Most of the courses are free in audit mode because I'd completed 55% of overall courses which is given below.
I'd started my online MOOC(Massive Online Open Courses) in Nov 2019. You can check me here whatever courses I did.
https://www.youracclaim.com/users/vikrant-singh.03191143

So I'm providing a complete analysis of all the best online platforms, educational blogs, tools you can use if you want to join the course, or want to build your own online learning platform.

1.Data Science Track

Month 1 - Data Analysis
Week 1 - Learn Python
EdX
https://www.edx.org/professional-certificate/python-data-science
https://www.edx.org/xseries/mitx-computational-thinking-using-python
Week 2 - Statistics & Probability
KhanAcademy https://www.khanacademy.org/math/statistics-probability
Week 3 Data Pre-processing, Data Visualization, Exploratory Data Analysis
EdX https://www.edx.org/course/introduction-to-computing-for-data-analysis
Week 4 Kaggle Project #1
Try your best at a competition of your choice from Kaggle.
https://www.kaggle.com/getting-started/148810Use Kaggle Learn as a helpful guide
Month 2 - Machine Learning
The math of Machine Learning Cheat Sheets
Statistics
Probability
Calculus
Linear Algebra
Week 1-2 - Algorithms & Machine Learning
Columbia https://courses.edx.org/courses/course-v1:ColumbiaX+DS102X+2T2018/course/
Week 3 - Deep Learning
Part 1 and 2 of DL Book https://www.deeplearningbook.org/
https://www.youtube.com/watch?v=vOppzHpvTiQ&list=PL2-dafEMk2A7YdKv4XfKpfbTH5z6rEEj3
Week 4 - Kaggle Project #2
Try your best at a competition of your choice from Kaggle. Make sure to add great documentation to your GitHub repository! Github is the new resume.
Month 3 - Real-World Tools
Week 1 Databases (SQL + NoSQL)
Udacity https://www.udacity.com/course/intro-to-relational-databases--ud197
EdX https://www.edx.org/course/introduction-to-nosql-data-solutions-2

Week 2 Hadoop & Map-Reduce + Spark

Udacity https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617

Spark Workshop https://stanford.edu/~rezab/sparkclass/slides/itas_workshop.pdf

Week 3 Data Storytelling

Edx https://www.edx.org/course/analytics-storytelling-impact-1

Week 4 Kaggle Project #3

Try your best at a competition of your choice from Kaggle.

2.Machine Learning Track

Month 1

Week 1 Linear Algebra

https://www.youtube.com/watch?v=kjBOesZCoqc&index=1&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/

Week 2 Calculus

https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr

Week 3 Probability

https://www.edx.org/course/introduction-probability-science-mitx-6-041x-2

Week 4 Algorithms

https://www.edx.org/course/algorithm-design-analysis-pennx-sd3x

Month 2

Week 1

Learn python for data science

https://www.youtube.com/watch?v=T5pRlIbr6gg&list=PL2-dafEMk2A6QKz1mrk1uIGfHkC1zZ6UU

Math of Intelligence

https://www.youtube.com/watch?v=xRJCOz3AfYY&list=PL2-dafEMk2A7mu0bSksCGMJEmeddU_H4D

Intro to Tensorflow

https://www.youtube.com/watch?v=2FmcHiLCwTU&list=PL2-dafEMk2A7EEME489DsI468AB0wQsMV

Week 2

Intro to ML (Udacity) https://eu.udacity.com/course/intro-to-machine-learning--ud120

Week 3-4

ML Project Ideas https://github.com/NirantK/awesome-project-ideas

Month 3 (Deep Learning)

Week 1

Intro to Deep Learning https://www.youtube.com/watch?v=vOppzHpvTiQ&list=PL2-dafEMk2A7YdKv4XfKpfbTH5z6rEEj3

Week 2

Deep Learning by Fast.AI http://course.fast.ai/

Week 3-4

Re-implement DL projects from Github https://github.com/llSourcell?tab=repositories

3.Deep Learning Track

https://drive.google.com/file/d/1DXdl4iPzYy7GEFRUROUv8cZRSxgUmu1E/view?usp=drivesdk

This folder contains all deep learning & Computer Science Track.

It contains links to Machine Learning & Data Science Courses, books, Practice Papers, Interview, Videos, Jupyter Notebooks of many projects everything you need to know. All links connect your best Medium blogs, Youtube, Top universities free courses.

We are really thankful to all contributors.

This is the link for interview practices

https://drive.google.com/file/d/1CL7Blkfelpcj3snyARvRXKVX6KDysLQs/view?usp=drivesdk

Booklists provided by MIT (most of them are free)

https://drive.google.com/file/d/1XRCbtNz2k-H5b_CXO-ZSAxnoD6S2NZTF/view?usp=drivesdk

Data Science Books(Probability, Linear Algebra, Statistics, Data Analytics....)

https://mega.nz/folder/0iZFXCbA#Rwh3Km42_YaRvgY_NOAvWw

https://mega.nz/folder/g2BRhaDJ#v2XWSegTk3sH6ZcLPNG-WA

Python & Machine Learning books(Programming, Applied Statistics with R...Etc)

https://mega.nz/folder/NmQRlaBa#0FKTDkkHYBmkSmcEu0kGoQ

If you want any other book & you don't want to purchase then please share the cover image of the book I will try to send a link to the complete pdf. I will upload the link of some of them in this blog in the future.

Complete Guide & Course of Quantum Machine Learning

If you are interested in Physics & Philosophy then you will get ultimate links of teachers, videos, courses, companies (IBM, Google, Microsoft..Etc) developments, quantum machine learning codes(Actually all are in the developing phase but you can start with Q# or q sharp with developing some understanding in Complex Numbers)...From here
https://drive.google.com/file/d/1Dy2oEsWazYlvKuqDPjEDh-79ywm_hiJX/view?usp=drivesdk

If you had just joined the kaggle

If you are starting from zero, you will get everything in my previous post.

1. Everything needs to know before Data Science

https://www.kaggle.com/getting-started/191220

2.Machine Learning basics

https://www.kaggle.com/getting-started/191390

3.Titanic Survival Project Solution for new learners

https://www.kaggle.com/vik2012kvs/titanic-survivals-project

4. Built a Chatbot in 9 Lineshttps://www.kaggle.com/getting-started/148810

https://www.kaggle.com/getting-started/191218

5. Built a face Recognition app in 9 Lines

https://www.kaggle.com/vik2012kvs/tutorial-face-recogination-in-9-lines

6.Interview Questions

https://www.kaggle.com/questions-and-answers/191039

I wish you will get some confidence after going through the above 6 links

Dependencies of Analysis

1.Alexa Ranking
2.300+ online Learning & Tech websites(Supporting Students for searching courses)
3.Rank #1 MOOC searching Engine ClassCentral(Dhawal Shah)
4.Millions of student reviews, enrollments, ratings, etc.
5.Support, length of courses, price, course materials, etc.
6.Pros & Cons

Top Online Learning Platforms

This pdf contains all trending & demanding Learning platforms. Here you get pros & cons and reviews which give you an idea before taking any course

https://drive.google.com/file/d/1J0ct16O9ULpqgrmEVazhiuHGISy2aQmI/view?usp=sharing

you can reach the learning platform by just clinking the links mentioned in the text to explore & more.

Top Online Learning Blogs

This is based on Alexa ranking, Number of Followers, Likes, Ratings..etc.

https://drive.google.com/file/d/1J5NTL7bKW9vkx-S07CqVE2A1DuIc4gTM/view?usp=sharing

From this text, you directly reach the live ranking, followers, rating, contact, email updates..etc.

It is updated whenever you reach any platform via given links.

https://www.kaggle.com/getting-started/1488105000+Online Courses

It is highly based on student's reviews who had taken the courses or going through courses. It is based on the #1 MOOC search engine ClassCentral.It is updated every day according to any change occurs in any platform in any course curriculums.

Number of Courses
1.Computer Science &bArtificial Intelligence - 1928
2.Data Science- 712
3.Programming - 1425
4.Mathmatics- 517
5.Bussiness - 3313
6.Science - 1616

and many more but it covers EdX,Coursera,Future Learn,MIT MOOC,Stanfords,Harvard Extension,IBM,Google,Microsoft,NPTEL,Udacity,Udemy,...etc(premium courses including online degrees).

https://drive.google.com/file/d/1JIw108xNUwdqv3CS3-FcxZmI9EwQ8Um0/view?usp=sharing

Online Platform Tools

If you are looking for making your own online academy then you may get from this text. It

contains the top 10 authoring tools. It contains pros, cons, price, ease of doing, etc.
https://drive.google.com/file/d/1J5NTL7bKW9vkx-S07CqVE2A1DuIc4gTM/view?usp=sharing

Recommended Sites to Learn Data Science

Coursera Online Learning

Coursera is an online learning platform that offers courses and degrees in a variety of areas, including machine learning. It works with universities to offer more than 2,000 courses.

Their courses topics include:

(i)Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks).

(ii) Unsupervised learning (clustering, dimensionality reduction, recommender systems, deep learning).

(iii) Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI). The course will also draw from numerous case studies and applications so that you'll also learn how to apply learning algorithms to building smart robots (perception, control), text understanding (web search, anti-spam), computer vision, medical informatics, audio, database mining, and other areas.

Visit: https://www.coursera.org/learn/machine-learning

Udacity Courses on Machine Learning

Udacity is a for-profit educational organization that offers Massive Open Online Courses online (MOOCs).

Learn foundational machine learning algorithms, starting with data cleaning and supervised models. Then, move on to exploring deep and unsupervised learning. At each step, get practical experience by applying your skills to code exercises and projects.

Visit: https://www.udacity.com/course/intro-to-machine-learning-nanodegree–nd229

DataQuest

Dataquest.io provides courses on Python, R, SQL, data visualization, data analysis, and machine learning.

Visit: https://www.dataquest.io/

Data Science Masters

They have collected many open-source materials online and have put together lists to learn Data Science, Math, Data Analysis, Python, and many more.

Visit: http://datasciencemasters.org/

Galvanize

The immersive data science curriculum includes a dive into machine learning and working on real problems in classification, regression, and clustering by utilizing structured and unstructured data sets. Students discover libraries like sci-kit-learn, NumPy, and SciPy, and use real-world case studies to root understanding of these libraries to real-world

applications.
Visit: https://www.galvanize.com/data-science

edX courses

This course is provided by Microsoft and forms part of their Professional Program Certificate in Data Science, although it can also be taken as a stand-alone course through EdX. Students are expected to have an “introductory” knowledge of R or Python – the two most popular languages for data science programming at the moment. Subjects covered include probability and statistics, data exploration, visualization, and an introduction to machine learning, using the Microsoft Azure framework. Although all of the course material is free, students can pay ($90 in this case) for an official certificate on completion.
Visit: https://www.edx.org/course/machine-learning-for-data-science-and-analytics

MIT OpenCourseWare

MIT has set up a site that includes all of its courses. It is offered at no cost to participants.
Visit: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-867-machine-learning-fall-2006/index.htm

Google Research Blog

Google researchers publish a variety of papers on topics related to machine learning and deep learning.
Visit: https://ai.googleblog.com/

Medium: Inside Machine Learning

This site gives you deep-dive articles on a wide range of machine learning topics. From weather predihttps://www.kaggle.com/getting-started/148810ctions to robots, you can explore the top machine learning case studies and get insights from industry experts.
Visit: https://medium.com/inside-machine-learning

10 CalTech – Learning from Data

California Institute of Technology provides a course that focuses on machine learning and is delivered as a series of video lectures along with homework assignments and a final exam.

Visit: http://work.caltech.edu/telecourse

Kaggle Wiki

The Kaggle Public Wiki is a resource for learning statistics, machine learning, and other data science concepts. It offers tutorials as well as a platform for data science competitions.
Visit: https://www.kaggle.com/

KDnuggets

KDnuggets is a popular site that provides a vast amount of information on analytics and a variety of information on data science. Check out the content at
Visit: https://www.kdnuggets.com/about/index.html

Data Science Central

Data Science Central is an online site for big data practitioners. It includes a community platform with technical forums for information exchange and technical support.
Visit: https://whttps://www.kaggle.com/getting-started/148810ww.datasciencecentral.com/

Cognitive Class
They provide learning paths for data science beginners to maximize their potential. They have online videos and a virtual lab environment to practice online. These classes are based on an IBM community initiative.
Visit: https://cognitiveclass.ai/
Data Science Weekly
Keep up to date on the latest meetups in your area, or join a virtual meetup featuring data science experts and sharing.
Visit: https://www.datascienceweekly.org/data-science-resources/data-science-meetups

Free Courses Data Science Learning Path from Newbie to Expert
Introduction to Data Science
ABOUT THIS COURSE
Find out the truth about what Data Science is. Hear from real practitioners telling real stories about what it means to work in data science. This course was formerly named Data Science 101.
TIME TO COMPLETE: 3 Hours

COURSE SYLLABUS
Module 1 – Defining Data Science
Module 2 – What do data science people do?
Module 3 – Data Science in Business
Module 4 – Use Cases for Data Science
Module 5 -Data Science People
Sign up: https://cognitiveclass.ai/courses/data-science-101/

Data Science Tools
ABOUT THIS COURSE
Get started with some of the most popular tools for collaborative data science, including RStudio IDE, Jupyter Notebooks, Apache Zeppelin notebooks, and IBM Watson Studio. Use the tools directly on Skills Network Labs, a cloud lab environment that brings powerful open data science tools together so you can analyze, visualize, explore, clean data, run models, and create apps.
TIME TO COMPLETE:4 hours

COURSE SYLLABUS
Module 1 -Introducing Skills Network Labs
Module 2 -Introducing Jupyter Notebooks
Module 3 – Introducing Zeppelin Notebooks
Module 4 – Introducing RStudio IDE
Sign up: https://cognitiveclass.ai/courses/data-science-hands-open-source-tools-2/

Data Science Methodology
ABOUT THIS COURSE

This course has one purpose, and that is to share a methodology that can be used within data science, to ensure that the data used in problem-solving is relevant and properly manipulated to address the question at hand.
Accordingly, in this course, you will learn:

The major steps involved in tackling a data science problem.
The major steps involved in practicing data science, from forming a concrete business or research problem to collecting and analyzing data, building a model, and understanding the feedback after model deployment.
How data scientists think!
TIME TO COMPLETE:5 Hours

AUDIENCE: Data Scientists, Data Engineers, Anyone with interest in Data Science

COURSE SYLLABUS
Module 1: From Problem to Approach
Module 2: From Requirements to Collection
Module 3: From Understanding to Preparation
Module 4: From Modeling to Evaluation
Module 5: From Deployment to Feedback
Sign up: https://cognitiveclass.ai/courses/data-science-methodology-2/

Statistics 101
ABOUT THIS COURSE
Split into five modules, this is a beginner's course covering the fundamentals of statistics. Start with mean, mode, and median. Then learn about standard deviation using examples from basketball. Learn about probability with dice. Learn what it means to group data by categorical variables, and how you can transform your data into appropriate graphs and charts.
In the final module, using an open dataset, learn whether good looking professors indeed get better teaching evaluations.

This course is taught using SPSS Statistics. No prior experience necessary.

TIME TO COMPLETE:6 Hours

AUDIENCE: Beginners in statistics

COURSE SYLLABUS
Module 1 – Welcome to Statistics!

Module 2 – Basic Statistics
Module 3 – Summarizing data
Module 4- Data Visualization
Module 5 – Does Beauty Pay?
Sign up: https://cognitiveclass.ai/courses/statistics-101/

Predictive Modeling Fundamentals I
ABOUT THIS COURSE
In this course, we will be focusing on predictive modeling fundamentals. These are the mathematical algorithms, which are used to "learn" the patterns hidden in data.
Learn the crucial step in the Big Data Lifecycle: using big data to make decisions!

Possess the modeling skills needed by companies all over the world to go beyond storing big data to understanding big data
Learn how to use these skills to make decisions such as cancer detection, fraud detection, customer segmentation, and predicting machine downtime.
Get introduced to the data mining process and modeling techniques using one of the most popular software, IBM's SPSS Modeler.
Learn how to build models on trained data, test the model with historical data, and use qualifying models on live data or other historical untested data.
Save or earn companies millions of dollars with your decisions!
TIME TO COMPLETE:5 Hours

AUDIENCE: Business Analysts, Management Consultants, Data Scientists, and Tech Professionals

COURSE SYLLABUS
Module 1 – Introduction to Data Mining
Module 2 – The Data Mining Process
Module 3 – Modeling Techniques
Module 4 – Model Evaluation
Module 5 – Deployment on IBM Bluemix
Sign up: https://cognitiveclass.ai/courses/predictive-modeling-fundamentals/

Python for Data Science
ABOUT THIS PYTHON COURSE
This introduction to Python will kickstart your learning of Python for data science, as well as programming in general. This beginner-friendly Python course will take you from zero to programming in Python in a matter of hours.
Upon its completion, you'll be able to write your own Python scripts and perform basic hands-on data analysis using our Jupyter-based lab environment. If you want to learn Python from scratch, this free course is for you.

You can start creating your own data science projects and collaborating with other data scientists using IBM Watson Studio. When you sign up, you get free access to Watson Studio. Start now and take advantage of this platform.

TIME TO COMPLETE:5 hours

AUDIENCE: Anyone interested in learning to program with Python for Data Science

COURSE SYLLABUS
Module 1 – Python Basics
Module 2 – Python Data Structures
Module 3 – Python Programming Fundamentals
Module 4 – Working with Data in Python
Sign up: https://cognitiveclass.ai/courses/python-for-data-science/

Data Analysis with Python
ABOUT THE COURSE
Learn how to analyze data using Python. This course will take you from the basics of Python to exploring many different types of data. You will learn how to prepare data for analysis, perform simple statistical analyses, create meaningful data visualizations, predict future trends from data, and more!
You will learn how to:

Import data sets
Clean and prepare data for analysis
Manipulate pandas DataFrame
Summarize data
Build machine learning models using scikit-learn
Build data pipelines
TIME TO COMPLETE:8 hours

AUDIENCE: Anyone who wants to use Python to analyze data

COURSE SYLLABUS
Module 1 – Importing Datasets
Module 2 – Cleaning and Preparing the Data
Module 3 – Summarizing the Data Frame
Module 4 – Model Development
Module 5 – Model Evaluation
Sign up: https://cognitiveclass.ai/courses/data-analysis-python/

Data Visualization with Python

ABOUT THIS DATA VISUALIZATION COURSE

“A picture is worth a thousand words”. We are all familiar with this expression. It especially applies when trying to explain the insight obtained from the analysis of increasingly large datasets. Data visualization plays an essential role in the representation of both small and large-scale data.

One of the key skills of a data scientist is the ability to tell a compelling story, visualizing data, and findings in an approachable and stimulating way. Learning how to leverage a software tool to visualize data will also enable you to extract information, better understand the data, and make more effective decisions. The main goal of this Data Visualization with Python course is to teach you how to take data that at first glance has little meaning and present that data in a form that makes sense to people. Various techniques have been developed for presenting data visually but in this course, we will be using several data visualization libraries in Python, namely Matplotlib, Seaborn, and Folium.

TIME TO COMPLETE:10 hours

AUDIENCE: Anyone interested in data science and has completed Python 101 and Data Analysis with Python

COURSE SYLLABUS

Module 1 – Introduction to Visualization Tools

Module 2 – Basic Visualization Tools

Module 3 – Specialized Visualization Tools

Module 4 – Advanced Visualization Tools

Module 5 – Creating Maps and Visualizing Geospatial Data

Sign up: https://cognitiveclass.ai/courses/data-visualization-with-python/

Machine Learning with Python

ABOUT THIS COURSE

This Machine Learning with Python course dives into the basics of Machine Learning using Python, an approachable and well-known programming language. You’ll learn about Supervised vs Unsupervised Learning, look into how Statistical Modeling relates to Machine Learning, and do a comparison of each.

Look at real-life examples of Machine Learning and how it affects society in ways you may not have guessed!

Explore many algorithms and models:

Popular algorithms: Classification, Regression, Clustering, and Dimensional Reduction.

Popular models: Train/Test Split, Root Mean Squared Error, and Random Forests.

More important, you will transform your theoretical knowledge into practical skills using many hands-https://www.kaggle.com/getting-started/148810on labs.

TIME TO COMPLETE:12 Hours

AUDIENCE: Anyone interested in Machine Learning and Python

COURSE SYLLABUS
Module 1 – Introduction to Machine Learning
Module 2 – Regression
Module 3 – Classification
Module 4 – Unsupervised Learning
Module 5 – Recommender Systems
Sign up: https://cognitiveclass.ai/courses/machine-learning-with-python/
Deep Learning Fundamentals
ABOUT THIS COURSE
Get a crash course on what there is to learn and how to go about learning more. Deep Learning presents a simplified explanation of some of the hottest topics in data science today:
What is Deep Learning?
What are convolutional neural networks?
Why is deep learning so powerful and what can it be used for?
Be part of a rapidly growing field in data science; there's no better time than now to get started with neural networks.
COURSE SYLLABUS
Module 1 – Introduction to Deep Learning
Module 2 – Deep Learning Models
Module 3 – Additional Deep Learning Models
Module 4 – Deep Learning Platforms and Software Libraries
What is a Deep Learning Platform?
H2O.ai
Dato GraphLab
What is a Deep Learning Library?
Theano
Caffe
TensorFlow
Sign up: https://cognitiveclass.ai/courses/introduction-deep-learning/

Deep Learning with TensorFlow
ABOUT THE COURSE
This Deep Learning with TensorFlow course focuses on TensorFlow. If you are new to the subject of deep learning, consider taking our Deep Learning 101 course first.

TensorFlow is one of the best libraries to implement deep learning. TensorFlow is a software library for numerical computation of mathematical expressional, using data flow graphs. Nodes in the graph represent mathematical operations, while the edges represent the multidimensional data arrays (tensors) that flow between them. It was created by Google and tailored for Machine Learning. In fact, it is being widely used to develop solutions with Deep Learning.

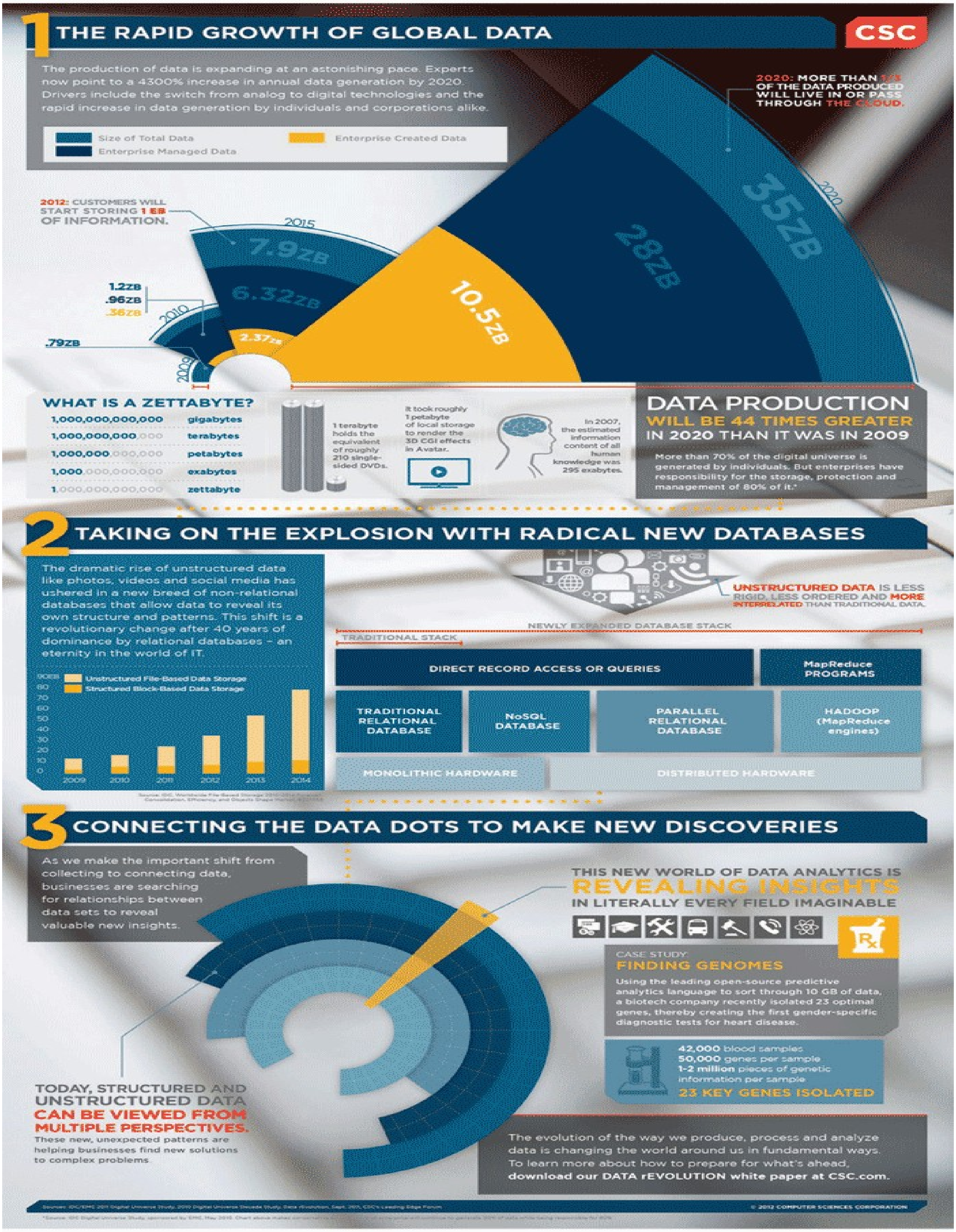
In this TensorFlow course, you will be able to learn the basic concepts of TensorFlow, the main functions, operations, and the execution pipeline. Starting with a simple “Hello World” example, throughout the course you will be able to see how TensorFlow can be used in curve fitting, regression, classification, and minimization of error functions. This concept is then explored in the Deep Learning world. You will learn how to apply TensorFlow for backpropagation to tune the weights and biases while the Neural Networks are being trained. Finally, the course covers different types of Deep Architectures, such as Convolutional Networks, Recurrent Networks, and Autoencoders.

TIME TO COMPLETE:10 Hours

AUDIENCE: Anyone interested in Machine Learning, Deep Learning, and TensorFlow

COURSE SYLLABUS
Module 1 – Introduction to TensorFlow
Module 2 – Convolutional Neural Networks (CNN)
Module 3 – Recurrent Neural Networks (RNN)
Module 4 – Unsupervised Learning
Module 5 – Autoencoders
Sign up: https://cognitiveclass.ai/courses/deep-learning-tensorflow/

Scope of Data Science & Rise of Data sources



CSC

1 THE RAPID GROWTH OF GLOBAL DATA

The production of data is expanding at an astonishing pace. Experts now point to a 4300% increase in annual data generation by 2020. Drivers include the switch from analog to digital technologies and the rapid increase in data generation by individuals and corporations alike.

WHAT IS A ZETTABYTE?

1,000,000,000,000	gigabytes
1,000,000,000,000	terabytes
1,000,000,000,000	petabytes
1,000,000,000,000	exabytes
1,000,000,000,000	zettabyte

1 terabyte holds the equivalent of roughly 210 single-sided DVDs.

It took roughly 1 petabyte of local storage to render the 3D CGI effects in Avatar.

In 2007, the estimated information content of all human knowledge was 295 exabytes.

DATA PRODUCTION WILL BE 44 TIMES GREATER IN 2020 THAN IT WAS IN 2009

More than 70% of the digital universe is generated by individuals. But enterprises have responsibility for the storage, protection and management of 80% of it.*

2 TAKING ON THE EXPLOSION WITH RADICAL NEW DATABASES

The dramatic rise of unstructured data like photos, videos and social media has ushered in a new breed of non-relational databases that allow data to reveal its own structure and patterns. This shift is a revolutionary change after 40 years of dominance by relational databases – an eternity in the world of IT.

90EB
80
70
60
50
40
30
20
10
0
Unstructured File-Based Data Storage
Structured Block-Based Data Storage
2009
2010
2011
2012
2013
2014

UNSTRUCTURED DATA IS LESS RIGID, LESS ORDERED AND MORE INTERRELATED THAN TRADITIONAL DATA.

NEWLY EXPANDED DATABASE STACK
TRADITIONAL STACK
DIRECT RECORD ACCESS OR QUERIES
MapReduce PROGRAMS
TRADITIONAL RELATIONAL DATABASE
NoSQL DATABASE
PARALLEL RELATIONAL DATABASE
HADOOP (MapReduce engines)
MONOLITHIC HARDWARE
DISTRIBUTED HARDWARE

3 CONNECTING THE DATA DOTS TO MAKE NEW DISCOVERIES

As we make the important shift from collecting to connecting data, businesses are searching for relationships between data sets to reveal valuable new insights.

THIS NEW WORLD OF DATA ANALYTICS IS REVEALING INSIGHTS IN LITERALLY EVERY FIELD IMAGINABLE

CASE STUDY:
FINDING GENOMES

Using the leading open-source predictive analytics language to sort through 10 GB of data, a biotech company recently isolated 23 optimal genes, thereby creating the first gender-specific diagnostic tests for heart disease.

42,000 blood samples
50,000 genes per sample
1-2 million pieces of genetic information per sample
23 KEY GENES ISOLATED

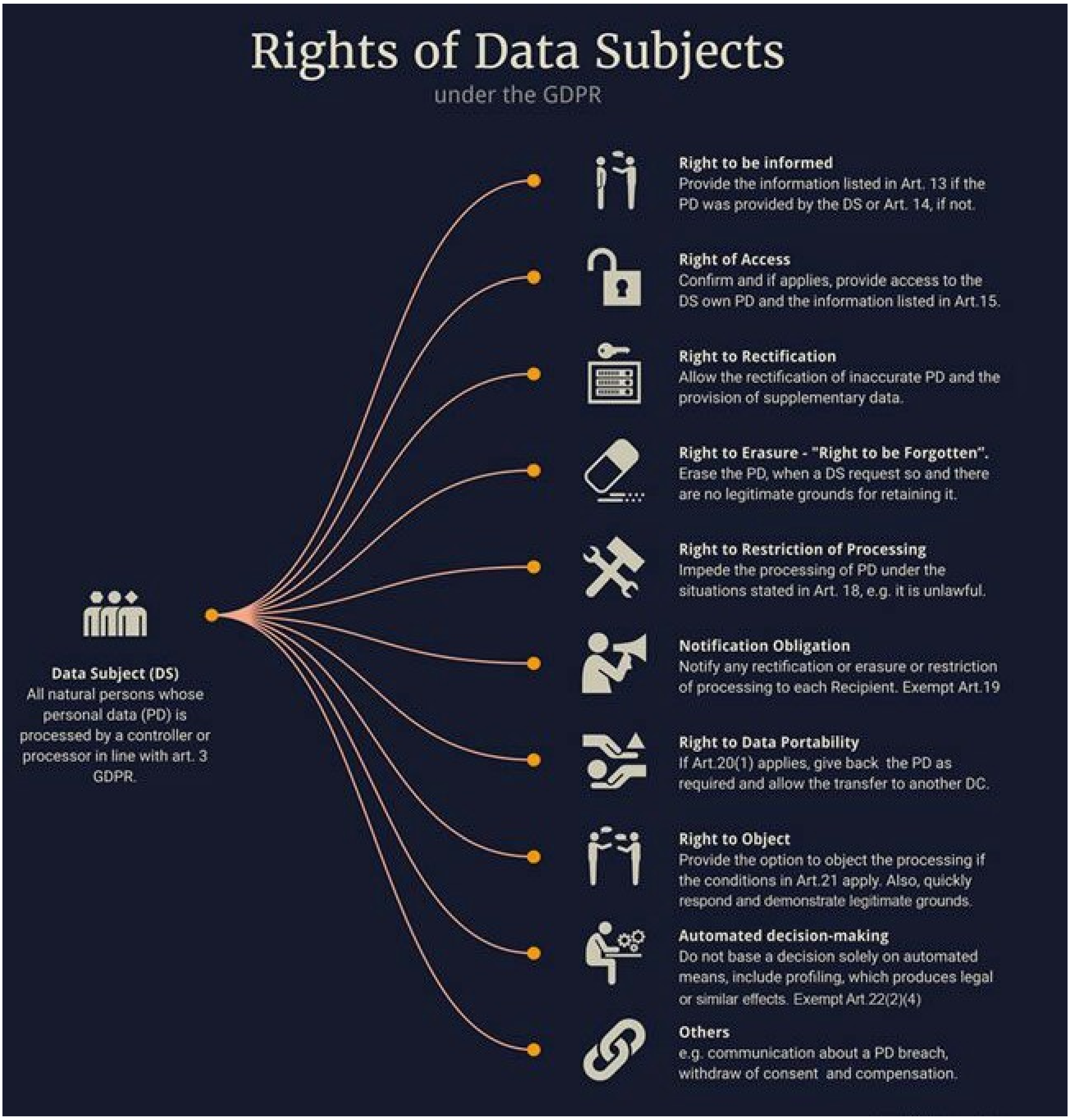
TODAY, STRUCTURED AND UNSTRUCTURED DATA CAN BE VIEWED FROM MULTIPLE PERSPECTIVES.
These new, unexpected patterns are helping businesses find new solutions to complex problems

The evolution of the way we produce, process and analyze data is changing the world around us in fundamental ways. To learn more about how to prepare for what's ahead, **download our DATA rEVOLUTION white paper at CSC.com.**

Sources: IDC/EMC 2011 Digital Universe Study, 2010 Digital Universe Decade Study, Data rEvolution, Sept. 2011, CSC's Leading Edge Forum

© 2012 COMPUTER SCIENCES CORPORATION

Rights of Data Science



Rights of Data Subjects

under the GDPR

Data Subject (DS)
All natural persons whose personal data (PD) is processed by a controller or processor in line with art. 3 GDPR.

Right to be informed
Provide the information listed in Art. 13 if the PD was provided by the DS or Art. 14, if not.

Right of Access
Confirm and if applies, provide access to the DS own PD and the information listed in Art.15.

Right to Rectification
Allow the rectification of inaccurate PD and the provision of supplementary data.

Right to Erasure - "Right to be Forgotten".
Erase the PD, when a DS request so and there are no legitimate grounds for retaining it.

Right to Restriction of Processing
Impede the processing of PD under the situations stated in Art. 18, e.g. it is unlawful.

Notification Obligation
Notify any rectification or erasure or restriction of processing to each Recipient. Exempt Art.19

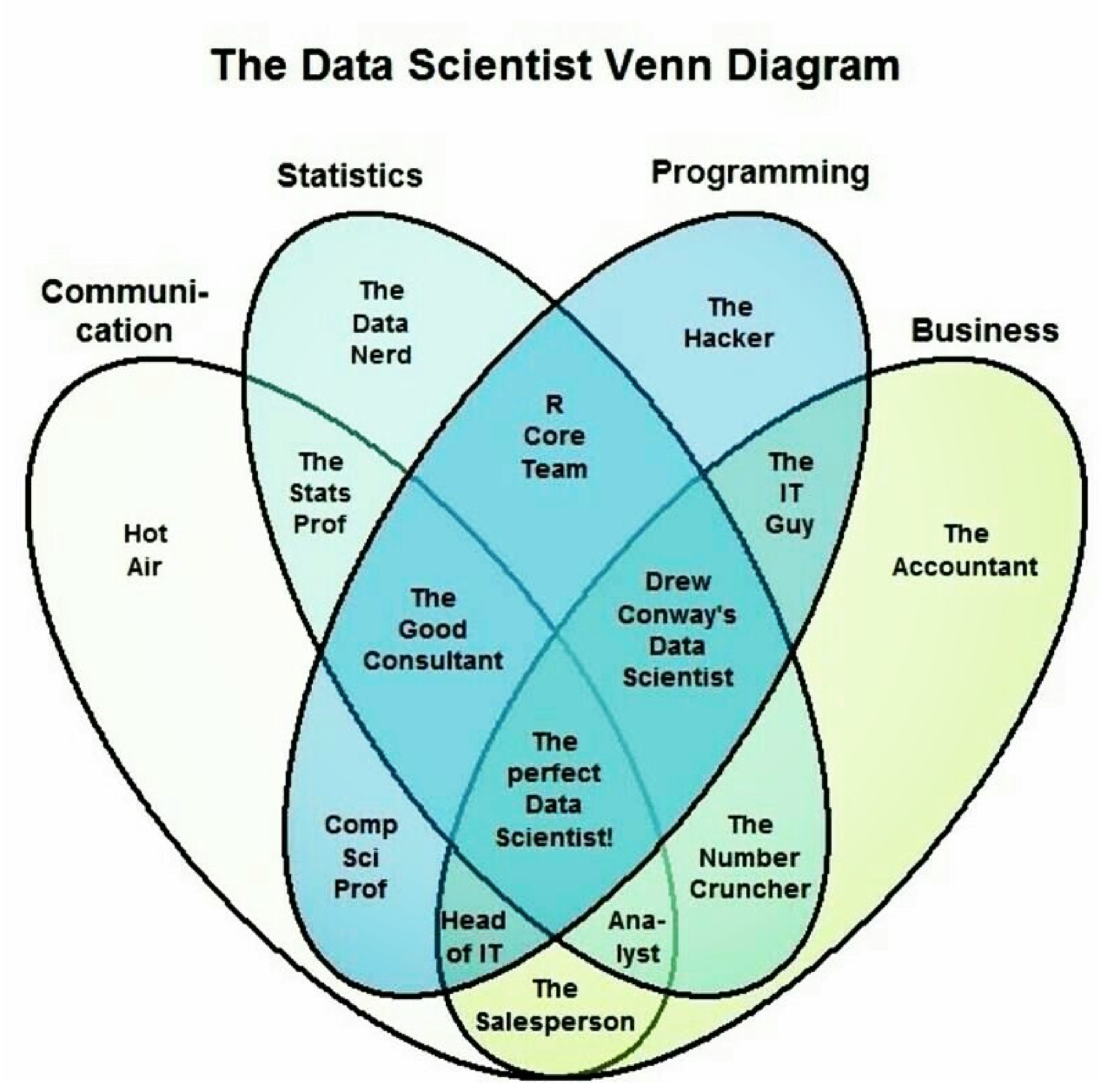
Right to Data Portability
If Art.20(1) applies, give back the PD as required and allow the transfer to another DC.

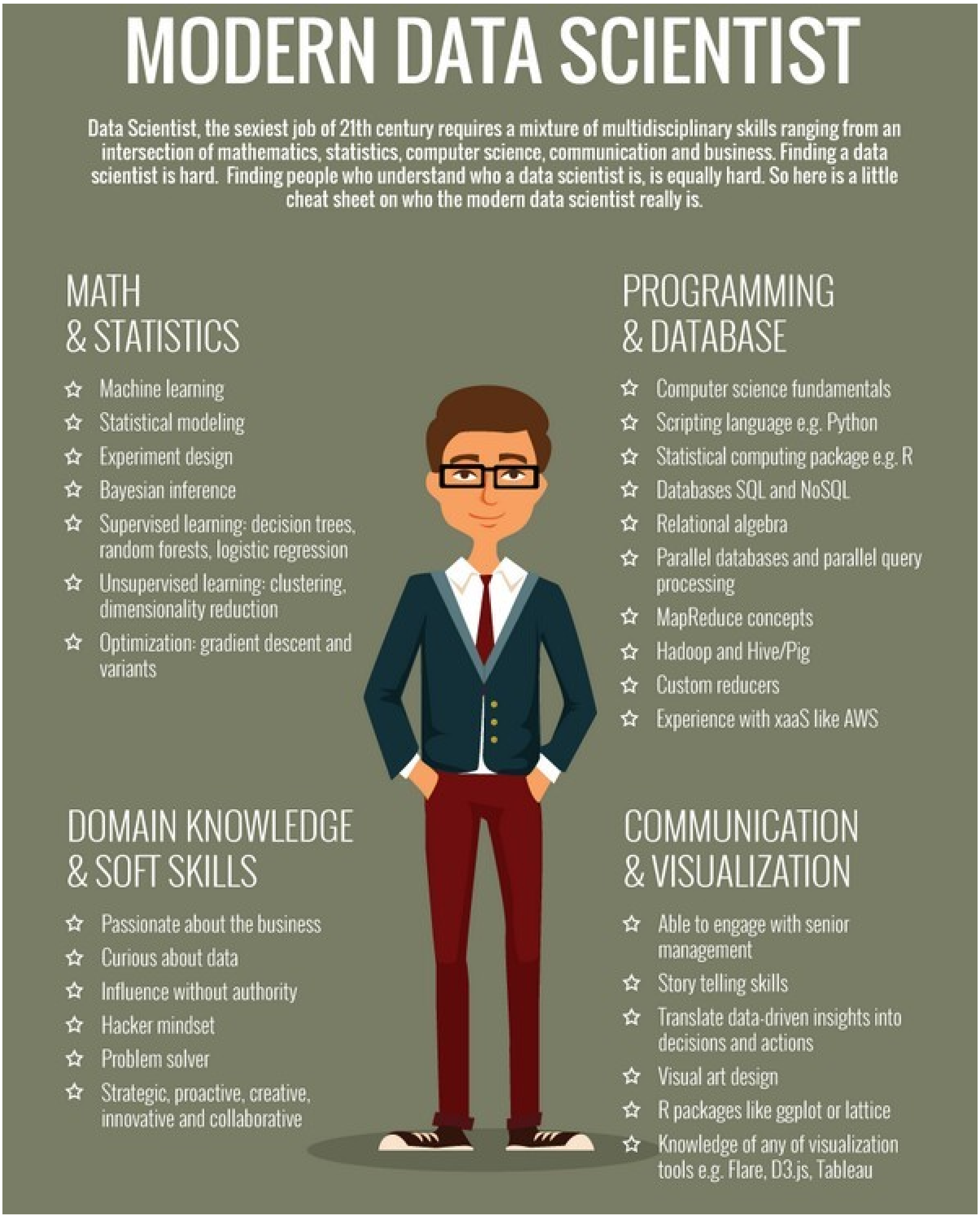
Right to Object
Provide the option to object the processing if the conditions in Art.21 apply. Also, quickly respond and demonstrate legitimate grounds.

Automated decision-making
Do not base a decision solely on automated means, include profiling, which produces legal or similar effects. Exempt Art.22(2)(4)

Others
e.g. communication about a PD breach, withdraw of consent and compensation.

Modern Data Scientist





MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

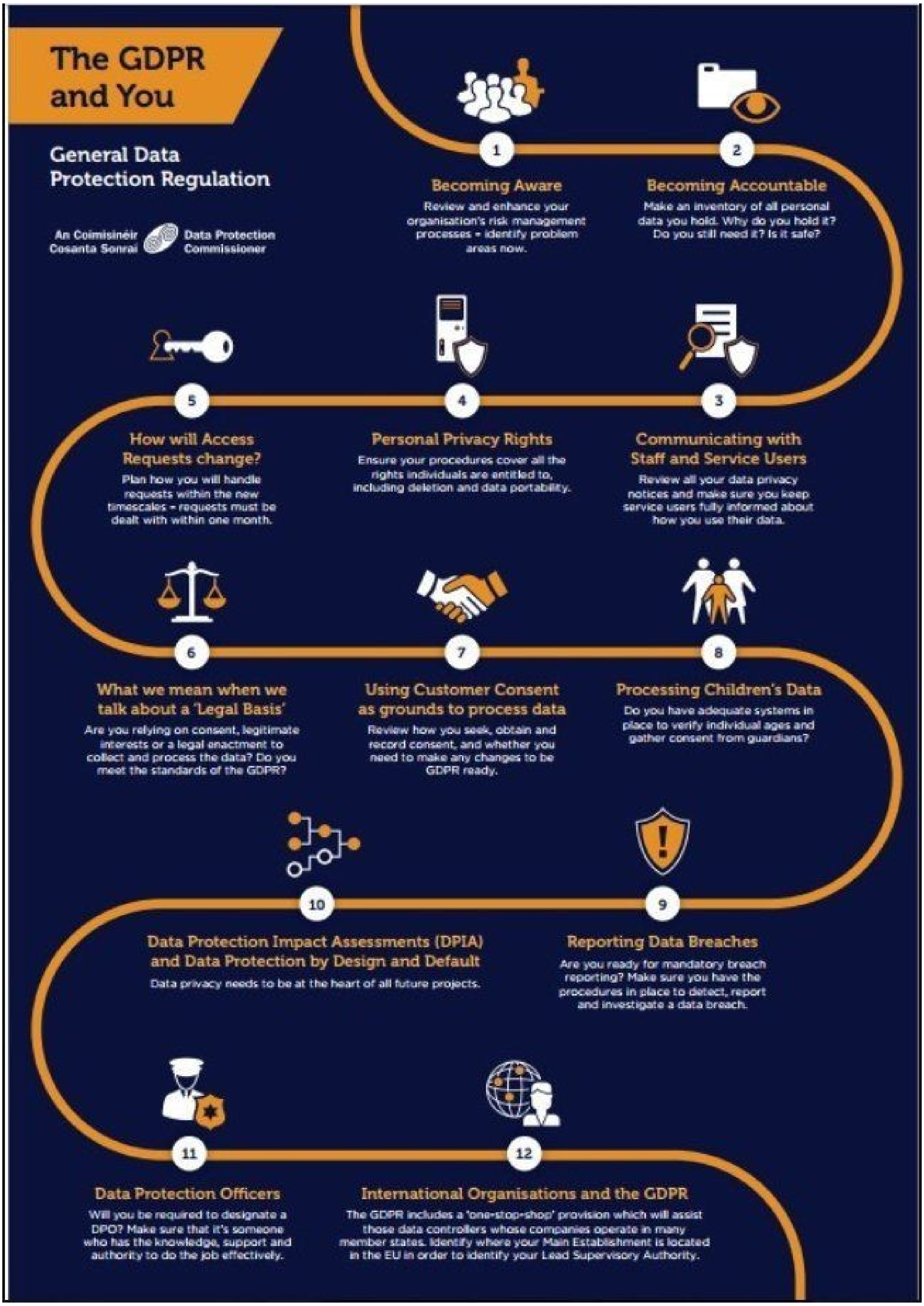
DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Regulations & Data Scientist



The GDPR and You

General Data Protection Regulation

An Coimisinéir Cosanta Sonraí
Data Protection Commissioner

1 **Becoming Aware**
Review and enhance your organisation's risk management processes – identify problem areas now.

2 **Becoming Accountable**
Make an inventory of all personal data you hold. Why do you hold it? Do you still need it? Is it safe?

3 **Communicating with Staff and Service Users**
Review all your data privacy notices and make sure you keep service users fully informed about how you use their data.

4 **Personal Privacy Rights**
Ensure your procedures cover all the rights individuals are entitled to, including deletion and data portability.

5 **How will Access Requests change?**
Plan how you will handle requests within the new timescales – requests must be dealt with within one month.

6 **What we mean when we talk about a 'Legal Basis'**
Are you relying on consent, legitimate interests or a legal enactment to collect and process the data? Do you meet the standards of the GDPR?

7 **Using Customer Consent as grounds to process data**
Review how you seek, obtain and record consent, and whether you need to make any changes to be GDPR ready.

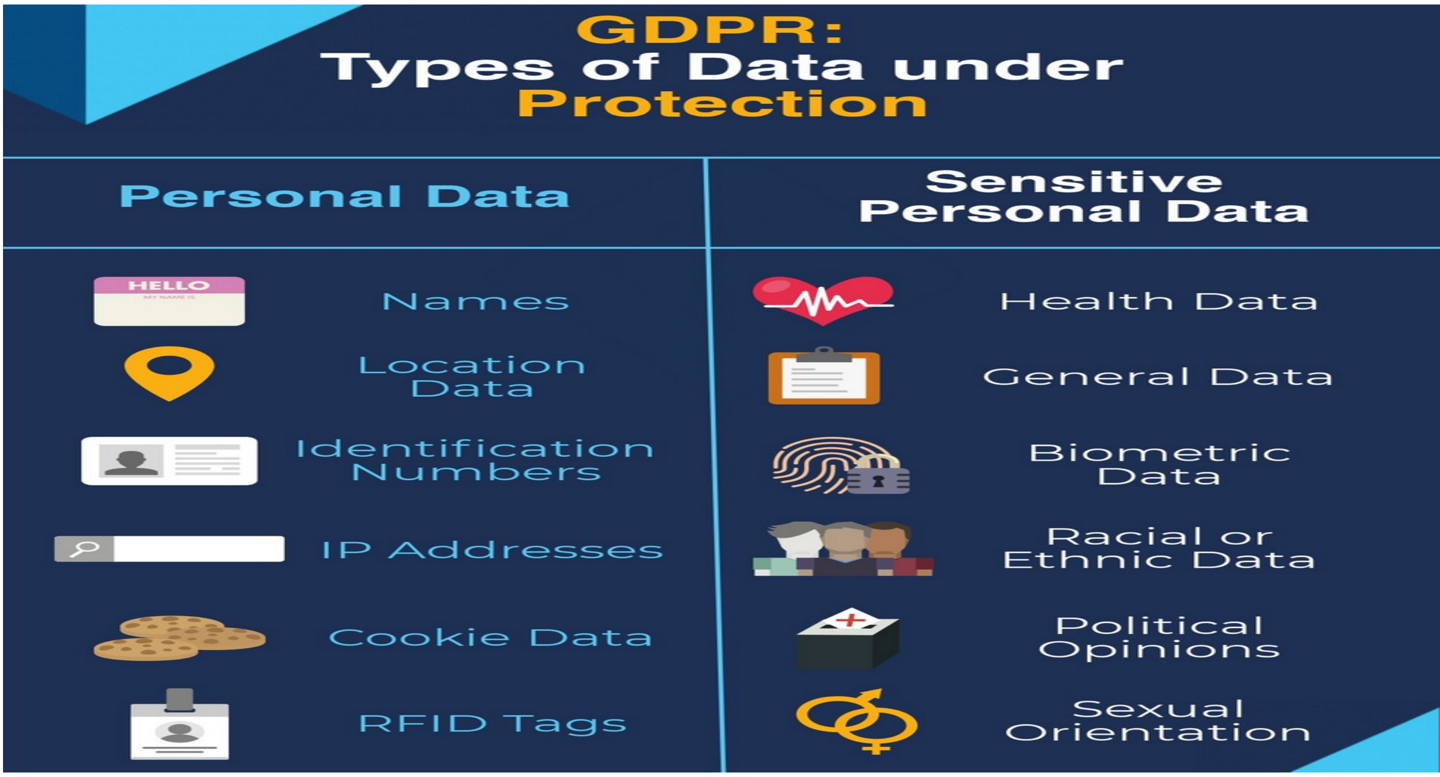
8 **Processing Children's Data**
Do you have adequate systems in place to verify individual ages and gather consent from guardians?

9 **Reporting Data Breaches**
Are you ready for mandatory breach reporting? Make sure you have the procedures in place to detect, report and investigate a data breach.

10 **Data Protection Impact Assessments (DPIA) and Data Protection by Design and Default**
Data privacy needs to be at the heart of all future projects.

11 **Data Protection Officers**
Will you be required to designate a DPO? Make sure that it's someone who has the knowledge, support and authority to do the job effectively.

12 **International Organisations and the GDPR**
The GDPR includes a 'one-stop-shop' provision which will assist those data controllers whose companies operate in many member states. Identify where your Main Establishment is located in the EU in order to identify your Lead Supervisory Authority.



GDPR: Types of Data under Protection

Personal Data	Sensitive Personal Data
Names	Health Data
Location Data	General Data
Identification Numbers	Biometric Data
IP Addresses	Racial or Ethnic Data
Cookie Data	Political Opinions
RFID Tags	Sexual Orientation

HELLO
MY NAME IS

General Data Protection Regulation (GDPR) in practice:

Privacy by Design

Guidance on how to define and implement privacy by design.

⚠ **Non-compliance with GDPR can result in fines up to €20 million or 4% of global turnover.** ⚠

> *Privacy by design is an approach to projects that promotes privacy and data protection compliance from the start. Under the GDPR, you have a general obligation to implement technical and organisational measures to show that you have considered and integrated data protection into your processing activities. Privacy should be 'built-in' to everything your organisation does and not bolted on as an afterthought.*

The ICO encourages organisations to ensure that privacy and data protection is a key consideration in the early stages of any project, and then throughout its lifecycle.

For example when:

- Building new IT systems for storing or accessing personal data.
- Developing legislation, policies or strategies that have privacy implications.
- Embarking on a data sharing initiative.
- Using data for new purposes.

Some of the benefits of a 'privacy by design approach' are:

- Potential problems are identified at an early stage, when addressing them will often be simpler and less costly.
- Increased awareness of privacy and data protection across an organisation.
- Organisations are more likely to meet their legal obligations and less likely to breach GDPR legislation.
- Actions are less likely to be privacy intrusive and have a negative impact on individuals.

What this could mean in practice:

- Conduct a data protection impact assessment (DPIA) to help you identify and minimise the data protection risks for each project. You should ensure these are documented. You must complete a DPIA for certain listed types of processing, or any other processing that is likely to result in a high risk to individuals' interests.
- The GDPR endorses the use of approved codes of conduct and certification mechanisms to demonstrate that you comply.

  Beyond compliance, these schemes can also:

  - Improve transparency and accountability.
  - Enable individuals to distinguish the organisations that meet the requirements of the law and they can trust with their personal data.
  - Provide mitigation against enforcement action.
  - Improve standards by establishing best practice.
- Consider appointing a Data Protection Officer (DPO) to assist you to monitor internal compliance, inform and advise on your data protection obligations, provide advice regarding Data Protection Impact Assessments (DPIAs) and act as a contact point for data subjects and the supervisory authority.
- DPOs must be independent, an expert in data protection, adequately resourced, and report to the highest management level. They can be an existing employee or externally appointed. There are certain criteria whereby you must appoint a DPO.

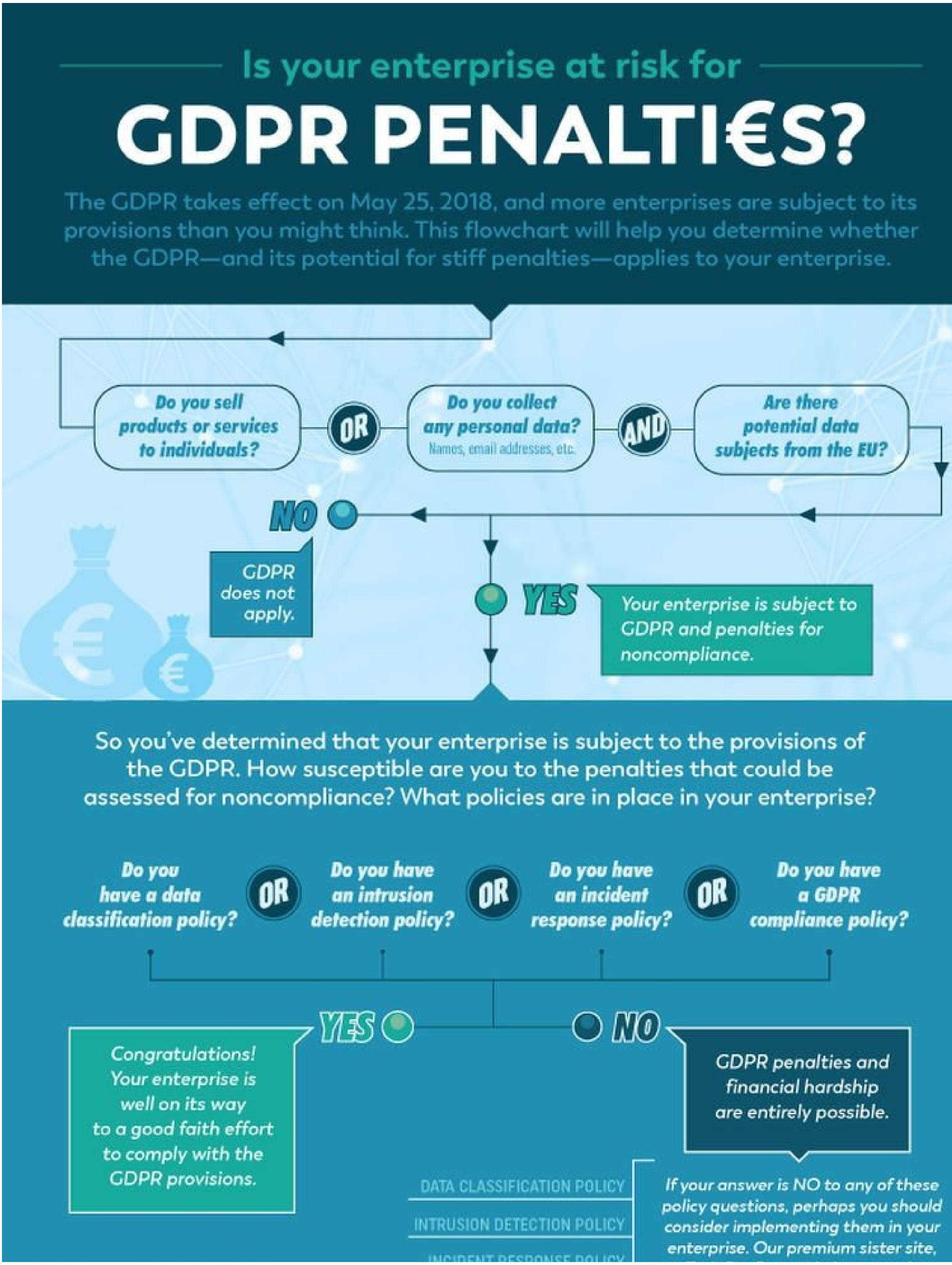
The Data Protection Officer (DPO) or relevant individual in your organisation will be able to inform you about privacy by design in your organisation.

Name of DPO or relevant individual:

Intended for unofficial guidance only.

For the complete official legislation and guidance on the General Data Protection Regulation (GDPR) visit **ico.org.uk**

GDPR
IN PRACTICE



Is your enterprise at risk for

GDPR PENALTI€S?

The GDPR takes effect on May 25, 2018, and more enterprises are subject to its provisions than you might think. This flowchart will help you determine whether the GDPR—and its potential for stiff penalties—applies to your enterprise.

Do you sell products or services to individuals?
OR
Do you collect any personal data?
Names, email addresses, etc.
AND
Are there potential data subjects from the EU?
NO
GDPR does not apply.
YES
Your enterprise is subject to GDPR and penalties for noncompliance.

So you've determined that your enterprise is subject to the provisions of the GDPR. How susceptible are you to the penalties that could be assessed for noncompliance? What policies are in place in your enterprise?

Do you have a data classification policy?
OR
Do you have an intrusion detection policy?
OR
Do you have an incident response policy?
OR
Do you have a GDPR compliance policy?
YES
Congratulations! Your enterprise is well on its way to a good faith effort to comply with the GDPR provisions.
NO
GDPR penalties and financial hardship are entirely possible.
DATA CLASSIFICATION POLICY
INTRUSION DETECTION POLICY
INCIDENT RESPONSE POLICY

If your answer is NO to any of these policy questions, perhaps you should consider implementing them in your enterprise. Our premium sister site,

General Data Protection Regulation (GDPR) in practice:

Glossary of terms for general understanding

⚠ **Non-compliance with GDPR can result in fines up to €20 million or 4% of global turnover.** ⚠

Accountability
One of the key data protection principles under GDPR - it makes an organisation responsible for complying with GDPR and says that an organisation must be able to demonstrate it's compliance. Accountability obligations are ongoing.

Consent
One of six lawful basis for processing personal data. Consent means offering individuals real choice and control over use of their personal data before it is collected or used.

Criminal Offence Data
Personal data about criminal convictions or offences. Organisations must have both a lawful basis under Article 6 and either legal authority or official authority for the processing under Article 10 of the GDPR.

Data Breach
Breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data. This includes breaches that are the result of both accidental and deliberate causes.

Data Controller/Controller
A controller determines the purposes and means of processing personal data.

Data Destruction Policy
A clear and simple document available to everyone in an organisation which explains your processes and measures for disposing of personal data.

Data Map
A document which shows how various types of data moves through an organisation along with how and why it is processed.

Data Minimisation
A GDPR principle which states that an organisation only holds personal data about an individual that is sufficient for the purpose it is holding it for in relation to that individual, and that it does not hold more information than it needs for that purpose.

DPA (Data Protection Act)
Previous data protection legislation to be replaced with the General Data Protection Regulations (GDPR).

Data Processor/Processor
A processor is responsible for processing personal data on behalf of a controller.

Data Protection Impact Assessment (DPIA)
A documented process to help you identify and minimise the data protection risks of a project prior to commencement.

Data Protection Officer (DPO)
An internal or external data protection expert to assist and monitor internal compliance, inform and advise on data protection obligations, provide advice regarding Data Protection Impact Assessments (DPIAs) and act as a contact point for data subjects and the ICO.

Data Retention Policy
A document which explains how long to retain different types of personal data for, for what purpose, and what happens to it when it is no longer to be retained. Personal data processed for any purpose(s) should not be kept for longer than is necessary for that purpose(s).

Data Retention Schedule
Often a pre-cursor to a data retention policy. It should list the types of data held and what the retention periods are for that data. It may not explain or justify as a data retention policy does, but it can demonstrate the thought process behind data retention generally.

Data Subject
Identified or identifiable natural person(s).

Due Diligence
Taking satisfactorily thorough steps to ensure that you have acted to protect the rights and freedoms of a data subject. For example, exercising 'due diligence' with a 3rd party may involve ensuring as far as it is reasonable that they operate in a GDPR compliant way before sharing or receiving any personal data with/from them.

Encryption
A mathematical function using a secret value - the key - which encodes data so that only users with access to that key can read the information. It may be an effective technical measure to minimise access to personal data within an organisation.

GDPR (General Data Protection Regulations)
Updated EU data protection legislation to replace the Data Protection Act (DPA).

ICO (Information Commissioners Office)
The supervisory authority enforcing GDPR within the UK.

Internal Data Breach Register
An internal log of data breaches which may be part of internal breach reporting procedures. It should document details of any breach and the likelihood/severity of the resulting risk to people's rights and freedoms. It should also note whether the breach has been reported to the ICO or the individual.

Lawful Basis
A valid basis for the processing of personal data under GDPR. There are six available lawful bases for processing. No single basis is 'better' or more important than the others – which basis is most appropriate to use will depend on your purpose and relationship with the individual.

Legal Obligation
A lawful basis for processing personal data whereby the processing of said data is necessary to comply with the law (not including contractual obligations).

Legitimate Interest
A lawful basis for processing personal data whereby the processing is necessary for your legitimate interests or the legitimate interests of a third party, providing they are in balance with an individual's rights and reasonable expectations.

Legitimate Interests Assessment
A document justifying reliance on legitimate interest as a lawful basis. An organisation's interests must balance against the individual's. If they would not reasonably expect the processing, or if it would cause unjustified harm, their interests are likely to override the legitimate interests of an organisation.

Personal Data
Any information relating to an identifiable person who can be directly or indirectly identified in particular by reference to an identifier.

Privacy By Design
An approach to projects that promotes privacy and data protection compliance from the start. Under GDPR, privacy should be 'built-in' to everything an organisation does and not bolted on as an afterthought.

Privacy Impact Assessment (PIA)
Superseded by Data Protection Impact Assessments under GDPR. A process which helps an organisation to identify and reduce the privacy risks of a project. An effective PIA will be used throughout the development and implementation of a project.

Privacy Policy
A concise, transparent, and easily accessible document which clearly explains how it intends to process personal data. It must be written in clear and plain language.

Processing
Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

Profiling
The use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

Pseudonymisation
A data management and de-identification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.

Public Tasks
A lawful basis for processing whereby the processing is necessary to perform a task in the public interest or for official functions, and the task or function has a clear basis in law.

Purpose
The reason why an organisation processes personal data.

Purpose Limitation
Once an organisation specifies the purpose or purposes for obtaining personal data, anything done with this data must be compatible with this purpose. This is a purpose limitation. It ensures that personal data is processed in line within reasonable expectations of the individual concerned.

Register of Personal Data
A log of what personal data is held within an organisation, for what purpose, and details of who has access.

Register of Processing Activities
A record of processing activities, covering areas such as processing purposes, data sharing and retention.

Right of Access
Individuals have the right to access their personal data and supplementary information. The right of access allows individuals to be aware of the nature and verify the lawfulness of the processing.

Right to be Informed
Individuals have the right to be informed about the collection and use of their personal data. This is a key transparency requirement under the GDPR.

Right to Data Portability
The right to data portability allows individuals to obtain and reuse their personal data for their own purposes across different services.

Right to Erasure
The GDPR introduces a right for individuals to have personal data erased. The right to erasure is also known as 'the right to be forgotten'.

Right to Object
Individuals have the right to object to processing based on legitimate interests or the performance of a task in the public interest/exercise of official authority (including profiling), direct marketing and processing for research purposes.

Right to Rectification
The GDPR includes a right for individuals to have inaccurate personal data rectified or completed if it is incomplete.

Right to Restrict Processing
Individuals have the right to request the restriction or suppression of their personal data.

Special Category Data
Sensitive personal data which could create more significant risks to a person's fundamental rights and freedoms. For example, by putting them at risk of unlawful discrimination. As such there are additional conditions in place for processing Special Category Data under GDPR.

Storage Limitation
Organisations are limited to storing personal data only for as long as it is needed for the purpose(s) for which it was collected. After this, it should be deleted unless there are other grounds for retaining it.

Subject Access Request
See 'Right of Access'. When individuals ask for details of what personal data is being held on them and how it is being processed, this is a 'Subject Access Request'. Organisations should not charge for this (unless requests are repetitive) and should respond within one month.

Supervisory Authority
The authority enforcing GDPR in a territory. In the case of the UK, it is the Information Commissioners Office (ICO).

Suppression
Restricting processing of personal data for a given purpose. For example - unsubscribing from emails may result in the suppression of personal data for the purposes of direct marketing.

Territorial Scope
The territory to which GDPR applies. This is not necessarily just within the EU - it also applies to processing outside of the EU if data is being processed in the context of the activities of an establishment of a data controller or a data processor in the EU.

Third Party
A natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data.

Transfer
The transfer of personal data across organisations or to outside of the European Union.

Vital Interests
A lawful basis for processing personal data whereby if it is required to process personal data to protect someone's life.

The Data Protection Officer (DPO) or relevant individual in your organisation is:

Intended for unofficial guidance only.

For the complete official legislation and guidance on the General Data Protection Regulation (GDPR) visit **ico.org.uk**

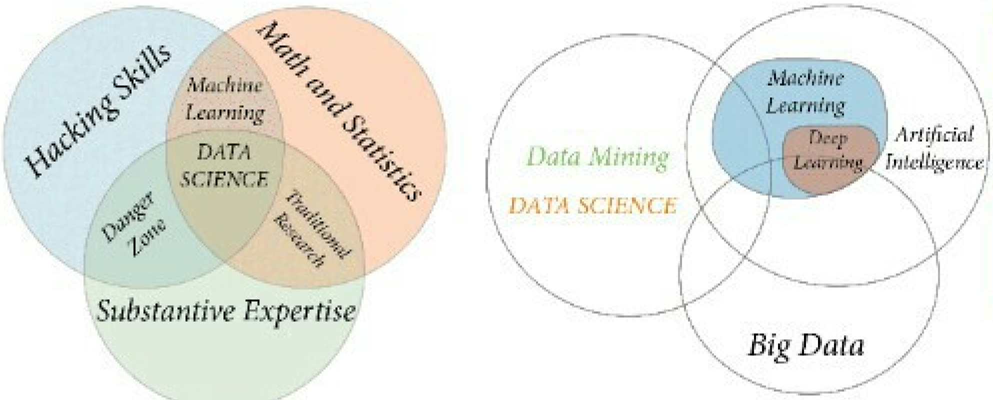
GDPR
IN PRACTICE

Steps for Beginner

LEARN DATA SCIENCE

8 (Easy) Steps

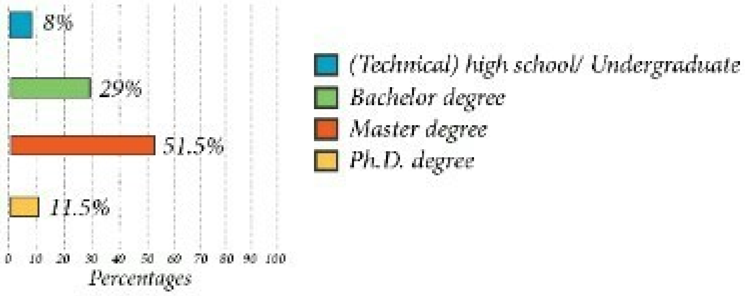
What's Data Science?



2010

2015

Data Scientists' Educational Background



"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician"
- Josh Wills

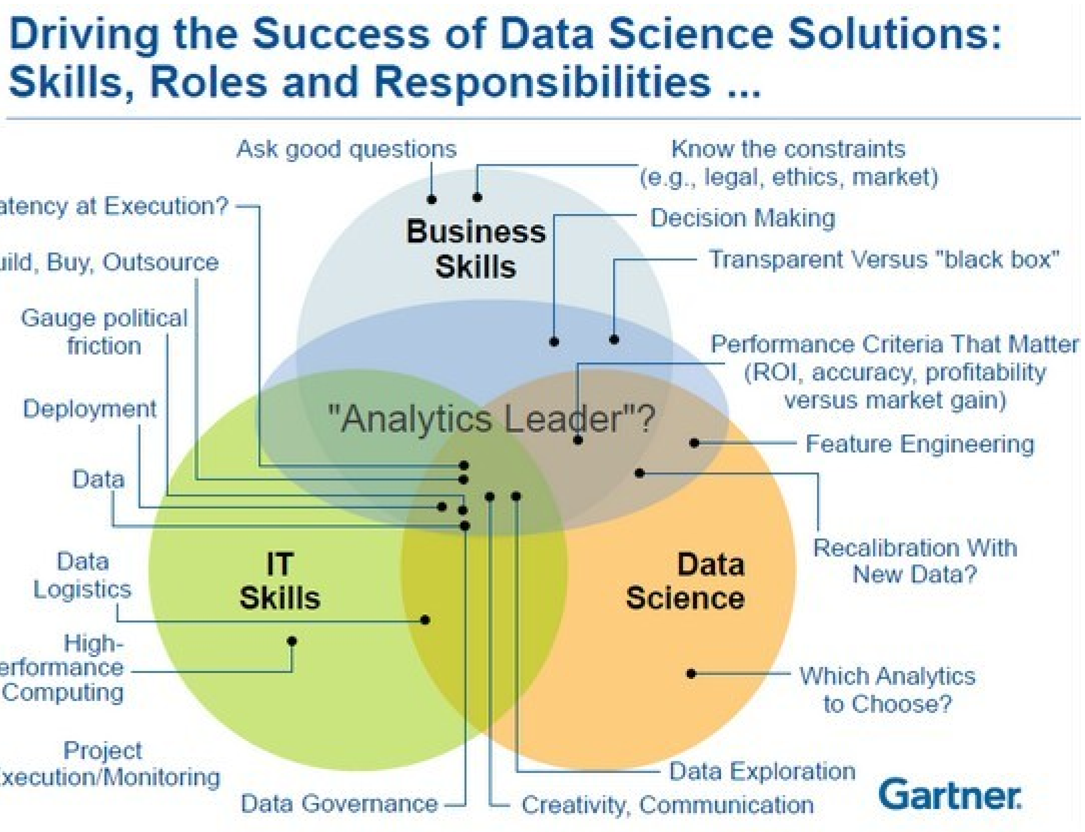
1 Get Good At Stats, Maths and Machine Learning

Math

Stats

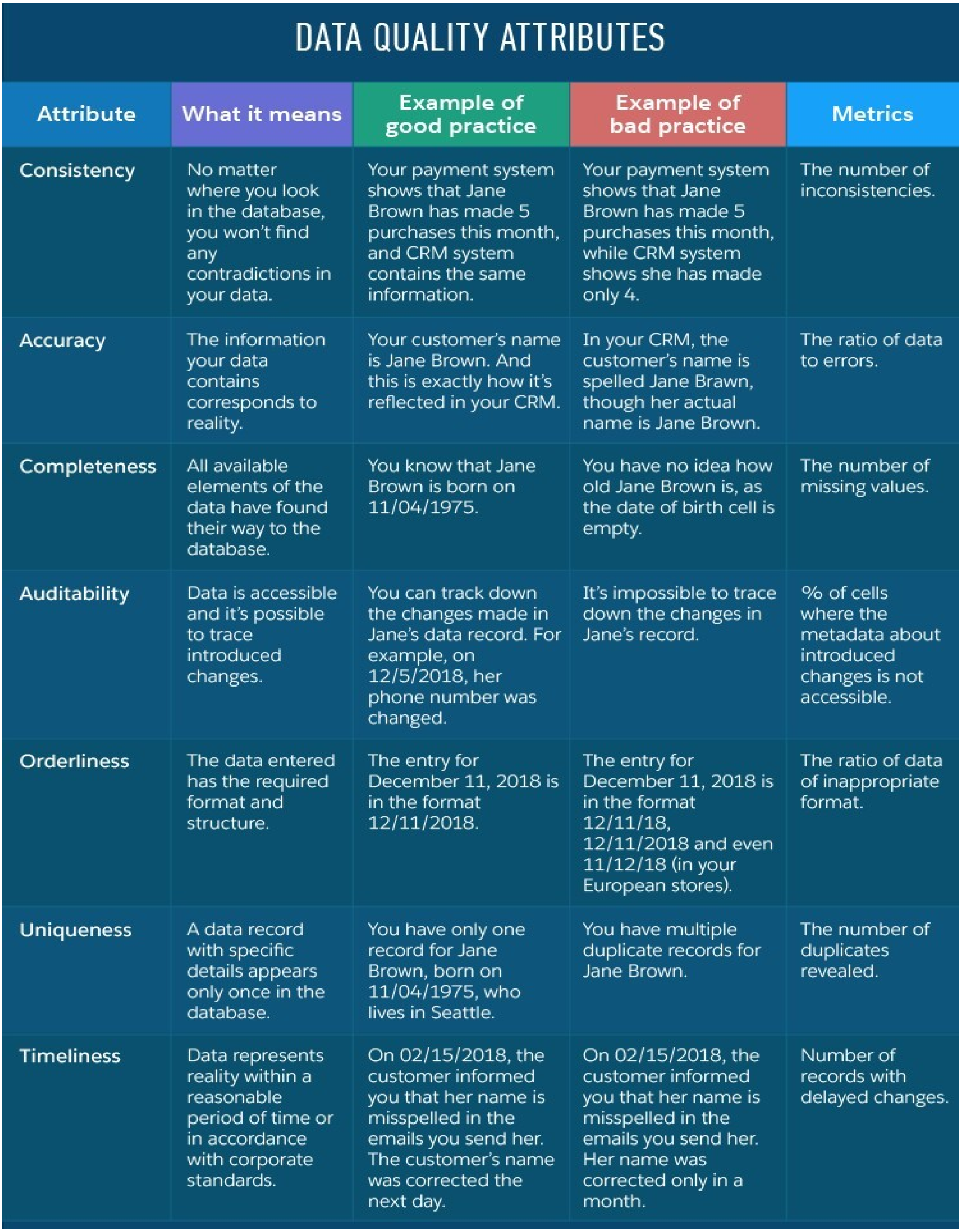
ML

Responsibilities



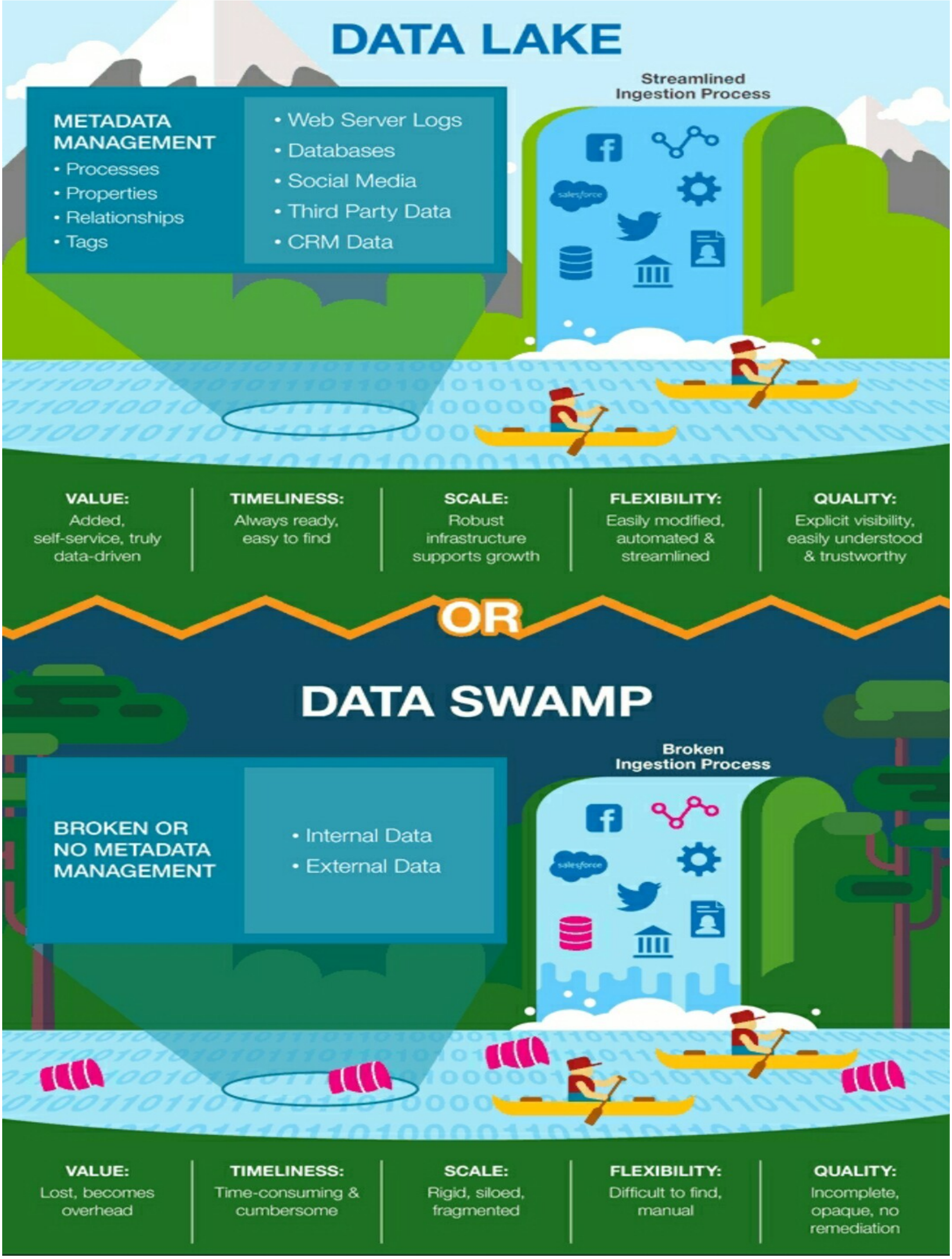
Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...

Ask good questions
Know the constraints (e.g., legal, ethics, market)
Latency at Execution?
Business Skills
Decision Making
Build, Buy, Outsource
Transparent Versus "black box"
Gauge political friction
Performance Criteria That Matter (ROI, accuracy, profitability versus market gain)
Deployment
"Analytics Leader"?
Feature Engineering
Data
Recalibration With New Data?
Data Logistics
IT Skills
Data Science
High-performance Computing
Which Analytics to Choose?
Project Execution/Monitoring
Data Exploration
Data Governance
Creativity, Communication
Gartner.



DATA QUALITY ATTRIBUTES

Attribute	What it means	Example of good practice	Example of bad practice	Metrics
Consistency	No matter where you look in the database, you won't find any contradictions in your data.	Your payment system shows that Jane Brown has made 5 purchases this month, and CRM system contains the same information.	Your payment system shows that Jane Brown has made 5 purchases this month, while CRM system shows she has made only 4.	The number of inconsistencies.
Accuracy	The information your data contains corresponds to reality.	Your customer's name is Jane Brown. And this is exactly how it's reflected in your CRM.	In your CRM, the customer's name is spelled Jane Brawn, though her actual name is Jane Brown.	The ratio of data to errors.
Completeness	All available elements of the data have found their way to the database.	You know that Jane Brown is born on 11/04/1975.	You have no idea how old Jane Brown is, as the date of birth cell is empty.	The number of missing values.
Auditability	Data is accessible and it's possible to trace introduced changes.	You can track down the changes made in Jane's data record. For example, on 12/5/2018, her phone number was changed.	It's impossible to trace down the changes in Jane's record.	% of cells where the metadata about introduced changes is not accessible.
Orderliness	The data entered has the required format and structure.	The entry for December 11, 2018 is in the format 12/11/2018.	The entry for December 11, 2018 is in the format 12/11/18, 12/11/2018 and even 11/12/18 (in your European stores).	The ratio of data of inappropriate format.
Uniqueness	A data record with specific details appears only once in the database.	You have only one record for Jane Brown, born on 11/04/1975, who lives in Seattle.	You have multiple duplicate records for Jane Brown.	The number of duplicates revealed.
Timeliness	Data represents reality within a reasonable period of time or in accordance with corporate standards.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. The customer's name was corrected the next day.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. Her name was corrected only in a month.	Number of records with delayed changes.



DATA LAKE

Streamlined Ingestion Process

METADATA MANAGEMENT

- Processes
- Properties
- Relationships
- Tags

- Web Server Logs
- Databases
- Social Media
- Third Party Data
- CRM Data

VALUE:	TIMELINESS:	SCALE:	FLEXIBILITY:	QUALITY:
Added, self-service, truly data-driven	Always ready, easy to find	Robust infrastructure supports growth	Easily modified, automated & streamlined	Explicit visibility, easily understood & trustworthy

OR

DATA SWAMP

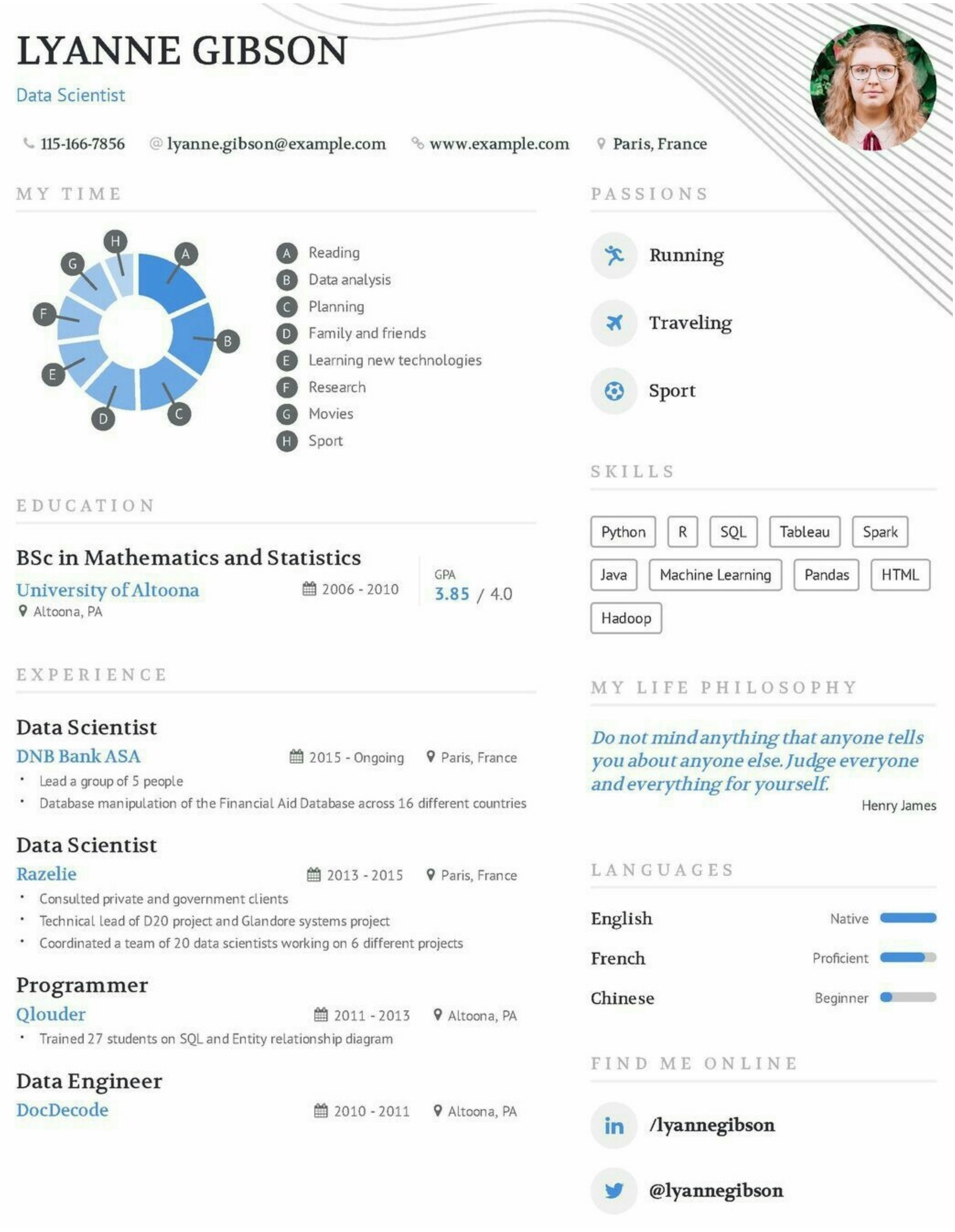
Broken Ingestion Process

BROKEN OR NO METADATA MANAGEMENT

- Internal Data
- External Data

VALUE:	TIMELINESS:	SCALE:	FLEXIBILITY:	QUALITY:
Lost, becomes overhead	Time-consuming & cumbersome	Rigid, siloed, fragmented	Difficult to find, manual	Incomplete, opaque, no remediation

Resume



LYANNE GIBSON

Data Scientist

☎ 115-166-7856 @ lyanne.gibson@example.com www.example.com Paris, France

MY TIME

- A Reading
- B Data analysis
- C Planning
- D Family and friends
- E Learning new technologies
- F Research
- G Movies
- H Sport

EDUCATION

BSc in Mathematics and Statistics

University of Altoona — 2006 - 2010 — GPA 3.85 / 4.0

Altoona, PA

EXPERIENCE

Data Scientist

DNB Bank ASA — 2015 - Ongoing — Paris, France

- Lead a group of 5 people
- Database manipulation of the Financial Aid Database across 16 different countries

Data Scientist

Razelie — 2013 - 2015 — Paris, France

- Consulted private and government clients
- Technical lead of D20 project and Glandore systems project
- Coordinated a team of 20 data scientists working on 6 different projects

Programmer

Qlouder — 2011 - 2013 — Altoona, PA

- Trained 27 students on SQL and Entity relationship diagram

Data Engineer

DocDecode — 2010 - 2011 — Altoona, PA

PASSIONS

Running

Traveling

Sport

SKILLS

Python R SQL Tableau Spark Java Machine Learning Pandas HTML Hadoop

MY LIFE PHILOSOPHY

> *Do not mind anything that anyone tells you about anyone else. Judge everyone and everything for yourself.*
>
> Henry James

LANGUAGES

English	Native
French	Proficient
Chinese	Beginner

FIND ME ONLINE

in /lyannegibson

@lyannegibson

Data Science is such a broad field that includes several subdivisions like data preparation and exploration, data representation and transformation, data visualization and presentation, predictive analytics, and machine learning, etc. For beginners, it’s only natural to raise the following question: What skills do I need to become a data scientist?

This article will discuss 10 essential skills that are necessary for practicing data scientists. These skills could be grouped into 2 categories, namely, technological skills (Math & Statistics, Coding Skills, Data Wrangling & Preprocessing Skills, Data Visualization Skills, Machine Learning Skills, and Real-World Project Skills) and soft skills (Communication Skills, Lifelong Learning Skills, Team Player Skills, and Ethical Skills).

Data science is an ever-evolving field, however mastering the foundations of data science will provide you with the necessary background that you need to pursue advanced concepts such as deep learning, artificial intelligence, etc. This article will discuss 10 essential skills for practicing data scientists.

1. Mathematics and Statistics Skills

(i) Statistics and Probability

Statistics and Probability is used for visualization of features, data preprocessing, feature transformation, data imputation, dimensionality reduction, feature engineering, model evaluation, etc. Here are the topics you need to be familiar with:

a) Mean

b) Median

c) Mode

d) Standard deviation/variance

e) Correlation coefficient and the covariance matrix

f) Probability distributions (Binomial, Poisson, Normal)

g) p-value

h) MSE (mean square error)

i) R2 Score

j) Baye’s Theorem (Precision, Recall, Positive Predictive Value, Negative Predictive Value, Confusion Matrix, ROC Curve)

k) A/B Testing

l) Monte Carlo Simulation

(ii) Multivariable Calculus

Most machine learning models are built with a data set having several features or predictors. Hence, familiarity with multivariable calculus is extremely important for building a machine learning model. Here are the topics you need to be familiar with:

a) Functions of several variables

b) Derivatives and gradients

c) Step function, Sigmoid function, Logit function, ReLU (Rectified Linear Unit) function

d) Cost function

e) Plotting of functions

f) Minimum and Maximum values of a function

(iii) Linear Algebra

Linear algebra is the most important math skill in machine learning. A data set is represented as a matrix. Linear algebra is used in data preprocessing, data transformation, and model evaluation. Here are the topics you need to be familiar with:

a) Vectors

b) Matrices

c) Transpose of a matrix

d) The inverse of a matrix

e) The determinant of a matrix

f) Dot product

g) Eigenvalues

h) Eigenvectors

(iv) Optimization Methods

Most machine learning algorithms perform predictive modeling by minimizing an objective function, thereby learning the weights that must be applied to the testing data to obtain the predicted labels. Here are the topics you need to be familiar with:

a) Cost function/Objective function

b) Likelihood function

c) Error function

d) Gradient Descent Algorithm and its variants (e.g., Stochastic Gradient Descent Algorithm)

Find out more about the gradient descent algorithm here: Machine Learning: How the Gradient Descent Algorithm Works.

1. Essential Programming Skills

Programming skills are essential in data science. Since Python and R are considered the two most popular programming languages in data science, essential knowledge in both languages is crucial. Some organizations may only require skills in either R or Python, not both.

(i) Skills in Python

Be familiar with basic programming skills in python. Here are the most important packages that you should master how to use:

a) Numpy

b) Pandas

c) Matplotlib

d) Seaborn

e) Scikit-learn

f) PyTorch

(ii) Skills in R

a) Tidyverse

b) Dplyr

c) Ggplot2

d) Caret

e) Stringr

(iii) Skills in Other Programming Languages

Skills in the following programming languages may be required by some organizations or industries:

a) Excel

b) Tableau

c) Hadoop

d) SQL

e) Spark

1. Data Wrangling and Preprocessing Skills

Data is key for any analysis in data science, be it inferential analysis, predictive analysis, or prescriptive analysis. The predictive power of a model depends on the quality of the data that was used in building the model. Data comes in different forms, such as text, table, image, voice, or video. Most often, data that is used for analysis has to be mined, processed and transformed to render it to a form suitable for further analysis.

i) Data Wrangling: The process of data wrangling is a critical step for any data scientist. Very rarely is data easily accessible in a data science project for analysis. It's more likely for the data to be in a file, a database, or extracted from documents such as web pages, tweets, or PDFs. Knowing how to wrangle and clean data will enable you to derive critical insights from your data that would otherwise be hidden.

ii) Data Preprocessing: Knowledge about data preprocessing is very important and include topics such as:

a) Dealing with missing data

b) Data imputation

c) Handling categorical data

d) Encoding class labels for classification problems

e) Techniques of feature transformation and dimensionality reduction, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

1. Data Visualization Skills

Understand the essential components of good data visualization.

a) Data Component: An important first step in deciding how to visualize data is to know what type of data it is, e.g., categorical data, discrete data, continuous data, time-series data, etc.

b) Geometric Component: Here is where you decide what kind of visualization is suitable for your data, e.g., scatter plot, line graphs, bar plots, histograms, Q-Q plots, smooth densities, boxplots, pair plots, heatmaps, etc.

c) Mapping Component: Here you need to decide what variable to use as your x-variable and what to use as your y-variable. This is important, especially when your dataset is multi-dimensional with several features.

d) Scale Component: Here you decide what kind of scales to use, e.g., linear scale, log scale, etc.

e) Labels Component: This includes things like axes labels, titles, legends, font size to use, etc.

f) Ethical Component: Here, you want to make sure your visualization tells the true story. You need to be aware of your actions when cleaning, summarizing, manipulating, and producing a data visualization and ensure you aren't using your visualization to mislead or manipulate your audience.

1. Basic Machine Learning Skills

Machine Learning is a very important branch of data science. It is important to understand the machine learning framework: Problem Framing, Data Analysis, Model Building, Testing & Evaluation, and Model Application. Find out more about the machine learning framework from here: The Machine Learning Process.

The following are important machine learning algorithms to be familiar with.

i) Supervised Learning (Continuous Variable Prediction)

a) Basic regression

b) Multi regression analysis

c) Regularized regression

ii) Supervised Learning (Discrete Variable Prediction)

a) Logistic Regression Classifier

b) Support Vector Machine Classifier

c) K-nearest neighbor (KNN) Classifier

d) Decision Tree Classifier

e) Random Forest Classifier

iii) Unsupervised Learning

a) KMeans clustering algorithm

1. Skills from Real World Capstone Data Science Projects

Skills acquired from course work alone will not make you a data scientist. A qualified data scientist must be able to demonstrate evidence of successful completion of a real-world data science project that includes every stage in data science and machine learning process such as problem framing, data acquisition and analysis, model building, model testing, model evaluation, and deploying models. Real-world data science projects could be found in the following:

a) Kaggle Projects

b) Internships

c) From Interviews

1. Communication Skills

Data scientists need to be able to communicate their ideas with other members of the team or with business administrators in their organizations. Good communication skills would play a key role here to be able to convey and present very technical information to people with little or no understanding of technical concepts in data science. Good communication skills will help foster an atmosphere of unity and togetherness with other team members such as data analysts, data engineers, field engineers, etc.

1. Be a Lifelong Learner

Data science is an ever-evolving field, so be prepared to embrace and learn new technologies. One way to keep in touch with developments in the field is to network with other data scientists. Some platforms that promote networking are LinkedIn, GitHub, and Medium (Towards Data Science and AI publications). The platforms are very useful for up-to-date information about recent developments in the field.

1. Team Player Skills

As a data scientist, you will be working in a team of data analysts, engineers, administrators, so you need good communication skills. You need to be a good listener, too, especially during early project development phases where you need to rely on engineers or other personnel to be able to design and frame a good data science project. Being a good team player will help you to thrive in a business environment and maintain good relationships with other members of your team as well as administrators or directors of your organization.

1. Ethical Skills in Data Science

Understand the implication of your project. Be truthful to yourself. Avoid manipulating data or using a method that will intentionally produce bias in results. Be ethical in all phases, from data collection and analysis to model building, analysis, testing, and application. Avoid fabricating results to mislead or manipulate your audience. Be ethical in the way you interpret the findings from your data science project.

Learn Free

HOW TO LEARN DATA SCIENCE FOR FREE

SELF LEARN
DATA SCIENCE

FREE CONTENT

fast.ai

Lectures on practical machine learning and deep learning

YOUTUBE

Recommended channels:
sentdex,
MIT OpenCourseWare

edX

edX offers up to 90% discount when you apply for financial assistance.

UDACITY

Udacity provides scholarship for their Nanodegree programs at no charge at all*.
*Depends on availability

coursera

Financial aids and scholarships are available.
No limit to the application but you have to apply for each course if you are taking a Specialization.

kaggle

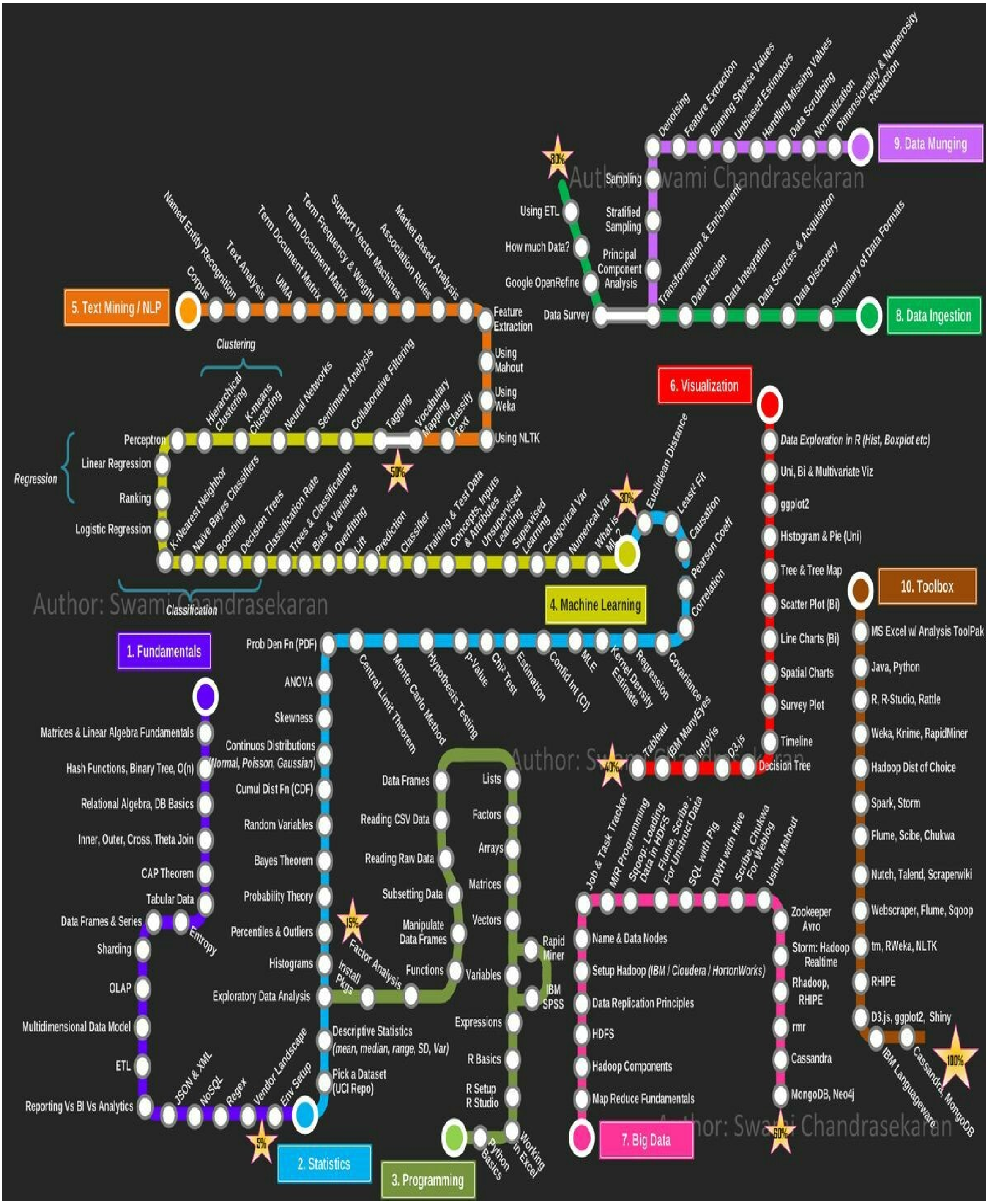
COURSES
Free courses to get you started

COMPETITIONS
Competitions to practice what you learn

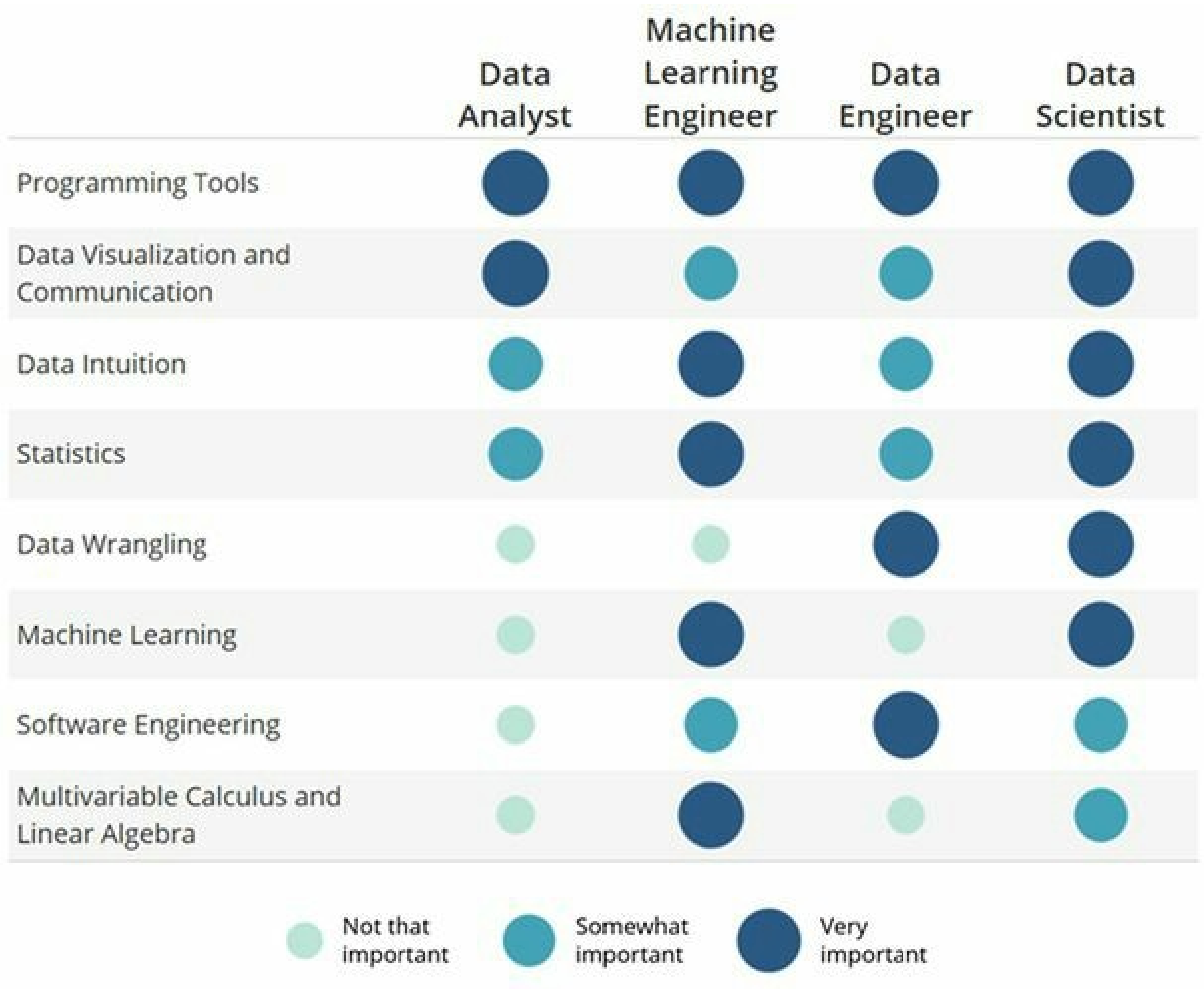
NOTEBOOKS
Notebooks to learn from others

SELF LEARN
DATA SCIENCE

Follow the Path

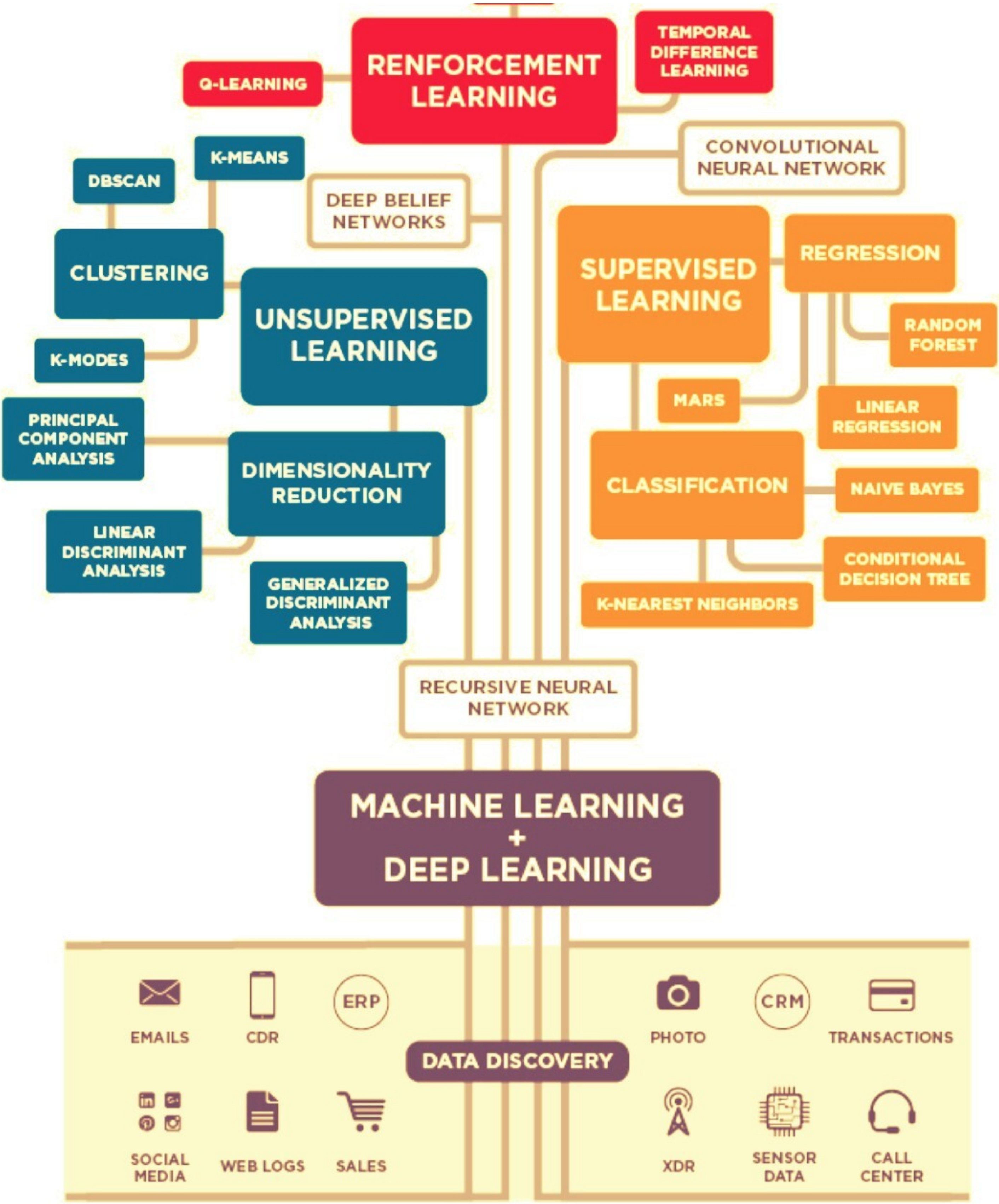


Differences

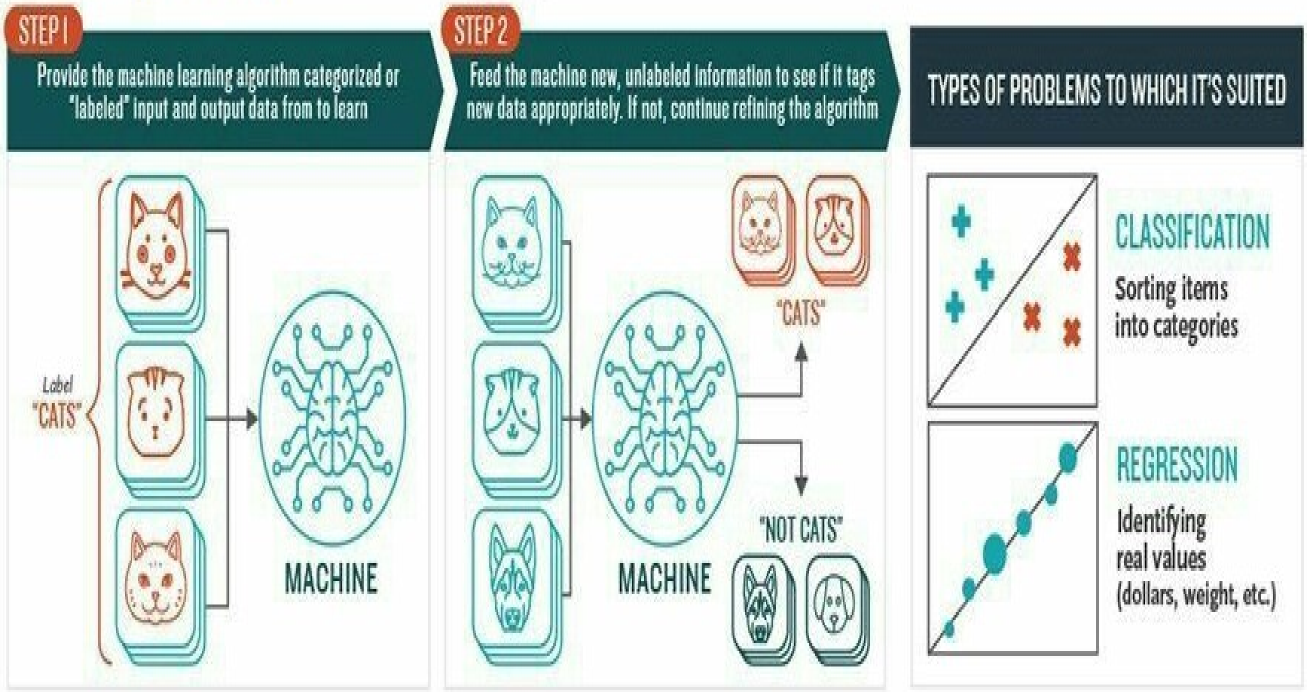


	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◉	◉	●
Data Intuition	◉	●	◉	●
Statistics	◉	●	◉	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◉	●	◉
Multivariable Calculus and Linear Algebra	○	●	○	◉

○ Not that important ◉ Somewhat important ● Very important


RENFORCEMENT LEARNING
Q-LEARNING
TEMPORAL DIFFERENCE LEARNING
CONVOLUTIONAL NEURAL NETWORK
K-MEANS
DBSCAN
DEEP BELIEF NETWORKS
CLUSTERING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REGRESSION
RANDOM FOREST
K-MODES
MARS
LINEAR REGRESSION
PRINCIPAL COMPONENT ANALYSIS
DIMENSIONALITY REDUCTION
CLASSIFICATION
NAIVE BAYES
LINEAR DISCRIMINANT ANALYSIS
CONDITIONAL DECISION TREE
GENERALIZED DISCRIMINANT ANALYSIS
K-NEAREST NEIGHBORS
RECURSIVE NEURAL NETWORK
MACHINE LEARNING + DEEP LEARNING
EMAILS
CDR
ERP
PHOTO
CRM
TRANSACTIONS
DATA DISCOVERY
SOCIAL MEDIA
WEB LOGS
SALES
XDR
SENSOR DATA
CALL CENTER

How Supervised Machine Learning Works



STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

TYPES OF PROBLEMS TO WHICH IT'S SUITED

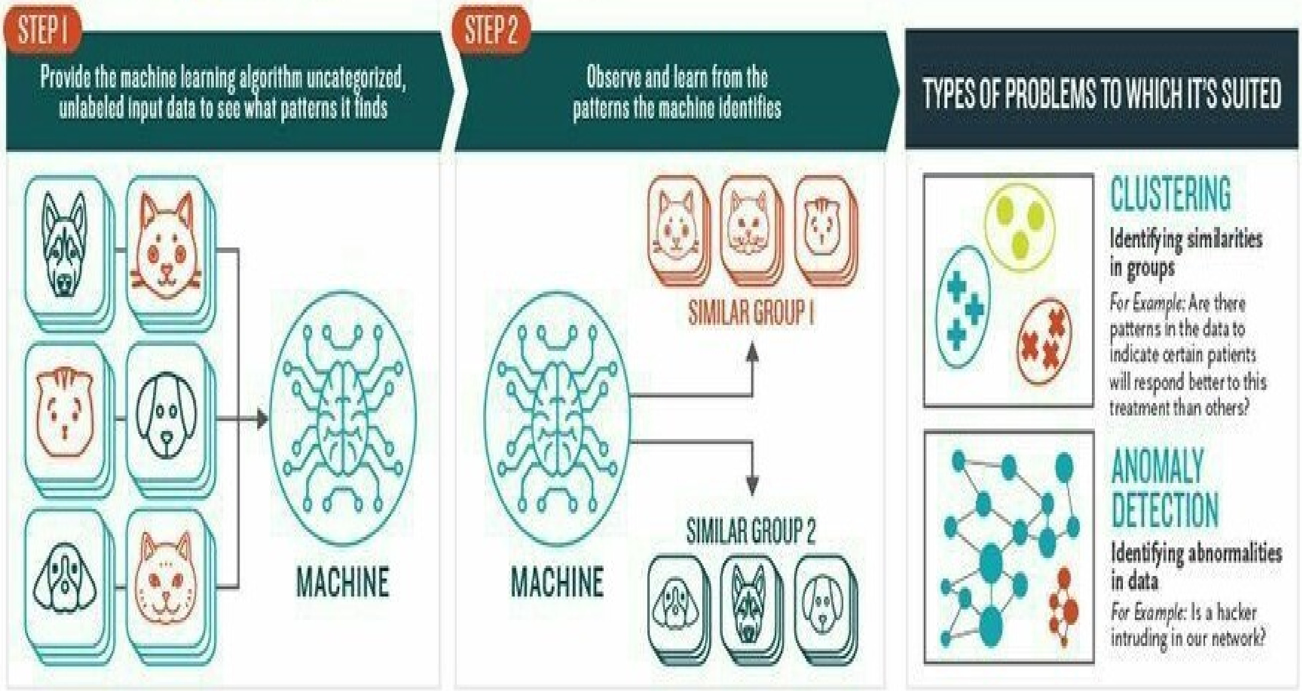
CLASSIFICATION

Sorting items into categories

REGRESSION

Identifying real values (dollars, weight, etc.)

How Unsupervised Machine Learning Works



STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2

Observe and learn from the patterns the machine identifies

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLUSTERING

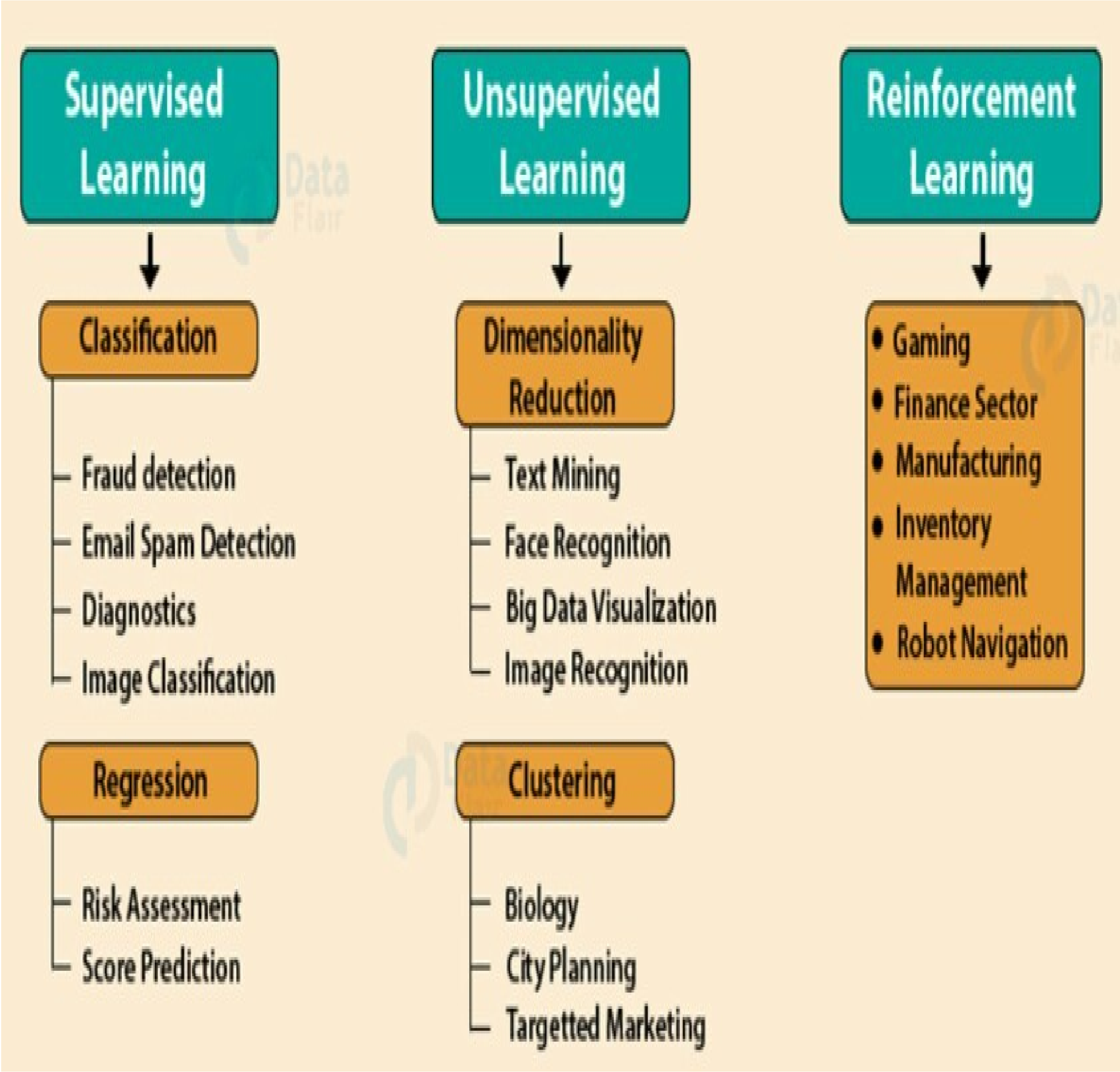
Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?

ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?


Supervised Learning
Classification
Fraud detection
Email Spam Detection
Diagnostics
Image Classification
Regression
Risk Assessment
Score Prediction
Unsupervised Learning
Dimensionality Reduction
Text Mining
Face Recognition
Big Data Visualization
Image Recognition
Clustering
Biology
City Planning
Targetted Marketing
Reinforcement Learning
Gaming
Finance Sector
Manufacturing
Inventory Management
Robot Navigation

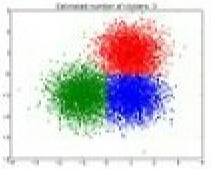
Machine learning algorithms are expected to replace 25% of the jobs across the world in the next 10 years.

Classification of Machine Learning Algorithms -

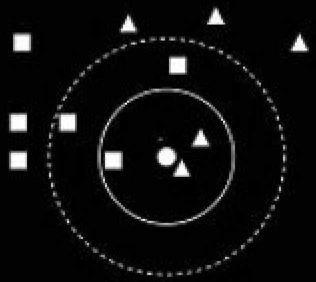
Supervised

Reinforcement

Unsupervised



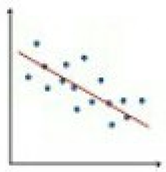
Naive Bayes Classifier Algorithm



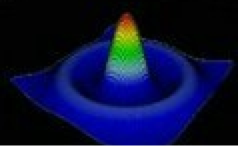
K Means Clustering Algorithm

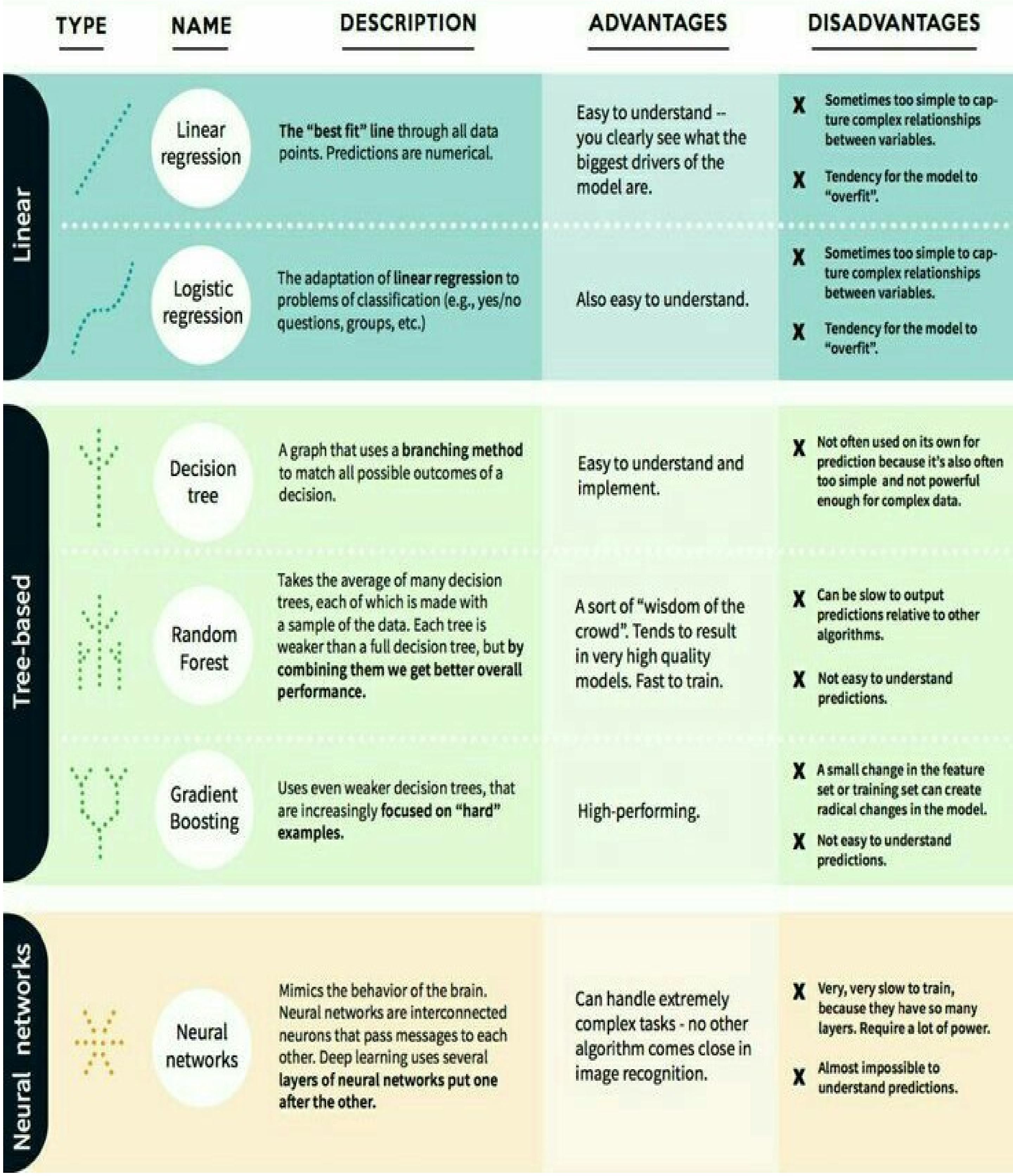
Support Vector Machine Algorithm

Apriori Algorithm

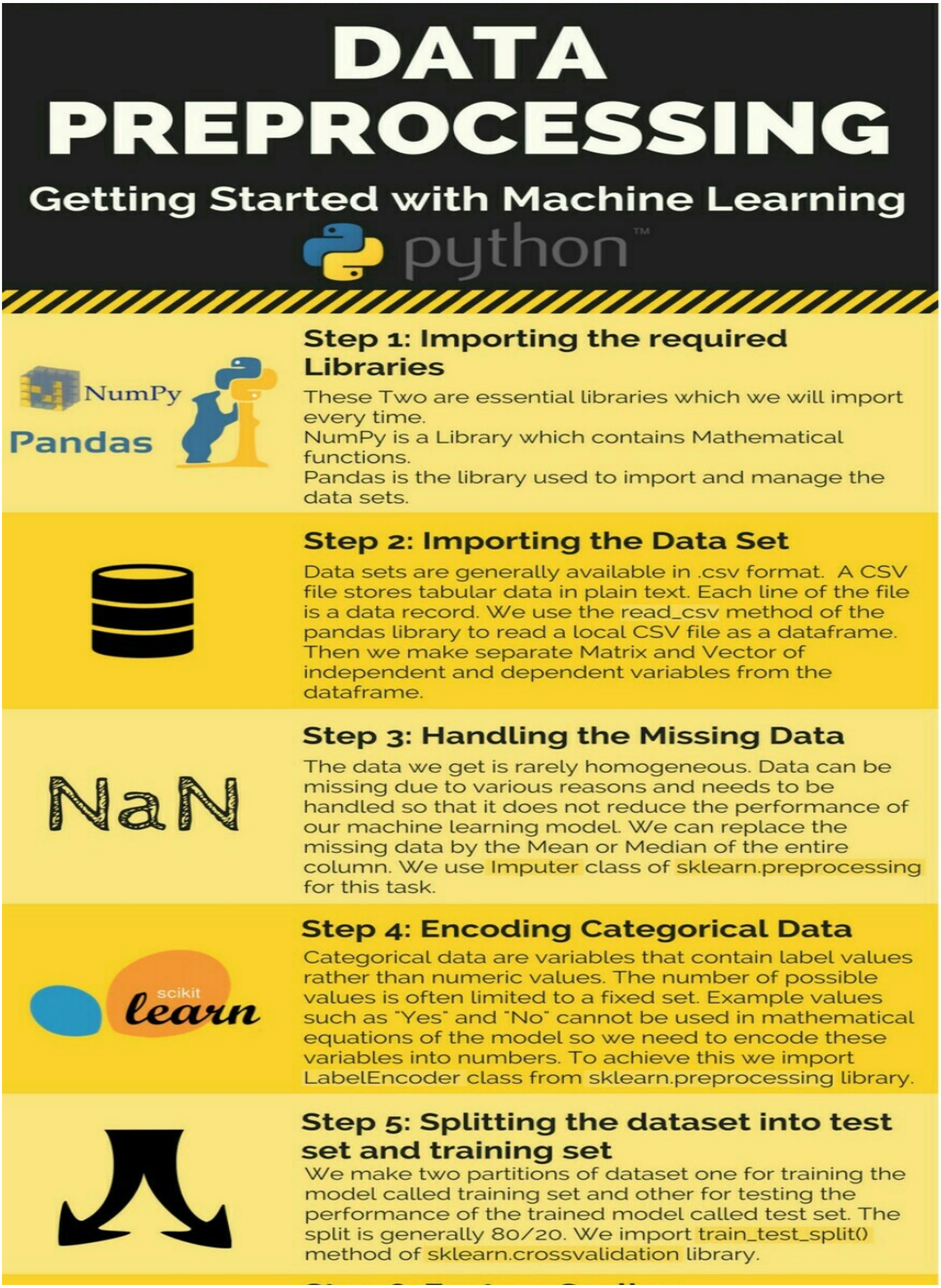


Linear Regression





TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	The **"best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Neural networks	Neural networks	Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several **layers of neural networks put one after the other.**	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	X Very, very slow to train, because they have so many layers. Require a lot of power. X Almost impossible to understand predictions.



DATA PREPROCESSING

Getting Started with Machine Learning

Step 1: Importing the required Libraries

These Two are essential libraries which we will import every time.

NumPy is a Library which contains Mathematical functions.

Pandas is the library used to import and manage the data sets.

Step 2: Importing the Data Set

Data sets are generally available in .csv format. A CSV file stores tabular data in plain text. Each line of the file is a data record. We use the read_csv method of the pandas library to read a local CSV file as a dataframe. Then we make separate Matrix and Vector of independent and dependent variables from the dataframe.

Step 3: Handling the Missing Data

The data we get is rarely homogeneous. Data can be missing due to various reasons and needs to be handled so that it does not reduce the performance of our machine learning model. We can replace the missing data by the Mean or Median of the entire column. We use Imputer class of sklearn.preprocessing for this task.

Step 4: Encoding Categorical Data

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. Example values such as "Yes" and "No" cannot be used in mathematical equations of the model so we need to encode these variables into numbers. To achieve this we import LabelEncoder class from sklearn.preprocessing library.

Step 5: Splitting the dataset into test set and training set

We make two partitions of dataset one for training the model called training set and other for testing the performance of the trained model called test set. The split is generally 80/20. We import train_test_split() method of sklearn.crossvalidation library.

SIMPLE LINEAR REGRESSION

Predicting a response using a single feature.

It is a method to predict dependent variable (Y) based on values of independent variables (X). It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

How to find the best fit line?

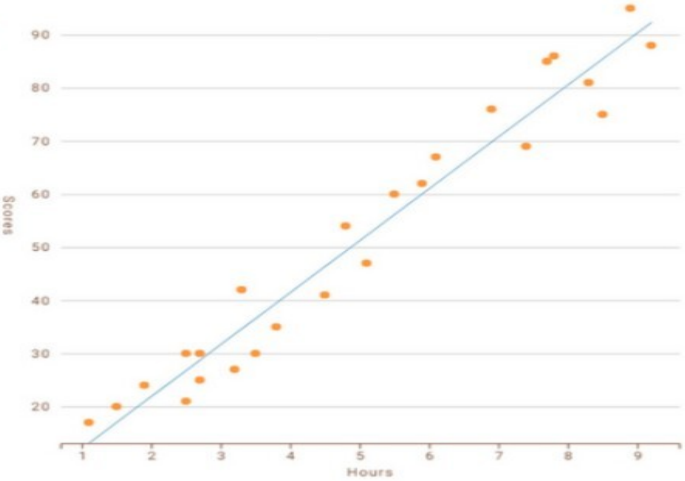
In this regression model, we are trying to minimize the errors in prediction by finding the "line of best fit" — the regression line from the errors would be minimal. We are trying to minimize the length between the observed value (Yi) and the predicted value from our model (Yp).

Predicted Value

$$y_i \quad y_p$$

Observed Value

$$\min\{\text{SUM}(y_i - y_p)^2\}$$



Dependent Variable

$$y = b_\circ + b_1 x_1$$

Independent Variable

In this regression task, we will predict the percentage of marks that a student is expected to score based upon the number of hours they studied.

Slope

$$Score = b_\circ + b_1 * hours$$

Y - intercept

STEP 1: PREPROCESS THE DATA

We will follow the same steps as in my previous infographic of Data Preprocessing.

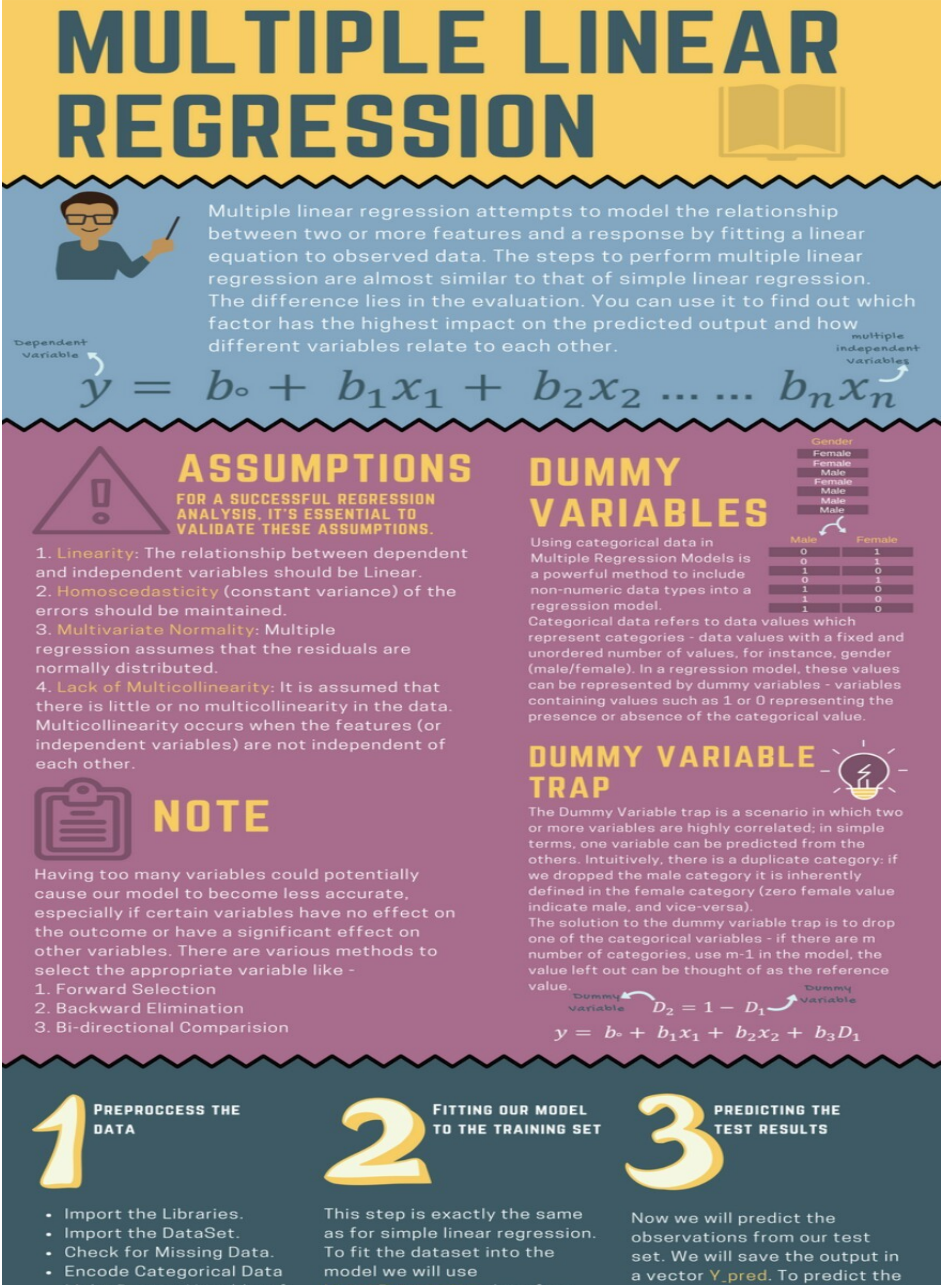
- Import the Libraries.
- Import the DataSet.
- Check for Missing Data.
- Split the DataSet.
- Feature Scaling will be taken care by the Library we will use for Simple Linear Regression Model.

STEP 2: FITTING SIMPLE LINEAR REGRESSION MODEL TO THE TRAINING SET

To fit the dataset into the model we will use LinearRegression class from sklearn.linear_model library. Then we make an object regressor of LinearRegression Class. Now we will fit the regressor object into our dataset using fit() method of LinearRegression Class.

STEP 3: PREDICTING THE RESULT

Now we will predict the observations from our test set. We will save the output in a vector



MULTIPLE LINEAR REGRESSION

Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data. The steps to perform multiple linear regression are almost similar to that of simple linear regression. The difference lies in the evaluation. You can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

Dependent variable

multiple independent variables

$$y = b_\circ + b_1x_1 + b_2x_2 \dots \dots b_nx_n$$

ASSUMPTIONS

FOR A SUCCESSFUL REGRESSION ANALYSIS, IT'S ESSENTIAL TO VALIDATE THESE ASSUMPTIONS.

1. Linearity: The relationship between dependent and independent variables should be Linear.
2. Homoscedasticity (constant variance) of the errors should be maintained.
3. Multivariate Normality: Multiple regression assumes that the residuals are normally distributed.
4. Lack of Multicollinearity: It is assumed that there is little or no multicollinearity in the data. Multicollinearity occurs when the features (or independent variables) are not independent of each other.

NOTE

Having too many variables could potentially cause our model to become less accurate, especially if certain variables have no effect on the outcome or have a significant effect on other variables. There are various methods to select the appropriate variable like -

1. Forward Selection
2. Backward Elimination
3. Bi-directional Comparision

DUMMY VARIABLES

Gender
Female
Female
Male
Female
Male
Male
Male

Male	Female
0	1
0	1
1	0
0	1
1	0
1	0
1	0

Using categorical data in Multiple Regression Models is a powerful method to include non-numeric data types into a regression model.

Categorical data refers to data values which represent categories - data values with a fixed and unordered number of values, for instance, gender (male/female). In a regression model, these values can be represented by dummy variables - variables containing values such as 1 or 0 representing the presence or absence of the categorical value.

DUMMY VARIABLE TRAP

The Dummy Variable trap is a scenario in which two or more variables are highly correlated; in simple terms, one variable can be predicted from the others. Intuitively, there is a duplicate category: if we dropped the male category it is inherently defined in the female category (zero female value indicate male, and vice-versa).

The solution to the dummy variable trap is to drop one of the categorical variables - if there are m number of categories, use m-1 in the model, the value left out can be thought of as the reference value.

Dummy variable

Dummy variable

$$D_2 = 1 - D_1$$

$$y = b_\circ + b_1x_1 + b_2x_2 + b_3D_1$$

1 PREPROCCESS THE DATA

- Import the Libraries.
- Import the DataSet.
- Check for Missing Data.
- Encode Categorical Data

2 FITTING OUR MODEL TO THE TRAINING SET

This step is exactly the same as for simple linear regression. To fit the dataset into the model we will use

3 PREDICTING THE TEST RESULTS

Now we will predict the observations from our test set. We will save the output in a vector Y_pred. To predict the

LOGISTIC REGRESSION

WHAT IS LOGISTIC REGRESSION

Logistic regression is used for a different class of problems known as classification problems. Here the aim is to predict the group to which the current object under observation belongs to. It gives you a discrete binary outcome between 0 and 1. A simple example would be whether a person will vote or not in upcoming elections.

How Does It Work?

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it's underlying logistic function.

Making Predictions

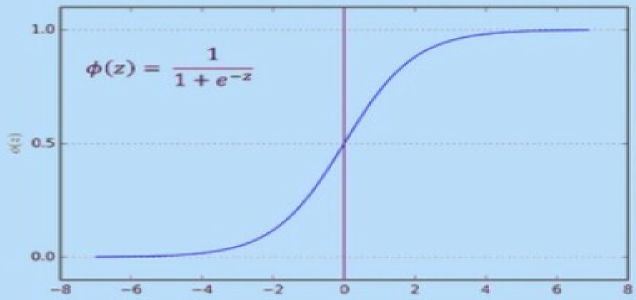
These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

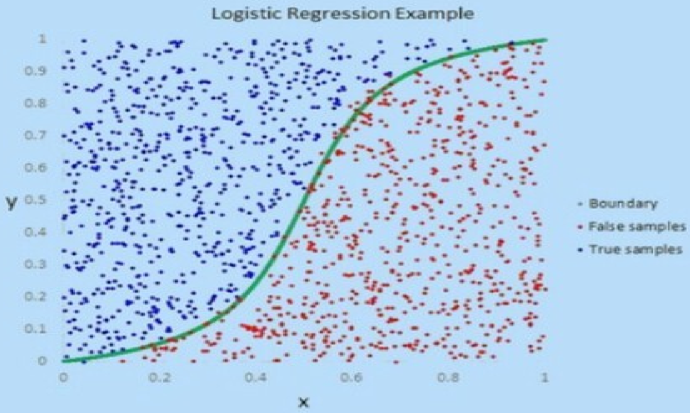
Logistic vs Linear

Logistic regression gives you a discrete outcome but linear regression gives a continuous outcome.

Sigmoid Function

The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits.





This infographic is just the Logistic regression intuition and is very brief. The mathematical logic and implementation part will be covered in another infographic.

K NEAREST NEIGHBOURS

An Intuition to K-NN Classification Algorithm

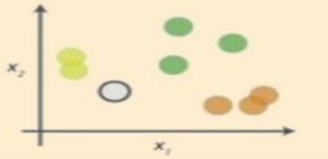
What is k-NN?

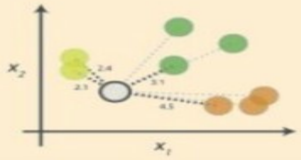
K-Nearest Neighbor algorithm is a simple yet most used classification algorithm. It can also be used for regression.

KNN is non-parametric (means that it does not make any assumptions on the underlying data distribution), instance-based (means that our algorithm doesnt explicitly learn a model. Instead, it chooses to memorize the training instances.) and used in a supervised learning setting.

k-NN is also called a lazy algorithm because it is instance based.

We want to classify the grey point into one of the three classes light green, green and red





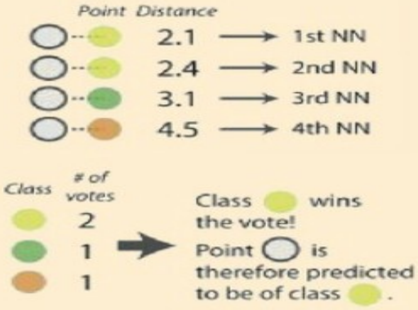
Start by calculating the distance between the grey point and k -nearest points

How Does k-NN Algorithm work?

k-NN when used used for classification — the output is a class membership (predicts a class — a discrete value).
There are three key elements of this approach: a set of labeled objects, e.g., a set of stored records, a distance between objects, and the value of k, the number of nearest neighbors.

Making Predections

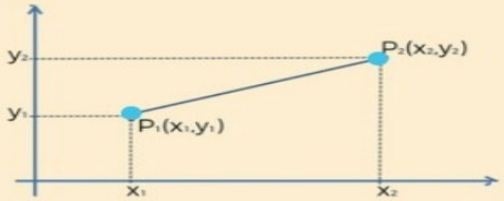
To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class label of the majority of nearest neighbors is then used to determine the class label of the object. For real-valued input variables, the most popular distance measure is Euclidean distance.



Point	Distance		
○--🟡	2.1	→	1st NN
○--🟡	2.4	→	2nd NN
○--🟢	3.1	→	3rd NN
○--🟠	4.5	→	4th NN

Class	# of votes
🟡	2
🟢	1
🟠	1

Class 🟡 wins the vote!
Point ○ is therefore predicted to be of class 🟡.



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The Distance

Euclidean distance is calculated as the square root of the sum of the squared differences between a new point and an existing point across all input attributes .
Other popular distance measures include:

- Hamming Distance
- Manhattan Distance
- Minkowski Distance

Value of k

SUPPORT VECTOR MACHINES

What is SVM?

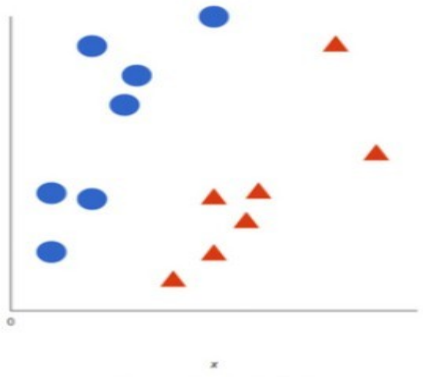
Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression. However, it is mostly used in classification problems.
In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate.

How is the data classified?

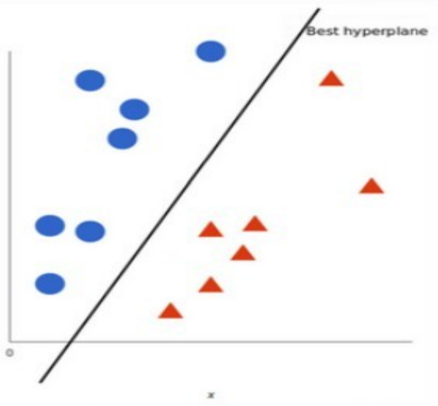
We perform classification by finding the hyperplane that differentiates the two classes very well. In other words the algorithm outputs an optimal hyperplane which categorizes new examples.

What is a optimal Hyper-Plane?

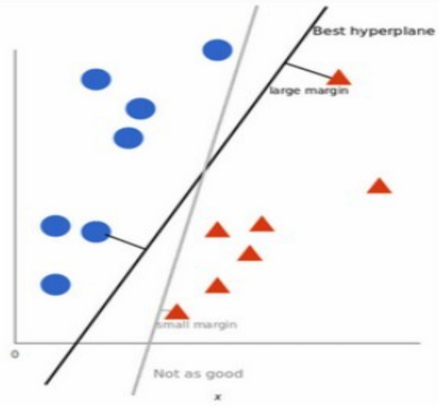
For SVM, it's the one that maximizes the margins from both tags. In other words: the hyperplane whose distance to the nearest element of each tag is the largest.



Our labeled data

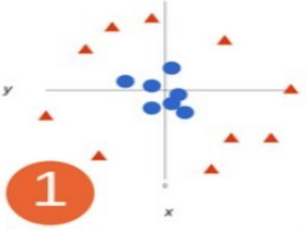


In 2D, the best hyper-plane is simply a line

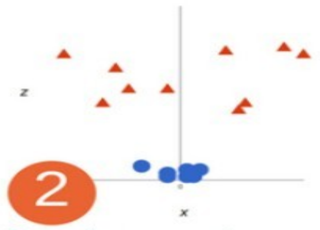


Not all hyper-planes are created equal

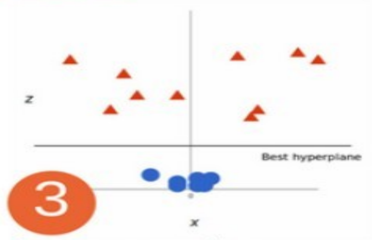
Nonlinear data



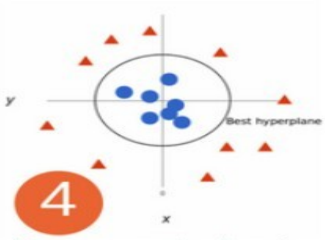
In above case we cannot draw a linear boundary.We will now add a third dimension. We create a new z dimension, and we rule that it be calculated a certain way that is convenient for us: $z = x^2 + y^2$ (equation for a circle).



This will give us a three-dimensional space. From a different perspective, the data is now in two linearly separated groups. All values for z would be always positive because z is the squared sum of both x and y.



Since we are in three dimensions now, the hyper-plane is a plane parallel to the x axis at a certain z. We choose the hyperplane, one that maximizes the margins from both of the Classes.



Now we map back to 2 dimensions.Our decision boundary is a circumference which separates both tags using SVM. We got a circle as a hyper-plane.

TUNING PARAMETERS

KERNEL

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role. Polynomial and exponential kernels calculates

REGULARIZATION

For large values of this parameter, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a

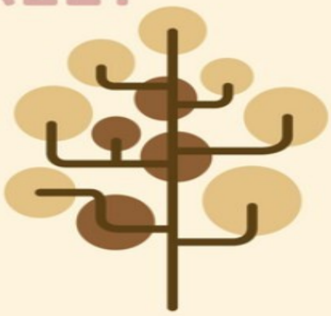
DECISION TREE

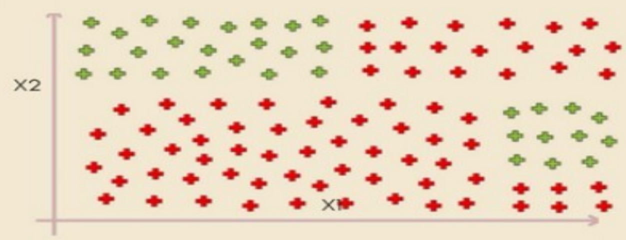
AN INTUITION ON DECISION TREE FOR CLASSIFICATION

1 WHAT IS A DECISION TREE?

It is a type of supervised learning algorithm that is mostly used in classification problems and works for both categorical and continuous input and output variables.

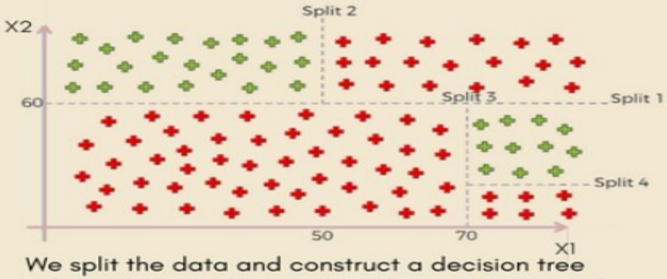
A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision.



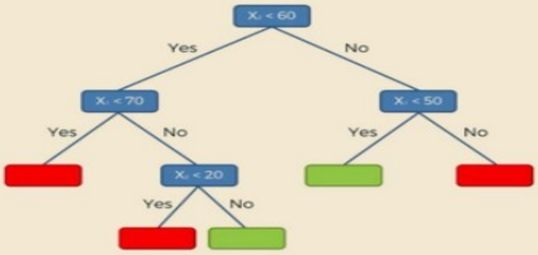


Here we've got an example with lots of points on our two dimensional scatter plot.
Now how does a decision tree work.

So what it is going to do is cut it up into slices in several iterations.



We split the data and construct a decision tree side by side which we will use later. This very task is achieved by using various algorithms. It builds a decision tree from a fixed set of examples and the resulting tree is used to classify future samples.



The resulting Tree (obtained by applying algorithms like CART, ID3) which will be later used to predict the outcomes

3 DECISION TREE ALGORITHM: ID3

ID3 stands for Iterative Dichotomizer 3. The basic idea is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node.
Sounds simple — but which node should we select to build the correct and most precise decision tree? How would we decide that? Well, we have some measures that can help us in selecting the best choice!

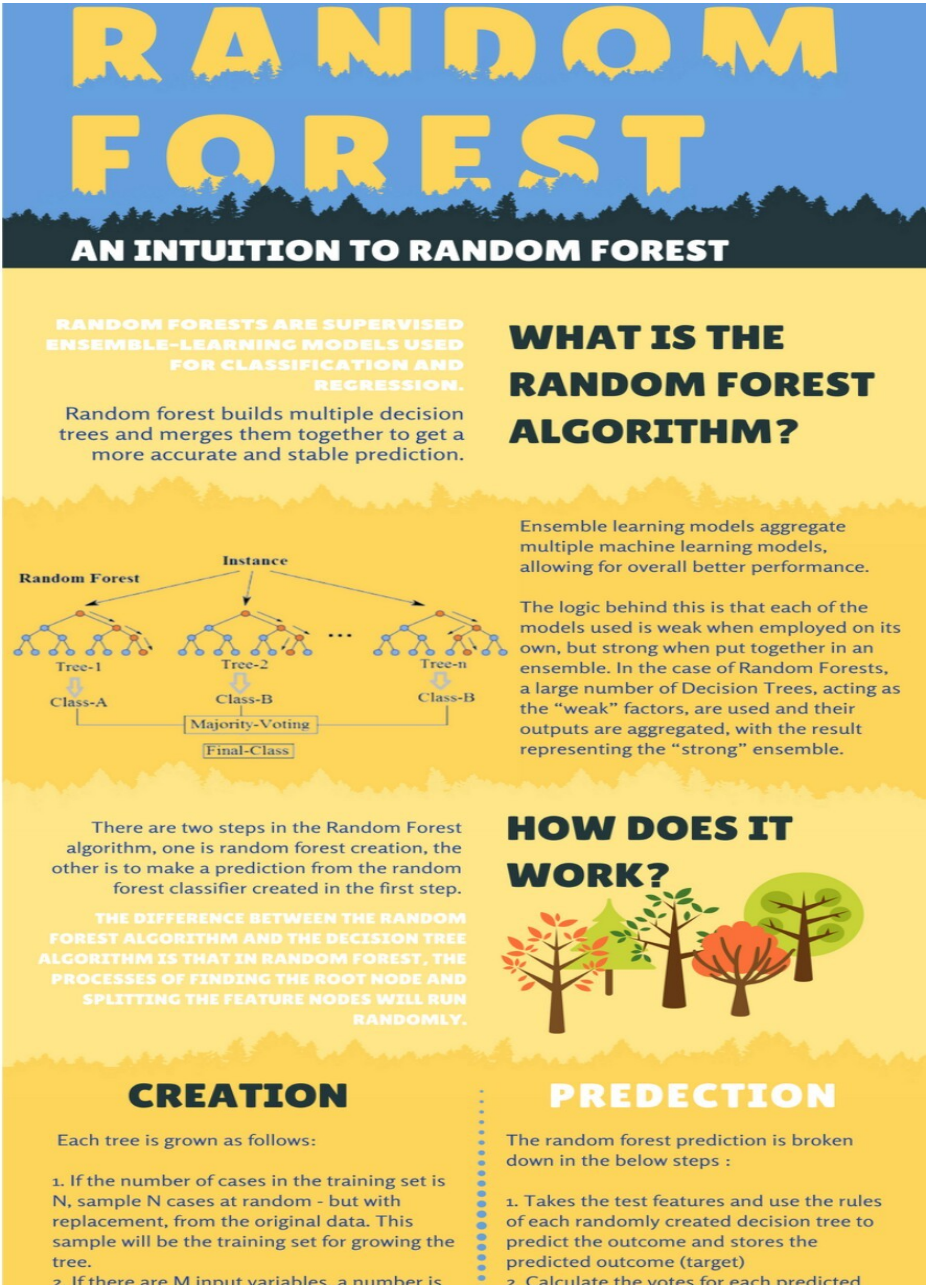
```
Loop:
    A –> Best Attribute
    Assign A as decision
attribute for node.
    For each value of A,
create a descendant of node.
    Sort training examples to
leaves.
    If
    examples perfectly
    classified:
    STOP
    Else:
    Iterate over leaves
```

INFORMATION GAIN

The best attribute is the one which gives us maximum Information Gain. Broadly speaking, it is a mathematical way to capture the amount of information we want by picking a particular attribute.

ENTROPY

Entropy in machine learning also carries almost the same meaning as it does in Thermodynamics. It is a measure of randomness.



RANDOM FOREST

AN INTUITION TO RANDOM FOREST

RANDOM FORESTS ARE SUPERVISED ENSEMBLE-LEARNING MODELS USED FOR CLASSIFICATION AND REGRESSION.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

WHAT IS THE RANDOM FOREST ALGORITHM?

Ensemble learning models aggregate multiple machine learning models, allowing for overall better performance.

The logic behind this is that each of the models used is weak when employed on its own, but strong when put together in an ensemble. In the case of Random Forests, a large number of Decision Trees, acting as the "weak" factors, are used and their outputs are aggregated, with the result representing the "strong" ensemble.

There are two steps in the Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first step.

THE DIFFERENCE BETWEEN THE RANDOM FOREST ALGORITHM AND THE DECISION TREE ALGORITHM IS THAT IN RANDOM FOREST, THE PROCESSES OF FINDING THE ROOT NODE AND SPLITTING THE FEATURE NODES WILL RUN RANDOMLY.

HOW DOES IT WORK?

CREATION

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number is

PREDICTION

The random forest prediction is broken down in the below steps :

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted

K-Means Clustering

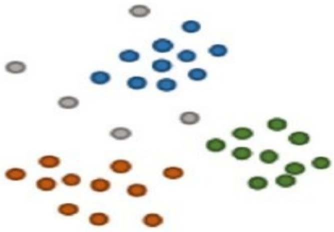
STARTING WITH UNSUPERVISED LEARNING

1.) WHAT IS UNSUPERVISED LEARNING

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. Unsupervised algorithms find patterns based only on input data. This technique is useful when we're not quite sure what to look for.

2.) CLUSTERING ALGORITHMS

Clustering Algorithms do the task of dividing the population or data points into a variety of groups such that data points within the same cluster are similar to other data points within the same cluster than those in other groups. Basically, the aim is to separate groups with similar traits and assign them into clusters.

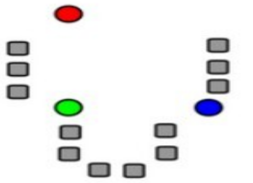


3.) K MEANS CLUSTERING

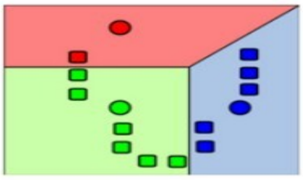
In this algorithm, we group the items into k clusters such that all items in the same cluster are as similar to each other as possible. And items not in the same cluster are as different as possible.

Distance measures(like Euclidean distance) are used to calculate similarity and dissimilarity between the data points. Each cluster has a centroid. Centroid can be thought as the point that is most representative of the cluster.

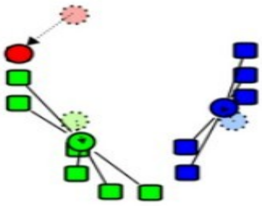
4.) HOW K-MEANS CLUSTERING WORKS



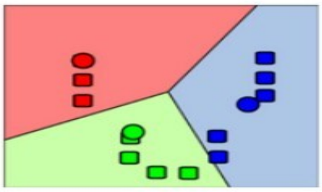
1. k initial "means" (in this case k=3) are randomly generated within the data domain.



2. k clusters are created by associating every observation with the nearest mean.



3. The centroid of each of the k clusters becomes the new mean.

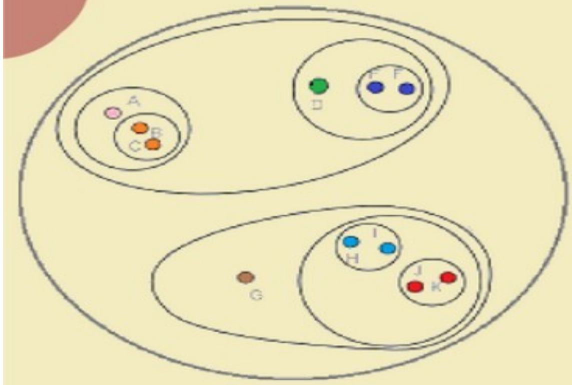


4. Steps 2 and 3 are repeated until convergence has been reached.

The objective of K-Means clustering is to minimize total intra-cluster

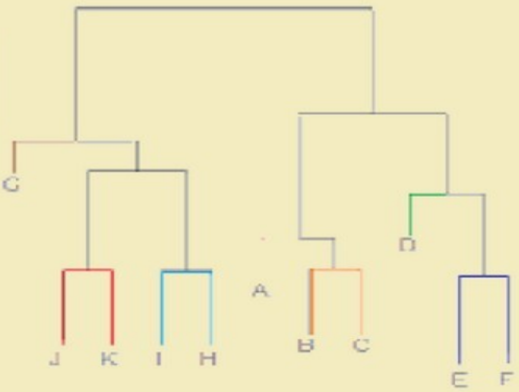
number of clusters
number of cases
centroid for cluster j
case i
k
n

HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING

Hierarchical clustering, as the name suggests is an algorithm that builds a hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. There are two types of hierarchical clustering, Divisive and Agglomerative.

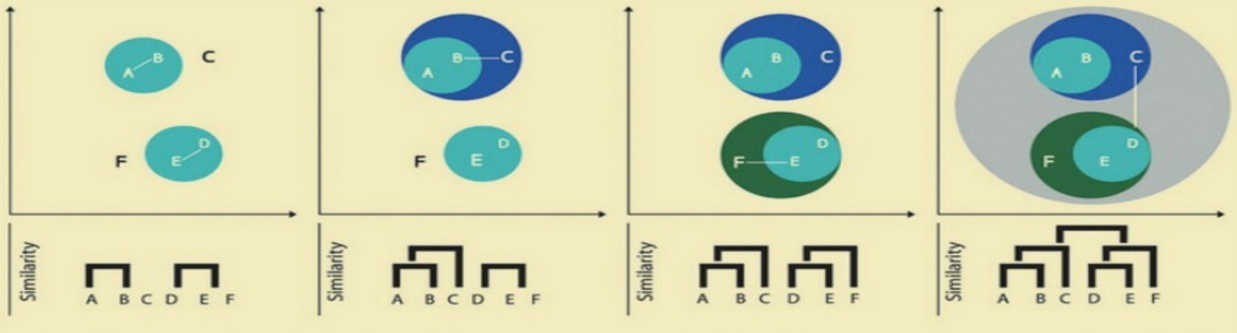


AGGLOMERATIVE HIERARCHICAL CLUSTERING

Here, each observation is initially considered as a cluster of its own (leaf). Then, the most similar clusters are successively merged until there is just one single big cluster (root). This hierarchy of clusters is represented as a tree (or dendrogram).

DENDROGRAM

The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.


A
B
C
D
E
F
Similarity
A B C D E F