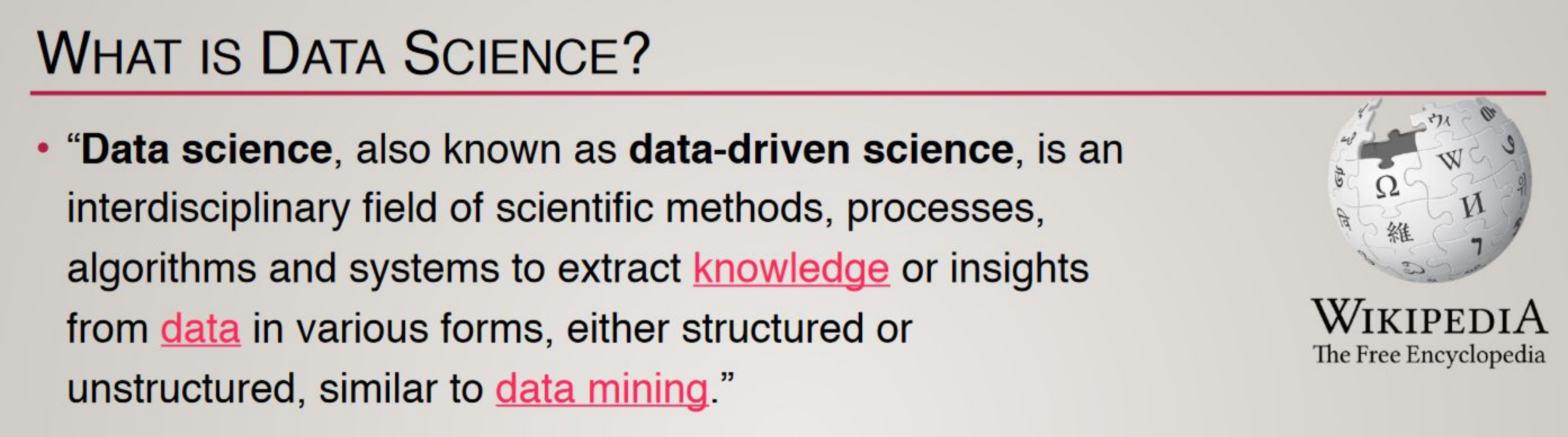
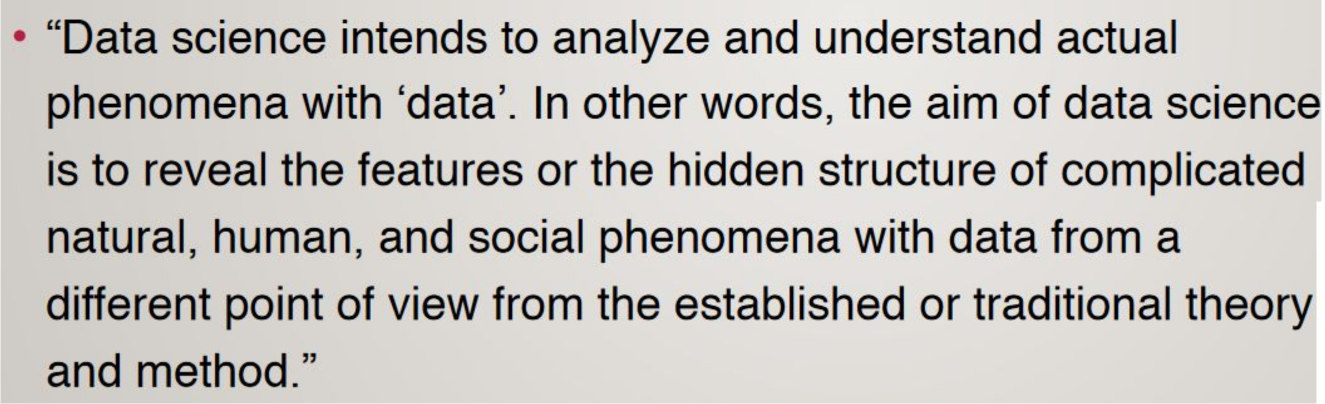
Data Science Fundamental
UCS 538



What is Data Science?

- “**Data science**, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.”

Wikipedia
The Free Encyclopedia

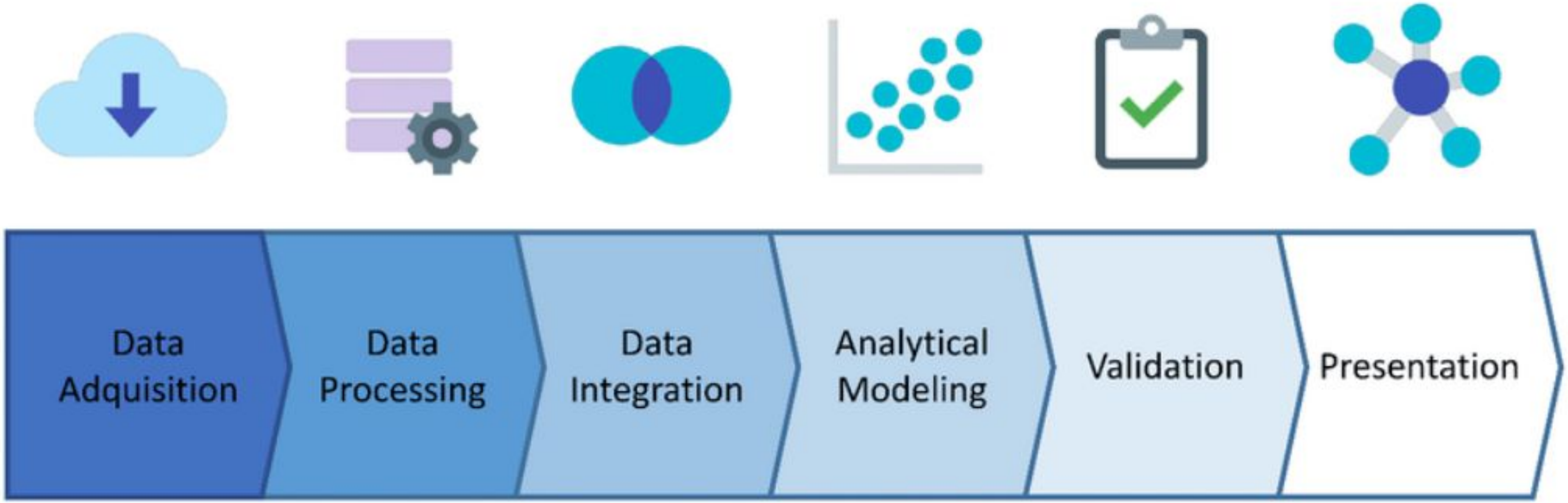


- “Data science intends to analyze and understand actual phenomena with ‘data’. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method.”

What is Important?

Need to solve a real problem using data…
No applications, no data science.

Pipeline



Data Adquisition

Data Processing

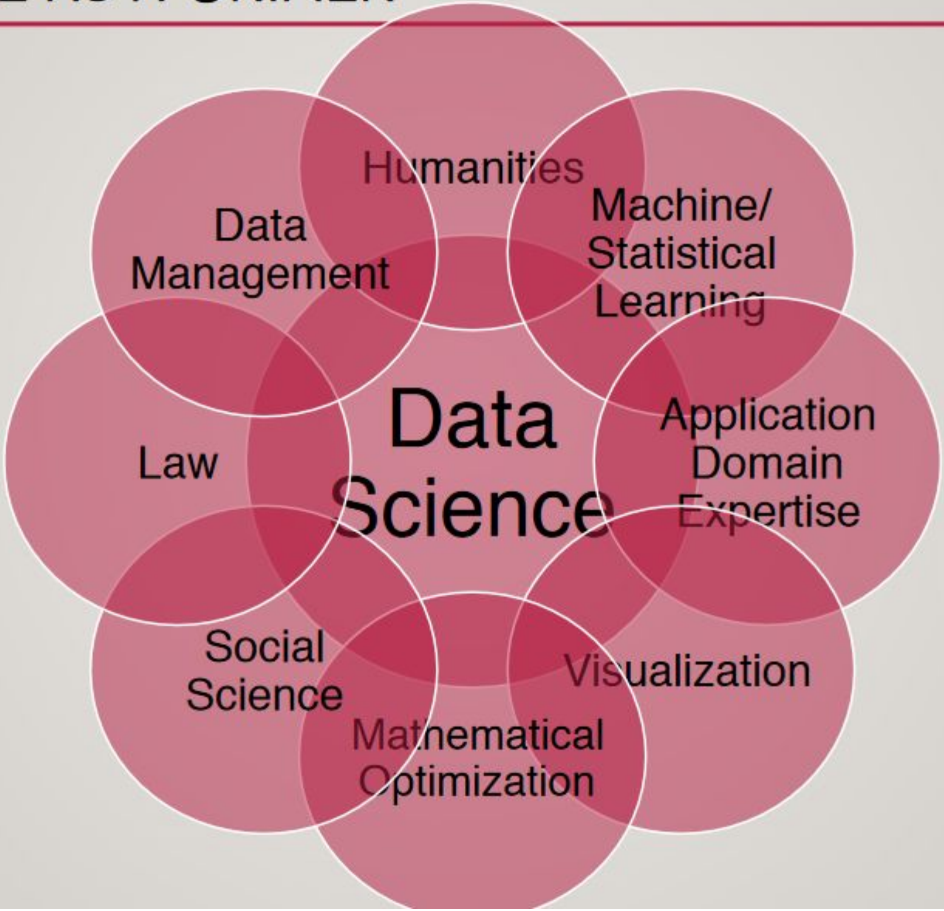
Data Integration

Analytical Modeling

Validation

Presentation

Data Science as a Unifier



Data Science

Humanities

Machine/ Statistical Learning

Application Domain Expertise

Visualization

Mathematical Optimization

Social Science

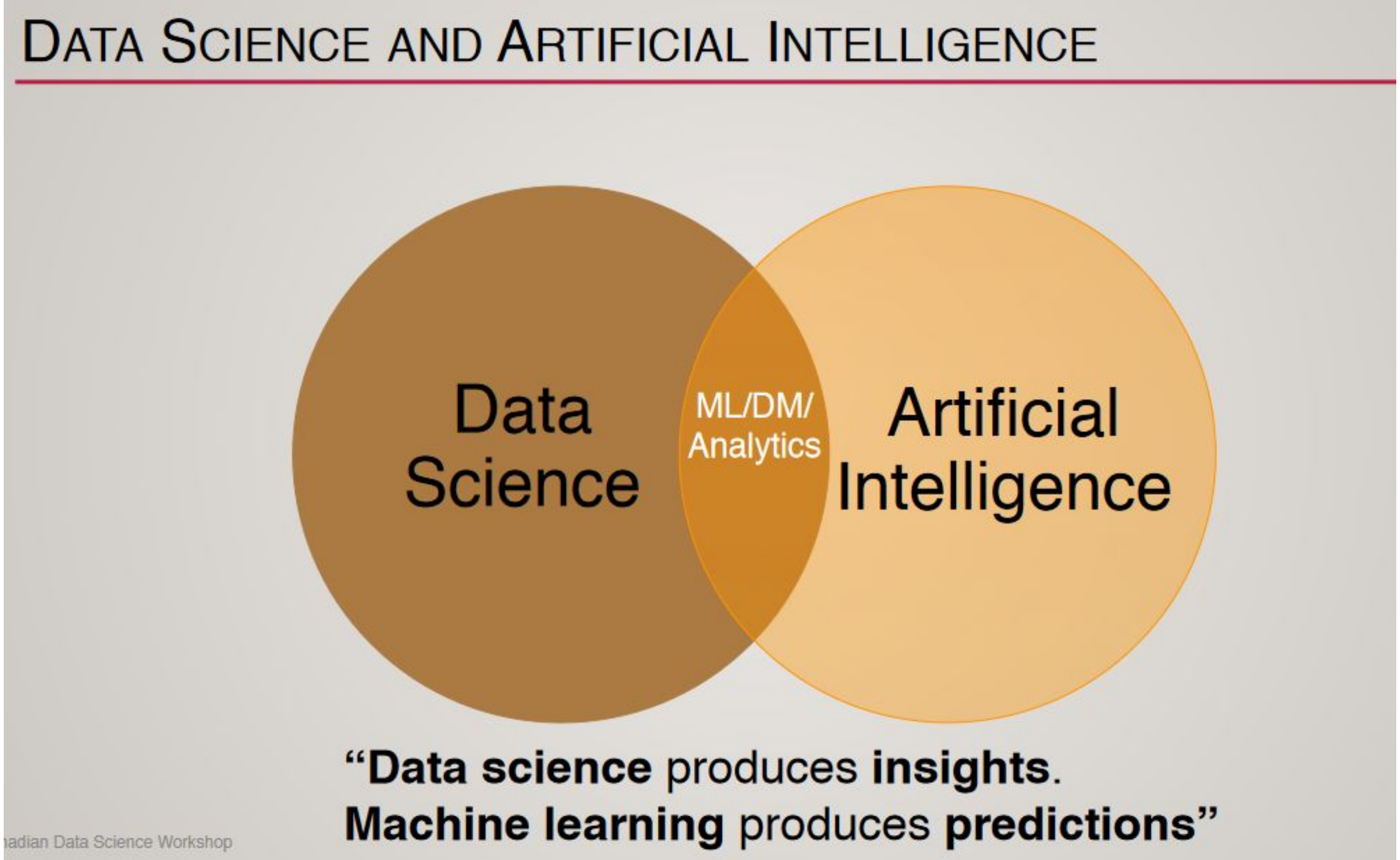
Law

Data Management

Data Science and Big Data

- They are not the “same thing”
- Big data = crude oil
  - Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”
- Data science is about refining the “crude oil”

Carlos Samohano
Founder, Data Science London



Data Science and Artificial Intelligence

“**Data science** produces **insights**.
Machine learning produces **predictions**”

DATA SCIENCE APPLICATION EXAMPLES

- Fraud detection
  - Investigate fraud patterns in past data
  - Early detection is important
    - Before damage propagates
    - Harder than late detection
  - Precision is important
    - False positive and false negative are both bad
  - Real-time analytics

Data Science Application Examples

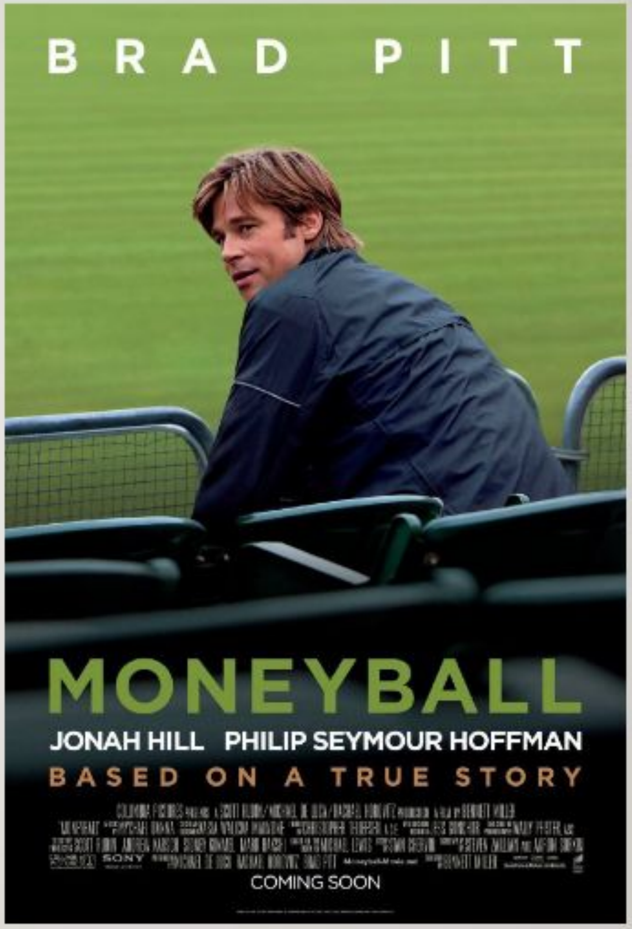
- Recommender systems
  - The ability to offer unique personalized service
  - Increase sales, click-through rates, conversions, …
    - Netflix recommender system valued at $1B per year
    - Amazon recommender system drives a 20-35% lift in sales annually
  - Collaborative filtering at scale

DATA SCIENCE APPLICATION EXAMPLES

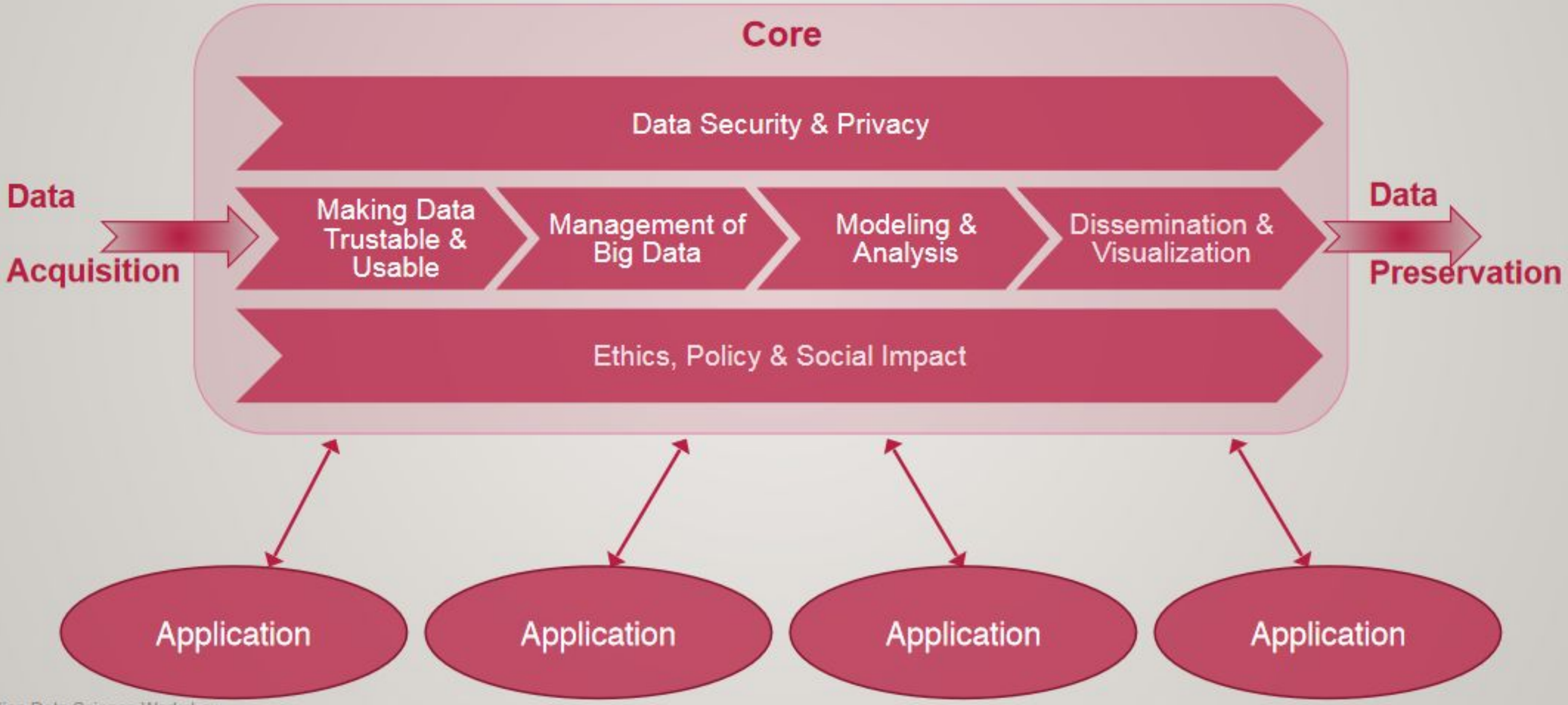
- Predicting why patients are being readmitted
  - Reduce costs
  - Improve population health
  - Find the “why” behind specific populations being readmitted
  - Data lakes of multiple data sources
  - Investigate ties between readmission an socioeconomic data points, patient history, genetics, …

Data Science Application Examples

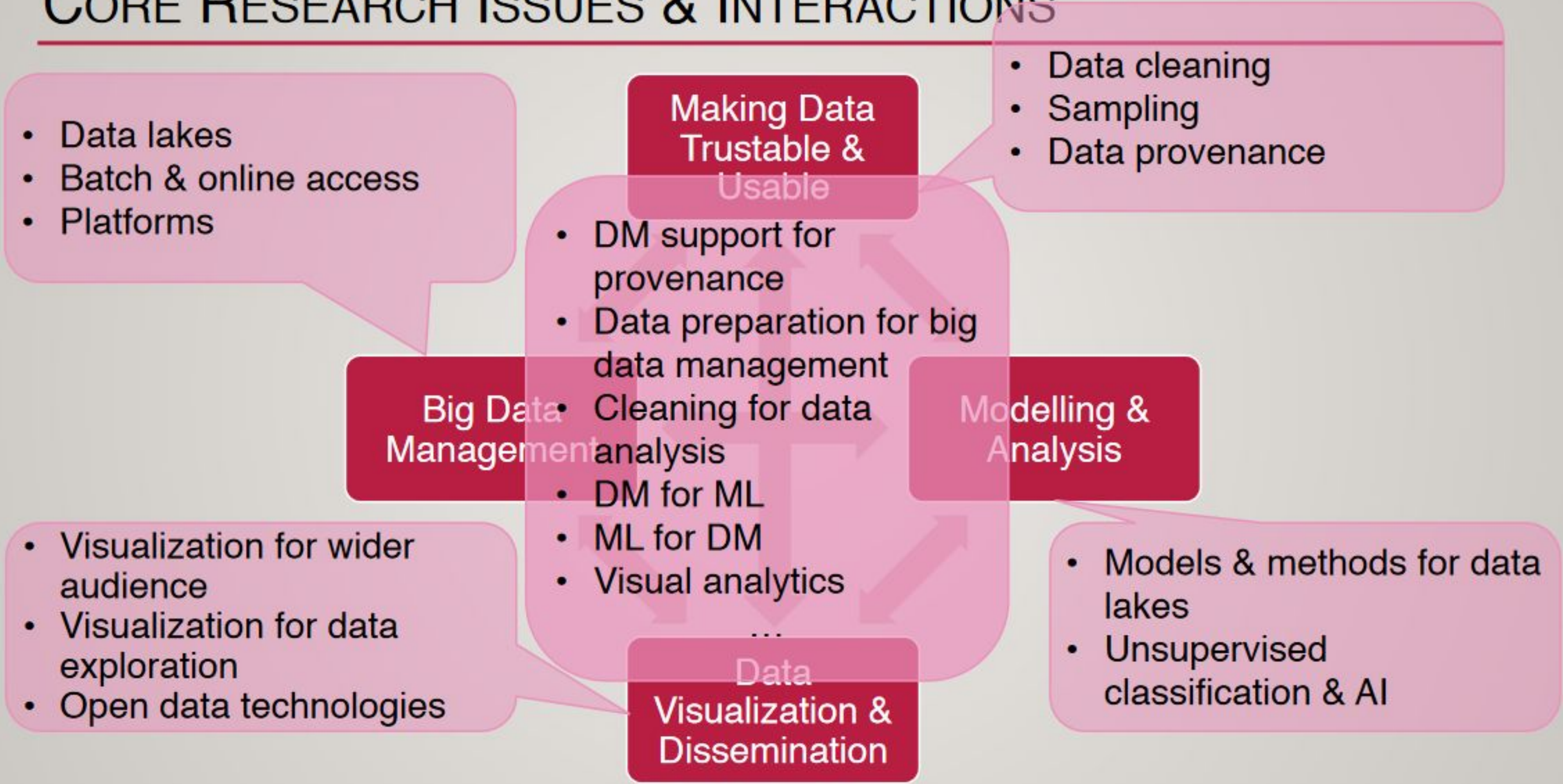
- Moneyball
  - How to build a baseball team on a very low budget by relying on data
  - *Sabermetrics*: the statistical analysis of baseball data to objectively evaluate performance
  - 2002 record of 103-59 was joint best in MLB
    - Team salary budget: $40 million
  - Other team: Yankees
    - Team salary budget: $120 million



Holistic Approach to Data Science



Core Research Issues & Interactions



Making Data Trustable & Usable

- Data cleaning
- Sampling
- Data provenance

Big Data Management

- Data lakes
- Batch & online access
- Platforms

- DM support for provenance
- Data preparation for big data management
- Cleaning for data analysis
- DM for ML
- ML for DM
- Visual analytics

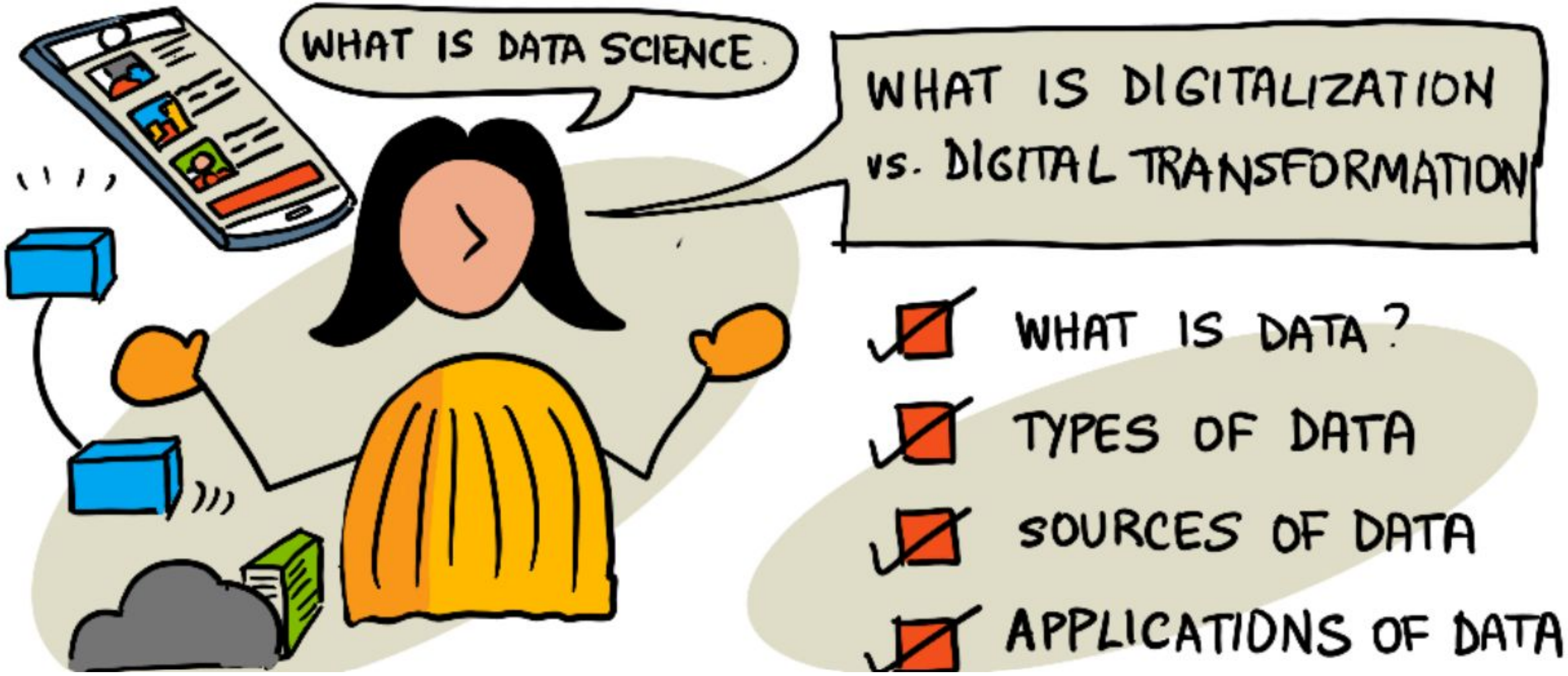
…

Modelling & Analysis

- Models & methods for data lakes
- Unsupervised classification & AI

Data Visualization & Dissemination

- Visualization for wider audience
- Visualization for data exploration
- Open data technologies


WHAT IS DATA SCIENCE.
WHAT IS DIGITALIZATION vs. DIGITAL TRANSFORMATION
☑ WHAT IS DATA?
☑ TYPES OF DATA
☑ SOURCES OF DATA
☑ APPLICATIONS OF DATA

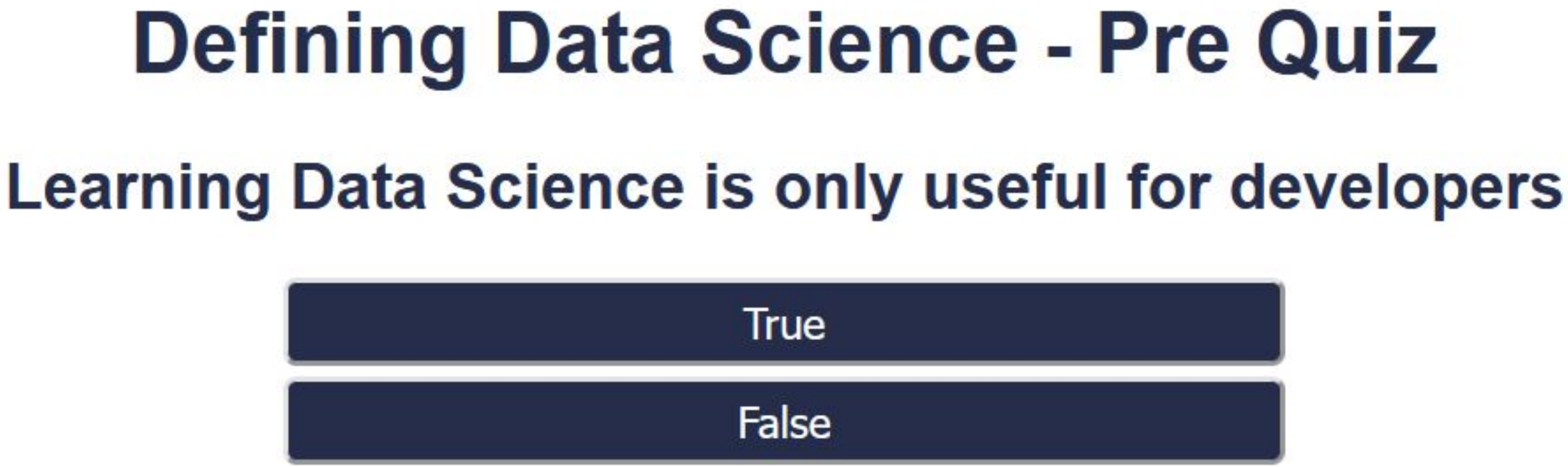
Defining Data Science - Pre Quiz

Why is the word _Science_ in Data Science?

It uses scientific methods to analyze data

Only people with academic degrees can understand it

To make is sound cool



Defining Data Science - Pre Quiz

Learning Data Science is only useful for developers

What do we need to demonstrate that basketball players are taller than average people?

- Collect some data
- Know some probability and statistics
- All of the above

What is Data?

In our everyday life, we are constantly surrounded by data.

- The list of phone numbers of your friends in your smartphone is data, as well as the current time displayed on your watch.
- As human beings, we naturally operate with data by counting the money we have or by writing letters to our friends.
- However, data became much more critical with the creation of computers. The primary role of computers is to perform computations, but they need data to operate on. Thus, we need to understand how computers store and process data.
- With the emergence of the Internet, the role of computers as data handling devices increased. If you think about it, we now use computers more and more for data processing and communication, rather than actual computations.
- When we write an e-mail to a friend or search for some information on the Internet - we are essentially creating, storing, transmitting, and manipulating data.

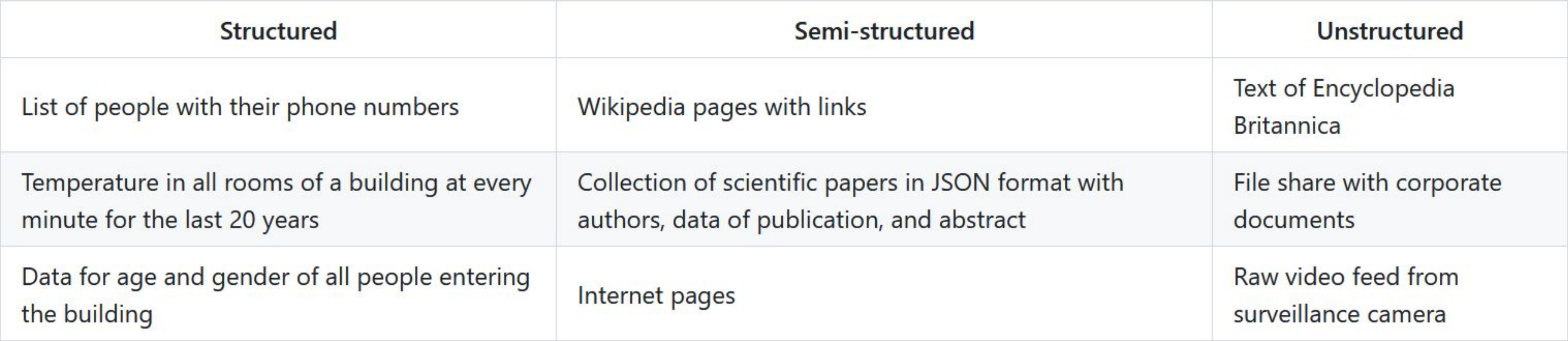
This definition highlights the following important aspects of data science:

- The main goal of data science is to **extract knowledge** from data, in other words - to **understand** data, find some hidden relationships and build a **model**.
- Data science uses **scientific methods**, such as probability and statistics. In fact, when the term *data science* was first introduced, some people argued that data science was just a new fancy name for statistics. Nowadays it has become evident that the field is much broader.
- Obtained knowledge should be applied to produce some **actionable insights**, i.e. practical insights that you can apply to real business situations.
- We should be able to operate on both **structured** and **unstructured** data. We will come back to discuss different types of data later in the course.
- **Application domain** is an important concept, and data scientists often need at least some degree of expertise in the problem domain, for example: finance, medicine, marketing, etc.

Types of Data

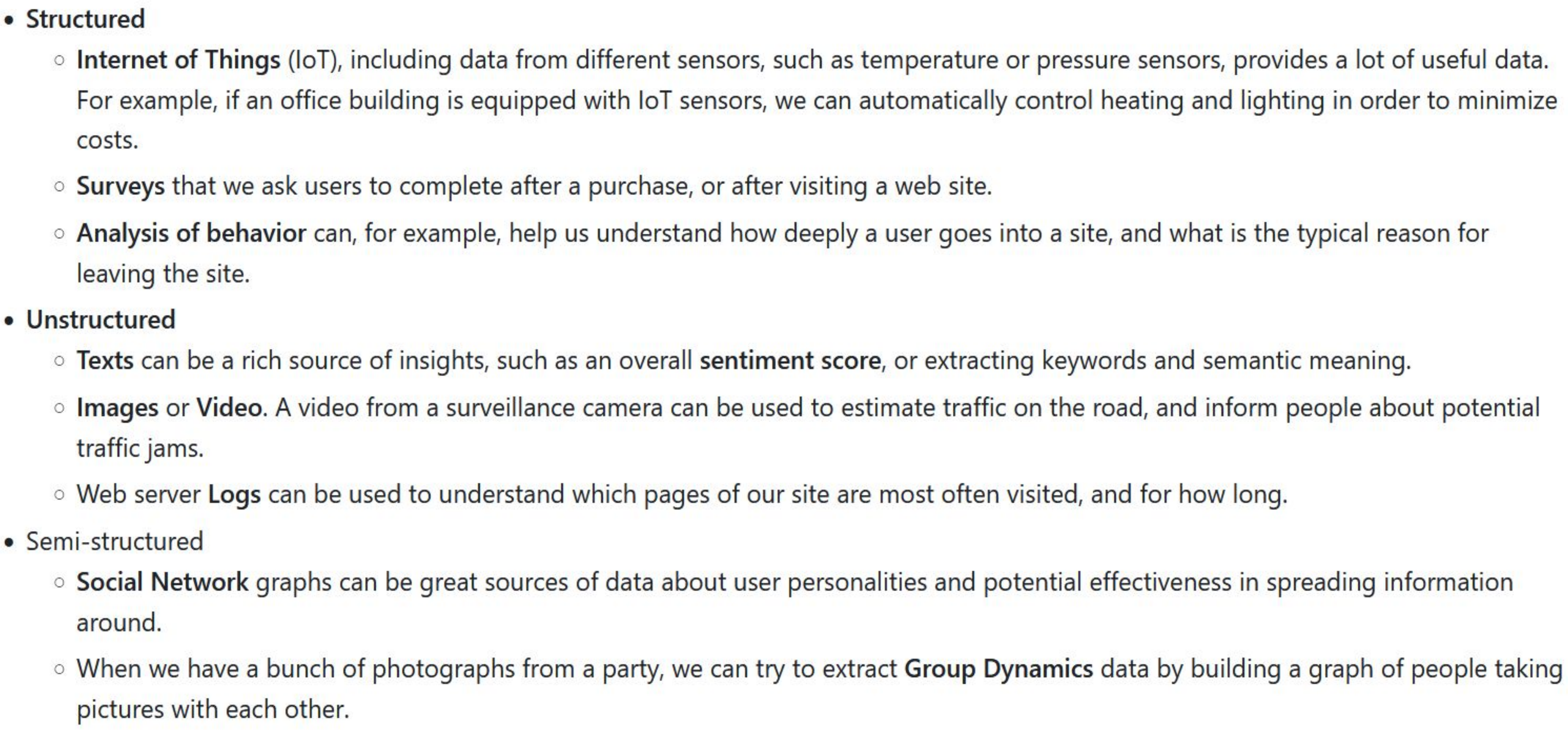
Data is everywhere. We just need to capture it in the right way!

It is useful to distinguish between **structured** and **unstructured** data.



Structured	Semi-structured	Unstructured
List of people with their phone numbers	Wikipedia pages with links	Text of Encyclopedia Britannica
Temperature in all rooms of a building at every minute for the last 20 years	Collection of scientific papers in JSON format with authors, data of publication, and abstract	File share with corporate documents
Data for age and gender of all people entering the building	Internet pages	Raw video feed from surveillance camera

Where to get Data



- **Structured**
  - **Internet of Things** (IoT), including data from different sensors, such as temperature or pressure sensors, provides a lot of useful data. For example, if an office building is equipped with IoT sensors, we can automatically control heating and lighting in order to minimize costs.
  - **Surveys** that we ask users to complete after a purchase, or after visiting a web site.
  - **Analysis of behavior** can, for example, help us understand how deeply a user goes into a site, and what is the typical reason for leaving the site.
- **Unstructured**
  - **Texts** can be a rich source of insights, such as an overall **sentiment score**, or extracting keywords and semantic meaning.
  - **Images** or **Video**. A video from a surveillance camera can be used to estimate traffic on the road, and inform people about potential traffic jams.
  - Web server **Logs** can be used to understand which pages of our site are most often visited, and for how long.
- Semi-structured
  - **Social Network** graphs can be great sources of data about user personalities and potential effectiveness in spreading information around.
  - When we have a bunch of photographs from a party, we can try to extract **Group Dynamics** data by building a graph of people taking pictures with each other.

Digitalization and Digital Transformation

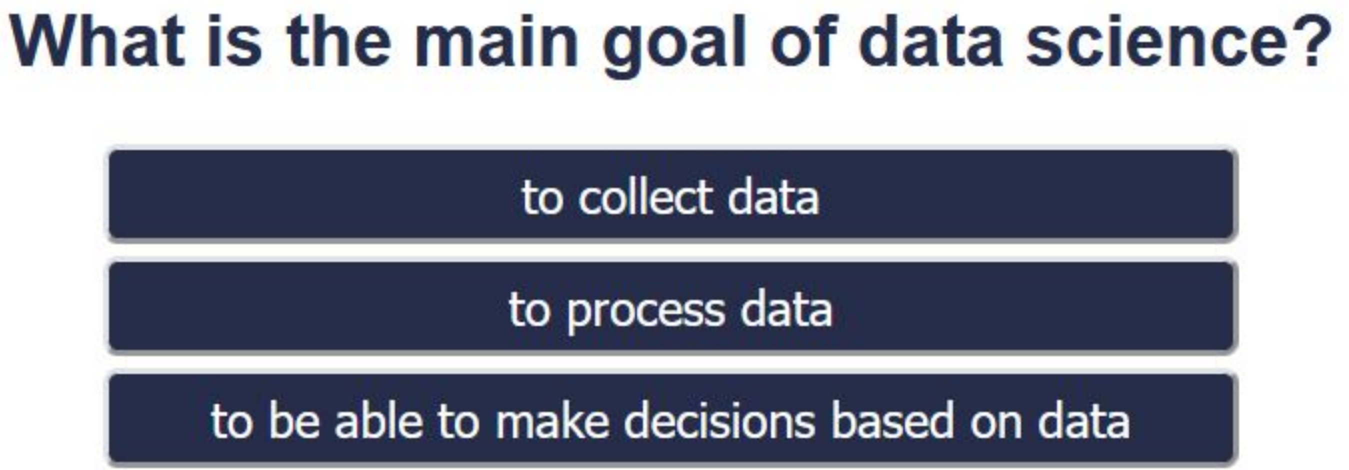
- To apply data science principles to running a business, one first needs to collect some data, i.e. translate business processes into digital form. This is known as **digitalization**.
- Applying data science techniques to this data to guide decisions can lead to significant increases in productivity (or even business pivot), called **digital transformation**.
- Let's consider an example. Suppose we have a data science course (like this one) which we deliver online to students, and we want to use data science to improve it. How can we do it?

Which of the following is an example of non-structured data?

List of students in class

Collection of student essays

Graph of friends of social network users



What is the main goal of data science?

to collect data

to process data

to be able to make decisions based on data