STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington

@AmyDWillis — Assistant Professor

@BryanDMartin_ — PhD Candidate

@paulinetrinh — PhD Candidate

STATISTICS

- Who wants to know more statistics?
- Why?

LEARNING OBJECTIVES

1. Introduce key concepts from statistics in a microbial setting
   - population, parameter, estimate, model…
2. Discuss estimation and properties of estimators
   - bias & variance
3. Discuss model misspecification and its implications for hypothesis testing
4. Learn about statistical transparency and ethics
5. Draw connections with your previous statistics exposure (if any)

LEARNING PROCESS

STATISTICS

- Two different types of statistics
  - Inferential statistics
  - Exploratory statistics

EXPLORATORY STATISTICS

- What does your data say?
- How do we show what it says?
- How to visualise it?
- *Descriptive statistics*
- Not the focus of this talk (many other talks will discuss these ideas!)

INFERENTIAL STATISTICS

- “Use your data to say something about a population”

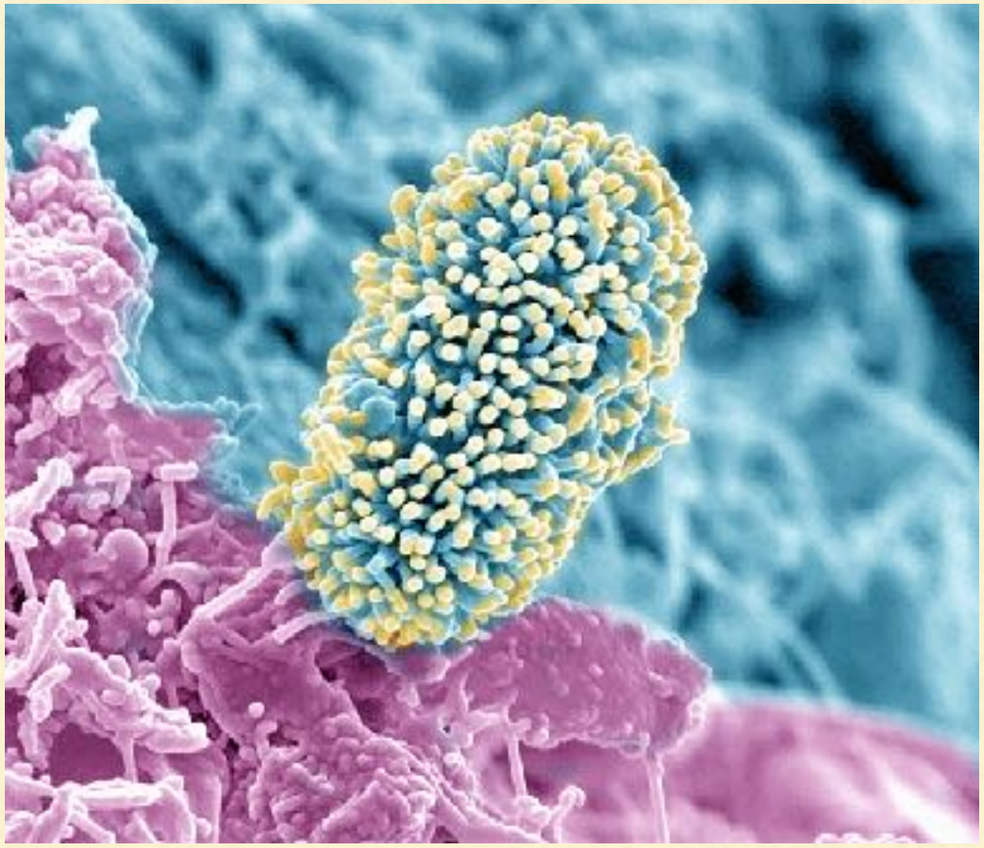
POPULATION

- Stat101
  - "The population of women with breast cancer"
  - "The population of American citizens with graduate degrees"
  - "The voting population of Massachusetts"

What is the population in microbial ecology?

MICROBIAL POPULATIONS

- The (A) microbiome is the (a) collection of microbes, and their genes and metabolites



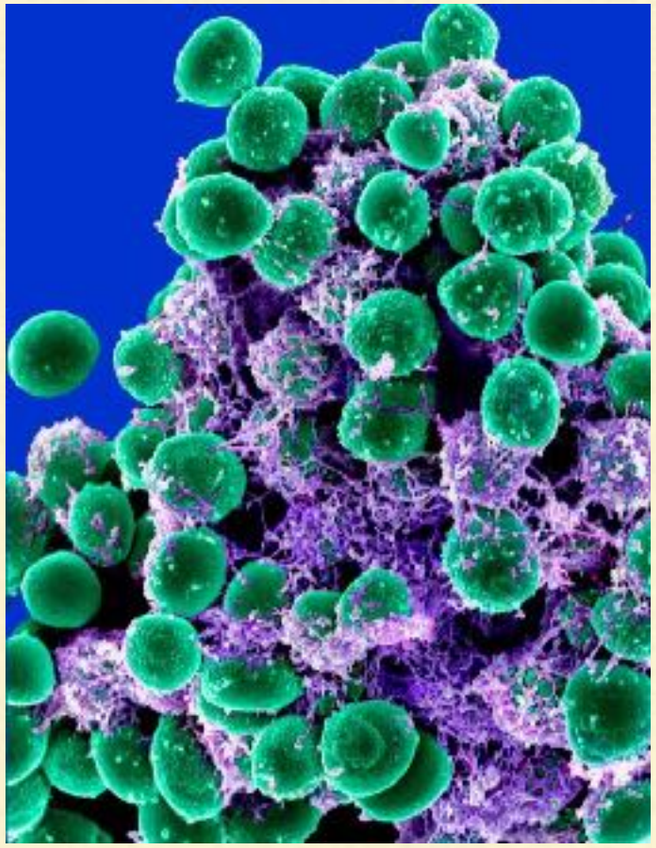


Photo credits: Pacific Northwest National Laboratory; NIAID

MICROBIAL POPULATIONS

- Are you interested in microbes living in the ocean?
  - Which ocean?
  - At what depth?
  - What time of year and day?
    - Or only those you can detect with your primers?

The population that you want to study may not be the population that you get to study

MICROBIAL POPULATIONS

- The 4 W's: **Who/What? Where? When? Why?**
  - **Who? What?** …the poop of white women 25-45 y.o. with a clinical IBD diagnosis?
  - **Where?** … who also live in the city that your study was conducted in?
  - **When?** … between January 2019-March 2019?
- Such observations can help us answer **why** certain patterns exist…
  - and why others don't....

EXPERIMENTAL DESIGN

The population that you want to study may not be the population that you get to study

- Before undertaking a microbiome study, think carefully about:
  - the question you want to answer,
  - the data you have access to, and
  - the questions you can answer with the data that you have access to

MICROBIAL POPULATIONS

- Group exercise: (5 minutes)
  - Come up with a microbiome-related question that ***you want to answer*** considering the following questions:
    - **Who/What? Where? When? Why?**
  - Come up with a microbiome-related question that ***you could study***
    - *How do (sequencing) technology and (bioinformatics) tools influence what populations you can study?*

POPULATIONS VERSUS SAMPLES

- The difference between a *population* and *a sample from it* is fundamental in statistics

- Inferential statistics: using information about the sample to infer something about the population

"SOMETHING ABOUT THE POPULATION"

- Statisticians have a formal concept of this
- "Parameter": a numerical characteristic of a model

PARAMETERS

- Statistical parameters are a way to connect reality to your data
- *You* need to decide on a reasonable model for reality

PARAMETERS

- Example of a model:
  - There are microbes in your saliva, and they all have a taxonomic label at the genus level
- Example of a parameter:
  - The genus-level relative abundance of Streptococcus in your saliva right now

PARAMETERS

- Example of a model:
  - Every #STAMPS2019 attendee carries with them methicillin-resistance S. aureus, or they don't
- Example of a parameter of this model
  - The fraction of #STAMPS2019 attendees carrying S. aureus that have methicillin-resistance in any abundance

PARAMETERS

- Other possible parameters of interest
  - The proportion of people in US hospitals carrying *S. aureus* that are methicillin-resistant
  - The mean phylum-level diversity of microbes on the hands of employees in the dining hall
- Who is the implicit population here? When?

PARAMETERS

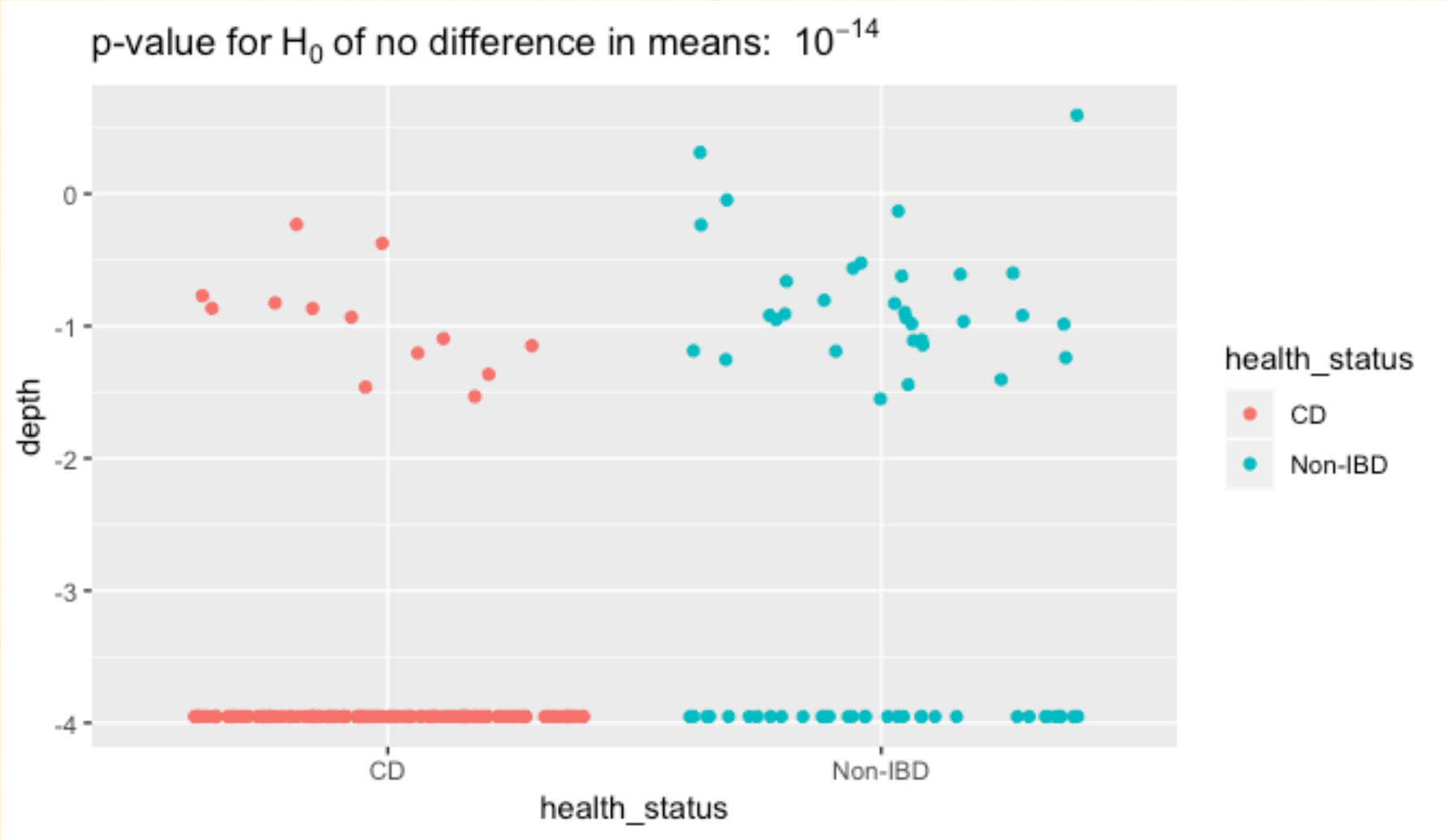
- Statisticians often estimate *means*
- Never forget that this is a broad-range summary
- Results about means say nothing about *individuals*!
- Trade off of inferential statistics: need a population

INFERENCE VS PREDICTION

- How exciting is a p-value of 10^{-14}?

INFERENCE VS PREDICTION

- How exciting is a p-value of 10^{-14}?



More on this later…

A DISTINCTION

genus-level relative abundance of Streptococcus in your saliva right now

$\neq$

the proportion of reads that map to Streptococcus that we would obtain from shotgun sequencing a sample

Why not? Discuss with the person next to you (3 minutes)

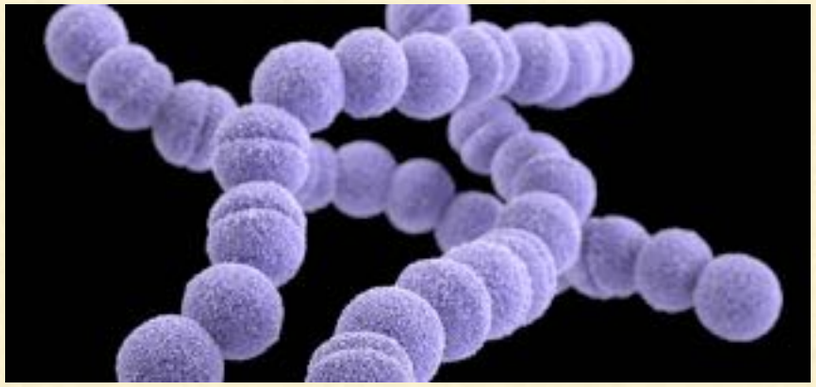
POPULATIONS VERSUS SAMPLES

- Key idea: *the sample is not the population*
- Inferential statistics: using information about the sample to infer something about the population
  - Use the observed data to estimate the parameters

"INFORMATION ABOUT THE SAMPLE"

- Estimates (n, pl) estimate (v) parameters (n)
- "Estimate": the number calculated from your data
  - “An estimate of the phylum-level relative abundance of Bacteroidetes in my gut right now is 30%”
- "Estimator": a function of your data
  - “My estimator of relative abundance is plug-in 16S relative abundance…”

EXAMPLE



- *Estimate the genus-level relative abundance of Streptococcus in your saliva using 16S data*
- Relative abundance is commonly estimated by the <u>observed</u> relative abundance of 16S copies from Streptococcus
- Is that the only estimate? Why does it seem like a good one?

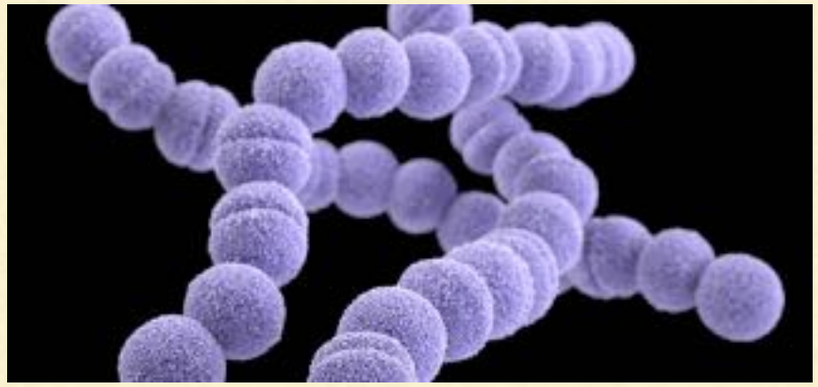
RELATIVE ABUNDANCE

- Suppose…
  - **n** = samples, indexed by **i** = 1, …, n
  - $\mathbf{p_i}$ = the relative abundance in each subject
  - $\mathbf{W_i}$ = # of observed sequenced copies from Strep
  - $\mathbf{M_i}$ = total # of sequenced copies
- Most common estimate of p_i is W_i/M_i

RELATIVE ABUNDANCE

- Why?
  - (Seems reasonable)
  - Under a model where each observed copy of the 16S gene is from Strep with probability p_i, and all copies are independent, this estimate is
    - consistent, normally distributed, efficient, unbiased, minimum variance out of all unbiased estimates…

EXAMPLE



- **Motivation**: Estimate the genus-level relative abundance of 16S copies from Streptococcus in *a group of people*
  - What if we have 10 people in our study?
  - What does relative abundance of Streptococcus mean now?
    - We will return to this question in a few slides…

PARAMETERS

- Two key concepts for evaluating estimators of parameters
  - bias: how far?
  - variance: how stable?
- Suppose we have a parameter θ and an estimate $\hat{\theta}$

ESTIMATES: NOTATION

- The parameter *Amy*:

ESTIMATES: NOTATION

- An estimate of the parameter *Amy*:

COMPARING ESTIMATES

- **Motivation**: "Estimate the genus-level relative abundance of Strep"
  - $\mathbf{W_i}$ = # observed sequenced reads mapping to Strep in person i
  - $\mathbf{M_i}$ = total # reads sequenced from person i
- *Consider the following two estimates*
  - Take everyone's relative abundance ($W_1/M_1, \ldots, W_n/M_n$), then average them
  - Add up everyone's Strep counts $W_{total} = W_1 + \ldots + W_n$, then add up everyone's total counts $M_{total} = M_1 + \ldots + M_n$, and divide the two: W_{total}/M_{total}

"BIAS"

- The estimate W_{total}/M_{total} is "biased towards" deeply sequenced samples
  - Could be fine! Could be great!
  - *It depends on your model…*

These are questions that occupy statisticians

Long version: W_{total}/M_{total} is a good estimate under a multinomial model, but under many types of hierarchical models, the mean of $W_1/M_1, \ldots, W_n/M_n$) is going to be closer to be "optimal" (though it depends on the type of hierarchical model)

BIAS

- If you care about a parameter θ, then the bias is the expected difference between the estimate and the true value

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

where

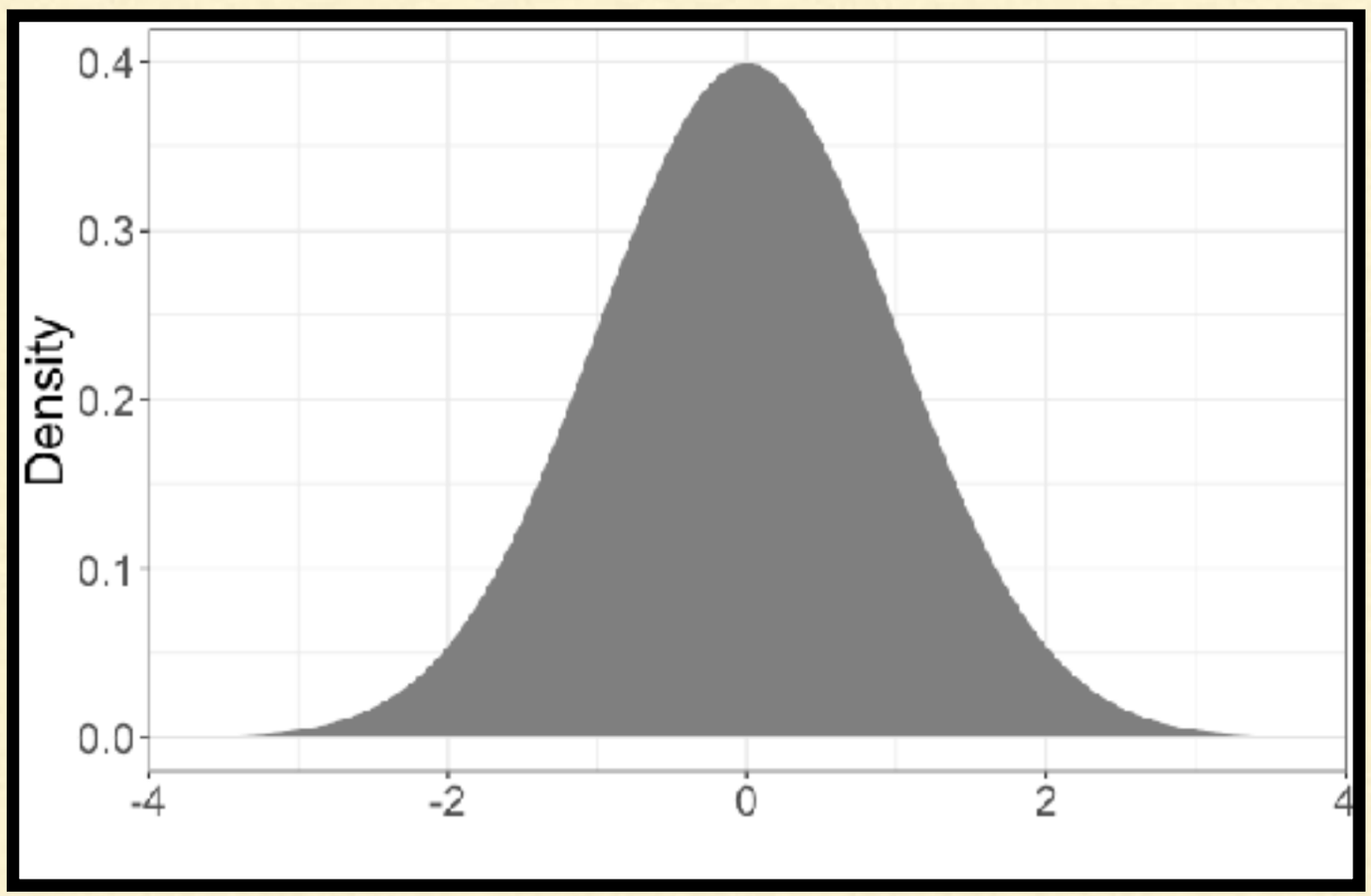
$$\mathbb{E}\hat{\theta} = \text{“expected value” of } \hat{\theta}$$

EXPECTED VALUE

- Expected value comes from the concept of a “distribution”
- It is the “middle” of the distribution

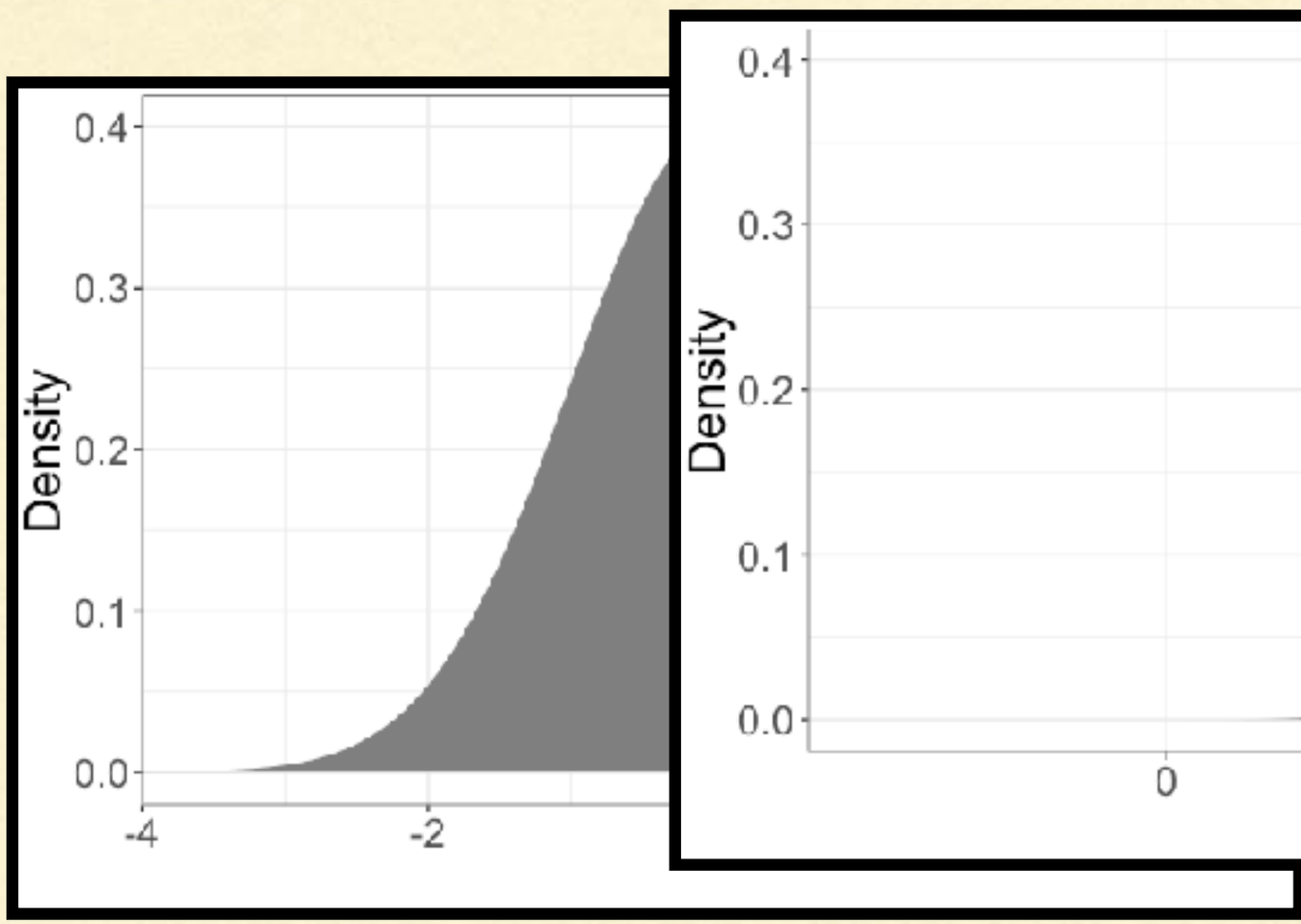
EXPECTED VALUE

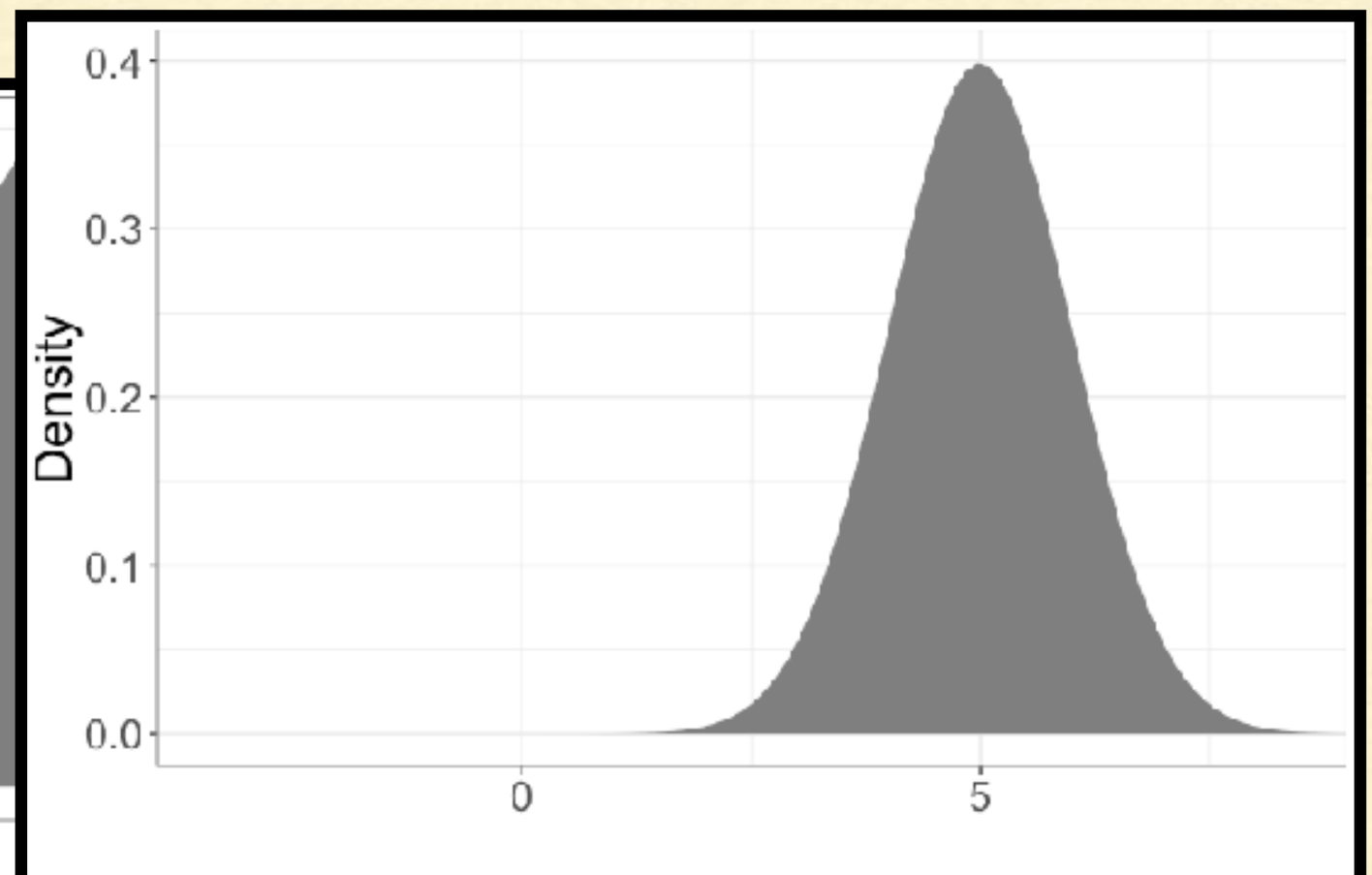
- What you think is the expected value of these distributions?



EXPECTED VALUE

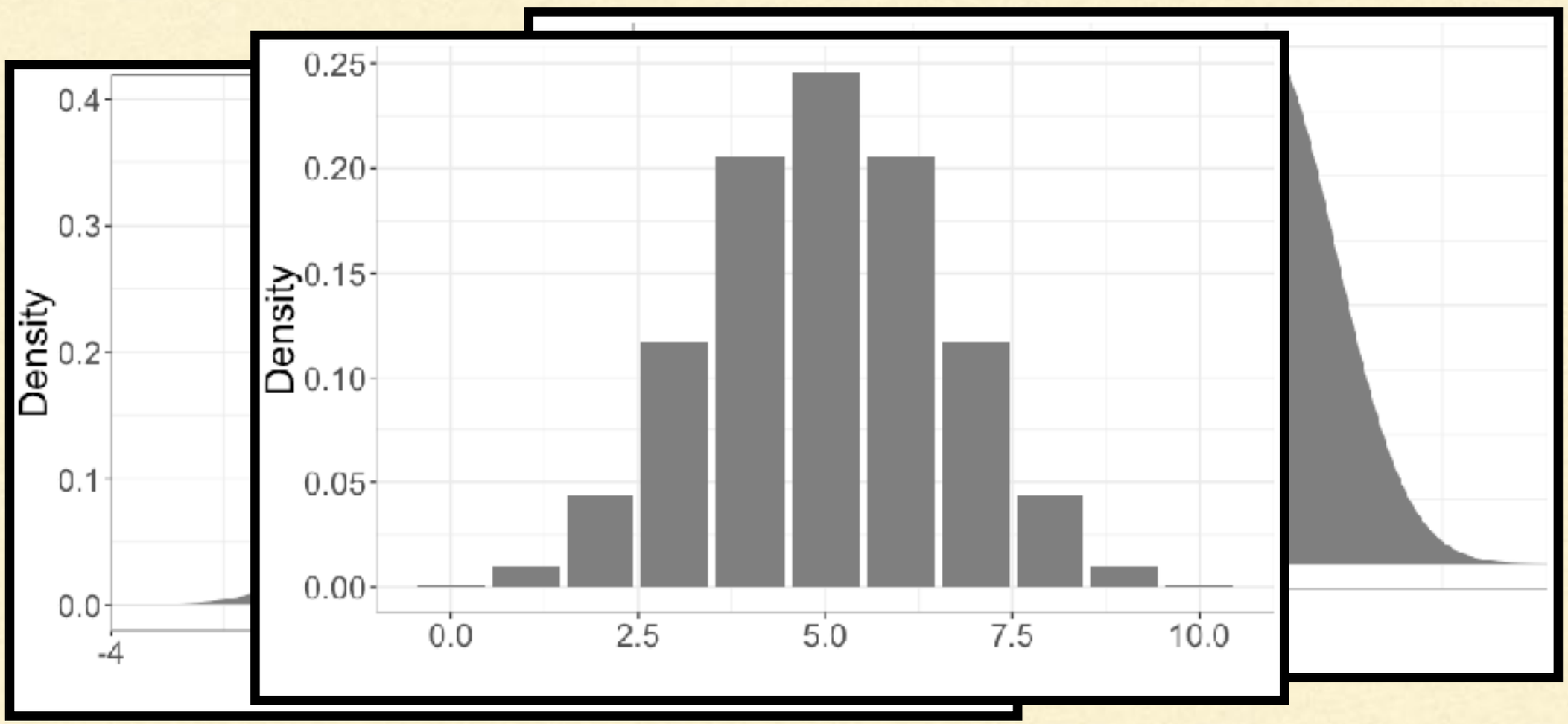
- What you think is the expected value of these distributions?





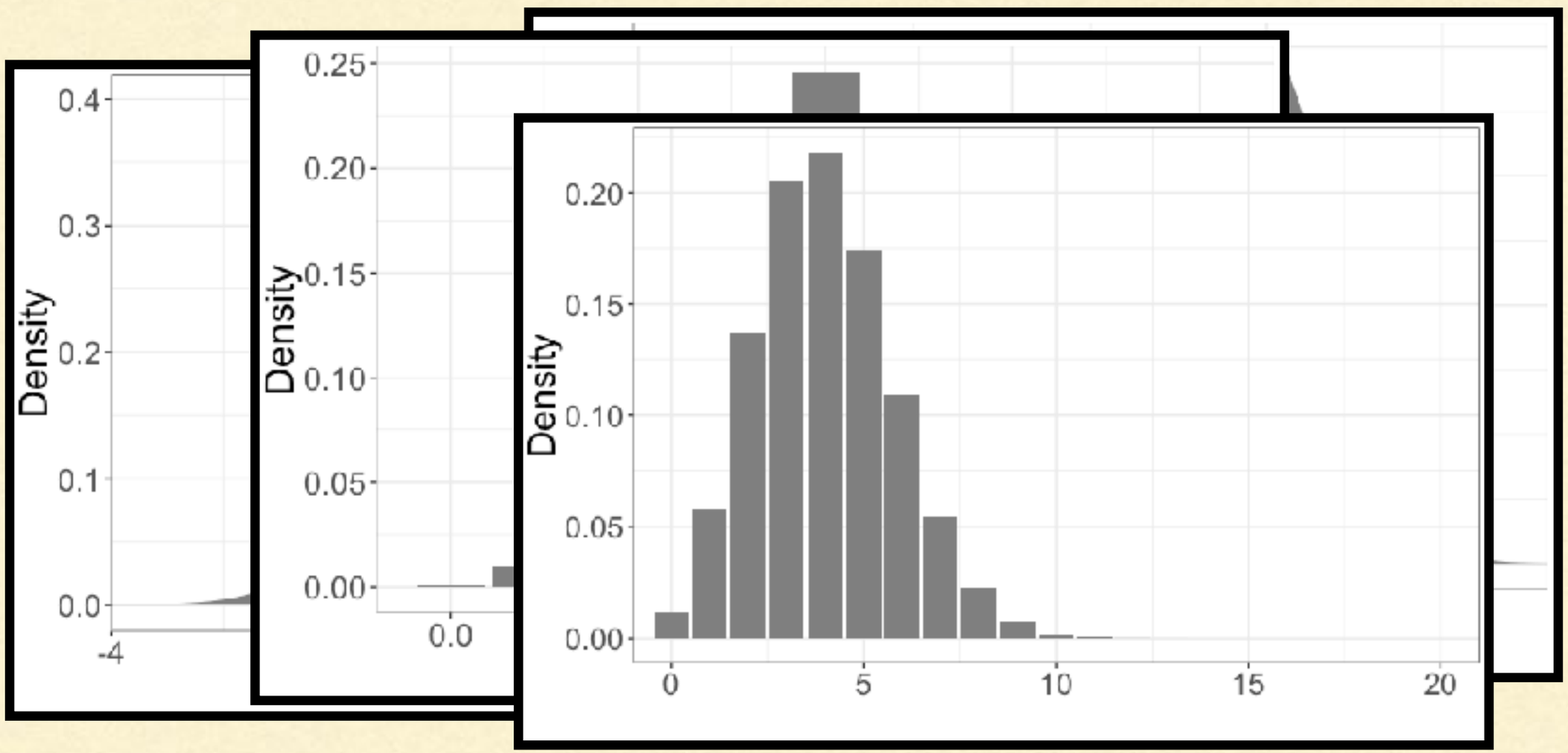
EXPECTED VALUE

- What you think is the expected value of these distributions?



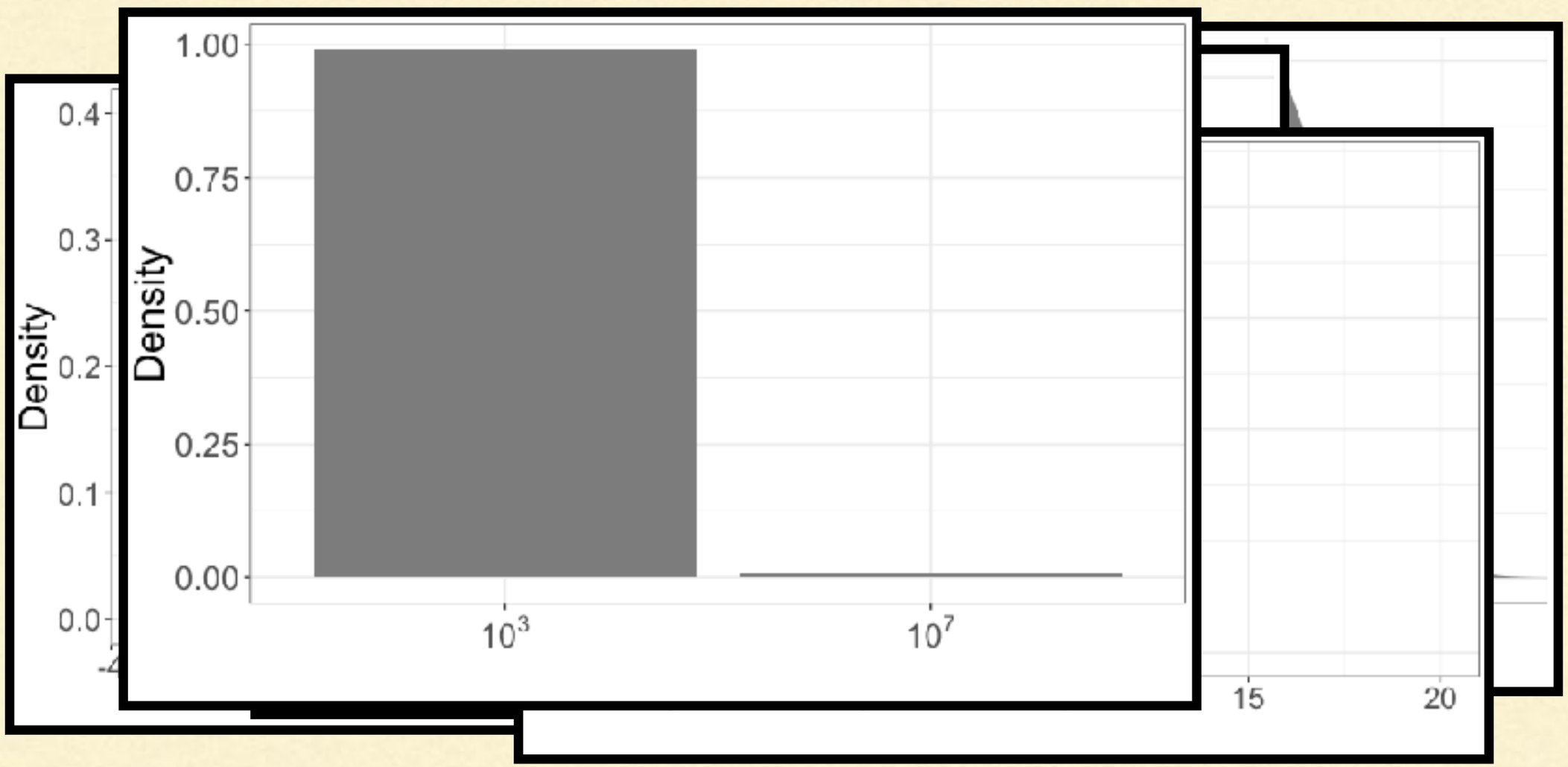
EXPECTED VALUE

- What you think is the expected value of these distributions?



EXPECTED VALUE

- What you think is the expected value of these distributions?



EXPECTED VALUE

- Suppose we have a distribution with discrete support on $x_1, \ldots, x_n$, and the probability of selecting each point is $p(x_1), \ldots, p(x_n)$. Then

$$\text{Expected value} = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n)$$

- Suppose we have a distribution with continuous support on x_{lower} to x_{upper}, and the density of a point x is f(x). Then

$$\text{Expected value} = \int_{x_{\text{lower}}}^{x_{\text{upper}}} x f(x) dx$$

Weighted average of values the random variable can take

BIAS

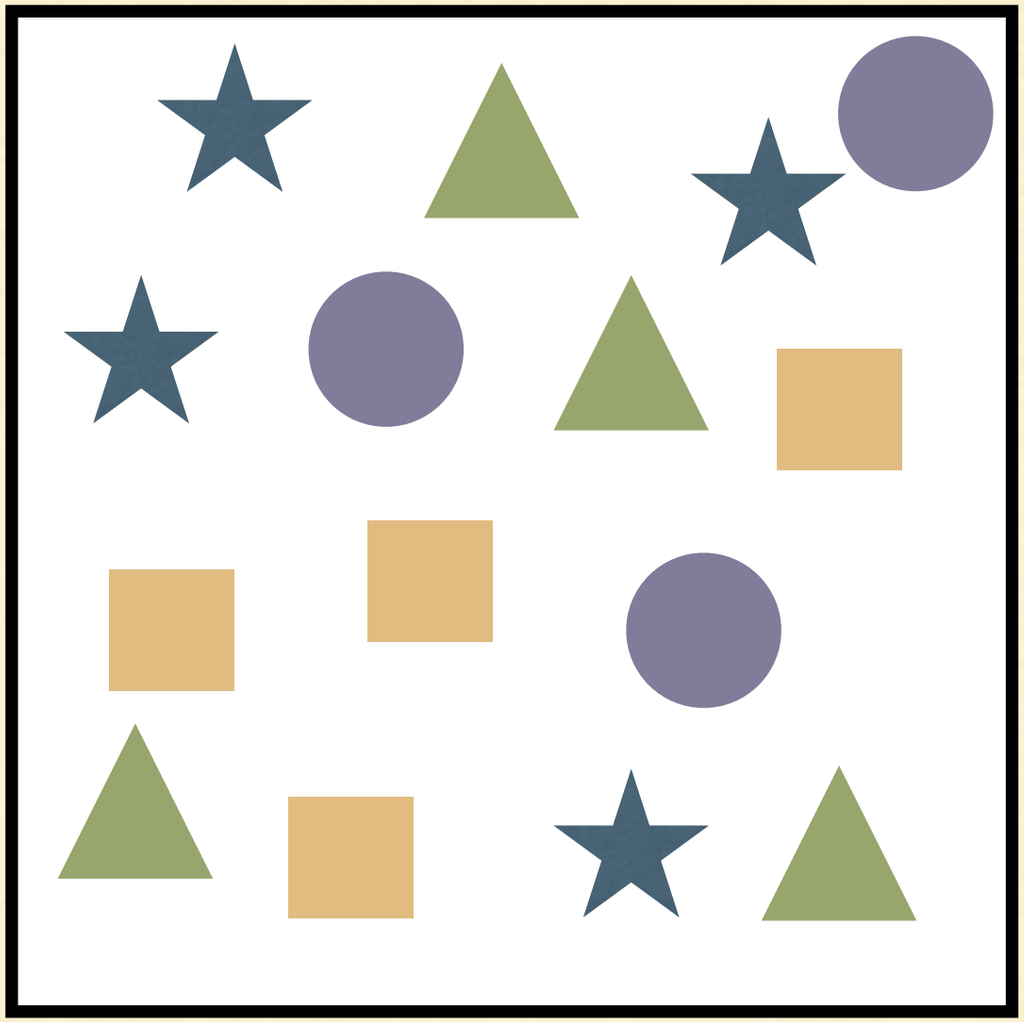
- An estimate is unbiased if its bias is zero
- The fine print:
  - An estimate of a parameter
    - is unbiased if
      - its bias is zero under the model
- The distribution of the estimate depends on the distribution of the data, and thus, on the model!

BIAS

- Be careful — this word is used frivolously
- Before being impressed, ask yourself
  - What is the estimate?
  - What is the model?
  - Is the model reasonable?
  - Why do they think its unbiased?

EXAMPLE

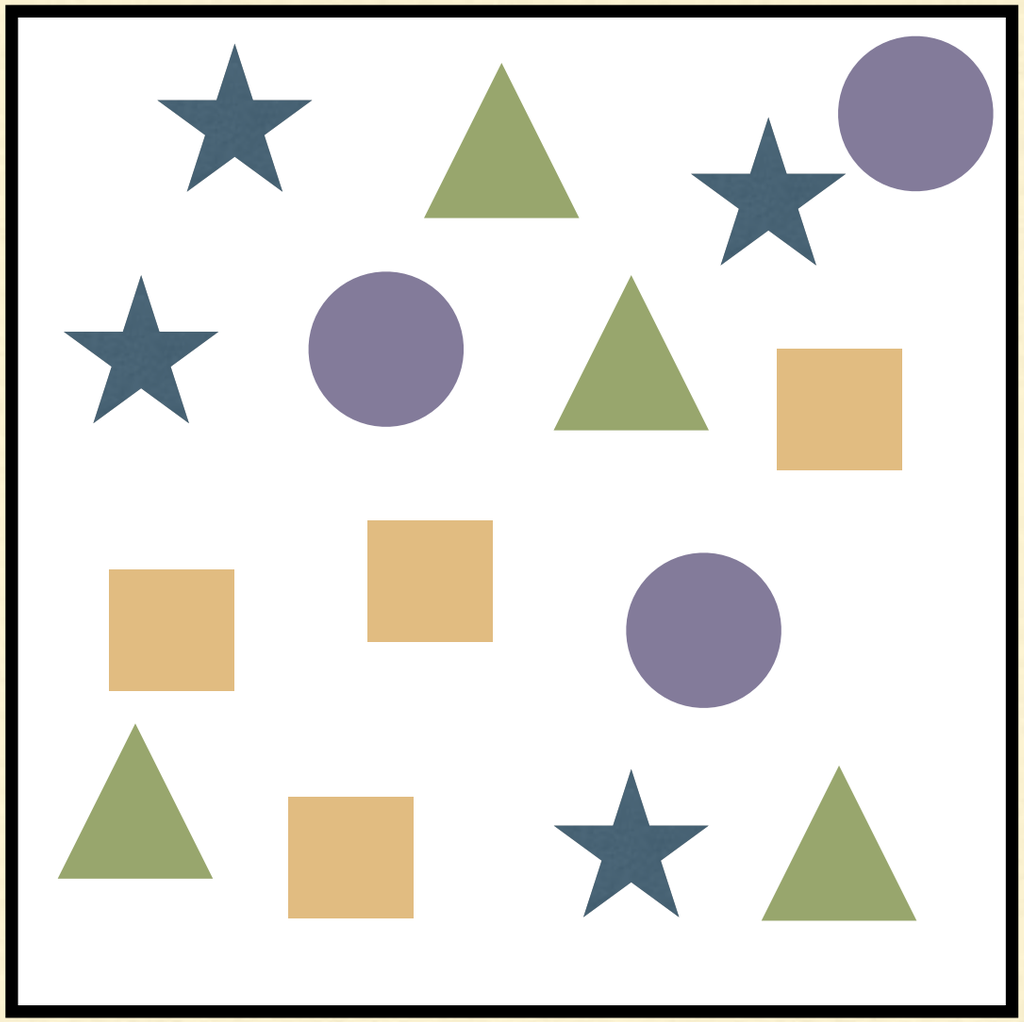
- (1 minute)
  - What are the <u>true</u> relative abundances in the community?
  - What's the (true) richness?
  - Shannon diversity?



EXAMPLE

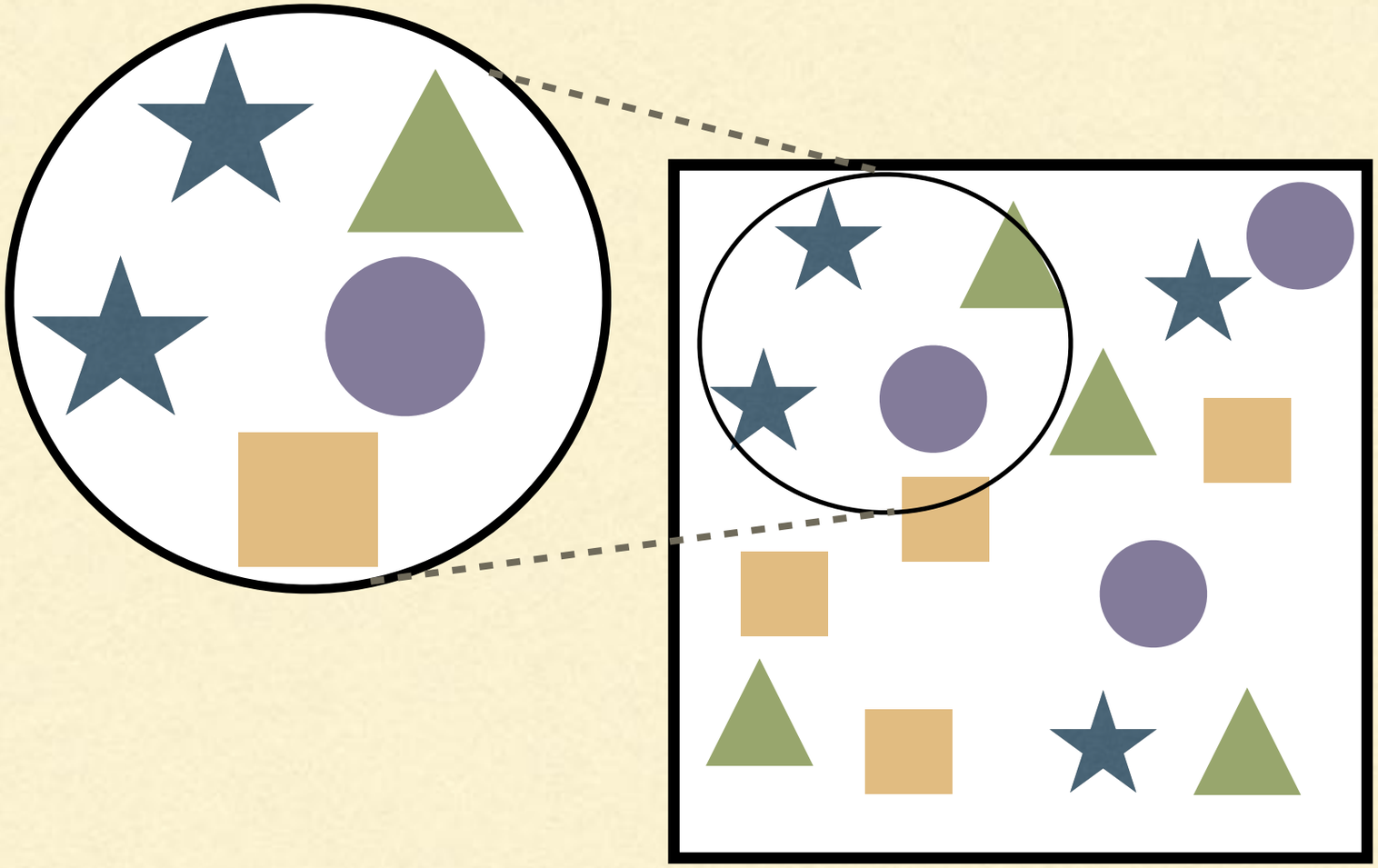
True abundances:

- ★ = 4/15
- ● = 3/15
- ▲ = 4/15
- ■ = 4/15



EXAMPLE

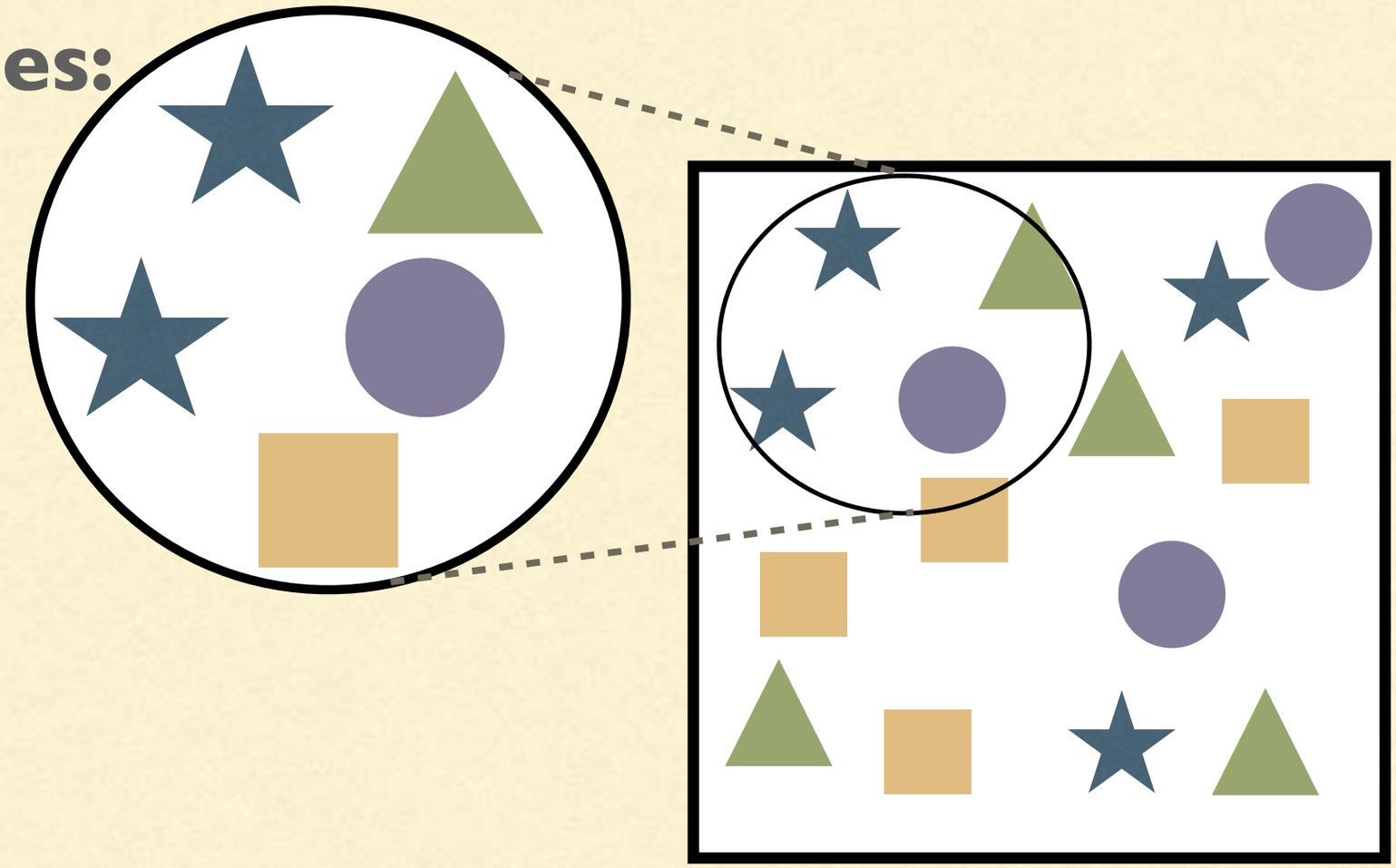
- (3 minutes)
  - Draw some nets. What are the <u>observed</u> relative abundances?



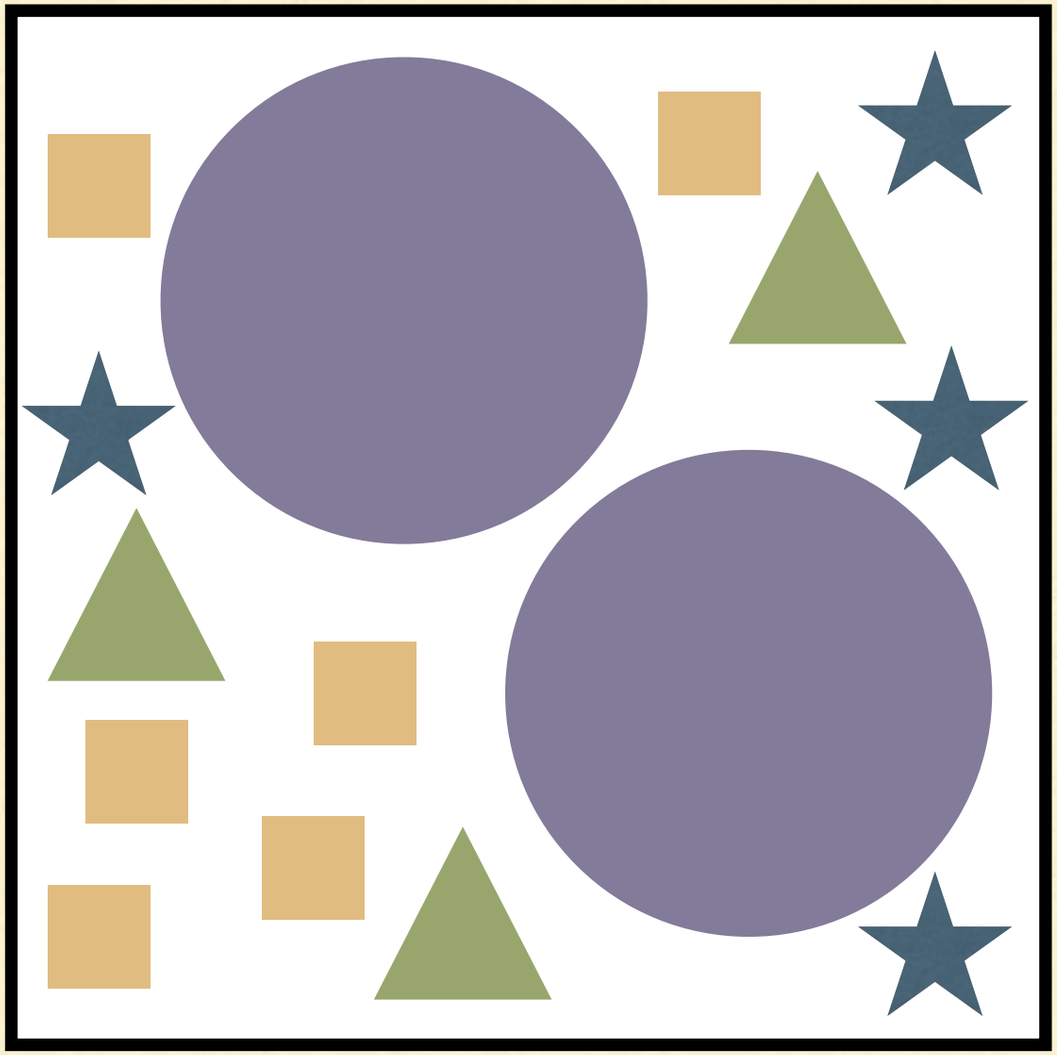
EXAMPLE

Observed abundances:

- ★ = 2/5
- ● = 1/5
- ▲ = 1/5
- ■ = 1/5



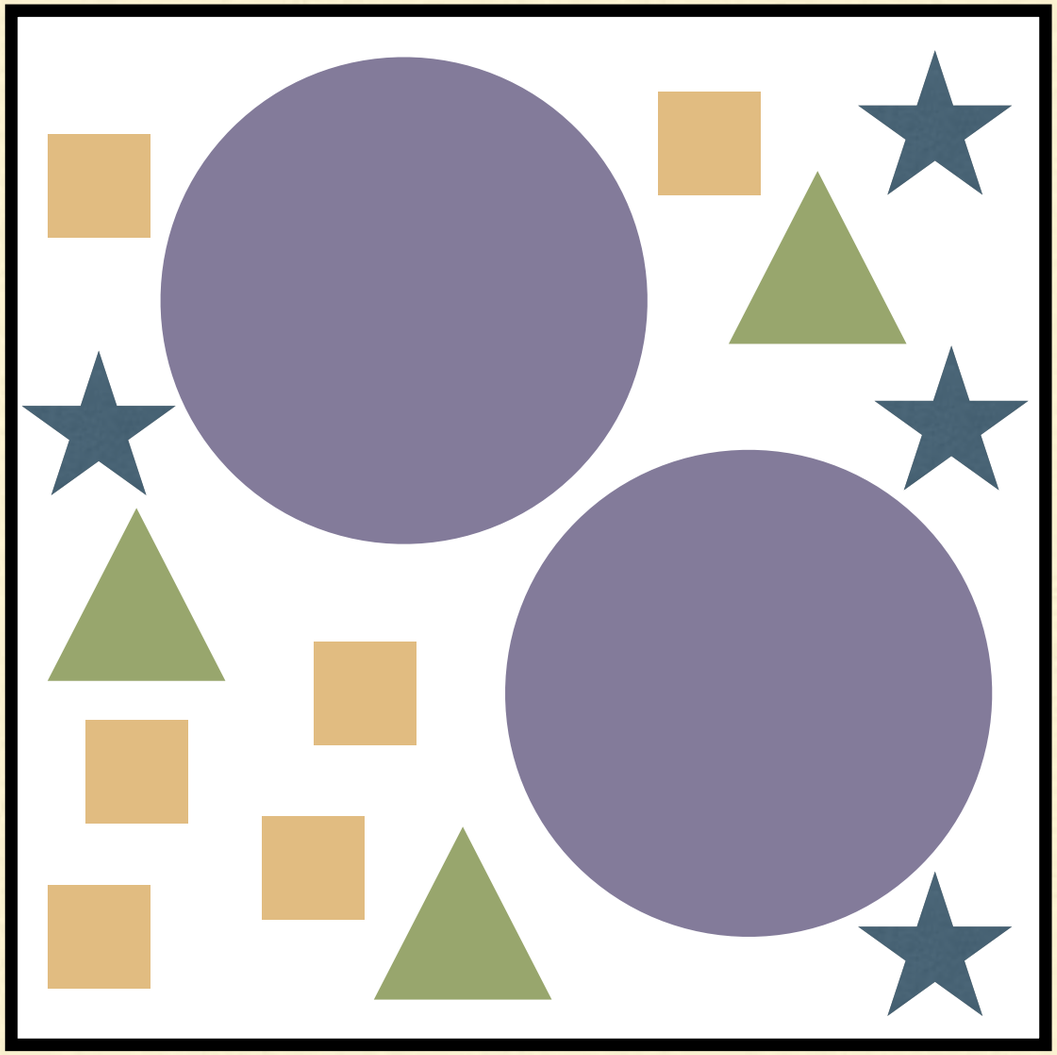
BIAS



- (1 minute)
  - What are the true relative abundances in the community?

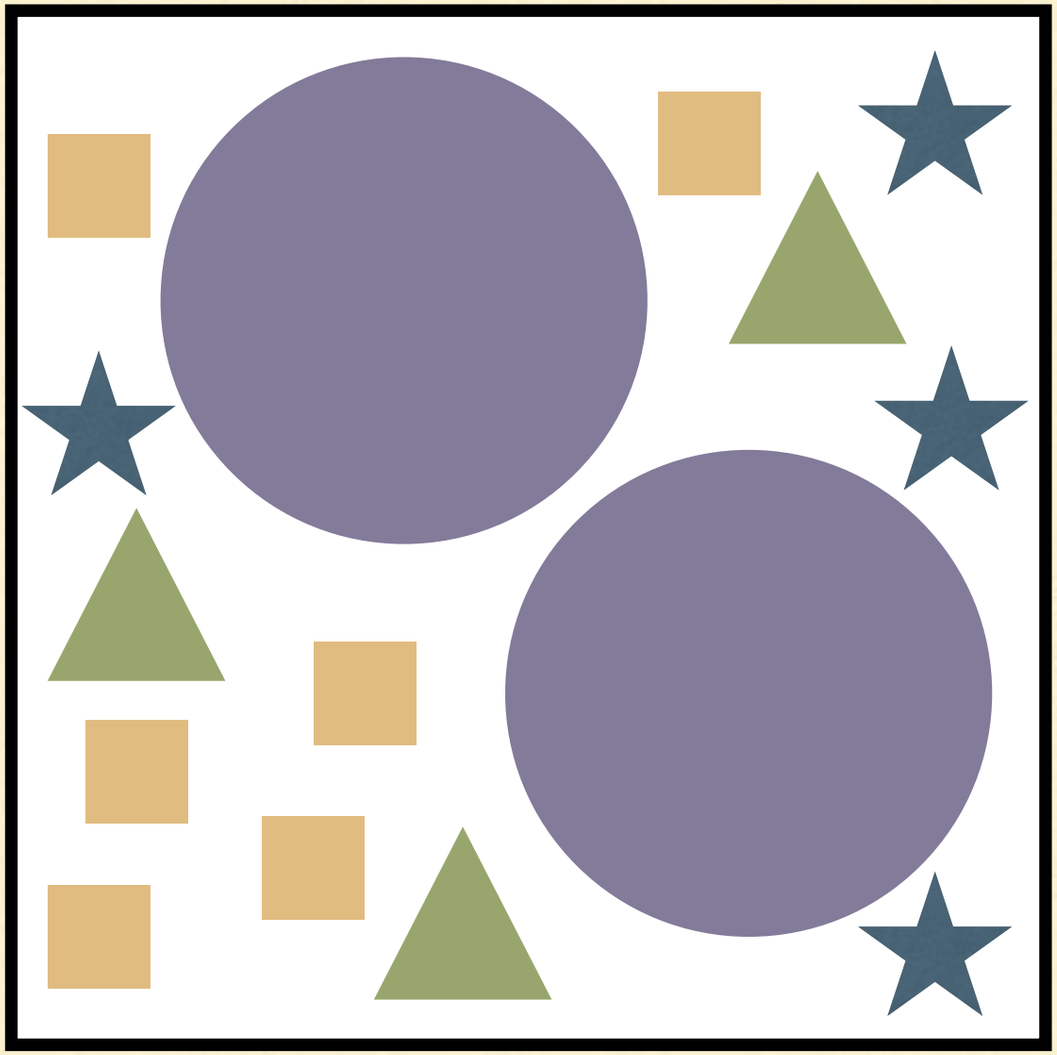
BIAS

- ★ = 4/15
- ● = 2/15
- ▲ = 1/5
- ■ = 2/5



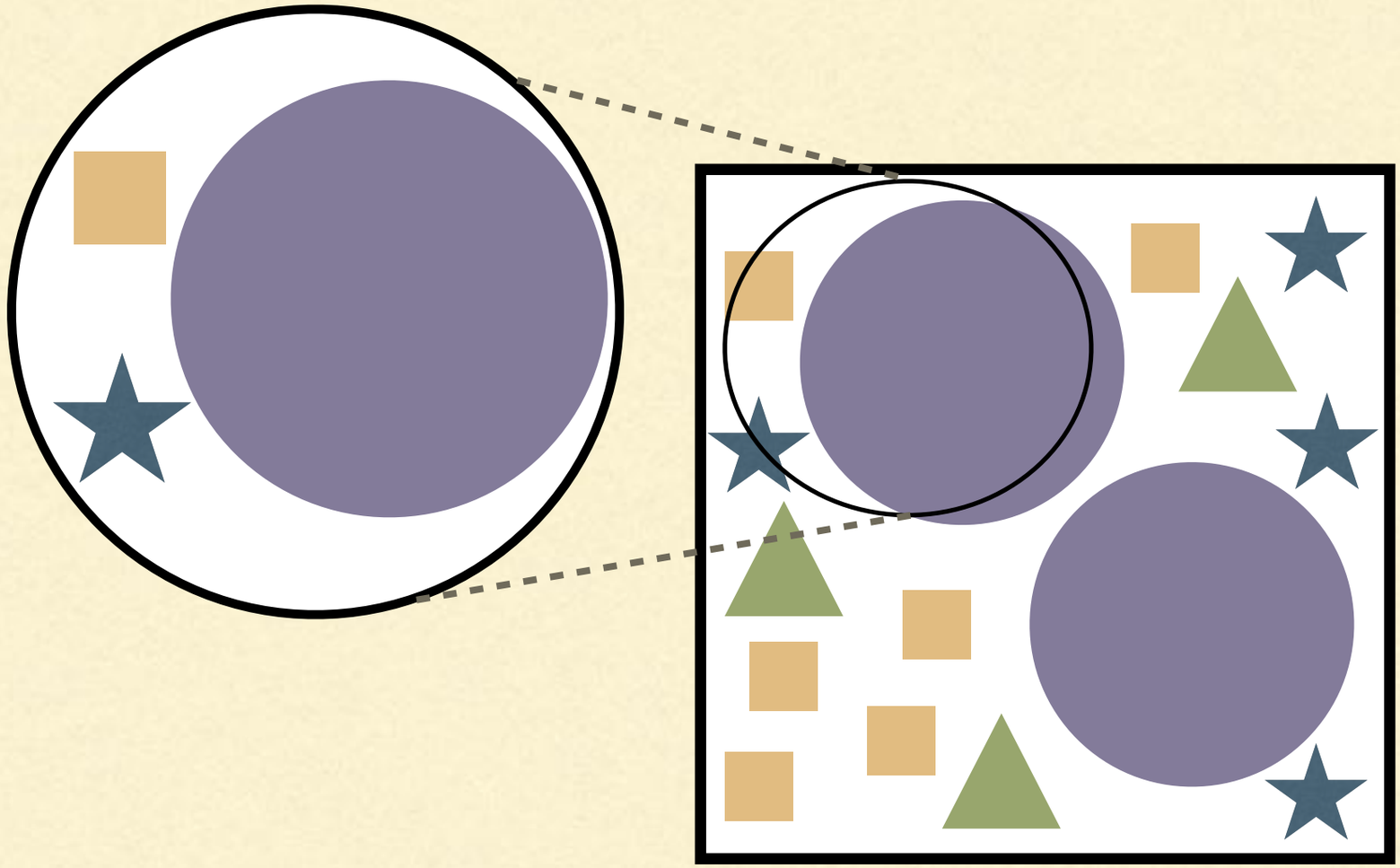
BIAS

- (3 minutes)
  - Draw some nets. What are the observed relative abundances?



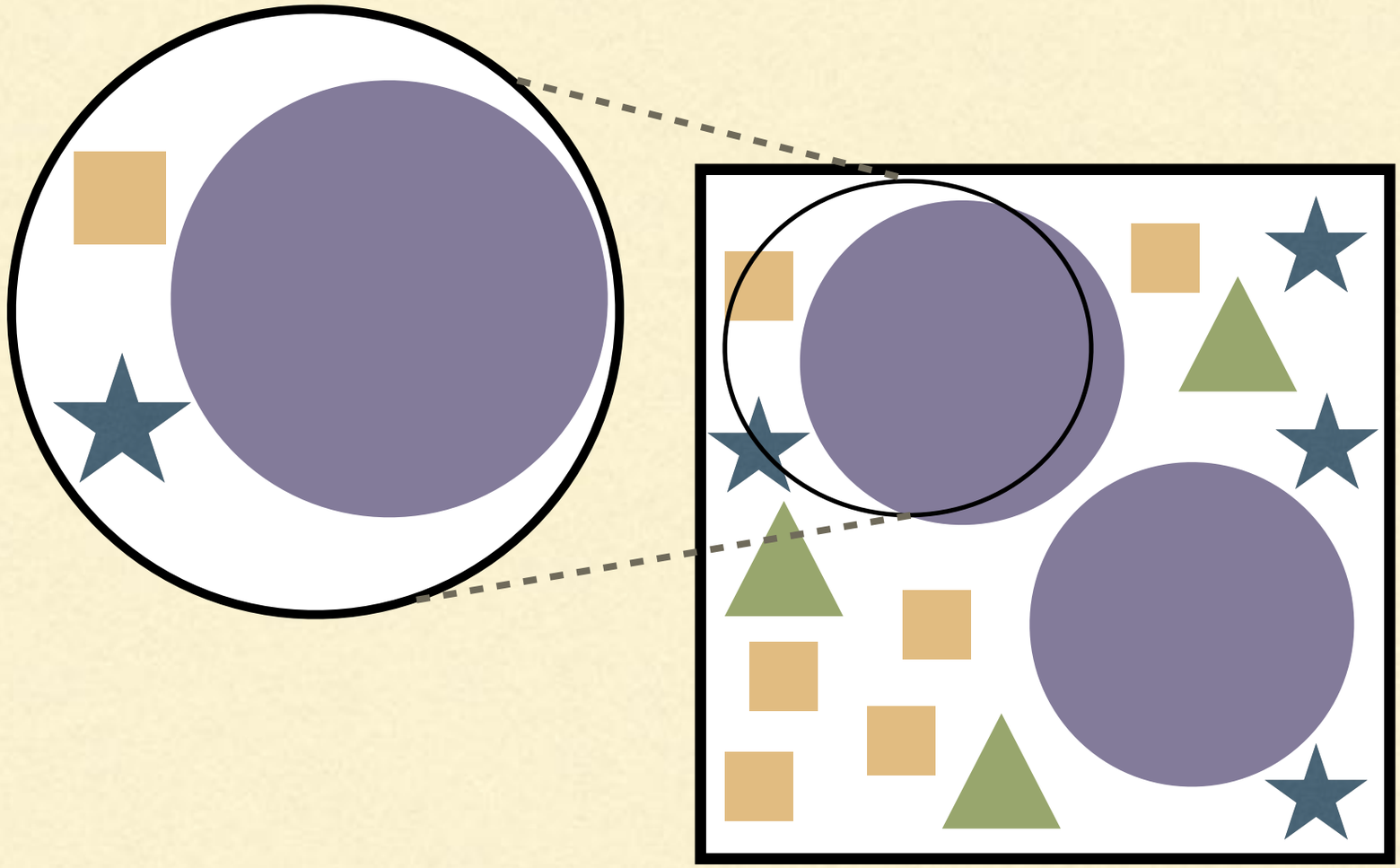
BIAS

- ★ = 1/3
- ● = 1/3
- ■ = 1/3



BIAS

- ★ = 5/15
- ● = 5/15
- ■ = 5/15



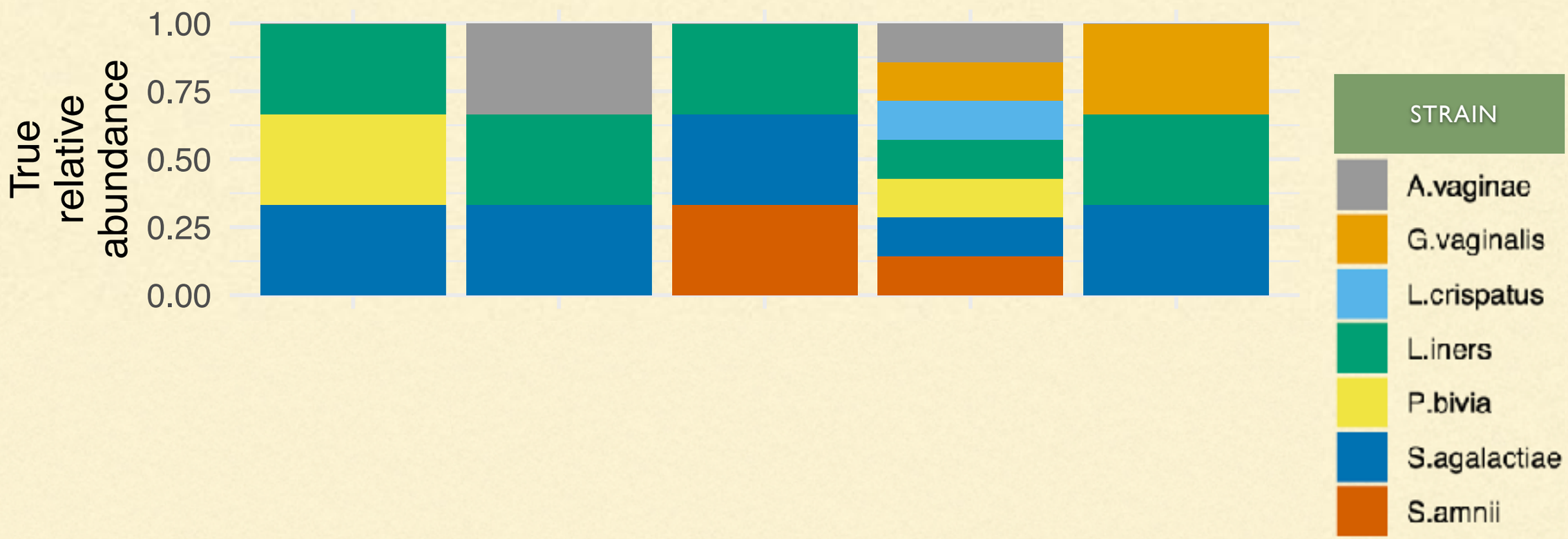
Truth:

★ = 4/15 ● = 2/15 ▲ = 3/15 ■ = 6/15

BIAS AND RELATIVE ABUNDANCE

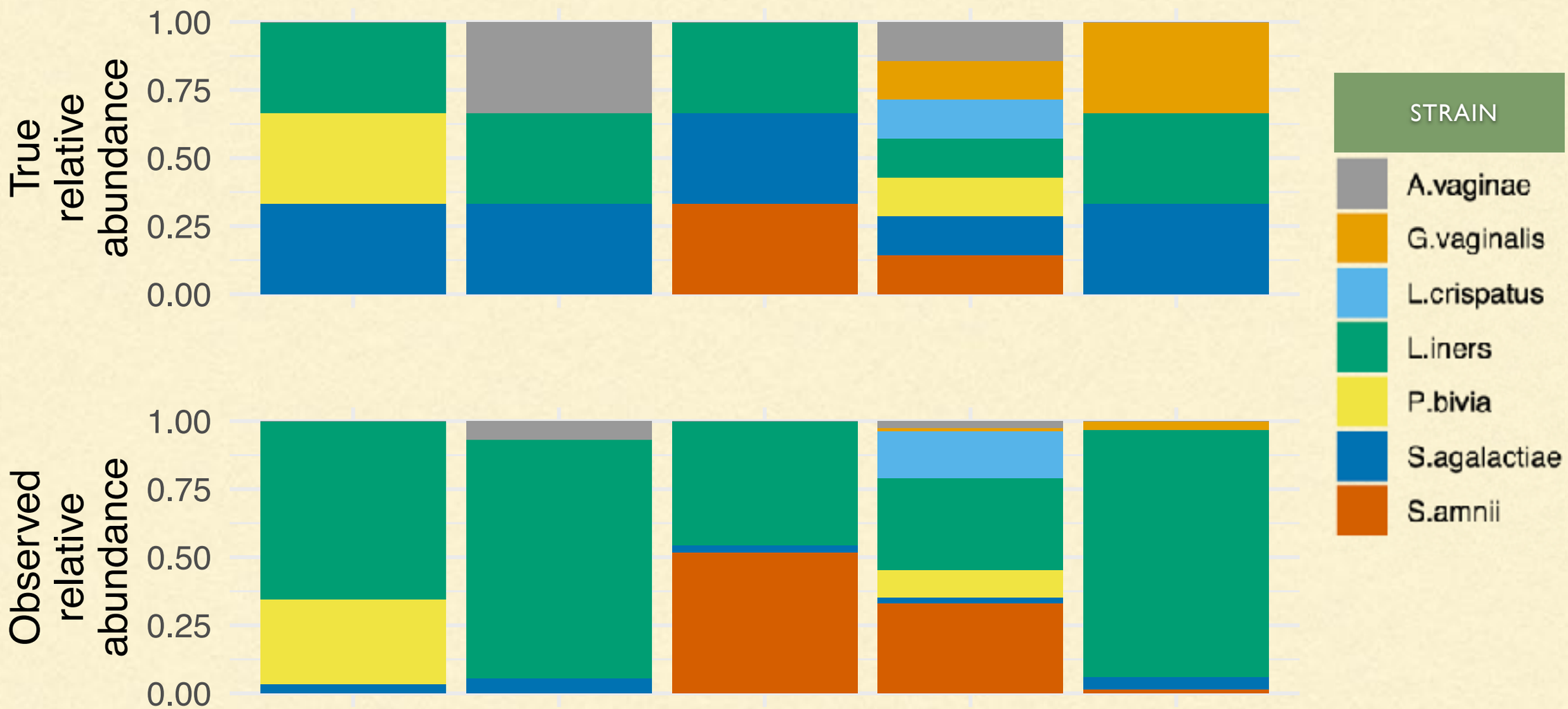
- Size of microbes doesn't drive bias in sequencing
- However, different species have different *catchability* that does drive bias in high throughput sequencing

BIAS AND RELATIVE ABUNDANCE



Data from Brooks *et al* (2015) BMC Microbiology

BIAS AND RELATIVE ABUNDANCE



Data from Brooks *et al* (2015) BMC Microbiology

BIAS AND RELATIVE ABUNDANCE

- What drives taxon-specific *catchability*?
  - Different efficiencies for DNA extraction
  - Differences in primer match
  - Differences in by informatics processes (e.g. representation in databases)

BIAS AND RELATIVE ABUNDANCE

Taxon specific efficiencies/biases/catchabilities

make

sample relative abundance

biased for

actual relative abundance

BIAS AND RELATIVE ABUNDANCE

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Consistent and correctable bias in metagenomic sequencing measurements

Michael R McLaren, Amy D Willis, Benjamin J Callahan

doi: https://doi.org/10.1101/559831

This article is a preprint and has not been peer-reviewed [what does this mean?].

More discussion on this tomorrow night

including how it's relevant to you!

RAREFACTION

- Rarefaction is another example of bias in sequencing
- Sample species richness underestimates total richness
  - Rarefying further underestimates total richness
    - More on this next Sunday...

BIAS

- Some estimates are biased
- Biased estimates are not good, especially if the bias depends on sample size
- The solution depends on the problem!
  - More discussion on Sunday when we talk about estimation

SUMMARY SO FAR

- *Statistical thinking*
  - Why we need models
  - Parameters of models
  - Estimates of parameters
  - Bias as a way to evaluate estimators

QUESTIONS

- We're about to take a break
- Any questions before we break?

BREAK

EVALUATING ESTIMATORS

- We want estimators to be
  - Accurate = correct on average = unbiased
  - Precise = usually close to its average = low variance
- Bias and variance are two criteria for evaluating estimators

EVALUATING ESTIMATORS

- e.g. Rarefying for diversity is low variance, high bias
- e.g. Estimating diversity with breakaway and DivNet is higher variance, lower bias
- You need to understand the tradeoffs of your analysis
  - More specifics next Sunday

VARIANCE

- Variance is a property of the estimate
- It describes how much the estimate varies
- Variance actually isn't about the parameter
- Definition:

$$\text{Variance}(\hat{\theta}) = \mathbb{E}\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2$$

VARIANCE IN REAL LIFE

- Definition:

$$\text{Variance}(\hat{\theta}) = \mathbb{E}\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2$$

- What does this expectation mean *in terms of your experiment*?

VARIANCE

- The variance reflects how far apart repeated estimates are
- If your estimates (from repeated experiments) are
  - 12, 12, 12, 12, 12… => variance is 0
  - 12, 12, 12, 13, 12… => variance is 0.2
  - 12, 12, 12, 13013, 12… => variance is 33805200
- A large change in the estimates equals a large variance

VARIANCE

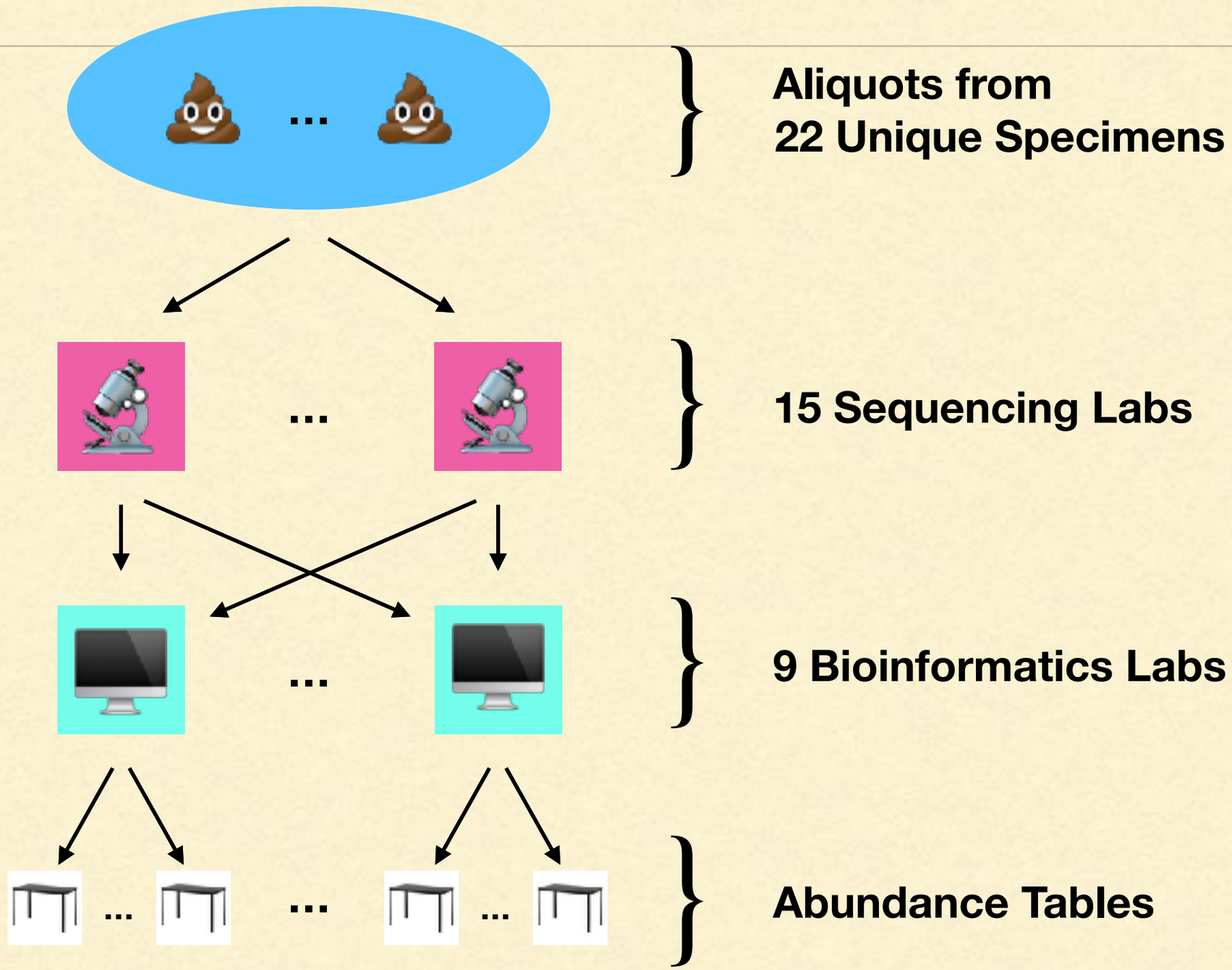
- Do the experiment, calculate the estimate: $\hat{\theta}_1$
- Repeat the experiment, calculate the estimate: $\hat{\theta}_2$
- …
- Let $\hat{\theta}_j$ be your estimate from the j-th time you do the experiment

$$variance = average_j(\hat{\theta}_j - average(\hat{\theta}))^2$$

VARIANCE

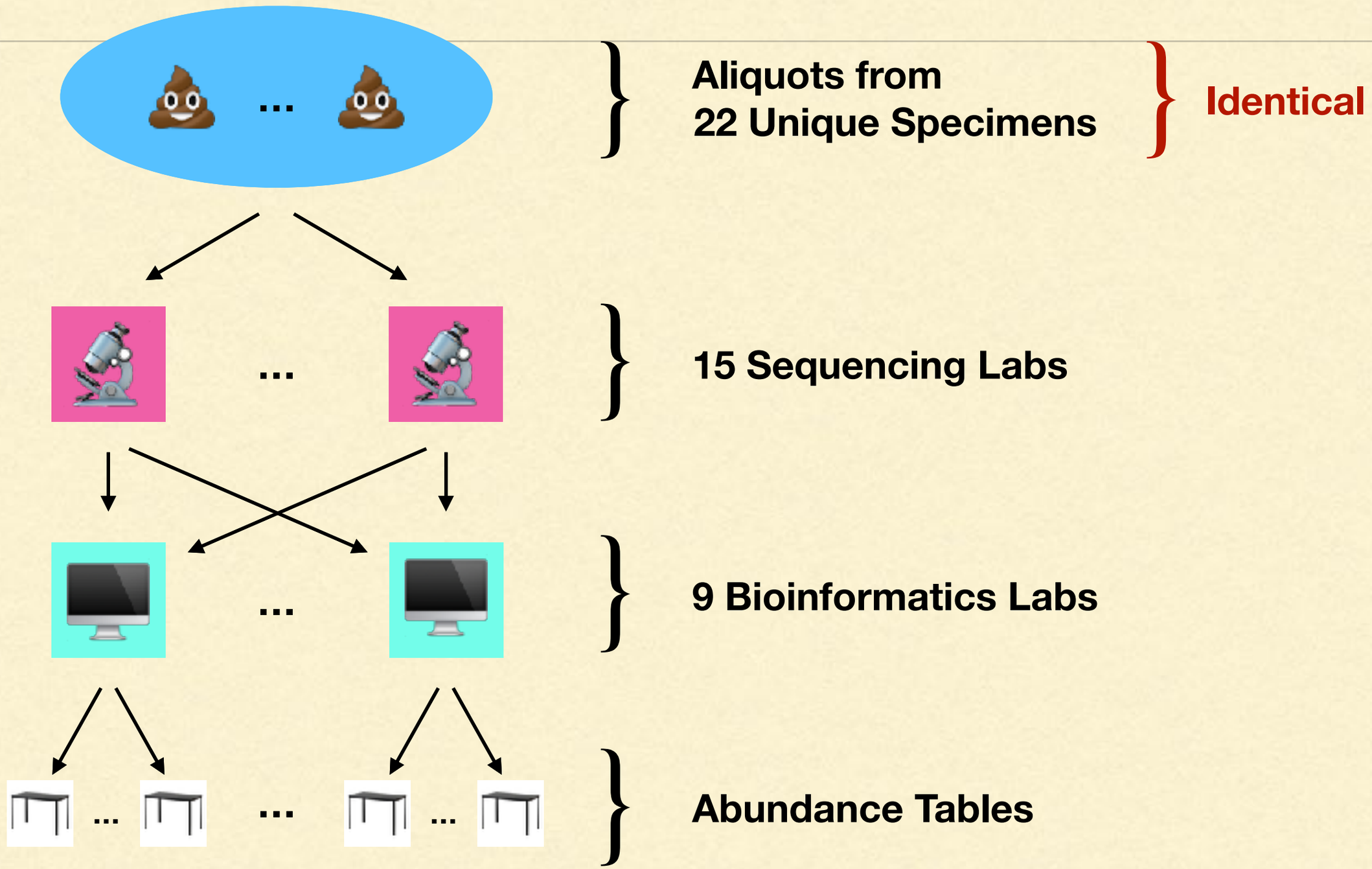
- What does “repeat the experiment” mean?
  - Expensive, impractical…
  - We rarely do this
- Interesting study: Microbiome Quality Control Project (MBQC)

MBQC DATA



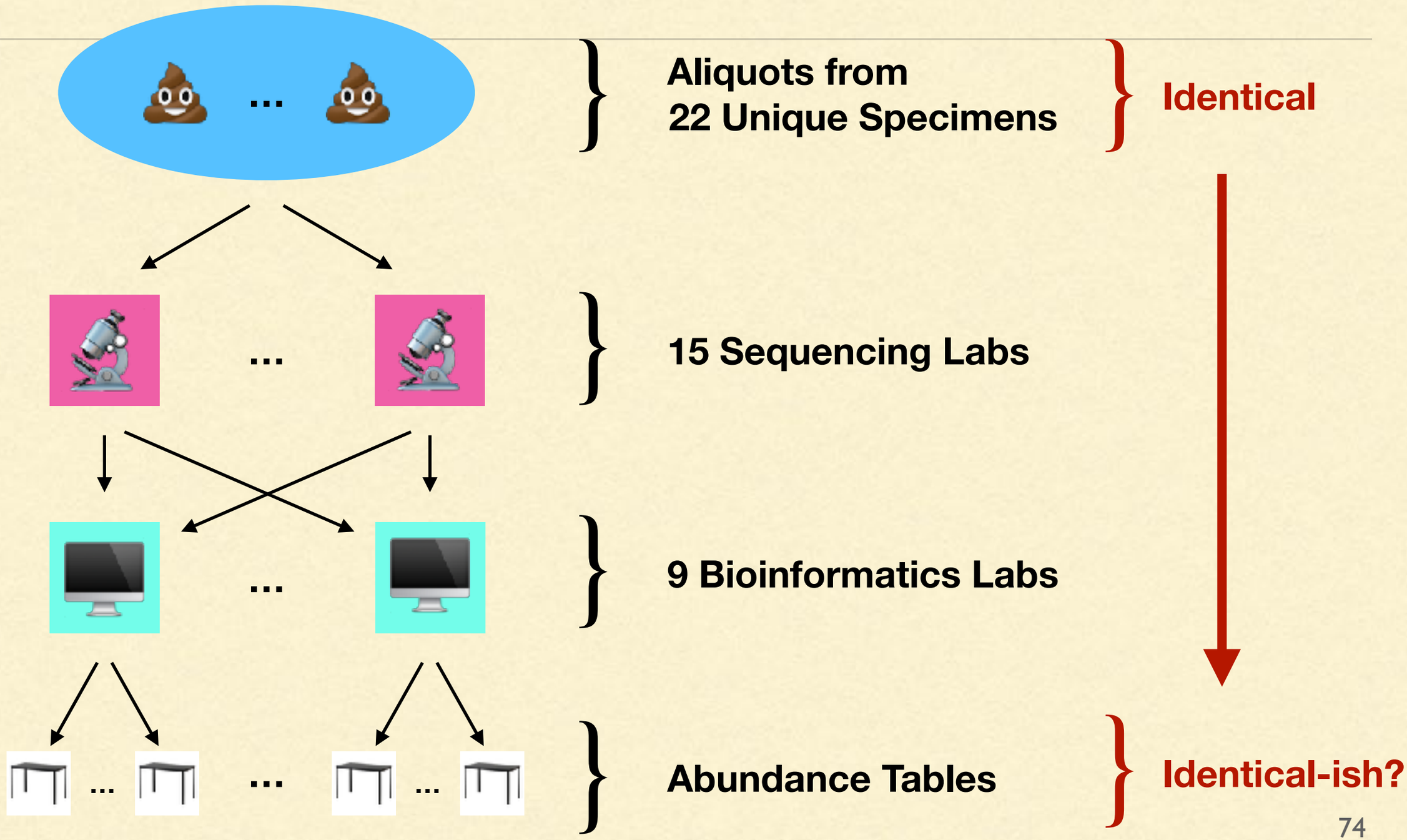
Data from Sinha *et al* (2017) *Nature Biotech*

MBQC DATA



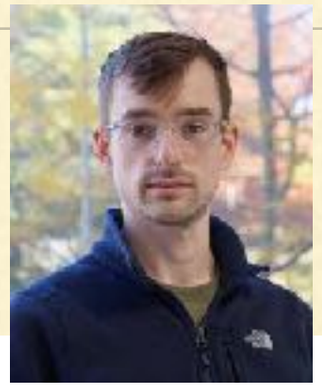
Data from Sinha *et al* (2017) *Nature Biotech*

MBQC DATA

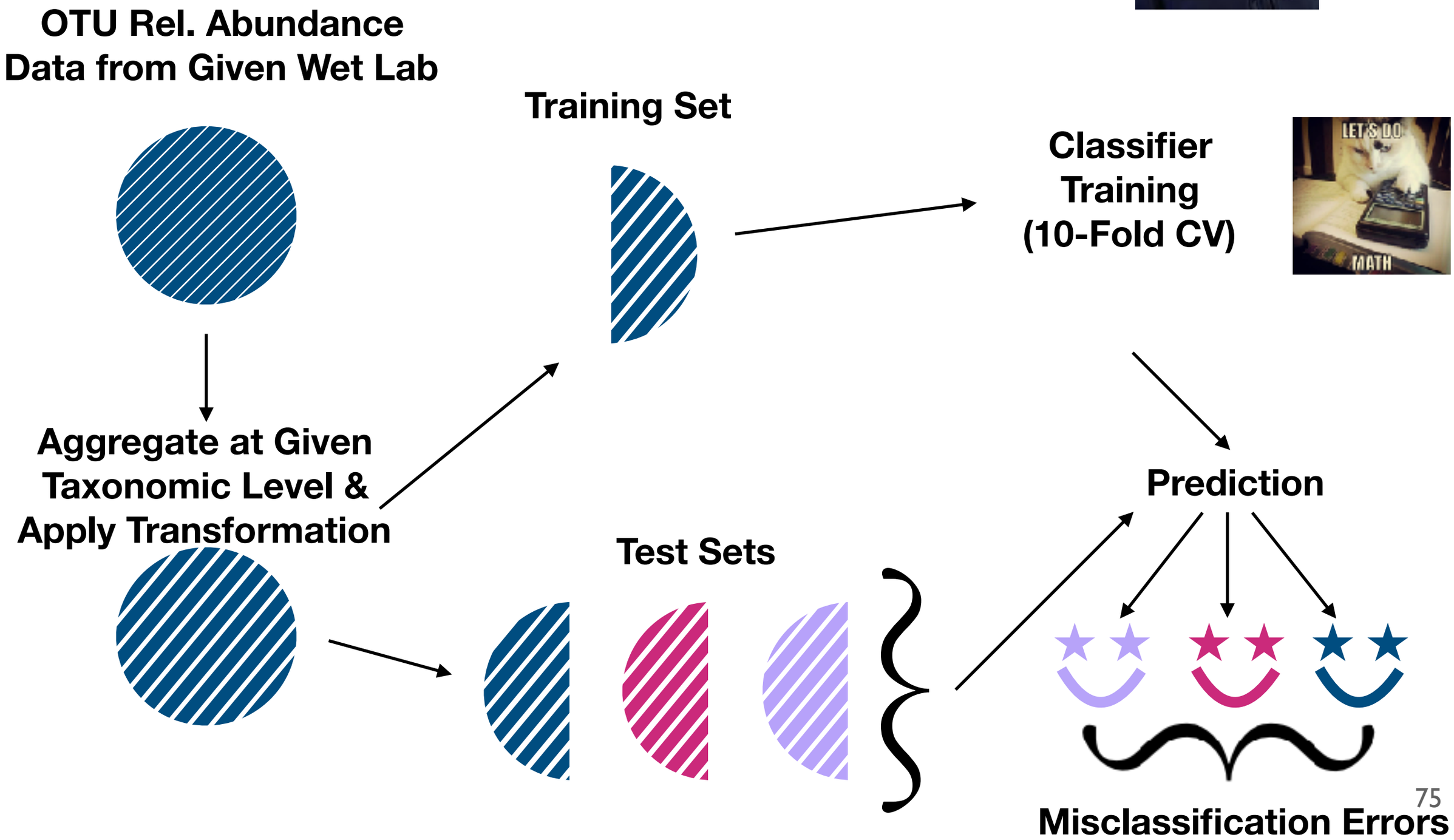


Data from Sinha *et al* (2017) *Nature Biotech*

APPROACH

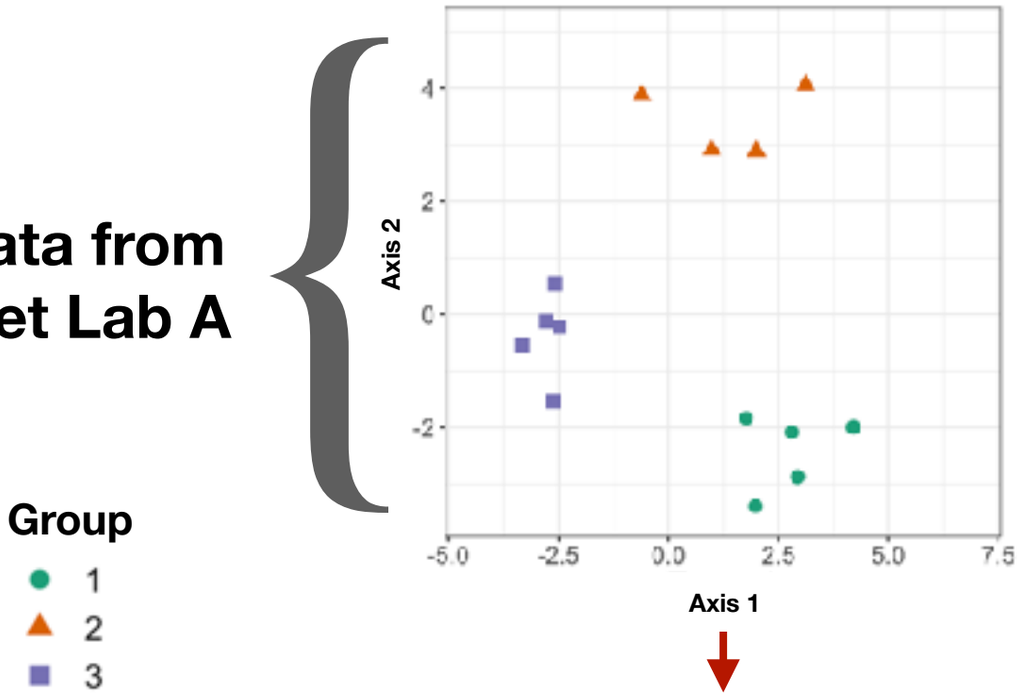


David Clausen (UW)

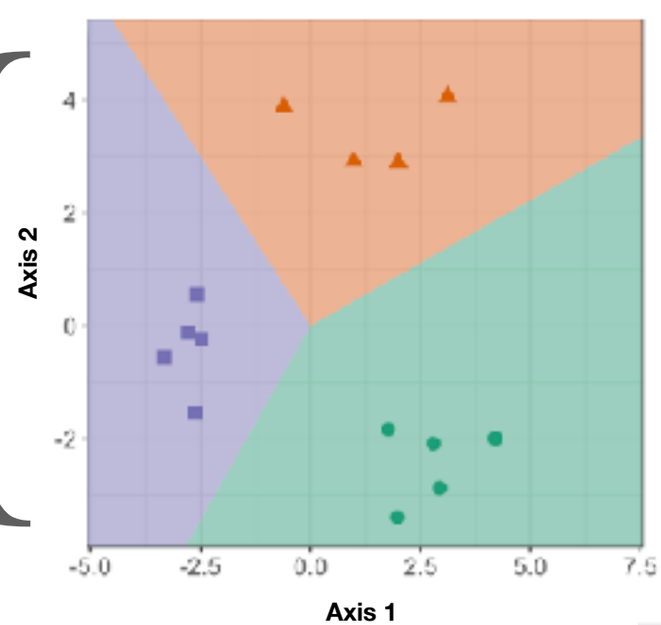


WHY CLASSIFICATION?

Data from Wet Lab A

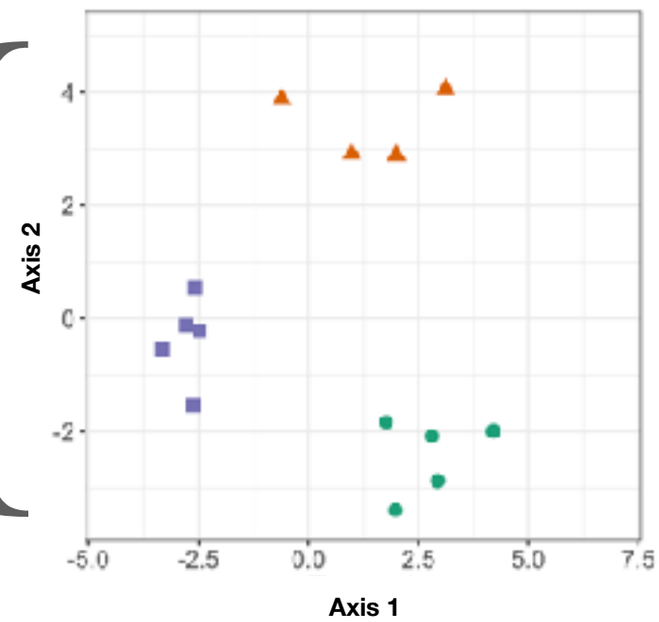


Class. Rule Based on Data from Wet Lab A



WHY CLASSIFICATION?

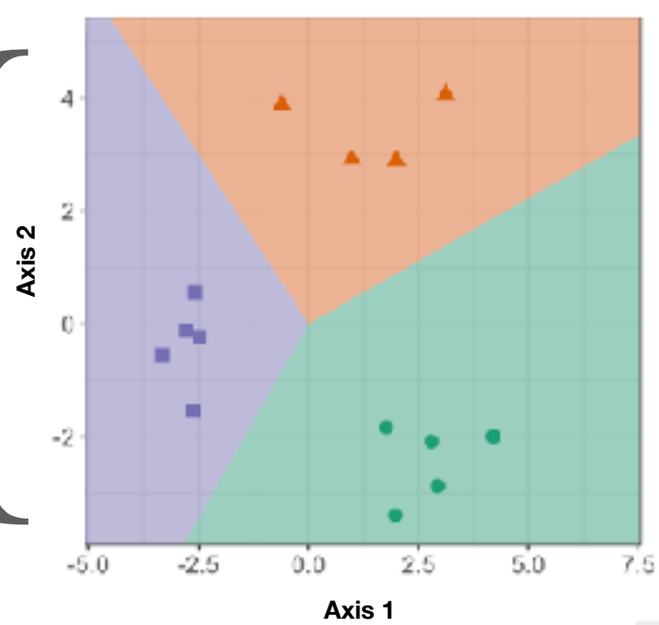
Data from Wet Lab A

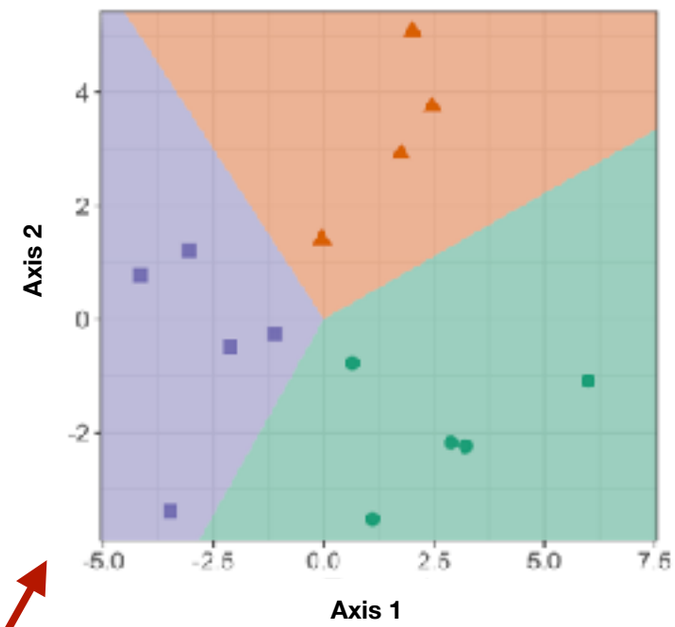


Group

- 1
- 2
- 3

Class. Rule Based on Data from Wet Lab A



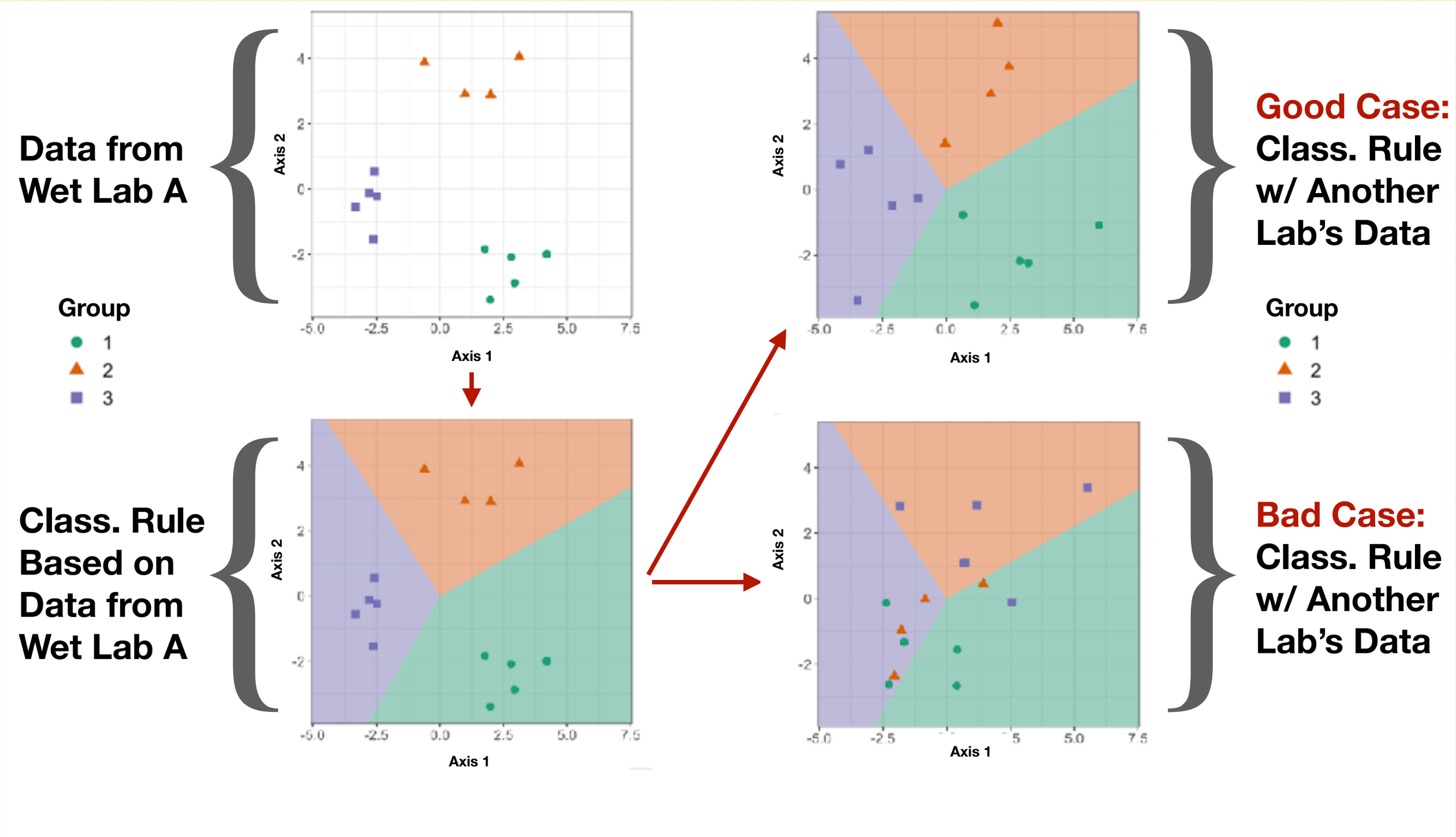


Good Case: Class. Rule w/ Another Lab's Data

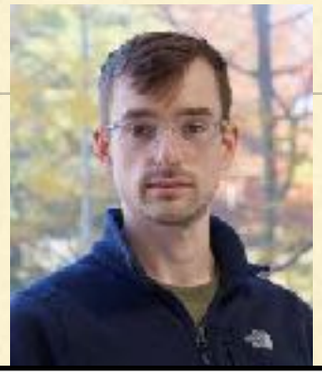
Group

- 1
- 2
- 3

WHY CLASSIFICATION?

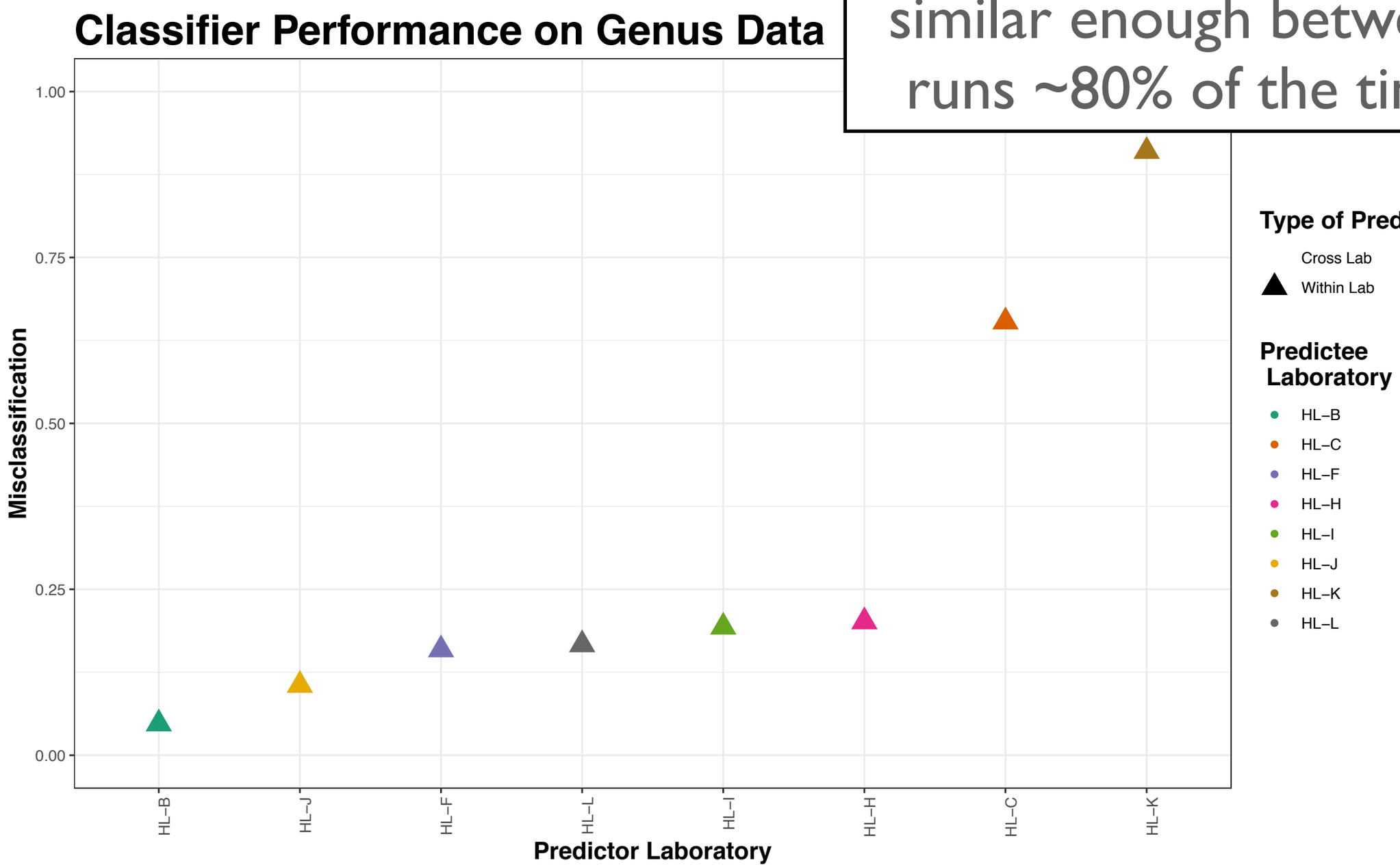


WITHIN-LAB VARIATION



David Clausen (UW)

Identical specimens are similar enough between runs ~80% of the time

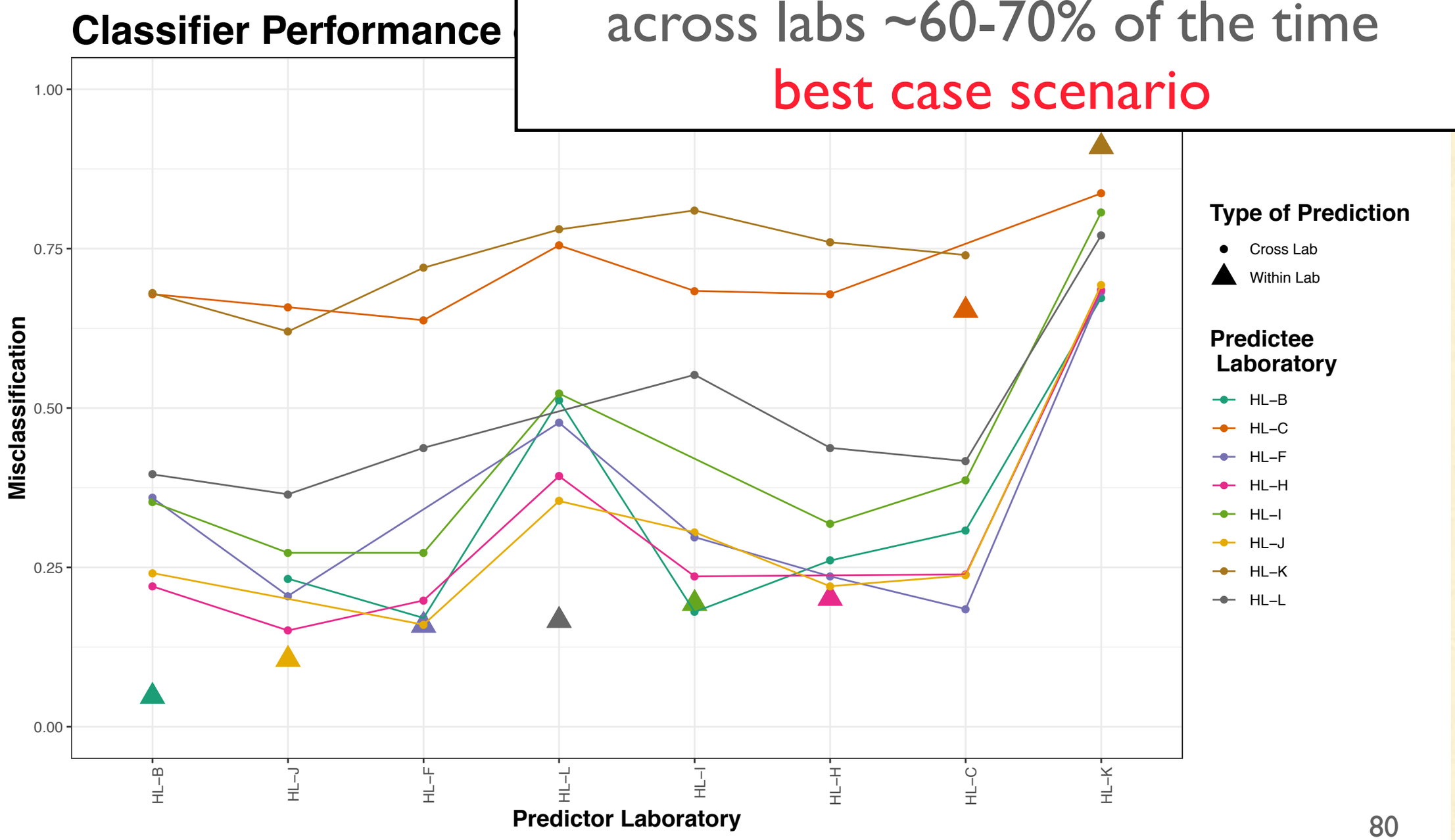


ACROSS-LAB VARIATION

David Clausen (UW)

Identical specimens are similar enough across labs ~60-70% of the time

best case scenario



SOURCES OF VARIANCE

- The technical variation that you see is far less than the technical variation that exists
- Good within-laboratory performance does not guarantee signal will replicate well in ***any*** other laboratory
- This is all due to measurement error — no sampling variability!

SOURCES OF VARIANCE

- MBQC was unique — usually difficult to know how much variation exists due to lab
- Usually our baseline variance (noise) is within-lab variance
  - Variance is usually understated
  - If you can't reproduce a published result, do not exclude technical variation as the cause

SOURCES OF VARIANCE

You need multiple approaches to validate your findings!

A single approach (16S/qPCR/shotgun/longread) is not sufficient to make conclusions!

STANDARD DEVIATION

- The standard deviation is the square root of the variance
  - Sometimes it's more convenient to work on the original scale of the data

- *What is a standard error?*

STANDARD ERROR

- **standard error = estimate of standard deviation**
- Distinction:
  - If you have one-dimensional data, it has a standard deviation
  - If you have a one-dimensional estimate, it has a standard error

STANDARD ERRORS

- "Model-based standard errors"
- are based on models!
- Your standard error is only as good as your model!

VARIANCE AND HYPOTHESIS TESTS

- Why is estimating variance important?
- Hypothesis testing!
- Most hypothesis tests take the form

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

VARIANCE AND HYPOTHESIS TESTS

- Wald test statistic

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

- Suppose your variance is half what it should be
- What happens to p-values?

WHAT IS A P-VALUE?

WHAT IS A P-VALUE?

- Tells us whether our result/hypothesis is right or wrong?

WHAT IS A P-VALUE?

- ~~Tells us whether our result/hypothesis is right or wrong?~~

WHAT IS A P-VALUE?

- ~~Tells us whether our result/hypothesis is right or wrong?~~
- Probability the null hypothesis is true?

WHAT IS A P-VALUE?

- ~~Tells us whether our result/hypothesis is right or wrong?~~
- ~~Probability the null hypothesis is true?~~

WHAT IS A P-VALUE?

- ~~Tells us whether our results/hypothesis is right or wrong?~~
- ~~Probability the null hypothesis is true?~~
- Inverse probability of a publication?

WHAT IS A P-VALUE?

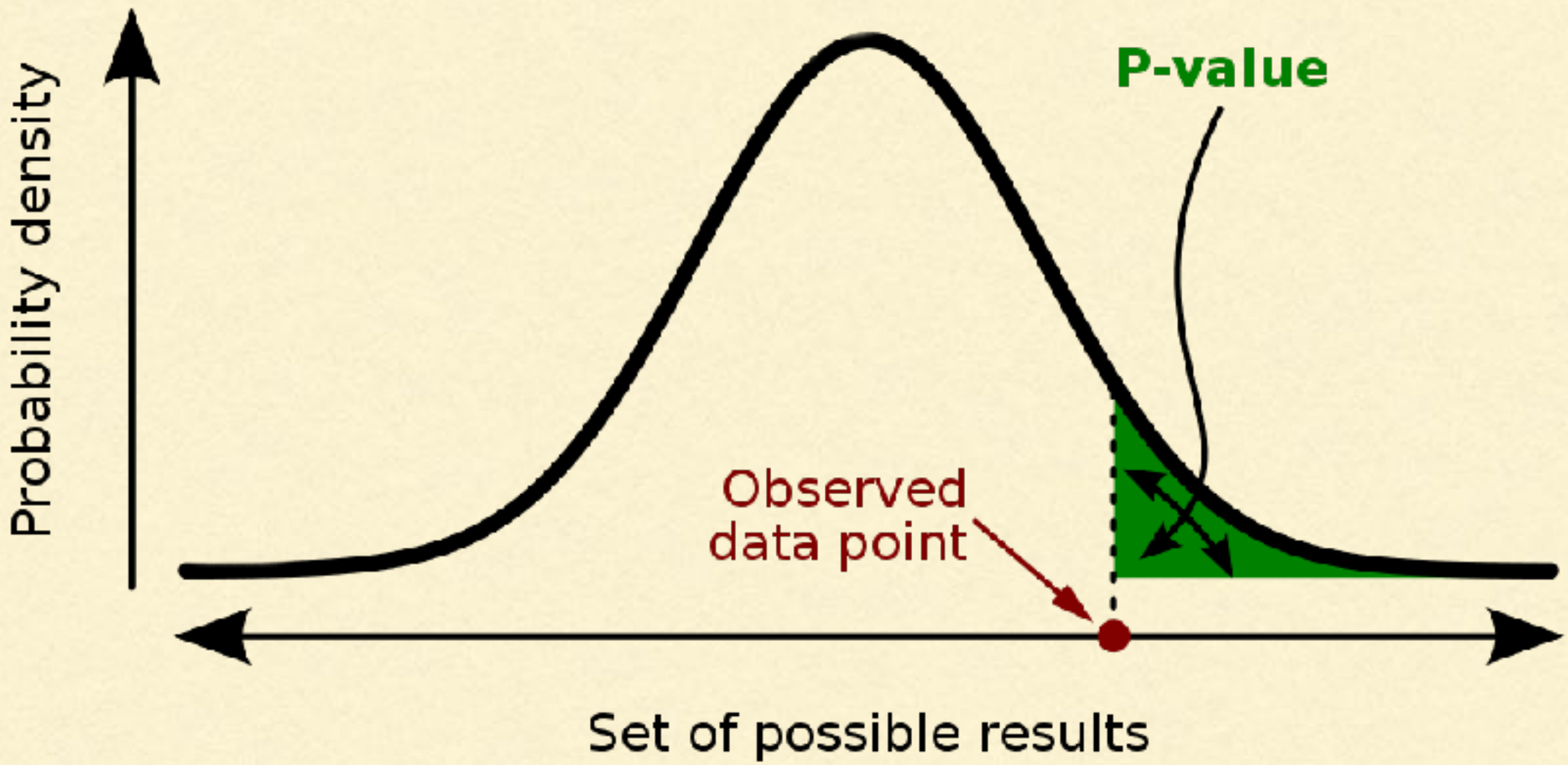
- ~~Tells us whether our results/hypothesis is right or wrong?~~
- ~~Probability the null hypothesis is true?~~
- ~~Inverse probability of a publication?~~

WHAT IS A P-VALUE?

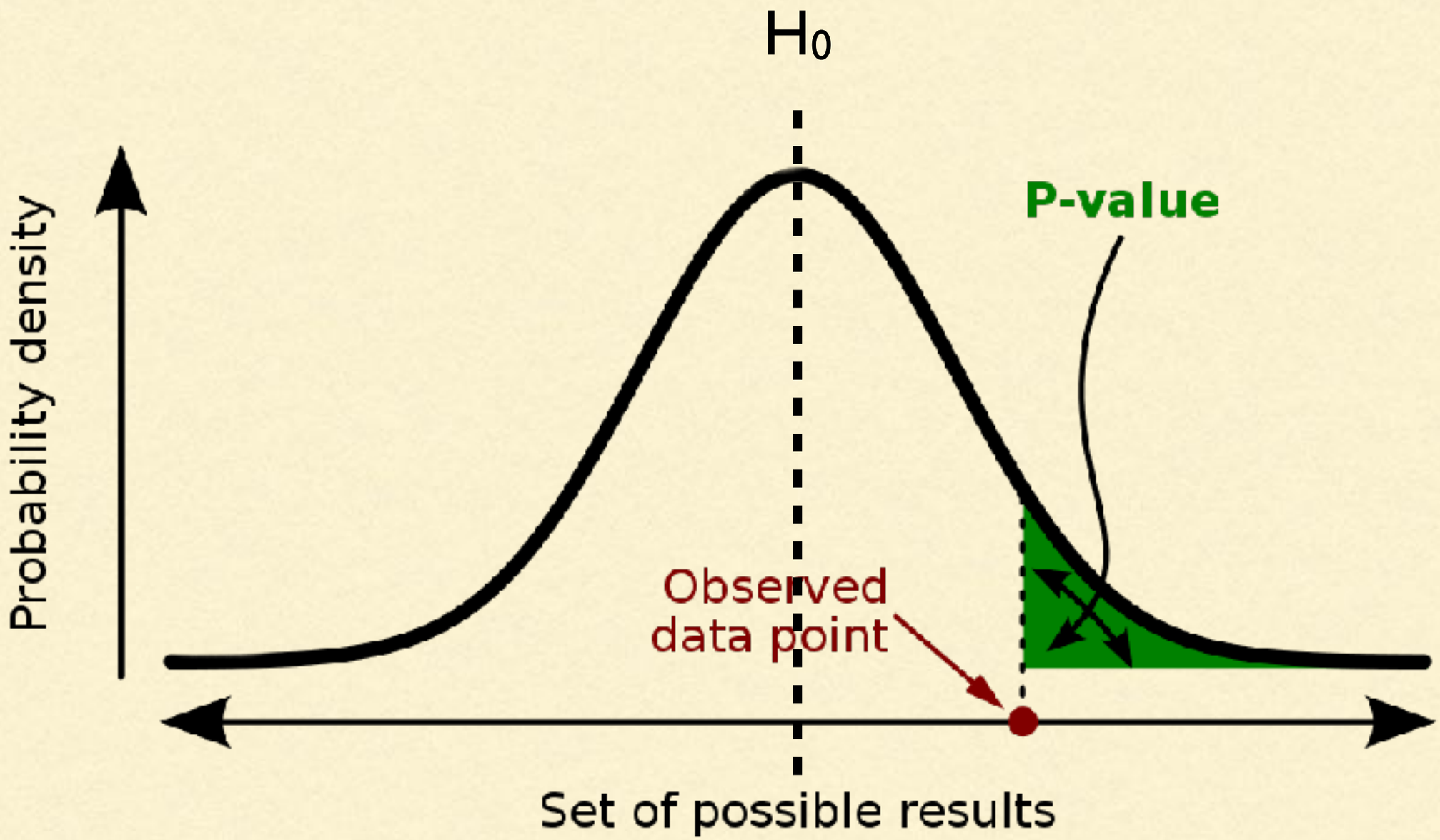
- ~~Tells us whether our results/hypothesis is right or wrong?~~
- ~~Probability the null hypothesis is true?~~
- ~~Inverse probability of a publication?~~
- You need statistics?

p-value: probability of observing a result as or more extreme than what we observed, assuming the null hypothesis is true

p-value: probability of observing a result *as or more extreme* than what we observed, assuming the null hypothesis is true


P-value
Probability density
Observed
data point
Set of possible results

p-value: probability of observing a result as or more extreme than what we observed, *assuming the null hypothesis is true*


H0
P-value
Probability density
Observed
data point
Set of possible results

p-value: probability of observing a result as or more extreme than what we observed, assuming the null hypothesis is true

Premise: If our results are extreme under the assumption of the null hypothesis, then perhaps our assumption is wrong

CONFIDENCE INTERVALS

- Interval estimate: estimate $\pm z \times$ s.e.
- Inform us about *uncertainty* in our estimates
- Provides more context than p-values alone
  - Statistical significance and scientific significance
  - Can give a sense of model fit and parameter interpretation

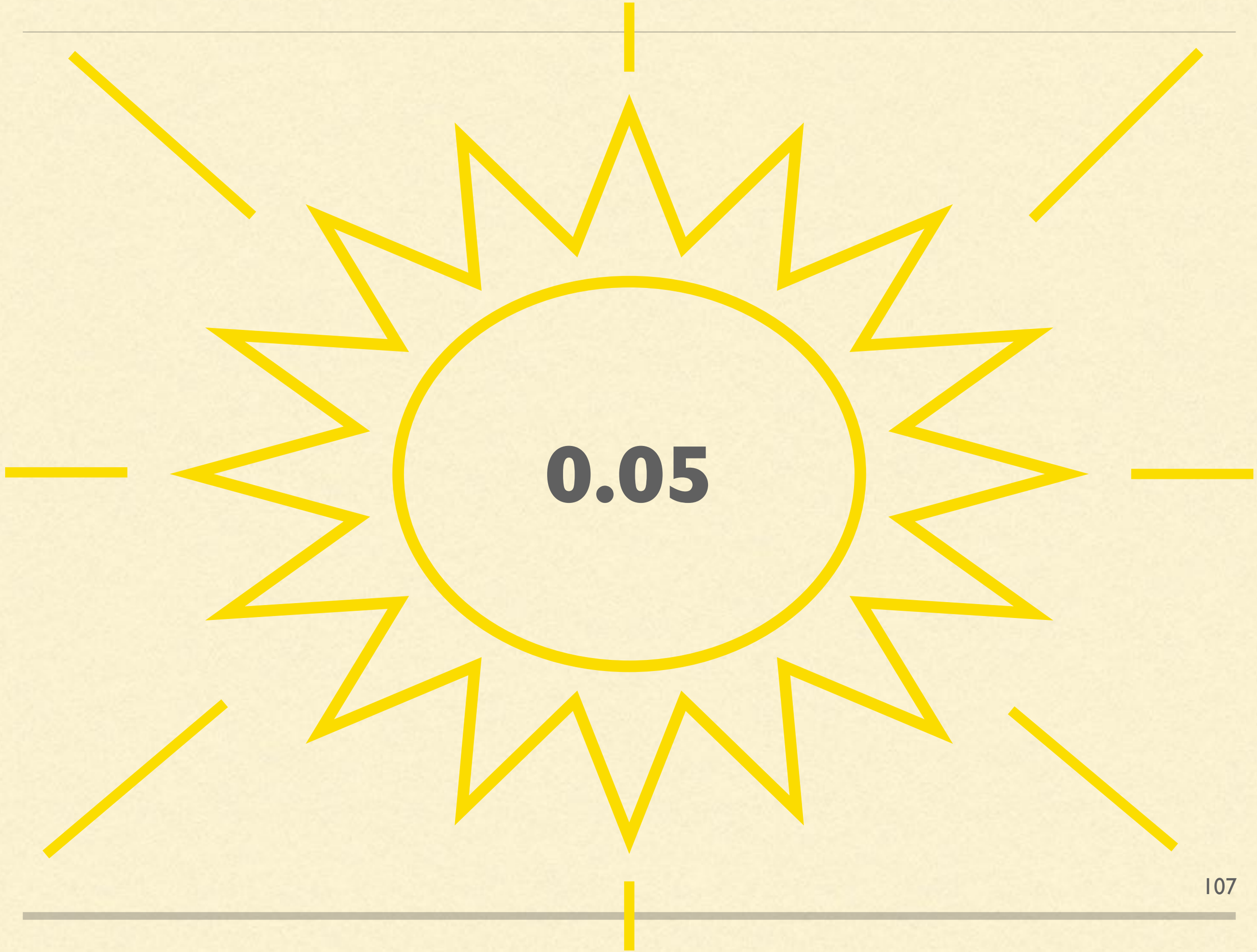
TYPE I ERROR

WHY WE SHOULD CAUTIOUS ABOUT SMALL P-VALUES

TYPE I ERROR

$$\alpha = P(\text{reject } H_0 \,|\, H_0)$$

$$1 - \alpha = P(\text{don't reject } H_0 \,|\, H_0)$$


0.05

TYPE I ERROR

$$.05 = P(\text{reject } H_0 \mid H_0)$$

$$.95 = P(\text{don't reject } H_0 \mid H_0)$$

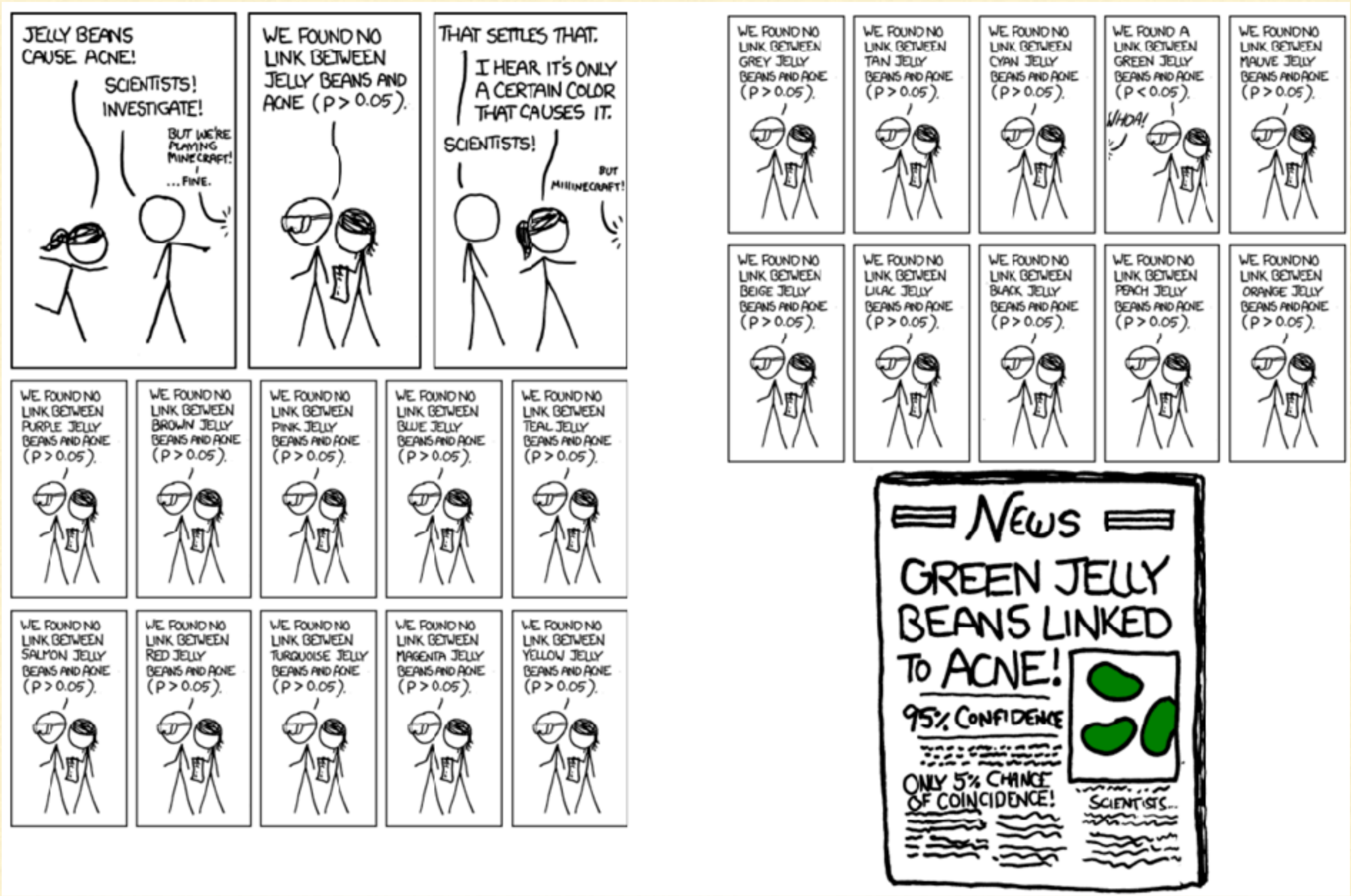
TYPE I ERROR

Type I Error

$$.05 = P(\text{reject } H_0 \mid H_0)$$

"false positive"

$$.95 = P(\text{don't reject } H_0 \mid H_0)$$


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

MULTIPLE TESTING

- 2 independent tests:
  - Don't reject H_0 for Test 1 $= .95$
  - Don't reject H_0 for Test 2 $= .95$
  - Don't reject H_0 for both tests $= .95 \times .95 = .9025$
- Don't make any type 1 errors: ~90%

MULTIPLE TESTING

- 3 independent tests:
  - Don't reject H_0 for all tests $= .95 \times .95 \times .95 = .8574$
- Don't make any type 1 error: 86%

MULTIPLE TESTING

m (independent) tests

$$(1-\alpha)^m = P(\text{don't reject any } H_0 \,|\, \text{all } H_0)$$

$$1-(1-\alpha)^m = P(\text{reject any } H_0 \,|\, \text{all } H_0)$$

MULTIPLE TESTING

m (independent) tests

$$(.95)^m = P(\text{don't reject any } H_0 \,|\, \text{all } H_0)$$

$$1 - (.95)^m = P(\text{reject any } H_0 \,|\, \text{all } H_0)$$

MULTIPLE TESTING

m	$1 - (.95)^m$
10	0.40
100	0.994
1,000	~1

MULTIPLE TESTING

- Two approaches to control type 1 error rate
  - **Family-wise Error Rate (FWER):** probability of at least one type 1 error
  - **False Discovery Rate (FDR):** the expected proportion of type 1 errors among the rejected hypotheses
- **You decide**: do you want to control FWER or FDR?

TYPE 2 ERROR

WHY WE SHOULD CAUTIOUS ABOUT LARGE P-VALUES

"How large of a sample size do I need to answer *X*?"

—everyone, all the time

TYPE 2 ERROR & POWER

- P-values tell us about the probability of a Type 1 error
- We can also commit Type 2 errors: "false negatives"
  - $P(\text{don't reject } H_0 \mid H_0 \text{ is false})$
- **Power:** probability of *correctly* rejecting the null hypothesis
  - $P(\text{reject } H_0 \mid H_0 \text{ is false})$
  - 1 - probability of type 2 error

TYPE 2 ERROR & POWER

- We can decrease the probability of a Type 2 errors by increasing our Type 1 error
  - **Exercise**: Why?
- We can decrease the probability of a Type 2 error *without sacrificing Type 1 error* by increasing our sample size

SUMMARY

- Be cautious of small p-values
- Be cautious of large p-values

SUMMARY

- Be ca
- Be ca

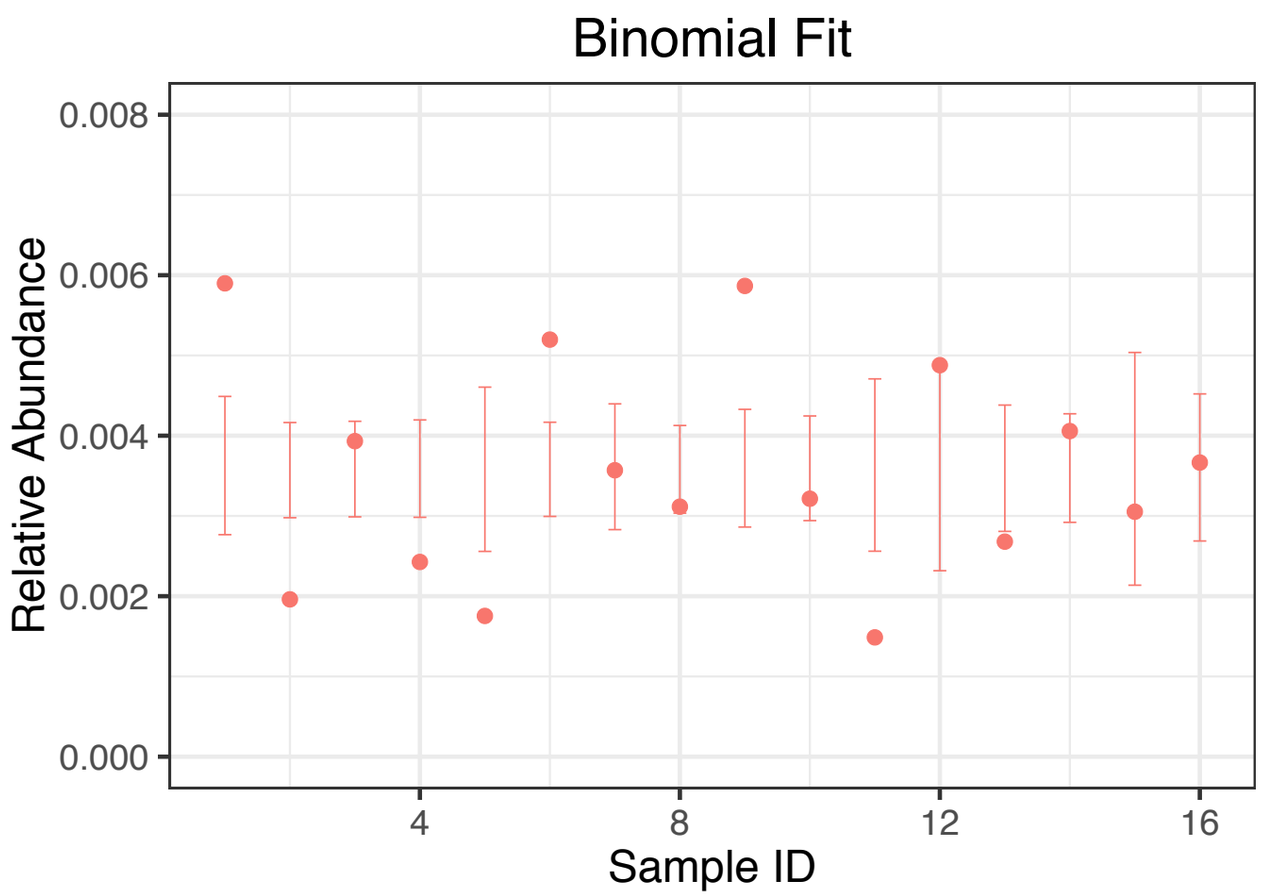
SUMMARY

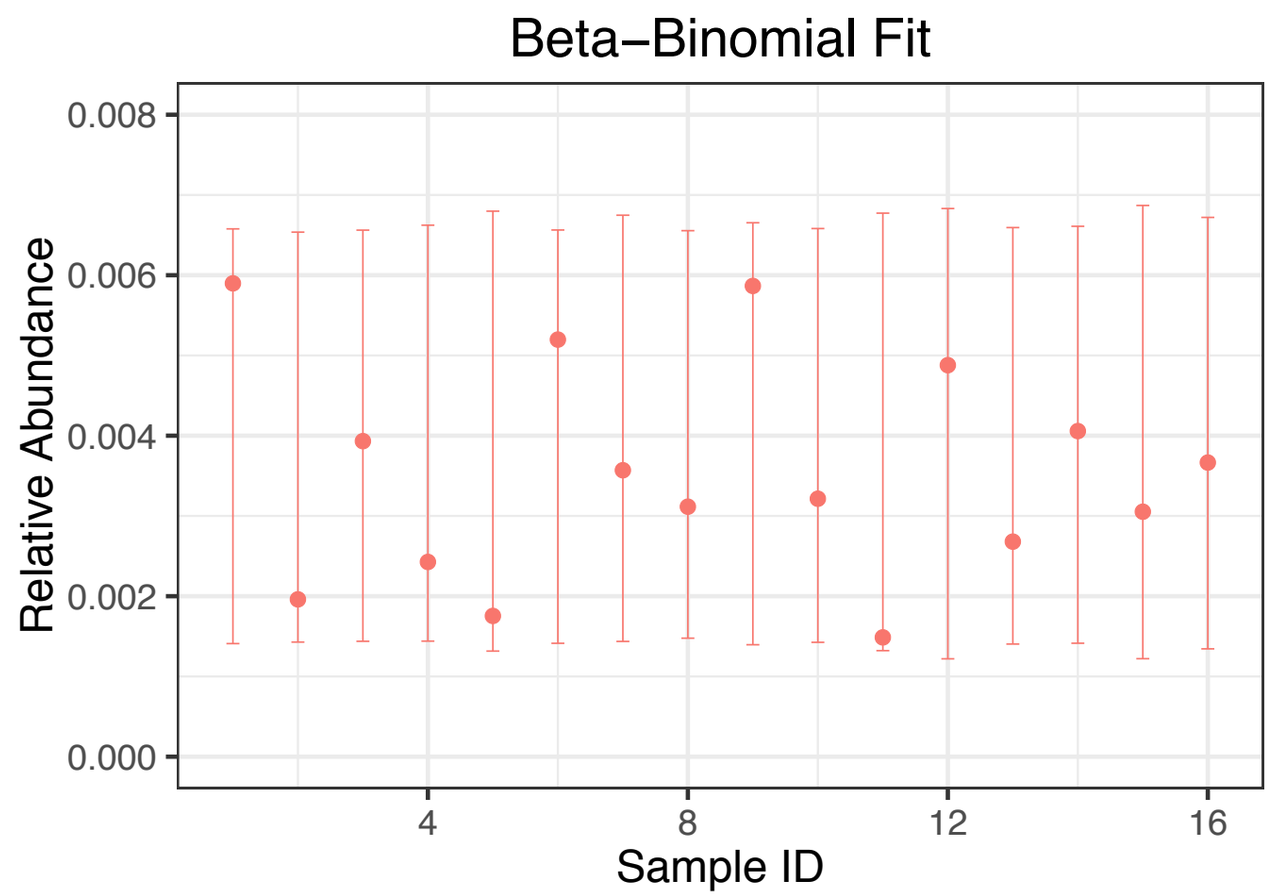
- Be cautious of small p-values
- Be cautious of large p-values
- Be cautious of medium p-values

MODEL MISSPECIFICATION

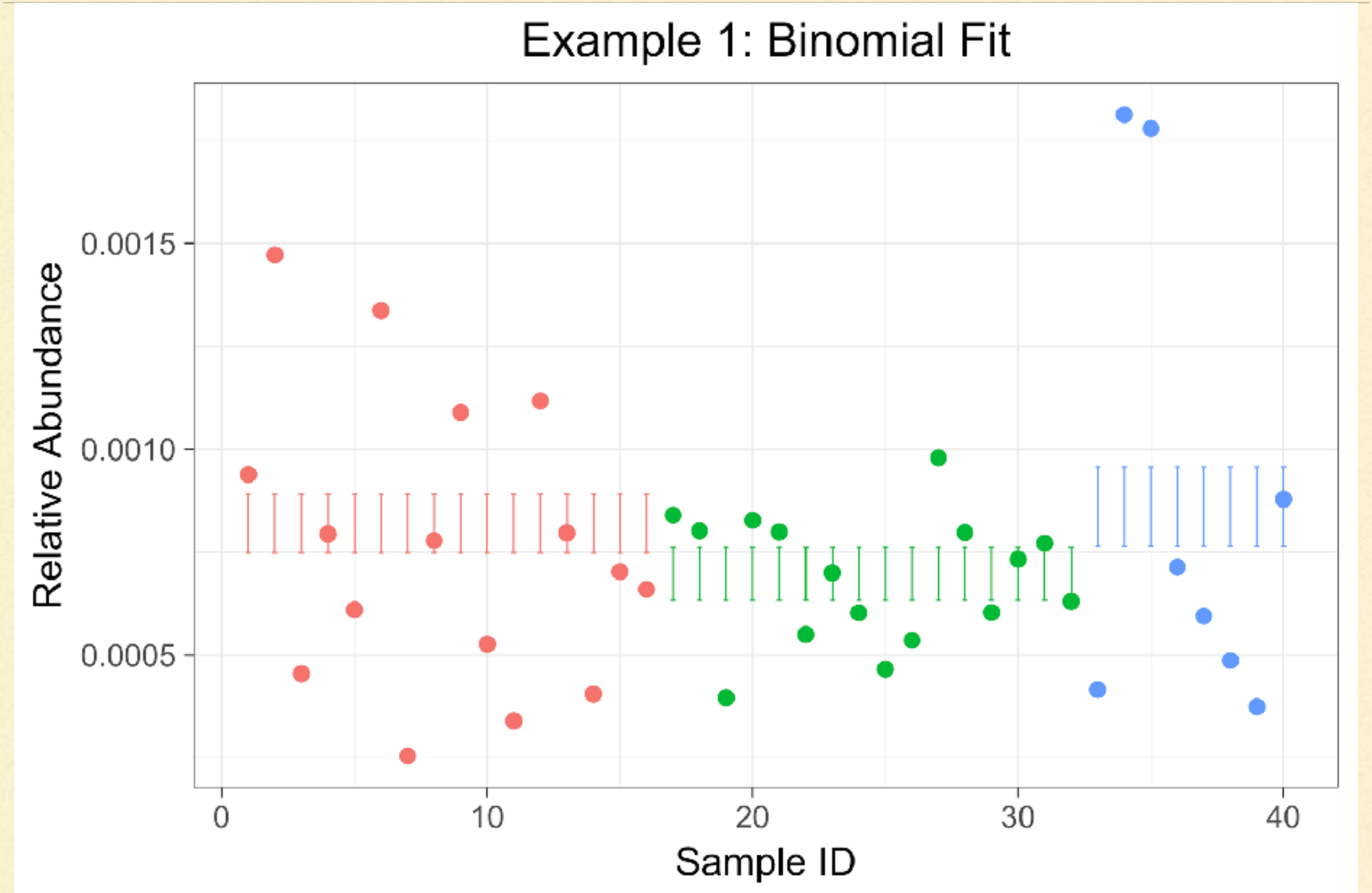
- p-values require estimates of the variance
- estimates of the variance require models
- wrong model → wrong variance → wrong p-value!

MODEL MISSPECIFICATION

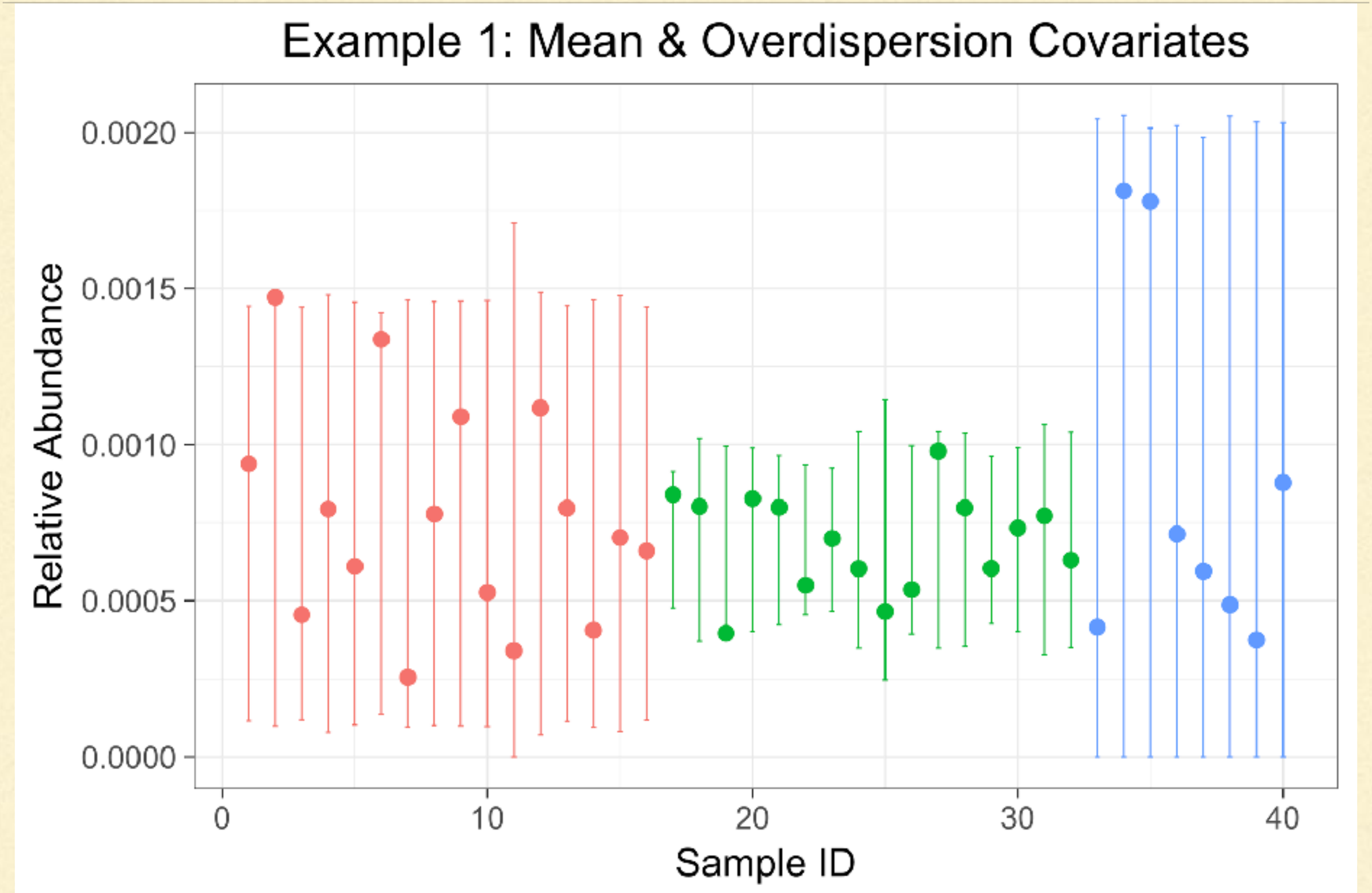




Martin, Witten, & Willis (2019)



p-value: 0.0058 😿😿



p-value: 0.3782 🌽🌽

P-VALUE ACTIVITY

- **Setting:** You are interested in differential abundance in the human gut microbiome in response to a disease
- After designing and executing your well planned study, you find that a statistical test of your most prized hypothesis has a p-value of 0.067.
- However, you notice the trend is stronger in men so you subset your data to only men and conduct a hypothesis test. Eureka! Your p-value is 0.031!
- **Exercise:** Discuss with your neighbors some problems with this approach

AMY'S RECAP ON P-VALUES

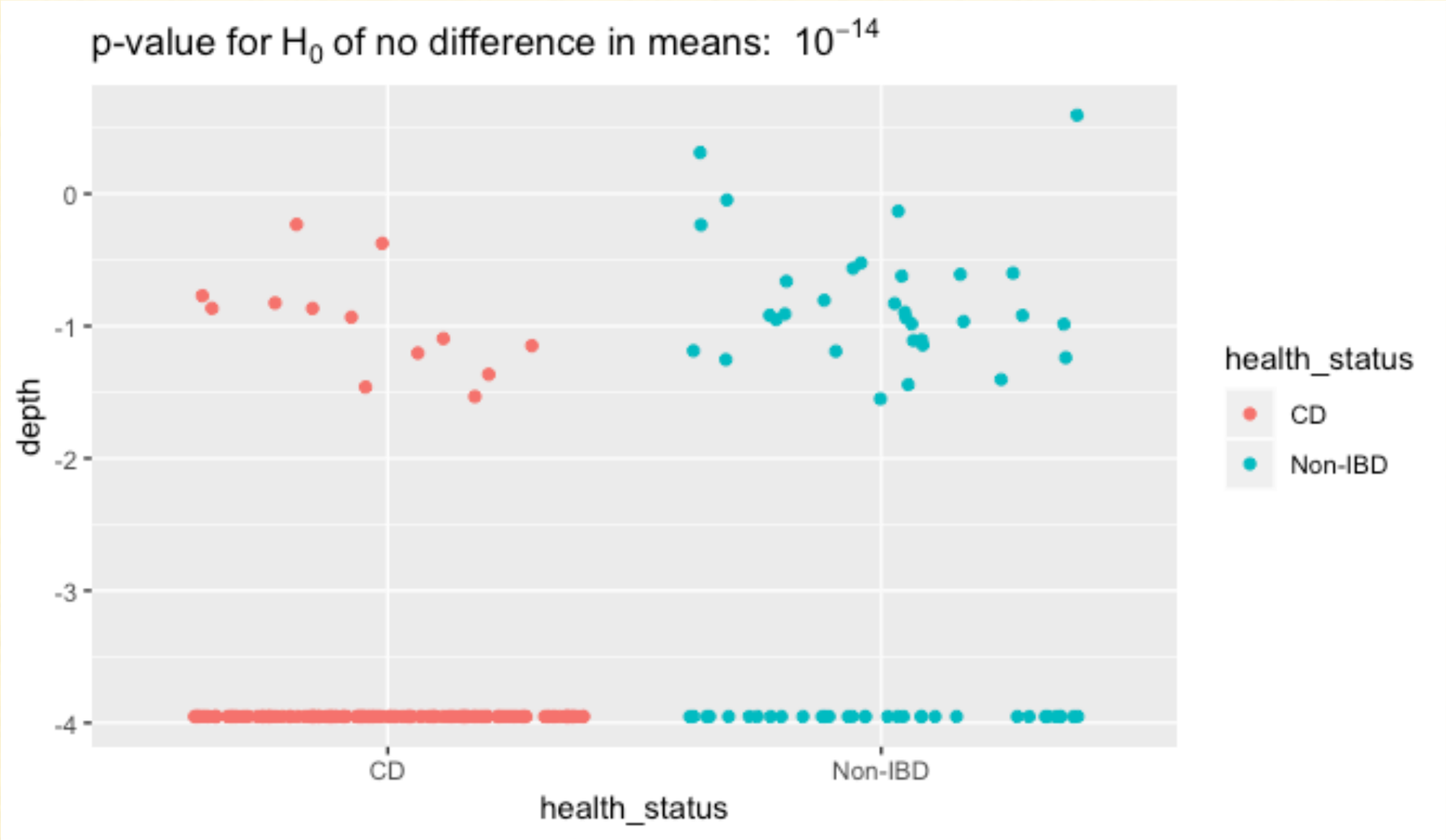
- Current conventions for statistical analysis in science are inherently unscientific
  - Yes, this is unfair
- You have some obligations
  - Know what a hypothesis test is doing: what is the parameter, model, null hypothesis…
  - Do not consider "p < [favourite number]" a benchmark for publication!

EVERY P-VALUE NEEDS A PLOT

- How exciting is a p-value of 10^{-14}?

EVERY P-VALUE NEEDS A PLOT

- How exciting is a p-value of 10^{-14}?



HOW DO I ASSESS WHERE VARIATION IS COMING FROM?

EXPERIMENTAL DESIGN & VARIABILITY

- Usually we want to understand biological variation
  - e.g. How does the microbiome change with [treatment]?
- However, some noise in our data may be the result of technical variation
  - Want less technical variation: signal vs noise

SOURCES OF VARIATION

- **technical variation:** variability due to the measuring equipment and protocols
  - e.g. not all reads are sequenced; this is a random process
- **biological variation:** variability across experimental units
  - e.g. samples from different field sites

TECHNICAL REPLICATES

- Experimental design determines the sources of variability you can assess
- Technical replicates help you assess technical variation (**important!**), but are useless for assessing biological variation
  - Samples from different patients help you understand patient-to-patient variability
  - Samples from different sites help you understand within-patient variability
  - Repeating the protocol on the same sample helps you understand variability due to protocol

VARIATION ACTIVITY

We have a protocol for extracting DNA and performing PCR. We want to assess the variability in taxonomic profiling of the human gut microbiome introduced at each step of our protocol.

Discuss: How might we go about assessing the variability introduced by each step of our protocol (DNA extraction & PCR)?

WHAT CAN WE DO?

- Think critically
  - What do you care about?
    - Do you care about biological variation? Do you know that your protocol has low technical variation?

WHAT ELSE CAN WE DO?

- Be scientific
- Be self-critical! Second guess your findings!
- "What else could have driven this result?"
- Confirming your findings using multiple approaches
  - e.g. multiple HTS technologies, qPCR, *in silico* modelling…

WHAT ELSE CAN WE DO?

- Understand your methods
- 🐤 test
- Use plots, not (just) p-values
- Ask the developers!

WHAT ELSE CAN WE DO?

- Be honest
  - Keep all analyses that you ran, not just the final one
- Write down all of the hypotheses that you care about
  - Before doing the experiment, before doing the analysis
- Your university might house a statistician; try to involve them...
  - ...in the entire process, not just calculating p-values

WHO, AGAIN?

statistical diversity lab

- We are the statistical diversity lab, and we develop…

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Therapeutics and Prevention

Rigorous Statistical Methods for Rigorous Microbiome Science

Amy D. Willis[a]

[a]Department of Biostatistics, University of Washington, Seattle, Washington, USA

ABSTRACT High-throughput sequencing has facilitated discovery in microbiome science, but distinguishing true discoveries from spurious signals can be challenging. The Statistical Diversity Lab develops rigorous statistical methods and statistical software for the analysis of microbiome and biodiversity data. Developing statistical methods that produce valid *P* values requires thoughtful modeling and careful validation, but careful statistical analysis reduces the risk of false discoveries and increases scientific understanding.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

WHO, AGAIN?

statistical diversity lab

We work on what we believe to be the most critical methodological needs in microbial science and the most series shortcomings of existing analysis methods. Along with our research, we see outreach, education, and collaboration as a core part of this mission.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

MANY THANKS

- Pauline Trinh and Bryan Martin
  - For being the best TAs evaaa
- Tracy & Mihai
  - For making this amazing workshop #STAMPS2019 happen
- YOU!
  - For engaging in reproducible and ethical science

RECAP

RECAP

STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington
@AmyDWillis — Assistant Professor
@BryanDMartin_ — PhD Candidate
@paulinetrinh — PhD Candidate

CASE STUDY 🤠🐄💩🐮💉

Who: A prospective cohort of 30 existing dairy workers, 30 new dairy workers, and 30 community controls

What: Fecal, nasal, blood samples collected at baseline enrollment, 3 month, 6 month, 12 month, and 24 month follow-ups. Survey data on demographics, antibiotic use, & food frequency questionnaire.

When: September 2017- December 2020.

Where: A conventional dairy farm in Yakima Valley, WA

🤠🐄💩🐮💉 CONT…

Hypothesis: Working on a dairy farm places people at a higher risk of acquiring antibiotic resistance genes.

Problem: You are interested in producing a pilot investigation into this hypothesis. However, you have a limited budget where you can only sequence a limited set of samples on one flow cell which produces about 300 million reads. You can just enough money for library prep of up to 30 samples.

Discuss: Which samples do you choose? Factors to consider include sequencing depth, relevant covariates, replicates…