Week 2 Class Notes

Objective: This document includes class notes: **summaries** of everything we learned, as well as **questions and answers** during the sessions, as well as any **resources** and links shared by students or otherwise. Also **assignments** and due dates. Other stuff might be included as well.

# Session 1

*Data science* is the field of study that combines:

1. domain expertise
2. programming skills
3. mathematics and statistics

to extract meaningful insights from data .

Before learning from data, programmers used to write algorithms to capture a long list of hand-crafted rules made by experts in their respective fields. Examples:

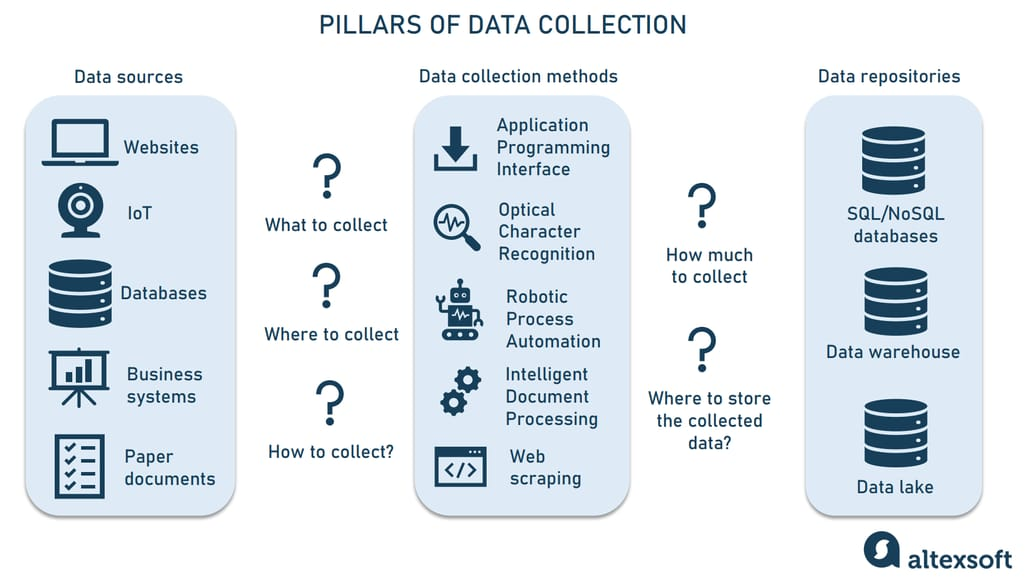
* Medical diagnosis: Systems would use a series of if-then rules based on symptoms, medical history, and test results to suggest possible diagnoses
* Robotics: Robot control systems were programmed with specific movements and actions based on sensor inputs

Problem: systems were often inflexible, and as problems became more complex, the number of rules grew exponentially, making maintenance difficult. Furthermore, expert knowledge wasn’t enough to capture all possible scenarios.

*Machine Learning* automated algorithm writing, which is especially useful when we have:

* **too many patterns**: too many cases for the algorithm to cover
* **too complex patterns**: impossible to write an algorithm for
* **changing patterns**: too costly, or impossible, to keep rewriting the algorithm

**Pillars of Data Collection**



Source: <https://www.altexsoft.com/blog/data-collection-machine-learning/>

**5 Characteristics of Data Quality**

1. **Accuracy:** Is the information correct in every detail?
2. **Completeness:** How comprehensive is the information?
3. **Reliability:** Does the information contradict other trusted resources?
4. **Relevance:** Do you really need this information?
5. **Timeliness:** How up- to-date is the information? Can it be used for real-time reporting?

Source: <https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality>

How Machine Learning works?

1. **Define** the problem.
2. **Identify** data sources.
3. **Prepare** the data.
4. **Select** ML algorithm.
5. **Train** the model (run the algorithm on the data to produce a trained model).
6. **Test** the model on unseen data.
7. **Integrate** with other systems.

Surveys:

* Open-ended questions vs closed questions (yes-no or MCQ)

Experiments:

* Control
* Randomization
* Ethical concerns: e.g., studying smoking and pregnancy. We can use Observational Studies in this case.

**Feature Engineering (Notebook)**: <https://github.com/HassanAlgoz/data-analytics/blob/main/techniques/feature_engineering.ipynb>