University of Exeter

Intro to Machine Learning

Part 3 - The machine learning pipeline

Course contents

Session 1

- Slides: what is machine learning?
- Tutorial: linear regression
- Slides: model selection and evaluation

Session 2

- Tutorial: model selection and evaluation
- **Slides: the machine learning pipeline**
- Tutorial: machine learning pipeline task

Session 3

- Continue with machine learning pipeline task
- Tutorial: unsupervised learning

The machine learning pipeline

University of Exeter

What is it?

- Treating machine learning problems as a pipeline of discrete tasks

Why bother?

- Improves reliability, repeatability and reproducibility of research, by…
  - Making processing, training and analysis steps very clear
  - Enabling testing of modular components
  - Encouraging modularisation and testing of code
  - Encouraging model versioning and tracking
- Leads the way to automation, deployment, and MLOps

The machine learning pipeline

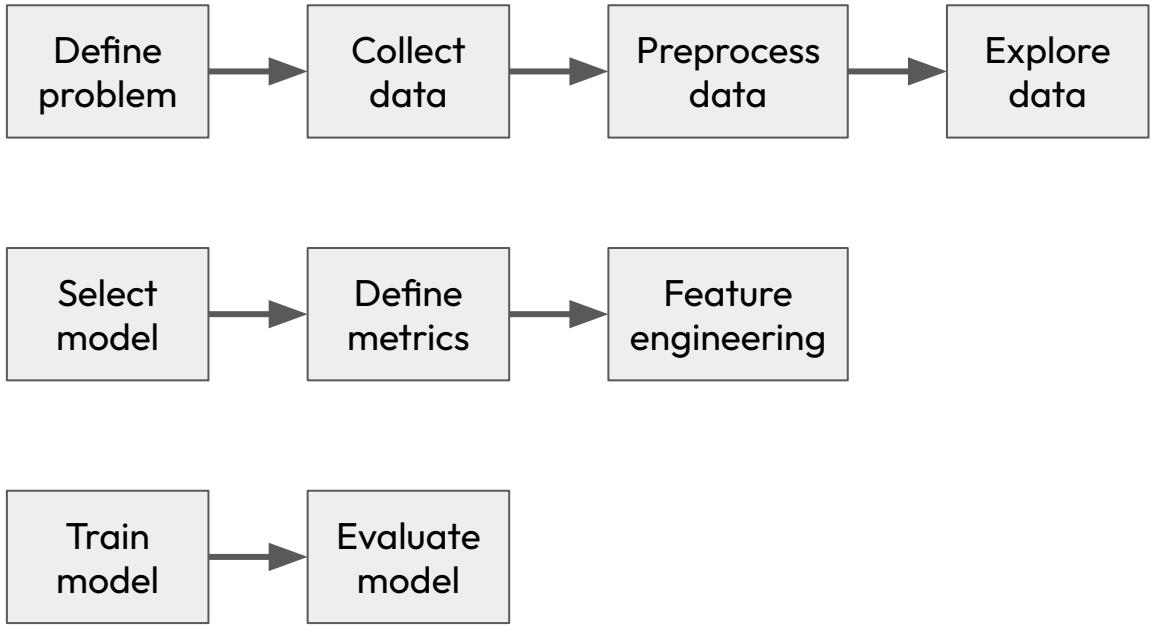
When to use it?

- On real problems, where you do not want to make mistakes

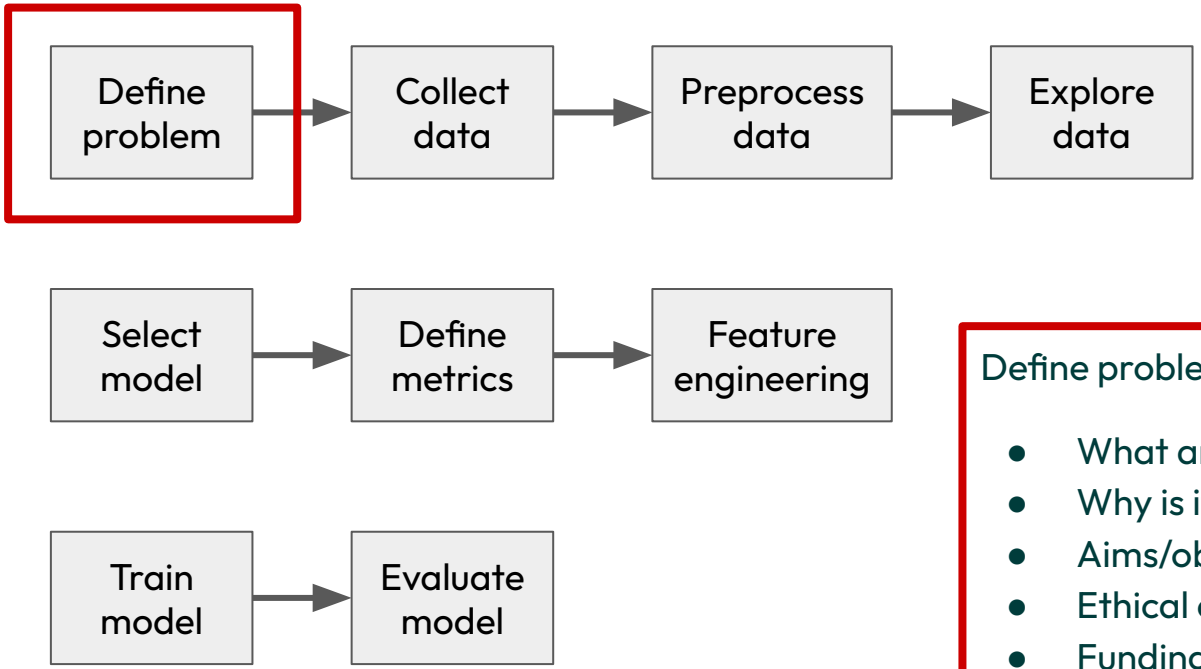
When not to use it?

- Potentially when you are playing around with a new technique or dataset
- However the pipeline stages are still useful to think about!

The machine learning pipeline



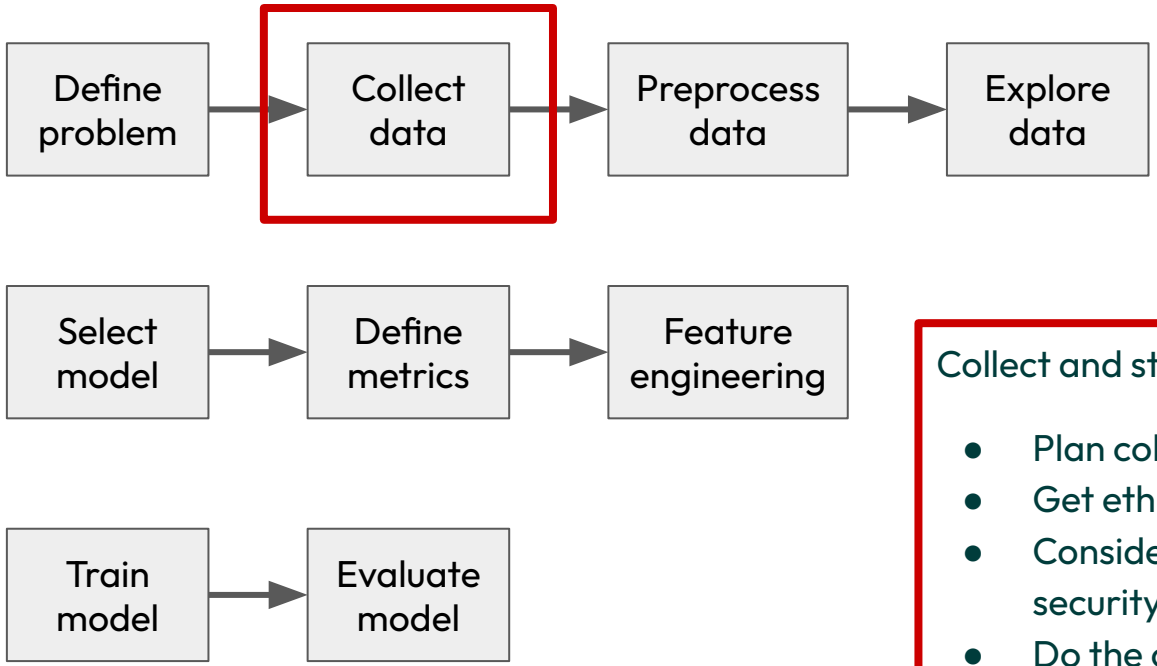
The machine learning pipeline



Define problem

- What are we doing?
- Why is it useful?
- Aims/objectives
- Ethical considerations
- Funding, business considerations

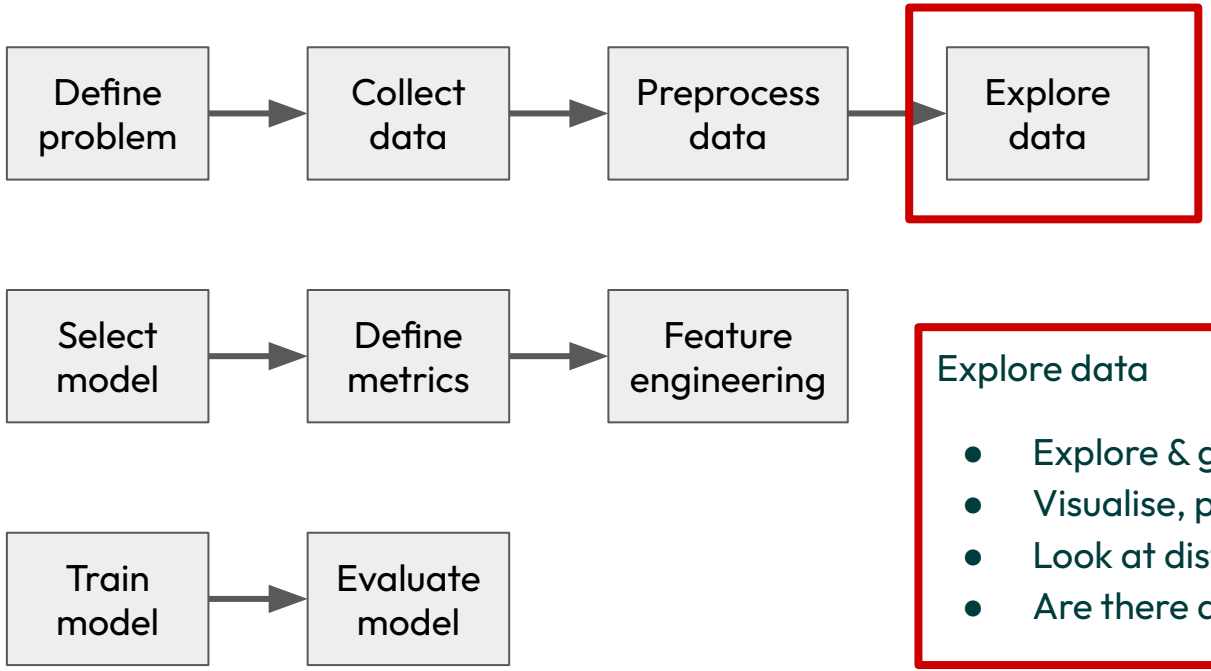
The machine learning pipeline



Collect and store data

- Plan collection details
- Get ethics approval
- Consider data management, storage, security, etc
- Do the collection, store, back it up

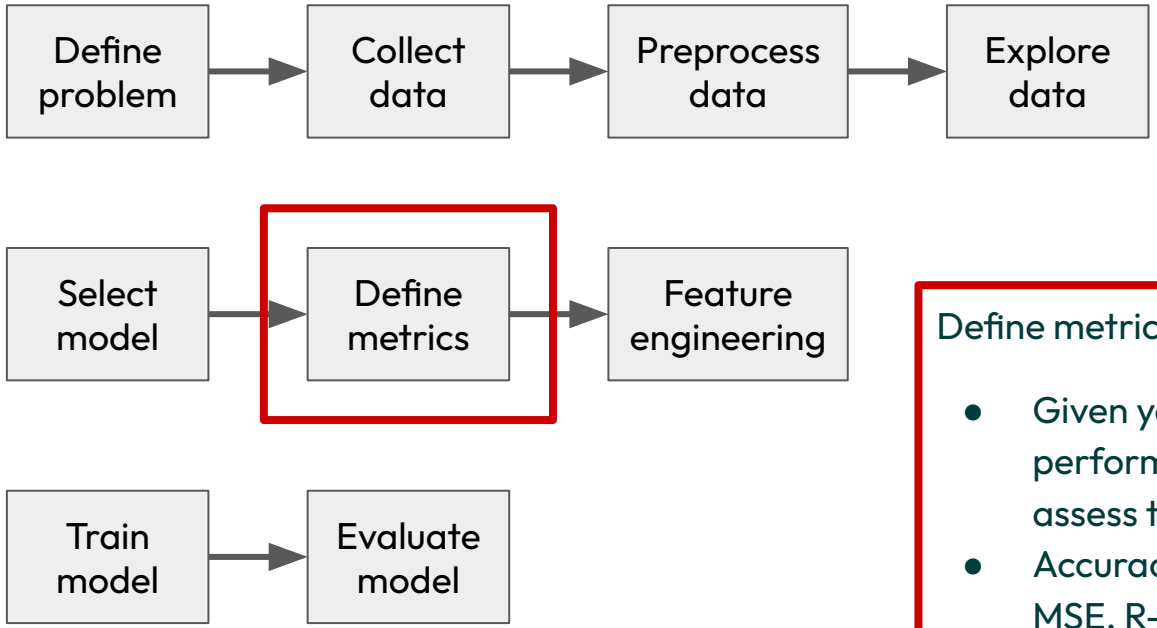
The machine learning pipeline



Explore data

- Explore & get to know your data
- Visualise, plot, transform, etc
- Look at distributions, bias
- Are there any obvious trends?

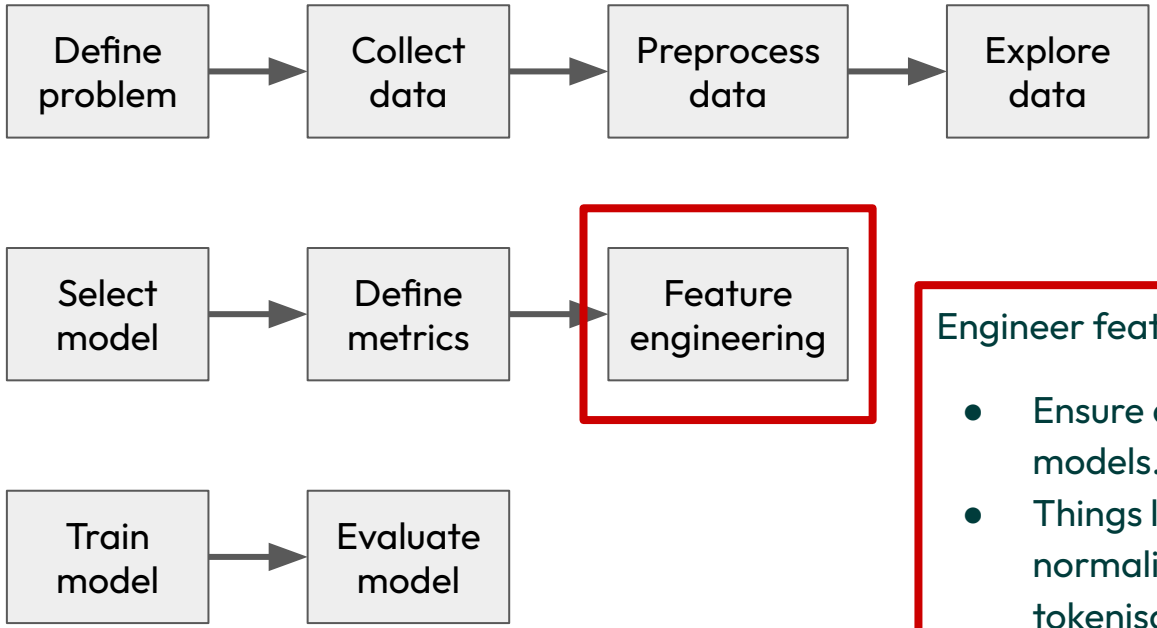
The machine learning pipeline



Define metrics and plan validation

- Given your initial chosen model, what performance metrics will you select to assess this model?
- Accuracy, f1-score, recall, precision, RMSE, MSE, R-squared etc
- Plan validation strategy here: i.e. cross validation, hyperparameter tuning

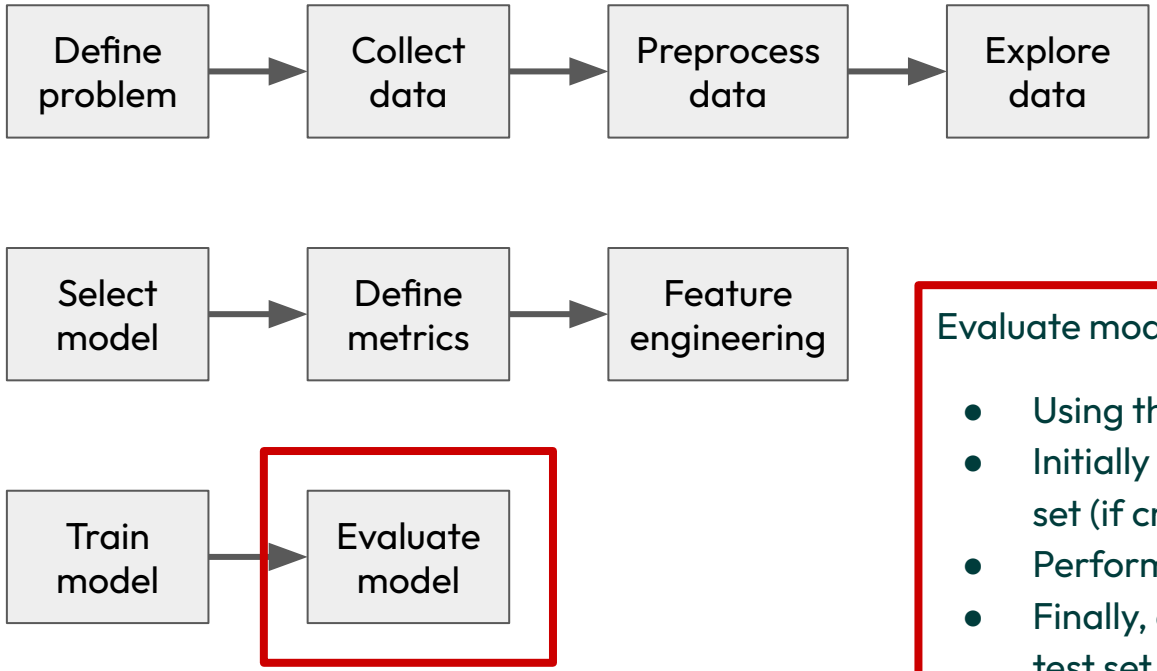
The machine learning pipeline



Engineer features

- Ensure data is encoded correctly for models.
- Things like one-hot encoding, scaling, normalisation, outliers, vectorisation, tokenisation, image processing, etc.
- Basically get the data ready for training.

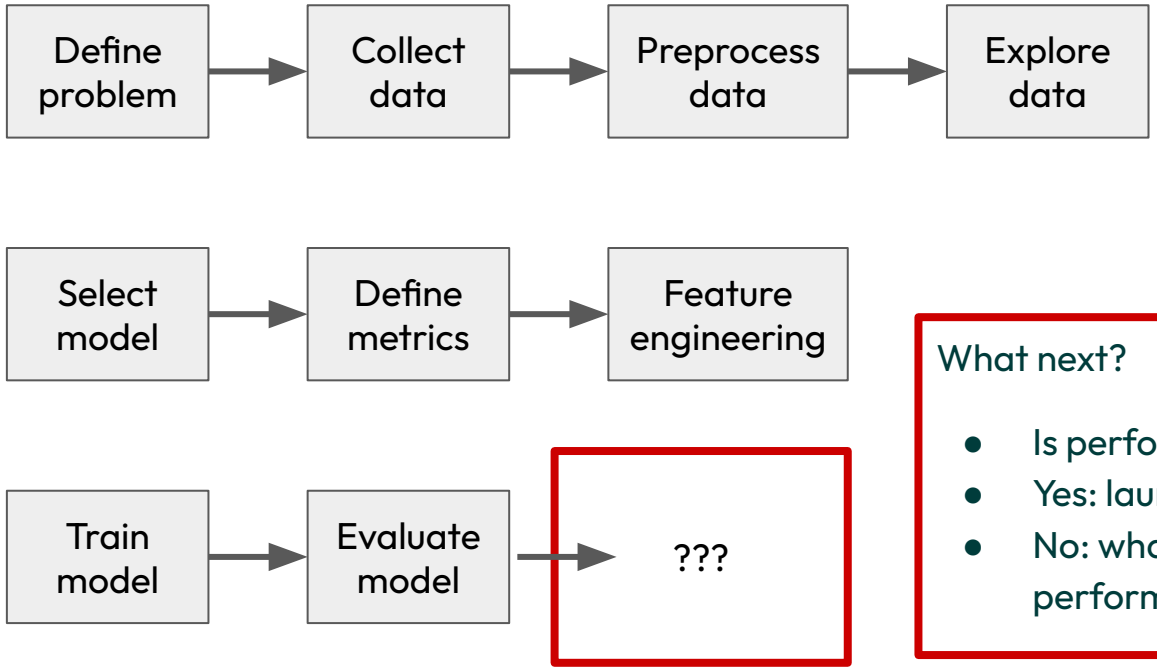
The machine learning pipeline



Evaluate model performance

- Using the metrics decided earlier
- Initially validate on your held out validation set (if cross validating)
- Perform fine tuning
- Finally, evaluate the model on the held out test set.

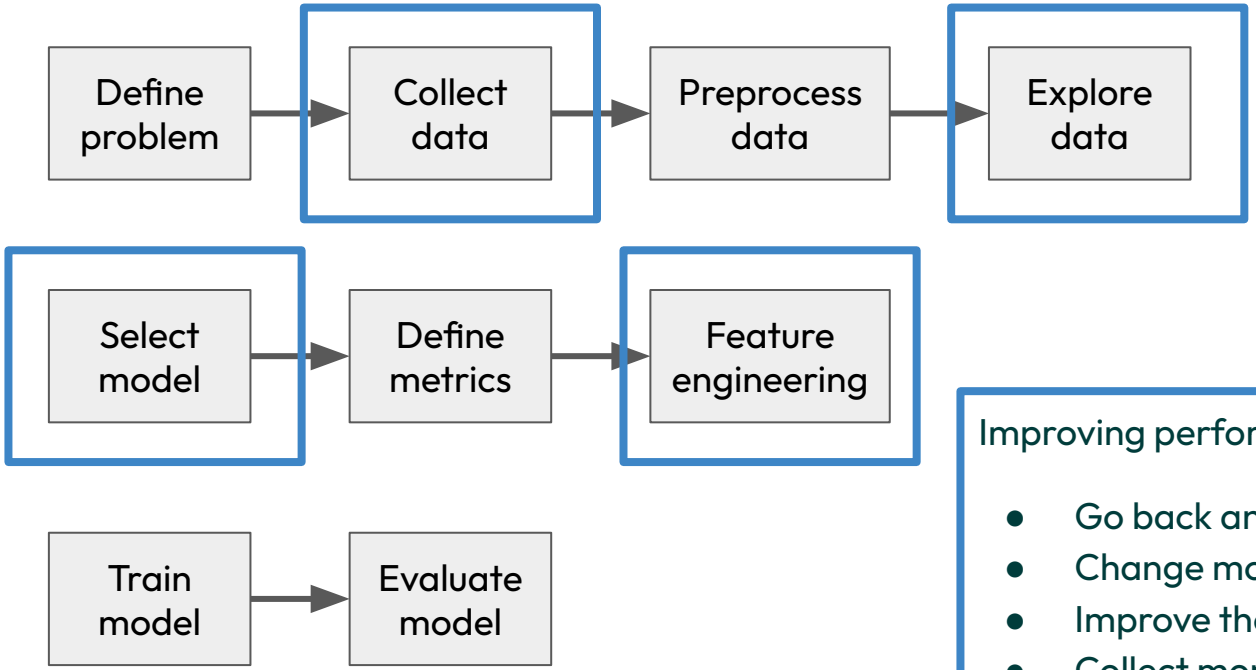
The machine learning pipeline



What next?

- Is performance good enough?
- Yes: launch model, product, research
- No: what can we do to improve performance?

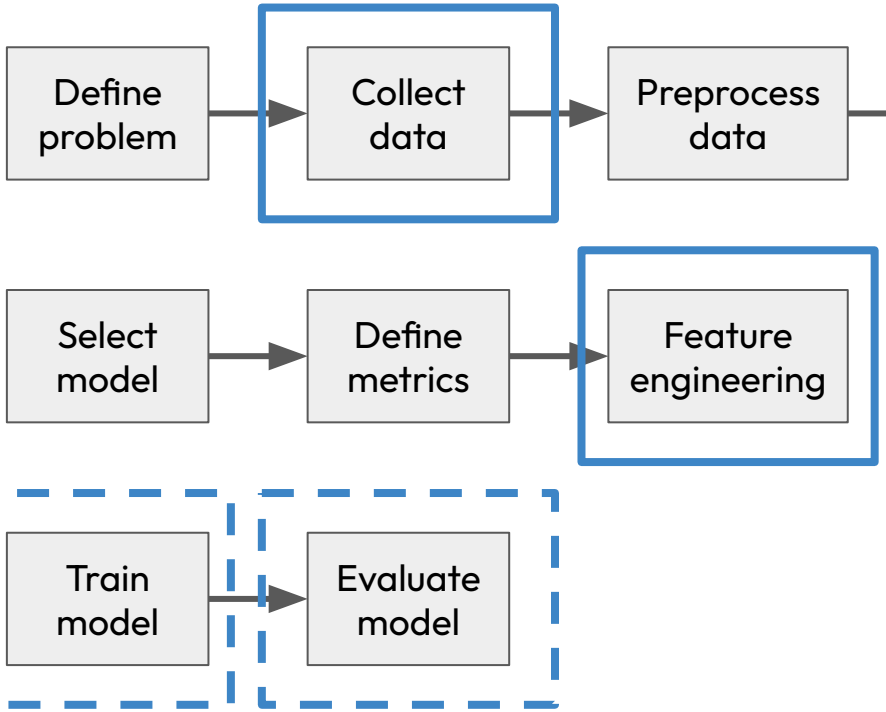
Improving model performance



Improving performance

- Go back and explore the raw data further
- Change model/approach
- Improve the feature engineering
- Collect more data

Where is most effort spent?



High effort:

- Data collection, feature engineering
- Maybe evaluation/fine tuning

Then:

- Data exploration/understanding problem
- Model training might also be hard, if high compute or RAM requirements.