Data Science

Modelling, concepts, techniques

Ayyub Sheikhi

Department of Statistics

Shahid Bahonar university of Kerman

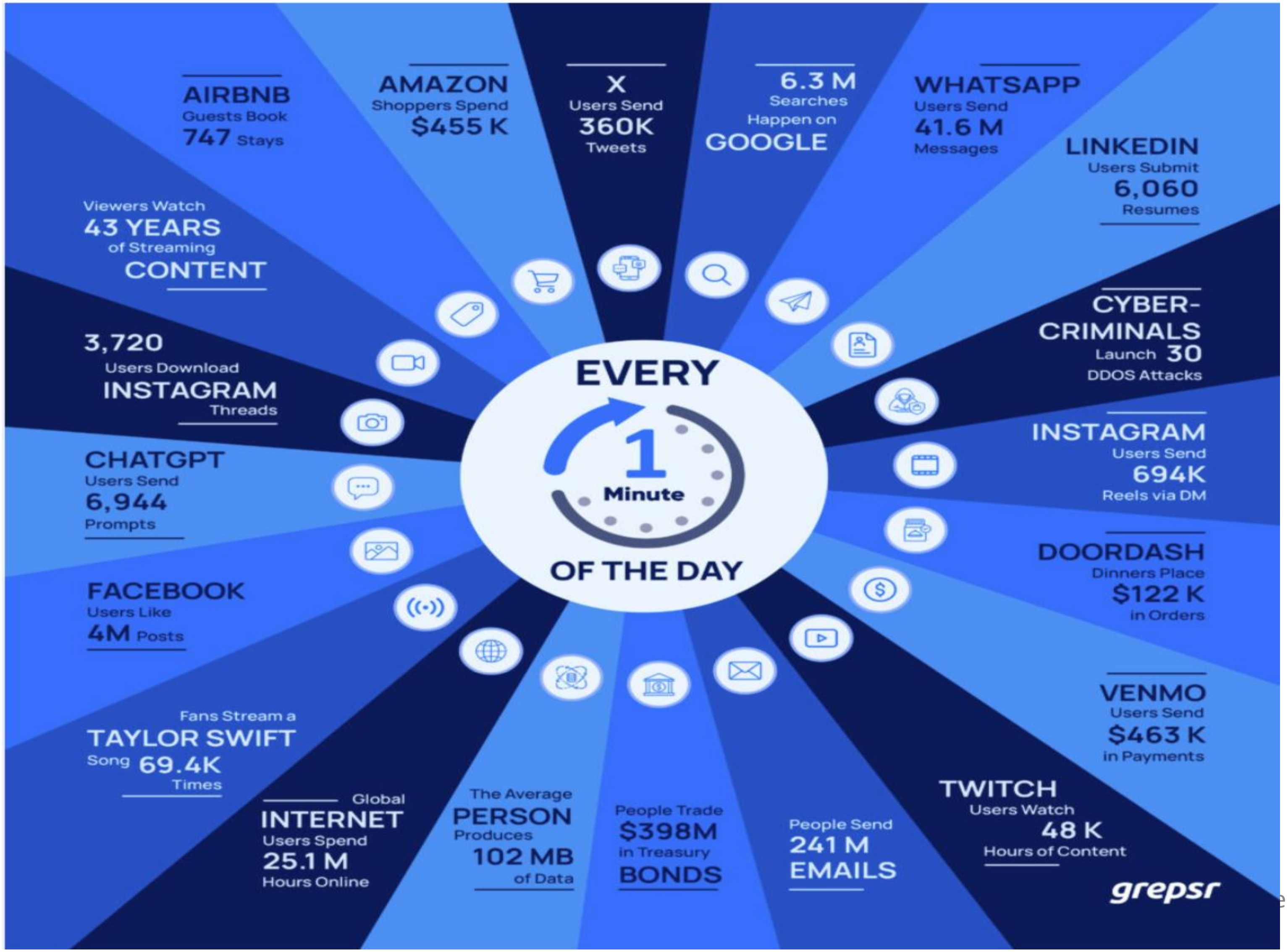
Goals

- ✓Know what Data Science is and learn the basic algorithms
- ✓Perform Data Science techniques
- ✓Know how to apply algorithms to real-world applications

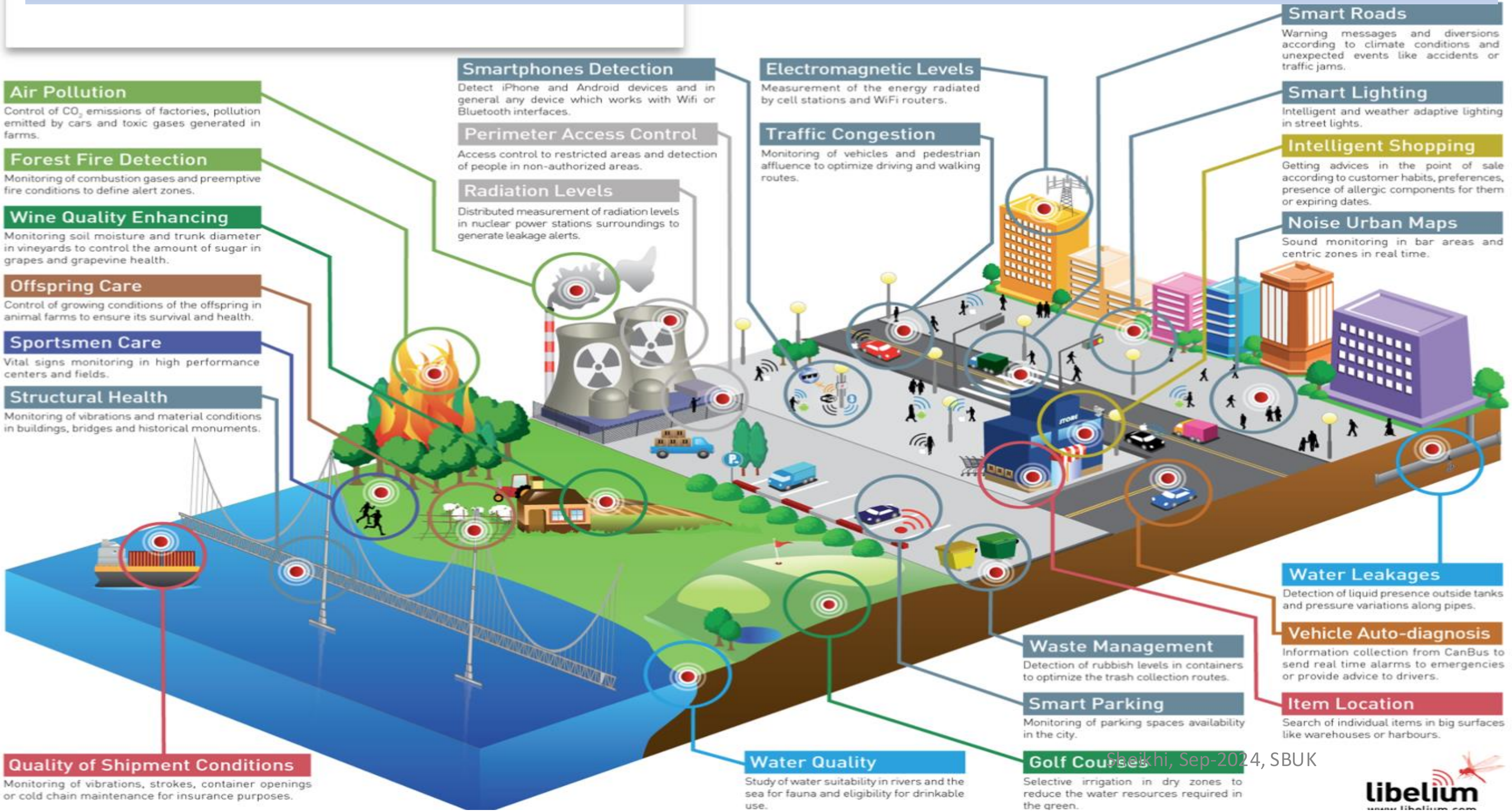
Chapter 1. Introduction

- Why Data Science ?
- What is Data Science ? How big is Data Science
- Characteristics of Data Science
- Top most popular Data Science algorithms
- Major issues in Data Science

Why Data Science


AIRBNB Guests Book 747 Stays
AMAZON Shoppers Spend $455 K
X Users Send 360K Tweets
6.3 M Searches Happen on GOOGLE
WHATSAPP Users Send 41.6 M Messages
LINKEDIN Users Submit 6,060 Resumes
Viewers Watch 43 YEARS of Streaming CONTENT
3,720 Users Download INSTAGRAM Threads
CYBER-CRIMINALS Launch 30 DDOS Attacks
INSTAGRAM Users Send 694K Reels via DM
CHATGPT Users Send 6,944 Prompts
EVERY
1
Minute
OF THE DAY
DOORDASH Diners Place $122 K in Orders
FACEBOOK Users Like 4M Posts
VENMO Users Send $463 K in Payments
Fans Stream a TAYLOR SWIFT Song 69.4K Times
Global INTERNET Users Spend 25.1 M Hours Online
The Average PERSON Produces 102 MB of Data
People Trade $398M in Treasury BONDS
People Send 241 M EMAILS
TWITCH Users Watch 48 K Hours of Content
grepsr

Why Data Science



Air Pollution
Control of CO_2 emissions of factories, pollution emitted by cars and toxic gases generated in farms.

Forest Fire Detection
Monitoring of combustion gases and preemptive fire conditions to define alert zones.

Wine Quality Enhancing
Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

Offspring Care
Control of growing conditions of the offspring in animal farms to ensure its survival and health.

Sportsmen Care
Vital signs monitoring in high performance centers and fields.

Structural Health
Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

Quality of Shipment Conditions
Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

Smartphones Detection
Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

Perimeter Access Control
Access control to restricted areas and detection of people in non-authorized areas.

Radiation Levels
Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

Electromagnetic Levels
Measurement of the energy radiated by cell stations and WiFi routers.

Traffic Congestion
Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

Water Quality
Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

Golf Courses
Selective irrigation in dry zones to reduce the water resources required in the green.

Waste Management
Detection of rubbish levels in containers to optimize the trash collection routes.

Smart Parking
Monitoring of parking spaces availability in the city.

Smart Roads
Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

Smart Lighting
Intelligent and weather adaptive lighting in street lights.

Intelligent Shopping
Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

Noise Urban Maps
Sound monitoring in bar areas and centric zones in real time.

Water Leakages
Detection of liquid presence outside tanks and pressure variations along pipes.

Vehicle Auto-diagnosis
Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

Item Location
Search of individual items in big surfaces like warehouses or harbours.

Some links of Data Science

https://everysecond.io

https://www.worldometers.info/coronavirus

http://irsc.ut.ac.ir

https://finance.yahoo.com/quote/BTC-USD/history

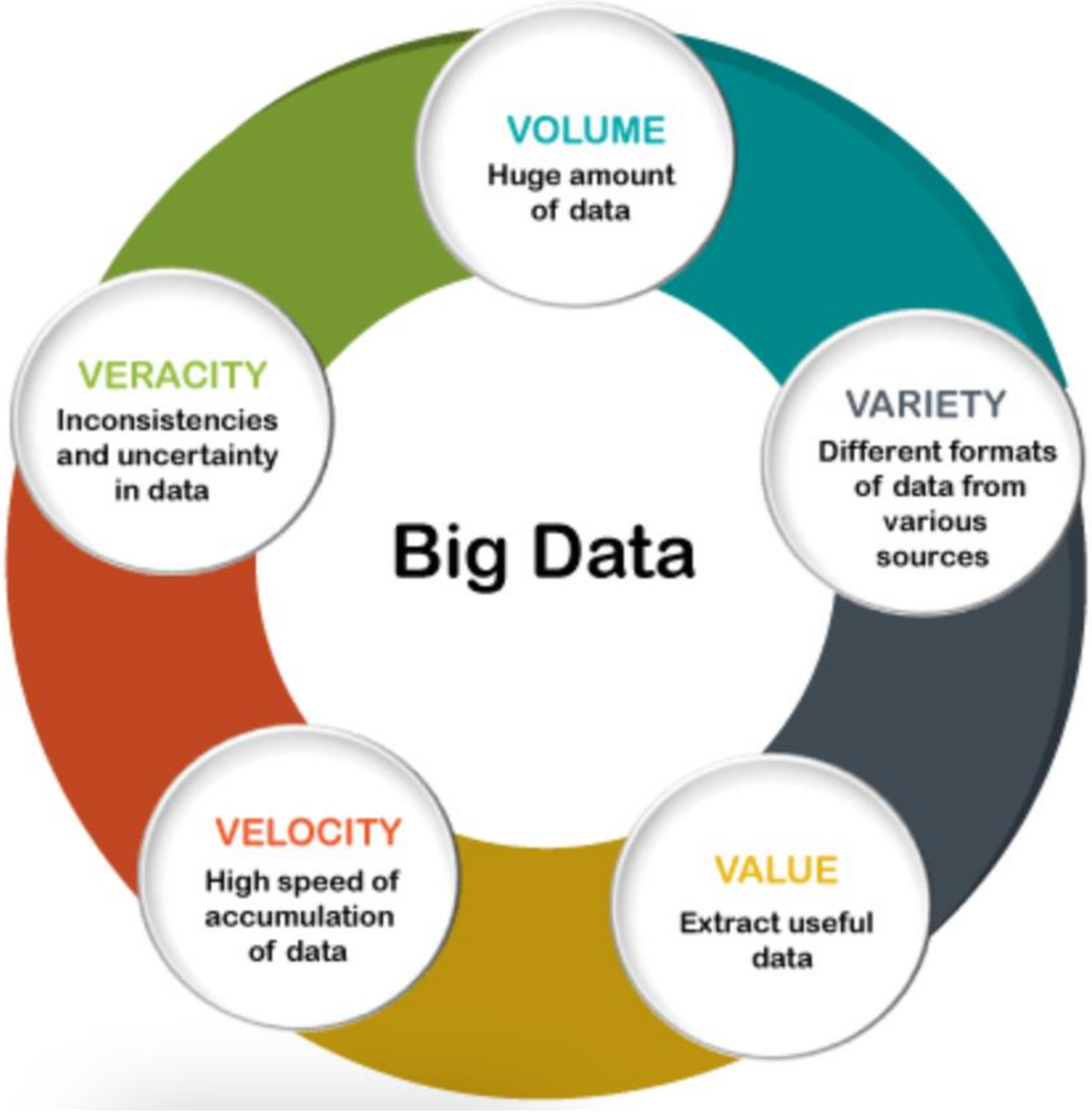
Why Data Science ?

- The Explosive Growth of Data: from terabytes to petabytes
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
    - Healthcares, recording patient symptom using online monitoring

Evolution of Sciences (from empirical sciences to data science)

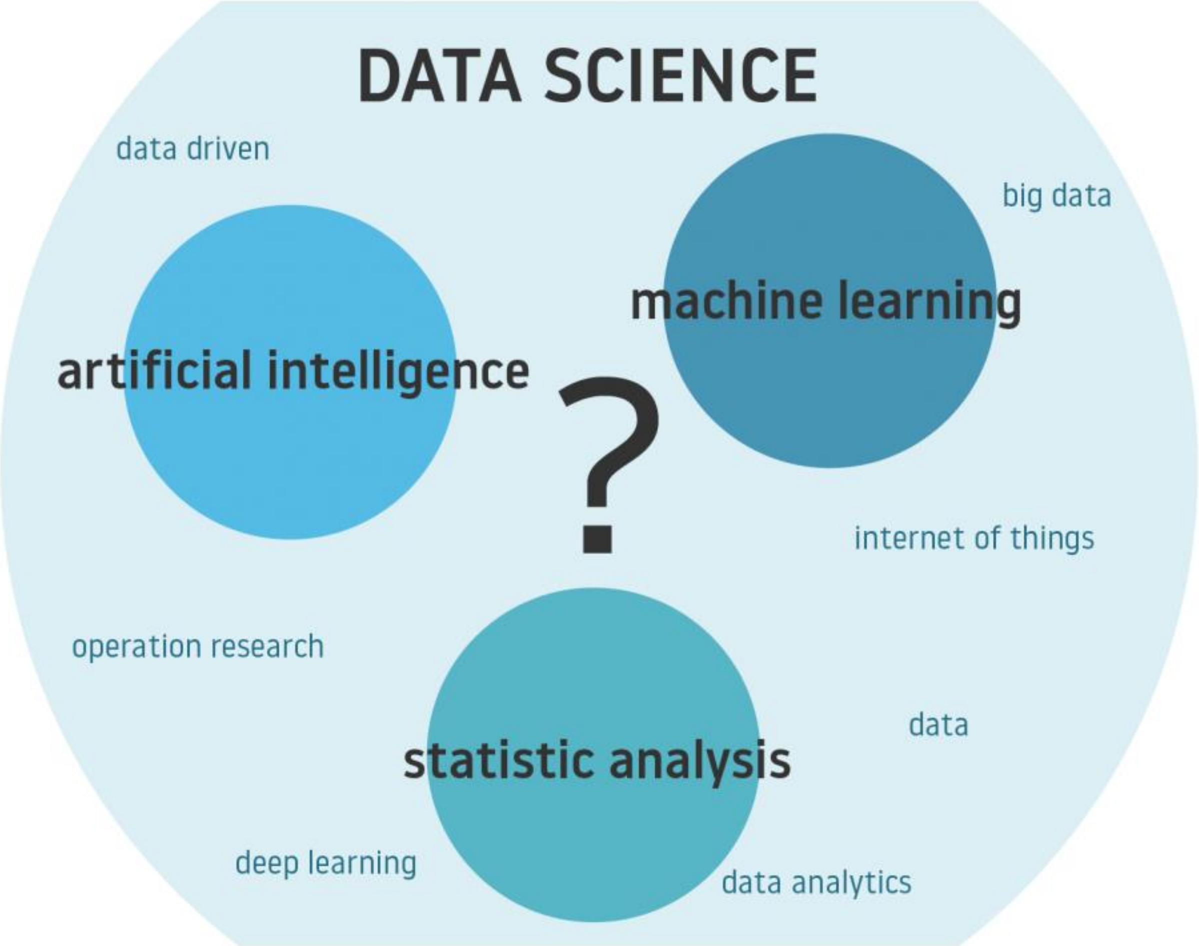
- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science/ Big Science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data Science is a major new challenge!

Characteristics of Data Science in big data ?

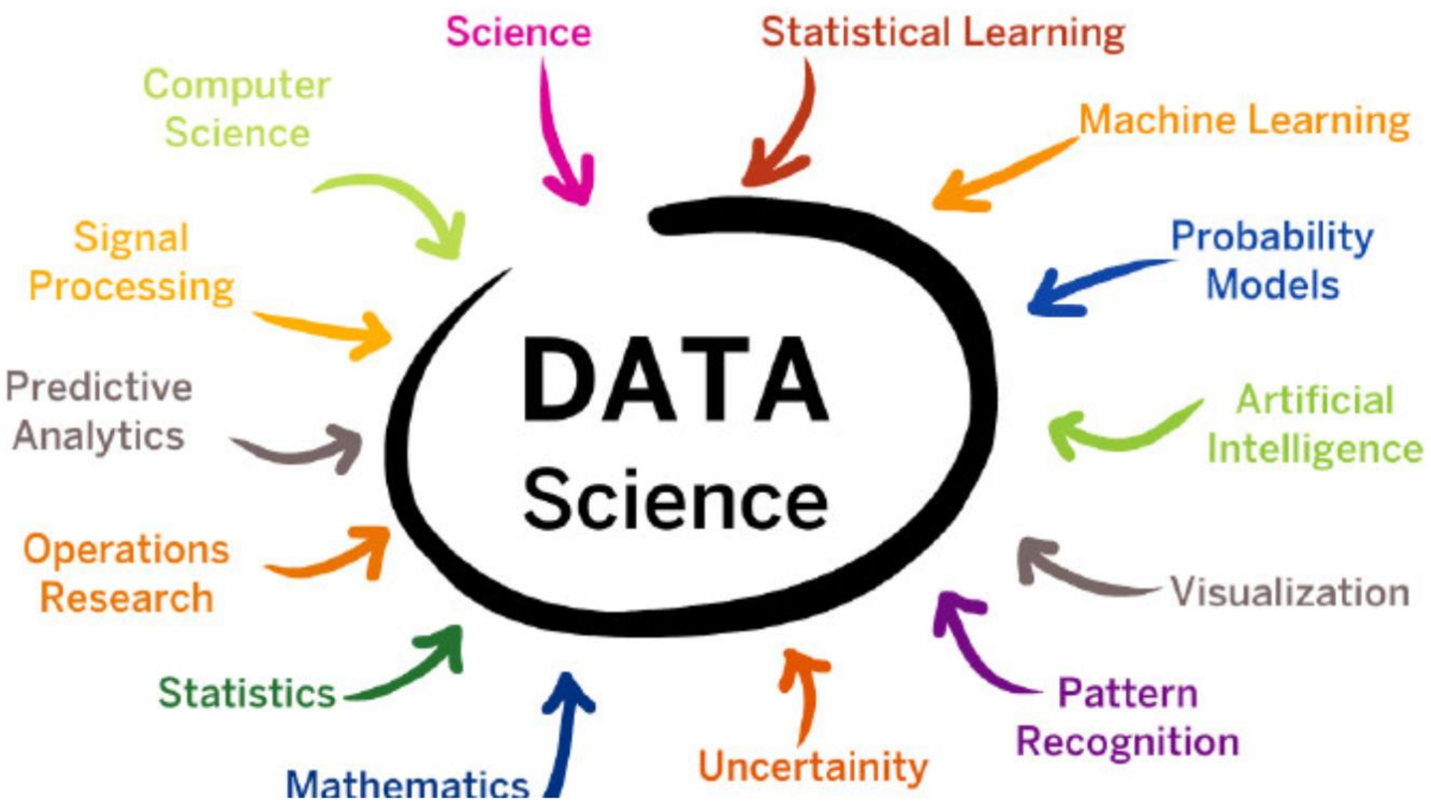


VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

Data Science: Confluence of Multiple Disciplines



Data Science: Confluence of Multiple Disciplines



Types of Variables

- ***Nominal*** **:** Name only--Gender, hair color, ethnicity
- ***Ordinal*** **:** Nominal categories with an implied order--Low, medium, high.
- ***Discrete***: Reflects a number obtained by counting—no decimal.
- ***Continuous***: Reflects a measurement; the number of decimal places depends on the precision of the measuring device.

Types of Variables

There are two basic types of variables:

categorical and ***numerical***.

Categorical Variables: variables defined by the classes or categories into which an individual member falls.

Numerical Variables: variables to which a number is assigned as a quantitative value.

Data and Variables

Data are often discussed in terms of variables, where a **variable** is:

> Any characteristic that ***varies*** from one member of a population to another.

A simple example is height in centimeters, which varies from person to person.

Definition of Variables in a data Matrix

- AGE: Age in years
- BMI: Body mass index, weight/height^2 in kg/m^2
- FFNUM: The average number of times eating “fast food” in a week
- TEMP: High temperature for the day
- GENDER: 1- Female 0- Male
- EXERCISE LEVEL: 1- Low 2- Medium 3- High
- QUESTION: Compared to others, what is your satisfaction rating of the National Practitioner Data Bank?

  1- Very Satisfied 2- Somewhat Satisfied 3- Neutral

  4- Somewhat dissatisfied 5- Dissatisfied

Data table/Data matrix

OBS	AGE	BMI	FFNUM	TEMP(^{0}F)	GENDER	EXERCISE LEVEL	QUESTION
1	26	23.2	0	61.0	0	1	1
2	30	30.2	9	65.5	1	3	2
3	32	28.9	17	59.6	1	3	4
4	37	22.4	1	68.4	1	2	3
5	33	25.5	7	64.5	0	3	5
6	29	22.3	1	70.2	0	2	2
7	32	23.0	0	67.3	0	1	1
8	33	26.3	1	72.8	0	3	1
9	32	22.2	3	71.5	0	1	4
10	33	29.1	5	63.2	1	1	4
11	26	20.8	2	69.1	0	1	3
12	34	20.9	4	73.6	0	2	3
13	31	36.3	1	66.3	0	2	5
14	31	36.4	0	66.9	1	1	5
15	27	28.6	2	70.2	1	2	2
16	36	27.5	2	68.5	1	3	3
17	35	25.6	143	67.8	1	3	4

Reading Web data

Co to the following address

https://raw.githubusercontent.com/amrrs/sample_revenue_dashboard_shiny/master/recommendation.csv

Account,Product,Region,Revenue--------------------→name of attributes
Axis Bank,FBB,North,2000
HSBC,FBB,South,30000
SBI,FBB,East,1000
ICICI,FBB,West,1000
Bandhan Bank,FBB,West,200
Axis Bank,SIMO,North,200
HSBC,SIMO,South,300
SBI,SIMO,East,100
ICICI,SIMO,West,100
Bandhan Bank,SIMO,West,200

Thanks for your attention