CLASSIFICATION AND VALUE ESTIMATION

ADVANCED DATA SCIENCE TRAINING I

"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)

LEARNING OBJECTIVES

CASE STUDY: MINNESOTA TAX AUDIT

CLASSIFICATION AND VALUE ESTIMATION

Data mining based tax audit selection: a case study of a pilot project at the Minnesota Department of Revenue

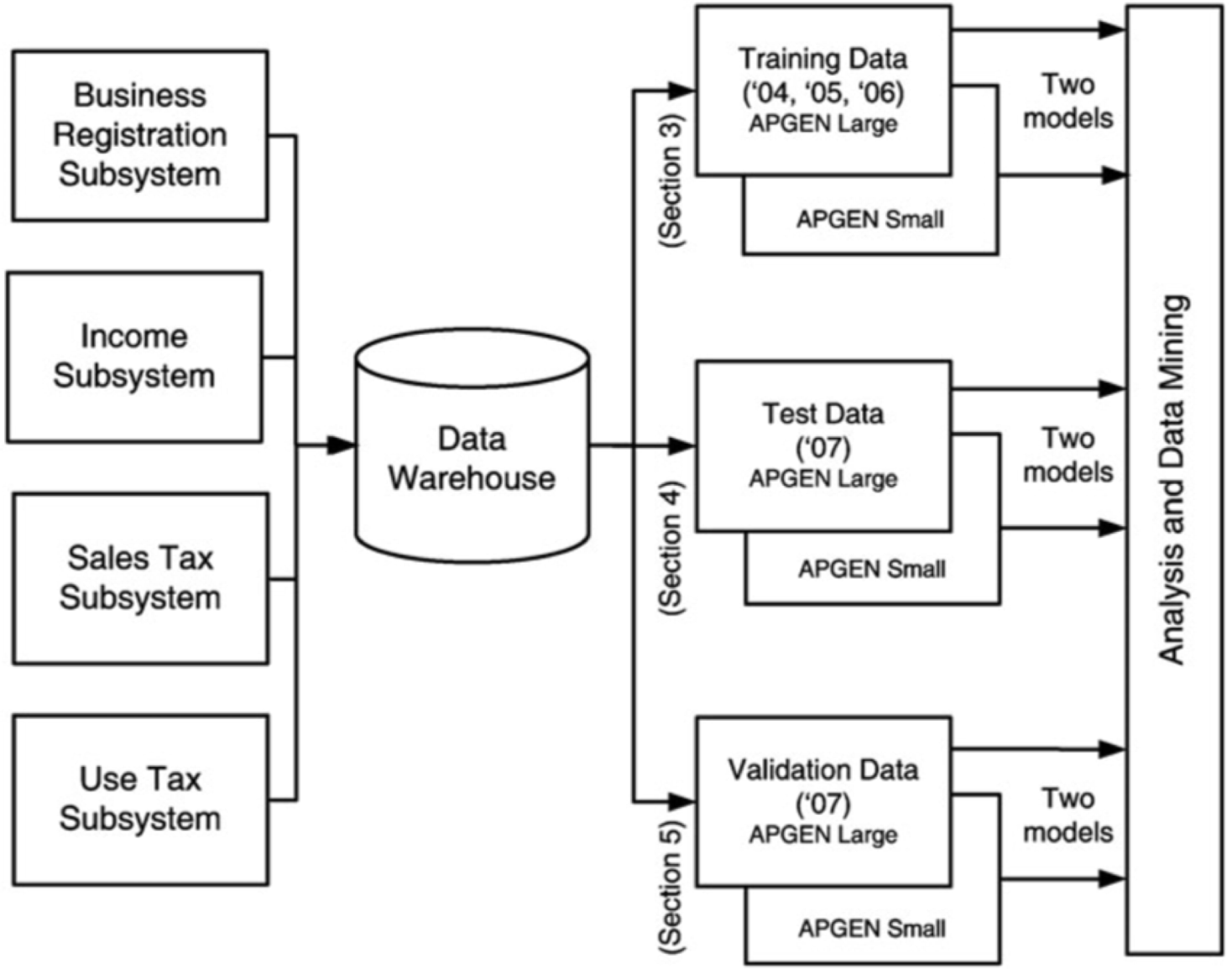
(Hsu, W., Pathak, N., Srivatsava, J., Tschida, Bjorklund, E. [2013], *Real Word Data Mining Applications*, *Annals of Information Systems,* v.17, Springer).

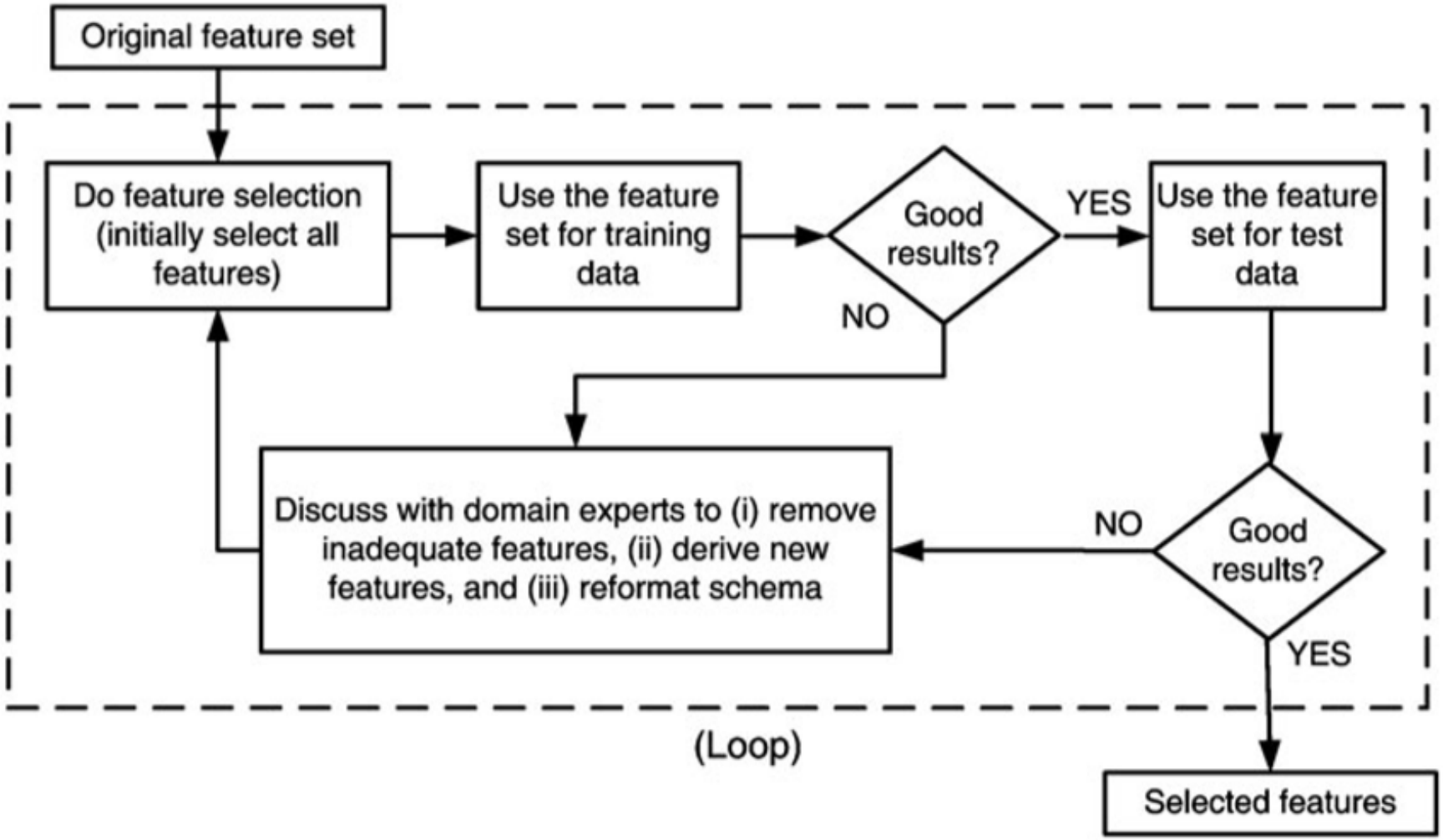
CONTEXT

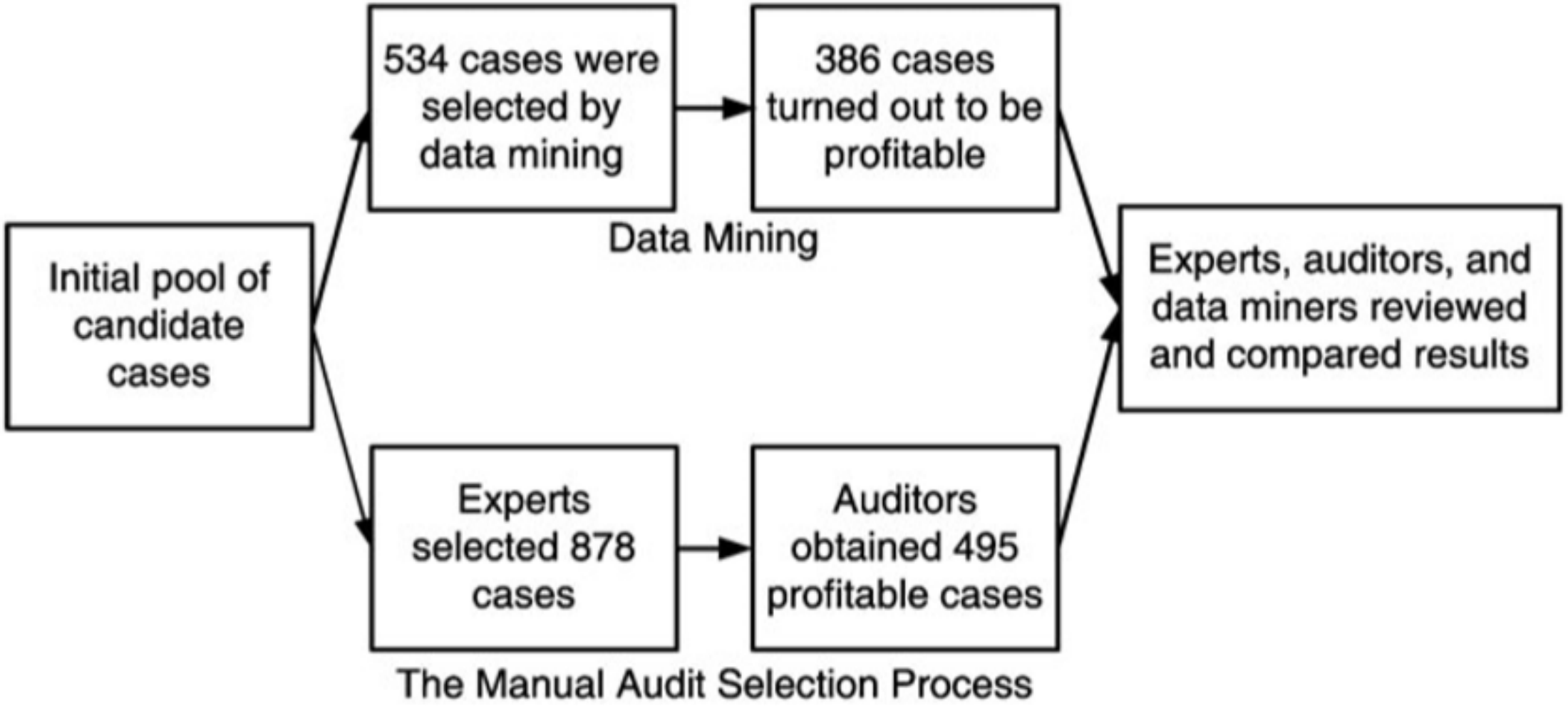
Large gaps between revenue owed (in theory) and revenue collected (in practice) are problematic for governments.

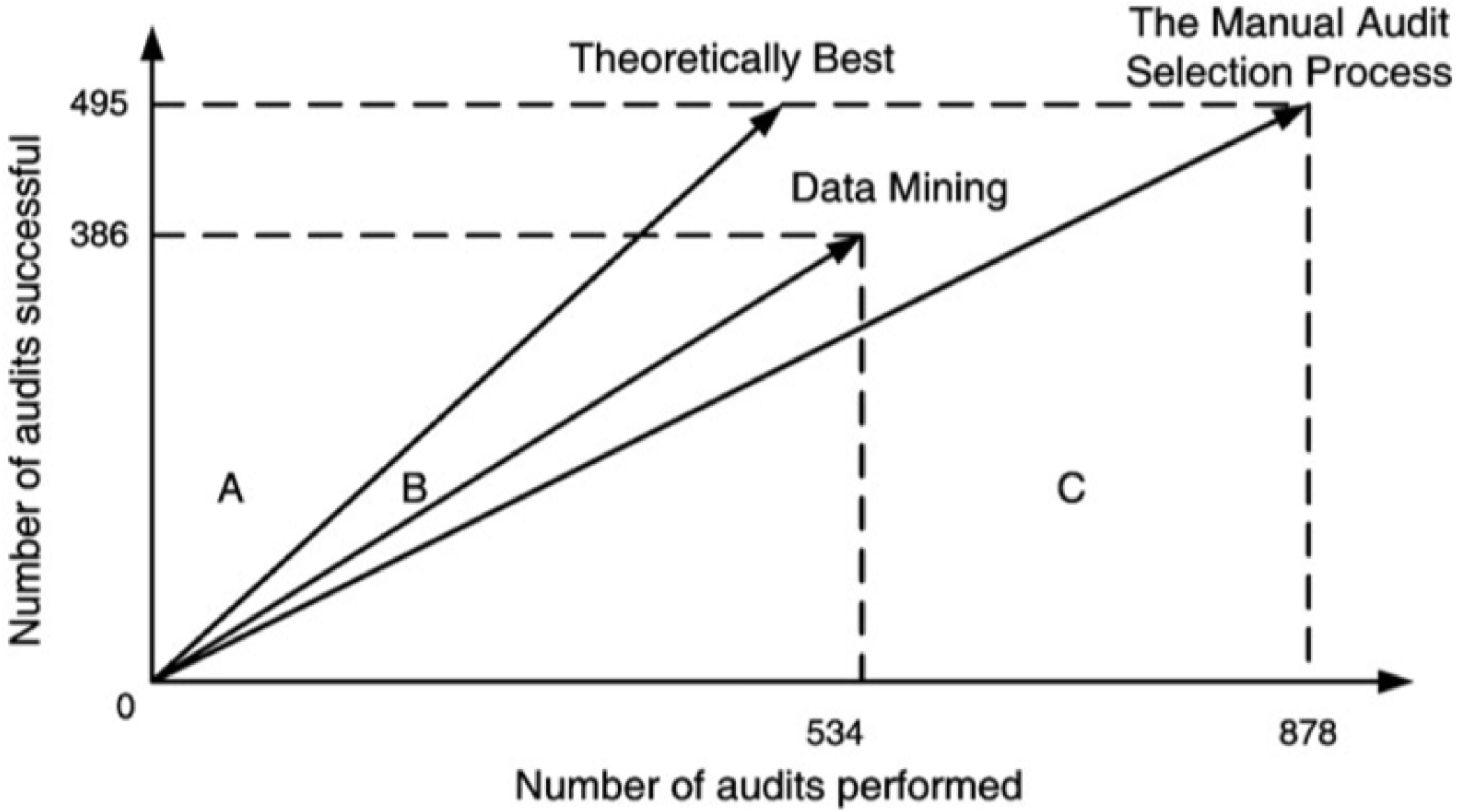
Revenue agencies implement various fraud detection strategies (such as audit reviews) to bridge that gap.

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


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


Original feature set
Do feature selection (initially select all features)
Use the feature set for training data
Good results?
YES
NO
Use the feature set for test data
Discuss with domain experts to (i) remove inadequate features, (ii) derive new features, and (iii) reformat schema
NO
Good results?
YES
(Loop)
Selected features


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


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

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = $5,577,431 (83.6 %) C = $177,560 (44 %)	109 (Use tax lost) R = $925,293 (13.9 %) C = $50,140 (12.4 %)
Actually bad	148 (costs wasted) R = $72,744 (1.1 %) C = $68,080 (16.9 %)	235 (costs saved) R = $98,105 (1.4 %) C = $108,100 (26.7 %)

DISCUSSION

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

CLASSIFICATION BASICS

CLASSIFICATION AND VALUE ESTIMATION

"We are much too much inclined in these days to divide people into permanent categories, forgetting that a category only exists for its special purpose and must be forgotten as soon as that purpose is served."

(Dorothy L. Sayers, *Are Women Human? Astute and Witty Essays on the Role of Women in Society*)

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

OTHER USES

Predicting that an object belongs to a particular class.

Organizing and grouping instances into categories.

Enhancing the detection of relevant objects

- avoidance: “this object is an incoming vehicle”
- pursuit: “this borrower is unlikely to default on her mortgage”
- degree: “this dog is 90% likely to live until it’s 7 years old”

In the absence of testing data, classification may be **descriptive** but not predictive.

EXAMPLES

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?

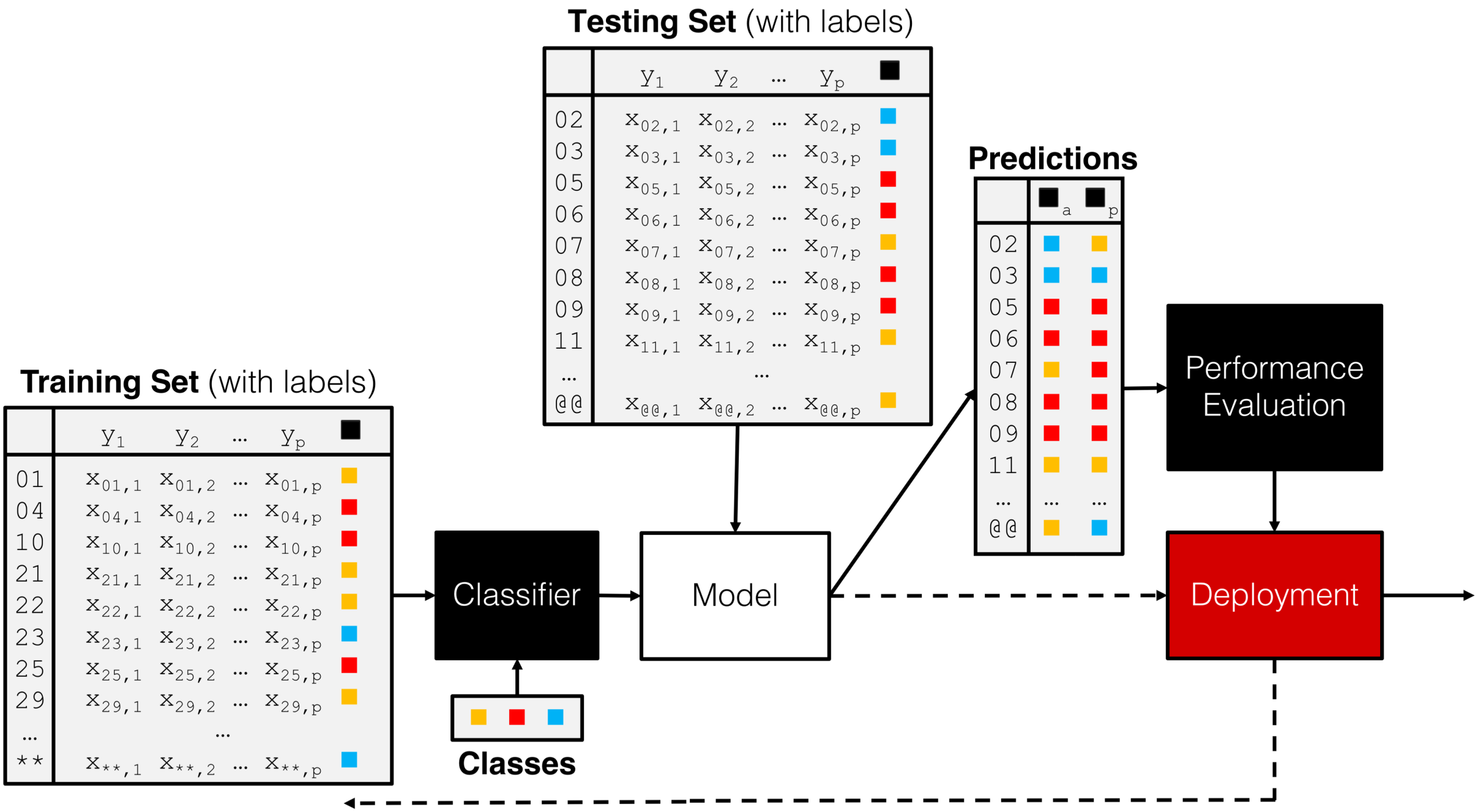
EXAMPLES

Scenario:

Customers who make a large number of calls to a mobile phone company's customer service number have been identified as churn risks. The company is interested in reducing said churn.

Questions: can we predict

- the overall lifetime value of a customer?
- which customers are more likely to churn in the near future?
- what retention offer a particular customer will best respond to?


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

EXERCISE

How would you use standard statistical modeling techniques to answer these questions?

CLASSIFICATION ALGORITHMS

CLASSIFICATION AND VALUE ESTIMATION

"The diversity of problems that can be addressed by classification algorithms is significant, and covers many domains. [...] It is difficult to comprehensively discuss all the methods in a single book."

(C.C. Aggarwal, *Data Classification: Algorithms and Applications*)

CLASSIFICATION SCHEMES

Logistic Regression

- classical model
- affected by variance inflation and variable selection process

Neural Networks

- hard to interpret
- requires all variables to be of the same type
- easier to train since backpropagation (chain rule)

Decision Trees

- may overfit the data if not pruned correctly (manually?)

CLASSIFICATION SCHEMES

Naïve Bayes Classifiers

- quite successful for text mining applications (spam filter)
- assumptions not often met in practice

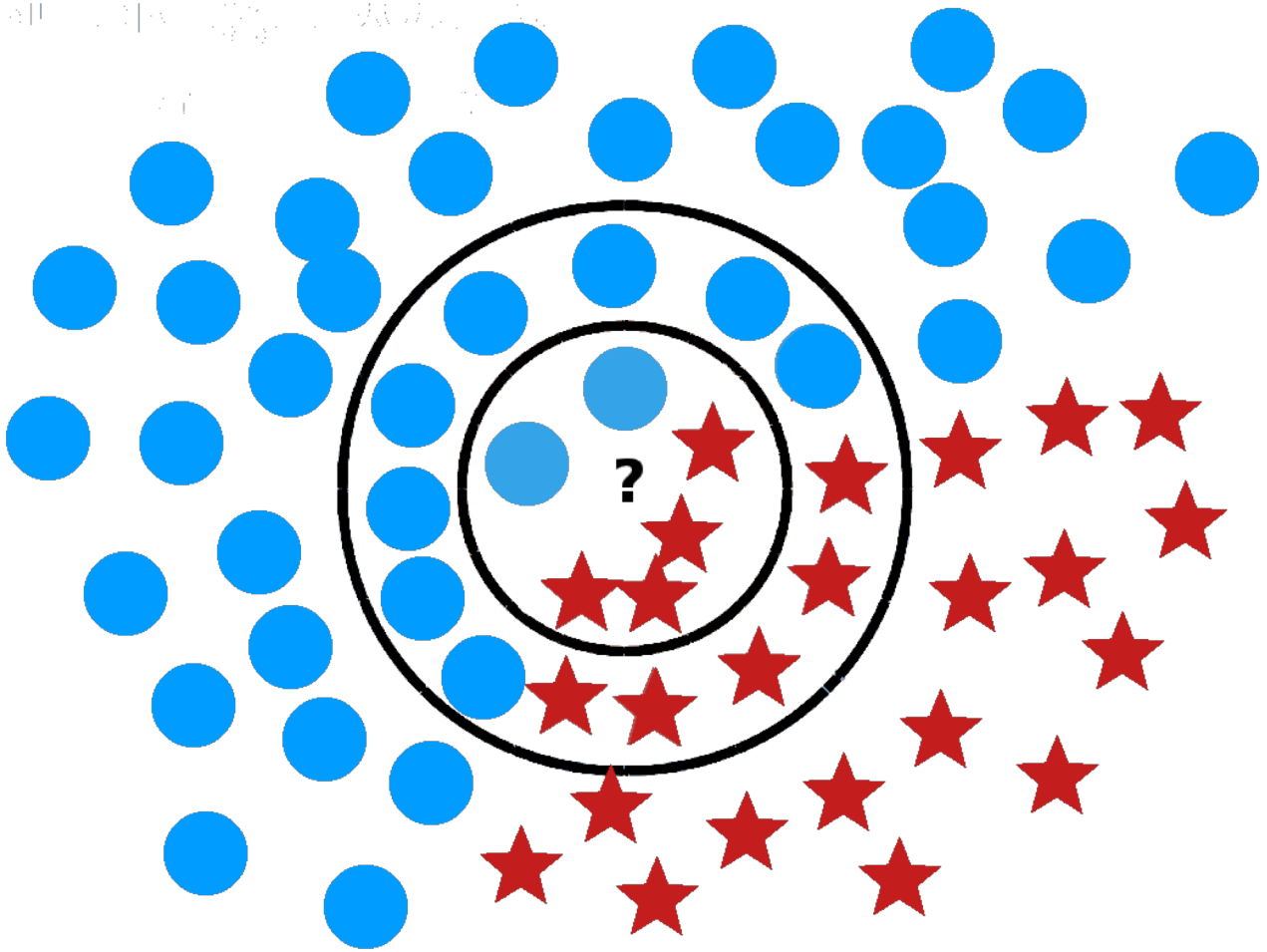
Support Vector Machines

- may be difficult to interpret (non-linear boundaries)
- can help mitigate big data difficulties

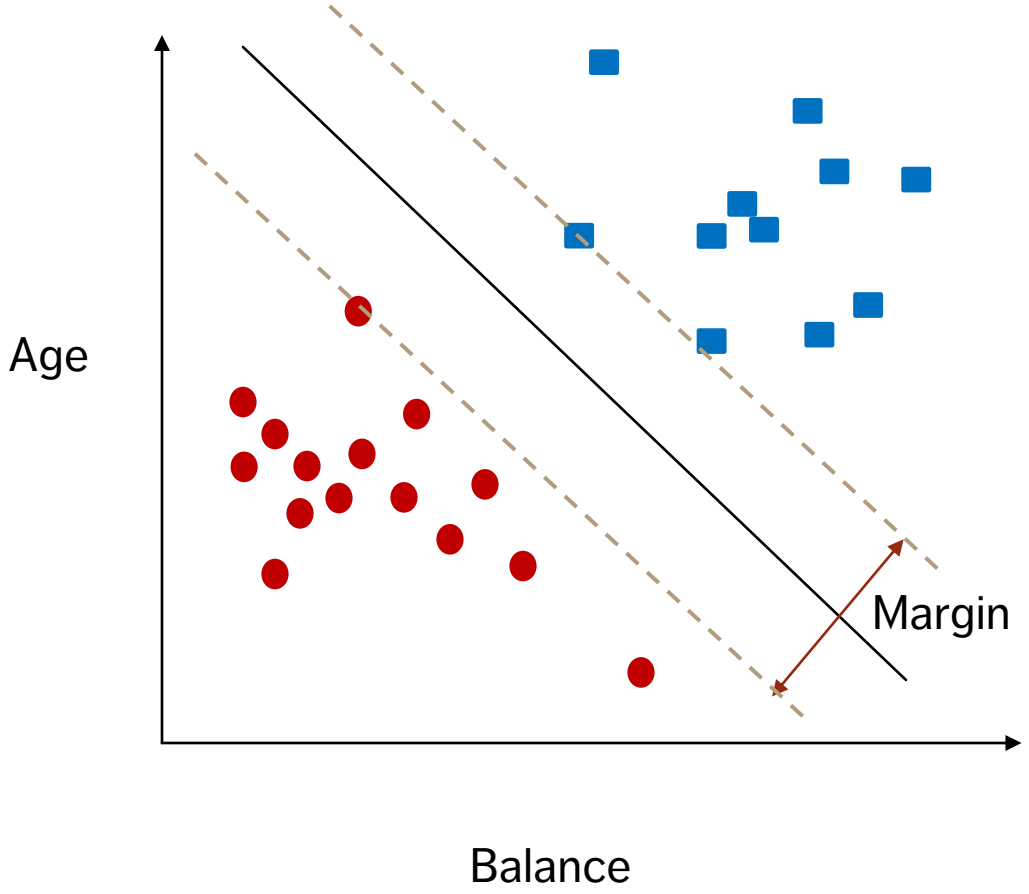
Nearest Neighbours Classifiers

- require very little assumptions about the data
- not very stable (adding points may substantially modify the boundary)

k NEAREST NEIGHBOURS

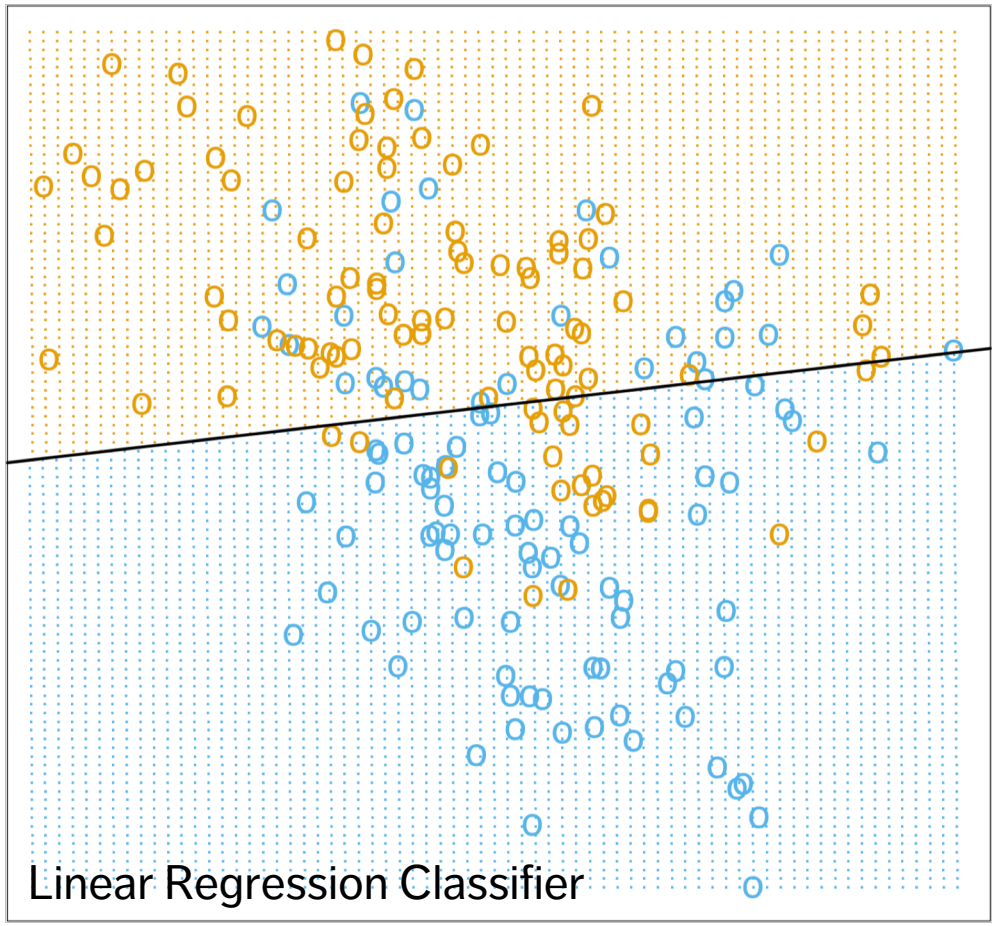


SUPPORT VECTOR MACHINES

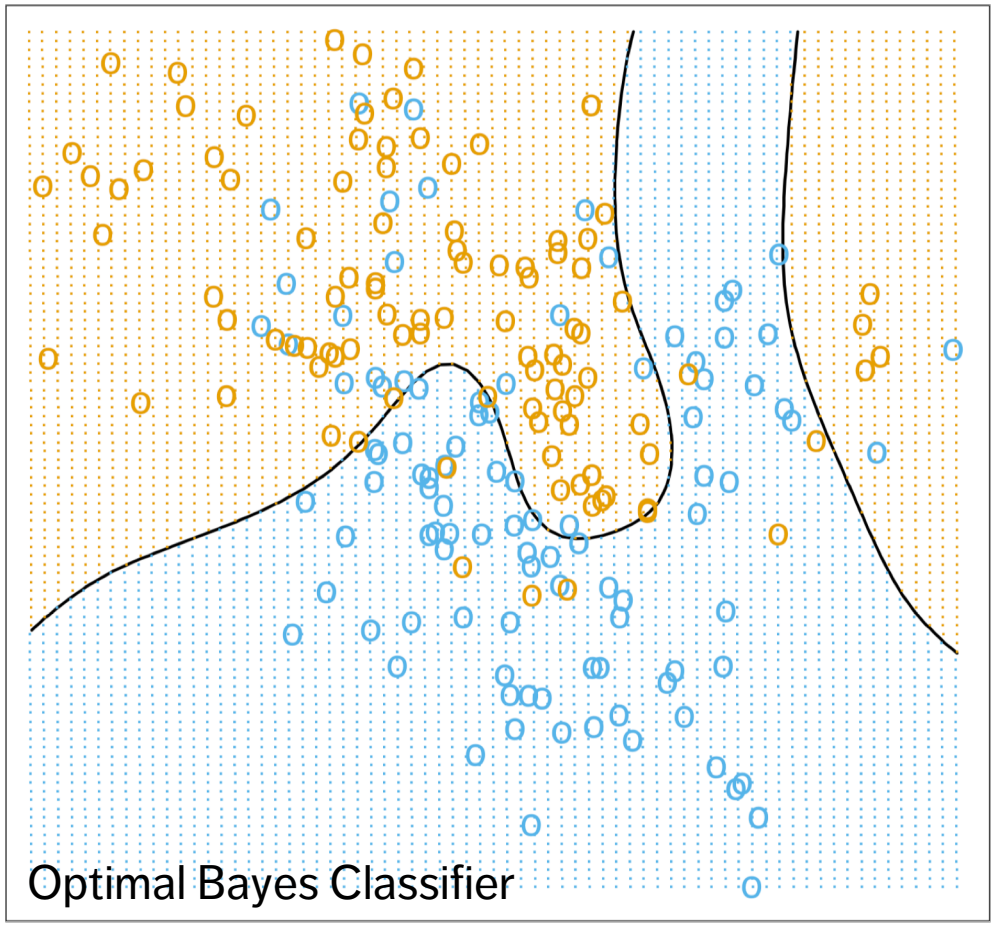


[Images taken from *The Elements of Statistical Learning*]

ILLUSTRATION



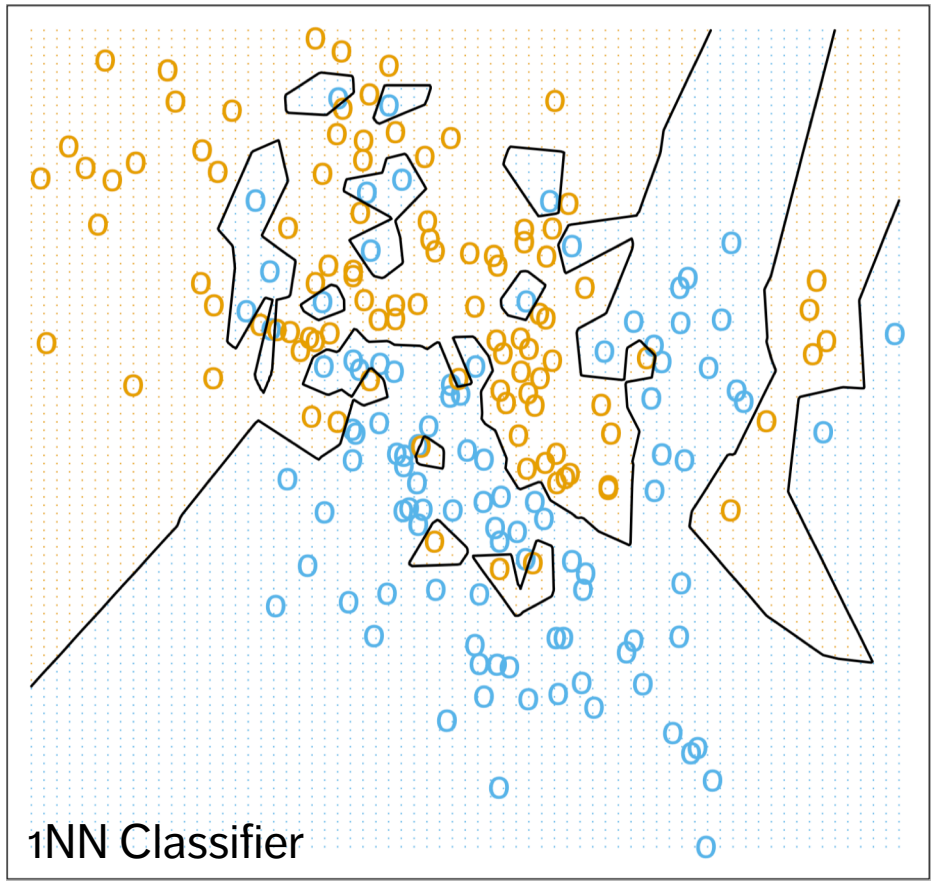
Linear Regression Classifier

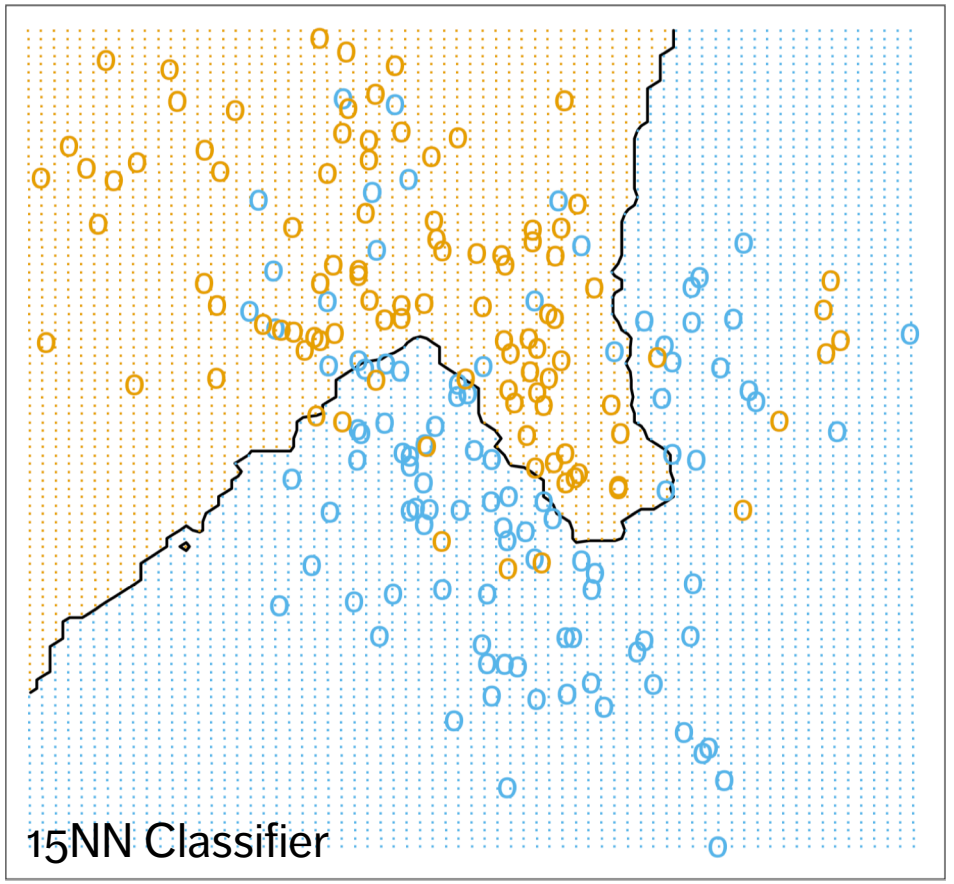


Optimal Bayes Classifier

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

ILLUSTRATION

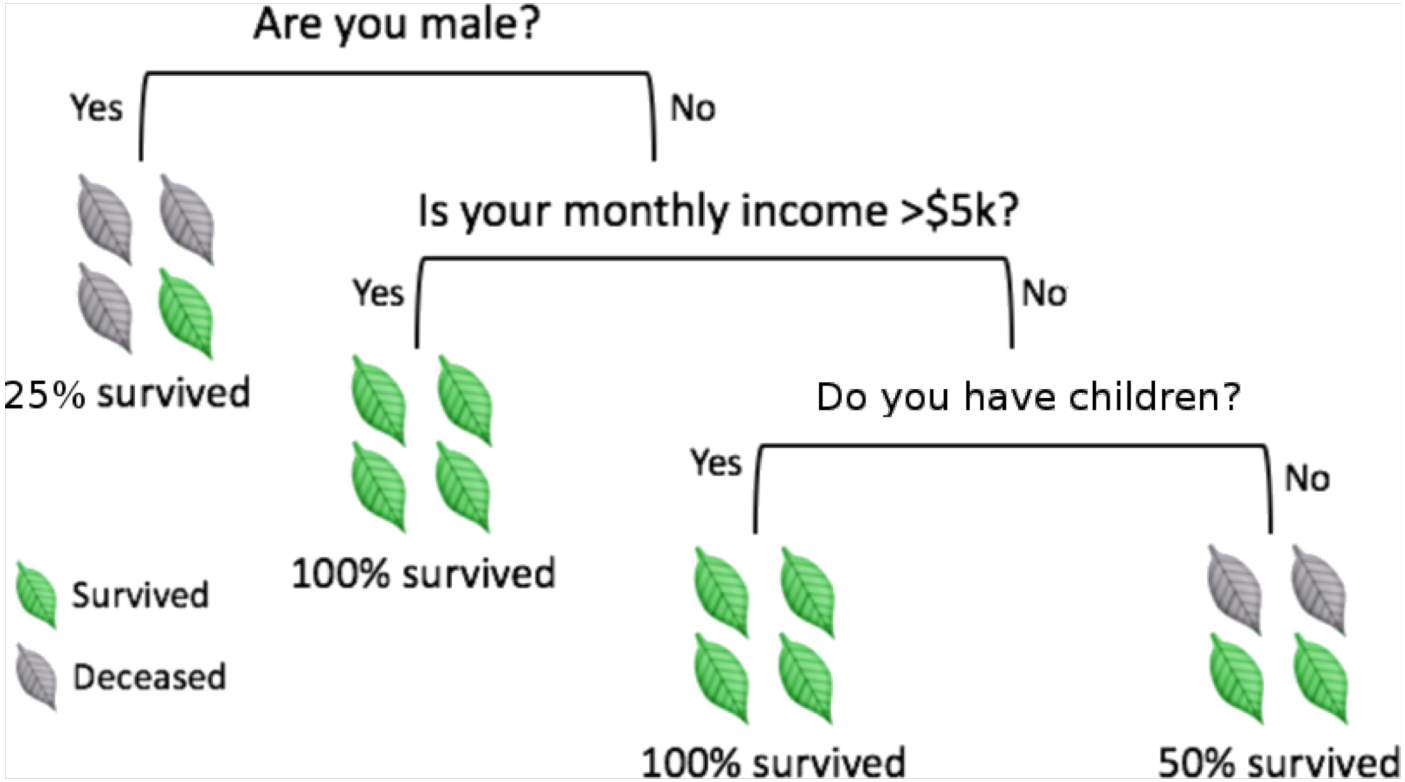




[adapted from A. Ng, K. Soo, *Numsense!*, Surviving a Disaster]

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.

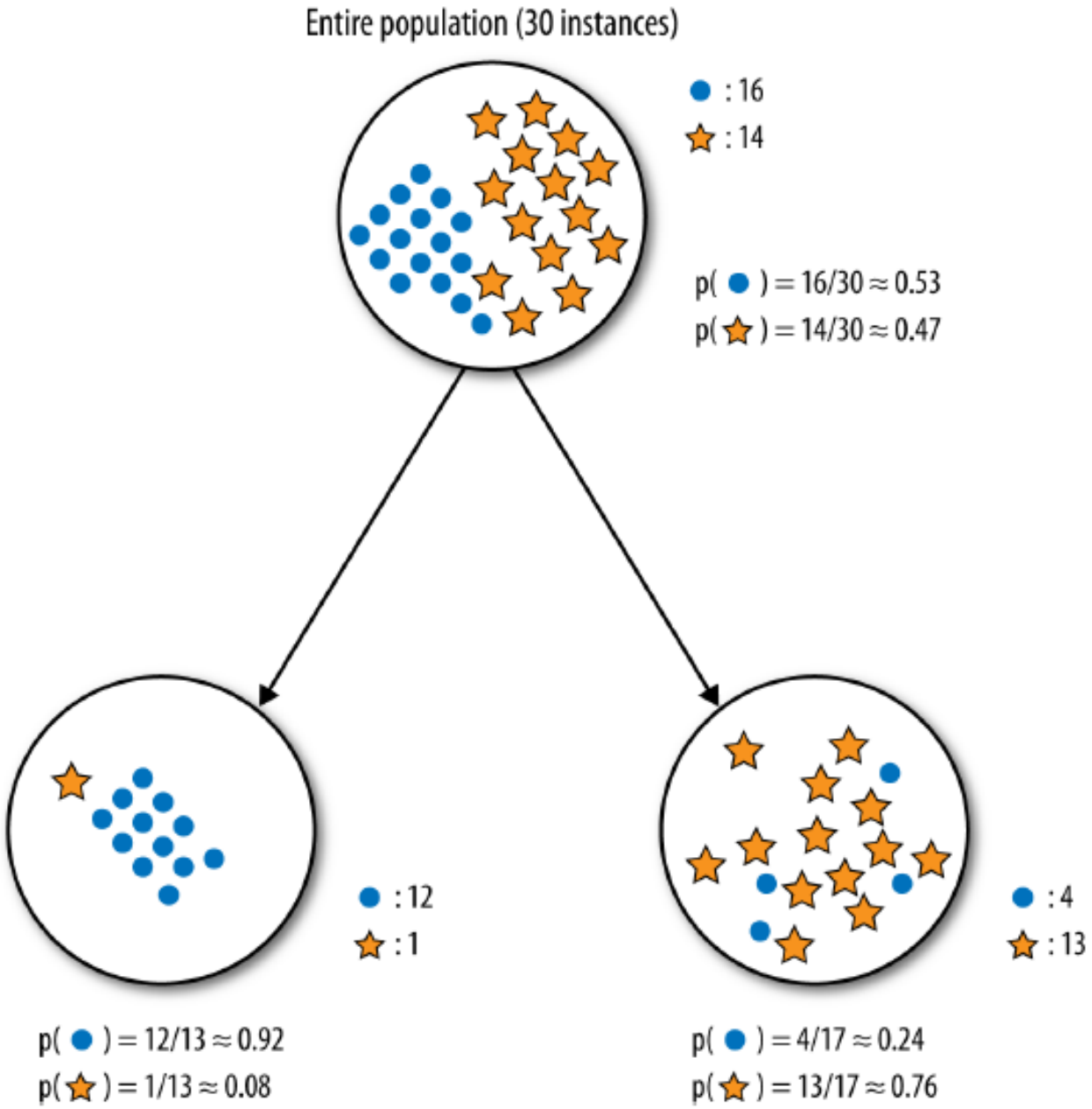
- other criteria are often used to prune trees, which may lead to **impure** leaves (i.e. with non-trivial entropy).

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?)



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39$$

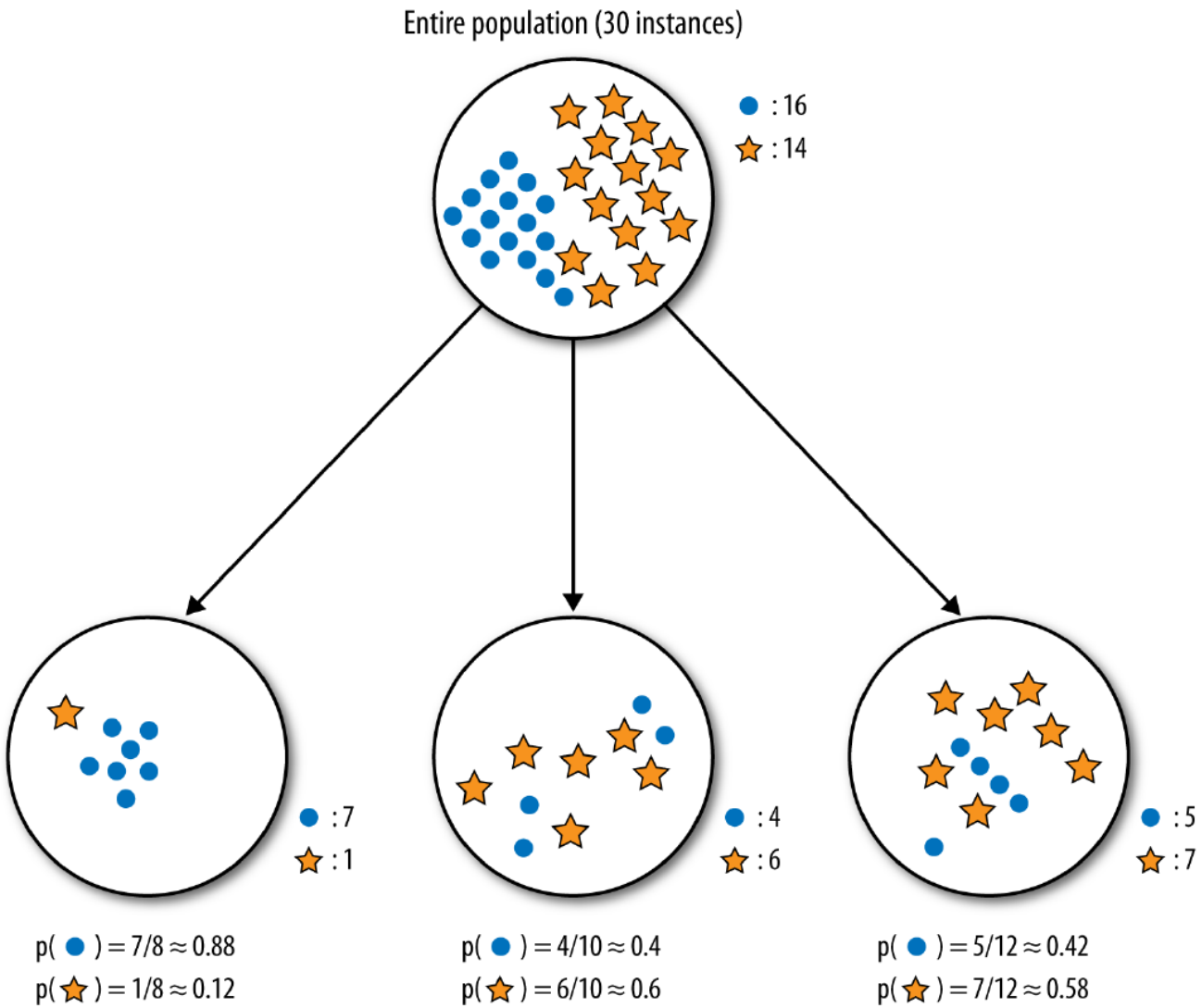
$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79$$

$$\text{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_R E(R)]$$

$$\approx 0.99 - \frac{1}{30}[13(0.39) + 17(0.79)]$$

$$\approx \mathbf{0.37}$$



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$

$$E(C) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$

$$= -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$

$$\mathrm{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)]$$

$$\approx 0.99 - \frac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)]$$

$$\approx \mathbf{0.13}$$

DISCUSSION

What split is most advantageous?

In what way does the choice of the algorithm(s) depend on the available data and data types, and on the prediction objective?

EXERCISE

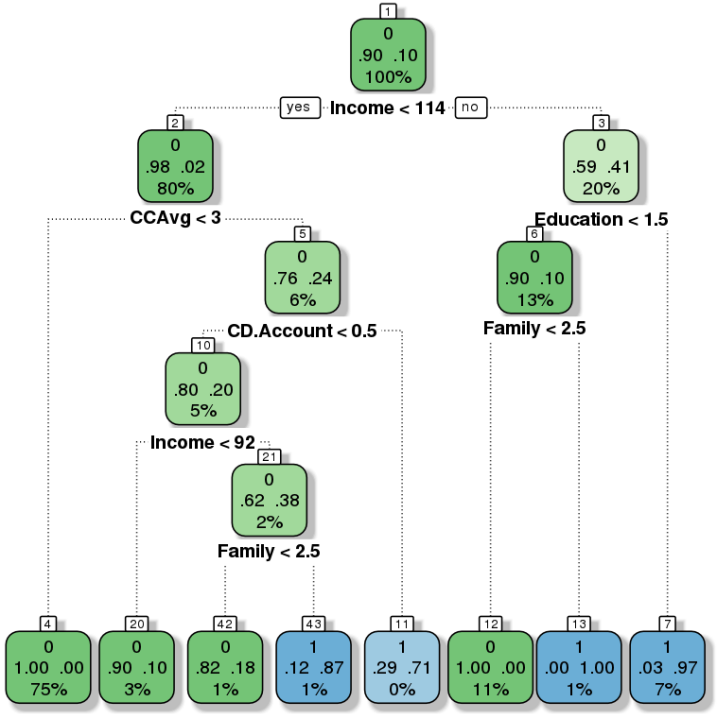
UniversalBank is looking at converting its **liability** customers (i.e., customers who only have deposits at the bank) into **asset** customers (i.e., customers who have a loan with the bank). In a previous campaign, *UniversalBank* was able to convert 9.6% of 5000 of its liability customers into asset customers. The marketing department would like to understand what combination of factors make a customer more likely to accept a personal loan, in order to better design the next conversion campaign.

UniversalBank's dataset contains data on 5000 customers, including the following measurements: age, years of professional experience, yearly income (in thousands of USD), family size, value of mortgage with the bank, whether the client has a certificate of deposit with the bank, a credit card, etc.

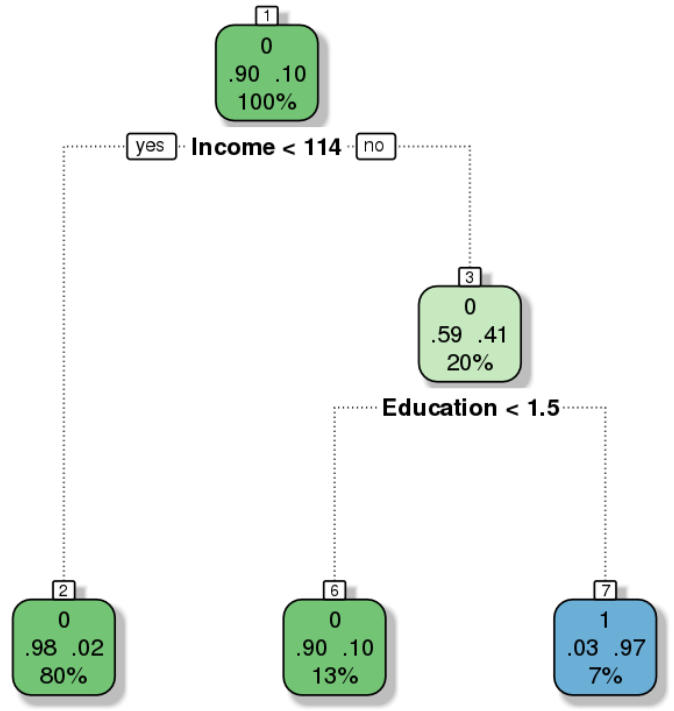
EXERCISE

We build 2 decision trees on a training subset of 3000 records to predict whether a customer is likely to accept a personal loan (1) or not (0).

Tree *A*



Tree *B*



EXERCISE

Explore the `UniversalBank.csv` dataset. Can you come up with a reasonable guess as to what each of the variables represent?

1. How many variables are used in the construction of tree A? Of tree B?
2. Is the following decision rule valid or not for tree A:
   IF (Income ≥ 114) AND (Education ≥ 1.5)
   THEN (Personal Loan $= 1$)?
3. Is the following decision rule valid or not for tree B:
   IF (Income < 92) AND (CCAvg ≥ 3) AND (CD.Account < 0.5)
   THEN (Personal Loan $= 0$)?

EXERCISE

4. What prediction would tree A make for a customer with:
   - yearly income of 94,000$USD (Income = 94),
   - 2 kids (Family = 4),
   - no certificate of deposit with the bank (CD.Account = 0),
   - a credit card interest rate of 3.2% (CCAvg = 3.2), and
   - a graduate degree in Engineering (Education = 3).
5. What about tree B?

NOTES AND VALIDATION

CLASSIFICATION AND VALUE ESTIMATION

"Random trees, and forests, and jungles, oh my!!"

(unknown)

DECISION TREES STRENGTHS

White box model

- predictions can always be explained by following the appropriate paths

Can be used with **incomplete** datasets

Built-in feature selection

- less relevant features don't tend to be used as splitting features

Makes **no assumption** about

- independence, constant variance, underlying distributions, co-linearity

DECISION TREES LIMITATIONS

Not as accurate as other algorithms (usually)

Not robust: small changes in the training dataset can lead to a completely different tree, with a completely different predictions

Particularly vulnerable to **overfitting** in the absence of pruning

- pruning procedures are typically convoluted

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.

PERFORMANCE EVALUATION

sensitivity = $TP/(TP + FN)$

specificity = $TN/(FP + TN)$

precision = $TP/(TP + FP)$

recall = $TP/(TP + FN)$

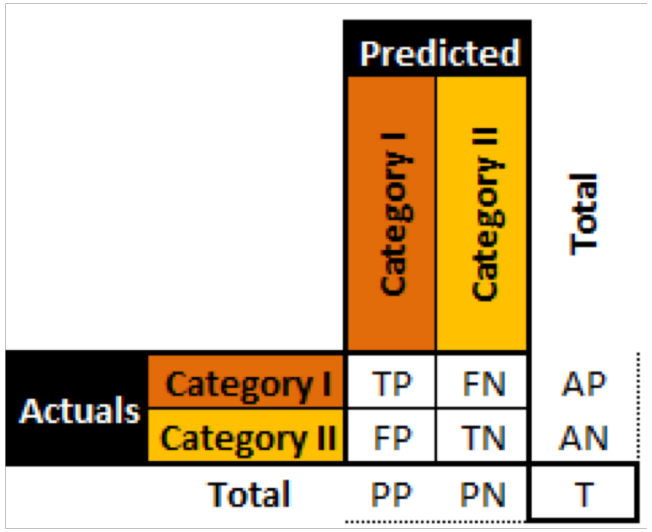
negative predictive value = $TN/(TN + FN)$

false positive rate = $FP/(FP + TN)$

false discovery rate = $FP/(FP + TP)$

false negative rate = $FP/(FN + TP)$

accuracy = $(TP + TN)/T$



Other metrics:

F_1-score, ROC AUC, informedness, markedness, Matthews' Correlation Coefficient (MCC), etc.

		Predicted			Total
		A	B		
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		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%
Total		70	11	81	
		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

DISCUSSION

How many predictive models should be built to deal with supervised tasks, in general?

EXERCISE

(*UniversalBank*, continued)

The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown here.

Tree A

		Predicted		Total	
		A	B		
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Tree B

		Predicted		Total	
		A	B		
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		

EXERCISE

6. Using the appropriate matrices, compute the 9 performance evaluation metrics for each of the trees (on the testing set).
7. If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should *UniversalBank*'s marketing group use to help maintain good customer relations?

EXAMPLE: KYPHOSIS

CLASSIFICATION AND VALUE ESTIMATION

"My misfortune is that I still resemble a man too much.
I should liked to be wholly a beast like that goat."

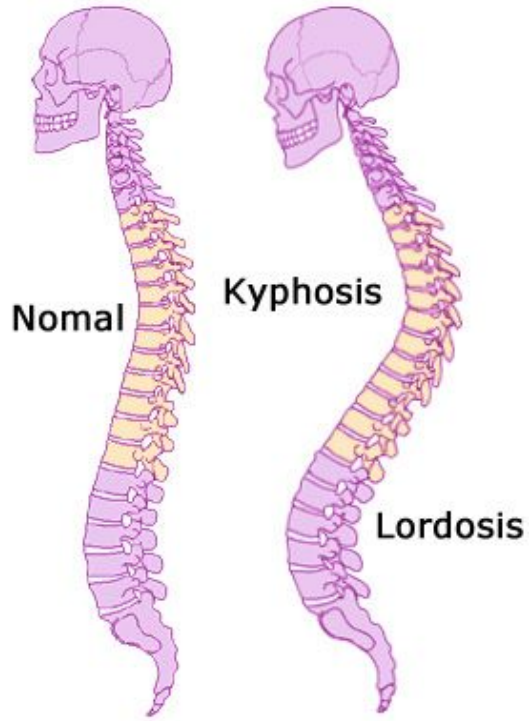
(V. Hugo, *The Hunchback of Notre Dame*)

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)

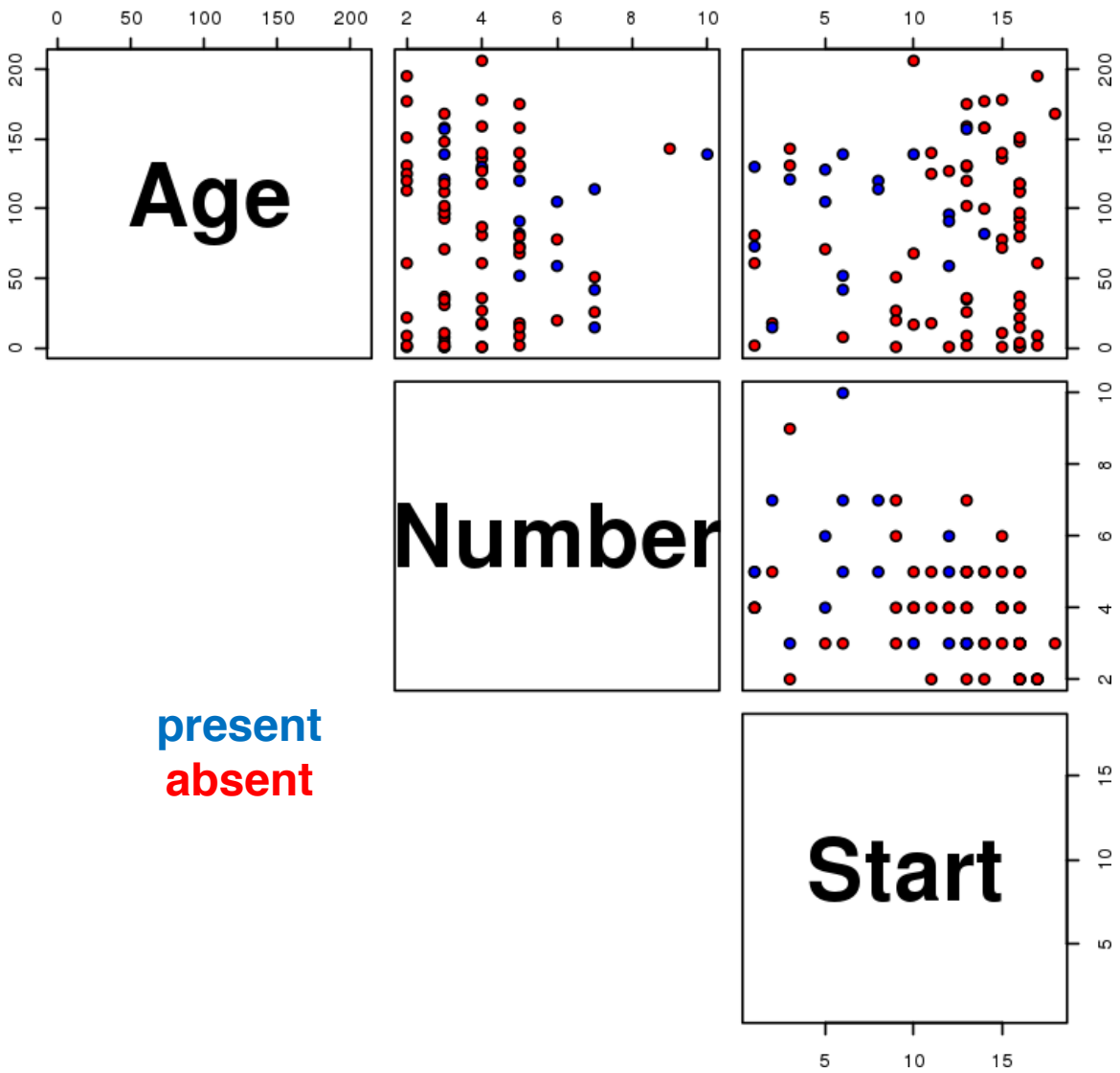


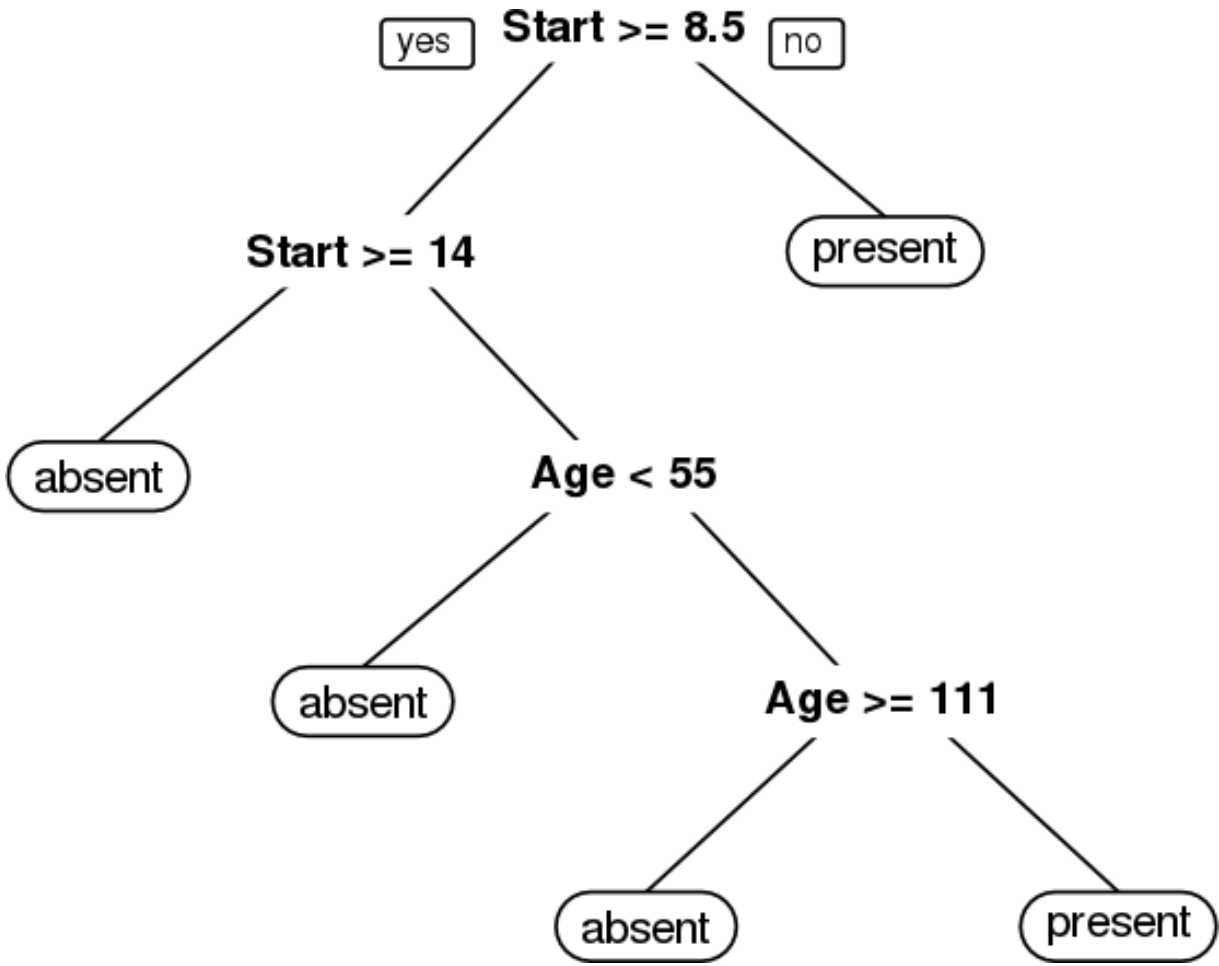
EXAMPLE – KYPHOSIS DATASET

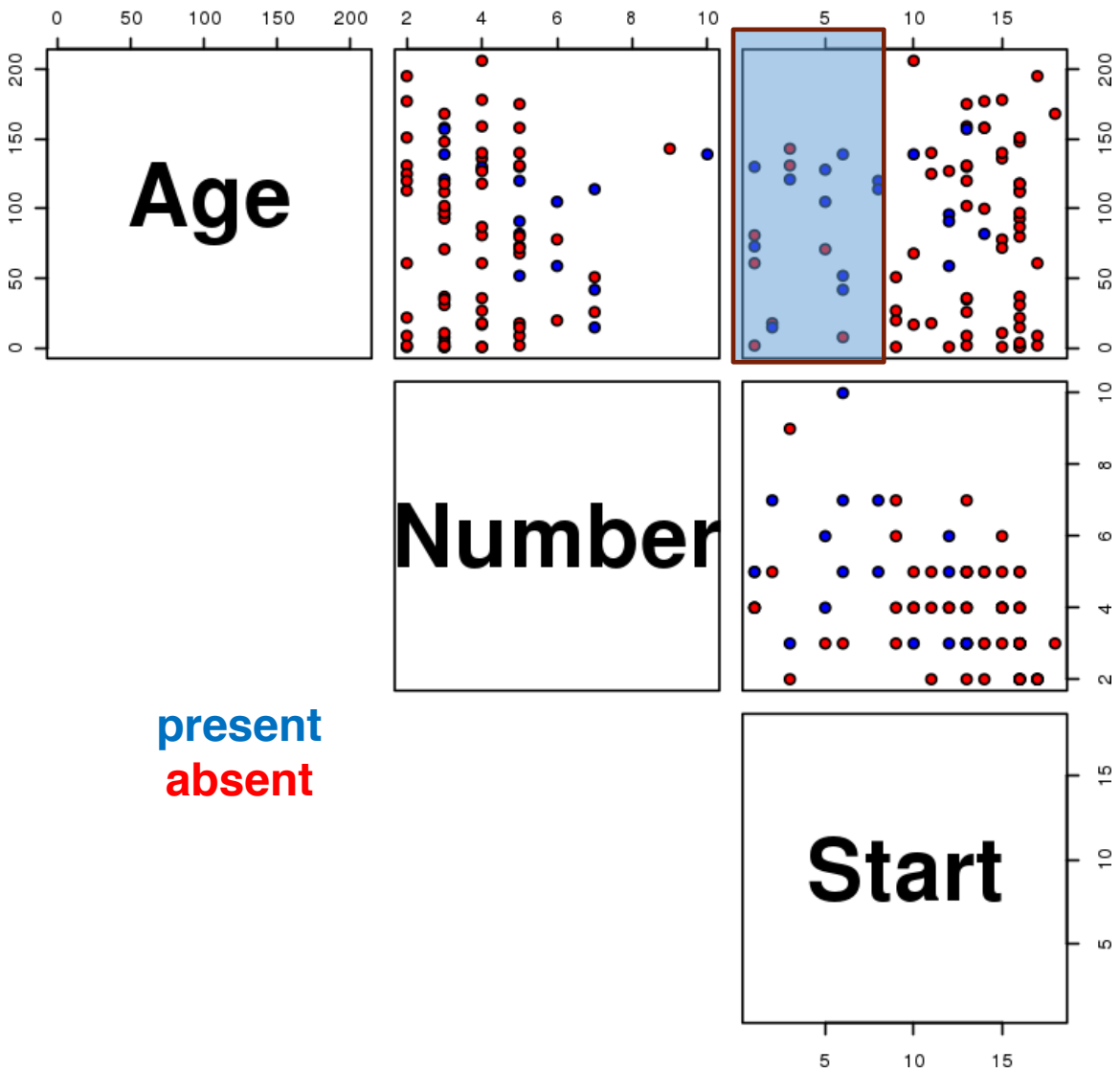
The question of interest for this natural dataset is how the three explanatory attributes might impact the operation's success.

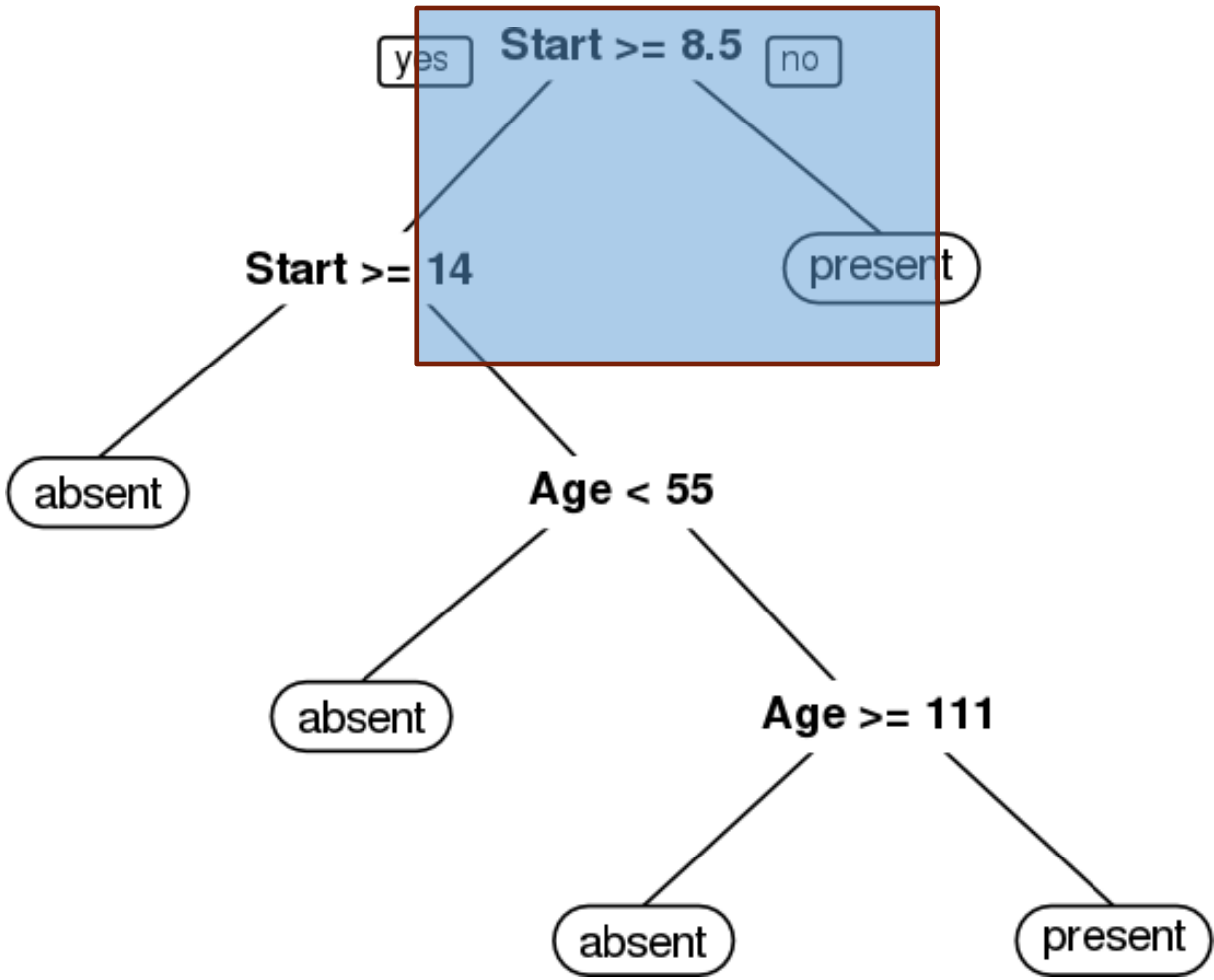
We use the rpart implementation of CART to generate candidate decision trees.

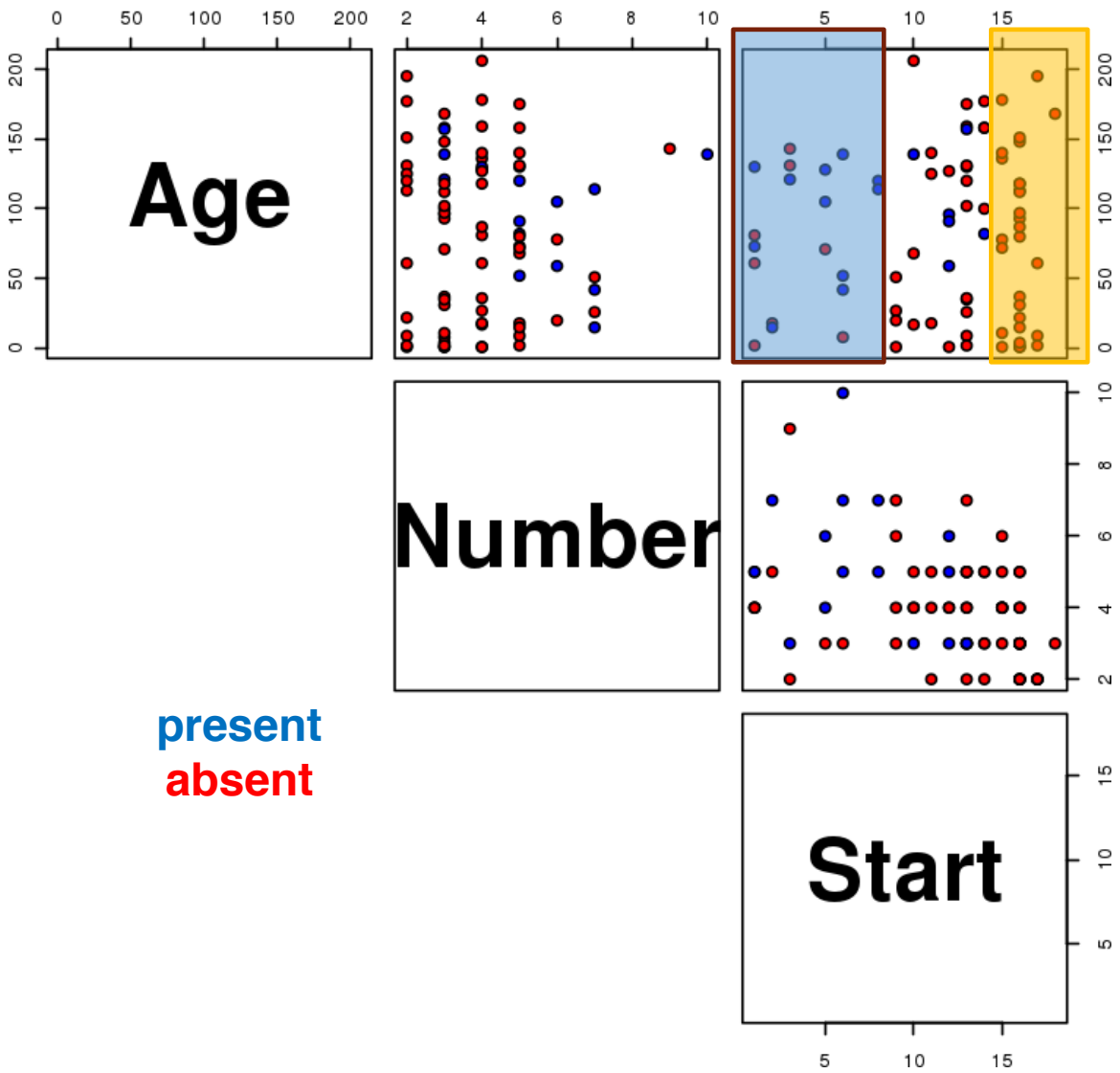
Strictly speaking, this is not a predictive supervised task as we treat the entire dataset as a training set (there are no hold-out testing observations for the time being).

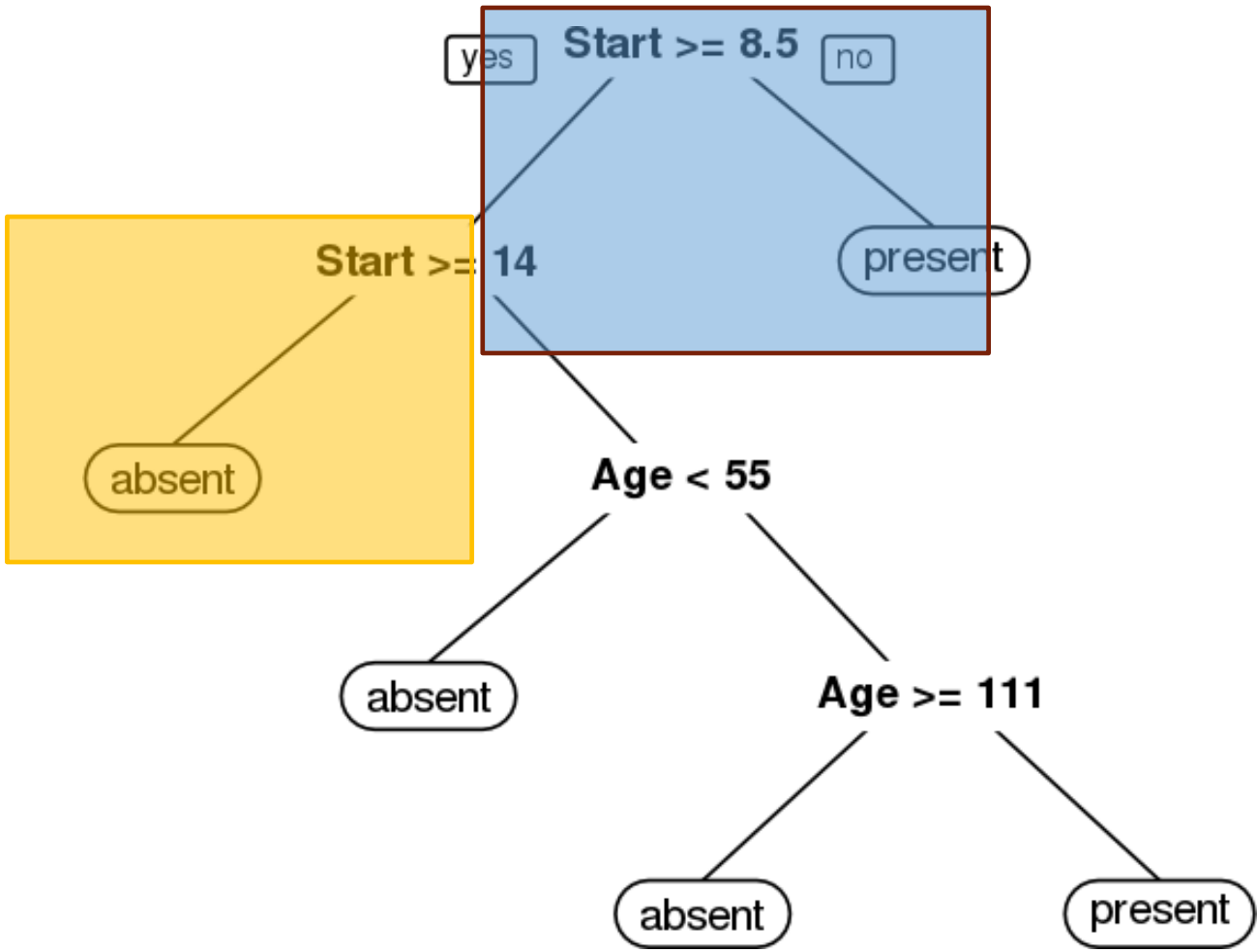

Age
Number
Start
present
absent

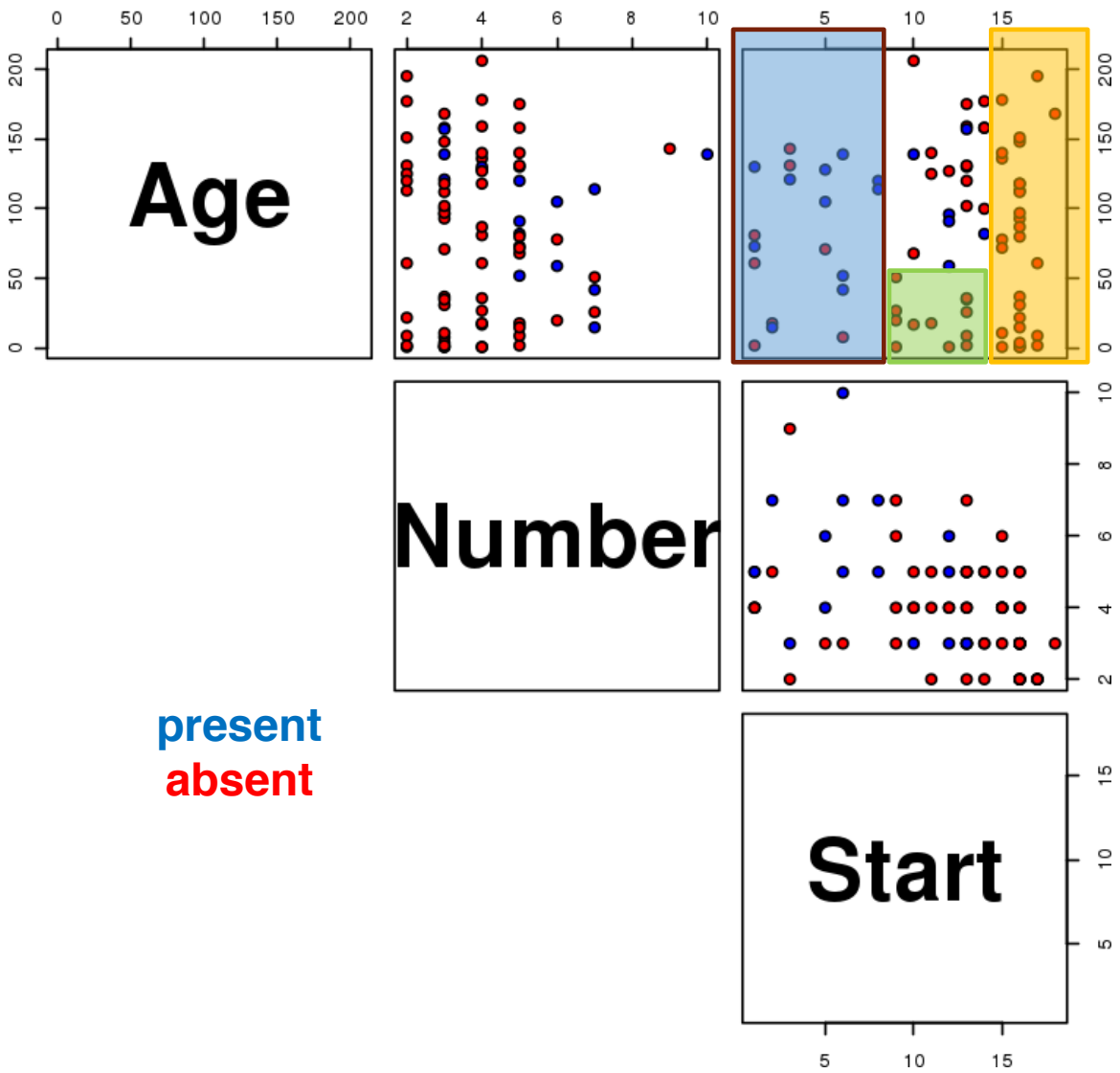

yes
Start >= 8.5
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present

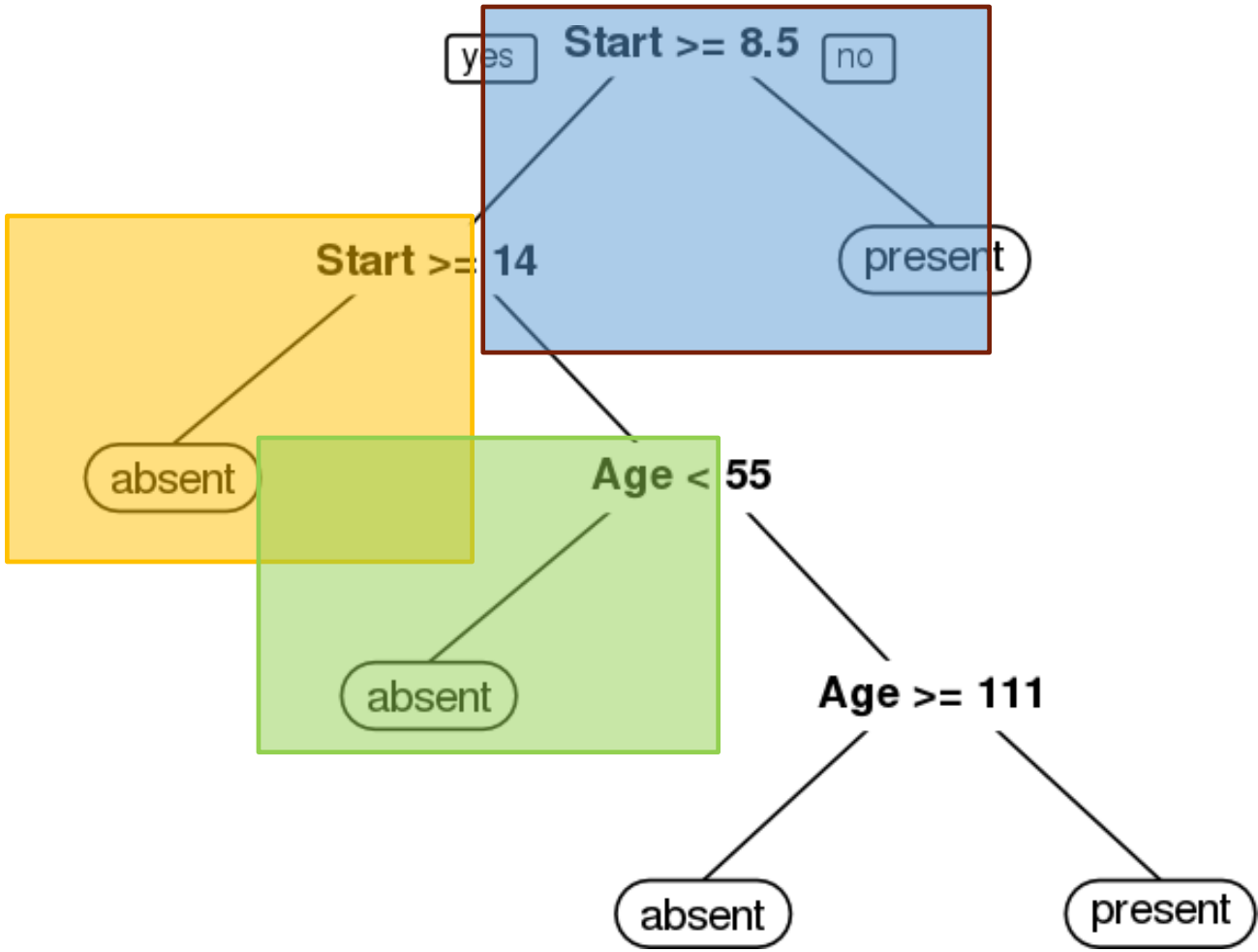

Age
Number
Start
present
absent

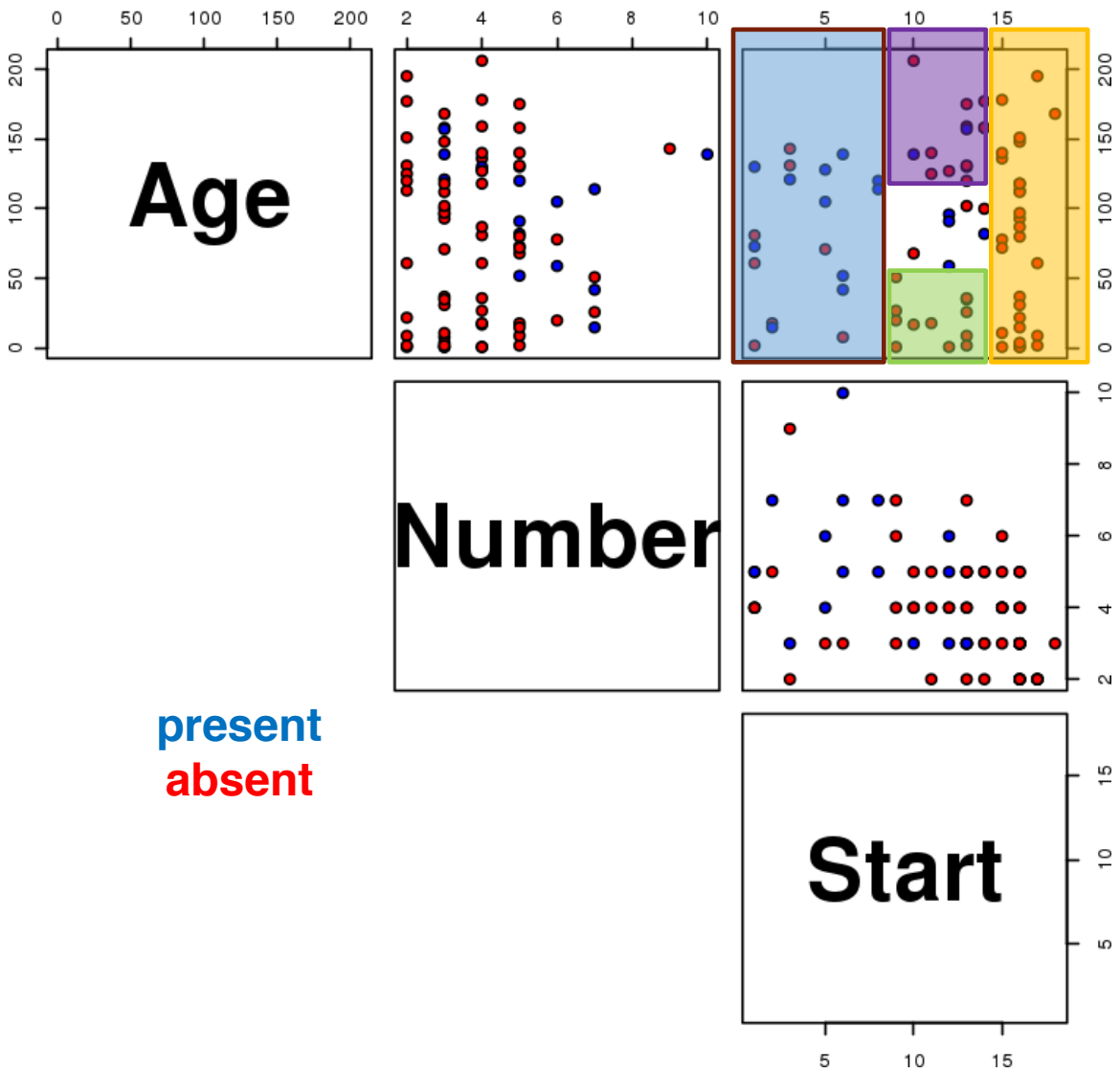

yes
Start >= 8.5
no
Start >= 14
present
absent
Age < 55
absent
Age >= 111
absent
present

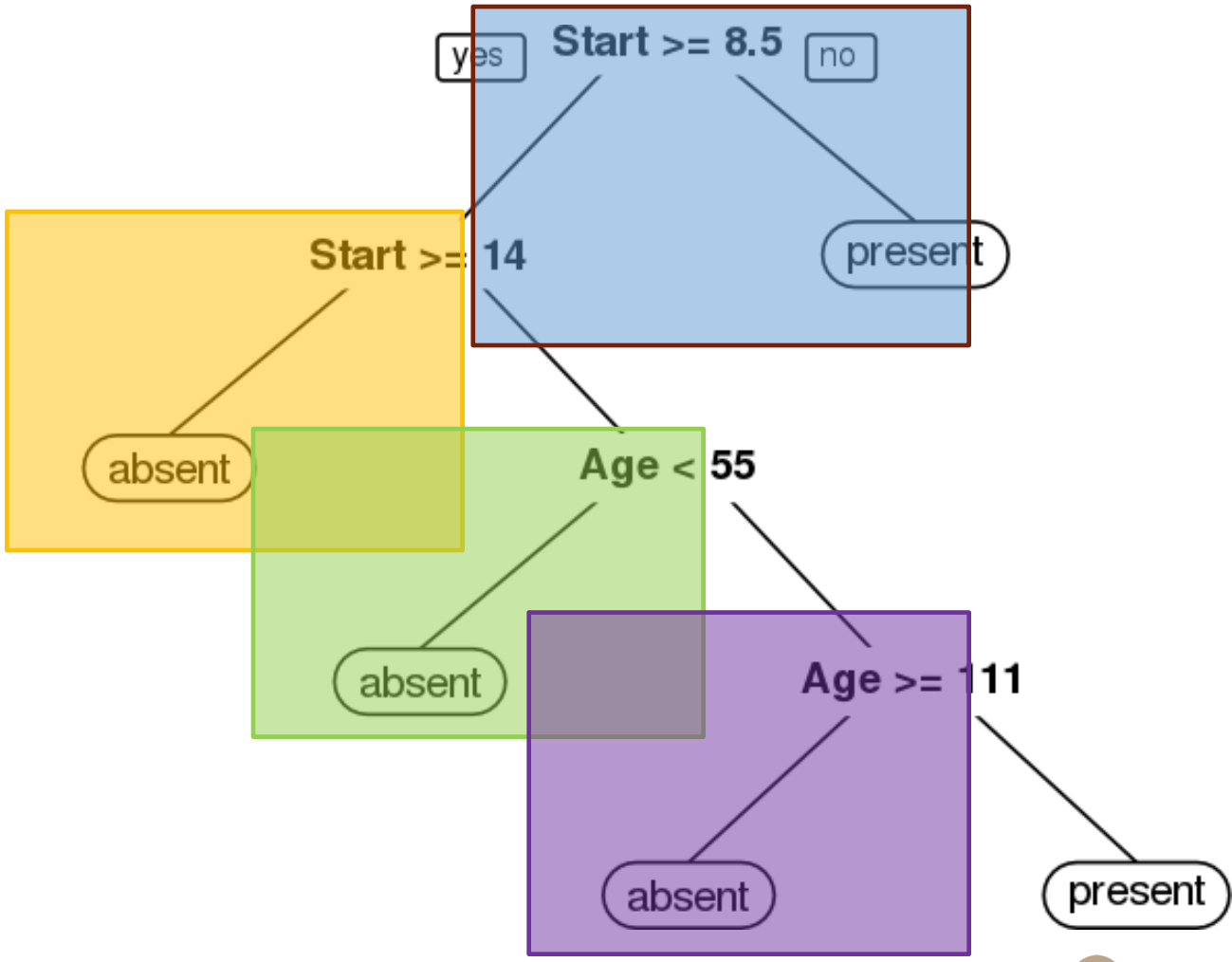

Age
Number
Start
present
absent

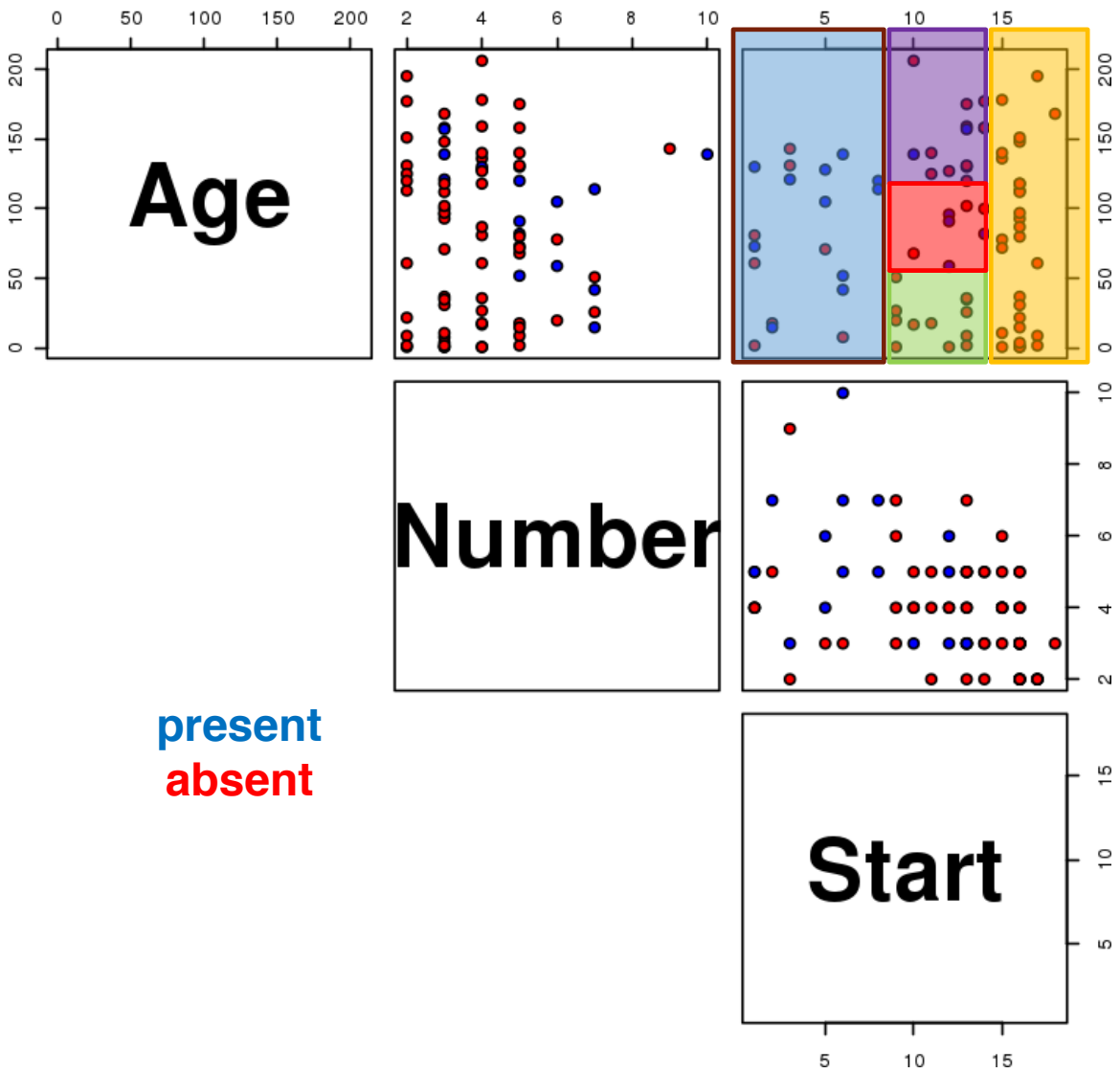

Start >= 8.5
yes
no
present
Start >= 14
absent
Age < 55
absent
Age >= 111
absent
present

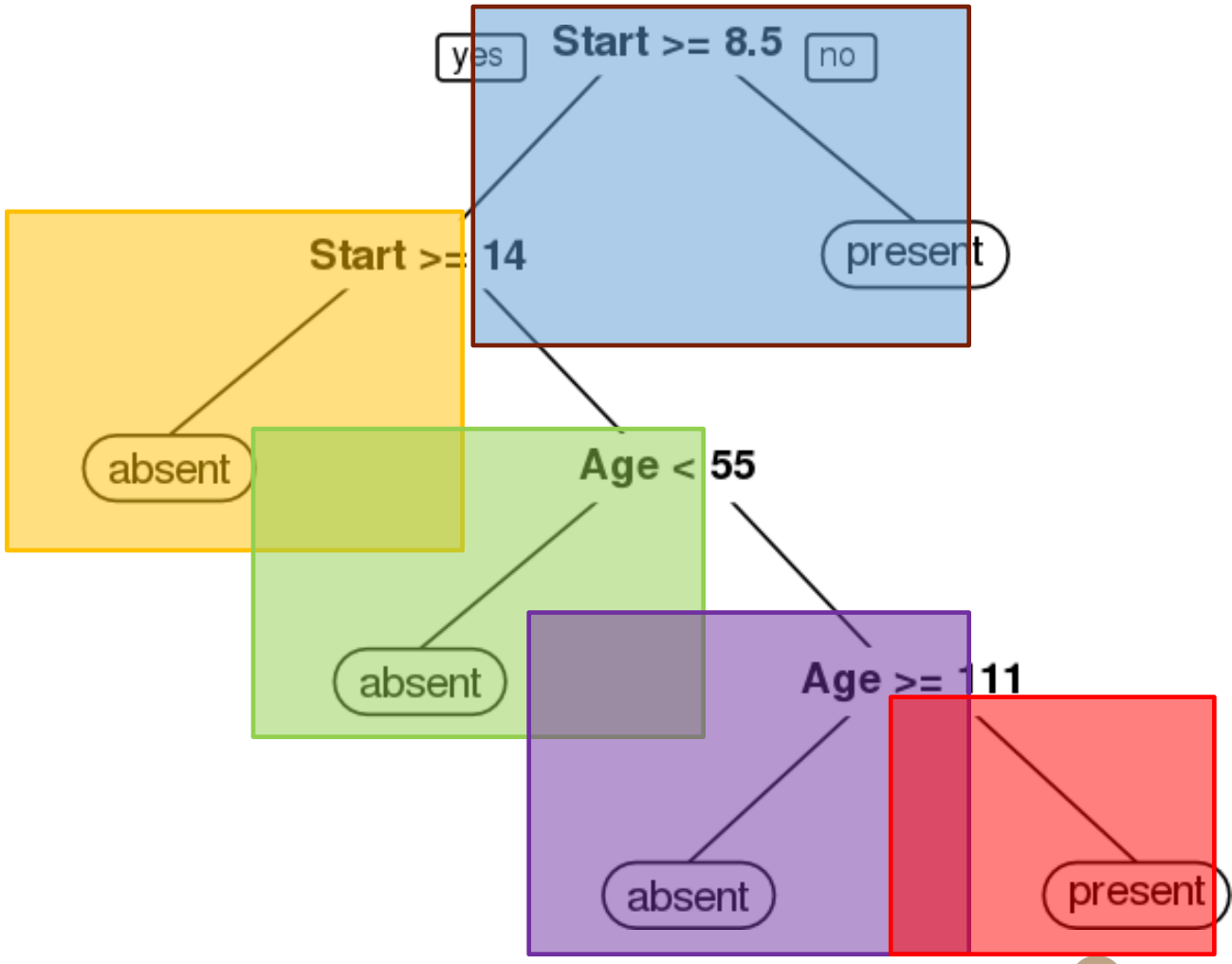

Age
Number
Start
present
absent


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

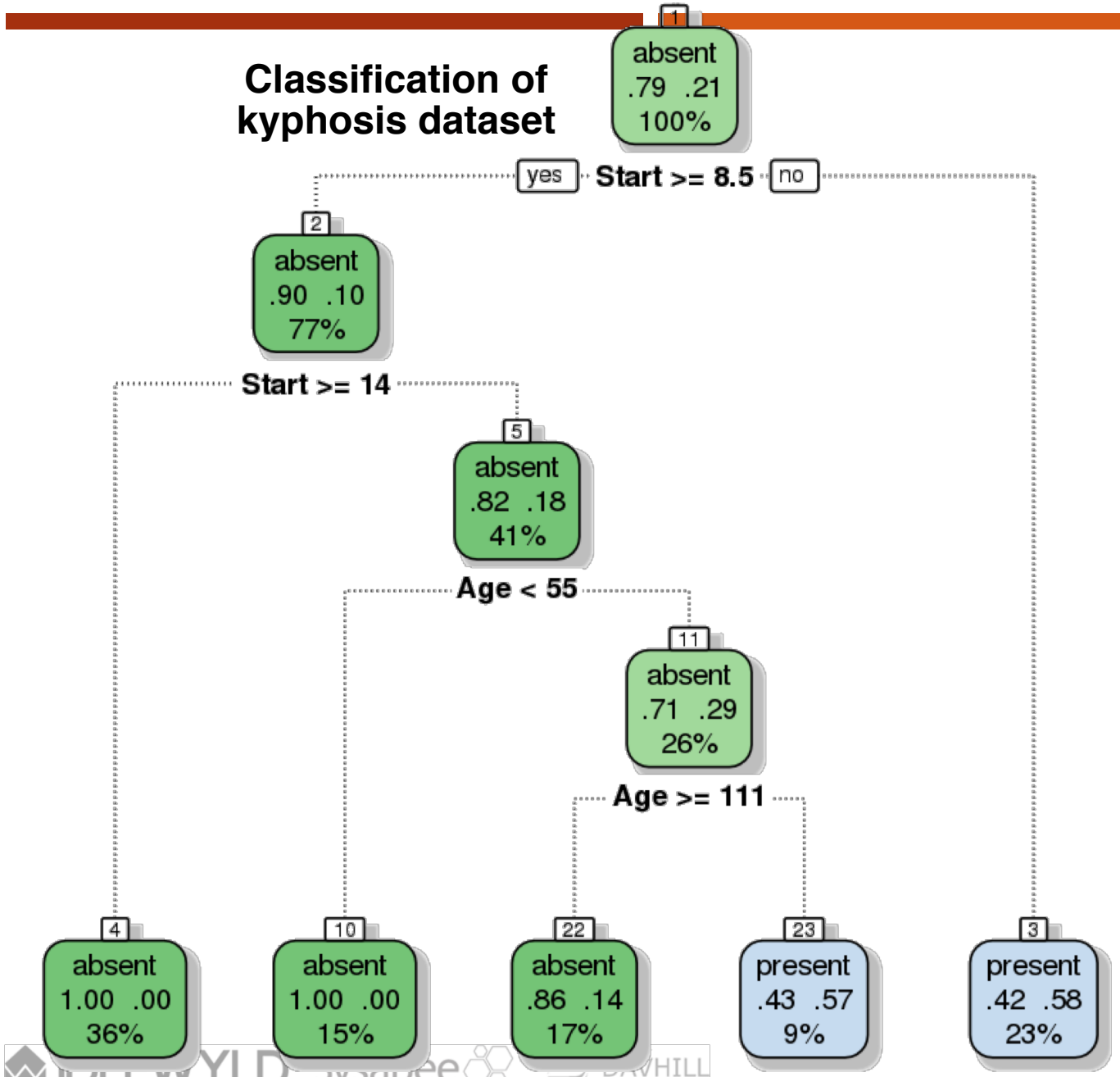

Age
Number
Start
present
absent


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

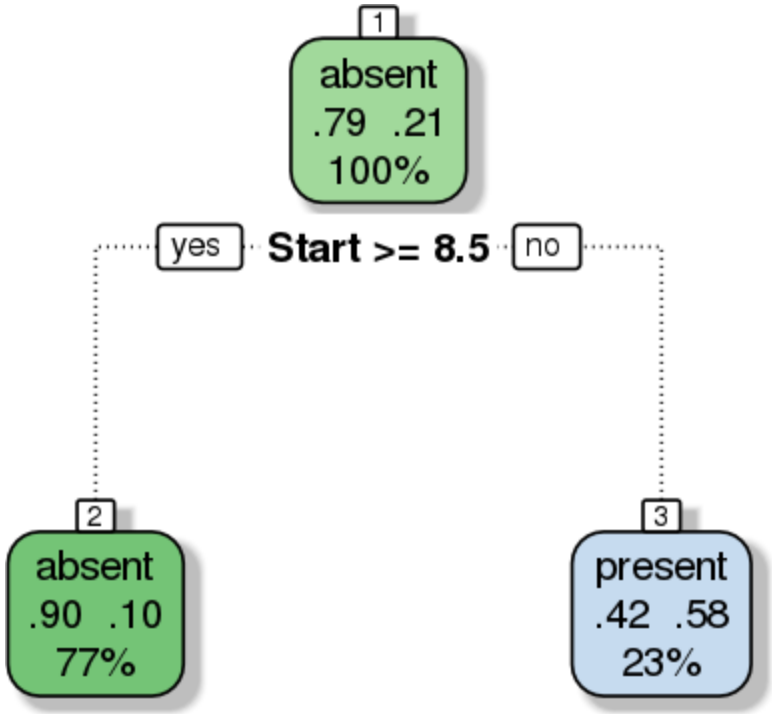

Age
Number
Start
present
absent


yes
Start >= 8.5
no
Start >= 14
present
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	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

DISCUSSION

Is this a good model?

Most performance metrics do not generalize to the multinomial case.

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

			Predicted: Maltreatment: Unfounded	Predicted: Maltreatment: Suspected	Predicted: Maltreatment: Substantiated	Predicted: Risk: No	Predicted: Risk: Yes	Predicted: Risk: Unknown	Total	
Actuals	Maltreatment	Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
		Suspected	-	965	-	29	2	-	*995*	*6.1%*
		Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
	Risk	No	894	-	763	949	19	9	*2,632*	*16.1%*
		Yes	123	-	520	122	111	5	*880*	*5.4%*
		Unknown	212	-	303	184	21	24	*745*	*4.6%*
	Total		*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
			35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

PERFORMANCE EVALUATION

For numerical targets y with predictions $\hat{y}$, metrics include:

- **mean squared** and **mean absolute errors**

$$\text{MSE} = \text{mean}\{(\hat{y}_i - y_i)^2\},\ \text{MAE} = \text{mean}\{|\hat{y}_i - y_i|\}$$

- **normalized mean squared** and **normalized mean absolute errors**

$$\text{NMSE} = \frac{\text{mean}\{(\hat{y}_i - y_i)^2\}}{\text{mean}\{(\bar{y} - y_i)^2\}}, \text{NMAE} = \frac{\text{mean}\{|\hat{y}_i - y_i|\}}{\text{mean}\{|\bar{y} - y_i|\}}$$

- **mean average percentage error** $\text{MAPE} = \text{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$

- **correlation** $\rho_{\hat{y},y}$

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.

REFERENCES

CLASSIFICATION AND VALUE ESTIMATION

SUPPLEMENTAL MATERIAL

Value Estimation Methods
https://www.data-action-lab.com/wp-content/uploads/2019/03/Value-Estimation-Methods.pdf

Logistic Regression
https://www.data-action-lab.com/wp-content/uploads/2019/03/Logistic-Regression.pdf

Naïve Bayes Classification
https://www.data-action-lab.com/wp-content/uploads/2019/03/Naïve-Bayes-Classification.pdf

REFERENCES

Kitts, B., Zhang, J., Wu, G., Brandi, W., Beasley, J., Morrill, K., Ettedgui1, J., Siddhartha, S., Yuan, H., Gao, F., Azo, P., Mahato, R. (in press), Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft, Annals of Information Systems Special Issue on Data Mining in Real-World Applications.

Kitts, B. (2013), The Making of a Large-Scale Ad Server, in Data Mining Case Studies Workshop and Practice Prize 5, Proceedings of the IEEE Thirteenth International Conference on Data Mining Workshops (ICDMW 2013), December, Dallas, TX, IEEE Press.

Fefilatyev, S., Kramer, K., Hall, L., Goldgof, D., Kasturi, R., Remsen, A., Daly, K. (2011), Detection of Anomalous Particles from Deepwater Horizon Oil Spill Using SIPPER3 Underwater Imaging Platform, in Data Mining Case Studies IV, Proceedings of the Eleventh IEEE International Conference on Data Mining, Vancouver, Canada

Kitts, B. (2005), Product Targeting From Rare Events: Five Years of One-to-One Marketing at CPI, Marketing Science Conference, Atlanta, June 2005.

REFERENCES

https://algobeans.com/2016/07/27/decision-trees-tutorial/

https://en.wikipedia.org/wiki/Decision_tree_learning

https://en.wikipedia.org/wiki/Predictive_analytics

https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines

https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Aggarwal, C.C. (ed.) [2015], *Data Classification: Algorithms and Applications*, CRC Press.

Leskovec, J., Rajamaran, A., Ullman, J.D. [2014], *Mining of Massive Datasets*, Cambridge Press.

Provost, F., Fawcett, T. [2013], Data Science for Business, O'Reilly.

REFERENCES

Naïve Bayes Classifier (Wikipedia)

Zhang, H. (2014) The optimality of Naïve Bayes

Domings, P., and Pazzani, M. (1997) Beyond independence: Conditions for the optimality of the simple Bayesian classifier

Markham, K. Scikit-learn video #3: Machine learning first steps with the Iris dataset

http://www.ee.columbia.edu/~vittorio/BayesProof.pdf

https://www.cs.cmu.edu/~epxing/Class/10701-08s/Lecture/lecture3-annotated.pdf

http://www.cogsys.wiai.uni-bamberg.de/teaching/ss05/ml/slides/cogsysII-9.pdf