

**The 49<sup>th</sup> Annual Meeting of the Association for  
Computational Linguistics: Human Language Technologies**



# **Interactive Topic Modeling**

---

**Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff  
{ynhu, bsonrisa}@cs.umd.edu, jbg@umiacs.umd.edu**

**University of Maryland**

**June 20, 2011**

# Outline

---

- Introduction of Topic Models
- Diagnosing Topic Models
- Encoding Feedback to Topic Models
- Strategies
- Experiments
- Conclusion
- Future Steps

# Outline

---

- Introduction of Topic Models
- Diagnosing Topic Models
- Encoding Feedback to Topic Models
- Strategies
- Experiments
- Conclusion
- Future Steps

# Why topic models?

---

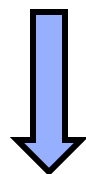
- A huge number of documents
- Want to know what's going on
- Don't have time to read



# Why topic models?

---

- A huge number of documents
- Want to know what's going on
- Don't have time to read



## **Topic Models**

- **A corpus-level view of major themes**
- **Unsupervised**

# Conceptual approach

- What topics are expressed throughout the corpus
- What topics are expressed by each document

## TOPIC 1

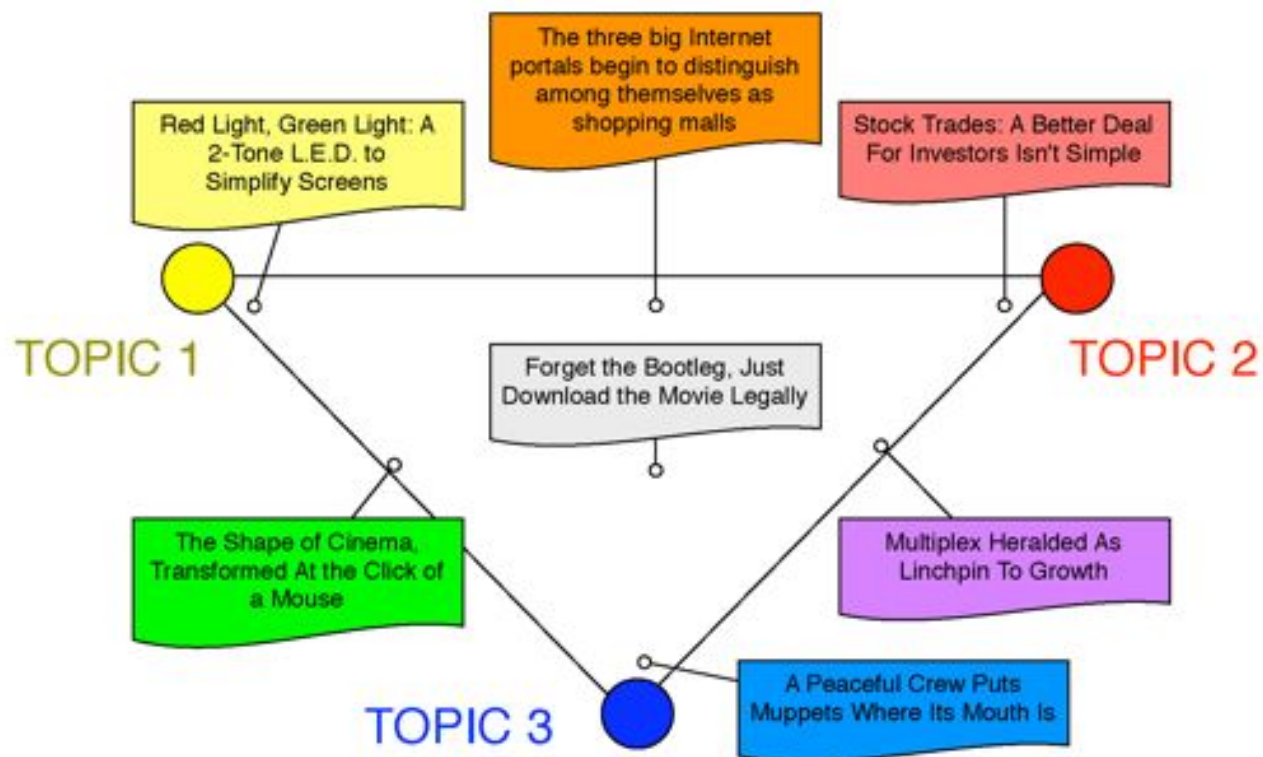
computer, site,  
technology, system,  
service, phone,  
internet, machine

## TOPIC 2

Sell, sale, market,  
product, business,  
advertising, store

## TOPIC 3

play, film, movie,  
theater, production,  
star, director, stage

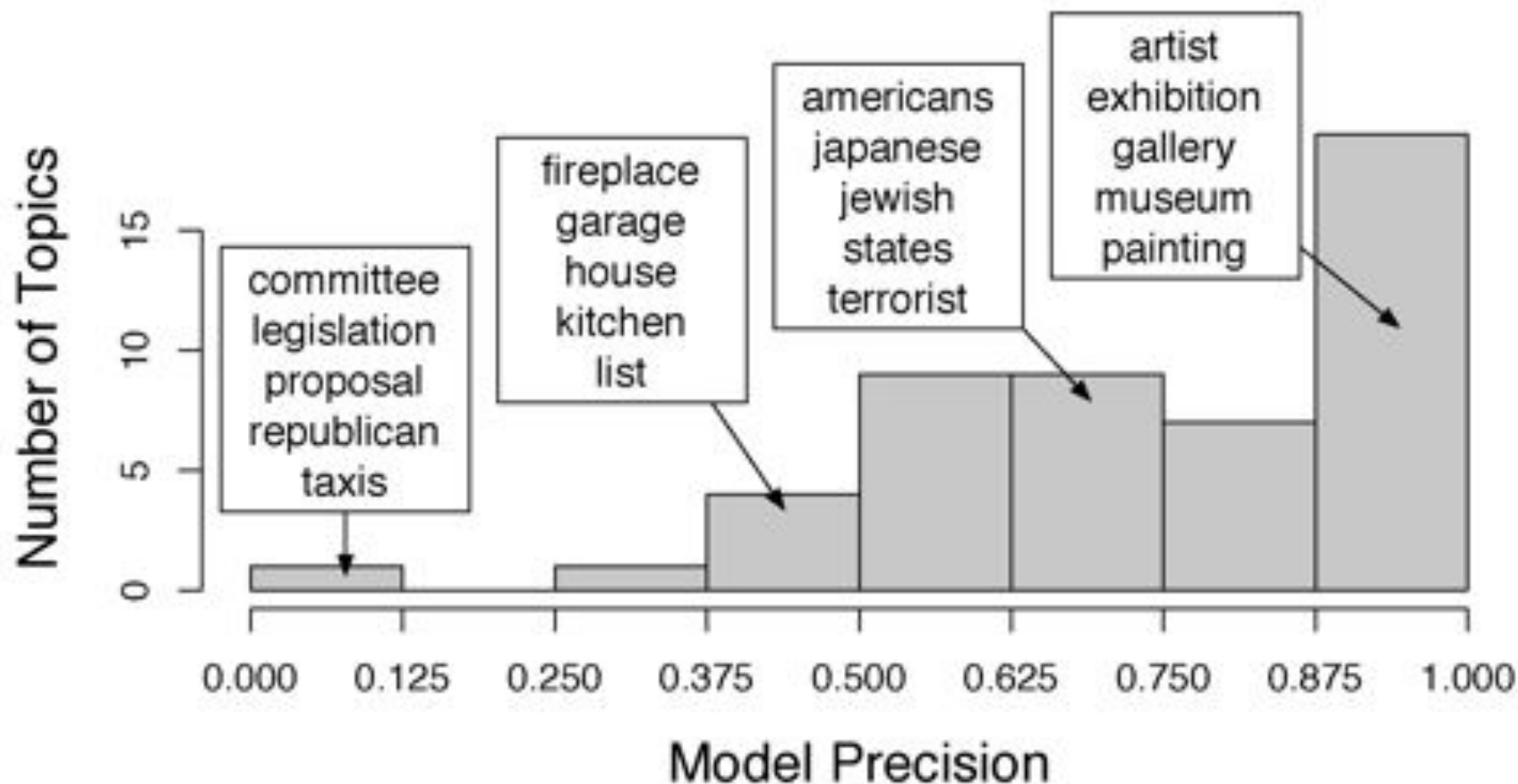


# What's Important?

---

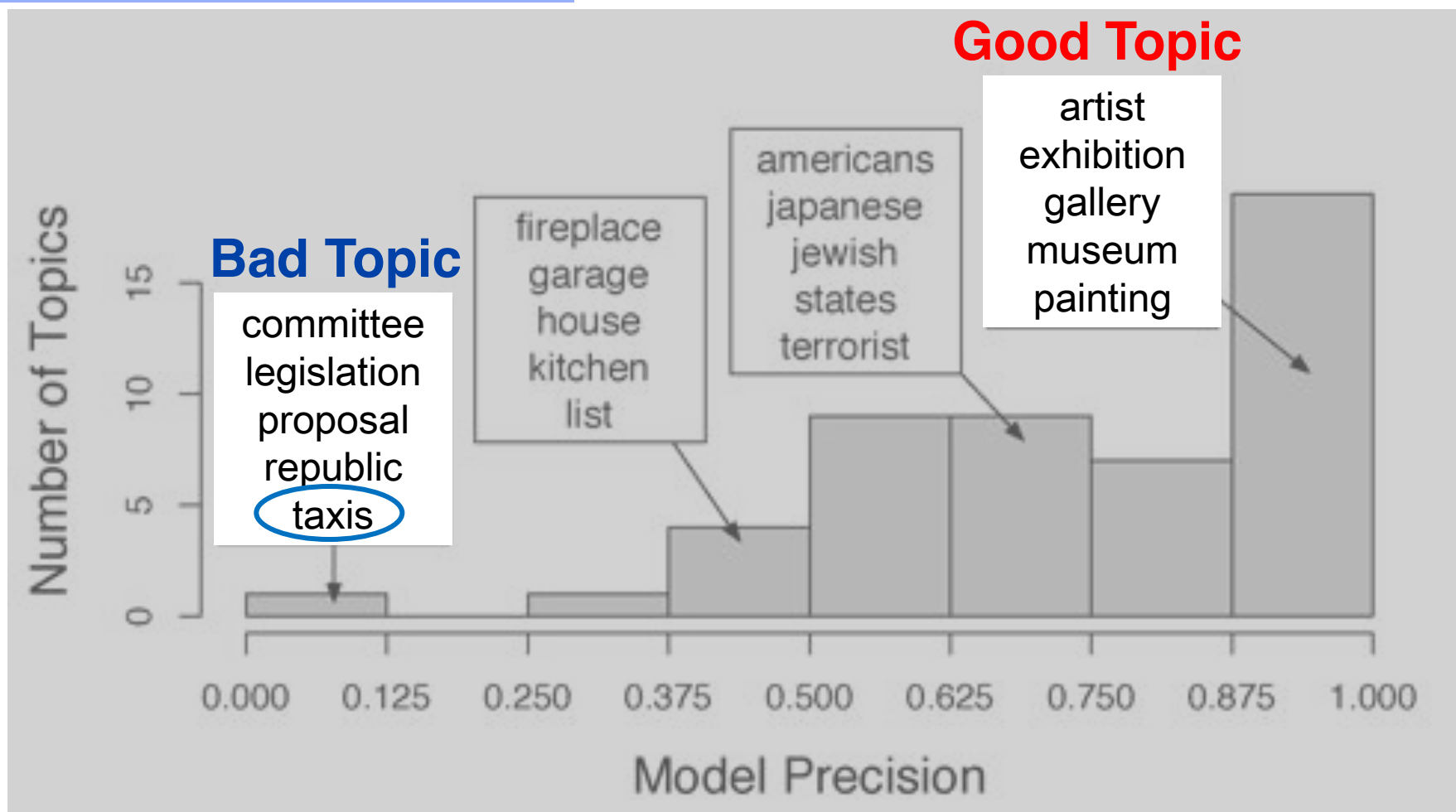
- A generative probabilistic model of documents that posits a hidden topic structure
- Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
  - A topic is a distribution over words
  - A document is a distribution over topics

# What's the problem?



- Measure topic quality (Chang et al., 2009), not all topics are good
- It is easy to be detected by humans

# What's the problem?



- Measure topic quality (Chang et al., 2009), not all topics are good
- It is easy to be detected by humans

# Outline

---

- Introduction of Topic Models
- **Diagnosing Topic Models**
- Encoding Feedback to Topic Models
- Strategies
- Experiments
- Conclusion
- Future Steps

# Diagnosing topic models

---

Topic 1	Topic 2
shuttle	NASA
launch	telescope
racket	quasar
battledore	saturn
backhand	space
astronaut	moon

# Diagnosing topic models

Topic 1	Topic 2
shuttle	NASA
launch	telescope
racket	quasar
battledore	saturn
backhand	space
astronaut	moon

shuttle, launch and NASA should be together.



# Diagnosing topic models

---

<b>Topic 3</b>
<b>bladder</b>
<b>spinal_cord</b>
<b>sci</b>
<b>spinal</b>
<b>urinary</b>
<b>urothelial</b>
<b>cervical</b>
<b>urinary_tract</b>
<b>lumbar</b>

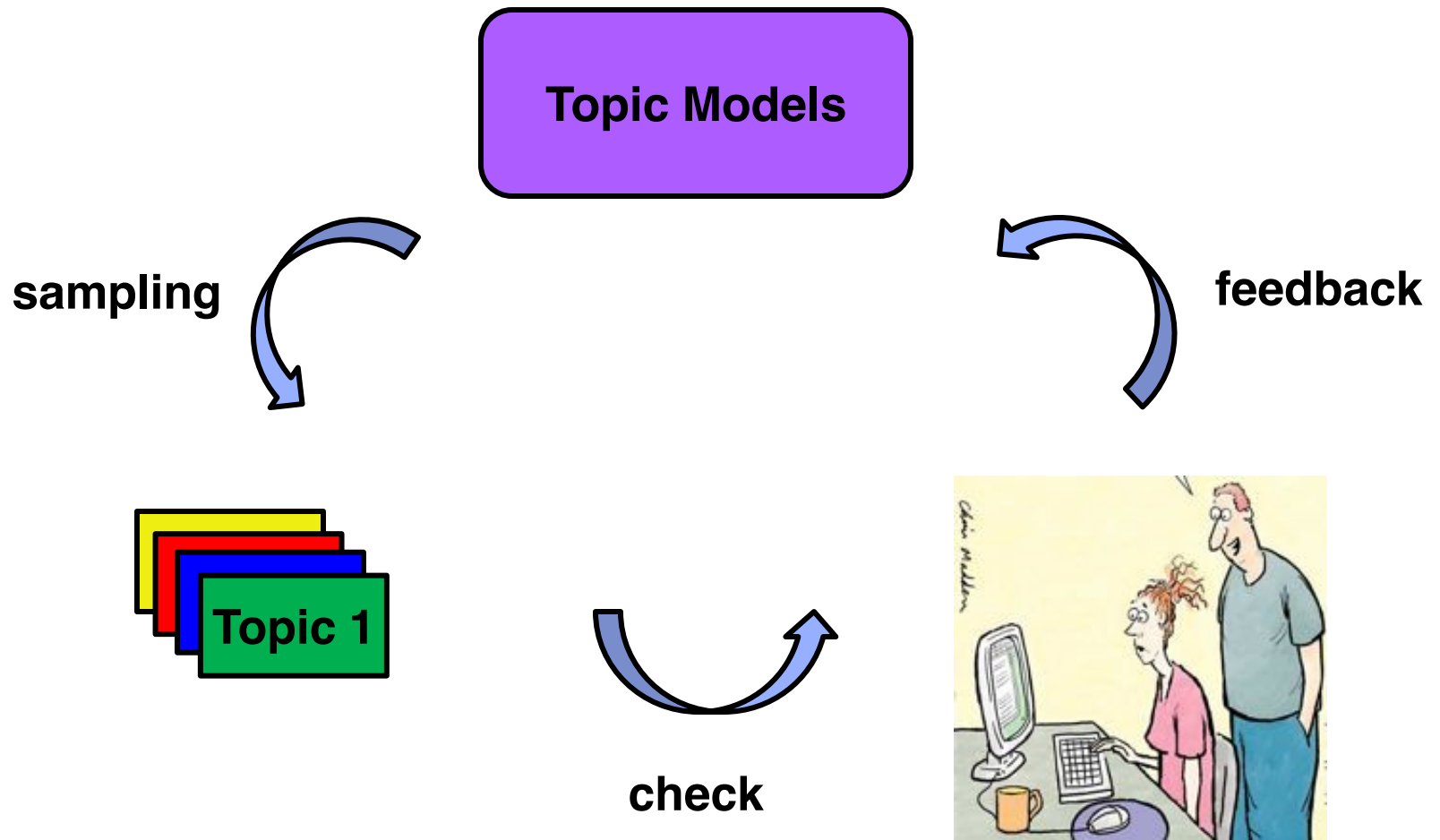
# Diagnosing topic models

Topic 3
<b>bladder</b>
spinal_cord
sci
spinal
<b>urinary</b>
<b>urothelial</b>
cervical
<b>urinary_tract</b>
lumbar

These words don't belong together!  
Should be separated.



# Simple interaction



# Outline

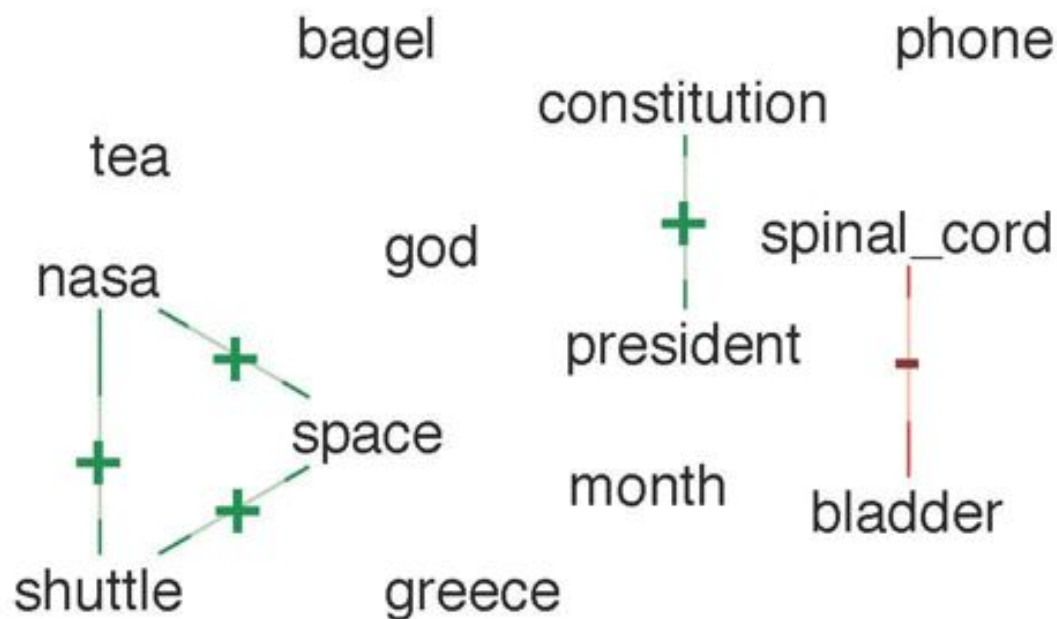
---

- Introduction of Topic Models
- Diagnosing Topic Models
- **Encoding Feedback to Topic Models**
- Strategies
- Experiments
- Conclusion
- Future Steps



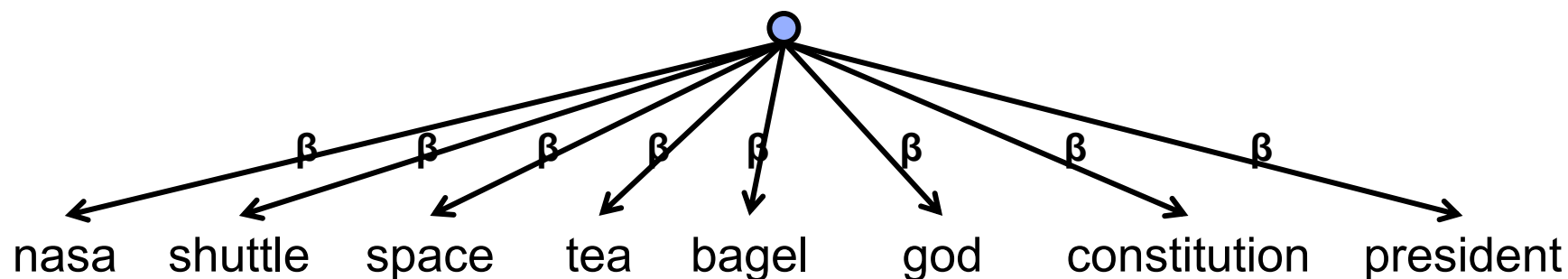
# What feedback?

- Topics are distributions over uncorrelated words
- Add Constraints: positive and negative correlations



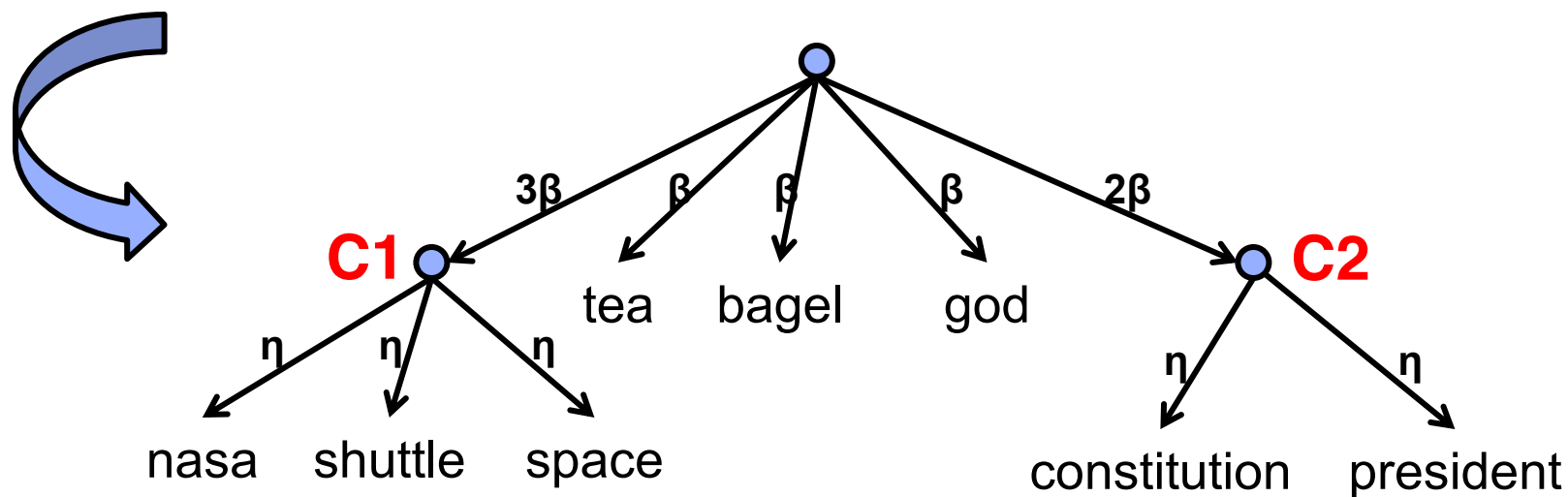
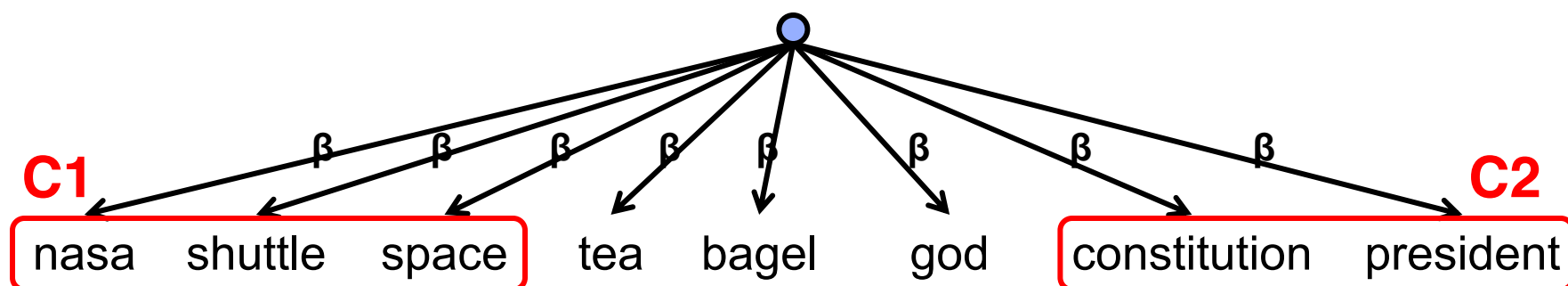
# Prior in normal LDA

- Same prior for all the words (Boyd-Graber et al., 2007)

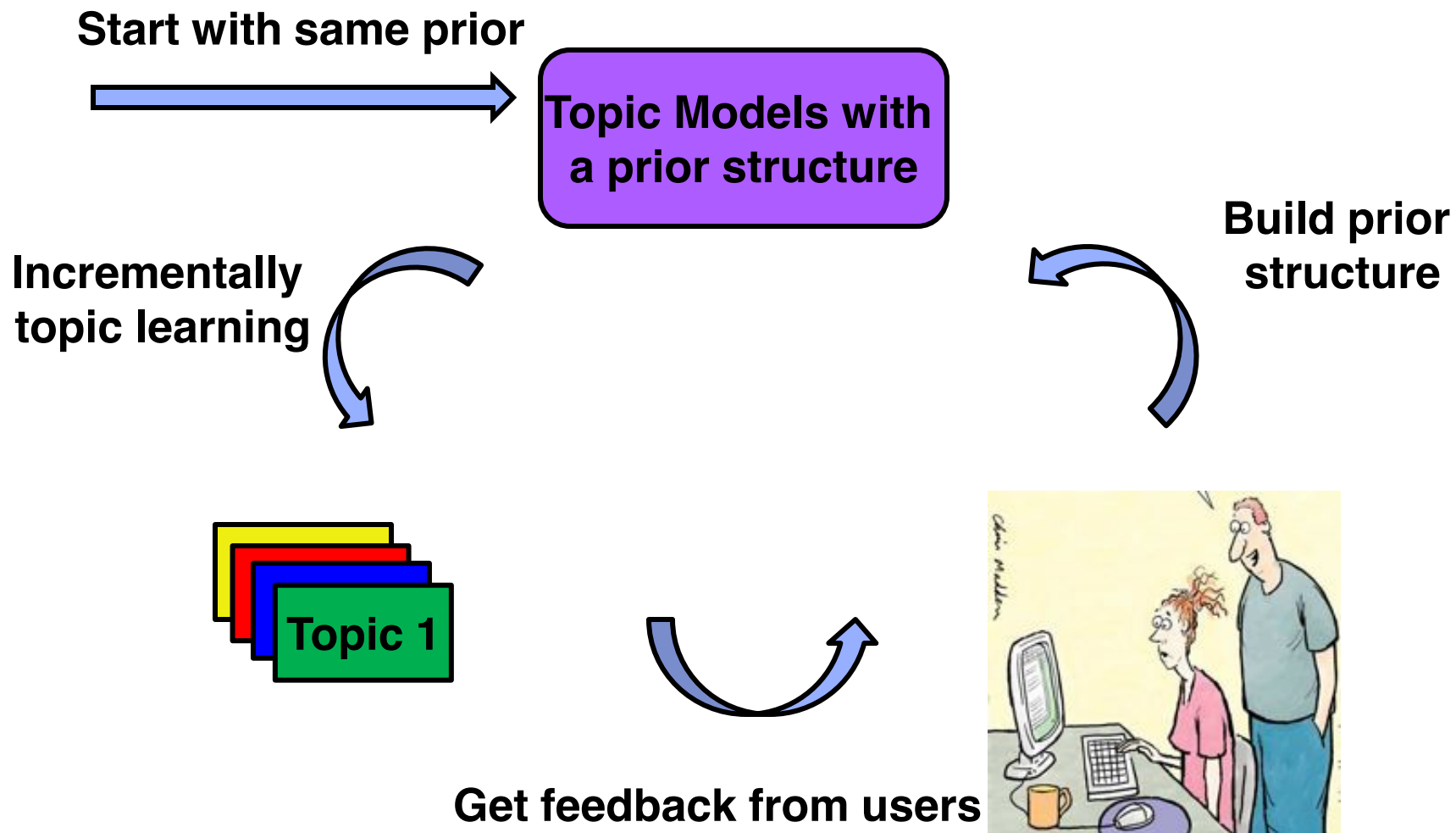


# Model constraints as prior

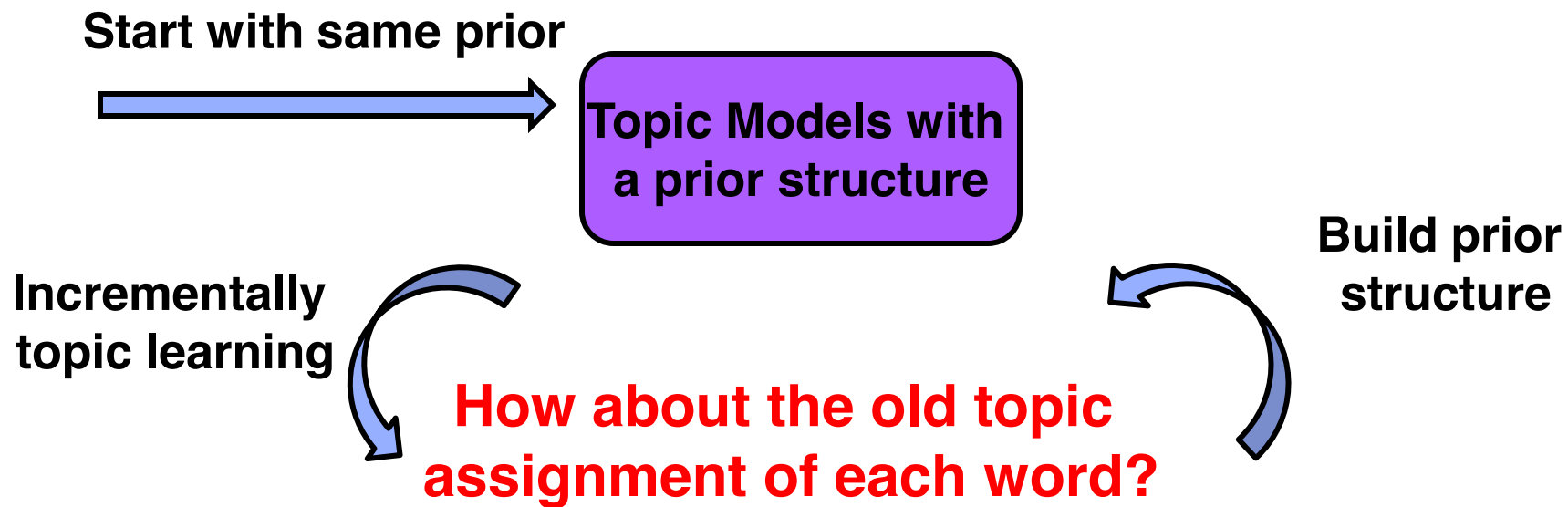
- Dirichlet Forest: prior tree structure (Andrzejewski et al. 2009)
- Positive constraints only in this paper



# How to incorporate feedback?



# How to incorporate feedback?



Get feedback from users



# Outline

---

- Introduction of Topic Models
- Diagnosing Topic Models
- Encoding Feedback to Topic Models
- **Strategies**
- Experiments
- Conclusion
- Future Steps

# Remember or forget?

---

- Four strategies
  - All
  - None
  - Doc
  - Term
- Toy example

# Toy example

Doc 1

nasa shuttle launch ...

Doc 2

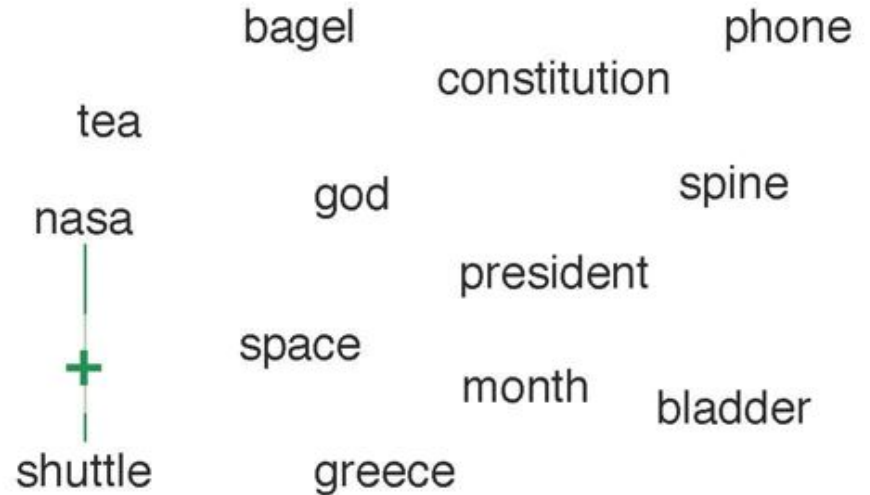
racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...



# Toy example: All

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea god spine  
nasa president  
space month bladder  
shuttle greece

## Strategy All

- Forget all topic assignments
- Start from the very beginning

# Toy example: None

Doc 1

nasa shuttle launch ...

Doc 2

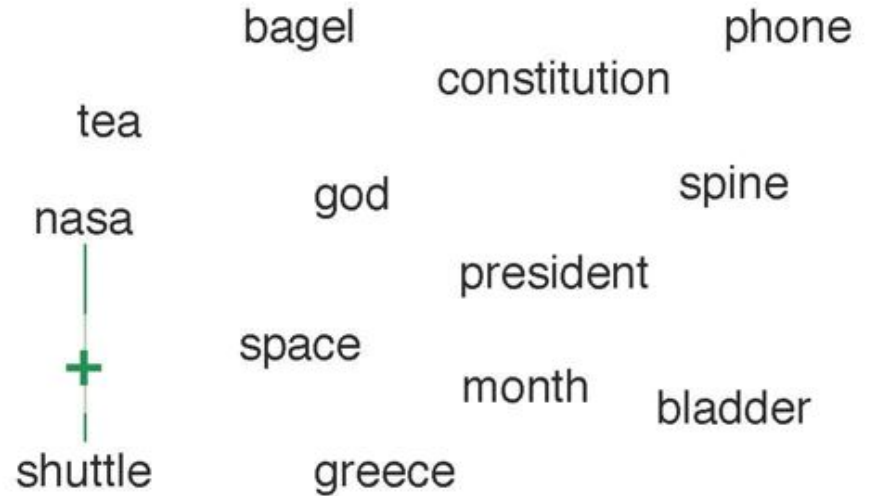
racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...



## Strategy None

- Remember everything
- Continue

# Toy example: Doc

Doc 1

nasa shuttle launch ...

Doc 2

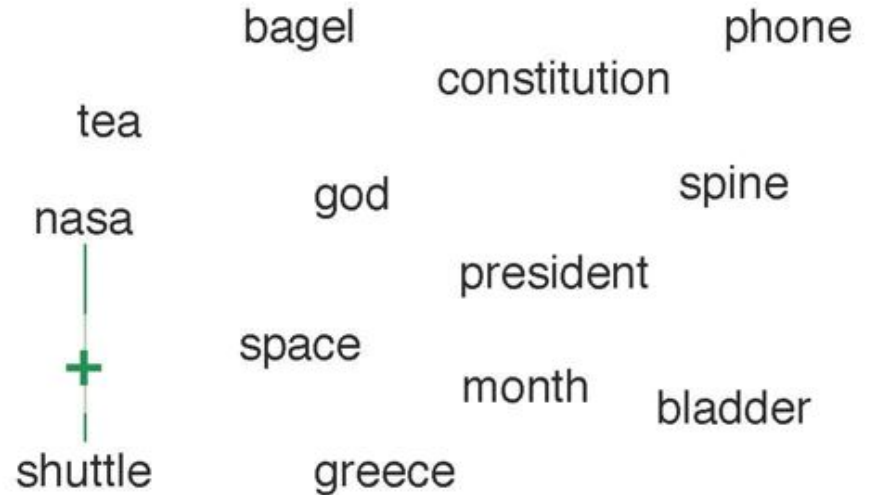
racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...



## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

# Toy example: Doc

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea god spine  
nasa president  
space month bladder  
shuttle greece

## Strategy Doc

- Forget the topic assignments for docs containing constraints
- Remember the others
- continue

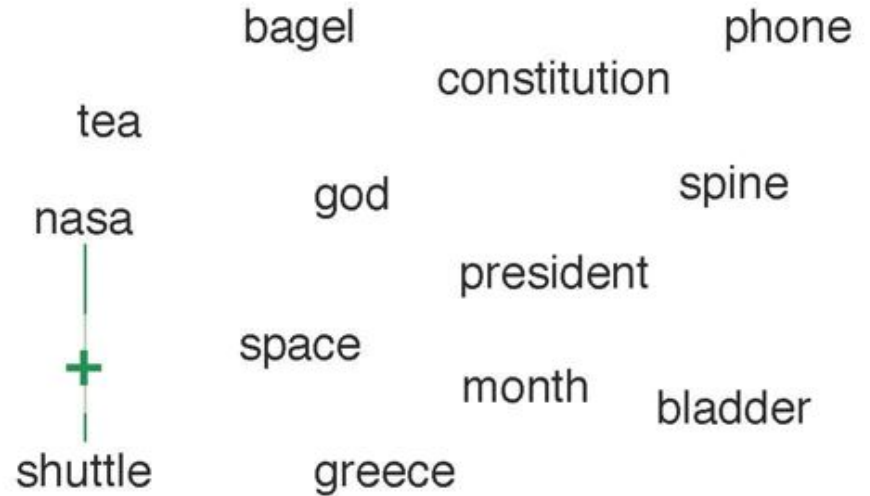
# Toy example: Doc

Doc 1  
nasa shuttle launch ...

Doc 2  
racket serve shuttle ...

Doc 3  
bladder pain bladder ...

Doc 4  
spine pain lumbar ...



## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

# Toy example: Doc

Doc 1  
nasa shuttle launch ...

Doc 2  
racket serve shuttle ...

Doc 3  
bladder pain bladder ...

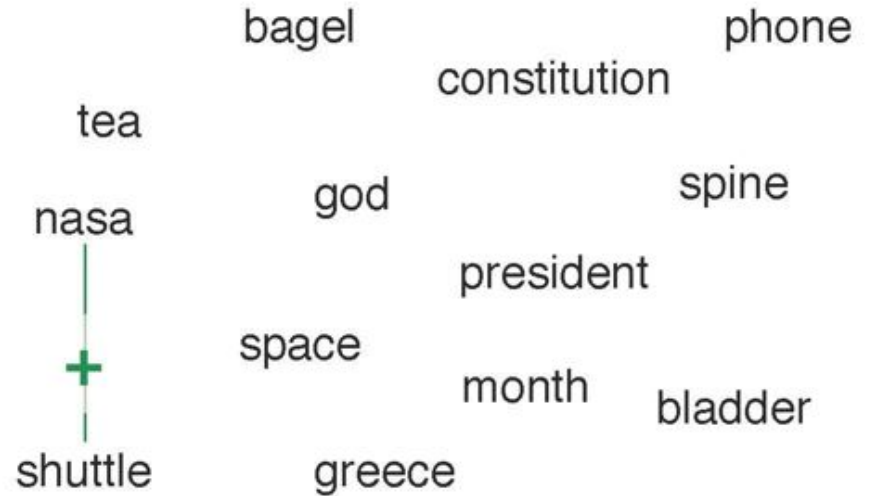
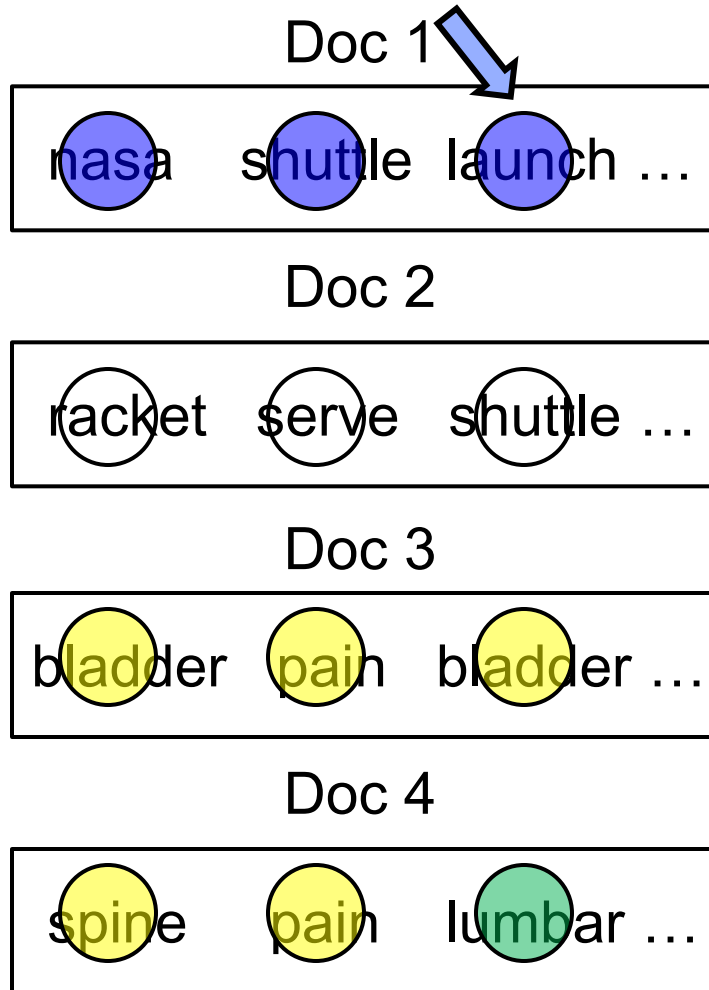
Doc 4  
spine pain lumbar ...

bagel constitution phone  
tea god spine  
nasa president  
shuttle + space month bladder  
greece

## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

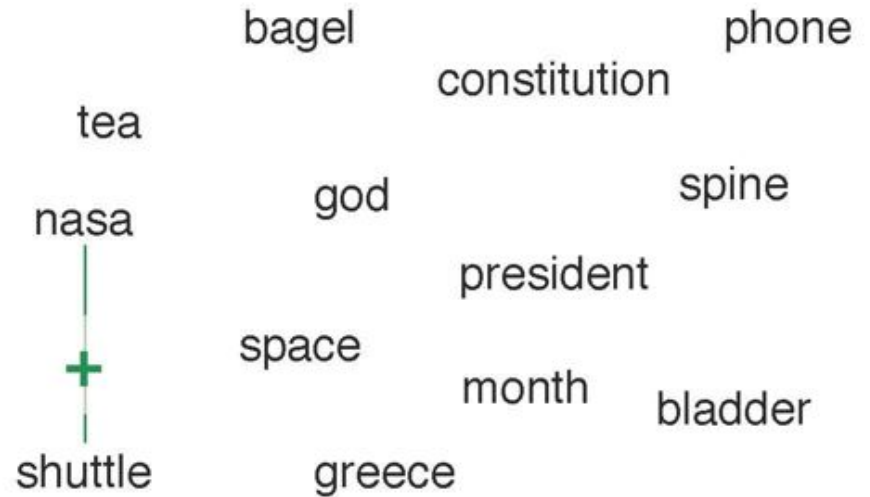
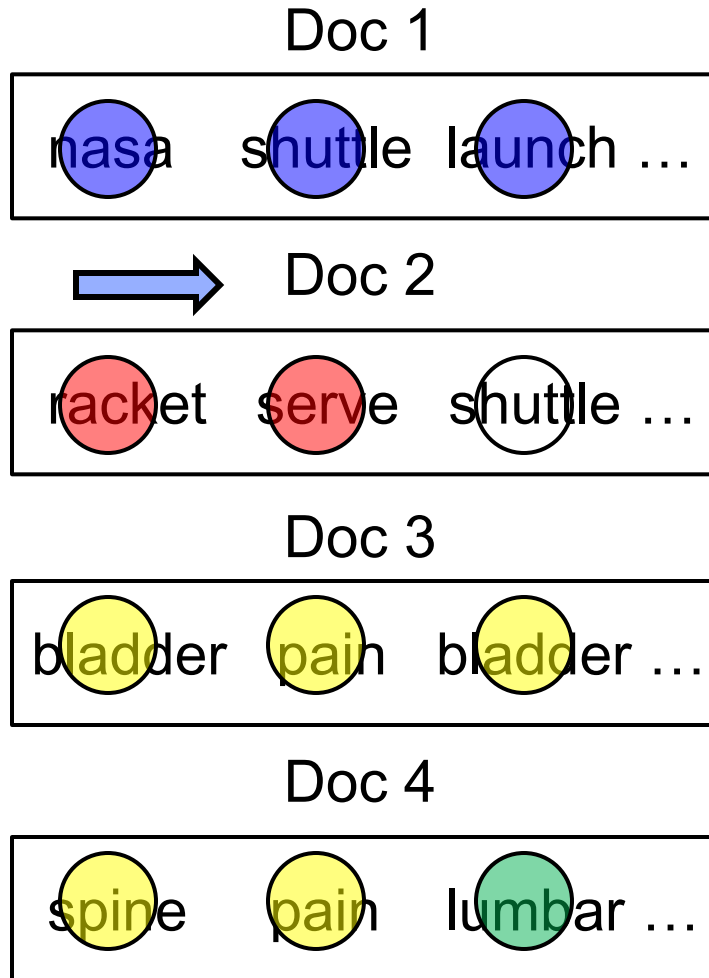
# Toy example: Doc



## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

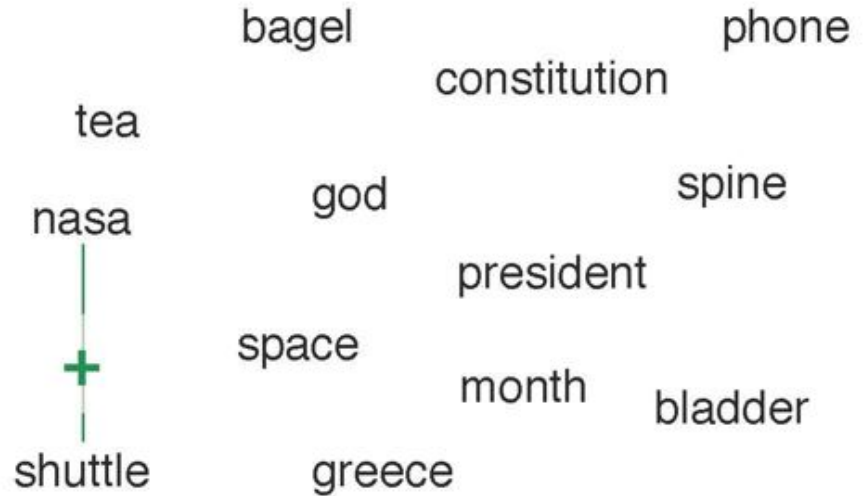
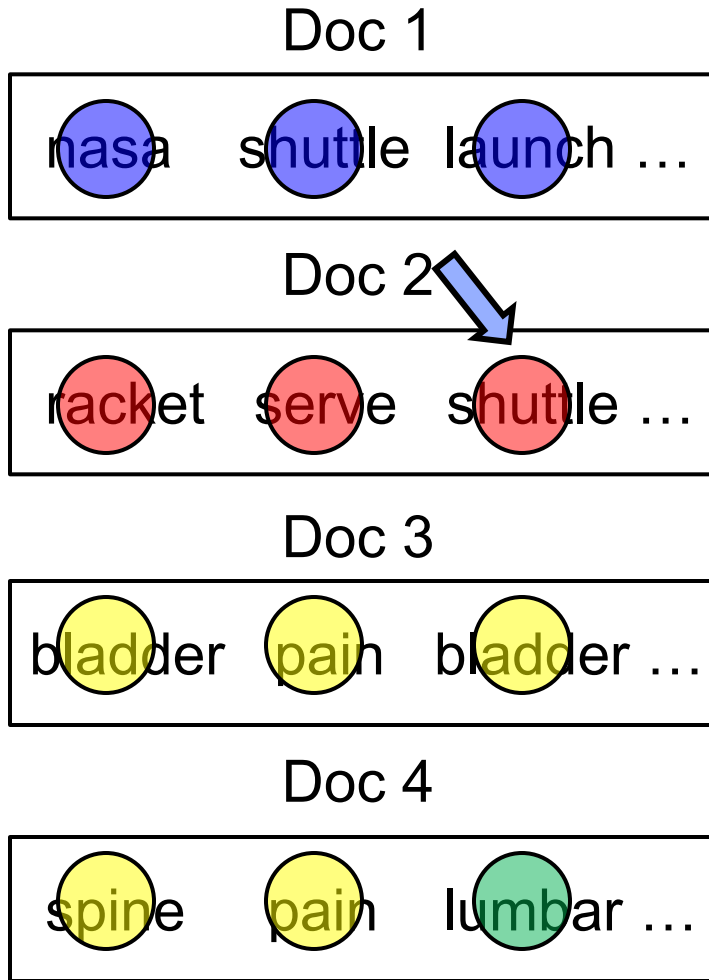
# Toy example: Doc



## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

# Toy example: Doc



## Round 1

- Positive constr: (nasa - shuttle)
- Strategy: Doc

# Toy example: Doc

Doc 1  
nasa shuttle launch ...

Doc 2  
racket serve shuttle ...

Doc 3  
bladder pain bladder ...

Doc 4  
spine pain lumbar ...



Doc 1  
nasa shuttle launch ...

Doc 2  
racket serve shuttle ...

Doc 3  
bladder pain bladder ...

Doc 4  
spine pain lumbar ...

# Toy example: Term

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea nasa god president spine  
space month bladder  
shuttle greece

## Round 2

- Negative constr: (spine - bladder)
- Strategy: Term

# Toy example: Term

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea nasa god president spine  
space month bladder  
shuttle greece

## Strategy Term

- Forget the topic assignments for the constraint words,
- Remember the others
- Continue

# Toy example: Term

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea god  
nasa god president spine  
space president  
month bladder  
shuttle greece

## Round 2

- Negative constr: (spine - bladder)
- Strategy: Term

# Toy example: Term

Doc 1

nasa shuttle launch ...

Doc 2

racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

spine pain lumbar ...

bagel constitution phone  
tea god  
nasa god president spine  
space president month bladder  
shuttle greece

## Round 2

- Negative constr: (spine - bladder)
- Strategy: Term

# Toy example: Term

Doc 1

nasa shuttle launch ...

Doc 2

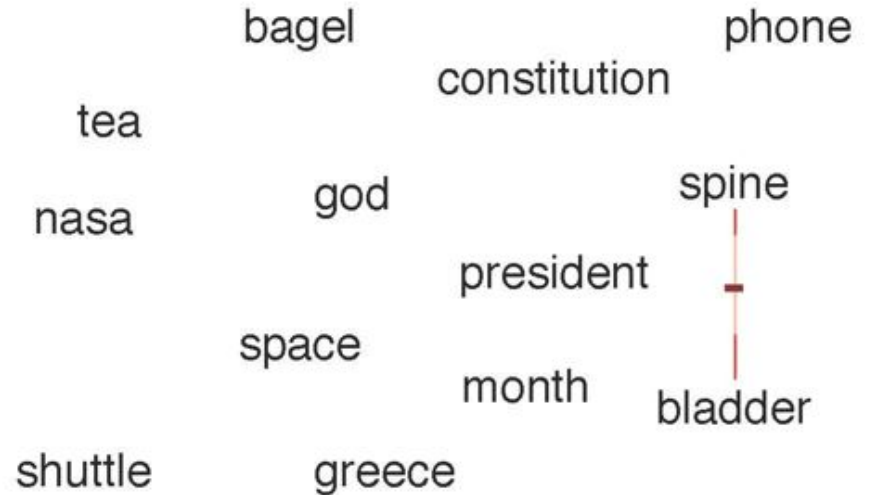
racket serve shuttle ...

Doc 3

bladder pain bladder ...

Doc 4

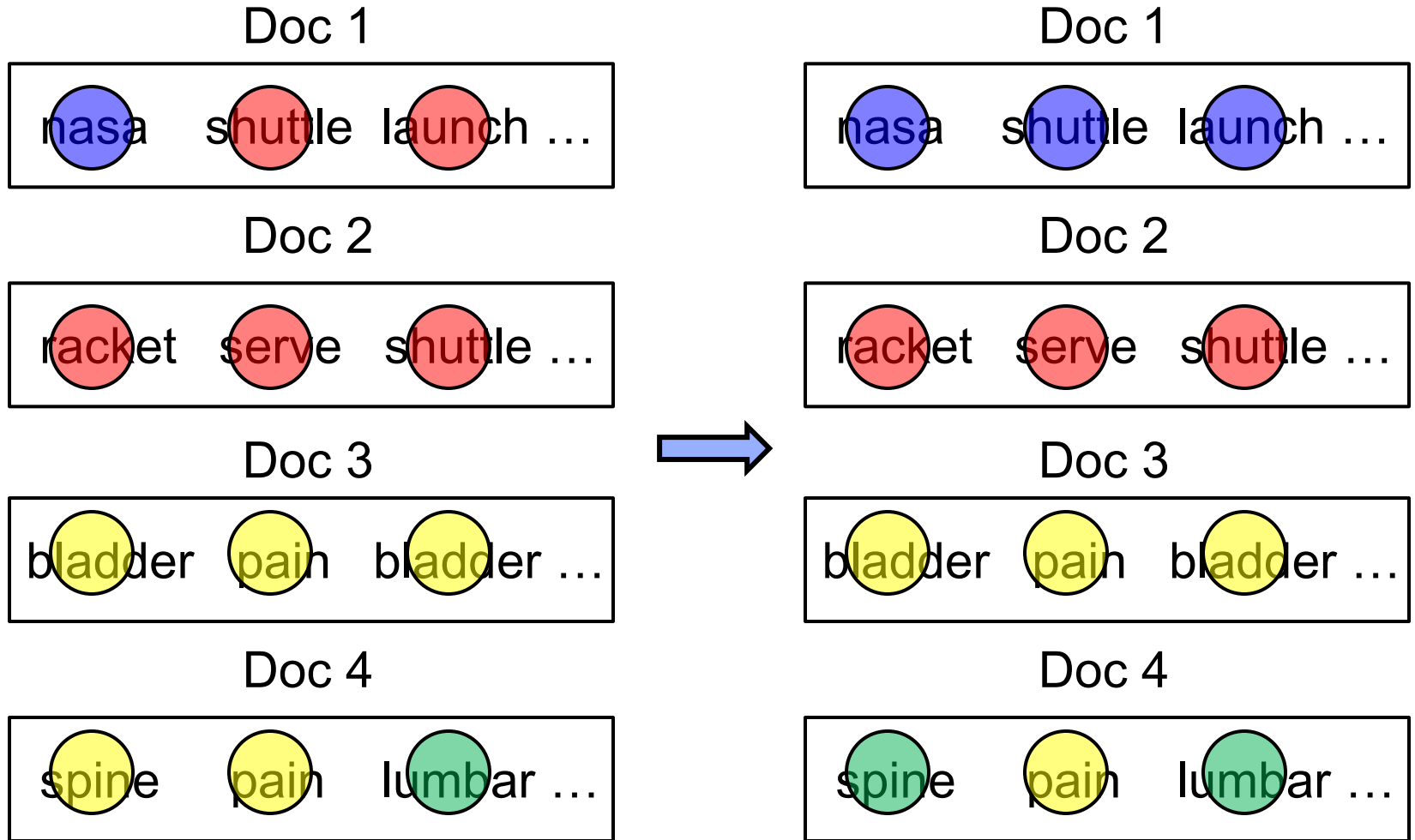
spine pain lumbar ...



## Round 2

- Negative constr: (spine - bladder)
- Strategy: Term

# Toy example



# Outline

---

- Introduction of Topic Models
- Diagnosing Topic Models
- Encoding Feedback to Topic Models
- Strategies
- **Experiments**
- Conclusion
- Future Steps

# Motivating example

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, David
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
...	...
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

# Motivating example

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, military, democracy, boris, country, south, years, month, government, vote, since, leader, presidential
...	...
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, ashington, western, bring, party

# Motivating example

Topic	Before
1	election, <b>yeltsin</b> , <b>russian</b> , political, party, democratic, <b>russia</b> , president, military, democracy, <b>boris</b> , country, south, years, month, government, vote, since, leader, presidential
...	...
20	<b>soviet</b> , lead, <b>gorbachev</b> , <b>union</b> , west, <b>mikhail</b> , reform, change, europe, leaders, poland, <b>communist</b> , know, old, right, human, ashington, western, bring, party

## Suggested constraint

**boris, communist, gorbachev, mikhail, russia, russian, soviet, union, yeltsin**

# Motivating example

Topic	Before	Topic	After
1	election, <b>yeltsin</b> , <b>russian</b> , political, party, democratic, <b>russia</b> , president, military, democracy, <b>boris</b> , country, south, years, month, government, vote, since, leader, presidential	1	election, democratic, south, country, president, party, <b>africa</b> , <b>lead</b> , <b>even</b> , democracy, leader, presidential, <b>week</b> , <b>politics</b> , <b>minister</b> , <b>percent</b> , <b>voter</b> , <b>last</b> , month, years
...	...	...	...
20	<b>soviet</b> , lead, <b>gorbachev</b> , <b>union</b> , west, <b>mikhail</b> , reform, change, europe, leaders, poland, <b>communist</b> , know, old, right, human, ashington, western, bring, party	20	<b>soviet</b> , <b>union</b> , economic, reform, <b>yeltsin</b> , <b>russian</b> , lead, <b>russia</b> , <b>gorbachev</b> , leaders, west, president, <b>boris</b> , <b>moscow</b> , europe, poland, <b>mikhail</b> , <b>relations</b> , <b>communist</b> , power

# Motivating example

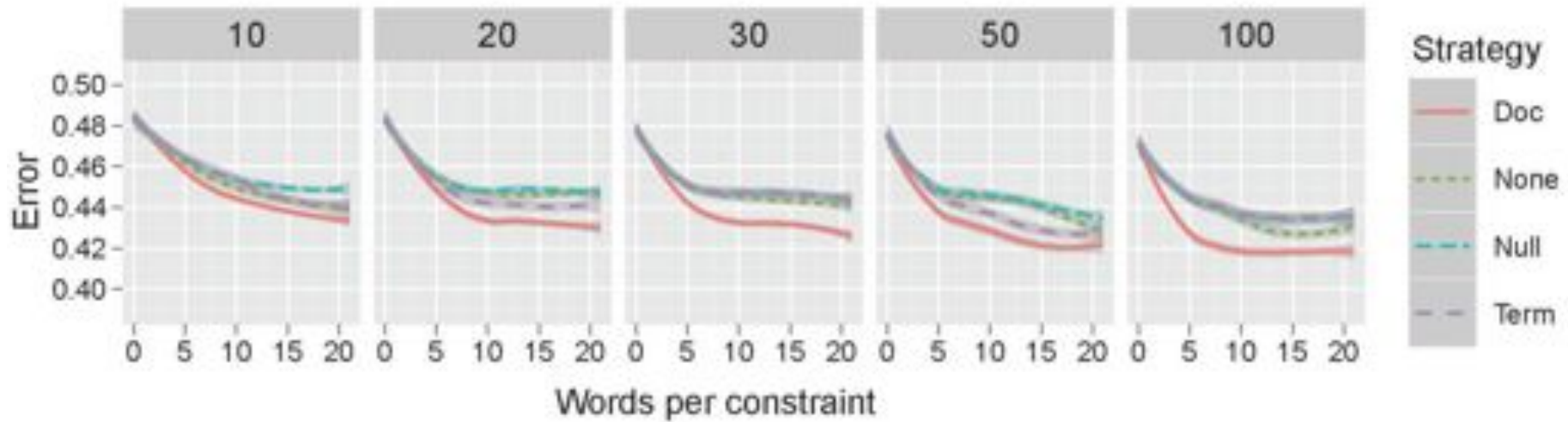
Topic	Before	Topic	After
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, David, rudolph, dinkins, lead, need, governor, legislature, pataki	2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, <b>dinkins</b> , legislature, plan, david, governor, <b>pataki, need, cut</b>
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, would, control, korea, intelligence, test, nation, country, testing	3	nuclear, arms, weapon, treaty, defense, <b>war</b> , missile, <b>may, come</b> , test, <b>american</b> , world, would, <b>need</b> , lead, <b>get, join</b> , yet, <b>clinton, nation</b>
4	president, bush, military, see, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international,	4	president, administration, bush, clinton, war, unite, force, reagan, american, america, <b>make, nation</b> , military, iraq, iraqi, <b>troops</b> , international, country, <b>yesterday, plan</b>

# Simulating an interactive user

---

- Dataset: 20 News groups
- Constraints from feature selection on training data
  - soc.religion.christian: “catholic, scripture, resurrection, pope, sabbath, spiritual, pray, divine, doctrine”
  - 20 classes: 20 constraint sets, 21 words per constraint set
- Add them to the topic model as positive constraints
  - Add one word per class each time, 21 rounds in total
- Train classifier on training data
  - Use topic distribution of each doc as the feature
- Measure classification error rate of test data

# Which strategy & how long to wait?



- Facet: number of iterations added per round
- Start with 100 iterations
- Null: no constraints, comparable iters
- “Doc” is best, run 30 or 50 iterations each round

# Put humans in the loop

Topic 41 [general](#) [attorney](#) [street](#) [like](#) [one](#) [know](#) [lead](#) [people](#) [something](#)  
[richard](#) [christmas](#) [sunday](#) [white](#) [wall](#) [get](#) [wear](#) [tree](#) [wrong](#)  
[look](#) [reporter](#)

Select words:

Topic 41	<a href="#">general</a> <a href="#">attorney</a> <a href="#">street</a> <a href="#">like</a> <a href="#">one</a> <a href="#">know</a> <a href="#">lead</a> <a href="#">people</a> <a href="#">something</a> <a href="#">richard</a> <a href="#">christmas</a> <a href="#">sunday</a> <a href="#">white</a> <a href="#">wall</a> <a href="#">get</a> <a href="#">wear</a> <a href="#">tree</a> <a href="#">wrong</a> <a href="#">look</a> <a href="#">reporter</a>
Topic 42	<a href="#">million</a> <a href="#">year</a> <a href="#">much</a> <a href="#">lead</a> <a href="#">years</a> <a href="#">spend</a> <a href="#">money</a> <a href="#">last</a> <a href="#">billion</a> <a href="#">project</a> <a href="#">help</a> <a href="#">space</a> <a href="#">welcome</a> <a href="#">america</a> <a href="#">cost</a> <a href="#">real</a> <a href="#">nearly</a> <a href="#">dollar</a> <a href="#">taxpayer</a> <a href="#">good</a>
Topic 43	<a href="#">soviet</a> <a href="#">russian</a> <a href="#">yeltsin</a> <a href="#">seem</a> <a href="#">union</a> <a href="#">reform</a> <a href="#">boris</a> <a href="#">president</a> <a href="#">russia</a> <a href="#">gorbachev</a> <a href="#">mikhail</a> <a href="#">europe</a> <a href="#">democracy</a> <a href="#">west</a> <a href="#">communist</a> <a href="#">moscow</a> <a href="#">party</a> <a href="#">week</a> <a href="#">change</a> <a href="#">years</a>
Topic 44	<a href="#">years</a> <a href="#">bring</a> <a href="#">history</a> <a href="#">reason</a> <a href="#">american</a> <a href="#">struggle</a> <a href="#">south</a> <a href="#">ago</a> <a href="#">month</a> <a href="#">recently</a> <a href="#">korea</a> <a href="#">seems</a> <a href="#">chile</a> <a href="#">times</a> <a href="#">order</a> <a href="#">military</a> <a href="#">north</a> <a href="#">position</a> <a href="#">secret</a> <a href="#">michael</a>
Topic 45	<a href="#">also</a> <a href="#">surprise</a> <a href="#">head</a> <a href="#">proposal</a> <a href="#">bond</a> <a href="#">want</a> <a href="#">bear</a> <a href="#">right</a> <a href="#">cause</a> <a href="#">lead</a> <a href="#">negotiation</a> <a href="#">strong</a> <a href="#">people</a> <a href="#">sound</a> <a href="#">edward</a> <a href="#">else</a> <a href="#">tamarkin</a> <a href="#">pressure</a> <a href="#">would</a> <a href="#">produce</a>
Topic 46	<a href="#">jail</a> <a href="#">paper</a> <a href="#">big</a> <a href="#">threaten</a> <a href="#">job</a> <a href="#">running</a> <a href="#">like</a> <a href="#">woman</a> <a href="#">black</a> <a href="#">manhattan</a> <a href="#">experience</a> <a href="#">throw</a> <a href="#">white</a> <a href="#">large</a> <a href="#">citizen</a> <a href="#">huge</a> <a href="#">cover</a> <a href="#">sell</a> <a href="#">hospital</a> <a href="#">shadow</a>

Currently selected words (click to deselect):  
attorney general jail

Existing links:  
(None)

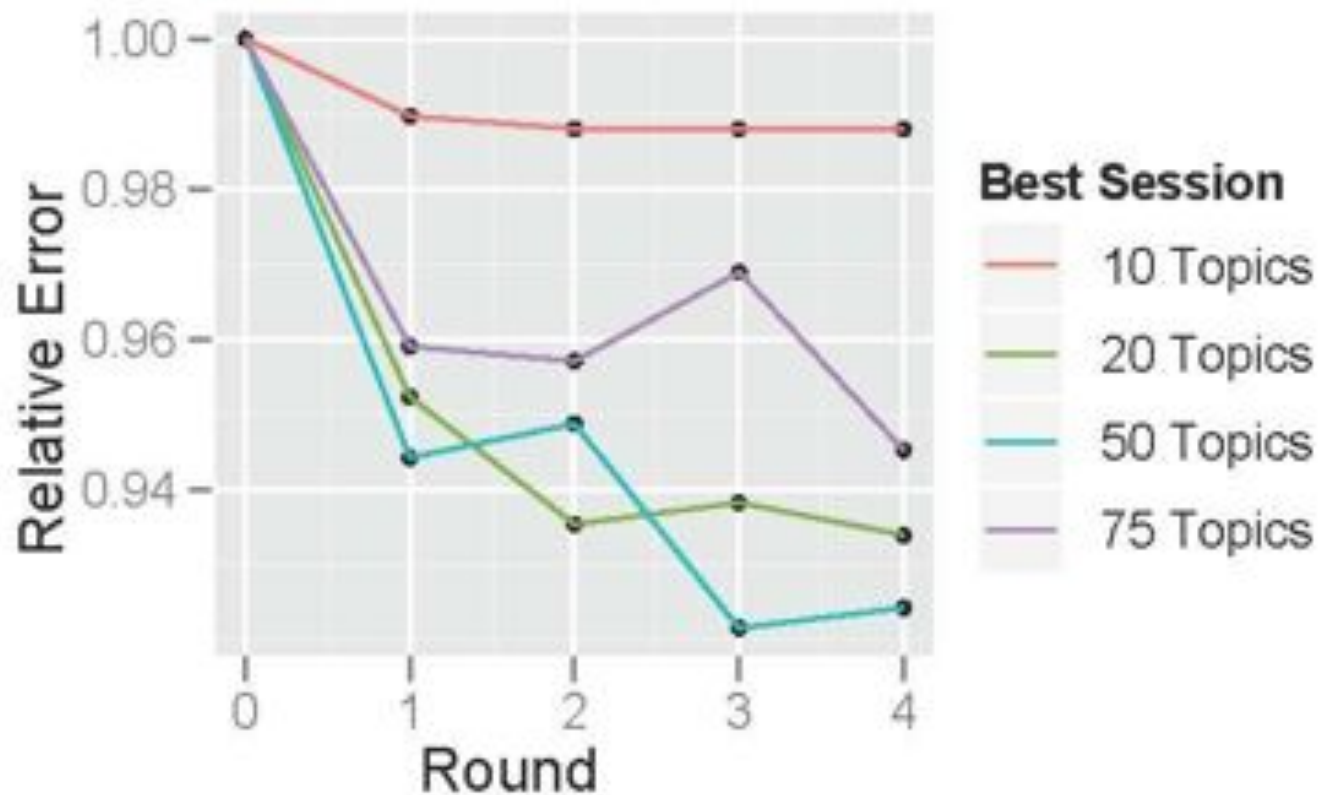
Currently selected words (click to deselect):

[attorney](#) [general](#) [jail](#)

Existing links:

(None)

# Put humans in the loop



# Put humans in the loop

---

- Some constraints users created
  - Inscrutable
    - better, people, right, take, things
    - fbi, let, says
  - Collocations
    - jesus, christ
    - solar, sun
    - even, number
    - book, list
  - Common instances (e.g. first names)
  - Soft constraint: mac, windows

# Negative constraints

- NIH data(700 topics)
- Negative constraint: bladder – spinal\_cord

Topic	Before	Topic	After
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial, injury, motor, recovery, reflex, cervical, urothelium, functional_recovery	318	sci, spinal_cord, spinal_cord_injury, spinal, injury, recovery, motor, reflex, urothelial, injured, functional_recovery, plasticity, locomotor, cervical, locomotion

# Conclusion

---

- An efficient way to refine and improve the topics discovered by topic models
- A paradigm for non-specialist consumers to refine models to better reflect their interests and needs
- Creating tools to do so
- We need users!

# Future steps

---

- Speed up
- Suggesting constraints
- Incorporating other domain knowledge
- Incorporating interaction to other models

# The 49<sup>th</sup> Annual Meeting of the Association for Computational Linguistics: Human Language Technologies



## Thank you! Any questions?

---

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff  
{ynhu, bsonrisa}@cs.umd.edu, jbg@umiacs.umd.edu

University of Maryland

June 20, 2011

# Constrained LDA

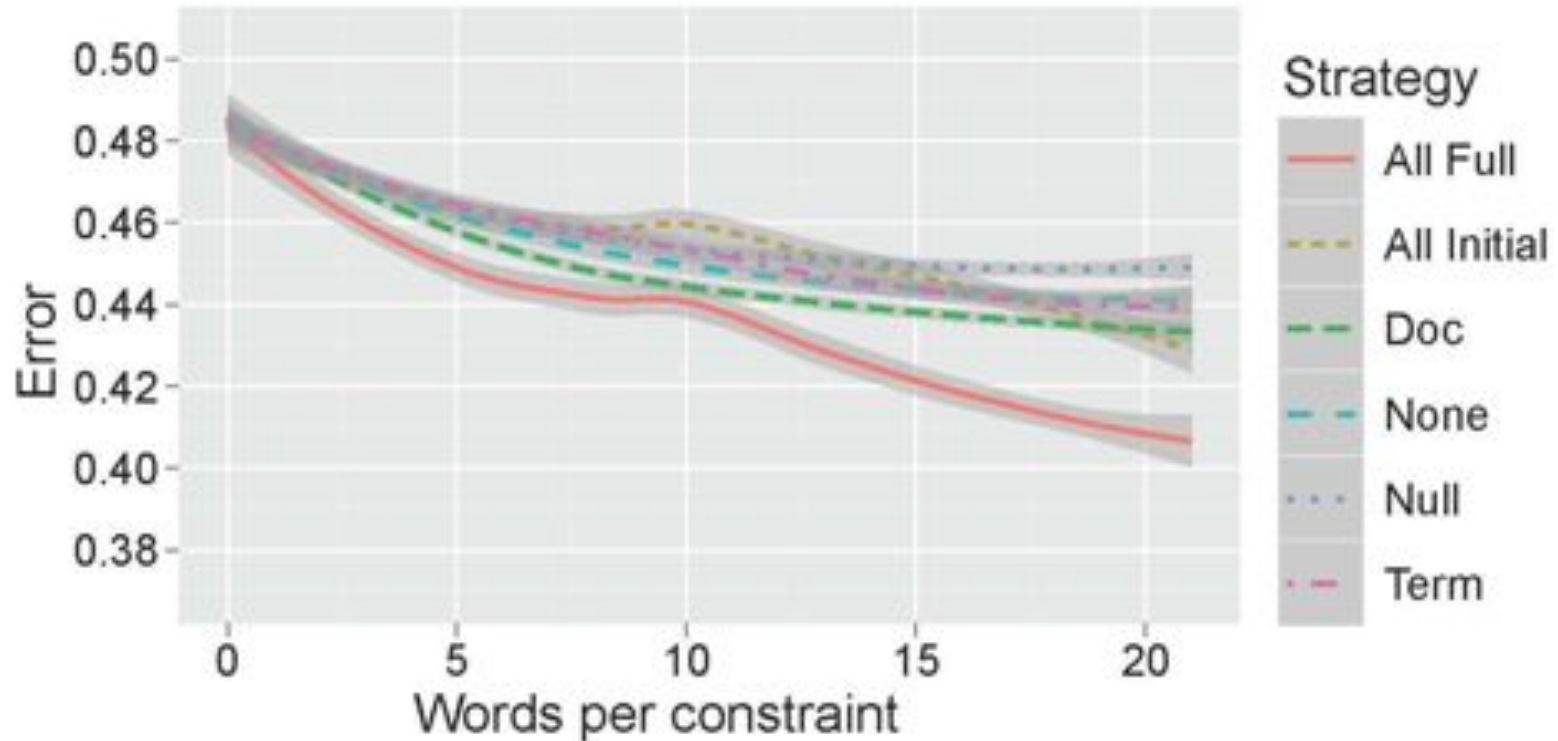
- Sampling equation

$$p(z_{d,n} = k | \mathbf{Z}_{-(d,n)}, \alpha, \beta, \eta)$$

$$\propto \begin{cases} \frac{T_{d,k} + \alpha}{T_{d,\cdot} + K\alpha} \frac{P_{k,w_{d,n}} + \beta}{P_{k,\cdot} + V\beta} & \text{if } \forall l, w_{d,n} \notin \Omega_l \\ \frac{T_{d,k} + \alpha}{T_{d,\cdot} + K\alpha} \frac{P_{k,l} + C_l\beta}{P_{k,\cdot} + V\beta} \frac{W_{k,l,w_{d,n}} + \eta}{W_{k,l,\cdot} + C_l\eta} & w_{d,n} \in \Omega_l \end{cases}$$

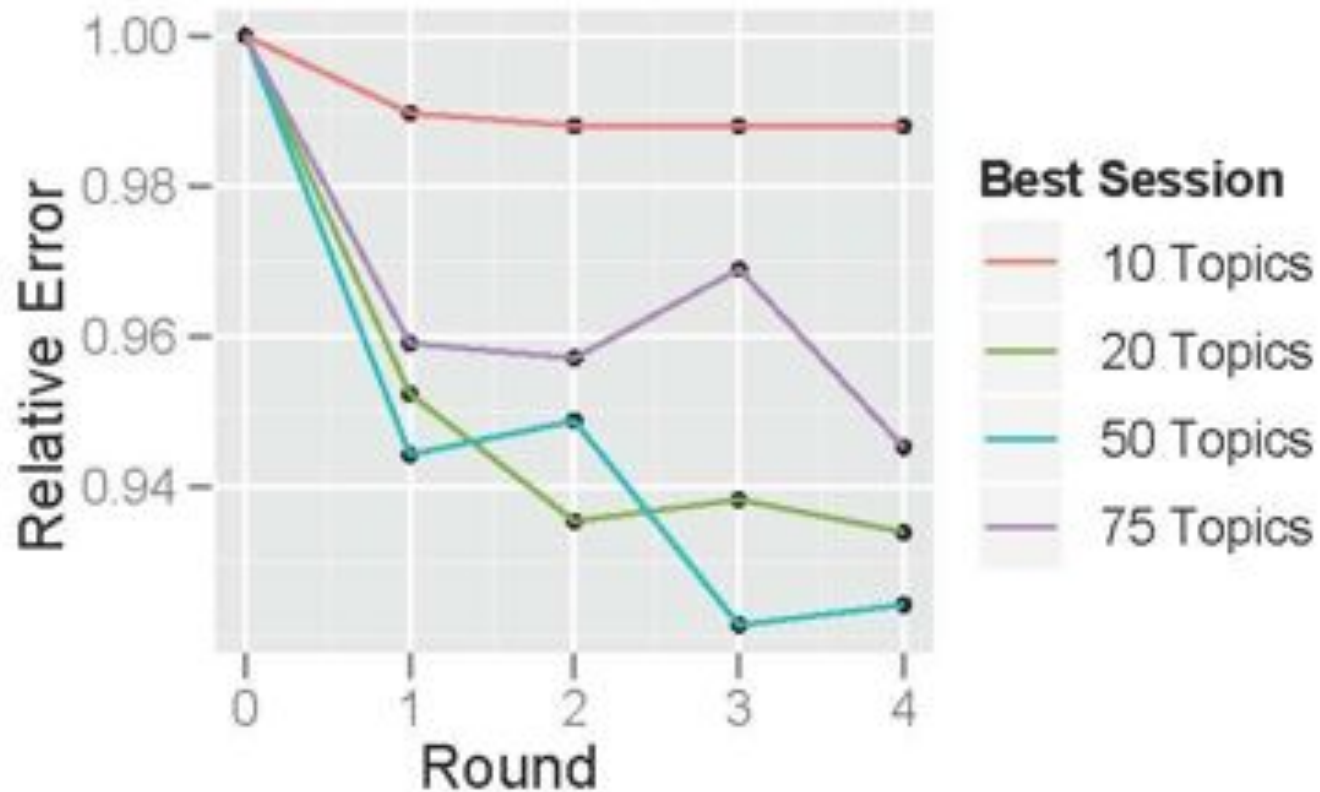
- $P_{k,w_{d,n}}$  number of times the unconstrained word  $w_{d,n}$  appears in topic  $k$
- $P_{k,l}$  number of times any word of constraint  $\Omega_l$  appears in topic  $k$
- $W_{k,l,w_{d,n}}$  the number of times word  $w_{d,n}$  appears in constraint  $\Omega_l$  in topic  $k$
- $V$  vocabulary size
- $C_l$  number of words in constraint  $\Omega_l$

# Which strategy?



- All Full: all constraints are known, comparable iters
- All Initial: all constraints are known, 100 iters
- Null: no constraints, comparable iters

# Put humans in the loop



# Reference

---

1. David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
2. Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.
3. David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of International Conference of Machine Learning*.
4. Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*.
5. Jonathan Chang. 2010. Not-so-latent dirichlet allocation: Collapsed gibbs sampling using human judgments. In *NAACL Workshop: Creating Speech and Language Data With Amazon's Mechanical Turk*.