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

Getting started...

- Coding
- Organizing a data analysis project
- Resources
- Wrap-up

Introductions

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

Introduce yourself to one other person in this class you've not met before:

Say your name, department/program

Introductions

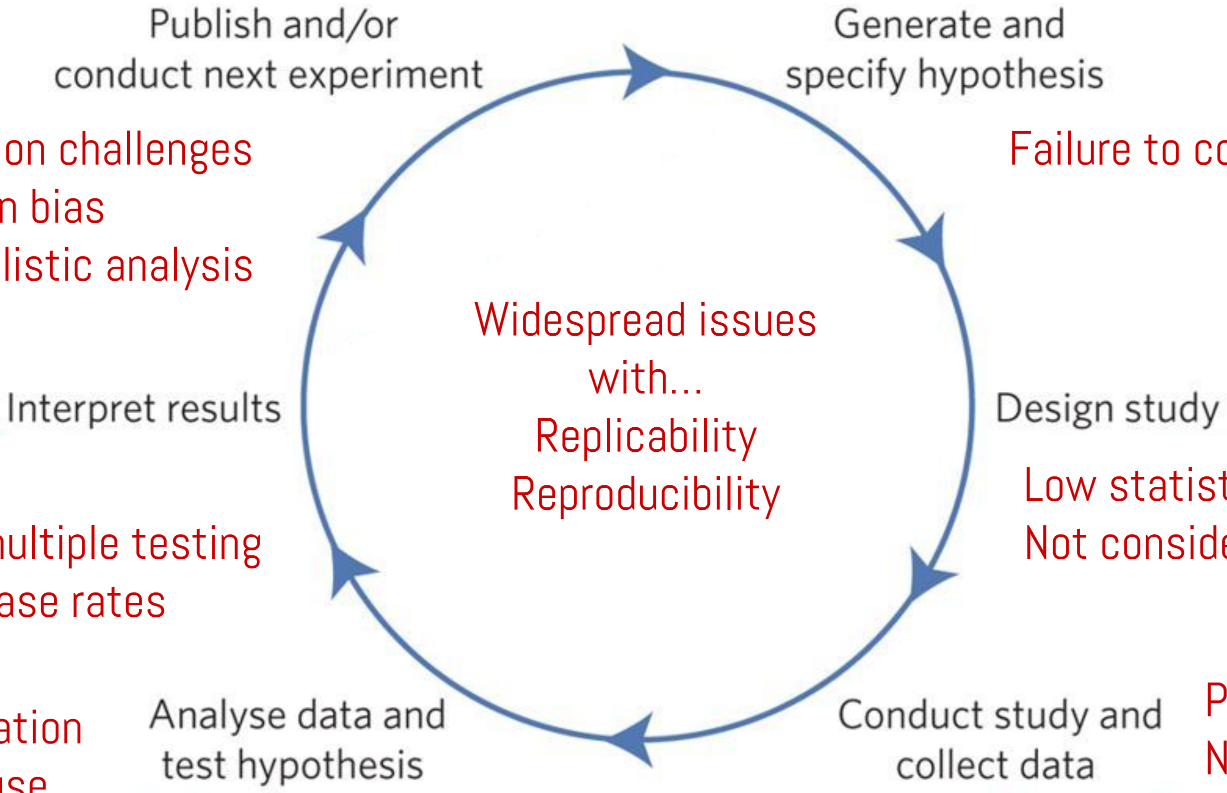
Introduce yourself on the #welcome channel on Slack with:

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

If you have not joined Slack, now is a good time to do it.

Come & talk to me if you've not received an invitation.

What's this course about?


Publish and/or conduct next experiment
Generate and specify hypothesis
Design study
Conduct study and collect data
Analyse data and test hypothesis
Interpret results

Widespread issues with...
Replicability
Reproducibility

Failure to control for bias

Low statistical power
Not considering confounders

Poor quality control
Not recording all metadata

No exploration
Model abuse
P-hacking

P-hacking
Ignoring multiple testing
Ignoring base rates

Visualization challenges
Publication bias
Lack of holistic analysis

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.

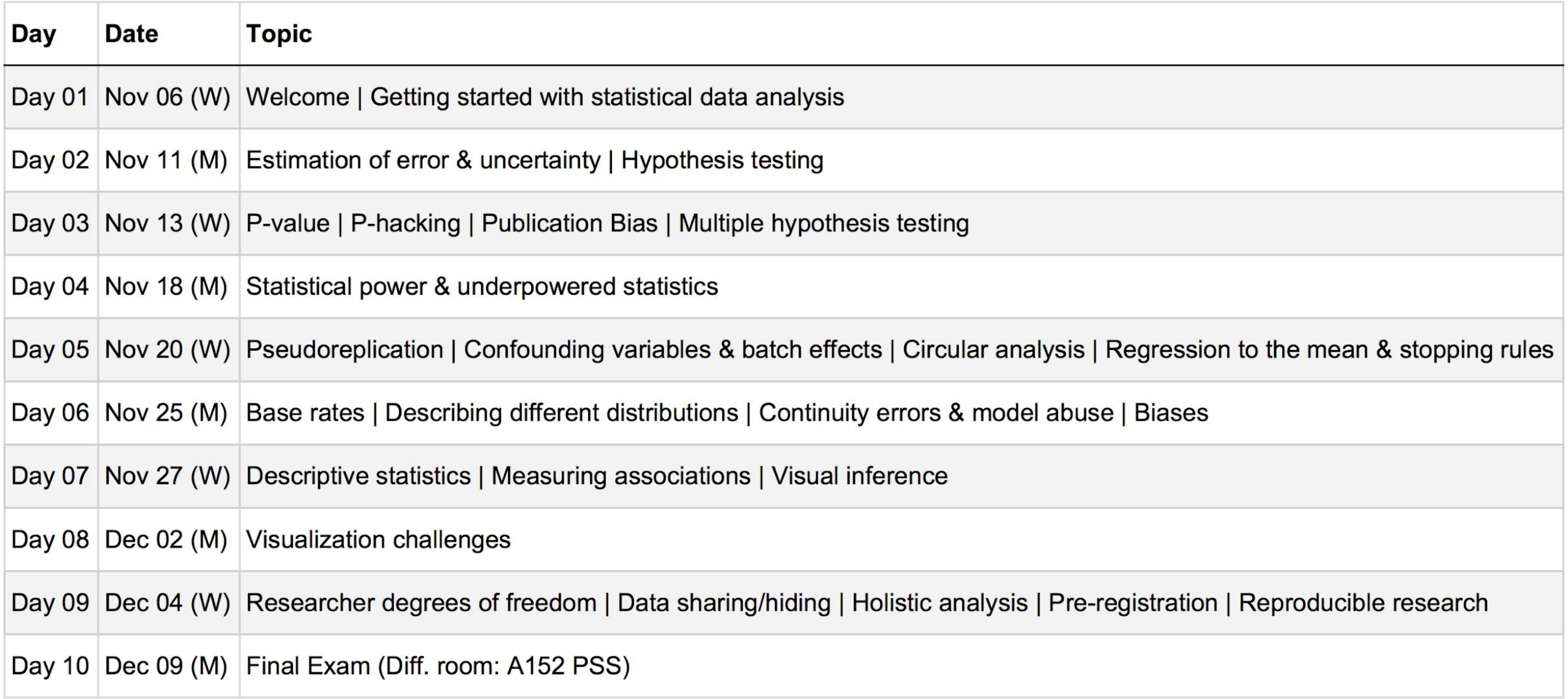
Prerequisites

We will assume:

1) Familiarity with basic statistics & probability

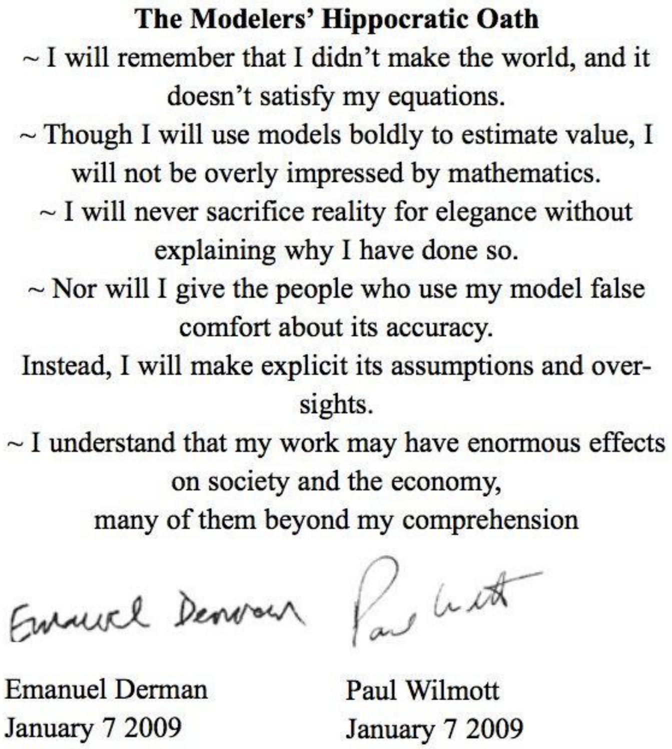
2) Ability to do basic data wrangling, analysis, & visualization using R or Python.

What’s this course about?



Day	Date	Topic
Day 01	Nov 06 (W)	Welcome | Getting started with statistical data analysis
Day 02	Nov 11 (M)	Estimation of error & uncertainty | Hypothesis testing
Day 03	Nov 13 (W)	P-value | P-hacking | Publication Bias | Multiple hypothesis testing
Day 04	Nov 18 (M)	Statistical power & underpowered statistics
Day 05	Nov 20 (W)	Pseudoreplication | Confounding variables & batch effects | Circular analysis | Regression to the mean & stopping rules
Day 06	Nov 25 (M)	Base rates | Describing different distributions | Continuity errors & model abuse | Biases
Day 07	Nov 27 (W)	Descriptive statistics | Measuring associations | Visual inference
Day 08	Dec 02 (M)	Visualization challenges
Day 09	Dec 04 (W)	Researcher degrees of freedom | Data sharing/hiding | Holistic analysis | Pre-registration | Reproducible research
Day 10	Dec 09 (M)	Final Exam (Diff. room: A152 PSS)

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.

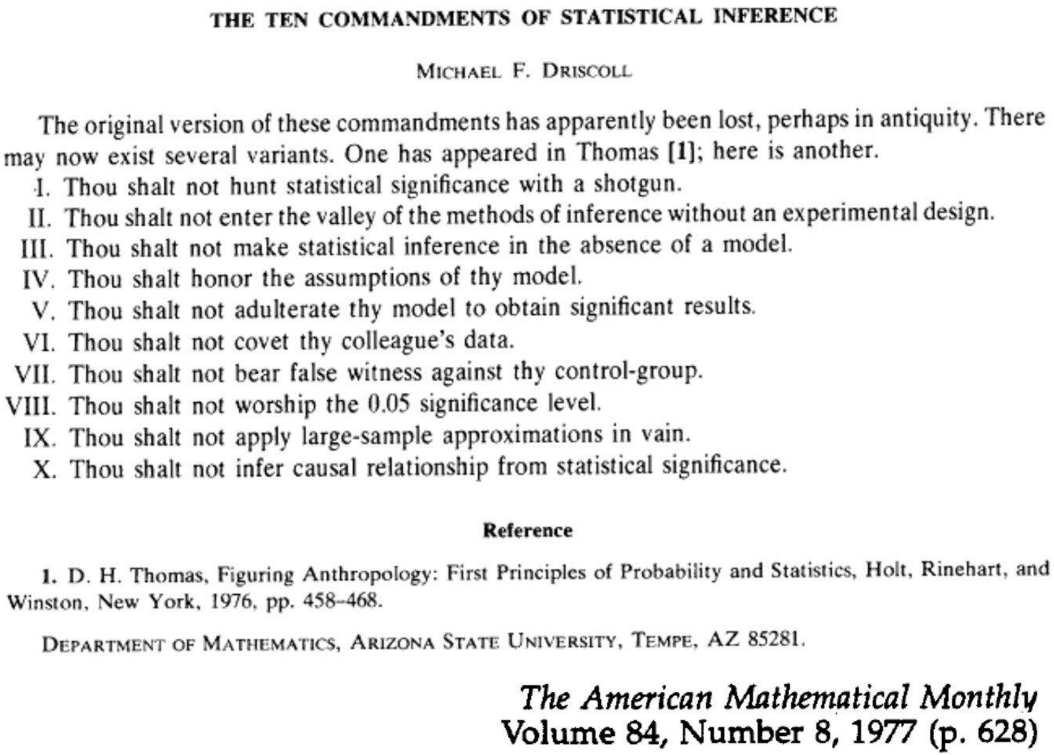
~ Nor will I give the people who use my model false comfort about its accuracy.
Instead, I will make explicit its assumptions and oversights.

~ I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension

Emanuel Derman   Paul Wilmott

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/statgaps2019

- 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

statgaps2019.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

#slides-materials

#assignments

#articles-tutorials

#blog-newsletter

#random

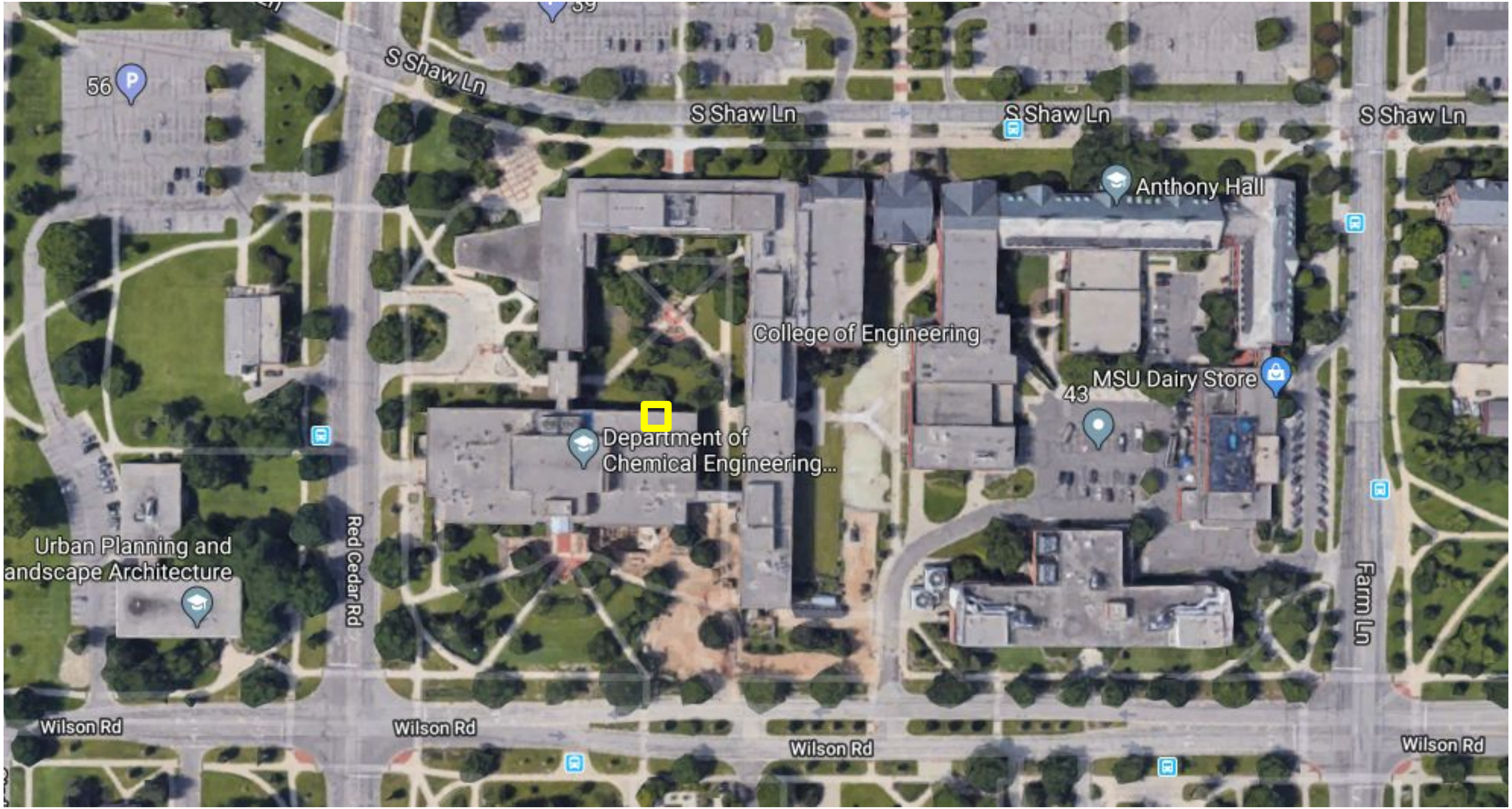
Offline discussion hours + Self-assessment

bit.ly/statgaps2019_incoming

- Select convenient hours for offline discussion
    - Will give preference to enrolled students
    - Happy to chat in-person but, many times, just messaging on Slack with your questions/concerns might work as well.
    - Happy to coordinate if you can't make it during this window for some reason. Again, just send message me on Slack.

Interest survey: bit.ly/statgaps2019_signup

My office: 2507H Engineering Building (2nd floor)



Course activities

- Assignments: ~25%
- Class participation: ~50%
- Final exam: ~25%

- Weekly blog/newsletter

Assignments

- Will be posted each Wednesday on Slack
- 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

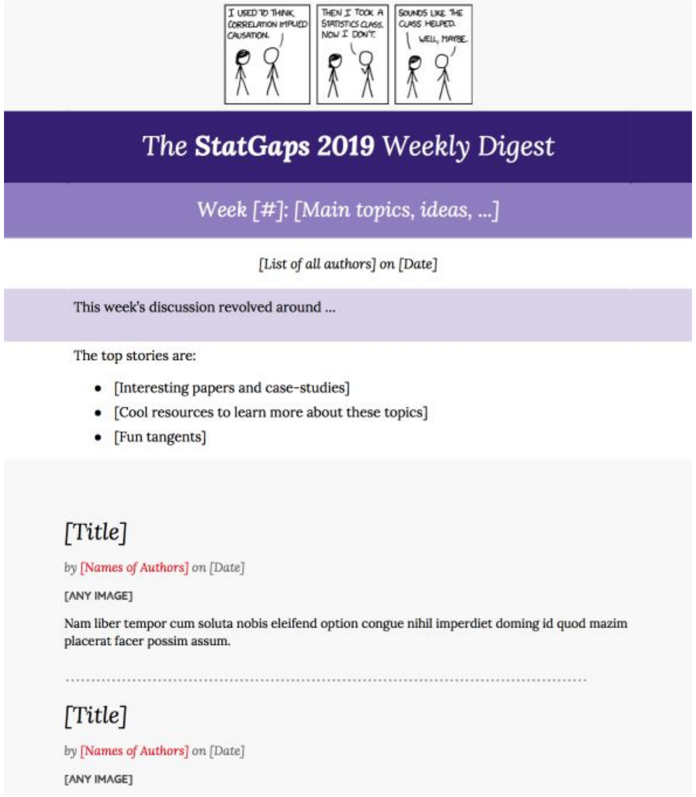
- Grand Rapids:
  - 5005 Grand Rapids Research Center
    - *Except Nov 20* when it in Room 2005
  - Join the #grand-rapids channel on Slack
  - Facilitator:
    - Joe Kochmanski, postdoc in the Bernstein Lab.
- East Lansing:
  - *Need a volunteer class facilitator*

Class participation

- Do the assignments and additional readings.
- Show up to class.
- Work in pairs/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: we ask for your active engagement with the class and its materials along with providing constructive feedback.

Weekly newsletter | 4 times

- A number of you have volunteered to contribute. You will receive a link to lead the effort once. You can contribute as many times as you like.
- Examples of good content:
  - Great learning resources
  - Fun bits of information, trivia, & asides
  - Case-studies
  - Examples of issues in stats/data-analysis in your own work


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

The **StatGaps 2019** *Weekly Digest*

Week [#]: [Main topics, ideas, ...]

[List of all authors] on [Date]

This week's discussion revolved around ...

The top stories are:

- [Interesting papers and case-studies]
- [Cool resources to learn more about these topics]
- [Fun tangents]

[Title]

by [Names of Authors] on [Date]

[ANY IMAGE]

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum.

[Title]

by [Names of Authors] on [Date]

[ANY IMAGE]

Final exam

- 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.
- The final “exam” will give you an opportunity to revisit many of the concepts discussed throughout the class and, in that process, do something practically useful to you in your future efforts with statistical data anlayses.
- We will discuss and nail the details when we meet in class.

My role

- <u>The most underrated part of teaching is learning</u>. 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.

Day 01

Welcome, Overview, Getting started

Welcome, overview

- Introductions
- Scope & topics
- Website / Communication
- Course Activities

Getting started...

- Coding
- Organizing a data analysis project
- Resources
- Wrap-up

Coding

You will be writing code to:

- read-in datasets,
- wrangle them into a convenient format,
- calling common statistical functions from standard packages/libraries to calculate mean, std. deviation, quantiles, correlation, etc.
- implementing some simulations/tests
  - random number generation
  - writing for/while loops
- making 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

Organizing a data analysis project

`project_directory`

- `data`
  - primary & processed data + `readme.txt` + `runlog.sh`
- `src`
  - all your code/scripts
- `bin`
  - all compiled code + installed binaries + `readme.txt`
- `doc`
  - literature notes + analysis notes + intermediate/final report
- `results`
  - YYYY-MM-DD sub_directories
    - `runlog.sh` + R/Python notebooks

Based on Noble (2006) PLoS Comp. Biol.

Organizing a data analysis project

`project_directory`

- `data`
  - primary & processed data + `readme.txt` + `runlog.sh`

No manual editing of data; Write scripts

Details on when & where data was downloaded

No code in this dir; Should point to & run code from `src`; this file should have all the command-lines used to run the code/scripts to process data here

- `src`
  - all your code/scripts

Including those used for data download, processing, and analysis; Well documented with detailed comments within the code + external documentation.

- `bin`
  - all compiled code + installed binaries + `readme.txt`

Details on when and from where external software was downloaded; also include installation instructions if it was not straightforward.

- `doc`
  - literature notes + analysis notes + intermediate/final report
- `results`
  - YYYY-MM-DD sub_directories
    - `runlog.sh` + R/Python notebooks

Based on Noble (2006) PLoS Comp. Biol.

Organizing a data analysis project

`project_directory`

- `data`
  - primary & processed data + `readme.txt` + `runlog.sh`
- `src`
  - all your code/scripts
- `bin`
  - all compiled code + installed binaries + `readme.txt`
- `doc`
  - literature notes + analysis notes + intermediate/final report dir
- `results`
  - YYYY-MM-DD sub_directories
    - `runlog.sh` + R/Python notebooks

One file named with YYYY-MM-DD date of each analysis; Should contain free-text details on the thoughts/ideas behind that day's analyses.

Used at the later stages of a project to pull all the results into a report/paper.

At each stage of an analysis, gather your results (as text files) & make plots to visualize & interpret.

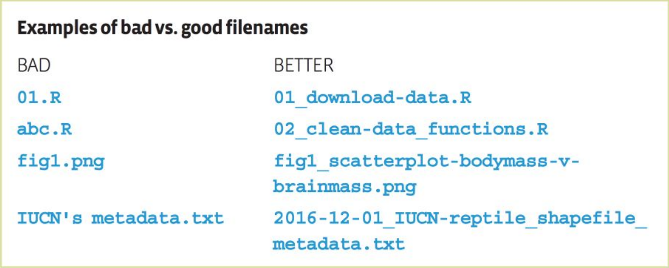
Should point to & run code from `src`; This file should have all the command-lines used to run the code/scripts to produce the results here.

Based on Noble (2006) PLoS Comp. Biol.

Managing data and code

Data

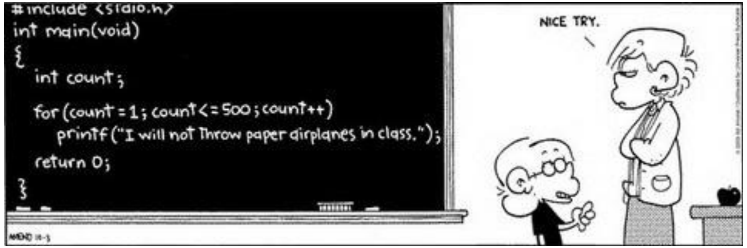
- Give all files meaningful, interpretable, & computable names
    - Machine readable, human readable, works well with default ordering.
- Do not tamper with original/source files
    - `readme.txt` should contain detailed information about when & from where each piece of data was obtained.
- Do not make changes by hand; Automate everything
    - Write scripts that read in the file and generates the desired file.
- Document everything
    - Keep track of all your commands (Linux & running code) in a `runlog.sh`.



Examples of bad vs. good filenames

BAD	BETTER
`01.R`	`01_download-data.R`
`abc.R`	`02_clean-data_functions.R`
`fig1.png`	`fig1_scatterplot-bodymass-v-brainmass.png`
`IUCN's metadata.txt`	`2016-12-01_IUCN-reptile_shapefile_metadata.txt`

https://speakerdeck.com/jennybc/how-to-name-files



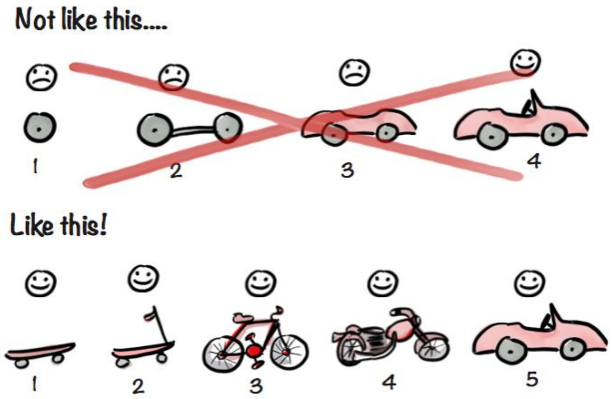
```
#include <stdio.h>
int main(void)
{
  int count;
  for (count = 1; count <= 500; count++)
     printf("I will not Throw paper airplanes in class.");
  return 0;
}
```

Managing data and code

Code

- Write code for both computers & humans.
  - Give descriptive, interpretable variable & function names.
  - Comment your code at the top: purpose, expected usage, example inputs/outputs, dependencies.
  - Record imports, constants, random seeds at the top.
  - Comment each block/function: the intended computation, arguments, return values.
- Properly acknowledge code borrowed from elsewhere; Check license.
- Program for the general case, and put the specifics outside the code as arguments & parameters.

Continuously functional & testable



Spotify

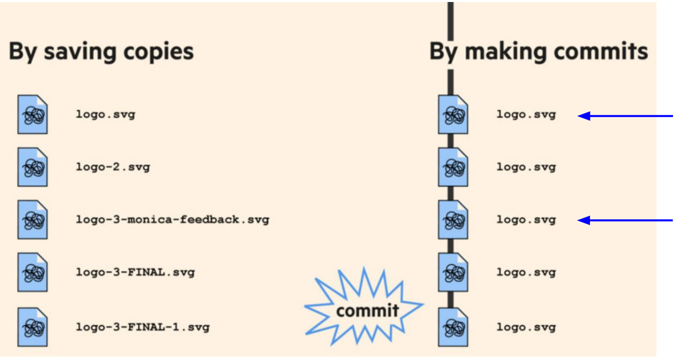
twitter.com/JennyBryan/status/952285541617123328

One of the most useful things I've learned from hanging out with (much) better programmers: don't wring hands and speculate. Work a small example that reveals, confirms, or eliminates something.

Managing data and code

Version control

- Storify your project
- Travel back in time
- Experiment with changes
- Backup your work
- Collaborate effectively



Arjun Krishnan
12:34pm January 3th 2018

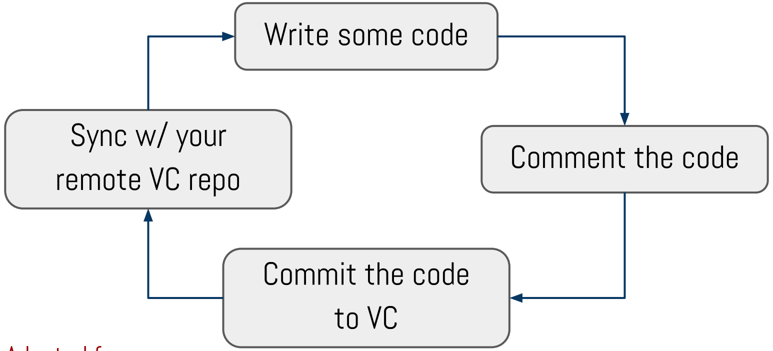
Updated background color

Changed background color to improve contrast.

Arjun Krishnan
9:15am January 4th 2018

Incorporated feedback from team

Made all changes based on team.org/feedback314



repository	Your project folder
commit	A snapshot of your repo
remote	A computer with the repository on it
clone	Get the repository from the remote for the first time
push	Send commits to a remote
pull	Get commits from a remote
merge	Combine two branches

Adapted from
@alicebartlett

Open science

Code: The field has dramatically shifted in thinking on how to publish code.

- Code used in research should be made available for research use free of charge.
- This is not just code for downloading & using. Original code must be made publicly available for others to use, review, and edit.
- Most common way to share code: GitHub.

Scientific publishing: Preprints

- Rapid publication of new science + free access (e.g. bioRxiv).
- Major source of cutting-edge research.
- Can have multiple (progressively better) versions of each manuscript.
- Preprints have NOT been peer-reviewed for quality and soundness of science. So, read/use with caution.

Resources @ MSU

Center for Statistical Training and Consulting

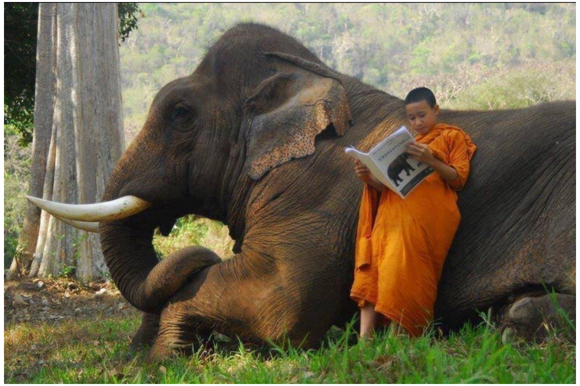
- Training resources: https://cstat.msu.edu/resources
- Events and workshops: https://cstat.msu.edu/events

Working/student groups

- R-Ladies: https://rladies-eastlansing.github.io/
- MSU Data Science: http://msudatascience.com/

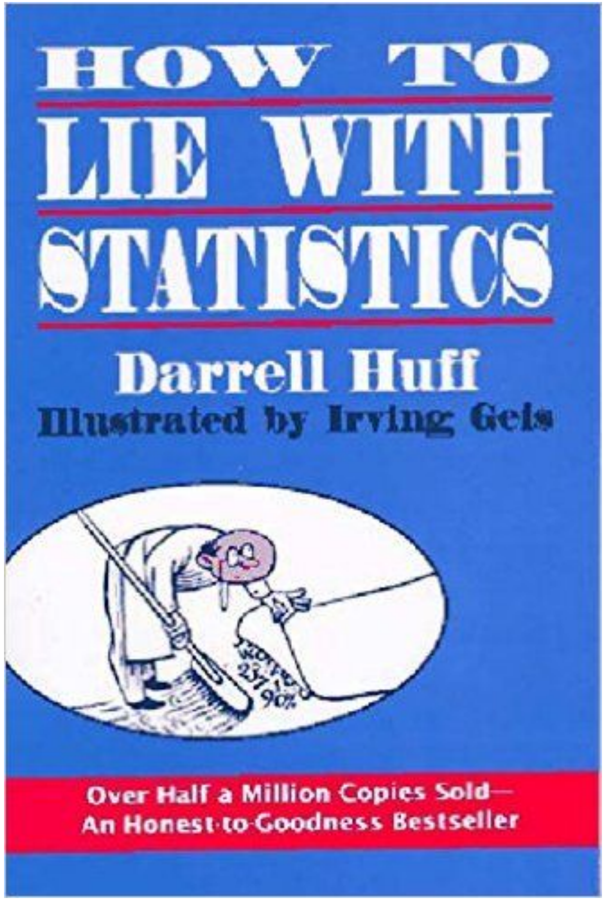
Getting help

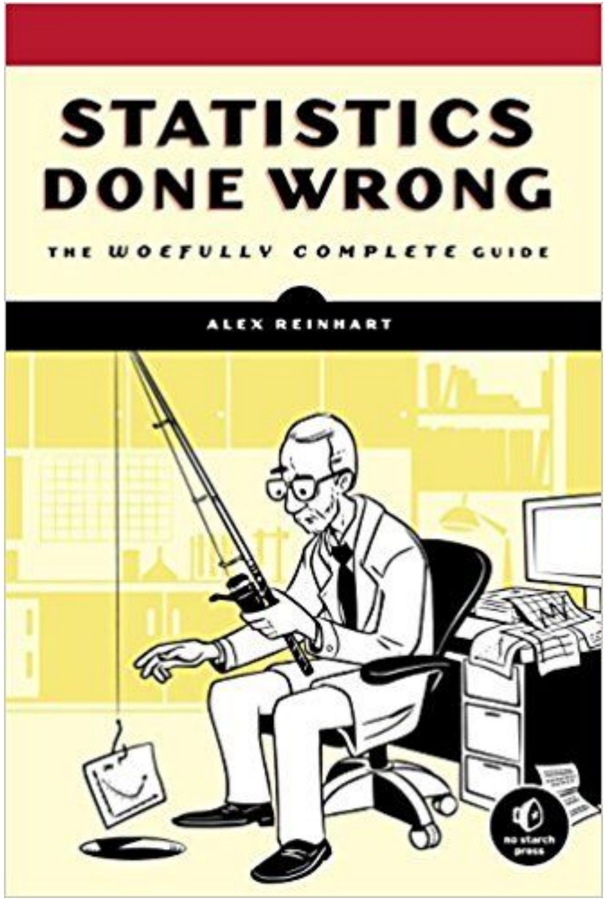
- **Linux** | rik.smith-unna.com/command_line_bootcamp, commandline.guide, & swcarpentry.github.io/shell-novice
- **Python** | Introduction: learnpythonthehardway.org/book & developers.google.com/edu/python | Data analysis: jakevdp.github.io/WhirlwindTourOfPython | Visualization: www.r-graph-gallery.com
- **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

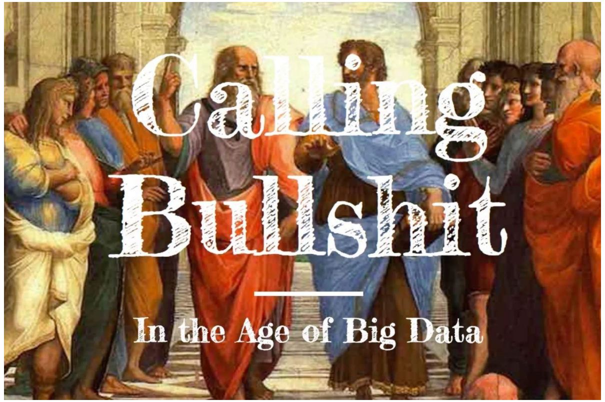


Google ... so many excellent blog posts!

Resources


HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller





Original research articles

Reviews

Blog posts

Podcasts

What you need to do before the next class

- R and Python (instructions will be posted on Slack)
    - Install R, RStudio, and Tidyverse (package); Get familiar with R Notebooks
    - Install Anaconda, Python 3.7, Jupyter Notebooks

- Concepts (resources will be posted on Slack & the class website)
    - Brush-up: Standard deviation/error, Confidence interval
    - Brush-up: Hypothesis testing and P-values

What you need to do before the next class

- Look out for messages on all channels: stagaps2019.slack.com
    - Instructions for next week
    - Blog/Newsletter sign-up sheet

- Read the course website: bit.ly/statgaps2019

- Fill-in the incoming + self-assessment survey: bit.ly/statgaps2019_incoming
    - If you haven't done so already, fill-in the interest survey: bit.ly/statgaps2019_signup