ISA 414 – Managing Big Data

Lecture 29 – Review Session

Dr. Arthur Carvalho
arthur.carvalho@miamioh.edu

MIAMI UNIVERSITY

Announcements

- Last lecture :~(
  - Stay in touch:
    - Arthur: https://www.linkedin.com/in/arthurgcarvalho/

Announcements

- **Project**
  - Deadline: Sunday, December 5th, 11:59 pm
    - Report + code + raw data set(s)
    - I must be able to run your code on the provided data sets and reproduce all of your results
  - Remember to fill in the "*Peer and Self Evaluation Form*" available on Canvas
    - You will only receive your individual project grade after submitting the evaluation form
    - Final, individual grades might differ from group grades
      - Group members who contributed to the project will have their grades raised
      - Group members who did not contribute as much to the project will have their grades lowered

Announcements

- **Conceptual final exam**
  - Available from Monday, December 6th (12 a.m.) to Saturday, December 11th (11:59 p.m.)
  - No need to come to campus
    - Please, let me know if you would like to use the PC lab
  - More on the format of the exam later in this lecture

Announcements

- **Course evaluations**
  - It is time to get some feedback from you
  - Changes based on previous course evaluations:
    - Course difficulty
      - Yes, the course is easier now than it was before
    - Hands-on activities with IBM Cloud
    - Final exam format
    - No more classes on relational database
    - More time for the final project

Announcements

- **Course evaluations**
  - What worked and what could be improved?
  - Please, let me know your thoughts (5-minute survey)
    1. Go to mymiami.miamioh.edu
    2. In the “My Courses” tab, click on “Course Evaluations” for the course **ISA 414/514 – Managing Big Data**

Lecture Objectives

- Review Assignment 4
- Introduction to the format of the final exam
- Review the major concepts learned in the course
  - Focus on the final exam

Final Exam

- Online
  - Available from December 6th (12 a.m.) to December 11th (11:59 p.m.)
- Duration: 2h
- You are allowed to use any search engine, slides, notes, …
  - Be very careful with the amount of time you spend on search engines or looking at notes
- No cell phones, second monitors
- No access to websites that allow for collaboration
- No more than one person in the room

Final Exam

- Format: similar to previous quizzes
  - Multiple-choice questions
  - Some questions have more than one correct answer
- 25 questions in total, 4 points each
  - Randomly drawn from a database containing 100 questions
  - Covers every single topic we discussed in class
    - No Python code

Final Exam

- Security measures (please, pay attention to this)
  1. We will use Proctorio for proctoring
  2. The orders of questions and answers are random
  3. You have only one shot
     - Your exam grade is final

REVIEW SESSION

Big Data

- Managing big data
  - Big data
    - Defined by a number of “Vs”
      - **V**olume: vast amount of data
      - **V**ariety: different forms of data (text, images, voice, geospatial, *etc*.)
      - **V**elocity: speed at which data is generated/analyzed
      - **V**eracity: biases, noise, abnormalities, uncertainties, truthfulness, trustworthiness of the data

CRISP-DM

- Key concept: CRISP-DM cycle

not covered in this course

MIAMI UNIVERSITY

Business Understanding

- Can we translate the business problem into a data analytics task (or many subtasks)?
- Supervised vs unsupervised learning techniques
  - Supervised means there is a clear target we use when training models
- Occasionally, we do not need statistical models
  - Descriptive analytics is enough to solve a business problem

covered in this course

Problem	Supervised	Unsupervised
Classification	X	
Regression	X	
Similarity Matching	X	
Clustering		X
Co-occurrence grouping		X
Profiling		X

Data Understanding and Collection

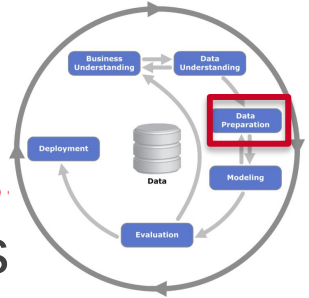
- We learned about:
  - Web & log scraping
    - Regular expressions
  - API
    - Request: REST
    - Response: JSON and XML
  - Database queries
    - Document-oriented databases (JSON)

Data Preparation/Preprocessing

- Raw data is cleaned/prepared for analysis
- Traditional steps:
  - Remove/impute invalid/missing values
  - Merge duplicate observations and/or different data sets
  - Transform/remove outliers
- Feature selection
  - Remove highly correlated predictors
  - Scaling
  - Dimensionality reduction

not covered in this course

Data Preparation/Preprocessing



- We have focused on preparing textual data for analys
  - Bag of words: each word is considered individually (“token”)
  - DTM: document-term matrix
    - Rows = documents
    - Columns = words
    - Cells = counts
      - TF: measures the popularity of a word inside a document
      - IDF: measures the popularity of a word in a corpus
      - TFIDF: measures the popularity of a word in a document and across the whole corpus

Data Preparation/Preprocessing

- Other common techniques
  - Remove stop words
  - Remove numbers
  - Stemming
  - Sparsity reduction
  - N-grams

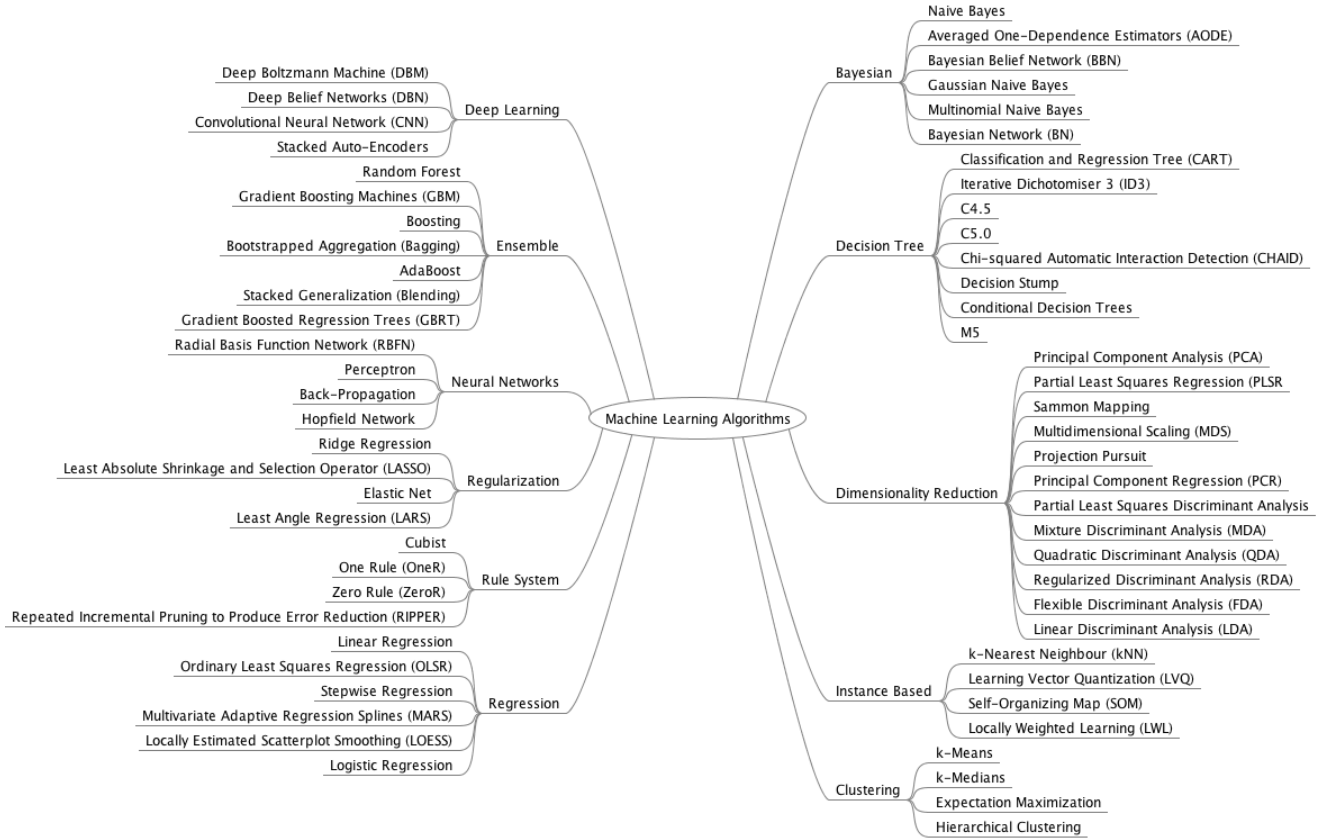
dimensionality reduction

dimensionality increase

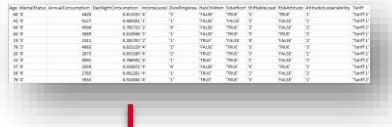
Data Modeling/Analysis

- We have learned a few techniques to analyze data
  - Textual data
    - Sentiment analysis
    - LDA
  - Statistical modeling
    - Classification: decision trees and random forests
    - Regression: regression trees and random forest
  - Keep in mind that there are many other techniques that are beyond the scope of this course

Data Modeling/Analysis



data



machine learning algorithm

model/ solution

MIAMI UNIVERSITY

Evaluation

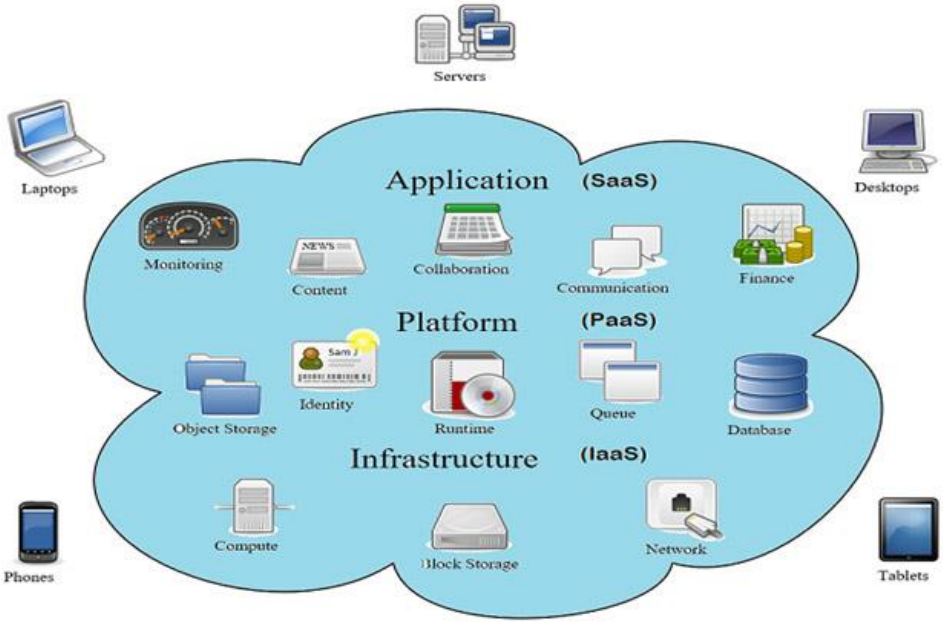
- Many different ways of evaluating a data analytics solution
  - Internal/external experts
  - Statistical analysis of errors when making predictions
    - Build the model using a training set
    - Evaluate the model using a test (holdout) set
    - Different metrics
      - Classification: overall accuracy, specificity, sensitivity, ROC area
      - Regression: MSE, RMSE, MAE, MAPE …

Review

- Big data brings new, complex challenges
  - Volume, variety, velocity, veracity
- Second part of the course: two big-data analytics enablers
  - Cloud computing and storage
    - IBM Cloud, Microsoft Azure, Google Cloud Platform, Amazon Web Services, …
  - Distributed storage and computing
    - Hadoop environment and Spark

Cloud Technologies

- Cloud = rental agreement
  - Different levels of engagement and servicing



Cloud Technologies

- Things to remember
  - Definitions
    - IaaS: Infrastructure as a Service
    - PaaS: Platform as a Service
    - SaaS: Software as a Service
    - XaaS: Everything as a Service
  - In-house vs cloud infrastructure
    - Considerations: Costs, Demand, Skills, Security

Cloud Technologies

- **Q1**: consider a company whose e-commerce sales grow massively during holiday season (*e.g.*, Best Buy). The company's in-house database system suffers tremendously from such an increase in online traffic, which causes the company's website to run slowly. When it comes to the IT infrastructure, what would likely be the most cost-effective and efficient approach to handle the extra online traffic during the holiday season?
  a) Sign an IaaS agreement: the company should move its whole infrastructure to the cloud
  b) Sign a PaaS agreement: the company should move its whole infrastructure to the cloud
  c) Sign a SaaS agreement: the company should follow a hybrid approach where the database workload is distributed between the in-house and cloud infrastructure
  d) Sign a PaaS agreement: the company should follow a hybrid approach where the database workload is distributed between the in-house and cloud infrastructure

Cloud Technologies

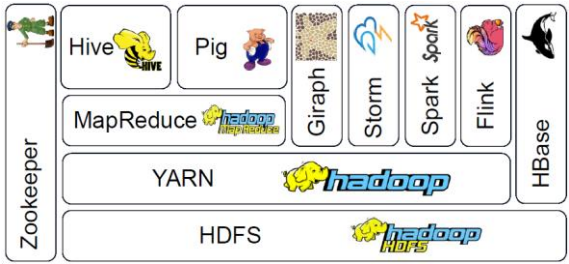
- **Q2**: consider a company that relies on data-analytics sporadically. As such, the company has no dedicated data scientists and no relevant skill set. Suppose that company is now planning a predictive-analytics project. Luckily, the data set to be used was already used in previous projects, meaning that it is ready for analysis. What would likely be the most cost-effective approach(es) for that company?
  a) Sign a PaaS agreement: the company should obtain a complete data-science platform, including IDE and distributed storage/computation frameworks
  b) Hire several seasoned data scientists to complete the task
  c) Sign an XaaS agreement that automates the development and deployment of statistical models

Hadoop

- Cluster of commodity computers
  - **Distributed storage**
    - Individual files are stored in different computers in a cluster
    - Massive files are broken down into chucks of data, which are stored in different computers (nodes) in a cluster
    - Sharing ‘disks’ (secondary storage)
  - **Distributed computation**
    - Complex, time-consuming computations are distributed across different nodes
    - Sharing CPUs/cores and main memory
    - Paradigm: “move computation to data”
      - Reduce network traffic

Hadoop

- **Hadoop**: framework used for distributed storage and computing
  - A tool that manages commodity clusters
  - A collection of technologies
    - HDFS: Hadoop Distributed File System: distributed storage
    - Yarn: schedule jobs/task over HDFS storage
    - MapReduce: programming model that simplifies parallel/distributed computing
    - Spark: built for real-time, in memory processing of data

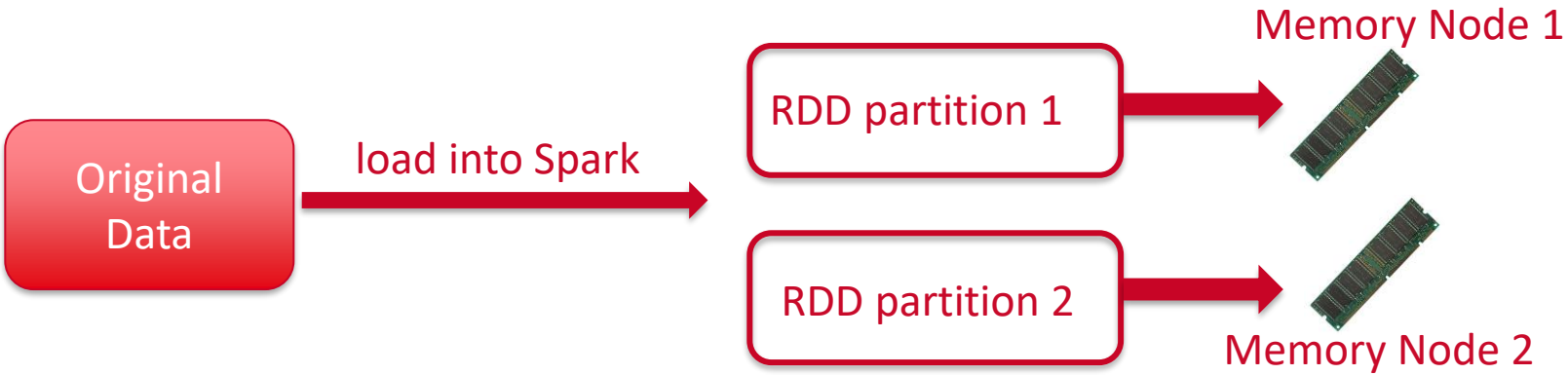


MapReduce

- Things to remember: a computational model to perform distributed computations
- Consists of three main steps: Mapping, Sort & Shuffling, and Reduce operations
  - The programmer only defines the mapping and reduce steps
    - Sort & shuffling is done behind the scenes
    - Literally, two functions: Map() and Reduce()
      - Map = computation to be executed in each block
        - **Uses key-value pairs**
      - Reduce = summary/grouping of the outputs produced by the nodes
        - **Uses key-value pairs**

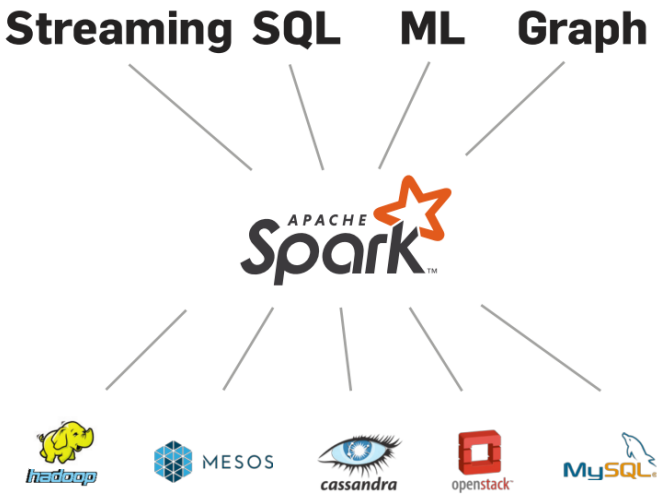
Spark

- The current “big thing” in data analytics
  - Allows for distributed computation
    - More flexible than MapReduce
    - In-memory processing



Spark

- In practice, it is very common to use Spark libraries, which are built on top of transformations/actions, in conjunction with programming languages such as R, Python, and Java

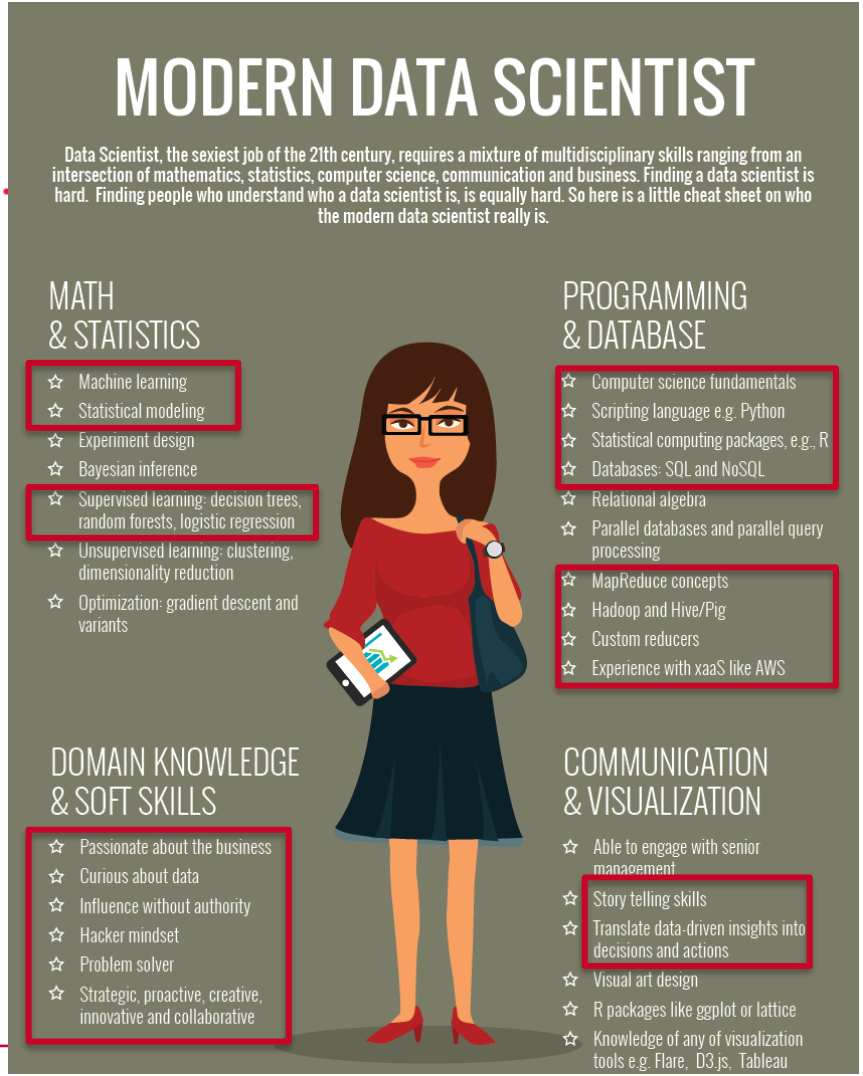


Spark

- Similar to MapReduce, one can only see the true power of Spark when working with big data sets and many nodes
- **Q3**: what are the reasons for the current trend where many companies are replacing MapReduce with Spark?
  a) Unlike MapReduce, Spark is free
  b) Spark tends to be 10x to 100x faster than MapReduce due to being an in-memory technology
  c) Spark does not require the underlying code to process data as key-value pairs
  d) Besides HDFS, Spark works well with other distributed-storage technologies

Final Considerations

- Analytics is not equal to statistics or computer science
  - Understanding technologies and business processes also matter
  - Things we experienced in this course
    - Make sure you share your experience with recruiters
    - Yes, there is a lot more to learn
  - Remember: the perfect data scientist is like a unicorn
    - The most important skill: **be passionate about learning**
      - Hacker mindset



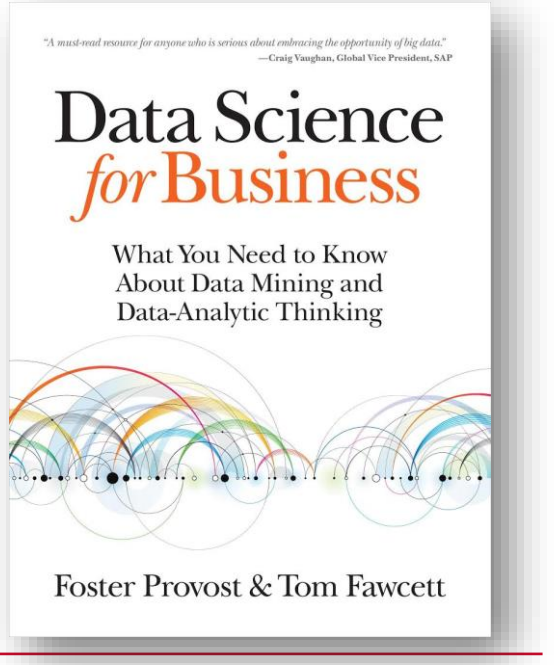
Final Considerations

- Always be familiar with the newest technologies and trends

Final Considerations

- Winter project idea:
  - Read the following book from cover to cover

- Easy to follow: no mathematics
- Several examples of how to apply data science to real-life business problems
- Authors are very famous analytics researchers
- Book I would use if I was teaching MBA-level analytics courses

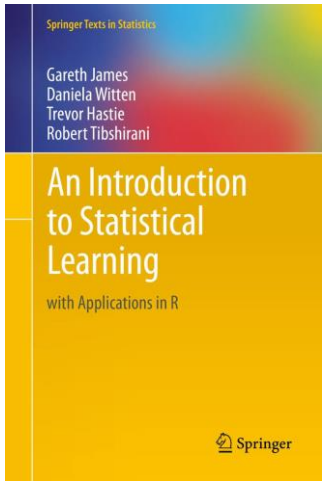


Final Considerations

- Winter project idea:
  - Read the following book from cover to cover
    - Free (and legal) pdf copy available at:

https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf

- Easy to follow: not a lot of mathematics
- Several examples (in R)
- Authors are very famous statisticians
- Book I would use if I was teaching ISA 491



Final Considerations