Issues in Training Data

1. What formula represents a dataset shift?
    - Ptrain(y|x)≠Pserve(y|x) and Ptrain(x)=Pserve(x)
    - Ptrain(y|x)=Pserve(y|x) and Ptrain(x)≠Pserve(x)
    - $P_{train}(y,x) \neq P_{serve}(y,x)$

    The most generic case of distribution skews is when the joint distribution of inputs and outputs differs between training and serving.

2. What measure is typically used to determine the degree of data drift?
    - Chebyshev distance (L-infinity)
    - Euclidean distance (L2)
    - Manhattan distance (L1)
    - Hamming distance

    Chebyshev distance is defined as maxi(|xi−yi|)

3. Distribution skew occurs when the distribution of the training dataset is significantly different from the distribution of the serving dataset, and is typically caused by: (check all that apply).
    - Occurs when serving and training data don't conform to the same schema. For example, int32 != float.
    - There is different logic for generating features between training and serving. For example, if you apply some transformation only in one of the two code paths.
    - Trend, seasonality, changes in data over time.

      Data distributions between training and serving often change and so this is another case of distribution skew.
    - Faulty sampling method that selects a sample for training which is not representative of serving data distribution.

      A faulty sampling mechanism that chooses a non-representative subsample is an example of distribution skew.
    - Different data sources for training and serving data.

      Data sources between training and serving often change and so this is another case of distribution skew.
    - A data source that provides some feature values is modified between training and serving time.

4. TensorFlow Data Validation (TFDV) helps TFX users maintain the health of their ML pipelines. TFDV can analyze training and serves data to:
    - Infer a schema.

      Nice going! In short, schemas describe the expectations for "correct" data and can thus be used to detect errors in the data.
    - Deploy pipeline to a mobile application.
    - Detect data anomalies.

      TFDV can check your data for error in the aggregate across an entire dataset or by checking for errors on a per-example basis.
    - Compute descriptive statistics.

      TFDV goes beyond computing relevant statistics, it also has nice browser-based visualization tools.
    - Perform feature selection.
    - Perform feature engineering.