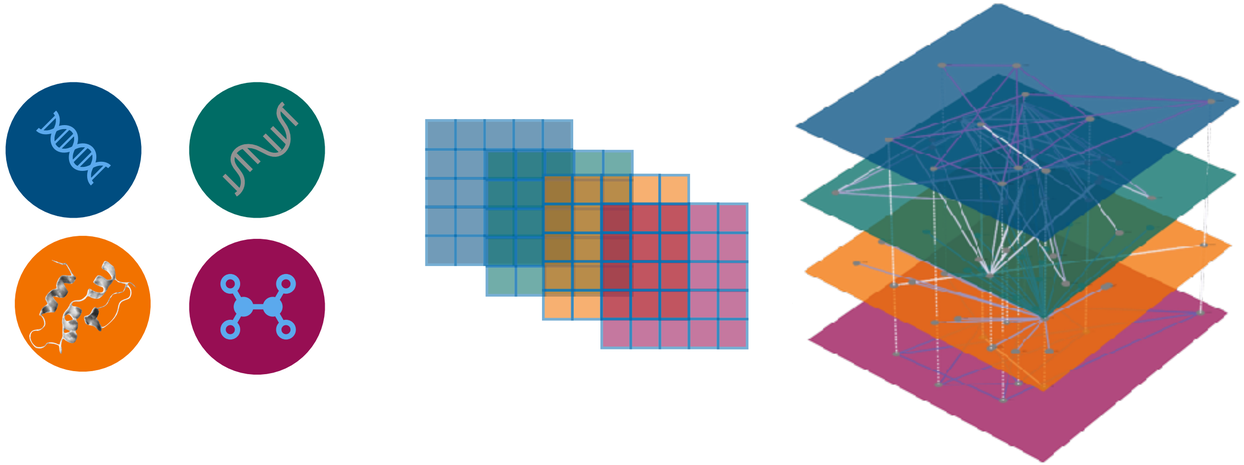


Graphs

AI-guided Protein Science

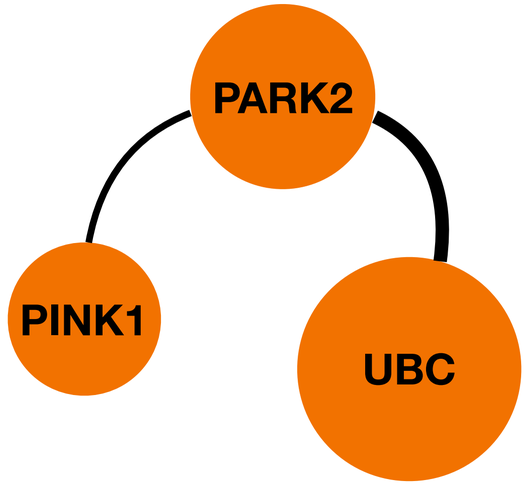
Alberto Santos — Multi-omics Network Analytics (MoNA)

The Novo Nordisk Foundation
Center for Biosustainability

Biology as a System

- Understanding biology **holistically** and **integratively** — structures, functions, and interactions.
- Biological systems are **complex systems** difficult to model and predict their responses to perturbations or interventions.

Protein	Intensity
PARK2	21421
PINK1	2456
UBC	77632



Modelling Biological **Systems**

Not walking away from complexity

- **Holistic understanding** — Studying biology as systems allows us to understand the **complex interactions** and **dependencies** among biological components, beyond isolated parts
- **Integrative approach** — Aggregating **data** from **multiple disciplines** and **technologies** helps to better understand life processes
- **Emergent properties analysis** — Many biological phenomena cannot be understood by studying components individually but **emerge from their interactions**
- **Real-world applications** — A systems perspective aims to **model complex global challenges**

Network Biology

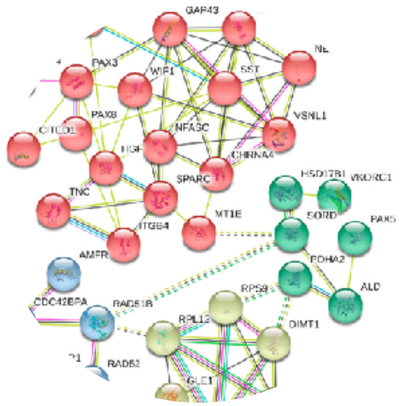
- Helps us to **visualise** and **analyse complex biological systems** as networks of interactions (e.g., protein-protein, gene regulatory, or metabolic pathways), aiding in understanding their structure and function
- Networks model **diverse biological datasets** into interpretable frameworks, enabling researchers to uncover **patterns** that might be overlooked otherwise (e.g., multi-omics data — genomics, proteomics, etc.)
- Researchers can use **network properties** to identify **critical nodes** or **pathways** disrupted in specific conditions like disease or define how biological systems maintain function, guiding strategies to enhance resilience or address fragility
- **Computational models** using network structures allow **prediction** of outcomes, reducing reliance on trial-and-error experiments

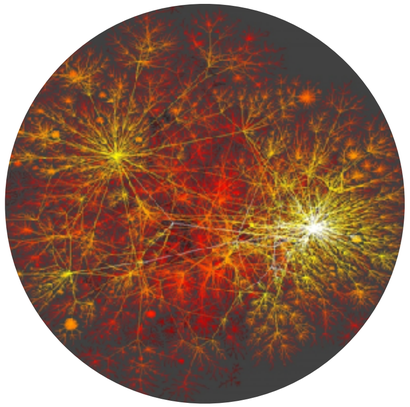
What is a Graph/Network?

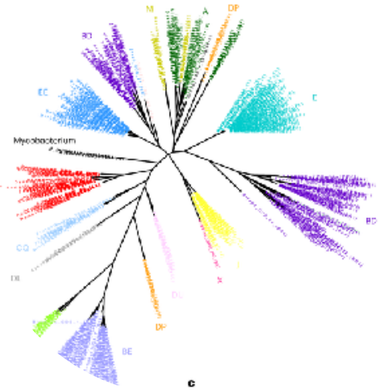
- Data structures of **components (nodes)** connected by **relationships (edges)**

Social networks

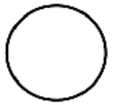
Biological networks





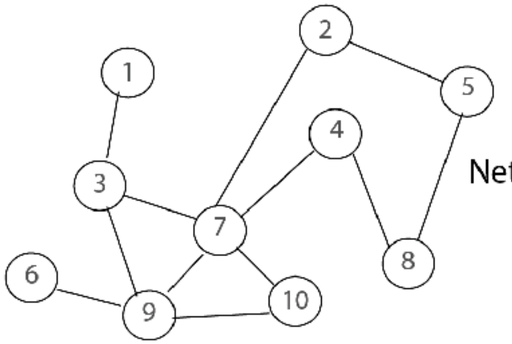


Graphs



Node

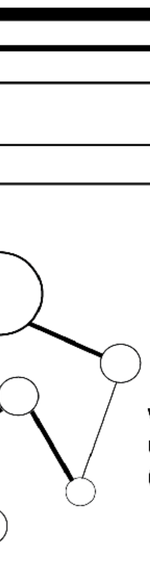
Edge



Network

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

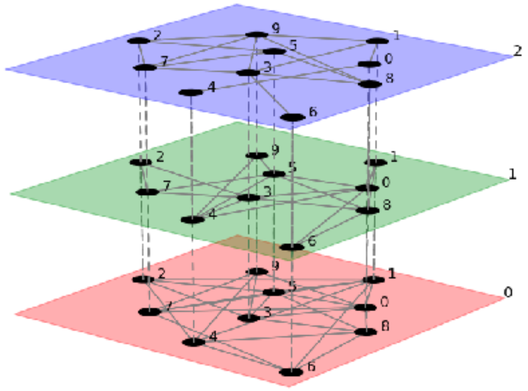
Adjacency matrix

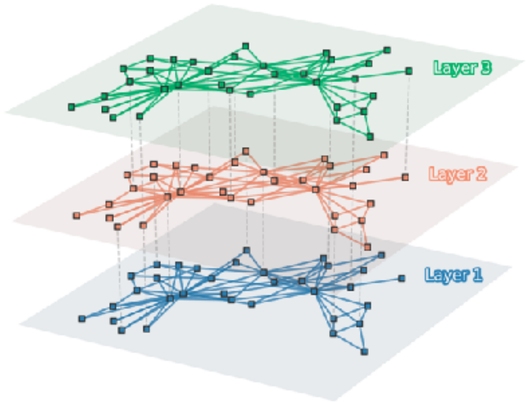


weighted
undirected network
(thickness)

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

Weighted
adjacency matrix





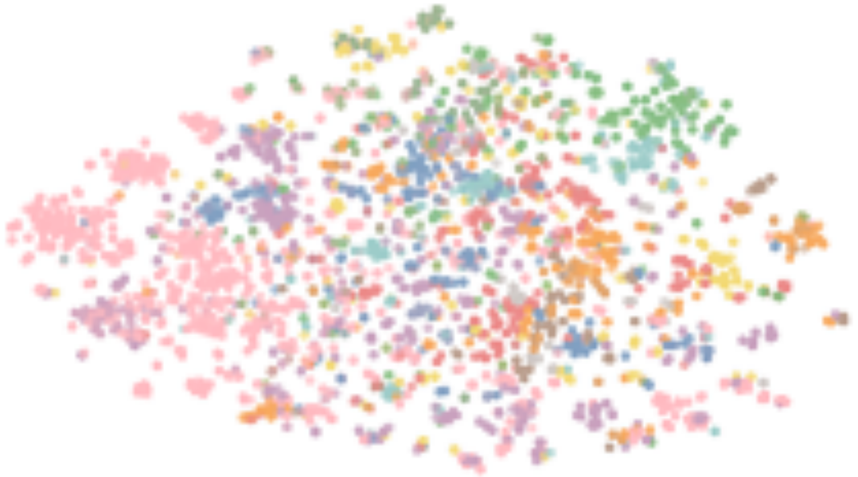
Why Graphs?

- 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
- Focus on **relationships** between points rather than the properties of individual points

How to Analyse Graph Structures

Using and Analysing Relationships

- **Graph Theory**: algorithms that allow you to extract relevant information from the topology of the graph.
  - **Topological Features:** Centrality, degree, clustering, etc.
- **Graph Machine Learning**:
  - Embeddings
  - Graph Neural Networks

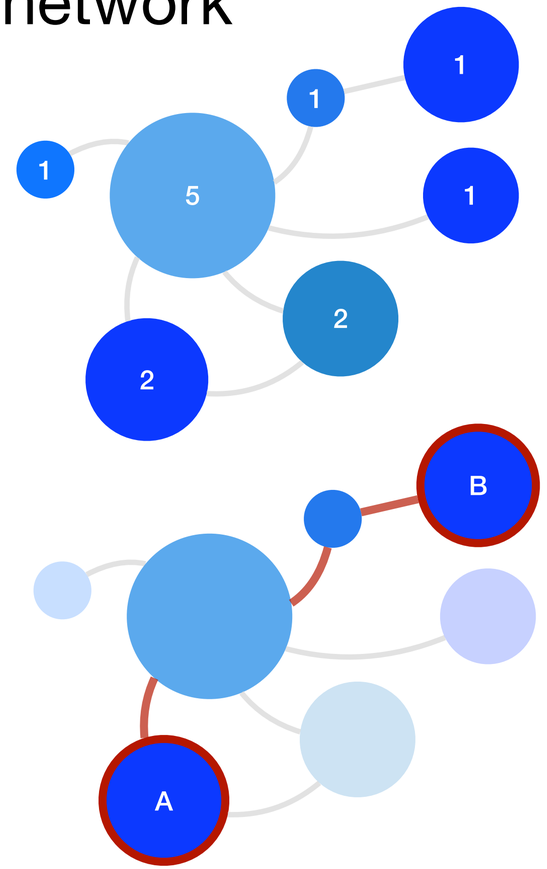


https://huggingface.co/blog/intro-graphml

Graph Theory — Some Topological Properties

Topological properties can help extract meaningful information and identify relevant structures within the network

- **Degree —** Measures the number of connections (edges) a node has in the network. Identifies highly connected nodes (hubs) that often represent critical molecules or interactions, such as essential proteins or highly expressed genes.
- **Path length —** Average shortest path between all pairs of nodes. Indicates the signal flow across the network, relevant for instance in signaling pathways.
- **Shortest Path —** The minimum number of edges required to traverse between two nodes. Essential for studying signal transduction, metabolic fluxes, and the efficiency of molecular or ecological interactions. Nodes with many shortest paths passing through them often have critical roles in the system.
- **Clustering Coefficient —** Measures the tendency of nodes to form tightly knit groups. High clustering often signifies functional specialization, such as protein-protein interaction clusters in cellular compartments.

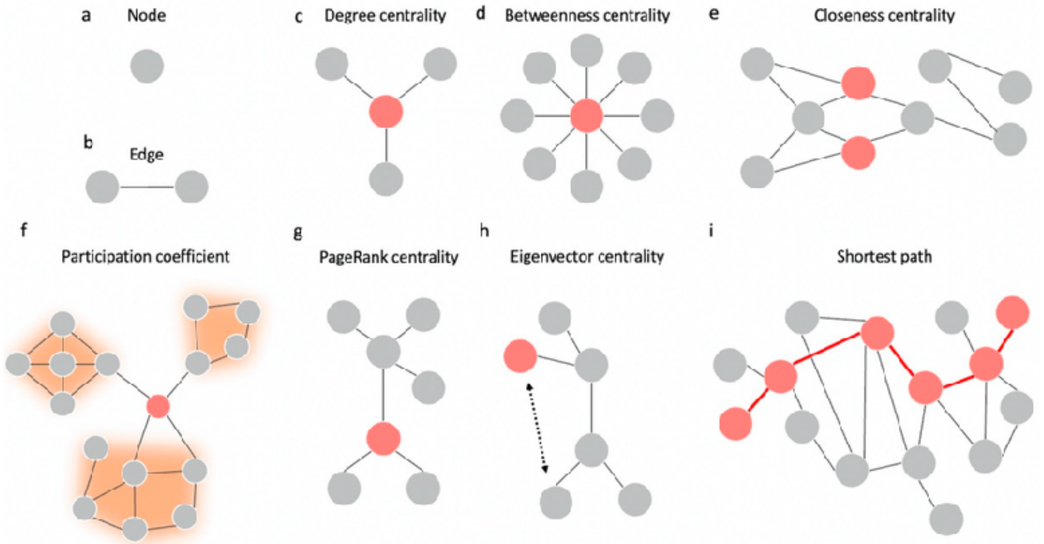


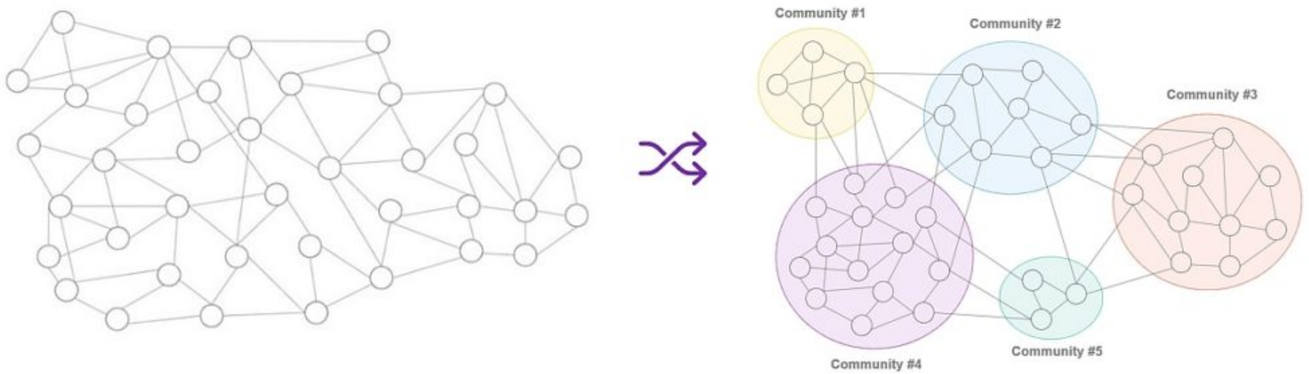
https://timbr.ai/community-detection-algorithm/

Graph Theory Measures and Their Application to Neurosurgical Eloquence. Cancers 2023

Some Topological Features

- **Centrality —** set of metrics that determine the importance or influence of a node within a network. Different centrality measures highlight different aspects of importance based on the network's structure
    - **Betweenness Centrality**: Reflects the frequency with which a node appears on the shortest paths between other nodes. Important for identifying key regulators or bottlenecks in pathways.
    - **Closeness Centrality**: Measures how close a node is to all other nodes, indicating its ability to quickly interact or influence others.
    - **Degree Centrality**: Identifies the most connected nodes, which may play pivotal roles in stability or robustness.
    - **Eigenvector Centrality**: Considers the influence of a node based on the importance of its neighbors, helping locate influential components in signaling or metabolic networks.
- **Community —** clusters or groups of nodes within a network that are more densely connected to each other than to nodes outside the group. These clusters, or communities, are also known as modules or sub-networks





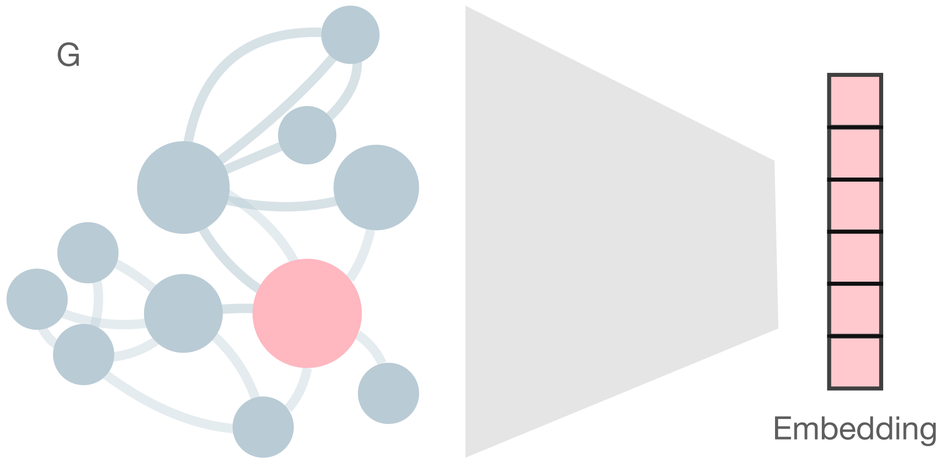
https://timbr.ai/community-detection-algorithm/

Graph Theory Measures and Their Application to Neurosurgical Eloquence. Cancers 2023

Embeddings

Representing graph structures

- **Large biomedical networks** require **high computation and space** —> **dimensionality reduction** techniques to represent nodes and edges
- **Graph Embedding**: **graphs**, **nodes** and **edges** represented as **numerical features** in a low dimensional space
- Use for:
  Node classification
  Node recommendation
  Link prediction
- Methods:
  - Shallow/Walk-based approaches: Node2Vec, Path2Vec, etc.
  - Graph Neural Networks
- **Python libraries:** NetworkX, DGL, StellarGraph, PyGraphML

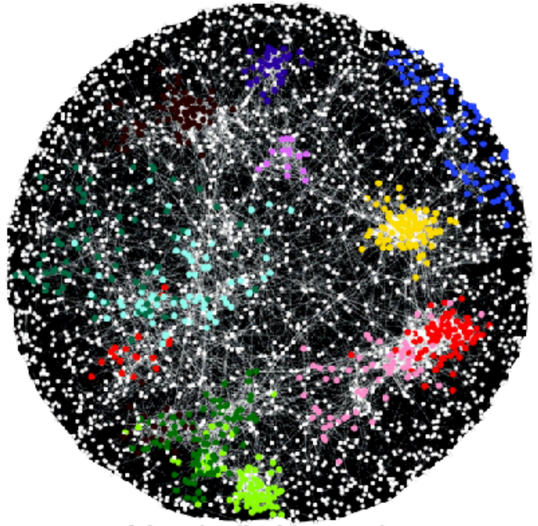


https://www.nature.com/articles/s41467-022-33026-0

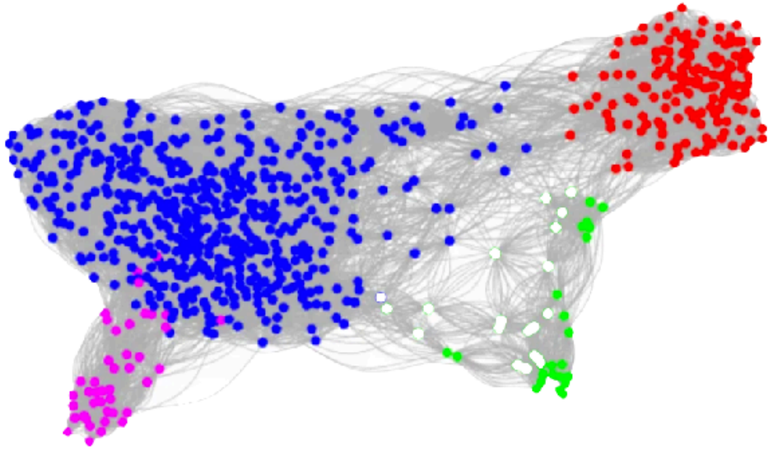
Graphs in Biology

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

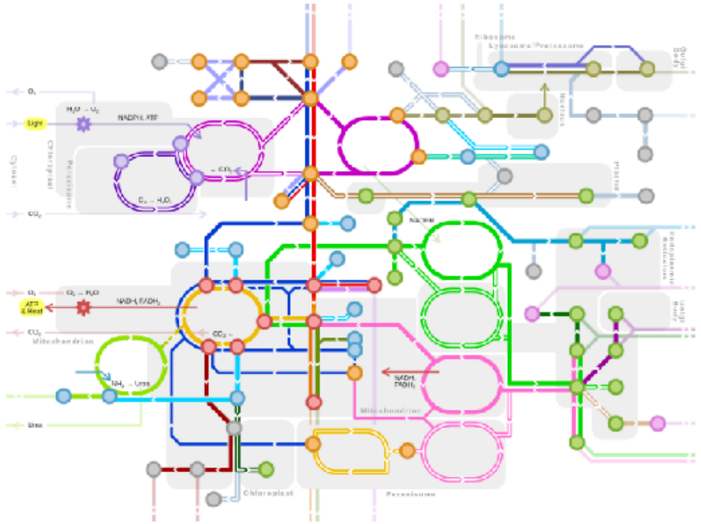


Single cell Networks

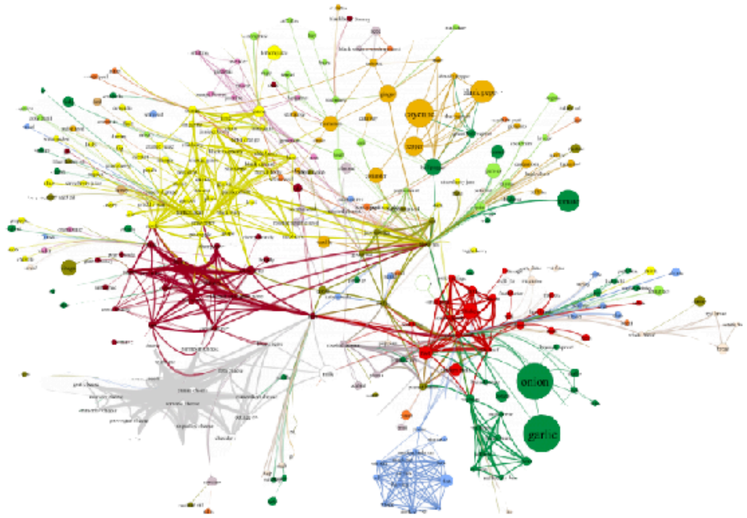


Disease Networks

Metabolic Networks



Food Networks



Diagnosis Progression Networks

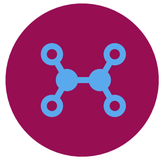
Omics Graphs

Types of Omics

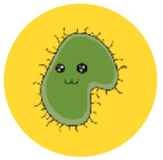
- **Genomics** Study of the genome, which includes all DNA within an organism
  - Sequence, structure, and function of genes
  - Key technology — Next-generation sequencing (NGS)

- **Transcriptomics** Study of the transcriptome, which is the complete set of RNA transcripts
  - Gene expression and regulation
  - Key technology — RNA sequencing (RNA-seq)

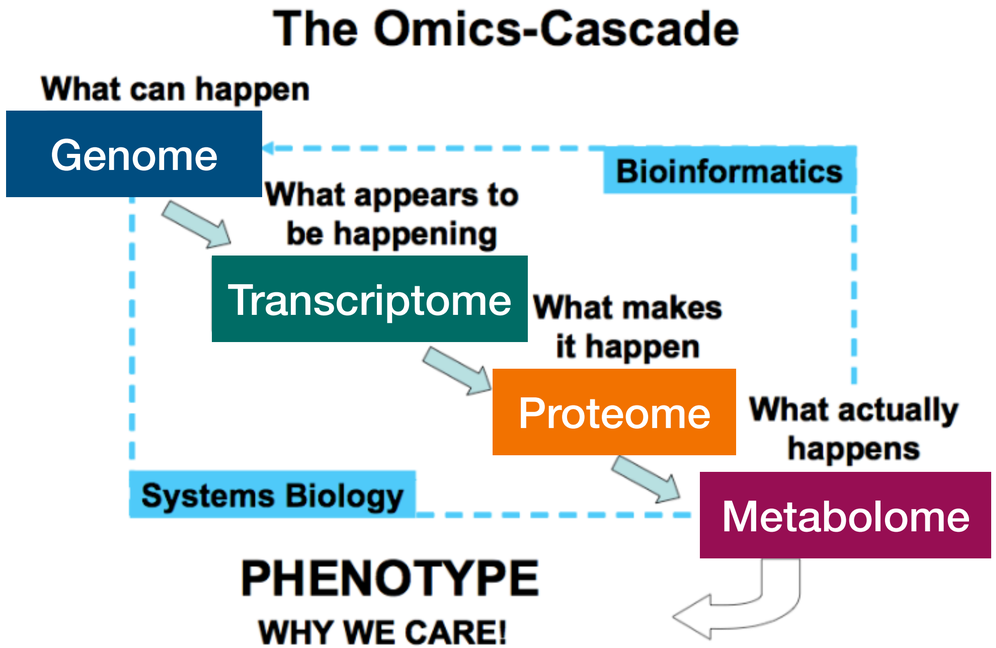
- **Proteomics** Study of the proteome, or the complete set of proteins in a cell or organism
  - Protein structure, function, interactions, and modifications
  - Key technology — Mass spectrometry (MS)



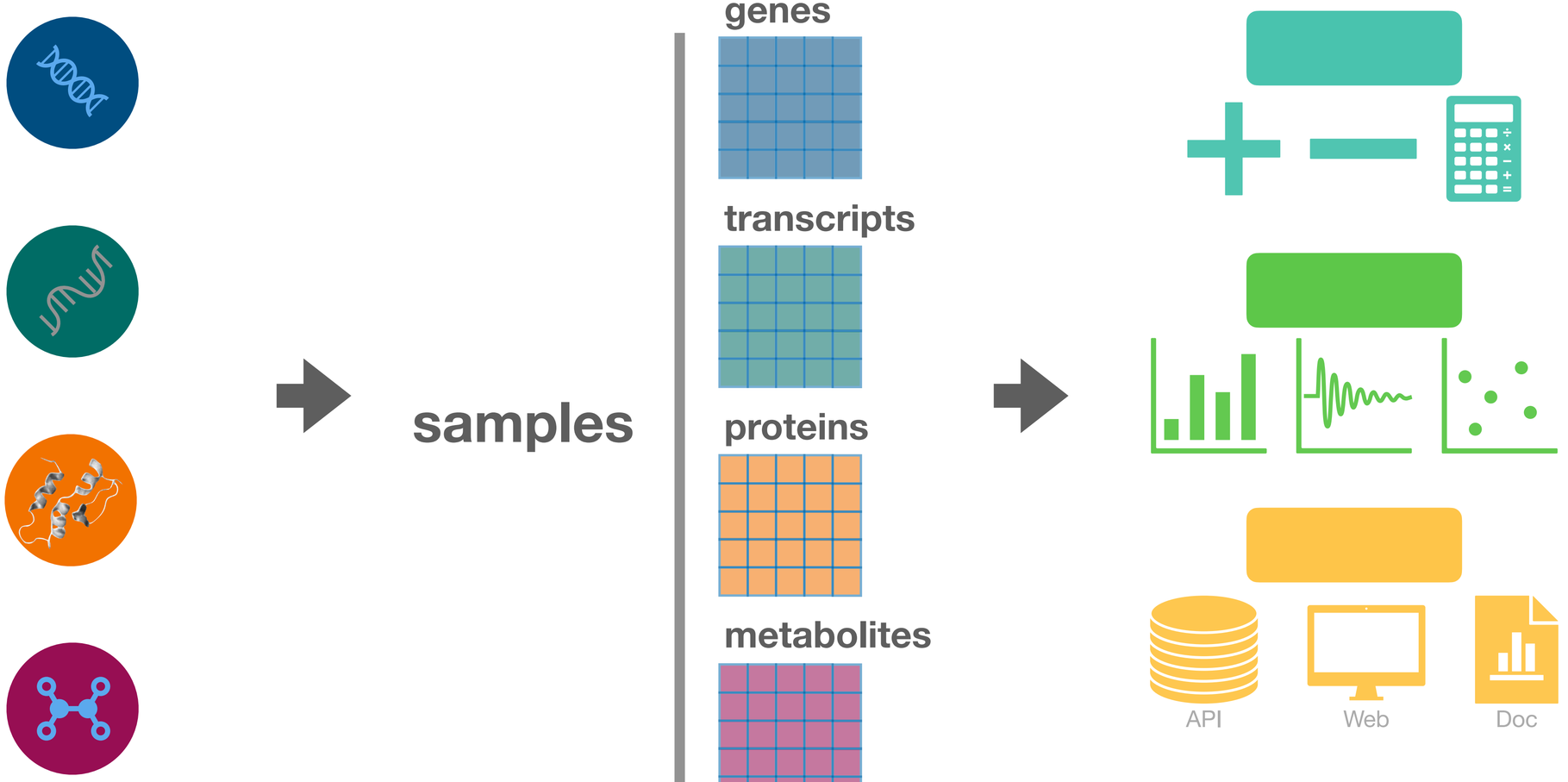
- **Metabolomics** Study of the metabolome, which includes all small-molecule metabolites in a cell or biological system
  - Cellular processes and metabolic pathways
  - Key technology — Mass spectrometry (MS)



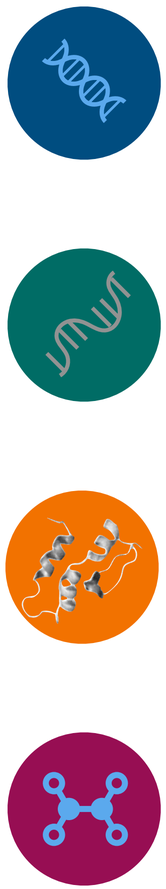
- **Metaomics** Studies the collective genetic material, proteins, metabolites, and other molecular components from entire communities of organisms in a specific environment, without needing to isolate or culture individual species.



Omics Data


samples
genes
transcripts
proteins
metabolites
API
Web
Doc

Omics Data



samples

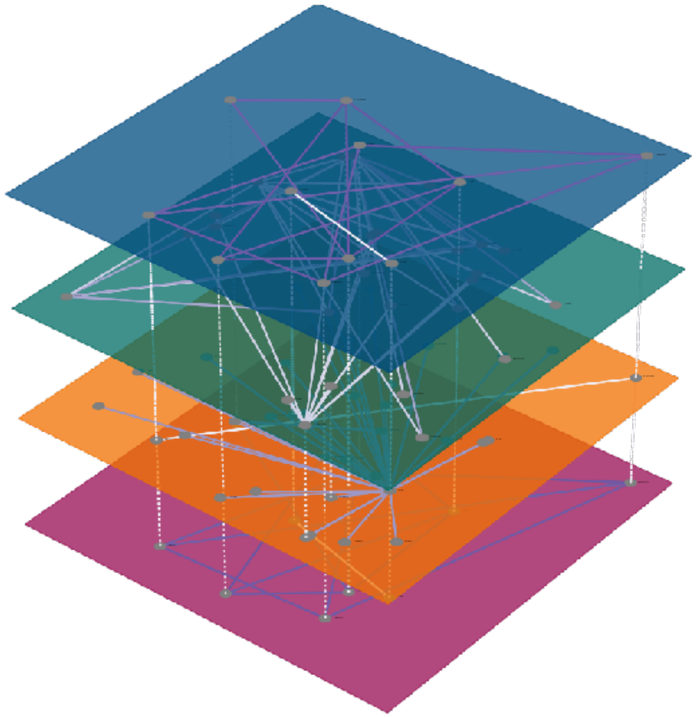
genes

transcripts

proteins

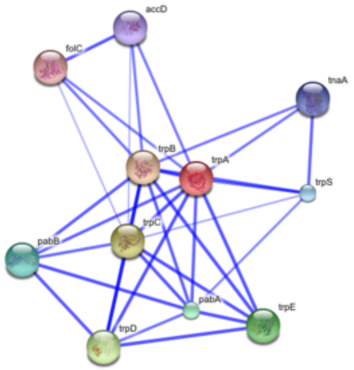
metabolites

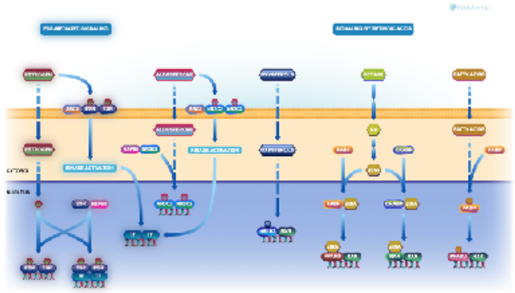
Graphs

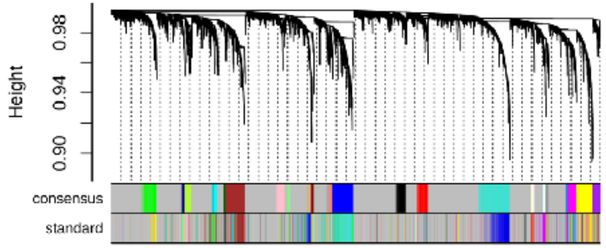


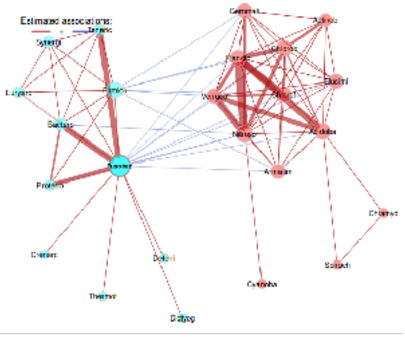
How to Build a Graph

Data to Graph





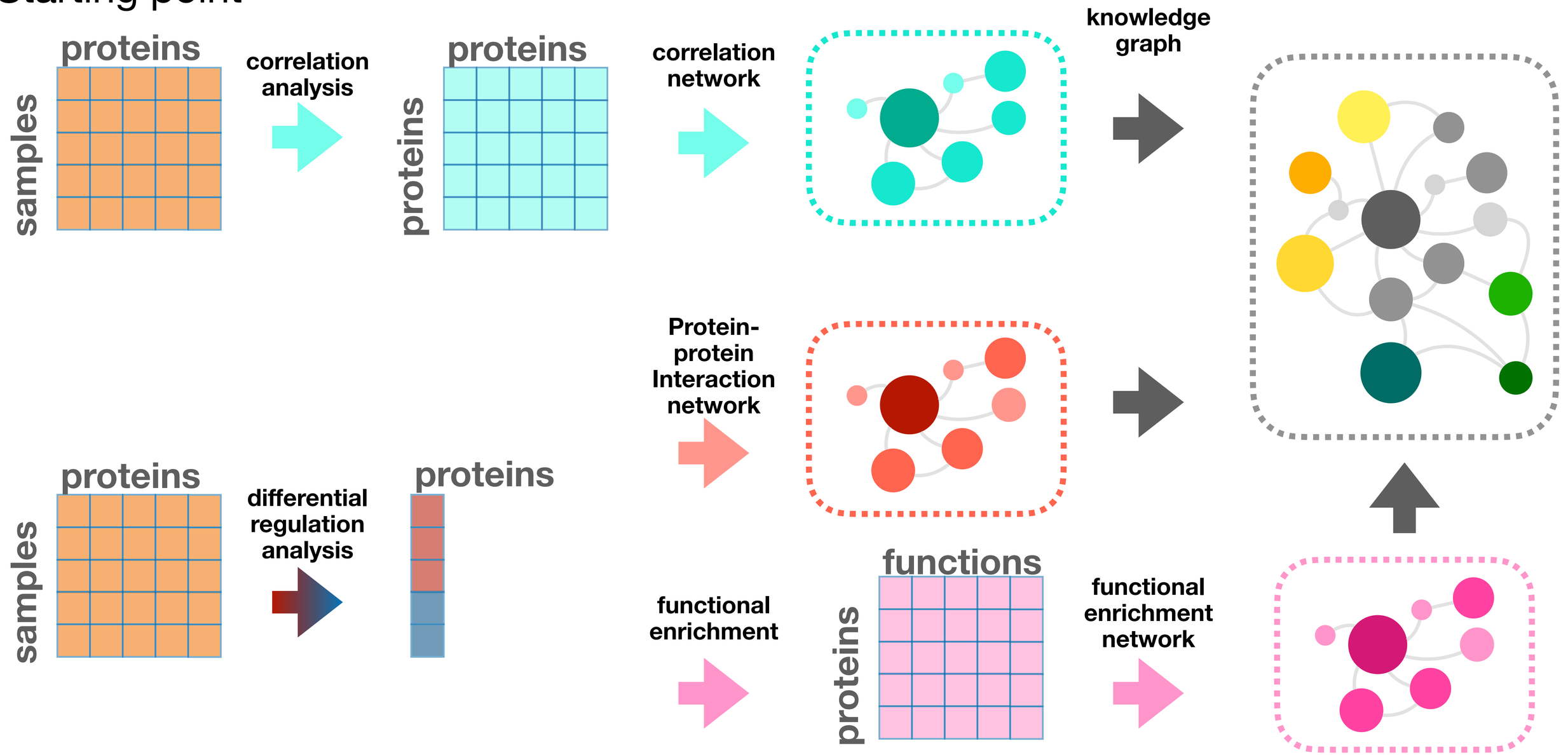




- **Data sources**
  - STRING — https://string-db.org/
  - BioGRID — https://thebiogrid.org/
  - IntAct — https://www.ebi.ac.uk/intact
  - REACTOME — https://reactome.org/
  - KEGG — https://www.genome.jp/kegg/
  - MINT — https://mint.bio.uniroma2.it/
- **Correlation-based networks —** constructed by calculating pairwise correlations between entities based on their expression profiles across multiple conditions, time points, or samples (Weighted gene co-expression network analysis (WGCNA), co-abundance networks)
- **Knowledge-base approaches** — also called knowledge graphs and built by integrating heterogeneous data from multiple sources —> **Knowledge Graphs**

How to Build a Graph

Starting point

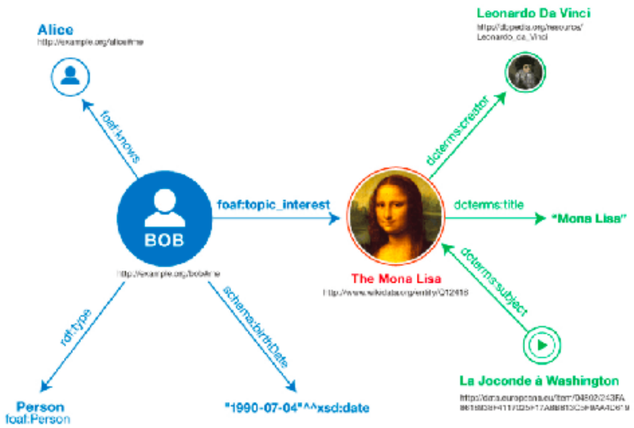


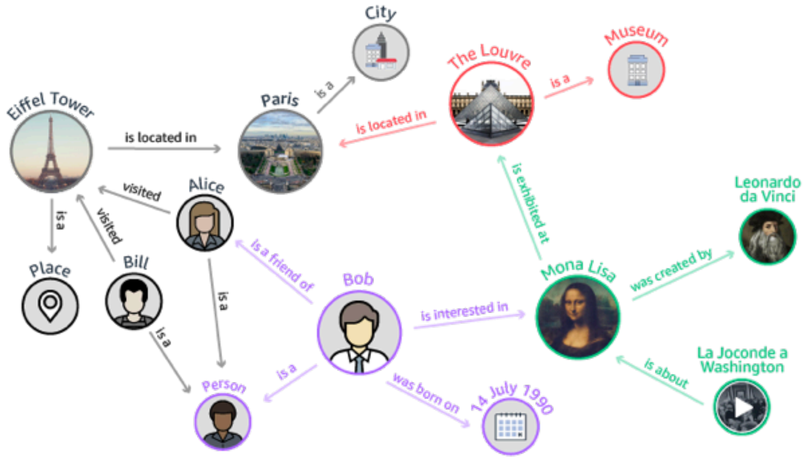
Knowledge Graphs

What is a Knowledge Graph (KG)

Relationships firsts everything else second

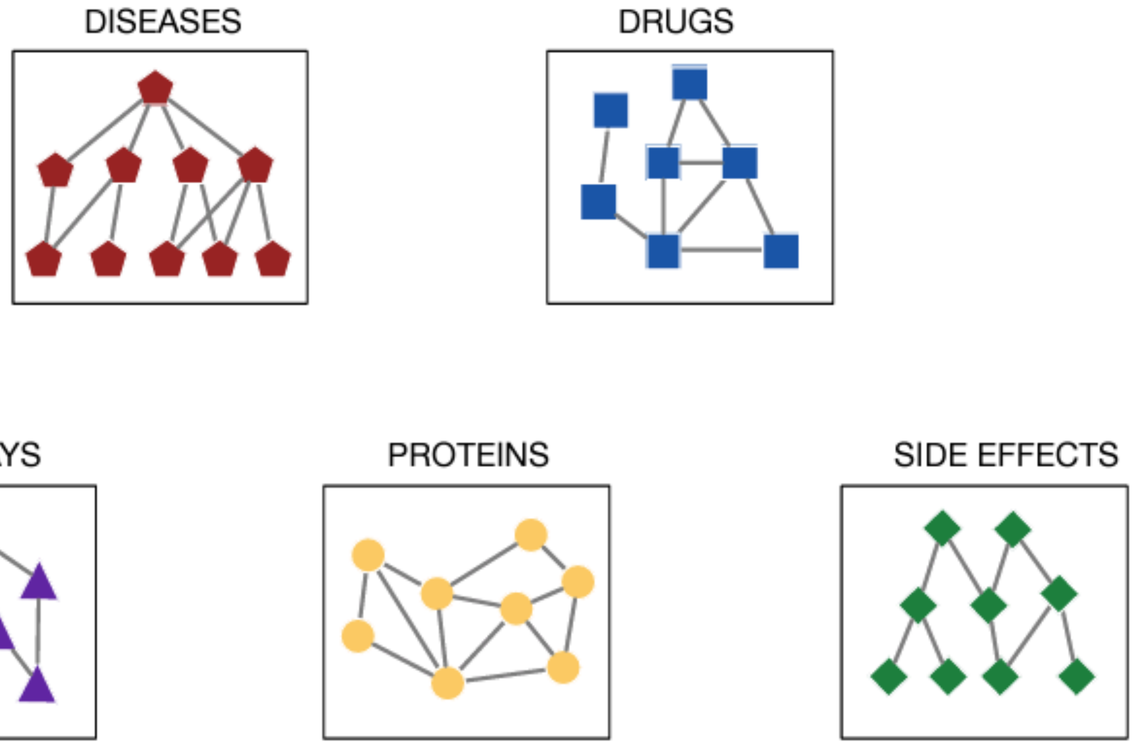
- A way to organise **knowledge**/information by defining **associations** or **relationships**
- These relationships facilitate **integration**, **management** and **enrichment** of data
- The **objective** when setting up a KG:
  - Standardisation / FAIRification
  - Reusability
  - Interpretability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





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

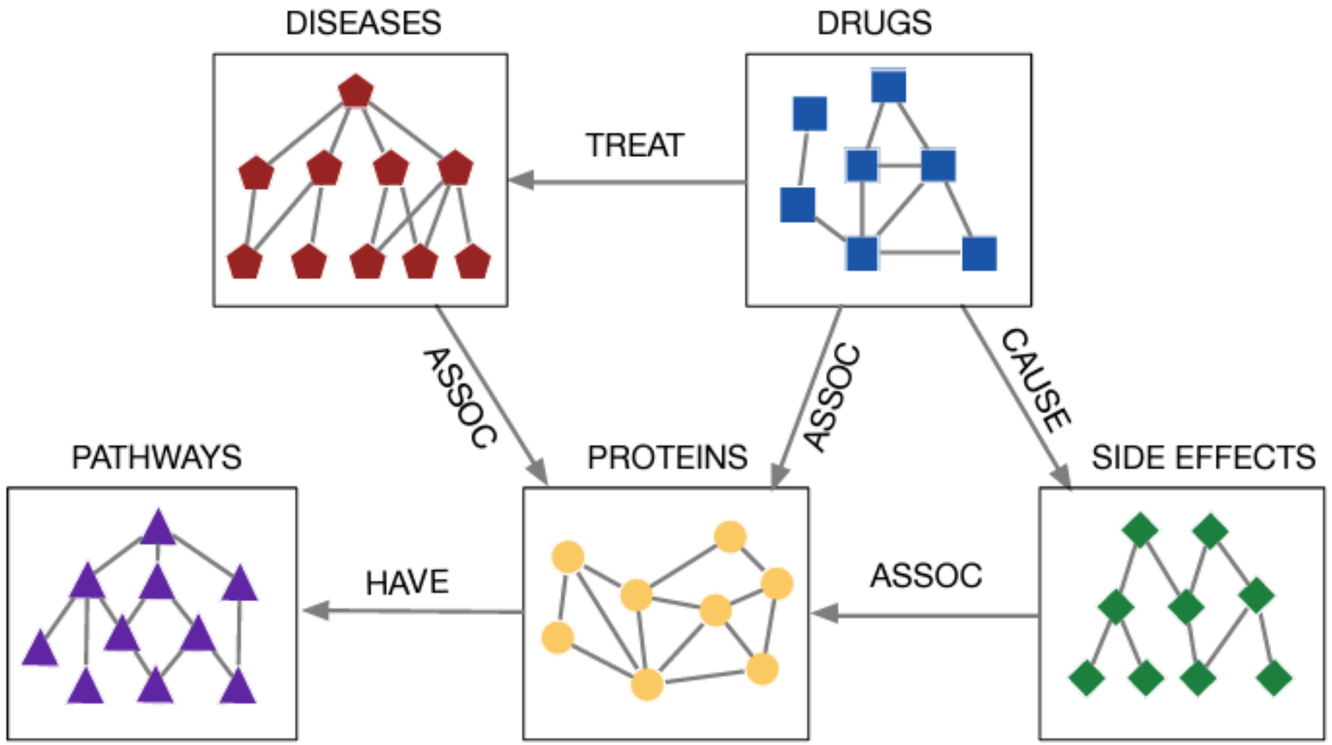
Knowledge Graph

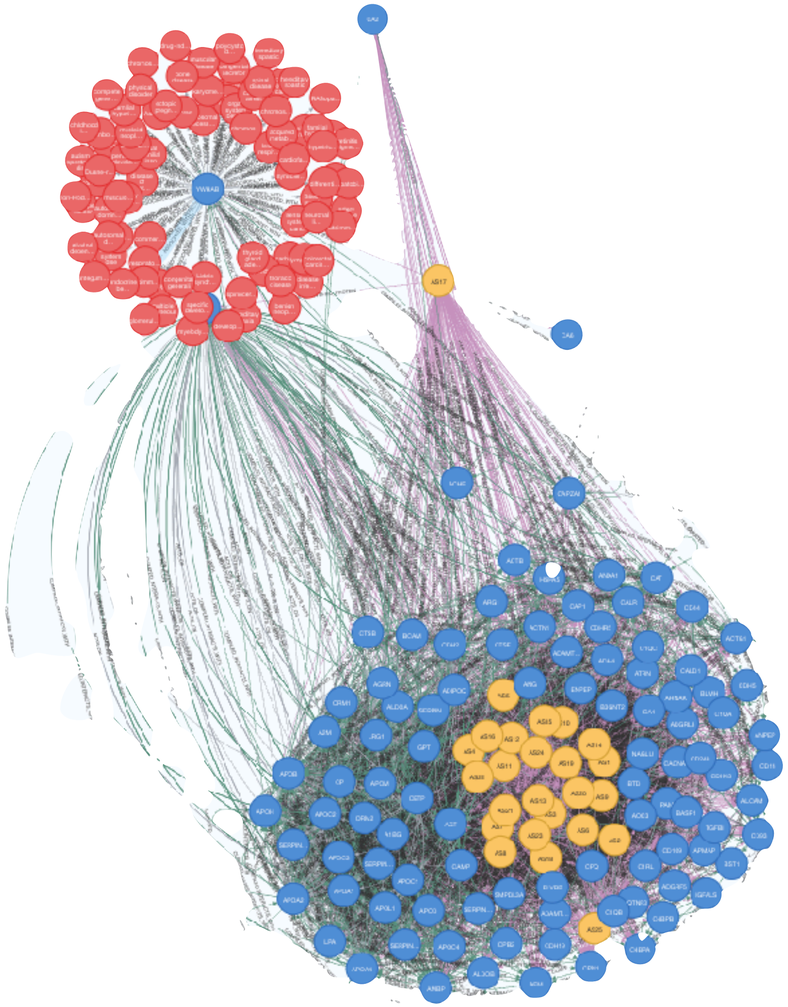


https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

Knowledge Graph

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

Building a Knowledge Graph

1. Define the **questions** you want to answer

2. Define **what data** can be used to answer these questions and **how it is linked** — Data model

3. Find **where to get these data**

4. Get the data, **standardise** it and **format** it

5. Generate the **graph**

6. **Query the graph** to answer the questions

Building a Knowledge Graph

Building a Knowledge Graph

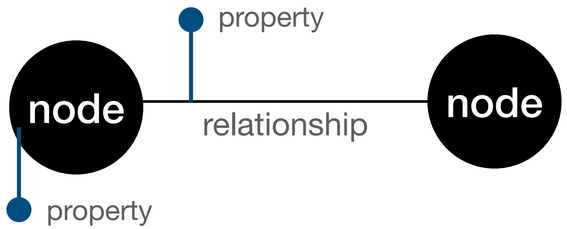
Exercise

Create a data model that allows us to answer the question:

What drugs related to our disease of interest target some of the proteins identified in our experiment or relevant protein complexes and pathways?

Graph Databases

- Knowledge Graphs became popular in **2012** thanks to **Google** (proprietary graphs)
- What made them accessible was the development of **open-source Graph Databases**
- Graph Databases are **NoSQL** databases that use **graph structures** to represent and store data
- Data is represented as **Nodes**, **Relationships** and **properties**

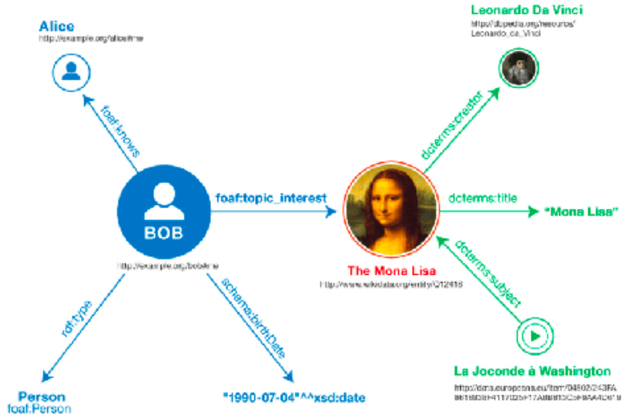


- They use their own **querying languages**: **Cypher**, **SPARQL**, GraphQL, Gremlin, etc.

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

Graph Data Models

Semantics vs Programmers

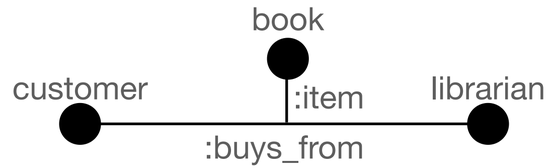


Semantic Graphs or Triple-stores

- The network represents **meaning** through semantic relationships, which simplifies **reasoning**
- Follows the **Resource Description Framework (RDF)** data model

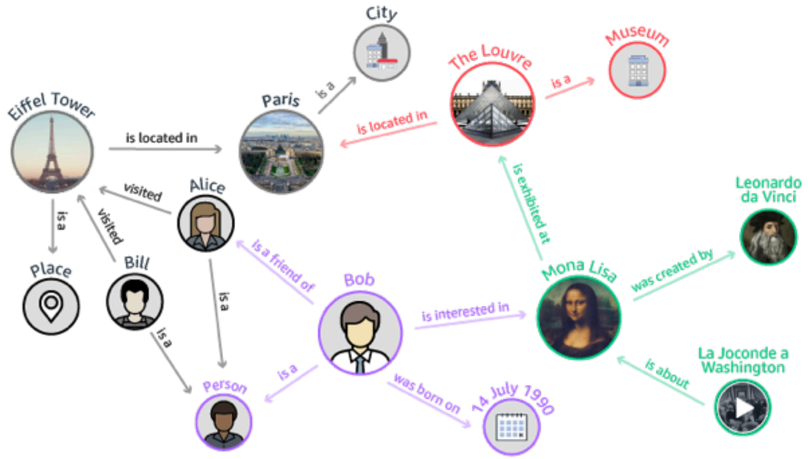
- **Properties** need to be represented as **nodes**
- Allows **n-array relationships** —>

- Uses **Uniform Resource Identifiers (URIs)** to identify concepts
- Used for **Ontologies** —> cancer — is_a — disease
- The **query language** used is **SPARQL**
- Vendors (I know):
  - STARDOG (https://www.stardog.com/)
  - PoolParty (https://www.poolparty.biz/)



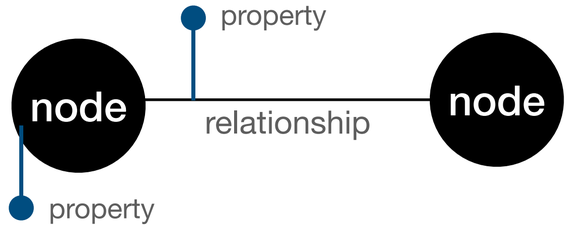
Graph Data Models

Semantics vs Programmers



Labelled Property Graphs (LPG)

- The network represents **relationships** between pairs of **nodes**, and they have *labels* describing id, type, class and other **properties**
- The number **nodes** is **reduced** by using properties instead
- **Does not allow n-array relationships**, instead properties —>

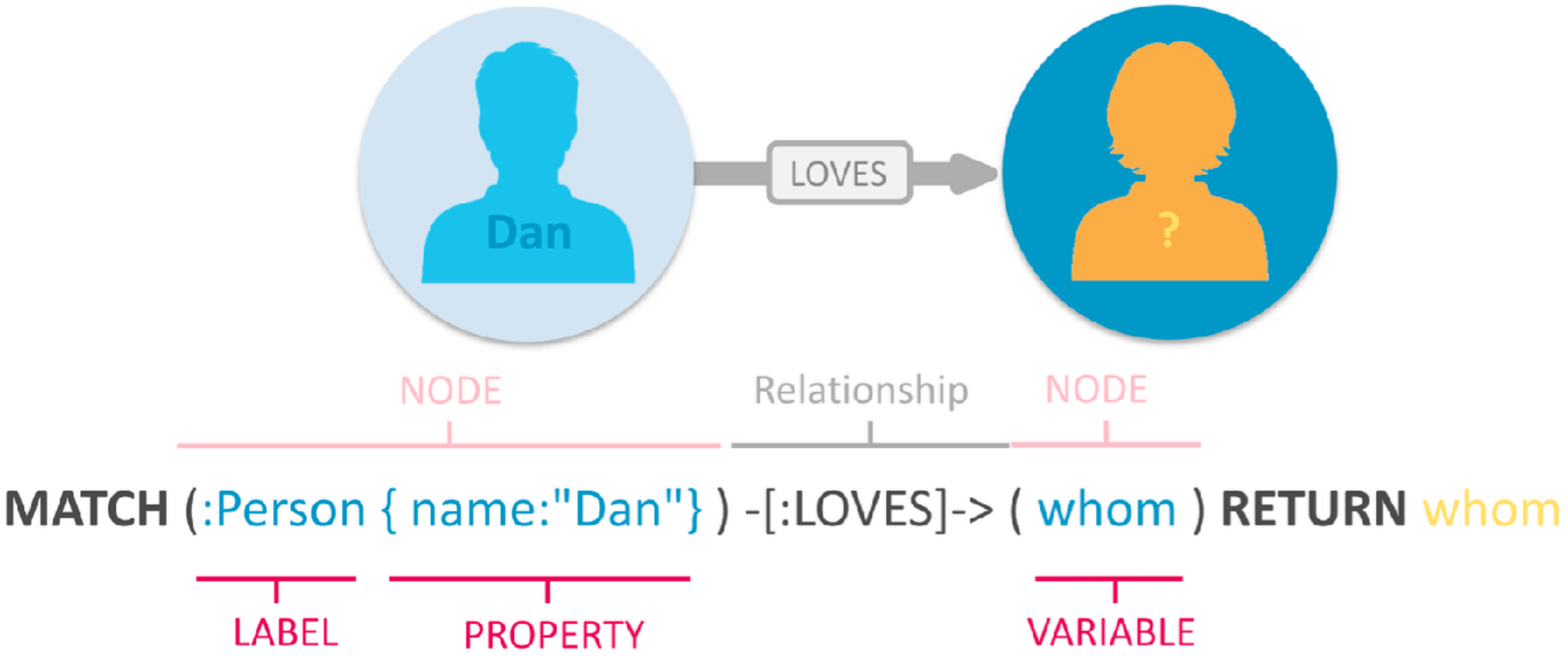


- The **query language** used is **Cypher** (not a standard but widely adopted)
- Vendors (I know):
  - Neo4j (https://neo4j.com/)
  - TypeDB (https://vaticle.com/)
  - Memgraph (https://memgraph.com/)
  - FalkorDB (https://www.falkordb.com/)
  - NebulaGraph (https://www.nebula-graph.io/)

https://www.oxfordsemantic.tech/fundamentals/what-is-a-labeled-property-graph

Cypher Query Language

Cypher is a graph query language that provides a visual way of matching patterns and relationships
(property graphs)



https://neo4j.com/developer/cypher/

Querying a Knowledge Graph

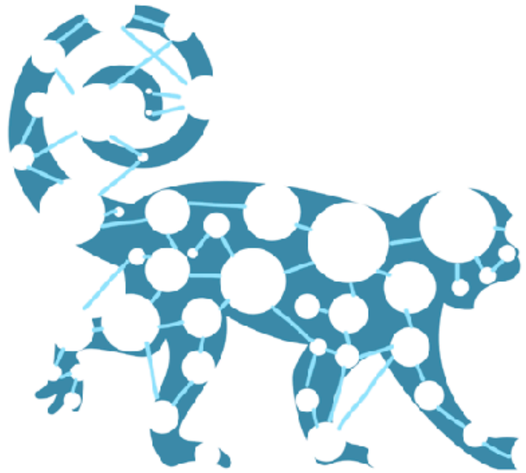
Common Exercise

Create a cypher query for the data model we created

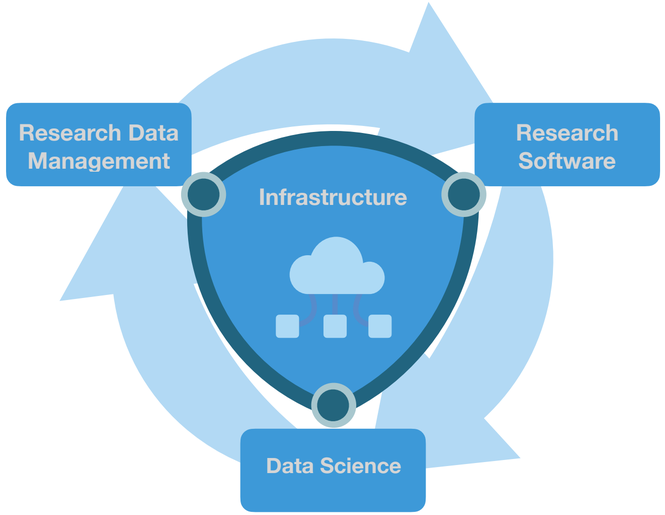
Questions?

Thank you

Multi-omics Network Analytics Research Group



Informatics Platform



The Novo Nordisk Foundation Center for Biosustainability

novo nordisk fonden

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

albsad@dtu.dk

@albsantosdel

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

https://multiomics-analytics-group.github.io/

The Novo Nordisk Foundation Center for Biosustainability