Training, validation, and test sets

Dataset

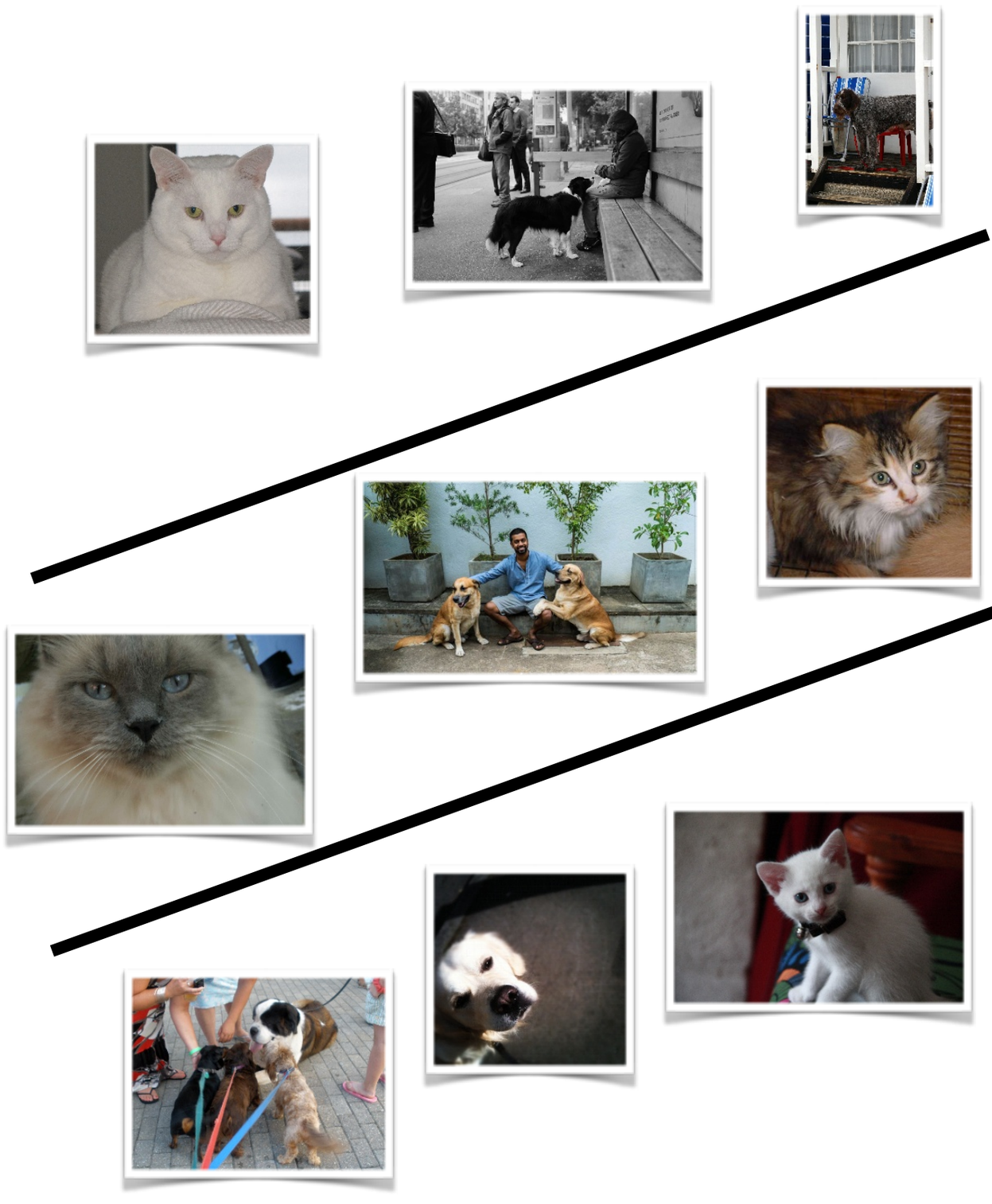
- Training set
  - Learn model parameters
- Validation set
  - Learn hyper-parameters
- Test set
  - Measure generalization performance





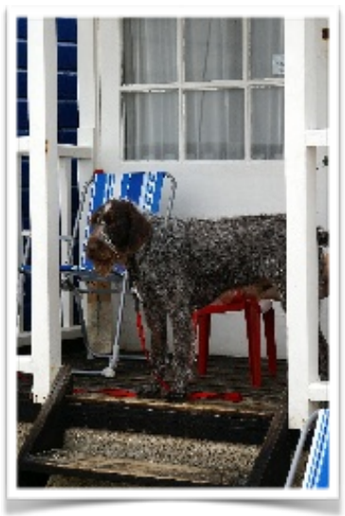

Why split the data?

- Overfitting
- Goal: Learn a model that works well in the real world
- Optimization objective: Learn a model that works well in training data



Training set

- Used to train all parameters of the model
- Model will work very well on training set
- Size: 60-80% of data



Validation set

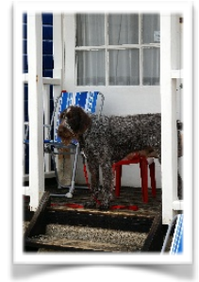
- Used to determine how well the model works
- Used to tune model and hyper-parameters
- Size: 10-20% of data

Testing set

- Used to measure performance of model on unseen data
- Used exactly once
- Size: 10-20% of data

How to split the data?

- Random sampling without replacement

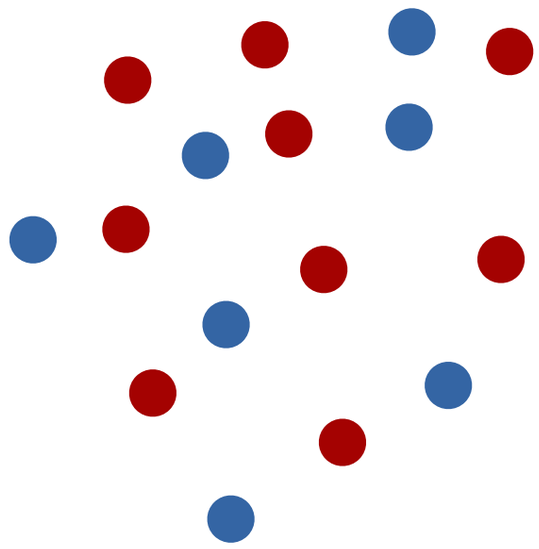







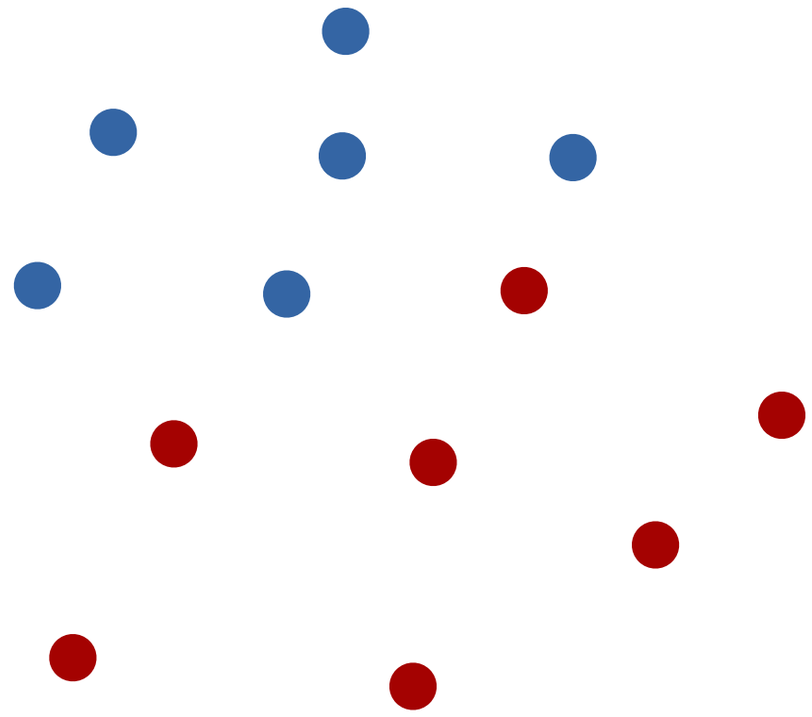

Distribution of data

Low dimensions



$$D_{data} \approx D_{train} \approx D_{valid} \approx D_{test}$$

High dimensions



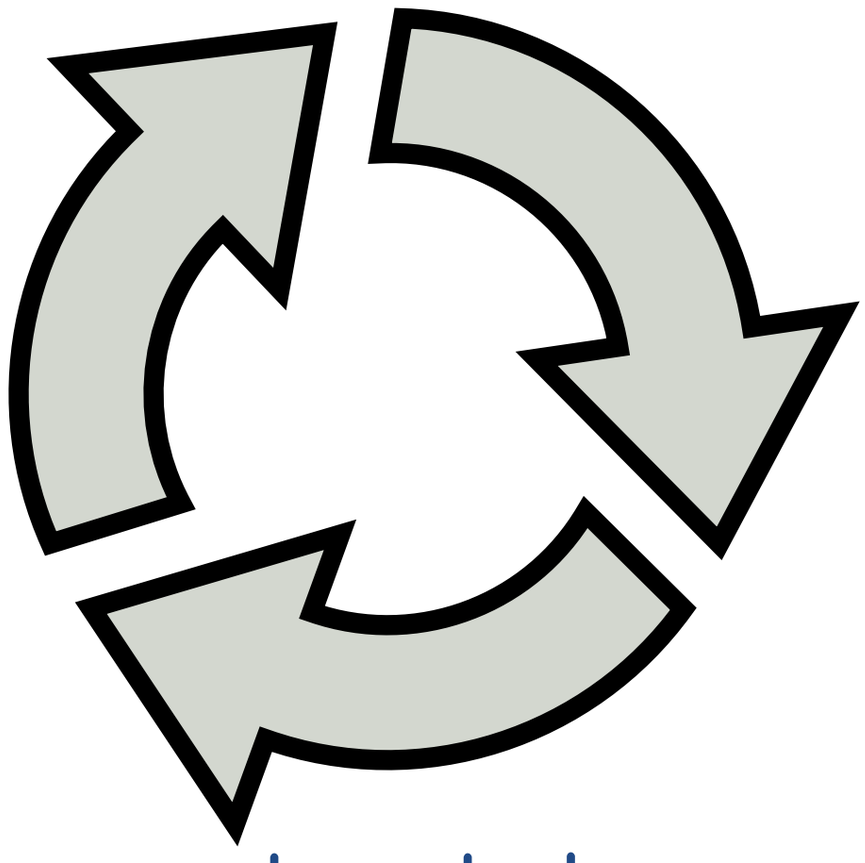
$$D_{data} \neq D_{train} \neq D_{valid} \neq D_{test}$$

Graduate student descent

Look at your
data / model output

semi-
automated

manual



Evaluate
your model on
validation set

Design and
train your model

automated