Exploiting Submodularity to Tame Information Overload

Khalid El-Arini

Dafna Shahaf

Yisong Yue

Carlos Guestrin

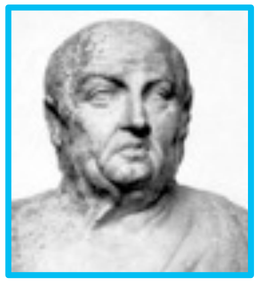
Select Lab

Carnegie Mellon

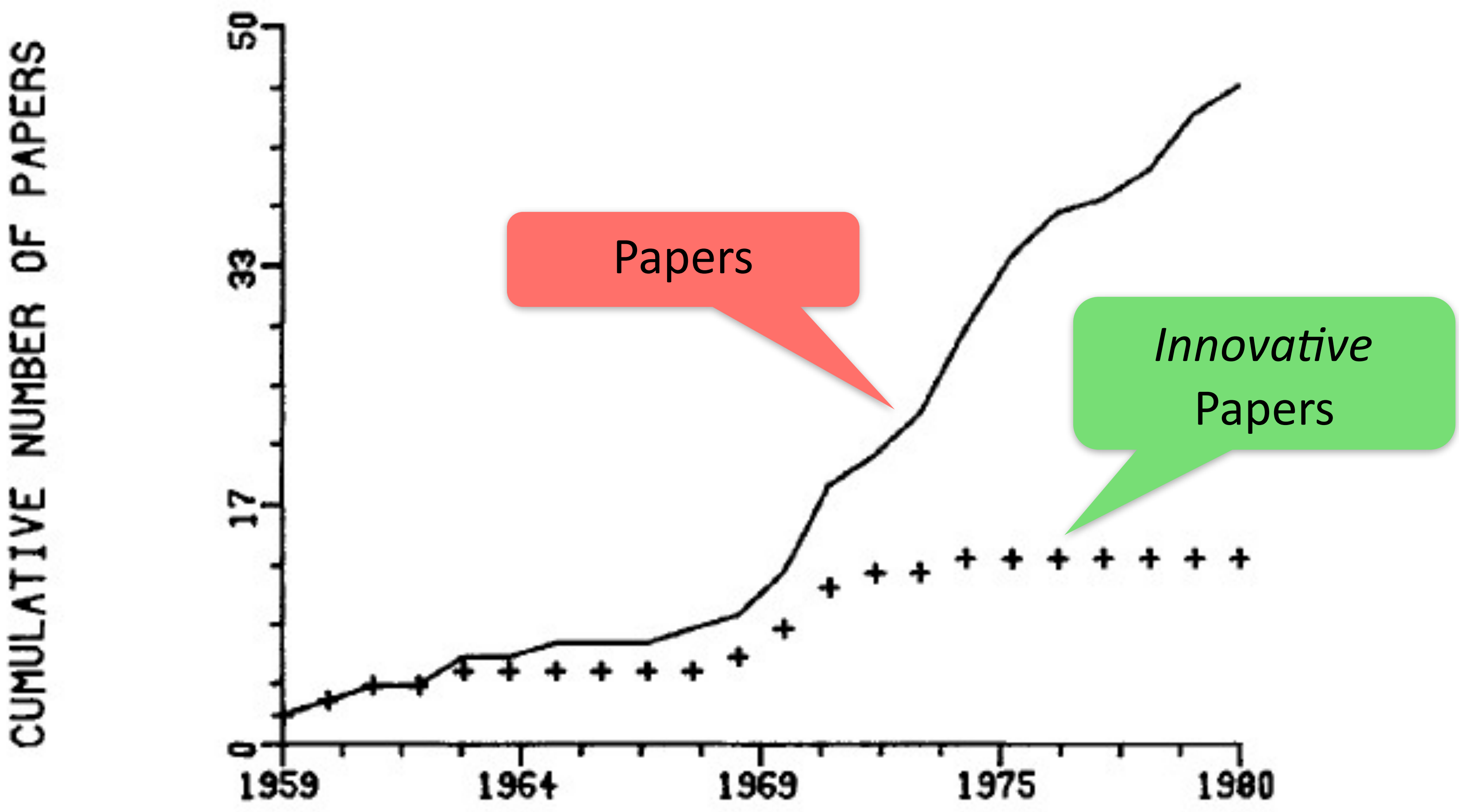
“ The abundance of books is a distraction ”

Lucius Annaeus Seneca

4 BC - 65 AD

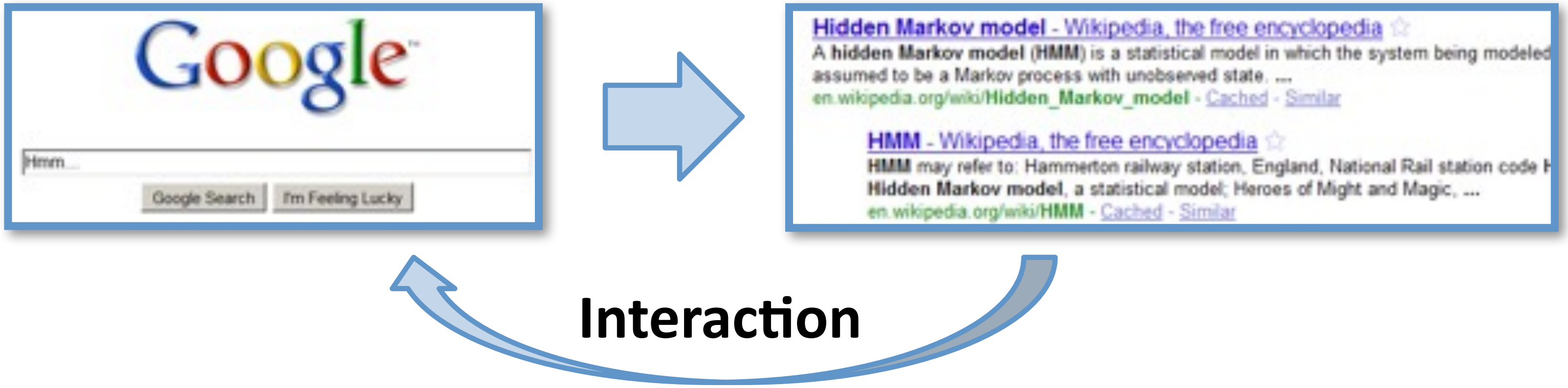


And it's getting worse... [Tague et al. 1981]



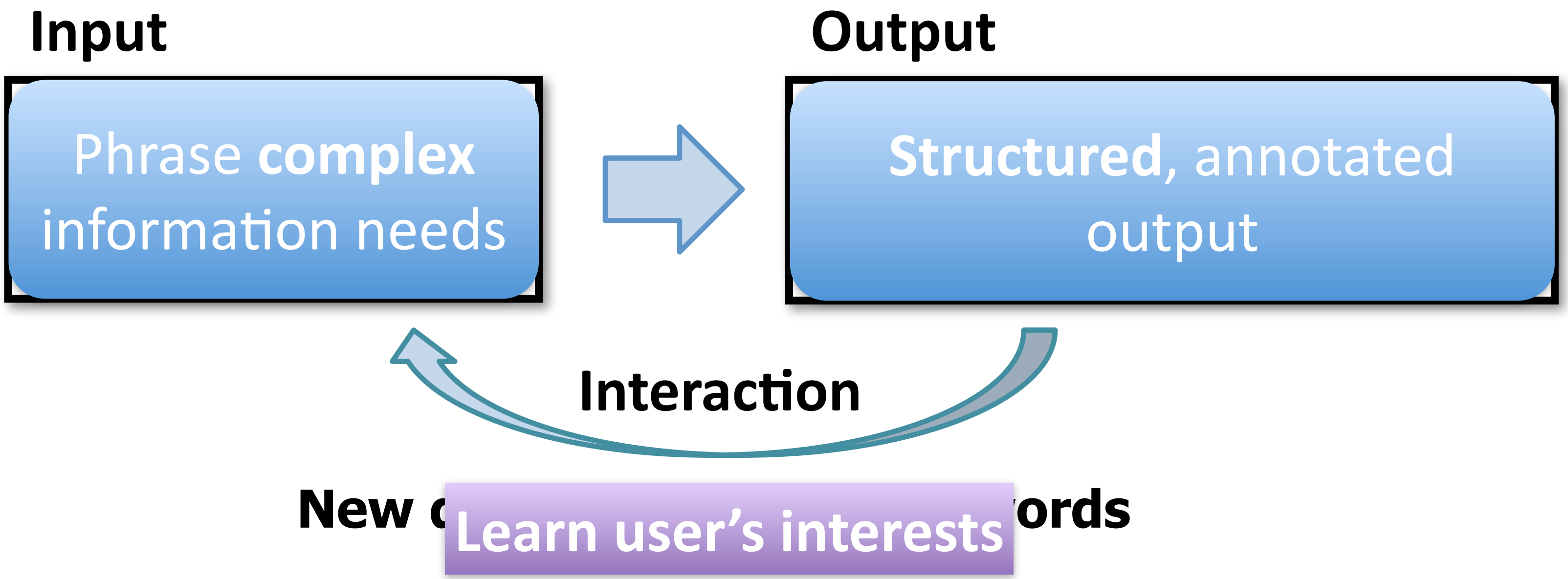
Search Limitations

Input **Output**



New query → Change keywords

Our Approach



sense learn act

TIME IN PARTNERSHIP WITH CNN

Thursday, Nov. 20, 2008

How Many Blogs Does the World Need?

By Michael Kinsley

- Millions of blog posts published every day
- Some stories become disproportionately popular
  - Hard to find information you care about

Palin Camp Takes Umbrage at

Sep 6, 2008, 4:34 p.m. EST

Treasury set to bail out Fannie Mae, Freddie Mac

Fiddling with Pig Lipstick While The Financial System Burns

Day of misery on Wall Street

Reuters

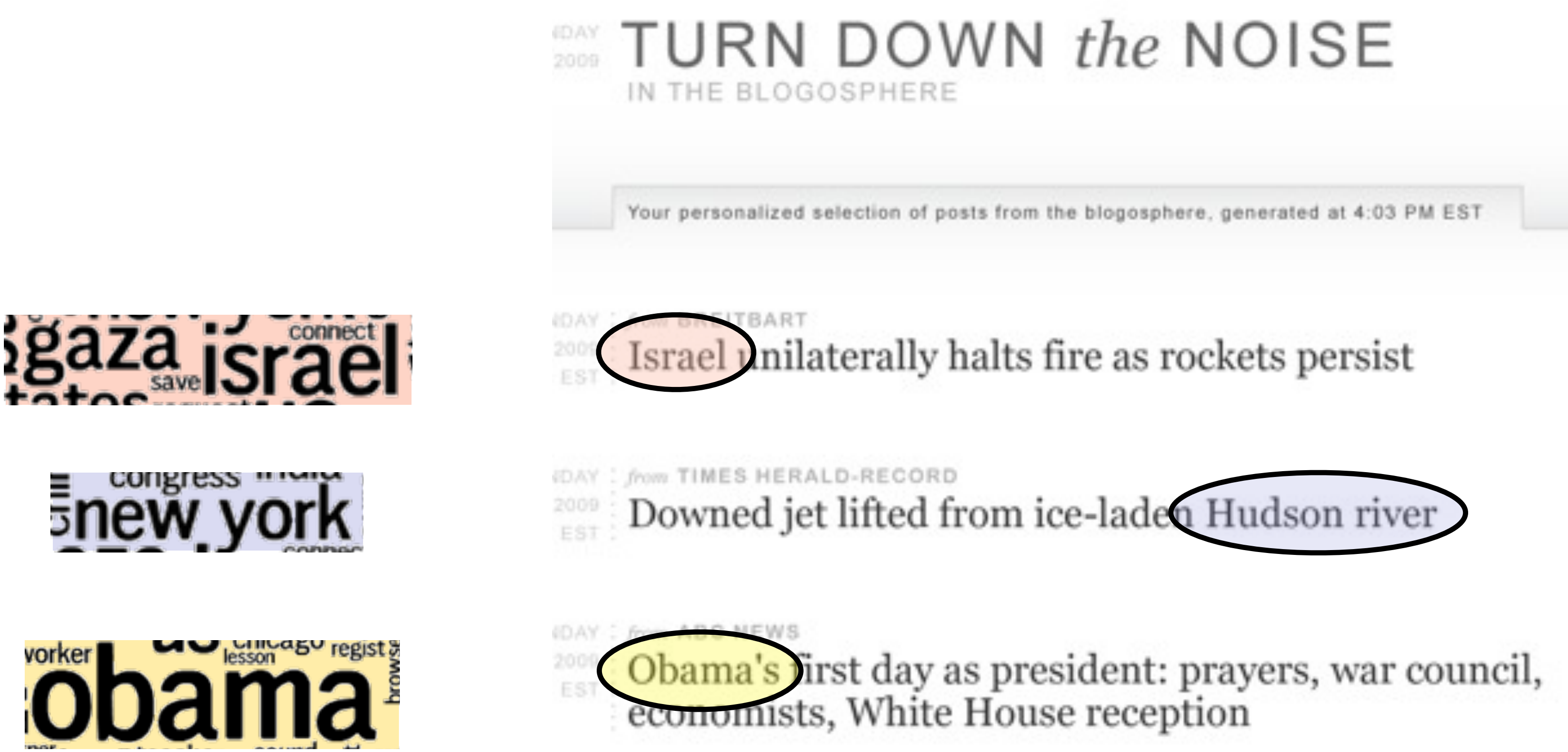
Our goal: **coverage**

- Turn down the noise in the blogosphere
  - **select a small set of posts that covers the most important stories**

January 17, 2009

Our goal: **coverage**

- Turn down the noise in the blogosphere
  - **select a small set of posts that covers the most important stories**



Our goal: **personalization**

- Tailor post selection to user tastes

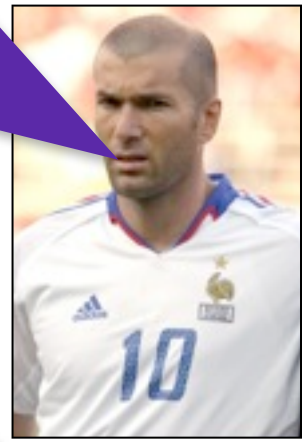
posts selected without personalization

But, I like sports! **I want articles like:**

Manchester United fold without a fight as Barcelona claim Champions League

Parker Scores 19 to Lead San Antonio Past Clippers

Parker scores 19 to lead Spurs over Clippers 106-84 despite only 8 points from Duncan

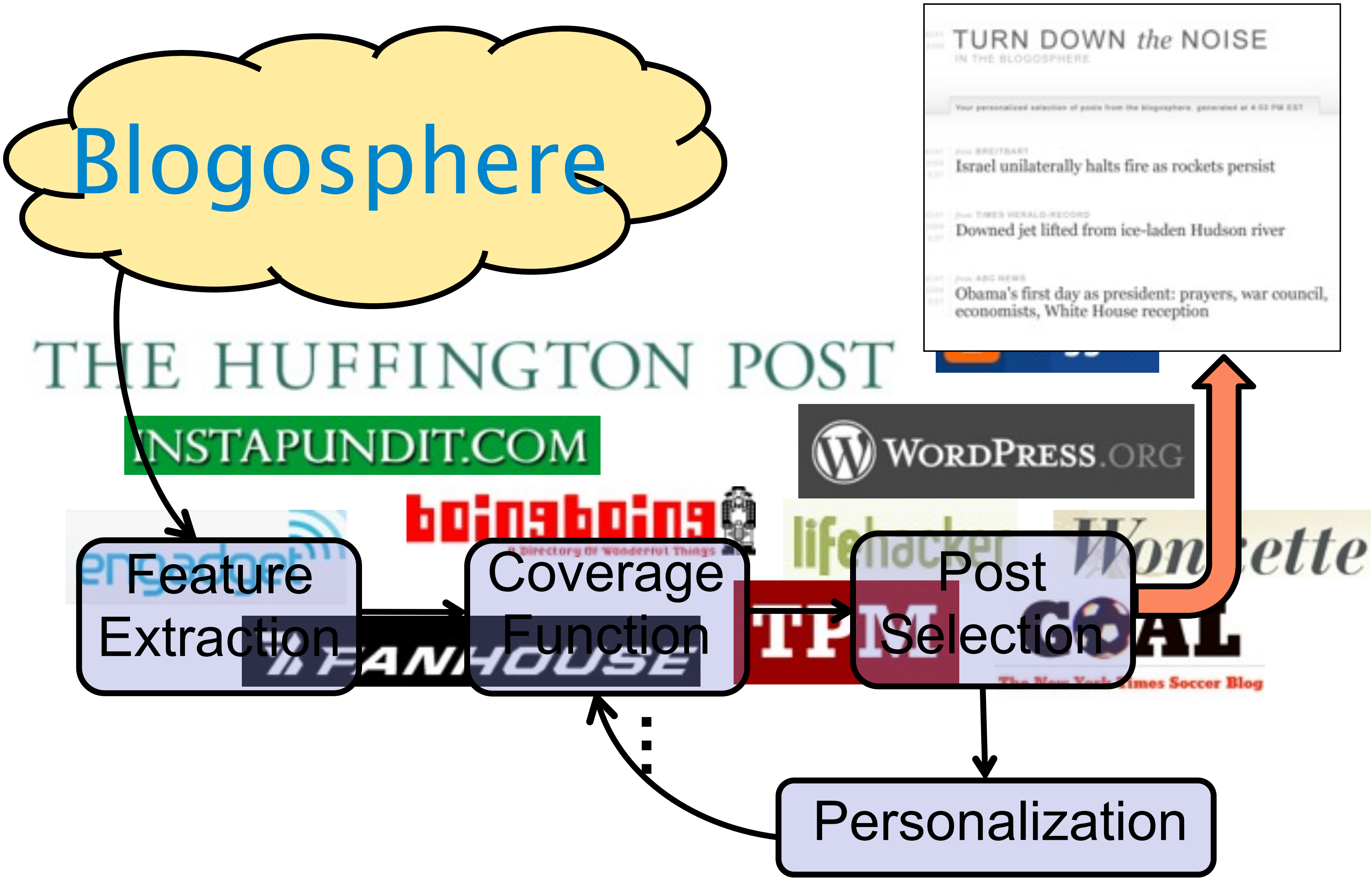


After personalization based on Zidane's feedback

TDN outline [El-Arini, Veda, Shahaf, G. '09]

- Coverage:
  - Formalize notion of covering the blogosphere
  - Near-optimal solution for post selection
  - Evaluate on real blog data and compare against:
- Personalization:
  - Learn a personalized coverage function
    - Algorithm for learning user preferences using limited feedback
  - Evaluate on real blog data

Approach Overview



sense learn act

Document Features

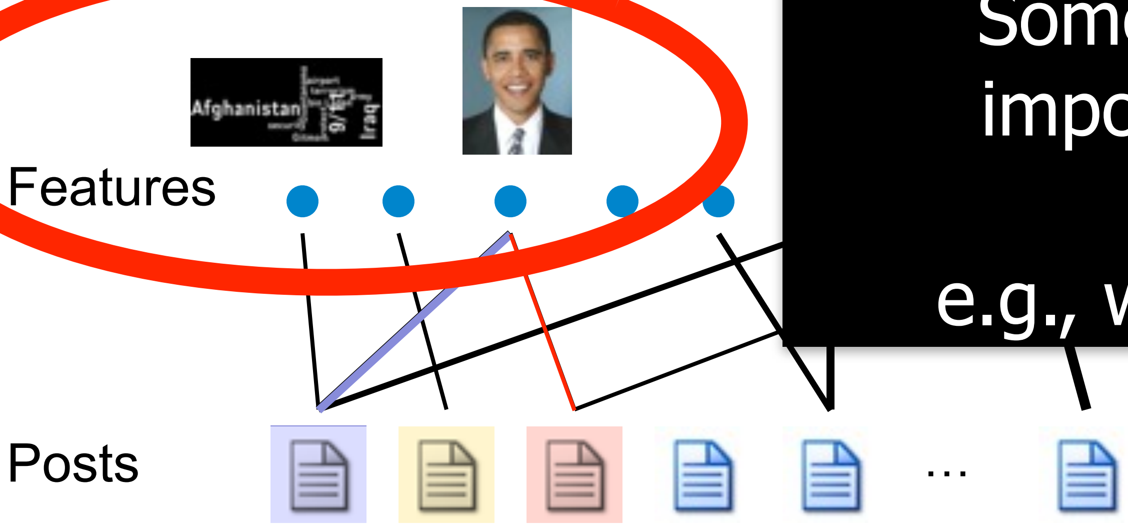
- Low level
  - Words, noun phrases, named entities
    - e.g., Obama, China, peanut butter

- High level
  - e.g., Topics
  - Topic = probability distribution over words

Inauguration Topic

National Security Topic

Coverage Function



Some features more important than others
⟹
e.g., weigh by frequency

- cover ( ) = amount by

$\text{cover}_d(f)$

- cover ( ) = amount

$\text{cover}_{\mathcal{A}}(f)$

A post never covers a feature completely
⟹
use soft notion of coverage, e.g., prob. at least one post in A covers feature f

Objective Function for Post Selection

- Want to select a set of posts $\mathcal{A}$ that maximizes

$$F(\mathcal{A}) = \sum_{f \in \mathcal{U}} w_f \, cover_{\mathcal{A}}(f)$$

Posts shown to user

feature set

weights on features

probability that set $\mathcal{A}$ covers feature f

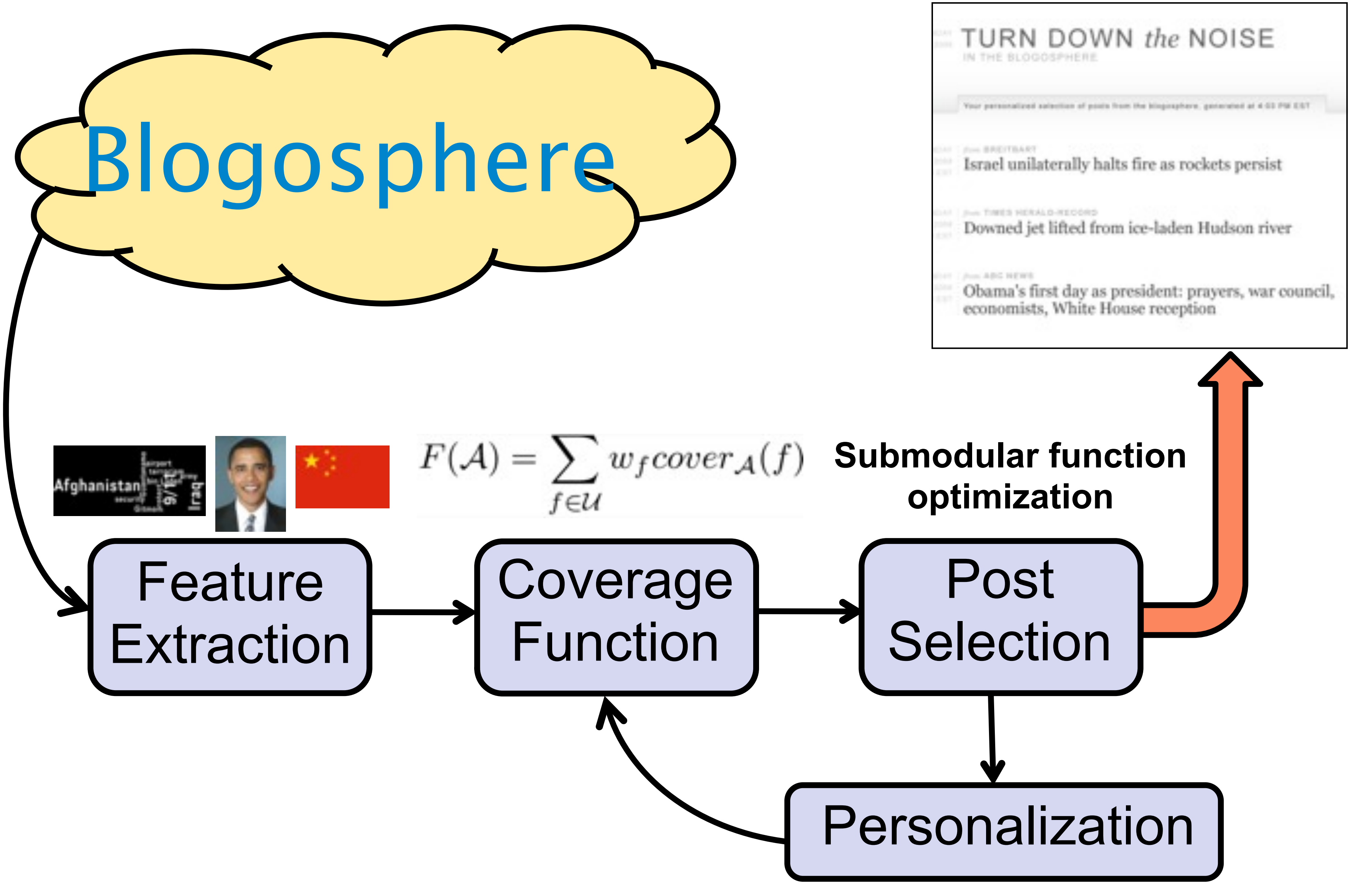
- Maximizing $F(\mathcal{A})$ is NP-hard!

$\boxed{F(\mathcal{A}) \text{ is submodular}}$

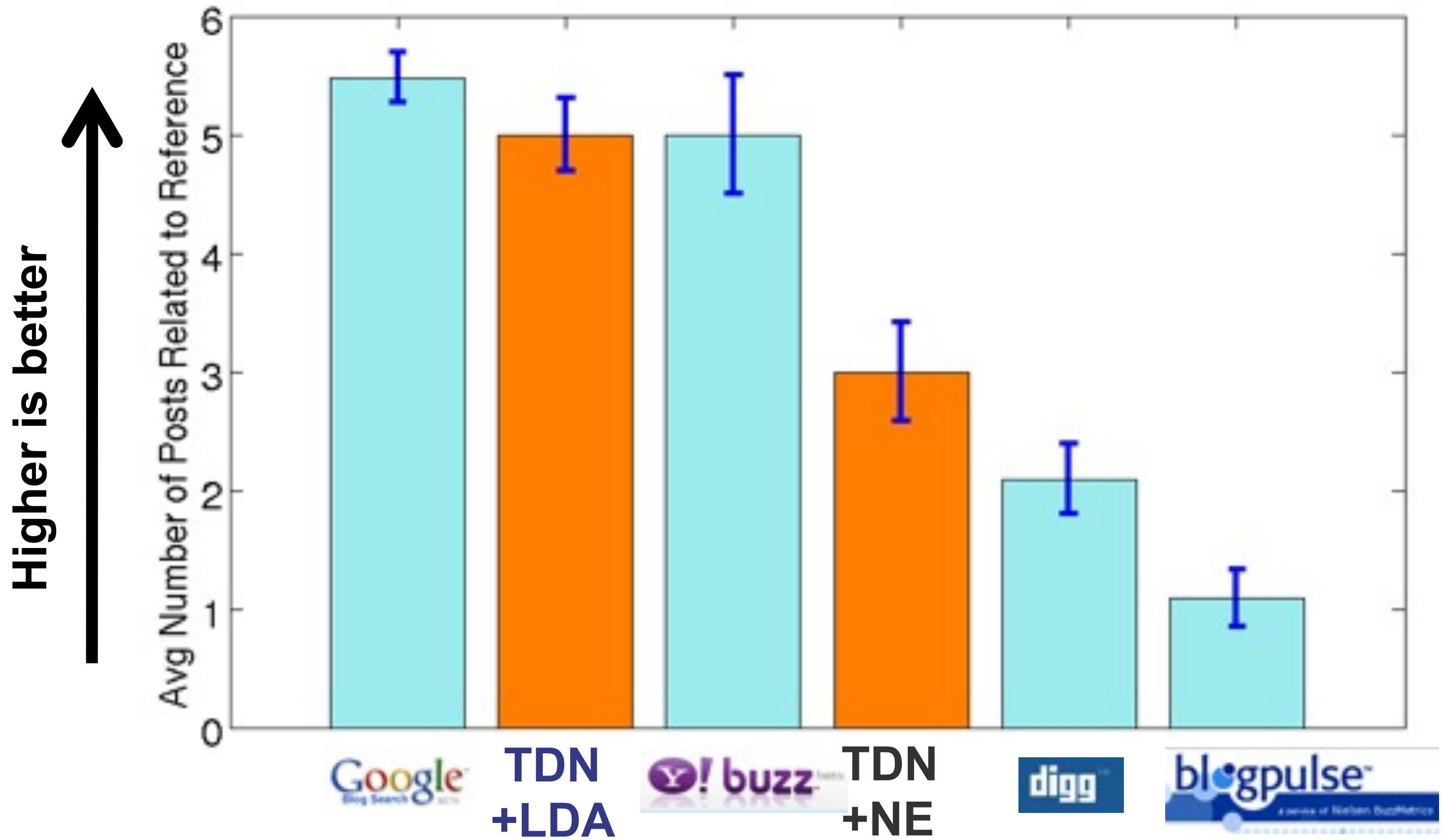
Greedy $\Rightarrow$ (1-1/e)-approximation

Lazy greedy (CELF) $\Rightarrow$ very fast, same guarantees

Approach Overview



User study:

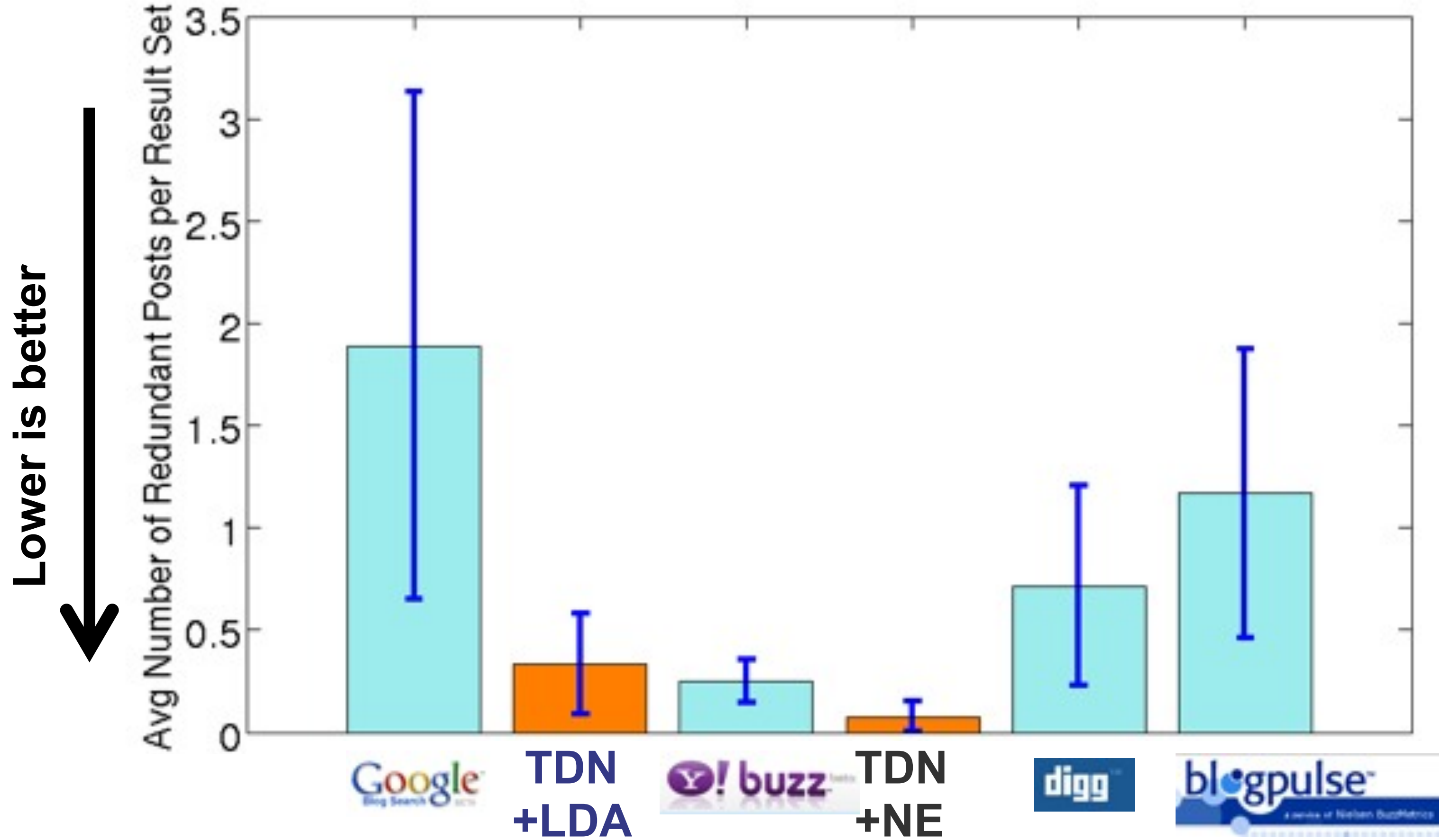


LDA topics as features

Named entities and common noun phrases as features

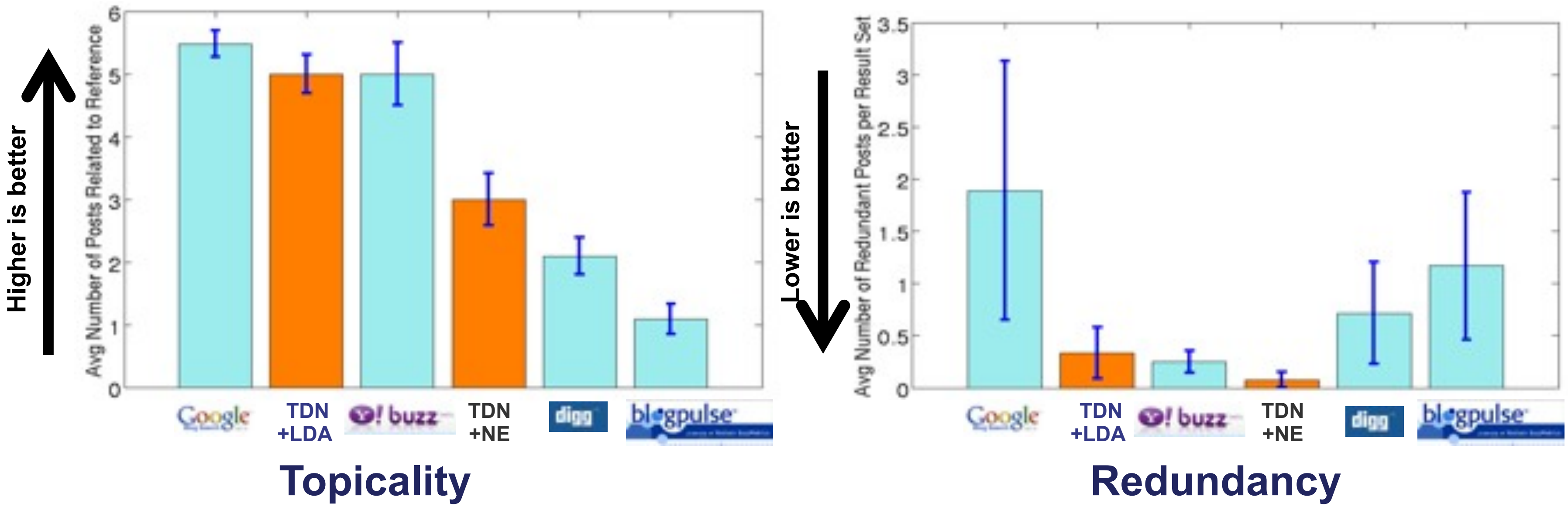
We do as well as Yahoo! and Google

User study:



Google performs poorly
We do as well as Yahoo!

User study summary



- Google: good topicality, high redundancy
- Yahoo!: performs well on both, but **uses rich features**
  - CTR, search trends, user voting, etc.

TDN performs as well as Yahoo!
using only post content

TDN outline

- Coverage:
  - Formalize notion of covering the blogosphere
  - Near-optimal solution for post selection
  - Evaluate on real blog data and compare against:

- Personalization:
  - Learn a personalized coverage function
    - Algorithm for learning user preferences using limited feedback
  - Evaluate on real blog data

Personalization

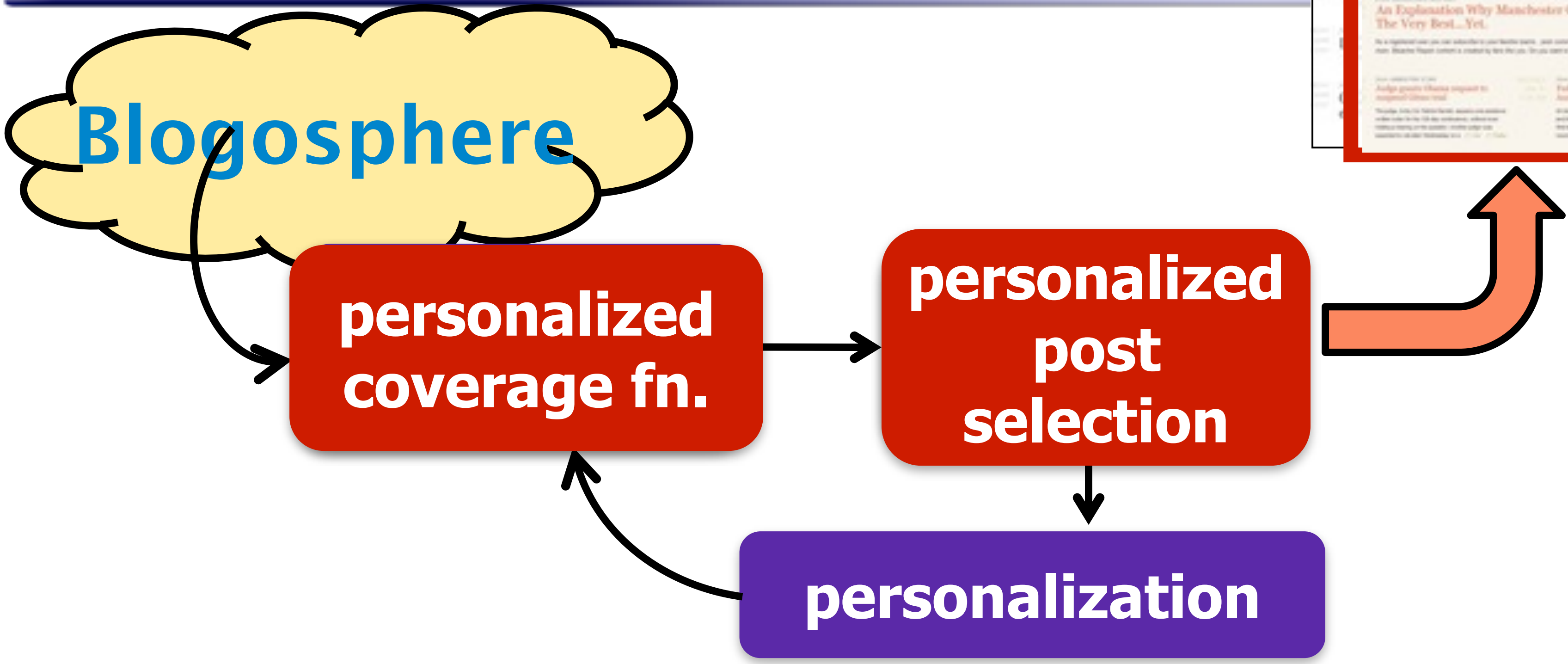
- People have varied interests

Barack Obama

Britney Spears

- **Our Goal**: Learn a personalized coverage function using limited user feedback

Personalize postings



learn your coverage function

online learning of submodular function problem

Modeling User Preferences

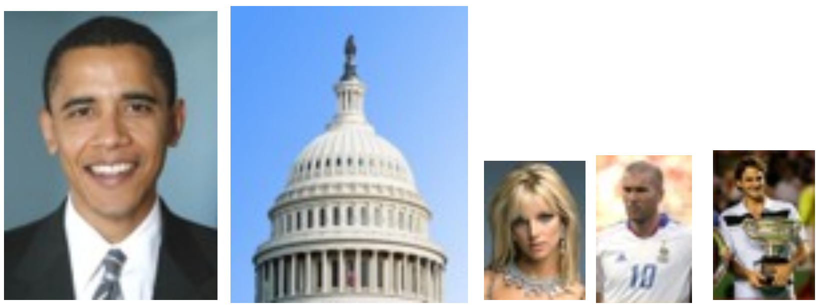
[Yue, G. 2011]

$$F_{\pi^*}(\mathcal{A}) = \sum_{f \in \mathcal{U}} \pi^*_f \cdot cover_{\mathcal{A}}(f)$$

Importance of feature in corpus

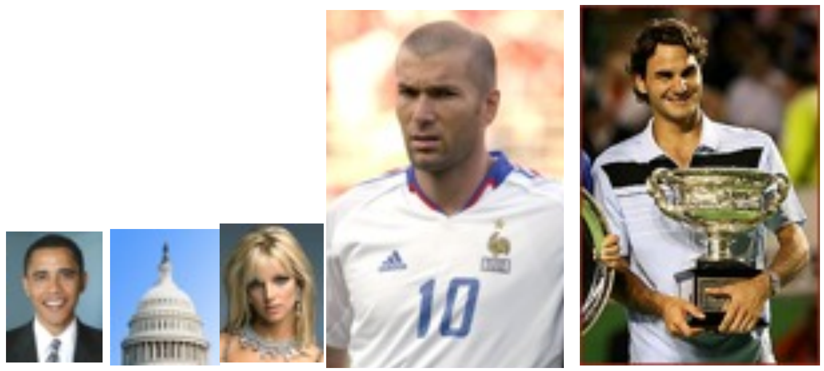
User preference

- π^*_f represents user preference for feature f
- Want to learn preference π^* over the features



π^*_1 π^*_2 π^*_3 π^*_4 π^*_5

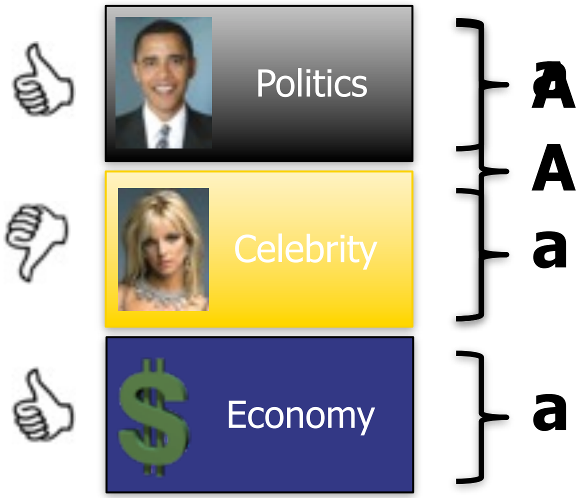
π^* for a politico



π^*_1 π^*_2 π^*_3 π^*_4 π^*_5

π^* for a sports fan

User Model



- User scans articles in order
- Stochastically generates feedback (reward)
- Independent of other feedback
- Depends on above articles

$$\mathbf{E}[r(a) \mid A] = \left(\pi^*\right)^T \Delta(a \mid A)$$

"Conditional Submodular Independence"

$$\Delta(a \mid A) = \begin{bmatrix} F_1(A \cup a) - F_1(A) \\ F_2(A \cup a) - F_2(A) \\ \vdots \\ F_D(A \cup a) - F_D(A) \end{bmatrix}$$

$$\mathbf{E}[r(A)] = E\ r\ a_i$$

Fitting User's Feedback

- Simple regression approach fits preference vector to expected reward:

Submodular advantage of article a_2 wrt each feature

$$\begin{bmatrix} \Delta(a_1 \mid A_1) \\ \Delta(a_2 \mid A_2) \\ \vdots \\ \Delta(a_t \mid A_t) \end{bmatrix} \begin{bmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_d \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_t \end{bmatrix}$$

Reward of article a_2

+ some regularization

Exploration vs Exploitation

- **Goal:** want to recommend content that user likes
  - Exploiting feedback from user, maximizing reward
- **However:** user only provides feedback on recommended content
  - Explore to collect feedback for new topics
    - Not addressed by [El-Arini, Veda, Shahaf, G. 2009]
- **Solution:** algorithm to balance exploration vs exploitation
  - Linear Submodular Bandits Problem

Balancing Exploration & Exploitation

Estimated coverage gain

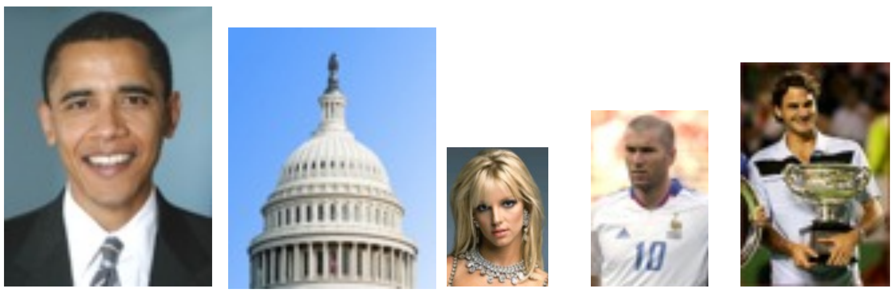
Uncertainty of estimate

$$\alpha_t \sqrt{\Delta(a \mid A_t)^T M_t^{-1} \Delta(a \mid A_t)}$$

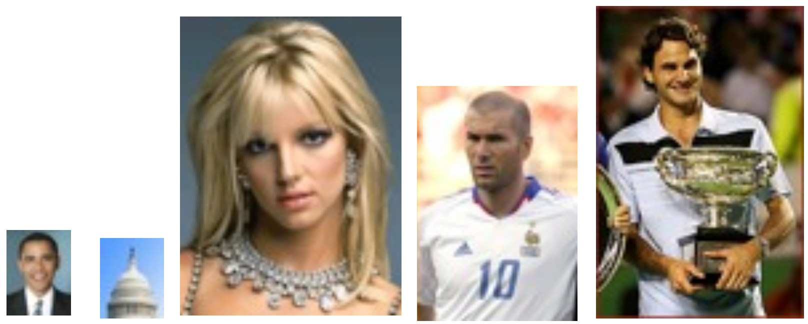
- For each slot, maximize trade-off
  - (pick article about **Tennis**)

Mean Estimate by Topic

Uncertainty of Estimate



$\hat{\pi}_1$ $\hat{\pi}_2$ $\hat{\pi}_3$ $\hat{\pi}_4$ $\hat{\pi}_5$



C_1 C_2 C_3 C_4 C_5

Learning User Preferences (Approach 2)

[Yue & Guestrin, 2011]

Explor
Tra

oit

Theorem: For Bandit Alg,

$$(1\text{-}1/e)\,\text{avg}(\ \ ) - \text{avg}(\ \ ) \longrightarrow 0$$

True **Learned**

π^* **π**

i.e., we achieve **no-regret**

Learns a good approximation of the true π^*

Rate: d/sqrt(kT),
recommending k documents, T rounds, d features

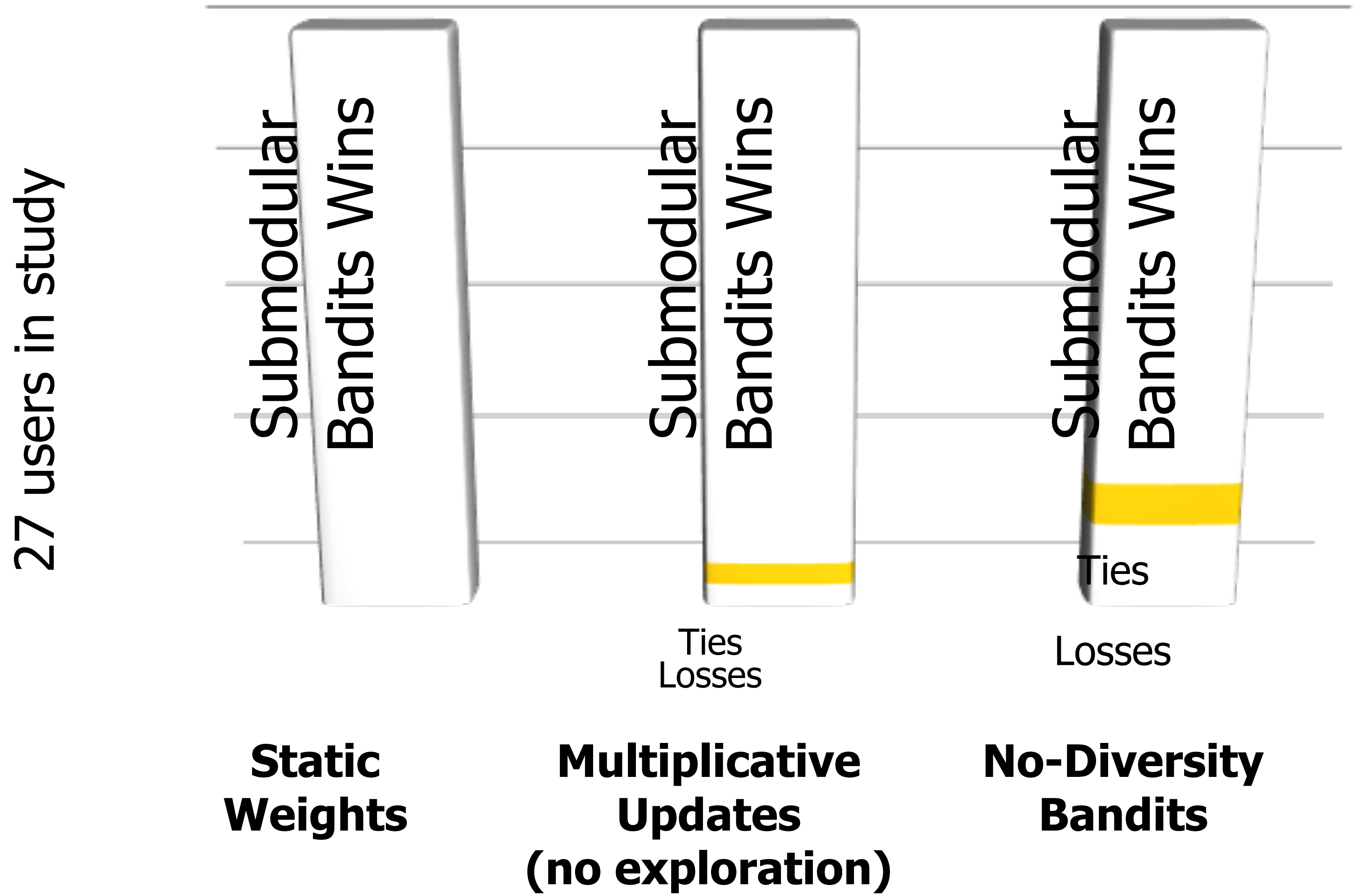
$\pi^{(0)}$ $\pi^{(1)}$ $\pi^{(2)}$

Before any feedback

After 1 day of personalization

After 2 days of personalization

User Study:


27 users in study
Submodular Bandits Wins
Ties
Losses
Submodular Bandits Wins
Ties
Losses
Submodular Bandits Wins
Static Weights
Multiplicative Updates (no exploration)
No-Diversity Bandits

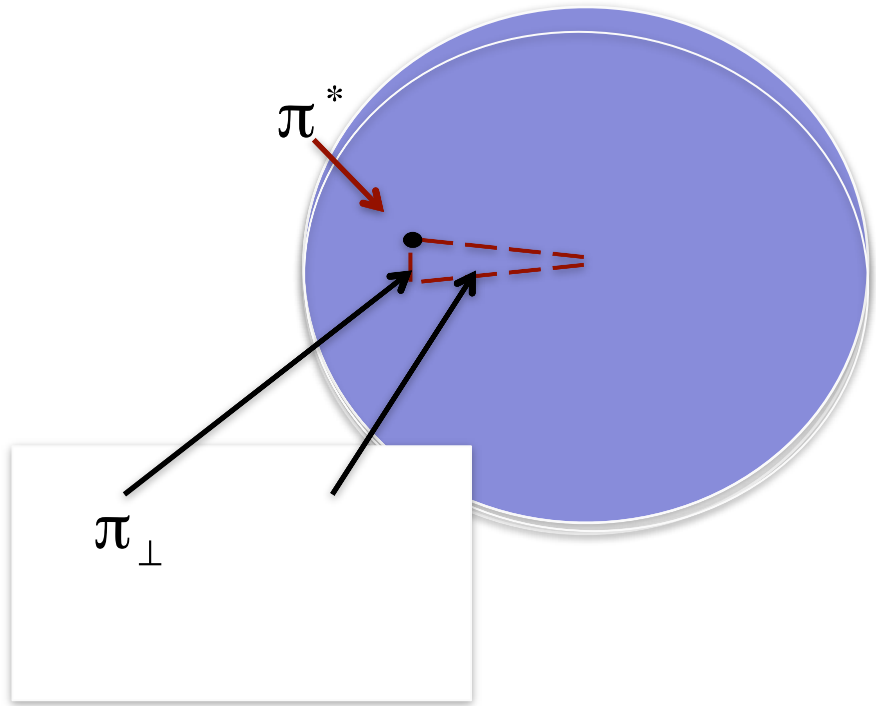
Problems in High Dimension

- Convergence rate has linear dependency on dimensionality:

$$\pi^*$$

Hierarchical Bandit Learning [Yue, Hong, G., 2012]

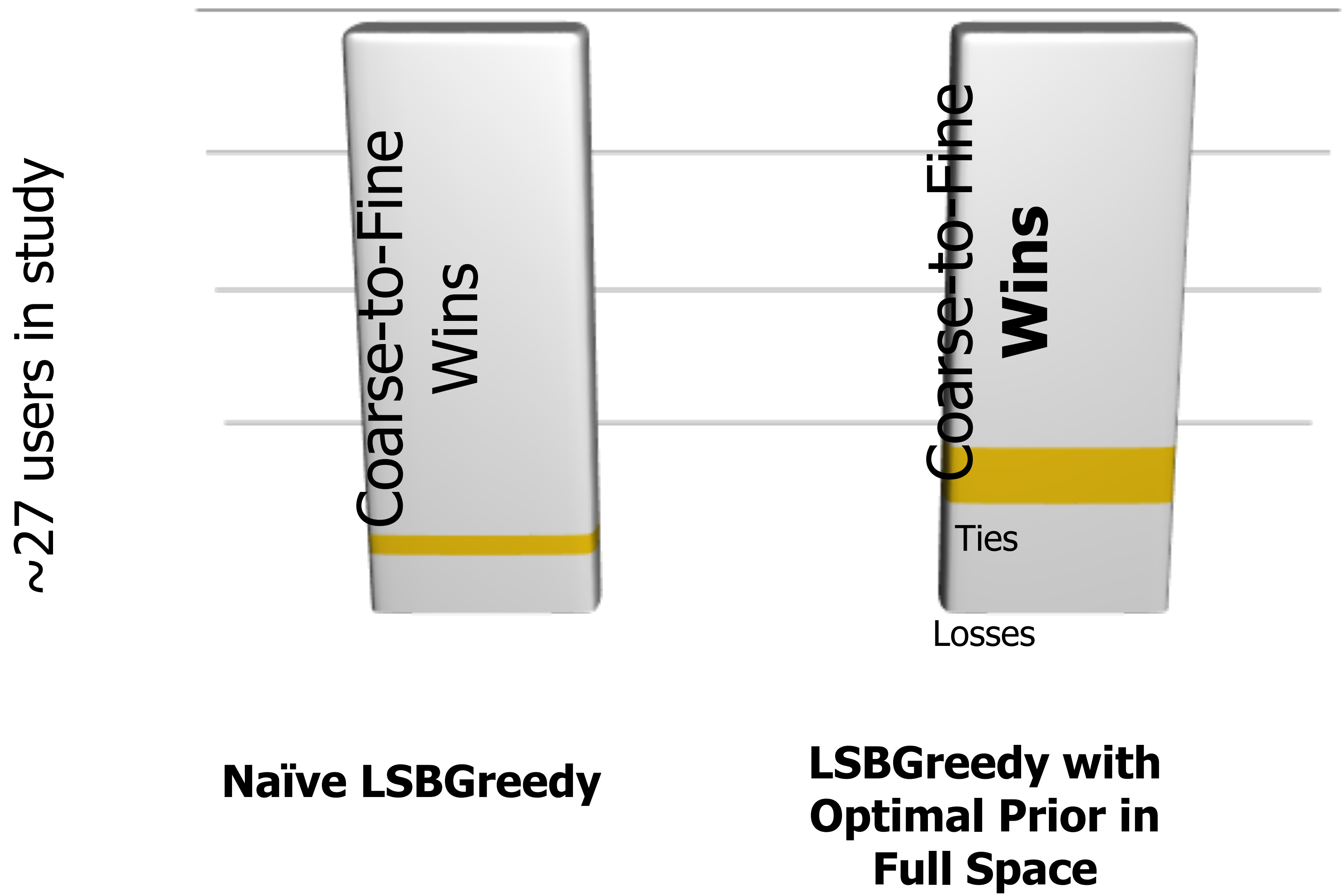
- Assume π^* mostly in subspace
  - Dimension $\ell << d$
  - E.g., Sports vs Politics
- Coarse to fine bandits
  - Two tiered exploration
  - Significantly fewer examples needed



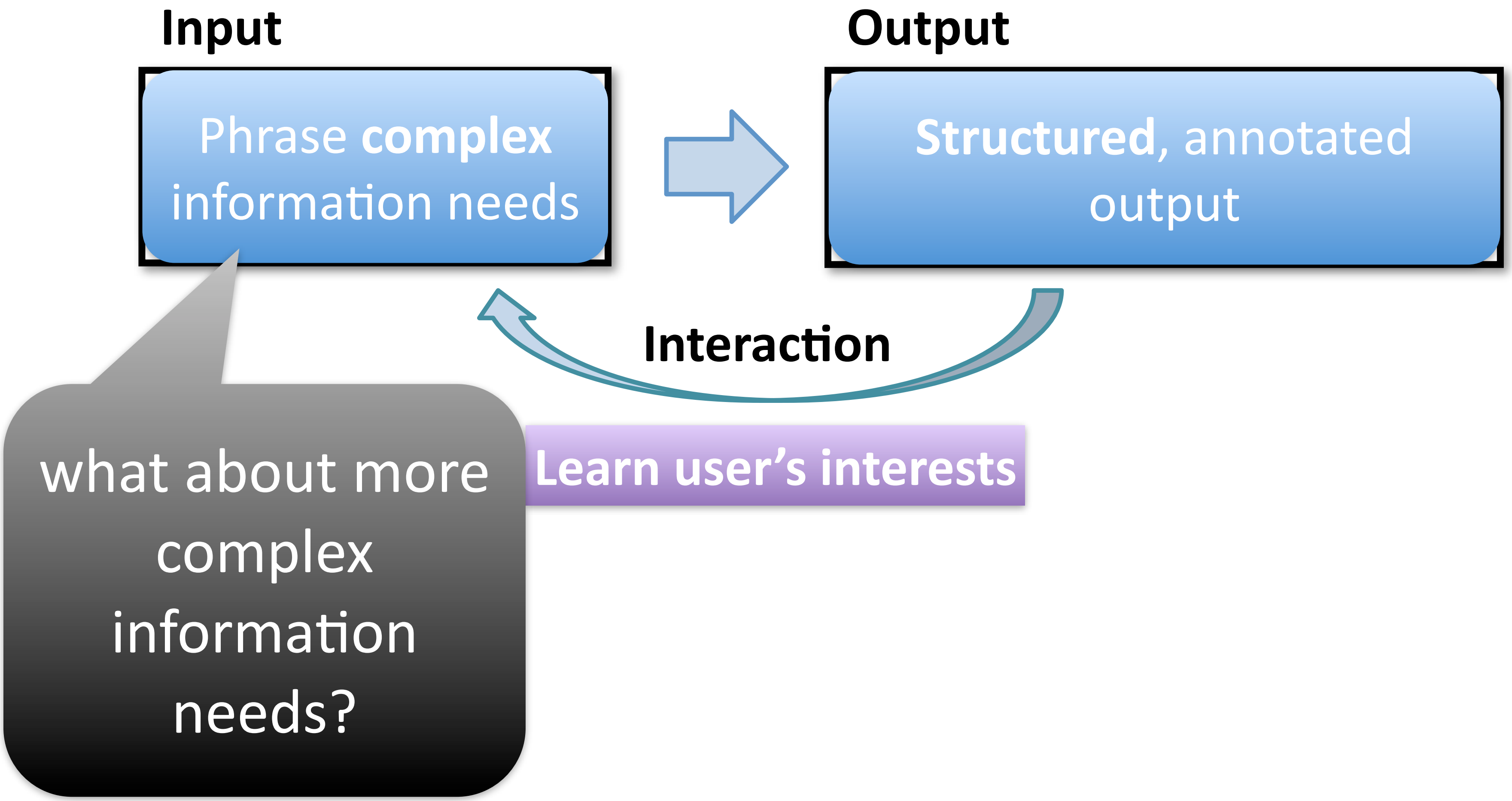
Original Guarantee:

π^*

User Study


~27 users in study
Coarse-to-Fine Wins
Coarse-to-Fine Wins
Ties
Losses
Naïve LSBGreedy
LSBGreedy with Optimal Prior in Full Space

Our Approach



A Prescient Warning

As long as the centuries...unfold, the number of books will grow continually...

as convenient to search for a bit of truth concealed in nature

as to find it hidden away in an immense multitude of bound volumes

-Dennis Diderot, Encyclopédie (1755)

Today: 10^7 papers in 10^5 conferences/journals*

How do we cope?

* Thomson Reuters Web of Knowledge

Motivation (1)

- Is there an approximation algorithm for the submodular covering problem that doesn't require an integral-valued objective function?

L.A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2:385–393, 1982.

Any recent papers influenced by this?

Motivation (2)

- It's 11:30pm Samoa Time. Your "Related Work" section is a bit sparse.

[16] T. Finin, A. Joshi, P. Kolari, A. Java, A. Kale, and A. Karandikar. The information ecology of social media and online communities. *AI Magazine*, 2008.

[17] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*, 2005.

[18] Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 2000.

[19] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[20] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 1999.

[21] M. Kinsley. How many blogs does the world need? *TIME Magazine*, 172(22), December 2008.

[22] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[23] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7, 2003.

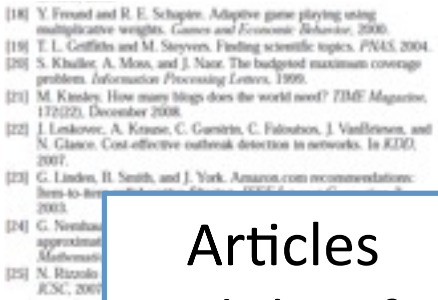
[24] G. Nemhauser, L. ... and M. Fisher. An analysis of the ... maximizing submodular set functions. ... 14:265–294, 1978.

... eling discriminative global inference. In

Here are some papers we've cited so far. Anything else?

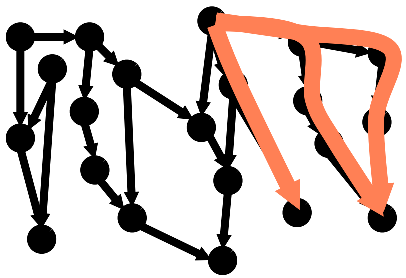
Recommending Scientific Articles

[El-Arini, G. '11]



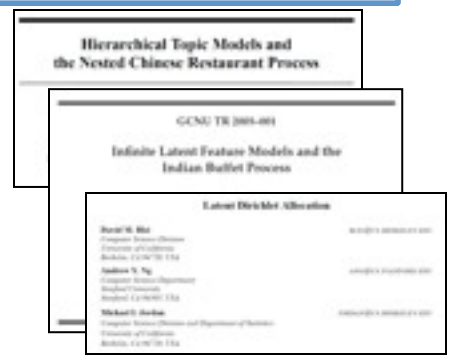
Articles read thus far

Model of influence in science



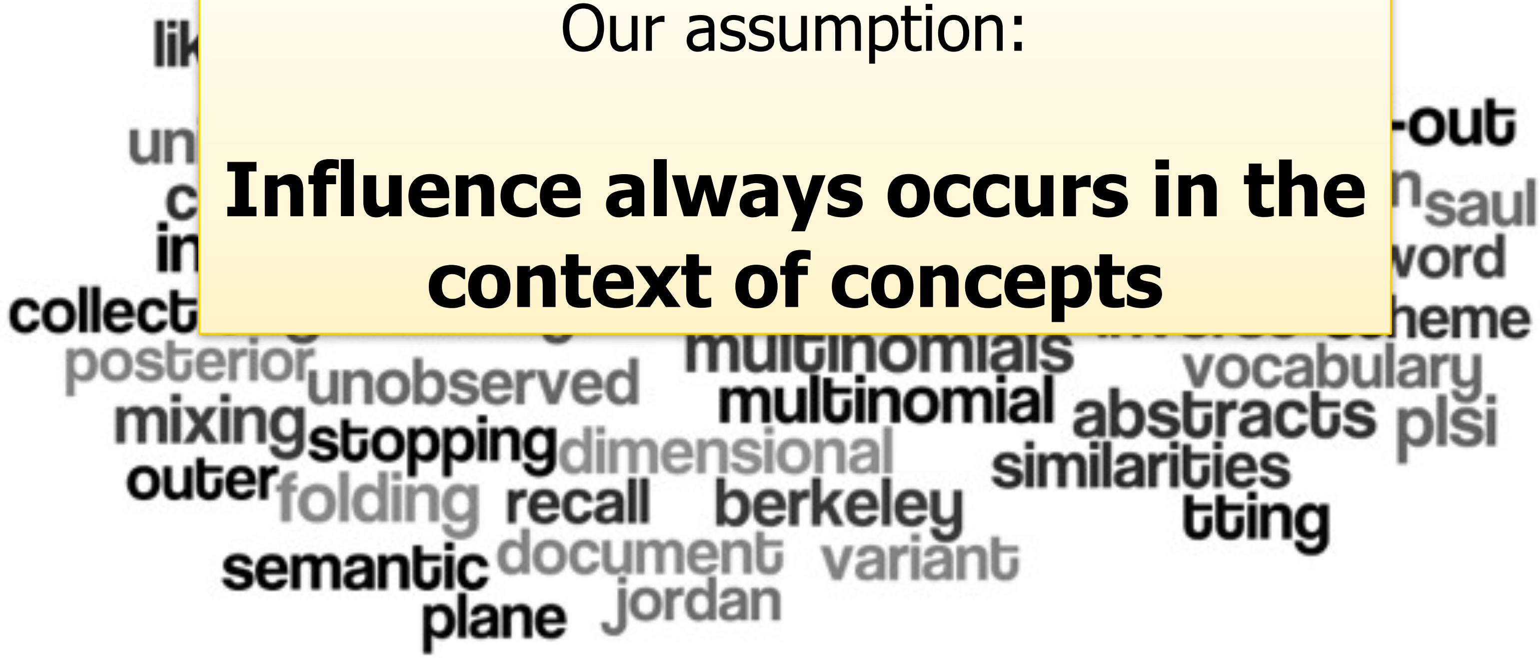
Submodular function optimization

Diverse set of recommended articles



Concept representation

- Words, phrases or important technical terms
- Proteins, genes, or other advanced features



Our assumption:

Influence always occurs in the context of concepts

Putting it all together

selected papers

$$F_Q(\mathcal{A}) = \sum_{q \in Q} \sum_{c \in \mathcal{C}} \gamma_q^{(c)} \, Influence_c(q \leftrightarrow \mathcal{A})$$

query set

concept set

prevalence of c in paper q

probability of influence between q and **at least one** paper in **A** wrt concept **c**

- Maximize $F_Q(\mathbf{A})$ s.t. $|A| \leq k$ (output k papers)
- **Submodular** maximization problem

But should all users get the same results?

Personalized trust

- Different communities trust different researchers for a given concept

e.g., network

Pearl

Kleinberg

Hinton

- **Goal:** Estimate personalized trust from limited user input

Specifying trust preferences

- Specifying trust should not be an onerous task
- Assume given (nonexhaustive!) set of trusted papers **B**, e.g.,
  - a **BibTeX** file of all the researcher's previous citations
  - a short list of **favorite conferences** and journals
  - **someone else's** citation history!

a committee member?
journal editor?
someone in another field?
a Turing Award winner?

Personalized Objective

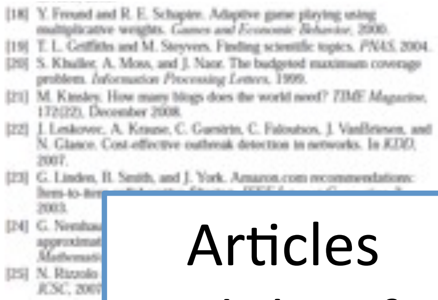
probability of influence between
q and **at least one** paper in **A**

$$F_{\mathcal{Q}|\mathcal{B}}(\mathcal{A}) \quad = \quad \sum_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} \gamma_q^{(c)} \, Influence_c(q \leftrightarrow \mathcal{A}|\mathcal{B})$$

Extra weight in Influence:
Does user trust **at least one** of authors of d with respect to concept c?

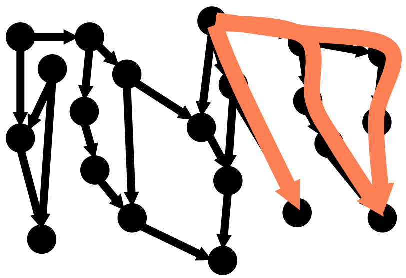
Recommending Scientific Articles

[El-Arini, G. '11]



Articles read thus far

Model of influence in science



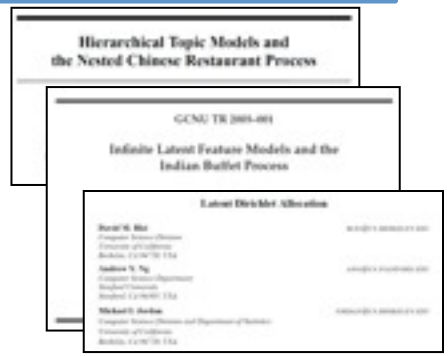
Submodular function optimization

Model user's interests

User's bibtex file

Personalized

Diverse set of recommended articles



Personalized Recommendations

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos — U.C. Riverside

Petros Faloutsos — U. of Toronto

Christos Faloutsos * — Carnegie Mellon Univ.

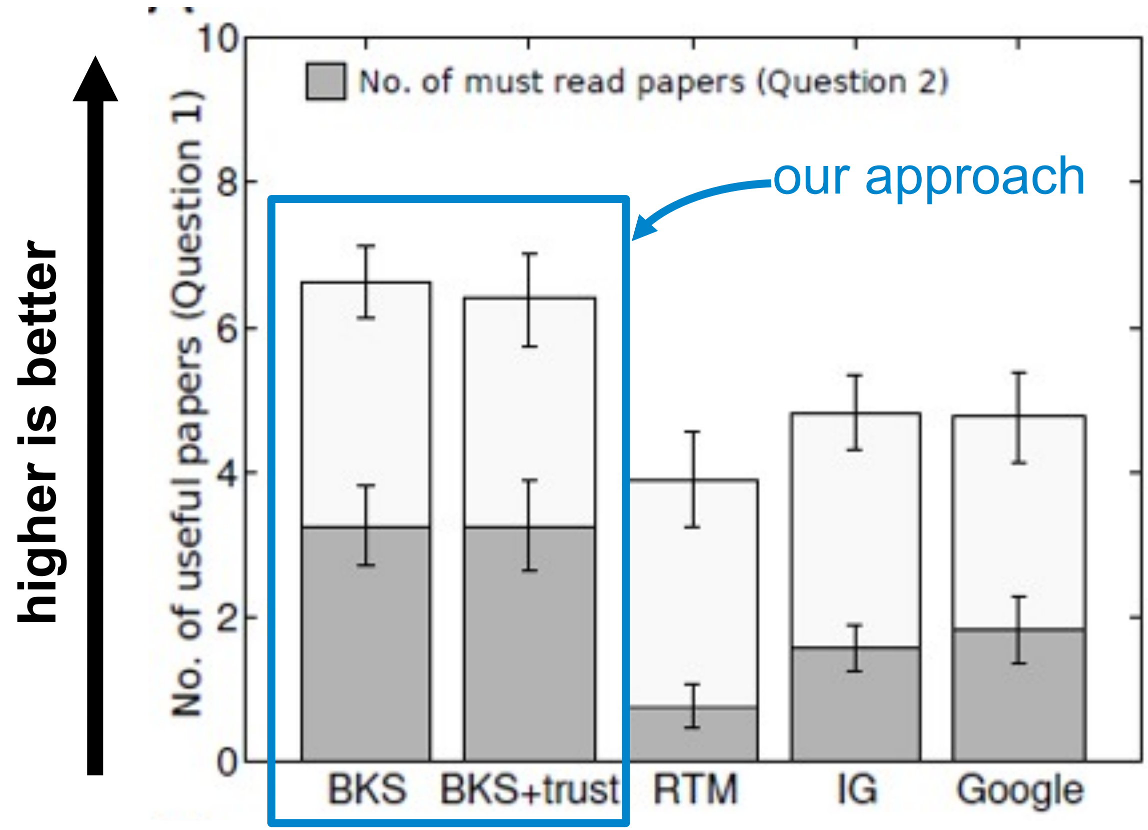
general recommendations:

personalized recommendations

User Study Evaluation

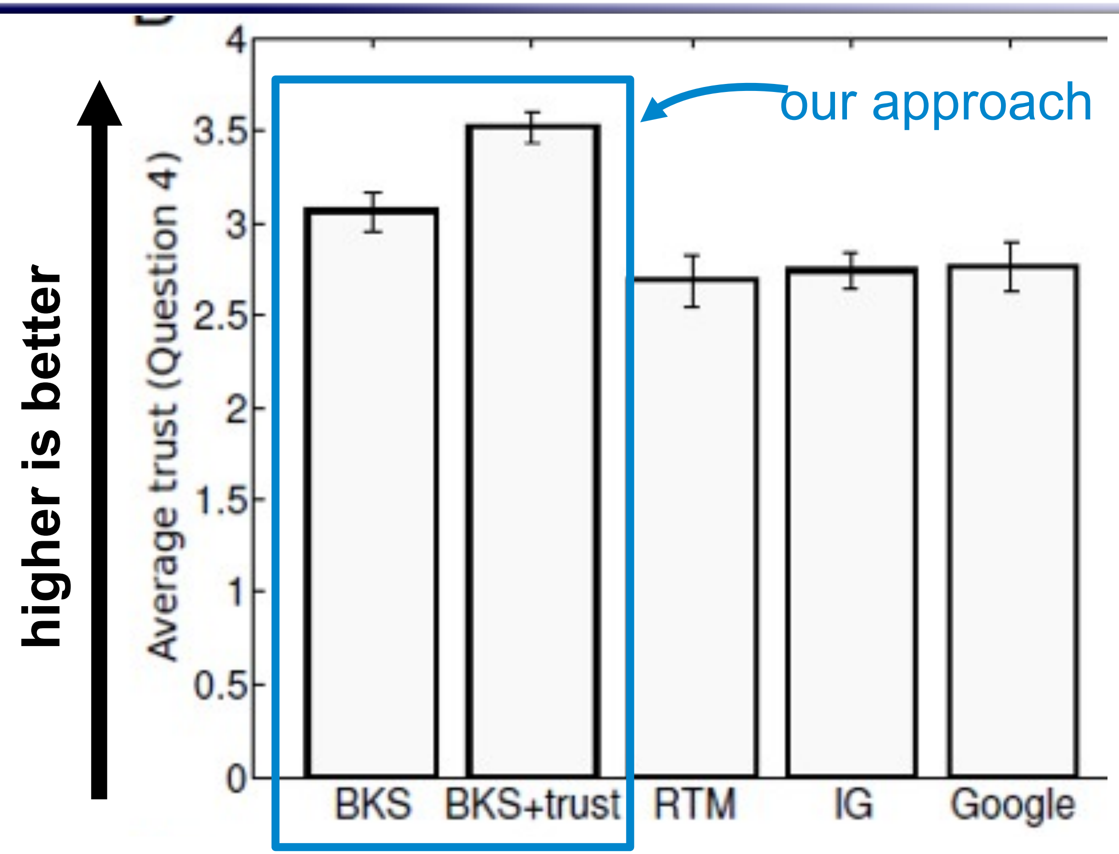
- 16 PhD students in machine learning
- For each:
  - Selected a recent publication of participant — the **study paper** — for which we find related work
  - Two variants of our methodology (w/ and w/o trust)
  - Three state-of-the-art alternatives:
    - Relational Topic Model (generative model of text and links) [Chang, Blei '10]
    - Information Genealogy (uses only document text) [Shaparenko, Joachims '07]
    - Google Scholar (based on keywords provided by coauthor)
- Double blind study where participant provided with title/author/abstract of one paper at a time, and asked several questions

Usefulness



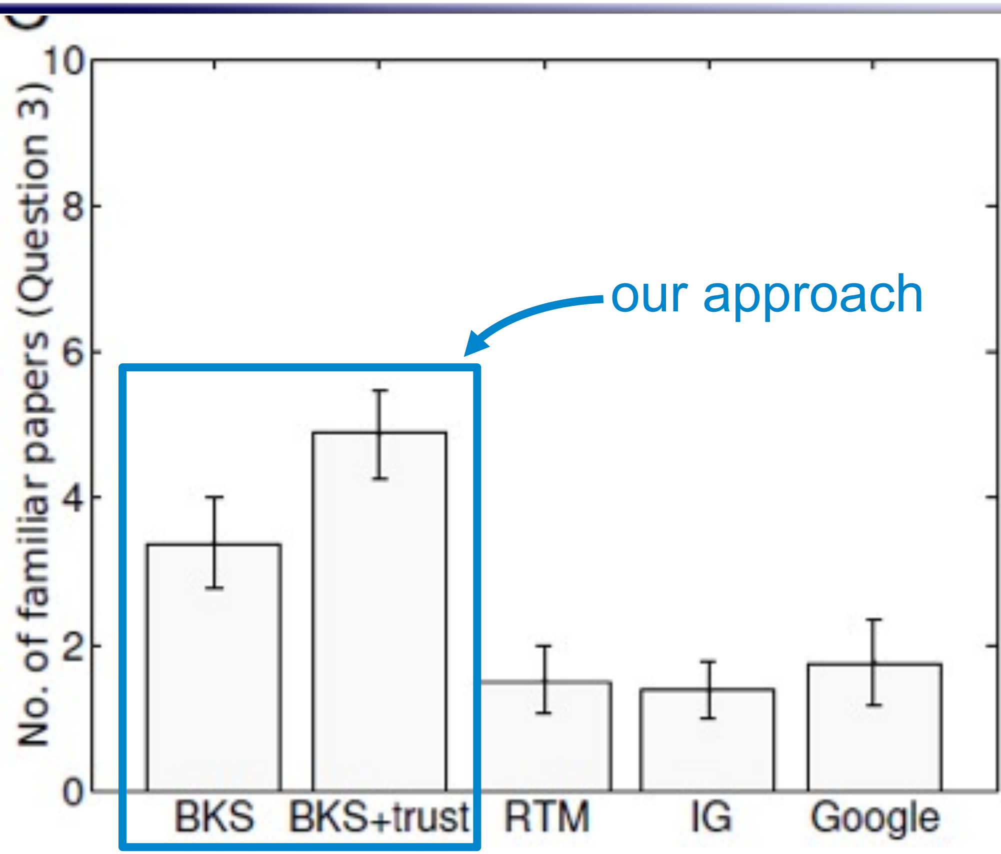
On average, our approach provides more useful and more must-read papers than comparison techniques

Trust



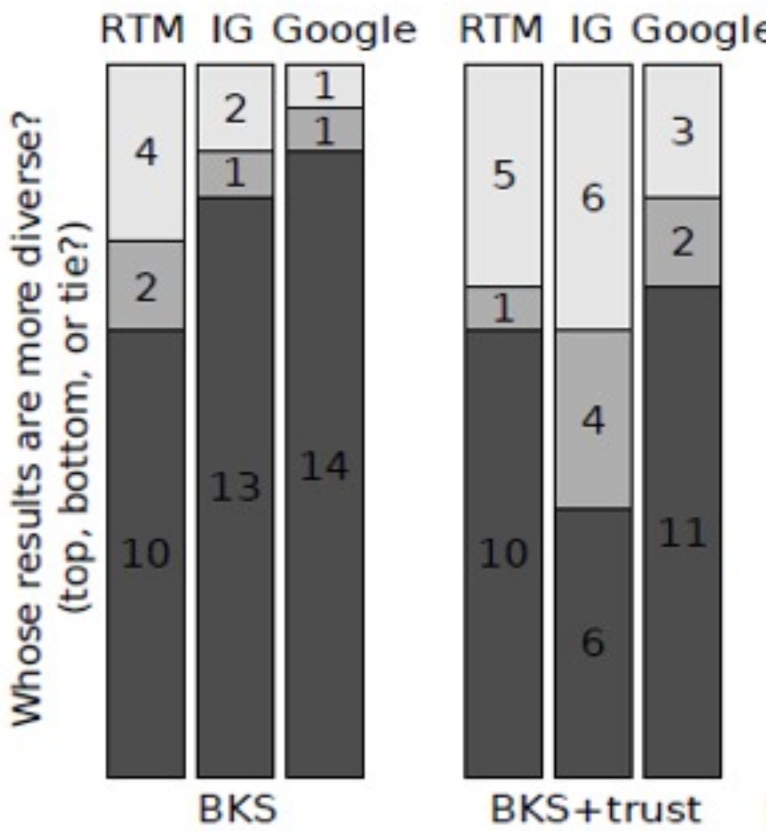
On average, our approach provides more trustworthy papers than comparison techniques, especially when incorporating participant's trust preferences

Novelty



On average, our approach provides more familiar papers than comparison techniques, especially when incorporating participant's trust preferences

Diversity



them

us

In pairwise comparison, our approaches produce more diverse results than the comparison techniques

Our Approach

Input

Phrase complex information needs

Output

Structured, annotated output

Interaction

Learn user's interests

what about structured outputs?

sense
learn
act

Connecting the Dots: News Domain

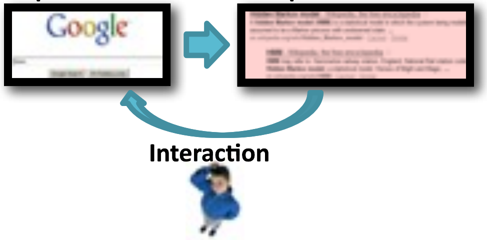
The New York Times

ECONOMIC SCENE

Can't Grasp Credit Crisis? Join the Club

3.19.2008

Housing Bubble

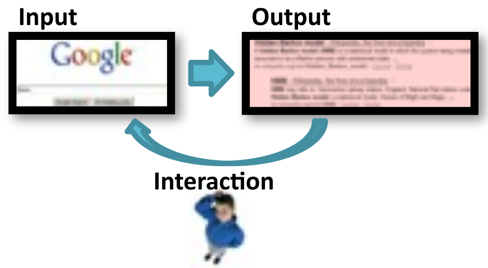


Input: Pick two articles

(start, goal)

Bailout

Housing Bubble



- Keeping

Input: Pick two articles
(start, goal)
Output: Bridge the gap
with a smooth chain of articles

- Bailout F al

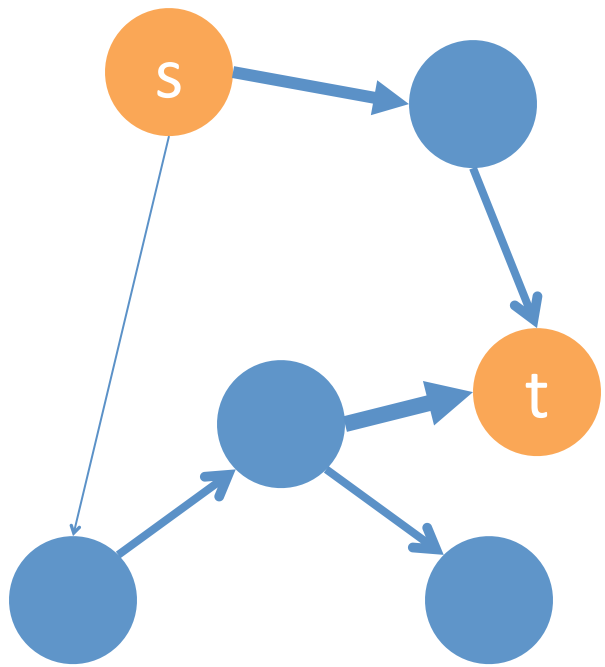
Bailout

Connecting the Dots [Shahaf, G. '10]

What is a Good Chain?

- What's wrong with shortest-path?

- Build a graph
  - Node for every article
  - Edges based on similarity
    - Chronological order (DAG)
  - Run BFS



sense learn act

Shortest-path

- A1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

Lewinsky

- A2: Judge Sides with the Government in **Microsoft Antitrust Trial**
- A3: Who will be the Next **Microsoft**?
  - trading at a **market** capitalization…
- A4: **Palestinians** Planning to Offer Bonds on **Euro. Markets**
- A5: Clinton Watches as **Palestinians Vote** to Rescind 1964 Provision
- A6: Contesting the Vote: The Overview; Gore asks Public For Patience;

Florida recount

Shortest-path

- A1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down
- A2: Judge Sides with the Government in **Microsoft Antitrust Trial**
- A3: Who will be **the Next Microsoft**?
  - trading at a **market capitalization**…
- A4: **Palestinians** Planning to Offer **Bonds on Euro. Markets**
- A5: Clinton Watches as **Palestinians Vote to Rescind** 1964 Provision
- A6: Contesting the Vote: The Overview; Gore asks Public For Patience;

Shortest-path

- A1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down
- A2: Juc **al**
- A3: W
  - tradi
- A4: **Palestinians** Planning to Offer Bonds on Euro. Markets
- A5: Clinton Watches as **Palestinians Vote** to Rescind 1964 Provision
- A6: Contesting the Vote: The Overview; Gore asks Public For Patience;

Stream of consciousness?

- Each transition is strong

- No global theme

More-Coherent Chain

- B1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

Lewinsky

- B2: **Clinton Admits Lewinsky** Liaison to Jury
- B3: G.O.P. Vote Counter in House Predicts **Impeachment of Clinton**
- B4: **Clinton Impeached**; He Faces a Senate Trial
- B5: **Clinton's Acquittal**; Senators Talk About Their Votes
- B6: Aides Say Clinton Is Angered As **Gore Tries to Break Away**
- B7: As **Election Draws Near**, the Race Turns Mean

Florida recount

Contesting the Vote: The Overview; Gore asks Public For Patience;

More-Coherent Chain

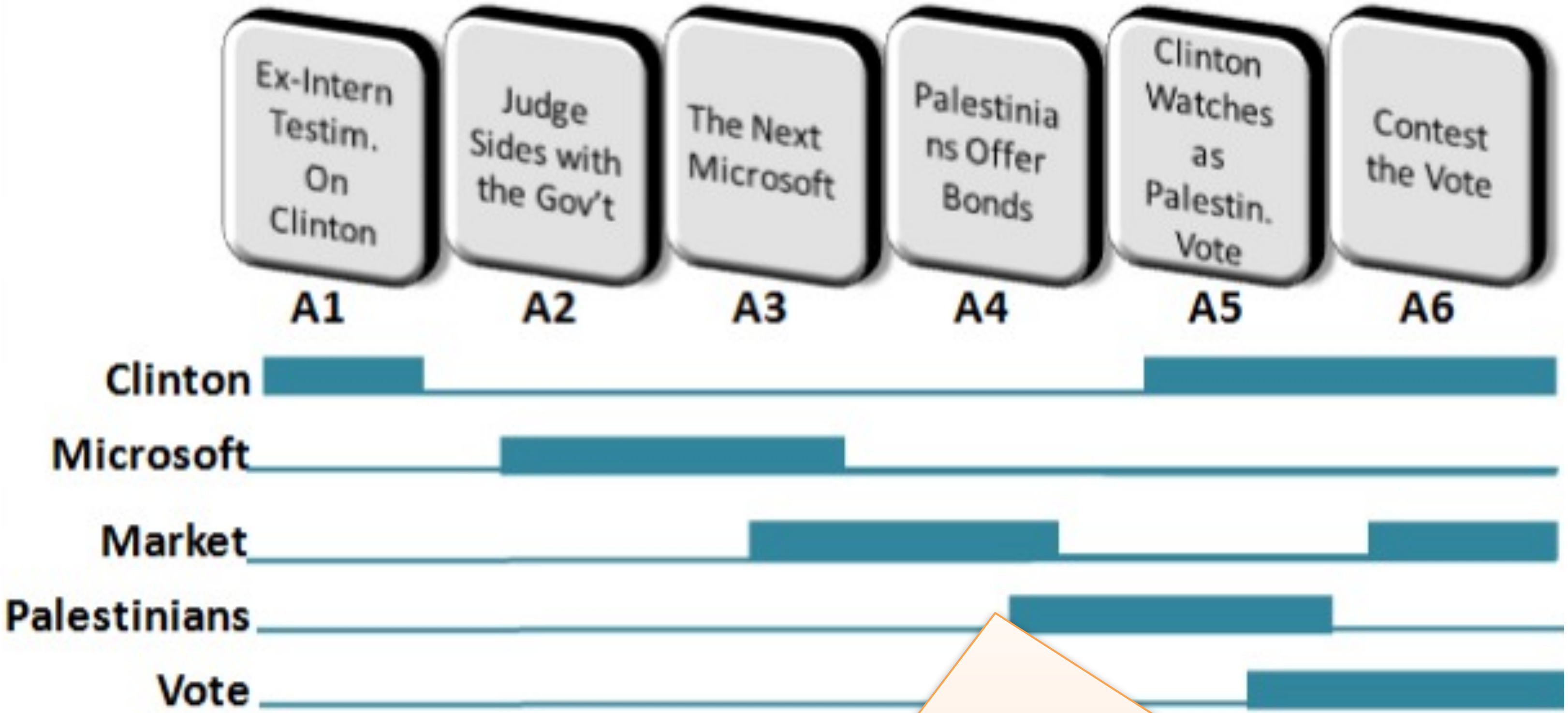
- B1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down
- B2: **Clinton Admits Lewinsky** Liaison to Jury
- B3: G.O.P. Vote Counter in House Predicts **Impeachment of Clinton**
- B4: **Cl**

What makes it coherent?

- B5: **Clinton's Acquittal**; Senators Talk About Their Votes
- B6: Aides Say Clinton Is Angered As **Gore Tries to Break Away**
- B7: As **Election Draws Near**, the Race Turns Mean
- B8: Contesting the Vote: The Overview; Gore asks Public For Patience;

Word Patterns

For Shortest Path Chain



Topic changes every transition (jittery)

Word Patterns

For Coherent Chain

Use this intuition to estimate coherence of chains

(LP-relaxation + randomized rounding)

Topic consistent over transitions

Interaction

Simpson Defense Drops DNA Challenge

Algorithmic ideas from online learning

- **Racial split** at the end

...

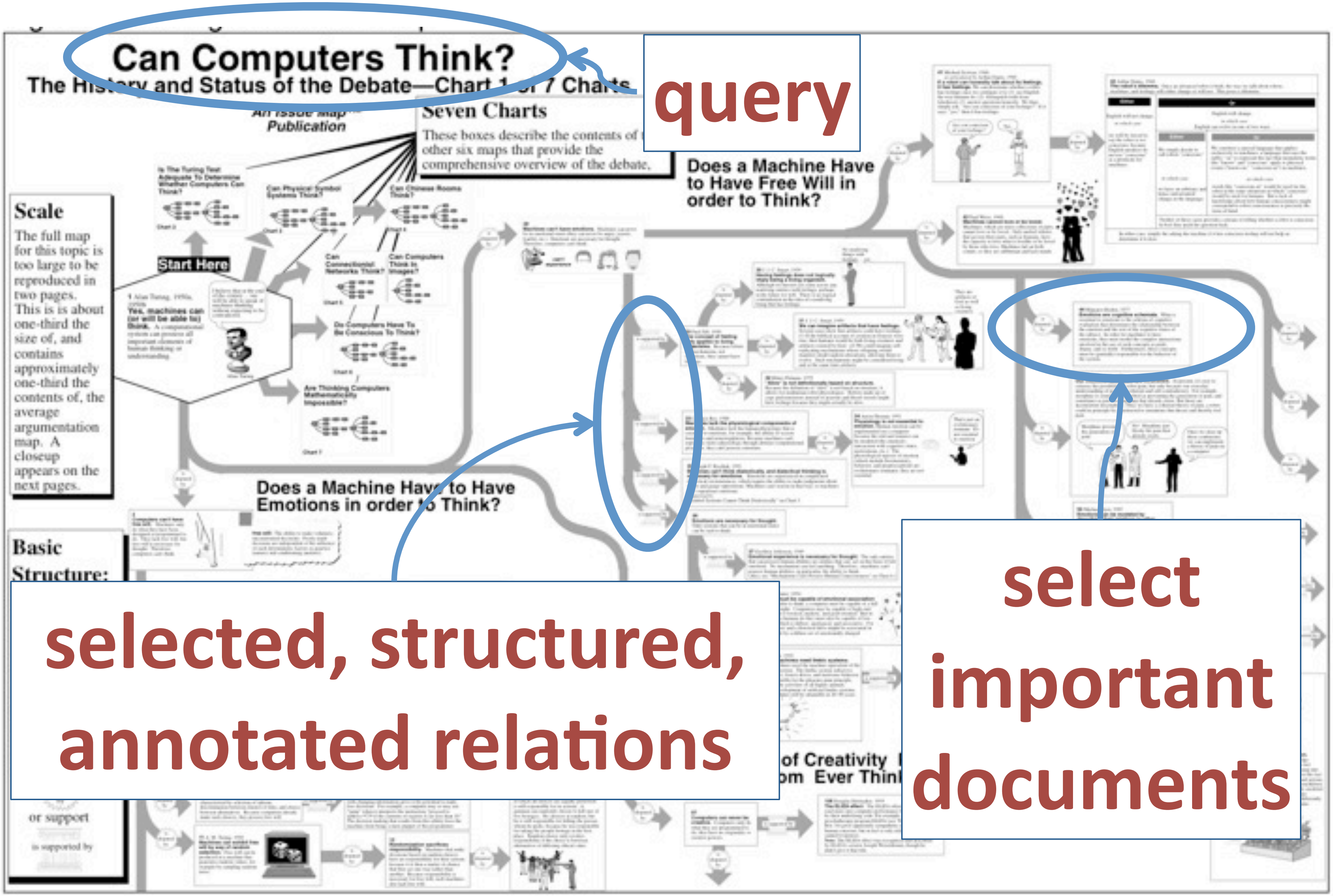
...

Simpson Verdict

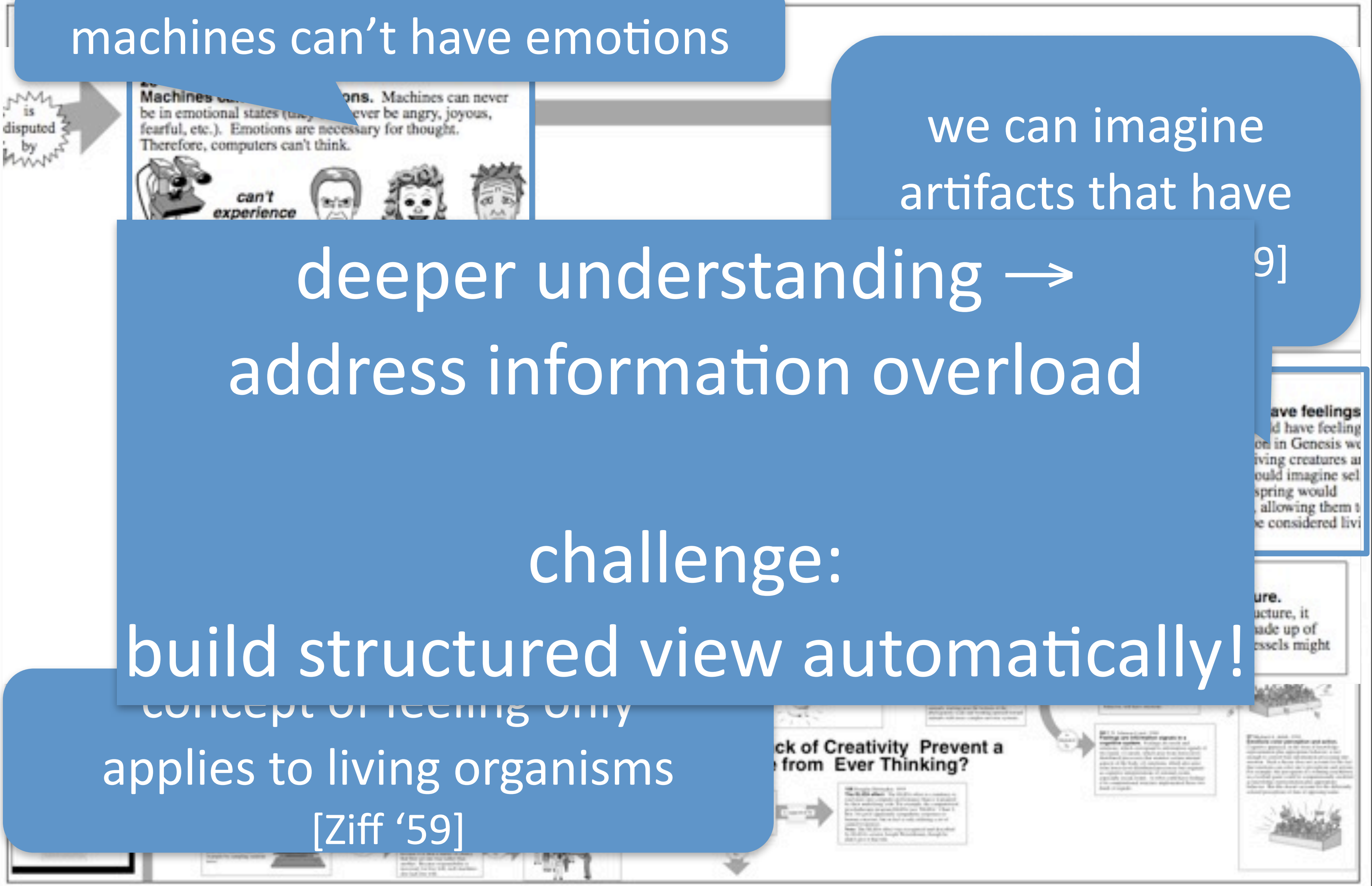
Verdict

Blood, glove

Moving Forward: Maps of Info



Moving Forward: Maps of Info



machines can’t have emotions

we can imagine artifacts that have ... 9]

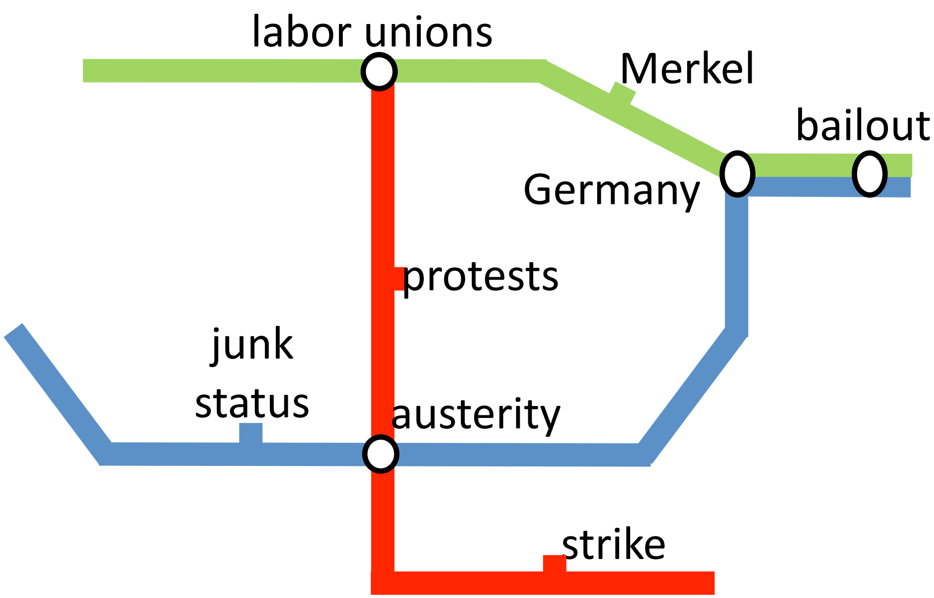
deeper understanding → address information overload

challenge:
build structured view automatically!

concept of feeling only applies to living organisms [Ziff ‘59]

Trains of Thought

- Given a set of documents
- Show important pieces of information
- … and how they relate



What makes a good map?

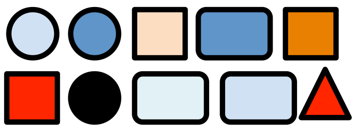
1. Coherence

2. Coverage

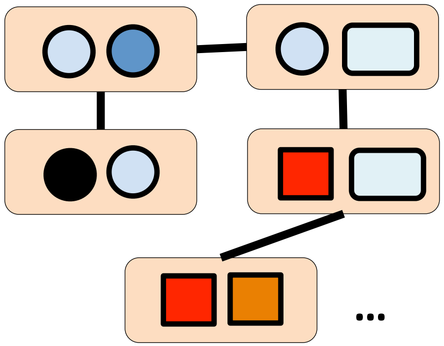
3. Connectivity

Approach overview

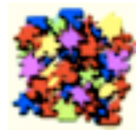
Documents D



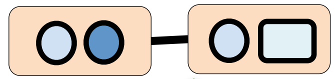
1. Coherence graph G



Encodes all coherent chains as graph paths



2. Coverage function f

f( **) = ?**

Find a set of paths that maximize coverage (submodular orienteering)

3. Find high-coverage, high-connectivity paths

First Step: Metro Maps of Science

[Shahaf, G. '12]

Example query: Reinforcement

Submodular optimization algorithm

Map of RL

Map of Science:

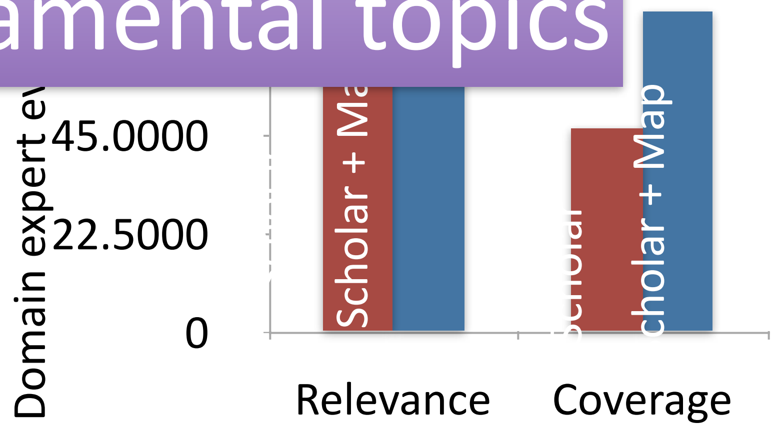
14% more relevant papers

58% more fundamental topics

[Kaelbling + al '96]

26 earlier-stage grad students

What are the most important topics and representative papers of RL today?



Taming Information Overload

set of query papers,
end points of chain

efficient algorithms with theoretical guarantees

smooth chain connecting the dots, metro maps, issue maps

Input

Phrase complex information needs

Output

Structured, annotated output

Interaction

like/dislike (online learning)
bibtex file (trust)
feedback on concepts

Learn user's inte

multiple user studies ⇒ promising direction for taming challenge of information overload