

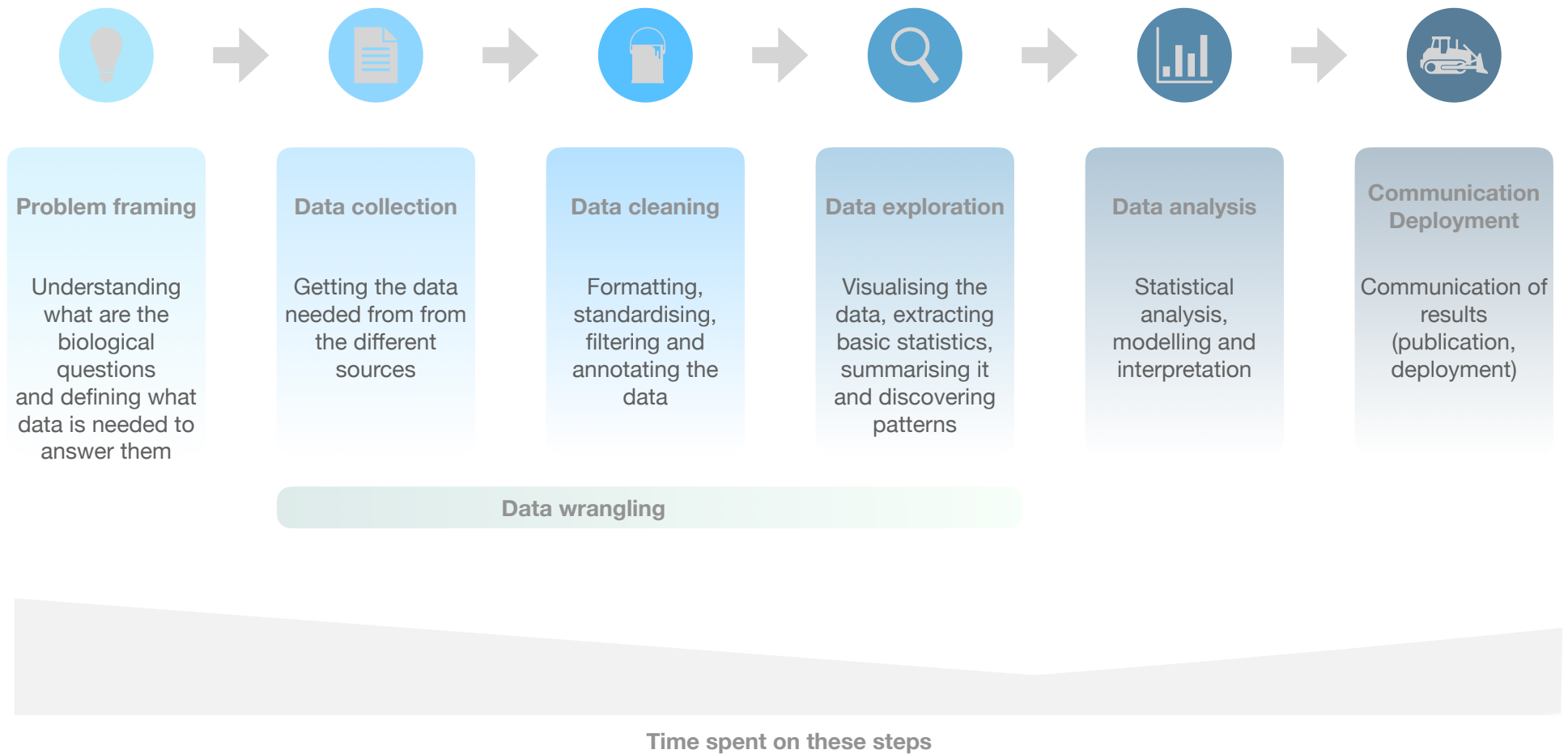
Data Wrangling

Data Club

Outline

- The Data Science Process
- Our Objective
- Data Wrangling
- Tools
- Problem framing: Project Paper
- Hands-on
- Discussion

The Data Science Process



Our Objective 2023

Developing a product



Data club




Streamlit



Data Wrangling



- **Data wrangling**

the process of transforming and preparing raw data into a format that is suitable for analysis. It involves several steps, including:

- **Data collection**

Data is collected from various sources, including databases, spreadsheets, web pages, and social media platforms.

- **Data cleaning**

Assess the quality of the data and fix any errors identified. It includes removing duplicate entries, correcting spelling and formatting errors, and dealing with missing data.

- **Data transformation**

This step involves converting data into a format that is easier to analyse. It includes data normalisation, aggregation, and filtering.

- **Data annotation**

This step involves adding additional data to the dataset. This could include merging data from different sources, adding calculated fields, or including metadata.

- **Data validation**

The final step involves verifying that the data has been transformed correctly and is ready for analysis. This step includes checking that the data is accurate, complete, and consistent.

Problem Framing: Project Paper

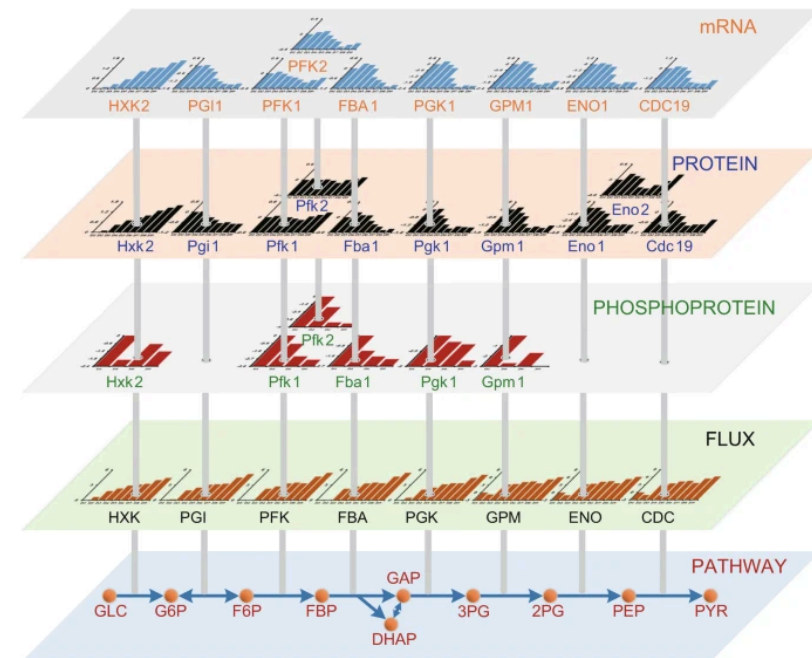
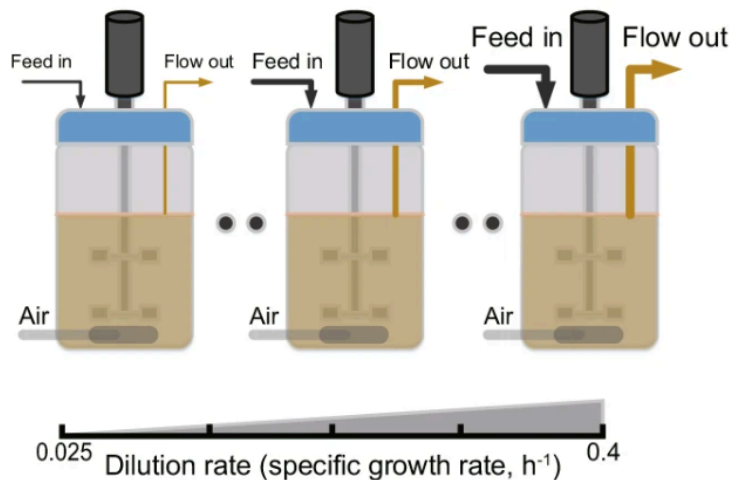
Proteome allocations change linearly with the specific growth rate of *Saccharomyces cerevisiae* under glucose limitation

[Jianye Xia](#), [Benjamin J. Sánchez](#), [Yu Chen](#), [Kate Campbell](#), [Sergo Kasvandik](#) & [Jens Nielsen](#) 

[Nature Communications](#) **13**, Article number: 2819 (2022) | [Cite this article](#)

Abstract

Saccharomyces cerevisiae is a widely used cell factory; therefore, it is important to understand how *Saccharomyces cerevisiae* organizes key functional parts when cultured under different conditions. Here, we perform a multiomics analysis of *S. cerevisiae* by culturing the strain with a wide range of specific growth rates using glucose as the sole limiting nutrient. Under these different conditions, we measure the absolute transcriptome, the absolute proteome, the phosphoproteome, and the metabolome. [...] Finally, using enzyme-constrained genome-scale modeling, we find that enzyme usage plays an important role in controlling flux in amino acid biosynthesis.



Problem Framing: Project Paper

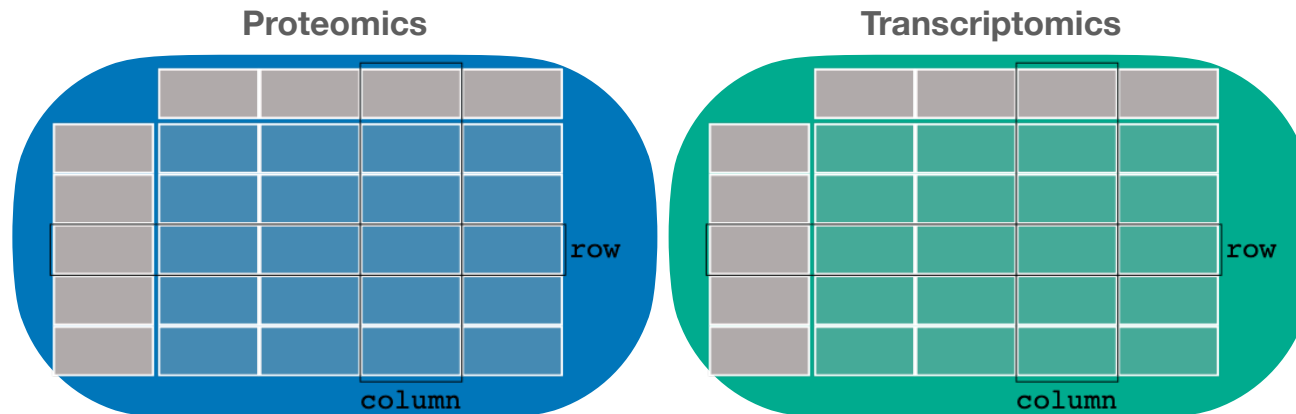
Proteome allocations change linearly with the specific growth rate of *Saccharomyces cerevisiae* under glucose limitation

[Jianye Xia](#), [Benjamin J. Sánchez](#), [Yu Chen](#), [Kate Campbell](#), [Sergo Kasvandik](#) & [Jens Nielsen](#) 

[Nature Communications](#) **13**, Article number: 2819 (2022) | [Cite this article](#)

Abstract

Saccharomyces cerevisiae is a widely used cell factory; therefore, it is important to understand how *Saccharomyces cerevisiae* organizes key functional parts when cultured under different conditions. Here, we perform a multiomics analysis of *S. cerevisiae* by culturing the strain with a wide range of specific growth rates using glucose as the sole limiting nutrient. Under these different conditions, we measure the absolute transcriptome, the absolute proteome, the phosphoproteome, and the metabolome. [...] Finally, using enzyme-constrained genome-scale modeling, we find that enzyme usage plays an important role in controlling flux in amino acid biosynthesis.



Hands-on

https://github.com/biosustain/data_club

Discussion