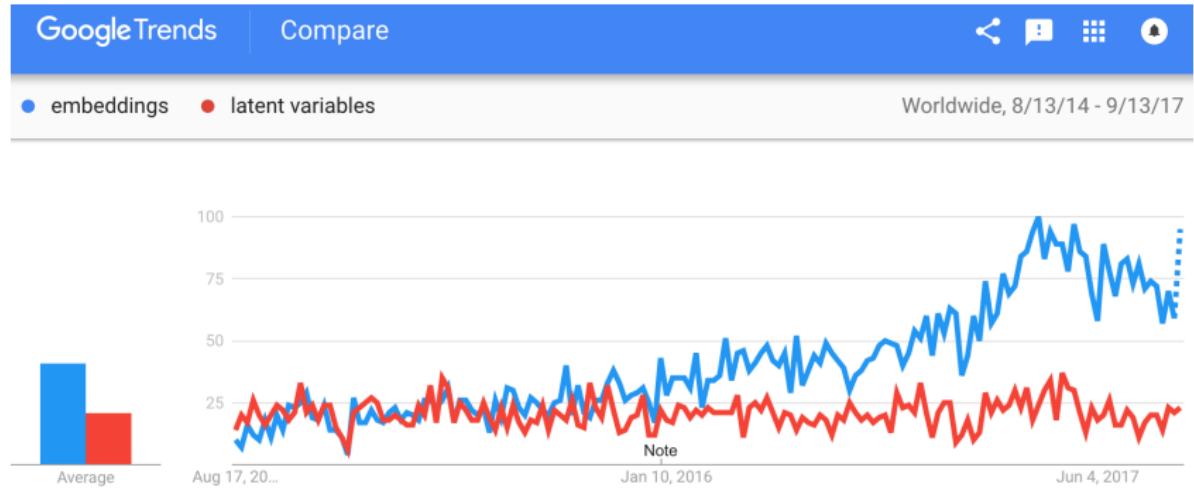


# Multi-Target Prediction Using Low-Rank Embeddings: Theory & Practice

Inderjit S. Dhillon  
UT Austin & Amazon

*ECML PKDD 2017*  
Skopje, Macedonia  
Sept 20, 2017

## Embeddings



# Outline

## 1 Motivation

## 2 Multi-Target Prediction

- Real-world Applications
- Linear Prediction
- Bilinear Prediction: Inductive Matrix Completion (IMC)
- Bilinear Prediction with Noisy Features
- Positive-Unlabeled Learning

## 3 Nonlinear Multi-Target Prediction

- Goal-Directed IMC: Two-layer Neural Network
- Deep Neural Networks

## 4 Conclusions and Future Work

# Sample Prediction Problems

## Predicting stock prices

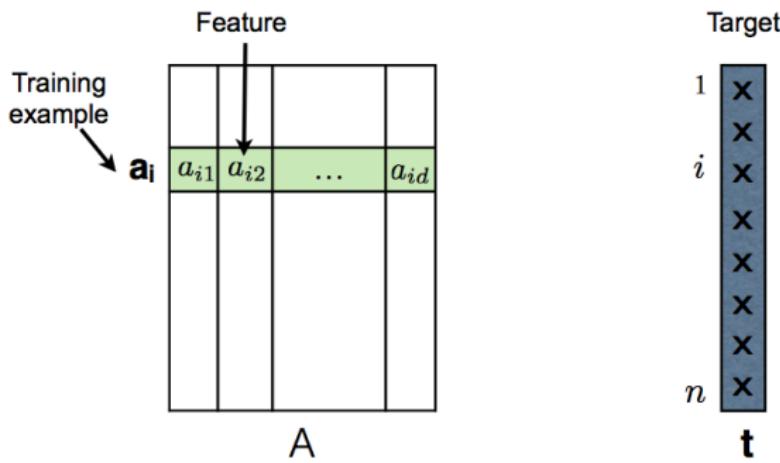


## Predicting risk factors in healthcare



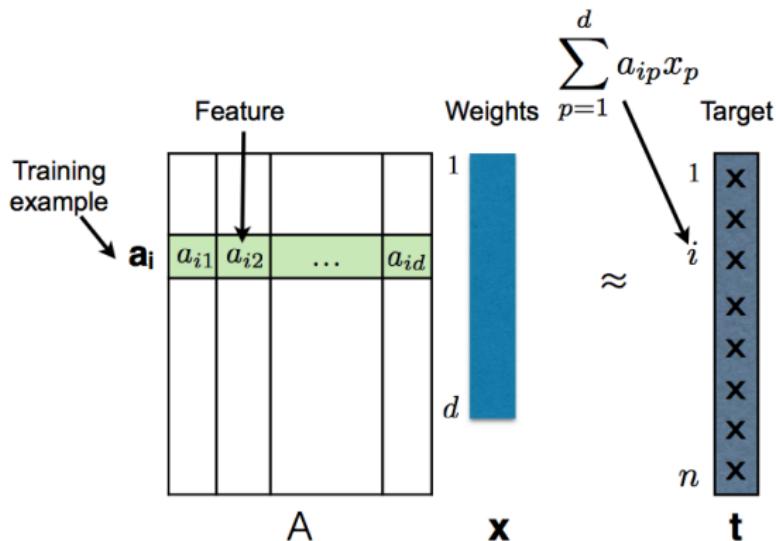
# Single-Target Regression

- Real-valued responses (target)  $t$
- Predict response for given input data (features)  $a$



# Linear Regression

- Estimate target by a linear function of given data  $\mathbf{a}$ , i.e.  $\mathbf{t} \approx \hat{\mathbf{t}} = \mathbf{a}^\top \mathbf{x}$ .

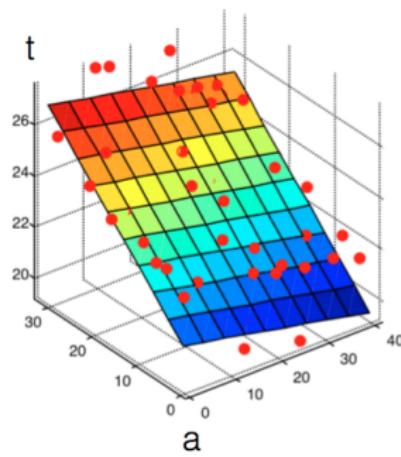
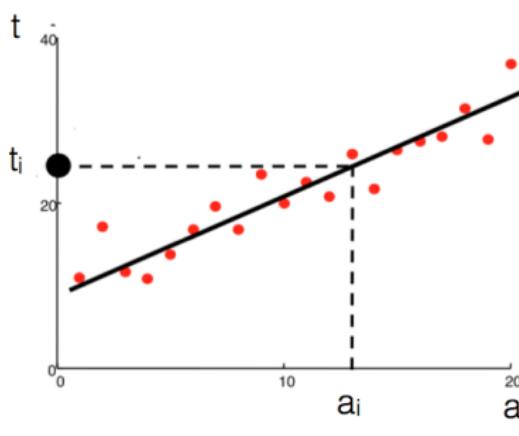


# Linear Regression: Least Squares

- Choose  $\mathbf{x}$  that minimizes

$$J_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^n (\mathbf{a}_i^T \mathbf{x} - t_i)^2$$

- Closed-form solution:  $\mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{t}$ .



# Classification

## Spam detection

Gmail ▾

COMPOSE

Inbox (8,439)

Starred

Important

Sent Mail

Drafts

Notes

Less ▾

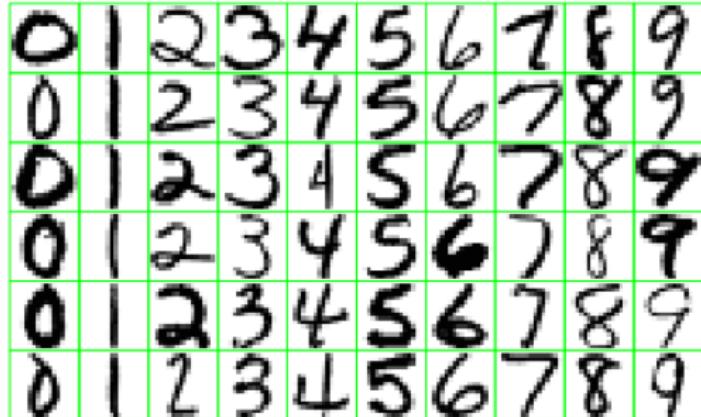
Chats

All Mail

Spam (298)

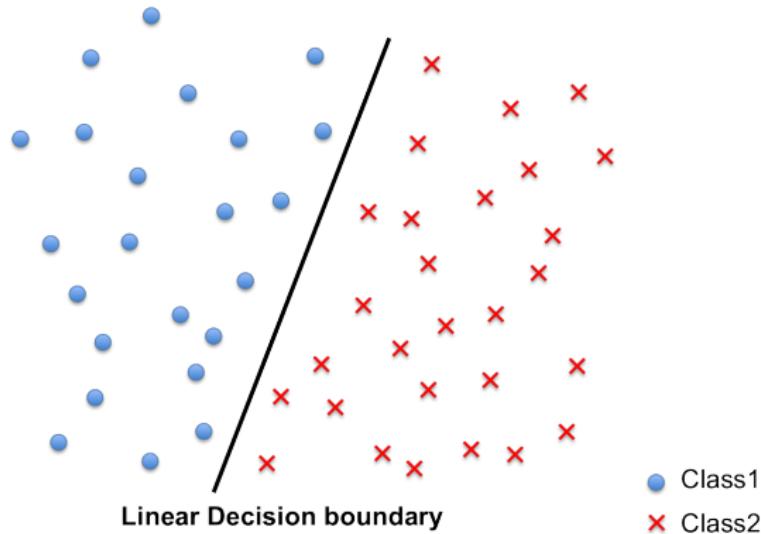
Trash

## Character Recognition



# Binary Classification

- Categorical responses (target)  $t$
- Predict response for given input data (features)  $a$
- Linear methods — decision boundary is a linear surface or hyperplane



# Linear Methods for Prediction Problems

## Regression:

- Ridge Regression:  $J_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{x} - t_i)^2 + \lambda \|\mathbf{x}\|_2^2$ .
- Lasso:  $J_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{x} - t_i)^2 + \lambda \|\mathbf{x}\|_1$ .

## Classification:

- Linear Support Vector Machines

$$J_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^n \max(0, 1 - t_i \mathbf{a}_i^\top \mathbf{x}) + \lambda \|\mathbf{x}\|_2^2.$$

- Logistic Regression

$$J_{\mathbf{x}} = \frac{1}{2} \sum_{i=1}^n \log(1 + \exp(-t_i \mathbf{a}_i^\top \mathbf{x})) + \lambda \|\mathbf{x}\|_2^2.$$

# Linear Prediction

Springer Series in Statistics

Trevor Hastie  
Robert Tibshirani  
Jerome Friedman

## The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

 Springer

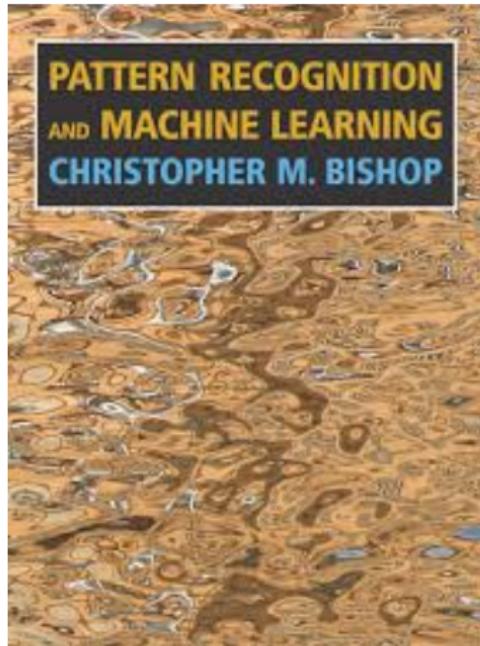
### 3 Linear Methods for Regression

3.1	Introduction	.	.	.	.
3.2	Linear Regression Models and Least Squares	.	.	.	.
3.2.1	Example: Prostate Cancer	.	.	.	.
3.2.2	The Gauss–Markov Theorem	.	.	.	.
3.2.3	Multiple Regression	.	.	.	.
	from Simple Univariate Regression	.	.	.	.
3.2.4	Multiple Outputs	.	.	.	.
3.3	Subset Selection	.	.	.	.
3.3.1	Best-Subset Selection	.	.	.	.

### 4 Linear Methods for Classification

4.1	Introduction	.	.	.	.
4.2	Linear Regression of an Indicator Matrix	.	.	.	.
4.3	Linear Discriminant Analysis	.	.	.	.
4.3.1	Regularized Discriminant Analysis	.	.	.	.
4.3.2	Computations for LDA	.	.	.	.
4.3.3	Reduced-Rank Linear Discriminant Analysis	.	.	.	.
4.4	Logistic Regression	.	.	.	.
4.4.1	Fitting Logistic Regression Models	.	.	.	.

# Linear Prediction



## 3 Linear Models for Regression

3.1	Linear Basis Function Models . . . . .
3.1.1	Maximum likelihood and least squares . . . . .
3.1.2	Geometry of least squares . . . . .
3.1.3	Sequential learning . . . . .
3.1.4	Regularized least squares . . . . .
3.1.5	Multiple outputs . . . . .
3.2	The Bias-Variance Decomposition . . . . .

## 4 Linear Models for Classification

4.1	Discriminant Functions . . . . .
4.1.1	Two classes . . . . .
4.1.2	Multiple classes . . . . .
4.1.3	Least squares for classification . . . . .
4.1.4	Fisher's linear discriminant . . . . .
4.1.5	Relation to least squares . . . . .
4.1.6	Fisher's discriminant for multiple classes . . . . .
4.1.7	The perceptron algorithm . . . . .
4.2	Probabilistic Generative Models . . . . .

# Multi-Target Prediction

# Modern Prediction Problems

## Ad-word Recommendation

The screenshot shows a Google search results page with a search bar containing 'beginner yoga classes'. The results are categorized by type: Web, Images, Maps, Videos, News, and More. The 'Web' results are highlighted with red boxes and a red arrow pointing to the first result. The results include:

- Laura Yoga Studio** (646) 702-4596  
[www.laurayoga.com](http://www.laurayoga.com)  
Great for beginners. Get the first 3 **classes** free! Call now.
- Youth Yoga Classes**  
[www.yogakids.com](http://www.yogakids.com)  
Yoga for all ages! We offer modern facilities and reasonable rates  
Yoga Kids Inc. – 610 McKenzie Boul. Denver, CO
- Yoga Accessories**  
[www.yogaaccessories.com](http://www.yogaaccessories.com)  
Experts or **beginners**, we have everything you need for **yoga**.
- Yoga Yoga Denver**  
[www.yogayogadenver.com](http://www.yogayogadenver.com)  
Yoga **classes** in denver. New to **Yoga**? Start here! Mommy & baby **yoga**!  
Map & directions to studio · Rent our Space · Energy/Exchange opportunities
- Yoga Basics: Your guide to the Practice of Yoga**  
[www.yogabasicsincredible.com](http://www.yogabasicsincredible.com)
- Hot Yoga Classes**  
[www.yogabears.com/hotyoga](http://www.yogabears.com/hotyoga)  
Dynamic, fun and cost effective!  
Special: 10 **classes** for \$100
- Yoga for beginners**  
[www.vinashiyoga.com](http://www.vinashiyoga.com)  
Burn calories and find peace.  
Small **classes**. First week free!  
(354) 555-0111 - Directions
- Lilac Yoga Studio**  
[www.lilacyogadenver.com](http://www.lilacyogadenver.com)  
Try our popular **yoga** sessions  
Limited time \$100 for 10!

At the bottom of the page, there are navigation icons for back, forward, and search.

# Modern Prediction Problems

## Ad-word Recommendation

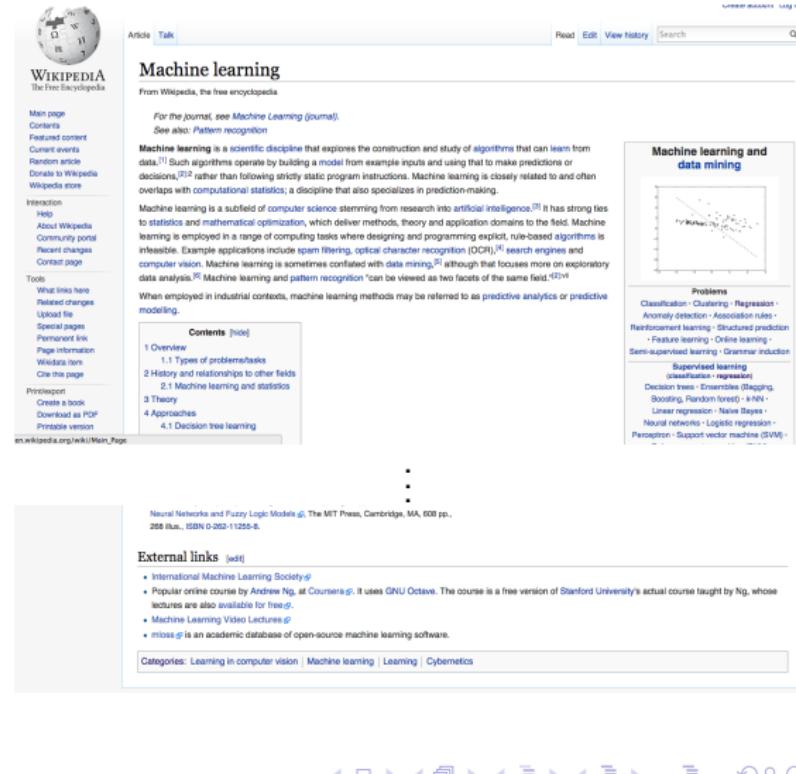
- geico auto insurance
- geico car insurance
- car insurance
- geico insurance
- need cheap auto insurance
- geico com
- car insurance coupon code



# Modern Prediction Problems

## Wikipedia Tag Recommendation

- Learning in computer vision
- Machine learning
- Learning
- Cybernetics



The screenshot shows the Wikipedia homepage with the 'Machine learning' article open. The article page includes a navigation bar with 'Article' and 'Talk' tabs, and a sidebar with 'Machine learning and data mining' categories. The main content of the article discusses machine learning as a scientific discipline and its relationship to other fields like statistics and computer science. It also mentions its applications in various domains. The sidebar lists sub-topics such as classification, clustering, regression, and various machine learning algorithms like decision trees, random forests, KNN, linear regression, naive Bayes, neural networks, logistic regression, and perceptrons. The page footer includes a 'External links' section and a 'Categories' section.

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data.<sup>[1]</sup> Such algorithms operate by building a model from example inputs and using that to make predictions or decisions,<sup>[2][3]</sup> rather than following strictly static program instructions. Machine learning is closely related to and often overlaps with computational statistics;<sup>[4]</sup> a discipline that also specializes in prediction-making.

Machine learning is a subfield of computer science stemming from research into artificial intelligence.<sup>[5]</sup> It has strong ties to statistics and mathematical optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible. Example applications include spam filtering, computer vision recognition (OCR),<sup>[6]</sup> search engines, and computer vision. Machine learning is sometimes conflated with data mining,<sup>[5]</sup> although that focuses more on exploratory data analysis.<sup>[6]</sup> Machine learning and pattern recognition<sup>[7]</sup> can be viewed as two facets of the same field.<sup>[4][5]</sup>

When employed in industrial contexts, machine learning methods may be referred to as predictive analytics or predictive modeling.

**Contents** [view]

- 1 Overview
- 2 History and relationships to other fields
- 2.1 Machine learning and statistics
- 3 Theory
- 4 Approaches
- 4.1 Decision tree learning

en.wikipedia.org/wiki/Main\_Page

Neural Networks and Fuzzy Logic Models<sup>[8]</sup>, The MIT Press, Cambridge, MA, 608 pp., 256 illus., ISBN 0-262-11255-8.

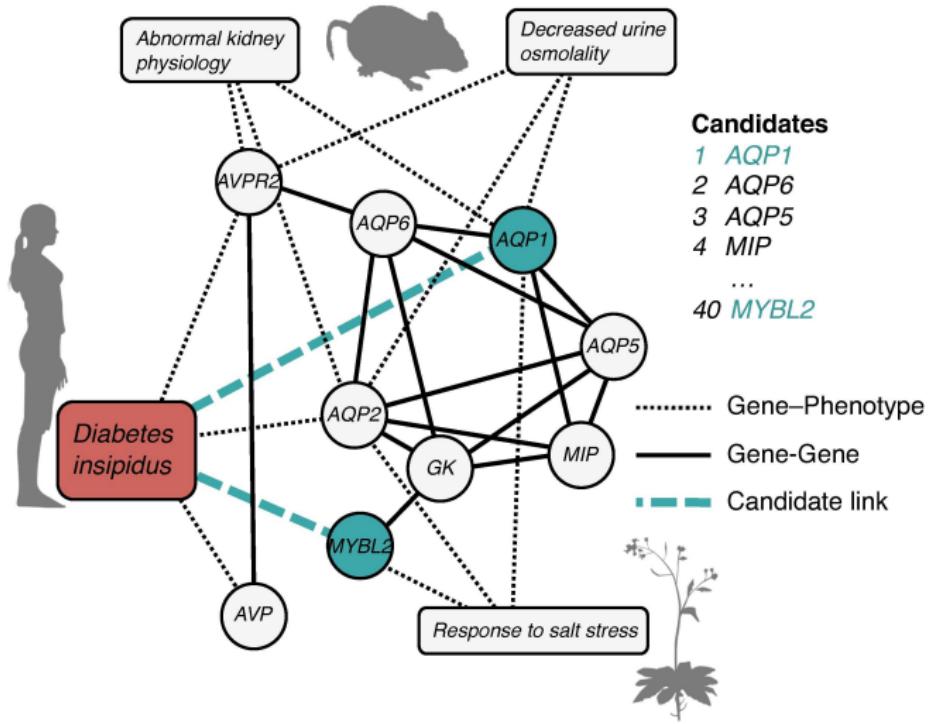
**External links** [edit]

- International Machine Learning Society<sup>[9]</sup>
- Popular online course by Andrew Ng, at Coursera<sup>[10]</sup>. It uses GNU Octave. The course is a free version of Stanford University's actual course taught by Ng, whose lectures are also available for free<sup>[11]</sup>.
- Machine Learning Video Lectures<sup>[12]</sup>
- miss<sup>[13]</sup> is an academic database of open-source machine learning software.

Categories: Learning in computer vision | Machine learning | Learning | Cybernetics

# Modern Prediction Problems

## Predicting associated disease genes



# Modern Prediction Problems

## Product Search

amazon Try Prime All ▾ yoga mat



prime student Get 50% off Prime

[See Color Options](#)

BalanceFrom GoYoga All-Purpose 1/2-Inch Extra Thick High Density Anti-Tear Exercise **Yoga Mat** with Carrying Strap

by BalanceFrom

\$17<sup>95</sup> - \$35<sup>09</sup>  prime

Some colors are Prime eligible

More Buying Choices

\$14.81 (6 used & new offers)

FREE Shipping on eligible orders

Show only BalanceFrom items

 5,714

[See Color Options](#)

AmazonBasics 1/2-Inch Extra Thick Exercise **Mat** with Carrying Strap

by AmazonBasics

\$17<sup>99</sup>  prime

Some colors are Prime eligible

FREE Shipping on eligible orders

Show only AmazonBasics items

 256

Reehut 1/2-Inch Extra Thick High Density NBR Exercise **Yoga Mat** for Pilates, Fitness & Workout w/ Carrying Strap

by Reehut

\$20<sup>99</sup> \$39.99  prime

Some colors are Prime eligible

FREE Shipping on eligible orders and **6 more promotions** 

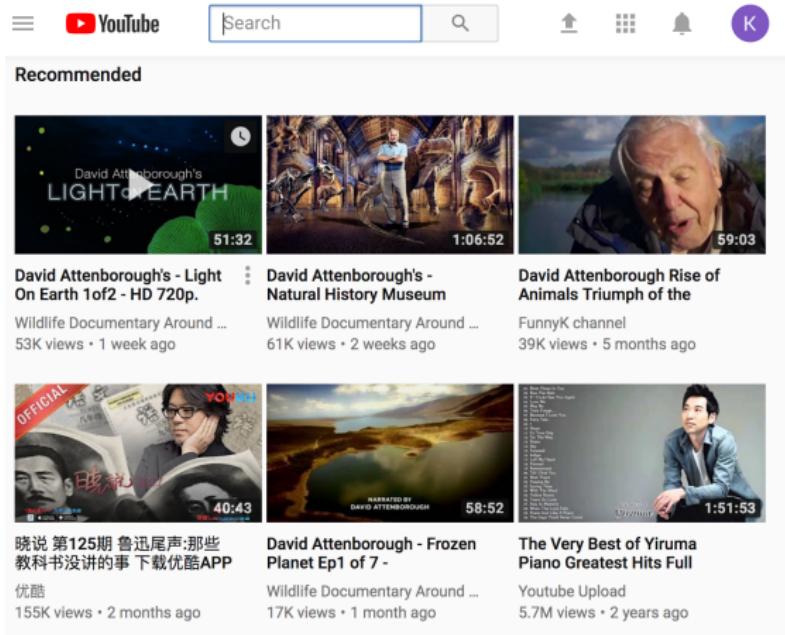
Show only Reehut items

 1,756



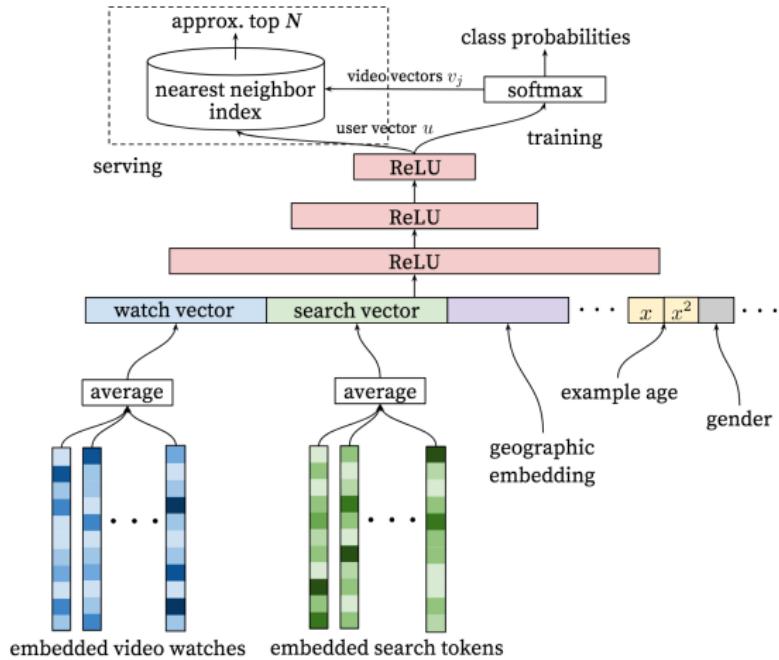
# Modern Prediction Problems

## Youtube recommendation



# Modern Prediction Problems

## Model architecture for youtube recommendation



P. Covington, J. Adams, and E. Sargin. *Deep neural networks for youtube recommendations*. Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.

# Modern Challenges in Multi-Target Prediction

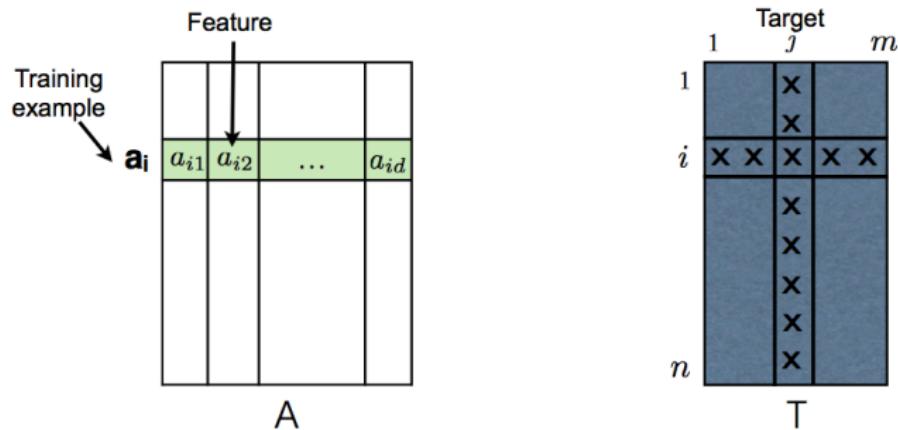
- Millions of correlated targets, and missing target values
- Targets have features
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

# Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

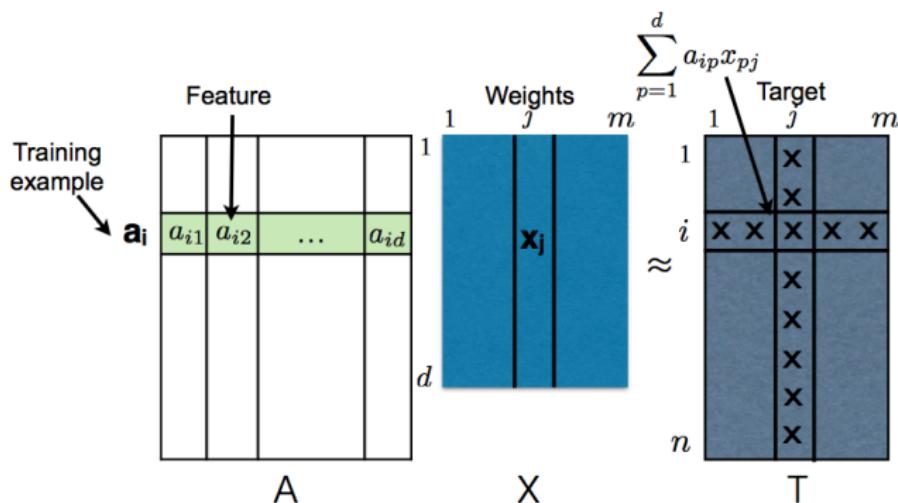
# Prediction with Multiple Targets

- Input data  $\mathbf{a}_i$  is associated with  $m$  targets,  $\mathbf{t}_i = (t_i^{(1)}, t_i^{(2)}, \dots, t_i^{(m)})$



# Multi-Target Linear Prediction

- Basic model: Treat targets independently
- Estimate regression coefficients  $x_j$  for each target  $j$



# Multi-Target Linear Prediction

- Assume targets  $\mathbf{t}^{(j)}$  are independent
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top \mathbf{X}$
- Objective for multi-target regression:

$$\min_{\mathbf{X}} \|\mathbf{T} - \mathbf{A}\mathbf{X}\|_F^2$$

# Multi-Target Linear Prediction

- Assume targets  $\mathbf{t}^{(j)}$  are independent
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top \mathbf{X}$
- Objective for multi-target regression:

$$\min_{\mathbf{X}} \|\mathbf{T} - \mathbf{A}\mathbf{X}\|_F^2$$

- Closed-form solution:

$$\mathbf{V}_A \Sigma_A^{-1} \mathbf{U}_A^\top \mathbf{T} = \arg \min_{\mathbf{X}} \|\mathbf{T} - \mathbf{A}\mathbf{X}\|_F^2$$

where  $\mathbf{A} = \mathbf{U}_A \Sigma_A \mathbf{V}_A^\top$  is the thin SVD of  $\mathbf{A}$

# Multi-Target Linear Prediction

- Assume targets  $\mathbf{t}^{(j)}$  are independent
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top \mathbf{X}$
- Objective for multi-target regression:

$$\min_{\mathbf{X}} \|\mathbf{T} - \mathbf{A}\mathbf{X}\|_F^2$$

- Closed-form solution:

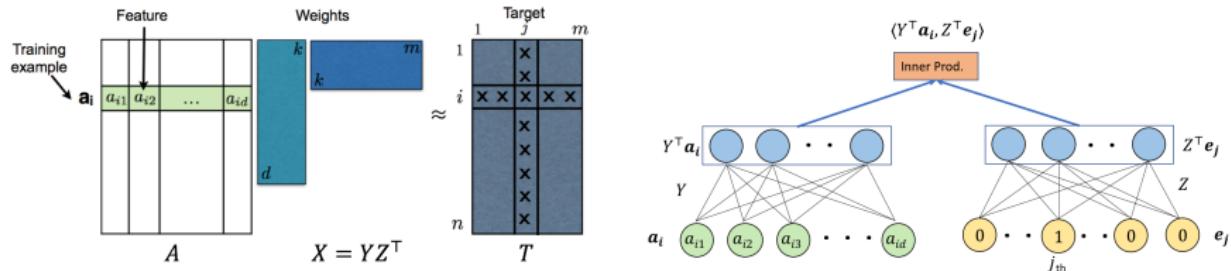
$$\mathbf{V}_A \Sigma_A^{-1} \mathbf{U}_A^\top \mathbf{T} = \arg \min_{\mathbf{X}} \|\mathbf{T} - \mathbf{A}\mathbf{X}\|_F^2$$

where  $\mathbf{A} = \mathbf{U}_A \Sigma_A \mathbf{V}_A^\top$  is the thin SVD of  $\mathbf{A}$

In multi-label classification: **Binary Relevance** (independent binary classifier for each label)

# Multi-Target Linear Prediction: Low-rank Model

- Exploit correlations between targets  $T$ , where  $T \approx AX$
- **Reduced-Rank Regression** [A.J. Izenman, 1974] — model the coefficient matrix  $X$  as *low-rank*



A. J. Izenman. *Reduced-rank regression for the multivariate linear model*. Journal of Multivariate Analysis 5.2 (1975): 248-264.

# Multi-Target Linear Prediction: Low-rank Model

- $X$  is rank- $k$
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression:

$$\min_{X: \text{rank}(X) \leq k} \|T - AX\|_F^2$$

# Multi-Target Linear Prediction: Low-rank Model

- $X$  is rank- $k$
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression:

$$\min_{X: \text{rank}(X) \leq k} \|T - AX\|_F^2$$

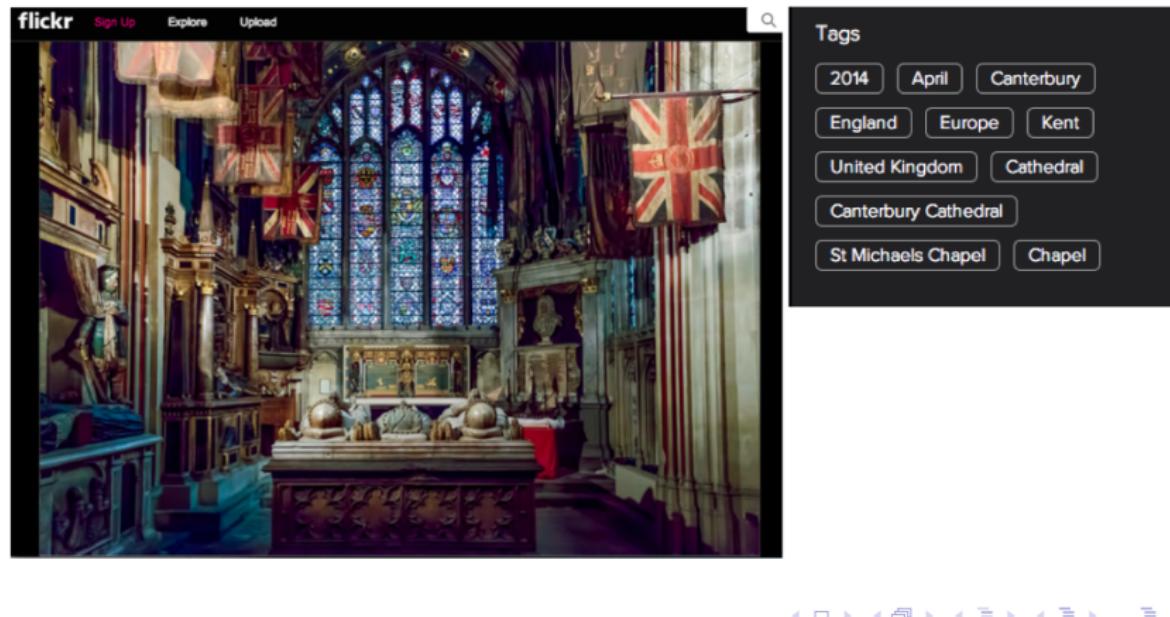
- Closed-form solution:

$$\begin{aligned} X^* &= \arg \min_{X: \text{rank}(X) \leq k} \|T - AX\|_F^2 \\ &= \begin{cases} V_A \Sigma_A^{-1} U_A^\top T_k & \text{if } A \text{ is full row rank,} \\ V_A \Sigma_A^{-1} M_k & \text{otherwise,} \end{cases} \end{aligned}$$

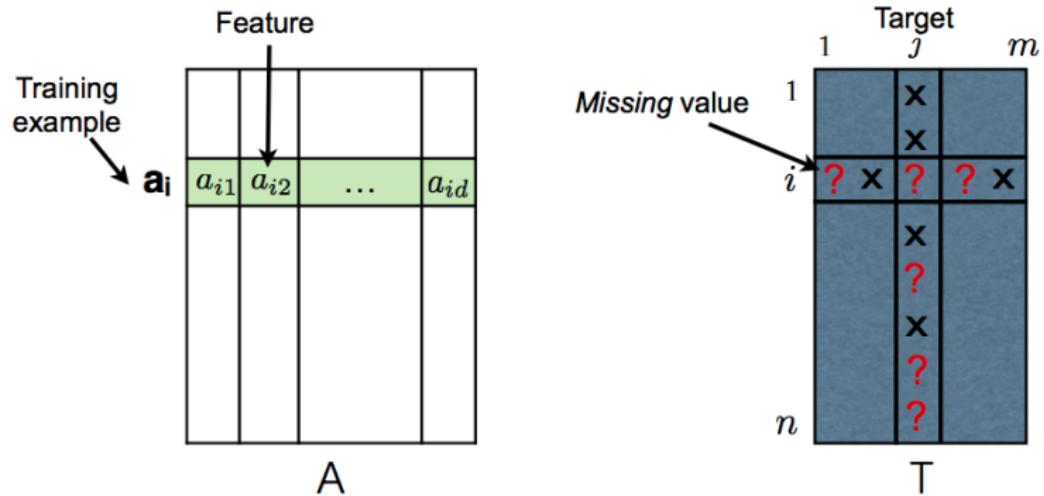
where  $A = U_A \Sigma_A V_A^\top$  is the thin SVD of  $A$ ,  $M = U_A^\top T$ , and  $T_k$ ,  $M_k$  are the best rank- $k$  approximations of  $T$  and  $M$  respectively.

# Multi-Target Prediction with Missing Target Values

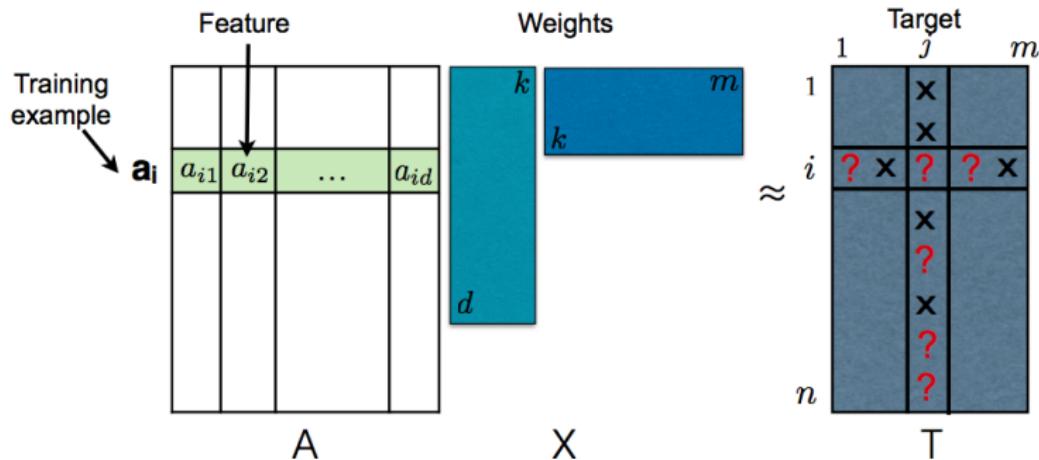
- In many applications, several observations (targets) may be *missing*
- E.g. Recommending tags for images and wikipedia articles



# Multi-Target Prediction with Missing Target Values



# Multi-Target Prediction with Missing Target Values



- Low-rank model:  $\mathbf{t}_i = \mathbf{a}_i^\top \mathbf{X}$  where  $\mathbf{X}$  is low-rank

# Multi-Target Prediction with Missing Target Values

- $X$  is rank- $k$
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression with missing target values:

$$\min_{X: \text{rank}(X) \leq k} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{e}_j - T_{ij})^2,$$

where  $\Omega$  is the set of observed targets.

# Multi-Target Prediction with Missing Target Values

- $X$  is rank- $k$
- Linear predictive model:  $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression with missing target values:

$$\min_{X: \text{rank}(X) \leq k} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{e}_j - T_{ij})^2,$$

where  $\Omega$  is the set of observed targets.

- No closed-form solution

# Multi-Target Prediction with Missing Values: Algorithms

- Algorithm 1 (LEML(Nuclear)): Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{e}_j - T_{ij})^2$$

- Convex Relaxation
- Algorithm 2 (LEML(ALS)): Alternating Least Squares

$$\min_{Y \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{m \times k}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top Y Z^\top \mathbf{e}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

- Alternately minimize w.r.t.  $Y$  and  $Z$
- Non-convex optimization
- Computationally cheaper than nuclear-norm method

# Application: Image Tag Recommendation

NUS-Wide Image Dataset



- 161,780 training images
- 107,879 test images
- 1,134 features
- 1,000 tags

# Application: Image Tag Recommendation

- Low-rank Model with  $k = 50$ :

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	<b>574</b>	<b>20.71</b>	<b>15.96</b>	<b>0.7741</b>
WSABIE	4,705	14.58	11.37	0.7658

- Low-rank Model with  $k = 100$ :

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	<b>1,097</b>	<b>20.76</b>	<b>16.00</b>	<b>0.7718</b>
WSABIE	6,880	12.46	10.21	0.7597

LEML: H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).  
WSABIE: J. Weston, S. Bengio, and N. Usunier. *Wsabie: Scaling up to large vocabulary image annotation*. in IJCAI (2011).

# Application: Wikipedia Tag Recommendation

## Wikipedia Dataset

The screenshot shows a Wikipedia article page for 'Machine learning'. The page title is 'Machine learning' with a sub-section 'Machine learning and data mining'. The main content discusses the field of machine learning as a scientific discipline that explores the construction and study of algorithms that can learn from data. It highlights that machine learning is a subfield of computer science stemming from research into artificial intelligence. The page also mentions its applications in various computing tasks like spam filtering, optical character recognition, search engines, and computer vision. A sidebar on the right provides a list of machine learning problems and sub-fields. The page footer includes categories like 'Learning in computer vision', 'Machine learning', 'Learning', and 'Cybernetics'.

- 881,805 training wiki pages
- 10,000 test wiki pages
- 366,932 features
- 213,707 tags



# Application: Wikipedia Tag Recommendation

- Low-rank Model with  $k = 250$ :

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	<b>9,932</b>	<b>19.56</b>	14.43	<b>0.9086</b>
WSABIE	79,086	18.91	<b>14.65</b>	0.9020

- Low-rank Model with  $k = 500$ :

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	<b>18,072</b>	<b>22.83</b>	<b>17.30</b>	<b>0.9374</b>
WSABIE	139,290	19.20	15.66	0.9058

LEML: H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).  
WSABIE: J. Weston, S. Bengio, and N. Usunier. *Wsabie: Scaling up to large vocabulary image annotation*. in IJCAI (2011).

# Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

# Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

# Bilinear Prediction: Inductive Matrix Completion (IMC)

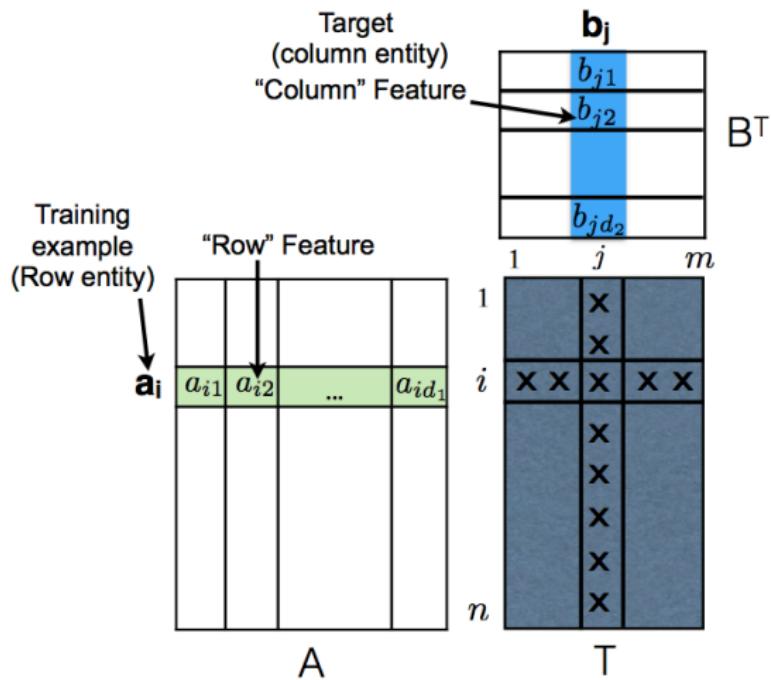
## Modern Prediction Problems

## Product Search

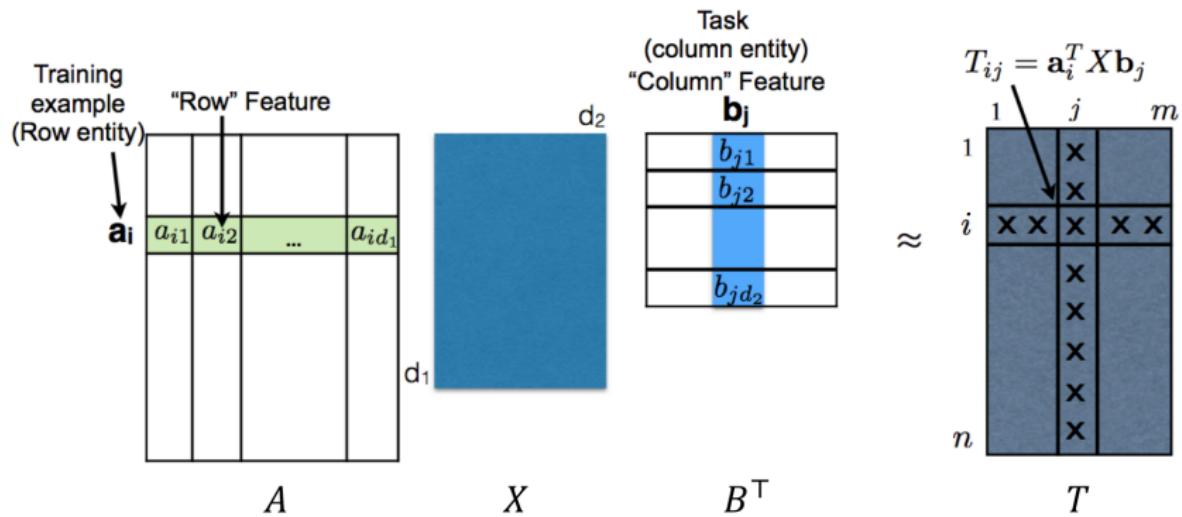
# Bilinear Prediction

- Augment multi-target prediction with *features* on targets as well
- Motivated by modern applications of machine learning — bioinformatics, recommendation system with item features
- Need to model *dyadic* or *pairwise* interactions
- Move from linear models to *bilinear* models — linear in input features *as well as* target features

# Bilinear Prediction



# Bilinear Prediction



# Bilinear Prediction

- Bilinear predictive model:  $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for bilinear predictive model

$$\min_X \|T - AXB^\top\|_F^2$$

# Bilinear Prediction

- Bilinear predictive model:  $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for bilinear predictive model

$$\min_X \|T - AXB^\top\|_F^2$$

- Closed-form solution:

$$V_A \Sigma_A^{-1} U_A^\top T U_B \Sigma_B^{-1} V_B^\top = \arg \min_X \|T - AXB^\top\|_F^2$$

where  $A = U_A \Sigma_A V_A^\top$ ,  $B = U_B \Sigma_B V_B^\top$  are the thin SVDs of  $A$  and  $B$

# Bilinear Prediction: Low-rank Model

- $X$  is rank- $k$
- Bilinear predictive model:  $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for low-rank bilinear model:

$$\min_{X: \text{rank}(X) \leq k} \|T - AXB^\top\|_F^2$$

# Bilinear Prediction: Low-rank Model

- $X$  is rank- $k$
- Bilinear predictive model:  $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for low-rank bilinear model:

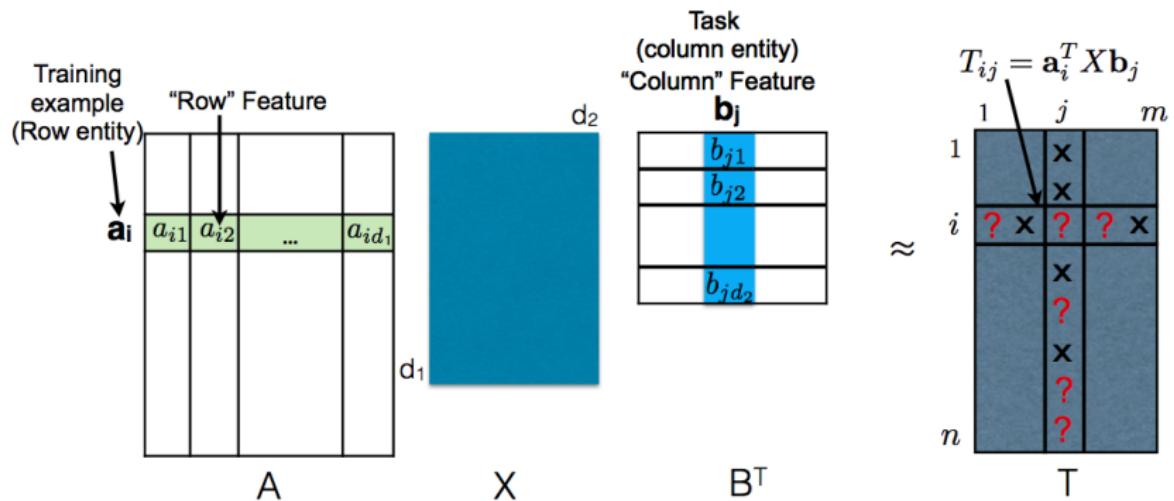
$$\min_{X: \text{rank}(X) \leq k} \|T - AXB^\top\|_F^2$$

- Closed-form solution:

$$\begin{aligned} X^* &= \min_{X: \text{rank}(X) \leq k} \|T - AXB^\top\|_F^2 \\ &= \begin{cases} V_A \Sigma_A^{-1} U_A^\top \mathbf{T}_k U_B \Sigma_B^{-1} V_B^\top & \text{if } A, B \text{ are full row rank,} \\ V_A \Sigma_A^{-1} \mathbf{M}_k \Sigma_B^{-1} V_B^\top & \text{otherwise,} \end{cases} \end{aligned}$$

where  $A = U_A \Sigma_A V_A^\top$ ,  $B = U_B \Sigma_B V_B^\top$  are the thin SVDs of  $A$  and  $B$ ,  $M = U_A^\top T U_B$ , and  $T_k$ ,  $M_k$  are the best rank- $k$  approximations of  $T$  and  $M$

# Bilinear Prediction with Missing Target Values



# Bilinear Prediction with Missing Target Values: Algorithms

- Algorithm 1: Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

- Convex Relaxation
- Algorithm 2: Alternating Least Squares (ALS)

$$\min_{Y \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{d \times k}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top Y Z^\top \mathbf{b}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

- Non-convex optimization

# Bilinear Prediction with Missing Target Values: Algorithms

- Algorithm 1: Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

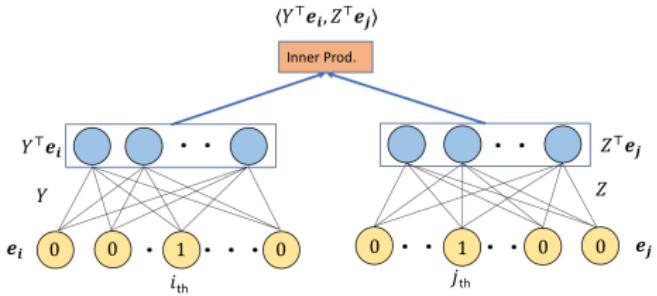
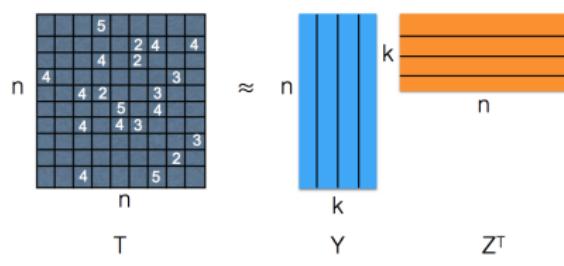
- Convex Relaxation
- Algorithm 2: Alternating Least Squares (ALS)

$$\min_{Y \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{d \times k}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top Y Z^\top \mathbf{b}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

- Non-convex optimization
- Can we recover the model? How many observations are required?

# Recovery Guarantees: Matrix Completion

- Matrix Completion:
  - Recover a low-rank matrix from partially observed entries
- Exact recovery requires  $\tilde{O}(kn)$  observed entries

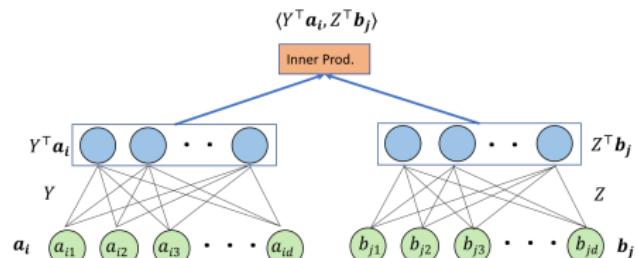
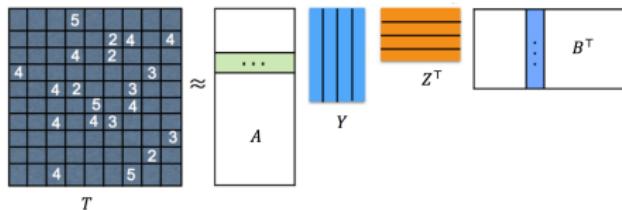


$\tilde{O}(n)$  hides  $\text{polylog}(n)$

E. J. Candes and B. Recht. *Exact matrix completion via convex optimization*. Foundations of Computational mathematics (2009).

# Inductive Matrix Completion

- Inductive Matrix Completion:
  - Recover a low-rank bilinear model from partially obtained targets
- Degrees of freedom in  $X$  are  $O(kd)$
- Can we get better sample complexity (than  $\tilde{O}(kn)$ )?



# Recovery Guarantees

## Theorem (Recovery Guarantees for Nuclear-norm Minimization)

Let  $X_* = U_* \Sigma_* V_*^\top \in \mathbb{R}^{d \times d}$  be the SVD of  $X_*$  with rank  $k$ , and  $T = AX_*B^\top$ . Let  $\mathcal{X} = \|X_*\|_*$ . Assume  $A, B$  are orthonormal matrices w.l.o.g., satisfying the incoherence conditions. Then if  $\Omega$  is uniformly observed with

$$|\Omega| \geq O(kd \log d \log n),$$

the solution of nuclear-norm minimization problem is unique and equal to  $X_*$  with high probability.

The incoherence conditions are

$$\mathbf{C1.} \max_{i \in [n]} \|\mathbf{a}_i\|_2^2 \leq \frac{\mu d}{n}, \max_{j \in [n]} \|\mathbf{b}_j\|_2^2 \leq \frac{\mu d}{n}$$

$$\mathbf{C2.} \max_{i \in [n]} \|U_*^\top \mathbf{a}_i\|_2^2 \leq \frac{\mu_0 k}{n}, \max_{j \in [n]} \|V_*^\top \mathbf{b}_j\|_2^2 \leq \frac{\mu_0 k}{n}$$

K. Zhong, P. Jain, I. S. Dhillon. [Efficient Matrix Sensing Using Rank-1 Gaussian Measurements](#). In ALT (2015).

# Recovery Guarantees

## Theorem (Convergence Guarantees for ALS )

Let  $X_*$  be a rank- $k$  matrix with condition number  $\beta$  and  $T = AX_*B^\top$ . Assume  $A, B$  are orthogonal w.l.o.g. and satisfy the incoherence conditions. Then if  $\Omega$  is uniformly sampled with

$$|\Omega| \geq O(k^4\beta^2 d \log d),$$

then after  $H$  iterations of ALS,  $\|Y_H Z_{H+1}^\top - X_*\|_2 \leq \epsilon$ , where  $H = O(\log(\|X_*\|_F/\epsilon))$ .

The incoherence conditions are:

$$\mathbf{C1}. \max_{i \in [n]} \|\mathbf{a}_i\|_2^2 \leq \frac{\mu d}{n}, \max_{j \in [n]} \|\mathbf{b}_j\|_2^2 \leq \frac{\mu d}{n}$$

$$\mathbf{C2'}. \max_{i \in [n]} \|Y_h^\top \mathbf{a}_i\|_2^2 \leq \frac{\mu_0 k}{n}, \max_{j \in [n]} \|Z_h^\top \mathbf{b}_j\|_2^2 \leq \frac{\mu_0 k}{n}, \forall h = 1, 2, \dots, H$$

# Sample Complexity for Recovery Guarantees

- Sample complexity of Inductive Matrix Completion (IMC) and Matrix Completion (MC).

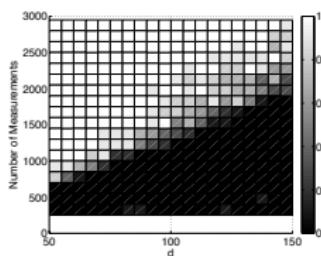
methods	IMC	MC
Nuclear-norm	$\tilde{O}(kd)$	$\tilde{O}(kn)$ (Recht & Cands, 2009)
ALS	$\tilde{O}(k^4\beta^2d)$	$\tilde{O}(k^3\beta^2n)$ (Hardt, 2014)

where  $\beta$  is the condition number of  $X$

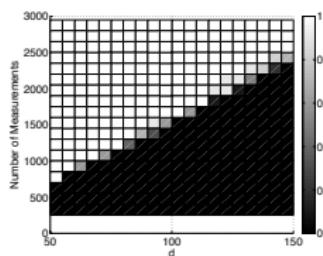
- In most cases,  $n \gg d$
- Incoherence conditions on  $A, B$  are required
  - Satisfied e.g. when  $A, B$  are Gaussian (no assumption on  $X$  needed)

# Sample Complexity Results

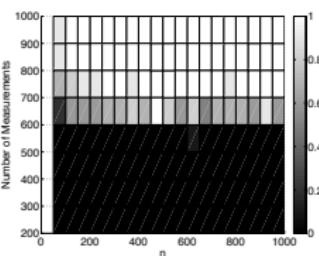
- All matrices are sampled from Gaussian random distribution.
- Left two figures: fix  $k = 5$ ,  $n = 1000$  and change  $d$ .
- Right two figures: fix  $k = 5$ ,  $d = 50$  and change  $n$ .
- Darkness of the shading is proportional to the number of failures (repeated 10 times).



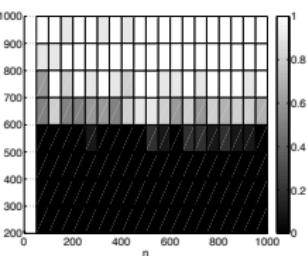
$|\Omega|$  vs.  $d$  (ALS)



$|\Omega|$  vs.  $d$  (Nuclear)



$|\Omega|$  vs.  $n$  (ALS)



$|\Omega|$  vs.  $n$  (Nuclear)

- Sample complexity is proportional to  $d$  while almost independent of  $n$  for both Nuclear-norm and ALS methods.

# Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

# Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

# Bilinear Prediction with Noisy Features

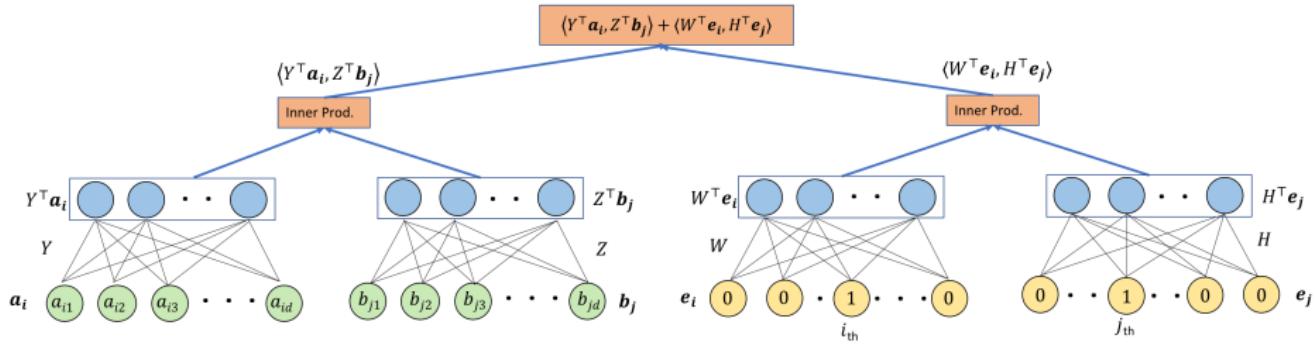
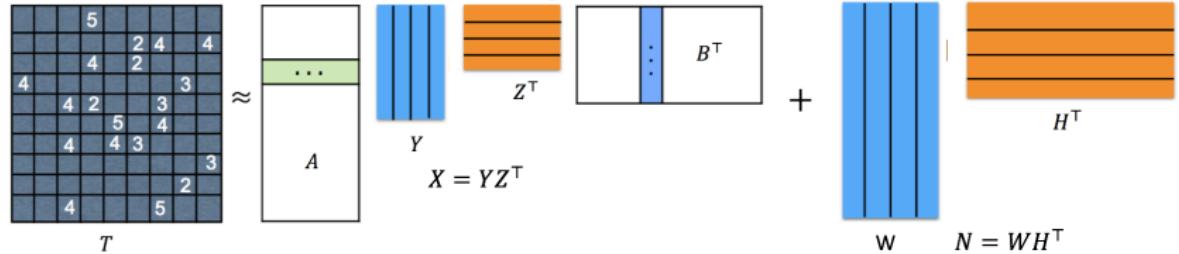
# Noisy Features

- IMC implicitly assumes that features are good, i.e.,

$$\text{col}(T) \subseteq \text{col}(A) \text{ and } \text{row}(T) \subseteq \text{col}(B)$$

- When features are not good, learn underlying matrix  $T$  jointly from two parts:
  1. Feature-covered part:  $AXB^\top$ .
  2. Residual part:  $N$ .  
⇒ Estimate  $T$  as  $AXB^\top + N$ . ([DirtyIMC](#))
- Both  $X$  and  $N$  are preferred to be low-rank.
  - 1  $X = YZ^\top$ , where  $Y \in \mathbb{R}^{d \times k}$ ,  $Z \in \mathbb{R}^{d \times k}$
  - 2  $N = WH^\top$ , where  $W \in \mathbb{R}^{n \times k}$ ,  $H \in \mathbb{R}^{n \times k}$

# Dirty Inductive Matrix Completion



# Dirty Inductive Matrix Completion

- Algorithm 1: Nuclear-norm constraint objective

$$\begin{aligned} \min_{X, N} \quad & \sum_{(i,j) \in \Omega} ((\mathbf{a}_i^\top X \mathbf{b}_j + N_{ij}) - T_{ij})^2 \\ \text{s.t.} \quad & \|X\|_* \leq \mathcal{X}, \|N\|_* \leq \mathcal{N} \end{aligned}$$

- Algorithm 2: Alternating Least Squares (ALS)

$$\begin{aligned} \min_{Y, Z, W, H} \quad & \sum_{(i,j) \in \Omega} ((\mathbf{a}_i^\top Y Z^\top \mathbf{b}_j + \mathbf{e}_i^\top W H^\top \mathbf{e}_j) - T_{ij})^2 \\ \text{s.t.} \quad & Y \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{d \times k}, W \in \mathbb{R}^{n \times k}, H \in \mathbb{R}^{n \times k} \end{aligned}$$

# Measuring Quality of Features

- Intuition: what is the meaning of good features?
  - $T$  lies mostly in the space spanned by features.
- A formal measurement:
  - Define linear projection onto  $\text{col}(A)$  and  $\text{col}(B)$ :

$$\bar{X} = \arg \min_{X} \|AXB^\top - T\|_F^2,$$

- The trace norm of residual is used for measuring quality of features.

$$\mathcal{N} = \|T - A\bar{X}B^\top\|_*$$

- Smaller  $\mathcal{N}$  implies a better (linear) feature set.

# Error Bound for Dirty IMC

## Theorem

Consider nuclear-norm objective with  $\mathcal{X} = \|\bar{X}\|_*$  and  $\mathcal{N} = \|T - A\bar{X}B^\top\|_*$ . Then with probability at least  $1 - \delta$ , the optimal solution  $(\hat{N}, \hat{X})$  satisfies:

$$\begin{aligned} \frac{1}{n^2} \|\hat{N} + A\hat{X}B^\top - T\|_F^2 \leq \min \left\{ L_\ell \mathcal{N} \sqrt{\frac{\log n}{|\Omega|}}, \sqrt{C L_\ell \mathcal{B} \frac{\mathcal{N} \sqrt{n}}{|\Omega|}} \right\} \\ + \frac{L_\ell d^2}{\gamma^2} \sqrt{\frac{\log n}{|\Omega|}} + \mathcal{B} \sqrt{\frac{\log \frac{1}{\delta}}{|\Omega|}}, \end{aligned}$$

where  $\mathcal{B}, C, \gamma, L_\ell$  are all numerical constants.

# Error Bound for Dirty IMC

What does this error bound mean?

- The feature quality measure,  $\mathcal{N} := \|T - A\bar{X}B^\top\|_*$ 
  - Case 1: when features are perfect,  $\mathcal{N} = 0$
  - Case 2: when features contain no information,  $\mathcal{N} = O(n)$
  - Case 3: when features are noisy but still informative,  $\mathcal{N} = o(n)$
- Consider Case 3, i.e.,  $\mathcal{N} = o(n)$ ,
  - ① if use IMC,

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \|A\hat{X}B^\top - T\|_F^2 \not\rightarrow 0$$

- ② if use DirtyIMC,

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \|\hat{N} + A\hat{X}B^\top - T\|_F^2 \rightarrow 0,$$

as long as  $|\Omega| = o(n^{1.5})$

## Applications: Sign Prediction

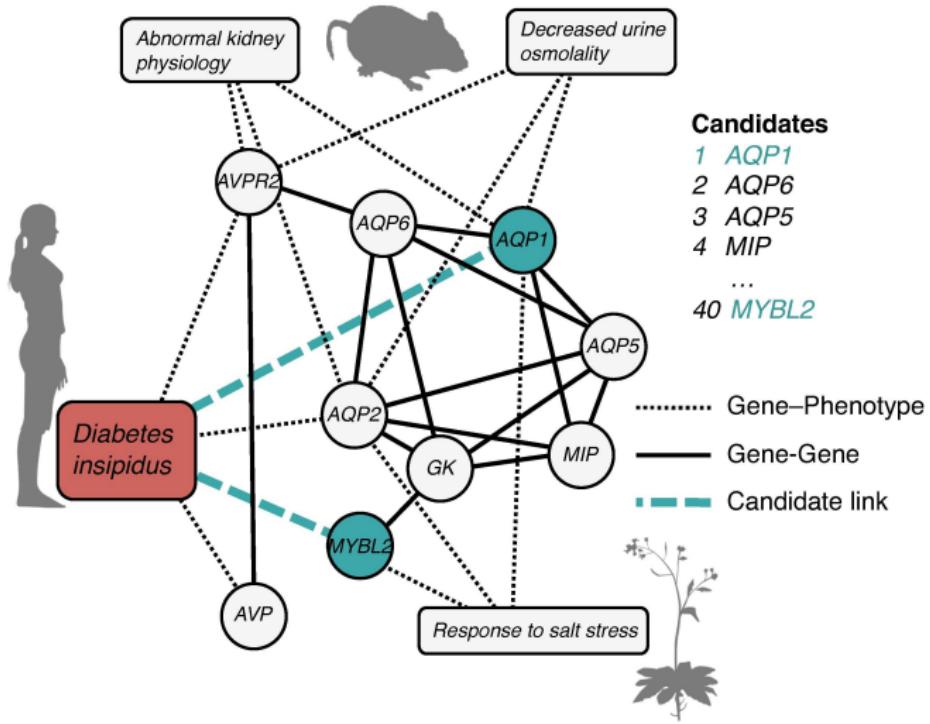
- Sign prediction task on a review sharing site Epinions ( $n \approx 105K$ ,  $|\Omega| \approx 807K$ ), where people can trust or distrust others.
- Low-rank MC and IMC yield state-of-the-art results on these problems.
- We also collect features  $\mathbf{a}_i \in \mathbb{R}^{41}$  for each person  $i$  from review history.
- Thus, we can replace MC with DirtyIMC by encoding side information.
- DirtyIMC performs the best in terms of both accuracy and AUC.

Method	Accuracy	AUC
DirtyIMC	<b>0.9474</b> $\pm 0.0009$	<b>0.9506</b>
MC	0.9412 $\pm 0.0011$	0.9020
IMC	0.9139 $\pm 0.0016$	0.9109

# Positive-Unlabeled Learning

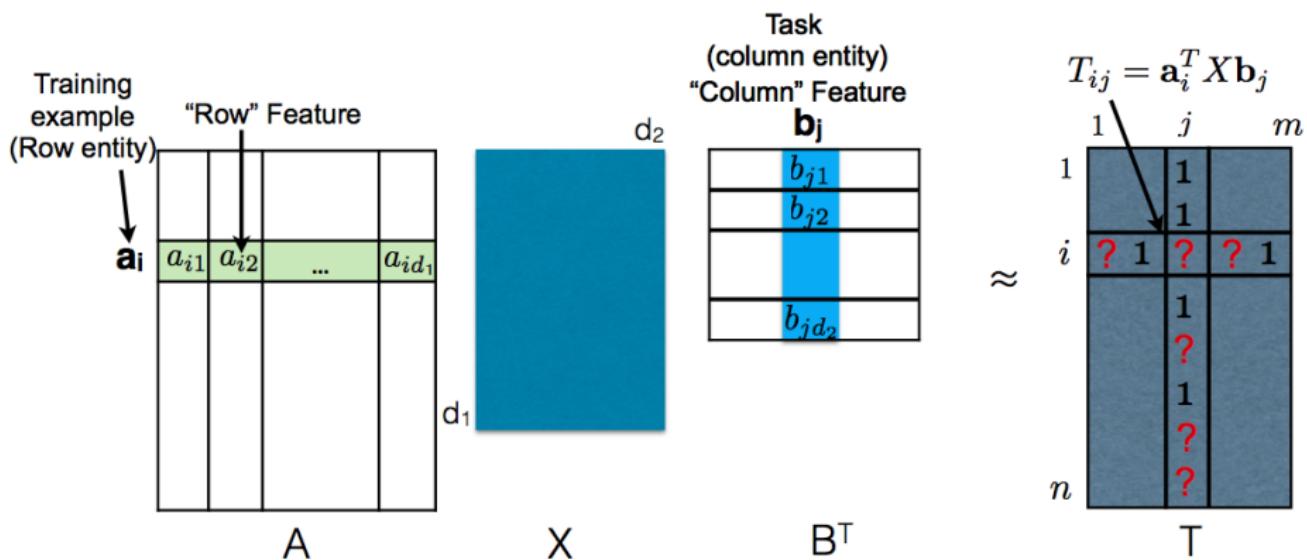
# Modern Prediction Problems

## Predicting related disease genes



# Bilinear Prediction: PU Learning

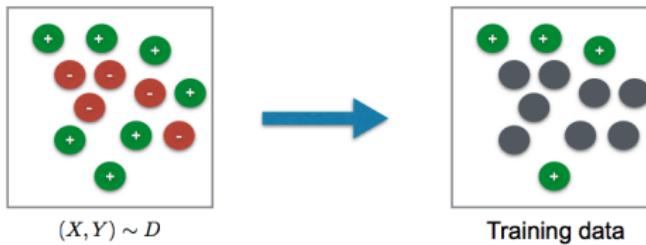
In many applications, only “positive” labels are observed



# PU Learning

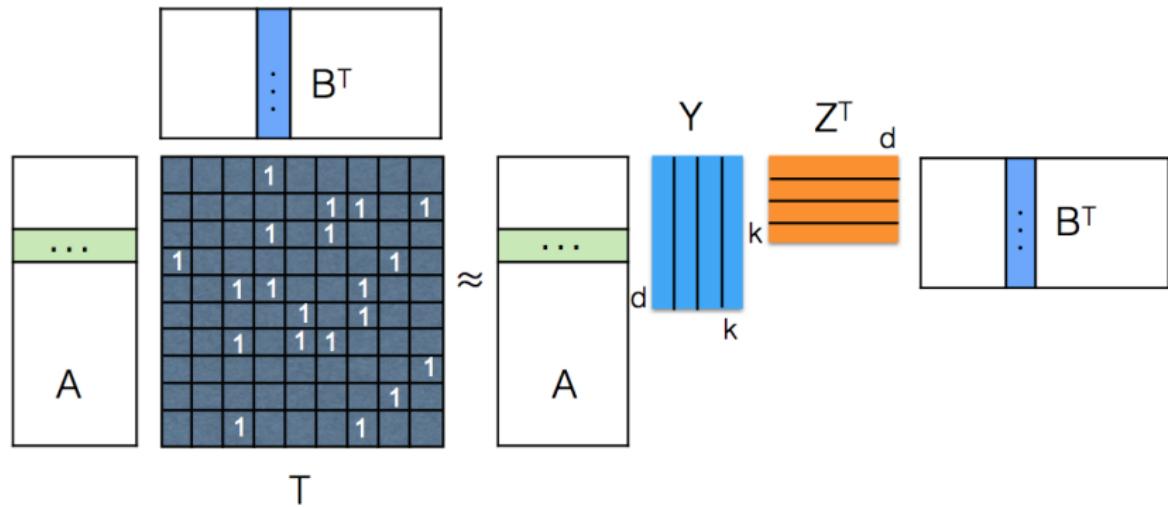
Learning Task	“Positives”	“Negatives”	“Unlabeled”
Supervised	✓	✓	
Semi-supervised	✓	✓	✓
Positive-Unlabeled (PU)	✓		✓
Unsupervised			✓

- No observations of the “negative” class available



# PU Inductive Matrix Completion

- Guarantees so far assume observations are sampled uniformly
- What can we say about the case when observations are all 1's (“positives”)?
- Typically, 99% entries are missing (“unlabeled”)



# PU Inductive Matrix Completion

- Inductive Matrix Completion:

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

- Commonly used PU strategy: Biased Matrix Completion

$$\min_{\|X\|_* \leq \mathcal{X}} \alpha \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2 + (1 - \alpha) \sum_{(i,j) \notin \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - 0)^2$$

Typically,  $\alpha > 1 - \alpha$  ( $\alpha \approx 0.9$ ).

# PU Inductive Matrix Completion

- Inductive Matrix Completion:

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

- Commonly used PU strategy: Biased Matrix Completion

$$\min_{\|X\|_* \leq \mathcal{X}} \alpha \sum_{(i,j) \in \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2 + (1 - \alpha) \sum_{(i,j) \notin \Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - 0)^2$$

Typically,  $\alpha > 1 - \alpha$  ( $\alpha \approx 0.9$ ).

- We can show theoretical guarantees for the biased formulation

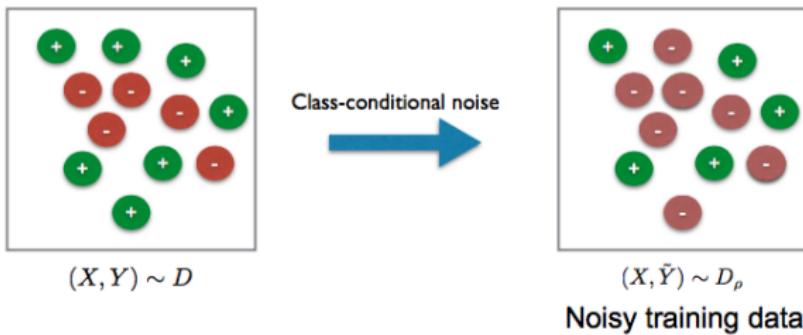
V. Sindhwani, S. S. Bucak, J. Hu, A. Mojsilovic. *One-class matrix completion with low-density factorizations*. ICDM, pp. 1055-1060. 2010.

# PU Learning: Random Noise Model

- Can be formulated as learning with “class-conditional” noise

$$P(\tilde{Y} = -1|Y = +1) = \rho_{+1}$$
$$P(\tilde{Y} = +1|Y = -1) = \rho_{-1}$$

Becomes PU learning  
when  $\rho_{-1} = 0$



N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. *Learning with Noisy Labels*. In Advances in Neural Information Processing Systems, pp. 1196-1204. 2013.

## PU Inductive Matrix Completion

## A deterministic PU learning model

M

0.2	0.1	0	0.8
0	0.6	0.1	0.9
0	0	0.8	0.1
0.9	0	0.2	0.1
0	0.6	0	1



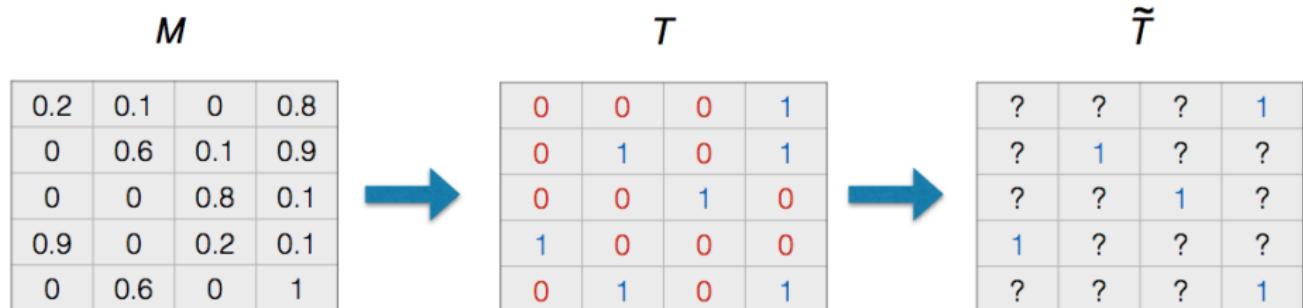
π

0	0	0	1
0	1	0	1
0	0	1	0
1	0	0	0
0	1	0	1

$$T_{ij} = \begin{cases} 1 & \text{if } M_{ij} > 0.5, \\ 0 & \text{if } M_{ij} \leq 0.5 \end{cases}$$

# PU Inductive Matrix Completion

A deterministic PU learning model



- $P(\tilde{T}_{ij} = 0 | T_{ij} = 1) = \rho$  and  $P(\tilde{T}_{ij} = 0 | T_{ij} = 0) = 1$ .
- We are given *only*  $\tilde{T}$  but *not*  $T$  or  $M$
- Goal: Recover  $T$  given  $\tilde{T}$  (recovering  $M$  is not possible!)

# PU Inductive Matrix Completion

## Theorem (Error Bound for PU IMC)

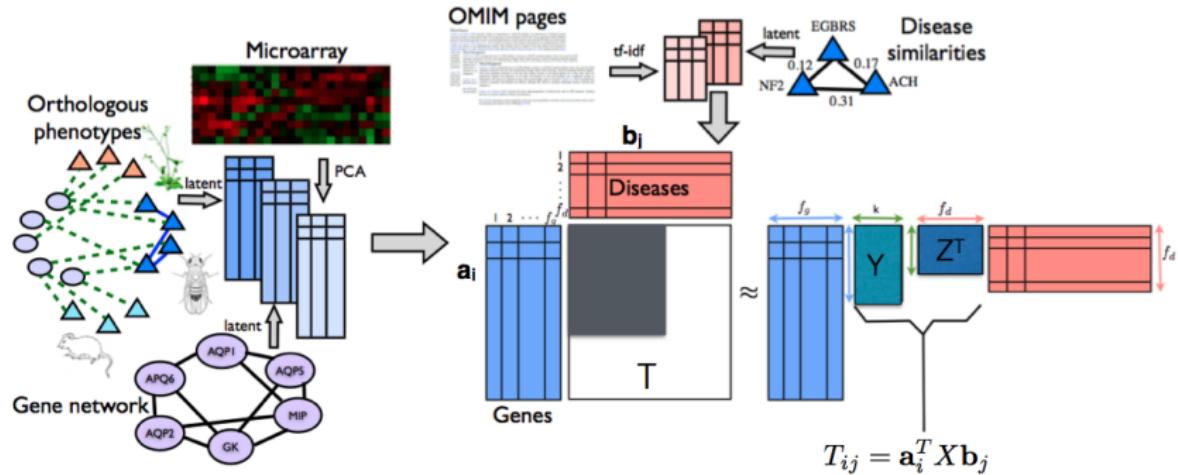
Assume ground-truth  $X$  satisfies  $\|X\|_* \leq \mathcal{X}$  (where  $M = AXB^\top$ ). Define  $\widehat{T}_{ij} = I[(A\widehat{X}B^\top)_{ij} > 0.5]$ ,  $\mathcal{A} = \max_i \|\mathbf{a}_i\|$  and  $\mathcal{B} = \max_i \|\mathbf{b}_i\|$ . If  $\alpha = \frac{1+\rho}{2}$ , then with probability at least  $1 - \delta$ ,

$$\frac{1}{n^2} \|T - \widehat{T}\|_F^2 = O\left(\frac{\sqrt{\log(2/\delta)}}{n(1-\rho)} + \frac{\mathcal{X}\mathcal{A}\mathcal{B}\sqrt{\log 2d}}{(1-\rho)n^{3/2}}\right)$$

- In other words, as long as  $1 - \rho = O\left(\frac{\log n}{n}\right)$ ,

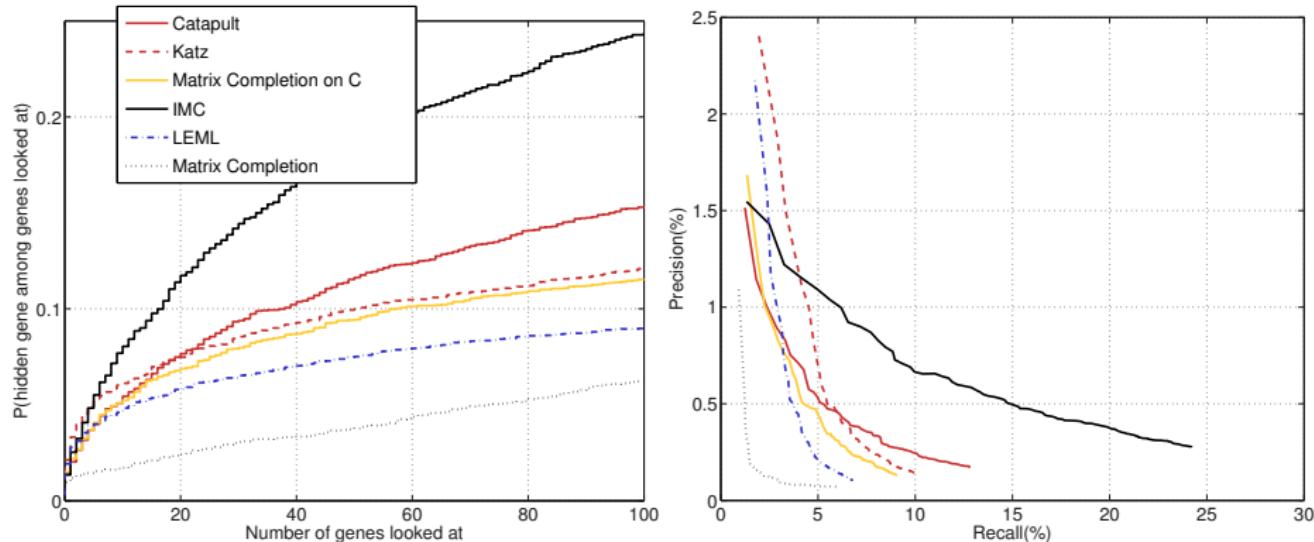
$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \|T - \widehat{T}\|_F^2 \rightarrow 0$$

# PU Inductive Matrix Completion: Gene-Disease Prediction



N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

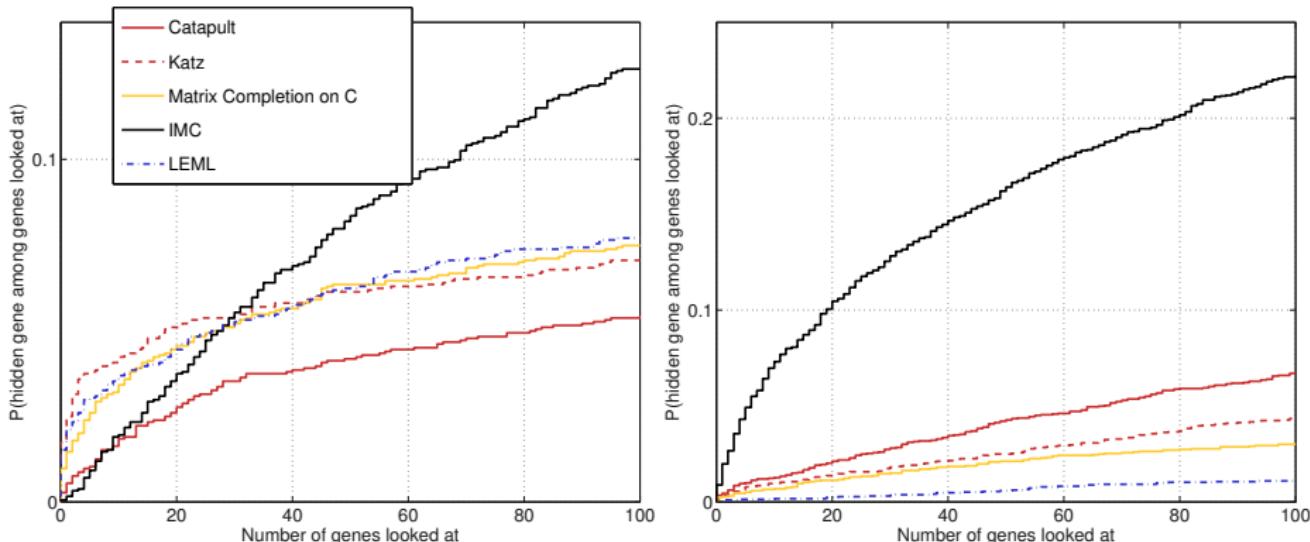
# PU Inductive Matrix Completion: Gene-Disease Prediction



Predicting gene-disease associations in the OMIM data set  
([www.omim.org](http://www.omim.org)).

N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

# PU Inductive Matrix Completion: Gene-Disease Prediction



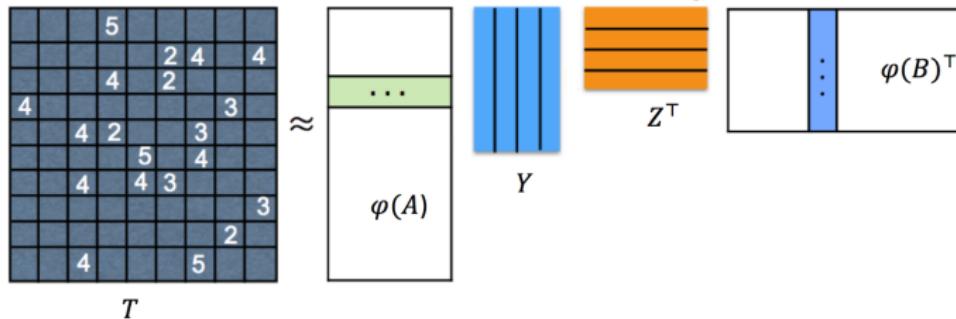
Predicting genes for diseases with *no* training associations.

N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

# Nonlinear Prediction

# Goal-Directed IMC: Two-layer Neural Network

- Key ideas for Goal-Directed IMC (GIMC):
  - Non-linear feature mapping.
    - $A \rightarrow \varphi(A)$  using non-linear mapping.
  - Learn the model and non-linear mapping simultaneously.
    - learn model and features jointly in a framework.
    - alternating minimization.

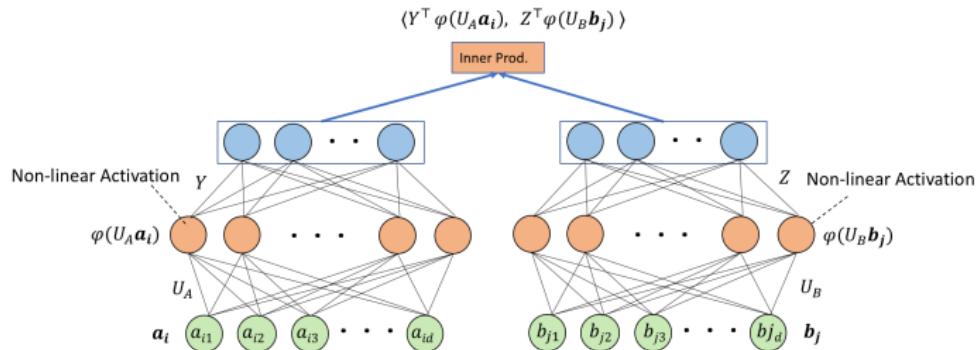


# Goal-Directed IMC: Two-layer Neural Network

- GIMC framework:

$$\min_{Y, Z, U_A, U_B} \sum_{(i, j) \in \Omega} (T_{ij} - (\varphi_{U_A}(A) Y Z^\top \varphi_{U_B}(B)^\top)_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2).$$

- $U_A$  and  $U_B$  are parameters for the non-linear feature mapping.
- $Y, Z$  are the model.



# Experimental Results: Multi-Label Learning

- Multi-label learning:
  - LEML[Yu et al. 2014]: an embedding based technique.
  - FASTXML[Prabhu and Varma, 2014]: a random forest approach.
  - SLEEC[Bhatia et al. 2015]: an ensemble of local distance preserving embeddings.
- Results (Precision @ 1, 3, 5):

	Delicious			NUS-WIDE			Delicious-large		
	P1 (%)	P3 (%)	P5 (%)	P1 (%)	P3 (%)	P5 (%)	P1 (%)	P3 (%)	P5 (%)
GIMC	<b>71.40</b>	<b>65.16</b>	<b>59.79</b>	<b>22.49</b>	<b>17.40</b>	<b>14.70</b>	46.13	40.32	38.15
SLEEC	68.38	61.50	56.35	17.67	14.20	12.07	<b>47.03</b>	<b>41.67</b>	<b>38.88</b>
FastXML	69.65	63.93	59.36	21.00	16.32	13.66	42.81	38.76	36.34
LEML	65.66	61.15	56.08	20.76	16.00	13.11	40.30	37.76	36.66

- On Delicious and NUS-WIDE, GIMC performs the best.
- On Delicious-large dataset, GIMC has similar accuracy with SLEEC, but takes much less time: 4,724 vs 25,289 seconds.

# Experimental Results: Semi-Supervised Clustering

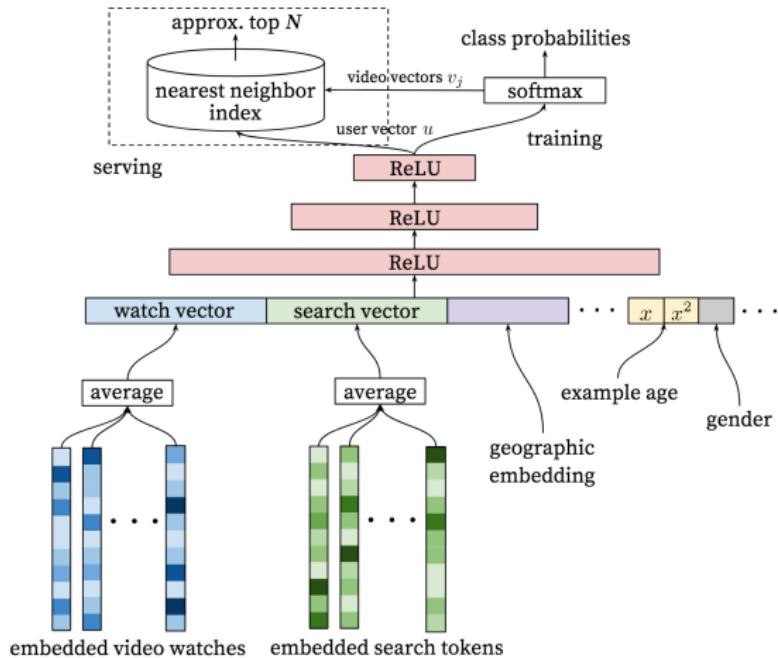
- Goal: find a clustering of  $n$  items given:
  - feature matrix  $A \in \mathbb{R}^{n \times d}$ .
  - pairwise constraints  $\{T_{ij} \mid (i, j) \in \Omega\}$  describing similar/dissimilar pairs.
- Earlier state-of-the-art: MCCC algorithm [Yi et. al, ICML 2013]:
  1. Learn a low rank similarity matrix  $S$  by conducting IMC on  $\Omega$ .
  2. Do  $k$ -means clustering on top- $k$  eigenvectors of  $S$ .
- Thus, we can replace IMC step in MCCC with GIMC.
- Results on clustering error rate:

Dataset	$n$	$d$	$k$	# constraints $ \Omega $	$n$	$k$ -means	MCCC	IMC	GIMC
Segment	2319	19	7	$n$	$n$	0.1433	0.0891	<b>0.0683</b>	0.0724
					$5n$	0.1347	0.0800	0.0580	<b>0.0570</b>
					$10n$	0.1363	0.0809	0.0650	<b>0.0446</b>
					$15n$	0.1362	0.0872	0.0678	<b>0.0402</b>
					$20n$	0.1330	0.0837	0.0564	<b>0.0380</b>
Covtype-sub	1711	54	7	$n$	$n$	0.2523	0.2498	<b>0.1840</b>	0.1898
					$5n$	0.2112	0.1772	0.1930	<b>0.1592</b>
					$10n$	0.2068	0.1708	0.1722	<b>0.1388</b>
					$15n$	0.2203	0.1677	0.1687	<b>0.1262</b>
					$20n$	0.2124	0.1607	0.1561	<b>0.1078</b>

# Deep Neural Networks

# YouTube recommendation

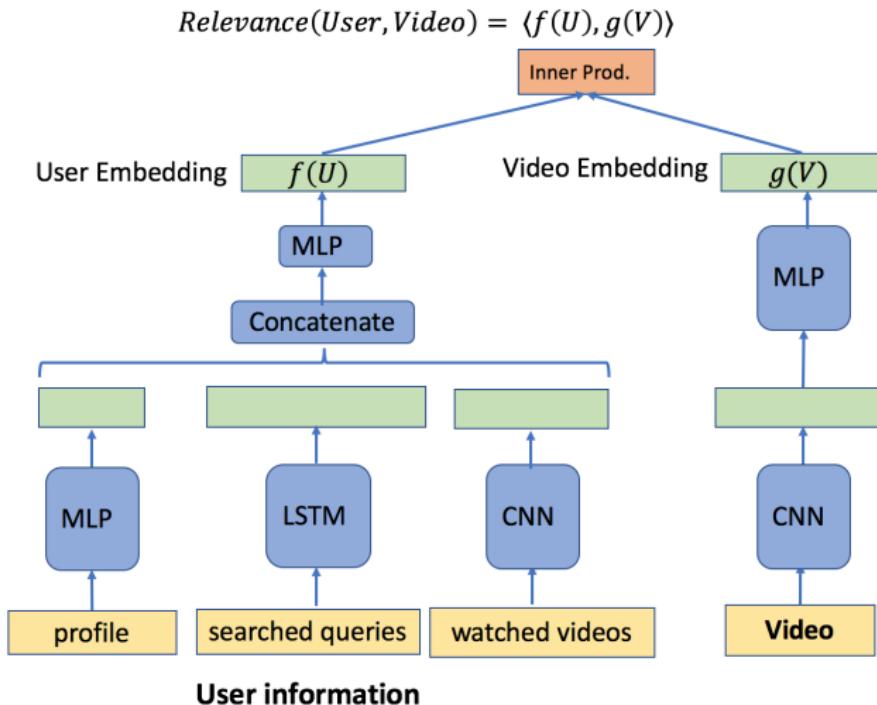
Recall model architecture:



P. Covington, J. Adams, and E. Sargin. *Deep neural networks for youtube recommendations*. Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.

# Multi-Target Prediction with Deep Neural Networks

## Video Recommendation

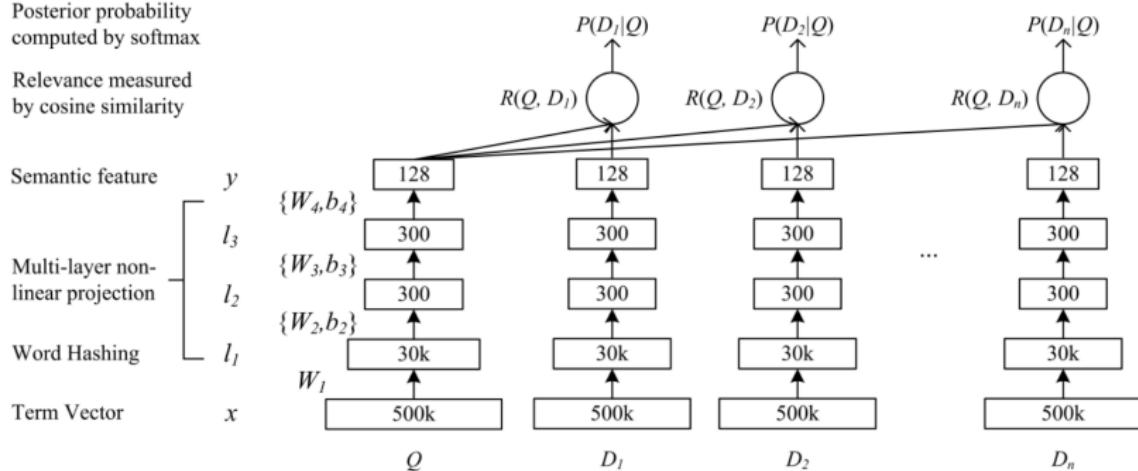


# Multi-Target Prediction with Deep Neural Networks

## Web Search (DSSM Model)

Posterior probability  
computed by softmax

Relevance measured  
by cosine similarity



P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, & L. Heck, [Learning deep structured semantic models for web search using clickthrough data](#). In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (2013)

# DSSM Model

## Web Search

- Training data: 100 million query-document pairs with rich click information sampled from one-year query log files of a commercial search engine.
- Evaluation Data: 16k queries, each with about 15 documents whose relevance scores are labeled as 0-4 by human.

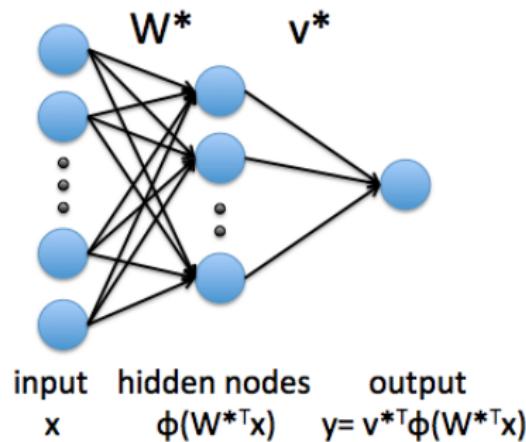
#	Models	NDCG@1	NDCG@3	NDCG@10
1	TF-IDF	0.319	0.382	0.462
2	BM25	0.308	0.373	0.455
3	WTM	0.332	0.400	0.478
4	LSA	0.298	0.372	0.455
5	PLSA	0.295	0.371	0.456
6	DAE	0.310	0.377	0.459
7	BLTM-PR	0.337	0.403	0.480
8	DPM	0.329	0.401	0.479
9	DNN	0.342	0.410	0.486
10	L-WH linear	0.357	0.422	0.495
11	L-WH non-linear	0.357	0.421	0.494
12	<b>L-WH DNN</b>	<b>0.362</b>	<b>0.425</b>	<b>0.498</b>

**Table 2:** Comparative results with the previous state of the art approaches and various settings of DSSM.

P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, & L. Heck, [Learning deep structured semantic models for web search using clickthrough data](#). In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (2013)

# Neural Networks: Theoretical Guarantees

- Small steps: we have shown recovery guarantees for single-target regression problems with one-hidden-layer neural network.



# Neural Networks: Theoretical Guarantees?

- Objective function

$$\widehat{f}_S(W) = \frac{1}{2n} \sum_{j \in [n]} \left( \sum_{i=1}^k v_i^* \varphi(\mathbf{w}_i^\top \mathbf{x}_j) - y_j \right)^2.$$

# Neural Networks: Theoretical Guarantees?

- Objective function

$$\widehat{f}_S(W) = \frac{1}{2n} \sum_{j \in [n]} \left( \sum_{i=1}^k v_i^* \varphi(\mathbf{w}_i^\top \mathbf{x}_j) - y_j \right)^2.$$

- We show **local strongly convexity** near the ground truth.
- Standard gradient descent will converge linearly to the ground truth when initialized appropriately.
- Future work: extend to multi-target with missing values.

# Summary

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

# Summary

- Inductive Matrix Completion:
  - Scales to millions of targets
  - Captures correlations among targets
  - Overcomes missing target values
  - Handles noisy features and non-linear features
  - Extends to PU learning

# Future Work

- Much work to do:
  - Other structures: low-rank+sparse, low-rank+column-sparse (outliers)?
  - Different loss functions?
  - Handling “time” as one of the dimensions — incorporating smoothness through graph regularization?
  - Efficient (parallel) implementations?
  - Guarantees for deep inductive matrix completion?

# Collaborators



Kai-Yang Chiang



Cho-Jui Hsieh



Prateek Jain



Nagarajan Natarajan



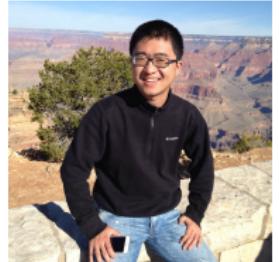
Nikhil Rao



Si Si



Hsiang-fu Yu



Kai Zhong

# References

- [1] P. Jain, and I. S. Dhillon. *Provable inductive matrix completion*. In arXiv preprint arXiv:1306.0626 (2013).
- [2] K. Zhong, P. Jain, I. S. Dhillon. *Efficient Matrix Sensing Using Rank-1 Gaussian Measurements*. In ALT (2015).
- [3] N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. In Bioinformatics, 30(12), i60-i68 (2014).
- [4] H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).
- [5] C-J. Hsieh, N. Natarajan, and I. S. Dhillon. *PU Learning for Matrix Completion*. In ICML (2015).
- [6] S. Si, K.-Y. Chiang, C.-J. Hsieh, N. Rao, and I.S.Dhillon *Goal-Directed Inductive Matrix Completion* In KDD, 2016.
- [7] K.-Y. Chiang, C.-J. Hsieh and I. S. Dhillon. *Matrix Completion with Noisy Side Information* In NIPS (2015).
- [8] K. Zhong, Z. Song, P. Jain, P. L. Bartlett & I. S. Dhillon. *Recovery Guarantees for One-hidden-layer Neural Networks*. In ICML (2017)