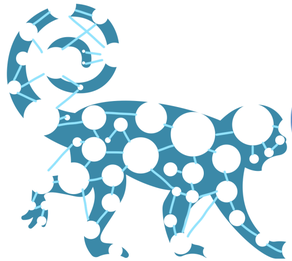
Leveraging Existing Biological Data and Making Sense of it

Finding, Accessing, Integrating and Reusing

Alberto Santos - Multi-omics Network Analytics

19 June 2023

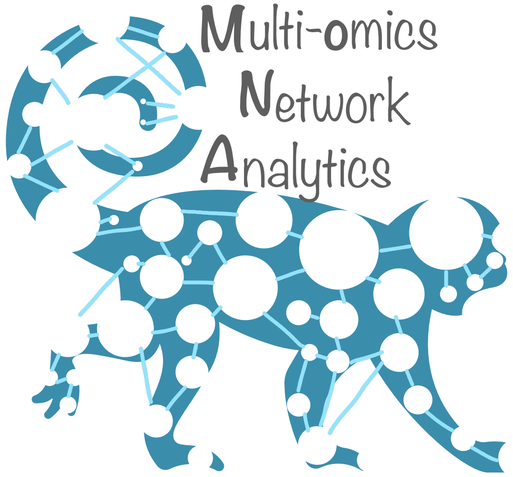


The Novo Nordisk Foundation
Center for Biosustainability

Outline

- Introduction
- No data, no fun
- Objectives / material
- Open science
- FAIR data
- Ontologies
- Data sources - databases
- Working with these databases - GUI/API
- Graphs/Networks
- Tools
- QA

Introduction



Mision

Getting a holistic view of **biological systems and their context** to understand their complexity and provide new insights and applications that can **benefit human and environmental health**

Located at: DTU Danmarks Tekniske Universitet — The Novo Nordisk Foundation Center for Biosustainability

Multi-omics

Using **multimodal data** to have a comprehensive view on **(micro) biology problems**

Network

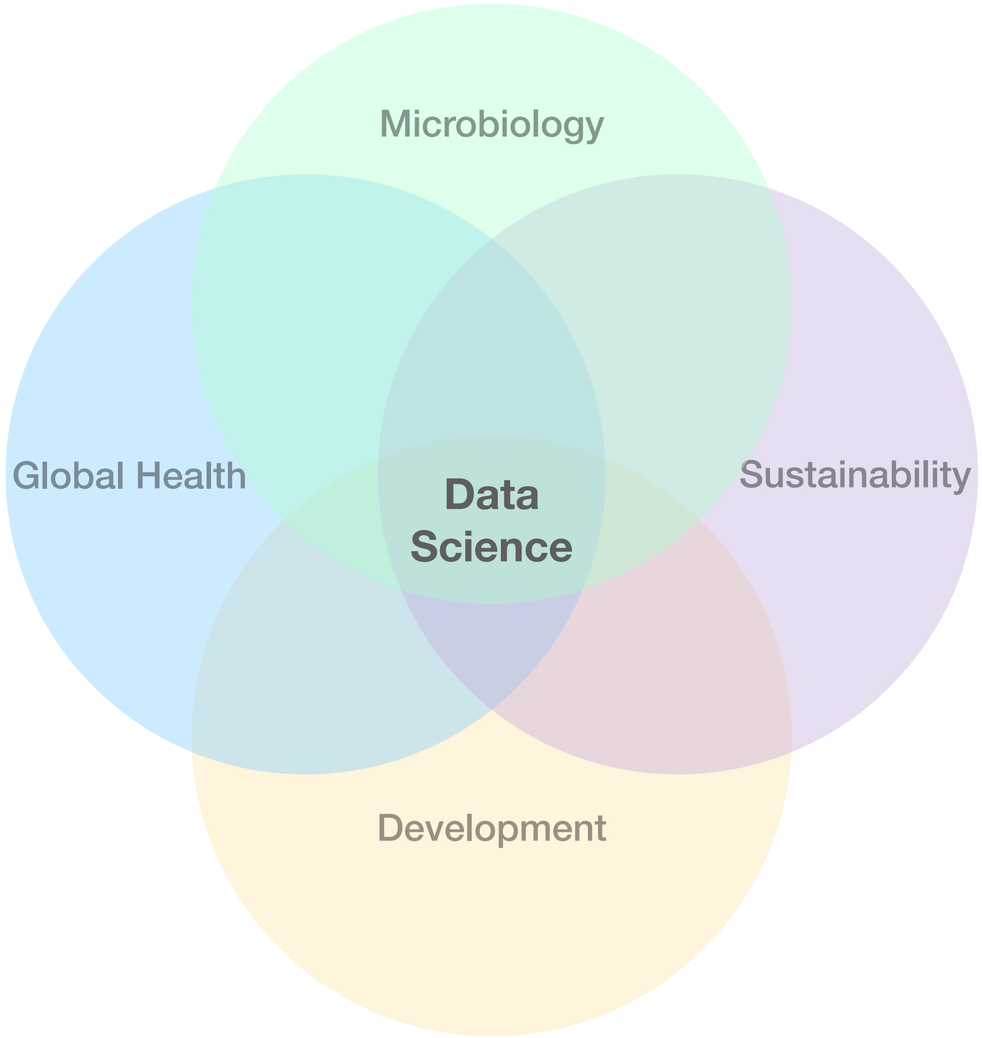
Exploiting **graphs** to **structure**, **represent**, **integrate** and **analyse** data

Analytics

Applying **Machine Learning** to answer **complex biological questions**

Research Areas

Data science at the intersection of microbiology, global health, sustainability, and development to tackle some of the world's most pressing challenges through data-driven approaches



Interconnected fields

Microbiology
Understanding the underlaying molecular factors that impact planetary and human health

Global health
Contributing to improving health outcomes in connection with climate change and biodiversity decline and with special interest in infectious diseases

Sustainability
Gathering and analysing data on environmental impacts, resource use, and the social and economic implications

Development
Efforts to improve the economic, social, and political well-being of communities and countries through education, analysis of inequality to climate change and resource depletion, etc.

Computational Lab

Omics

Internal
External
Meta-omics
Proteomics and **metabolomics**
Protein-protein Interactions (PPI)

Technology

Knowledge Graphs
Data integration
Natural Language Processing
Graph Machine Learning
Databases

Education

Data Science democratisation
Open science
FAIR data

Data

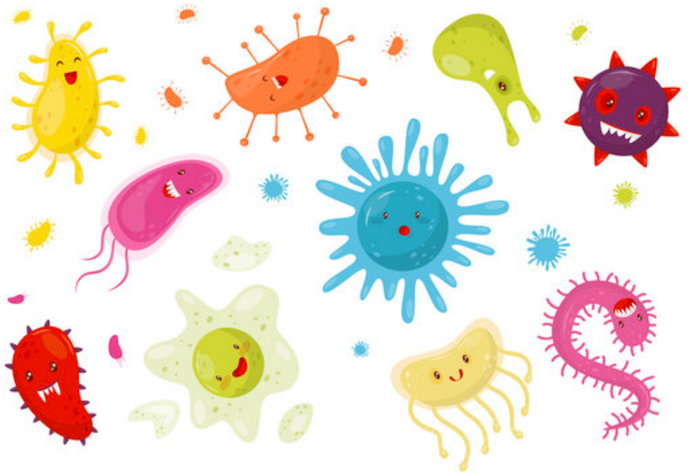
Objectives

- Show
  - The **value** of **Open** and **standardised data**
  - How to **generate** and **find** these data
  - How to **access** and **use** them
  - Some **examples**

Objectives

- Show
  - The **value** of **Open** and **standardised data**
  - How to **generate** and **find** these data
  - How to **access** and **use** them
  - Some **examples**

How to become a data parasite



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

Course website

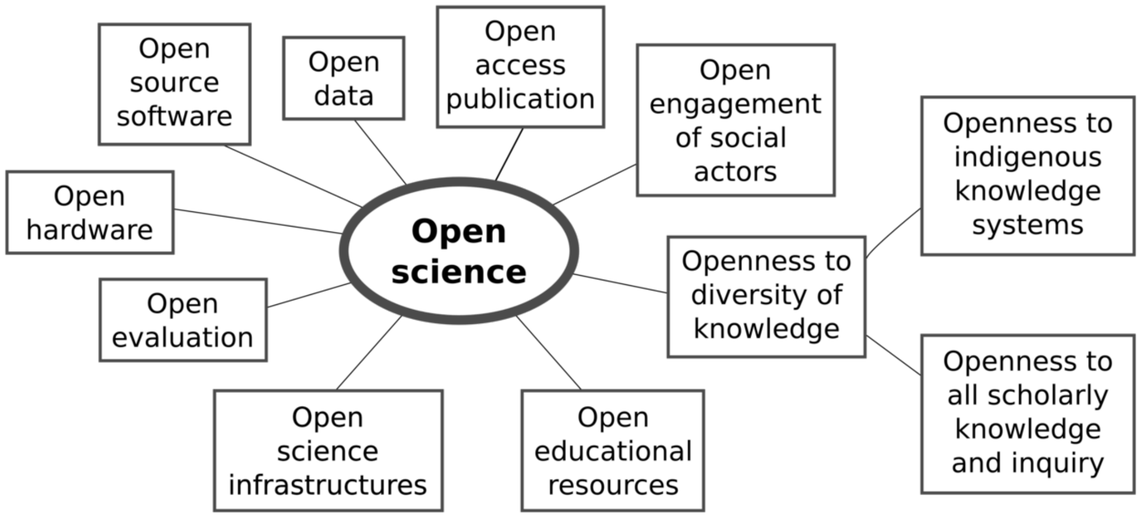
https://github.com/Multiomics-Analytics-Group/course_synthetic_biology_data

Open Science

What is Open Science

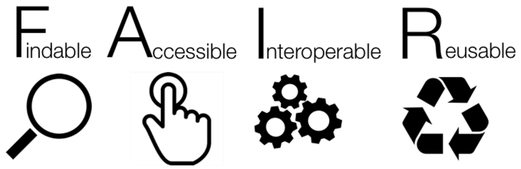
Impact, Contribution, Trust

- 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

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

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

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

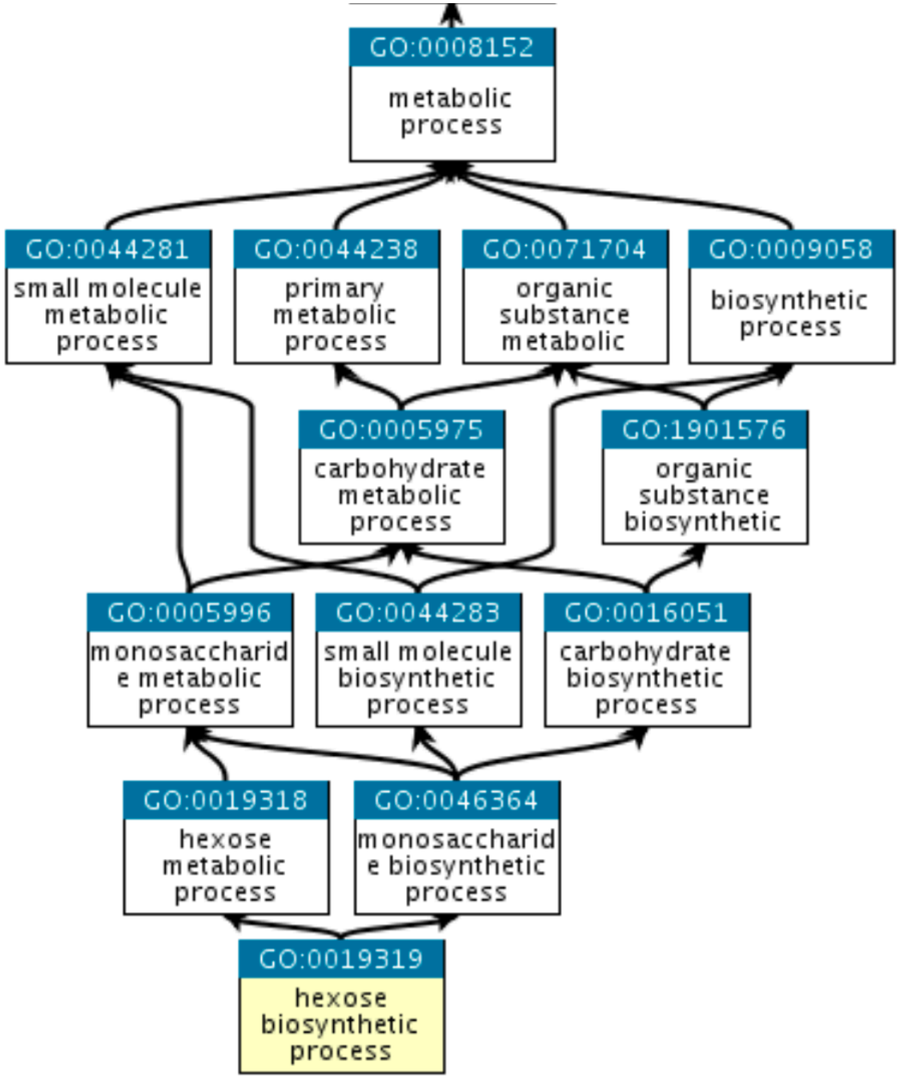
  A collection of terms and their definitions for a specific domain

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

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

Ontologies

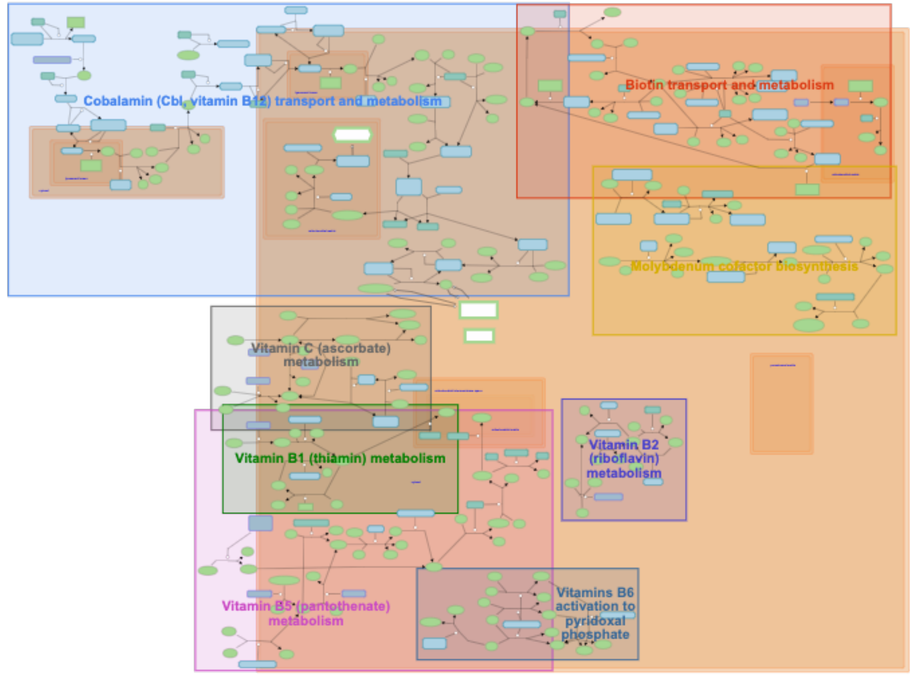
Gene Ontology



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



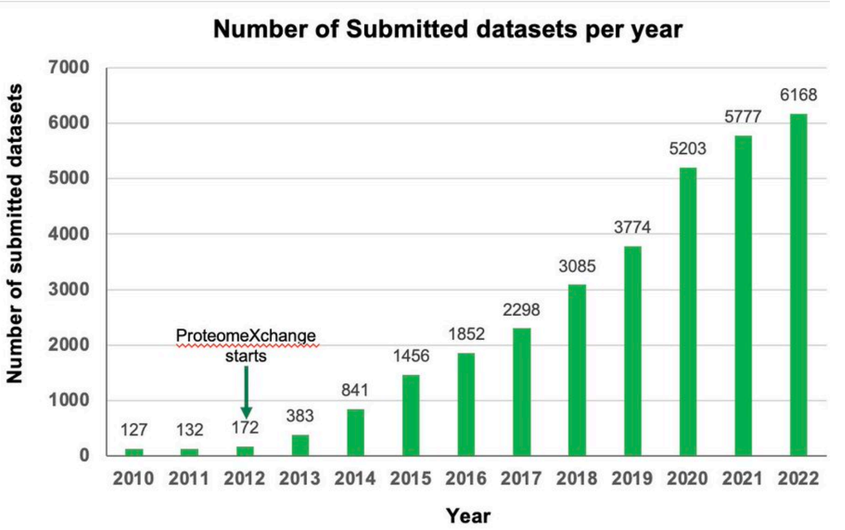
https://www.ebi.ac.uk/ols/ontologies

https://reactome.org/

http://geneontology.org/

Publicly Available Resources

PRIDE
PRoteomics IDEntifications Database



- **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/

Data Sources

NCBI

Organisms Database

- The **NCBI Taxonomy** database allows **browsing** of the **taxonomy** tree, which contains a classification of organisms
- Provides information about:
  - **Taxonomic identifier** for organisms
  - **Genomic overview**
  - Link-outs to **domain-specific databases**

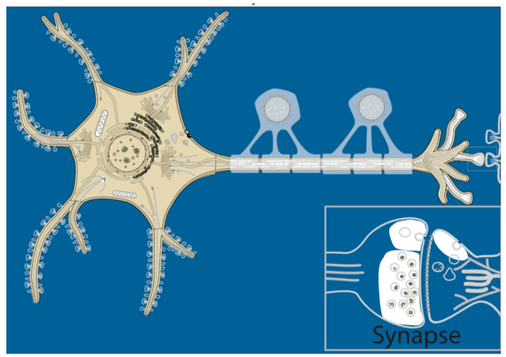
Entrez records			
Database name	Subtree links	Direct links	Links from type
Nucleotide	2,721,504	2,578,488	7,410
Protein	22,445,169	20,989,652	-
Structure	2,632	1,230	-
Genome	1	1	-
Popset	1,043	1,043	-
Conserved Domains	10	8	-
GEO Datasets	6,794	4,153	-
PubMed Central	111,167	111,167	-
Gene	128,094	3,304	-
SRA Experiments	41,306	33,055	-
GEO Profiles	141,600	141,600	-
Protein Clusters	7,547	7,547	-
Identical Protein Groups	3,556,802	3,524,631	-
BioProject	2,696	2,004	-
BioSample	56,942	49,071	15
Assembly	21,782	21,326	9
PubChem BioAssay	12,338	10,607	-
Taxonomy	383	1	-

https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

UniProt

Protein Database

- Comprehensive view on **proteins**
- Aggregates information about:
  - **Function**
  - **Sequence features**
  - **Subcellular localisation**
  - **Tissue expression**
  - **Disease association**
  - **Variants and PTMs**
  - **Interactions**
  - **Structure**

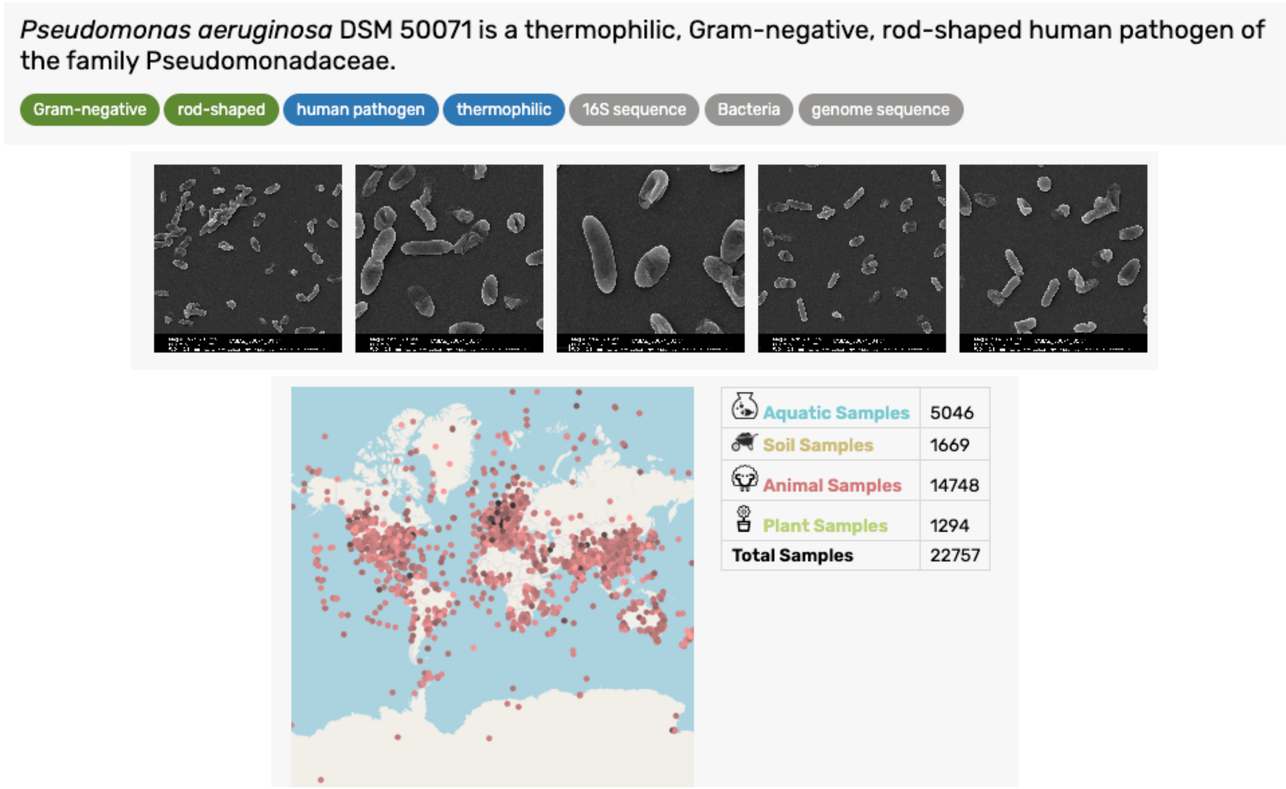


https://www.uniprot.org/

BacDive

Bacterial Database

- Largest database for **standardised bacterial phenotypic information**
- Collects information about:
  - **Bacterial classification**
  - **Morphology**
  - **Culture** and **growth** conditions
  - **Physiology** and **metabolism**
  - **Environment**
  - **Genome-based predictions**



https://bacdive.dsmz.de/

The Microbial Metabolites Database

MiMeDB

- The **human microbiome** is believed to produce or process **>55,000 different compounds** – many of which **affect** human **health**, **behavior** and **disease**
- **Microbes synthesise primary metabolites** required for their own survival, but they also **produce** other **compounds** arising **from substrates or host-derived food sources**

E.g., microbes transform xenobiotics from food constituents, food additives, phytochemicals, drugs, cosmetics and other exogenous or man-made chemicals

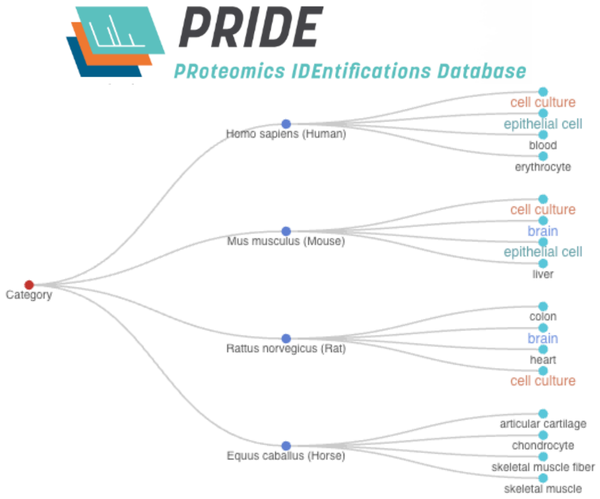
- **MiMeDB** is a **database** of **small molecule metabolites found** in the **human microbiome**
- Provides links between **metabolites**, **microbes**, **hosts**, **health** and **exposure** data

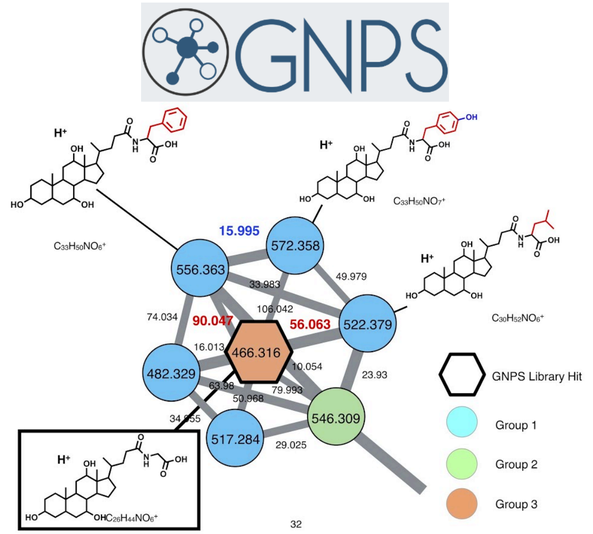
https://mimedb.org/

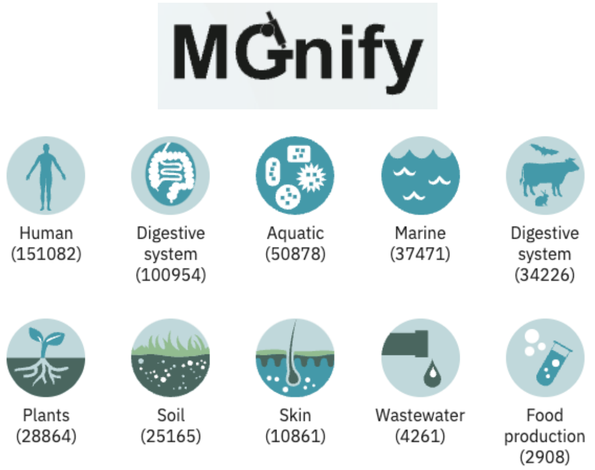
Metadatabases

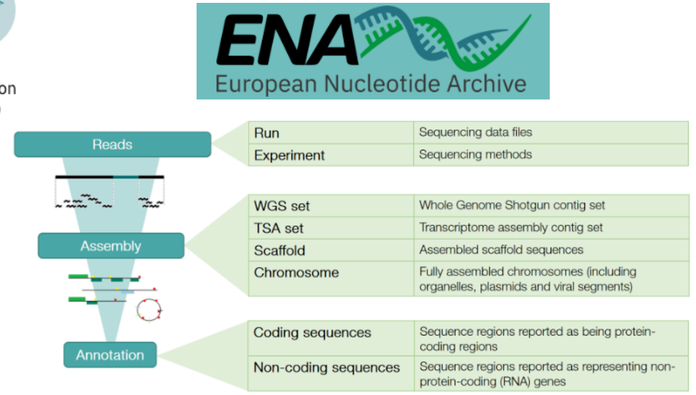
Data Repositories

- **PRIDE**: Mass Spectrometry-based (MS-based) Proteomics
- **GNPS:** MS-based Metabolomics
- **MGnify**: Microbiome data
- **ENA:** sequencing information, covering raw sequencing data, sequence assembly information and functional annotation







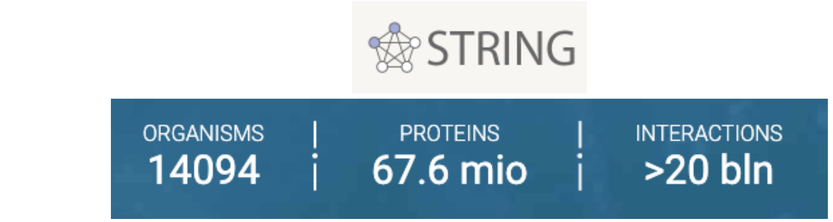


Protein-Protein Interactions

Interactions

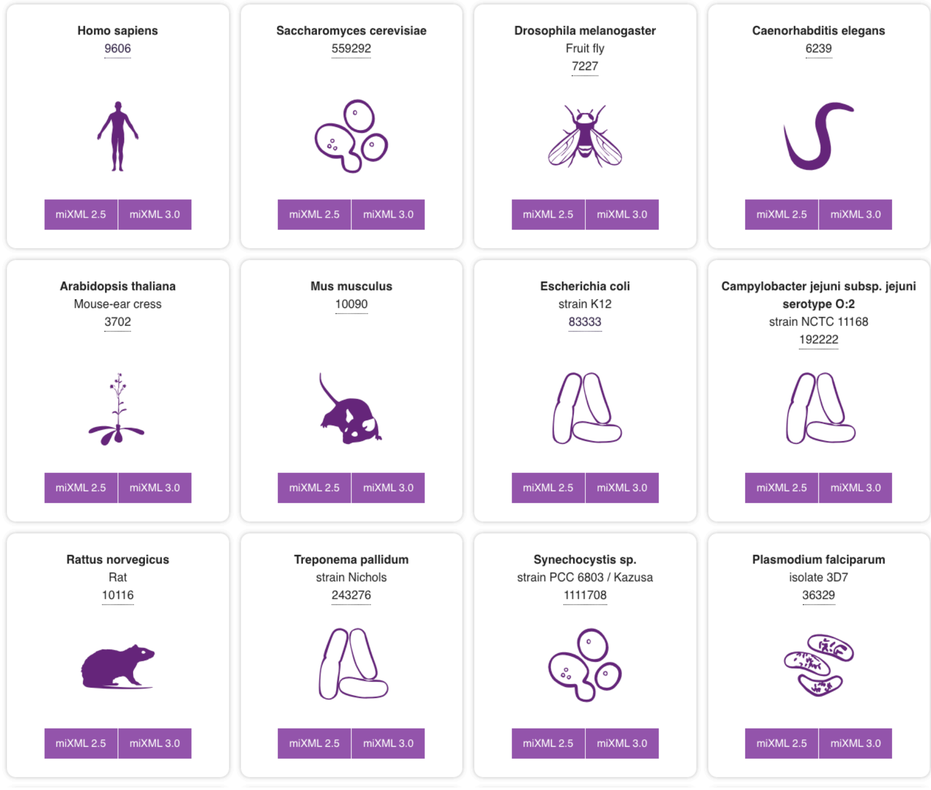
- **STRING**: PPIs and functional enrichment

- **Intact:** intra- and inter-species PPIs



https://string-db.org/

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



Other Resources

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

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/

FooDB: most comprehensive resource on food constituents, chemistry and biology https://foodb.ca/

Access

GUI vs API

Graphical **U**ser **I**nterface

- **View** the data
- **Facilitates** initial work
- Limited to the **interface**

Application **P**rogramming **I**nterface

- **Access** the data
- **Free** to do whatever you want
- Limited to **provided data**

Jupyter Notebooks

https://jupyter.org/

- **Web-based development environment** for **creating**, **running** and **sharing** **Python** (and other languages) code
- A **notebook** is an interactive document that combines **live code**, **equations**, **text or markdown**, and **visualisations** (output of your code)
- Notebooks are divided into **cells** that **run sequentially**! (Need to pay attention)
- It **requires** having **Python installed** on your local machine

Colab Notebooks

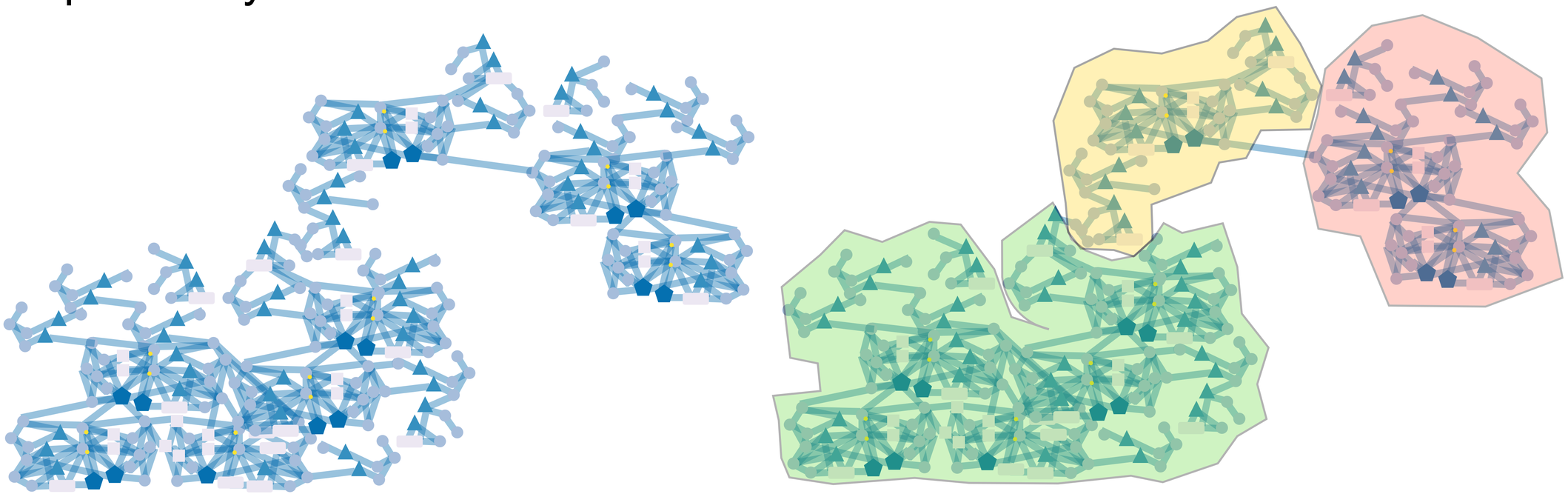
https://research.google.com/colaboratory/faq.html

- Google Colab is based on **Jupyter Notebook** open source project **hosted on Google's servers**
- Advantages:
  - Requires **no setup** to use (no python installation)
  - Provides **free** access to **computing resources** on Google's servers including GPUs
  - **Notebooks** can be **shared** just as you would with Google Docs or Sheets.
  - You can **import** existing **Jupyter notebooks**
- **Own data and notebooks** need to be accessed through **Google Drive** — Need Google account

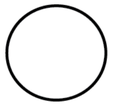
Graphs

What is a Graph/Network?

- Data structures of **components (nodes)** connected by **relationships (edges)**
- Allows us to **model** and **analyse data**
- Provides a **mathematical framework** to analyse complex relationships: Graph theory

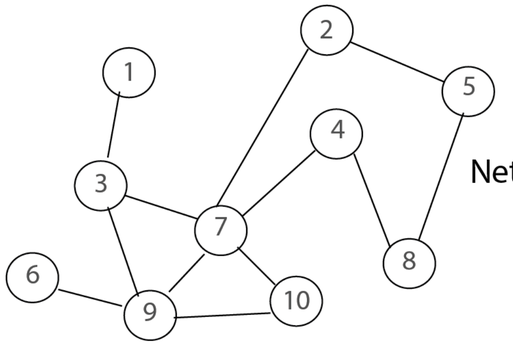


Graphs



Node

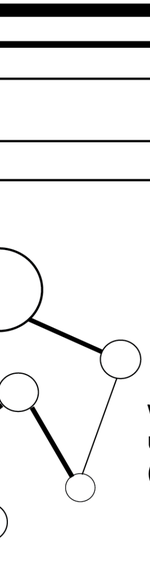
Edge



Network

$$\begin{bmatrix}
0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&1&0&1&0&0&0\\
1&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&0&1&1&0&0\\
0&1&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&1&0\\
0&1&1&1&0&0&0&0&1&1\\
0&0&0&1&1&0&0&0&0&0\\
0&0&1&0&0&1&1&0&0&1\\
0&0&0&0&0&0&1&0&1&0
\end{bmatrix}$$

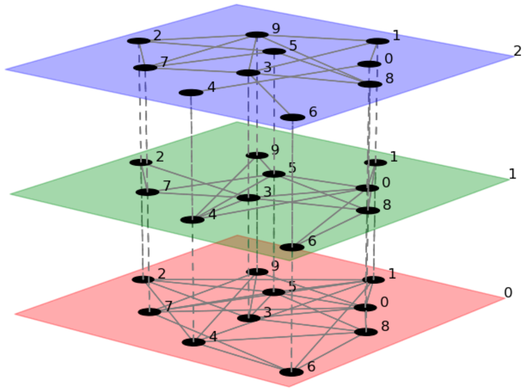
Adjacency matrix

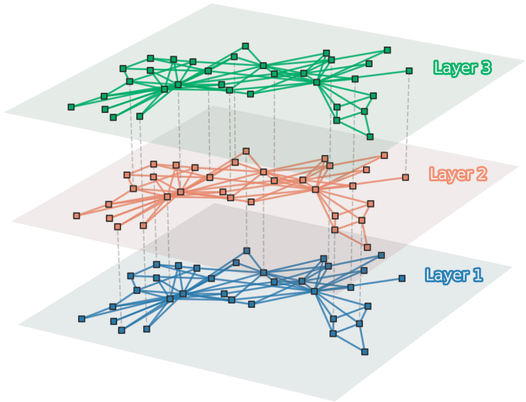


weighted
undirected network
(thickness)

$$\begin{bmatrix}
0&0&w_{13}&0&0&0&0&0&0&0\\
0&0&0&0&0&w_{25}&0&w_{27}&0&0\\
w_{31}&0&0&0&0&0&w_{37}&0&w_{39}&0\\
0&0&0&0&0&0&w_{47}&w_{48}&0&0\\
0&w_{52}&0&0&0&0&0&w_{58}&0&0\\
0&0&0&0&0&0&0&0&w_{69}&0\\
0&w_{72}&w_{73}&w_{74}&0&0&0&0&w_{79}&w_{7,10}\\
0&0&0&w_{84}&w_{85}&0&0&0&0&0\\
0&0&w_{93}&0&0&w_{96}&w_{97}&0&0&w_{9,10}\\
0&0&0&0&0&0&w_{10,7}&0&w_{10,9}&0
\end{bmatrix}$$

Weighted
adjacency matrix

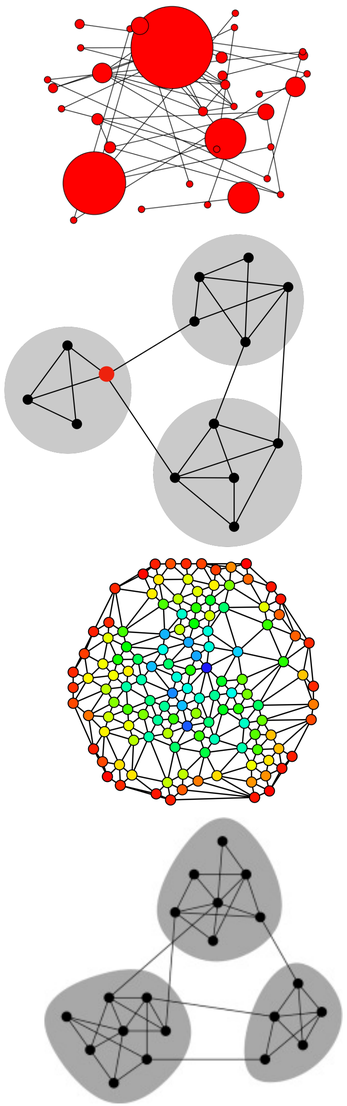




Why Graphs?

- These structures allow:
  - Quick **integration** of **heterogeneous data** based on relationships
  - **Graph theory** methods can be used to **analyse** and **interpret** data, e.g., topological properties can be used to explain:
    - The possible **role** of specific components
    - The **flow** of information
    - The **robustness** of the system
- **Visualize** data

Graph Metrics



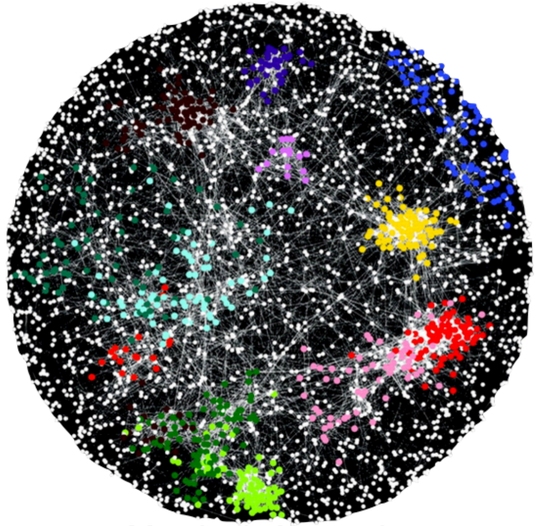
- **Degree**: number of edges in (in-degree) and out (out-degree) of a node
  node importance
- **Connectivity**: minimum number of elements (nodes or edges) that need to be removed to get two or more isolated subgraphs
  graph resilience
- **Distance:** length of shortest path between two nodes
- **Centrality**: ranking of nodes according to their position
  node importance
    - **Degree centrality**
    - **Closeness centrality:** the average length of the shortest path between the node and all other nodes in the graph
    - **Betweenness centrality:** number of times a node acts as a bridge along the shortest path between two other nodes
    - **Eigenvector centrality:** It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more
    - **Katz centrality:** number of all nodes that can be connected through a path (distant nodes are penalised)
- **Clustering:** partitioning a graphs into different groups that share some form of similarity

https://en.wikipedia.org/wiki/Degree_(graph_theory)
https://en.wikipedia.org/wiki/Distance_(graph_theory)
https://en.wikipedia.org/wiki/Centrality
https://en.wikipedia.org/wiki/Connectivity_(graph_theory)
https://www.sciencedirect.com/topics/computer-science/clustering-graph

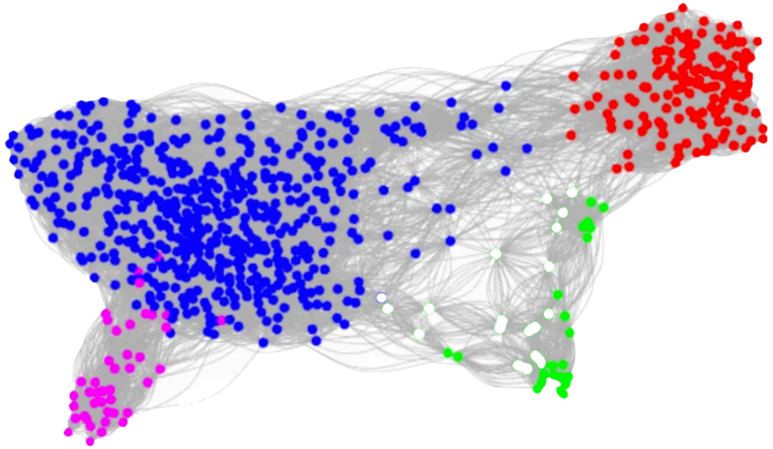
Graphs in Biology

https://towardsdatascience.com/umap-for-data-integration-50b5cfa4cdcd
http://snap.stanford.edu/deepnetbio-ismb/ipynb/Human+Disease+Network.html
https://cytoscape.org/cytoscape-tutorials/presentations/ppi-tools1-2017-mpi.html#/
https://en.wikipedia.org/wiki/Metabolic_network
https://www.scienceandfood.org/the-flavor-network/

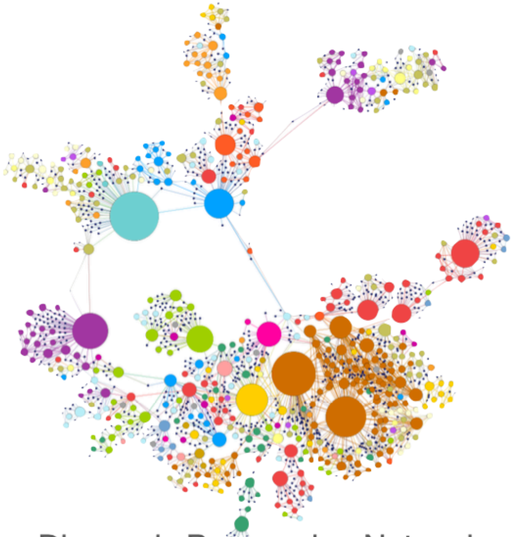
Protein-protein Interaction Networks



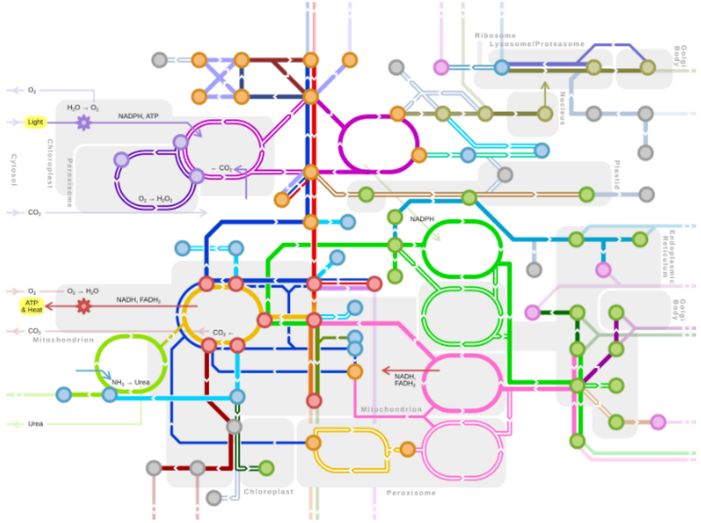
Single cell Networks



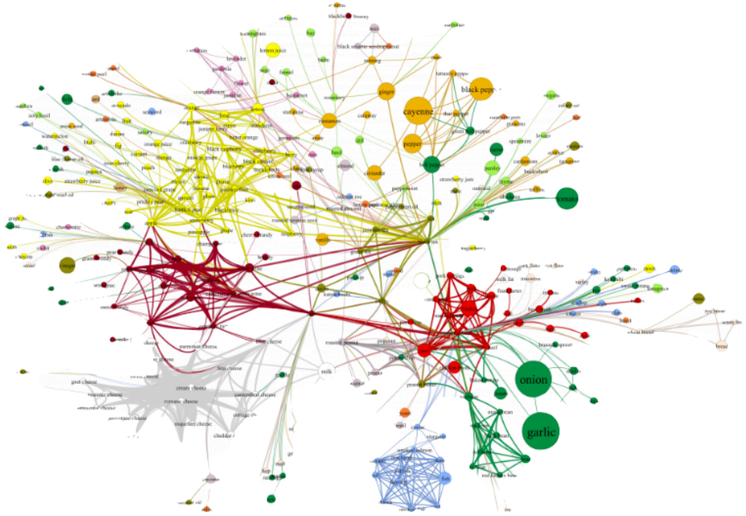
Disease Networks



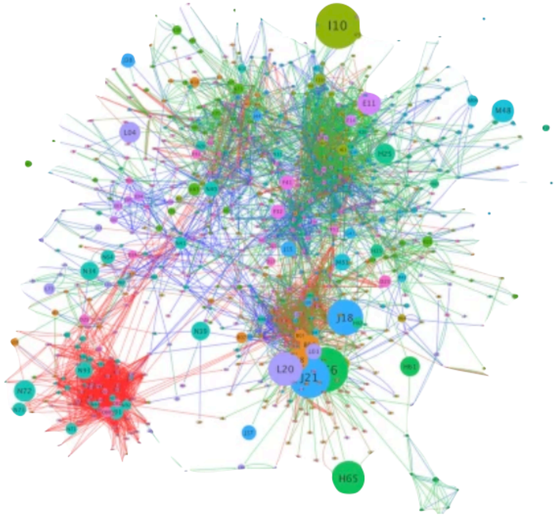
Metabolic Networks



Food Networks



Diagnosis Progression Networks

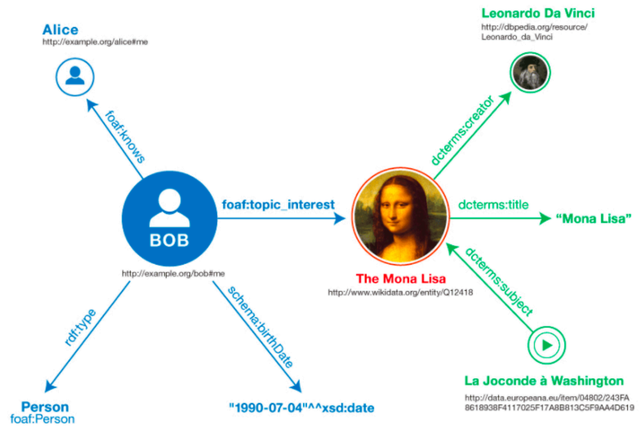


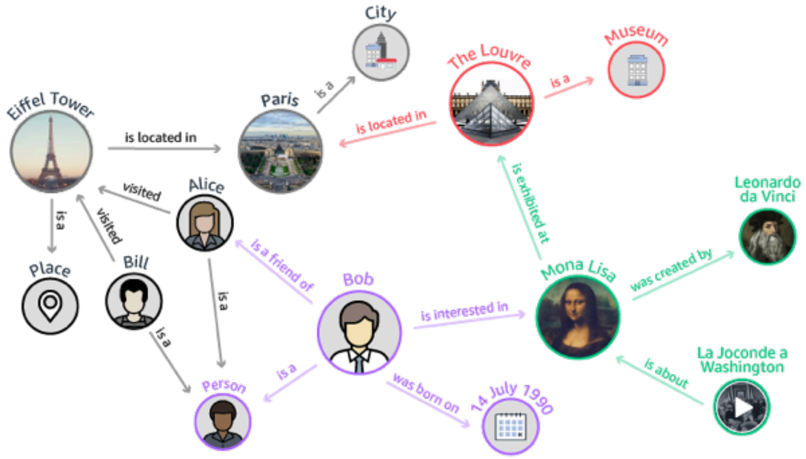
Knowledge Graphs

What is a Knowledge Graph (KG)

Relationships firsts everything else second

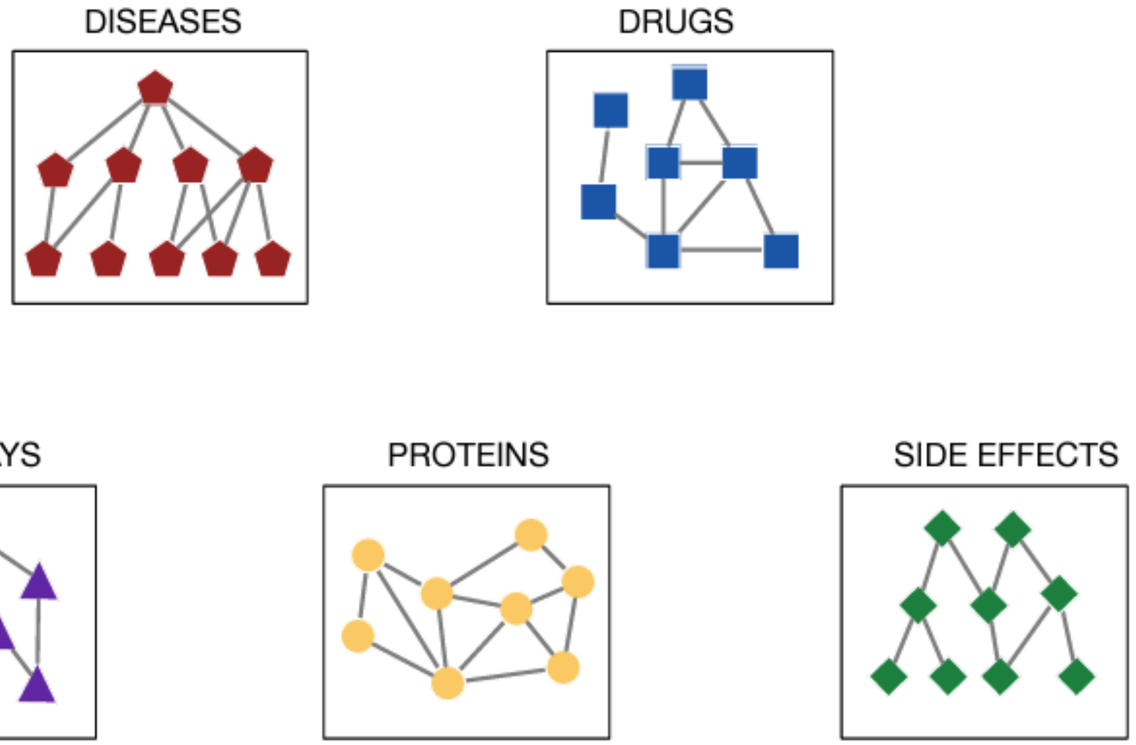
- A way to organise **knowledge**/information by defining **associations** or **relationships**
- These relationships facilitate the **integration**, **management** and **enrichment** of data
- The **objectives** when setting up a KG:
  - Standardisation
  - Quality
  - Reusability
  - Interpretability
  - Automation
  - Representation/Visualisation





The Knowledge Graph Cookbook. Andreas Blumauer and Helmut Nagy. 2020

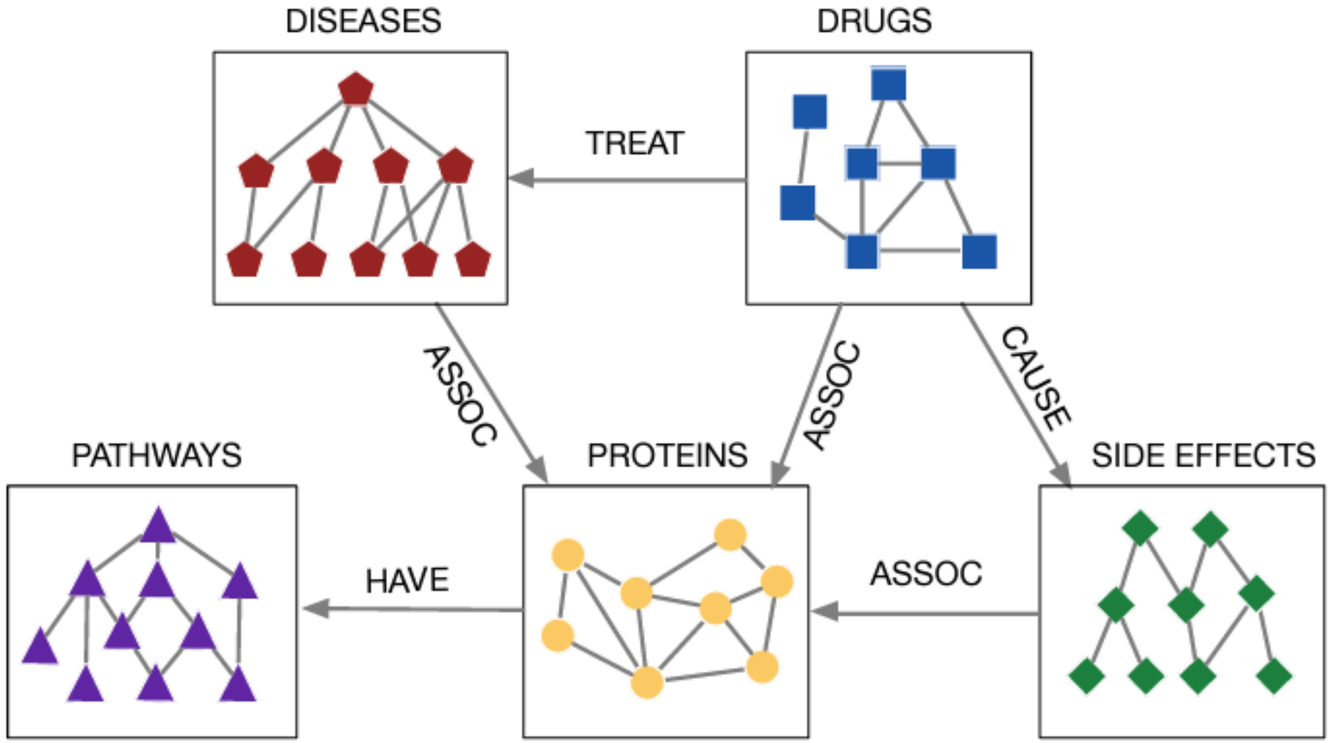
Knowledge Graph vs Graph Database

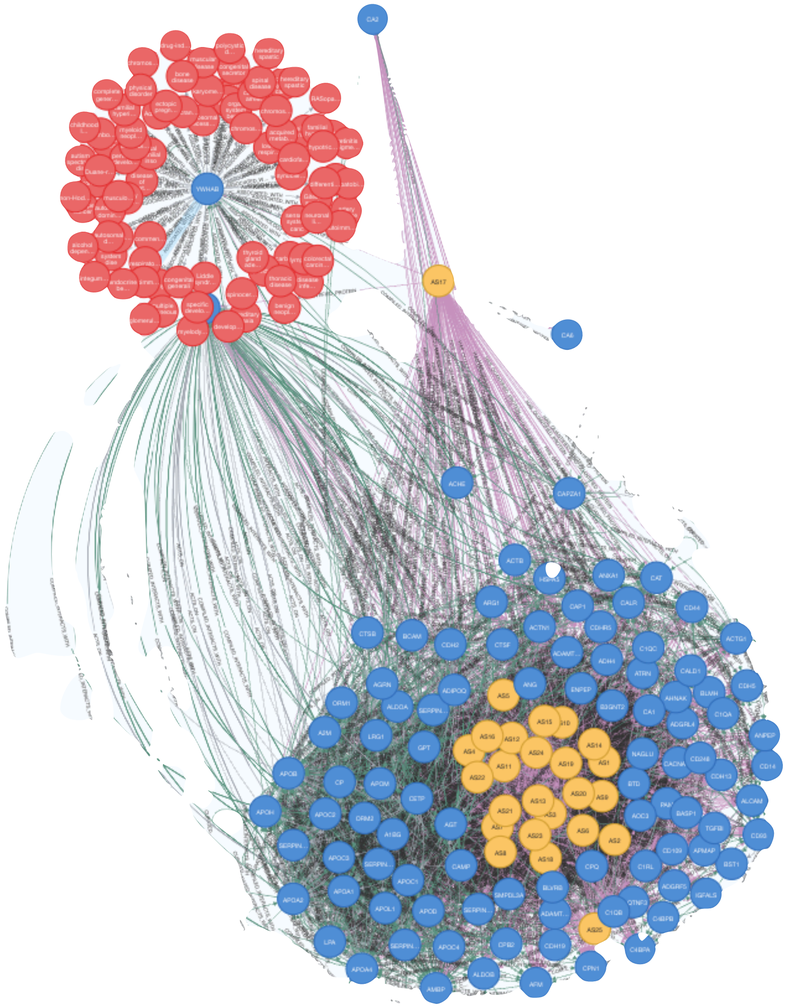


https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

Knowledge Graph vs Graph Database

Focus on **data integration** to **represent complex biological systems** and be able **to reason over them**





https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

Tools

Cytoscape

https://cytoscape.org/

- An open source **software platform** for **visualising** and **analysing** complex **networks**
- Used for any kind of networks but **specialised** on **biological domains**:

  e.g, Molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data.
- **Additional features** are available as freely available **Apps** (https://apps.cytoscape.org/)

Getting Started

Gephi

https://gephi.org/

- Open source **platform** for **graph visualisation** and **analysis**
- Used across **many domains** and by **many industries**
- **Performs well** for large graphs
- **Connects to Graph databases** such as Neo4j
- Some of the **visualisations and analysis** require **plugins**

Getting Started

Code

https://networkx.org/

https://r.igraph.org/

QA

Thank you

albsad@dtu.dk

@albsantosdel

https://github.com/albsantosdel