BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 1: INTRODUCTION AND THINKING ABOUT DATA

COURSE DETAILS

- Sept. 17th to November 5th
  - Due to pre-planned travel, there will be no class on Sept. 24th, October 1st, or Oct. 22nd
  - These lectures will also be recorded (still working on a Canvas site)
- Office hours are by request
  - Email: Eric.Lofgren@wsu.edu
- Grading:
  - Three problem sets, each worth 13.33% of your grade (40% total)
  - *Participation* is worth 30% of your grade
    - This class calls for active and engaged learning, and you are not guaranteed full participation marks
    - If you need an accommodation in this area, please arrange a meeting
  - A final project, worth 30% of your grade

PROBLEM SETS

- Problem sets will be assigned at the end of a Wednesday class, due *before* the start of class the next Wednesday
- Collaboration is encouraged on all problem sets
- While you can work in groups, the work must be individual to you – that is, it should be written in your own words, reflecting your own understanding of the solutions your group arrived at
- If you did work with a group, please list the names of the other students you worked with

THE FINAL

- The final project is to select a topic that is of interest to you, based on your own work, or as part of the material that we've covered.
- Identify a problem
- Discuss the prior research in this area – what is yet to be solved?
- Outline how a statistical approach might aid you in answering it
- All in the form of a 5-page paper
  - One-inch margins, reasonable font sizes, etc. apply.
- The final must be completed independently
- This is due two-weeks after our last class on Wednesday Nov. 19th at 11:59 PM PST.

LATE WORK

- Assignments submitted after the beginning of class will receive one-half of the total graded points
- A late *final* will lose 10% off the final grade for each day it is late, beginning on Nov. 20th at 12:01 AM PST.

ACADEMIC HONESTY

- Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course. Students who violate WSU's Academic Integrity Policy (identified in Washington Administrative Code (WAC) 504-26-010(3) and -404) will fail the course, will not have the option to withdraw from the course pending an appeal, and will be reported to the Office of Student Conduct.

- Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students, WAC 504-26-010(3). You need to read and understand all of the definitions of cheating: http://app.leg.wa.gov/WAC/default.aspx?cite=504-26-010. If you have any questions about what is and is not allowed in this course, you should ask course instructors before proceeding. If you wish to appeal a faculty member's decision relating to academic integrity, please use the form available at https://conduct.wsu.edu/.

ARTIFICIAL INTELLIGENCE

- Artificial intelligence is an evolving field, with a great deal of promise and a great many pitfalls
- At the moment, LLMs *cannot* look up information – they use statistical models to predict what this sort of thing *should* be
- They excel at giving you answers that sound right, but are not
- At the moment, LLMs are extensively trained on data where the authors of that data have not given their consent, been compensated, etc.
- AI tools also consume a tremendous amount of both electricity and fresh water, and have a substantial environmental footprint
- I will not prohibit the usage of AI tools in my class (if for no other reason than this is a rabbit hole of "what counts"), but use with caution

WRITE ME A SPEAKER INTRODUCTION FOR ERIC LOFGREN, THE EPIDEMIOLOGIST. DO NOT USE PLACEHOLDERS, AND USE SPECIFICS FOR WHERE HE WORKS, AND HIS AREAS OF EXPERTISE.

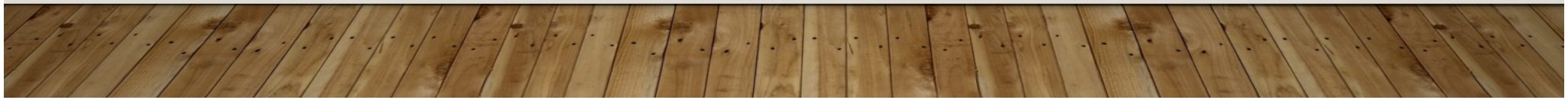
Ladies and gentlemen,

It is my distinct pleasure to introduce our next speaker, Dr. Eric Lofgren. Dr. Lofgren is an esteemed epidemiologist and assistant professor at the School of Public and International Affairs at Virginia Tech. He holds a Ph.D. in Epidemiology from the University of California, Berkeley, where he specialized in the mathematical modeling of infectious diseases.

Dr. Lofgren's research has made significant contributions to our understanding of epidemic dynamics and the global impact of infectious diseases. His expertise in mathematical modeling is crucial for forecasting and managing epidemic outbreaks, with a particular focus on diseases such as influenza and emerging pathogens. His innovative models help guide public health officials and policymakers in crafting effective response strategies.

In addition to his groundbreaking research, Dr. Lofgren is dedicated to bridging the gap between academic research and practical public health applications, ensuring that scientific discoveries are translated into actionable policies. His work not only advances the field of epidemiology but also plays a pivotal role in shaping responses to current and future public health challenges.

Please join me in welcoming Dr. Eric Lofgren.



WRITE ME A SPEAKER INTRODUCTION FOR ERIC LOFGREN, THE EPIDEMIOLOGIST. DO NOT USE PLACEHOLDERS, AND USE SPECIFICS FOR WHERE HE WORKS, AND HIS AREAS OF EXPERTISE.

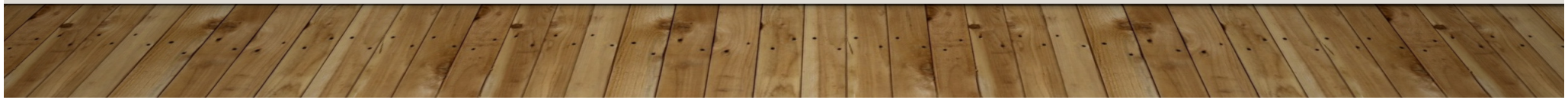
Ladies and gentlemen,

It is my distinct pleasure to introduce our next speaker, Dr. Eric Lofgren. Dr. Lofgren is an esteemed epidemiologist and **assistant professor** at the **School of Public and International Affairs at Virginia Tech**. He holds a **Ph.D. in Epidemiology** from the **University of California, Berkeley**, where he specialized in the **mathematical modeling of infectious diseases**.

Dr. Lofgren's research has made significant contributions to our understanding of epidemic dynamics and the global impact of infectious diseases. His expertise in mathematical modeling is crucial for forecasting and managing epidemic outbreaks, with a particular focus on **diseases such as influenza and emerging pathogens**. His innovative models help guide public health officials and policymakers in crafting effective response strategies.

In addition to his groundbreaking research, Dr. Lofgren is dedicated to bridging the gap between academic research and practical public health applications, ensuring that scientific discoveries are translated into actionable policies. His work not only advances the field of epidemiology but also plays a pivotal role in shaping responses to current and future public health challenges.

Please join me in welcoming Dr. Eric Lofgren.

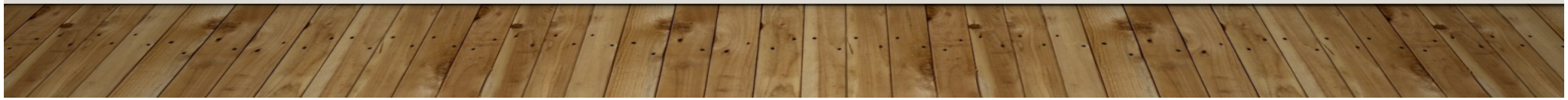


WRITE ME A SPEAKER INTRODUCTION FOR ERIC LOFGREN, THE EPIDEMIOLOGIST. DO NOT USE PLACEHOLDERS, AND USE SPECIFICS FOR WHERE HE WORKS, AND HIS AREAS OF EXPERTISE.

It is our pleasure to introduce Dr. Eric Lofgren, a renowned epidemiologist at the Washington State University Paul G. Allen School for Global Health, where he serves as an Associate Professor. Dr. Lofgren received his PhD in Epidemiology from the University of California, Davis, and has since become a leading expert in the field of zoonotic diseases, with a particular focus on the dynamics of disease transmission at the animal-human interface. His research has taken him to various parts of the world, from studying the epidemiology of influenza in Southeast Asia to investigating the risk factors of disease transmission in African wildlife. Today, Dr. Lofgren will share his insights on the complexities of emerging infectious diseases and global health security, drawing on his extensive experience in the field and his passion for improving health outcomes worldwide. Please join us in welcoming Dr. Eric Lofgren.

WRITE ME A SPEAKER INTRODUCTION FOR ERIC LOFGREN, THE EPIDEMIOLOGIST. DO NOT USE PLACEHOLDERS, AND USE SPECIFICS FOR WHERE HE WORKS, AND HIS AREAS OF EXPERTISE.

It is our pleasure to introduce Dr. Eric Lofgren, a renowned epidemiologist at the **Washington State University Paul G. Allen School for Global Health**, where he serves as an **Associate Professor**. Dr. Lofgren received his **PhD in Epidemiology** from the **University of California, Davis**, and has since become a **leading expert in the field of zoonotic diseases**, with a particular focus on the dynamics of disease transmission at the animal-human interface. His research has taken him to various parts of the world, from studying the epidemiology of **influenza in Southeast Asia** to investigating the risk factors of **disease transmission in African wildlife**. **Today, Dr. Lofgren will share his insights on the complexities of emerging infectious diseases and global health security**, drawing on his extensive experience in the field and his passion for improving health outcomes worldwide. Please join us in welcoming Dr. Eric Lofgren.



LAUREN'S PROMISE

- On October 22, 2018, Lauren McCluskey, 21 years old, was murdered by a man she briefly dated on the University of Utah campus, where she was a student. Lauren was raised in Pullman, Washington. Together with her parents, who are professors at WSU, this university community stands firmly behind Lauren's Promise: **WSU will listen and facilitate support and reporting options if someone is threatening you.**
- WSU prohibits discrimination and harassment. This includes discriminatory harassment, hate crimes, sexual discrimination, sex-based harassment, stalking, dating violence, domestic violence, sexual assault, and all types of sexual violence.
- If you are in immediate danger, call 911.
- If you have experienced or have witnessed discriminatory behavior, you can contact the WSU Compliance and Civil Rights (CCR) and/or the WSU Title IX Coordinator. CCR can provide information on reporting options, including confidential resources available to you, and facilitate supportive measures. To contact CCR:
- Online: Online Reporting Form
- Email: ccr@wsu.edu
- Phone: 509-335-8288
- For more information, see the WSU Policy Prohibiting Discrimination and Harassment (Executive Policy 15), WSU Standards of Conduct for Students (Chapter 504-26 WAC), and the WSU Notice of Nondiscrimination.

ON EMAIL…

- Faculty are very busy people, and their inboxes are usually flooded
- I have 300+ unread emails on an ordinary Monday
- If I haven't answered you, it is 100% not because I hate you, think your email is dumb, or am silently judging you – it's because I looked at it, didn't get a chance to answer, and it got buried
- *Please* if I have not responded to an email from you about the course, email me *again*
- I will never be upset about a gentle reminder or nudging something urgent to the top of the pile

A FINAL DISCLAIMER

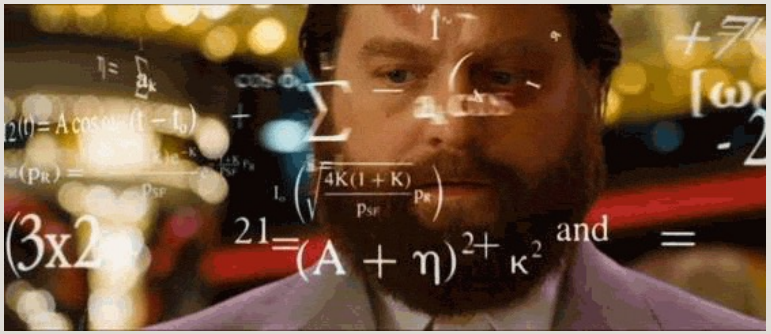
- There will be a great many Elder Millennial cultural references in these lectures
- I recognize that over time these will be less intelligible to my students
- I'm not going to stop

WHAT IS "STATISTICAL THINKING"?

- "The great body of physical science ... [is] only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen ... it is necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write." – H.G. Wells

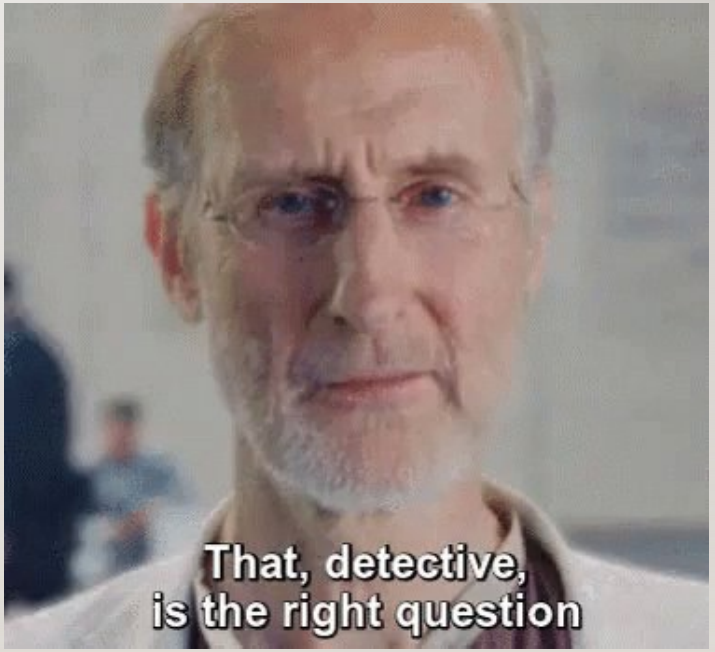
WHAT IS "STATISTICAL THINKING"?

- Statistical thinking is both a toolset and a mentality for looking at the world in a way where we are trying to understand both the world's underlying processes – and that there is uncertainty surrounding those processes
- It is not just *doing* statistics, but understanding why, and approaching the design of experiments in a way that will yield meaningful data



WHY THIS CLASS?

- An essential element of becoming an independent scientist is being able to articulate not only *what* you did, but why
- In your defense, etc. "Because [My PI] told me to…" is not really a sufficient answer
- I have seen grant proposals, etc. dinged for being vague about statistics and taking a "we'll sort that out later" approach
- If you think about statistics from the outset of your research, you are likely going to generate better data that will be easier to analyze, and give you a clearer picture of what you're doing
  - It'll also help you ask the right question

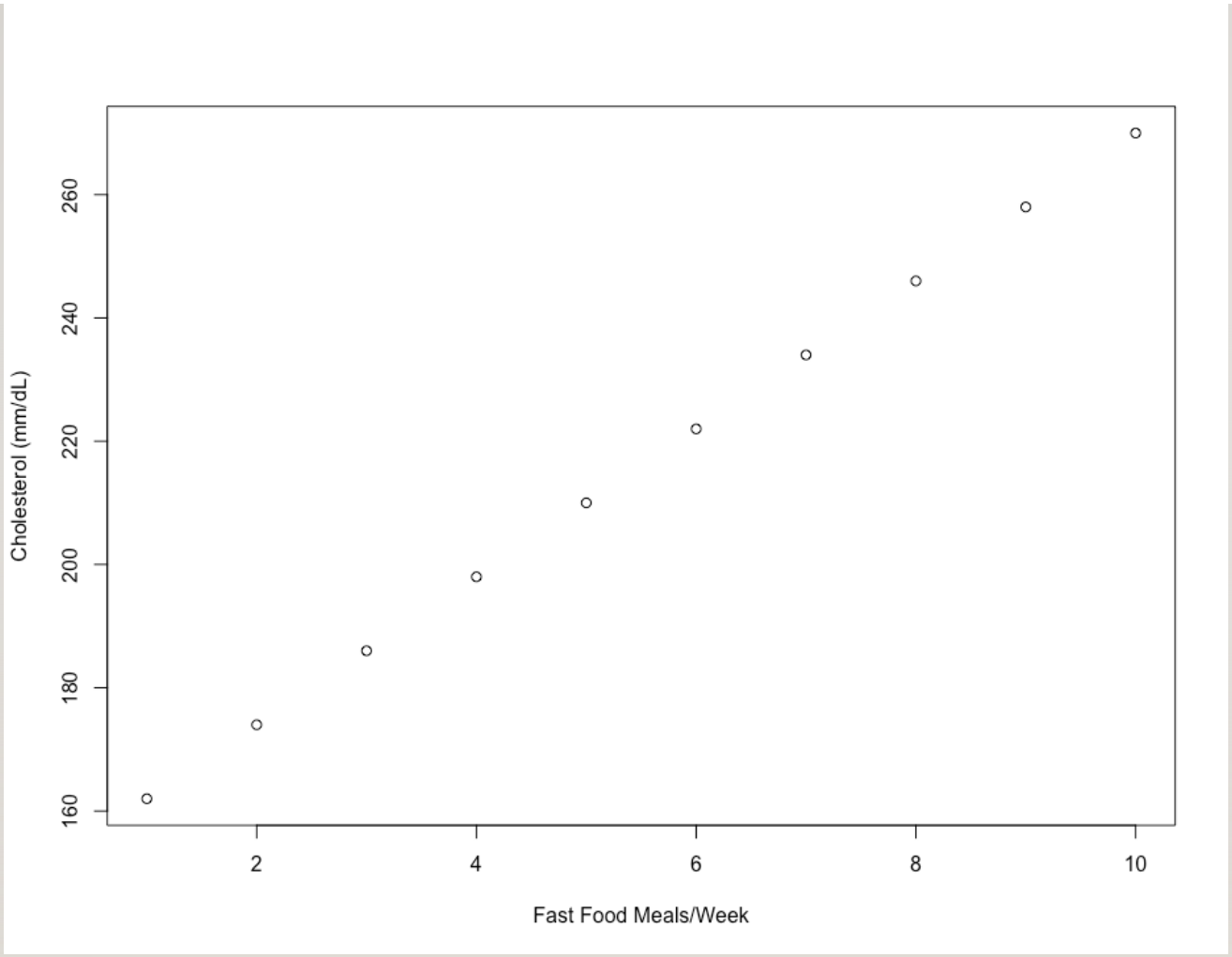


WHAT THIS CLASS ISN'T

- A "Stats" class
  - A single credit class isn't nearly enough to teach you the mechanics of statistical analysis in the biological sciences
  - Arguably, a *series* of classes isn't necessarily enough

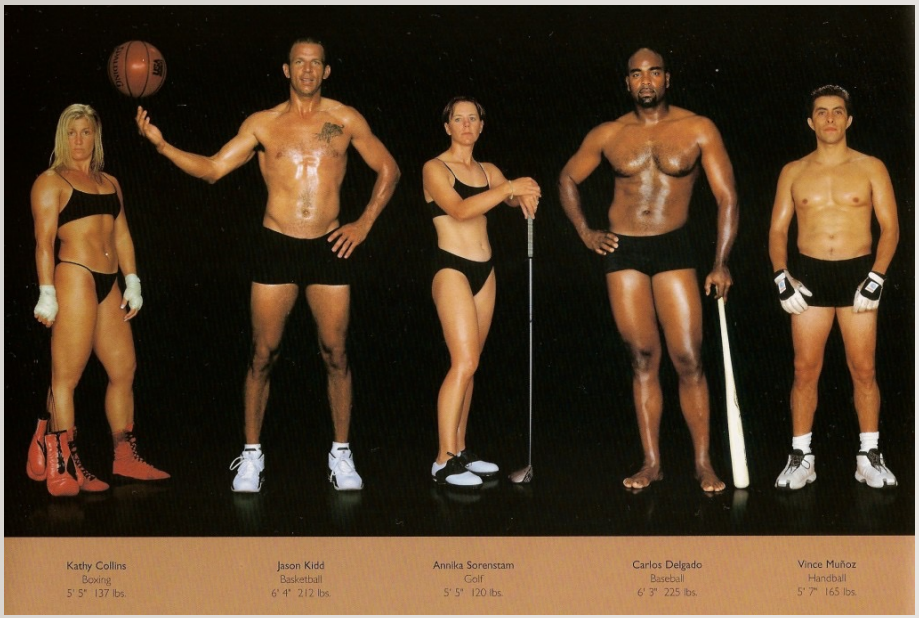
WHY DO WE NEED STATISTICS AT ALL?

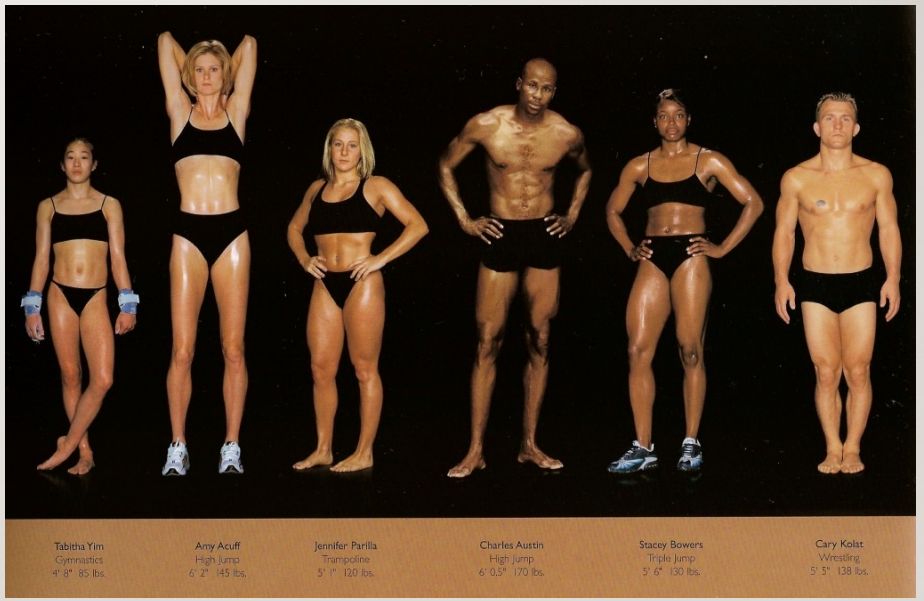
- A thought experiment: Does the consumption of fast-food cause high cholesterol
  - Measure cholesterol levels in human blood
  - Subjects self report the number of fast food meals they eat
  - Compare the cholesterol level in the blood by how many self reported meals


Cholesterol (mm/dL)
160
180
200
220
240
260
2
4
6
8
10
Fast Food Meals/Week

VARIATION IN THE BIOMEDICAL SCIENCES

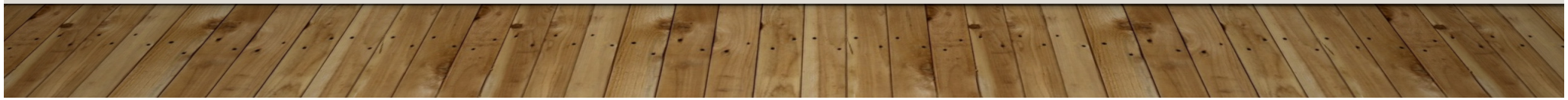
- Biomedical science studies living organisms
- There's inherent variation in said organisms
- No two individuals are ever identical, and it's very rare that we can measure every individual we're interested in
- If you look into the history of statistics, a *lot* of it is motivated by biology – the two fields have always been connected

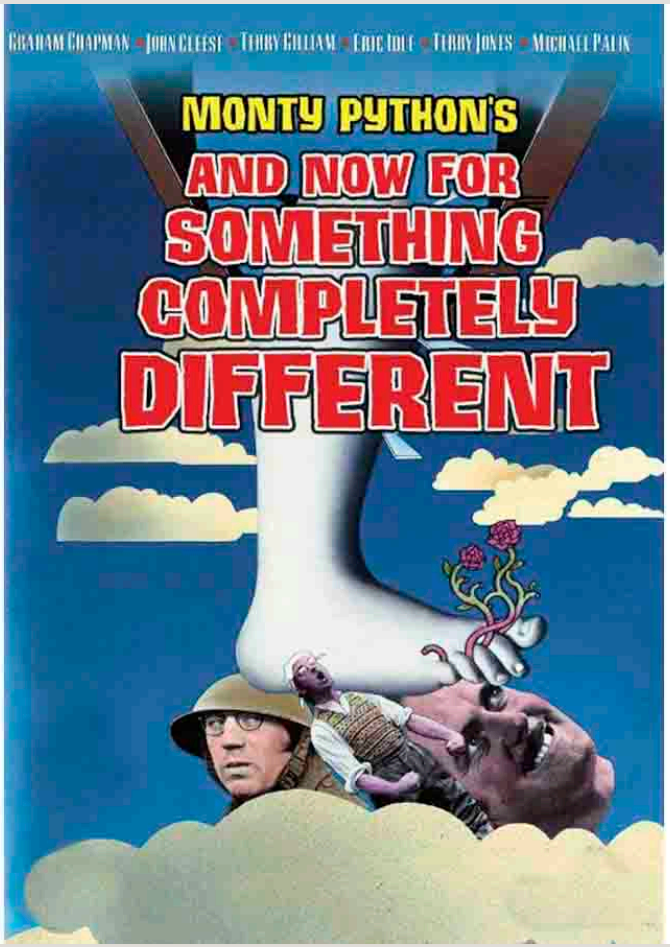




"The Athlete" by Howard Schatz and Beverly Ornstein, 2002

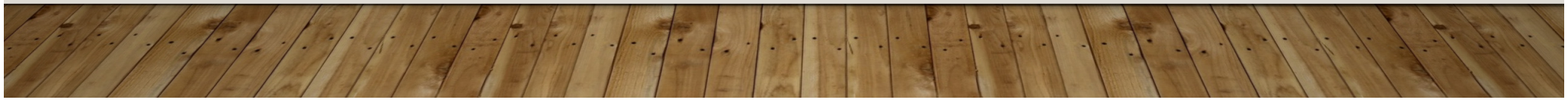
QUESTIONS SO FAR?




GRAHAM CHAPMAN • JOHN CLEESE • TERRY GILLIAM • ERIC IDLE • TERRY JONES • MICHAEL PALIN
MONTY PYTHON'S
AND NOW FOR
SOMETHING
COMPLETELY
DIFFERENT

COMPUTATIONAL TOOLS IN DATA

- Statistics involves coding
- But you can be good at statistics and *bad* at coding
- This is, if you ask some people, the norm
- There are some basics that are good practice for research code
- Three years after you graduate, you don't want to be looking for analysis_final_final2_edited_finalfinal.R somewhere on your laptop
- We solve this with something called *version control*



GITHUB

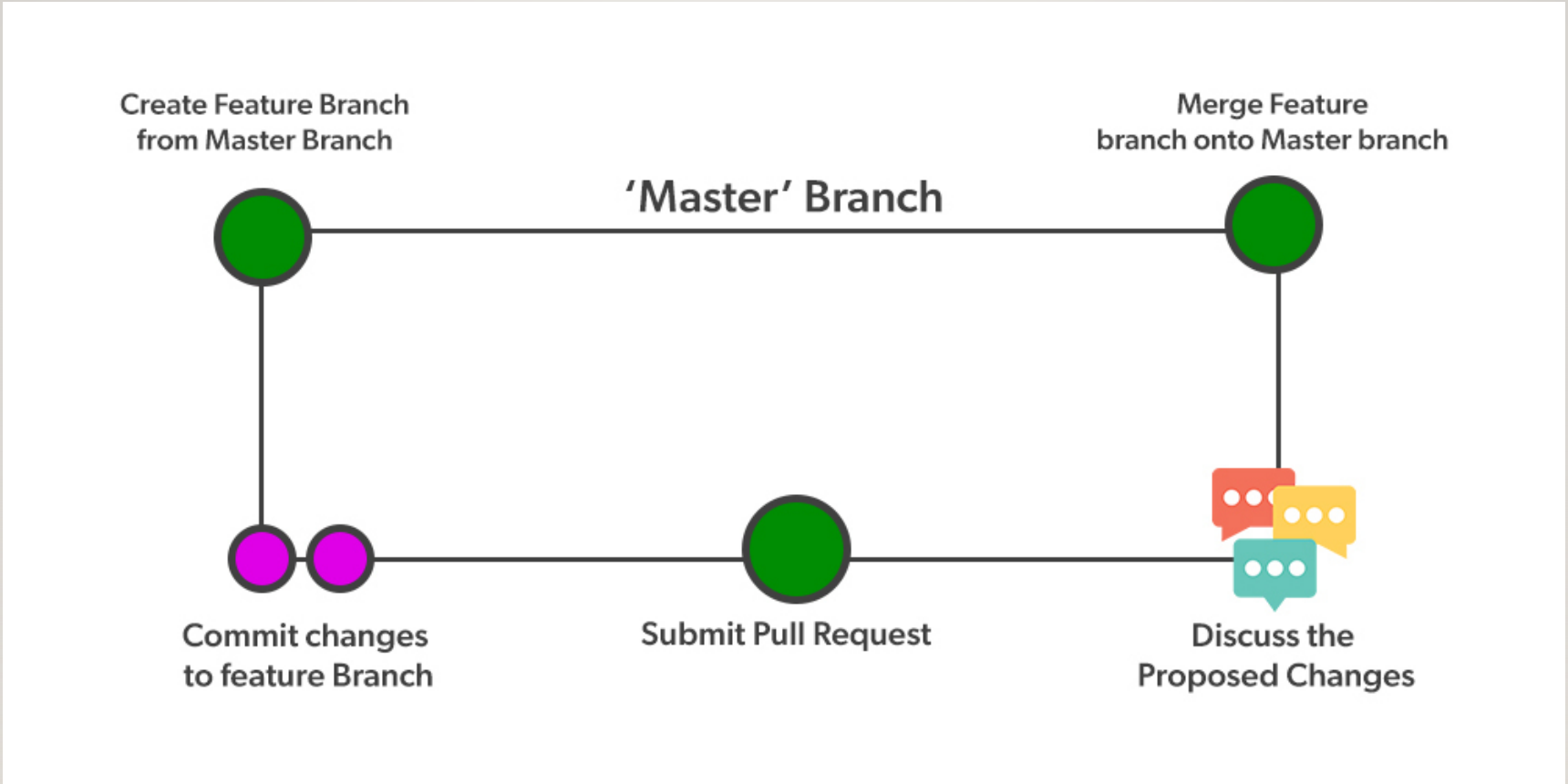
- Github.com is one of the major platforms for version control, and the one we're going to use in this class
- It is built on top of the version control system *git*
- It is *wildly* powerful, and has very sophisticated features, but even a basic use of it will improve your organization, let you undo mistakes, etc.
- Lots of resources for students

WHY USE VERSION CONTROL?

- Added redundancy for keeping your code safe
  - I used to work at an Apple store a very long time ago, you would be shocked how many times the Genius Bar got asked to try to save someone's dissertation
- Being able to go backwards to fix code, etc.
- Designed around multiple people being able to interact with a code base at the same time
  - No "can you stop editing the document so I can add my section?", etc.
- Much easier to share than email

ESSENTIAL TERMINOLOGY

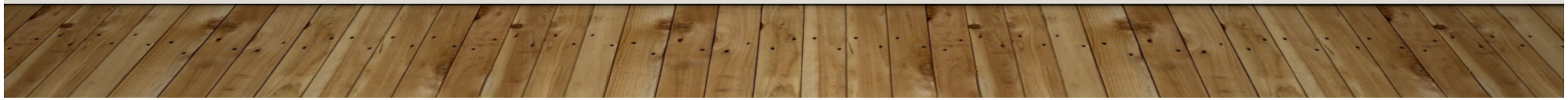
- Repository ("Repo"): A collection of code or other files, and the basic organizational unit of GitHub. Think of this like the main directory for a project
- Commit: A single addition of new code, changes, etc. to a repository
- Branch: A sort of side repository that can be worked on without making changes to the main repository, with the intention of them being reincorporated later
- Fork: Creating an entirely new repository based on an original repository where the intention is not necessarily to reincorporate them
- Push: Taking a commit and changing the main repository
- Pull Request: Asking the person who controls a repository to review and incorporate your code


Create Feature Branch
from Master Branch
Merge Feature
branch onto Master branch
'Master' Branch
Commit changes
to feature Branch
Submit Pull Request
Discuss the
Proposed Changes

GOOD GITHUB TUTORIALS

- https://docs.github.com/en/get-started
  - When in doubt, go to the source
- https://www.datacamp.com/tutorial/github-and-git-tutorial-for-beginners
  - Slightly more complex
- https://swcarpentry.github.io/git-novice/
  - Occasionally someone at WSU offers a Software Carpentry class. If you get the chance, take it
- Note: Many of these tutorials will talk about both Git and GitHub, which involves some work in the command line

DEMO

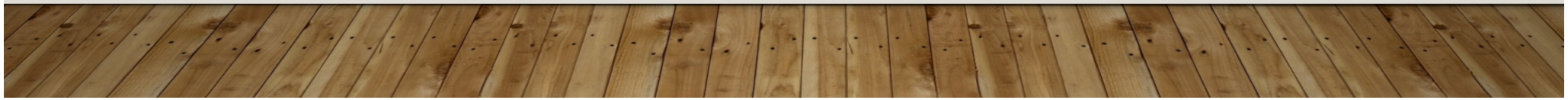


QUESTIONS?

A NOTE ON THE PROBLEM SET

- Due next Wednesday
- Again, you can work in groups, but your work should be your own
- For all problem sets, there might not be *a* right answer

WHAT IS DATA?



WHAT IS DATA?

- A "datum" is a piece of information, so it stands to reason that data is a collection of pieces of information about something
- Data has to be in some way *systematically* collected
  - Otherwise it's just anecdotes
- In the biological sciences, data usually come from *populations* and are most often *samples* of those populations
  - We're going be spending a whole class talking about sampling, but we'll cover it here briefly

WHAT'S A POPULATION?

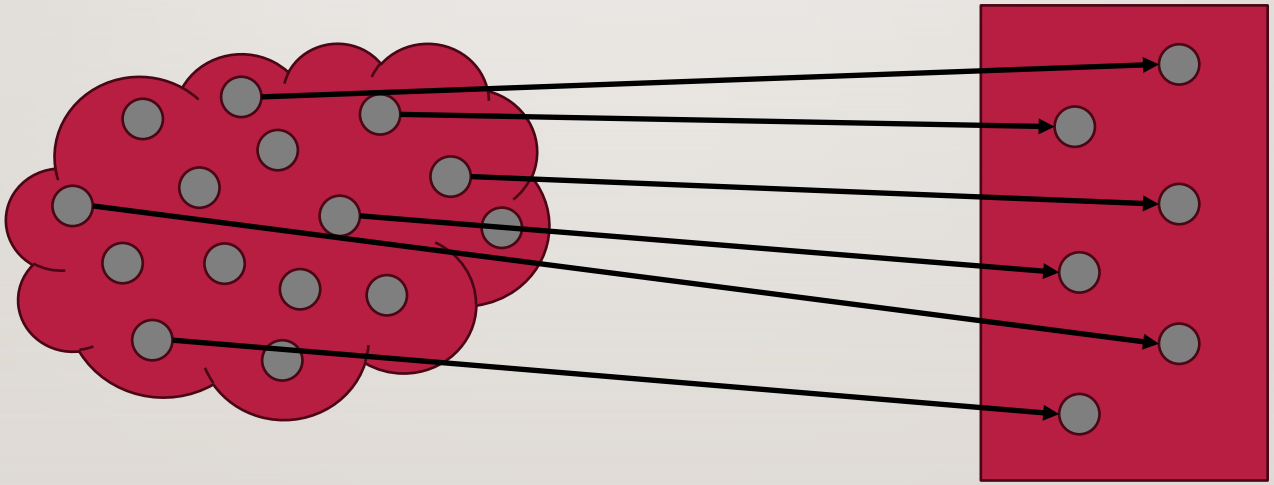
WHAT'S A POPULATION

- A group of *things* that you want to study
- Conceptually, these can be very specific, or very vague
  - *Klebsiella pneumoniae*
  - *Klebsiella pneumoniae* in intensive care units in Chicago, Illinois
  - Goats
  - Goats owned by small holder farmers in Tanzania
  - Humans
  - U.S. Marines in the 1st and 3rd Marine Expeditionary Forces
- I'm going to refer to these from now on as *source populations*

WHAT'S A SAMPLE?

WHAT'S A SAMPLE

- A smaller part of the source population that has been selected and/or is available for study



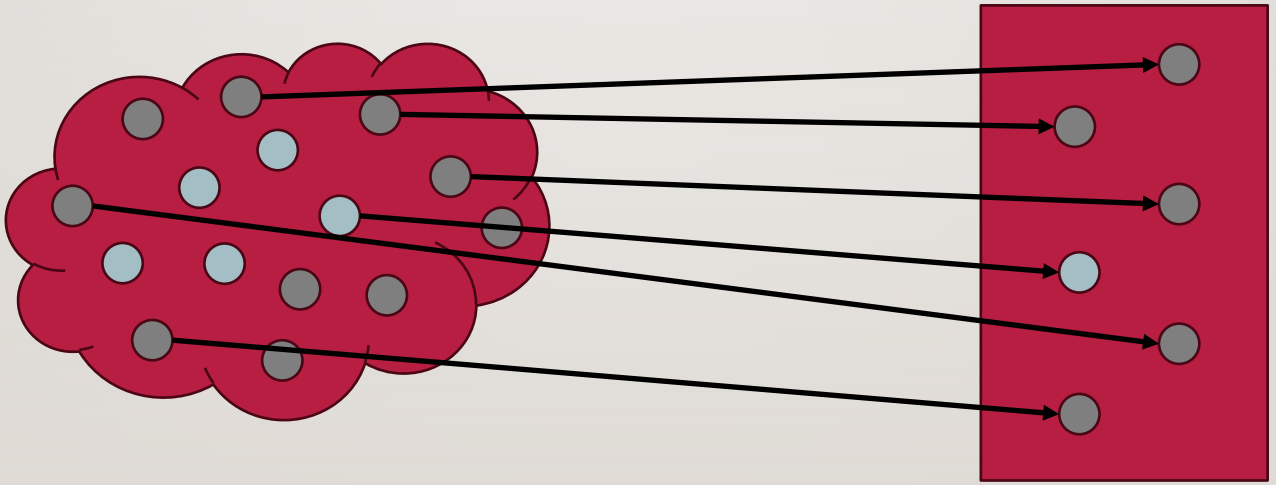
WHY DO WE NEED TO SAMPLE?

WHY DO WE NEED TO SAMPLE?

- Logistics
  - It's possible that sampling everyone in a population is simply impossible
  - It might also merely be very expensive
    - A full human genome sequence is about $600
    - There are 8.2 billion people in the world
    - $4,920,000,000,000
    - A mere 163 times the NIH budget in 2024
  - It may be hard to reach some parts of the population
    - This loops back to the expensive part
  - Presumably, you would all also like to graduate at some point
- Ethics
  - Often study participation involves some risk to the participants, and it is our ethical responsibility as researchers to minimize the number of people we expose to that risk

THE PERILS OF SAMPLING

Sampling Error

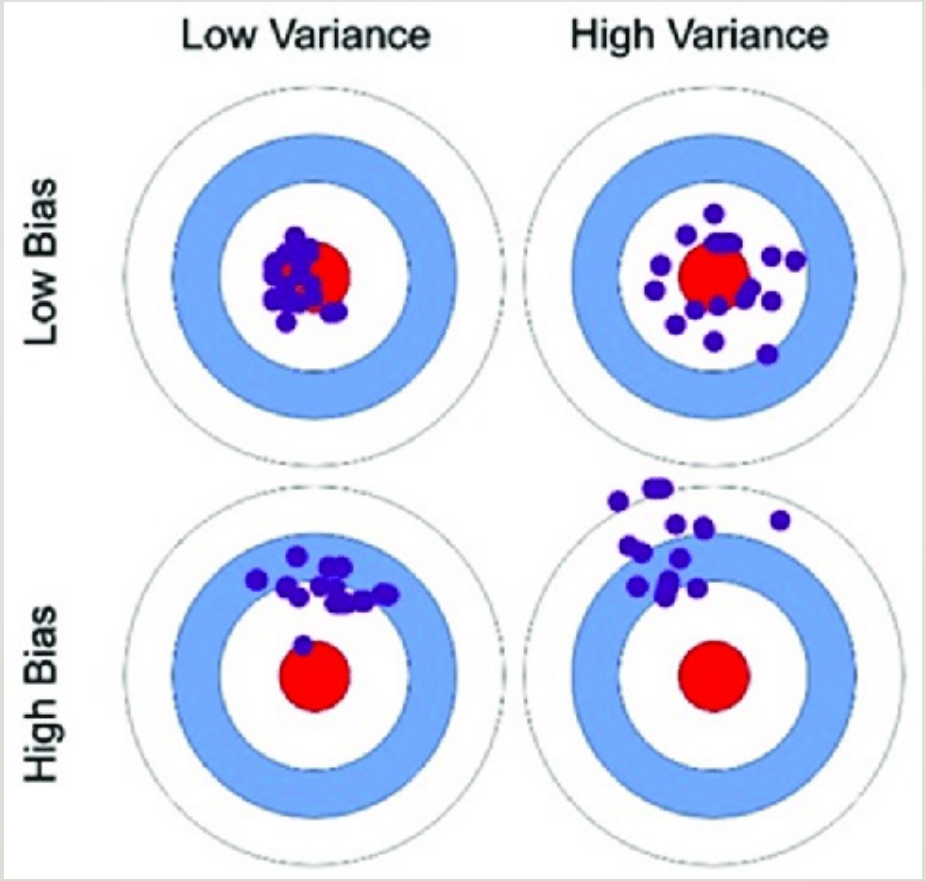


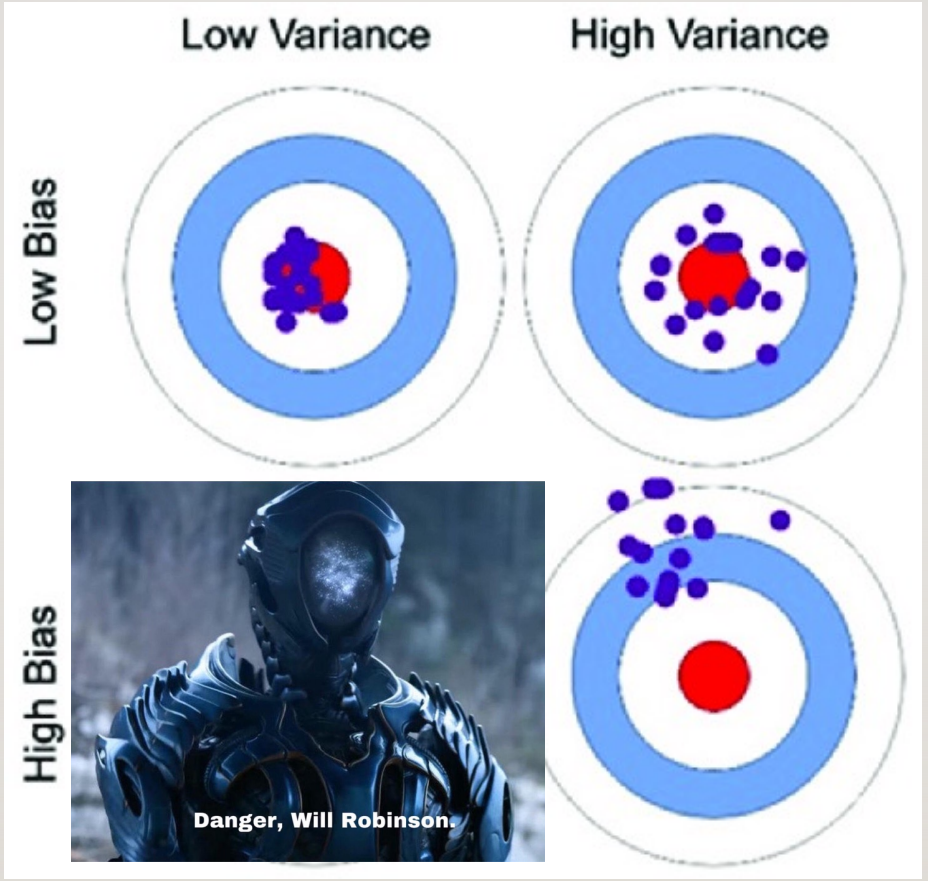
Population

Sample

SAMPLING ERROR IS OKAY!

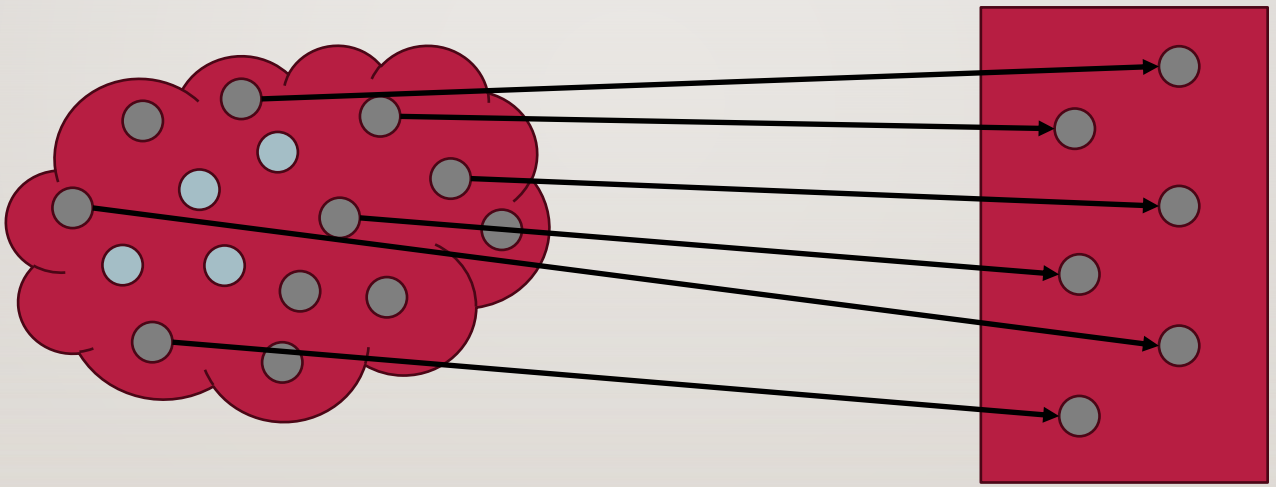
- And, to be blunt, inevitable
- Random variation pervades all of the biological sciences
- Sampling error creates *uncertainty* but not *bias*
- Bigger samples, more studies, meta-analysis, etc. can help reduce that uncertainty

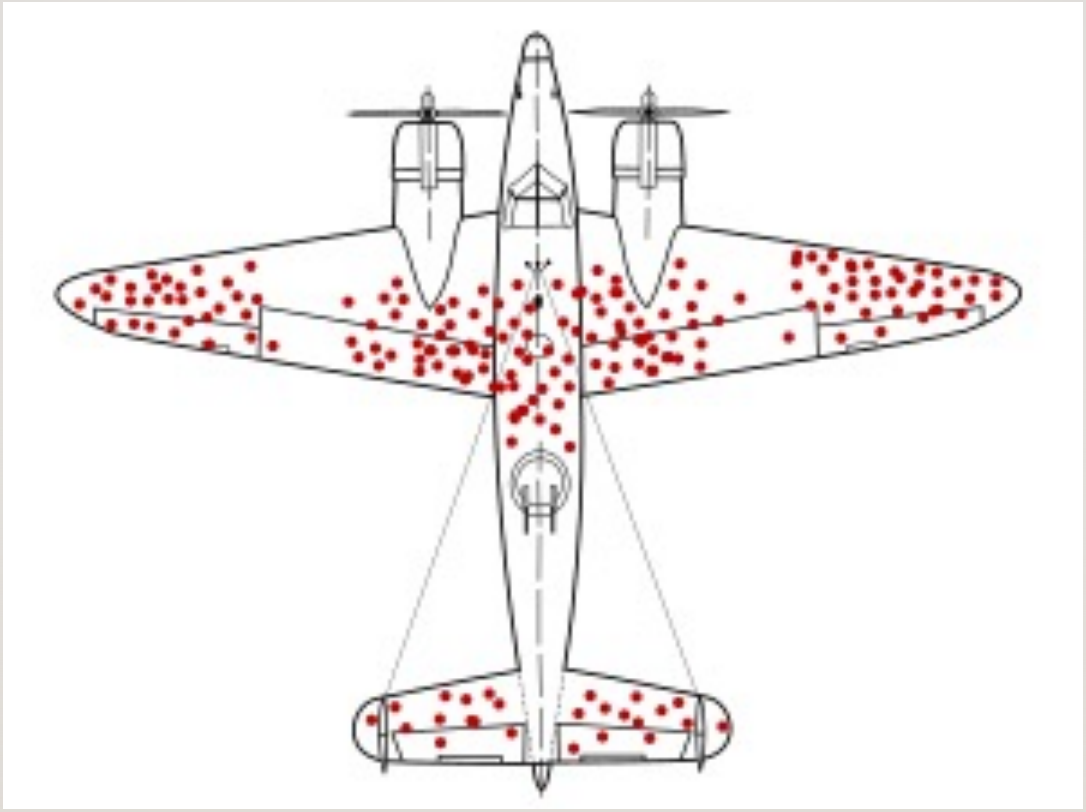

Low Variance
High Variance
Low Bias
High Bias


Low Variance
High Variance
Low Bias
High Bias
Danger, Will Robinson.

THE PERILS OF SAMPLING

Biased Sample





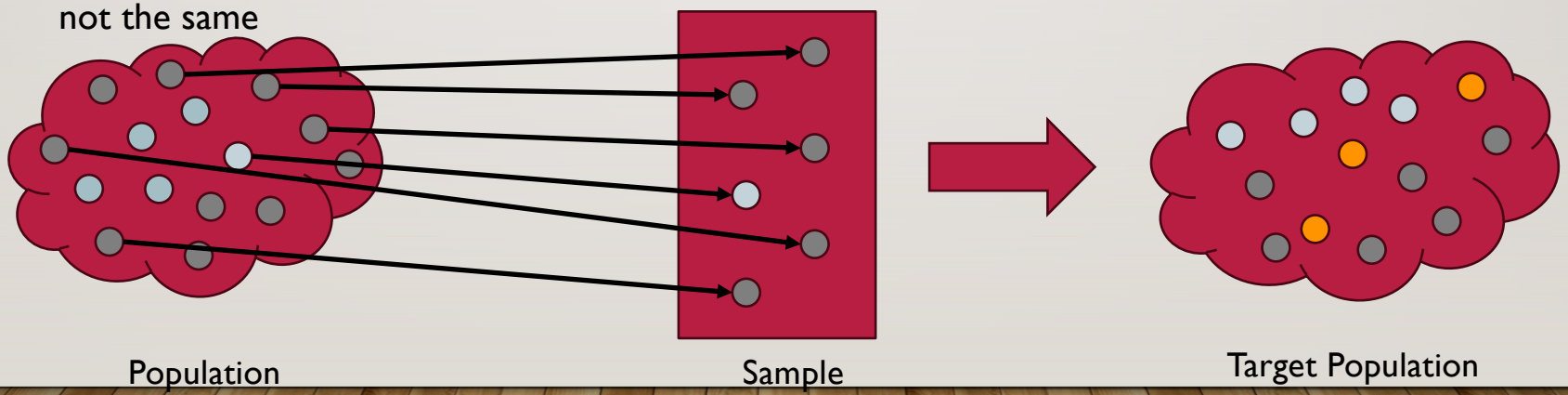

Lockheed B-34 Lexington
Cockpit and controls
Shots here could injure the pilots, or sever control lines, dooming the plane to an unavoidable crash.
Engines and radiator
Damaging these would bring the plane down quickly.
Wing extremities
Bullets hitting here would often cut clean through the wing, doing little damage more than leaving behind a nice round hole.
Midpoint of the wing
The midpoints of the wings undergo the most stress. Cannon shots here could snap the wing in half.
Self-sealing fuel tanks
When punctured, a chemical reaction in the fuel tank liner would plug up any holes. As a result, shots to the fuel tank were very survivable.
Thinnest point of plane fuselage
The waist and tail cone were the thinnest -- and often weakest -- parts of the plane's body
Gun turrets
The gunner crew were the bomber's only form of self-defense. If they were hit, the plane was left vulnerable.

SAMPLES OF CONVENIENCE

- These are samples that are *easy* for researchers to get
- A classic example is psychology studies conducted on psychology students
- Why might this be a problem?

TARGET POPULATIONS

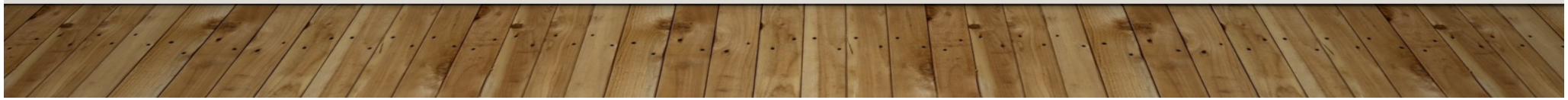
- Some studies have an additional population to think about – the target population
- This is the population we want to apply our results to
- This is easy if we just want to know things about our study population
- This can be *very* hard if the target population and the population the study is drawn from are not the same



"VALIDITY"

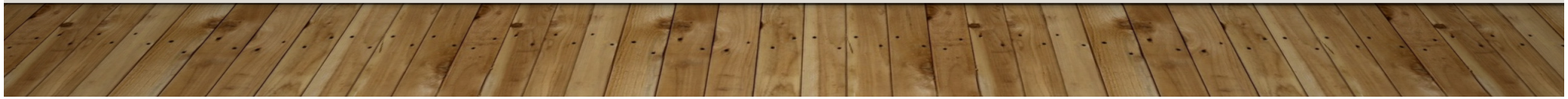
- Internal Validity: Are the results of your study unbiased – within your sample, can we be confident that your results are "correct"
- External Validity: How well can the results of your study be applied to other populations?
- Historically, we have emphasized internal validity
- Target Validity: This is a joint measure of internal and external validity
  - Relatively new concept
  - Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. Am J Epidemiol. 2019 Feb 1;188(2):438-443. doi: 10.1093/aje/kwy228. PMID: 30299451; PMCID: PMC6357801.

QUESTIONS?

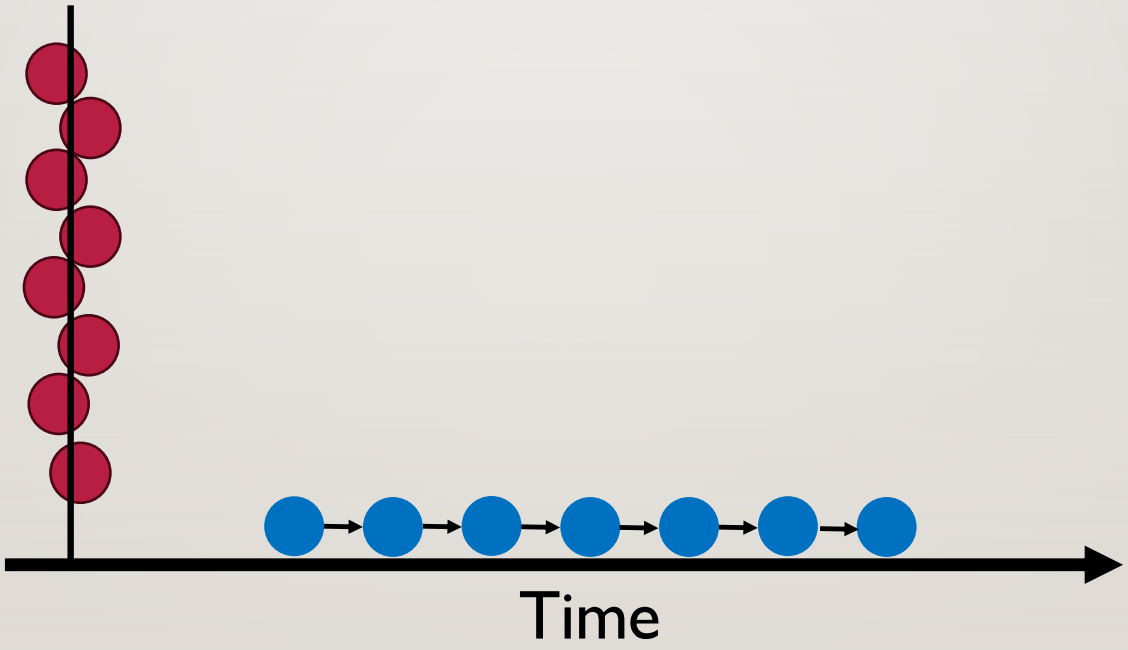


TYPES OF DATA

An Incomplete List



LONGITUDINAL VS. CROSS-SECTIONAL



CROSS-SECTIONAL DATA

- One or more groups examined at a particular point in time
- Gives a good "snapshot" of the study population
- These study designs are often very efficient
- One of two assumptions:
  - "Now" is inherently important in some way
  - "Now" represents at least a window of time

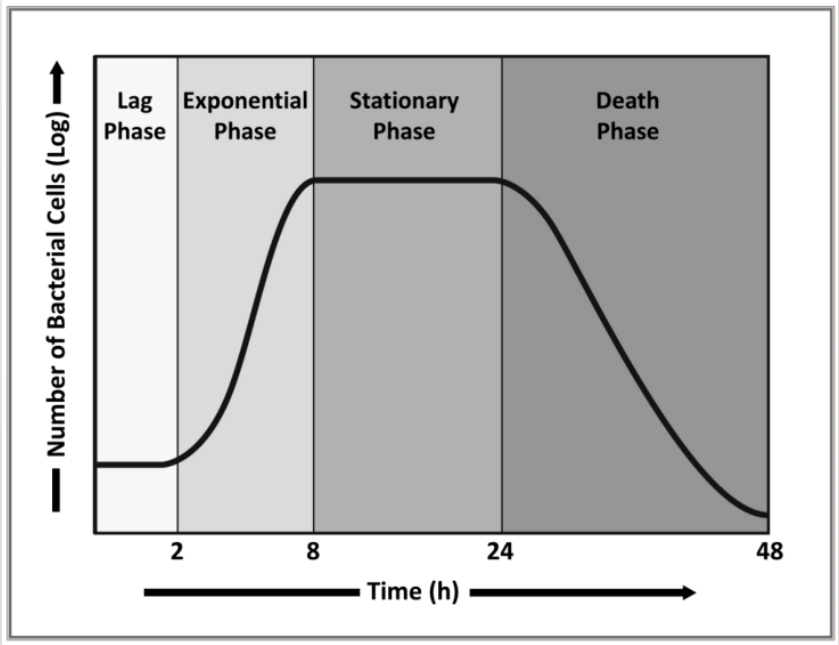
LONGITUDINAL DATA

- One or more groups are followed for a period of time
- This type of data allows for analysis with a time component to it
- It is often much more difficult and much more expensive
- This is true at most scales

SPECIAL TYPES OF LONGITUDINAL DATA

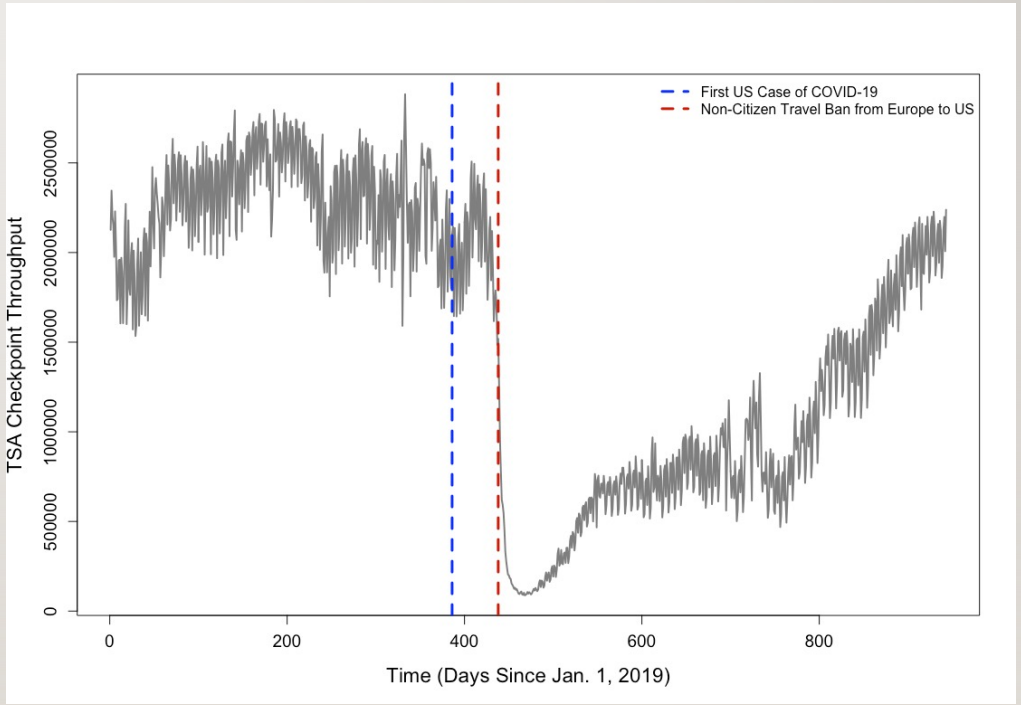
GROWTH DATA

- Longitude data about the growth of a population
- Some special dynamics about this type of data
  - Often characterized by exponential or logistic functions, depending on if the population is somehow constrained
- Applications outside biomedical science



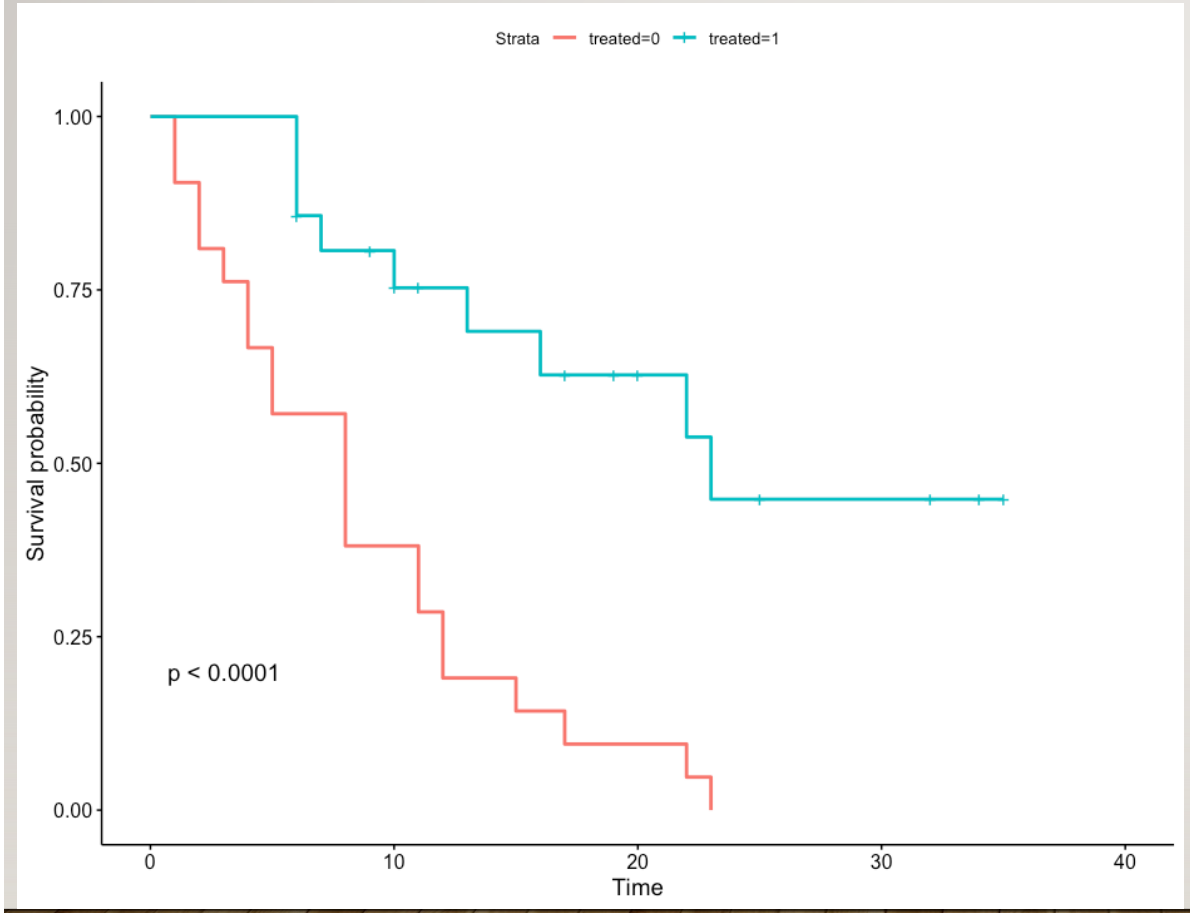
TIME SERIES DATA

- Data where time itself is of interest
- Very common in analysis of policy, natural experiments, etc.
- Also things like weather, the stock market, etc.
- Often longitudinal on a *very* high frequency
- Often aggregated (if a population) and each value in time is what we're interested in



TIME TO EVENT DATA

- Longitudinal data where what is of interest is the conversion of what's being studied from one state to the other
- HIV seroconversion
- All cause mortality
- Elimination of rabies in a particular area
- Often a very powerful type of data, but sometimes tricky


Strata
treated=0
treated=1
1.00
0.75
0.50
0.25
0.00
Survival probability
p < 0.0001
0
10
20
30
40
Time


BOLLYPOP.IN
MY FAVOURITE

PROSPECTIVE VS. RETROSPECTIVE DATA

- A concept for longitudinal data collection
- Prospective data: The outcome of interest has not occurred when data collection begins
- Retrospective: The outcome of interest *has* occurred when data collection begins
- Retrospective vs. Prospective is typically assessed from the perspective of the researcher
- Lots of data can be *collected* prospectively (i.e. it is about the present time when it is collected) but will end up being part of a retrospective study
- Your records from a medical appointment today will be part of a prospective study if it starts today, and a retrospective study if it starts a year from now
- This is murkier than a lot of people give it credit for

WHY RETROSPECTIVE DATA?

WHY RETROSPECTIVE DATA?

- It's already been collected, which usually means it's cheaper
  - This isn't *always* true – for example, using a new technique, assay, etc. on banked samples
- The answer can be obtained relatively rapidly
  - While subjects are followed for a long time potentially, that time has already happened
  - For prospective data, you have to bide your time
- There are pitfalls to analyzing retrospective data that are beyond the scope of today
- That does not mean prospective data is easy
  - Collecting it is often very hard

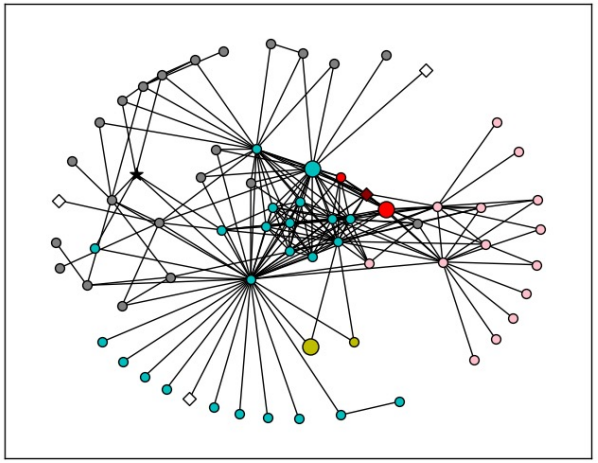
ACTIVE VS. PASSIVELY COLLECTED DATA

ACTIVE VS. PASSIVELY COLLECTED DATA

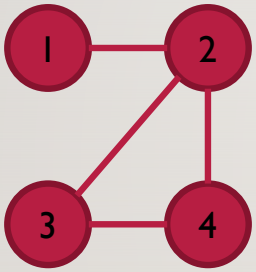
- Actively collected data has to be directly and deliberately collected
  - I sort of view this as "it takes effort to collect this data"
- Passively collected data is somehow gathered automatically
  - Pulling from records collected for other purposes, etc.
- This does not necessarily suggest "intent"
  - You can have very focused passive data collection

NETWORK DATA

- Data that is specifically collected around relationships
- What is a "network"
  - A conceptual way of representing relationships between things
  - Nodes: *Things*. People, places, etc.
  - Edges: Links between nodes
  - Occasionally these are called graphs, vertexes and arcs
    - Network science co-evolved in several different fields at about the same time
- Networks can be represented in a number of different ways
- A network's structure is sometimes called its "topology"
- There are whole classes on this



REPRESENTING NETWORKS



Diagram

1	2
2	3
2	4
3	4

Edge List

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Adjacency Matrix

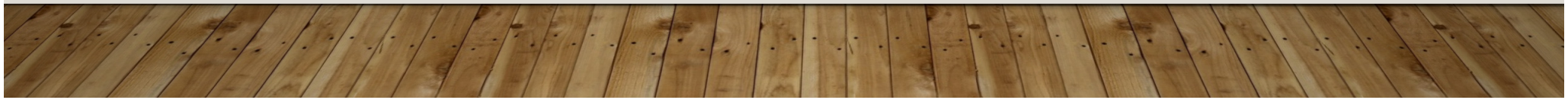
REPRESENTING NETWORKS

- Diagrams
  - Pros: Easily visualize the network structure, often look really cool
  - Cons: Can get difficult to interact with rapidly, "hairball" networks, not easily machine readable
- Edge Lists
  - Pros: Compact, expressive, easily machine readable
  - Cons: Less "human readable"
- Adjacency Matrix:
  - Pros: Matrix operations unlock all kinds of cool analysis techniques
  - Cons: Also less human reasable, less machine readable than edge lists
- Easy to go back and forth

SYNTHETIC DATA

- "Fake" data
- Data that is made up by a researcher
- There's actually a lot of utility to this type of data
  - When the data is generated, because we're generating it, we know its properties
  - This lets us check to make sure our tools give us the right answer
  - We can also make it go wrong in known ways
- Easy to share, do development on, etc. in a way that protects subject privacy
  - Big deal for humans, less of a thing for animals
- Can let us study populations that we would never be able to sample empirically

BIG DATA



VARYING DEFINITIONS

- Technical:
  - Data of a size where the full set of data cannot be held in RAM
  - This isn't normally what people mean when they say "Big Data"
    - More "Data that which is large"
- 2016 Silicon Valley Venture Capitalist
  - Massive, largely passively collected data
  - Large numbers of both columns (individuals) *and* rows (variables)
  - These also fit the first definition

PERILS OF BIG DATA

- Almost no "Big Data" is purpose built for what biomedical researchers want to use it for
  - Much of it is commercial
  - "Data of Convenience" – Jan Dasgupta
- Lots of data, tons of variable, etc. tends to force the use of automated methods for variable selection, etc.
- Computational issues – loops, sorting, etc. become hard, as does storage, querying, visualization, etc.
- Very high levels of precision
  - This is both very good (we can actually talk about rare diseases, etc.)
  - It's also dangerous (we can be very certain about being wrong)

THE PROMISE OF BIG DATA

- Rarity is less of a problem when you have massive amounts of data
  - A one-in-a-million condition is unlikely to show up in a 5,000 person sample
  - There's several hundred of them in something that captures the population of the United States
- Having *lots* of variables means potentially uncovering new and unexpected associations
  - Some of these are spurious
- Very high frequency data and automated analysis can potentially show us new insights
  - Video data, etc.