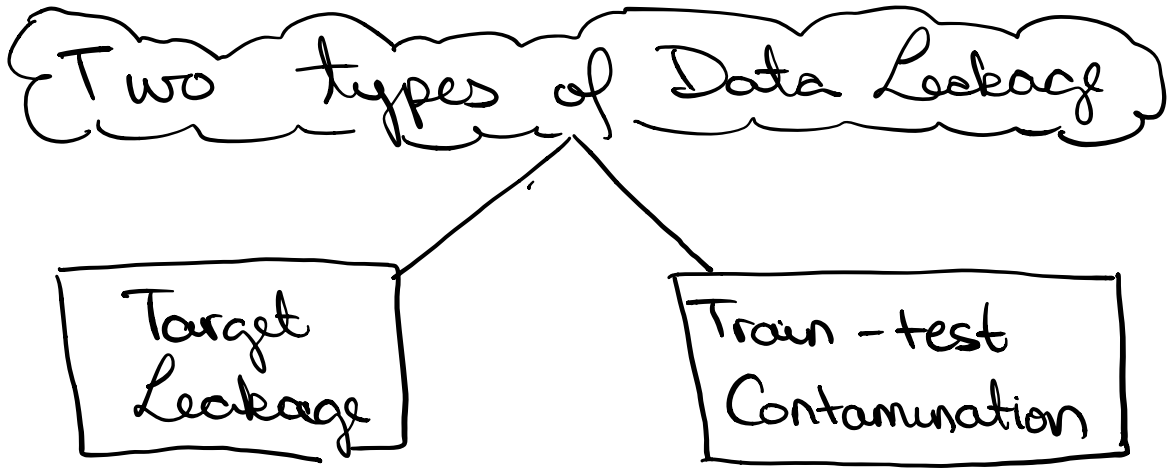
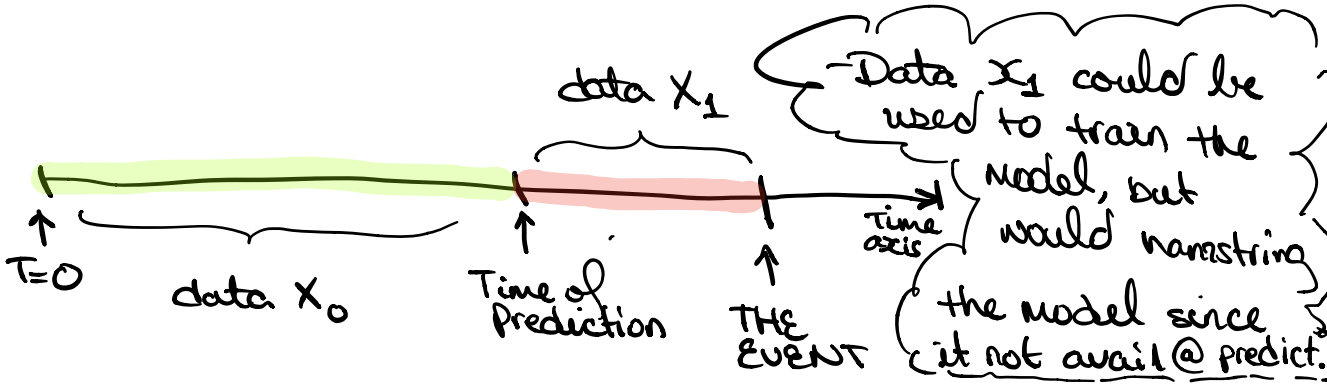
- Creating models based on unavailable or scarce data
  - ↳ may result in high model accuracy but model will not perform with the data begotten by the application of the model.



Target Leakage

- using data or updated data which will not be available at the time of prediction.
  - ↳ the data might not be absent. the data schema could be isomorphic — but the data in the sample may contain chronologically unavailable data in the timeline of the prediction.



Train-Test Contamination

- using data unavailable at the time of prediction. Particularly better generalizations — begotten by using test or validation data with the training data. Often this happens by feature engineering; for example, if you were to impute training values by considering the data in the validation/test/rest of the population.

  ↳ Is the answer, → not expediency, get more data with the categorical/numerical range you require, can't impute because while it provides the feature range, those feature sets aren't accompanied by their empirical associations.

  ↳ For example having the distribution or set of salaries without the features that cause that feature minimizes or removes one's ability to even draw conclusions about the gradient of salaries and the significance of anyone particular feature.

(*) remember validation data — is meant to measure the generality of your model

A good rule of thumb is to exclude valid/test data from any type of fitting — including the fitting of of preprocessing steps.

⇒ scikit-learn pipelines

this espec. important when leveraging x-validati. because the test-train contamination would be exacerbated by a similar magnitude.