Microbial Co-Exclusion and Co-Occurrence:

Making it All Add Up

Andrew Fernandes
Jean Macklaim
Gregor Reid
Gregory Gloor

Western

Schulich
MEDICINE & DENTISTRY

International Human Microbiome Congress

Hyatt Regency Vancouver
March 9-11,2011

Competitive Exclusion Principle

“Complete Competitors Can’t Coexist”

They live all Packed together

Or we can have <u>mutualists</u> that *love* to coexist!

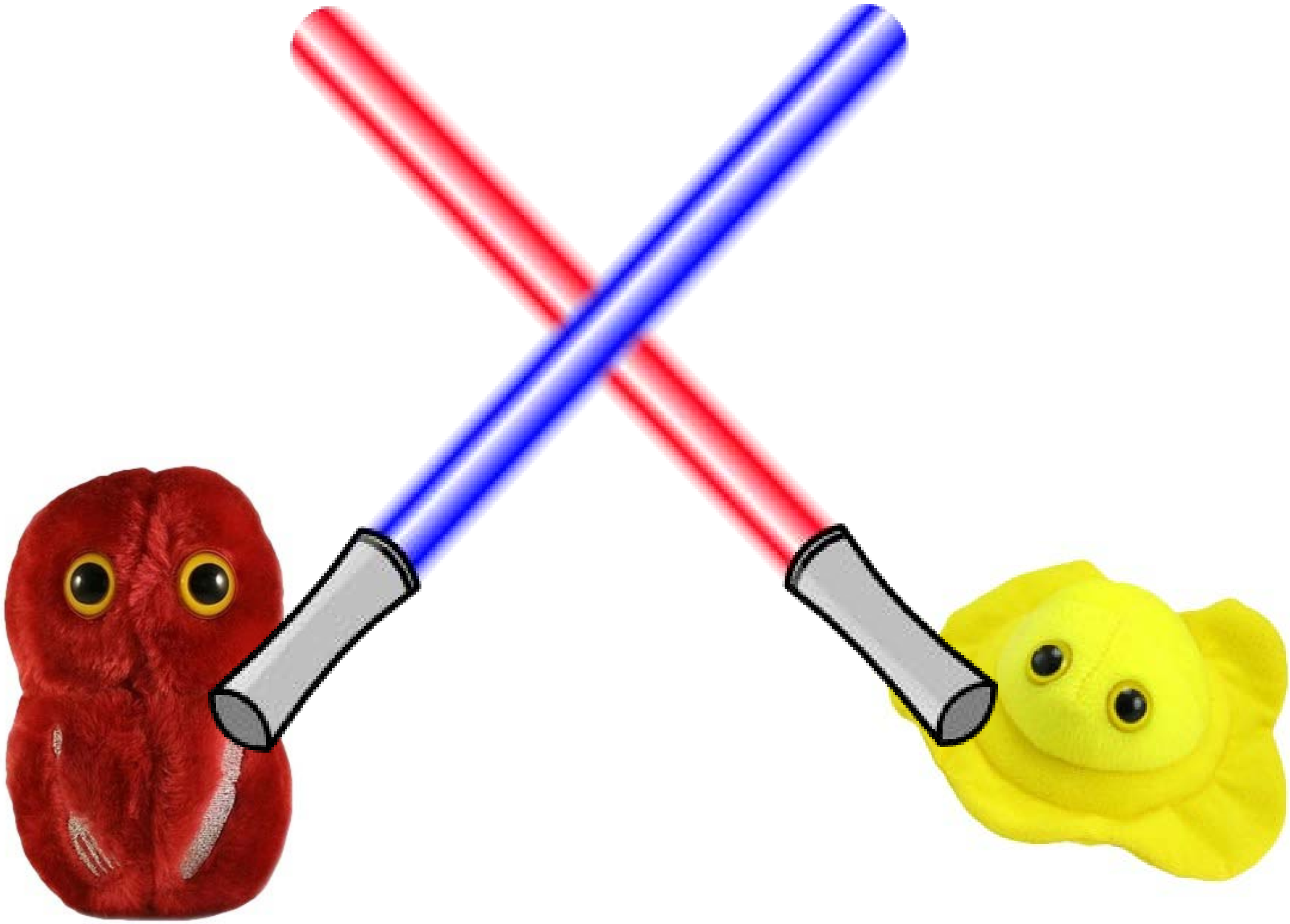
How do they interact?

Consider just
two species…

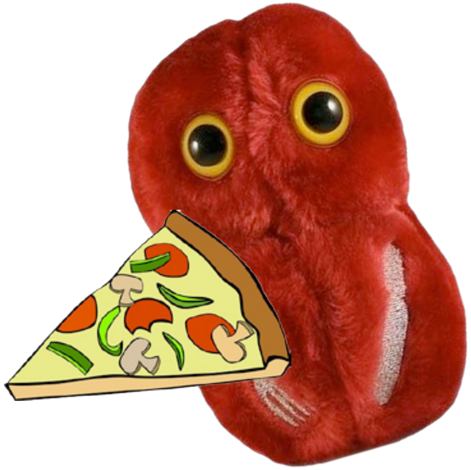
… living together
in an environment.

Co-Occurrence?

or Co-Exclusion?



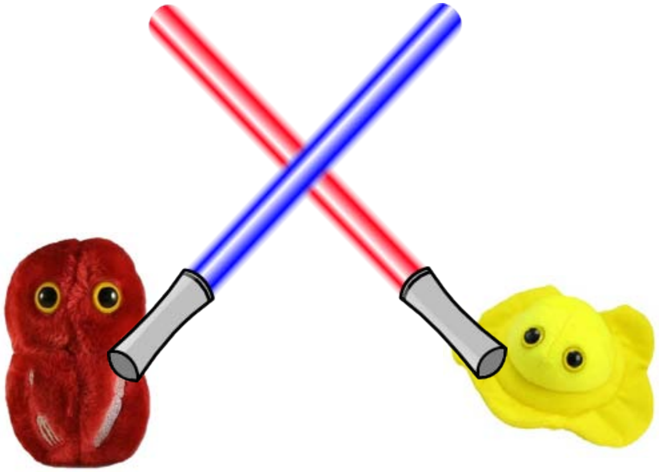
Indifference?



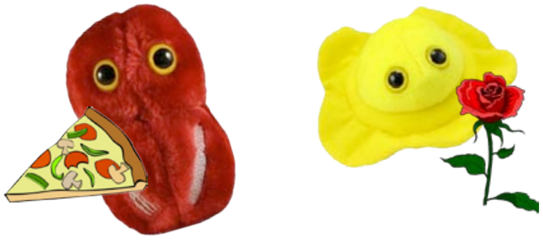
How can we tell *which* of these is occurring?

Co-Occurrence

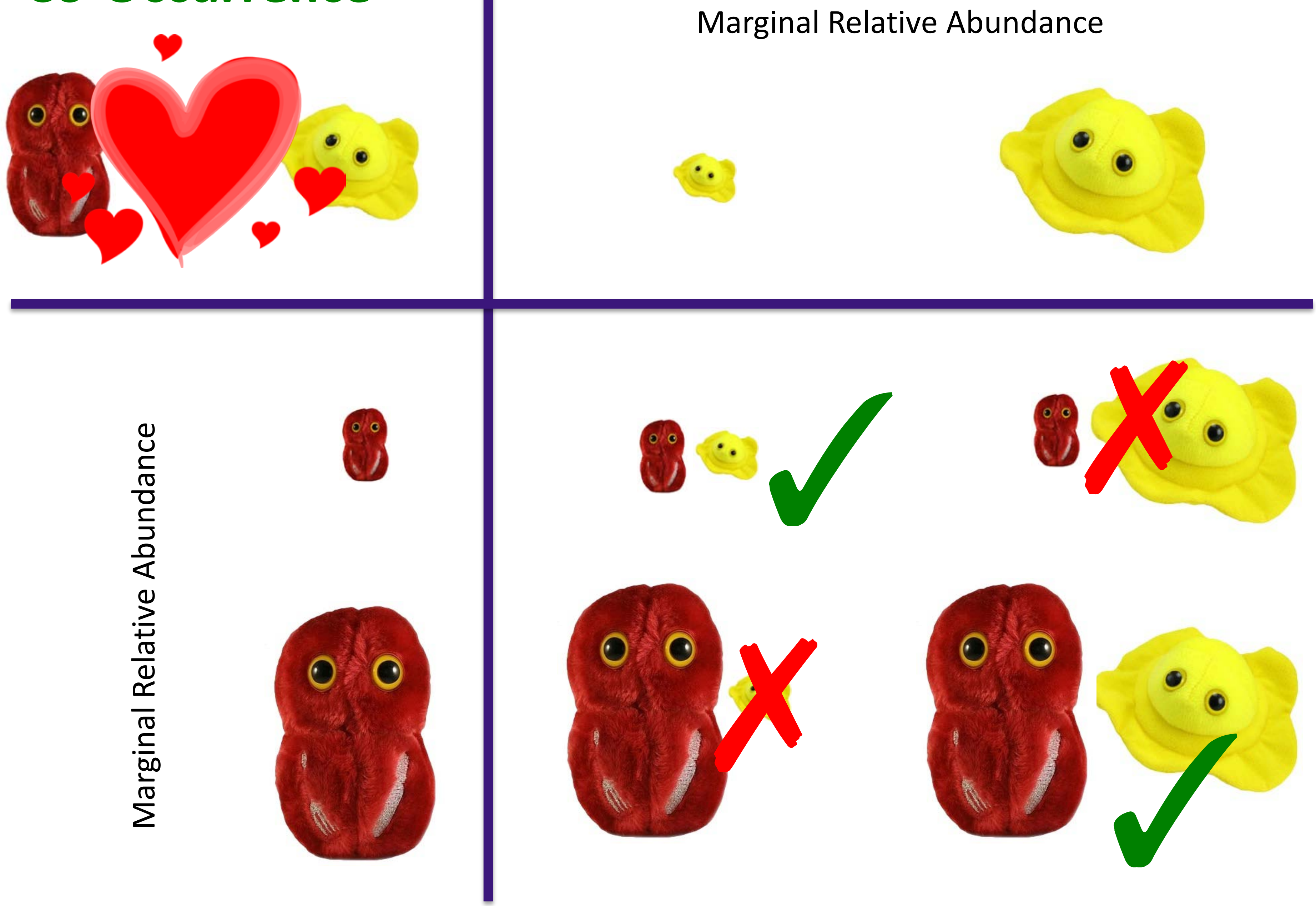
Co-Exclusion



Indifference



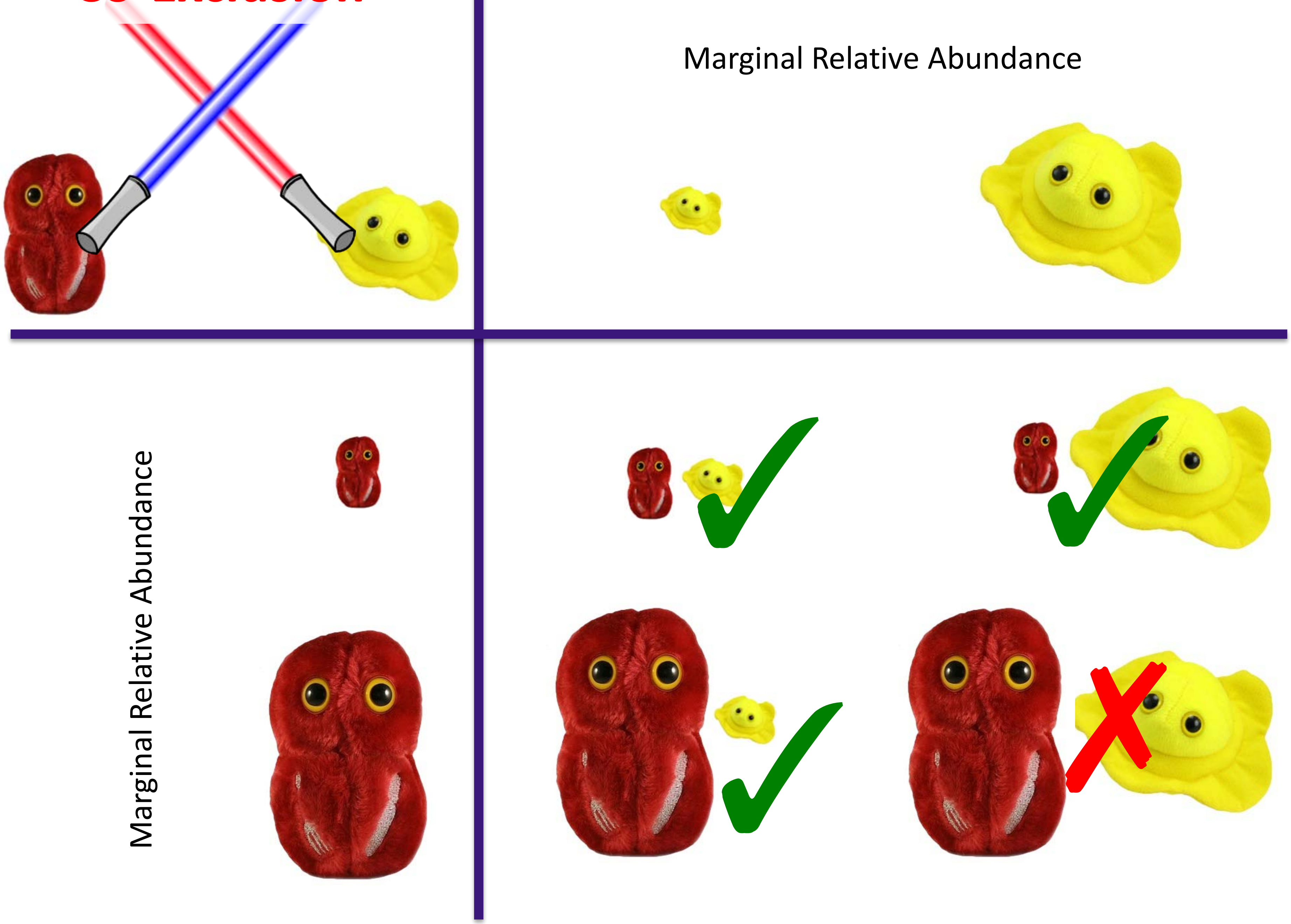
Co-Occurrence



Marginal Relative Abundance

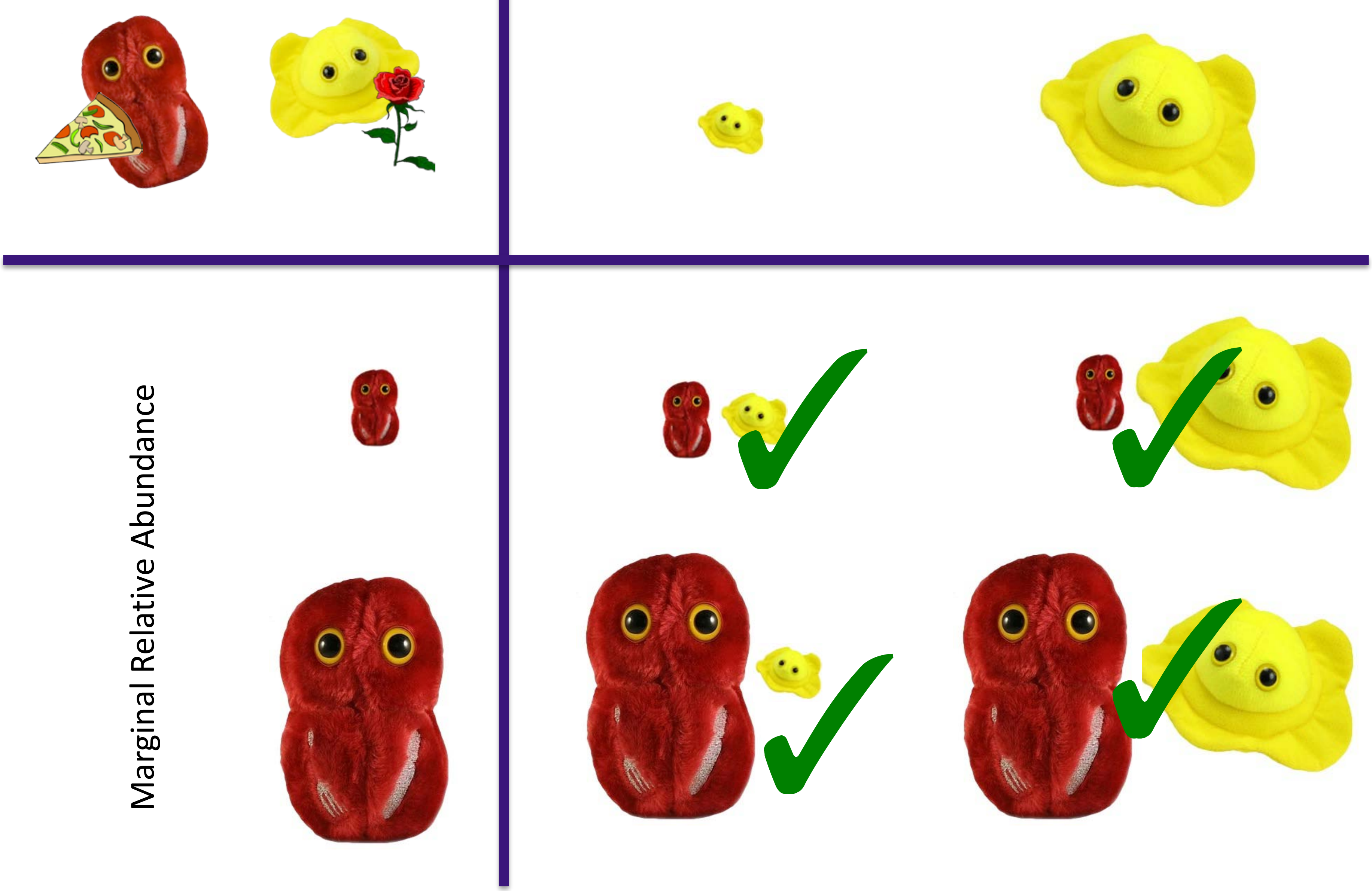
Marginal Relative Abundance

Co-Exclusion



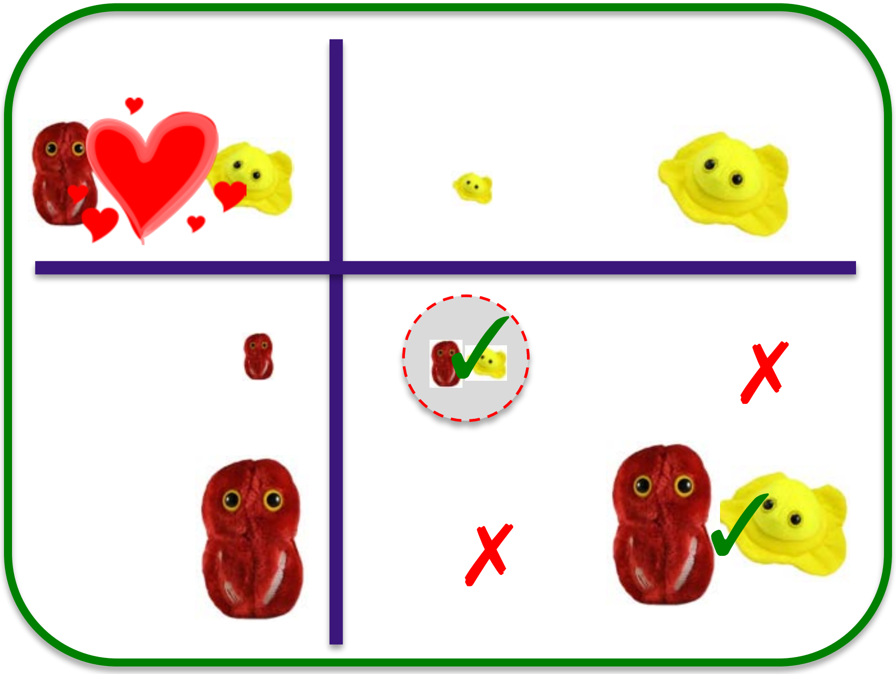
Indifference

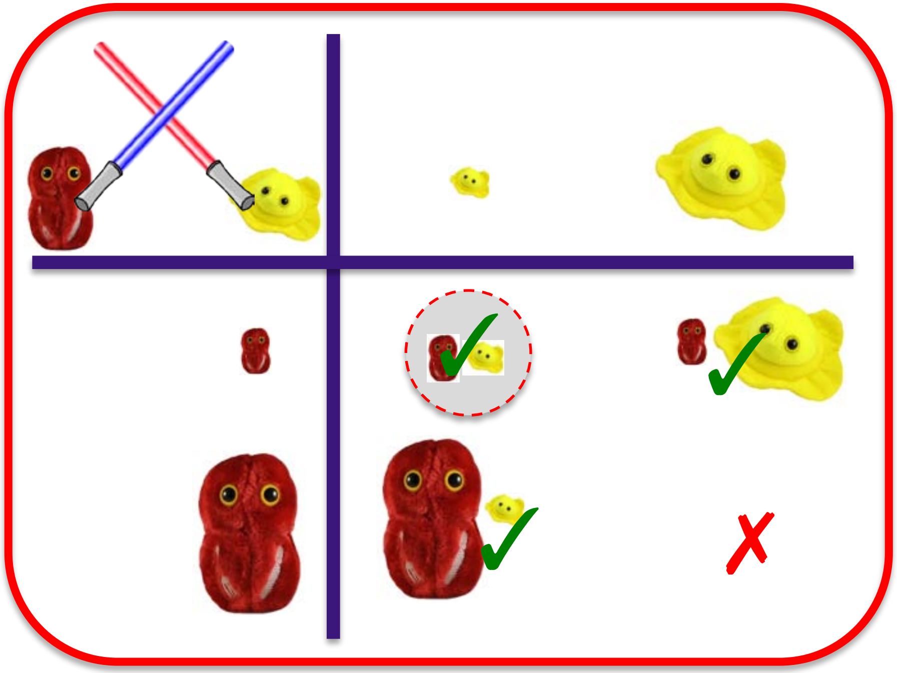
Marginal Relative Abundance

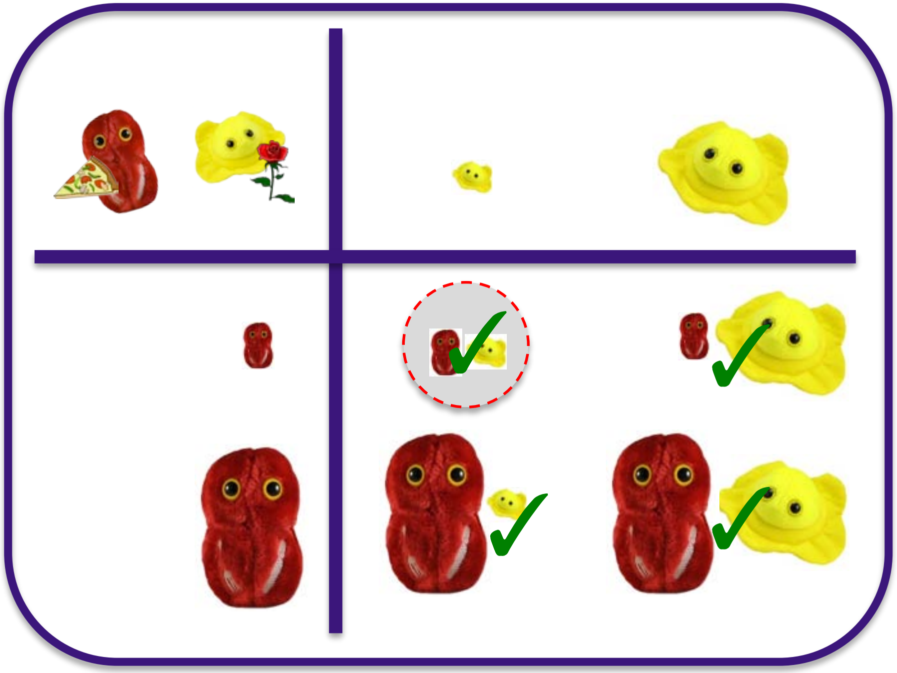


Marginal Relative Abundance

Let’s recap…





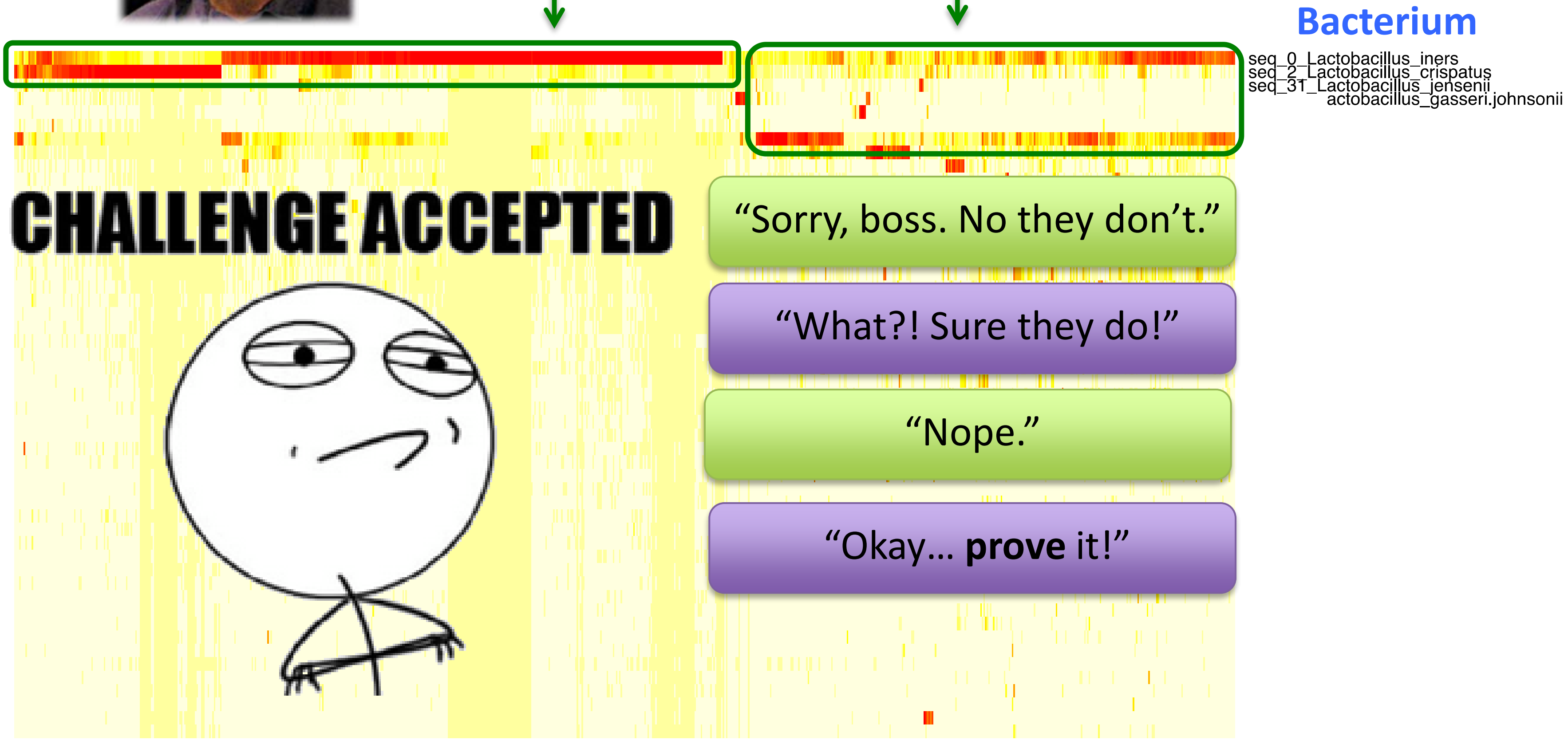


These are __subtle__ differences!

Human Vaginal Microbiome

“Hey! Look at these *great* results! They show **obvious** co-exclusion!” *

(normal *vs.* disease state)



“Sorry, boss. No they don’t.”

“What?! Sure they do!”

“Nope.”

“Okay… **prove** it!”

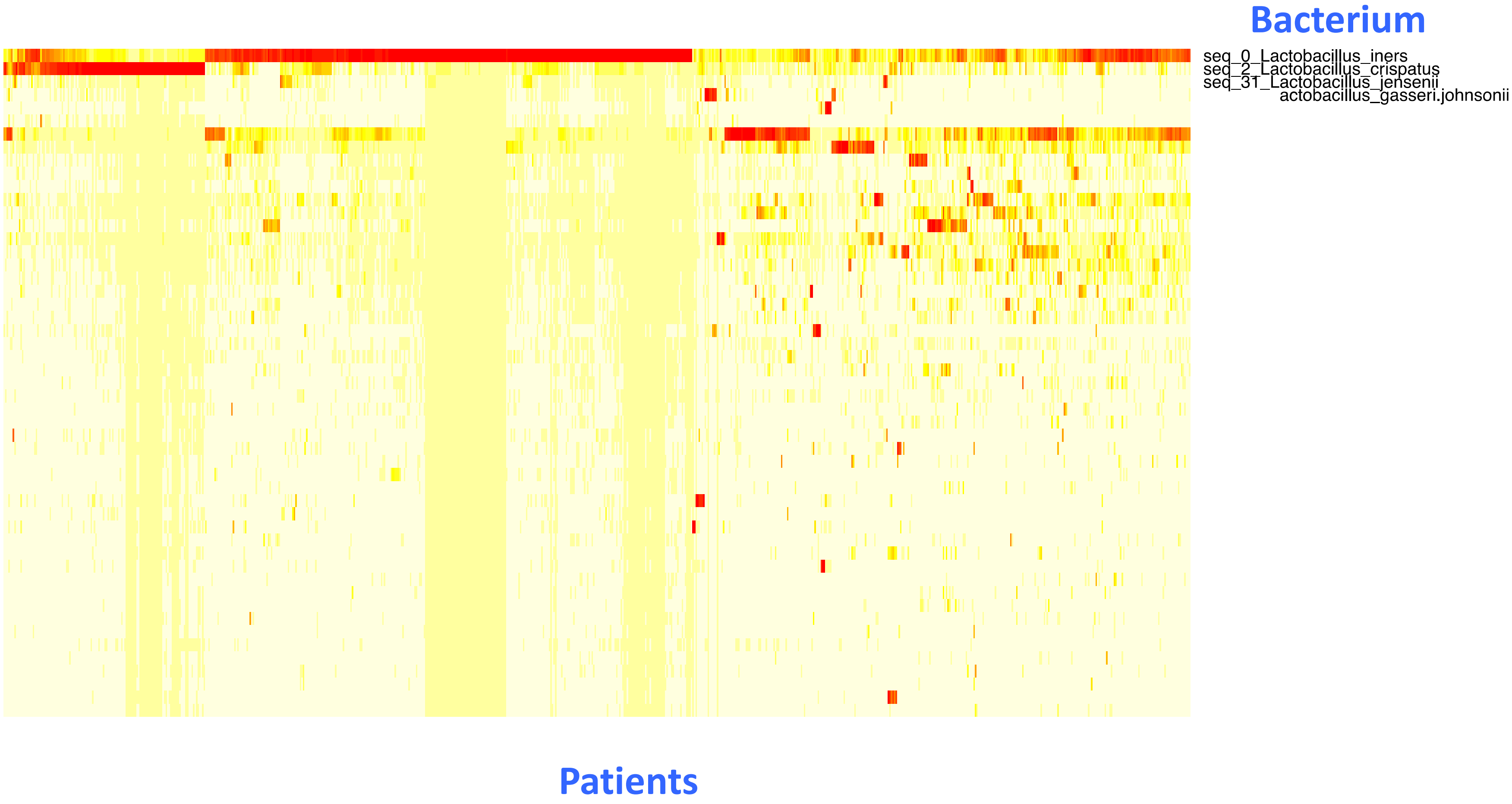
- Gregory B. Gloor, Ruben Hummelen, Jean M. E. Macklaim, **Andrew D. Fernandes**, and Gregor Reid (2010, *Accepted*) Community Microbiome Profiling by Combinatorial Barcoding with Illumina Sequencing. *PLoS One* (Manuscript PONE-D-10-00044R1). Archived at ⟨http://arxiv.org/abs/1007.5075v1⟩.

- Ruben Hummelen, **Andrew D. Fernandes**, Jean M. E. Macklaim, Russell J. Dickson, John Changalucha, Gregory B. Gloor, and Gregor Reid (2010, *Accepted*) Deep Sequencing of the Vaginal Microbiota of Women With HIV. *PLoS One* (Manuscript 10-PONE-RA-19937).

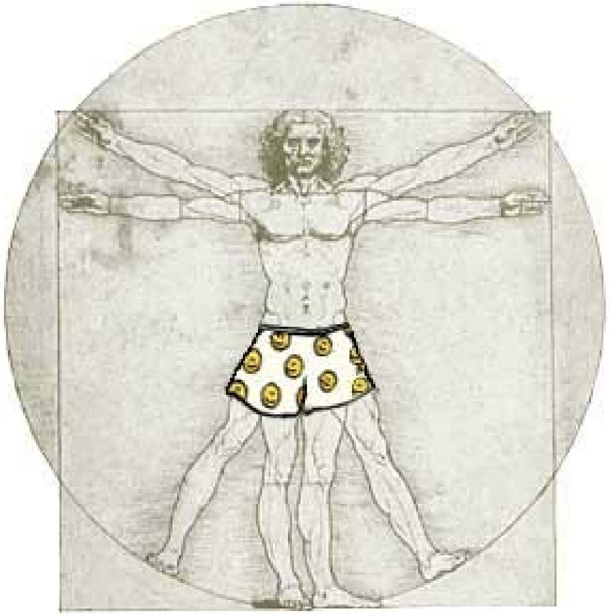
*Dramatic Re-interpretation

Where do results like this come from?

Claim: Zeros, Logarithms, and PCA!



Zeros



Clinical Subject

(well, female...)

Swab Sample

PCR

Multiplex Next-Gen Sequence

ion torrent

(actually Illumina & Solid...)

	Lactobacillus iners	*Lactobacillus crispatus*	*Lactobacillus jensenii*	*Lactobacillus gasseri*
Patient 1	49679	3177	21389	135
Patient 2	7755	29752	989	368
Patient 3	3286	5955	549	397
Patient 4	2265	3263	13742	148
Patient 5	10239	2926	226	100
Patient 6	16376	20706	1037	79
Patient 7	27313	4878	5320	92
Patient 8	33006	1103	1186	176
Patient 9	20504	1771	346	161

V6 Read Counts

LONDON REGIONAL GENOMICS CENTRE

Western

LAWSON HEALTH RESEARCH INSTITUTE

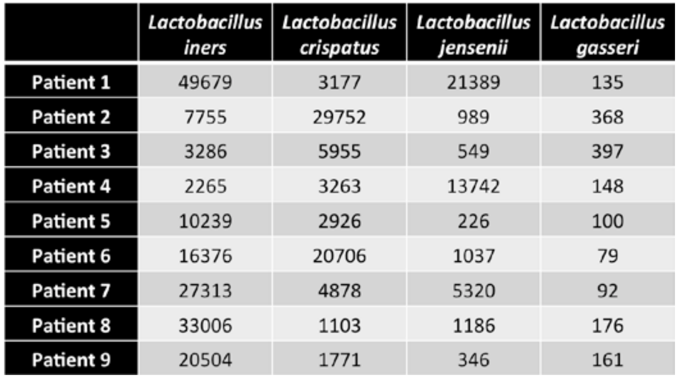
robarts RESEARCH

Counts *between samples (patients)* are **meaningless**!

We are interested in sample **proportions**, only!

	Lactobacillus iners	***Lactobacillus crispatus***	***Lactobacillus jensenii***	***Lactobacillus gasseri***	…
Patient 1	49679	3177	21389	135	…
Patient 2	7755	29752	989	368	…
Patient 3	3286	5955	549	397	…
Patient 4	2265	3263	13742	148	…
Patient 5	10239	2926	226	100	…
Patient 6	16376	20706	1037	79	…
Patient 7	27313	4878	5320	92	…
Patient 8	33006	1103	1186	176	…
Patient 9	20504	1771	346	161	…
⋮	⋮	⋮	⋮	⋮	

But proportions and counts are kind-of-almost the same thing, aren't they?



	Lactobacillus iners	*Lactobacillus crispatus*	*Lactobacillus jensenii*	*Lactobacillus gasseri*
Patient 1	49679	3177	21389	135
Patient 2	7755	29752	989	368
Patient 3	3286	5955	549	397
Patient 4	2265	3263	13742	148
Patient 5	10239	2926	226	100
Patient 6	16376	20706	1037	79
Patient 7	27313	4878	5320	92
Patient 8	33006	1103	1186	176
Patient 9	20504	1771	346	161

$$p_i \approx \frac{n_i}{\sum_j n_j}$$

Nope.

ONLY VALID IF n_i IS NOT SMALL!

So where does $p_i \approx \frac{n_i}{\sum_j n_j}$ come from?

Consider a biome with only **two** species.

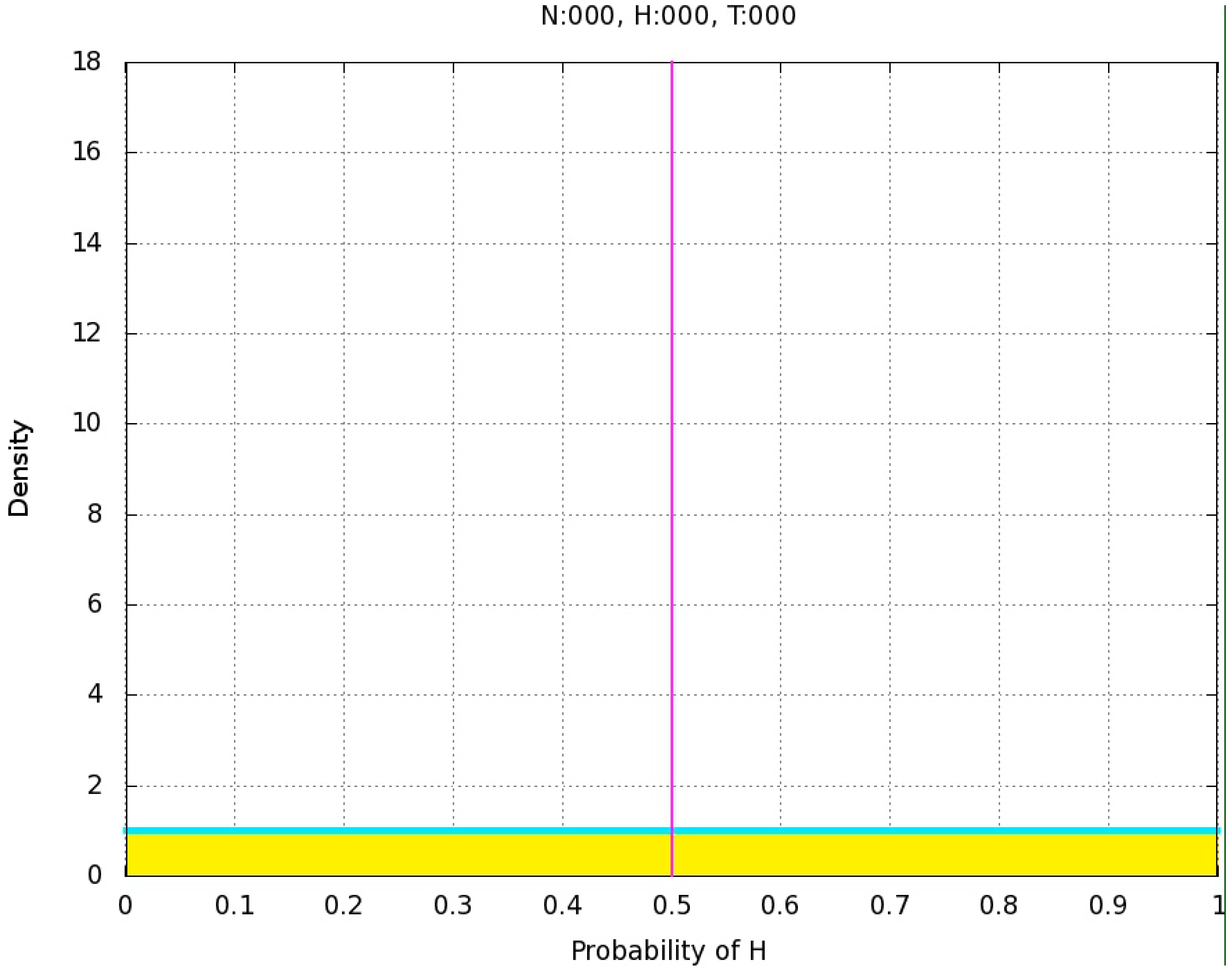
$$\Pr\left(p_H, p_T | n_H, n_T\right) = \frac{(n_H + n_T)!}{n_H!\, n_T!}\, p_H^{n_H}\, p_T^{n_T}$$

$$\begin{aligned}
\log\left[\Pr\left(p_H, p_T | n_H, n_T\right)\right] &= \log\left[\tfrac{(n_H+n_T)!}{n_H!\,n_T!}\right] + n_H \log\left(p_H\right) + n_T \log\left(p_T\right) \\
&= \log\left[\tfrac{(n_H+n_T)!}{n_H!\,n_T!}\right] + n_H \log\left(p_H\right) + \left(n - n_H\right) \log\left(1 - p_H\right)
\end{aligned}$$

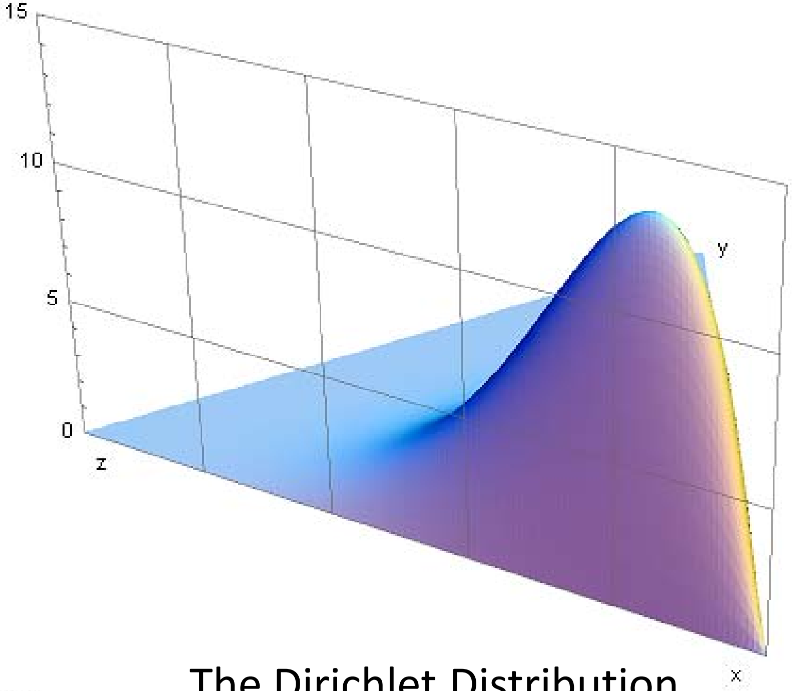
$$\partial / \partial p_H = 0 \qquad \Rightarrow \qquad \frac{n_H}{p_H} - \frac{n - n_H}{1 - p_H} = 0$$

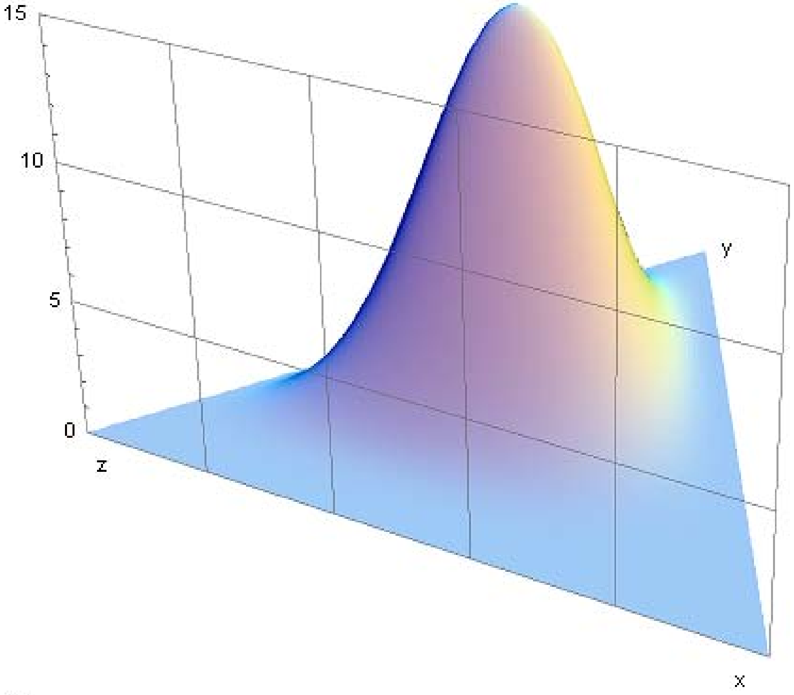
$$\Rightarrow \qquad \Pr\left(p_H | n, n_H\right) \text{ is maximzed when } p_H = \frac{n_H}{n}$$

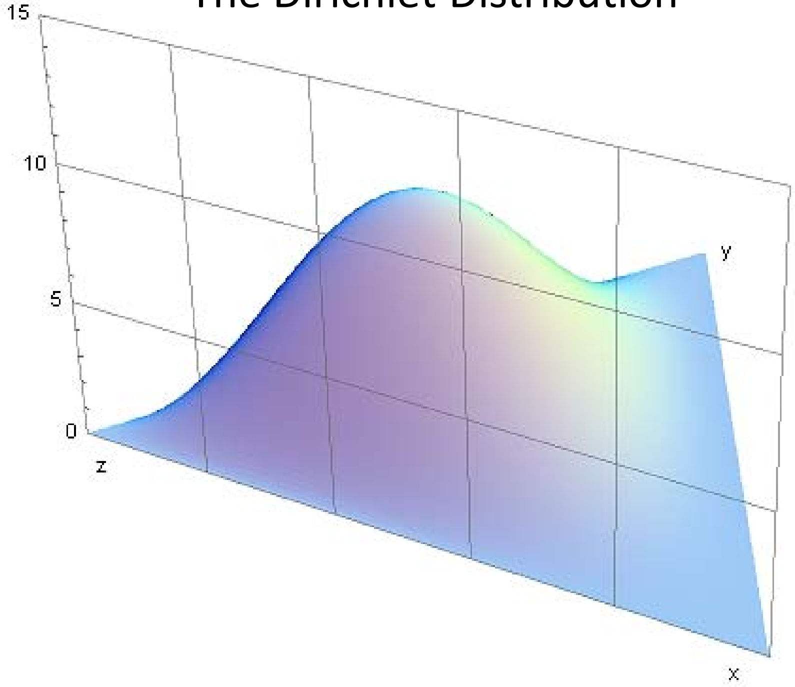
That's a **lot** of math for something that most people think is rather obvious…

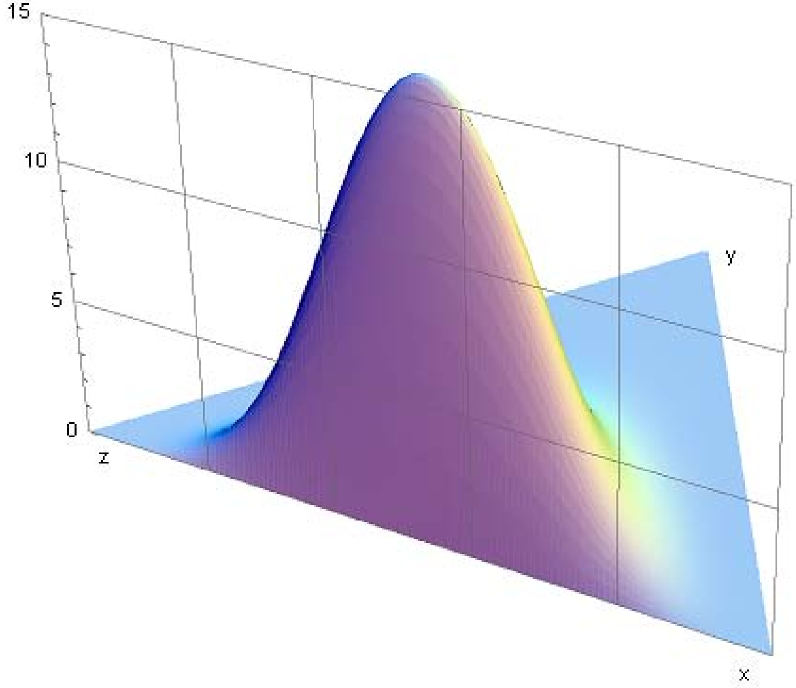

N:000, H:000, T:000
Density
Probability of H
18
16
14
12
10
8
6
4
2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

The Dirichlet Distribution









The **point**… mathematically, even though

$$p_i \approx \frac{n_i}{\sum_j n_j} \qquad n_i = 0 \nRightarrow p_i = 0$$

There is a **enormous** difference between a *biological* zero and a *mathematical* zero!

(Distinguishing “rare” *vs.* “impossible” is almost the entire basis of Shannon’s Theory of Information…)

$$\mathbb{E}\left[\boxed{\log(p_i)}\right] = \psi(\alpha_i) + \psi\left(\sum_j \alpha_j\right)$$

(Never Zero!)

$$\psi(z) = \frac{d}{dz}\ln\Gamma(z) \qquad \alpha_i = n_i + \frac{1}{2}$$

You can think of this as a “pseudo-count”

(You would be completely wrong, but at least you would feel comforted…)

Logarithms

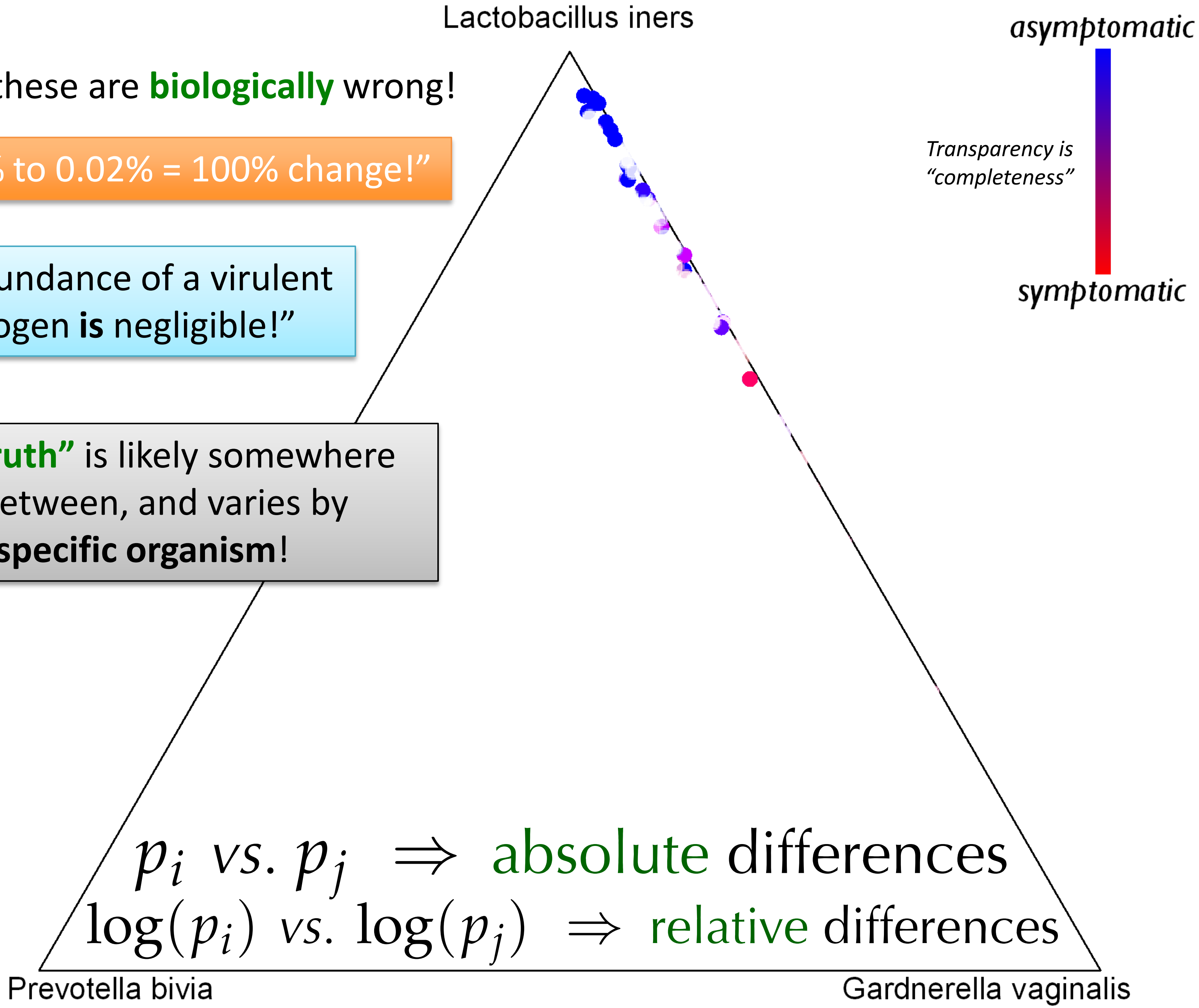
Why use $\log(p_i)$?

Both of these are **biologically** wrong!

"0.01% to 0.02% = 100% change!"

"1% abundance of a virulent pathogen **is** negligible!"

The **"truth"** is likely somewhere in-between, and varies by **specific organism**!

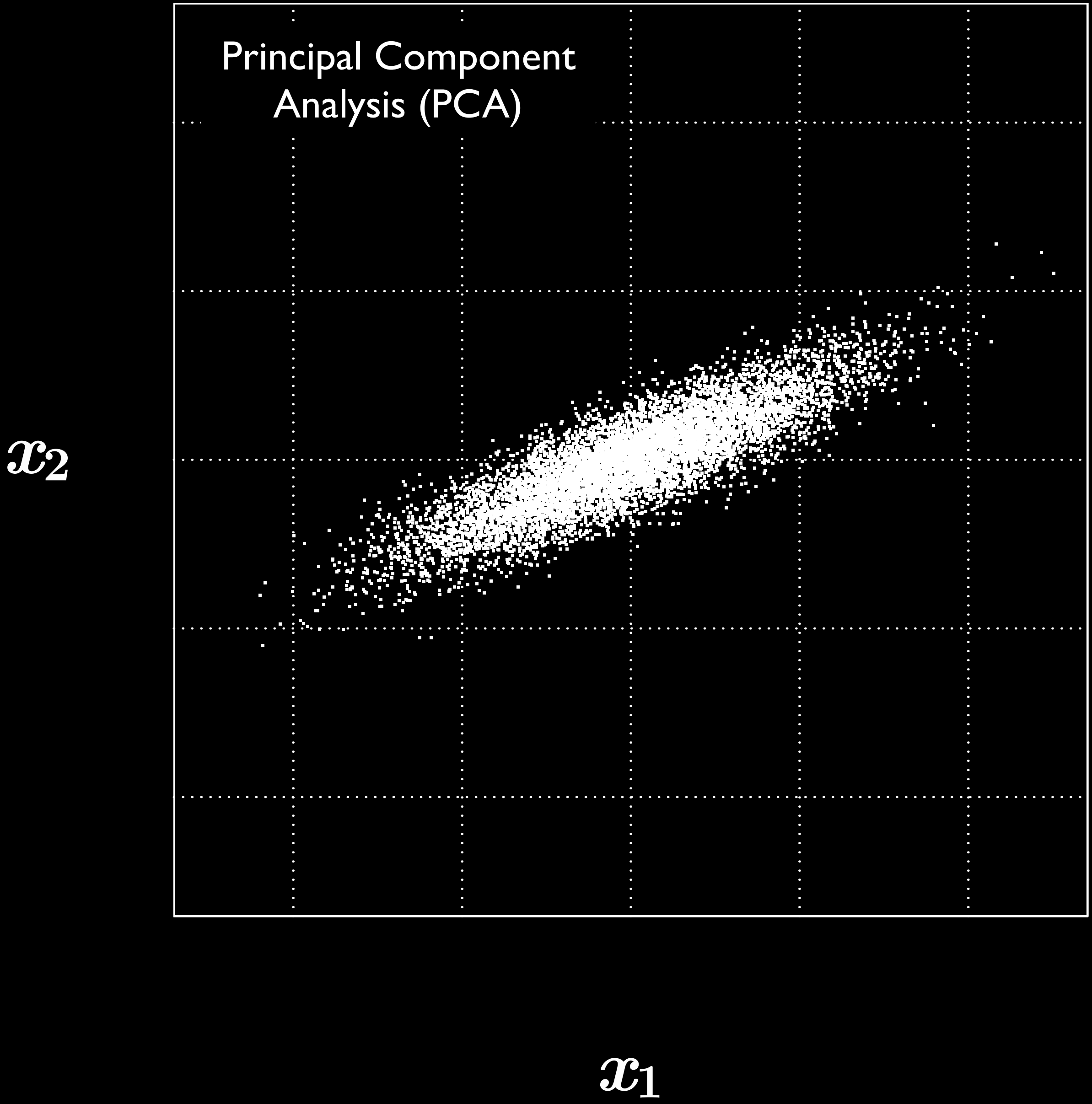


$p_i \ vs. \ p_j \Rightarrow$ absolute differences

$\log(p_i) \ vs. \ \log(p_j) \Rightarrow$ relative differences

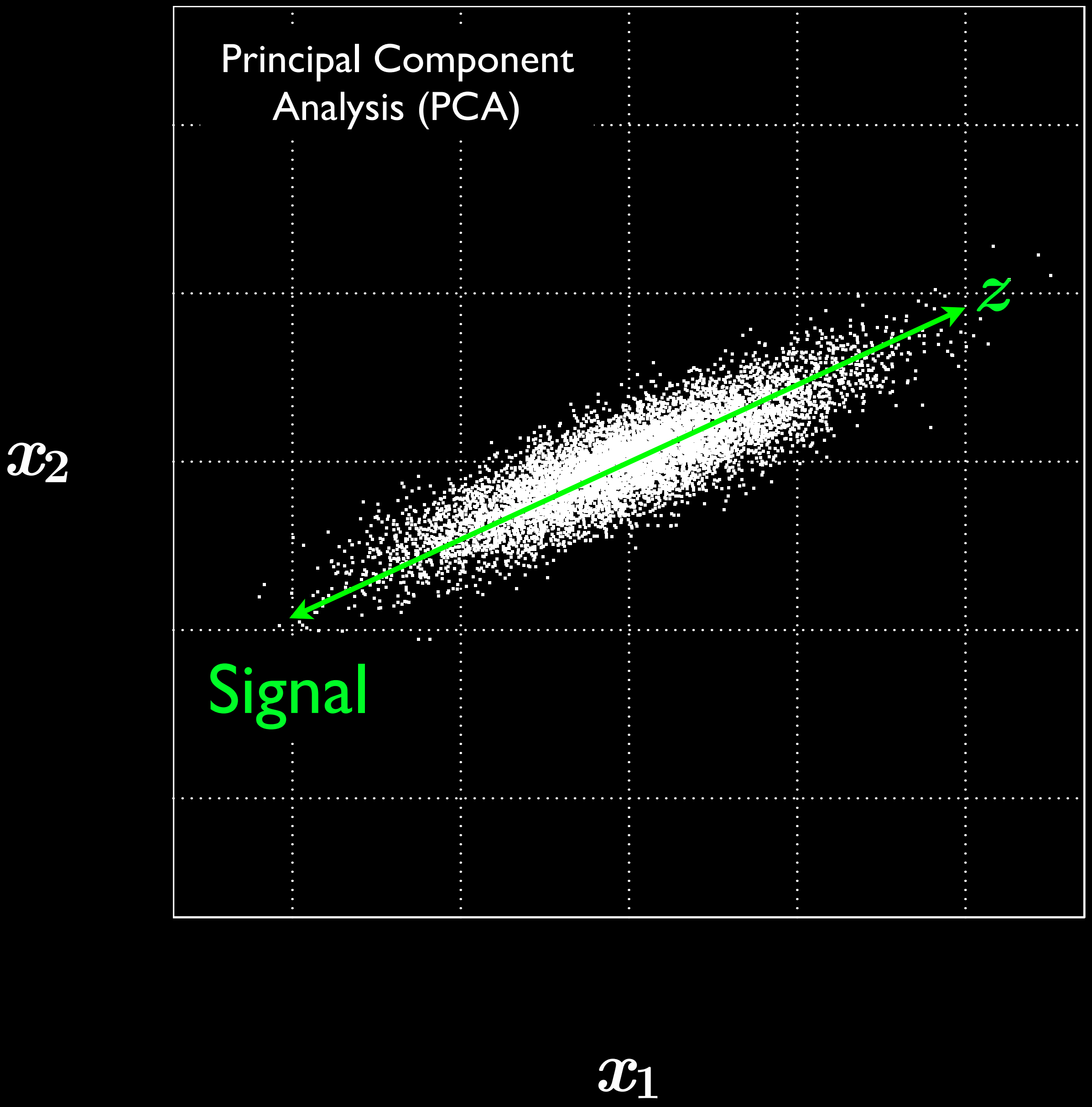
Principal Component Analysis

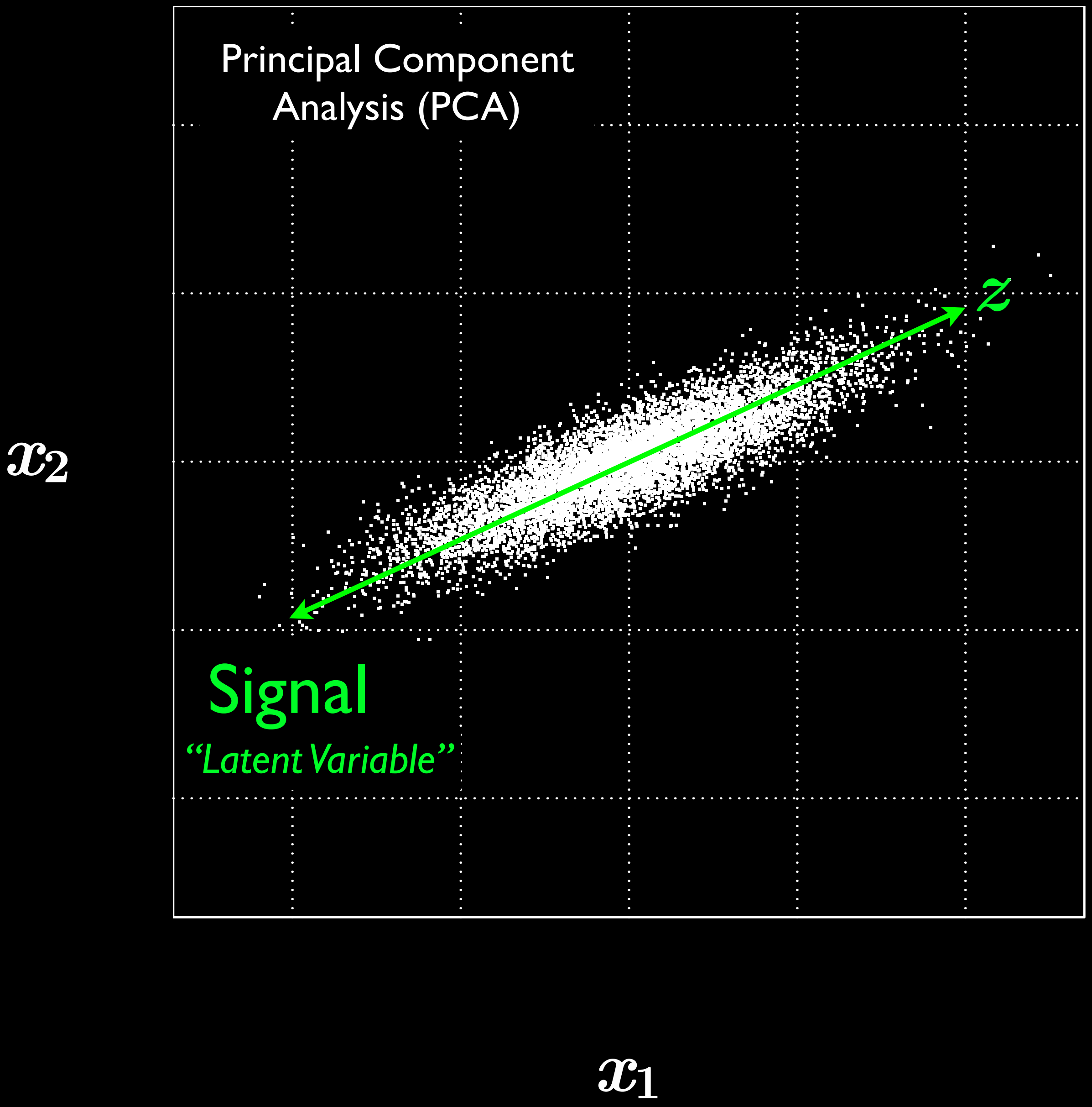
Why use $\log(p_i)$?



Principal Component Analysis (PCA)

Principal Component Analysis (PCA)



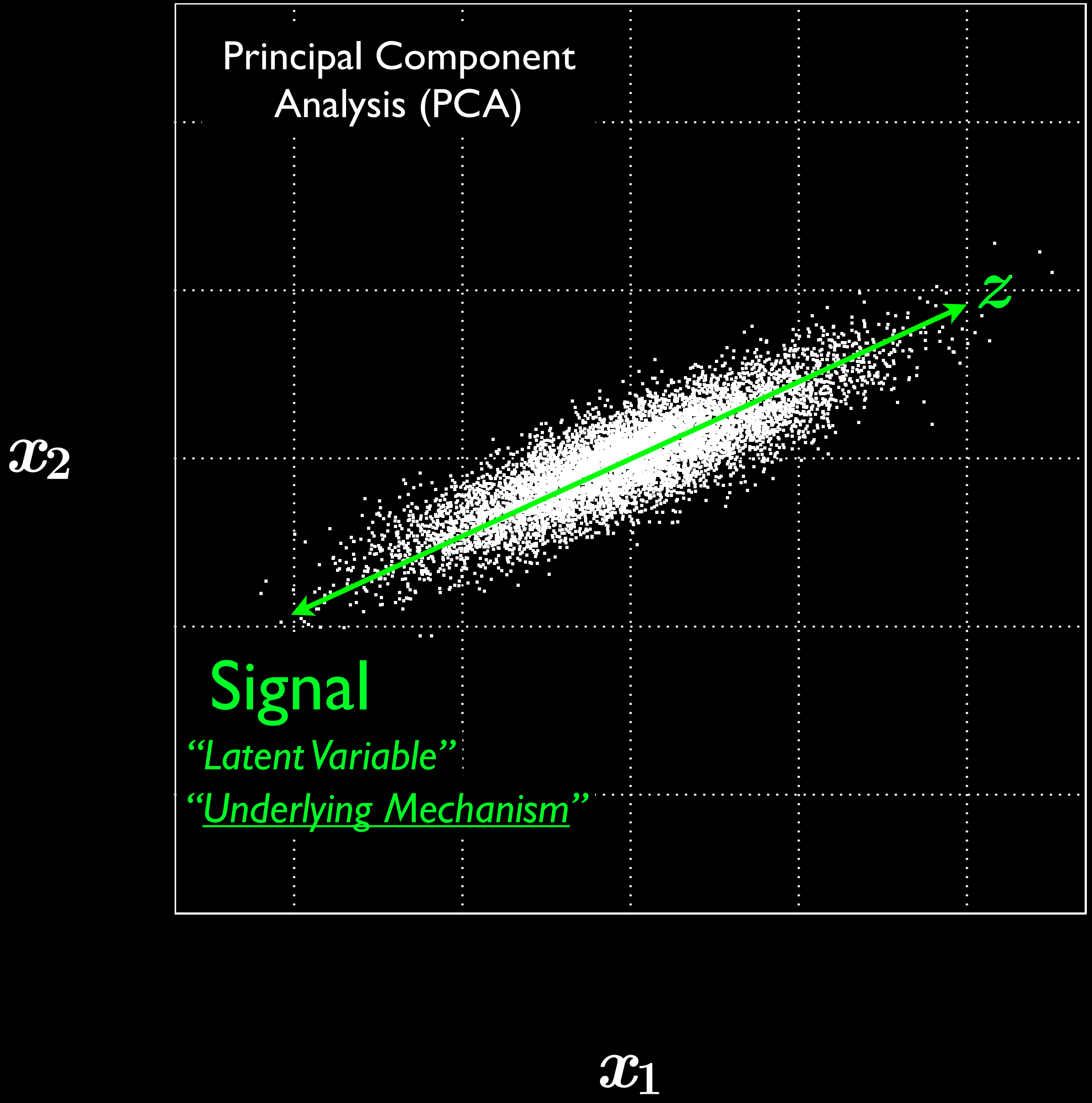


Principal Component Analysis (PCA)

Signal

“Latent Variable”

Principal Component Analysis (PCA)

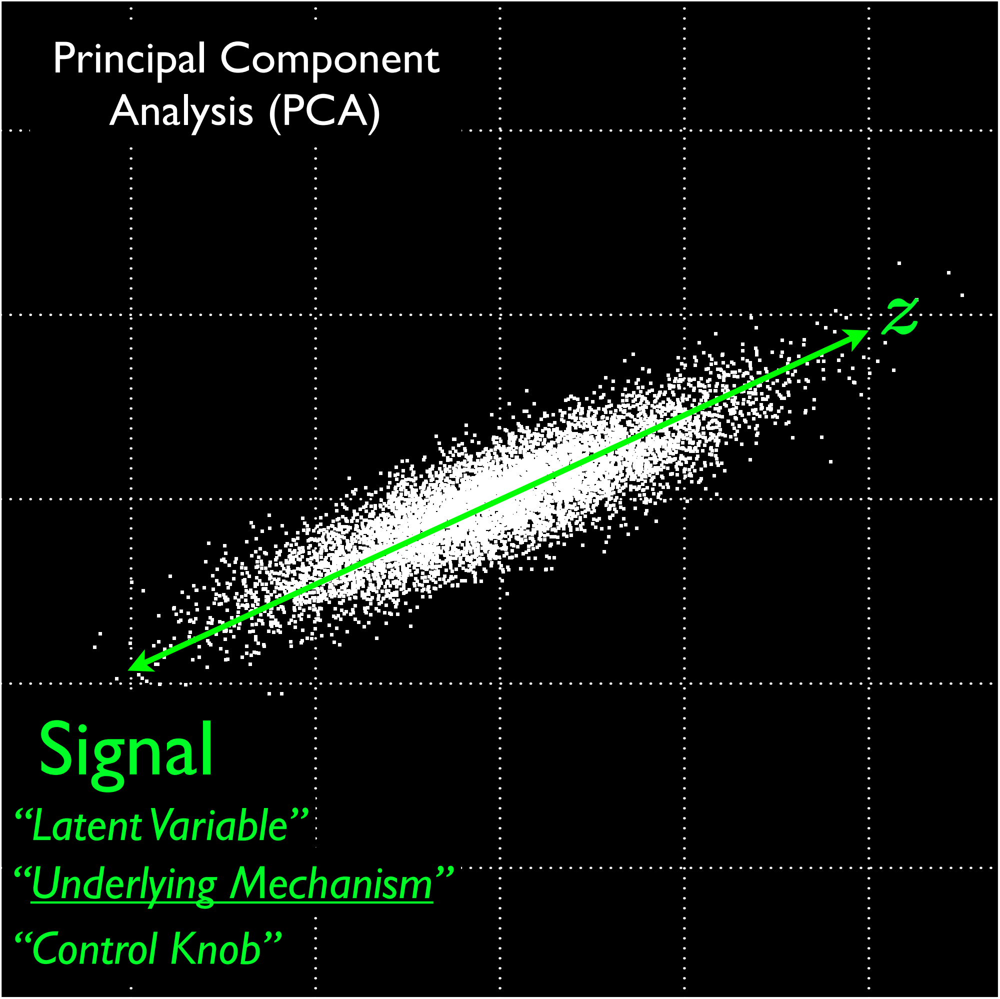


Signal

“Latent Variable”

“Underlying Mechanism”

Principal Component Analysis (PCA)



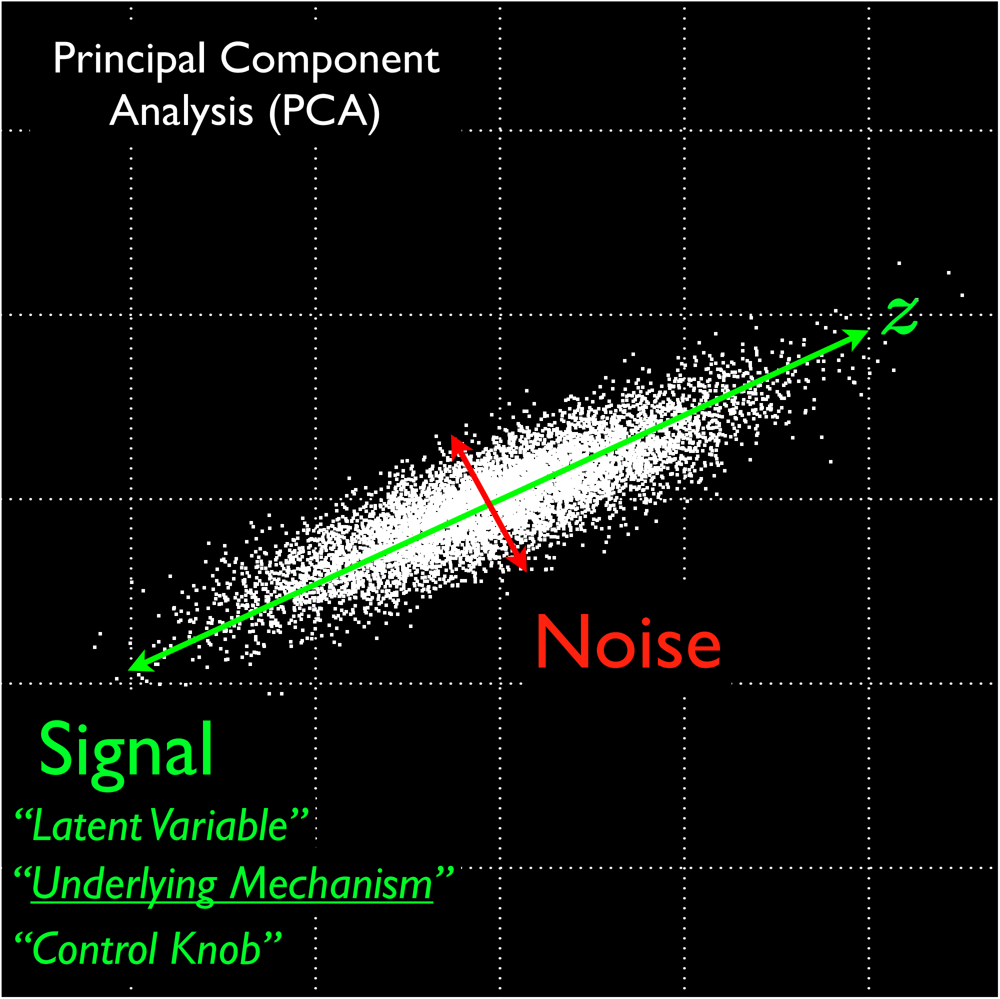
Signal

"Latent Variable"

"Underlying Mechanism"

"Control Knob"

Principal Component Analysis (PCA)



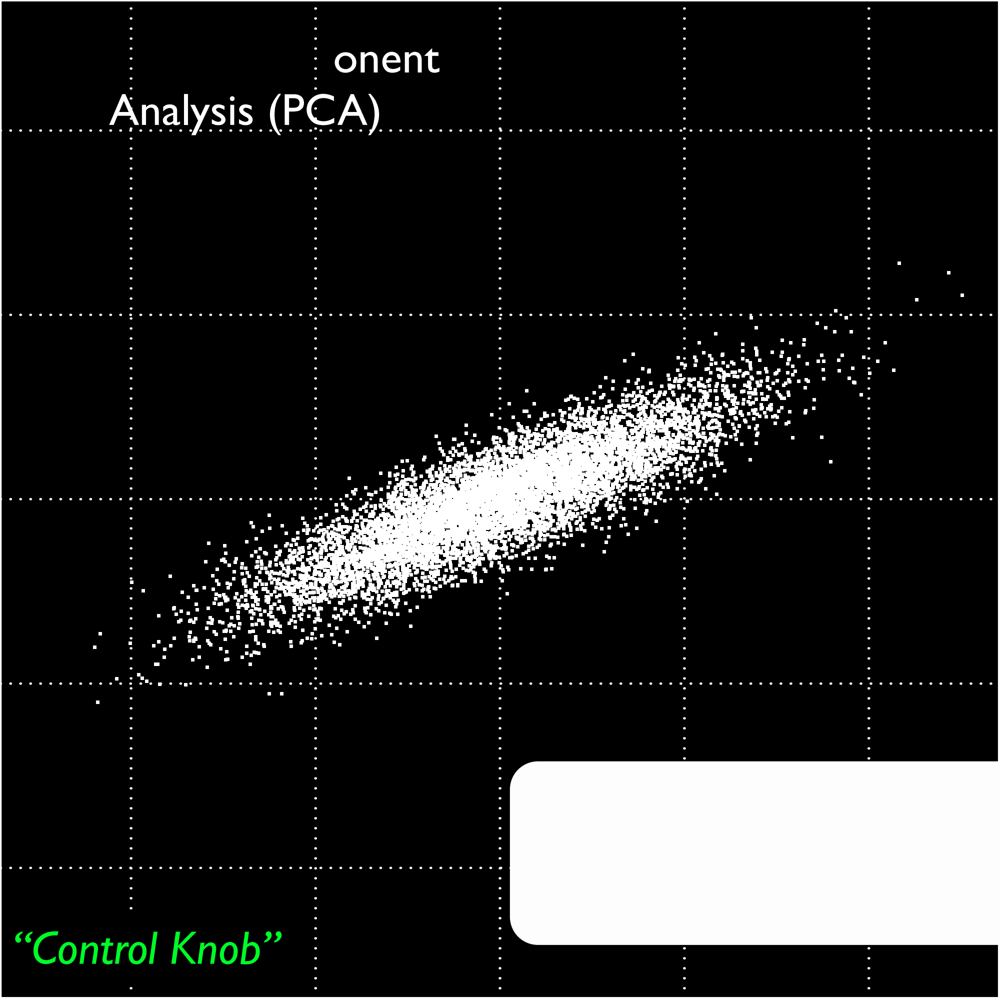
Noise

Signal

“Latent Variable”

“Underlying Mechanism”

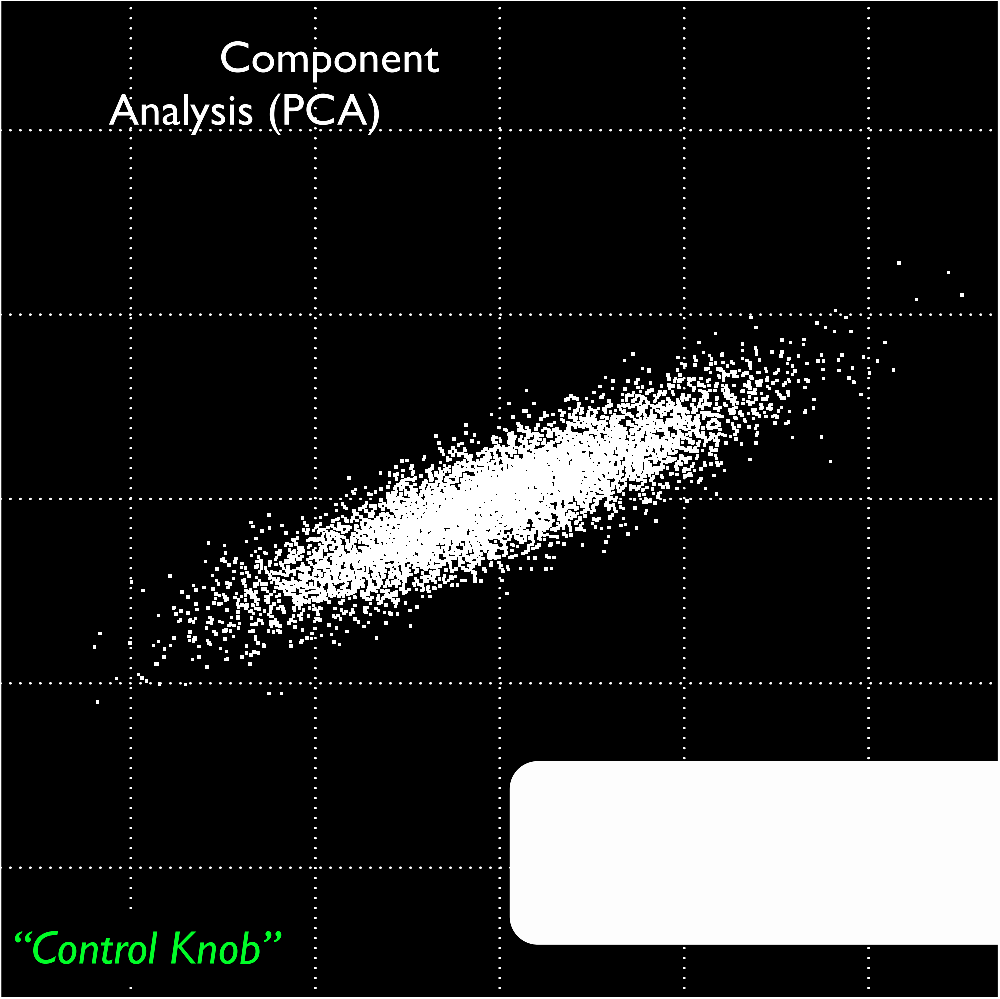
“Control Knob”



onent

Analysis (PCA)

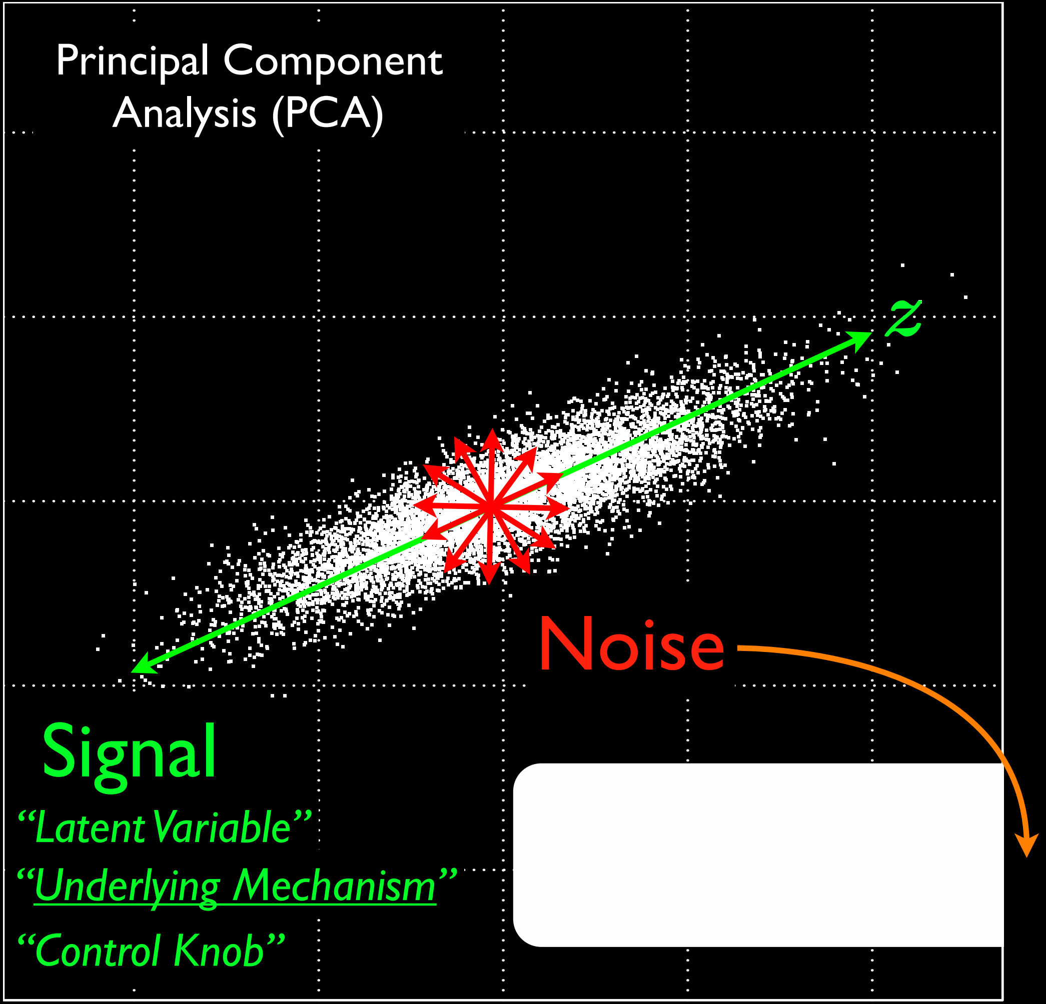
“Control Knob”



Component
Analysis (PCA)

“Control Knob”

Principal Component Analysis (PCA)



Signal

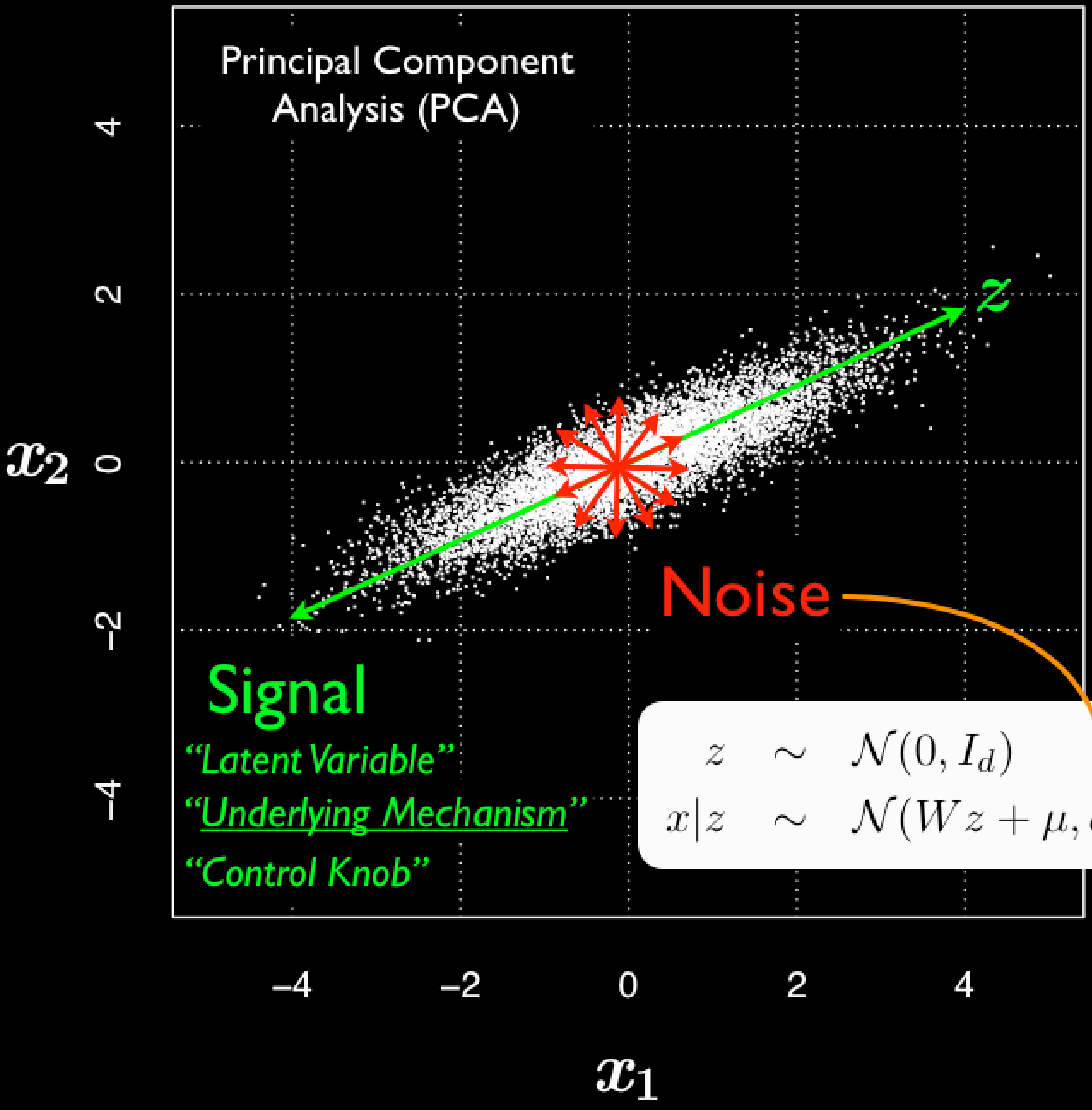
“Latent Variable”

“<u>Underlying Mechanism</u>”

“Control Knob”

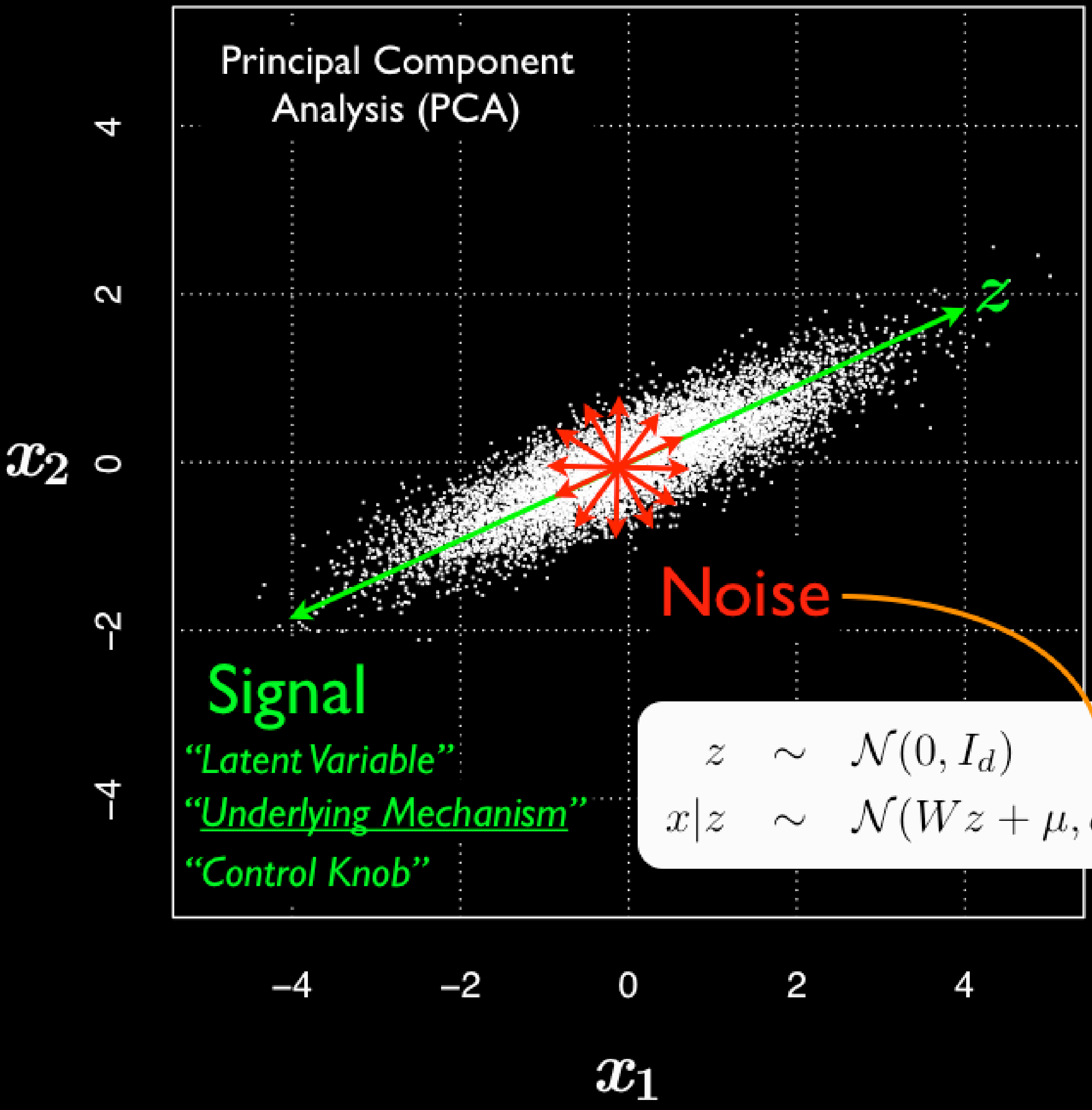
The PCA variance decomposition has **strange** biological implications:

① Every fraction has the **same** proportion of stochastic variation



$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \\ x|z &\sim \mathcal{N}(Wz+\mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md} \end{aligned}$$

Tipping and Bishop (1999)



Principal Component Analysis (PCA)

Signal

"Latent Variable"

"Underlying Mechanism"

"Control Knob"

Noise

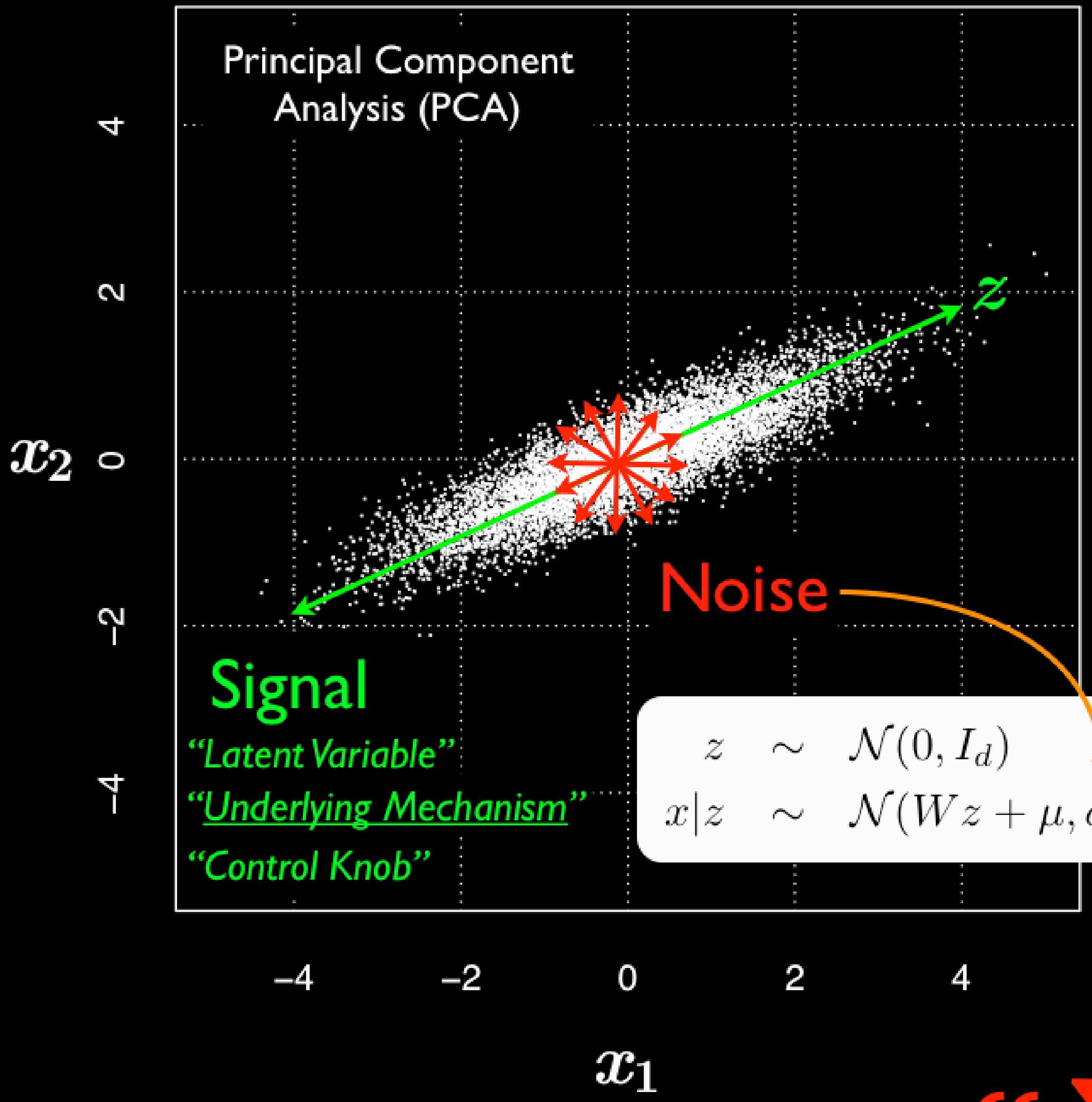
$$
\begin{aligned}
z &\sim \mathcal{N}(0, I_d) \\
x|z &\sim \mathcal{N}(Wz + \mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md}
\end{aligned}
$$

Tipping and Bishop (1999)

The PCA variance decomposition has **strange** biological implications:

① Every fraction has the **same** proportion of stochastic variation

② Results depend on the units of measurement



Noise

Signal

“Latent Variable”

“Underlying Mechanism”

“Control Knob”

The PCA variance decomposition has **strange** biological implications:

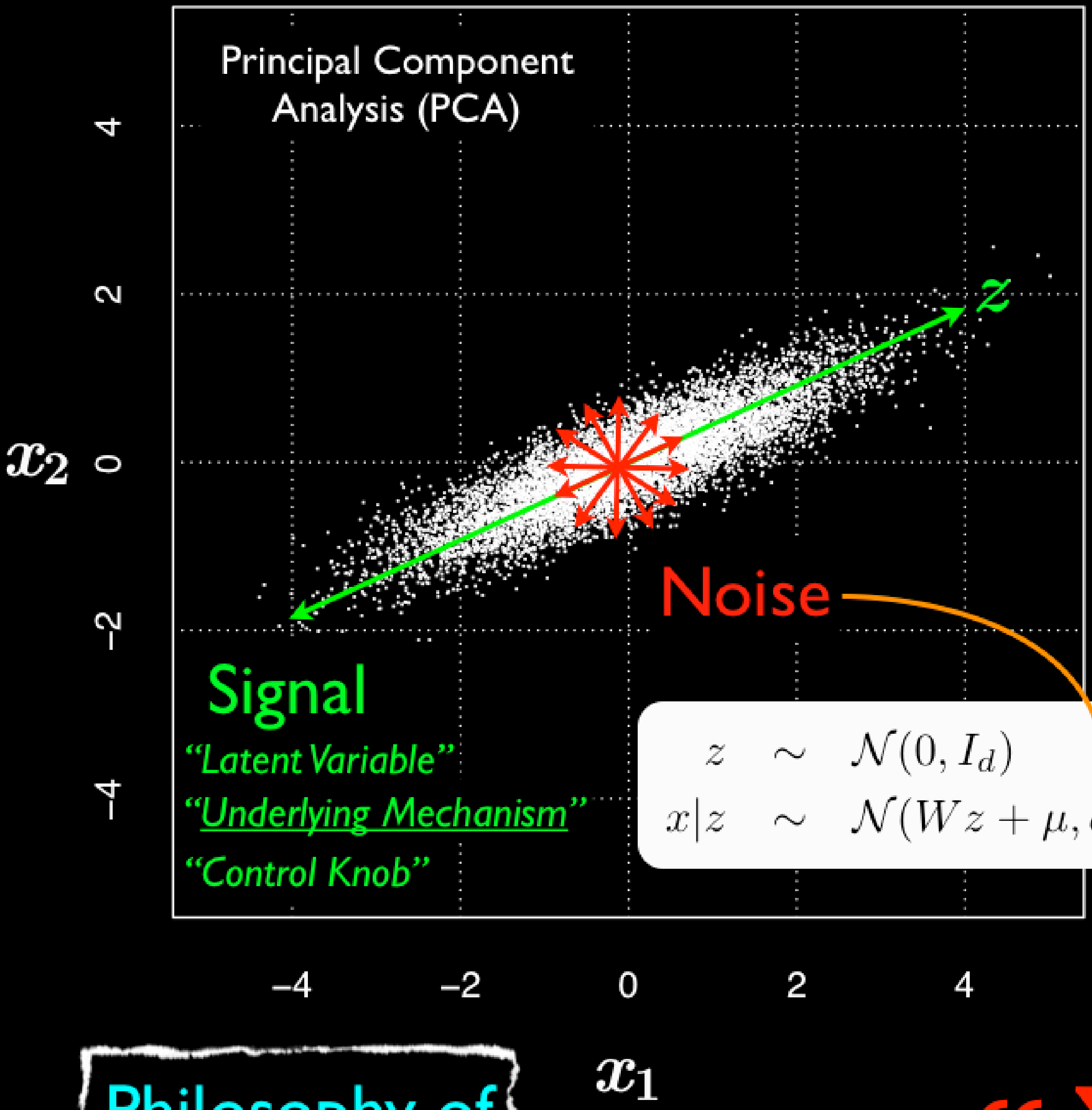
① Every fraction has the **same** proportion of stochastic variation

② Results depend on the units of measurement

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \\ x|z &\sim \mathcal{N}(Wz+\mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md} \end{aligned}$$

Tipping and Bishop (1999)

“ Σ or ρ ” ?



Noise

Signal

"Latent Variable"

"Underlying Mechanism"

"Control Knob"

$$
\begin{aligned}
z &\sim \mathcal{N}(0, I_d) \\
x|z &\sim \mathcal{N}(Wz+\mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md}
\end{aligned}
$$

Tipping and Bishop (1999)

The PCA variance decomposition has **strange** biological implications:

① Every fraction has the **same** proportion of stochastic variation

② Results depend on the units of measurement

Philosophy of Modelling

" Σ or ρ " ?

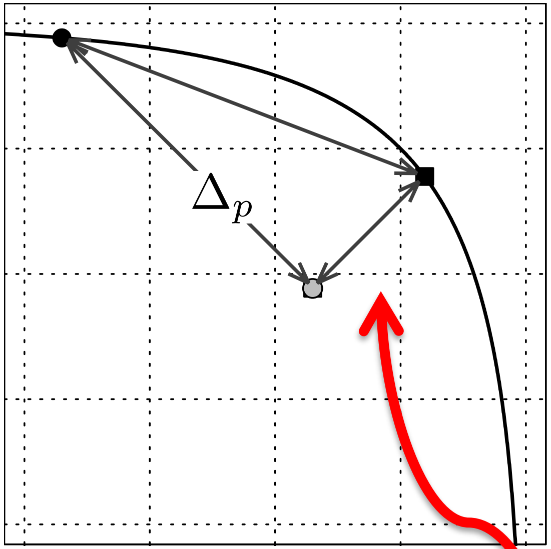
PCA also introduces **systematic distortions** in the analysis!

$$\begin{bmatrix} p_i \\ q_i \\ r_i \end{bmatrix} \Rightarrow \begin{bmatrix} p_i/(p_i+q_i+r_i) \\ q_i/(p_i+q_i+r_i) \\ r_i/(p_i+q_i+r_i) \end{bmatrix}$$

$$\begin{bmatrix} \log(p_i/(p_i+q_i+r_i)) \\ \log(q_i/(p_i+q_i+r_i)) \\ \log(r_i/(p_i+q_i+r_i)) \end{bmatrix} \Rightarrow \begin{bmatrix} \log(p_i)-\log(p_i+q_i+r_i) \\ \log(q_i)-\log(p_i+q_i+r_i) \\ \log(r_i)-\log(p_i+q_i+r_i) \end{bmatrix}$$

$$\begin{bmatrix} \log(p_i)-\log(p_i+q_i+r_i) \\ \log(q_i)-\log(p_i+q_i+r_i) \\ \log(r_i)-\log(p_i+q_i+r_i) \end{bmatrix} \Rightarrow \begin{bmatrix} \log(p_i) \\ \log(q_i) \\ \log(r_i) \end{bmatrix} - \cancel{k}$$

(arbitrary)

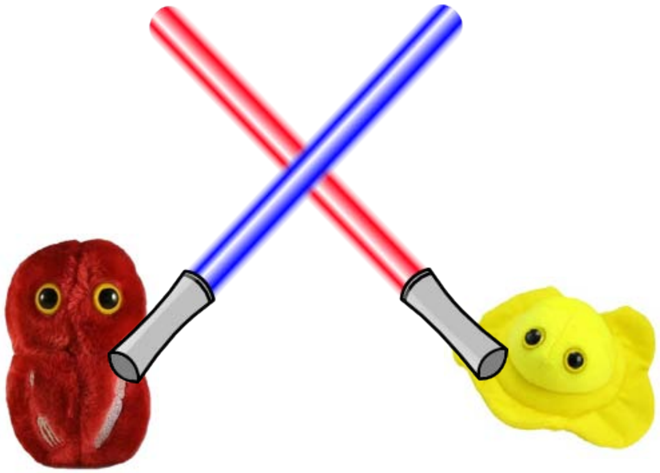


Distortions are **routinely between** **5% to 50% to 500%** of datum distances!

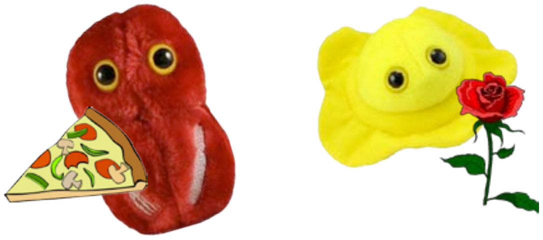
How can we tell *which* of these is occurring?

Co-Occurrence

Co-Exclusion



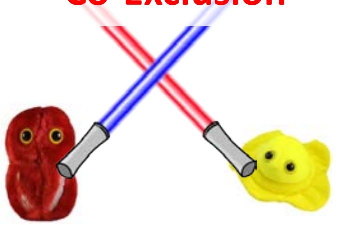
Indifference



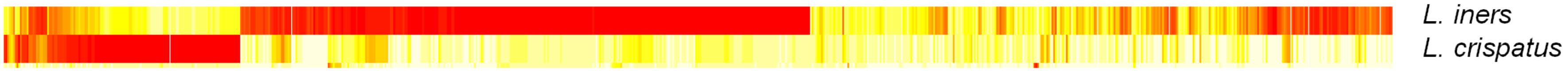
… taking into account …

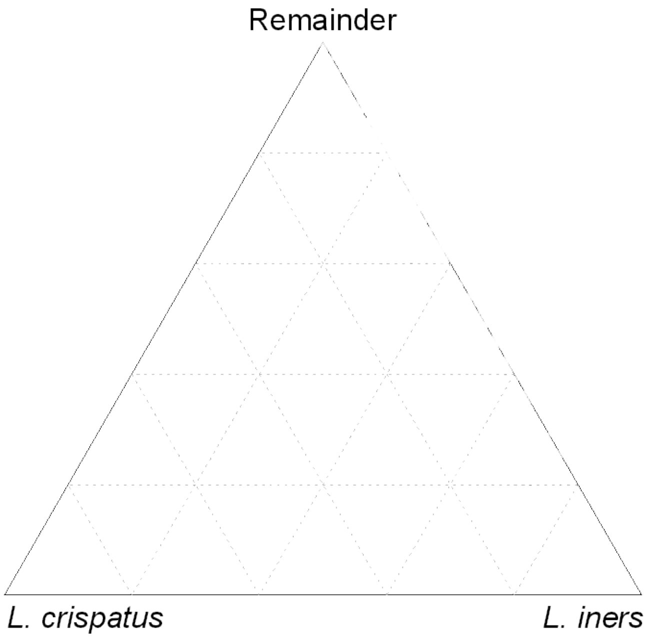
- Zero-counts (not zero-proportions)
- Not-really-absolute and not-really-relative
- Removal of systematic distortions
- **Reparameterization invariance**

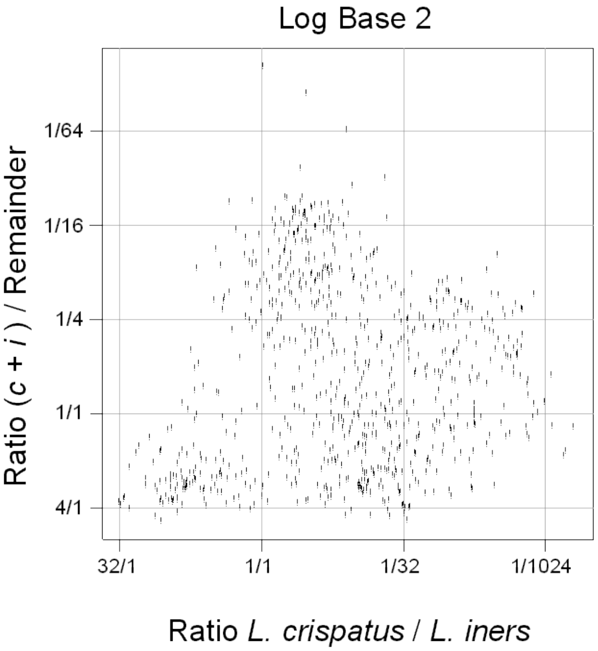
Co-Exclusion

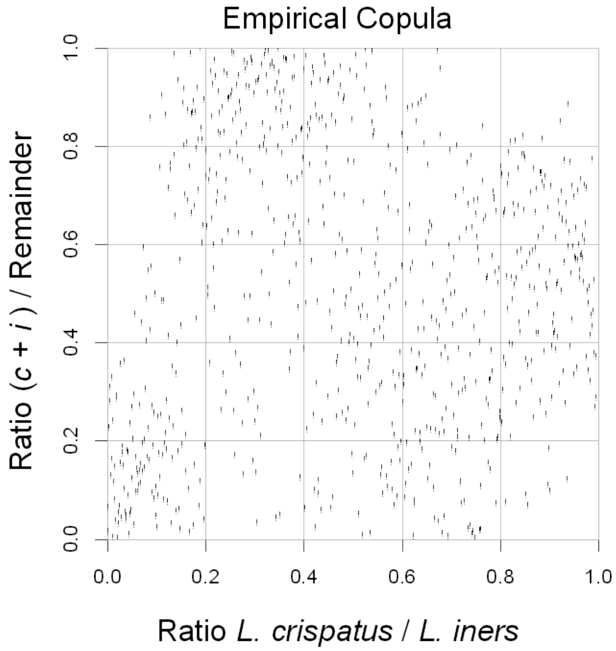


?



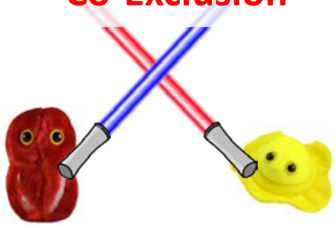




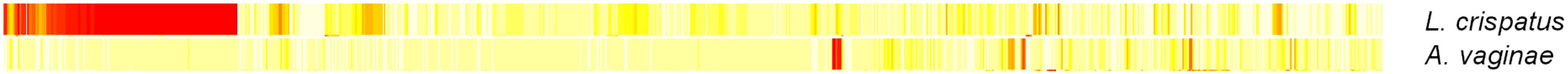


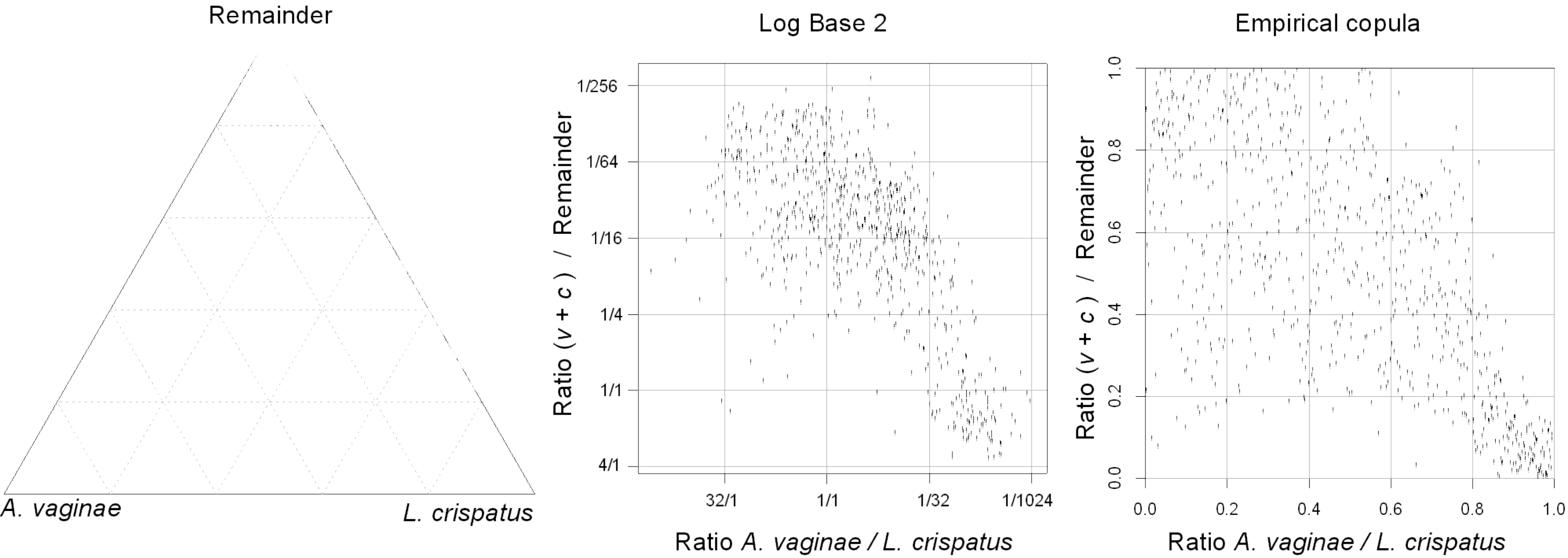
No!

Co-Exclusion

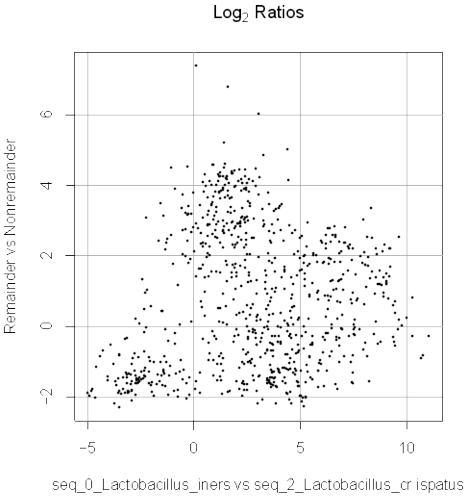


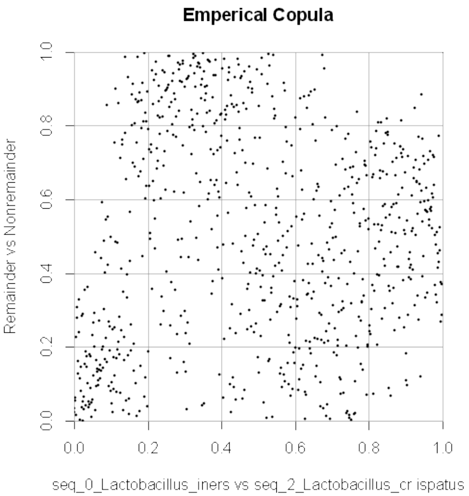
?

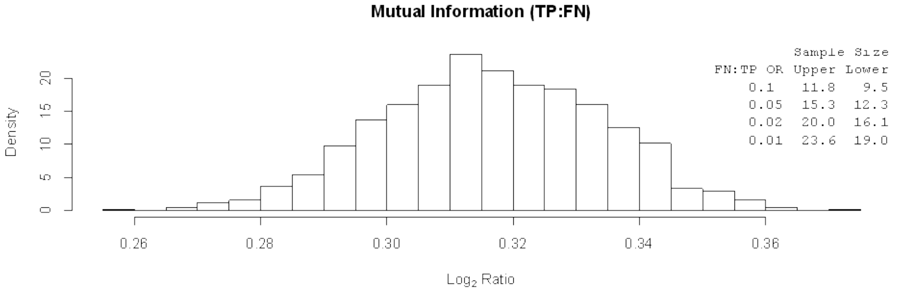




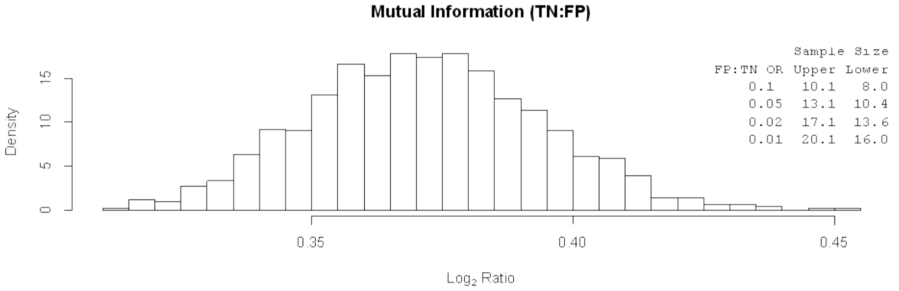
Yes! **(weakly)**





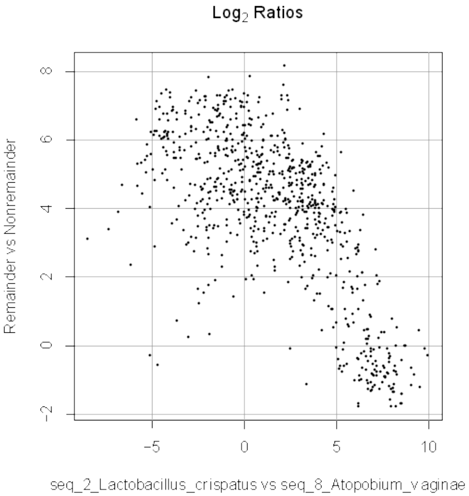


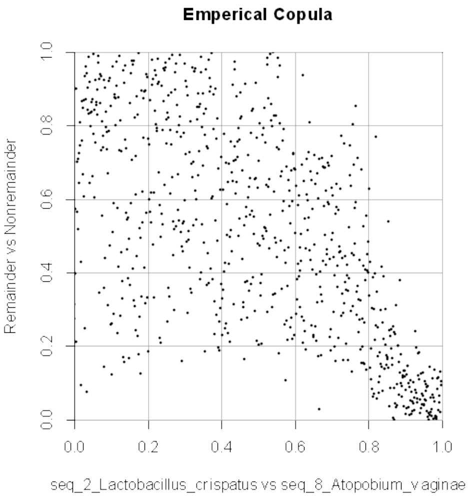
	Sample	Size
FN:TP OR	Upper	Lower
0.1	11.8	9.5
0.05	15.3	12.3
0.02	20.0	16.1
0.01	23.6	19.0

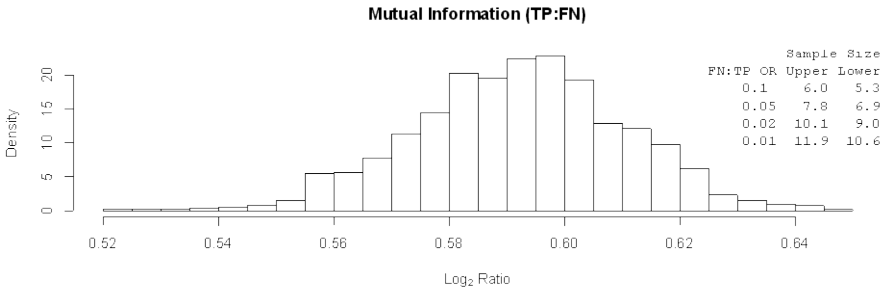


	Sample	Size
FP:TN OR	Upper	Lower
0.1	10.1	8.0
0.05	13.1	10.4
0.02	17.1	13.6
0.01	20.1	16.0

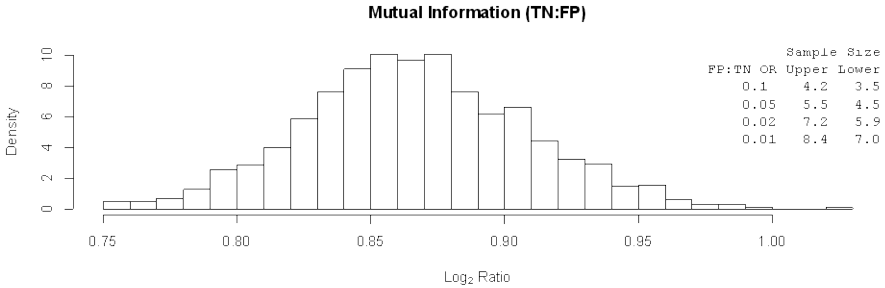
Quantifying “Uniformity”



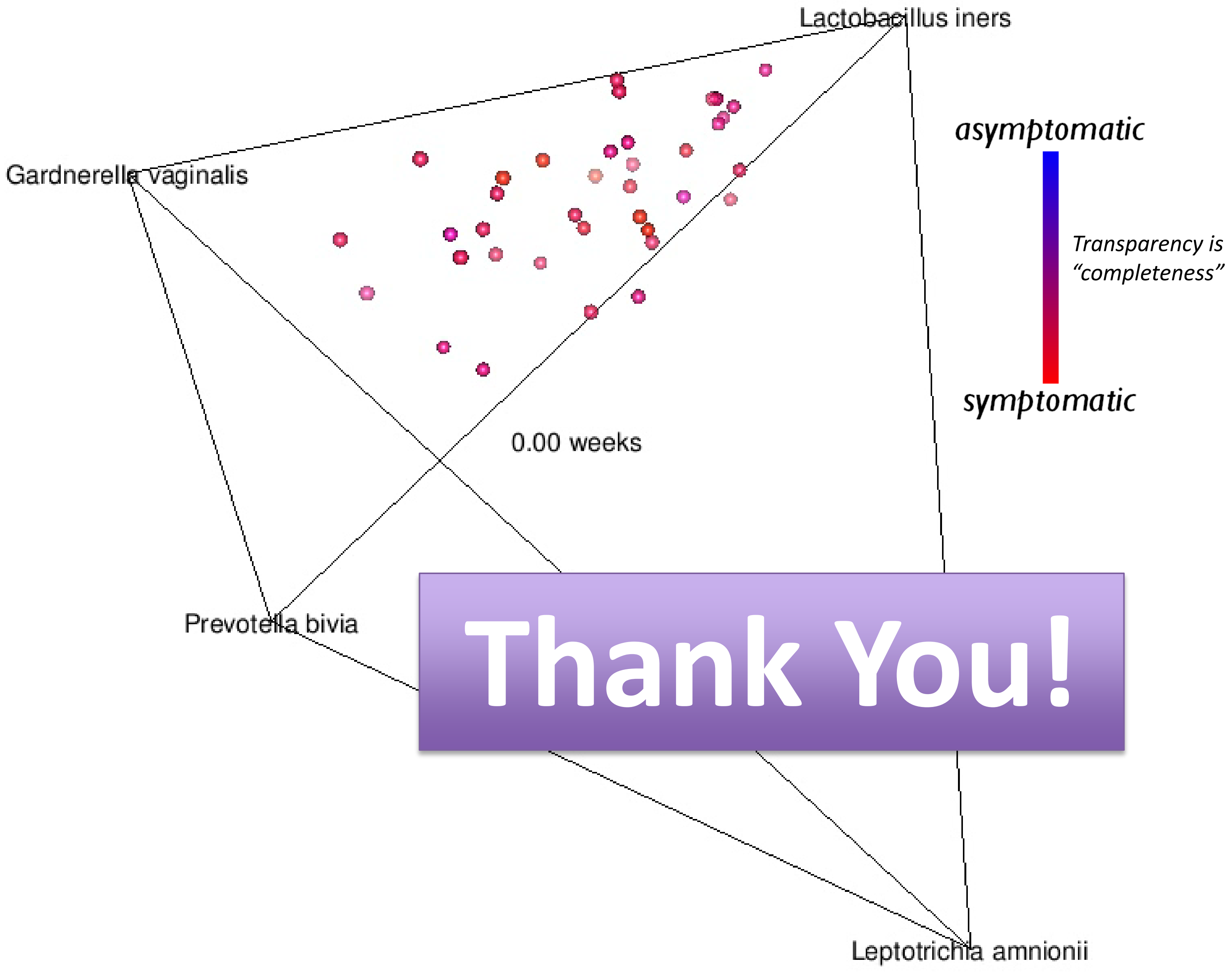




	Sample	Size
FN:TP OR	Upper	Lower
0.1	6.0	5.3
0.05	7.8	6.9
0.02	10.1	9.0
0.01	11.9	10.6



	Sample	Size
FP:TN OR	Upper	Lower
0.1	4.2	3.5
0.05	5.5	4.5
0.02	7.2	5.9
0.01	8.4	7.0


Lactobacillus iners
Gardnerella vaginalis
asymptomatic
Transparency is "completeness"
symptomatic
0.00 weeks
Prevotella bivia
Leptotrichia amnionii

Thank You!