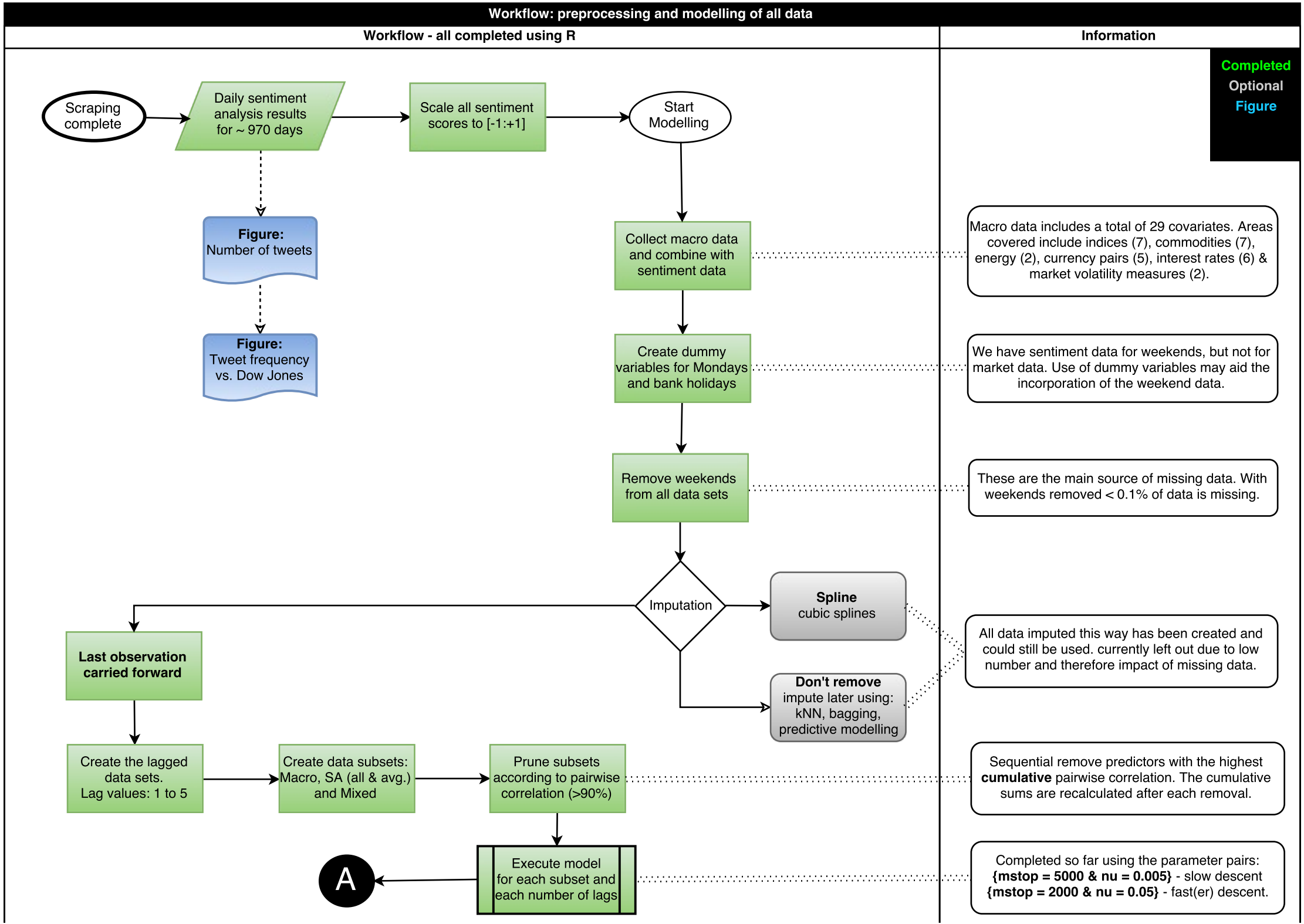
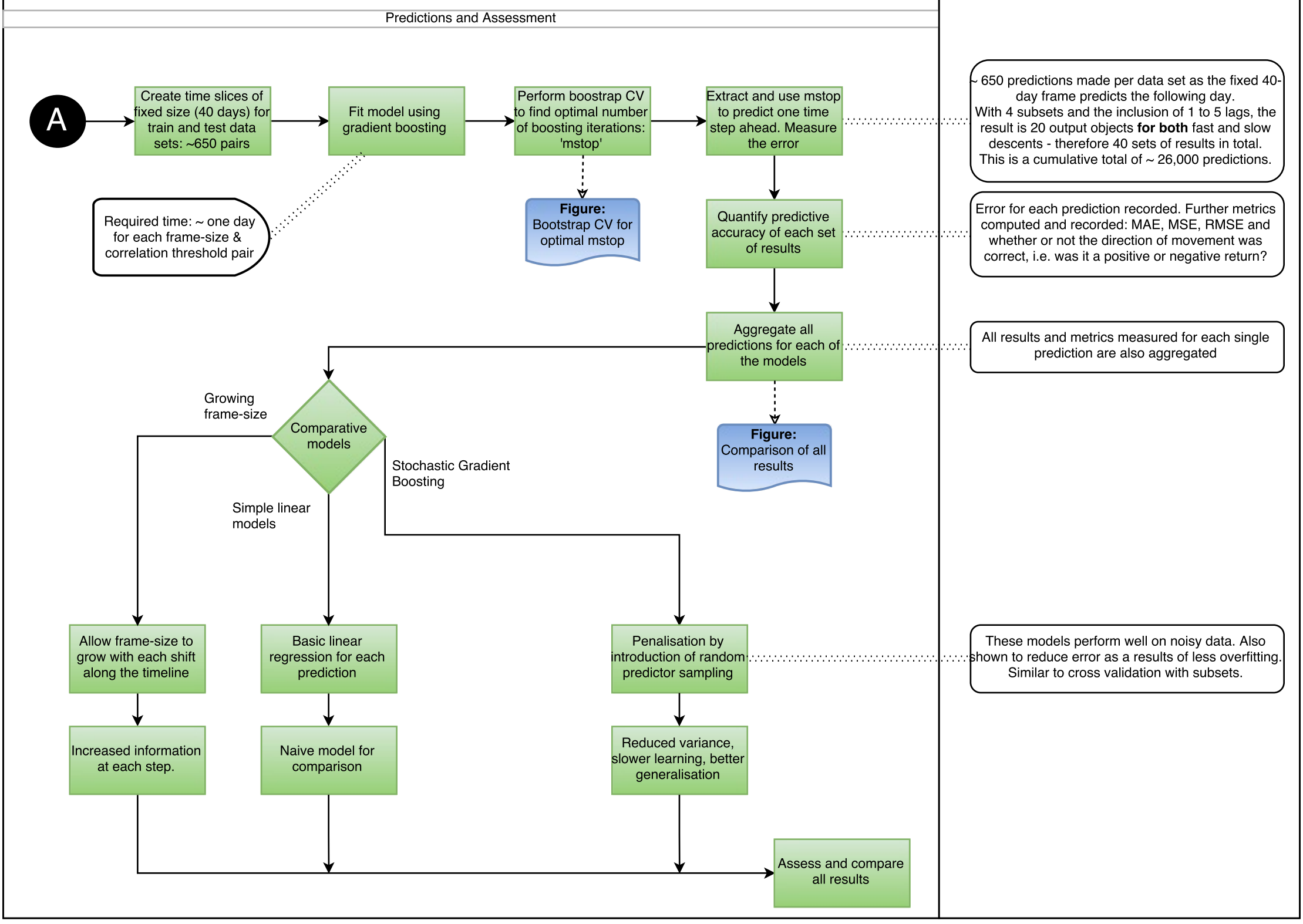

Workflow: preprocessing and modelling of all data
Workflow - all completed using R
Information
Completed
Optional
Figure
Scraping complete
Daily sentiment analysis results for ~ 970 days
Scale all sentiment scores to [-1:+1]
Start Modelling
Figure: Number of tweets
Figure: Tweet frequency vs. Dow Jones
Collect macro data and combine with sentiment data
Macro data includes a total of 29 covariates. Areas covered include indices (7), commodities (7), energy (2), currency pairs (5), interest rates (6) & market volatility measures (2).
Create dummy variables for Mondays and bank holidays
We have sentiment data for weekends, but not for market data. Use of dummy variables may aid the incorporation of the weekend data.
Remove weekends from all data sets
These are the main source of missing data. With weekends removed < 0.1% of data is missing.
Imputation
Spline
cubic splines
Last observation carried forward
Don't remove
impute later using:
kNN, bagging,
predictive modelling
All data imputed this way has been created and could still be used. currently left out due to low number and therefore impact of missing data.
Create the lagged data sets.
Lag values: 1 to 5
Create data subsets:
Macro, SA (all & avg.)
and Mixed
Prune subsets according to pairwise correlation (>90%)
Sequential remove predictors with the highest cumulative pairwise correlation. The cumulative sums are recalculated after each removal.
A
Execute model for each subset and each number of lags
Completed so far using the parameter pairs:
{mstop = 5000 & nu = 0.005} - slow descent
{mstop = 2000 & nu = 0.05} - fast(er) descent.


Predictions and Assessment
A
Create time slices of fixed size (40 days) for train and test data sets: ~650 pairs
Fit model using gradient boosting
Perform boostrap CV to find optimal number of boosting iterations: 'mstop'
Extract and use mstop to predict one time step ahead. Measure the error
~ 650 predictions made per data set as the fixed 40-day frame predicts the following day. With 4 subsets and the inclusion of 1 to 5 lags, the result is 20 output objects for both fast and slow descents - therefore 40 sets of results in total. This is a cumulative total of ~ 26,000 predictions.
Required time: ~ one day for each frame-size & correlation threshold pair
Figure: Bootstrap CV for optimal mstop
Quantify predictive accuracy of each set of results
Error for each prediction recorded. Further metrics computed and recorded: MAE, MSE, RMSE and whether or not the direction of movement was correct, i.e. was it a positive or negative return?
Aggregate all predictions for each of the models
All results and metrics measured for each single prediction are also aggregated
Figure: Comparison of all results
Growing frame-size
Comparative models
Stochastic Gradient Boosting
Simple linear models
Allow frame-size to grow with each shift along the timeline
Basic linear regression for each prediction
Penalisation by introduction of random predictor sampling
These models perform well on noisy data. Also shown to reduce error as a results of less overfitting. Similar to cross validation with subsets.
Increased information at each step.
Naive model for comparison
Reduced variance, slower learning, better generalisation
Assess and compare all results