

Data analysis

Lecture 1

General practices

First project with generated data

Kateryna Pantiukh
pantiukh@ut.ee

GitHub





Please set up your GitHub repository
and code editor (Visual Studio Code)

GitHub

<https://github.com>

Visual Studio Code

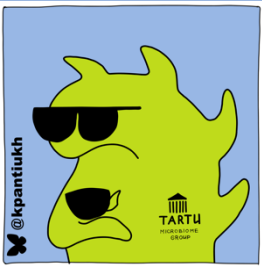
<https://code.visualstudio.com>



Please clone our working directory
from GitHub

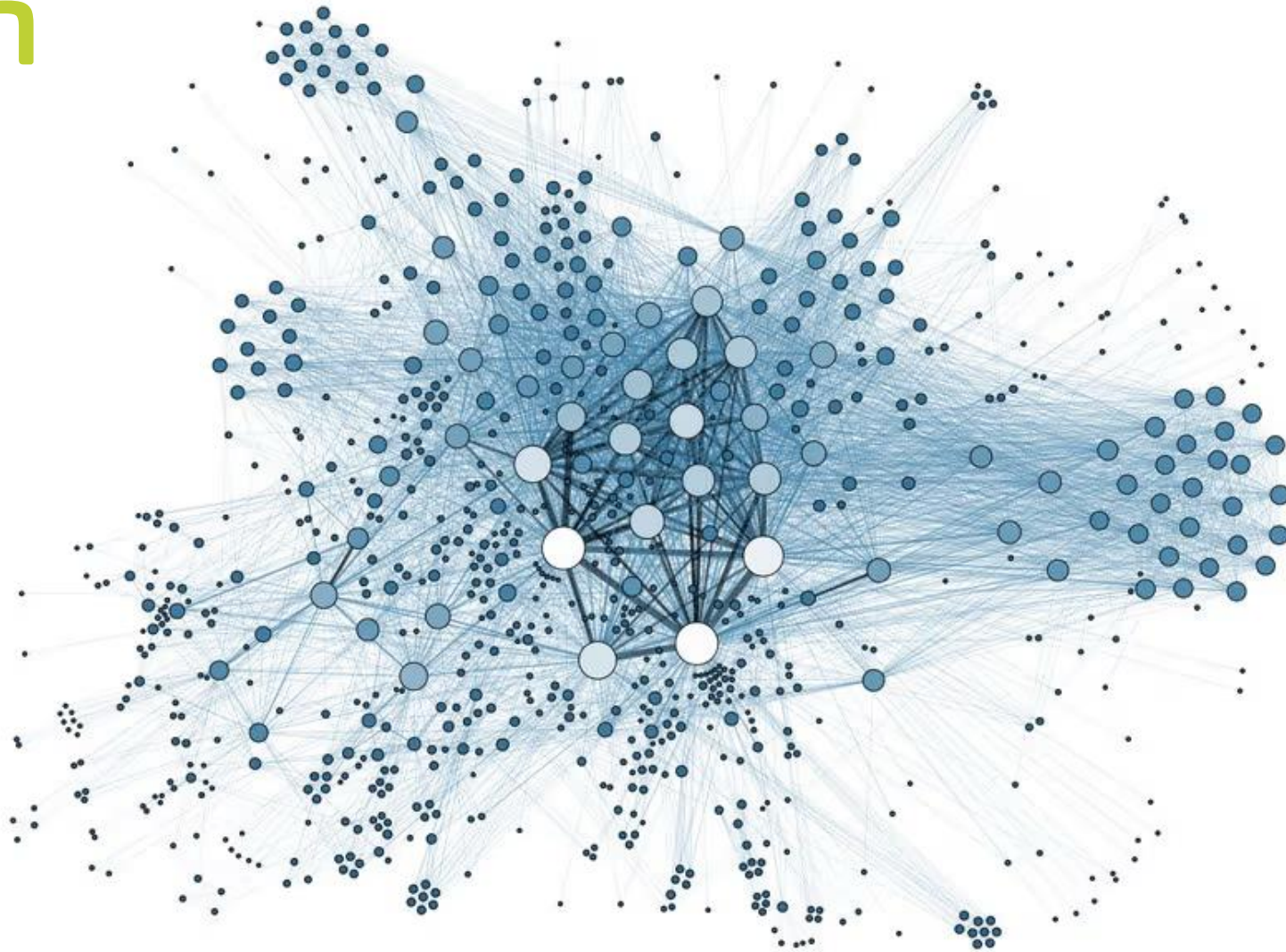
GitHub

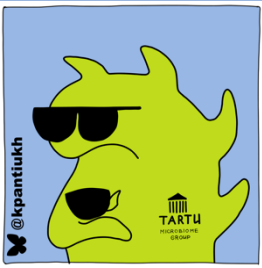
https://github.com/Chartiza/KSE_microbiome



Introduction

The **microbiome** is a **complex system** composed of interacting microbial populations whose collective behaviour emerges from their interactions.

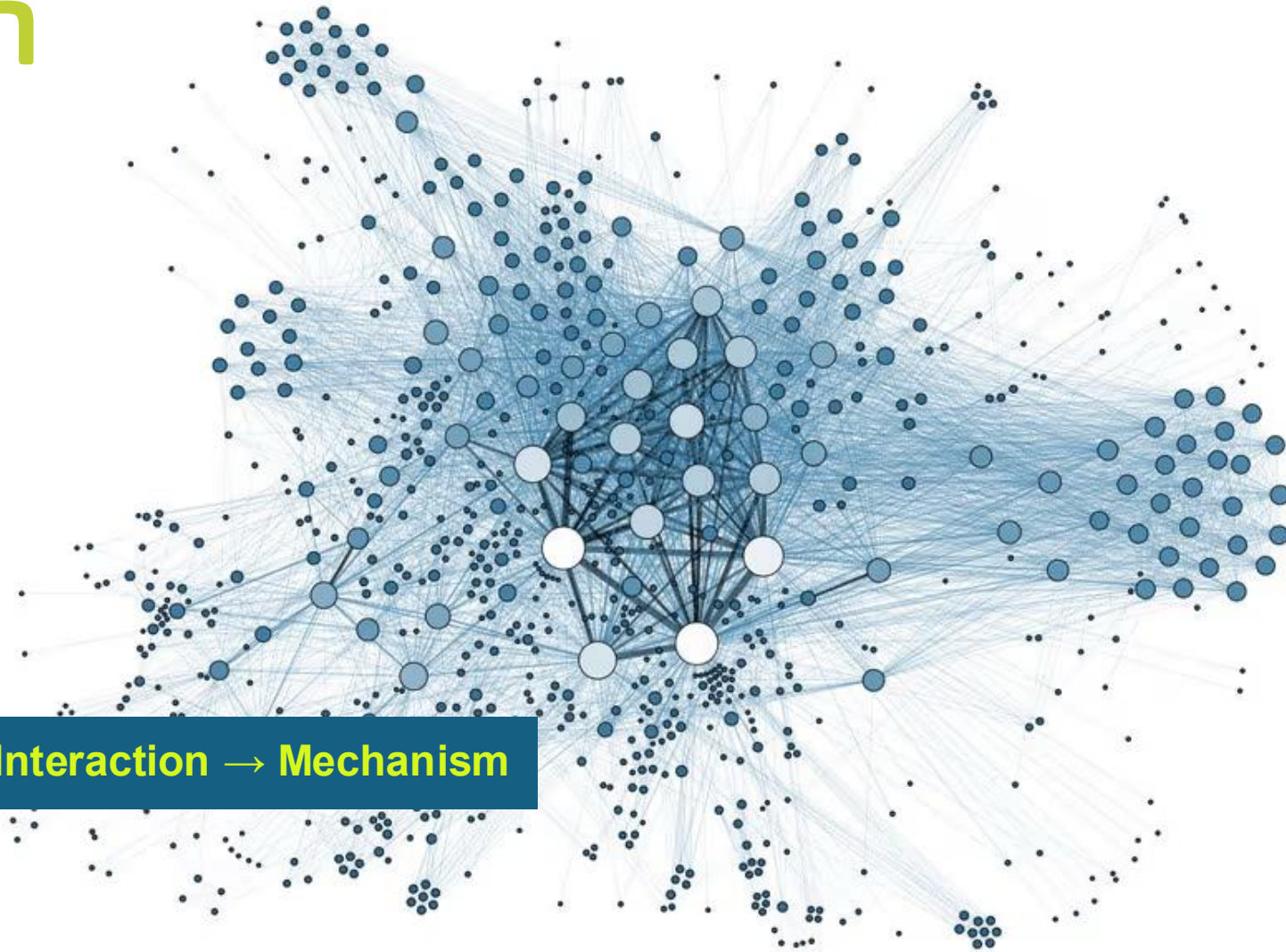




Introduction

The **microbiome** is a **complex system** composed of interacting microbial populations whose collective behaviour emerges from their interactions.

Structure → Interaction → Mechanism





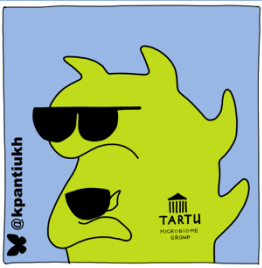
Introduction

Structure → Interaction → Mechanism

You identify:

- **Entities** or agents (who is there?)
- **Scales** (spatial, temporal, organizational).
- State **variables** describing each entity





Coding task

Analysis of Human Gut Microbiome Community Structure

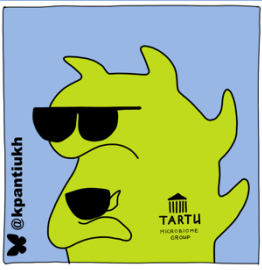
- **Dataset:**

- https://github.com/Chartiza/KSE_microbiome/tree/main/lecture_01
- You are provided with a synthetic dataset representing a cohort of human gut microbiome samples. The dataset includes three core data types commonly used in microbiome studies:
- An abundance table with microbial species abundances per sample
- A taxonomic annotation table linking species to higher taxonomic ranks
- Sample metadata describing host characteristics
-

- **Overall Goal:**

The goal of this assignment is to understand how different microbiome data types are structured, how they relate to each other, and what biological insights can be extracted from them. You will write code to explore the data, compute summary statistics, generate visualizations, and prepare a short analytical report describing your findings and conclusions.

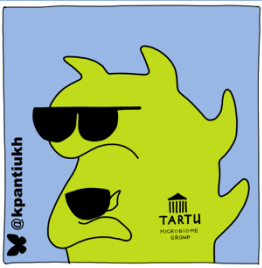




Abundance Table Analysis

- Which species appear to dominate the gut microbiome in this dataset?
- Is there evidence for a core microbiome, defined as species present in the majority of samples?
- Do samples appear broadly similar, or is there high variability between individuals?

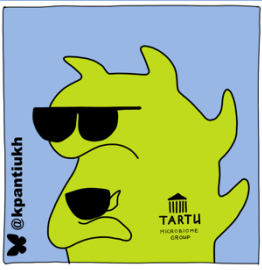




Sample Metadata Analysis

- Does this dataset appear representative of a general human population?
- What limitations do you see when trying to generalize conclusions from this dataset to a broader population?
- Which metadata variables are missing or underrepresented, and how could that affect interpretation?

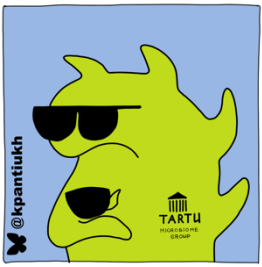




Taxonomic Table Analysis

- Whether gut microbes in this dataset are dominated by a small number of phyla or spread evenly across many phyla?





HW-01

For this homework assignment, you are expected to submit:

- Well-documented and reproducible code, including clear variable names and informative comments
- A concise written report summarizing your analyses, key results, biological interpretations, and limitations of the dataset

Deadline: 05-02-26

