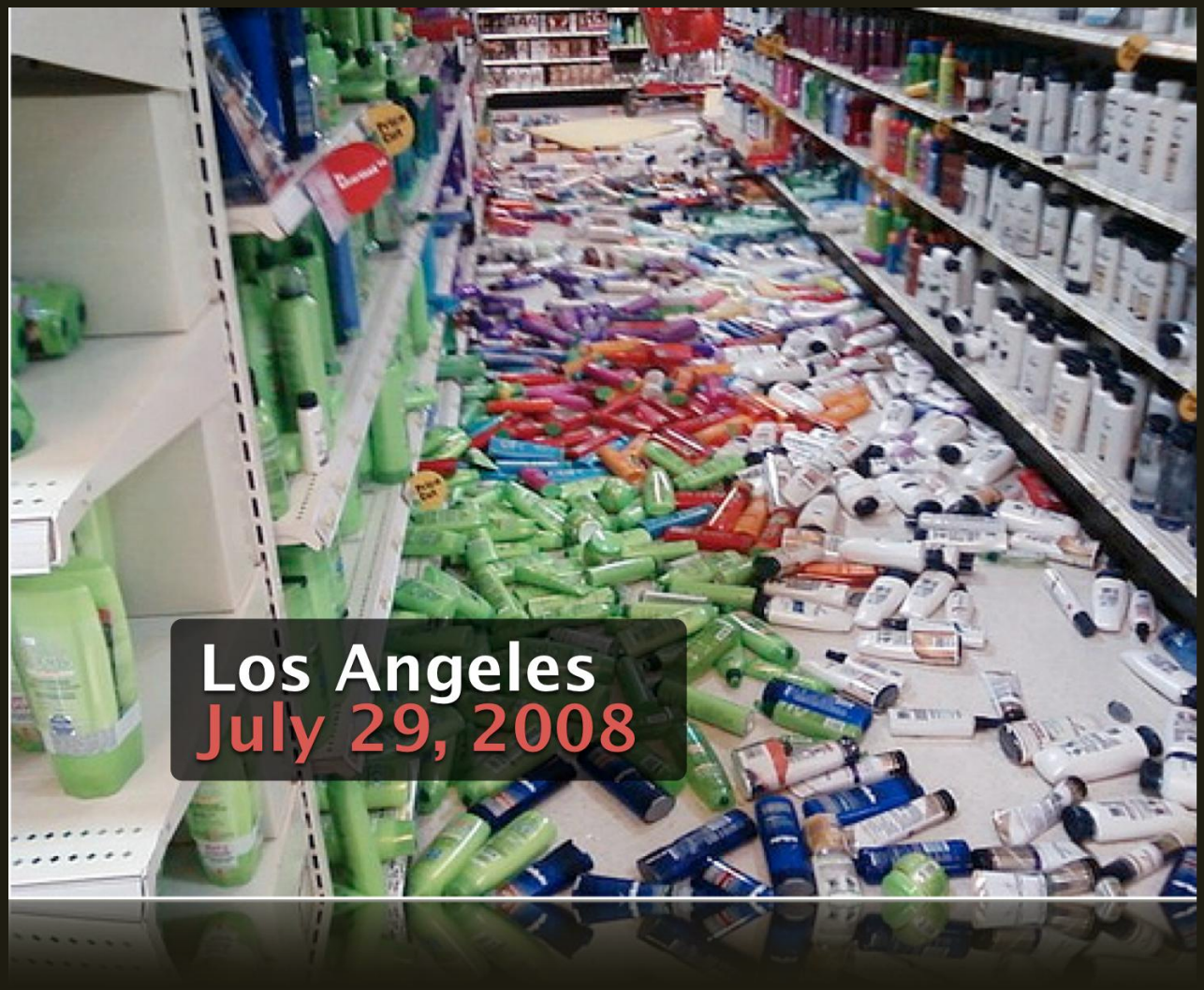
Large Scale (Machine) Learning at Twitter

Aleksander Kołcz

Twitter, Inc.



Earthquake: 0 sec

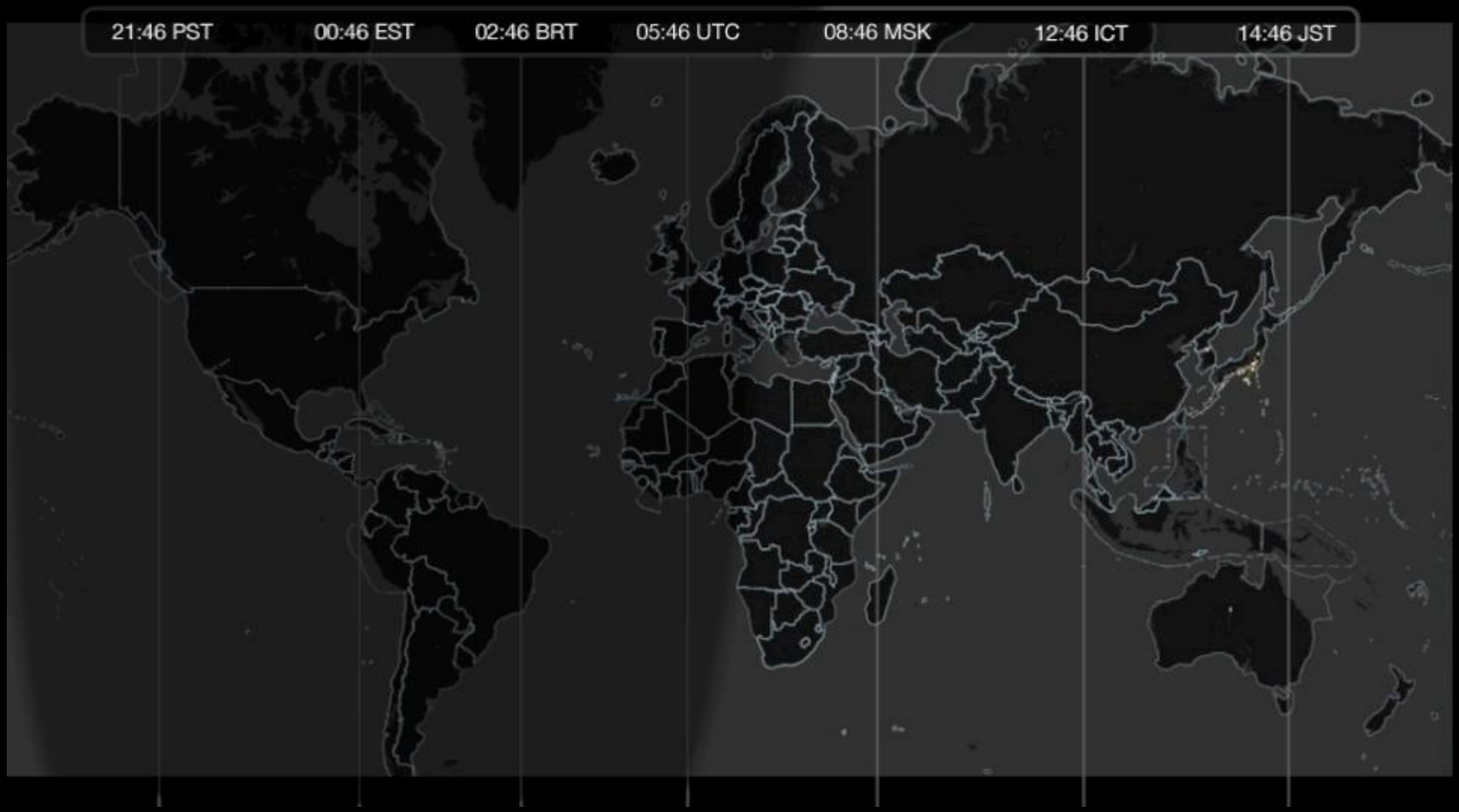
First Tweet: 5 sec

Local News: 4 min

AP Wire: 9 min

Book of Tweets

twitter



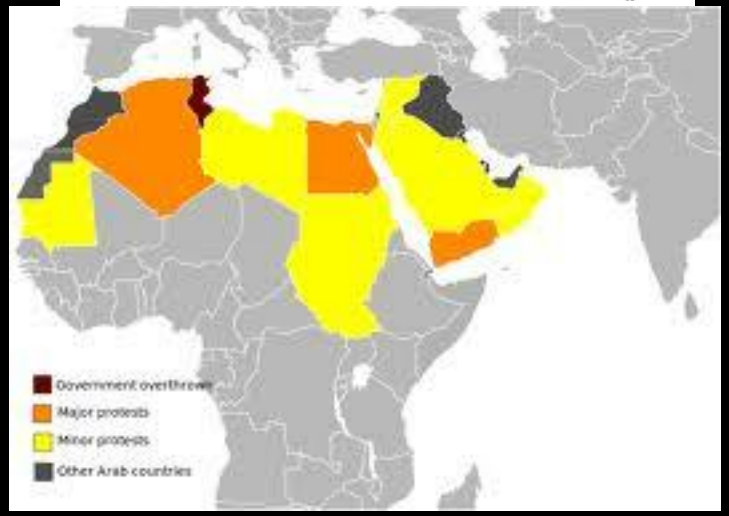
twitter

twitter

What is Twitter

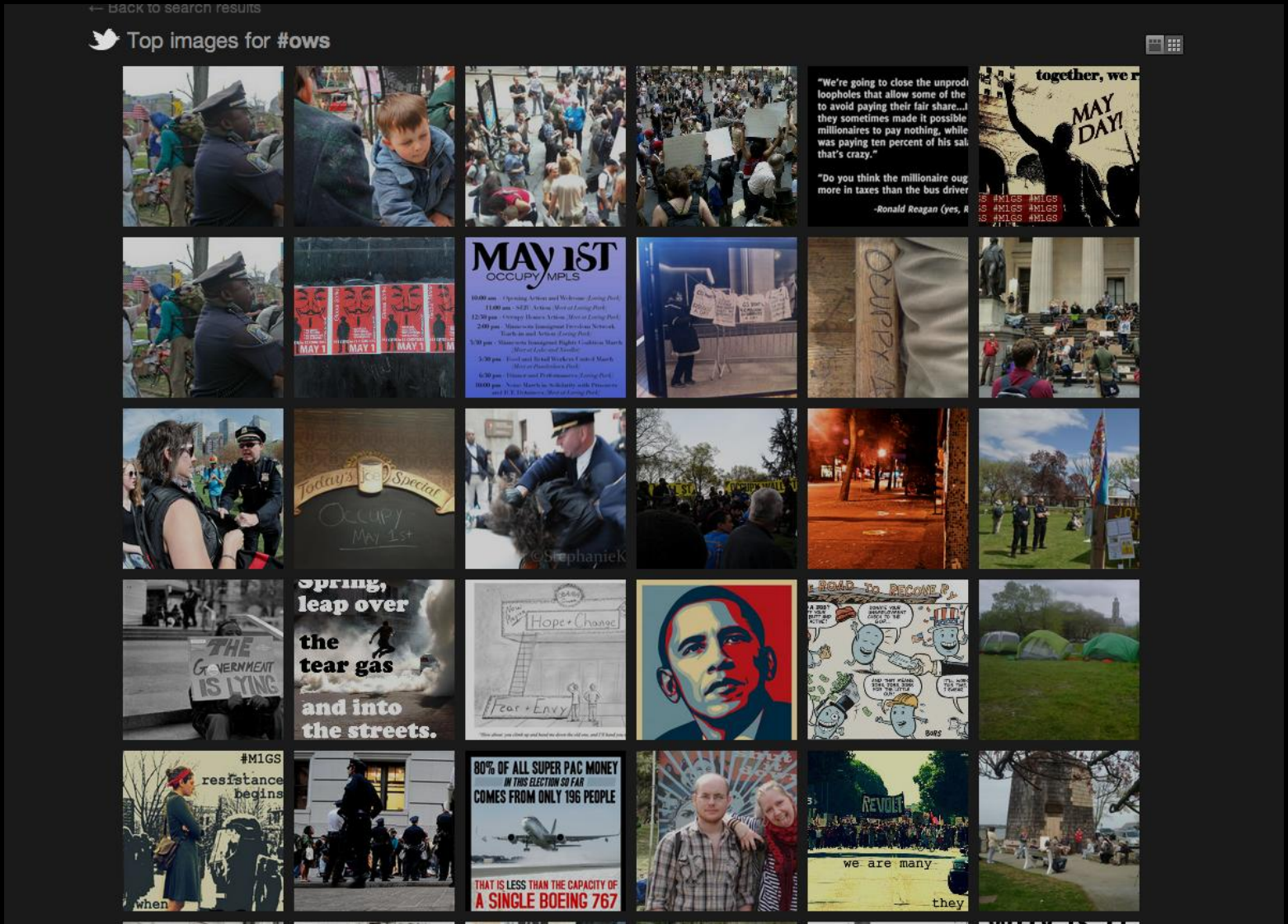
- Micro-blogging service
- Open exchange of information/opinions
  - Private communication possible via DMs
- Content restricted to 140 characters
- Seems tiny but …

#egypt, #tunesia, #libya, #syria …



twitter

#ocws, #occupywallstreet, …



The Scale of Twitter

- Twitter has more than 140 million active users
- 340 million Tweets are sent per day
- Over 400 million monthly unique visitors to twitter.com
- 50 million people log into Twitter every day

twitter

Large scale infrastructure of information delivery

- Users interact via web-ui, sms, and various apps
- Over 55% of our active users are mobile users
- Real-time redistribution of content
- Plus search and other services living on top of it
  - E.g., 2.3B search queries/day (26K/sec)

twitter

Support for user interaction

- Search
  - Relevance ranking
- User recommendation
  - WTF or Who To Follow
- Content recommendation
  - Relevant news, media, trends

- ML is an important component but really just a part of a larger whole

twitter

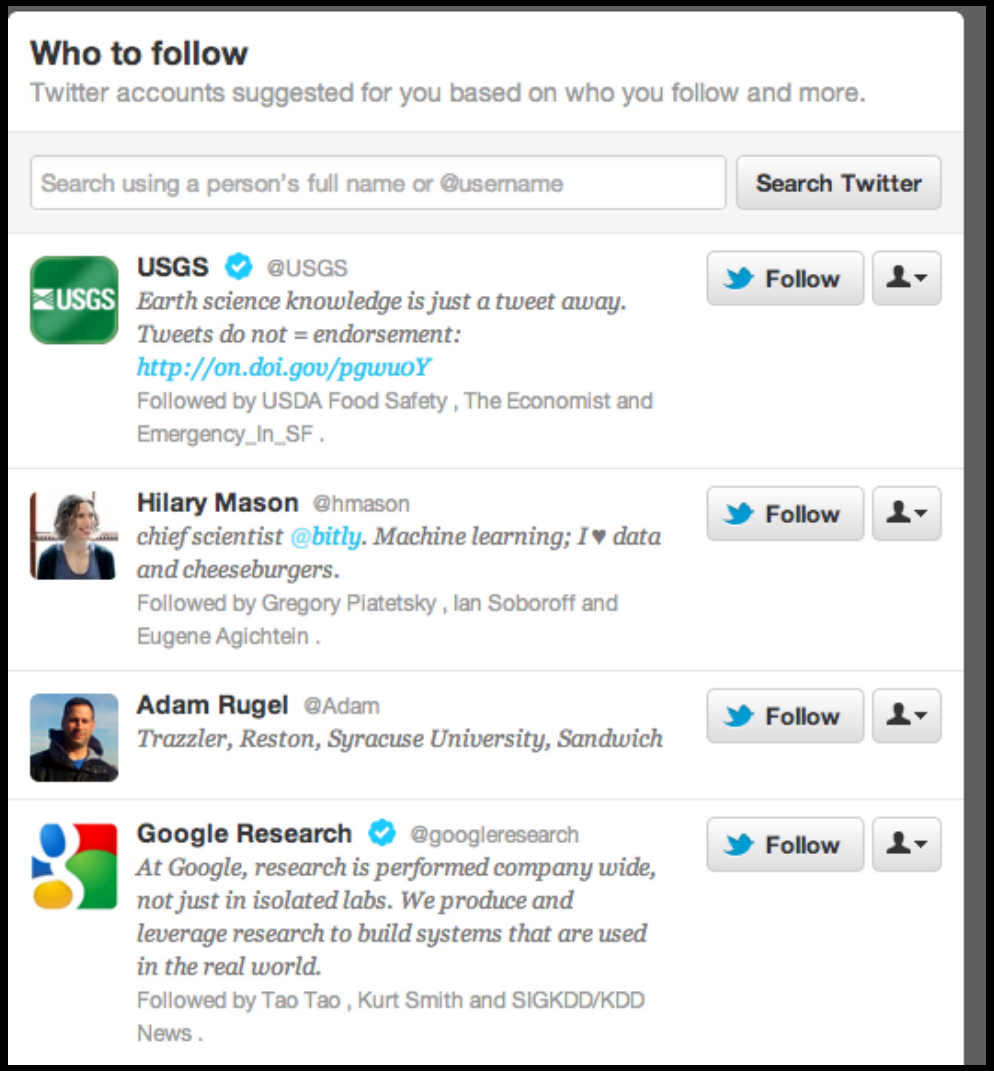
Problems we are trying to solve

- Relevance
  - ranking in search

twitter

Problems we are trying to solve

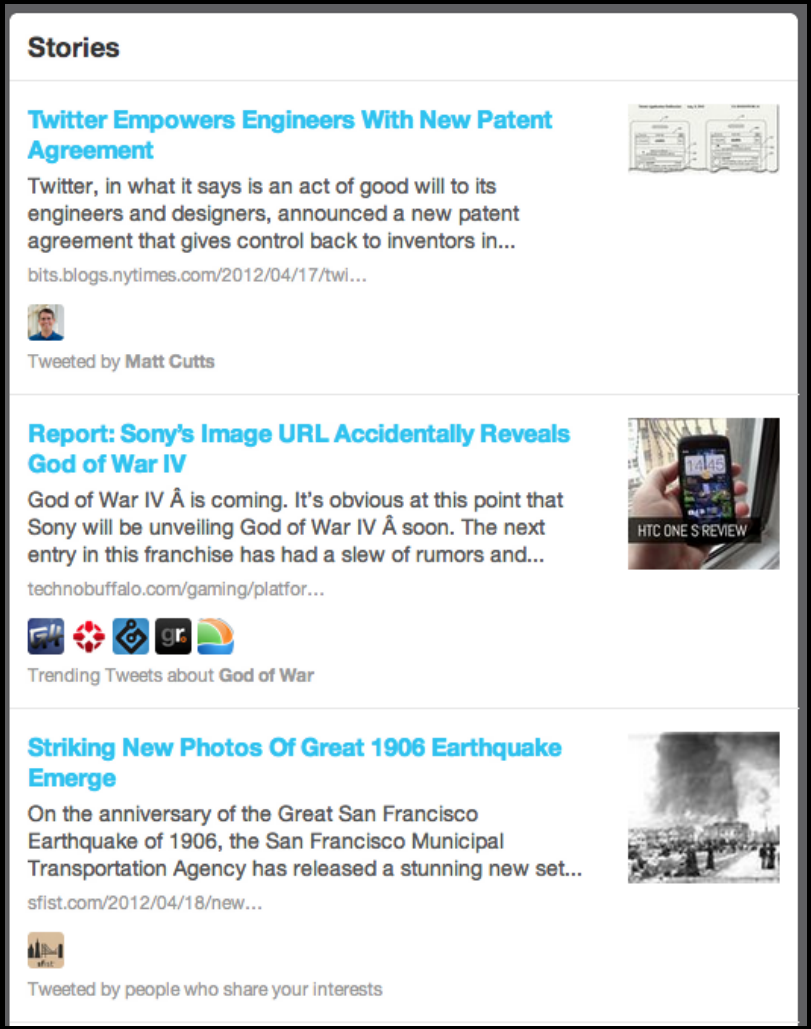
- Who to follow



twitter

Problems we are trying to solve

Content recommendation (stories/media)



twitter

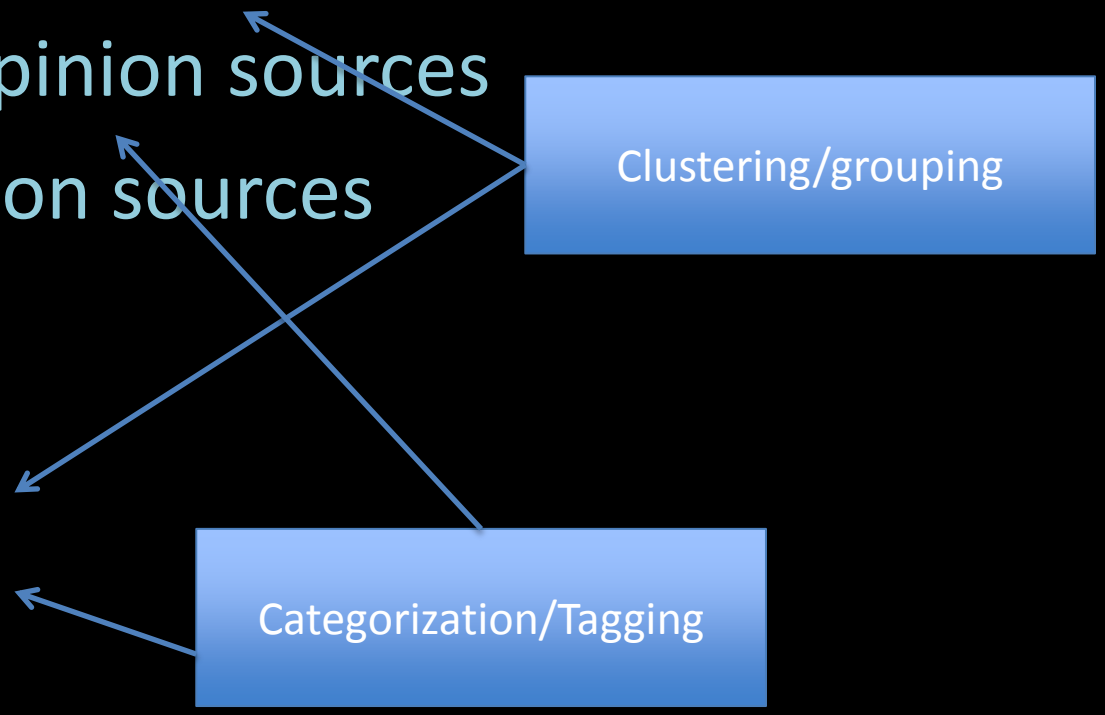
(other) problems we are trying to solve

- Trending topics
- Language detection
- Anti-spam
- Revenue optimization
- User interest modeling
- Growth optimization

twitter

Recommendation/Personalization

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



twitter

Is this BIG data ?

twitter

Challenges

- Twitter is International
  - 70% of accounts are outside the US
  - Twitter supports more than 28 different languages
- Twitter is real time
  - The “now” factor is very important for people interacting with Twitter
  - The concept of relevance is highly time dependent
- 140 characters : “documents” are very short and vocabulary contains many non-standard acronyms
  - *@XXXX_ Cool*
  - *Please please please*

twitter

What type of machine learning?

- At this point ML is an extensive and diverse field
- Algorithms differ widely in terms of implementational and operational complexity
- Decisions which techniques to use are driven by the nature of the problems and infrastructure/operational constraints

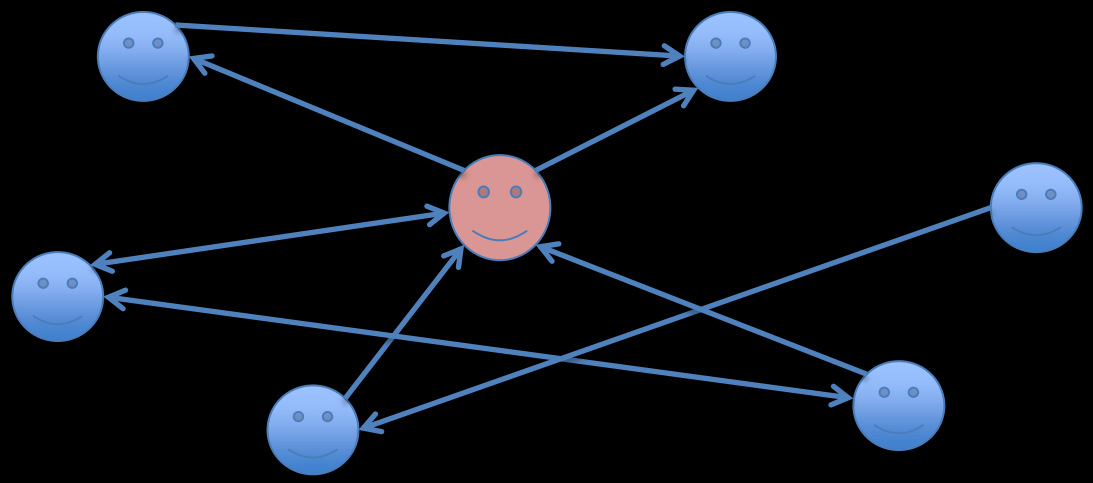
twitter

ML for social networks

- Power of simple models
  - It has been found again and again that with enough data relatively simple models (e.g., Naïve Bayes or Logistic Regression) work remarkably well
  - Ease of integration with data processing flows (both training and inference) trumps model sophistication

ML for social networks

- Power of graph aggregation
  - Even an imperfect signal (e.g., classifier) can be useful when its effect on a graph node is taken in the context of the node's neighborhood



twitter

ML for social networks

- Challenges of scalability and adaptive processing
  - Learning and inference need to handle data streams and combinations thereof
  - Models need to be able to quickly adapt to data stream change

Are there wheels not to reinvent?

- Temptation is always there ….
- Data processing infrastructure?
- Core ML algorithm libraries ?
- Data pre/post processing ?
- Visualization ?
- Glue code ?

Analytics Ecosystem

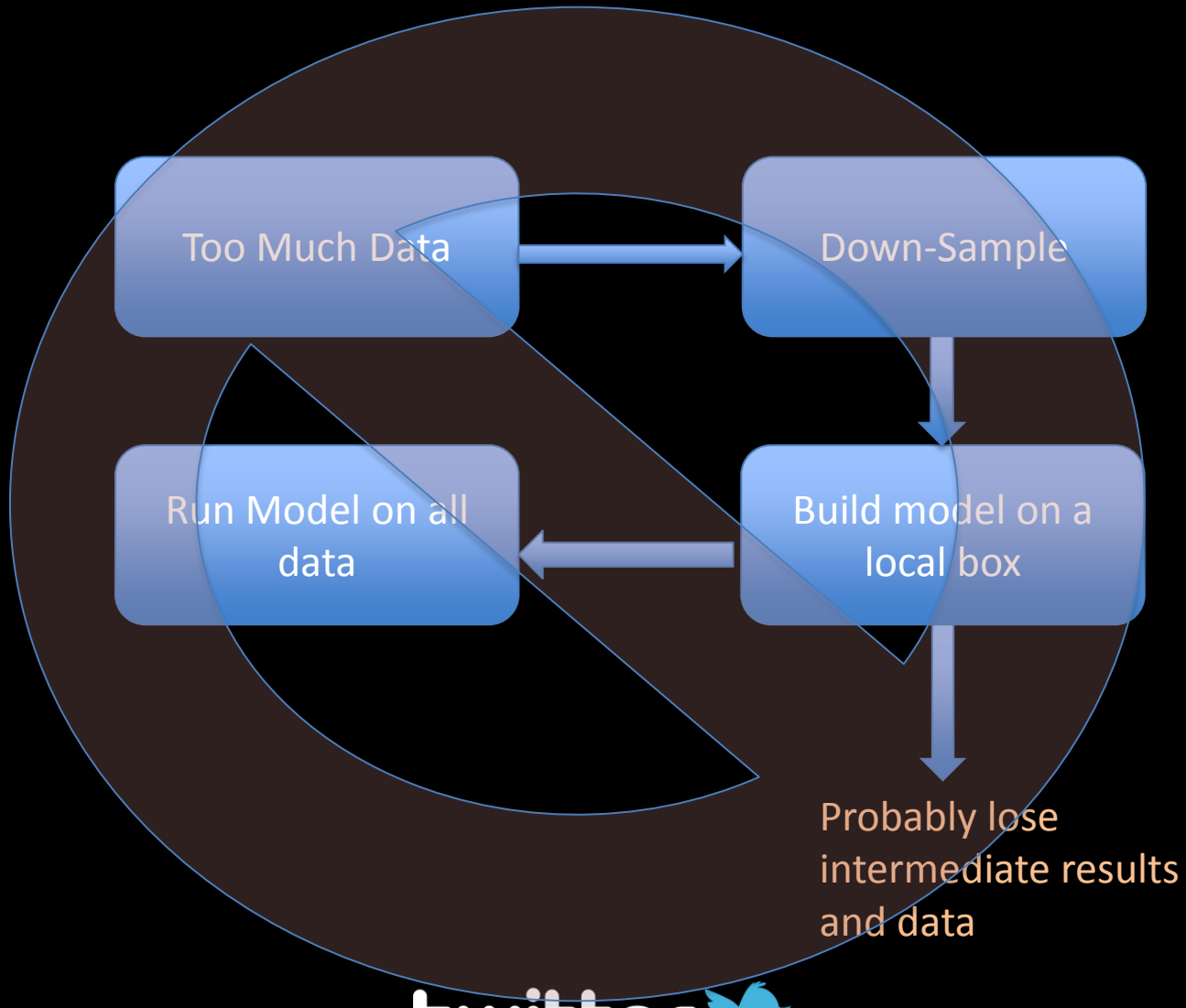
- Apache Mesos

Maximizing the use of Hadoop

- We cannot afford too many diverse computing environments
- Most of analytics job are run using Hadoop cluster
  - Hence, that's where the data live
  - It is natural to structure ML computation so that it takes advantage of the cluster and is performed close to the data

twitter

AVOID: “janky” analysis of messy data



twitter

Leveraging off-line tools

- While the data is big, a lot of useful feature components can be learned with smaller datasets
- The ML tool ecosystem provides a wide range of options for optimization and tuning various models
- Once tuned, the models can be applied to big data in a distributed fashion
  - often not as final models but as feature extractors
- The key is not to rely primarily on ad-hockery in production pipelines

twitter

Large scale learning frameworks

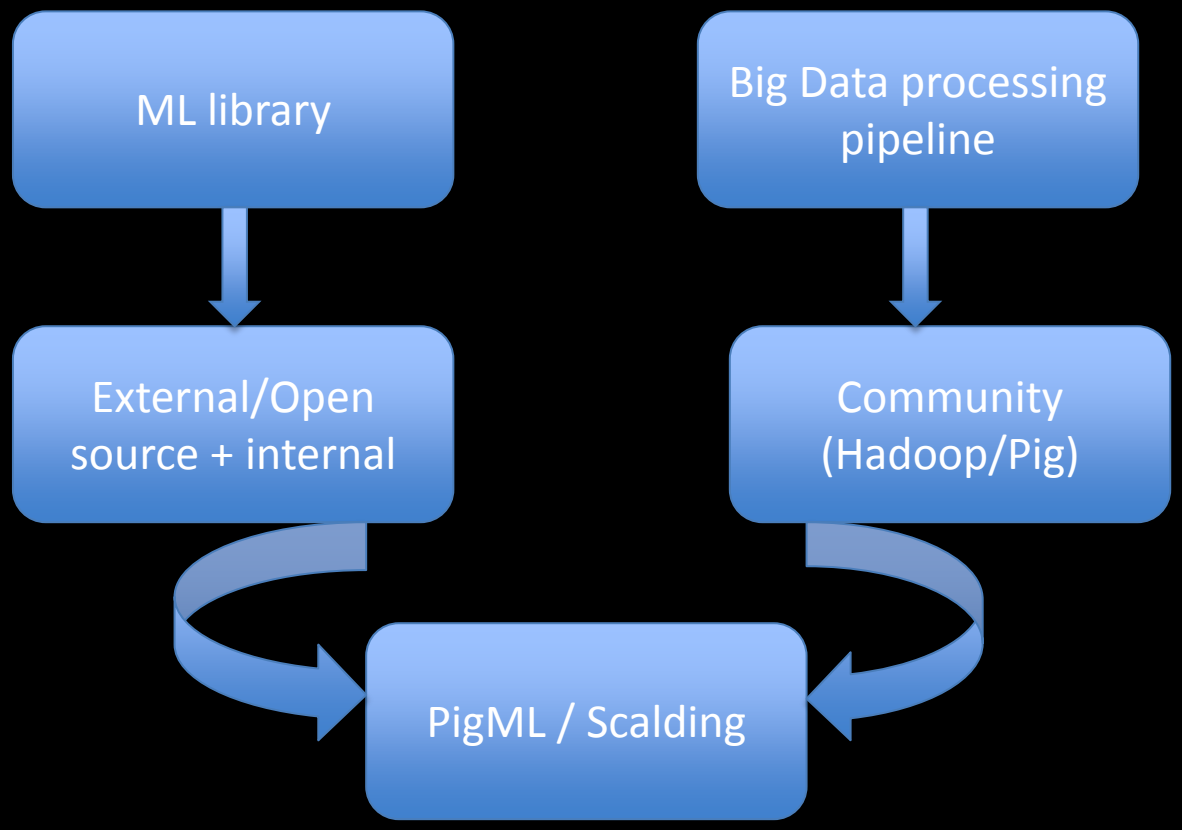
- Participation in the open source community
  - There are a number of initiatives for using ML over Hadoop (e.g., Mahout)
- Upsides
  - Larger support and developer network
- Downsides
  - Not always convenient to integrate with internal analytics and data processing flows

twitter

Our extensions

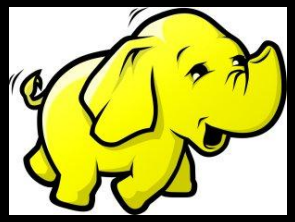
- PigML
  - A library of UDFs wrapping the ML functionality
- Scalding/PyCascading
  - Cascading with Scala/Jython
- ML Java lib
  - Used in both cases to capture the low level ML functionality

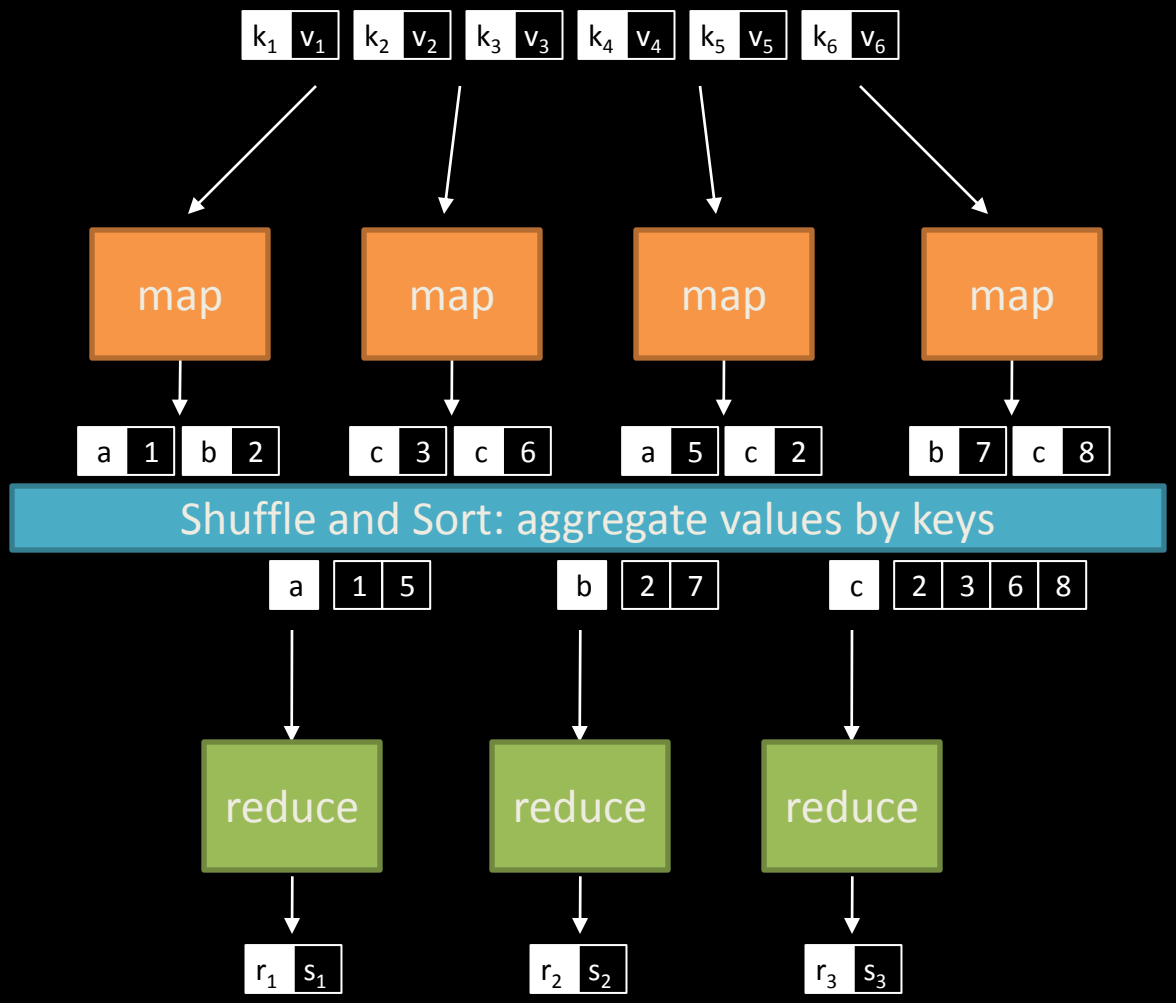
Build/reuse/integrate



twitter

MapReduce





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';
```

twitter

PigML

- Embedding ML learning/eval functionality directly in Pig
- Flows naturally with other data processing operations
- Minimal learning curve for users

Training a model in Pig

```
training = load ’trn_data' using
    piggybank.ml.Storage() as (target: double,
    features: map[]);

store training into 'model-LR' using
    piggybank.ml.train.online.LRClassifierBuilder('with
    Pegasos withLambda:0.1');
```

Training a model in Pig

```
training = load ’trn_data' using
    piggybank.ml.Storage() as (target: double,
    features: map[]);

store training into 'model-LR' using
    piggybank.ml.train.online.LRClassifierBuilder('with
    Pegasos withLambda:0.1');
```

Training a model in Pig

It’s just a store function!

```
training = load ’trn_data' using
    piggybank.ml.Storage() as (target: double,
    features: map[]);

store training into 'model-LR' using
    piggybank.ml.train.online.LRClassifierBuilder('with
    Pegasos withLambda:0.1');
```

Applying a model in Pig

```
DEFINE Classify
    piggybank.ml.classify.ClassifyFeaturesWithLRClassifier('model-LR');

data = load 'test_data' using piggybank.ml.Storage() as (target: double,
    features: map[]);

data = foreach data generate target, Classify(features) as prediction;

results = foreach data generate (target == prediction.label ? 1 : 0) as
    matching;

dump results;
```

Applying a model in Pig

```
DEFINE Classify
    piggybank.ml.classify.ClassifyFeaturesWithLRClassifier('model-LR');

data = load 'test_data' using piggybank.ml.Storage() as (target: double,
    features: map[]);

data = foreach data generate target, Classify(features) as prediction;

results = foreach data generate (target == prediction.label ? 1 : 0) as
    matching;

dump results;
```

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:
  - 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)

twitter

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!

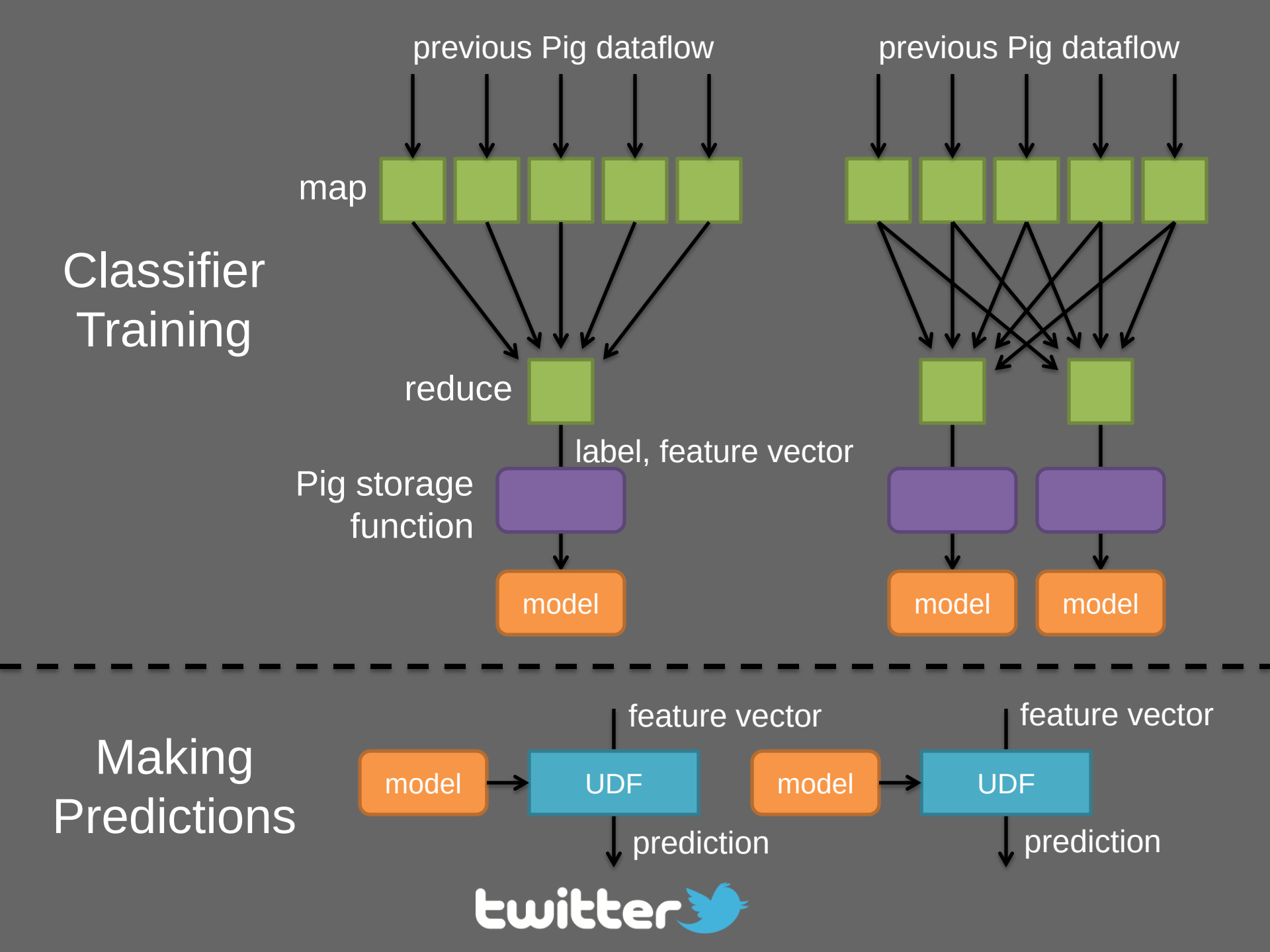
Solves the iteration problem!

What about the single reducer problem?

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

twitter


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
twitter

Ensembles: continued

- Ensembles of linear classifiers
  - E.g., each trained using SGD over different subset of features
- Ensembles of decision trees (random forest)
  - Each tree is seeing only a subset of the dataset and node split variables are randomized at each split

Further advantages of parallelism

- Although stream based learners look at all of the data, a number of them can be executed in parallel
  - Effective for tuning hyper-parameters
- Generative models such as Naïve Bayes are naturally well suited to distributed learning
  - Just counting

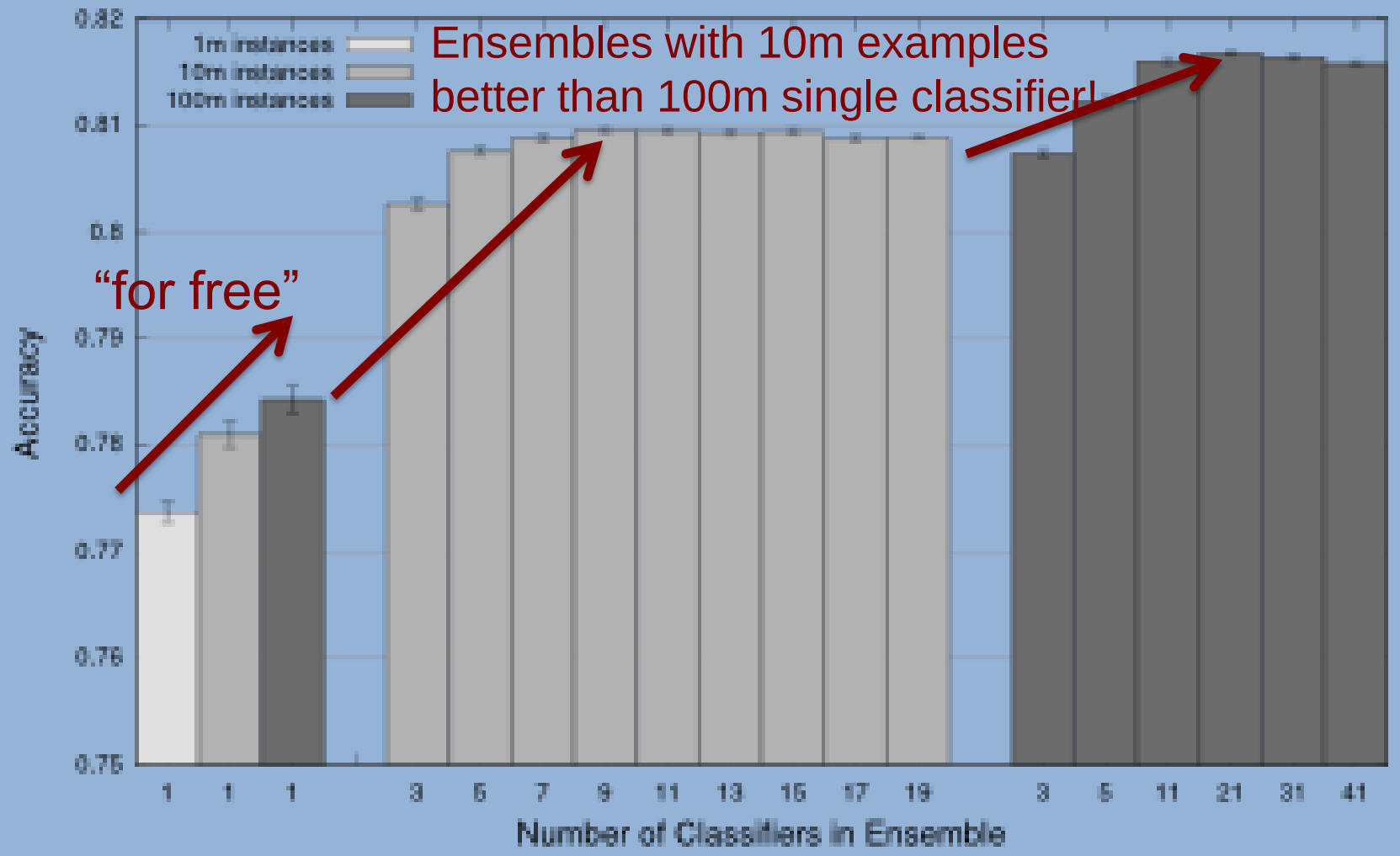
twitter

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

twitter

Diminishing returns…



single classifier 10m ensembles 100m ensembles

twitter

Iterative algorithms

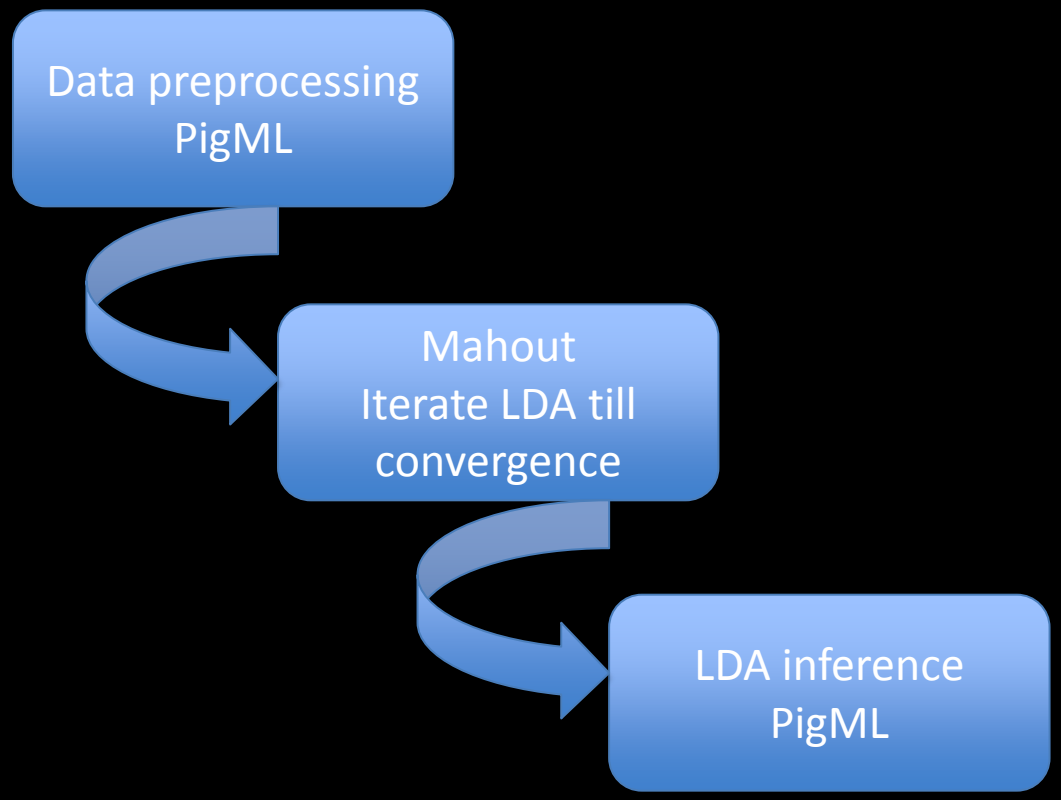
- What if one **REALLY** wants to iterative over big data till convergence?
  - Unrolling of small loops in Pig
  - Implementing the algorithm in Cascading with Scala or Python
  - Using Python Control Flow in Pig
  - Offload to custom MR job (e.g., Mahout)


kenblackwell
Schwarzenegger
LockheedMartin
the_USO
PJTV
The_RGA
NCPA
CarlRinsch
PajamasMedia
NolteNC
thecatholicsun
redcounty
cnalive
coasttocoastam
stephenkruiser
RonPaul_2012
lewrockwell
RCP_Articles
TaoMannaDon
hipEchik
ConservativeGal
Peggynoonannyc
Daggy1
MelissaTweets
US_Air_Force
ByronYork
wbconservative
AlyatFox
NationalDebt
ShannonBream
elemonader
PatrickRuffini
IngrahamAngle
HeyTammyBruce
TheFlaCracker
kathrynlopez
LPNational
hughhewitt
SarahPalin_Blog
JewishTweets
EdMorrissey
ExJon
DeptVetAffairs
Politisite
WTNH
amandacarpenter
RealAlexJones
foxnewsalert
BreakingNews
NYPDretCOP
infowarsstories
theblaze
FBNStossel
NTU
E_Stampede
thejointstaff
mkhammer
RepMikePence
GarySinise
procopy
gopconference
foxnation
ScottBrownMA
kilmeade
AIPAC
RickSantorum
reasonmag
GovChristie
PolijAM
EricCantor
GovBrewer
biggovt
Jennafnc
TuckerCarlson
IndyEnigma
THEHermanCain
marklevinshow
marcelostrike
TeaPartyNJ
gabriel_oficial
SenJohnMcCain
oreillyfactor
foxnewspolitics
NRANews
VersosSencillos
AmSol
C4Liberty
foxandfriends
hannityshow
brunasald
FredBarnes
GB_TV
gretawire
USCG
GovernorPerry
GOPLeader
FoxNews
newtgingrich
kmita3
marcorubio
RepRonPaul
DLoesch
mises
politico
ConNews
AllenWest
fredthompson
Drudge_Report
Lileks
SpeakerBoehner
AndrewBreitbart
johnboehner
sdoocy
laurawbush
newsbusters
DRUDGE
USArmy
secupp
camposwell
OSumido
johnthune
FoxBusiness
NRSC
AnnCoulter
ChuckGrassley
livedesk
crnc
AFPAA
DanaPerino
BarackObama
netanyahu
AACONS
andylevy
unfollowtt
KarlRove
TexGov
GovMikeHuckabee
sanuzis
Talkmaster
megynkelly
f396
Bret_Baier
krauthammer
RepPaulRyan
JeffFlake
JorgeBernal
glennbeck
PrAbnerFerreira
stephenfhayes
seanhannity
JIDF
edutrevisan
Israel
RasmussenPoll
STU_GBP
JebBush
limbaugh
theMRC
nikkihaley
InfowarsFeed
SarahPalinUSA
rollcall
RonPaul
TedNugent
HaleyBarbour
MittRomney
michellemalkin
thehill
MicheleBachmann
OrrinHatch
Ericatwitts
USCCB
Sarah_Palin
BluegrassPundit
NavyNews
JimDeMint
Heritage
allahpundit
FreedomWorks
USMC
marthamaccallum
NTARC
RealSheriffJoe
DeptofDefense
DennisDMZ
Dr_Rose
RonPaulcom
RumsfeldOffice
grivno
Senate_GOPs
JordanSekulow
Judgenap
George_WBush
BillSchulz
TeamBachmann
SenateFloor
AmSpec
keibosite
fxnopinion
usairforce
scrowder
Michael_Yon
mattklewis
IDFSpokesperson
julianperretta
newadvent
mikeallen
foxheadlines
BraveLad
AM_National
JohnCornyn
israelconsulate
davidicke
CatholicNewsSvc
latimestot
RepublicanGOP
AlinskyDefeater
KOSMOSNET
kimkomando
ErnieFowlke
infowars
FLGovScott
jasoninthehouse
tedcruz
andilinks
MegWhitman
CarrollStandard
NBCConnecticut
lauraingle
Justicereport
HumanEvents
Wizard101
NRCC
ScottWGraves
GroverNorquist
Conservativeind
SarahLPalinNews

Example: modeling topic distribution

- Latent Dirichlet Allocation (LDA)
  - Bayesian soft-clustering technique
  - Each document can naturally belong to multiple clusters with degree-of-membership
  - Popular in topics modeling for text data
- Why should we care?
  - Topics capture user interests
  - Knowledge of interests makes recommendation easier

Mahout/PigML integration



twitter

LDA applications

- User interest modeling facilitates user matching as well as predicting engagement.
- It does not directly solve the more general problem of finding users who might be interested in a tweet, story, web-page, etc
- LDA offers a probabilistic model of user interests (e.g., based on what they tweet about).

Anchoring LDA

- Runs of LDA can lead to significantly different results from one run to the next
- This may hamper the usefulness/interpretability of the results
- Anchoring clusters based on the known tag labels regularizes the clustering procedure
  - Also known as Labeled LDA (LLDA)

twitter

ML/Data Mining we contribute to

- Cassowary
  - Large graph mining in Java (single box)
- Pig
- Scalding
  - cascading with Scala
- PyCascading
  - Cascading with Python
- Mahout

twitter

ML outside of Twitter

- What does the academic community consider important?
  - Sentiment detection
  - Event tracking (e.g., epidemics)
  - Politics
  - Locality detection
  - Spam detection

Publications mentioning …

- Google Scholar (late 2011)

Query	Result count
Twitter spam	9,800
Web spam	52,300
Email spam	41,600
IM spam	30,000
SMS spam	10,500

Quick search

- Number of titles on Amazon (early 2011)

Query	Result count
Twitter marketing	544
Twitter profits	100
Twitter money	236
Twitter rich	36
Twitter seo	21

Spam/spammer modeling

- Why spam Twitter (or other social media)?
- The types of spam seen
  - @replies/@mentions
  - Trend/search spam
  - Follow spam
- Long term vs. short term
  - Spammers favor “pump and dump”
  - Swift reaction to attacks is key

twitter

Example: normal interactions

Example: spammy interactions

ML @ Twitter

- Kumman Chellapila
- Jimmy Lin
- Yue Lu
- Jake Mannix
- Gilad Mishne
- Ram Ravichandran
- Miguel Rios
- Andy Schlaikjer
- Yifan Shi
- +others

THANK YOU

twitter