Uncertainty propagation with heterogenous data.

Susan Holmes
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford University

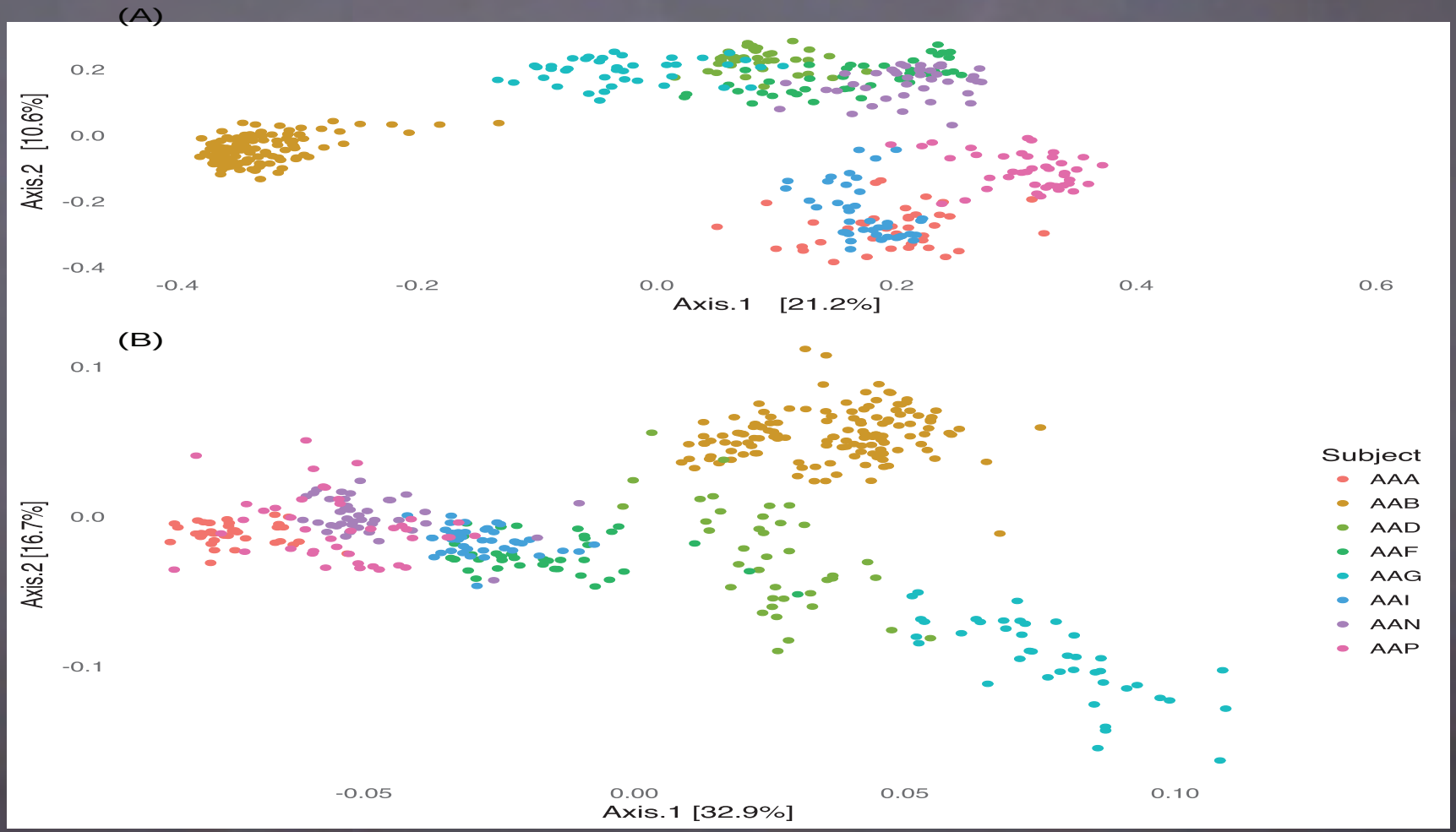
Toulouse, September 2, 2019

Part I

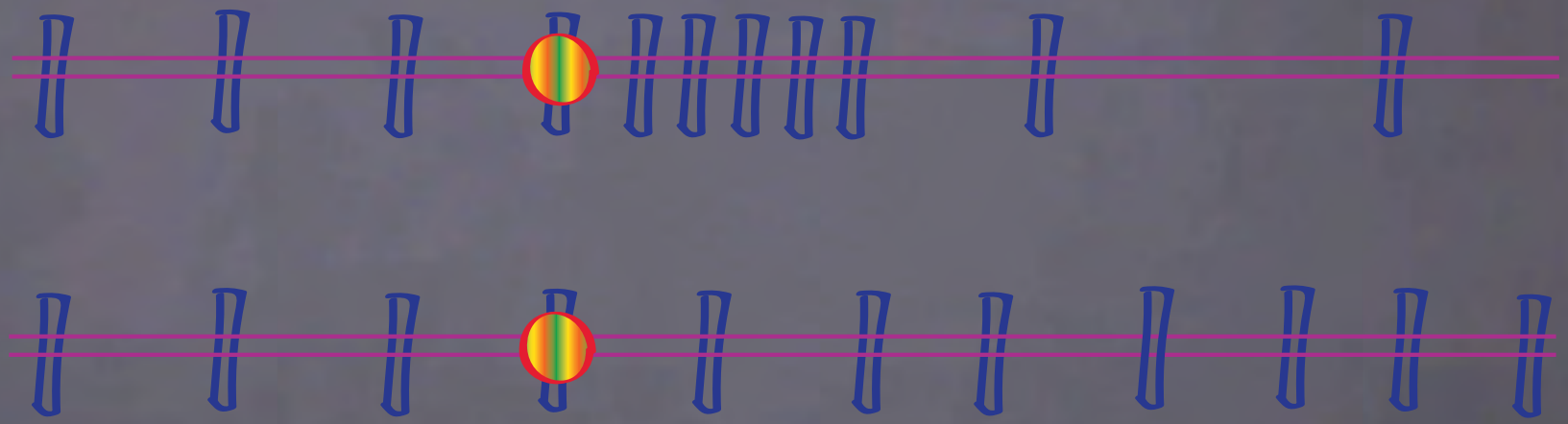
Experimental Design and data

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0	D	791	0	79	108	11
1	D	1616	0	1413	192	31
2	D	1323	0	915	165	23
3	D	1846	0	1366	170	31
4	D	2314	0	689	135	26
5	D	2244	0	776	310	175
6	D	1652	0	609	235	181

Subject to Subject variation is largest source of variation



Not equally distant time points.



Between point variation should be equal.

See Peter Diggle's text : Analysis of Longitudinal Data, 2002.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV)

generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p_{tt}} = (p_1, p_2, \ldots p_J) \qquad \text{For } J \ ASV's$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \ldots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV)

generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p_{tt}} = (p_1, p_2, \dots p_J) \qquad \text{For } J \ ASV's$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \dots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors. We need to quantify the uncertainty we have on the parameters.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV)

generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p_{tt}} = (p_1, p_2, \ldots p_J) \qquad \text{For } J \ ASV's$$

$$\mathbf{p_{ctl}} = (p_1, p_2, \ldots p_J) \qquad \Delta = \text{diff}(\mathbf{p_{tt}} - \mathbf{p_{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.

Models for noise: hierarchical Gamma-Poisson: we know how to transform the data to stabilize the variance (Delta-method).

McMurdie and Holmes (2014) "Waste Not, Want Not: Why

Read data are counts, the data are not compositional.

We do not summarize them to ratios or “relative abundance”.

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (`decontam`).
- Estimating depth bias requires read numbers.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Transform the data to equalize the variance.

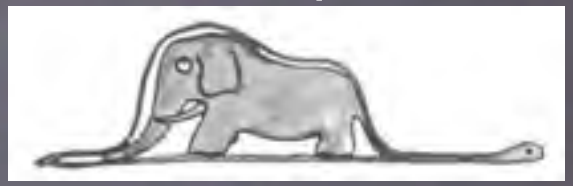
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.

hidden variables.

Paths in thinking about these heterogeneous systems

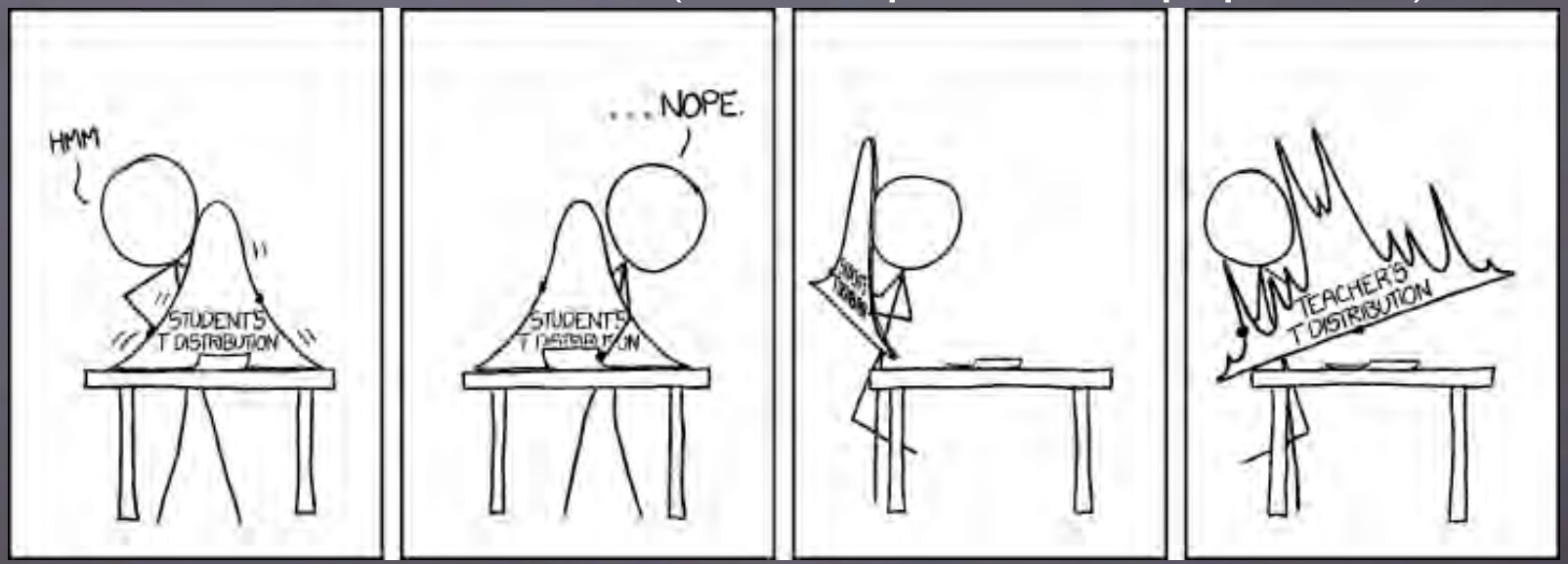
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

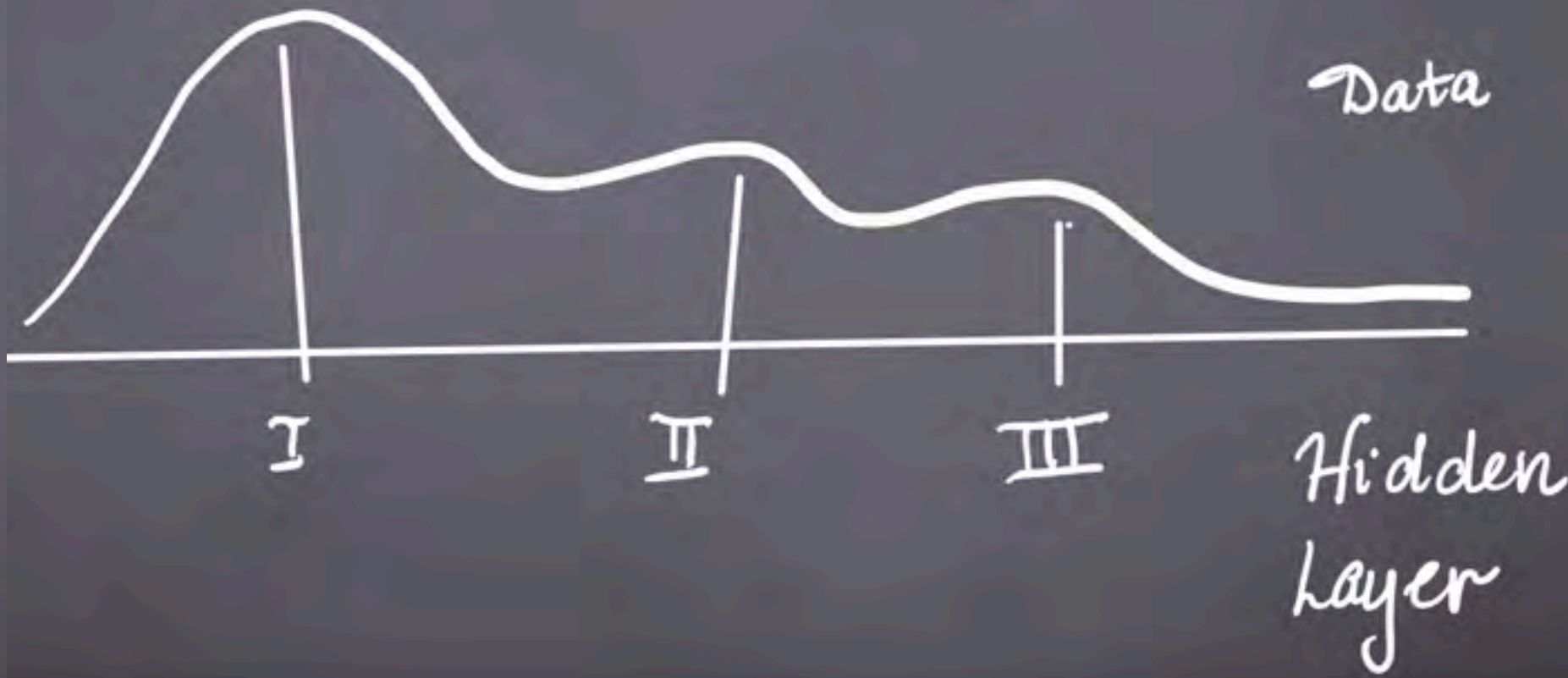
Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).



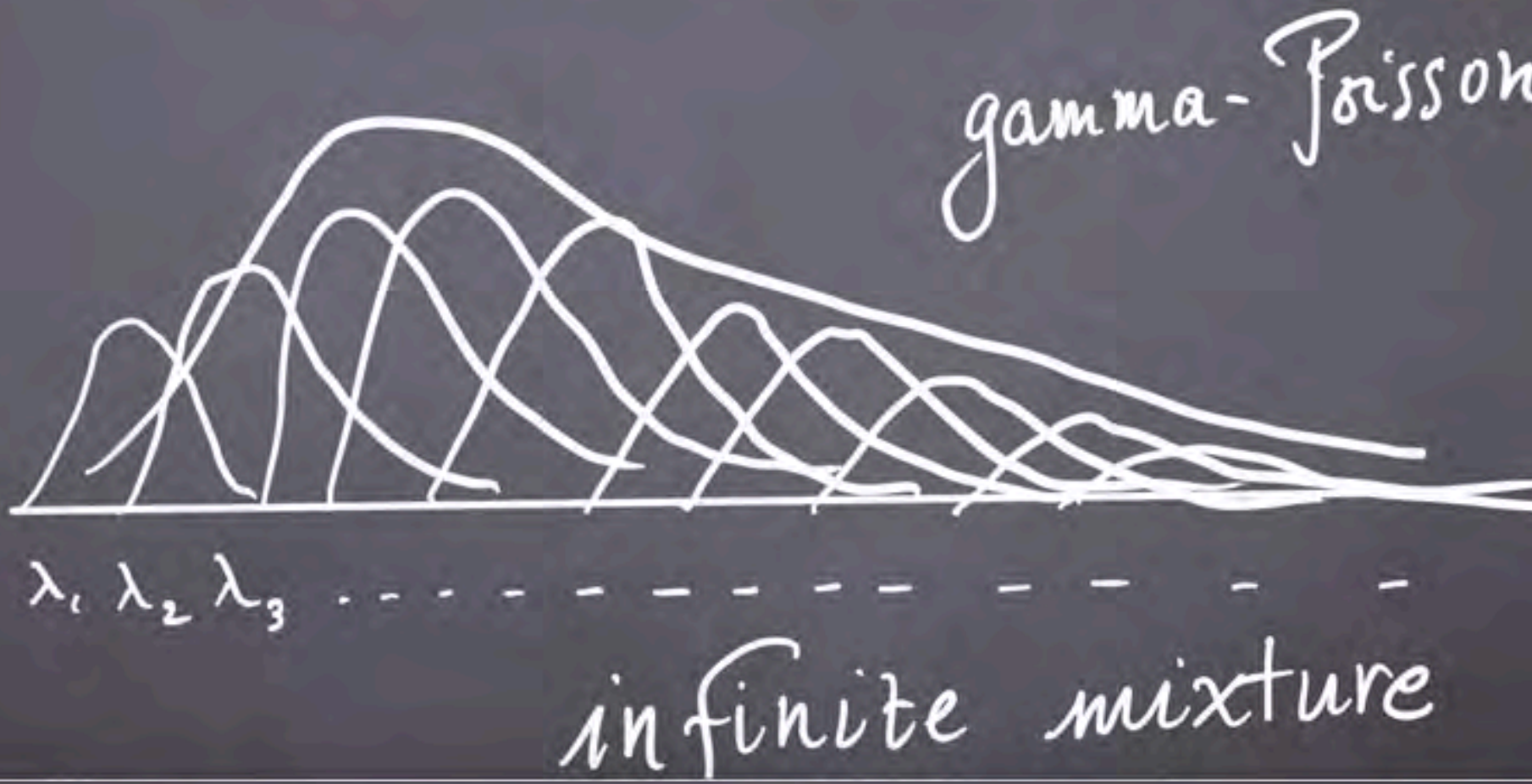
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation



Part II

Models for Microbial Communities over time.

Pregnancy data: perturbation, stability and preterm birth

A case-control study of 49 pregnant women:

- 15 delivered preterm.
- From 40 of these women: 3,766 specimens collected weekly during gestation, and monthly after delivery.
- Sites:vagina, distal gut, saliva, and tooth/gum.
- 9 women: validation set collected after the first study was complete.

Methods used: variance stabilization through negative binomial, testing perturbations through linear mixed-effects modeling. Preterm prediction through medoid-based clustering and simple Markov chain. Provided: Simple community temporal trends, community structure, and vaginal community state transitions.

Attention to detail

- Careful noise models (`dada2`) and variance stabilization (DESeq2, vsn, voom).
- Random effects, mixed models.
- Finite State Markov chains.
- Differential abundance testing provides biomarkers for preterm birth.

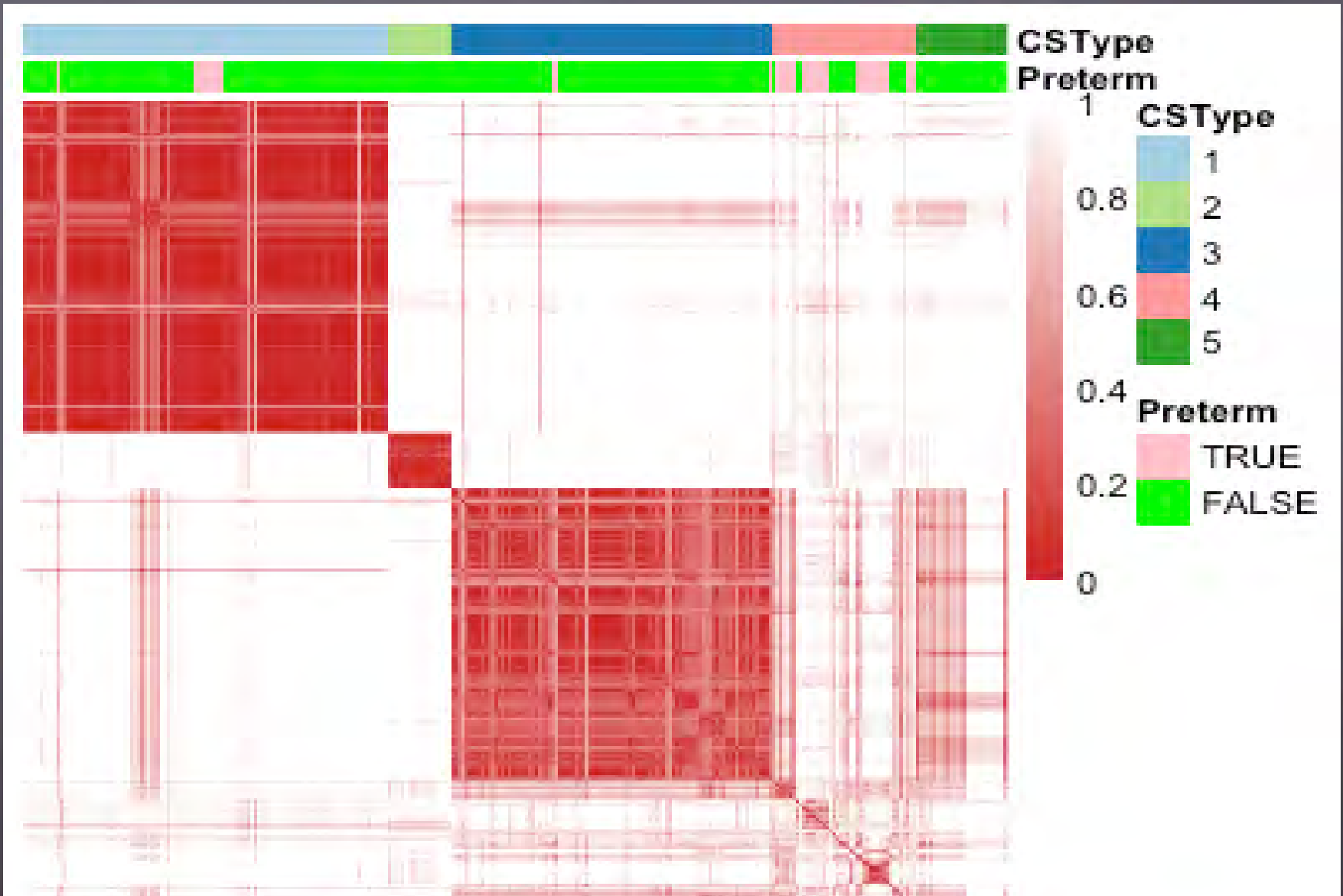
DiGiulio DB, Callahan BJ, McMurdie PJ, ... & Holmes, SP and Relman, DA
Temporal and spatial variation of the human microbiota during pregnancy. PNAS, 2015,112(35):11060-5.

Co-occurrence networks

Dual networks:

- Edges are created between taxa if in more than a certain proportion of samples share that taxa.
  This can be seen as a geometric graph with the distance being the Jaccard distance.
- Edges are created between samples if they share more than a certain proportion of taxa in common.

Communities of bacteria organize into 5 different types

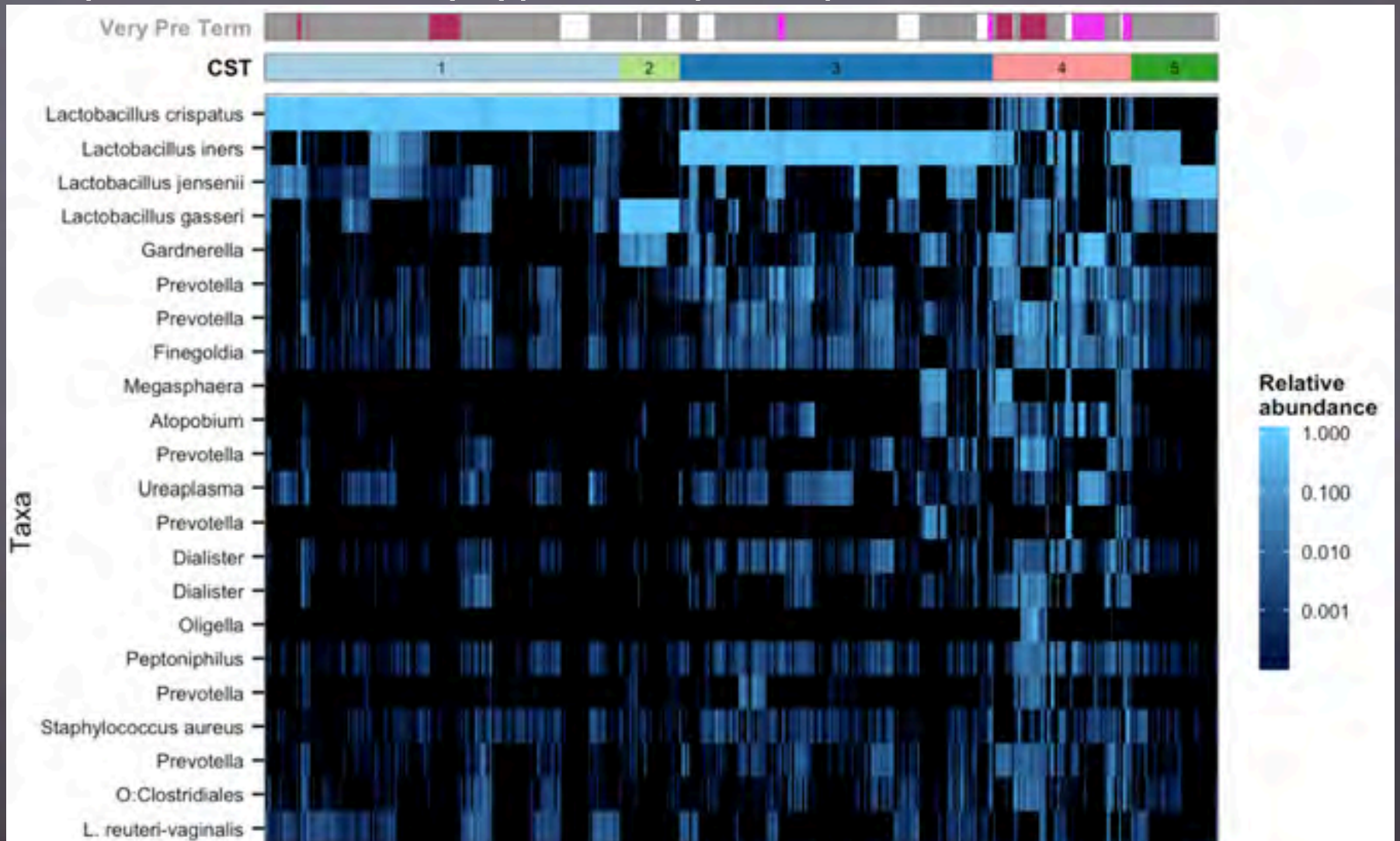


Questions asked?

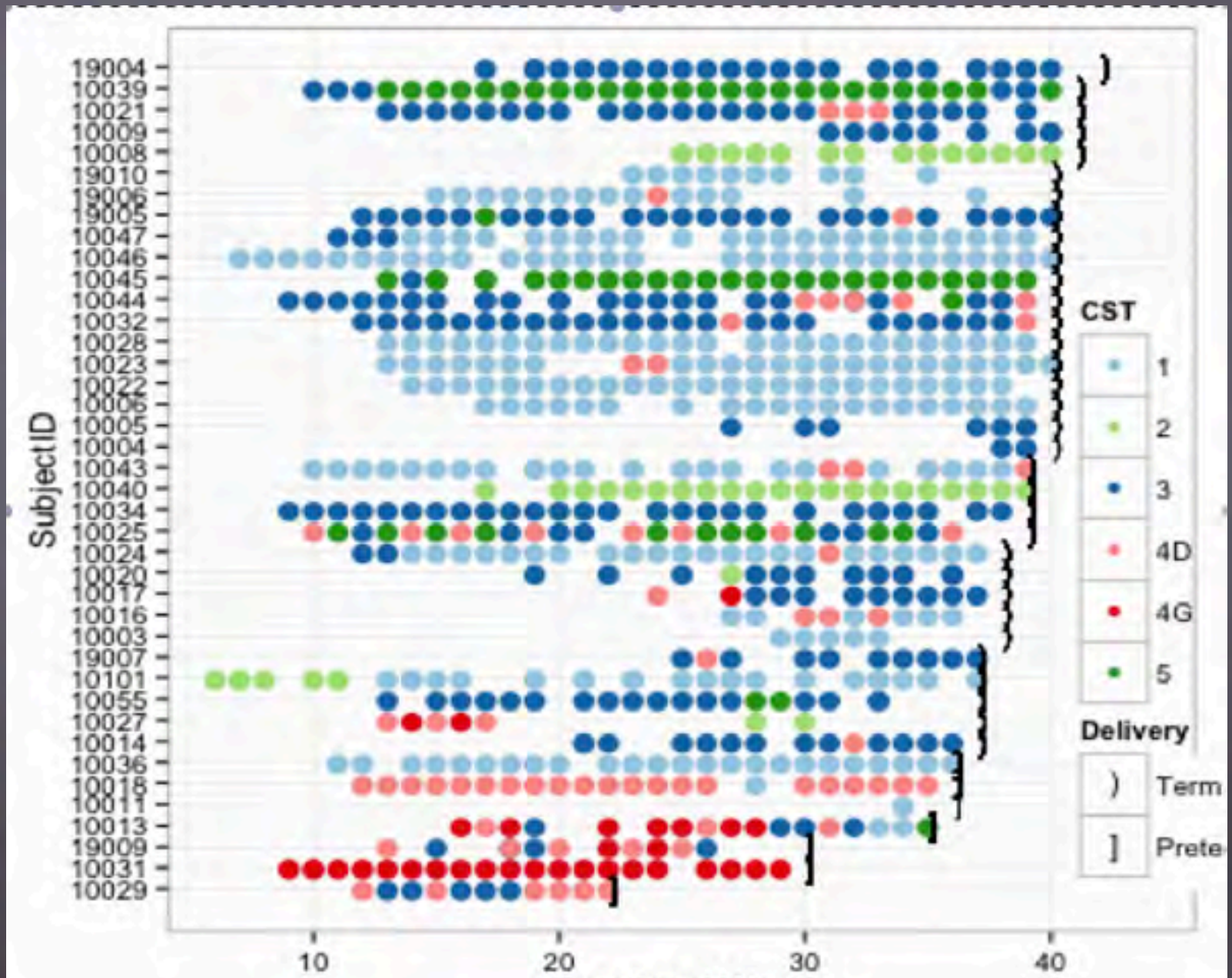
- Are the community state types the same as seen in previous studies?
- How stable are the communities within each individual during pregnancy?
- What alterations of the vaginal microbiome predict preterm birth?
- How early do these alterations occur?

Previously known Microbial Community State Types: Latent categorical variable.

Samples into community types and species patterns associated.

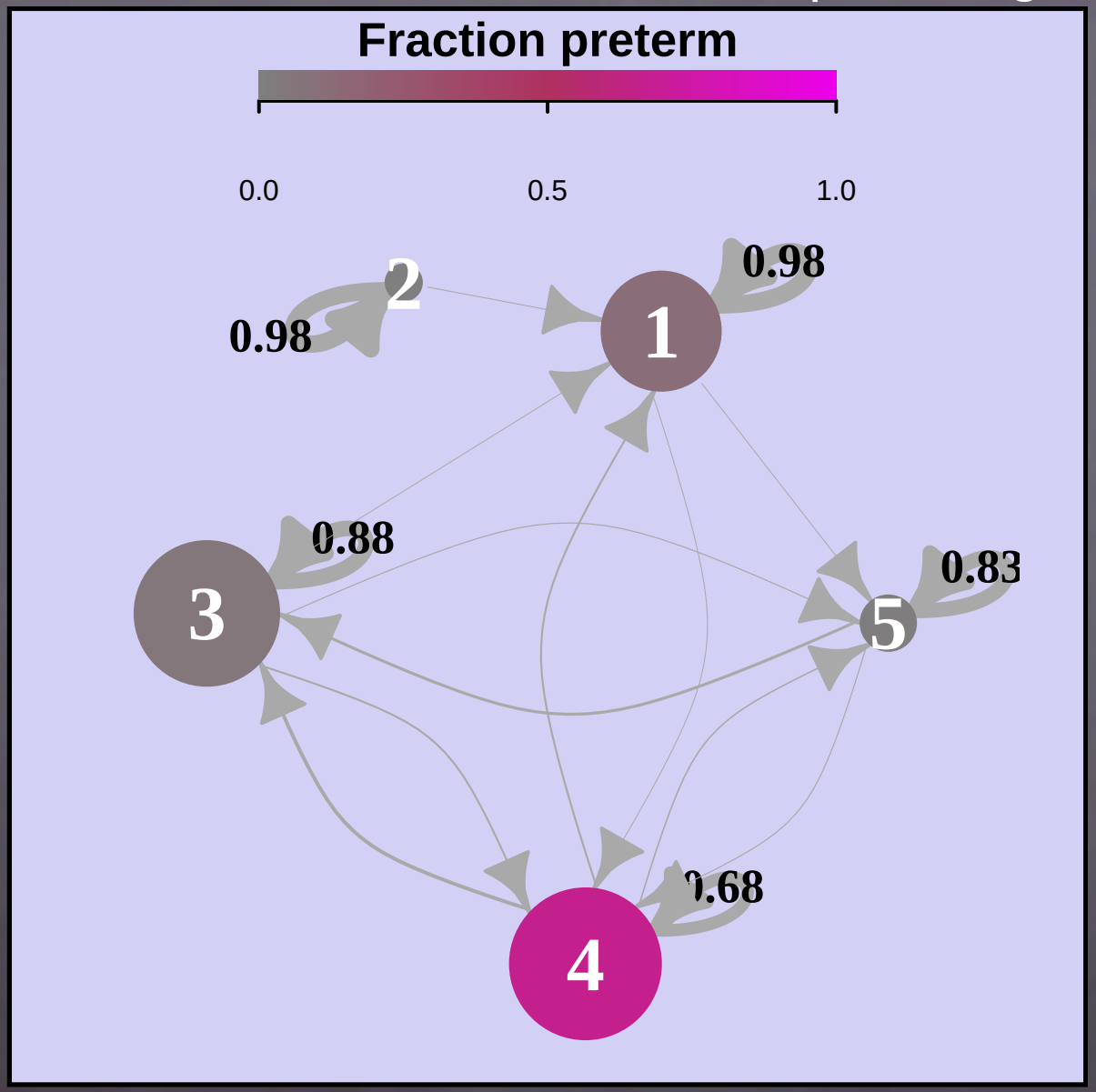


Longitudinal Analyses



Markov Chain Model

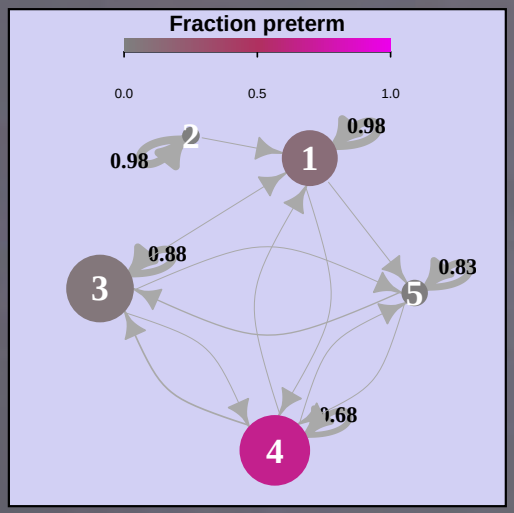
Transitions between states, as in simple ecological models.

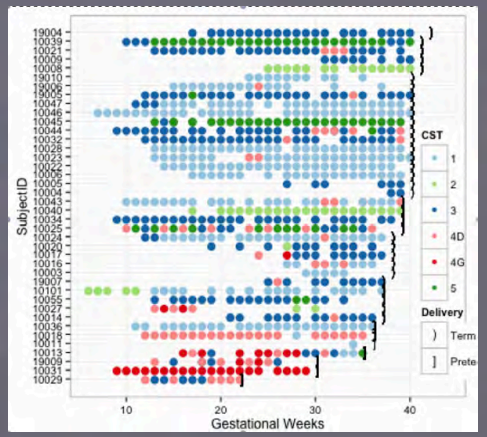


Conclusions for this study

- Microbiota community and diversity stable during pregnancy.
- Prevalence of a Lactobacillus-poor vaginal community state type (CST 4) was inversely correlated with gestational age at delivery ($p=0.0039$).
  Risk for preterm birth was more pronounced for subjects with CST 4 accompanied by elevated Gardnerella or Ureaplasma abundances.
- Finding validated with a separate diagnostic set of 246 vaginal specimens from nine women (four of whom delivered preterm).

Illustration through Analyses

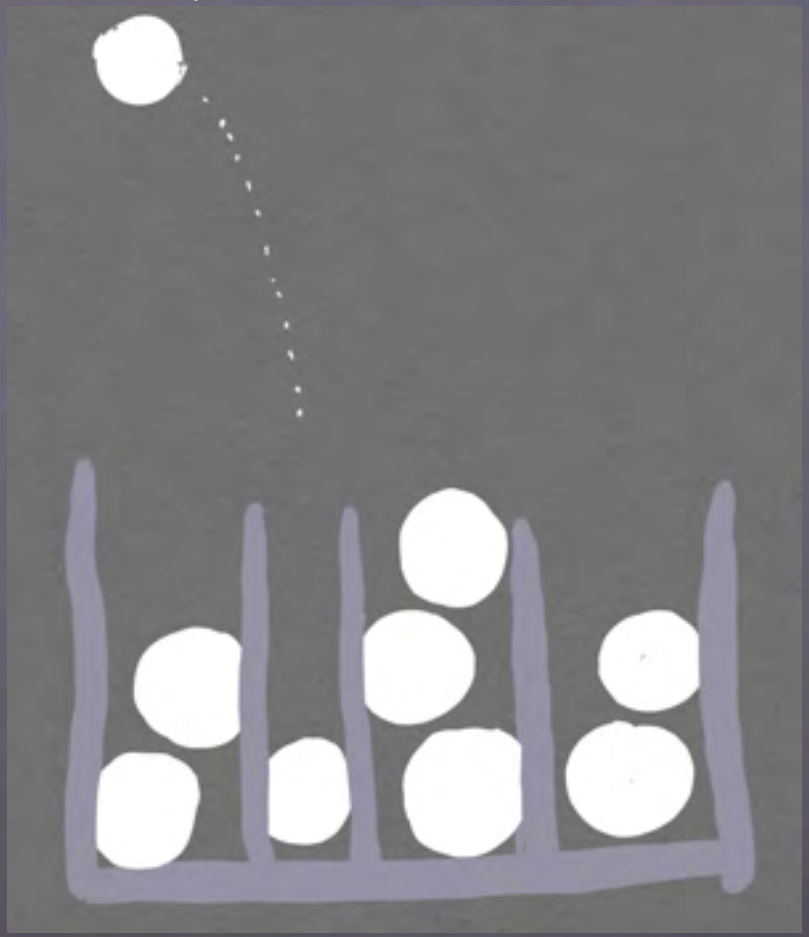




- Delivery Perturbation
- Preterm Prediction
- Stability

Part III

The Dirichlet for the multinomial



General Ideas about the multinomial

- Balls in boxes, not necessarily the same size.
- The number of balls is the number of reads, the boxes are the ASVs.
- Multinomial model gives the probability of seeing say (4,2,3,1) if the probabilities of the four boxes are $p_1 = 0.3, p_2 = 0.2, p_3 = 0.4, p_4 = 0.1$ this number is:

```
> dmultinom(c(4,2,3,1),prob=c(0.3,0.2,0.4,0.1))
[1] 0.02612736
```

- Apart from the fact that if a lot of balls fall in the first box there will be less balls for the other boxes, the boxes' contents are independent: that is BAD.

Dirichlet

Make the p's vary randomly.
Hierarchical Model:
$ps \sim Dirichlet(\alpha, \alpha, \alpha, \alpha)$
Uniform on the simplex (four cornered pyramid).

```
x <- round(gtools::rdirichlet(5, c(1,1,1,1) ),2)
> x
     [,1] [,2] [,3] [,4]
[1,] 0.06 0.50 0.08 0.36
[2,] 0.20 0.57 0.18 0.05
[3,] 0.07 0.20 0.55 0.18
[4,] 0.57 0.04 0.00 0.39
[5,] 0.02 0.16 0.27 0.55
```

Multinomial needs to be modified

Multivariate dependencies in bacterial communities

Data depart from a multinomial distribution within each row:

- Some taxa are quasi-exclusive (*Lactobacillus crispatus* and *Gardnerella*).
- Co-occurrence through syntrophy, in which a molecular hydrogen-consuming species (typically a methanogen, like *Methanobrevibacter smithii* in the human gut) enhances the growth of a molecular hydrogen-producing species (any of a number of secondary fermenters in the gut).
- In the mouth (subgingival crevice), where in cases of moderate to severe periodontitis, a methanogen (Methanobrevibacter oralis) is always found with a syntrophic partner.
- There are not a finite number of taxa a priori, taxa evolve, some are sample-specific.

Part IV

Interpretability: latent variables and topic analysis

Discrete/disconnected Community state types are rare

Each sample is assigned to only one type of community.

Need a more nuanced model: mixtures.

Mixture models

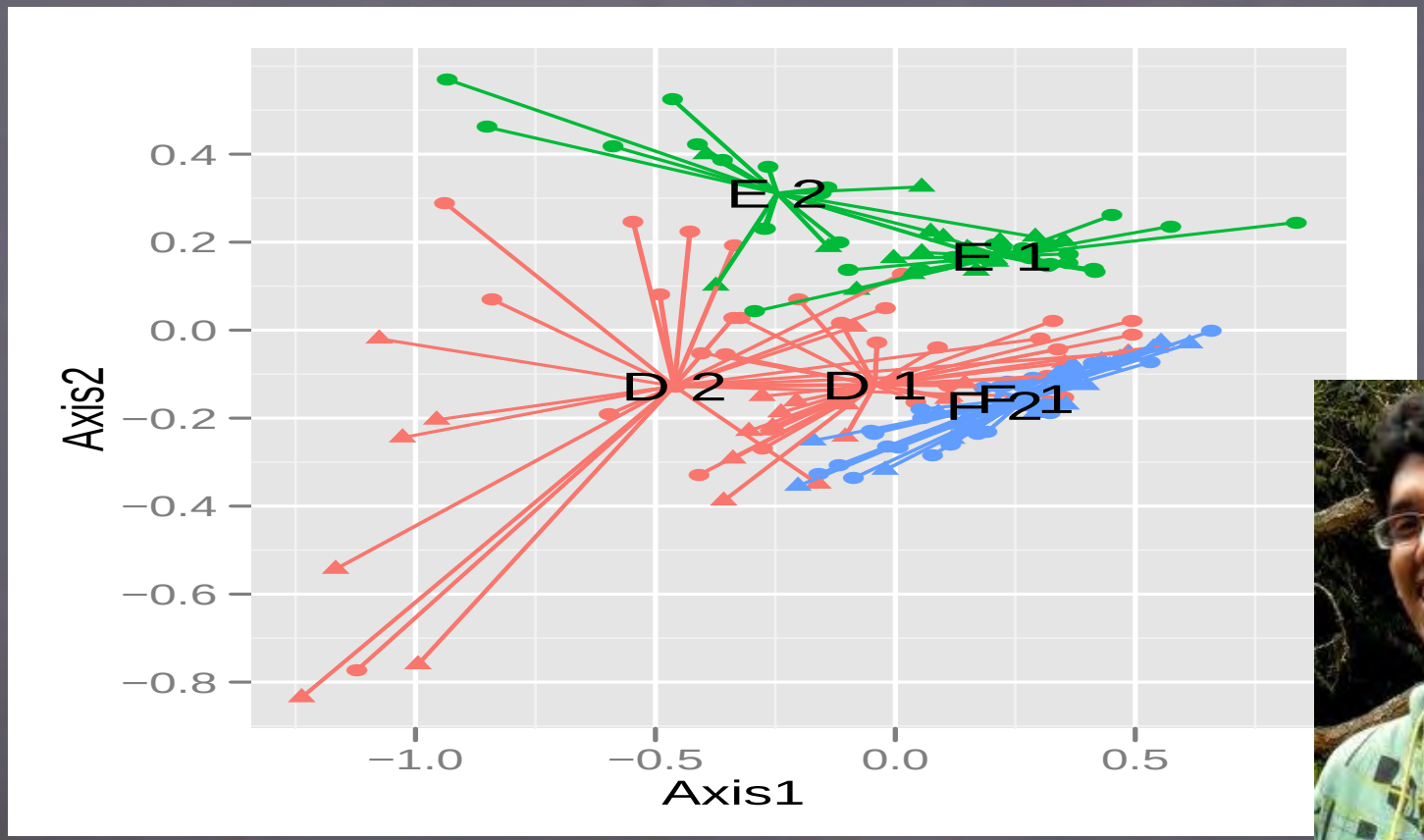
- In clustering and hidden discrete categorical variables, every sample belonged to a community state type.
- In a topic mixture model, every sample can be composed of several topics.

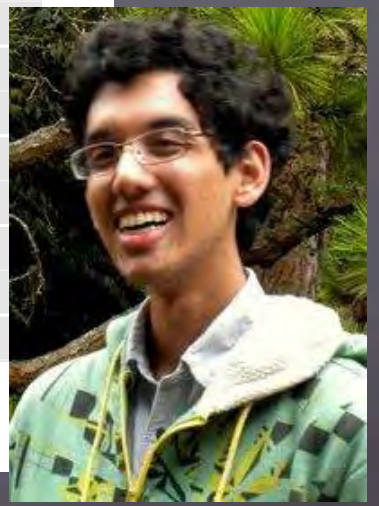
Most useful parallel: natural language processing.

Generative model

- Pick topics at random among a certain number of topics.
- Each topic corresponds to a probability distribution for many words.
- Pick a word at random according to the chosen topic.

How to understand the the taxa involved in the perturbation?





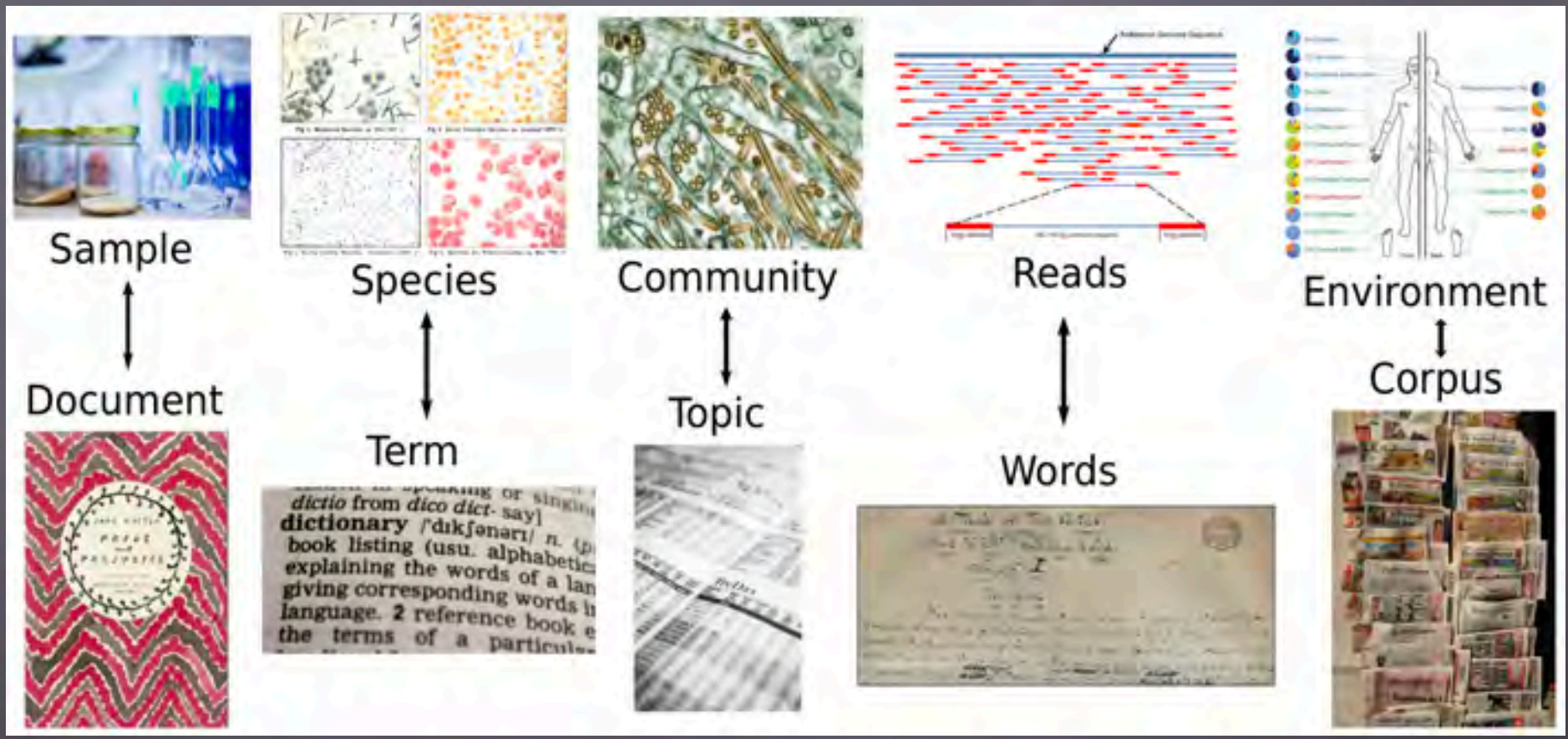
Biostatistics, 2018,

Kris Sankaran

Latent Variable Modeling for the Microbiome.

Kris Sankaran's Topic Page

Parallel between topic and community analyses



Credit: Kris Sankaran

Parallel between topic and community analyses

index	book	elizabeth	darcy	bennet	miss	jane	bingley	time
0	P & P	0	0	4	0	1	3	0
1	P & P	1	0	5	0	1	4	0
2	P & P	0	0	6	0	0	5	1
3	P & P	1	4	5	1	0	9	1
4	P & P	3	3	5	4	4	5	3
5	P & P	3	0	0	2	1	6	1
6	P & P	0	6	6	7	1	5	1

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0	D	791	0	79	108	11
1	D	1616	0	1413	192	31
2	D	1323	0	915	165	23
3	D	1846	0	1366	170	31
4	D	2314	0	689	135	26
5	D	2244	0	776	310	175
6	D	1652	0	609	235	181

Statistical Model

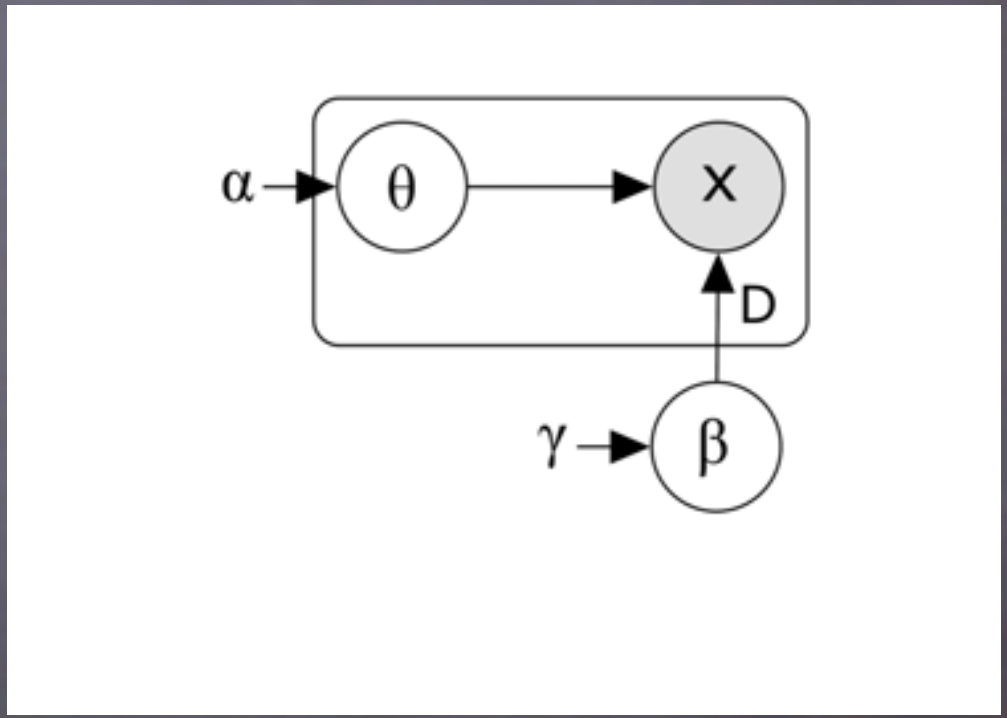
Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling.

It assumes samples have mixed memberships across topics.
(See Pritchard et. al 2000, Blei et. al. 2003)

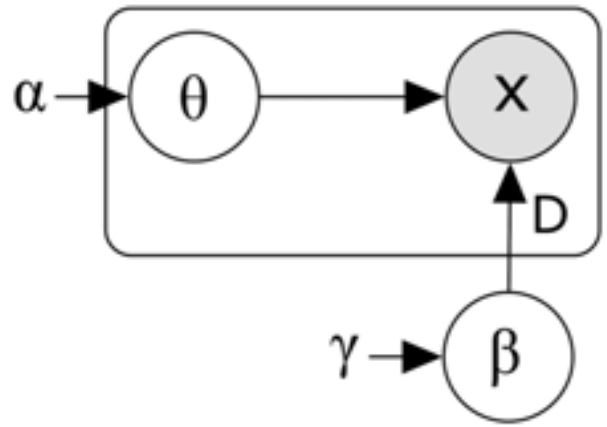
Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling.

Observed microbiomes $\sim$ mixtures of underlying community types.

Statistical Model



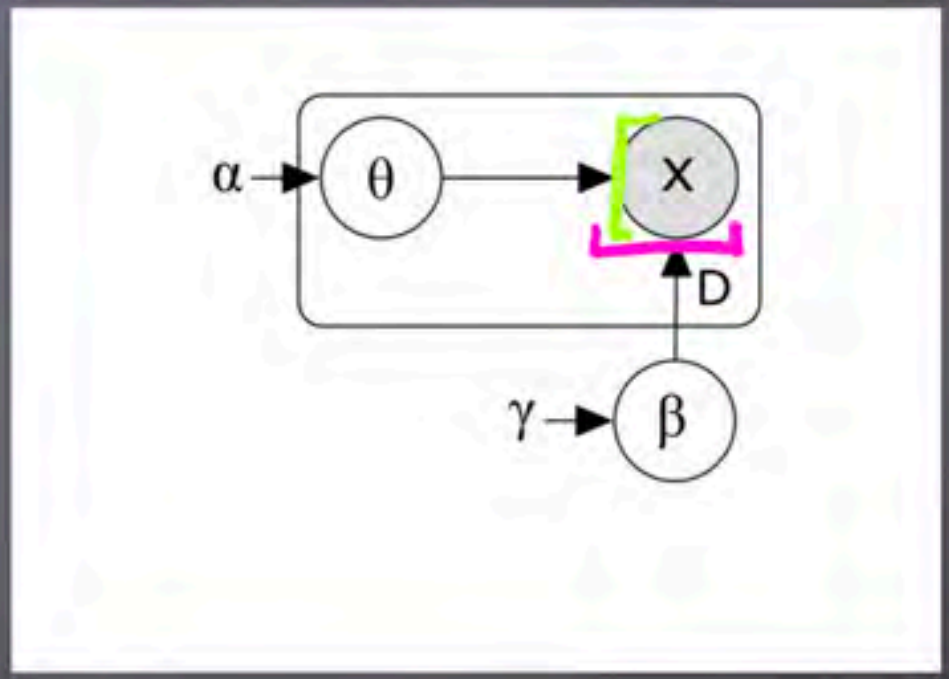
Statistical Model



Statistical Model

samples layer observa. { sample 1 ⋯ ⋮ ⋮ sample n ⋯ } rows (X)

hidden layer for taxa

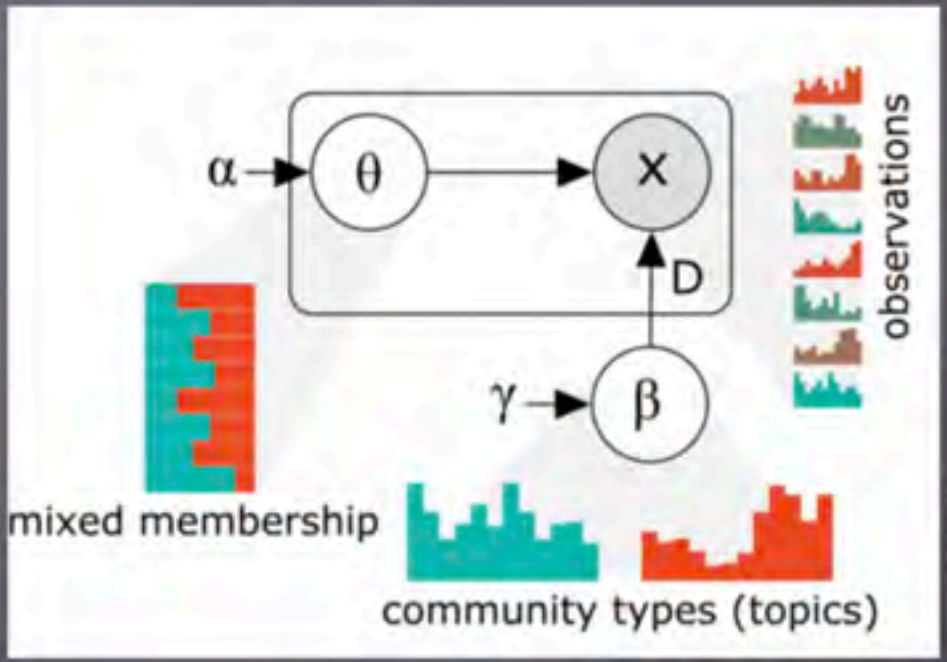


D Documents
K communities or topics

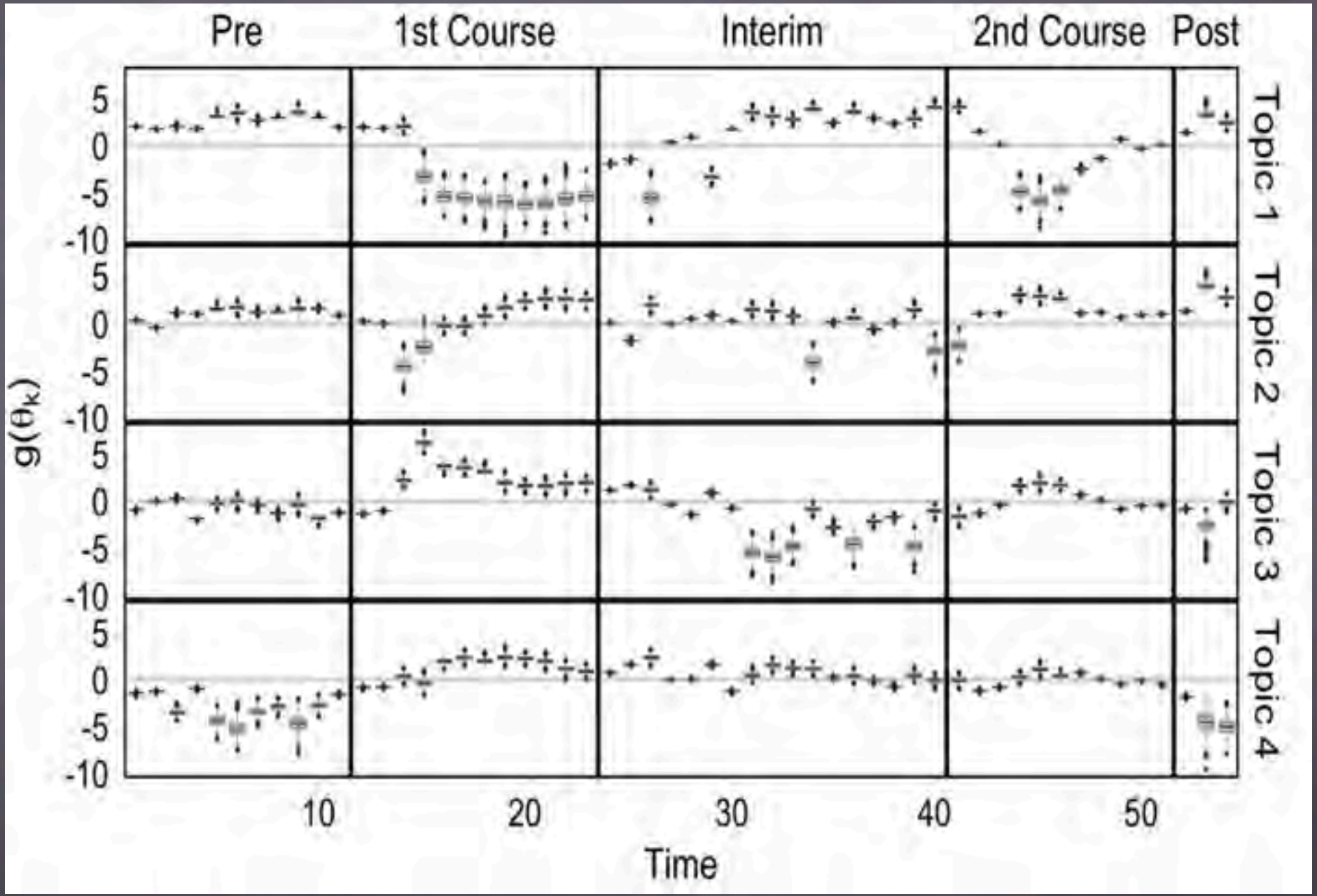
Statistical Model

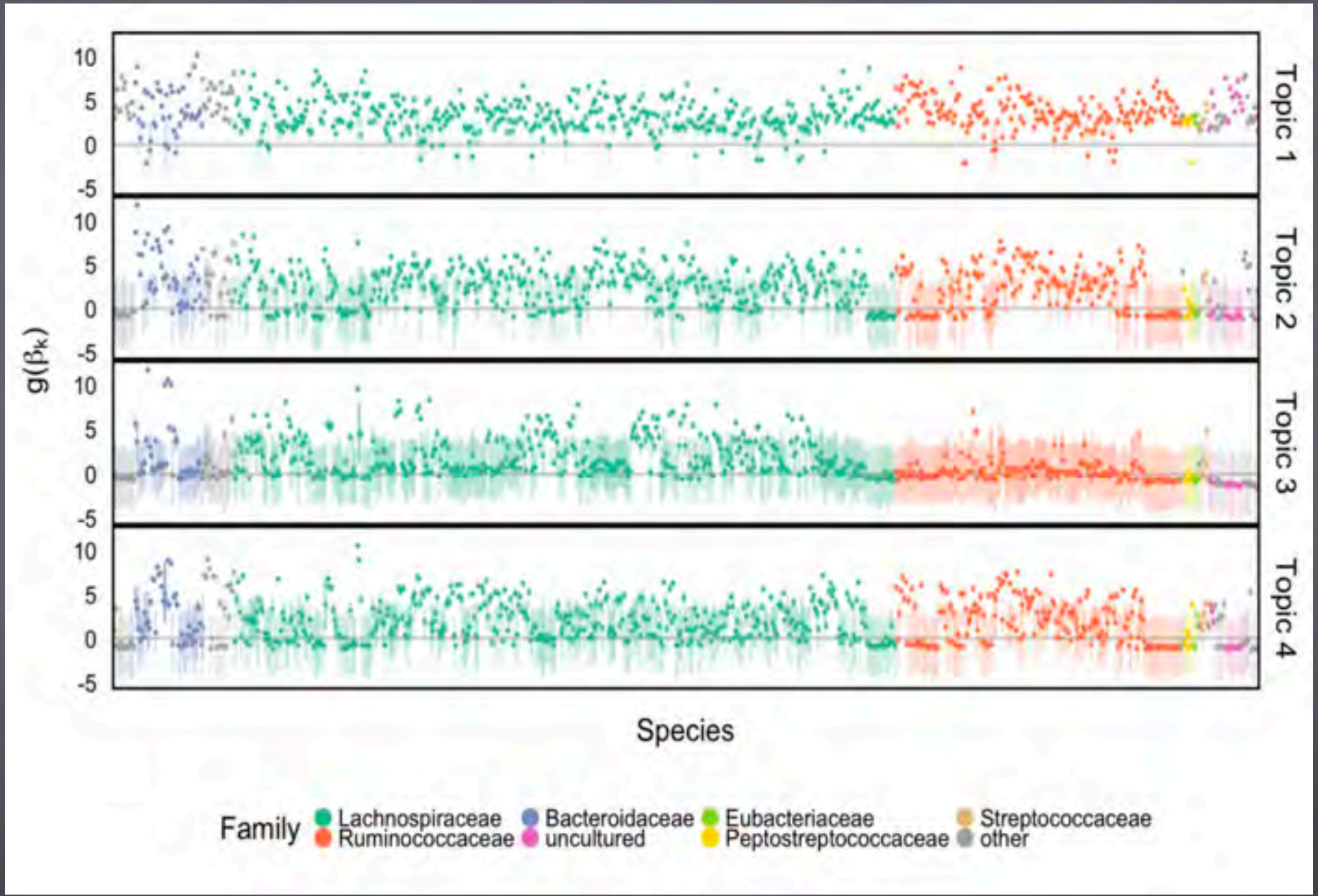
$$x_{d\cdot} \mid \beta \sim Mult(N_d, B\theta_d)$$

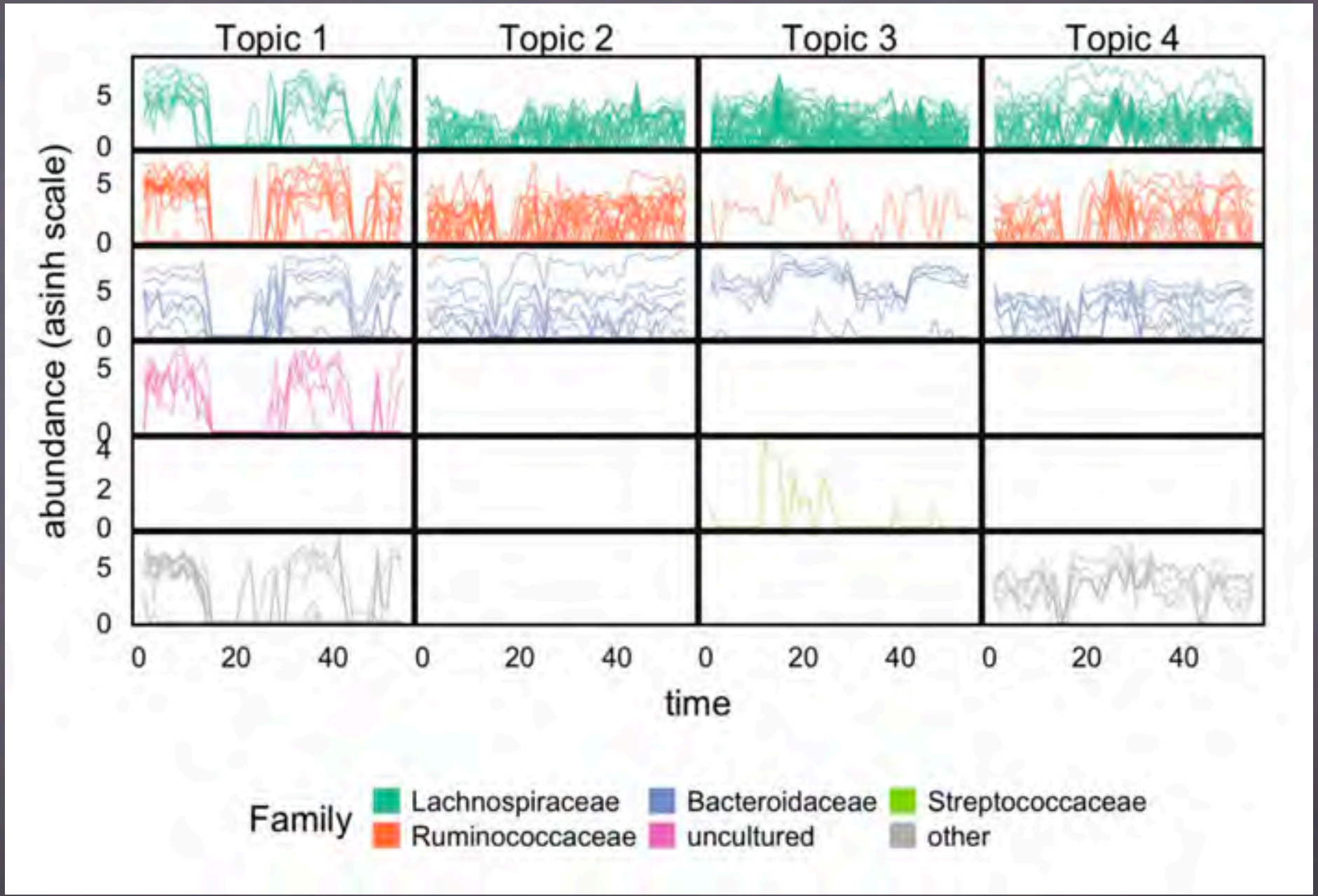
$$\theta_d \sim Dir(\alpha)$$
$$d = 1 \ldots D$$



$$\beta_k \sim Dir(\gamma), \quad k = 1 \ldots K$$

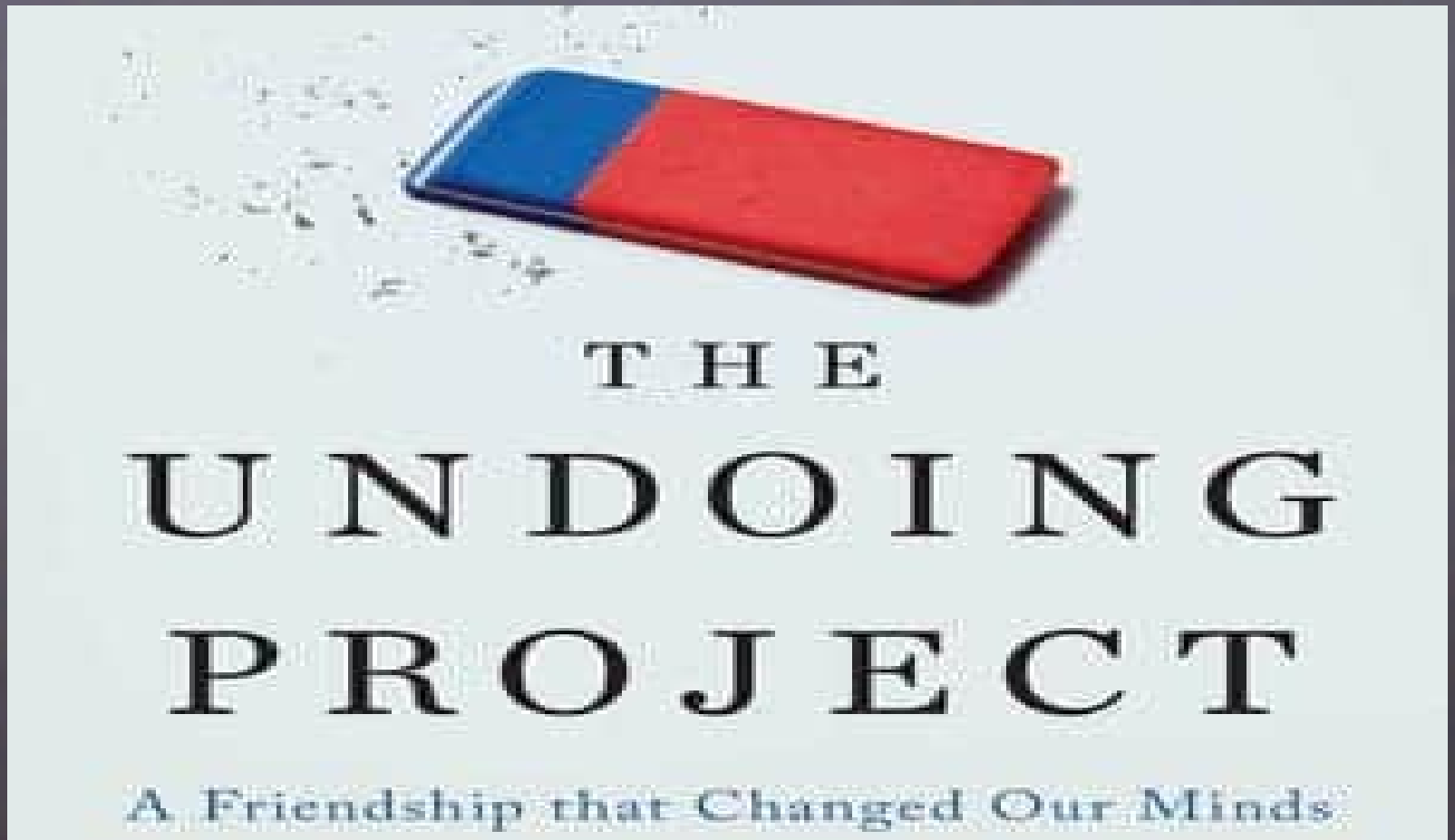

Pre
1st Course
Interim
2nd Course
Post
Topic 1
Topic 2
Topic 3
Topic 4
g(θk)
5
0
-5
-10
10
20
30
40
50
Time


Topic 1
Topic 2
Topic 3
Topic 4
10
5
0
-5
g(βk)
Species
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Eubacteriaceae
Peptostreptococcaceae
Streptococcaceae
other


Topic 1
Topic 2
Topic 3
Topic 4
abundance (asinh scale)
5
0
5
0
5
0
5
0
4
2
0
5
0
0
20
40
time
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Streptococcaceae
other

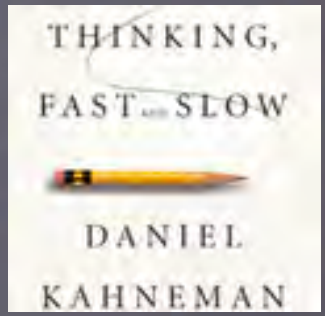
Part V

Distances cannot provide all the information



But distances are not everything....remember the baseline

Amos Tversky and Danny Kahnneman



Heuristics and Biases, more particularly the representativeness heuristic.

Heuristics are described as ”judgmental shortcuts that generally get us where we need to go - and quickly - but at the cost of occasionally sending us off course.”

Heuristics are useful because they use effort-reduction and simplification in decision-making.

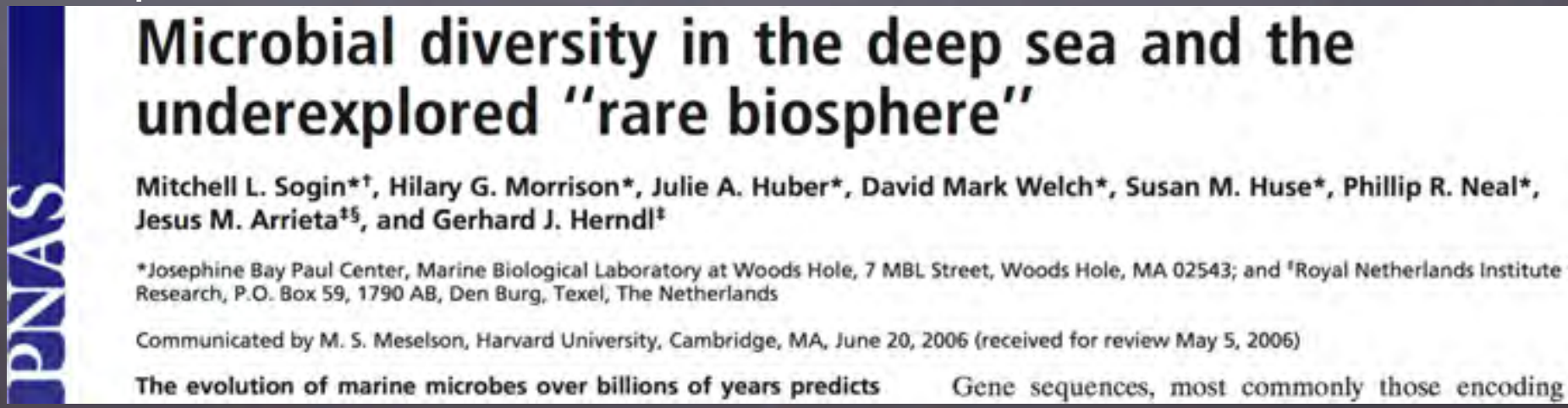
For representativenes of an event, similarity or a small distance is not enough, the baseline frequencies (ie probability) are essential to conclude.

We need careful realistic probability models for treespace, no real data has ever been uniform, no multivariate data is ever multivariate normal.

Diversities in the microbiome depend on the number of taxa

- α-diversity: Number of 'species'-taxa in a biological sample (from one location).
- β-diversity: Differentiation in diversity among different samples from different locations.

Extremely sensitive to noise.
Fake species:



PNAS

Microbial diversity in the deep sea and the underexplored "rare biosphere"

Mitchell L. Sogin*†, Hilary G. Morrison*, Julie A. Huber*, David Mark Welch*, Susan M. Huse*, Phillip R. Neal*, Jesus M. Arrieta‡§, and Gerhard J. Herndl‡

*Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and ‡Royal Netherlands Institute Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

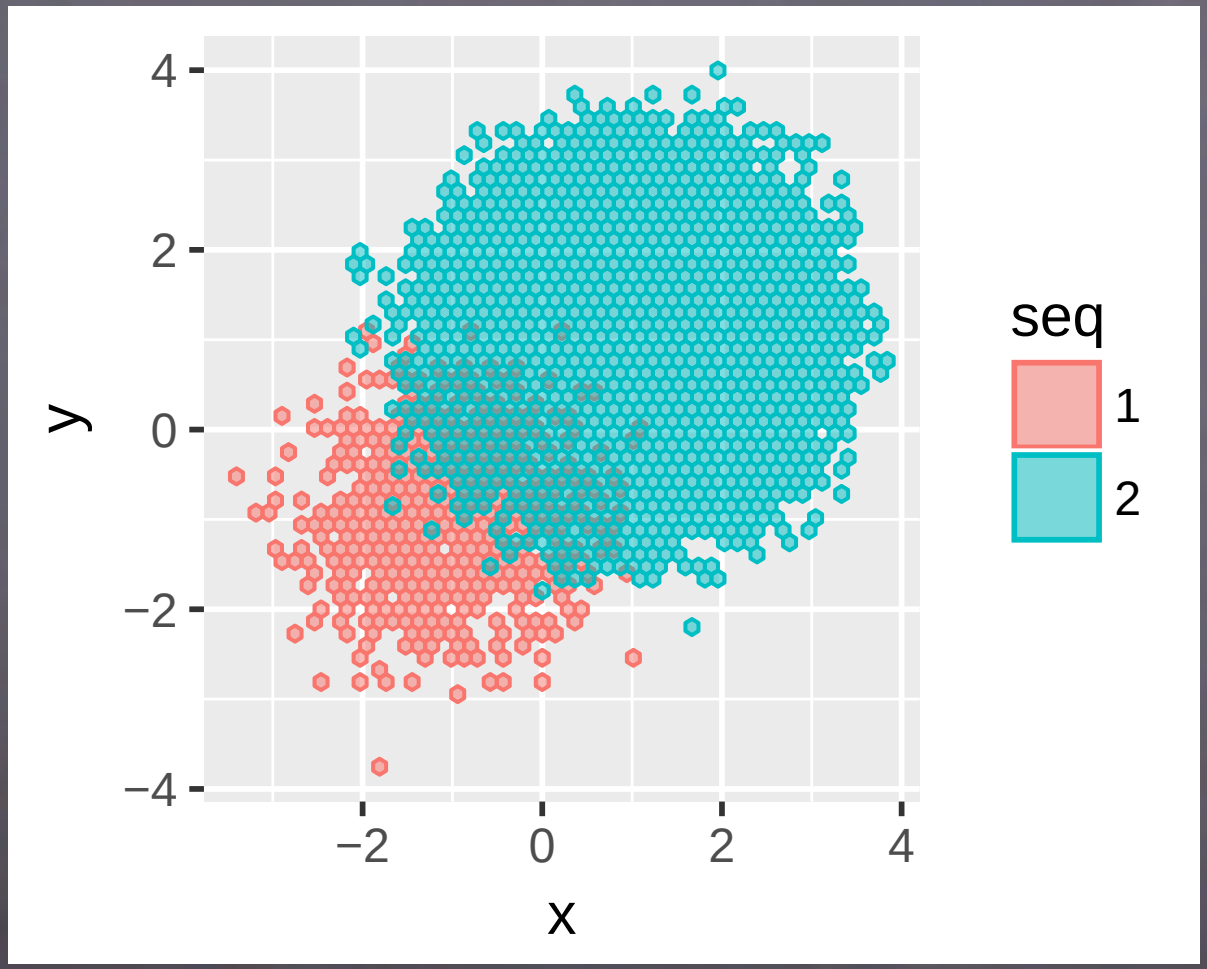
Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

The evolution of marine microbes over billions of years predicts Gene sequences, most commonly those encoding

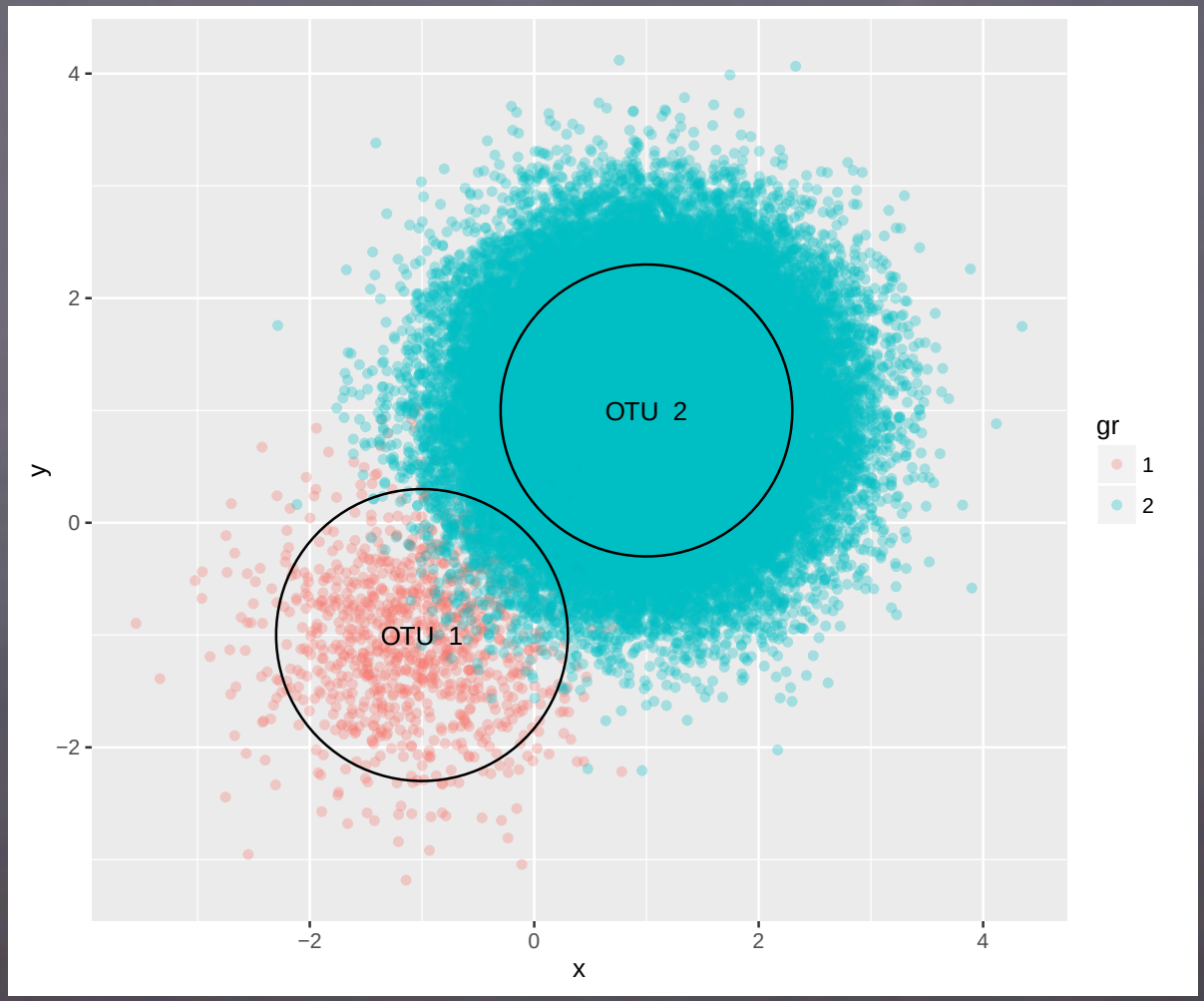
How many words does Professor D. know?

- Maybe 15,000, 20,000?
- Start sampling..... banana, bannana, bannanna, orange, orenge, muscle, musel, muscel, foreign, forene, forane,.....
- How many real words does Prof D. know?
- Use more information than the spelling....

From reads to Operational Taxonomic Units

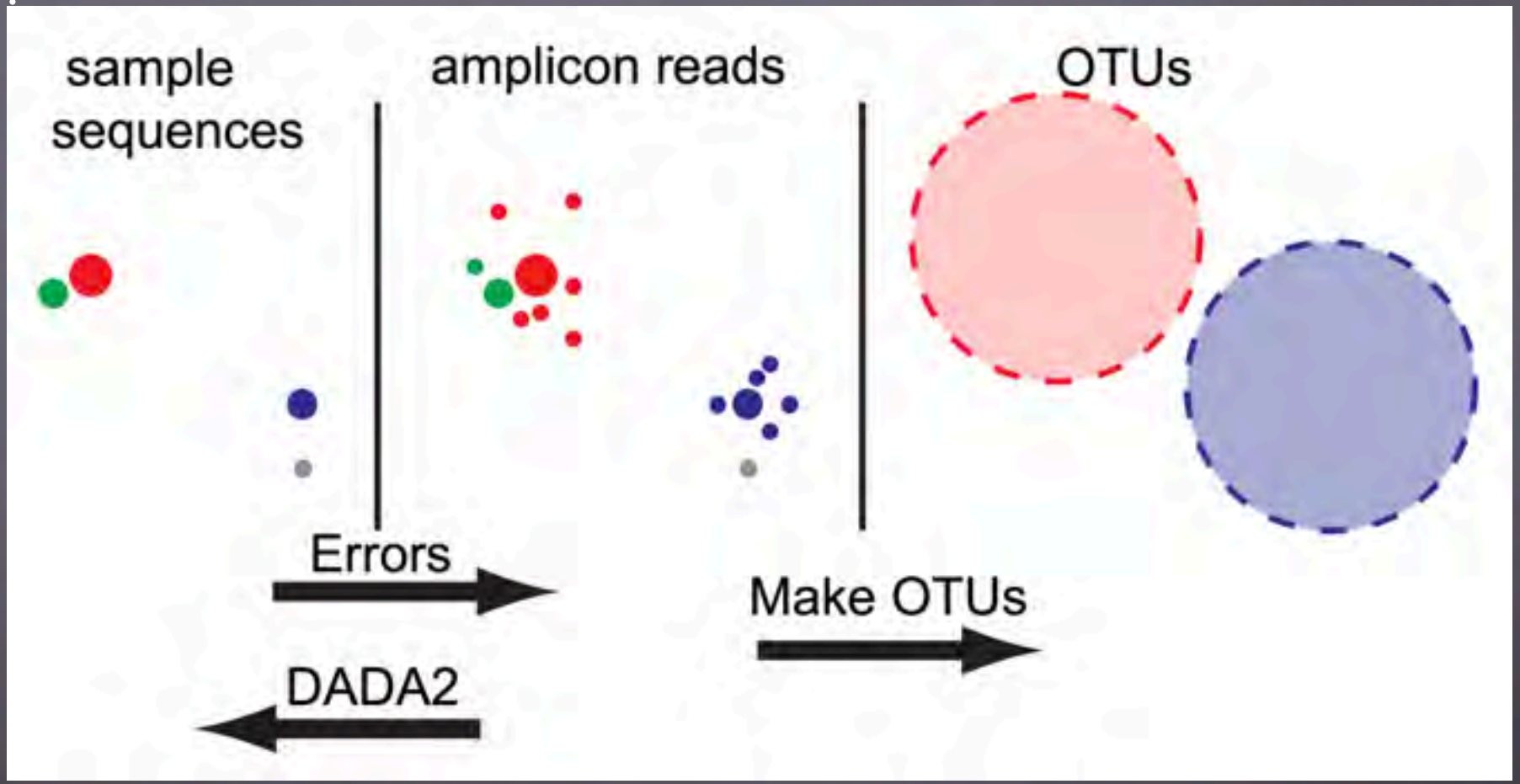


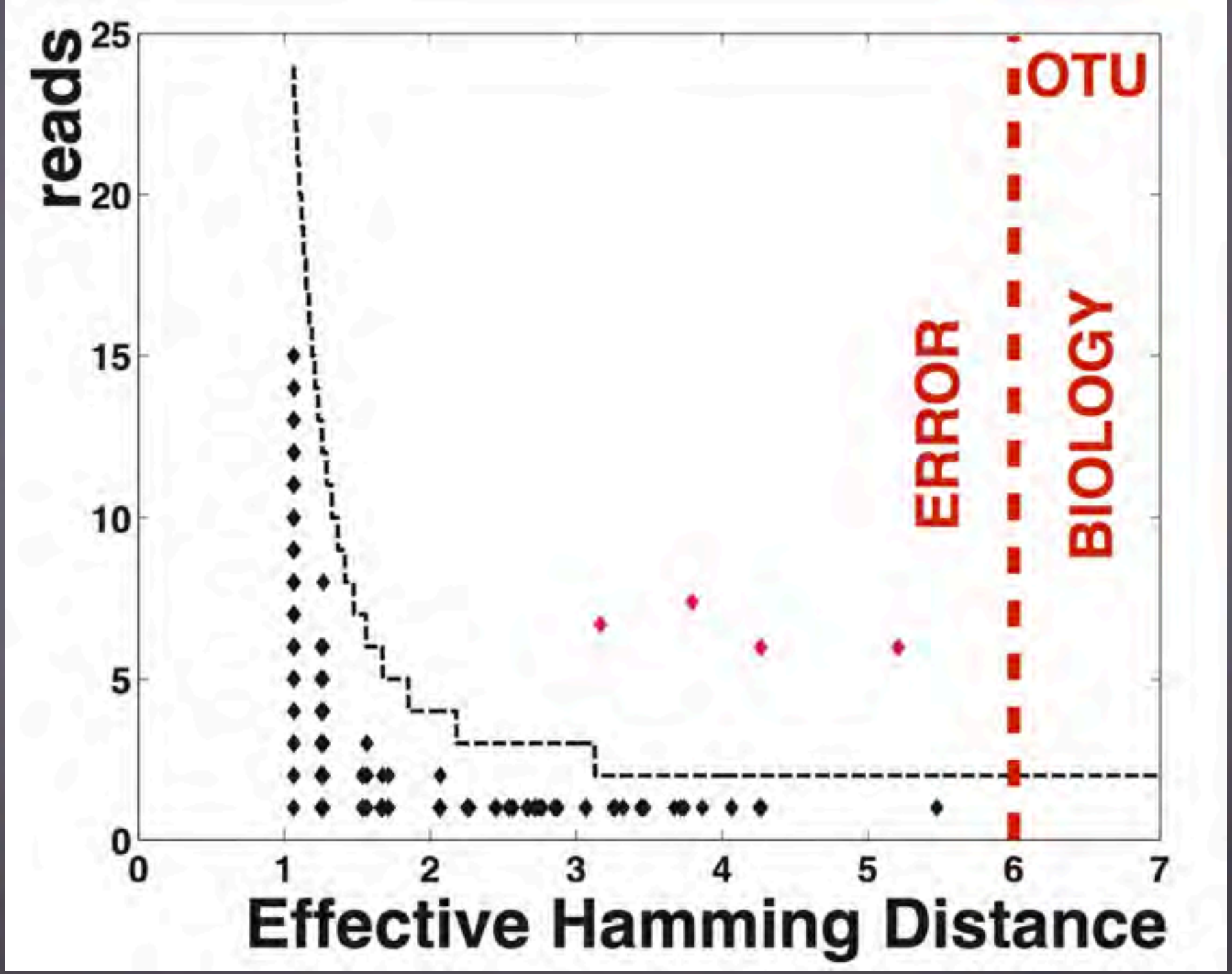
From reads to Operational Taxonomic Units

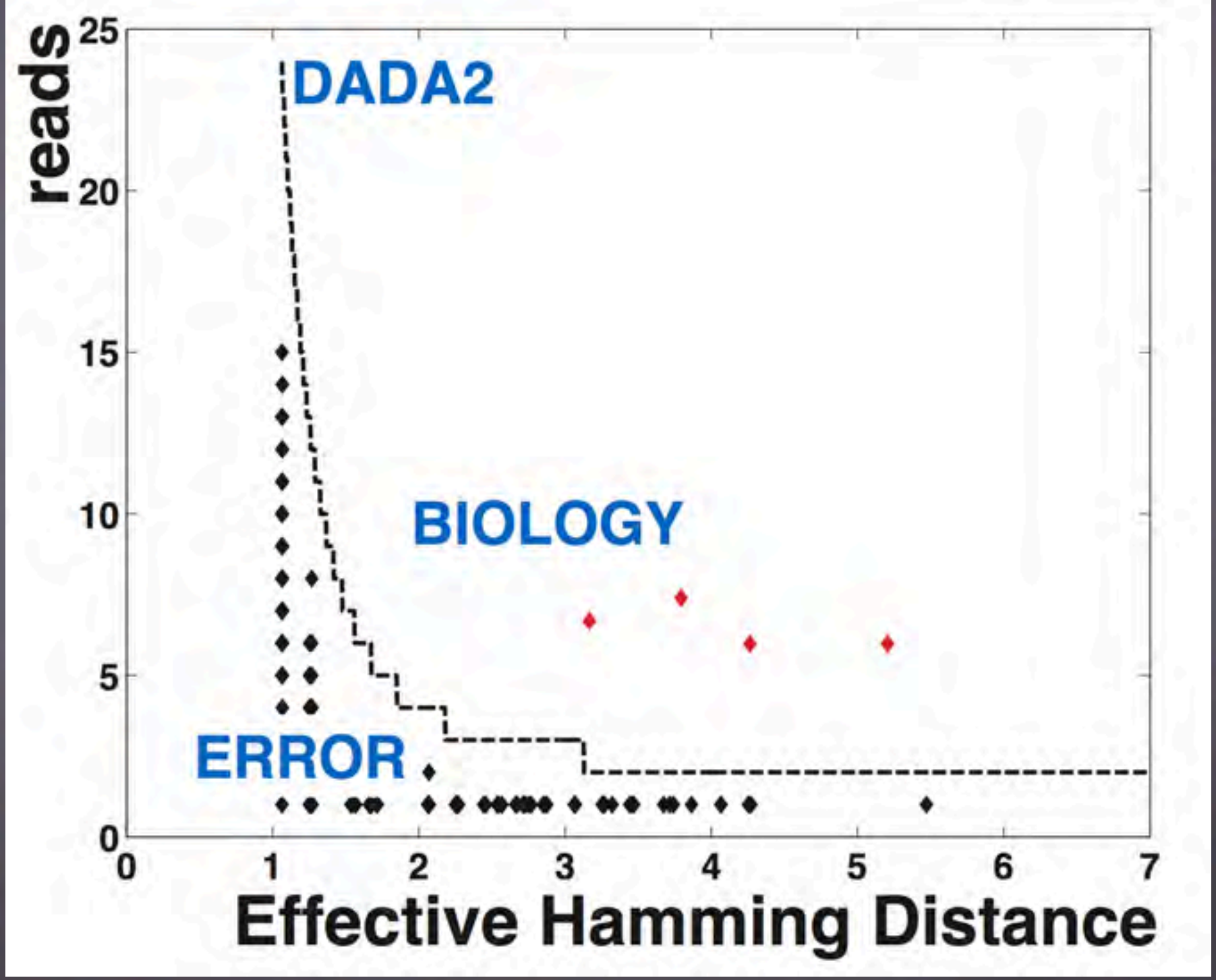


Current practice (`qiime`, `mothur`, `rdp`, ...): 97% similarity

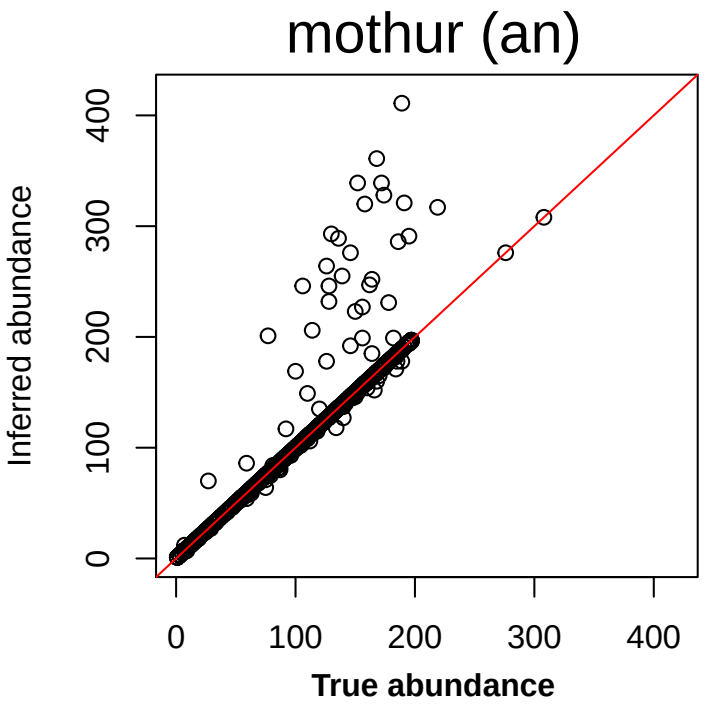
Probabilistic Model



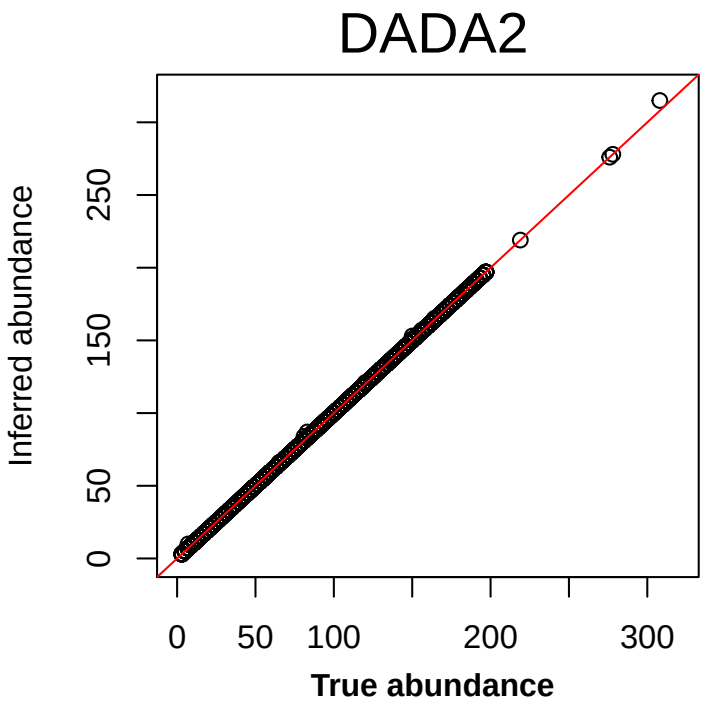

reads
25
20
15
10
5
0
0
1
2
3
4
5
6
7
Effective Hamming Distance
OTU
ERROR
BIOLOGY


reads
25
20
15
10
5
0
DADA2
BIOLOGY
ERROR
0
1
2
3
4
5
6
7
Effective Hamming Distance

Accuracy: Simulated data



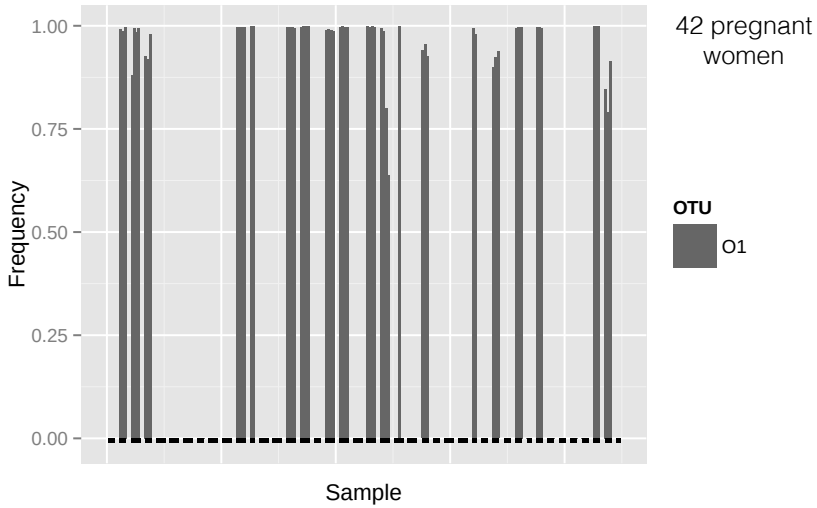
TP: 978
FP: 272
FN: 77
cor: 0.935



TP: 1042
FP: 0
FN: 13
cor: 0.999

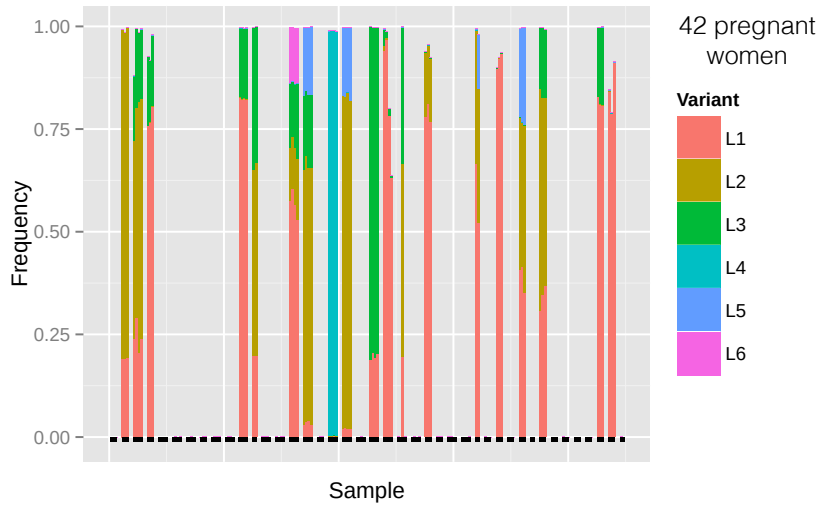
Data: Kopylova, et al. mSystems, 2016.

Resolution: L. crispatus



Data: MacIntyre et al. Scientific Reports, 2015.

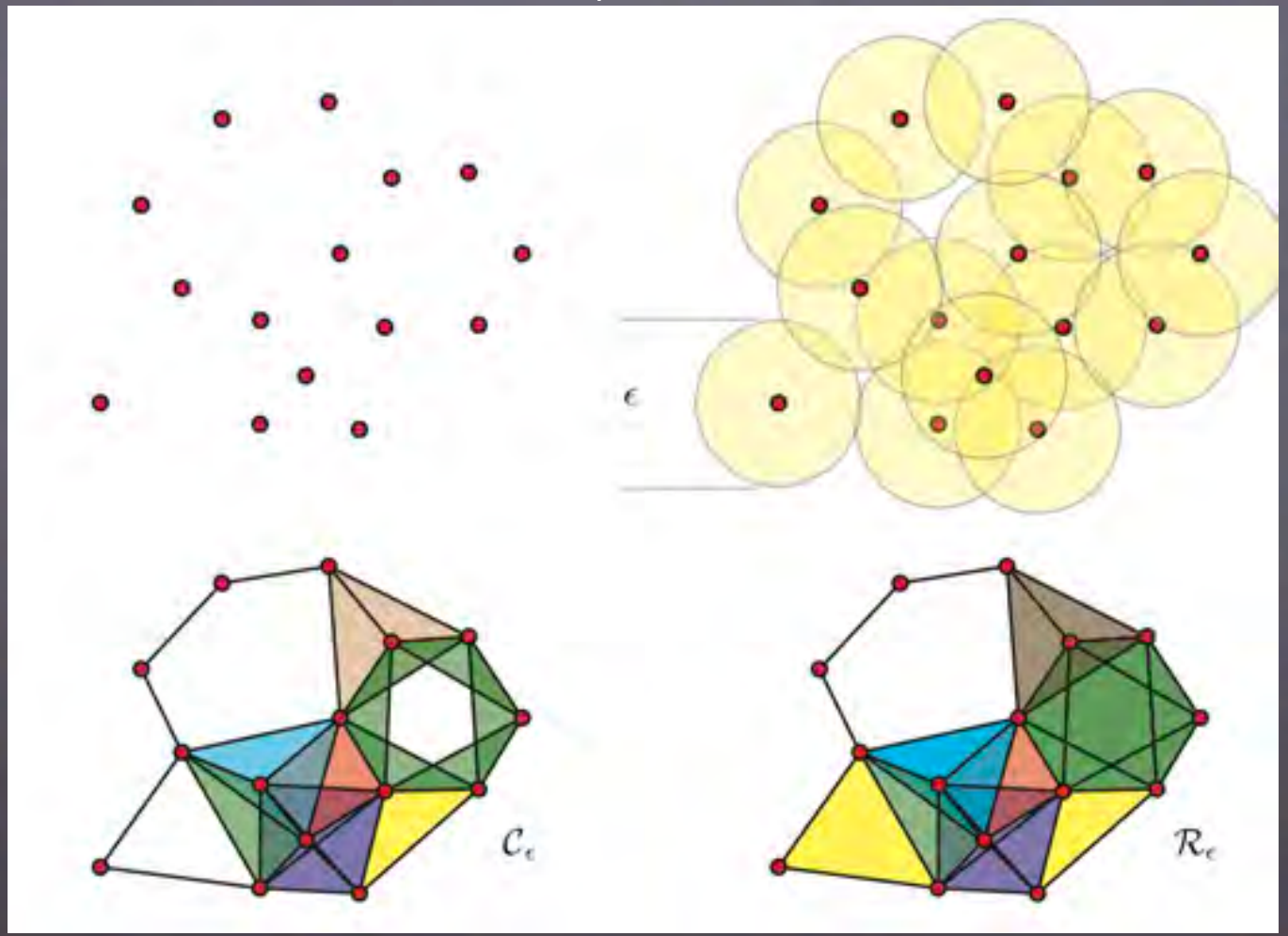
Resolution: L. crispatus



Data: MacIntyre et al. Scientific Reports, 2015.

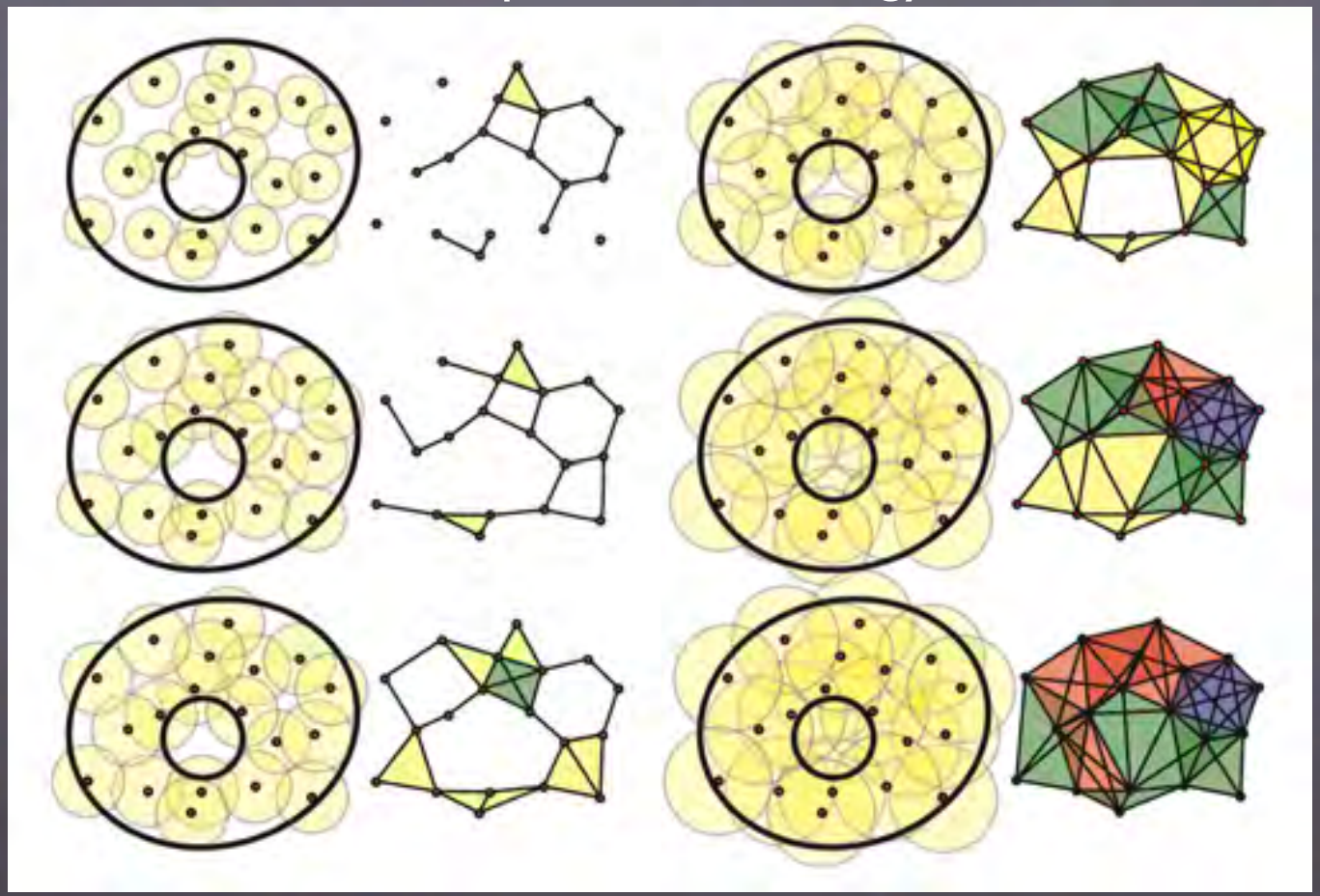
Mathematicians also have baseline measure problems

Examples in Topological Data Analysis (Edelsbrunner, Carlsson, Zoromodian, Ghrist et al.).



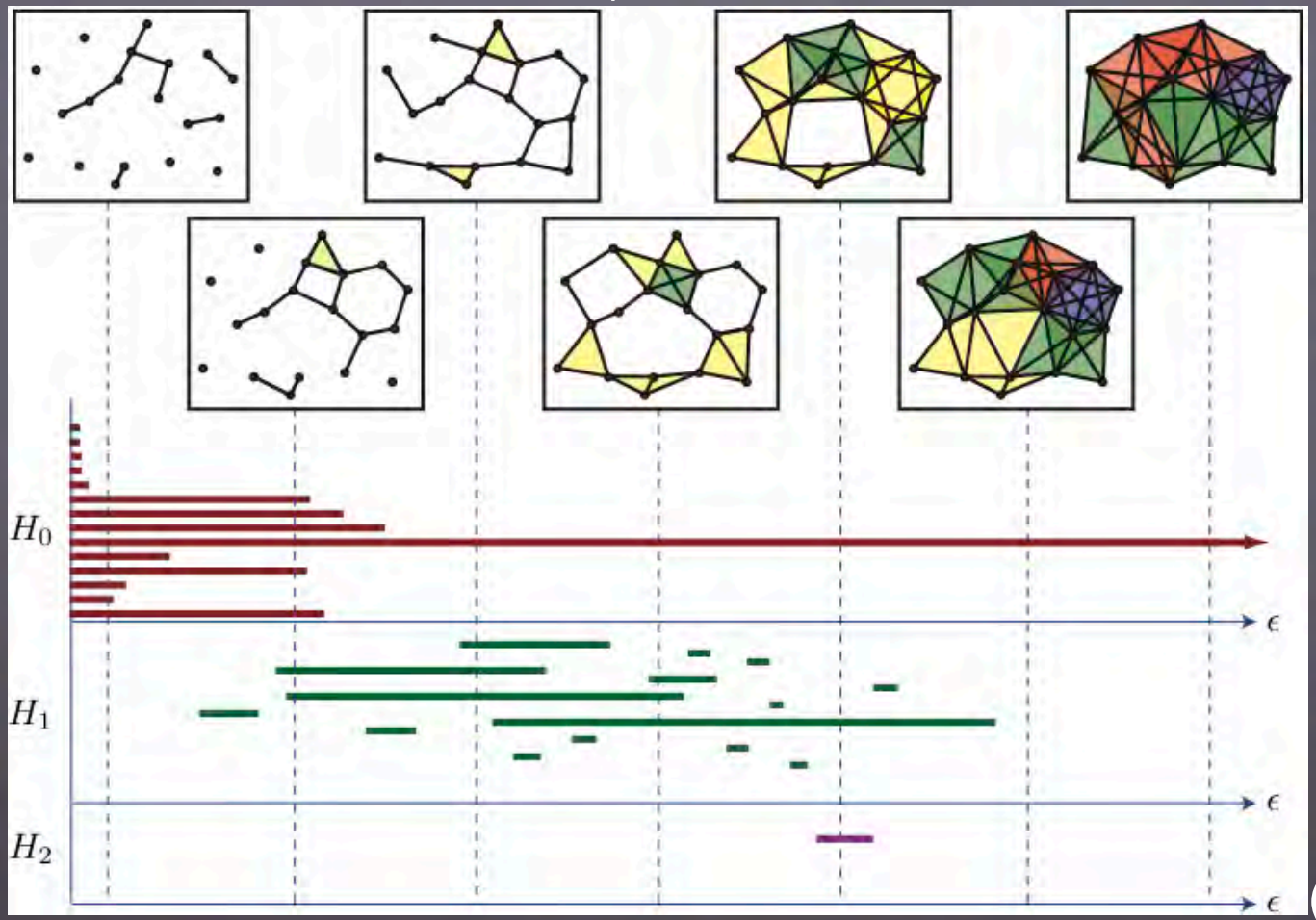
Mathematicians also have baseline measure problems

Ghrist, R. Barcodes: persistent toology of data, AMS, 2008



Mathematicians also have baseline measure problems

Examples in Topological Data Analysis (Edelsbrunner, Carlsson, Zoromodian, Ghrist et al.).

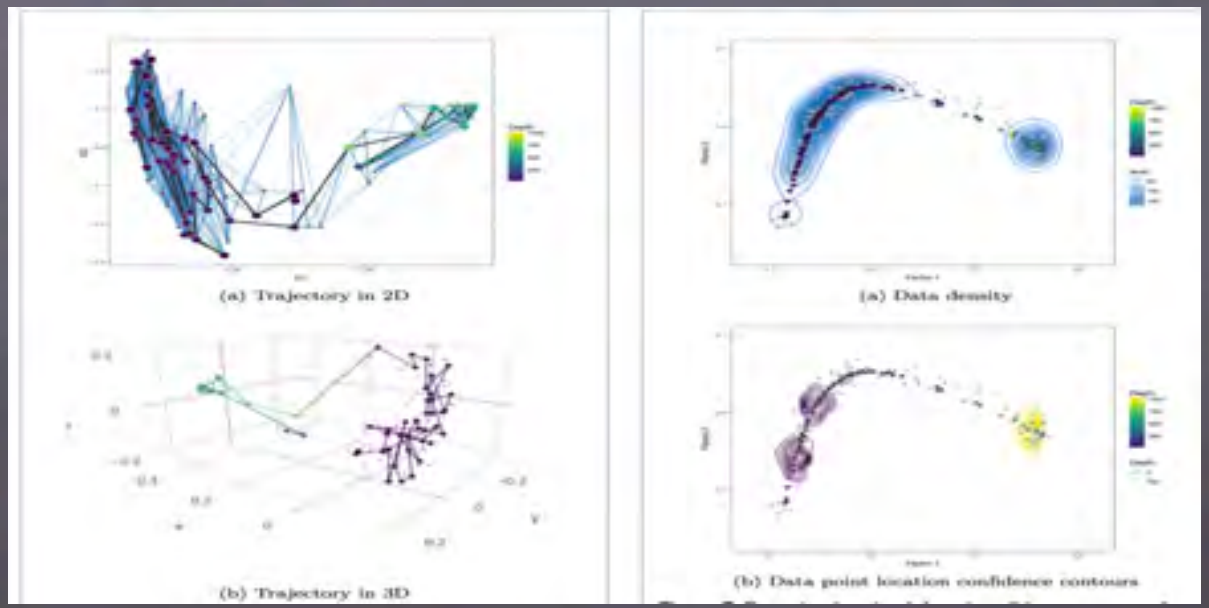


(Ghrist)

Open Question: How to make a method designed for uniformly

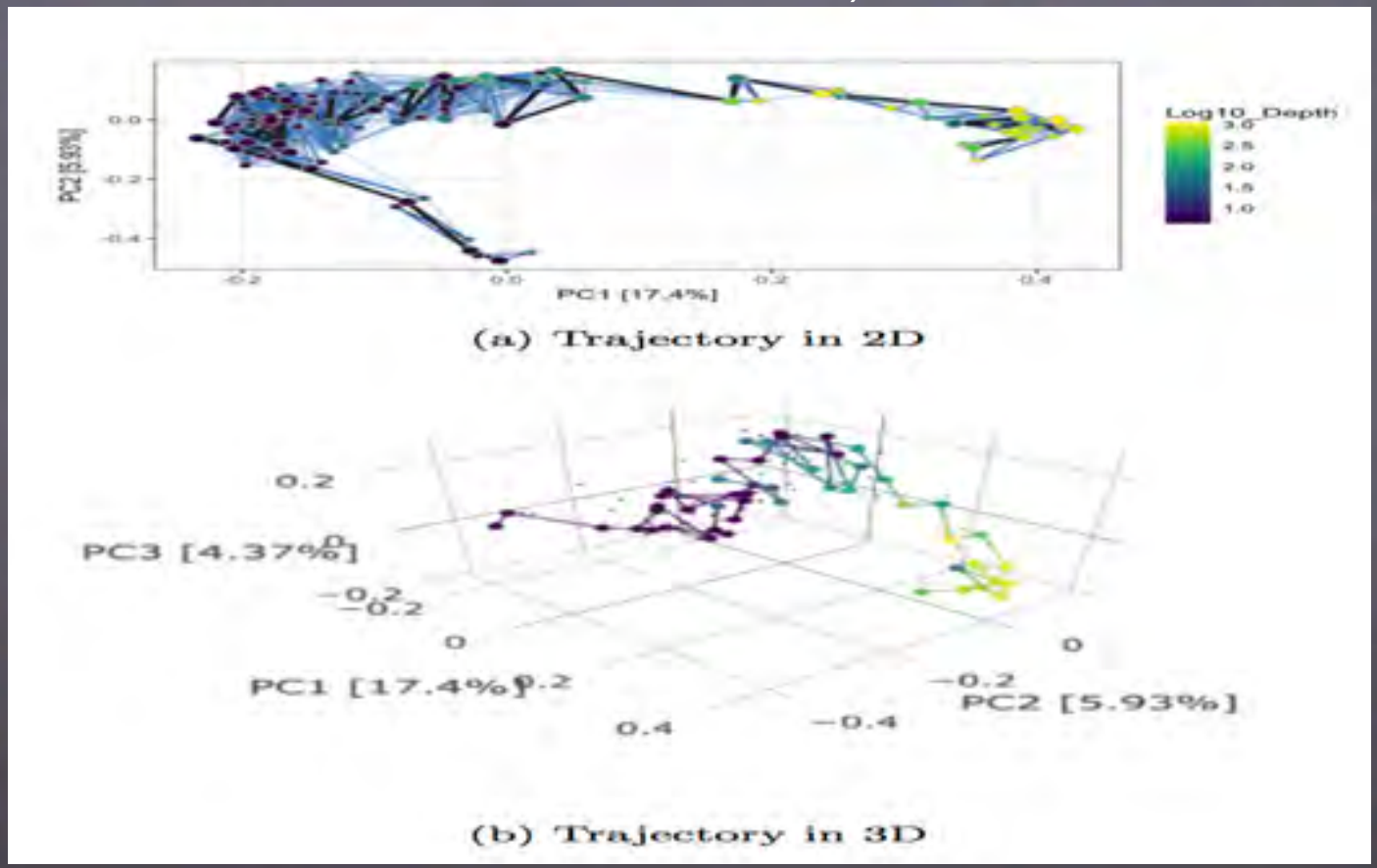
Part VI

Uncertainty quantification for latent gradients



Uncertainty Quantification for rankings and gradients

Bayesian Unidimensional Scaling (Lan Huong Nguyen and Susan Holmes, 2017, BMC Bioinformatics).



Bayesian model for distances

$$
\begin{aligned}
d_{ij}|\delta_{ij} &\sim \text{Gamma}[\mu_{ij} = \delta_{ij},\, \sigma^2_{ij} = s^2_{ij}\sigma^2_\epsilon], \qquad (1)\\
\delta_{ij} &= |\tau_i - \tau_j|,\\
\tau_i|\alpha_\tau, \beta_\tau &\sim \text{Beta}(\alpha_\tau, \beta_\tau),\\
\alpha_\tau &\sim \text{Cauchy}^+(1, \gamma_\tau),\\
\beta_\tau &\sim \text{Cauchy}^+(1, \gamma_\tau),\\
\sigma_\epsilon &\sim \text{Cauchy}^+(0, \gamma_\epsilon),
\end{aligned}
$$

Modeling the heteroscedasticity

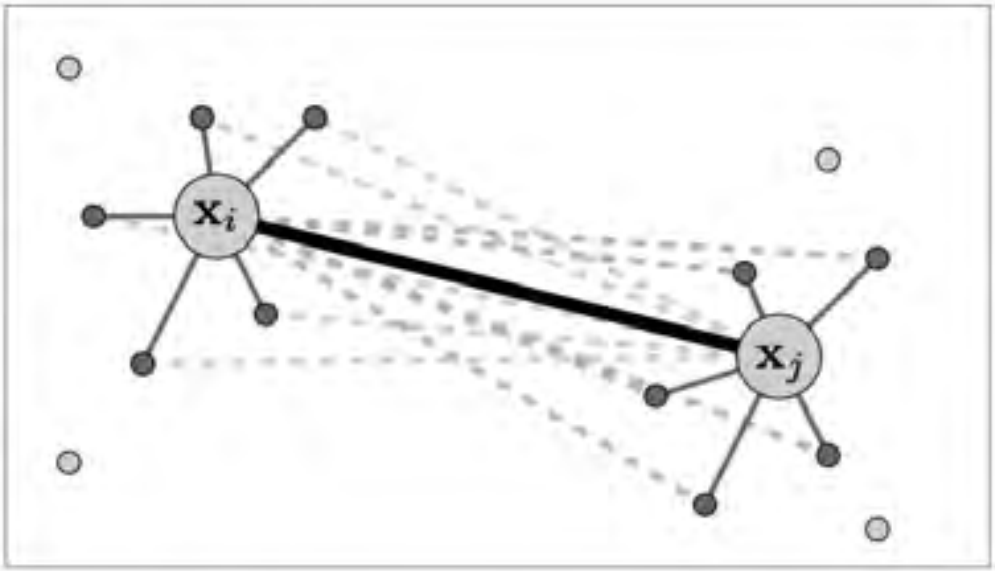
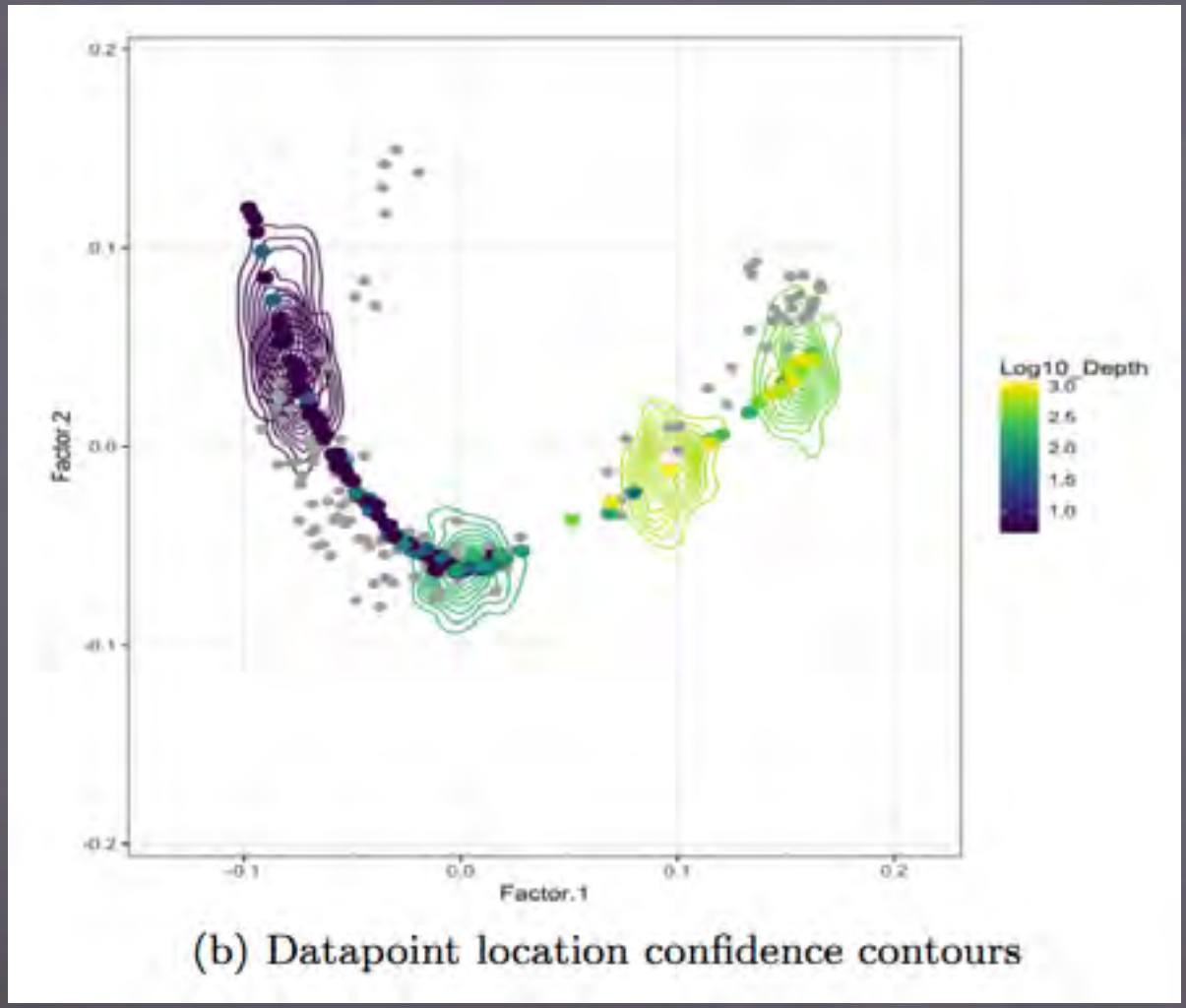


Figure 1 Graphical representation of points $\mathbf{x}_i$ and $\mathbf{x}_j$ together with their neighbors. The set of (dashed) distances from $\mathbf{x}_i$ to the K-nearest-neighbors of $\mathbf{x}_j$, and from $\mathbf{x}_i$ to the K-nearest-neighbors of $\mathbf{x}_i$ is used to compute $\hat{s}^2(d_{ij})$, the estimate of the variance of d_{ij}. Here we chose $K = 5$.

$$s(\hat{d}_{ij}) = \frac{1}{|D_{ij}^K|} \sum_{d \in D_{ij}^K} (d - \bar{d}_{ij}^K)^2$$

Scale parameter for the error term: $s_{ij}^2 = s(\hat{d}_{ij}) / s(\bar{\hat{d}}_{ij})$.



(b) Datapoint location confidence contours

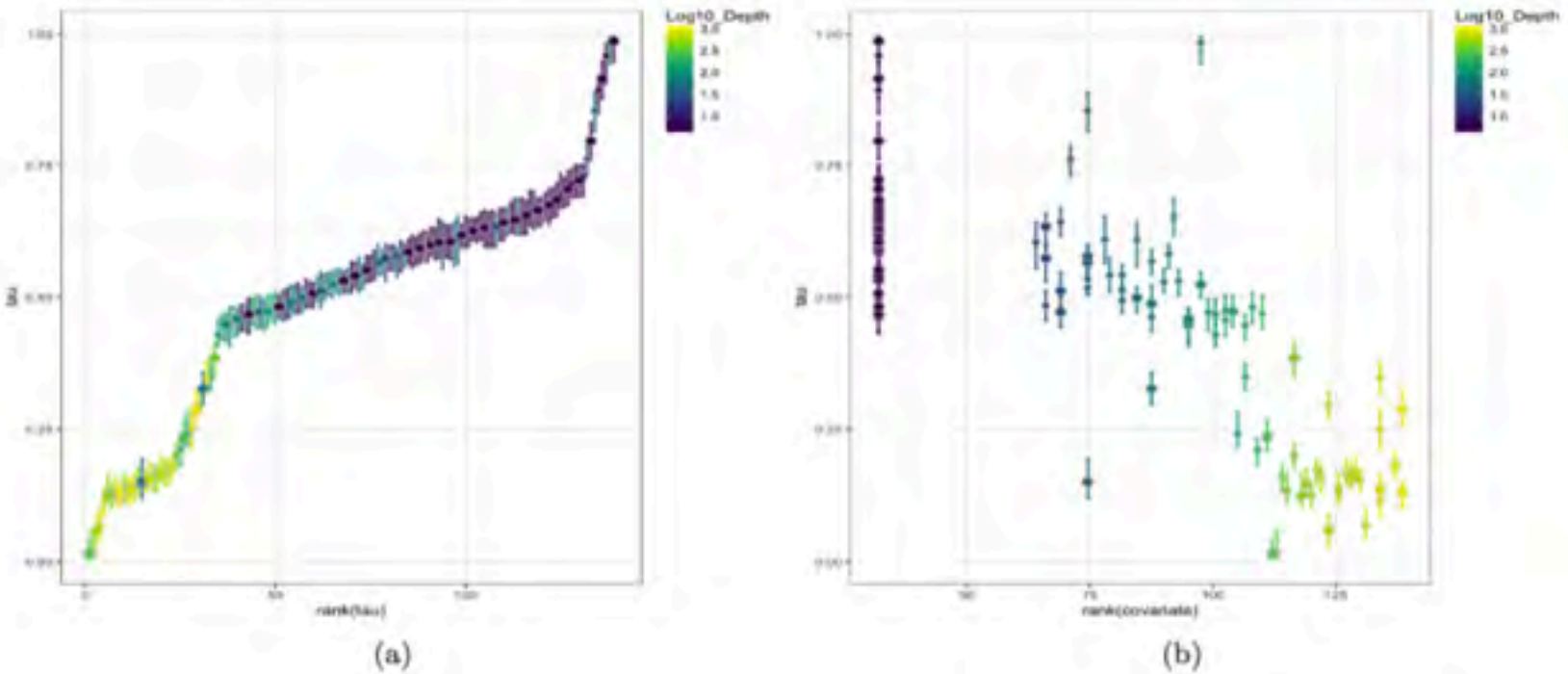


Fig. 4
Latent ordering in TARA Oceans dataset shown with uncertainties. The differences in the slope of plot (**a**) indicate varying data coverage along the underlying gradient. Correlation between the water depth and the latent ordering in microbial composition data is shown in (**b**). Coloring corresponds to log10 of the water depth (in meters) at which the ocean sample was collected

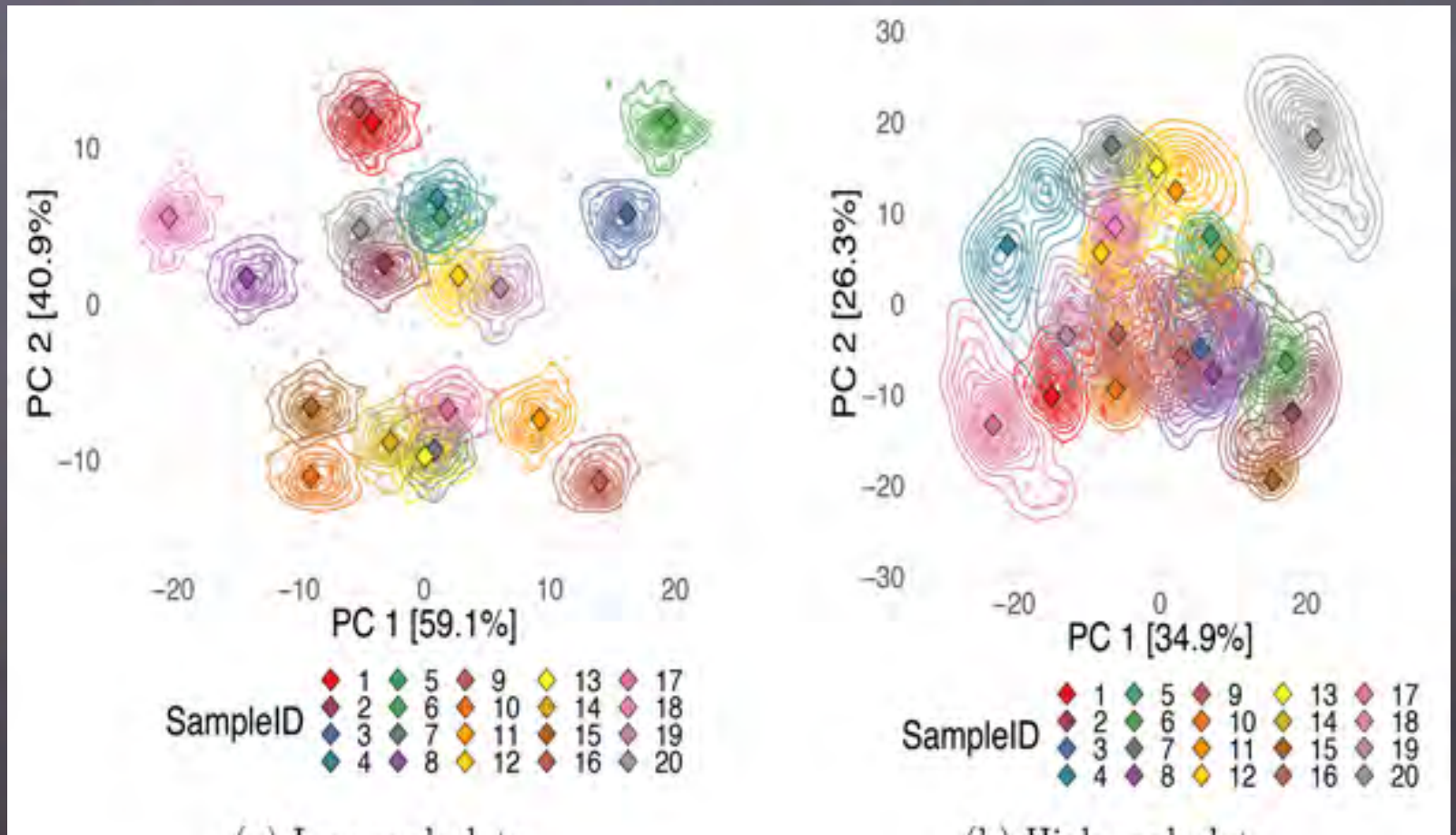
Code using `stan`

```
fit_buds <- function(D, K = NULL,
                     method = c("vb", "mcmc"),
                     hyperparams = list(
                       "gamma_tau"= 2.5,
                       "gamma_epsilon" = 2.5,
                       "gamma_bias" = 2.5,
                       "gamma_rho" = 2.5,
                       "min_sigma" = 0.03),
                     init_from = c("random", "principal_cu
                     seed = 1234, max_trials = 20, ...) {
```

`buds` package on github.

Part VII

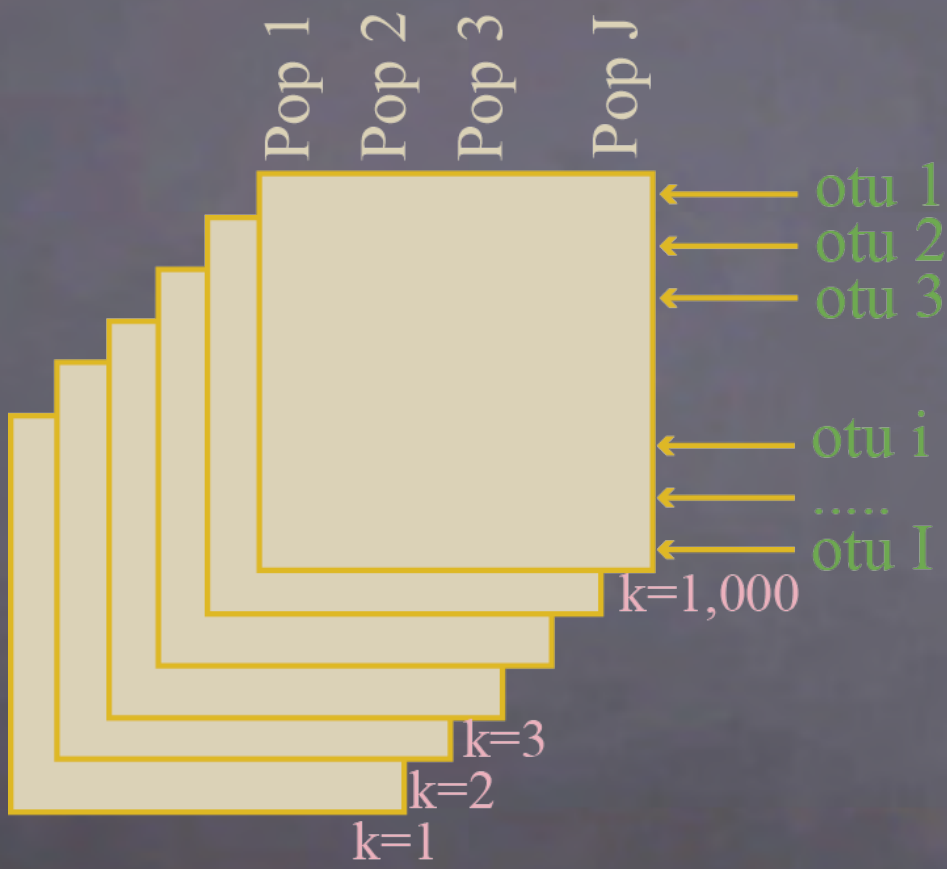
Uncertainty quantification for latent factors



Full Bayesian nonparametric model

- We do not know the number of OTUs.
- We suppose underlying low dimensional latent variables for the sample P^j's.
- We use dependent microbial distributions, marginal priors of discrete distributions are built using manipulation of a Gaussian process and then extending this to multiple correlated distributions.

Generalization: Bayesian posterior uncertainty measures



Parameters for samples
$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$
Define a joint prior on these factors through the Gram matrix
$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$
The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of ASVs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j},$$
$$\epsilon_{i,j} \quad \text{iid Normal}$$

Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren, Bacallado, Favaro, Holmes, Trippa (2017, JASA).

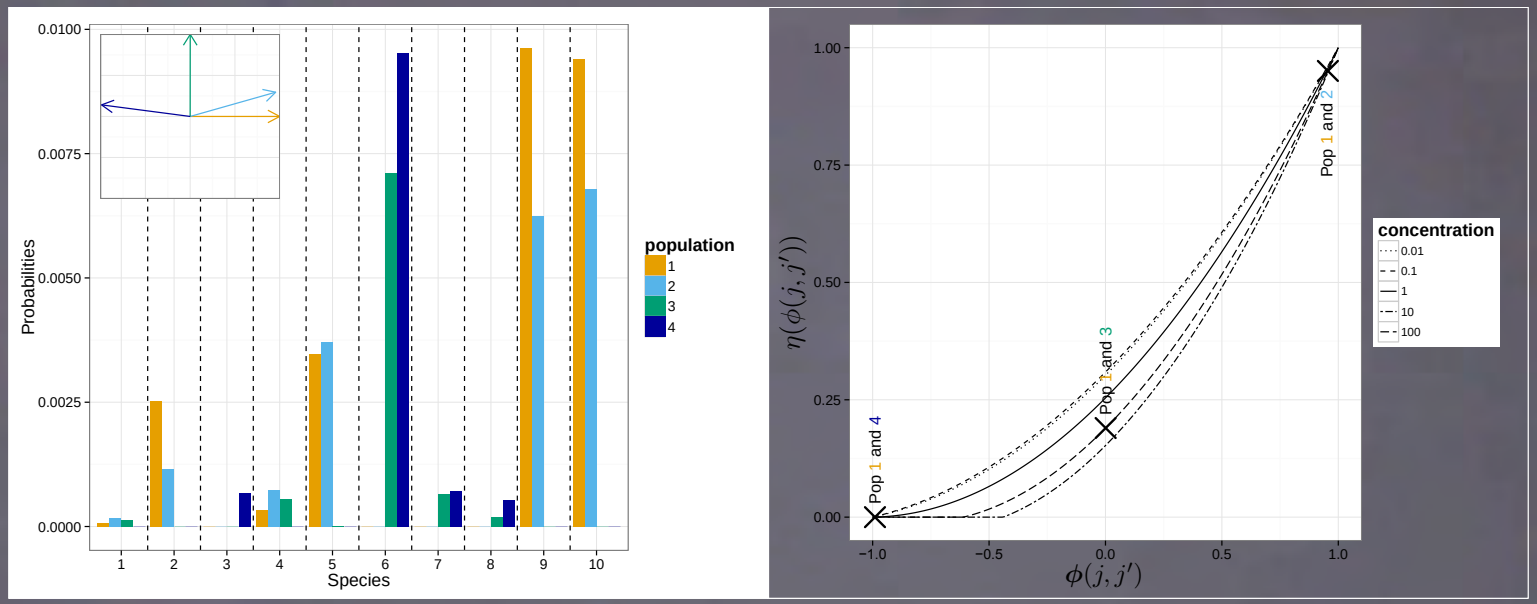


Figure: **Left panel**: realization of 4 microbial distributions from a dependent Dirichlet processes with 10 OTUs **Right panel**: correlation of two random probability measures when the cosine $\phi(j, j')$ between $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$ varies from -1 to 1. (Ren et al, JASA, 2017).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \qquad (1)$$

where the $\epsilon_{i,j}$ are independent Normal variables.

The degree of similarity between the discrete distributions $\{P^j; j \in \mathcal{J}\}$ is summarized by the Gram matrix $(\phi(j, j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle; j, j' \in \mathcal{J})$.
The dependent Dirichlet processes is defined by setting

$$P^j(A) = \frac{\sum_i \mathbb{I}(Z_i \in A) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}, \qquad \forall j \in \mathcal{J}, \qquad (2)$$

for every $A \in \mathcal{F}$. Here the sequence $(Z_1, Z_2, \ldots)$ and the array $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$, contain independent and identically distributed random variables, while σ is a Poisson process on the unit interval defined by using a prior on $\sigma = (\sigma_1, \sigma_2, \ldots)$, the distribution of ordered points $(\sigma_i > \sigma_{i+1})$ in a Poisson process on $(0, 1)$ with intensity

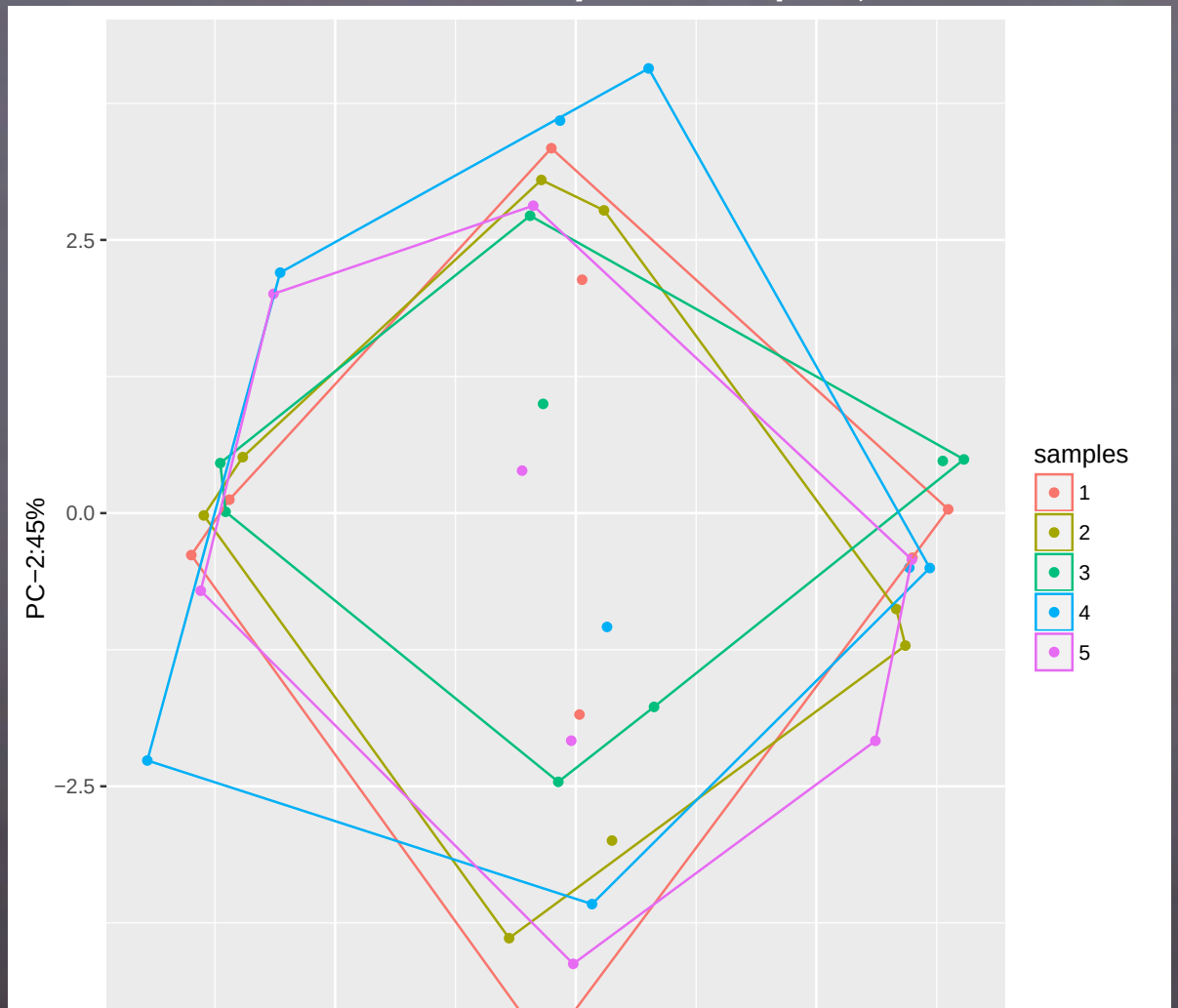
$$\nu(\sigma) = \alpha \sigma^{-1} (1 - \sigma)^{-1/2}, \qquad (3)$$

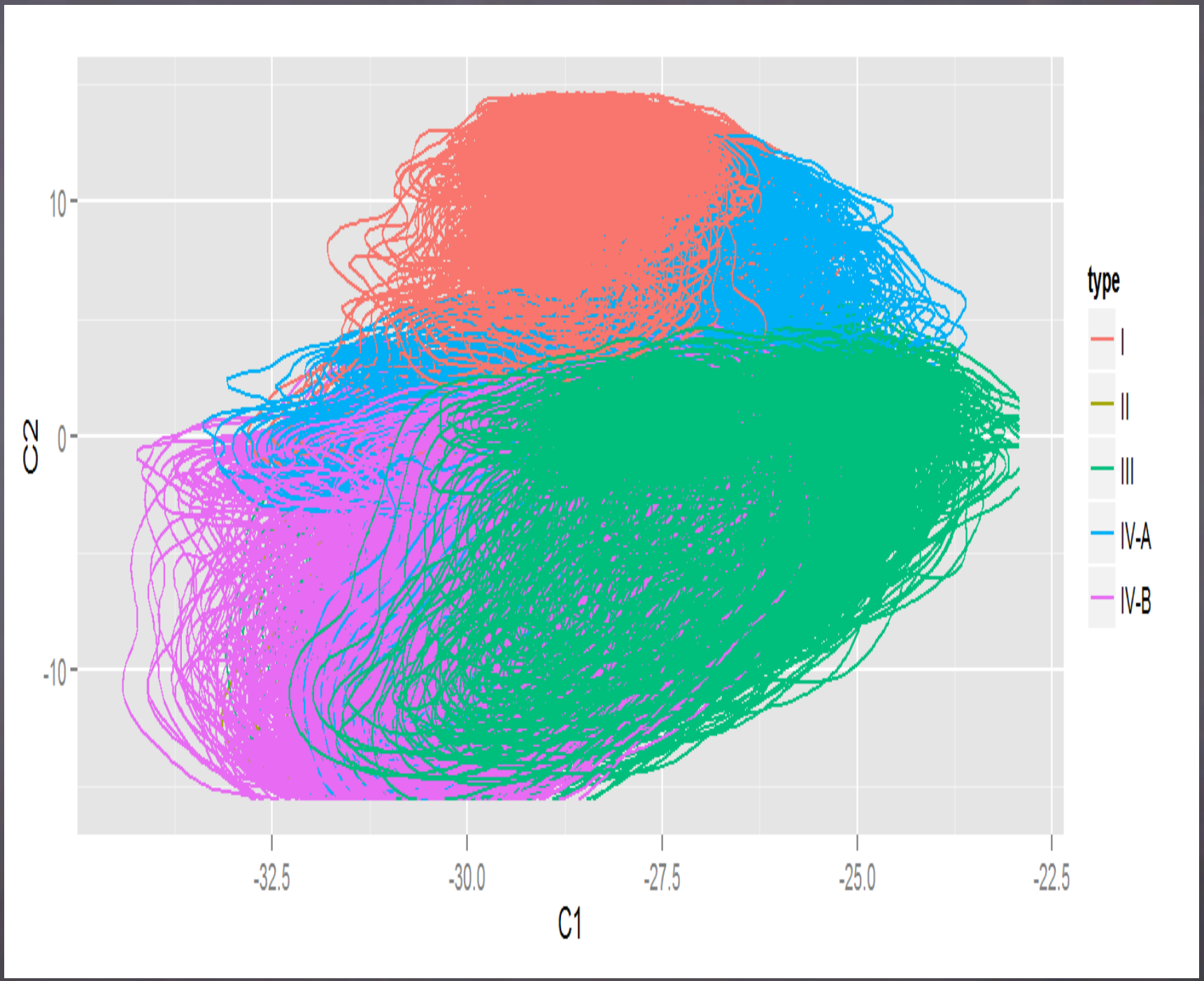
where $\alpha > 0$ is a concentration parameter.
We will use the notation $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$.

The methods that we consider here are all related to PCA and use the normalized Gram matrix **S** between biological samples.
S is the correlation matrix of $(\mathrm{Q}_{i,1}, \ldots, \mathrm{Q}_{i,J})$. Based on a single posterior instance of **S**, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.

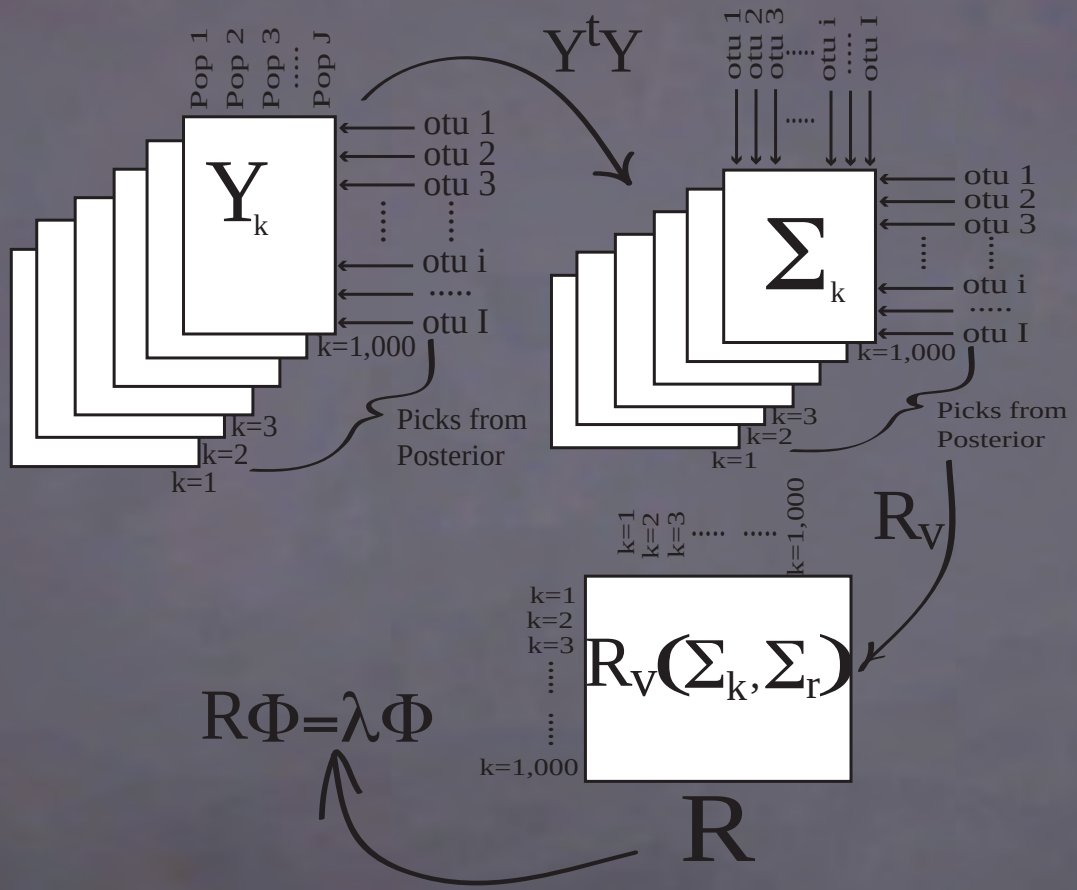
A projection approach

Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of **S** on the same plot *could* show the variability of the projections.




10
0
-10
C2
-32.5
-30.0
-27.5
-25.0
-22.5
C1
type
I
II
III
IV-A
IV-B

Alternatively



We identify a consensus lower dimensional space for all posterior samples using STATIS (Escoufier, 1980, see Holmes, 2005). We list the three main steps used to visualize the variability of **S**.

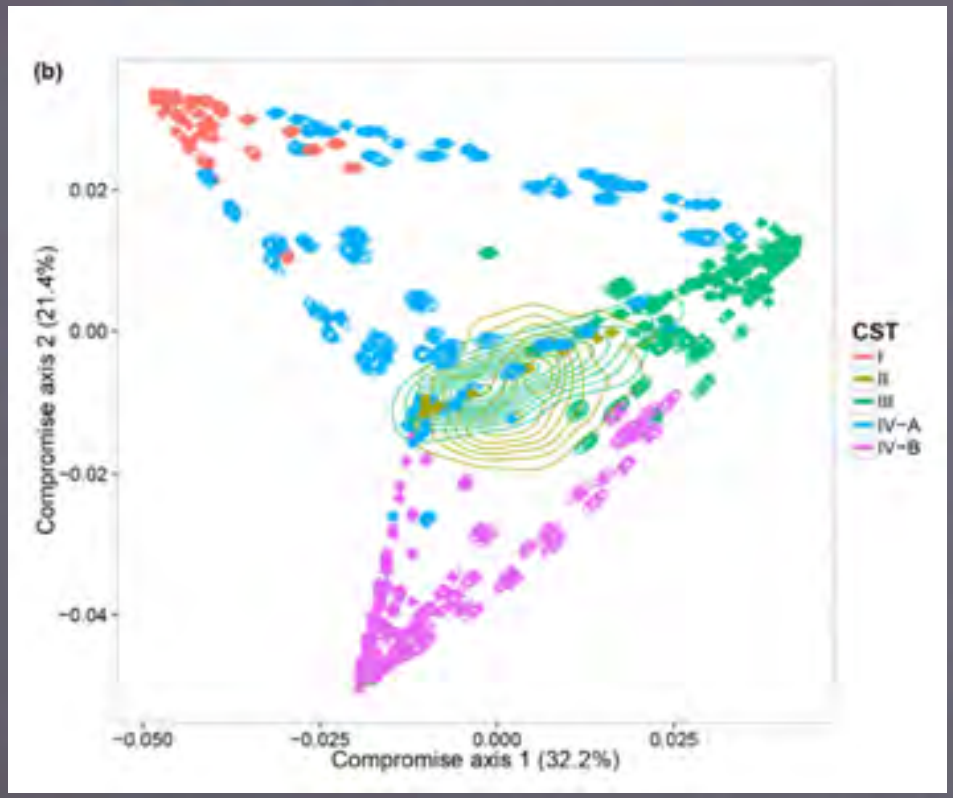
Registration: Find $\mathbf{S}_0$

Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.
We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.
We use the **RV** similarity metric between two symmetric square matrices **A** and **B**

$$RV(\mathbf{A}, \mathbf{B}) = Tr(\mathbf{AB})/\sqrt{Tr(\mathbf{AA})Tr(\mathbf{BB})}$$

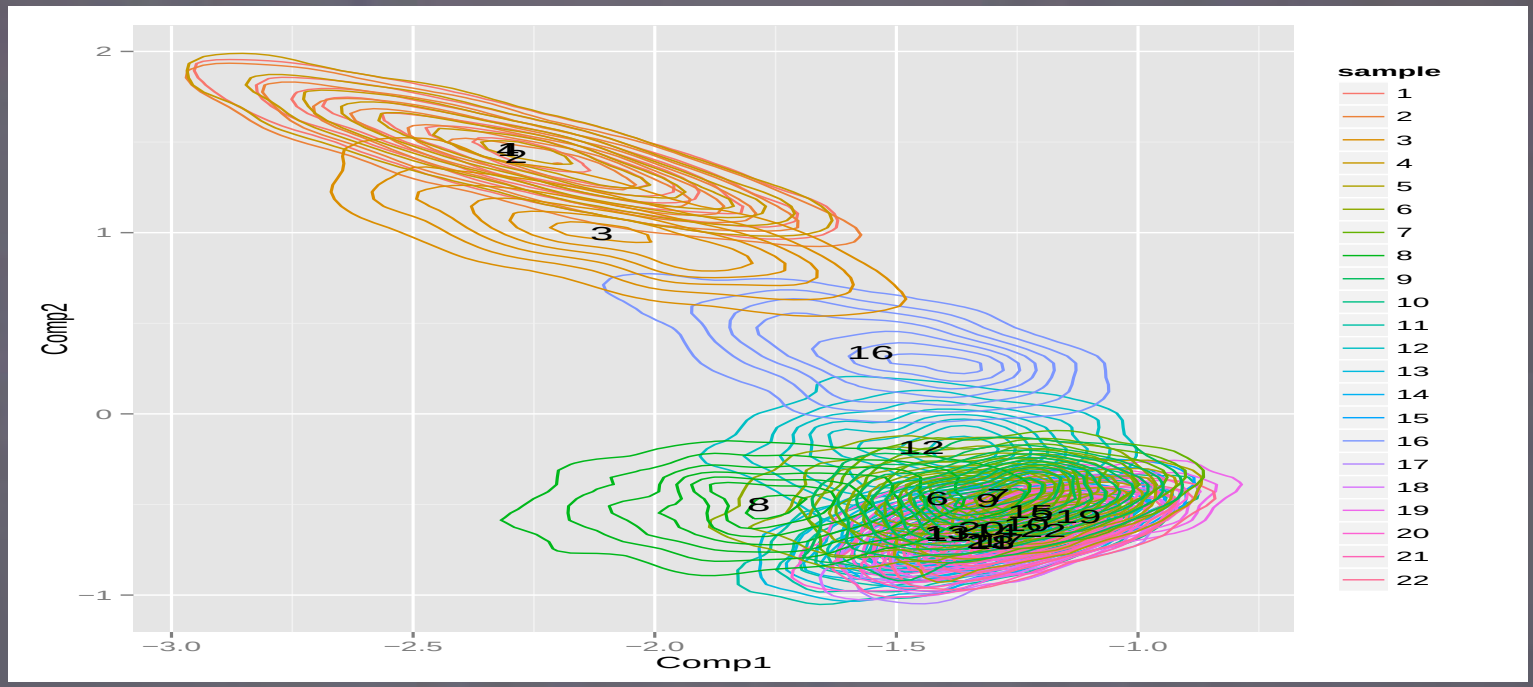
We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

We can see the uncertainties



Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2017 (JASA).

A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space V.



A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space V.

R packages and resources

phyloseq: http://bioconductor.org/packages/stats/bioc/phyloseq/

dada2: http://bioconductor.org/packages/stats/bioc/dada2/

treelapse: https://krisrs1128.github.io/treelapse/

treelapse antibiotics http://statweb.stanford.edu/~kriss1/antibiotic.html

microbiome_pvlm: https://github.com/krisrs1128/microbiome_plvm

decontam: https://github.com/benjjneb/decontam/

adaptiveGPCA: https://cran.r-project.org/web/packages/adaptiveGPCA/index.html

bootLong: https://github.com/PratheepaJ/bootLong/blob/master/vignettes/Workflow.Rmd

Modern Statistics for Modern Biology

http://bios221.stanford.edu/book/

Solutions for microbiome analyses: respect the data.

- Poor data quality, information $\longrightarrow$ quality scores & probability.
- Maintain all information $\longrightarrow$ sequences are names.
- Interpretation $\longrightarrow$ latent variables (gradients or clusters).
- Nonlinearity: gradients $\longrightarrow$ t-sne and buds for manifold estimation.
- Reproducibility $\longrightarrow$ complete code source.
- Heterogeneity $\longrightarrow$ multicomponent objects:phyloseq.
- Training and collaboration $\longrightarrow$ Rmd and html.
- Find the right "statistic" to bootstrap or compute posterior distribution for.

Benefitting from the tools and schools of Statisticians.....

Thanks to the `R` and `Bioconductor` community and to co-authors.

Wolfgang Huber, Joey McMurdie, Ben Callahan, JJ Allaire and Rob Gentleman.

Lab Group and David Relman

Postdoctoral Fellows Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Pratheepa Jeganathan, Nina Miolane. Students: John Cherian, Diana Proctor, Daniel Sprockett, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran, Claire Donnat. Funding from NIH TR01 and NSF-DMS.

phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.
How can I (my students, my postdocs...) learn more?
Ask me.
http://www-stat.stanford.edu/~susan/