# Stage 2. ML problem formulation – Data

**Supporting materials:**

- **Machine Learning: The Basics. Chapter 2.1.**
- **Lecture notes CS 329S Lecture 3, Sections “Sampling”, “Class imbalance".**
- **Lecture notes CS 329S Lecture 10, Section “Data distribution shift".**

Now we are continuing to develop our project by defining the first component of ML – Data.

**Introduction***

*You can copy-paste/ edit from “Machine Learning - when and why?” task based on received feedback. Alternatively, you can choose completely different problem.

**Problem Formulation**

Formalise the application an ML problem:

- Explain the source of the dataset. How did you collect (sample) data (you can make up some possible scenario for your type of data)? Is your sampling biased (e.g. due to non-probability sampling)? Discuss class (im)balance if applicable.
- Clearly explain the data points, features and labels of this ML problem. Indicate type of data (continuous variable, categorical or ordinal values etc.) and units of measurement when applicable.
- Explain your feature selection process (no theoretical justification needed).
- Do you need to continuously collect (update) data, or do you use static dataset?