UNIVERSITY OF THE PUNJAB

DATA SCIENCE

Engr. Dr. Muhammad Nadeem Majeed

Today's Agenda

- About Myself
- Course Information and Protocols
- Data Data Everywhere
- Categories of Data
- What is Data Science?
- Factors making Data Science Ubiquitous
- Applications of Data Science
- Discussion on Course Matrix

Engr. Dr. Muhammad Nadeem Majeed

Associate Professor
Department of Data Science

- PhD Computer Engineering
- MS Computer Engineering

Research interests:
Mispronunciation Detection,
Disease classification.
Network and Communications.
Machine Learning

Certifications:

1. **Project Management Professional (PMP)®**
2. **PRINCE2 Agile Practitioner**
3. **Professional Scrum Master (PSM)**
4. **Certified Lean Six Sigma Green Belt (CSSC)**
5. **ITIL Certified**
6. **Cisco Certified Network Associate (CCNA)**
7. **Juniper Networks Certified Internet Associate (JNCIA)**
8. **IBM Data Engineer**

Experience: 22 Years of Teaching & Network Management

- University of The Punjab, Lahore
- University of Engineering and Technology Taxila.
- NUST School of Electrical Engineering and Computer Science. Islamabad.
- University of Arid Agriculture Rawalpindi.

Course Information and Protocols

Course Info

- **Textbook(s):** Python for Data Analysis, by, Wes McKinney, 2nd Edition, Published in 2017, ISBN-13: 9781491957660
- **Lectures Slides Available at:** http://arifbutt.me
- **Video Lectures Available at:** https://youtube.com/learnwitharif
- **Codes Hosted at:** https://github.com/arifpucit/data-science
- **Grades Website:** http://online.pucit.edu.pk
- **Prerequisites: Basic Programming skills**
- **Office:** Room #37, Building-A, FCIT (NC)
- **Students Counseling hours:**
  - Mon: 1500 hrs – 1600 hrs
  - Wed: 1500 hrs – 1600 hrs
- **24 hour turnaround for email: nadeem.majeed@pucit.edu.pk**

Who cares to get an A

Final exam: 40

Mid-exam: 35

Sessionals: 25

- Quizzes: 30%
- Programming Assignments : 30%
- Research Papers : 40%

MPhil.

Minimum GP to pass a course: GP >=2.3 [C+ or 61 mks]

Degree Completion Requirement: CGPA >=2.5

Probation: 2.3 <= CGPA < 2.5 [Only one probation allowed]

Dropped out: CGPA < 2.3

Ph.D.

Minimum GP to pass a course: GP >=2.7 [B- or 65 mks]

Degree Completion Requirement: CGPA >=2.8

Probation: 2.8 <= CGPA < 3.0

Dropped out: CGPA < 2.8

Cheating Policy

- Academic integrity
- Both the cheater and the student who aided the cheater will be held responsible for the cheating
- The instructor may take actions such as:
  - require repetition of the subject work,
  - assign 'zero' or may be 'negative' marks for the subject work,
  - for serious offenses, assign an F grade for the course

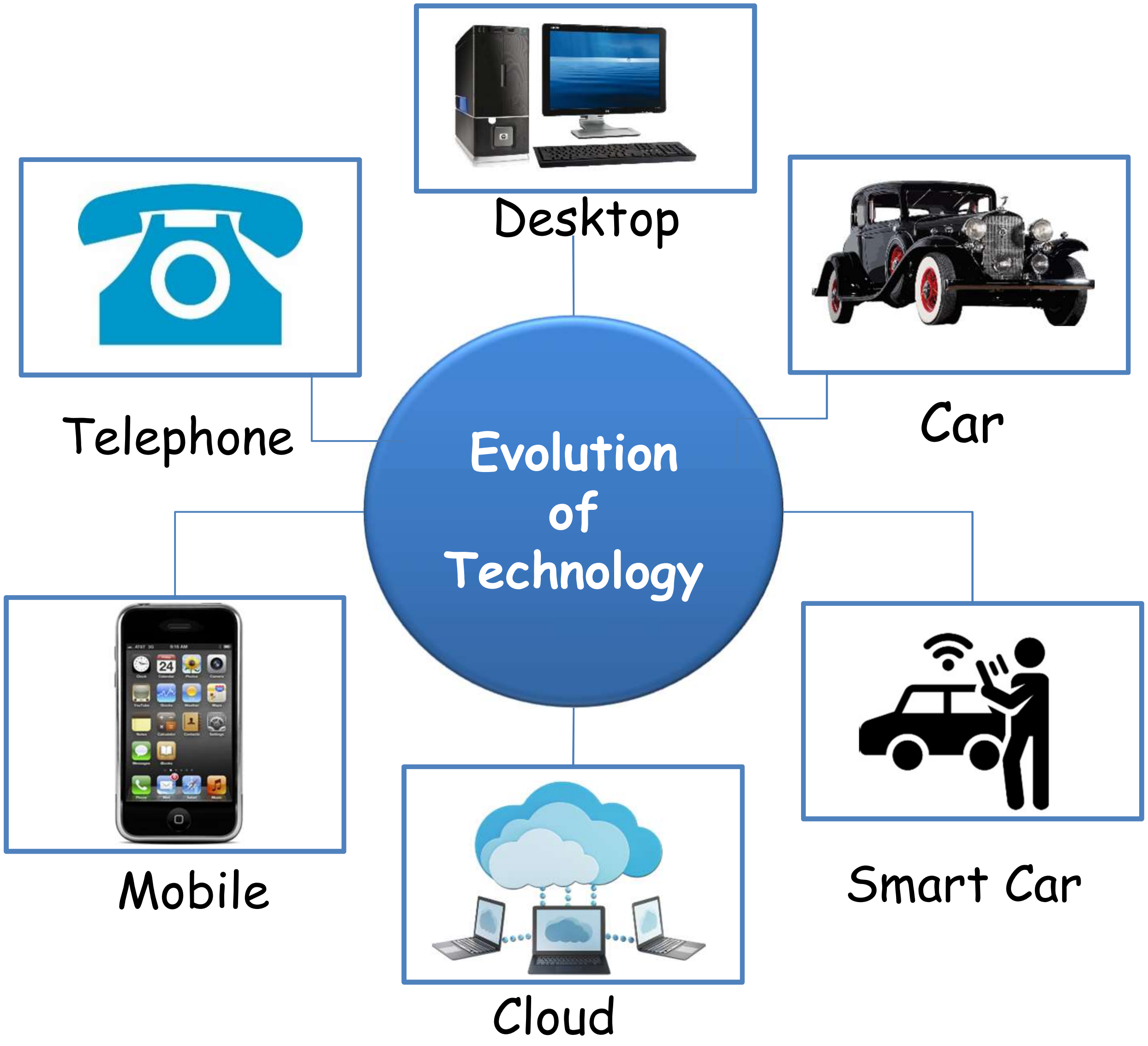
Late Policy for Home Works and PA

- Late policy for Assignment, Quizzes, and other deliverables
  - No late Assignment submissions!
  - No late quizzes or exams!
- Sticking to dates is your responsibility!
  - Check announcements on lecture notes regularly
- Your best strategy is to play it safe – submit everything on time

Lecture Format

Data Data Everywhere
Data Sources

Data Sources: Evolution of Technology



Evolution of Technology

Desktop

Car

Smart Car

Cloud

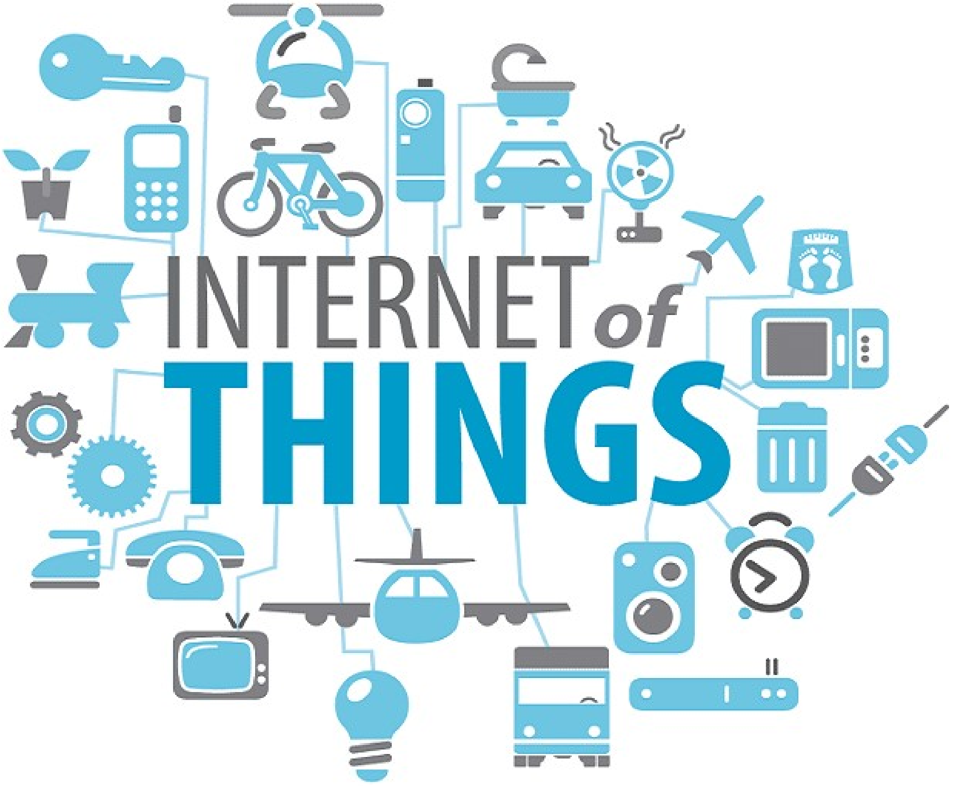
Mobile

Telephone

Data Sources: IoT

Collection of interconnected devices that communicate and transfer data through the Internet

IoT



As per CISCO recent survey, IoT is generating more than 500 ZB of data per year

Data Sources: Social Media

347,222 tweets

1,736,111 pictures

4,166,667 likes & 200,000 photos

204,000,000 emails

YouTube

300 hrs of video uploaded
2.78 million video views

342,000 apps downloaded

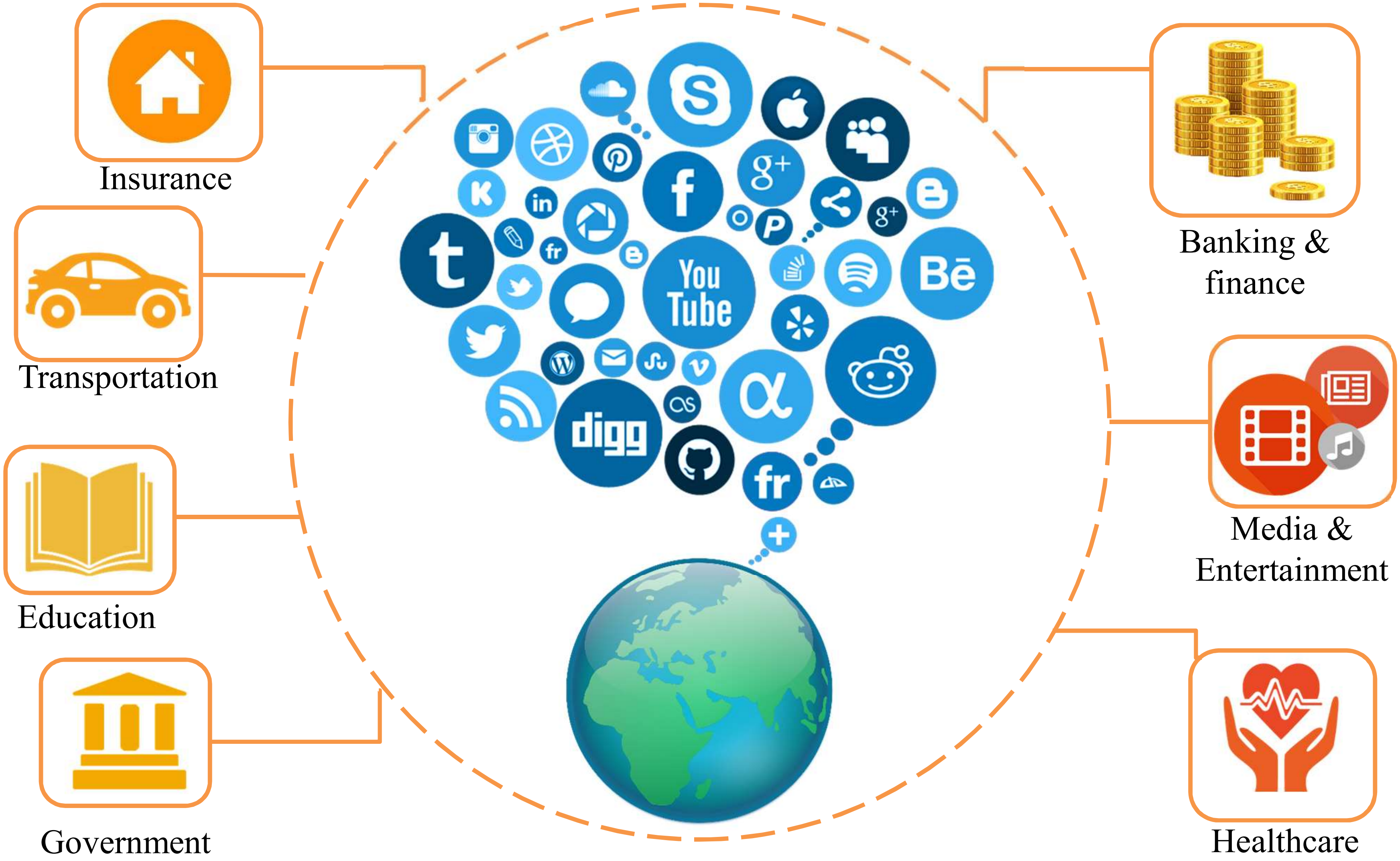
NETFLIX

70,017 hours watched

2.4 million search queries

Imagine processing & analyzing this much data, and then trying to figure out important insights from it

Data Sources: Other Factors



Insurance

Transportation

Education

Government

Banking & finance

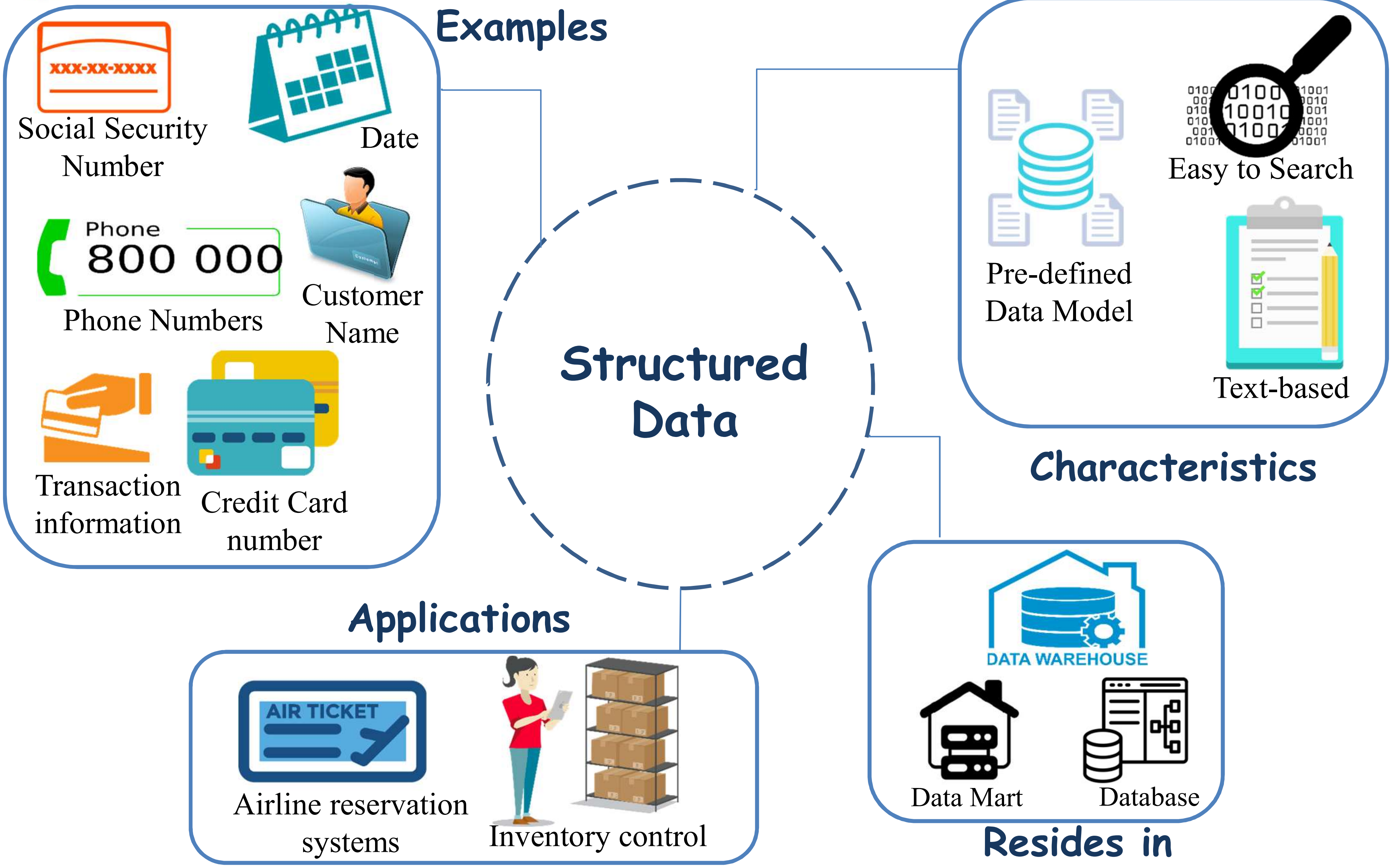
Media & Entertainment

Healthcare

Data Science is all about extracting the useful insights from data and using it to grow your business

Categories of Data

Structured Data



Structured Data

Examples

Social Security Number

Date

Phone Numbers

Customer Name

Transaction information

Credit Card number

Characteristics

Pre-defined Data Model

Easy to Search

Text-based

Applications

Airline reservation systems

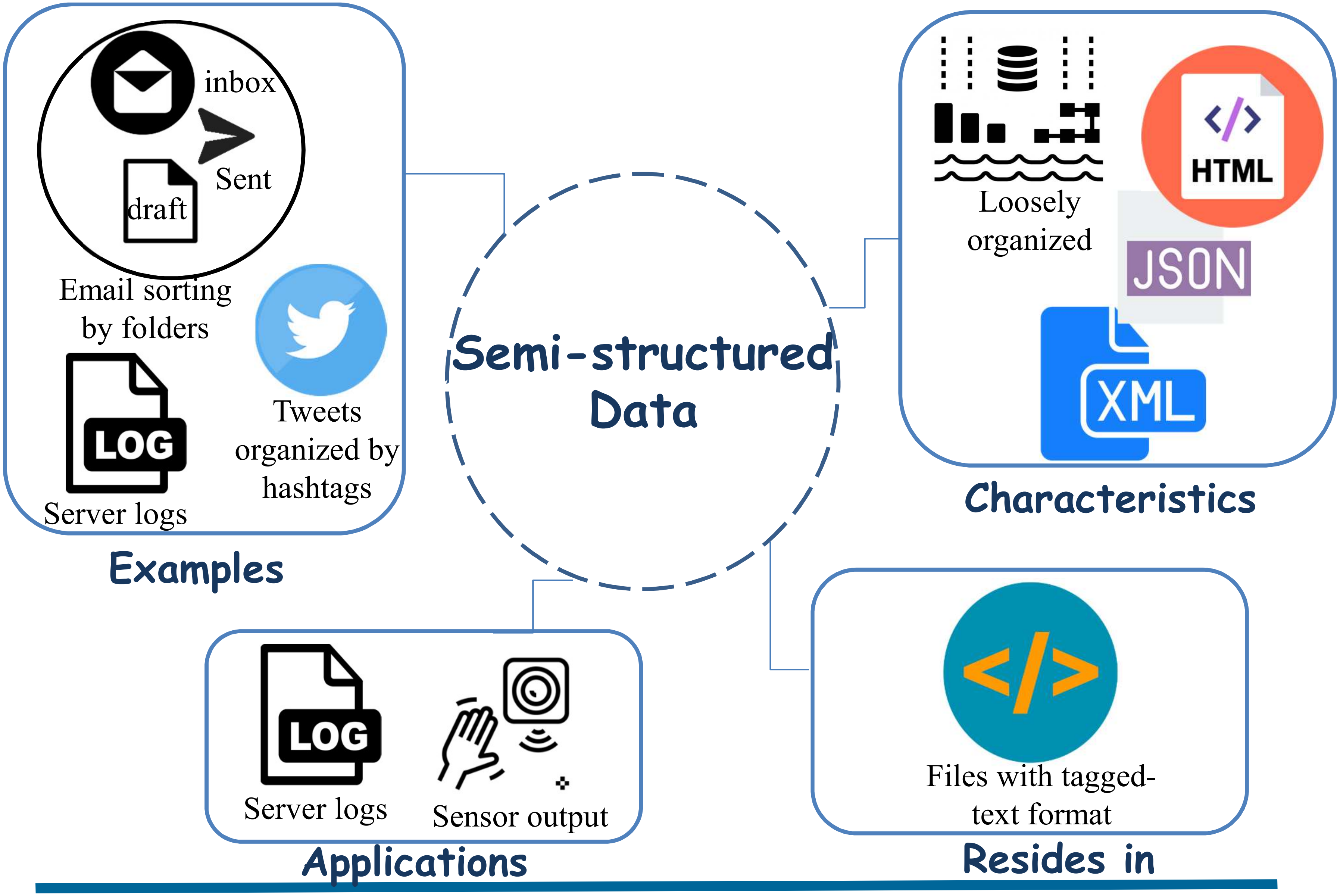
Inventory control

Resides in

Data Mart

Database

Semi-structured Data



Semi-structured Data

Examples

Email sorting by folders

Tweets organized by hashtags

Server logs

Characteristics

Loosely organized

Applications

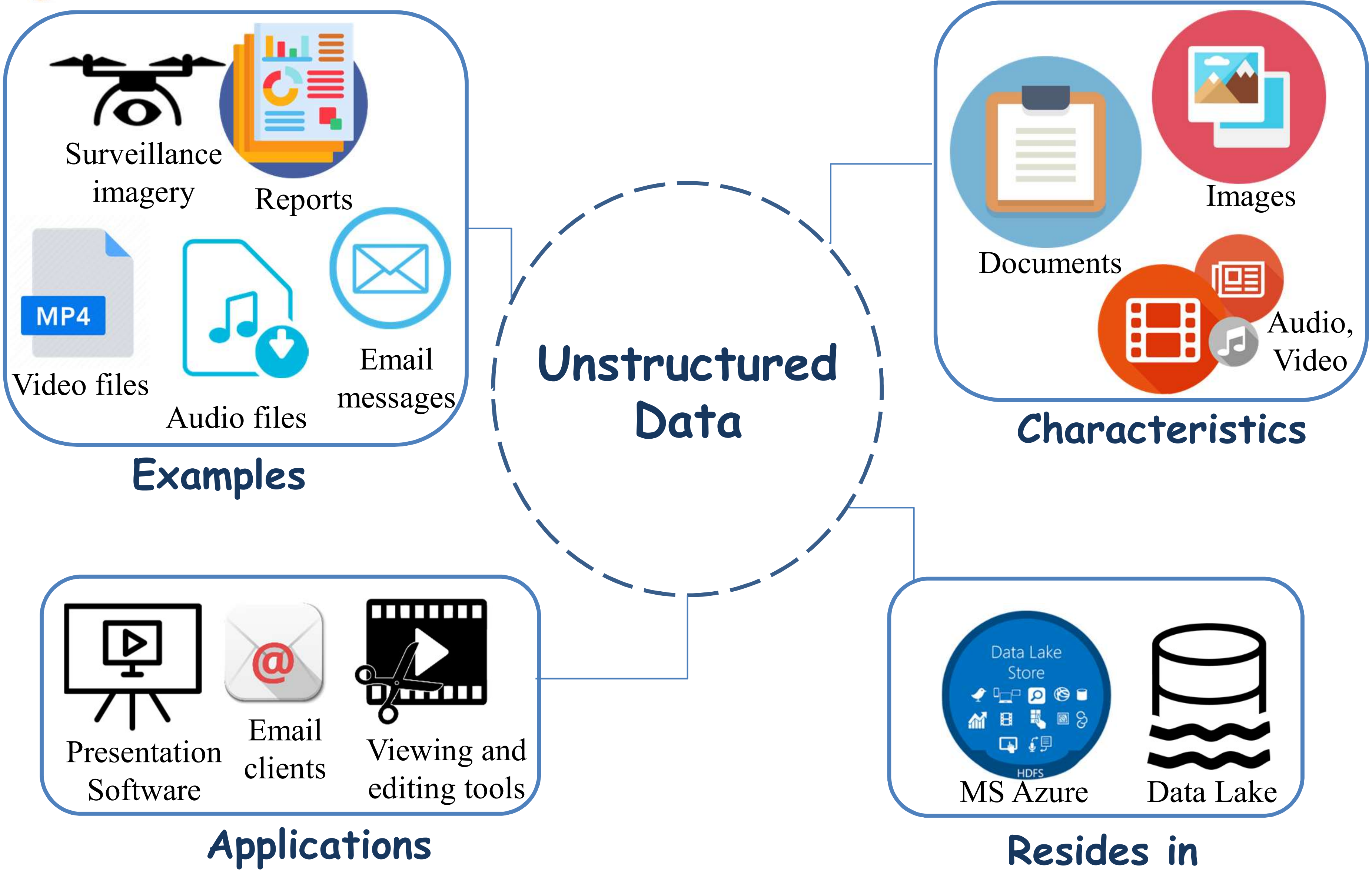
Server logs

Sensor output

Resides in

Files with tagged-text format

Unstructured Data



Unstructured Data

Examples

- Surveillance imagery
- Reports
- Video files
- Audio files
- Email messages

Characteristics

- Documents
- Images
- Audio, Video

Applications

- Presentation Software
- Email clients
- Viewing and editing tools

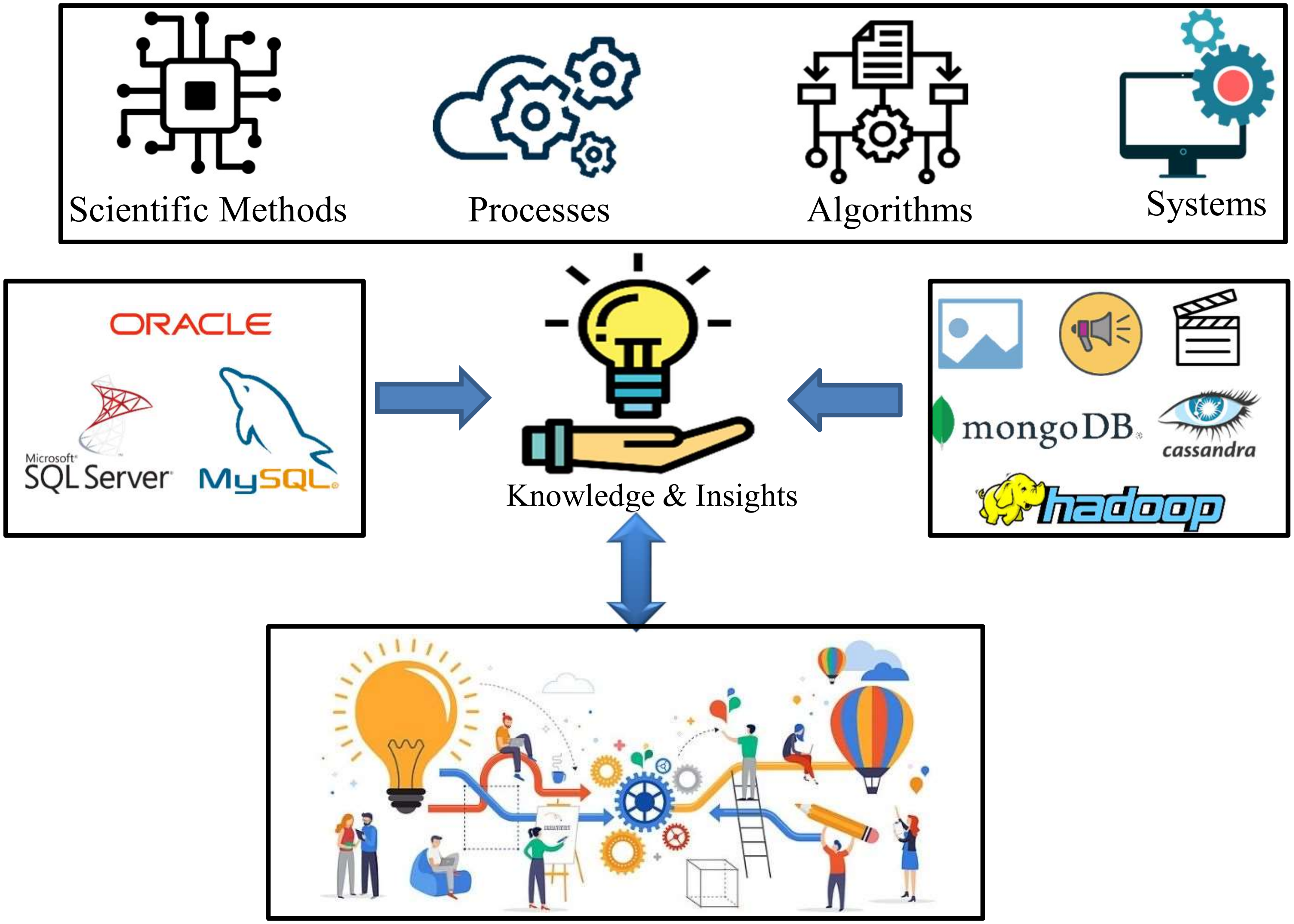
Resides in

- MS Azure
- Data Lake

What is Data Science?

What is Data Science?

Data Science is an Inter-Disciplinary Field that uses



Scientific Methods

Processes

Algorithms

Systems

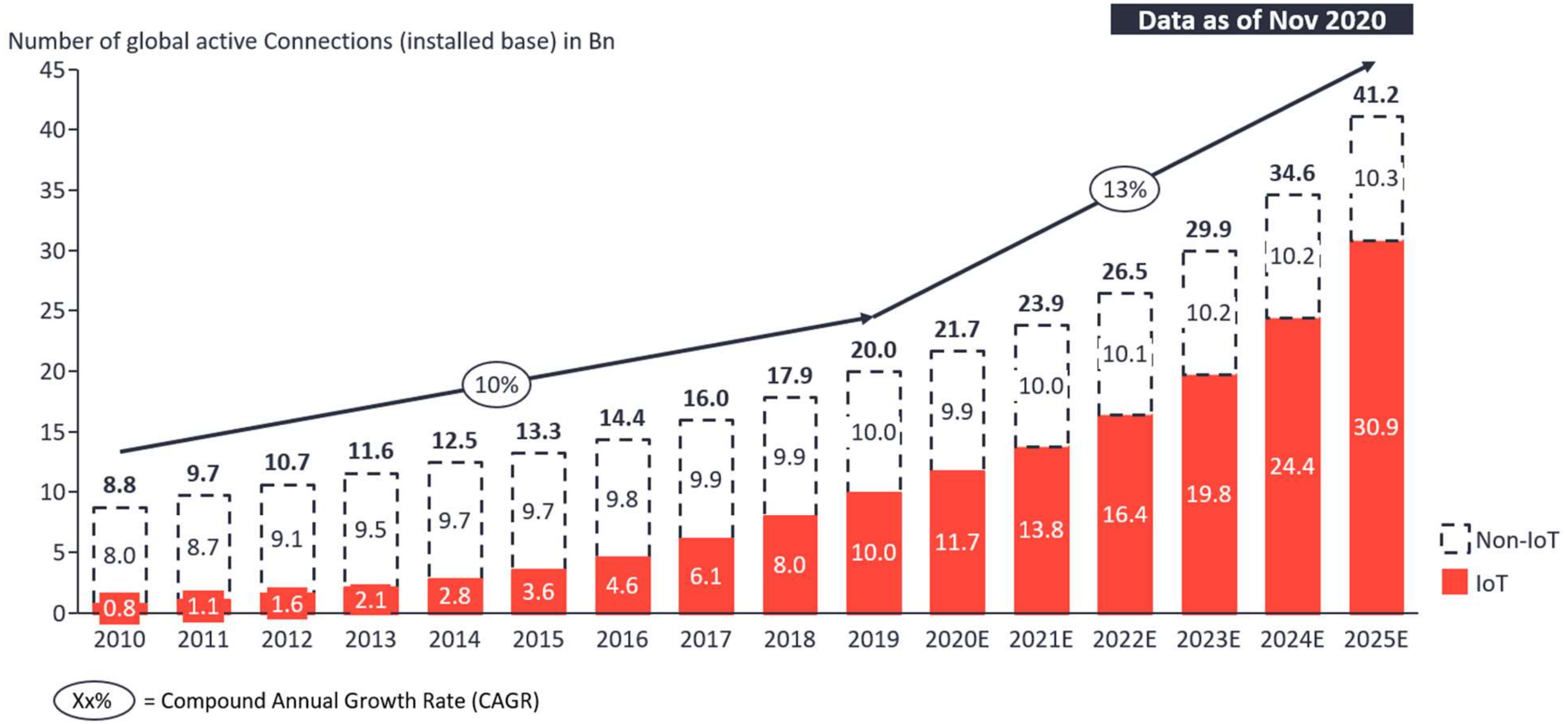
Knowledge & Insights

Factors Making Data Science Ubiquitous

Increasing Number of Connected Devices

Total number of device connections (incl. Non-IoT)

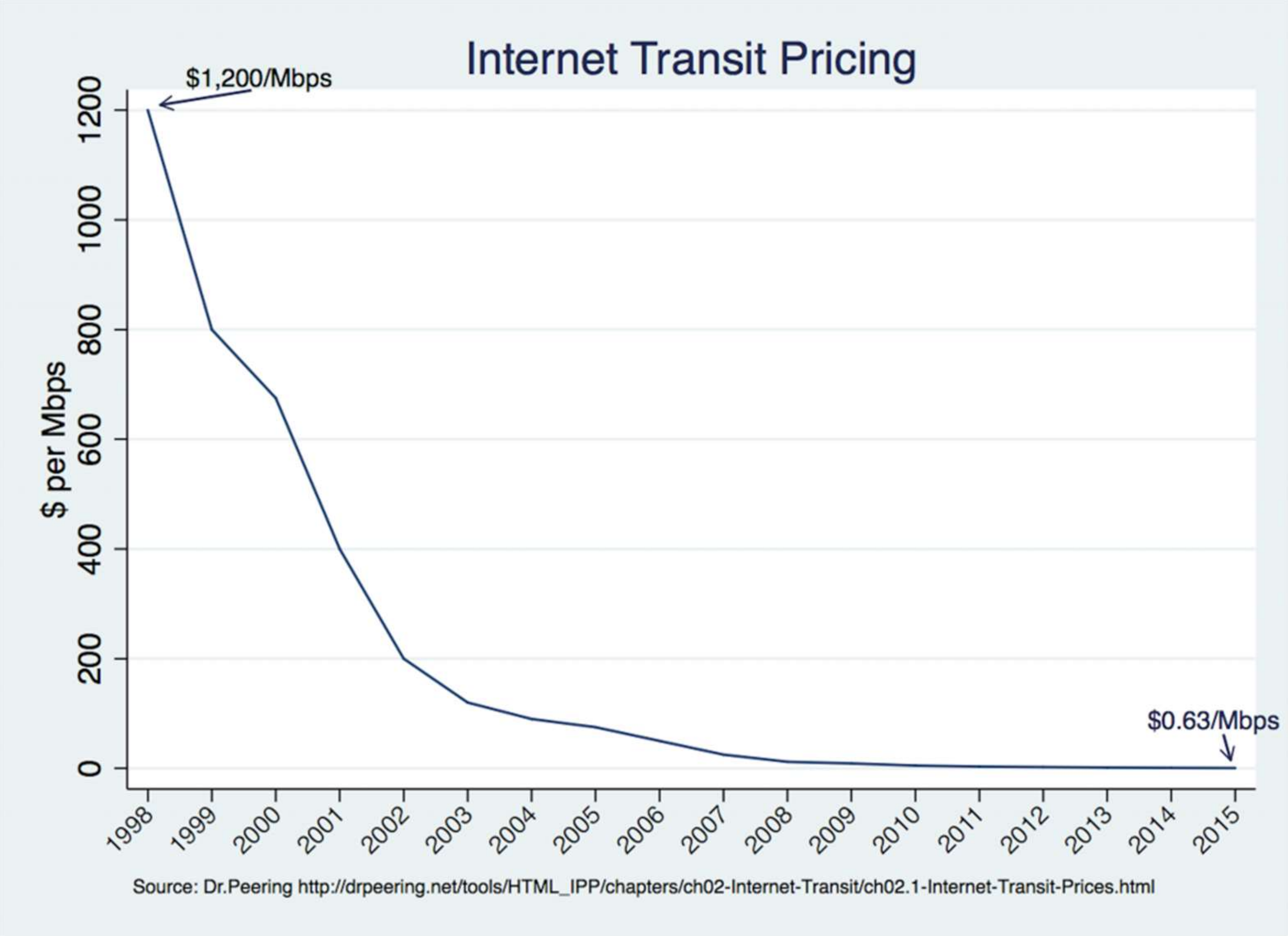
20.0Bn in 2019– expected to grow 13% to 41.2Bn in 2025


Data as of Nov 2020

Note: Non-IoT includes all mobile phones, tablets, PCs, laptops, and fixed line phones. IoT includes all consumer and B2B devices connected – see IoT break-down for further details

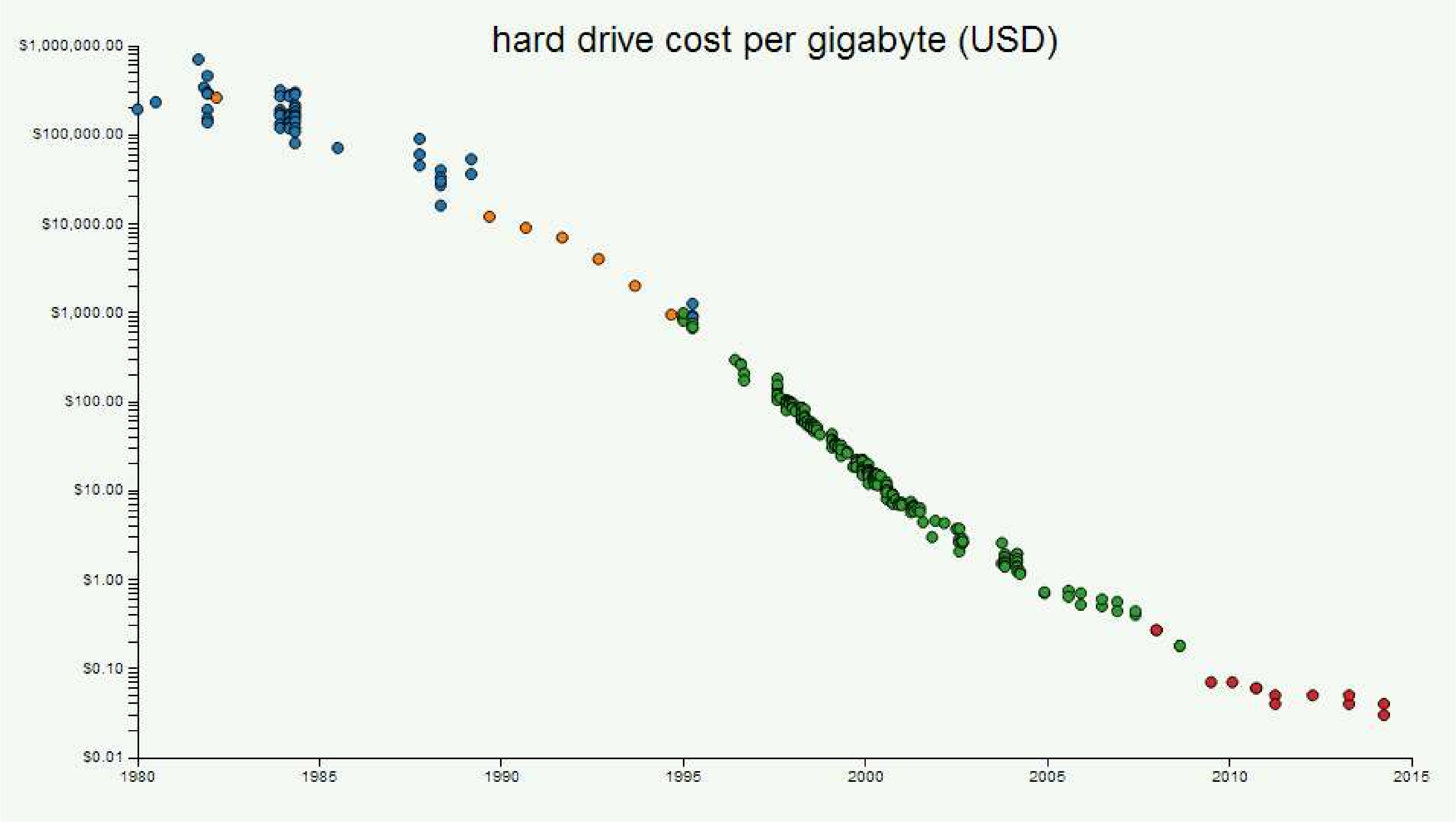
Source(s): IoT Analytics - Cellular IoT & LPWA Connectivity Market Tracker 2010-25

Decreasing Internet Transit Pricing



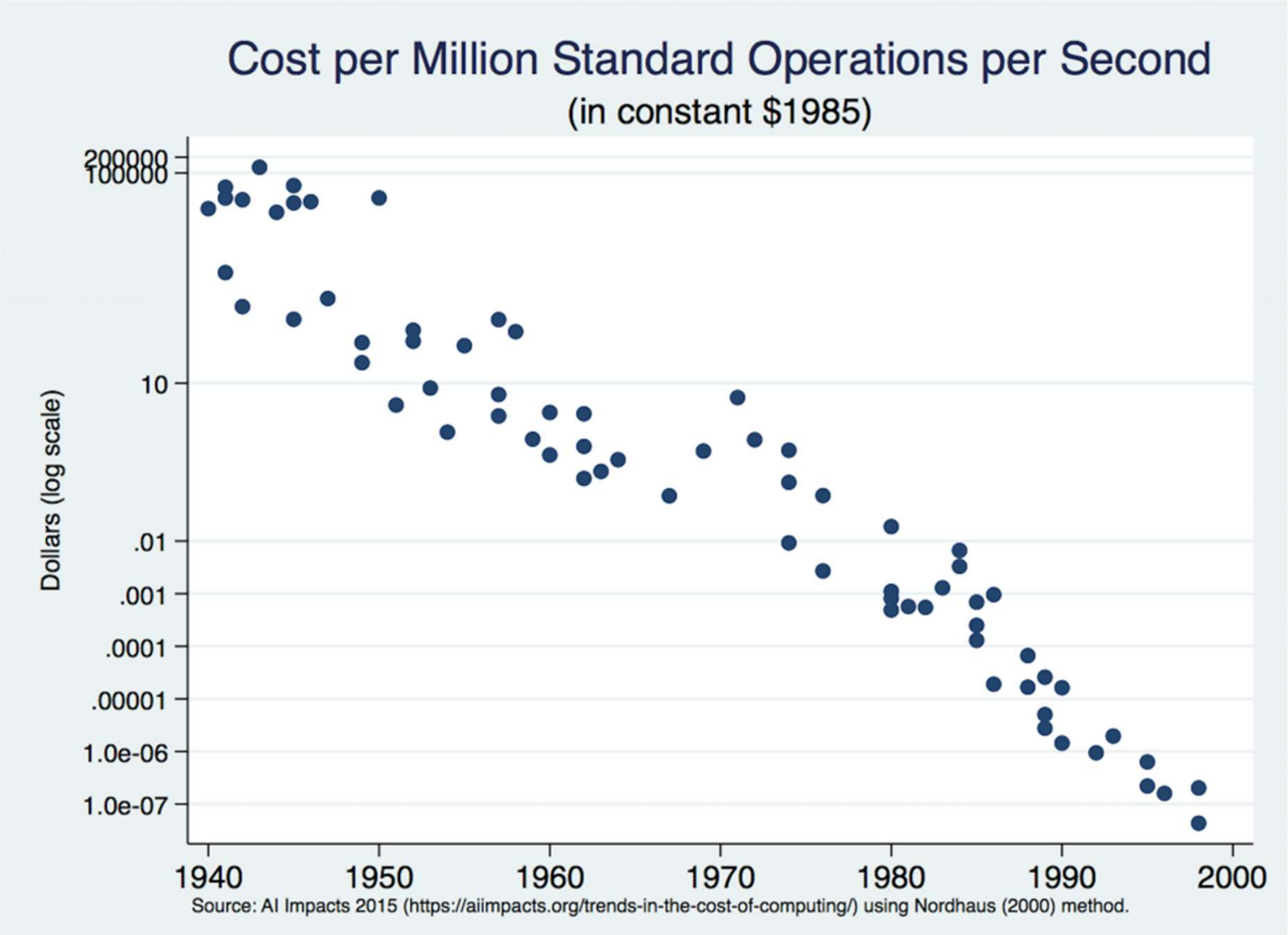
Source: Dr.Peering http://drpeering.net/tools/HTML_IPP/chapters/ch02-Internet-Transit/ch02.1-Internet-Transit-Prices.html

Decreasing Costs for Data Storage



Source: https://community.spiceworks.com

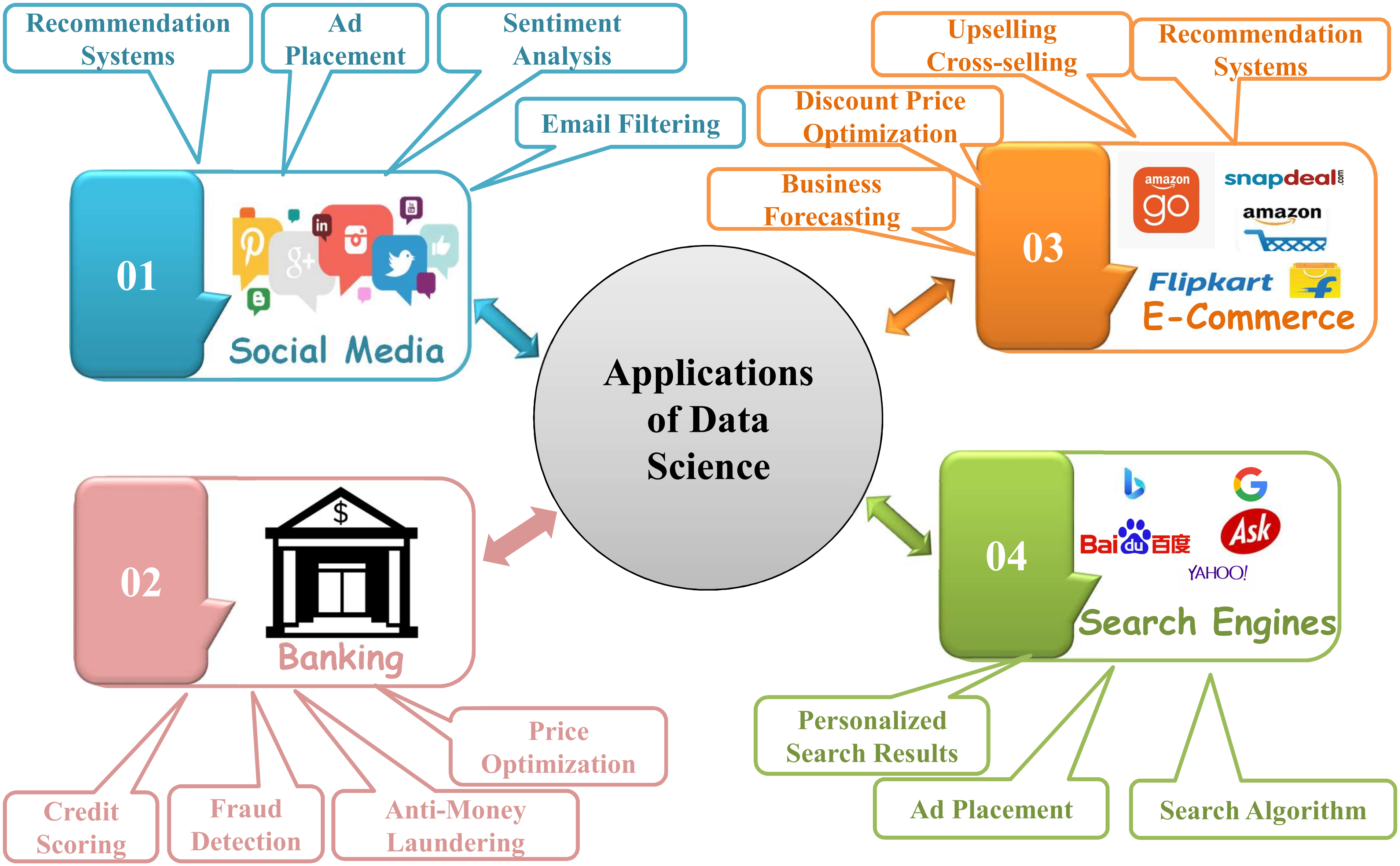
Decreasing Computational Costs



Source: AI Impacts 2015 (https://aiimpacts.org/trends-in-the-cost-of-computing/) using Nordhaus (2000) method.

Applications of Data Science?

Applications of Data Science



Applications of Data Science

01 Social Media

Recommendation Systems

Ad Placement

Sentiment Analysis

Email Filtering

02 Banking

Credit Scoring

Fraud Detection

Anti-Money Laundering

Price Optimization

03 E-Commerce

Upselling Cross-selling

Recommendation Systems

Discount Price Optimization

Business Forecasting

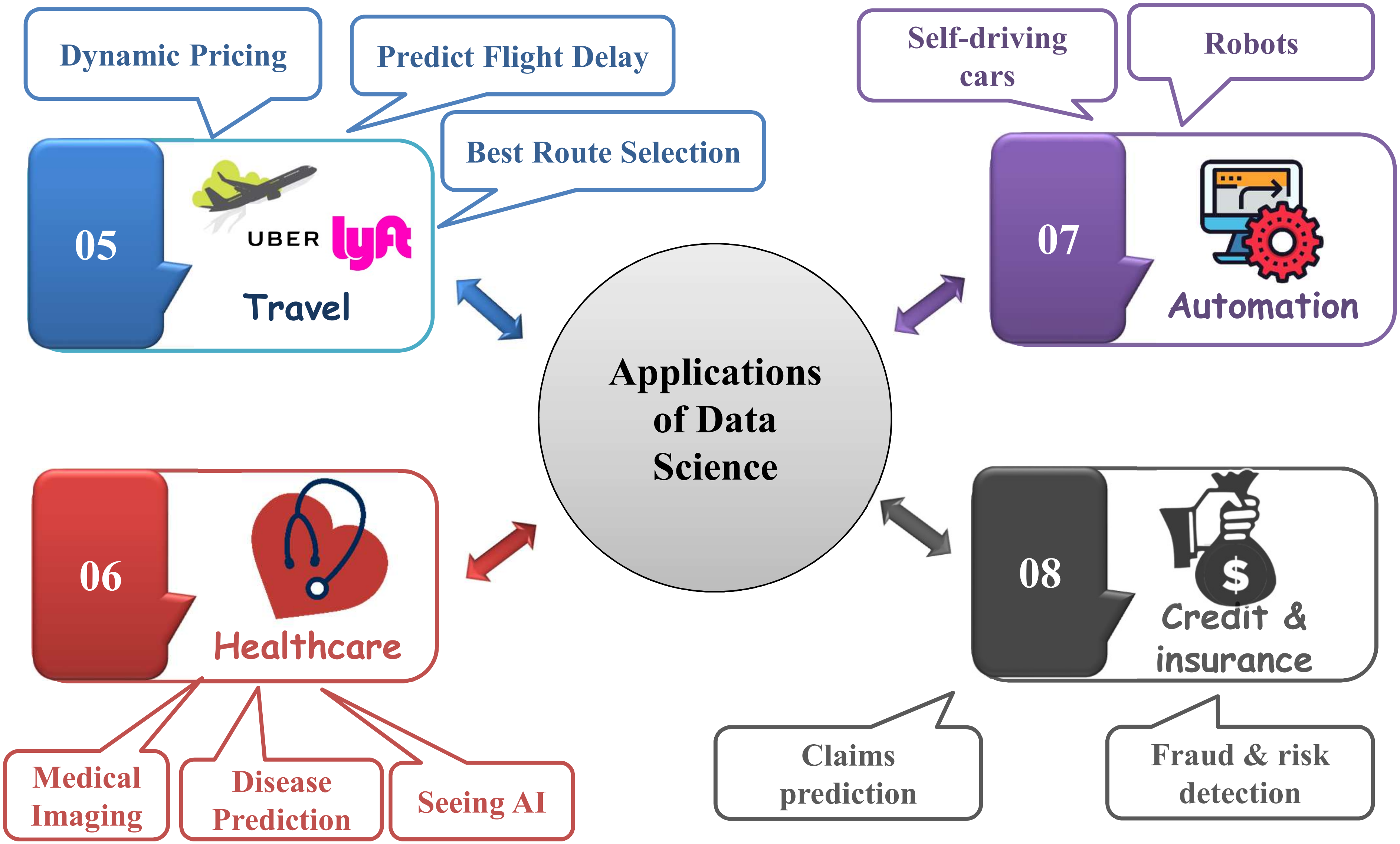
04 Search Engines

Personalized Search Results

Ad Placement

Search Algorithm

Applications of Data Science (cont…)



Dynamic Pricing

Predict Flight Delay

Best Route Selection

05

UBER lyft

Travel

Self-driving cars

Robots

07

Automation

Applications of Data Science

06

Healthcare

Medical Imaging

Disease Prediction

Seeing AI

08

$

Credit & insurance

Claims prediction

Fraud & risk detection

Discussion on Course Matrix

What we will do in this course?

Module 1: (Overview of the course)

- What is Data Science?
- Why/How to do Data Science?
- Structured vs Unstructured data
- Applications of Data Science
- Tools and Technologies for Data Science
- Life Cycle of a Data Science Project
- Job Roles in the Industry
- Data Science Use Cases from real life
- Git and Github for Data Scientists

Reading Tasks:

- …

What we will do in this course?

Module 2: (Basics of Python Programming)

- Overview of Python programming language
- Python programming environments
- Python intrinsic data types and operators
- Python data structures
- Selection and Repetition structures
- Functions in Python
- Exception handling
- Modules, packages and libraries
- Basic file handling in Python

Reading Tasks:

- …

What we will do in this course?

Module 3: (Python for Data Scientists)

- Overview of Python libraries for Data Science
- Reading data in Python (csv, xlsx, json)
- Data manipulation with NumPy
- Scientific computation with SciPy
- Data manipulation with Pandas
- Visualization with Matplotlib and Seaborn

Reading Tasks:

- …

What we will do in this course?

Module 4: (Mathematics for Data Scientists)

- Applied Linear Algebra for Data Scientists
- Applied Calculus for Data Scientists
- Applied Descriptive Statistics for Data Scientists
- Applied Inferential Statistics for Data Scientists

Reading Tasks:

- …

What we will do in this course?

Module 5: (Data Acquisition)

- Overview of Data Acquisition
- Data Acquisition from SQL Databases
- Data Acquisition from NoSQL Databases
- Data Acquisition from Websites

Reading Tasks:

- …

What we will do in this course?

Module 6: (Data Wrangling and EDA)

- Acquire data sets from different sources
- Understand the datasets
- Transform to appropriate format
- Perform data cleaning
- Perform data wrangling
- Carry out Exploratory Data Analysis
- Identify pattern/trends using different visualization and statistical tools
- Preparing the dataset ready to be used by Machine Learning Engineer

Reading Tasks:

- …

What we will do in this course?

Module 7: (Machine Learning)

- Overview of Machine Learning
- Categories of Machine Learning Types and Algorithms
- Python for Machine Learning (Scikit-learn)
- Will do hands on practice for
  - ✓ Model creation
  - ✓ Model training
  - ✓ Model evaluation
  - ✓ Feature engineering
  - ✓ Dimensionality reduction

Reading Tasks:

- …

What we will do in this course?

Module 8: (Deep Learning: A Bird's-eye View)

- Machine Learning vs Deep Learning
- Overview of Deep Learning Models (CNN vs RNN)
- Deep Learning Applications
  - ✓ Natural Language Processing
  - ✓ Image recognition
  - ✓ Self-driving cars
  - ✓ Language translation services
- A Hello World on Deep Learning Project using
  - ✓ TensorFlow/Keras/Theano/Torch/Caffe

Reading Tasks:

- …

What we will do in this course?

Module 9: (Big Data: A Bird's-eye View)

- What is Big Data?
- Big Data Storage and Processing Frameworks
  - ✓ Apache Hadoop with MapReduce (used by Alibaba, AOL)
  - ✓ Apache Storm (used by Twitter, Spotify)
  - ✓ Apache Spark (used by Netflix, Yahoo, eBay)
  - ✓ Apache Hive (used by Facebook, Walmart)
- An Overview of Hadoop Ecosystem
  - ✓ Data Storage (HDFS, HBASE)
  - ✓ Data Processing (YARN, Map Reduce)
  - ✓ Data Access (Hive, Pig, Mahout, Avro, Sqoop)
  - ✓ Data Management (Oozie, Chukwa, Flume, ZooKeeper)

Reading Tasks:

- …

Things To Do

- Memorize and follow class protocols
- Should have a very clear understanding of different data sources, its types and storage
- Must know the applications of data science in different domains
- Visit the resource web-sites and download all uploaded resources
- While going through todays lecture slides click all the tools and technologies, which have been hyperlinked to respective web sites.
- You should be able to give a single line description of each tool and technology mentioned in lecture slides

Coming to office hours does NOT mean you are academically weak!