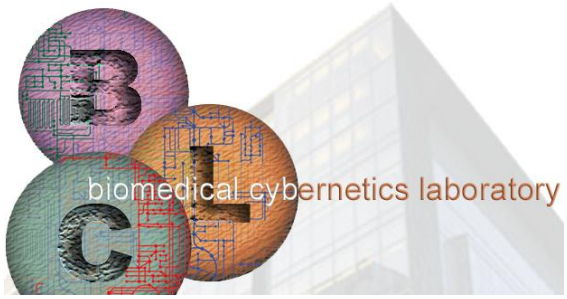
Prediction of Disease by Pathway-Based Integrative Genomic and Demographic Analysis

Skanda Koppula[14], Amin Zollanvari[123], Gil Alterovitz[1234*]

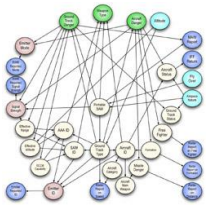
PRIMES Conference
May 18, 2013



[1] Center for Biomedical Informatics, Harvard Medical School [Boston, MA 02115].
[2] Children's Hospital Informatics Program at Harvard-MIT Division of Health Science [Boston, MA 02115].
[3] Partners Healthcare Center for Personalized Genetic Medicine [Boston, MA 02115].
[4] Dept.of Electrical Engineering and Computer Science at MIT [Cambridge, MA 02139].
* Corresponding author. Contact: gil@mit.edu

Introduction

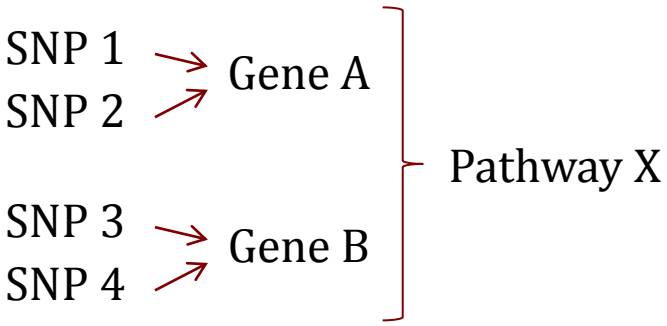
- Why prediction-based analysis of data?
  - ✓ Flexible model types
  - ✓ Gauge effect of feature on phenotype
  - ✓ …effective diagnostic tools!



Introduction

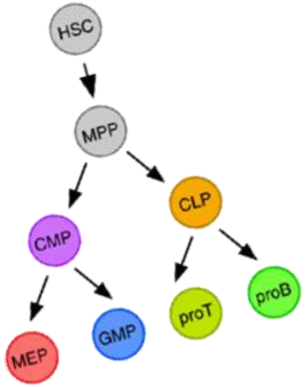
- Why prediction-based analysis of data?
  - ✓ Flexible model types
  - ✓ Gauge effect of feature on phenotype
  - ✓ …effective diagnostic tools!

- Try analysis on a different level!



Introduction

- Why prediction-based analysis of data?
  - ✓ Flexible model types
  - ✓ Gauge effect of feature on phenotype
  - ✓ …effective diagnostic tools!



- Try analysis on a different level?
  - ✓ Use inter-gene relations!
  - ✓ No black-box around disease mechanism
  - ✓ More knowledge about features *with no data*

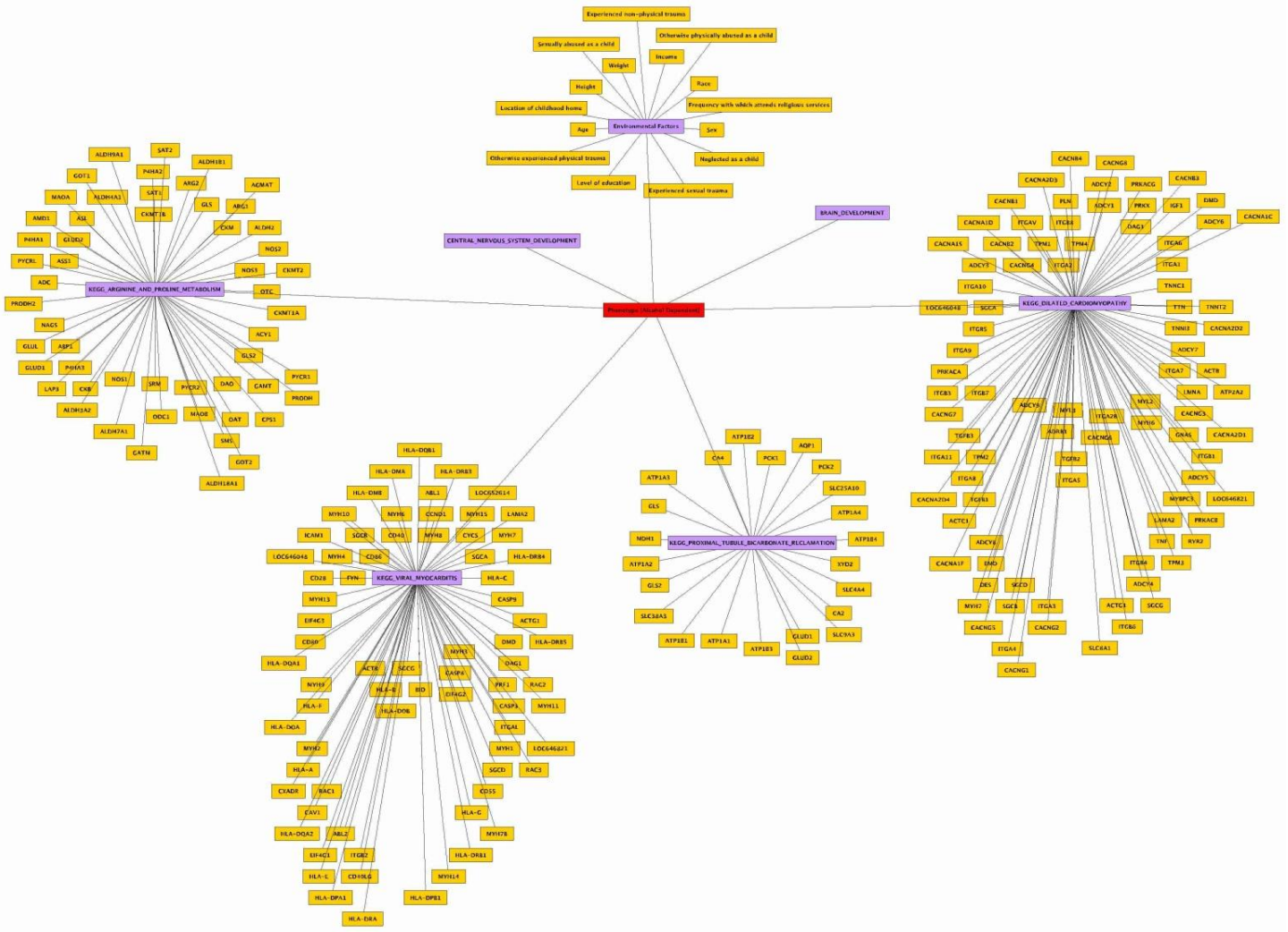
Introduction

- Why prediction-based analysis of data?
  - ✓ Flexible models [data type, number of features]
  - ✓ Easy to measure effect of feature on phenotype
  - ✓ Effective diagnostic tool

- Try analysis on a different level?

Pathway-based predictive models

Predictive Framework : TAN and Naïve Bayes



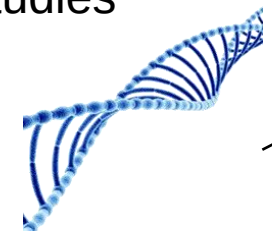
Alcoholism

2.5 million 14%

"increasing consumption of alcohol even in face of adverse consequences"

---

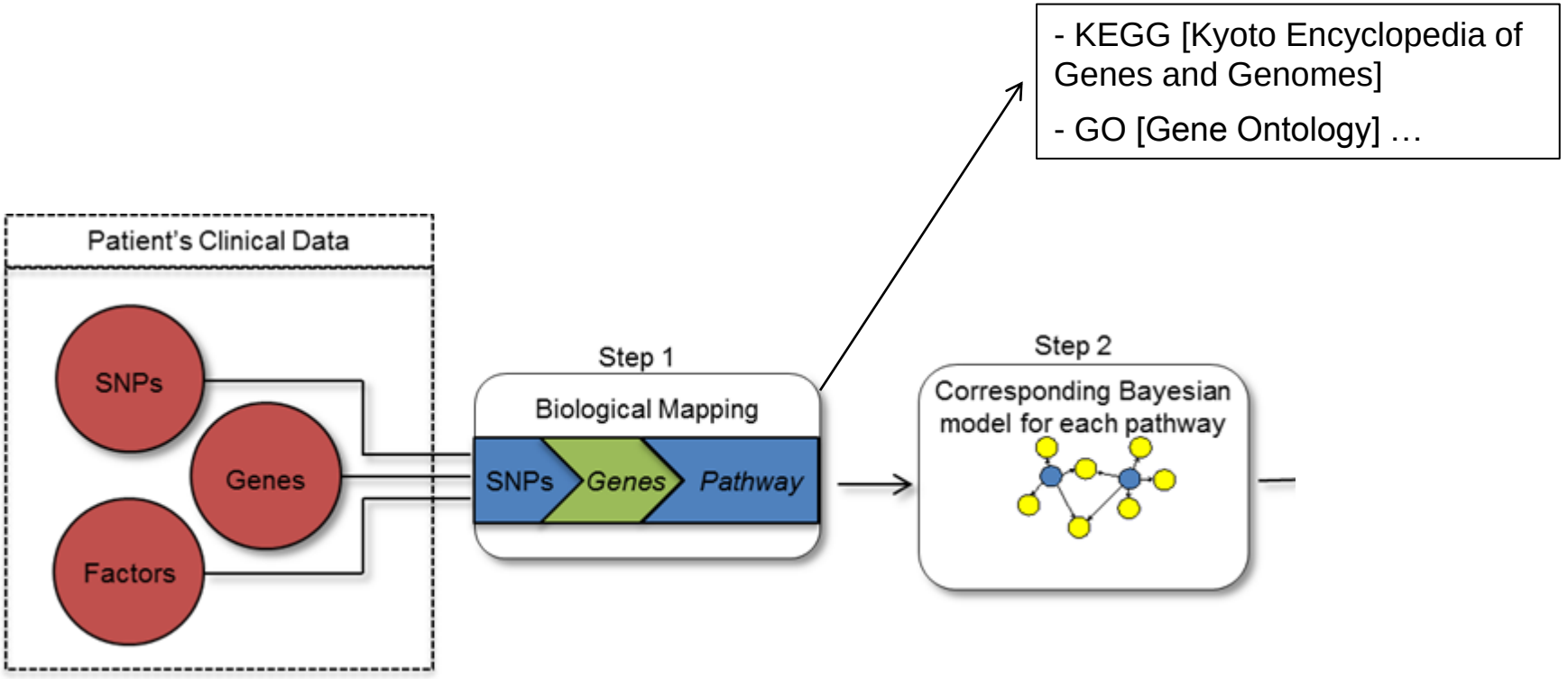
twin adoption studies

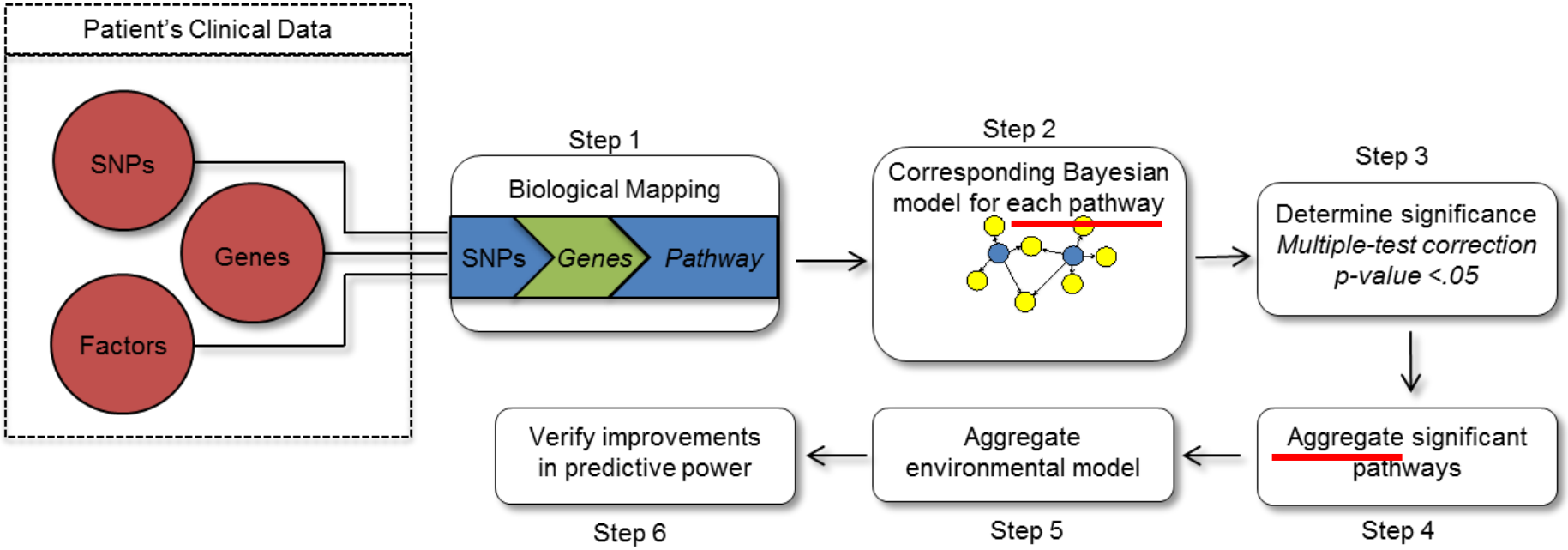


environmental studies

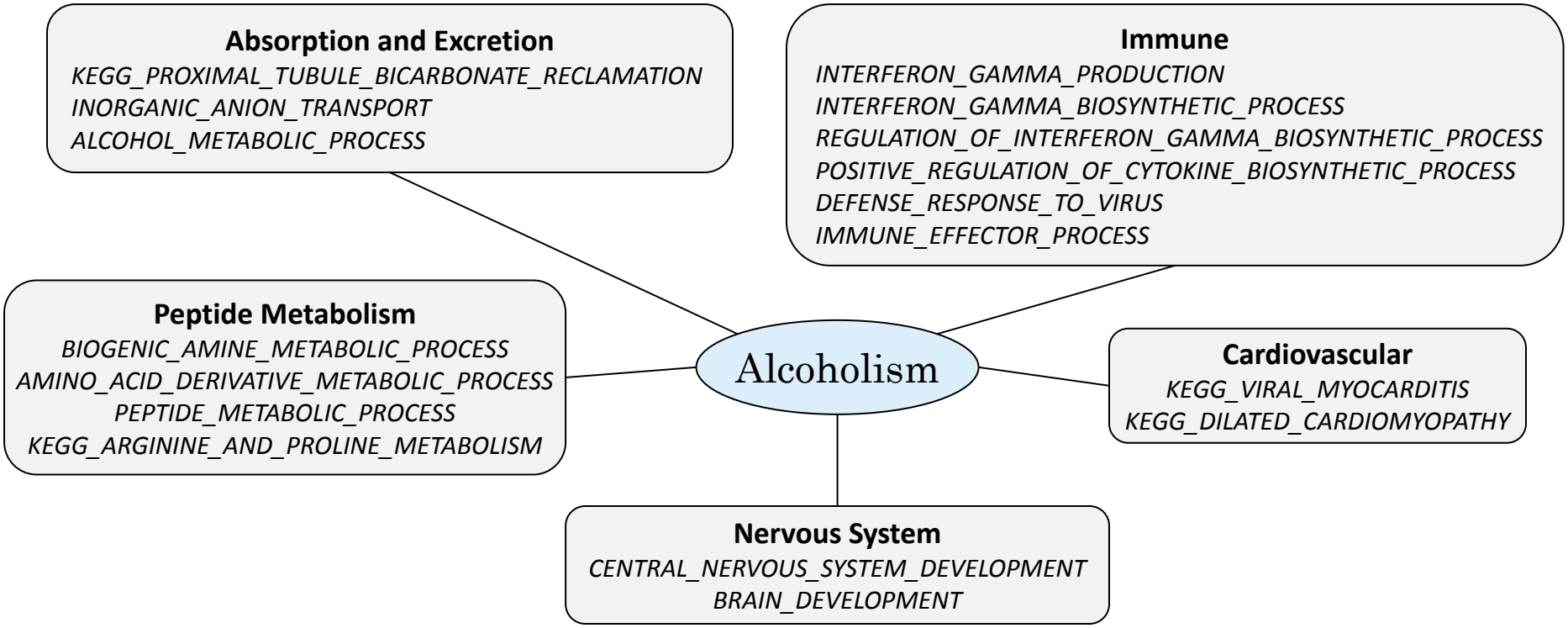
The datasets:

- COGA (1653 patients)
- COGEND (1350 patients)


- KEGG [Kyoto Encyclopedia of Genes and Genomes]
- GO [Gene Ontology] …
Patient's Clinical Data
SNPs
Genes
Factors
Step 1
Biological Mapping
SNPs
Genes
Pathway
Step 2
Corresponding Bayesian model for each pathway


Patient's Clinical Data
SNPs
Genes
Factors
Step 1
Biological Mapping
SNPs
Genes
Pathway
Step 2
Corresponding Bayesian model for each pathway
Step 3
Determine significance
Multiple-test correction
p-value <.05
Aggregate significant pathways
Step 4
Aggregate environmental model
Step 5
Verify improvements in predictive power
Step 6

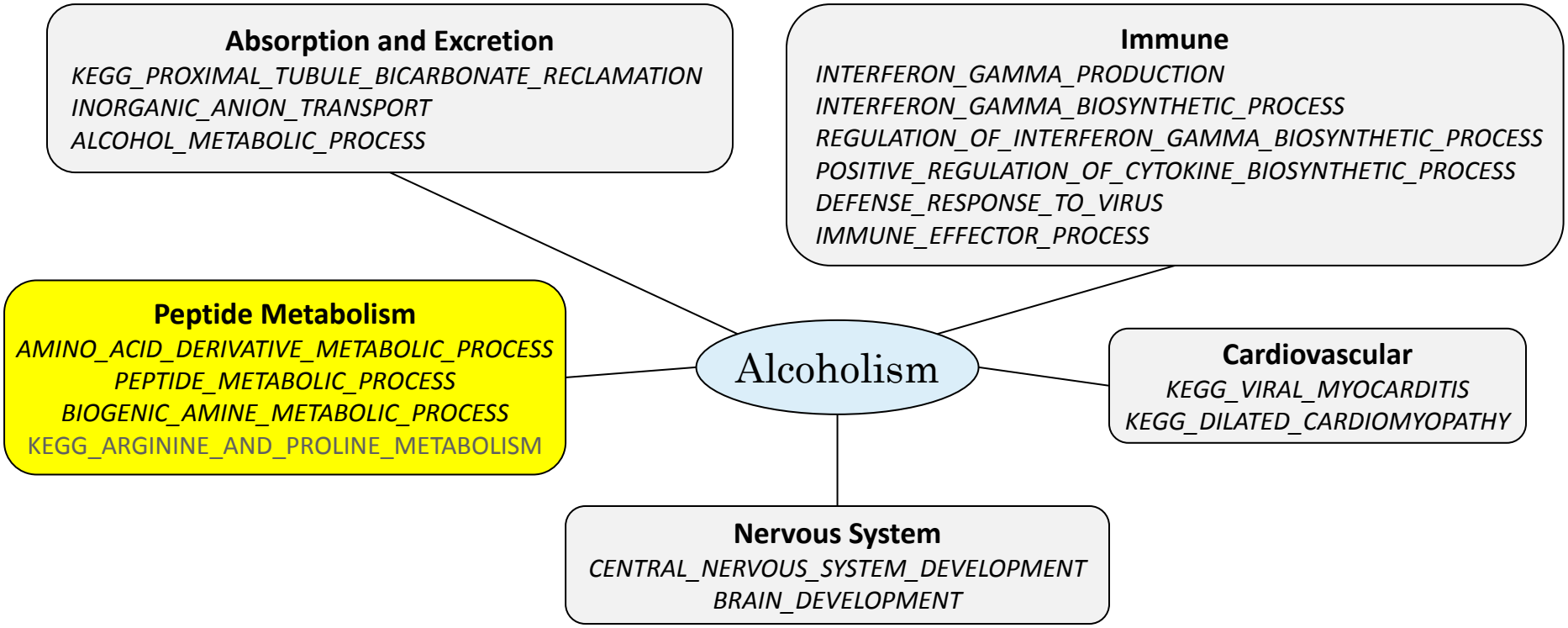
Genetic-Only Model



AUROC = 0.83

$p < 10^{-3}$

Genetic-Only Model



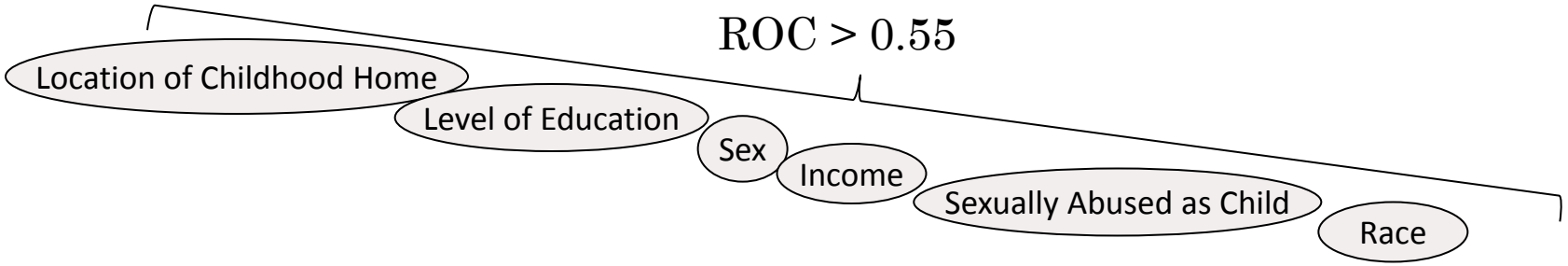
✸ AUROC = 0.83

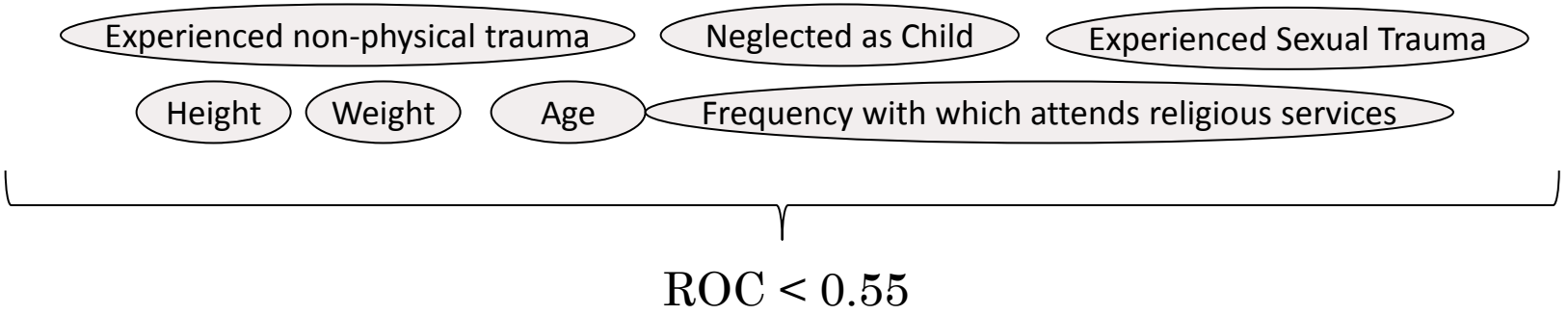
$p < 10^{-3}$

Genetic-Demographic Model

AUROC = 0.90

$p < 10^{-5}$





Genetic-Demographic Model

- Increase due to more # features?
  - ✓ No! Replacement increases accuracy by 2.8%

- Why?
  - ✓ Genes and demo. factors boost each other
    - ☞ *Inorganic Anion Transport contains {CLCNX gene group} on X-chromosome*

Lung Cancer

Pathway	AUROC
Estrogen receptor regulation (carm1 and -er)	0.75
Eukaryote Translation Initiation Factor (eif4, eif2)	0.73
rnaPathway	0.73
ST_Tumor_Necrosis_Factor_Pathway	0.72
vegfPathway	0.67
MAP00010_Glycolysis_Gluconeogenesis	0.66
P53_UP	0.66

$$\text{AUROC} = 0.85$$

$$p < 10^{-5}$$

Next Steps

1. Insight from inter-feature relationships?
2. Application for layman to use predictive framework?
3. *In vitro* validation of identified pathways
4. Other learning structures?

Acknowledgements

- PRIMES program for providing me with this opportunity
  - ✓ Dr. Gerovitch, Professor Etingof, and Professor Khovanova
- Professor Alterovitz
- NIH Grants:
  - ✓ 5R21DA025168-02 (G. Alterovitz)
  - ✓ 1R01HG004836-01 (G. Alterovitz)
  - ✓ 4R00LM009826-03 (G. Alterovitz)

Thank You! Questions?