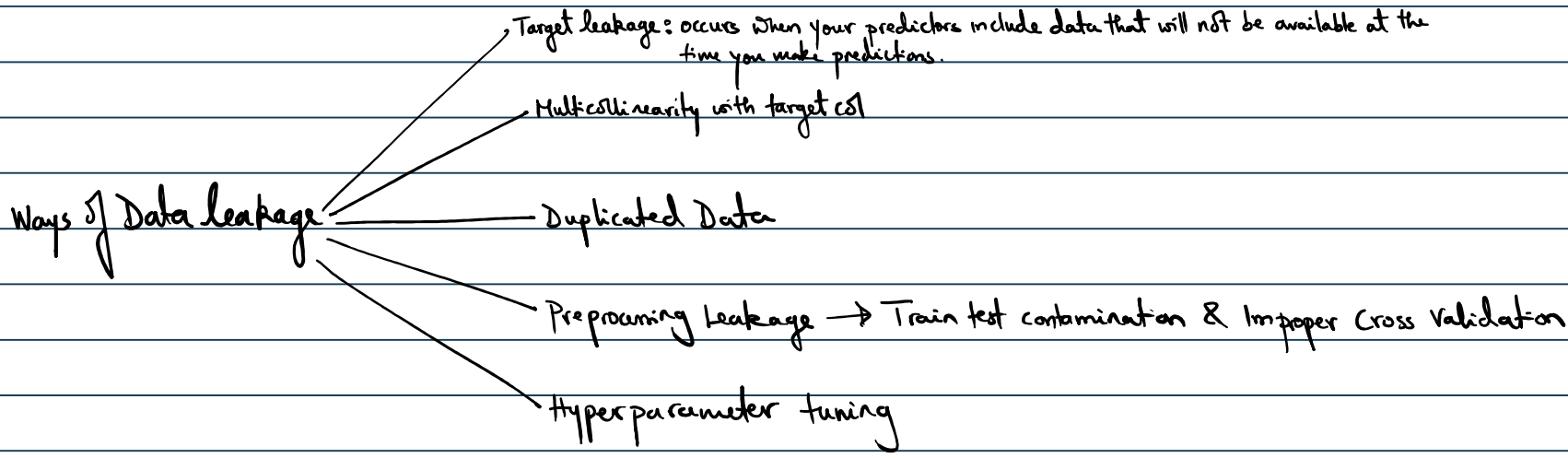
Data Leakage

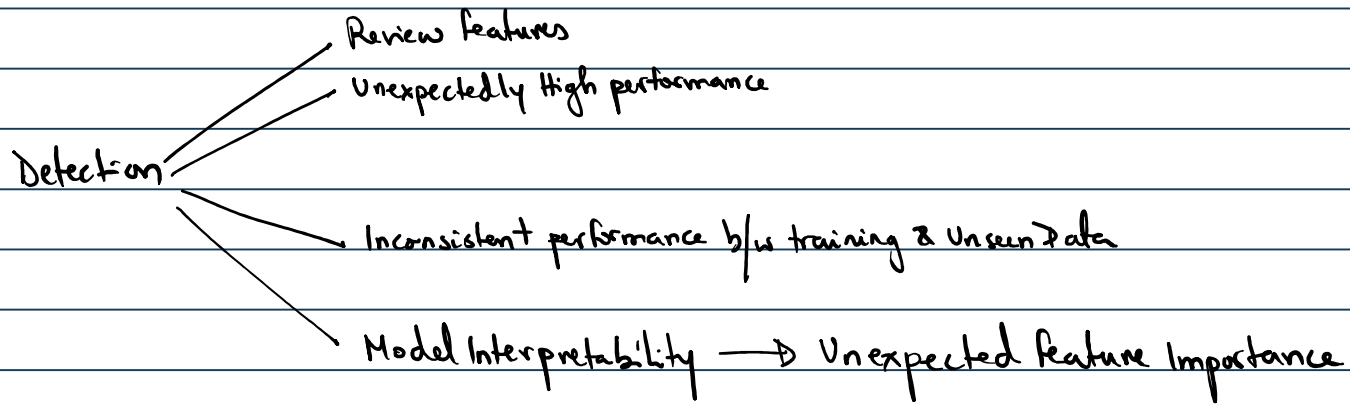
Sunday 28 April 2024 4:29 PM

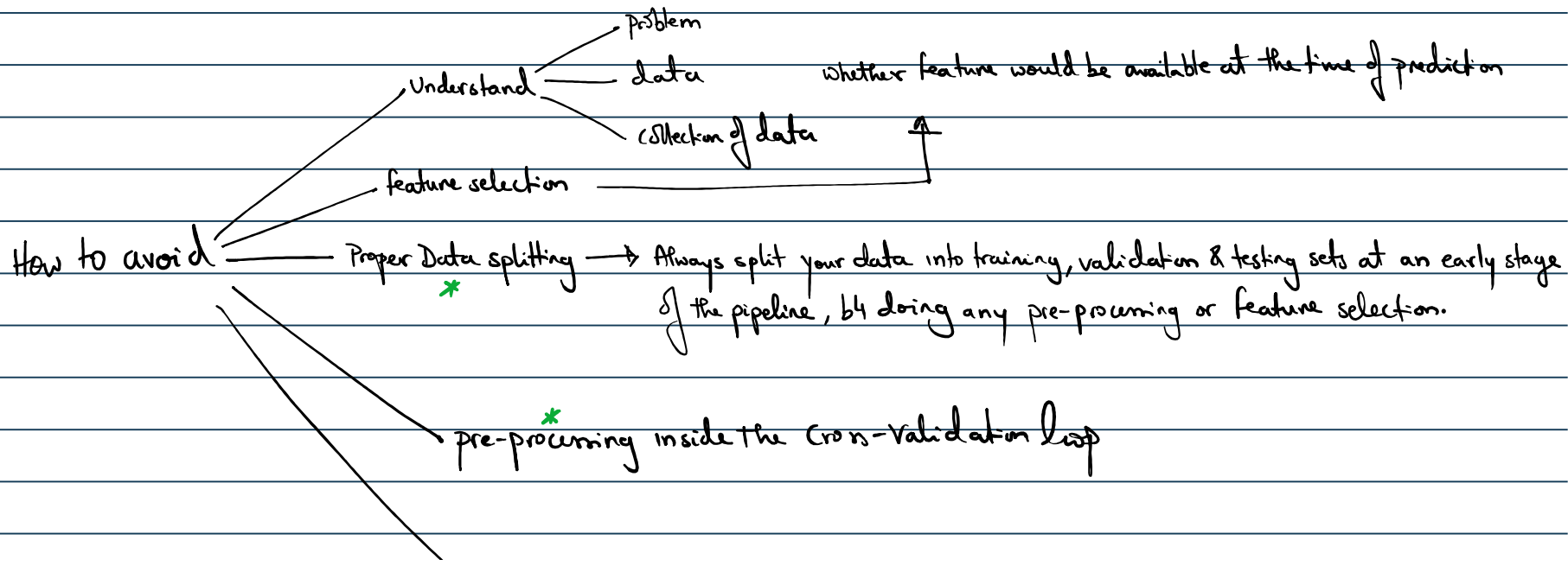
→ Refers to a problem where info from outside the training dataset is used to create the model.

→ It is information that the model wouldn't have access to when it's used for prediction in a real-world scenario.

→ This can lead to overly optimistic performance estimates during training & validation, as the model has access to extra information. However, when the model is deployed in a production environment, the additional information is no longer available, & the performance of the model can drop significantly.







of the pipeline, b4 doing any pre-processing or feature selection.

- pre-processing inside the Cross-Validation loop

- Avoid overlapping Data