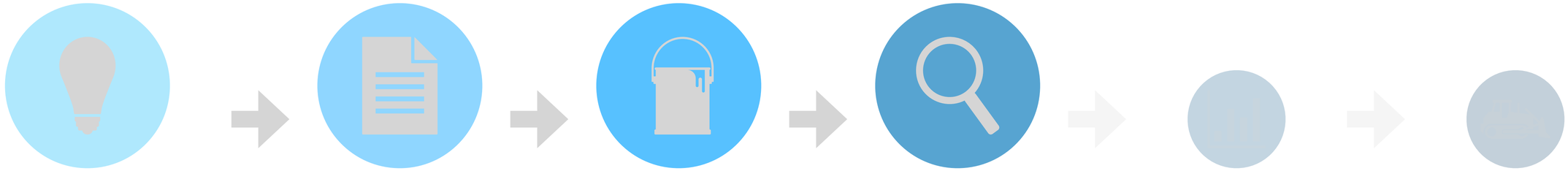


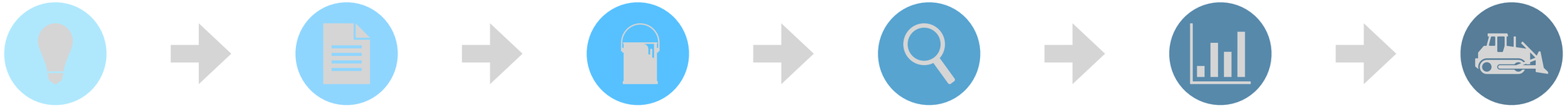
Data Annotation

Data Club

Alberto Santos

8/11/2023

The Data Science Process



Problem framing

Understanding what are the biological questions and defining what data is needed to answer them

Data collection

Getting the data needed from from the different sources

Data cleaning

Formatting, standardising, filtering and annotating the data

Data exploration

Visualising the data, extracting basic statistics, summarising it and discovering patterns

Data analysis

Statistical analysis, modelling and interpretation

Communication Deployment

Communication of results (publication, deployment)

Data wrangling

Time spent on these steps

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.

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.

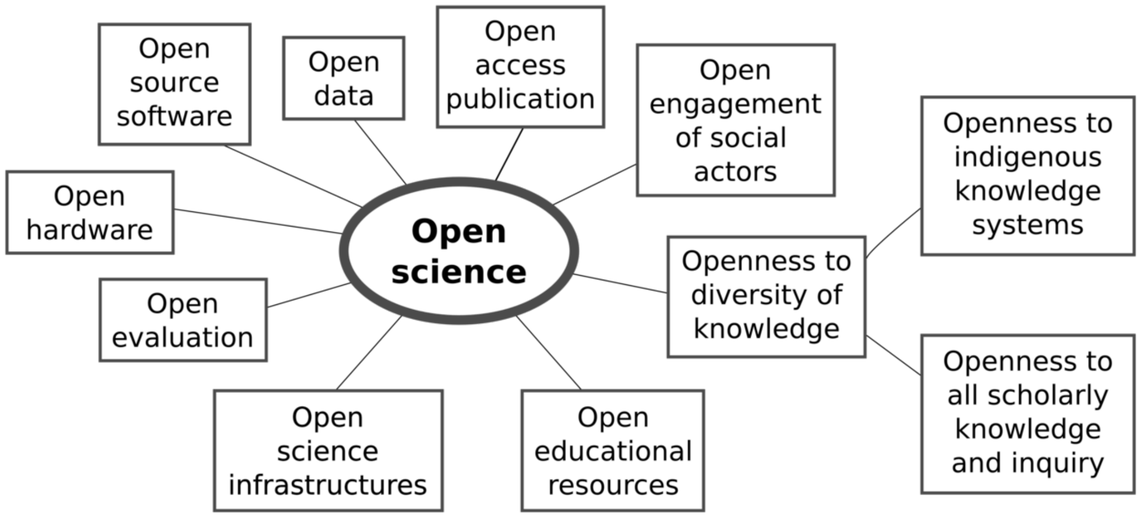
Data Annotation

- Helps us and machines to use and make sense of the data — **Data interpretation**
- Annotation is the process of **adding metadata**, **knowledge** or **labels** to the data e.g., the biology associated with the data
- Some annotations are:
  - **Free text** — appropriate for human interpretation
  - **Structured** using an **ontology** — human and machine understanding

What is Open Science

Impact, Contribution, Trust

- Make scientific research **accessible** to all levels of society:
  - Publications
  - Samples
  - Methods
  - **Software**
  - **Data**
- Advantages:
  - **Reproducibility** and **replicability**
  - Societal **responsibility** — publicly funded, publicly available
  - **Multi-purpose** of research outputs
- Disadvantages: concerns of data **misuse**



https://en.wikipedia.org/wiki/Open_science

Challenges Sharing and Reusing

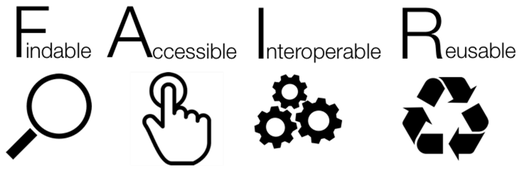
The marshmallow test — delayed gratification

- **Open does not mean FAIR**
- Requires an **effort**
- **Metadata** becomes **the most important data**
- In many cases there are **no standards or multiple ones**
- **Most** of the **data** out there **not FAIR**

https://imgflip.com/memegenerator

https://en.wikipedia.org/wiki/Stanford_marshmallow_experiment

FAIR Data and Software



- **F**indable and **A**ccessible
  - Add enough **metadata** — data about your data
  - Deposit your data in **public repositories** or make them available in **databases**
- **I**nteroperable:
  - Use **standard** and **open formats**
  - Provide **all data needed** to reproduce your analysis
- **R**eusable:
  - **Describe** your data well, e.g., good metadata but also
  - Attach a **license**

Minimum Information for Biological and Biomedical Investigations

Zenodo
Figshare
Pride
Metabolights
GEO
GitHub

Provide README files describing the data
Use descriptive column headers for the data tables

Standardisation and Ontologies

- Data **standardisation** requires defining **terminologies** and **vocabularies** that:
  - Assign **unique identifiers** to entities/concepts such as proteins, genes, diseases
  - **Describe** those entities/concepts and **provide meaning**
  - **Relate** those concepts to other terms
  - Classify those entities/concepts into **categories**
- **Solution —> Ontologies**
- **Ontology**:

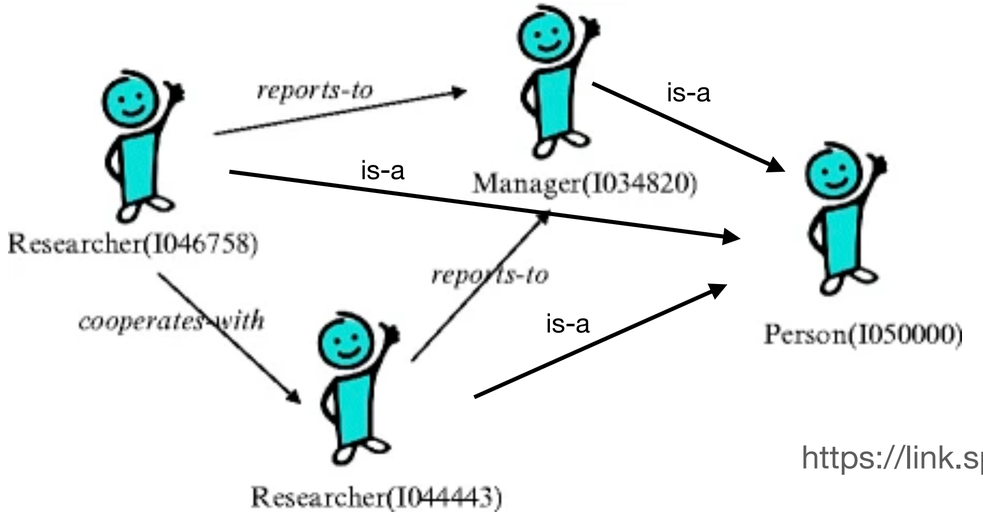
  formal way of representing knowledge in which concepts are described both by their meaning and their relationship to each other

  A collection of terms and their definitions for a specific domain

https://www.nature.com/articles/nrg1295

Ontologies

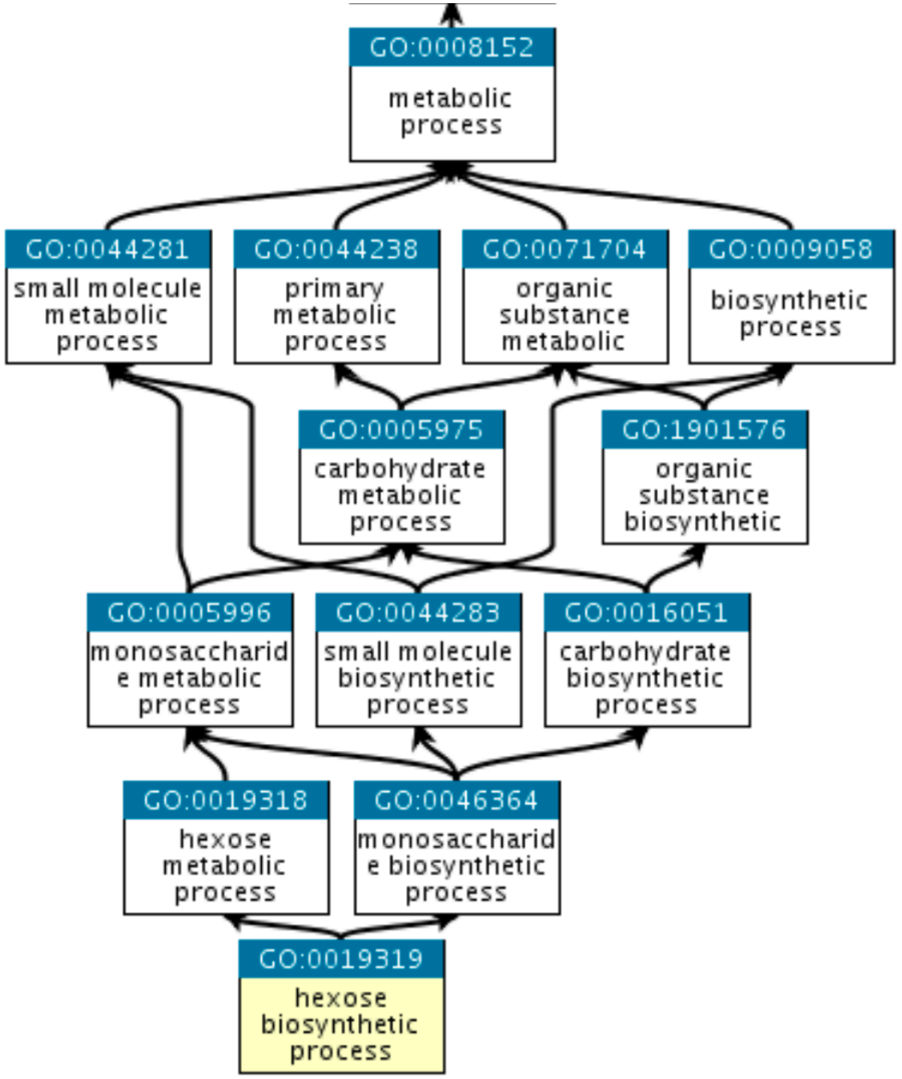
- An ontology is a formal **description of concepts** and **relationships formally modelling the structure of a system**
- The notion of ontologies is crucial for **enabling knowledge sharing and reuse**
- The backbone of an ontology consists of a **generalization/specialization hierarchy of concepts**



https://link.springer.com/chapter/10.1007/978-3-540-92673-3_0

Ontologies

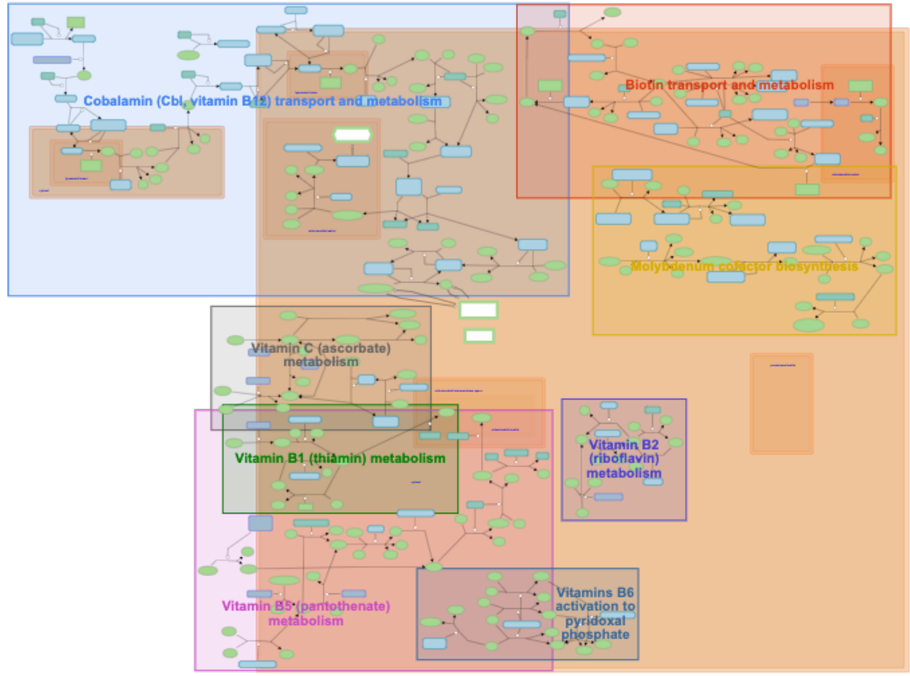
Gene Ontology



Disease Ontology

- disease
  - disease by infectious agent
    - bacterial infectious disease
    - fungal infectious disease
    - infective endocarditis
      - ornithine translocase deficiency
    - osteomyelitis
    - parasitic infectious disease
    - pelvic inflammatory disease
    - viral infectious disease
  - disease of anatomical entity
    - cardiovascular system disease
    - endocrine system disease
    - gastrointestinal system disease
    - hematopoietic system disease
    - immune system disease
    - integumentary system disease
    - musculoskeletal system disease
    - nervous system disease
    - reproductive system disease
    - respiratory system disease
    - thoracic disease
    - urinary system disease
  - disease of cellular proliferation
  - disease of mental health
  - disease of metabolism
  - genetic disease
  - physical disorder
  - syndrome

REACTOME Pathways



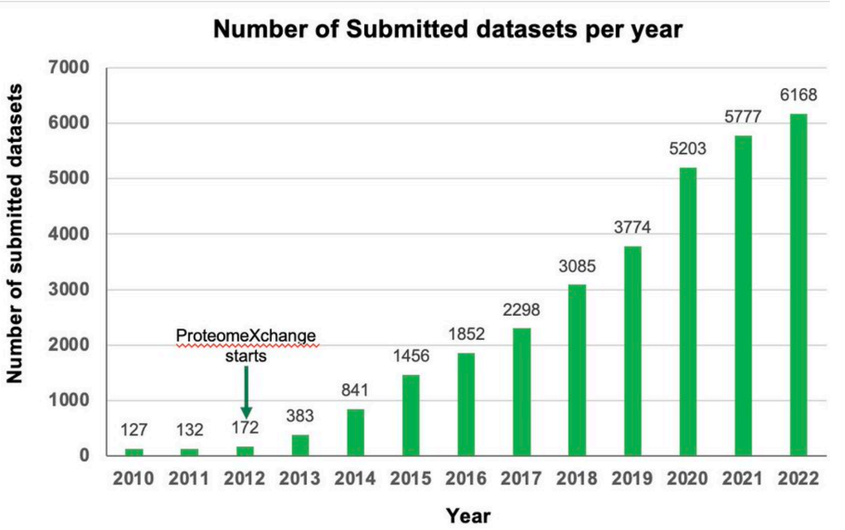
https://www.ebi.ac.uk/ols/ontologies

https://reactome.org/

http://geneontology.org/

Publicly Available Resources

Be a Data Parasite



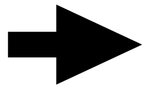
- **Do not reinvent the wheel**
- **Extend** the **life** and **purpose** of publicly available **data**
- Build **in-silico hypotheses** before jumping into experiments (cheaper, higher success rate)
- **Download — Use — Test — Transform — Upload**
- **Growing number** of **resources** and **datasets** available

https://www.ebi.ac.uk/pride/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258825/

Publicly Available Resources

- **3 main options** to use these resources:
  1. **Website queries**
  2. **Data download**
  3. **API**
- Also, **scraping** — https://realpython.com/python-web-scraping-practical-introduction/



Note: Do it responsibly — totally ok, but excessive use (too many request) can impact the platforms being scraped
Contacting the people behind the platform for collaboration may be more efficient

Some Resources

ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation https://aledb.org/

MiMeDB: the Human Microbial Metabolome Database https://mimedb.org/

Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes https://metatlas.nersc.gov/wom/project-begin.view

MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes http://www.liwzlab.cn/microphenodb

BacDive in 2022: the knowledge base for standardized bacterial and archaeal data https://bacdive.dsmz.de/

MASI: microbiota—active substance interactions database http://www.aiddlab.com/MASI/

iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning https://imodulondb.org/index.html

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters https://mibig.secondarymetabolites.org/

UniprotKB: is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. https://www.uniprot.org/

Saccharomyces Genome Database (SGD): provides comprehensive integrated biological information for the budding yeast Saccharomyces cerevisiae https://www.yeastgenome.org/

Use Case 1 — API

UniProt

- Provides material to learn the full potential of the database — Training

- All the information can be:
  - Queried online — https://www.uniprot.org/
  - Downloaded from the FTP server — https://ftp.uniprot.org/pub/databases/uniprot/
  - Accessed programmatically — https://www.uniprot.org/help/programmatic_access

Use Case 2 — Data Download

Gene Ontology (GO) Knowledgebase

- The world's largest source of information on the functions of genes
- All the information can be:
  - Queried online — https://geneontology.org/
  - Downloaded from the web — https://current.geneontology.org/products/pages/downloads.html
  - Accessed programmatically — https://geneontology.org/docs/tools-guide/#programmatic-download-bdbag

Molecular Function	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport".
Cellular Component	A location, relative to cellular compartments and structures, occupied by a macromolecular machine.
Biological Process	The larger processes, or 'biological programs' accomplished by multiple molecular activities.

Our Objective 2023

Developing a product

Data club

Streamlit

Problem Framing: Project Paper

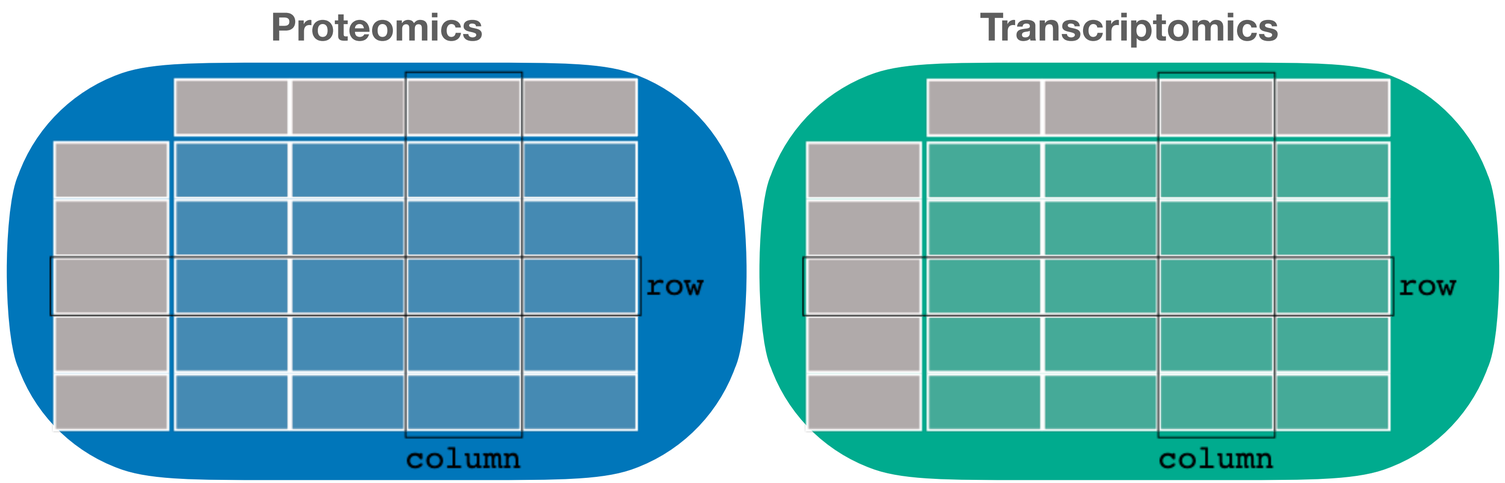
Proteome allocations change linearly with the specific growth rate of *Saccharomyces cerevisiae* under glucose limitation

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.



Practical

https://github.com/biosustain/data_club/tree/main/notebooks/data_annotation