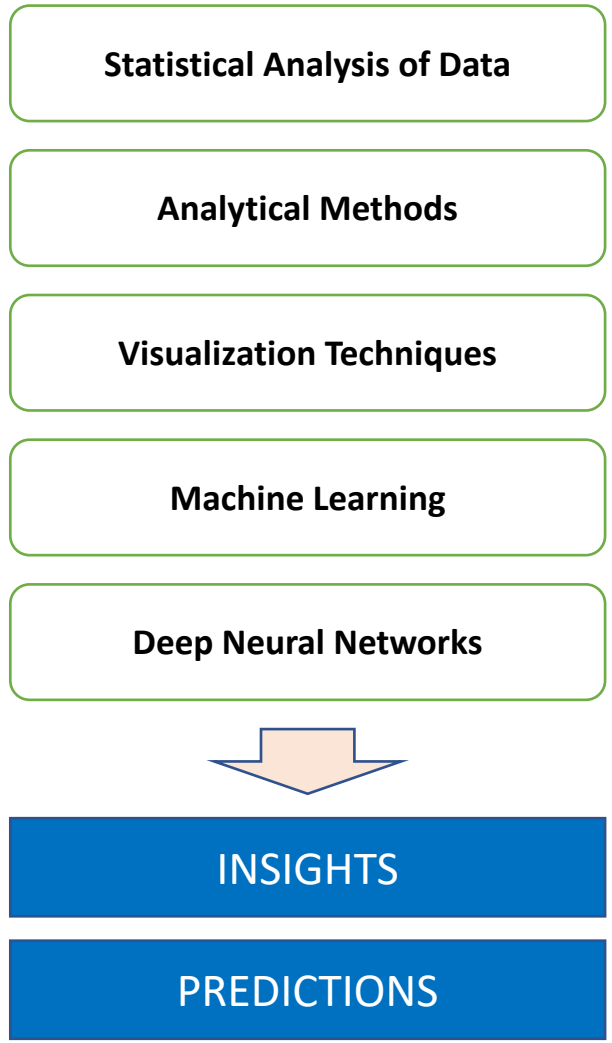
Practical Data Science using Python
Introduction

Contents

What is Data Science

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.



Data Science Use Cases

Banks are creating solutions for making on-the-spot decisions to loan applicants using machine learning-powered credit risk models.

E-Commerce companies are using Machine Learning to build powerful product recommendation engines that helps suggest relevant products to its website visitors.

Digital media companies have started using Data Science based solutions to use Media Analytics and predictive models to smartly embed media to the users' taste and preferences in the right channel at the right timing for maximising viewership and revenue.

Healthcare companies are developing solutions to scan medical imaging to automatically detect diseases without doctor's intervention.

Manufacturing companies are using Analytics and Prediction on machine-sensor data to prevent faults in the machinery y detecting them before they break down.

Data Science is positively impacting all Industries

Who oversees Data Science in an Organization

Business Leaders: They work with the Data Science teams to define business problems and develop strategies to solve them. They may be the head of a line of business, such as marketing, finance, or sales, and have a data science team reporting directly or indirectly to them.

IT Leaders: Senior IT Leaders are responsible for the infrastructure and architecture that will support data science operations. They are continually monitoring operations and resource usage to ensure that data science teams operate efficiently and securely. They may also be responsible for building and updating IT environments for data science teams.

Data Science and Analytics Leaders: They oversee the Data Science teams and their day-to-day work. They are team builders who can balance team development with project planning and monitoring.

But the most important player in this process is the Data Scientist.

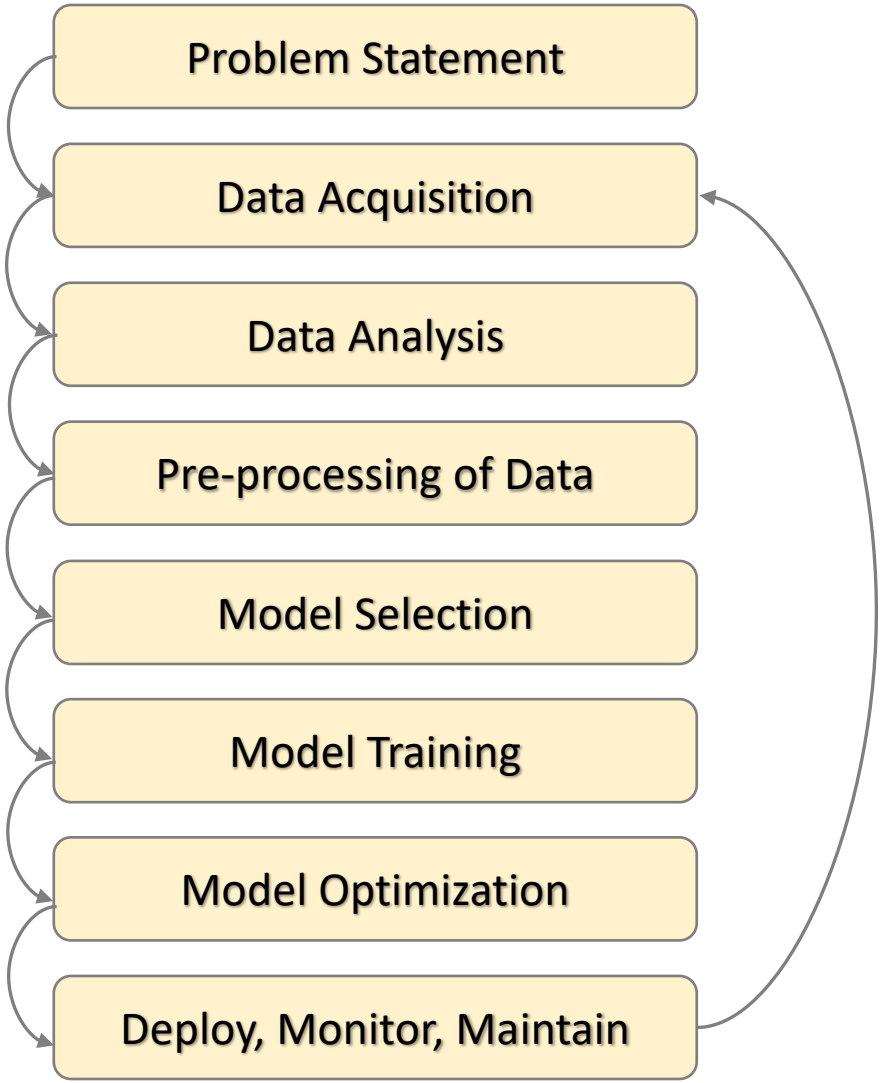
Who is a Data Scientist

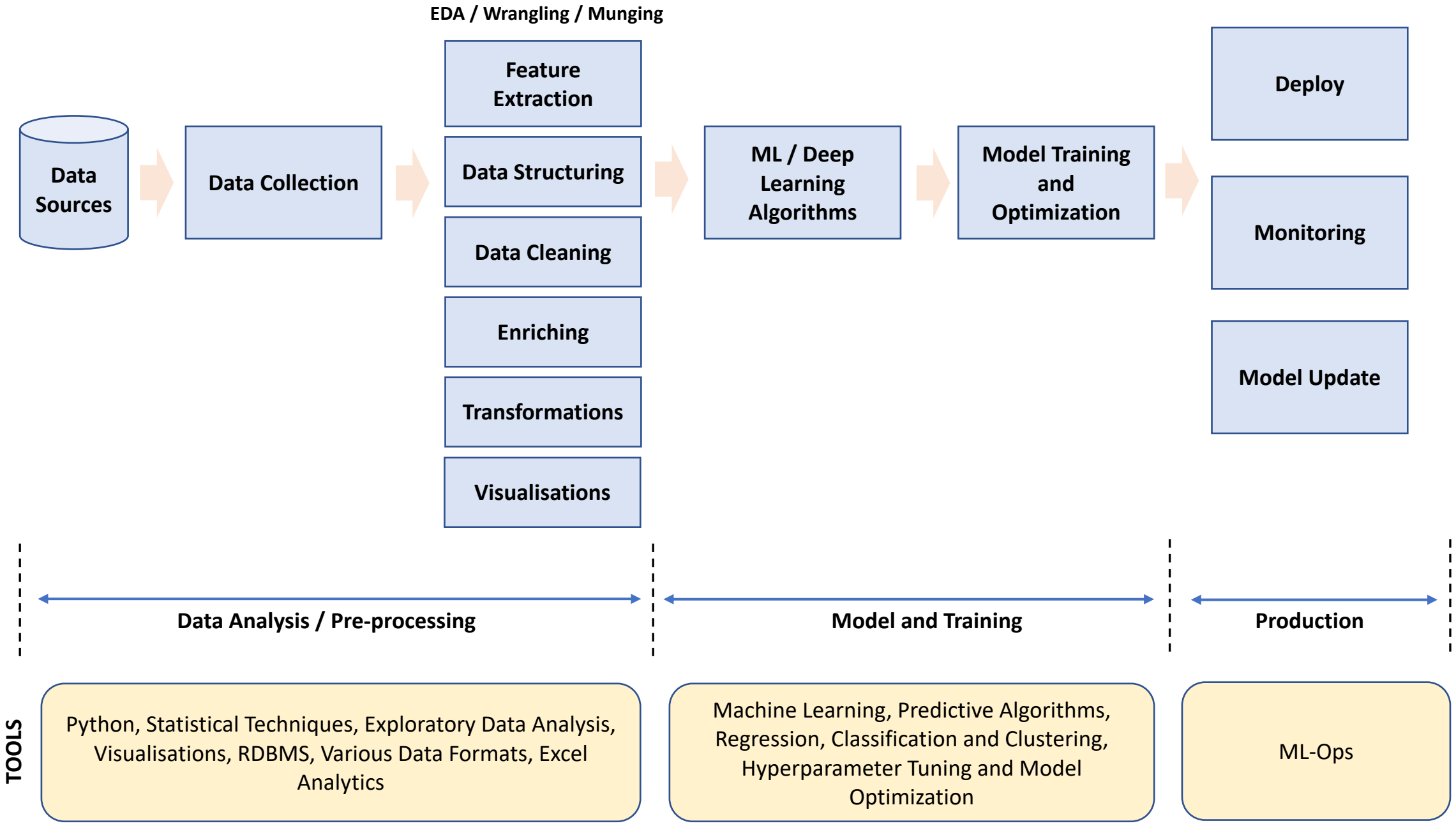
- A Data Scientist's duties can include devel**oping strategies for analysing data, preparing data for analysis, exploring, analysing, and visualizing data, building models** with data using programming languages, such as Python and R, and deploying models into applications.

- The Data Scientist **works in teams**. In addition to a data scientist, this team might include a **business analyst** who defines the problem, a **data engineer** who prepares the data and how it is accessed, an **IT architect** who oversees the underlying processes and infrastructure, and an **production engineer** who deploys the models or outputs of the analysis into applications and products.

Data Science / Machine Learning Project Lifecycle

Every Machine Learning/Data Science initiative or project follow certain steps. There would be variations depending on the business goal and the type of Machine Learning algorithm used. However, there are some broad steps that will almost always be applicable. In the following few sections, we will look at these steps forming a complete Machine Learning project.




EDA / Wrangling / Munging
Data Sources
Data Collection
Feature Extraction
Data Structuring
Data Cleaning
Enriching
Transformations
Visualisations
ML / Deep Learning Algorithms
Model Training and Optimization
Deploy
Monitoring
Model Update
Data Analysis / Pre-processing
Model and Training
Production
TOOLS
Python, Statistical Techniques, Exploratory Data Analysis, Visualisations, RDBMS, Various Data Formats, Excel Analytics
Machine Learning, Predictive Algorithms, Regression, Classification and Clustering, Hyperparameter Tuning and Model Optimization
ML-Ops

Data Science Tools and Technologies

Type of Data Science Tools	Tool Name	Functions
Programming Languages	Python	Base Programming Language for Data Science and ML Tasks. Massive amount of Libraries and community support available.
	R Language	Base Programming Language. Very mature and supported well. More specialized for Statistical problem solving.
	Julia	Based Programming Language, more suited for scientific and mathematical problems.
ML Libraries	TensorFlow	Most popular Tensor Management Library from Google. This is the backend for most Deep Learning tasks.
	Microsoft CNTK	ML Backend Library from Microsoft.
	Theano	ML Backend Library, more suited for Scientific applications.
ML APIs	Keras	User friendly Machine Learning API that can tap TensorFlow, CNTK, Theano.
	PyTorch	ML User API and can work with all the ML backend Libraries.
Data Libraries	Numpy	Array management library
	Pandas	Data Structure Library (DataFrame and Series Objects)
Visualization Libraries	Matplotlib	Most used Visualization Library and backend for may modern Libraries
	Seaborn	Visualization Library built on top of MarplotLib
	Bokeh	A more modern and elegant Visualization Library
Analytics Tools	Tableau	Analytics Application for Enterprises
	SAS	Analytics Application for Enterprises
	Qlik Sense	Analytics Application for Enterprises

Scope of Study in this Course

- Python Language
- Fundamentals of Statistics
- Data Analytics Techniques
- Data Visualization Techniques
- Machine Learning

- Types of Machine Learning Algorithms
- Model Building
- Model Evaluation
- Model Optimization

Data to Wisdom – Purpose of Data Science

Data	⟹	Raw, in primary format, collected at source
Information	⟹	Finding relationships within data
Knowledge	⟹	When it is useful for a purpose
Wisdom	⟹	When it is helpful in making decisions

Data Science aims to derive wisdom out of raw Data through a series of steps

Analytics vs. Data Science

- **Analytics is performed manually on Data** by setting up rules of information extraction. Data Science (Predictive) models perform **automatic pattern recognition** on the data.
- The above also means that the Predictive or **Machine Learning Models are dynamic** based on changing data while Analytics patterns are not.
- In analytics, testing is performed to check that the defined outcome is achieved as expected, while in machine learning, model is tested against prior labels/outcomes to optimize its design depending on the nature of the data.
- Techniques and tools used to develop analytics models and machine learning models differ. Machine learning modelling techniques are much more advanced and are built on statistical and mathematical principles related to how the machine will learn to optimize the model performance.

Data Analyst

- BI Tools
- Intermediate Statistics
- RDBMS
- Excel
- Data Visualisations
- Data Analysis
- Data Mining

Data Scientist

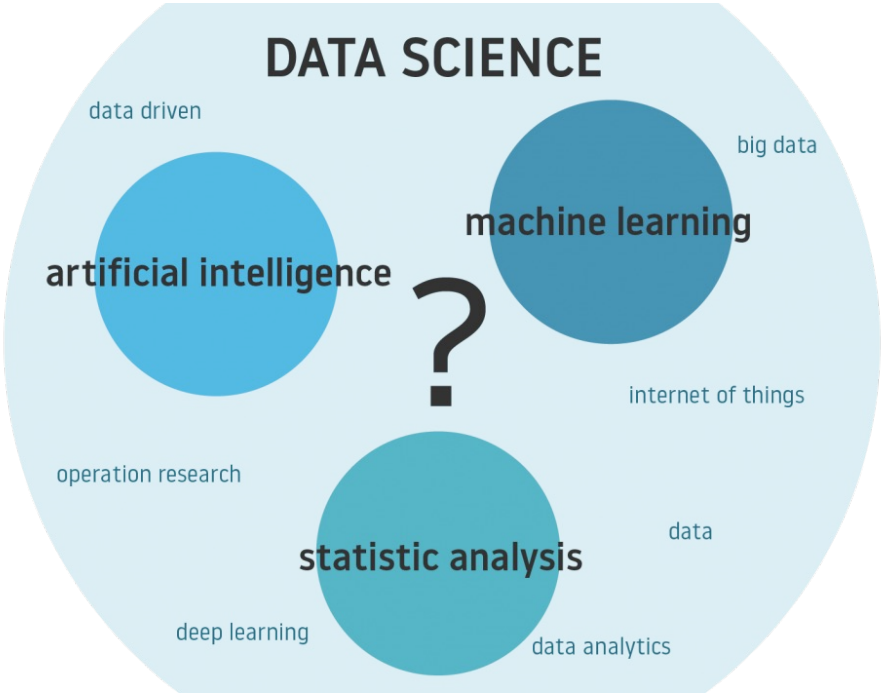
- Advanced Statistical Methods
- Machine Learning
- Data Analysis
- Visualizations
- Predictive Analytics
- Data Mining
- ML-Ops
- Algorithm Development

Data Science vs. Machine Learning

- Data Science encompasses a number of tools, techniques and technologies in dealing with data, especially in bringing out wisdom from raw data.

- Machine Learning, Neural Networks, Statistical Analysis are some of the areas that comprise Data Science.

- Some of the Techniques that are part of Data Science are
  - Data Collection
  - Data Mining
  - Data Extraction
  - Data Cleaning/refining
  - Data Analysis
  - Data Visualization
  - Predictive Modelling/Machine Learning



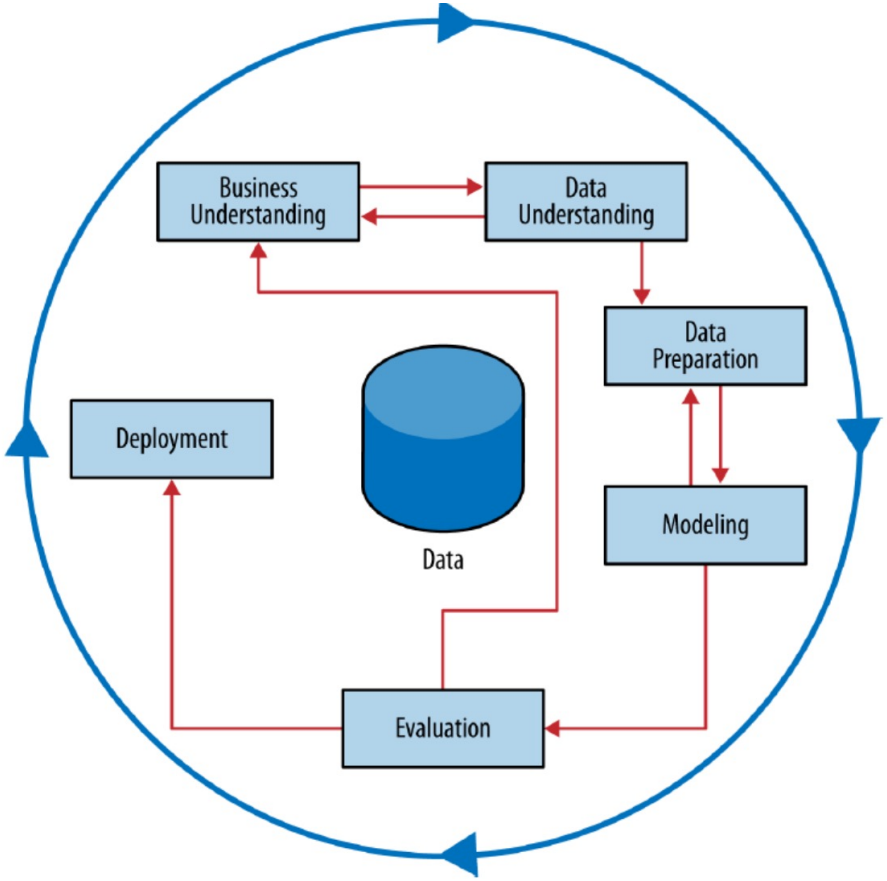
Sources of Data in Data Science

Social Data ⟹ Gathered from the Likes, Tweets & Retweets, Comments, Video Uploads, and general media that are uploaded and shared via social media platforms. This kind of data provides invaluable insights into consumer behaviour and sentiment and can be enormously influential in marketing analytics.

Machine Data ⟹ These are data generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behaviour. This type of data is expected to grow exponentially as the internet of things grows ever more pervasive and expands around the world. Sensors such as medical devices, smart meters, road cameras, satellites, games and the rapidly growing Internet Of Things will deliver high velocity, value, volume and variety of data in the very near future.

Transactional Data ⟹ These are generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts – all are characterized as transactional data.

Data Science Project Life Cycle: CRISP-DM Model



Cross-Industry Standard Process for Data Mining (CRISP-DM)

The CRISP-DM model is flexible and can be customized.

For example, if your organization aims to detect money laundering, it is likely that you will sift through large amounts of data without a specific modelling goal. Instead of modelling, your work will focus on data exploration and visualization to uncover suspicious patterns in financial data.

In such a situation, the modelling, evaluation, and deployment phases might be less relevant than the data understanding and preparation phases.

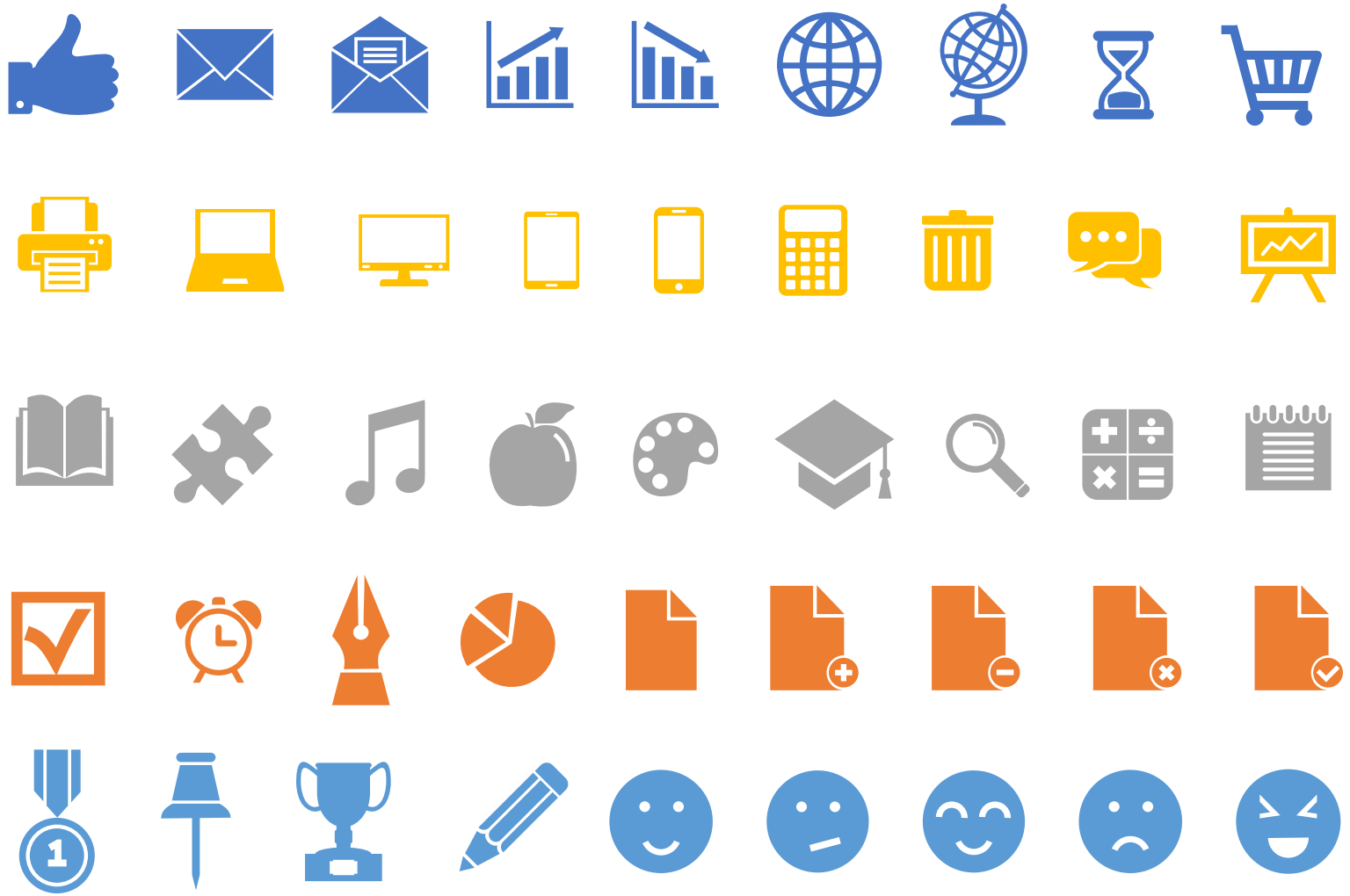
Fully Editable Icon Sets: A

You can Resize without losing quality

You can Change Fill Color &
Line Color

FREE
PPT TEMPLATES

www.allppt.com



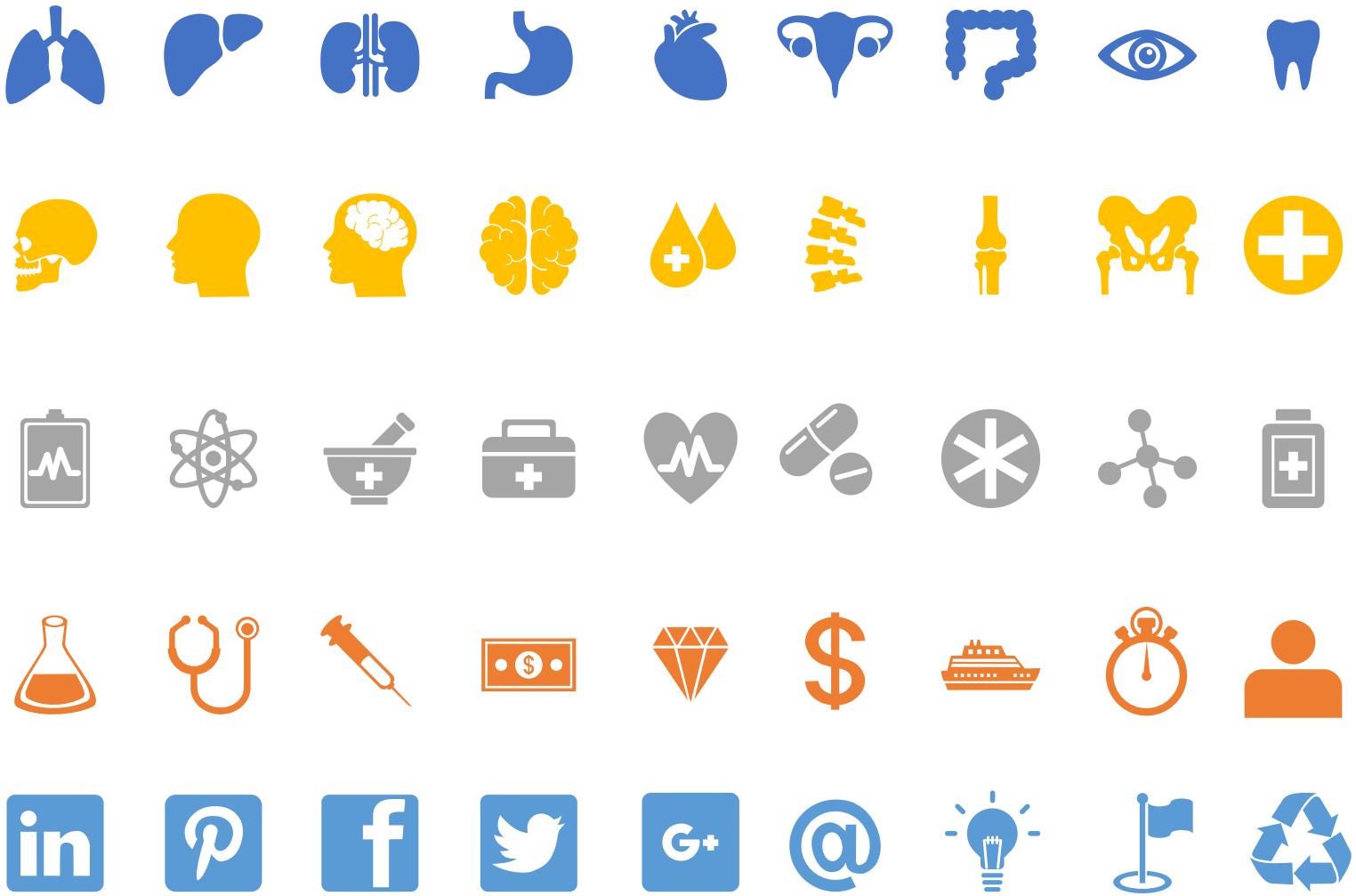
Fully Editable Icon Sets: B

You can Resize without losing quality

You can Change Fill Color &
Line Color

FREE
PPT TEMPLATES

www.allppt.com



Fully Editable Icon Sets: C

You can Resize without losing quality

You can Change Fill Color &
Line Color

FREE
PPT TEMPLATES

www.allppt.com