Discovering Adenoid Cystic Carcinoma Biomarkers Using a Purpose-Built Hypergraph Database and Link Prediction

SYSTEMS IMAGINATION, INC.

PIETER DERDEYN

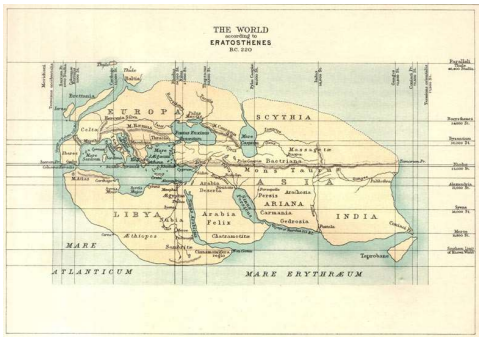
CHRIS YOO, PH.D.

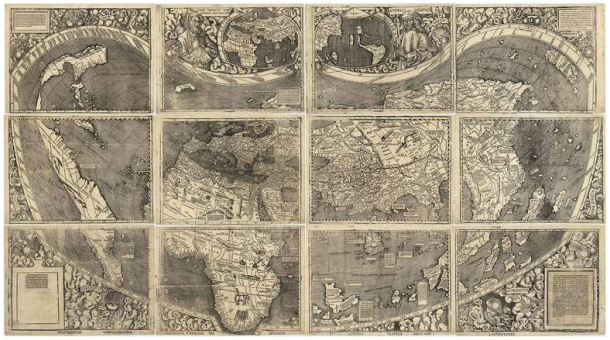
Systems Imagination

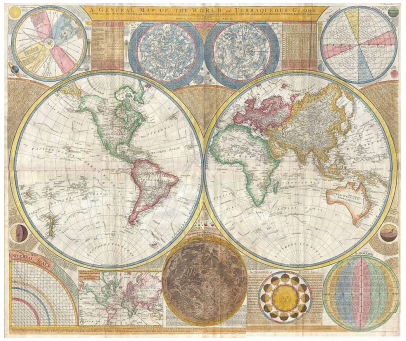
Mapping Big Data

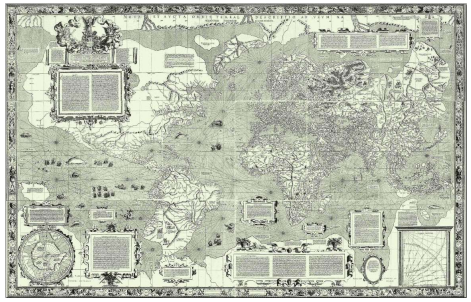
Systems Imagination

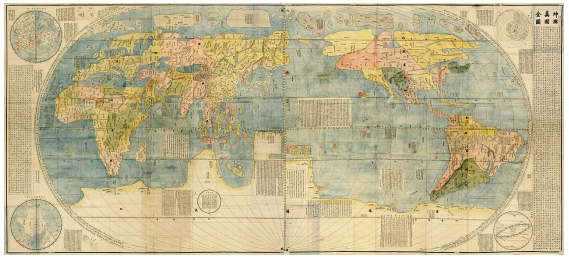
Maps throughout history…





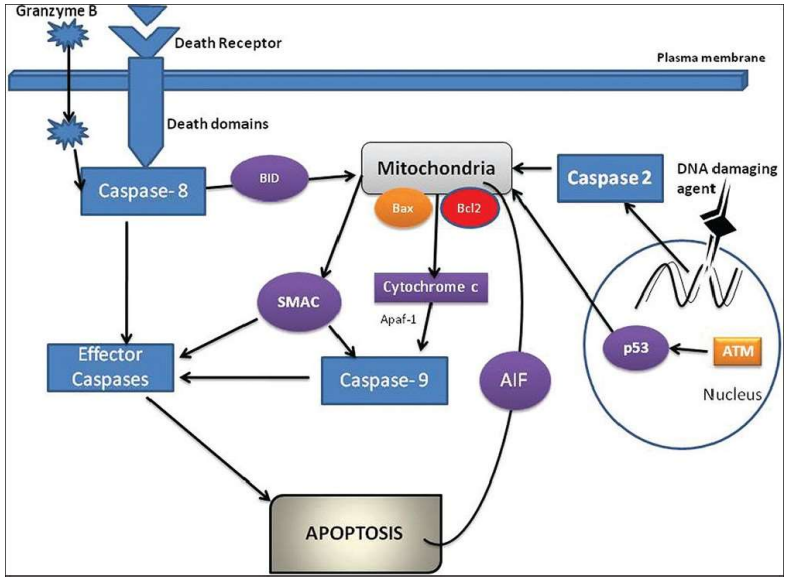






Systems Imagination

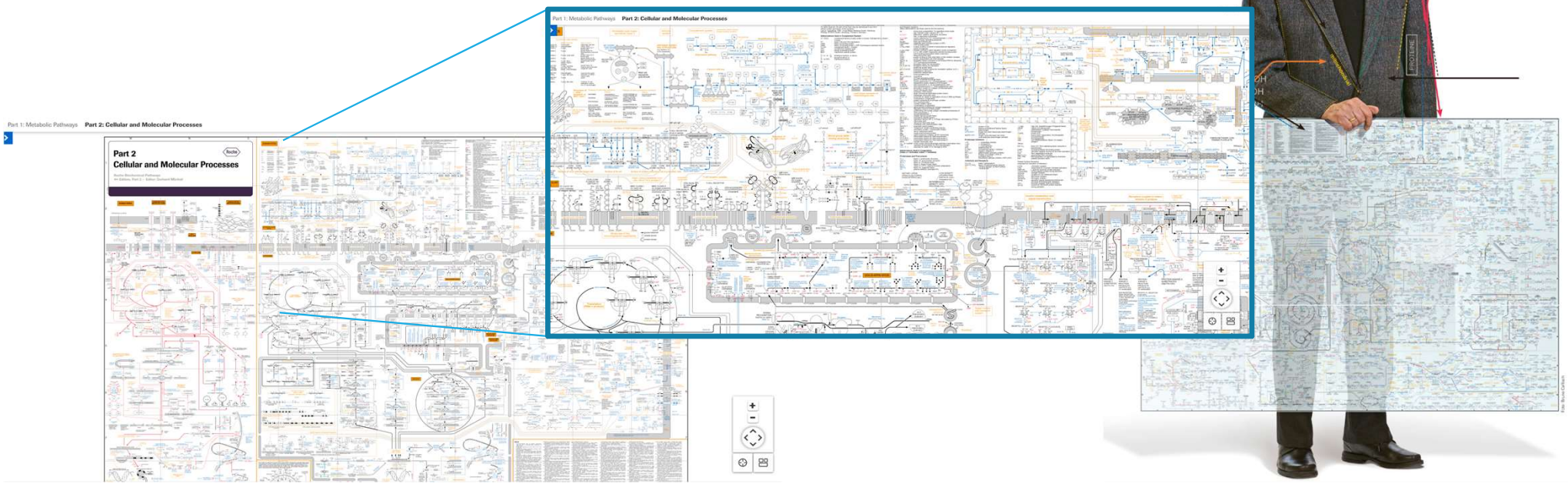
How do we map cancer?



Systems Imagination

Dr. Michel, 40 years of biochemistry in one map

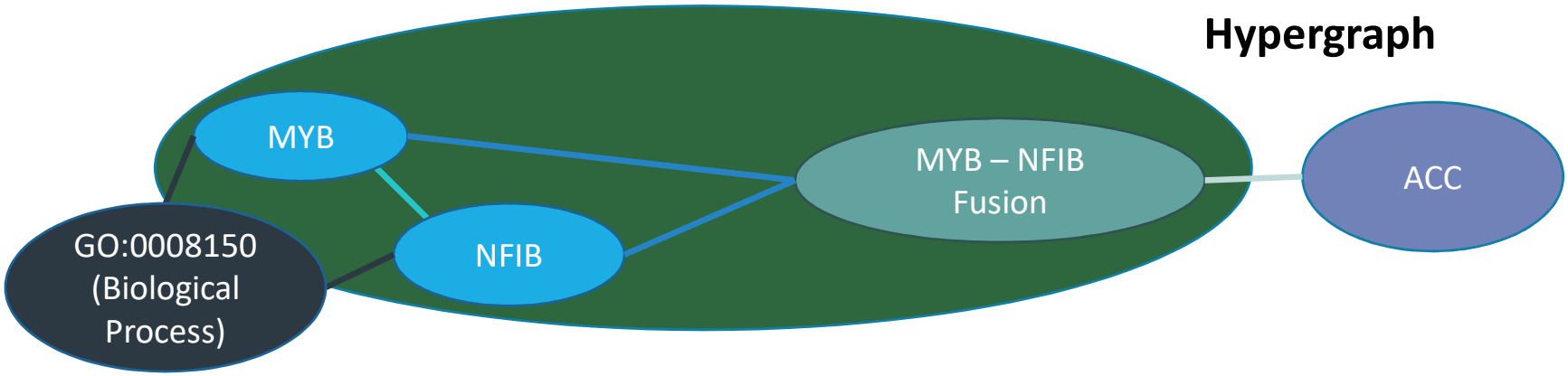
- Data management - linking data into a useful framework
- Interpretation of the meaning of the data in context



Dr Gerhard Michal,Editor of the Roche Biochemical Pathways

A Hypergraph Map of Cancer

Represent the data as knowledge – what's the best way?



Populating the hypergraph

Multiple sources

Requires harmonization

Use Case: Adenoid cystic carcinoma (ACC)

Rare (~1200/yr in US)

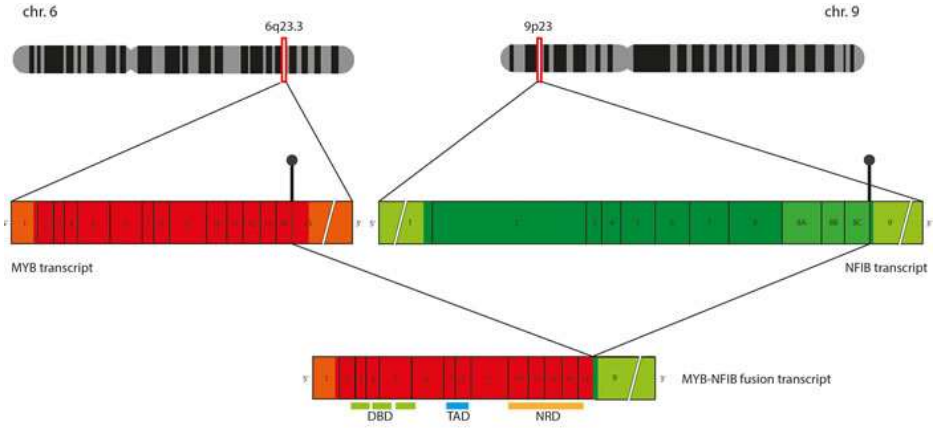
Majority of ACC cases display activation of MYB, commonly through genomic translocation event with NFIB, both transcription factors

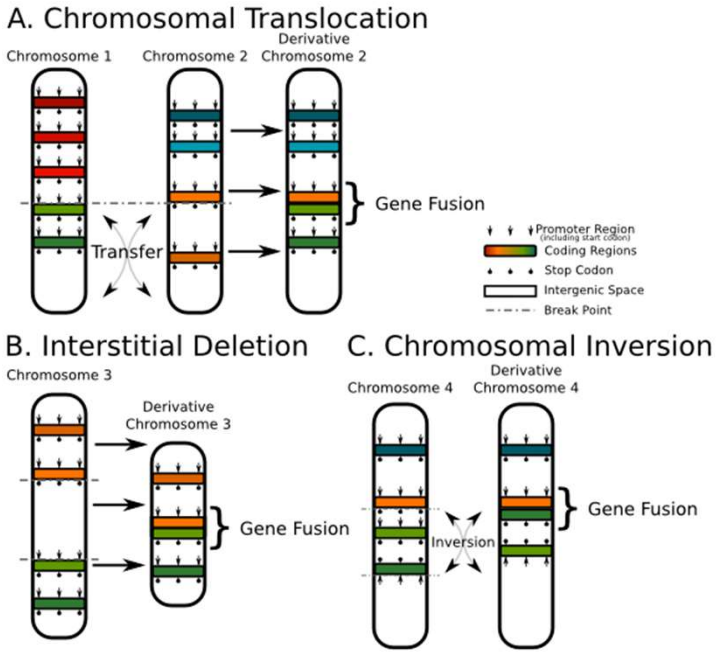
Initial prognosis with surgery is good (5yr: 89%) but long term follow up indicates aggressive recurrence (15yr: 40%)

What data can be examined to find hypotheses to explain these results?

Target: Gene fusions

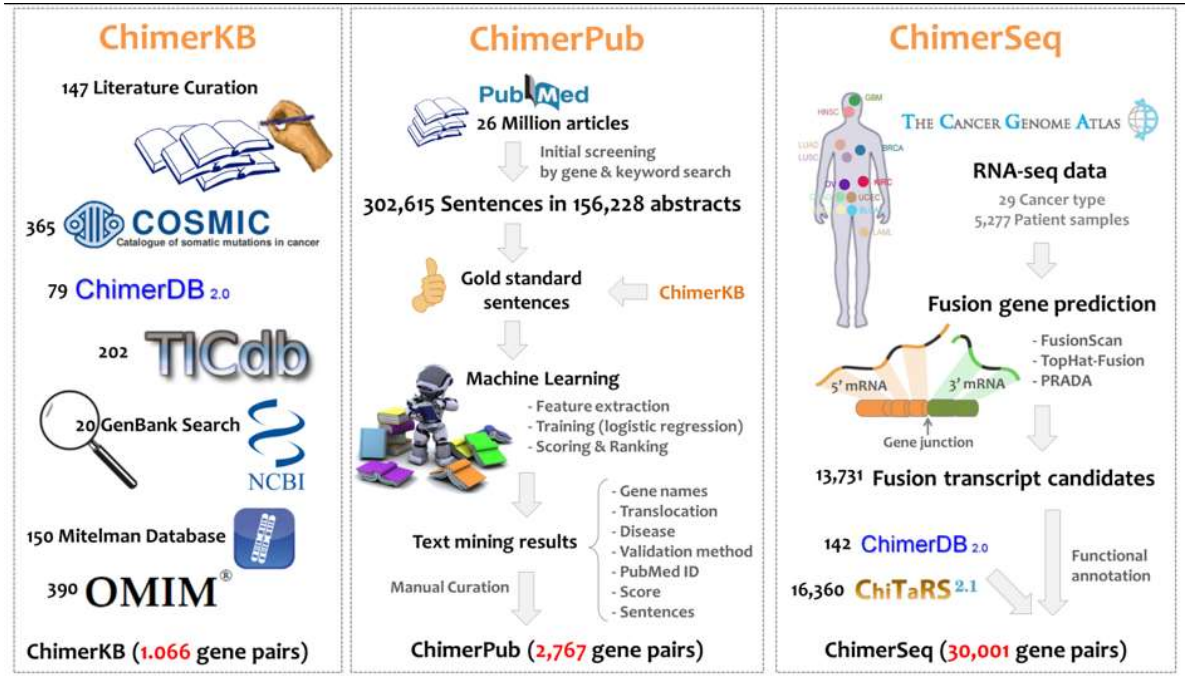
- Hybrid gene from two previously separate genes (Wikipedia)
- Are often oncogenes because they lead to much more active abnormal proteins than normal genes
- MYB+MYB1+NFIB





Systems Imagination

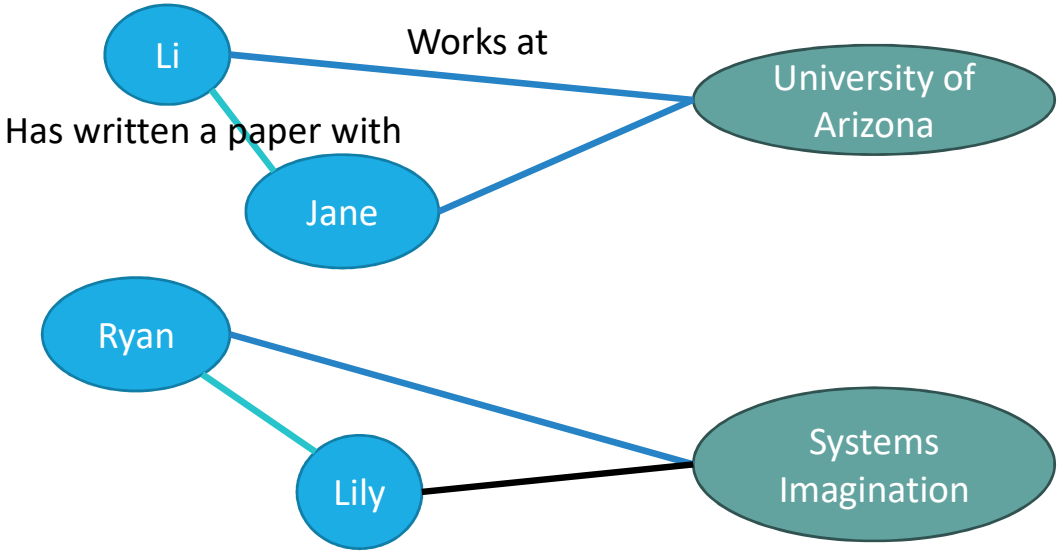
Gene fusions – Data Sources



Link Prediction

For a given pair of nodes, we would like to predict whether they have a certain edge type connecting them

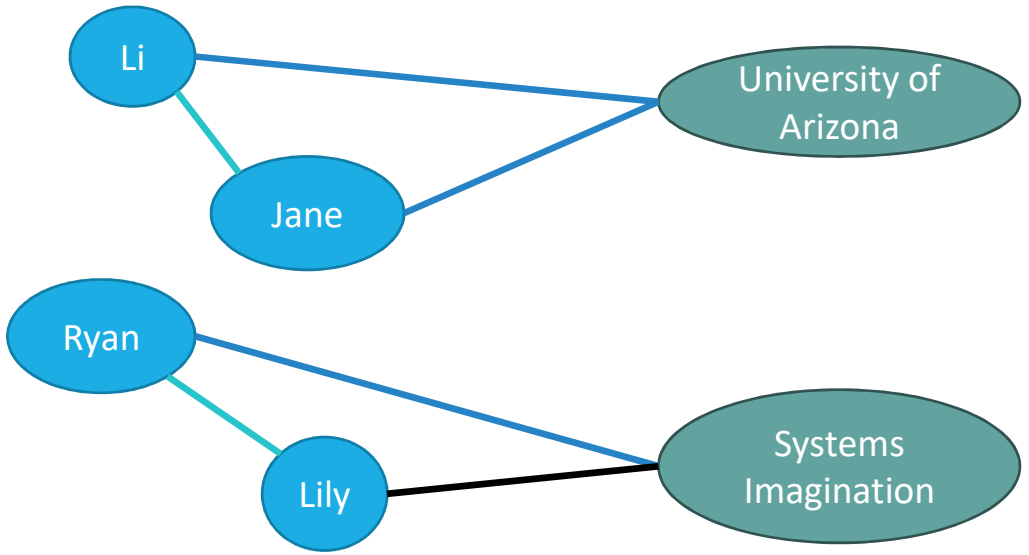
For example, what is the likelihood that Lily works at Systems Imagination?



Link Prediction

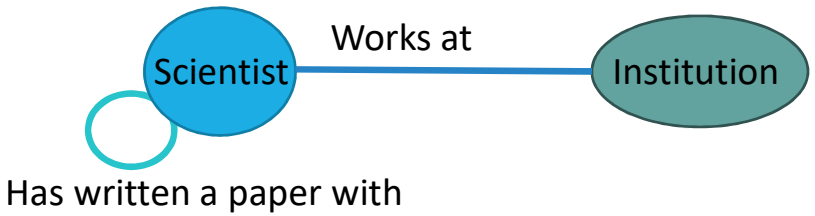
Train a supervised learning model using topological features like:

- Path counts
- Metapath counts

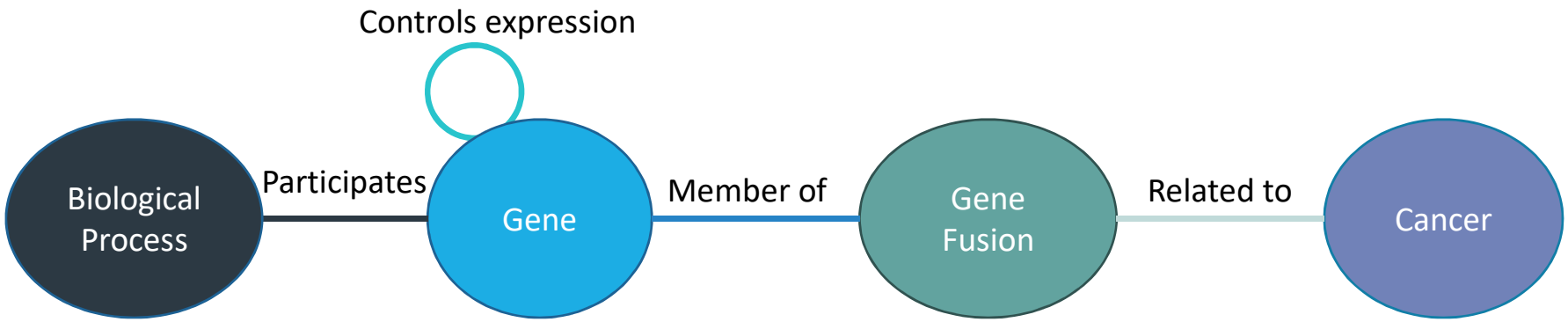


Link Prediction

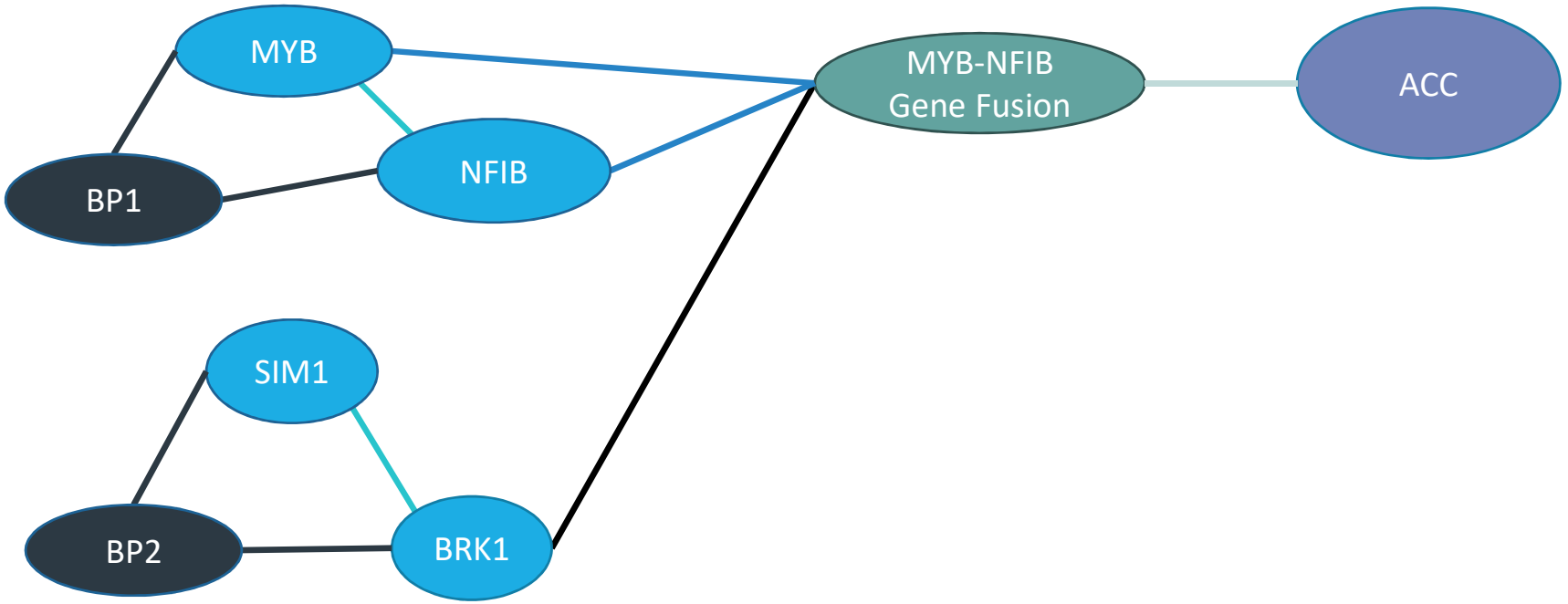
Network Schema – a representation of all node types (metanodes) and the edge types (metaedges) between them



Link Prediction – Gene Fusions

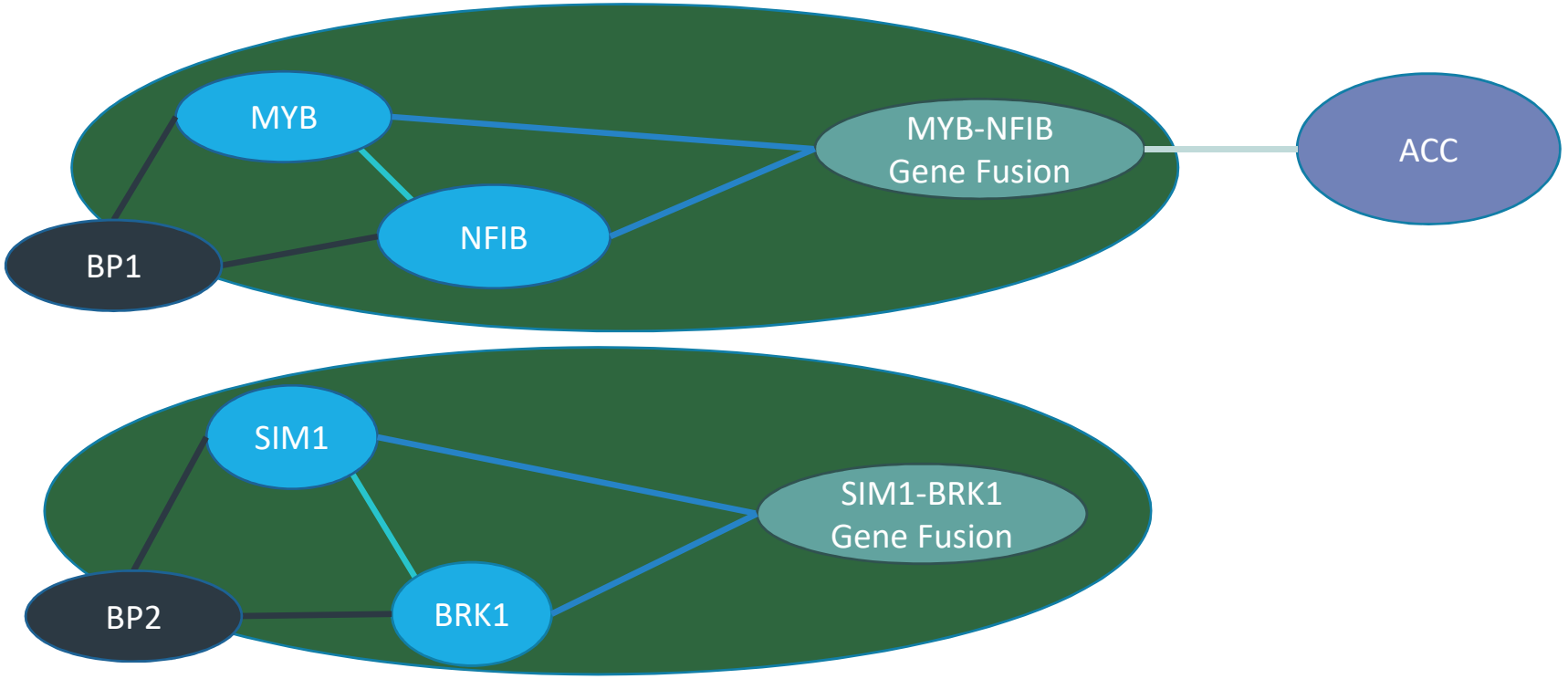


Link Prediction – Gene Fusions



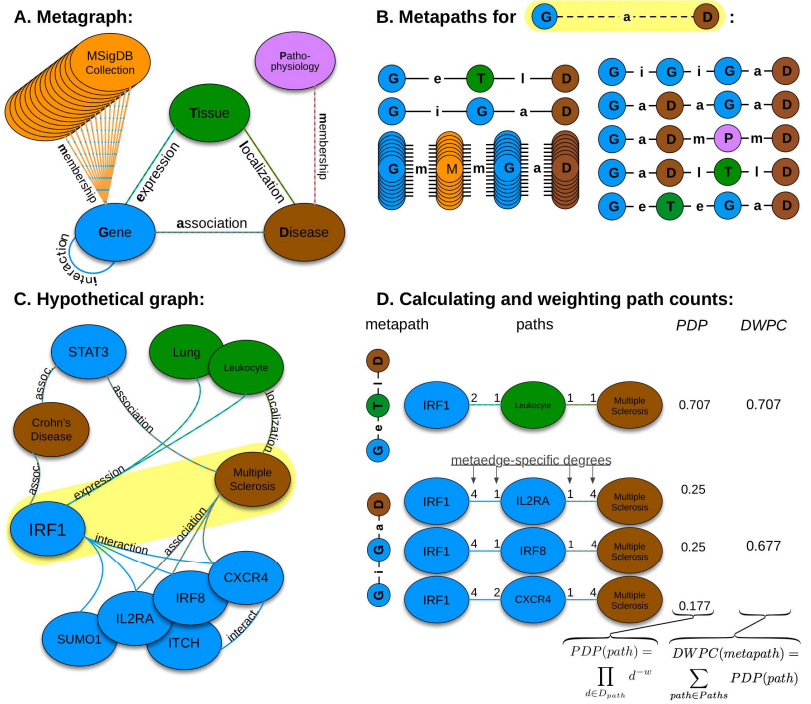
Systems Imagination

Hyperedge Prediction – Gene Fusions



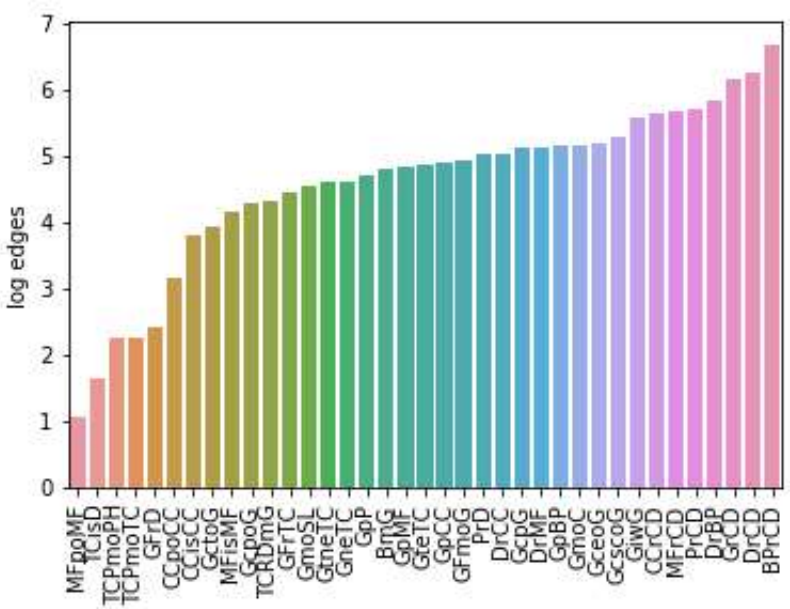
Mining Heterogenous Information Networks

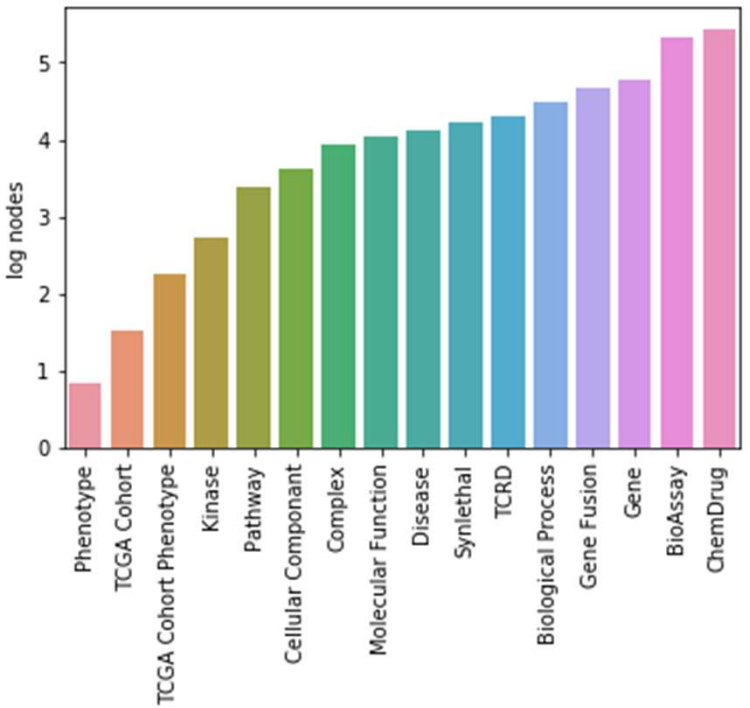
Hetionet	Cancer Research Hypergraph
David Himmelstein et al.	Systems Imagination, Inc.
47,000 nodes	695,464 nodes
11 metanodes	16 metanodes
2,250,000 edges	12,007,912 edges
24 metaedges	41 metaedges



Himmelstein et al

Nodes and Edges





Paths

Predictions

Gene Fusion Prediction Pipeline

For a given pair of genes, are they in a gene fusion or not?

Dataset:
Cancer Research Hypergraph Database

Features:
DWPC (Degree Weighted Path Count), Degrees of nodes, prior likelihood of gene fusion

Supervised Learning Models:
Random Forest, Logistic Regression, Decision Trees, XGBoost, Neural Networks

Model Interpretation:
Assess predictions, feature analysis

Challenges

Data integration:
Integrating data from dozens of sources and converting between 3 different formats

Feature computation:
10 times the data, 100 times the computational cost

Systems Imagination

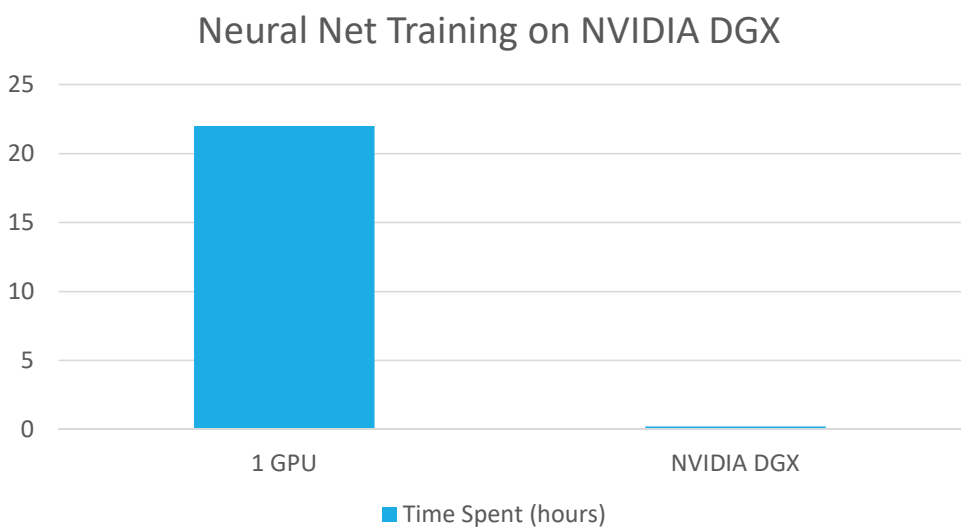
Strategies

NVIDIA DGX

- 40 CPUs
- 256 GB RAM
- 4 x Tesla V100 GPUs (64GB memory total)

Can do production level computation locally

Results



Systems Imagination Benchmarking

Dense NN built with mxnet and keras

7 hidden layers with 200-700 neurons each

33,658,931 rows of data
18 features
6 classes

Strategies

Multi-processing: 3 lines of python code sped processing up by 6 times

GPU acceleration:

- Accelerated numpy computations by 10 times by moving to CuPy
- Accelerated deep learning 20 times by using mxnet on NVIDIA DGX

Profiling and Debugging code: what is the bottleneck and how can I relieve it

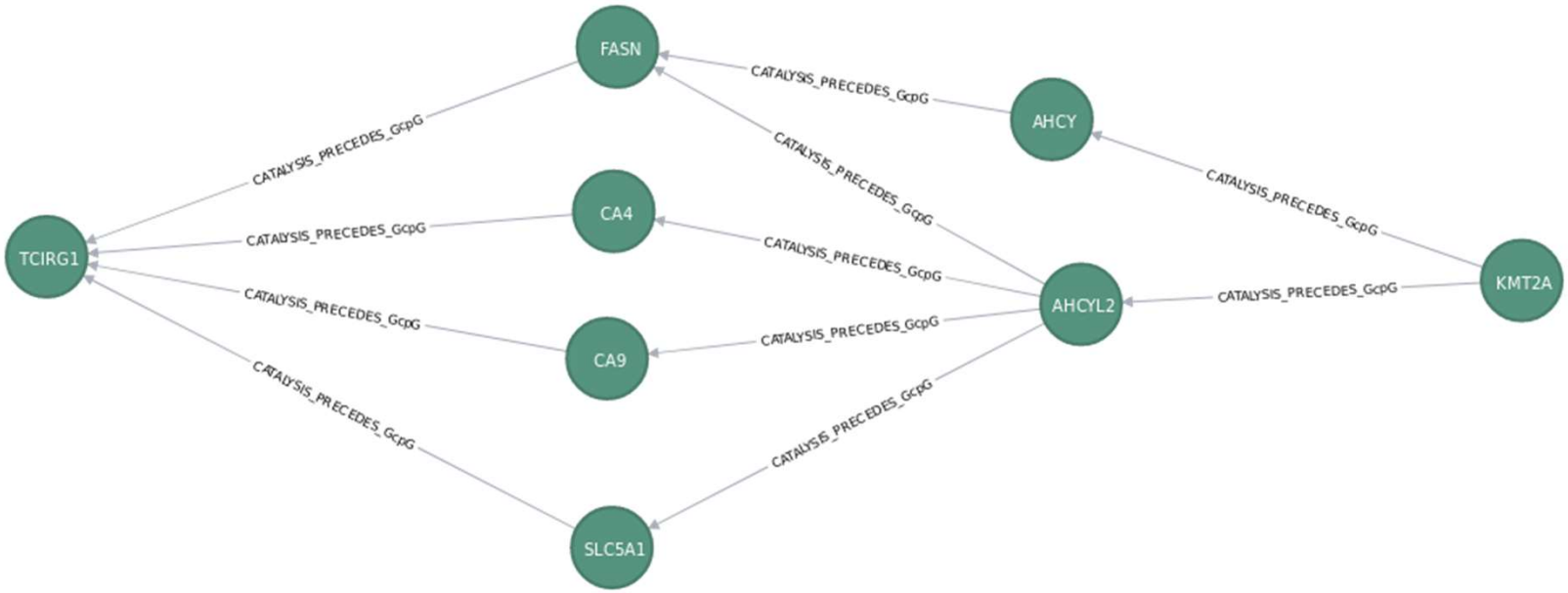
- Rabbit hole: Not optimizing just the code, but optimizing the time spent developing and running the code

Results

Gene 1	Gene 2	Probability of Gene Fusion
EWSR1	HMGA2	0.929894619
BBS9	KMT2A	0.928350421
IQCJ	KMT2A	0.927711269
CYP11B1	KMT2A	0.926647616
KMT2A	TCIRG1	0.912818918
KMT2A	VEPH1	0.911844923
CCR6	KMT2A	0.873986434
KCNQ1	KMT2A	0.834963505
ACSL1	KMT2A	0.834097153
EWSR1	RUNX1T1	0.832868597

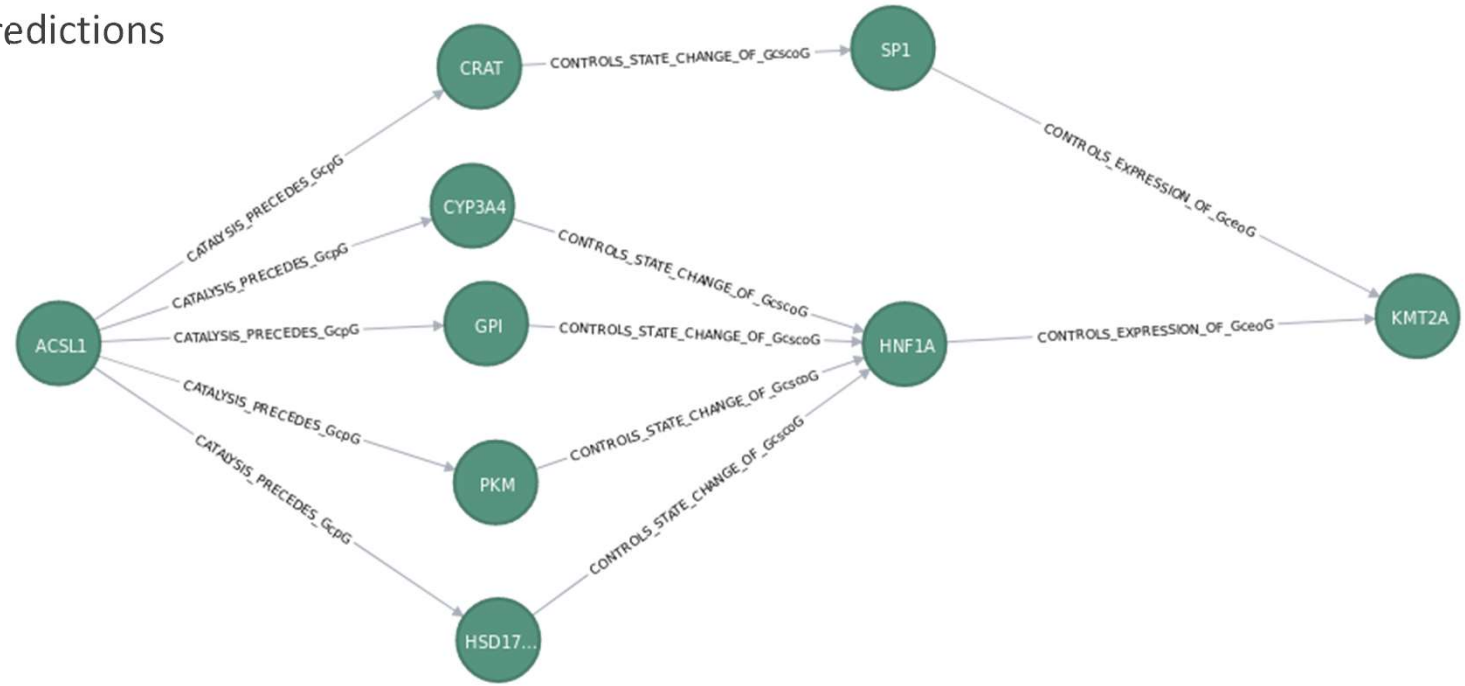
Results

Predictions



Results

Predictions



Systems Imagination

The Team