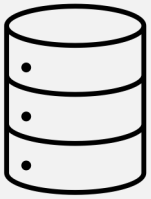
GEOG5927 Predictive Analytics: Machine Learning

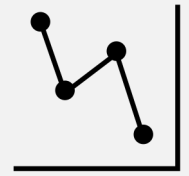
Rachel Oldroyd

R.Oldroyd@leeds.ac.uk

@r_oldroyd

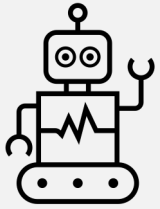
Machine Learning



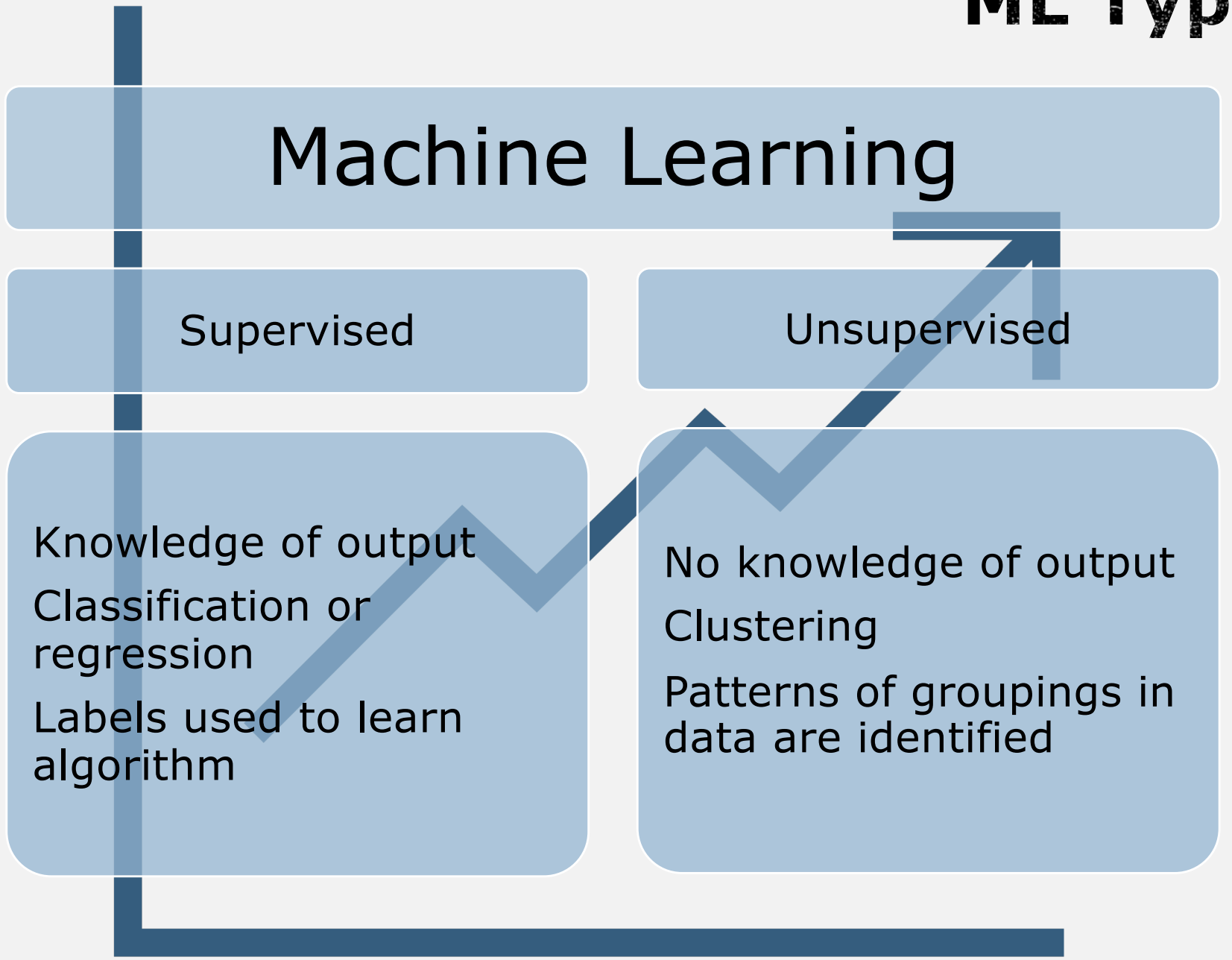


Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed

(https://www.expert.ai/blog/machine-learning-definition/)



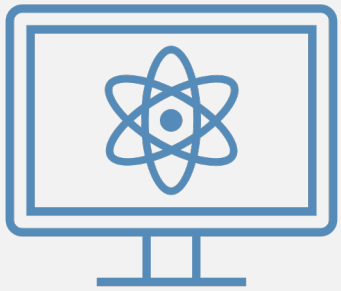
ML Types


Machine Learning
Supervised
Unsupervised
Knowledge of output
Classification or regression
Labels used to learn algorithm
No knowledge of output
Clustering
Patterns of groupings in data are identified

ML Algorithms

Commonly used supervised models:

- Support Vector Machine
- Naïve Bayes
- K Nearest Neighbour
- Logistic Regression
- Linear Regression
- Random Forest

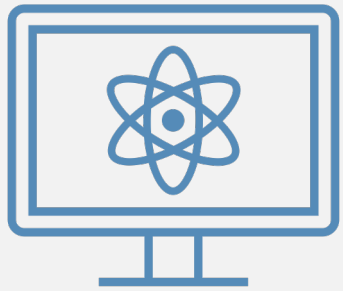


More Info:
https://www.ibm.com/cloud/learn/supervised-learning#toc-supervised-QVA1W1YW

ML Algorithms

Commonly used supervised models:

- Support Vector Machine
- Naïve Bayes
- K Nearest Neighbour
- Logistic Regression
- Linear Regression
- Random Forest

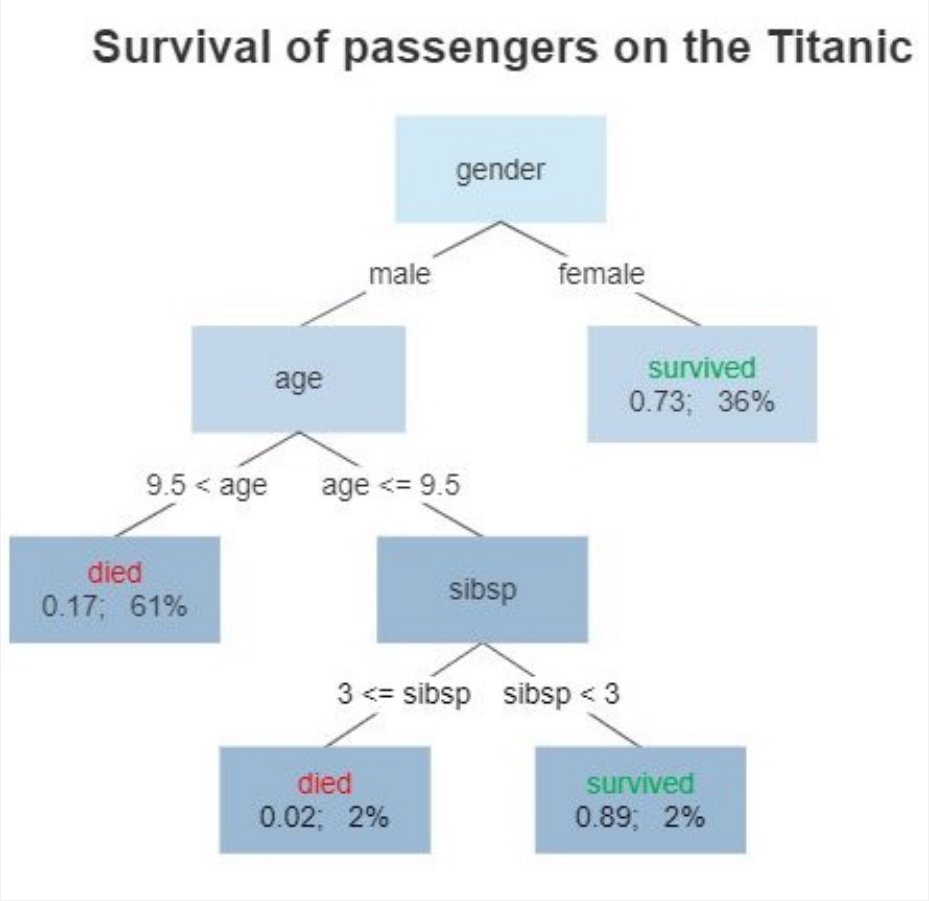


More Info:
https://www.ibm.com/cloud/learn/supervised-learning#toc-supervised-QVA1W1YW

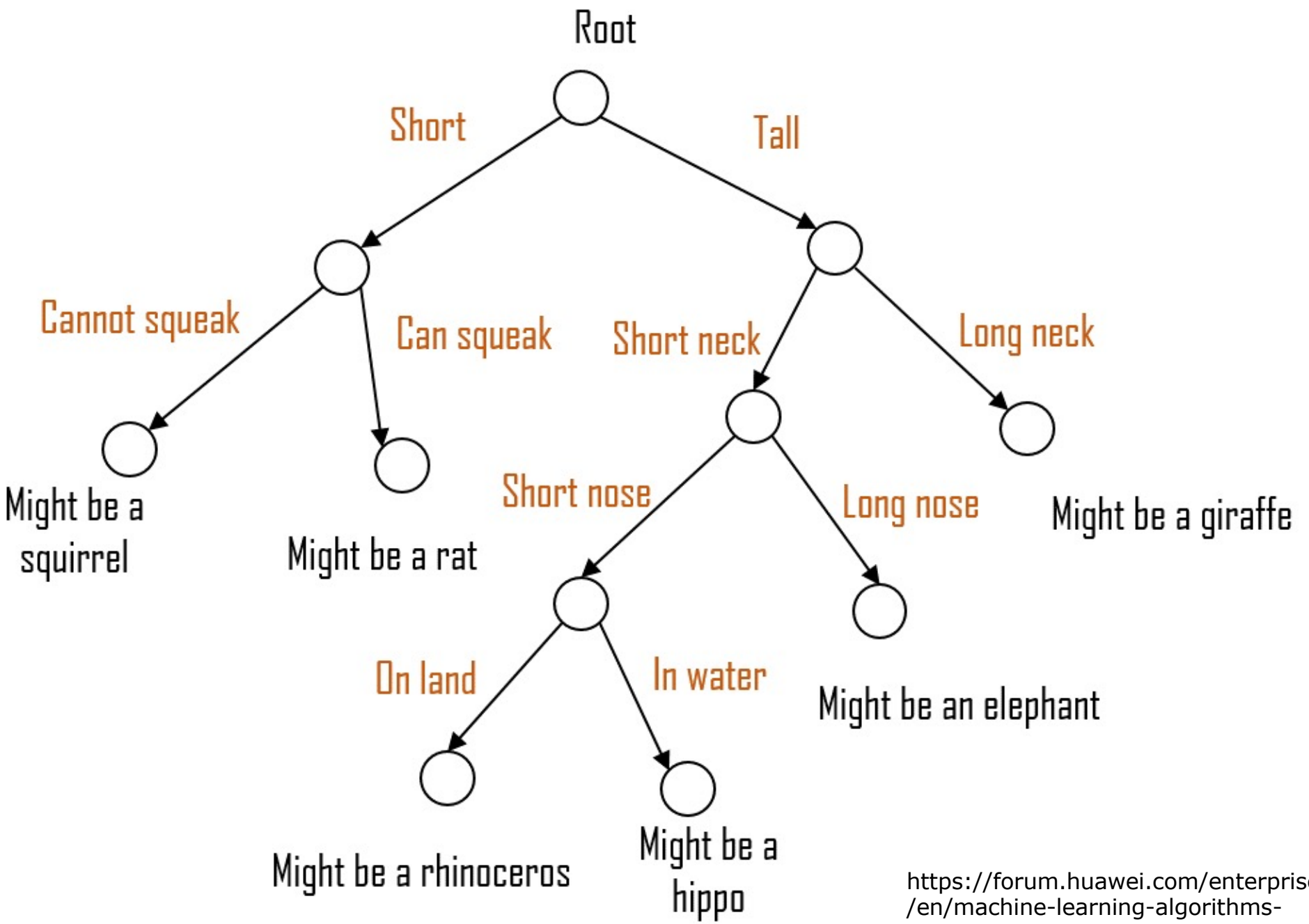
Decision Trees

Predict the class / outcome of a target variable

- The relationship / association between outcome and predictors is represented as a tree
- A set of splitting rules is determined to build the tree
- Each split attempts to minimize entropy (randomness in the data)



https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:Decision_Tree.jpg



https://forum.huawei.com/enterprise/en/machine-learning-algorithms-decision-trees/thread/710283-895

Decision Trees Pros & Cons

Pros:

- Easy to build
- Easy to conceptualise / visualise
- Minimal data prep
- Both continuous & categorical data

Cons:

- Prone to overfitting = small changes in data can lead to large changes in the tree structure.

Random Forest

Can address problems with overfitting…

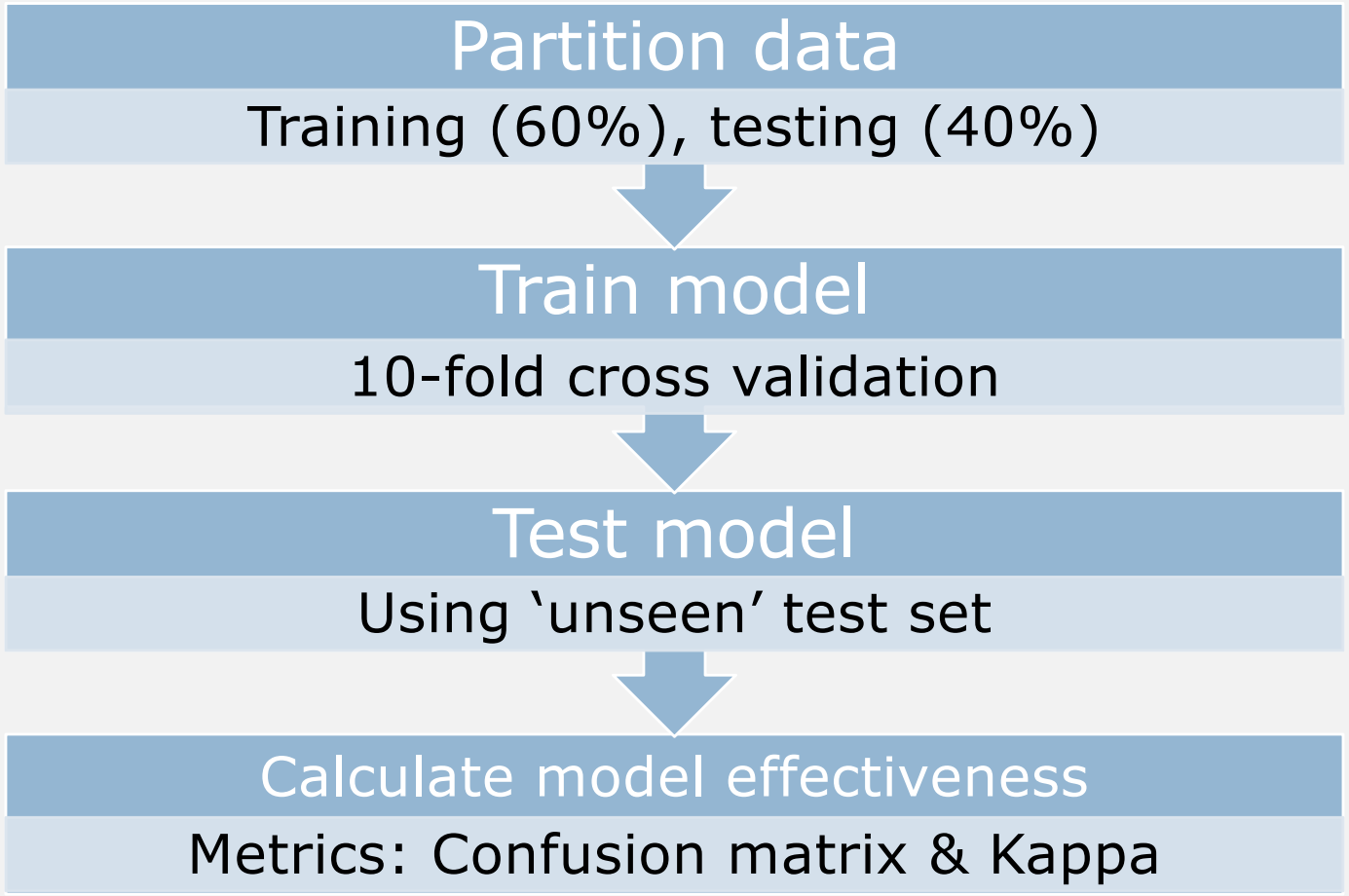
- Random Forest fits hundreds of decision trees to average the best outcomes (bagging)
- A subset of predictor vars used for each tree -> prevents overfitting

This week's practical

Use Random Forest to predict 'overseas airport' outcome variable -> Palma Mallorca

- Use synthetic data generated last week
- Method used for targeted marketing strategies
- Input variables:
  - Age band
  - Sex (m,f)
  - Number of children

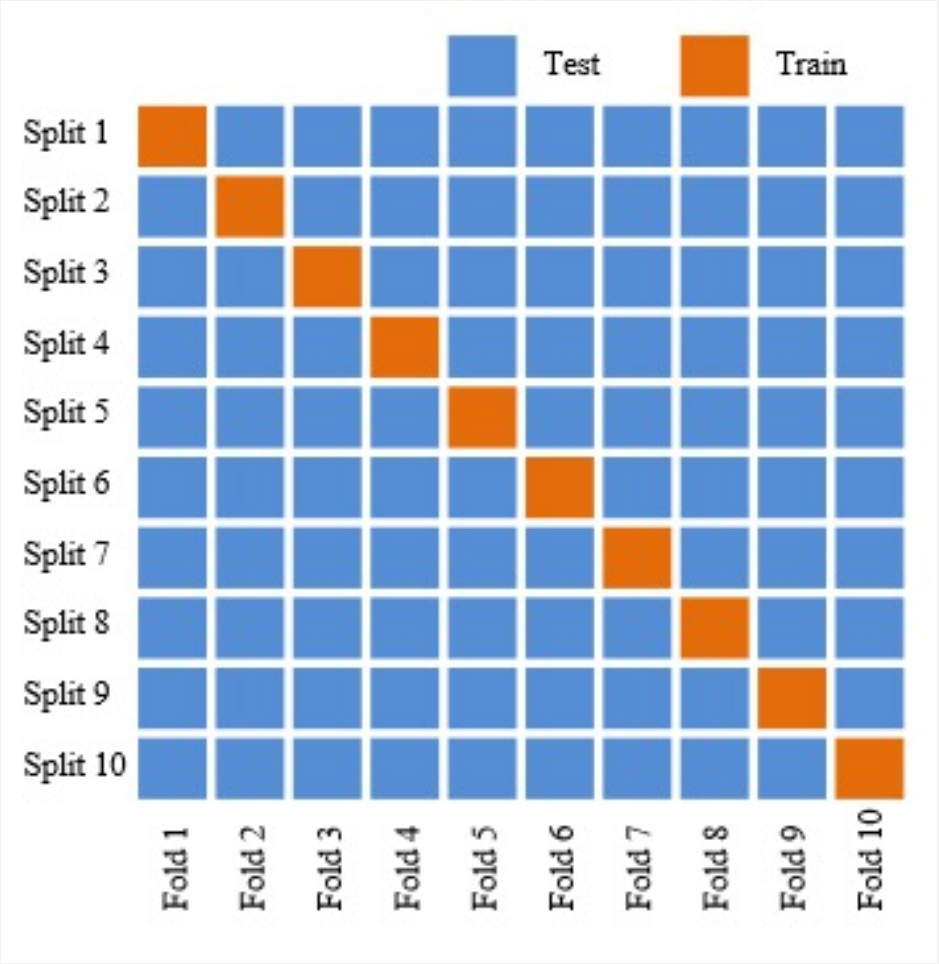
Strategy


Partition data
Training (60%), testing (40%)
Train model
10-fold cross validation
Test model
Using ‘unseen’ test set
Calculate model effectiveness
Metrics: Confusion matrix & Kappa

Cross-validation

Data split into 10 equal folds (stratified sampling)

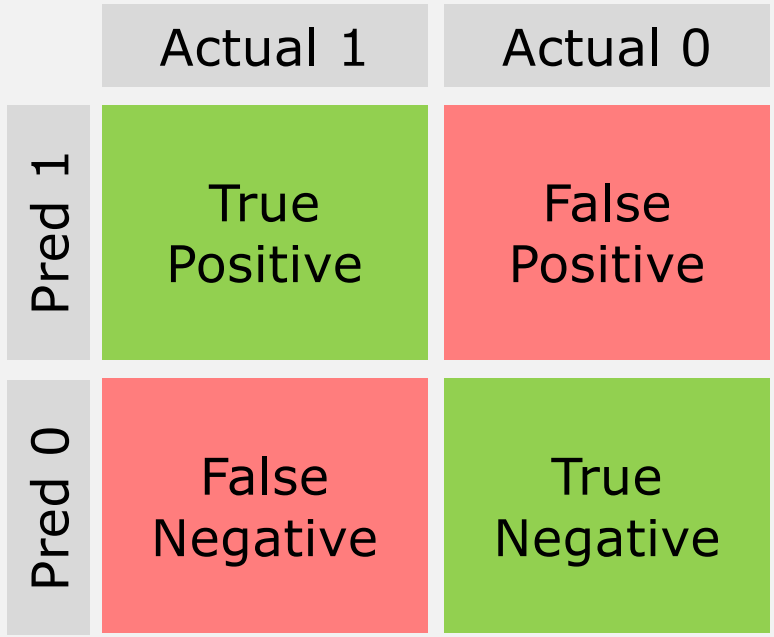
- 1 fold set aside to test effectiveness of algorithm
- Remaining 9 folds used to learn algorithm (i.e. determine split points in data)
- Best performing algorithm chosen as final model



Confusion matrix

Confusion Matrix

- Compare predicted & actual labels
- TP- Predicted positive, actual positive
- FP- predicted positive, actual negative
- TN- predicted negative, actual negative
- FN, predicted negative, actual positive



1 = positive class (PMI)
0 = negative class (not PMI)

Kappa

An accuracy metric which takes into account class size (useful for imbalanced problems):

$$k = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

$$k = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

$k = 1$ indicates exact agreement between predicted and actual labels. Important to interpret alongside the CM.

Class imbalance

PMI contributes 8% of overseas airport variable

- Model can achieve 92% accuracy by labelling all records as 0 (not PMI)
- Not helpful!
- We are often interested in identifying a small number of data records in a bigger set.
- Need to resample our data to minimise the class imbalance problem.
  - i.e. 1:1 or 3:2 ratio

Practical Overview

Summary:

- Partition data
- Train RF algorithm using cross validation
- Apply to 'unseen' test set (prediction)
- Evaluate performance
  - 1st model suffers class imbalance problems

- Resample training data
- Retrain model
- Re-evaluate performance
  - 2nd model performs better

Script, instructions and R workspace provided

- Consider ways in which model might be improved

Predicting non-compliant food outlets in England and Wales using neighbourhood characteristics: a machine learning approach

Dr Rachel Oldroyd, Dr Michelle Morris, Prof Mark Birkin

An ESRC Data Investment

Background

Local Authorities (LAs) enforce food standards

- Overseen by Food Standards Agency (FSA)

Every food serving business is inspected by a Food Hygiene Officer

- Food Hygiene Rating Scheme (FHRS) score*

*Scotland operates a pass / fail system

Rationale 1

LAs are struggling to meet their inspection targets:

- Only 2% of LA's in the UK have no overdue inspections*
- 18% of LA's have over 20% of businesses overdue an inspection*
- Recent work suggests this has worsened during the pandemic

* National Audit Office 2019

Rationale 2

Business owners not receiving support
Consumers are exposed to unknown levels of risk

Extremely problematic -> 60% of foodborne illness is contracted outside the home

Foodborne illness affects ~2 million people annually

- At a cost of 1.6 billion GBP



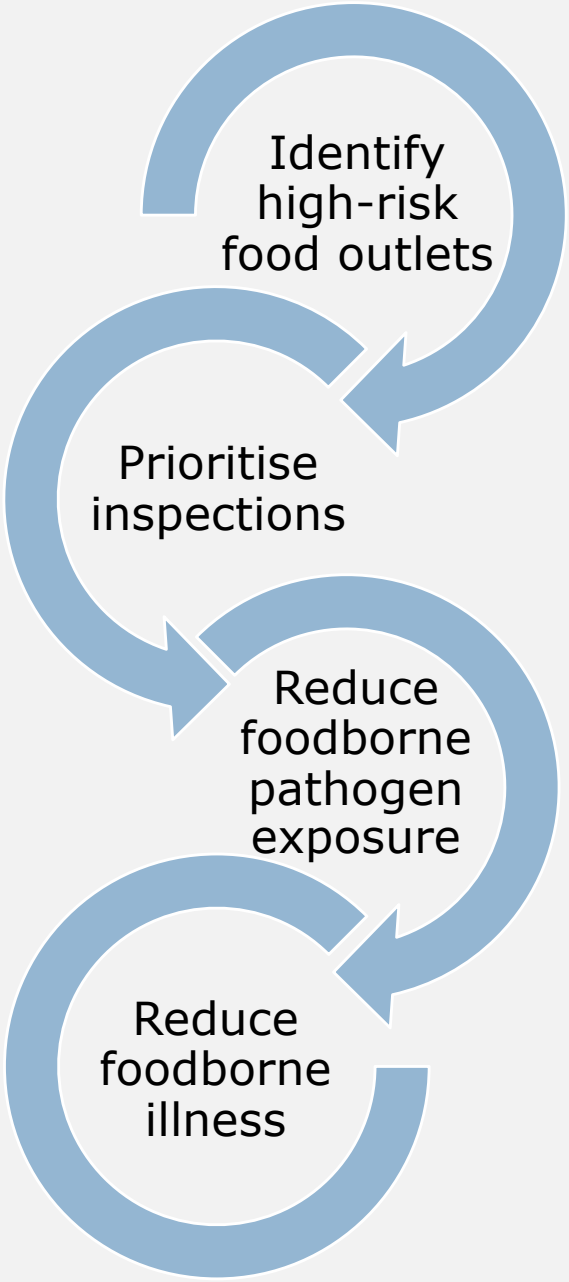
Non-compliance

5	Hygiene standards are very good
4	Hygiene standards are good
3	Hygiene standards are generally satisfactory
2	Some improvement is necessary
1	Major improvement is necessary
0	Urgent improvement is required

Not broadly compliant = FHRS ≤ 2

Broadly compliant = FHRS ≥ 3

Aim



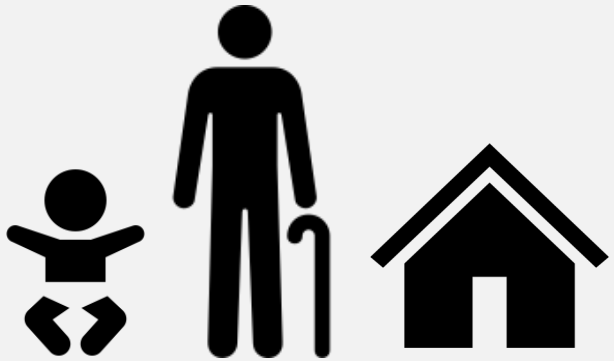
Explore the utility of machine learning to predict high-risk (non-compliant) food outlets in England and Wales

- Using neighbourhood characteristics

Context

Previous studies have reported significant associations between food outlet compliance and neighbourhood characteristics:

- Urbanness
- Demographics
  - Age
  - Ethnicity
  - Deprivation



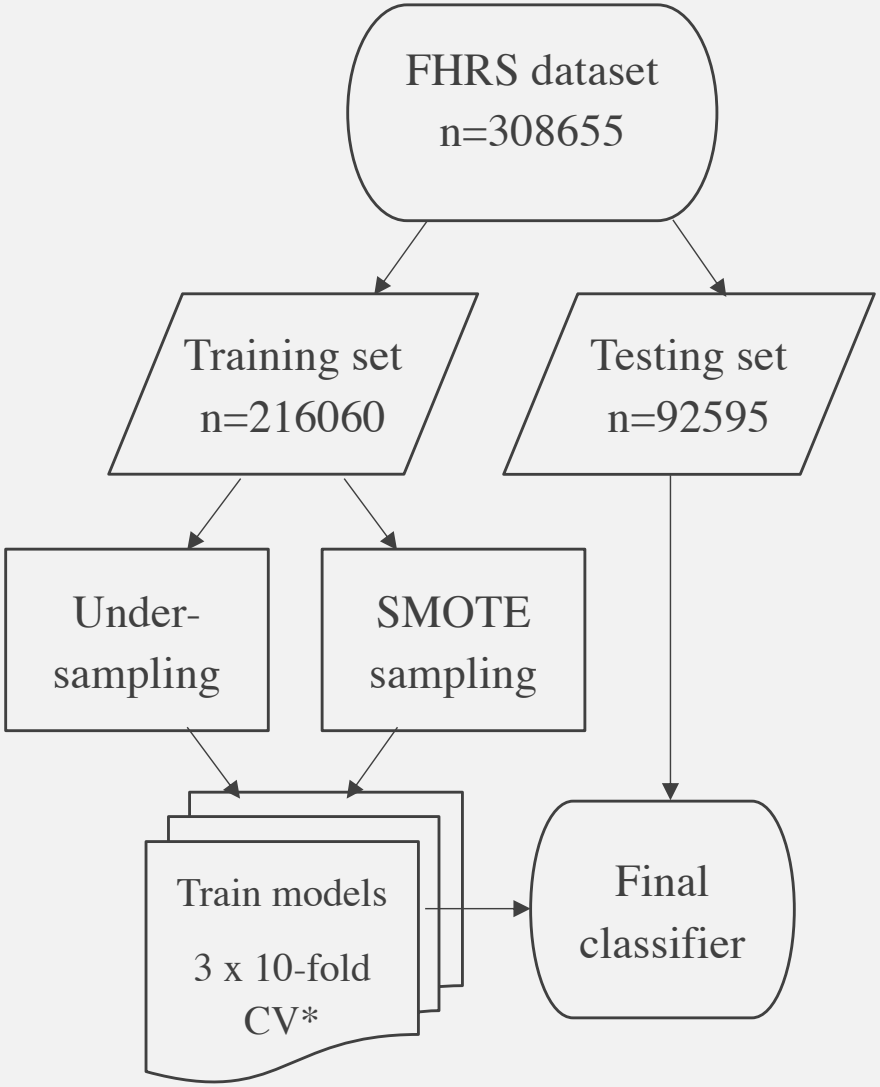
Data

Outcome variable:
FHRS (0, 1)

Predictor variables:

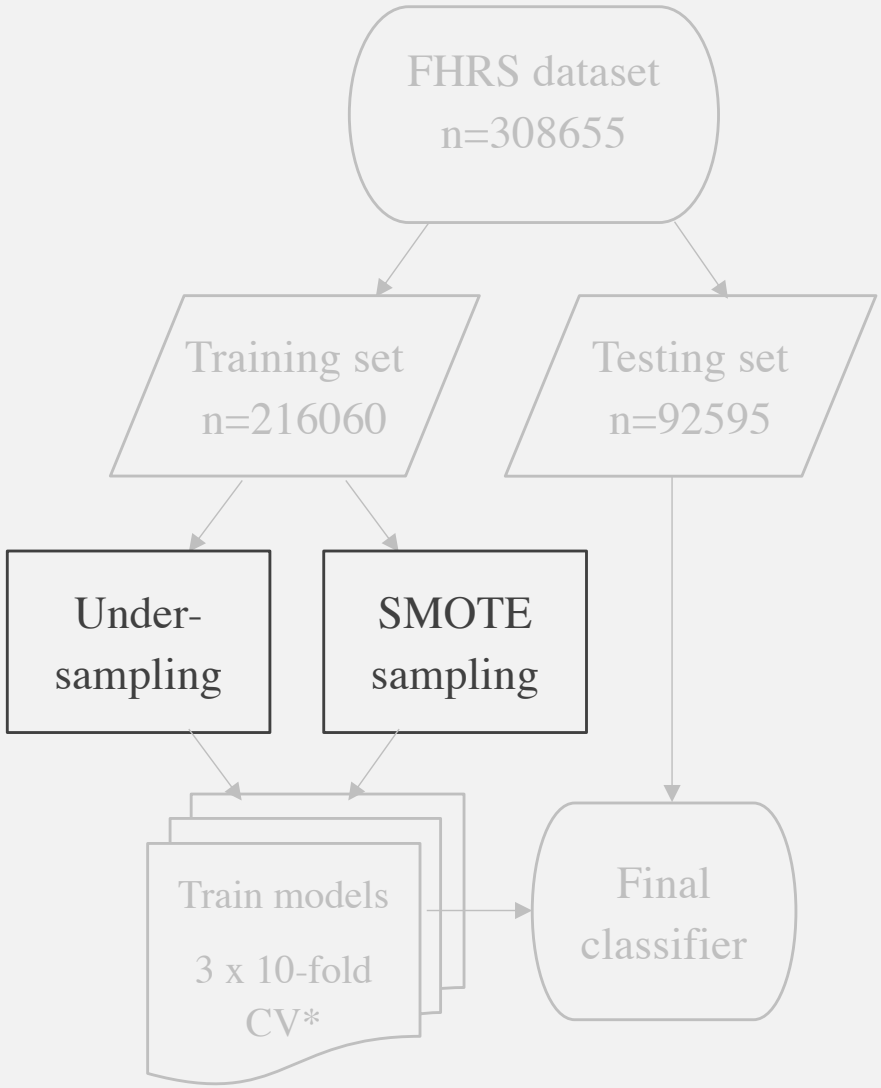
- Business type
- Region
- Age (% individuals)
- Ethnicity (% individuals)
- No car access (% households)
- Renting (% households)
- Overcrowding (% households)
- Unemployment (% individuals)
- Rural Urban Classification (RUC)
- Output Area Classification (OAC)

Method Overview



1. Split data using stratified sampling:
   - Training set (70%)
   - Test set (30%)
2. Resample training set to overcome class imbalance
3. Train models using cross-validation
4. Test model on ‘unseen’ data – test set.
5. Calculate model metrics

Sampling strategies



Different sampling strategies & ratios to address class imbalance (7% non-compliant outlets)

- Under-sampling
  - Straight forward
  - Reduces size of training set
- Synthetic Minority Over Sampling Technique (SMOTE)
  - Add synthetic data points to minority class –KNN
  - Under-sample majority class
  - Maximises data for training
  - Time intensive

Final Model

Repeat model training & testing across sampling strategies (Under sampling & SMOTE), ratios (e.g. 1:1, 3:2, 2:3) and three algorithms:

- Linear SVM
- Radial SVM
- Random Forest

=33 models in total run on High Performance Computer (HPC)

SMOTE, Random Forest, 1:1 adopted as final model

Results

	RF Set 1 n=92595		RF unsampled n=92595	
	unweighted	weighted	unweighted	weighted
Probability Threshold	0.603	0.481	0.067	0.021
AUC	0.87	0.87	0.796	0.796
Sensitivity	0.759	0.843	0.661	0.859
Specificity	0.858	0.745	0.797	0.481
True Positives	4624	5139	4029	5903
False Positives	12264	21676	17571	77591
True Negatives	74235	64823	68928	8908
False Negatives	1472	957	2067	193
Kappa	0.338	0.230	0.210	0.010
Precision	0.274	0.192	0.187	0.071

Results summary

85% non-compliant outlets identified by model

Highly predictive variables:

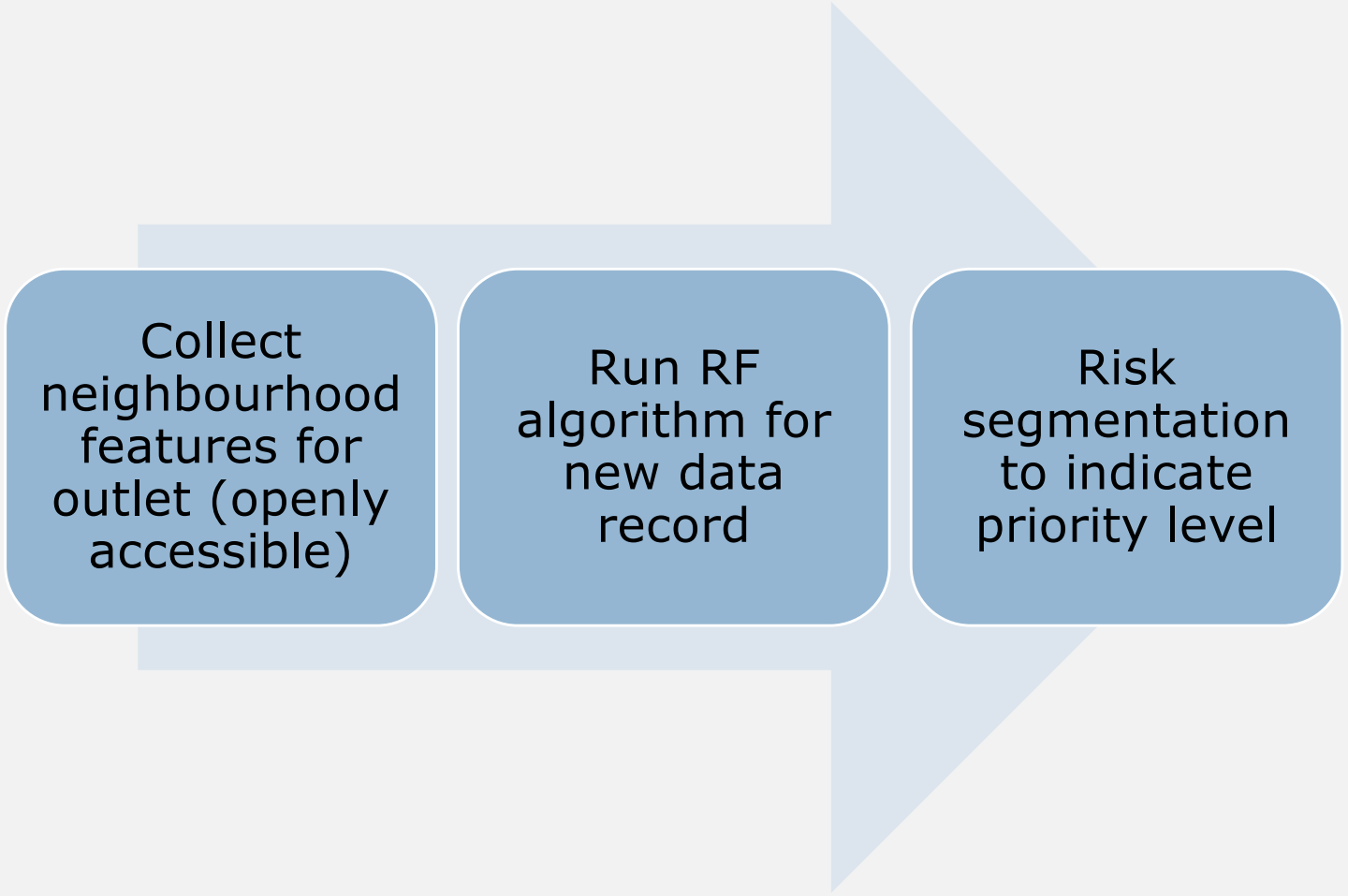
- Deprivation
- Non-white ethnicities
- Some age variables
- Large urban areas
- Takeaways / sandwich shops

Further research required to unpick these relationships

- High population turnover -> high staff turnover
- FHRS score display -> incentive to improve
- The role of cuisine type

Real world application

For a newly opened outlet or routine inspection:



Limitations

- Data
  - FHRS -> Snapshot in time (some inspection data > 5 years old)
  - Inspection bias (deprivation, ethnicity)
  - Census 2011 outdated (esp. in large urban areas)

- Model doesn’t take behaviours into account
  - Food hygiene in the home
  - Habits of eating outside the home

- Problems with entropy based classification
  - Future work with look at alternative algorithms
  - Partial permutations (Altman et al. 2010), unbiased trees (Painsky and Rosset 2017)

Publication

Open Access Article

Predicting Food Safety Compliance for Informed Food Outlet Inspections: A Machine Learning Approach

by Rachel A. Oldroyd [1,2,*], Michelle A. Morris [1,3,4] and Mark Birkin [1,2,4]

[1] Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9JT, UK
[2] School of Geography, University of Leeds, Leeds LS2 9JT, UK
[3] School of Medicine, University of Leeds, Leeds LS2 9JT, UK
[4] Alan Turing Institute, London NW1 2DB, UK
[*] Author to whom correspondence should be addressed.

Academic Editor: Paul B. Tchounwou

Int. J. Environ. Res. Public Health **2021**, *18*(23), 12635; https://doi.org/10.3390/ijerph182312635

Questions