BE700 SPRING 2021

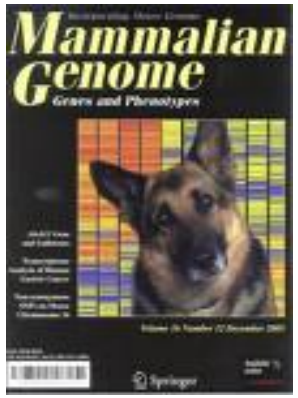
SYSTEMS BIOLOGY AND ARTIFICIAL INTELLIGENCE (AI)

USING AI TO ADVANCE PERSONALIZED MEDICINE

Nature unveiled….

Human genome…… chimp genome ……dog genome ….. human mutations……

2001

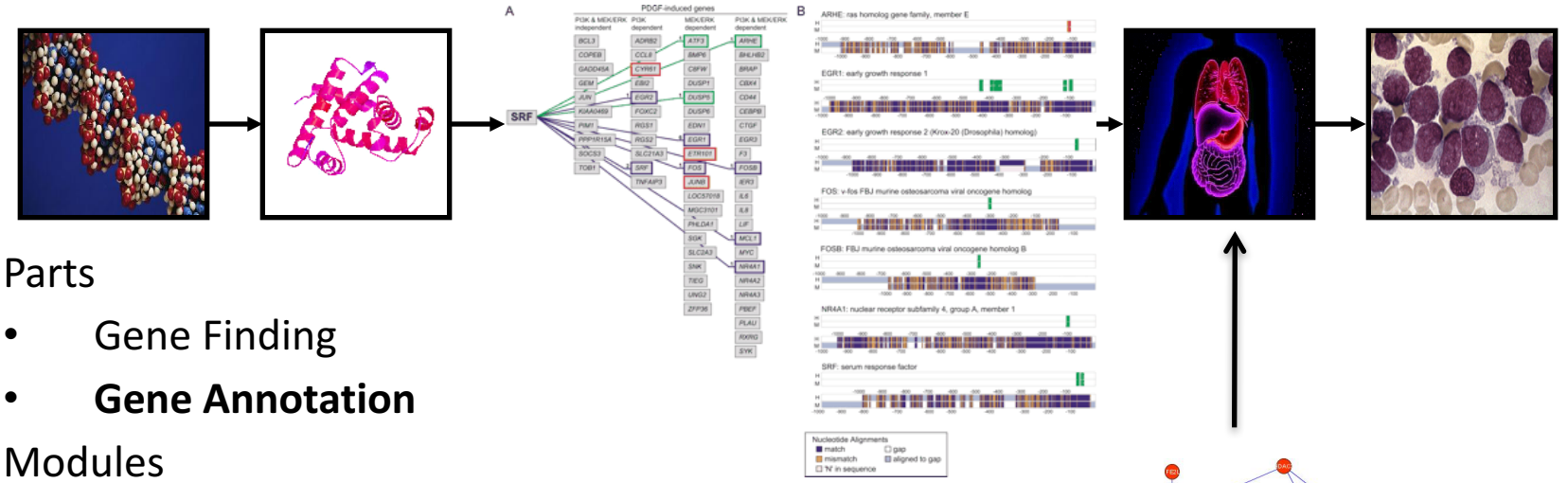


What next?

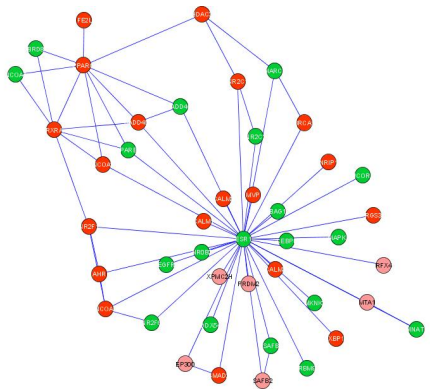
Deciphering Nature

F?

Genome → **Phenotype**



1. Parts
   - Gene Finding
   - **Gene Annotation**
2. Modules
   - Gene Regulation
   - Protein-Protein Interaction Networks
3. Pathways: Discovering and Modeling Gene Modules
4. **Associating Pathways/Modules with Clinical Phenotypes**
5. Causal Network Models of Disease

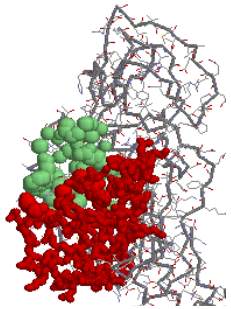


"Imagination is better than knowledge", Albert Einstein

This quote is the inspiration for much of creative advances in computer science in recent years. When I started working in Computational and Systems Biology twenty years ago I was frustrated with the amount of knowledge one needs to make high impact advances and talked to my friend David Lipman (Director of NCBI).

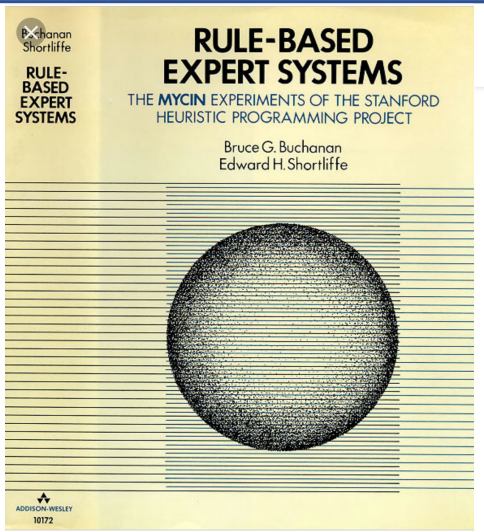
"But perhaps nature has a bigger imagination", David Lipman

It is humbling !



Pictures borrowed from the WWW

AI and Medicine



Cardiogram

Three Challenges for Artificial Intelligence in Medicine

... particularly deep learning — self-driving cars, Siri, AlphaGo, Google Translate, computer vision — the effect on medicine has been nearly nonexistent.

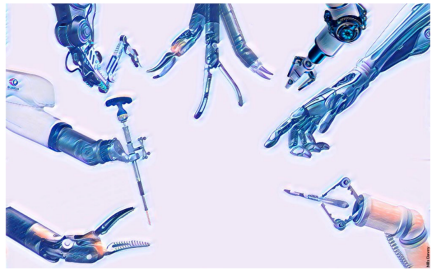
Images may be subject to copyright. Learn More

INVESTOR'S BUSINESS DAILY

Sign In

TECHNOLOGY

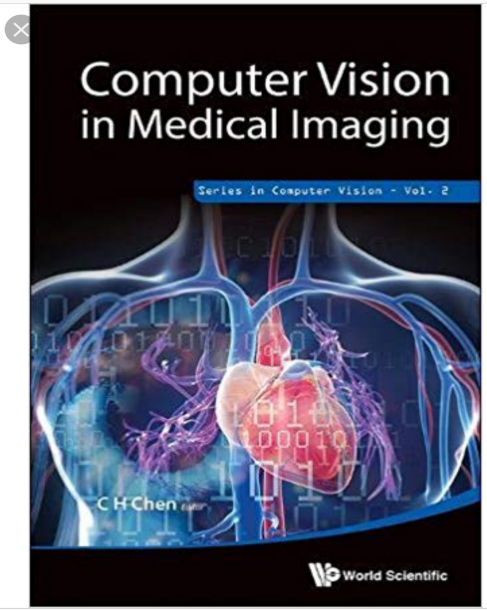
AI And Robotics Are Taking Robotic Surgery To New Levels



The robotic surgery market is about to get a lot bigger as new systems are approved for sale. (Nils Davey for IBD)

ALLISON GATLIN | 6/12/2018

With artificial intelligence now firmly entrenched in many hospital operating rooms, the field of robotic surgery is starting to get competitive.



Amazon.com

Computer Vision in Medical Imaging

google.com

healthitanalytics.com

HEALTH IT ANALYTICS
xtelligent HEALTHCARE MEDIA

Has Google Cracked EHR Speech Recognition for Medical Conversations?

Two new speech recognition models from Google may offer a way to reduce EHR burnout by accurately recording medical conversations in natural settings.

Computational/Evolutional Thinking about Biology and Disease

- *“Biology is a software process. Our bodies are made up of trillions of cells, each governed by this process. You and I are walking around with outdated software running in our bodies, which evolved in a very different era”.*
- *Ray Kurzweil*

What is AI ?

- AI has many subfields ☺
- Core applications: speech, language, vision, robotics
- Core AI
  - Machine Learning (ML) : Inductive Inference, Classification
  - Knowledge Representation and Acquisition
  - Automated Reasoning (AR): Deductive Inference
  - Planning, Decision Making
- Machine learning is very trendy
- Reasoning and Planning are equally IF NOT MORE challenging
- Many types of Reasoning
  - Probabilistic (e.g. graphical models, Bayes)
  - Logic
- Our focus:
  - INTEGRATION OF ML + AR + BIOLOGY

AI IN ACTION:

GLIMMER: A MACHINE LEARNING SYSTEM THAT IDENTIFIES GENES IN NEW GENOMES

GLIMMER is a **Microbial** Gene Finder
GLIMMER LEARNS FROM DATA in a Semi-Supervised Fashion Using a Speech Understanding Technique (Interpolated Markov Models – from Speech)
PREDICTS regions of the genome that code for proteins (e.g. ENZYMES).
GLIMMER contributed the majority of the microbial genes we know today (millions of enzymes in thousands of organisms)
Microbial Enzymes are very important for medicine: CRISPR, OPTOGENETICS..

Implemented by Arthur Delcher
(Art was my 1st PHD Student at Johns Hopkins)
In collaboration with Steven Salzberg and Owen White at TIGR

98% accuracy on gene identification

Planning Beyond Human Reach

- ***Computer Is Pushed to Edge To Solve Old Chess Problem***
- **By THE ASSOCIATED PRESS 1990**
- A 25-year-old graduate student has solved a long-standing chess problem by taking a computer to places no computer has gone before.
- The double feat, by **Lewis Stiller**, a computer science student at Johns Hopkins University, not only settled an old chess conundrum, but also opened the door for analysis once believed to be too complicated for even the fastest computers.
- By performing one of the **largest computer searches** ever conducted, Mr. Stiller found that a king, a rook and a bishop can defeat a king and two knights in a **maximum of 223 moves**, ending centuries of uncertainty over whether the position is a draw.

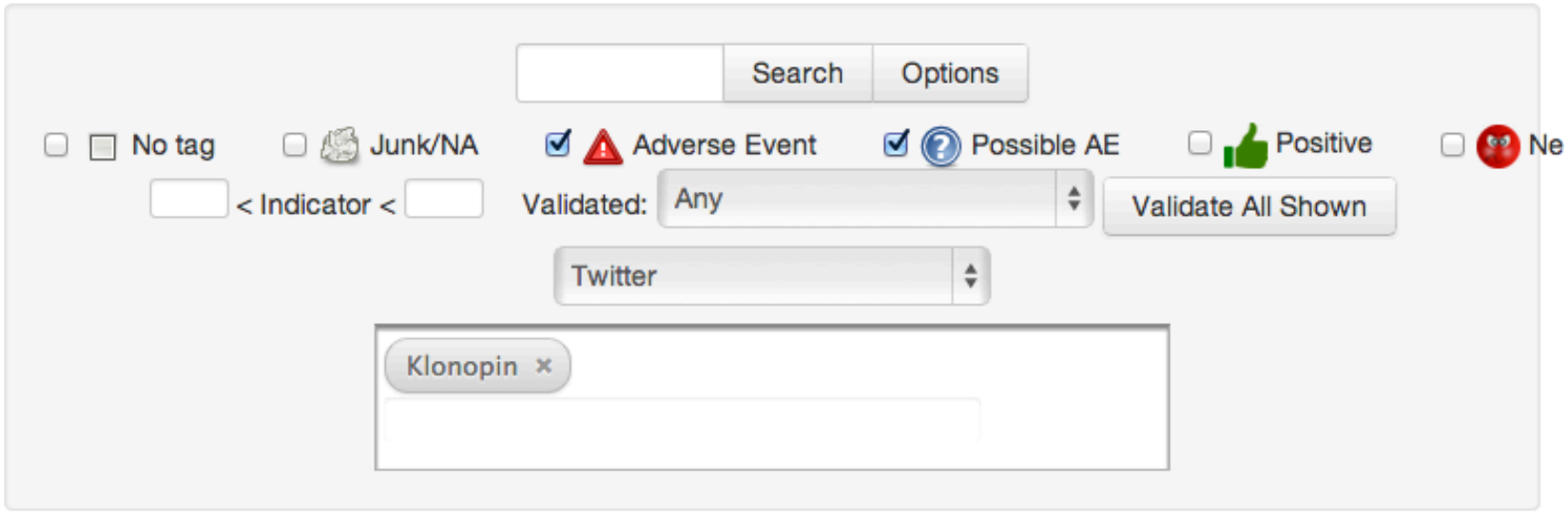
- Lewis was my 3rd PhD Student ☺
- He programmed on the Connection Machine Built by Danny Hillis (MIT)

MedWatcher: A Machine Learning System that Mines Social Networks for Adverse Drug Events

- Acquisition: collect posts from online forums (e.g. **Twitter)** via search for product **generic** and **brand** names.
- Apply **natural language processing and machine learning algorithms** to filter posts and identify adverse events.

Clark Freifeld, PhD 2014 MS Media Lab
(co-advised by J. Brownstein (HMS) and S. Kasif)

Curation Tool



0.926	Weed couldn't get me high lol I've taken morphine since I was 6th grader y'all don't know about my kinda high	4 Mar 04:09:36 2014
0.938	Leg super painful this morning, so have taken a couple of paracetamol & codeine tablets, now feeling fuzzy all over, but no leg pain, hurrah	4 Mar 04:08:30 2014
0.997	If only I'd known my vaccination would make my arm swell up, I probably would have asked for it in the arm I don't have to sleep on... #doh!	4 Mar 04:08:13 2014
0.955	@I[redacted]s i honestly think i died.... ambien makes me feel the same when i fight the sleep and start getting wavy. lol	4 Mar 04:06:01 2014

Discovery of a Wellness Network ?

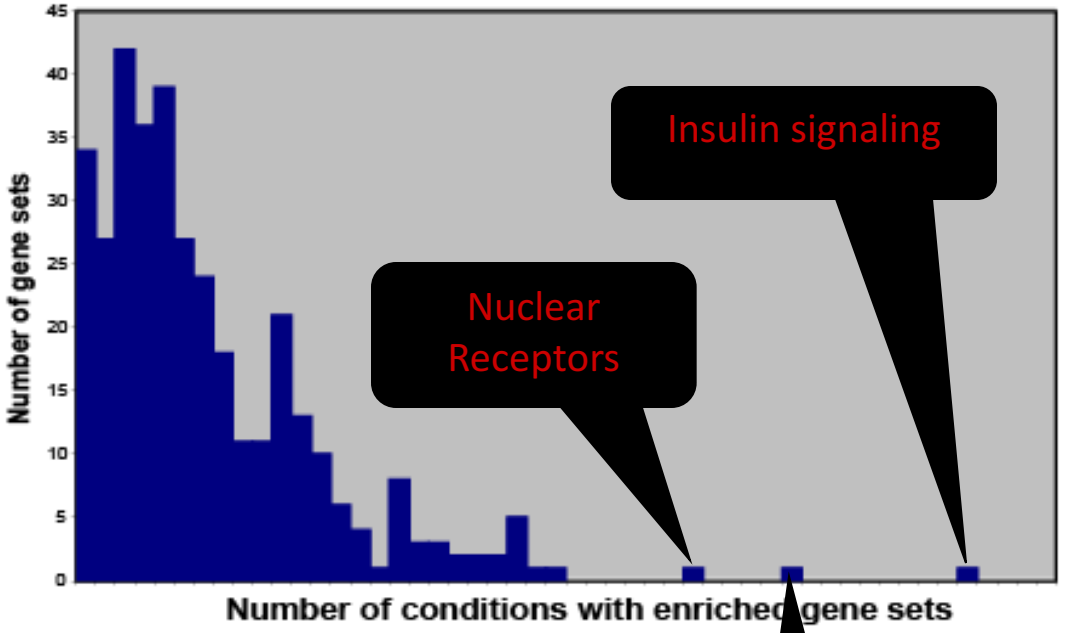
Manway Liu (double major MIT)

co-mentored by Zak Kohane (HMS) and Simon Kasif

1. Insulin signaling, interleukins (inflammation), and nuclear receptors (hormone receptors).

2. Insulin signaling is consistent with the given disease models. Was not identified using standard techniques.

3. Interleukins and nuclear receptors consistent with the inflammation and disordered metabolism associated with type 2 diabetes



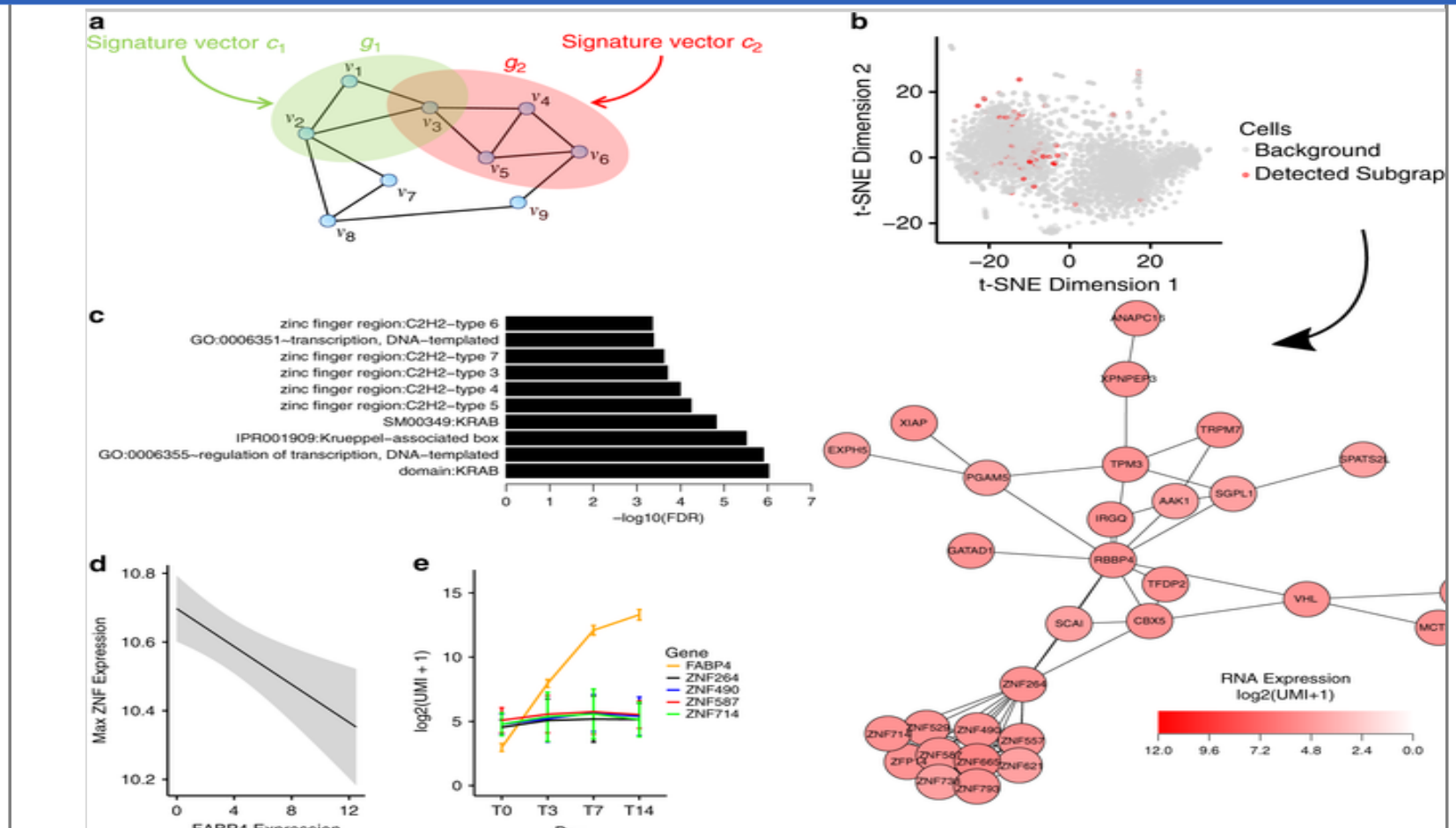
Nuclear receptors: 31 of 67 (many hormone receptors)
Interleukins: 38 of 67 JNK, SMAD
Insulin signaling: 45 of 67.

Discovery of a Network that Inhibits Differentiation?

Alfred Ramirez et al, to appear 2020

Alfred is a double major Biology & Physics from MIT

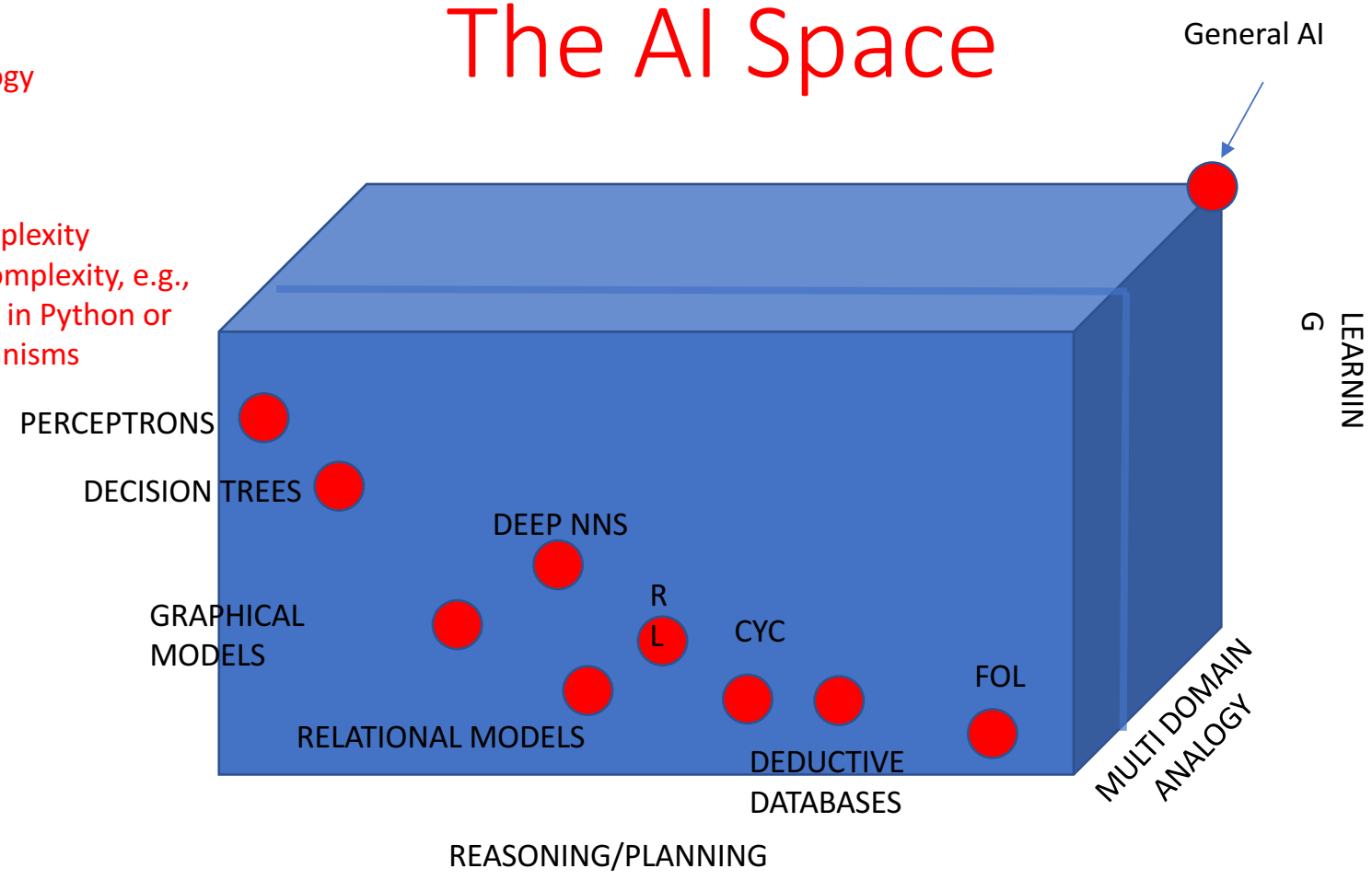
co-mentored by Simon Kasif and Ron Kahn (HMS,Joslin)



AI Dimensions

Learning
Reasoning/Planning
Multi-Domain Analogy
Abstraction
Explanation
Self Reflection
Computational Complexity
Representational Complexity, e.g., learning to program in Python or synthesize new organisms

The AI Space



AI and Biology: Slice of the Past

1. Protein Structure
2. Bacterial Gene Identification
3. Drosophila Gene Finding
4. Human Gene Finding
5. Human Genome
6. **Mouse Genome ****
7. Network Based Function Prediction
8. Interaction Prediction
9. Bacterial Regulatory Prediction
10. Human Regulatory Network Prediction
11. Cancer Module Prediction
12. Gene X Phenotype Prediction
13. Bacterial Systems
14. Integration
15. Drug Discovery

1. Neural Nets Sejnowsky et al /
2. Interpolated Markov Models Delcher et al
3. Neural Nets, Haussler et al
4. Generalized HMMs, Burge et al
5. Human Genome
6. Paired HMMs, Brent et al, Berger/Lander et al
7. MRFs, Diffusion, Gene Mania (Mostafavi et al)
8. Bayes Nets, Troyanskaya et al, Gerstein et al
9. Bayes Nets, Many methods, Faith et al
10. Causal Inference Califano et al, BNs - Peer et al
11. Segal/Koller/Friedman
12. Too many to list …
13. Too many to list, Overbeek et al, Peter Karp
14. Bayes Nets, Kasif et al, Gifford/Jakkola et al
15. Many and most recently Jim Collins et al MIT

**** Tarjei Mikkelsen et al**
Tarjei was an undergrad at MIT when he joined my HGP group ☺

…MUCH MORE ……. ☺

Logistics

- No Exams
- Homework
- Project in Groups (SUBJECT TO CHANGE)
- Reading Assignments
- Project Presentations
- Grading
  - Homework: 75%
  - Project 25%
  - Class Participation Extra 10%
- TA ?
- All email with the subject line: BE700 Fall 2020
- All class materials will be posted on BB

Machine Learning INTRODUCTION

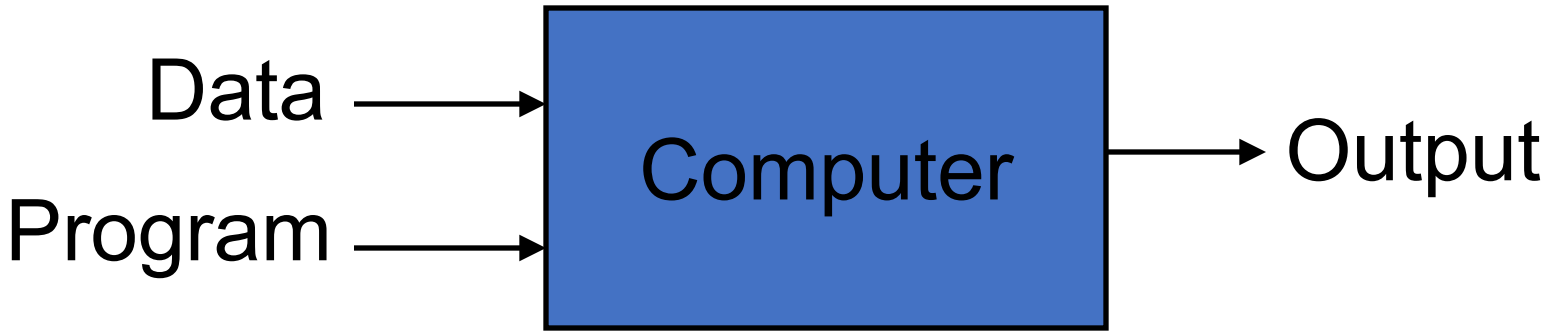
BOOKS
NOT REQUIRED

- T. Mitchell, ***Machine Learning,*** McGraw-Hill
- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley *(A CLASSIC BUT MORE TECHNICALLY CHALLENGING)*
- *Readings provided on the BB (required)*

Machine Learning in Popular Press

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning today is one of the hottest aspects of computer science” (Steve Ballmer, CEO, Microsoft)

Computer Programming



Machine Learning in the future

Examples

- Past Tense Learning
  - Input → Output
  - Go → Went
  - Do → Did
  - Love → Loved
  - Learn → Learned
  - Forget → ??
  - Memorize → ??

Systems Biology Example

Gene 1	Gene 2	Class
0.9	0.88	Normal
2.3	-1.3	Cancer
1.2	1.1	Normal
2.8	-1.5	Cancer
2.5	-1.2	??

Representation

- Memories
- Decision trees
- Rules / Logic
- Classifiers
- Hierarchies
- Graphical models (Bayes Networks)
- Neural networks with many layers
- Perceptrons and Support vector machines
- Programs

Evaluation

- Accuracy (true/false positives and negatives)
- Precision and recall
- Error
- Likelihood
- Posterior probability
- Entropy
- K-L divergence
- AUC
- More

ML Current State

- Hundreds of thousands of machine learning programs
- Thousand of papers every year
- Most machine learning algorithms
  - **Choosing knowledge representation**
  - **Clever Algorithm – mostly standard with tweaks**
  - **Training / Testing / Evaluation of Accuracy**
  - **Optimization**
- **High salaries !**

Optimization

- Combinatorial optimization
  - E.g.: Simulated Annealing or search
- Convex optimization
  - E.g.: Stochastic Gradient descent
- Constrained optimization
  - E.g.: Quadratic programming

Types of Learning

- **Supervised (inductive) learning**
  - Training data includes class labels or output for every input
- **Unsupervised learning**
  - Training data does not include class labels or output for every input
- **Semi-supervised learning**
  - Training data includes some labelled and many un-labelled inputs
- **Other Learning (not covered in class)**

Inductive Learning

- **Given** examples of a function *(X, F(X))*
- **Predict** function *F(X)* for new examples *X*
  - Discrete *F(X)*: Classification
  - Continuous *F(X)*: Regression
  - *F(X)* = Probability(*X*): Probability estimation

What We'll Cover

- **Supervised learning**
  - Decision tree induction
  - Rule induction
  - Instance-based learning
  - Bayesian learning
  - Neural networks
  - Support vector machines
  - Model ensembles
  - Learning theory
- **Unsupervised learning**
  - Clustering
  - Dimensionality reduction

ML in Practice

- Choosing Representations (or Models) the ML program will learn
- Data integration, feature selection, pre-processing, normalization, identifying outliers, and more .
- Learning
- Interpreting results (key emphasis in the class)
- Deployment and Improvement

Supervised Learning From Measurements

	Gene1	Gene2	Gene3....	Class Label
Patient 1	2.1	1.1	1.3	normal
Patient 2	10.2	0.51	3.2	cancer
Patient 3				
…				
…				
Patient N….				

Aim: Automatically Produce a classifier that predicts class label given the input gene expression profile of a biopsy

Simple Learning Method of Gene/Protein Function using 1-NN

- Given a new gene expression profile X produce the class label of X, f(X) as follows:
- Identify the most similar profile X’ in the training set D
- Most similar = smallest Euclidean distance i.e., X’ = min |X-X’| of all X’ in D.
- Let f(X’) → f(X) Done !!

K-NN