Data Science

DAI-101 Spring 2025-26

Dr. Devesh Bhimsaria

Office: F9, Old Building

Department of Biosciences and Bioengineering

Indian Institute of Technology–Roorkee

devesh.bhimsaria@bt.iitr.ac.in

About the Course

- Contact Hours (Hrs. per week): L:3 T:1 P:0
- Total contact hours: 42
- Credits: 4
- Prerequisite: None.
- Taught by: Dr. Devesh Bhimsaria & Dr. Deepak Sharma
- Ran in 5 Batches

Course Outline

- Introduction to Data Science: Latest and greatest in data science
- Python programming
- Data Analysis Foundation: Types of data (data matrix, numeric, categorical datasets), data preparation: data cleaning, data reduction and transformation
- Exploratory Data Analysis and Visualization: Univariate and bivariate analysis, data visualization
- Statistical Analysis: Confidence Intervals, Hypothesis Testing, p-values, Bias and Variance trade-off
- Machine Learning: introduction to supervised and unsupervised methods, model training, overfitting and underfitting, bias and variance, introduction to supervised methods: regression and classification (Linear regression, logistic, decision trees, SVM), Clustering, K-means, PCA
- Deep learning and Big Data: Gradient Descent, Neural nets, Convolutional Neural Networks, Big Data technologies (MapReduce, HDFS)

Evaluation

- Mid-term exam
- End-term exam
- Assignments
- Attendance
- If there is any modification, you'll be informed in advance

Rules and other points

- Maintain class decorum.
- Timely Submission of assignments.
- For any help related to the course or otherwise – a) You can email me, b) ask me during lecture/tutorial.
- If urgent, CR may call or message.

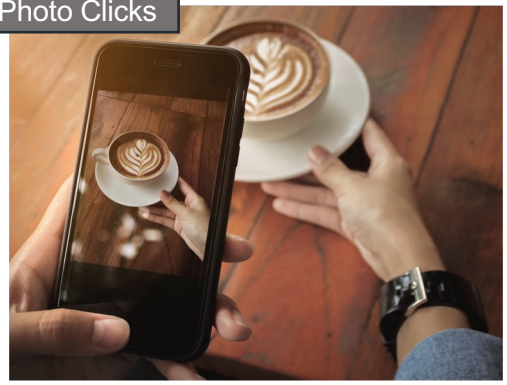
The Age of Data

Our every action generates data

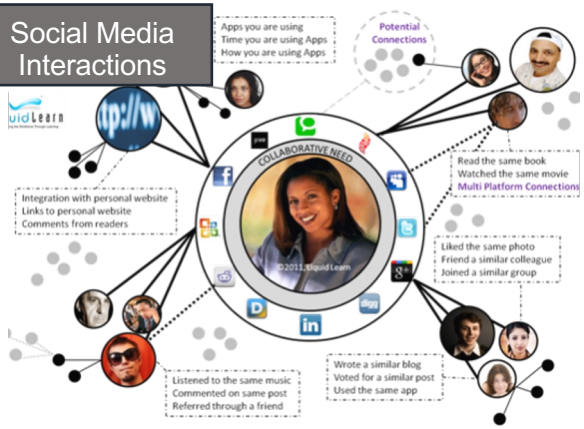
Experiments

Smart Wearables

Photo Clicks



Social Media Interactions



Google Searches

Stock Market Shares

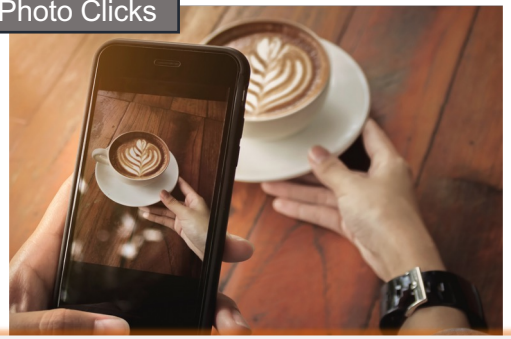
The Age of Data

Our every action generates data

Experiments

Smart Wearables

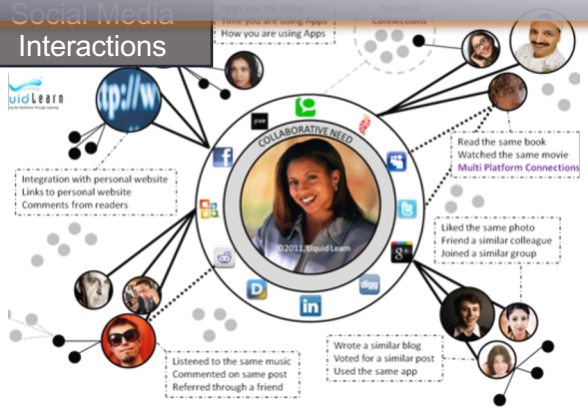
Photo Clicks



328,770,000 TB of data generated per day

How do we make sense of it?

Social Media Interactions



Google Searches

Stock Market Shares

What is Data Science?

- Definition: An interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- Key Components:
  - Data Collection
  - Data Cleaning and Preparation
  - Data Analysis and Visualization
  - Machine Learning/AI

Importance

- Real-world applications:
  - Healthcare: Predicting diseases
  - Business: Customer segmentation
  - Finance: Fraud detection
  - Social Media: Recommendation systems
- Industry growth and demand for data professionals

Skills Required for DS analysis

- Technical Skills:
  - Programming: Python, R, SQL
  - Data Manipulation: Pandas, NumPy
  - Visualization: Matplotlib, Seaborn
  - Machine Learning: Scikit-learn, TensorFlow

Challenges in Data Science

- Data Quality Issues: Missing, noisy, or inconsistent data
- Data Privacy & Ethics: Ensuring compliance with regulations
- Model Interpretability: Explaining complex models
- Scalability: Handling large datasets

Thank You

- All my slides/notes excluding third party material are licensed by various authors including myself under https://creativecommons.org/licenses/by-nc/4.0/