University of
South Australia

COMP 2019

Week 7
ML Training

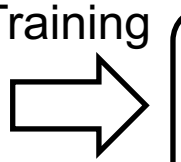
Learning Objectives

- Explain the machine learning process (CO3)
- Explain how data is prepared (CO3)
- Explain how ML models are trained (CO3)
- Explain how ML models are evaluated (CO3)

Supervised Learning from Data

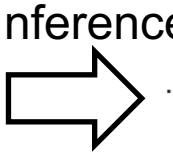
x1	xn	y

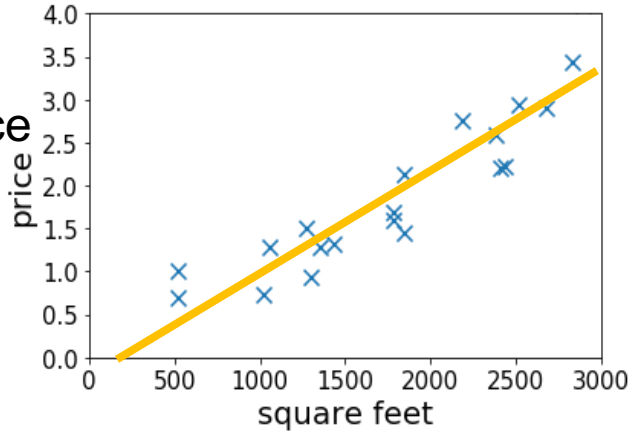
Training



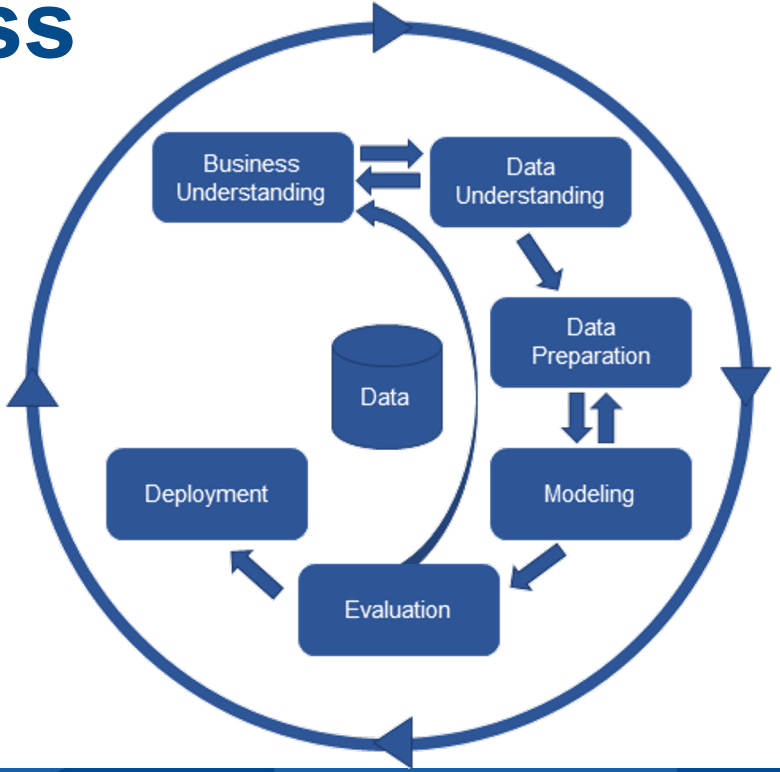
Model
y = f(X)

Inference





ML Process



https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

Understanding Data

- Key attribute distribution
- Label distribution
- Relationships between key attributes
- Attributes of important sub-populations
- Simple aggregation results
- Simple analysis of statistics

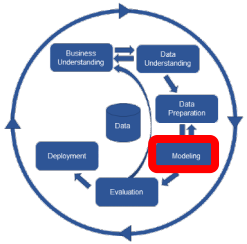
Examine the Quality of Data

- Is the data complete, covering all the required cases?
- Is the data correct?
  - How often do errors occur?
  - What is the nature of errors?
- Does the data contain missing values?
  - Where do they occur?
  - How they are they represented?
  - How frequent?
  - Systematic or random?

Preparing Data

- Cleaning the data
  - Rectify data quality issues
- Construct features
- Generate records
  - Negative cases may not be represented in the data
- Integration: combine data from multiple sources
- Aggregation

Data Sets



x1	…	xn	y

Never use the same data for training and testing

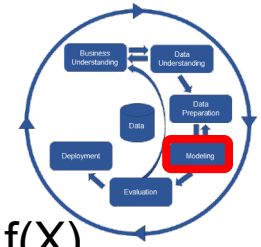
Training Dataset

x1	…	xn	y

Test Dataset

x1	…	xn	y

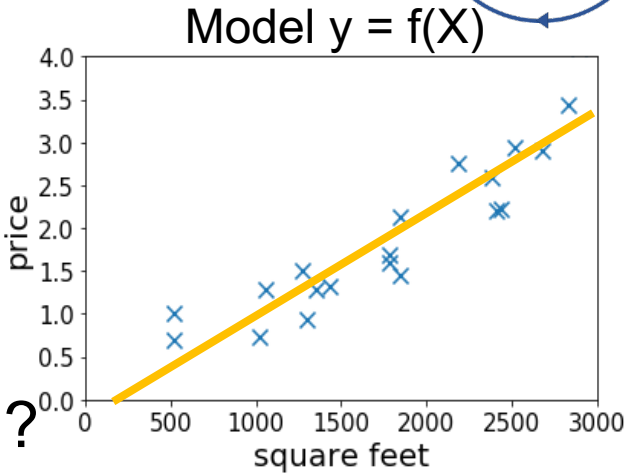
Model Fitting



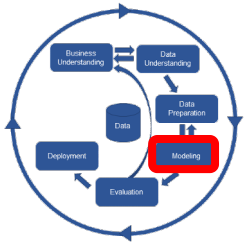
Training Dataset

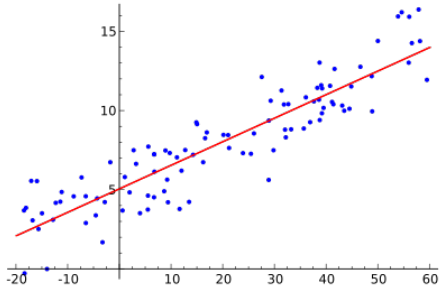
x1	…	xn	y

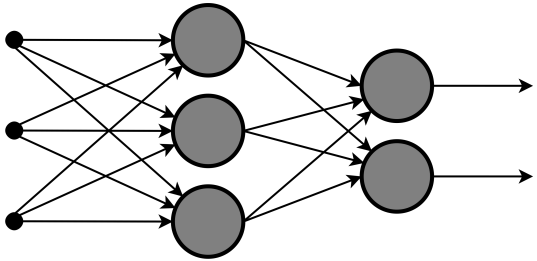
- Type of model?
- Optimisation goal?
- Learning algorithm?

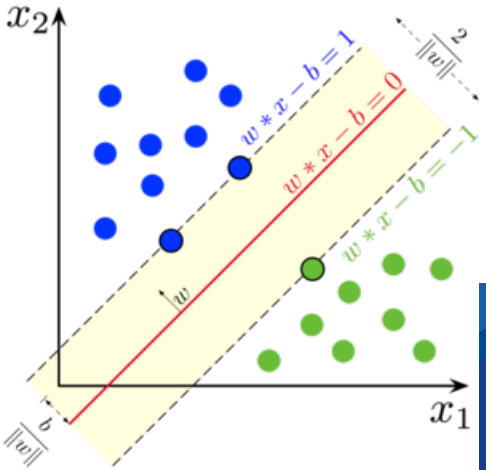


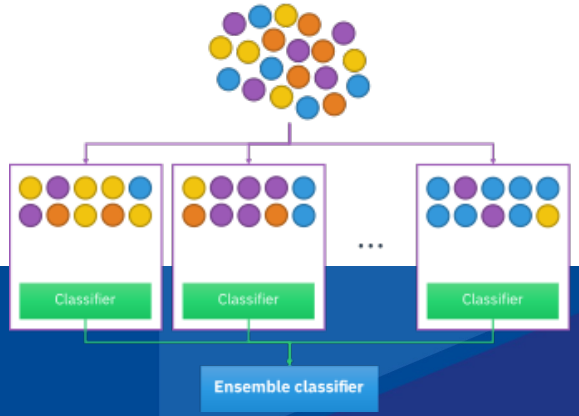
Types of Model

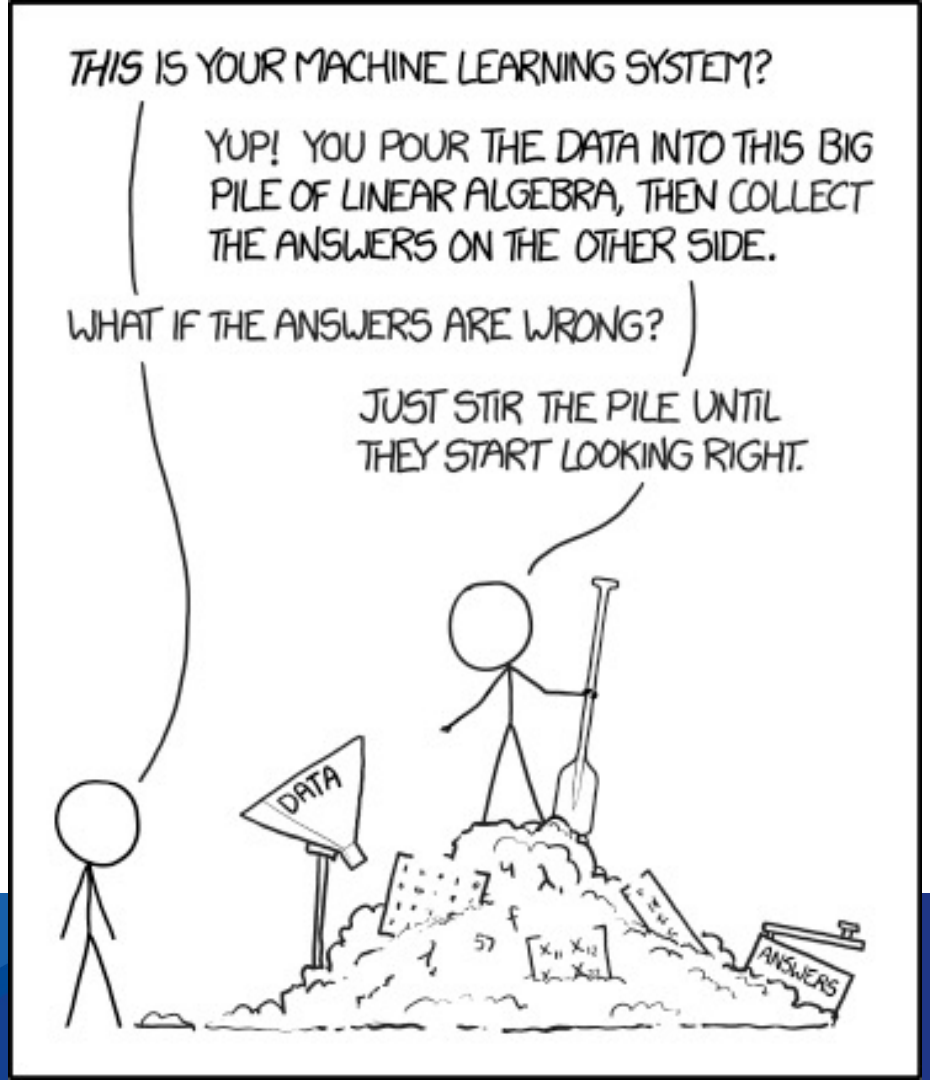




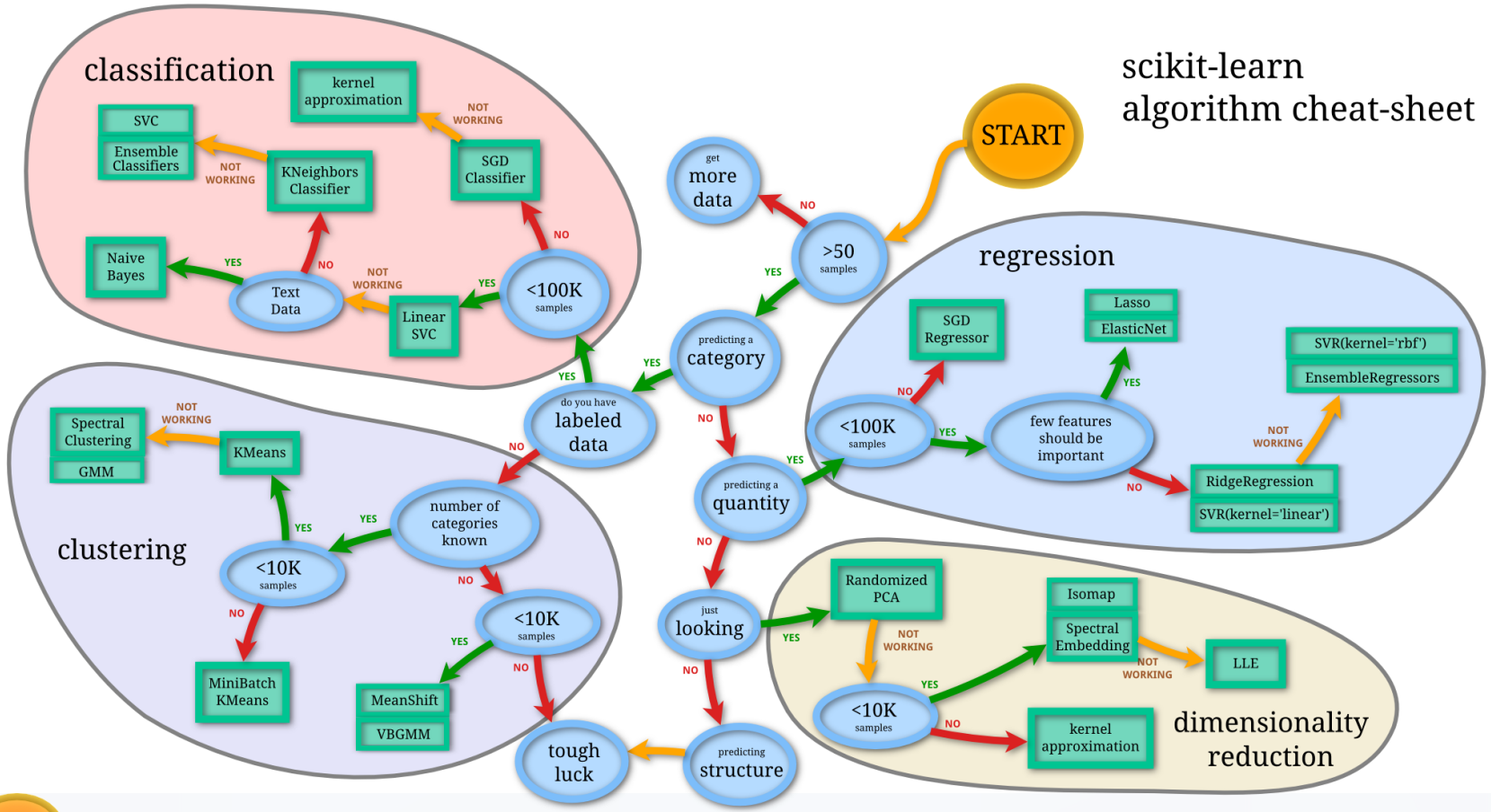






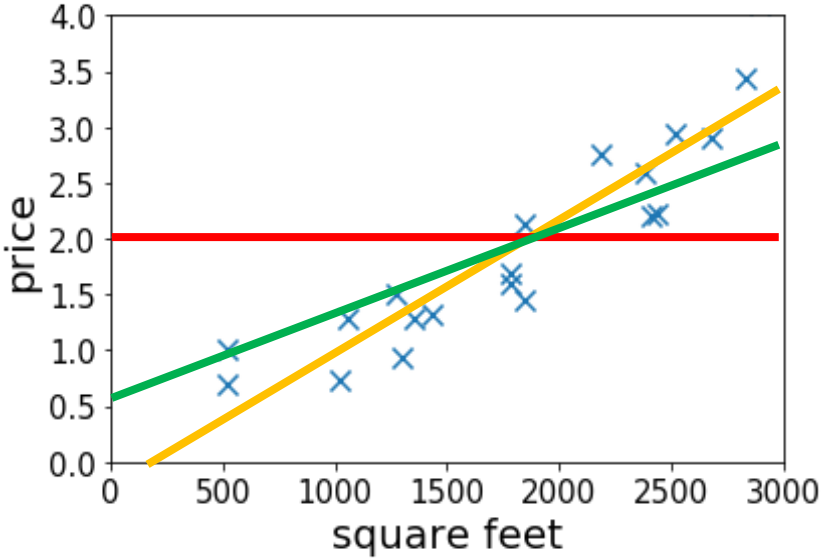


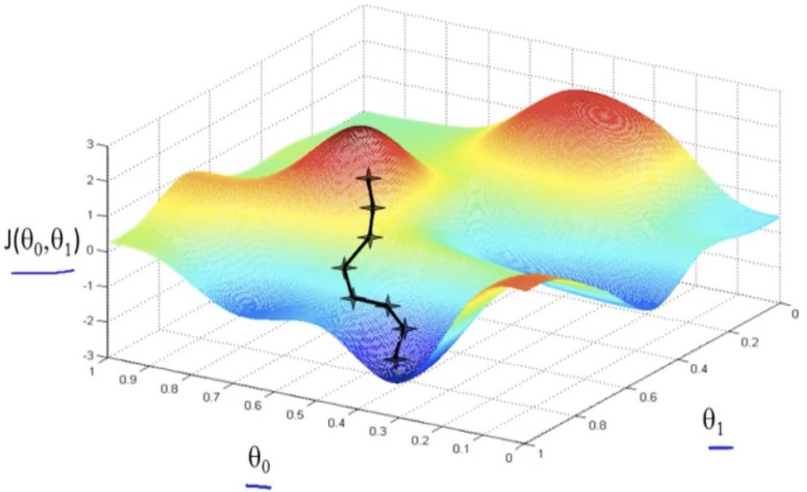
https://xkcd.com/1838/

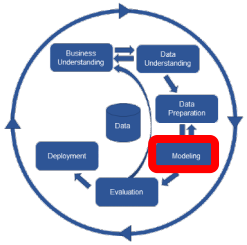

scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
just looking
YES
NO
predicting structure
tough luck
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
YES
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
regression
SGD Regressor
NO
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
YES
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
Back
scikit learn

Training as Optimisation

$$price = \theta_0 \times \text{sqFeet} + \theta_1$$

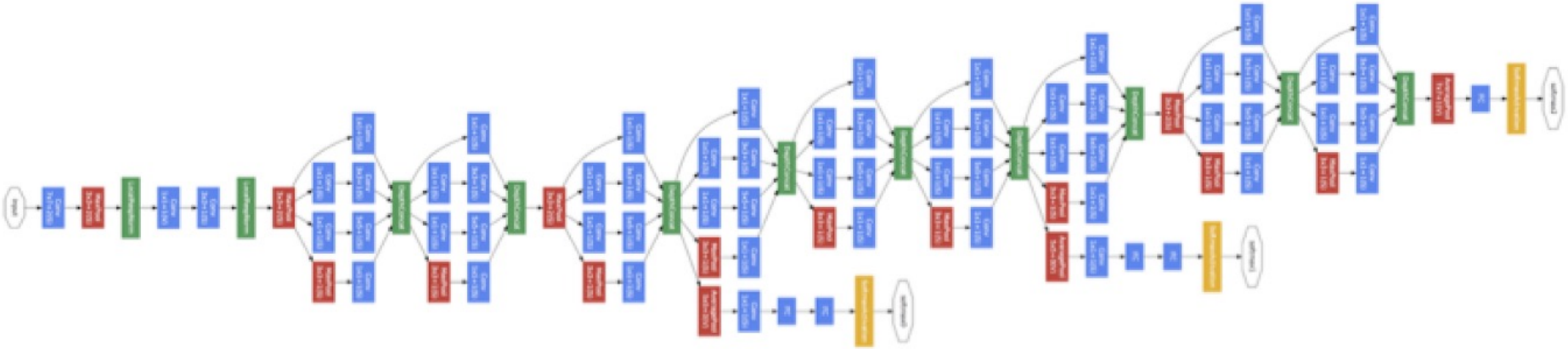






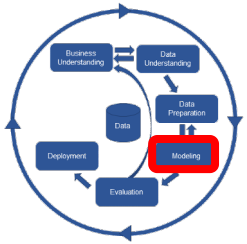
https://zitaoshen.rbind.io/project/optimization/1-min-of-machine-learning-gradient-decent/

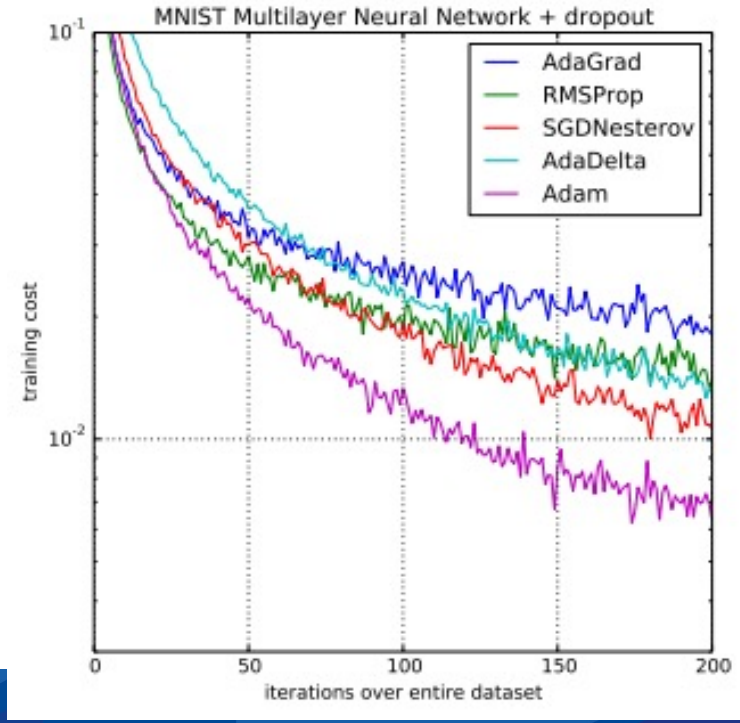
Deep Neural Net



GoogLeNet/InceptionV1 (2014) has 22 layers and 4 million parameters

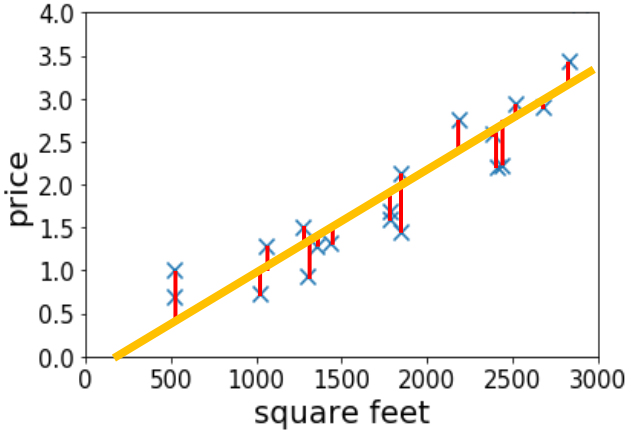
Optimisation Algorithm

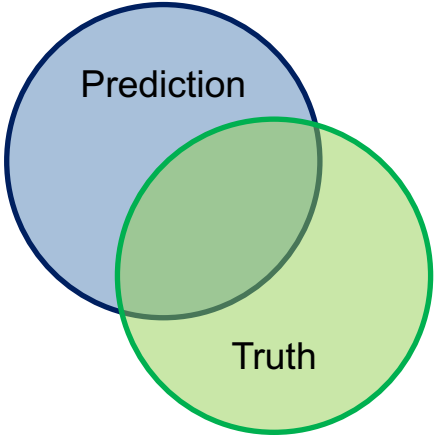




https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Metrics and Loss

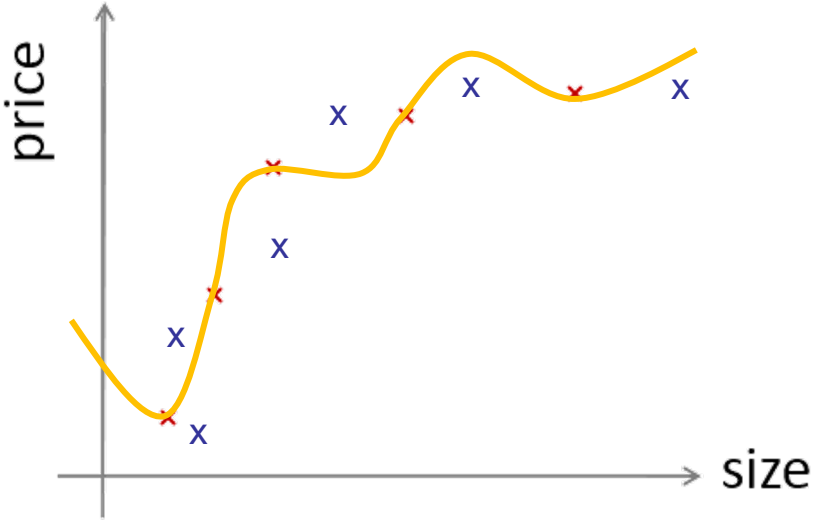


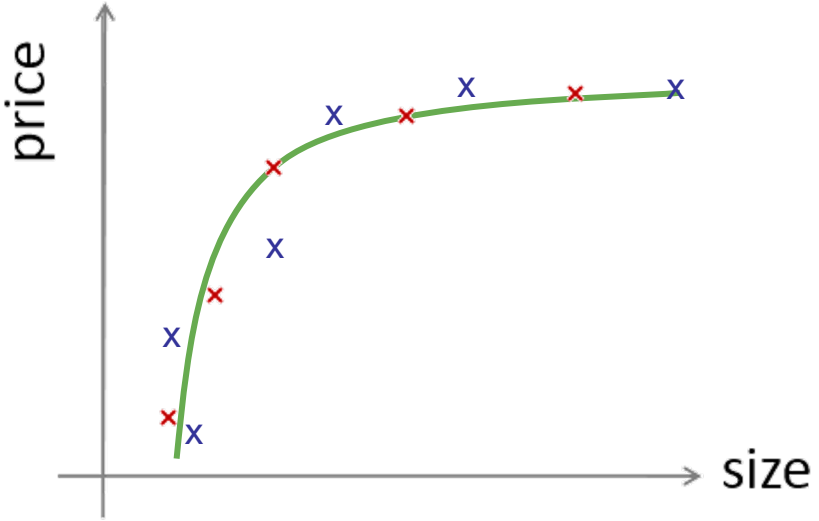


https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Learning = Generalisation

- How well does the model perform on **UNSEEN** data?

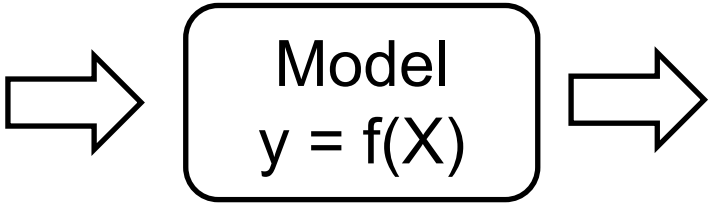




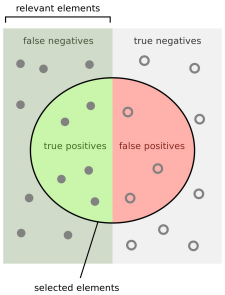
Model Evaluation

Test Dataset

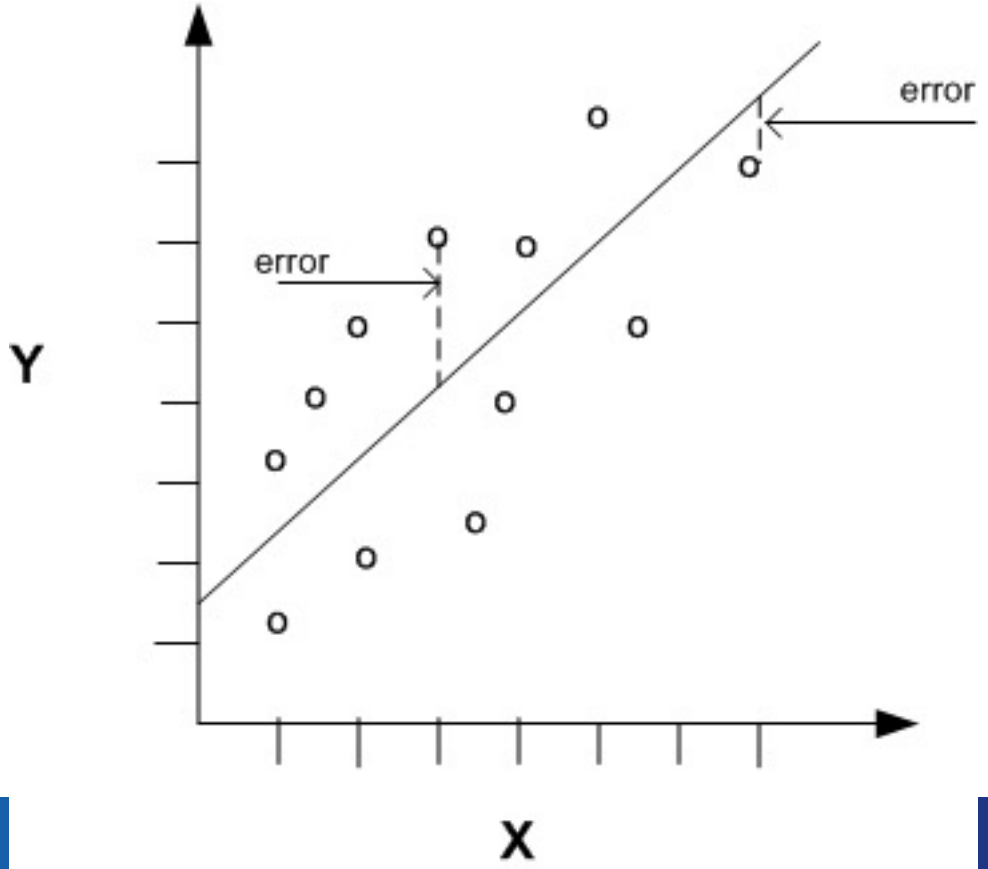
x1	...	xn	y

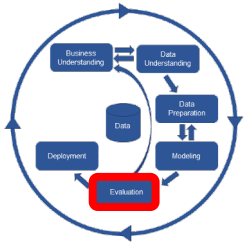


Metrics

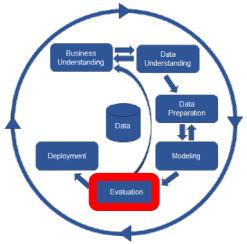


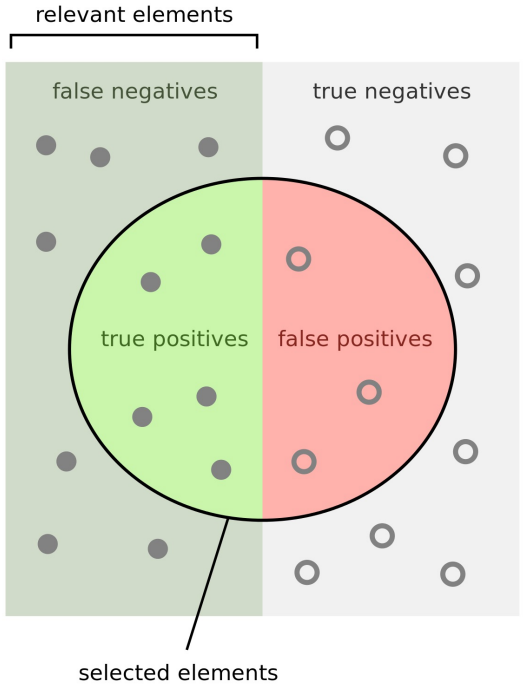
(R)MSE


error
error
Y
X


Business Understanding
Data Understanding
Data Preparation
Data
Deployment
Modeling
Evaluation

Accuracy and Precision





How many decisions are correct?

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{all elements}}$$

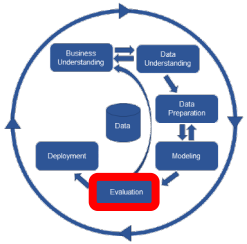
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

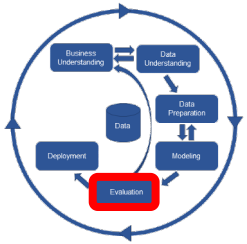
Precision AND Recall?



	Precision	Recall
Classifier 1	0.5	0.4
Classifier 2	0.7	0.1
Classifier 3	0.02	1.0

- Usually there is a trade-off between Precision and Recall
- Mean ($\frac{P+R}{2}$) is not meaningful
  - Classifier 3 has highest mean but predicts ‘positive’ all the time

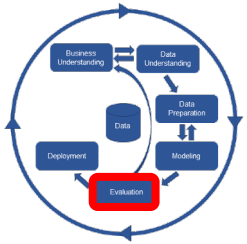
F_1 Score

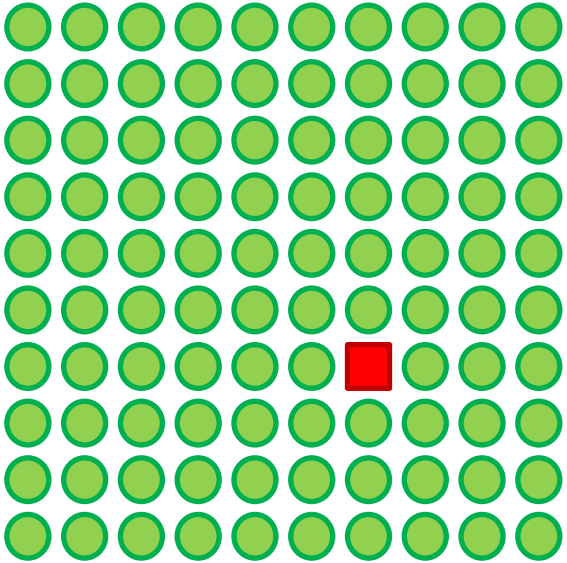


	Precision	Recall	F_1
Classifier 1	0.5	0.4	0.444
Classifier 2	0.7	0.1	0.175
Classifier 3	0.02	1.0	0.039

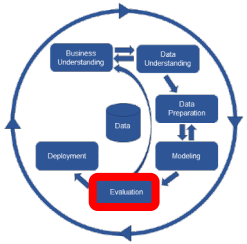
- $F_1 = 2\frac{P \times R}{P+R}$
  - Higher is better
  - F_1 is 0 if P=0 or R=0
  - F_1 is 1 if P=1 and R=1

Accuracy Paradox





Contingency Tables: > 2 Classes



	Predicted C1	Predicted C2	Predicted C3
Actual C1	100	40	25
Actual C2	25	50	4
Actual C3	1	0	7

Summary

- ML systems are created in a 6 step process
  - Data collection & preparation is where most effort is spent
- ML Models are trained and evaluated on separate data sets
- Training means choosing parameters that optimise a metric
- Knowing the distribution of (unseen) data is important for selecting models and metrics

University of
South Australia

Questions?