Introduction

- Quick introduction from the lecturers
- Who are we?
- What are the expectations from this course?
- Which skills would you like to gain at the end of this course?

MAT386E - Computational Data Science

Topics	Weeks
Data Science: Big data and project management	1
Machine learning: Linear Regression, Polinomial Regression	2
Machine learning: Classification	3
Machine learning: Clustering	4
Deep learning: Basic deep learning method and applications	5
Text Processing: What is NLP? NLP Techniques, Current Technologies and Applications	6
Big Data Platforms: Architecture, Tools and frameworks	7
Data Storage	8
Data extraction, transformation and loading (ETL)	9
Batch Data Processing	10
Streaming Data Processing	11
Spark ML: Variable selection and transformation, regression, clustering model evaluation	12
Data visualization & Scoring	13
Project Presentations	14

MAT386E, Computational Data Science - Grading

- 4 Homeworks: %30
- 1 Term Project: %30
- 1 Final Exam: %40
- Attendance: Not Required

Garanti BBVA Technology Coordinators:

Sena Tezcan – senatezcan96@gmail.com Musa Bayır – bayirmusa45@gmail.com

Textbook: Géron A., Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, Inc, 2017.

Garanti BBVA Technology

ITU Computational Data Science

Week 1

Sep'23

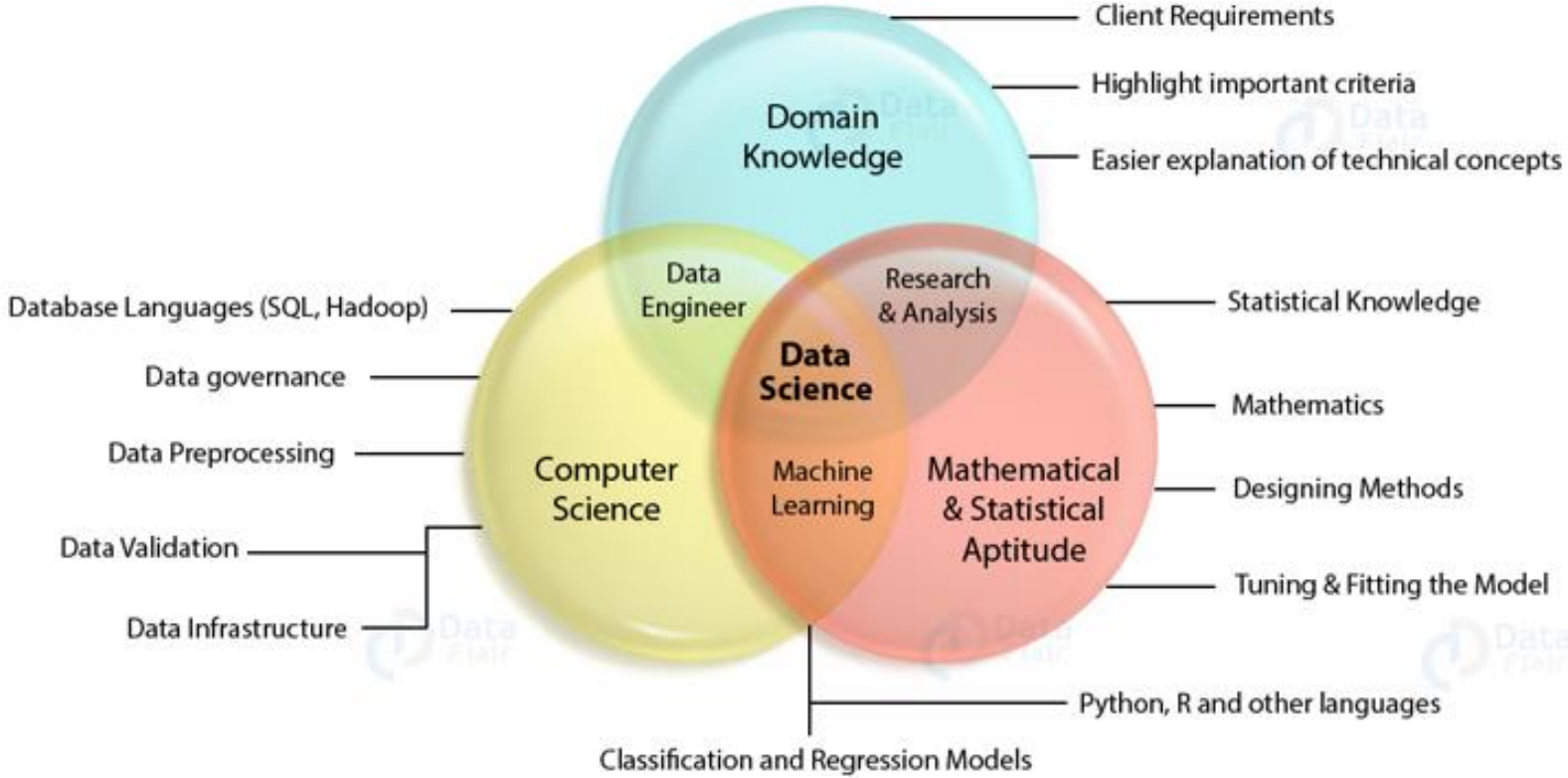
Big Data & Advanced Analytics Unit

Providing advanced analytical solutions for business problems by using all available data sources

- Sales & Marketing Analytics
- Digital & Customer Analytics
- Process Analytics
- Natural Language, Image & Speech Processing
- Big Data Technologies

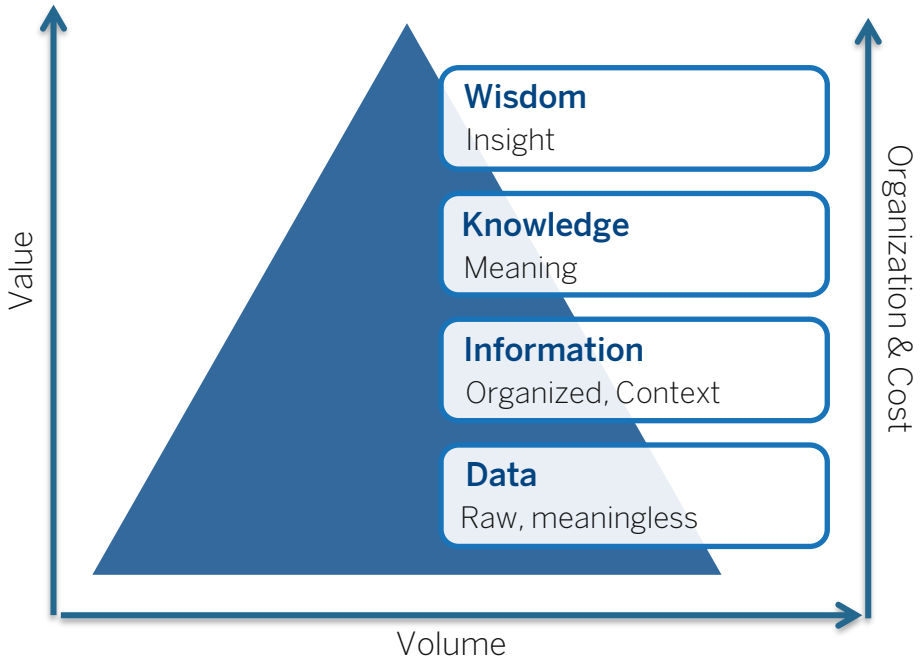
Educational Backgrounds

What is Data Science?



Data Scientist: Creating Knowledge & Wisdom

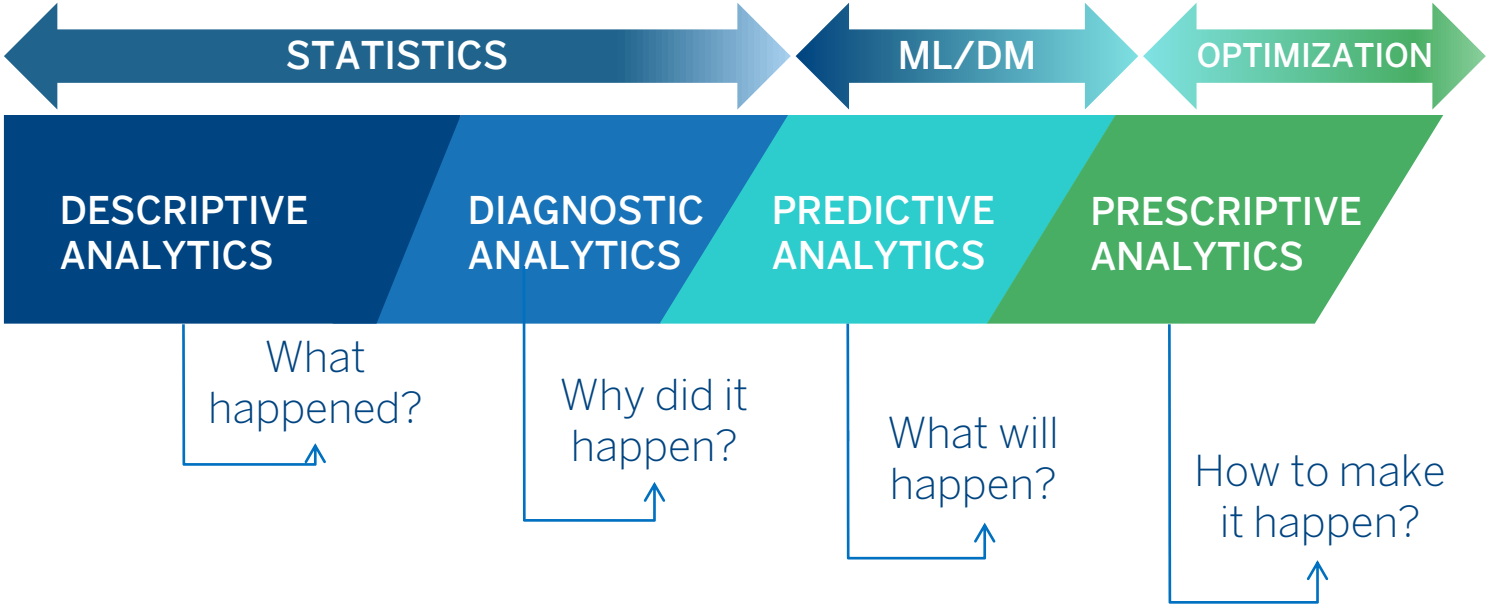
DIKW Pyramid:



Each step up,

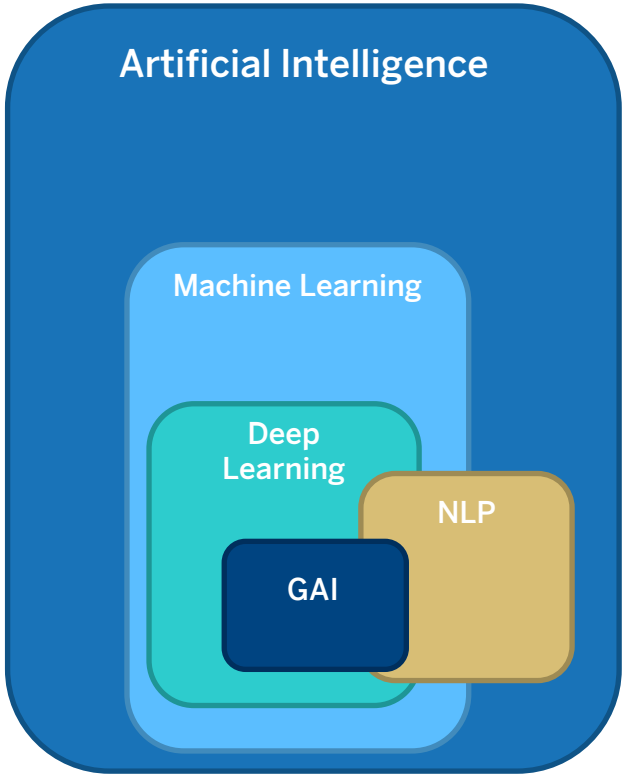
- The pyramid adds value to the initial data
- Enrich data with context & meaning
- Answers questions
- Guide to make better decisions

Statistics, ML and Deep Learning



- 1- Show the reality captured via statistics
- 2- Generalize sample conclusions to the entire population and study the relationships between variables and compare hypotheses
- 3- Determine future data via historical data
- 4- Recommend the suitable action and its consequences

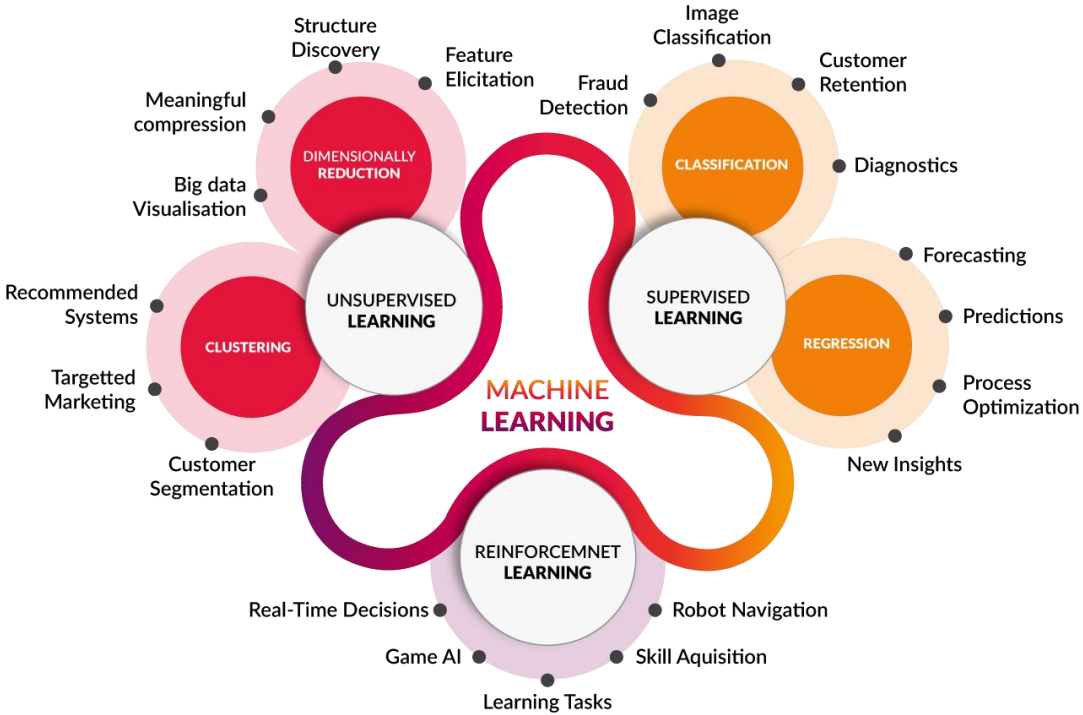
AI vs Machine Learning (ML) vs Deep Learning (DL)



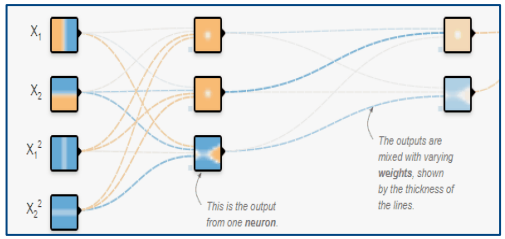
- **AI** involves machines that can perform tasks that are characteristic of human intelligence» (John McCarthy)
  - General AI
  - Narrow AI
- **Machine learning** is a way of achieving AI. «The ability to learn without being explicitly programmed» (Arthur Samuel)
  - Supervised
  - Unsupervised
  - Reinforcement
- Machine learning uses algorithms to parse data, learn from that data, and make informed decisions based on what it has learned
- **Deep learning** is one of many approaches to machine learning
- Deep learning structures algorithms in layers to create an "artificial neural network" that can learn and make intelligent decisions on its own

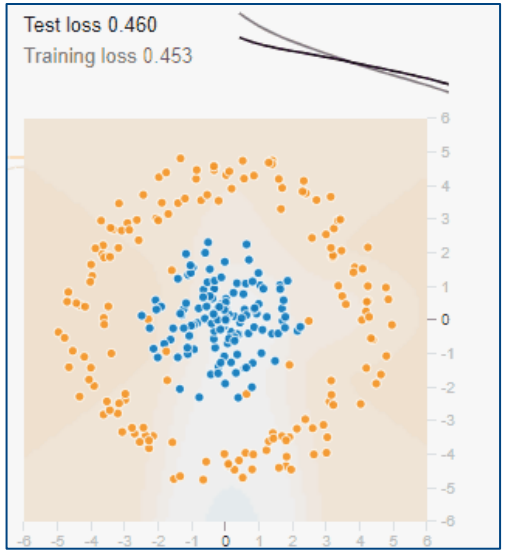
Source: https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Overview of ML Algorithms



Deep Learning Theory (DL)





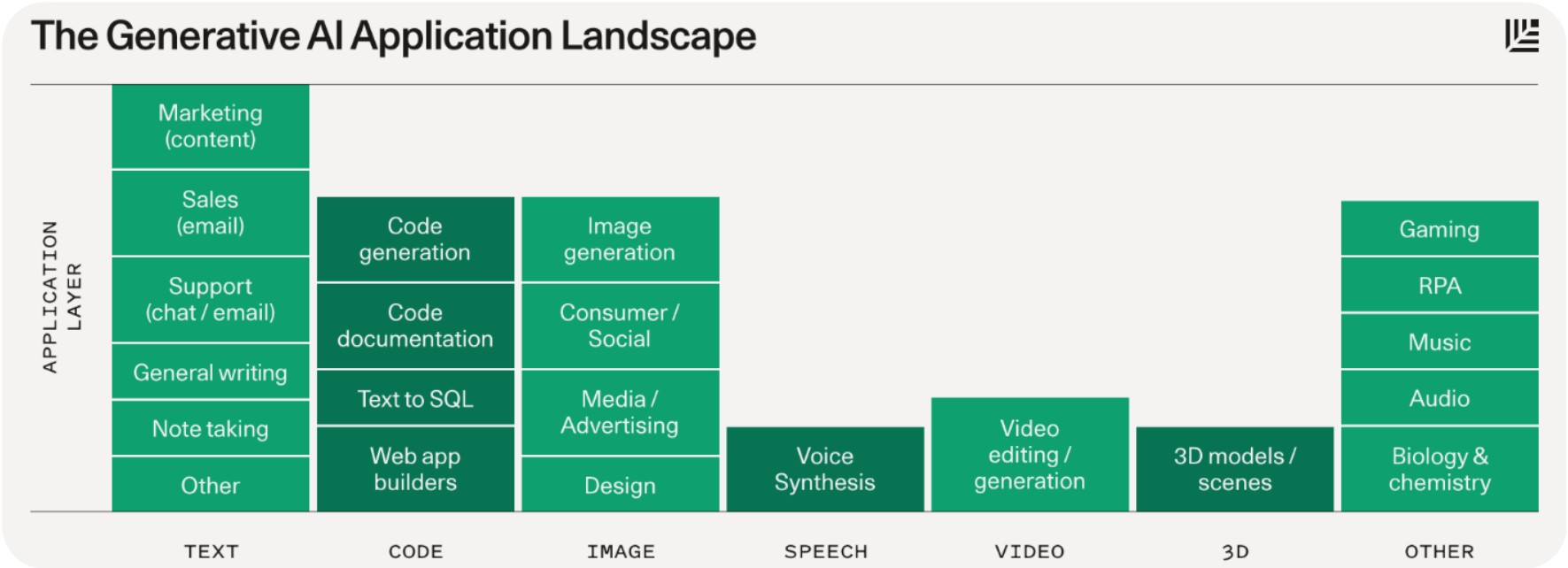
- A subset of machine learning
- Inspired by the biology of human brain to make its own intelligent decisions
- A tool for achieving Artificial Intelligence fed by large amounts of data

Business Value:

- Natural language processing (text to speech, speech to text, language translation)
- Voice processing (virtual assistants - AI robot interact with customers and gathering data about their behavior)
- Text processing (chat-bots or service-bots providing customer service on online banking)
- Image processing (face recognition, eye detection)
- Anomaly detection
- Recommendation engines

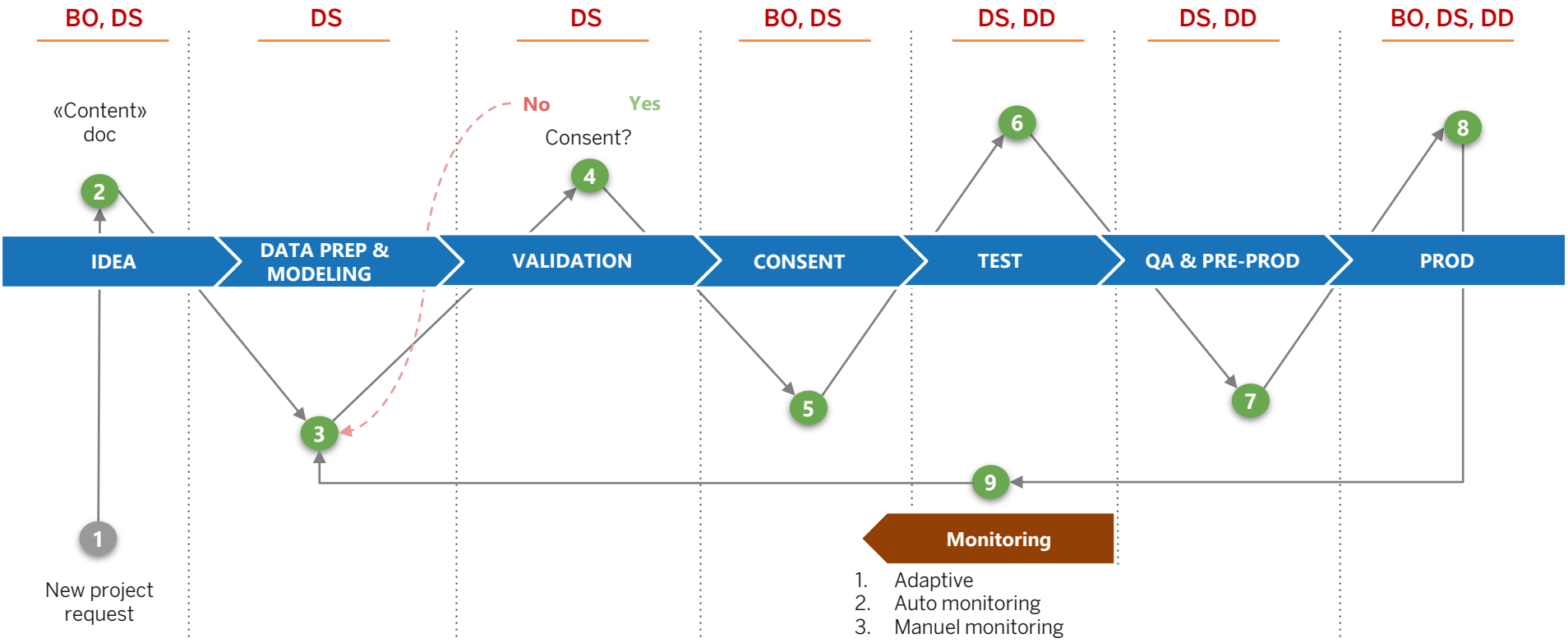
Generative AI

Application Landscape



Analytical Modeling Life Cycle & Roles

BO: Business Owner
DS: Data Scientist
DD: Data Developer



DS Modelling Platforms

SQL

- Oracle
  - Toad
  - Sql Developer
- Microsoft-SQL
- MySQL
- Postgre-SQL

Open-Sourced Products

- Python
  - Spyder
  - Pycharm
  - JupyterNotebooks
  - JupyterLab
- R
  - R Studio
- Spark
  - Scala
  - PySpark
  - Java

Licensed Products

- SAS – Enterprise Guide
- SAS – Enterprise Miner
- SPSS
- Oracle Data Miner
- KNIME

Python Libraries

Core Libraries & Statistics

- Numpy
- Scipy
- Pandas
- StatsModels

Machine Learning

- Scikit-learn
- Xgboost/LightGBM/CatBoost
- Eli5

Natural Language Processing

- NLTK
- SpaCy
- Gensim

Visualization

- Matploblib
- Seaborn
- Plotly
- Bokeh
- Pydot

Deep Learning

- Tensowflow
- Pytorch
- Keras

Distributed Deep Learning

- Dist-keras
- elephas
- Spark-deep-learning

Data Scraping

- Scrapy

Data Governance

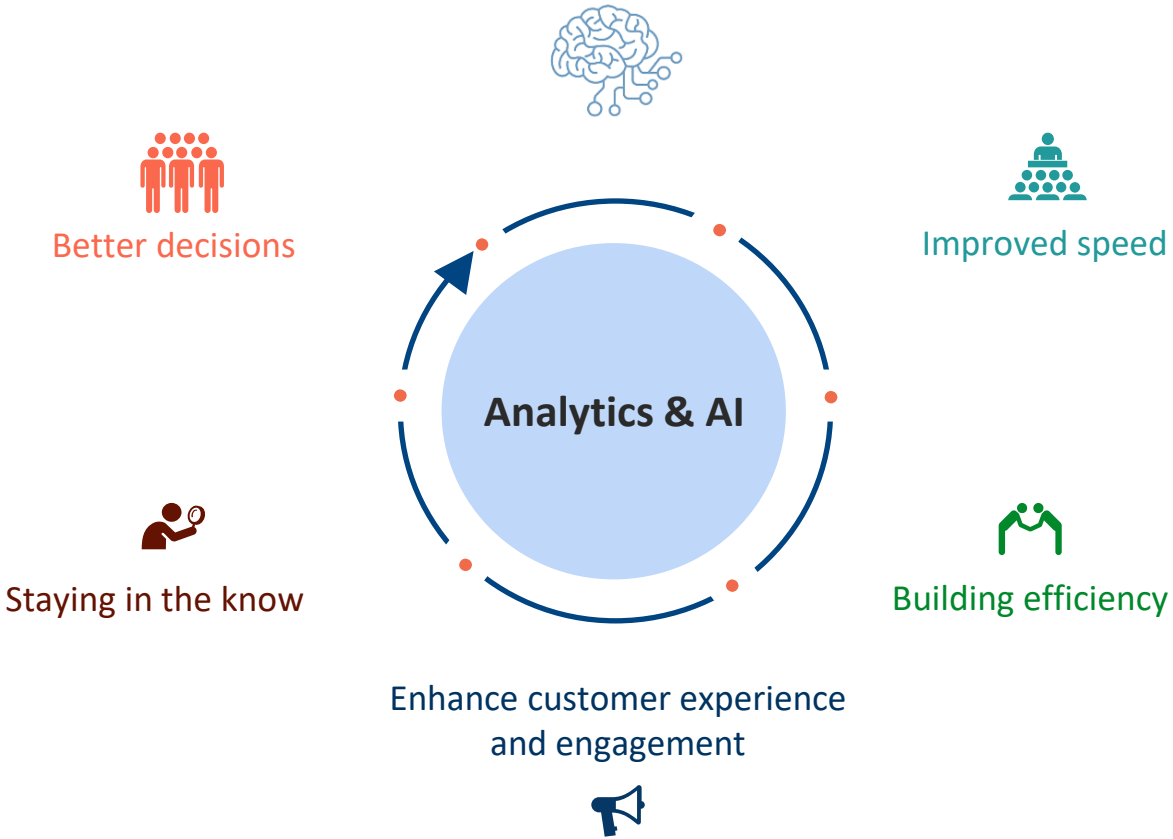
- Data Governance is the set of policies, processes, rules, roles and responsibilities of data.
- Being a data-driven company requires not only understanding or processing data but also managing them.
- Data Governance allows strategic decisions to be based on complete, high quality and reliable data.

Transparent | **Reliable** | **Traceable** | **Accessible**

Components of Data Governance

Data Content	Creating the content of the data with all functional (definition, owner, security, etc.) and technical (location, type, etc.) information
Data Quality	Ensuring that the data has the required quality, with the help of data quality tools.
Data Traceability	Documentation and monitoring of all processes from the origination of data to annihilation of data.
Data Accessibility	Ensuring the accessibility of data by all relevant teams in line with customer privacy and security / authorization procedures

What Business Goals Are Organizations Pursuing With AI



Source: https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Technology Skills Enable Use Case Imagination

Search	Sentiment	Detecting Credit Card Fraud	Spam Filtering for Email
Writing Recognition	Speech Understanding	Stock Analysis	Structural Health Monitoring
Syntactic Pattern Recognition	Topic Spotting	Weather Prediction	Face Detection
Finance – Derivatives Trading	Game Playing	Software as a Service	Customer Segmentation
Machine Translation	Medical Diagnosis	Mood Analysis	Brain Machine Interface
Optical Character Recognition	Recommen-dation Systems	Robot Locomotion	Advertising - Targeting
Bioinformatics	Automatic Word Completion	Classifying DNA Sequences	Computer Vision

Retail	▪ Segmentation, marketing, customer experience
Automotive	▪ Computer vision in cabin and roadway
Logistics	▪ Automated packing, transportation, planning, safety improvement
Healthcare	▪ Lifestyle management, drug discovery, resource planning, patient experience
Medical	▪ Diagnoses, pharmaceutical optimization, computer vision
Oil and Gas	▪ Increase production and to optimize costs

Source: https://www.zendesk.com/blog/machine-learning-and-deep-learning/

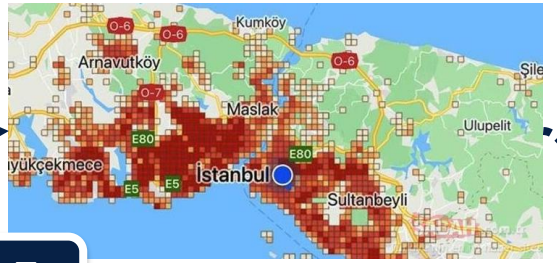
Use of Analytics in Different Business Areas

1 Can we trust?

2 How much s/he earns?

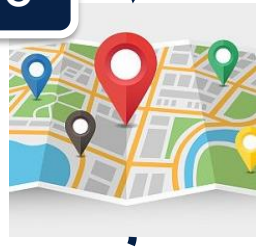
3 May be Fraud?

4 Know your customer



5 Harvest your data !

Where to locate? 6



7 How much money we need?

8 Whom to serve first?

Use of Analytics in Different Business Areas

9 Why s/he calls?

10 Lets chat!

11 Automate & speed up..

12 What to sell?

13 Channels to serve?

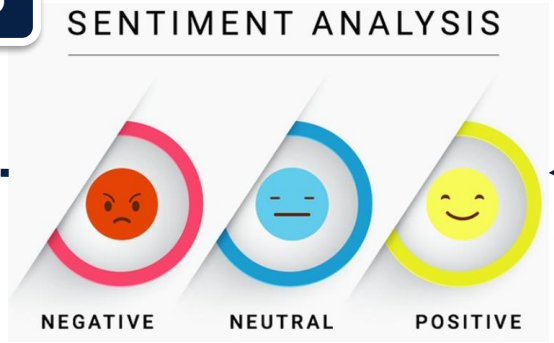
14 Is it OK to sell?

15 Don't go ☹

Use of Analytics in Different Business Areas

17 Know beforehand..

16 What they talk about?



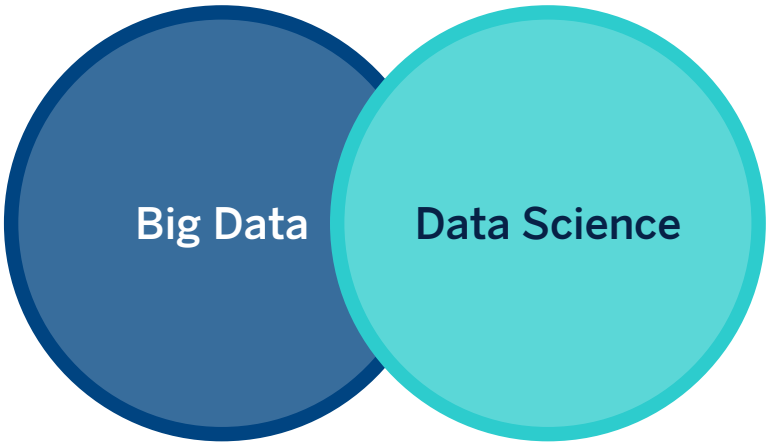
18 Make smart offers..

19 Farewell ☹

Data Science & Big Data

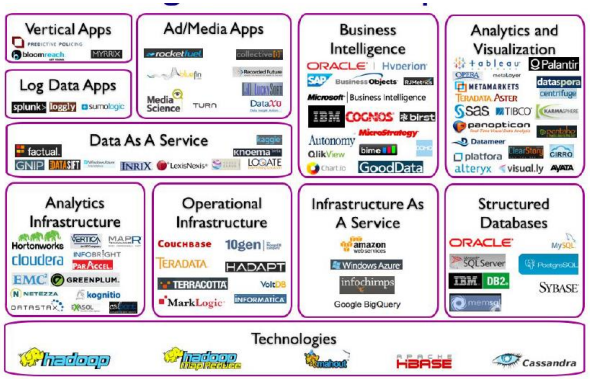
Are Big Data and Data Science the same thing?

- I wouldn't say so...
- Data Science can be done on small data sets.
- And not everything done using Big Data would necessarily be called Data Science.
- But there certainly is a substantial overlap!



Big Data Components

Technological Comparison 2012 vs 2017



~704

Online Educational Platforms

https://www.coursera.org/

https://www.kaggle.com/

https://www.datacamp.com/

https://www.edx.org/

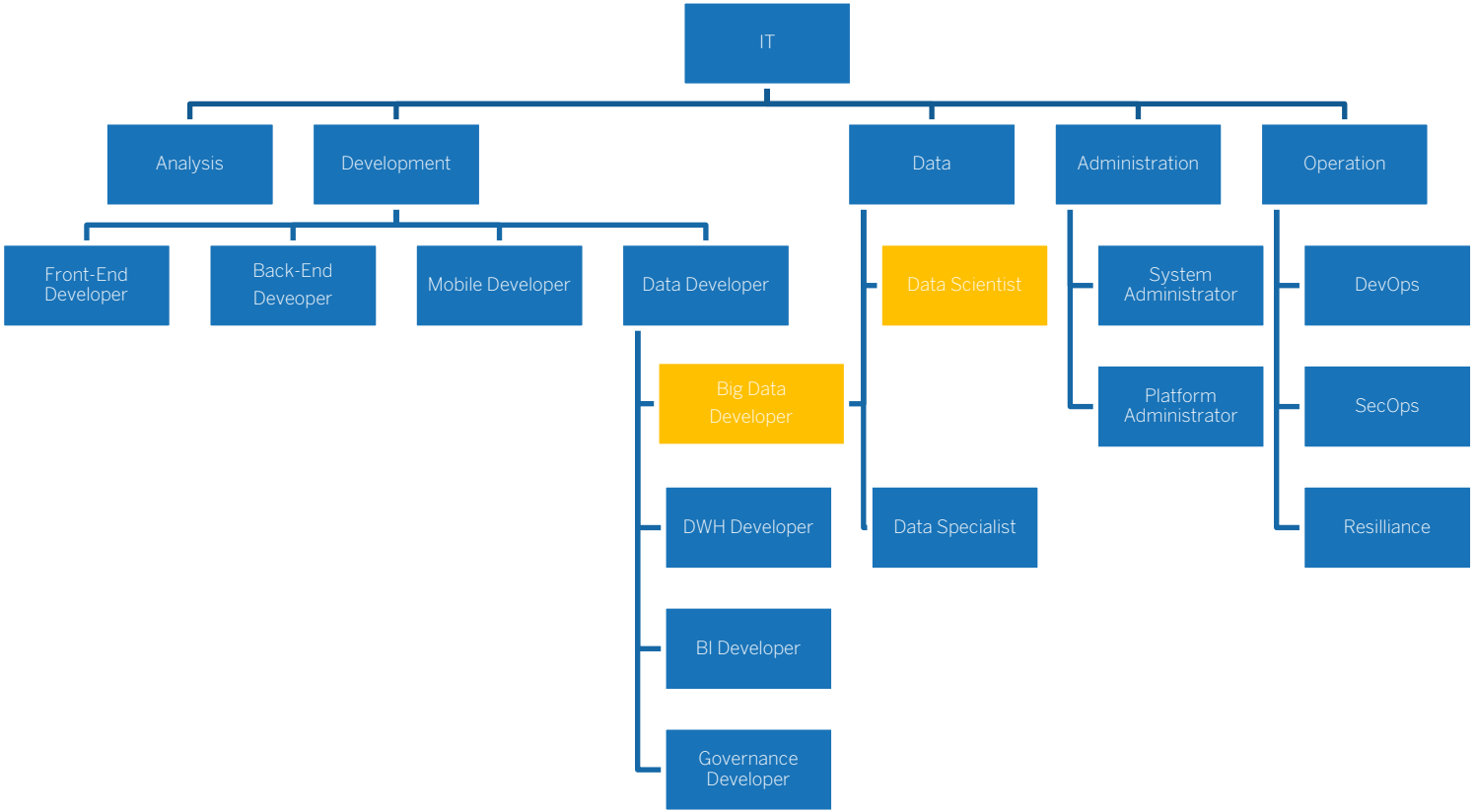
https://www.udemy.com/

https://colab.research.google.com/

https://machinelearningmastery.com/

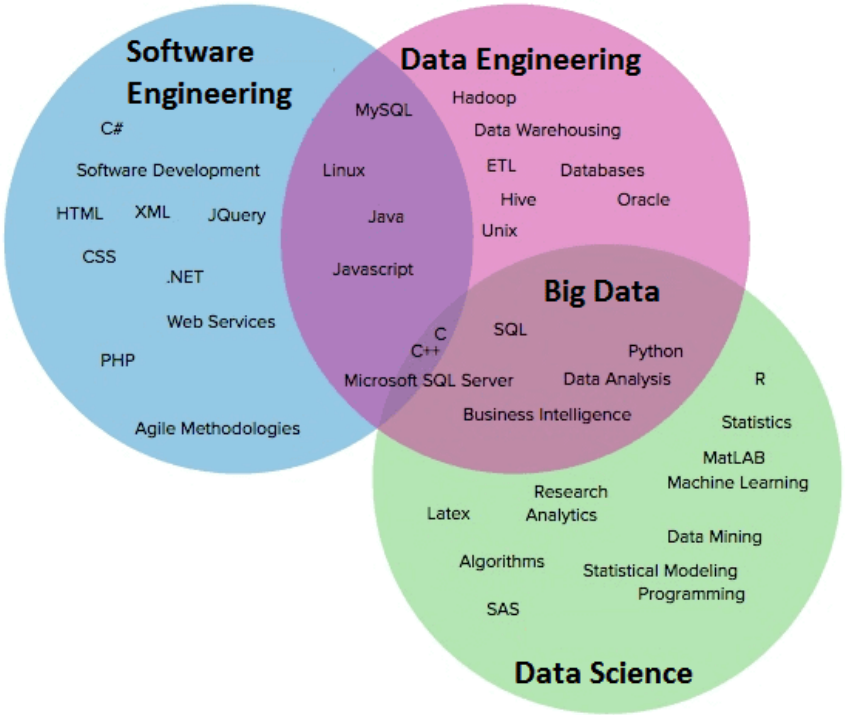
https://online.stanford.edu/courses/cs221-artificial-intelligence-principles-and-techniques

Data Engineer? Data Developer? Big Data Engineer?



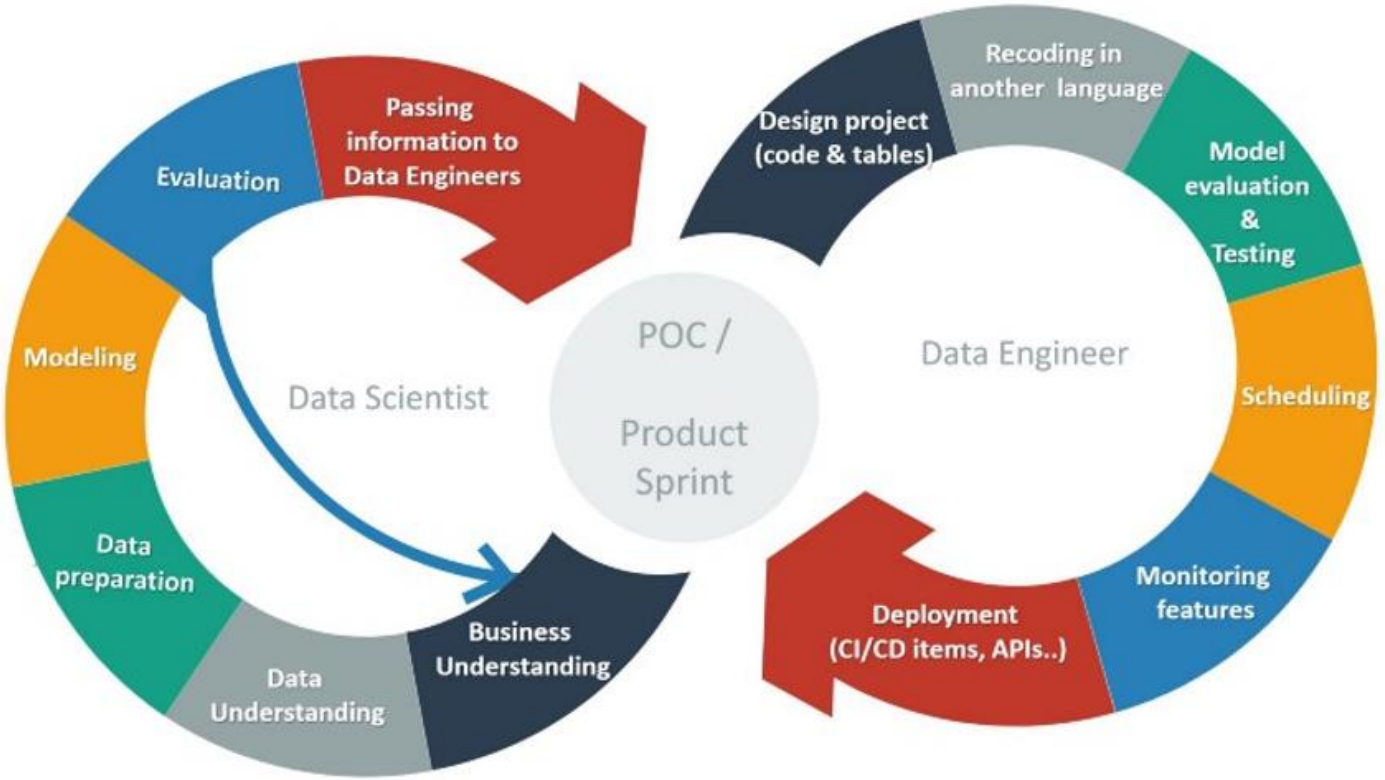
Data Engineer vs Software Engineer

Similar Skills, Different Professions



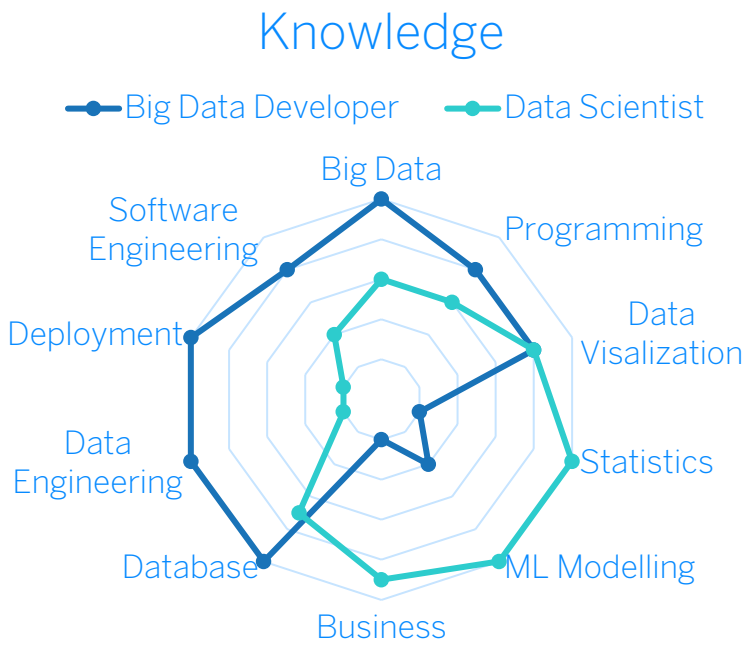
Data Engineer vs Data Scientist

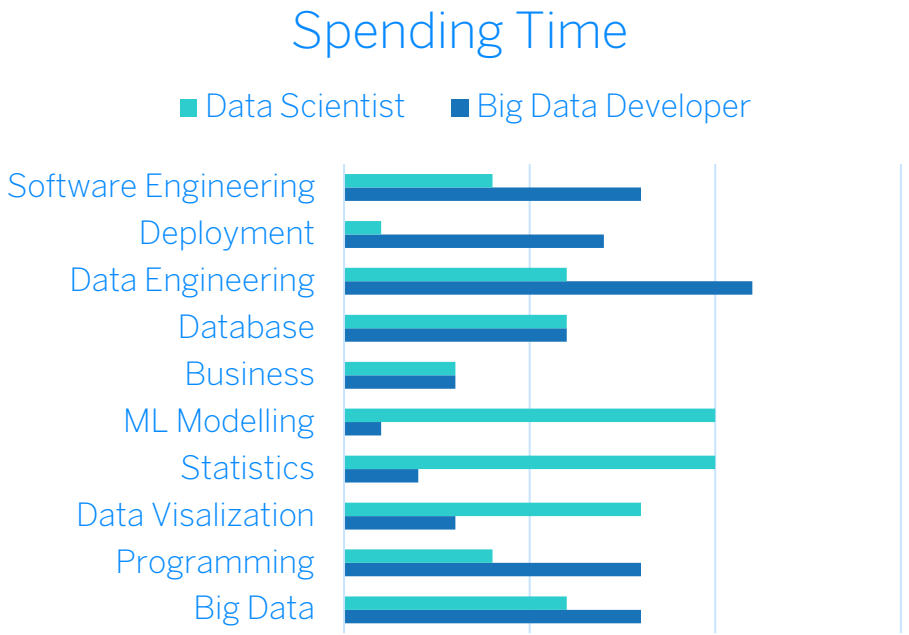
Two pieces jigsaw



Big Data Engineer vs Data Scientist

Similar Skills, Different Professions





How to Start Your Career in Data Engineering

Before beginning

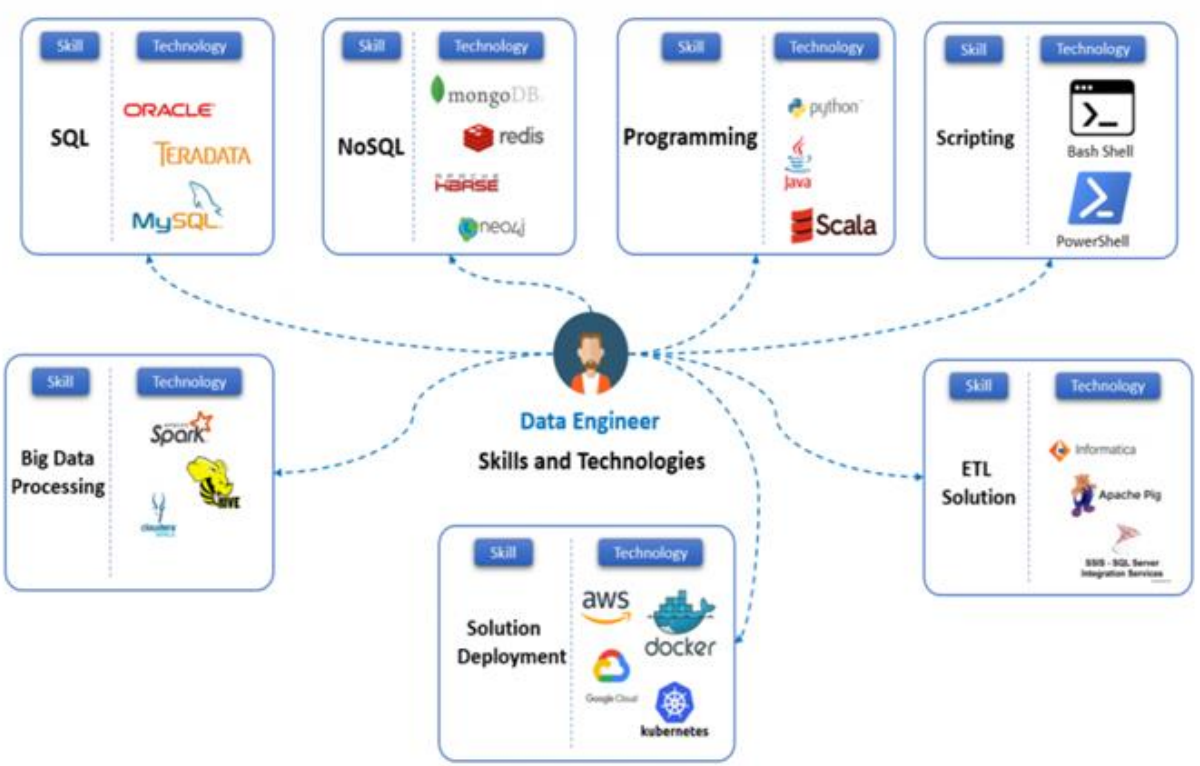
- Start this because you want to be a data engineer, not because it's popular
- Know yourself
- Do not need to know much, know that how to reach content you needed
- Do not stuck on which programming language is perfect
- Avoid over engineering

After beginning

- Learn storage platforms
- Learn how to ingest and transform data
- Learn distributed systems and cloud
- Learn automation and scripting
- Keep yourself up to date

An Ordinary Day for Big Data Engineers

- Architecture knowhow usage in wide perspective
- Design
- Data manipulations
- Coding in different languages
- Maintenance & monitoring
- Integration & impact analysis
- Reverse engineering



Sample Big Data Ecosystem