Data Collection

1. In ML, data are first-class citizens?
   - Yes
   - No

   Just right! Data are entities that support all operations generally available to other entities.

2. A data pipeline is a series of data processing steps such as:
   - Data collection

     Data collection is the first step in building ML systems.
   - Data ingestion

     Data ingestion is the process of absorbing data from different sources and transferring it to a target site where it can be deposited and analyzed.
   - Data Analysis
   - Data Preparation

     Data Preparation consists of data formatting, engineering and feature extraction.

3. Is the Data pipeline vital for the success of the production ML system?
   - Yes
   - No

   It consists of the incredibly important steps to the production ML system success.

4. What do you apply to maximize predictive signals in your data?
   - Data coverage
   - Data formatting
   - Feature selection
   - Feature engineering

   Feature engineering is the process of using domain knowledge to extract features with high levels of predictive signal from raw data.

5. Your training data should reflect the diversity and cultural context of the people who will use it. What can be done to mitigate inherent biases in a given data set?

- Collect data from equal proportions from different user groups.
- Commit to fairness.
- Adapt to continuously changing data
- Engineer better features

Balanced sampling from different user groups helps avoid inherent biases.

6. More often than not, ML systems can fail the users it serves. In this context, what is representational harm?
   - The amplification or negative reflection of certain groups stereotypes.
   - Making predictions and decisions that preclude certain groups from accessing resources or opportunities.
   - Giving skewed outputs more frequently for certain groups of users
   - Inferring prejudicial links between certain demographic traits and user behaviours.

   This is a prototypical way an ML system may fail the users it serves.

7. Accurate labels are necessary to properly train supervised models. Many times, human subjects known as raters perform this labeling effort. What are the main categories of human raters? (check all that apply).
   - Generalists

     Good choice! Generalists usually come from crowdsourcing sites.
   - Subject matter experts

     A classical example is radiologists labeling medical images for automated diagnosis tools.
   - Your users

     Users can provide labels within your application. A classical example is photo tagging.
   - Loggers
   - Aggregators
   - Classifiers