Introduction to Emerging Technologies

Chapter Two

Data Science

Outline

- **An Overview of Data Science**
- Data Types and Data Representation
- Data Value Chain
- Basic Concepts of Big Data

Introduction

- In the previous chapter, the concept of the role of data for emerging technologies was discussed.
- In this chapter, you are going to learn more about data science, data vs. information, data types and data representation, data value chain, and basic concepts of big data.

What is Data Science?

No single definition

Components:

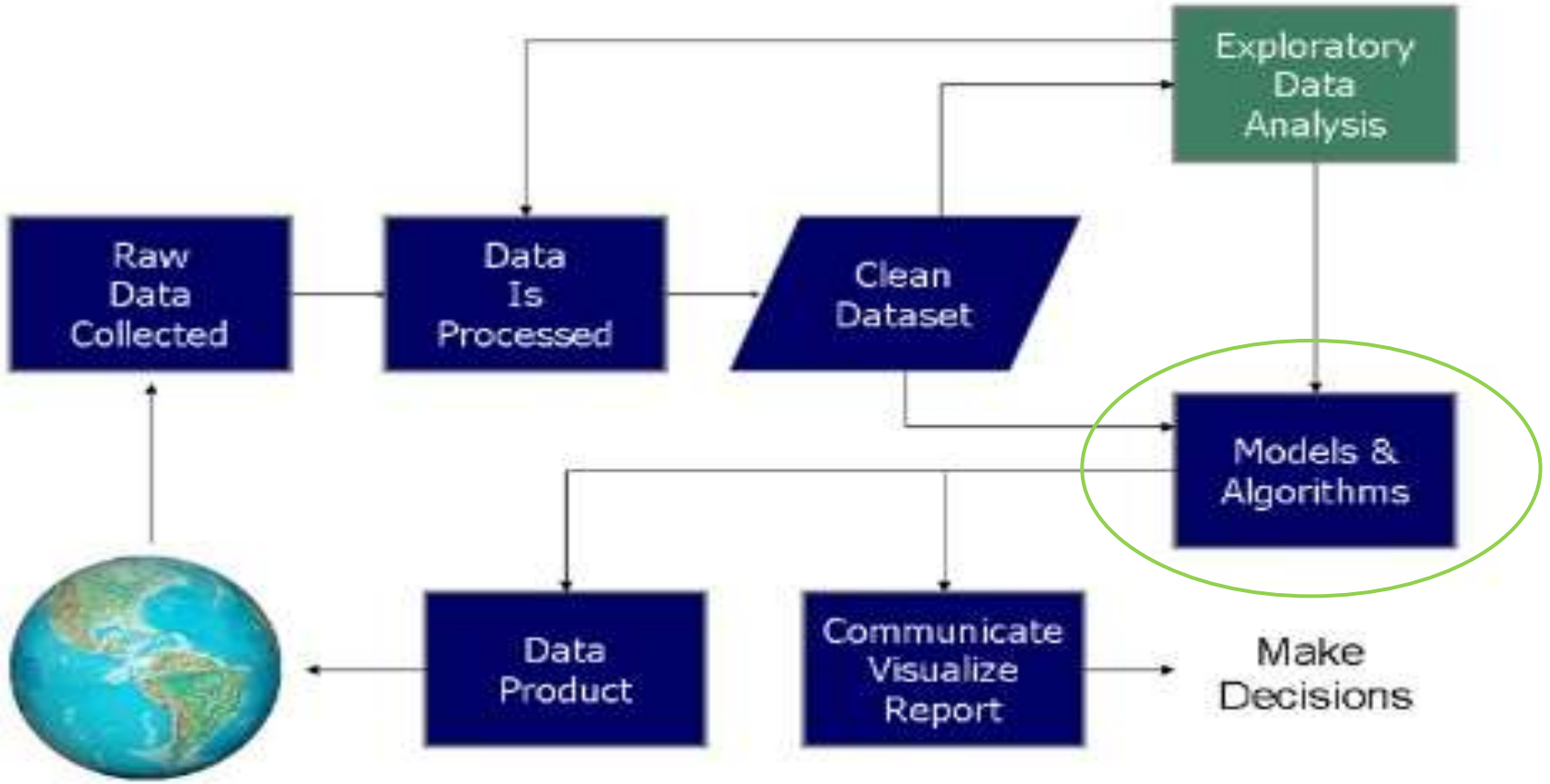
- Data-driven (the more the better)
- Interdisciplinary (math, stat, CS, …)
- Extract knowledge from observed data

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information."

Data Science is about the whole processing pipeline to extract (and utilize) information out of data.

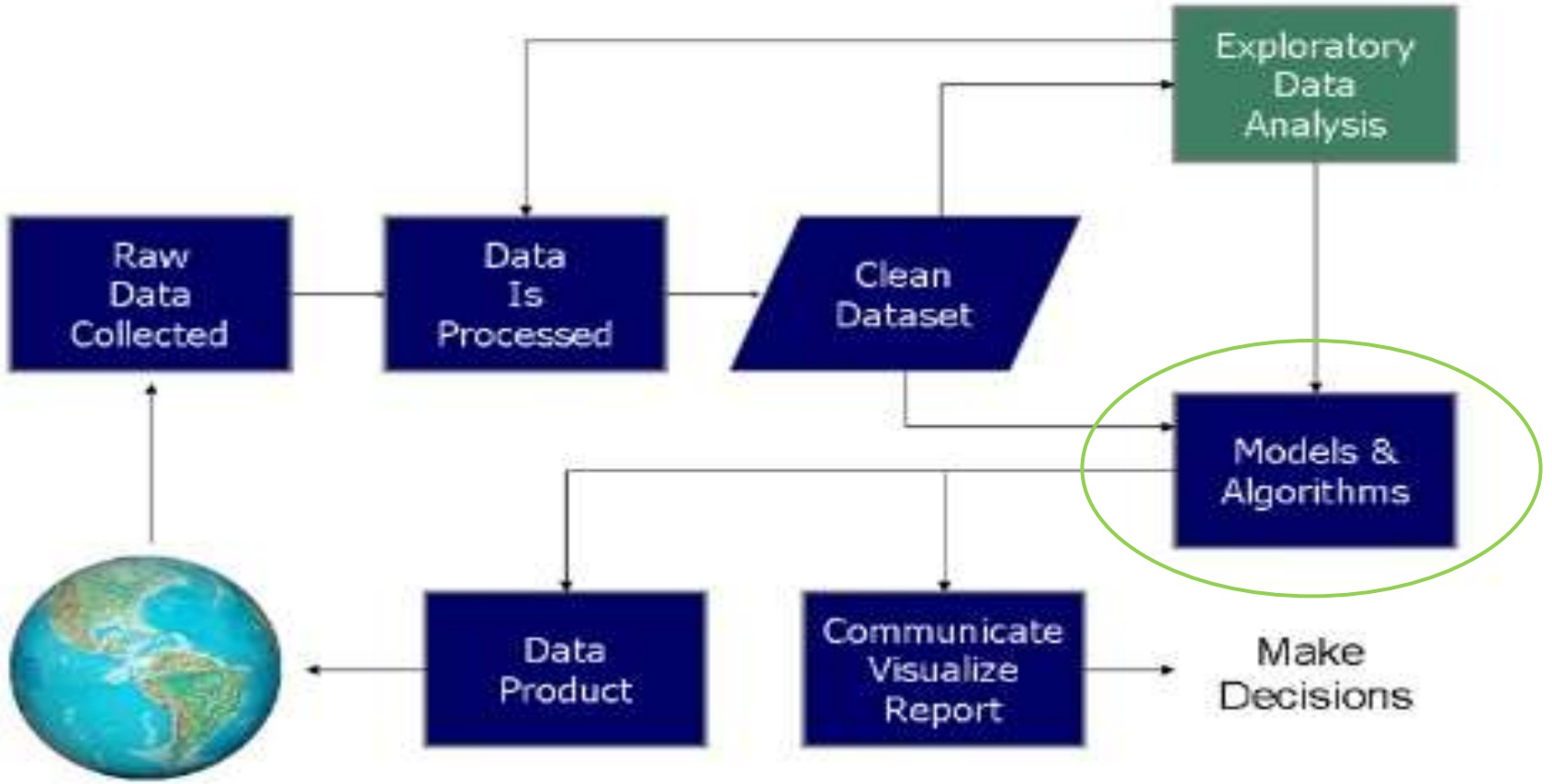
Data Scientists understand and care about the whole data pipeline.

Data Science: process



https://en.wikipedia.org/wiki/Data_analysis

Data Science: process



https://en.wikipedia.org/wiki/Data_analysis

The field of algorithms has traditionally assumed that the input data to a problem is presented in **random access memory**, which the algorithm can repeatedly access. This is not feasible for problems involving **enormous amounts of data**. **The streaming model** and other models have been formulated to reflect this. In this setting, **sampling** plays a crucial role and, indeed, we have to **sample on the fly**.

DATA SCIENCE IS ABOUT *DATA PRODUCTS*

- **Data-driven apps**
  - **Spellcheckers**
  - **Language Translators**
  - **Automatic Image Captioning apps**
- **Interactive visualizations**
  - **Google flu application**
  - **Global Burden of Disease**
- **Online Databases**
  - **Enterprise data warehouse**

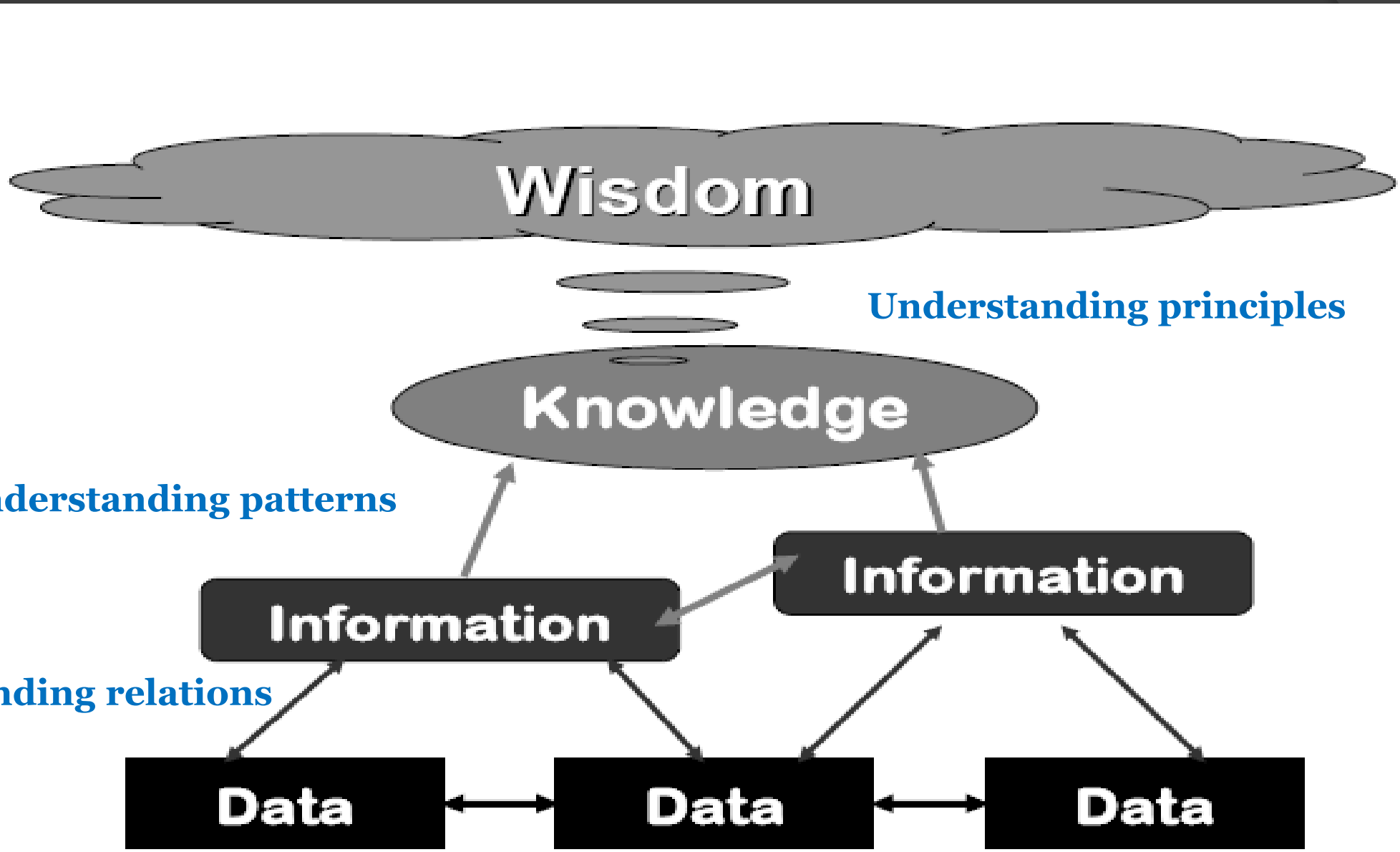
What is data?

A representation of **facts**, **concepts**, or **instructions** in a formalized manner, which should be **suitable for communication**, **interpretation**, or **processing** by human or electronic machine.

Data can be described as **unprocessed facts** and **figures.**

It can also be defined as groups of **non-random symbols** in the form of **text**, **images**, and **voice** representing quantities, actions and objects

Information Hierarchy



Definition of Information

- Information = organized data
  - Formatted, filtered, organized, structured, interpreted, summarized data
  - data + relations (context) = information
  - Relates to a description, definition or perspective (what, who, when , where)

Definition of Knowledge

- Knowledge = information that has been organized, internalized and integrated with experience, study, or intuition
  - Case, rule, process, model, ideas
  - Rules and procedures that guide decisions and actions
  - Information + application = knowledge
  - Comprises of strategy, practice, method, or approach (how)

Reading assignment: Read about Data Processing Cycle

Data types and data representations

In computer science and computer programming, a **data type** or simply type is a **structure** of data which tells the compiler or interpreter **how the programmer intends to use the data.**

Common data types include:

Integers, **Boolean**, **Characters**, **Floating-Point Numbers**, **Strings.**

These data types **define the operations that can be done on the data**, the **meaning** of each data attribute, and **the way values of those types can be stored.**

Data Types Continued…

- A **data type** defines a collection of data values and a set of predefined operations on those values.
- Data types that are not defined in terms of other types are called **primitive data types**. Nearly all programming languages provide a set of primitive data types. Some of the primitive types are merely reflections of the hardware—for example, most integer types.

Data Types Continued…

- Programming languages provide high-level **data types** such as truth values, integers, characters, records, and arrays, together with operations over these types. Target machines provide only machine 'types' such as bits, bytes, words, and double-words, together with low-level arithmetic and logical operations. To bridge the semantic gap between the source language and the target machine, the implementer must decide how to *represent* the source language's types and operations in terms of the target machine's types and operations.
- The *primitive types* of a programming language are those whose values are primitive, i.e., cannot be decomposed into simpler values.

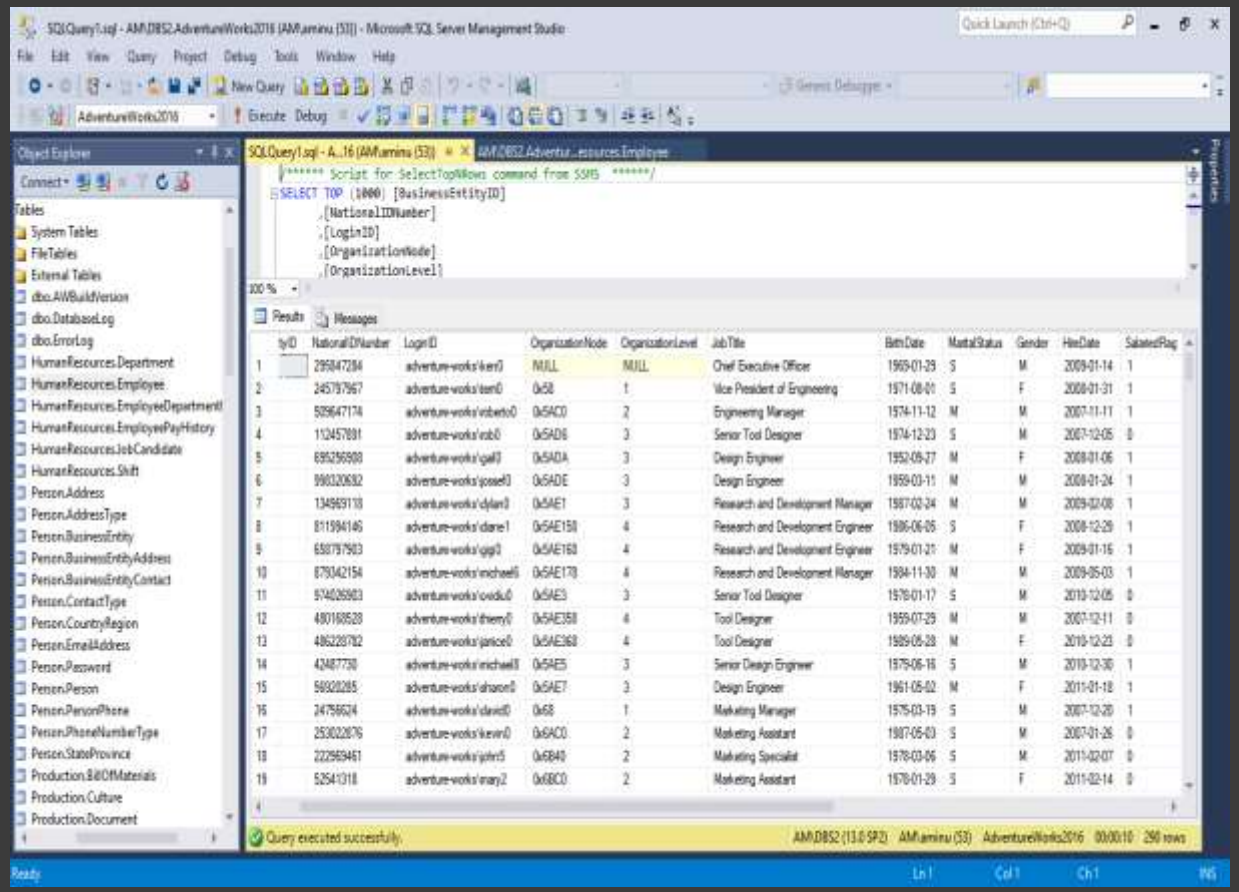
Programming Language Processors in Java, Watt & Brown

Types of Data from Data Analytics perspective

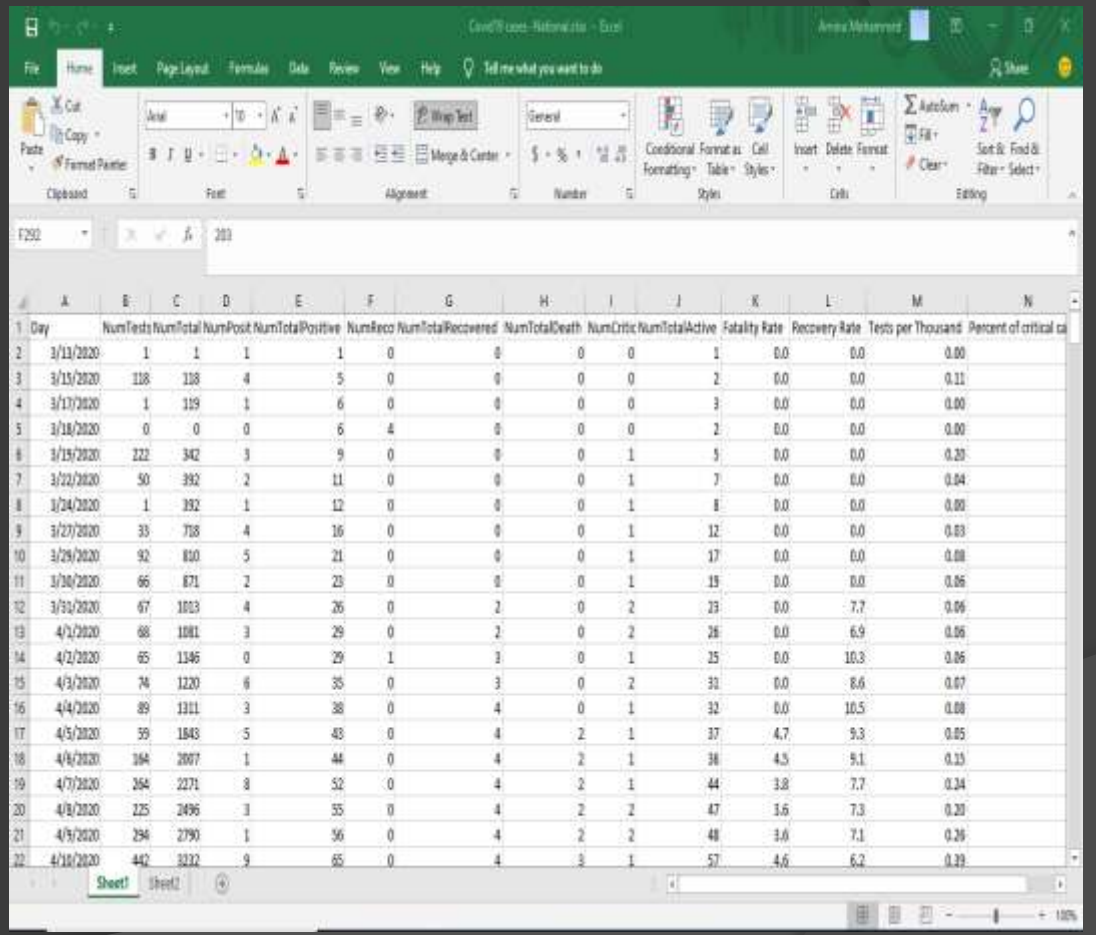
Structured, **Unstructured**, and **Semi-structured** data types

Structured Data -- examples

- SQL Data

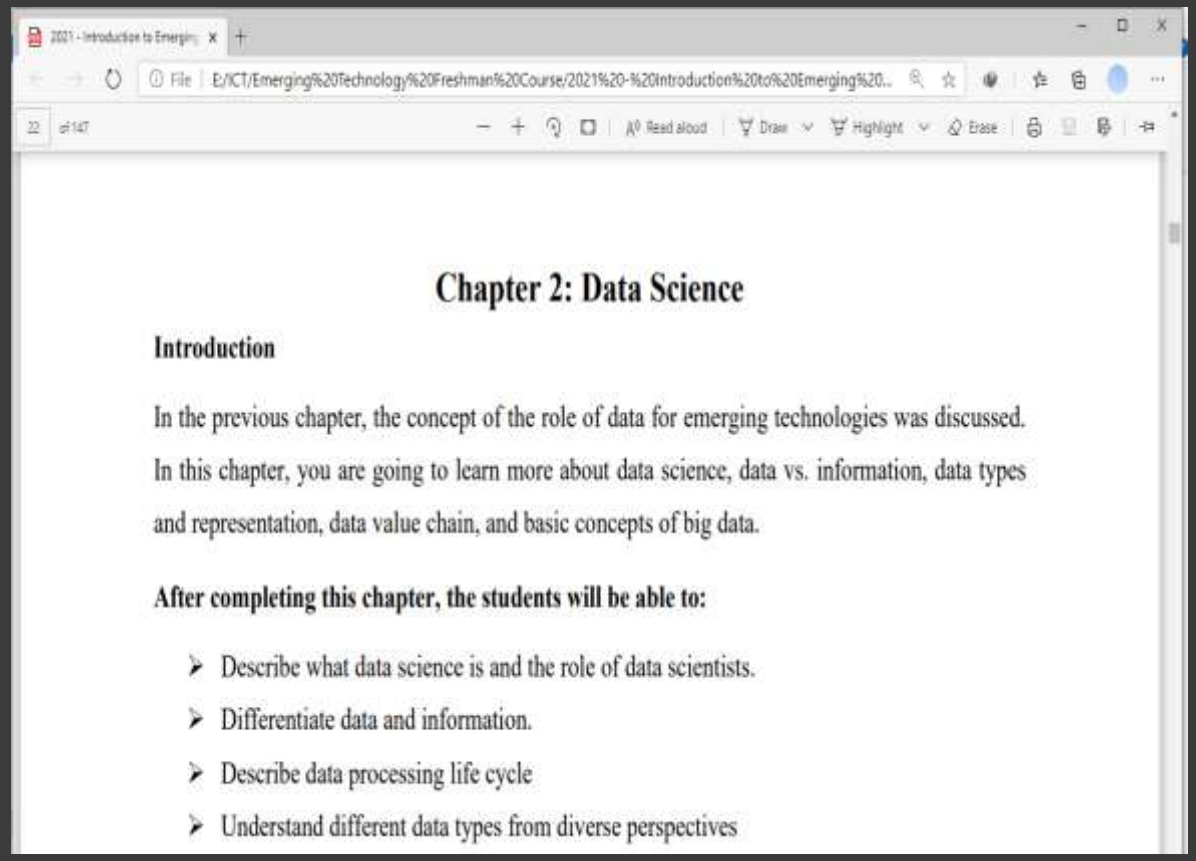


- Excel File



Unstructured Data -- examples

- Pdf files



Chapter 2: Data Science

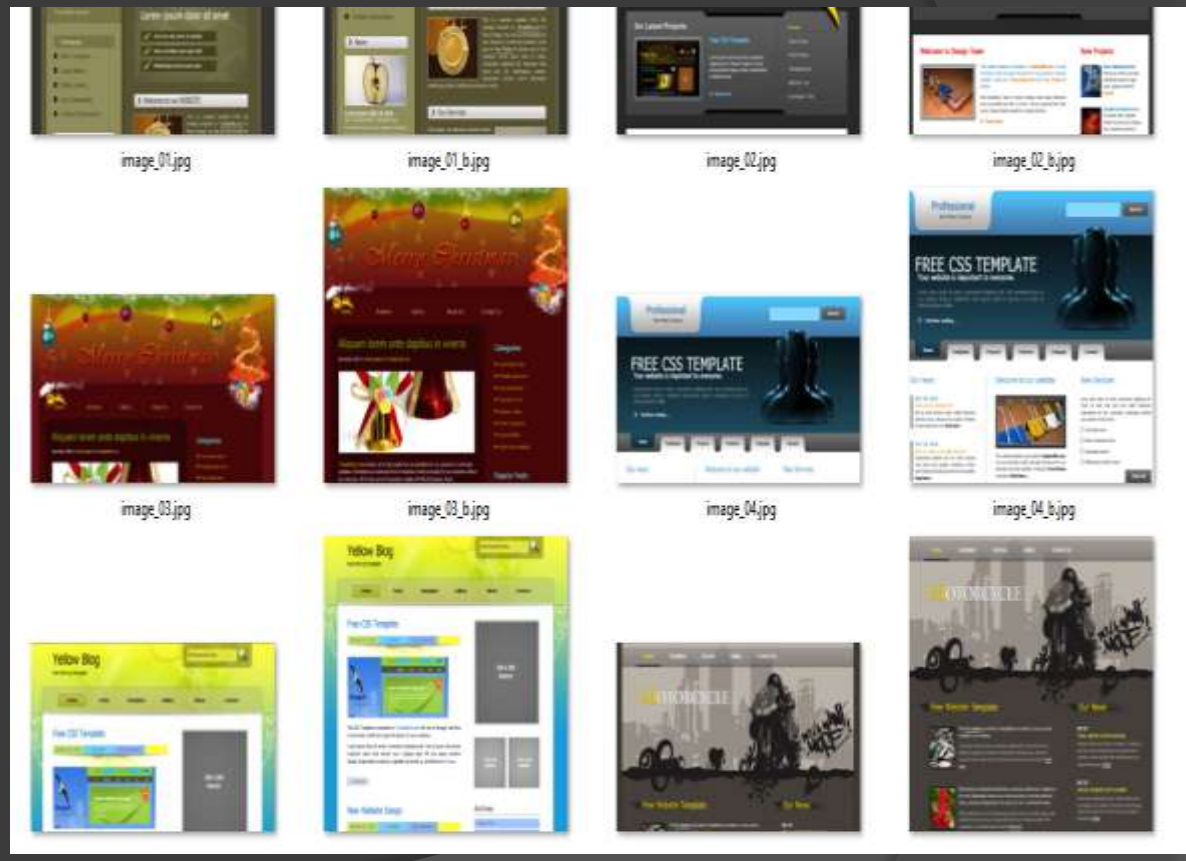
Introduction

In the previous chapter, the concept of the role of data for emerging technologies was discussed. In this chapter, you are going to learn more about data science, data vs. information, data types and representation, data value chain, and basic concepts of big data.

After completing this chapter, the students will be able to:

- Describe what data science is and the role of data scientists.
- Differentiate data and information.
- Describe data processing life cycle
- Understand different data types from diverse perspectives

- Images



Semi-structured Data -- examples

Examples of semi-structured data

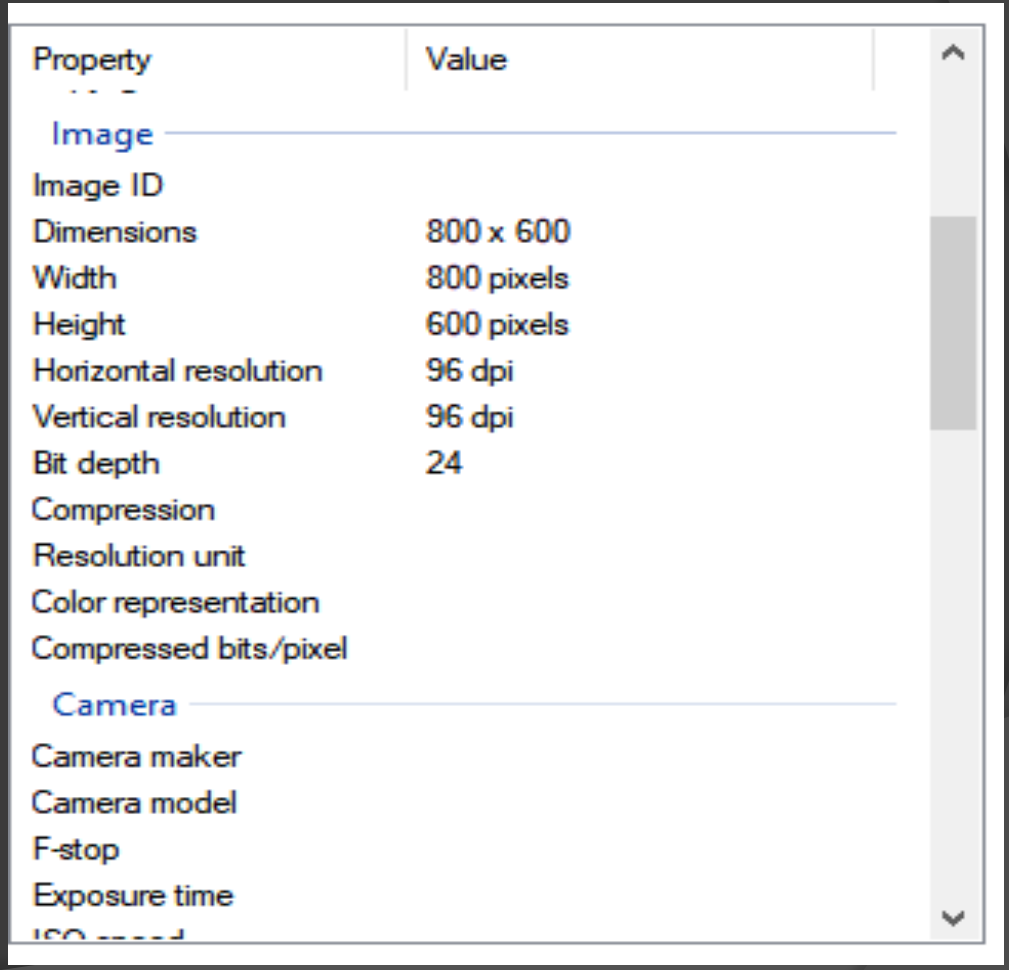
JSON and XML

```
{"employees":[
  { "firstName":"John", "lastName":"Doe" },
  { "firstName":"Anna", "lastName":"Smith" },
  { "firstName":"Peter", "lastName":"Jones" }
]}
```

```
<employees>
  <employee>
    <firstName>John</firstName> <lastName>Doe</lastName>
  </employee>
  <employee>
    <firstName>Anna</firstName> <lastName>Smith</lastName>
  </employee>
  <employee>
    <firstName>Peter</firstName> <lastName>Jones</lastName>
  </employee>
</employees>
```

Metadata -- example

- **Metadata about an image**



Property	Value
Image	
Image ID	
Dimensions	800 x 600
Width	800 pixels
Height	600 pixels
Horizontal resolution	96 dpi
Vertical resolution	96 dpi
Bit depth	24
Compression	
Resolution unit	
Color representation	
Compressed bits/pixel	
Camera	
Camera maker	
Camera model	
F-stop	
Exposure time	

Data Science Continued

- **Data Value Chain**
- **Basic Concepts of Big Data**

Data Value Chain

Describes the information flow within a big data system as a **series of steps** needed **to generate value** and **useful insights** from data.

The **Big Data Value Chain** identifies the following key high-level activities:

Data Acquisition, **Data Analysis**, **Data Curation**, **Data Storage**, **Data Usage**

Data Acquisition

The infrastructure required for data acquisition must

- deliver **low**, **predictable latency** in both capturing data and in executing queries
- be able to **handle very high transaction volumes**, often in a distributed environment
- support **flexible** and **dynamic data structures**

Data Analysis

Involves **exploring**, **transforming**, and **modelling** data with the goal of **highlighting relevant data**, **synthesizing** and **extracting** useful hidden information with high potential from a business point of view.

Related areas include **data mining**, **business intelligence**, and **machine learning.**

What is synthesizing?

Data Curation

Data curation processes can be categorized into different activities

- **content creation**, **selection**, **classification**, **transformation**, **validation**, and **preservation**
- ensuring that data are **trustworthy**, **discoverable**, **accessible**, **reusable**, and **fit** for purpose

A key trend for the curation of big data **utilizes community** and **crowd sourcing** approaches (**attribute of IR4**).

Data Storage

Relational databases that guarantee database transactions, **lack flexibility** with regard to **schema changes**, **performance** and **fault tolerance** when data volumes and complexity grow, making them **unsuitable for big data scenarios**

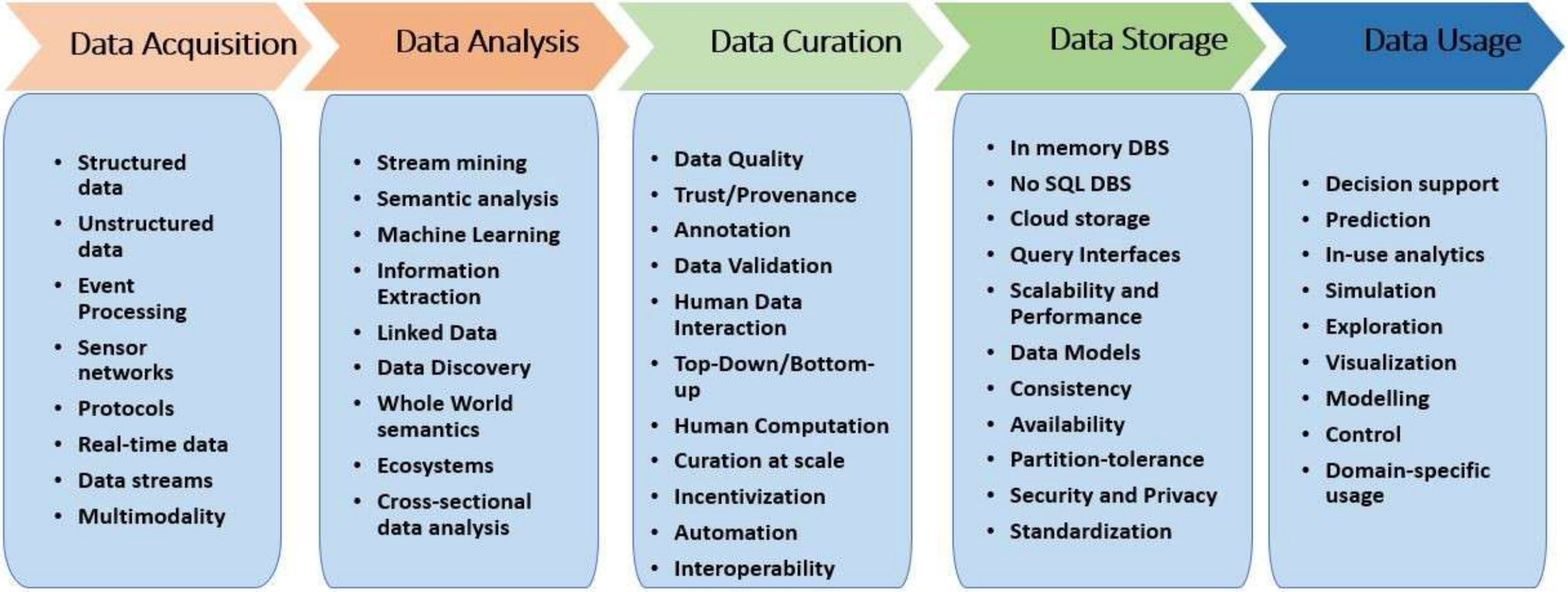
NoSQL technologies have been designed with the **scalability** goal in mind and present a wide range of solutions based on **alternative data models**

Data Usage

Covers the **data-driven business activities** that need **access** to the curated data, its **in-use analytics**, and the tools needed to **integrate** the data analyses within the business activity.

In business decision-making , it can **enhance competitiveness** through **reduction of costs**, **increased added value**, or any other parameter that can be measured against existing performance criteria

Big Data Value Chain Main Activities



Basic concepts of big data

Big data is a blanket term for the **non-traditional** strategies and technologies needed to gather, organize, process, and gain **insights** from large **datasets**. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the **pervasiveness**, **scale**, and **value** of this type of data have greatly expanded in recent years.

What Is Big Data?

A collection of datasets **so large and complex** that it becomes difficult to process using on-hand database management tools or traditional data processing applications and individual machines. ***Wikipedia***

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" ***Laney 2001***

"When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam"

Loukides 2010

The 4 Vs Characterizing Big Data.

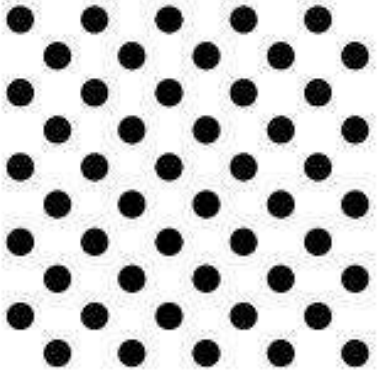
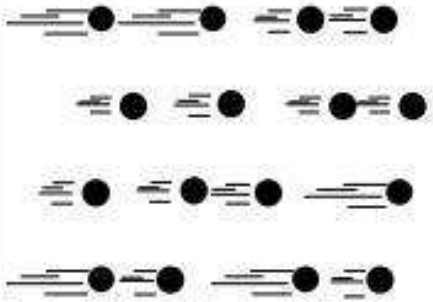
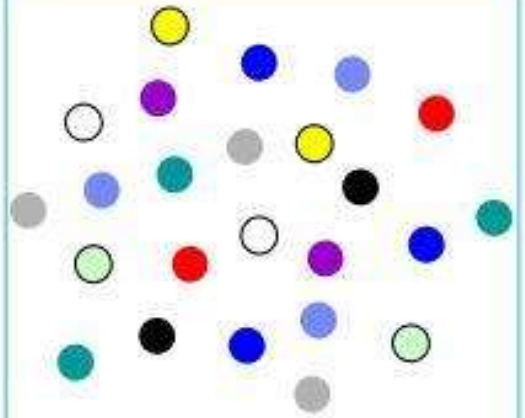
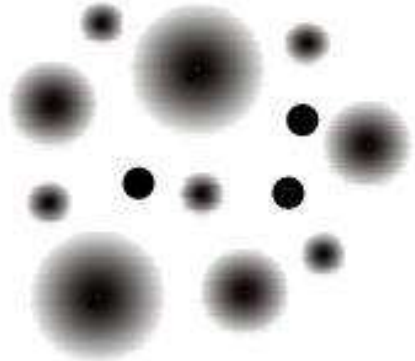
Volume: large amounts of data in yottabytes or zetabytes, (massive datasets)

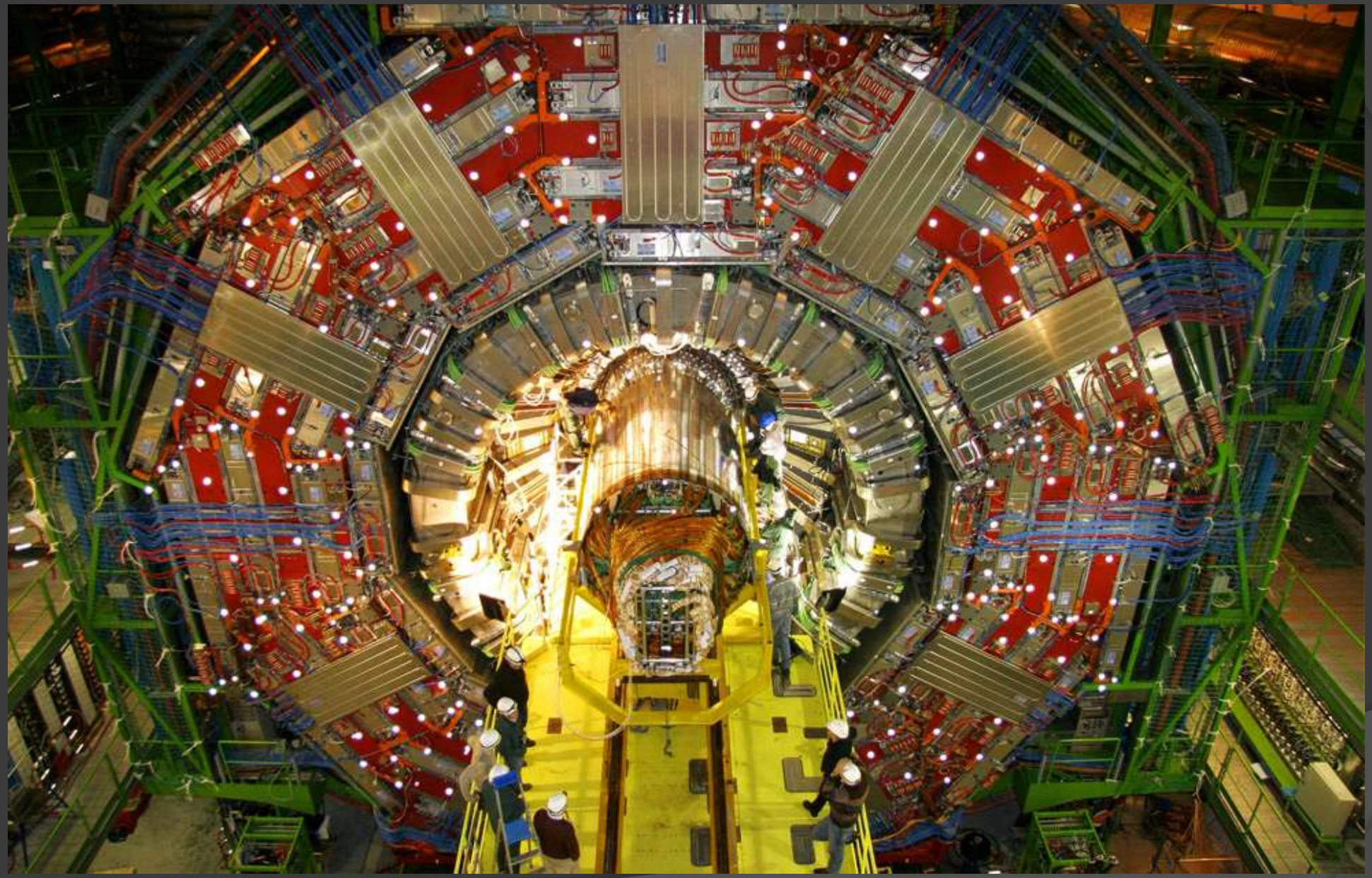
Velocity: Data is live streaming or in motion

Variety: data comes in many different forms / from diverse sources and formats

Veracity: can we trust the data? How accurate is it? Doubt in data, etc.

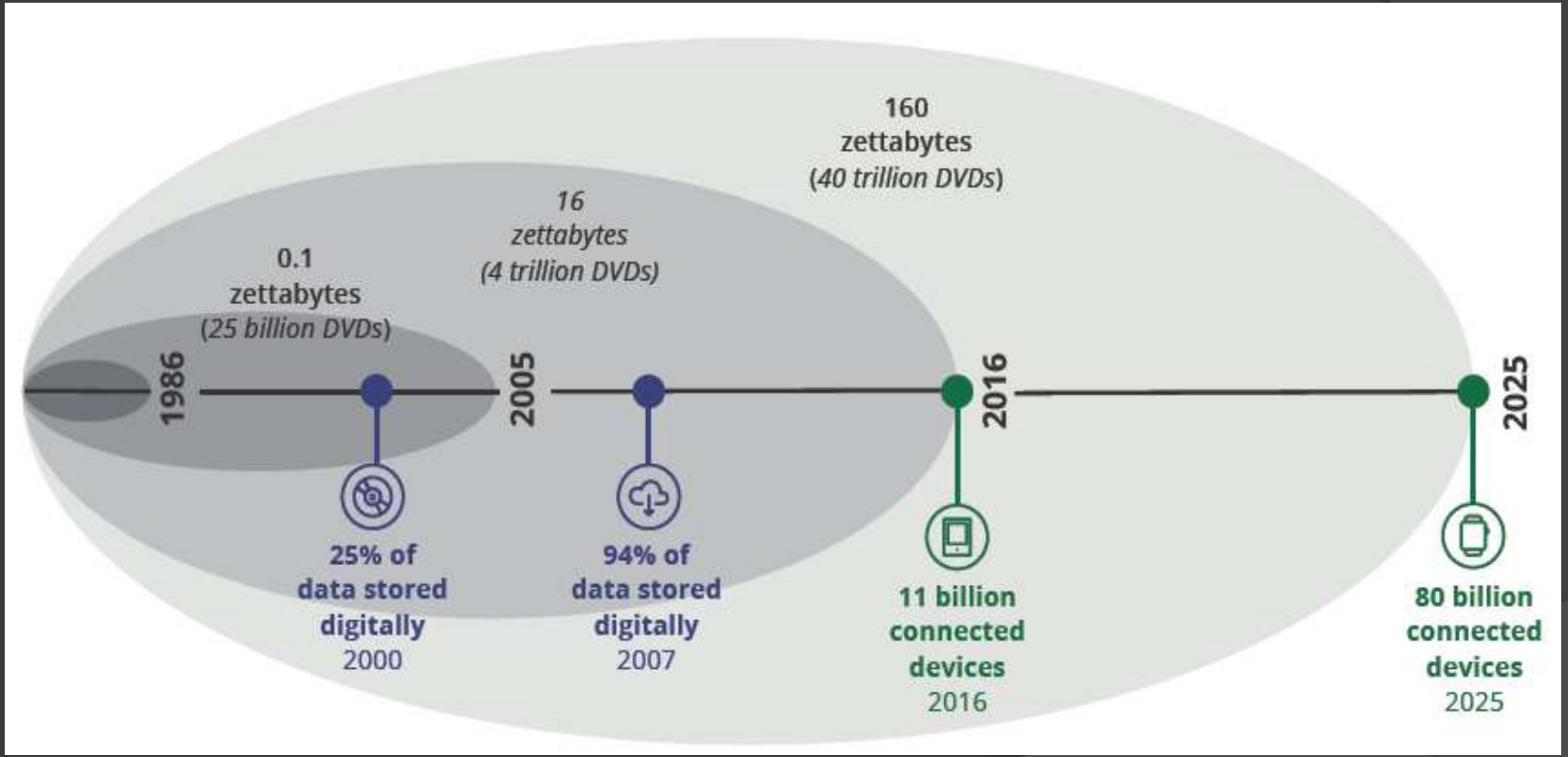
The 4 Vs

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations



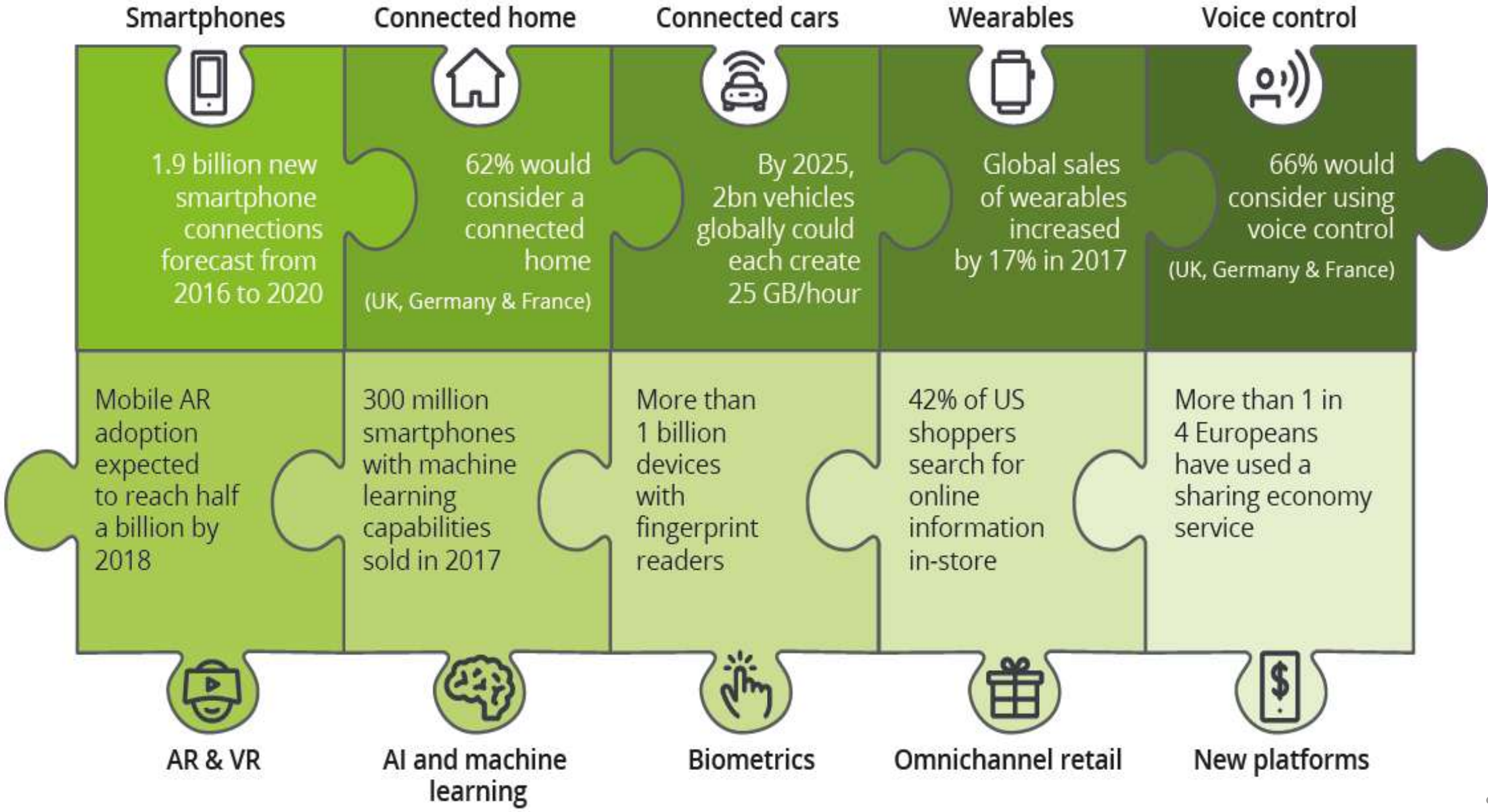
CERN’s Large Hydron Collider (LHC) generates 15 PB a year

Illustration of global data growth trends over time



1 zettabyte = 1,000 exabytes = 1,000,000 petabytes = 1,000,000,000 terabytes = 1,000,000,000,000 gigabytes

Drivers of data growth (Big data drivers)



Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data) in Databases
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Networks, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc.)

To extract knowledge➔ all these types of data need to be linked together

Velocity (Speed)

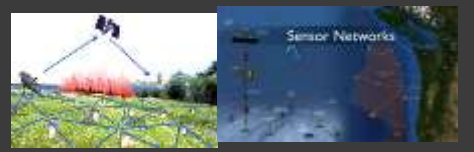
- Data is being generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- **Example**
  - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and computer networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion.

Assignment 1

- Write an article explaining the difference between relational database systems (RDBMS) and HDFS. Include other Hadoop equivalent systems as much as possible in your discussion. Read about other types of file systems (reading assignment).
  - Individual , in 10 days (on or before November 10).
  - Follow Standard formatting rules.
  - Similar works are not allowed; shall be rejected.
  - Lamination not required; Maximum of 5 pages (you decide).

Clustered Computing and Hadoop Ecosystem

- Because of the qualities of big data, individual computers are often inadequate for handling big data at most stages.
- To address the high storage and computational needs of big data, computer clusters are a better fit.

Clustered Computing Continued…

- Big data clustering software **combines** the **resources** of many **smaller machines**, seeking to provide the following benefits:
  - **Resource pooling**: Storage, CPU, Working memory
  - **High availability**: Fault tolerance, Availability guaranties
  - **Easy scalability**: Easy to scale horizontally by adding more machines

Clustered Computing Continued…

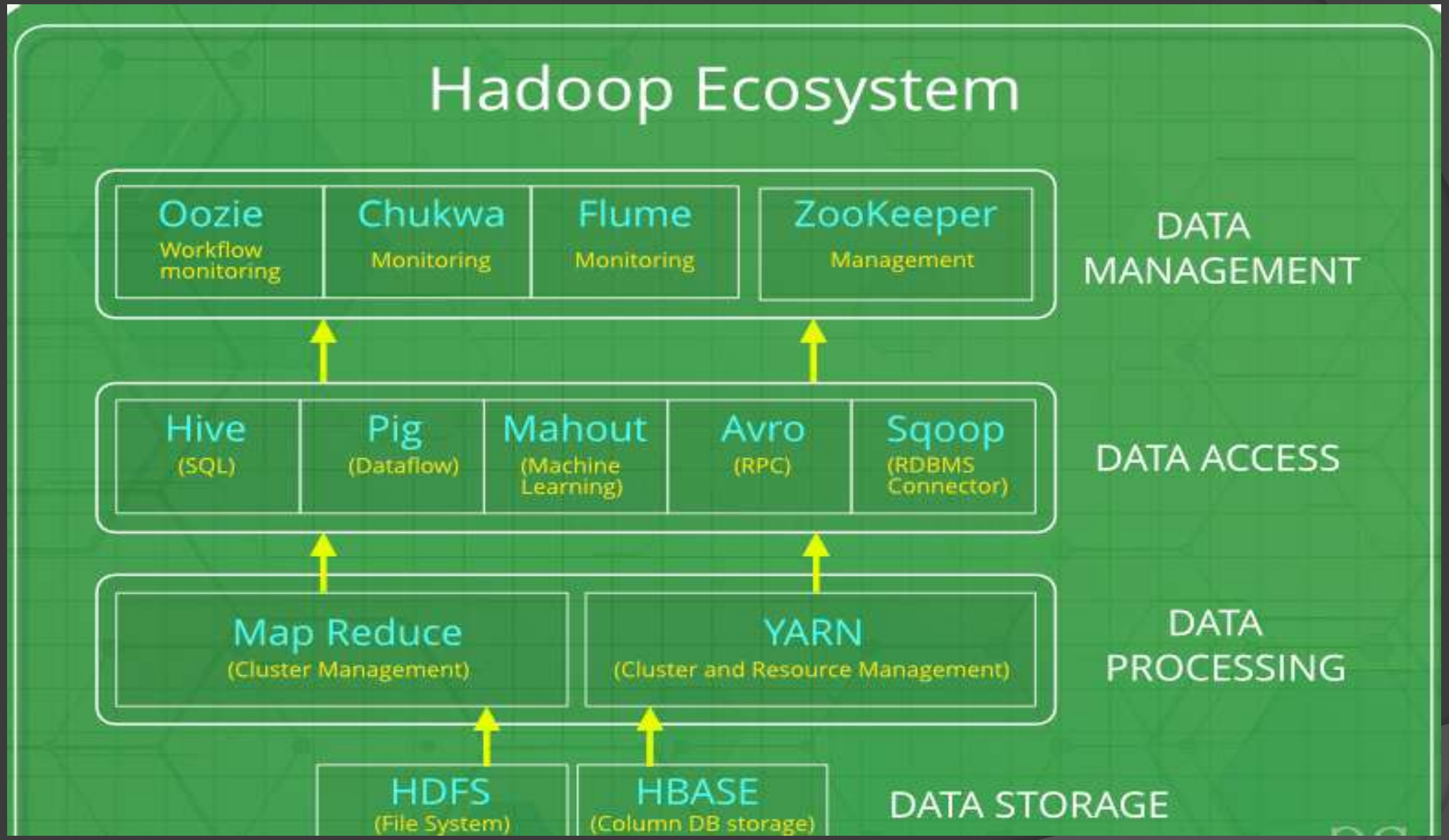
- Using clusters requires a **solution** for **managing cluster membership**, coordinating **resource sharing**, and **scheduling actual work** on individual nodes. Cluster membership and resource allocation can be handled by software like **Hadoop's YARN** (Yet Another Resource Negotiator).
- The machines involved in the computing cluster are also typically involved with the management of a distributed storage system (**distributed computing**).

Hadoop and its Ecosystem

- Hadoop is an open-source **framework** intended **to make interaction with big data easier**. It is a framework that allows distributed processing of large datasets across **clusters** of computers using simple programming models. It is inspired by a technical document published by Google.

Hadoop and its Ecosystem Cont…

- Four key **characteristics** of Hadoop are:
  - **Economical:** highly economical as ordinary computers can be used for data processing.
  - **Reliable**
  - **Scalable**
  - **Flexible**



Hadoop Ecosystem

Big Data Life Cycle with Hadoop

- **Ingesting data into the system:** data is transferred to **Hadoop** from various sources such as relational databases, systems, or local files. **Sqoop** transfers data from RDBMS to HDFS, whereas **Flume** transfers event data.
- **Processing data for storage and (out from storage)**
- **Computing and analyzing data**
- **Visualizing the results**