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)

Two paradigms for data collection

3 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 a model
  - Models depend on unknown parameters
  - The parameters are estimated from the data
  - 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 from a model
  - Models depend on unknown parameters
  - The parameters are estimated from the data

Case Study:
Microbial abundance parameters

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

Case Study: Microbial abundance parameters

- Model: "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 $p_{ij} := \frac{Y_{ij}}{\sum_j Y_{ij}}$
- …

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

There are *many* parameters that you could care about

Number of distinct species present,
mean total abundance,
differences in relative abundance,
rates of presence
evolutionary rates,
closest relatives
many others…

You decide!

Why consider parameters?

- Once you know what parameter you care about, you connect it to your data via a model

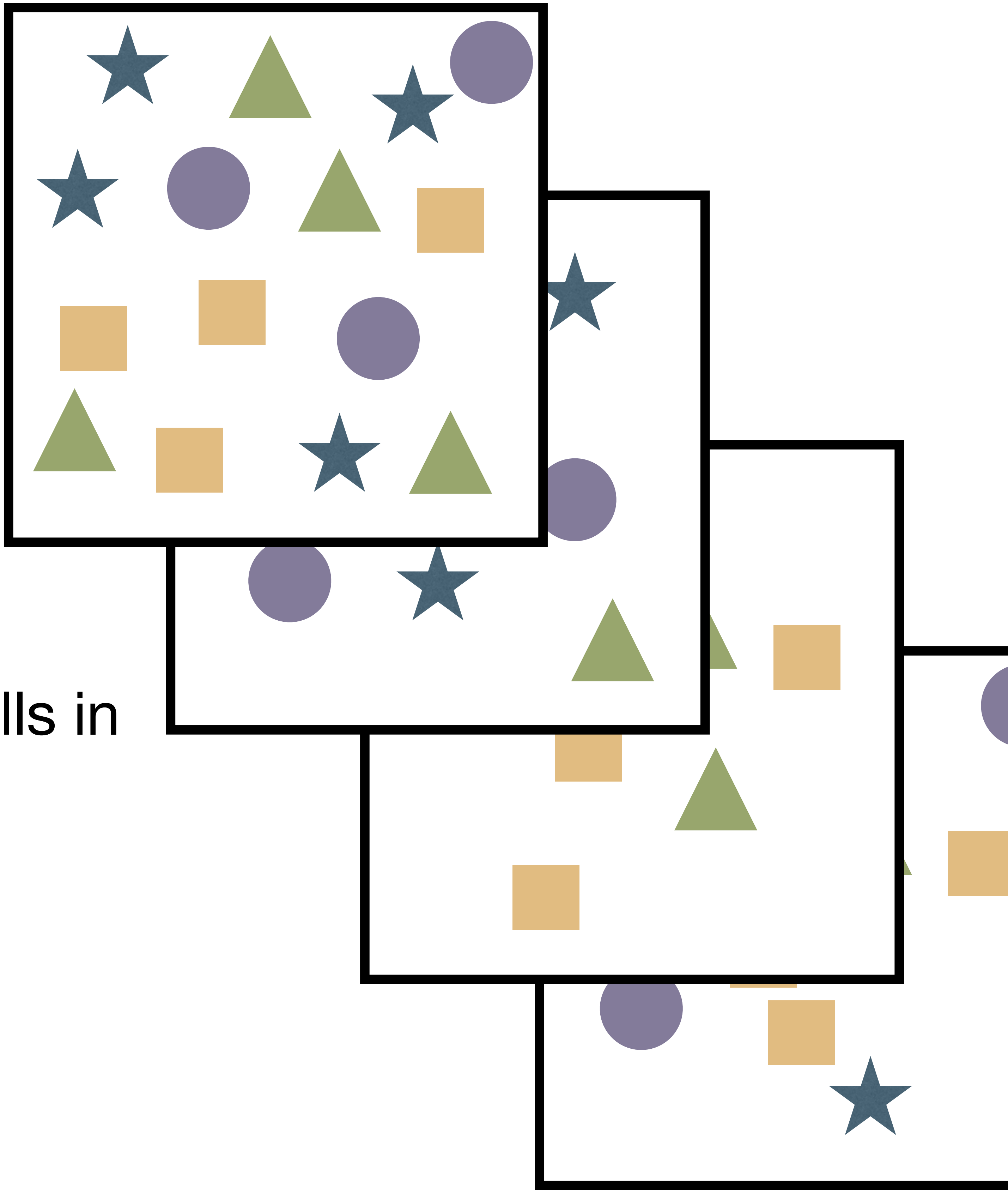
Case Study:
Microbial abundance models

- 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				

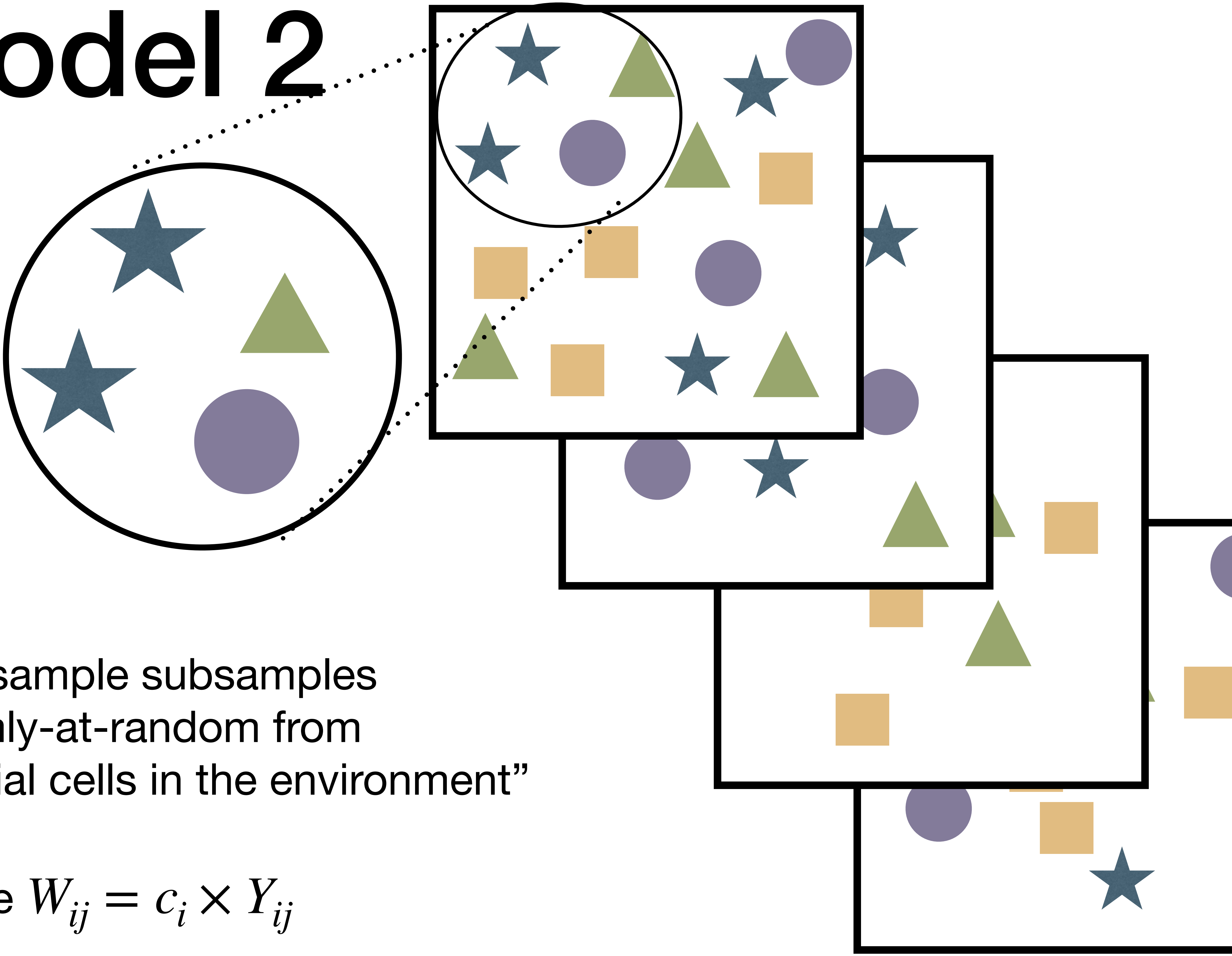
How do we connect the W_{ij}'s to the $Y'_{ij}s$?

Model 1



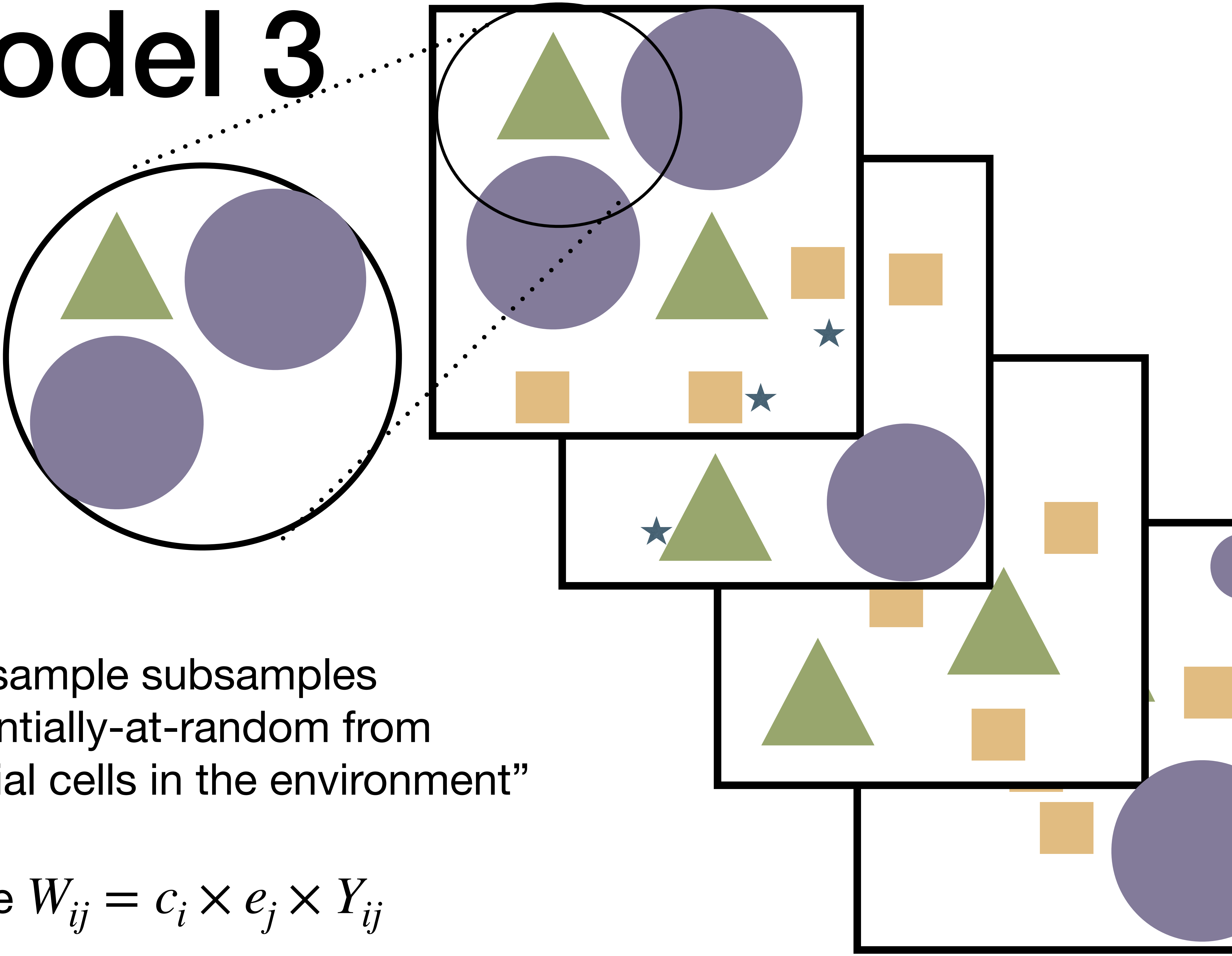
- “Each sample accurately counts all the microbial cells in the environment”
- $W_{ij} = Y_{ij}$

Model 2



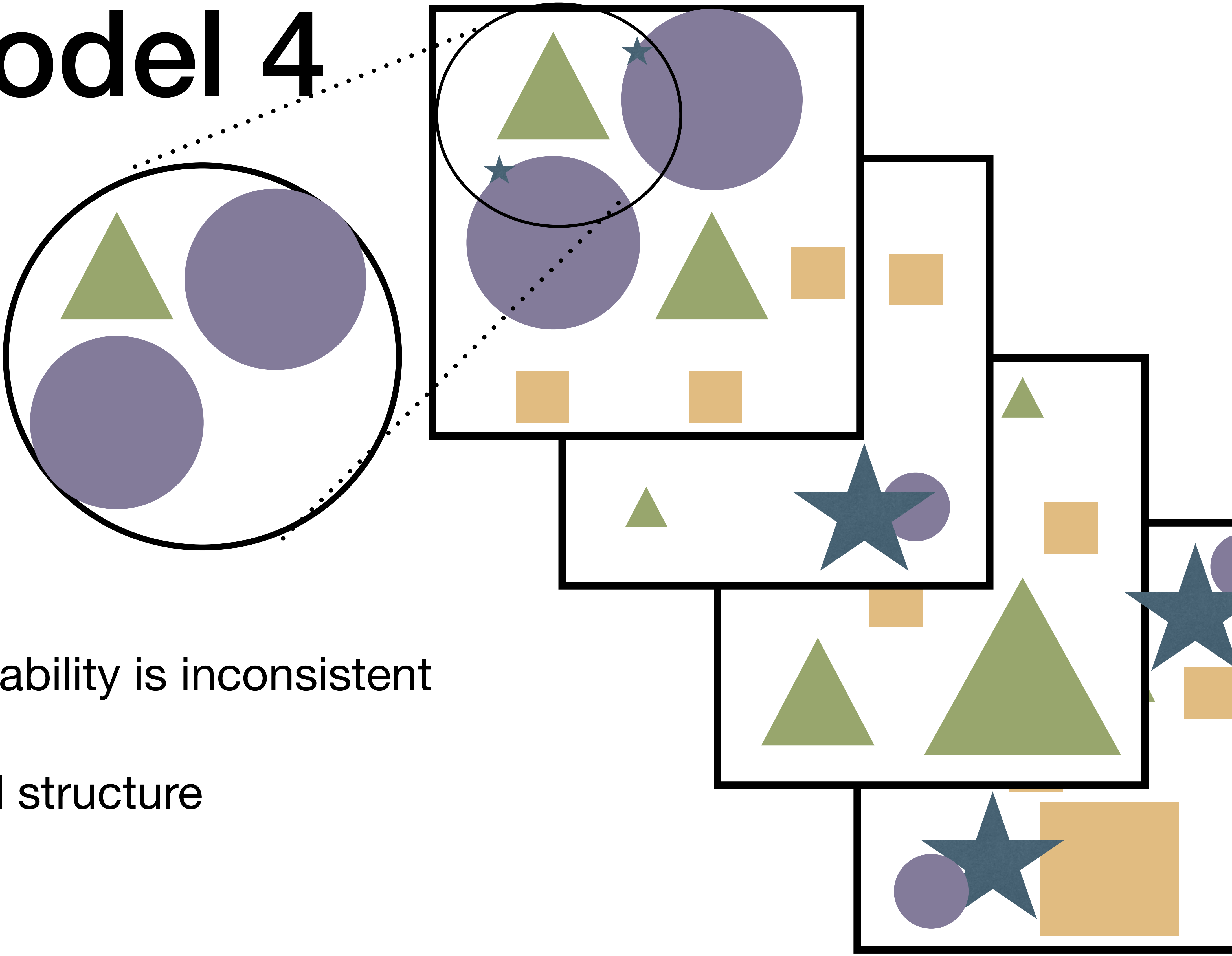
- “Each sample subsamples uniformly-at-random from microbial cells in the environment”
- average $W_{ij} = c_i \times Y_{ij}$

Model 3



- “Each sample subsamples preferentially-at-random from microbial cells in the environment”
- average $W_{ij} = c_i \times e_j \times Y_{ij}$

Model 4



- Detectability is inconsistent
- Spatial structure
- ???

Models, algebraically

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

Can data be perfect?

Discuss in small groups!

(4 minutes)

Can data be useless?

Discuss in small groups!

(3 minutes)

Can models be perfect?

Discuss in small groups!

(4 minutes)

Can models be useless?

Discuss in small groups!

(3 minutes)

Models

- A good model is one that

    1. You understand

    2. Captures the most important features of your model 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

- Once you've decided on your parameter and model, you need to estimate your parameters
    - Hope a statistician has done this for you!
- We (the StatDivLab) are always excited to connect you to what's out there, or to hear about new parameters / the need for better models…

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*

We propose parameters, suggest models, and develop estimators!

- 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🦤**

github.com/statdivlab

We propose parameters, suggest models, and develop estimators!

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

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

*As distinct from a summary of… you know… the facts?

Summary

of my personal opinions

- Data doesn’t need to be perfect to be useful
- Good models connect data to reality
- Great models connect data to something you care about

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

– *me*