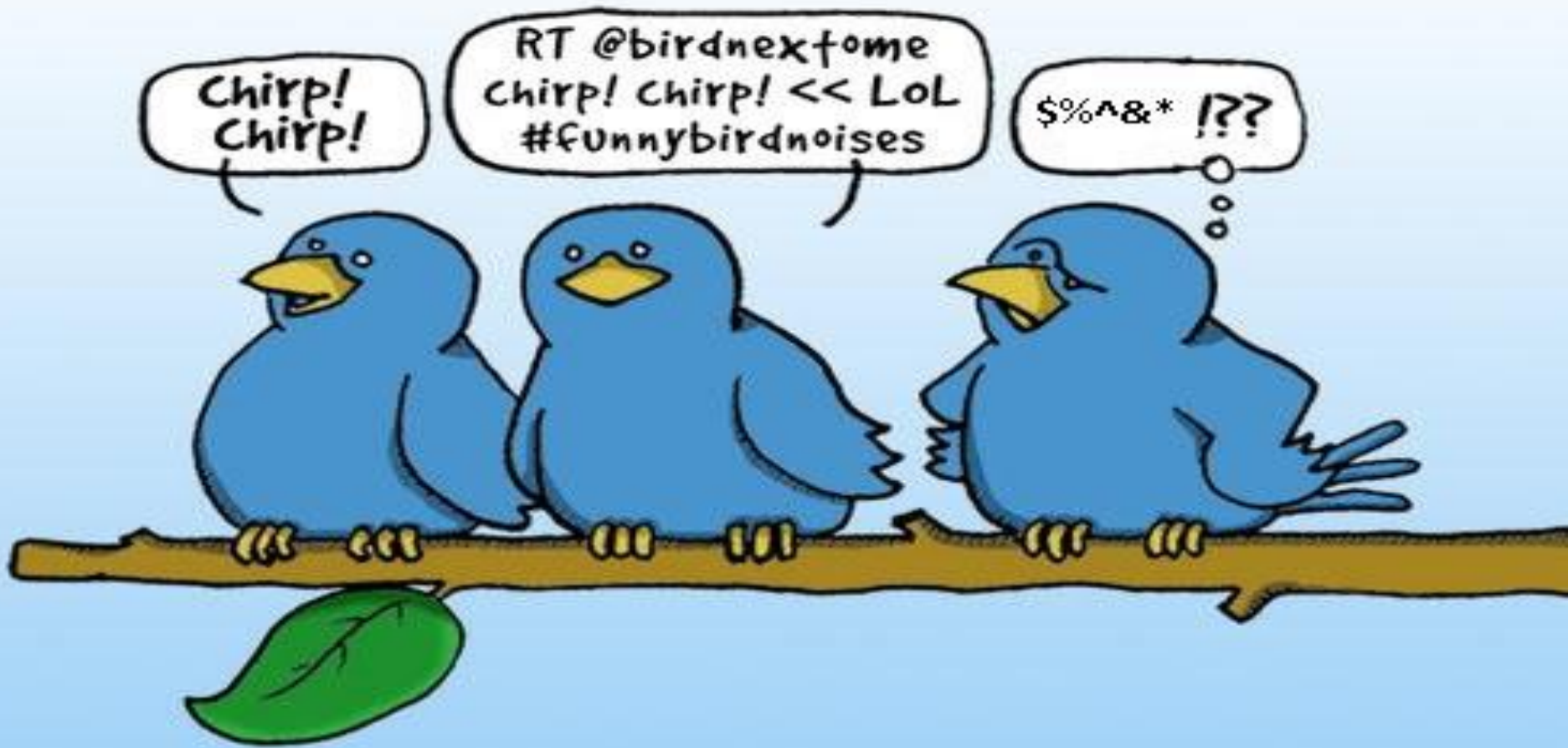
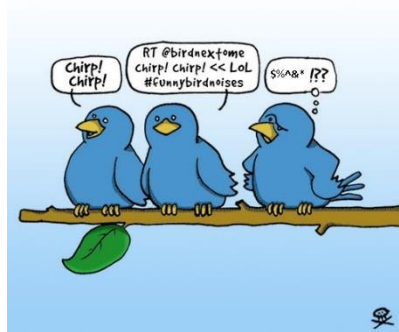


Large-Scale Machine Learning at Twitter

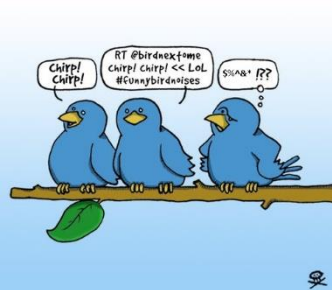




Large-Scale Machine Learning at Twitter

Jimmy Lin and Alek Kolcz
Twitter, Inc.





Large-Scale Machine Learning at Twitter

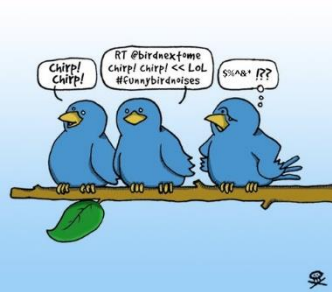
Outline

Outline

- Is twitter big data?
- How can machine learning help twitter?
- Existing challenges?

- Existing literature of large-scale learning
- Overview of machine learning
- Twitter analytic stack
- Extending pig

- Scalable machine learning
- Sentiment analysis application



Large-Scale Machine Learning at Twitter

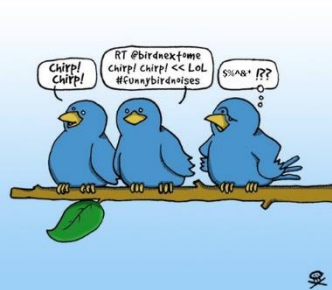
Focus of talk..

What we will not talk about :

- Different “useful” application of twitter
- Why Twitter is a great product and one of its kind

What we will talk about :

- Challenges faced while making it a good product
- Solution approach by “Insiders”



Large-Scale Machine Learning at Twitter

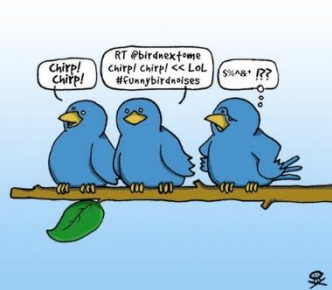
Some twitter bragging ..

The Scale of Twitter

- Twitter has more than 280 million active users
- 500 million Tweets are sent per day
- 50 million people log into Twitter every day
- Over 600 million monthly unique visitors to twitter.com

Large scale infrastructure of information delivery

- Users interact via web-ui, sms, and various apps
- Over 70% of our active users are mobile users
- Real-time redistribution of content
- At Twitter HQ we consume 1,440 hard boiled eggs weekly
- We also drink 585 gallons of coffee per week



Large-Scale Machine Learning at Twitter

Problems in hand ..

Support for user interaction

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

(other) problems we are trying to solve

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

Results for **president obama**

Tweets Top / All / Timeline

- Obama 2012** @Obama2012 11h
"In this country, prosperity does not trickle down. Prosperity grows from the bottom up." — President Obama speaking in Elyria, Ohio today
- Barack Obama News** @ObamaNews 7m
Press Release: **President Obama Signs Hawaii Disaster Declaration** bit.ly/14dN0q
- Barack Obama News** @ObamaNews 4h
Blog Post: **President Obama Talks About Investing in Training American Workers** bit.ly/13JN1j
- Donna Brazile** @donnabrazile 4h
For the record, I support **President Obama's** re-election efforts. But, I am not a surrogate for the campaign or the spokesperson for the DNC.
- Barack Obama** @BarackObama 5h
President Obama met with some Ohioans who are benefiting from community college job training programs today. OFA.BO/Wmcw83

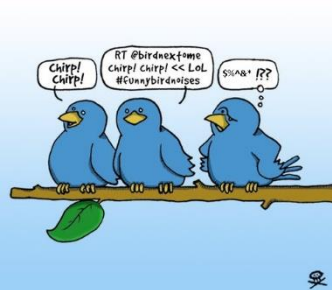
Who to follow
Twitter accounts suggested for you based on who you follow and more.

Search using a person's full name or @username

- USGS** @USGS
Earth science knowledge is just a tweet away. Tweets do not = endorsement: <http://on.doi.gov/pguuoY>
Followed by USDA Food Safety, The Economist and Emergency_in_SF
- Hilary Mason** @hlmason
chief scientist @bitly. Machine learning; I ♥ data and cheeseburgers.
Followed by Gregory Platetsky, Ian Soboroff and Eugene Ajichten
- Adam Rugel** @AdamTrazler, Reston, Syracuse University, Sandwich
- Google Research** @googlerearch
At Google, research is performed company wide, not just in isolated labs. We produce and leverage research to build systems that are used in the real world.
Followed by Tao Tao, Kurt Smith and SIGKDD/TKDD News

Stories

- Twitter Empowers Engineers With New Patent Agreement**
Twitter, in what it says is an act of good will to its engineers and designers, announced a new patent agreement that gives control back to inventors in... bit.ly/drogs.nytimes.com/2012/04/17/new...
Tweeted by Matt Cutts
- Report: Sony's Image URL Accidentally Reveals God of War IV**
God of War IV is coming. It's obvious at this point that Sony will be unveiling God of War IV. The next entry in this franchise has had a slew of rumors and... technobuffalo.com/gaming/platform...
Trending Tweets about God of War
- Striking New Photos Of Great 1906 Earthquake Emerge**
On the anniversary of the Great San Francisco Earthquake of 1906, the San Francisco Municipal Transportation Agency has released a stunning new set... sfist.com/2012/04/18/new...
Tweeted by people who share your interests



Large-Scale Machine Learning at Twitter

To put learning formally ..

Supervised classification in a nutshell

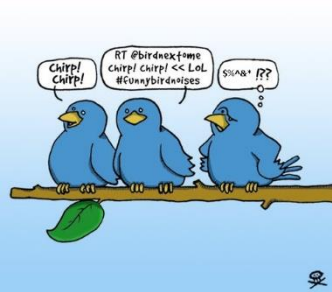
Given $D = \left\{ \left(\mathbf{x}_i, y_i \right) \right\}_i^n$ label
(sparse) feature vector
Induce $f : X \rightarrow Y$ s.t. loss is minimized
empirical loss = $\frac{1}{n} \sum_{i=0}^n \ell(f(\mathbf{x}_i), y_i)$ loss function

Consider functions of a parametric form:

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

model parameters

Key insight: machine learning as an optimization problem!
(closed form solutions generally not possible)



Large-Scale Machine Learning at Twitter

Literature..

Literature

- Traditionally, the machine learning community has assumed sequential algorithms on data fit in memory (which is no longer realistic)
- Few publication on machine learning work-flow and tool integration with data management platform

Google – adversarial advertisement detection

Predictive analytic into traditional RDBMSes

Facebook – business intelligence tasks

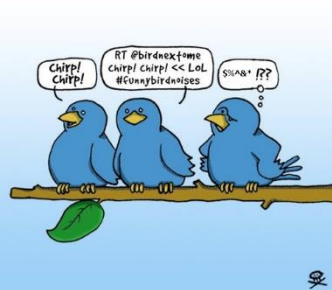
LinkedIn – Hadoop based offline data processing

But they are not for machine learning specifically.

Spark

ScalOps

But they result in end-to-end pipeline.



Large-Scale Machine Learning at Twitter

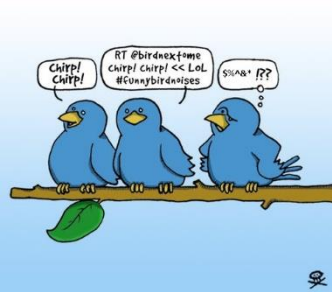
What is author's contribution ..

Contribution

- Provided an overview of Twitter's analytic stack
- Describe pig extension that allow seamless integration of machine learning capability into production platform
- Identify stochastic gradient descent and ensemble methods as being particularly amenable to large-scale machine learning

Note that,

No fundamental contributions to machine learning

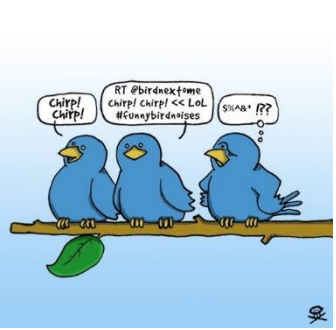


Large-Scale Machine Learning at Twitter

Scalable Machine Learning

Scalable Machine learning

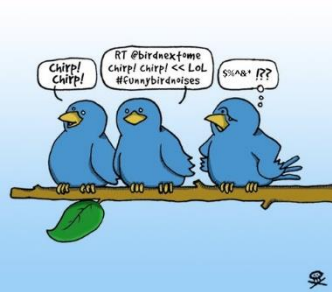
- Techniques for large-scale machine learning
- Stochastic gradient descent
- Ensemble method



Large-Scale Machine Learning at Twitter

Gradient Descent..





Large-Scale Machine Learning at Twitter

Gradient Descent..

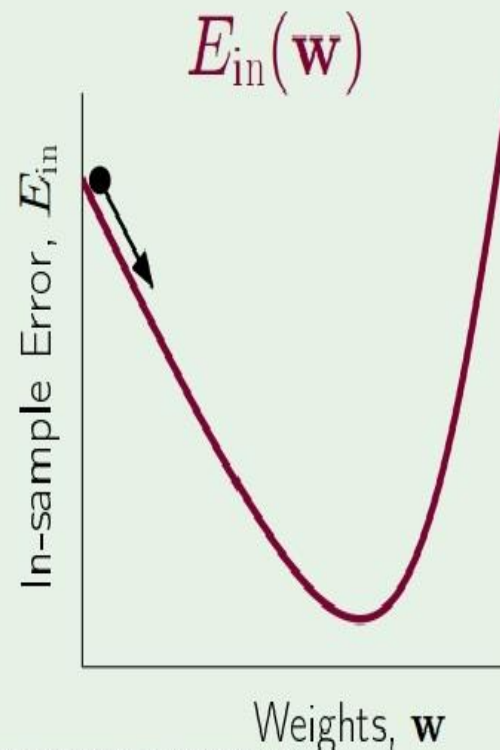
General method for nonlinear optimization

Start at $\mathbf{w}(0)$; take a step along steepest slope

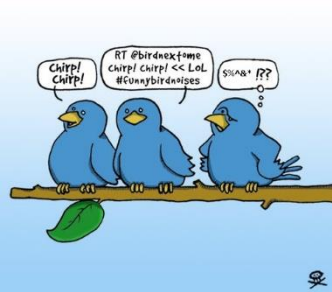
Fixed step size:
$$\mathbf{w}(1) = \mathbf{w}(0) + \eta \hat{\mathbf{v}}$$

Next Weight = Current Weight + move
Move = Step Size * Unit Vecor

What is the direction $\hat{\mathbf{v}}$?



Creator : Yaser Abu Mostafa: Cal tech



Large-Scale Machine Learning at Twitter

Gradient Descent..

Formula for the direction $\hat{\mathbf{v}}$

$$\Delta E_{\text{in}} = E_{\text{in}}(\mathbf{w}(0) + \eta \hat{\mathbf{v}}) - E_{\text{in}}(\mathbf{w}(0))$$

$$= \eta \nabla E_{\text{in}}(\mathbf{w}(0))^T \hat{\mathbf{v}} + O(\eta^2)$$

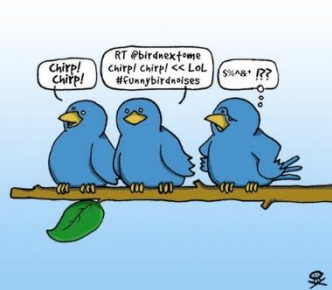
Using Taylor's series expansion

Because the surface
non linear

$$\geq -\eta \|\nabla E_{\text{in}}(\mathbf{w}(0))\|$$

Since $\hat{\mathbf{v}}$ is a unit vector,

$$\hat{\mathbf{v}} = - \frac{\nabla E_{\text{in}}(\mathbf{w}(0))}{\|\nabla E_{\text{in}}(\mathbf{w}(0))\|} \rightarrow \text{Descent along gradient of error..}$$



Large-Scale Machine Learning at Twitter

Stochastic Gradient Descent (SGD)

sto·chas·tic
stə'kastik/
adjective

1. randomly determined; having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely.

Stochastic gradient descent

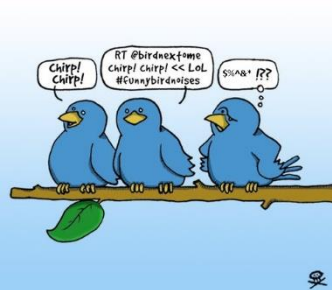
GD minimizes:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \underbrace{e(h(\mathbf{x}_n), y_n)}_{\ln(1+e^{-y_n \mathbf{w}^T \mathbf{x}_n})} \leftarrow \text{in logistic regression}$$

by iterative steps along $-\nabla E_{\text{in}}$:

$$\Delta \mathbf{w} = -\eta \nabla E_{\text{in}}(\mathbf{w})$$

∇E_{in} is based on all examples (\mathbf{x}_n, y_n)



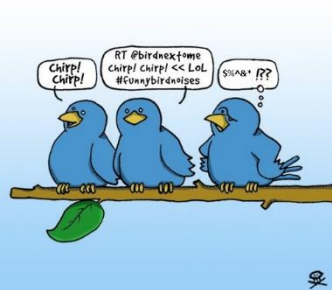
Stochastic gradient descent

Gradient Descent

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

Compute the gradient in the loss function by optimizing value in dataset. This method will do the iteration for all the data in order to one a gradient value.

Inefficient and everything in the dataset must be considered.



Large-Scale Machine Learning at Twitter

Stochastic Gradient Descent (SGD)

Stochastic gradient descent

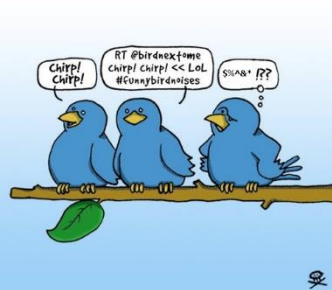
Approximating gradient depends on the value of gradient for one instance.

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

Solve the iteration problem and it does not need to go over the whole dataset again and again.

Stream the dataset through a single reduce even with limited memory resource.

But when a huge dataset stream goes through a single node in cluster, it will cause network congestion problem.



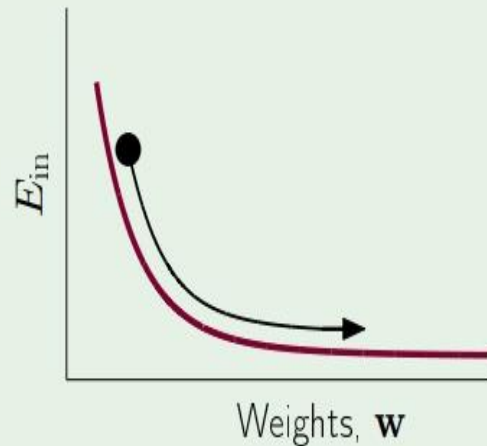
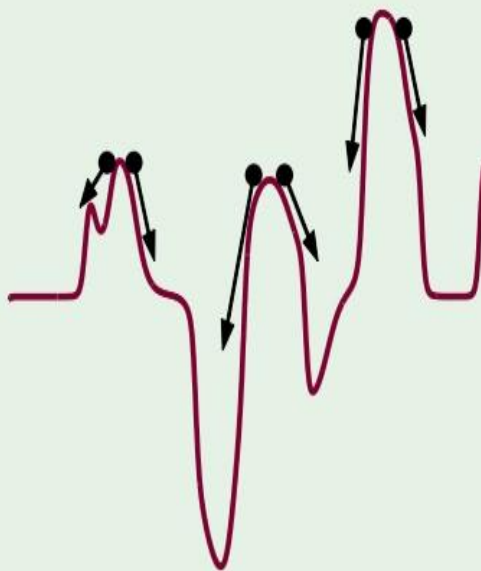
Large-Scale Machine Learning at Twitter

Stochastic Gradient Descent (SGD)

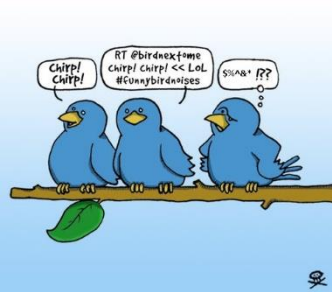
Benefits of SGD

1. cheaper computation
2. randomization
3. simple

Rule of thumb:



randomization helps

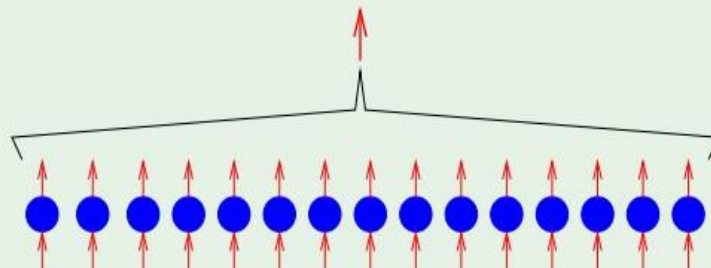


Large-Scale Machine Learning at Twitter

Aggregation a.k.a Ensemble Learning

What is aggregation?

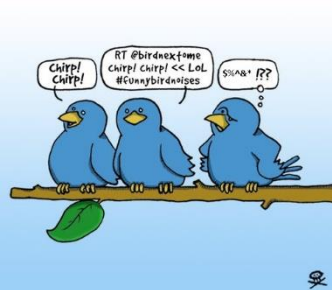
Combining different solutions h_1, h_2, \dots, h_T that were trained on \mathcal{D} :



Regression: take an average

Classification: take a vote

a.k.a. *ensemble learning* and *boosting*

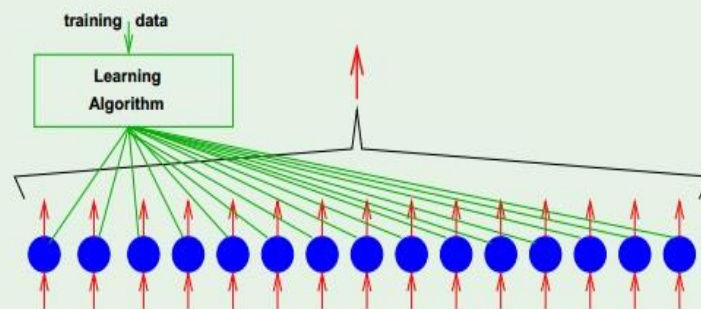


Large-Scale Machine Learning at Twitter

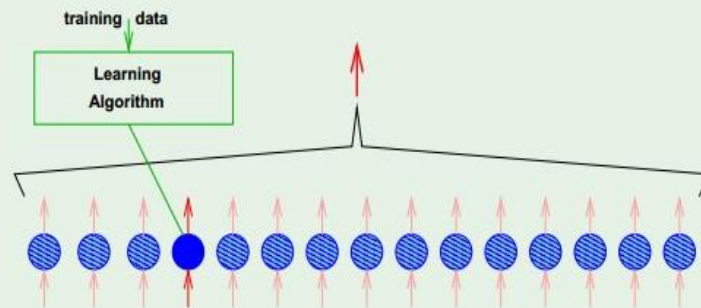
Aggregation a.k.a Ensemble Learning

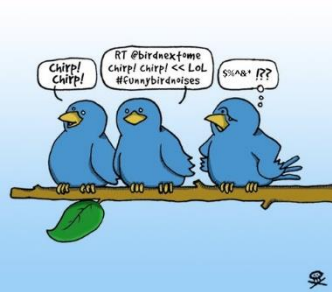
Different from 2-layer learning

In a 2-layer model, all units learn **jointly**:



In aggregation, they learn **independently** then get combined:





Large-Scale Machine Learning at Twitter

Ensemble Learning..

Ensemble Methods

Classifier ensembles: high performance learner

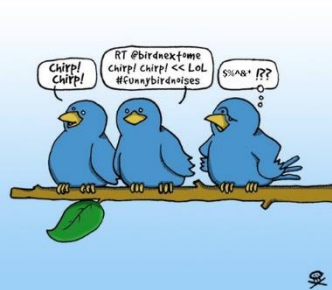
Performance: very well

Some rely mostly on randomization

-Each learner is trained over a subset of features and/or instances of the data

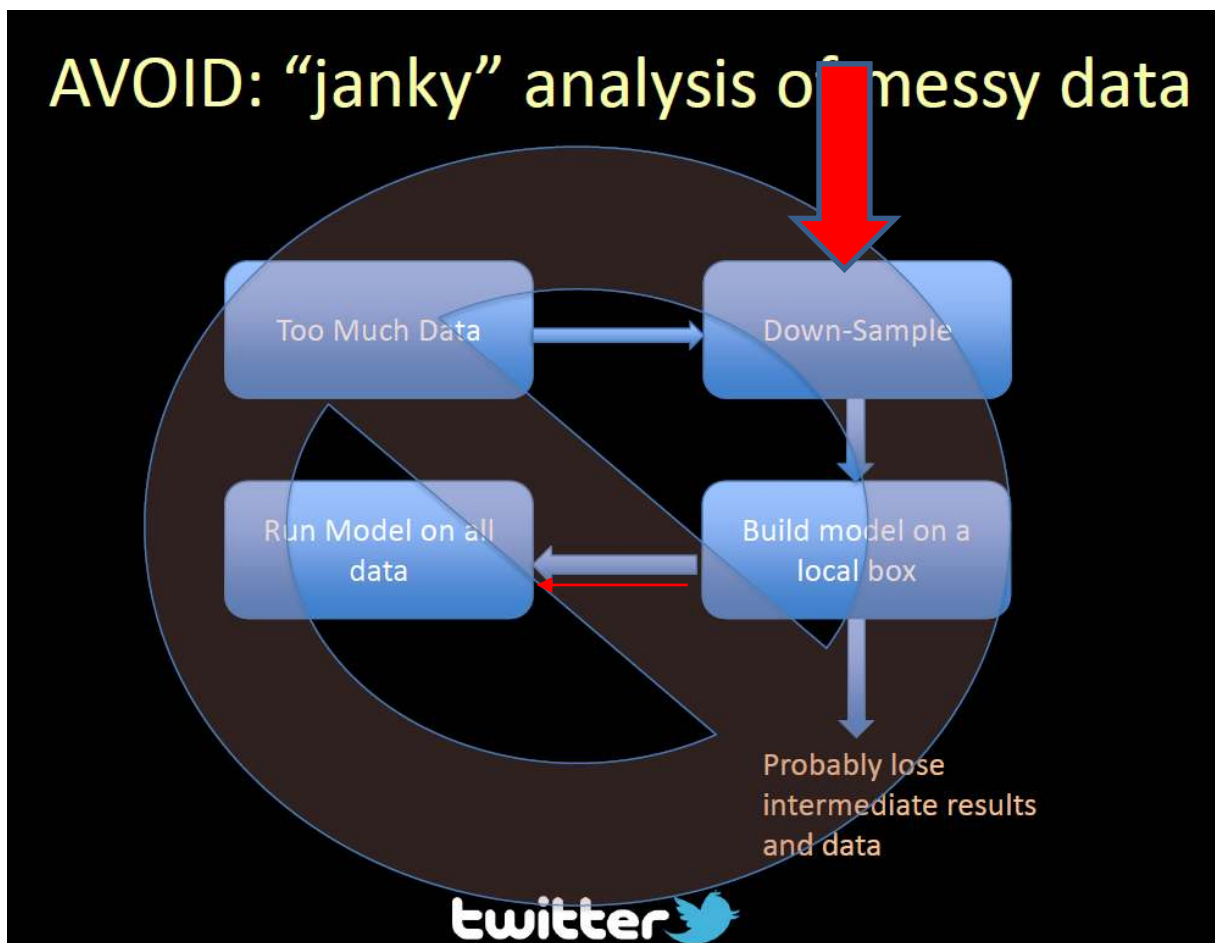
Ensembles of linear classifiers

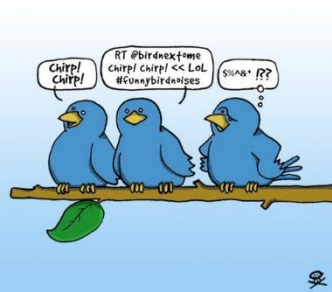
Ensembles of decision trees (random forest)



Large-Scale Machine Learning at Twitter

At Twitter ...





Large-Scale Machine Learning at Twitter

Hoeffding's Inequality

In a big sample (large N), ν is probably close to μ (within ϵ).

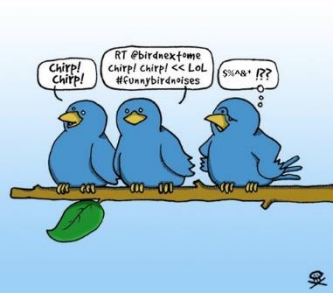
Formally,

$$\mathbb{P} [|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Sample frequency ν is likely close to bin frequency μ .

This is called **Hoeffding's Inequality**. Slide taken from Caltech's Learning from Data Course : Dr Yaser Abu Mostafa

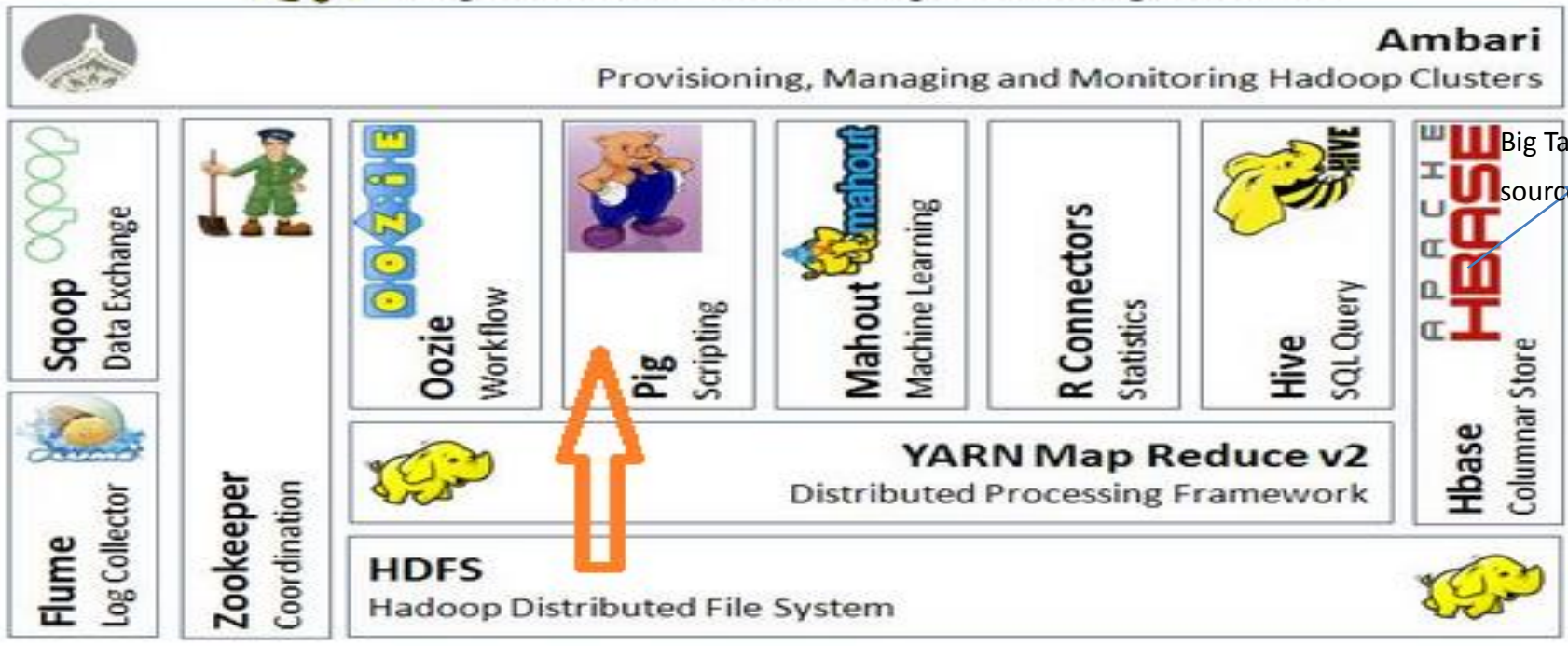
Large-Scale Machine Learning at Twitter



Hadoop Ecosystem

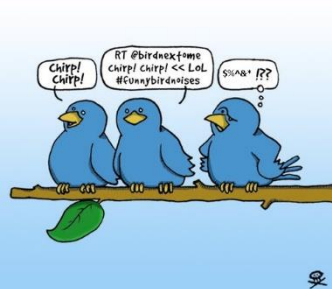


Apache Hadoop Ecosystem



Big Table open source version

Image Source: Apache Yarn Release



Large-Scale Machine Learning at Twitter

Hadoop Ecosystem at Twitter..

- Oink:**
- Aggregation query
Standard business intelligence tasks
 - Ad hoc query
One-off business request
Prototypes of new function
Experiment by analytic group

Database

Application log

Other sources



Real-time processes

Batch processes

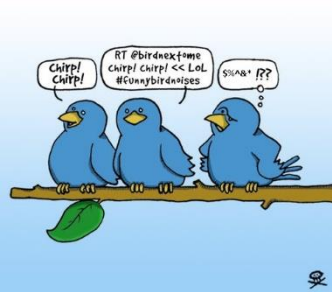


HDFS

Serialization Protocol buffer /Thrift



Hadoop cluster



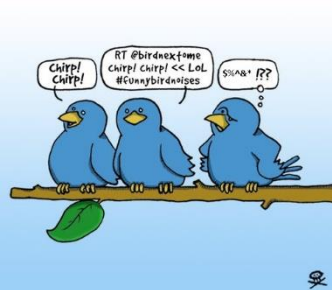
Large-Scale Machine Learning at Twitter

Glorifying PIG

Why use Pig?

- **Suppose you have user data in one file, website data in another, and you need to find the top 5 most visited sites by users aged 18 - 25**





Large-Scale Machine Learning at Twitter

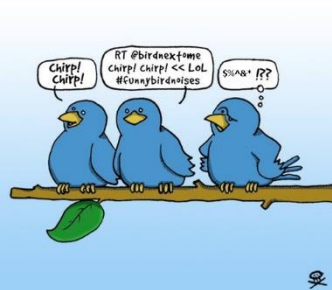
Glorifying PIG

In Pig Latin

```
Users = load 'input/users' using PigStorage(',') as (name:chararray, age:int);
Fltrd = filter Users by age >= 18 and age <= 25;
Pages = load 'input/pages' using PigStorage(',') as (user:chararray,
url:chararray);
Jnd = join Fltrd by name, Pages by user;
Grpd = group Jnd by url;
Smmr = foreach Grpd generate group,COUNT(Jnd) as clicks;
Srted = order Smmr by clicks desc;
Top5 = limit Srted 5;
store Top5 into 'output/top5sites' using PigStorage(',');
```

9 lines of code, 15 minutes to write

170 lines to 9 lines of code



Large-Scale Machine Learning at Twitter

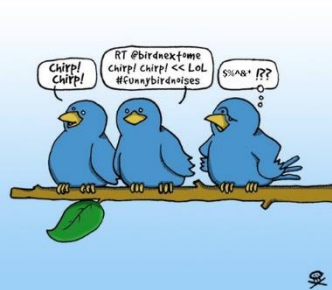
Maximizing the use of Hadoop ..

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

Seamless scaling to large datasets
Integration into production workflows





Large-Scale Machine Learning at Twitter

What authors contributed technically ..

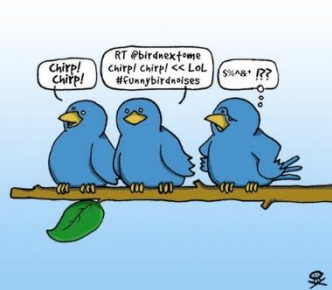
Core libraries:

Core Java library

Basic abstractions similar to existing packages (weka, mallet, mahout)

Lightweight wrapper

Expose functionalities in Pig



Large-Scale Machine Learning at Twitter

PIG Functions..

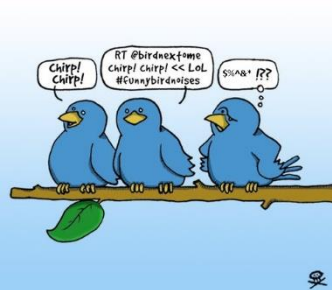
Training models:

```
training = load 'training.txt'  
            using SVMLightStorage()  
            as (target: int, features: map[]);  
store training into 'model/'  
            using FeaturesLRClassifierBuilder();
```



Storage function

```
-1 1:0.43 3:0.12 9284:0.2 # abcdef
```



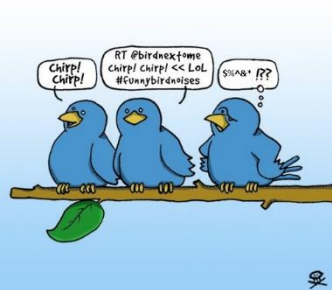
Large-Scale Machine Learning at Twitter

PIG Functions..

Shuffling data:

```
training = foreach training generate  
    label, features, RANDOM() as random;  
training = order training by random parallel 1;
```

```
data = foreach data generate target, features,  
    RANDOM() as random;  
split data into training if random <= 0.9,  
    test if random > 0.9;
```



Large-Scale Machine Learning at Twitter

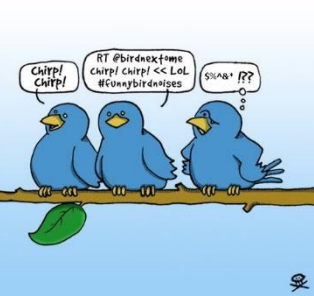
PIG Functions..

Using models:

```
define Classify ClassifyWithLR('model/');
data = load 'test.txt' using SVMLightStorage()
      as (target: double, features: map[]);
data = foreach data generate target,
      Classify(features) as prediction;
```


Large-Scale Machine Learning at Twitter

HortonWorks Way..



My Scripts Query history

You have gone full screen. [Exit full screen \(F11\)](#)

My scripts

[NEW SCRIPT](#)

pigScript

Settings

Email notification

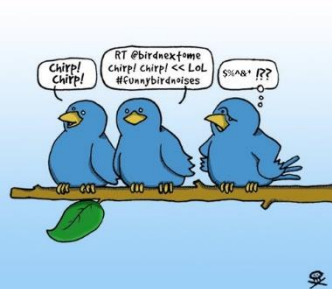
[USER-DEFINED FUNCTIONS](#)

[Upload UDF Jar](#)

Title: pigScript

Pig script. [PIG helper](#)

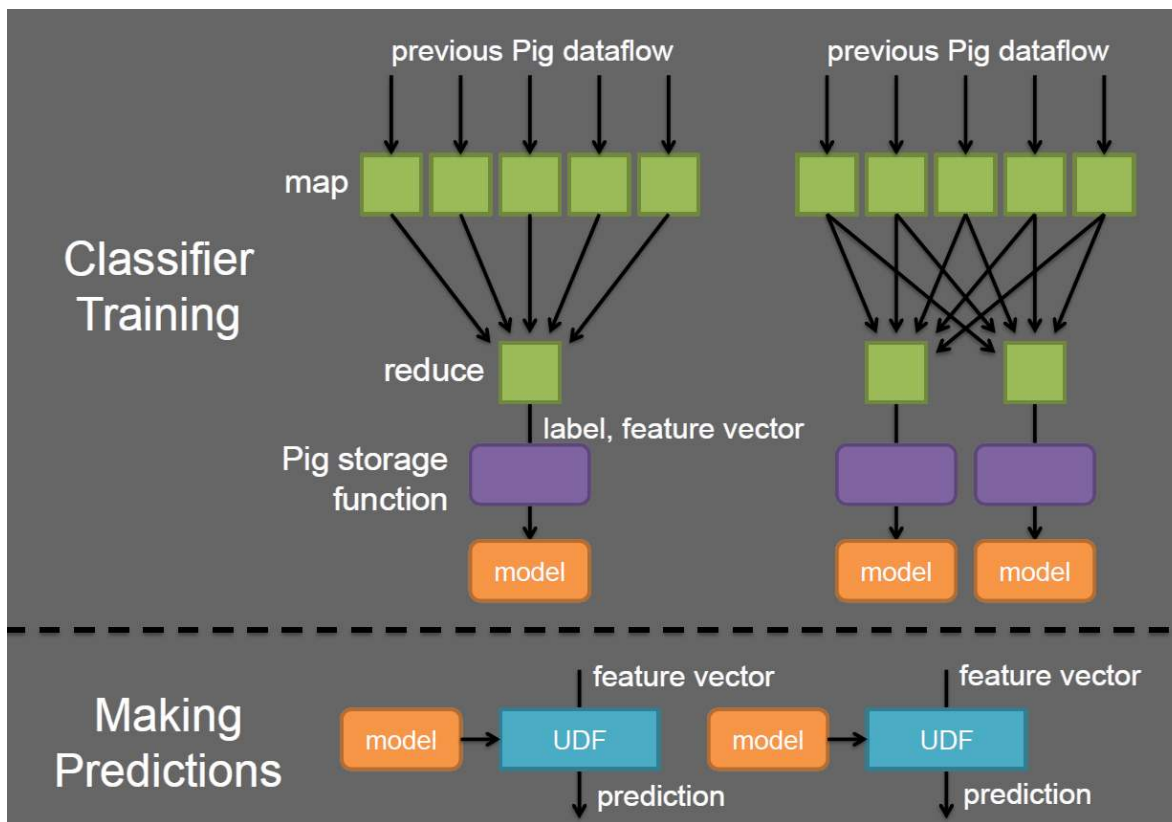
```
1 batting = load 'Batting.csv' using PigStorage(',');
2 runs = FOREACH batting GENERATE $0 as playerID, $1 as year, $8 as runs;
3 grp_data = GROUP runs by (year);
4 max_runs = FOREACH grp_data GENERATE group as grp,MAX(runs.runs) as max_runs;
5
6 join_max_run = JOIN max_runs by ($0, max_runs), runs by (year,runs);
7 join_data = FOREACH join_max_run GENERATE $0 as year, $2 as playerID, $1 as runs;
8 dump join_data;
9
```

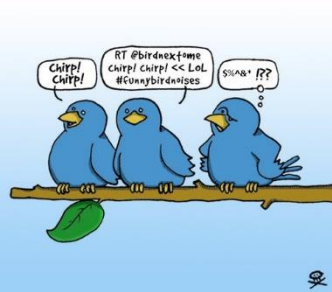


Large-Scale Machine Learning at Twitter

Final Model which works!!!

Final Learning - Ensemble Methods





Large-Scale Machine Learning at Twitter

Use case..

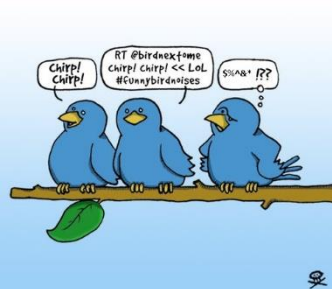
Example: Sentiment Analysis

Emotion Trick 😊 ☹️

Test dataset: 1 million English tweets, minimum 20 letters-long

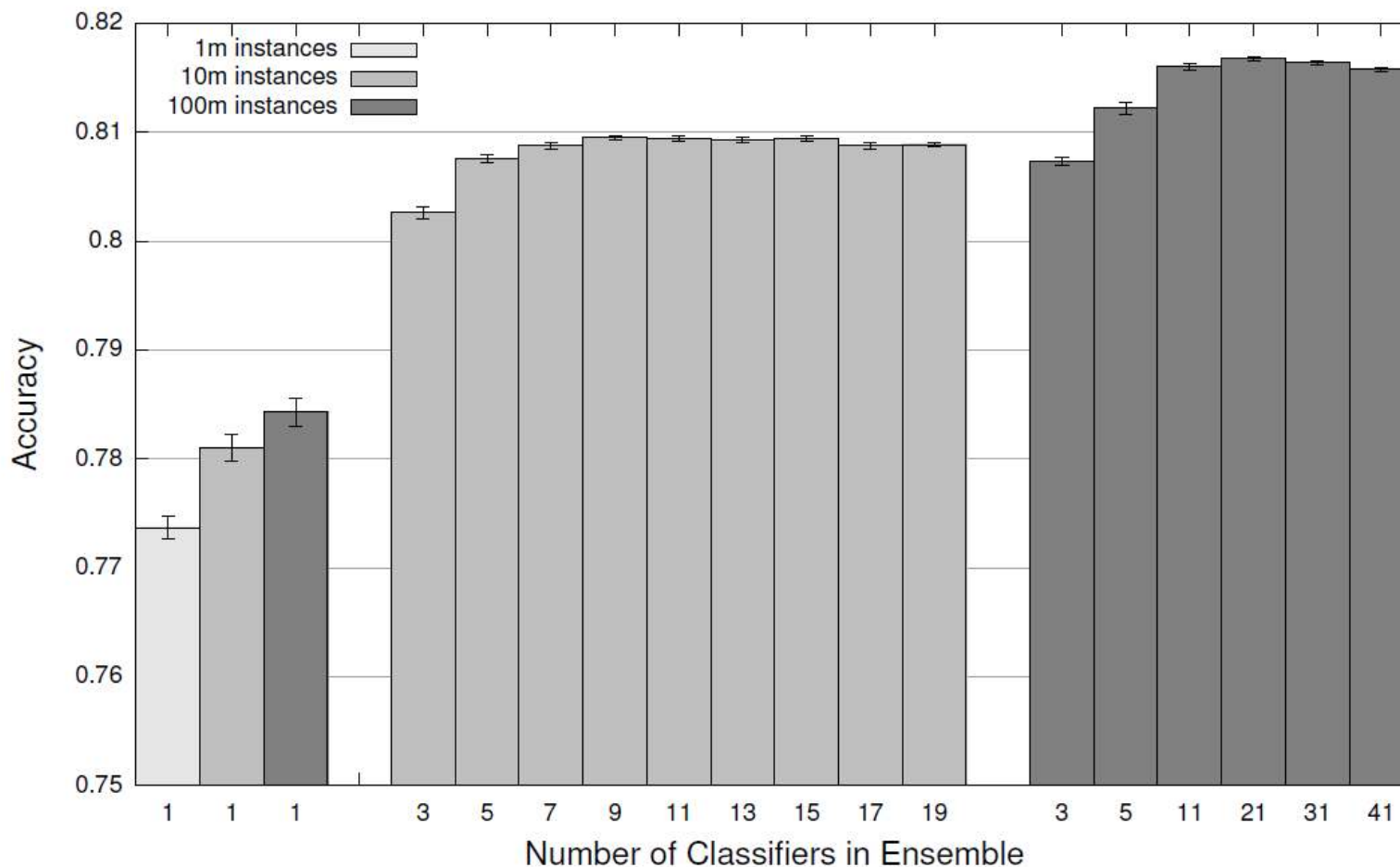
Training data: 1 million, 10 million and 100 million English training examples

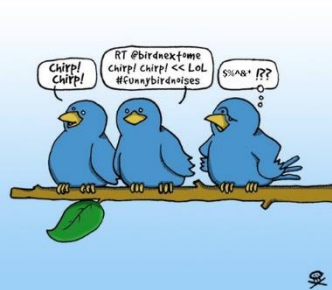
Preparation: training and test sets contains equal number of positive and negative examples, removed all emoticons.



Large-Scale Machine Learning at Twitter

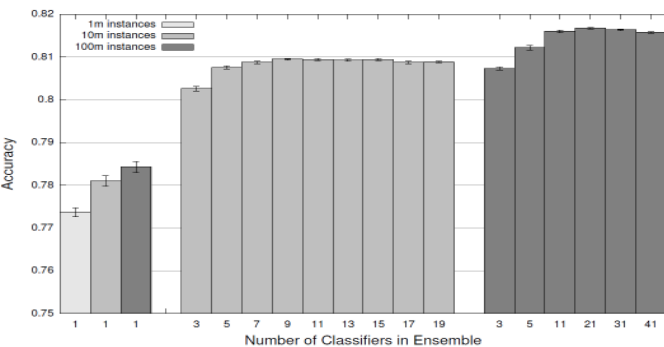
Finally a graph ..



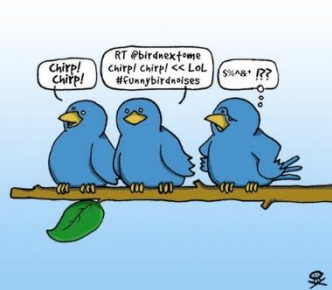


Large-Scale Machine Learning at Twitter

Explaining a bit more of graph ..



1. The error bar denotes 95% confidence interval
2. The leftmost group of bars show accuracy when training a single logistic regression classifier on {1, 10, 100} million training examples.
3. 1-10 Change Sharp , 10 – 100 million : Not that sharp
4. The middle and right group of bars in Figure 2 show the results of learning ensembles
5. Ensembles lead to higher accuracy—and note that an ensemble trained with 10 million examples outperforms a single classifier trained on 100 million examples
6. No accurate running time reported as experiments were run on production clusters – but informal observations are in sync with what the logical mind suggests (ensemble takes shorter to train because models are learned in parallel)
7. In terms of applying the learned models, running time increases with the size of the ensembles—since an ensemble of n classifiers requires making n separate predictions.

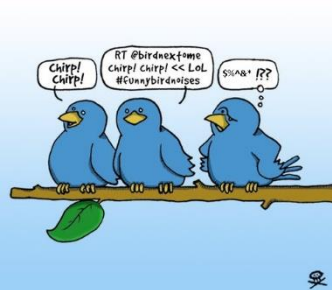


Large-Scale Machine Learning at Twitter

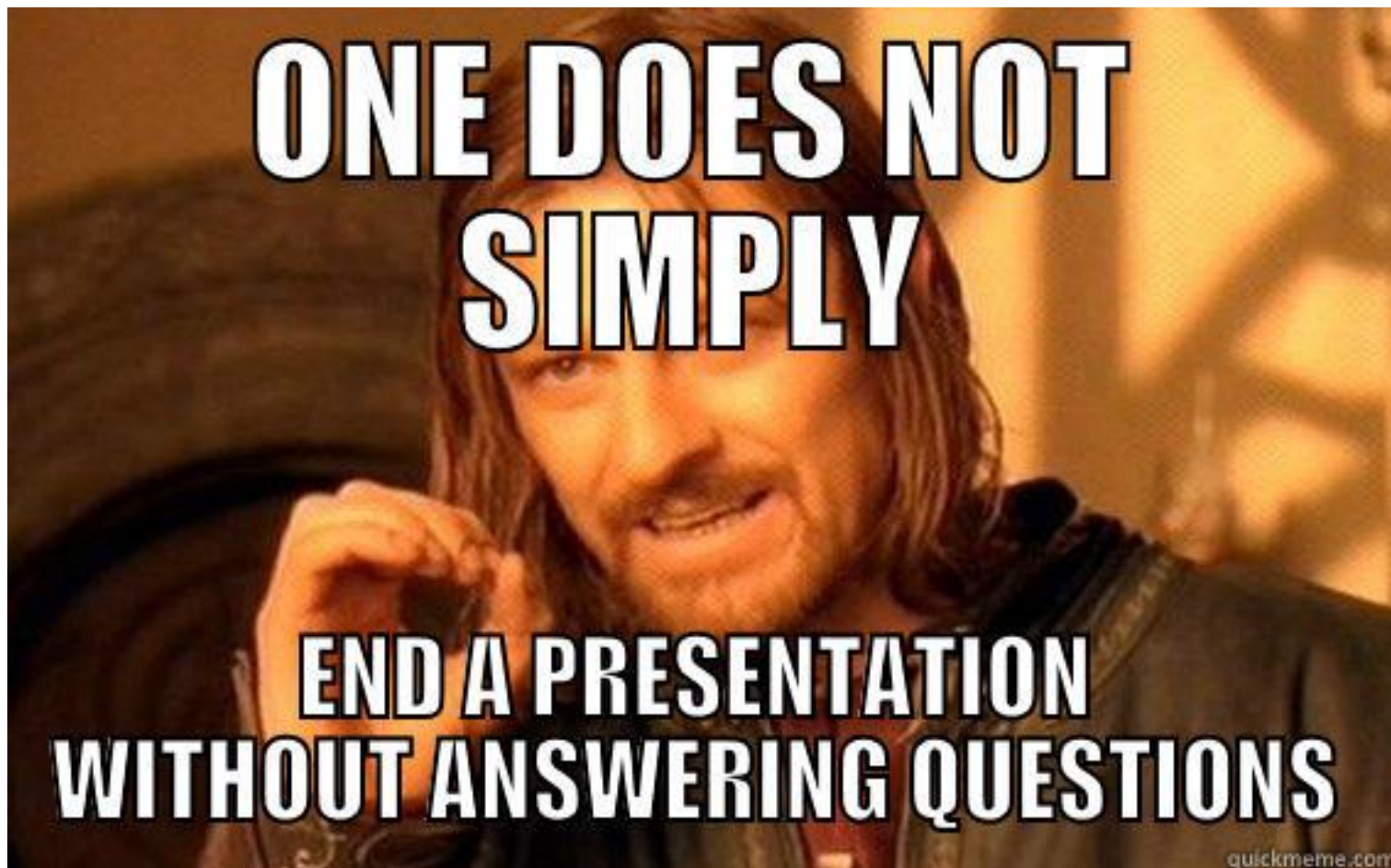
Conclusion

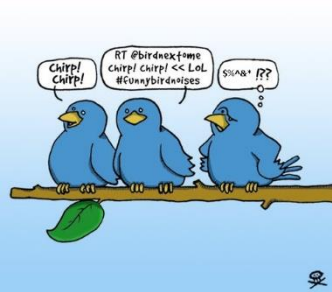
What I loved about paper : I understood it 😊 ?

“our goal has never been to make fundamental contributions to machine learning, we have taken the pragmatic approach of using off-the shelf toolkits where possible. Thus, the challenge becomes how to incorporate third-party software packages along with in-house tools into an existing workflow”..



Large-Scale Machine Learning at Twitter





Large-Scale Machine Learning at Twitter

