CLASSIFICATION AND VALUE ESTIMATION

SETTING THE STAGE

"Data science does not replace statistical modeling and data analysis; it augments them."

(P. Boily)

"Data is not information, information is not knowledge,
knowledge is not understanding, understanding is not wisdom."

(attributed to Cliff Stoll in Keeler's *Nothing to Hide: Privacy in the 21st Century*, 2006)

Hsu, W., et al. [2013], Data mining based tax audit selection: a case study of a pilot project at the Minnesota Department of Revenue, *Real Word Data Mining Applications*, *Annals of Information Systems,* v.17, Springer).

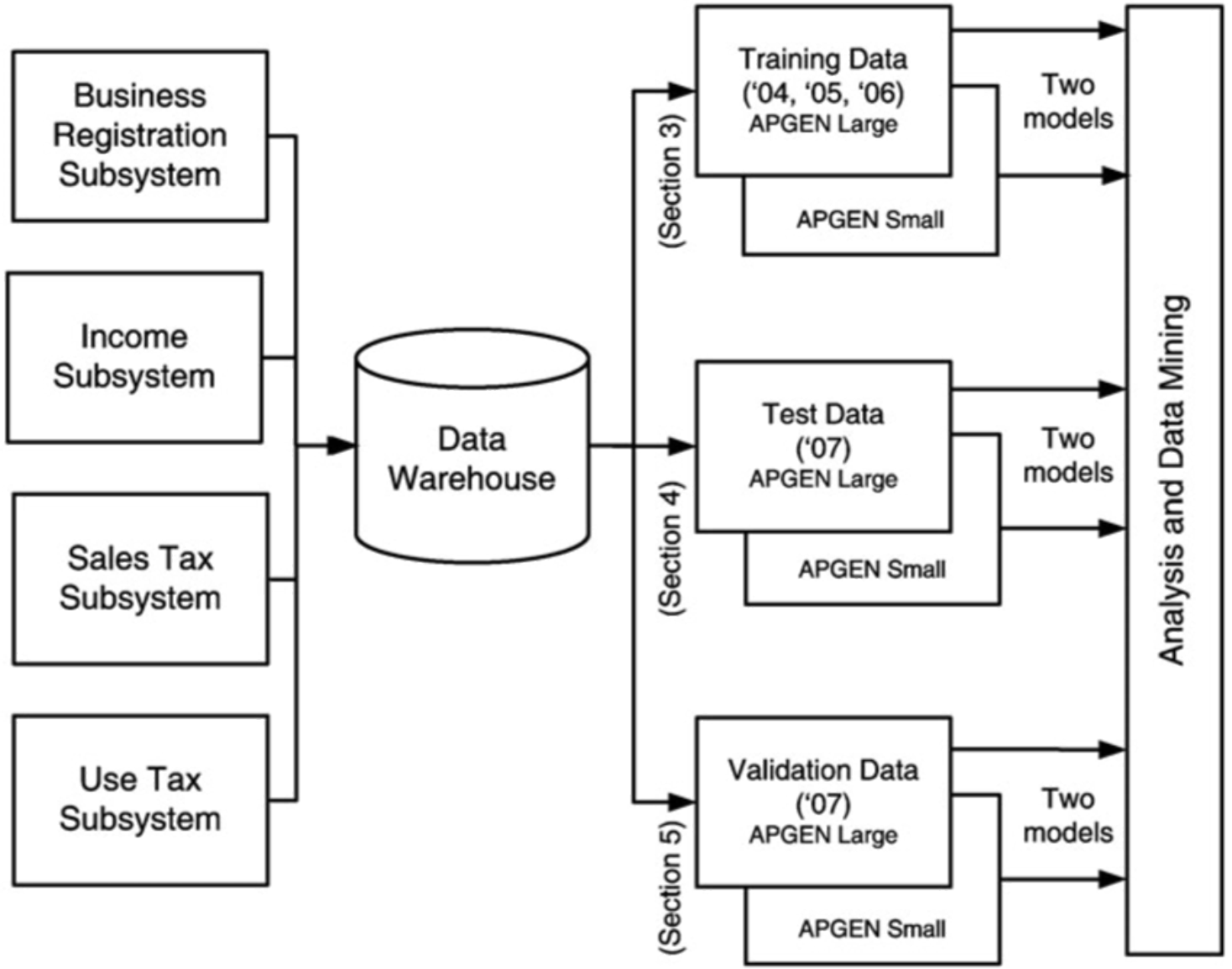
CASE STUDY: MINNESOTA TAX AUDIT

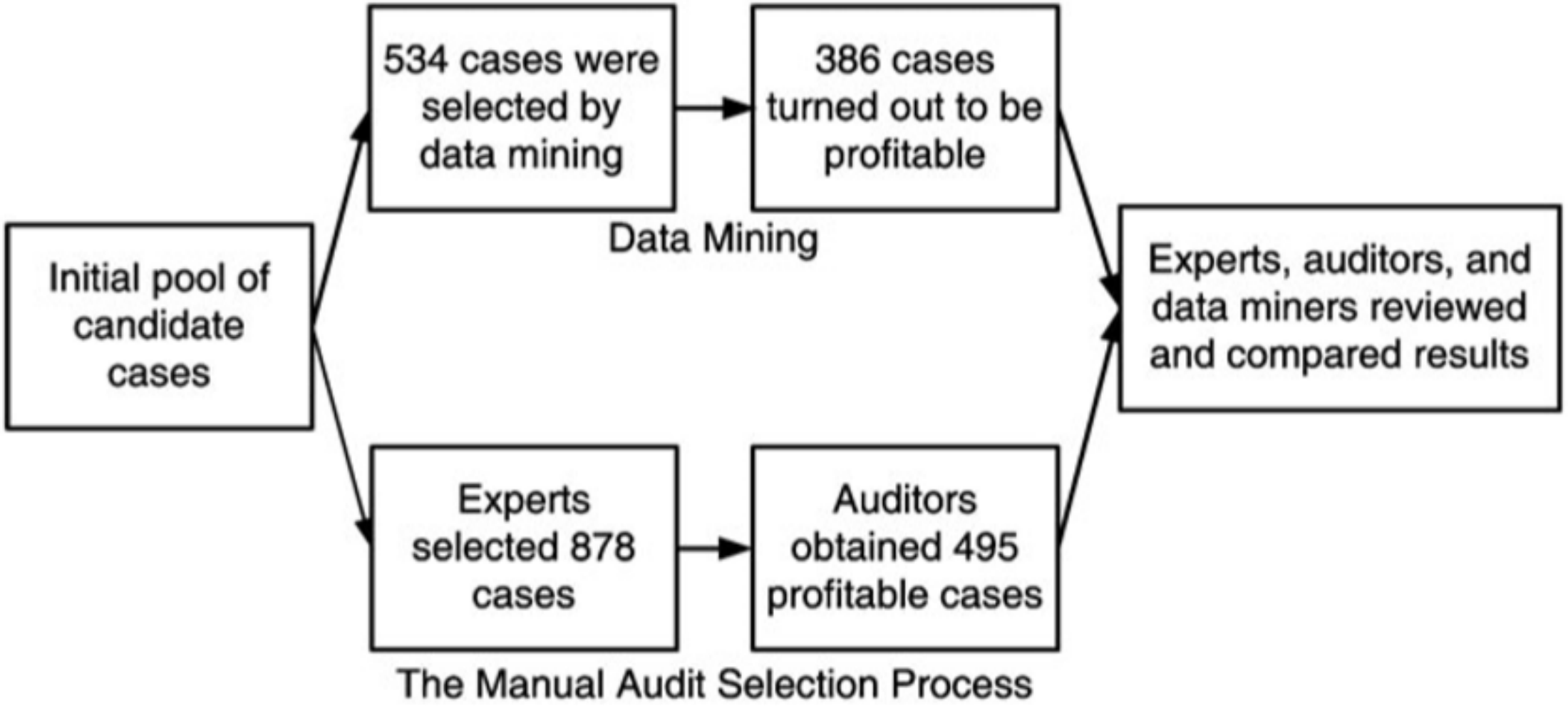
Large gaps between revenue owed (in theory) and revenue collected (in practice) are problematic for governments.

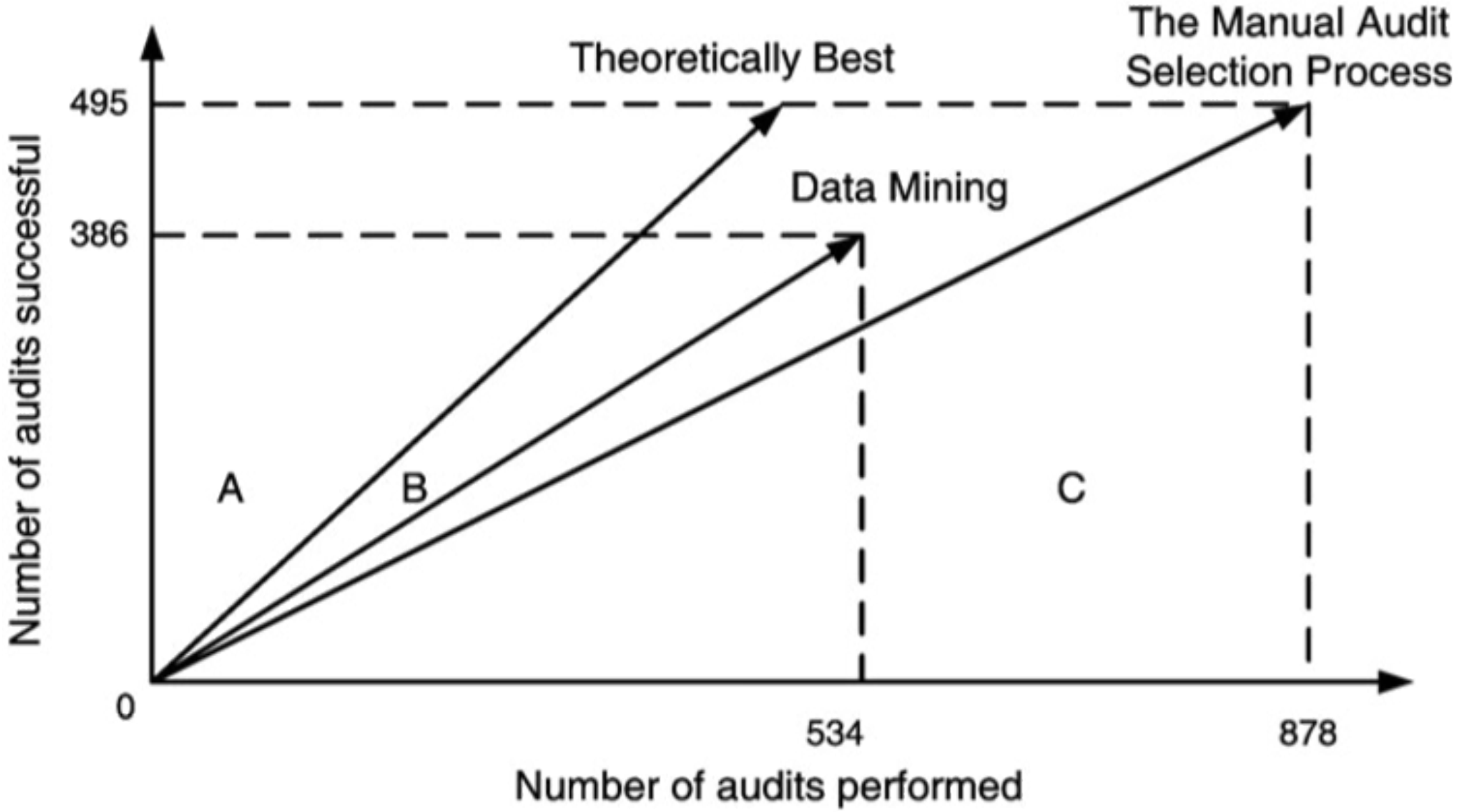
Revenue agencies implement various fraud detection strategies (such as audit reviews) to bridge that gap.

Business audits are costly – are there **algorithms that can predict whether an audit is likely to be successful or a waste of resources**?

Should a tax collection agency seek to maximize its revenues/profits or to ensure compliance?


Business Registration Subsystem
Income Subsystem
Sales Tax Subsystem
Use Tax Subsystem
Data Warehouse
(Section 3)
(Section 4)
(Section 5)
Training Data ('04, '05, '06) APGEN Large
APGEN Small
Test Data ('07) APGEN Large
APGEN Small
Validation Data ('07) APGEN Large
APGEN Small
Two models
Two models
Two models
Analysis and Data Mining


534 cases were selected by data mining
386 cases turned out to be profitable
Data Mining
Initial pool of candidate cases
Experts, auditors, and data miners reviewed and compared results
Experts selected 878 cases
Auditors obtained 495 profitable cases
The Manual Audit Selection Process


The Manual Audit
Selection Process
Theoretically Best
Data Mining
495
386
Number of audits successful
A
B
C
0
534
878
Number of audits performed

CLASSIFICATION OVERVIEW

In **classification**, a sample set of data (the **training** set) is used to determine rules and patterns that divide the data into pre-determined groups, or classes (supervised learning; predictive analytics).

The training data usually consists of a **randomly** selected subset of the **labeled** (target) data.

Value estimation (regression) is akin to classification when the target variable is numerical.

CLASSIFICATION OVERVIEW

In the **testing** phase, the model is used to assign a class to observations for which the label is hidden, but ultimately known (the **testing** set).

The performance of a classification model is evaluated on the testing set, **never** on the training set.

Technical issues include:

- selecting the features to include in the model
- selecting the algorithm
- etc.

APPLICATIONS

Medicine and Health Science

- predicting which patient is at risk of suffering a second, fatal heart attack within 30 days based on health factors (blood pressure, age, sinus problems, etc.)

Social Policies

- predicting the likelihood of requiring assisting housing in old age based on demographic information/survey answers

Marketing and Business

- predicting which customers are likely to switch to another cell phone company based on demographics and usage

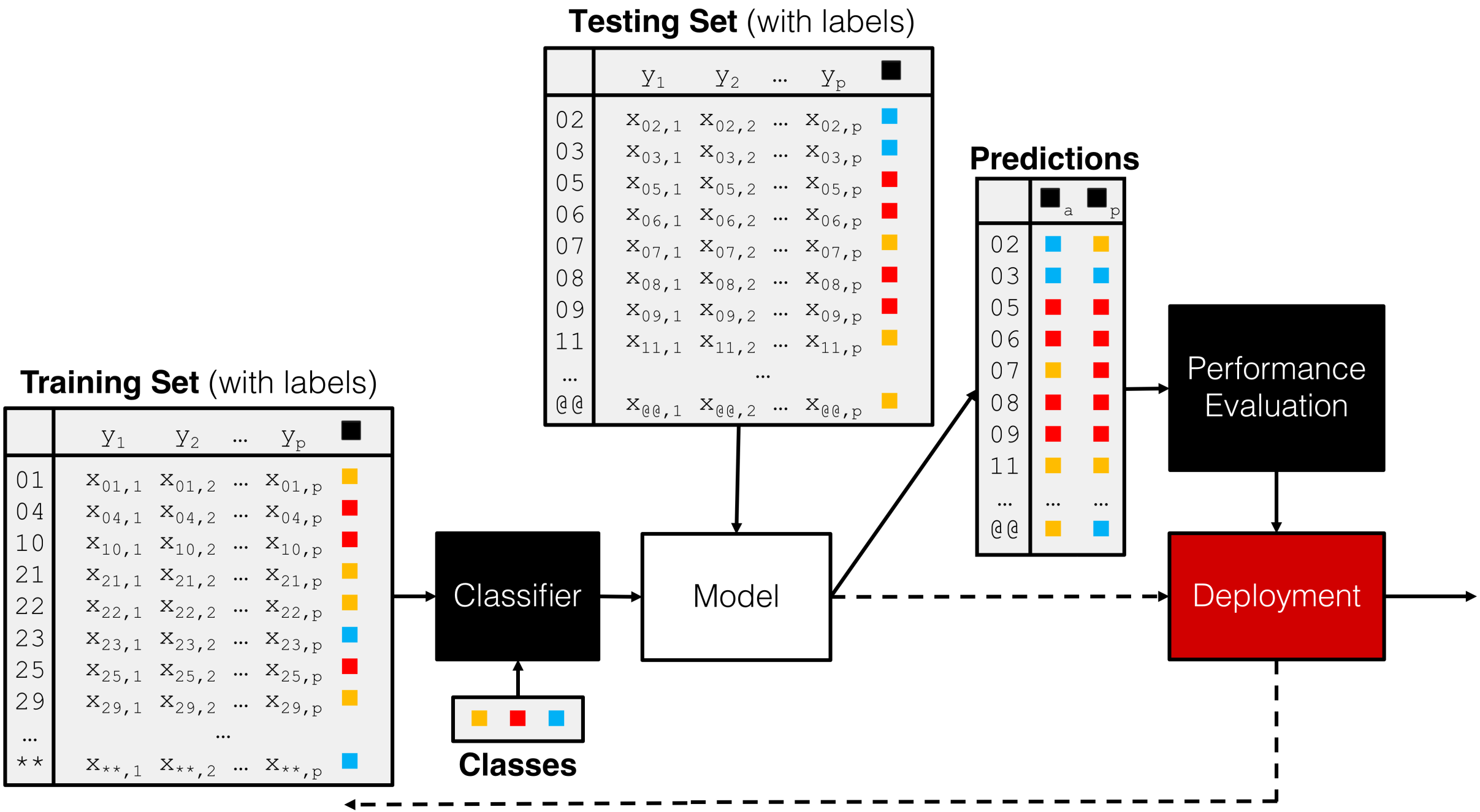
EXAMPLE

Scenario:

A motor insurance company has a fraud investigation dept. that studies up to 30% of all claims made, yet money is still getting lost on fraudulent claims.

Questions: can we predict

- whether a claim is likely to be fraudulent?
- whether a customer is likely to commit fraud in the near future?
- whether an application for a policy is likely to result in a fraudulent claim?
- the amount by which a claim will be reduced if it is fraudulent?


Testing Set (with labels)
Predictions
Training Set (with labels)
Classifier
Model
Performance Evaluation
Deployment
Classes

CLASSIFICATION SCHEMES

Logistic Regression

Neural Networks

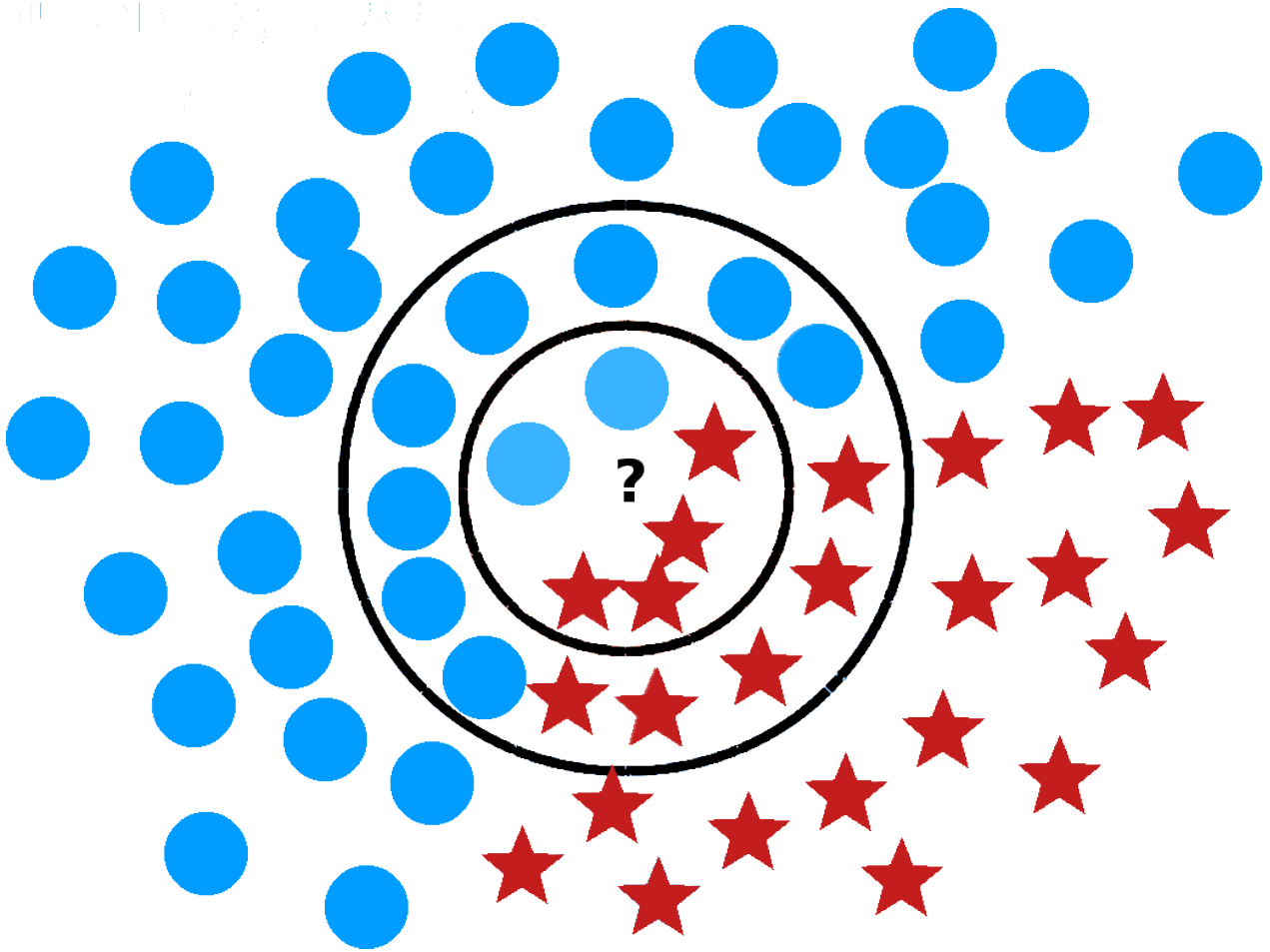
Decision Trees

Naïve Bayes Classifiers

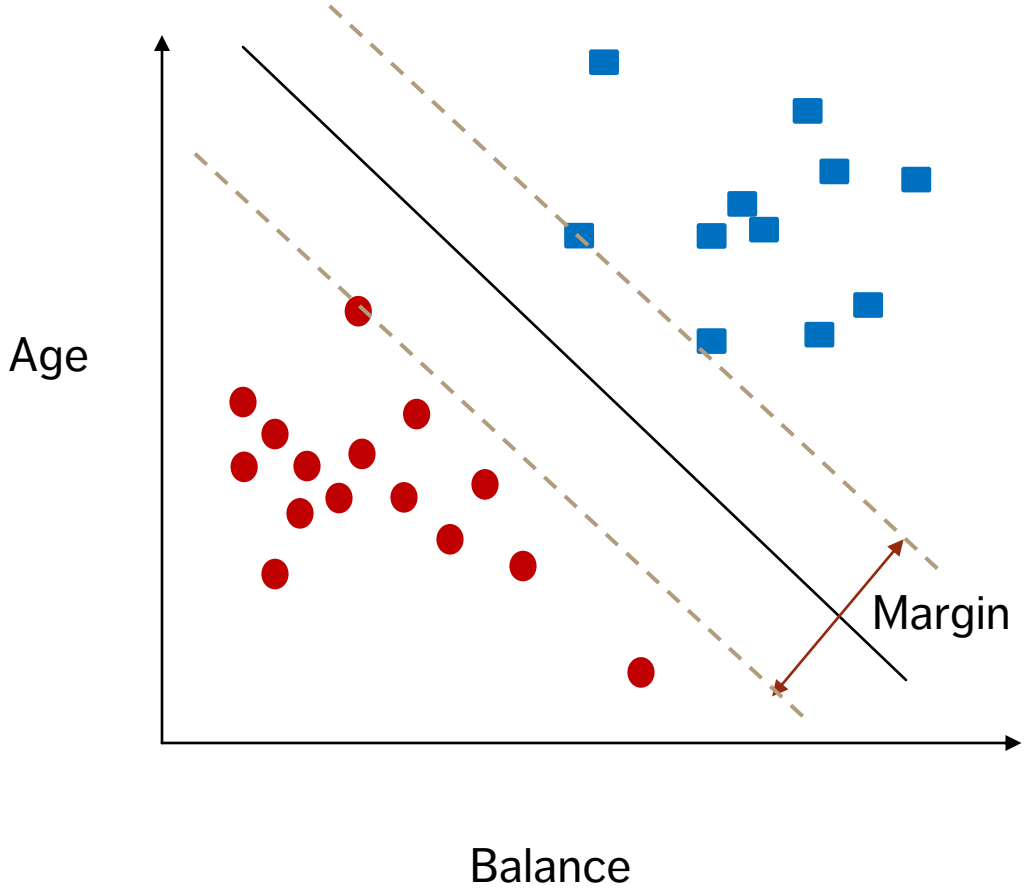
Support Vector Machines

Nearest Neighbours Classifiers

k NEAREST NEIGHBOURS

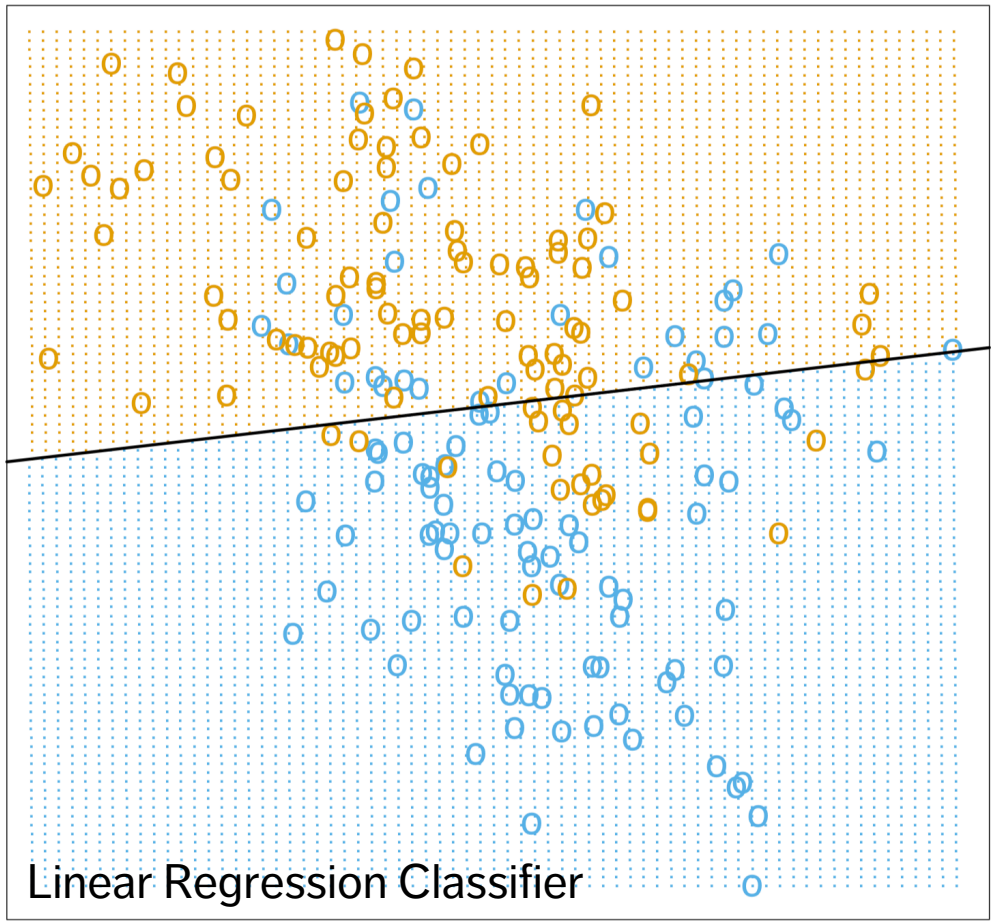


SUPPORT VECTOR MACHINES

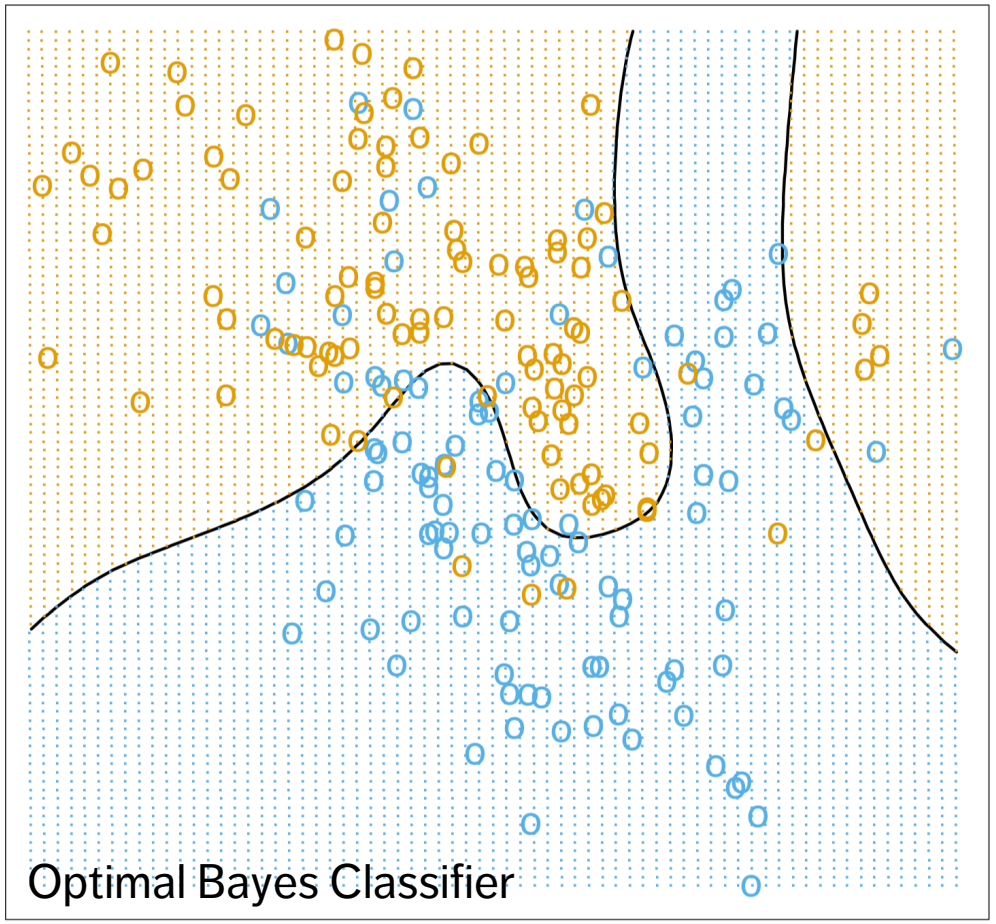


[Images taken from *The Elements of Statistical Learning*]

ILLUSTRATION



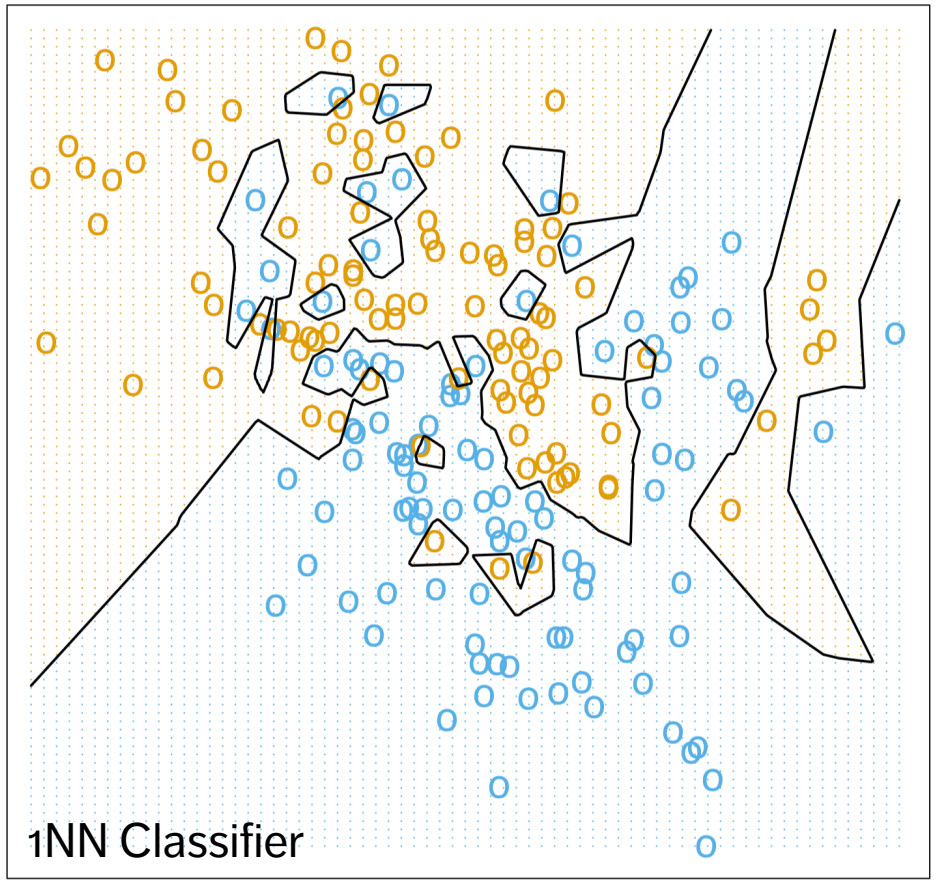
Linear Regression Classifier

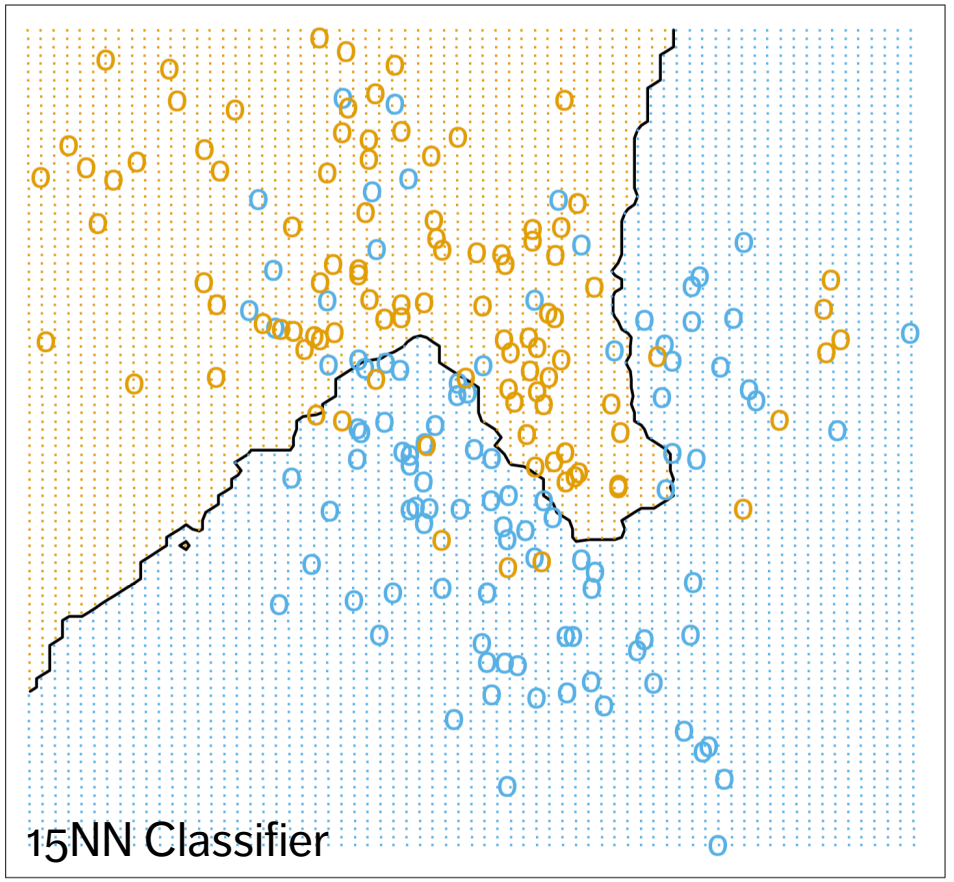


Optimal Bayes Classifier

[Images taken from *The Elements of Statistical Learning*]

ILLUSTRATION

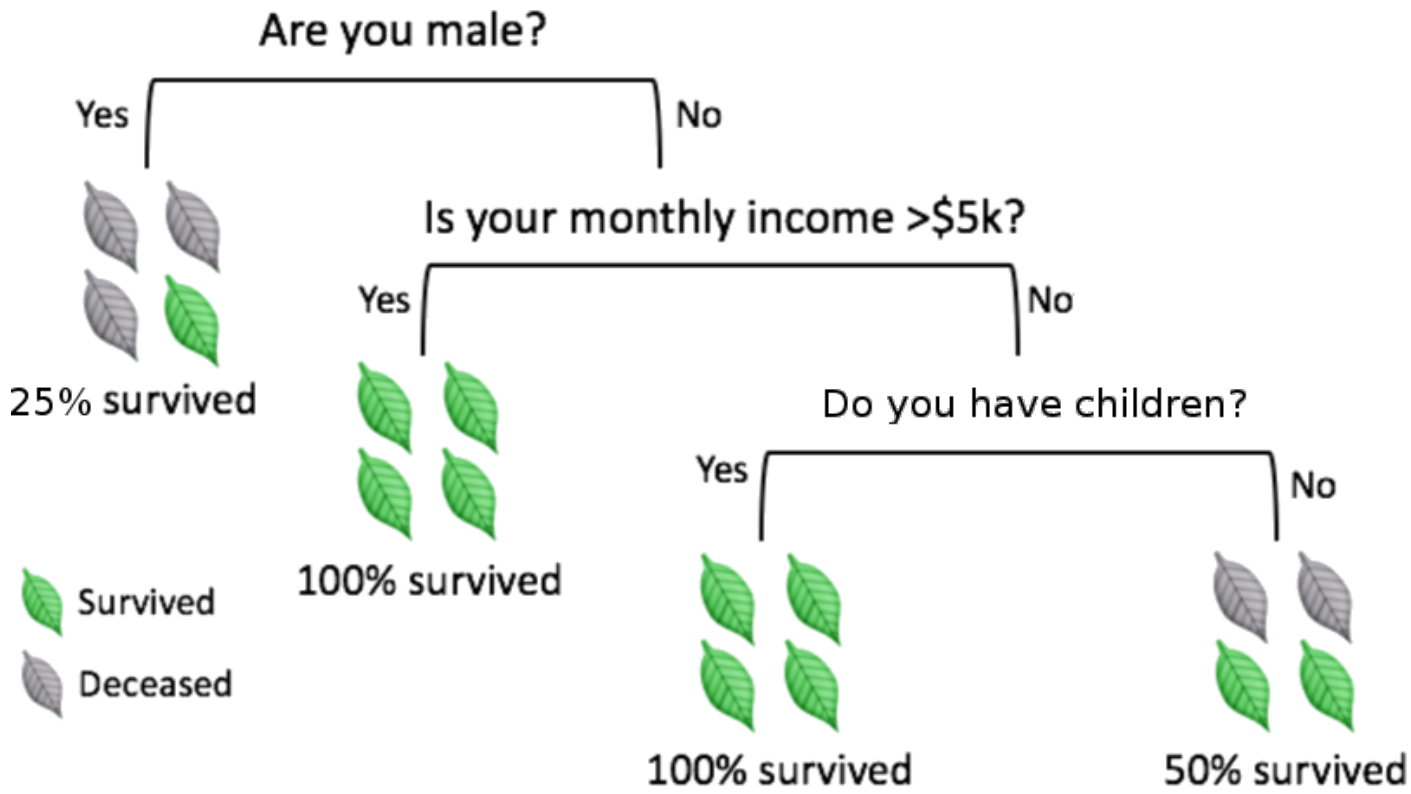




[adapted from A. Ng, K. Soo, *Numsense!*, Surviving a Disaster]

DECISION TREES

Decision trees are perhaps the most **intuitive** of these methods: classification is achieved by following a path up the tree, from its **root**, through its **branches**, and ending at its **leaves**.



DECISION TREES

To make a **prediction** for a new instance, follow the path down the tree, reading the prediction directly once a leaf is reached.

Creating the tree and traversing it might be **time-consuming** if there are too many variables.

Prediction accuracy can be a concern in trees whose growth is **unchecked**. In practice, the criterion of **purity** at the leaf-level is linked to bad prediction rates for new instances.

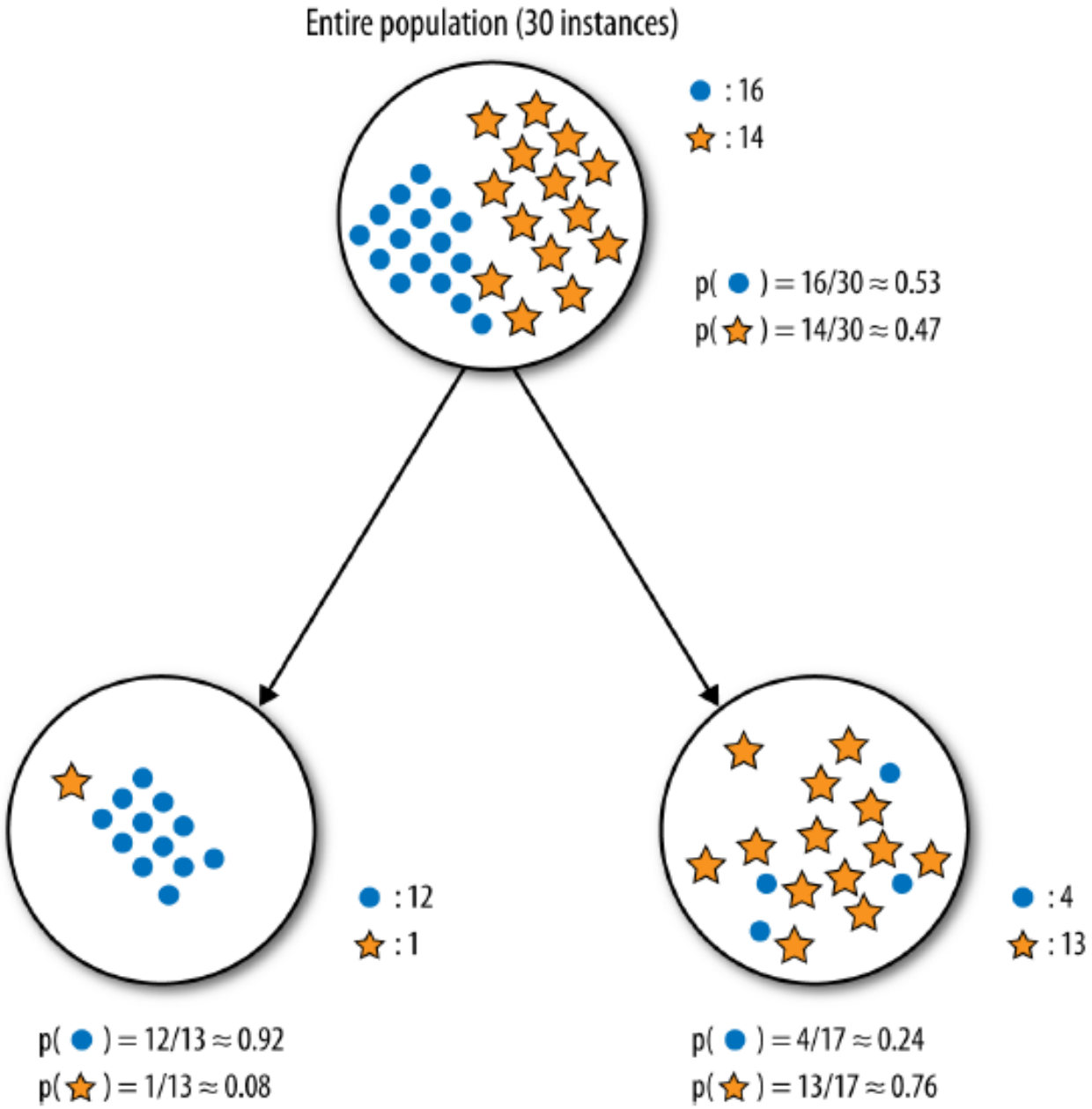
- other criteria are often used to prune trees, which may lead to **impure** leaves (i.e. with non-trivial entropy).

DECISION TREE ALGORITHM (ID3)

Task: grow a decision tree using a training set (a subset of the data for which the correct classification of the target is known).

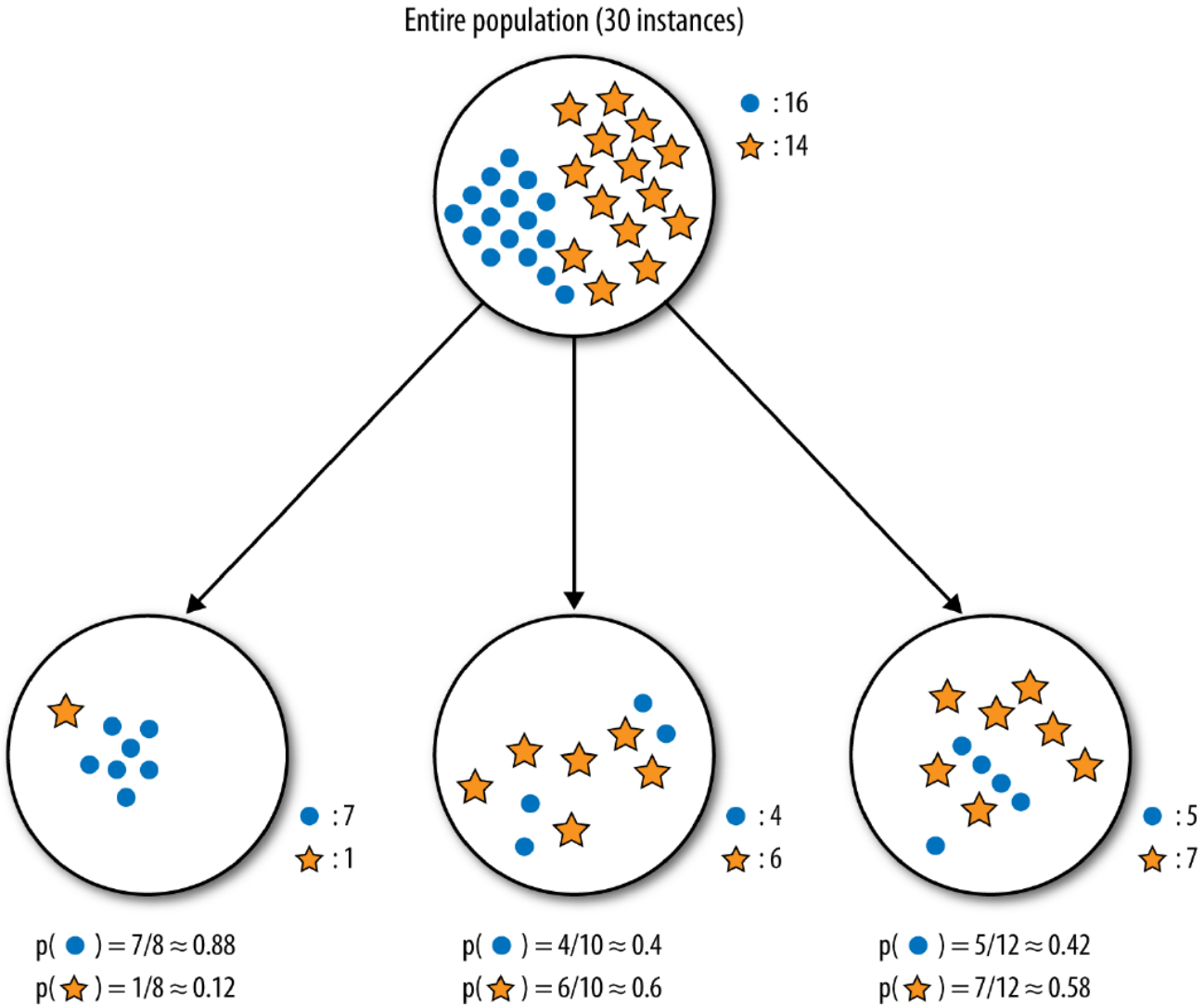
Overview:

1. Split the training data (**parent**) set into (**children**) subsets, using the different levels of a particular attribute
2. Compute the **information gain** for each subset
3. Select the **most advantageous** split
4. Repeat for each node until some **leaf** criterion is met (each item in the leaf has the same classification?)



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$
$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39$$
$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79$$

$$\text{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_R E(R)]$$
$$\approx 0.99 - \frac{1}{30}[13(0.39) + 17(0.79)]$$
$$\approx \mathbf{0.37}$$



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$
$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$
$$E(C) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$
$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$
$$\text{IG} = E(S) - \tfrac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)]$$
$$\approx 0.99 - \tfrac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)]$$
$$\approx \mathbf{0.13}$$

DECISION TREES STRENGTHS AND LIMITATIONS

White box model

Can be used with **incomplete** datasets

Built-in feature selection

Makes **no assumption** about

Not as accurate as other algorithms (usually)

Not robust

Particularly vulnerable to **overfitting**

Optimal decision tree learning is **NP-complete**

Biased towards categorical features with high number of levels

DECISION TREES NOTES

Splitting Metrics

- information gain, Gini impurity, variance reduction, etc.

Common Algorithms

- Iterative Dichotomiser 3, C4.0, C4.5, CHAID, MARS, conditional inference trees, CART

Decision trees can also be combined together using boosting algorithms (**AdaBoost**) or **Random Forests**, providing a type of voting procedure (Ensemble Learning).

OTHER POINTS TO PONDER

Classification is linked to **probability estimation**

- approaches based on regression models could prove fruitful

Rare occurrences (often more interesting or important) continue to plague classification attempts

- historical data at Fukushima's nuclear reactor prior to the meltdown could not have been used to learn about meltdowns

No Free-Lunch Theorem: no classifier works best for all data.

With big datasets, algorithms must also consider efficiency.

PERFORMANCE EVALUATION

Classifiers are evaluated on the testing set.

Ideally, a good classifier would have high rates of both **True Positives** (TP) and **True Negatives** (TN), and low rates of both **False Positives** (FP, Type I error) and **False Negatives** (FN, Type II error).

Evaluation metrics mean very little on their own: context requires comparison with other classifiers, and other evaluation metrics.

		Predicted			Total
		A	B		
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
		60	21	81	
Total		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted			Total
		A	B		
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
		70	11	81	
Total		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

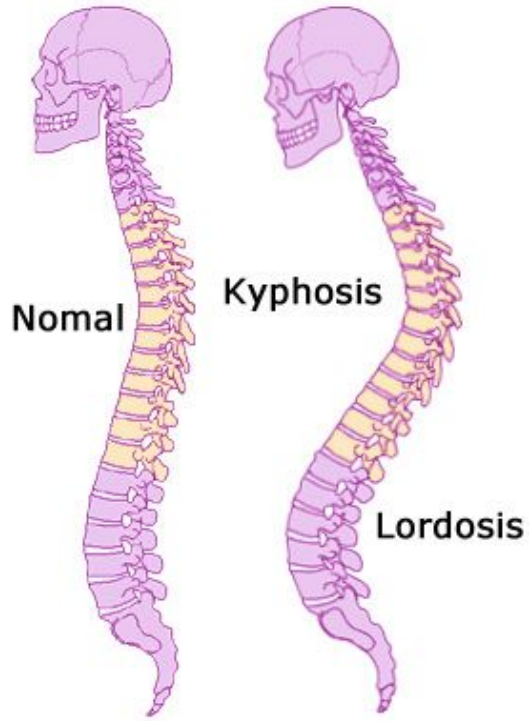
Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

EXAMPLE – KYPHOSIS DATASET

Kyphosis is a medical condition related to the excessive convex curvature of the spine. Corrective spinal surgery is at times performed on children.

The dataset has 81 observations and 4 attributes:

- **kyphosis** (absent or present after operation)
- **age** (at time of operation, in months)
- **number** (of vertebrae involved)
- **start** (topmost vertebra operated on)

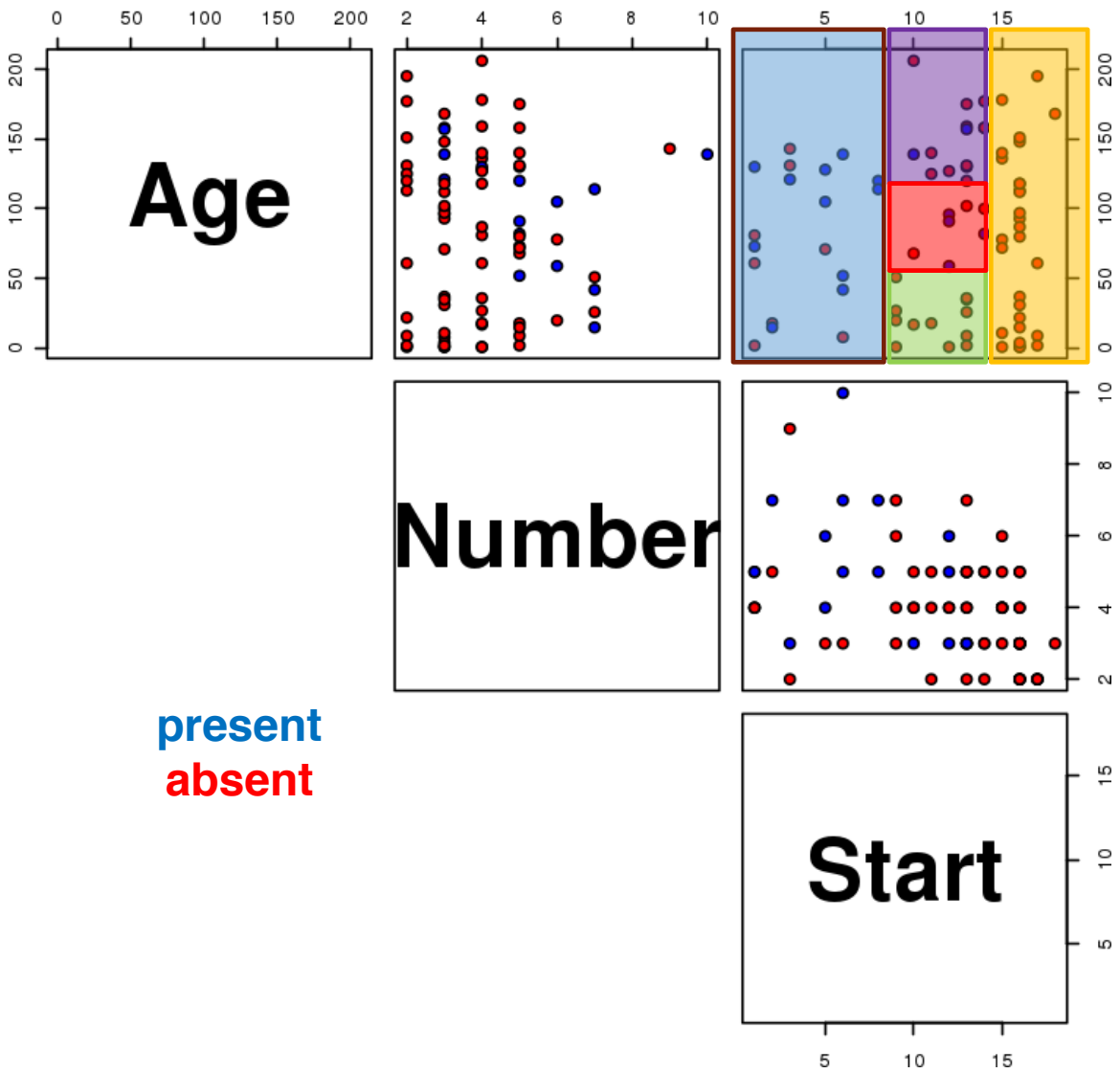


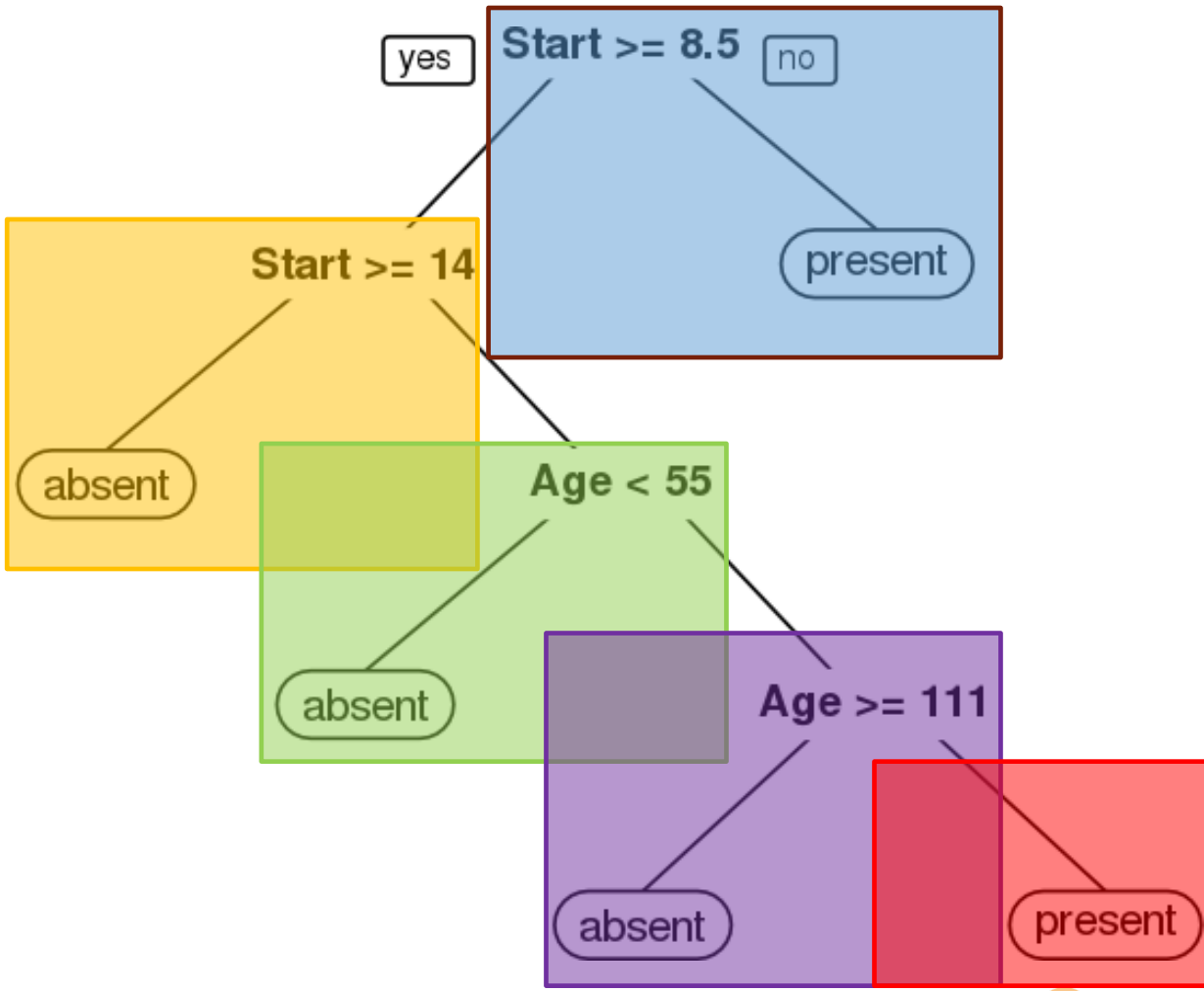
EXAMPLE – KYPHOSIS DATASET

The question of interest for this natural dataset is how the three explanatory attributes might impact the operation's success.

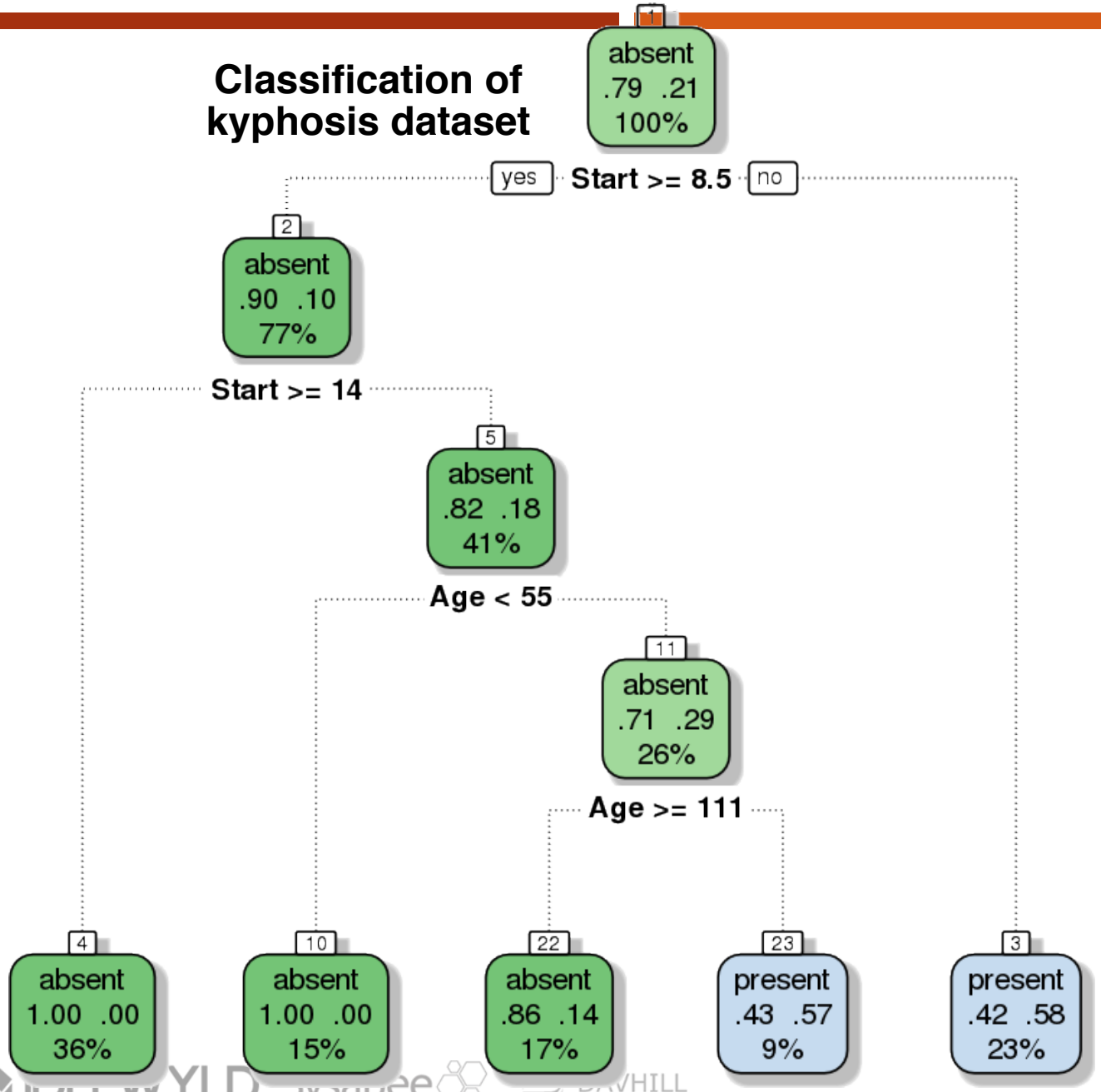
We use the rpart implementation of CART to generate candidate decision trees.

Strictly speaking, this is not a predictive supervised task as we treat the entire dataset as a training set (there are no hold-out testing observations for the time being).

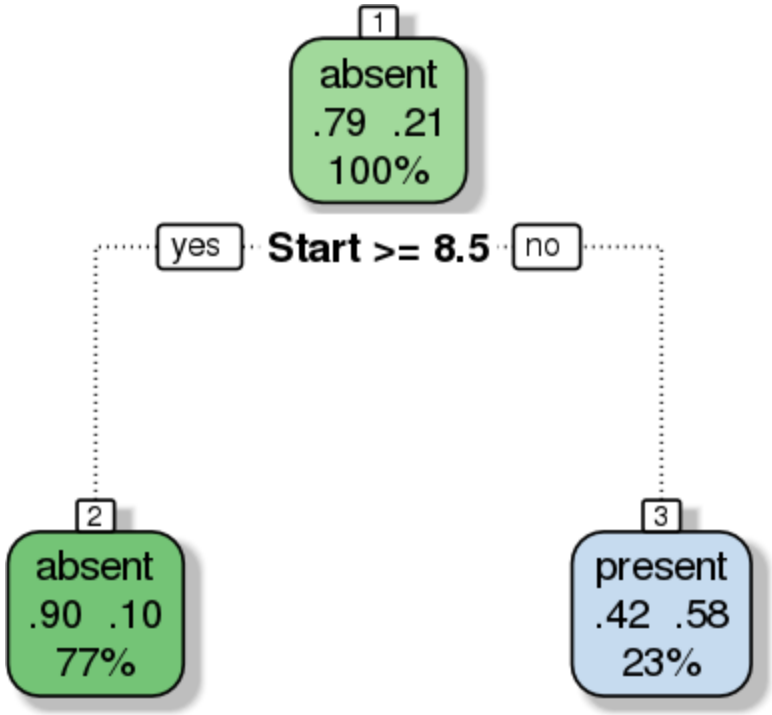

Age
Number
Start
present
absent


yes
Start >= 8.5
no
present
Start >= 14
absent
Age < 55
absent
Age >= 111
absent
present

Classification of kyphosis dataset



Pruned classification of kyphosis dataset



EXAMPLE – KYPHOSIS DATASET

We train a model on 50 observations (selected randomly) and evaluate the performance on the remaining 31 observations.

		Predicted A	Predicted B	Total	
Actuals	A	23	3	26	83.9%
	B	3	2	5	16.1%
Total		26	5	31	
		83.9%	16.1%		

Classification Rates	
Sensitivity:	0.88
Specificity:	0.40
Precision:	0.88
Negative Predictive Value:	0.40
False Positive Rate:	0.60
False Discovery Rate:	0.12
False Negative Rate:	0.12

Performance Metrics	
Accuracy:	0.81
F1-Score:	0.88
Informedness (ROC):	0.28
Markedness:	0.28
M.C.C.:	0.28
Pearson's chi2:	0.00
Hist. Stat:	0.00

Is this a good model?

PERFORMANCE EVALUATION

In both the categorical and numerical estimation problem, isolated performance metric does not provide enough of a rationale for model validation, unless it has first been normalized.

There is (a lot) more to be said on the topic of model selection.