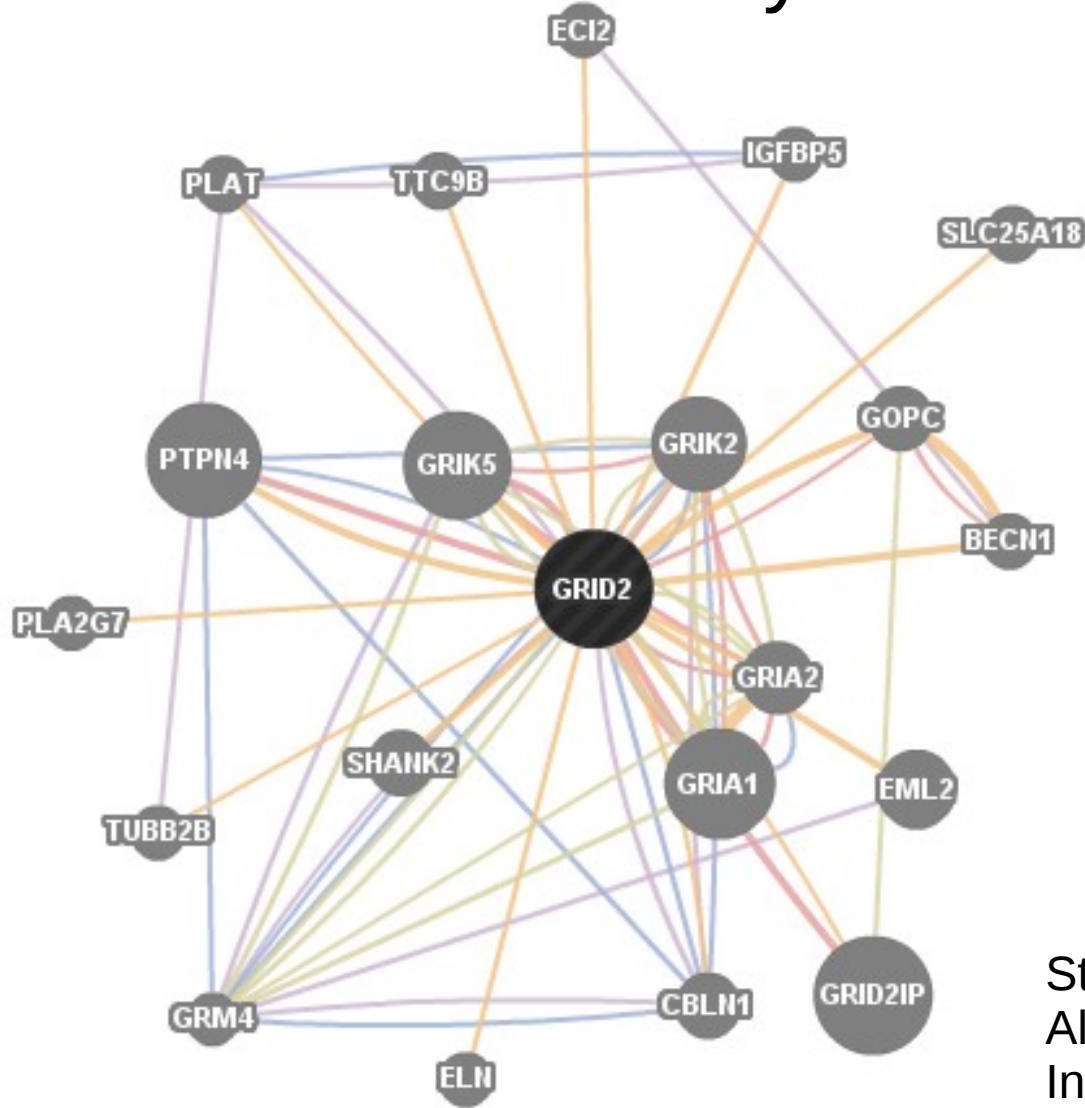


Machine learning in disease subnetwork discovery



Student: Bogdan Kirillov,
Algorithmic Bioinformatics,
Institute of Bioinformatics

Research supervisor: Son Pham, Salk Institute

Why?

- The causes of most neuronal diseases are still unknown
- Neuronal diseases (f. e. autism) are severe and currently incurable
- We have a large whole genome sequencing dataset
- We can explore it with graph theory and machine learning

The Data

- A large (~PetaB) dataset of whole genome sequencing data
- A genemania.org homo sapiens combined gene interaction network

Directed multigraph with 19264 nodes and 7290094 edges

```
>>> print(c.graph.summary())
IGRAPH DNW- 19264 7290094 --
+ attr: name (v), group (e), network (e), weight (e)
>>> []
```

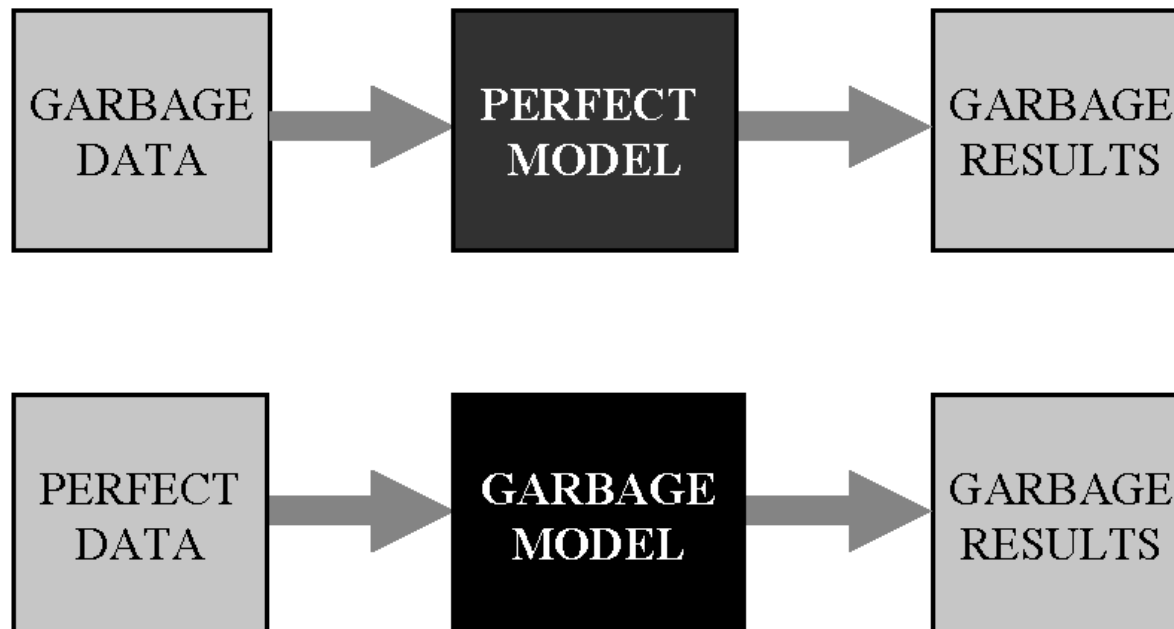
What needs to be done?

- Gene interaction network verification

We have to check whether the network is correct, because of the first rule of data analysis: check your data.

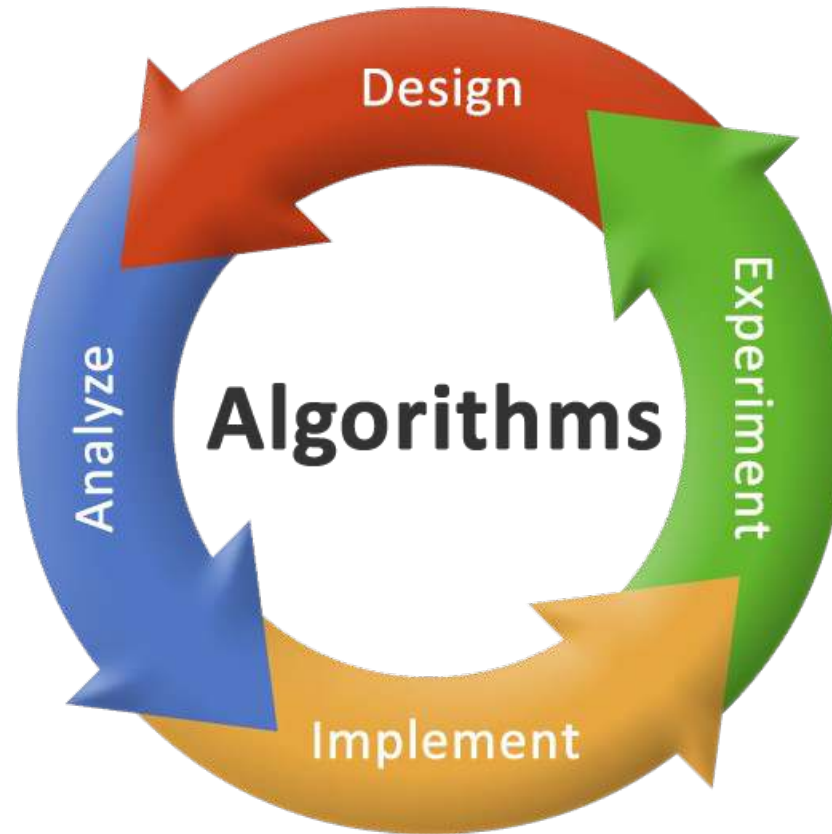
MODEL CALCULATIONS

”Garbage In-garbage Out” Paradigm



What needs to be done?

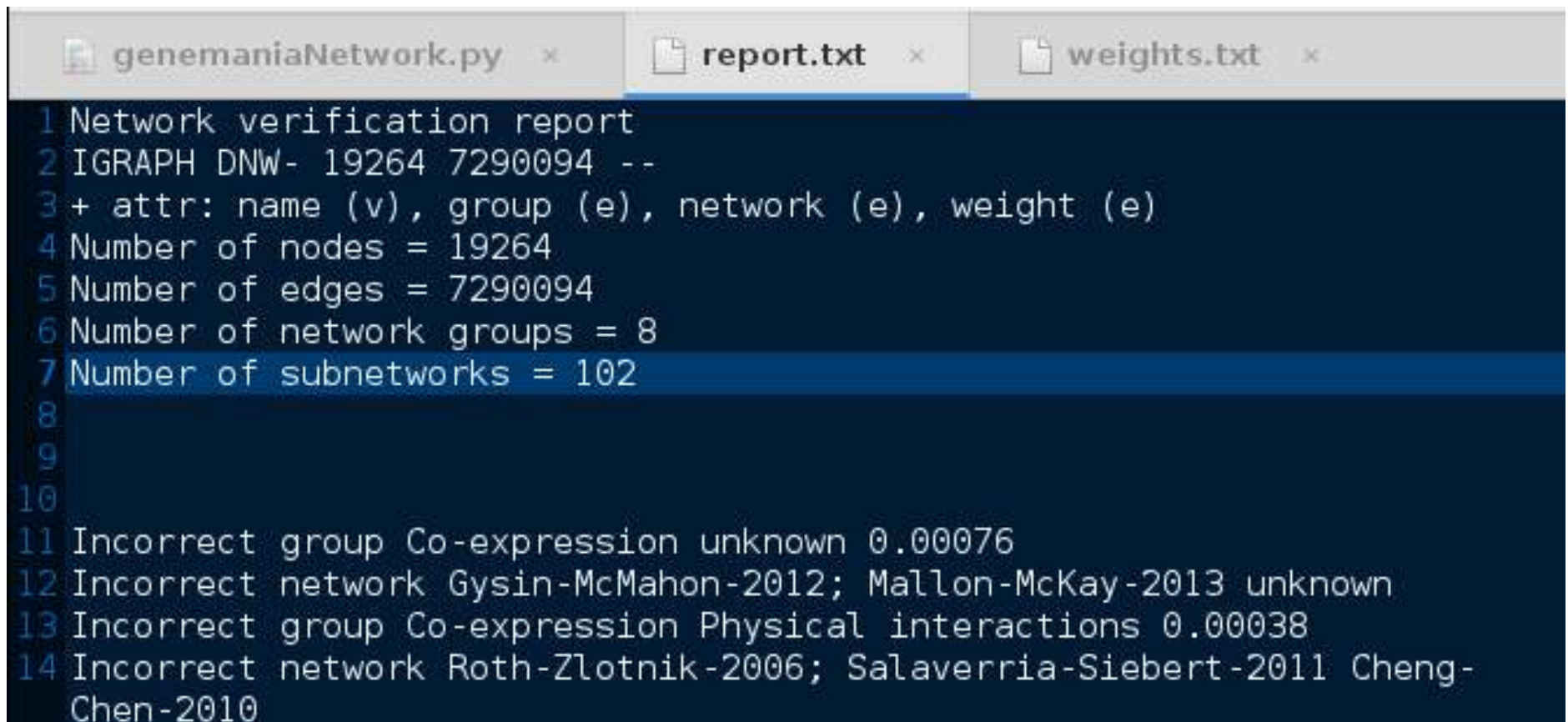
- Algorithm design and implementation



What was done?

- Genemania network verifier

<https://github.com/kspham/diseasenetwork>



```
genemaniaNetwork.py x report.txt x weights.txt x
1 Network verification report
2 IGRAPH DNW- 19264 7290094 --
3 + attr: name (v), group (e), network (e), weight (e)
4 Number of nodes = 19264
5 Number of edges = 7290094
6 Number of network groups = 8
7 Number of subnetworks = 102
8
9
10
11 Incorrect group Co-expression unknown 0.00076
12 Incorrect network Gysin-McMahon-2012; Mallon-McKay-2013 unknown
13 Incorrect group Co-expression Physical interactions 0.00038
14 Incorrect network Roth-Zlotnik-2006; Salaverria-Siebert-2011 Cheng-
    Chen-2010
```

These “Incorrect network” and “Incorrect groups” are yet to deal with

Plan till December

- Get rid of incorrects in the verifier
- Design and implement an algorithm

