Data science languages

- **SQL**
  - SQL = Structured Query Language
  - SQL databases : MySQL, IBM Db2, PostgreSQL, Apache OpenOffice Base, SQLite, Oracle, MariaDB, Microsoft SQL Server
  - RDBMS Tools
    - IBM DB2, MySQL, Oracle Database, and PostgreSQL
  - NoSQL Tools
    - Redis, MongoDB, Cassandra, and Neo4J
  - Data warehouses Tools
    - Oracle Exadata, IBM Db2 Warehouse on Cloud, IBM Netezza Performance Server, and Amazon RedShift
- **Java**
  - Java applications are compiled to bytecode
  - run on the Java Virtual Machine, or "JVM."
  - data science tools built with Java include
    - Java-ML machine learning library
    - Apache MLlib makes machine learning scalable
    - Weka, for data mining
    - Deeplearning4j for deep learning
    - Apache Hadoop, Java-built application, manages data processing and storage for big data applications running in clustered systems
- **Scala**
  - a general-purpose programming language
  - support for functional programming and a strong static type system
  - runs on the JVM
  - Scala = "scalable language
  - data science = Apache Spark designed with scala
  - Spark is a fast and general-purpose cluster computing system
  - Provides APIs that make parallel jobs easy to write, and an optimized engine that supports general computation graphs
  - Spark includes
    - Shark, which is a query engine
    - MLlib, for machine learning
    - GraphX, for graph processing
    - Spark Streaming
- R
  - R tools for data visualization
    - Libraries: Ggplot, plotly, Lattice, Leaflet
    - Functions: Plot,

Data Science Processes

- **Data Science Methodology, How to think?**

From problem to approach:

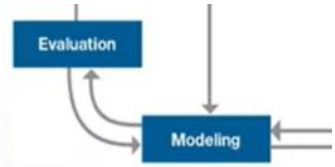
1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

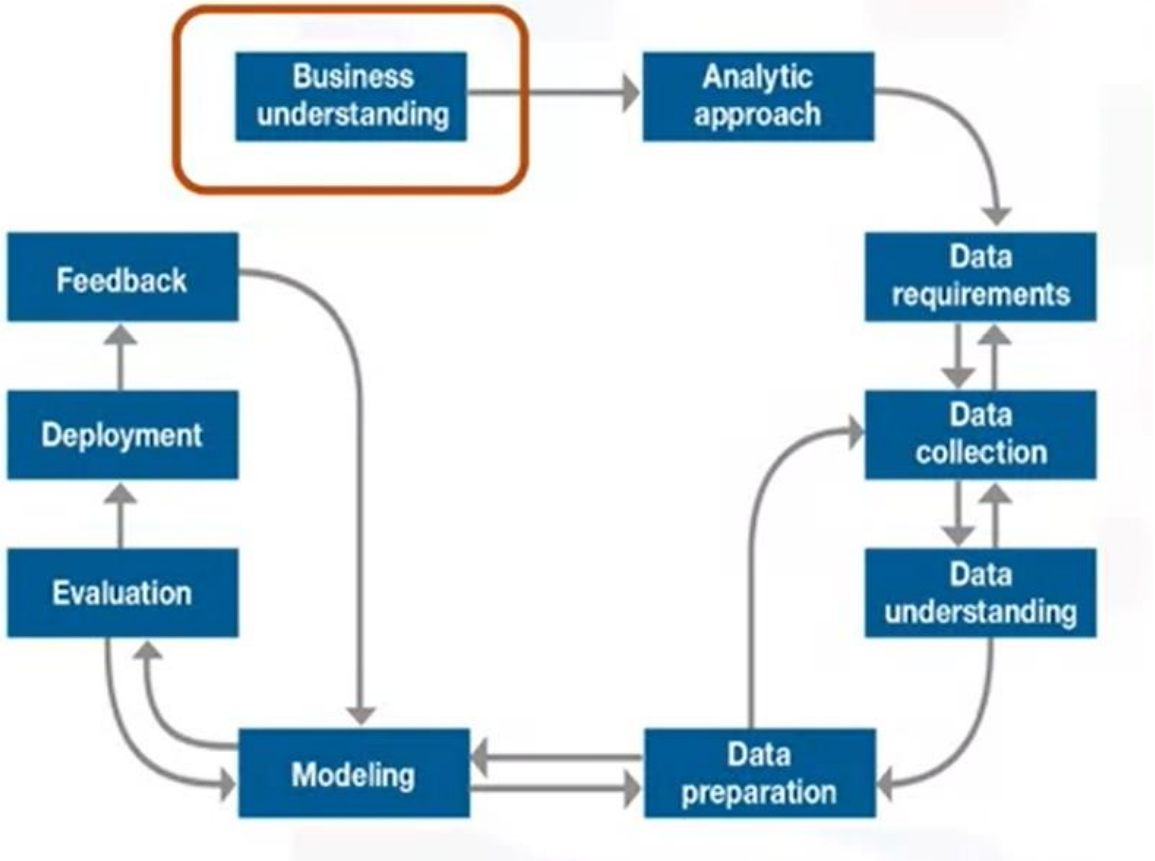
Working with the data:

3. *What data do you need to answer the question?*
4. *Where is the data coming from (identify all sources) and how will you get it?*
5. *Is the data that you collected representative of the problem to be solved?*
6. *What additional work is required to manipulate and work with the data?*

Deriving the answer:

7. *In what way can the data be visualized to get to the answer that is required?*
8. *Does the model used really answer the initial question or does it need to be adjusted?*
9. *Can you put the model into practice?*
10. *Can you get constructive feedback into answering the question?*





- **Action and Model relation**
  - Probability of Action: Predictive Model
  - Show Relation: Descriptive Approach
  - Binary Answer (Yes/No): Classification Model
  - Multiple Classification : Decision tree Classification Model
- **From Problem to Approach**
  - Business Understanding
    - clarify the goal

        - Why we are doing this analysis?
    - Analytical Approach
        - How we will do this analysis?
        - Which statistical and machine-learning technique method will be used?
- **Working with Data**
    - Data Requirements
        - What type of data can be used to address this findings?
    - Data Collection
        - From where we will collect the data?
    - Data Understanding
        - What are there in data?
        - Missing Values, Data Distribution on histogram, invalid values, False values, Related and not related parameters
    - Data Preparation
        - Consume 70-90% of time
        - Data must be cleaned and should be in the format that can be used directly
        - Replace Yes/No to 1/0 to apply mathematical computation
        - Replace missing values by average or other methods
        - Identify Wrong values and outliers
        - Remove duplicate data
- **Deriving the Answer**
    - Modelling
        - Built predictive or destructive models
        - Predictive model outcome is Binary Yes/No, Greater/Less
        - Descriptive model classify the outcome in multiple range and relation can be also defined like proportional / inversely proportional / exponential
    - Evaluation
        - Does the model used really answer the initial question or does it need to be adjusted?
        - ROC curve: Plot of true-positive rate against the false-positive rate for different values of the relative misclassification cost
        - Maximum distance points are best choice
    - Development
        - Involve stakeholders, users, IT persons to deploy model in reality and train all users to use it
        - Initially it can be used in small cluster to identify issues and areas of improvements before its deployment for all users
    - Feedback
        - Collect new data and check their outcomes with the existing model to improve the efficiency

        - From the feedback of stack holders and new data improve the performance of the model

- **Data process cycles in ML**
    - pre-processing of data
    - feature selection
    - feature extraction
    - train test splitting
    - defining the algorithms
    - fitting models
    - tuning parameters
    - prediction
    - evaluation
    - exporting the model
- **Data management tools**
    - relational databases : MySQL and PostgreSQL
    - NoSQL databases : MongoDB Apache, CouchDB, Apache Cassandra
    - File-based tools : Hadoop File System
    - Cloud File systems : Ceph
    - Text data: Elasticsearch
    - Commercial tools:
        - Oracle Database
        - Microsoft SQL Server
        - IBM Db2
    - Cloud based tools
        - Amazon DynamoDB (Data structure is Json)
        - Cloudant based on Apache couch db
        - IBM DB2
- **Data integration and transformation**
    - ETL: extract, transform, and load
    - ELT : Extract, Load, Transform
    - data refinery and cleansing
    - tools: Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, NodeRED
    - Commercial Tools:
        - Informatica Powercenter
        - IBM InfoSphere DataStage
    - Cloud based tools
        - Informatics
        - IBM data Refinery
- **data visualization tools**
    - Hue: create visualizations from SQL queries
    - Kibana : data exploration and visualization web application
    - Apache Superset : data exploration and visualization web application

    - Commercial Tools:
        - Tableau
        - Microsoft Power BI
        - IBM Cognos Analytics
    - Cloud based tools
        - Datameer
        - IBM data Refinery
        - IBM Cognos Analytics
- **Model deployment**
    - Apache Prediction IO: supports Apache Spark ML models for deployment
    - Seldon : supports nearly every framework TensorFlow, Apache SparkML, R, and scikit-learn
    - MLeap: support SparkML
    - TensorFlow can serve : TensorFlow service
    - TensorFlow Lite : Low computing resources
    - TensorFlow dot JS : browser based
    - Commercial Tools:
        - SPSS Modeler
        - SAS
    - Cloud based tools
        - IBM Watson Machine Learning
        - Google cloud AI platform training
- **Code asset management or version management or version control**
    - Git
    - GitHub
    - GitLab: fully open source platform
    - Bitbucket
- **Development environments**
    - Jupyter: interactive Python programming, now supports 100+ different programming languages
    - JupyterLab : next generation of Jupyter Notebooks, more modern and modular
    - Apache Zeppelin: doesn't require coding
    - Rstudio: statistics and data science
    - Spyder : Basics
- **Extension environments**
    - cluster-computing framework Apache Spark: used by Fortune 500 companies, linear scalability, batch (file) data processing engine
    - Apache Flink : Updates of spark, stream (live) data processing engine
    - Ray : large-scale deep learning model training
- **Fully integrated and visual tools**
    - Support data integration, transformation, data visualization, and model building
    - KNIME : drag-and-drop capabilities

- Orange : less flexible than KNIME, but easier to use
- commercial tool support majority of data tasks
  - IBM Watson studio
  - Whatson open scale
  - H2o.ai
- Cloud based tools
  - Microsoft Azure
  - IBM Watson studio
  - Whatson open scale
  - H2o.ai

- **Python libraries**
  - **Scientific computing libraries**
    - Pandas: cleaning, manipulation, and analysis, works with two-dimensional table consisting of columns and rows “DataFrame”
    - NumPy: based on arrays, enabling you to apply mathematical functions to these arrays. Pandas is actually built on top of NumPy
  - **Data visualization Libraries**
    - Matplotlib : graphs and plots
    - Seaborn: based on matplotlib, heat maps, time series, and violin plots
  - **Machine learning Libraries**
    - Scikit-learn: statistical modeling, regression, classification, clustering
  - **Deep learning Libraries**
    - Keras: standard deep learning model
    - TensorFlow: low-level framework used in large scale production of deep learning models
    - Pytorch: used for experimentation, making it simple for researchers to test their ideas
  - **General-purpose cluster-computing framework**
    - Apache Spark
      - Enables you to process data using compute clusters
      - Process data in parallel, using multiple computers simultaneously
      - Similar functionality as Pandas Numpy Scikit-learn Apache Spark
      - Python, R, Scala, SQL can be used
- **API: Application Programming Interfaces**
  - API is used to communicate with the other software
  - Software to software
  - Code to Software
  - **REST API**
    - RE: Representational
    - S: State
    - T: Transfer

        - REST API enable you to communicate using the internet
        - Advantages of storage, greater data access, artificial intelligence algorithms, and integration of many other resources
        - REST APIs user program = client
        - Client communicate with web services / resources
        - To send request from client, we use HTTP method containing JSON file
        - And services operate on JSON file and revert back the JSON file to client
- **Data sets**
    - Structured collection of data
    - as text, numbers, or media such as images, audio, or video files
    - Collection of rows and columns
    - **CSV: comma separated values**
        - text file where each line represents a row and data values are separated by a comma
    - **Hierarchical or network data structures**
        - Used to represent relationships between data
        - Hierarchical data is organized in a tree-like structure
        - Network data might be stored as a graph
            - Graph: Connections between people on a social networking
    - **raw data**
        - Images or audio.
        - MNIST dataset : images of handwritten digits and is commonly used to train image processing
- **Data resources**
    - - Data Portals (datacatalogs.org)
    - Data.gov
    - data.europa.eu
    - UNdata
    - Find Open Datasets and Machine Learning Projects | Kaggle
    - Dataset Search (google.com)

Machine learning models

- Supervised learning
    - Data is labelled with correct output
    - Model tries to identify relationships and dependencies between the input data and the correct output
    - Used to solve regression and classification problems
        - Regression/Estimation
            - Used to predict a numeric or real value
            - Geographic location, size, number of bedrooms, price
        - Classification
            - Prediction for something belongs to a category/class
            - emails as spam or not

- Unsupervised learning
  - Data is not labelled with correct output
  - Analyze the data and try to identify patterns and structure within the data based only on the characteristics of the data itself
  - Clustering
    - Used to divide each record of a data set into one of a small number of similar groups
    - Purchase recommendations based on past shopping behaviour and the content of a shopping basket
  - Anomaly
    - Anomaly detection identifies outliers in a data set
    - fraudulent credit card transactions
    - Suspicious online log-in attempts
  - Dimension reduction
    - Used to reduce the size of data
  - Density Estimation
  - Market basket analysis / Association technique
    - Used for finding items or events that often co-occur
    - Grocery items that are usually bought together
- Reinforcement learning
  - Similarly, the way human beings and other organisms learn
  - Reinforcement learning model learns the best set of actions to take, given its current environment, in order to get the most reward over time
  - Used in games such as go, chess, and popular strategy video games
- Sequence mining
  - Used for predicting the next event
  - Live Stream in websites
- Recommendation systems
  - Associates' people's preferences with others who have similar tastes
  - Recommends new items to them, such as books or movies.

Deep learning

- Emulate the way the human brain solves a wide range of problems
- Natural language, spoken and text, images, audio, and video, forecast time series data
- Requires very large data sets of labelled data to train a model, and compute-intensive
- Model zoo: contained pre trained deep learning models
- IBM's MXA : Model assets exchange to get deep learning pre trained models
- Models use : PyTorch, Keras, TensorFlow

Data Science Libraries

- NumPy
  - Math library to work with N-dimensional arrays in Python

- Working with arrays, dictionaries, functions, datatypes and working with images
- SciPy
  - Collection of numerical algorithms and domain specific toolboxes
  - Signal processing, optimization, statistics
  - Good library for scientific and high performance computation
- Matplotlib
  - Plotting package that provides 2D and 3D plotting
- Pandas
  - High-level Python library that provides high performance easy to use data structures
  - Used for data importing, manipulation and analysis
- SciKit Learn
  - Collection of algorithms and tools for machine learning
  - Free Machine Learning Library for the Python programming language
  - Includes Classification, regression and clustering algorithms

AI

- Artificial Intelligence, Augmented Intelligence
- Examples
  - Siri, google assistance,
  - Cortana, Alexa
  - Humanoid Robots, Driverless car
  - Chatbot, Voice based CC,
  - Voice to Speech conversation
  - Image or video based detection or identification
- AI is a system which mimic the human behaviour, intelligence and abilities
- AI is used to compute human tasks with more efficiency and accuracy
- AI is also used for time complex tasks to perform to save time of humans
- Types of AI
  - **Weak or Narrow AI or Applied AI**
    - Applied to a specific domain
    - Language translators, virtual assistants, self-driving cars, AI-powered web searches, recommendation engines, intelligent spam filters
    - Can perform specific tasks, but not learn new ones
    - making decisions based on programmed algorithms, and training data
  - **Strong AI or Generalized AI**
    - Can interact and operate a wide variety of independent and unrelated tasks
    - Can learn new tasks to solve new problems, and it does this by teaching itself new strategies
    - Combine many AI strategies that learn from experience and can perform at a human level of intelligence
  - **Super AI or Conscious AI**

        - Similar to human-level consciousness
        - Self-aware capabilities

Data Science Job posts

- Data Engineer
    - Convert row data into usable data
    - Place all toys from truck to inside the shop to be ready to sale
- Data Analyst
    - Use data to generate insights
    - Identify the top toys for each customer and explain it meaning fully
- Data Scientist
    - Design ML and AL models to use the data and to forecast future predictions
    - Oder toys need to be purchased for the next festivals
- BI Analyst
    - Focused for business application and development
    - Decides the price of the each toys based on demand supply ratio

Data Analysis

- Gather, clean, analyse and mine data
- Interpret results and report finding
- Find patterns and their correlation and based on those insights are derived and conclusion are made
- Types of Data Analytics techniques
    - Descriptive Analysis
        - What happened
    - Diagnostic Analysis
        - Why did it happen
    - Predictive Analysis
        - What will happen next
    - Prescriptive Analysis
        - What should be done after
- Process involved in any Data Analysis tasks
    - Understanding the problem and desired output
    - Define method of measurement of output
    - Gathering data
    - Decide tools
    - Cleaning data for accurate analysis
    - Analyse and mine data
    - Interpreting results
    - Presenting findings
- Data Analysis
    - Can be nonnumeric data
    - detailed examination of the elements or structure of something
    - Past event analysis

- Data Analytics
    - Must be numeric data
    - the systematic computational analysis of data or statistics
    - Future event analysis
- Skill set for Data analytics
    - Technical skills
        - Spreadsheets / Microsoft Excel / Google Sheets
        - Statistical analysis
        - Visualization tools
        - Programming languages such as R / Python / C++ / Java / MATLAB
        - SQL and NoSQL databases
        - Big Data processing tools Hadoop, Hive, and Spark
        - Analyse your data, Validate your analysis, Identification of errors
        - Problem understanding and solving skills
        - Manage the process, people, dependencies, and timelines
    - Soft skills
        - Collaboration in work
        - Communicate effectively to report and present your findings
        - Convincing story
        - Curious
        - New questions to surface and challenge your assumptions and hypotheses
        - Intuition for possible cause and result

Big Data

- Definition
    - Large and dynamic data
    - Created by machines, humans and tools
    - To derive insights for business, risk, performance, profit, management
    - Data is collected, stored and analytically processed
- V's of Big data
    - Value : Profit, social benefit, Improvement in current ongoing process
    - Variety : Data generated from various sources and types of data
    - Velocity : speed of data generation
    - Veracity : Quality of data, accurate, consistent
    - Volume : the size of the data
- Tools
    - Apache Spark
    - Hadoop

Version Control

- Tools: Git, Github, GitLab, BitBucket, and Beanstalk
- SSH protocol is a method for secure remote login from one computer to another
- Repository contains your project folders that are set up for version control
- Fork is a copy of a repository

- Pull request is the way you request that someone reviews and approves your changes before they become final
- Working directory contains the files and subdirectories on your computer that are associated with a Git repository
- Commands of git
    - Init: When starting out with a new repository
    - Add: moves changes from the working directory to the staging area
    - Status: allows you to see the state of your working directory and the staged snapshot of your changes
    - Commit: takes your staged snapshot of changes and commits them to the project
    - Reset: undoes changes that you've made to the files in your working directory
    - Log: enables you to browse previous changes to a project
    - Branch: lets you create an isolated environment within your repository to make changes
    - Checkout: lets you see and change existing branches
    - Merge: lets you put everything back together again
-