

Scalable Machine Learning at Yahoo

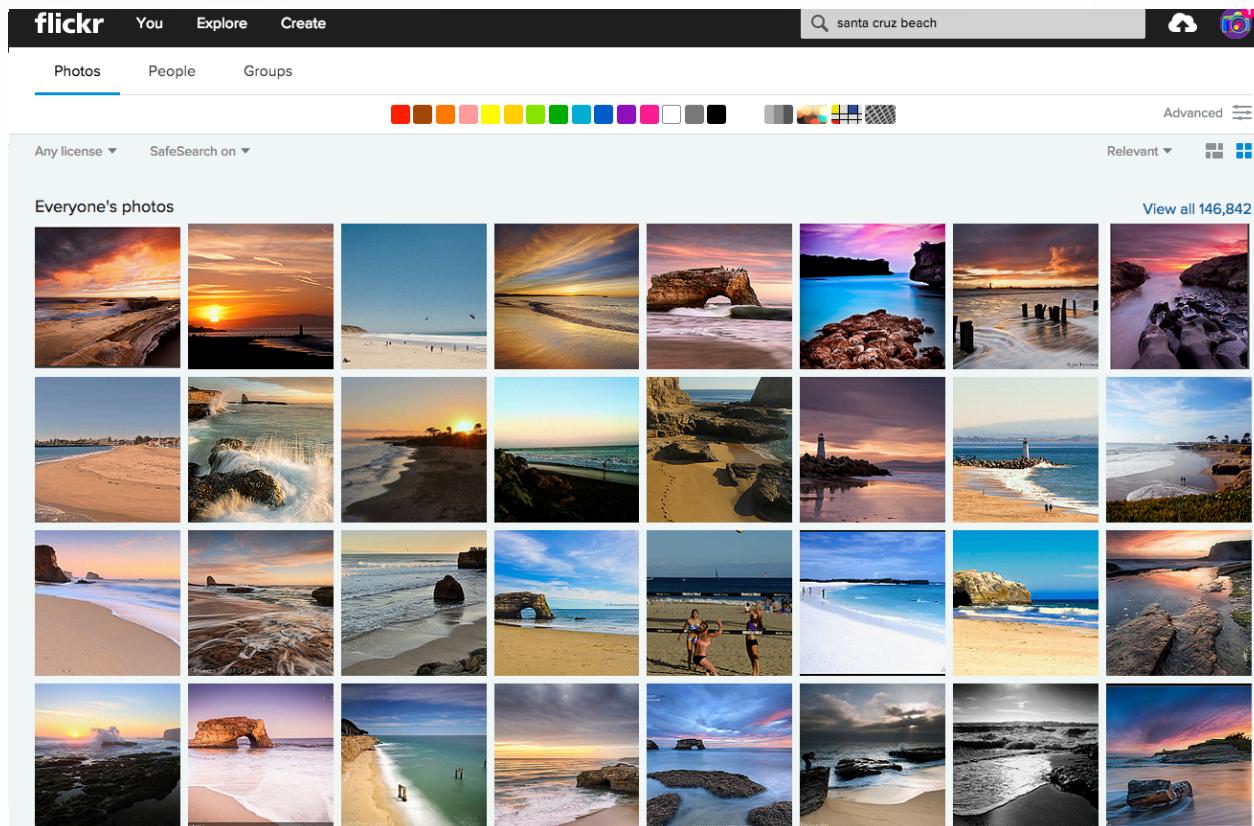
Andy Feng

VP Architecture, Yahoo

Agenda

- Use Cases
- Challenge
- Our Approach
- Algorithms
 - 1. Logistic Regression
 - 2. Word Embedding
 - 3. Deep Learning

ML User Case: Flickr



- 1TB free storage
- 10 billion photos
 - › 7.5 million per day
 - › No tags for most photos
- Search: private/public
 - › Computed tag
- Scene detection
 - › Beach, Sunset, ...
- Object recognition
 - › Lighthouse, Sea arch, ...

YAHOO!

ML Use Case: Yahoo Weather



■ Beauty

- › Computational assessed

■ Relevant scene

- › Location
- › Time
- › Mostly cloudy
- › Partly cloudy
- › Shower
- › ...

YAHOO!

ML Use Case: Search

YAHOO!

Web Images Video News Local Answers Shopping More

Anytime Past day Past week Past month

Ad related to 2015 mercedes technology package

Mercedes-Benz® GLA250 - See Offer Details.
www.MBUSA.com/GLA/250
Lease the 2015 GLA250 4MATIC® for \$369 Per Month.

Bay Area Dealerships **Build Your Own**
[Compare Vehicles](#) [Coupes](#)
[Current Special Offers](#) [Sedans](#)

2015 Mercedes S-class coupe wraps high-tech, high luxury in ...
www.autonews.com/article/20141017/OEM04/141019879/2015... Cached
The 2015 Mercedes-Benz S-class coupe embodies the automaker's new design philosophy of modern luxury, says Gorden Wagener, the brand's design chief.

2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A ...
[cars.findthebest.com](http://cars.findthebest.com/Motors/2015-Acura-RDX-Technology-Package-4dr-SUV-AWD-3-5L-6cyl-6A) > Motors
Compare Cars: 2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A) vs Mercedes-Benz GLA-Class GLA45 AMG 4MATIC 4dr SUV AWD (2.0L 4cyl Turbo 7AM). In this side ...

2015 Mercedes Technology Package - Image Results

[More 2015 Mercedes Technology Package images](#)

2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A ...
[cars.findthebest.com](http://cars.findthebest.com/Motors/2015-Acura-RDX-Technology-Package-4dr-SUV-AWD-3-5L-6cyl-6A) > Motors
Compare Cars: 2015 Acura RDX Technology Package 4dr SUV AWD (3.5L 6cyl 6A) vs Mercedes-Benz GLK-Class GLK350 4MATIC 4dr SUV AWD (3.5L 6cyl 7A). See all the ...

Andrew Corp Mail

See more ads for:
[2015 mercedes technology package](#)
[2015 mercedes c class](#)
[disney world vacation packages 2015](#)
[mercedes gla 2015](#)
[mercedes benz gla 2015](#)
[disney vacation packages 2015](#)

Ads

2015 Mercedes-Benz S550
[FleetRates.com](#)
New Rebate Now \$699/mo \$82,275
Low Prices Nationwide Delivery

2015 Mercedes | [Edmunds.com](#)
[www.Edmunds.com](#)
 78 reviews
Research, Reviews, Pricing & More.
Edmunds.com® - Ask the Car People!

Top-Rated SUVs 2015
[educatehow.com/Savings](#)
Explore the Best New SUVs for 2015.
Compare Prices & Specs.

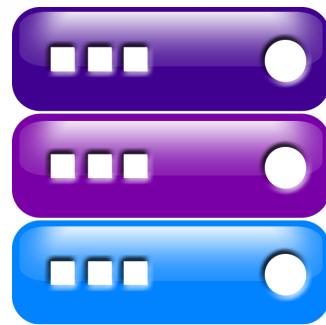
- Intention understanding
- Content ranking
- Query-Ads matching
- Ad click prediction

YAHOO!

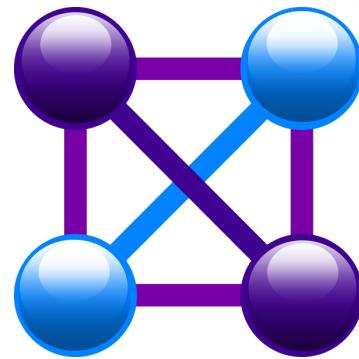
ML Challenge: Scale

1. Massive amount of examples
 - › Naïve solutions take days/weeks
2. Billions of features
 - › Model exceeds memory limits of 1 computer
3. Variety of algorithms
 - › Different solutions required for scale-up

Massive Hadoop at Yahoo



600 PB
HDFS



40K
Computers



MACHINE
LEARNING

Scalable ML in Action

 **RUNNING Applications**

Cluster Metrics																	
	Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	
Applications	1043967	10	194	1043763	2559	18.65 TB	41.46 TB	0 B	32201	231600	0	965	0	3	2	0	

Show 20 entries Search: af

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
application_1431026891795_302886	afeng	ps-matrix-server-afeng	org.apache.slider	gridops	Thu, 14 May 2015 18:29:46 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1044013	afeng	query2ads-afeng	SPARK	gridops	Wed, 27 May 2015 20:57:31 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1043971	afeng	[mx3_hcat_uad_traffic/201504192115]	MAPREDUCE	gridops	Wed, 27 May 2015 20:56:19 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1043972	dfsload	[mx3_hcat_uad_traffic/201504192130]	MAPREDUCE	gridops	Wed, 27 May 2015 20:56:18 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1043969	dfsload	[mx3_hcat_uad_traffic/201504192230]	MAPREDUCE	gridops	Wed, 27 May 2015 20:56:10 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1043967	dfsload	[mx3_hcat_uad_traffic/201504192200]	MAPREDUCE	gridops	Wed, 27 May 2015 20:56:10 GMT		RUNNING	UNDEFINED		ApplicationMaster
application_1431026891795_1043966	dfsload	[mx3_hcat_uad_traffic/201504192215]	MAPREDUCE	gridops	Wed, 27 May 2015 20:56:08 GMT		RUNNING	UNDEFINED		ApplicationMaster

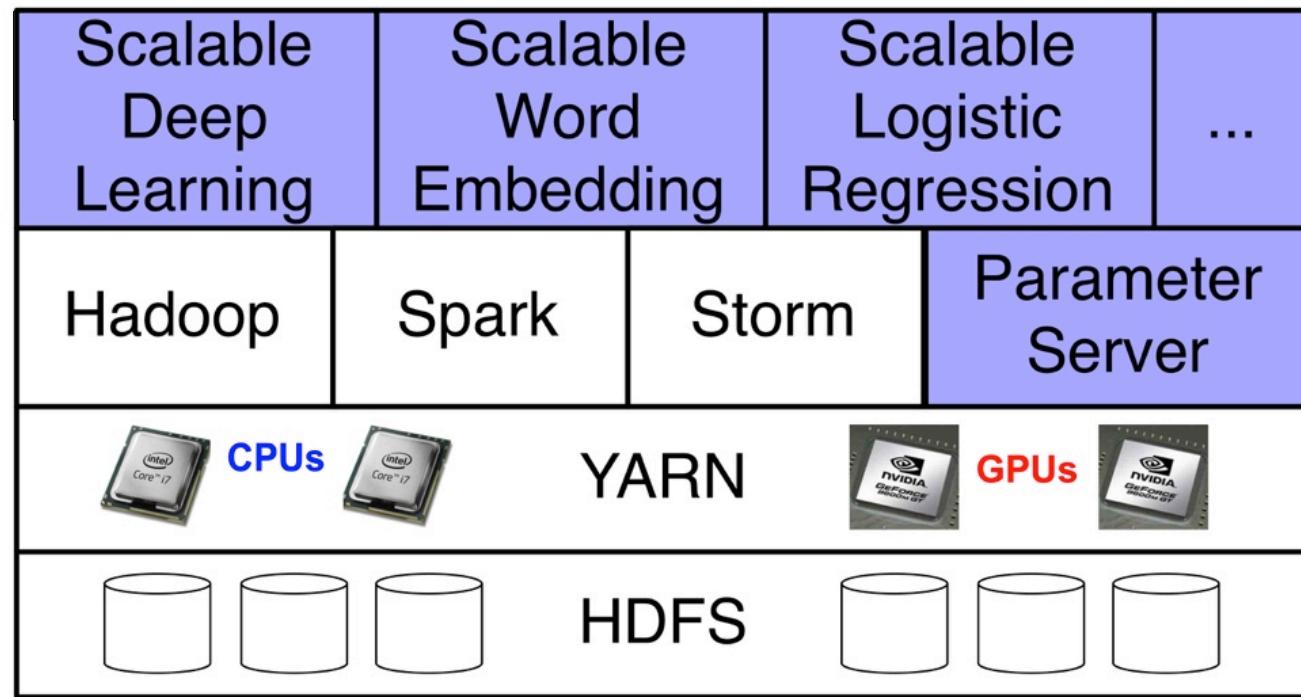
Showing 1 to 7 of 7 entries (filtered from 143 total entries) First Previous 1 Next Last

ML server

ML learner

Map Reduce

Architecture for Scalable ML



Examples of ML Algorithms

1. Logistic Regression

- › Problem: Model size
- › Solution: Spark + Parameter Server

2. Ad-Query Vectors

- › Problem: Model size + Training latency
- › Solution: Spark + Parameter Server

3. Deep Learning

- › Problem: Training latency
- › Solution: Spark + GPU + MPI

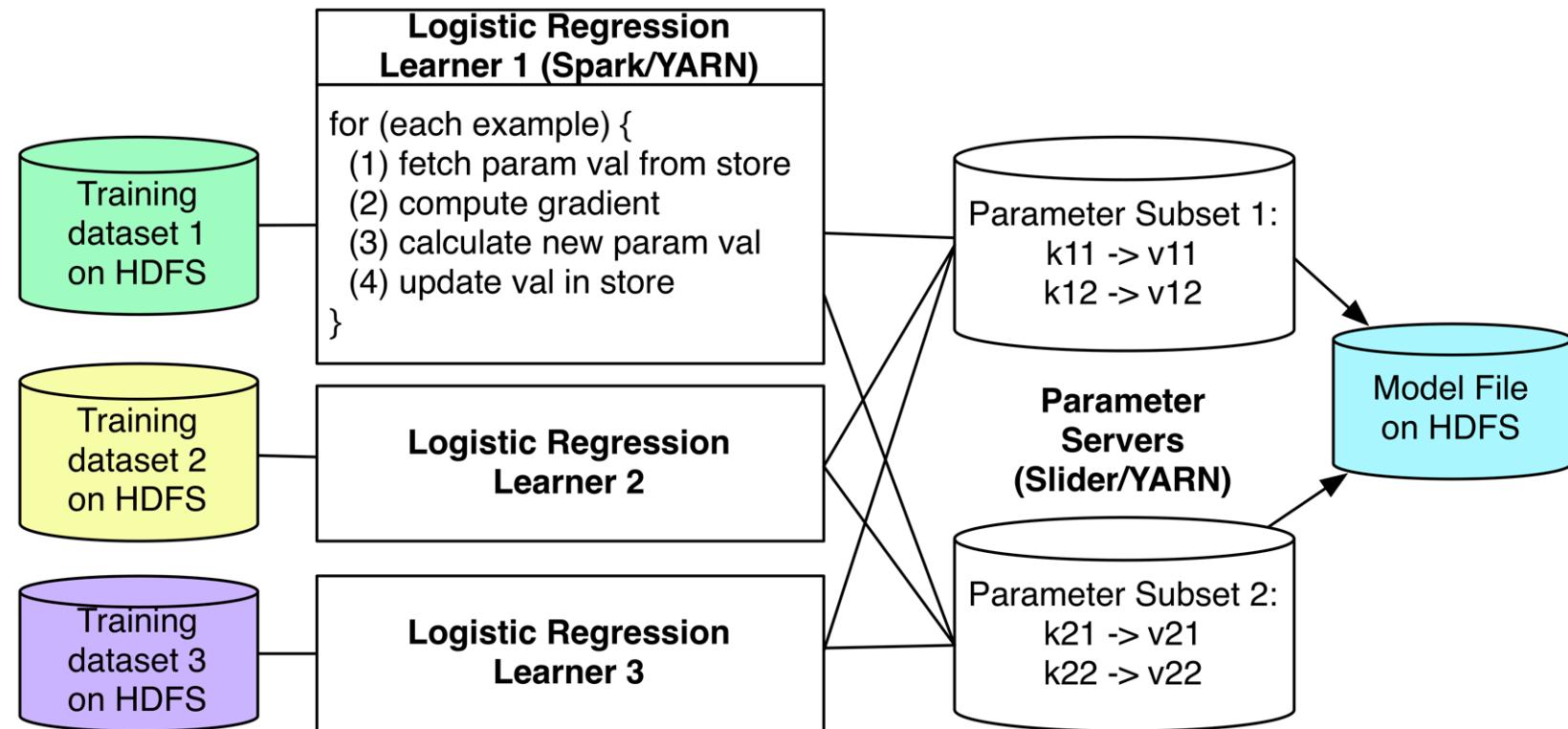
Algorithm 1: Logistic Regression

$$p(X_1, X_2, \dots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}}$$

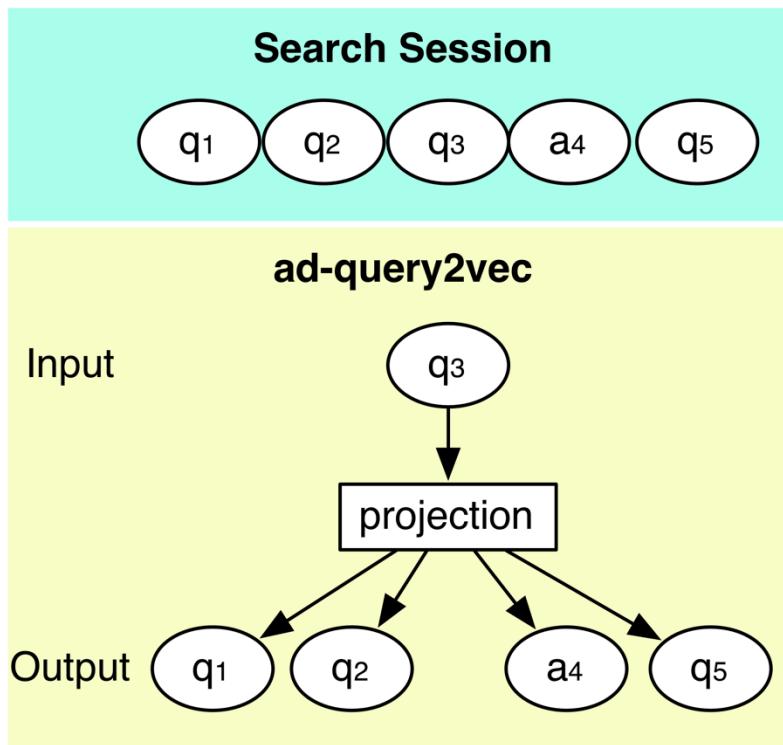
When $|\beta| > 100B$,

- › 100 Billion * 16 Bytes = 1.6 TB
- › β exceeds memory limit of 1 computer

Logistic Regression: 1000x Scale-up



Algorithm 2: Ad-Query Vectors



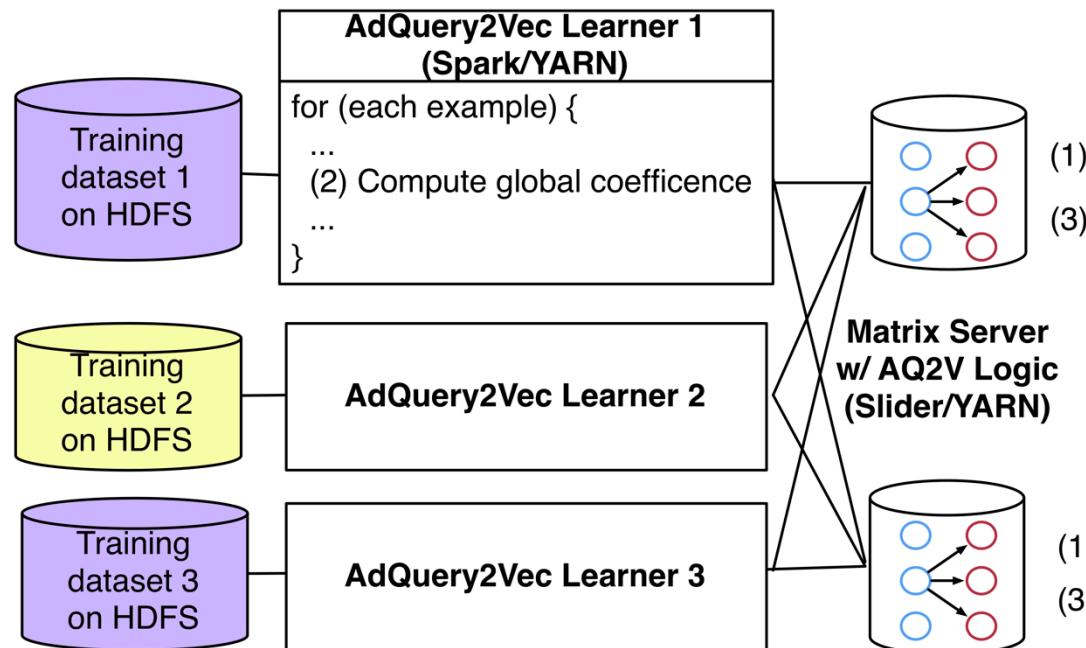
- Vector: numeric representation of queries/ads
 - › $\text{Vector}(\text{"san jose weather"}) \approx \text{Vector}(\text{"weather 95113"}) \approx \text{Vector}(\text{ad123})$
- Model size
 - › $1 \text{ Billion}^* 300 \text{ dimensions} = 2.4\text{TB}$
- Vector computation (X^*Y , $aX+Y$)
 - › Took weeks for small datasets

13

* Yahoo Labs: <http://bit.ly/1G3f6L2>

YAHOO!

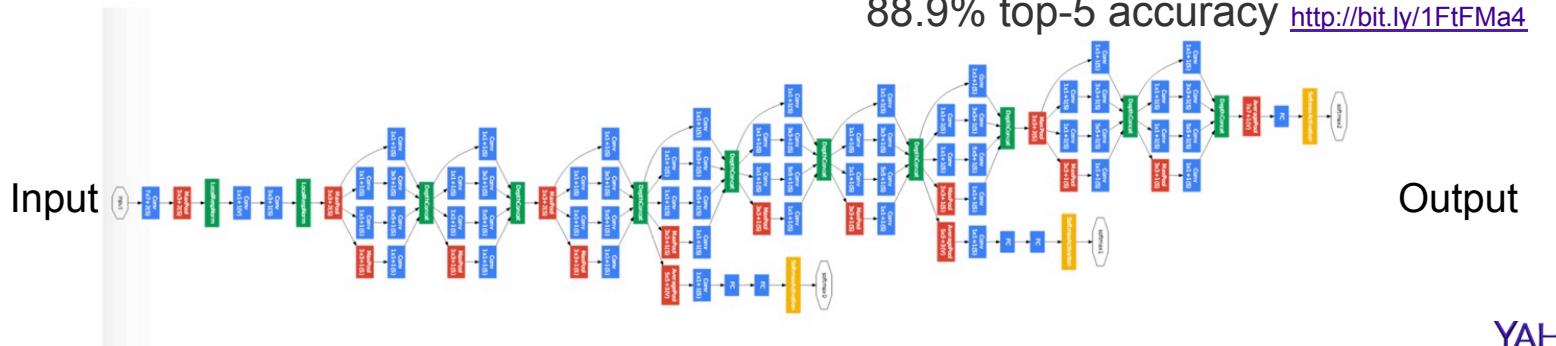
Ad-Query2Vec: 100x Speed/Scale-up



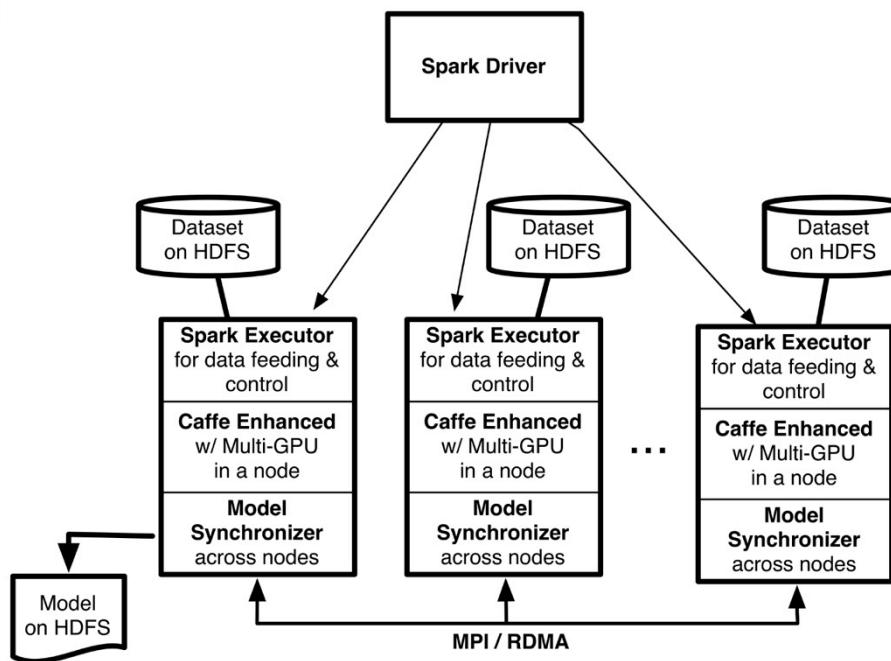
- Computation on servers
 - › (1) Negative sampling
 - › (1) Compute gradient: X^*Y
 - › (3) Adjust vectors: $Y = aX + Y$
- Daily training enabled
 - › weeks → hours

Algorithm 3: Deep Learning

- ILSVRC classification
 - › Classify images to one of 1,000 categories
 - › Dataset
 - 1.2 million images for training
 - 50,000 images for validation
- GoogLeNet: 2014 Winner
 - › 22 layer deep network
 - Convolution, Pooling,..., Loss
 - › Trained on Google DistBelief cluster (16,000 CPU cores)
- Training latency
 - › **47 days** for 1 K40c GPU to achieve 88.9% top-5 accuracy <http://bit.ly/1FtFMa4>



Deep Learning: Caffe-on-Spark with GPU

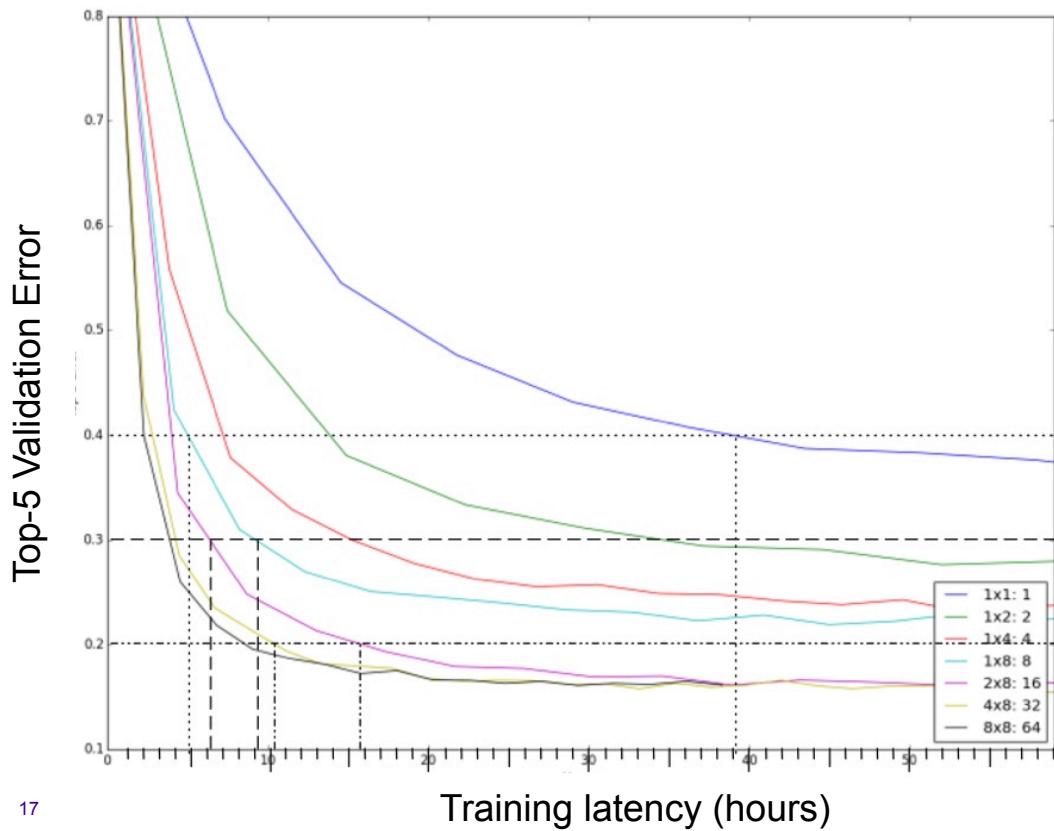


- GPU + Infiniband
 - › 10x computing speedup over CPU
 - › 10x network speedup over Ethernet

- Caffe-on-Spark

- › Apache Spark (spark.apache.org)
 - open source for cluster computing
- › Caffe (caffe.berkeleyvision.org)
 - Open source for deep learning

Deep Learning: 19x Speedup



- 1x1 GPU
 - › 39 hours ... 60% top-5 accuracy
- 4x8 GPUs
 - › 10 hours ... 80% top-5 accuracy
 - › 19x speedup estimated over 1x1 GPU
 - We are working on larger speedup

Summary

- Machine learning is critical for Yahoo business
 - › Search, advertisement, Flickr, weather, recommendation, security, ...
- Scalable ML platforms built on Hadoop & Spark
 - › Heterogeneous cluster: CPU + GPU
 - › Approximate computing

Additional Info

- Scalable machine learning
 - › <http://bit.ly/1NW3GvO>
 - › <http://bit.ly/1VedY0f>
- Flickr use case
 - › <http://bit.ly/1KEytuV>
 - › <http://bit.ly/1KIDfof>
- Yahoo Hadoop blog
 - › <http://yahoothadoop.tumblr.com/>

Thanks!

We are hiring.

bigdata@yahoo-inc.com

