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

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

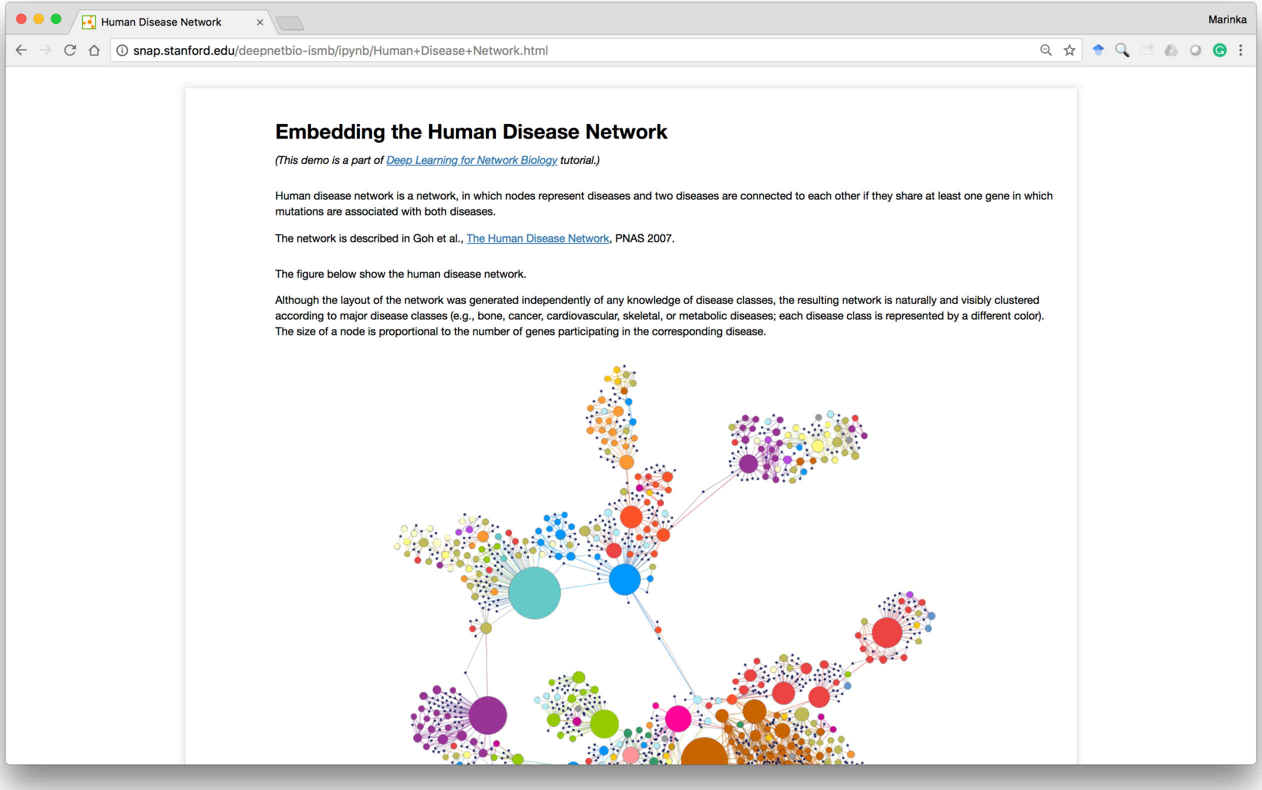
Outline of This Section

1. Practical advice and demos

2. Future directions & conclusion

Practical Advice and Demos

Demo: Diseases


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.)

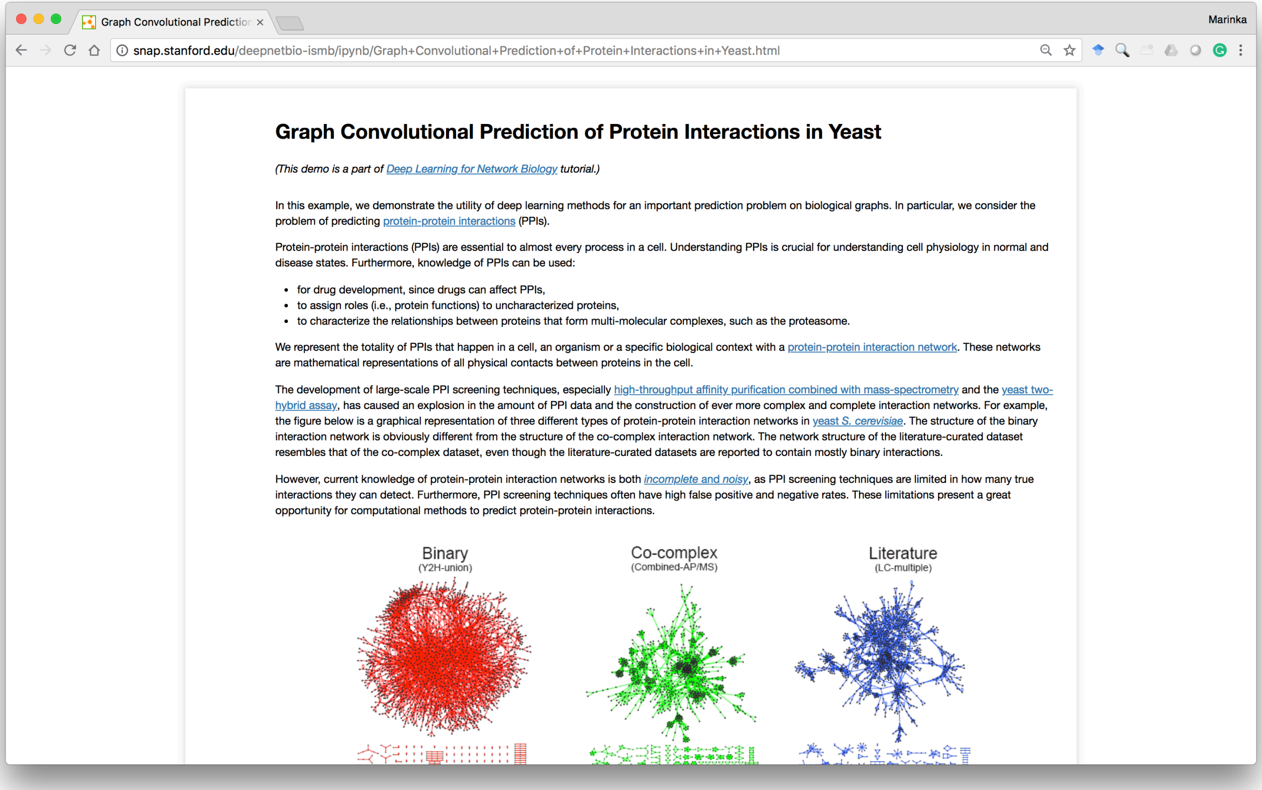
Human disease network is a network, in which nodes represent diseases and two diseases are connected to each other if they share at least one gene in which mutations are associated with both diseases.

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.

Demo: Protein Interactions


Graph Convolutional Predictior ×
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.

General Tips

1) Network data preprocessing is important:
   - renormalization tricks
   - variance-scaled initialization
   - network data whitening
2) Use the ADAM optimizer:
   - ADAM naturally takes care of decaying the learning rate
3) ReLU (activation function) often works really well
4) No activation function at your output layer:
   - Easy mistake if you build layers with a shared function
5) Include bias term in every layer
6) Graph convolution layer of size 64 or 128 is plenty

Debugging Deep Networks

- Debug?!:
  - Loss/accuracy not converging during training
- Important for model development:
  - **Overfit on training data:**
    - Accuracy should be essentially 100% or error close to 0
    - If neural network cannot overfit a single data point, something is wrong
  - **Scrutinize your loss function!**
  - **Scrutinize your visualizations!**

Future Directions and Opportunities

Material based on:

- Zitnik et al. 2018. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities.
- Camacho et al. 2018 Next-Generation Machine Learning for Biological Networks. *Cell.*

Learning Hierarchies

Hierarchical structures are ubiquitous in network biology

Challenges:

- How to infer hierarchies from pairwise similarity scores?
- How to learn continuous representations of hierarchies?
- How to exploit the properties of networks' hyperbolic geometry?

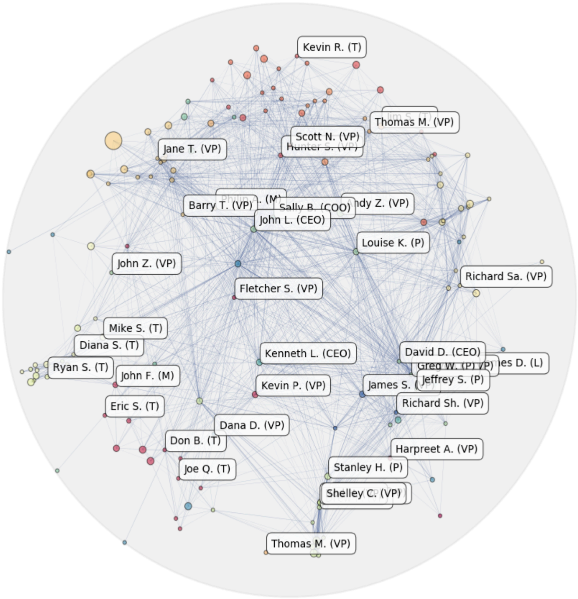
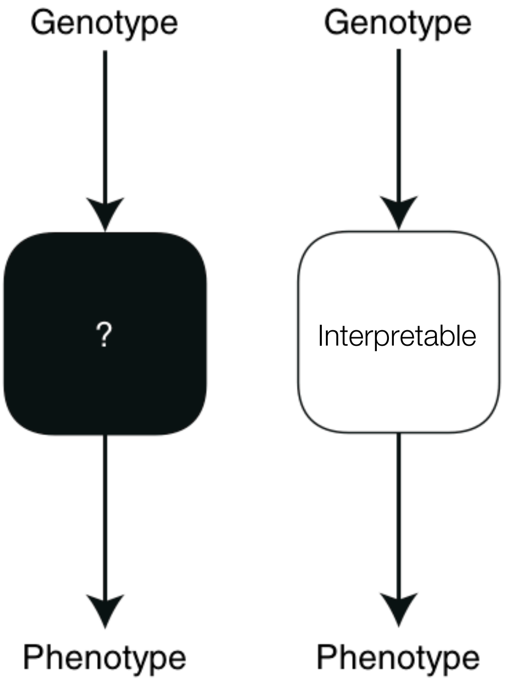


Image from: Nickel et al. 2018. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. *ICML.*

Explainability



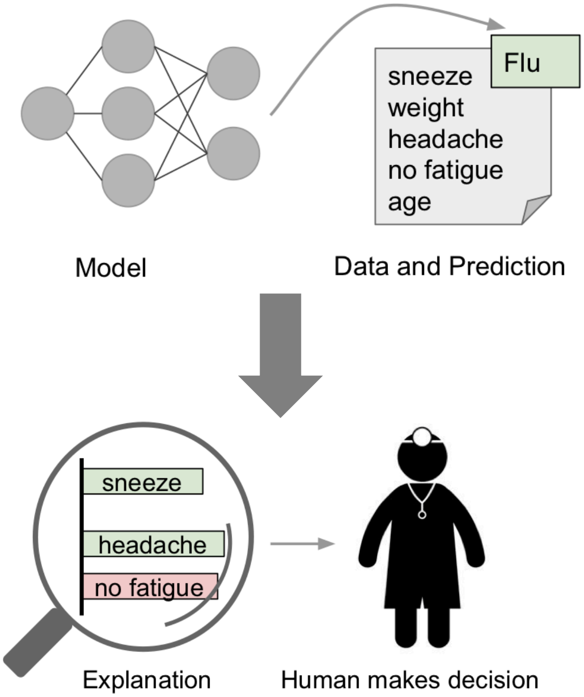
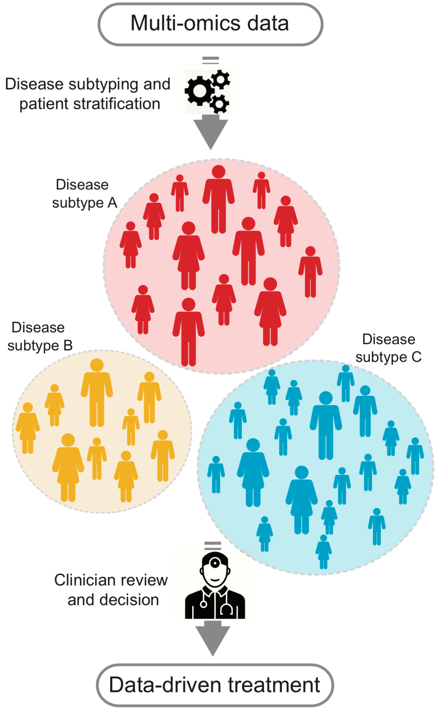
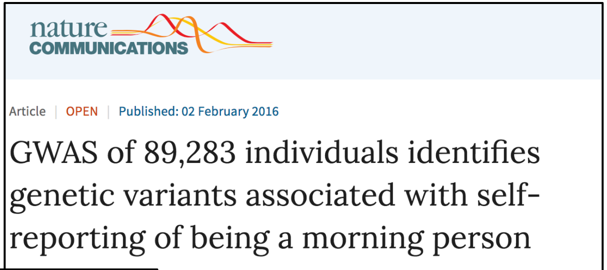


Image from: Ma et al. 2018. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods.* Ribeiro et al. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *KDD*.

Internet-Based Phenotyping

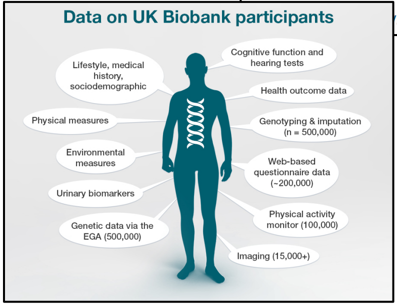



nature COMMUNICATIONS

Article | OPEN | Published: 02 February 2016

GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person

gelska, David Tran, Nicholas Eriksson, Joyce Y. Tung & David A. Hinds


Data on UK Biobank participants
Lifestyle, medical history, sociodemographic
Physical measures
Environmental measures
Urinary biomarkers
Genetic data via the EGA (500,000)
Cognitive function and hearing tests
Health outcome data
Genotyping & imputation (n = 500,000)
Web-based questionnaire data (~200,000)
Physical activity monitor (100,000)
Imaging (15,000+)

Self-reported & ecological data

UK Biobank: A prospective cohort of 500 K people to support the investigation of risk factors for major diseases of middle and old age

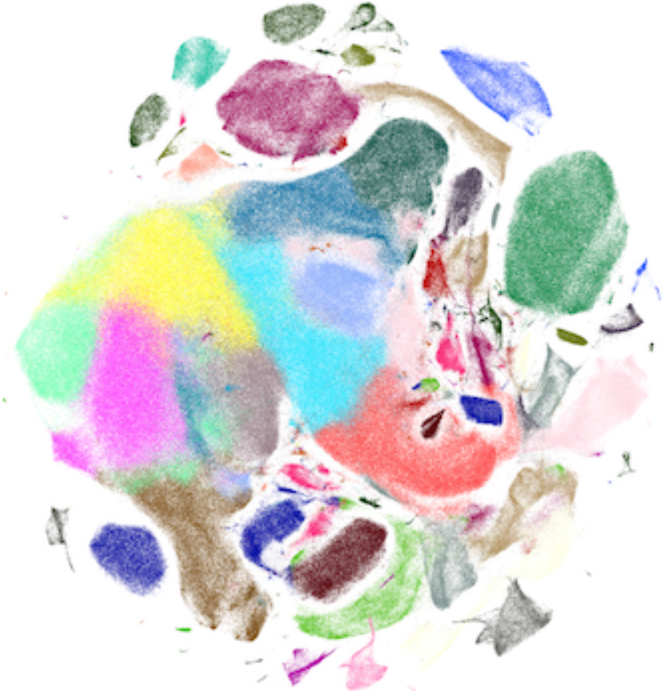
Image from: Zitnik et al. 2018. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities.

Giga-Scale Network Data

Goal: Handle massive graphs

Challenge: Existing methods do not scale to new high-throughput datasets

Idea: Use graph neural networks with efficient batch optimization and parameter sharing



E.g., **The Human Cell Atlas:** cell-cell similarity networks

Image from: Wolf et al. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*.

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

Deep Learning for Network Biology

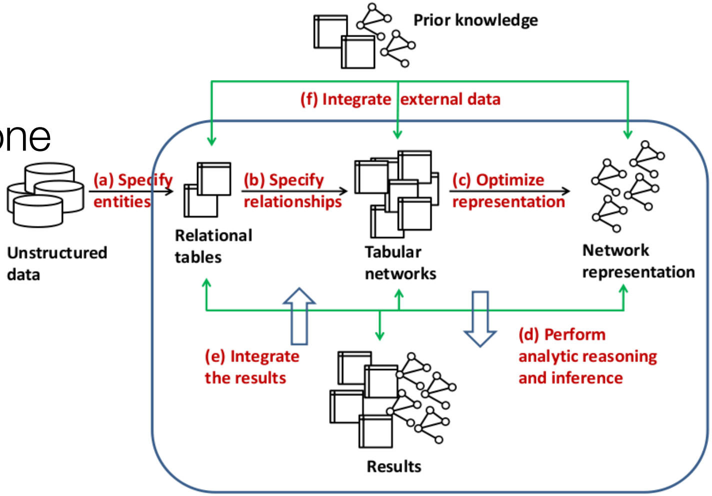
How to Start?

Tutorial Resources

- Network **analytics tools** in SNAP
- 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 data:**
  - snap.stanford.edu/projects.html:
    - CRank, Decagon, MAMBO, NE, OhmNet, Pathways, and many others

Network Analytics with **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
(d) Perform analytic reasoning and inference
Results

BioSNAP: Network Data

COMING SOON

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

- Node2vec:
  - https://github.com/aditya-grover/node2vec (Python)
  - https://github.com/snap-stanford/snap/tree/master/examples/node2vec (C++)
- Graph Convolutional Networks (GCNs):
  - https://github.com/tkipf/gcn (Tensorflow)
  - https://github.com/tkipf/pygcn (PyTorch)
  - https://github.com/tkipf/keras-gcn (Keras)
- GraphSAGE:
  - https://github.com/williamleif/GraphSAGE (Tensorflow)
  - https://github.com/williamleif/graphsage-simple (Pytorch)
- Metapath2vec and metapath2vec++ (Python):
  - https://ericdongyx.github.io/metapath2vec/m2v.html
- OhmNet (Python):
  - https://github.com/marinkaz/ohmnet
- Decagon (Tensorflow):
  - https://github.com/marinkaz/decagon

Deep Learning for Network Biology

Next-Generation Machine Learning for Networks in Biology and Medicine

PhD Students

Claire Donnat

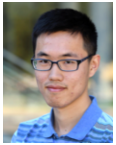
Mitchell Gordon

David Hallac

Emma Pierson

Geet Sethi

Himabindu Lakkaraju


Rex Ying

Tim Althoff

Will Hamilton

Alex Porter

Post-Doctoral Fellows

Baharan Mirzasoleiman

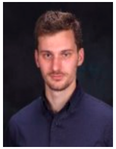
Marinka Zitnik

Michele Catasta

Srijan Kumar

Stephen Bach

Research Staff


Adrijan Bradaschia

Rok Sosic

Industry Partnerships

Ford
azumio
UNDER ARMOUR
facebook
twitter
Pinterest
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
Stanford | Stanford Data Science Initiative

Collaborators

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