



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data Mining

Learning from Large Data Sets

Lecture 11 – Adaptive Recommendations

263-5200-00L

Andreas Krause

Course organization

- **Retrieval**

- Given a query, find “most similar” item in a large data set
- Determine relevance of search results
- *Applications:* GoogleGoggles, Shazam, ...

- **Supervised learning** (Classification, Regression)

- Learn a concept (function mapping queries to labels)
- *Applications:* Spam filtering, predicting price changes, ...

- **Unsupervised learning** (Clustering, dimension reduction)

- Identify clusters, “common patterns”; anomaly detection
- *Applications:* Recommender systems, fraud detection, ...

- **Interactive data mining**

- Learning through experimentation / from limited feedback
- *Applications:* Online advertising, opt. UI, learning rankings, ...

Sponsored search



squash rackets

Search

[Advanced Search](#)
[Preferences](#)

Web [Shopping](#)

Results 1 - 10 of about 326,000 for [squash rackets](#). (0.31 seconds)

Shopping results for [squash rackets](#)

[Slazenger Squash Racket : Xtreme Blast](#) \$27.77 - [ACA Sports](#)

[2008 - Dunlop Tempo Squash Racquet](#) \$28.95 - [SquashGear.com](#)

[Prince O3 Hybrid UltraLite Squash Racquet](#) \$99.99 - [Joe's Sports](#)


[Squash & Tennis Rackets from Just-Rackets UK and Worldwide online ...](#)

Squash, tennis, badminton, and racquetball specialist. Online retailer specialising in **rackets**, clothing, and accessories.

[justackets.com/](#) - 61k - [Cached](#) - [Similar pages](#) - 


[Squash Gear - Squash Equipment - squash racquets - squash rackets ...](#)

27 Dec 2008 ... **Squash** gear and **squash** equipment: **squash** racquets, **squash** rackets, bags, shoes, and balls from Adidas,Asics,Ashaway,Prince,Dunlop,Wilson, ...

[www.squashgear.com/](#) - 21k - [Cached](#) - [Similar pages](#) - 


[Squash Rackets, Badminton Rackets, Tennis Rackets from UK Rackets](#)

Shop for **Squash Rackets**, **Badminton Rackets** and **Tennis Racquets** within the UK.

[www.ukrackets.com/](#) - 9k - [Cached](#) - [Similar pages](#) - 

[Tennis, Badminton & Squash Rackets, Shoes, Clothing, Bags, Grips ...](#)

tennisnuts.com - the UK **racket** sports superstore, specialising in tennis, badminton and **squash**. Order on-line, mail order by ringing 0845 602 7062 or visit ...

[www.tennisnuts.com/](#) - 85k - [Cached](#) - [Similar pages](#) - 

[sportdiscount.com™ - Discounted squash rackets, badminton rackets ...](#)

Sponsored Links

[Dunlop Squash Racquets](#)

Full range of Dunlop **rackets**, balls gear. Free shipping for any order
[www.KiwiSquashGear.com](#)

[Dunlop Squash Rackets](#)

Aerogels, M-Fils, All Dunlop **Squash** Great Service and Lowest Prices
[www.JustSquash.com](#)

[Squash Racket Specialist](#)

Dunlop Aerogel 2 for 1 offers
ProGT Ultimate Lee Beachhill Ice
[www.just-rackets.co.uk](#)

[Squash Rackets](#)

Alles rund ums Thema Outdoor günstig aus privater Hand
[www.kijiji.de](#)

[Squash Rackets?](#)


Top Preise: Squashschläger. Jetzt Günstig **Squash rackets** kaufen!
[www.evita.de/Squash+Rackets](#)


Which ads should be displayed to maximize revenue?


Which news should we display?


The screenshot shows the Yahoo! News homepage. At the top is the "YAHOO! NEWS" logo and a search bar. Below the logo is a navigation bar with categories: HOME, U.S., WORLD, BUSINESS, ENTERTAINMENT, SPORTS, TECH, POLITICS, SCIENCE, and HEALTH. Underneath the navigation bar is a "Top Stories" section with several tabs: "Top Stories", "ABC News", "Latest News", "Slideshows", "AP", "Reuters", and "AFP". The main content area displays four news items, each with a thumbnail image, a headline, a source, and a timestamp.

Top Stories | [ABC News](#) | [Latest News](#) | [Slideshows](#) | [AP](#) | [Reuters](#) | [AFP](#)

 **Everest weekend death toll reaches 4** AP - 2 hrs 7 mins ago
Climbers have reported seeing another body on Mount Everest, raising the death toll to four for one of the worst days ever on the world's highest mountain. [More »](#)

 **Colombia Secret Service prostitution scandal spreads to DEA** ABC News - 8 hrs ago
The Drug Enforcement Administration announced that at least three of its agents are under investigation for allegedly hiring prostitutes in Cartagena. [More »](#)

 **Obama: U.S. can't wait for Afghanistan to be 'perfect'** The Ticket - 7 hrs ago
President Obama acknowledged "risks" in his decision to withdraw U.S. combat forces from Afghanistan by the end of 2014 but said war-weary Americans can't wait for that strife-torn country to be "perfect." [More »](#)

 **Why ex-Rutgers student got 30-day sentence in spycam case** Christian Science Monitor - 9 hrs ago
A former Rutgers University student was sentenced to serve 30 days in jail in a case of webcam spying that drew national attention to issues of online privacy, suicide, and anti-gay bullying. [More »](#)

The UCB1 algorithm [Auer et al '02]

- Set $\hat{\mu}_1 = \dots = \hat{\mu}_k = 0$ $n_1 = \dots = n_k = 0$
- For $t = 1:T$
 - For each arm i calculate $UCB(i) = \hat{\mu}_i + \sqrt{\frac{2 \ln t}{n_i}}$
 - Pick arm $j = \arg \max_i UCB(i)$ and observe y_t
 - Set $n_j \leftarrow n_j + 1$ and $\hat{\mu}_j \leftarrow \hat{\mu}_j + \frac{1}{n_j} (y_t - \hat{\mu}_j)$
- “Optimism in the face of uncertainty”

Performance of UCB

- Theorem [Auer et al 2002]

- Suppose the optimal mean payoff is $\mu^* = \max_i \mu_i$

and for each arm let $\Delta_i = \mu^* - \mu_i$

- Then it holds that

$$\mathbb{E}[R_T] = \underbrace{\left[8 \sum_{i: \mu_i < \mu^*} \left(\frac{\ln T}{\Delta_i} \right) \right]}_{O(k \ln T)} + \underbrace{\left(1 + \frac{\pi^2}{3} \right) \left(\sum_{i=1}^k \Delta_i \right)}_{O(\epsilon)}$$

$$\Rightarrow O\left(\frac{R_T}{T}\right) = \left(\frac{k \ln T}{T}\right)$$

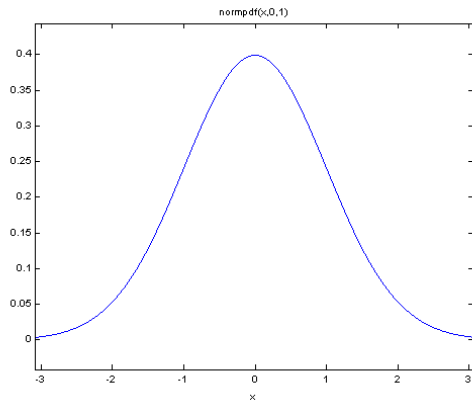
Challenges in recommendation

- Number of recommendations k to choose from large
 - Similar ads → similar click-through rates!
- Performance depends on query / context
 - Similar queries → similar click-through rates!
- Need to compile sets of k recs. (instead of only one)
 - Similar sets → similar click-through rates!
- **Need to model and exploit “similarity”**

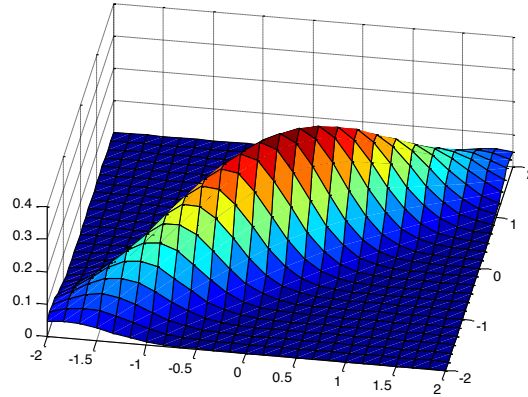
Stochastic ∞ -armed bandits

- (Possibly infinite) Set X of choices
- Class F of functions on X
- Each choice x in X associated with (unknown) probability distribution P_x supported in $[0,1]$ with means $\mu_x = f(x)$ for some $f \in F$
- Play game for T rounds
- In each round t , we pick an arm x , and obtain a random sample Y_t from P_x independent of previous samples
- Our goal is to maximize $\sum_{t=1}^T Y_t$

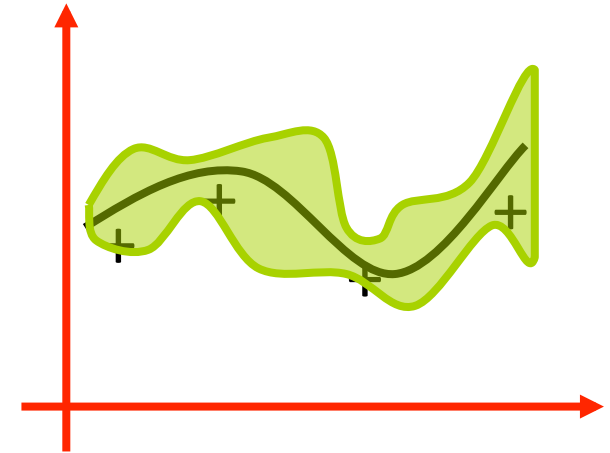
Gaussian Processes to model payoff f



Normal dist.
(1-D Gaussian)



Multivariate normal
(n-D Gaussian)



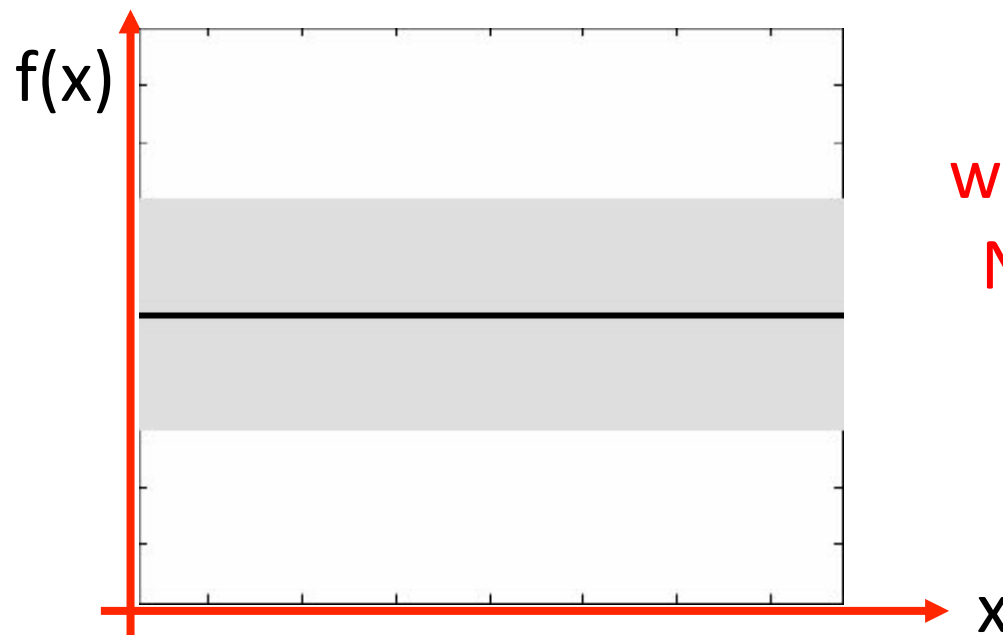
Gaussian process
(∞ -D Gaussian)

- **Gaussian process (GP)** = normal distribution over *functions*
- Finite marginals are multivariate Gaussians
- Closed form formulae for Bayesian posterior update exist
- Parameterized by *covariance function* $K(x, x') = \text{Cov}(f(x), f(x'))$

Upper confidence sampling

Pick input that maximizes upper confidence bound:

$$x_t = \arg \max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$



How should we choose β_t ??
Need theory!

Naturally trades off exploration and exploitation

Does not waste samples (with high prob.)

Guarantees for GP-UCB

Theorem: [Srinivas et al, ICML '10]

Choose $\beta_t = O(\log t)$. Then, for the following kernels:

- Linear: $\frac{R_T}{T} = \mathcal{O}^* \left(\frac{d}{\sqrt{T}} \right)$
- Squared-exponential: $\frac{R_T}{T} = \mathcal{O}^* \left(\frac{(\log T)^{d+1}}{\sqrt{T}} \right)$
- Matérn with $\nu > 2$, $\frac{R_T}{T} = \mathcal{O}^* \left(T^{\frac{\nu+d(d+1)}{2\nu+d(d+1)} - 1} \right)$

Bandits for recommendation

- Number of recommendations k to choose from large
 - Similar ads → similar click-through rates!




- Performance depends on query / context
 - Similar queries → similar click-through rates!


- Need to compile sets of k recs. (instead of only one)
 - Similar sets → similar click-through rates!


Which news should we display?


The screenshot shows the Yahoo! News homepage. At the top is the "YAHOO! NEWS" logo and a search bar. Below the logo is a navigation bar with categories: HOME, U.S., WORLD, BUSINESS, ENTERTAINMENT, SPORTS, TECH, POLITICS, SCIENCE, and HEALTH. Underneath the navigation bar is a "Top Stories" section with several tabs: Top Stories, ABC News, Latest News, Slideshows, AP, Reuters, and AFP. The main content area displays four news items, each with a thumbnail image, a headline, a source, and a timestamp.

Top Stories | [ABC News](#) | [Latest News](#) | [Slideshows](#) | [AP](#) | [Reuters](#) | [AFP](#)

 **Everest weekend death toll reaches 4** AP - 2 hrs 7 mins ago
Climbers have reported seeing another body on Mount Everest, raising the death toll to four for one of the worst days ever on the world's highest mountain. [More »](#)

 **Colombia Secret Service prostitution scandal spreads to DEA** ABC News - 8 hrs ago
The Drug Enforcement Administration announced that at least three of its agents are under investigation for allegedly hiring prostitutes in Cartagena. [More »](#)

 **Obama: U.S. can't wait for Afghanistan to be 'perfect'** The Ticket - 7 hrs ago
President Obama acknowledged "risks" in his decision to withdraw U.S. combat forces from Afghanistan by the end of 2014 but said war-weary Americans can't wait for that strife-torn country to be "perfect." [More »](#)

 **Why ex-Rutgers student got 30-day sentence in spycam case** Christian Science Monitor - 9 hrs ago
A former Rutgers University student was sentenced to serve 30 days in jail in a case of webcam spying that drew national attention to issues of online privacy, suicide, and anti-gay bullying. [More »](#)

LinUCB for personalized recommendation

- Every round receive **context** [Li et al WWW '10]
 - User features (e.g., articles viewed before, ...)

- Linear model for each article's click through rate

More generally: Contextual bandits

In each round t do:

- Observe context $z_t \in Z$
- Pick $x_t \in D$
- Observe $y_t = f(x_t, z_t) + \epsilon_t$
- Incur regret

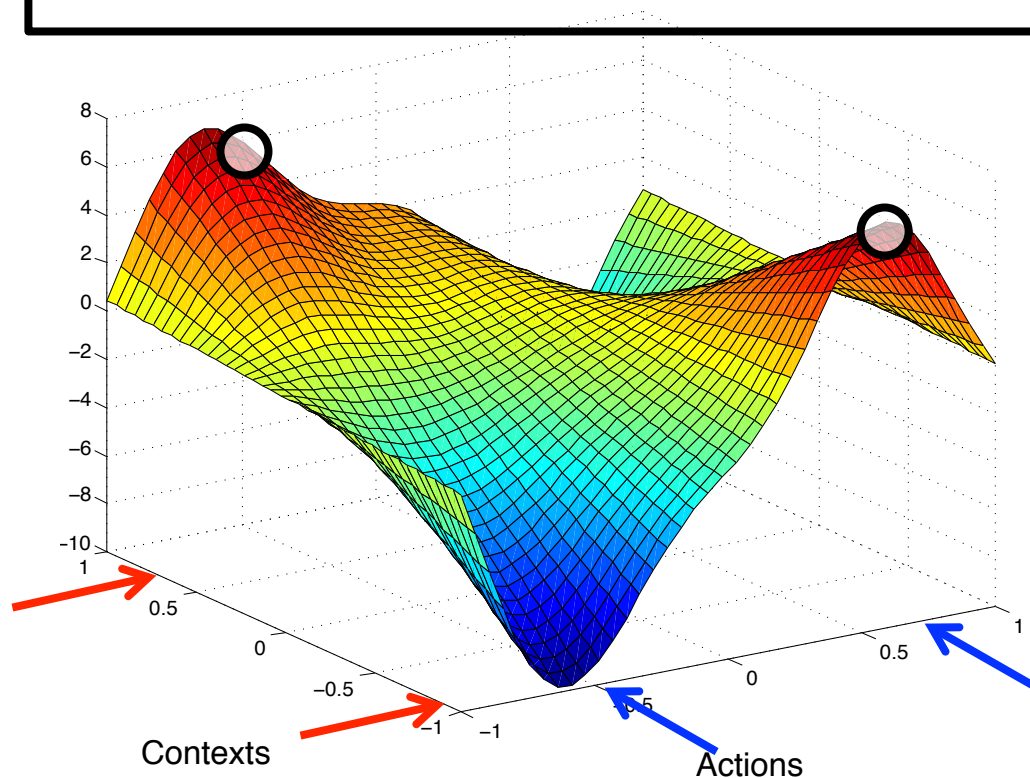
- Cumulative **contextual** regret $R_t = \sum_{t=1}^T r_t$

- Obtaining sublinear regret requires learning **optimal mapping** from contexts to actions!

CGP-UCB

Pick input that maximizes upper confidence bound
at current context

$$x_t = \arg \max_{x \in D} \mu_{t-1}(x, z_t) + \beta_t \sigma_{t-1}(x, z_t)$$



Similar regret bounds as in «context-free» setting.

Algorithm 1 LinUCB with disjoint linear models.

0: Inputs: $\alpha \in \mathbb{R}_+$

[Li et al '10]

1: **for** $t = 1, 2, 3, \dots, T$ **do**

2: Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$

3: **for all** $a \in \mathcal{A}_t$ **do**

4: **if** a is new **then**

5: $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d -dimensional identity matrix)

6: $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$ (d -dimensional zero vector)

7: **end if**

8: $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$

9: $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$

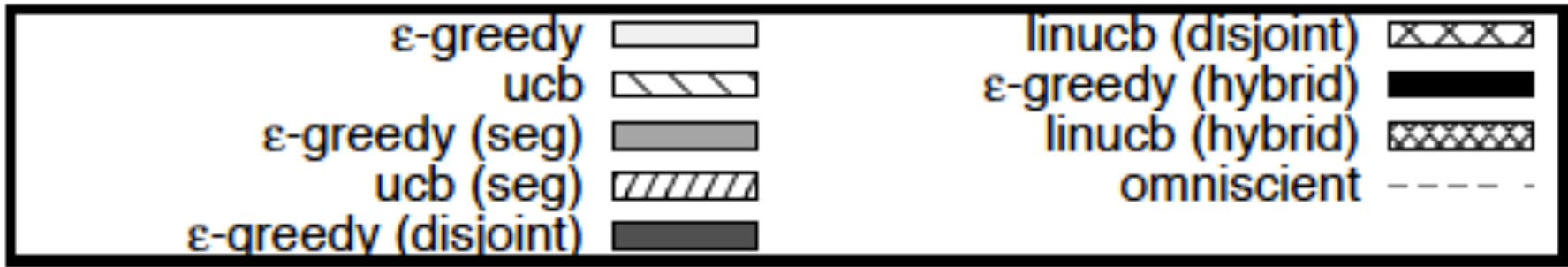
10: **end for**

11: Choose arm $a_t = \arg \max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t

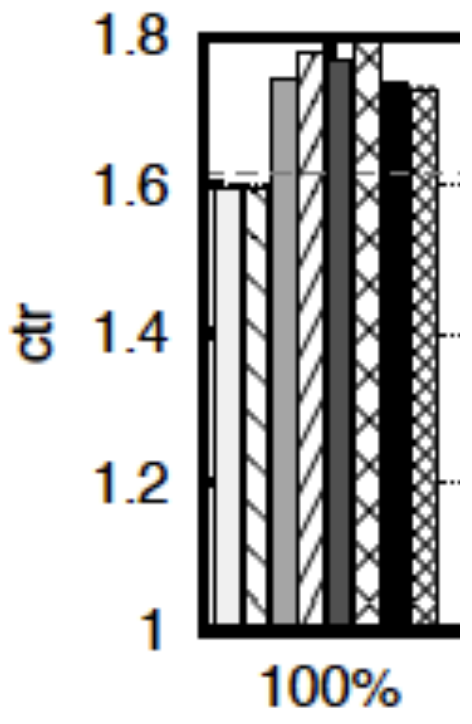
12: $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$

13: $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$

14: **end for**



[Li et al WWW '10]



(a) CTRs in the deployment bucket.

Bandits for recommendation

- Number of recommendations k to choose from large

- Similar ads → similar click-through rates!



- Performance depends on query / context

- Similar queries → similar click-through rates!



- Need to compile sets of k recs. (instead of only one)

- Similar sets → similar click-through rates!

Sponsored search

The image shows a Google search results page for the query "cars". The search bar at the top contains "cars" and a "Search" button. Below the search bar, it indicates "About 1,440,000,000 results (0.15 seconds)" and a link to "Advanced search".

On the left side, there are navigation options: "Everything", "Images", "Videos", "News", "Shopping", and "More". Below these are filters for "Any time" (Latest, Past 24 hours, Past week, Past 2 weeks, Past month, Past year, Custom range...) and "All results" (Sites with images, Wonder wheel, Related searches, More search tools). At the bottom left, there are suggestions for "Something different" (bikes, boats, tractors, jeeps, honda).

The main search results are listed in the center. The first result is "New & Used Cars for Sale, Auto Dealers, Car Reviews and Car ...". The second is "Auto Trader UK - Used cars for sale". The third is "Used Cars - New Cars - Search New & Used Cars For Sale - carsales ...". The fourth is "Cars (2006) - IMDb". The fifth is "Cars 2 | Disney". The sixth is "New Cars, Used Cars - Find Cars at AutoTrader.com".

On the right side, there is a red-bordered box labeled "Ads" containing five sponsored search results:

- Cars**: Electric car Think City at m-way: Economic, safe & eco-friendly www.m-way.ch/cars
- Suzuki online**: Suzuki fahren, Treibstoff sparen. Jetzt Probefahrt vereinbaren! www.suzukiautomobile.ch/cars
- Cars**: Spielwaren bei Brack.ch bis 15 Uhr bestellt - Versand noch heute! www.brack.ch/Spielwaren
- Cars in Switzerland**: Find Cars in Switzerland. Browse through ads now! www.tutti.ch/Vehicles
- Second Hand Cars**: Participate in Used Car Auctions In The USA. Cars Shipped Worldwide! exporttrader.com is rated ★★★★☆ www.exporttrader.com

At the bottom of the red box, there is a link: "See your ad here »".

Which set of ads should be displayed to maximize revenue?

News recommendation

Google news

Search News

Search the Web

Advanced news search

U.S. edition ▼ Add a section »

Top Stories

Dirk Nowitzki
Oprah Winfrey
American Idol
Wells Fargo
Kim Kardashian
Michael Brown
Yemen
Arizona
Congresswoman
Gabrielle Giffords
Steve Ballmer
Boston Bruins

Starred ☆

World
Zurich, Switzerland
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight

All news

Headlines
Images

Top Stories



CBC.ca

Budget tactics spark Senate GOP divide

Politico - 37 minutes ago

Mitch McConnell says he won't pressure colleagues to vote for Paul Ryan's budget plan. | AP Photo Close By MANU RAJU & JENNIFER HABERKORN | 5/26/11 4:36 AM EDT Senate Republicans are scrambling to regain the political edge in the budget debate, ...

Gingrich stops short of backing Ryan's Medicare plan
Boston Globe

The Note's Must-Reads for Thursday, May 26th, 2011
ABC News (blog)

See all 2,357 sources »



Related
Paul Ryan »



Globe and ...

Many searching for loved ones in tornado-ravaged Joplin

CNN International - 1 hour ago

By the CNN Wire Staff Joplin, Missouri (CNN) -- Authorities in hard-hit Joplin on Thursday will release a list of people unaccounted for as the community deals with the aftermath of a tornado that tore homes and families apart.



Fox News

Is Sheriff Arpaio's political future at stake?

AZFamily - 4 hours ago

PHOENIX - In recent month the Maricopa County Sheriff's Office has been accused of mispending nearly \$100 million, allowing corruption and mismanagement to run rampant within the organization and failing to investigate hundreds of ...



ABC News

How Scotty McCreery Won 'American Idol'

ABC News - 2 hours ago

Scotty McCreery wins "American Idol," after an evening of star-studded performances by Beyonce, Lady Gaga, Judas Priest, Tom Jones, Tony Bennett, and Bono singing alongside "Idol" contestants.

Recent

28 dead, 700 flee Mexico gang battles

USA Today - 22 minutes ago

Storm-Wearry Midwest Hoping to Dodge More Devastation

Fox News - 45 minutes ago

World stocks rise from two-month low; euro firmer

Reuters - 56 minutes ago

Zurich, Switzerland » - Edit

FC Basel retain Swiss title, St Gallen relegated

Los Angeles Times - 12 hours ago

Accelero XTREME Plus II - Higher versatility. Greater compatibility

HEXUS - 18 hours ago

Starlets out to shine in Zurich

Fifa.com - May 24, 2011

Spotlight »

Harold Camping "Judgement Day": Christian radio mogul nowhere to be found ...

Slate Magazine - May 22, 2011

Why did no one notice the boy was missing?

CNN International - May 22, 2011

Where are India's millions of missing girls?

BBC News - May 20, 2011

Fox News Chief Roger Ailes Thinks Sarah Palin Is 'Stupid': New York Magazine

Huffington Post - May 22, 2011

The Stone: The Flight of Curiosity

New York Times (blog) - May 22, 2011

Meet the blackjack player who beat the Trop for \$6 million, Borgata for \$5 and ...

Press of Atlantic City - May 22, 2011

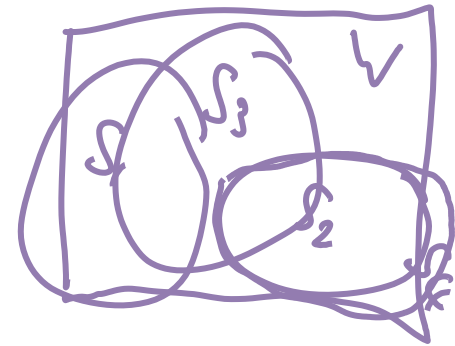
Relevance vs. Diversity

- Want to choose a set that caters to as many users as possible
- Users may have different interests / queries may be ambiguous
- Want to optimize both relevance and diversity

Simple abstract model

- Suppose we're given a set W of users and a collection V of ads / documents
- Each ad i is relevant to a set of users $S_i \subseteq W$
- For each set of ads define

$$F(A) = \left| \bigcup_{i \in A} S_i \right|$$



- Want to select k ads to maximize “users covered”

$$\max_{|A| \leq k} F(A)$$

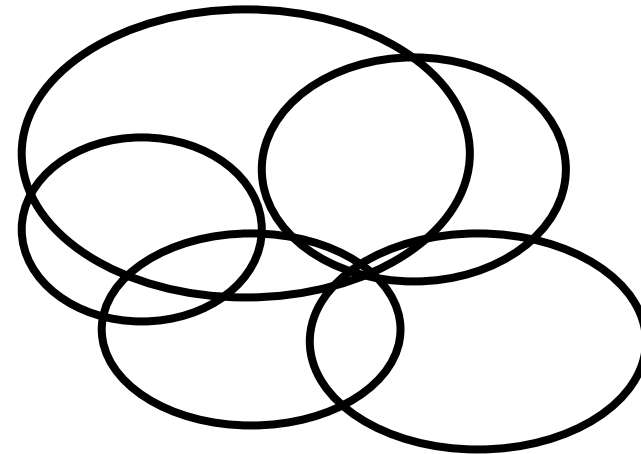
- Number of sets A grows exponential in k !
- Finding optimal A is NP-hard ☹️

Maximum coverage

- Given: Collection V of sets, utility function $F(A)$
Want: $A^* \subseteq V$ such that

$$A^* = \operatorname{argmax}_{|\mathcal{A}| \leq k} F(\mathcal{A})$$

NP-hard!



Greedy algorithm:

Start with $A_0 = \{\}$

For $i = 1$ to k

$$s^* = \operatorname{argmax}_s F(A \cup \{s\})$$

$$A_i = A_{i-1} \cup \{s^*\}$$

How well does
this simple
heuristic do?

Approximation guarantee

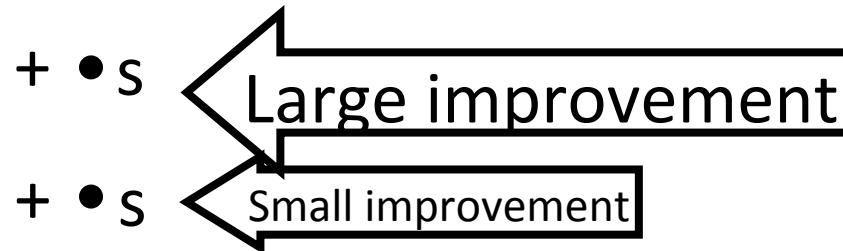
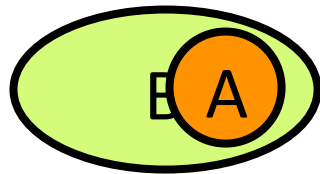
- **Theorem:** Greedy algorithm produces a solution A where $F(A) \geq (1-1/e)$ of optimal value (~63%)
[Nemhauser, Fisher, Wolsey '78]

- Claim holds for functions F with 2 properties:
 - F is **monotone**:
if $A \subseteq B$ then $F(A) \leq F(B)$ and $F(\{\})=0$
 - F is **submodular**:
adding element to a set gives less improvement than adding to one of subsets

Submodularity

- Diminishing returns
- Set function F on V is called **submodular** if:
For all $A \subseteq B$, $s \notin B$:

$$\underbrace{F(A \cup \{s\}) - F(A)}_{\text{Gain of adding } s \text{ to a small set}} \geq \underbrace{F(B \cup \{s\}) - F(B)}_{\text{Gain of adding } s \text{ to a large set}}$$



Example: Set cover

- F is submodular: $A \subseteq B$

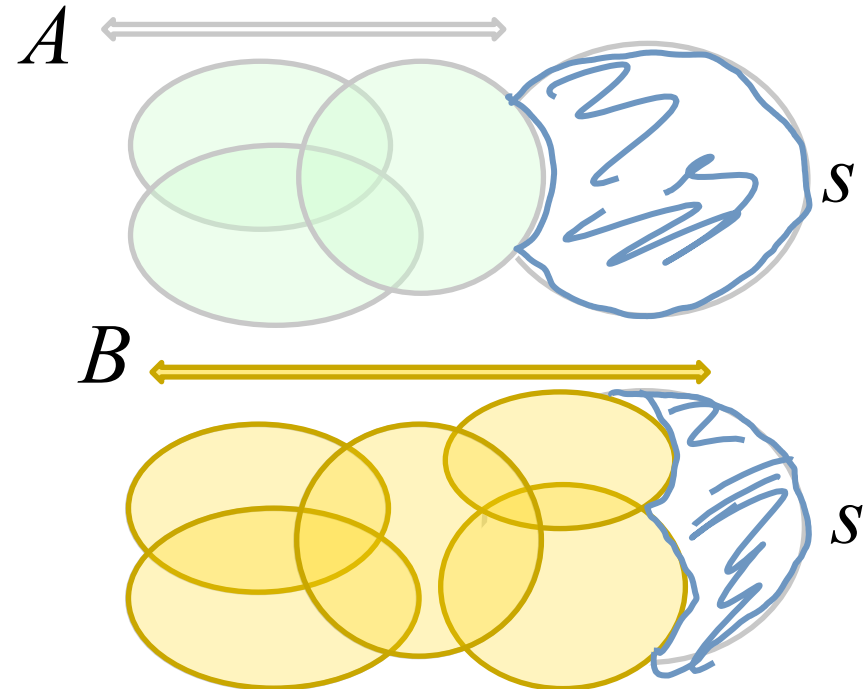
$$\underbrace{F(A \cup \{s\}) - F(A)}_{\text{Gain of adding a set } s \text{ to a small solution}} \geq \underbrace{F(B \cup \{s\}) - F(B)}_{\text{Gain of adding a set } s \text{ to a large solution}}$$

Gain of adding a set s to a small solution

Gain of adding a set s to a large solution

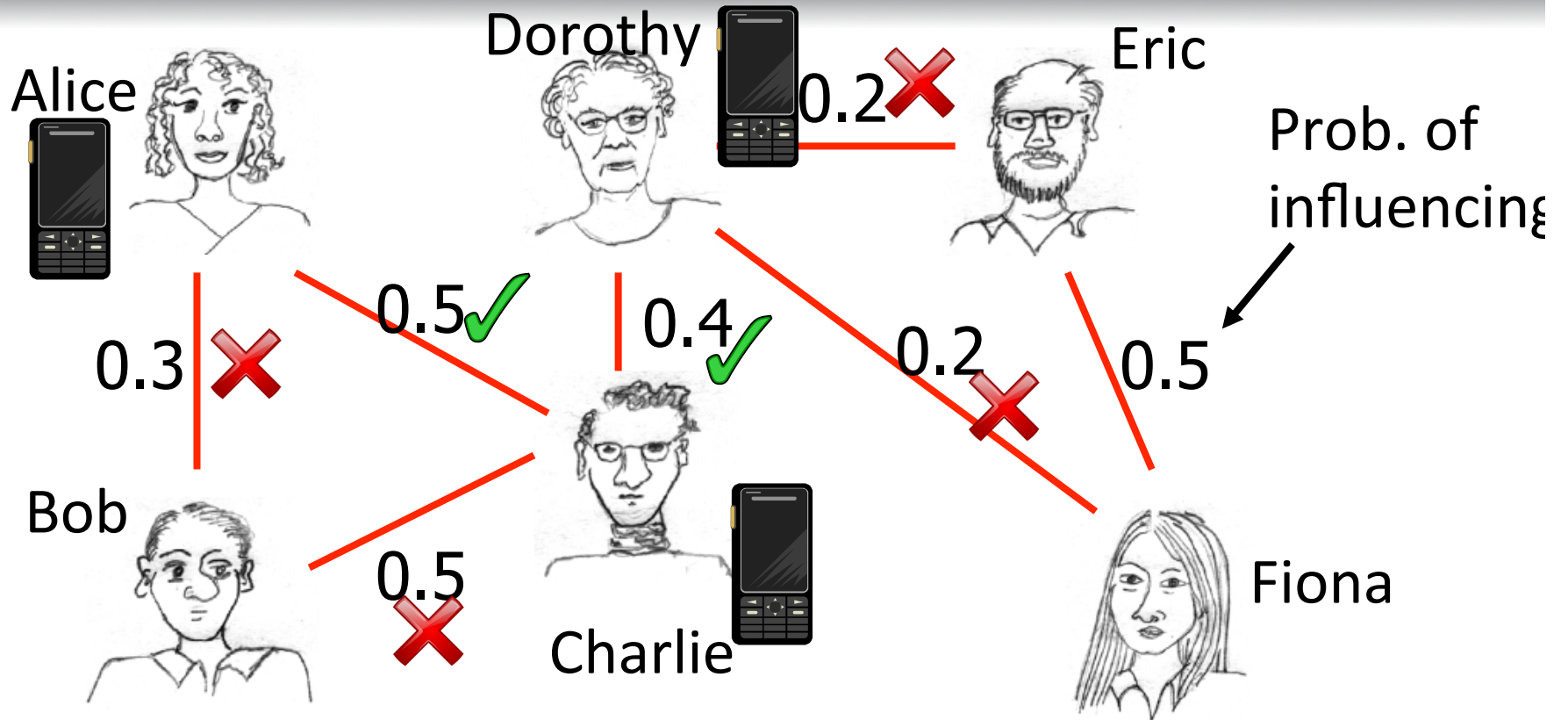
- Natural example:

- Sets s_1, s_2, \dots, s_n
- $F(A)$ = size of union of s_i
(size of covered area)



Example: Influence in social networks

[Kempe, Kleinberg, Tardos KDD '03]



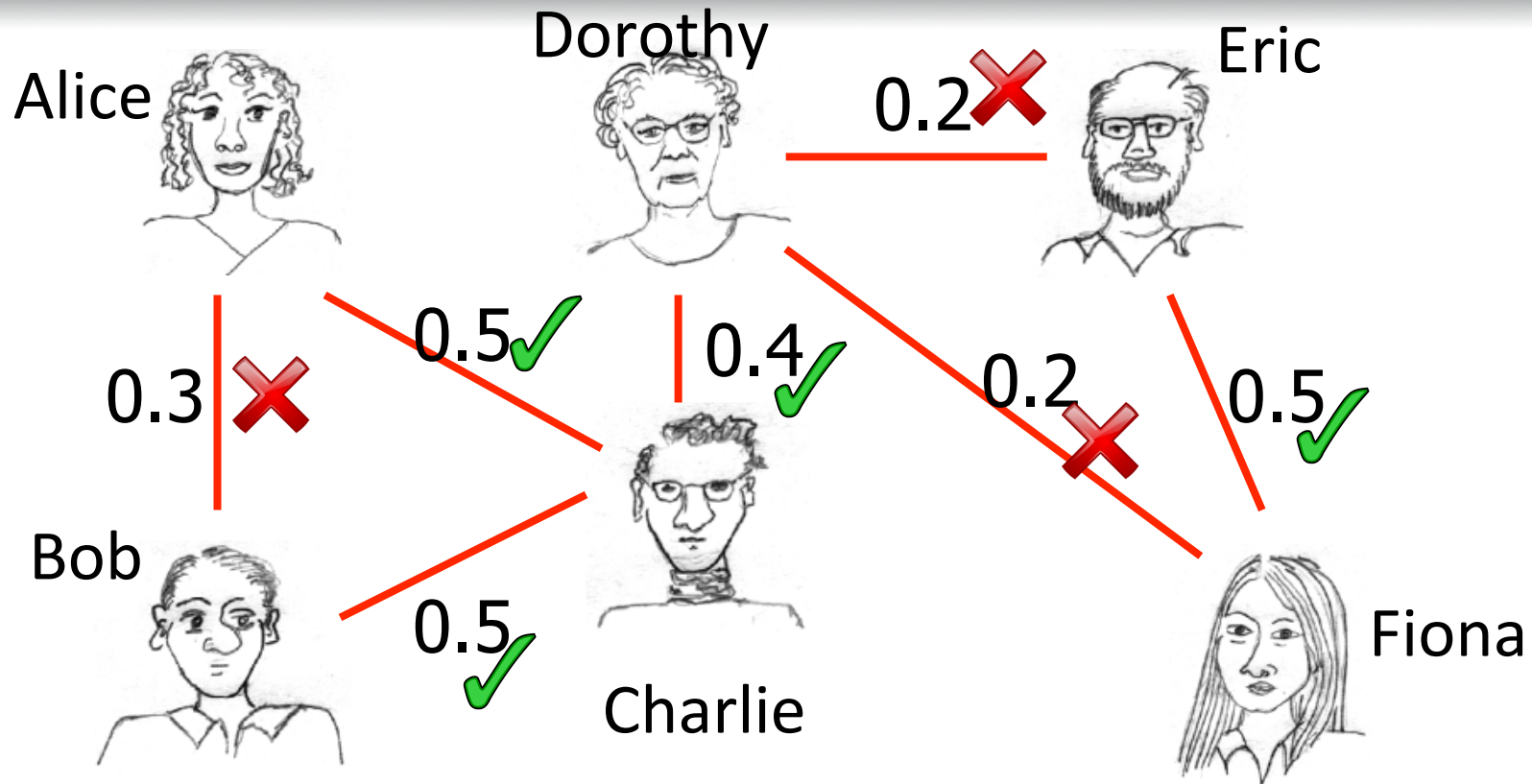
Who should get free cell phones?

$V = \{\text{Alice, Bob, Charlie, Dorothy, Eric, Fiona}\}$

$F(A) =$ Expected number of people influenced when targeting A

Influence in social networks is submodular

[Kempe, Kleinberg, Tardos KDD '03]



Key idea: Flip coins \mathbf{c} in advance \rightarrow “live” edges

$F_{\mathbf{c}}(A)$ = People influenced under outcome \mathbf{c} (set cover!)

$F(A) = \sum_{\mathbf{c}} P(\mathbf{c}) F_{\mathbf{c}}(A)$ is submodular as well!

Closedness properties

- F_1, \dots, F_m submodular functions on V
and $\lambda_1, \dots, \lambda_m \geq 0$
- Then: $F(A) = \sum_i \lambda_i F_i(A)$ is submodular!
- Submodularity closed under **nonnegative linear combinations**
- Extremely useful fact:
 - $F_\theta(A)$ submodular $\Rightarrow \sum_\theta P(\theta) F_\theta(A)$ submodular!
 - Multicriterion optimization:
 F_1, \dots, F_m submodular, $\lambda_i > 0 \Rightarrow \sum_i \lambda_i F_i(A)$ submodular

Question

I have 10 minutes. Which blogs should I read to be most up to date?

[Leskovec-Krause et al. '07]



Thursday, Nov. 20, 2008

How Many Blogs Does the World Need?

By Michael Kinsley

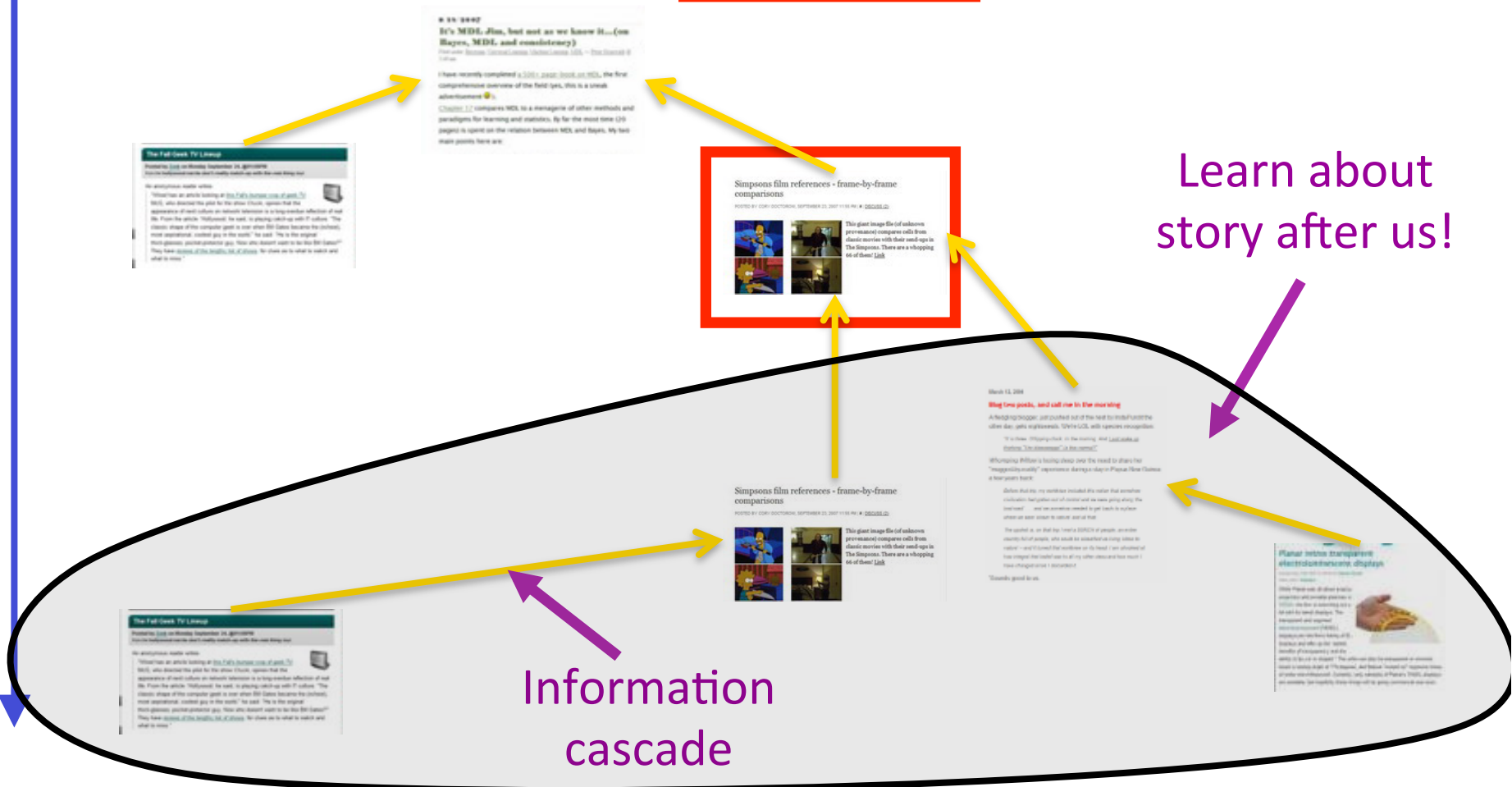
Detecting Cascades in the Blogosphere



Machine Learning
(Theory)



Time



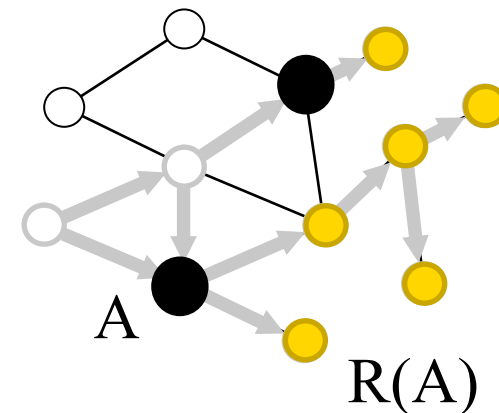
Learn about story after us!

Information cascade

Which blogs should we read to learn about big cascades early?

Modeling reward in information cascades

- Maximize the number of nodes that read the story after us:
 - If A are the monitored nodes, let $R(A)$ denote the number of nodes we “beat”



Optimization problem

- Given:
 - Graph $G(V,E)$, budget k
 - Data on how cascades $1,\dots,N$ spread over time
- Select a set of nodes A maximizing the reward

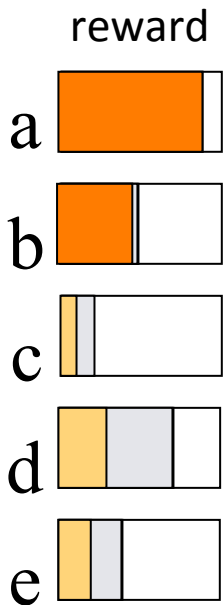
$$\max_{A \subseteq V} \sum_{i \in A} \text{Prob}(i) \underbrace{R_i(A)}_{\text{Reward for detecting cascade } i}$$

subject to $|A| \leq k$

F(A)

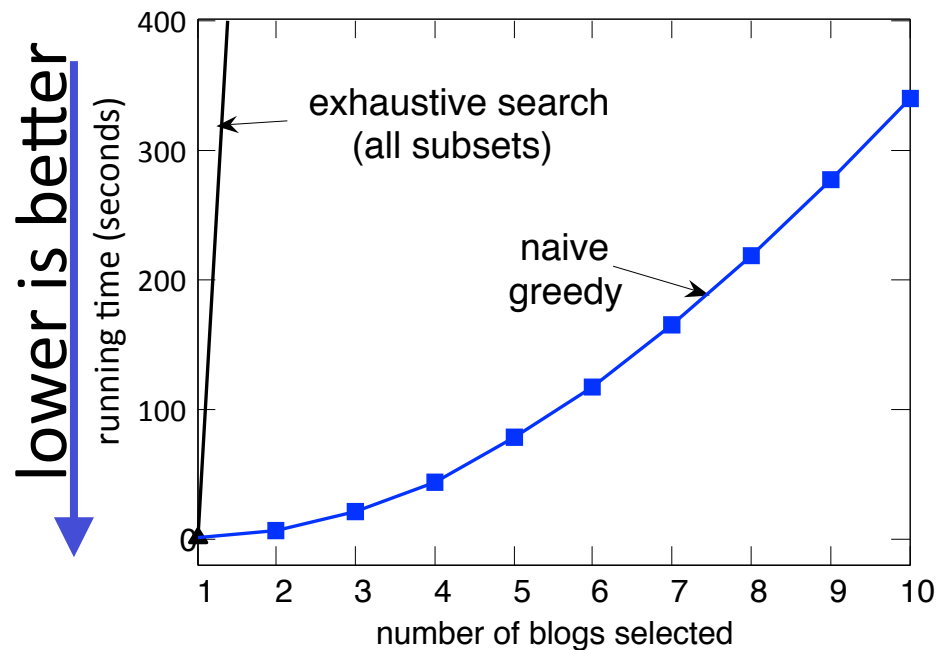
Greedy algorithm

Greedy algo



Add element with highest marginal gain

- Greedy algorithm is **slow**:
 - At each iteration we need to re-evaluate gains of all blogs
 - It scales as $O(n \cdot k)$



Scaling up greedy algorithm

- In round $i+1$:

- have so far picked $A_i = \{s_1, \dots, s_i\}$
- pick $s_{i+1} = \operatorname{argmax}_s F(A_i \cup \{s\}) - F(A_i)$
i.e., maximize “marginal benefit” $\delta_s(A_i)$
 $\delta_s(A_i) = F(A_i \cup \{s\}) - F(A_i)$

- Observation:** Submodularity implies
 $i \leq j \Rightarrow \delta_s(A_i) \geq \delta_s(A_j)$

$$A_0 = \{\}$$

$$A_0 \subseteq A_1 \subseteq A_2 \subseteq \dots \subseteq A_i$$

$$\delta_s(A_i) = F(A_i \cup \{s\}) - F(A_i)$$

$$\geq$$

$$\delta_s(A_j) = F(A_j \cup \{s\}) - F(A_j)$$

Marginal benefits δ_s never increase!

$$\delta_s(A_i) \geq \delta_s(A_{i+1})$$

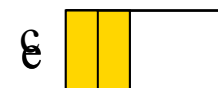


“Lazy” greedy algorithm

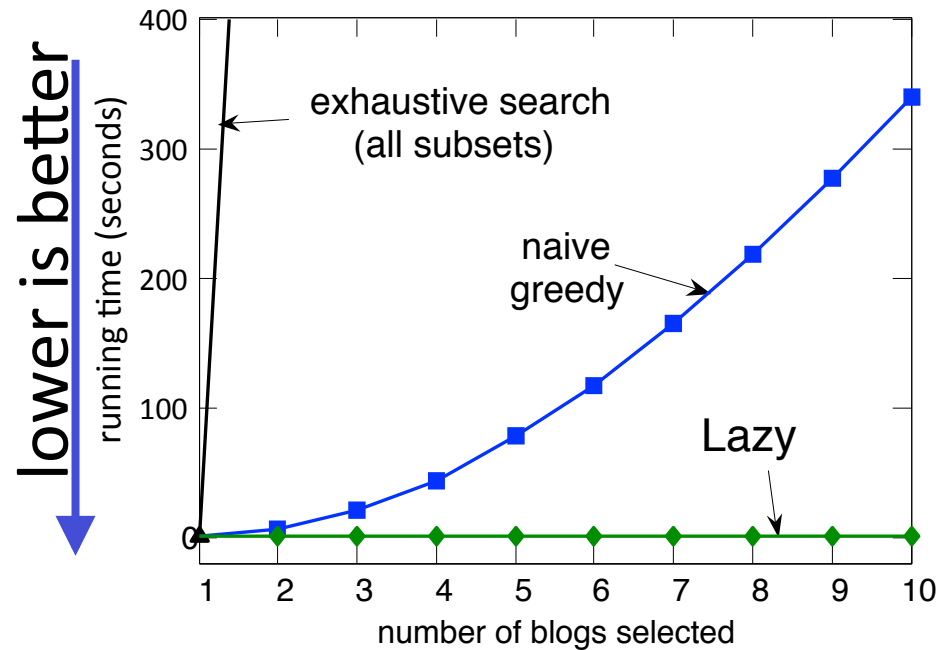
Lazy greedy algorithm:

- First iteration as usual
- Keep an ordered list of marginal benefits δ_i from previous iteration
- Re-evaluate δ_i only for top element
- If δ_i stays on top, use it, otherwise re-sort

Benefit $\delta_s(A)$

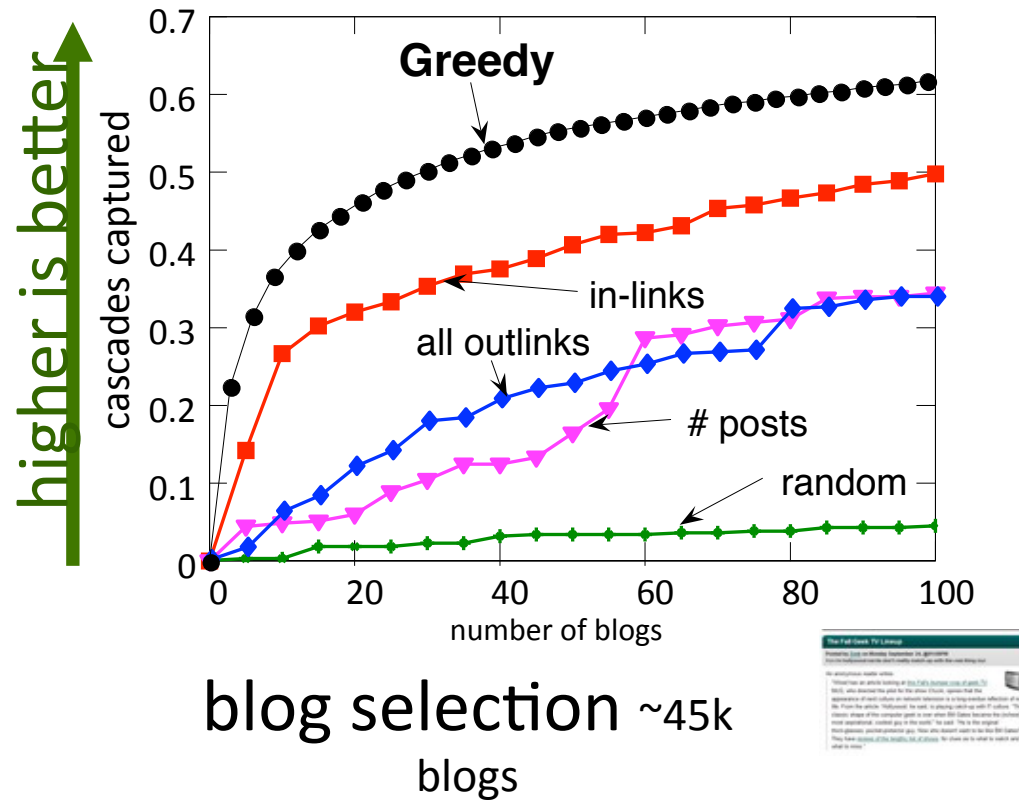


Result of lazy evaluation



- Using “lazy evaluations”
 - 700 times performance improvement!

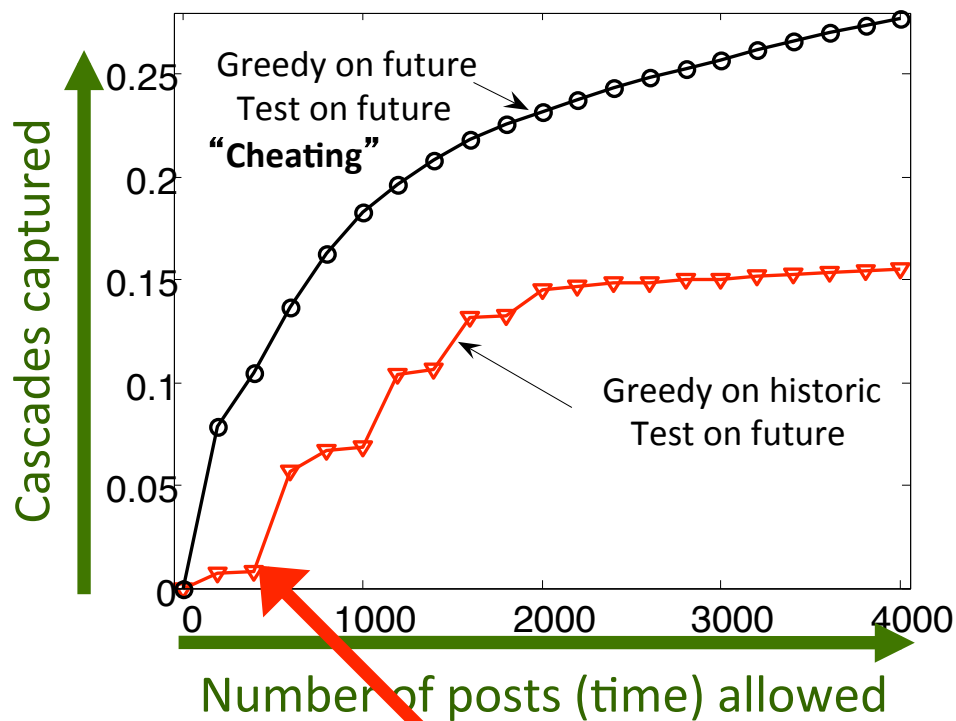
Performance on Blog selection



- Submodular formulation outperforms heuristics

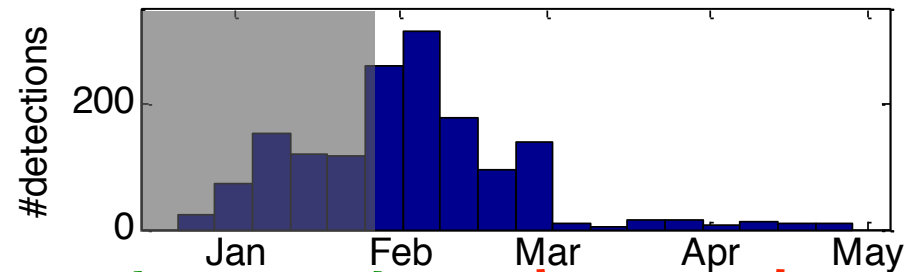
Predicting informativeness

- Want blogs that *will be* informative in the future
- Split data set; train on historic, test on future



Let's see what goes wrong here.

Detects on training set



Blog selection "overfits" to training data!

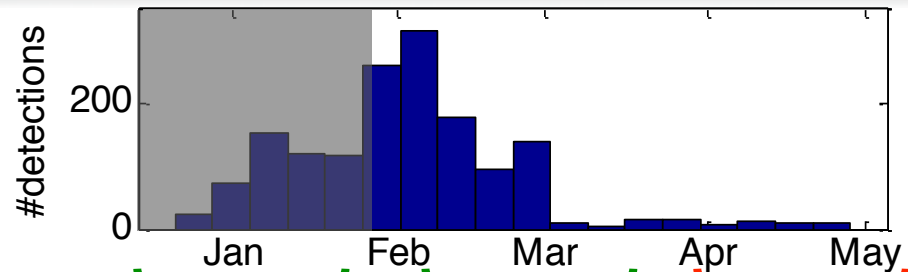
Poor generalization!

Want blogs that continue to do well!

Online optimization

“Overfit” blog
selection **A**

$F_i(A)$ = detections
in interval i



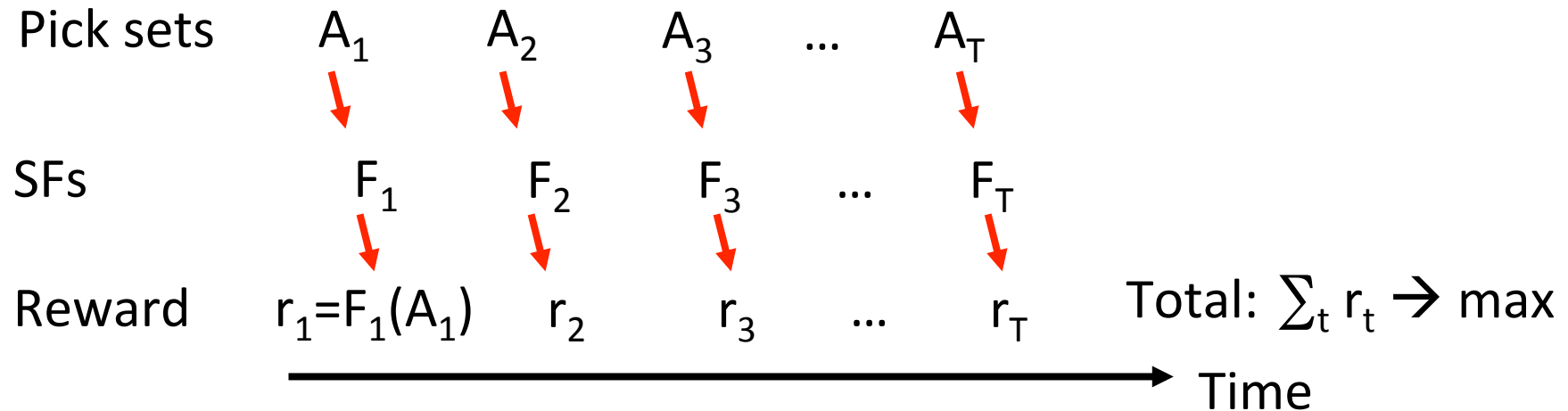
→ $F_1(A) = .5$ $F_3(A) = .6$ $F_5(A) = .02$
 $F_2(A) = .8$ $F_4(A) = .01$

Online optimization:

$$\max \sum_{t=1}^T F_t(A_t)$$

Online submodular maximization

[Streeter, Golovin NIPS '08]



This is an online learning / multi-armed bandit problem, with one arm for each set of blogs to recommend!

Number of arms is exponential!!

Online maximization of submodular functions

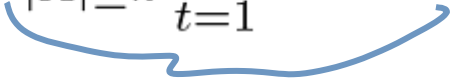
- Suppose we wish to pick k out of n documents
- Initialize k multi-armed bandit (MAB) algorithms
- OnlineGreedy algorithm: In each round do
 - For $i = 1:k$
 - MAB i picks item x_i
 - Feed back reward $F_t(\{x_1, \dots, x_i\}) - F_t(\{x_1, \dots, x_{i-1}\})$
- Example: $F_t(A)$
 - Counts the number of clicks (0 or 1) on ads in A
 - Counts the cascades detected by blogs A
 - ...

Online submodular maximization

[Streeter, Golovin NIPS '08]

Theorem

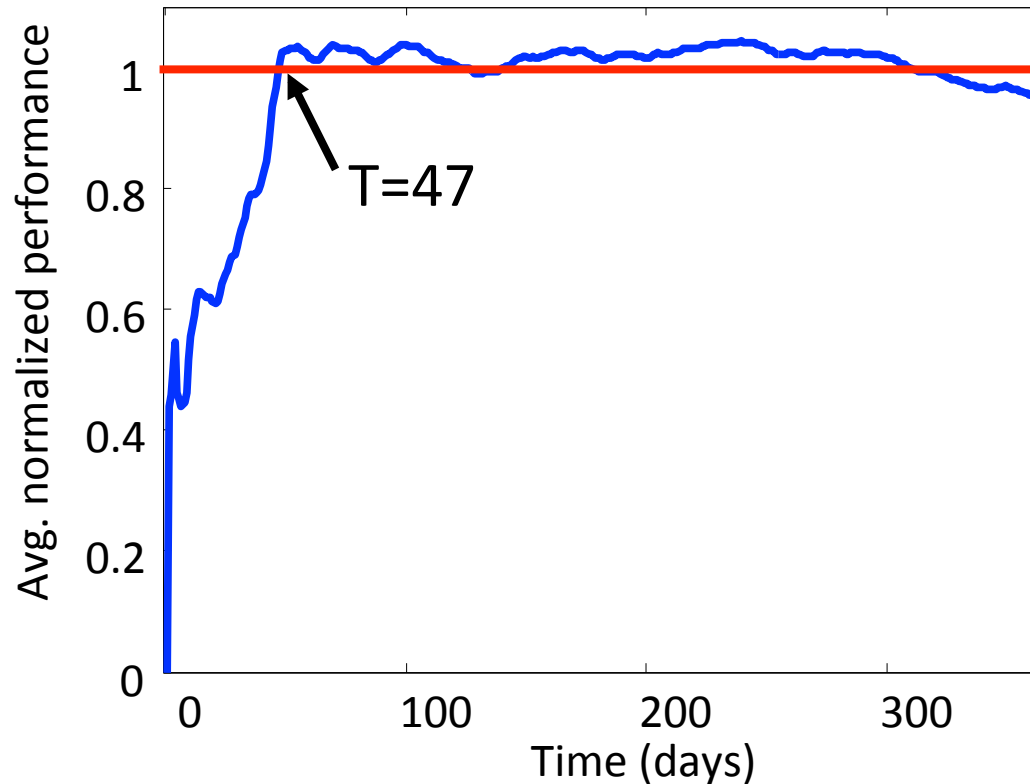
OnlineGreedy chooses A_1, \dots, A_t s.t. in expectation

$$\frac{1}{T} \sum_{t=1}^T F_t(A_t) \geq \frac{1 - 1/e}{T} \max_{|A| \leq k} \sum_{t=1}^T F_t(A)$$


for any sequence F_i , as $T \rightarrow \infty$

Can get 'no-regret' over "cheating" greedy algorithm

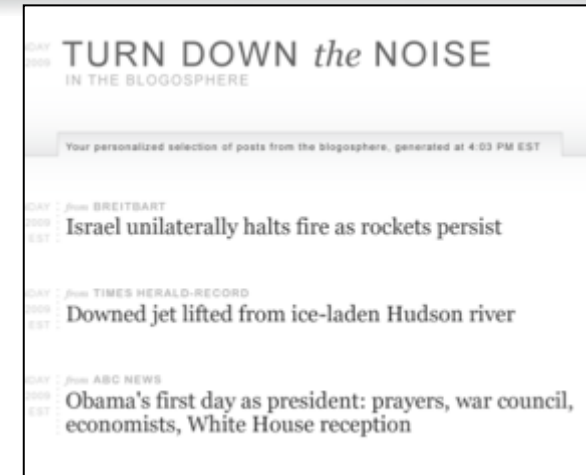
Results on blogs



Performance of online algorithm converges quickly to “cheating” offline greedy algorithm!

Turning down the noise in the blogosphere [El-Arini et al '09]

Blogosphere



Afghanistan
airport
terror
Iraq
9/11
security
Globe



$$F(\mathcal{A}) = \sum_{f \in U} w_f \text{cover}_{\mathcal{A}}(f)$$

Submodular function optimization

Feature
Extraction

Coverage
Function

Post
Selection



Evaluating Coverage [El-Arini et al '09]

- Evaluate on **real blog data** from Spinn3r
 - 2 week period in January 2009
 - ~200K posts per day (after pre-processing)
- Two variants of the algorithm



TDN+LDA: High level features
Latent Dirichlet Allocation
topics



TDN+NE: Low level
features

- User study involving 27 subjects to evaluate:

Topicality &
Redundancy

Topicality = Relevance to current events



The Washington Post



REUTERS

ESPN People

...

Reference Stories
(ground truth)

Downed jet lifted from
ice-laden Hudson
River

NEW YORK (AP)—

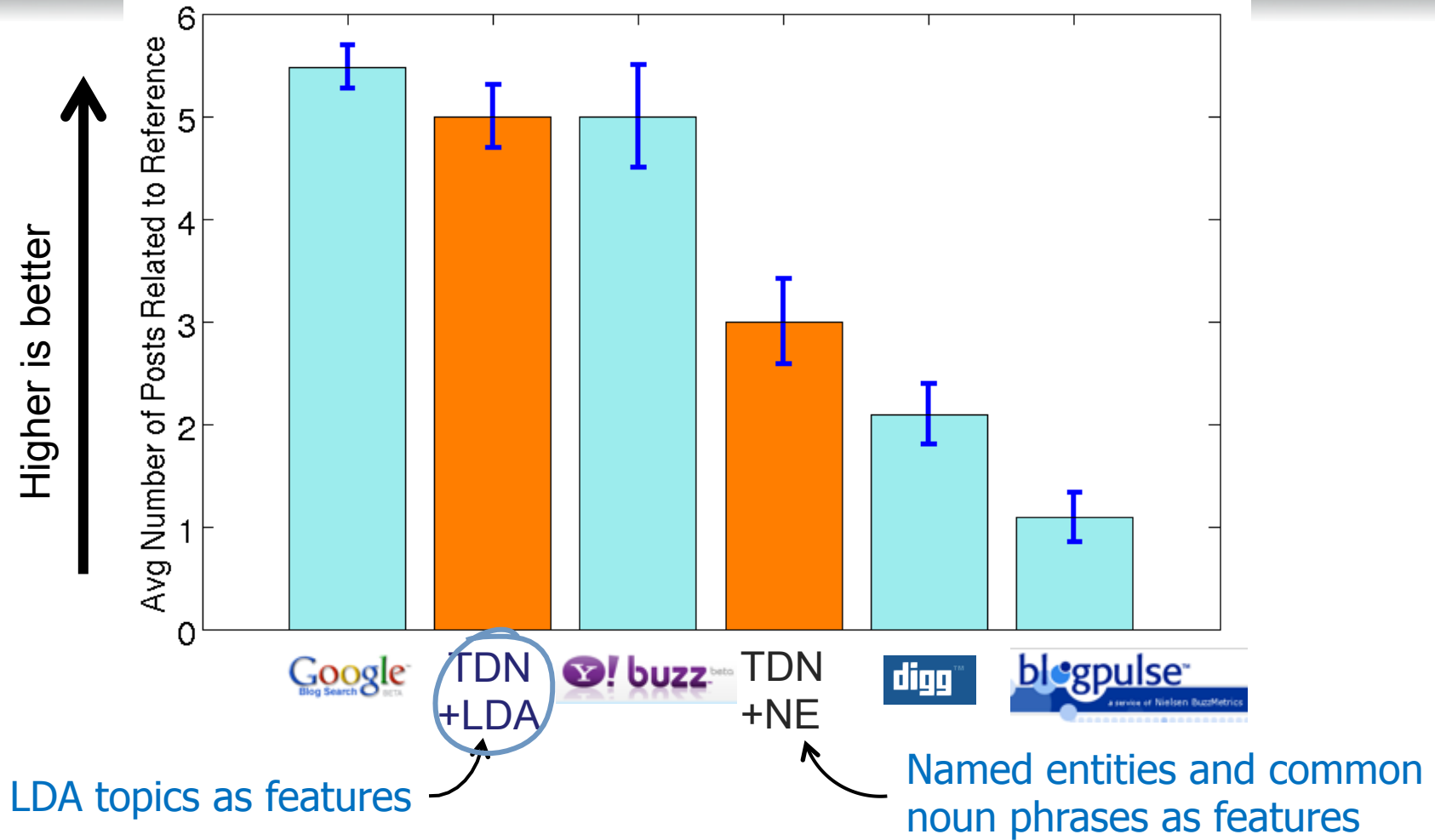
The airline jet was

Is this post topical? i.e.,
is it related to any of the
major stories of the day?

Post for evaluation

[El-Arini et al '09]

Results: Topicality



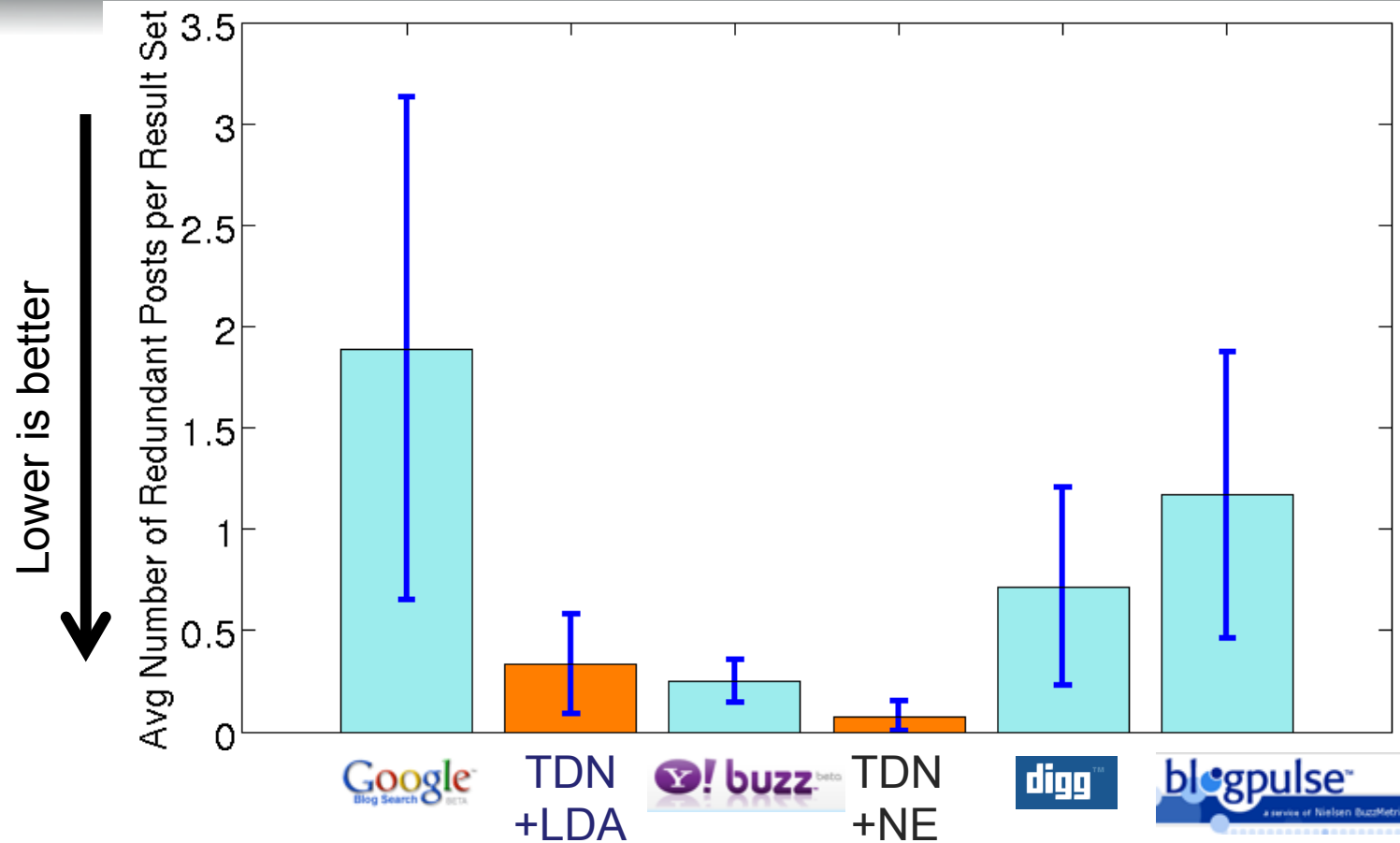
Evaluation: Redundancy

1. Israel unilaterally halts fire as rockets persist
2. Downed jet lifted from ice-laden Hudson River
3. Israeli-trained Gaza doctor loses three daughters and niece to IDF tank shell
4. ...

Is this post redundant with respect to any of the previous posts?

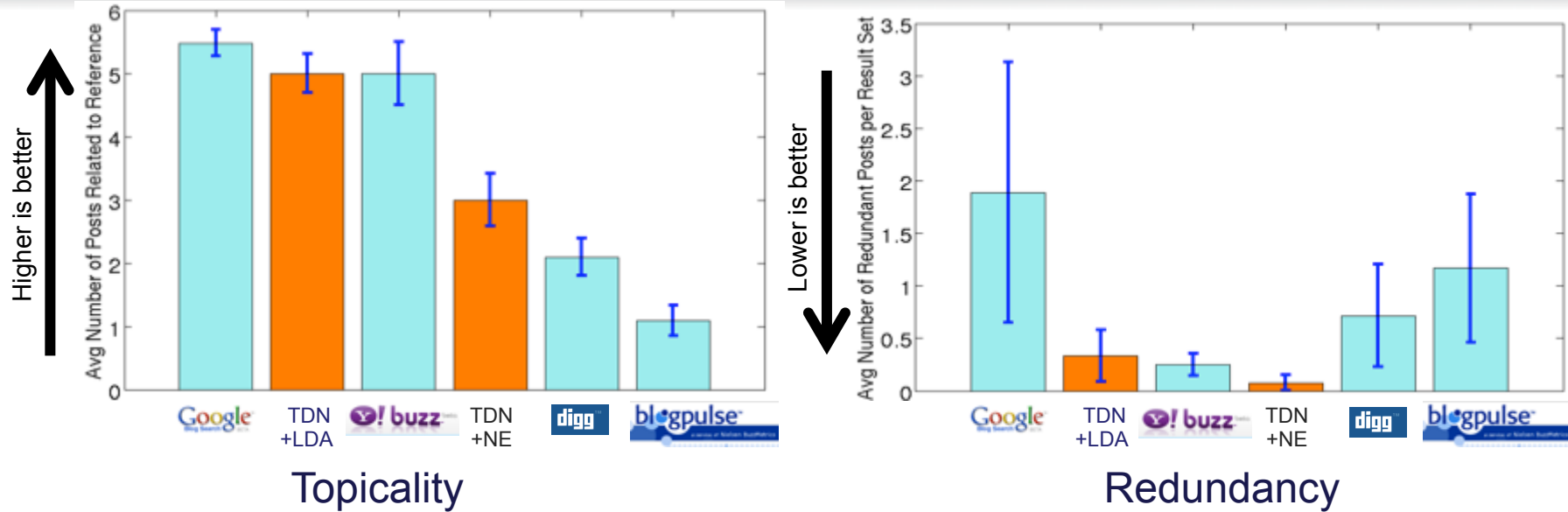
[El-Arini et al '09]

Results: Redundancy



[El-Arini et al '09]

Results: Coverage



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

Summary

- Submodular functions arise in many problems
- Simple greedy algorithm performs well
- Can combine with multi-armed bandit algorithms

You learned a lot

- **Retrieval**

- Nearest-neighbor, min-hashing, locality sensitive hashing, ...

- **Supervised learning**

- Online SVM, online logistic regression, online convex programming, parallel online learning, ...

- **Unsupervised learning**

- Online k-means and EM, coresets, active learning, uncertainty sampling, informative sampling, ...

- **Interactive data mining**

- Multi-armed bandits, epsilon-greedy, UCB, GP optimization, submodular functions, ...

Common insights

- Large data → New computing paradigm
 - Distributed / map-reduce style computations
 - Dealing with data streams
- Large data → Simple algorithms
 - Random sampling (hashing, dim. reduction, coresets, ...)
 - Stochastic gradient descent (online SVM, online EM, ...)
 - Greedy algorithms (UCB, submodularity)
 - In many cases, can prove that these simple algorithms work well (often the better the larger the data)
- Large data → Attention as a scarce resource
 - Need methods to cope with the information overload (active learning, exploration-exploitation tradeoffs, ...)

Where to learn more

- Probabilistic Artificial Intelligence (Fall term)
How can we build systems that
 - do well in unknown environments and unforeseen situations?
 - exhibit “intelligent” behavior, without explicit rules?
 - learn from experience in order to improve their performance?
 - Will cover modeling techniques and algorithms from statistics, optimization, planning, and control and study applications in areas such as sensor networks, robotics, and the Internet.
- Conference proceedings
 - Data Mining: KDD, SDM, ICDM, ...
 - Machine learning: ICML, NIPS, AISTATS, ...
 - Web mining: WWW, WSDM, ...
- Master’s thesis 😊

Acknowledgments

- These slides are partially based on material from Carlos Guestrin and Jure Leskovec