Data Science

(Definition, Importance, Career-Paths & Challenges)

DR. JASMEET SINGH
ASSISTANT PROFESSOR,
CSED, TIET

What is Data Science ?

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.
- Relatively young field
- The term often broadly used, due to its hype
- “Definition” (e.g., what it is, what it includes) often hotly debated; a controversial topic!

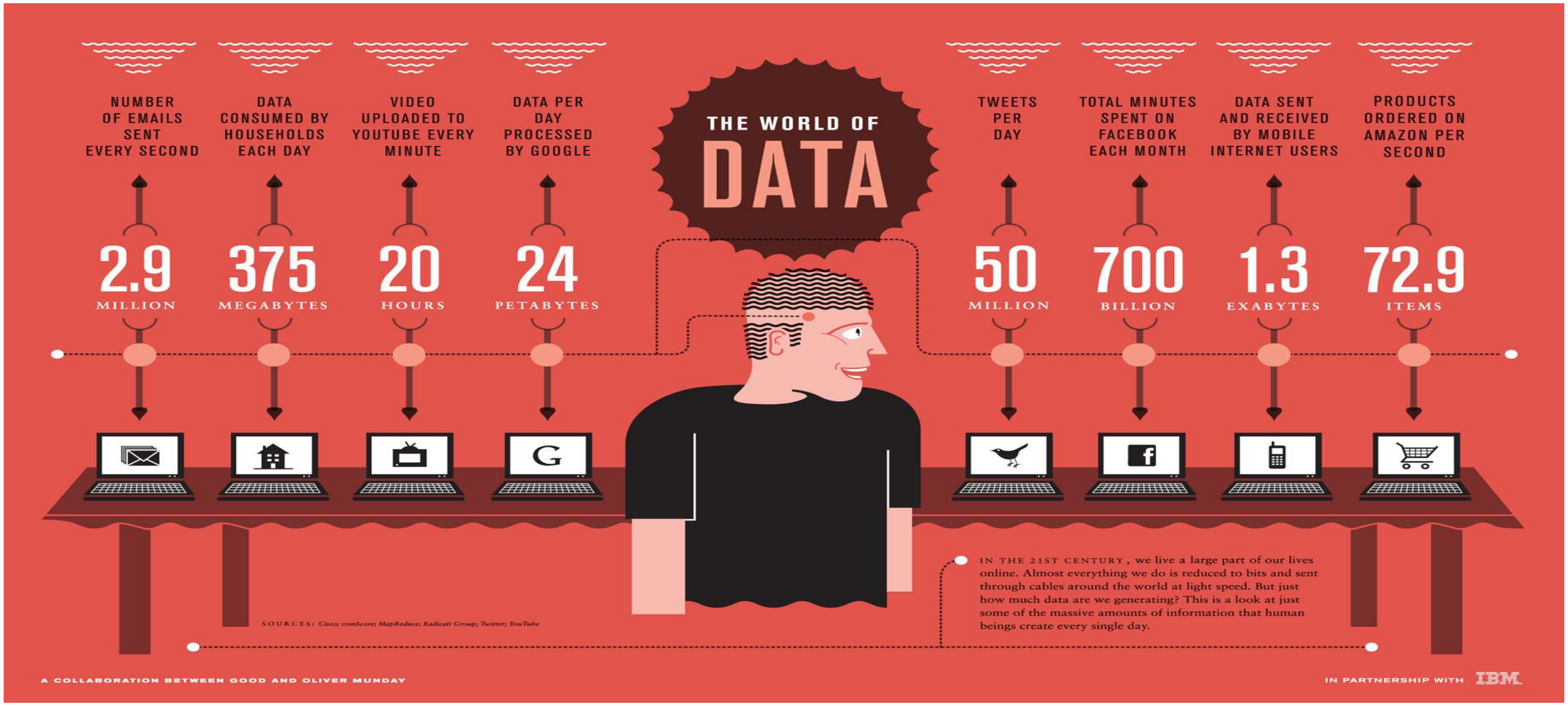
What is Data Science?

Why so complex?

- Data science is interdisciplinary, due to its goal to aid discoveries, decision making, etc.
  - Applicable to many domains (e.g., sciences, finance, healthcare, etc.)
  - Often requires analysis of large amount of data

    (helpful to use scalable algorithms, distributed computing, etc.)
  - Mathematics and Statistics is a pillar of data science
  - Users may need to explore data, and understand and present analysis results (benefits from visualization, good user interface design, etc.)

Why is Data Science Important?


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
G
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
f
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

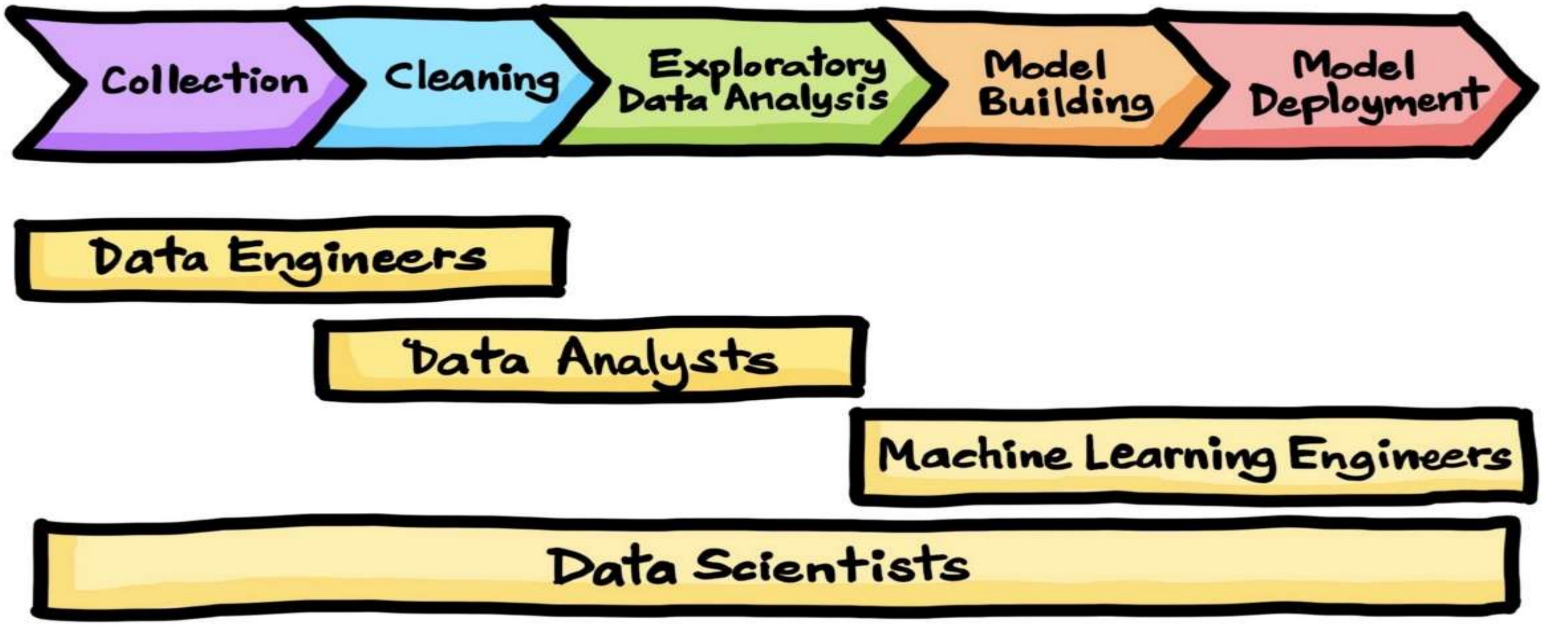
SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

Why is Data Science Important?

- Data Science **unlocks potential** of data in solving **societal challenges** and **large-scale complex problems across domains**, from business, technology, science, engineering, healthcare, to government, and many more.
- As data continues to grow in volume, velocity and complexity, there is a **strong demand** for data science talents to help design the best solutions.
- Data-analytic thinking helps you
  - Assess whether and how data can address a problem
  - Envision opportunities for data-driven decision making
  - See data-oriented competitive threats

What does a Data Scientist do ?

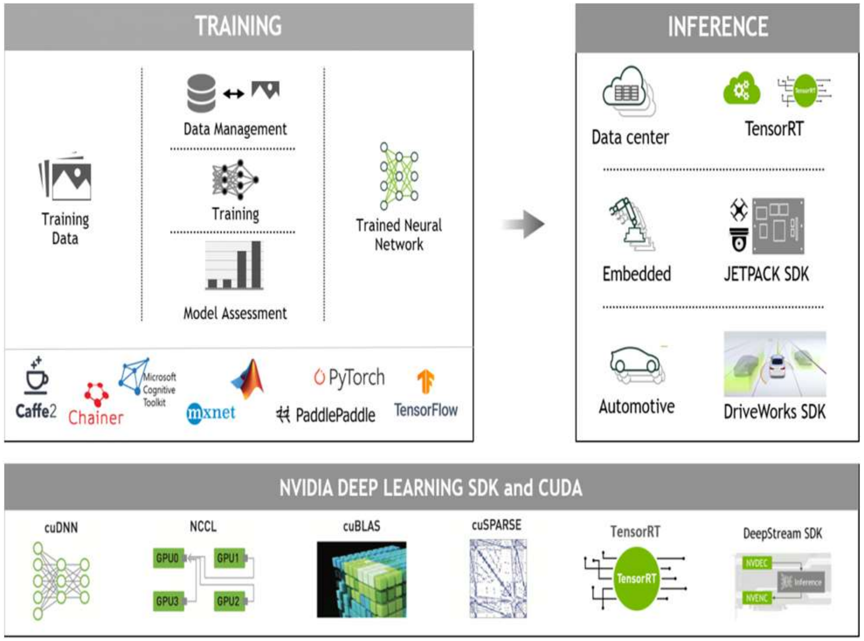
(Data Science Process)



What does a Data Scientist do ?

(Data Science Process)

- As a data scientist you might be responsible for any aspect of data science including Data Collection, Data Preprocessing, Data Visualization, Data Analytics & Application
- This figure shows NVIDIA’s ecosystem framework for processing data with deep learning with support of NVIDIA tools such as the deep learning SDK and CUDA, and different deep learning platforms like Google’s TensorFlow and Apache Mxnet.
- These tools work together to accomplish data analysis such as TRAINING and INFERENCE, and implement applications such as Autodrive.



Data Science Process- Data Collection

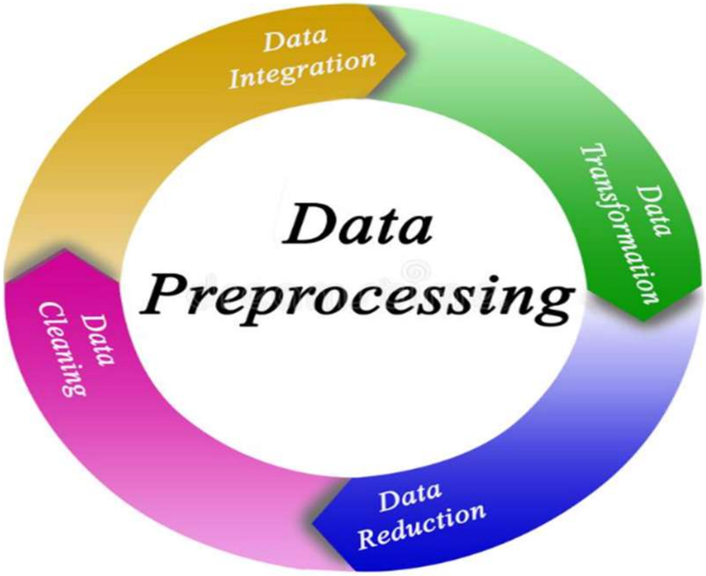
- One of the first task for data scientist is to collect data in terms of features of project tasks.
- Large data sets ("Big data") can be already be available for different projects such as computer vision and natural language processing with the development of IoT.
- We however should be careful to collect useful data since more and more "noise" comes with more and more data.

Data Science Process- Data Collection

- One of the first task for data scientist is to collect data in terms of features of project tasks.
- Large data sets ("Big data") can be already be available for different projects such as computer vision and natural language processing with the development of IoT.
- We however should be careful to collect useful data since more and more "noise" comes with more and more data.

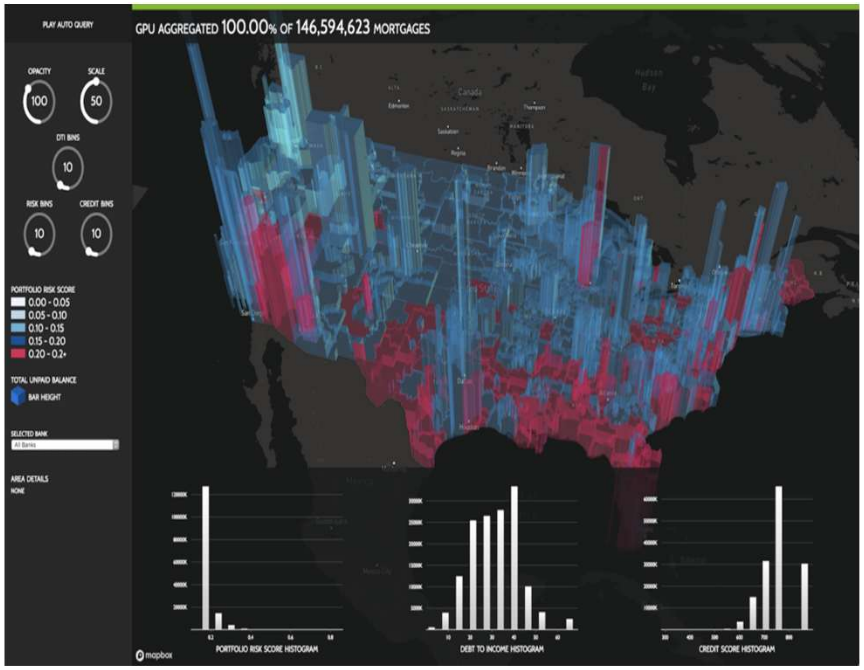
Data Science Process- Data Collection

- Data Preprocessing is a complex task to prepare raw data for data analysis.
- It includes subtasks such as data integration, data cleaning, data reduction, and data transformation.
  - For example, data integration will merge different sources of data.
- Each subtask is a hot topic on data science research.



Data Science Process- Exploratory Data Analysis

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- Data Visualization is visualizing data collected to present data features with figures.
- Here is an example of visualization of mortgage data risk assessment.



GPU AGGREGATED 100.00% OF 146,594,623 MORTGAGES

OPACITY 100

SCALE 50

DTI BINS 10

RISK BINS 10

CREDIT BINS 10

SELECTED BANK All Banks

AREA DETAILS NONE

Data Science Process- Building Applications

- Finally, with the help of data collection, data preprocessing, and data visualization, data scientists can build applications.

- Data scientists then design and verify machine and deep learning models on data collected.

Skills Needed to Be a Data Scientist

- As a data scientist you will need to work with other domain experts for many projects.
- You therefore will need different skills which includes
  - Analytical skills,
  - Communication skills,
  - Critical and Logical thinking skills,
  - Maths skills,
  - Programming skills

Skills Needed to Be a Data Scientist (Contd…)

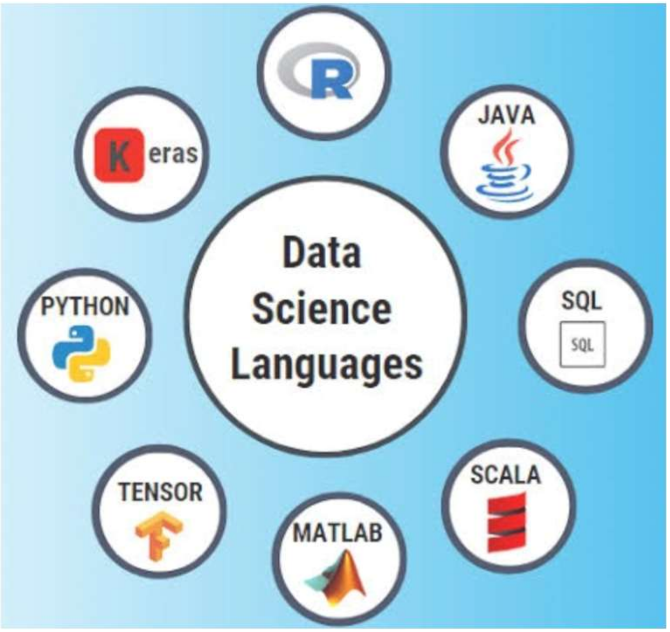
- **Analytical Skills**
  - Analytical skills will help you investigate a problem and find the ideal solution in a timely, efficient manner.
  - They will help you complete detecting patterns, brainstorming, observing, interpreting data, integrating new information, theorizing, and making decisions.
  - Finally, with these skills, you will design solutions.
- **Communication Skills**
  - You need to be an effective communicator to discuss data patterns, conclusions, and recommendations with others involved in the project.
  - Such communication skills include problem sensitivity, teamwork, oral and written communication, and presentation skills.

Skills Needed to Be a Data Scientist (Contd…)

- **Critical and Logical Thinking**
  - Critical thinking will help resolve problems through evaluating information collected.
  - For data analysis, it is preferable to learn skills in information evaluation, deductive reasoning, data interpretation and inductive reasoning.
- **Maths Skills**
  - Math skills are very important to data analysis since it can help design models to resolve problems.
  - This includes calculus, algebra, probability, Stochastic process, and dynamic programming.
  - For example, deep learning models are built through combining methods from calculus, algebra, information theory, and probability.

Skills Needed to Be a Data Scientist (Contd…)

- **Computer Programming Competency**
  - Finally, you have to implement the solutions (models) to solve the problems by programming with certain programming languages.
  - So far it seems the Python programming language dominates data analysis with AI techniques.
  - For biomedical data analysis, R is viewed by many as best programming language to complete various tasks as it provides different powerful packages and libraries.
  - Keras is a platform that simplifies programming since it provides high-level interfaces to implement deep learning models efficiently.



Related Jobs in Data Science

Data Engineer

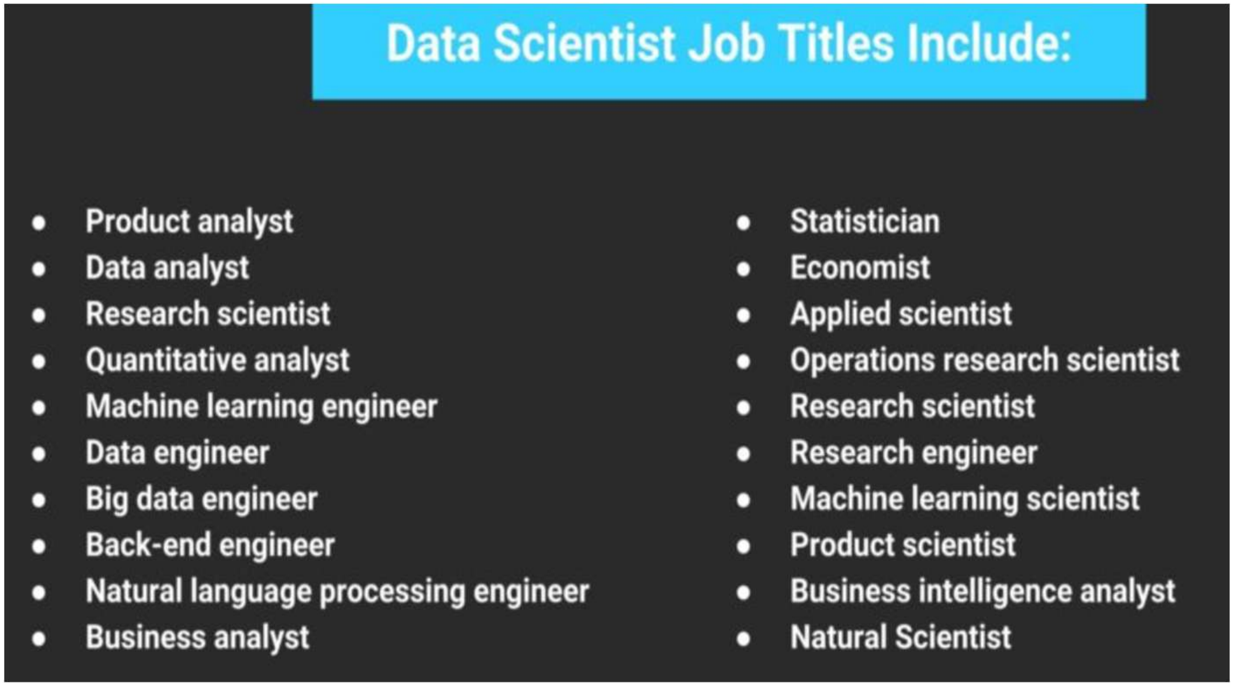
Data analyst

Machine learning engineer

Data Scientist

Big data engineer

…



Data Scientist Job Titles Include:

- Product analyst
- Data analyst
- Research scientist
- Quantitative analyst
- Machine learning engineer
- Data engineer
- Big data engineer
- Back-end engineer
- Natural language processing engineer
- Business analyst
- Statistician
- Economist
- Applied scientist
- Operations research scientist
- Research scientist
- Research engineer
- Machine learning scientist
- Product scientist
- Business intelligence analyst
- Natural Scientist

Challenges in Data Science Careers

1. **Insights not used in Decision Making**: Insights obtained from data are not very useful in decision making because of their weak interpretable features.
2. **Data Privacy, Veracity, Unavailability**: Big data is not available with respect to the data privacy in some fields like medical domains.
3. **Limitations of tools to scale/deploy**: Few of tools are available to visualize high-dimensional data.
4. **Wrong Questions Asked**: Sometimes, it is not easy to formulate the problems, and they will lead to incorrect directions.

Data Analytics-Building Blocks

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination

Can skip some

Can go back (two-way street)

- Data types inform visualization design
- Data informs choice of algorithms
- Visualization informs data cleaning (dirty data)
- Visualization informs algorithm design
  (user finds that results don’t make sense)