Introduction to Applied Data Science

Kaustubh Kulkarni

Assistant Professor

Department of Computer Engineering

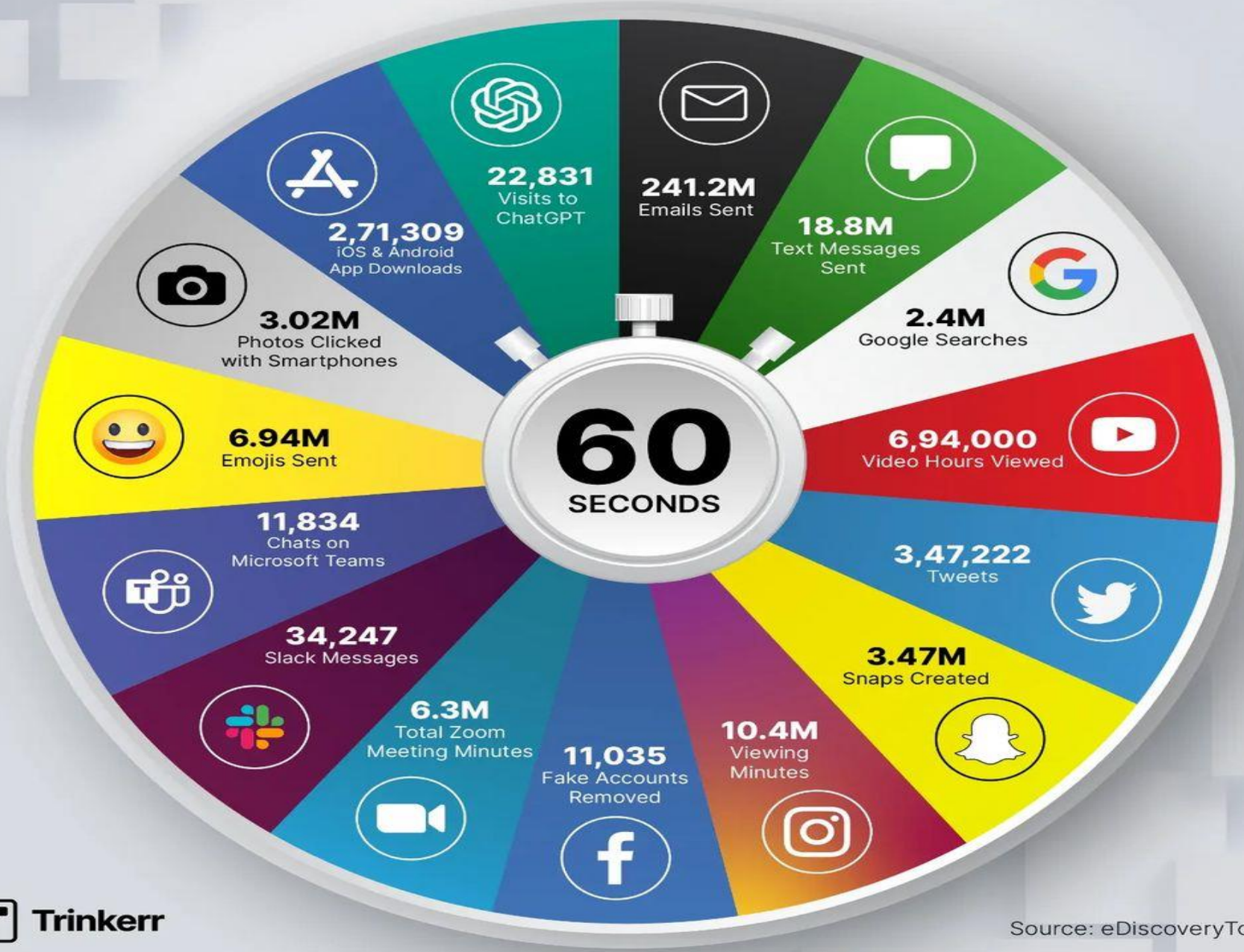
K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

Outline

- Datafication- Data everywhere
- Big Data
- What is Data Science?
- Big Data and Data Science
- Current landscape of perspectives
- Data Scientist Skill sets
- Challenges and skill Sets needed and various applications areas.
- Impact of applying Data Science in business scenario
- Estimation and validation for added value due to data science

EVERY MINUTE OF INTERNET IN 2023



Trinkerr

Source: eDiscoveryToday

Datafication

Definition

- Datification is about taking a process or activity that was previously invisible and turning it into data.
- That data can then be tracked, monitored, and optimized, leading to new opportunities — and new challenges.

Datafication

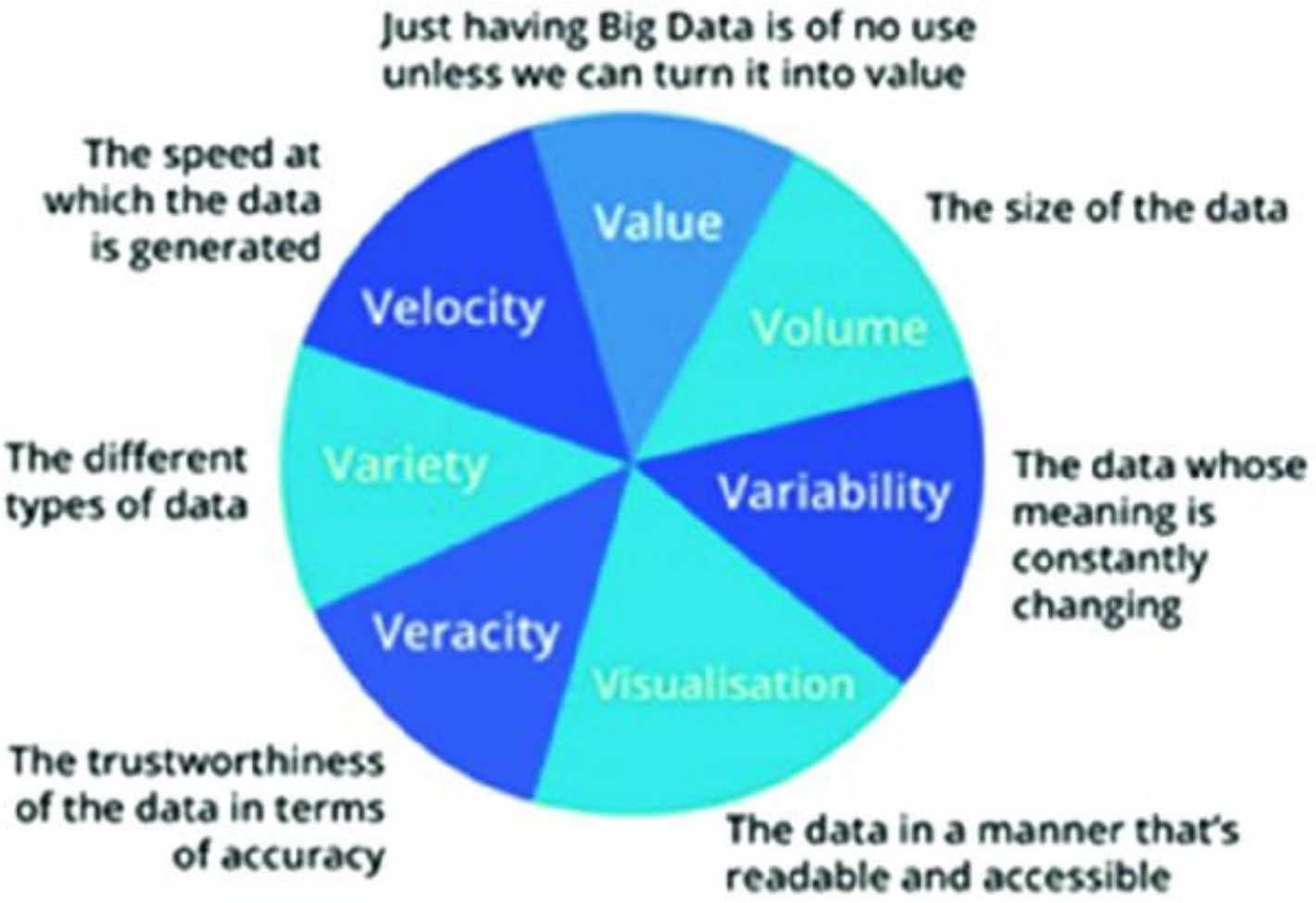
Example

- Quantify friends with ‘likes’
- Twitter
- LinkedIn
- Browsing web, with cookies
- Walk in store, street we are datafied via sensors, cameras, etc.
- Taking part of social media experiment

What is Big Data?

- The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity.
- This is also known as the three Vs.
- Two more Vs have emerged over the past few years: **value** and **veracity**.
- Data has intrinsic value. But it's of no use until that value is discovered.
- Equally important: How truthful is your data—and how much can you rely on it?

7 V's of Big Data



What is data science?

- Data science is the study of data to extract meaningful insights for business.
- It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.
- This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

What are the data science techniques?

- **Classification**
- Classification is the sorting of data into specific groups or categories.
- Computers are trained to identify and sort data.
- Known data sets are used to build decision algorithms in a computer that quickly processes and categorizes the data.
- For example:·
  - Sort products as popular or not popular·
  - Sort insurance applications as high risk or low risk·
  - Sort social media comments into positive, negative, or neutral.

What are the data science techniques?

- **Regression**
- Regression is usually modeled around a mathematical formula and represented as a graph or curves.
- When the value of one data point is known, regression is used to predict the other data point.
- For example:
  - The rate of spread of air-borne diseases.
  - The relationship between customer satisfaction and the number of employees.
  - The relationship between the number of fire stations and the number of injuries due to fire in a particular location.

What are the data science techniques?

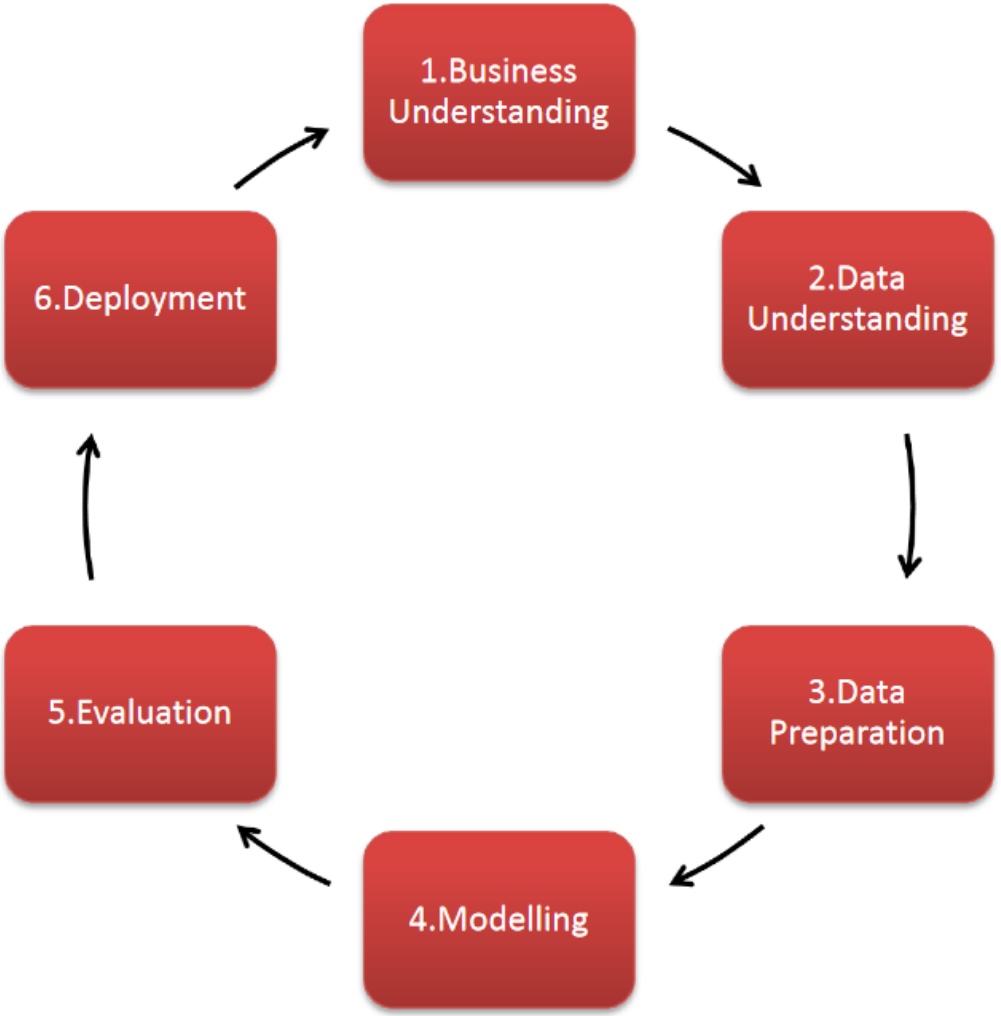
- **Clustering**
- Sometimes the data cannot be accurately classified into fixed categories.
- Hence the data is grouped into most likely relationships.
- New patterns and relationships can be discovered with clustering.
- For example: ·
  - Group customers with similar purchase behavior for improved customer service.·
  - Group network traffic to identify daily usage patterns and identify a network attack faster.
  - Cluster articles into multiple different news categories and use this information to find fake news content.

A Data Science Profile

- https://www.coursera.org/articles/data-scientist-skills

Data Science Process – CRISP-DM

- **Cross-Industry Standard Process for Data Mining**



Question ?