Leveraging Social Context for Modeling Topic Evolution

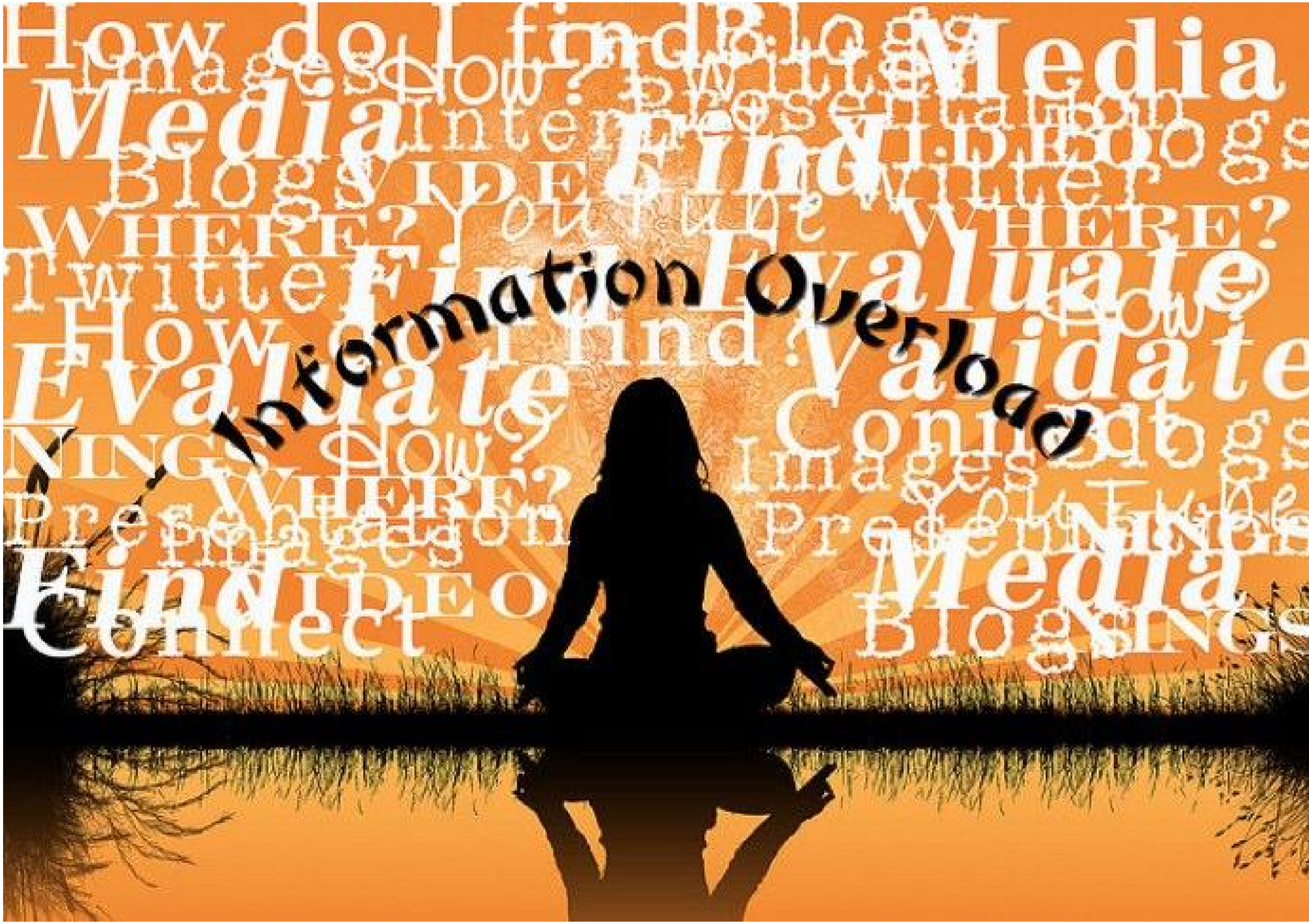
Janani Kalyanam, Amin Mantrach, Diego Saez-Trumper,
Hossein Vahabi, Gert Lanckriet

UC San Diego

YAHOO! LABS

Introduction

Introduction



Introduction

Topic Modeling

- NMF-based
- Bayesian (like LDA)

Bird flu outbreak; everything you need to know goo.gl/F1dnfk #birdflu

U.S to review protocols following birdflu outbreak goo.gl/X88iSe #birdflu

U.S poultry devastated by birdflu outbreak goo.gl/1gX8FC #birdflu

Bird flu outbreak; everything you need to know goo.gl/F1dnfk #birdflu

U.S to review protocols following birdflu outbreak goo.gl/X88iSe #birdflu

U.S poultry devastated by birdflu outbreak goo.gl/1gX8FC #birdflu

Bird flu outbreak; everything you need to know goo.gl/F1dnfk #birdflu

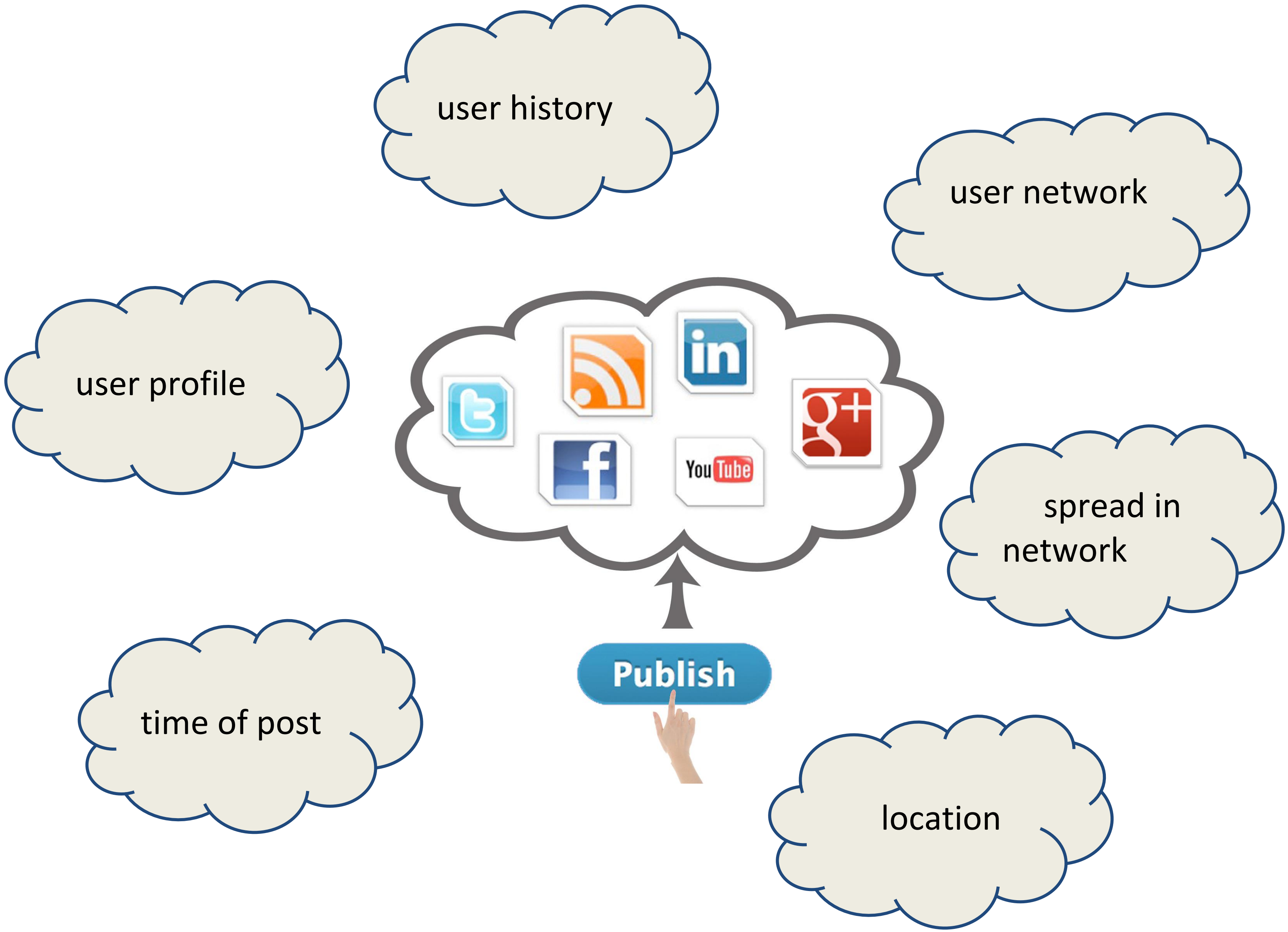
U.S to review protocols following birdflu outbreak goo.gl/X88iSe #birdflu

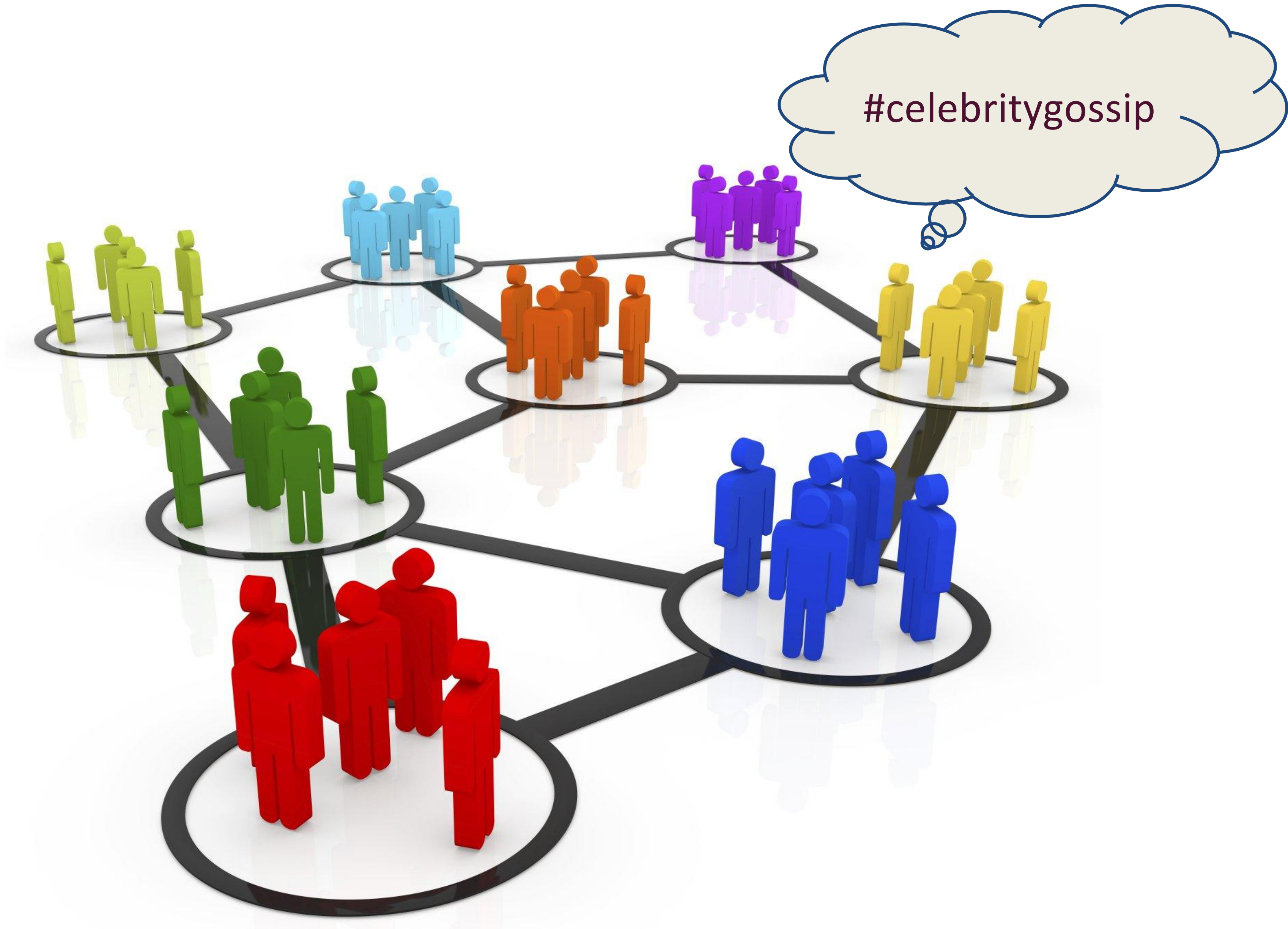
U.S poultry devastated by birdflu outbreak goo.gl/1gX8FC #birdflu

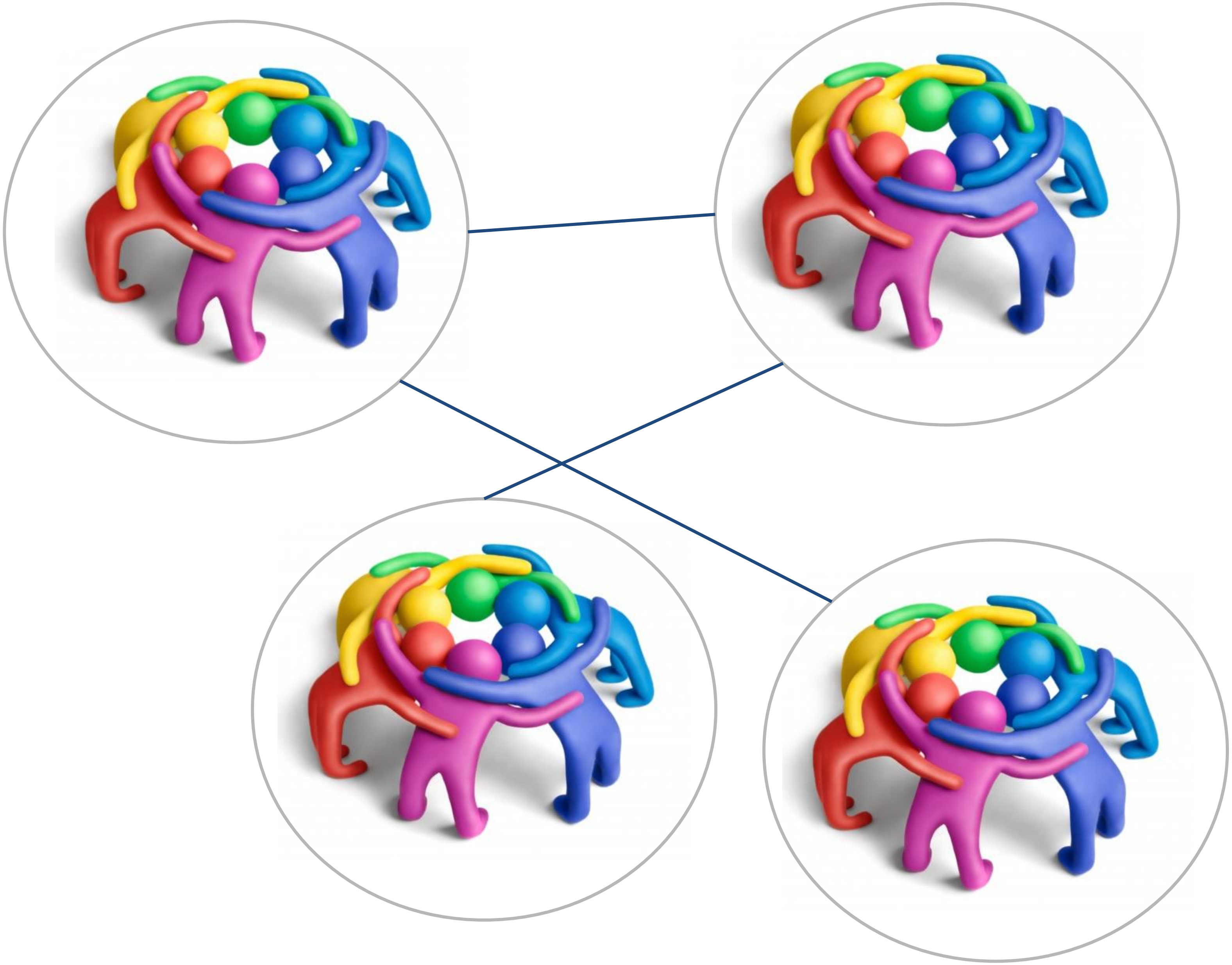
Kim Kardashian: pregnant again! goo.gl/Ir1knd #celebritygossip

Selina Gomez and Justin Bieber: “just friends” goo.gl/M9dlhj #celebritygossip

Lindsay Lohan messed up contract with Oprah goo.gl/Ir1knd #celebritygossip


user history
user network
user profile
spread in network
time of post
Publish
location


#celebritygossip



Topic Modeling

- NMF-based
- Bayesian (like LDA)

Generally focus on content

What’s needed

in addition to textual content, use context and meta data that surrounds the text to discover the latent topics

Our goal

Does user interactions, and temporal evolution help detect better topics?

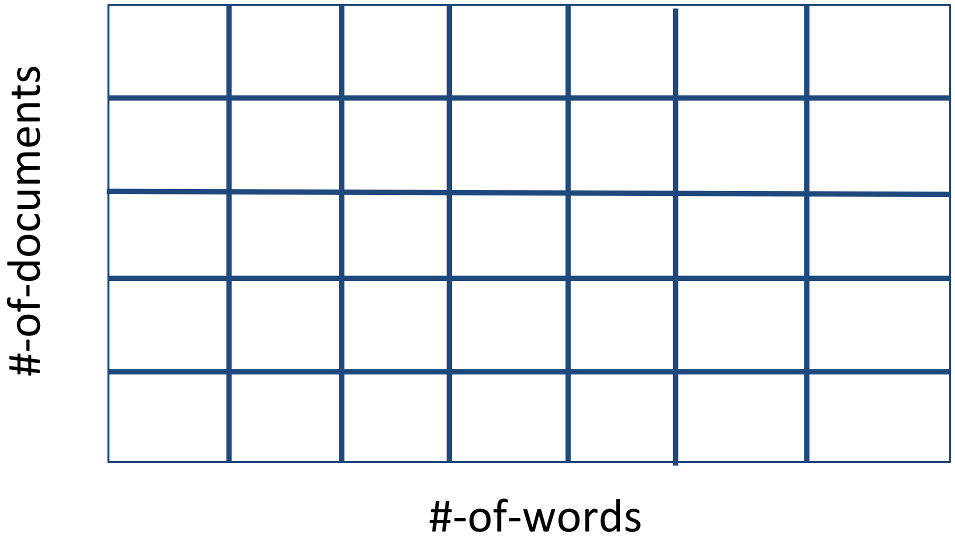
How do we do this?

How do we approach this?

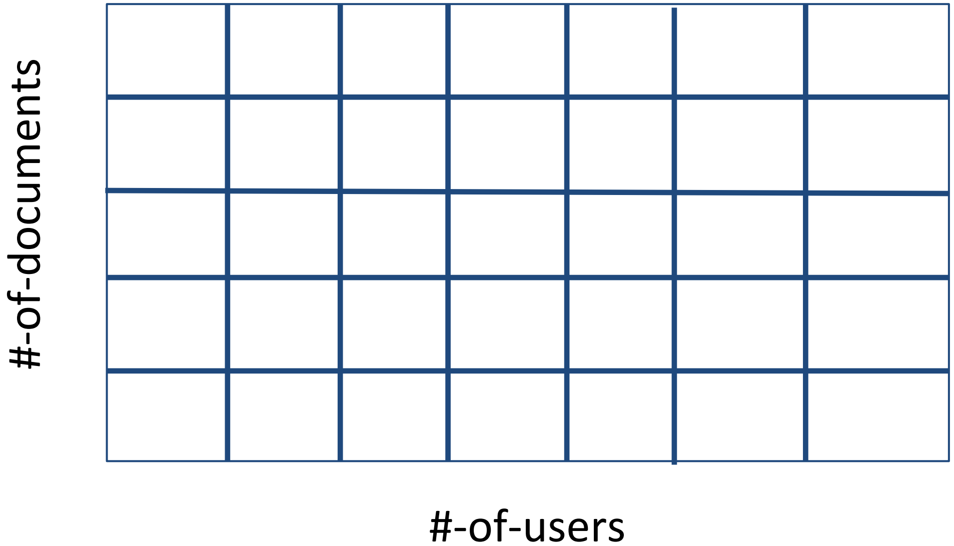
- Non Negative Matrix Factorization based method.
- Start with the classical NMF objective..
- build on it..

Notation

X^t

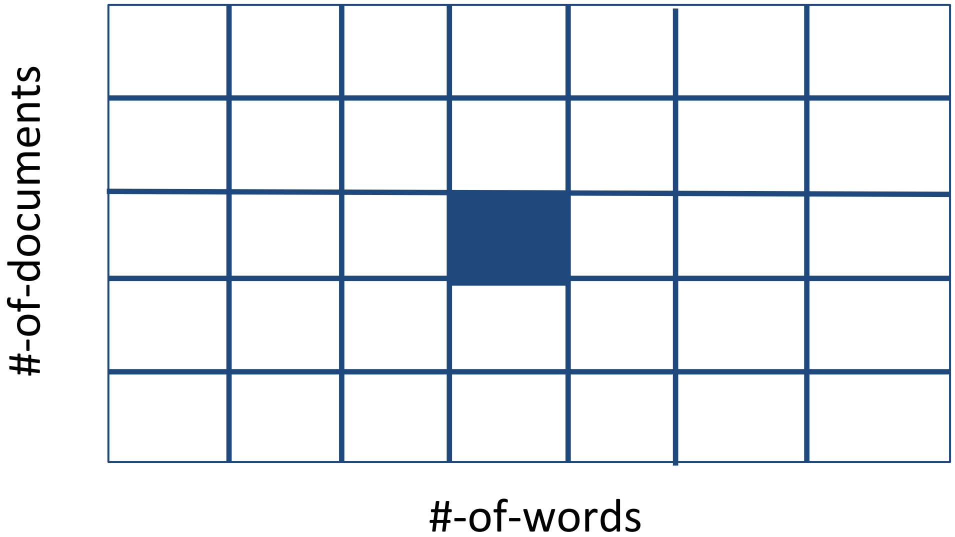


U^t

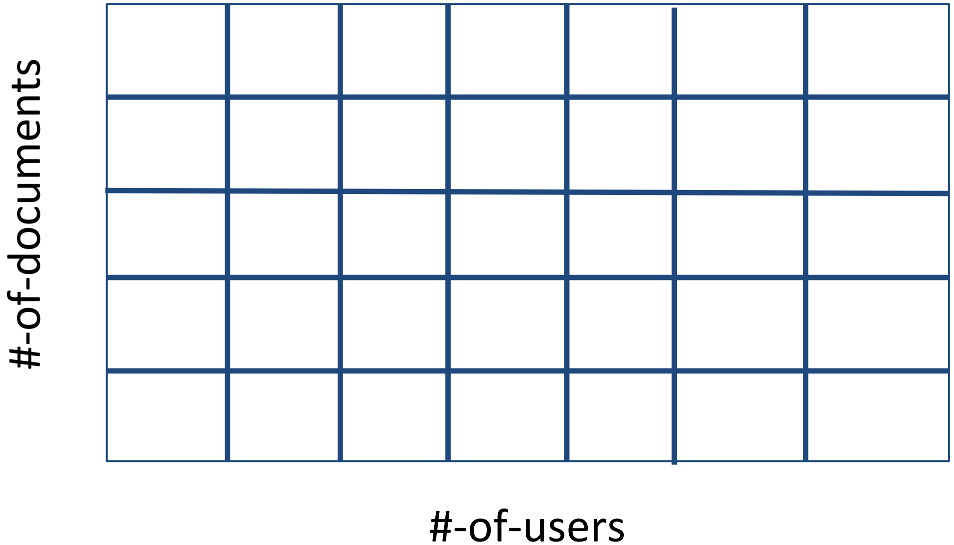


Notation

X^t

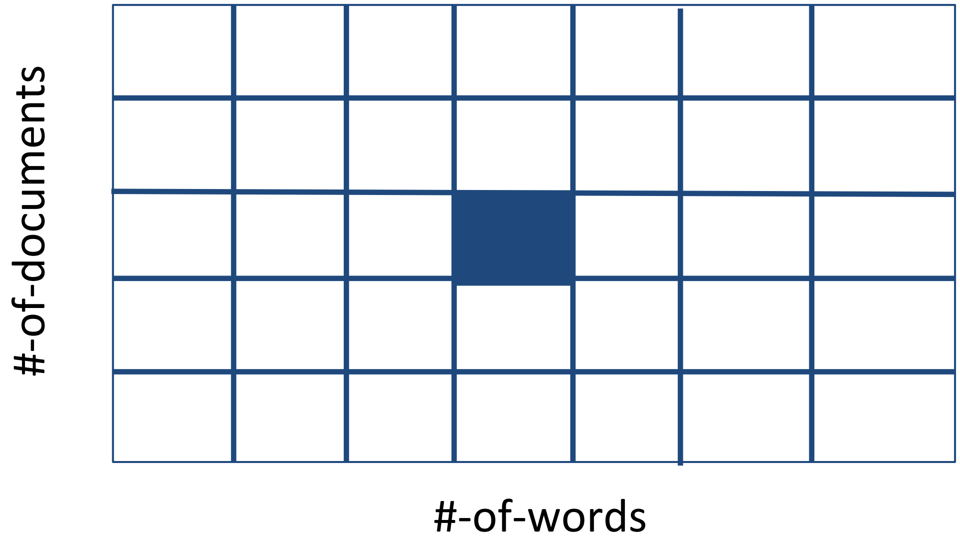


U^t

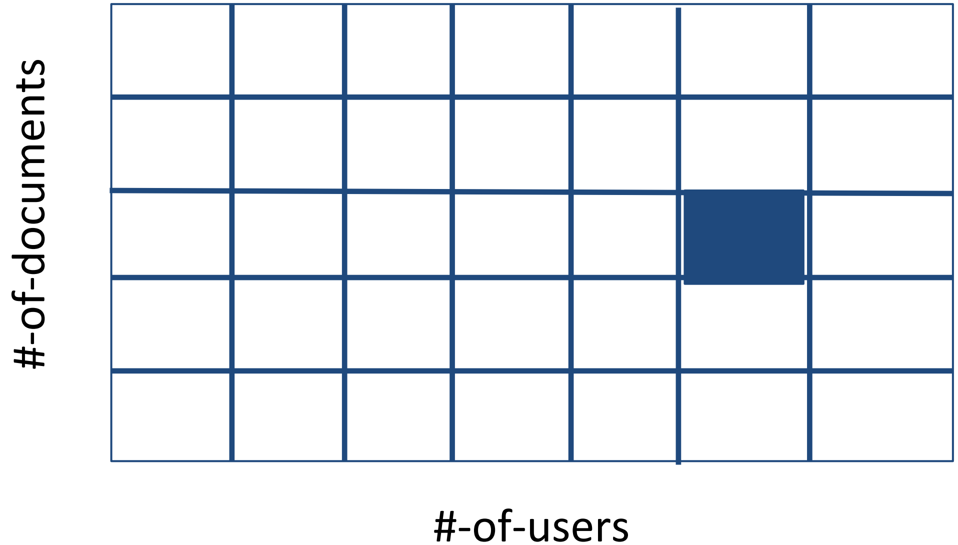


Notation

X^t

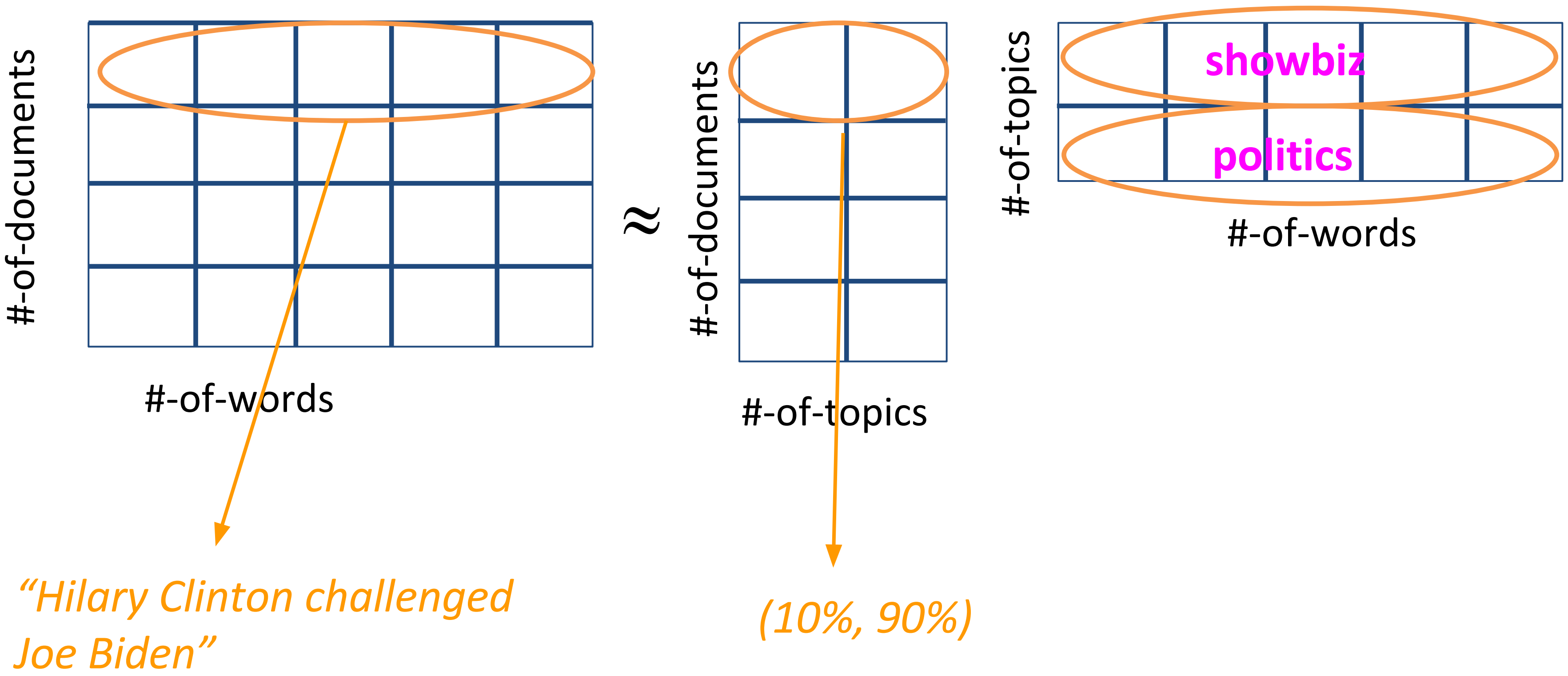


U^t



How do we approach this?

$$X^t \approx W^t H^t$$



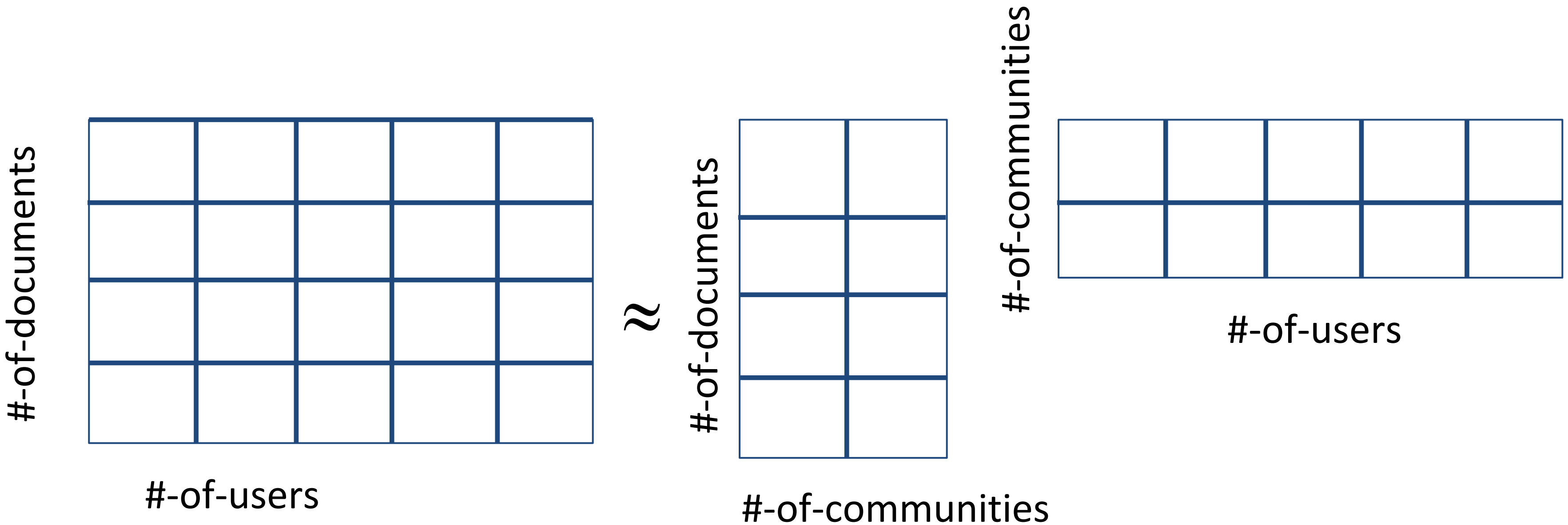
Ingredients of Objective Function

$$\|X^t - W^t H^t\|^2$$

Variables are W^t H^t

How do we approach this?

$$U^t \approx W^t G^t$$



Ingredients of Objective Function

$$\|X^t - W^t H^t\|^2 \qquad + \|U^t - W^t G^t\|^2$$

Variables are W^t H^t G^t

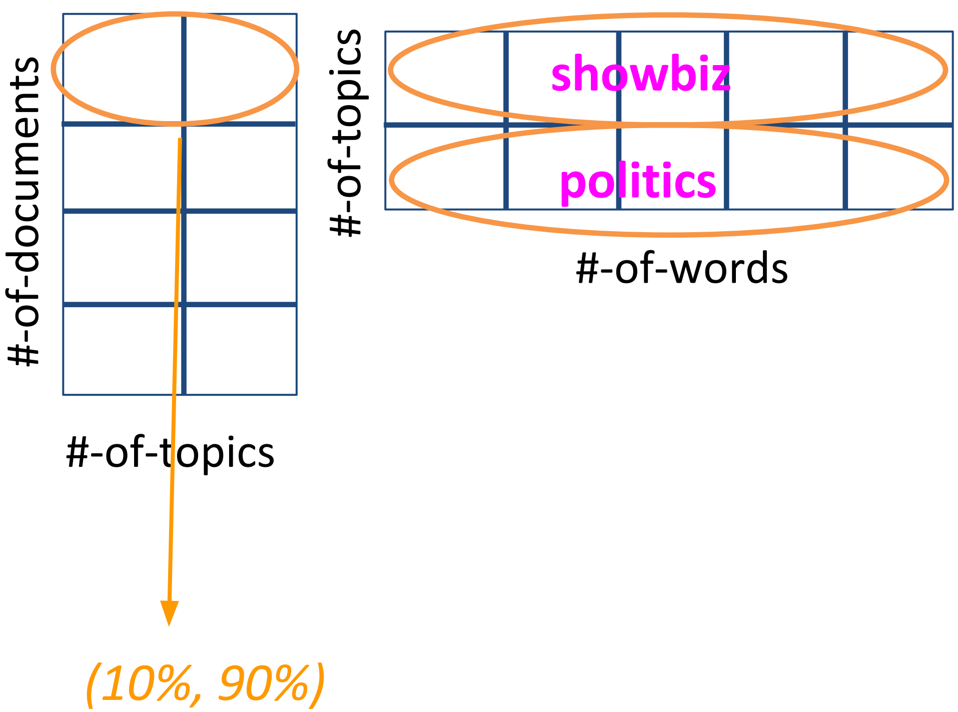
Key Assumption

$$X^t \approx W^t H^t \qquad U^t \approx W^t G^t$$

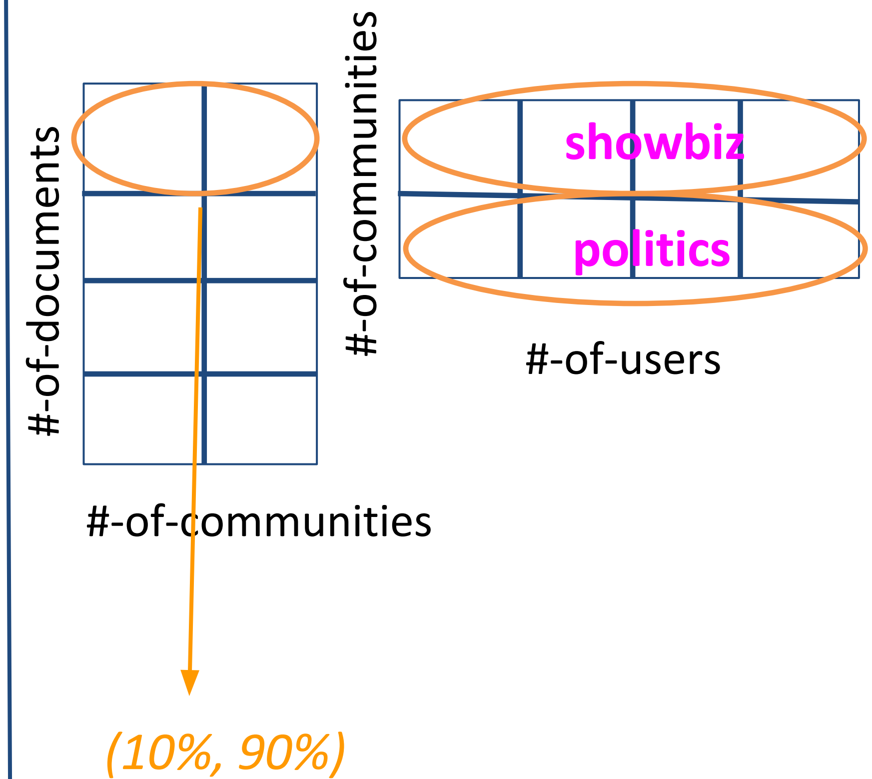
The W^t matrix is common to both decompositions.

Key Assumption

$$X^t \approx W^t H^t$$



$$U^t \approx W^t G^t$$



Evolution Over Time

$$X^t \approx W^t M_T^t H^{t-1}$$

Evolution Over Time

$$X^t \approx W^t \underbrace{M_T^t H^{t-1}}_{H^t}$$

M_T^t Evolution matrix

Ingredients of Objective Function

$$\underbrace{\|X^t - W^t H^t\|^2 + \|X^t - W^t M_T^t H^{t-1}\|^2}_{L_T} + \underbrace{\|U^t - W^t G^t\|^2 + \|U^t - W^t M_C^t G^{t-1}\|^2}_{L_C}$$

content part

community part

Variables are W^t H^t G^t M_T^t M_C^t

Loss Function

$$L = \mu L_T + (1-\mu) L_C + R$$

μ importance parameter

R regularization

How to evaluate?

How do we do this?

Split into three categories..

How do we do this?

Split into three categories..

- “good topics”, *CONTENT STABLE TOPICS*

How do we do this?

Split into three categories..

- “good topics”, *CONTENT STABLE TOPICS*
- “difficult topics”, *COMMUNITY STABLE TOPICS*

How do we do this?

Split into three categories..

- “good topics”, *CONTENT STABLE TOPICS*
- “difficult topics”, *COMMUNITY STABLE TOPICS*
- a mixture of the above two or *MIXED STABLE TOPICS*

How do we do this?

In each category, evaluate how much does adding the contextual information and temporal information really help..

Data

- Content
  - News articles from CNN, BBC, Al jazeera
- Community
  - All tweets which linked to the articles
    - Collect username publishing the tweet
    - Collect the hashtag in the tweet

Baseline Approaches

- LTECS: Learning Topic Evolution from Content and Social Media activity

- Link-PLSA-LDA (Nallapati et. al. KDD 2008): lacks temporal element

Baseline Approaches

- Online LDA (AlSumait et. al. ICDM 2008): lacks community element
- Joint Past Present Decomposition (Vaca Ruiz et. Al. WWW 2014): lacks of community
- CMF (Recsys 2014): lacks of temporal element

Results (Community Stable)

LTECS

	K = 5	K = 10	K = 15	K = 20
NDCG	0.4081	0.4800	0.5029	0.5129
MAP	0.2653	0.3637	0.4007	0.4173
	μ = 0.01	μ = 0.5	μ = 0.5	μ = 0.5

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.3699	0.4496	0.4608	0.4138
MAP	0.2191	0.3596	0.3462	0.3420

Baseline Approach; NO TEMPORAL MODELING

	K = 5	K = 10	K = 15	K = 20
NDCG	0.3454	0.4338	0.4771	0.4827
MAP	0.2044	0.3190	0.3757	0.3665

Results (Content Stable)

LTECS

	K = 5	K = 10	K = 15	K = 20
NDCG	0.6888	0.6055	0.6317	0.6623
MAP	0.5655	0.4784	0.5115	0.5559
	$\mu = 1$	$\mu = 1$	$\mu = 0.75$	$\mu = 0.75$

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.6888	0.6055	0.4885	0.6504
MAP	0.5655	0.4784	0.3089	0.5411

Baseline Approach; NO TEMPORAL MODELING

	K = 5	K = 10	K = 15	K = 20
NDCG	0.5846	0.4919	0.4455	0.4327
MAP	0.4423	0.3207	0.2556	0.2557

Results (Mixed Stable)

LTECS

	K = 5	K = 10	K = 15	K = 20
NDCG	0.9005	0.8868	0.9249	0.9089
MAP	0.7783	0.7965	0.8964	0.8845
	$\mu = 0.25$	$\mu = 0.75$	$\mu = 0.25$	$\mu = 0.25$

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.8771	0.8762	0.4251	0.4580
MAP	0.7762	0.7783	0.3232	0.3644

Baseline Approach; NO TEMPORAL MODELING

	K = 5	K = 10	K = 15	K = 20
NDCG	0.6712	0.8768	0.8905	0.8765
MAP	0.5329	0.8223	0.8499	0.8337

Conclusion

- Using community side information helps with “noisy” topics.

Thank You!