What is Hadoop?

1. Replicate each piece of information, send pieces to thousands of computers in a cluster, each computer would run the process on its file, send it back.
    a. Results then sorted and redistributed to another process
        i. First process called a "map" or "mapper", second called a "reduce process"
    b. Scale linearly – Many servers, twice performance, handle twice amount of data
2. Doug Cutting: Working on clone/copy of Google's Big data architecture – Hadoop
3. Bottom of Data Science: algebra, linear algebra, programming and databases
4. Now we apply machine learning
    a. Instead of taking a sample from large data sets to test hypotheses, we look for patterns
    b. Not hypotheses testing, rather generating hypotheses
5. Decision Sciences = probability, statistics and mathematics
    a. Deep learning just added
    b. Neural networks around 20-30 years

How Big Data is Driving Digital Transformation

1. Support needed from CEO, CIO and emerging role of Chief Data Officer

Data Science Skills and Big Data

1. My definition of big data is data that is large enough and has enough volume and velocity that you cannot handle it with traditional database systems
2. Started when Google tried to figure out how to solve their page rank algorithm

Establishing Data Mining Goals

1. Identify key questions that need answers
2. Identify costs and benefits of the exercise
3. Determine expected level of accuracy and usefulness
    a. High levels of accuracy cost more
    b. Cost trade-off for desired level of accuracy are important considerations

Selecting Data

1. Data-mining exercise depends upon quality of data being used
    a. Large databases: customer purchases and demographics
    b. Not readily available: surveys may be needed
2. Type, size and frequency of collection have a direct bearing on the cost
3. Identifying the right kind of data is crucial

Processing Data

1. Raw data often messy, erroneous or irrelevant
2. With relevant data, information can be missing
3. In the pre-processing stage you identify irrelevant attributes of data and expunge
4. Identifying erroneous aspects of data set and flagging them is necessary

   a. Human error may lead to inadvertent merging or incorrect parsing of information between columns
   b. Subject data to checks to ensure integrity
5. Develop formal method of dealing with missing data and whether they are missing randomly or inadvertently
   a. Randomly: simple set of solutions would suffice
   b. Systemic way: determine impact of missing data on results
6. Must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it

Transforming Data

1. Once relevant attributes are retained, determine the appropriate format in which the data must be stored
2. Aim: Reduce the number of attributes needed to explain phenomena
   a. Data reduction algorithms (Principal Component Analysis)
   b. Variables may need to be transformed to help explain phenomenon being studied (e.g. all types of income revenues – rentals, salary, etc)
3. Transform continuous variables into categorical variables. This could help capture non-linearities in the underlying behaviour

Storing Data

1. Stored in a format that is conducive for data mining
   a. Unrestricted and immediate read/write privileges to the data scientist
2. During mining, new variables are created, then written back to the original database, this is why data storage schemes should facilitate efficiently reading from and writing to the database
3. Store data on serves or media that keeps data secure and prevents data mining algorithm from unnecessarily searching for pieces of data scattered of different serves or storage media
   a. Safety and privacy should be a prime concern

Mining Data

1. Once processed, transformed and stored it is subjected to data mining
2. This covers data analysis methods
   a. Parametric and non-parametric methods
   b. Machine-learning algorithms
3. Good start is data visualization
   a. Multidimensional views of data using advanced graphing capabilities of data mining software = preliminary understanding of trends hidden in data sets

Evaluating Mining Results

1. Once mined, you do a formal evaluation of the results
2. Includes:
   a. Testing predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data

      i. Known as “in-sample forecast”
   b. Results shared with key stakeholders for feedback
      i. Then incorporated in later iterations to improve process
3. Data mining, evaluating results becomes an iterative process analysts can use better and improved algorithms in light of feedback from key stakeholders.

Deep Learning and Machine Learning

1. Big data: Massive, quickly built, varied – not formed with a traditional database
   a. Described in the 5 V’s
   b. Data Mining: process of automatically searching and analysing data, discovering previously unrevealed patterns.
      i. Pre-processing data to prepare it, transforming it into appropriate format
      ii. Insights and patterns are mined and extracted using various tools and techniques (data visualization, machine learning, statistical models)
2. Machine Learning: Subset of AI that uses computer algorithms to analyze data and make intelligent decisions without being explicitly programmed
   a. Trained with large sets of data they learn from examples
   b. Do not follow rules-based algorithms
3. Deep learning: Subset of machine learning uses layered neural networks to simulate human decision-making
   a. Label and categorize info
   b. Enables AI systems to learn on the job – improve quality and determine whether decisions were correct
   c. Artificial neural networks = neural networks
      i. Small computing units called neurons that take data and learn to make decisions over time
      ii. Layer-deep: become more efficient as data sets increase in volume
      iii. Other machine learning algorithms plateau as data increases
4. Data Science: Process and method for extracting knowledge from insights from large volumes of disparate data
   a. Multi-disciplinary: mathematics, statistical analysis, data visualization and more
   b. Appropriate info, see patterns, find meaning from large volumes of data and drive business
   c. Can use AI techniques – learning algorithms and deep learning models
   d. Broad term that encompasses the entire data processing methodology – while AI includes everything that allows computers to learn to solve problems and make decisions

Neural Networks and Deep Learning

1. Computer Sciences: attempt to mimic neurons
2. Neural networks: mimic how our brins use nerons to process things
    - Neurons and synapses – build complex networks that can be trained
    - Start with inputs and outputs to see what kind of outputs. Done repeatedly in a way this network should converge
    - Computationally very intensive
3. Deep Learning: 4/5 years ago
    - Neural networks on steroids
    - Multiple neural networks using lots of computing power
    - Needs matrix and linear algebra calculations
    - Speech, people, faces, images, classifying images
    - GPU: Graphics processing unit = 600 cores of processing cores
    - Speech recognition
    - Doesn't have to be taught
    - Learn Linear AEGRBRA

Applications of Machine Learning

1. Predictive analytics: Area of machine learning
    a. Recommender systems, cluster analysis, market basket analysis
    b. Decision trees, Bayesian Analysis, naïve Bayes
    c. E.G. Don't have to understand how they used or how to do them but must understand what their meanings are
2. Recommendations: Recommending based on your previous decisions
    a. E.g. Investment ideas that are similar
    b. Similar asset, company, technique, etc.
    c. Fraud detection: Machine learning problem – look at previous transactions, build a model that looks at each charge that comes through

How Data Science is Saving Lives

1. Targeted info to give best treatment to patients
    a. Data mining, data modelling, statistics and machine learning
    b. Factors for a disease
        i. Gene markers, associated conditions and environmental factors
        ii. Recommends tests, trials and treatments
2. Natural disasters
    a. Warick university used social media to track development of floods, hurricanes and weather events
    b. Weather stations

How should Companies Get Started In Data Science?

1. First thing a company must do is to start capturing data
    a. Costs, labour, material, products, revenue
    b. Capture it, archive it, do not overwrite on your old data – data never gets old
2. Then apply algorithms and apply algorithms
3. Data science inside a company is only as valuable as the data collected
4. Put together a team of data scientists

Application of Data Science

1. 2011 McKinsey & Company: data science key basis of competition. New waves of productivity, growth and innovation
2. 2013: UPS new route guidance system

The Final Deliverable

1. Ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy.
    a. Analytics summarize findings in tables and plots
    b. Data scientists then uses narrative to communicate findings
        i. Academia: Form of essays and reports (1000-7000 words)
        ii. Consulting and business: small documents (1500) with tables and plots, or comprehensive document
2. Discussed scope of the final deliverable
    a. Deliberated the key message of the report
    b. Looked for the data and analytics

The Report Structure

1. Cover page, table of contents, executive summary, detailed contents, acknowledgments, references and appendices (if needed)
2. Cover Page
    a. Title, names of authors, their affiliations, contacts, name of the institutional publisher (if any), date of publication.
3. Executive summary
    a. Abstract (executive summary), introductionary section, review of available relevant research on subject matter,
    b. Methodology
        i. Introduce research methods and data sources
        ii. New data? Explain the data collection exercise in some detail
    c. Results
        i. Present empirical findings

            1. Descriptive statistics
            2. Illustrative graphs
            3. Regression models or categorical analysis
            4. Empirical techniques that fall under data mining
                a. Mostly they rely on illustrative graphics
    d. Discussion
        i. Craft main arguments
        ii. Rely on narrative to communicate thesis
        iii. Refer to the research question and knowledge gaps
        iv. Highlight findings and missing piece to the puzzle
    e. Conclusion
        i. Generalize specific findings – take on a marketing approach
        ii. Identify future developments in research and applications
    f. References
4. Publication
    a. Have you told readers, at the outset, what they might gain by reading your paper?
    b. Have you made the aim of your work clear?
    c. Have you explained the significance of your contribution?
    d. Have you set your work in the appropriate context by giving sufficient background (including a complete set of relevant references) to your work?
    e. Have you addressed the question of practicality and usefulness?
    f. Have you identified future developments that might result from your work?
    g. Have you structured your paper in a clear and logical fashion?