POOJA PREMNATH
CSE-B

Unit 1 - Introduction to Data Science

- **Data Science**: involves methods to analyze massive amounts of data and extract the knowledge that it contains

- **Big Data**: a blanket term for any set of data so large and complex that it becomes difficult to process them using traditional RDBMS systems.

Characteristics of big data:
(i) Volume - how much data there is
(ii) Variety - how diverse the different types of data are
(iii) Velocity - how fast new data is generated
(iv) Veracity - how accurate the data is

- Differences between a data scientist and a statician - A data scientist has the ability to work with big data, and is experienced in machine learning, computing and algorithm building, may use tools like Hadoop, R, Spark, Python, Java.

Benefits and uses of data science and big data

1. Used by companies: to gain insights on their customers, products, can be used for targeted ads. eg. Google AdSense
2. by HR professionals: use people analytics, text mining to screen candidates monitor the mood of employees, study informal networks among coworkers.
3. by Financial institutions: to predict stocks, to predict the risk of pending money, and how to attract new clients
4. By govt. orgs & NGOs
5. By universities: for research, and to enhance the study experience of students
6. Massive Open Online courses MOOCs can study how this type of learning can complement traditional classes

Facets of Data

(1) Structured data: • resides within fixed fields in a record
• in the form of Excel files or SQL tables
• Hierarchical data is also structured but it is harder to store in a traditional RDBMS

• Data isn't always structured, it is imposed upon by humans and machines.

(2) Unstructured data: • doesn't fit within a data model, as it is content specific
• for eg. emails. There are some structured elements like the sender, title etc, the content itself is highly unstructured.

(3) Natural Language: • a special type of unstructured data
• requires knowledge of special data science techniques and linguistics for it to be processed.
• NLP used for entity recognition, summarization, text completion sentiment analysis.
• Models trained in one domain do not translate well to another model

(4) Machine Generated data: • info that is automatically generated by a computer, w/o human intervention.
• analysis of machine generated data done with highly scalable tools, due to its large volume.
• eg. call records, web server records, telemetry

(5) Graph based or network data

- A graph is a mathematical structure used to model pair-wise relationships between objects.
- It focuses on the relationship or adjacency of objects.
- Graph structures use nodes, edges and properties to represent and store graphical data.
- Graph based data is used to represent social networks, and its structure allows for the calculation of specific metrics like the influence of a person and the shortest path between people.
- Graph databases store graph-based data and are queried with specialized query languages like SPARQL.

(6) Audio, image and video

- are datatypes that are difficult for computers to interpret
- processing of audio, image and video can be used for in-game analytics / developing an algorithm to play video games

(7) Streaming Data

- can be like any other form of data, except for the fact that data flows into the system as and when an event happens instead of being loaded batchwise.
- Although the type of data isn't that different; it has to be treated differently.
- For eg. the trends on Twitter / the stock market / live sporting events

The Big Data Ecosystem

Components of the big data ecosystem and the tools used.

① Distributed File Systems

- similar to a normal file system, except that it runs on multiple servers at the same time.
- can store, read and delete files, just like normal file system.

Distributed file systems have several advantages

(i) can store files larger than any one computer disk

(ii) Files are automatically replicated across multiple servers for redundancy or parallel operations

(iii) can be easily scaled.

Example: HDFS – Hadoop File System

Vertical and Horizontal Scaling : vertical scaling: moving data to another server with more memory

horizontal scaling : add additional smaller servers

② Distributed Programming Framework

- must be used to exploit the data stored in the distributed file syster
- Data is not brought to the program, but rather the program is moved to the data.

③ Data Integration Framework

- to add data, and move data from one source to another
- eg. Apache Sqoop and Apache Flume excel.

④ Machine Learning Frameworks

- to extract insights from the data using ML models
- for eg. neural networks are learning algorithms that mimic the human brain in learning mechanics and complexity. They are advanced and black box, can use PyBrain
- Natural Language Toolkit (NLTK) : a library to work with NLP

Other libraries = Pylearn 2 & TensorFlow

(5) NoSQL Databases

software required to manage huge volumes of data.

No stands for 'Not only', NoSQL databases have implemented version of SQL, that solve several of the problems of traditional databases.

NoSQL databases allow for the virtually endless growth of data.

Types of databases

i) Column databases: data stored in columns, allows algorithms to perform faster queries.

ii) Document stores: each observation is stored in a documents, allows for a much more flexible data scheme.

iii) Streaming data: data is collected in real time

(iv) Key-value stores: data is not stored in a table, a value is assigned for every key. Easy to scale, but most of the implementation has to be done by the developer

(v) SQL on Hadoop: Batch-wise queries in Hadoop in an SQL-like language.

vi) New SQL: combines the scalability of NOSQL databases with the advantages of relational databases.

(6) Scheduling Tools: automate repetitive tasks and trigger jobs based on an event such as adding a new file to a folder.

(7) Benchmarking Tools: optimize big data installation by providing standardized profiling suites. A profiling suite is taken from a representative set of big data jobs.

(8) System deployment: helps to set up a big data infrastructure, largely automate the installation and configuration of big data components

(9) ~~Suite~~ Service Programmings: to allow others to use one's application. service tools expose big data applications to other applications through a service, foreg. REST service, REST= representational state transfer, used to feed websites with data.

(10) Security : Big data security tools allow one to have a centra
fine grained control over access to data.

The Data Science Process

1. setting the research goal
2. retrieving data
3. data preparation
4. data exploration
5. data modelling
6. presentation and automation

(1) Setting the research goal

A project charter must be prepared. It should contain:

(i) what one is going to research and the goal
(ii) how the company would benefit from it
(iii) how the analysis would be carried out
(iv) the data and resources required
(v) a timetable
(vi) deliverables

(2) Retrieving data

- To obtain data from within the company / from third party organizations
- The data can be in the form of text files or tables in a database

Data within the company : - data may be stored in official repositories like databases, data marts, data warehouses and data lakes.

(i) Database : primary goal is to store data
(ii) Data warehouse : reads and analyzes that data
(iii) Data mart : subset of data warehouse, serves a specific business unit.
(iv) Data lakes : contains data in its natural or raw format, unlike data warehouses and data marts, that have preprocessed data.

- Finding data within the company may also be difficult, as it may be scattered.

Getting access is also difficult - there are policies in place which prevent everyone from having access to all data.

- These policies translate into physical and digital barriers called Chinese walls.

Data from outside the company

- Data can be obtained from 3rd party companies like Nielsen and GfK.
- Data is also released by governments for public access.

(3) **Data Preparation** - consists of 3 phases, data cleansing, data integration and data transformation

- (i) **Data cleansing** - removing false values and inconsistencies from data sources.

Types of errors: (a) **Interpretation error** - value of the data is taken for granted, eg. saying that a person's age > 300yrs.

(b) **Inconsistencies**: where the data is different from the standardized values. eg. putting 'Female' in one table and 'F' in another, though they represent the same thing.

Errors pointing to false values in a data set

Error	Possible solution
1. mistakes during data entry	manual overrule
2. Redundant white space.	use string functions
3. Impossible values	manual overrule
4. Missing values	remove value
5. Outliers	validate, and if erroneous, treat as a missing value.

Errors pointing to inconsistencies between data sets

Error	Possible solution
1. Deviations from a code book	match on keys / use manual overrules
2. Different units of measurement	recalculate
3. Different levels of aggregation	bring to same level of measurement by aggregation or extrapolation.

Techniques to handle missing data.

Technique	Advantage	Disadvantage
1. Omit the value	easy to perform	lose info. from observation
2. Set value to null	easy to perform	some models cannot handle null values
3. Impute a static value like 0 or the mean	easy to perform, info not lost	can lead to false estimations from a model
4. Impute a value from an estimated or theoretical distribution	does not disturb the model too much	harder to execute, data assumptions are made
5. Modelling the value (non dependent)	does not disturb the model too much	harder to execute, data assumptions are made, can lead to too much confidence in the model, can artificially raise dependence among the variables.

Why data cleansing is needed:

Decision makers may make costly mistakes based on incorrect data.

(ii) If errors are not found early on, cleansing will have to be done for every project that uses the data.

(iii) Data errors may point to a business process that doesn't work as designed.

(iv) may cause / point to defective equipment, such as broken transmission lines and defective sensors.

(v) Data errors can point to bugs in software.

Data Integration: can be done in 2 ways: joining or appending/ stacking

① Joining: enriching an observation from one table with data from another table.

- To join tables, variables that represent the same object in both tables are used. These tables are called fields.
- When a key uniquely describes a record, called a primary key.

② Appending tables: adding observations from one ~~obs~~ table to another table.

- equivalent to a union in set theory, and SQL.

* Usage of views: A view is a virtual table, that can combine different tables, but physically does not exist.

Advantage - lesser storage space

Disadvantage - a table join is done only once, a view join has to be recreated every time the table is queried ⇒ more processing power.

Note: Data enrichment can also be done by adding calculated information to the table.

Data Transformation : ensures that the data is in a suitable to be used in one's models.

- It can be done by transforming the input variables, by

(a) reducing the number of variables

(b) turning variables into dummies - they can take only 2 values, true(1) or false(0). Separate ~~classes~~ columns are made for the classes stored in one variable, and it is indicated by 1 if the class is present and zero otherwise.

(4) Data Exploration - To understand how variables interact with one another, the distribution of data, and checking if there are outliers

- It uses descriptive statistics, visual techniques and simple modelling.
- called EDA - exploratory data analysis.

EDA Techniques

(i) Brushing and linking - combine, link different graphs and tables, so changes in one graph are automatically visible in another graph.

(ii) Visual representation - (a) histograms - a variable is cut into discrete categories and the number of occurrences in each category is summed up and shown in the graph.

(ii) Boxplot : shows the distribution between categories, like the maximum, minimum, median values.

(5) Model Building : Model building involves the following steps

(i) Selection of a modelling technique - select a technique from the field of statistics, machine learning, operations research etc.

(ii) Execution of the model - Building a model is an iterative process that involves selecting the variables for the model, executing the model

~~(iii) Model diagnostic & comparison~~ selection (contd).

- if the model has to be moved to a production environment
- if the model is difficult to maintain

Model execution (contd):

Model fit: use of the R-squared or adjusted R-squared technique, which is a measure of the amount of variation in the data that is captured by the model.

(iii) Model Diagnostics and comparison

(a) use holdout samples: they are a part of the data that is left out of the model building, so that it can be used to evaluate the model afterwards, works on the principle that the model should work on unseen data.

Other checks: mean square error: check for every prediction how far it was from the truth, square the error, and sum it up

(6) Presentation and automation

(i) in the form of presentations and research reports

(ii) automate the execution of the process so that the insights gained can be used in another project or enable an operational process to use the outcome of that model.