Statistical analysis of microbiome data

A primer

Amy D Willis PhD
Principal Investigator, Statistical Diversity Lab
Associate Professor, Department of Biostatistics
University of Washington

Context: Data and statistics

- My usual delineation
  - Bioinformatics turns “raw” sequence data into quantitative data
  - Quantitative data =
    - Some sort of *units*
    - Sometimes, some sort of *counts* of the units
  - Statistics usually happens on *quantitative* data

Why do we collect data?

Discuss in small groups!

(4 minutes)

Why do we collect data?

- [summary of what you said]

Three approaches to analyzing data

1. Inferential statistics

    - My data reflects a greater mechanism. What can I say about the mechanism?

2. Predictive modeling

    - What will happen next time?

3. Exploratory analysis

    - How can I explore patterns/surprises in my data?

Inferential statistics is concerned with *parameters*

- In the inferential paradigm
  - Data is generated from some *complex process*
  - We are interested in *summaries* of this complex process
  - These summaries are usually numbers. They are unknown. They are called parameters.
  - Parameters are *estimated* from the data
  - To estimate them, we make assumptions about the complex process
  - A hypothesis about the parameter's value can be tested

Exploratory statistics is concerned only with *data*

- Alternative approach
  - "My data reflects no greater mechanism"
  - "I'll just analyze the data”
- Normalize, rarefy, transform, compute distances, plot...
- Exploratory approach is *incompatible* with hypothesis testing

Inferential vocab

- In the inferential paradigm…
  - Data is generated
  - <u>Parameters</u> summarize the data generation process
  - We make <u>assumptions</u> to <u>estimate</u> the parameters from data

Case Study: Microbial abundance parameters

- Starting point: ”There is some number of a given biological quantity in any environment"
  - Biological quantity = some biological or genetic unit
  - Context-dependent
    - genomes, gene copies, sequence variants, k-mers, gene transcripts...

Case Study: Microbial abundance parameters

- Starting point: ”There is some number of a given biological quantity in any environment"

  - "There are 54,601 *S epidermidis* cells on my index finger"

  - "There are 874,455,469 copies of the k-mer ATGCCTAGGGA circulating in my blood"

  - "There are 0 transcripts of the gene *Core RC1 subunit PsaA* on my desk"

Case Study: Microbial abundance parameters

- Y_{ij} = true number of unit j in sample i

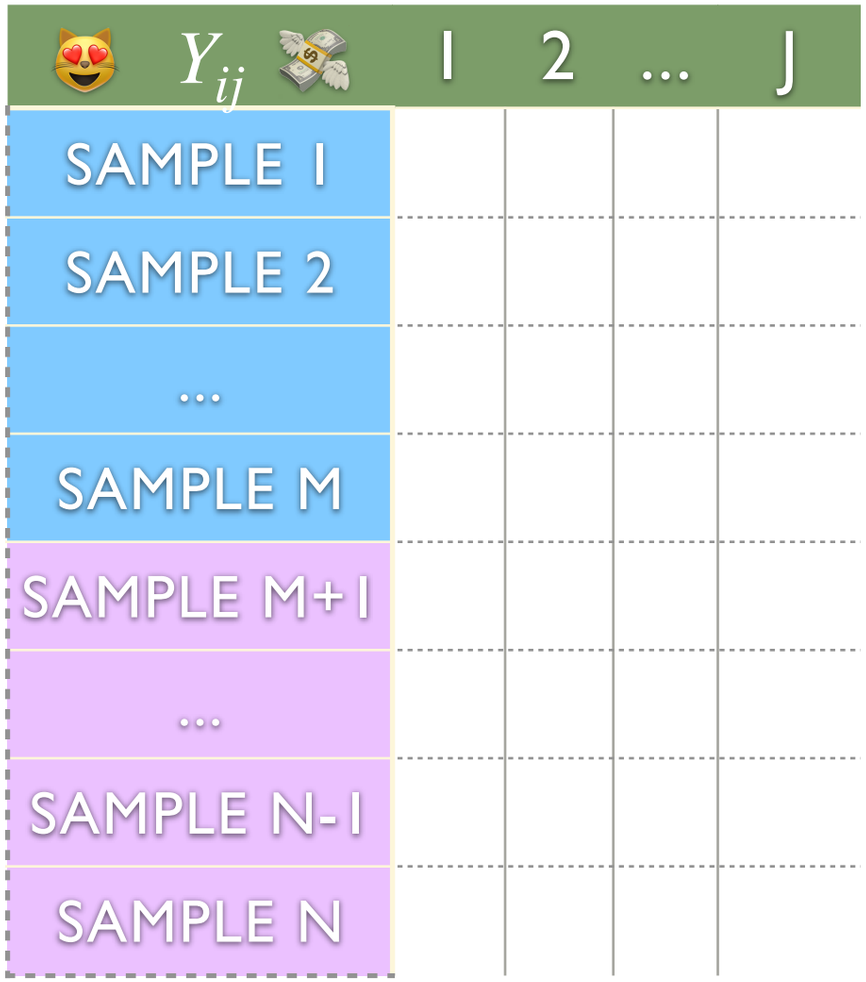
- X_i = environment types (e.g., treatment vs control, low- vs high-rainfall…)

😍 Y_{ij} 💸	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

If you *knew* the Y_{ij}'s, what would you do with them?

Case Study: Microbial abundance parameters

- Average of Y_{i4} across environments
- % of environments in which $Y_{i2} > 0$
- $\#\{j : Y_{ij} > 0\}$
- $-\sum_{j=1}^{J} p_{ij} \log p_{ij}$ for $p_{ij} := \frac{Y_{ij}}{\sum_j Y_{ij}}$
- ...



There are *many* parameters that you could care about

Number of distinct species present

Evolutionary rates

Mean total abundance

Differences in relative abundance

Closest Relatives

Rates of presence

Many others…

You decide!

Why consider parameters?

- Once you know what parameter you care about, you can evaluate what assumptions are reasonable to estimate it

- Assumptions connect your data with parameters

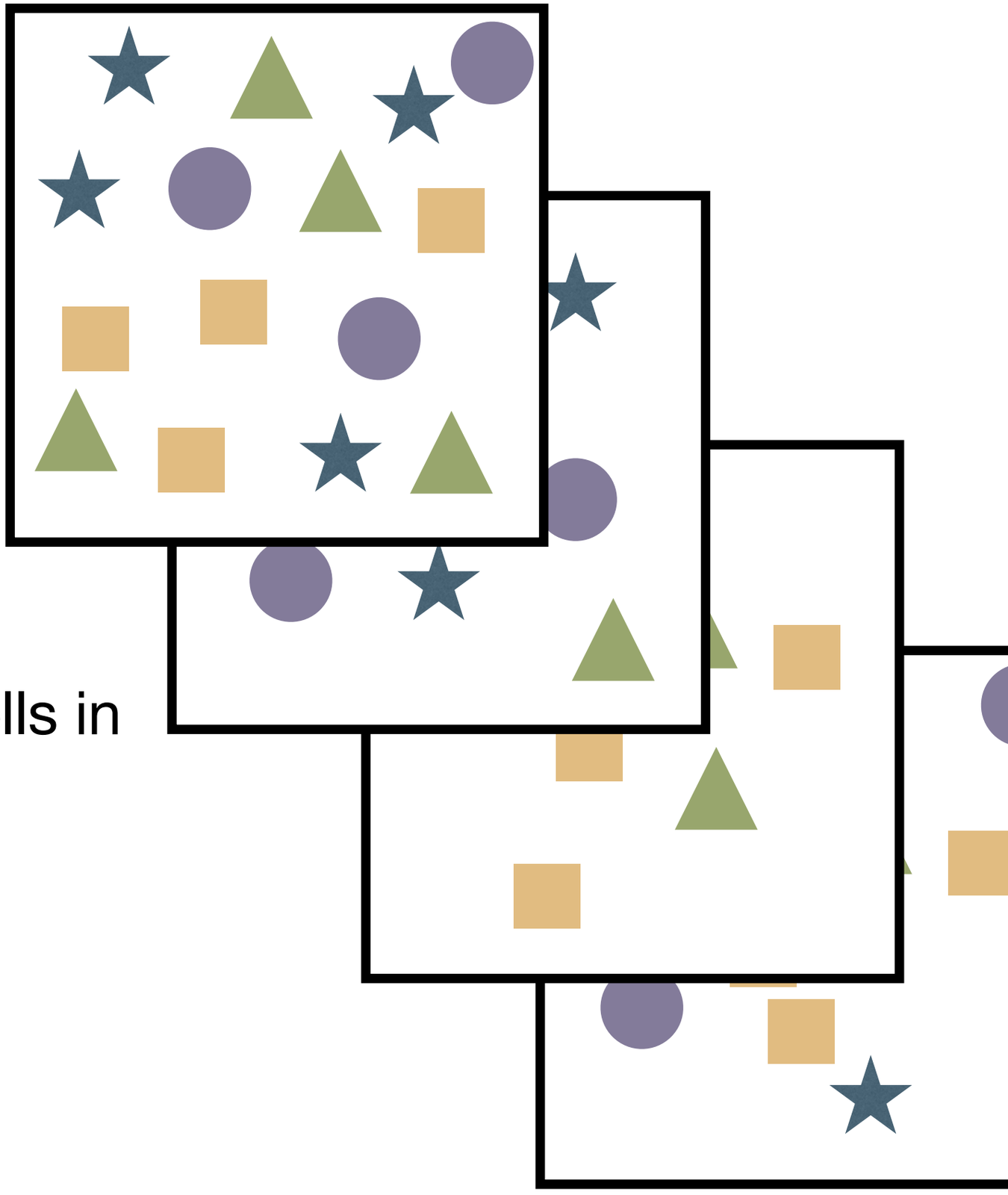
Case Study: Microbial abundance models

- Y_{ij} = true number of unit j in sample i
  - We don't observe the Y_{ij}'s
- W_{ij} = number of times unit j observed in sample i from HTS

🌧️ W_{ij} 😿	1	2	...	J
SAMPLE 1				
SAMPLE 2				
...				
SAMPLE M				
SAMPLE M+1				
...				
SAMPLE N-1				
SAMPLE N				

How do we connect the W_{ij}'s to the $Y_{ij}'s$?

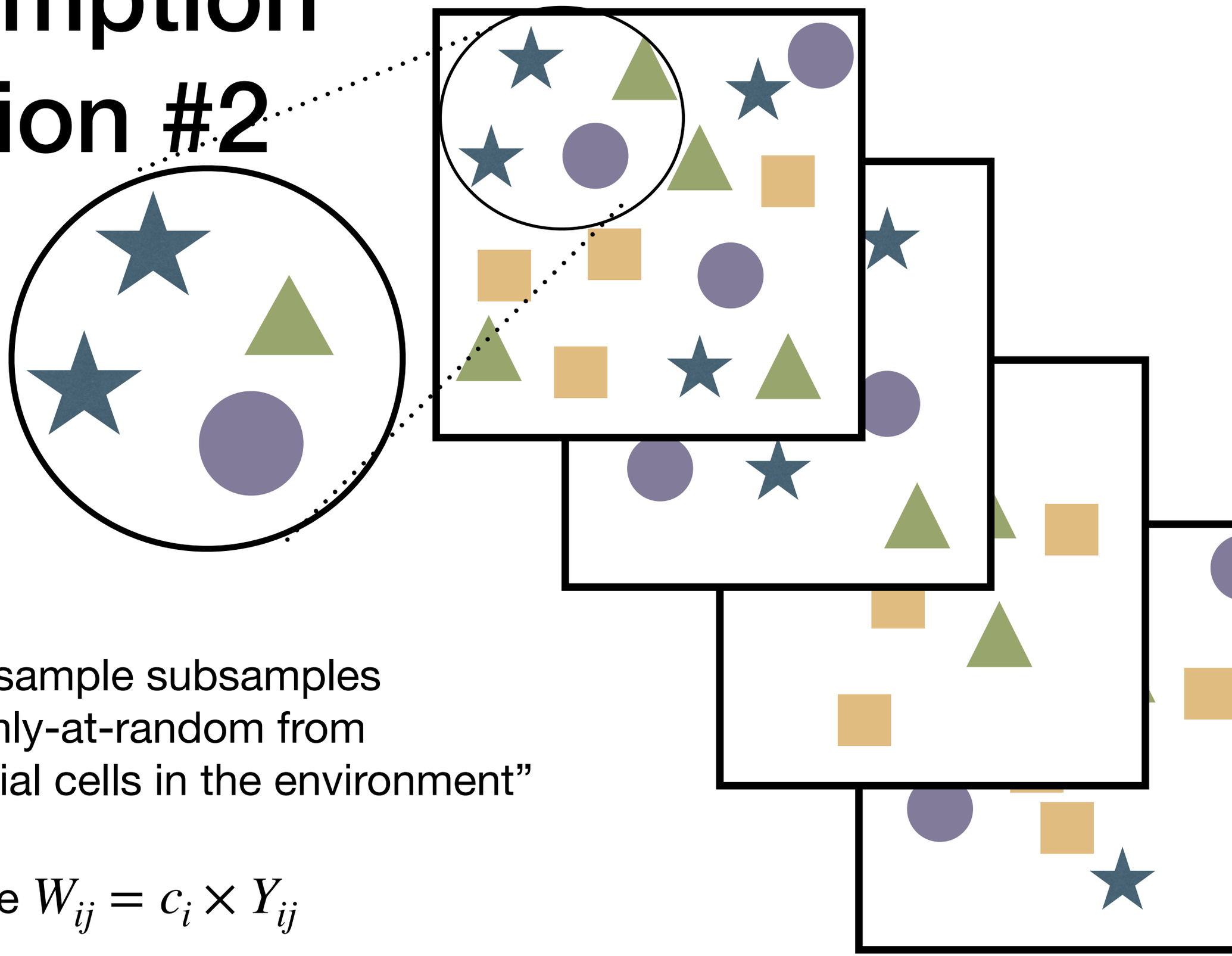
Assumption Option #1



- "Each sample accurately counts all the microbial cells in the environment"

- $W_{ij} = Y_{ij}$

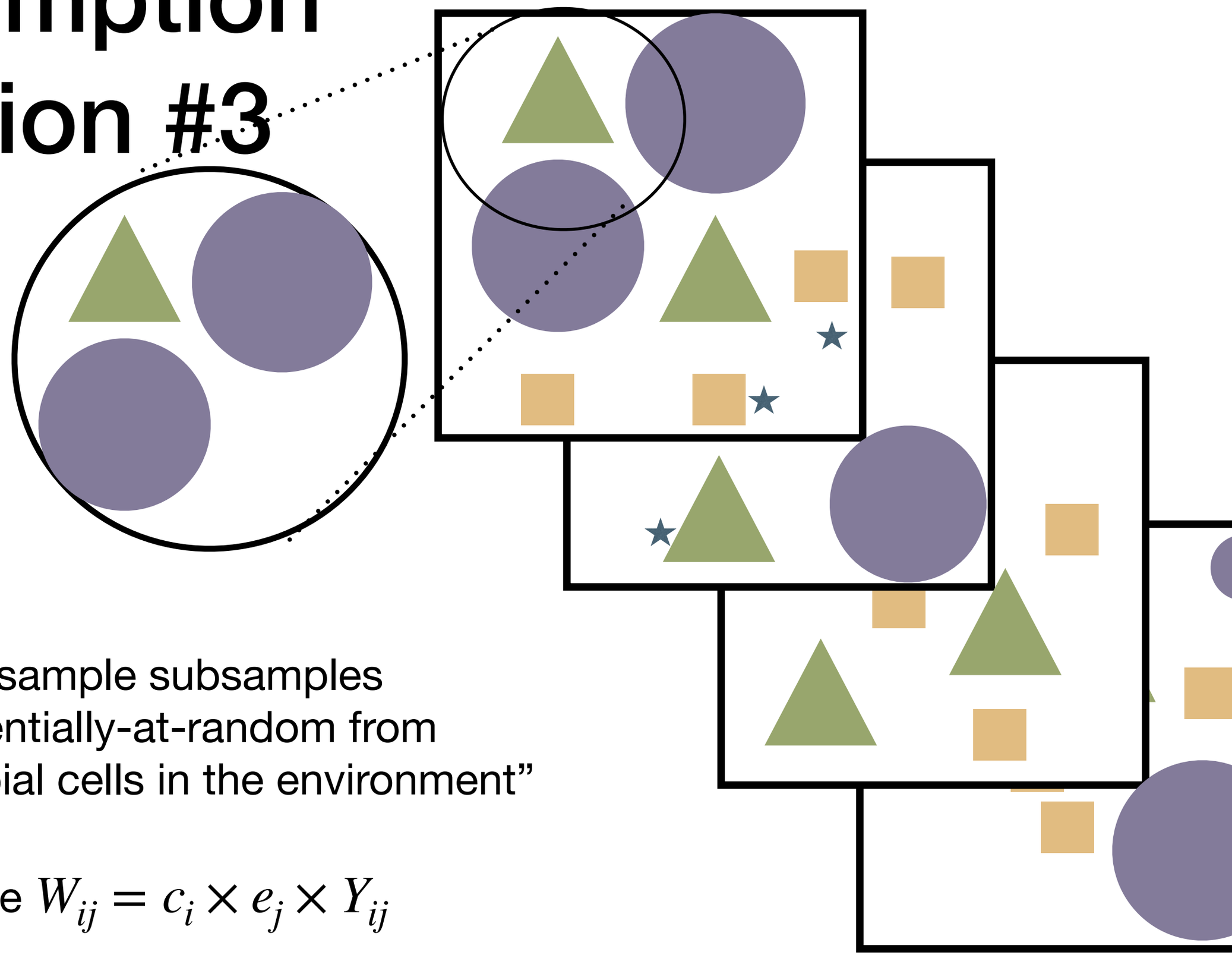
Assumption Option #2



- “Each sample subsamples uniformly-at-random from microbial cells in the environment”

- average $W_{ij} = c_i \times Y_{ij}$

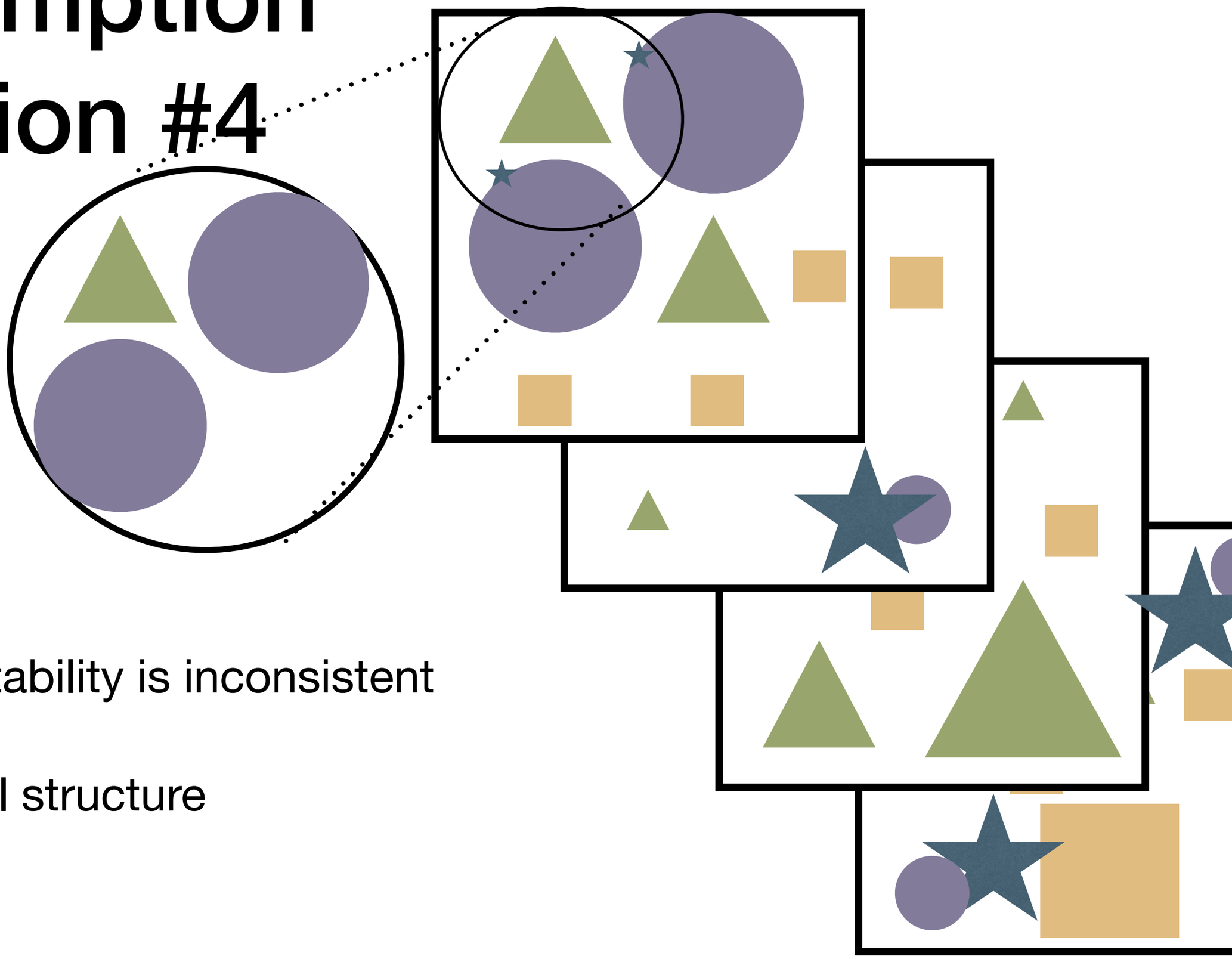
Assumption Option #3



- “Each sample subsamples preferentially-at-random from microbial cells in the environment”

- average $W_{ij} = c_i \times e_j \times Y_{ij}$

Assumption Option #4



- Detectability is inconsistent
- Spatial structure
- ???

Assumptions, algebraically

- Assumption Option 1: $W_{ij} = Y_{ij}$
- Assumption Option 2: average $W_{ij} = c_i \times Y_{ij}$
- Assumption Option 3: average $W_{ij} = c_i \times e_j \times Y_{ij}$
- Assumption Option 4: something about averages, something about co-occurance, something about inconsistent detectabilities…
- …

Models

is just a fancy word for

assumptions

Can data be perfect?
Can data be useless?
Can models / assumptions be perfect?
Can models / assumptions be useless?

Discuss in small groups!

(5-ish minutes)

Models

- A good model is one that

    1. You understand

    2. Captures the most important features of both the universe and data

    3. Answers a question that you have about biology

- More complex models are not always better

- There are not “universally” best models

Estimation

- The bridge between parameters and estimates is filled by statistical tools
- In practice: computational methods to generate a number
- Key: that number can mean *something*

Example: implausible assumptions

- Parameter: Average number of live *E coli* cells in the gut microbiome of [this] group of people

- Assumption: “Each sample accurately counts all the microbial cells in the environment”

- Estimator: Average the observed number of 16S counts attributed to *E coli*

- Estimate: 20,471.8

Example: plausible assumptions

- Parameter:
  - "Average number of *E coli* cells in the gut mb of with diarrhea" divided by "average number of *E coli* cells … without diarrhea"
- Assumptions:
  - "Each sample subsamples preferentially-at-random from microbial cells in the environment"
  - "Some samples are sequenced more deeply than others"
  - "The average ratio (across bacteria) is 1"
- Estimator: [radEmu output]
- Estimate: 1.42 (95% confidence interval: 0.86, 1.57)

Example: implausible assumptions

- Parameter: Average number of live *E coli* cells in the gut microbiome of this group of people
- Assumption: “Each sample accurately counts all the microbial cells in the environment”
- Estimator: Average the number of 16S counts attributed to E coli
- Estimate: 20,471

Vocab: Estimators

- *Parameters* are unknown summaries of the universe
- We *estimate* parameters using our data
- We call these functions of our data *estimators*

Which paradigm?

- Exploratory vs predictive vs inferential
- It’s up to you!
  - Summarise data
  - Learn about biology/the universe

Which parameter?

- It’s up to you!
- Choose based on your *questions*

Statisticians suggest parameters and assumptions

- Estimating and modeling species richness 💰breakaway💰 & 🐠betta🐠
- Estimating and modeling Shannon diversity 🕸️DivNet🕸️
- Estimating and modeling relative abundances 🌽corncob🌽
- **Estimating and modeling presence/absence 🔵happi🟣**
- Estimating detection efficiencies of HTS relative to qPCR data ⛑️paramedic⛑️
- Decontaminating relative abundance & estimating differential detection w/ mock communities 🧛 tinyvamp🧛
- **General purpose regression models with robust hypothesis testing 📈rigr📈**
- **Investigating gene-phylogenies alongside your phylogenomic tree 🌴groves🎄**
- **Estimating fold-changes in absolute abundances from HTS data 🦤radEmu🦤 and 🎈fastEmu🎈**

github.com/statdivlab

Statisticians suggest parameters and assumptions

- Estimating and modeling species richness 💰breakaway💰 & 🐠betta🐠
- Estima
- Estima
- **Estima**
- Estima
- Decon
  🧛 tiny
- **Gener**
- **Investigating gene-phylogenies alongside your phylogenomic tree 🌴groves🎄**
- **Estimating fold-changes in absolute abundances from HTS data 🦆radEmu🦆 and 🎈fastEmu🎈**

We are going to go into more detail about <u>specific</u> parameters, models, & estimators next Monday on…

⭐ stats day 😻

github.com/statdivlab

Summary

of my personal opinions

- There’s no such thing as perfect data
- There’s no such thing as perfect models in biology

Summary

of my personal opinions

- Data doesn’t need to be perfect to be useful
- Models connect data to a scientific mechanism
- Great models connect data to a scientific mechanism you care about
  - “Do stuff” vs “Answer questions”

A simple model that you understand is far better than a complex model that you don’t

– *me*

Statistical analysis of microbiome data

Questions, please, I beg you…

Amy D Willis PhD
Principal Investigator, Statistical Diversity Lab
Associate Professor, Department of Biostatistics
University of Washington

An exercise

that is way too hard

and will be a disaster.

Let’s not do this.

Hopefully it’s dinner time.