BIOMED SCI 552: STATISTICAL THINKING

LECTURE 2: THINKING ABOUT DATA

QUESTIONS FROM TUESDAY?

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

A NOTE ON THE PROBLEM SET

- Due next Thursday
- Should have a Canvas site up on Friday
- 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?

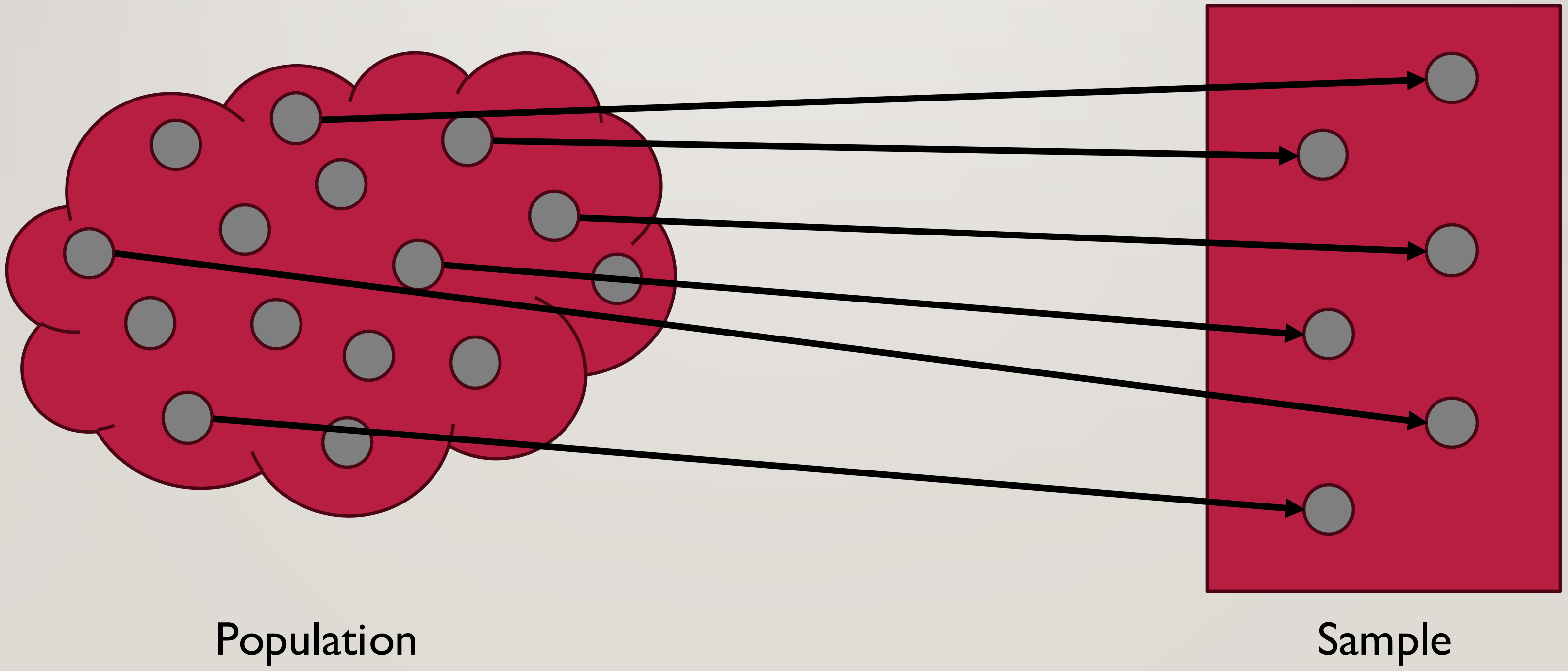
WHAT'S A POPULATION

- 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
  - United States Marines deployed to Afghanistan during the Global War on Terror
- 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



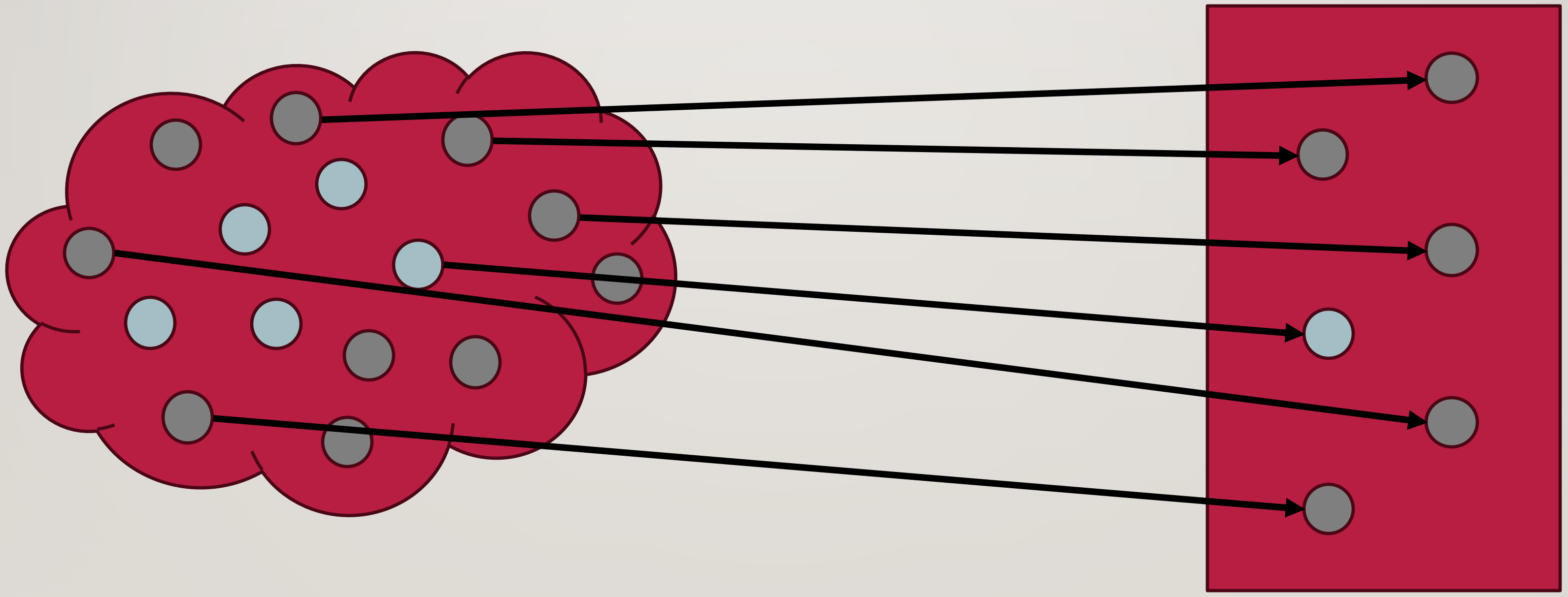
WHY DO WE NEED TO SAMPLE?

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
  - 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

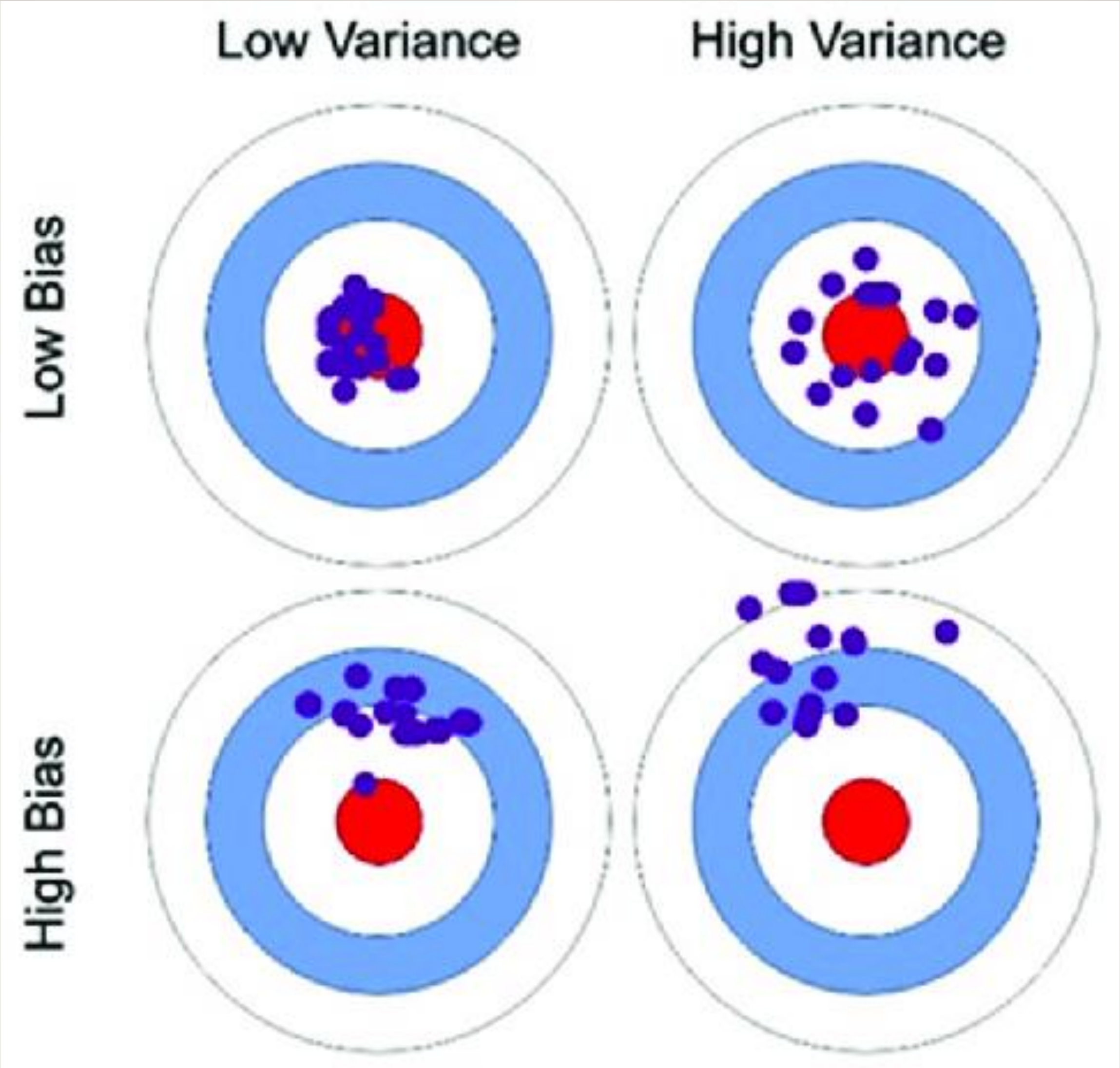


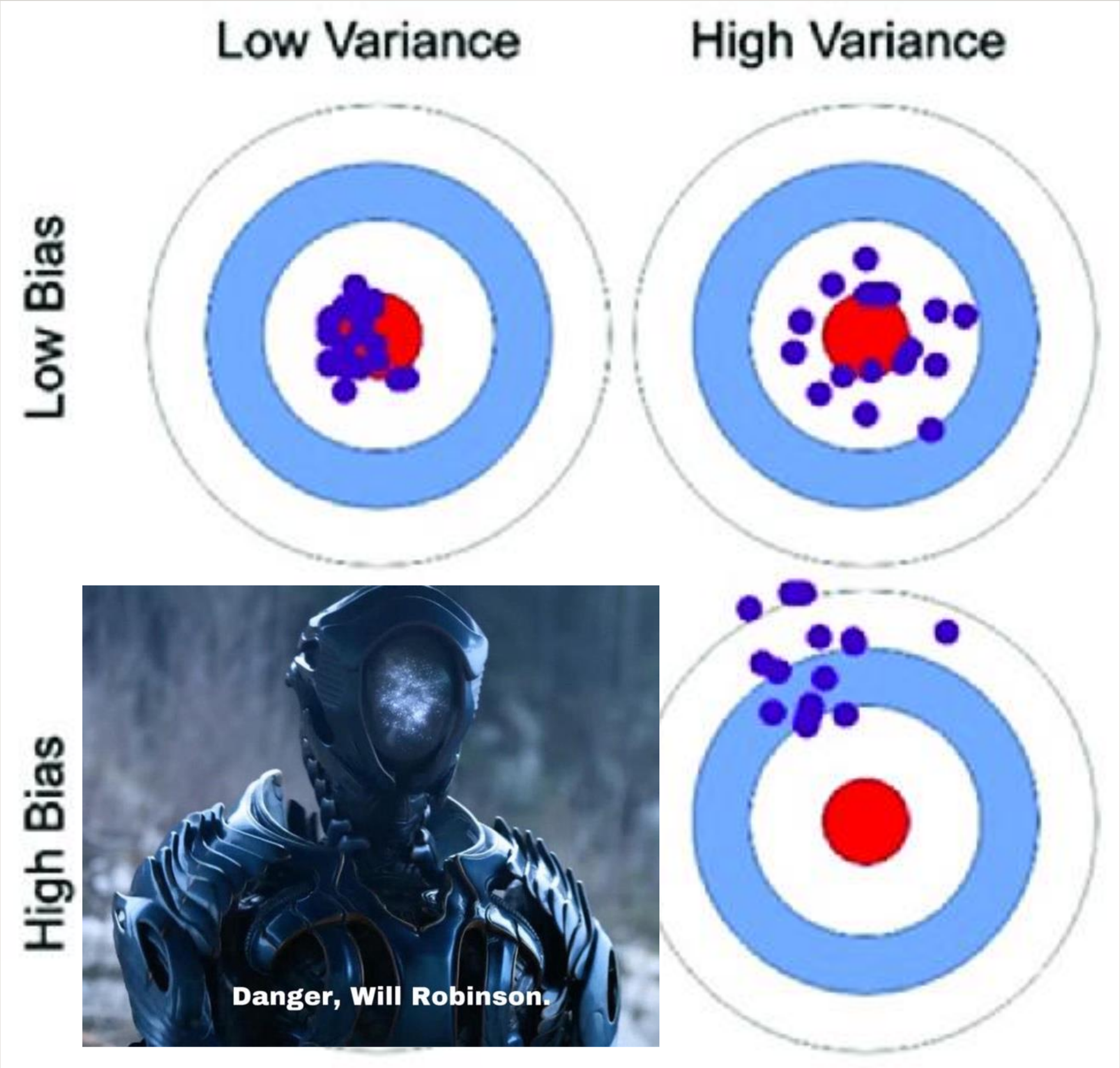
Population

Sample

SAMPLING ERROR IS OKAY!

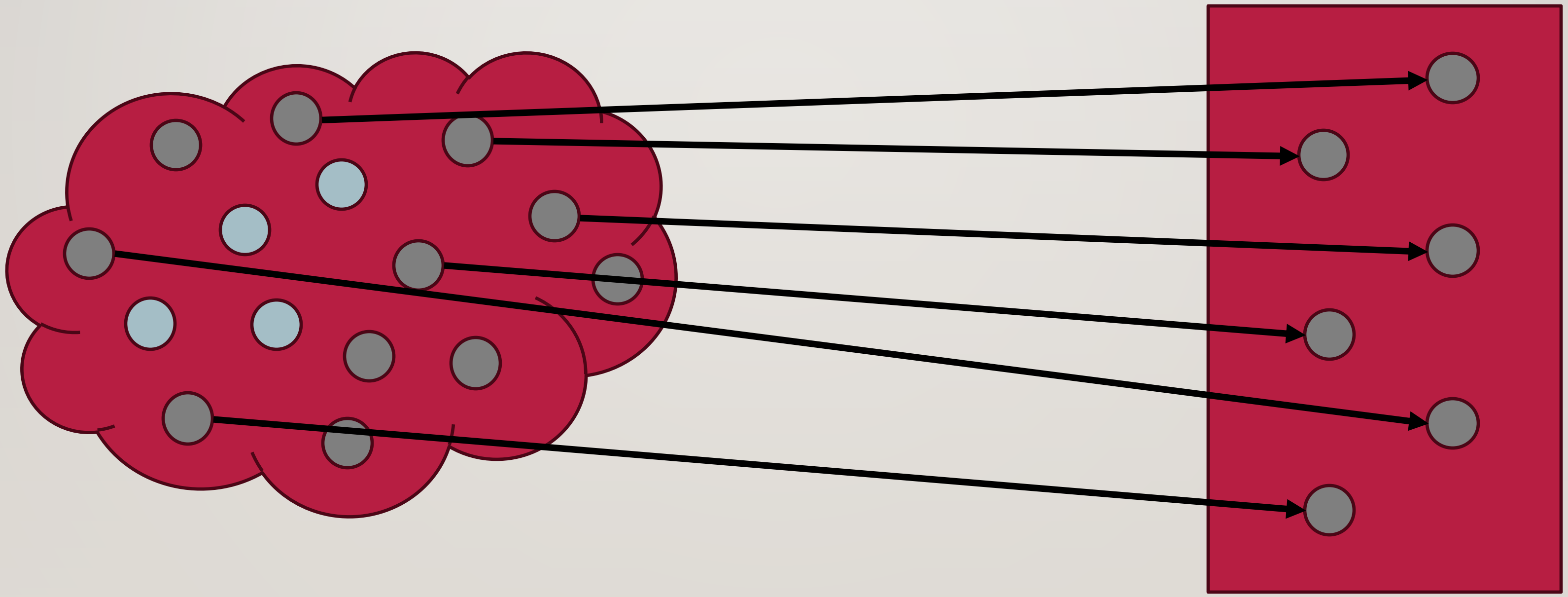
- And, to be blunt, inevitable
- Random variation pervades all of the biological sciences
- Sampling error creates *uncertainty* but not *bias*
- Bigger samples, more studies, meta-analysis, etc. can help reduce that uncertainty


Low Variance
High Variance
Low Bias
High Bias


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

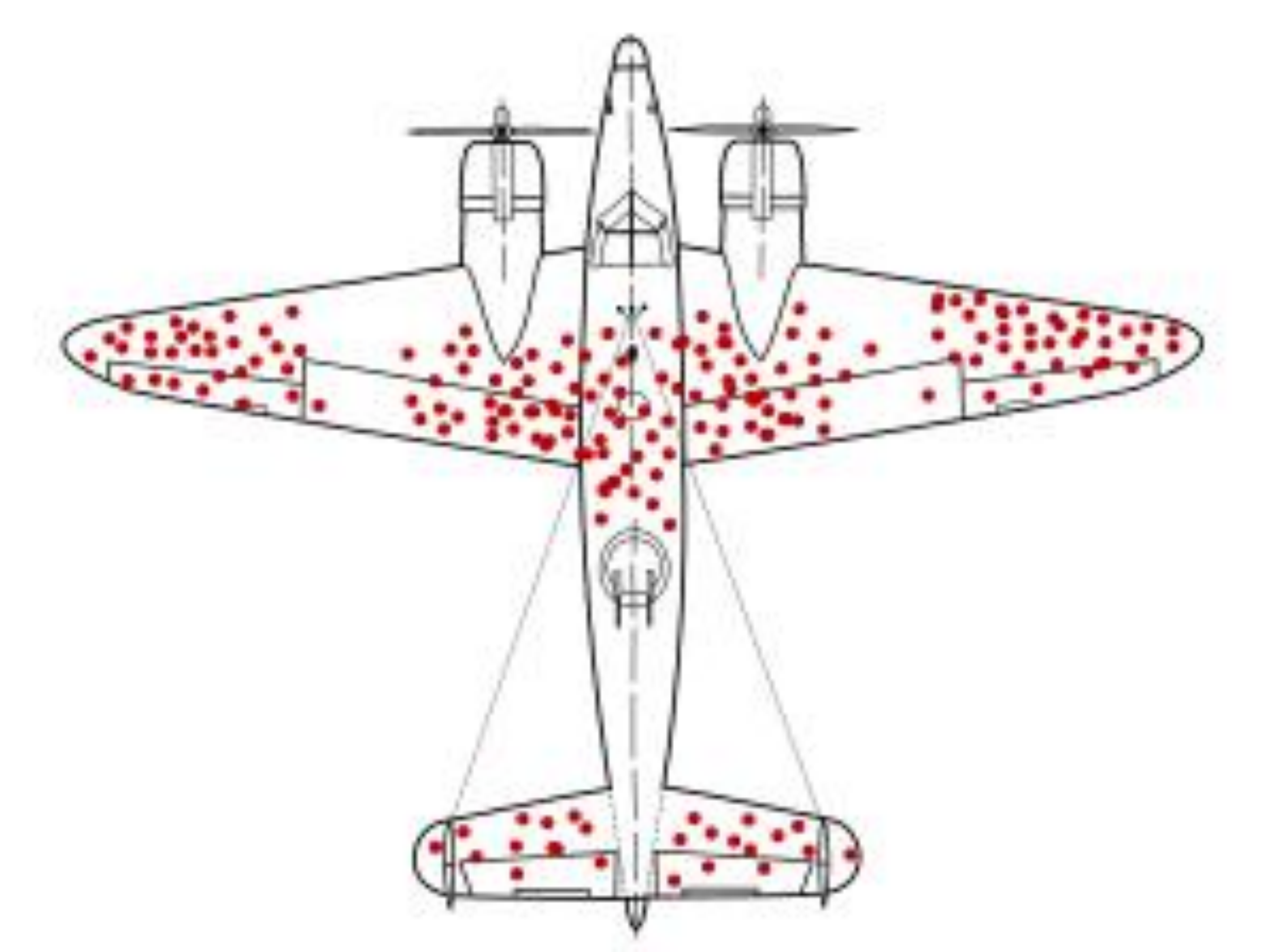
THE PERILS OF SAMPLING

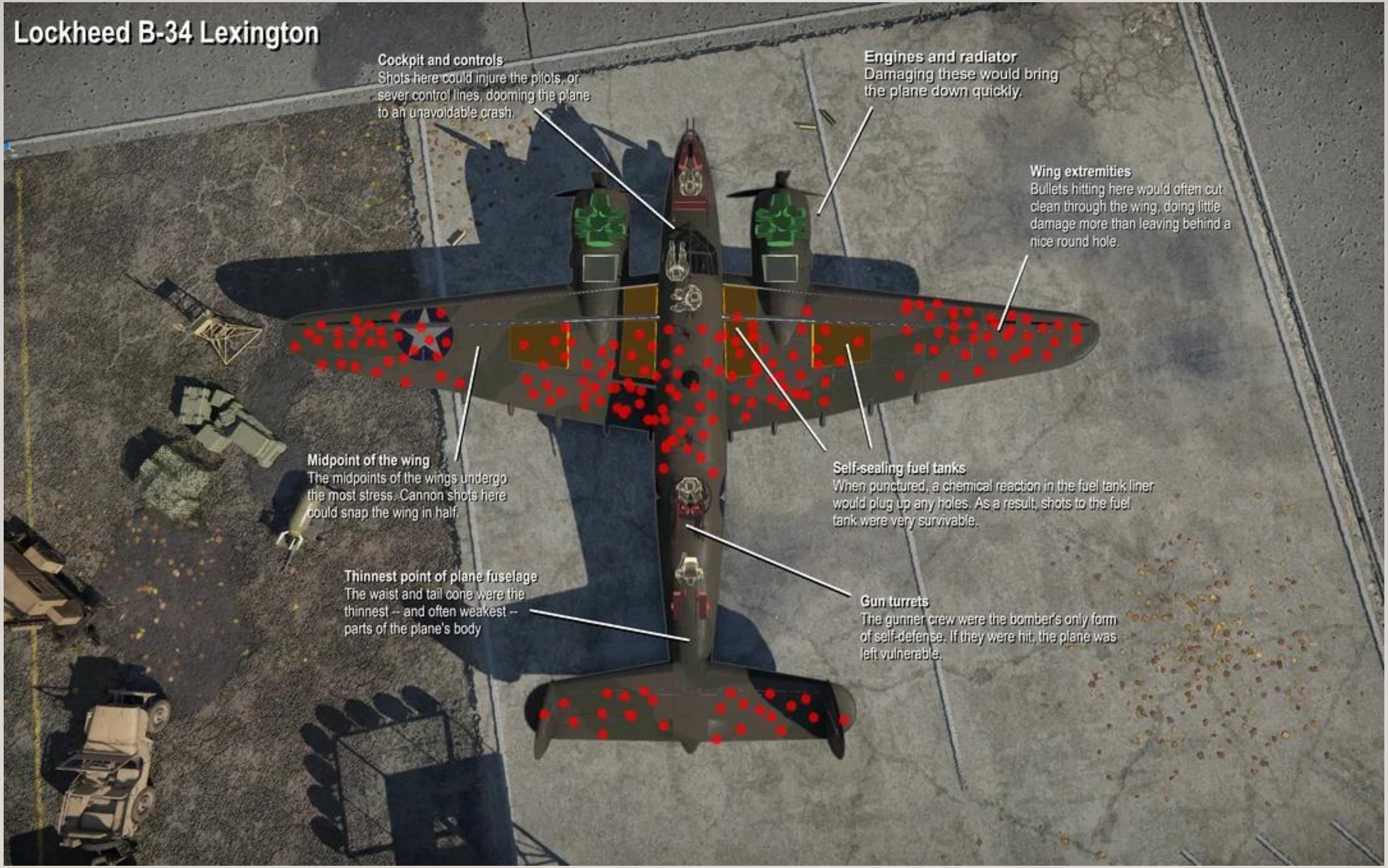
Biased Sample



Population

Sample



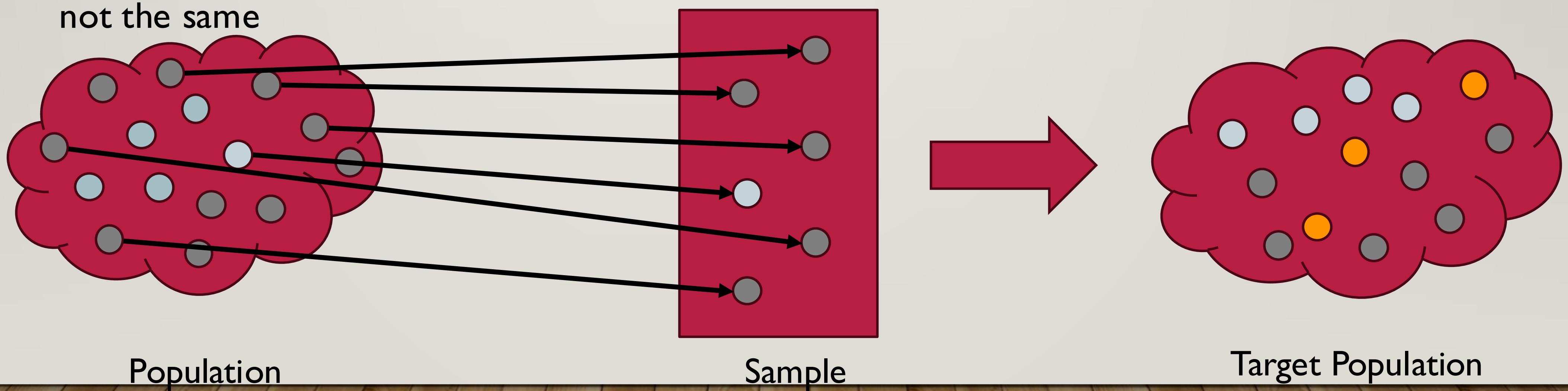

Lockheed B-34 Lexington
Cockpit and controls
Shots here could injure the pilots, or sever control lines, dooming the plane to an unavoidable crash.
Engines and radiator
Damaging these would bring the plane down quickly.
Wing extremities
Bullets hitting here would often cut clean through the wing, doing 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"

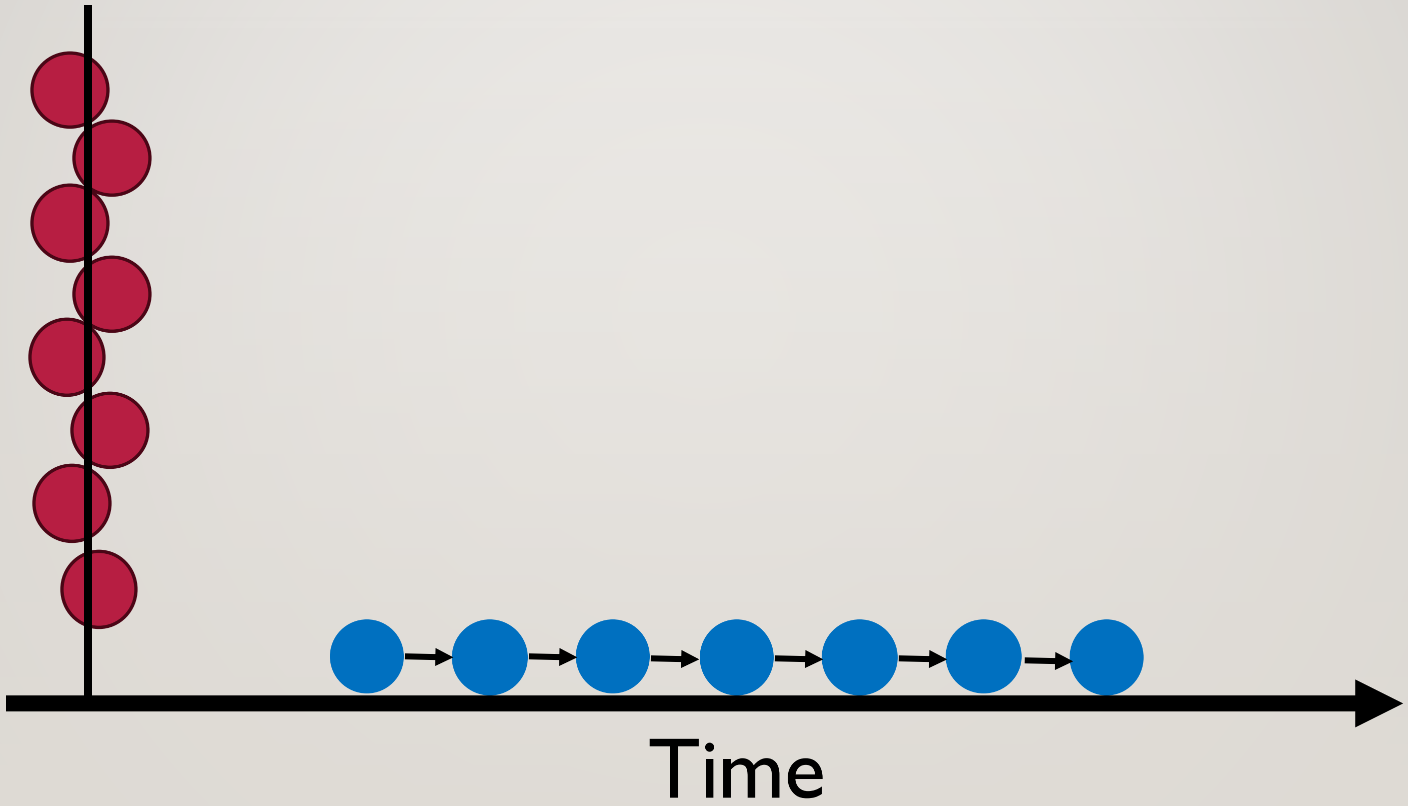
- Internal Validity: Are the results of your study unbiased – within your sample, can we be confident that your results are "correct"
- External Validity: How well can the results of your study be applied to other populations?
- Historically, we have emphasized internal validity
- Target Validity: This 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

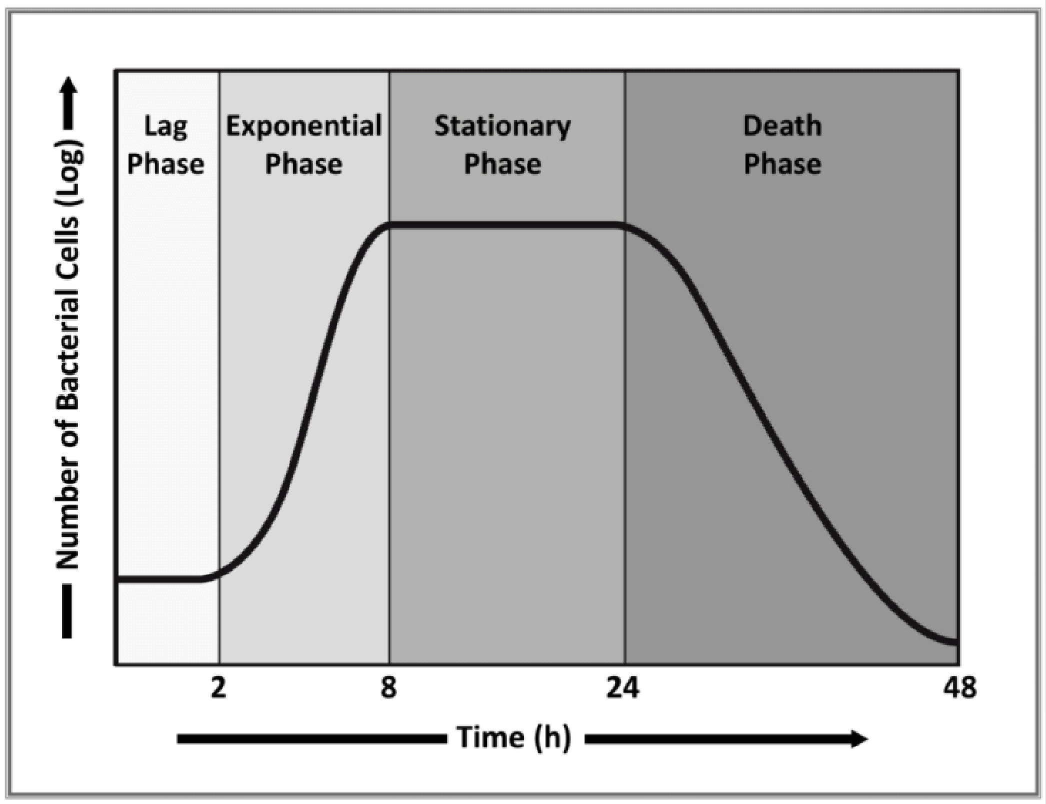
LONGITUDINAL DATA

- 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

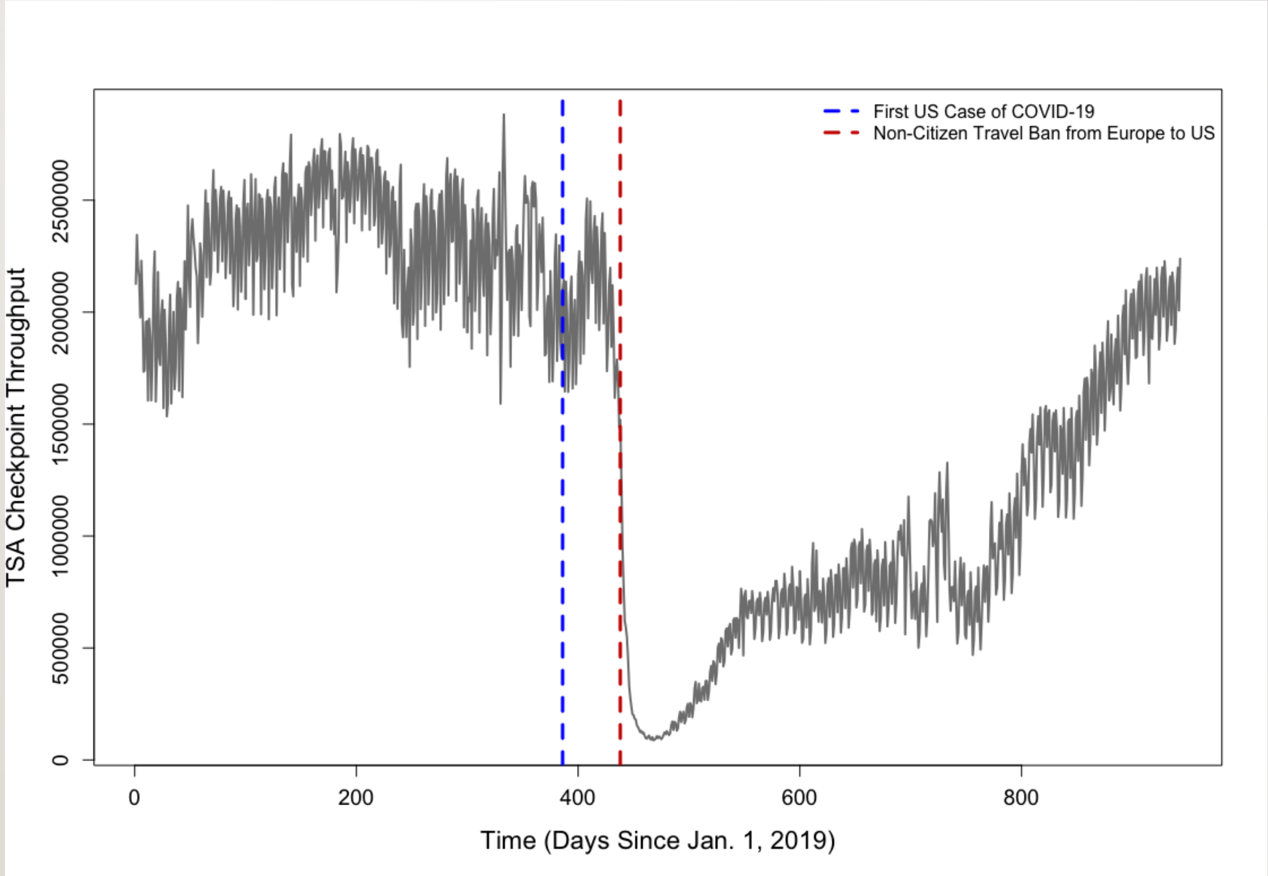
GROWTH DATA

- 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



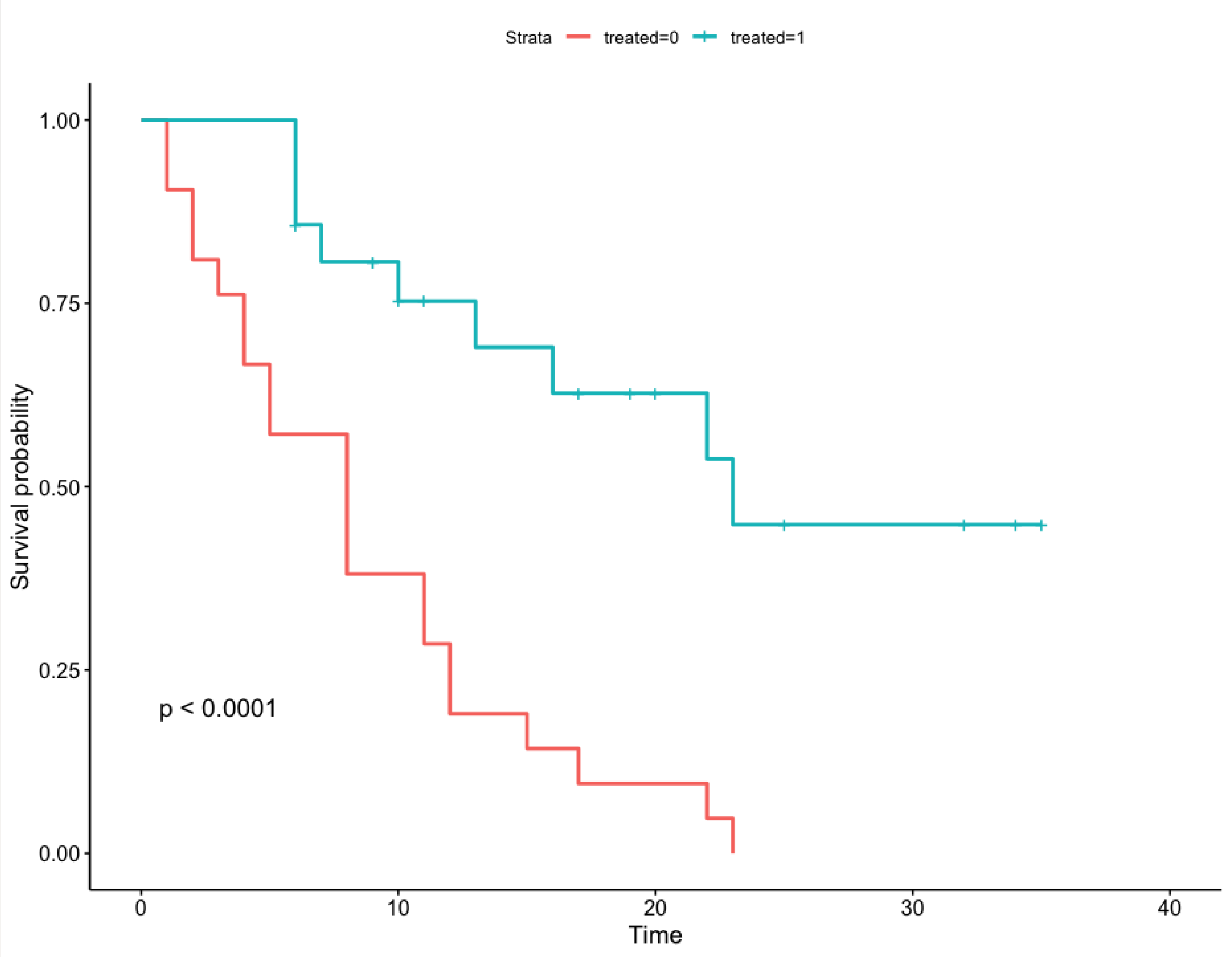
TIME SERIES DATA

- 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

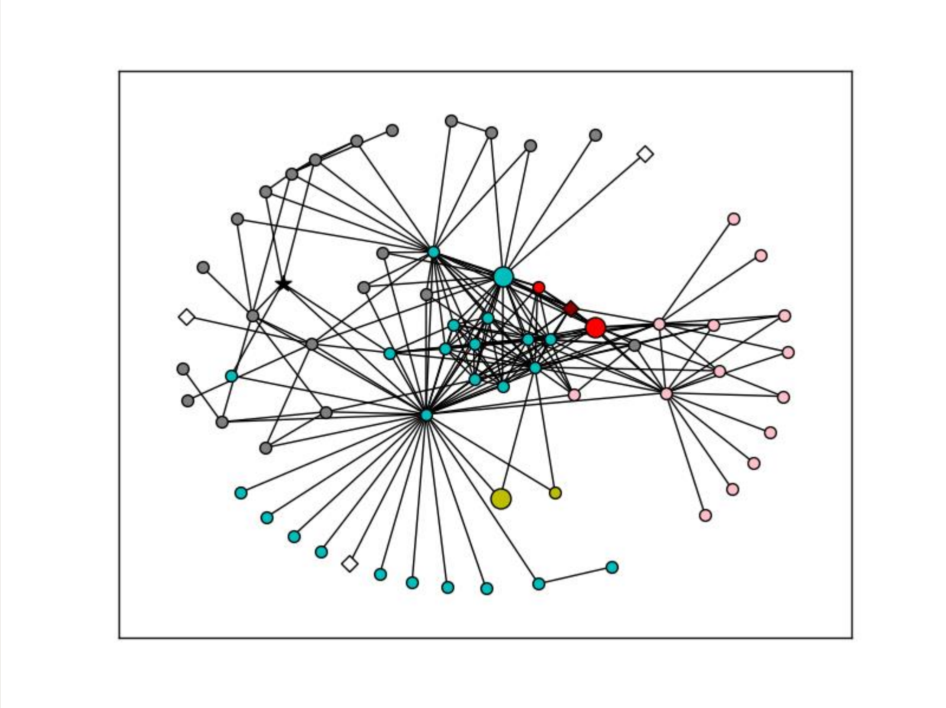
ACTIVE VS. PASSIVELY COLLECTED DATA

ACTIVE VS. PASSIVELY COLLECTED DATA

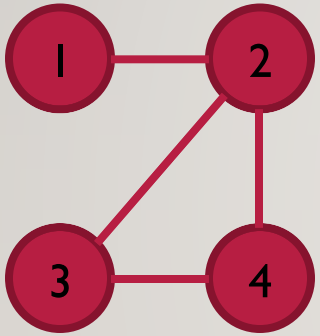
- Actively collected data has to be directly and deliberately corrected
  - I sort of view this as "it takes effort to collect this data"
- Passively collected data is somehow gathered automatically
  - Pulling from records collected for other purposes, etc.
- This does not necessarily suggest "intent"
  - 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
- Very high frequency data and automated analysis can potentially show us new insights
  - Video data, etc.