ICDS Innovation Center for Data Science

Data Science RoadMap

Masoud Mazloom

20-07-2022

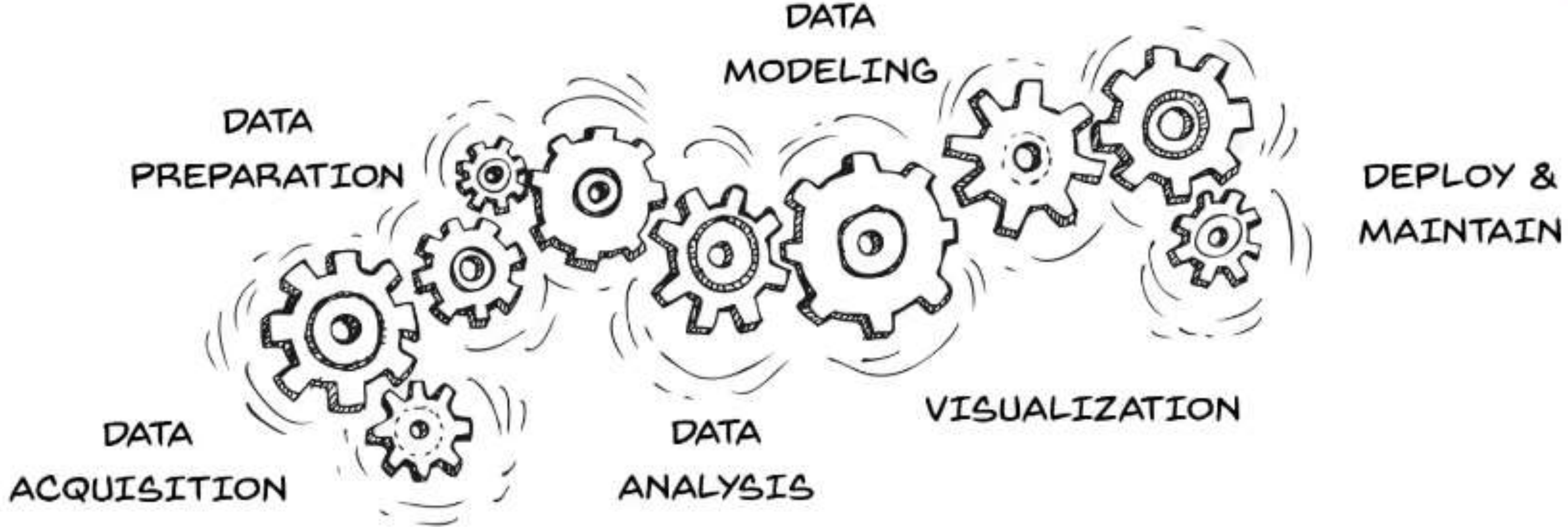
Data is the currency of the digital economy
Get the analytical skills you need to cash in

Although data is the lifeblood of the digital economy, many companies are blind to the value of the data they create. It’s time for that to change.

- Interdisciplinary field that focuses on **analyzing massive amounts of data** to **automatically identify inherent patterns**, **extract underlying models**, and **make relevant predictions**.

- Impacting virtually all areas of the economy, including **science, engineering, medicine, banking, finance, sports and the arts**.

- Exciting real-world applications include credit card fraud detection, speech recognition, predictive medical diagnosis, and self-driving cars.



We will tell you how does it really work under the hood!

What is the data science learning roadmap?

- Charts out multi-level skills map with details on
  - **What** skills you want to hone,
  - **How** you will measure the outcome at each level
  - and **techniques** to further master each skill

1

BUSINESS PROBLEM!

1 BUSINESS PROBLEM

WHY?....WHY?....WHY?....

ONE OF THE MANY TRAITS OF

A GOOD DATA SCIENTIST!

② DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- API'S
- ONLINE REPOSITORIES

③

DATA PREPARATION

DATA CLEANING

INCONSISTENT DATATYPES

MISSPELLED ATTRIBUTES

MISSING AND DUPLICATE VALUES

TRANSFORMATION

(4) EXPLORATORY DATA ANALYSIS

(4) EXPLORATORY DATA ANALYSIS

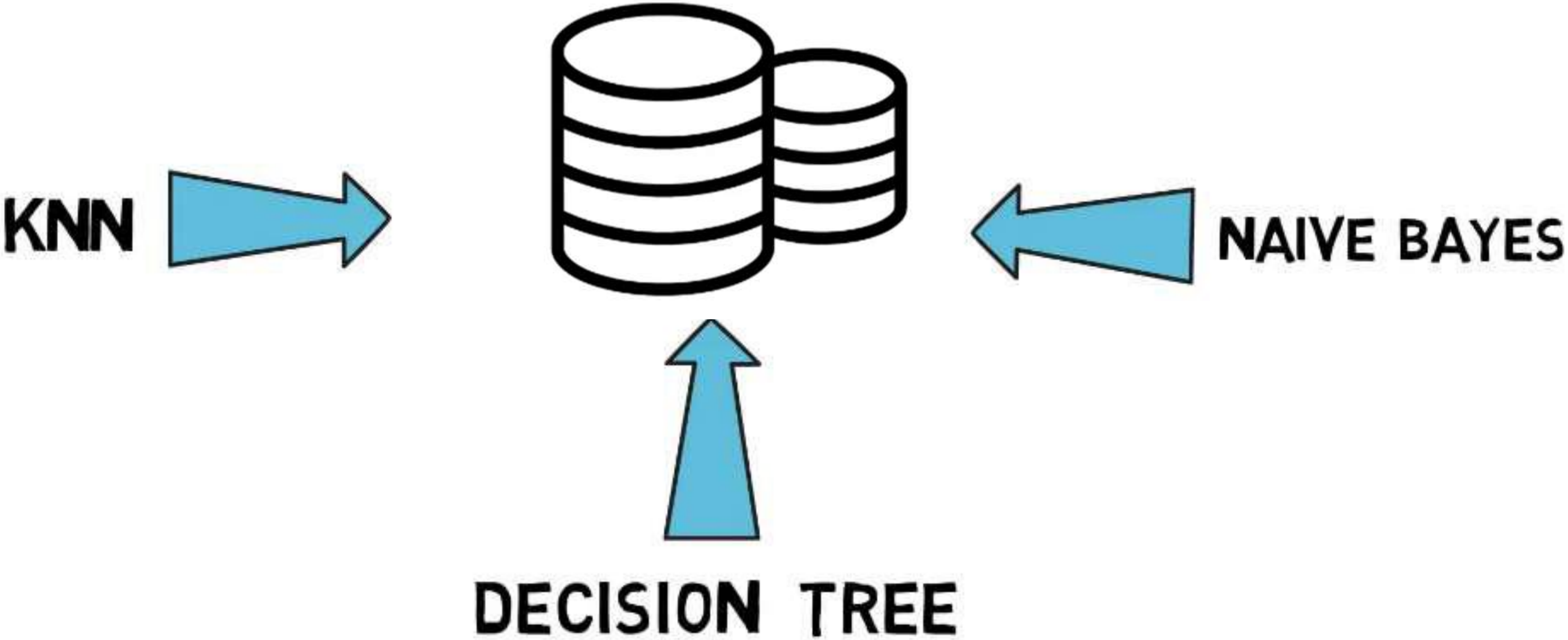
DEFINES AND REFINES

THE SELECTION OF FEATURE

VARIABLES THAT WILL BE USED

IN THE MODEL DEVELOPMENT

5 DATA MODELING



KNN

NAIVE BAYES

DECISION TREE

IDENTIFY THE MODEL THAT BEST FITS THE BUSINESS REQUIREMENT

TRAINS THE MODELS ON THE TRAINING DATASET AND TEST

SELECT THE BEST PERFORMING MODEL

python™

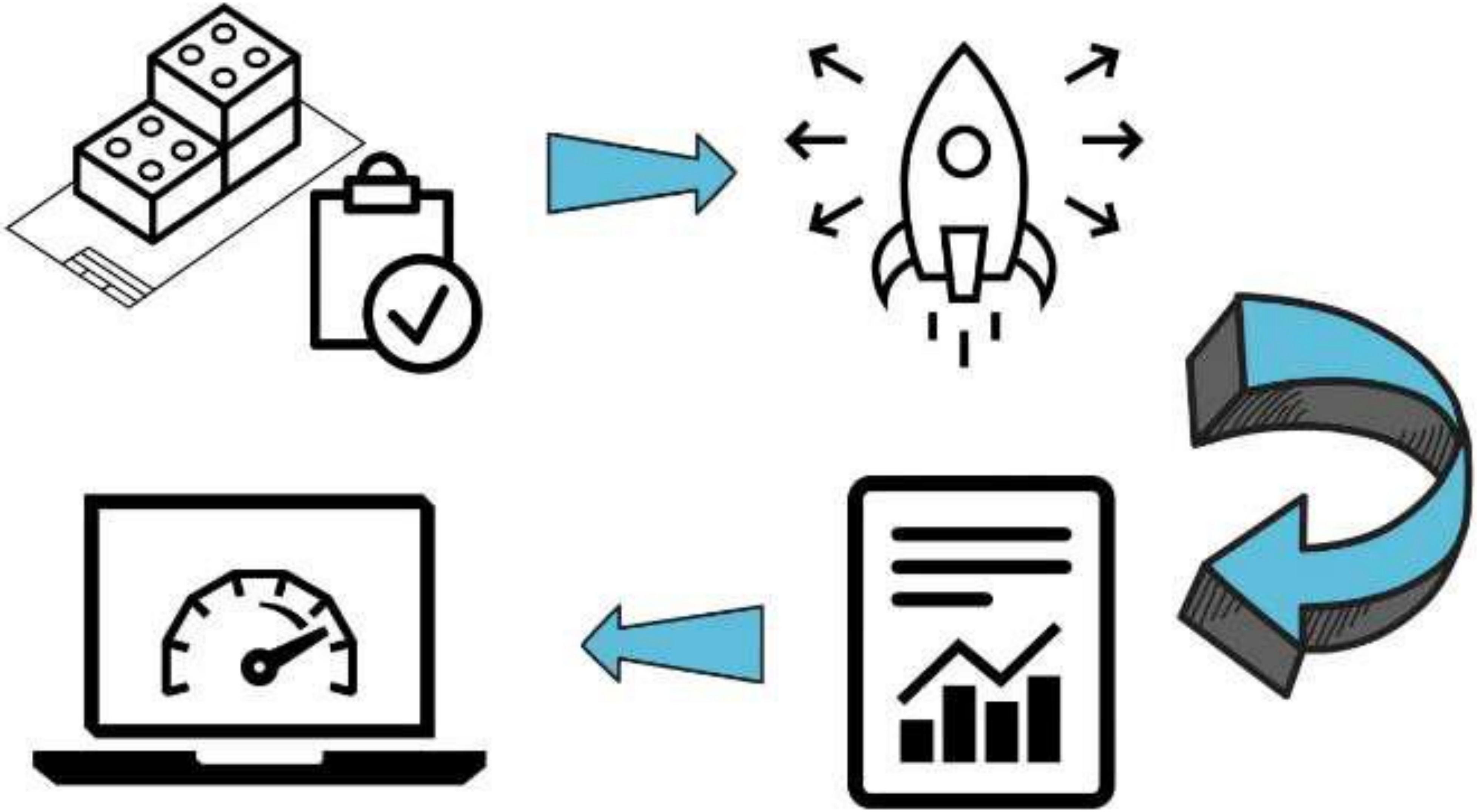
R

sas.

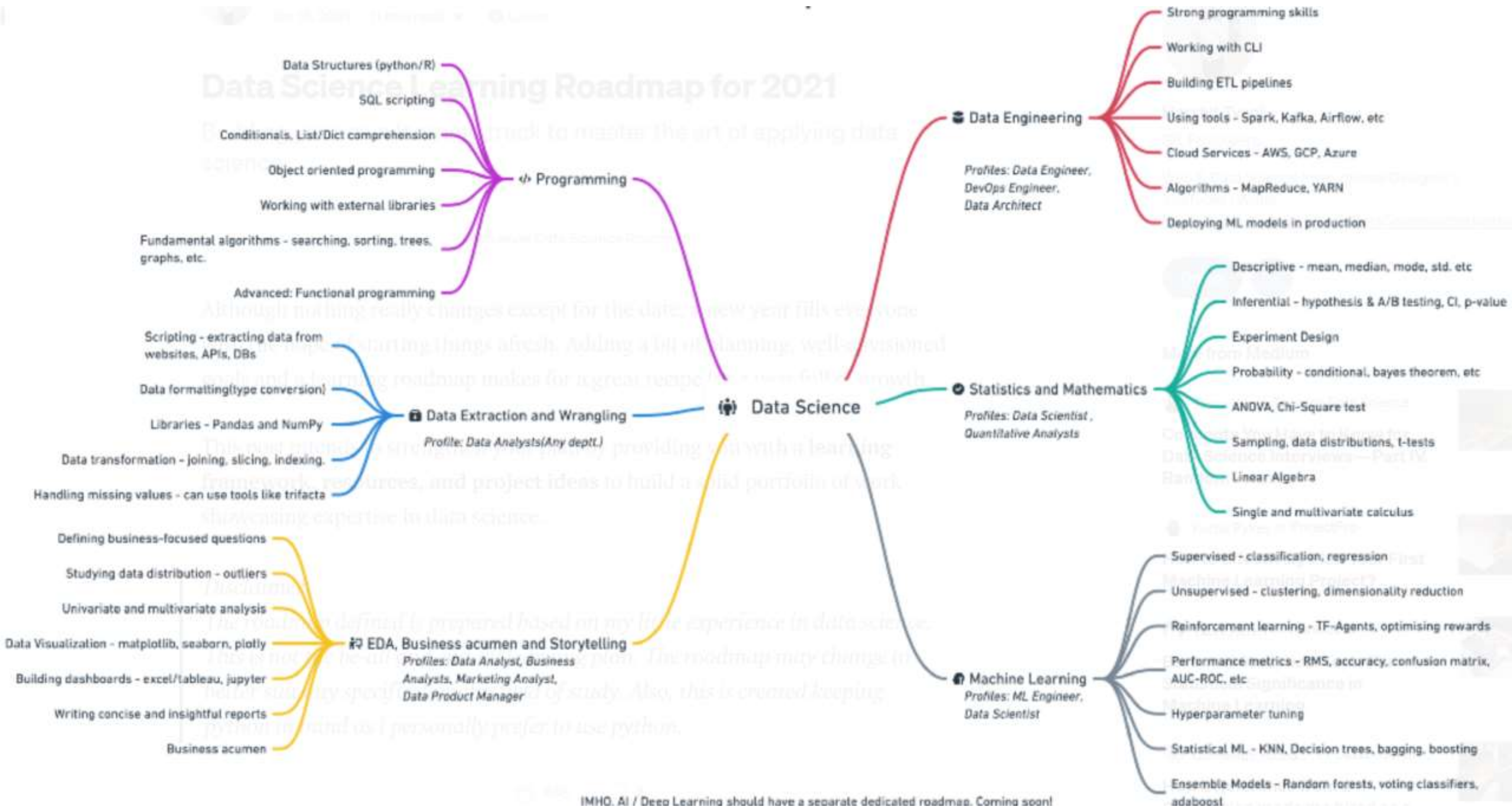
6 VISUALIZATION AND COMMUNICATION

tableau Power BI QlikView

7 DEPLOYS AND MAINTAINS



THE DAILY ROUTINE OF A DATA SCIENTIST IS A WHOLE LOT OF FUN, HAS A LOT OF INTERESTING ASPECTS AND COMES WITH ITS OWN SHARE OF CHALLENGES



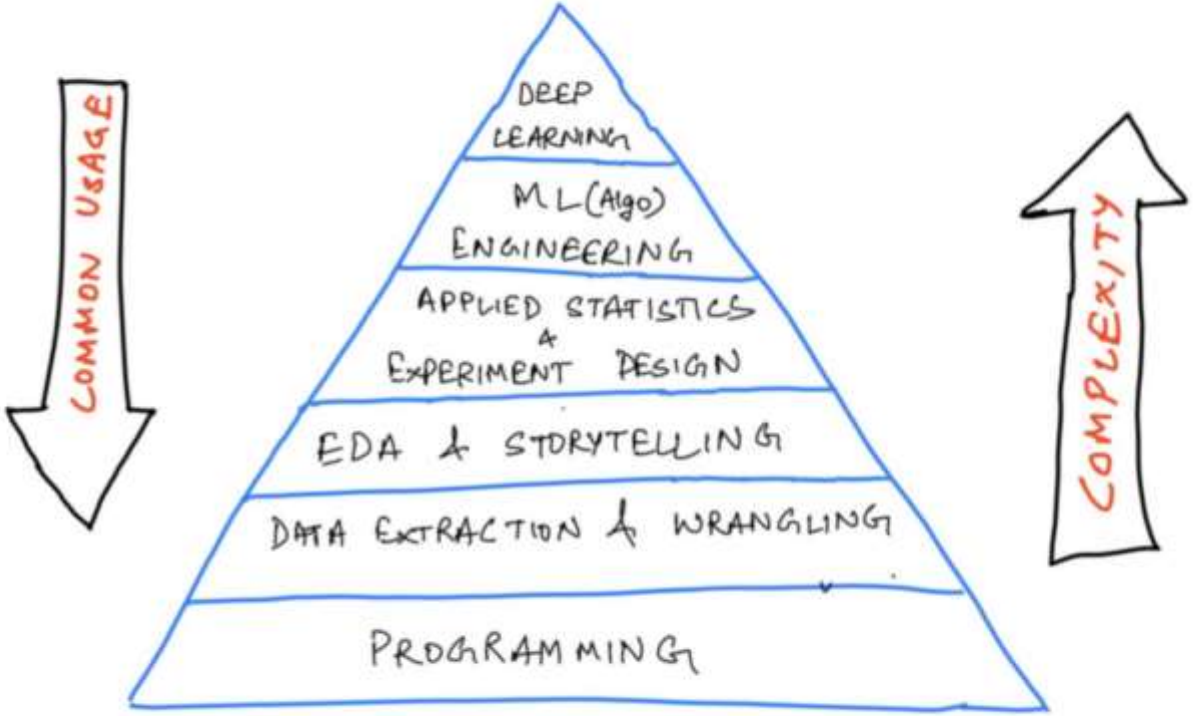
Data Science

- Programming
  - Data Structures (python/R)
  - SQL scripting
  - Conditionals, List/Dict comprehension
  - Object oriented programming
  - Working with external libraries
  - Fundamental algorithms - searching, sorting, trees, graphs, etc.
  - Advanced: Functional programming
- Data Extraction and Wrangling
  - Profile: Data Analysts(Any dept.)
  - Scripting - extracting data from websites, APIs, DBs
  - Data formatting(type conversion)
  - Libraries - Pandas and NumPy
  - Data transformation - joining, slicing, indexing
  - Handling missing values - can use tools like trifacta
- EDA, Business acumen and Storytelling
  - Profiles: Data Analyst, Business Analysts, Marketing Analyst, Data Product Manager
  - Defining business-focused questions
  - Studying data distribution - outliers
  - Univariate and multivariate analysis
  - Data Visualization - matplotlib, seaborn, plotly
  - Building dashboards - excel/tableau, jupyter
  - Writing concise and insightful reports
  - Business acumen
- Data Engineering
  - Profiles: Data Engineer, DevOps Engineer, Data Architect
  - Strong programming skills
  - Working with CLI
  - Building ETL pipelines
  - Using tools - Spark, Kafka, Airflow, etc
  - Cloud Services - AWS, GCP, Azure
  - Algorithms - MapReduce, YARN
  - Deploying ML models in production
- Statistics and Mathematics
  - Profiles: Data Scientist, Quantitative Analysts
  - Descriptive - mean, median, mode, std. etc
  - Inferential - hypothesis & A/B testing, CI, p-value
  - Experiment Design
  - Probability - conditional, bayes theorem, etc
  - ANOVA, Chi-Square test
  - Sampling, data distributions, t-tests
  - Linear Algebra
  - Single and multivariate calculus
- Machine Learning
  - Profiles: ML Engineer, Data Scientist
  - Supervised - classification, regression
  - Unsupervised - clustering, dimensionality reduction
  - Reinforcement learning - TF-Agents, optimising rewards
  - Performance metrics - RMS, accuracy, confusion matrix, AUC-ROC, etc
  - Hyperparameter tuning
  - Statistical ML - KNN, Decision trees, bagging, boosting
  - Ensemble Models - Random forests, voting classifiers, adaboost

IMHO, AI / Deep Learning should have a separate dedicated roadmap. Coming soon!

Complexity and common usage

- Weights to each level based on the complexity and commonality of application in the real-world



Programming or software engineering

- Every data science job description would ask for programming expertise in at least one of the languages (Python/R)
  - Common data structures(data types, lists, dictionaries, sets, tuples), writing functions, logic, control flow, searching and sorting algorithms, object-oriented programming, and working with external libraries
- SQL scripting: Querying databases using joins, aggregations, and subqueries
- Comfortable with using the Terminal, version control in Git, and using GitHub

- Resources for python:
  - learnpython.org
  - Kaggle
  - freecodecamp on YouTube
- SQL:
  - Intro to SQL and Advanced SQL on Kaggle
  - Datacamp also offers many courses on SQL
- Git:
  - Guide for Git and GitHub

Data collection, extraction and wrangling

- A significant part of the data science work is centered around finding apt data that can help you solve your problem
- You can collect data from different legitimate sources:
  - scraping(if the website allows)
  - APIs
  - Databases
  - Publicly available repositories
- Data is rarely clean and formatted for use in the “real world”. Pandas and NumPy are the two libraries that are at your disposal to go from dirty data to ready-to-analyze data.

- Resources:
  - Data Manipulation using pandas
  - Data Cleaning course by Kaggle
  - freecodecamp course on learning Numpy, pandas

Exploratory Data Analysis

- Drawing insights from data and communicating to the management in simple terms
  - **Exploratory data analysis:**
    - Defining questions, handling missing values, outliers, formatting, filtering, univariate and multivariate analysis
  - **Data visualization:**
    - Plotting data using libraries like matplotlib
    - Knowledge to choose the right chart to communicate the findings from the data
  - **Developing dashboards:**
    - Use Excel or a specialized tool like Power BI and Tableau to build dashboards that summarize/aggregate data to help the management in making decisions
  - **Business acumen**:
    - Work on asking the right questions to answer, ones that actually target the business metrics
    - Practice writing clear and concise re

- Resources:
  - Career track on Data Analysis
  - Data Analysis with Python
  - Data Visualization in Spreadsheets, Excel, Tableau, Power BI

Statistics and Mathematics

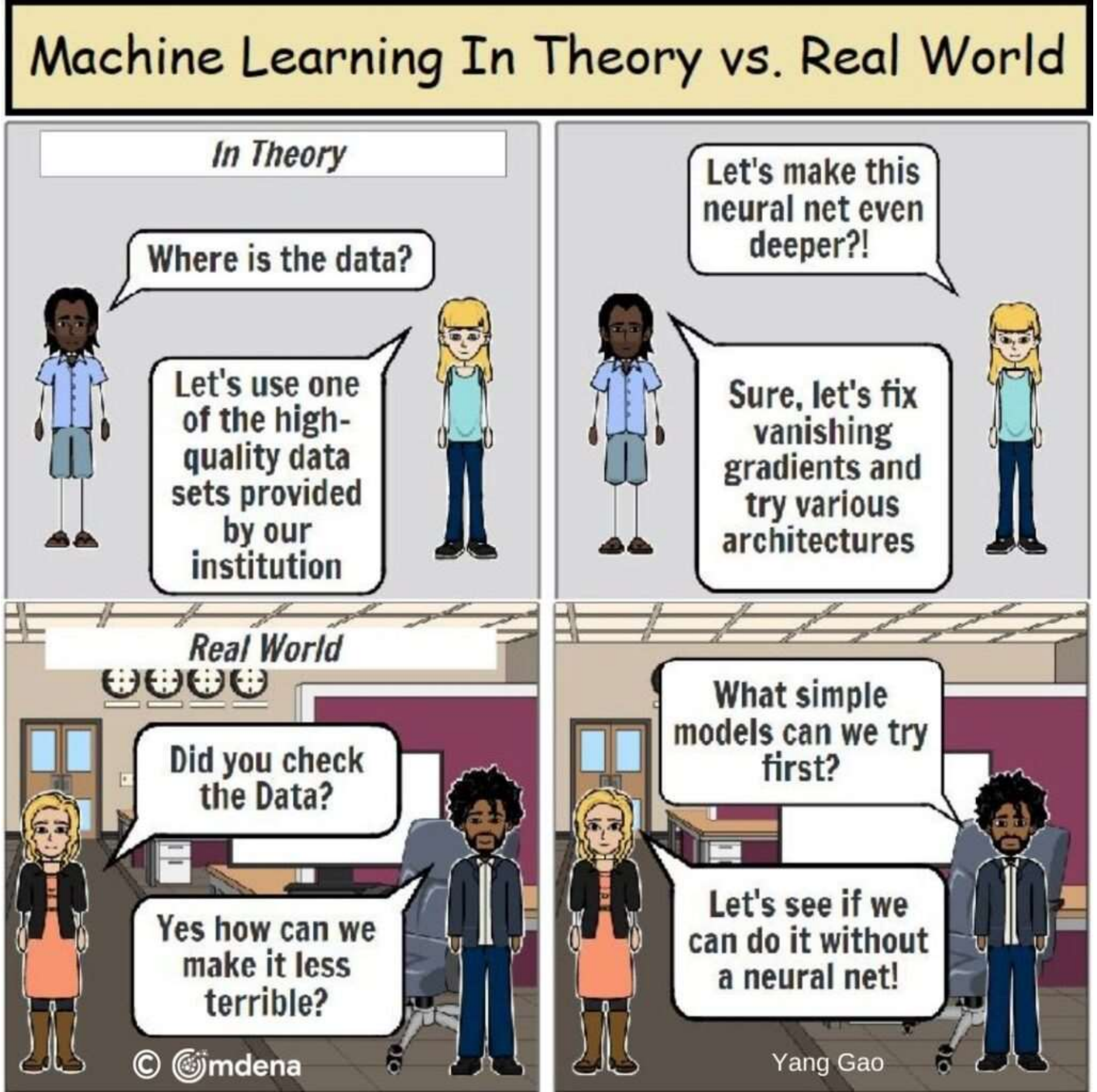
- Statistical methods are a central part of data science
- Focus more on descriptive and inferential statistics
  - **Descriptive Statistics**: to be able to summarise the data is powerful but not always. Learn about estimates of location(mean, median, mode, weighted statistics, trimmed statistics), and variability to describe the data.
  - **Inferential statistics:** designing hypothesis tests, A/B tests, defining business metrics, analyzing the collected data and experiment results using confidence interval, p-value, and alpha values.
  - **Linear Algebra, Single and multi-variate calculus** to understand loss functions, gradient, and optimizers in machine learning.

- Resources:
  - [Book]Practical statistics for data science**(highly recommend)**
  - Statistical thinking in Python
  - Intro to Descriptive Statistics
  - Inferential Statistics
  - Probability and Statistics for Data Science (Series) on Medium
  - Three Blue One Brown Lecture Series

Machine learning

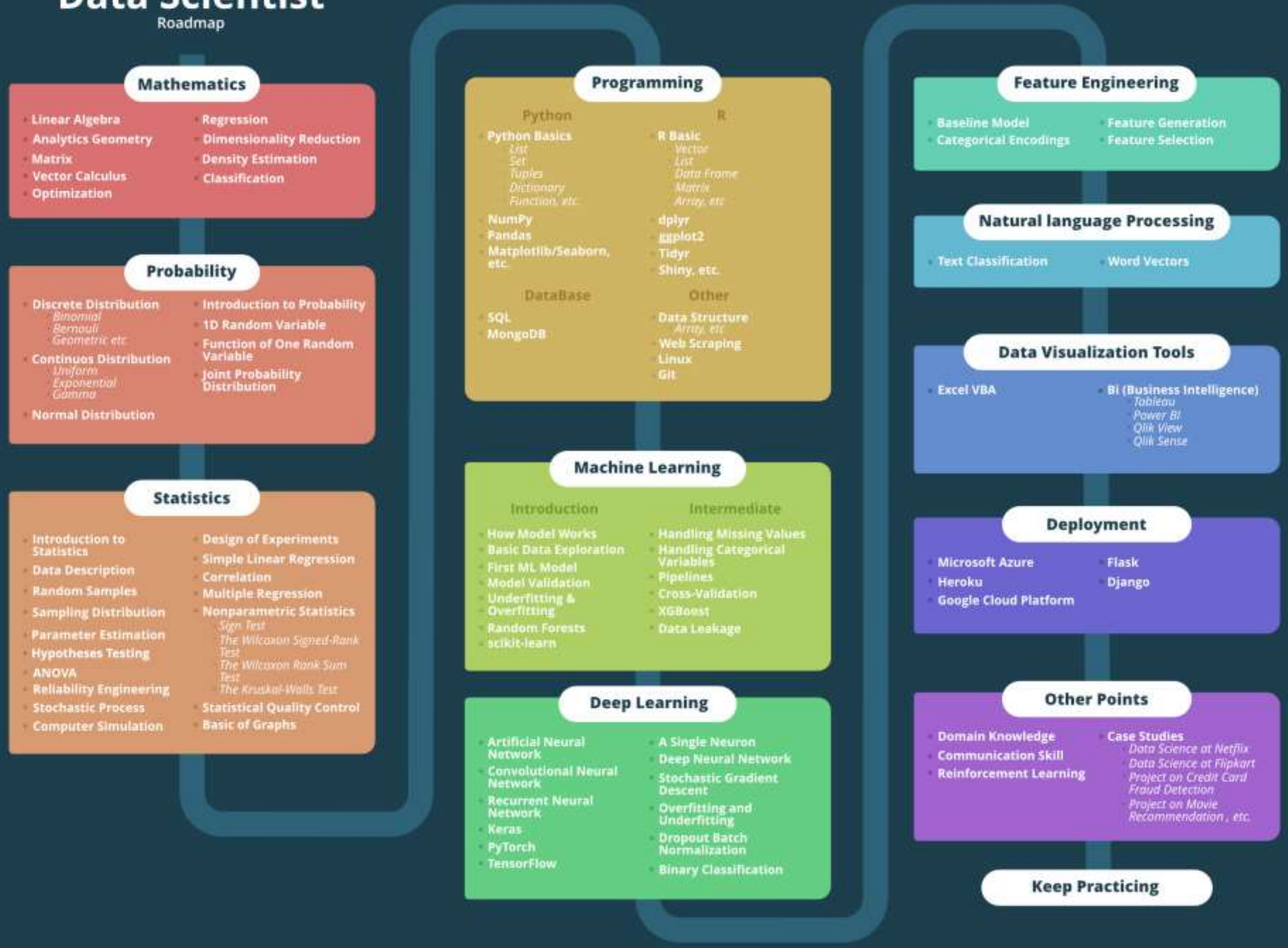
- There are three major types of learning:
  1. **Supervised Learning** — includes regression and classification problems. Study simple linear regression, multiple regression, polynomial regression, naive Bayes, logistic regression, KNNs, tree models, ensemble models. Learn about evaluation metrics.
  2. **Unsupervised Learning** — Clustering and dimensionality reduction are the two widely used applications of unsupervised learning. Dive deep into PCA, K-means clustering, hierarchical clustering, and gaussian mixtures.
  3. **Reinforcement learning**(can skip*) — helps you build self-rewarding systems. Learn to optimize rewards, using the TF-Agents library, creating Deep Q-networks, etc.

- Resources:
  - [book]Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition
  - [book]Pattern recognition and Machine Learning
  - Machine Learning Course by Andrew Ng
  - Introduction to Machine Learning
  - Supervised learning with Python


Machine Learning In Theory vs. Real World
In Theory
Where is the data?
Let's use one of the high-quality data sets provided by our institution
Let's make this neural net even deeper?!
Sure, let's fix vanishing gradients and try various architectures
Real World
Did you check the Data?
Yes how can we make it less terrible?
What simple models can we try first?
Let's see if we can do it without a neural net!
© Omdena
Yang Gao

Data Scientist

Roadmap



Mathematics

- Linear Algebra
- Analytics Geometry
- Matrix
- Vector Calculus
- Optimization
- Regression
- Dimensionality Reduction
- Density Estimation
- Classification

Probability

- Discrete Distribution
  - *Binomial*
  - *Bernoulli*
  - *Geometric etc*
- Continuos Distribution
  - *Uniform*
  - *Exponential*
  - *Gamma*
- Normal Distribution
- Introduction to Probability
- 1D Random Variable
- Function of One Random Variable
- Joint Probability Distribution

Statistics

- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypotheses Testing
- ANOVA
- Reliability Engineering
- Stochastic Process
- Computer Simulation
- Design of Experiments
- Simple Linear Regression
- Correlation
- Multiple Regression
- Nonparametric Statistics
  - *Sign Test*
  - *The Wilcoxon Signed-Rank Test*
  - *The Wilcoxon Rank Sum Test*
  - *The Kruskal-Walls Test*
- Statistical Quality Control
- Basic of Graphs

Programming

Python

- Python Basics
  - *List*
  - *Set*
  - *Tuples*
  - *Dictionary*
  - *Function, etc.*
- NumPy
- Pandas
- Matplotlib/Seaborn, etc.

R

- R Basic
  - *Vector*
  - *List*
  - *Data Frame*
  - *Matrix*
  - *Array, etc*
- dplyr
- ggplot2
- Tidyr
- Shiny, etc.

DataBase

- SQL
- MongoDB

Other

- Data Structure
  - *Array, etc*
- Web Scraping
- Linux
- Git

Machine Learning

Introduction

- How Model Works
- Basic Data Exploration
- First ML Model
- Model Validation
- Underfitting & Overfitting
- Random Forests
- scikit-learn

Intermediate

- Handling Missing Values
- Handling Categorical Variables
- Pipelines
- Cross-Validation
- XGBoost
- Data Leakage

Deep Learning

- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- PyTorch
- TensorFlow
- A Single Neuron
- Deep Neural Network
- Stochastic Gradient Descent
- Overfitting and Underfitting
- Dropout Batch Normalization
- Binary Classification

Feature Engineering

- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural language Processing

- Text Classification
- Word Vectors

Data Visualization Tools

- Excel VBA
- BI (Business Intelligence)
  - *Tableau*
  - *Power BI*
  - *Qlik View*
  - *Qlik Sense*

Deployment

- Microsoft Azure
- Heroku
- Google Cloud Platform
- Flask
- Django

Other Points

- Domain Knowledge
- Communication Skill
- Reinforcement Learning
- Case Studies
  - *Data Science at Netflix*
  - *Data Science at Flipkart*
  - *Project on Credit Card Fraud Detection*
  - *Project on Movie Recommendation , etc.*

Keep Practicing

Soft skills (people behavior skills)

- Commonly used to “***refer to the “emotional side” of human beings*** in opposition to the IQ
- ***Character traits*** and ***interpersonal skills*** that characterize a person's relationships with others
  - Help employees interact with others and succeed in the workplace
- Describe a **person's emotional quotient** (EQ) as opposed to intelligence quotient (IQ)
- Soft skills include:
  - Communication skills
  - Mentor your coworkers
  - Leadership skills
  - Follow instructions, and get a job done on time
  - Team building and Teamworking skills
  - Problem-solving skills
  - Analytical skills
  - Collaboration

97% of employers say that soft skills are either as important or more important than hard skills

80% of companies' success is due to soft skills

Resources

- https://towardsdatascience.com/data-science-learning-roadmap-for-2021-84f2ba09a44f
- https://skaf.medium.com/data-scientist-roadmap-2022-3e247fe6fe87
- https://www.mltut.com/data-science-with-python-roadmap/
- https://towardsdatascience.com/become-a-data-scientist-in-2022-a-practical-52-week-course-8244cc18284e
- https://omdena.com/blog/data-science-road-map/