Foundations of AI & Data Science

Unit 1

LEARNING OBJECTIVES

- To excite students about the potential that resides in data and the value that data analytics can add to business processes
- To impart skills related to data cleaning/wrangling, data transformation/preprocessing, and data comprehension through statistical analysis
- To impart skills related to analytical (mathematical) data modeling

Course overview

- Introduction to Data Science & AI
- Data Preparation and Cleaning
- Classification Techniques
- Supervised & Unsupervised Learning Methods
- Data and Customer Segmentation
- Time Series Forecasting

Software and data repositories

- Python
- Data on Kaggle Website
  - http://www.kaggle.com/

Books

- Data Science for Business by Provost and Fawcett (2013)
- Data Mining Concepts and Techniques by Han and Kamber (2011)

MOOCs on Data Sciences

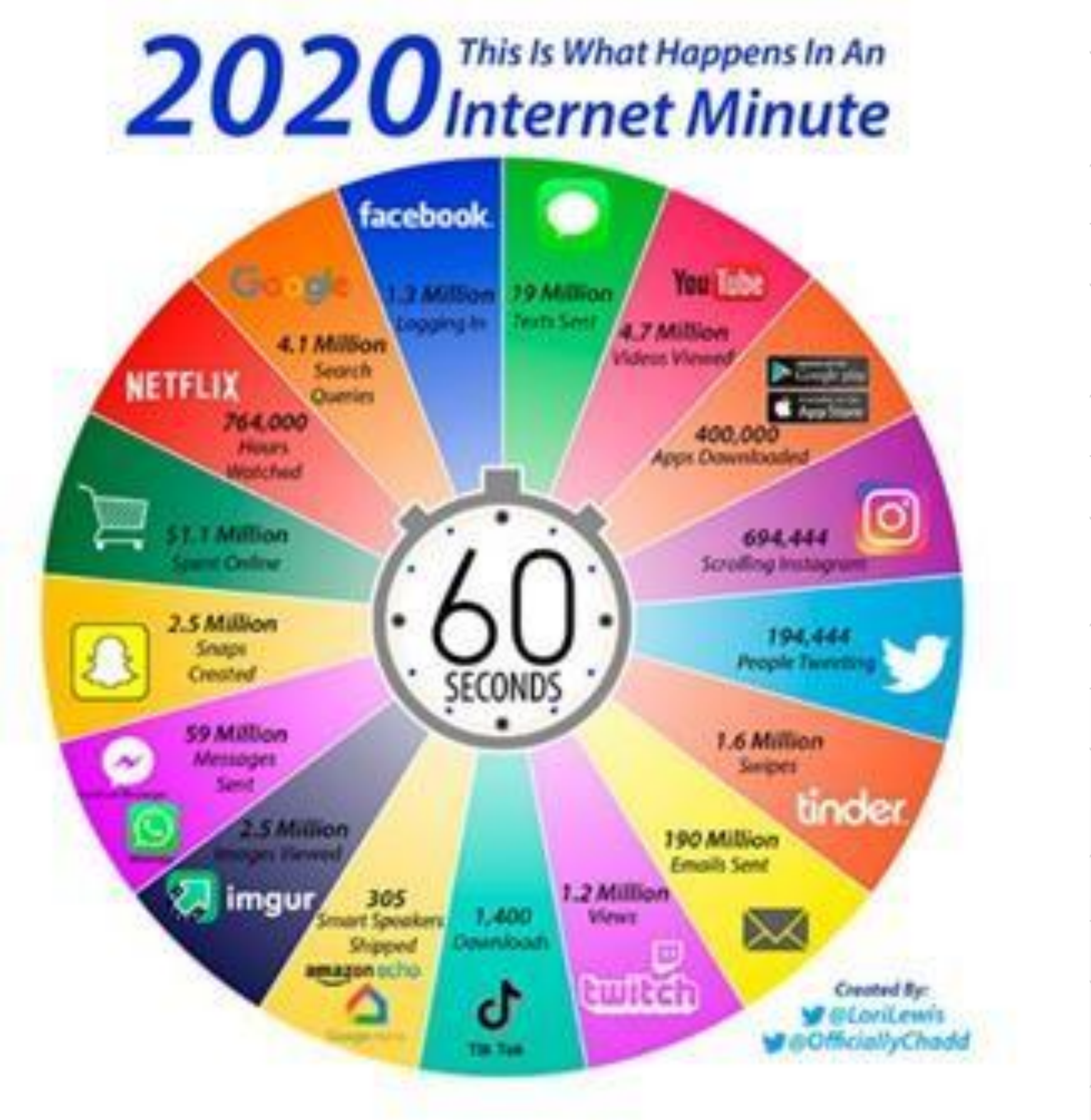
- In the past couple of years, Data Science related courses and specialization have been extremely popular on MOOCs websites:
  - John Hopkins University (Coursera)
  - University of Washington (Coursera)
  - Google (UdaCity)
  - UC Berkley (EdX)
  - University of Toronto (Coursera)
  - And many others……..

Artificial Intelligence

Origin and
Taxonomy

Data is Everywhere!

- Lots of data is being collected and warehoused
  - Web data
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Competitive Pressure is Strong
  - Provide better, customized services



What Is Big Data?

- There is not a consensus as to how to define big data

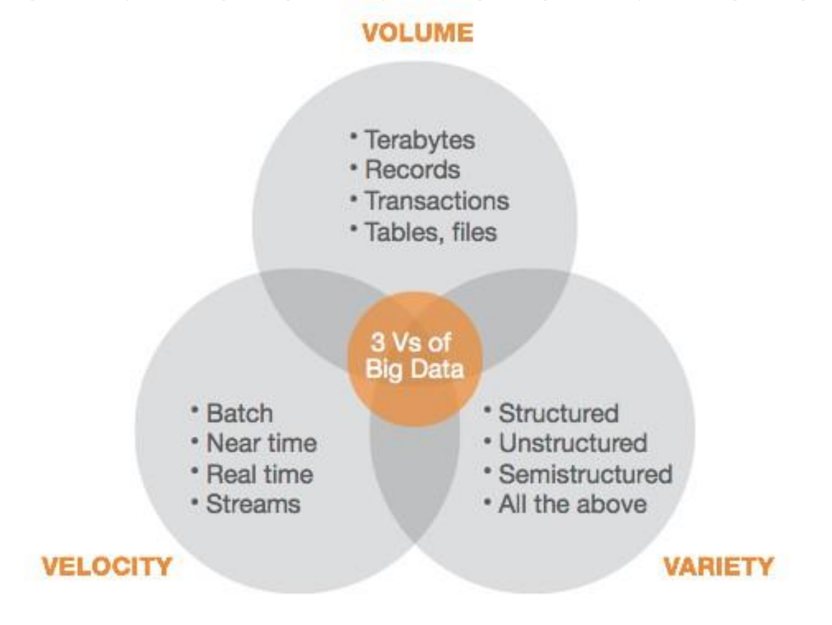
> “Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.”

- The McKinsey Global Institute, 2011

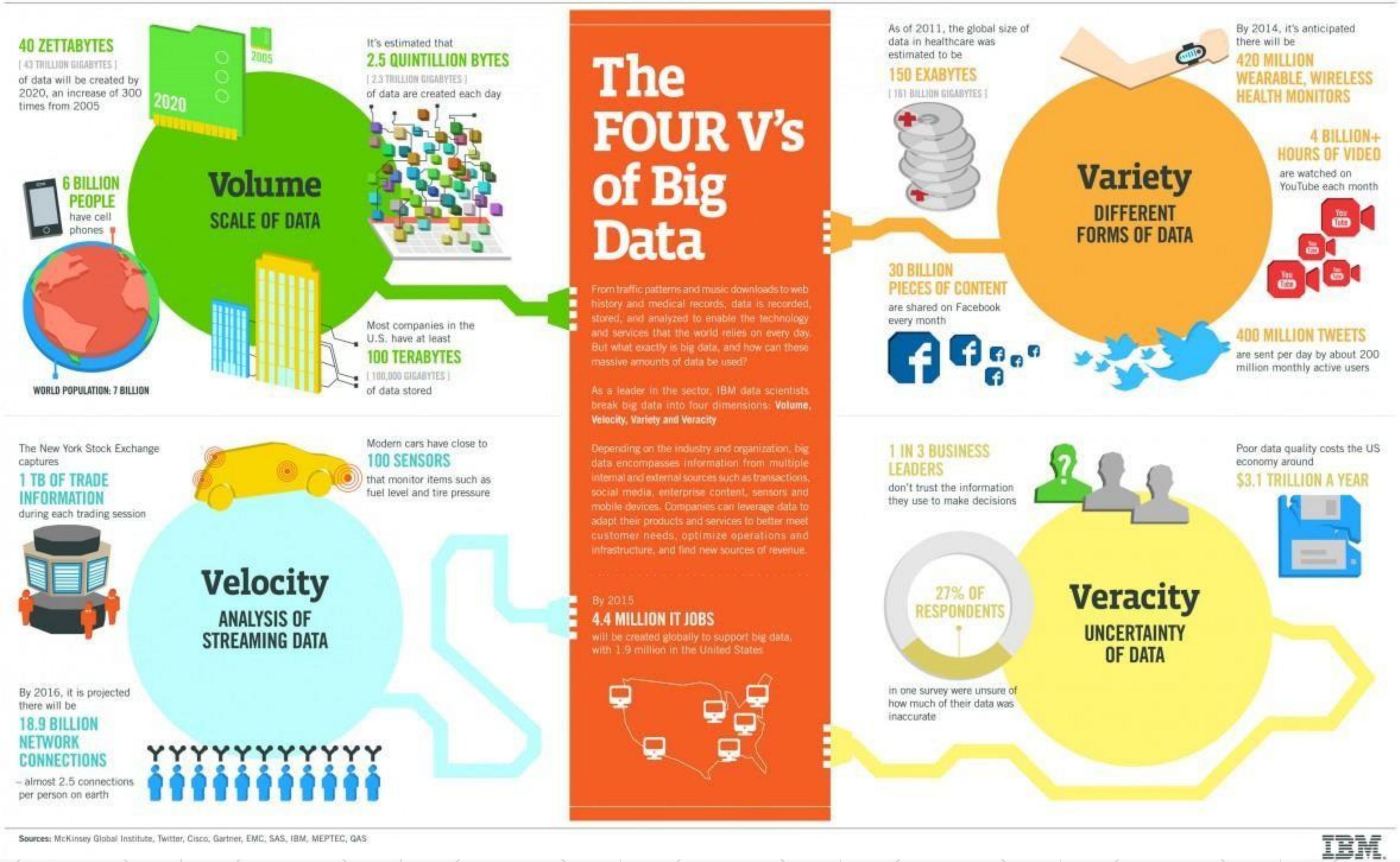
- One reasonable definition is that it’s data which can’t comfortably be processed on a single machine.

3 V’s of Big Data

- Doug Laney was the first one in talking about 3 V's in Big Data management:
  - **Volume**: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process
  - **Variety**: there are many different types of data, as text, sensor data, audio, video, graph, and more
  - **Velocity**: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time



4 V’s by IBM (2014)



The FOUR V’s of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It’s estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it’s anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don’t trust the information they use to make decisions

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

What to do with Data?

- Disease Prediction
- Drug Discovery
- Virtual Assistant
- Image Analysis

Health Care

- Fraud Detection
- Credit Risk Modeling

Banking

- Produc Recommendation
- Fake review detection

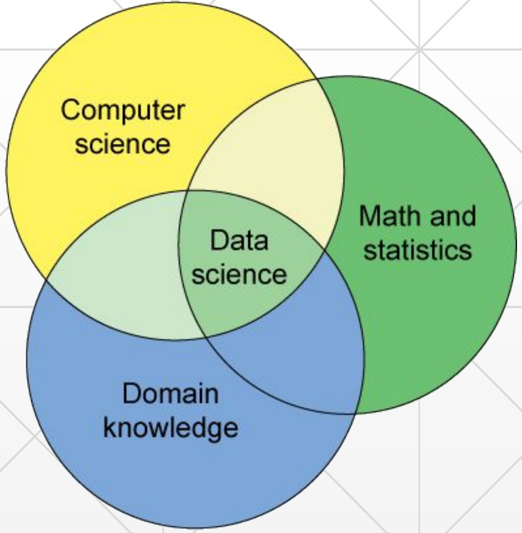
Ecommerce

- Self Driving Cars
- Traffic control

Transport

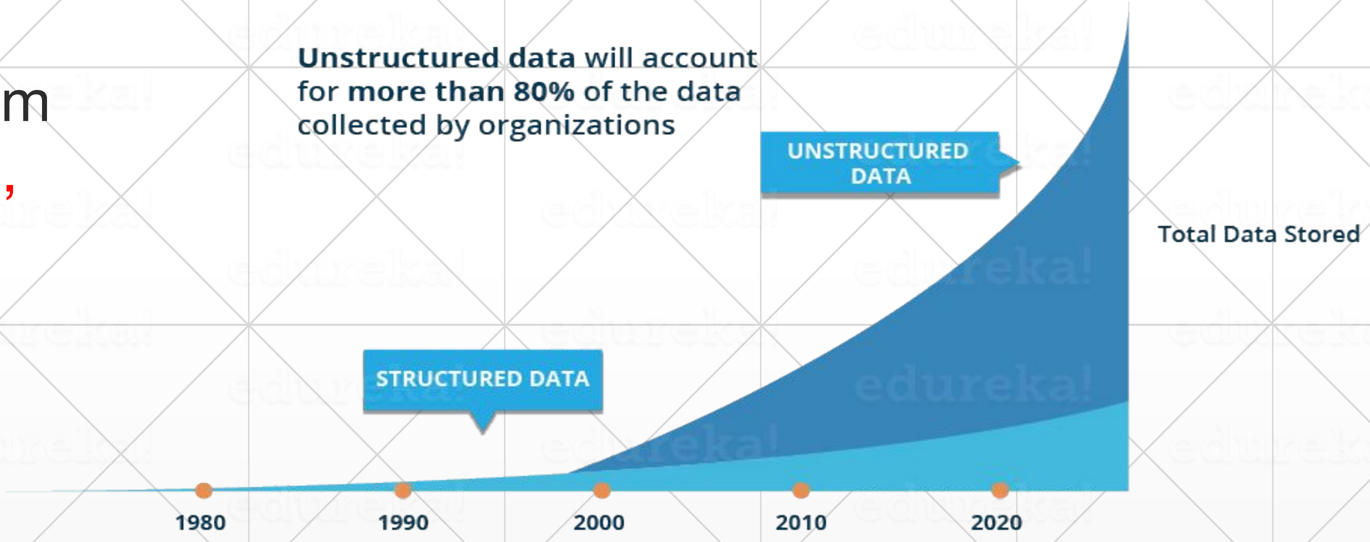
Data Science

- Data science incorporates varying elements and builds on techniques and theories from many fields, including ***mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing*** with the goal of extracting meaning from data and creating data products. *(From Wikipedia)*
- A practitioner of data science is called a data scientist.



Why Data Science?

- Traditional Data – Analyzed using simple BI tools
- Unstructured data is generated from different sources like financial logs, text files, multimedia forms, sensors, and instruments.
- Huge volume and variety of data requires - Complex and advanced analytical tools and algorithms for meaningful insights out of it.



Source:
https://www.edureka.co/blog/what-is-data-science/