Big Data is Worthless without Artificial Intelligence

Austin, TX May 2016

Erin LeDell Ph.D.
Machine Learning Scientist
H2O.ai

Introduction

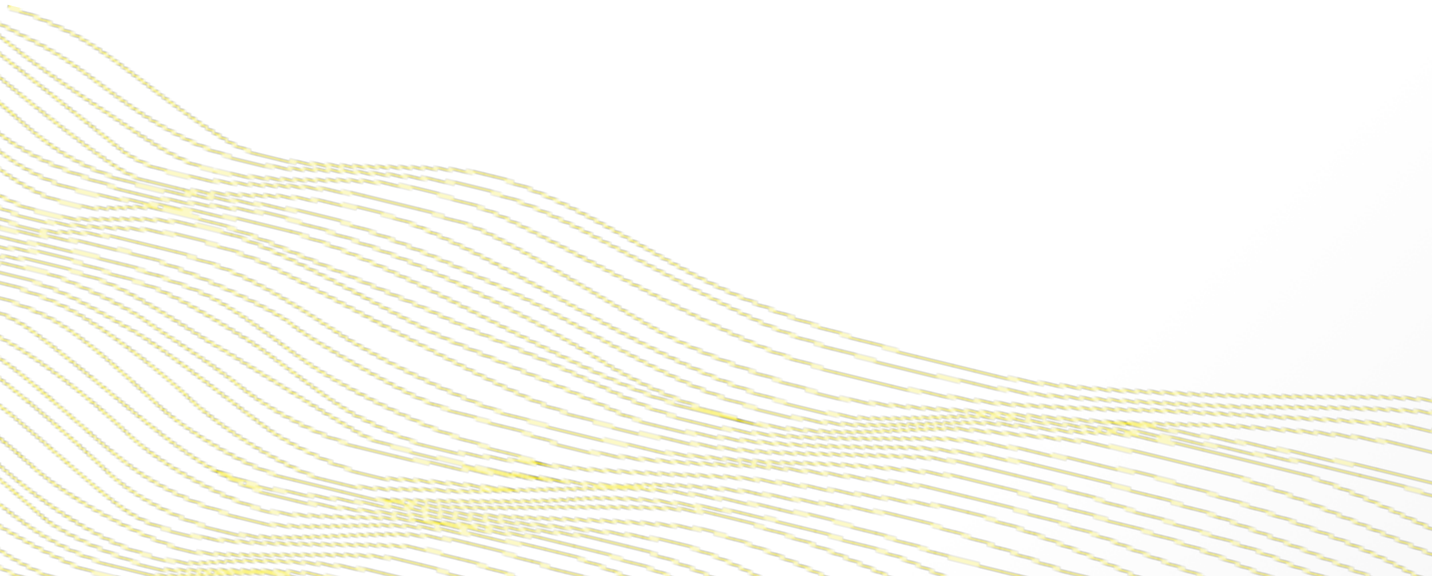
- Statistician & Machine Learning Scientist at H2O.ai in Mountain View, California, USA
- Ph.D. in Biostatistics from UC Berkeley (Machine Learning)
- Worked as a data scientist at several startups
- Developer of open source machine learning software

Berkeley UNIVERSITY OF CALIFORNIA H_2O.ai

Agenda

- What is “Big Data” Anyway?
- Big Data Platforms
- Big Data Analytics 1.0
- Artificial Intelligence & Machine Learning
- Get More Value from Your Data
- Big Data Machine Learning

“Big Data”



What is “Big Data”?

“Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.”

— Wikipedia (2016)

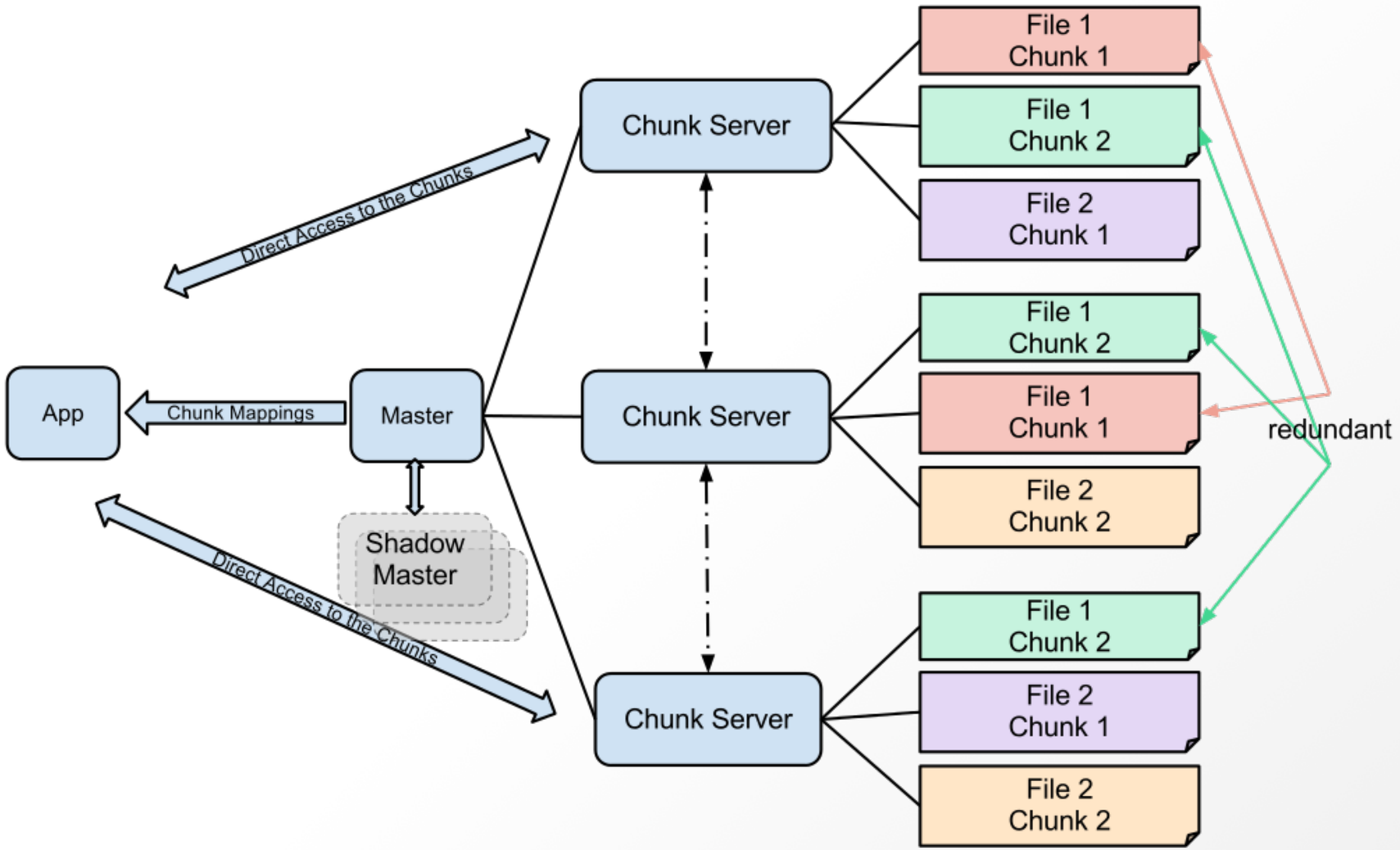
“Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.”

— Edd Dumbill, O’Reilly Media (2012)

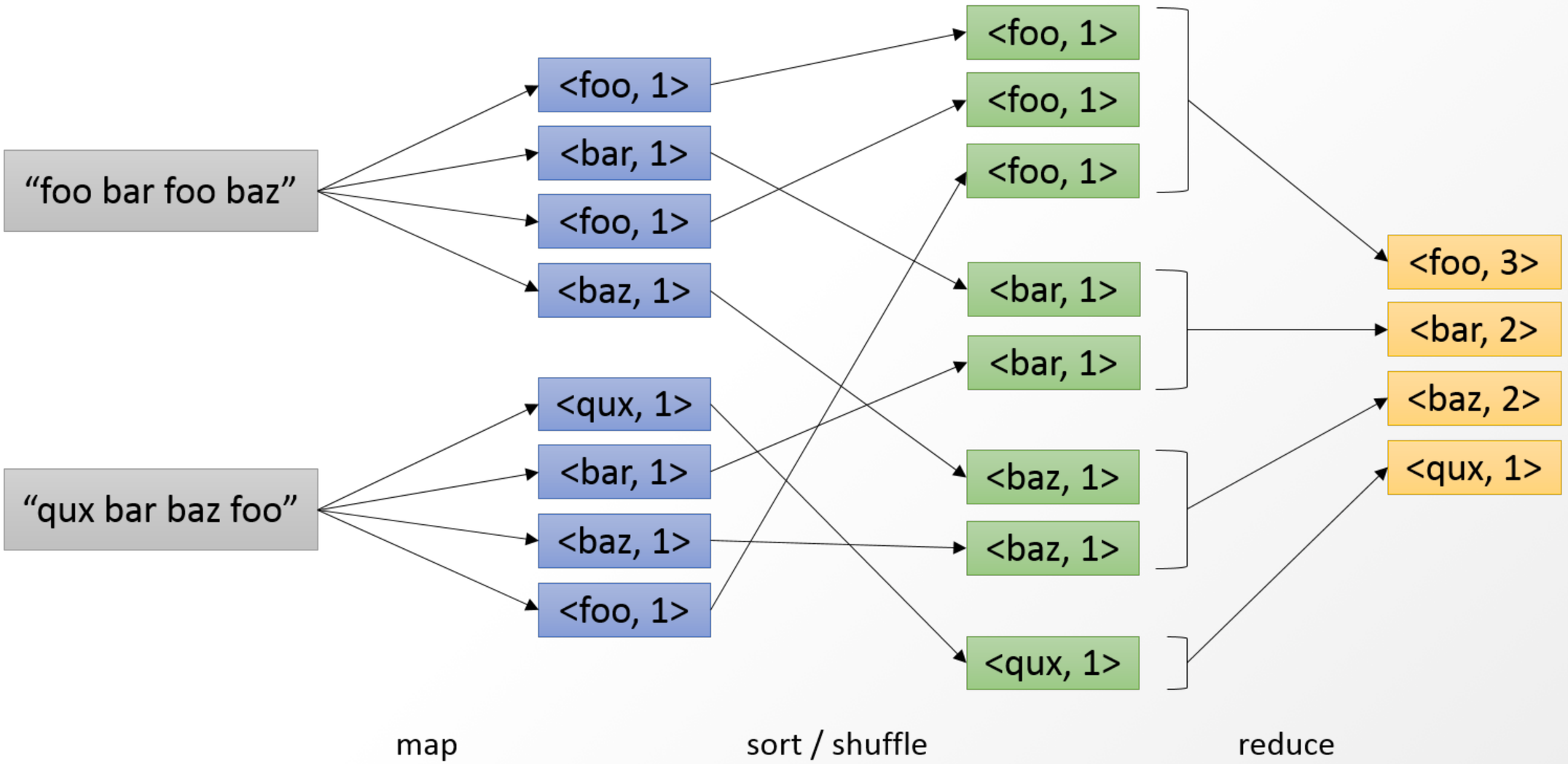
Where did Big Data come from?

- Distributed File Systems: Google File System (2003)
- Google File System grew out of an earlier Google effort, "BigFiles", developed by Larry Page and Sergey Brin while still at Stanford.
- One of the earliest and most popular programming paradigms in big data is “MapReduce”, developed at Google.
- *MapReduce: Simplified Data Processing on Large Clusters* by Jeff Dean and Sanjay Ghemawat

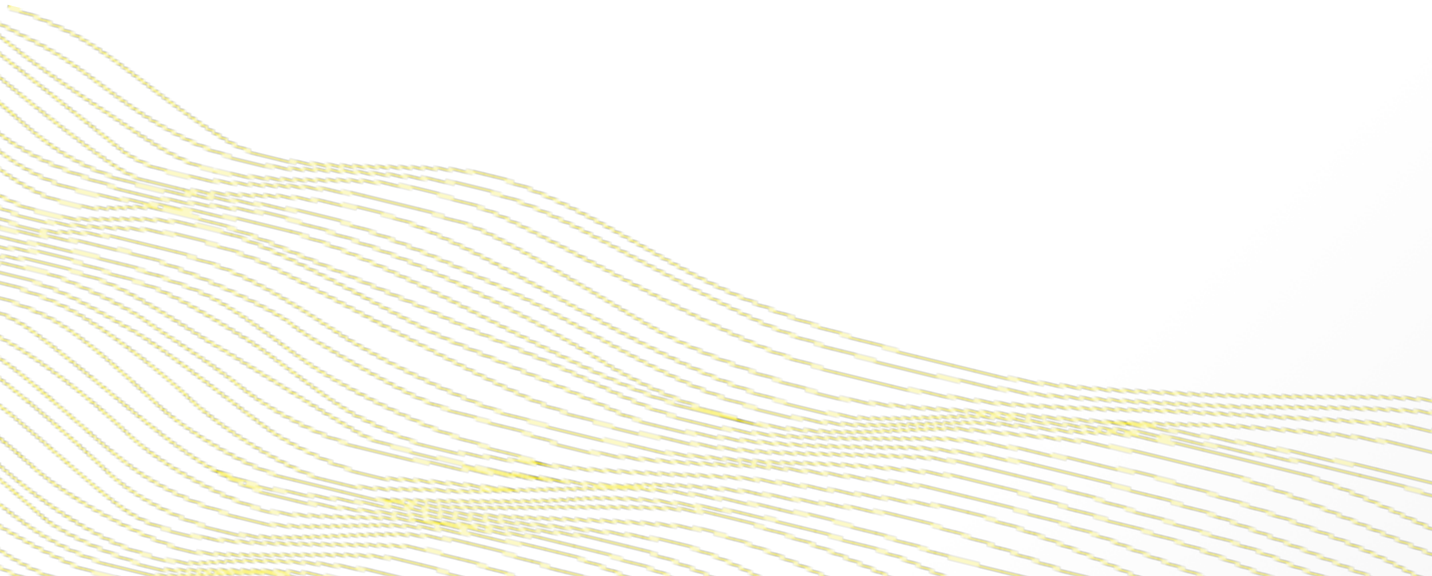
Google File System (GFS)



MapReduce by Example (Word Count)



Big Data Platforms



Big Data Platforms

- Hadoop uses MapReduce for large scale batch processing.
- Reads data from disk, processes data and writes data back to disk to store results.
- YARN eliminates some of the bottlenecks in Hadoop's MapReduce.

- Spark makes it possible to perform streaming, batch processing and machine learning all in the same cluster.
- Spark is about 10-100 times faster at batch processing than Hadoop MapReduce.
- Spark code is easy to write.

Apache Hadoop and HDFS

- Hadoop is an open source implementation of the MapReduce paradigm with HDFS (vs the proprietary GFS).
- It was released in 2006 by Doug Cutting at Yahoo!
- The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework.
- HDFS uses TCP/IP sockets for communication.

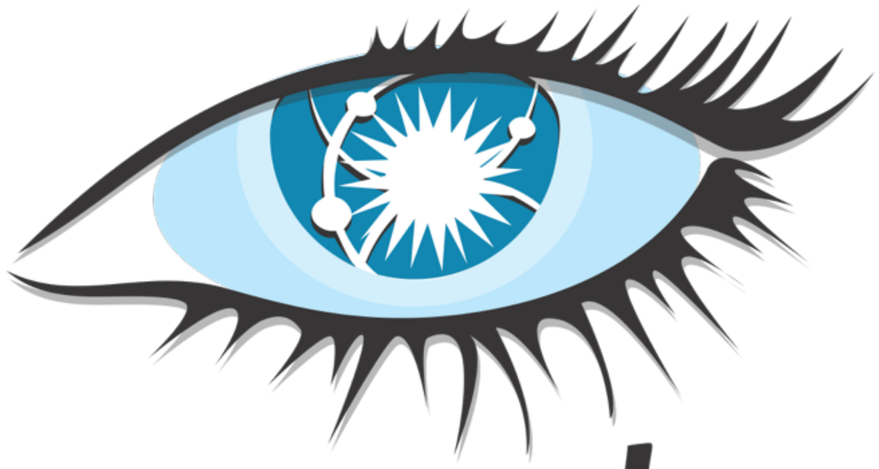
Apache Spark

- Apache Spark is a fast, in-memory, general-purpose cluster computing system, developed at UC Berkeley's AMPLab by Matei Zaharia and first released in 2014.
- It was developed in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs.
- It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs.

Spark Streaming

- Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics.
- It ingests data in mini-batches and performs transformations on those mini-batches of data.
- This design enables the same set of application code written for batch analytics to be used in streaming analytics, on a single engine.

NoSQL Databases



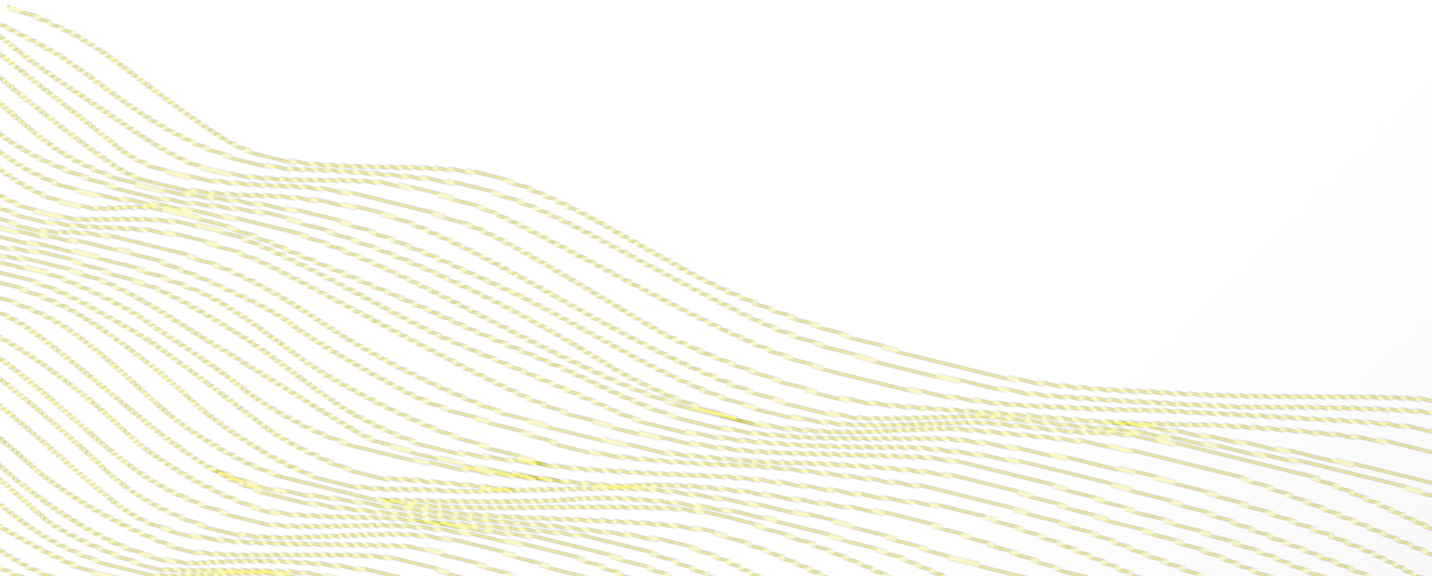
cassandra

APACHE HBASE

mongoDB®

In addition to frameworks like Hadoop and Spark, there are also a whole slew of document-oriented and/or distributed databases that specialize in unstructured data.

Big Data Analytics 1.0



Big Data Analytics 1.0

```
context = HiveContext(sc)
results = context.sql(
  "SELECT * FROM people")
names = results.map(lambda p: p.name)
```

- The original set of big data analytics tools centered around scalable, yet basic, querying of the data, similar or equivalent to SQL queries.
- Examples of this are Hadoop's PIG and HIVE query engines.
- Spark includes "Spark SQL" for querying structured data, and supports unmodified HIVE queries.

Big Data Analytics 1.0

For NoSQL databases, there are a number of different query engines.

Big Data Analytics 1.0

According to Gartner, 64% of organizations had purchased big data systems in 2013, yet only 8% actually deployed big data technology.

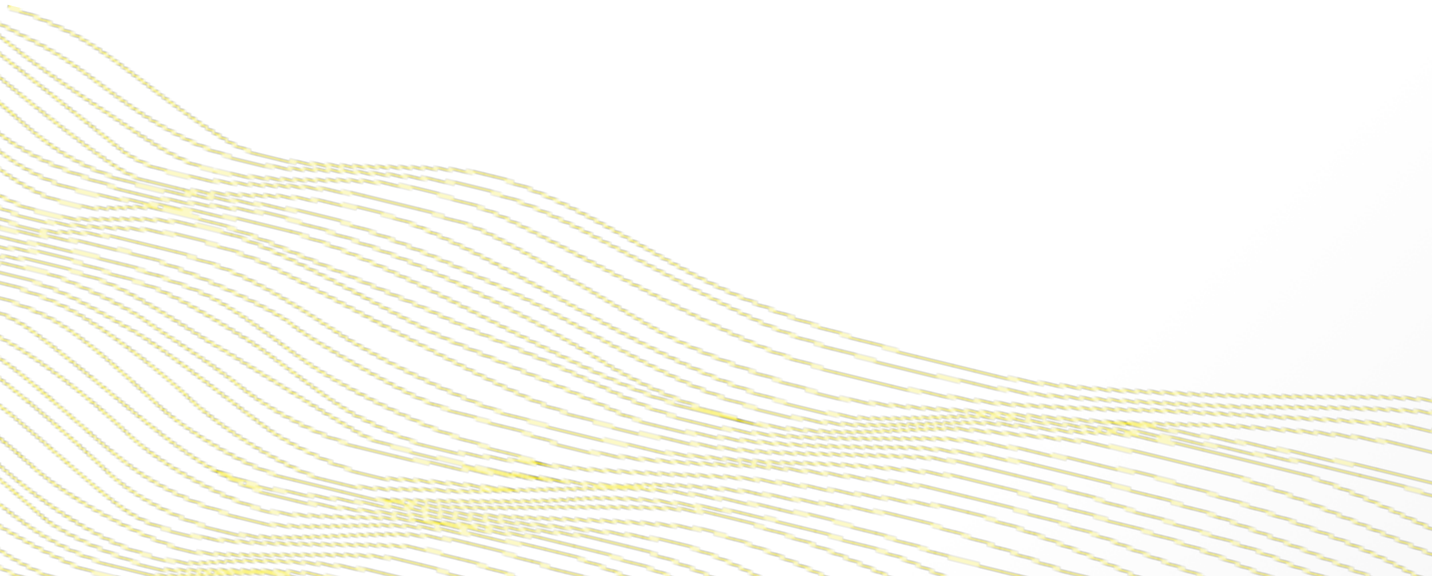
Before companies start collecting big data, they should have a clear idea of what they want to do with it with from a business sense.

Big Data Analytics 1.0

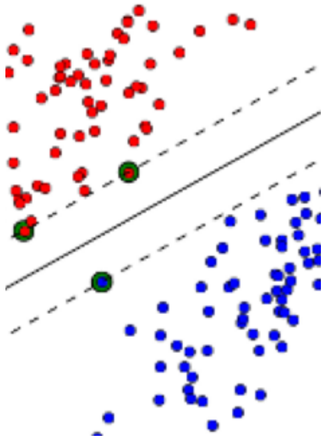
The first generation of Big Data Analytics tools failed to deliver on the promise of producing huge financial returns for companies.

> “Instead of offering a clear windshield, the big data phenomenon is more like a big rear-view mirror telling us nothing about the future.”
>
> — Philippe Silberzahn & Milo Jones, *Constructing Cassandra* (2013)

Artificial Intelligence & Machine Learning



What is Machine Learning?



"Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel, 1959

Unlike rules-based systems which require a human expert to hard-code domain knowledge directly into the system, a machine learning algorithm learns how to make decisions from the data alone.

Machine Learning Tasks

Regression

- Predict a real-valued response (e.g. viral load, price)
- Gaussian, Gamma, Poisson, etc. distributed response
- Evaluate with MSE or R^2

Classification

- Multi-class or binary classification
- Ranking (e.g. Google Search results order)
- Evaluate with Classification Error or AUC

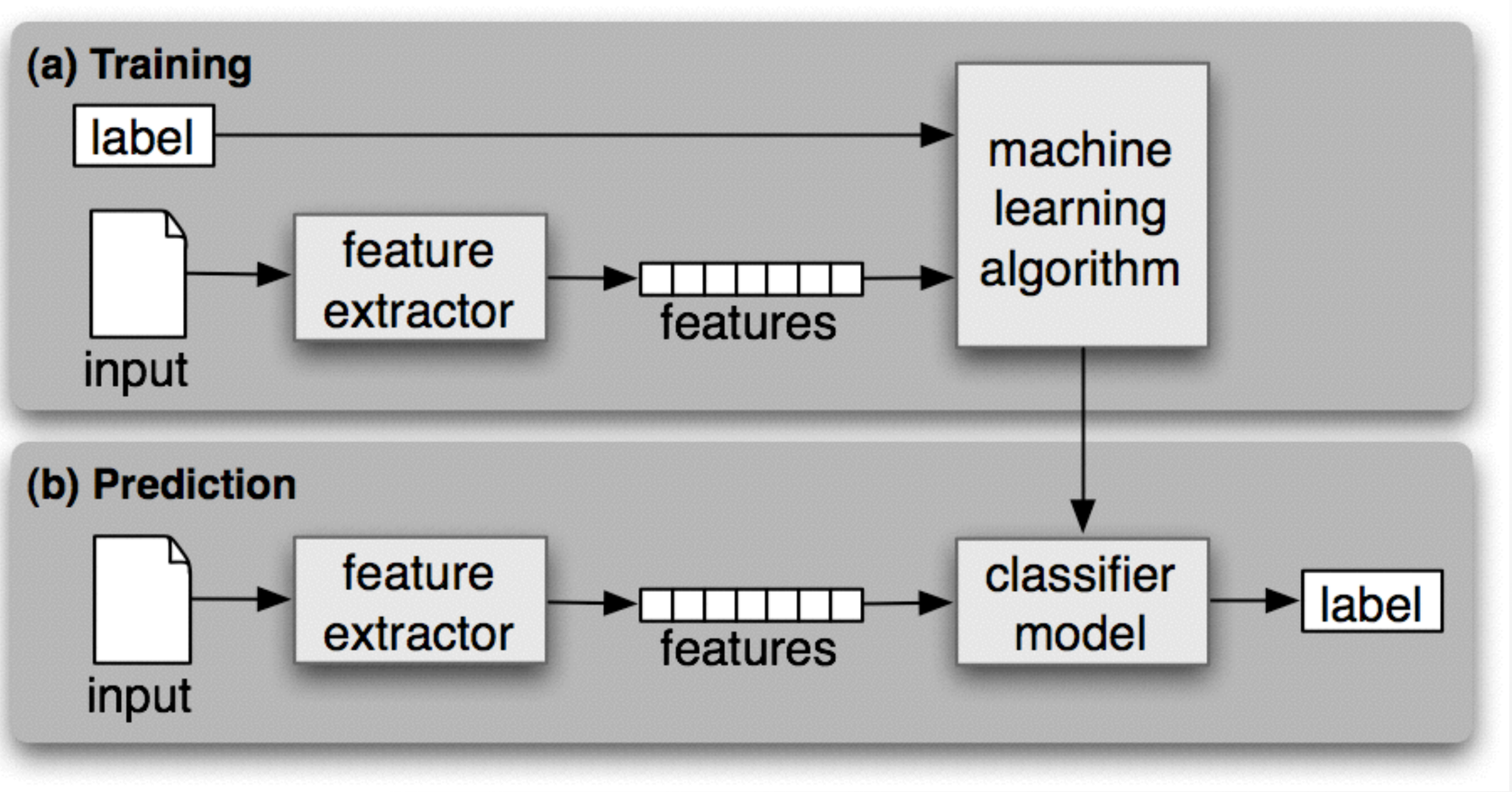
Clustering

- Unsupervised learning (no training labels)
- Partition the data; identify clusters or sub-populations
- Evaluate with AIC, BIC or Total Sum of Squares

Train, Validation and Test Set

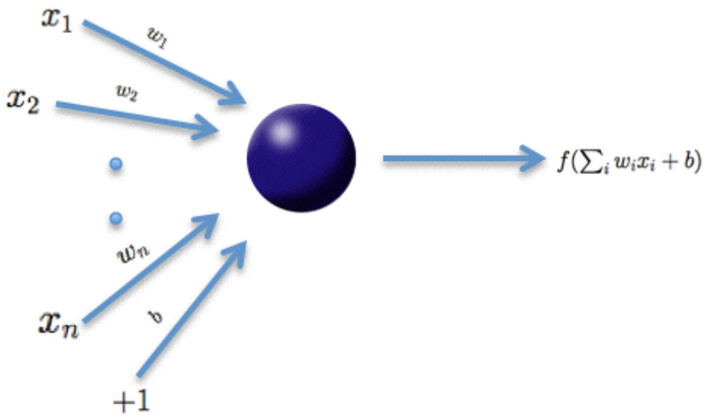
- If you plan on doing any model tuning, you should split your dataset into three parts: Train, Validation and Test
- There is no general rule for how you should partition the data and it will depend on how strong the signal in your data is, but an example could be:
  50% Train, 25% Validation and 25% Test
- The validation set is used strictly for model tuning and the test set is used to make a final estimate of the generalization error.

Machine Learning Workflow



Training and Prediction
in machine learning

What is Deep Learning?

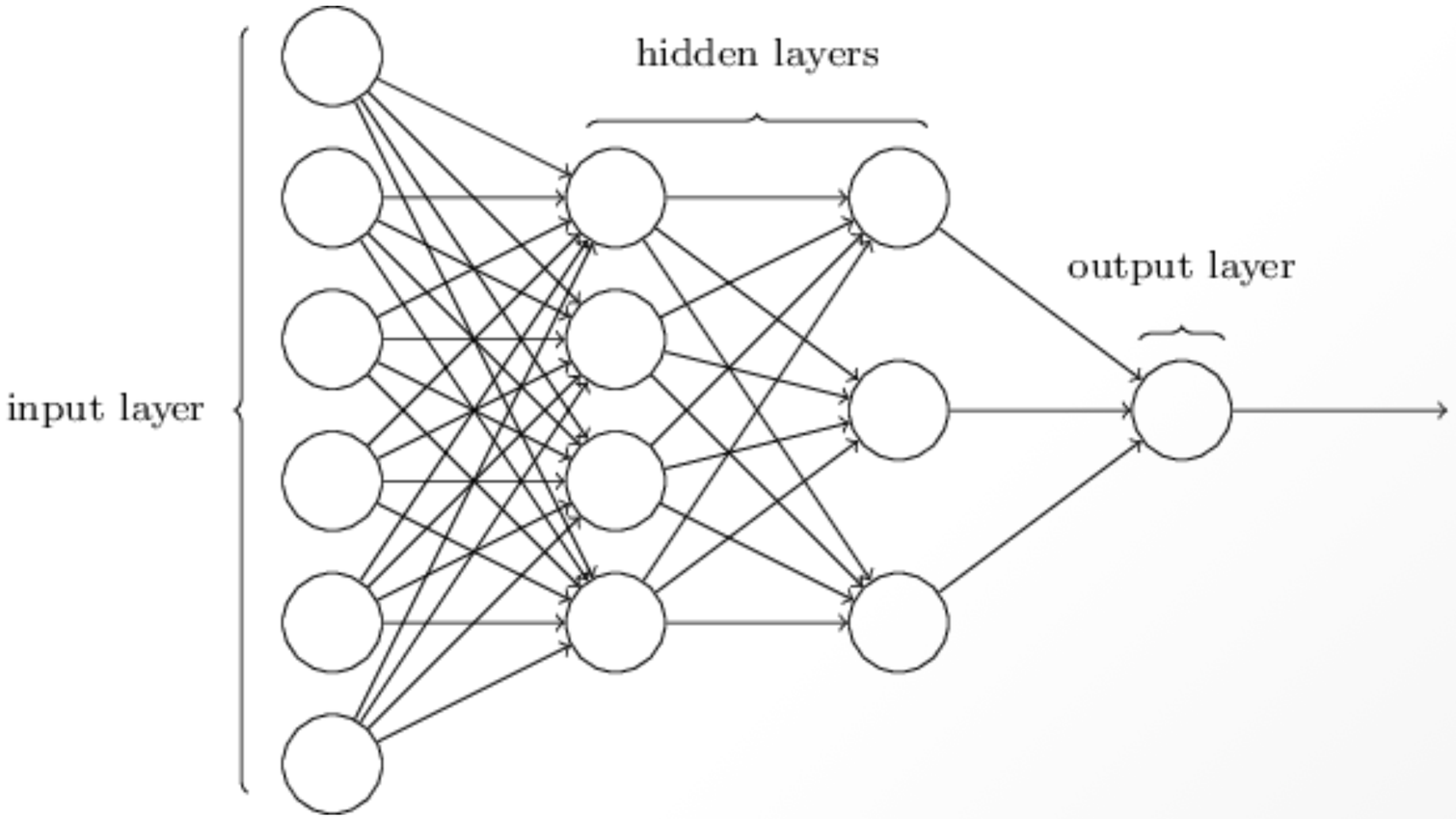


"A branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, composed of multiple non-linear transformations."

— Wikipedia (2015)

- Deep neural networks have more than one hidden layer in their architecture. That's why they are called "deep" neural networks.
- Very useful for complex input data such as images, video, audio.

What is Deep Learning?



- Deep learning architectures, specifically artificial neural networks (ANNs) have been around since 1980.
- However, there were breakthroughs in training techniques that lead to their recent resurgence in the mid 2000's.
- Combined with modern computing power, they are quite effective.

What is Ensemble Learning?

"Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms."

— Wikipedia (2016)

- Random Forests and Gradient Boosting Machines (GBM) are both ensembles of decision trees.
- Stacking, or Super Learning, is technique for combining various learners into a single, powerful learner using a second-level metalearning algorithm.

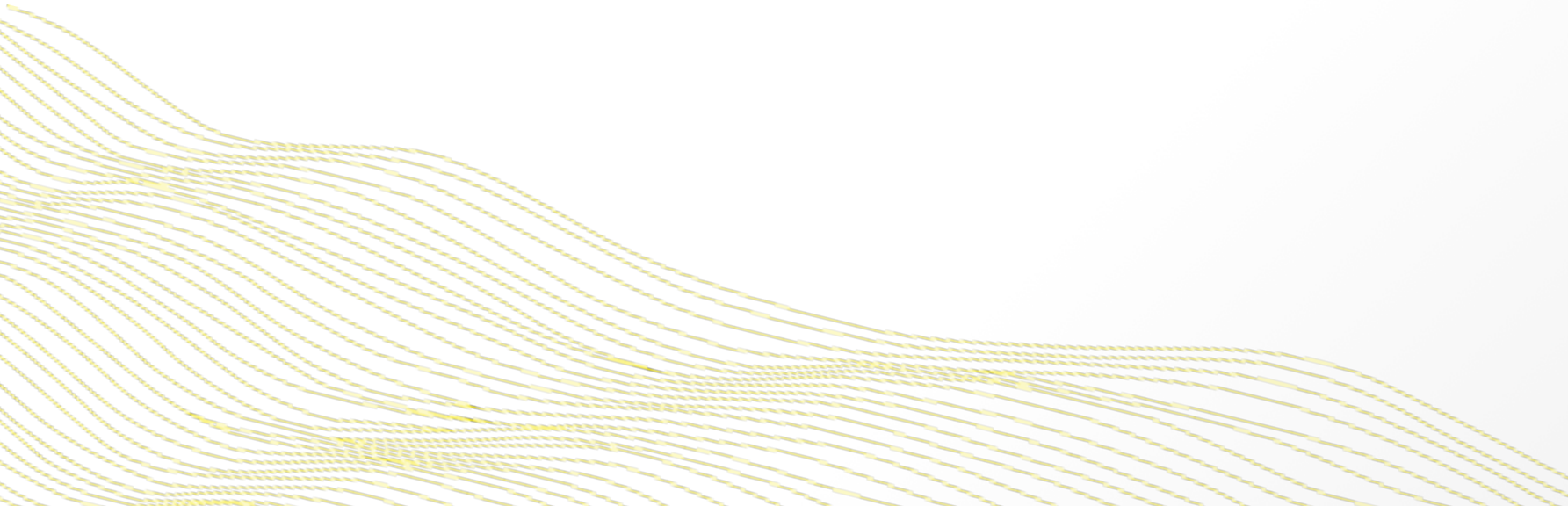
No Free Lunch

"Even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience."

— David Hume (1711-1776)

- No general purpose algorithm to solve all problems.
- No right answer on optimal data preparation.
- Some algorithms may have such strong biases that they can only learn certain kinds of functions.

Get More Value From Your Data



Turn Data into Valuable Insights

Applications of Machine Learning

CIO

Home > Business Analytics > Analytics

TODAY'S TOP STORIES

Starwood taps machine learning to dynamically price hotel rooms

Credit: Thinkstock

Starwood Hotels & Resorts Worldwide uses an analytics engine to alter hotel pricing rates on the fly, improving demand forecasting by 20 percent.

By Clint Boulton | Follow
CIO | May 13, 2016 1:01 PM PT

MORE

MORE LIKE THIS ::


Will hotel room keys and desk check-in soon be obsolete?

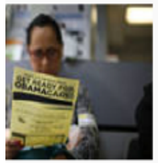

Make analytics pay off for you and your customers


How Hilton delivers greater convenience to hotel guests

on IDG Answers
Can someone else use my mobile data?

Applications of Machine Learning

THE WALL STREET JOURNAL.



Insurer Highmark Sues U.S. Over Affordable Care Act

Apple Looks to India for Growth

Shale Driller Key to Survival: Efficiency

CIO JOURNAL.

The Security Download: Anticipating Cyberattacks with Machine Learning

By RACHAEL KING

Mar 9, 2015 10:35 am ET

0 COMMENTS

CONTENT

Applications of Machine Learning

HNGN
TUESDAY, MAY 17, 2016
HEADLINES & GLOBAL NEWS

STARS & THEIR PETS EXCLUSIVE: Angie Everhart And Her Dog Duk

HNGN's Rescue Pet Of The Week: Adoptable Guinea Pigs Felix A...

Lions And Tigers And Bears, Oh My! Leo, Shere Khan And Baloo...

Machine Learning Helps Scientists Uncover New Materials With Desirable Properties

Researchers from Los Alamos National Laboratory have successfully used a machine-learning algorithm to discover new materials with specific desirable properties.

By Tyler MacDonald | May 10, 2016 10:01 AM EDT

Applications of Machine Learning

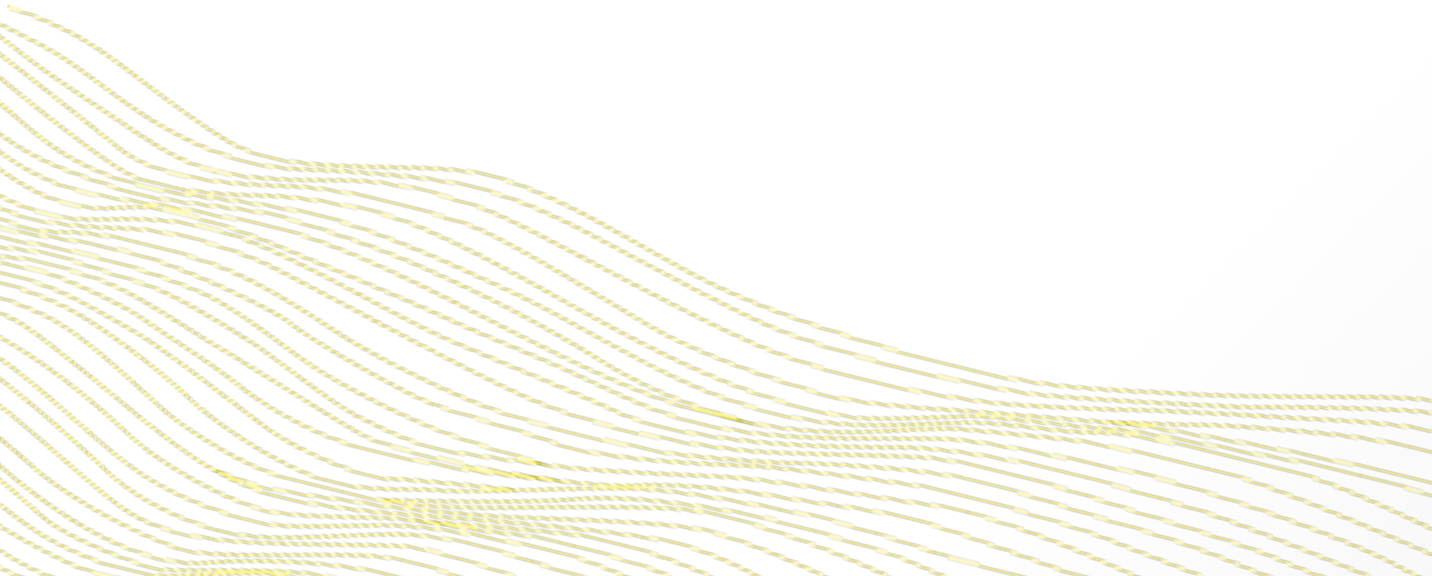
MUST READ SAP'S SAPPHIRE NOW CHALLENGE: SELLING EMPATHY OVER ECONOMICS

Telstra using machine learning to deal with 'distressed' customers

Telstra's general manager for analytics has said the telco giant has been employing machine learning techniques in attempt to reduce the frustration of its distressed customers.

By Asha Barbaschow | April 21, 2016 -- 00:50 GMT (17:50 PDT) | Topic: Telcos

Big Data
Machine Learning

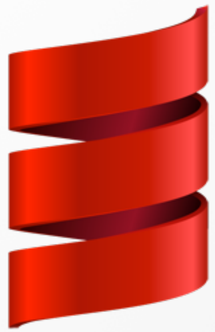


Open Source Big Data Machine Learning

H2O Platform Overview

- Distributed implementations of cutting edge ML algorithms.
- Core algorithms written in high performance Java.
- APIs available in R, Python, Scala, REST/JSON.
- Interactive Web GUI.




H2O Platform Overview

- Write code in high-level language like R (or use the web GUI) and output production-ready models in Java.
- To scale, just add nodes to your H2O cluster.
- Works with Hadoop, Spark, databases and your laptop.

H2O R & Python Code Tutorial

http://tinyurl.com/h2o-chicago

Thank you!

@ledell on Github, Twitter

erin@h2o.ai

http://www.stat.berkeley.edu/~ledell