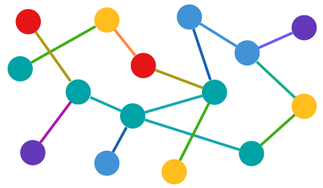
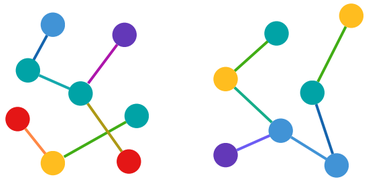
Implementation of bioinformatics approaches to predict gene-phenotype interactions using heterogeneous data and machine learning algorithms

CeDRE
Antoine TOFFANO
Supervisor : Christophe Héligon

Outline



- Dataset and objective



- Train / Test split

- Algorithms

- Evaluation

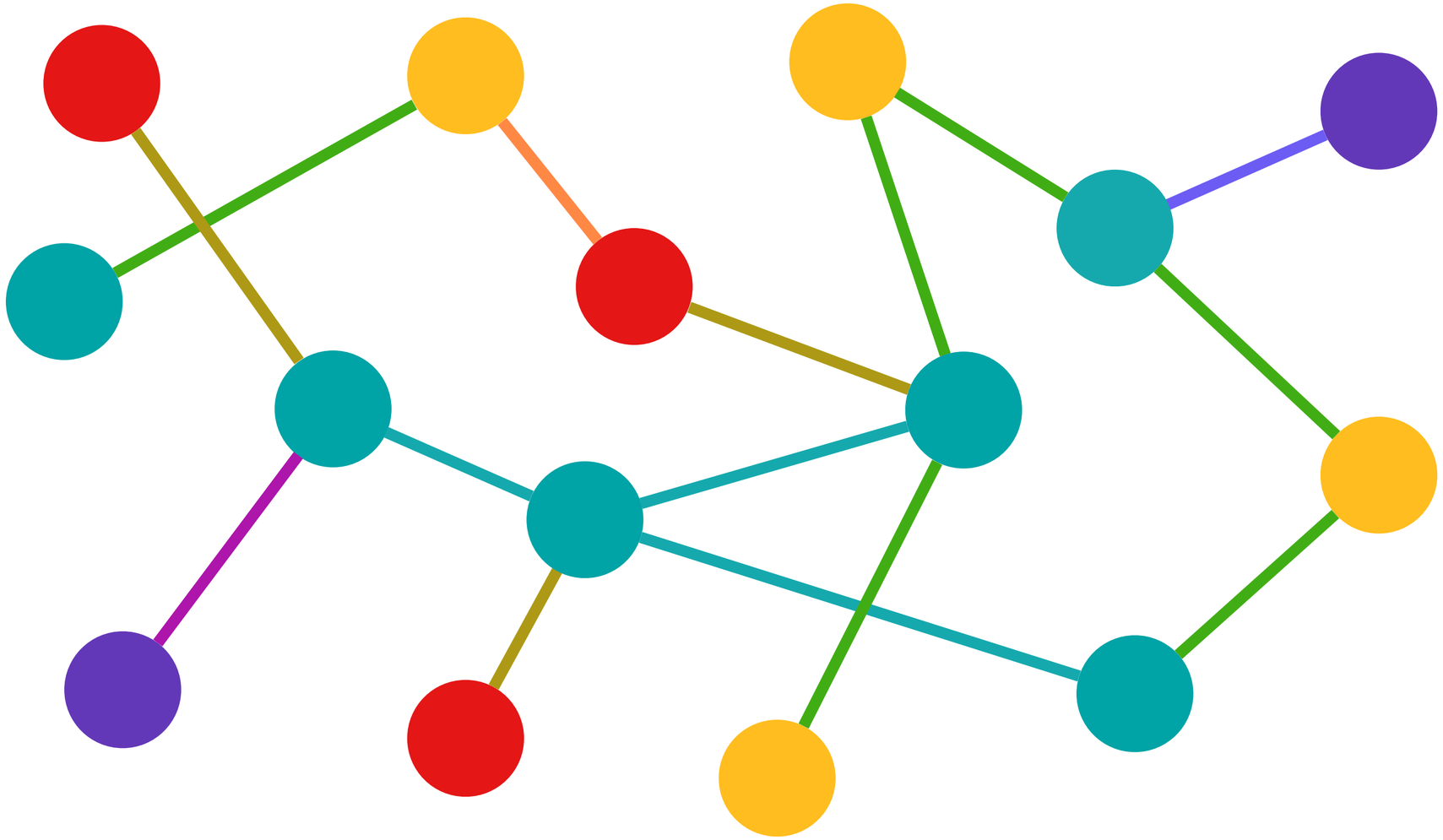
Dataset and objective

WormBase

- *Protein - Protein interactions*
- *Gene - Phenotype associations*
- *Gene - Disease associations*
- *Gene symbols and alternative symbols*
- *Gene - disease by orthology associations*
- *Gene - GO terms associations*
- *Gene RNA expression (whole individual bulk FPKM)*

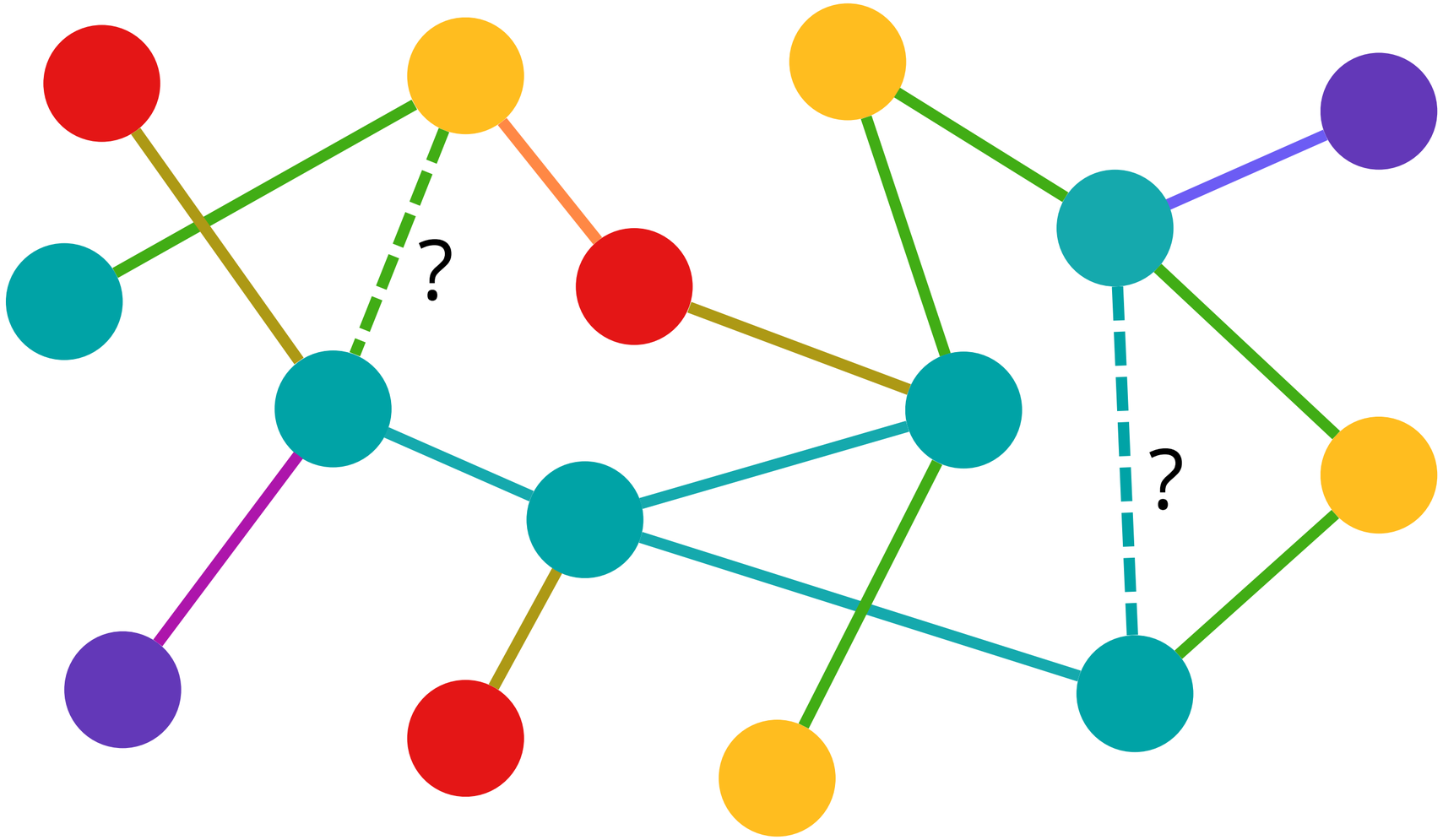
igdrX

Dataset and objective



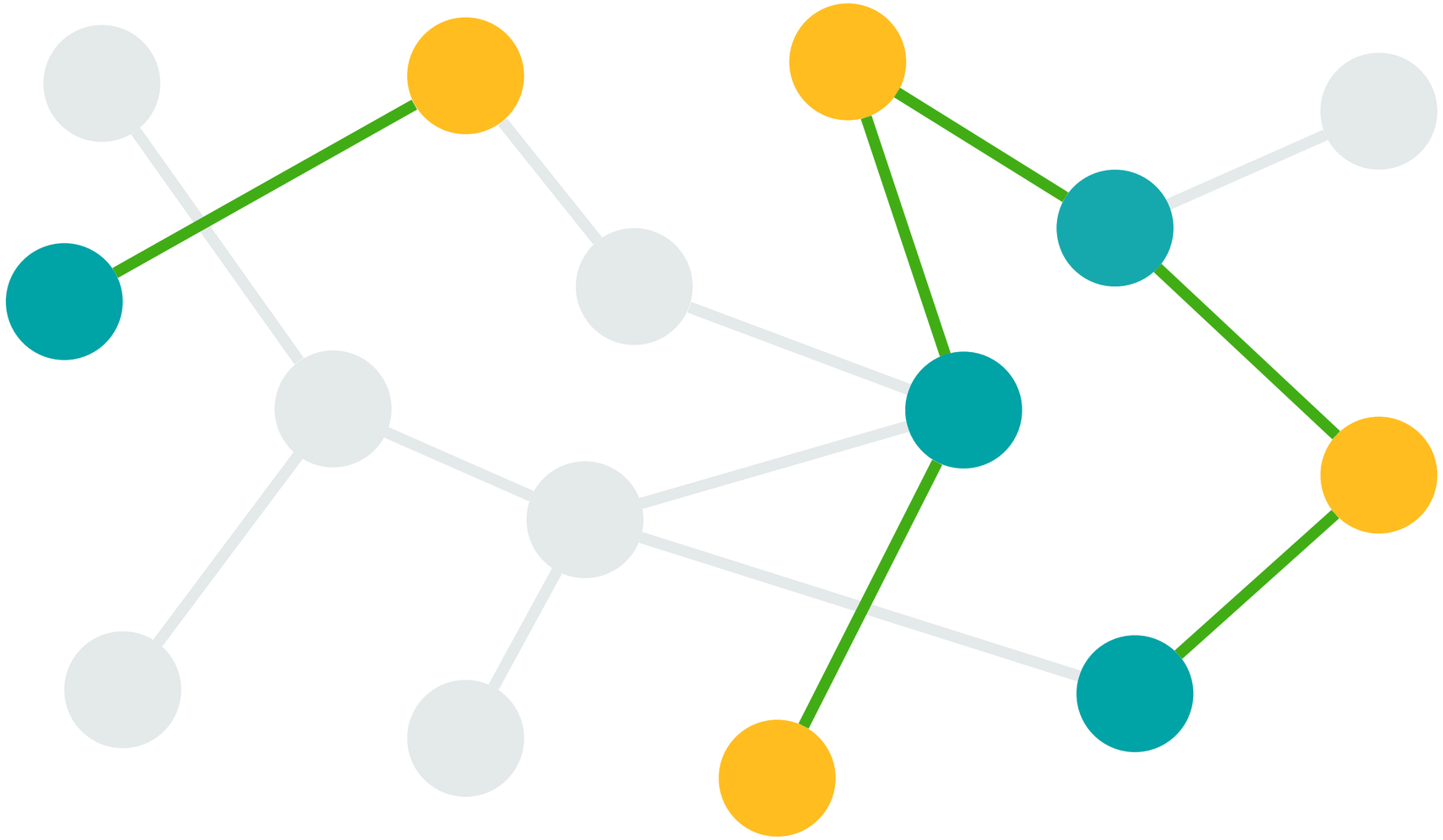
Gene
Phenotype
Disease
Go

Dataset and objective



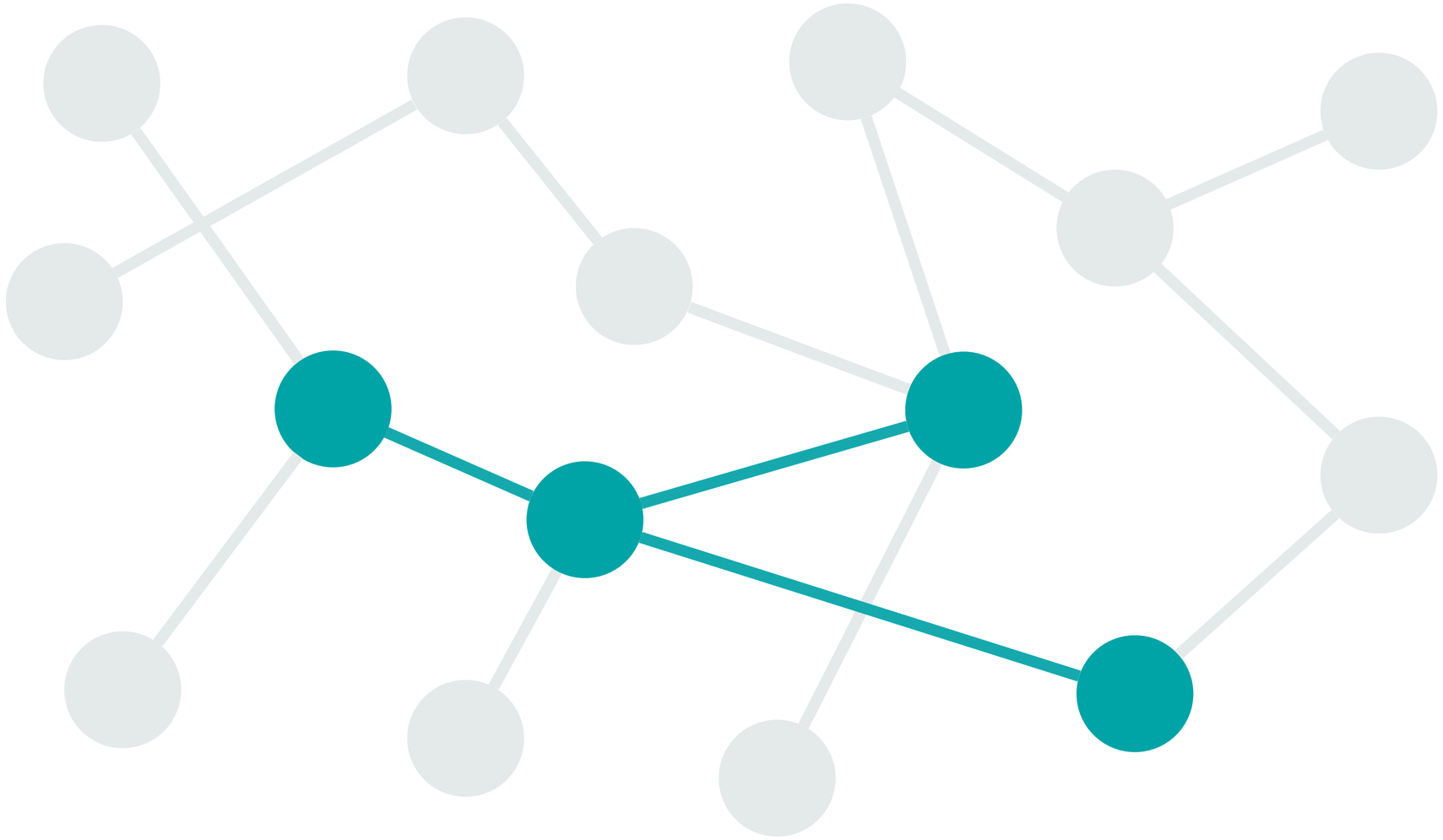
Gene
Phenotype
Disease
Go

Dataset and objective



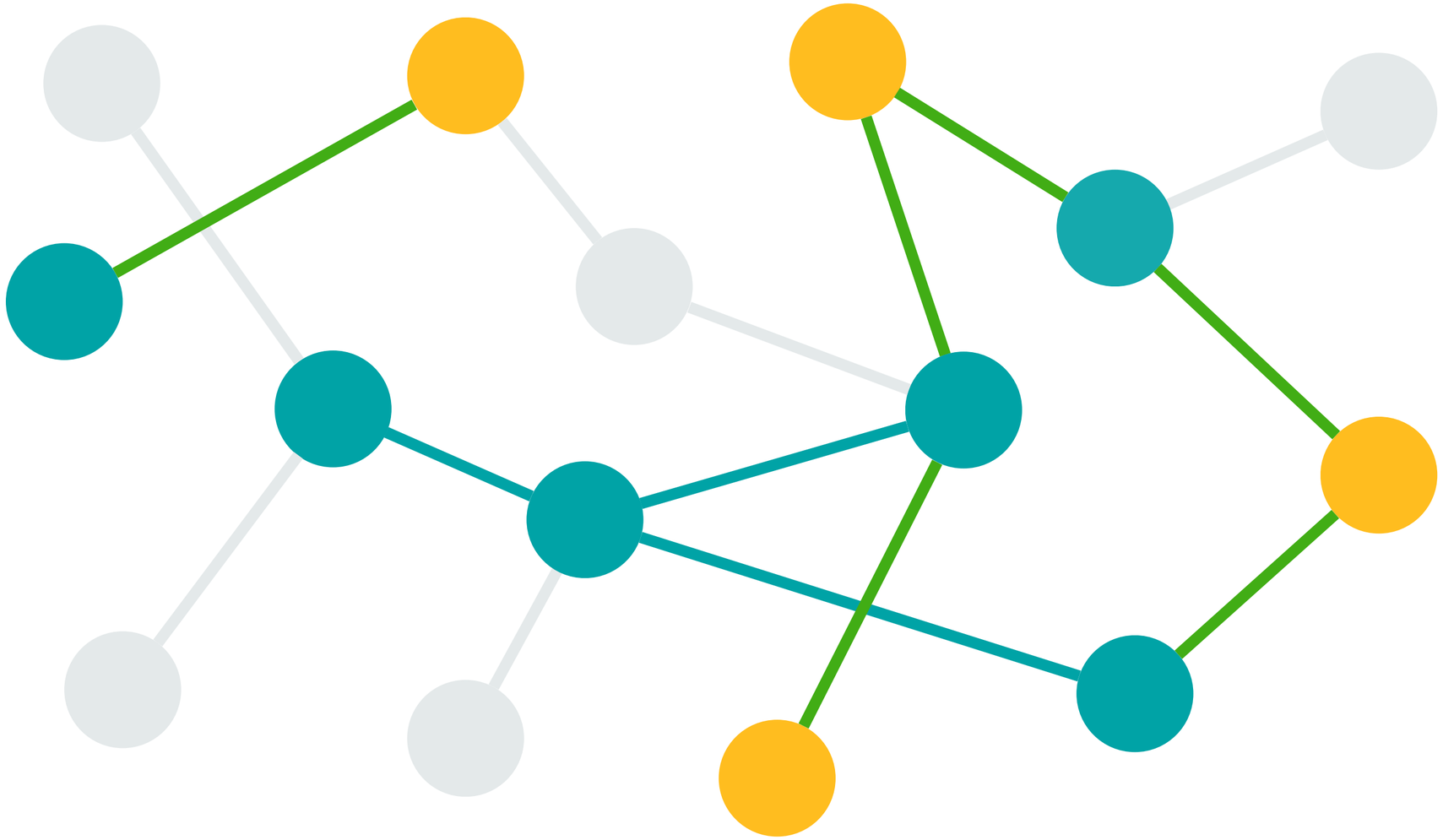
Protein
Gene
Phenotype
Disease
Go

Dataset and objective



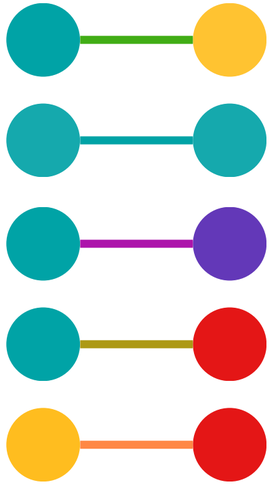
- Protein
- Gene
- Phenotype
- Disease
- Go

Dataset and objective



Protein
Gene
Phenotype
Disease
Go

Dataset and objective



Gene - Phenotype

Gene - Gene

Gene - GO

Gene - Disease

Phenotype - Disease

What should we add in priority ?

Date of paper publication

Author / Paper

Type of evidence (eco)

Chemical Entities (ChEBI)

Protein data (UniProtKB)

Phylogeny (Panther)

Enzyme nomenclature (Expasy)

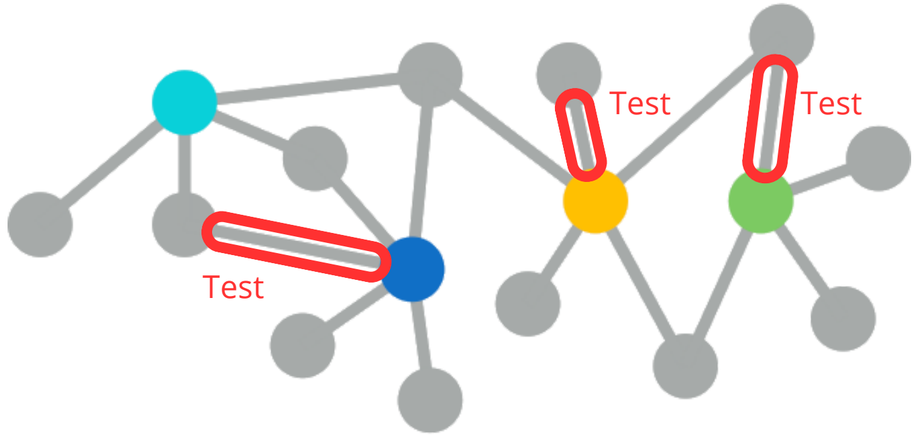
Chemical Reactions (RHEA)

Protein Family (InterPro)

AlphFold DB

Not implemented yet

Train / Test split



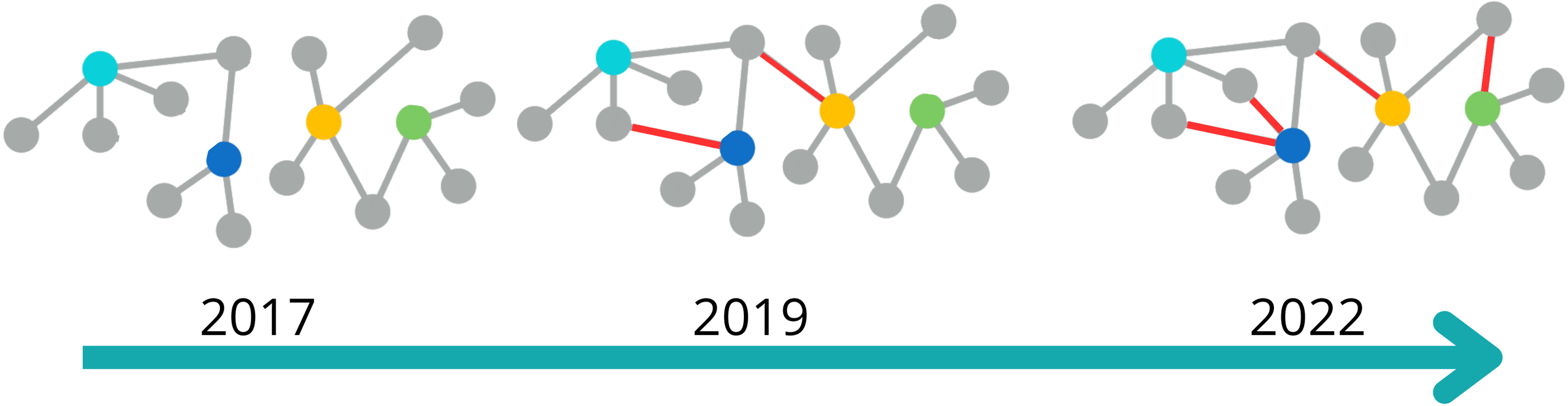
Pros

- Easy to avoid bias in test set
- Easy to implement

Cons

- Not representative of the evolution of a database

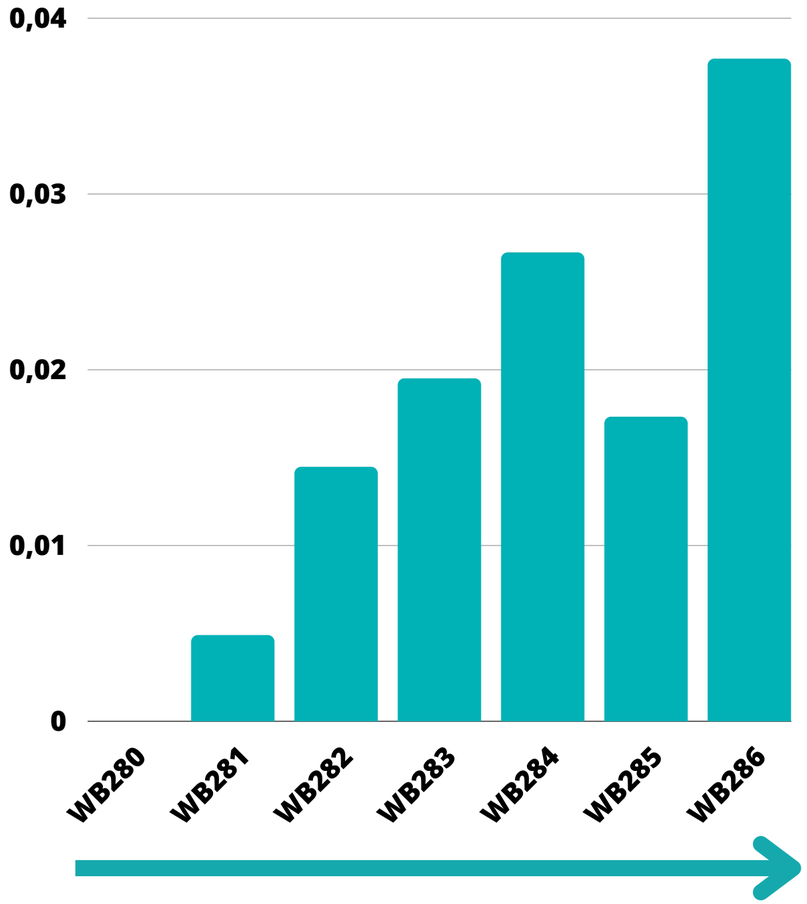
Train / Test split



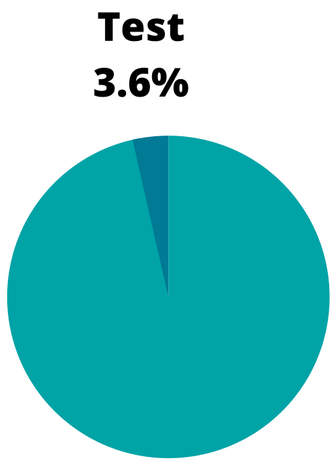
Train on a old version, and test on edges that have been added afterwards.

Train / Test split

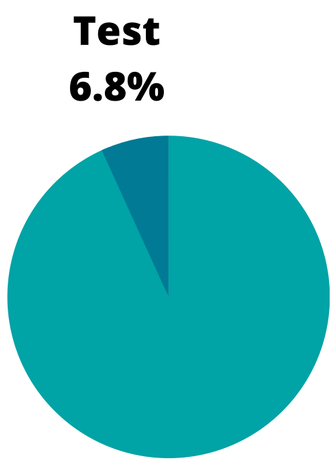
Wormbase size aumentation since WB280



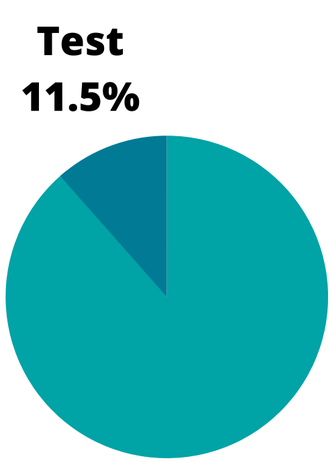
Wormbase



WN18RR



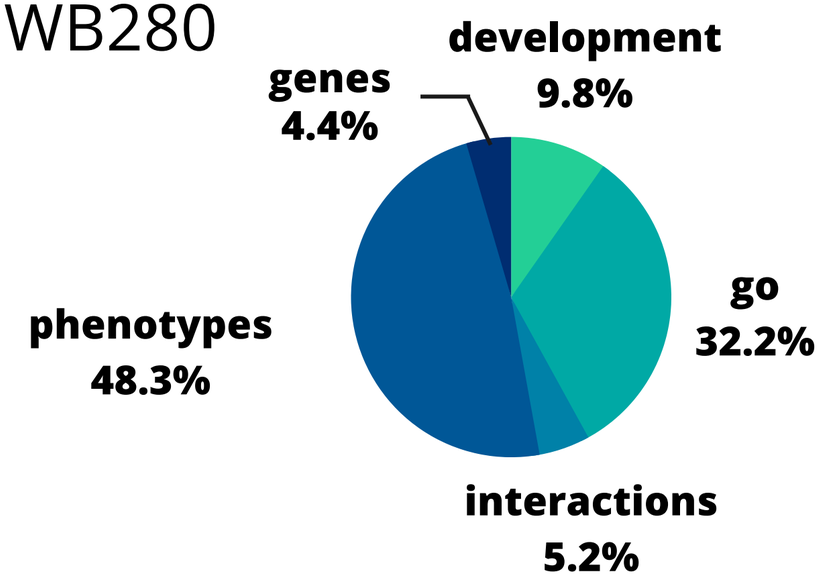
FB15K-237

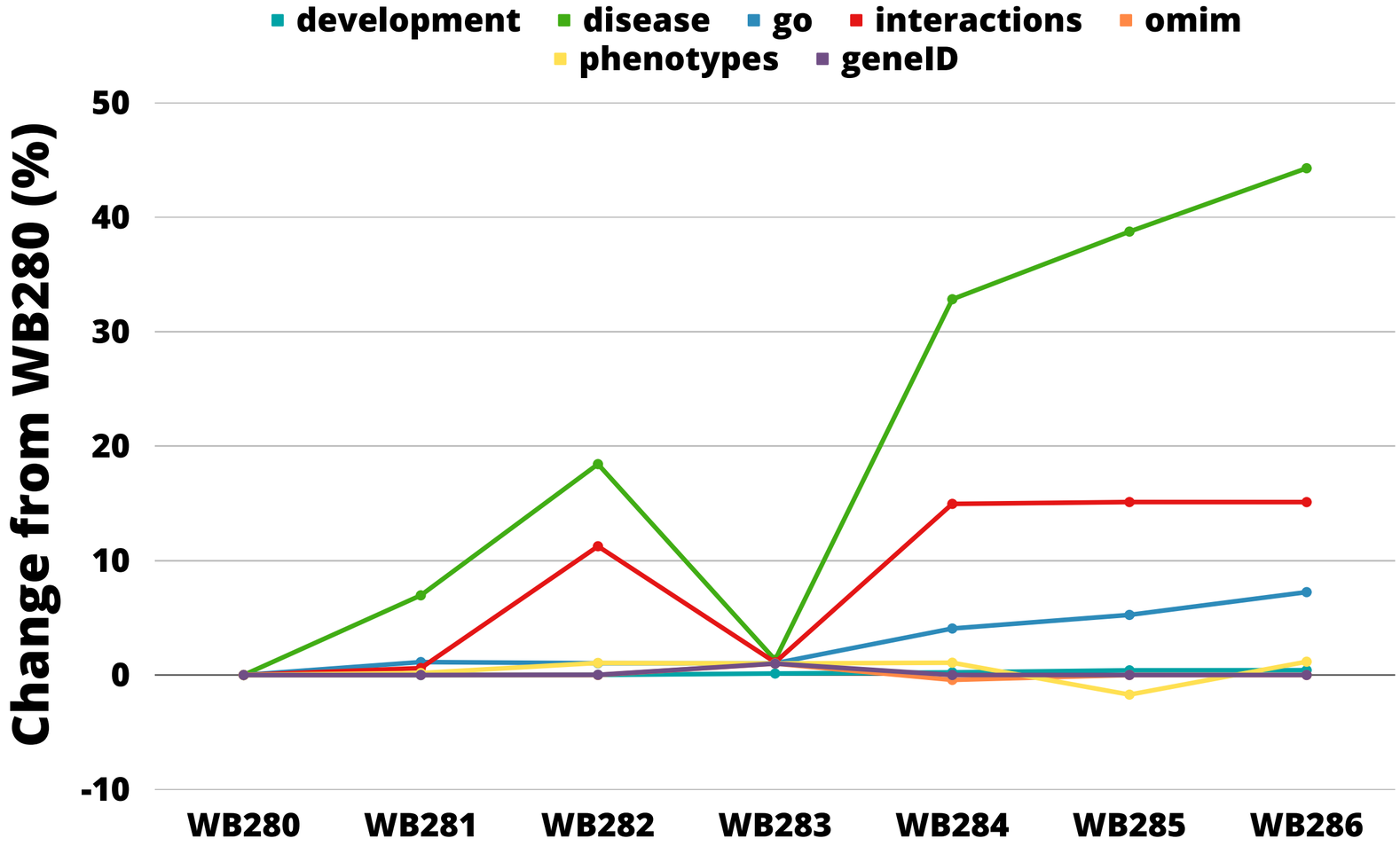


Do we need to go back further ?

Train / Test split

WB280





How much of a problem is this imbalance ?

Algorithms

Shared goal : find good embeddings of graph elements

Random Walk

- Runs on CPU
- High time complexity
- Multiple examples of being used on biological data

1 implemented
1 underway

Machine learning - Baseline

- Better time complexity
- Time-tested
- Usually low amount of hyperparameters

1 pipeline implemented:
(10 methods)

Deep learning

- Time complexity (?)
- Higher performances (?)
- More recent, often SoTA of a particular dataset

2 methods implemented
2 underway

igdrX

Scoring methods

Goal : compare methods to each other

Hits@1

-

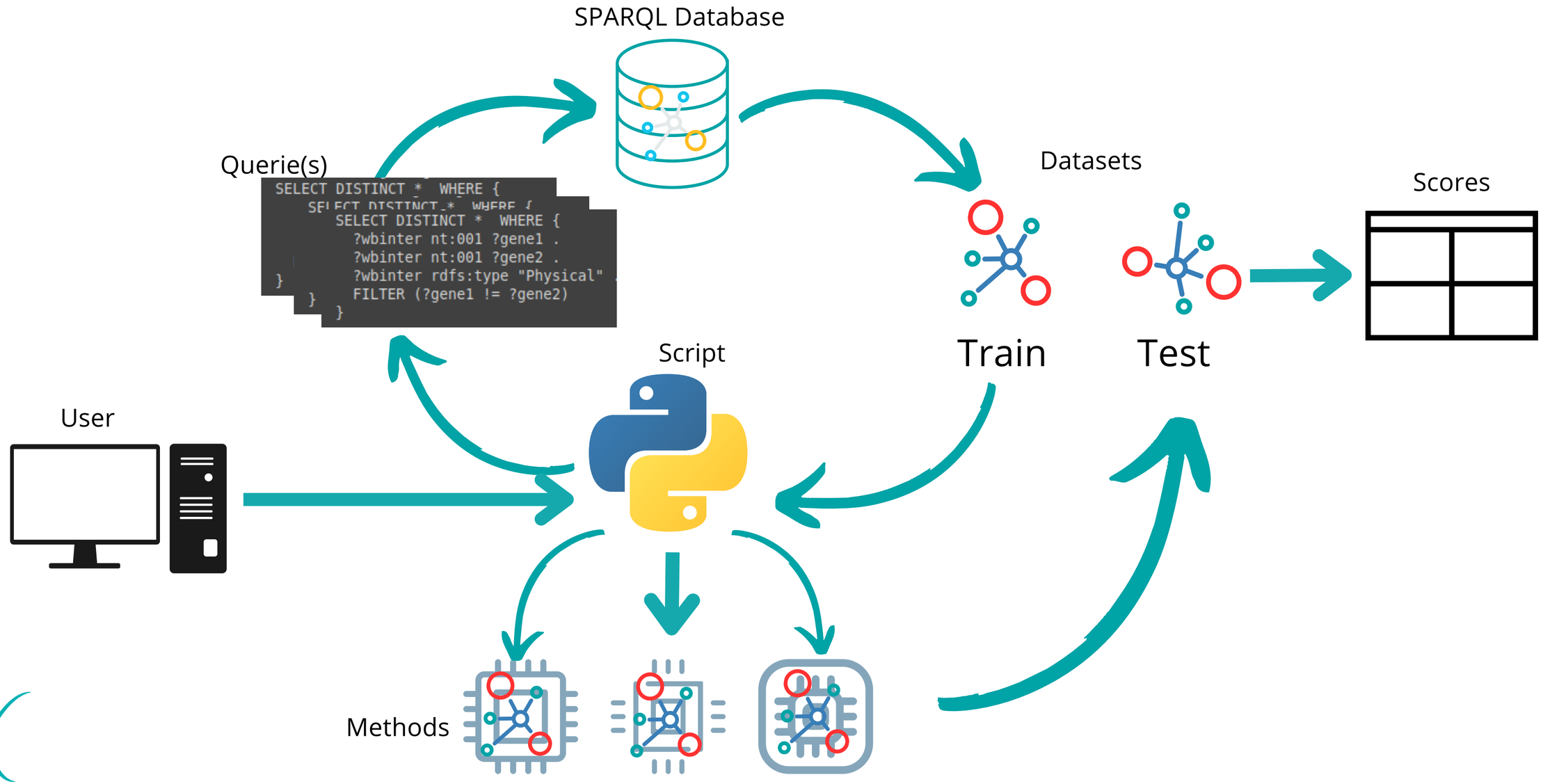
Hits@10

-

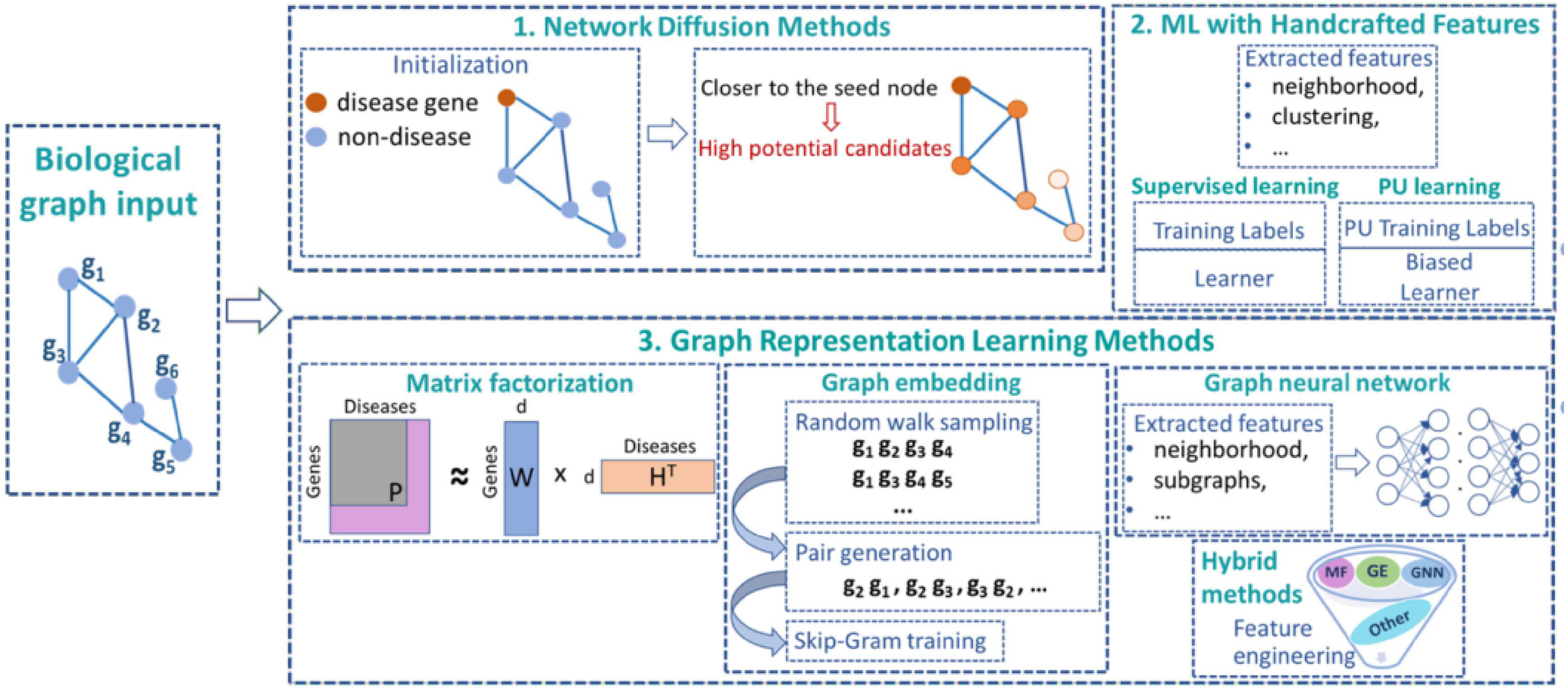
MRR

-

Process overview

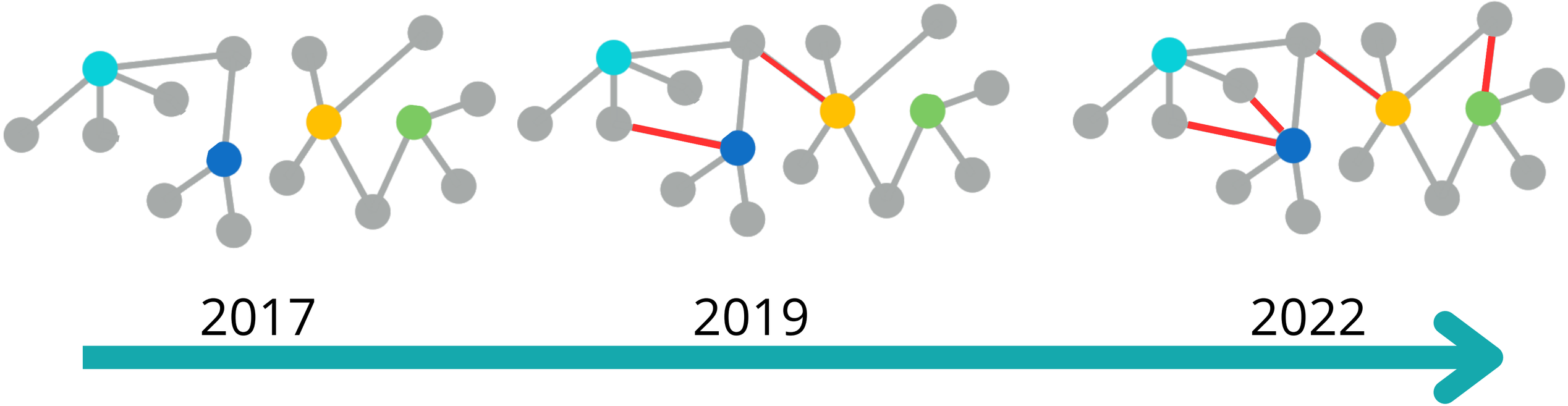


Benchmarking methods



- Recent advances in network-based methods for disease gene prediction, Ata et al., 2021, 10.1093/bib/bbaa303

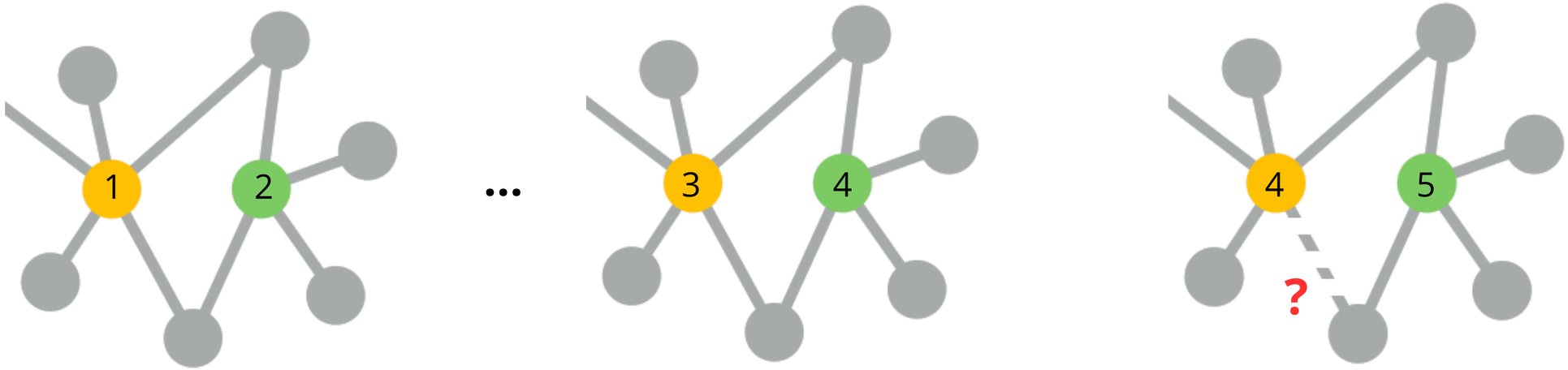
Train / Test split



Train on a old version, and test on edges that have been added afterwards.

Gene-phenotype prediction

Internship goal : Link prediction between gene and phenotype in *C.Elegans*



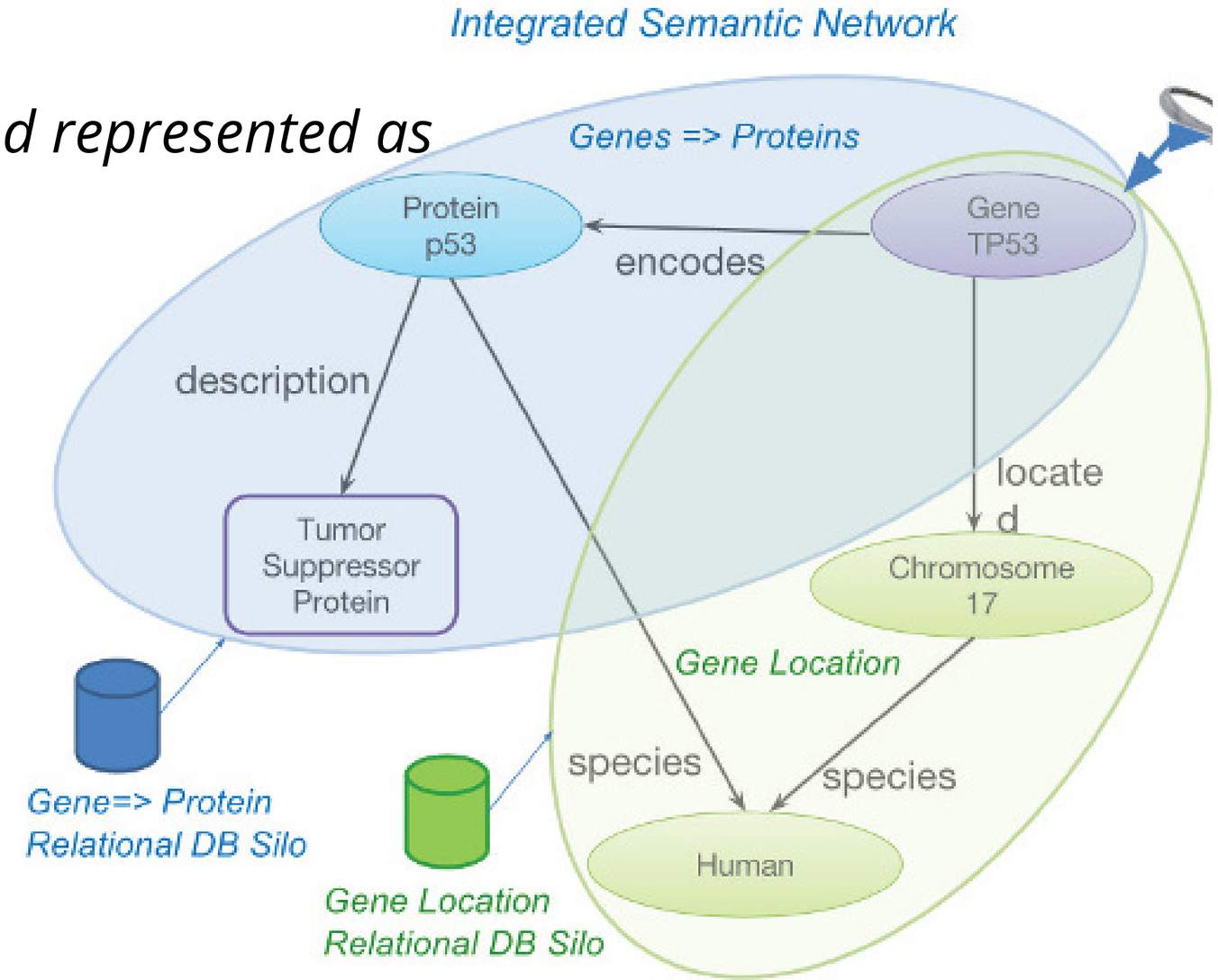
Learn 'patterns' in the graph

RDF and semantic web

*Data is structured in a 'rdf' format and represented as a **graph***

- *Object -> Predicate -> Subject*
- *Node1 -> Edge type -> Node2*

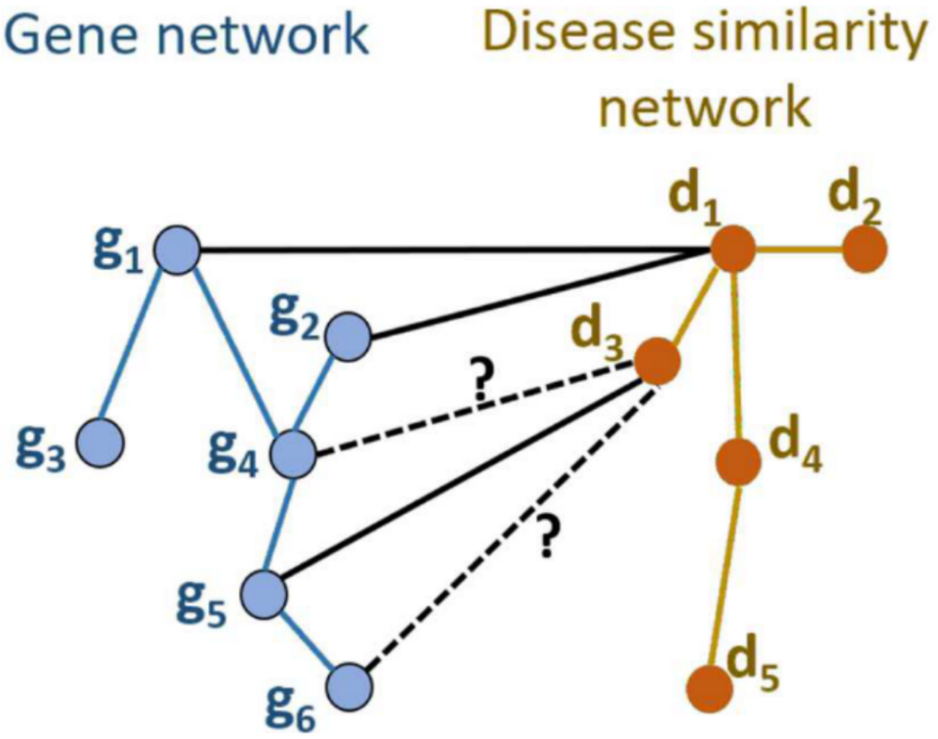
```
12933 wbdata:go-786 nt:001 wbgene:00000054 .
12934 wbdata:go-786 dcterms:source wbref:00046480 .
12935 wbdata:go-786 dcterms:source pmid:21873635 .
12936 wbdata:go-786 ro:0002331 go:0034220 .
12937 wbdata:go-786 sio:000067 go:0008150 .
12938 wbdata:go-786 sio:000772 eco:0000318 .
12939 wbdata:go-786 sio:001403 FLYBASE:FBgn0000036 .
12940 wbdata:go-786 sio:001403 FLYBASE:FBgn0000039 .
12941 wbdata:go-786 sio:001403 FLYBASE:FBgn0039840 .
12942 wbdata:go-786 sio:001403 FLYBASE:FBgn0264908 .
12943 wbdata:go-786 sio:001403 MGI:MGI:95621 .
```



- https://www.americanlaboratory.com/180191-Integration-The-Missing-Link-in-Data-Rich-Research-Interview-With-Robert-Stanley-of-IO-Informatics/

Gene-phenotype prediction

Internship goal : Link prediction between gene and phenotype in *C.Elegans*



Edges: 12 086 627

Nodes : 2 977 525

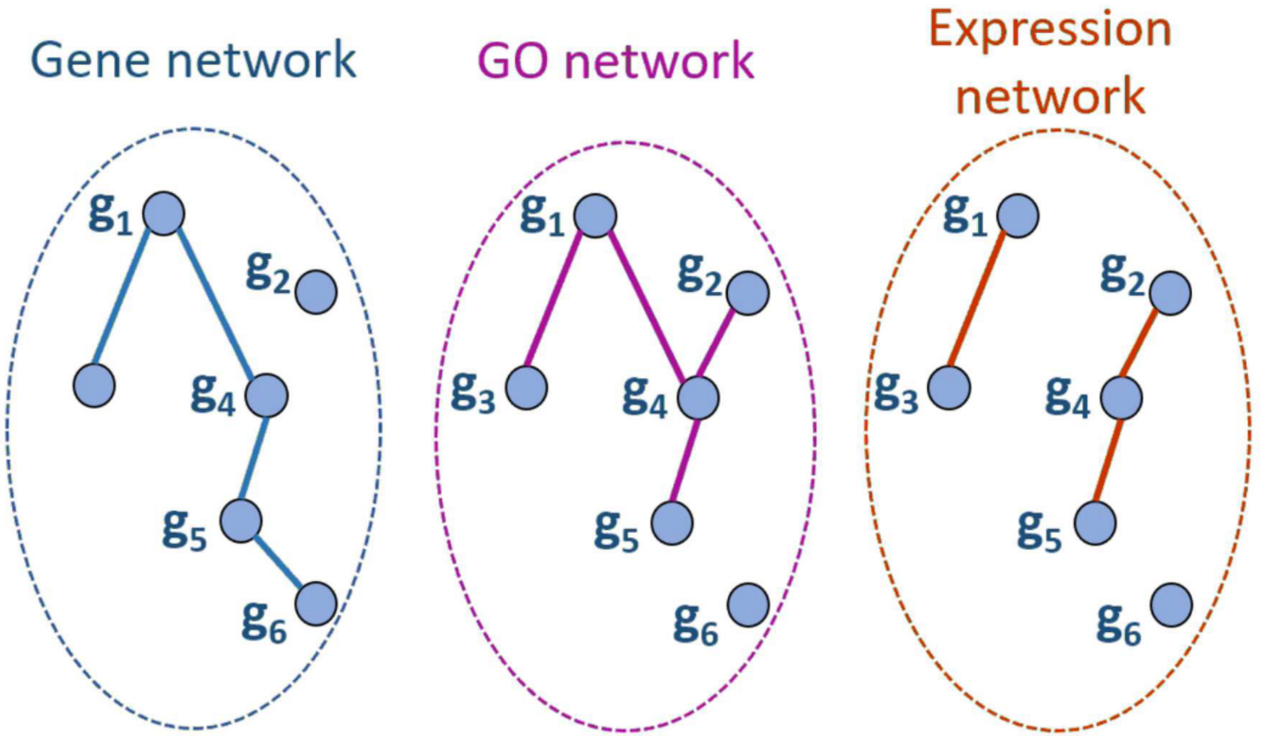
- Recent advances in network-based methods for disease gene prediction, Ata et al., 2021, 10.1093/bib/bbaa303

Features Selection

Find the most relevant features by using different subsets of the *C.Elegans* dataset

Edges: 12 086 627

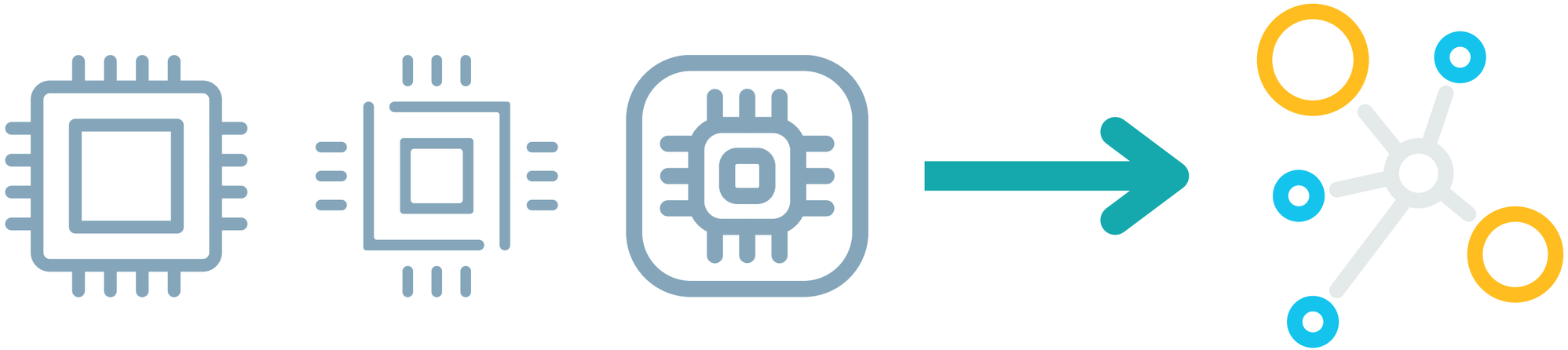
Nodes : 2 977 525



- Recent advances in network-based methods for disease gene prediction, Ata et al., 2021, 10.1093/bib/bbaa303

Benchmarking methods

Finally, compare methods on our dataset



To consider:

- Time complexity of the method
- **… On what do we evaluate our methods ?**

2 approaches:

- Random split
- Time-based split

Institut de Génétique & Développement de Rennes

Thank you for your attention !