**Master degree in Physics of Data** - Academic Year 2024/2025
Final Project for the course of:
Laboratory of Computational Physics - mod A
**Teacher**: Marco Zanetti


UNIVERSITÀ DEGLI STUDI DI PADOVA

# Data Analysis of Mice Gut Microbiome

**Group ID**: 08

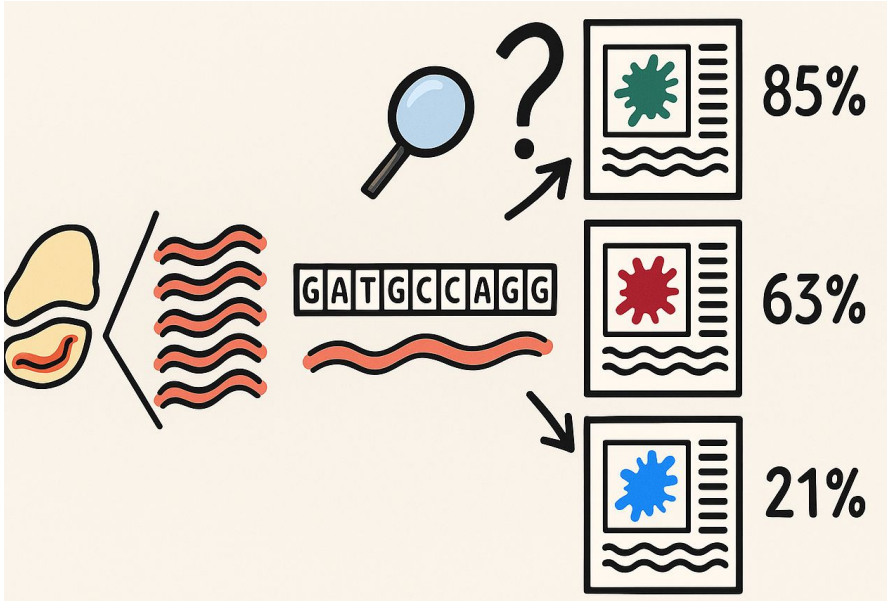
**Supervisor:** Samir Simon Suweis | samir.suweis@unipd.it

**Students name | Student ID | Student email**

Bortolato, Angela | 2156562 | angela.bortolato.2@studenti.unipd.it
Fasiolo, Giorgia | 2159992 | giorgia.fasiolo@studenti.unipd.it
Volpi, Luca | 2157843 | luca.volpi@studenti.unipd.it
Zara, Miriam | 2163328 | miriam.zara@studenti.unipd.it

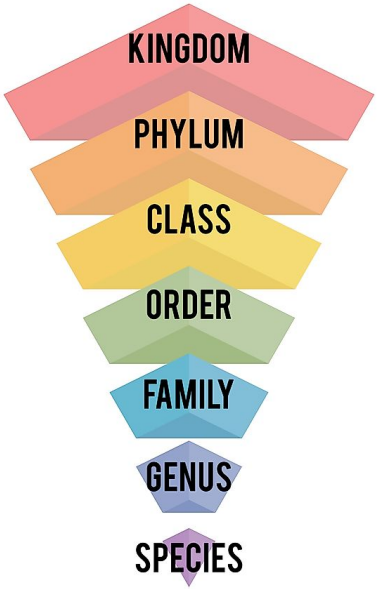
# 1. The Data

- 8 mice, born from the same parents and raised in the same cage
- A fecal sample taken from each of them every few days (~ 4-7)
- Bacteria in it are identified with **16s rRNA sequencing technique**




selective PCR amplification

database comparison



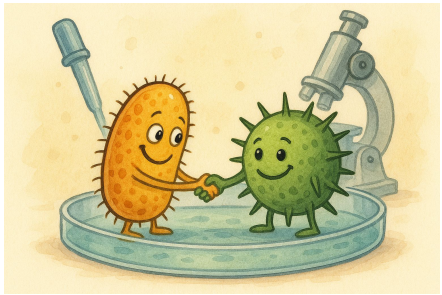
# 2. The Data

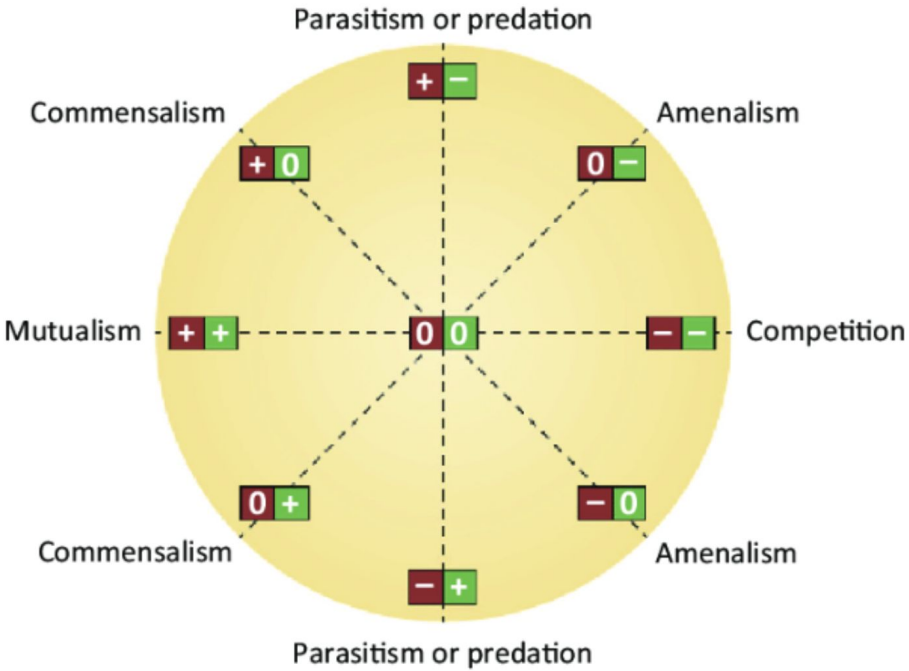
| OTU | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|
| 00001 | Bacteroidia | Bacteroidales | Prevotellaceae | **Prevotella** | Prevotella (79.62%) |
| 00002 | Bacilli | Lactobacillales | Lactobacillaceae | **Lactobacillus** | Lactobacillus taiwanensis (100%) |

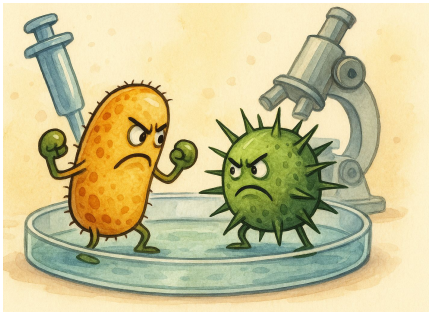
# 3. Learning Ecological Interactions

Treating *Clostridium difficile* Infection With Fecal Microbiota Transplantation

Bakken, Johan S. et al. Clinical Gastroenterology and Hepatology, Volume 9, Issue 12, 1044 - 104, 2011 DOI: 10.1016/j.cgh.2011.08.0149

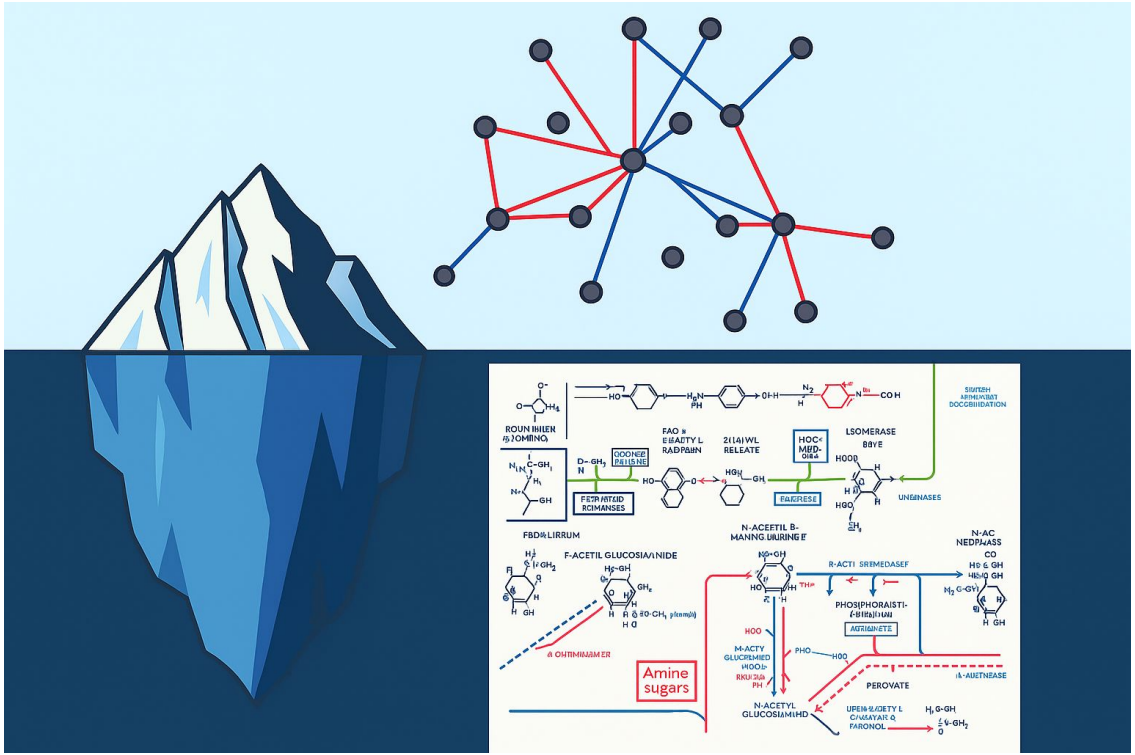




images generated with AI

# 4. Learning Ecological Interactions




Complex network:

- species are nodes
- edges are weighted and signed to represent the interaction

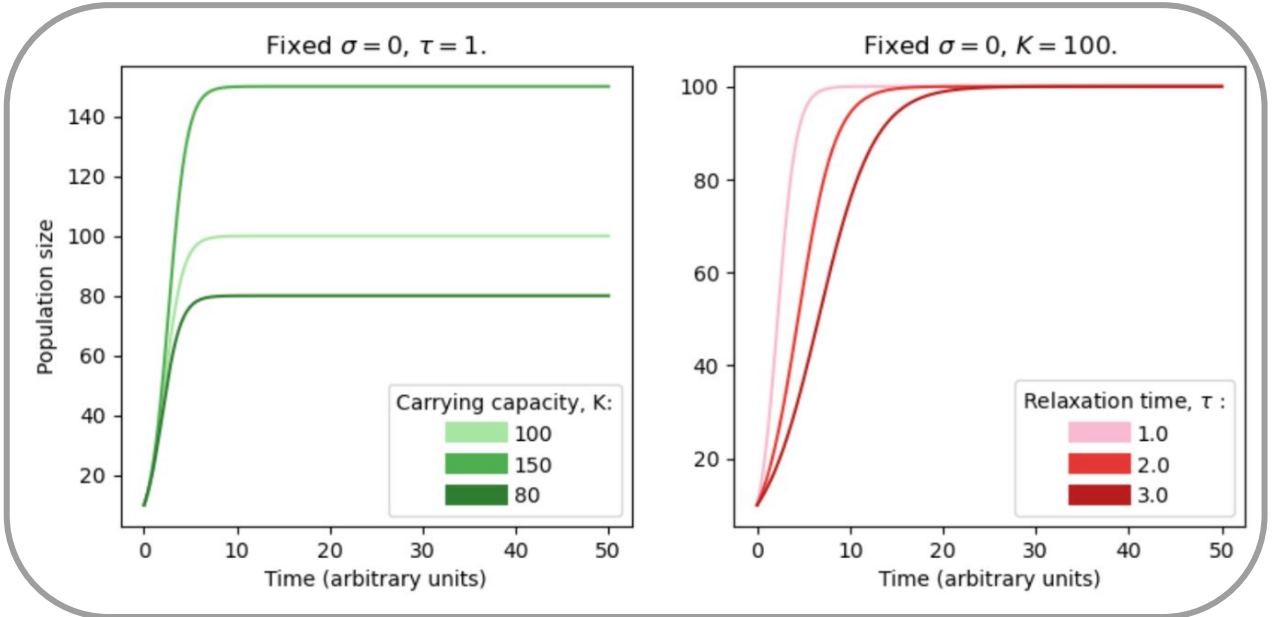
Goal: predict how the ecosystem will respond to external perturbation.

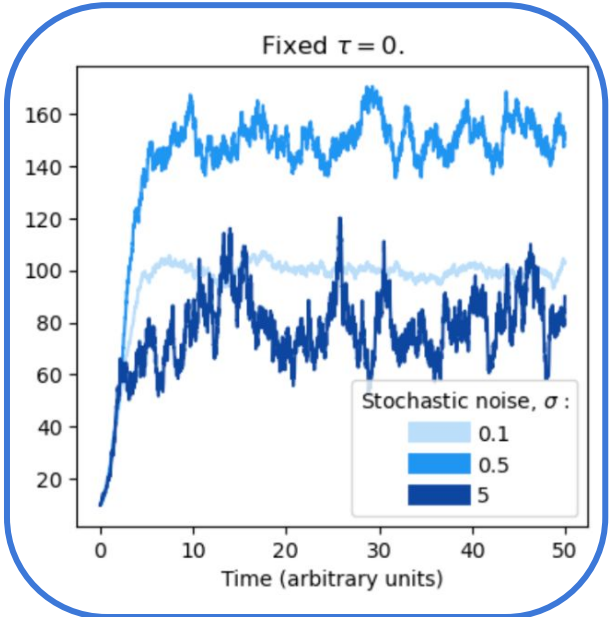
…. Does the available data even show evidence of interactions?

# 5. Logistic Model vs Stochastic Logistic Model

$$\dot{\lambda}(t) = \lambda(t)\left[\frac{1}{\tau}\left(1-\frac{\lambda(t)}{K}\right)+\sqrt{\frac{\sigma}{\tau}}\xi(t)\right]$$

N = Population size
K= Carrying capacity
τ = Relaxation time
σ = Environmental noise intensity
ξ(t) = Gaussian white noise



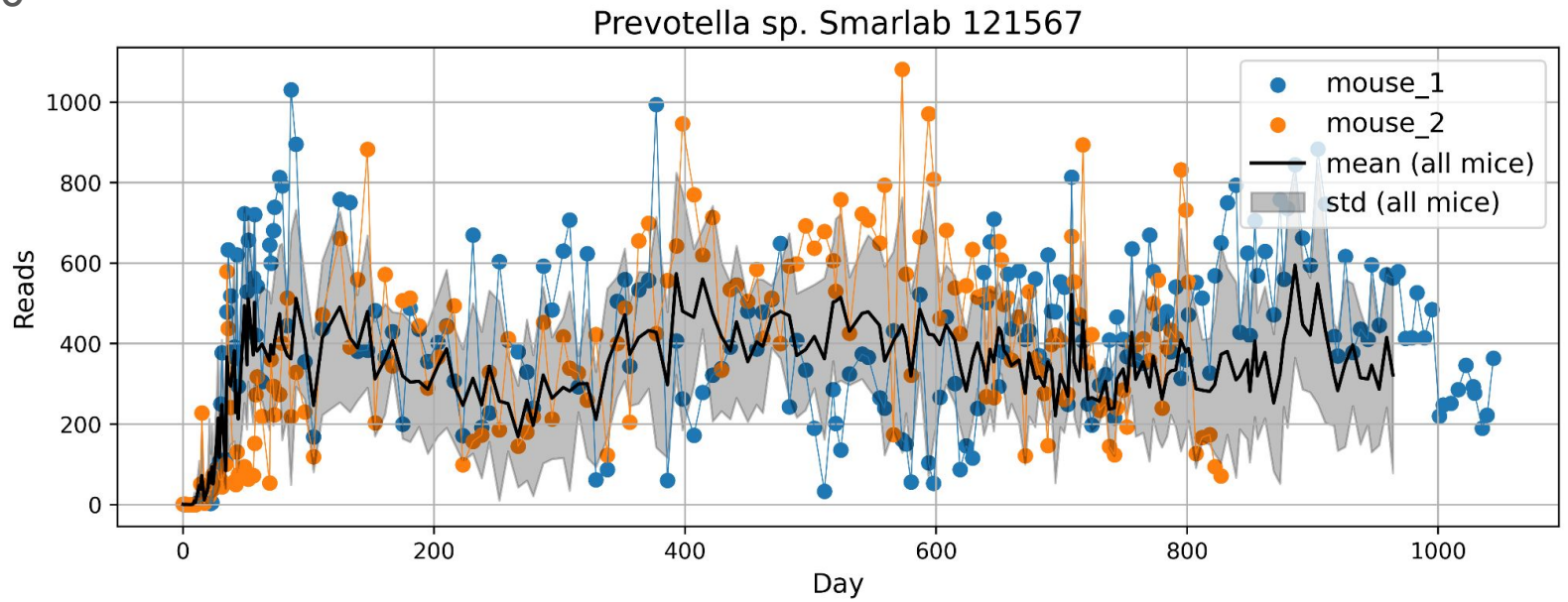

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

# Data Preliminary Analysis

# 6. Data visualization

Preprocessing step: aggregate the reads for OTUs assigned to the same species

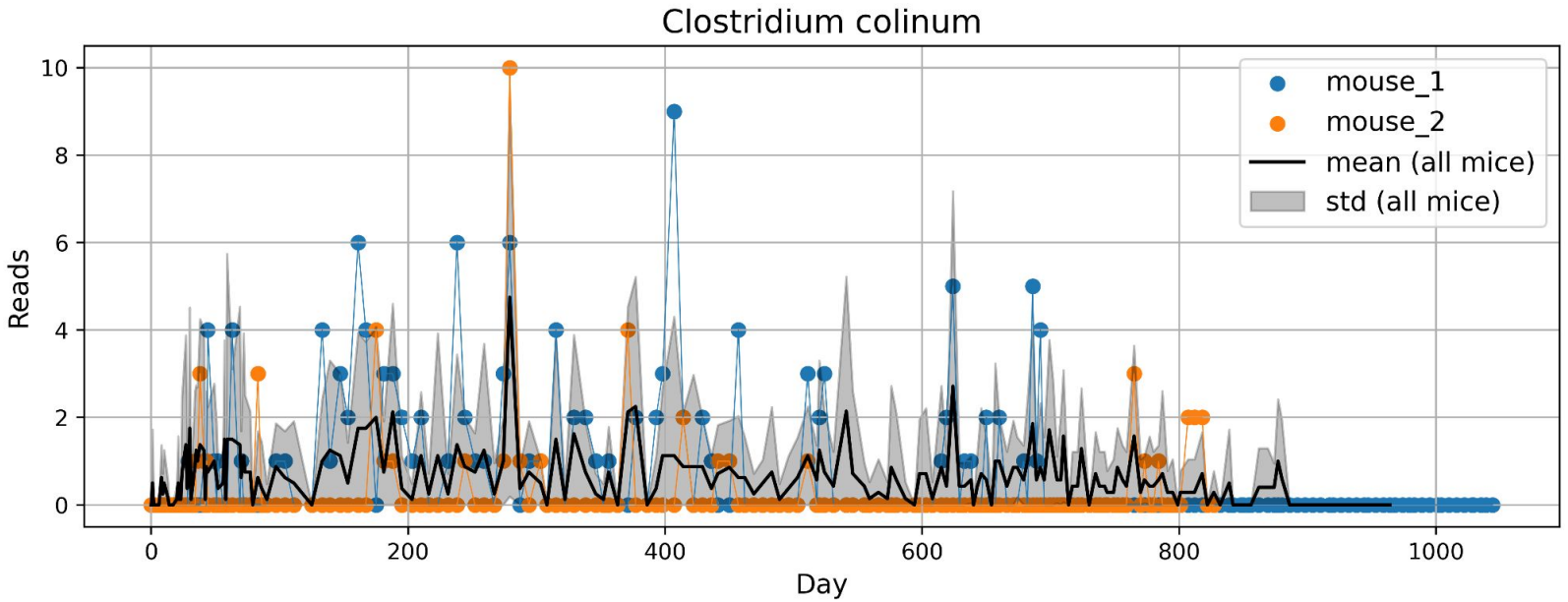
OTU queries: 21.768
*Species*: 1.260
*Genus*: 412
*Family*: 141
*Order*: 66
*Class*: 37



# 7. Data visualization

OTU: 21.768
*Species*: 1.260
*Genus*: 412
*Family*: 141
*Order*: 66
*Class*: 37

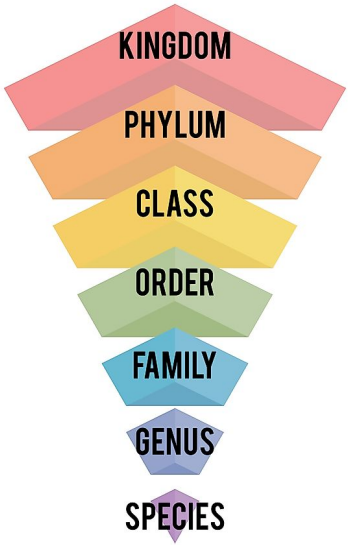
Threshold?

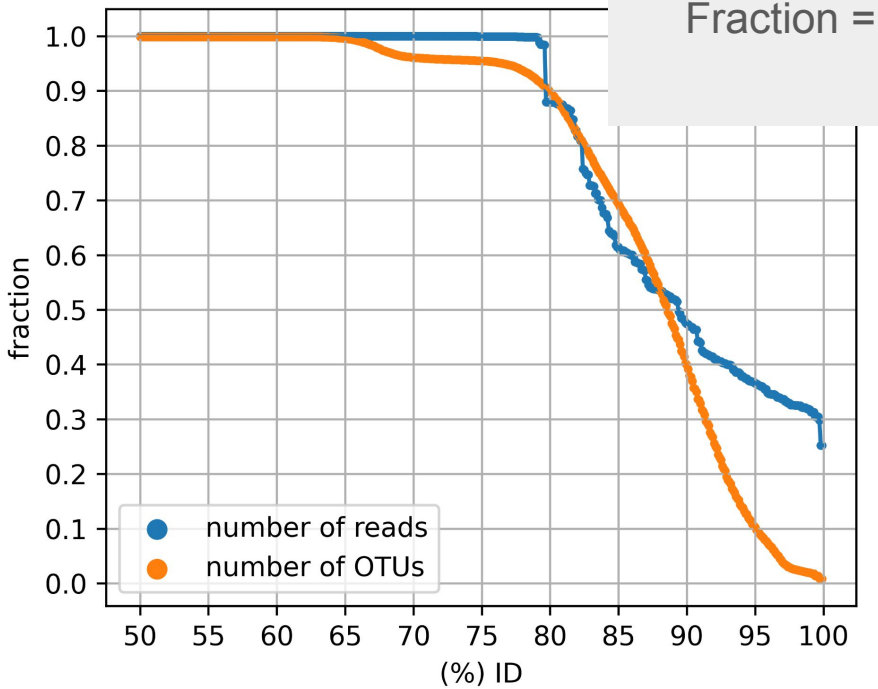


# 8. Resolution of OTU assignation

How reliable is the identification of OTUs at the *Species* taxonomic level ?

$$\text{Fraction} = \frac{\#\text{OTU/reads satisfying the \%ID}}{\text{total \#OTU/reads}}$$



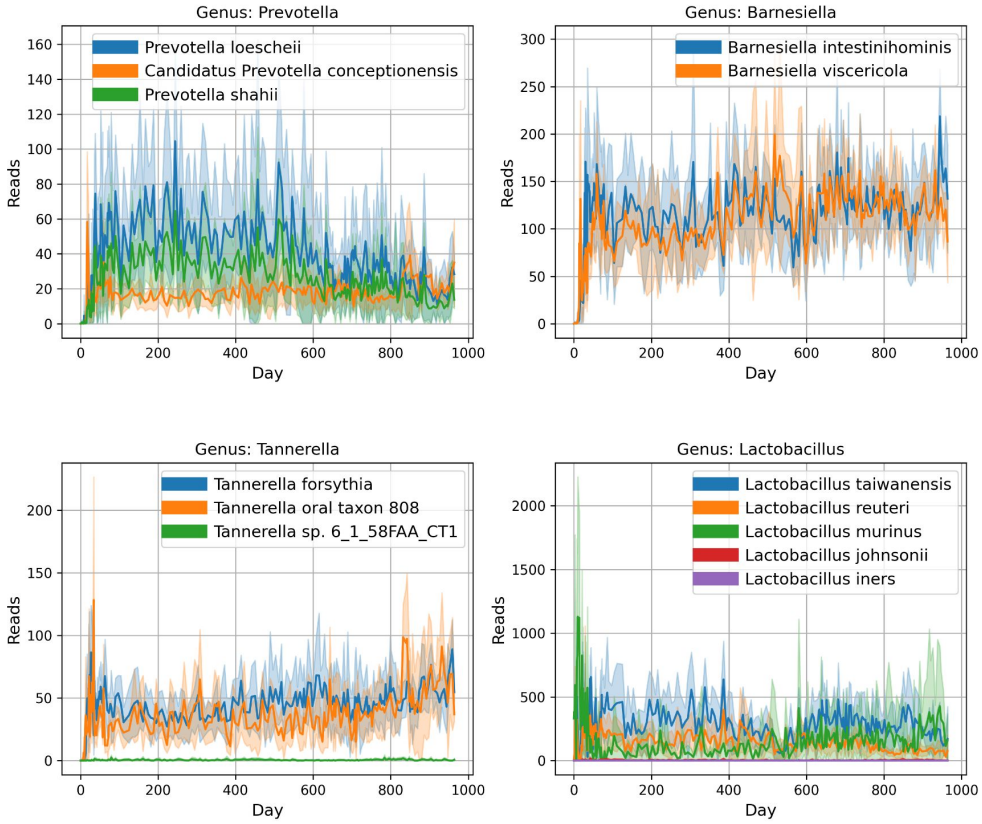



# 9. Data Analysis - Measure uncertainties

Why aggregating by *Genus*?

- More robust taxonomic classification at this level (more bases are preserved)
- Better network inference when fewer nodes, while keeping taxonomic diversity

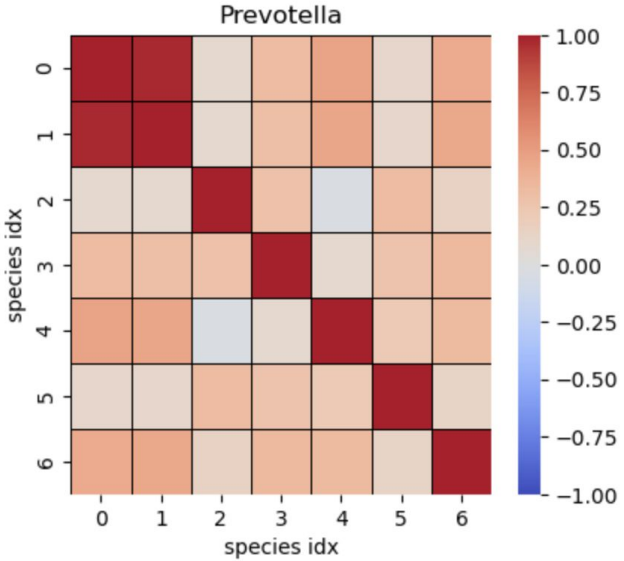
*Species*: 1.260
*Genus*: 412



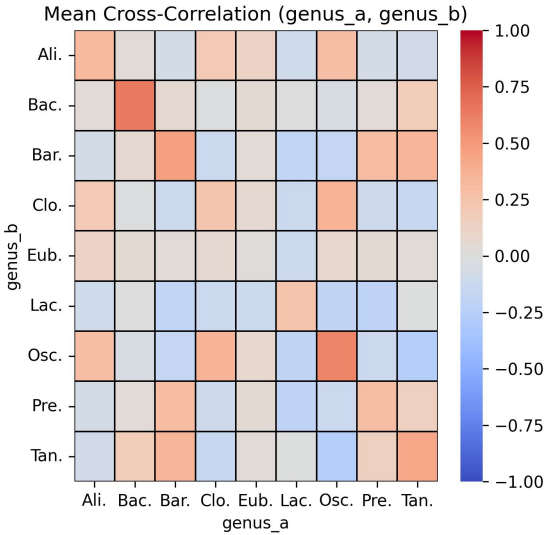
# 10. Data Aggregation by “Genus”

Restricting to:

- genera with at least 2 species
- species with mean counts > 3




Mean cross-correlation for genus Prevotella = 0.28

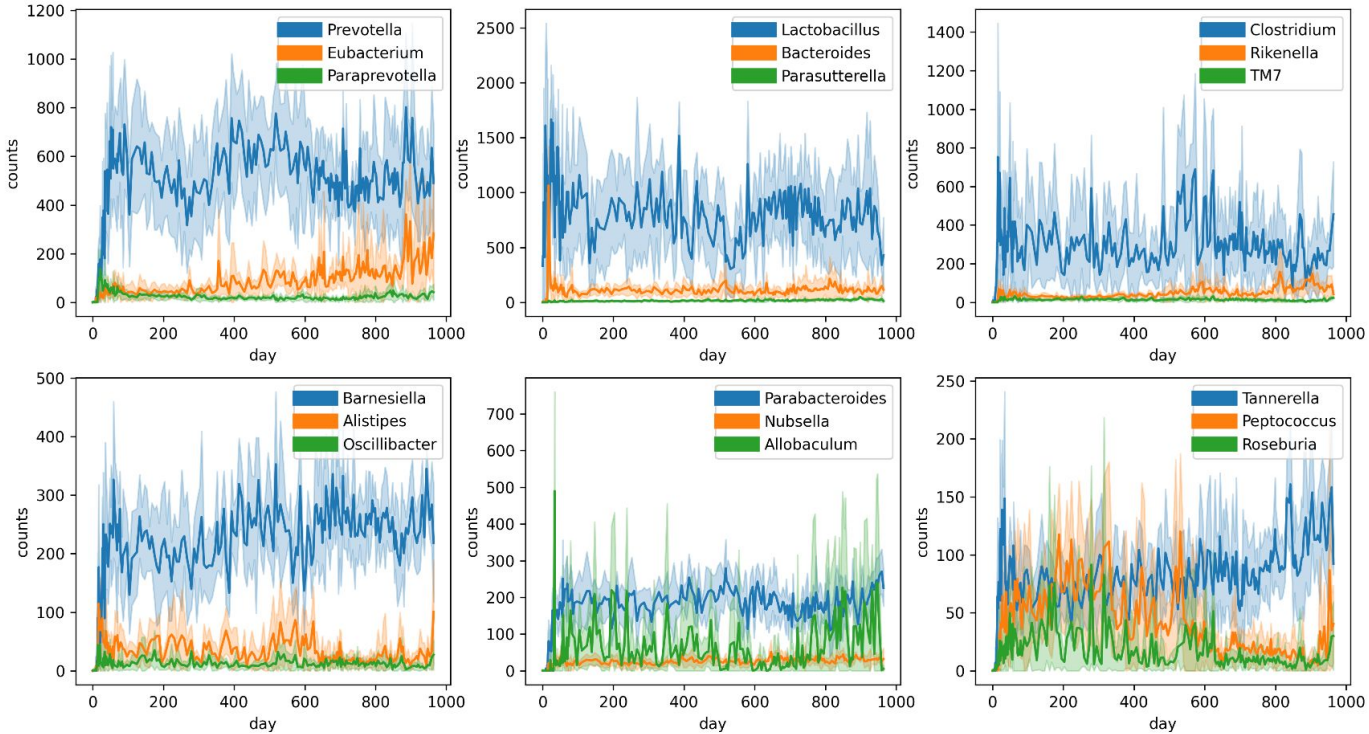



Mean cross-correlation:

- inter genus: ~ 0.008
- intra genus: ~ 0.356

# 11. Data Aggregation by “Genus”

Examples of Genus time series (18 most abundant).



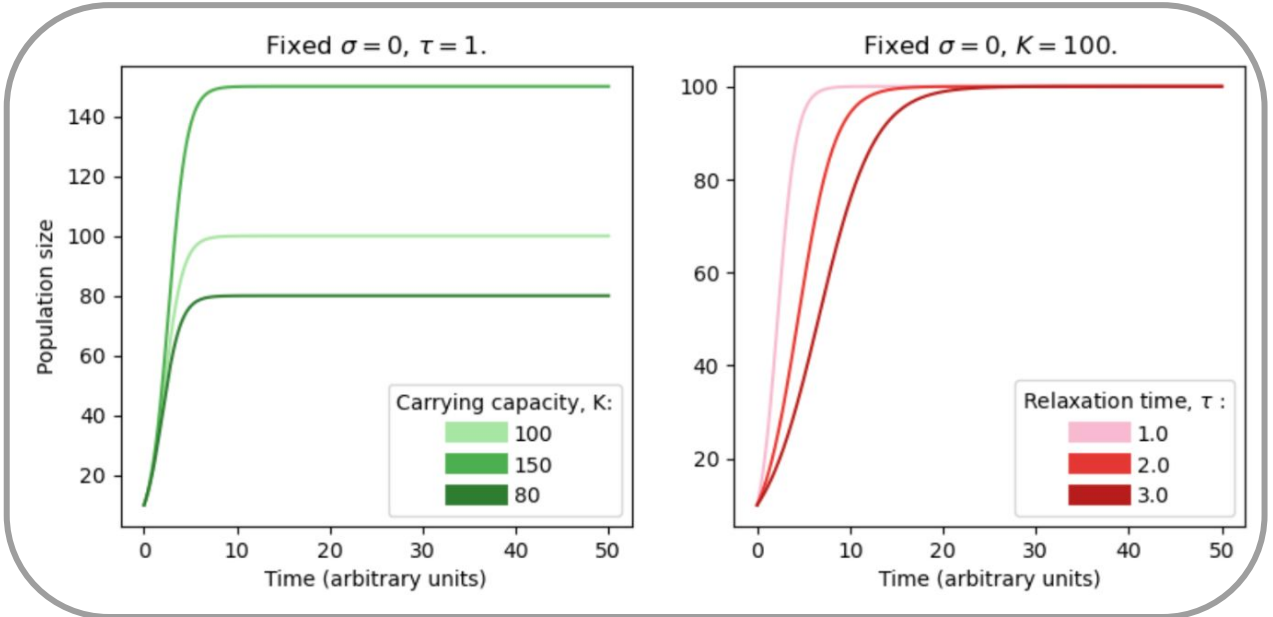

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

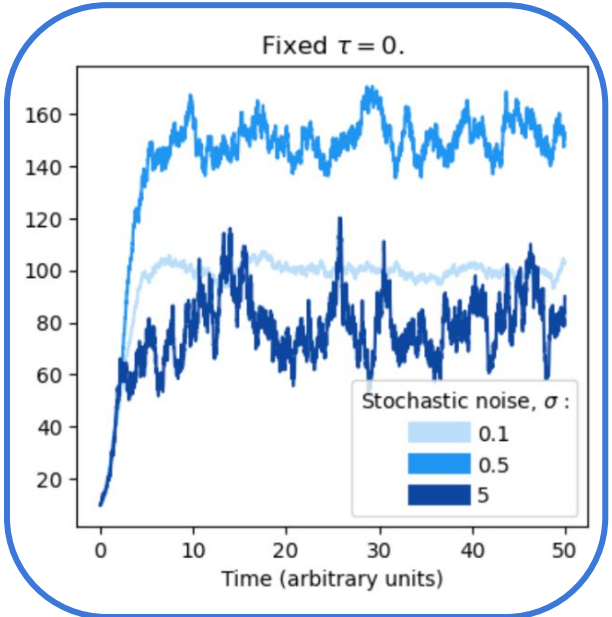
# Time Series Analysis

# 5. Logistic Model vs Stochastic Logistic Model

$$\dot{\lambda}(t) = \lambda(t)\left[\frac{1}{\tau}\left(1-\frac{\lambda(t)}{K}\right)+\sqrt{\frac{\sigma}{\tau}}\xi(t)\right]$$

N = Population size
K= Carrying capacity
τ = Relaxation time
σ = Environmental noise intensity
ξ(t) = Gaussian white noise



# 13. Workflow

The aim of this project is to **validate a quantitative model** that describes the **intra-host variability** and **intra-genera dynamics** of bacterial populations in the **mouse gut microbiome,** as proposed by Zaoli and Grilli (2021).

1

**DISSIMILARITY**

How abundances change overtime

2

**THRESHOLD**

75th percentile of dissimilarity slopes

3

**GAMMA DISTRIBUTION**

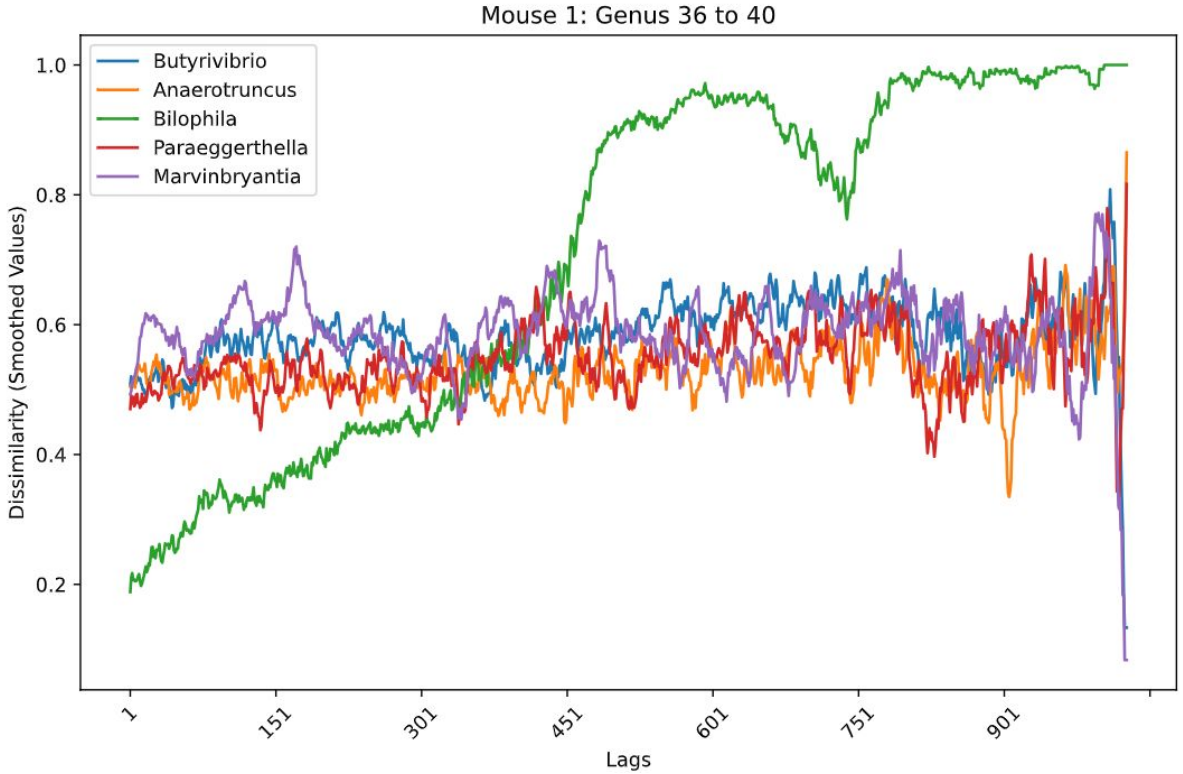
For stationary Genus

*Analysis is performed separately for every mouse and genus.*

# a) Dissimilarity

DISSIMILARITY

$$\Phi_i(t, T) = \left( \frac{\lambda_i(t) - \lambda_i(t+T)}{\lambda_i(t) + \lambda_i(t+T)} \right)^2$$

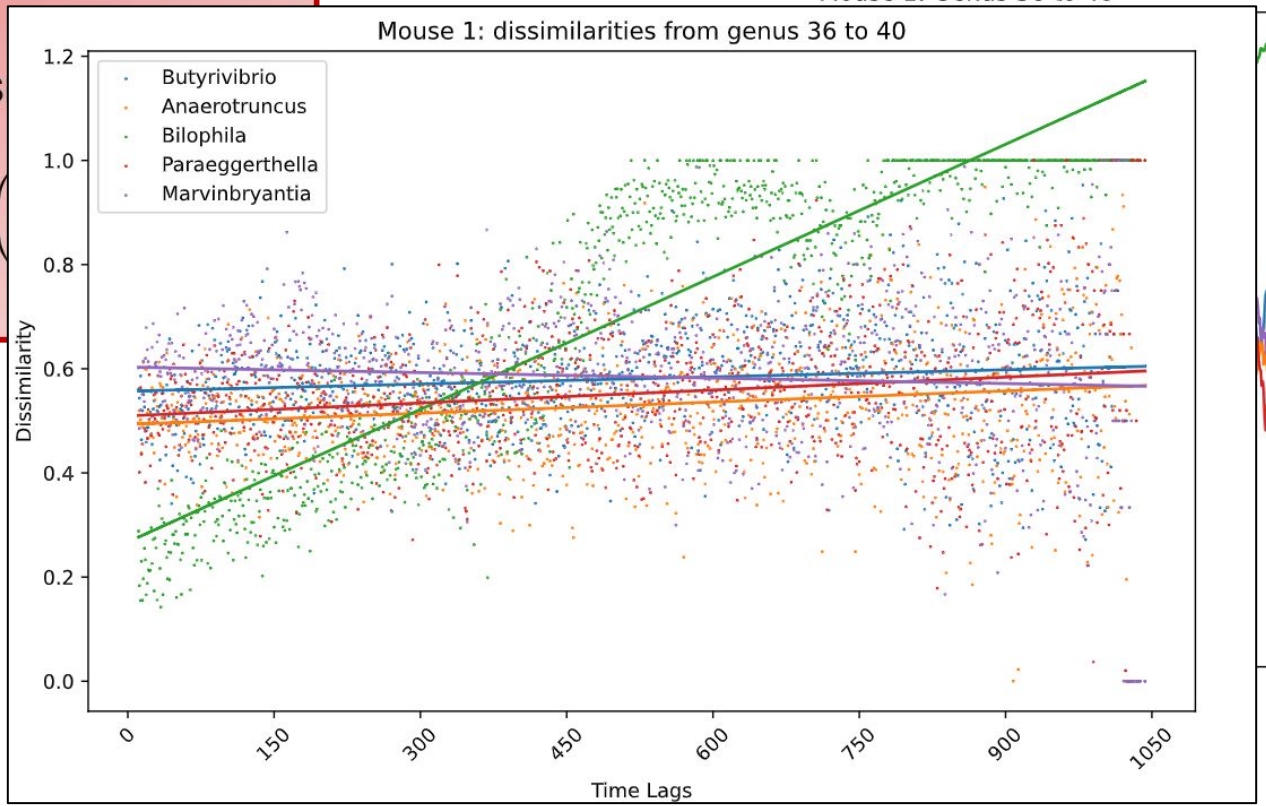


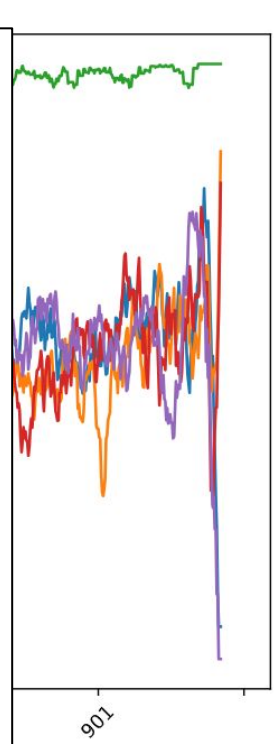
# a) Dissimilarity

1

DISS

$\Phi_i(t, T) =$ (



# b) Thresholds

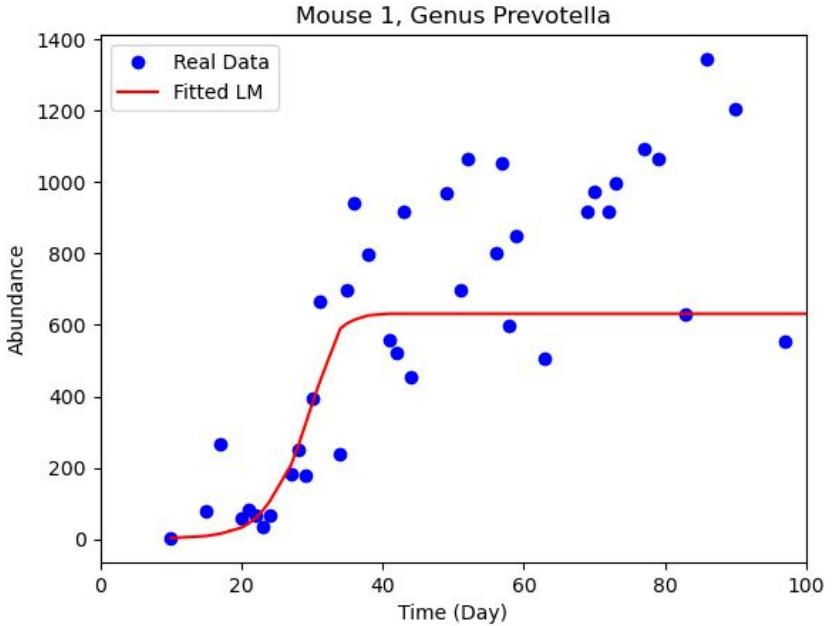
| Fit | Parameters | Simulation | Sampling | Thresholds |
|---|---|---|---|---|
| We fitted data with Logistic Model to find the transient end | Discarded the transient, we estimated parameters | We simulated the dynamics of each Genus according to the SLM with parameters equal to the parameters estimated | From simulated these time series, we sampled only for days for which the individual was sampled in real data | We then computed simulated dissimilarity and its slope. We defined as threshold the 75% quantile of the slopes obtained for that individual |

# b) Thresholds

**Fit** | **Parameters**

We fitted data with Logistic Model to find the transient end

Discarded the transient, we estimated parameters K and σ (amplitude of fluctuations)



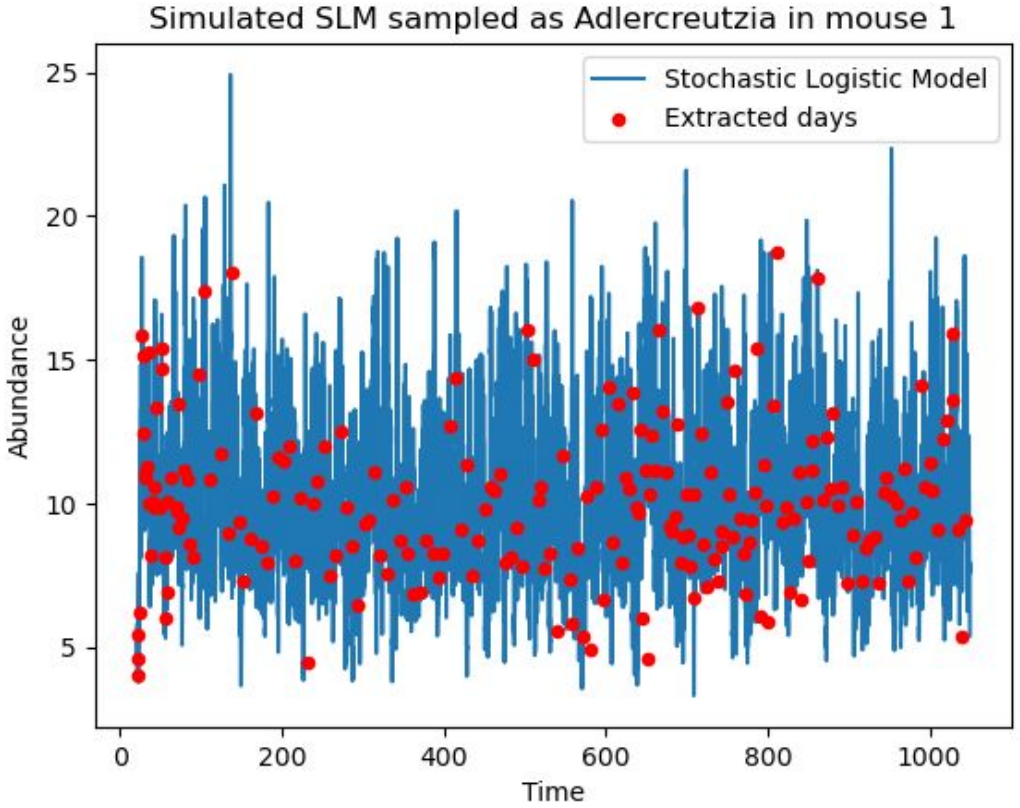

$$\langle \lambda \rangle = K \frac{2-\sigma}{2} \qquad Var(\lambda) = \frac{\sigma \langle \lambda \rangle^2}{2-\sigma}$$

# b) Thresholds

**Simulation** **Sampling**

We simulated the dynamics of each Genus according to the SLM with parameters equal to the parameters estimated

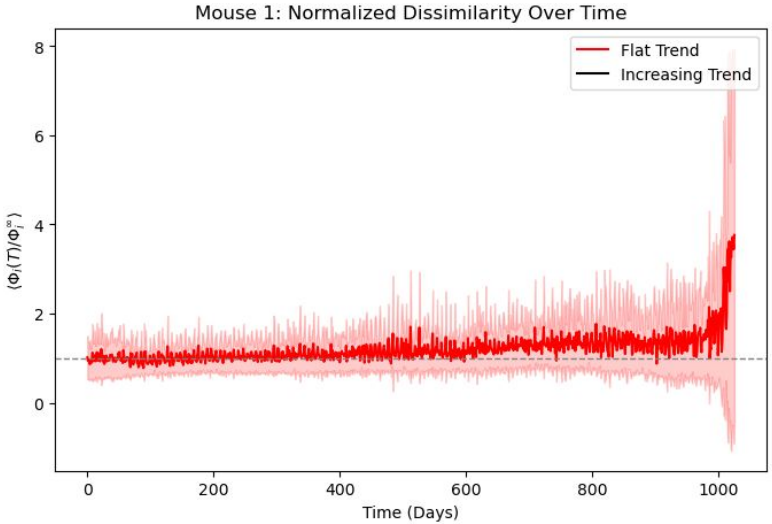
From simulated these time series, we sampled only for days for which the individual was sampled in real data

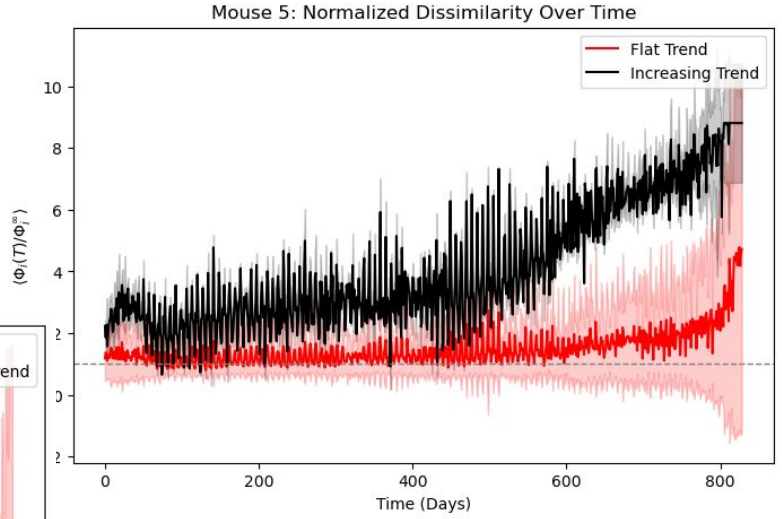


# b) Thresholds

## Thresholds

We then computed simulated dissimilarity and its slope. We defined as threshold the 75% quantile of the slopes obtained for that individual



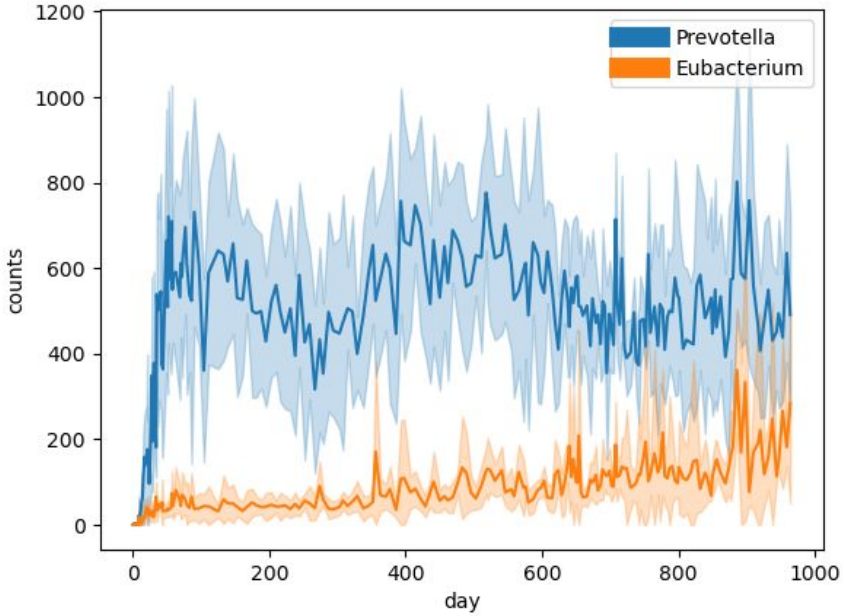




$$\mathbb{E}(\phi_i^{\infty}) = \frac{\sigma}{4-\sigma} \qquad \left\langle \frac{\phi_i(T)}{\phi_i^{\infty}} \right\rangle$$

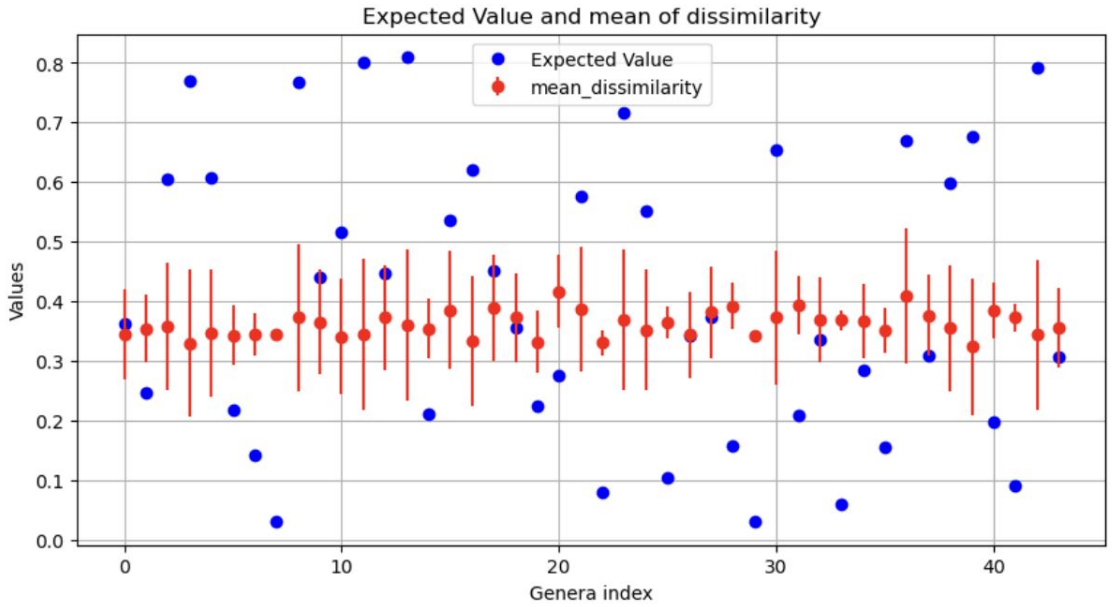
# b) Thresholds

With thresholds calculated for each mouse, we classified Genera by “**flat**” and “**increasing**”.
Almost all Genera are classified “flat” except:

| | |
|---|---|
| Mouse 1 | - |
| Mouse 2 | Candidatus Arthromitus, Eubacterium |
| Mouse 3 | - |
| Mouse 4 | - |
| Mouse 5 | Aureibacter, Coprobacter |
| Mouse 6 | - |
| Mouse 7 | Aureibacter, Coprobacter, Eubacterium, Parasutterella |
| Mouse 8 | Coprobacter |



# c) Gamma Distribution

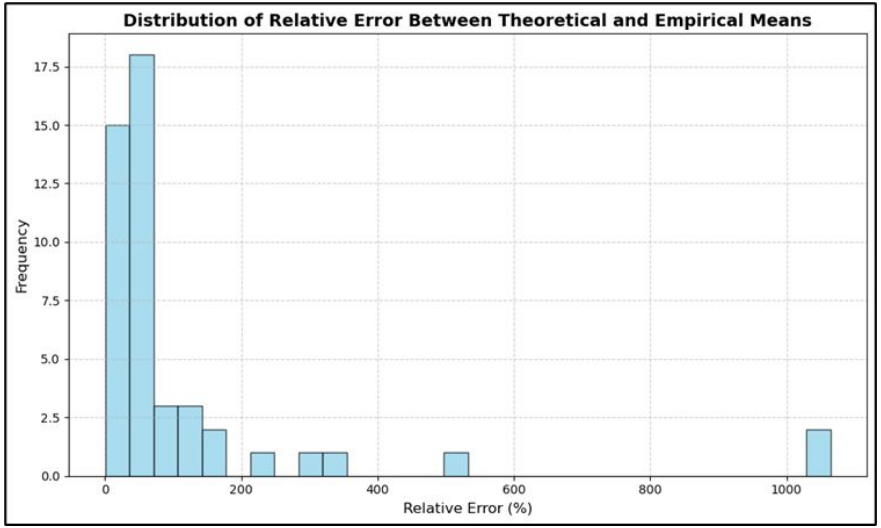



$$E[\Phi_\infty] = \frac{\sigma}{4-\sigma}$$

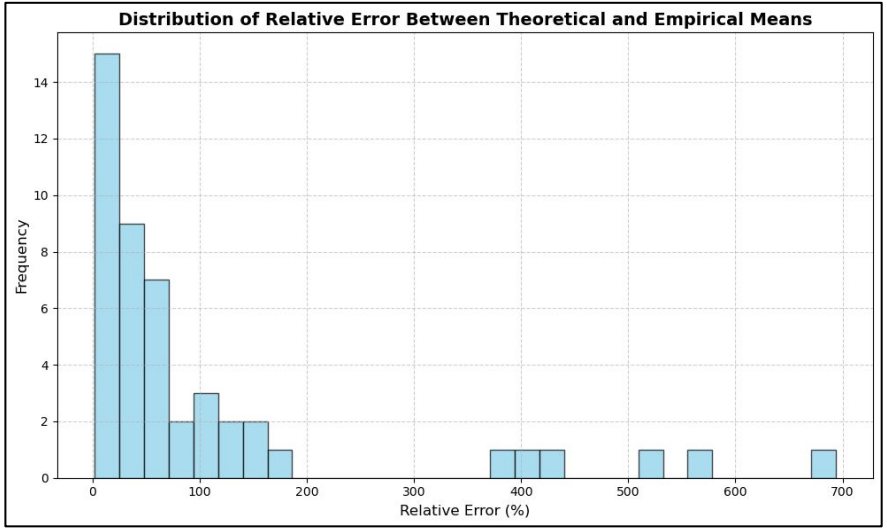
$$P(\lambda; K, \sigma) = \frac{1}{\Gamma\left(\frac{2}{\sigma} - 1\right)} \cdot \left(\frac{2-\sigma}{K}\right)^{\frac{2}{\sigma}-1} \cdot \lambda^{\frac{2}{\sigma}-2} \cdot \exp\left(-\frac{2}{\sigma K} \cdot \lambda\right)$$

# 14. Relative error

## Mouse 1




## Mouse 2



# 15. Conclusions and Improvements

Stochastic Logistic Model is likely not an appropriate or sufficiently robust representation of the complex biological system within the mouse gut

Further improvements would be:

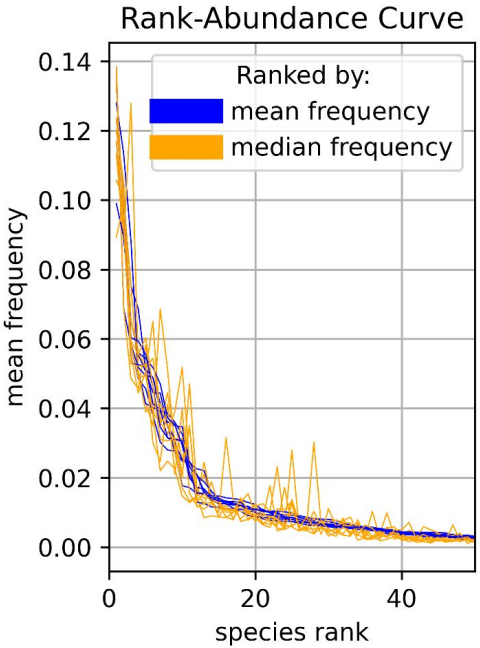
- K-jumps
- Outliers inspection
- Using more complex models

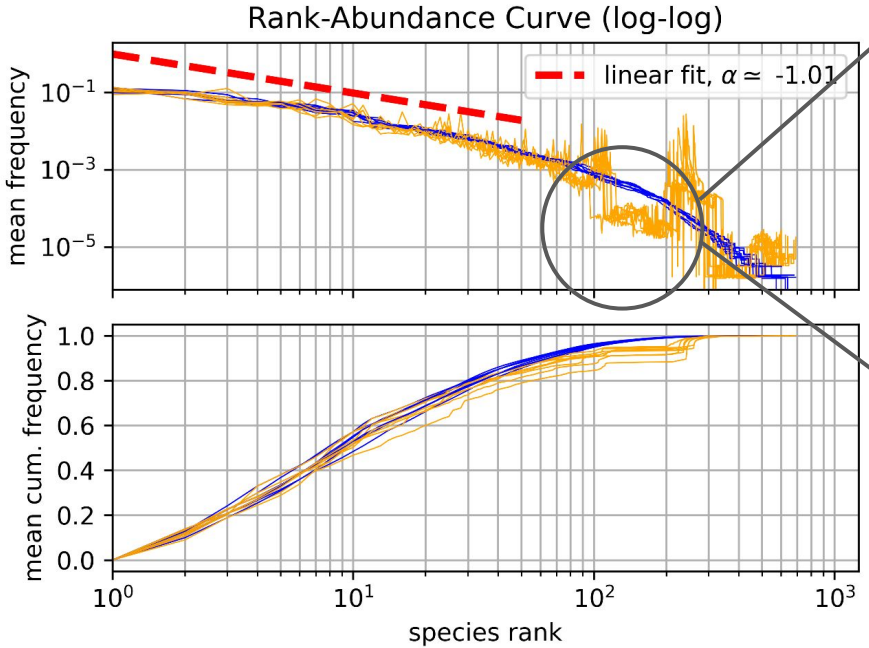

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

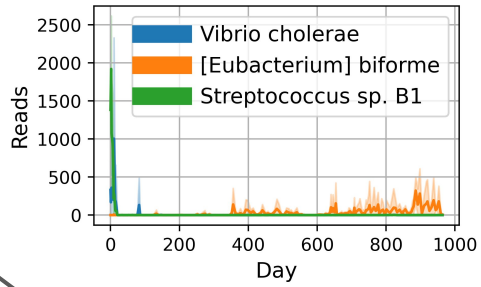
# Backup Slides

# 16. Sample composition - RAD

$$\text{frequency} = c \cdot \text{rank}^{\alpha}$$



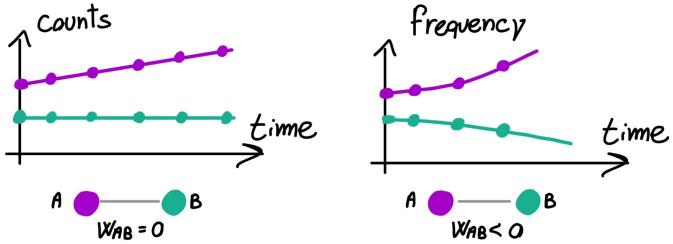

non-stationary series

# 17. Criticalities

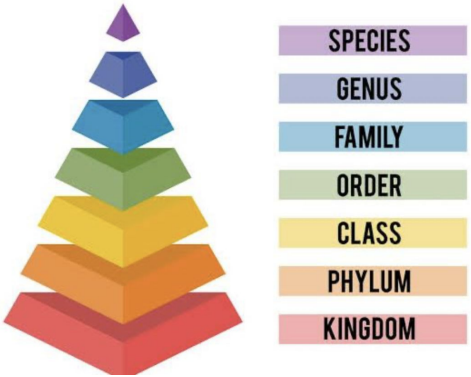
finite sequencing depth:

a species that is rare could still have a big influence on the others. By imposing a threshold (exiplicitly or implicity) we exclude possibly vital information.

the sample is fixed in size: measures are frequencies, not counts -> correlations may arise as statistical artifacts but have no correspondent physical reality



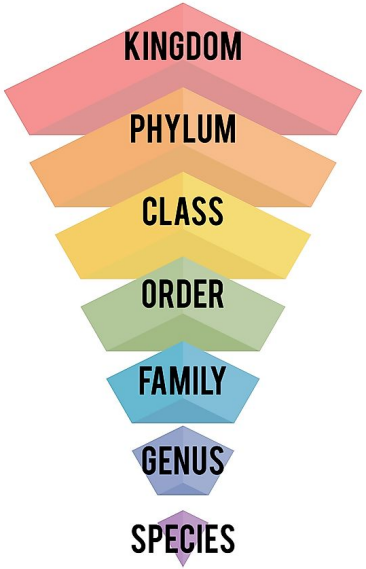

- 16s rRNA sequences was found to be efficient at identifying the high-order taxonomy, but less efficient at low-level taxonomy

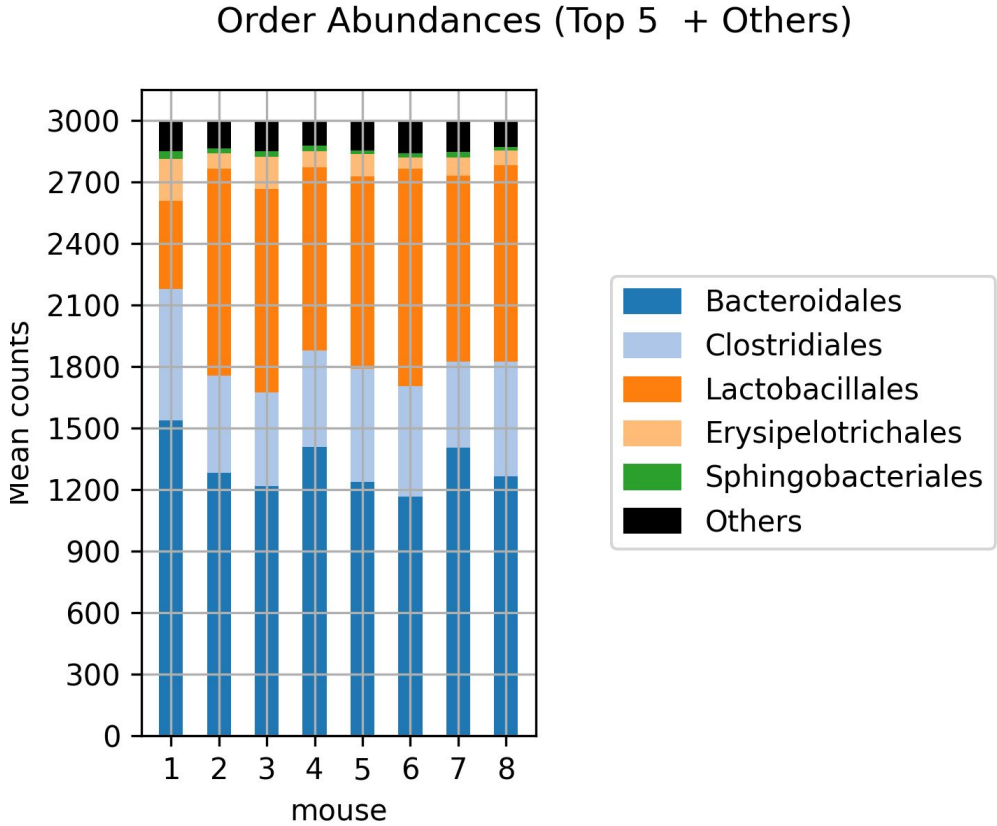


# 18. Sample Composition

OTU queries: 21.768

*Species*: 1.260
*Genus*: 412
*Family*: 141
*Order*: 66
*Class*: 37



# 19. Sample Composition