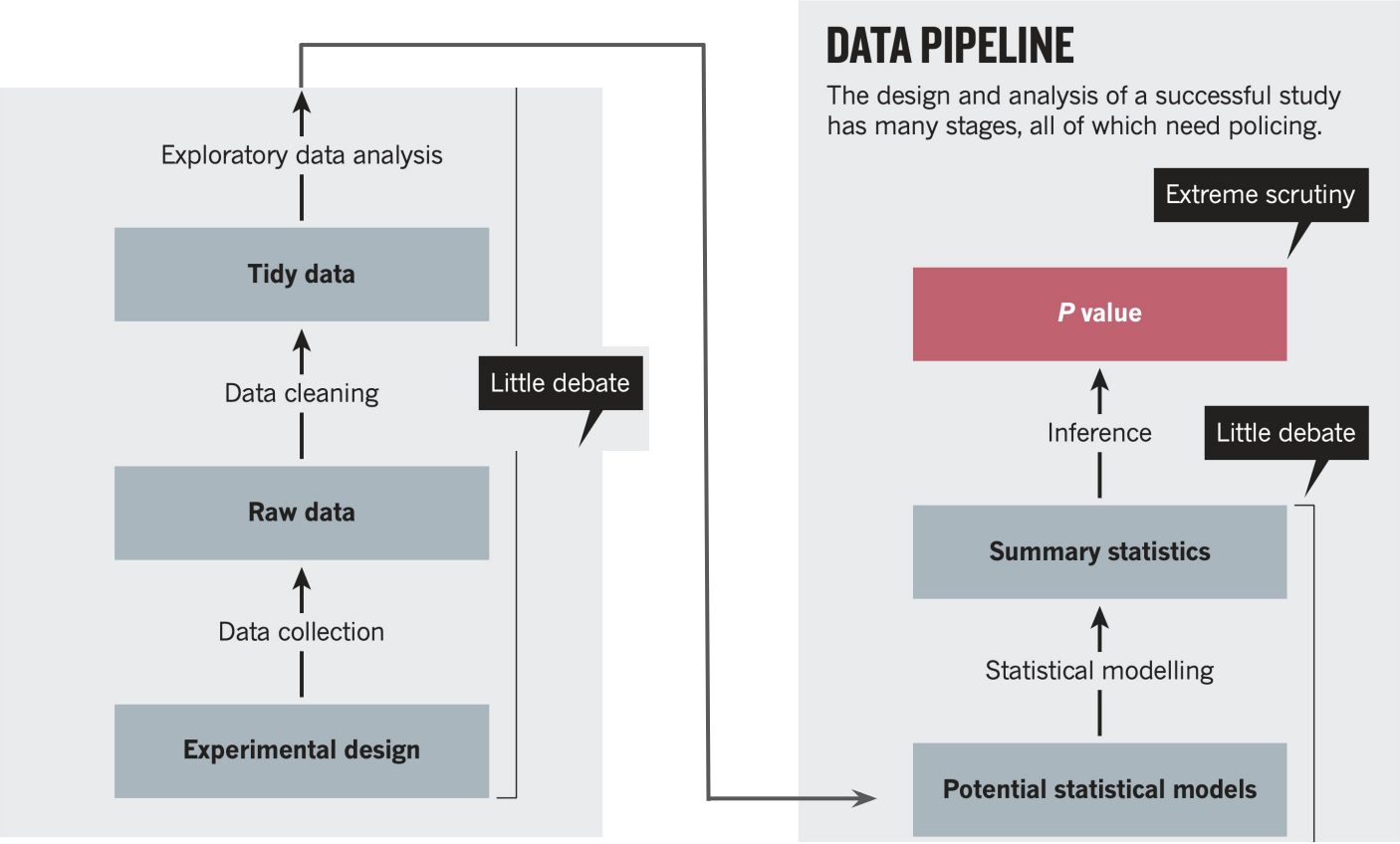
If your experiment needs statistics, you ought to have done a better experiment.

– Ernest Rutherford

He uses statistics as a drunken man uses lamp-posts... for support rather than illumination.

– Andrew Lang

What's this course about?

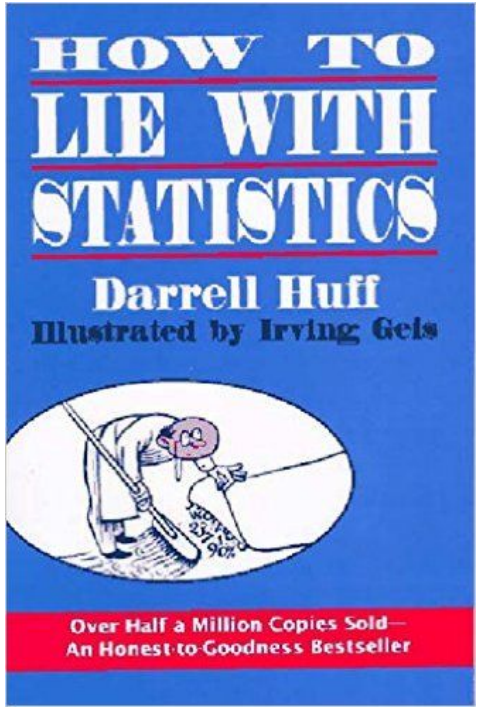


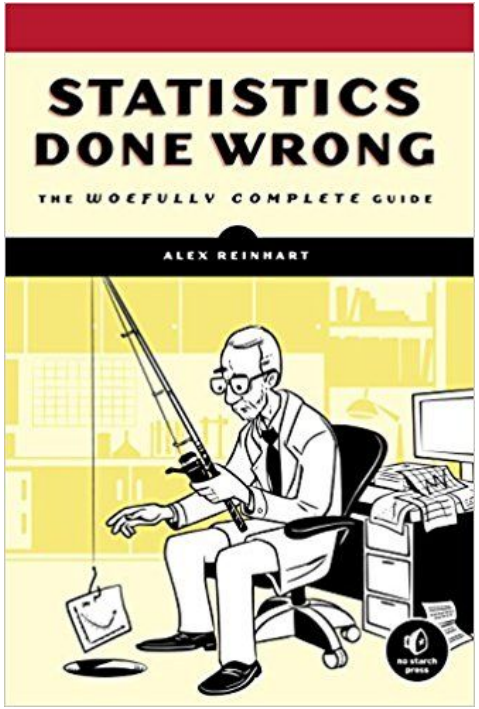
JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

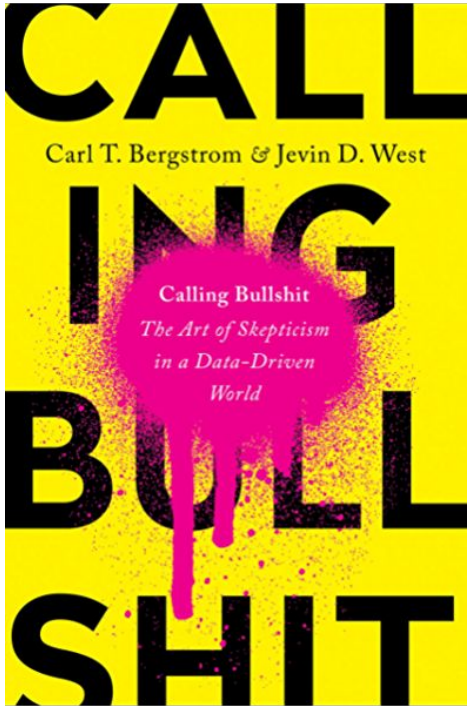
What's this course about?

- Estimation of error & uncertainty, Sampling biases
- P-value & P-hacking, Multiple hypothesis correction
- Statistical power & Underpowered statistics
- Pseudoreplication, Confounding variables & batch effects
- Circular analysis, Regression to the mean & stopping rules, Cognitive biases
- Base rates & conditional probabilities
- Measuring association with continuous variables
- Visualization challenges
- Researcher degrees of freedom, Data sharing, Reproducible research

Resources







Original research articles | Reviews | Blog posts | Podcasts

What's this course about?

The Modelers' Hippocratic Oath

~ I will remember that I didn't make the world, and it doesn't satisfy my equations.

~ Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.

~ I will never sacrifice reality for elegance without explaining why I have done so.

~ Nor will I give the people who use my model false comfort about its accuracy.

Instead, I will make explicit its assumptions and over-sights.

~ I understand that my work may have enormous effects on society and the economy,

many of them beyond my comprehension

Emanuel Derman
January 7 2009

Paul Wilmott
January 7 2009

THE TEN COMMANDMENTS OF STATISTICAL INFERENCE

MICHAEL F. DRISCOLL

The original version of these commandments has apparently been lost, perhaps in antiquity. There may now exist several variants. One has appeared in Thomas [1]; here is another.

I. Thou shalt not hunt statistical significance with a shotgun.
II. Thou shalt not enter the valley of the methods of inference without an experimental design.
III. Thou shalt not make statistical inference in the absence of a model.
IV. Thou shalt honor the assumptions of thy model.
V. Thou shalt not adulterate thy model to obtain significant results.
VI. Thou shalt not covet thy colleague's data.
VII. Thou shalt not bear false witness against thy control-group.
VIII. Thou shalt not worship the 0.05 significance level.
IX. Thou shalt not apply large-sample approximations in vain.
X. Thou shalt not infer causal relationship from statistical significance.

Reference

1. D. H. Thomas, Figuring Anthropology: First Principles of Probability and Statistics, Holt, Rinehart, and Winston, New York, 1976, pp. 458–468.

DEPARTMENT OF MATHEMATICS, ARIZONA STATE UNIVERSITY, TEMPE, AZ 85281.

The American Mathematical Monthly
Volume 84, Number 8, 1977 (p. 628)

Course website

bit.ly/statgaps2020

- Contact information
- Course outline and materials
- Schedule, location, calendar, & offline hours
- Website and communication
- Course activities
- Grading information
- Attendance, conduct, honesty, and accommodations

- Lecture slides
- Learning materials
- Assignments
- Notes

Communication

statgaps2020.slack.com

- The primary mode of communication in this course (including major announcements) will be the course Slack account.
- All of you should have invitations to join this account in your MSU email.

#announcements	#talks-seminars
#slides-materials	#fun-breaks
#primers-articles	#random

Offline discussion hours

bit.ly/statgaps2020_incoming

- Select convenient hours for offline discussion
  - Will give preference to enrolled students when picking the time.
  - Even if you're not able to make it to the designated hours, just messaging on Slack with your questions/concerns will work as well.

Course activities

- Assignments: 40%
- Class participation: 60%

Just like in the real world:

- There are no tests of memory. I strongly encourage you to talk to your fellow learners, peers, mentors, and me. Also, everything is open-internet. You can refer to anything you like.

Assignments

- Will be posted on Slack a week before it is due.
- The goal is to prepare for the discussions the following week:
  - Concepts in statistics / data-analysis to brush-up
  - R and Python commands, functions, packages to brush-up

Class participation

- Do the assignments and additional readings.
- Show up to class.
- Work in groups during in-class discussion sessions.
- Contribute to material in-class and on slack.
- No one will have the perfect background + the topics are all non-straightforward at all.
    - Ask questions about statistical or biological concepts.
- Postdocs, researchers, & faculty-members: I'm asking for your active engagement with the class & its materials, along with any feedback.

Class participation

This course is heavily **discussion-based**.

- I would *really* like your help in sustaining healthy discussions.
  - Stop and ask questions.
  - Feel free to interrupt me to share your thoughts. If you prefer, you're welcome to raise your hand and I can pause.
  - To the maximum extent possible, please keep your videos on.

33andMe

- The most underrated part of teaching is learning. I design courses that help me learn.
- Things to note:
  - I do not have a PhD in Statistics. I consider myself as an almost-power-user!
  - I will tell what parts of my understanding of these topics/ideas are works in progress and, hence, known-incomplete. I will try to be explicit about where the limits of my knowledge & understanding are.
  - I have no problem saying "Hmm, I'm not sure. Let me think about this & get back to you" or "I have no clue now but, if you're interested, we can read a couple of sources together & revisit this."
  - Correct me if/when I'm wrong.

Coding

You will be working with code to:

- read-in existing datasets or generate mock datasets,
- wrangle them into a convenient format,
- call common statistical functions from standard packages/libraries to calculate mean, std. deviation, quantiles, correlation, etc.
- implement some simulations/tests
  - random number generation
  - writing for/while loops
- make plots (scatterplot, histograms, boxplots, etc.)

Coding

Language, IDE, Notebook	R | RStudio | R Notebook	Python | Rodeo | Jupyter
Pre-built external packages	CRAN, Bioconductor	PyPI, Biopython
Scientific computing	In-built + Hundreds of packages	NumPy, SciPy + Hundreds of packages
Data wrangling & visualization	Tidyverse	Pandas, Seaborn

There are hundreds of software packages for statistical data analysis written in various languages (C, C++, R, & Python) that can be run from the command-line.

- Linux command-line
  - Navigating the file system
  - Running code
  - Manipulating data
  - Writing shell scripts

What you need to do before the next class

PART 1

Complete the incoming survey: bit.ly/statgaps2020_incoming

Among other things, this will help in finding a time for offline discussions.

If you have not already done so, I would also like you to complete the incoming survey: bit.ly/statgaps2020_signup

What you need to do before the next class

PART 2

Install R or Python

- Install R, RStudio, and Tidyverse (package); Get familiar with R Notebooks, or
- Install Anaconda, Python 3.7, Jupyter Notebooks

Resources with detailed instructions are on the class website.

Resources @ MSU

Center for Statistical Training and Consulting

- Training resources: https://cstat.msu.edu/resources
- Events and workshops: https://cstat.msu.edu/events

Working/student groups

- R-Ladies: https://rladies-eastlansing.github.io/

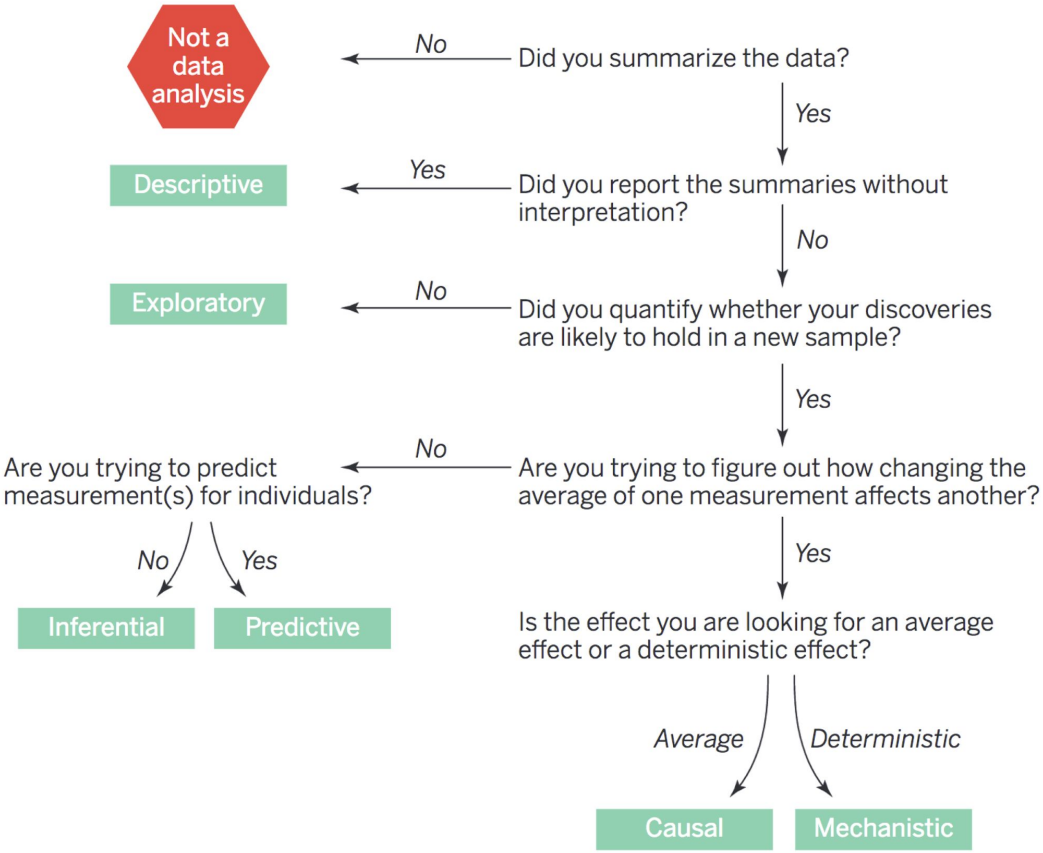
What you need to do before the next class

PART 3

- A major goal of this course is to prepare your ability to perform and critique statistical data analysis and to present your ideas and results effectively.
- bit.ly/statgaps2020_assignment01
- This assignment will give you an opportunity to revisit many statistical concepts and set the tone for this course.

Look out for messages on all channels: stagaps2020.slack.com

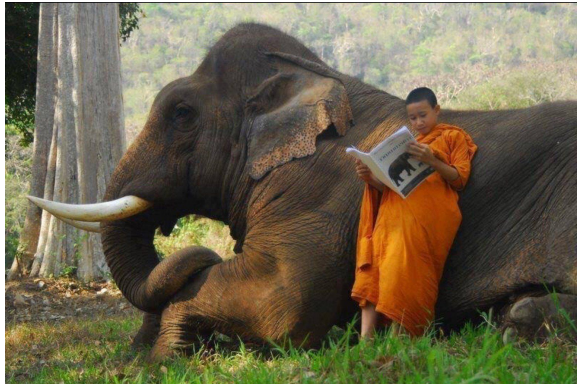
How to pick an analysis/result to focus on?



JT Leek, RD Peng What is the question?

Getting help

- **Linux** | rik.smith-unna.com/command_line_bootcamp, commandline.guide, & swcarpentry.github.io/shell-novice
- **Python** | Introduction: learnpythonthehardway.org/book & developers.google.com/edu/python | Data analysis: jakevdp.github.io/WhirlwindTourOfPython | Visualization: www.r-graph-gallery.com
- **R** | Introduction: swcarpentry.github.io/r-novice-inflammation & swirlstats.com ('R Programming' & 'Data Analysis') | Data analysis: r4ds.had.co.nz | Visualization: python-graph-gallery.com
- **Git & GitHub** | swcarpentry.github.io/git-novice/, speakerdeck.com/alicebartlett/git-for-humans, & rogerdudler.github.io/git-guide/
- **Probability and Statistics** | Nature Collection (Statistics for Biologists | Practical Guides | Points of Significance): www.nature.com/collections/qghhqm



Google ... so many excellent blog posts!

Getting help – Additional reading

- Fantastic resources on Reproducible code, Data management, Getting published, and Peer review
  http://www.britishecologicalsociety.org/publications/guides-to/
- A Quick Guide to Organizing Computational Biology Projects
  https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424
- A Quick Introduction to Version Control with Git and GitHub
  http://dx.plos.org/10.1371/journal.pcbi.1004668
- Ten Simple Rules for Taking Advantage of Git and GitHub
  http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004947