Machine Learning for Big Data "Complexity" in Biomedical Data Analytics

Yanjun (Jane) Qi, PhD
Department of Computer Science,
University of Virginia

2014.03.05

OUR DATA-RICH WORLD

- Biomedicine
  - Patient records, brain imaging, MRI & CT scans, …
  - Genomic sequences, protein-structure, drug effect info, …
- Science
  - Historical documents, scanned books, databases from astronomy, environmental data, climate records, …
- Social media
  - Social interactions data, twitter, facebook records, online reviews, …
- Business
  - Stock market transactions, corporate sales, airline traffic, …
- Entertainment
  - Internet images, Hollywood movies, music audio files, …

BIG DATA CHALLENGES

- Data capturing (sensor, smart devices, medical instruments, et al.)
- Data transmission
- Data storage
- Data management
- High performance data processing
- Data visualization
- Data security & privacy (e.g. multiple individuals)
- ……

Today

- Data analytics
  - How can we convert this big data wealth to knowledge ?
  - E.g. Machine learning

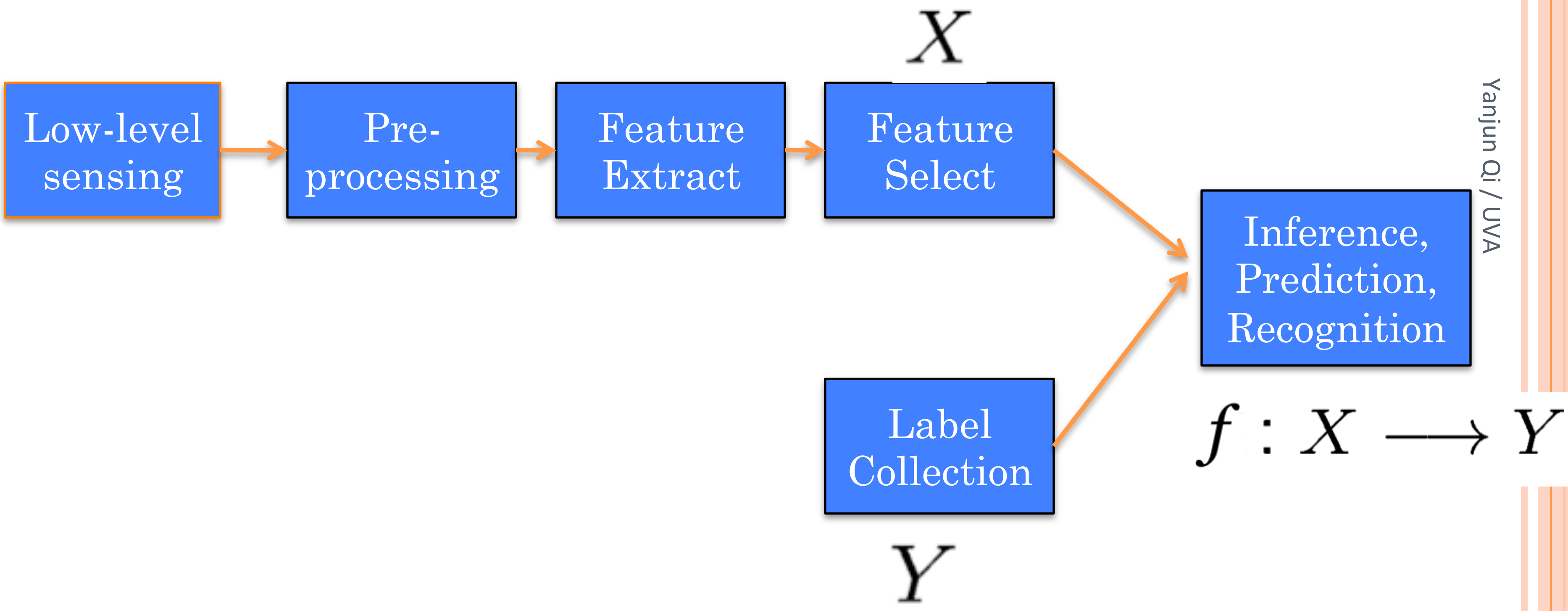
BASICS OF MACHINE LEARNING

- "The goal of machine learning is to build computer systems that can learn and adapt from their experience." – Tom Dietterich

- "Experience" in the form of available data examples (also called as instances, samples)

- Available examples are described with properties (data points in feature space X)

TYPICAL MACHINE LEARNING SYSTEM

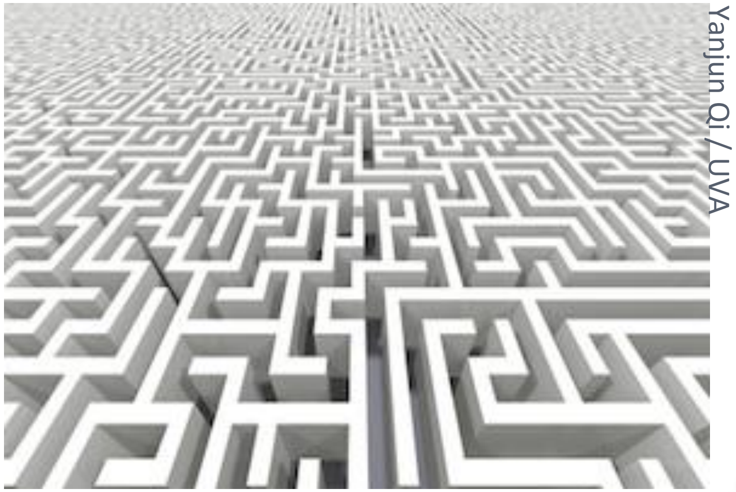


$$f : X \longrightarrow Y$$

BIG DATA CHALLENGES FOR MACHINE LEARNING

LARGE-SCALE

Highly Complex



Yanjun Qi / UVA

Today

The situations / variations of both X (feature, representation) and Y (labels) are complex !

When to use Machine Learning (Adapt to / learn from data) ?

- 1. Extract knowledge from data
  - Relationships and correlations can be hidden within large amounts of data
  - The amount of knowledge available about certain tasks is simply too large for explicit encoding (e.g. rules) by humans

- 2. Learn tasks that are difficult to formalise
  - Hard to be defined well, except by examples

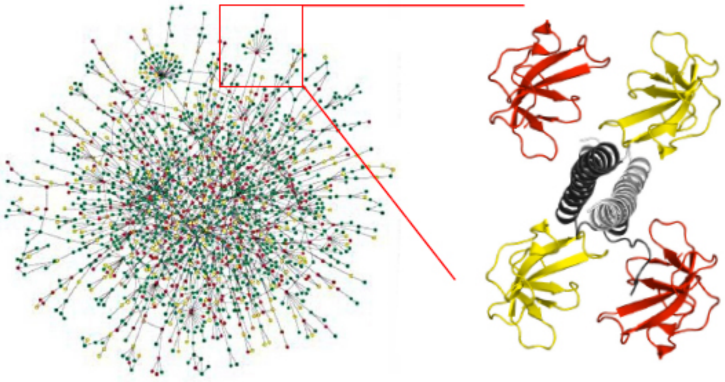
- 3. Create software that improves over time
  - New knowledge is constantly being discovered.
  - Rule or human encoding-based system is difficult to continuously re-design "by hand".

Interesting Data Challenges in BioMed for Machine Learning

- Noisy measurements (e.g. weak/partial labels)
- Structured input (e.g. vector, strings, graphs)
- Structured output (e.g. trees, sequences, graphs)
- Combination of different data types is essential (e.g. information fusion)
- Large amount of data (e.g. lots of next generation sequencing data)

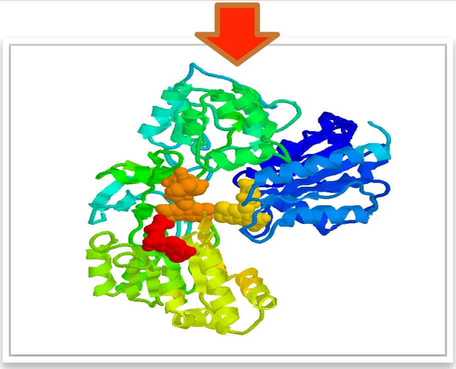
THIS TALK COVERS

I.

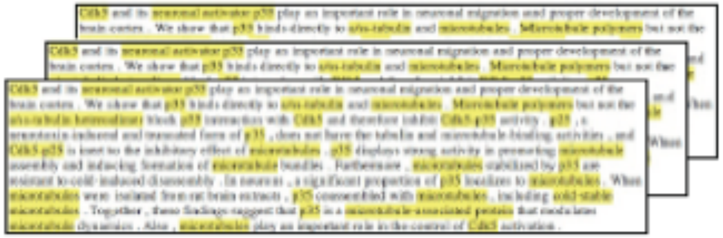


II.

MTYKLILNGKTKGETTTEAVD...



III.

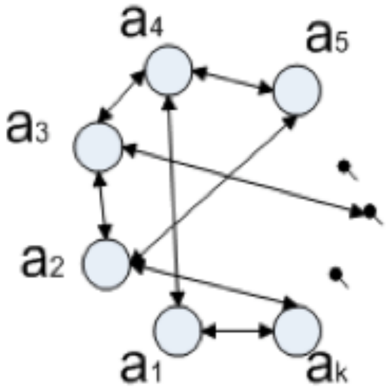


Cdk5 and its neuronal activator p35 play an important role in neuronal migration and proper development of the brain cortex . We show that p35 binds directly to α/β-tubulin and microtubules . Microtubule polymers but not the α/β-tubulin heterodimers block p35 interaction with Cdk5 and therefore inhibit Cdk5/p35 activity . p25 , a neurotoxin-induced and truncated form of p35 , does not have the tubulin and microtubule-binding activities , and Cdk5/p25 is inert to the inhibitory effect of microtubules . p35 displays strong activity in promoting microtubule assembly and inducing formation of microtubule bundles . Furthermore , microtubules stabilized by p35 are resistant to cold-induced disassembly . In neurons , a significant proportion of p35 localizes to microtubules . When microtubules were isolated from rat brain extracts , p35 coassembled with microtubules , including cold-stable microtubules . Together , these findings suggest that p35 is a microtubule-associated protein that modulates microtubule dynamics . Also , microtubules play an important role in the control of Cdk5 activation .

Cdk5 and its neuronal activator p35 play an important role in neuronal migration and proper development of the brain cortex . We show that p35 binds directly to α/β-tubulin and microtubules . Microtubule polymers but not the

Cdk5 and its neuronal activator p35 play an important role in neuronal migration and proper development of the brain cortex . We show that p35 binds directly to α/β-tubulin and microtubules . Microtubule polymers but not the

IV.



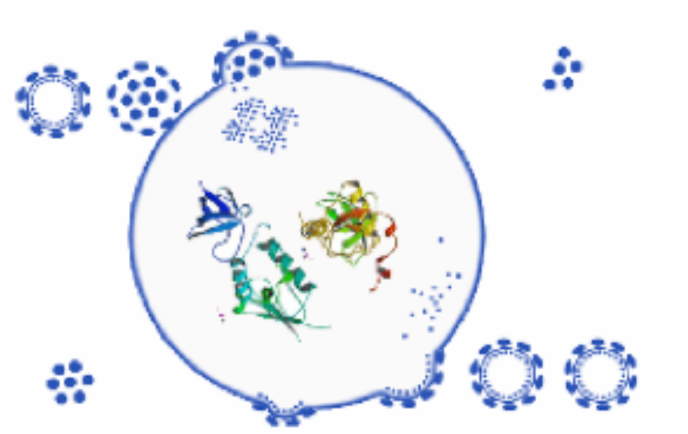
THIS TALK COVERS

	Project Topic	Complexity	HOW ?
I	Protein interaction identification	Y	Training with auxiliary labels
II	Protein structure prediction	X & Y	Unified feature learning for multiple related tasks
III	Biomedical text mining	X	Add semi-supervision on features
IV	Conditional dependency graph among Genes / TFs	X	Model data example with feature interactions

Yanjun Qi / UVA

VIRUS VS. HUMAN PROTEIN INTERACTION

- Human Immuno-deficiency Viruses, (e.g. HIV-1 Virus), can cause life-threatening infectious diseases (like AIDS)
- Virus must communicate with the host to invade and infect
- Typical communication through interactions between virus and human host proteins (*potential drug/vaccine targets*)

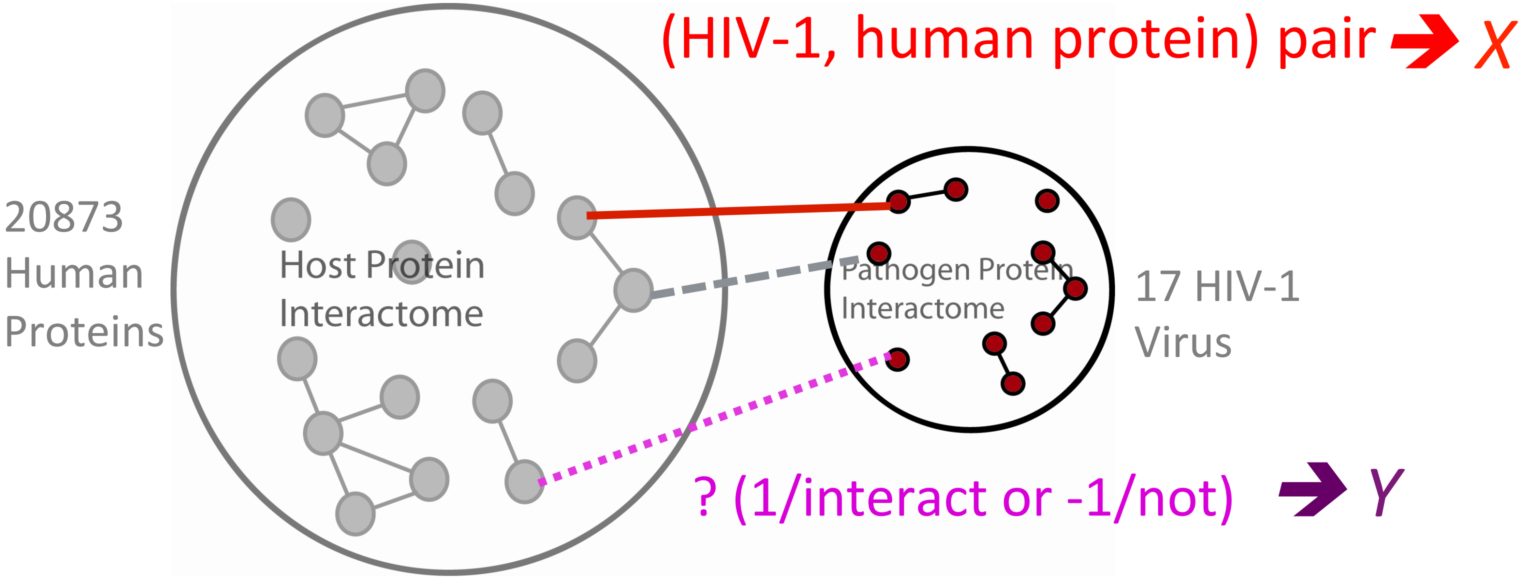


[Y. Qi, et al, Bioinformatics 2010]
[Y. Qi, et al, Proteomics 2009]

Objective & Previous Work

- **GOAL: to discover unknown direct physical interactions between HIV-1 and human proteins**

➔ **(Help biologist prioritize potential interaction pairs)**



(HIV-1, human protein) pair ➔ *X*

20873 Human Proteins

17 HIV-1 Virus

? (1/interact or -1/not) ➔ *Y*

Yanjun Qi / UVA

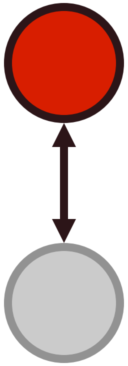
- Model each (HIV-1, human protein) pair with (X, Y)
- State-of-the-art performance: Random forest (Tastan et al. (PSB 2009))

Simplified view: lost spatial / temporal information of interaction pairs

[Y. Qi, et al, Bioinformatics 2010]   [Y. Qi, et al, Proteomics 2009]

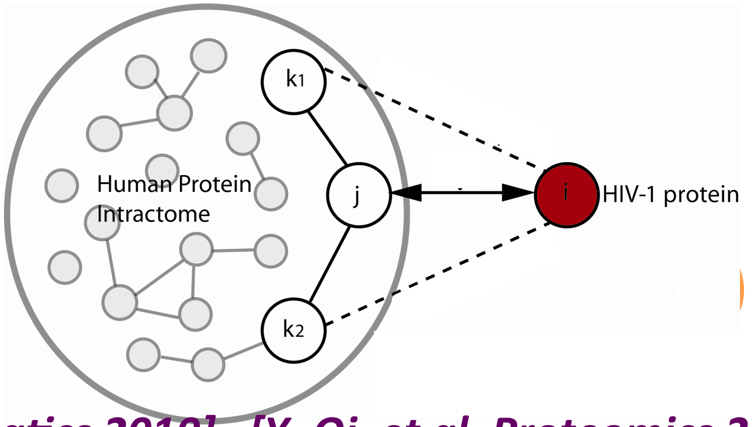
Background: 18 Features describing each pair

- □ Differential gene expression in HIV infected vs uninfected cells (4)
- □ Human protein expression in HIV-1 susceptible tissues (1)
- □ Similarity of the two proteins in terms of (4)
  - Cellular location
  - Molecular process
  - Molecular function
  - Sequence



- □ ELM-ligand feature (1)
- □ Human PPI interactome features (8)
  - □ Similarity of HIV-1 protein to human protein’s interaction partner (5)
  - □ Topological properties of human interaction graph (3)

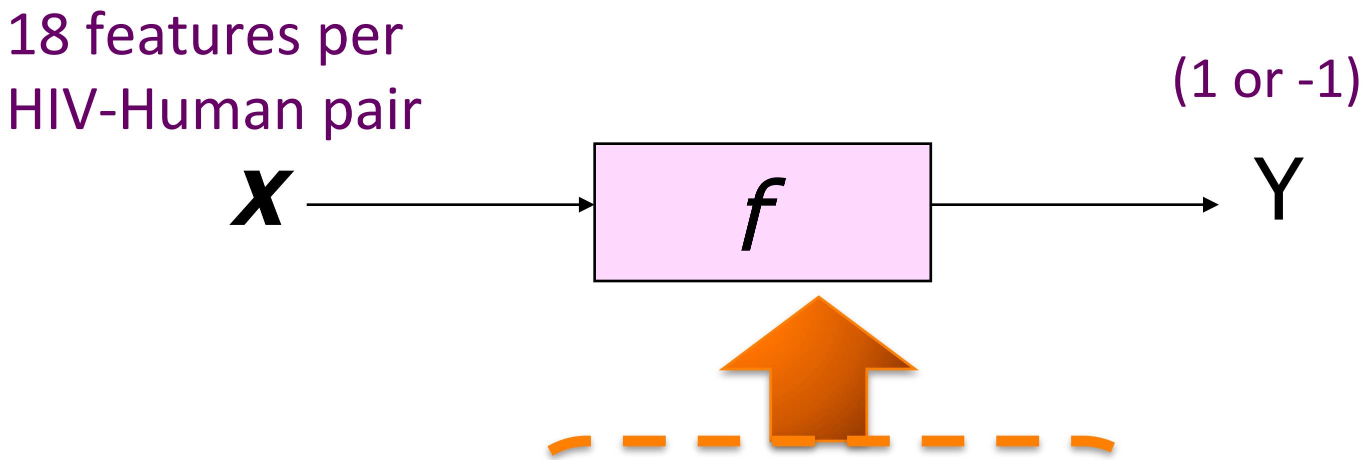
Yanjun Qi / UVA



[Y. Qi, et al, Bioinformatics 2010] [Y. Qi, et al, Proteomics 2009]

Label Complexity: Auxiliary "Partial" Labels Y'

→ Improve with multiple tasking and semi-supervised learning



18 features per HIV-Human pair

(1 or -1)

Positive Y	Partial Positive Y'	Remaining Y?
~200	~2000	~350,000
Expert annotated	Literature Extracted	

- Highly skewed class distribution (much more non-interacting pairs than interacting pairs)

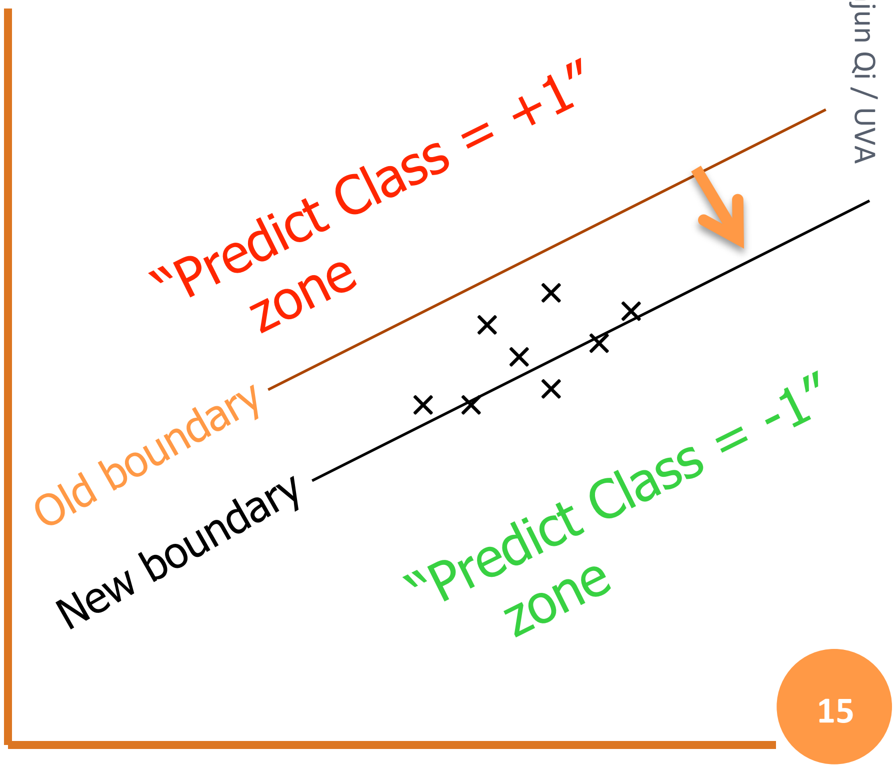
[Y. Qi, et al, Bioinformatics 2010] [Y. Qi, et al, Proteomics 2009]

Method: How to Utilize "Partial" Labels Y' ?

- **Multi-Tasking**
  - **Supervised Classification (using Y)**
  - **Auxiliary Task (using Y')**

✓Main Task: a candidate pair interacts OR not ?

✓Auxiliary Task: e.g. a pair is more likely than random pairs to interact OR not ?



× denotes Y'

[Y. Qi, et al, Bioinformatics 2010] [Y. Qi, et al, Proteomics 2009]

Method: Main Classification + Three Possible Auxiliary Tasks

To Optimize : $$\sum_{i=1}^{L} \ell(f(x_i), y_i) + \lambda \;\; \text{Loss (Auxiliary Task)}$$

Auxiliary task added as a regularizer on the supervised main task

Main: MLP classification	$\sum_{i=1}^{L} \ell(f(x_i), y_i) = \sum_{i=1}^{L} \max(0, 1 - y_i f(x_i)).$
Auxiliary (1): SMLC classification	$\text{Loss (Auxiliary Task)} = \sum_{j=L+1}^{L+U} \max(0, 1 - y'_j g(x_j))$
Auxiliary (2): SMLR pairwise ranking	$\text{Loss(Aux.)} = \sum_{p \in P} \sum_{n \in N} \max\left(0, 1 - f(x_p) + f(x_n)\right)$
Auxiliary (3): SMLE embedding	$Loss(Aux.) = \sum_{i,j=1}^{L+U} L(f(x_i), f(x_j), W_{ij})$

[Y. Qi, et al, Bioinformatics 2010] [Y. Qi, et al, Proteomics 2009]

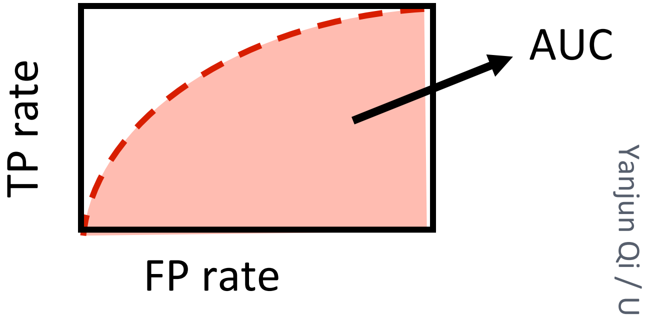
Yanjun Qi / UVA

Evaluation: Performance Comparison

- Improved performance to Random Forest classifier

METHOD	AUC 50	AUC
SMLR	**0.310**	**0.919**
RF-P	0.230	0.896
MLP-P	0.229	0.893

ROC curve



Yanjun Qi / UVA

- Validation and confirmed by multiple recent available functional assay related to HIV (siRNA data & Virion data)

- Extra: similar framework applied to look for human protein partners for receptor proteins
  - Five of our predictions were chosen for experimentally tests and three were verified ➔ 3 out of 5
  - If purely random chosen ➔ 1 out of ~20,000

[Y. Qi, et al, Bioinformatics 2010]   [Y. Qi, et al, Proteomics 2009]

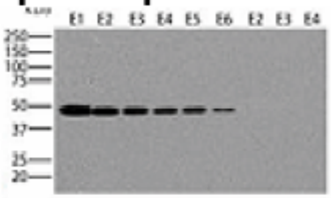
Evaluation: Experimental Validation of Predicted PPI wrt Human Membrane Receptors

➔ **(Help biologist prioritize potential interaction pairs)**

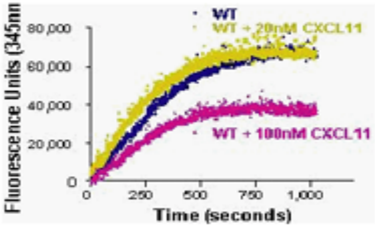
- Five of our top predictions were chosen for experimentally tests and three were verified
  - EGFR with HCK (pull-down assay)
  - EGFR with Dynamin-2 (pull-down assay)
  - RHO with CXCL11 (functional assays, fluorescence spectroscopy, docking)

  - Experiments @ U.Pitt School of Medicine

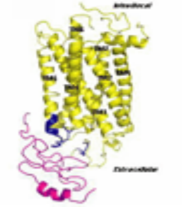
co-immuno-precipitation



functional assay



docking



Details in the paper

Y. Qi, et al Proteomics2009

THIS TALK COVERS

	Project Topic	Complexity	HOW ?
I	Protein interaction identification	Y	Training with auxiliary labels
II	Protein structure prediction	X & Y	Unified feature learning for multiple related tasks
III	Biomedical text mining	X	Add semi-supervision on features
IV	Conditional dependency graph among Genes / TFs	X	Model data example with feature interactions

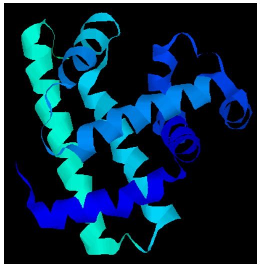
Target Problem

To Save Cost ➔ very time-consuming and expensive to measure protein structures

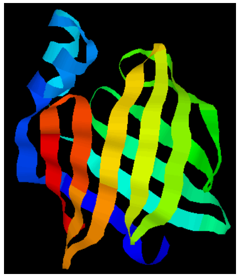
PROTEIN SEQUENCE ➔ STRUCTURAL SEGMENTS

- Input X : Primary sequence

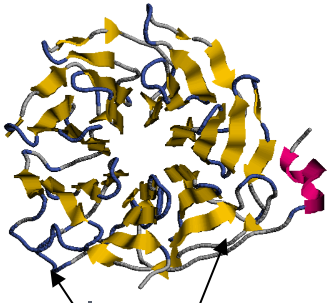
MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE



helices



strands



loops

Yanjun Qi / UVA

- Output Y :
  - Secondary structure (SS)
  - Solvent accessibility (SAR)
  - Coiled coil regions (CC)
  - DNA binding residues (DNA)
  - Transmembrane topology (TM)
  - Signal peptide (SP)
  - Protein binding residue detection (PPI)
  - ……

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Target Problem

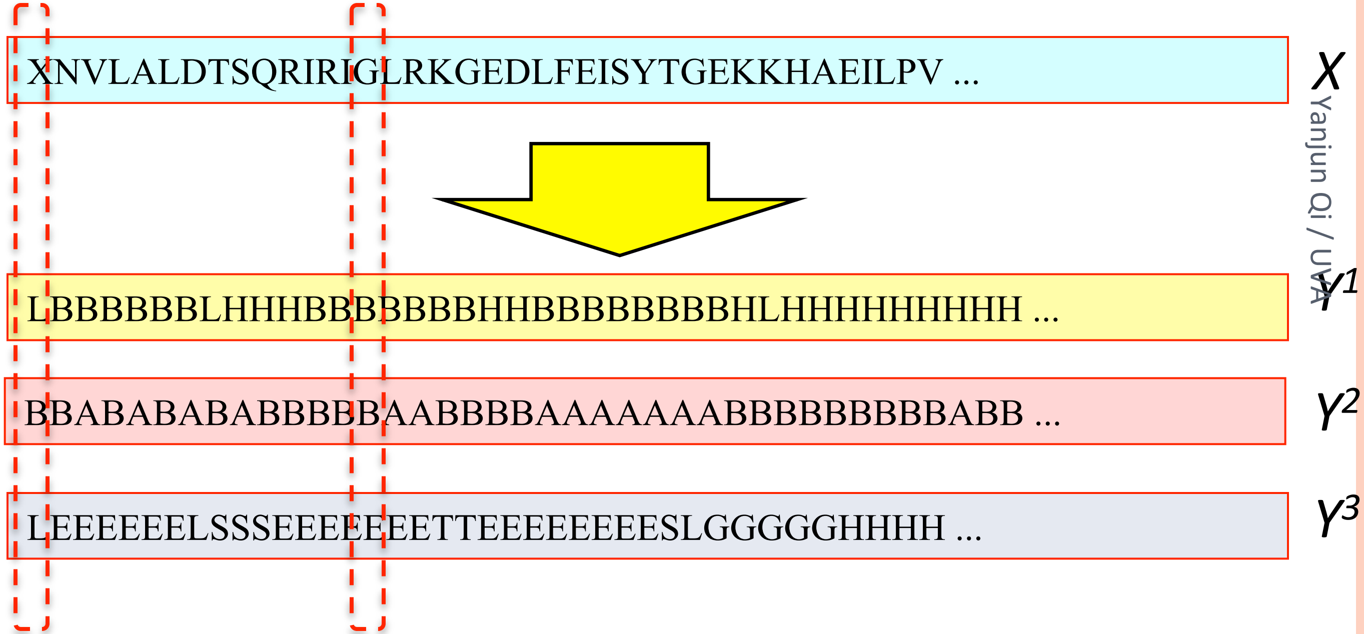
- ✓INPUT: A STRING OF AMINO ACIDS (AA)
- ✓OUTPUT: A STRING OF CLASS LABELS (OF AA)

Multiple Targets:

Secondary structures

Solvent accessibility

……



XNVLALDTSQRIRIGLRKGEDLFEISYTGEKKHAEILPV ...

LBBBBBBLHHHBBBBBBBHHBBBBBBBBBHLHHHHHHHHH ...

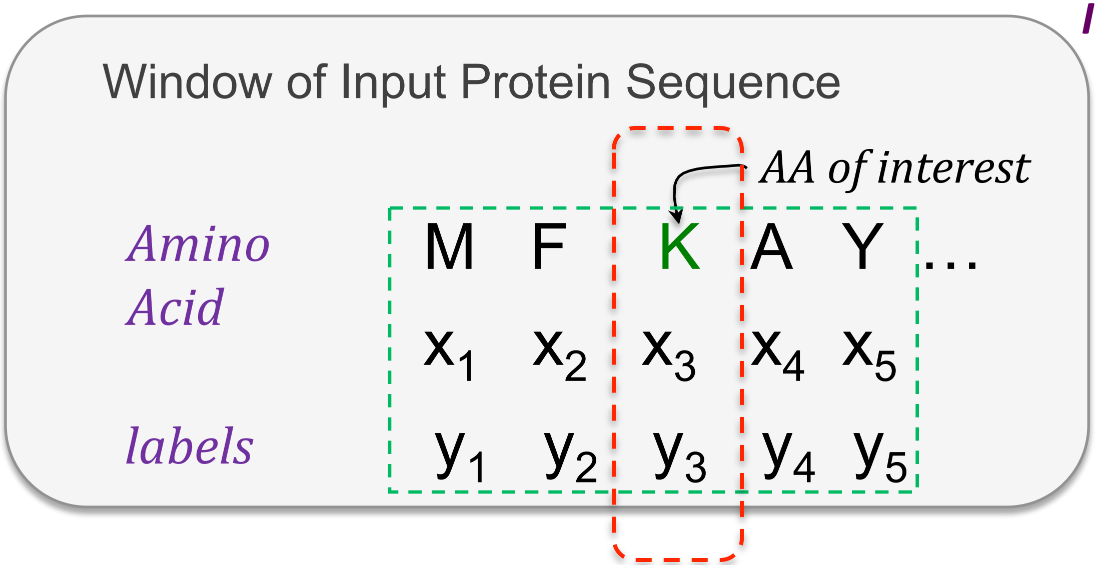
BBABABABABBBBBAABBBBAAAAAAABBBBBBBBBABB ...

LEEEEEELSSSEEEEEEETTEEEEEEEEESLGGGGGHHHH ...

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Essentially Sequence Labeling/Tagging Tasks

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM14, ECIR14



Window of Input Protein Sequence

AA of interest

Amino Acid M F K A Y …

x_1 x_2 x_3 x_4 x_5

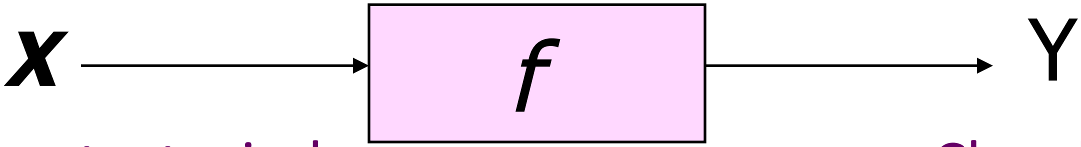
labels y_1 y_2 y_3 y_4 y_5

Yanjun Qi / UVA

+

- **Labeling each residue amino acid (AA) using its context windows:**

Using task “SS” as one example:



Each AA + its context window

$x=(x_{1,\dots,}\ x_5)$

Class label in terms of “SS” for current AA

$y=y_3$

Previous systems : Issue (1)

- Previous approaches focus on one task at a time
- Tasks exhibit strong inter-task dependencies, e.g.
  - ✓ Most transmembrane protein segments are alpha helice
  - ✓ Signal peptide prediction can be viewed as prediction of a particular type of transmembrane segment

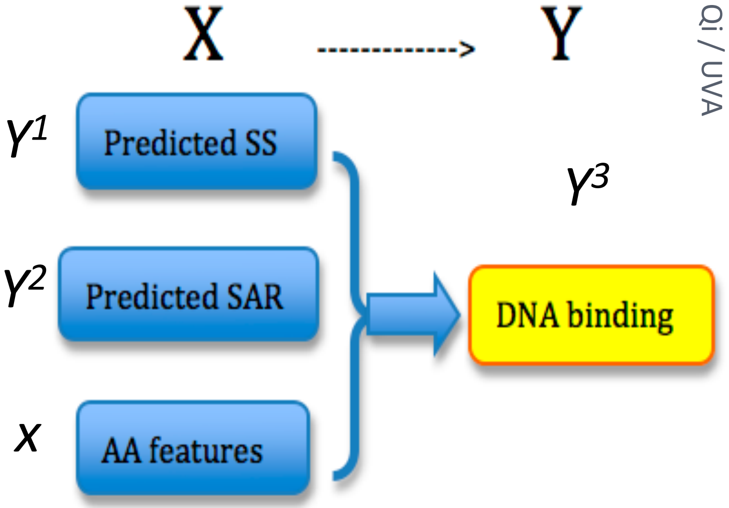
➔ Improve with multiple task learning

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Previous systems : Issue (2)

- Previous work makes use of these dependencies in a pipelined fashion,
  - ✓ Hand-craft feature engineering for each task
  - ✓ Errors from one classifier get propagated to downstream classifiers

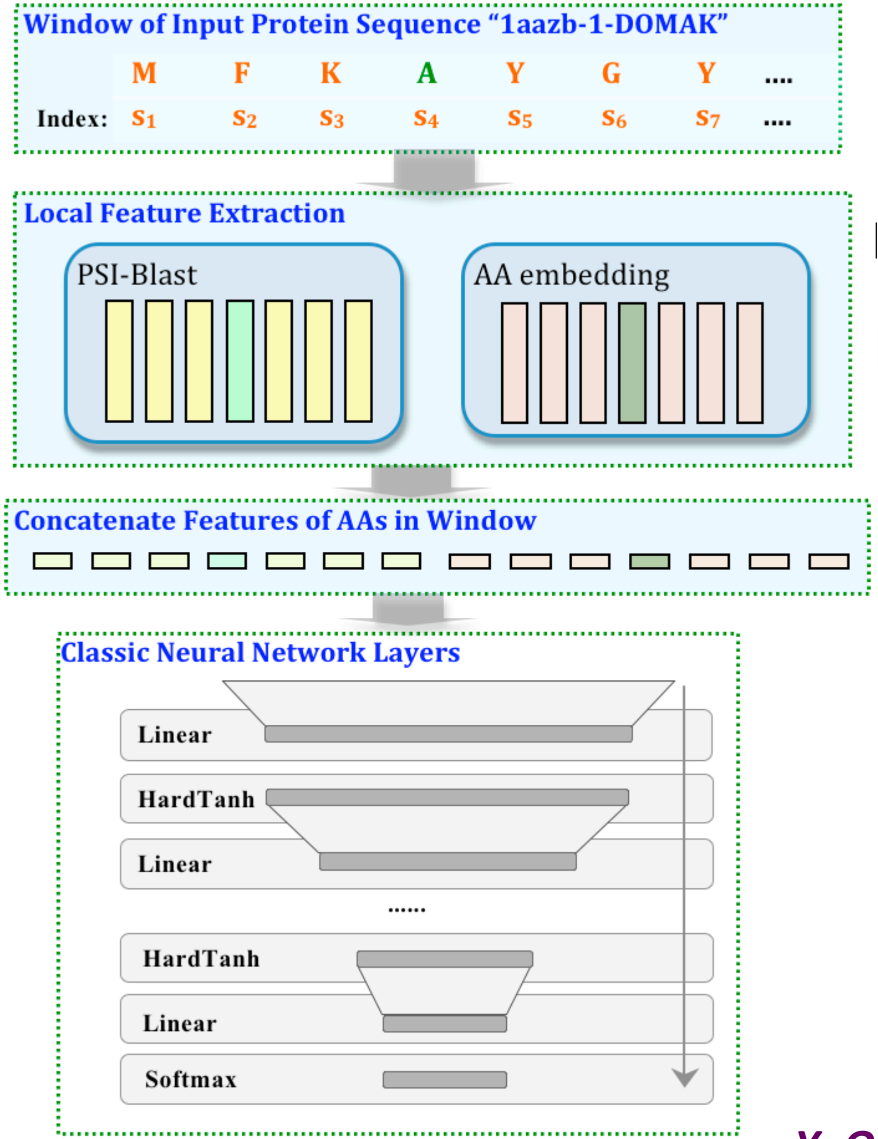
→ Improve with feature / representation learning



Yanjun Qi / UVA

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Method: Adapt deep CNN for Each Sequence Modeling Task


Window of Input Protein Sequence "1aazb-1-DOMAK"
M F K A Y G Y ....
Index: S1 S2 S3 S4 S5 S6 S7 ....
Local Feature Extraction
PSI-Blast
AA embedding
Concatenate Features of AAs in Window
Classic Neural Network Layers
Linear
HardTanh
Linear
......
HardTanh
Linear
Softmax

Learn Feature Representation for each amino acid

Learn Representation for each segment around current position

Learning function to map from representation to TAG/class label

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Method: Multi-Tasking to train a single, joint model for Ten tasks

XNVLALDTSQRIRIGLRKGEDLFEISYTGEKKHAEILPV ... X

Multiple Targets:

Secondary structures LBBBBBBLHHHBBBBBBBBHHBBBBBBBBBHLHHHHHHHHH ... Y^1

Solvent accessibility BBABABABABBBBBAABBBBAAAAAAABBBBBBBBBABB ... Y^2

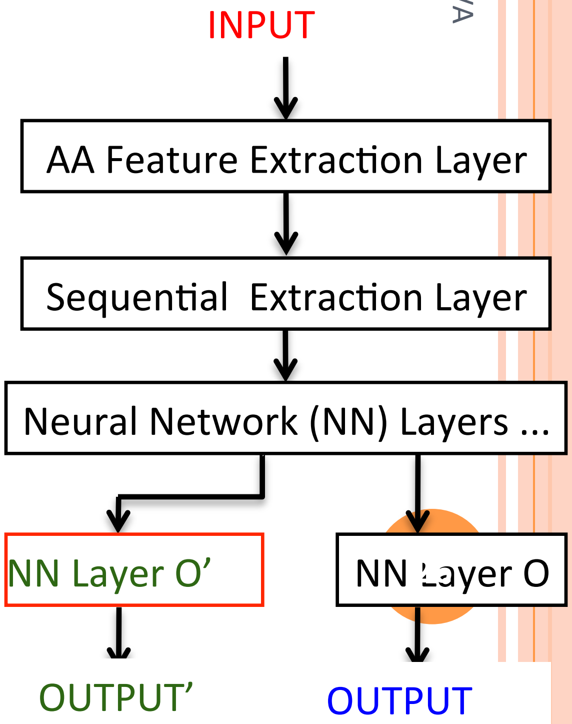
...... LEEEEEELSSSEEEEEEEETTEEEEEEEEESLGGGGGHHHH ... Y^3

Parameters to learn (assuming total T tasks)

$$\Theta_t = \{W, L^1, L^2, \ldots, L^{L-1}, L_t^L\}$$

By optimize

$$\sum_{t=1}^{T} \sum_{n_t=1}^{N_t} E_t(\Theta_t, x_{n_t}, y_{n_t})$$


INPUT
AA Feature Extraction Layer
Sequential Extraction Layer
Neural Network (NN) Layers ...
NN Layer O'
NN Layer O
OUTPUT'
OUTPUT

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Method: Backpropagation & Stochastic Gradient Descent

- **Backpropagation**
  - Using backward recurrence it jointly optimizes all parameters
  - Requires all activation functions to be differentiable
  - Enables flexible design in deep model architecture
  - Gradient descent is used to (locally) minimize objective:

$$W^{k+1} = W^{k} - \eta \frac{\partial L}{\partial W^{k}}$$

- **Stochastic Gradient Descent (SGD)** (first-order iterative optimization)
  - SGD is an **online learning** method
  - Approximates “true” gradient with a gradient at one data point
  - Attractive because of low computation requirement
  - Rivals **batch learning** (e.g., SVM) methods on large datasets

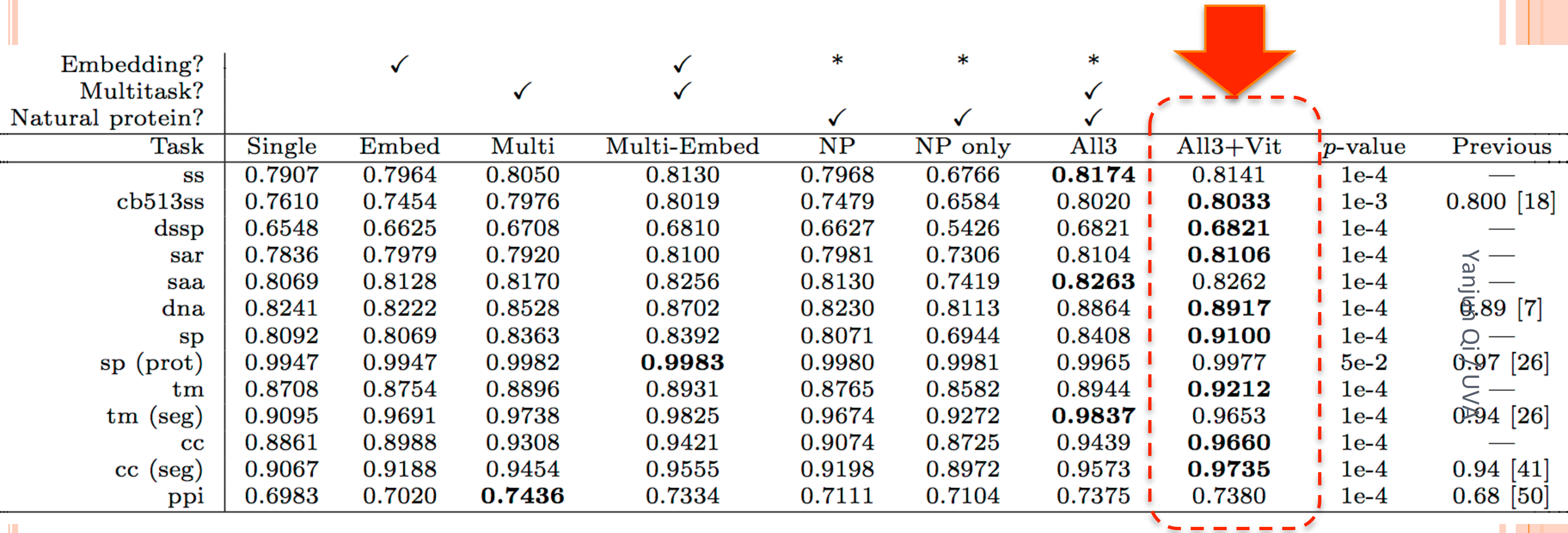
Y. LeCun et al. 1998. Efficient BackProp.

Olivier Bousquet and Ulrike von Luxburg. 2004. Stochastic Learning.

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Evaluation: Summary of Performance Comparison tasks

Multitask + Embedding + Pretrain + Viterbi



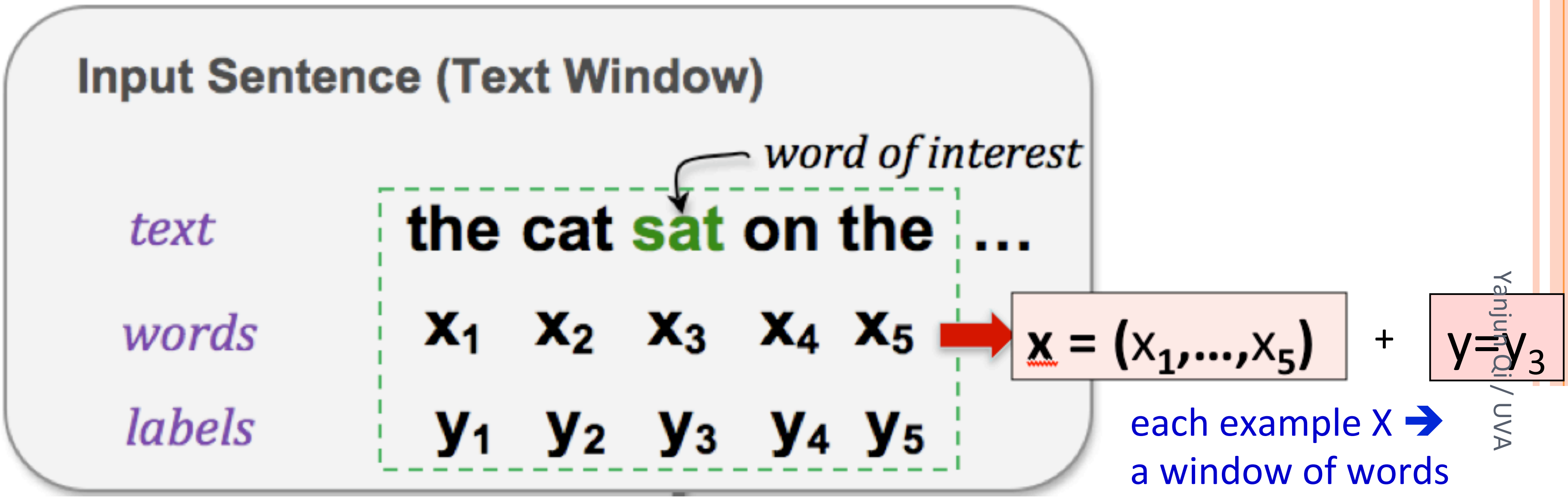
Embedding?		✓		✓	*	*	*			
Multitask?			✓	✓			✓			
Natural protein?					✓	✓	✓			
Task	Single	Embed	Multi	Multi-Embed	NP	NP only	All3	All3+Vit	*p*-value	Previous
ss	0.7907	0.7964	0.8050	0.8130	0.7968	0.6766	**0.8174**	0.8141	1e-4	—
cb513ss	0.7610	0.7454	0.7976	0.8019	0.7479	0.6584	0.8020	**0.8033**	1e-3	0.800 [18]
dssp	0.6548	0.6625	0.6708	0.6810	0.6627	0.5426	0.6821	**0.6821**	1e-4	—
sar	0.7836	0.7979	0.7920	0.8100	0.7981	0.7306	0.8104	**0.8106**	1e-4	—
saa	0.8069	0.8128	0.8170	0.8256	0.8130	0.7419	**0.8263**	0.8262	1e-4	—
dna	0.8241	0.8222	0.8528	0.8702	0.8230	0.8113	0.8864	**0.8917**	1e-4	0.89 [7]
sp	0.8092	0.8069	0.8363	0.8392	0.8071	0.6944	0.8408	**0.9100**	1e-4	—
sp (prot)	0.9947	0.9947	0.9982	**0.9983**	0.9980	0.9981	0.9965	0.9977	5e-2	0.97 [26]
tm	0.8708	0.8754	0.8896	0.8931	0.8765	0.8582	0.8944	**0.9212**	1e-4	—
tm (seg)	0.9095	0.9691	0.9738	0.9825	0.9674	0.9272	**0.9837**	0.9653	1e-4	0.94 [26]
cc	0.8861	0.8988	0.9308	0.9421	0.9074	0.8725	0.9439	**0.9660**	1e-4	—
cc (seg)	0.9067	0.9188	0.9454	0.9555	0.9198	0.8972	0.9573	**0.9735**	1e-4	0.94 [41]
ppi	0.6983	0.7020	**0.7436**	0.7334	0.7111	0.7104	0.7375	0.7380	1e-4	0.68 [50]

Ten different tasks

- ✓ All reach state-of-the-art performance
  - ▪ Unsupervised pretrain + Supervised pretraining (with large tasks)
- ✓ One unified framework for all task
  - ▪ Simple + powerful !
- ✓ No need for task-specific feature engineering

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Similar Models Applied Successfully on NLP Tagging Tasks


Input Sentence (Text Window)
word of interest
text
the cat sat on the ...
words
x_1 x_2 x_3 x_4 x_5
labels
y_1 y_2 y_3 y_4 y_5
$x = (x_1,\ldots,x_5)$
+
$y = y_3$

each example X → a window of words

- Similar as natural language processing (NLP) tagging tasks (e.g. part-of-speech, name entity recognition)
- Similar deep models have achieved state-of-art results on NLP tagging of English, German, Chinese

Y. Qi, et al, PLoS ONE (2012), ICDM10, CIKM10, SDM 14, ECIR 14

Yanjun Qi / UVA

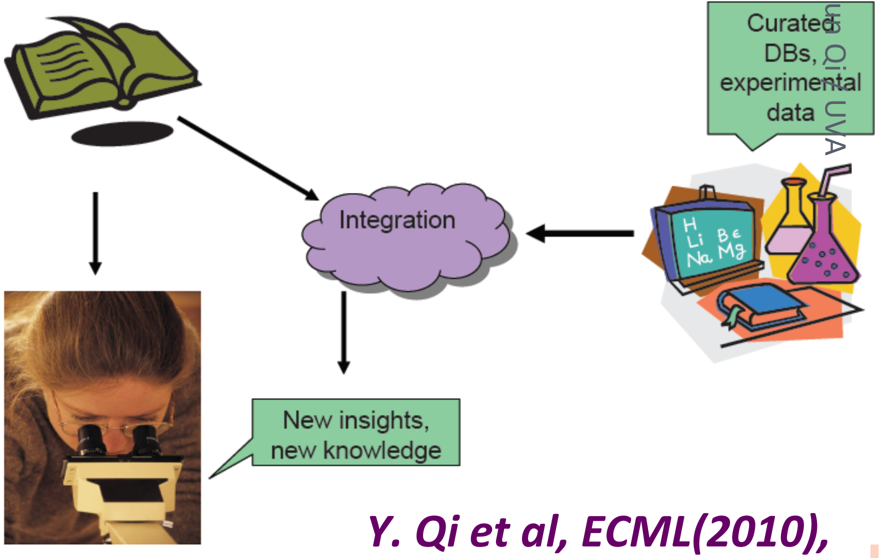
THIS TALK COVERS

	Project Topic	Complexity	HOW ?
I	Protein interaction identification	Y	Training with auxiliary labels
II	Protein structure prediction	X & Y	Unified feature learning for multiple related tasks
III	Biomedical text mining	X	Add semi-supervision on features
IV	Conditional dependency graph among Genes / TFs	X	Model data example with feature interactions

Why Text Mining for Biomedicine ?

- Data Situation
  - MEDLINE: over 70 million queries every month and about 20 million publications
  - new terms (genes, proteins, chemical compounds, drugs) and discoveries constantly created/added in
    - **Impossible to annotate manually**

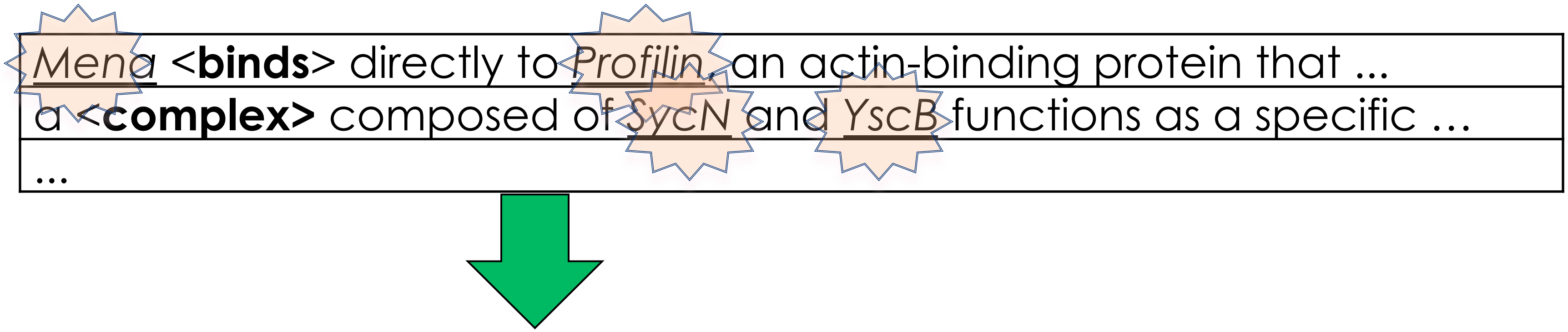
- Linking text to bio-databases and ontologies is crucial, for
  - Efficient access and discovery of facts and events in biosciences



Y. Qi et al, ECML(2010), SDM(2011), TRECMED(2012),

➔ Need text mining to (help) analyze / organize biomedical literature

Two Benchmark Tasks



Mena <**binds**> directly to *Profilin*, an actin-binding protein that …

a <**complex**> composed of *SycN* and *YscB* functions as a specific …

…

Protein	Protein	Relation	Reference
Mena	Profilin	bind to	PubMed
SycN	YscB	complex	PubMed

- Related Tasks
  - Protein Name Recognition
  - Protein Interaction Event Recognition

Y. Qi et al, ECML(2010), SDM(2011), TRECMED(2012),

➔ **Many Similar Tasks**

- **Bio-Entity recognition (e.g. chemical terms, disease names,)**
- **Bio-Relational extraction (e.g. genetic interaction, disease to phenotype)**

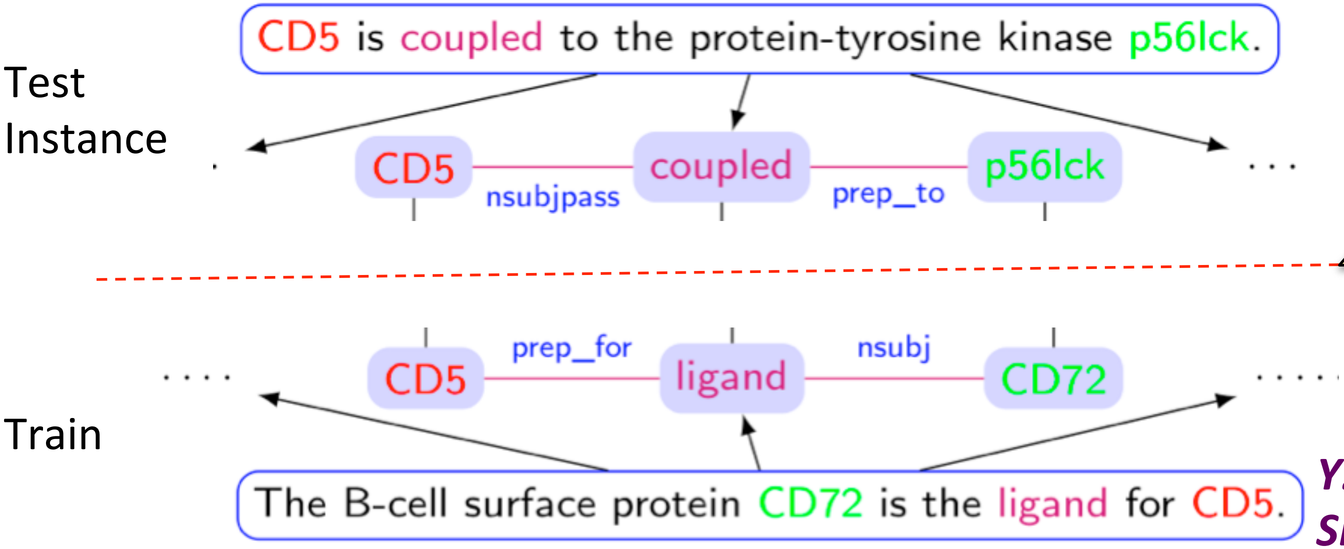
Challenges

How to improve current approaches by learning from unlabeled examples X* (e.g. Pubmed articles) ?

- Annotated training sets are small
  - Hardly cover words in vocabulary (~2 million in PubMed)
- Millions of Pubmed articles freely available

- To design learning methods able to measure semantic similarity between words or word sequences
  - Rigid symbolic matching could not capture such similarity

Yanjun Qi / UVA



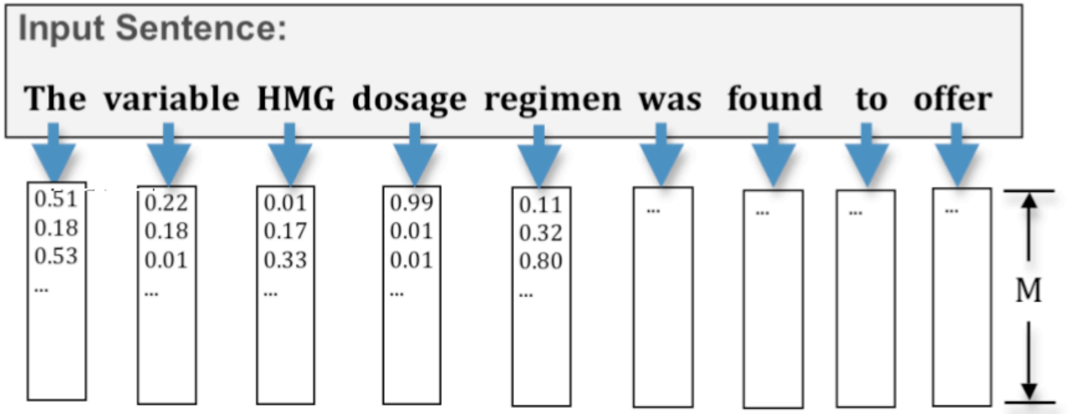
Test Instance
CD5 is coupled to the protein-tyrosine kinase p56lck.
CD5
nsubjpass
coupled
prep_to
p56lck
Train
CD5
prep_for
ligand
nsubj
CD72
The B-cell surface protein CD72 is the ligand for CD5.

To search for similar word sequences in training for a new test sequence

Y. Qi et al, ECML(2010), SDM(2011), TRECMED(2012

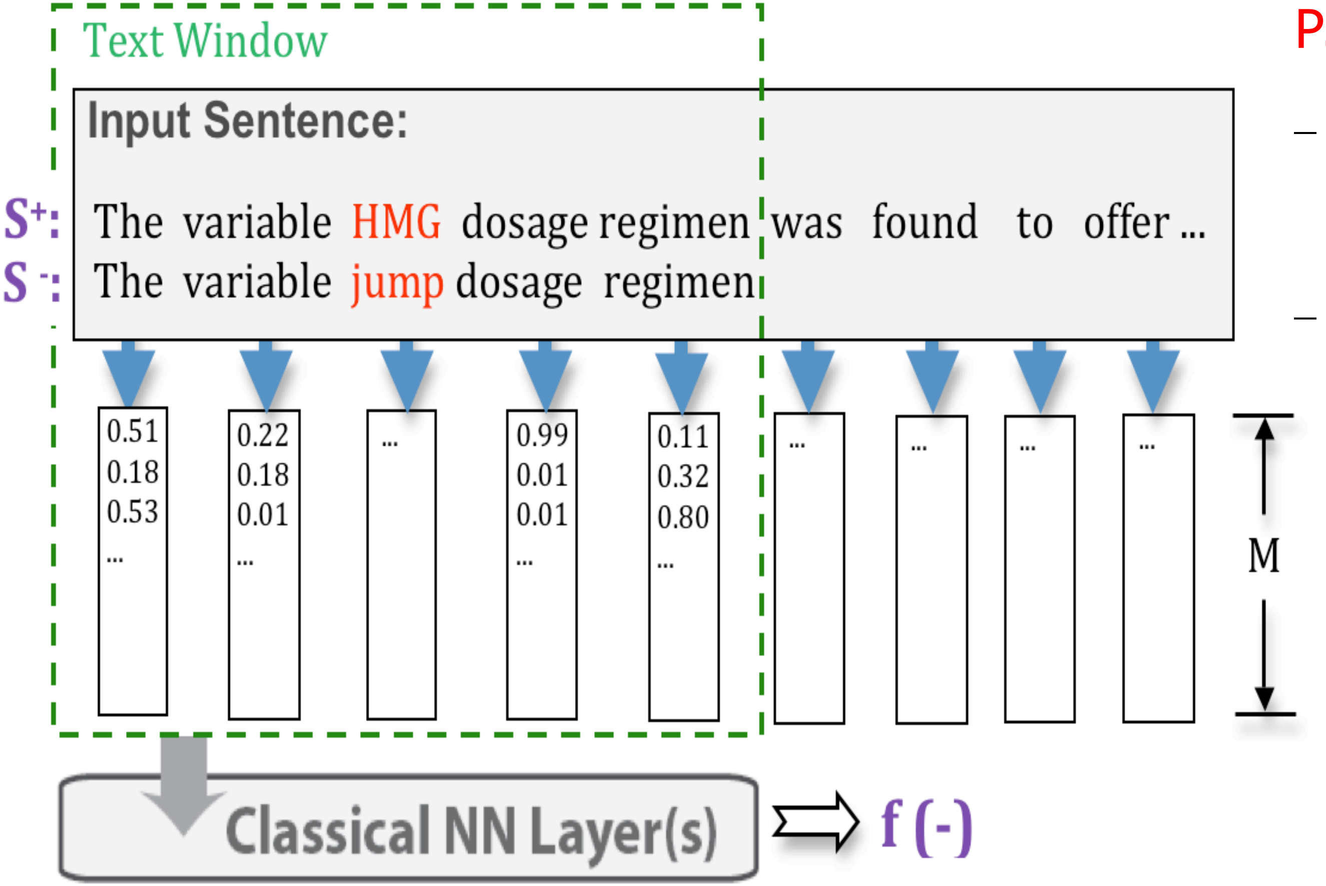
Learn Word Representation Reflecting Semantic Similarity

- Learn to embed each word into a vector of real values (with dimensionality M)
  - Based on unlabeled data (i.e. PubMed abstracts 1995-2009, ~1.3G word tokens, ~4.5M abstracts)
  - Semantically similar words have closer embedding representations



Y. Qi, et al, NIPS(2009), ICDM(2009), ECML(2010), CIKM(2011), SDM(2011), TRECMED(2012), NIPS(2012), ECML(2012), SDM (2014)

Local Embedding Based on Pattern of Short Text Window



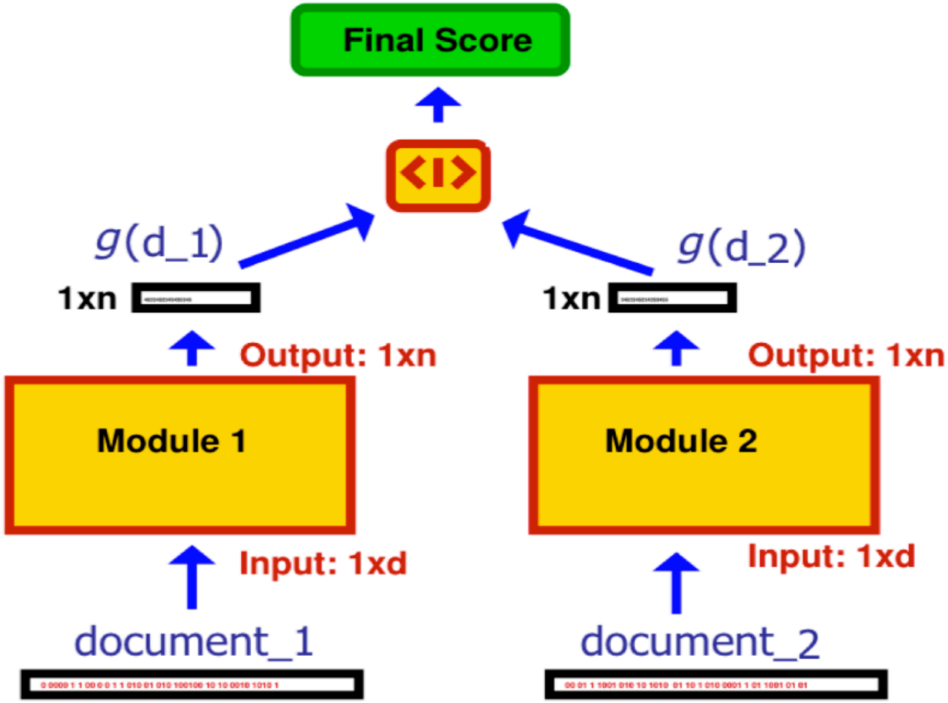
Pseudo supervised signals

- Positive examples: Text window extracted from unlabeled corpus randomly
- Negative examples: Text window with middle word replaced by a random word

Y. Qi et al, ECML(2010), SDM(2011), TRECMED(2012)

- Build a paiwise ranking task to train word embedding (first layer in deep neural network)
  - f(-) measures how likely a word segment exist in Pubmed ?
  - Pairwise rank loss to optimize: $\sum \max\left(0, 1 - f(s^{+}) + f(s^{-})\right)$

Global Embedding using Similarity between Text Documents



- Pseudo supervised signals by splitting each Pubmed abstract into two documents (each with half)
  - Similar if from the same
  - Dissimilar otherwise

- g(-) ➔ learned representation of each text document
  - first layer of g(-) maps to “global” word embedding
  - Each document is represented as “bag-of-words”
- Learning g(-) by forcing g(-) of two documents
  - with similar meanings to have closer representations,
  - with different meanings to be dissimilar

Y. Qi et al, ECML(2010), CIKM(2011), SDM(2011), TRECMED(2012),

Results: Nearest Words of Sample Query Word

Query	Local Embed	Global Embed
protein	ligand, subunit, receptor, molecule	proteins, phosphoprotein, isoform,
medical	surgical, dental, preventive, reconstructive	hospital, investigated, research, urology
interact	cooperate, compete, interfere, react	interacting, member, associate, ligand
immunoprecipitation	co-immunoprecipitation, EMSA, autoradiography, RT-PCR	coexpression, two-hybrid, phosphorylated, tbp

Y. Qi, et al, NIPS(2009), ICDM(2009), ECML(2010), CIKM(2011), SDM(2011), TRECMED(2012), NIPS(2012), ECML(2012), SDM (2014)

Results: Performance

- Achieved the state-of-the-art performance (by using large amount of unlabeled data from Pubmed)
  - With word features only
  - Added on single base classifier (string kernel + SVM)

- Previous best systems used complex combination of many classifiers with many more linguistic features, dictionaries, and etc

- Semi-supervision **IMPROVES** both benchmark tasks
  - Bio-Entity tagging (genes, proteins, etc)

  - Protein-Protein Interaction (PPI) event extraction

Y. Qi, et al, NIPS(2009), ICDM(2009), ECML(2010), CIKM(2011), SDM(2011), TRECMED(2012), NIPS(2012), ECML(2012), SDM (2014)

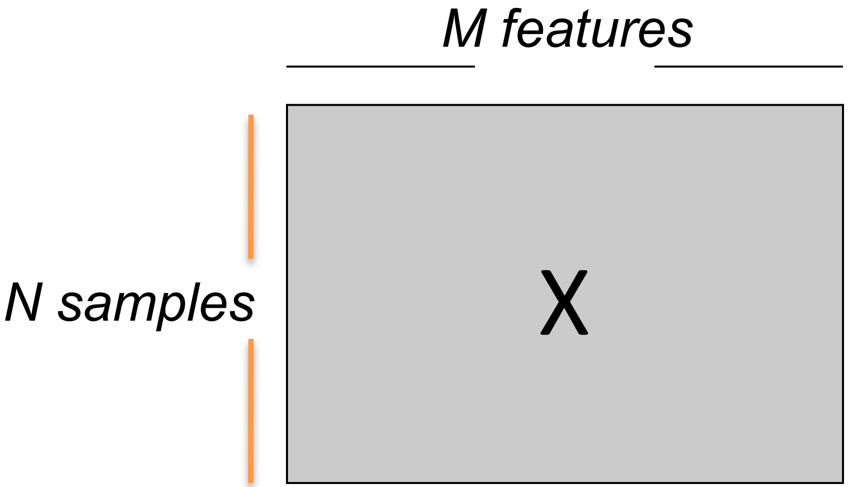
THIS TALK COVERS

	Project Topic	Complexity	HOW ?
I	Protein interaction identification	Y	Training with auxiliary labels
II	Protein structure prediction	X & Y	Unified feature learning for multiple related tasks
III	Biomedical text mining	X	Add semi-supervision on features
IV	Conditional dependency graph among Genes / TFs	X	Model data example with feature interactions

MODEL FEATURE DEPENDENCY TO GET BETTER FEATURES

- Feature variables have correlations or high-order conditional dependency relationship
  - E.g. genes work with other genes together to affect certain disease

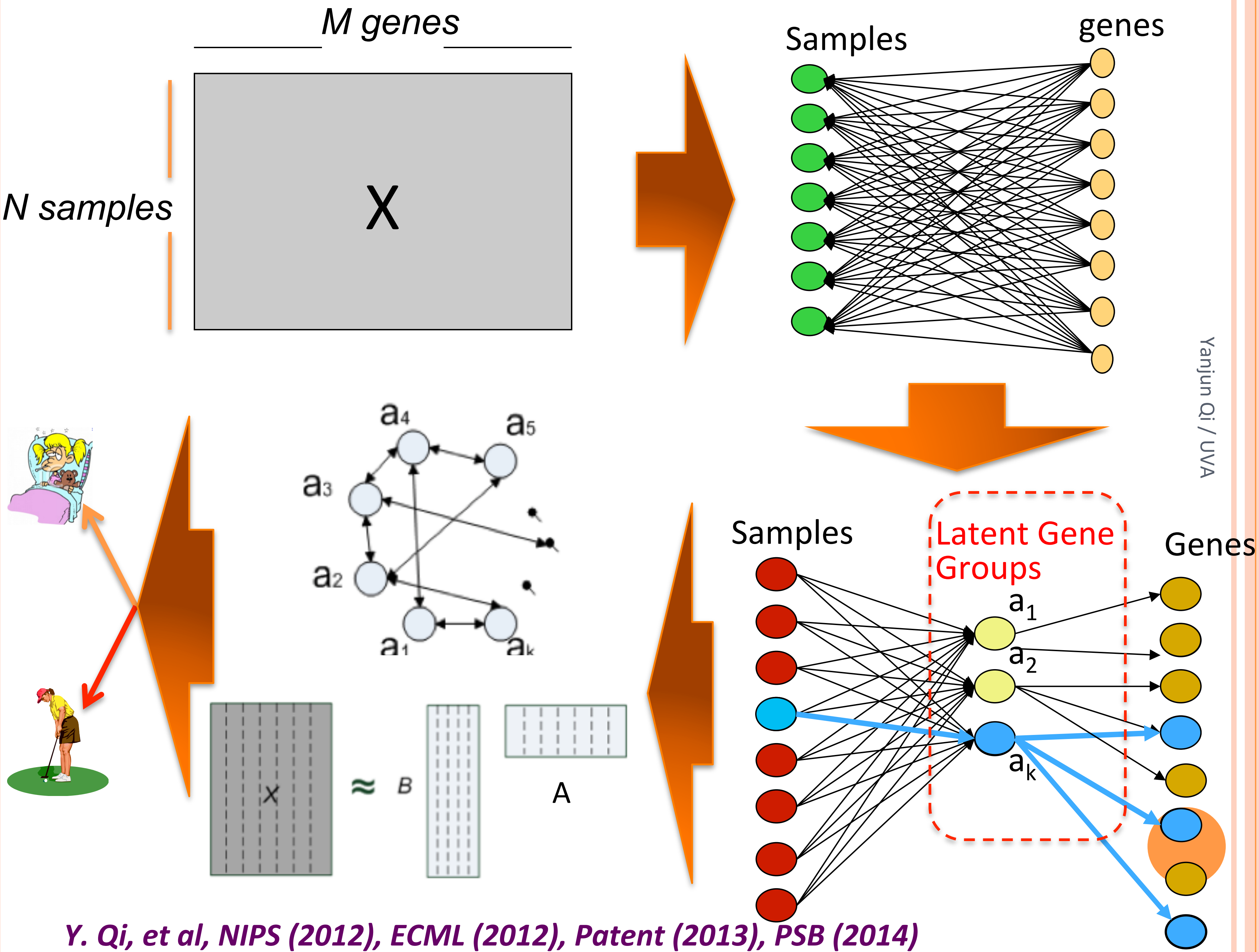
Yanjun Qi / UVA



Hypothesis:

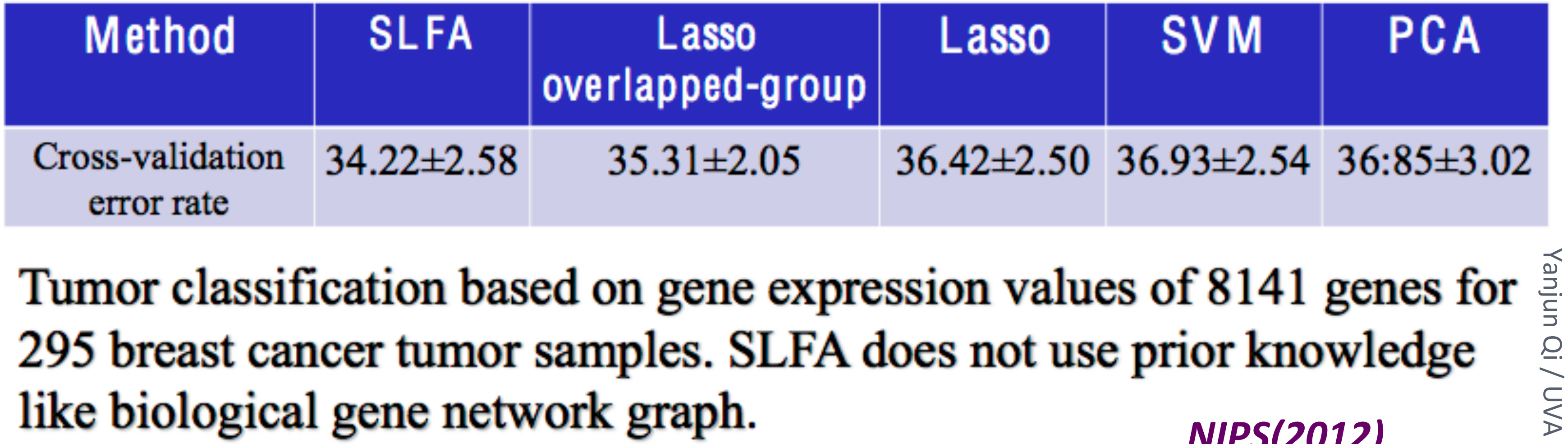
➔ May model samples better if considering feature dependencies / interactions

Y. Qi, et al, NIPS (2012), ECML (2012), Patent (2013), PSB (2014)



Y. Qi, et al, NIPS (2012), ECML (2012), Patent (2013), PSB (2014)

Task: Learning Dependency between Hidden Feature Groups



Method	SLFA	Lasso overlapped-group	Lasso	SVM	PCA
Cross-validation error rate	34.22±2.58	35.31±2.05	36.42±2.50	36.93±2.54	36:85±3.02

Tumor classification based on gene expression values of 8141 genes for 295 breast cancer tumor samples. SLFA does not use prior knowledge like biological gene network graph.

NIPS(2012)

Yanjun Qi / UVA

Same model successfully applied to learn dependency between text topics for modeling text documents

NIPS (2012)

A similar / simpler model successfully applied to learn conditional dependency between transcription factors using ENCODE data

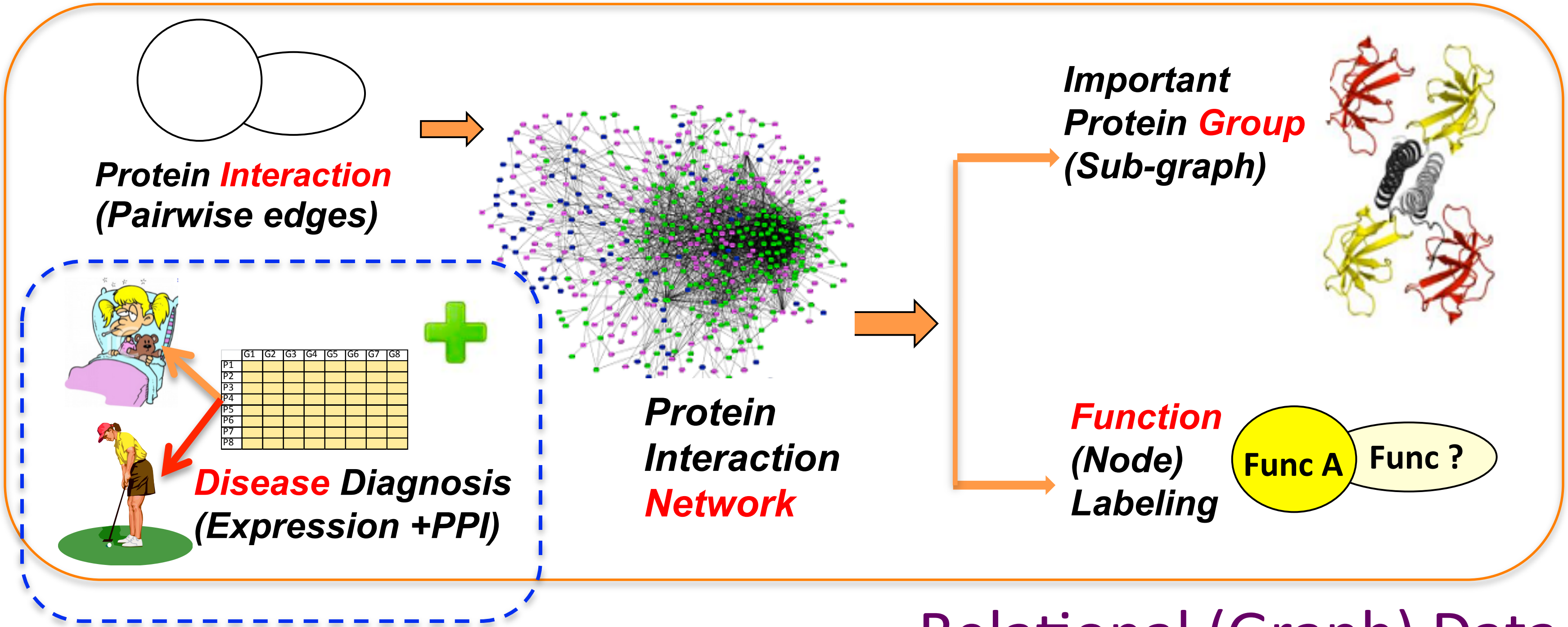
Patent (2013)

THIS TALK COVERS

	Project Topic	Complexity	HOW ?
I	Protein interaction identification	Y	Training with auxiliary labels
II	Protein structure prediction	X & Y	Unified feature learning for multiple related tasks
III	Biomedical text mining	X	Add semi-supervision on features
IV	Conditional dependency graph among Genes / TFs	X	Model data example with feature interactions

MORE NOT COVERED OF MY PROJECTS

www.cs.virginia.edu/yanjun



Protein Interaction (Pairwise edges)

Protein Interaction Network

Important Protein Group (Sub-graph)

Function (Node) Labeling

Disease Diagnosis (Expression +PPI)

	G1	G2	G3	G4	G5	G6	G7	G8
P1								
P2								
P3								
P4								
P5								
P6								
P7								
P8								

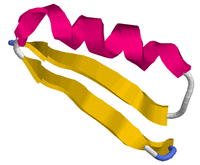
Relational (Graph) Data

Applications are diverse but methods are generic

Yanjun Qi / UVA

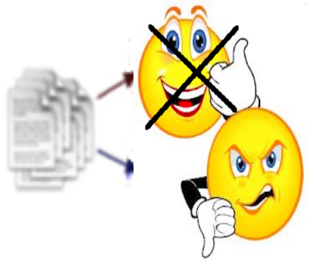
MORE NOT COVERED OF MY PROJECTS

www.cs.virginia.edu/yanjun



CCEEEEEECCCCCHHHCCC...

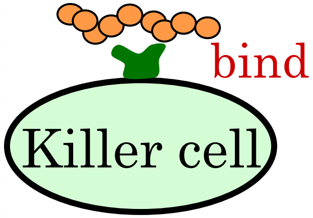
Tagging Protein Sequence



Classifying Social Text Sentiment

Retrieving Medical Records

Entity & Relation Recognition

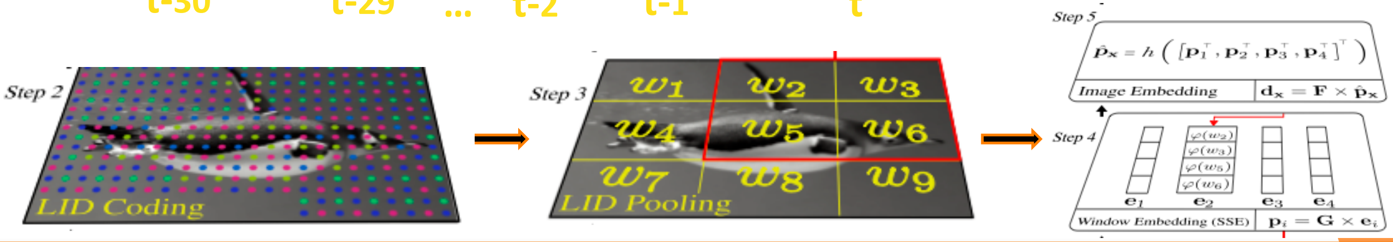


MHC binding Peptide Prediction

Sequential Data

Video segmentation; Video retrieval,

Image Classification



Multimedia Data

Applications are diverse but methods are generic

www.cs.virginia.edu/yanjun

Actively Looking for collaborations !

Contact: yanjun@virginia.edu

www.cs.virginia.edu/yanjun/