Modeling Temporal Dynamics & Geographical Language Variations in Social Streams

Google

Nov. 14, 2012

Liangjie Hong, Ph.D. Candidate
Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA

Temporal Dynamics & Geographical Language Variations

- Motivation
- Modeling Social Streams
- Future work

Temporal Dynamics & Geographical Language Variations

- Motivation
- Modeling Social Streams
  - Modeling Popular Messages [WWW 2011]
  - Modeling Personal Decision & Content [WSDM 2013]
  - Empirical Topic Modeling Study [SOMA 2010]
  - Temporal Dynamics [KDD 2011]
  - Geographical Language Variations [WWW 2012]
- Future work

Temporal Dynamics & Geographical Language Variations

- Motivation
- Modeling Social Streams
  - Modeling Popular Messages [WWW 2011]
  - Modeling Personal Decision & Content [WSDM 2013]
  - Empirical Topic Modeling Study [SOMA 2010]
  - **Temporal Dynamics [KDD 2011]**
  - **Geographical Language Variations [WWW 2012]**
- Future work

Temporal Dynamics & Geographical Language Variations
Motivation

The Blossom of Social Media

Temporal Dynamics & Geographical Language Variations
Motivation

Social Streams

Temporal Dynamics & Geographical Language Variations
Motivation

Social Streams

Temporal Dynamics & Geographical Language Variations
Motivation

Social Streams

Temporal Dynamics & Geographical Language Variations
Motivation

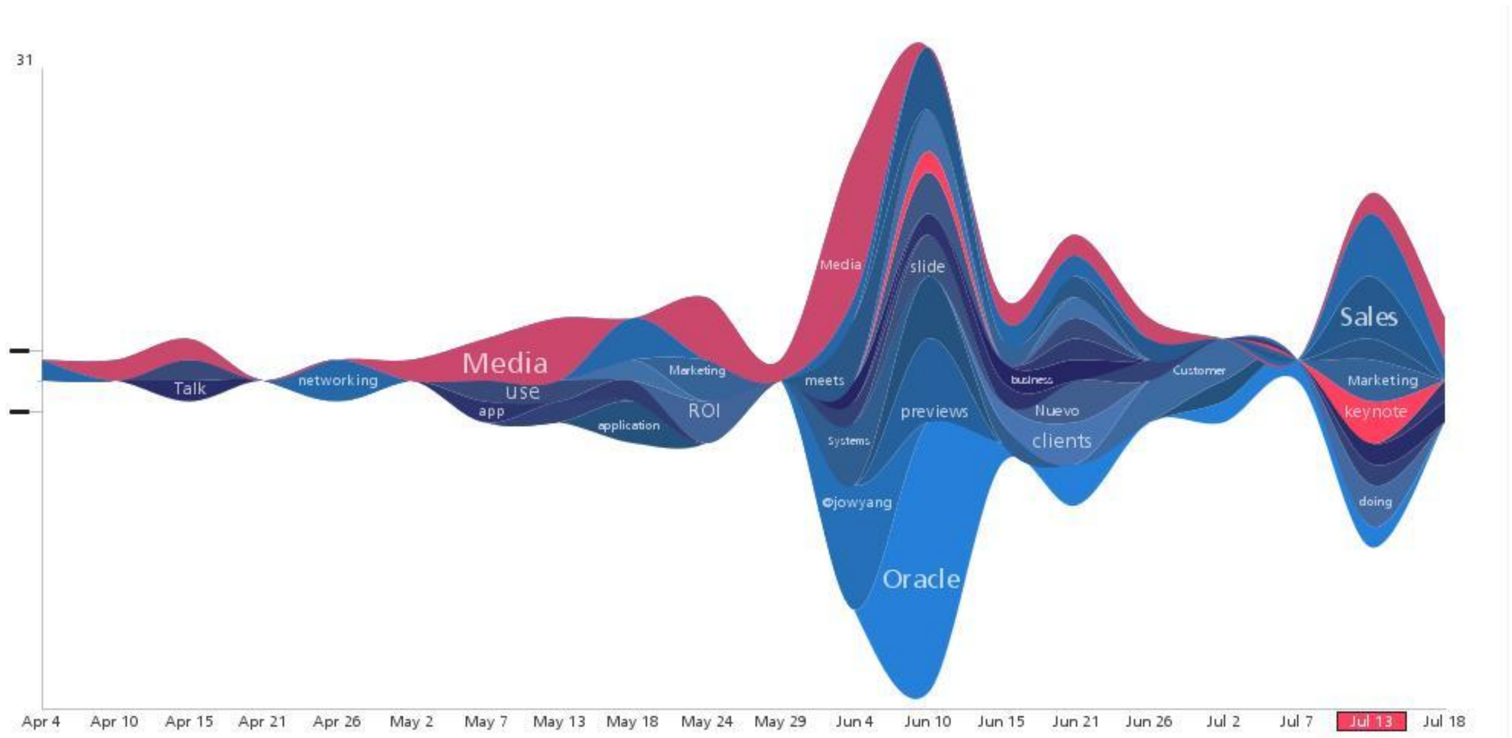
Social Streams

Temporal Dynamics & Geographical Language Variations Motivation

Challenges

Temporal Dynamics & Geographical Language Variations
Motivation

Challenges: Temporal Dynamics



Challenges: Multiple Sources

Temporal Dynamics & Geographical Language Variations
Motivation

Challenges: Meta-data

Temporal Dynamics & Geographical Language Variations
Motivation

Challenges: Meta-data

Temporal Dynamics & Geographical Language Variations
Motivation

Challenges

- Temporal dynamics
- Multiple sources
- Geographical locations

Temporal Dynamics & Geographical Language Variations
Motivation

Technical Challenges

- Multi-facet data
- Large scale
- Incorporate other research advances

Temporal Dynamics + Multiple Sources

Geographical Language Variations

Interesting Questions

- Are here any common topics among multiple media sources?

Modeling Temporal Dynamics
Motivation

Interesting Questions

- Are here any common topics among multiple media sources?
- How can we find them, automatically?

Modeling Temporal Dynamics
Motivation

Interesting Questions

- Are here any common topics among multiple media sources?
- How can we find them, automatically?
- Are there transferred from one source to another?

Modeling Temporal Dynamics
Motivation

Interesting Questions

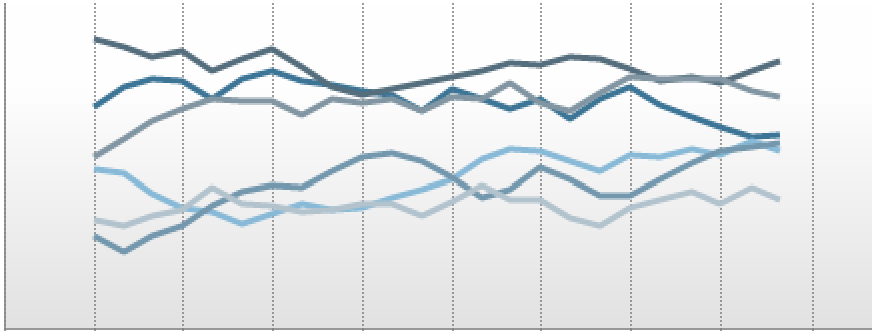
- Are here any common topics among multiple media sources?
- How can we find them, automatically?
- Are there transferred from one source to another?

[Zhao et al., ECIR 2011]

Modeling Temporal Dynamics
Motivation

Applications

Data Visualization



Trend Prediction

Modeling Temporal Dynamics
Motivation

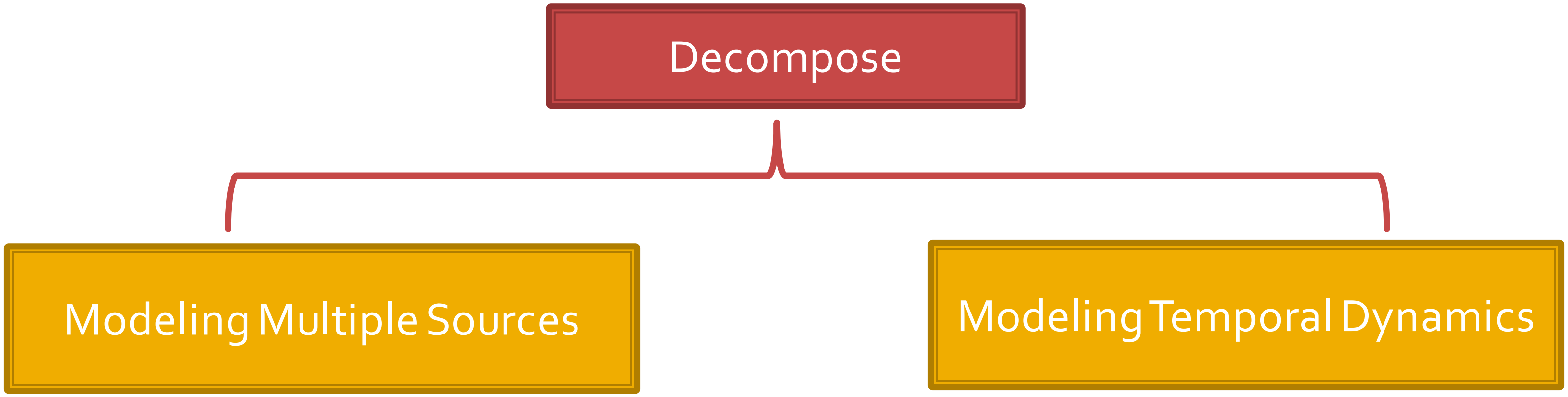
Goal

- identify common/local topics from multiple streams
- characterize their temporal dynamics

- *principled way*

Modeling Temporal Dynamics
Our Proposed Model

Our Approach



Our Approach

Decompose

Modeling Multiple Sources

Modeling Temporal Dynamics

Introduce common topics & local topics

Modeling Temporal Dynamics
Our Proposed Model

Our Approach

Decompose

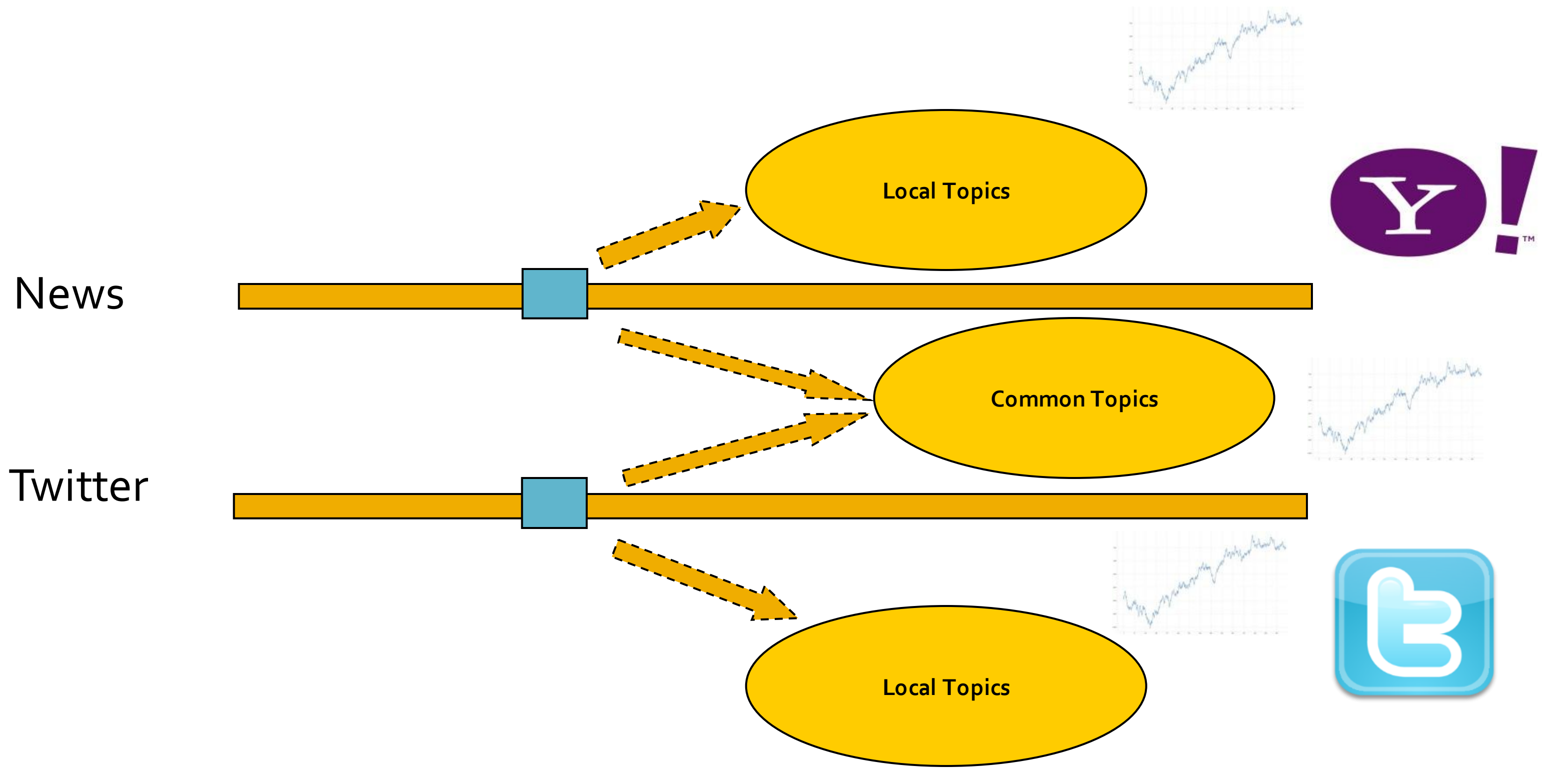
Modeling Multiple Sources

Introduce common topics & local topics

Modeling Temporal Dynamics

Introduce temporal dependent priors

Modeling Temporal Dynamics
Basic Setting



Modeling Temporal Dynamics Handling Multiple Sources

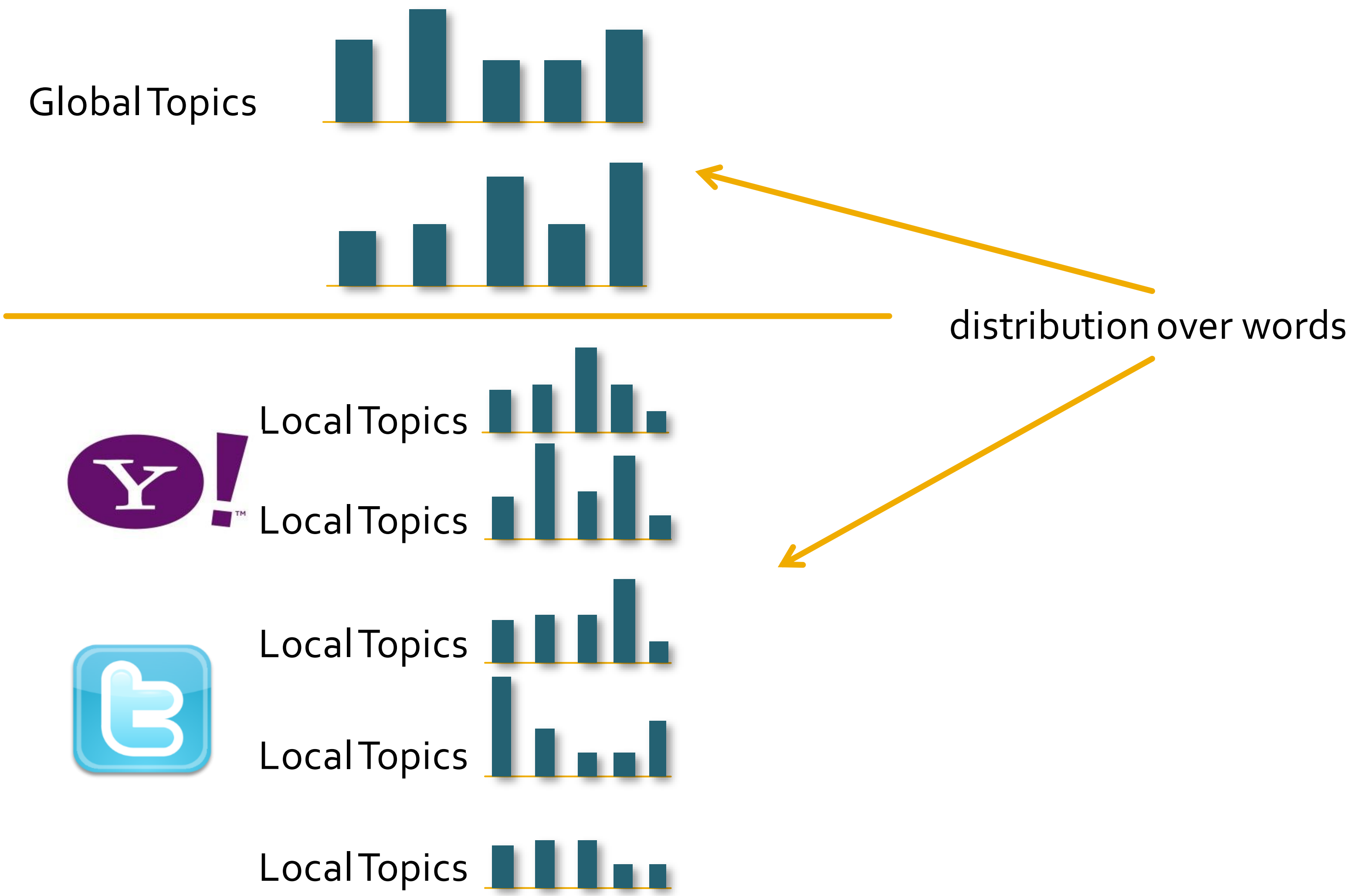
- Basic Intuitions
  - Some topics are shared.

  Tsunami, Super bowl, NBA…
  - Some topics are specific to a certain stream.

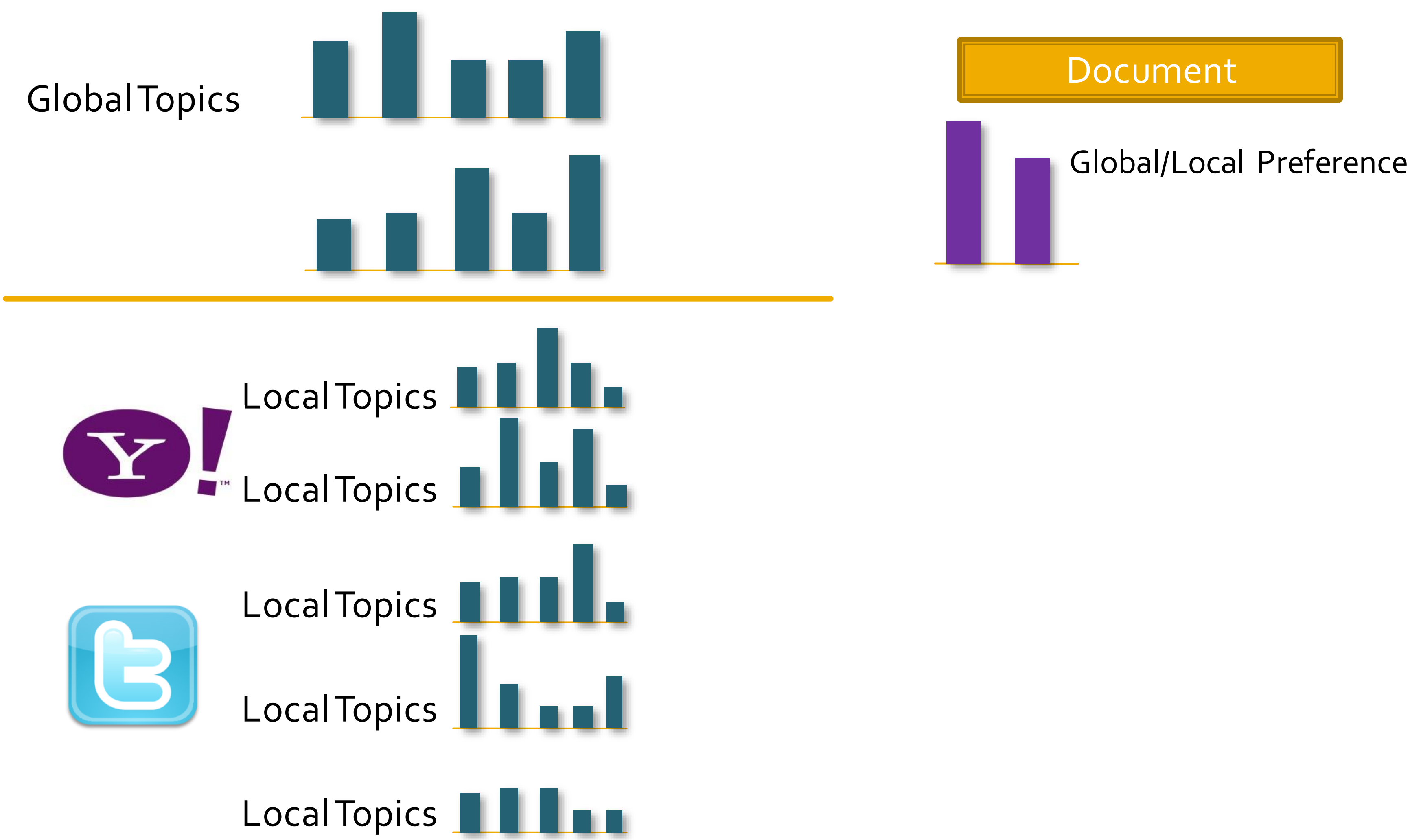
  Local news, Personal opinions…

  - Each stream is a mixture of them.

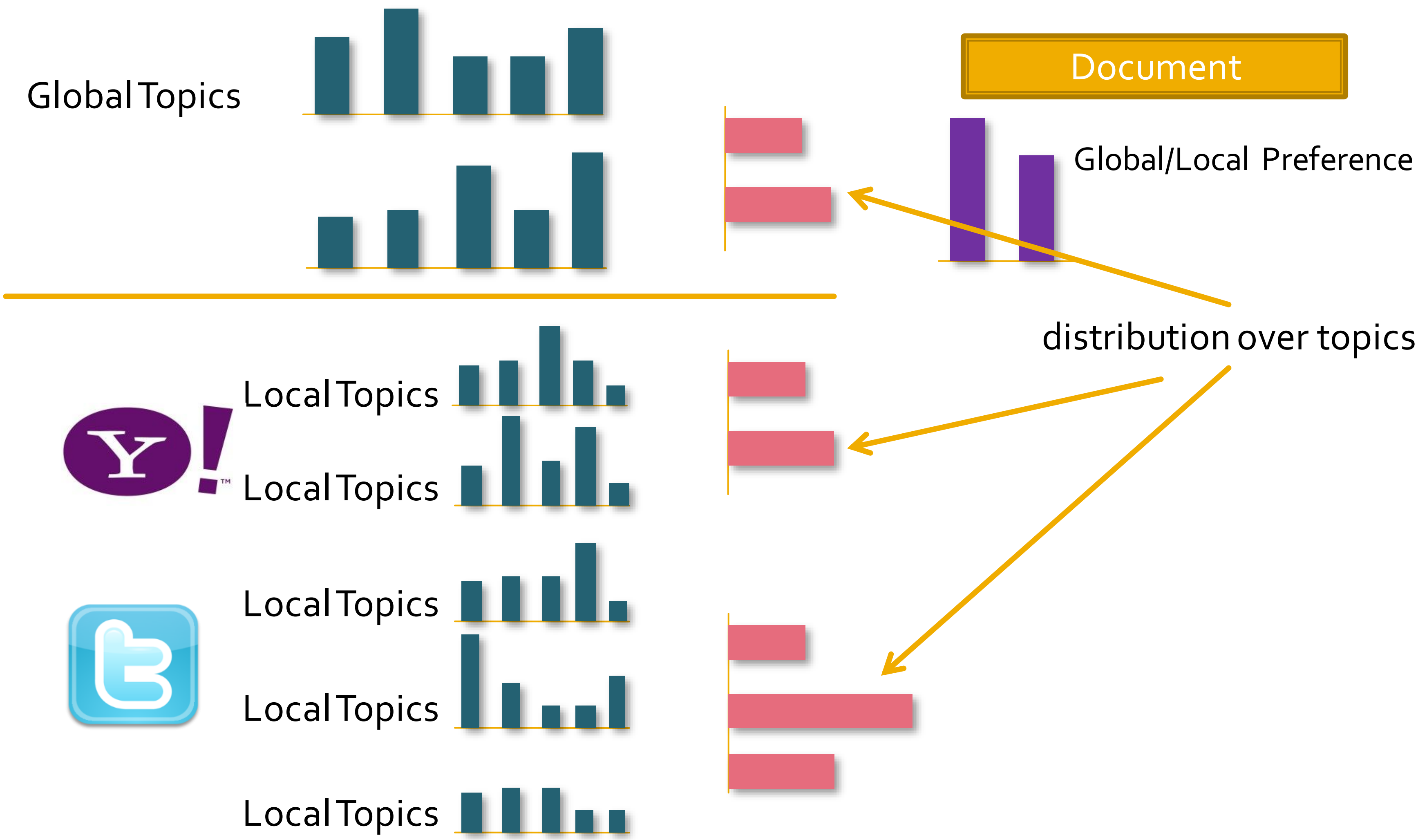
Modeling Temporal Dynamics
Handling Multiple Sources



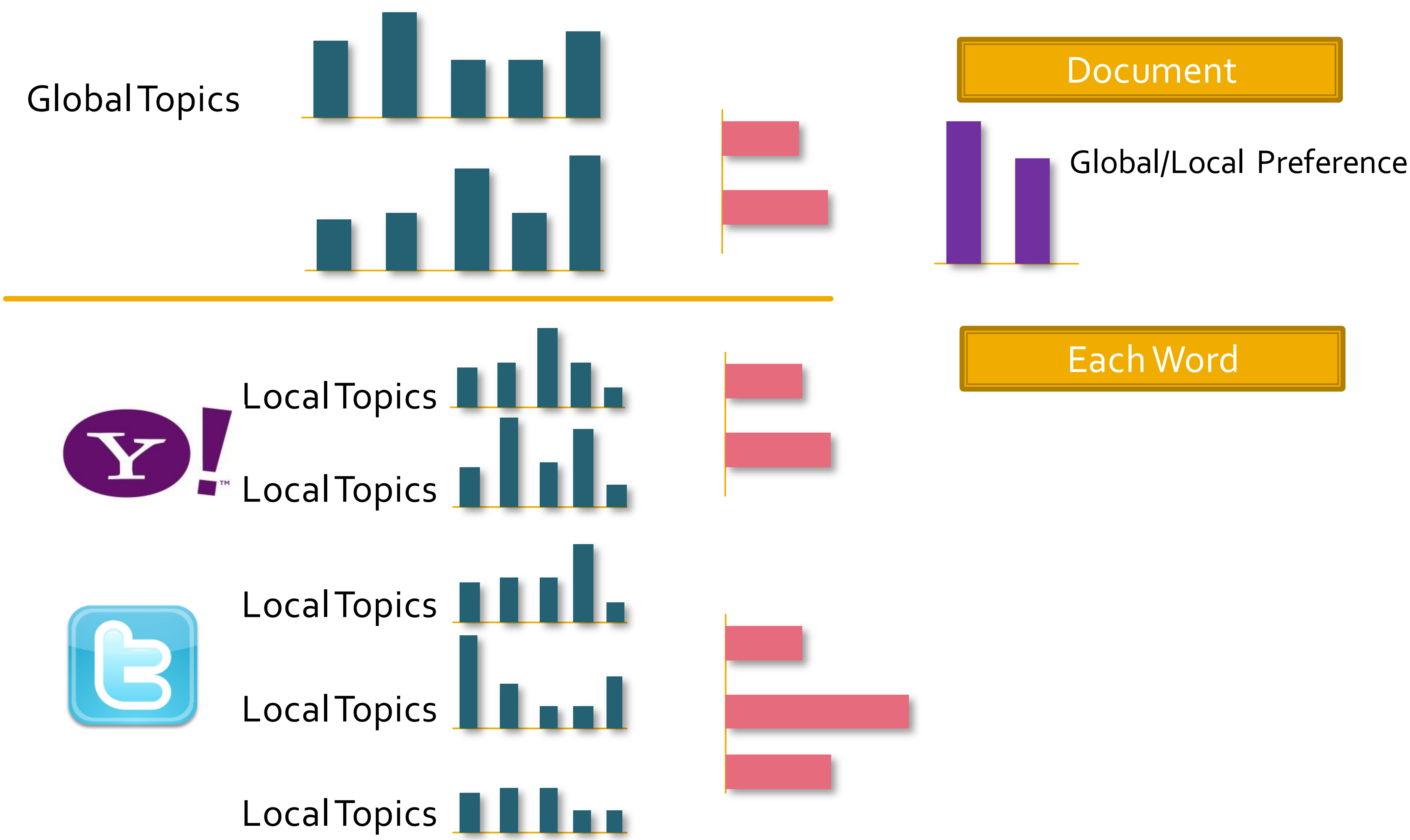
Modeling Temporal Dynamics
Handling Multiple Sources



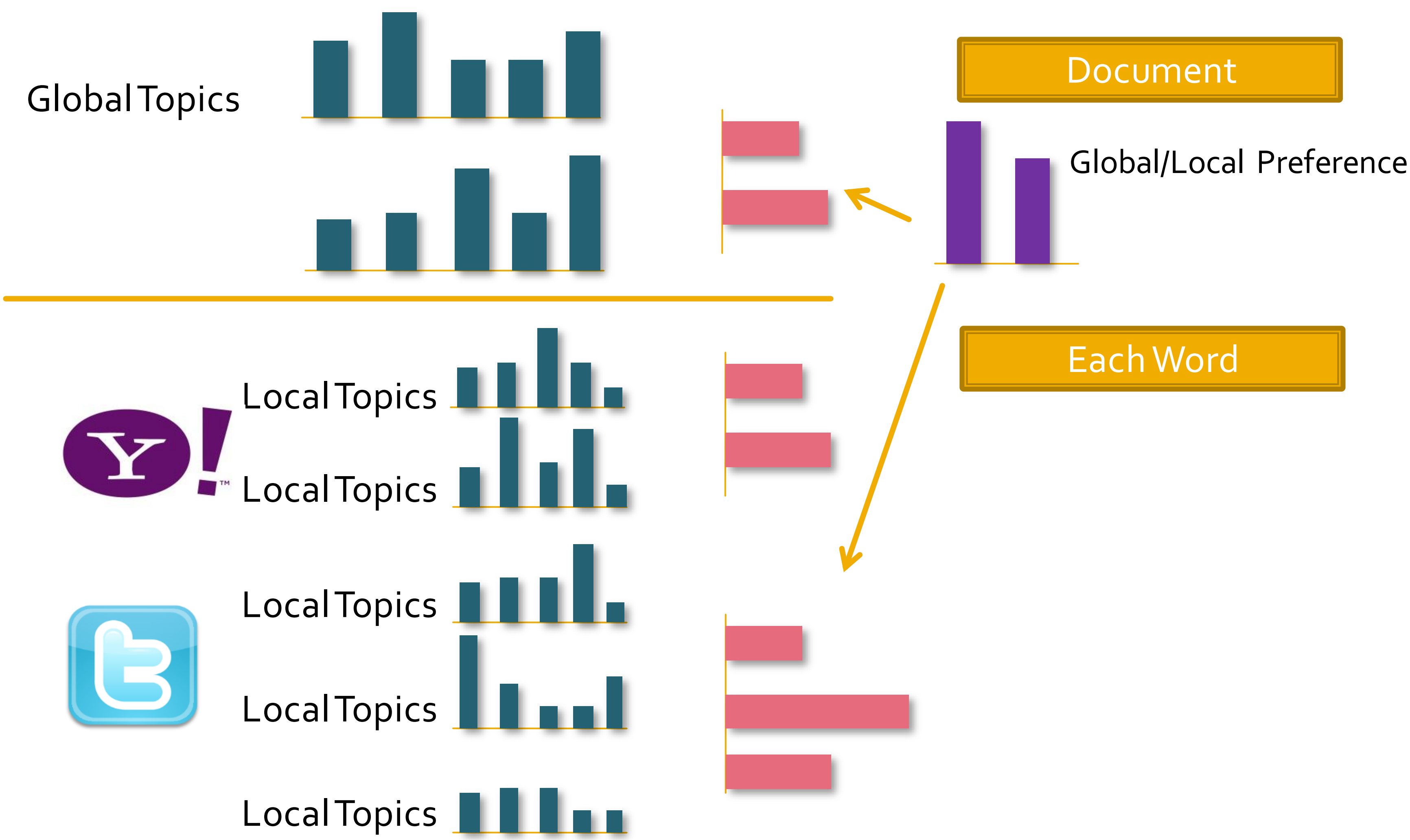
Modeling Temporal Dynamics
Handling Multiple Sources



Modeling Temporal Dynamics
Handling Multiple Sources



Modeling Temporal Dynamics
Handling Multiple Sources



Modeling Temporal Dynamics
Handling Multiple Sources

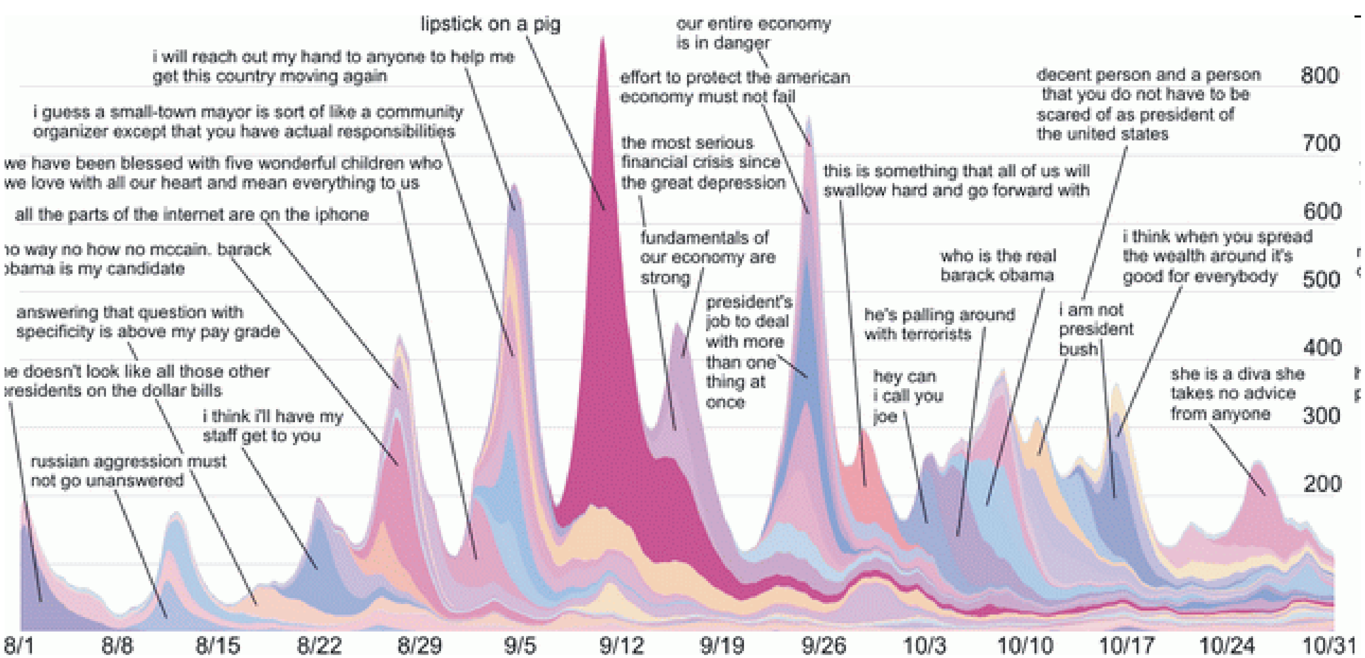
Modeling Temporal Dynamics
Handling Multiple Sources

Modeling Temporal Dynamics
Handling Multiple Sources

- Generative Process Summary:
  - Per-stream
    - Global/Local Preference Prior
    - Topic Prior
    - Language Model
  - Per-document
    - Global/Local Preference
    - Topic proportion
  - Per-token
    - Global/Local Choice
    - Topic Choice

Modeling Temporal Dynamics
Temporal Modeling

Meme Tracking



Modeling Temporal Dynamics
Temporal Modeling

Intuitions

Modeling Temporal Dynamics
Temporal Modeling

Intuitions

- Markovian Assumption

Modeling Temporal Dynamics
Temporal Modeling

Intuitions

- Markovian Assumption
  [Blei and Lafferty, ICML 2007]
  [Wang et al., UAI 2008]
  [Wei et al., IJCAI 2007]

Modeling Temporal Dynamics
Temporal Modeling

Intuitions

- Markovian Assumption
- Use a function to characterize the changes of topic proportions over time

[Wang and McCallum, KDD 2006]
[Yin et al., ICDM 2011]

Modeling Temporal Dynamics
Temporal Modeling

Intuitions

- Markovian Assumption
- Use a function to characterize the changes of topic proportions over time
  - At certain time t, we will have higher prior probability to choose some

Modeling Temporal Dynamics
Temporal Modeling

Assumptions

- Each topic only has one peak
- All topics are "trending"

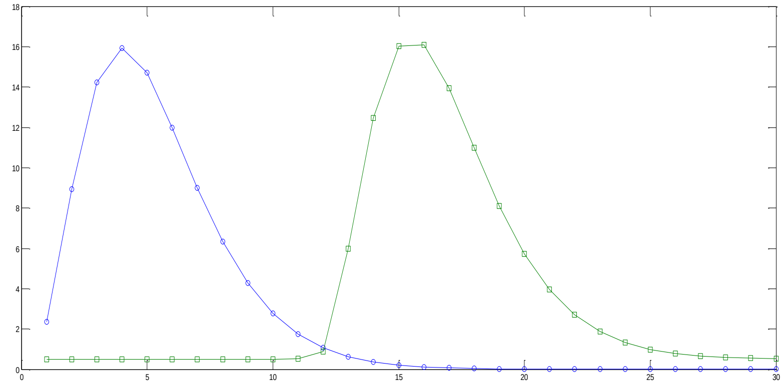
Assumptions

- Each topic only has one peak
- All topics are “trending”

- Yes, it’s naïve & simplified & unrealistic…

Temporal Function

$$\alpha_{t,k} = A_k t^{M_k} \exp(-L_k t)$$

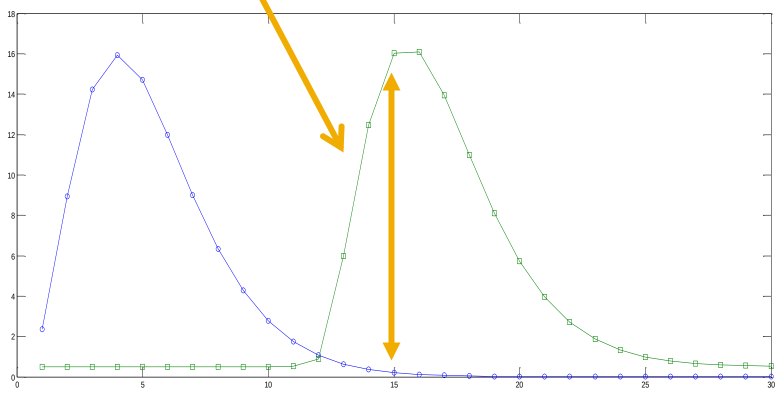


[Yang and Leskovec, WSDM 2011] [Leskovec et al., KDD 2009]

Modeling Temporal Dynamics
Temporal Modeling

Temporal Function

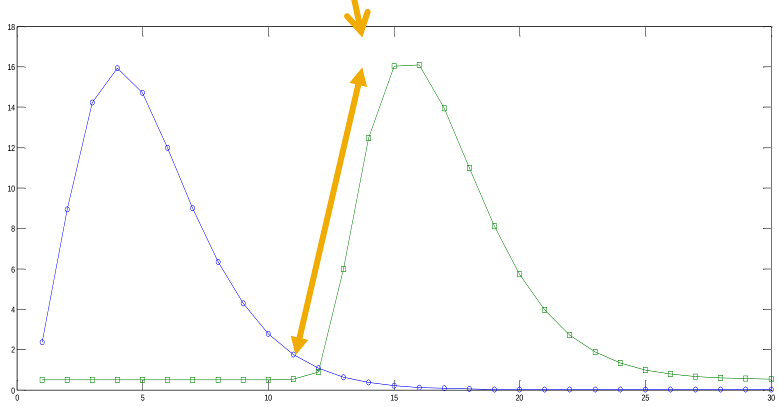
$$\alpha_{t,k} = A_k t^{M_k} \exp(-L_k t)$$



Modeling Temporal Dynamics
Temporal Modeling

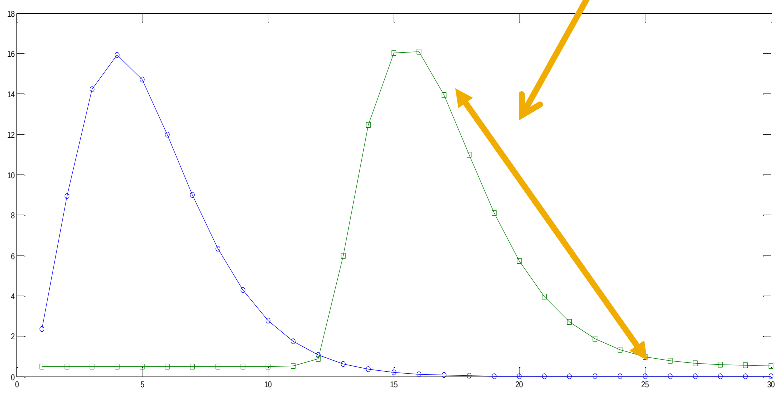
Temporal Function

$$\alpha_{t,k} = A_k t^{M_k} \exp(-L_k t)$$



Temporal Function

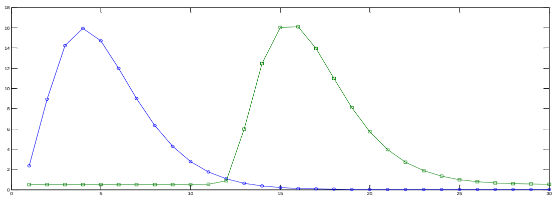
$$\alpha_{t,k} = A_k t^{M_k} \exp(-L_k t)$$

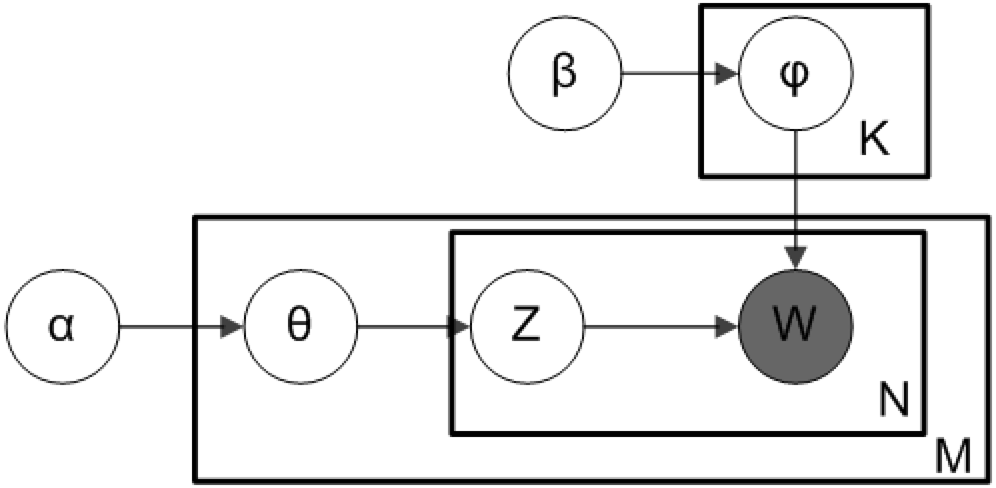


Modeling Temporal Dynamics
Temporal Modeling

Temporal Function

$$\alpha_{t,k} = A_k t^{M_k} \exp(-L_k t)$$





Temporal Modeling

- Overall Algorithm
  - EM-Style Algorithm
    - Gibbs Sampling in E-step
    - Functional Optimization in M-step
      Non-linear Least Square fit

Temporal Modeling

- Experiments & Conclusions
  - News & Tweets
  - 233,488 News articles
  - 1,736,350 Tweets
  - 720 hours in May, 2010

Temporal Modeling

- Experiments & Conclusions

Comparison of Top Ranked Common Topics between LDA (Left) and Temporal Collection (Right)			
Title	Top Terms	Title	Top Terms
“finance”	percent billion bank market greece financial banks debt	“finance”	percent billion bank greece financial debt banks euro crisis
“crime”	police car times vehicle found york square street bomb	“oil spill”	oil gulf spill coast drilling mexico water louisiana
“junk”	link cont via #jobs #fb album super live wii #tcot #news	“world cup”	world cup team league final players south season club
“oil spill’	oil gulf spill coast mexico gas drilling sea water	“health care”	health medical care cancer hospital patients study research
“junk”	dont people cant thats youre bad look tell talk	“UK election”	minister party prime cameron political leader president

Comparison of Local Topics between News (Left) and Twitter (Right)			
News		Twitter	
Title	Top Terms	Title	Top Terms
“crime”	police car times vehicle found york square street	“social media”	blog video post check news via twitter online facebook
“US election”	election party law president vote political campaign	“hash tags”	#fb info #quote #fail #ge #lol #ff #twibbon cont
“China”	minister china south india north chinese korea indian	“non-English”	les pas pour sur une cest est qui avec bien suis tout faire
“jobs”	budget tax million money pay bill federal increase cuts	“junk”	cant this wait watch next believe gonna watching just
“education”	school students schools board education district college	“junk”	that would have could never were wish there

Temporal Modeling

- Experiments & Conclusions

Comparison of Top Ranked Common Topics between LDA (Left) and Temporal Collection (Right)

Title	Top Terms	Title	Top Terms
“finance”	percent billion bank market greece financial banks debt	“finance”	percent billion bank greece financial debt banks euro crisis
“crime”	police car times vehicle found york square street bomb	“oil spill”	oil gulf spill coast drilling mexico water louisiana
“junk”	link cont via #jobs #fb album super live wii #tcot #news	“world cup”	world cup team league final players south season club
“oil spill’	oil gulf spill coast mexico gas drilling sea water	“health care”	health medical care cancer hospital patients study research
“junk”	dont people cant thats youre bad look tell talk	“UK election”	minister party prime cameron political leader president

Comparison of Local Topics between News (Left) and Twitter (Right)

News		Twitter	
Title	Top Terms	Title	Top Terms
“crime”	police car times vehicle found york square street	“social media”	blog video post check news via twitter online facebook
“US election”	election party law president vote political campaign	“hash tags”	#fb info #quote #fail #ge #lol #ff #twibbon cont
“China”	minister china south india north chinese korea indian	“non-English”	les pas pour sur une cest est qui avec bien suis tout faire
“jobs”	budget tax million money pay bill federal increase cuts	“junk”	cant this wait watch next believe gonna watching just
“education”	school students schools board education district college	“junk”	that would have could never were wish there

Temporal Modeling

- Experiments & Conclusions

Comparison of Top Ranked Common Topics between `LDA` (Left) and `Temporal Collection` (Right)

Title	Top Terms	Title	Top Terms
"finance"	percent billion bank market greece financial banks debt	"finance"	percent billion bank greece financial debt banks euro crisis
"crime"	police car times vehicle found york square street bomb	"oil spill"	oil gulf spill coast drilling mexico water louisiana
"junk"	link cont via #jobs #fb album super live wii #tcot #news	"world cup"	world cup team league final players south season club
"oil spill'	oil gulf spill coast mexico gas drilling sea water	"health care"	health medical care cancer hospital patients study research
"junk"	dont people cant thats youre bad look tell talk	"UK election"	minister party prime cameron political leader president

Comparison of Local Topics between News (Left) and Twitter (Right)

News		Twitter	
Title	**Top Terms**	**Title**	**Top Terms**
"crime"	police car times vehicle found york square street	"social media"	blog video post check news via twitter online facebook
"US election"	election party law president vote political campaign	"hash tags"	#fb info #quote #fail #ge #lol #ff #twibbon cont
"China"	minister china south india north chinese korea indian	"non-English"	les pas pour sur une cest est qui avec bien suis tout faire
"jobs"	budget tax million money pay bill federal increase cuts	"junk"	cant this wait watch next believe gonna watching just
"education"	school students schools board education district college	"junk"	that would have could never were wish there

Temporal Modeling

- Experiments & Conclusions

Comparison of Top Ranked Common Topics between LDA (Left) and Temporal Collection (Right)			
Title	**Top Terms**	**Title**	**Top Terms**
"finance"	percent billion bank market greece financial banks debt	"finance"	percent billion bank greece financial debt banks euro crisis
"crime"	police car times vehicle found york square street bomb	"oil spill"	oil gulf spill coast drilling mexico water louisiana
"junk"	link cont via #jobs #fb album super live wii #tcot #news	"world cup"	world cup team league final players south season club
"oil spill'	oil gulf spill coast mexico gas drilling sea water	"health care"	health medical care cancer hospital patients study research
"junk"	dont people cant thats youre bad look tell talk	"UK election"	minister party prime cameron political leader president

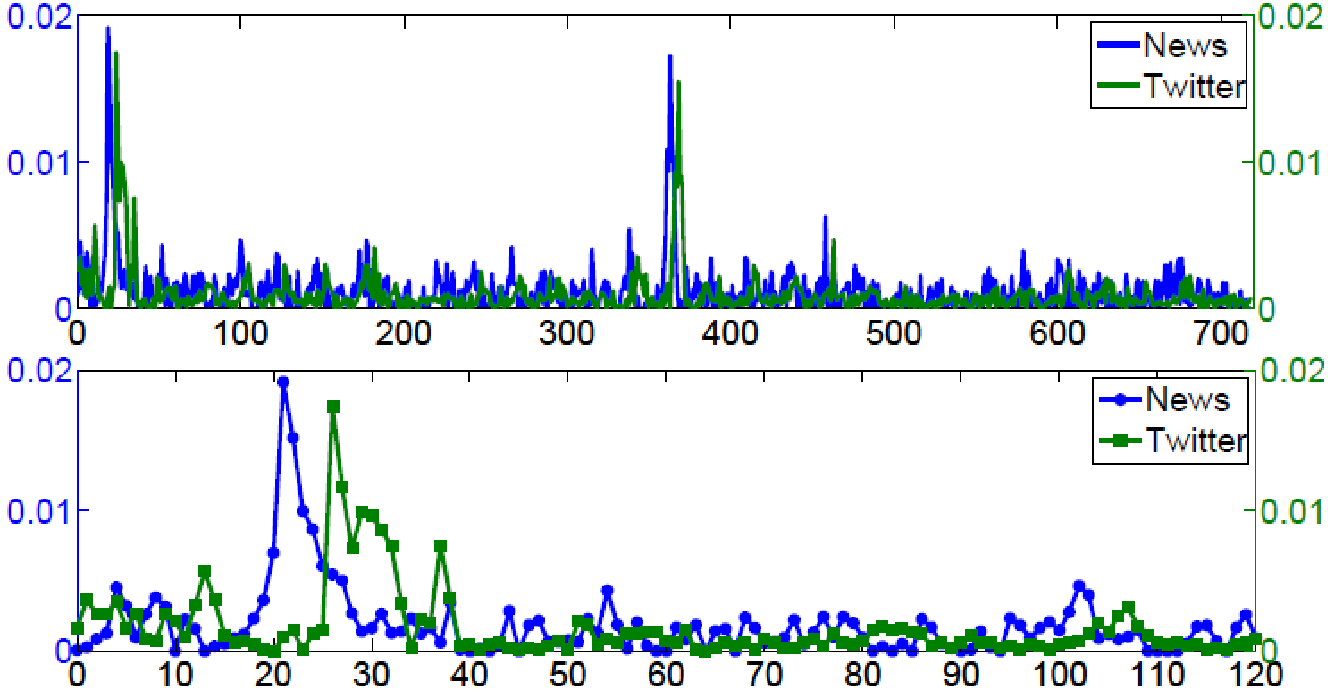
Comparison of Local Topics between News (Left) and Twitter (Right)			
News		Twitter	
Title	**Top Terms**	**Title**	**Top Terms**
"crime"	police car times vehicle found york square street	"social media"	blog video post check news via twitter online facebook
"US election"	election party law president vote political campaign	"hash tags"	#fb info #quote #fail #ge #lol #ff #twibbon cont
"China"	minister china south india north chinese korea indian	"non-English"	les pas pour sur une cest est qui avec bien suis tout faire
"jobs"	budget tax million money pay bill federal increase cuts	"junk"	cant this wait watch next believe gonna watching just
"education"	school students schools board education district college	"junk"	that would have could never were wish there

Temporal Modeling

- Experiments & Conclusions
  Case Study on A Common Topic "Kentucky Derby"
  - Select a common topic which ranks the following terms high:
    "derby", "race", "borel", "kentucky" and "horse" …
  - Tracking temporal dynamics of a topic

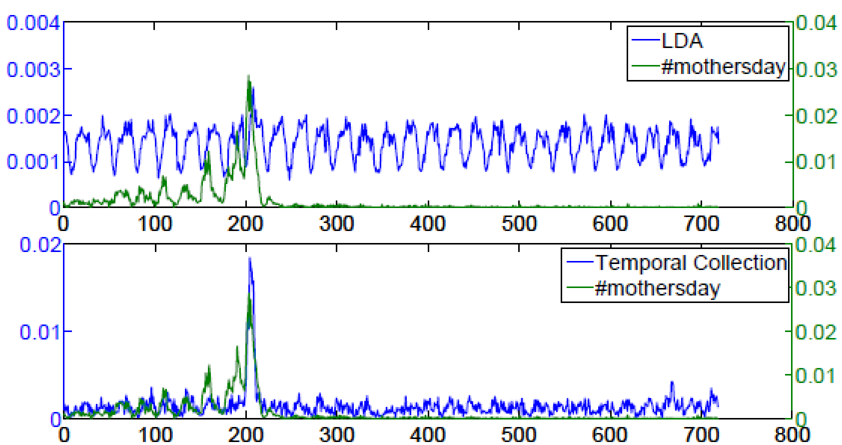
Temporal Modeling

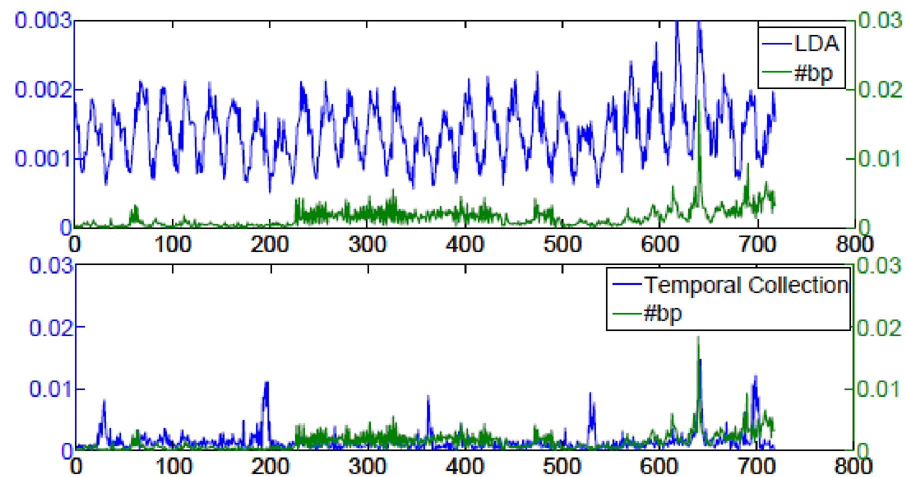
- Experiments & Conclusions



Temporal Modeling

- Experiments & Conclusions





Conclusion

- A framework for modeling temporal dynamics for multiple sources.
- *Principled* way to tackle the problem.
- Bridge topic modeling & Information cascading.

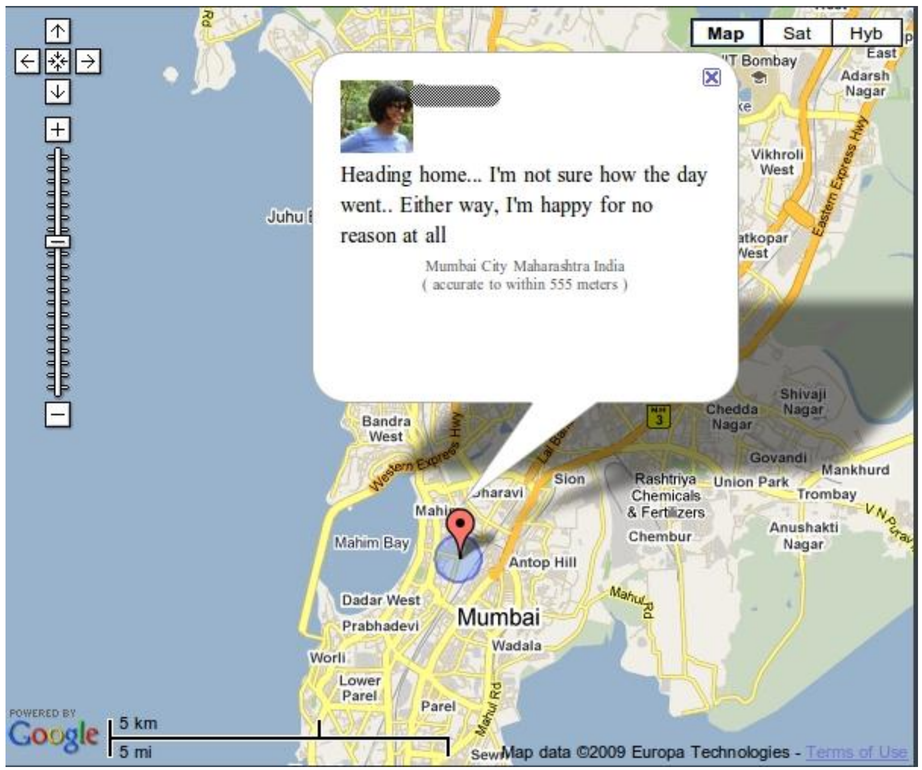
Temporal Dynamics & Geographical Language Variations

Temporal Dynamics + Multiple Sources

Geographical Language Variations

Modeling Geographical Language Variations

Social Stream + Locations



Interesting Questions

- How is information created and shared in different geographic locations? What is the inherent geographic variability of content?
- What are the spatial and linguistic characteristics of people? How does this vary across regions?
- Can we discover patterns in users' usage of micro-blogging services?

Interesting Questions

- How is information created and shared in different geographic locations? What is the inherent geographic variability of content?
- What are the spatial and linguistic characteristics of people? How does this vary across regions?
- Can we discover patterns in users' usage of micro-blogging services?

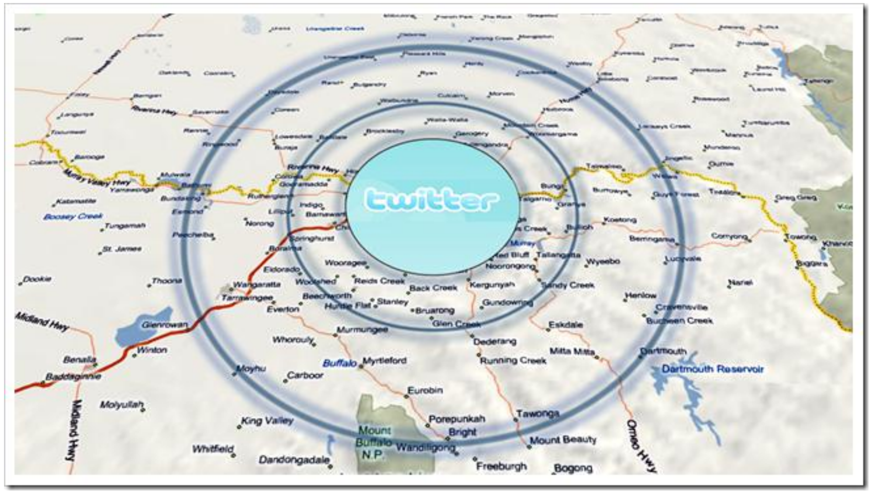
- Can we predict user location from tweets?

Modeling Geographical Language Variations

Applications

Behavioral targeting and user modeling

Better local information filtering



Modeling Geographical Language Variations

Technical Challenges

- Tweets
  - noisy and short (140 characters)
- Only 1% of tweets geo-tagged
  - Can we predict locations for non-tagged tweets?
- Many intuitions to be combined
  - Background, regional language models, topics
  - Personal preferences, regional preferences…

…

Modeling Geographical Language Variations

Can we really infer locations for a tweet?

Yes via tweet decomposition

Modeling Geographical Language Variations

Just landed after a long flight. It is raining here at Lyon though!

What is the user's location?

Modeling Geographical Language Variations

Just landed after a long flight. It is raining here at Lyon though!

background

just
after
It
be
the
can
cant
will

Modeling Geographical Language Variations

Just landed after a long flight. It is raining here at Lyon though!

Travel	background
landed	just
flight	after
delay	It
TSE	be
Gate	the
terminal	can
	cant
	will

Modeling Geographical Language Variations

Just landed after a long flight. It is raining here at Lyon though!

Travel/airport	background	SE airport area
landed	just	Lyon
flight	after	Saint
delay	It	Exupery
TSE	be	convention
Gate	the	center
terminal	can	raining
	cant	
	will	

Modeling Geographical Language Variations

Semantic Topic

Travel/airport

landed
flight
delay
TSE
Gate
terminal

Background Language Model

background

just
after
It
be
the
can
cant
will

Regional Language Model

SE airport area

Lyon
Saint
Exupery
convention
center
raining

Modeling Geographical Language Variations

Delayed again at the TSE check point and might miss my flight. way to go SF!

Travel/airport	background	SFO
landed	just	SF
flight	after	SFO
delay	It	San
TSE	be	Francisco
Gate	the	airport
terminal	can	
	cant	
	will	

Modeling Geographical Language Variations

Can we always do that?

Modeling Geographical Language Variations

Life is good! Feeling great today!

Modeling Geographical Language Variations

Life is good! Feeling great today!

Daily life

life
feeling
good
today
morning

background

just
after
It
be
the
can
cant
will

?

Modeling Geographical Language Variations

Life is good! Feeling great today!

If we know something extra about the context and **user location preferences**, perhaps we can do better than random guessing!

Modeling Geographical Language Variations

Previous work

- Simple regional language models
  - No factorization
- No personal preferences
- Complicated inference algorithms
  - Usually two step process
  - Fails to learn coherent regions

Modeling Geographical Language Variations

- Motivations
- Our Proposed Model
- Experiments
- Conclusions

Modeling Geographical Language Variations

- A novel probabilistic model considers
  - Regional language models
  - Global topics
  - Personal preferences
- Sparse modeling + Bayesian treatment
- An efficient inference algorithm

Modeling Geographical Language Variations

- Basic Intuition
  - Regions
  - Topics
  - Users
  - Tweets
- The generative process
  - Intuition
  - Glory details

Modeling Geographical Language Variations
Basic Intuition: Region

- Must be coherent
  - There is enough traffic in it
  - Affects the way we write tweets
    - Has preference over what topic discussed
    - Specific keywords
  - Area over the map
  - Example
    - An airport
    - A park
    - A mall
    - A city

Modeling Geographical Language Variations
Basic Intuition: Topic

- Classify the content of the tweet
- Might not tell us the location
- Puts a distribution over words
- Examples
  - Sports
  - Politics
  - Travel
  - Daily life, etc

Modeling Geographical Language Variations
Basic Intuition: User

- Has preferences over locations
  - Where he usually spends his/her time
- Has preference over topics
  - What he tweets about

Modeling Geographical Language Variations
Basic Intuition: Tweet

- Written by a given user
- At a specific location (region)
  - Depends on the user
- About a specific topic
  - Depends on
    - What the user talks about
    - What is being discussed at this location
- Composed of a bag of words from
  - Topic + location + background language models

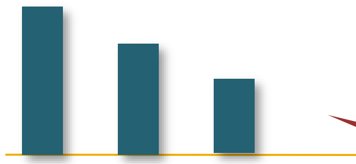
Modeling Geographical Language Variations
The Model

- Basic Intuition
  - Regions
  - Topics
  - Users
  - Tweets
- The generative process
  - Intuitive explanation
  - Glory details

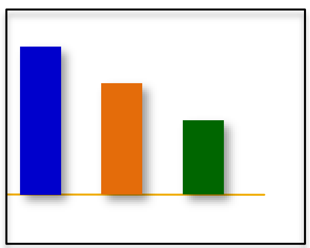
Modeling Geographical Language Variations
How a tweet is being generated?

- Pick a location
- Pick a topic
- Generate the words

Modeling Geographical Language Variations
How a tweet is being generated?

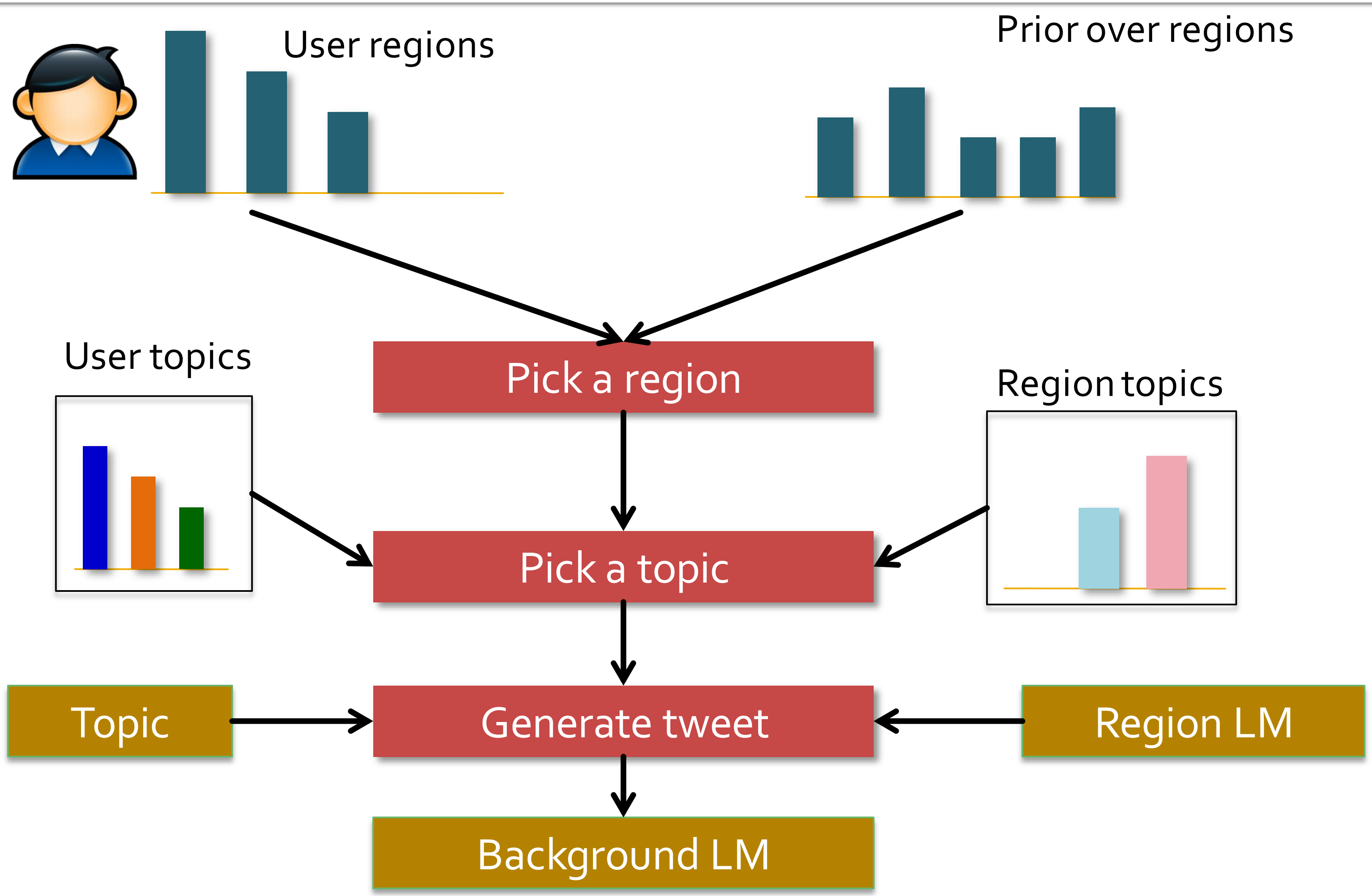


Preferences over regions
Regions are unsupervised
Just an area over the map

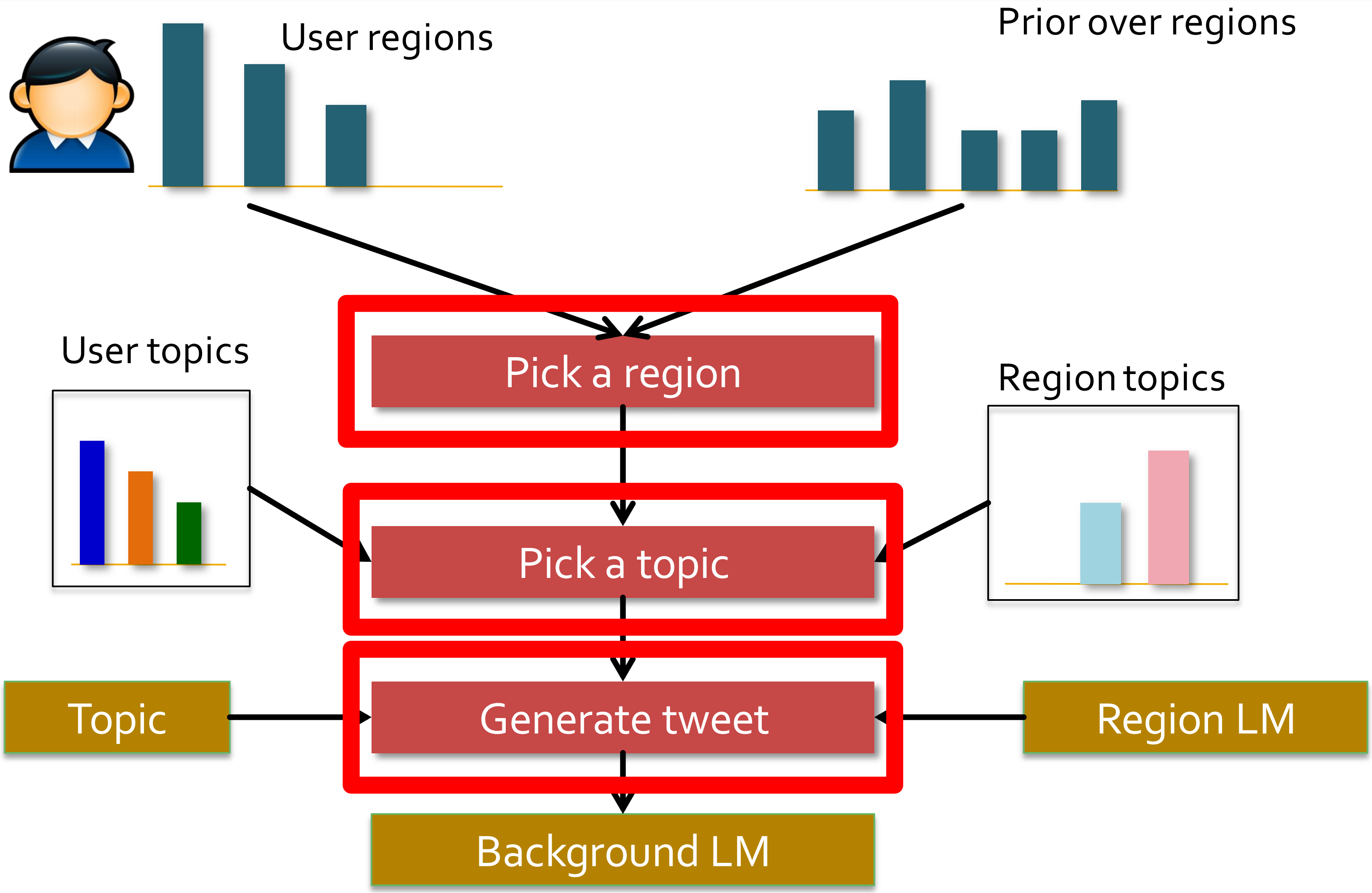


Preference over topics:
What he likes to talk about

Modeling Geographical Language Variations
How a tweet is being generated?



Modeling Geographical Language Variations
How a tweet is being generated?



Modeling Geographical Language Variations
Discrete Additive Models

- Switch-based models
  - Normalized distributions
  - Pick one distribution
  - Sample from it
- SAGE [Eisenstein et. al, 2011]
  - Un-normalized distribution
    - Log frequencies
  - Add them all together
  - Exponentiate and sample

Modeling Geographical Language Variations
SAGE

An Additive model for discrete distributions

- Discrete distribution via natural parameters
  Example:

$$p(v|\boldsymbol{\phi}) = \exp\left(\boldsymbol{\phi}_v - g(\boldsymbol{\phi})\right) \text{ where } g(\boldsymbol{\phi}) = \log \sum_v \exp\left(\boldsymbol{\phi}_v\right)$$

- Log-frequency differences
- Addition of multiple models
  Example:

$$P(v|\boldsymbol{\phi}_0, \boldsymbol{\phi}_u, \boldsymbol{\phi}_g) := p(v|\boldsymbol{\phi}_0 + \boldsymbol{\phi}_u + \boldsymbol{\phi}_g)$$

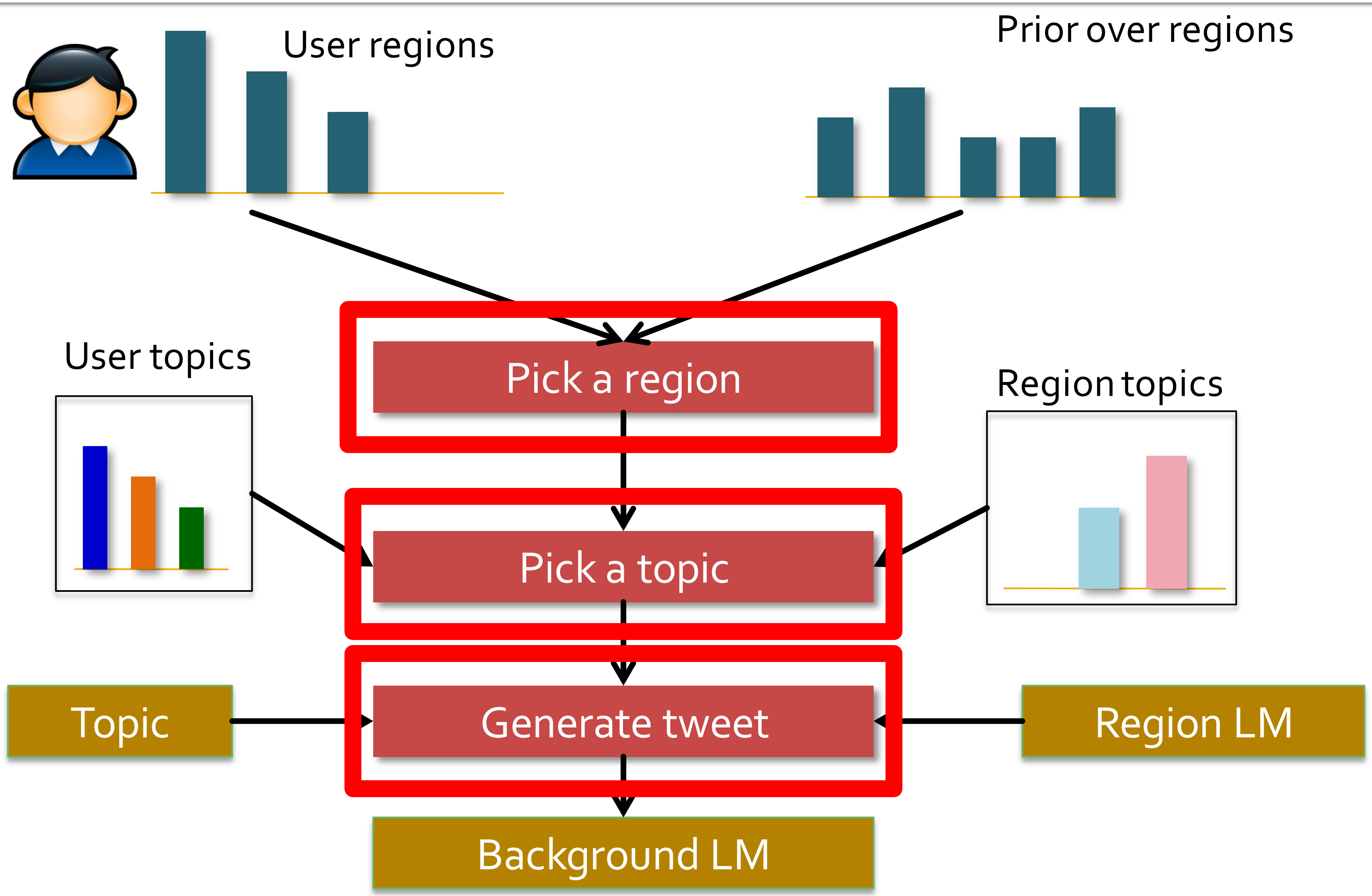
Modeling Geographical Language Variations SAGE

Use `SAGE` to replace “switch” variables to enable us incorporate multiple sources in different levels of our model easily

- Language models
  Example: background, regional, global…
- User preferences
  Example: global, regional, personal…

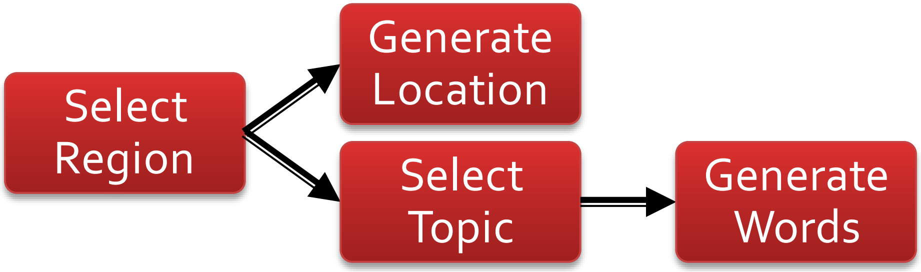
…

Modeling Geographical Language Variations
SAGE



Modeling Geographical Language Variations
Recap

- Generative Process



- Sparse Modeling
  - L_1 regularization (Laplace priors)
- Geographical Modeling
  - Bayesian treatment

Modeling Geographical Language Variations
Inference Algorithm

- A variant of Monte Carlo EM
  - "E-Step": Sample latent discrete variables
  - "M-step": Update all model parameters

- Sparse update of gradients
- L_1 regularization: `ISTA` algorithm
- Initialize regions with `K-means` algorithm

Modeling Geographical Language Variations Experiments

Dataset

- Twitter data
  - Randomly sample 1,000 users
  - All tweets from Jan 2011 to May 2011
  - 573,203 distinct tweets
- Twitter geographical data
  - Locations + Twitter Places

Modeling Geographical Language Variations Experiments

Location Prediction

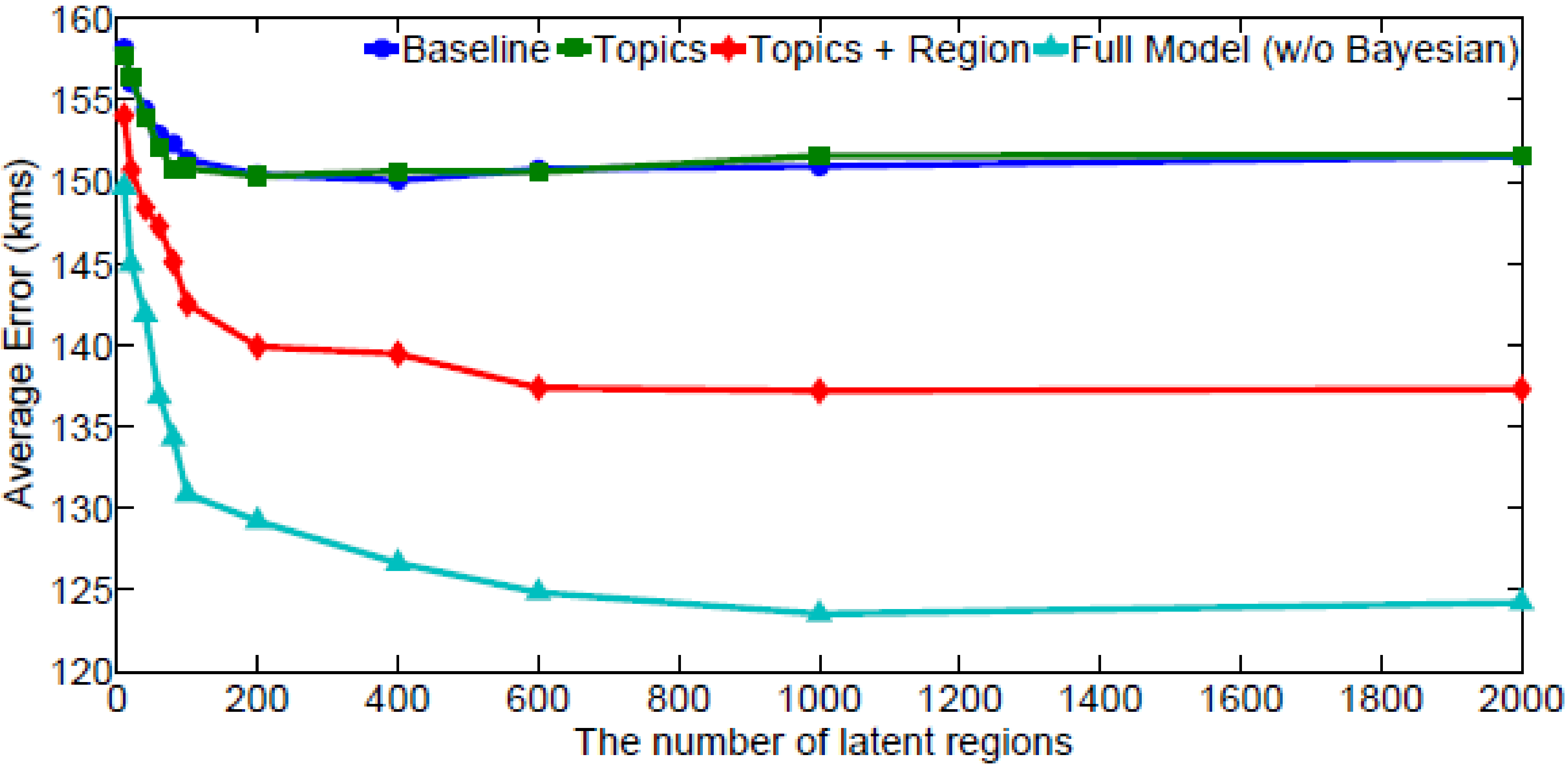
- Metric
  - average error distance
  - Kilometers

Modeling Geographical Language Variations Experiments

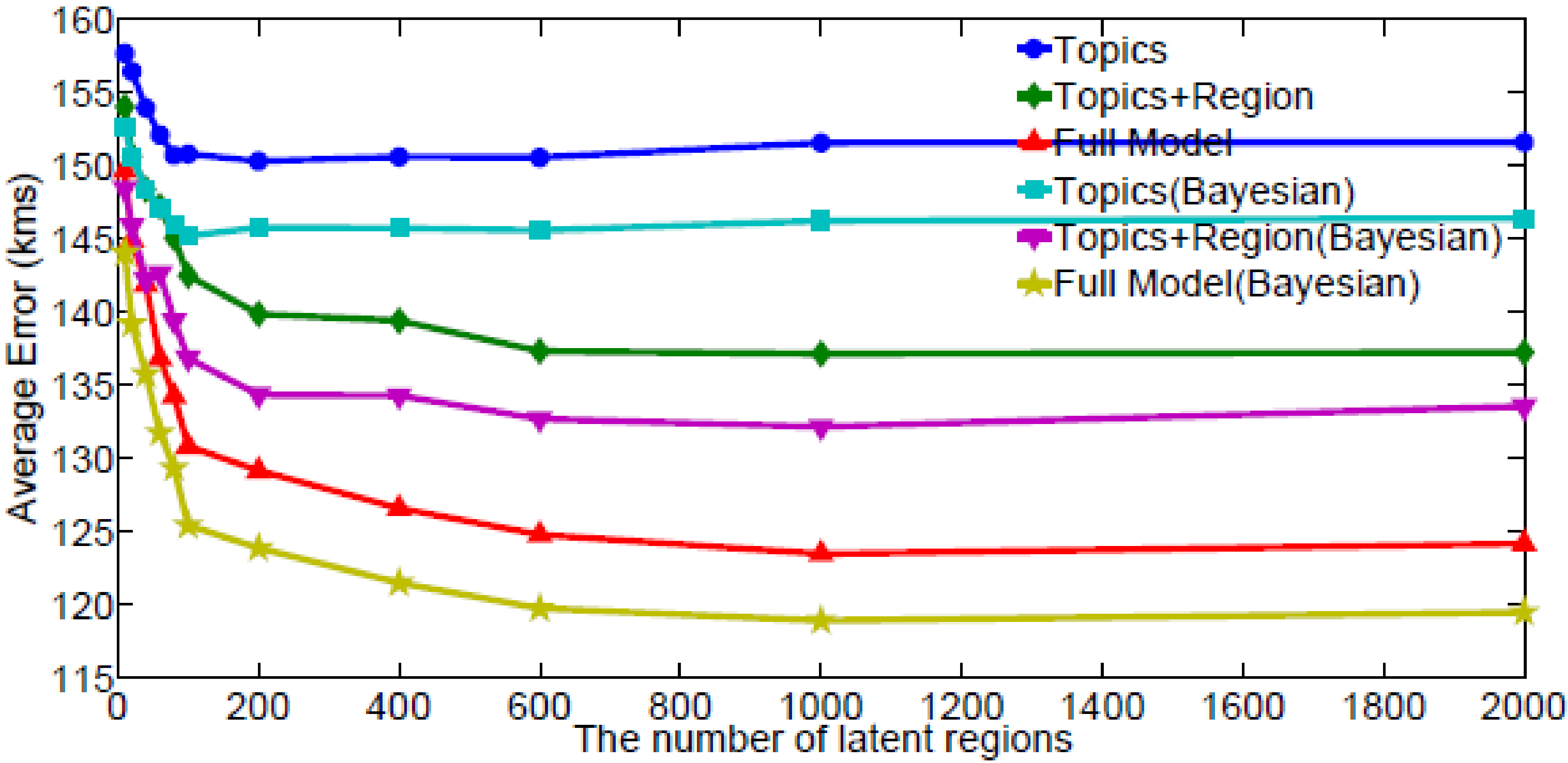
Location Prediction

- Baselines
  - [Yin et al. WWW 2011] paper
    - PLSA formalism
    - No personalization
  - Our model without ϕ^{geo}, η^{user} and θ^{user}
    - Similar to Yin et al.'s formalism but SAGE model
  - Our model without η^{user} and θ^{user}

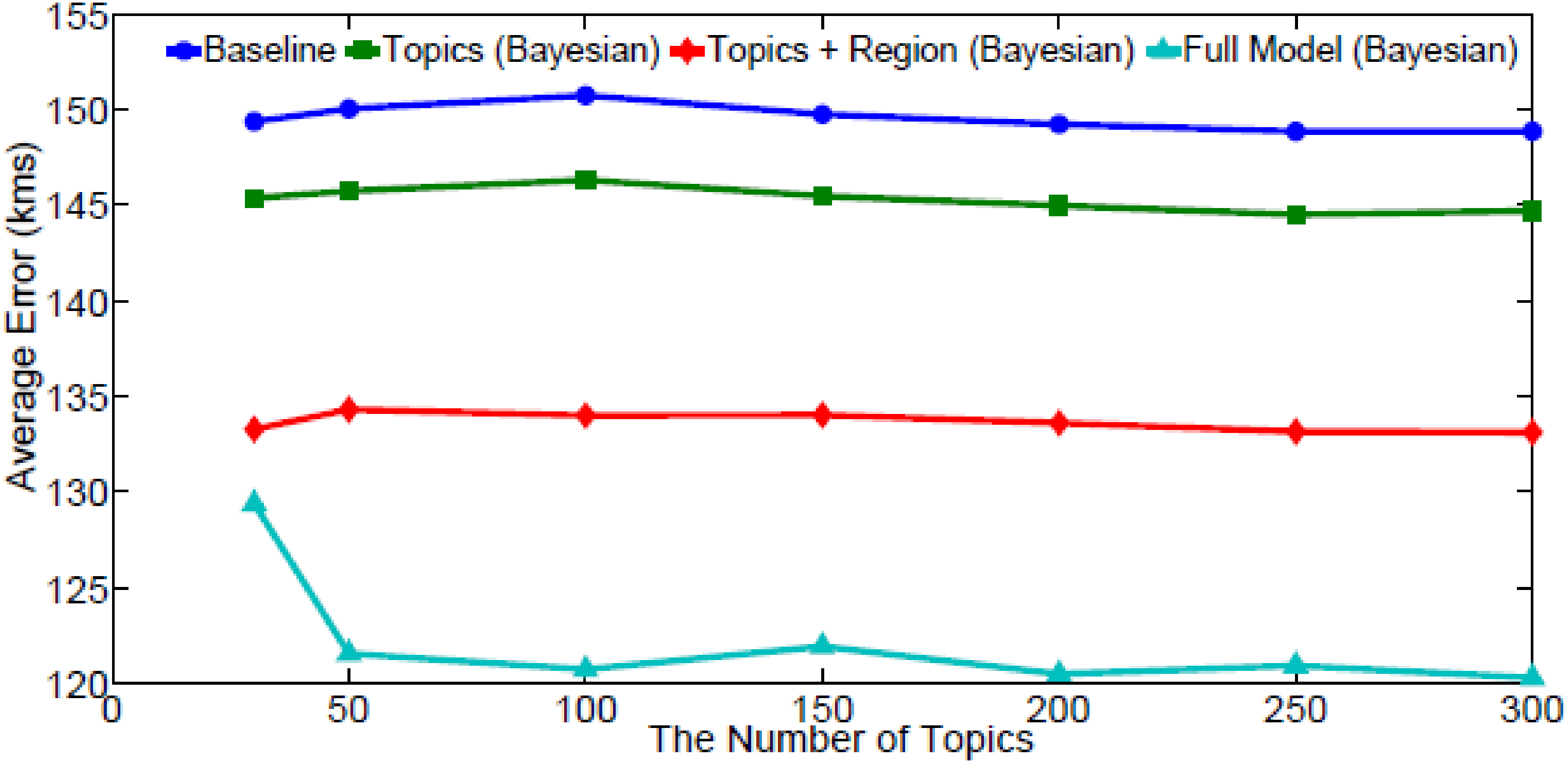
Modeling Geographical Language Variations
Location Prediction



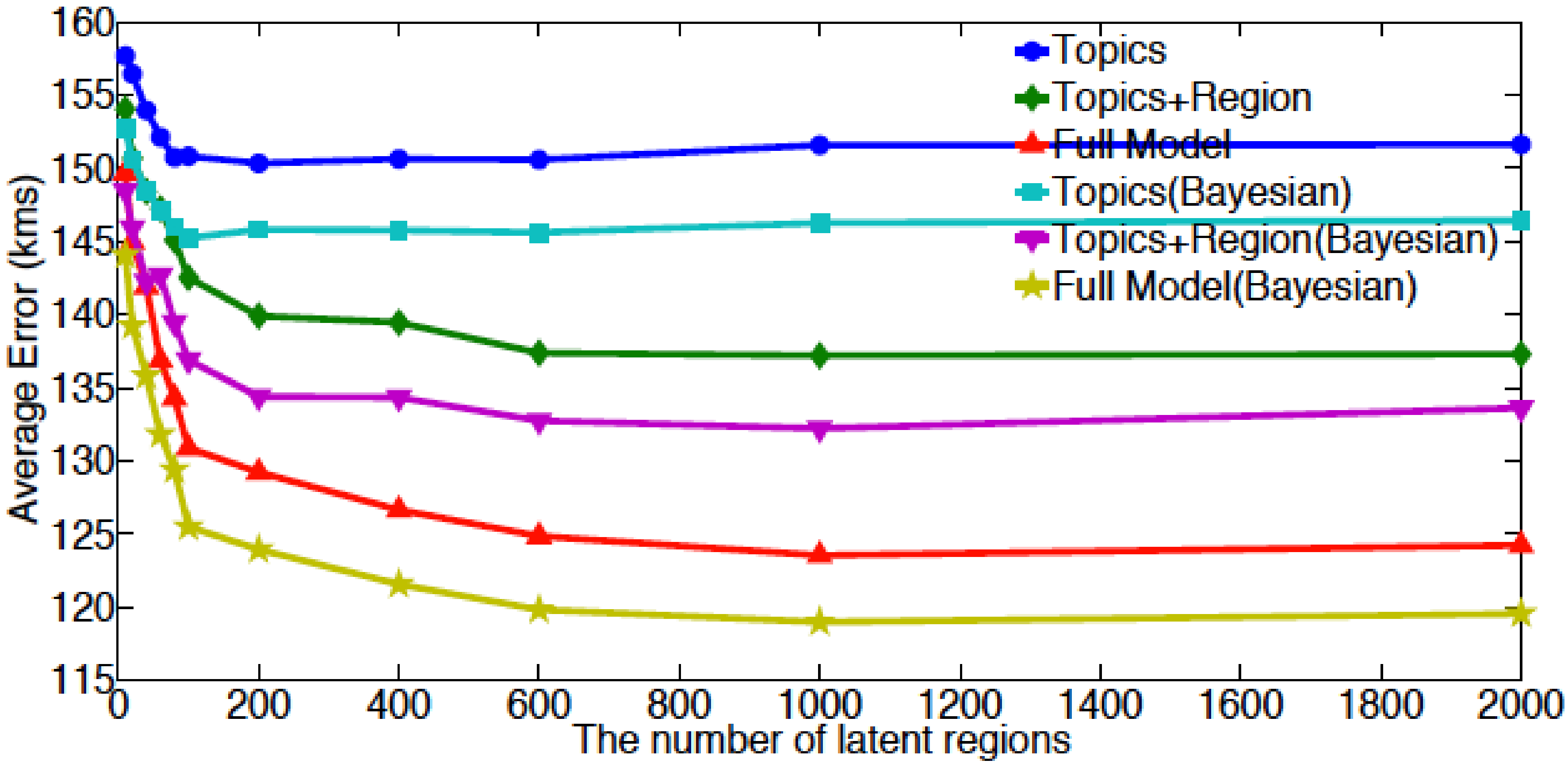
Modeling Geographical Language Variations
Bayesian Treatment



Modeling Geographical Language Variations
Number of Topics



Modeling Geographical Language Variations
Number of Regions

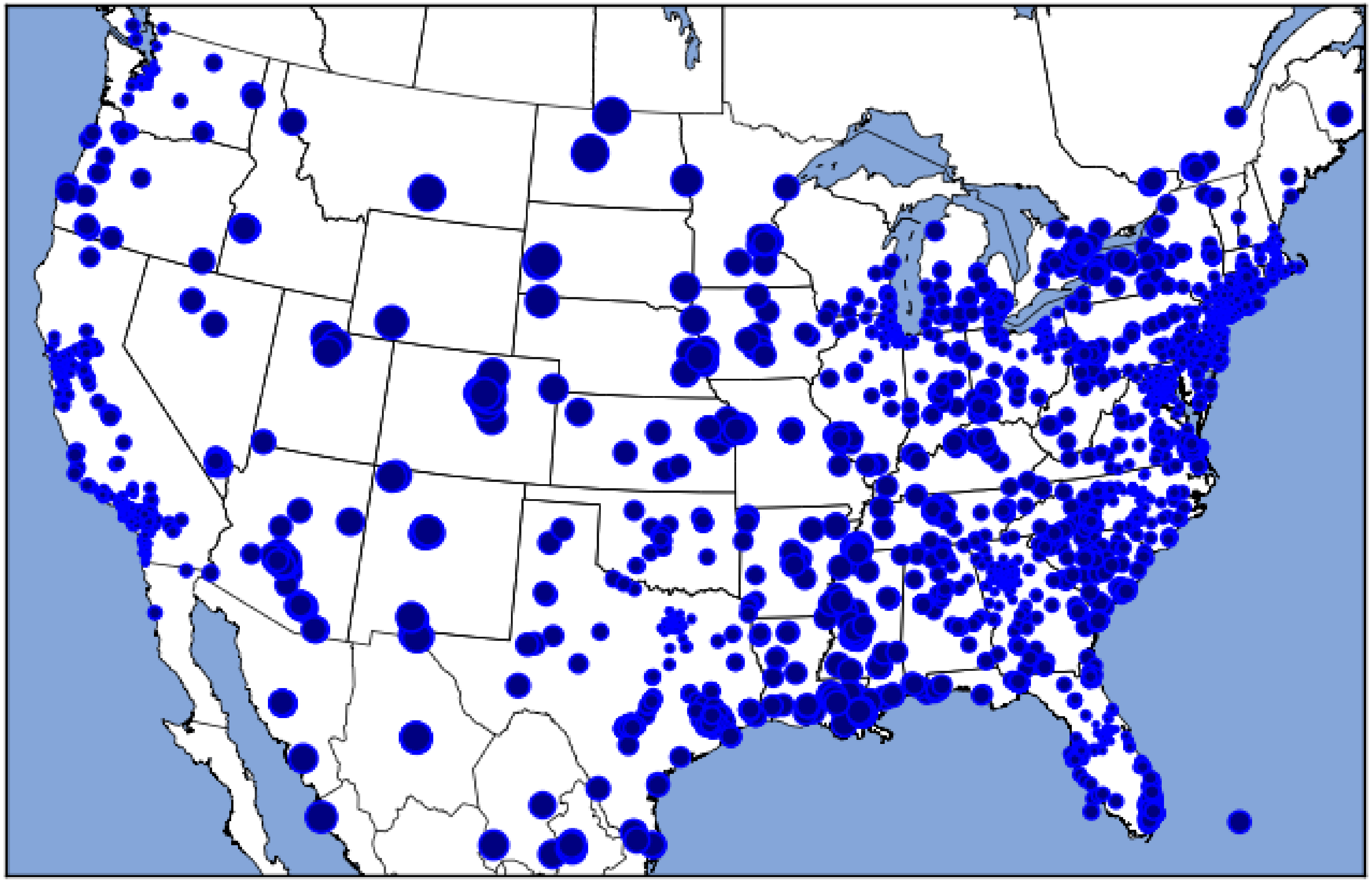


Modeling Geographical Language Variations Experiments (Public Data)

# of regions	[3]	[2]	[1]	Topics	Topics + Region	Full Model
10	494	479	501	540.60	481.58	449.45
20	494	479	501	522.18	446.03	420.83
40	494	479	501	513.06	414.95	395.13
60	494	479	501	507.37	410.09	380.04
80	494	479	501	499.42	408.38	374.01
100	494	479	501	498.94	407.78	**372.99**

[1] Eisenstein et al. EMNLP 2010.
[2] Wing and J. Baldridge. ACL 2011.
[3] Eisenstein, Ahmed, Xing ICML 2011.

Modeling Geographical Language Variations
Error Analysis



Modeling Geographical Language Variations
Global and local topics

Entertainments
lady bieber album music beats artist video listen itunes apple produced movies #bieber lol new songs
Sports
yankees match nba football giants wow win winner game weekend horse #nba
Politics
obama election middle east china uprising egypt russian tunisia #egypt afghanistan people eu

Location with Top Ranked Terms
United States->New York->Brooklyn
brooklyn ave flatbush avenue mta prospect 5th #brooklyn spotlight carroll bushwick museum broadway madison vanderbilt coney slope eastern subway new york pkwy #viernesnayobon #mets otsego greenwich starbucks
United States->California->San Francisco
sfo francisco san airport international millbrae terminal flight burlingame bart mateo boarding bayshore telecommute landed heading bay airlines united bound flying #sfo camino groupon caltrain moon tsa baggage california engineer valley
United States->Pennsylvania->Philadelphia
philadelphia #philadelphia phl #jobs market others #job street philly walnut septa chestnut the cherry sansom arch spruce citizens locust btw temple pennsylvania rittenhouse passyunk bitlyetq7a6 bookrenters pike international
United Kingdom->England->London
winds lhr hounslow terminal the cloudy mph ickenham bath heathrow temperature airport car only airways uxbridge sun splendid fair london british lounge tothers harmondsworth speedbird whens for stars day flight dominos navigation brunel
Australia->New South Wales->Sydney
sydney #sydney bondi george street mascot domestic syd surry station cnr platforms harbour darlinghurst qantas hoteloxford eddy haymarket terminal wales australia chalmers uts pitt #marketing junction darling centre #citijobs citigroup druitt

Modeling Geographical Language Variations
Conclusions

- Probabilistic model for geographical information
  - Regional variations
  - Personal preferences
- Effective inference algorithm
- Best location prediction
- Discriminatively learned language models
- Future work
  - Hierarchical model
  - Hash tags
  - Temporal location model

A little bit more about me …

- Ph.D. candidate at Lehigh (5.5 years)
- Published 10+ technical papers
  - KDD (3), SIGIR (2[1]), WWW (1[2]), WSDM (1) , AAAI (1) and CIKM ([1])
  - Best Poster in WWW 2011
- Four internships
  - Yahoo! Labs (2010, 2011)
  - LinkedIn (2011)
  - A local software company (2008)
- Collaborated with
  Alex Smola (Google) , Amr Ahmed (Google), Marco Pennacchiotti (eBay), Siva Gurumurthy (Twitter), Kostas Tsioutsiouliklis (Twitter), Jian Guo (Harvard Univ.), Ron Bekkerman (LinkedIn), Brian D. Davison (Lehigh Univ.), Dawei Yin (Lehigh Univ.), Ovidiu Dan (Microsoft), Zaihan Yang (Lehigh Univ.), Zhenzhen Xue (Google)

Questions?

Generative Process

1. For all common topics T_c, draw $\phi^{(c)} \sim \text{Dir}(\beta^{(c)})$
2. For a particular stream s
   (a) For all local topics T_s, draw $\phi^{(s)} \sim \text{Dir}(\beta^{(s)})$
   (b) For each document d in s
      i. Draw Bernoulli parameter $\eta_{s,d} \sim \text{Beta}(\gamma_s^{(s)}, \gamma_s^{(c)})$
      ii. Draw $\theta_d^{(s)} \sim \text{Dir}(\alpha_s)$
      iii. Draw $\theta_d^{(c)} \sim \text{Dir}(\alpha_c)$
      For each word position i in document d
         A. Draw $x_{di} \sim \text{Bernoulli}(\eta_{s,d})$
         B. Draw a topic $z_{di} \sim \text{Multinomial}(\theta_d^{(x_{di})})$
         C. Draw a word $w_{di} \sim \text{Multinomial}\ (\phi_{z_{di}}^{(x_{di})})$

Approximate Inference

- Gibbs Sampling

$$p(x_{di} = s, z_{di} = t) \propto$$
$$\frac{c_{d,s-i} + \gamma_s}{N_d + \gamma_s + \gamma_c - 1} \frac{m_{d,z-i} + \alpha_z}{\sum_{z \in T_s} m_{d,z-i} + \alpha_z} \frac{n_{z,w-i} + \beta_w^{(s)}}{\sum_w^V n_{z,w-i} + \beta_w^{(s)}}$$
$$p(x_{di} = c, z_{di} = t) \propto$$
$$\frac{c_{d,c-i} + \gamma_c}{N_d + \gamma_s + \gamma_c - 1} \frac{m_{d,z-i} + \alpha_z}{\sum_{z \in T_c} m_{d,z-i} + \alpha_z} \frac{n_{z,w-i} + \beta_w^{(c)}}{\sum_w^V n_{z,w-i} + \beta_w^{(c)}}$$

Temporal Modeling

- Temporal Function

$$V(t+1) = cf[V(t)]\delta(t)$$

  - *V(t)*: volume of the story
  - *f(v)*: a function of volume, encode “popularity”
  - *σ(t)*: a function of time, encode “decay”

Temporal Modeling

- Temporal Function

For some choices of function f and σ, we can analytically solve volume *V(t)*:

$$A_k t^{M_k} \exp(-L_k t)$$

Temporal Modeling

- Overall Algorithm

initialize Gibbs Sampler
while not converge **do**
  E-step
  For all documents in all text streams, update topic assignments using Equation (1)
  M-step
  Update α, β and γ values through the method introduced in [16]
  for each all local and common topics **do**
    1) Fit "Gaussian" function to α values
    2) Fit "Temporal Gamma" function by using the parameters from the previous step
    3) Re-calculate α values for topic k by using fitted function
  end for
end while

Modeling Temporal Dynamics

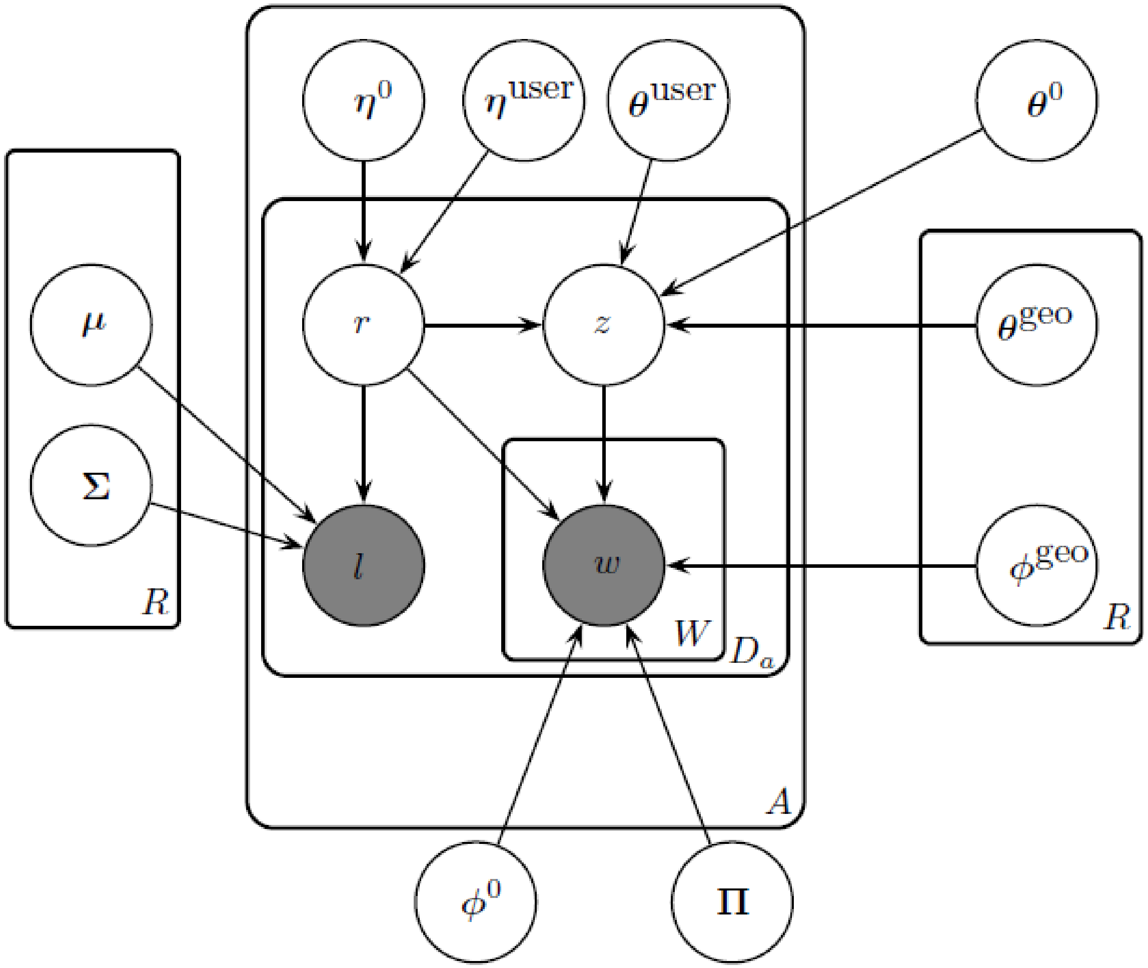
Hashtag	Top Terms of Mapped Topic
[a] Hashtag Mapping for LDA model	
#mothersday	family home life children mother son friends
#memorialday	event june call center community club park
#bp	oil gulf spill coast mexico gas drilling
#kentuckyderby	race car track kentucky win top cars
#gaga & #justinbieber	justin lady super try bieber ider rio gaga jonas
[b] Hashtag Mapping for Temporal Collection model	
#mothersday	family children day home life church mother
#memorialday	memorial event day june community center
#bp	oil gulf spill coast drilling mexico water louisiana
#kentuckyderby	derby race borel kentucky horse super
#gaga & #justinbieber	bieber music video song gaga album lady

Modeling Geographical Language Variations Generative Process: Details

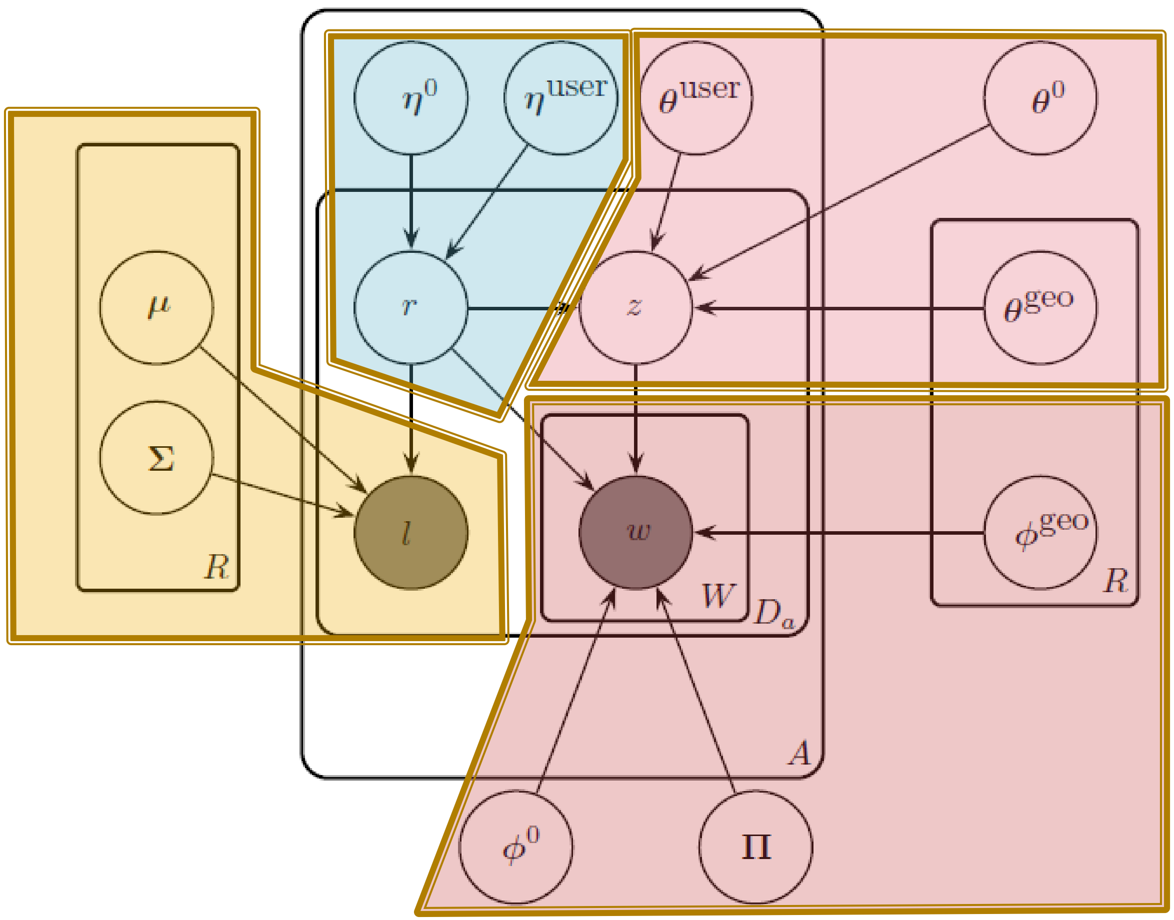
Notations

Symbol	Size	Usage
$\boldsymbol{\eta}^0$	$1 \times \mathbb{R}$	global region distribution
$\boldsymbol{\eta}^{\text{user}}$	$\mathbb{U} \times \mathbb{R}$	user-dependent region distribution
$\boldsymbol{\theta}^0$	$1 \times \mathbb{K}$	global topic distribution
$\boldsymbol{\theta}^{\text{geo}}$	$\mathbb{R} \times \mathbb{K}$	region-dependent topic distribution
$\boldsymbol{\theta}^{\text{user}}$	$\mathbb{U} \times \mathbb{K}$	user-dependent topic distribution
$\boldsymbol{\phi}^0$	$1 \times \mathbb{V}$	global term distribution
$\boldsymbol{\phi}^{\text{geo}}$	$\mathbb{R} \times \mathbb{V}$	region-dependent term distribution
$\boldsymbol{\Pi}$	$\mathbb{K} \times \mathbb{V}$	a global topic matrix
$\boldsymbol{\mu}$	$\mathbb{R}^2$	mean location of a latent region
$\boldsymbol{\Sigma}$	$\mathbb{R}^{2\times 2}$	covariance matrix of a latent region

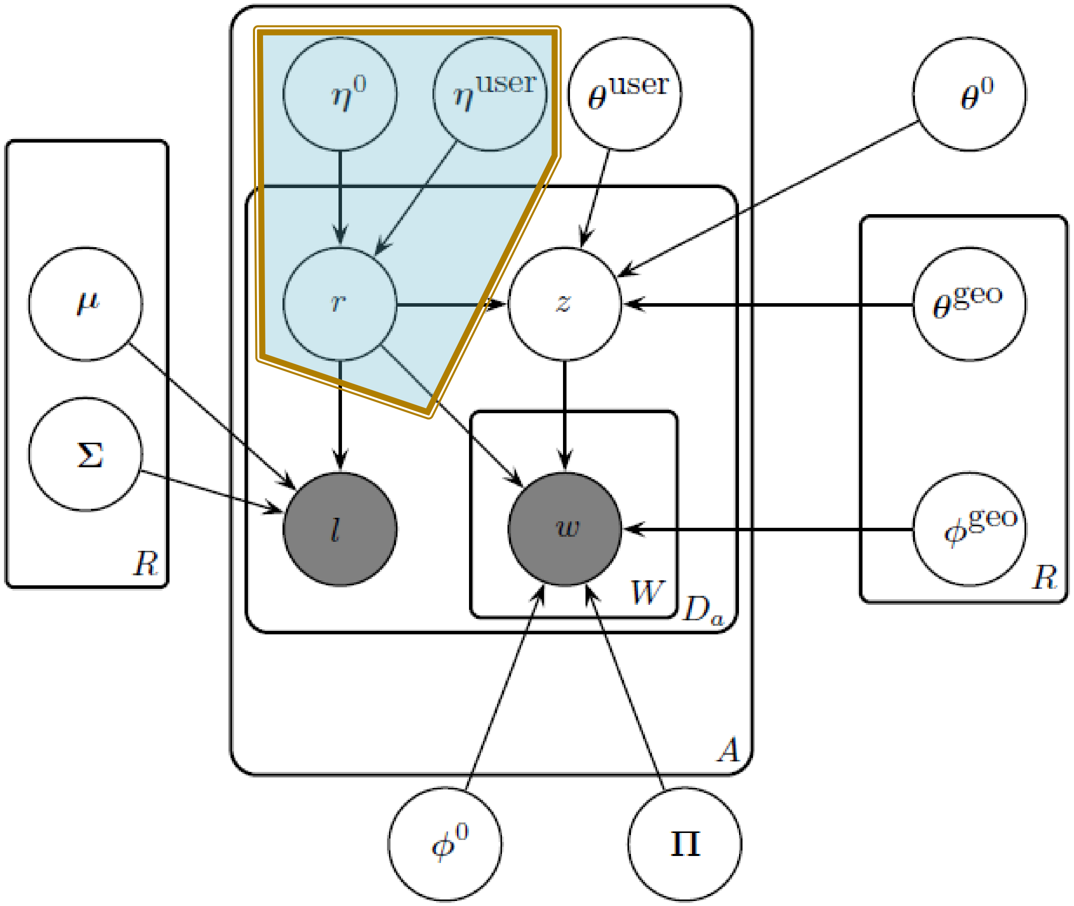
Modeling Geographical Language Variations
The Graphical Model



Modeling Geographical Language Variations
The Graphical Model



Modeling Geographical Language Variations
Region Selection



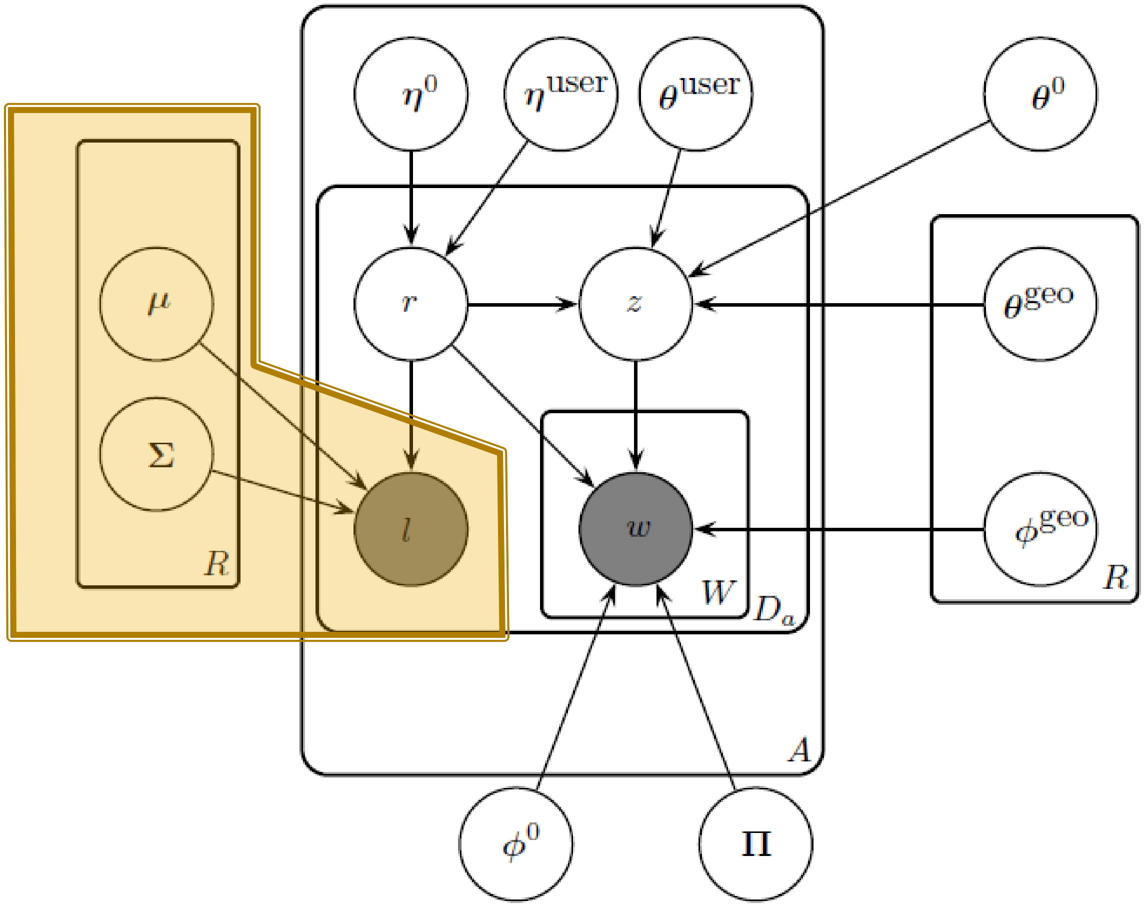
Modeling Geographical Language Variations
Region Selection

Step-by-Step

- Users tend to appear in a handful geographical locations.

$$P\left(r|\boldsymbol{\eta}^{0}, \boldsymbol{\eta}_{u}^{\text{user}}\right) = p\left(r|\boldsymbol{\eta}^{0} + \boldsymbol{\eta}_{u}^{\text{user}}\right)$$

Modeling Geographical Language Variations
Location Generation

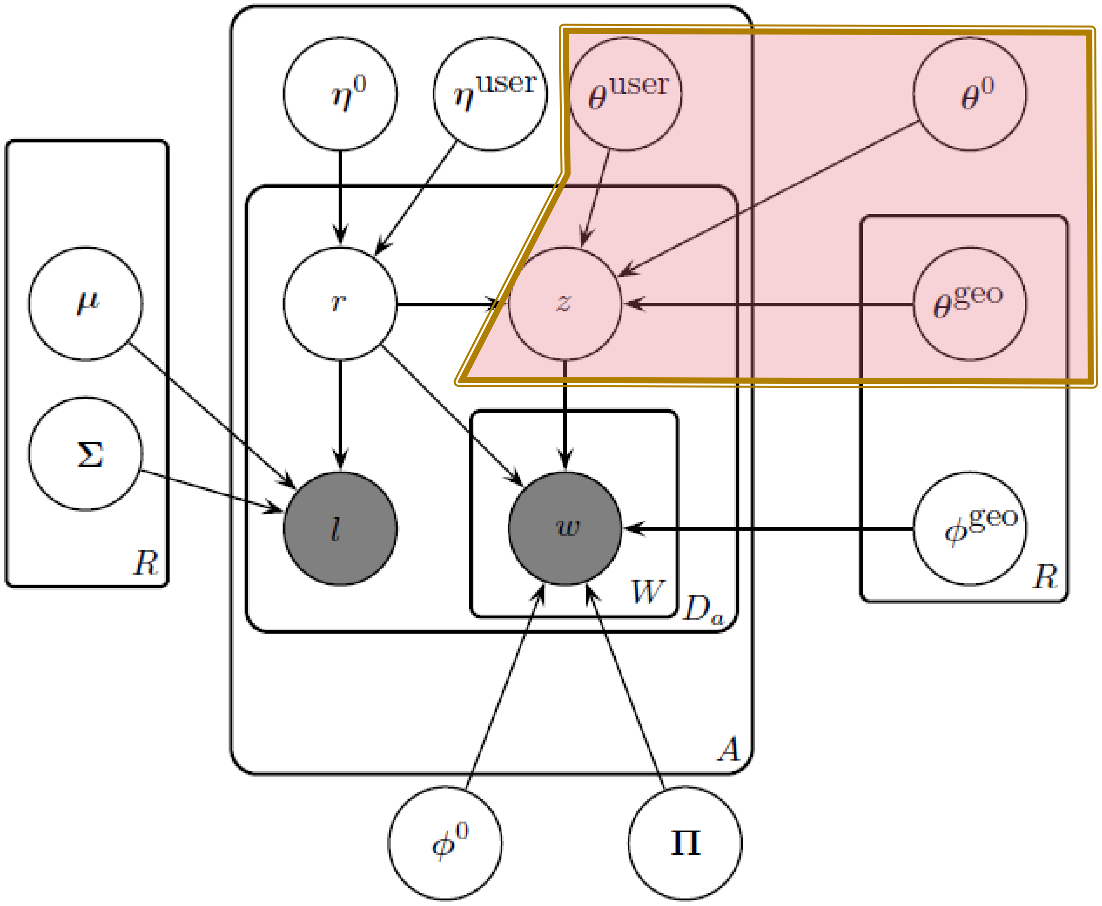


Modeling Geographical Language Variations
Location Generation

- Once a region is selected, locations can be generated.

$$\mathbf{l}_d \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r).$$

Modeling Geographical Language Variations
Topic Selection

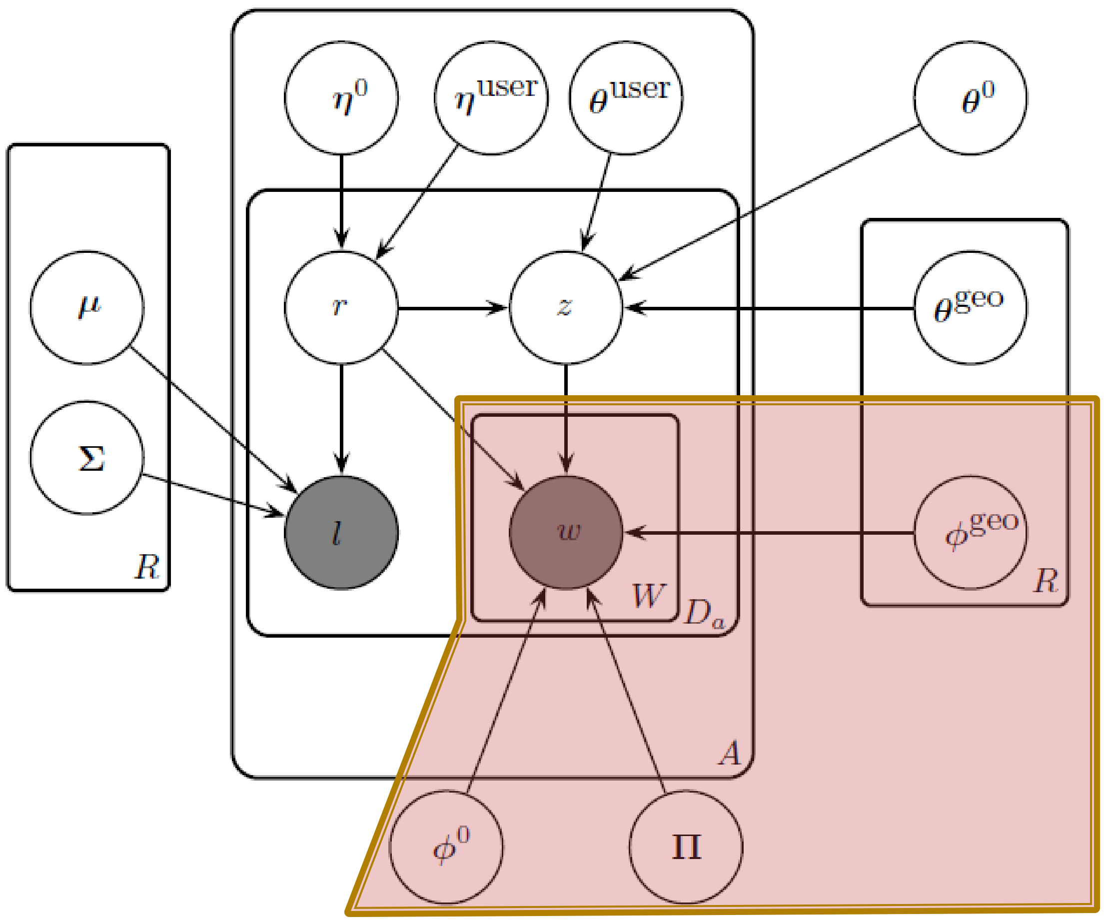


Modeling Geographical Language Variations
Topic Selection

- Topics have different chances to be discussed in different regions by different users

$$P\left(z|\theta^0, \theta_u^{\text{user}}, \theta_r^{\text{geo}}\right) = p\left(z|\theta_j^0 + \theta_{u,j}^{\text{user}} + \theta_{r,j}^{\text{geo}}\right)$$

Modeling Geographical Language Variations
Word Generation



Modeling Geographical Language Variations
Word Generation

- Words used in a tweet depend on both the location and topic of the tweet.

$$P\left(w|z, \phi^{0}, \phi_{r}^{\text{geo}}, \mathbf{\Pi}_{z}\right) = p\left(w|\phi^{0} + \phi_{r}^{\text{geo}} + \mathbf{\Pi}_{z_d}\right)$$

Modeling Geographical Language Variations
Sparse Modeling

- Laplace Priors

$$\eta_r^0 \sim \mathcal{L}(0, \omega^0) \qquad \eta_{u,r}^{\text{user}} \sim \mathcal{L}(0, \omega_u)$$

$$\theta_z^{\text{geo}} \sim \mathcal{L}(0, \lambda_l) \qquad \theta_{u,z}^{\text{user}} \sim \mathcal{L}(0, \lambda_u) \qquad \theta_{r,z}^{\text{geo}} \sim \mathcal{L}(0, \lambda_r)$$

$$\phi_v^0 \sim \mathcal{L}(0, \psi^0) \qquad \phi_{r,v}^{\text{geo}} \sim \mathcal{L}(0, \psi_l)$$

$$\Pi_{z,v} \sim \mathcal{L}(0, \psi_t)$$

Sparsity results in predictive models

Modeling Geographical Language Variations
Bayesian treatment

- Prior distributions over mean and covariance matrix
- Jeffery prior

$$\mu \sim \mathrm{Unif.}$$

$$P(\Sigma) \propto |\Sigma|^{-(3/2)}.$$

- Penalize large regions
    - We want region to be predictive as much as the data supports