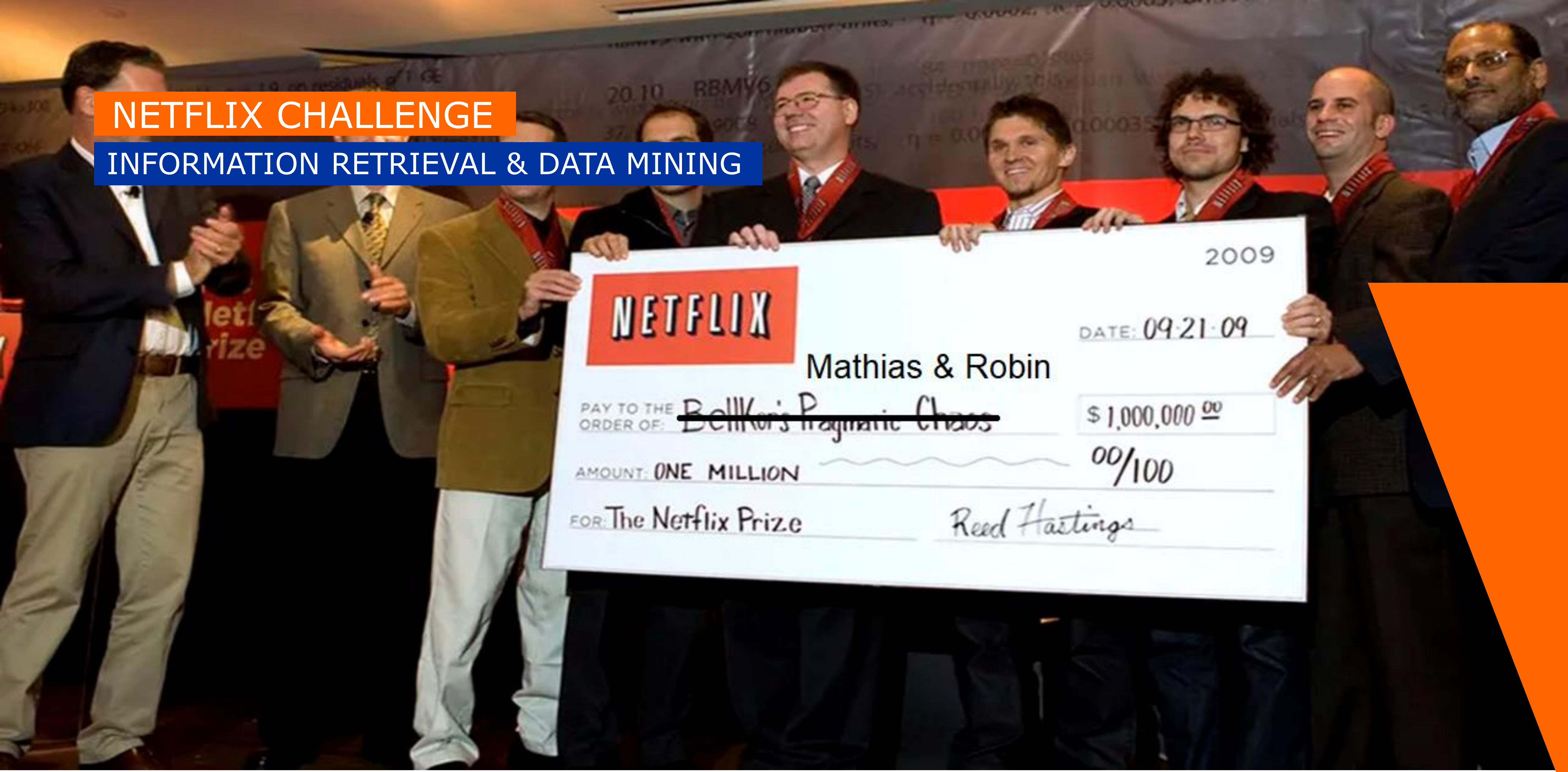
NETFLIX CHALLENGE

INFORMATION RETRIEVAL & DATA MINING



VUB
VRIJE UNIVERSITEIT BRUSSEL

- Robin De Haes
- Mathias Mallinus

THE NETFLIX CHALLENGE DATASET

DATASET OVERVIEW

SPARSITY

	Minimum	Maximum	# Distinct Entries
Movie ID	1	17 770	17 770
User ID	1	2 649 429	480 189
Rating	1	5	100 480 507

$$\Rightarrow \quad Density: \frac{100\,480\,507}{17\,770 \times 480\,189} = 0.012$$

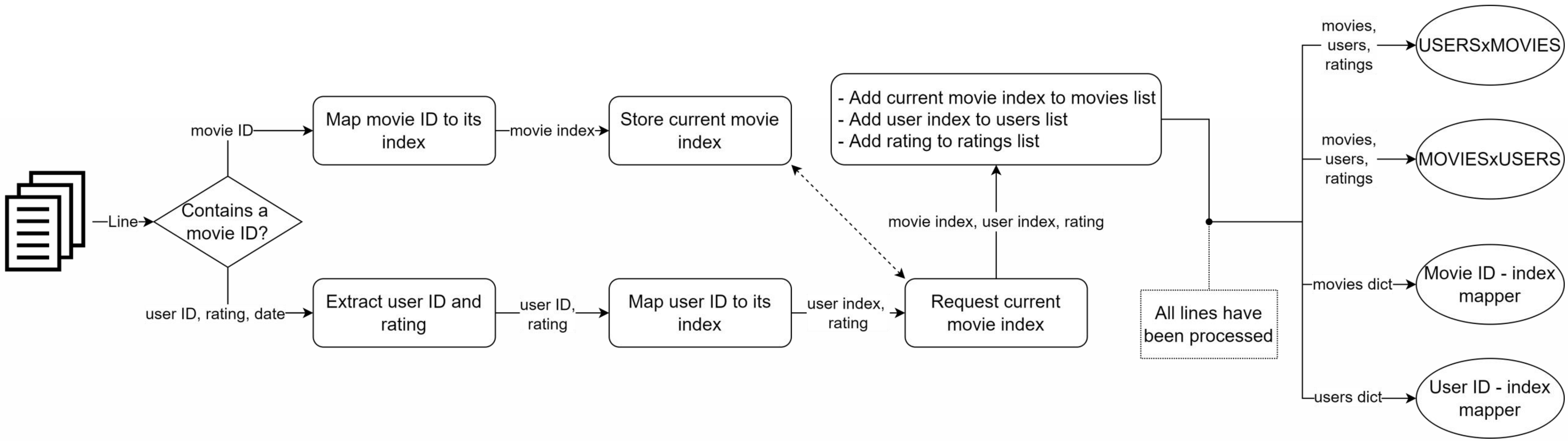
DATASET
FORMAT

4 text files

MovieID1:
CustomerID11,Rating11,Date11
CustomerID12,Rating12,Date12
...
MovieID2:
CustomerID21,Rating21,Date21
CustomerID22,Rating22,Date22

LOADING THE DATASET

PARSING PIPELINE



SPLITTING THE DATASET

PROBE FILE

1 probe file

MovieID1:
CustomerID11
CustomerID12
...
MovieID2:
CustomerID21
CustomerID22

SPLITTING THE DATASET

PROBE MASK

Original Sparse Matrix

	1	**2**	**3**	**…**	**480189**
1	1	5	4		
2		2	1		2
3			1		
…	3				
17769		3			
17770			1		

Index-based Mask

	1	**2**	**3**	**…**	**480189**
1		1			
2		1			1
3					
…					
17769					
17770			1		

VRIJE UNIVERSITEIT BRUSSEL

SPLITTING THE DATASET

PROBE OVERVIEW

	# Movie IDs	# User IDs	# Ratings	% of Full Dataset
Training	17 770	480 189	9 907 112	98.6
Test	16 938	462 858	1 408 395	1.4
Full	17 770	480 189	100 480 507	100

SPLITTING THE DATASET

RANDOM SAMPLING

Original Sparse Matrix

	1	2	3	…	480189
1	1	5	4		
2		2	1		2
3			1		
…	3				
17769		3			
17770			1		

Test Sparse Matrix

	1	2	3	…	480189
1		5			
2		2			2
3					
…					
17769					
17770			1		

DIMSUM

SINGULAR VALUE DECOMPOSITION
DIMSUM APPROACH

$A^T A = V\Sigma^2 V^T$

- Map-reduce approximation for scalability on tall skinny matrices
- Importance sampling for safeguarding singular values

DIMSUMMapper(r_i)

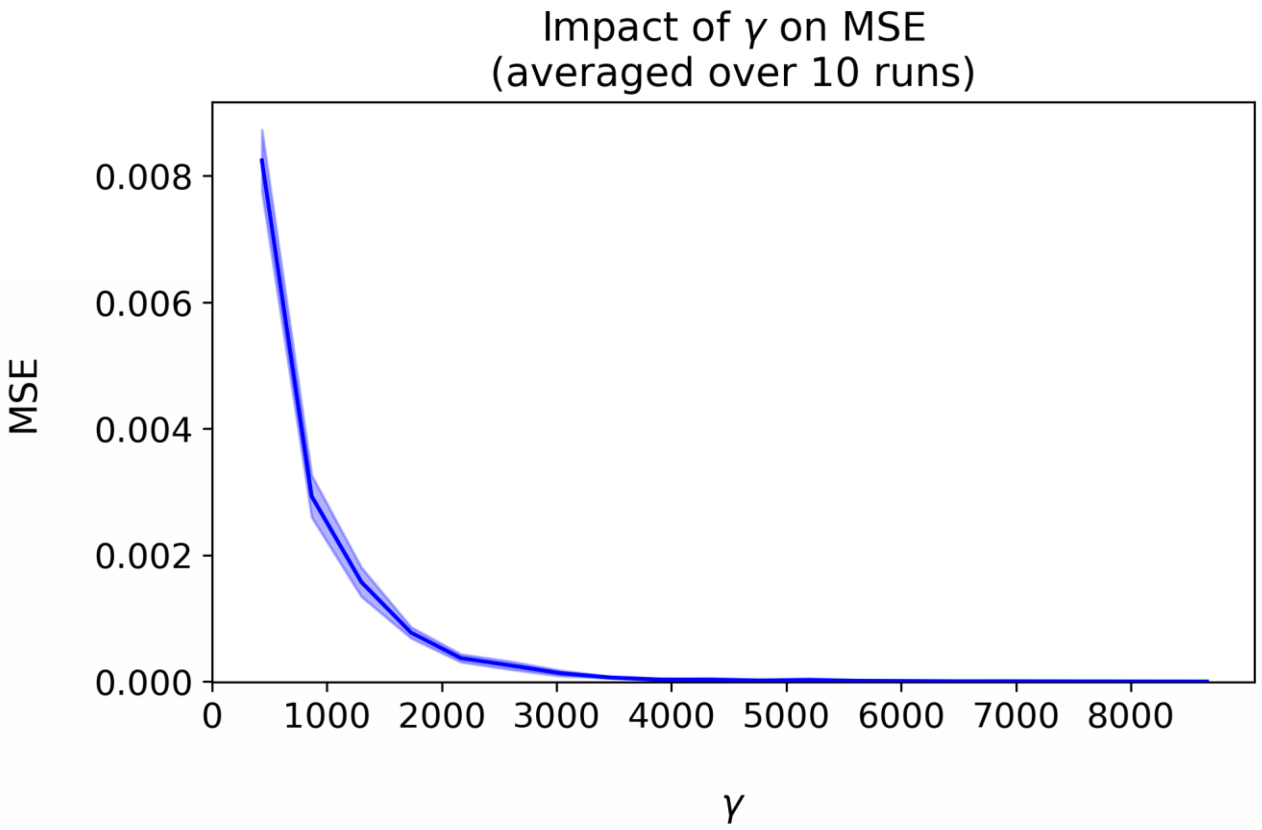
for all pairs (a_{ij}, a_{ik}) in r_i **do**
With probability
$$\min\left(1, \boxed{\gamma \frac{1}{||c_j||||c_k||}}\right)$$
emit $((c_j, c_k) \rightarrow a_{ij}a_{ik})$
end for

DIMSUMReducer($(c_i, c_j), \langle v_1, \ldots, v_R \rangle$)

if $\boxed{\frac{\gamma}{||c_i||||c_j||}} > 1$ **then**
output $b_{ij} \rightarrow \frac{1}{||c_i||||c_j||} \sum_{i=1}^{R} v_i$
else
output $b_{ij} \rightarrow \frac{1}{\gamma} \sum_{i=1}^{R} v_i$
end if

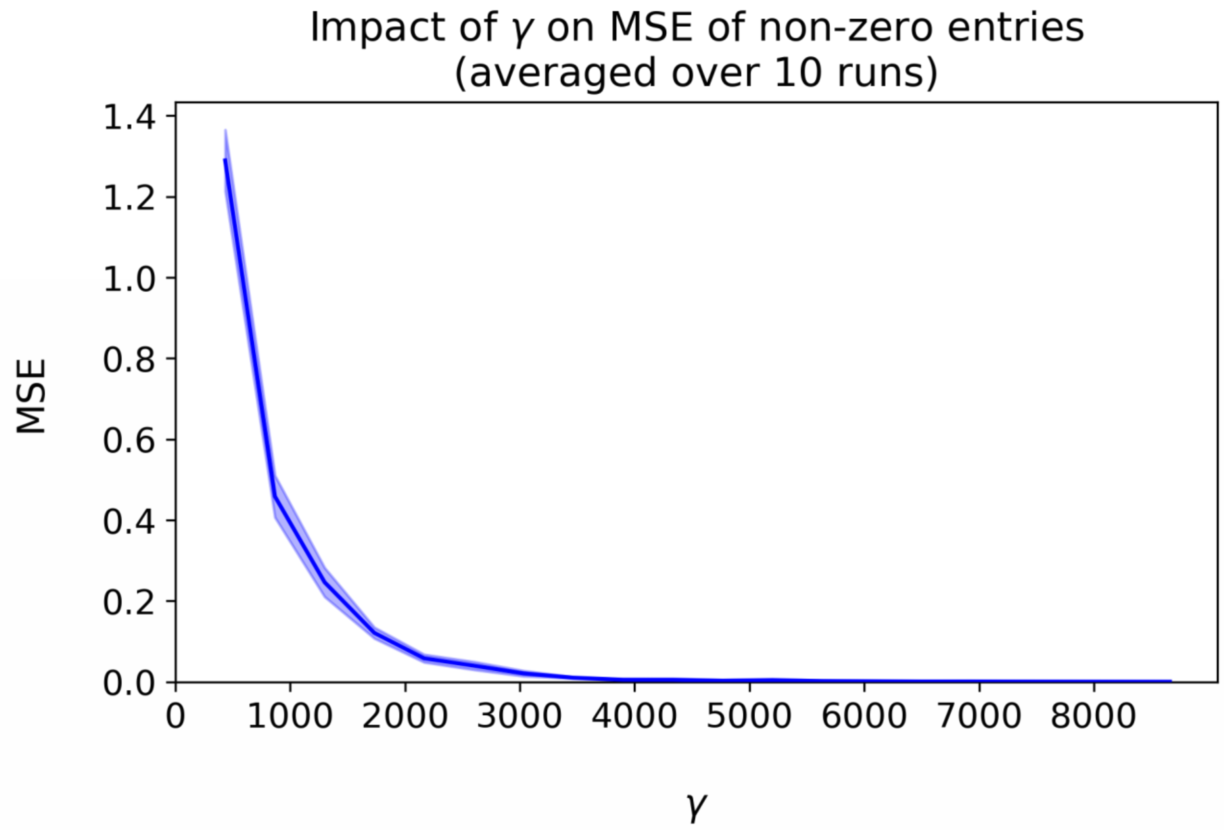
DIMSUM

IMPACT OF γ



DIMSUM

IMPACT OF γ



GRADIENT DESCENT WITH LATENT FACTORS

SINGULAR VALUE DECOMPOSITON

LATENT FACTORS

- $A = U\Sigma V^T = QP^T$
  - $Q = U$: latent movie information
  - $P^T = \Sigma V^T$: latent user information
  - $Q_m P_u^T$: predicted rating for movie m by user u

P^T

	1	…	u	…	480189
1					
2					
3					

Q

	1	2	3
1			
…			
m			
…			
17770			

	1	…	u	…	480189
1	1	5	4		
…		2	1		2
m			1		
…	3				
17770			1		

GRADIENT DESCENT

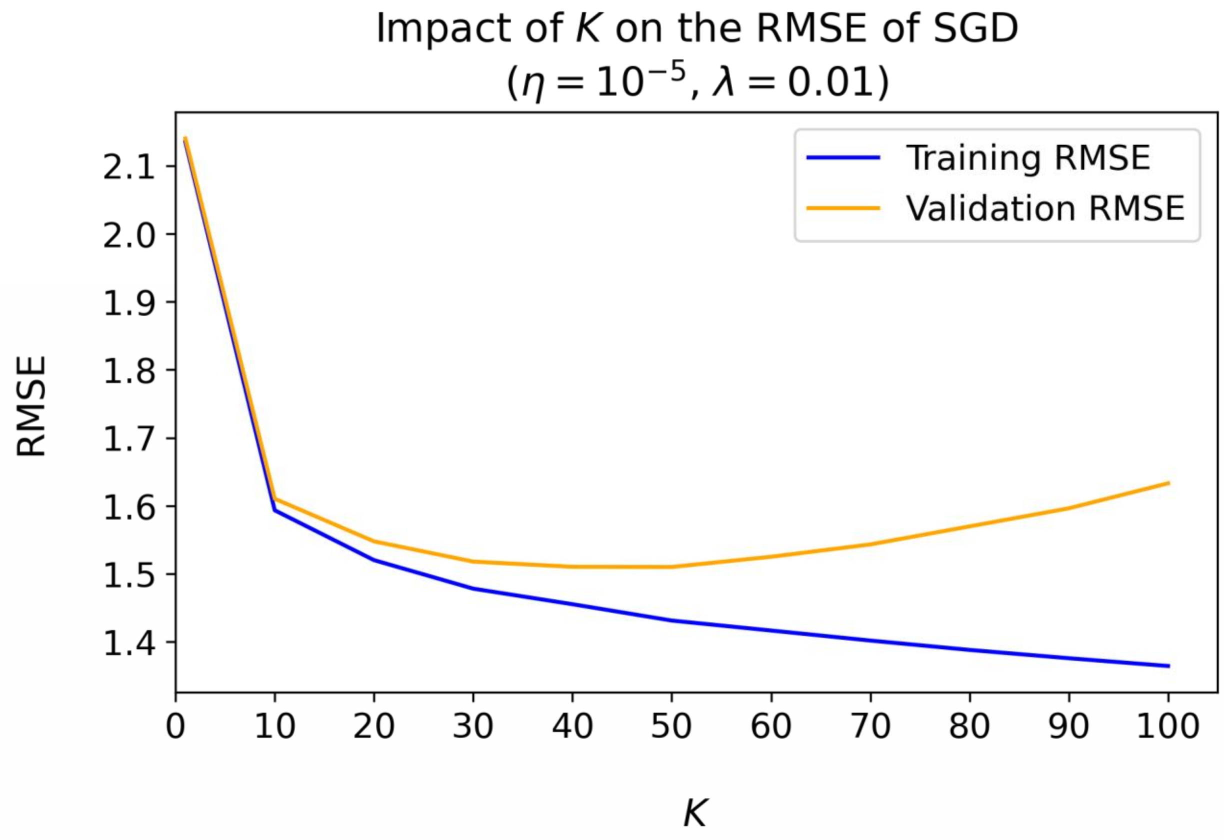
ITERATIVE REFINEMENT

- Objective function: $\min_{p,q} \sum_{(m,u)\in K} \left(r_{mu} - q_m p_u\right)^2 + \lambda \left(\|q_m\|^2 + \|p_u\|^2\right)$
- Stochastic gradient descent *(with random shuffling)*
- Batch gradient descent
- Hyperparameters
  - Regularization parameter λ
  - Learning rate η
  - # epochs (2-5)
  - # latent factors K

VRIJE UNIVERSITEIT BRUSSEL

STOCHASTIC GRADIENT DESCENT

IMPACT OF K

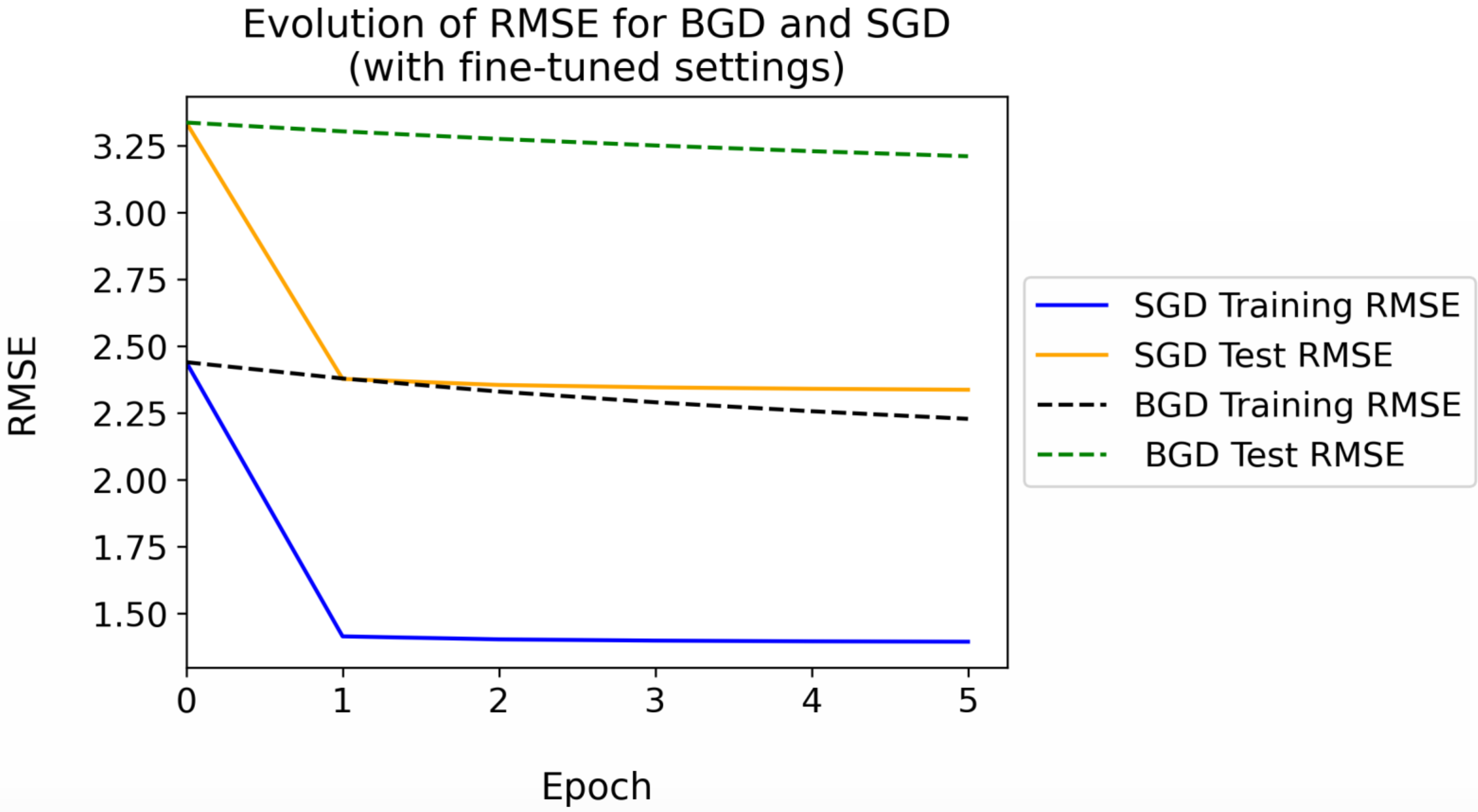


GRADIENT DESCENT

COMPARISON OF SGD AND BGD

Hyperparameters

- K = 50
- $\lambda = 0.01$
- *SGD*: $\eta = 10^{-5}$
- *BGD*: $\eta = 0.1/|\text{dataset}|$



REFERENCES
BIBLIOGRAPHY

[1] *BellKor's Pragmatic Chaos Wins $1 Million Netflix Prize by Mere Minutes*. (2009). Retrieved May 31, 2022, from https://www.wired.com/2009/09/bellkors-pragmatic-chaos-wins-1-million-netflix-prize

[2] Zadeh, R. B., & Carlsson, G. (2013). Dimension independent matrix square using mapreduce. *arXiv preprint arXiv:1304.1467.*

[3] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

QUESTIONS?