# Unit 1: Introduction and Motivation

## Videos:

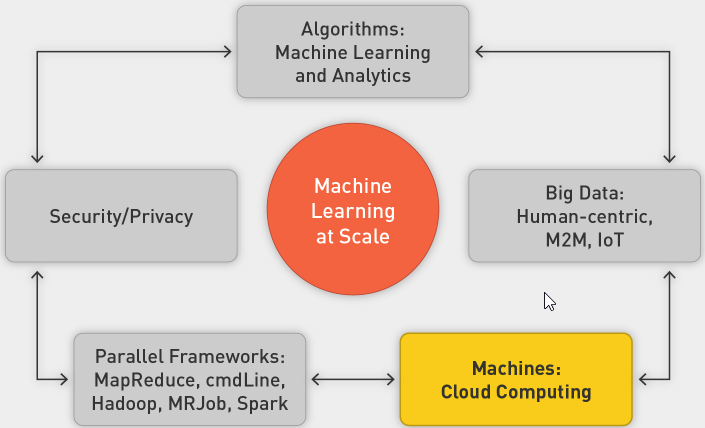
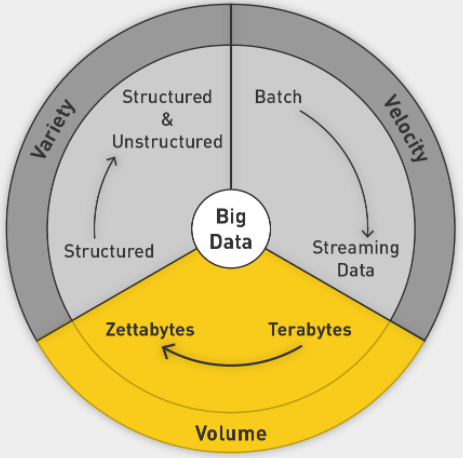
Introduction

1. Intro
2. Big Data
3. Modern-day DS
4. Seven steps in data modelling pipeline
5. Large scale machine learning
6. Tricks of the trade
7. MR using the command line

Readings

1. Doing Data Science Chapter 1 and 2: on Safari
2. An introduction to information retrieval: Chapter 13, downloaded; Info on Naïve Bayes and Text Classification
3. Elements of Statistical Learning, Chapter 7: downloaded and read

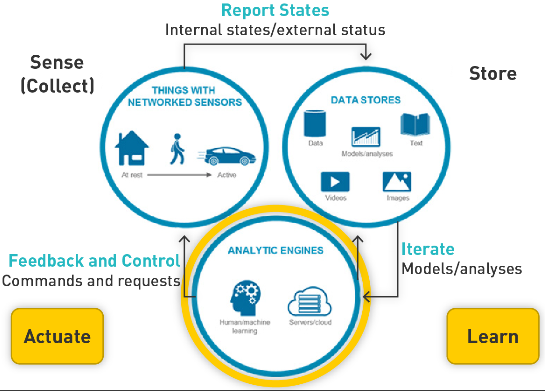
Big Data Definitions

1. ML at Scale
   1. 
2. Big Data Definition
   1. Big data is broad term for data sets so large or complex that traditional data-processing applications are inadequate:
   2. Processing
   3. Laptop (8–16 GB of memory, 1 TB hard drive) overwhelmed with 4–5 GB
   4. Storage
   5. Laptop only 1 TB
   6. Throughput
   7. Three hours to read 1 TB on laptop
   8. Other Challenges
      1. Analysis
      2. Capture
      3. Data curation
      4. Search
      5. Sharing
      6. Storage
      7. Transfer
      8. Visualization
      9. Security
      10. Information privacy
   9. Three Vs
      1. 
      2. Add Veracity: uncertainty of data
3. Sources Driving Big Data
   1. Everything is recorded online
   2. User-generated world, social media
   3. IoT
   4. Scientific Computing
4. Areas of Application
   1. NYSE
   2. FB
   3. CERN
5. Data in Zettabytes (ZB)
   1. 1ZB is 10^21
   2. 2020: 40 ZB

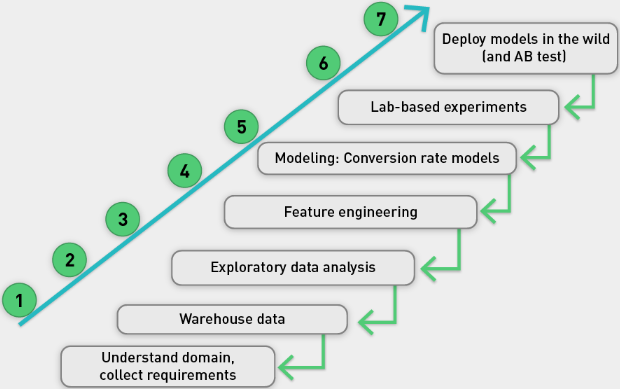
Sources of Big Data

1. Applications of Big Data
2. Societal and Personal Sources of Data
3. Online Society
4. Advertising
   1. Represents 2% of GDP in United States since 1900s
   2. Increased spending online as opposed to on traditional channels
   3. Mobile advertising surpasses others
5. Dating
6. Life Logging
7. Health Care and how tech can help
8. Genomics

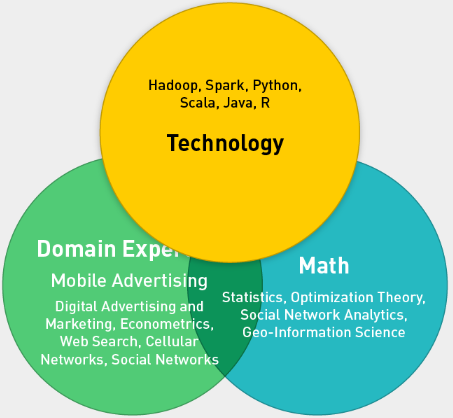
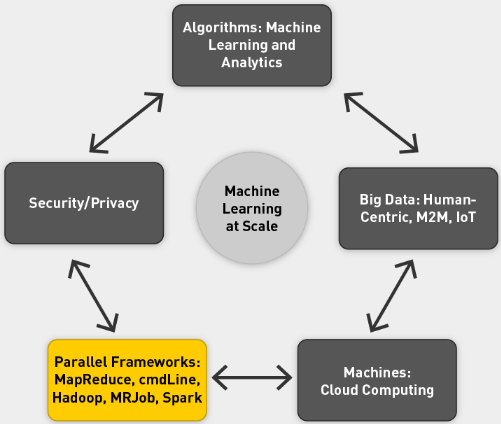
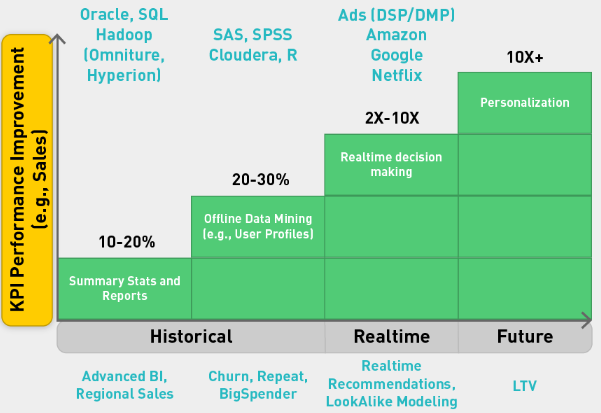
IoT

1. IoT
   1. More data coming from machines than from humans
   2. Machine to machine communication
   3. Attach sensors to all the things
   4. 2020: 26b things on network
   5. Everyone has a unique identifier and data is automatically sent across the internet
2. Japan Earthquake
   1. 14 seconds before, all trains and plants shut down
   2. Mesh of sensors from connected computers to form an early warning system
3. Tracking Nature in the Wild
   1. Sea lions with GPS tracking system
   2. Underwater hub picks up signals from devices
   3. Serengeti
4. Challenges
   1. Battery life
   2. Storage
   3. Network connectivity
   4. Privacy and security
5. Smart homes
6. Smart Cities
   1. 
7. Autonomous Vehicles

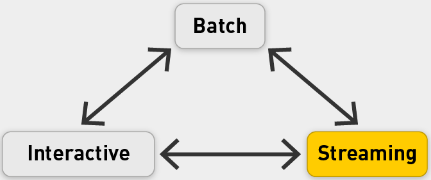
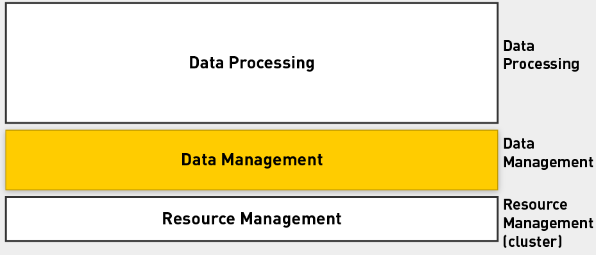
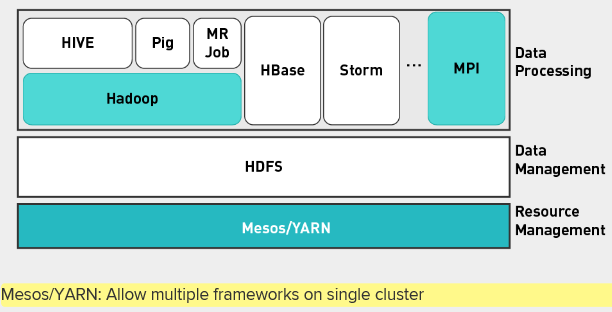
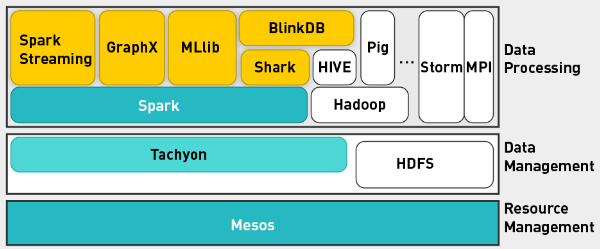
Data Modelling Pipeline

1. Intro
   1. Massive challenge for machine learning
   2. ML is not in a vacuum
   3. ML in real world
2. Modelling
   1. 
   2. Very iterative, might go back at every step
   3. 80% of work is spent in getting, prepping and engineering data
3. Example US insurance
   1. Half of it is property and casualty (auto, home, commercial)
   2. Highly competitive rates
      1. Special rates for good drivers
      2. Score computed through accident history over five years
      3. Static, annual, batch process, easy
4. Progressive’s Snapshot product
   1. Bad driver surveillance through telematics
   2. Small box plugged into steering wheel that records, sends information
   3. Features:
   4. How many miles driven
   5. How many miles driven between midnight and 4 a.m.
   6. Use of sharp or gentle braking
   7. Data used to analyze driving patterns, risk to insurance company
   8. Potential reduced premiums for user
5. Privacy
   1. Huge data generated
   2. Data not transitory
   3. Caution needed with who can access and how is it stored
   4. Who owns the data?

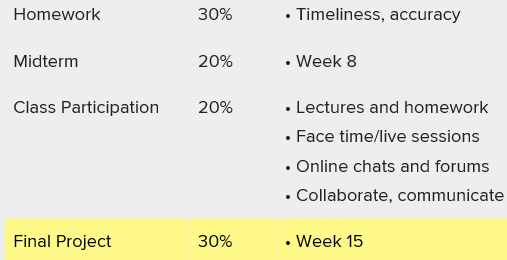
Data Scientist: Post This Class

1. Three skill sets
   1. 
2. Understand Models
3. Understand Security and Privacy
4. ML at scale
   1. 
5. DS Team with different experts in the three fields
6. DS improves KPIs dramatically
   1. 

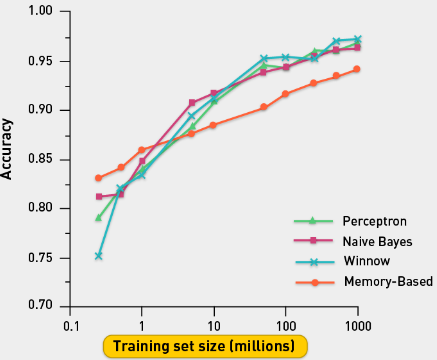
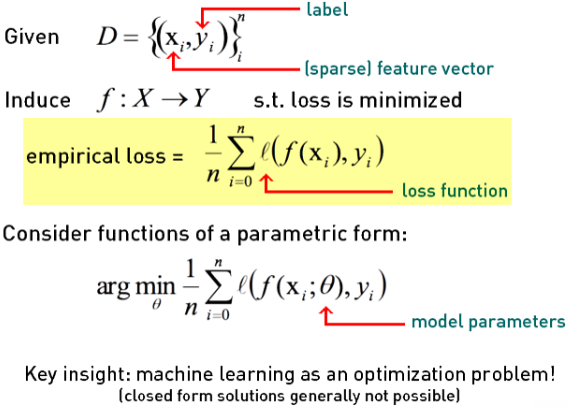
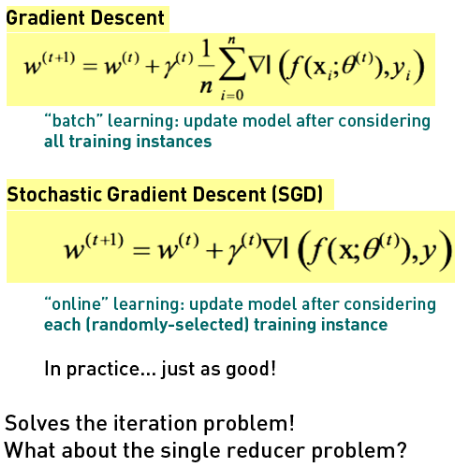
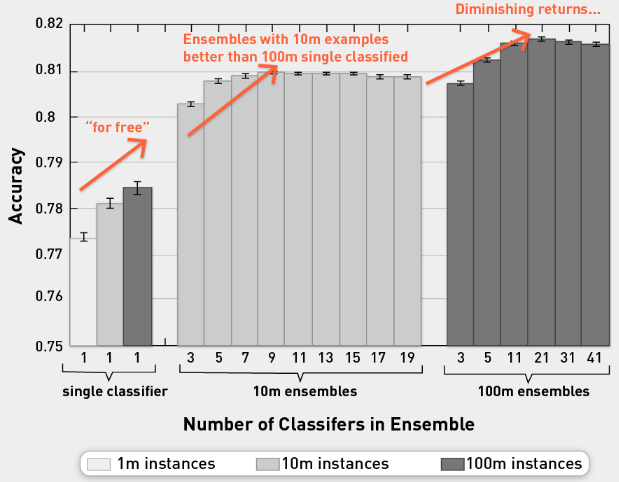
Goals of Class

1. ML at Scale
2. Large-Scale ML
   1. Needs supporting infrastructure
   2. Parallel computing frameworks
3. MapReduce Framework
   1. Allows us to divide and conquer huge data problems
   2. Four frameworks
      1. Cmd
      2. Hadoop
      3. MRJob
      4. Spark
4. Embarrassingly Parallel Problems
   1. Most ML problems
   2. Divide into tasks that can easily be distributed and parallelized
   3. Divided into subproblems
   4. Require little or no communication
   5. Are easily distributed
   6. Have linear computational flow and scale
5. Framework Requirements
   1. Scalable
   2. Fault tolerant
   3. Iterative
   4. Interactive
   5. 
   6. Frameworks that empower us to do above
   7. OS ecosystem
6. Four MR frameworks
   1. CMD based
   2. Hadoop: Storing and manipulating data
   3. MRJob: Convenient Python application programming interface (API) for writing MapReduce programs
   4. Spark: memory backed with integrated framework; Specialized libraries for many things
7. Data analytics stack
   1. 
   2. Populating the stack
   3. 
   4. Add Spark
   5. 
8. Class Phases
   1. Phase 0: CMD
   2. Phase 1: Hadoop/HDFS
   3. Phase 2: MRJob
   4. Phase 3: Spark
9. ML Algorithms
   1. Adopt them for parallel framework
   2. Supervised machine learning (convex optimization, gradient descent, linear regression, decision trees, ensembles of models, support vector machines)
   3. Unsupervised (expectation maximization, matrix multiplication, alternating least squares)
   4. Graphs (random walks, PageRank, graph search algorithms such as breadth-first search, shortest path)
   5. Hybrid algorithms (supervised machine learning and random walks)
   6. Applications (digital advertising, social media, health care, e-commerce, entertainment, metrics, statistics)

Logistics and Performance Evaluation

1. The more you put in, the more you get back
2. Structure
   1. 13 lectures over 14 weeks
   2. Two exam weeks
   3. Week 8 midterm
   4. Week 15 final project
3. Performance evaluation
   1. 

Large Scale ML

1. Do you need more data scientists or do you need more data?
2. ML and Data Study
   1. Filling in confusable words in sentences
   2. Use ML
   3. Vary size of training data
   4. 
   5. Different algorithms perform differently if they have small number of training examples, but similarly with large number of examples
   6. Conclusion: more data leads to 10-20% boost in performance
3. Supervised Classification in a nutshell
   1. 
   2. 
4. Complexity of models
   1. 
   2. Moving from single to ensemble, you get huge boost in performance
   3. However, if you use 100m examples, you get further boost of 1-2%.
   4. Using more data we get amazing improvements, but we hit diminishing returns
5. More data or more DS?
   1. Trade-off between bias and variance
   2. Bias from data scientists
   3. Variance from more data
   4. With more data can reduce variance

Bias-Variance Background