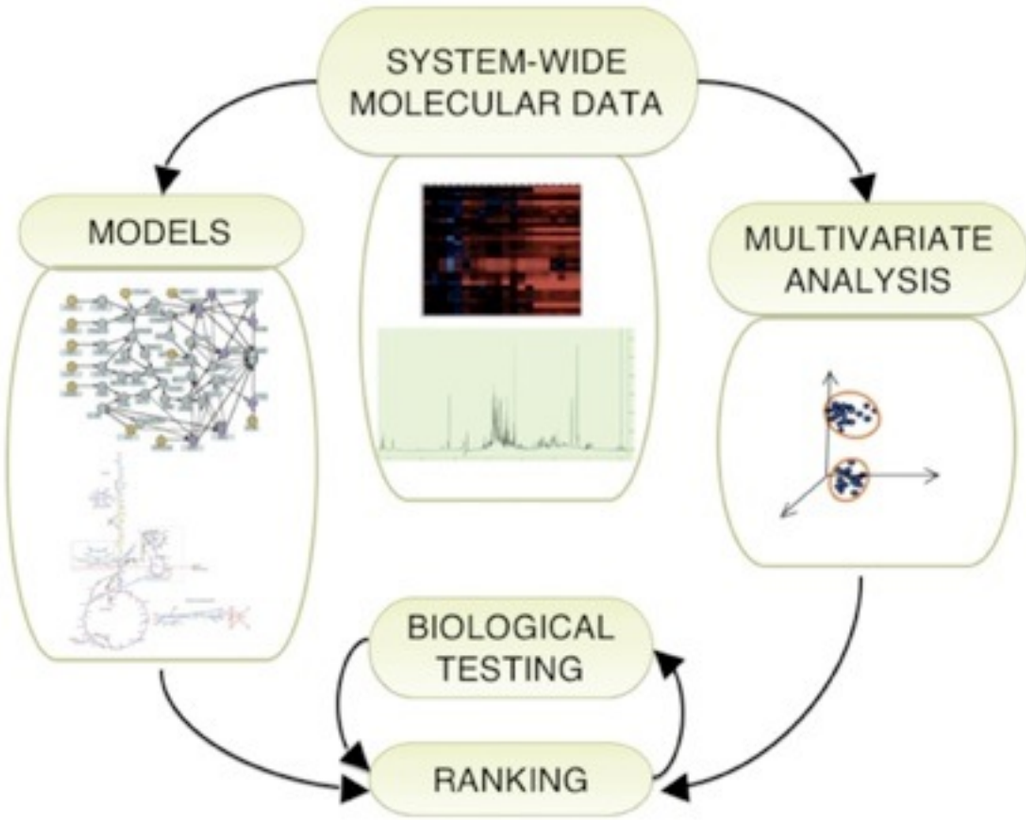
Informatics in Computational Medicine

Inderjit S. Dhillon
Director, Center for Big Data Analytics
UT Austin
Computer Science, Mathematics & ICES

ICES Computational Medicine Day
May 13, 2014

Introduction

- Computational Medicine — Quantitative approach to understanding, detecting and treating diseases

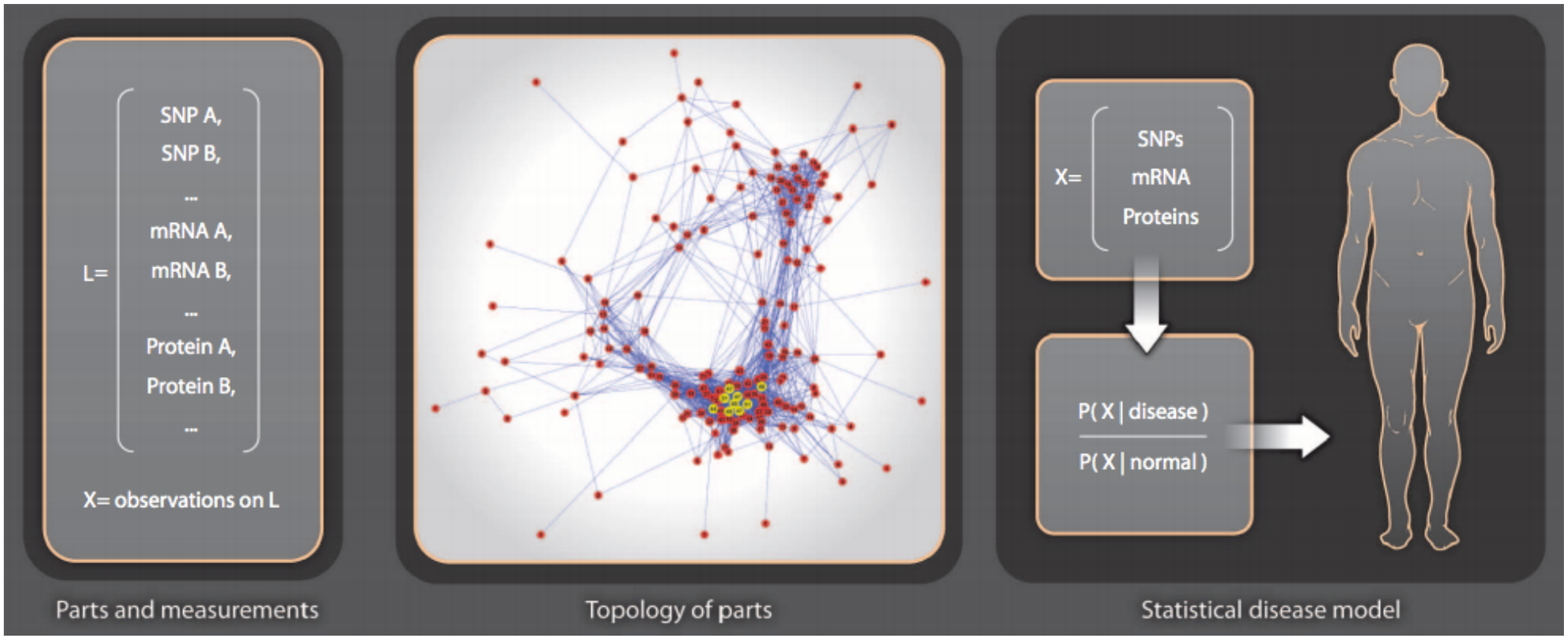


- Need to understand and control biological processes involved in human diseases via systemic measurements and computational analyses

Calzolari D, Bruschi S, Coquin L, Schofield J, Feala JD, et al. (2008) Search Algorithms as a Framework for the Optimization of Drug Combinations. PLoS Comput Biol 4(12): e1000249

Medical Genetics

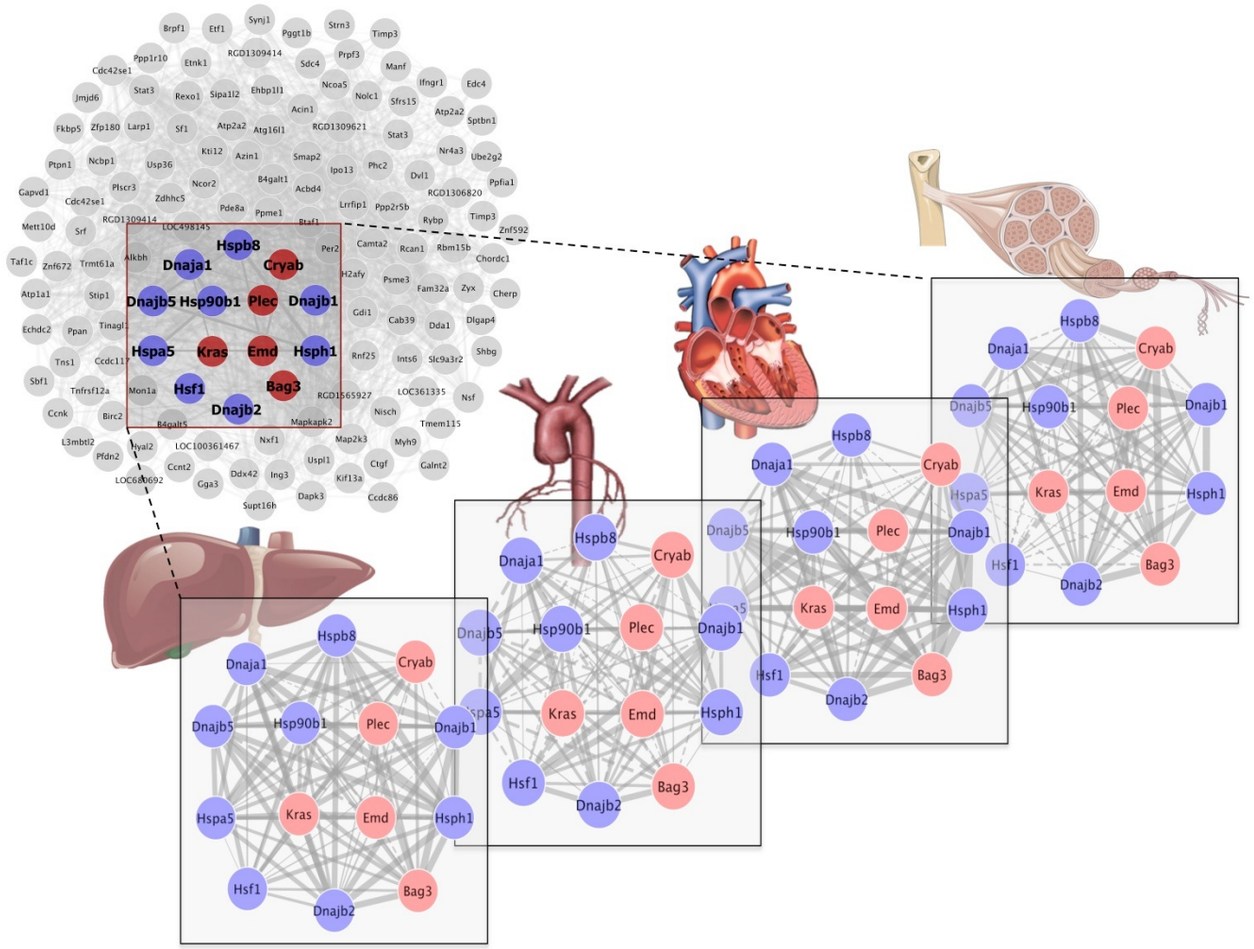
- Analysis begins at the level of genes — study of human genetics & its application to medical care
- Eventual goals: Gene therapy for cure — personalized medicine



Winslow, Raimond L., et al. "Computational medicine: translating models to clinical care." Science Translational Medicine 4.158 (2012)

Gene Networks

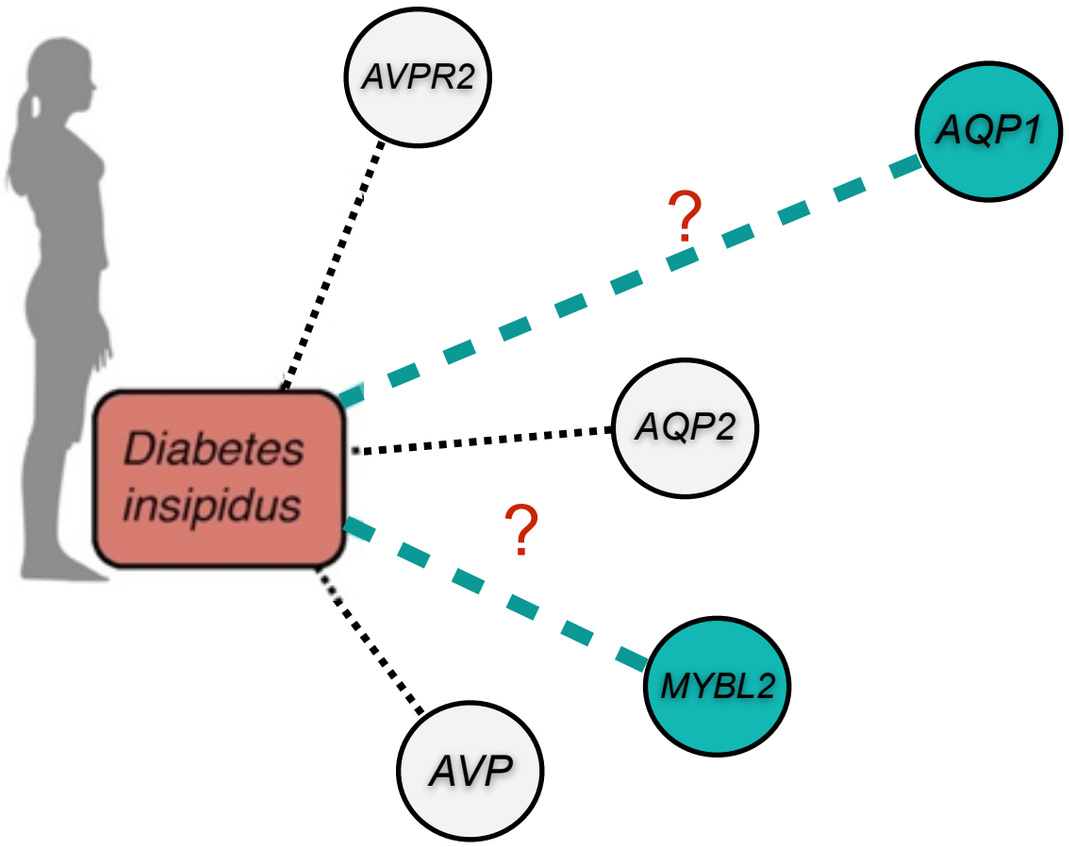
- Multiple genes collectively influence the likelihood of developing many common and complex diseases.



- Important to understand how genes interact with each other (co-expression gene networks)

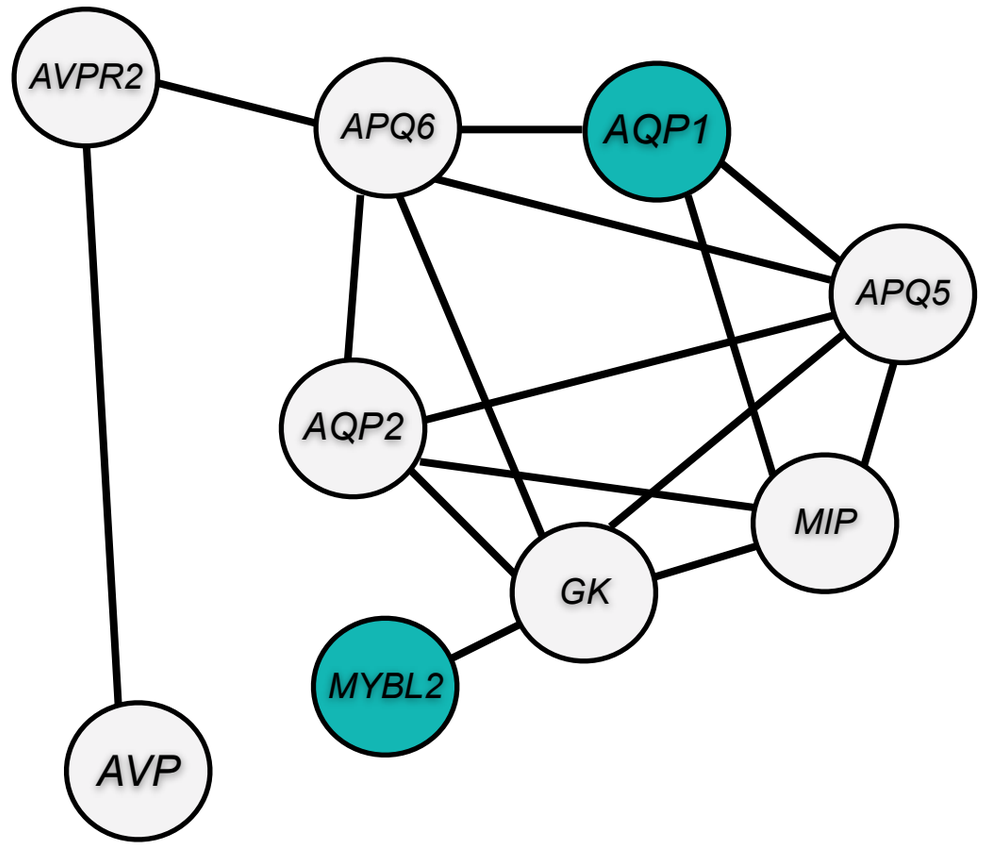
Xiao, X. et al., 2014. Multi-tissue analysis of co-expression networks by Higher-Order generalized singular value decomposition identifies functionally coherent transcriptional modules. PLoS Genetics 10 (1), e1004006+.

Predicting gene-disease links



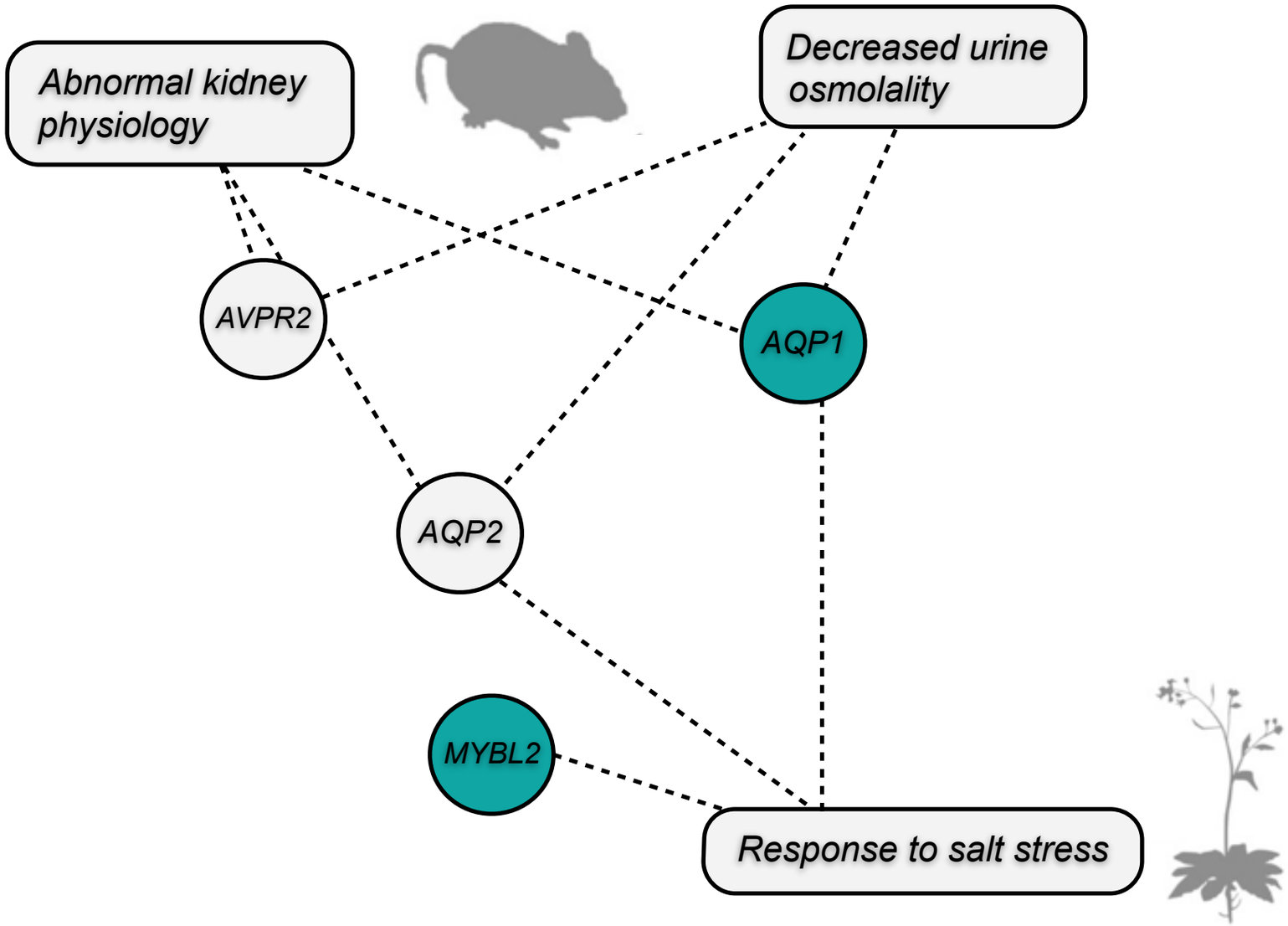
- Goal: Discover human gene-disease associations
- Biologists prefer a short list of potentially relevant genes for further studies

Gene Network



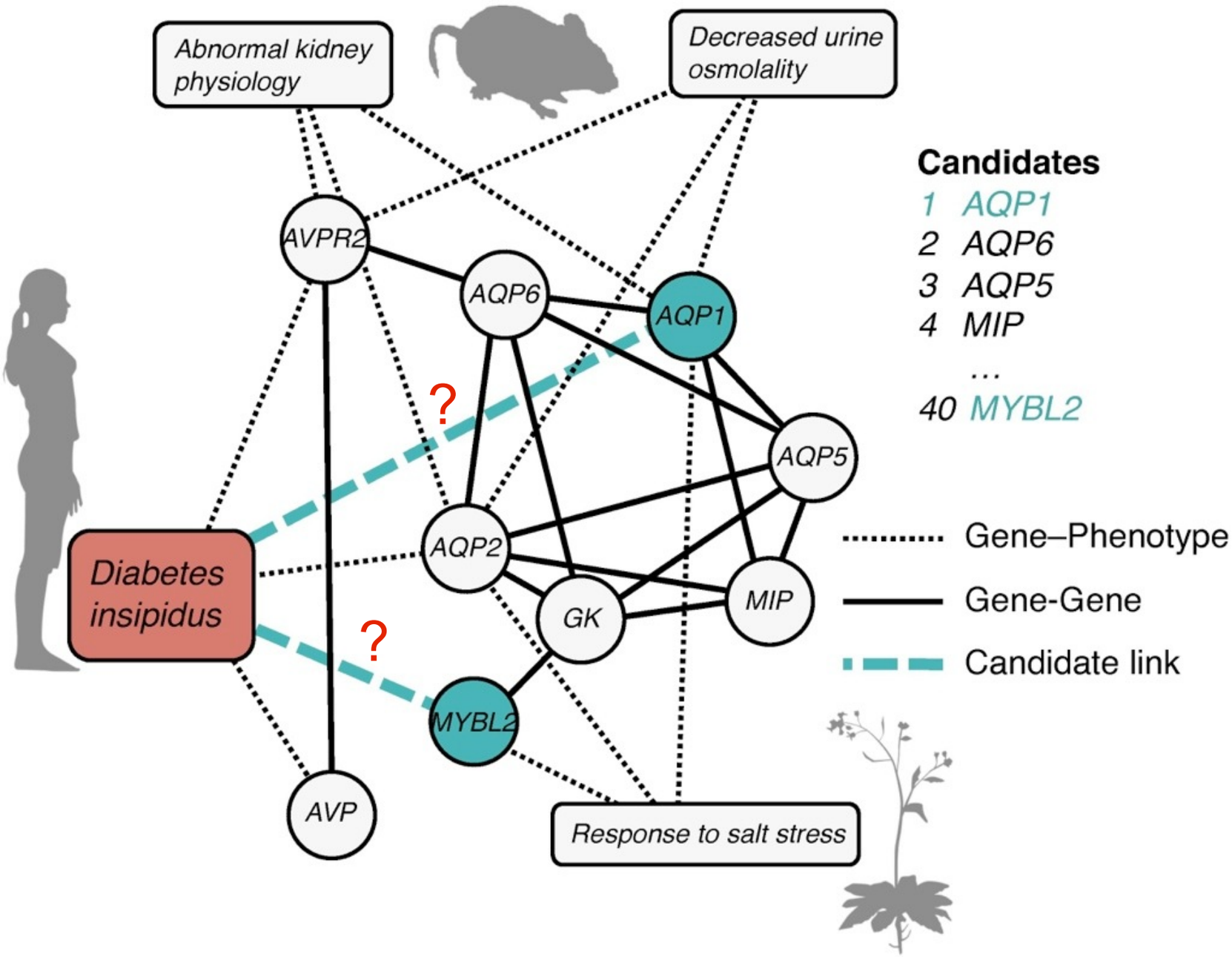
- Functional interactions between genes (e.g. HumanNet)

Gene-Phenotype Networks

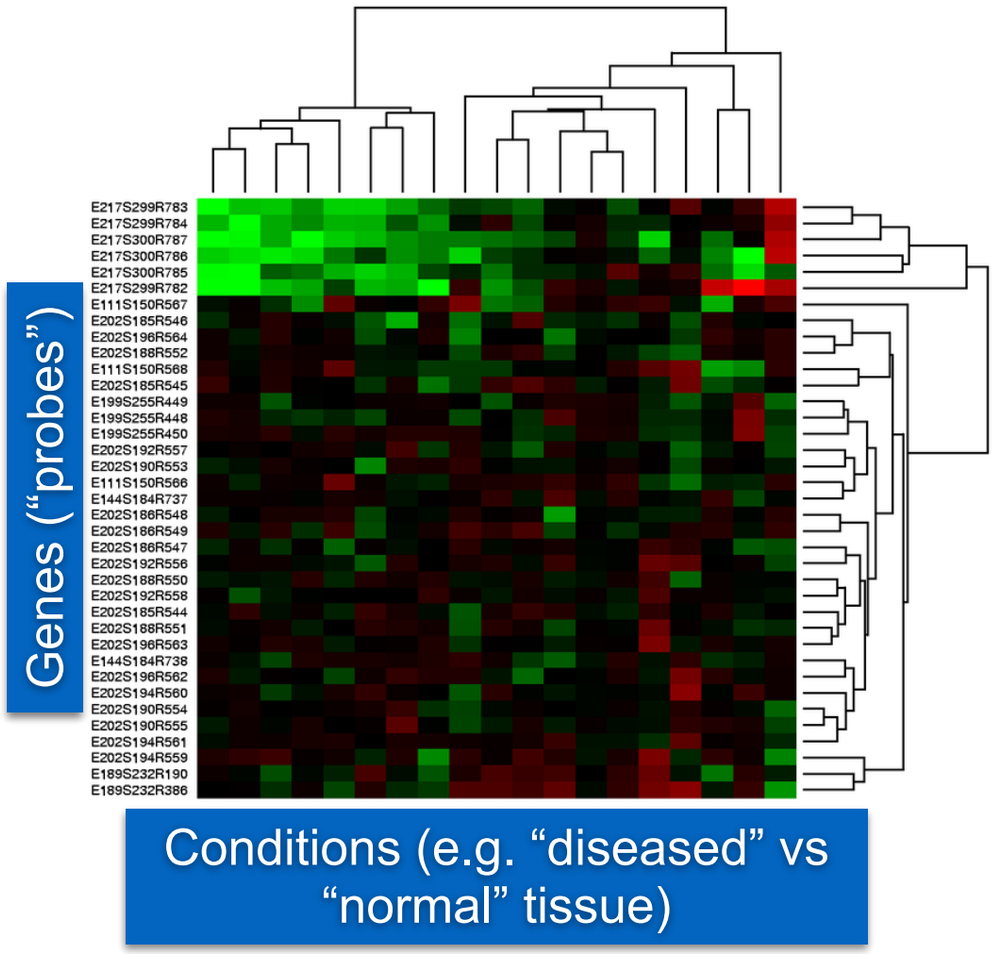


- "Orthologous" phenotypes in other model species can shed light on gene functions and in turn disease-causing genes

The Prediction Problem



Other data sources: Genomic data

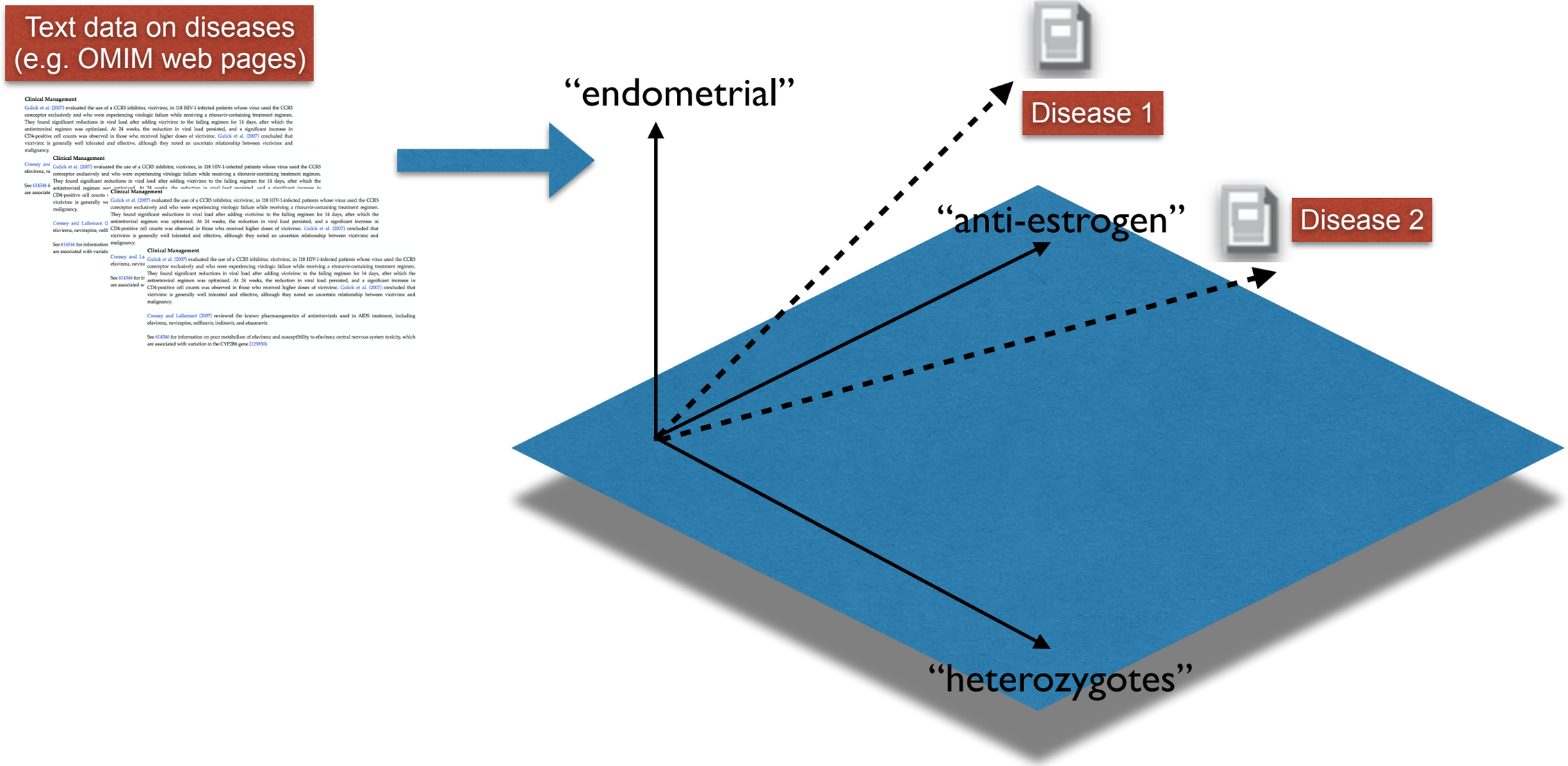


- Abundant expression data available (sources such as BioGPS)
- Co-expression reveals gene function modules ("Eigengenes")

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. BMC systems biology, 1(1), 54.

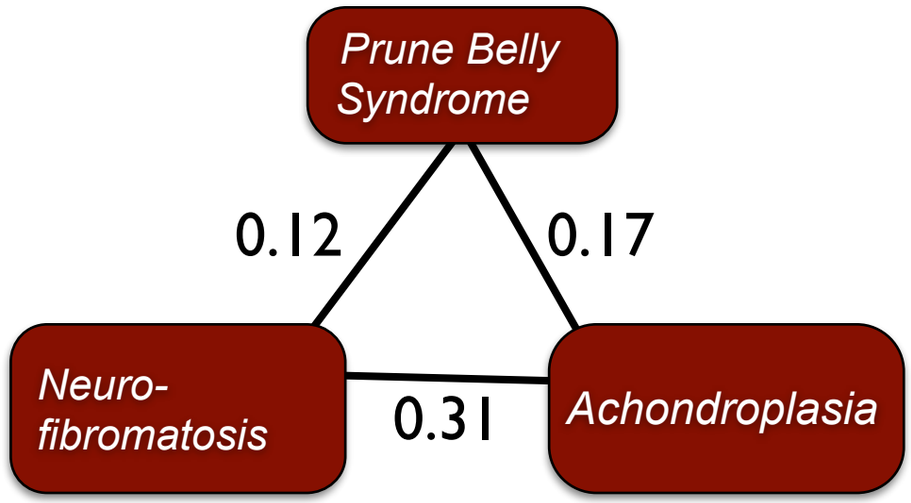
Other data sources: Text data

- Descriptions of diagnosis, clinical features and management in text articles provide information on diseases
- Represent diseases by document term frequencies



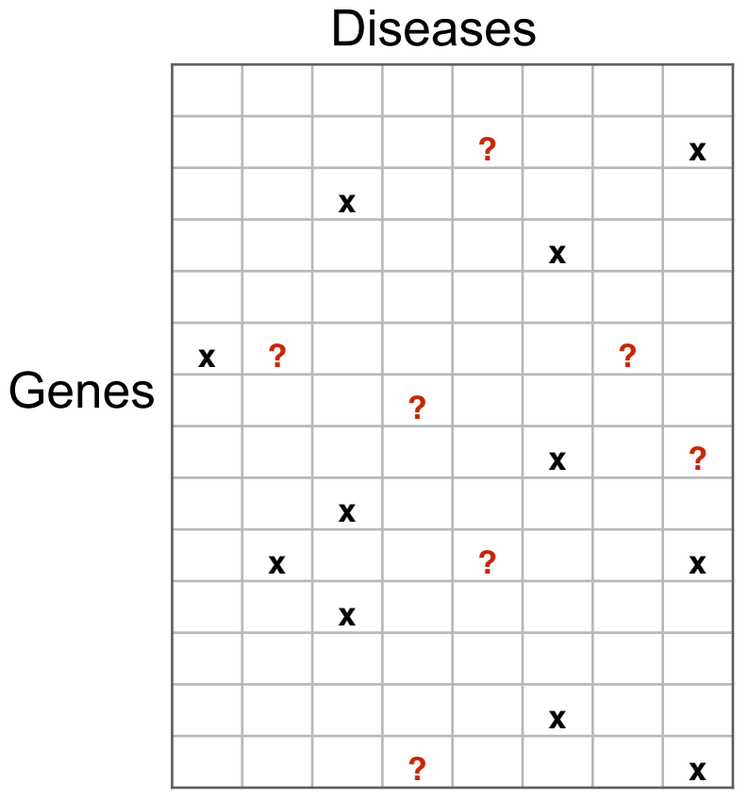
Other data sources: Co-morbidity

- Similarity network between diseases



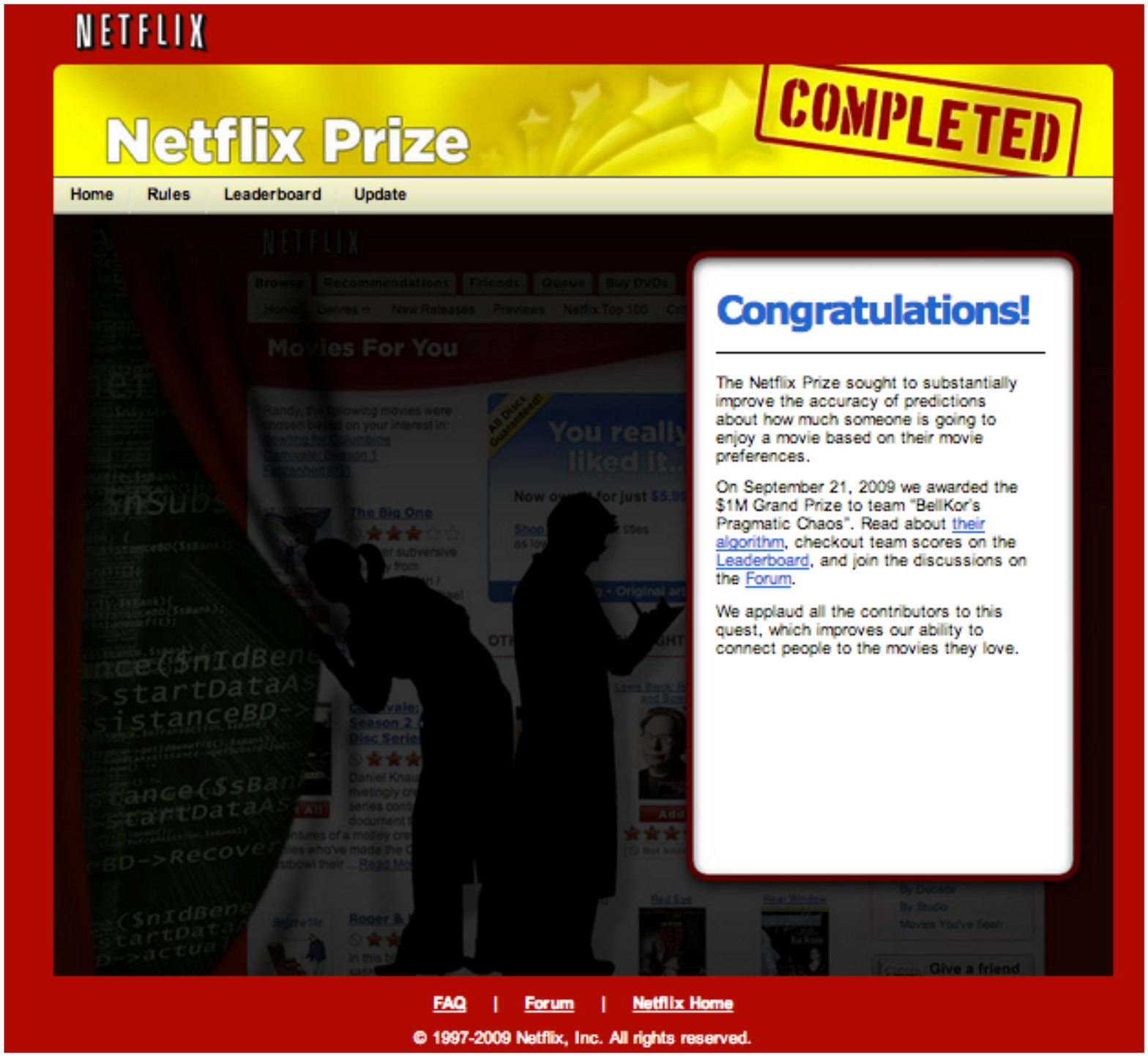
- Similarities can be computed from data available on diseases (such as text, symptoms, drug responses, etc)

Problem Formulation

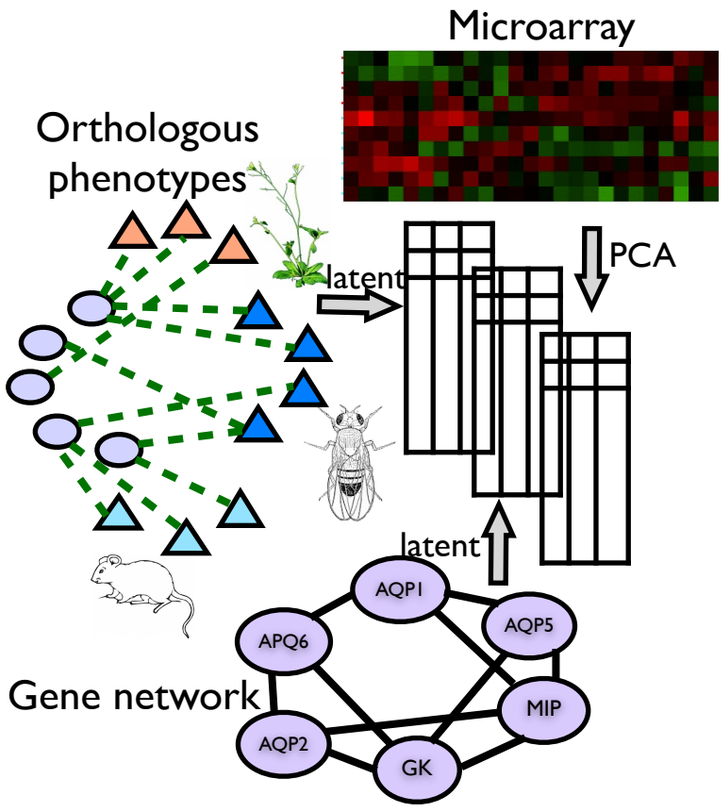


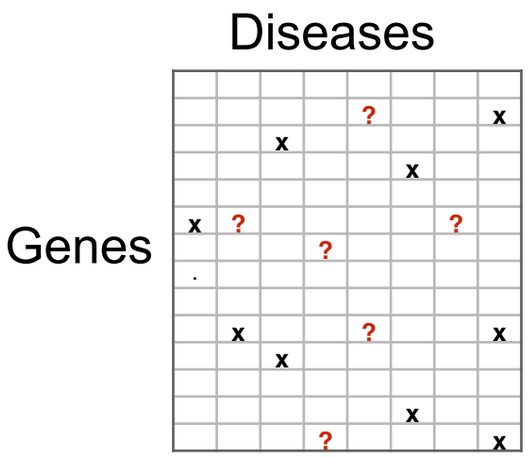
- We want to predict “missing” associations denoted by ?

The Netflix Problem

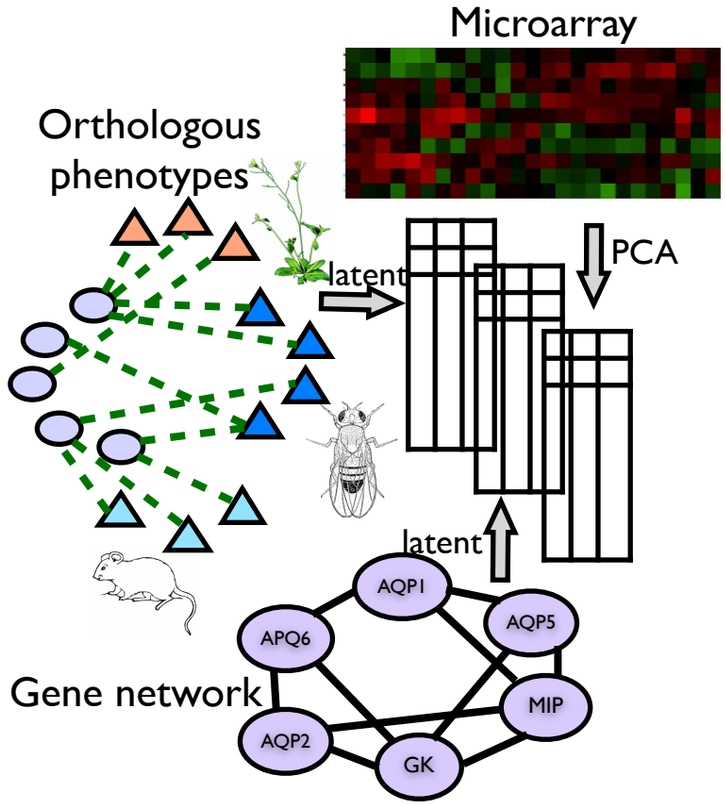


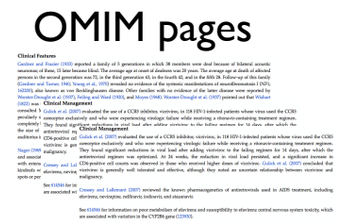
Inductive Matrix Completion

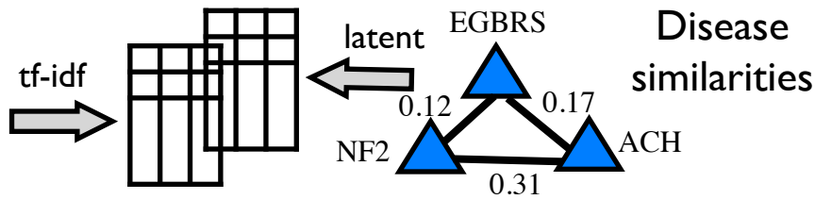


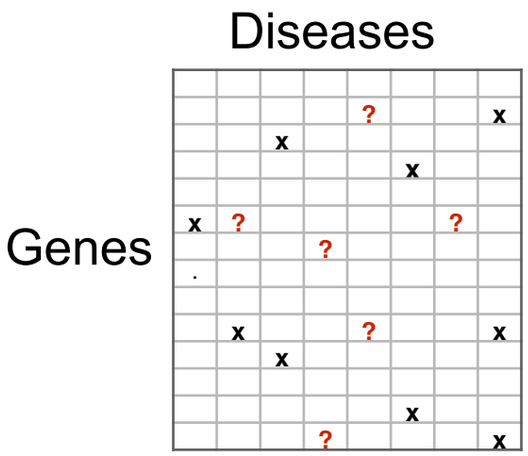


Inductive Matrix Completion

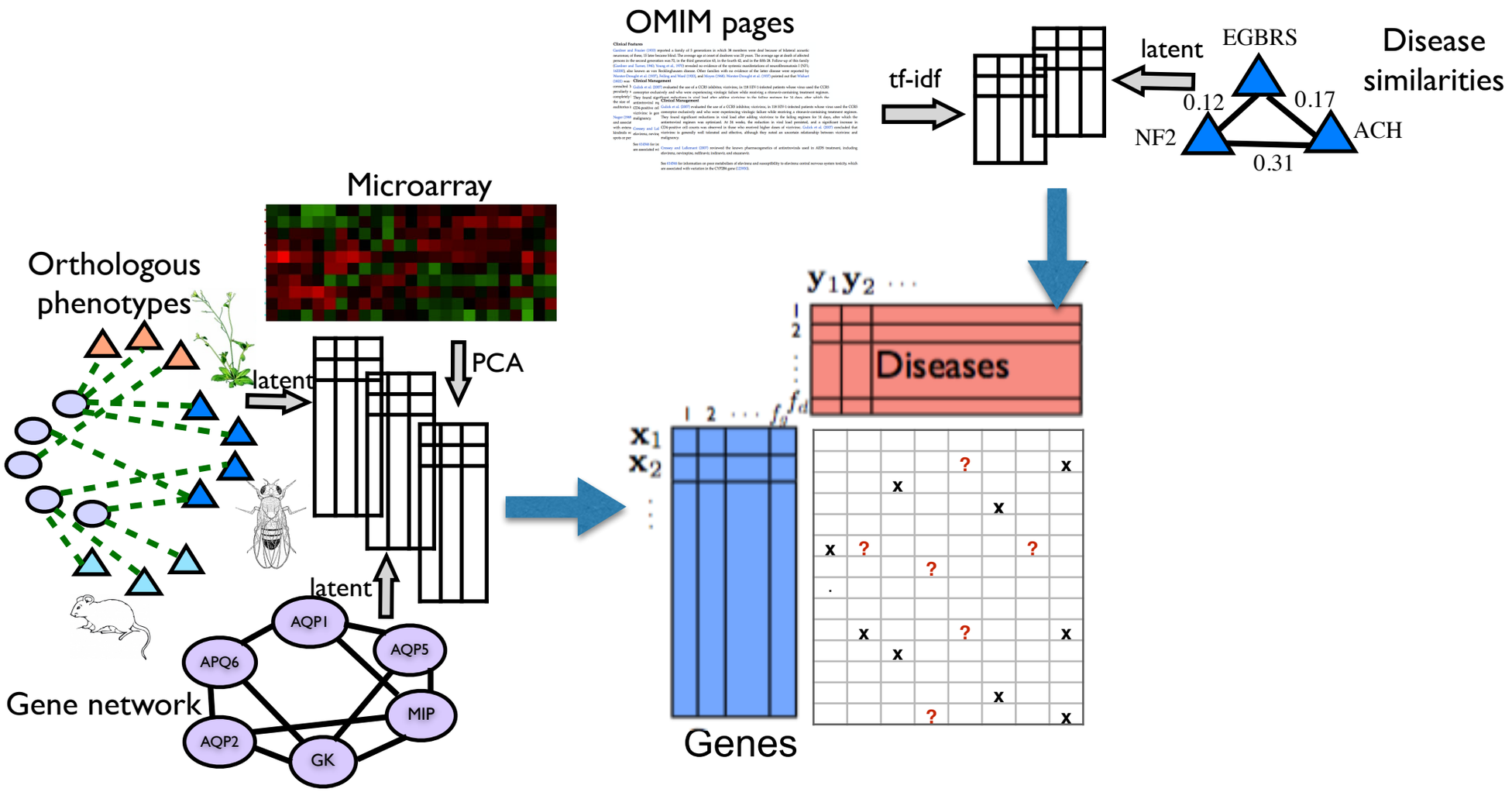




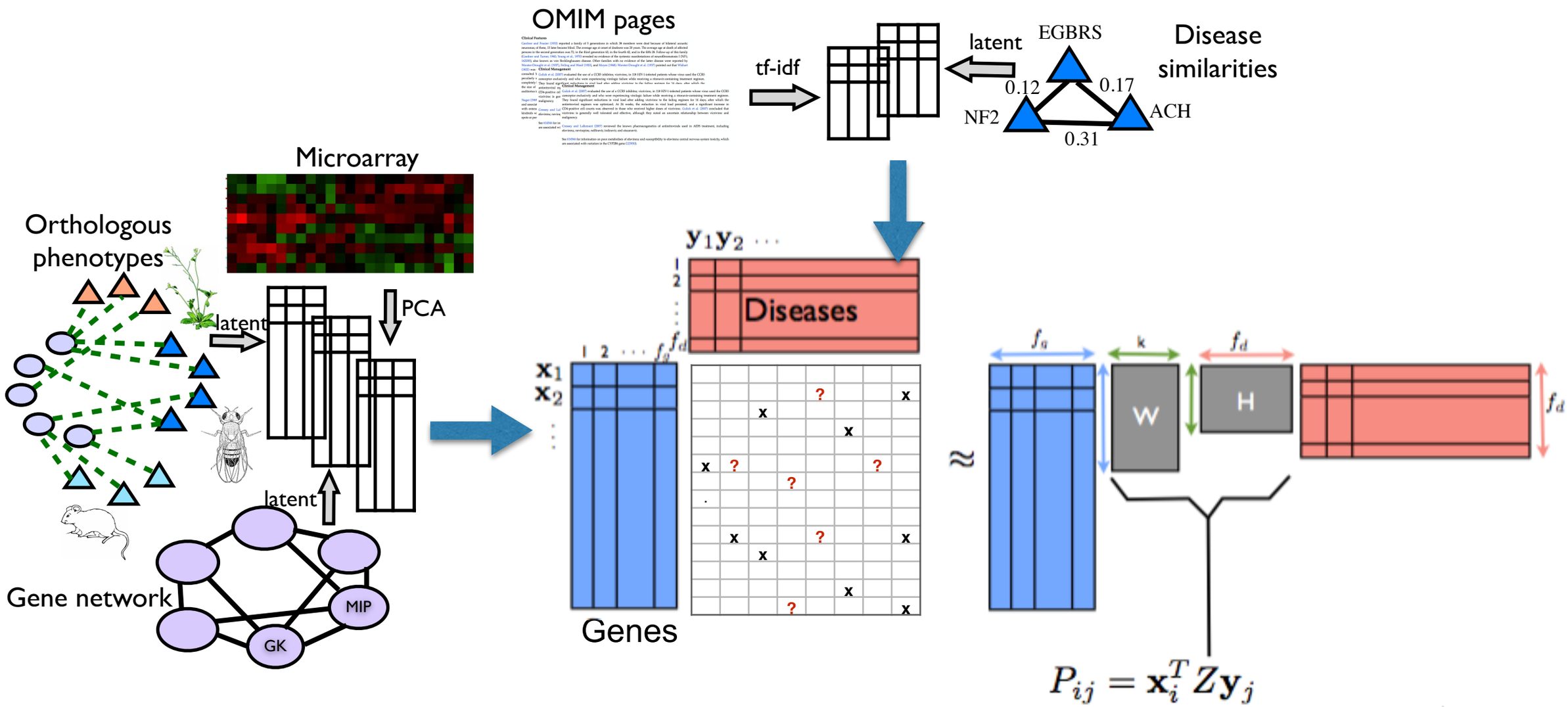




Inductive Matrix Completion

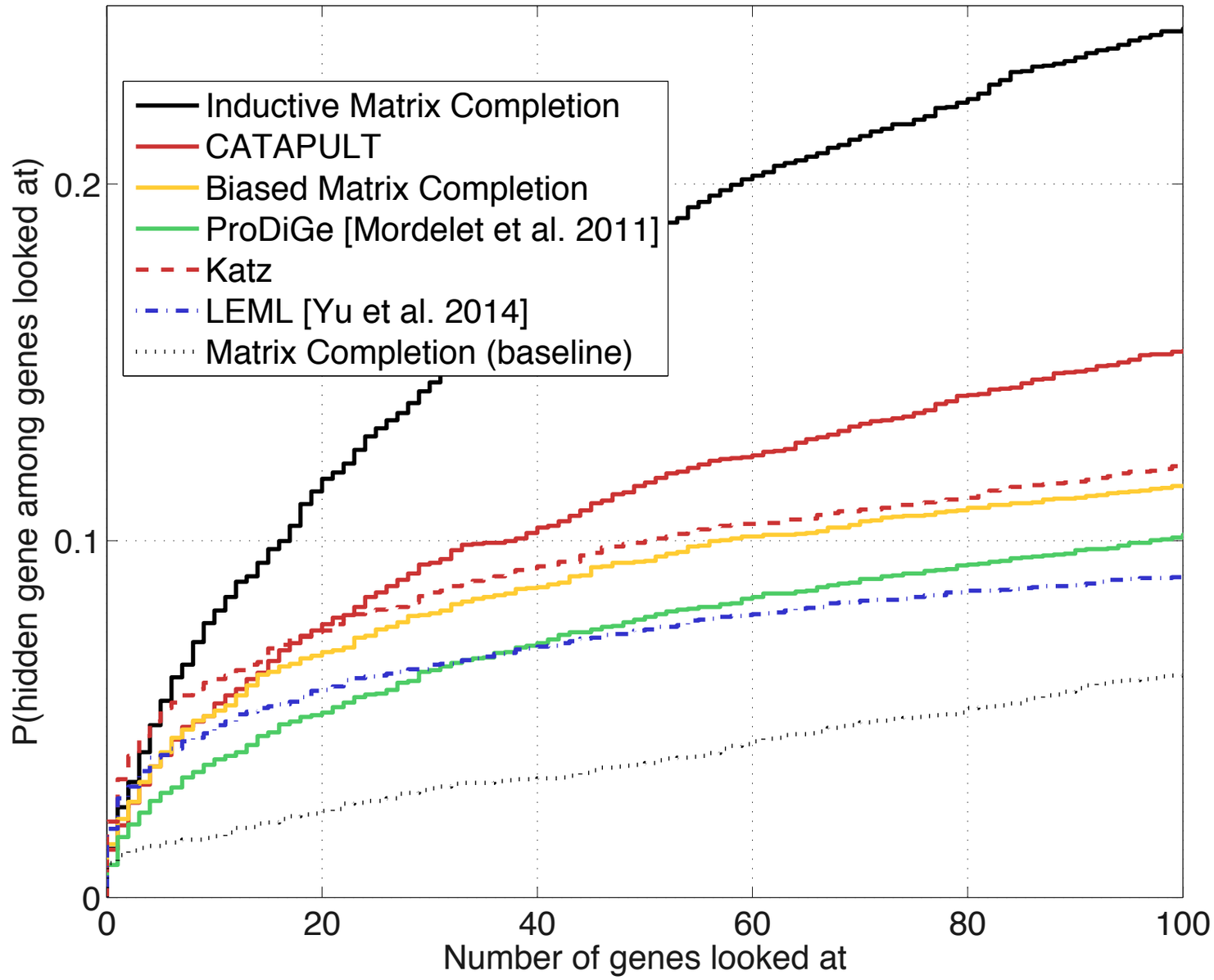


Inductive Matrix Completion



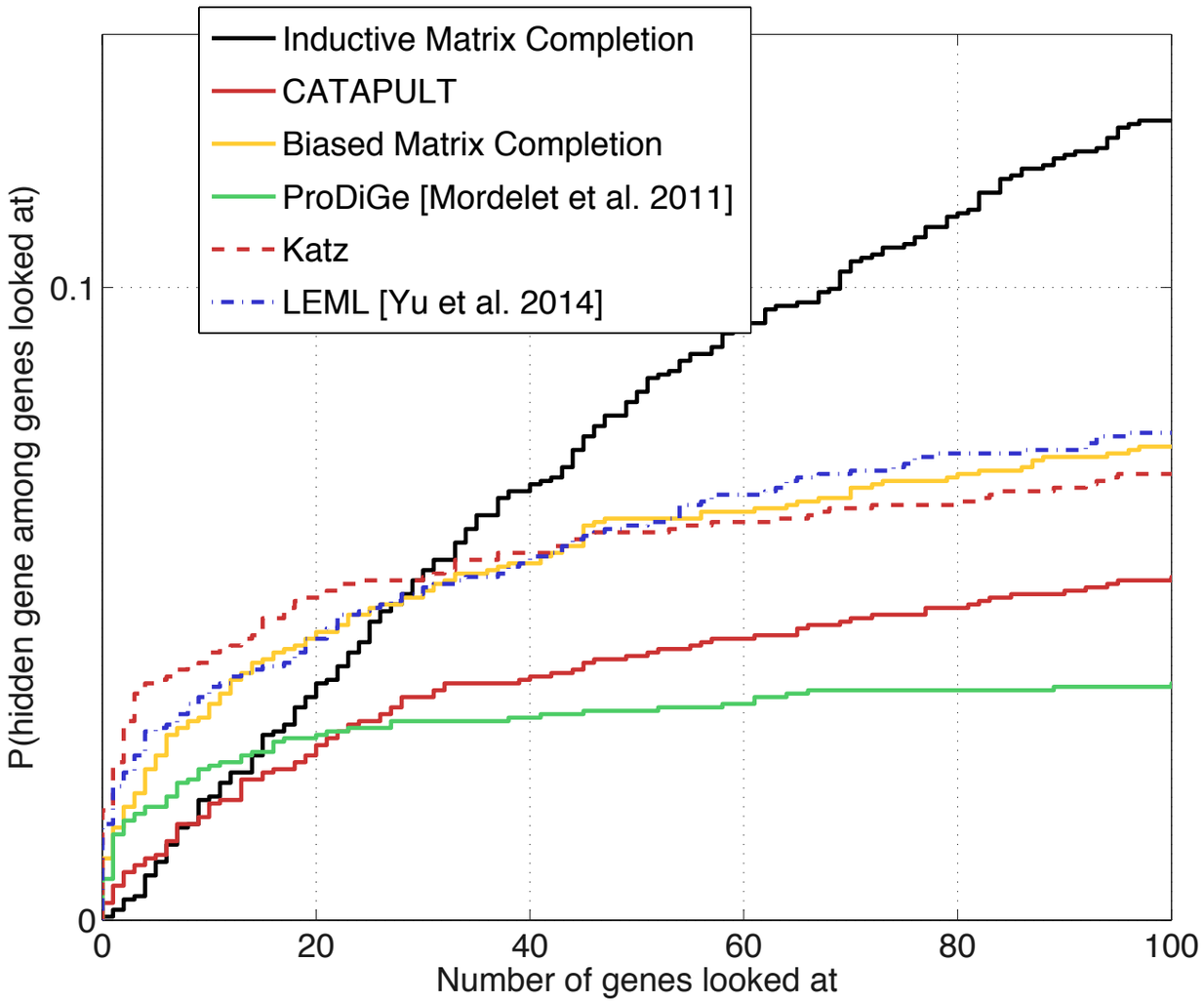
Natarajan N, Dhillon IS. Inductive Matrix Completion for Predicting Gene-Disease Associations. To appear in Bioinformatics, 2014.

Results on OMIM data

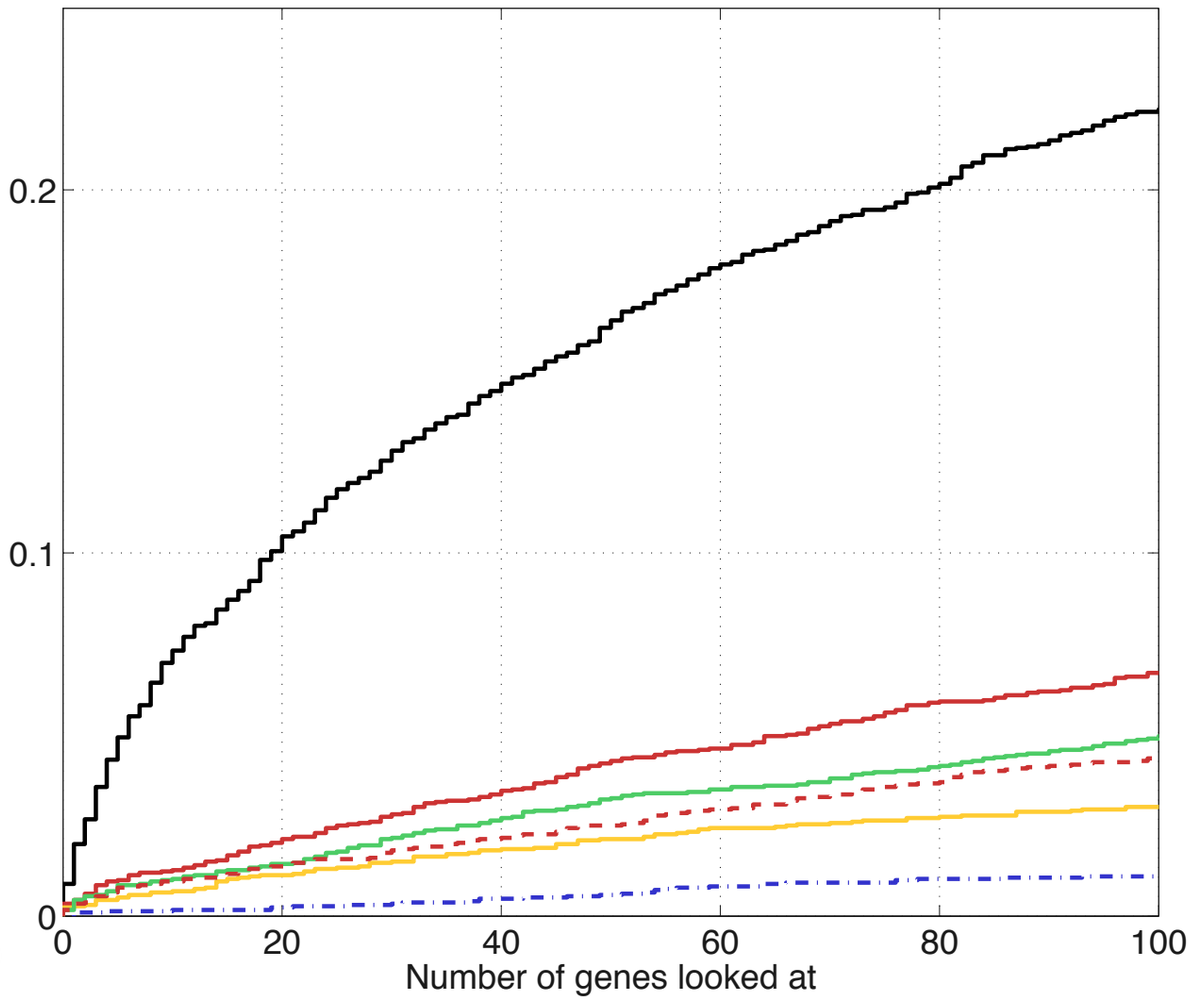


1. Online Mendelian Inheritance in Man: www.omim.org
2. F. Mordelet and J.-P. Vert. ProDiGe: PRioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics 2011.
3. Yu, Hsiang-Fu, Prateek Jain, and Inderjit S. Dhillon: Large-scale Multi-label Learning with Missing Labels. ICML 2014.

Results on OMIM data



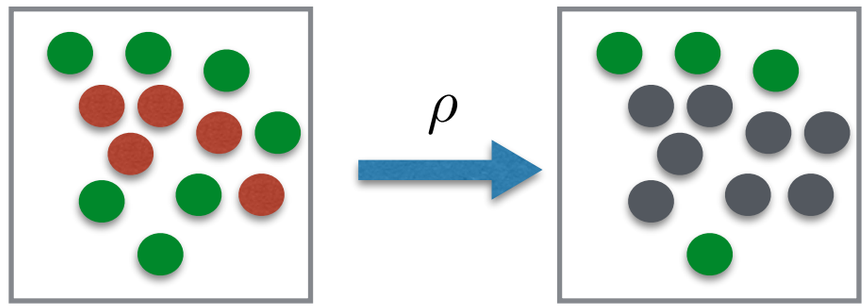
New genes



New diseases

Positive-Unlabeled Learning

- The prediction problem gives rise to a novel machine learning problem called “PU learning” —- learning in the absence of negative examples
- Methods such as “biased SVM” and “biased matrix completion” shown to perform well empirically
- Can analyze theoretically using random noise models



- For biased matrix completion, we can show

Theorem: Let ρ be the “noise rate”. With probability at least 1 - δ, solution X to the biased matrix completion is “close” to the true (n x n) matrix Y:

$$\frac{1}{n^2}\sum_{ij}(X_{ij}-Y_{ij})^2 = O\left(\frac{1}{(1-\rho)n}\right)$$

Conclusions

- Informatics for Computational Medicine is crucial
- Need novel statistical techniques —
  - Analysis of extremely high-dimensional data from high-throughput studies
  - Learning from diverse information sources
  - Interpretable computational models
- Can obtain better gene-disease associations than state-of-the-art