# **Introduction to Data Science (DS1101)**

# ***Lecturer: Dr. UA Piumi Ishanka***

## LOS

* Explain relevant data science theories and concepts
* Design and implement an experiment incorporating data science principles

## Evaluation

* Quizes (04) - 20
* Assignment (01) - 10 -(replaces with)> Mid-Exam - 20
* Final Exam - 70

## What is Data Science?

Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education

* Compute Science
  + Pattern Recognition, visualization, data warehousing, High performance computing, Databases, AI
* Mathematics
  + Mathematical Modeling
* Statistics
  + Statistical and Stochastic modeling, Probability.

Definitions:

* **Data Science** is a set of fundamental principles, processes and techniques that guide the extraction of knowledge from data with the goal of improving decision-making
* It is an interdisciplinary academic field that is bases on:
  + Mathematics
  + Statistics
  + Machine learning and Artificial Intelligence
  + Specialized Programming
* **Data mining** is the extraction of knowledge from data, via phonolites that incorporate data science principles

## Types of Data Science

## Data-driven decision-making

Data-driven decision-making (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on intuition.

Some decision can be made automatically (finance, recommendations)

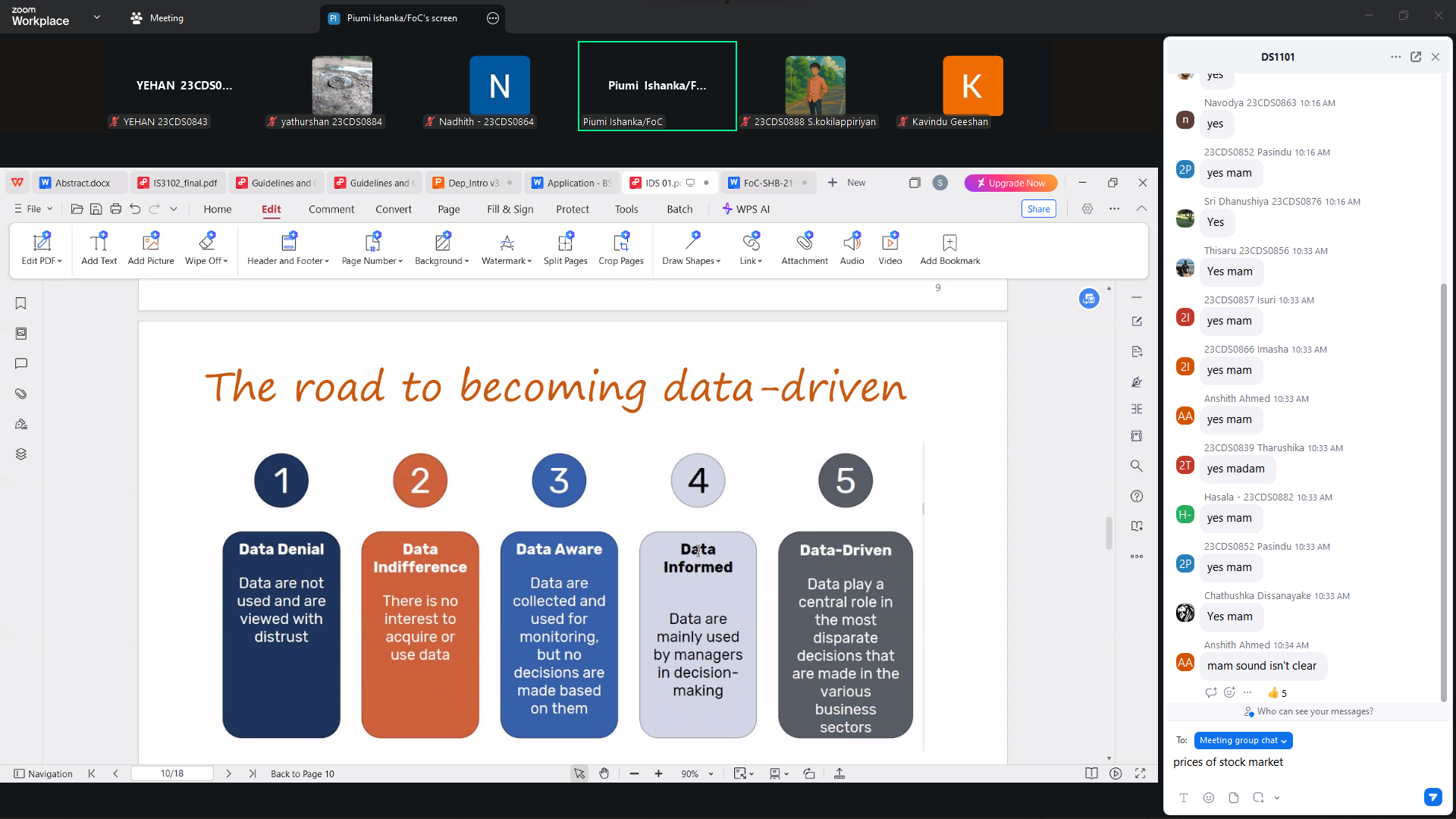
Data engineering and processing support many data oriented business tasks but do not necessarily involve extracting knowledge or data-driven decision making

Data, and the capability to extract useful knowledge from data, should be regarded as key strategic asset

* Need to invest to acquire the right data (even lose money)
* Understand data science, even if you will not do it

## 

## The road to becoming data-drive



1 - DATA DENIAL: Data are not used and are viewed with distrust

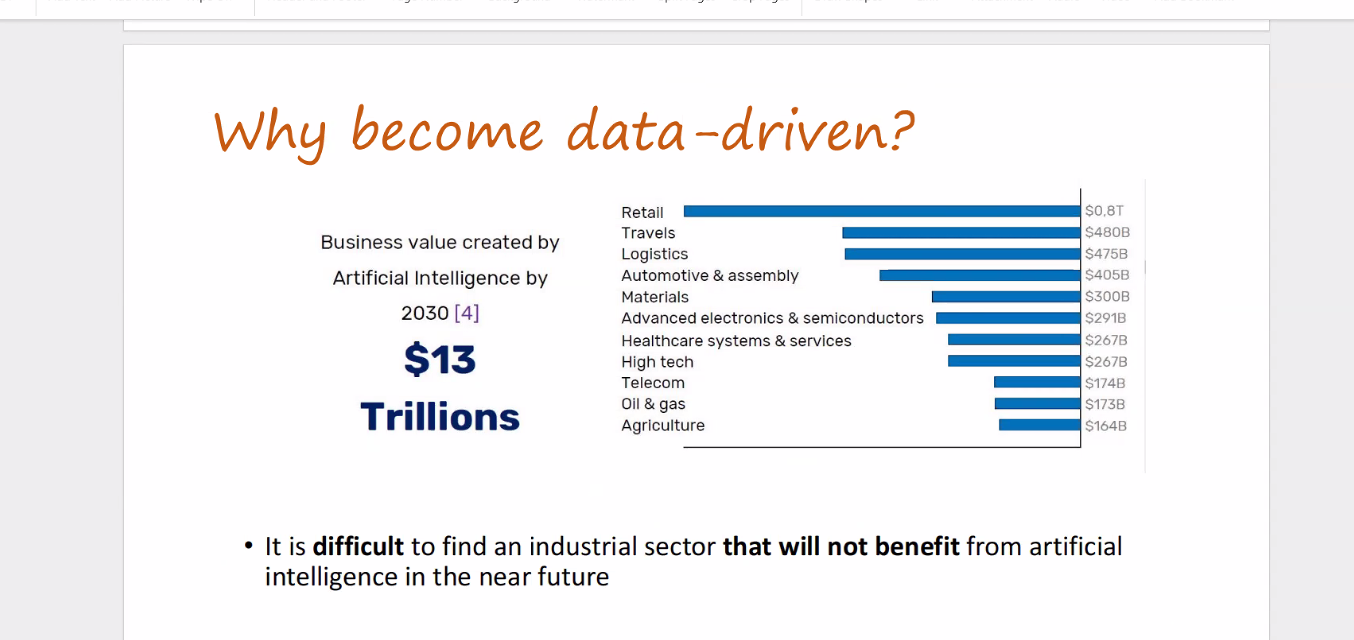
2 - DATA INDIFFERENCE - There is no interest to acquire or use data

3 - DATA AWARE- Data are collected and used for monitoring, but no decisions are made based on them

4 - DATA INFORMED - Data are mainlyused by managers in decision-making

5 - DATA DRIVEN - Data play a central role in the most disparate decisions that are madein the various business sectors

## Why become data driven?



It is **difficult** to find an industrial sector **that will not benefit** from artificial intelligence in the near future

What is data, talk about benefits of analyzing data, (ex was hate speech from social media, covid19 vacciness results)

## What are data?

We refer to data as any piece of information that has been collected and stored in a computer.

Examples:

- Sensor measurements

- Customer information

- Transaction history

- Social media posts

## 

## Types of Data: Structured vs Unstructured

### **Structured data**

### 

### Data that are organized following a predefined scheme and stored in tabular formats(Excel sheets, SQL databases…)

### 

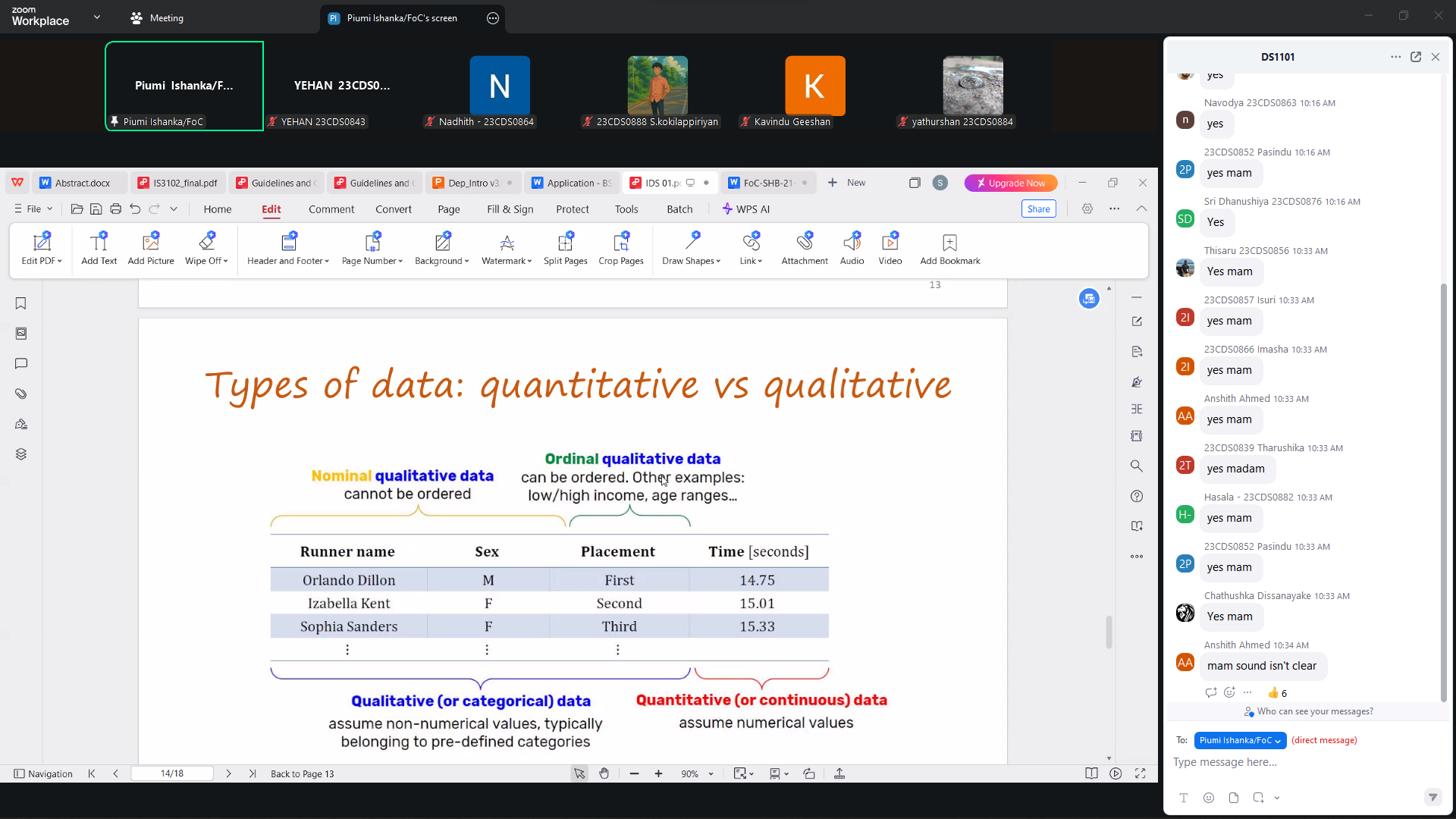
### **Unstructured data**

### 

* Data that can have an internal structure but do not follow a predefined data model or scheme

### 

## Types of Data: quantitative vs qualitative

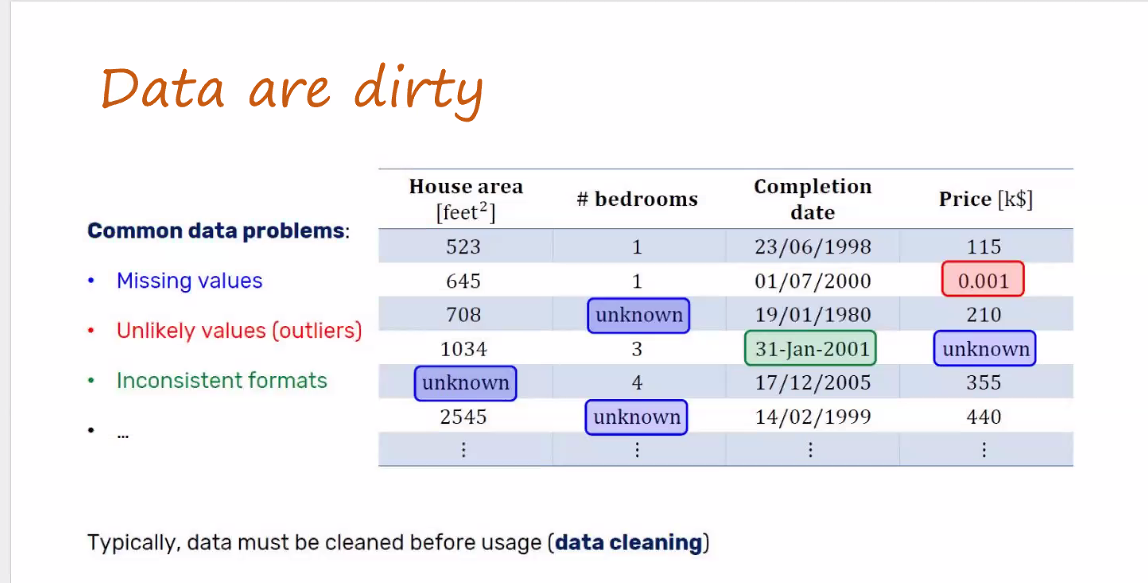


## 

## Types of Data

* Relational Data (tables/transaction/legacy data)
* Text data (web)
* Semi-structured data (XML)
* Graph Data
* Social Network, Semantic Web(RDF)
* Streaming Data
* You can afford to scan the data once

## Data are dirty



## Common data problems:

* Missing values
* Unlikely values (outliers)
* Inconsistent formats

Typically, data must be cleaned before usage (**Data Cleaning**)

## References C. Shah, A Hands-On Introduction to Data Science, 1st edition. 2020 -

## [Download link (epub)](https://archive.org/download/a-hands-on-introduction-to-data-science-chirag-shah/hands-introduction-data-science-chirag-shah.epub)

## [Download link 1 (pdf)](http://103.203.175.90:81/fdScript/RootOfEBooks/E%20Book%20collection%20-%202023%20-%20G/CSE%20%20IT%20AIDS%20ML/chirag-shah-a-hands-on-introduction-to-data-science-2020.pdf) [Download link 2 (pdf)](https://archive.org/download/a-hands-on-introduction-to-data-science-chirag-shah/hands-introduction-data-science-chirag-shah.pdf)

## 