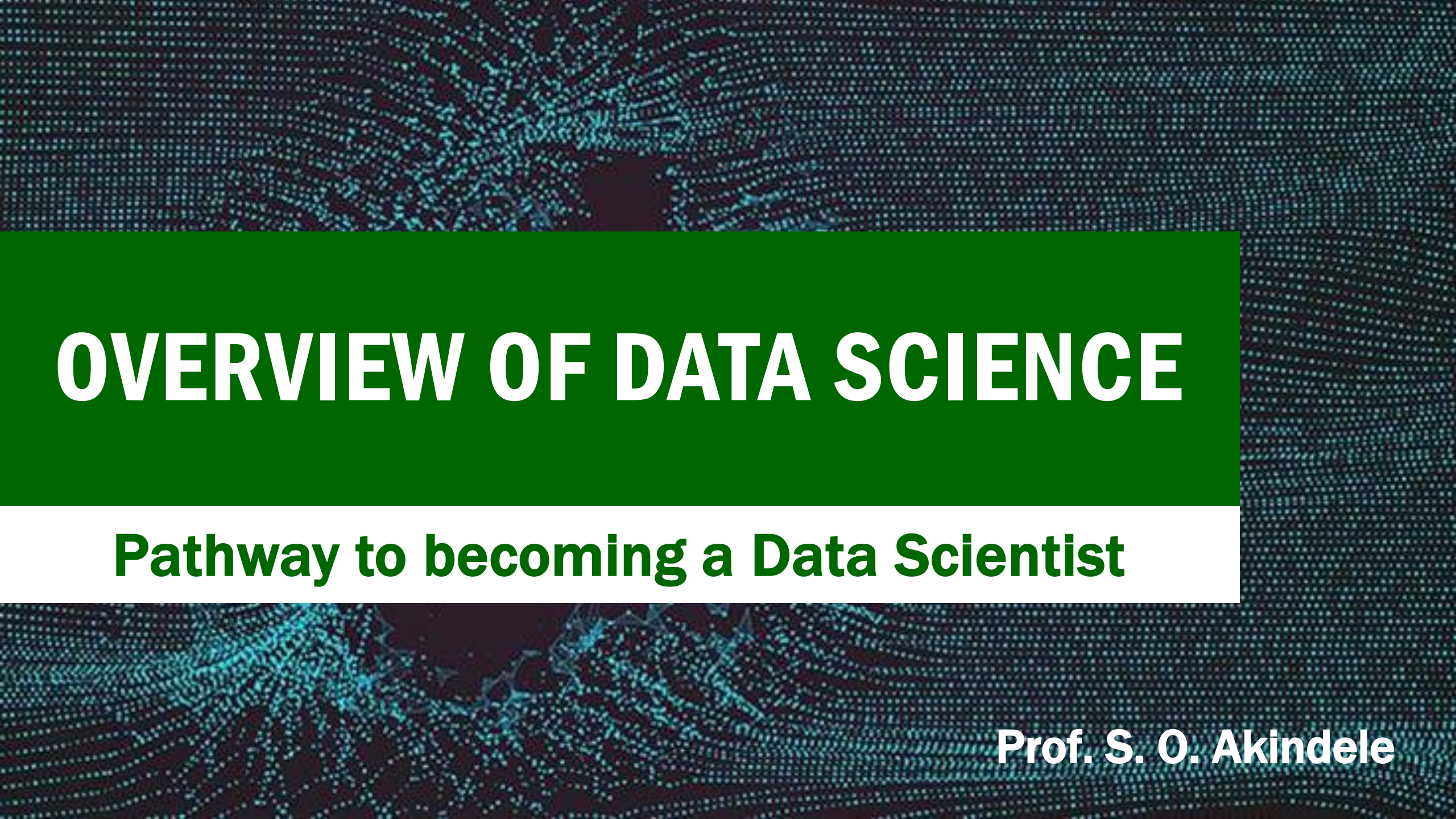


OVERVIEW OF DATA SCIENCE

Pathway to becoming a Data Scientist

Prof. S. O. Akindele

DEFINITION OF DATA SCIENCE

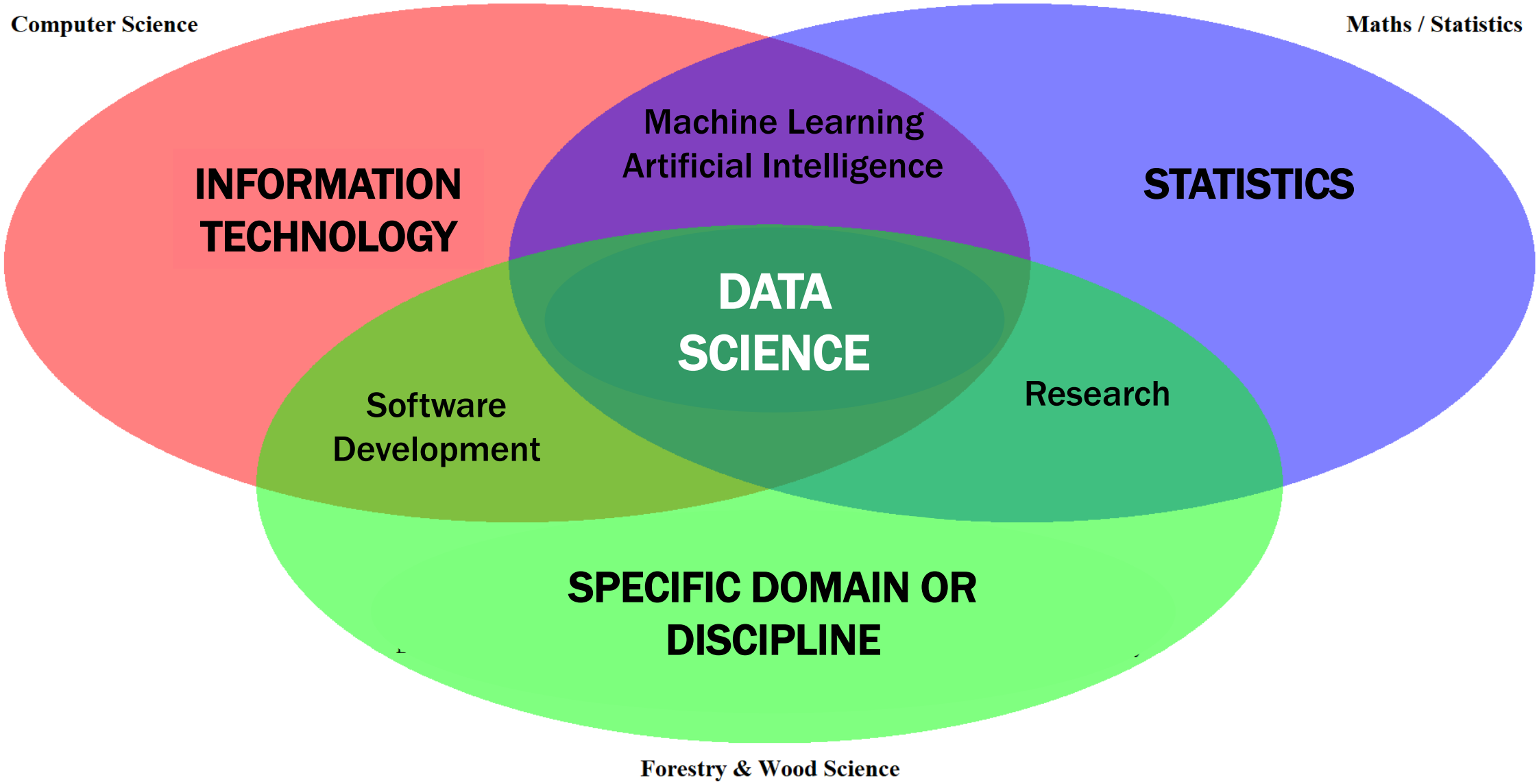
Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, and interact with data to create data products.

From the definition, we can see that Data Science is

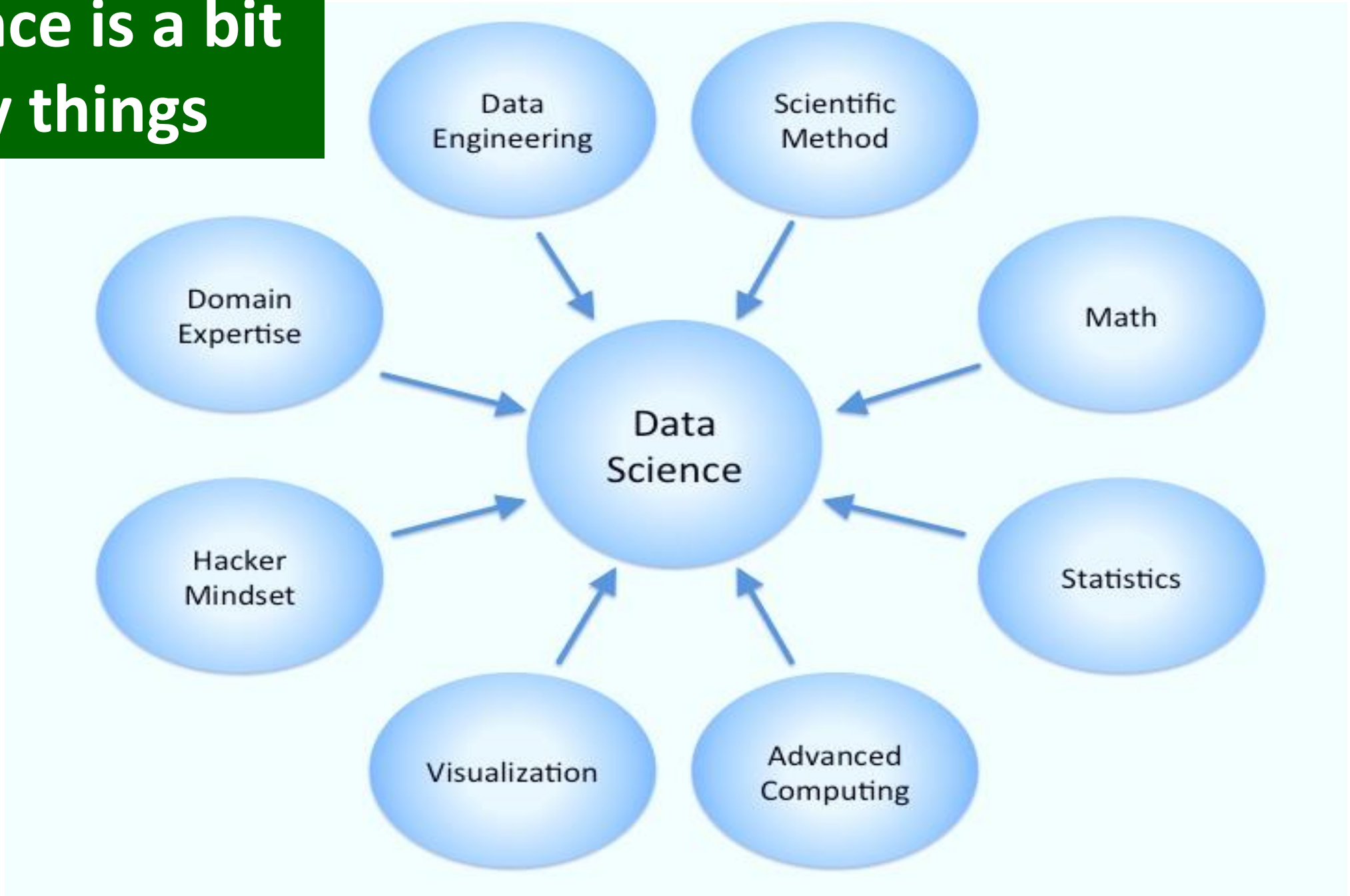
- Multidisciplinary
- Data-driven
- Computerized

Based on notes from CS194 at UC Berkeley by Michael Franklin, John Canny, and Jeff Hammerbacher

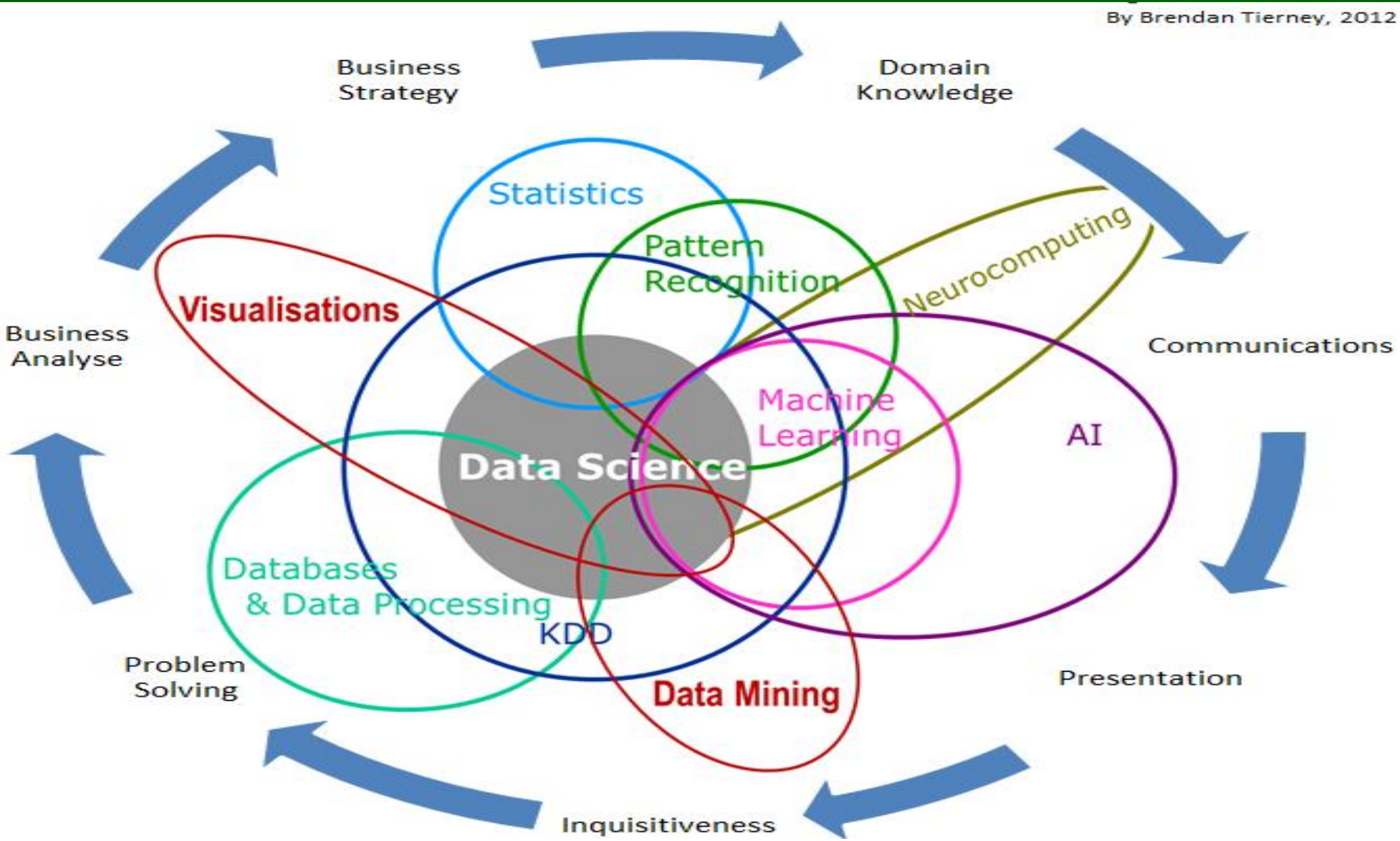
MULTIDISCIPLINARY NATURE OF DATA SCIENCE



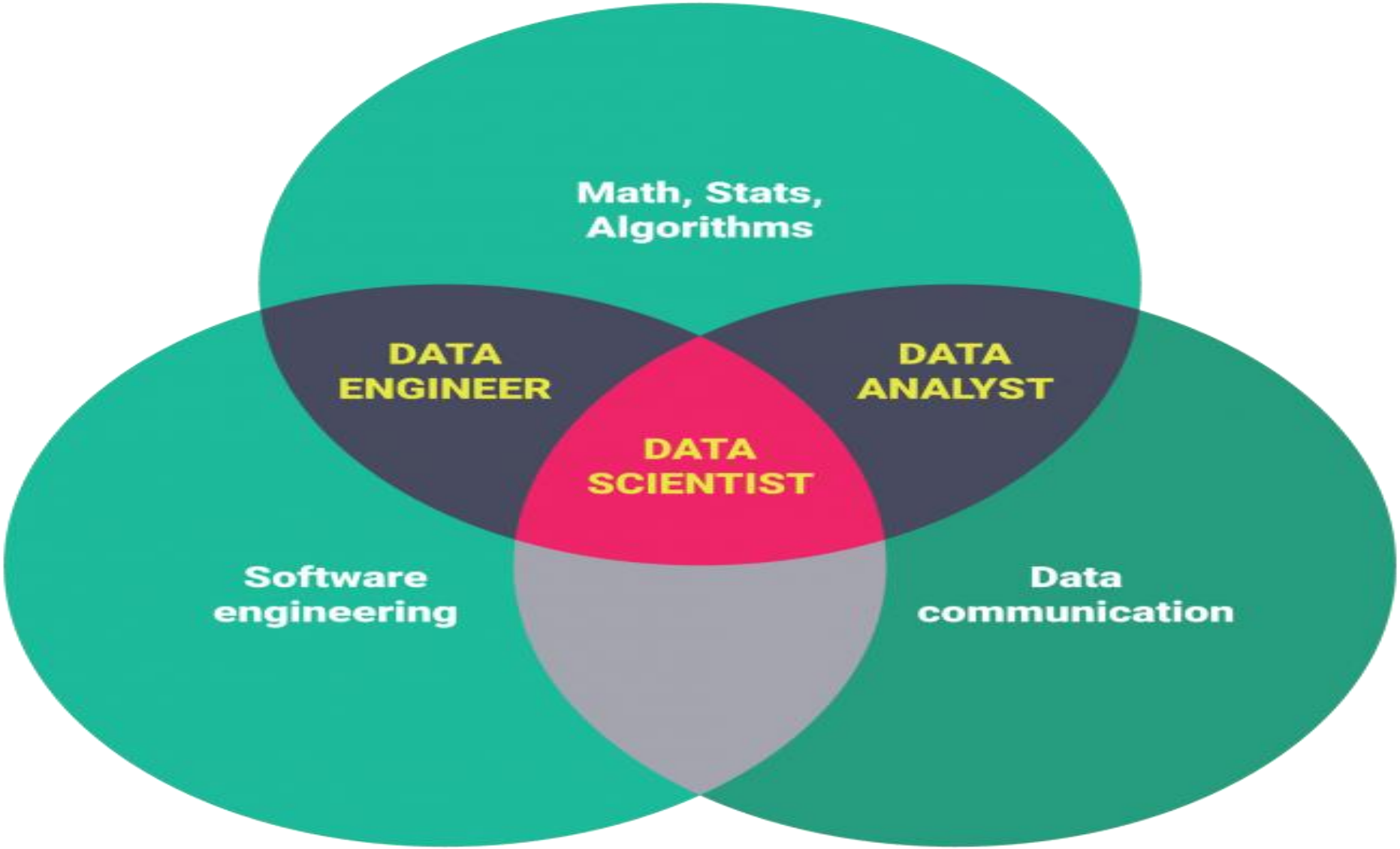
Data Science is a bit of many things



MULTIDISCIPLINARY NATURE OF DATA SCIENCE



MULTIDISCIPLINARY NATURE OF DATA SCIENCE



Designations of Data Specialists

Data Analyst	Data Engineer	Data Scientist
Data Analyst analyzes numeric data and uses it to help companies make better decisions.	Data Engineer is involved in preparing data. He develops, constructs, tests & maintains complete data architecture.	Data Scientist analyzes and interprets complex data. They are data wranglers who organize (big) data.

https://www.edureka.co/blog/data-analyst-vs-data-engineer-vs-data-scientist/

Roles and Responsibilities

Data Analyst	Data Engineer	Data Scientist
• Pre-processing and data gathering	• Construct, develop, test and maintain data architectures	• Responsible for developing Operational Models
• Emphasis on representing data via reporting and visualization	• Understand programming and its complexity	• Carry out data analytics and optimization using machine learning & deep learning
• Responsible for statistical analysis & data interpretation	• Deploy Machine Learning & statistical models	• Involved in strategic planning for data analytics
• Ensures data acquisition & maintenance	• Building pipelines for various data Extract, Transfer and Load (ETL) operations	• Integrate data & perform ad-hoc analysis
• Optimize Statistical Efficiency & Quality	• Ensures data accuracy and flexibility	• Fill in the gap between the stakeholders and customer

https://www.edureka.co/blog/data-analyst-vs-data-engineer-vs-data-scientist/

Skill-Sets for Data Specialists

Data Analyst	Data Engineer	Data Scientist
• Data Warehousing	• Data Warehousing & ETL	• Statistical & Analytical skills
• Adobe & Google Analytics	• Advanced programming knowledge	• Data Mining
• Programming knowledge	• Hadoop-based Analytics	• Machine Learning & Deep learning principles
• Scripting & Statistical skills	• In-depth knowledge of SQL/ database	• In-depth programming knowledge (SAS/R/ Python coding)
• Reporting & data visualization	• Data architecture & pipelining	• Hadoop-based analytics
• SQL/ database knowledge	• Machine learning concept knowledge	• Data optimization
• Spread-Sheet knowledge	• Scripting, reporting & data visualization	• Decision making and soft skills

https://www.edureka.co/blog/data-analyst-vs-data-engineer-vs-data-scientist/

TERMS AND TECHNOLOGIES COMMONLY USED BY DATA SCIENTISTS

- **Data preparation**: the process of converting raw data into another format so it can be more easily consumed.
- **Data visualization:** the presentation of data in a pictorial or graphical format so it can be easily analyzed.
- **Machine learning:** a branch of artificial intelligence based on mathematical algorithms and automation.
- **Deep learning**: an area of machine learning research that uses data to model complex abstractions.
- **Pattern recognition**: technology that recognizes patterns in data (often used interchangeably with machine learning).
- **Text analytics**: the process of examining unstructured data to glean key business insights.

https://www.sas.com/en_ca/insights/analytics/what-is-a-data-scientist.html

TYPICAL TASKS FOR DATA SCIENTISTS

- Collecting large amounts of unruly data and transforming it into a more usable format.
- Solving business-related problems using data-driven techniques.
- Working with a variety of programming languages, including SAS, R and Python.
- Having a solid grasp of statistics, including statistical tests and distributions.
- Staying on top of analytical techniques such as machine learning, deep learning and text analytics.
- Communicating and collaborating with both IT and business.
- Looking for order and patterns in data, as well as spotting trends that can help a business's bottom line.

https://www.sas.com/en_ca/insights/analytics/what-is-a-data-scientist.html

THINGS TO LEARN IN DATA SCIENCE

1. **Python**
   a. Python Fundamentals (Basic syntax, functions, control flow, loops, modules and classes)
   b. Data Analysis with Python (Numpy, Pandas, Matplotlib)
   c. Python for Machine Learning (scikit-learn)

2. **Structured Query Language (SQL)**
   a. MySQL
   b. Google Cloud BigQuery

3. **R** (R Studio, R Markdown)

4. **Software Engineering** (to help your coding skill)

https://towardsdatascience.com/how-to-learn-data-science-for-free-eda10f04d083

THINGS TO LEARN IN DATA SCIENCE

5. **Deep Learning**
    a. Introduction to Machine Learning
    b. Practical Deep Learning
    c. Computational Linear Algebra
    d. Introduction to Natural Language Processing

6. **Mathematics**
    a. Calculus
    b. Linear Algebra

7. **Statistics**
    a. Descriptive Statistics
    b. Inferential Statistics
    c. Experimental Design
    d. Linear and Nonlinear Regression
    e. Classification (Cluster Analysis)

https://towardsdatascience.com/how-to-learn-data-science-for-free-eda10f04d083

Pathway to learning Data Science

There is no particular pathway that is applicable to all because people have different learning styles.

1) To start learning, go to the Kaggle Data Science Micro Courses https://www.kaggle.com/learn/overview The courses introduce you to the basics of Python, Machine Learning and Deep Learning.
2) Go through other people's Kaggle workbook and start writing your own projects. Kaggle has a community where you can ask questions. There are practical exercises to help understand Data Science projects.
3) Next course to take is the Google Machine Learning Crash Course. The course has a video component that is very informative. https://developers.google.com/machine-learning/crash-course
4) The third course is Fast.ai to learn Deep Learning concepts. https://www.fast.ai/

To learn Python, you can also learn from the Python Documentation Website.
https://wiki.python.org/moin/BeginnersGuide/Programmers

https://www.youtube.com/watch?v=Ip50cXvpWY4

Data Science through MOOC Programmes

To start learning about Data Science, take advantage of available MOOC (Massive Open Online Course) programmes. You can audit the courses for FREE but to earn a Certificate, you need to pay.

Harvard University Professional Certificate in Data Science
https://www.edx.org/professional-certificate/harvardx-data-science

University of Michigan Applied Data Science with Python Specialization
https://www.coursera.org/specializations/data-science-python

Professional Certificate in IBM Data Science
https://www.edx.org/professional-certificate/ibm-data-science

Codecademy Data Science
https://www.codecademy.com/catalog/subject/data-science

Pathway to becoming a Data Scientist

1. Each Data Scientist has a different story of the pathway suitable to them.
2. The Internet has a lot of resources to help you but you must do the searching.
3. Self-discipline and commitment to create time to learn is key.
4. Self motivation and passion are required to push on when the journey becomes tough.
5. Learn along with others can help you maintain focus and encourage you to press on.

The following links will be of help:

- https://www.learnhowtobecome.org/data-scientist/
- https://www.dataquest.io/blog/how-to-become-a-data-scientist/
- https://www.kdnuggets.com/2020/11/data-science-without-degree.html

Data Science Curriculum for self-study:

https://www.kdnuggets.com/2020/02/data-science-curriculum-self-study.html

THANK YOU