Module 1: Introduction to Data Science

CSCI1360

Outline

Data Science Definition

Why Data Science is Important

Reasons Behind Data Growth

Applications

Data Types

Data Preprocessing

Data Representation

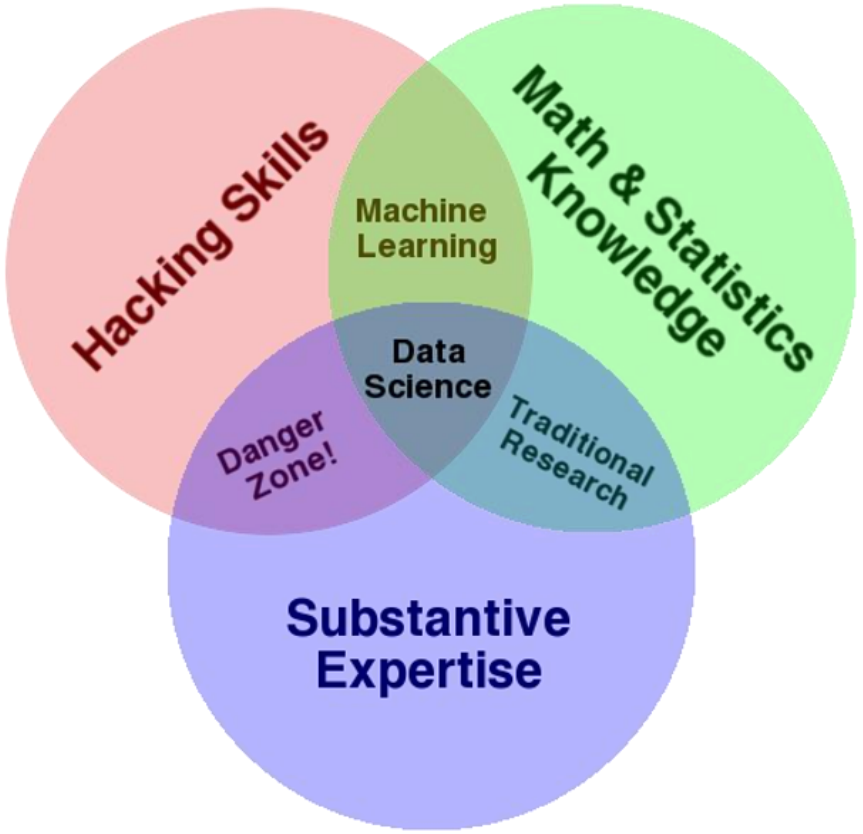
Data Visualization

Data Science Tools

Ethical Considerations

Data Modeling

Data Science

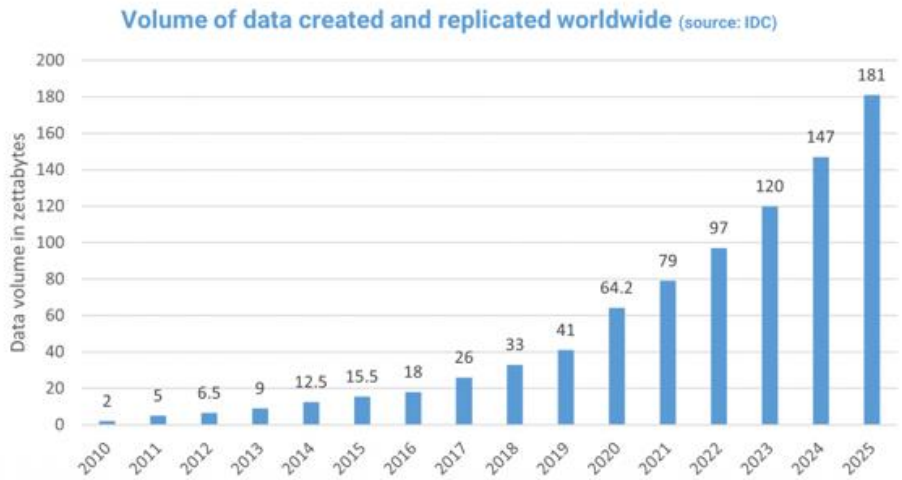


Definition

- Data Science as a proper field of study is the confluence of three major aspects:
  - Hacking skills: the ability to code, and knowledge of available tools.
  - Math and statistics: strong quantitative skills that can be implemented in code.
  - Substantive expertise: some specialized area of emphasis.

Why Data Science is Important?

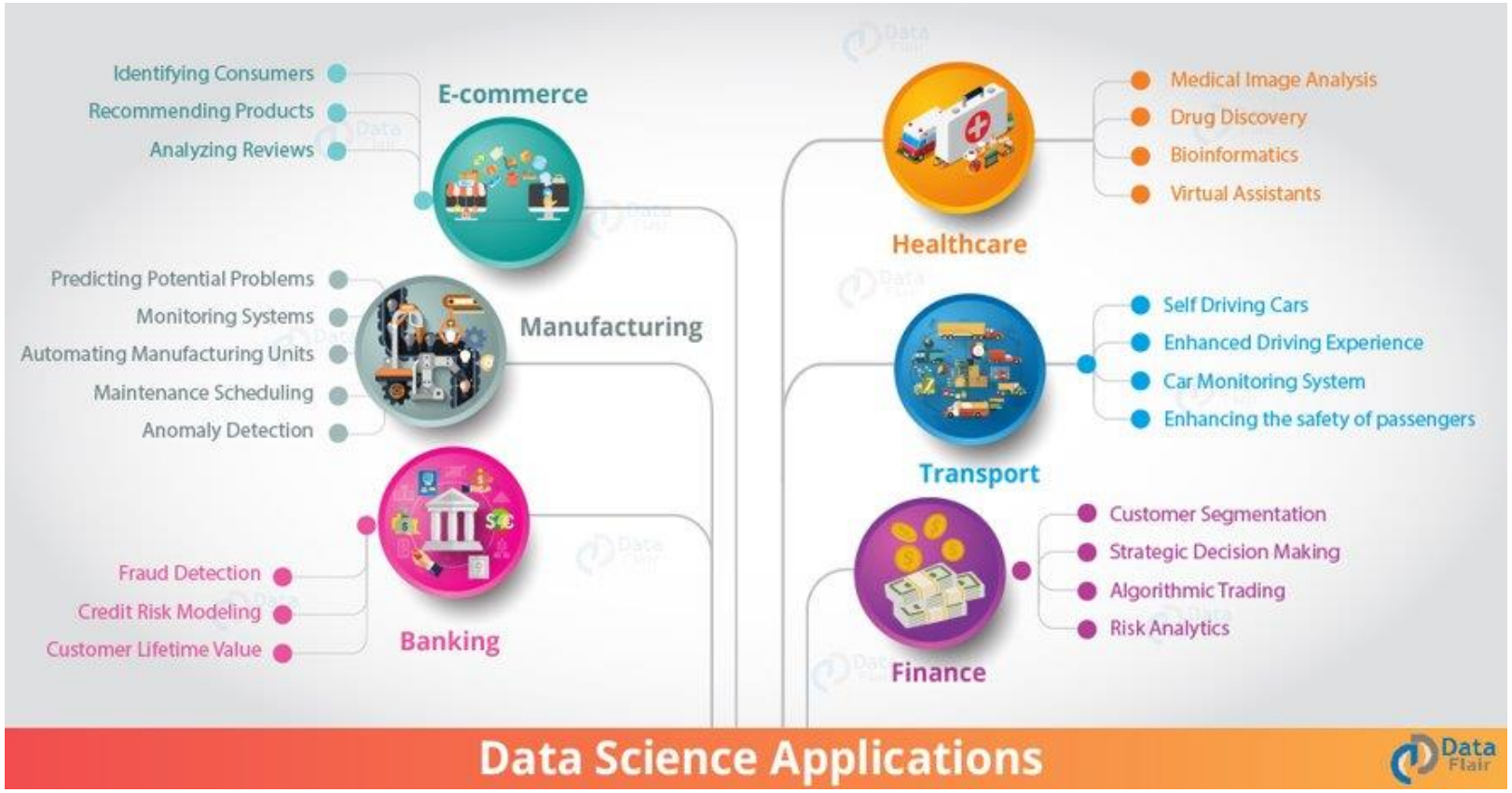
- Observe data growth
- Zettabyte = 1 Billion Terabyte
- Data Science is important because:
  - Helps extracting meaningful information from data
  - Enhances decision making
  - Process automation



Reasons Behind Data Growth

- Storage Technologies and Prices
  - Fast on-premise storage
  - Cheap cloud storage
- Computation Technologies and Prices
  - CPUs and GPUs capabilities
- Data Collection Technologies
  - Wide usage of IoT device: Sensors and Cameras
  - Social Media Usage

Data Science Applications



https://data-flair.training/blogs/data-science-applications/

Data Types and Sources

- Data types:
  - Structured: Excel Spreadsheet
  - Un-structured: Audio
  - Semi-structured: Log file

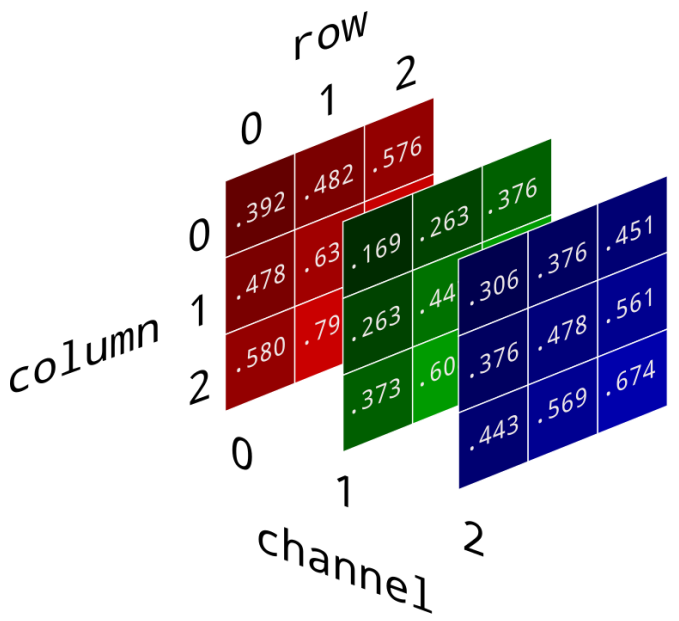
- Data types by format examples:
  - Numerical Data: Bank transactions
  - Text: Movie reviews
  - Images: X-Ray images

Data Preprocessing

- Data cleaning is an important step before running data science algorithms
- Preparing data for analysis incorporates:
  - Removing corrupted values
  - Imputing or removing missing values
  - Transforming data to a numerical format
  - Scaling values

Data Transformation

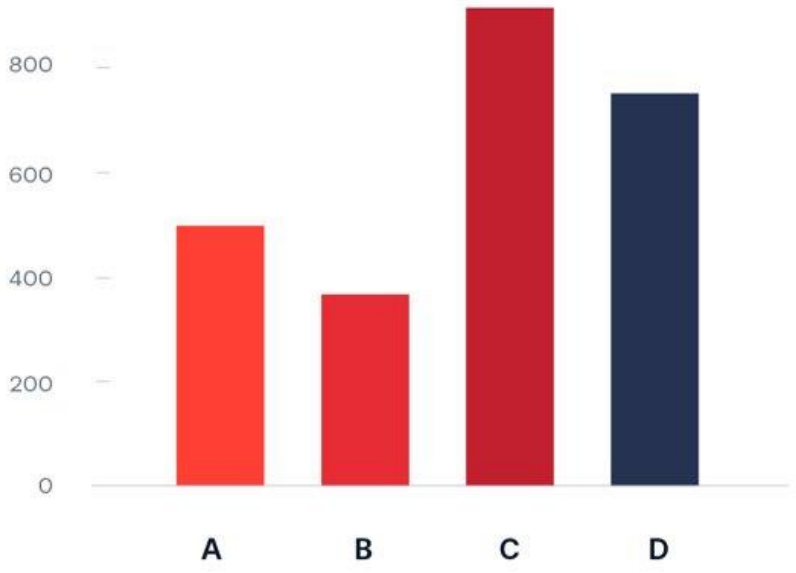
- Computers can process data in the form of numbers
- Images can be represented in the Red-Green-Blue (RGB) format
- Text data can be represented using the bag of words model



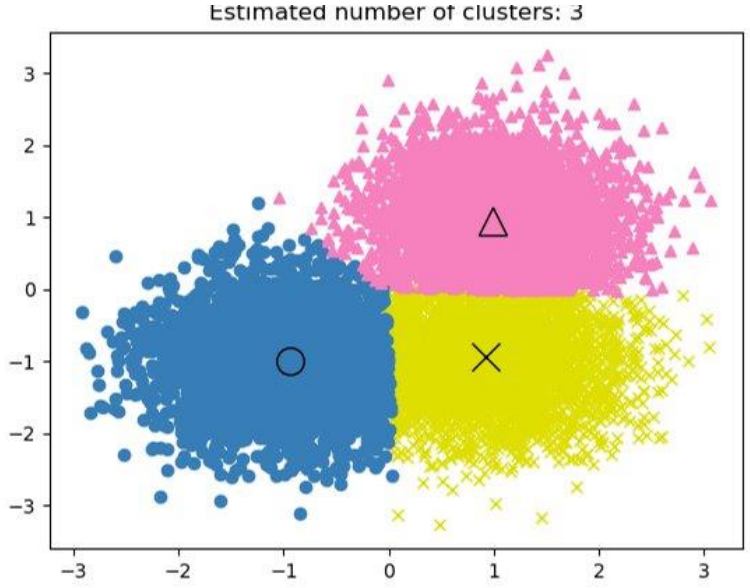
Data Visualization

- Why visualization is important?
  - To understand data
  - To share data easily
  - To visualize patterns

Visualization Examples



Bar Chart

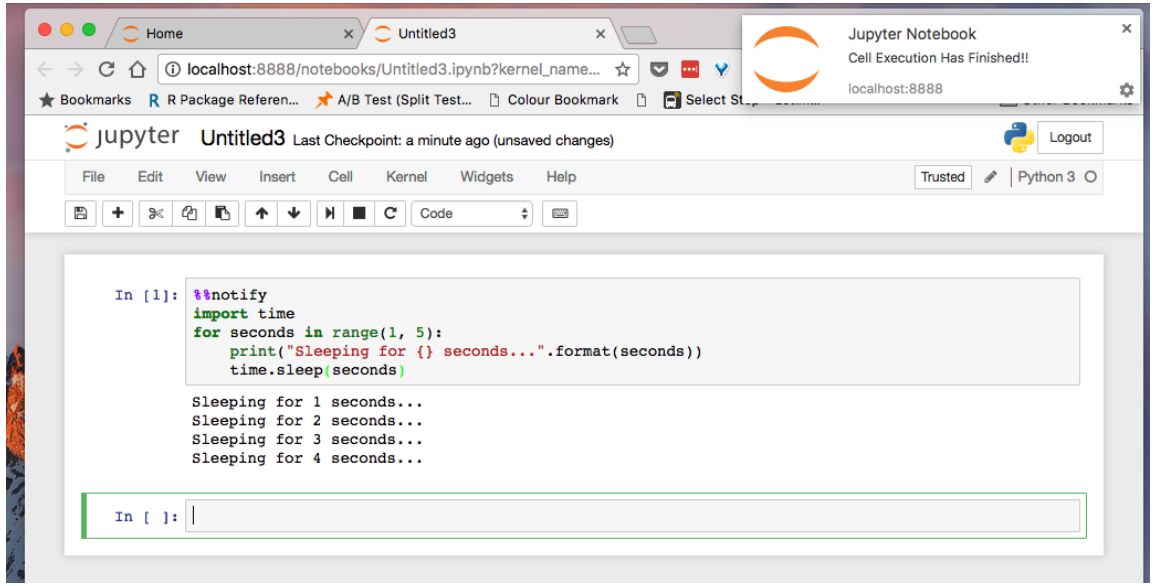


Scatter Plot

Visualization Examples

Word Cloud

Image Segmentation



Data Science Tools

- Programming language:
  - Python, R
- Code Editor:
  - Jupyter Notebook, Visual Studio Code
- Algorithm packages:
  - Scikit Learn, Pandas, Numpy
- Visualization:
  - Matplotlib

Ethical Considerations

Privacy

- Anonymizing sensitive data to prevent the identification of individuals.

Data Ownership

- Clarifying who owns the data

Bias and Fairness

- Mitigating algorithms bias to certain group

Data Modeling

- Discriminative modeling
  - Learn the boundaries between data samples
- Generative modeling (beyond the scope of the course)