**Data Science Literacy Fundamentals**

Table of Contents

[Course 1: Data Science for Everyone 2](#_Toc91528986)

[Course 2: Machine Learning for Everyone 3](#_Toc91528987)

[Course 3: Data Visualization for Everyone 6](#_Toc91528988)

[Course 4: Data Engineering for Everyone 7](#_Toc91528989)

[Course 5: Cloud Computing for Everyone 8](#_Toc91528990)

# Course 1: Data Science for Everyone

* Data Science Workflow: Data Collection and Storage 🡪 Data Preparation 🡪 Exploration and Visualization 🡪 Experimentation and Prediction

Diagram

Description automatically generated

* Applications of Data Science: Traditional Machine Learning, Internet of Things (IoT), and Deep Learning
* Roles: Data Engineer, Data Analyst, Data Scientist, and Machine Learning Scientist
  + Data Engineer: Build data pipelines, Phase 1, SQL, Java, Python, Scala, Shell, Cloud Computing
  + Data Analyst: Simpler analysis, dashboards, and reports, Phase 2 and 3, Spreadsheets, SQL, BI tools (Tableau, Power BI, Looker), Python, R
  + Data Scientist: Stats, experiments and analysis for insights, traditional machine learning, Phase 2,3, and 4, SQL, Python, and R
  + Machine Learning Scientist: Predictions and Explorations, Deep Learning, Phase 2,3, 4 (mainly 4), Python and R (TensorFlow and Spark)

Graphical user interface, application

Description automatically generated

* Data Sources: Company Data (Web Events, Survey Data, Net Promoter Score [how likely to recommend]) and Open Data (APIs Application Programming Interface, Public Records)
* Data Types: Quantitative, Qualitative, Image, Text, Geospatial, Network data
* Data Storage: Unstructured (Document Database: Email, text, video, sound, web page, social media data) and Tabular (Relational Database)
* Retrieval: NoSQL for Document Database and SQL for Relational Database
* Data Pipeline: how data moves
* Step 1 Data Pipeline: ETL (Extract, Transform, Load); Automate pipeline
* Step 2 Tidy Data (row observation; column variable)
* Step 3 Data Exploration (EDA: Exploratory Data Analysis): Descriptive and Visualization, look for outliers
* Step 4: Experimentation and Prediction
* A/B Testing: Champion/Challenger Testing using stat inference
* A/B Testing Steps: Pick a metric to track, determine sample size, run experiment, check for significance
* Larger sample sizes allow us to detect smaller differences / effect sizes
* Predictive Modelling
* Time Series data
* Seasonality in time series data: cyclical occurrence of specific events
* Machine learning: prediction from data
* Supervised machine learning: predictions from data with labels and features
* Example of supervised machine learning: churning (loss) of costumers

Timeline

Description automatically generated

* Unsupervised machine learning: only features; example: clustering
* You usually must select number of clusters

# Course 2: Machine Learning for Everyone

* Artificial Intelligence (AI): set of tools to make computers intelligently
* Machine learning most prevalent subset of AI
* Machine learning: a set of tools to make inferences and predictions from data
* Use machine learning to: predict future events, infer causes of events and behaviors, and infer patterns
* Data science: making discoveries and creating insights from data
* Machine learning is an important tool for data science
* Machine learning mode: Input 🡪 Model 🡪 Output
* 3 types of machine learning: reinforcement learning, supervised learning, and unsupervised learning
* Machine Learning workflow:Diagram

  Description automatically generated
* Supervised machine learning classification: target has labels ; classify into category
* Use support vector machine (linear classifier vs polynomial classifier)
* Regression: continuous variable values prediction
* Clustering models : K-means (specify number of clusters) and DBSCAN (density-based spatial clustering of applications with noise) : specify what constitutes a cluster and it outputs the number of clusters with data in them

Diagram

Description automatically generated with medium confidence

* Step 4: Evaluate performance
* Problem: Overfitting: performs great on training data but poorly on testing data [doesn’t generalize well]
* Confusion matrix:

Table

Description automatically generated

* Accuracy = correctly classified (true positive+ true negative) / all cases
* Sensitivity: true positives / (true positive + false negatives)
* Specificity: true negatives / (true negatives + false positives)
* Difficult to evaluate unsupervised learning
* Improving performance options: dimensionality reduction, hyperparameter tuning, and ensemble methods
* Dimensionality reduction: get rid of irrelevant and group those with high correlation into one feature or keep one feature
* Hyperparameter tuning: change many parameters of model (SVM algorithm) such as kernel, C, degree, gamma, shrinking, coef0, tol, …
* Ensemble methods: use more than one model (used in classification [what majority of models say] and regression [mean of what models say])
* Deep learning: Neural network using nodes
* Neural networks figure out the relationship between neurons and figures out the neurons in the middle

Diagram

Description automatically generated

* Use deep learning when you have lots of data, access to processing power, lack of domain knowledge, and complex problems (computer vision and natural language processing)
* Image data: pixels (width and height) and RGB so width x height x 3
* Natural language processing: Bag of words (n-grams), translation, sentiment analysis, chatbots
* Limitations: data quality (garbage in, garbage out), explainability (black box)

# Course 3: Data Visualization for Everyone

* Histogram: single continuous variable (see distribution / frequency)
* Binwidth
* Modality (unimodal, bimodal, trimodal)
* Skewness [symmetricity]: left-skewed, symmetric, right-skewed
* Kurtosis [how many extreme values]: mesokurtic (normal distribution), leptokurtic (narrow peak and lots of extreme values), platykurtic (broad peak, few extreme values)
* Boxplot: continuous variable split by categorical variable
* Scatter plots: 2 continuous variables
* Add trend lines and smooth trend lines; consider logarithmic scaling
* Line plots: 2 continuous vars, and consecutive observations are somehow connected (time)
* Bar plots: categorical variables
* Vertical bars, stacking bars
* Box plot: spread VS bar: count
* Dot plot: categorical variable, display numeric score for category on log scale or multiple numeric scores for each category
* For higher dimensions (>2) use aesthetics: color, shape, size, transparency, line color, line type (most useful is color) or lots of panels
* 3 types of color scales: qualitative (use different hues to distinguish unordered categories), sequential (vary chroma or luminance to show ordering), or diverging (vary chroma or luminance with 2 hues to show above or below midpoint like neutral)
* When plotting many variables at once use pair plot (cat + cont)
* If both cont use correlation heatmap
* Use parallel coordinates plot if you have lots of continuous variables, you want to find patterns across these variables, or you want to visualize clusters of observations
* Polar coordinates or pie charts rarely use (unless for time or things that a circular) – rose plot
* Bar plot + polar coordinates = pie chart / Histogram + polar coordinates = rose plot
* Dual axes are misleading; better to use multiple panes
* Make sure 0 included in y axis starting point
* Remove chartjunk (pictures, skeuomorphism [reflections, 3D, shadow], extra dimensions, ostentatious colors or lines) to avoid sensory overload

# Course 4: Data Engineering for Everyone

* Data engineers focus on step 1 of data workflow
* Work with big data
* 5 Vs of big data: Volume (how much), variety (what kind), velocity (how frequent), veracity (how accurate), and value (how useful)
* Data scientist: steps 2 to 4
* Data pipeline ETL: Extract, Transform, and LoadDiagram

  Description automatically generated
* Data structures: structure data, semi-structured data, unstructured data
* SQL (Structured Query Language) is the industry standard for RDBMS (relational databases management systems)
* Data lake: stores raw data and stores all data structures
* Data warehouse: stores mainly structured data for specific use; type of database
* We have data catalogues for data lakes
* Data processing: converting from raw to meaningful data
* Scheduling data: manual scheduling, time scheduling, and sensor (event-triggered) scheduling
* Batches (group records at intervals) vs streams (sends records right away)
* Parallel computing: split task into subtasks and run them over different computers
* Cloud computing: servers on cloud ( for file storage, computation, and databases) ; can have multicloud to avoid relying on one vendor

# Course 5: Cloud Computing for Everyone

* Cloud computing: delivery of technology services (including compute, storage, databases, networking, software, and many more) over the internet with pay as you go
* Cloud computing characteristics: virtualization, scalability, cost, speed, performance, growth, reliability, and security
* Vertical scaling: one instance with more power / horizontal scaling: more than one instance
* Cloud service models: On premise, IaaS (Infrastructure as a service), PaaS (platform as a service), SaaS (software as a service)

A screenshot of a computer

Description automatically generated with low confidence

Table

Description automatically generated

A picture containing text, businesscard

Description automatically generated

A picture containing text, businesscard

Description automatically generated

Graphical user interface, text

Description automatically generated

* FaaS: function as a service
* Cloud deployment models: Private, public, and hybrid
* Other deployment models: multicloud, community
* Example of regulation on the cloud: GDPR
* Personal data
* New cloud roles: cloud architect, cloud engineer, DevOps engineer, and security engineer
* Cloud providers examples: Amazon AWS, Microsoft Azure, Google Cloud , Alibaba, IBM, Oracle