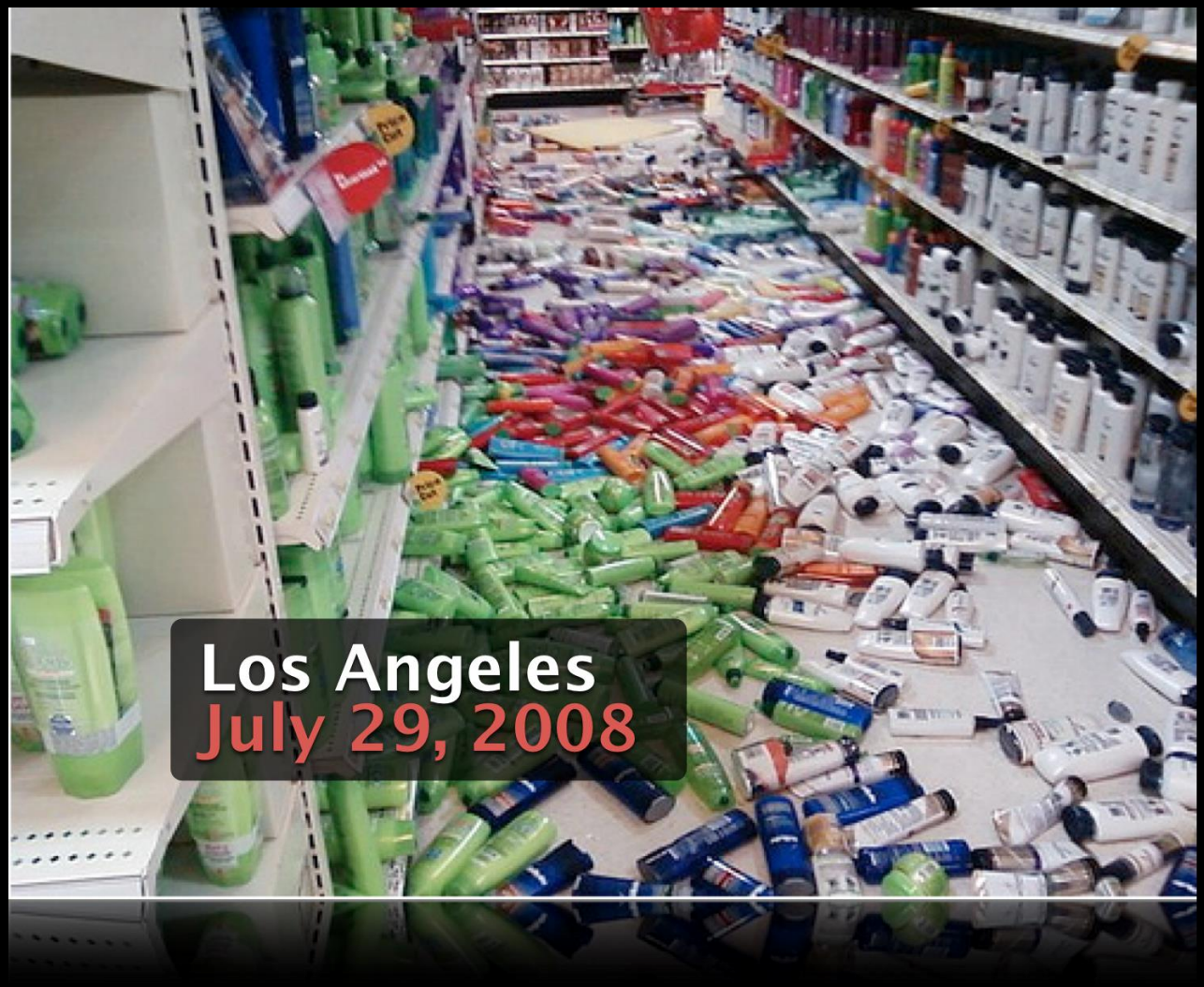
Research at Twitter

Aleksander Kołcz

Twitter, Inc.



Earthquake: 0 sec

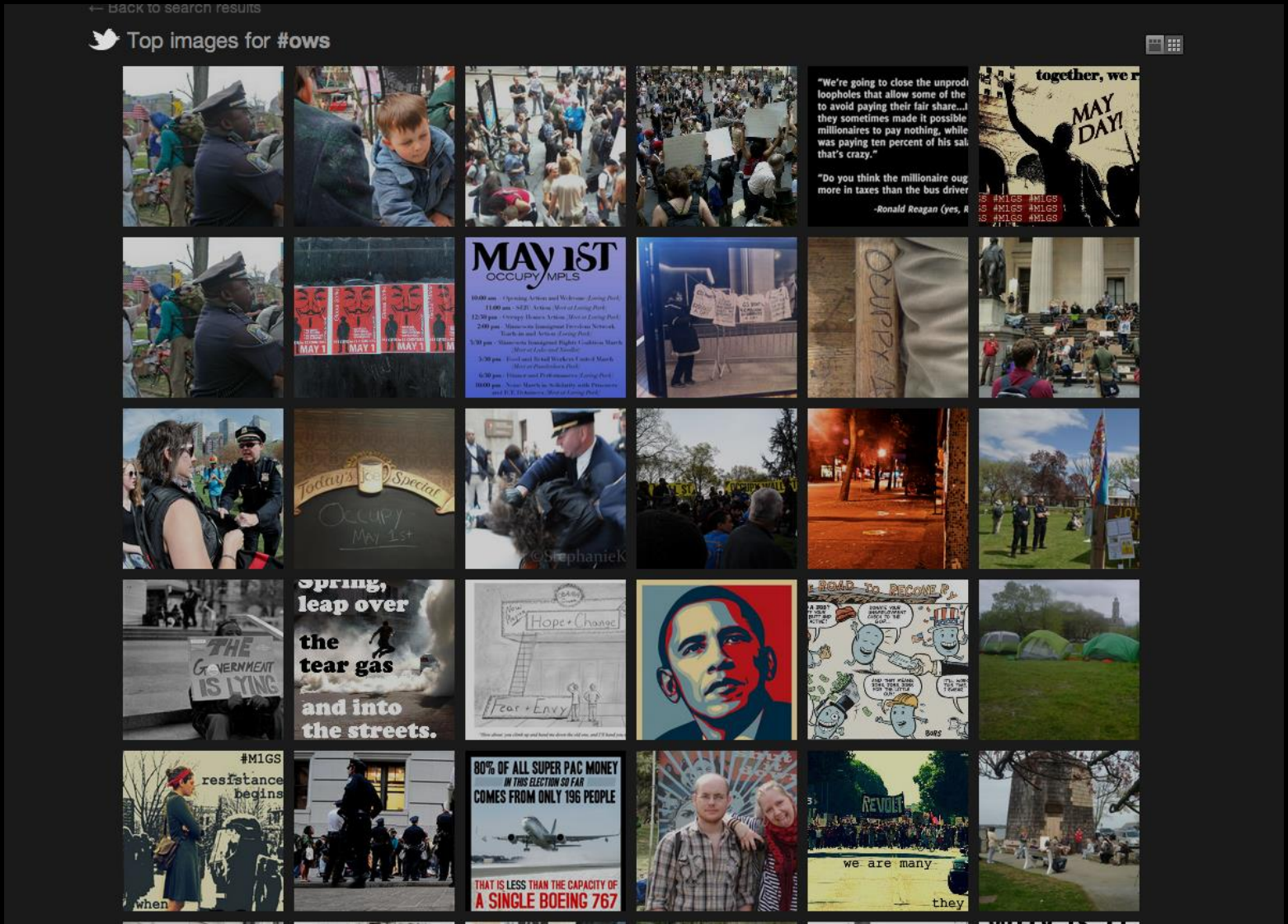
First Tweet: 5 sec

Local News: 4 min

AP Wire: 9 min

Book of Tweets

#ocws, #occupywallstreet, …



The Scale of Twitter

- > 200M active users
- Approx. 400M Tweets sent/day
- > 400M/month unique visitors to twitter.com
- Support for > 35 languages
- 70% of users outside US
- 33,388 record TPS

Internal vs External R&D

- A lot of Twitter data is open (in small quantities)
- It is easy to create small-scale specialized collections and institutions can acquire larger collections via Gnip
- There are lots of academic publications analyzing interesting properties of Twitter data
- Internal R&D is focused on improving the user experience

Overview

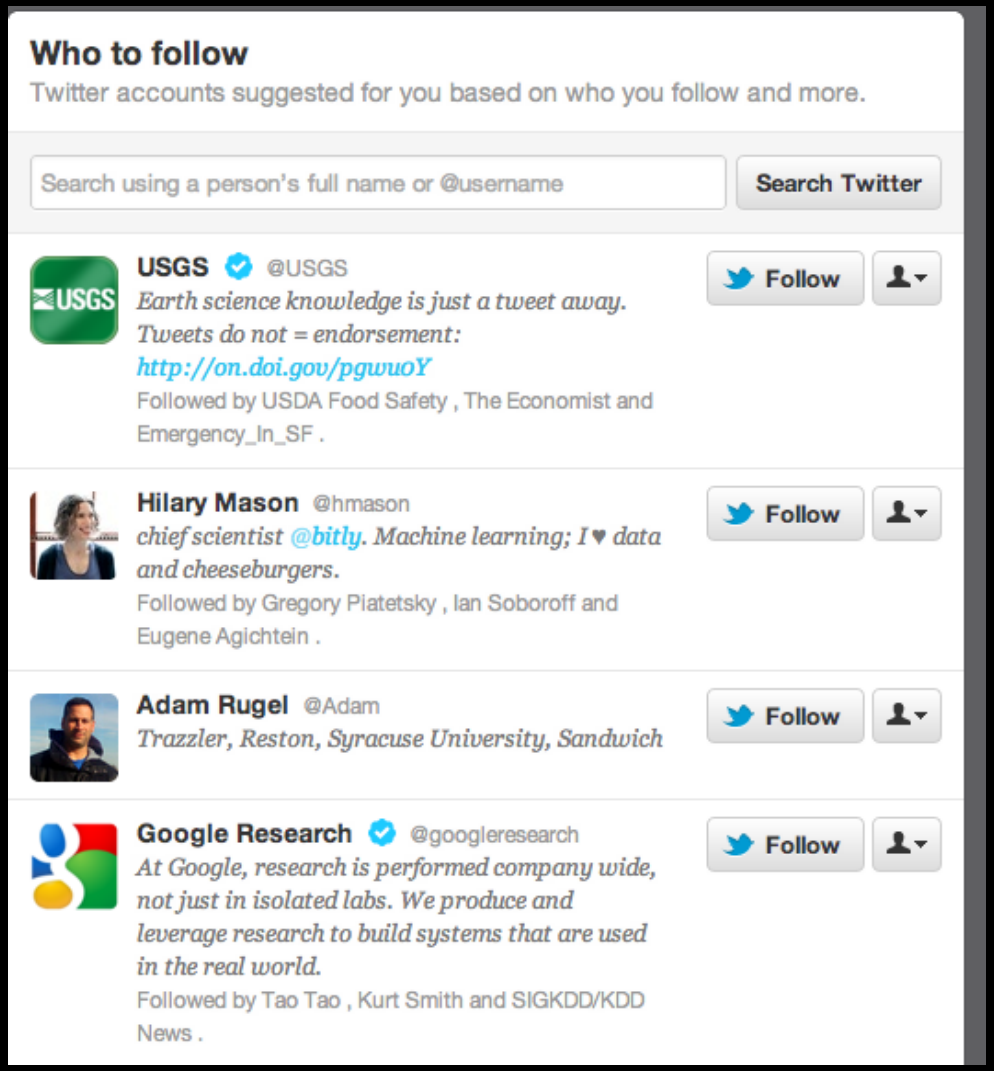
- Key problem areas
  - Product features
  - Infrastructure
- Key research problems/subproblems
  - Evolving landscape
- Collective experiences
  - A few lessons learned

Problems we are trying to solve

- Relevance
  - ranking in search

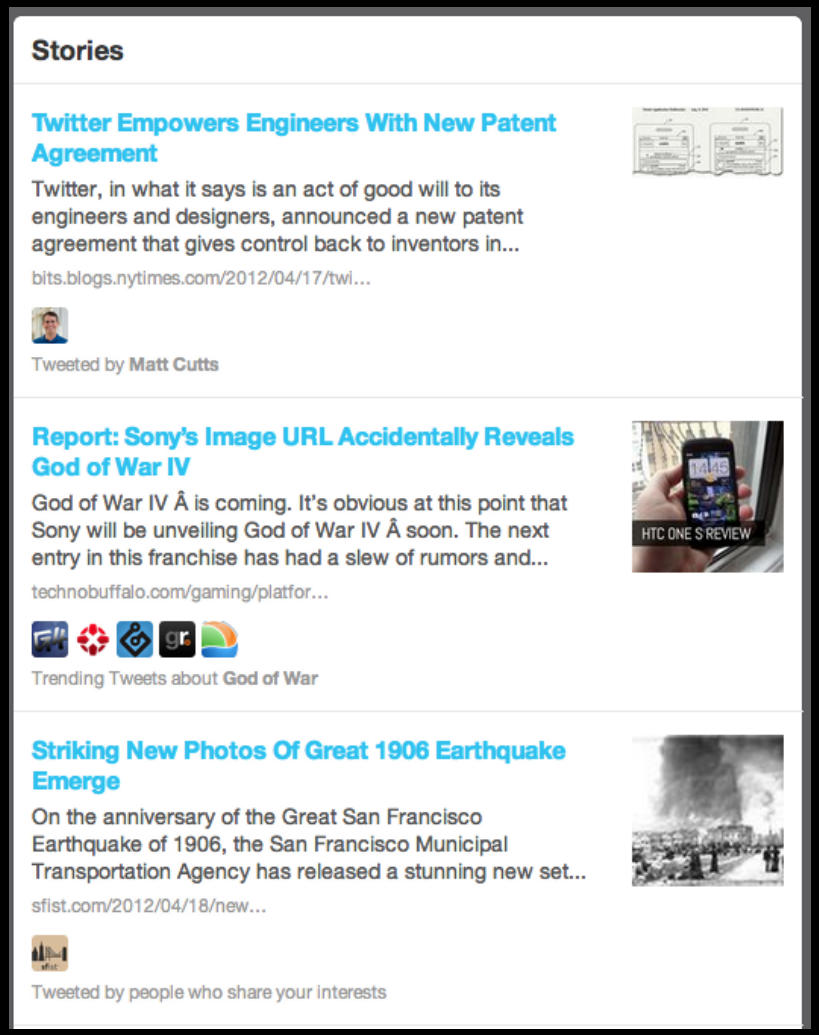
Problems we are trying to solve

- Who to follow



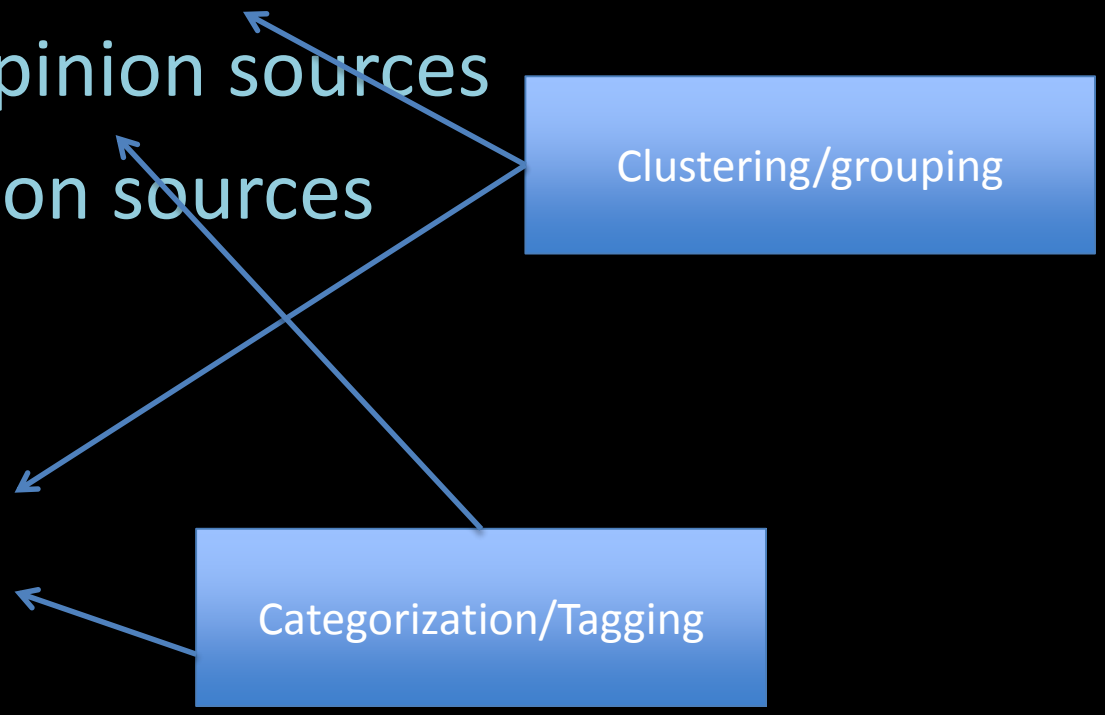
Problems we are trying to solve

Content recommendation (stories/media)



Recommendation/Personalization

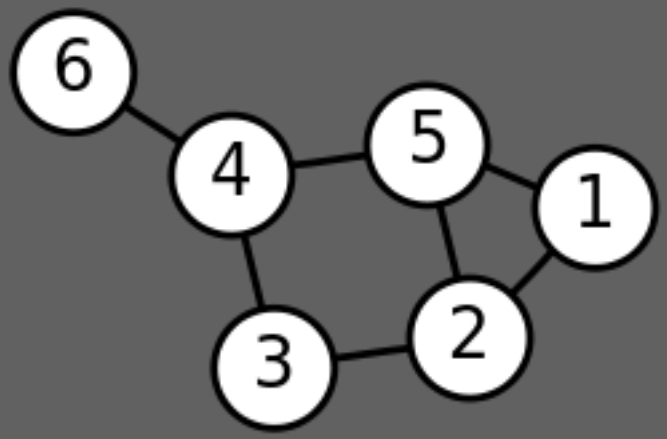
- Other users – friends
- Other users – opinion sources
- News/information sources
- Specific tweets
- Specific urls
- Lists/hashtags



Key Research Areas

- Graph mining
- Content/user tagging/classification
- Content/user recommendation
- Search/IR
- Integration/Infrastructure (best practices)

Graph mining



Graph mining (Follow graph)

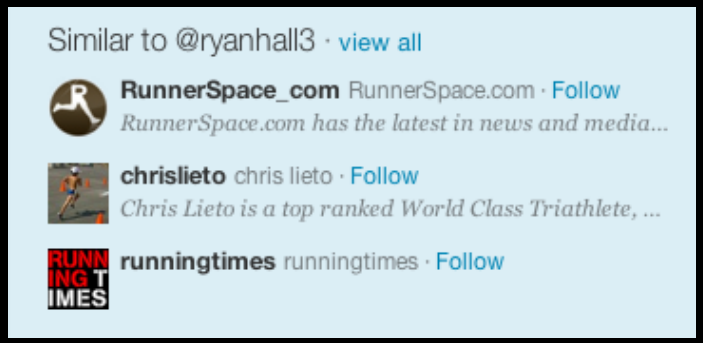
- How does it vary across locales
- How to use it to judge user importance
- How to use it to provide user recommendations
- The role of reciprocated edges (mutual follows)
- How to use it to analyze information spread (diffusion)

Follow graph

- 20B edges connecting active users
- Over 1K users having 1M+ followers
- > 25 users having 10M+ followers

wtf

(who to follow)

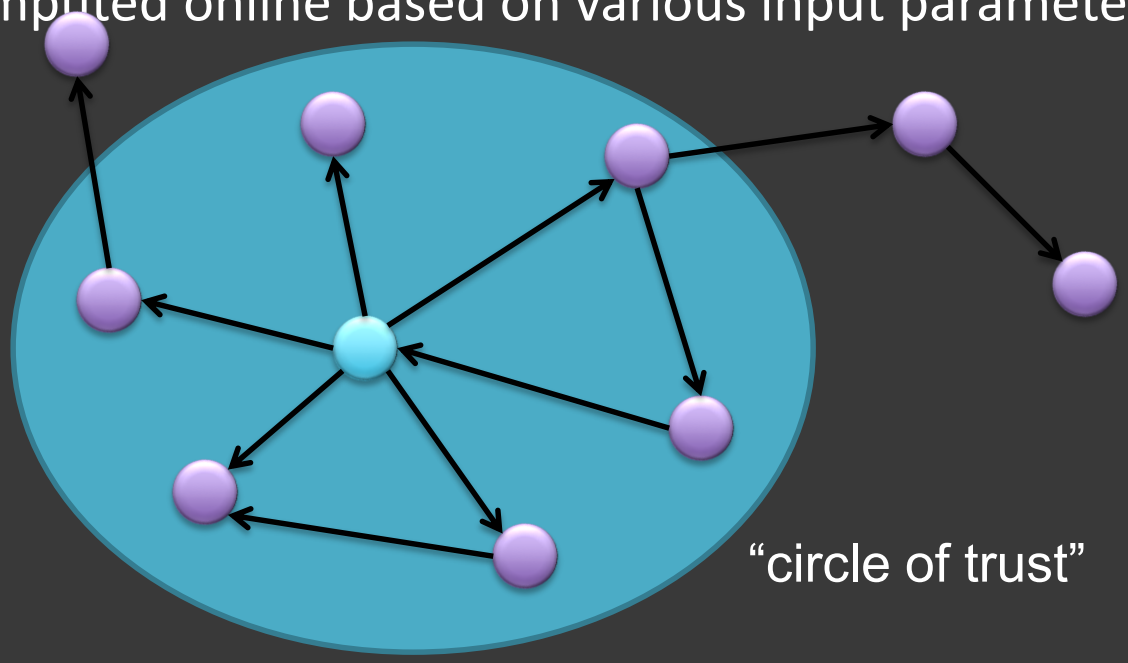


Source: Gupta et al. "WTF: The Who to Follow Service at Twitter", WWW'13

“Circle of Trust”

Ordered set of important neighbors for a user

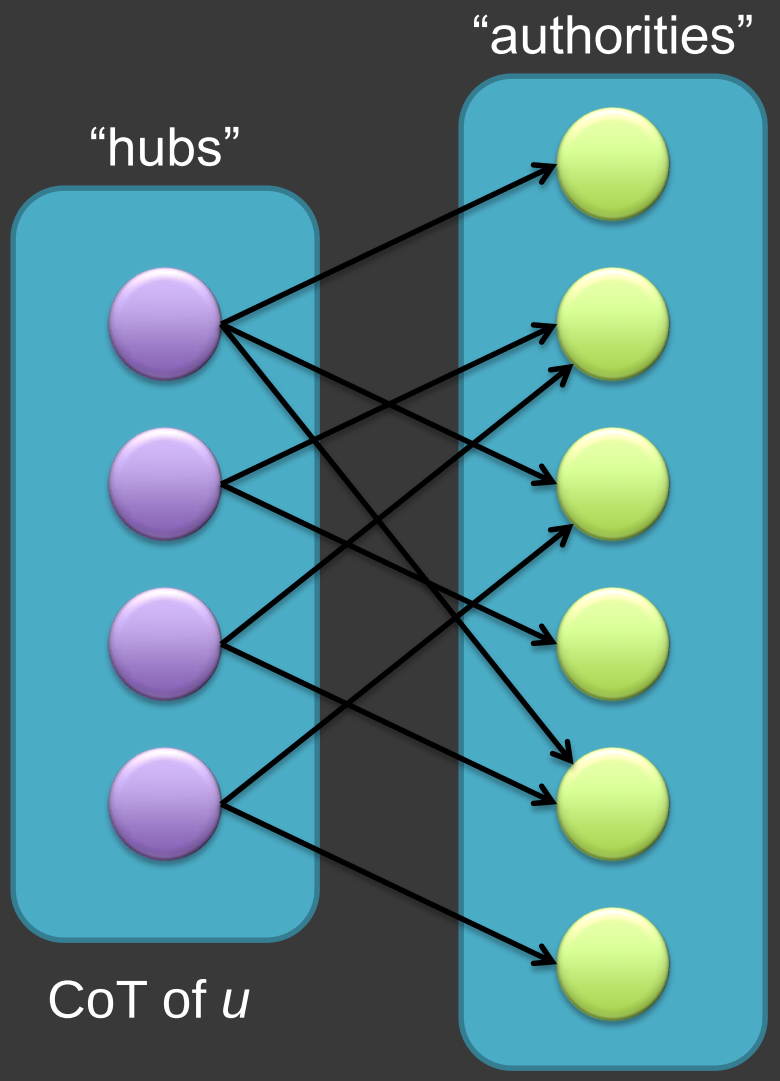
- Result of egocentric random walk
- Computed online based on various input parameters



One of the features used in search

Source: Gupta et al. "WTF: The Who to Follow Service at Twitter", WWW'13

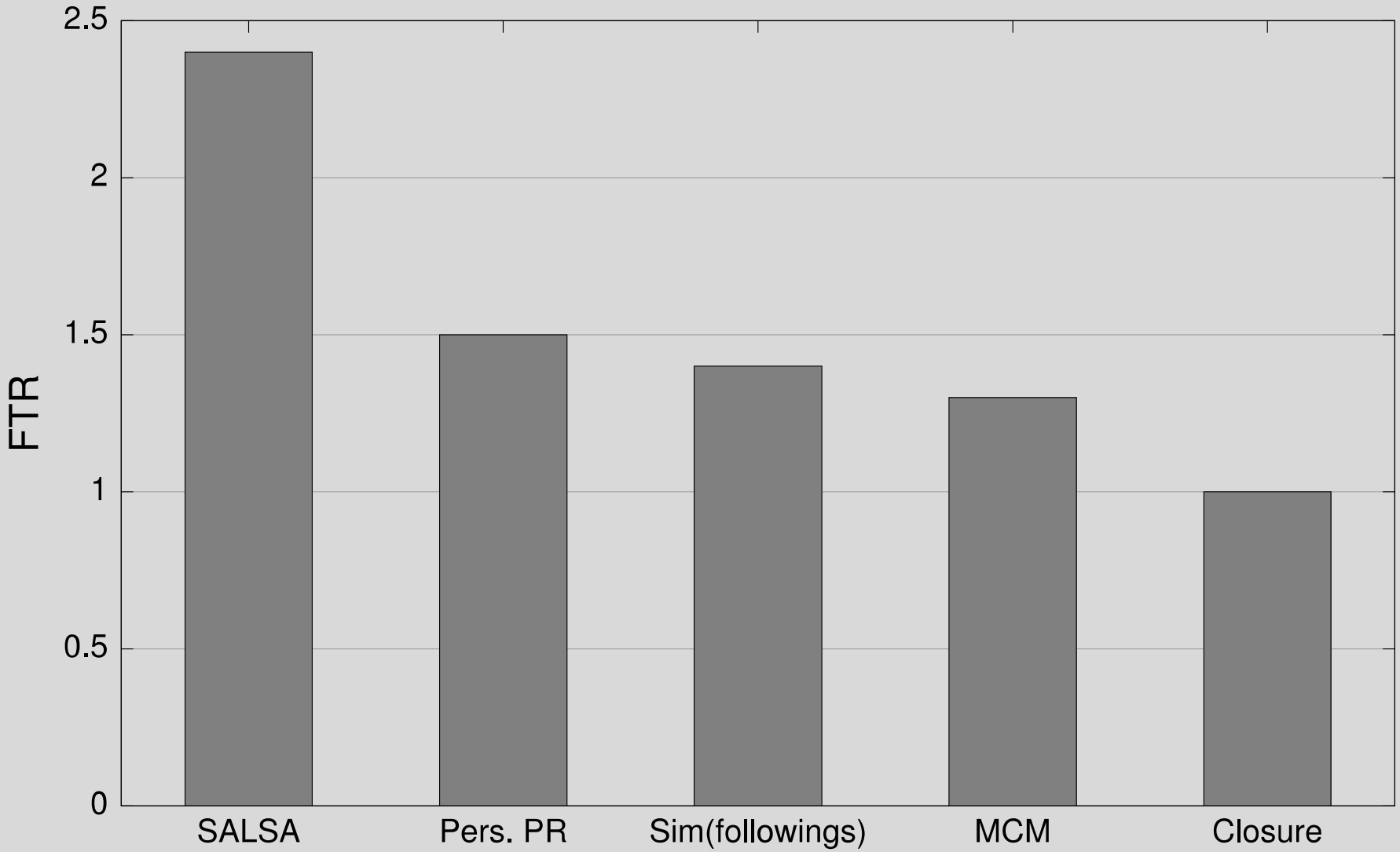
SALSA for Recommendations



hubs scores:
similarity scores to *u*

authority scores:
recommendation scores for *u*

Source: Gupta et al. "WTF: The Who to Follow Service at Twitter", WWW'13



Source: Gupta et al. "WTF: The Who to Follow Service at Twitter", WWW'13

Graph mining (Engagement)

- Connecting user actions to content (users, tweets, urls, images, video)
- Implicit relevance judgment
- Optimizing the performance of search and recommender systems

Graph mining (Heterogeneous)

- Recommendation of users/content may depend not just on user connectivity according to follow behavior but also on their actions and perceptions by others
- What sources if information to combine (clicks, tweets, favs, lists)
- How best to integrate this information (weighting factors, combination method)

User/content categorization

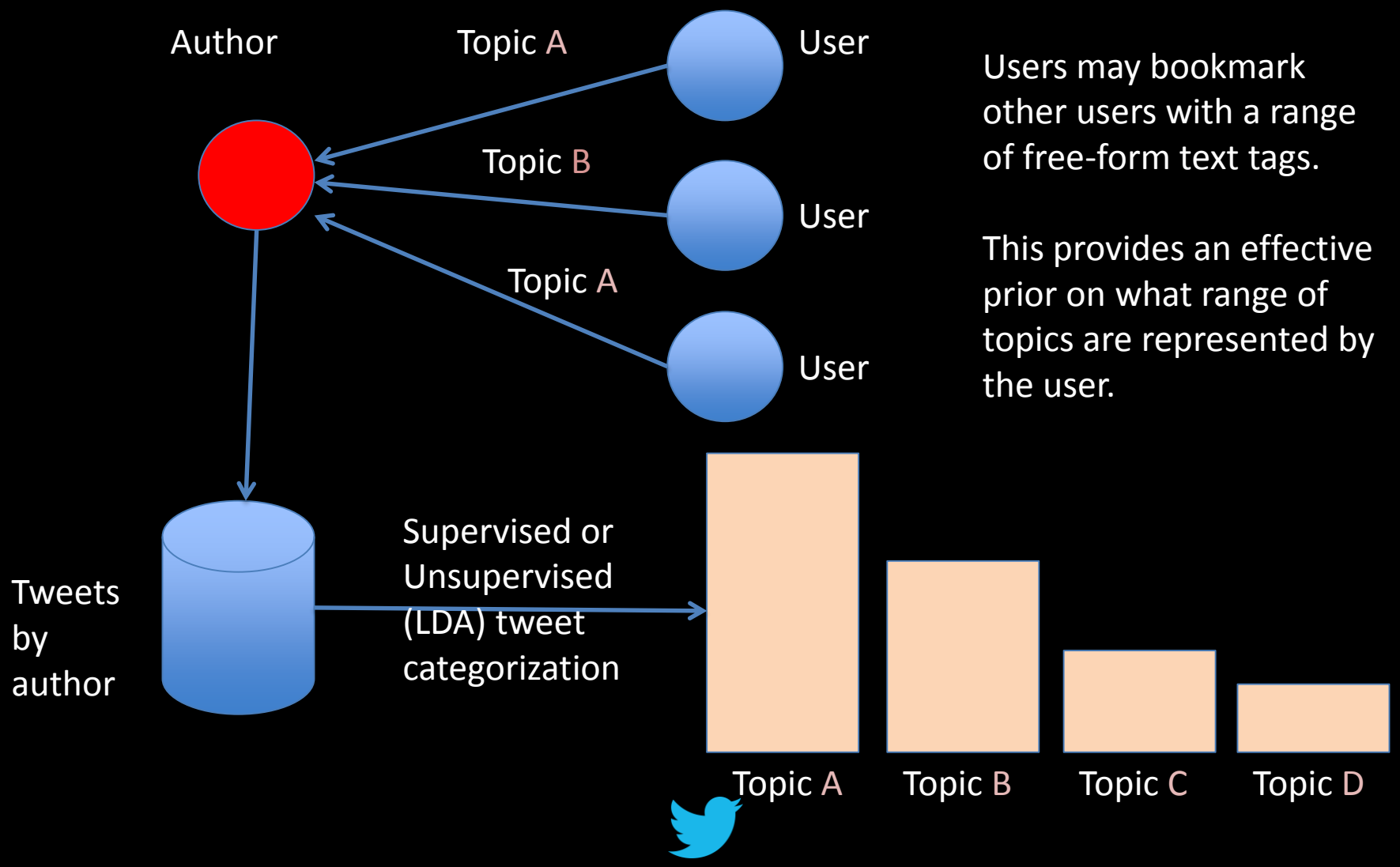
User Interest Modeling

- Users are interested in topics embodied by users they follow
- Users interact (read, retweet, fav) with tweets that seem interesting
- Users click/read urls corresponding to interesting content
- Users search for items that are of interest

Challenges

- A user may be followed for a number of reasons, e.g., family/friend relationships
  - Can we infer the reason?
- A user may tweet on a range of topics, and often "chatter" about their personal life
  - It is hard to assign labels to tweets by using the author
- Tweets are short and noisy
  - Can we infer the topic just looking at a tweet without context?
- User interests are time sensitive
  - Vary with trends, fads, current events, etc

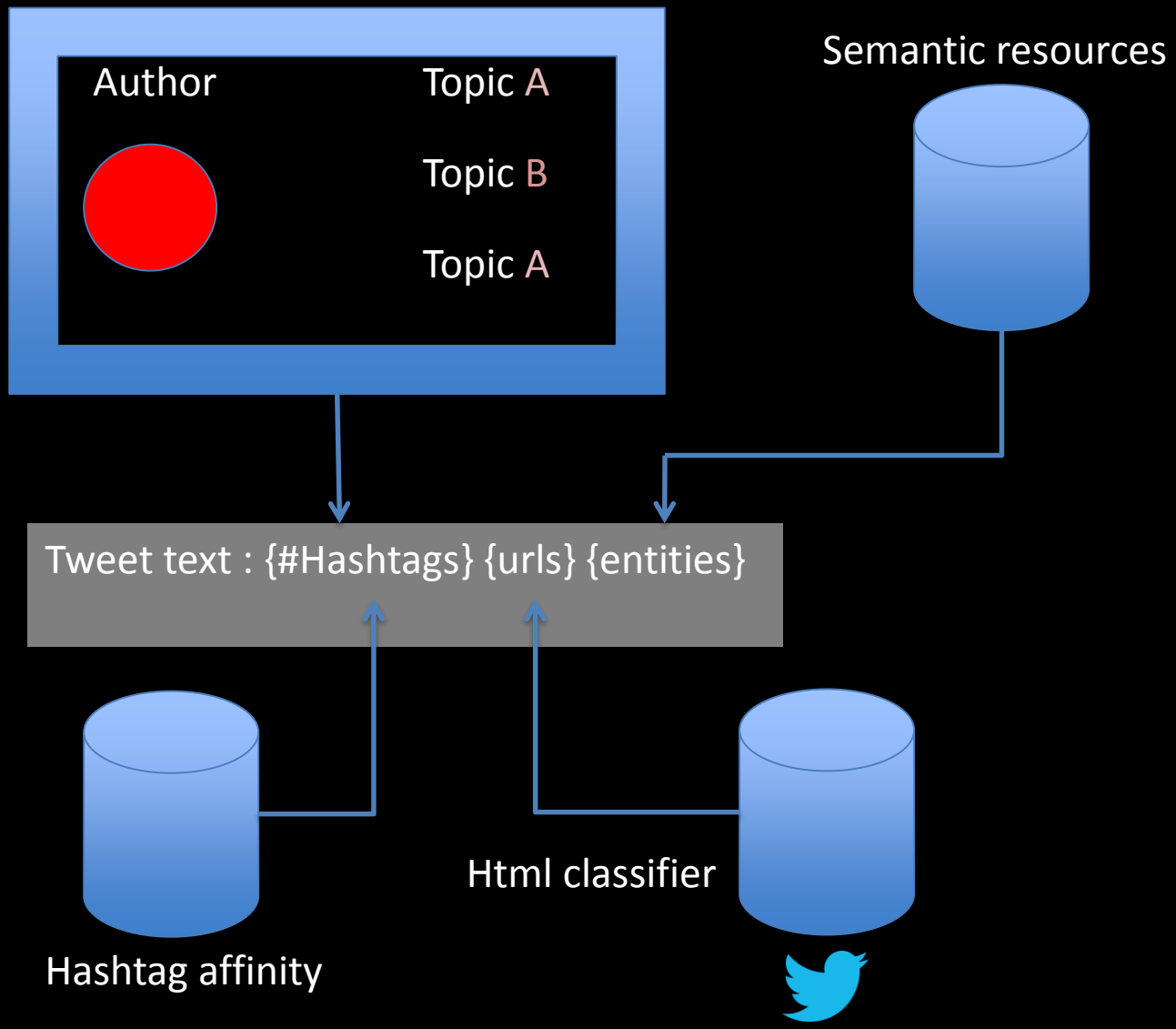
User topic attribution: producer side



Users may bookmark other users with a range of free-form text tags.

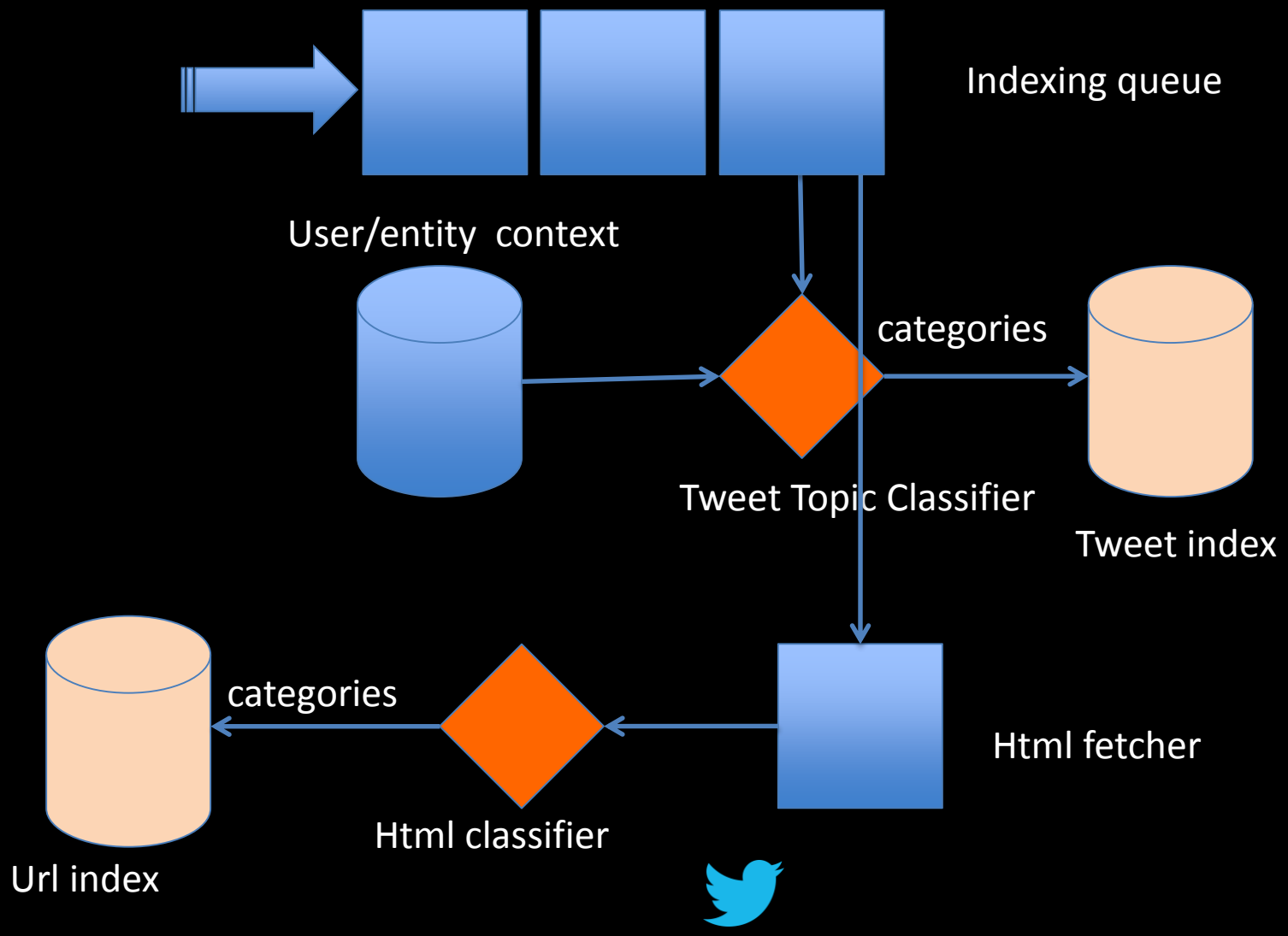
This provides an effective prior on what range of topics are represented by the user.

Tweet topic attribution



Tweets themselves are short and therefore in addition to raw tweet text it is important to consider their context as the source of potential features

Integration of tweet/url models



Multi-label classification at scale - 1

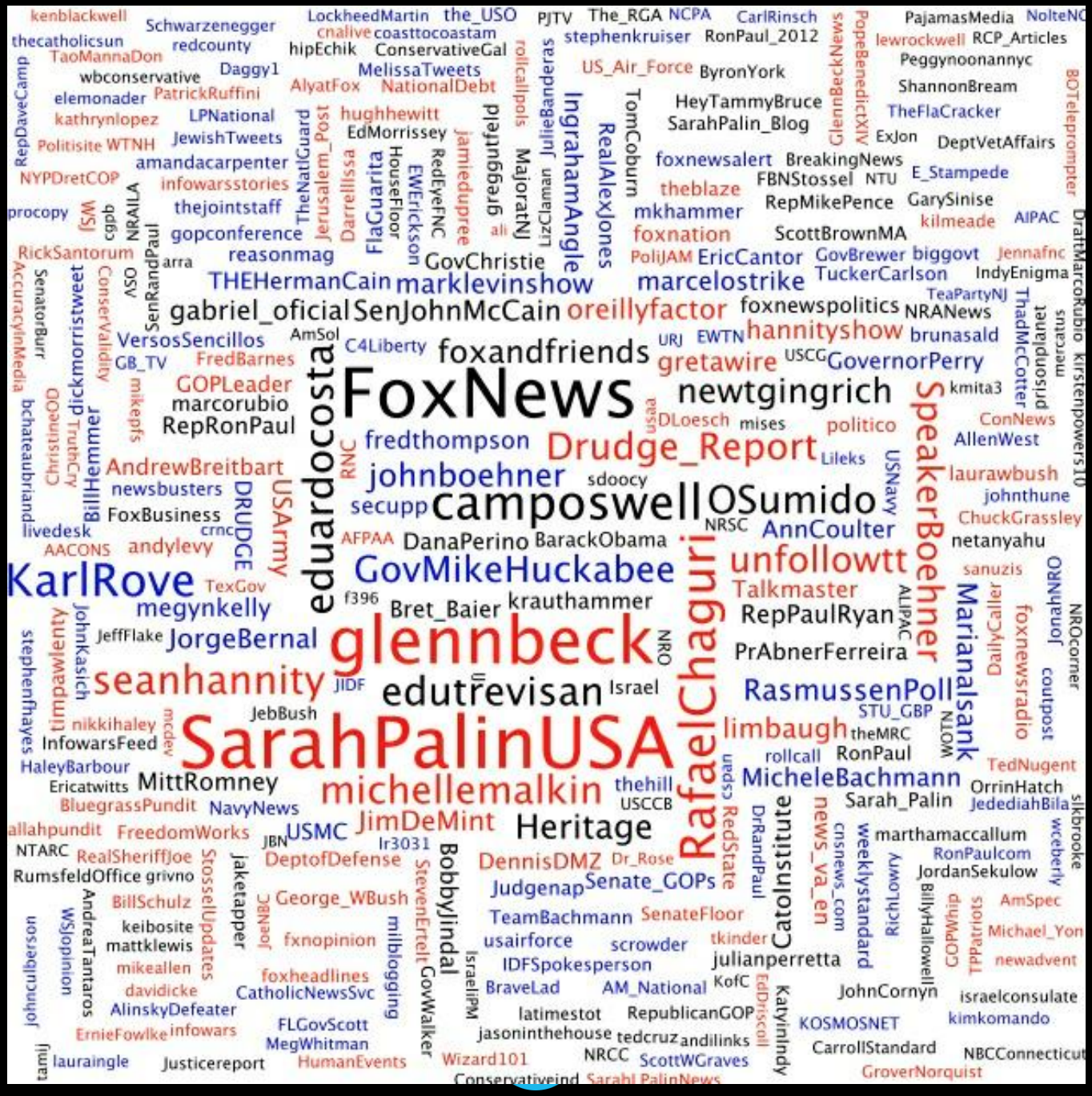
- Tweets are tagged with millions of rapidly evolving hashtags
- Users are tagged with millions of unique list names
- These are just two examples of folksonomy-type tag spaces that can be used to derive models to classify new content
- Each item can naturally belong to multiple categories, but it is unlikely that an item is labeled with all categories it can belong to
- How to curate labeled data to facilitate training and evaluation
- How to deal with label correlation

Multi-label classification at scale - 2

- Challenges of running tens of thousand of models (parallel evaluation)
- Automatic detection of synonymous topics (related hashtags, lists) even if surface similarity of tags is low
- Dealing with concept hierarchy (auto-generation)
- The role of external semantic resources (Wikipedia, Freebase, etc)

Chatter detection

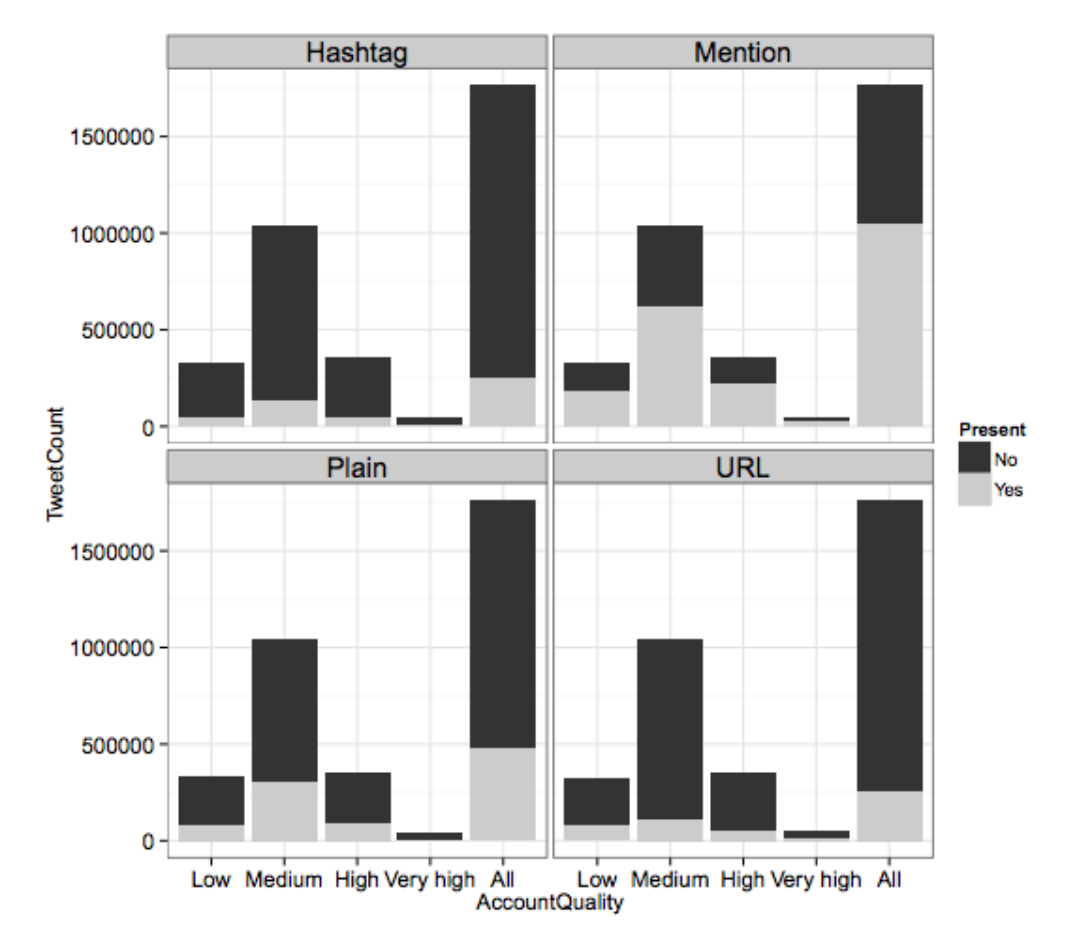
- A lot of tweets correspond to personal updates
- These are certainly important but perhaps to a rather small audience
- In recommendation it is important to identify items of general interests (e.g., oriented about well defined topics)


FoxNews
SarahPalinUSA
glennbeck
KarlRove
RafaelChaguri
SpeakerBoehner
Drudge_Report
eduardocosta
camposwell
OSumido
unfollowtt
GovMikeHuckabee
seanhannity
edutrevisan
michellemalkin
JimDeMint
Heritage
CatoInstitute
MicheleBachmann
RasmussenPoll
Marianalsank
limbaugh
newtgingrich
johnboehner
foxandfriends
oreillyfactor
marklevinshow
THEHermanCain
SenJohnMcCain
gabriel_oficial
marcelostrike
hannityshow
gretawire
GovernorPerry
IngrahamAngle
RealAlexJones
TomCoburn
foxnewsalert
theblaze
mkhammer
foxnation
EricCantor
GovChristie
reasonmag
TuckerCarlson
foxnewspolitics
NRANews
brunasald
laurawbush
AnnCoulter
DanaPerino
BarackObama
krauthammer
Bret_Baier
RepPaulRyan
PrAbnerFerreira
JorgeBernal
JeffFlake
megynkelly
andylevy
AndrewBreitbart
newsbusters
FoxBusiness
secupp
fredthompson
RonPaul
MittRomney
Ericatwitts
InfowarsFeed
nikkihaley
JebBush
Israel
USCCB
thehill
RedState
DrRandPaul
DennisDMZ
Dr_Rose
Senate_GOPs
Judgenap
TeamBachmann
SenateFloor
usairforce
scrowder
IDFSpokesperson
BraveLad
AM_National
KofC
julianperretta
JohnCornyn
israelconsulate
kimkomando
KOSMOSNET
CarrollStandard
NBCConnecticut
GroverNorquist
RepublicanGOP
latimestot
jasoninthehouse
tedcruz
andilinks
NRCC
ScottWGraves
Wizard101
HumanEvents
Justicereport
lauraingle
FLGovScott
MegWhitman
infowars
ErnieFowlke
AlinskyDefeater
davidicke
CatholicNewsSvc
foxheadlines
fxnopinion
George_WBush
BillSchulz
keibosite
mattklewis
mikeallen
AndreaTantaros
WSJopinion
johnculberson
RumsfeldOffice
grivno
RealSheriffJoe
NTARC
allahpundit
FreedomWorks
BluegrassPundit
NavyNews
USMC
JBN
lr3031
DeptofDefense
StossselUpdates
jaketapper
BobbyJindal
StevenErtelt
GovWalker
milblogging
IsraeliPM
JordanSekulow
RonPaulcom
marthamaccallum
weeberly
slkbrooke
JedediahBila
Sarah_Palin
news_va_en
weeklystandard
RichLowry
BillyHallowell
AmSpec
Michael_Yon
newadvent
TPPatriots
COPMhip
KatyInIndy
EdDriscoll
CatholicNewsSvc
AlinskyDefeater
FlaGuarita
HouseFloor
EWErickson
RedEyeFNC
jamiedupree
greggutfeld
MajoratNJ
LizClaman
juliegbanderas
rollcallpols
GlennBeckNews
PopeBenedictXVI
BOTeleprompter
DraftMarcoRubio
kirstenpowers10
ConNews
AllenWest
johnthune
ChuckGrassley
netanyahu
TedNugent
OrrinHatch
RonPaul
rollcall
theMRC
STU_GBP
foxnewsradio
coutpost
NROcorner
JonahNRO
DailyCaller
ALIPAC
NRO
RedState
cspan
DrRandPaul
Israel
JIDF
JohnKasich
timpawlenty
stephenfhayes
HaleyBarbour
mcdev
f396
TexGov
AACONS
livedesk
crnc
DRUDGE
USArmy
BillHemmer
TrueCity
ChristineO'O
AccuracyinMedia
SenatorBurr
bchateaubriand
mikepfs
GOPLeader
marcorubio
RepRonPaul
FredBarnes
GB_TV
VersosSencillos
dickmorristweet
ConserValidity
SenRandPaul
NRAILA
WSJ
procopy
RepDaveCamp
kathrynlopez
Politisite
WTNH
NYPDretCOP
amandacarpenter
infowarsstories
thejointstaff
gopconference
RickSantorum
TheNatlGuard
Jerusalem_Post
Darrellissa
EdMorrissey
hughhewitt
AlyatFox
NationalDebt
MelissaTweets
Daggy1
wbconservative
elemonader
PatrickRuffini
LPNational
JewishTweets
thecatholicsun
TaoMannaDon
redcounty
cnalive
coasttocoastam
hipEchik
ConservativeGal
LockheedMartin
the_USO
PJTV
The_RGA
NCPA
CarlRinsch
stephenkruiser
RonPaul_2012
US_Air_Force
ByronYork
HeyTammyBruce
SarahPalin_Blog
ExJon
DeptVetAffairs
BreakingNews
FBNStossel
NTU
E_Stampede
RepMikePence
GarySinise
kilmeade
AIPAC
ScottBrownMA
GovBrewer
biggovt
Jennafnc
IndyEnigma
TeaPartyNJ
ThadMcCotter
mercats
snuats
kmita3
politico
mises
DLoesch
usaa
USNavy
Lileks
sdoocy
secupp
NRSC
AFPAA
f396
ALIPAC
sanuzis
foxnewsradio
coutpost
NROcorner
JonahNRO
DailyCaller
WOTN
TedNugent
OrrinHatch
slkbrooke
weeberly
RonPaulcom
JordanSekulow
AmSpec
Michael_Yon
newadvent
israelconsulate
kimkomando
NBCConnecticut
GroverNorquist
PajamasMedia
NolteNC
lewrockwell
RCP_Articles
Peggynoonannyc
ShannonBream
TheFlaCracker
DeptVetAffairs
E_Stampede
GarySinise
AIPAC
Jennafnc
IndyEnigma
ThadMcCotter
kirstenpowers10
ConNews
AllenWest
johnthune
ChuckGrassley
netanyahu
TedNugent
OrrinHatch
JedediahBila
slkbrooke
weeberly
RonPaulcom
JordanSekulow
AmSpec
Michael_Yon
newadvent
israelconsulate
kimkomando
NBCConnecticut
GroverNorquist

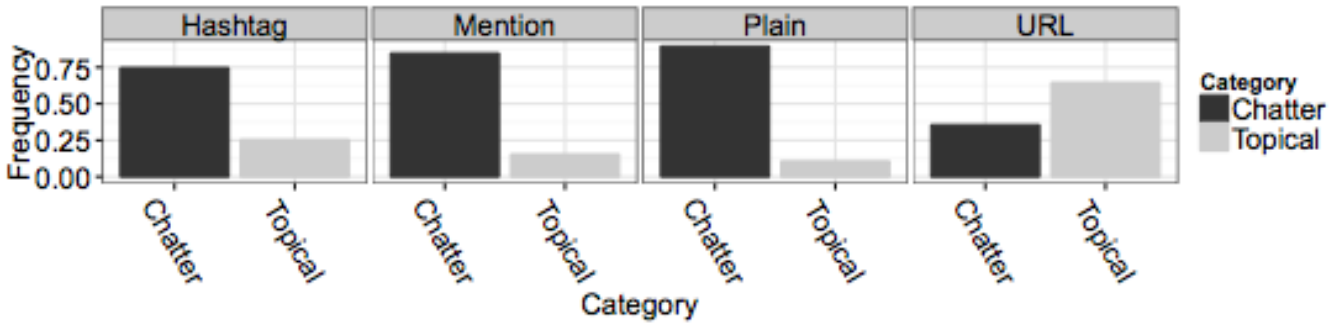
Topical vs chatter

Topical Topics
* potter, harry, tumblr, fandom, weirdly, gifs, gif, hogwarts, fictional, rewatch, hp, rowlingpuns, obvs, deathly
* rep., ballot, officials, mayor, gov, votes, investigation, ballots, mayor's, committee, capitol, district, courthouse, senate
* donation, charity, donated, awareness, donations, donate, autism, raise, help, funds, fundraiser, fundraising, donating, support
* investors, financial, banks, debt, markets, goldman, finance, banking, stocks, economic, earnings, ipo, bank, equity, investment
Chatter Topics
* foh, lmaooo, lmaooooo, lml, lmaoo, deadass, henny, lmaooooo, niggas, djzeeti, smfh, nah, nigga, tho
* coworker, washer, dangit, dryer, yeah, i've, beeping, oh, 6ish, i'll, 10ish, nope, 4am, kinda, probably
* thanks, enjoyed, congrats, big, next, week, soon, coming, everyone, well, incredible, wow, meeting, achievement, weekend
* someone, because, hate, sometimes, anymore, she, person, tell, her, aren't, saying, sleep, enough, without, ask, real, money

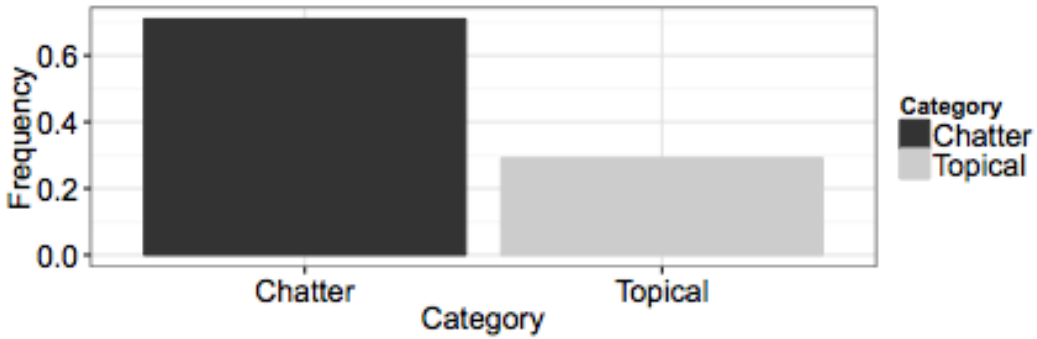
Chatter detection



Chatter detection



(b) Chatter incidence across tweets with different characteristics.



(c) Overall sample chatter incidence.

International support

- Over 70% of users reside outside of US
- Localization is important (over 35 languages)
- Content modeling in a variety of languages
- Topic distribution not the same for different languages
- Quantity (skew) of data across different languages
- Many users are multi-lingual

Bootstrapping topic models

- English tends to be the first choice in developing supervised or unsupervised topic models
- Acquisition of labeled data is expensive and time consuming
- Can an English corpus augmented with automatic translation technology be build models in other languages
- If additional labeled data in the target languages is needed (e.g., for evaluation), how to minimize the extra amount

Spammy interactions

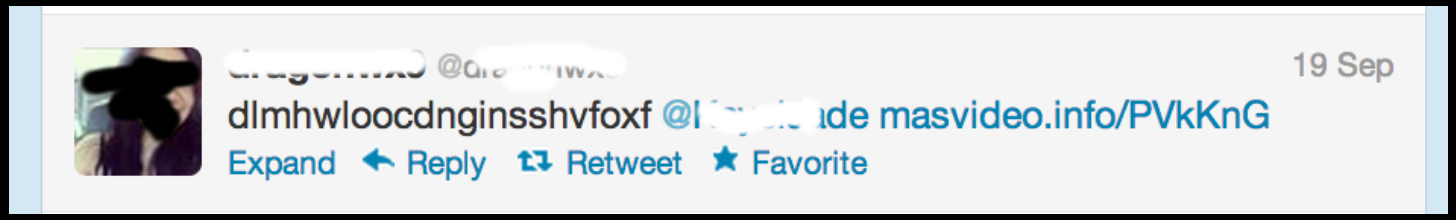
Adversarial Classification (spam)

- Avoidance of being gamed by spammers
- Research into how spammers operate
- Types of spam
- Account creation vs purchasing

Adversarial games

- Spammers try to evade detection defenses
- The tug-of-war leads to the shift of tactics over time
- Tricks cover randomization, hiding of payload and trying to appear to be “normal”

Tweet level


19 Sep
dlmhwloocdnginsshvfoxf masvideo.info/PVkKnG
Expand
Reply
Retweet
Favorite


#syria OMG!! This helped me laugh =)) HAHAHAHAHAHAHA
bit.ly/RlxnFT
View summary
Reply
Retweet
Favorite


16 Sep
Hi there
ar Shirley had said you might be sincerely
interested in
motion pofile
Expand
16 Sep
Whats up
Zachery had said you will be interested in
motion pofile
Expand
Reply
Retweet
Favorite

Tweet-history level

swing 081
Expand

Whoa what a awesome day it is!
Expand

@Z... you've been chosen to be in the next ... video
click @Be... and follow the directions
View conversation

Whoa its such a awesome day;)

once barnett1
Expand Reply Retweet Favorite

@Th... We are filming near you for a brand new movie
that stars Justin Bleber ! We want you to be in it >>> @...
View conversation

Tess swanson @Tess87ldn 18 Sep
in order that Catton123
Expand

This year?s report has just been released and there are some
interesting trends that personal branders can draw from the report
Expand

I will give you 5000 twitter followers in less than 24 hours
bit.ly/OQssfM @...

If you are a Coca cola lover or a Coca cola drinker then you must see
this post in this post i have shown
Collapse Reply Retweet Favorite

Account sales

Pricing Per Thousand Accounts

$15-60

$5

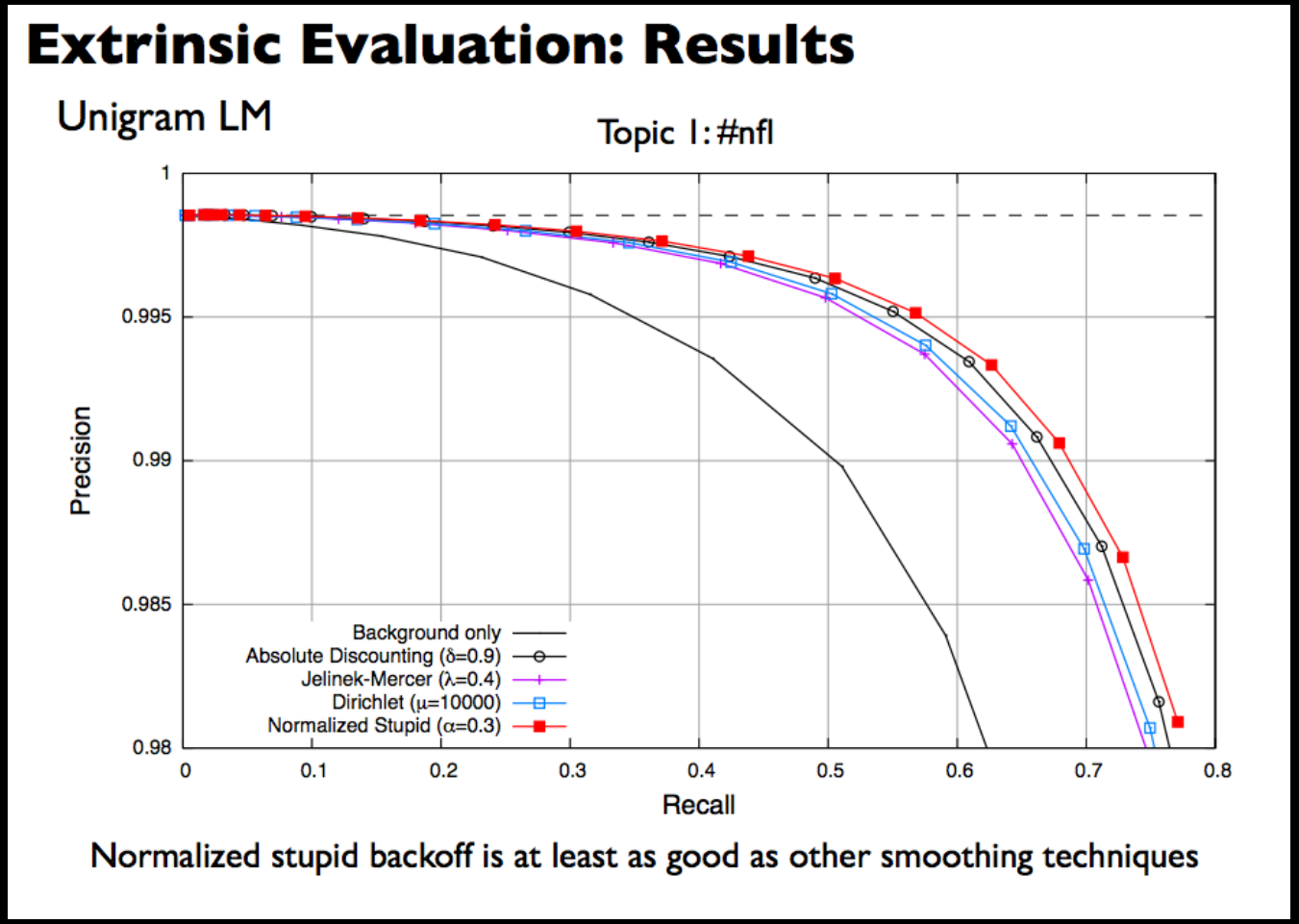
$100

Legitimacy signals

- Email confirmation rate: 80%
- Unique IP signup rate: 87%
- Accounts are “pre-aged”
- Markets exist for both blank accounts and accounts with a social network
  - Prices vary

Big Data vs Fast Data

Hashtag topic tracking



Source: Lin et al. "Smoothing Techniques for Adaptive Online Language Models: Topic Tracking in Tweet Streams", KDD 2011

“Real-time” human computation

- Spiking queries often correspond to new events/topics.
  - Cold start problem
- Semantics of the intent may be quite different from the “usual” semantics of the query.
  - “Clint Eastwood” during the 2012 presidential debate
- The lifetime of the spike is short-lived.
  - A lot can be learned from user actions clicks, but by that time the query may have stopped trending

“Real-time” human computation (crowdsoursing)

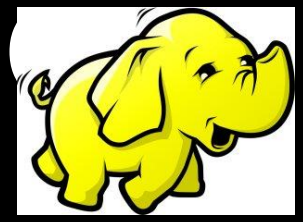
- Solutions – ask humans!
- Crowdsourcing systems allow one to have question answered with short turnaround time
  - Low cost allows one to ask many questions.
- Humans-in-the-loop can be automated
  - Use of MTurk and internal APIs
- Quality control
  - Custom pool of judges

Plumbing

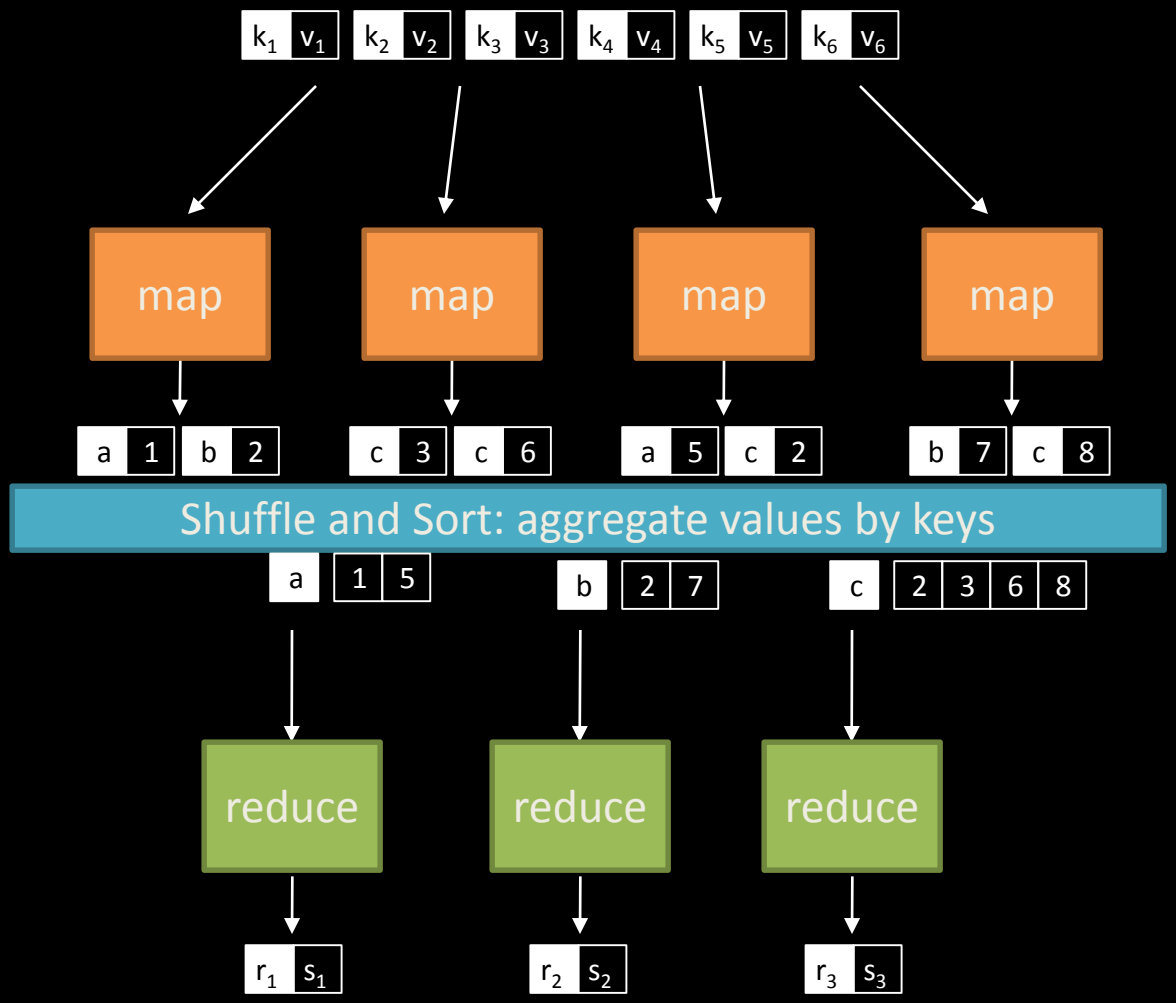
Scalable analytics/infrastructure

- Logging/archiving and computing at the same time
- Ability to recover from and catch up from disruptions (crashes, outages, etc)
- Planning for the needs of an ever-growing number of analytics jobs

MapReduce/Hadoop



Pig



Source: Lin and Dyer (2010)

```
visits        = load '/data/visits' as (user, url, time);
gVisits       = group visits by url;
visitCounts   = foreach gVisits generate url, count(urlVisits);
urlInfo       = load '/data/urlInfo' as (url, category, pRank);
visitCounts   = join visitCounts by url, urlInfo by url;
gCategories   = group visitCounts by category;
topUrls       = foreach gCategories generate
    top(visitCounts,10);

store topUrls into '/data/topUrls';
```

Scalable machine learning

- Large scale machine learning on Hadoop
  - Big data streaming with stochastic gradient descent
  - Parallel training with data randomization (linear and tree-based ensembles)
- Trade-offs between learning with true MR jobs vs using MR for pre-processing and doing the learning part on a cluster of large boxes with hdfs access
  - There a limitation of RAM size of a reducer (e.g., 3-6GB)
  - Hadoop clusters tend to be composed from smaller boxes
  - Sometimes it is necessary to iterate over a dataset of significant size (e.g., 100-200G B)

What kind of models?

- Most of our data is text based
- Models should be able to consume large amount of training data
  - and possibly a large number of categories
- Emphasis on simplicity and speed

Model training UDF internals

- A single node in the Hadoop cluster does not have extensive memory resources
- The learner cannot cache too much data
- Natural fit for (not iterating over all data):
  - Stochastic gradient descent (SGD), possibly with mini-batching
  - Effective for streaming the whole dataset through a single learner

Supervised classification in a nutshell

Given $D = \{(\mathrm{x}_i, y_i)\}_i^n$

label

(sparse) feature vector

Induce $f : X \rightarrow Y$ s.t. loss is minimized

empirical loss = $\frac{1}{n}\sum_{i=0}^{n} \ell(f(\mathrm{x}_i), y_i)$

loss function

Consider functions of a parametric form:

$$\arg\min_{\theta} \frac{1}{n}\sum_{i=0}^{n} \ell(f(\mathrm{x}_i;\theta), y_i)$$

model parameters

Key insight: machine learning as an optimization problem!

(closed form solutions generally not possible)

Gradient Descent

$$w^{(t+1)} = w^{(t)} + \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla \mathrm{l}\left(f(\mathrm{x}_i; \theta^{(t)}), y_i\right)$$

"batch" learning: update model after considering all training instances

Stochastic Gradient Descent (SGD)

$$w^{(t+1)} = w^{(t)} + \gamma^{(t)} \nabla \mathrm{l}\left(f(\mathrm{x}; \theta^{(t)}), y\right)$$

"online" learning: update model after considering *each* (randomly-selected) training instance

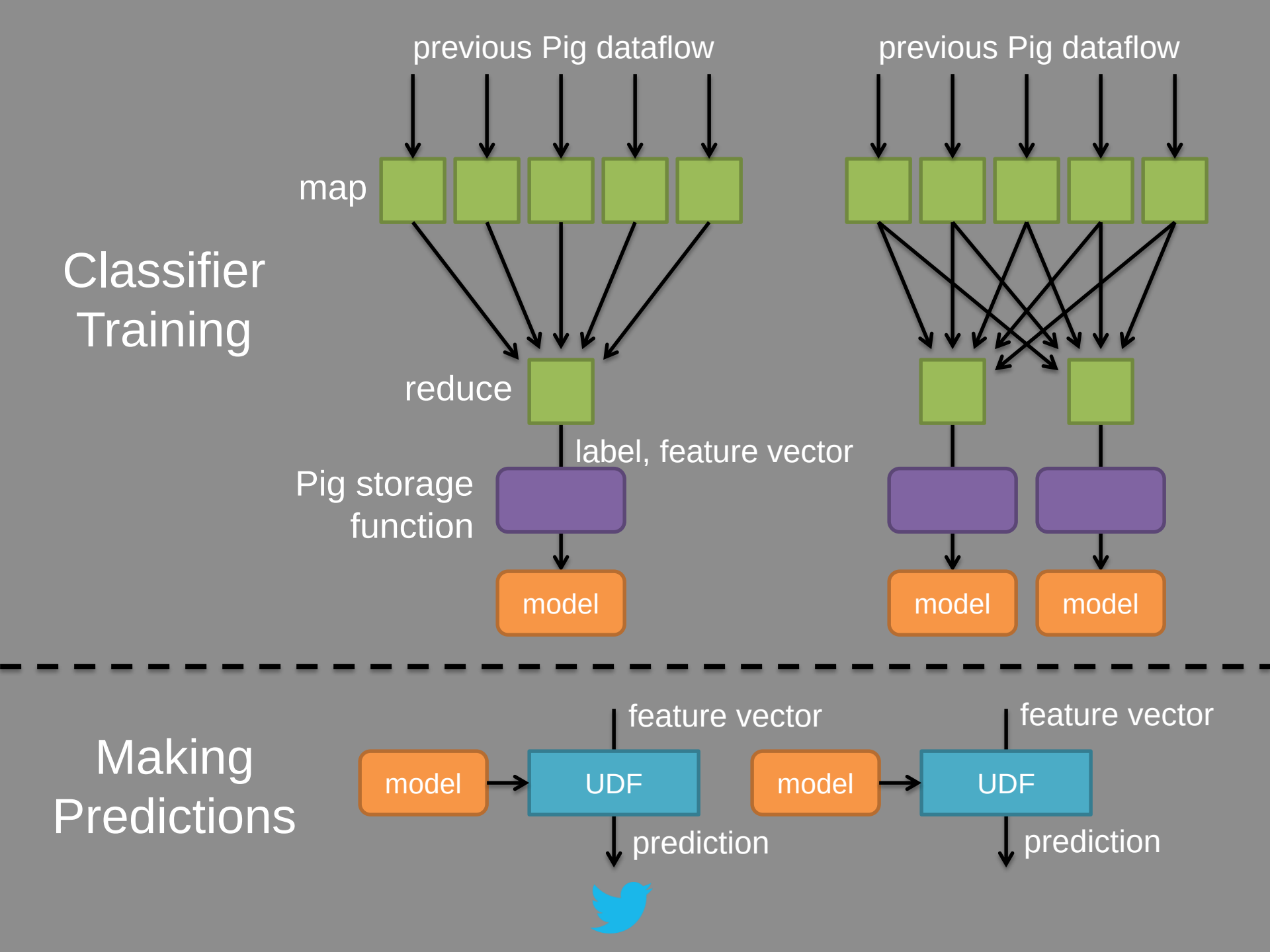
In practice… just as good!

Solves the iteration problem!

What about the single reducer problem?

Another way - Ensembles

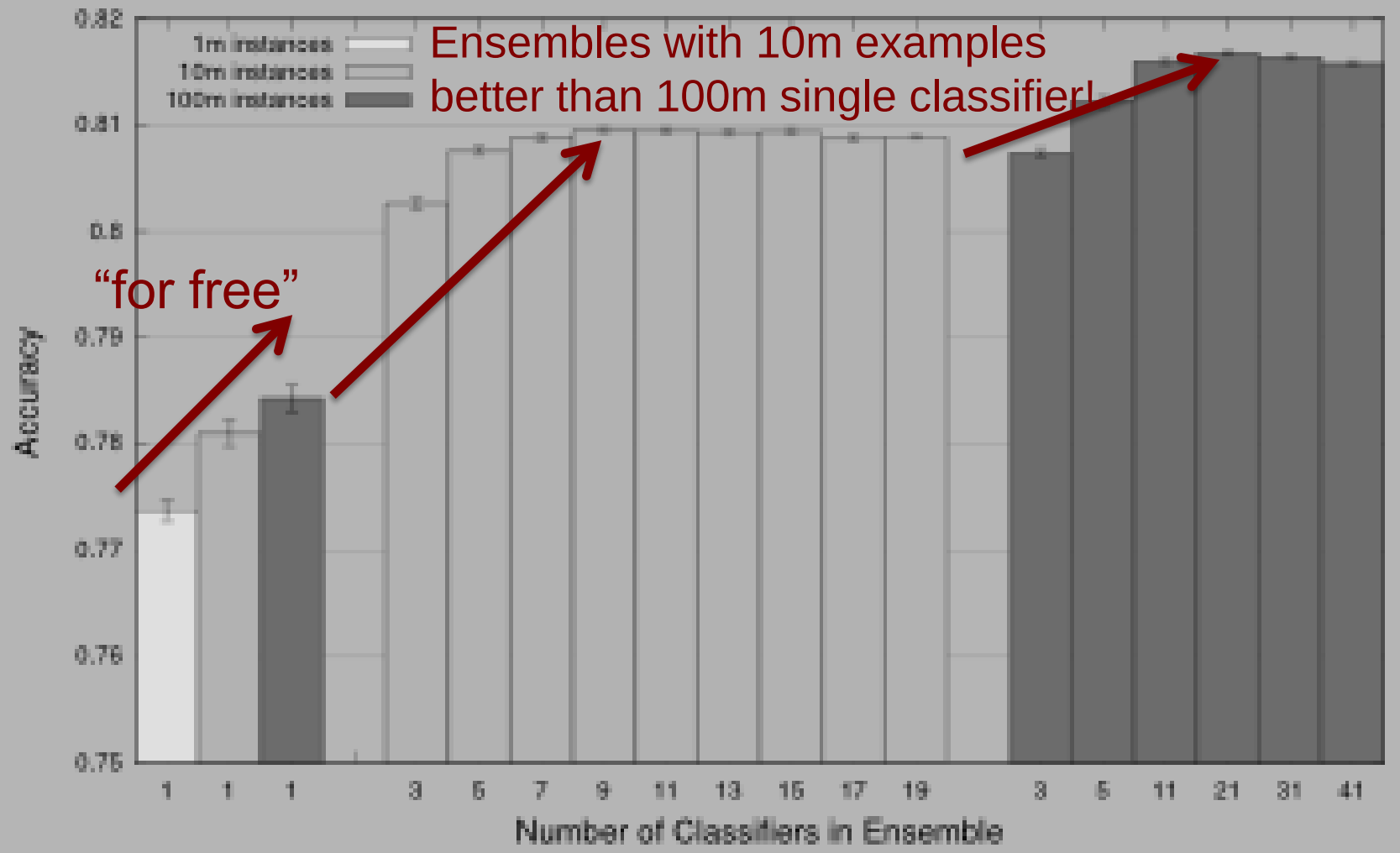
- Classifier committees are one of the best performing types of learners
- Some of these algorithms are sequential (not very MR friendly)
  - Boosting
- But others rely mostly on randomization
  - Each learner is trained over a different split (features and/or instances) of the data


Classifier Training
previous Pig dataflow
previous Pig dataflow
map
reduce
label, feature vector
Pig storage function
model
model
model
Making Predictions
feature vector
feature vector
model
UDF
model
UDF
prediction
prediction

Example: tweet sentiment detection

- Training/Test data: tweets with *emoticons*
  - ☺ ☹
- Emoticons provide surrogate labels for training
  - Emoticons are removed from the data
- Logistic Regression trained over character 4grams
- Single classifier vs. use of ensembles

Diminishing returns…



single classifier 10m ensembles 100m ensembles

Stream-based machine learning

- Twitter is largely about real-time
- Global and personal relevance models need to respond quickly to the changes of content distribution as well as user-action history.
- Technologies to handle real-time streams
  - Storm (open sourced)
  - Several application specific custom solutions
  - Write-through caching
- Learning
  - Updating models vs. updating features
  - Stochastic gradient descent (e.g., with Logistic Regression)

Big boxes have a role

- Cassowary
  - Twitter's WTF service served by redundant single box systems 2010-2012
- Big-box supervised learning
  - 200GB 12-core boxes at the upper end of "commodity"

THANK YOU