Deep Learning for Network Biology

Marinka Zitnik and Jure Leskovec

Stanford University

This Tutorial

snap.stanford.edu/deepnetbio-ismb

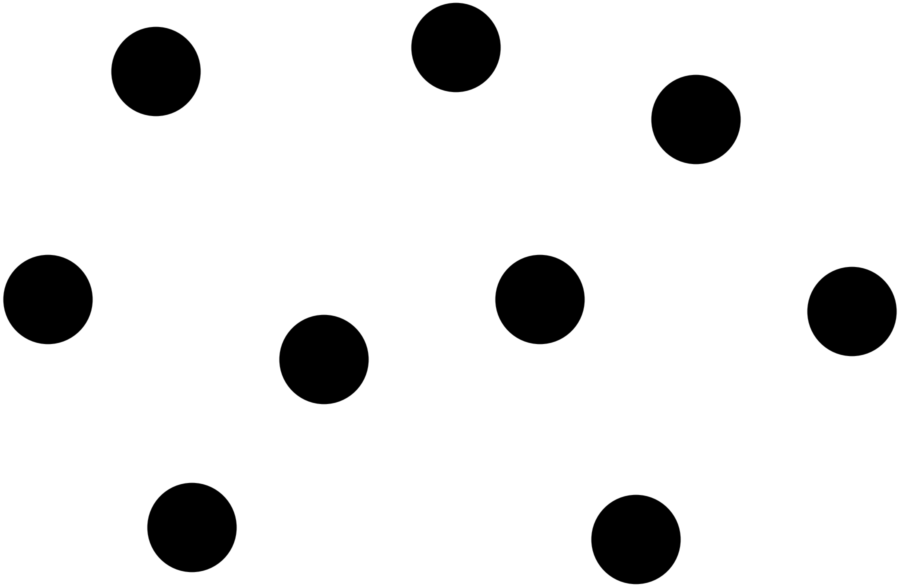
ISMB 2018

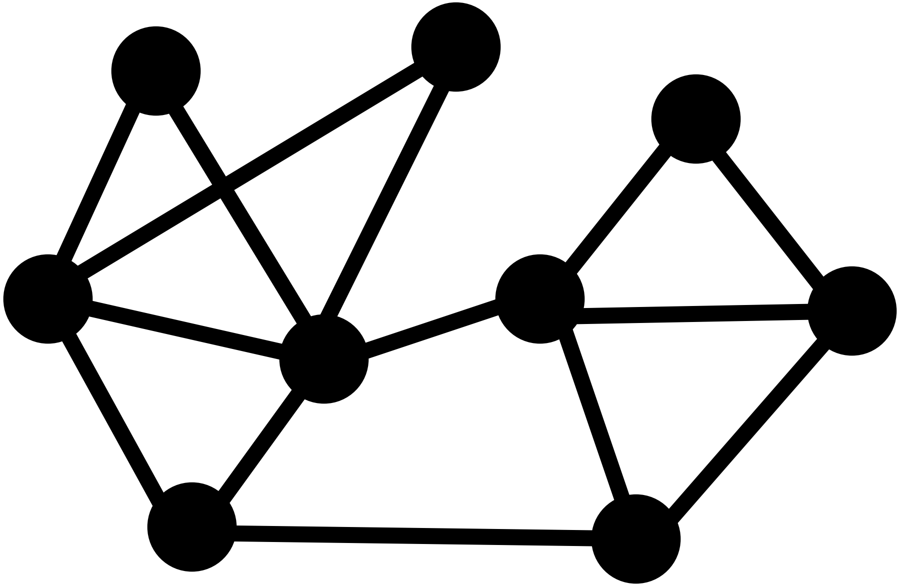
July 6, 2018, 2:00 pm - 6:00 pm

ISMB 2018
CHICAGO JULY 6-10
HYATT REGENCY HOTEL
iSCB
26TH CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY

Why networks?

Networks are a general language for describing and modeling complex systems





Network!

Why Networks? Why Now?

- **Question:** How are human genetic diseases and the corresponding disease genes related to each other?
- **Findings:** Genes associated with similar diseases are likely to interact and have similar expression

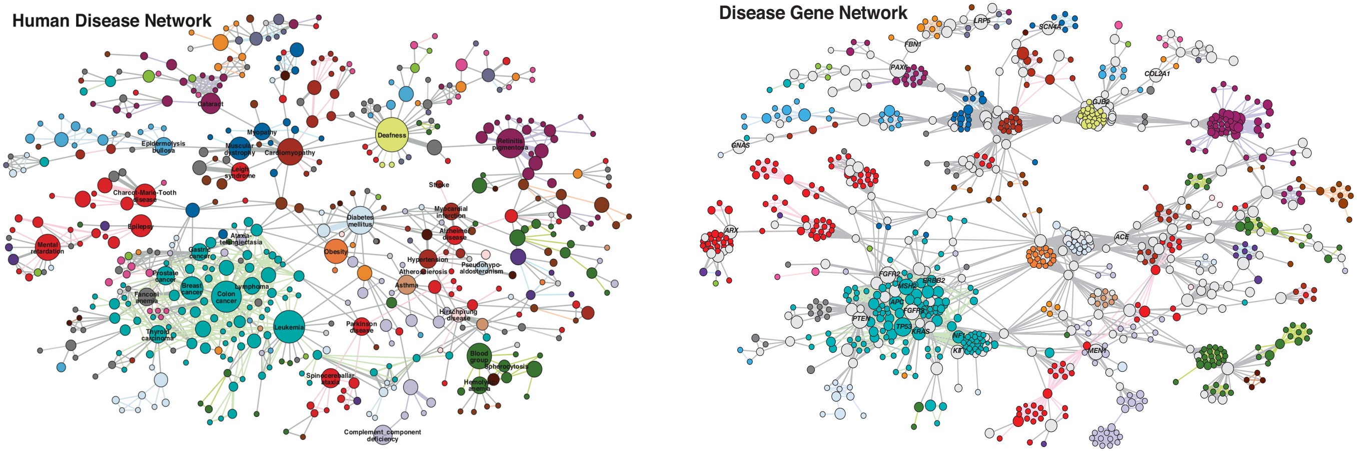


Image from: Goh et al. 2007. The human disease network. *PNAS*.

Why Networks? Why Now?

- **Question:** How to simulate a basic eukaryotic cell?
- **Findings:** Simulations reveal molecular mechanisms of cell growth, drug resistance and synthetic life

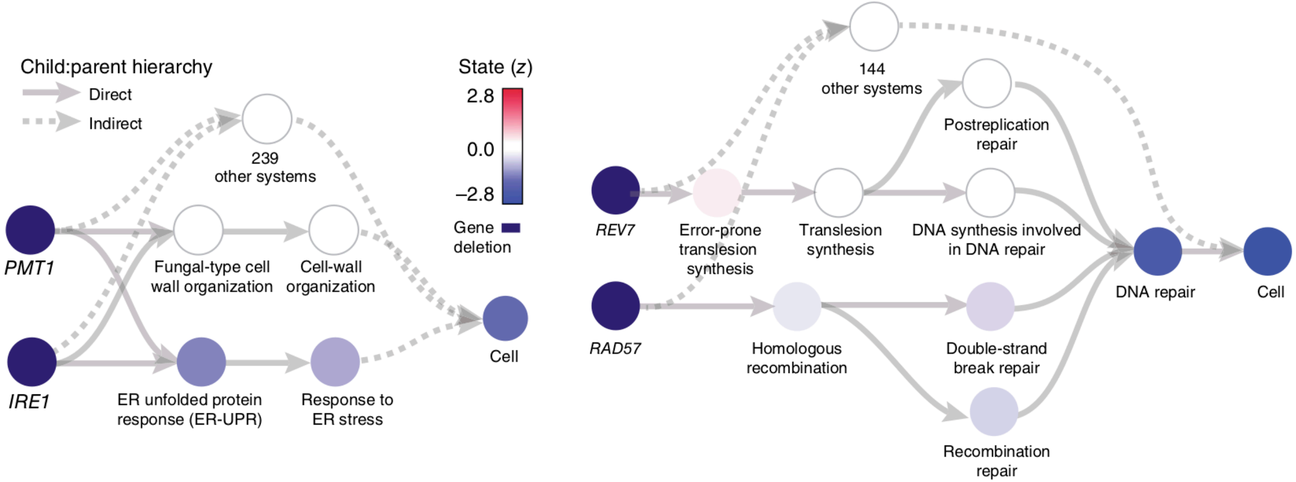
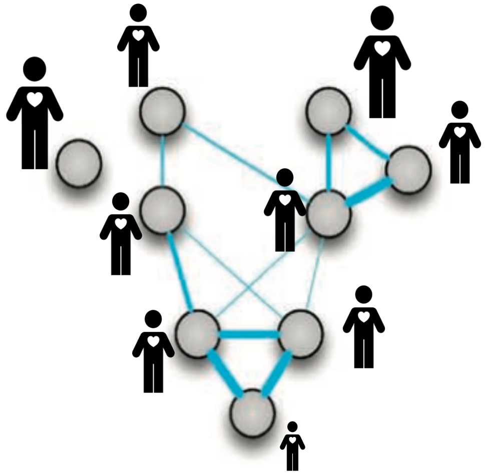


Image from: Ma et al. 2018. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*.

Why Networks? Why Now?

- **Question:** How to discover heterogeneity of cancer?
- **Findings:** Analysis identifies new cancer subtypes with distinct patient survival



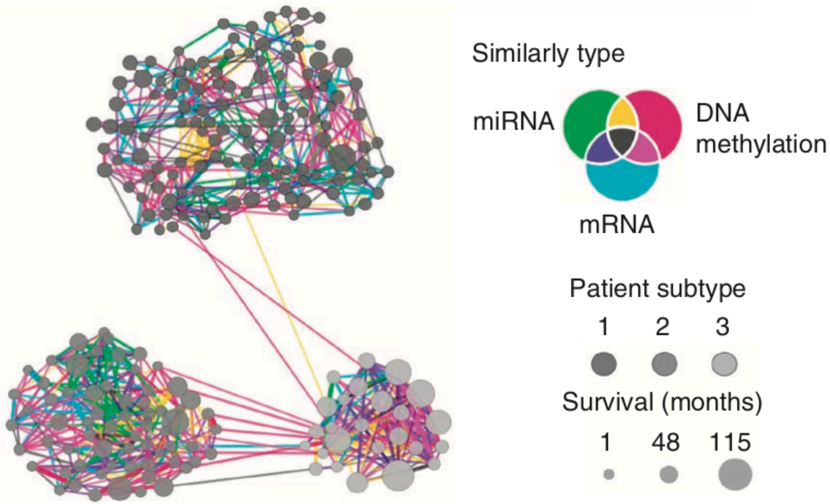
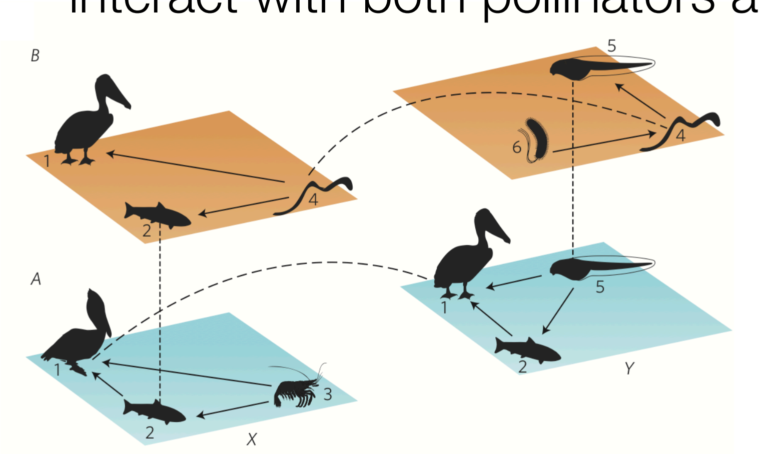


Image from: Wang et al. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*.

Why Networks? Why Now?

- **Question:** How to study ecological systems?
- **Findings:** Pollinators interact with flowers in one season but not in another, and the same flower species interact with both pollinators and herbivores



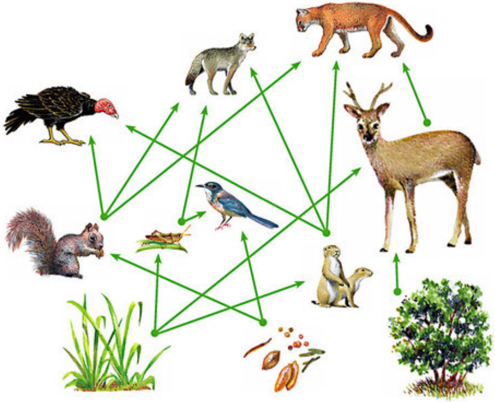


Image from: Pilosof et al. 2017. The multilayer nature of ecological networks. *Nature Ecology and Evolution*.

Why Networks? Why Now?

- **Question:** What are features of human microbiome?
- **Findings:** Microbiota reflects the seasonal availability of different types of food and differentiate industrialized and traditional populations

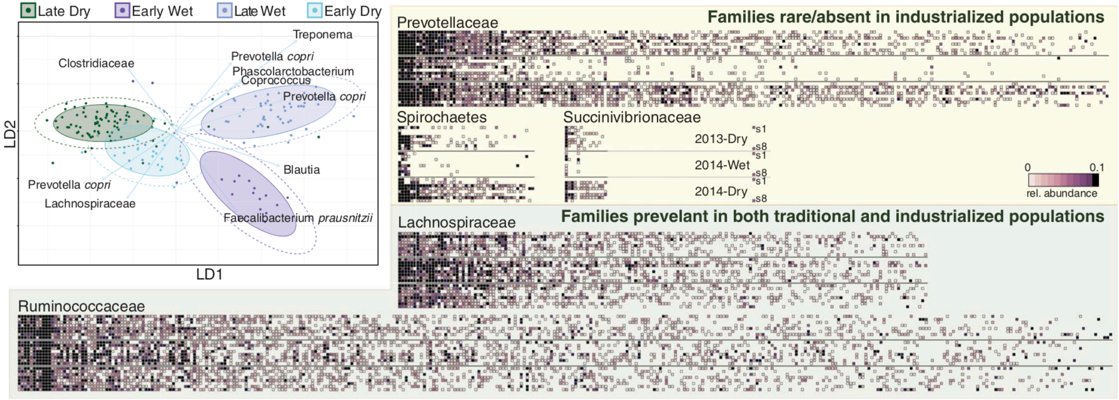
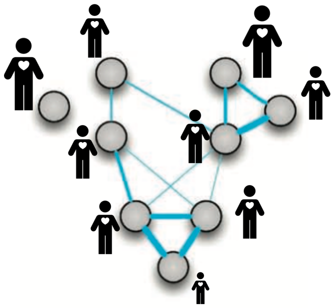
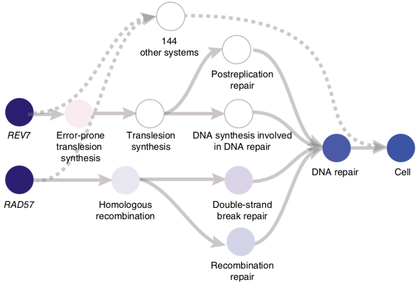


Image from: Smits et al. 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science*.

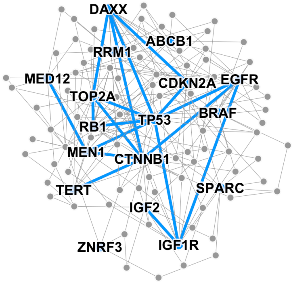
Many Data are Networks



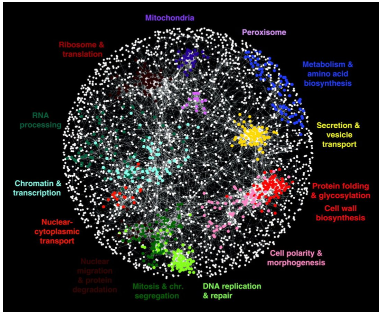
Patient networks



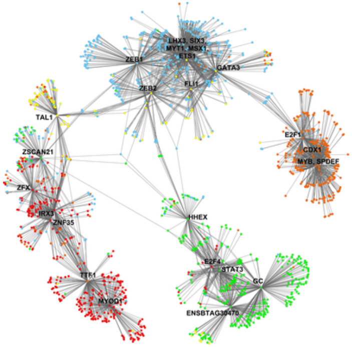
Hierarchies of cell systems



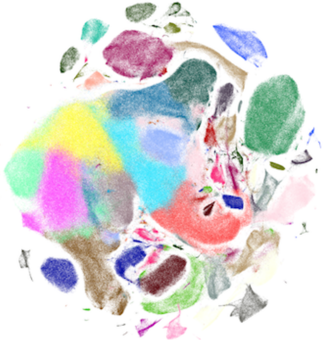
Disease pathways



Genetic interaction networks



Gene co-expression networks

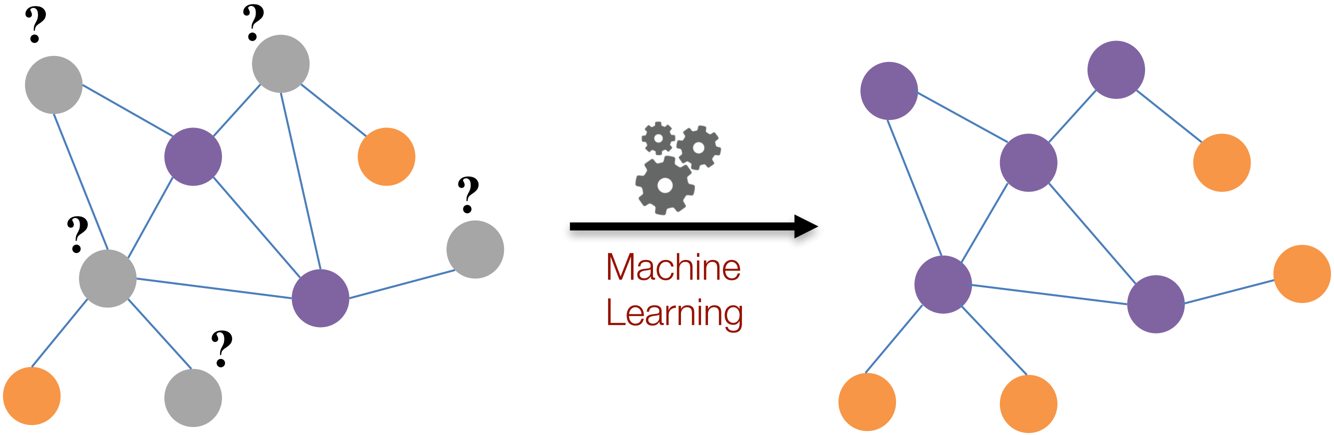


Cell-cell similarity networks

Ways to Analyze Networks

- Predict a type of a given node
  - Node classification
- Predict whether two nodes are linked
  - Link prediction
- Identify densely linked clusters of nodes
  - Community detection
- How similar are two nodes/networks
  - Network similarity

Example: Node Classification



Example: Node Classification

Classifying the function of proteins in the interactome!

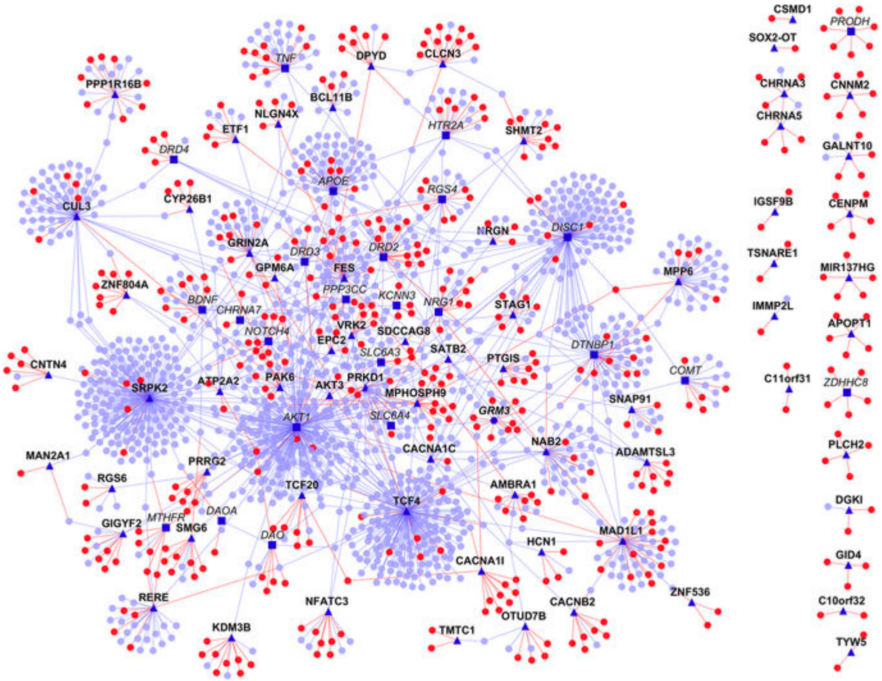
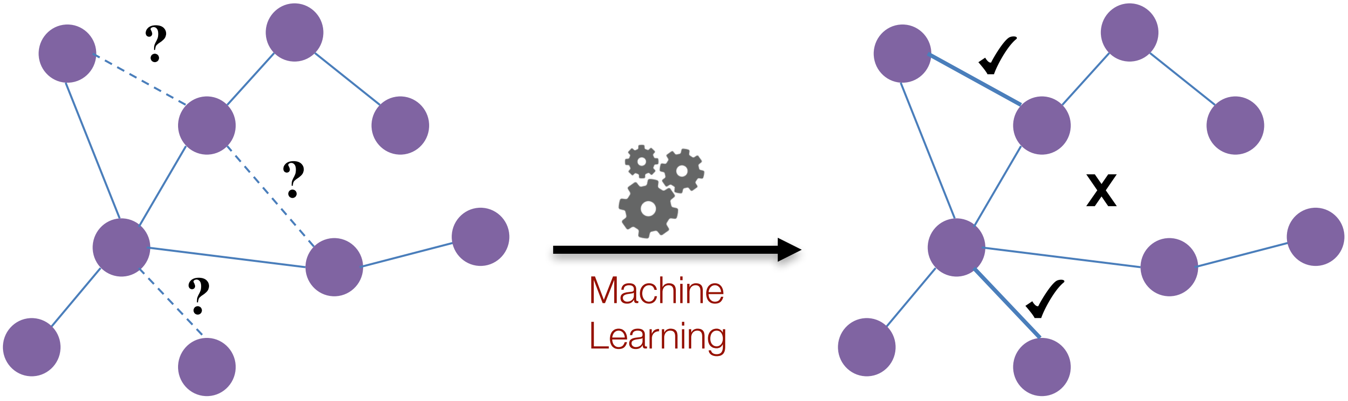


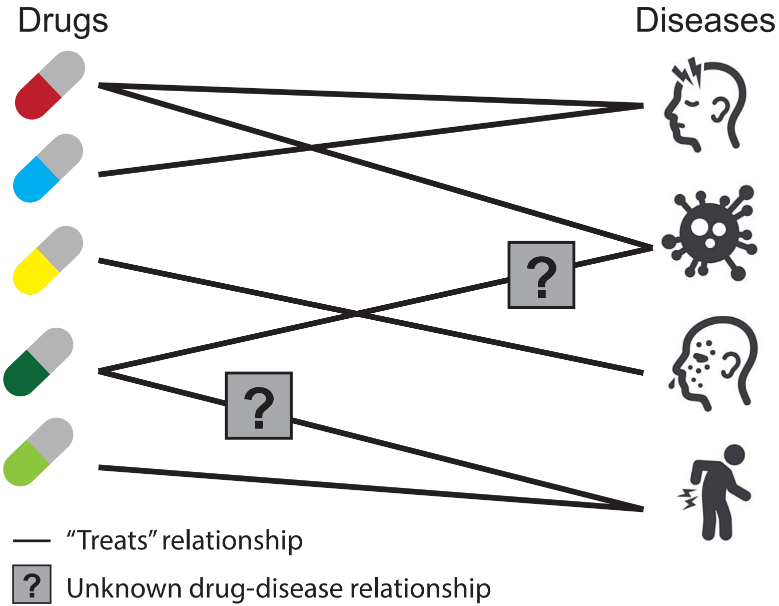
Image from: Ganapathiraju et al. 2016. Schizophrenia interactome with 504 novel protein–protein interactions. *Nature*.

Example: Link Prediction

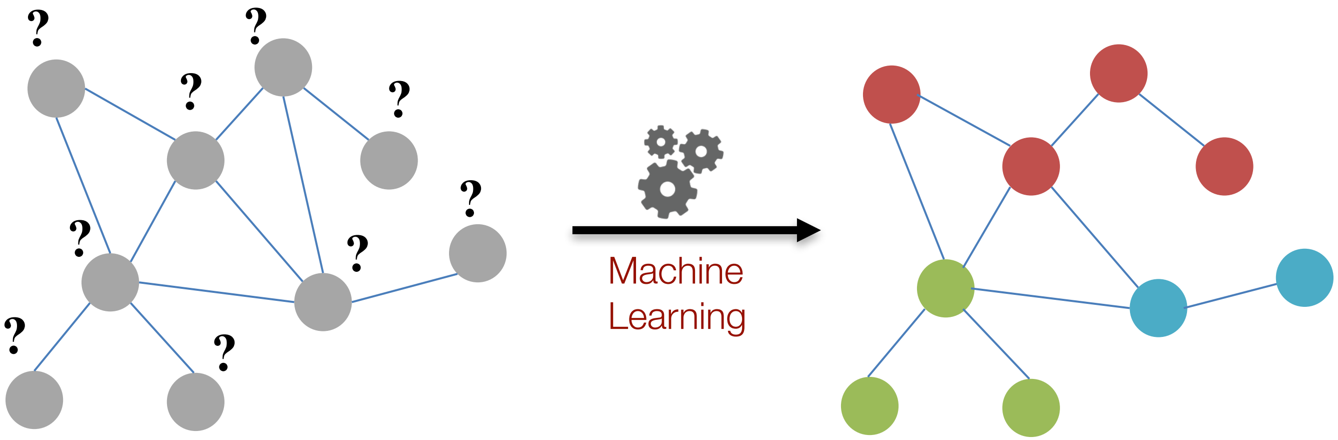


Example: Link Prediction

Predicting which diseases a new molecule might treat!



Example: Community Detection



Example: Community Detection

Identifying disease proteins in the interactome!

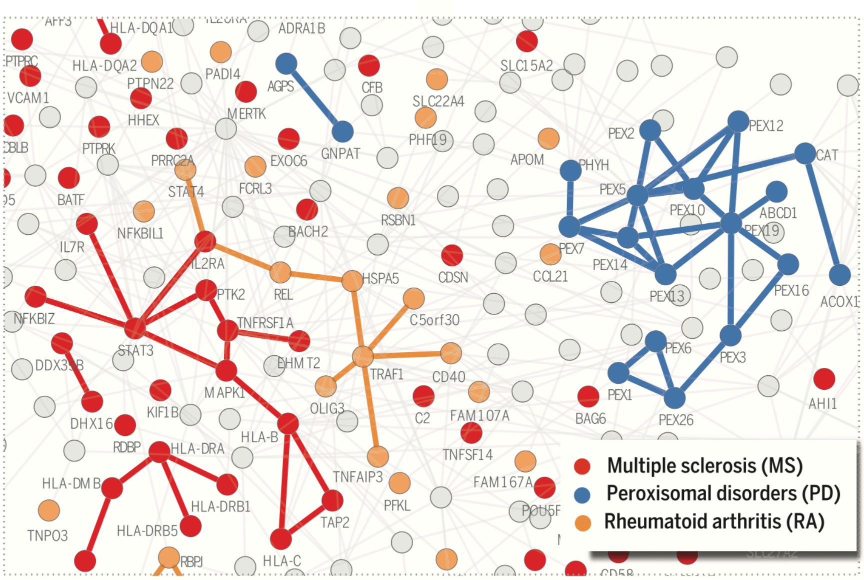
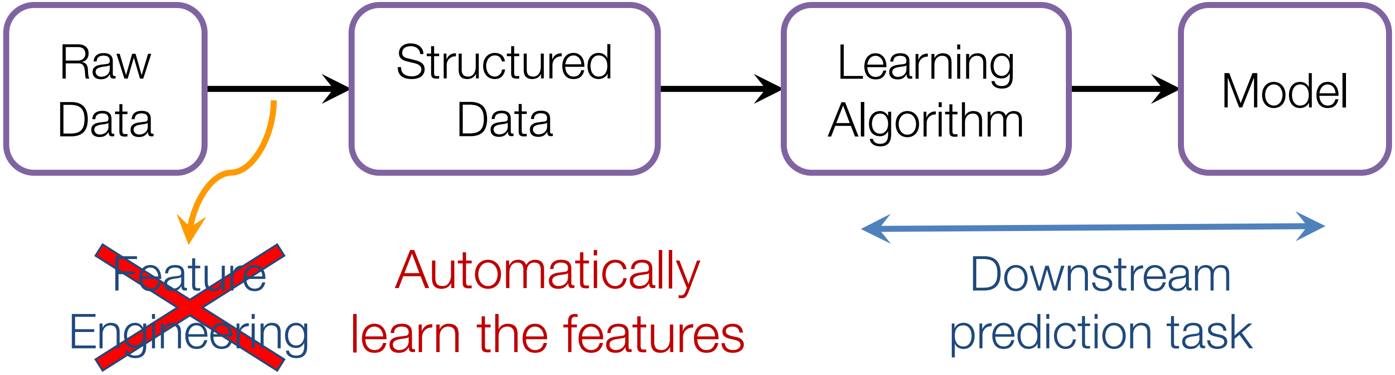


Image from: Menche et al. 2015. [Uncovering disease-disease relationships through the incomplete interactome](). *Science*.

Network Analytics Lifecycle

- (Supervised) Machine Learning Lifecycle: This feature, that feature. Every single time!



Raw Data

Structured Data

Learning Algorithm

Model

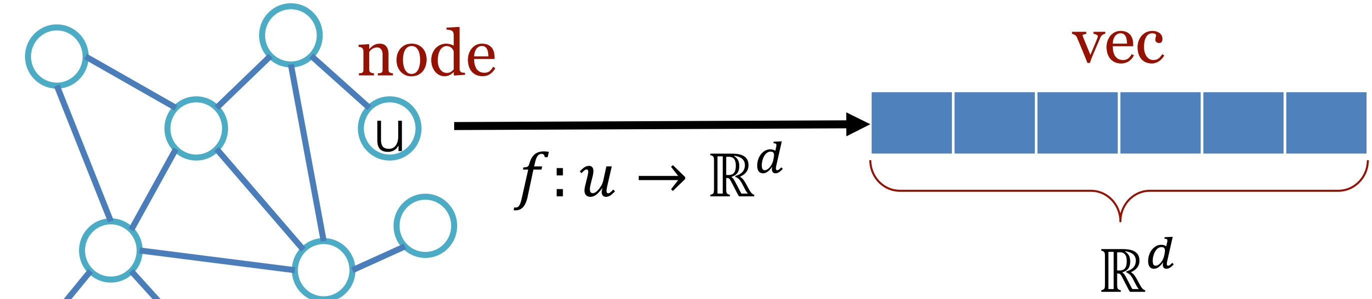
~~Feature Engineering~~

Automatically learn the features

Downstream prediction task

Feature Learning in Graphs

Goal: Efficient task-independent feature learning for machine learning in networks!

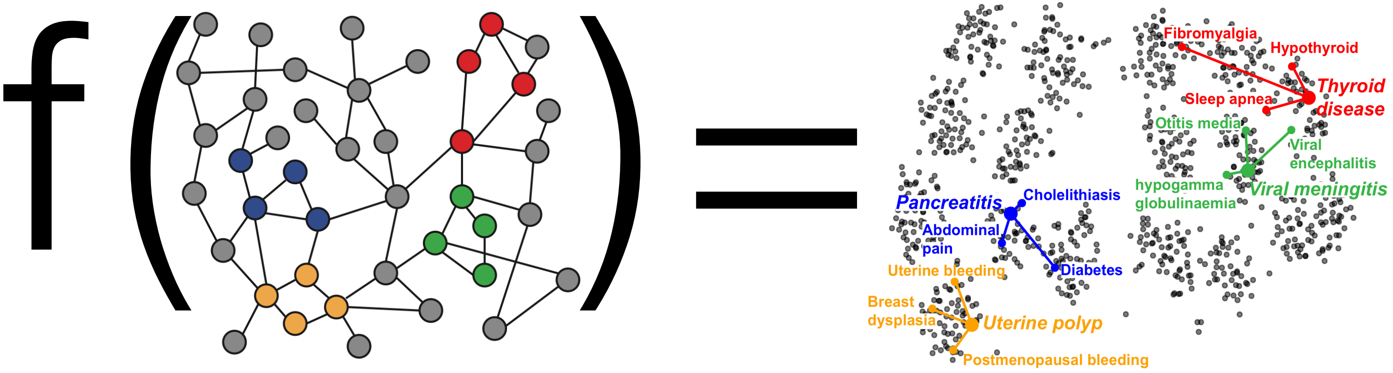


$f: u \rightarrow \mathbb{R}^d$

$\mathbb{R}^d$

Feature representation, embedding

Feature Learning in Graphs



Disease similarity network

2-dimensional node embeddings

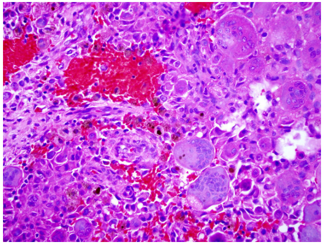
Input

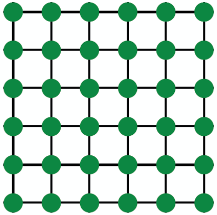
Output

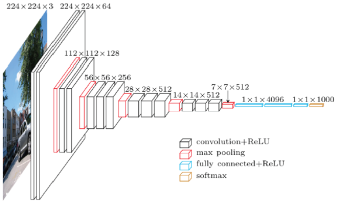
How to learn mapping function *f*?

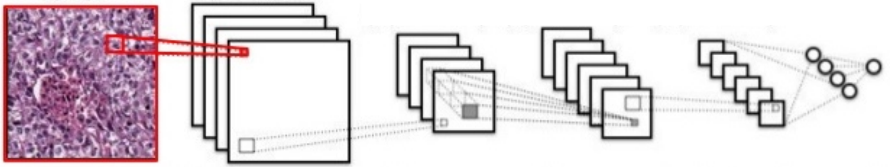
Why Is It Hard?

- Modern deep learning toolbox is designed for grids or simple sequences
  - Images have 2D grid structure
  - Can define convolutions (CNN)



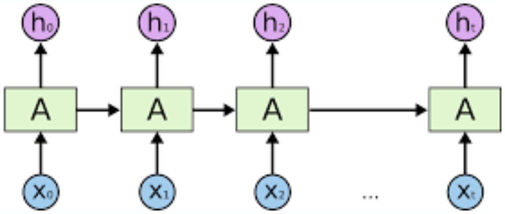






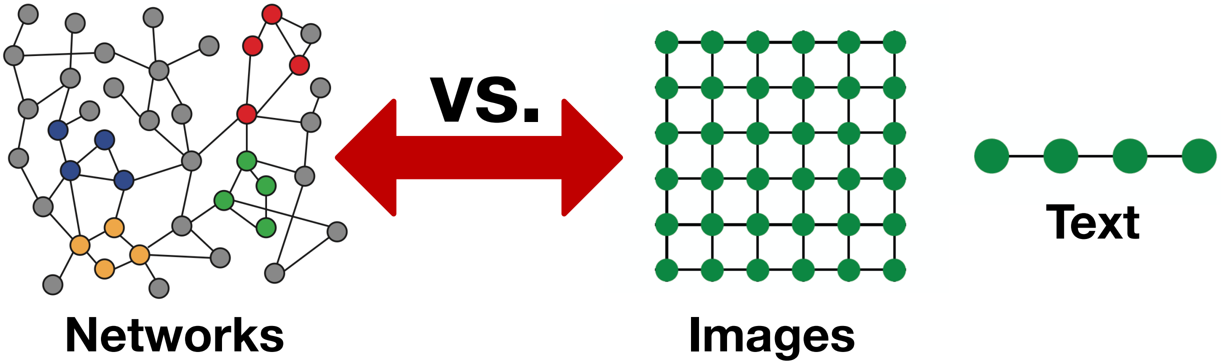
Why Is It Hard?

- Modern deep learning toolbox is designed for grids or simple sequences
  - Text and sequences have linear 1D structure
  - Can define sliding window, RNNs, word2vec, etc.



Why Is It Hard?

- But networks are far more complex!
  - Arbitrary size and complex topological structure (i.e., no spatial locality like grids)



  - No fixed node ordering or reference point
  - Often dynamic and have multimodal features

This Tutorial

1) Node embeddings

- Map nodes to low-dimensional embeddings
- *Applications:* PPIs, Disease pathways

2) Graph neural networks

- Deep learning approaches for graphs
- *Applications:* Gene functions

3) Heterogeneous networks

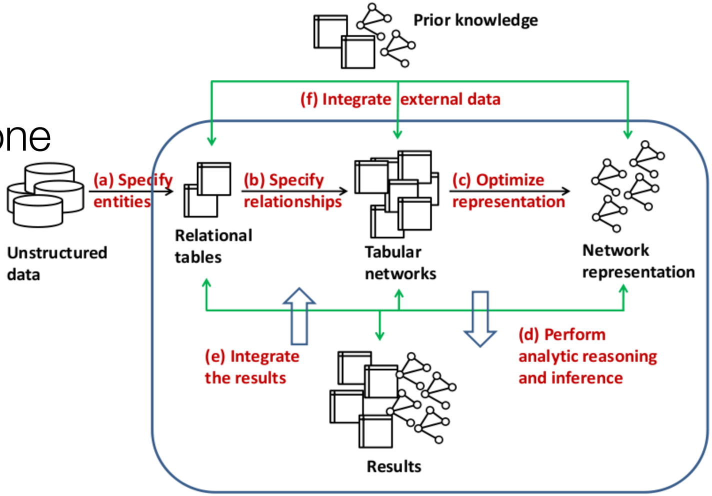
- Embedding heterogeneous networks
- *Applications:* Human tissues, Drug side effects

Tutorial Resources

- Network **analytics tools** in SNAP
- **Network data:**
  - snap.stanford.edu/projects.html:
    - CRank, Decagon, MAMBO, NE, OhmNet, Pathways, and many others
- Deep learning **code bases:**
  - End-to-end examples in Tensorflow/PyTorch
  - Popular code bases for graph neural nets
  - Easy to adapt and extend for your application

Network Analytics in **SNAP**

- **S**tanford **N**etwork **A**nalysis **P**latform (SNAP) is our general purpose, high-performance system for analysis and manipulation of large networks
    - http://snap.stanford.edu
    - Scales to massive networks with hundreds of millions of nodes and billions of edges
- **SNAP software**: C++, Python
- **Software requirements**: none


Prior knowledge
(f) Integrate external data
Unstructured data
(a) Specify entities
Relational tables
(b) Specify relationships
Tabular networks
(c) Optimize representation
Network representation
(e) Integrate the results
Results
(d) Perform analytic reasoning and inference

COMING SOON

BioSNAP: Network Data

Biomedical network dataset collection:

- Different types of biomedical networks
- Ready to use for:
  - Algorithm benchmarking
  - Method development
  - Knowledge discovery
- Easy to link entities across datasets

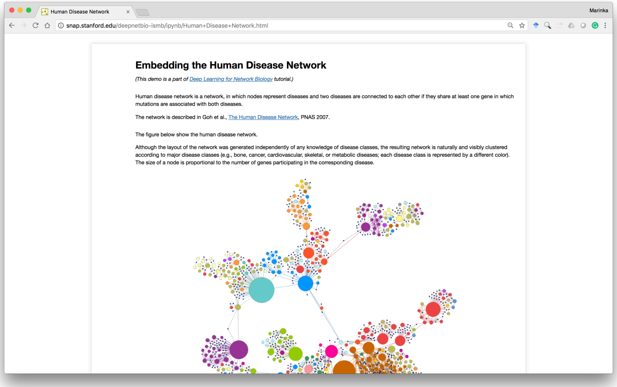
Total: 250M entities, 2.2TB raw network data

Dataset	**#Items**	**Raw Size**
DisGeNet	30K	10MB
STRING	10M	1TB
OMIM	25K	100MB
CTD	55K	1.2GB
HPRD	30K	30MB
BioGRID	64K	100MB
DrugBank	7K	60MB
Disease Ontology	10K	5MB
Protein Ontology	200K	130MB
Mesh Hierarchy	30K	40MB
PubChem	90M	1GB
DGIdb	5K	30MB
Gene Ontology	45K	10MB
MSigDB	14K	70MB
Reactome	20K	100MB
GEO	1.7M	80GB
ICGC (66 cancer projects)	40M	1TB
GTEx	50M	100GB
Many more…		

Deep Learning Code Bases

This tutorial: Using graph neural networks:

- End-to-end examples in Tensorflow/PyTorch
- Popular code bases for graph neural nets
- Easy to adapt and extend for your application


Human Disease Network
Marinka
snap.stanford.edu/deepnetbio-ismb/ipynb/Human+Disease+Network.html

Embedding the Human Disease Network

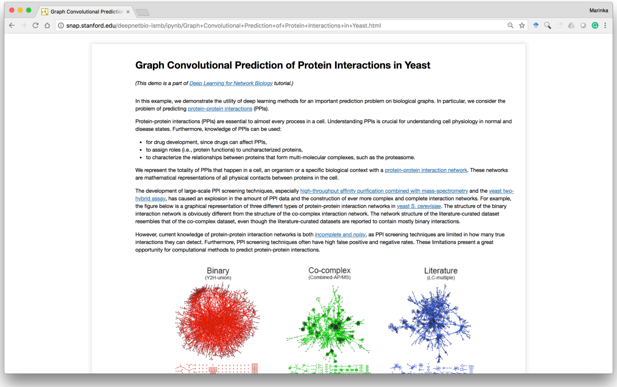
(This demo is a part of Deep Learning for Network Biology tutorial.)

Human disease network is a network, in which nodes represent diseases and two diseases are connected to each other if they share at least one gene in which mutations are associated with both diseases.

The network is described in Goh et al., The Human Disease Network, PNAS 2007.

The figure below show the human disease network.

Although the layout of the network was generated independently of any knowledge of disease classes, the resulting network is naturally and visibly clustered according to major disease classes (e.g., bone, cancer, cardiovascular, skeletal, or metabolic diseases; each disease class is represented by a different color). The size of a node is proportional to the number of genes participating in the corresponding disease.


Graph Convolutional Predictio
Marinka
snap.stanford.edu/deepnetbio-ismb/ipynb/Graph+Convolutional+Prediction+of+Protein+Interactions+in+Yeast.html

Graph Convolutional Prediction of Protein Interactions in Yeast

(This demo is a part of Deep Learning for Network Biology tutorial.)

In this example, we demonstrate the utility of deep learning methods for an important prediction problem on biological graphs. In particular, we consider the problem of predicting protein-protein interactions (PPIs).

Protein-protein interactions (PPIs) are essential to almost every process in a cell. Understanding PPIs is crucial for understanding cell physiology in normal and disease states. Furthermore, knowledge of PPIs can be used:

- for drug development, since drugs can affect PPIs,
- to assign roles (i.e., protein functions) to uncharacterized proteins,
- to characterize the relationships between proteins that form multi-molecular complexes, such as the proteasome.

We represent the totality of PPIs that happen in a cell, an organism or a specific biological context with a protein-protein interaction network. These networks are mathematical representations of all physical contacts between proteins in the cell.

The development of large-scale PPI screening techniques, especially high-throughput affinity purification combined with mass-spectrometry and the yeast two-hybrid assay, has caused an explosion in the amount of PPI data and the construction of ever more complex and complete interaction networks. For example, the figure below is a graphical representation of three different types of protein-protein interaction networks in yeast *S. cerevisiae*. The structure of the binary interaction network is obviously different from the structure of the co-complex interaction network. The network structure of the literature-curated dataset resembles that of the co-complex dataset, even though the literature-curated datasets are reported to contain mostly binary interactions.

However, current knowledge of protein-protein interaction networks is both *incomplete and noisy*, as PPI screening techniques are limited in how many true interactions they can detect. Furthermore, PPI screening techniques often have high false positive and negative rates. These limitations present a great opportunity for computational methods to predict protein-protein interactions.

PhD Students

Claire Donnat

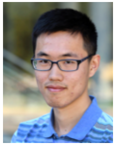
Mitchell Gordon

David Hallac

Emma Pierson

Geet Sethi

Himabindu Lakkaraju


Rex Ying

Tim Althoff

Will Hamilton

Alex Porter

Post-Doctoral Fellows

Baharan Mirzasoleiman

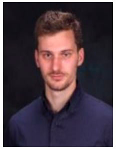
Marinka Zitnik

Michele Catasta

Srijan Kumar

Stephen Bach

Research Staff


Adrijan Bradaschia

Rok Sosic

Industry Partnerships

Ford
azumio
facebook
twitter
Pinterest
UNDER ARMOUR
LinkedIn
HYUNDAI
BOEING
WIKIPEDIA The Free Encyclopedia
NVIDIA
amazon.com
NTT docomo
spinn3r
HUAWEI
SAP
HITACHI
BOSCH Invented for life
CRISIS TEXT LINE |

Funding

MURI ARO
DARPA
NSF
NIH
CHAN ZUCKERBERG INITIATIVE
IARPA

Collaborators

Stanford | Stanford Data Science Initiative

Dan Jurafsky, Linguistics, Stanford University
Christian Danescu-Miculescu-Mizil, Information Science, Cornell University
Stephen Boyd, Electrical Engineering, Stanford University
David Gleich, Computer Science, Purdue University
VS Subrahmanian, Computer Science, University of Maryland
Sarah Kunz, Medicine, Harvard University
Russ Altman, Medicine, Stanford University
Jochen Profit, Medicine, Stanford University
Eric Horvitz, Microsoft Research
Jon Kleinberg, Computer Science, Cornell University
Sendhill Mullainathan, Economics, Harvard University
Scott Delp, Bioengineering, Stanford University
Jens Ludwig, Harris Public Policy, University of Chicago

STANFORD INFOLAB

WE’RE HIRING!

Many interesting high-impact projects
in Machine Learning and Large Biomedical Data

Applications: Precision Medicine & Health, Drug Repurposing, Drug Side Effect modeling, Network Biology, and many more