6.034 Introduction to Artificial Intelligence

Tommi Jaakkola
MIT CSAIL

The world is drowning in data...

The world is drowning in data...

... access to information is based on recommendations

Recommending news feeds

- Lots of venues (and articles) ... challenging to find the few articles that you are actually interested in reading

Recommending news feeds

- Training examples and corresponding ratings

news articles


Romney Tells Evangelicals Their Values Are His, Too
By ASHLEY PARKER
Speaking at Liberty University, Mitt Romney sought to quell concerns among evangelical voters by offering a forceful defense of Christian values and faith in public life.

x_1


U.S. May Scrap Costly Efforts to Train Iraqi Police Force
By TIM ARANGO 12:14 AM ET
The State Department could jettison a multibillion-dollar training effort by the end of 2012 that has emerged as the latest high-profile example of America's waning influence in the country.

x_2


Candidate in Egypt Makes an Insider's Run for President
By KAREEM FAHIM
Amr Moussa, a former Egyptian foreign minister who served under President Hosni Mubarak, is trying to make a strength from the liability of his long government career.

x_3


Member of Afghan Peace Council Is Assassinated
By ROD NORDLAND and JAWAD SUKHANYAR
7 minutes ago
Arsala Rahmani, a former Taliban minister and current member of Afghan High Peace Council, was shot dead by an unknown gunman in Kabul on Sunday morning, a Kabul police official confirmed.

x_4

$\cdots$

rating	+1	-1	+1	-1	$\cdots$
	y_1	y_2	y_3	y_4	

Recommending news feeds

- Training examples and corresponding ratings

news articles

Romney Tells Evangelicals Their Values Are His, Too
By ASHLEY PARKER
Speaking at Liberty University, Mitt Romney sought to quell concerns among evangelical voters by offering a forceful defense of Christian values and faith in public life.

x_1

U.S. May Scrap Costly Efforts to Train Iraqi Police Force
By TIM ARANGO 12:14 AM ET
The State Department could jettison a multibillion-dollar training effort by the end of 2012 that has emerged as the latest high-profile example of America's waning influence in the country.

x_2

Candidate in Egypt Makes an Insider's Run for President
By KAREEM FAHIM
Amr Moussa, a former Egyptian foreign minister who served under President Hosni Mubarak, is trying to make a strength from the liability of his long government career.

x_3

Member of Afghan Peace Council Is Assassinated
By ROD NORDLAND and JAWAD SUKHANYAR
7 minutes ago
Arsala Rahmani, a former Taliban minister and current member of Afghan High Peace Council, was shot dead by an unknown gunman in Kabul on Sunday morning, a Kabul police official confirmed.

x_4

$\cdots$

feature vectors $\phi(x_1)$ $\phi(x_2)$ $\phi(x_3)$ $\phi(x_4)$ $\cdots$

rating +1 -1 +1 -1 $\cdots$

y_1 y_2 y_3 y_4

Recommending news feeds

- Training examples and corresponding ratings

news articles

Romney Tells Evangelicals Their Values Are His, Too
By ASHLEY PARKER
Speaking at Liberty University, Mitt Romney sought to quell concerns among evangelical voters by offering a forceful defense of Christian values and faith in public life.

U.S. May Scrap Costly Efforts to Train Iraqi Police Force
By TIM ARANGO 12:14 AM ET
The State Department could jettison a multibillion-dollar training effort by the end of 2012 that has emerged as the latest high-profile example of America's waning influence in the country.

Candidate in Egypt Makes an Insider's Run for President
By KAREEM FAHIM
Amr Moussa, a former Egyptian foreign minister who served under President Hosni Mubarak, is trying to make a strength from the liability of his long government career.

Member of Afghan Peace Council Is Assassinated
By ROD NORDLAND and JAWAD SUKHANYAR
7 minutes ago
Arsala Rahmani, a former Taliban minister and current member of Afghan High Peace Council, was shot dead by an unknown gunman in Kabul on Sunday morning, a Kabul police official confirmed.

news articles	x_1	x_2	x_3	x_4	$\cdots$
feature vectors	$\phi(x_1)$	$\phi(x_2)$	$\phi(x_3)$	$\phi(x_4)$	$\cdots$
rating	+1	-1	+1	-1	$\cdots$
	y_1	y_2	y_3	y_4	

Articles as feature vectors

- Does the word order matter?

White House officials consulted with the Justice Department in preparing a list of U.S. attorneys who would be removed.

(NYT 03/13/07)

x

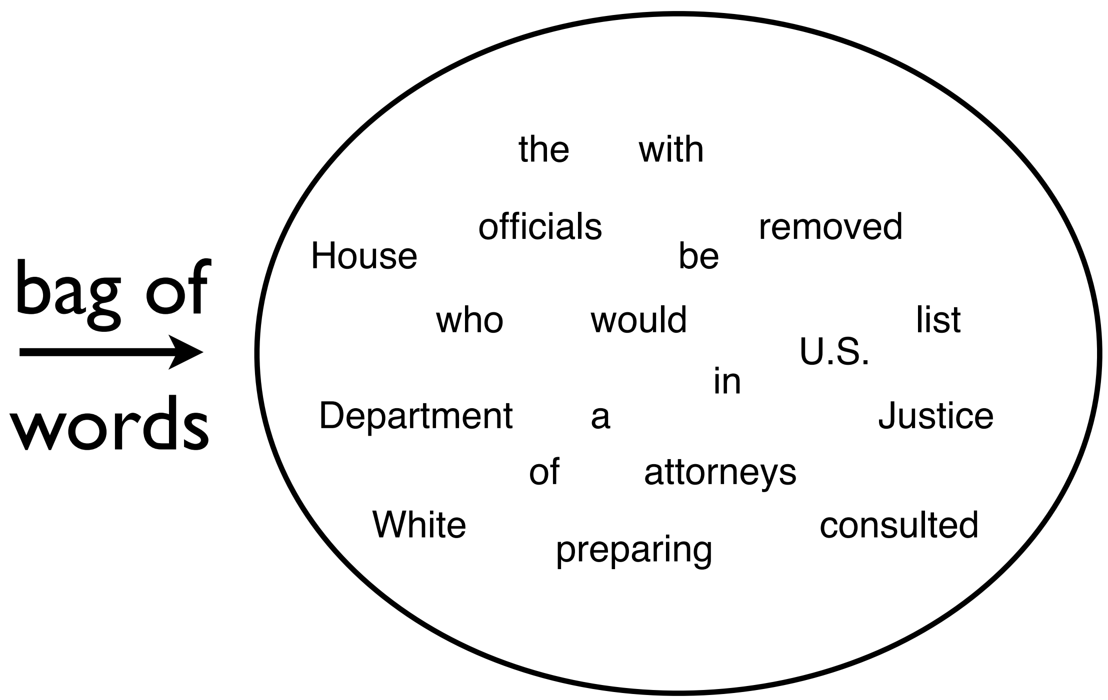
Articles as feature vectors

- Does the word order matter?

White House officials consulted with the Justice Department in preparing a list of U.S. attorneys who would be removed.

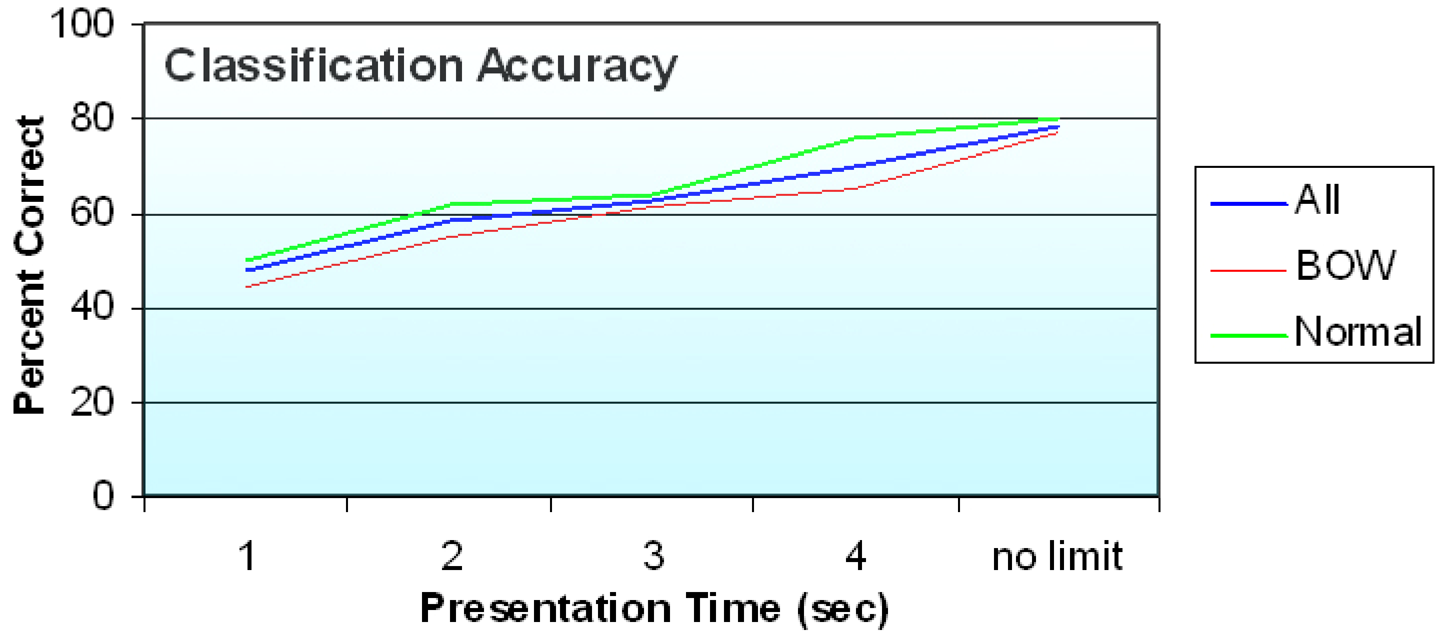
(NYT 03/13/07)

x



Does the word order matter?

- Not for every task...



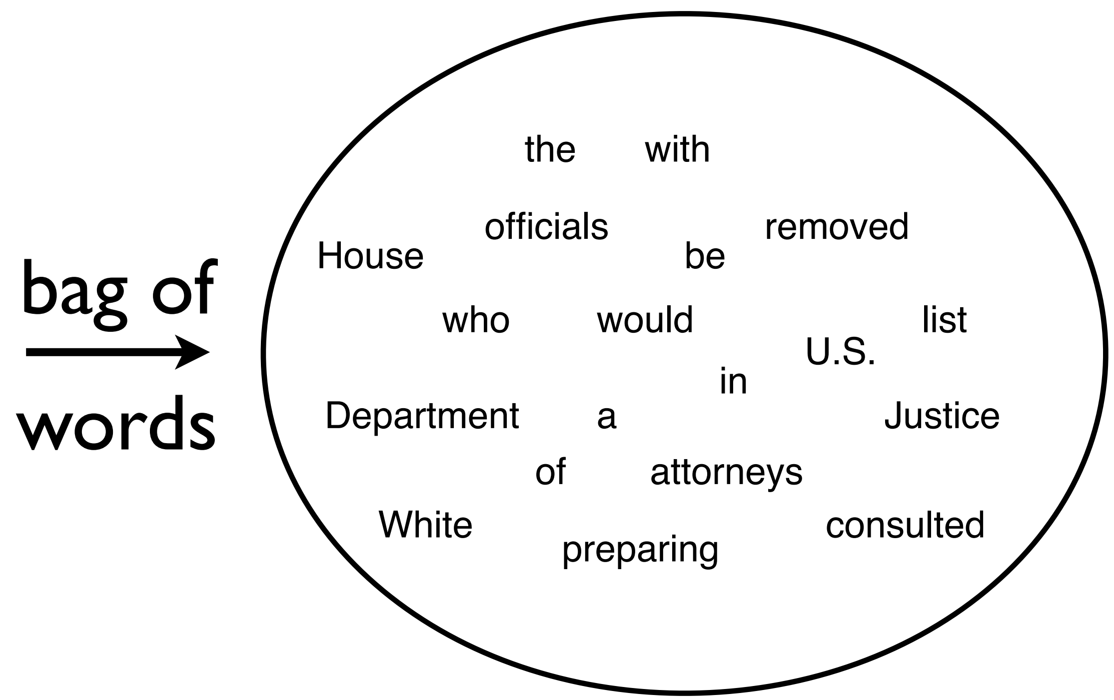
(Wolf et al. 2006)

Articles as feature vectors

White House officials consulted with the Justice Department in preparing a list of U.S. attorneys who would be removed.

(NYT 03/13/07)

x

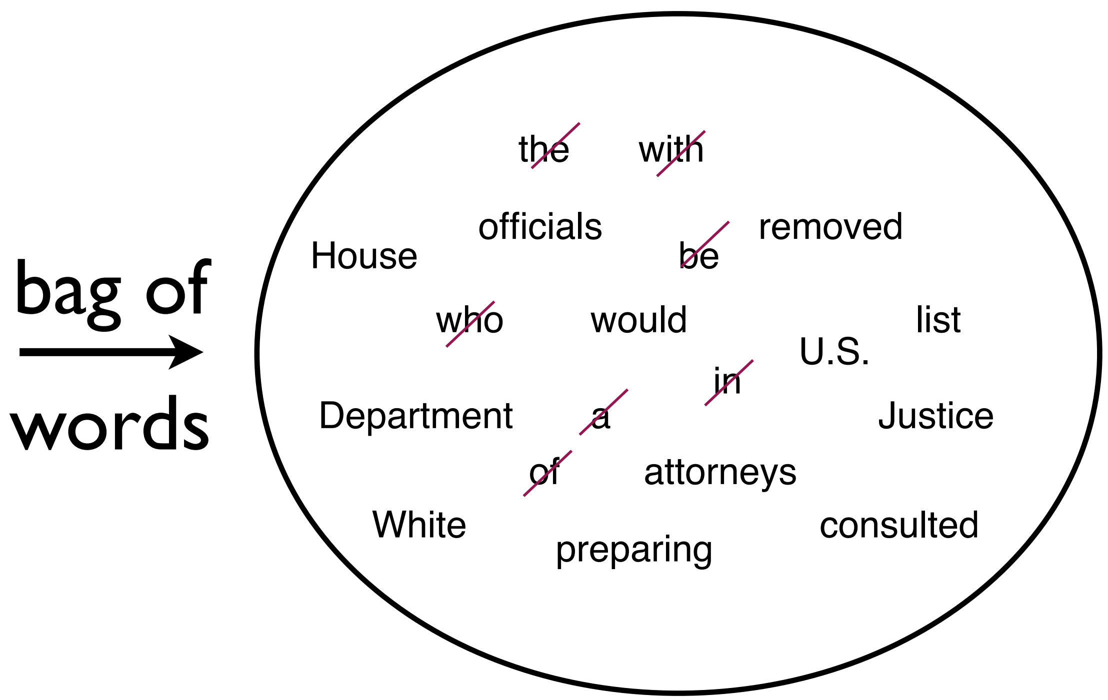


Articles as feature vectors

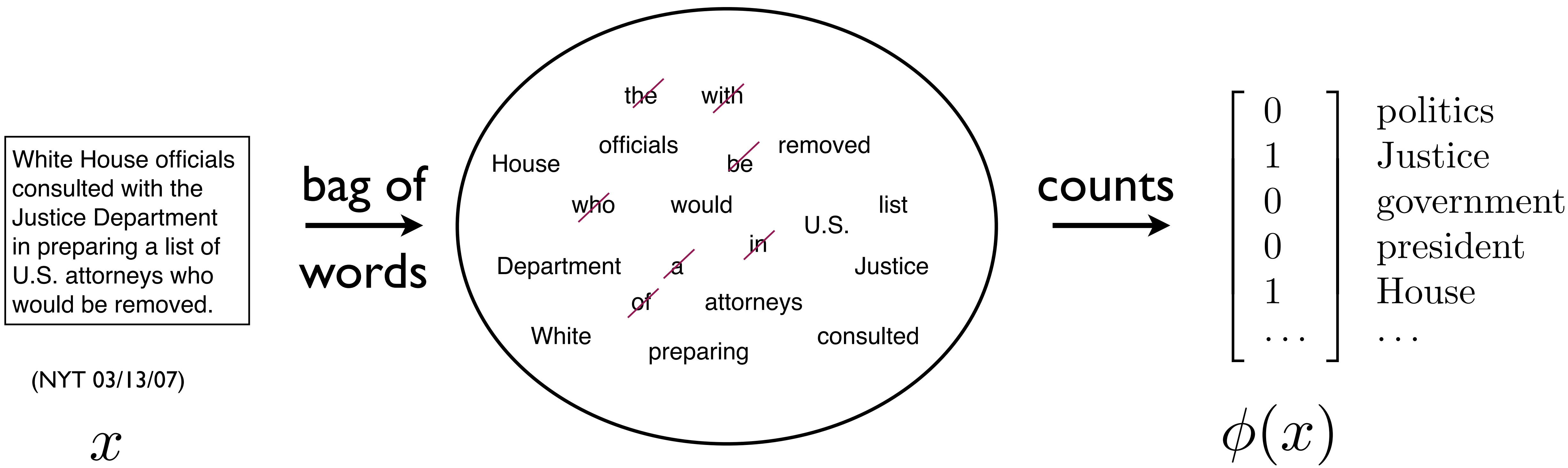
White House officials consulted with the Justice Department in preparing a list of U.S. attorneys who would be removed.

(NYT 03/13/07)

x

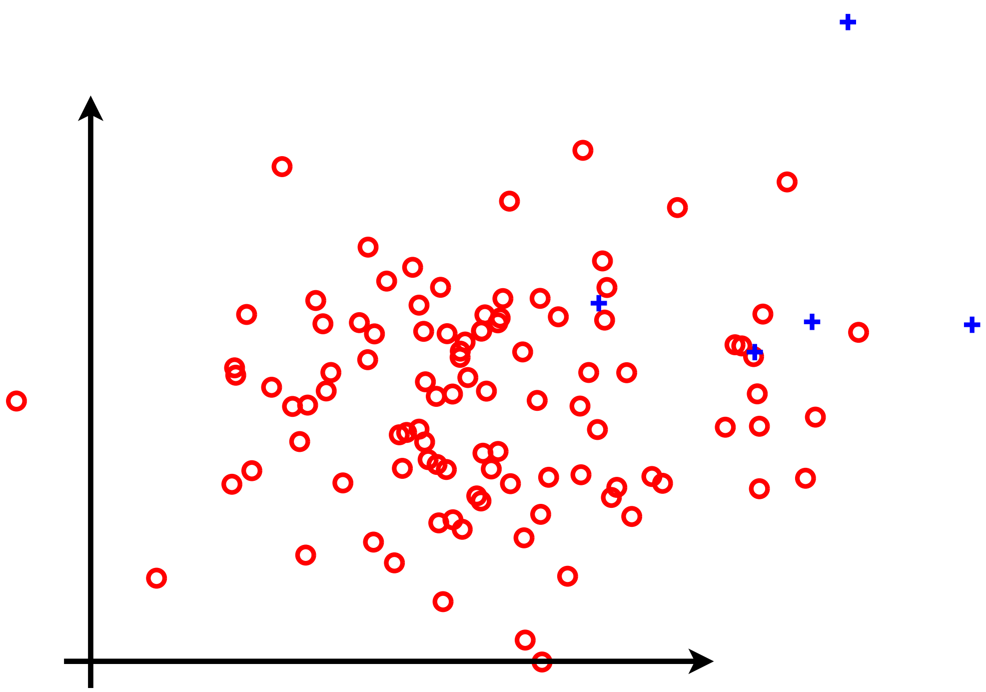


Articles as feature vectors



Recommending news feeds

- A few examples of articles that we'd like to read (+1)
- Potentially a large number of unwanted articles (-1)

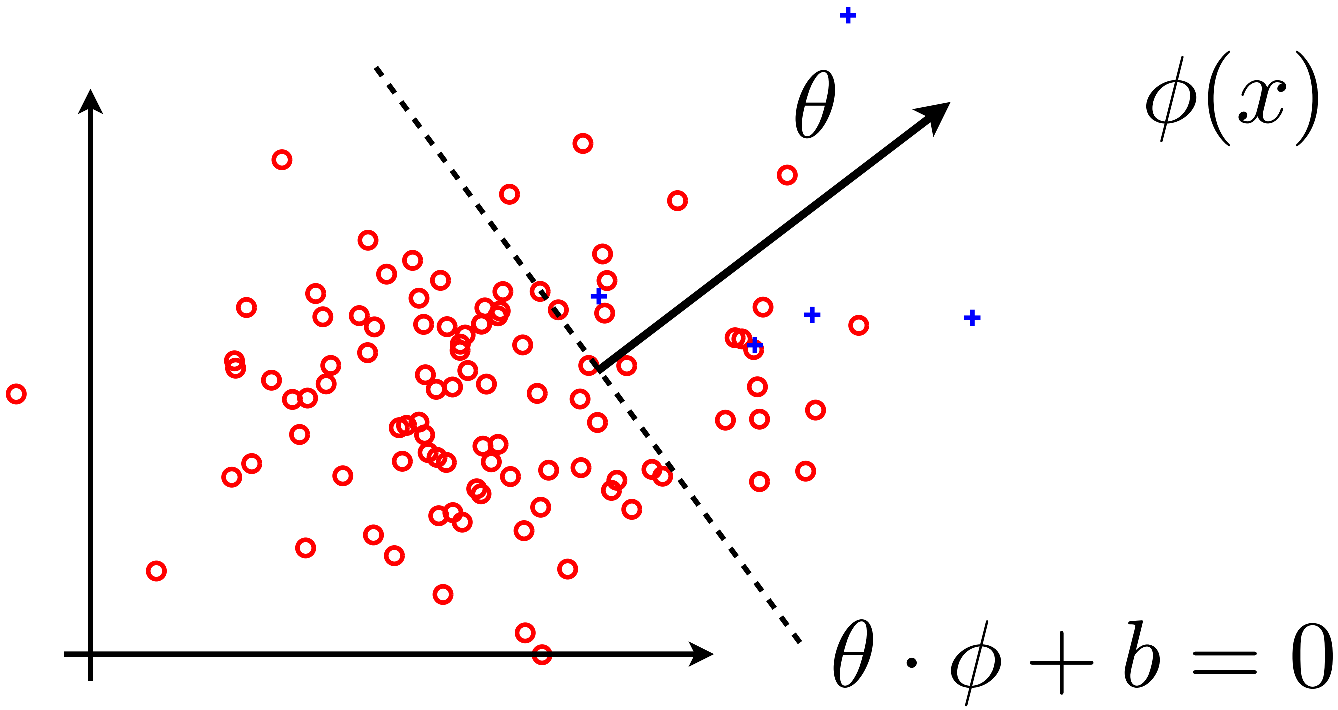


$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \cdots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \cdots \end{matrix}$$

Recommending news feeds

- A few examples of articles that we'd like to read (+1)
- Potentially a large number of unwanted articles (-1)

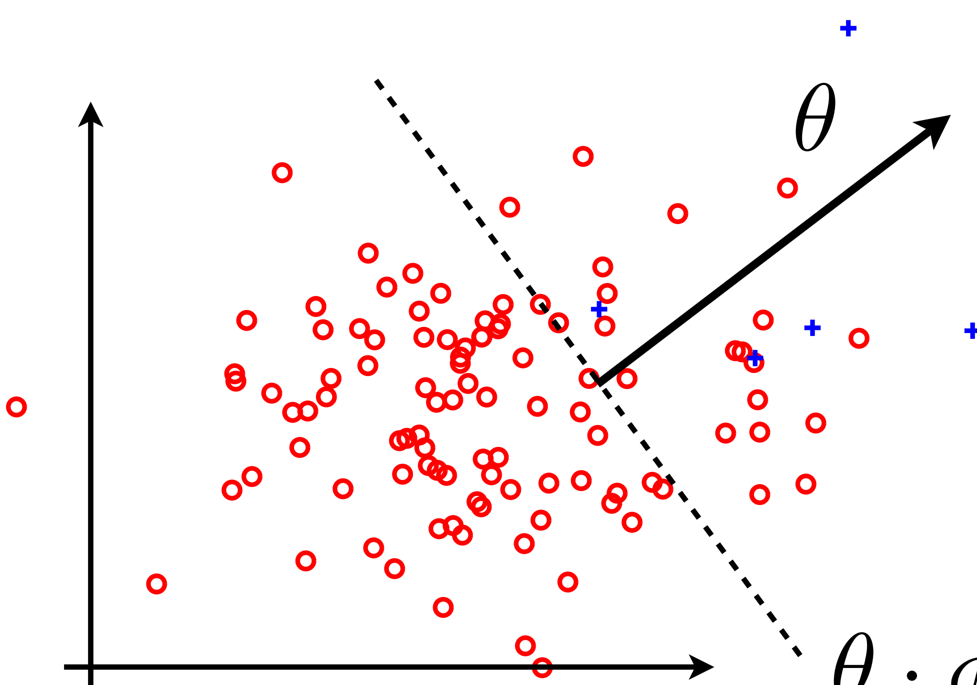
linear preferences $y(x) = \theta \cdot \phi(x) + b$



$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \dots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \dots \end{matrix}$$

Recommending news feeds

- Why is the problem challenging?
  - lots of possible words
  - only a small subset appears in any particular article
  - most frequent words are not content words
  - meaningful classes of articles are typically tied to words that occur relatively infrequently
  - any two articles in the same meaningful class may have only a few content words in common



$$\theta \cdot \phi + b = 0$$

$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \ldots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \ldots \end{matrix}$$

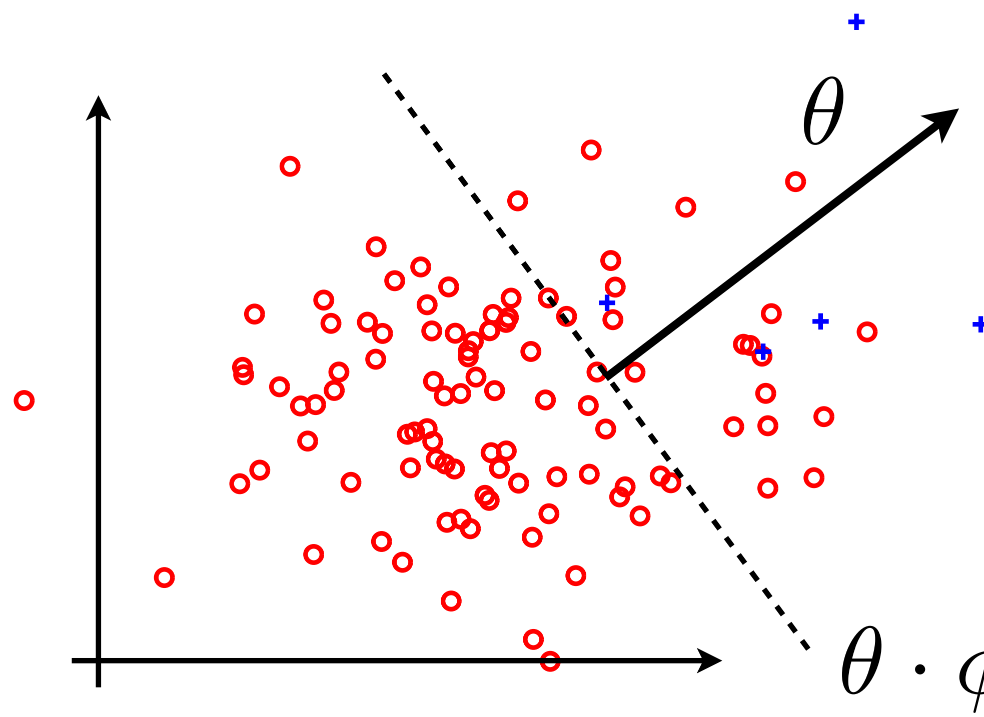
Some tricks

- We can transform the counts in the feature vectors so as to emphasize more "relevant" words
- TFIDF weighting

$$\phi_w(\mathbf{x}) \quad = \quad \overbrace{\left(\begin{array}{c} \text{freq. of word} \\ w \text{ in doc. } \mathbf{x} \end{array} \right)}^{\text{TF}} \cdot \overbrace{\log \left[\frac{\text{\# of docs}}{\text{\# of docs with word } w} \right]}^{\text{IDF}}$$

Recommending news feeds

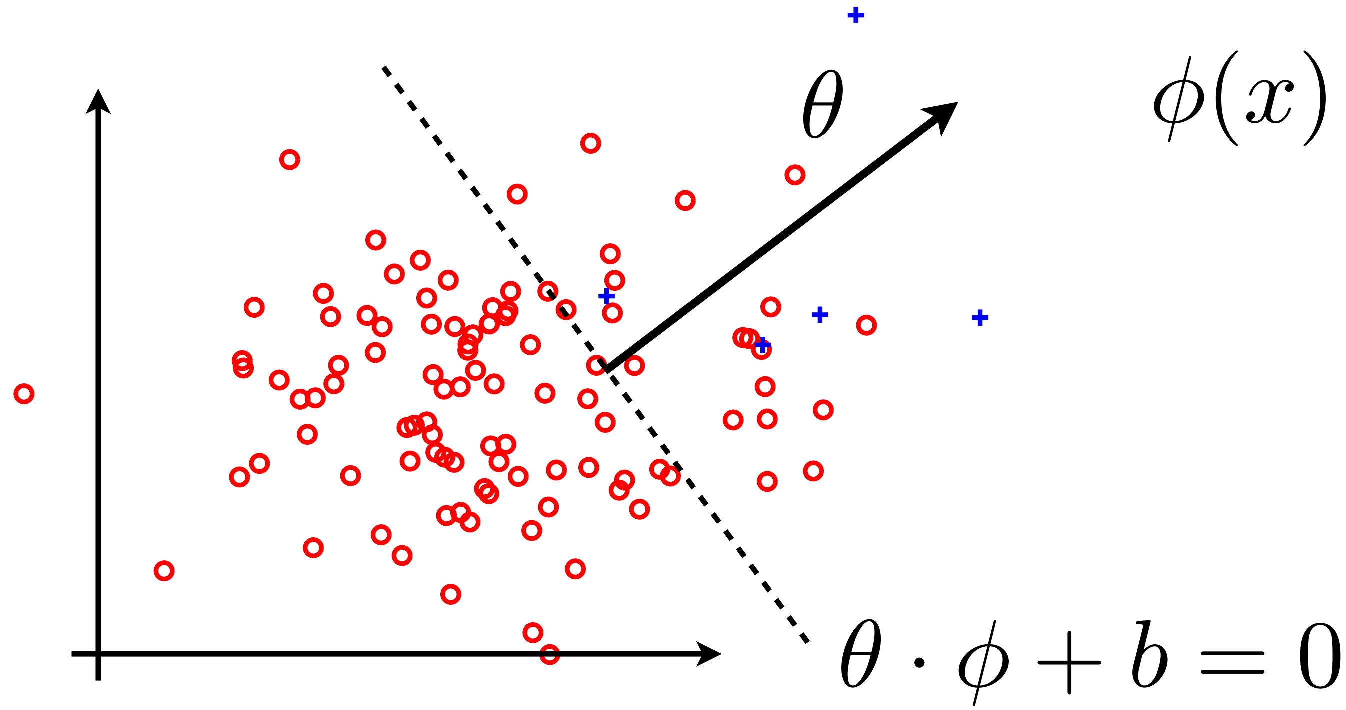
linear preferences $y(x) = \theta \cdot \phi(x) + b$



$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \ldots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \ldots \end{matrix}$$

Recommending news feeds

linear preferences $y(x) = \theta \cdot \phi(x) + b$



$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \cdots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \cdots \end{matrix}$$

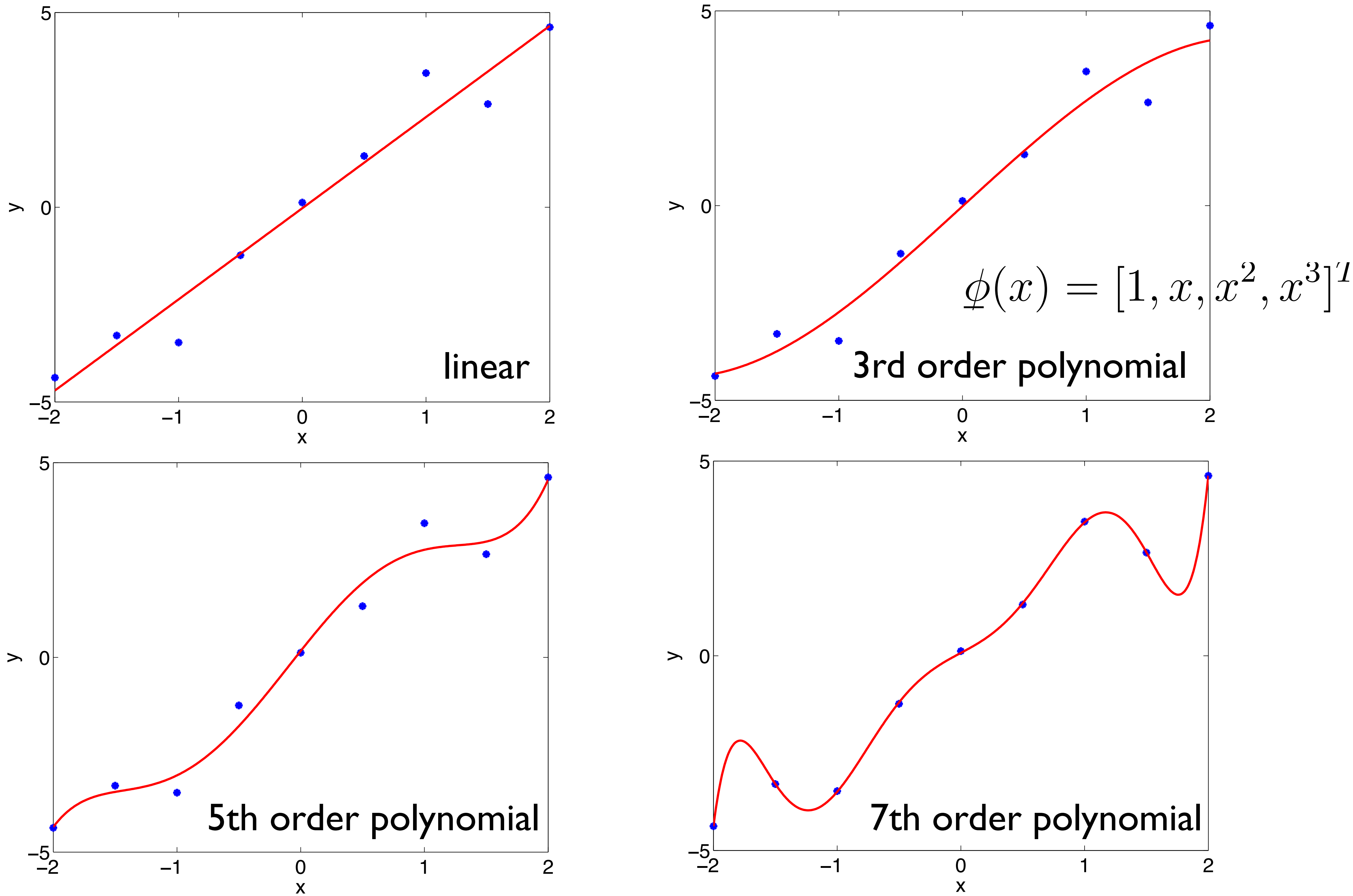
$$J(\theta, b) = \sum_{t=1}^{n} (y_t - \theta \cdot \phi(x_t) - b)^2$$

sum over the training examples

squared prediction error on each example

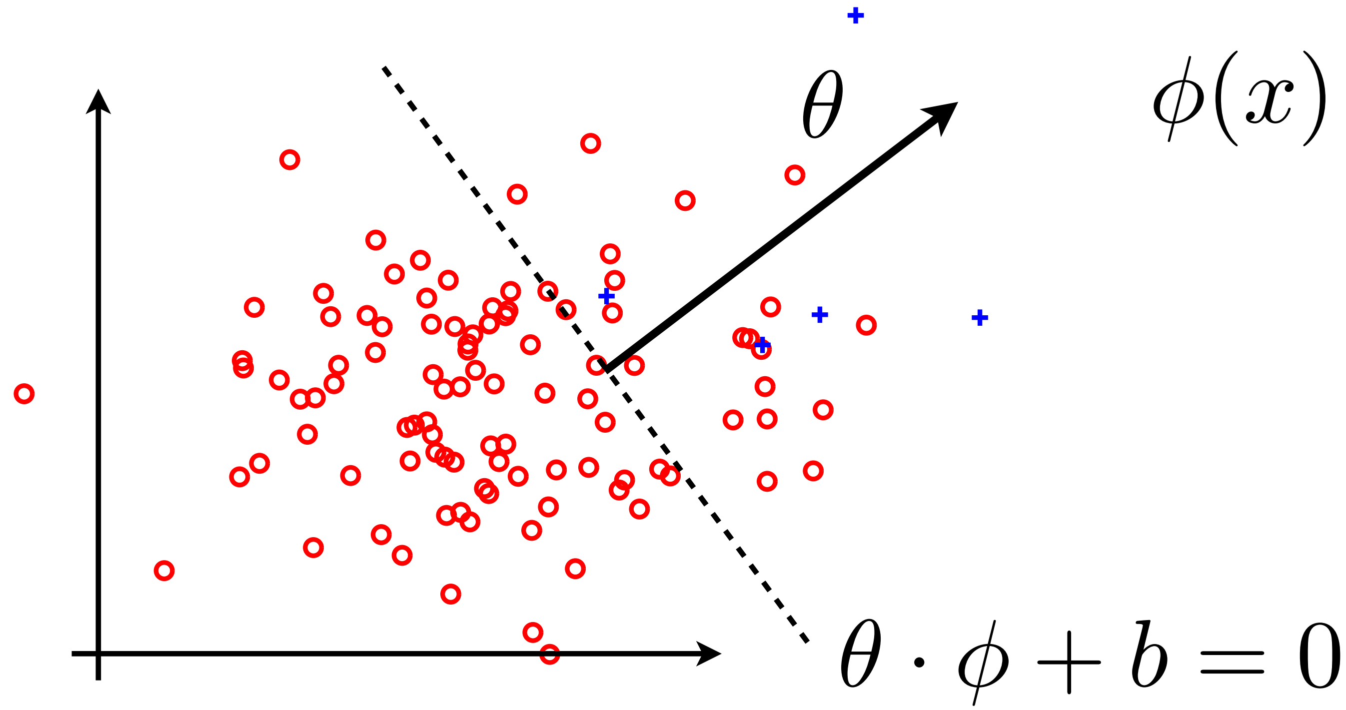
Linear regression, complexity

- We can easily obtain (too) complex regression functions by considering different feature mappings



Recommending news feeds

linear preferences $y(x) = \theta \cdot \phi(x) + b$



$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \cdots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \cdots \end{matrix}$$

$$J(\theta, b) = \sum_{t=1}^{n} (y_t - \theta \cdot \phi(x_t) - b)^2$$

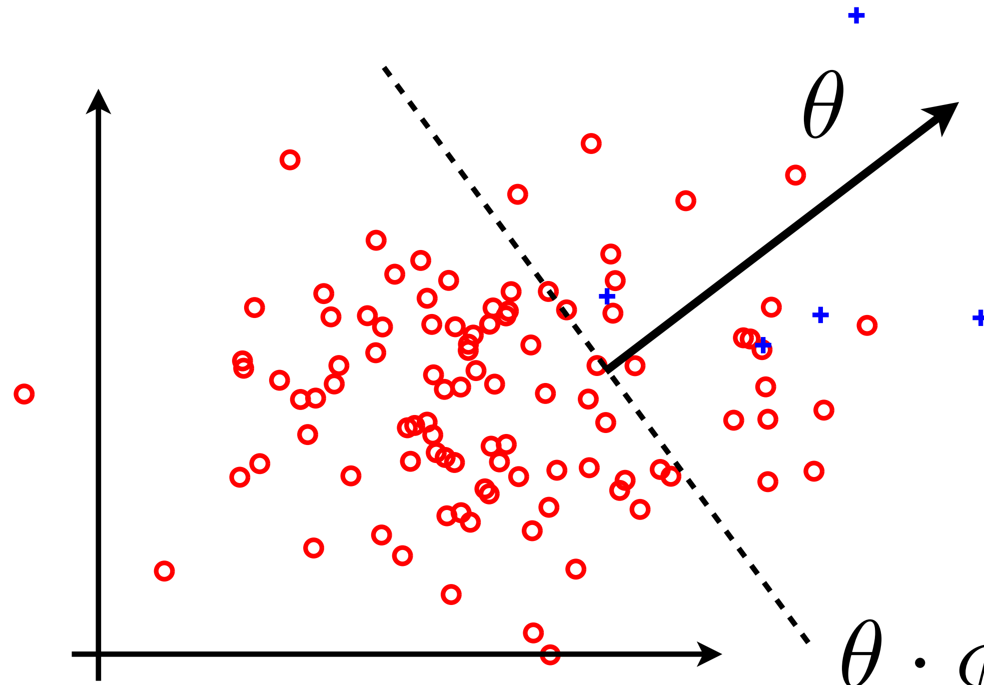
sum over the training examples

squared prediction error on each example

Recommending news feeds

linear preferences $y(x) = \theta \cdot \phi(x) + b$

$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \cdots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \cdots \end{matrix}$$



$$J(\theta, b) = \sum_{t=1}^{n} (y_t - \theta \cdot \phi(x_t) - b)^2 \quad + \lambda \|\theta\|^2$$

sum over the training examples

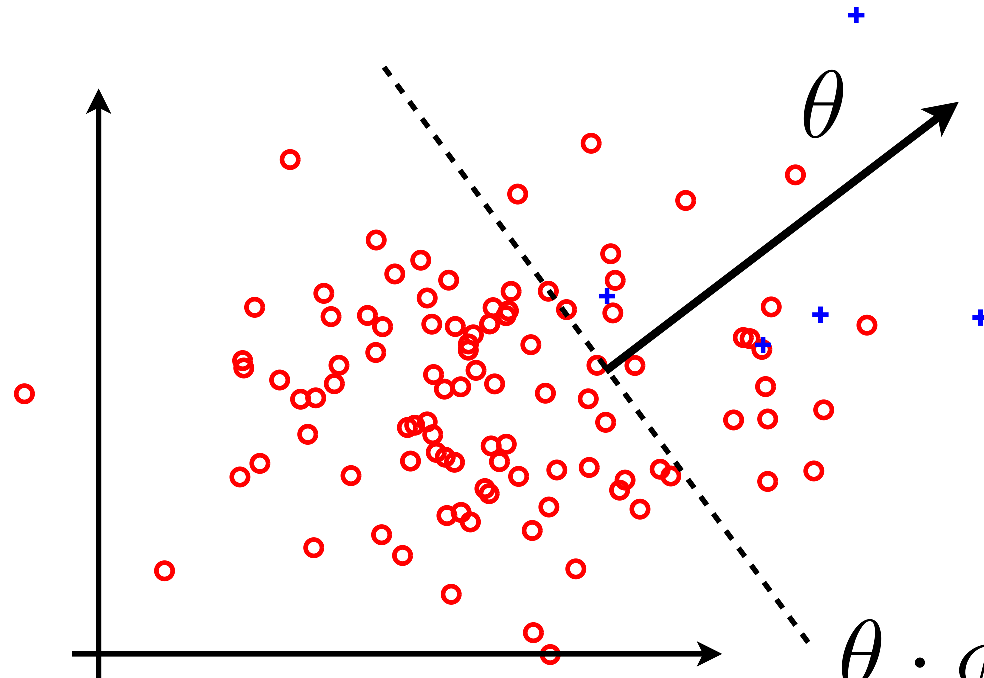
squared prediction error on each example

regularization term

Recommending news feeds

linear preferences $y(x) = \theta \cdot \phi(x) + b$

$$\phi(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \cdots \end{bmatrix} \begin{matrix} \text{politics} \\ \text{Justice} \\ \text{government} \\ \text{president} \\ \text{House} \\ \cdots \end{matrix}$$



$$\theta \cdot \phi + b = 0$$

$$J(\theta, \cancel{b}) = \sum_{t=1}^{n} (y_t - \theta \cdot \phi(x_t) - \cancel{b})^2 \quad + \lambda \|\theta\|^2$$

sum over the training examples

squared prediction error on each example

regularization term

Today's topics

- Preface: regression for recommendation problems
- Collaborative filtering
  - setup, regression formulation
  - matrix factorization

Collaborative filtering

- Consider the problem of predicting how n users rate m movies
- Known ratings (training data) are arranged in a partially filled nxm data matrix
- The goal is to predict the remaining entries

m movies

n users

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

Collaborative filtering

- Consider the problem of predicting how n users rate m movies
- Known ratings (training data) are arranged in a partially filled nxm data matrix
- The goal is to predict the remaining entries
- Basic intuition: similar users can complete each others experience

m movies

n users

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

Collaborative filtering

- Consider the problem of predicting how n users rate m movies
- Known ratings (training data) are arranged in a partially filled nxm data matrix
- The goal is to predict the remaining entries
- Basic intuition: similar users can complete each others experience

m movies

n users

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5		4	5	1		4			
		5				5		3	

Collaborative filtering

- Consider the problem of predicting how n users rate m movies
- Known ratings (training data) are arranged in a partially filled nxm data matrix
- The goal is to predict the remaining entries
- Basic intuition: similar users can complete each others experience
- Key part of the problem is to couple the estimation tasks across users / movies

m movies

n users

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5		4	5	1		4			
		5				5		3	

Collaborative filtering

- Our goal is to fill the data matrix, i.e., accurately predict values for unobserved entries
- Computational issues:
  - a typical matrix is very large, e.g., n=400K, m=17K
- Statistical issues:
  - the matrix is very sparse, e.g., 1% known ratings
  - ratings may be diverse and under-sampled (?)
- Formulation issues:
  - many interpretations for missing entries

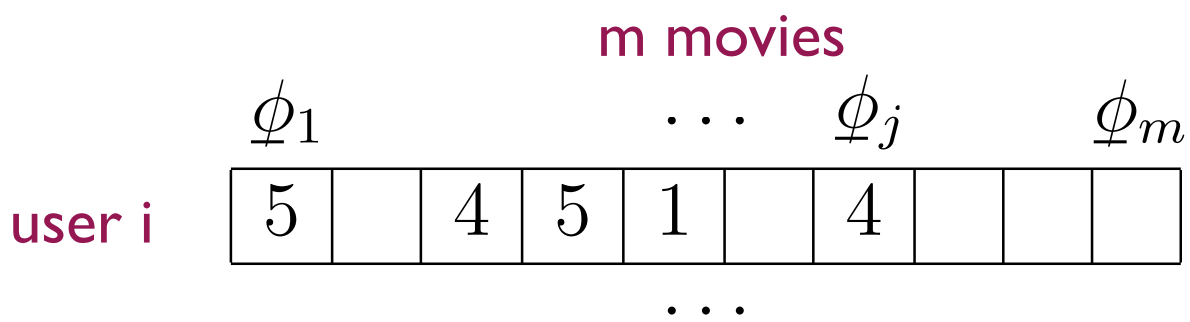
m movies

n users

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

Single user predictions

- We could try to solve the problem separately for each user using simple linear regression models for ratings



$$dim(\underline{\theta}_i) = dim(\underline{\phi}_j) = d$$

$$J_i(\underline{\theta}_i) = \sum_{j \in M_i} (Y_{ij} - \underline{\theta}_i \cdot \underline{\phi}_j)^2 + \lambda \|\underline{\theta}_i\|^2$$

known entries for user i — rating matrix — user i parameters — feature vector for movie j

Single user predictions

- We could try to solve the problem separately for each user using simple linear regression models for ratings

m movies

	ϕ_1				$\cdots$		ϕ_j			ϕ_m
user i	5		4	5	1		4			

$\cdots$

$$dim(\underline{\theta}_i) = dim(\underline{\phi}_j) = d$$

$$J_i(\underline{\theta}_i) = \sum_{j \in M_i} (Y_{ij} - \underline{\theta}_i \cdot \underline{\phi}_j)^2 + \lambda \|\underline{\theta}_i\|^2$$

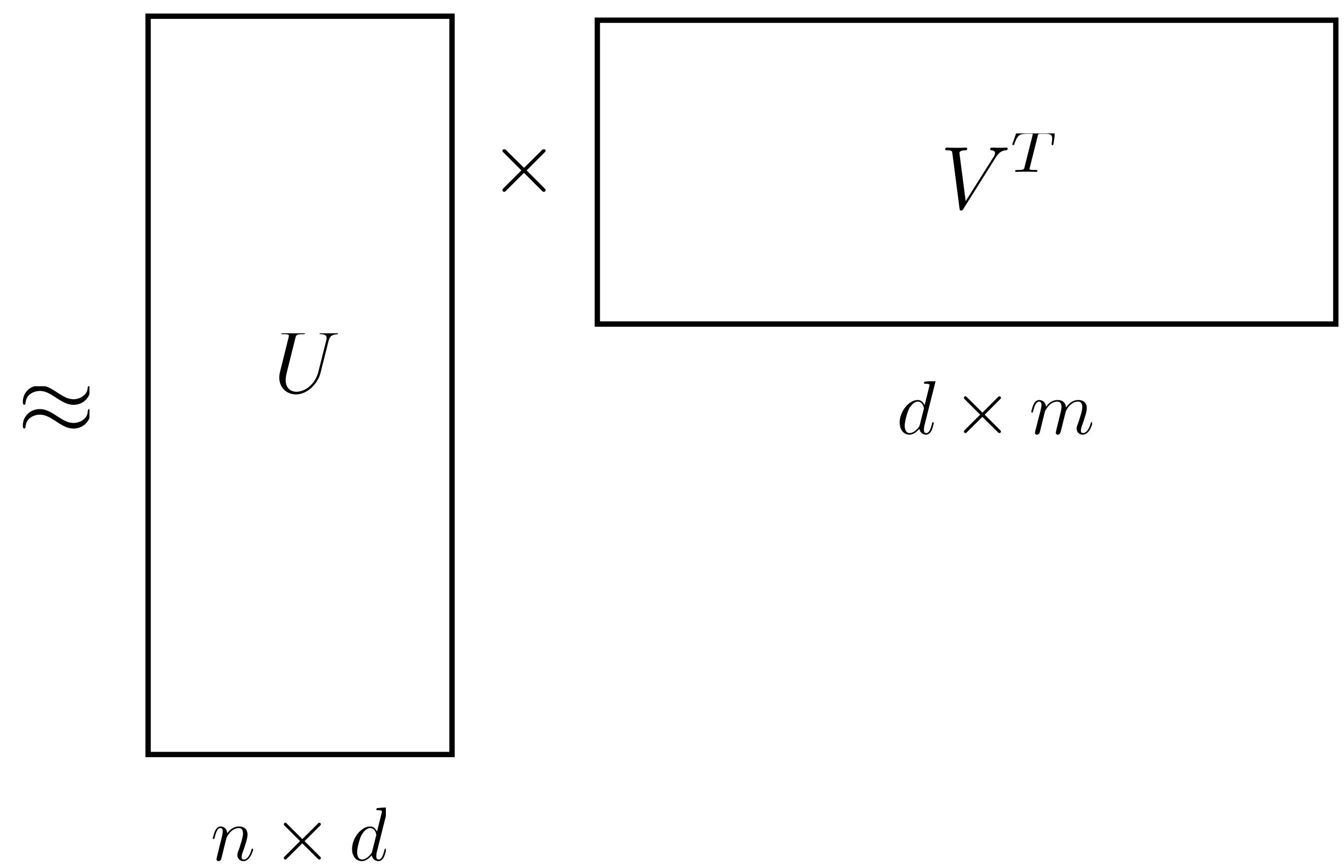
- $j \in M_i$: known entries for user i
- Y_{ij}: rating matrix
- $\underline{\theta}_i$: user i parameters
- $\underline{\phi}_j$: feature vector for movie j

- But
  - reasonable feature vectors may be hard to obtain
  - each user may have only a few ratings
  - no help from similar users

Matrix factorization

- We can approximate the rating matrix as a product of two lower rank matrices

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	



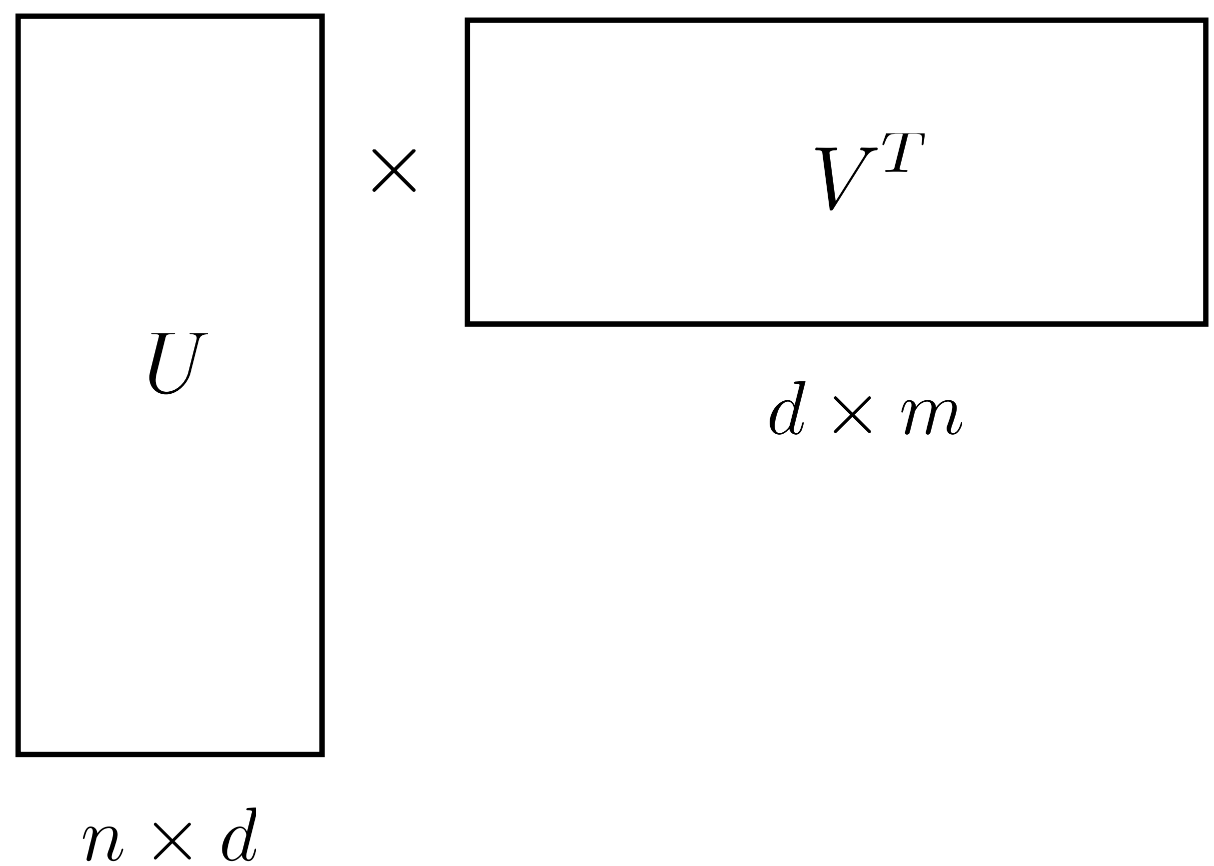
$$Y_{ij} \approx [UV^T]_{ij}$$

Matrix factorization

- We can approximate the rating matrix as a product of two lower rank matrices

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

$\approx$



$$Y_{ij} \approx [UV^T]_{ij}$$

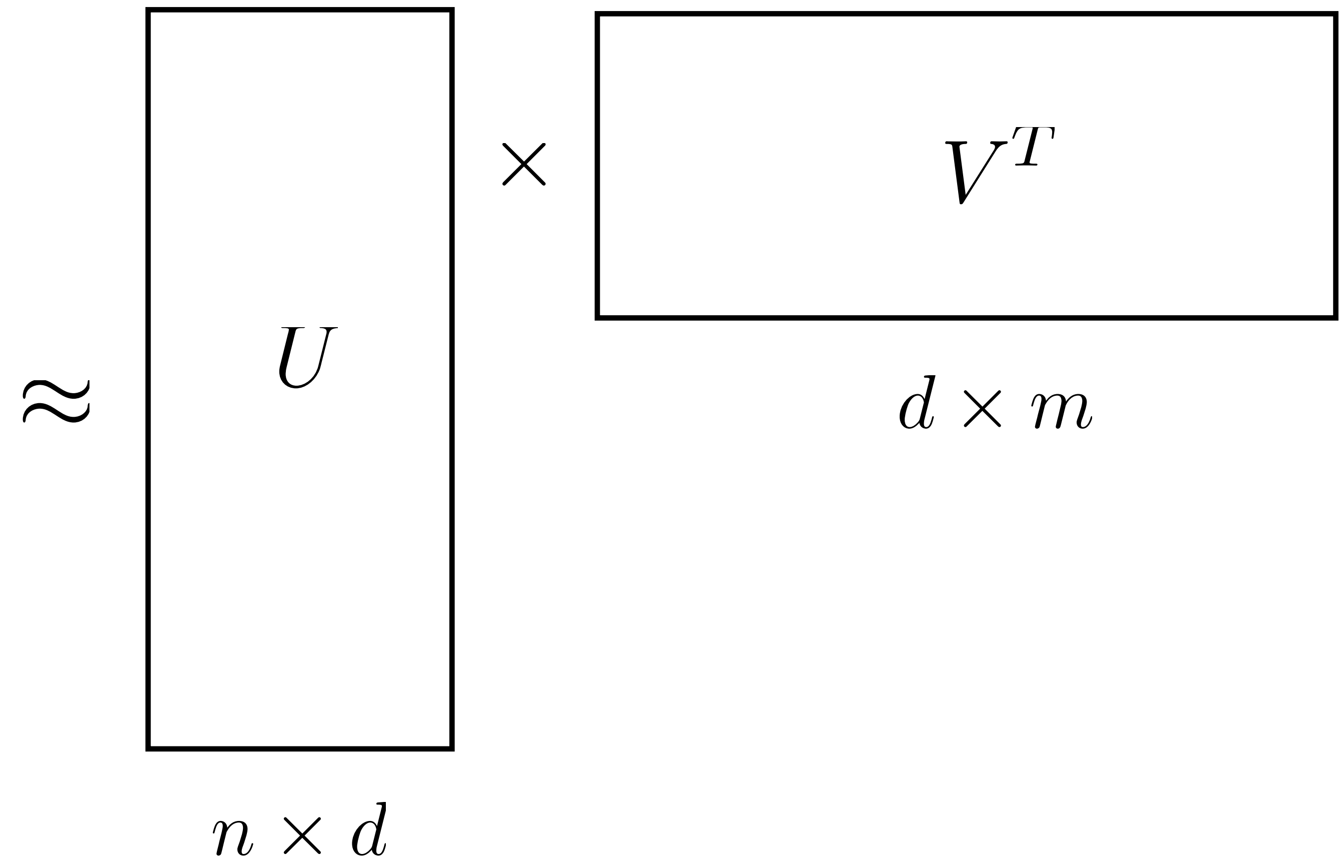
$$\min_{U,V} \sum_{ij \in D} (Y_{ij} - [UV^T]_{ij})^2 .$$

observed entries

Matrix factorization

- We can approximate the rating matrix as a product of two lower rank matrices

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	



$$Y_{ij} \approx [UV^T]_{ij}$$

$$\min_{U,V} \sum_{ij \in D} (Y_{ij} - [UV^T]_{ij})^2 .$$

observed entries

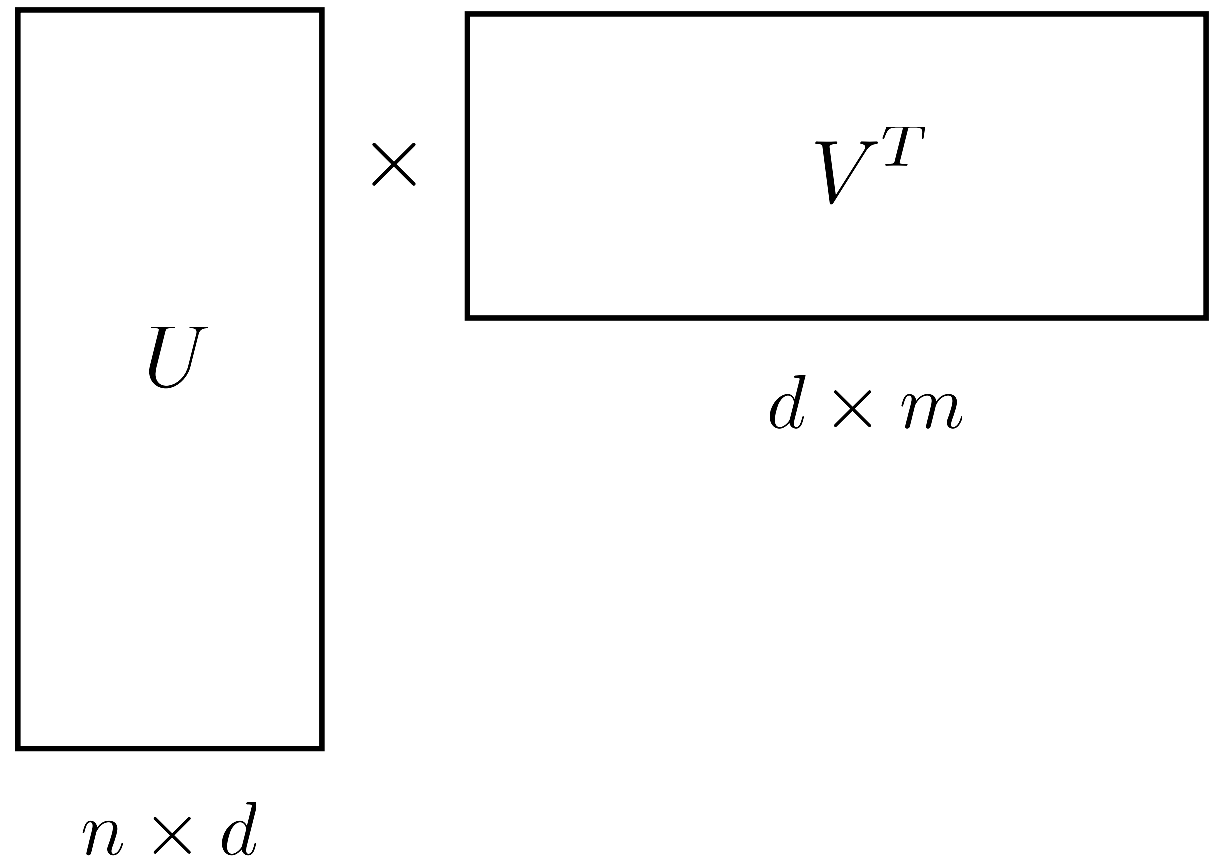
the only complexity control would be the rank d

Matrix factorization

- We can approximate the rating matrix as a product of two lower rank matrices

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

$\approx$



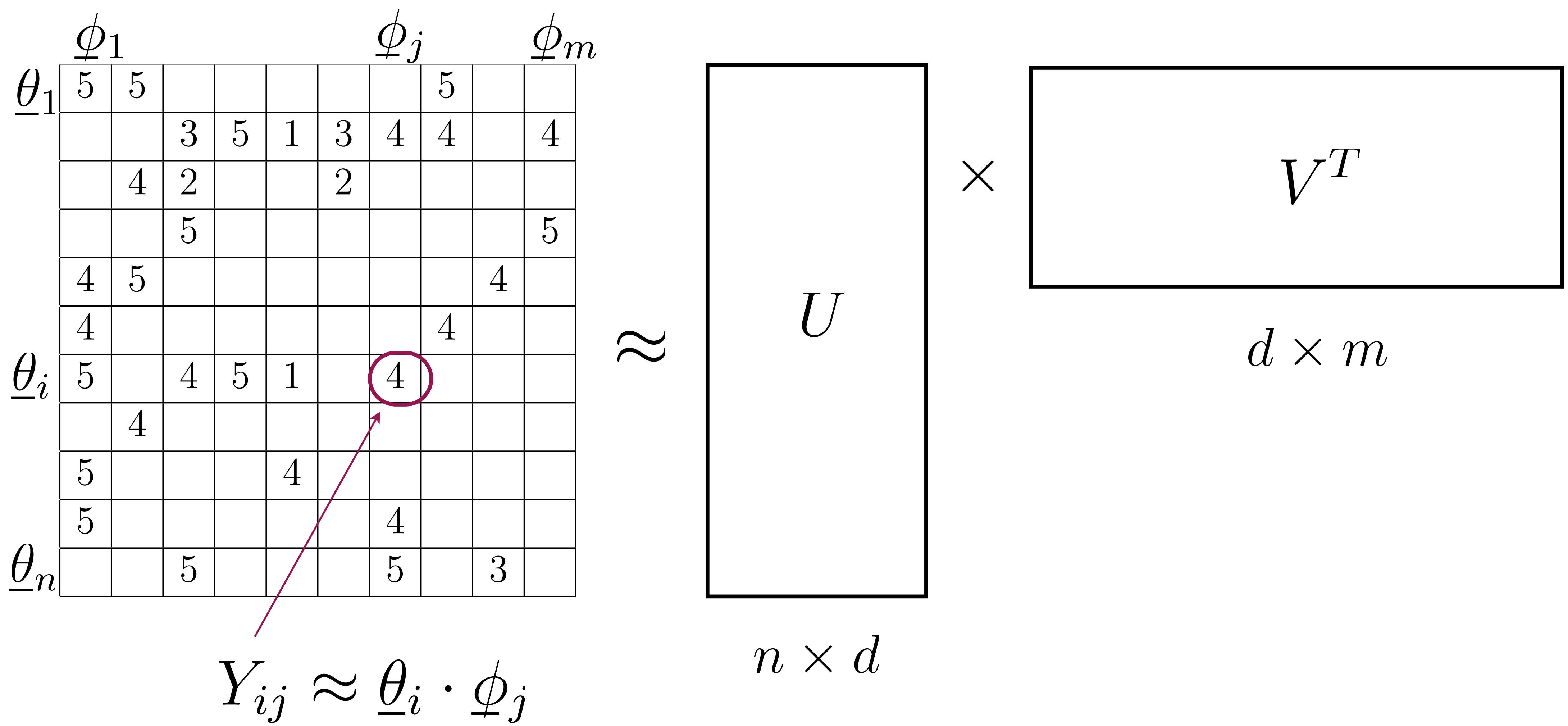
$$Y_{ij} \approx [UV^T]_{ij}$$

$$\min_{U,V} \sum_{ij \in D} (Y_{ij} - [UV^T]_{ij})^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

observed entries

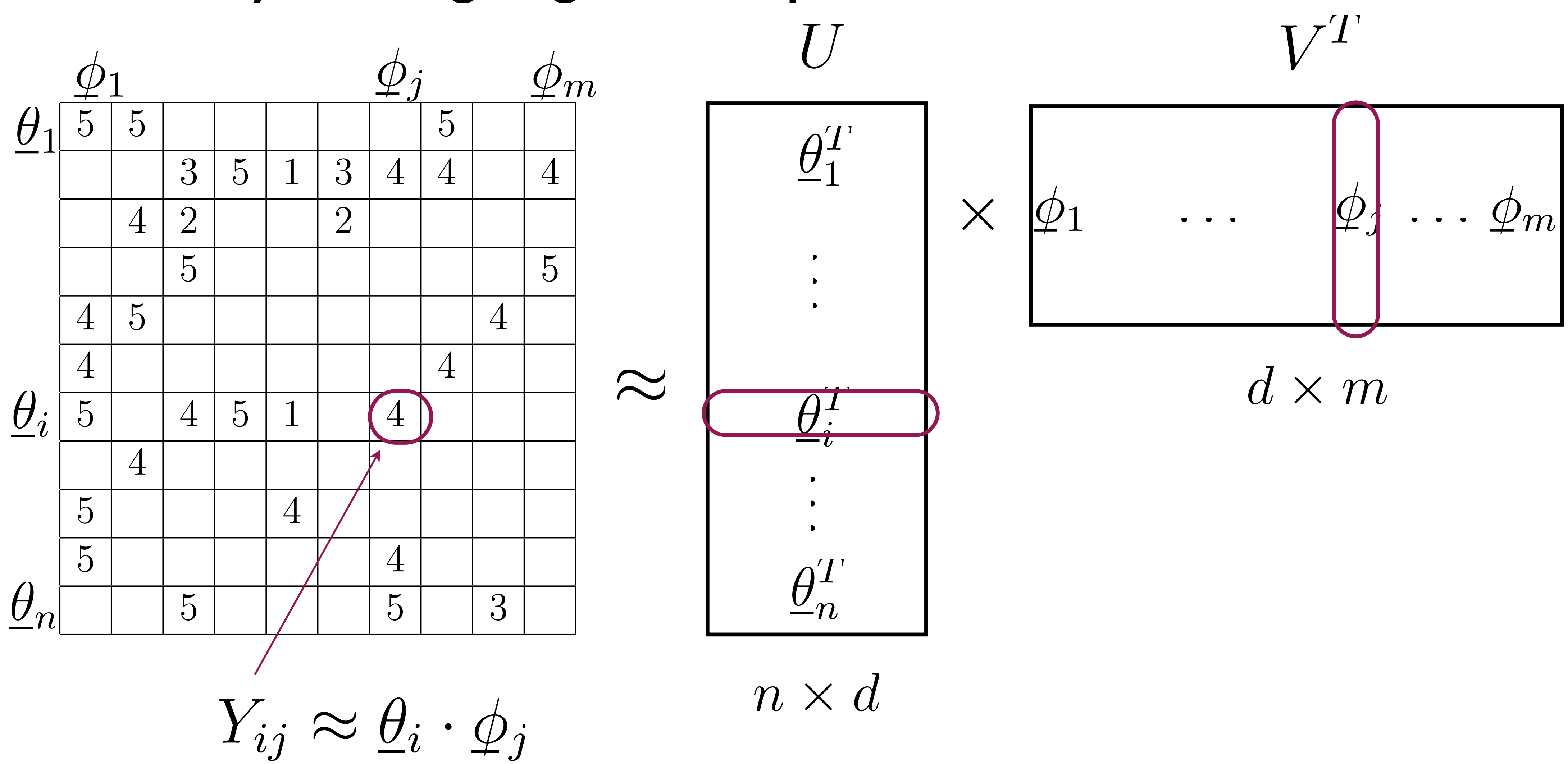
Matrix factorization

- The matrix factorization approach can be interpreted as iteratively solving regression problems for users/movies



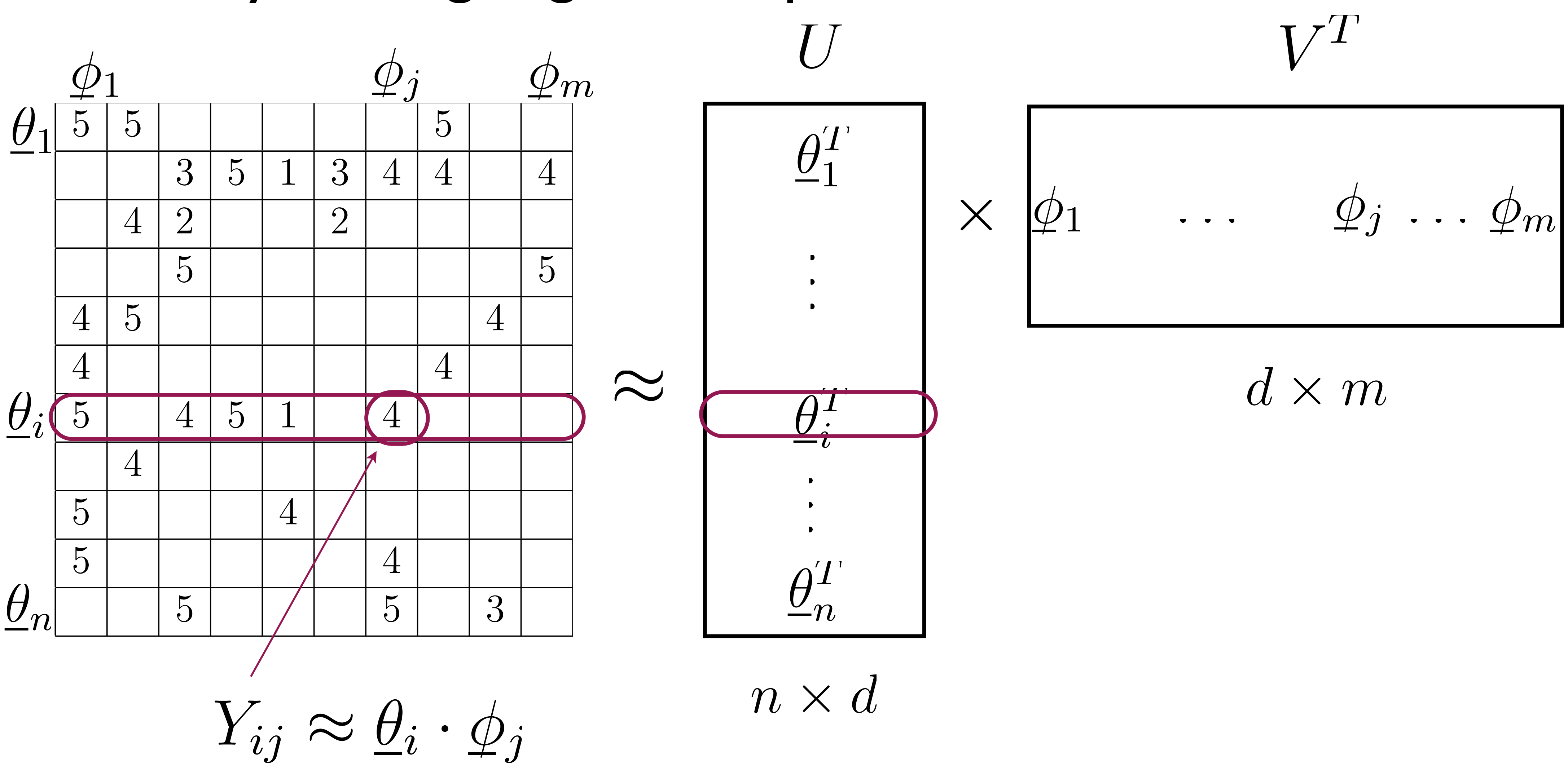
Matrix factorization

- The matrix factorization approach can be interpreted as iteratively solving regression problems for users/movies



Matrix factorization

- The matrix factorization approach can be interpreted as iteratively solving regression problems for users/movies

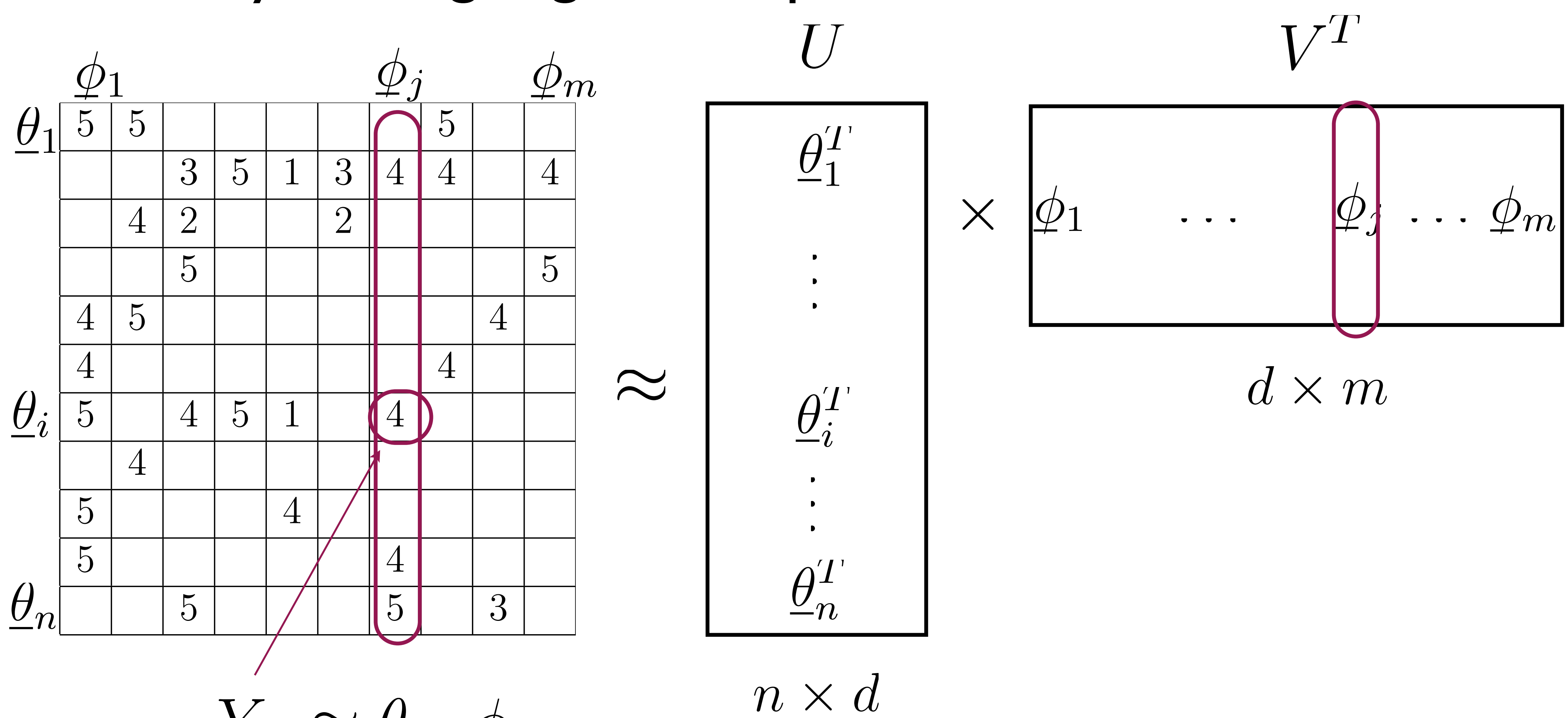


$$J_i(\underline{\theta}_i) = \sum_{j:ij \in D} (Y_{ij} - \underline{\theta}_i \cdot \underline{\phi}_j)^2 + \lambda \|\underline{\theta}_i\|^2$$

regression problem for each user with fixed movie features

Matrix factorization

- The matrix factorization approach can be interpreted as iteratively solving regression problems for users/movies



$$Y_{ij} \approx \underline{\theta}_i \cdot \underline{\phi}_j$$

$$J_j(\underline{\phi}_j) = \sum_{i:ij \in D} (Y_{ij} - \underline{\theta}_i \cdot \underline{\phi}_j)^2 + \lambda \|\underline{\phi}_j\|^2$$

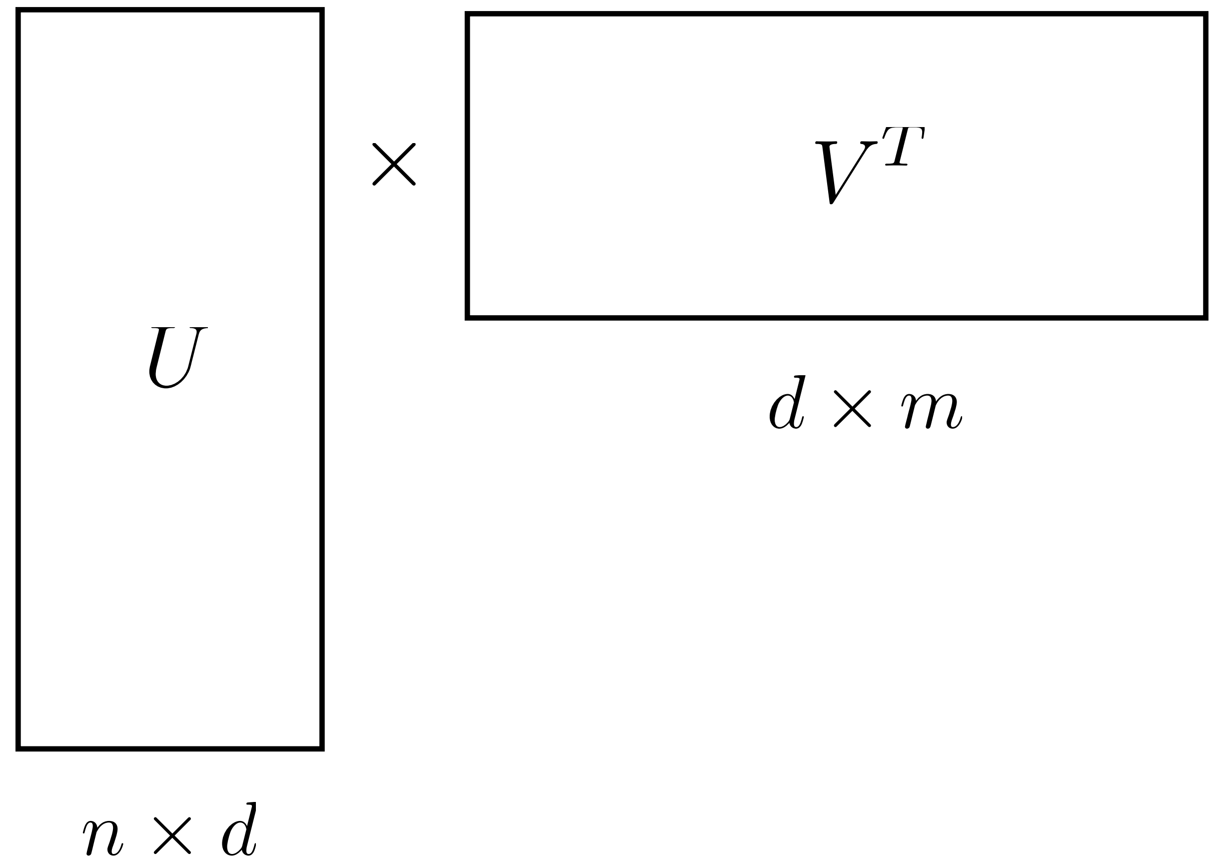
regression problem for each movie with fixed user features

Matrix factorization cont'd

- We can approximate the rating matrix as a product of two lower rank matrices

5	5						5		
		3	5	1	3	4	4		4
	4	2			2				
		5							5
4	5							4	
4							4		
5		4	5	1		4			
	4								
5				4					
5						4			
		5				5		3	

$\approx$



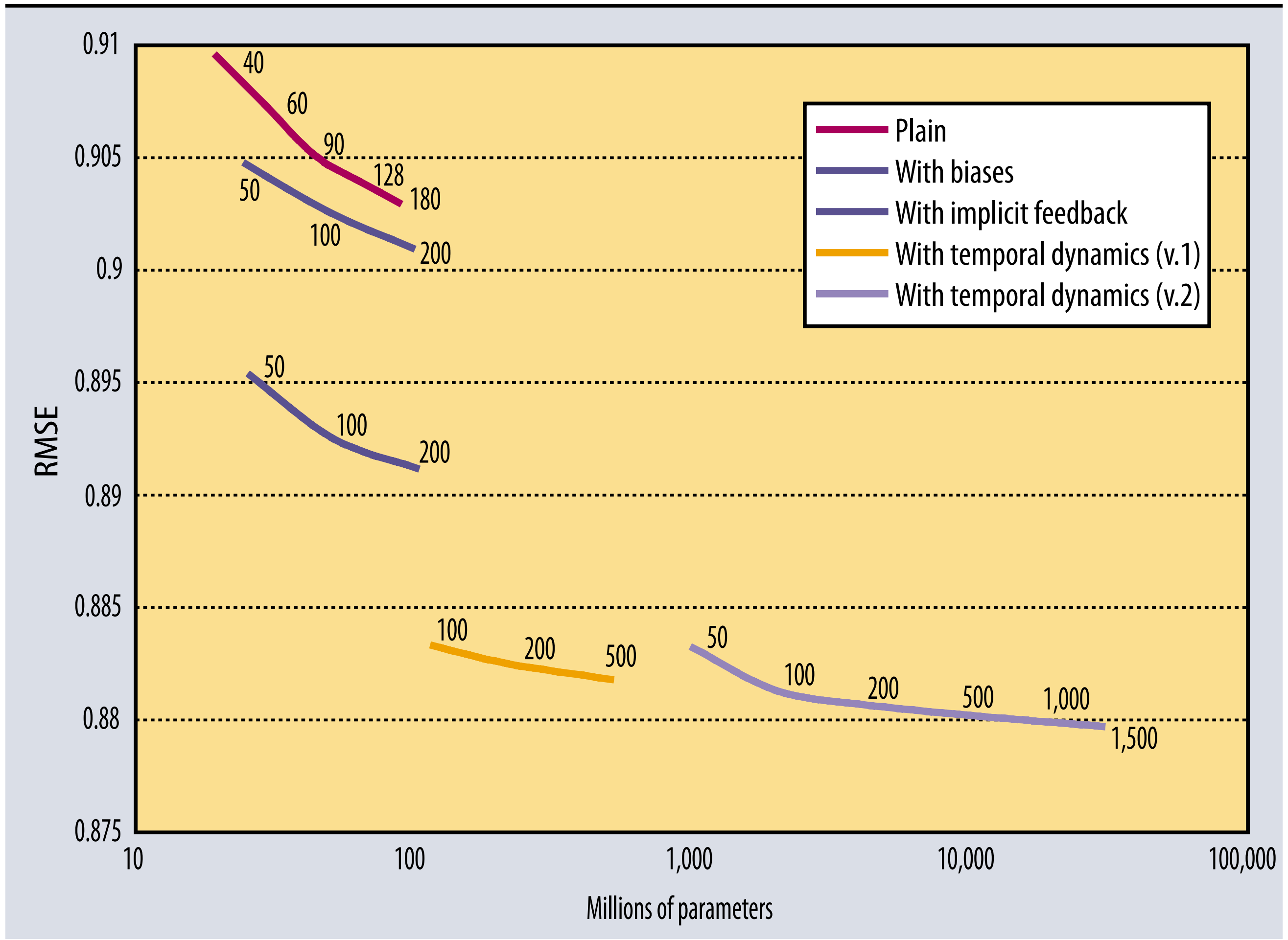
$Y_{ij} \approx [UV^T]_{ij}$

$$\min_{U,V} \sum_{ij \in D} (Y_{ij} - [UV^T]_{ij})^T + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

observed entries

CF and the Netflix Price

- Progress using different matrix factorization methods



(Koren et al., 2009)

- (to win the price, one had to combine hundreds of different methods)

Matrix factorization

- We try to find the best rank d approximation to the rating matrix based on the observed entries

  minimize $$\frac{1}{2}\sum_{ij\in D}(Y_{ij}-[UV^T]_{ij})^2+\frac{\lambda}{2}\|U\|_F^2+\frac{\lambda}{2}\|V\|_F^2$$

  where U is $n \times d$ and V is $m \times d$

  - rank d can be used for complexity control along with the regularization parameter lambda
  - the optimization problem is not jointly convex in U and V. However, it is convex in U if we fix V, and vice versa
  - an alternating minimization algorithm, i.e., iteratively solving user / movie regression problems, may get stuck in a locally optimal solution (initialization is important)
  - algorithms that sequentially add simple rank-1 components at a time are typically better.