FRAMING

WHAT IS SEQUENCE COUNT DATA?

Multivariate count data $\mathbf{Y_{ij}}$ representing the number of transcripts of type ***j*** sequenced in sample ***i***

Examples:

- 16s rRNA sequencing
- RNA-seq (± Single Cell)
- T-cell receptor sequencing
- [hand waving] tumor subtypes

Extended Applications

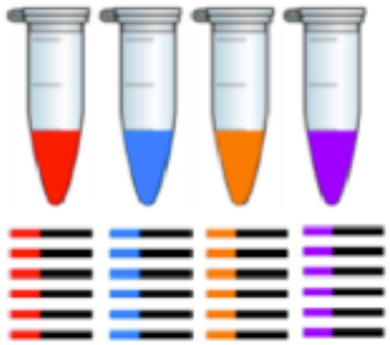
[Beyond Sequencing]:

- Multiparametric Flow Cytometry
- Political Polling

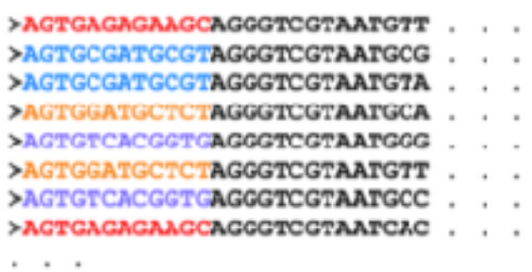
DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection
and Storage

DNA Extraction
PCR Amplification



Sequencing



Assign Sequences
to Samples

Denoise Reads
or Cluster

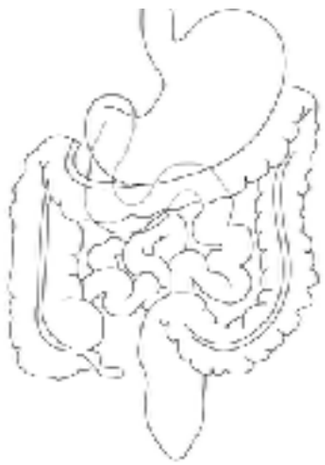
	Species 1	Species 2	Species 3	Spe
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Make Count Table

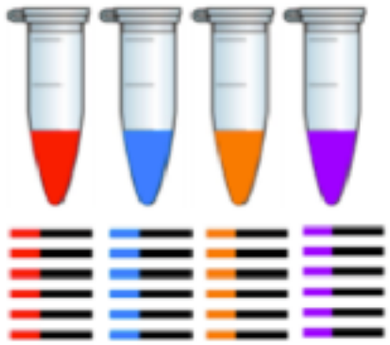
Adapted from Hamady. et al., *Nature Methods*, 2008

DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection
and Storage



DNA Extraction
PCR Amplification

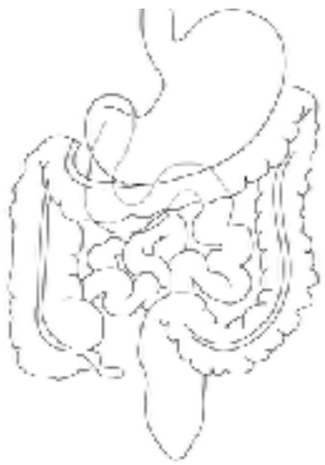


Sequencing

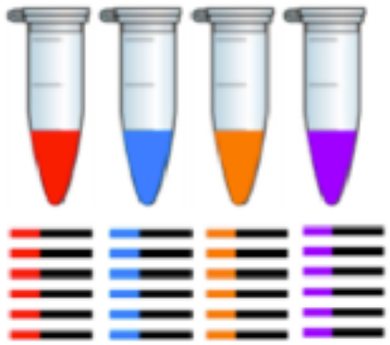
Adapted from Hamady. et al., *Nature Methods*, 2008

DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection and Storage



DNA Extraction PCR Amplification



Sequencing

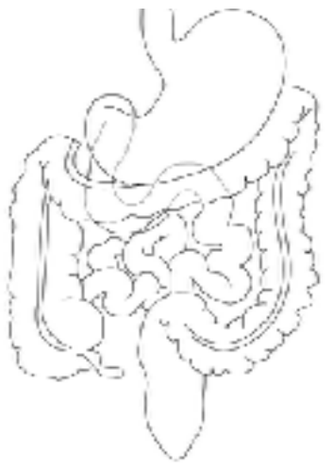
TECHNICAL VARIATION AND BIAS

COUNTING AND BIAS

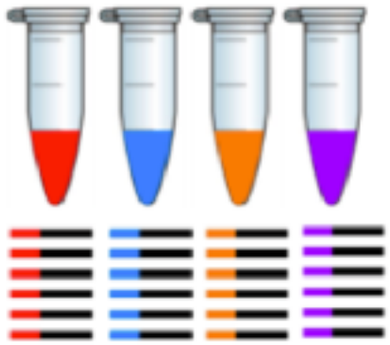
Adapted from Hamady. et al., *Nature Methods*, 2008

DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection and Storage



DNA Extraction PCR Amplification



Sequencing

TECHNICAL VARIATION AND BIAS

COUNTING AND BIAS

RANDOM SUBSAMPLING

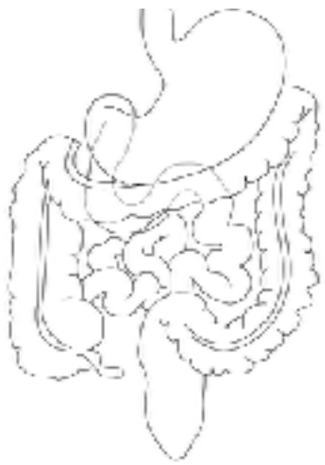
RANDOM SUBSAMPLING

RANDOM SUBSAMPLING

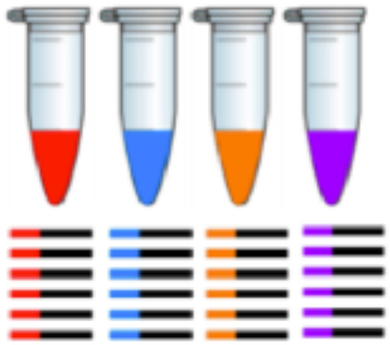
Adapted from Hamady. et al., *Nature Methods*, 2008

DATA COLLECTION AND SAMPLE PROCESSING

Sample Collection and Storage



DNA Extraction PCR Amplification



Sequencing

BIOLOGICAL VARIATION AND SIGNAL

TECHNICAL VARIATION AND BIAS

COUNTING AND BIAS

RANDOM SUBSAMPLING

RANDOM SUBSAMPLING

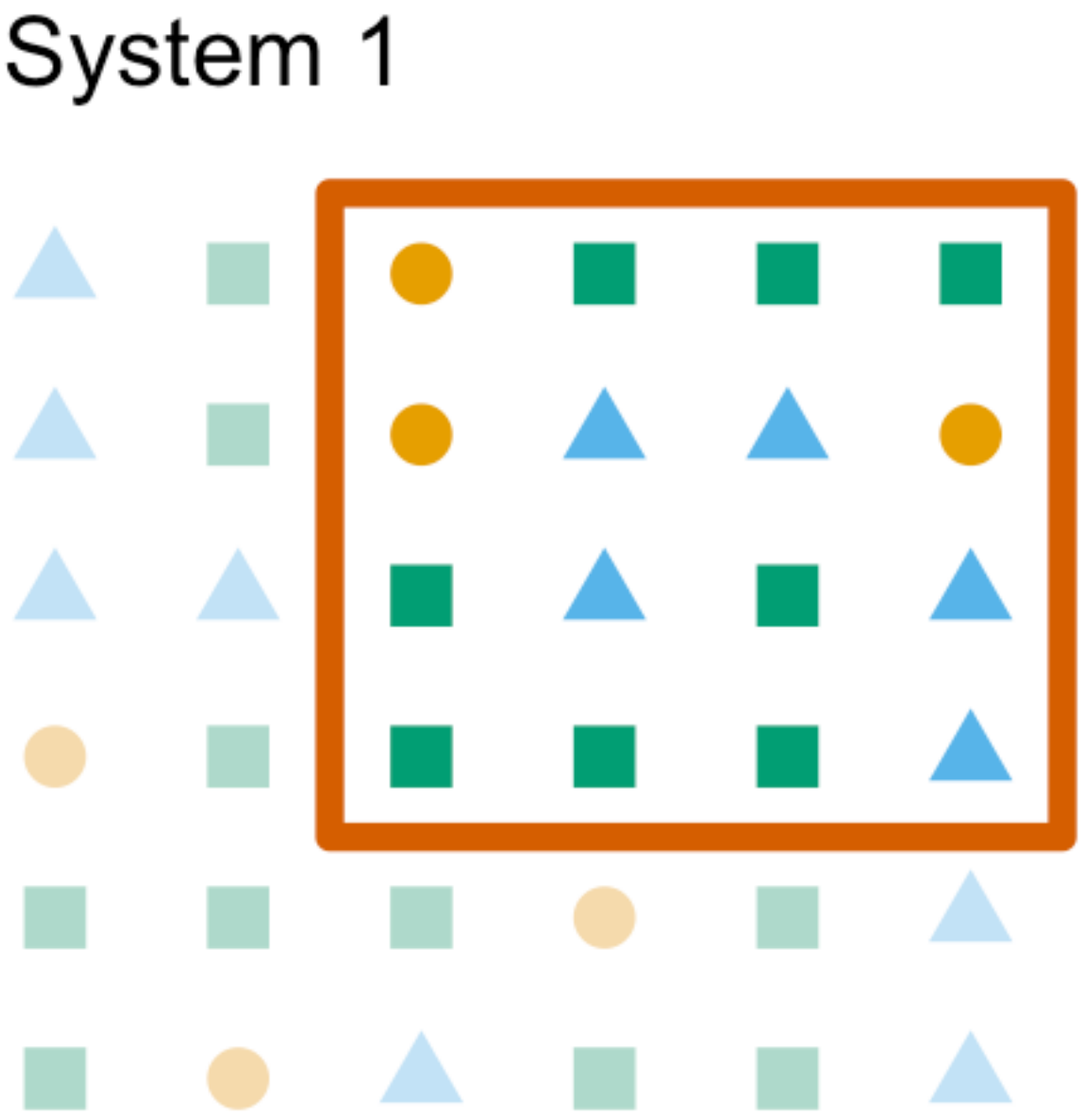
RANDOM SUBSAMPLING

Adapted from Hamady. et al., *Nature Methods*, 2008

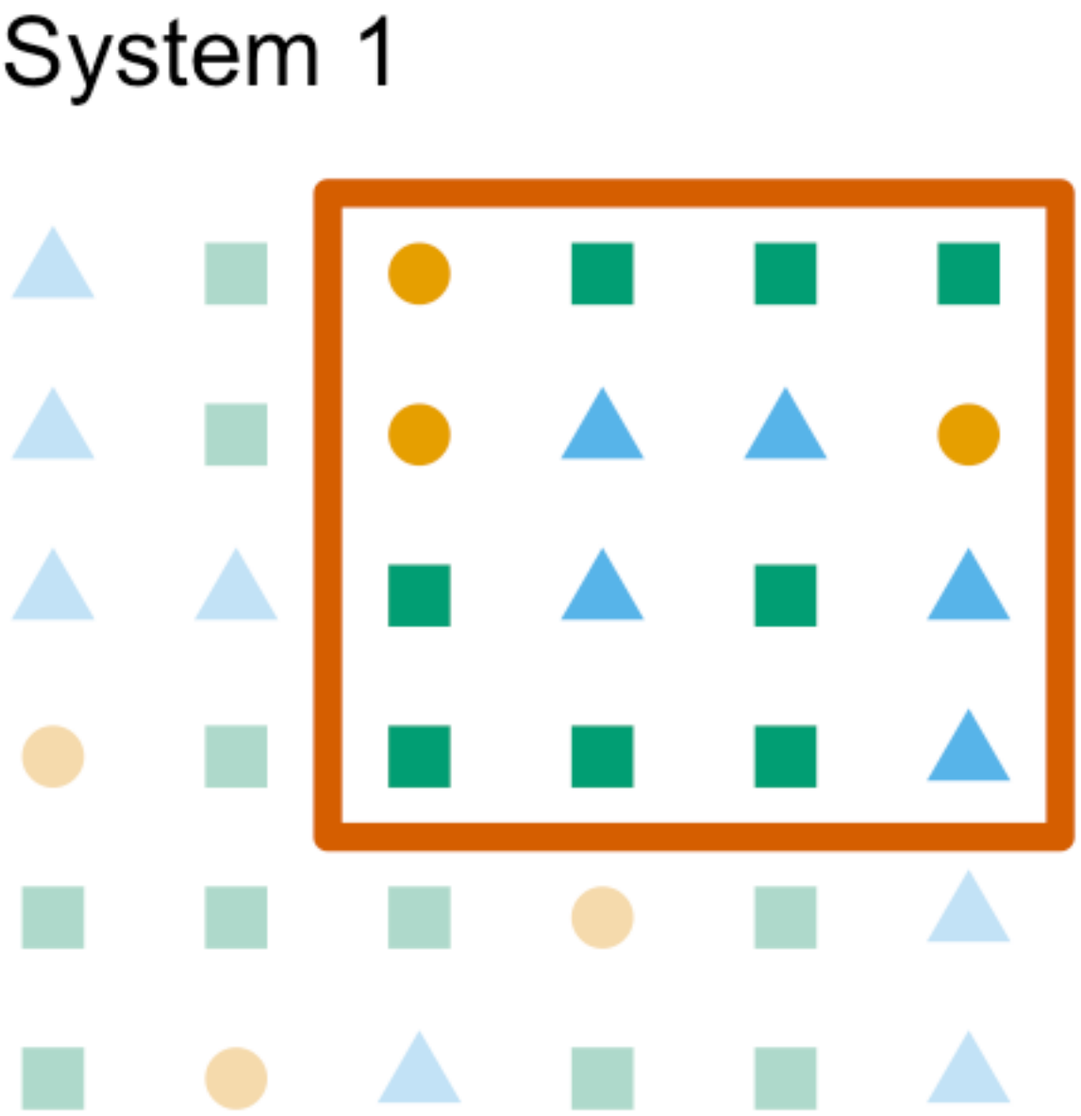
KEY POINT

- Sequencing depth does not seem to correlate with microbial load.

PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING

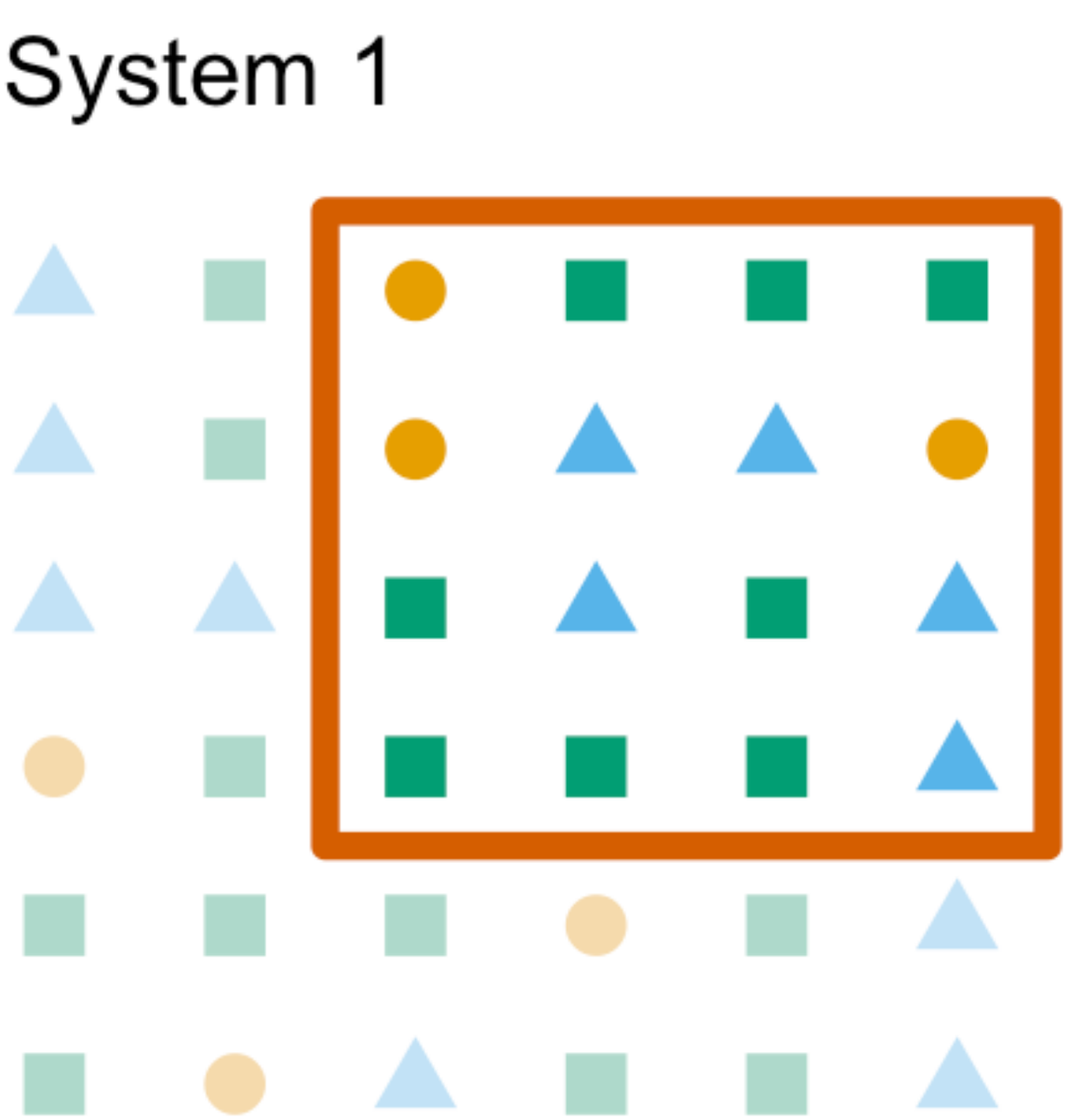


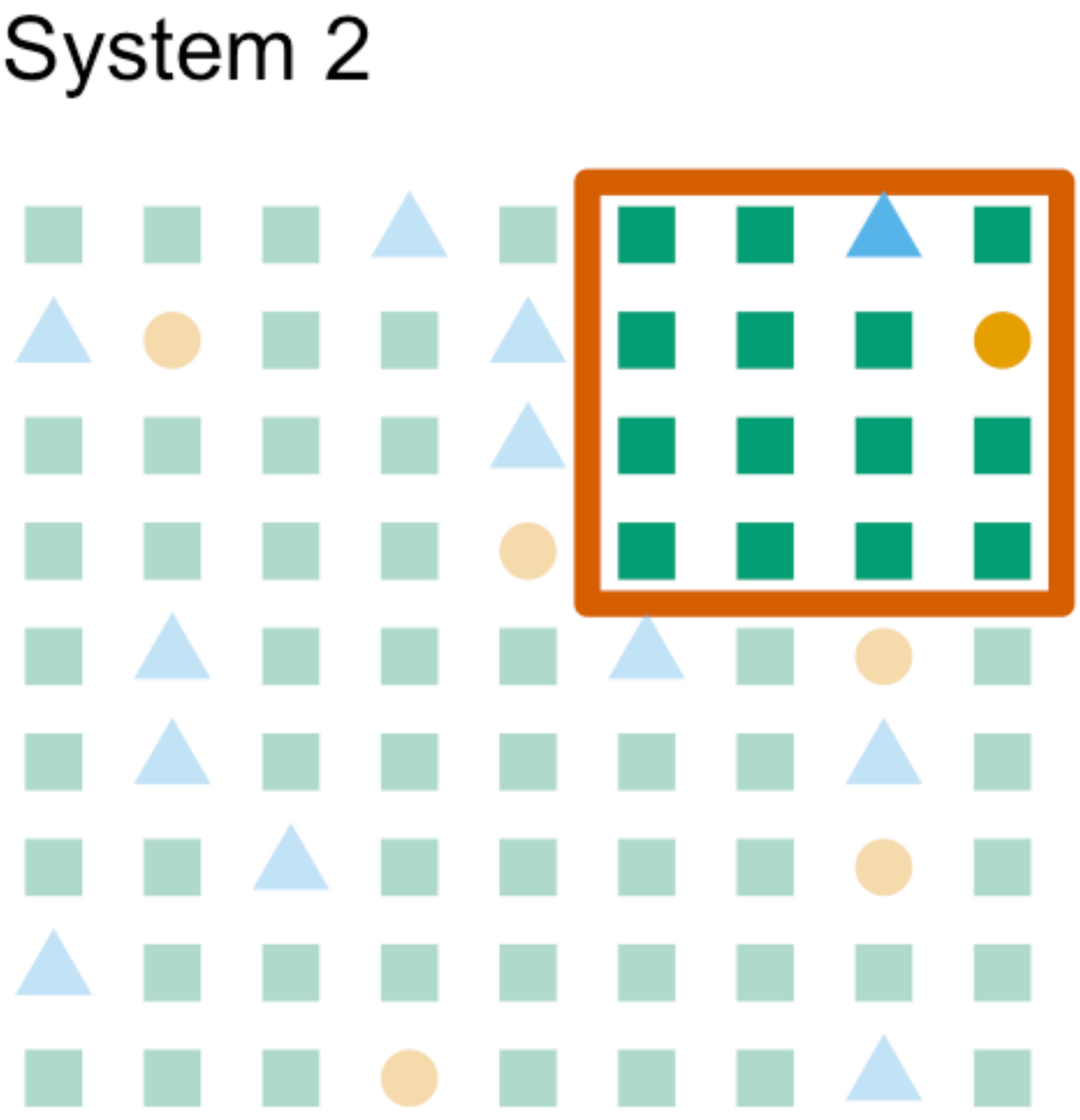
PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING



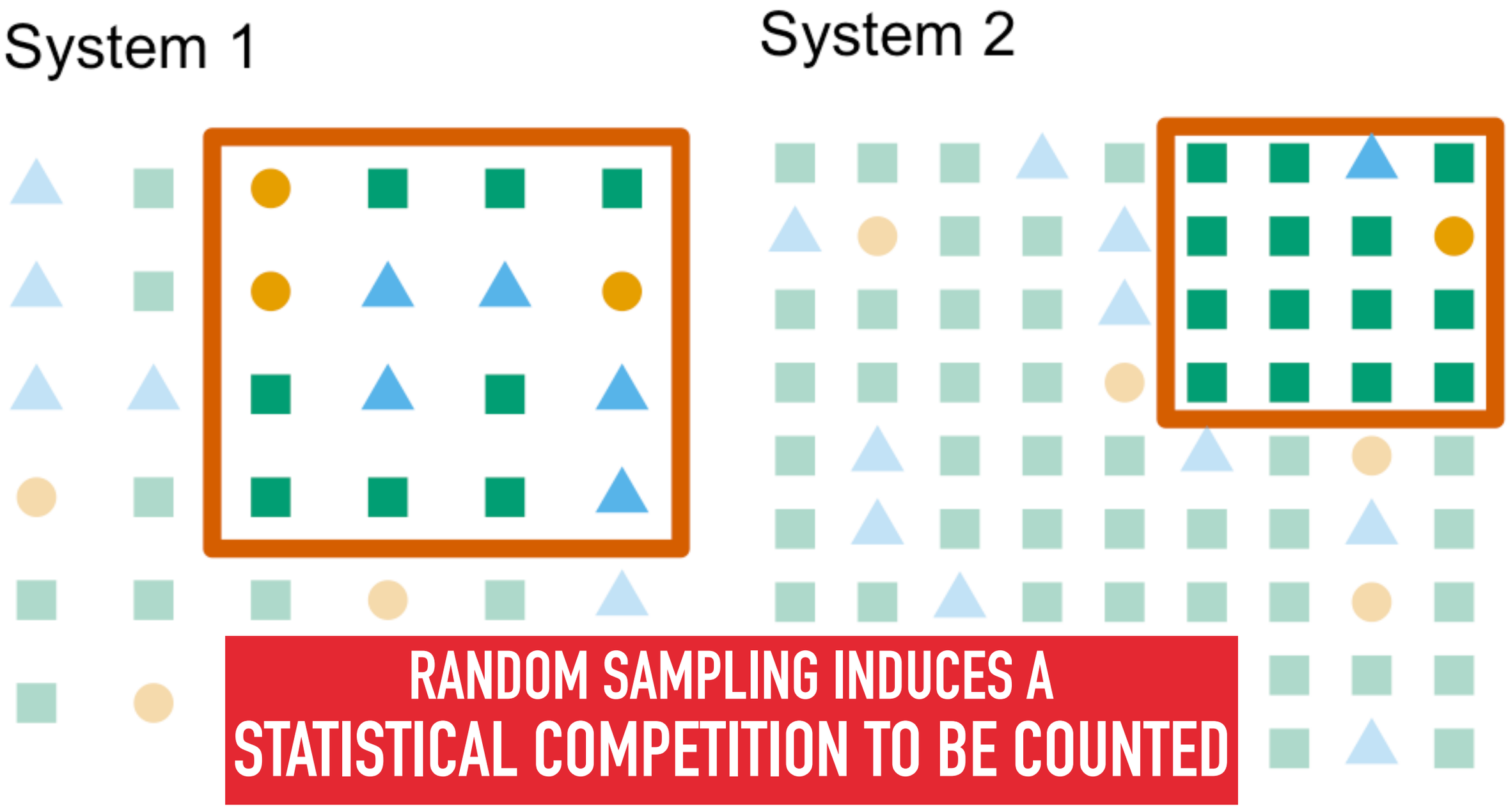
% Blue
% Orange
% Green

PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING

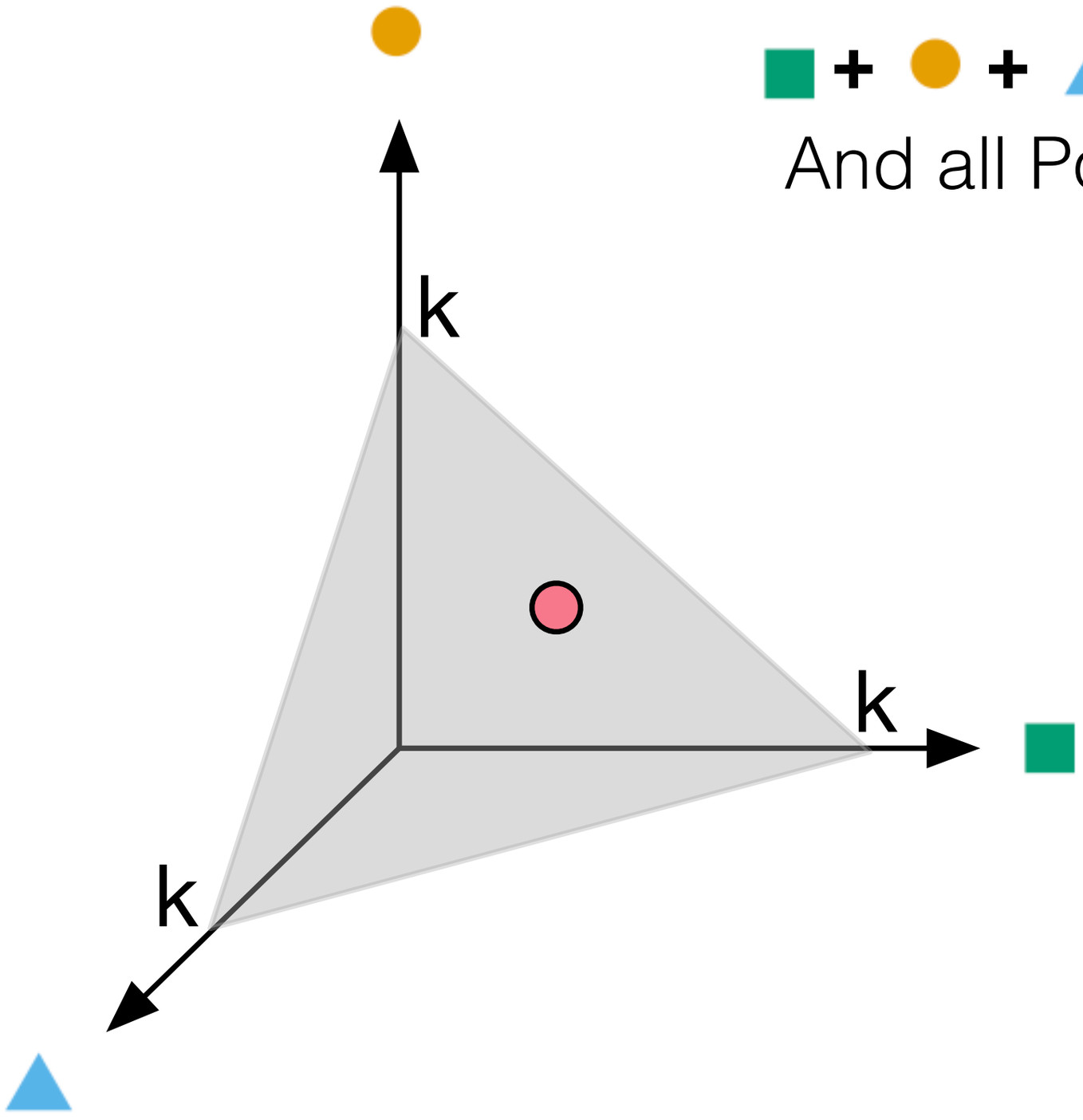




PROBLEM WITH MULTIVARIATE RANDOM SUBSAMPLING



THE SPACE OF RELATIVE DATA

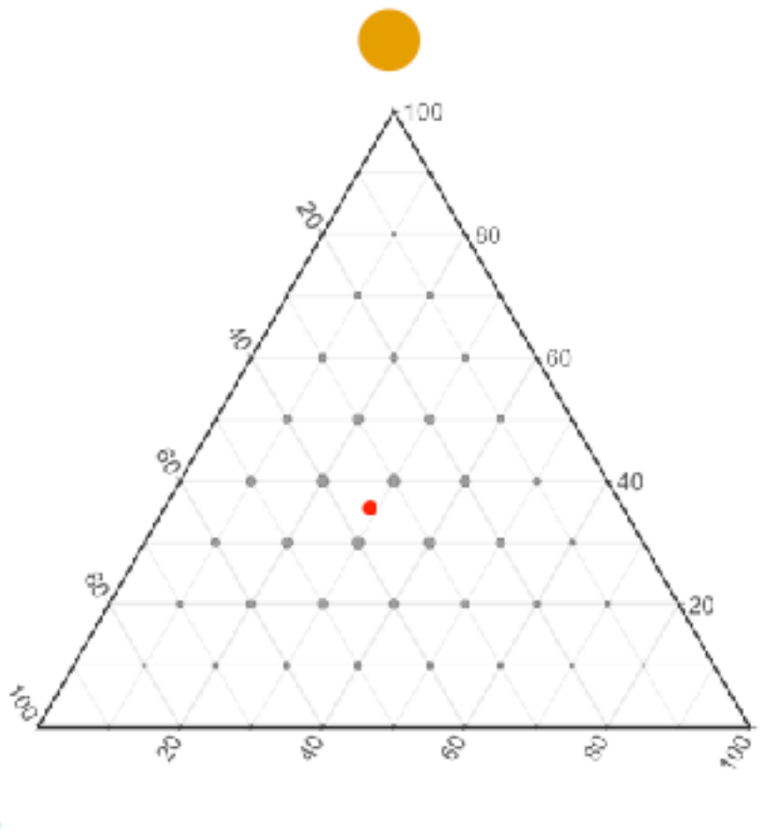


■ + ● + ▲ =k

And all Positive

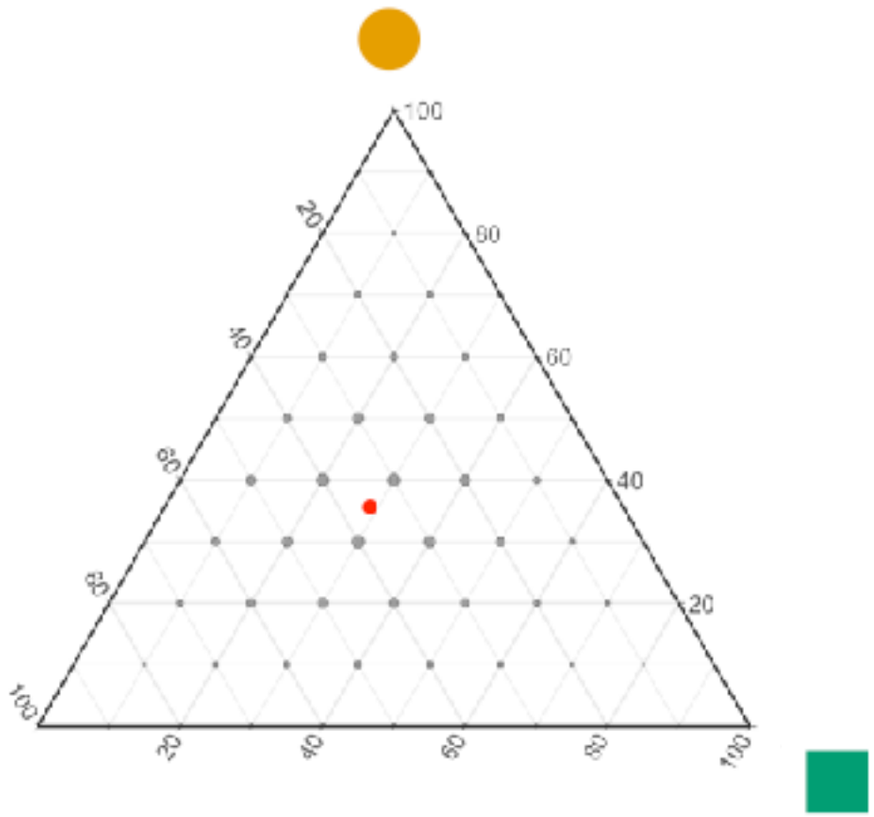
COUNTING INTRODUCES UNCERTAINTY

N=10

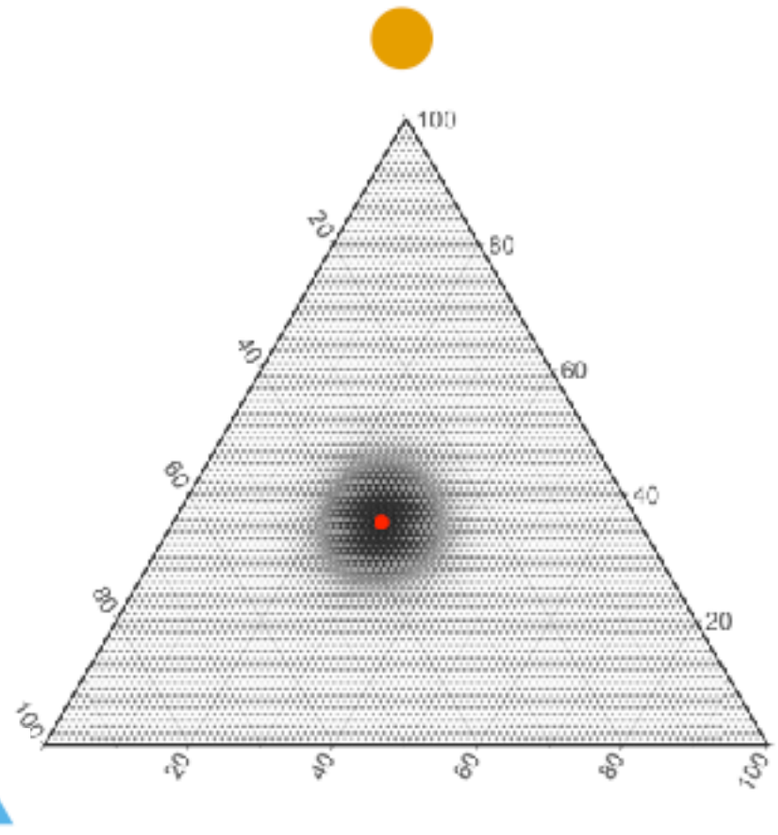


COUNTING INTRODUCES UNCERTAINTY

N=10



N=100



THE PROBLEM WITH FEW COUNTS OR SMALL COUNTS

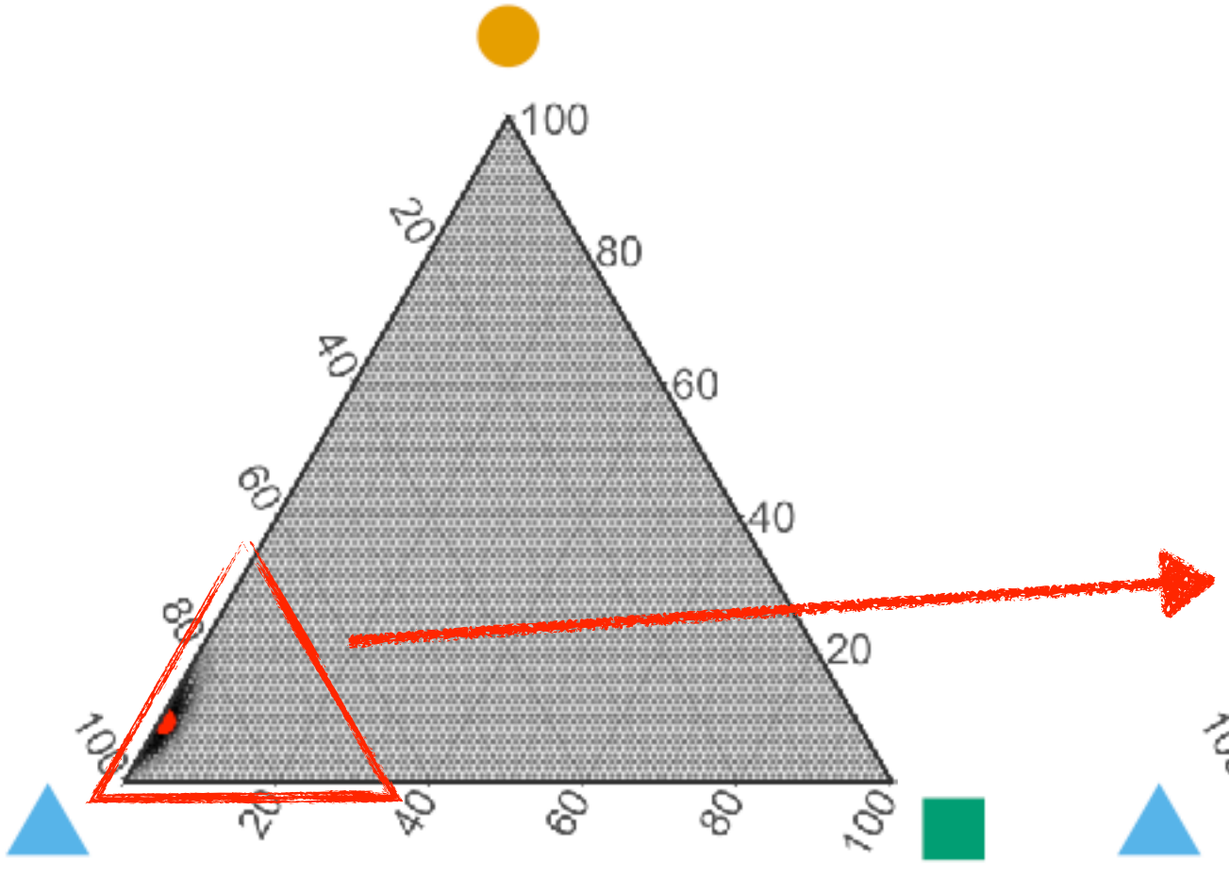
N=100

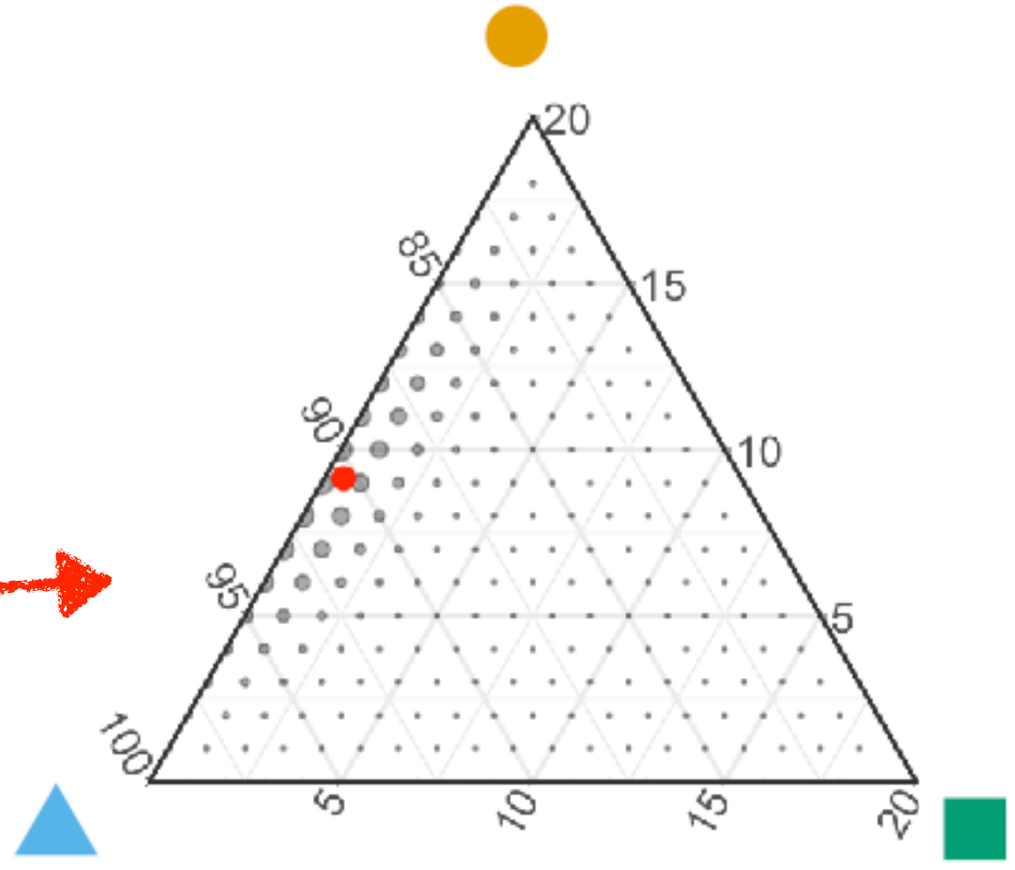
P=(90.3%, 9.2%, 0.5%)

THE PROBLEM WITH FEW COUNTS OR SMALL COUNTS

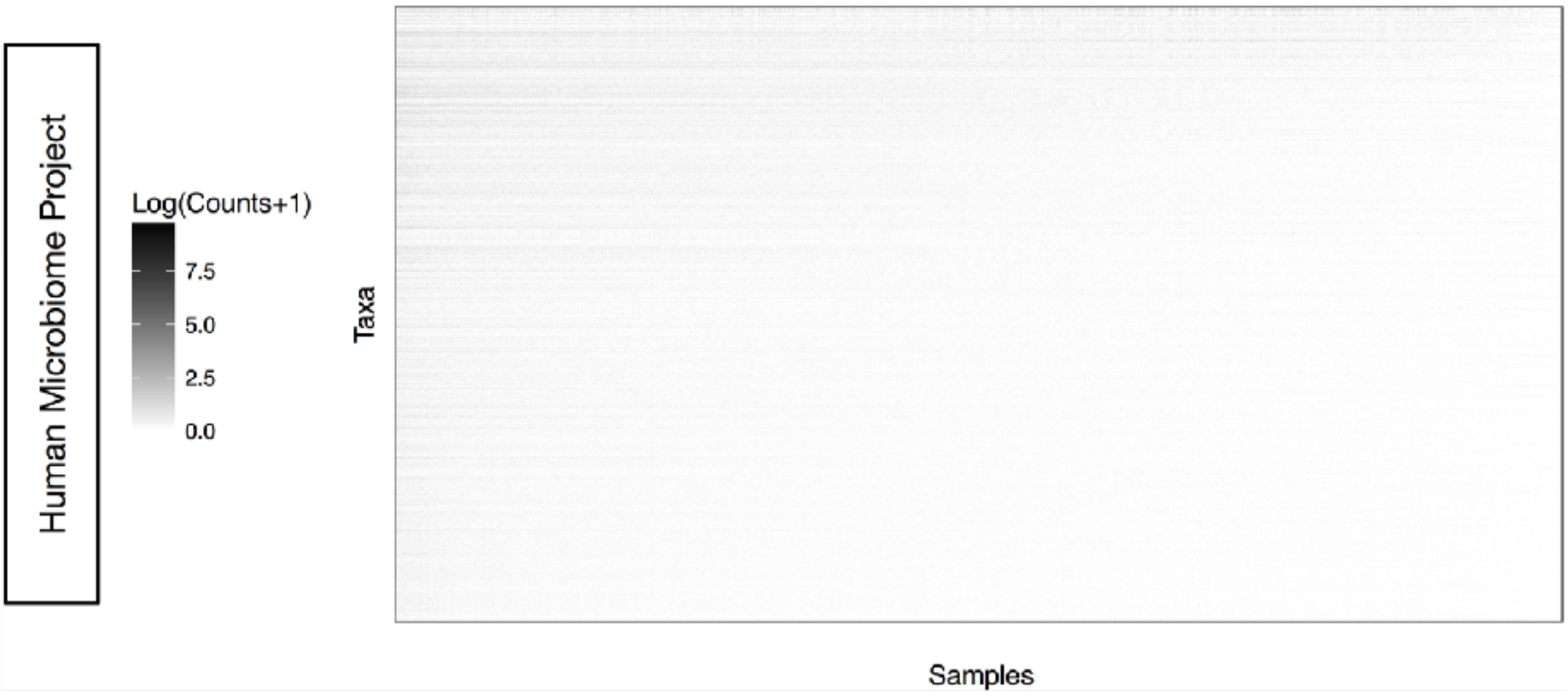
N=100

P=(90.3%, 9.2%, 0.5%)





MICROBIOME DATA IS SPARSE



Silverman, et al., *eLife* 2017

TAKE HOME MESSAGES

- Counts (with many zero or small counts)
- Compositional Information Only
- The data is multivariate not univariate!
- Lots of variation (technical and biological)
  - Variation is not always random (then we call it "bias")

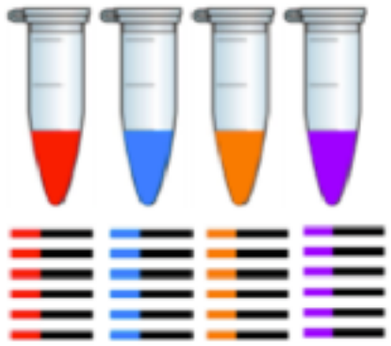
MODELING

GENERATIVE MODELING

Sequencing

$$Y_i \sim \text{Multinomial}(n_i, \pi_i)$$

DNA Extraction
PCR Amplification



?

Sample Collection
and Storage

?

Adapted from Hamady. et al., *Nature Methods*, 2008

FIVE (NEARLY) EQUIVALENT STATEMENTS

FIVE (NEARLY) EQUIVALENT STATEMENTS

- The Aitchison Geometry in the Simplex is the relevant space to model our systems in

FIVE (NEARLY) EQUIVALENT STATEMENTS

- The Aitchison Geometry in the Simplex is the relevant space to model our systems in
- "Our Systems Multiply" + The information in our data is relative

GROUP THEORY / VECTOR SPACE

FIVE (NEARLY) EQUIVALENT STATEMENTS

- The Aitchison Geometry in the Simplex is the relevant space to model our systems in
- "Our Systems Multiply" + The information in our data is relative

GROUP THEORY / VECTOR SPACE

- Conclusions should be drawn from [Log]-Ratios

INTUITIVE

FIVE (NEARLY) EQUIVALENT STATEMENTS

- The Aitchison Geometry in the Simplex is the relevant space to model our systems in
- "Our Systems Multiply" + The information in our data is relative

GROUP THEORY / VECTOR SPACE

- Conclusions should be drawn from [Log]-Ratios

INTUITIVE

- The Logistic-Normal Distribution is the CLT for our unobserved system(s)

STATISTICAL

FIVE (NEARLY) EQUIVALENT STATEMENTS

- The Aitchison Geometry in the Simplex is the relevant space to model our systems in
- "Our Systems Multiply" + The information in our data is relative

GROUP THEORY / VECTOR SPACE

- Conclusions should be drawn from [Log]-Ratios

INTUITIVE

- The Logistic-Normal Distribution is the CLT for our unobserved system(s)

STATISTICAL

- All methods for analyzing relative data should adhere to three principles (1) Scale Invariance, (2) Permutation Invariance, (3) Subcompositional Coherence

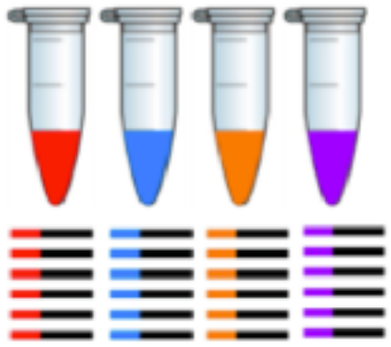
AXIOMATIC

GENERATIVE MODELING (LIKELIHOOD ONLY)

Sequencing

$$Y_i \sim \text{Multinomial}(n_i, \pi_i)$$

DNA Extraction
PCR Amplification



$$\pi_i \sim \text{Logistic Normal}(\mu_i, V)$$

Sample Collection
and Storage

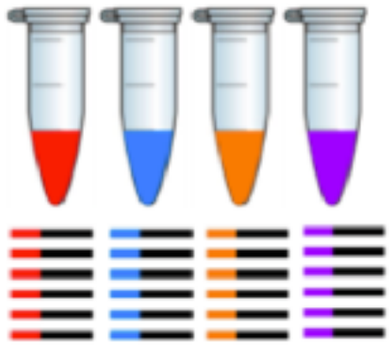
$$\mu_i \sim \text{Logistic Normal}(a, V)$$

Adapted from Hamady. et al., *Nature Methods*, 2008

WHAT IS MISSING?

Sequencing

DNA Extraction
PCR Amplification



Sample Collection
and Storage

Assign Sequences
to Samples

```
>AGTGAGAGAAGCAGGGTCGTAATGTT . . .
>AGTGCGATGCGTAGGGTCGTAATGCG . . .
>AGTGCGATGCGTAGGGTCGTAATGTA . . .
>AGTGGATGCTCTAGGGTCGTAATGCA . . .
>AGTGTCACGGTGAGGGTCGTAATGGG . . .
>AGTGGATGCTCTAGGGTCGTAATGTT . . .
>AGTGTCACGGTGAGGGTCGTAATGCC . . .
>AGTGAGAGAAGCAGGGTCGTAATCAC . . .
. . .
```

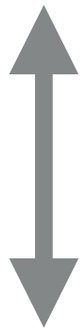
Denoise Reads
or Cluster

	Species 1	Species 2	Species 3	Sp
Sample 1	23	53	2	
Sample 2	69	64	70	
Sample 3	33	100	68	
Sample 4	5	63	57	
Sample 5	76	80	46	
Sample 6	58	7	37	
Sample 7	10	87	32	
Sample 8	21	89	73	

Adapted from Hamady. et al., *Nature Methods*, 2008

MULTINOMIAL–LOGISTIC NORMAL (OR NORMAL ON THE SIMPLEX)

$$Y \sim \text{Multinomial}(\pi)$$
$$\pi \sim \text{Logistic Normal}(\rho, \Xi)$$

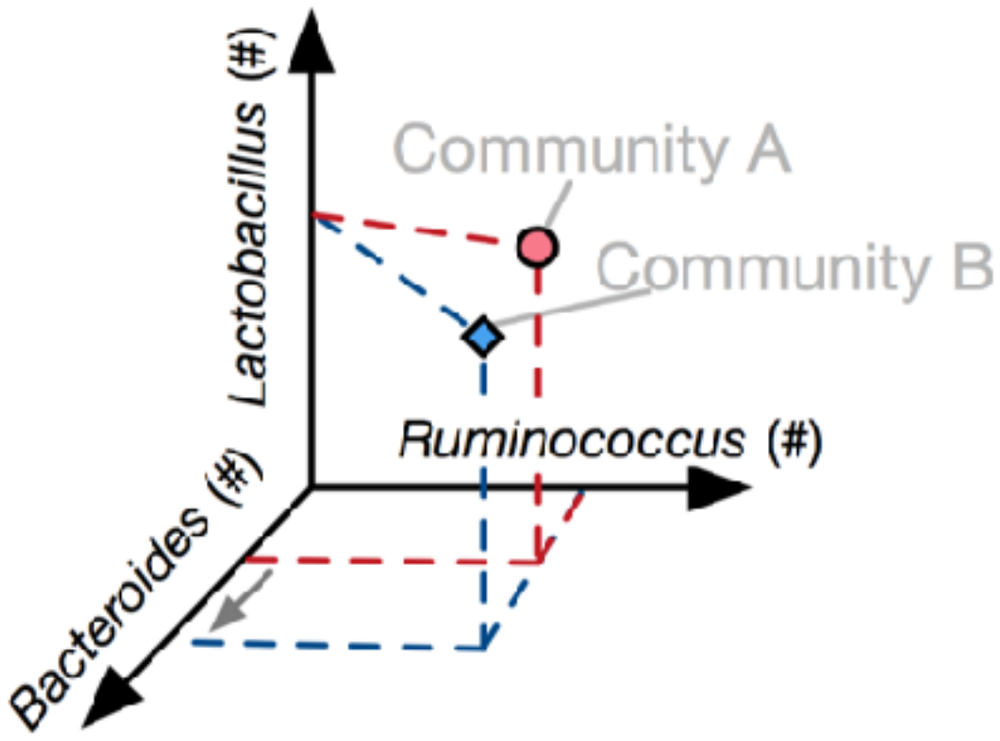


$$Y \sim \text{Multinomial}(\pi)$$
$$\pi = \text{ILR}^{-1}(\eta)$$
$$\eta \sim \text{Multivariate Normal}(\mu, \Sigma)$$

INTERPRETING THE MODEL

ISOMETRIC LOGRATIO TRANSFORM – AN ORTHONORMAL BASIS

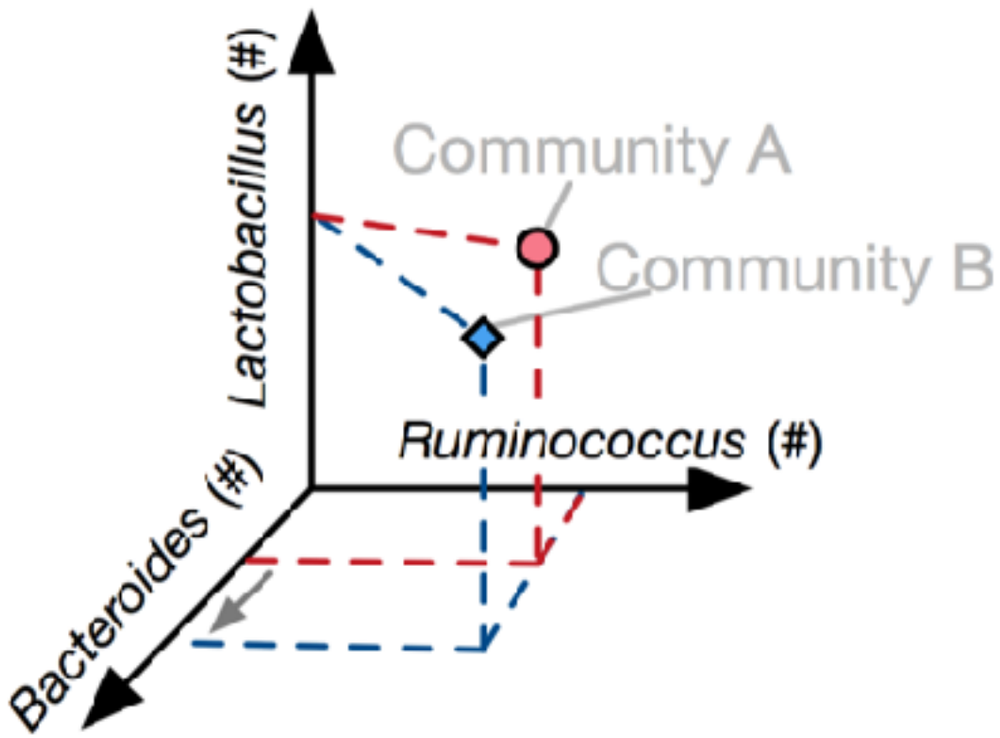
UNOBSERVED ABSOLUTE ABUNDANCES



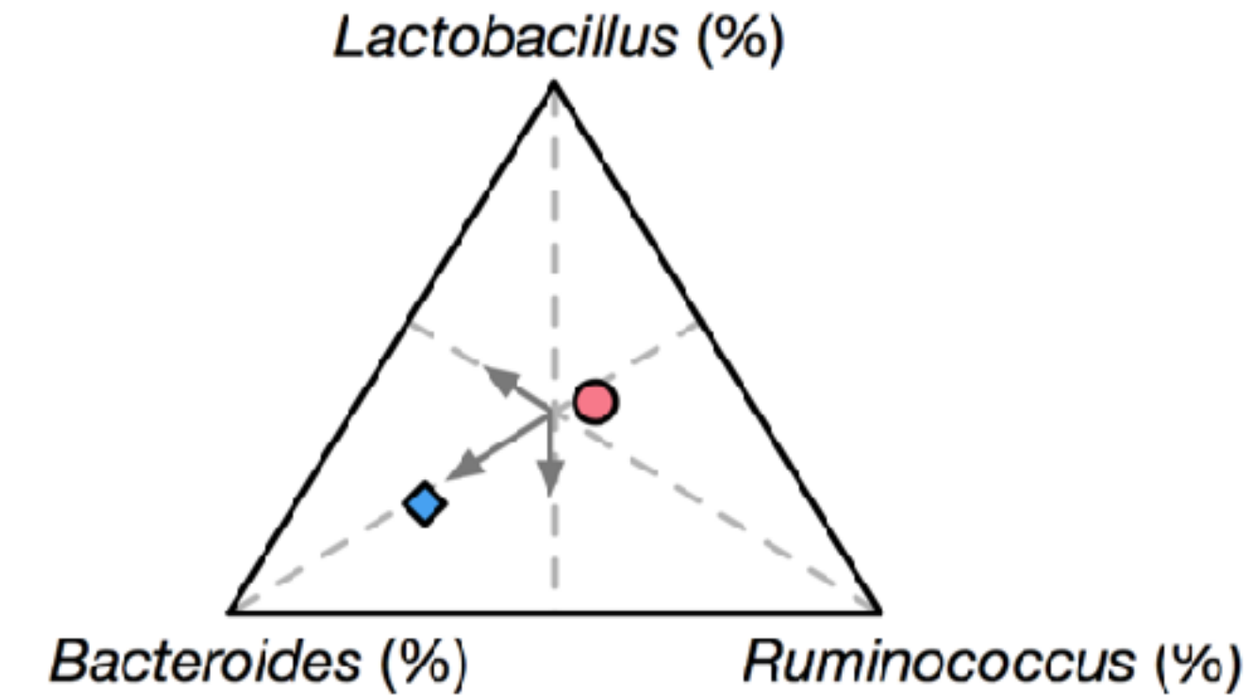
Silverman, et al., *eLife* 2017

ISOMETRIC LOGRATIO TRANSFORM – AN ORTHONORMAL BASIS

UNOBSERVED ABSOLUTE ABUNDANCES



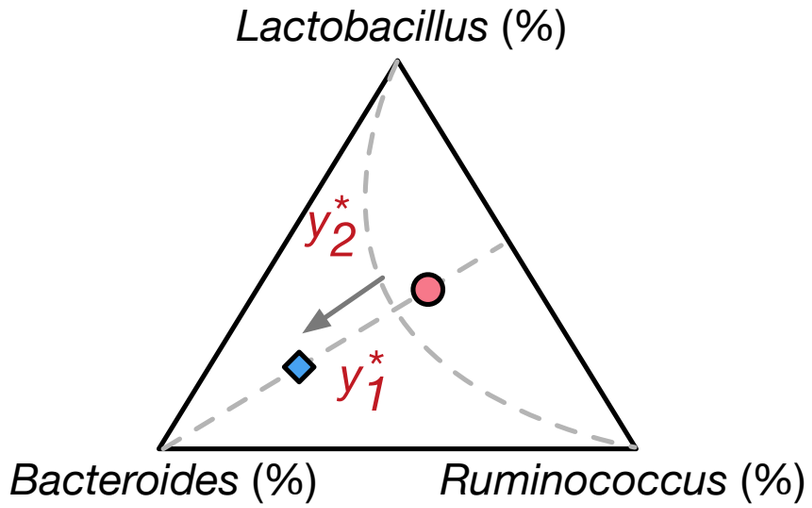
CORRESPONDING COMPOSITION



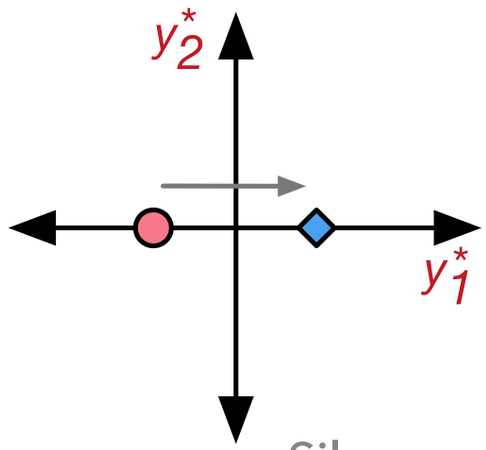
Silverman, et al., *eLife* 2017

ISOMETRIC LOGRATIO TRANSFORM

ORTHONORMAL BASIS IN SIMPLEX



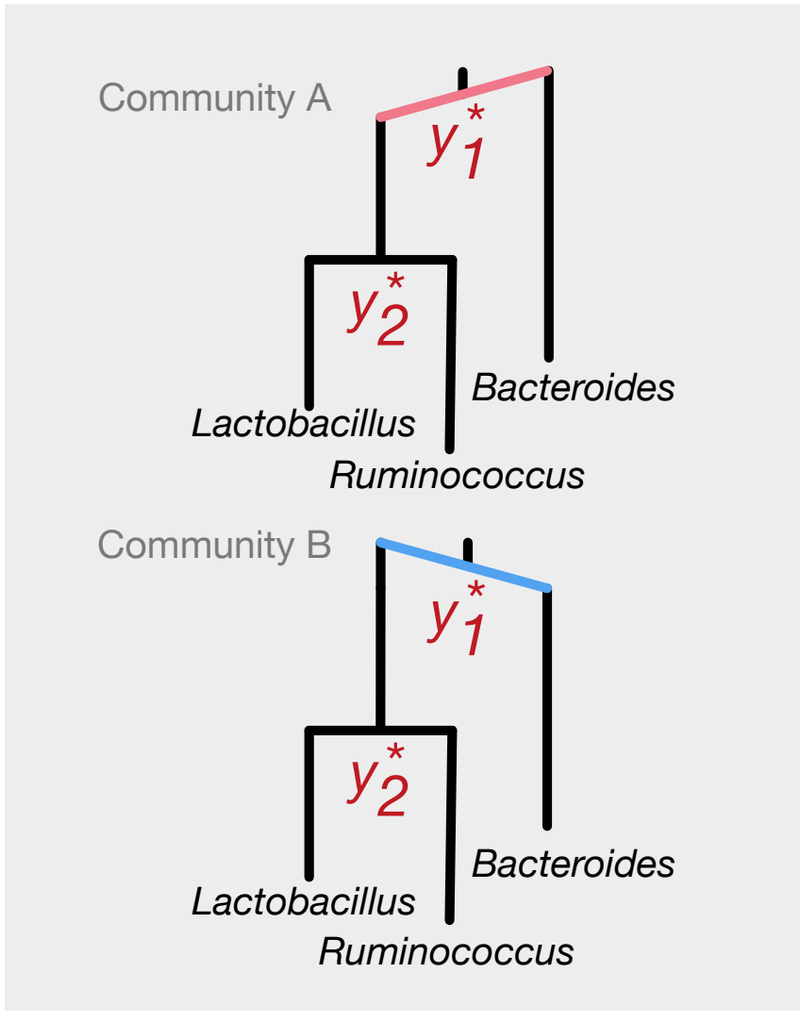
DATA PROJECTED ONTO BASIS



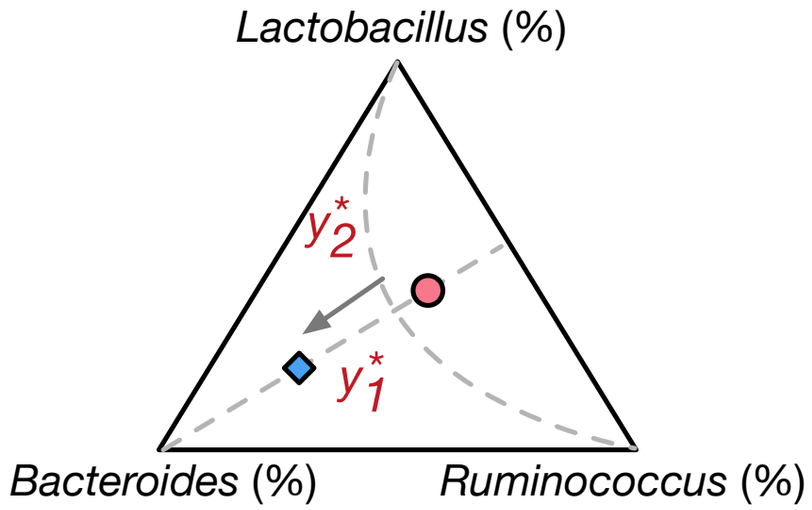
Silverman, et al., *eLife* 2017

PHYLOGENIC ISOMETRIC LOGRATIO (PHILR) TRANSFORM

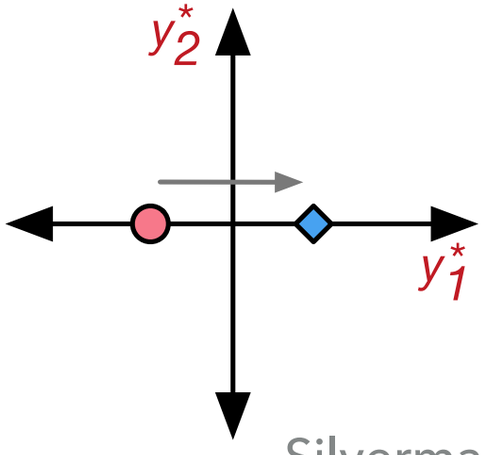
PHYLOGENETIC BALANCES



ORTHONORMAL BASIS IN SIMPLEX



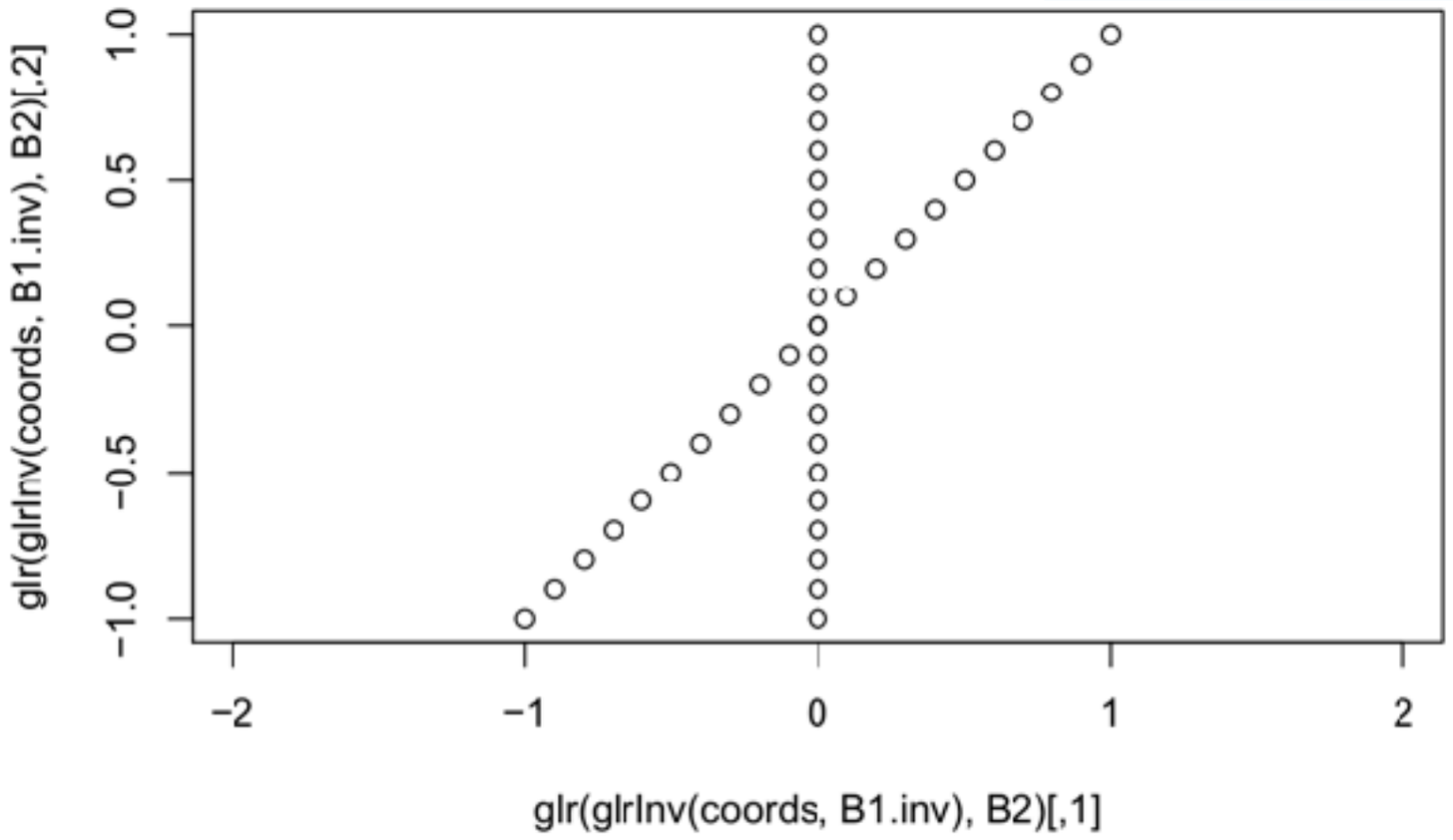
DATA PROJECTED ONTO BASIS



Silverman, et al., *eLife* 2017

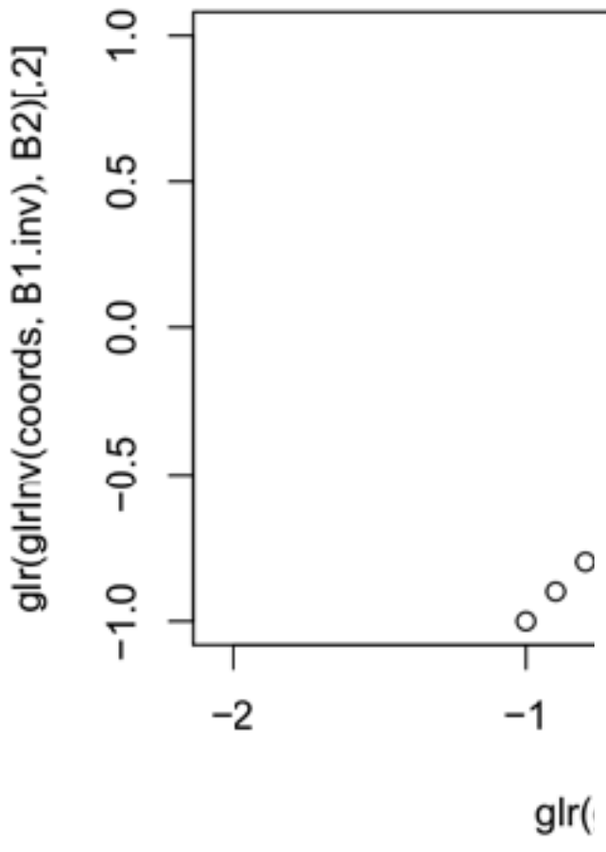
WHY AN ORTHONORMAL BASIS?

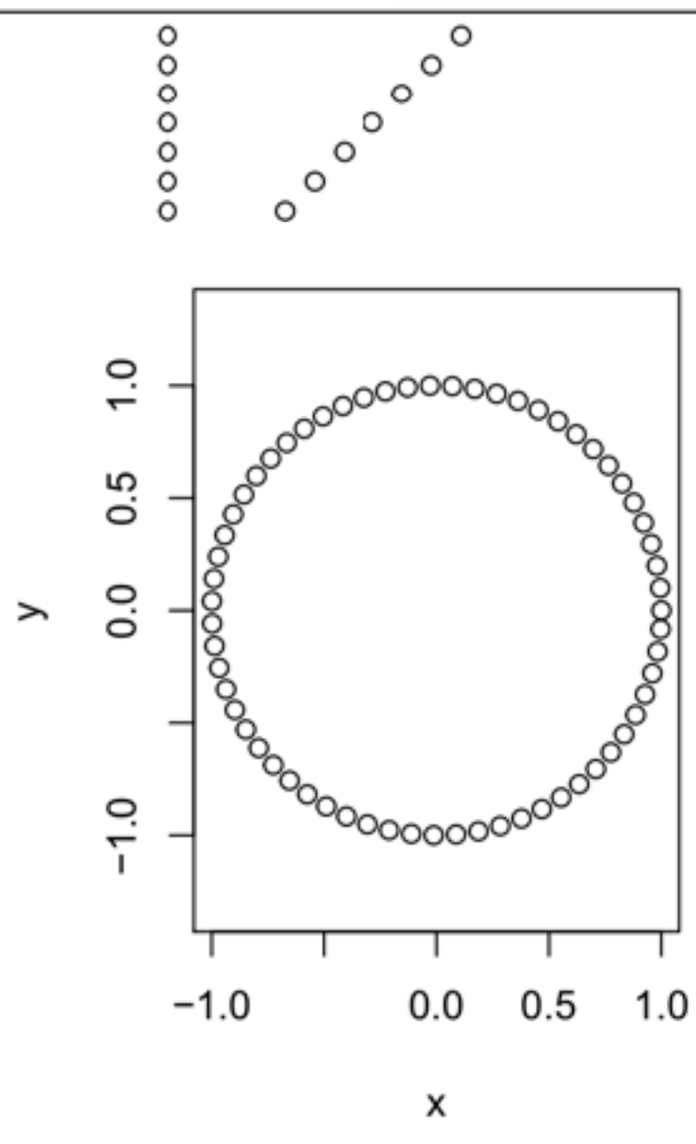
CURRENT STATISTICAL STANDARD "IDENTIFIED SOFTMAX" (AKA INVERSE ALR)

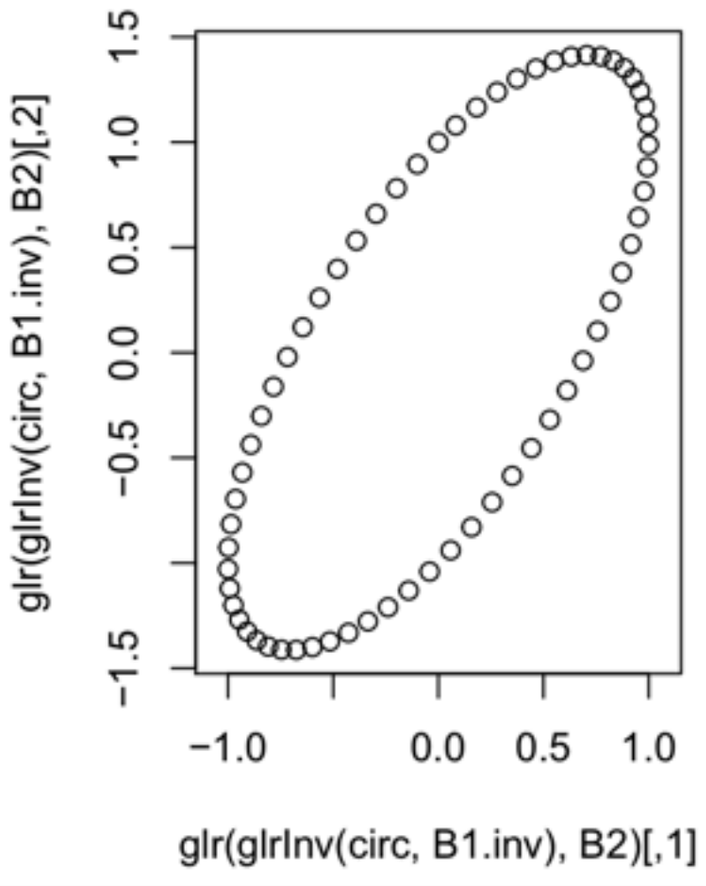


WHY AN ORTHONORMAL BASIS?

CURRENT STATISTICAL STANDARD "IDENTIFIED SOFTMAX" (AKA INVERSE ALR)

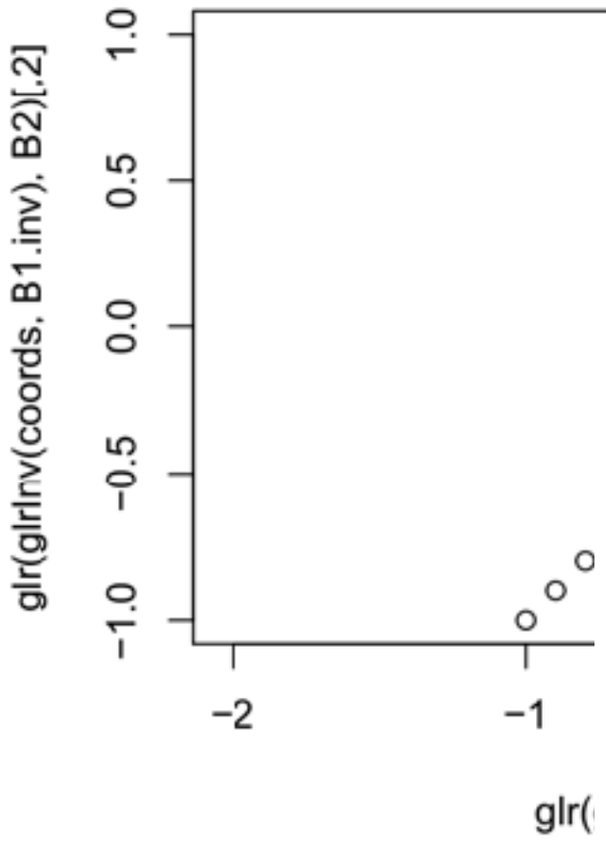


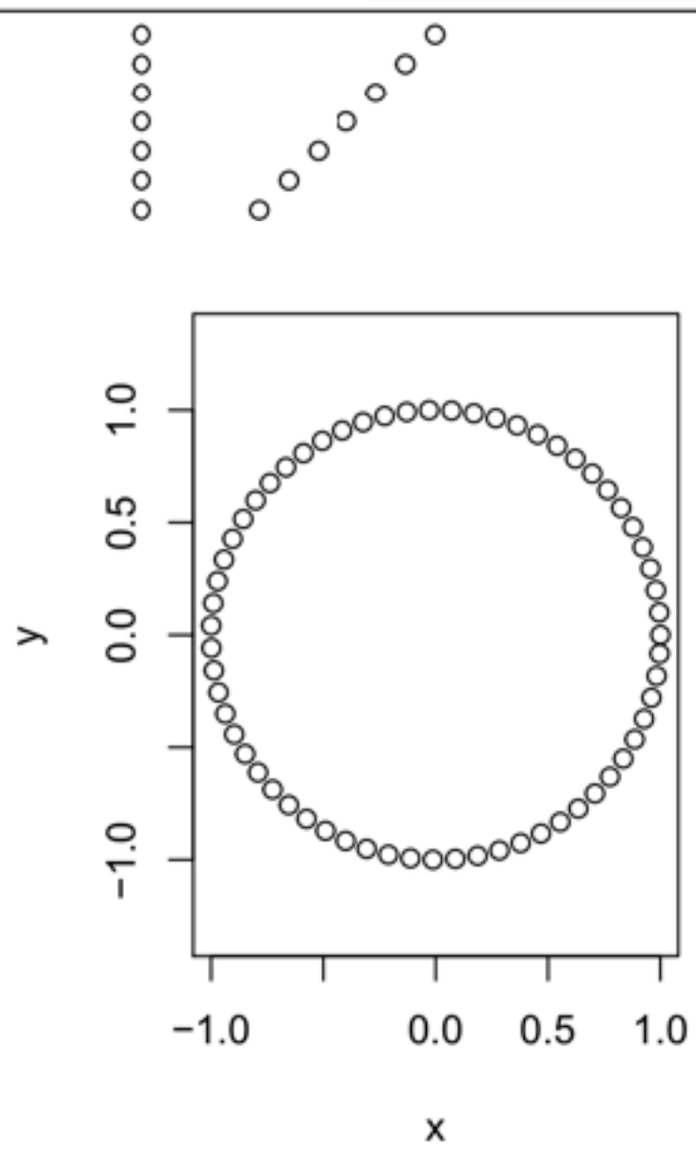


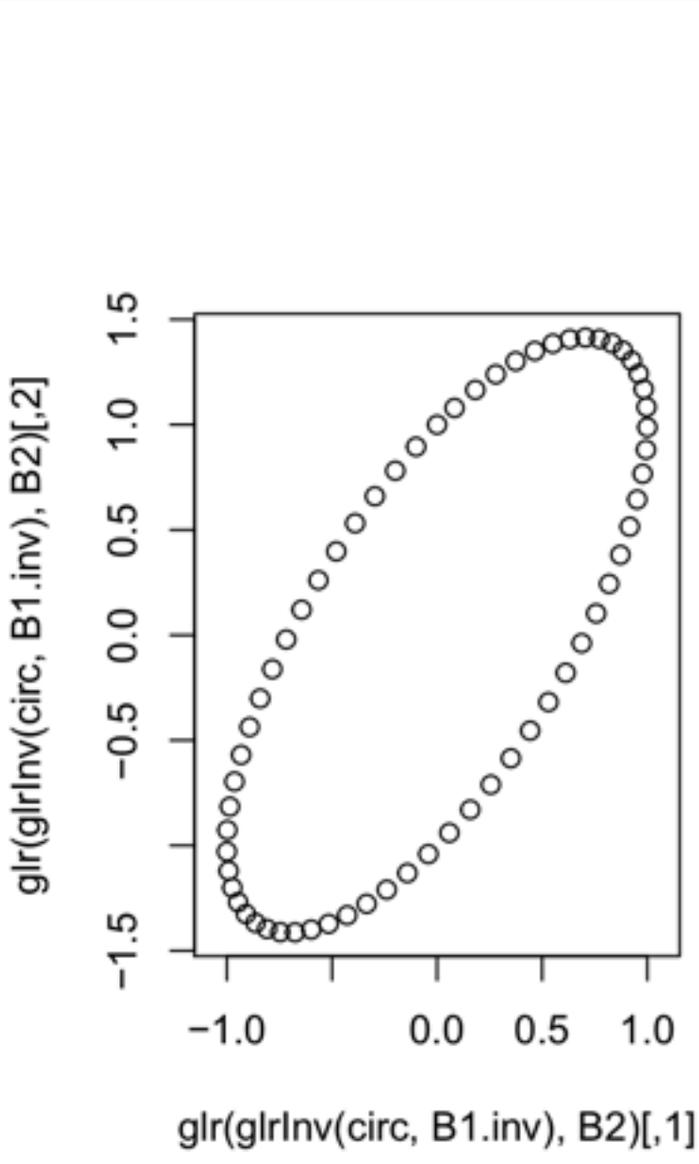


WHY AN ORTHONORMAL BASIS?

CURRENT STATISTICAL STANDARD "IDENTIFIED SOFTMAX" (AKA INVERSE ALR)

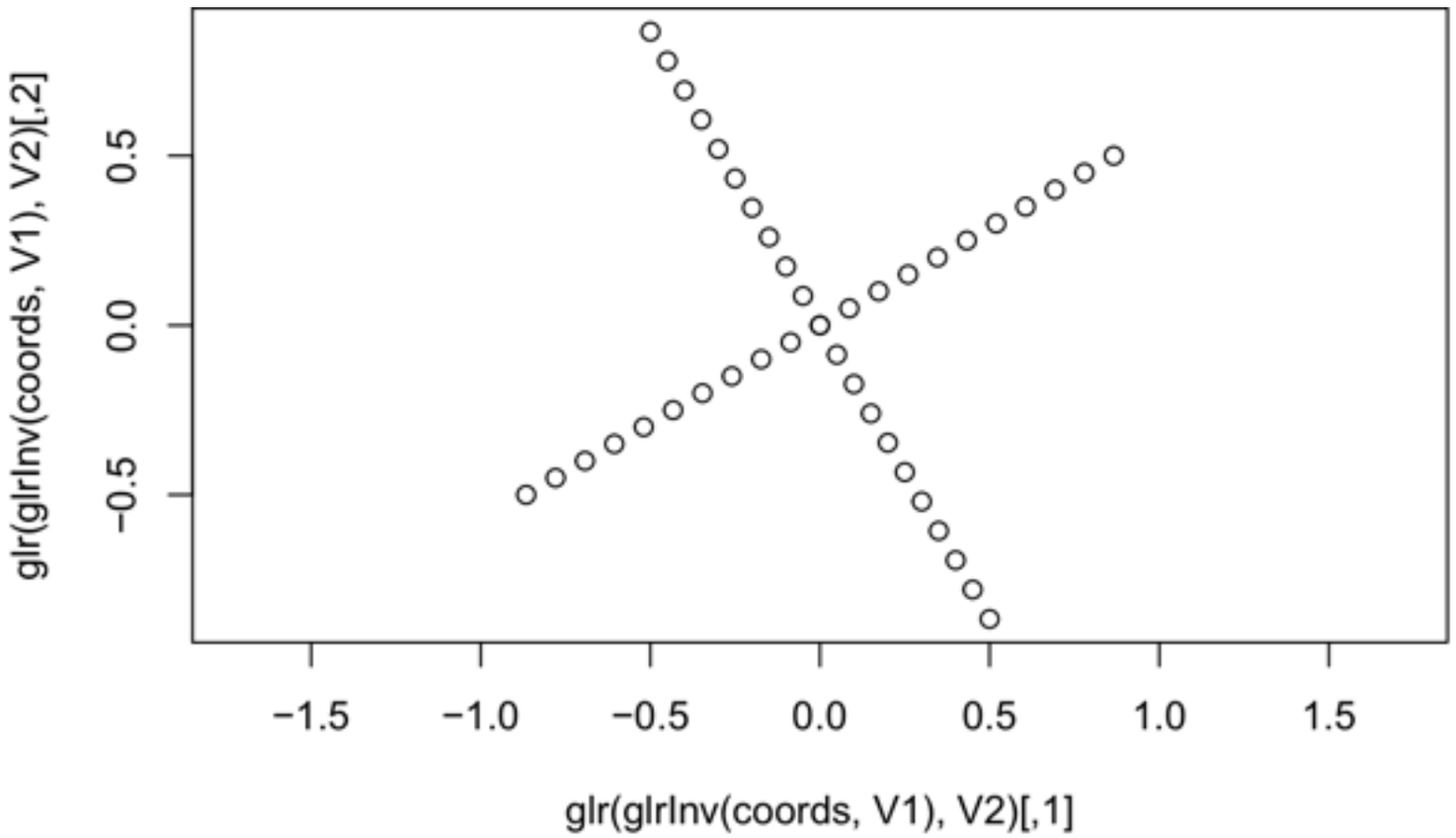






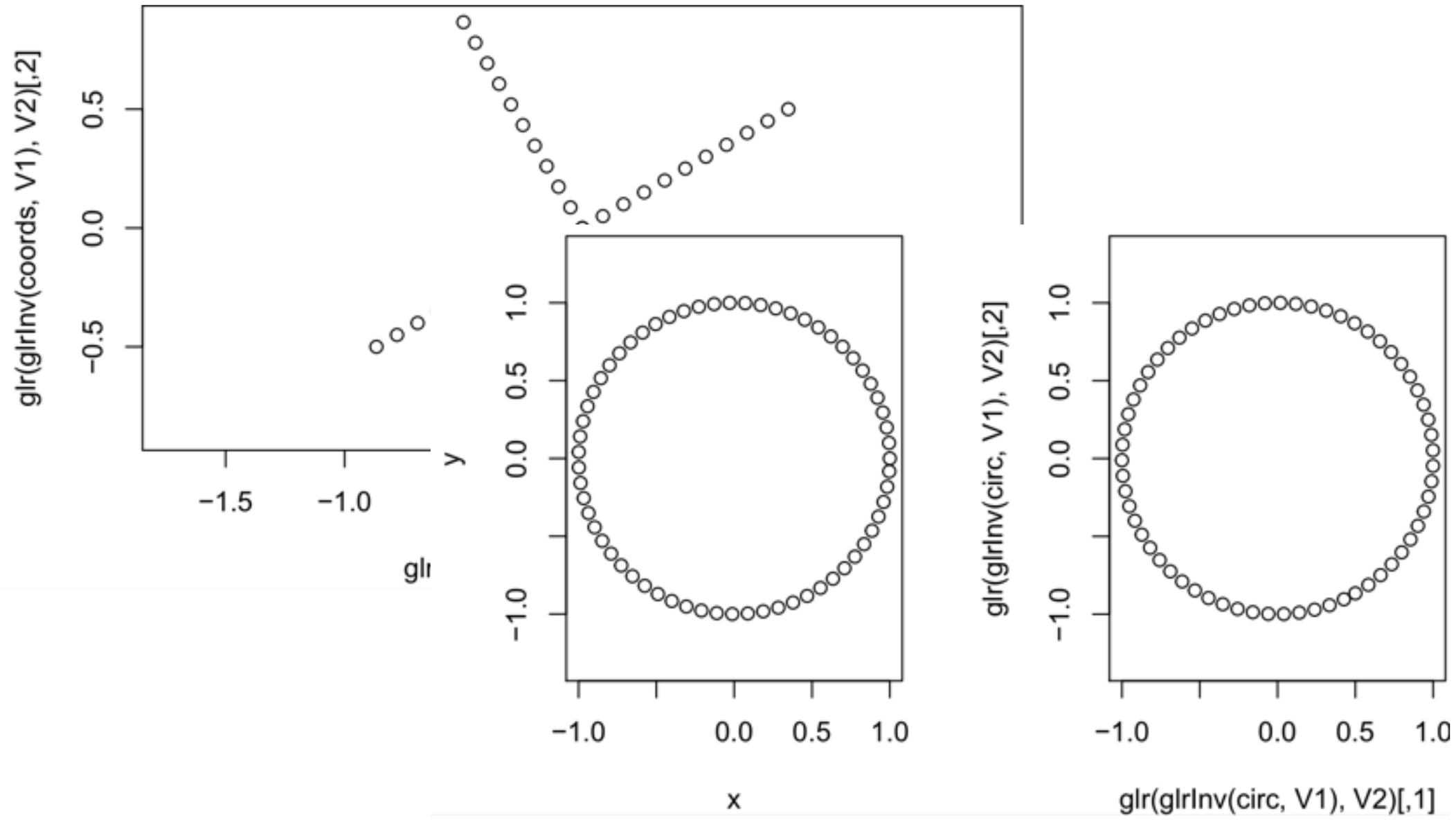
WHY AN ORTHONORMAL BASIS?

ORTHONORMAL BASIS



WHY AN ORTHONORMAL BASIS?

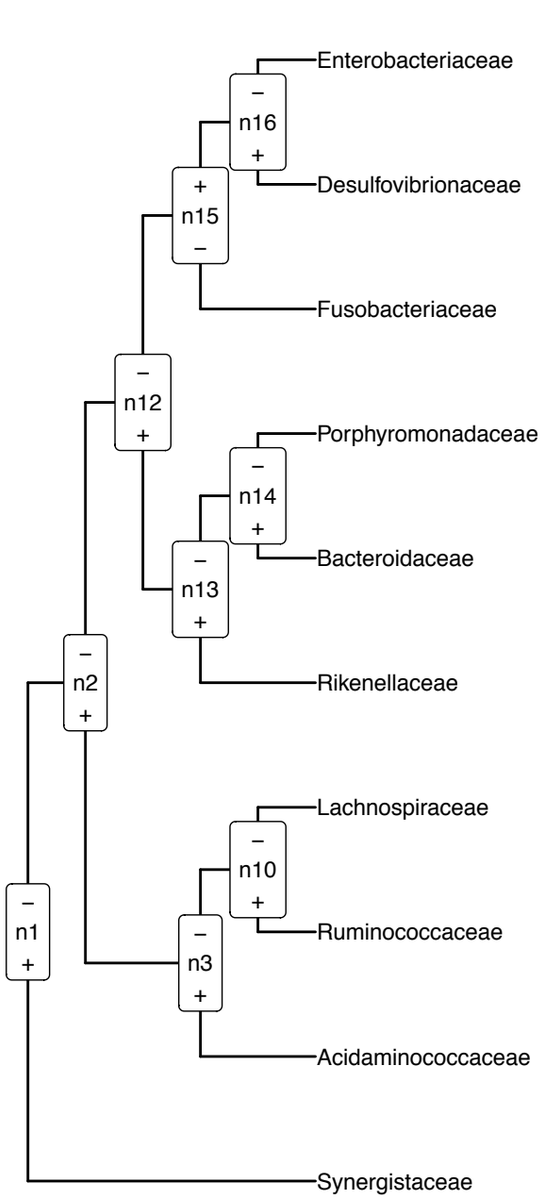
ORTHONORMAL BASIS

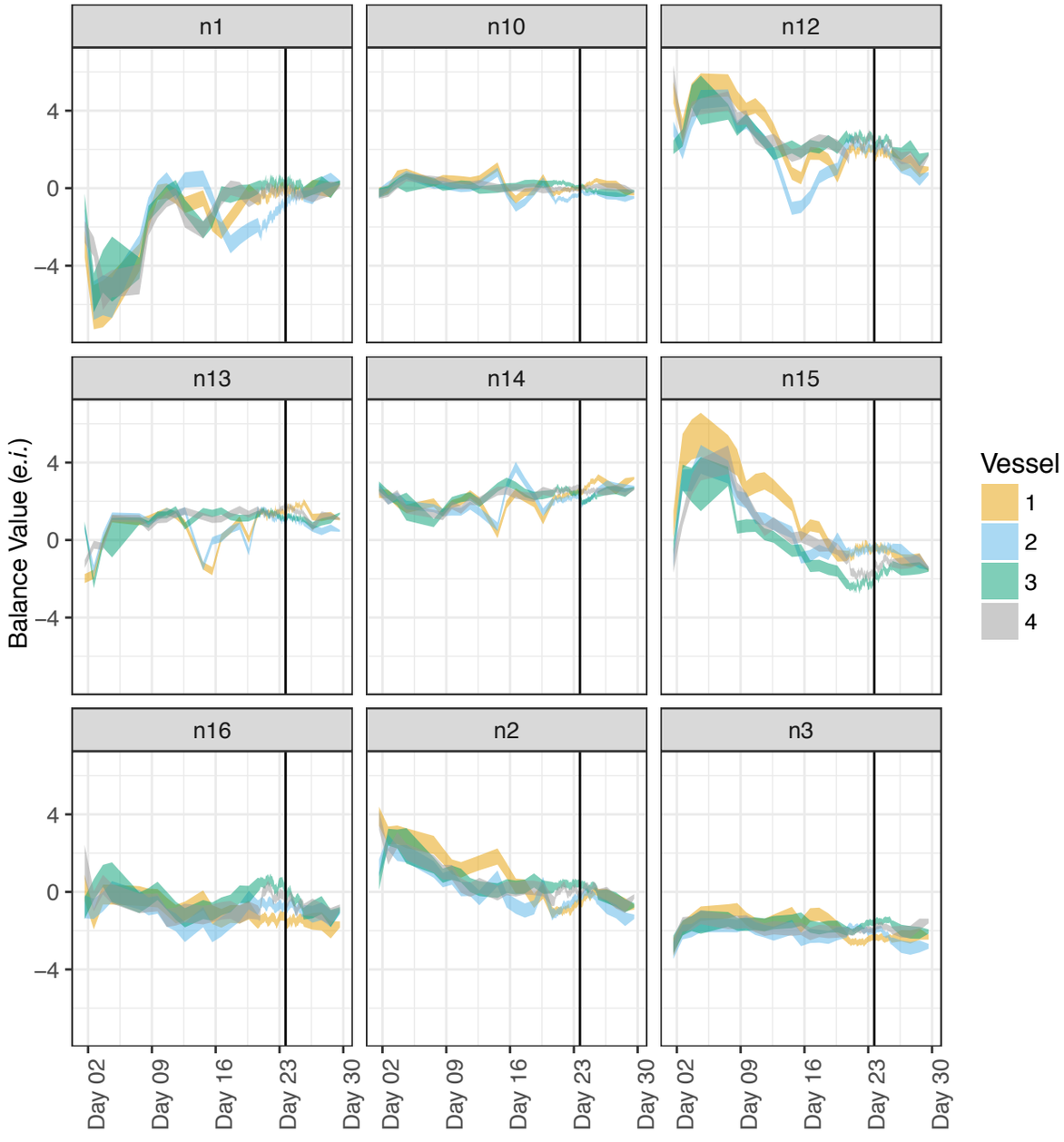


WHY AN ORTHONORMAL BASIS?

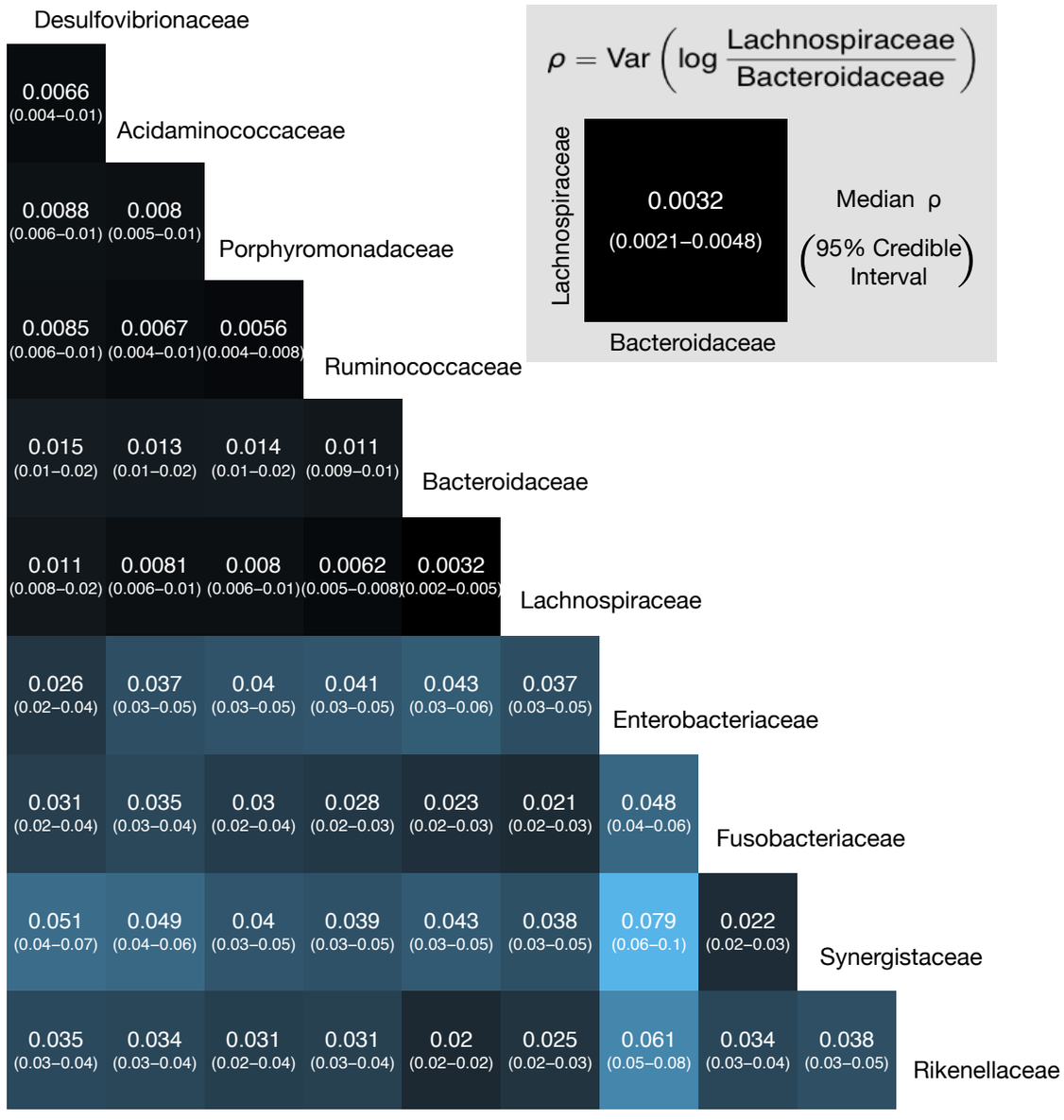
- Ability to rotate your view requires orthonormal basis.
- Interpretability of low-dimensional projections requires orthonormal basis
- "Objects change when you look at them differently with non-orthonormal bases"
- "Units" Require an orthonormal Basis (Evidence Information)

PHILR BASIS

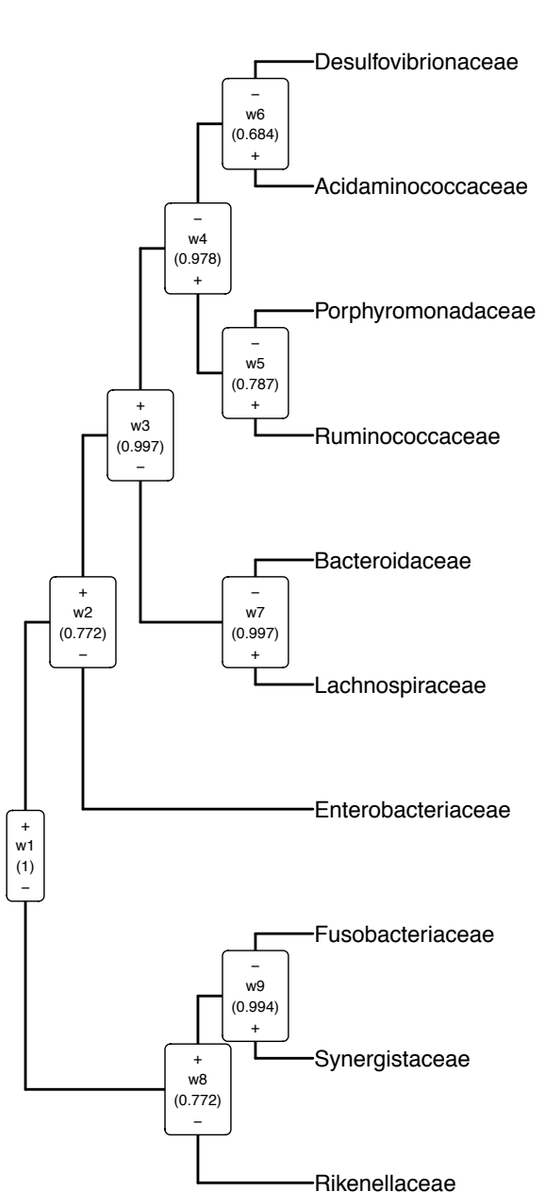

Enterobacteriaceae
n16
Desulfovibrionaceae
n15
Fusobacteriaceae
n12
Porphyromonadaceae
n14
Bacteroidaceae
n13
Rikenellaceae
n2
Lachnospiraceae
n10
Ruminococcaceae
n3
Acidaminococcaceae
n1
Synergistaceae

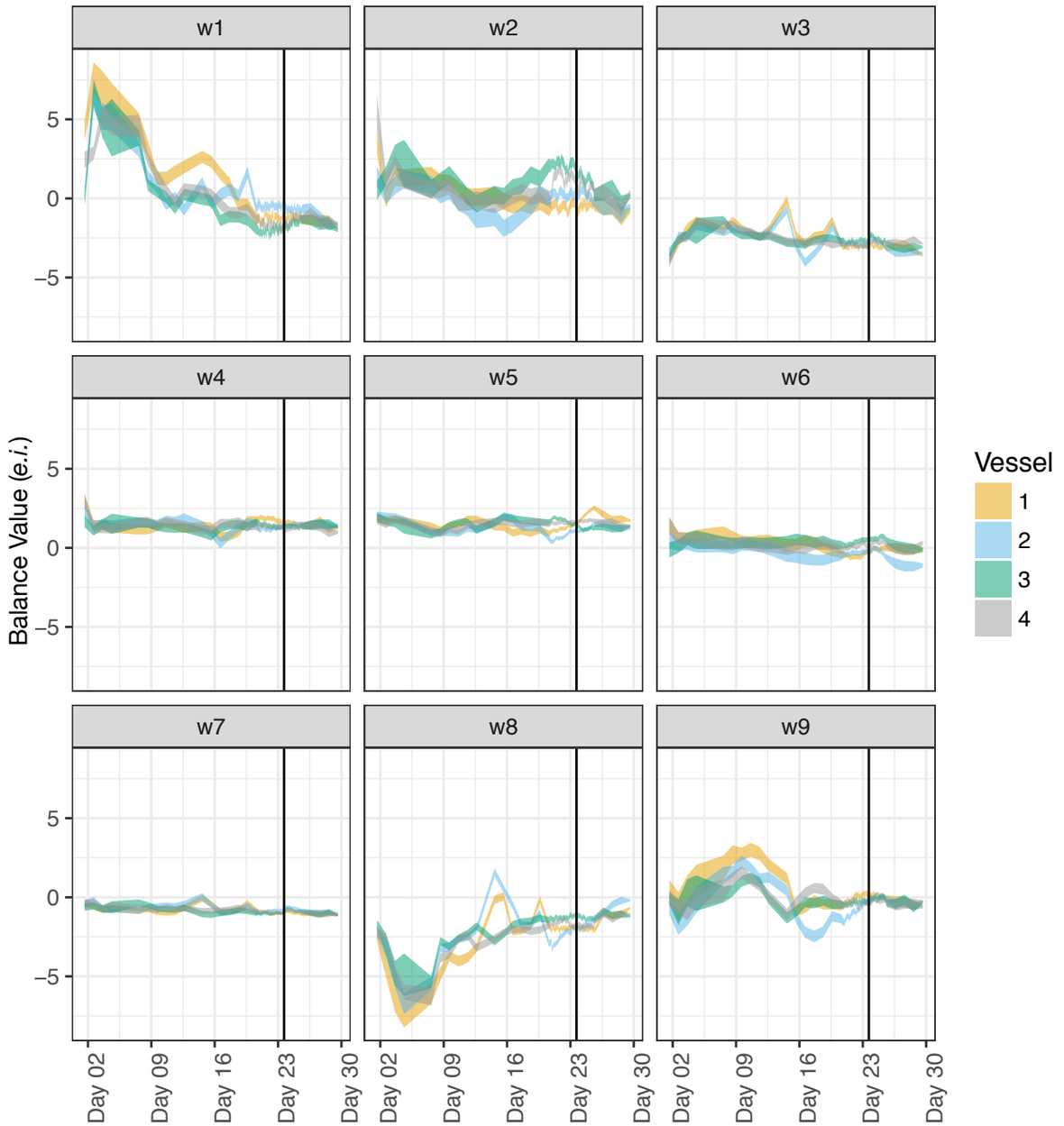

n1
n10
n12
n13
n14
n15
n16
n2
n3
Balance Value (e.i.)
4
0
−4
Day 02
Day 09
Day 16
Day 23
Day 30
Vessel
1
2
3
4

VARIATION ARRAY

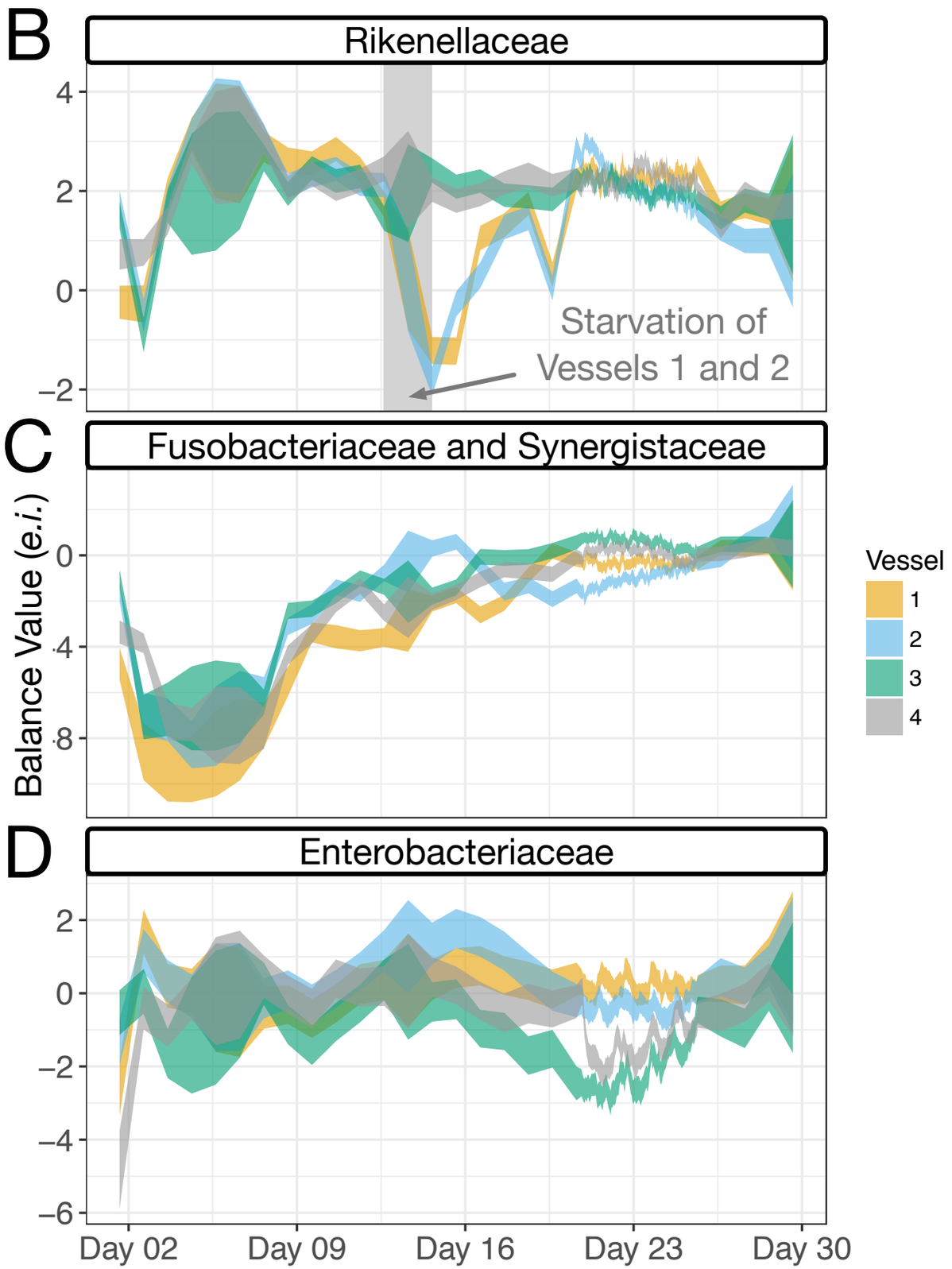


PRINCIPLE BALANCE ANALYSIS





MANUAL CURATION



EXAMPLE APPLICATIONS

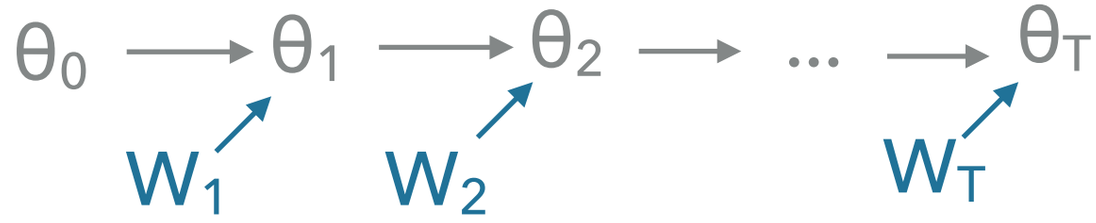
MODELING TIME-EVOLUTION WITH MALLARD

True State with Biological Noise

$$\theta_0 \longrightarrow \theta_1 \longrightarrow \theta_2 \longrightarrow \dots \longrightarrow \theta_T$$

MODELING TIME-EVOLUTION WITH MALLARD

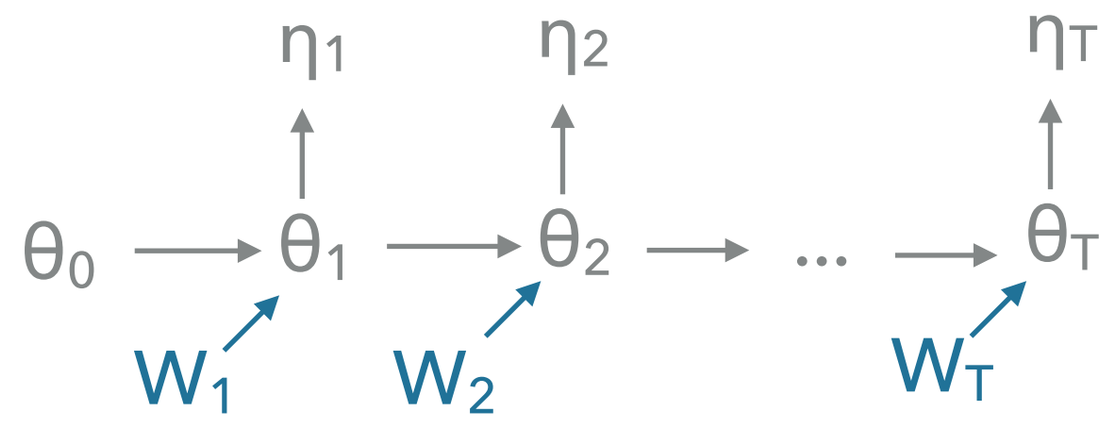
True State with Biological Noise



MODELING TIME-EVOLUTION WITH MALLARD

Addition of Technical Noise

True State with Biological Noise

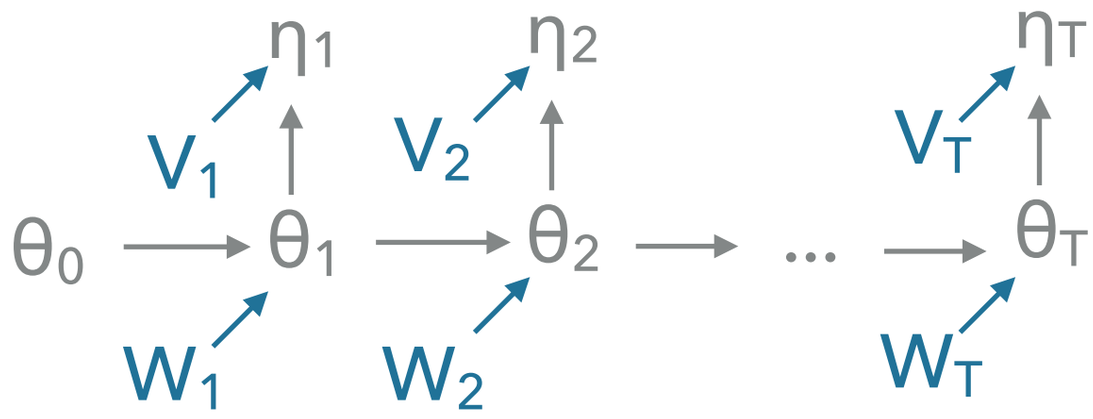


MODELING TIME-EVOLUTION WITH MALLARD

Addition of Technical Noise

↑

True State with Biological Noise



MODELING TIME-EVOLUTION WITH MALLARD

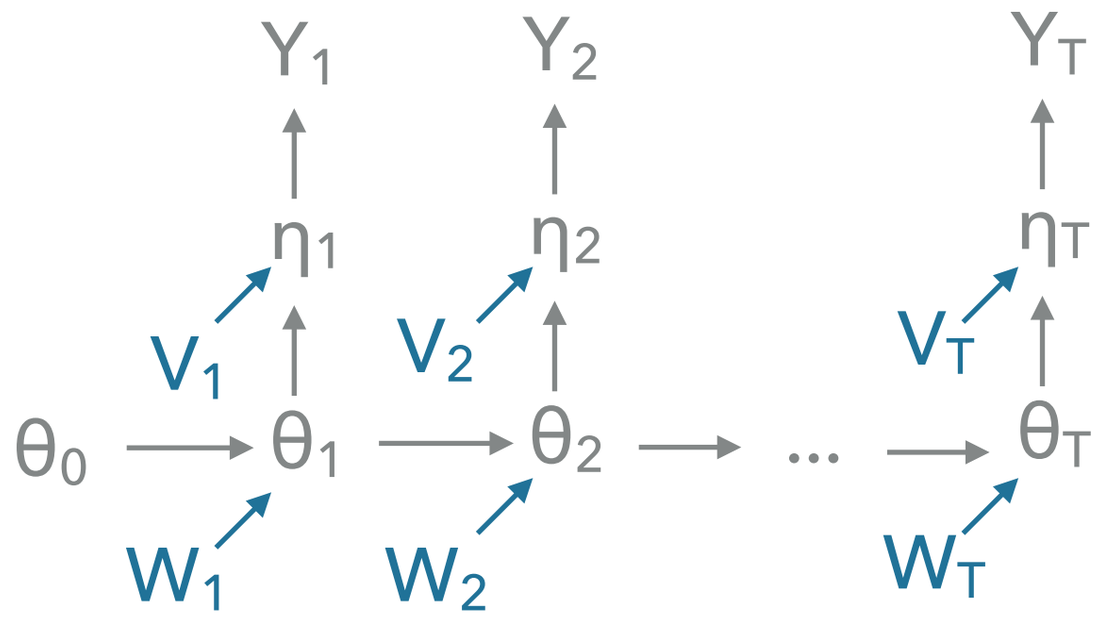
Observed Counts

↑

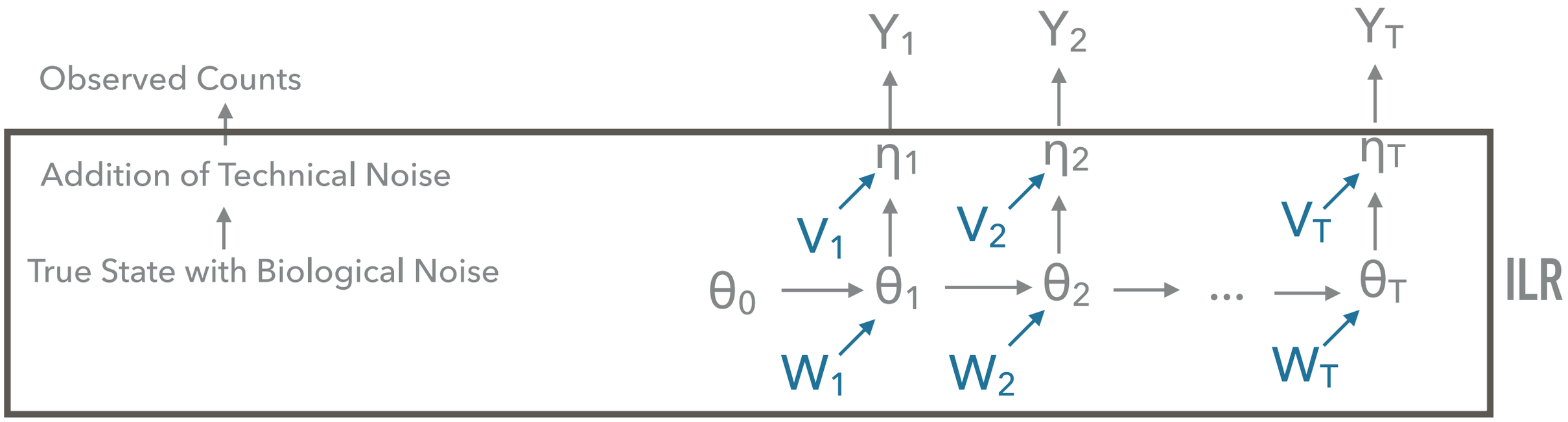
Addition of Technical Noise

↑

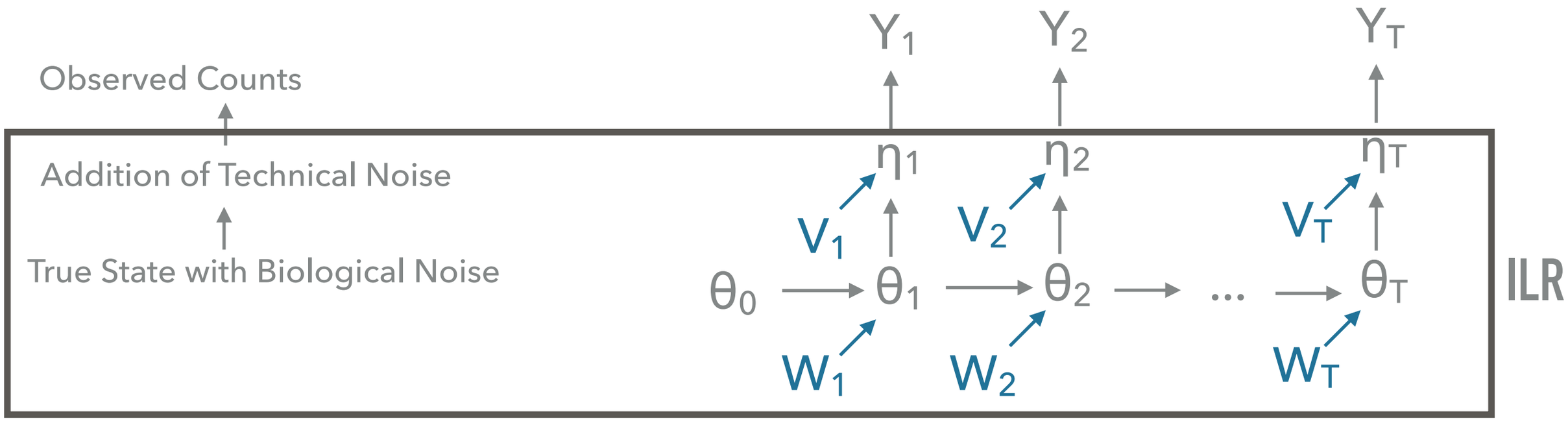
True State with Biological Noise

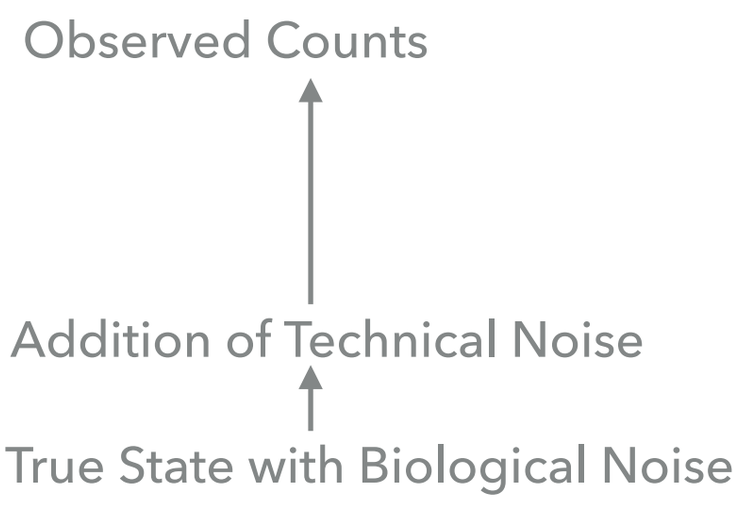


MODELING TIME-EVOLUTION WITH MALLARD



MODELING TIME-EVOLUTION WITH MALLARD





$$\boldsymbol{Y}_t \sim \text{Multinomial}(\boldsymbol{\pi}_t)$$

$$\boldsymbol{\pi}_t = \text{ILR}^{-1}(\boldsymbol{\eta}_t)$$

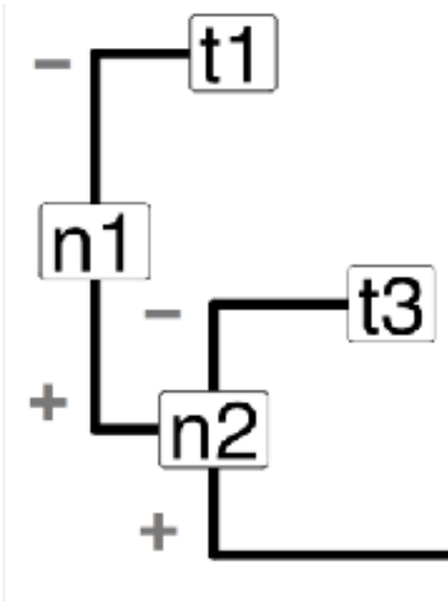
$$\boldsymbol{\eta}_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t \qquad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{V}_t)$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \qquad \boldsymbol{\omega}_t \sim N(0, \boldsymbol{W}_t)$$

A SIMPLE SIMULATION

$$
\begin{aligned}
Y_t &\sim \text{Multinomial}(\pi_t) \\
\pi_t &= \text{ILR}^{-1}(\eta_t) \\
\eta_t &= \mu_t + v_t & v_t &\sim N(0, V) \\
\mu_t &= \mu_{t-1} + \omega_t & \omega_t &\sim N(0, W)
\end{aligned}
$$

A SIMPLE SIMULATION

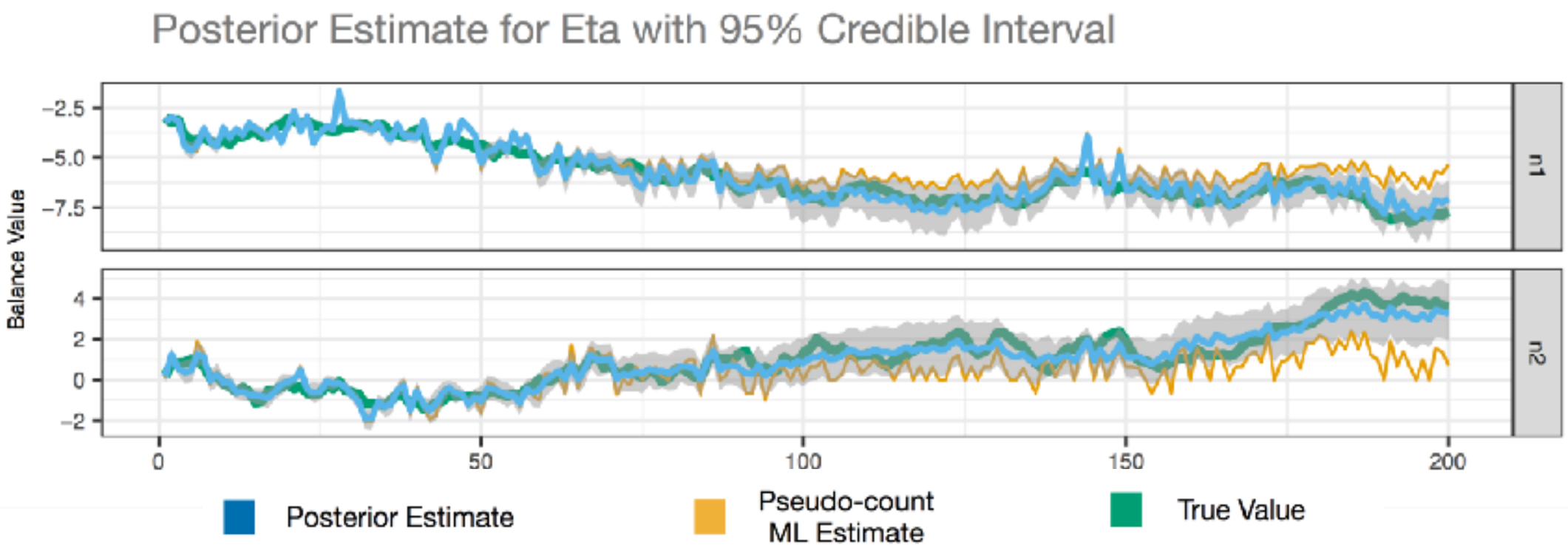


$$Y_t \sim \text{Multinomial}(\pi_t)$$

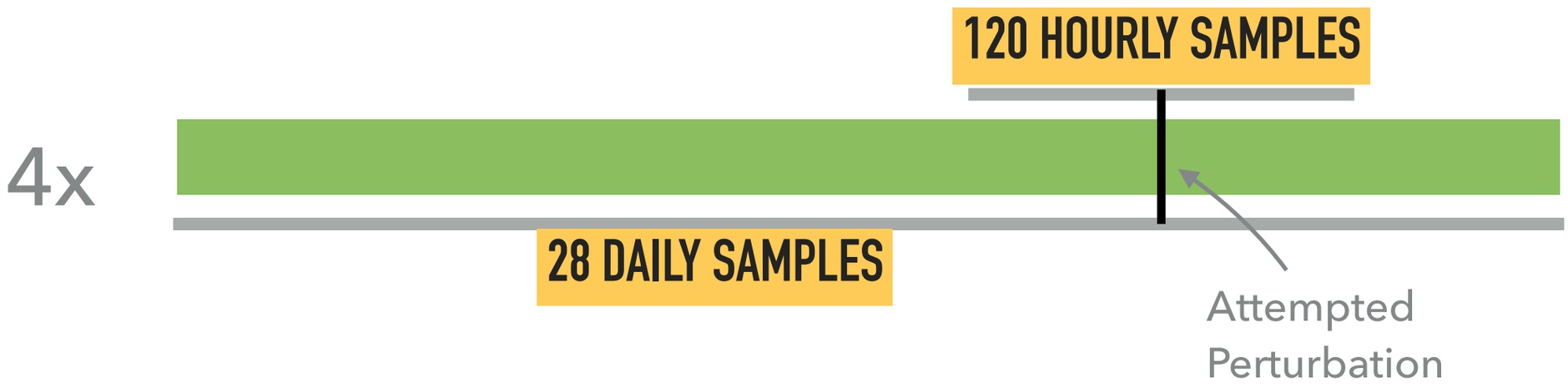
$$\pi_t = \text{ILR}^{-1}(\eta_t)$$

$$\eta_t = \mu_t + v_t \qquad v_t \sim N(0, V)$$

$$\mu_t = \mu_{t-1} + \omega_t \qquad \omega_t \sim N(0, W)$$



LONGITUDINAL MODEL WITH REPEATED OBSERVATIONS

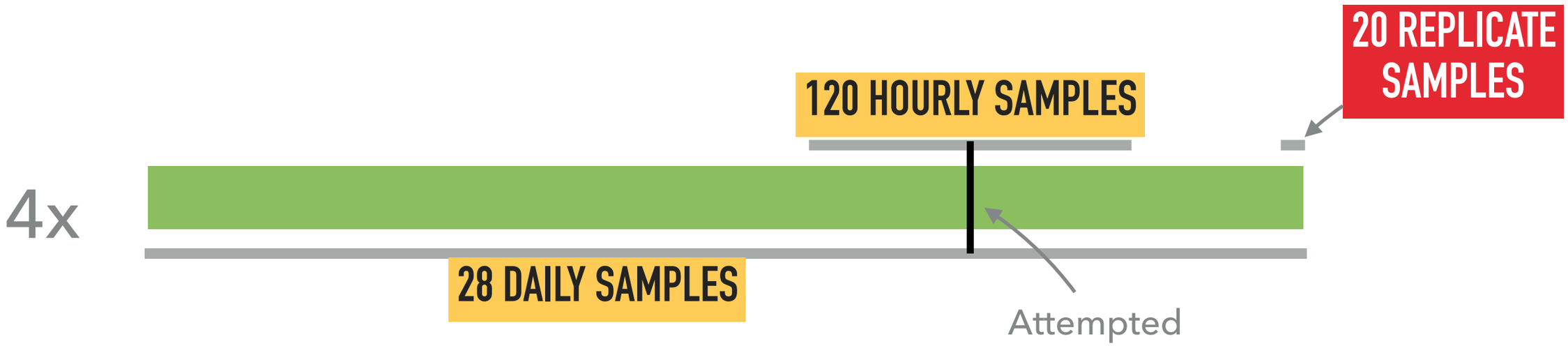


LONGITUDINAL MODEL WITH REPEATED OBSERVATIONS

STANDARD LONGITUDINAL MODEL

$$
\begin{aligned}
Y_t &\sim \text{Multinomial}(\pi_t) \\
\pi_t &= \text{ILR}^{-1}(\eta_t) \\
\eta_t &= \mu_t + v_t & v_t &\sim N(0, V) \\
\mu_t &= \mu_{t-1} + \omega_t & \omega_t &\sim N(0, W)
\end{aligned}
$$

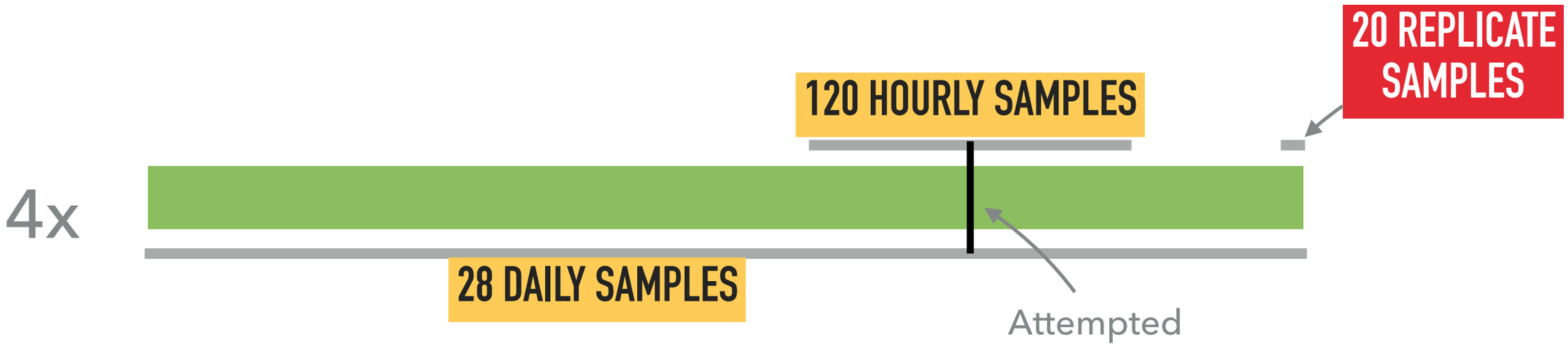
LONGITUDINAL MODEL WITH REPEATED OBSERVATIONS



STANDARD LONGITUDINAL MODEL

$$
\begin{aligned}
Y_t &\sim \text{Multinomial}(\pi_t) \\
\pi_t &= \text{ILR}^{-1}(\eta_t) \\
\eta_t &= \mu_t + v_t & v_t &\sim N(0, V) \\
\mu_t &= \mu_{t-1} + \omega_t & \omega_t &\sim N(0, W)
\end{aligned}
$$

LONGITUDINAL MODEL WITH REPEATED OBSERVATIONS



STANDARD LONGITUDINAL MODEL

$$
\begin{aligned}
Y_t &\sim \text{Multinomial}(\pi_t) \\
\pi_t &= \text{ILR}^{-1}(\eta_t) \\
\eta_t &= \mu_t + v_t \qquad v_t \sim N(0, V) \\
\mu_t &= \mu_{t-1} + \omega_t \qquad \omega_t \sim N(0, W_t)
\end{aligned}
$$

CONDITION TO HANDLE REPLICATES

$$
W_t = \begin{cases} 0 & \text{if } t \text{ is a replicate of } t-1; \\ W & \text{otherwise;} \end{cases}
$$

ESTIMATING "SIGNAL-TO-NOISE" RATIO

Biological Noise to Technical Noise Ratio

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V)}$$

Percent of Noise (Excluding Counting) Attributable to Biology

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V) + \mathrm{Tr}(W)}$$

ESTIMATING "SIGNAL-TO-NOISE" RATIO

Biological Noise to Technical Noise Ratio

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V)}$$

Percent of Noise (Excluding Counting) Attributable to Biology

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V) + \mathrm{Tr}(W)}$$

STANDARD LONGITUDINAL MODEL

$$Y_t \sim \text{Multinomial}(\pi_t)$$
$$\pi_t = \text{ILR}^{-1}(\eta_t)$$
$$\eta_t = \mu_t + \nu_t \qquad \nu_t \sim N(0, V)$$
$$\mu_t = \mu_{t-1} + \omega_t \qquad \omega_t \sim N(0, W_t)$$

CONDITION TO HANDLE REPLICATES

$$W_t = \begin{cases} 0 & \text{if } t \text{ is a replicate of } t-1; \\ W & \text{otherwise;} \end{cases}$$

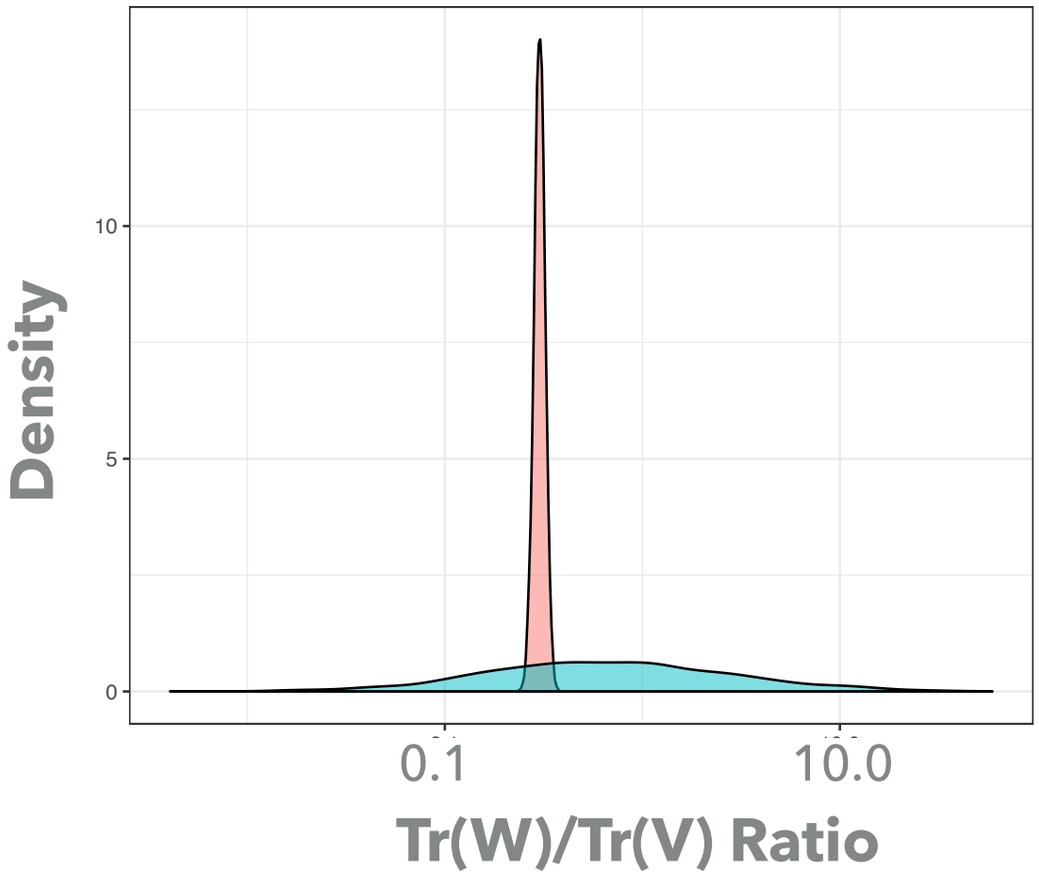
DOMINATED BY TECHNICAL NOISE AT HOURLY INTERVALS

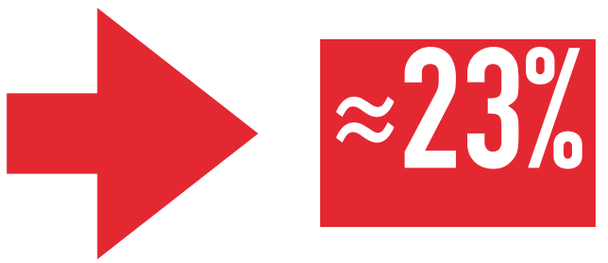
Biological Noise to Technical Noise Ratio

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V)}$$

Percent of Noise (Excluding Counting) Attributable to Biology

$$\frac{\mathrm{Tr}(W)}{\mathrm{Tr}(V) + \mathrm{Tr}(W)}$$





BIOLOGICAL AND TECHNICAL VARIATION CAN HAVE DIFFERENT SHAPES