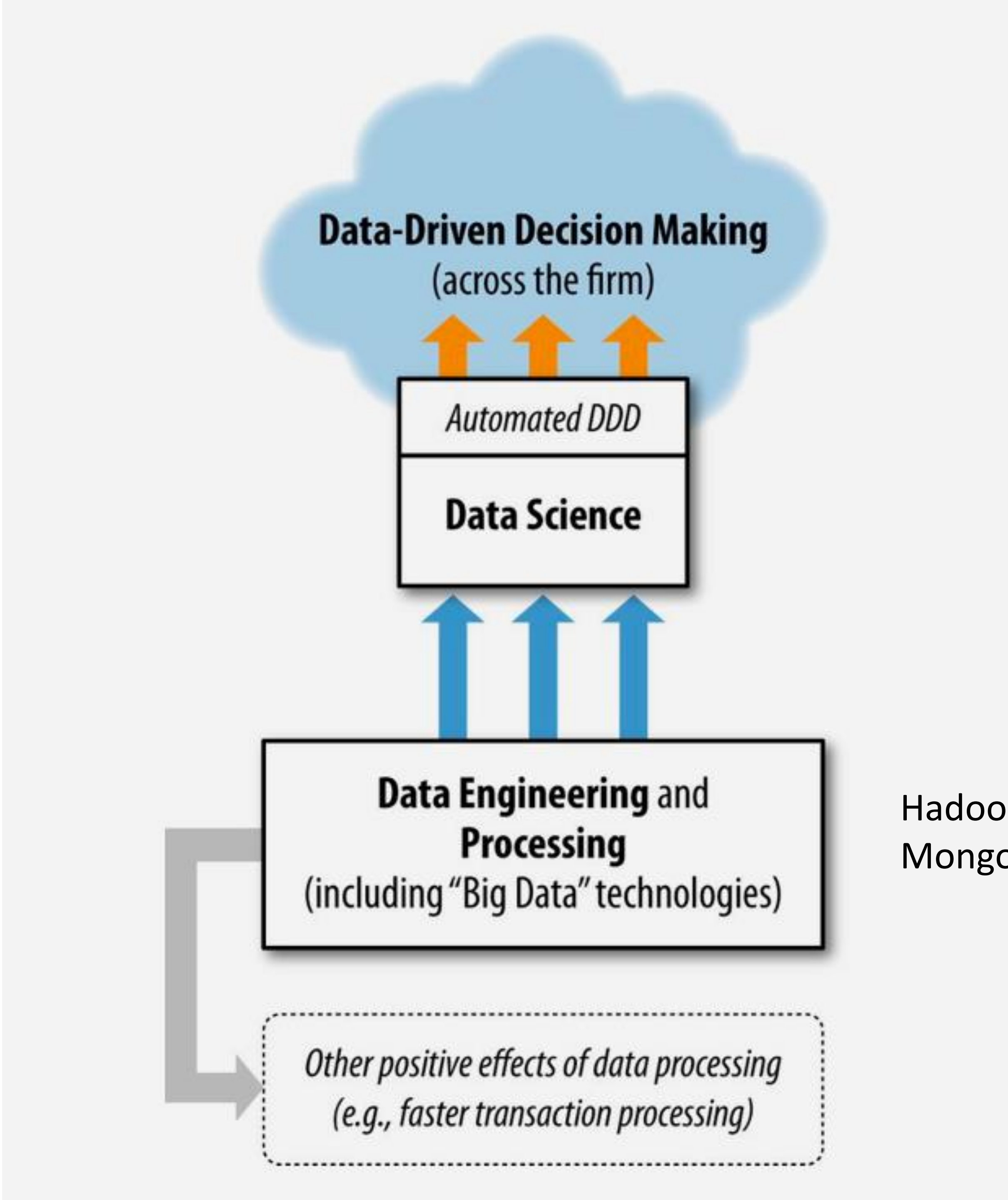
Data Science for Business

陳炫碩
中央大學 ERP中心 主任

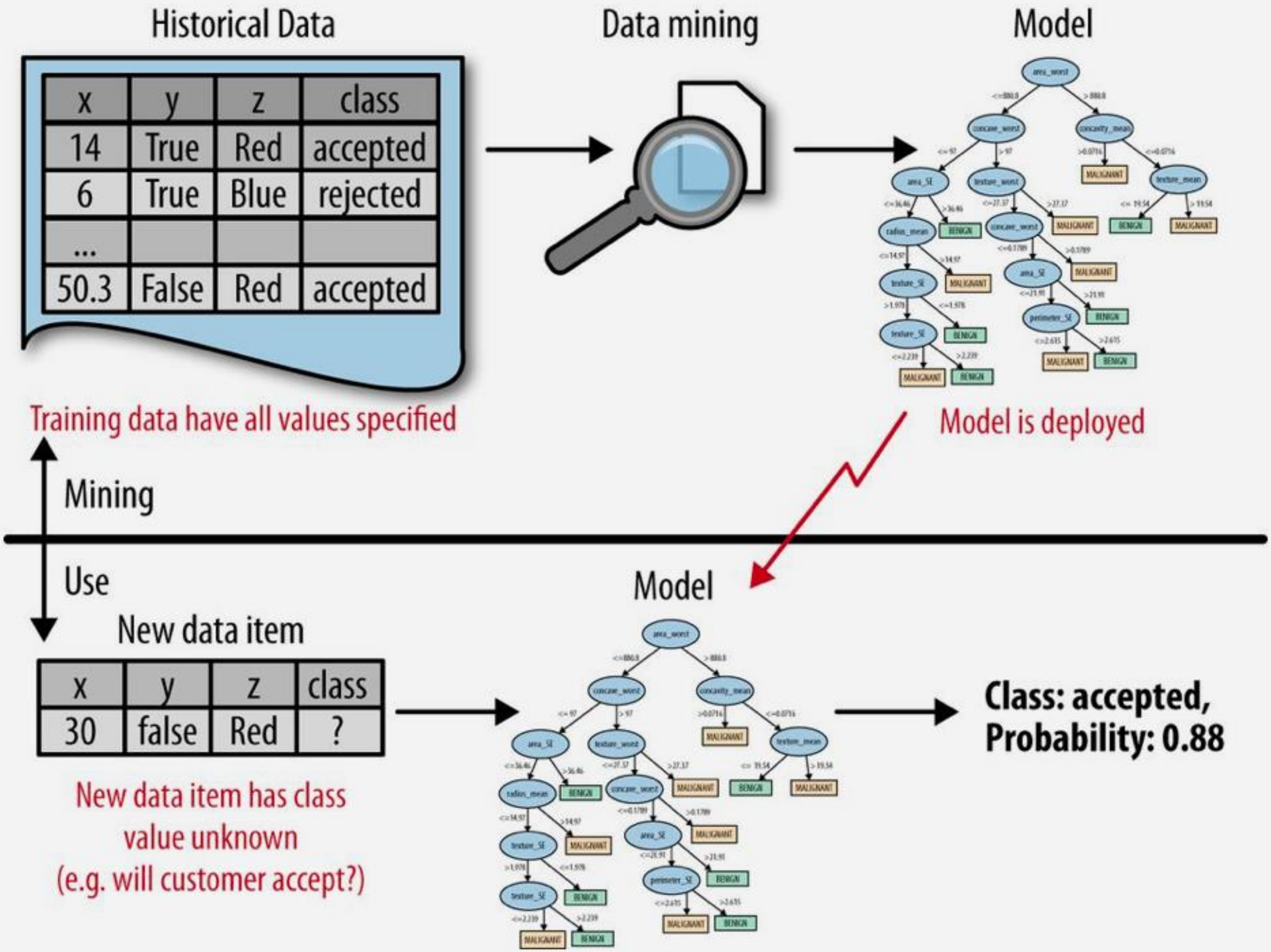
Extracting knowledge from data to solve business problems

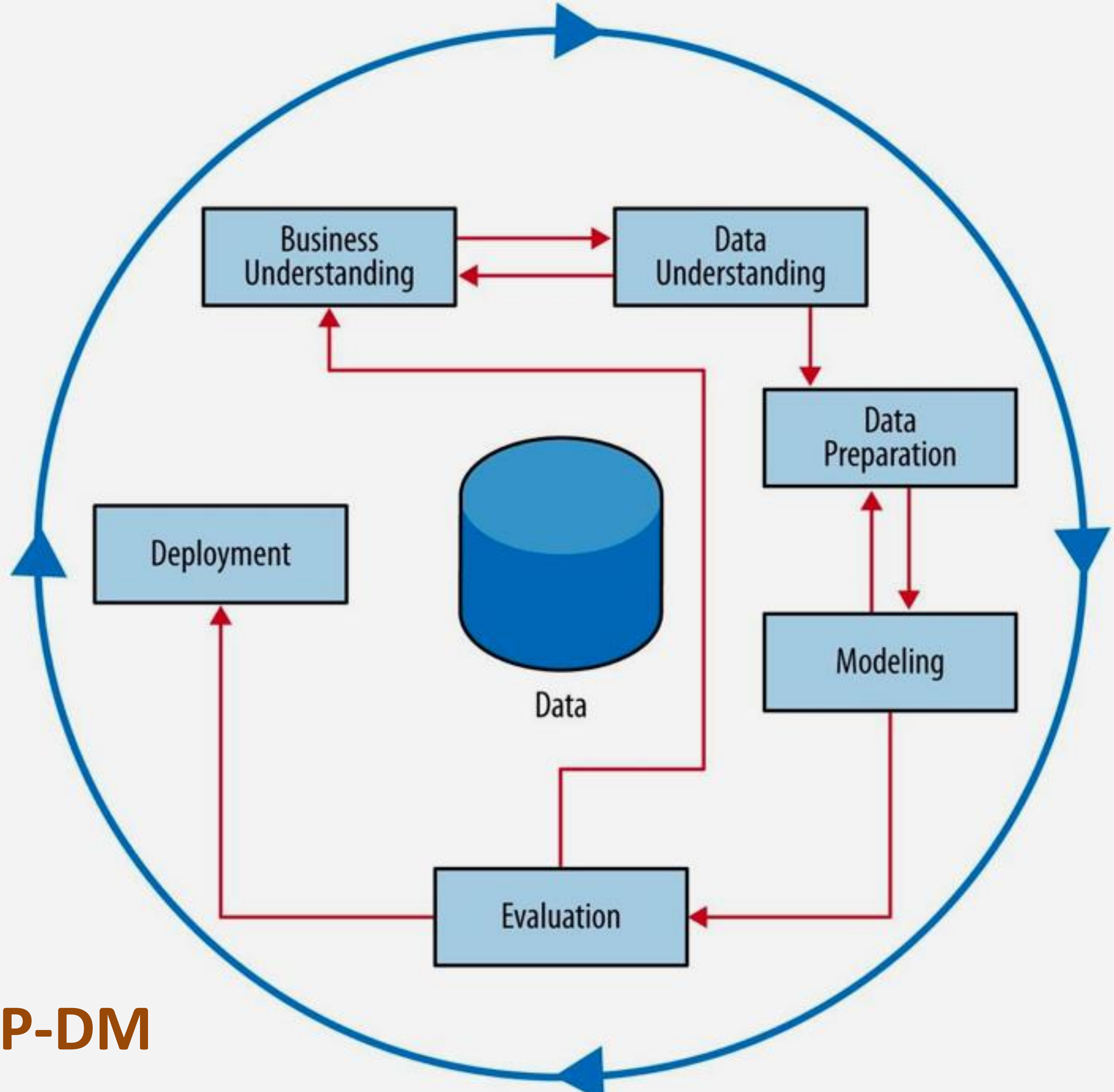


Hadoop, Hbase,
MongoDB, SAP HANA

Data Science & Data Mining

- Data science is a set of fundamental principles that guide the extraction of knowledge from data.
- Data mining is the extraction of knowledge from data, via technologies that incorporate these principles.


Historical Data
x
y
z
class
14
True
Red
accepted
6
True
Blue
rejected
...
50.3
False
Red
accepted
Data mining
Model
Training data have all values specified
Model is deployed
Mining
Use
Model
New data item
x
y
z
class
30
false
Red
?
New data item has class
value unknown
(e.g. will customer accept?)
Class: accepted,
Probability: 0.88


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM

Modeling

Predictive model

- Informative variables

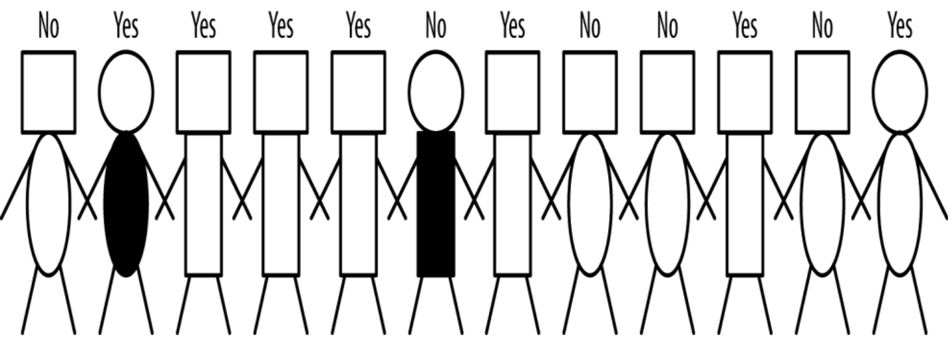
  - provide information that reduces uncertainty about something

- Model: simplified representation of reality created to serve a purpose

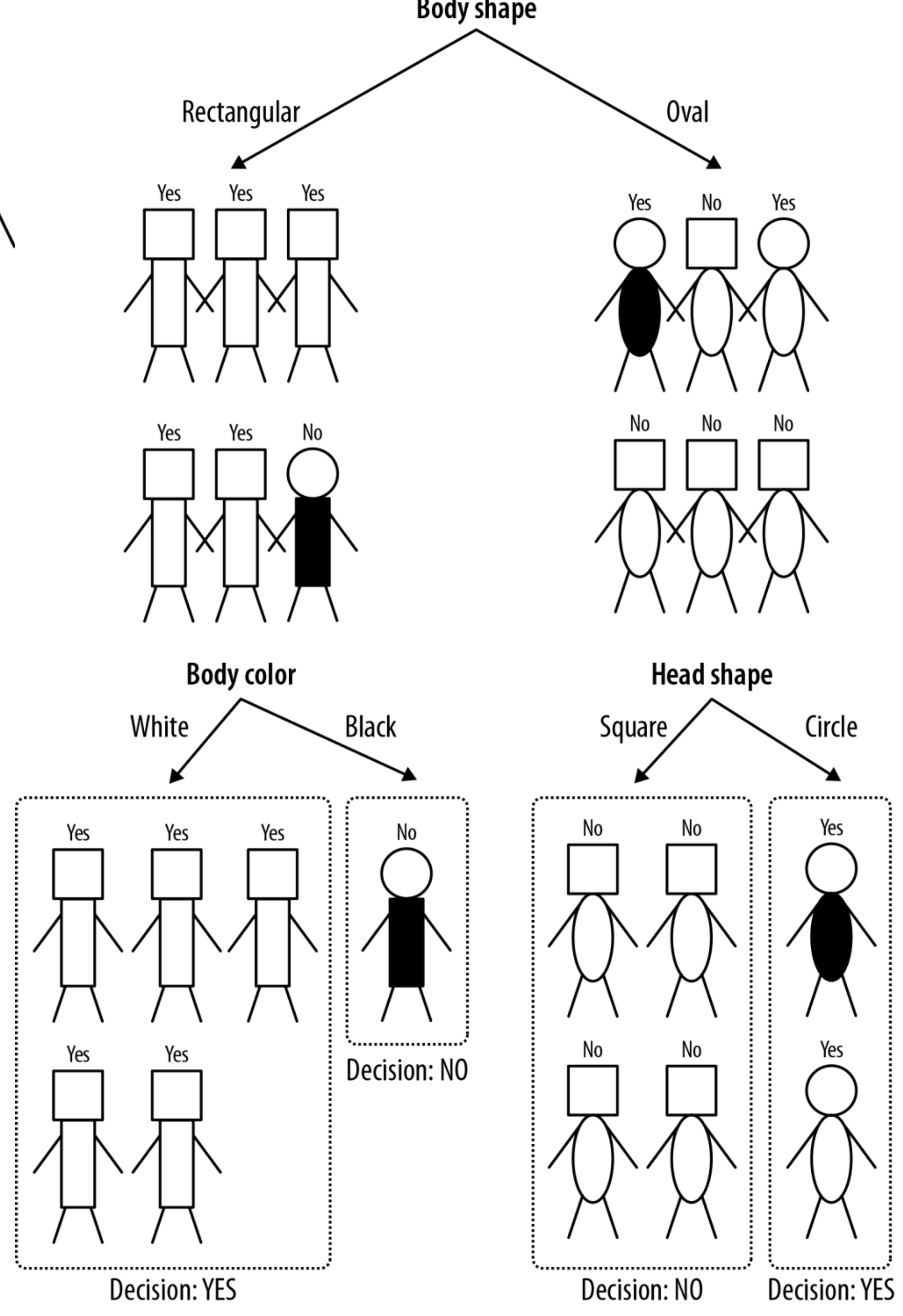
- A predictive model is a formula for estimating the unknown value of interest. (the formula could be mathematical or a logical statement.)

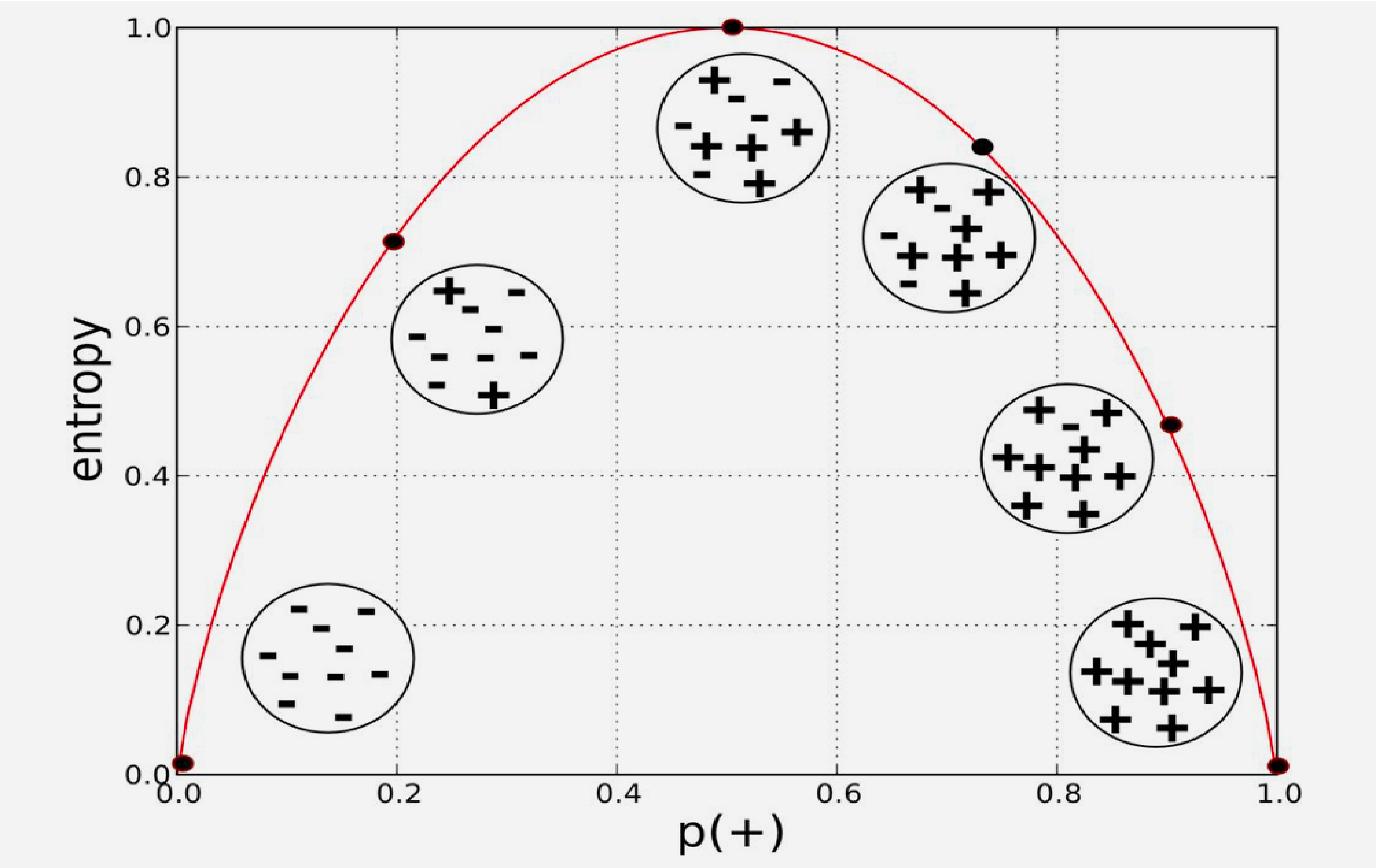
Prediction model

- Classification model
- Class-probability estimation model
- Regression model



Classification





Entropy = $\sum_i - p_i \log_2 p_i$

Information gain

- How much pure are the split sets compared to the original set?
- How much an attribute improves entropy?

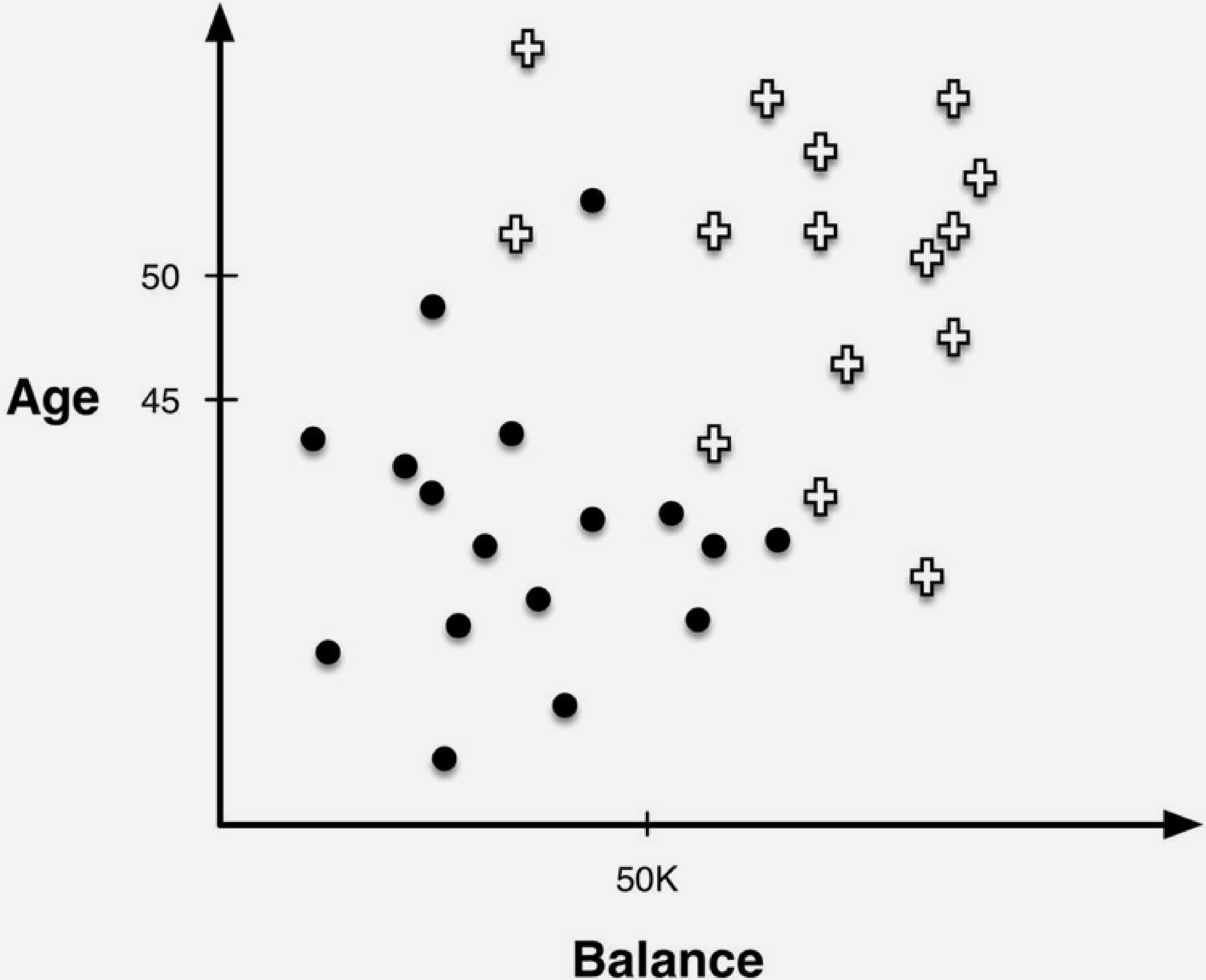
Decision Tree

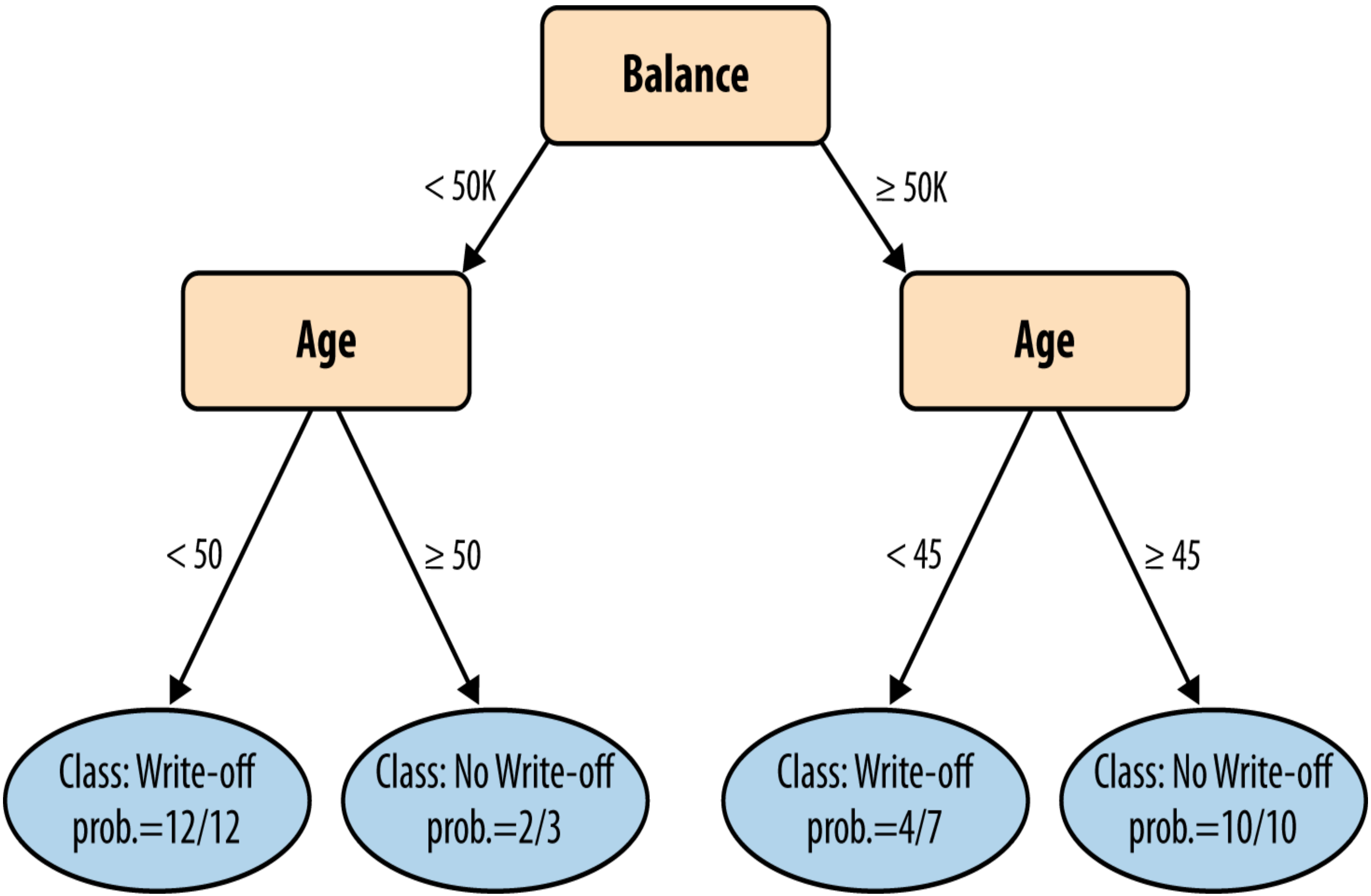
Attributes

Target attribute

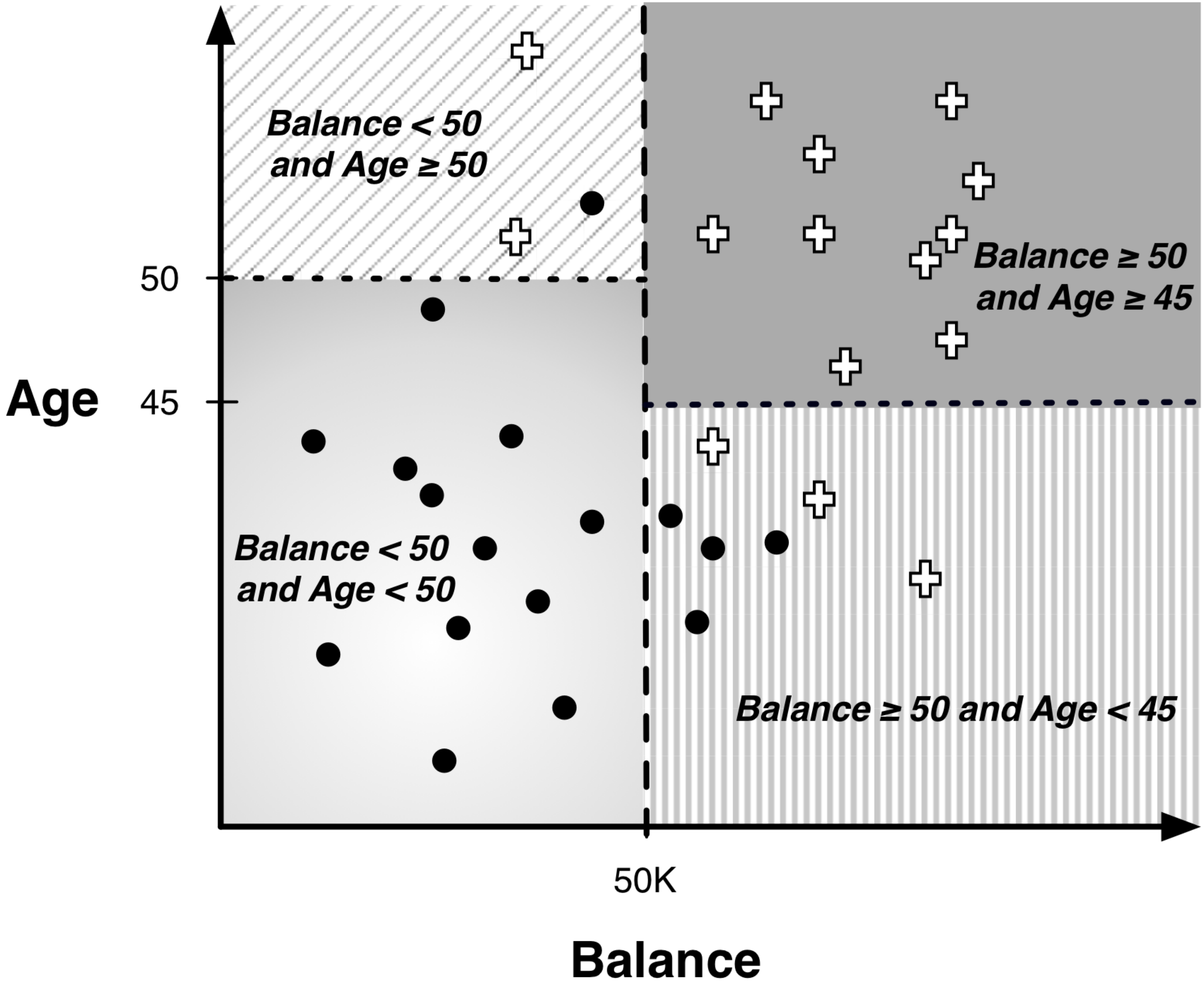
Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

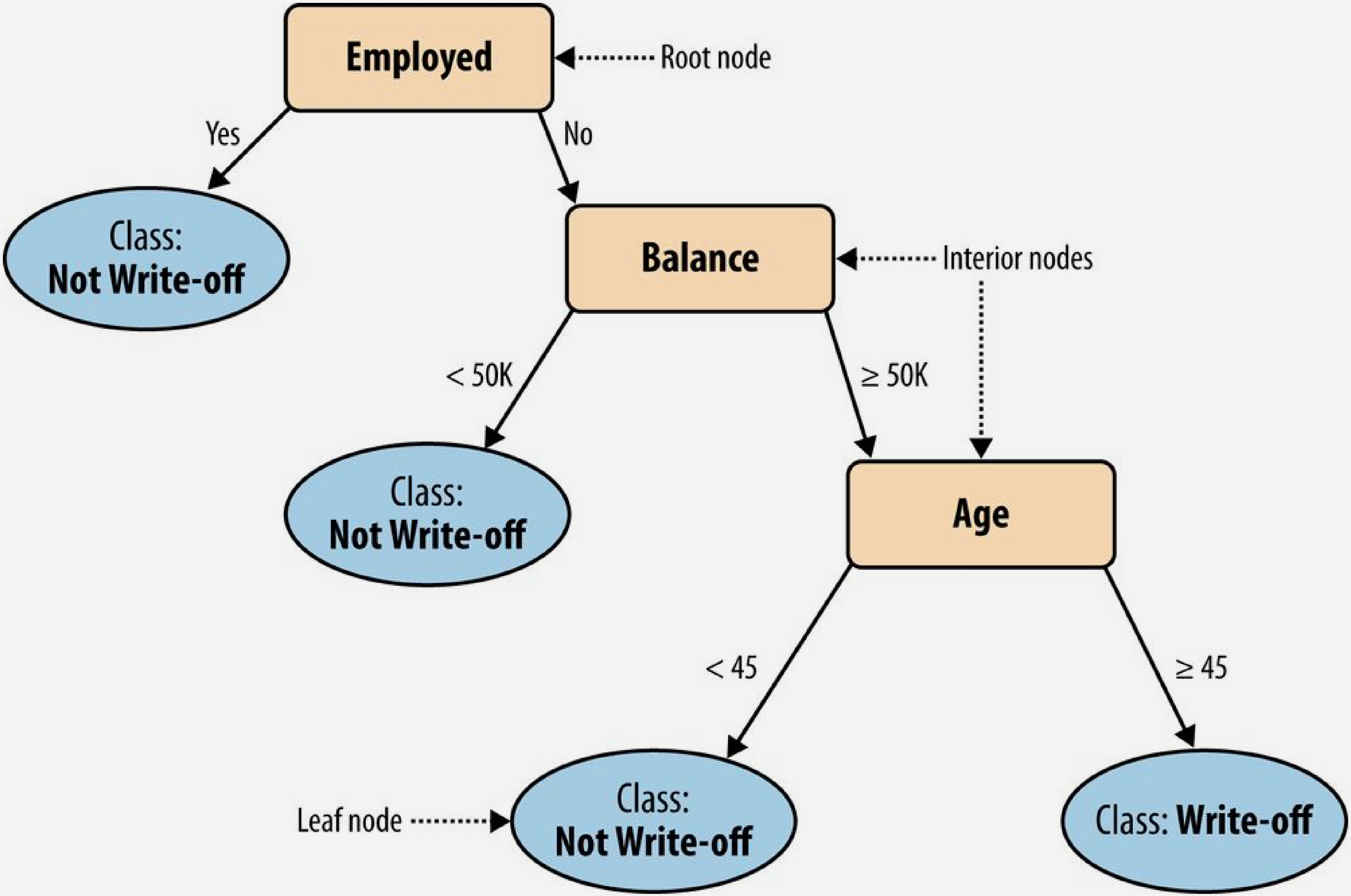
This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**


Age
50
45
50K
Balance



Balance
< 50K
≥ 50K
Age
Age
< 50
≥ 50
< 45
≥ 45
Class: Write-off
prob.=12/12
Class: No Write-off
prob.=2/3
Class: Write-off
prob.=4/7
Class: No Write-off
prob.=10/10


Balance < 50
and Age ≥ 50
Balance ≥ 50
and Age ≥ 45
Balance < 50
and Age < 50
Balance ≥ 50 and Age < 45
50
45
Age
50K
Balance

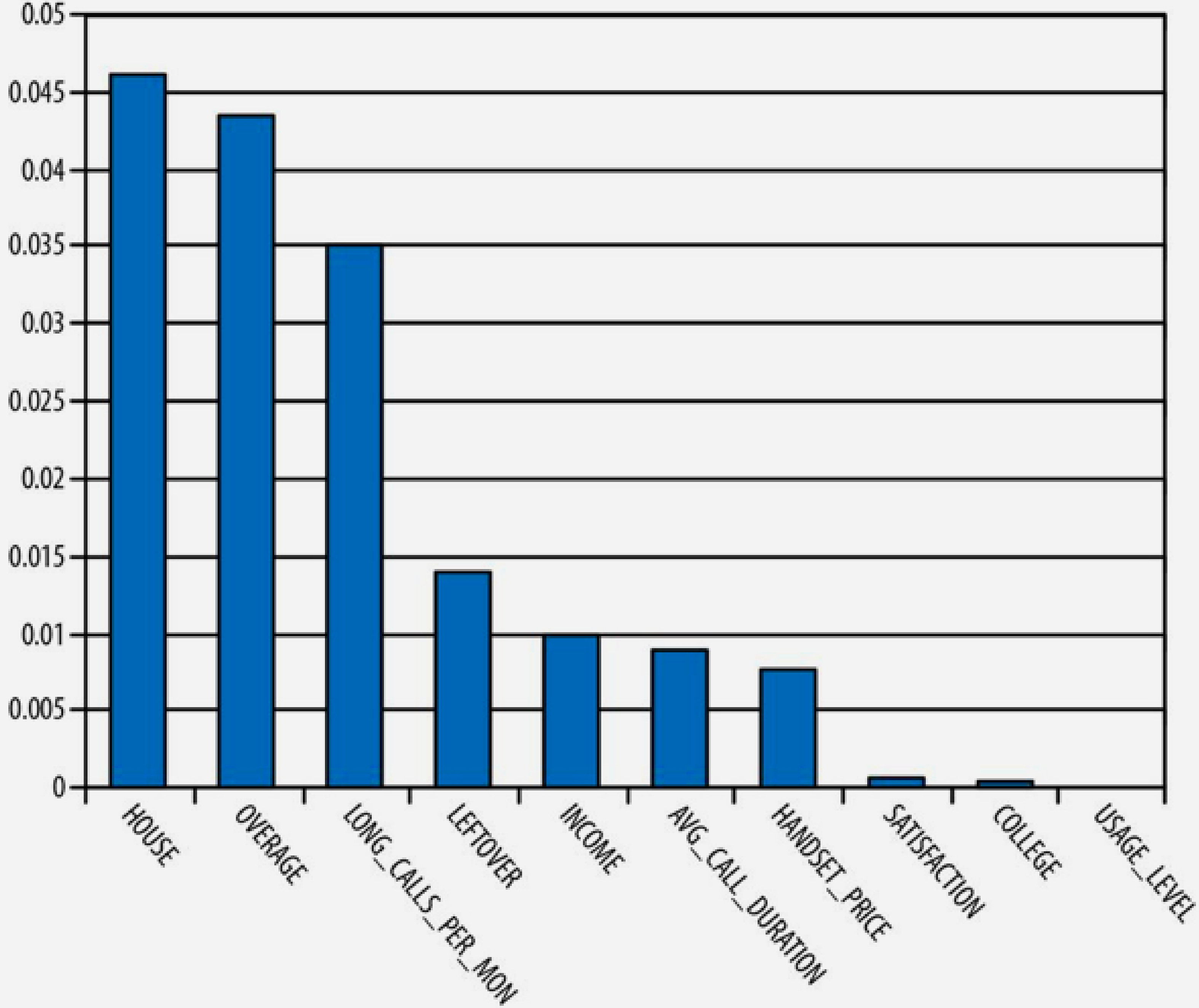

Employed
Root node
Yes
No
Class:
Not Write-off
Balance
Interior nodes
< 50K
≥ 50K
Class:
Not Write-off
Age
< 45
≥ 45
Leaf node
Class:
Not Write-off
Class: Write-off

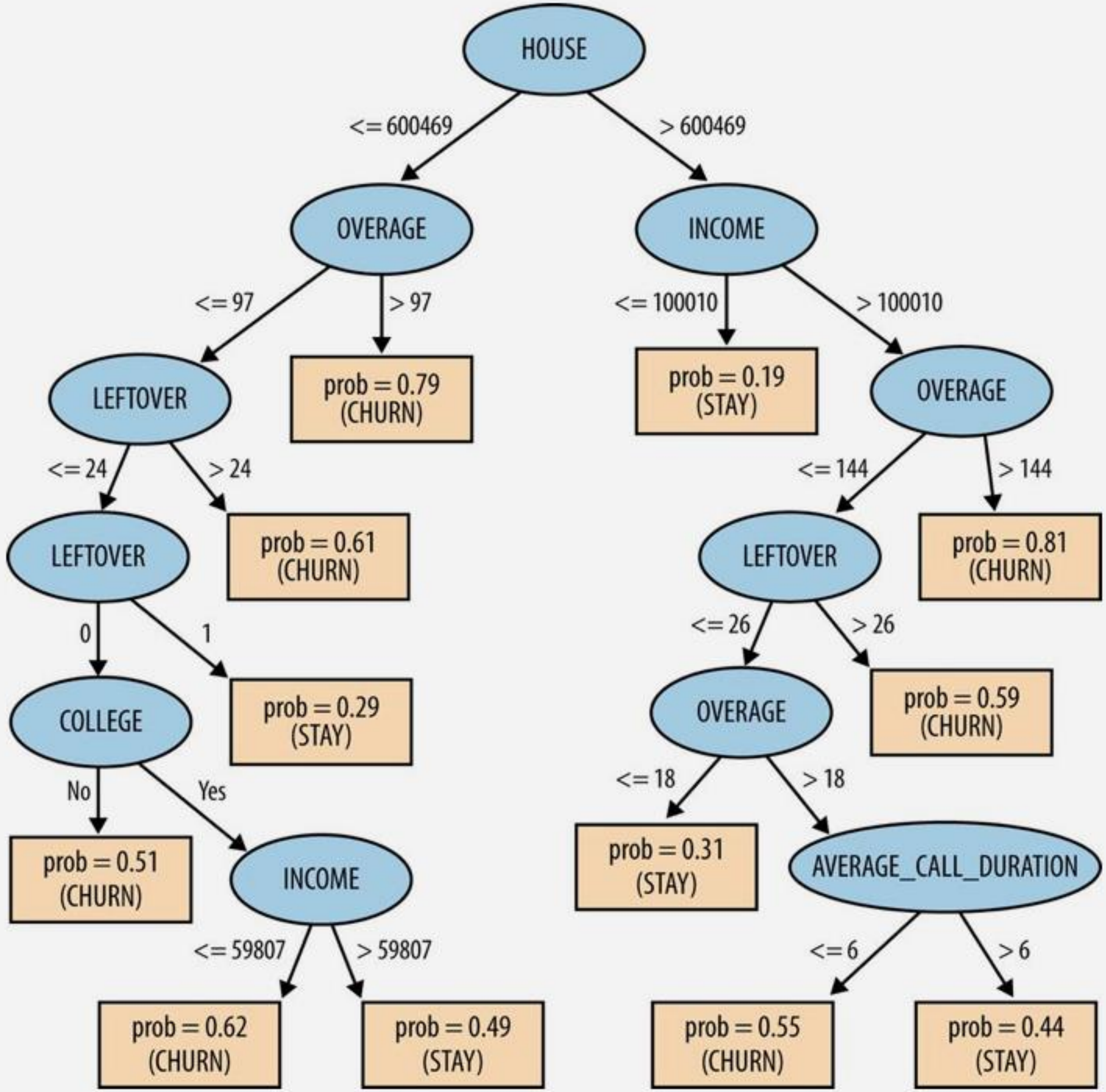
Balance < 50
and Age ≥ 50
Balance ≥ 50
and Age ≥ 45
Balance < 50
and Age < 50
Balance ≥ 50 and Age < 45
Age
50
45
50K
Balance

Class-probability estimation model

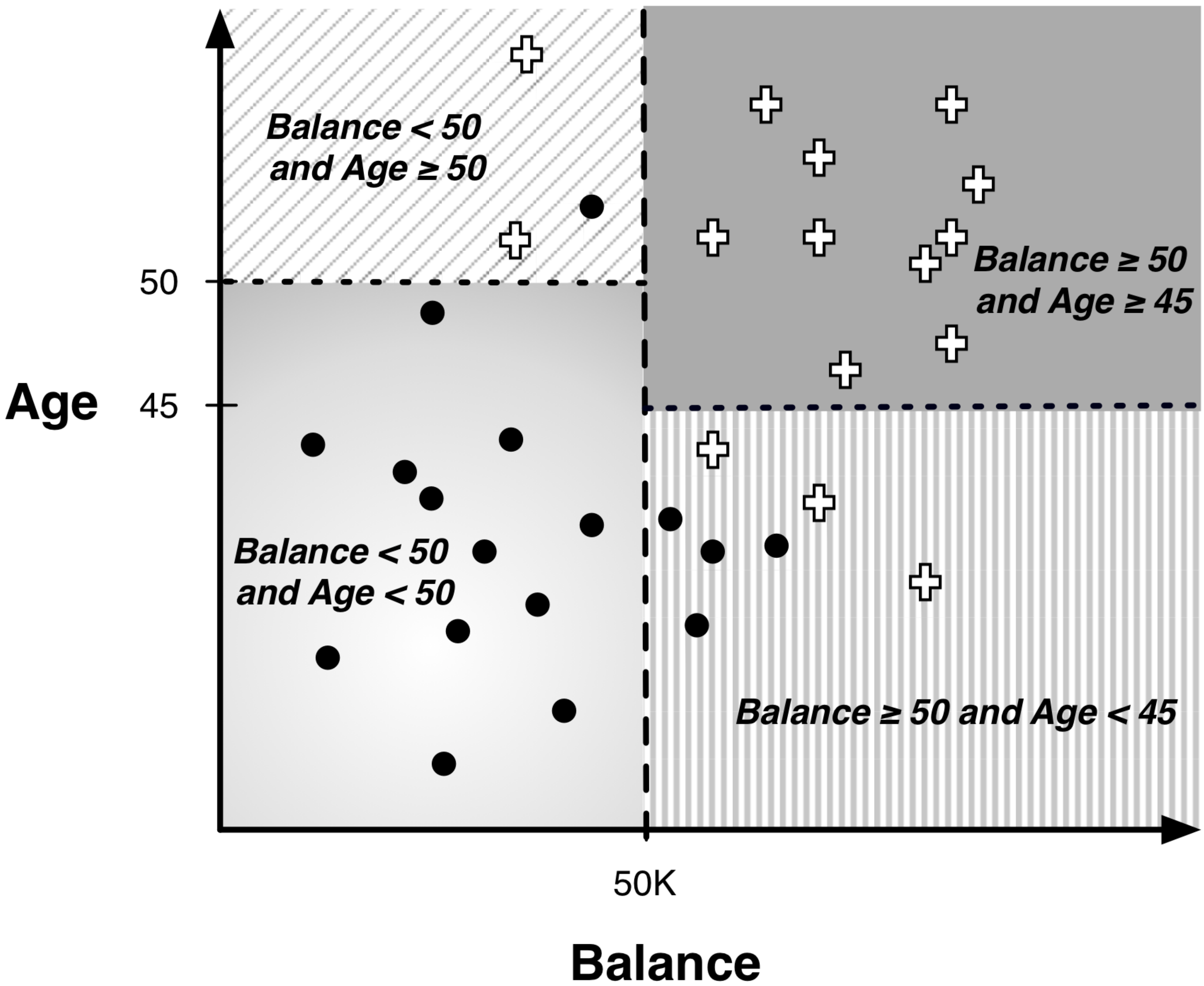
- Frequency-based estimate of class membership probability.
- Laplace correction: to moderate the influence of leaves with only a few instances.

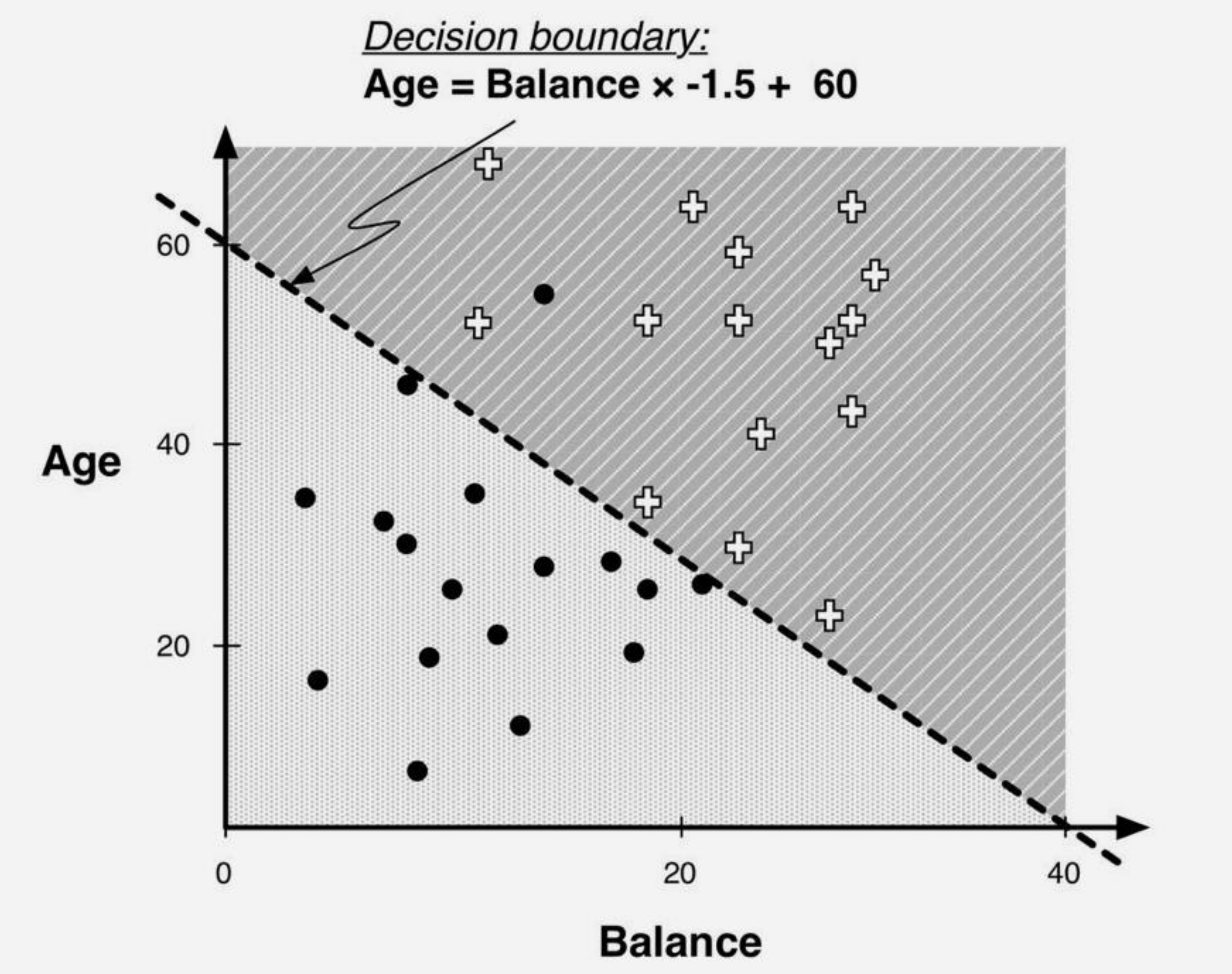
  $p(c) = (n+1)/(n+m+2)$


0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
HOUSE
OVERAGE
LONG_CALLS_PER_MON
LEFTOVER
INCOME
AVG_CALL_DURATION
HANDSET_PRICE
SATISFACTION
COLLEGE
USAGE_LEVEL


HOUSE
<= 600469
> 600469
OVERAGE
INCOME
<= 97
> 97
<= 100010
> 100010
LEFTOVER
prob = 0.79
(CHURN)
prob = 0.19
(STAY)
OVERAGE
<= 24
> 24
<= 144
> 144
LEFTOVER
prob = 0.61
(CHURN)
LEFTOVER
prob = 0.81
(CHURN)
0
1
<= 26
> 26
COLLEGE
prob = 0.29
(STAY)
OVERAGE
prob = 0.59
(CHURN)
No
Yes
<= 18
> 18
prob = 0.51
(CHURN)
INCOME
prob = 0.31
(STAY)
AVERAGE_CALL_DURATION
<= 59807
> 59807
<= 6
> 6
prob = 0.62
(CHURN)
prob = 0.49
(STAY)
prob = 0.55
(CHURN)
prob = 0.44
(STAY)

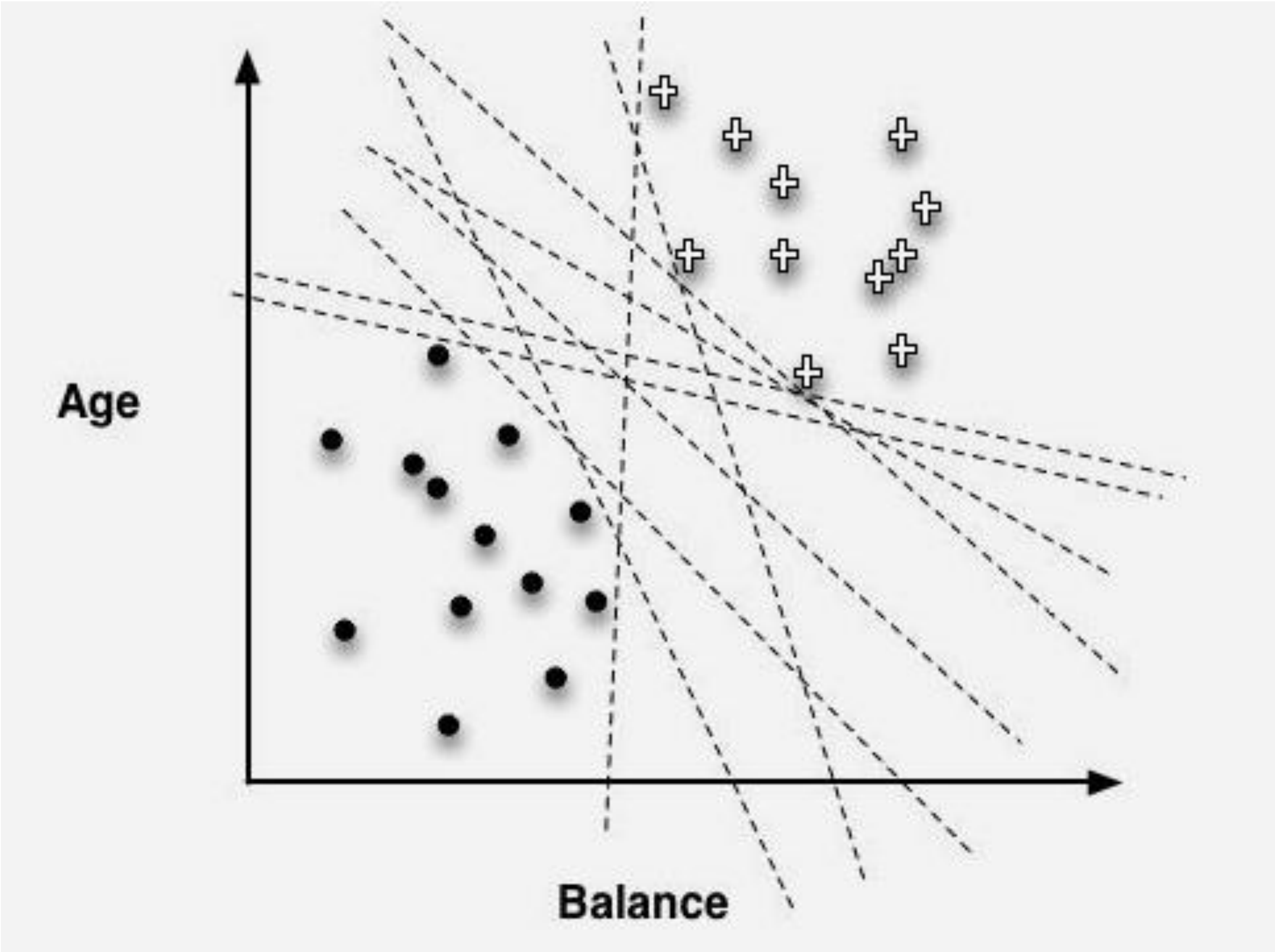
Linear Model


Balance < 50
and Age ≥ 50
Balance ≥ 50
and Age ≥ 45
50
45
Age
Balance < 50
and Age < 50
Balance ≥ 50 and Age < 45
50K
Balance



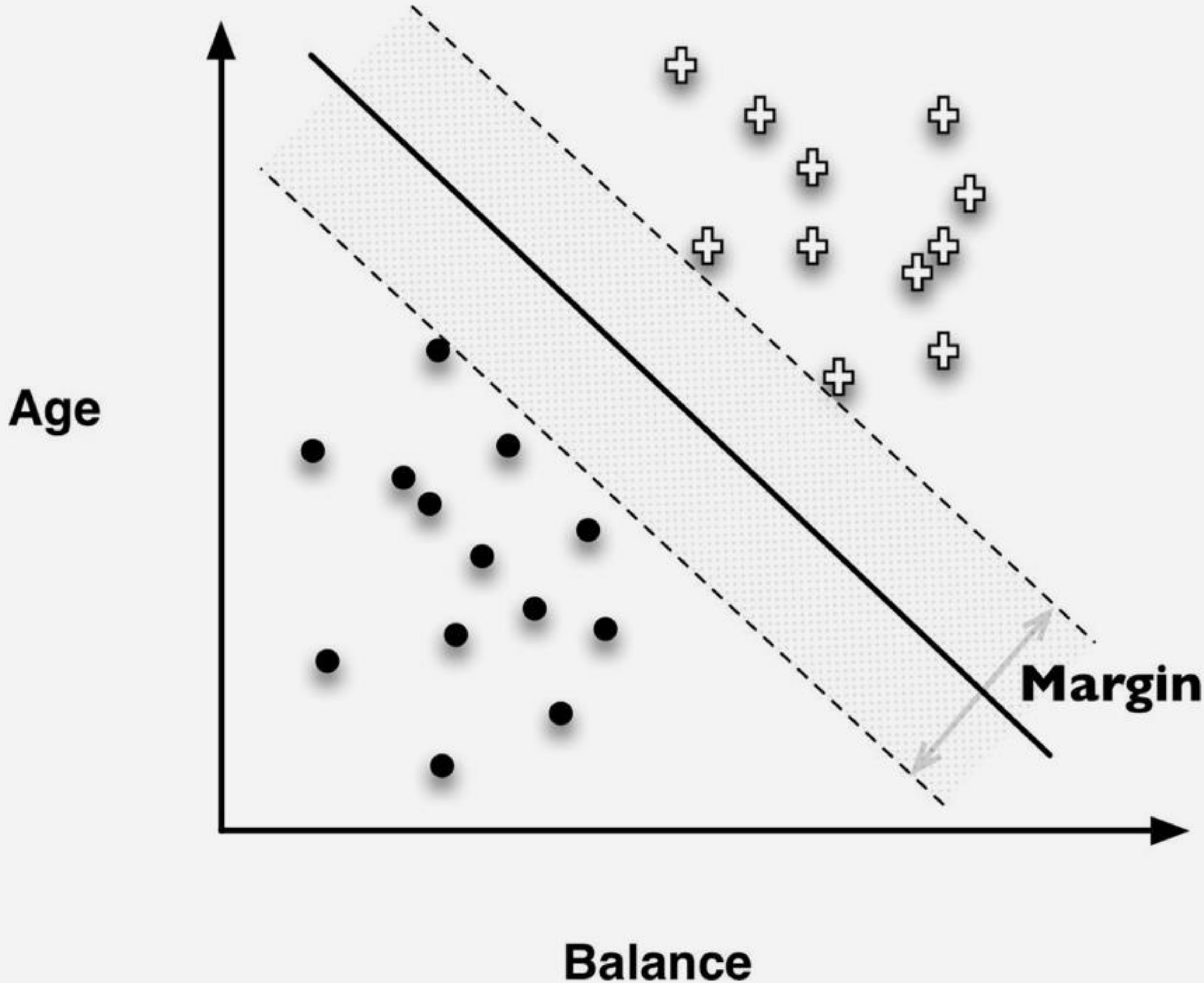
$$f(\mathbf{x}) = 60 - 1.0 \times Age - 1.5 \times Balance$$

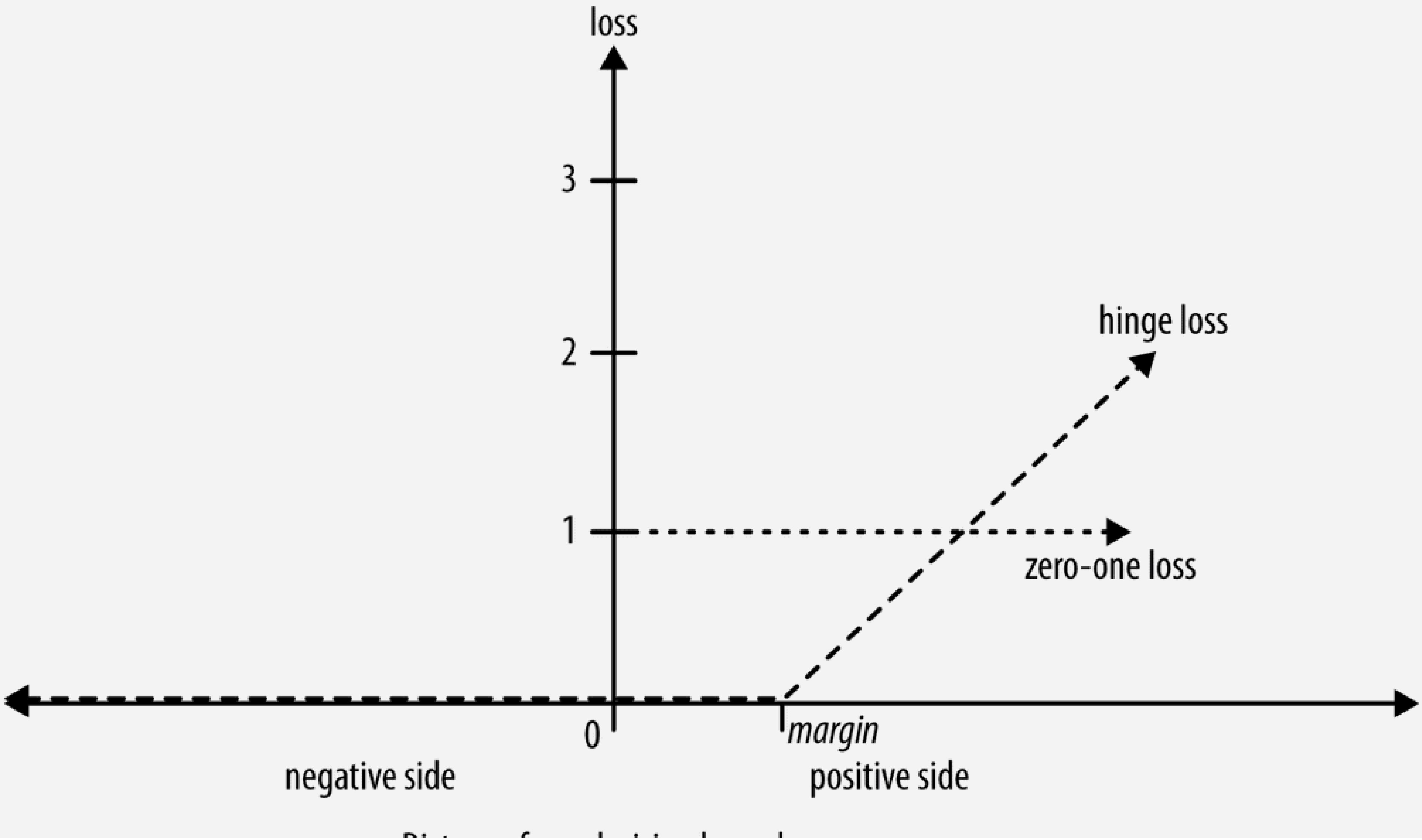
Which is the best line to fit the date?



Objective function

- Linear regression, support vector machines (SVM) and logistic regression are all trying to fit a linear model to data.
- The key difference is that each uses a different objective function.
- Linear regression -> least square
- SVM -> Maximizing the margin
- Logistic regression -> Maximum likehood


Age
Margin
Balance


loss
3
2
1
0
hinge loss
zero-one loss
margin
negative side
positive side

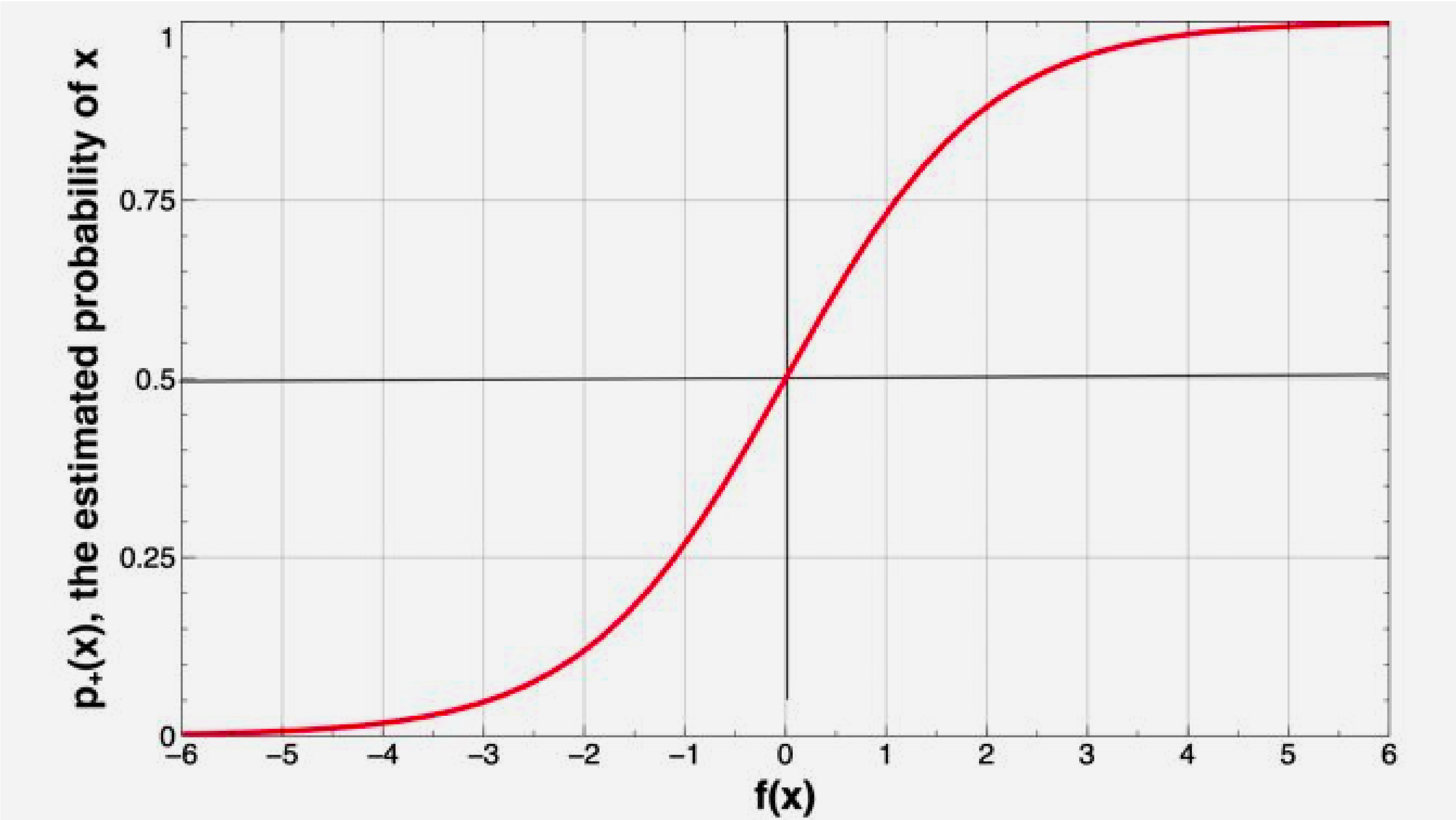
Class probability estimation

- What is the problem with simply using basic linear model to estimate the class probability?
- f(x) gives the distance from the separating boundary, ranging from $-\infty$ to $+\infty$, while a probability should range from 0 to 1.

Probability	Odds	Log-odds
0.5	50:50 or 1	0
0.9	90:10 or 9	2.19
0.999	999:1 or 999	6.9
0.01	1:99 or 0.0101	−4.6
0.001	1:999 or 0.001001	−6.9

$$\log\left(\frac{p_{+}(\mathbf{x})}{1-p_{+}(\mathbf{x})}\right) = f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots$$

$$p_{+}(\mathbf{x}) = \frac{1}{1+e^{-f(\mathbf{x})}}$$



p+(x), the estimated probability of x
1
0.75
0.5
0.25
0
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
f(x)

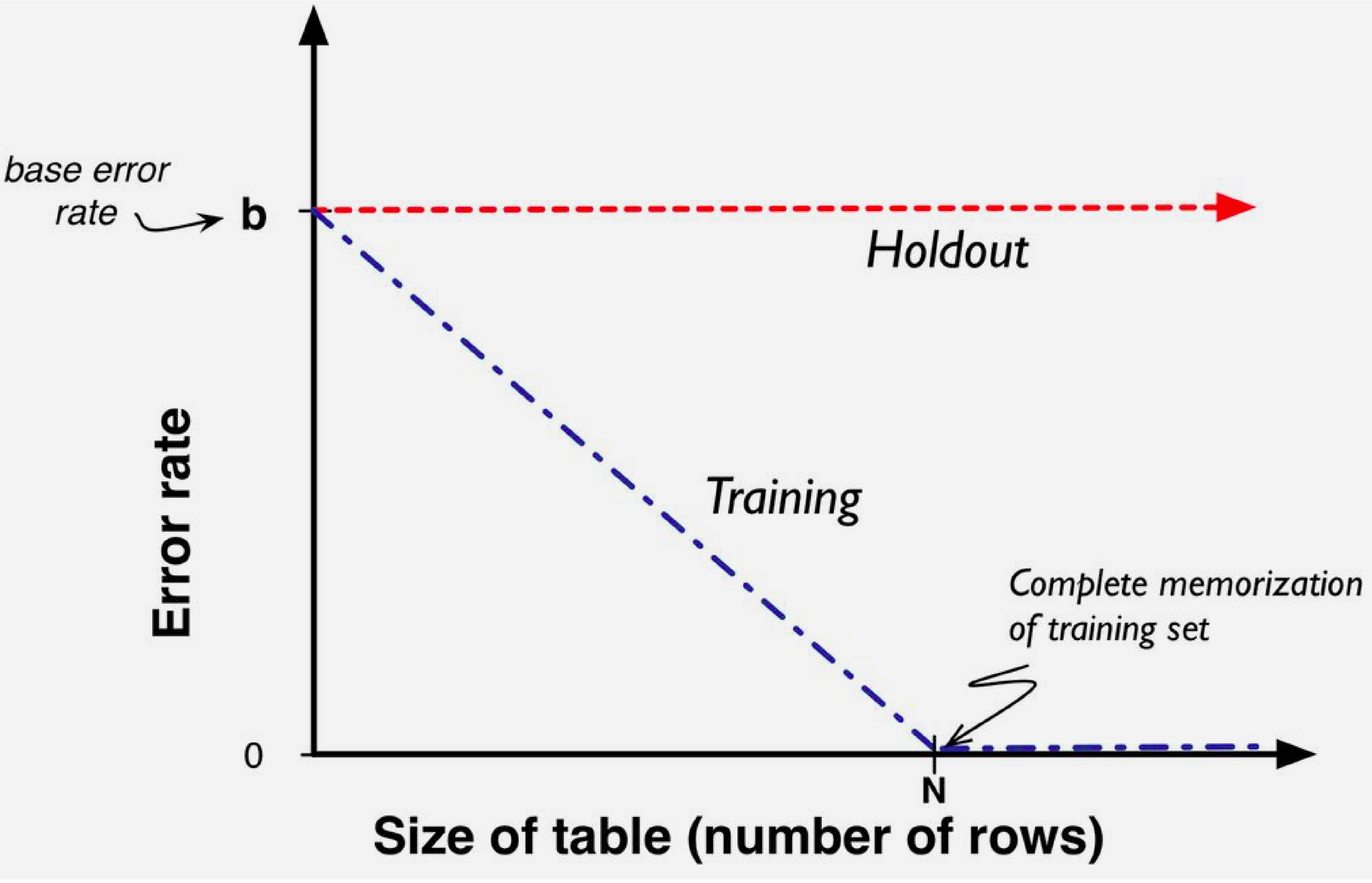
Objective function for Logistic Regression

$$g(\mathbf{x},\mathbf{w}) = \begin{cases} p_+(\mathbf{x}) & \text{if } \mathbf{x} \text{ is a } + \\ 1 - p_+(\mathbf{x}) & \text{if } \mathbf{x} \text{ is a } \bullet \end{cases}$$

Overfitting

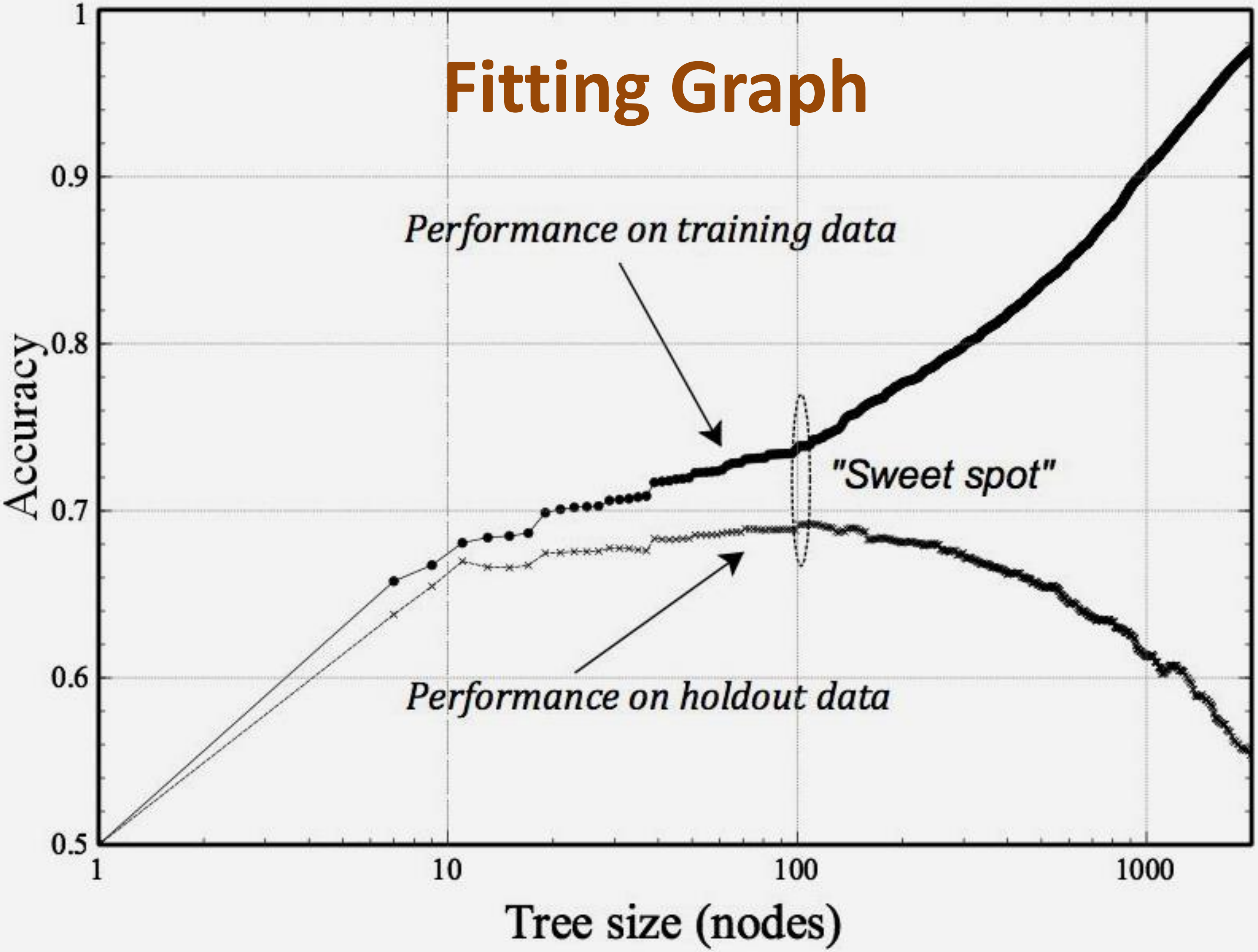
Overfitting

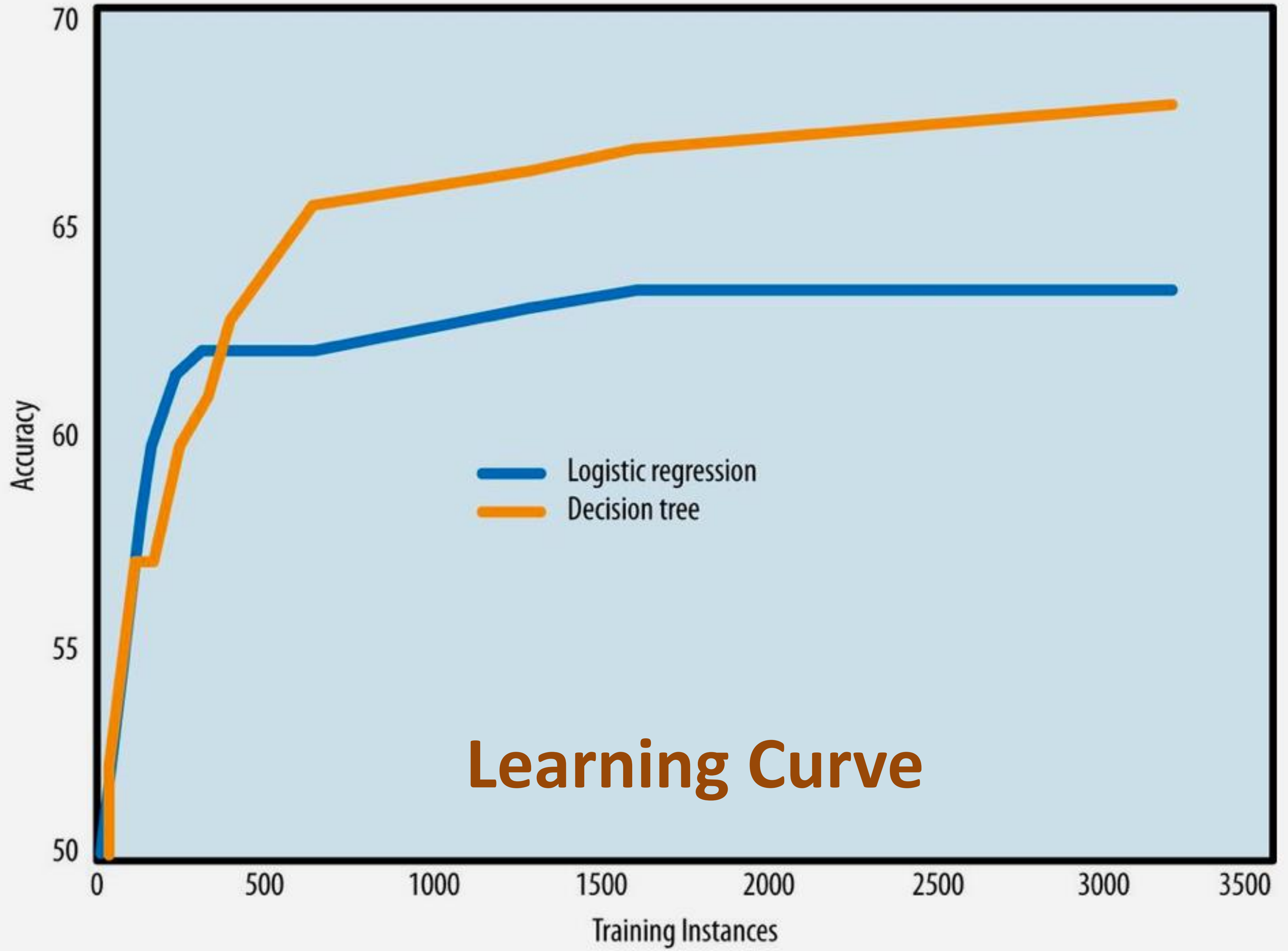
- If we allow ourselves enough flexibility in searching for patterns in a particular dataset, we will find patterns.
- These “patterns” may be just chance occurrences in the data, while we are interested in patterns that generalize – that predict well for instances that we have not yet observed.
- For example, a table model memorizes the training data and performs no generalization.



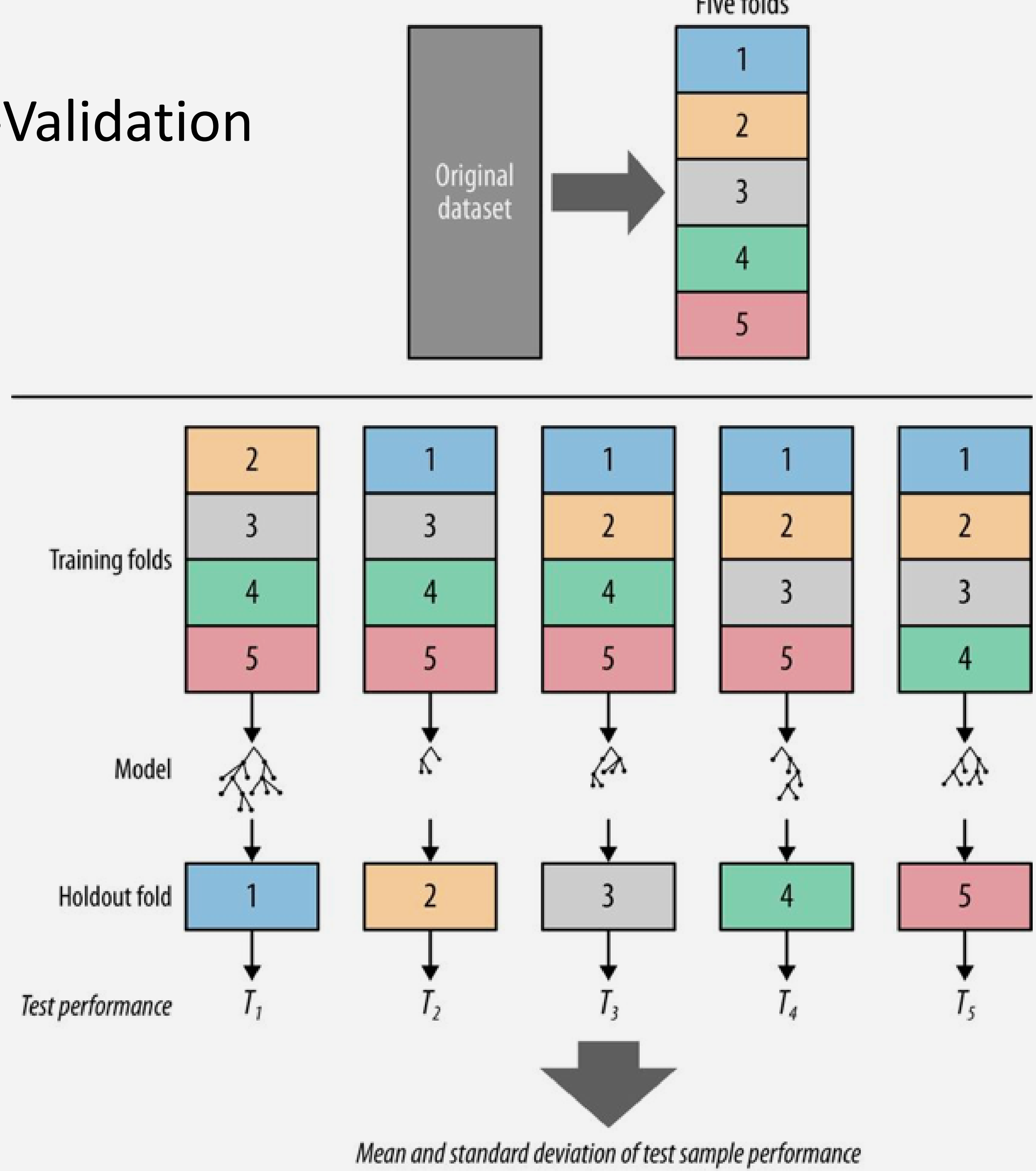
base error rate
b
Holdout
Training
Error rate
Complete memorization of training set
0
N
Size of table (number of rows)

Fitting Graph




70
65
60
55
50
Accuracy
0
500
1000
1500
2000
2500
3000
3500
Training Instances
Logistic regression
Decision tree
Learning Curve

Cross-Validation



What is a good model?

Evaluation

- Classifiers
- Ordering
- Class probability

Bad Positive & Harmless Negative

- Positive: worthy of attention
- Negative: uninteresting
- A classifier is screening through a large population consisting mostly of negatives and looking for a small number of positive instances.
- The number of mistakes made on negative examples (false positive) may be dominant, though the cost of each mistakes made on positive example (false negative) will be higher.

Accuracy?

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

Confusion Matrix

	p	**n**
Y	True positives	False positives
N	False negatives	True negatives

Problems with unbalanced classes

- A Tele company has churn rate of 0.01%.
- Is a model with 95% accuracy good?

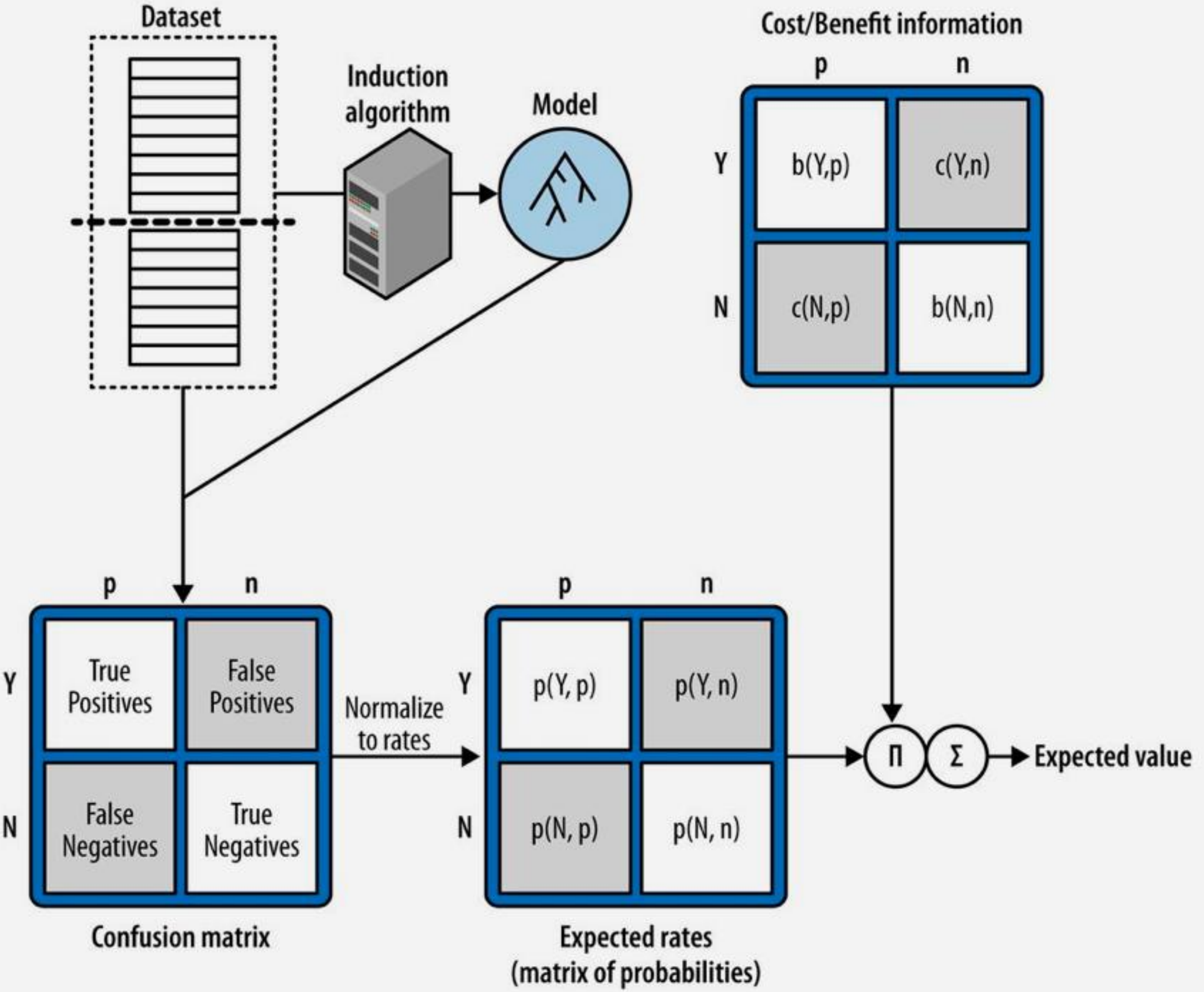
Problems with unequal costs and benefits

- How much we care about different errors? (false negatives & false positives)
- What are the good results for the correct prediction? (true positives & true negatives)
- The expected value is the weighted average of the values from different possible outcomes where the weight to each value is the probability of occurrence.

Example

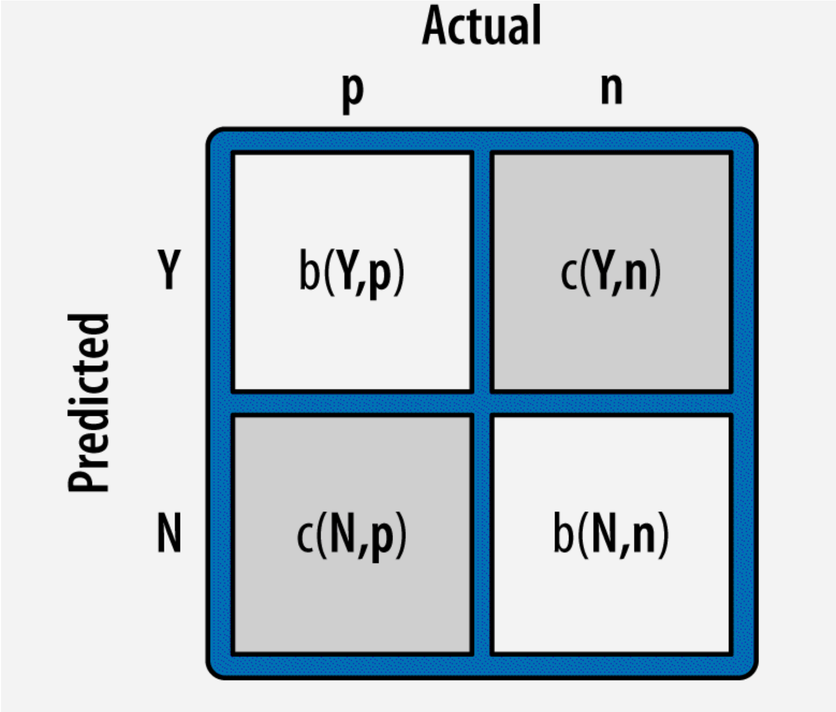
- Promotion cost: $1
- Product-related cost: $100
- Product price: $150

Who are the customers this campaign should target to make a profit?



Dataset
Induction algorithm
Model
Cost/Benefit information
p
n
Y
b(Y,p)
c(Y,n)
N
c(N,p)
b(N,n)
p
n
Y
True Positives
False Positives
N
False Negatives
True Negatives
Confusion matrix
Normalize to rates
p
n
Y
p(Y, p)
p(Y, n)
N
p(N, p)
p(N, n)
Expected rates (matrix of probabilities)
Π
Σ
Expected value

	p	**n**
Y	56	7
N	5	42



T = 110	
$p(\mathbf{Y},\mathbf{p}) = 56/110 = 0.51$	$p(\mathbf{Y},\mathbf{n}) = 7/110 = 0.06$
$p(\mathbf{N},\mathbf{p}) = 5/110 = 0.05$	$p(\mathbf{N},\mathbf{n}) = 42/110 = 0.38$

$$Expected\ profit = p(\mathbf{Y,p}) \cdot b(\mathbf{Y,p}) + p(\mathbf{N,p}) \cdot b(\mathbf{N,p}) + p(\mathbf{N,n}) \cdot b(\mathbf{N,n}) + p(\mathbf{Y,n}) \cdot b(\mathbf{Y,n})$$

$$p(x,\ y) = p(y) \cdot p(x \mid y)$$

$$\text{Expected profit} = p(\mathbf{Y|p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{Y,\ p}) + p(\mathbf{N|p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{N,p}) \cdot p(\mathbf{N|n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{N,n}) + p(\mathbf{Y|n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{Y,n})$$

$$\text{Expected profit} = p(\mathbf{p}) \cdot [p(\mathbf{Y|p}) \cdot b(\mathbf{Y,p}) + p(\mathbf{N|p}) \cdot b(\mathbf{N,p})] \cdot p(\mathbf{n}) \cdot [p(\mathbf{N|n}) \cdot b(\mathbf{N,n}) + p(\mathbf{Y|n}) \cdot b(\mathbf{Y,n})]$$

Correct & Error Rate

- True positive rate
- True negative rate
- False positive rate
- False negative rate

Correct & Error Rate

When the instance is actually positive

- True positive rate
- False negative rate

When the instance is actually negative

- True negative rate
- False positive rate

T = 110	
P = 61	N = 49
$p(\mathbf{p}) = 0.55$	$p(\mathbf{n}) = 0.45$
tp rate = 56/61 = 0.92	*fp rate* = 7/49 = 0.14
fn rate = 5/61 = 0.08	*tn rate* = 42/49 = 0.86

Sensitivity & Specificity

- Sensitivity: 有病者被檢出為有病的機率 (true positive rate)

  有病者: TP + FN

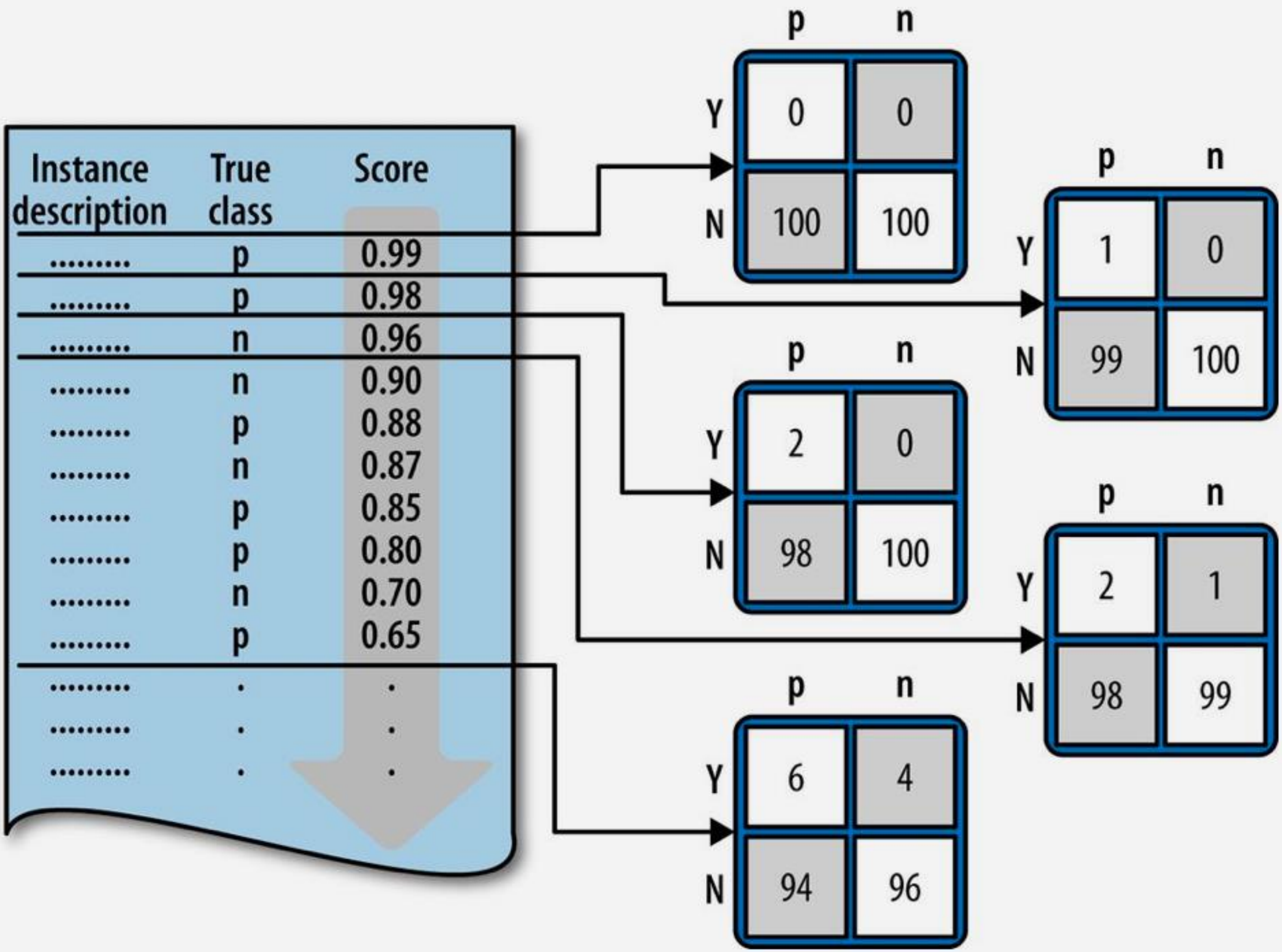
  Sensitivity 高 = FN 低, 即 false negative 少, 只要是檢驗結果是negative, 沒病機率高
- Specificity: 沒病者被檢出為沒病的機率

  沒病者: TN + FP

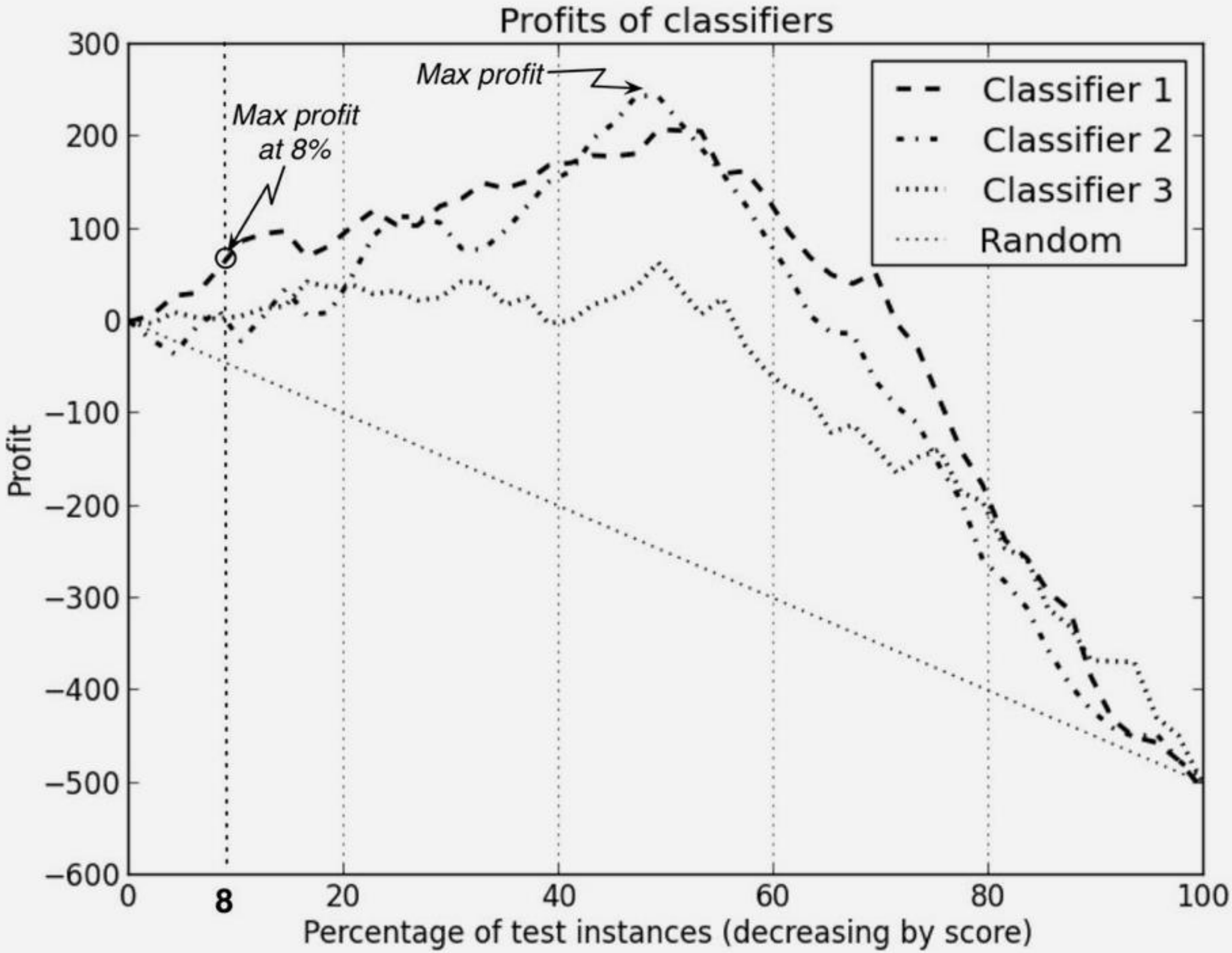
  Specificity 高 = FP低, 即 false positive 少,只要是檢驗結果是positive, 有病機率高

Evaluation

- Classifiers
- Ordering
- Class probability



Instance description
True class
Score
.... p 0.99
.... p 0.98
.... n 0.96
.... n 0.90
.... p 0.88
.... n 0.87
.... p 0.85
.... p 0.80
.... n 0.70
.... p 0.65
p
n
Y
N
0
0
100
100
1
0
99
100
2
0
98
100
2
1
98
99
6
4
94
96



Profits of classifiers
Max profit
Max profit
at 8%
Classifier 1
Classifier 2
Classifier 3
Random
300
200
100
0
−100
−200
−300
−400
−500
−600
Profit
0
8
20
40
60
80
100
Percentage of test instances (decreasing by score)

Which model do you choose?

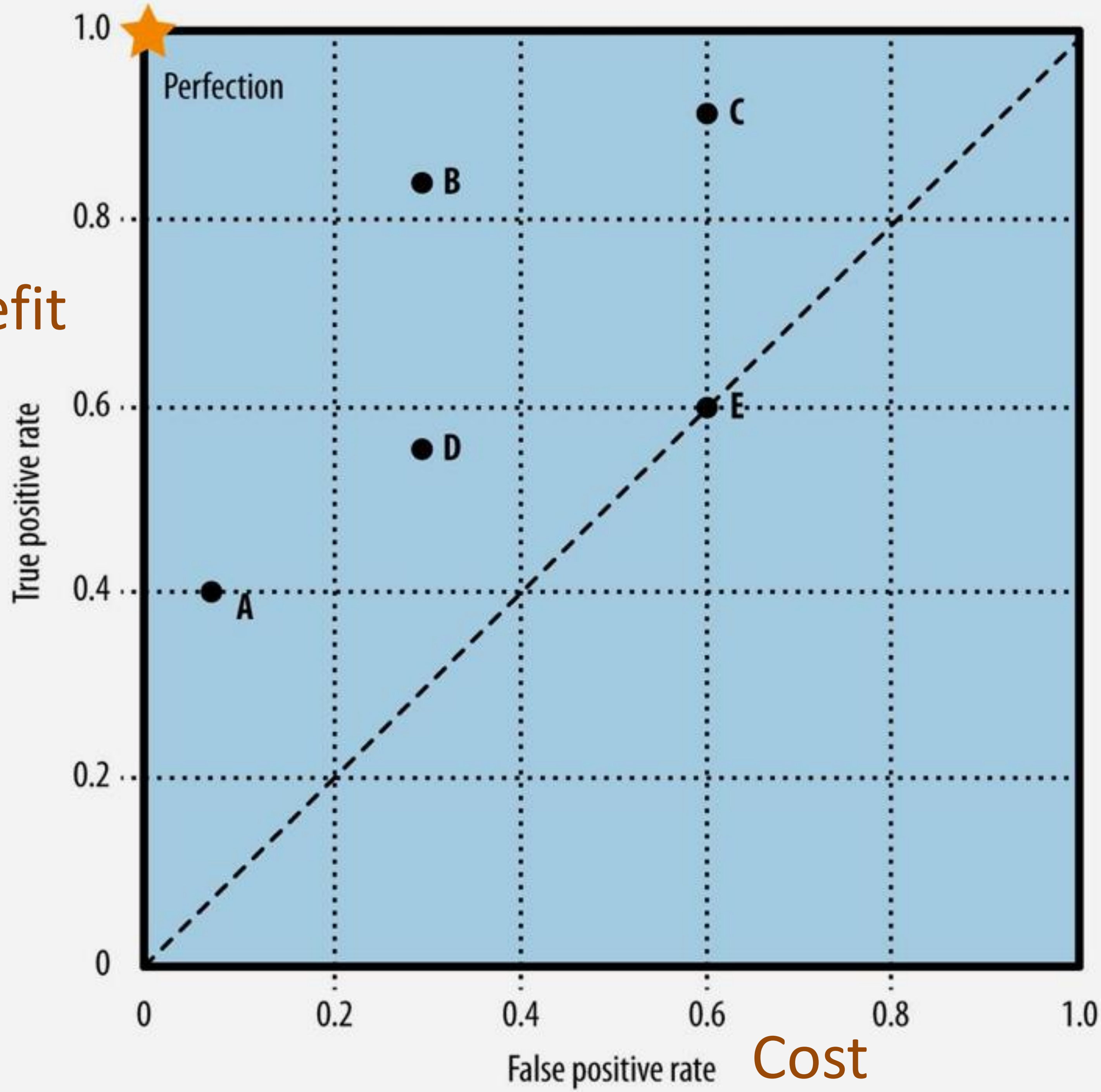
- If you have unlimited budget
- If you have 100,000 customers and \$40,000 budget (each offer costs \$5)

Limits of profit curve

- The class priors p(p) & p(n)

- The costs and benefits

ROC

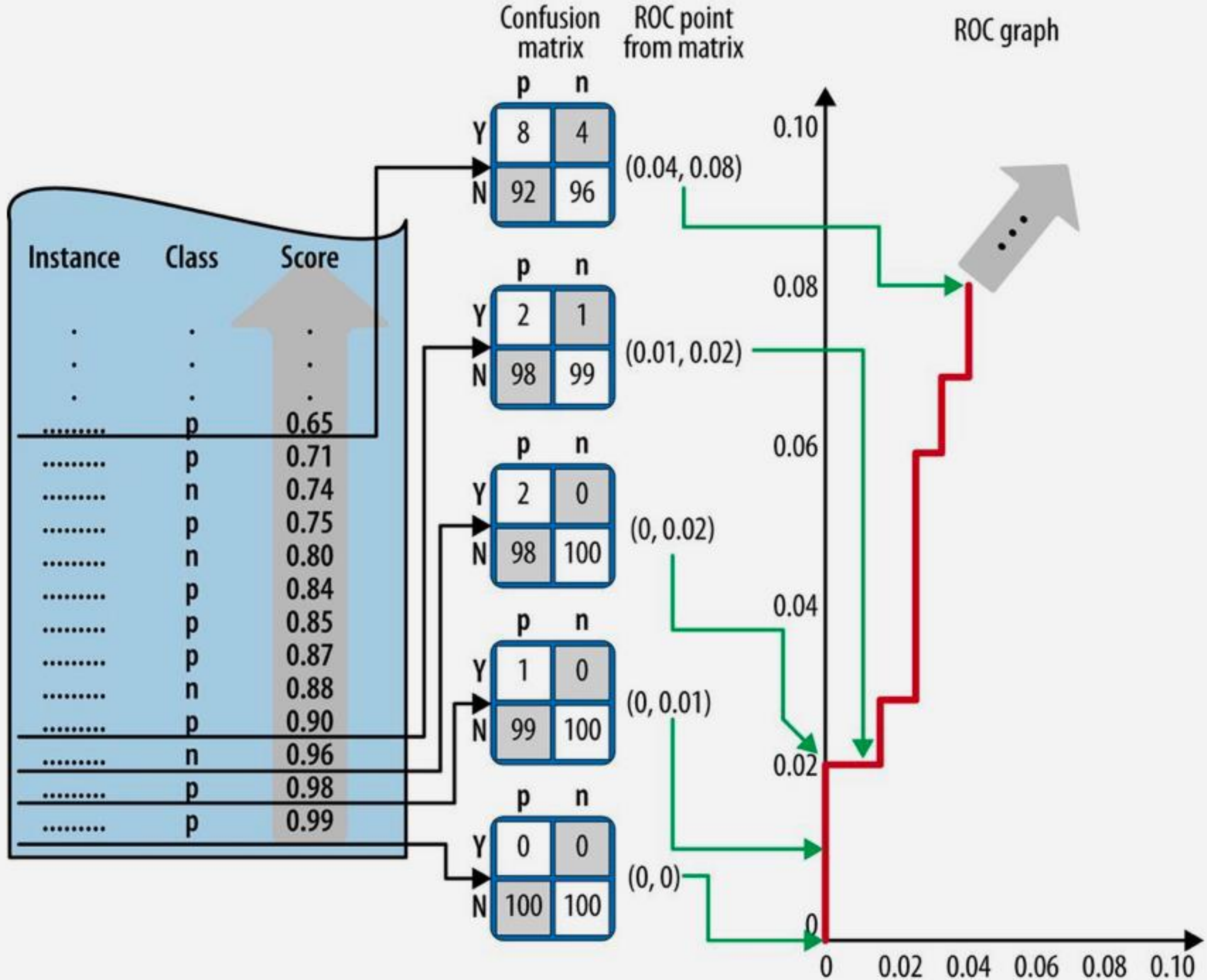


Benefit

Cost

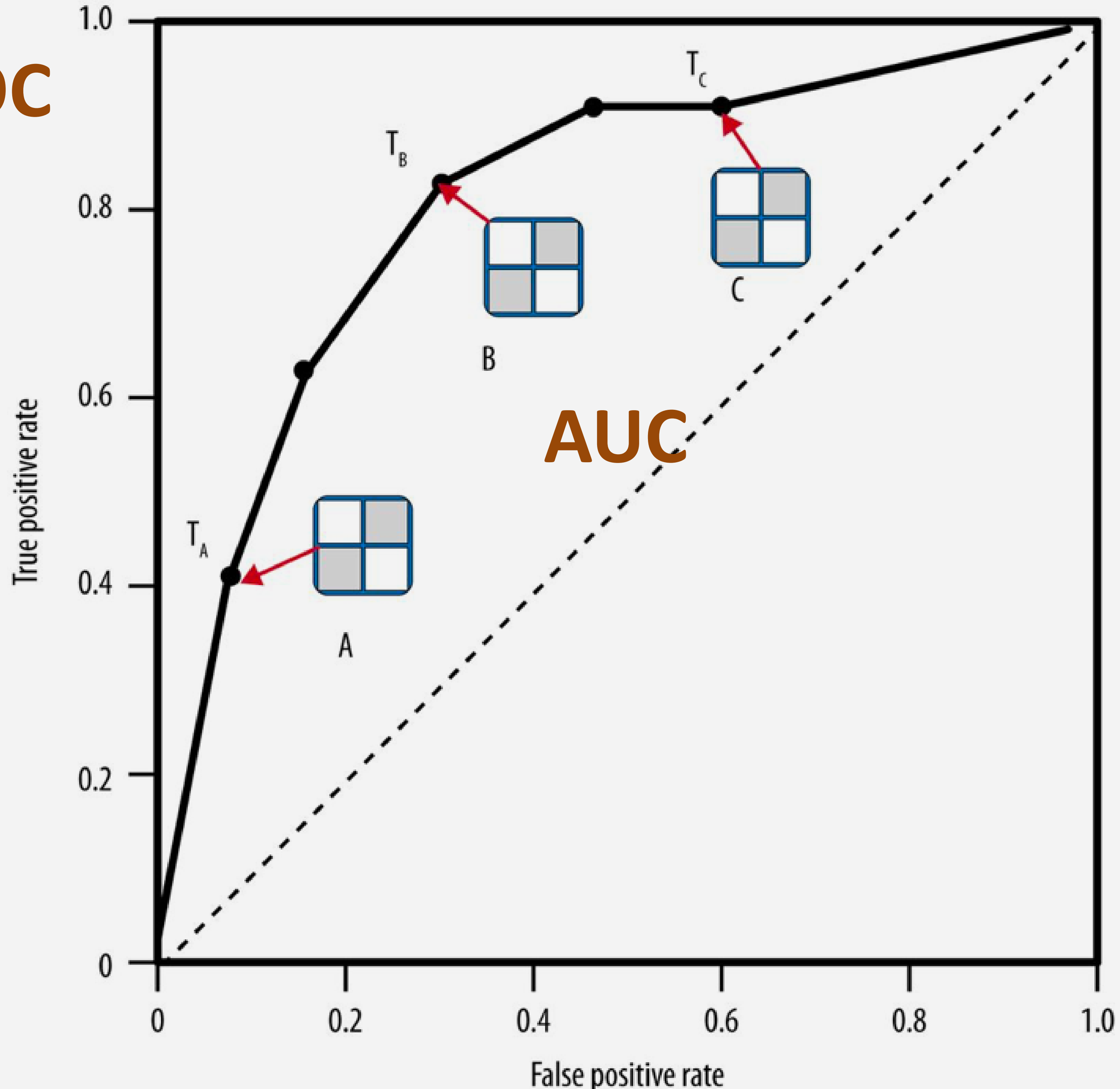
ROC

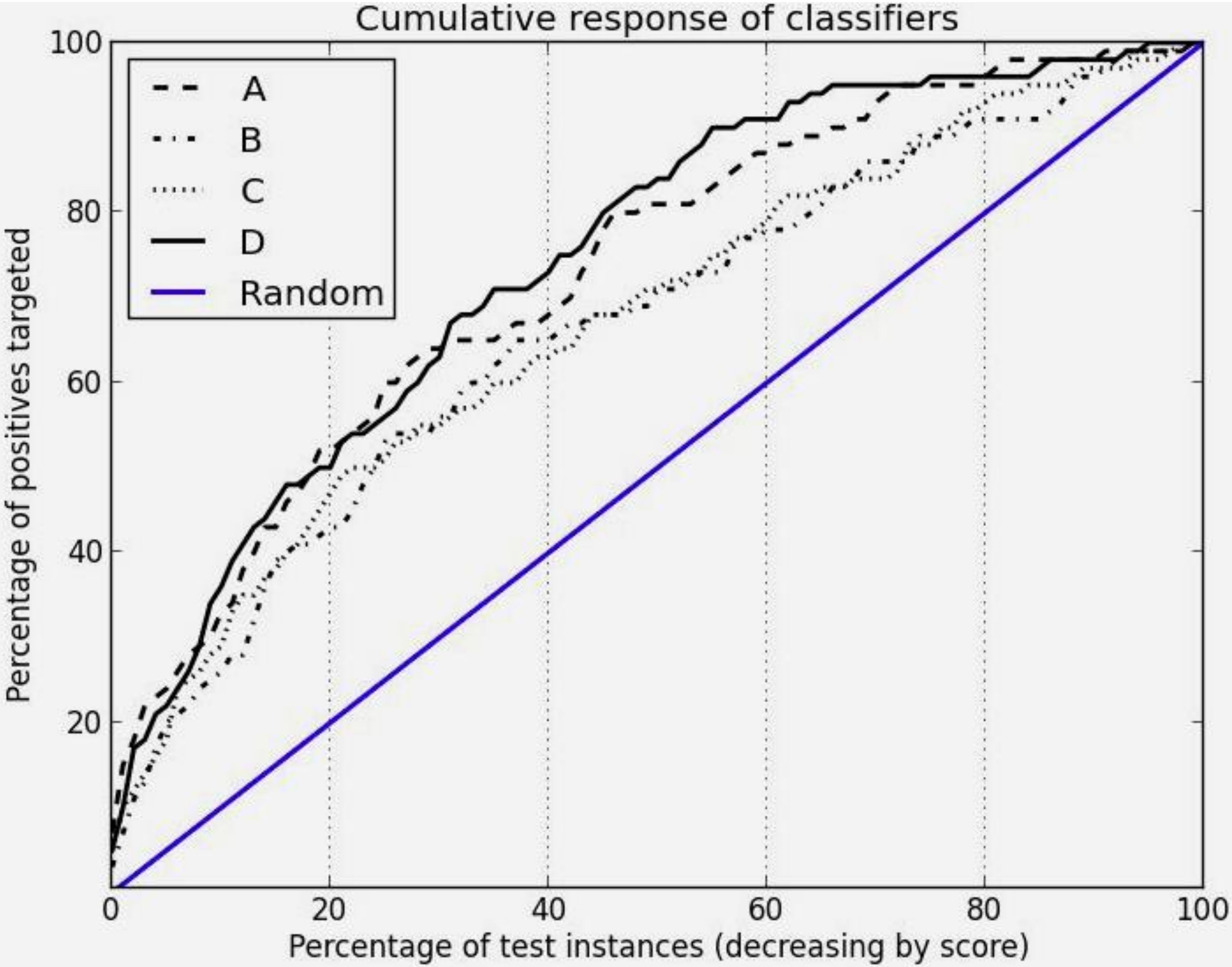
- (0,0): never issuing a positive
- (1,1): unconditionally issuing positives
- (1,0): perfect classification
- (0.5,0.5): guessing positive 50% of the time
- (0.9,0.9): guessing positive 90% of the time
- Many real-world domains are dominated by negative cases, so the performance in the far left-hand side of the ROC graph is more interesting than elsewhere.



Confusion matrix
ROC point from matrix
ROC graph
Instance
Class
Score
p n
Y 8 4
N 92 96
(0.04, 0.08)
p n
Y 2 1
N 98 99
(0.01, 0.02)
p n
Y 2 0
N 98 100
(0, 0.02)
p n
Y 1 0
N 99 100
(0, 0.01)
p n
Y 0 0
N 100 100
(0, 0)
p 0.65
p 0.71
n 0.74
p 0.75
n 0.80
p 0.84
p 0.85
p 0.87
n 0.88
p 0.90
n 0.96
p 0.98
p 0.99
0.10
0.08
0.06
0.04
0.02
0
0 0.02 0.04 0.06 0.08 0.10

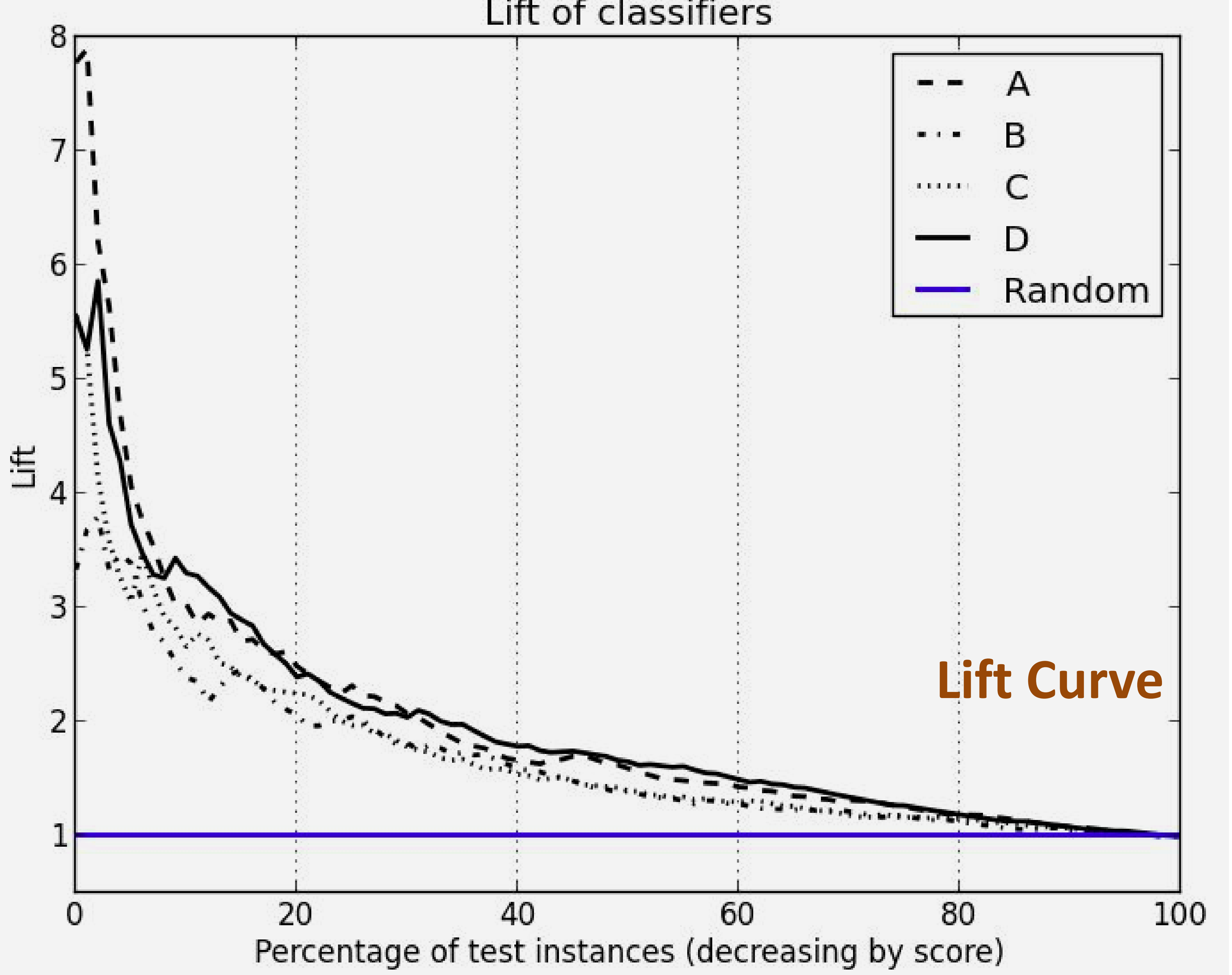
ROC


AUC
True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1.0
T_A
T_B
T_C
A
B
C


Cumulative response of classifiers
A
B
C
D
Random
Percentage of positives targeted
Percentage of test instances (decreasing by score)
0
20
40
60
80
100

Lift Curve

- The lift of a classifier represents the advantage it provides over random guessing.
- Consider a list of 100 customers, half of them are positive instances. If you scan down the list and stop halfway……
- If the list is sorted randomly, you would expect to see only half the positives. (a lift of 1).
- If the classifier is perfect, the lift is 2.
- Unlike for ROC curves, the lift curve assumes that the test set has exactly the same target class priors as the population to which the model will be applied.


Lift of classifiers
Lift
Percentage of test instances (decreasing by score)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
A
B
C
D
Random
Lift Curve

Business Problems and Analysis Tasks

Data-Analytic Thinking

- Business problems rarely are classification problems, regression problems or clustering problems.
- View business problems from a data perspective with structure and principles to guide you to the solutions.
- Bring together intuition, creativity, common sense and domain knowledge.

Targeting the best prospects for a charity mailing

- Would we want to model the probability of responding to the offer for each prospective donor.
- Would we like to maximize the total amount of donations? (the amount for this campaign or for the lifetime)
- Would we like to maximize the total amount of profit?

$$\text{Expected benefit of targeting} = p(R \mid \mathbf{x}) \cdot v_R + [1 - p(R \mid \mathbf{x})] \cdot v_{NR}$$

$$\text{Expected benefit of targeting} = p(R \mid \mathbf{x}) \cdot v_R(\mathbf{x}) + [1 - p(R \mid \mathbf{x})] \cdot v_{NR}(\mathbf{x})$$

$$\begin{aligned}
p(R \mid \mathbf{x}) \cdot (d_R(\mathbf{x}) - c) + [1 - p(R \mid \mathbf{x})] \cdot (-c) &> 0 \\
p(R \mid \mathbf{x}) \cdot d_R(\mathbf{x}) - p(R \mid \mathbf{x}) \cdot c - c + p(R \mid \mathbf{x}) \cdot c &> 0 \\
p(R \mid \mathbf{x}) \cdot d_R(\mathbf{x}) &> c
\end{aligned}$$

Selection Bias

- The data we are using to build the model may well be biased – meaning that they are selected randomly from the population to which you intend to apply the model.
- The data are from past donation – from the individual who did respond in the past.
- Some people may donate $10 each and every time they're asked. Others may give $100 and then feel they need not donate for a while.

Targeting the offer to customers who would churn when their contracts expire

- Do we really want to target our offer to those with the highest probability of defection?
- Targeting those with the highest value?
- Targeting those whom we would lost the most value If they were to leave.
- Assessing the expected benefit of not targeting.

$$EB_T(\mathbf{x}) = p(S \mid \mathbf{x},\ T) \cdot (u_s(\mathbf{x}) - c) + [1 - p(S \mid \mathbf{x},\ T)] \cdot (u_{NS}(\mathbf{x}) - c)$$

$$EB_{notT}(\mathbf{x}) = p(S \mid \mathbf{x},\ notT) \cdot u_s(\mathbf{x}) + [1 - p(S \mid \mathbf{x},\ notT)] \cdot u_{NS}(\mathbf{x})$$

$$\begin{aligned} VT &= p(S \mid \mathbf{x},\ T) \cdot u_S(\mathbf{x}) - p(S \mid \mathbf{x},\ notT) \cdot u_S(\mathbf{x}) - c \\ &= [p(S \mid \mathbf{x},\ T) - p(S \mid \mathbf{x},\ notT)] \cdot u_S(\mathbf{x}) - c \\ &= \Delta(p) \cdot u_S(\mathbf{x}) - c \end{aligned}$$

Competitive Advantage

Differences

- Netflix vs. Blockbuster
- Amazon vs. Barnes & Noble
- Amazon vs. eBay

Sustaining competitive advantages via data science

- Data and data science capabilities are assets to every company? (Web to Dell vs. Web to Compaq)
- Your competitors enjoy the same value from these assets?
- Formidable historical advantage; unique intellectual property; complementary assets

KSF

- Data and the capability to extract useful knowledge from data should be regarded as key strategic assets.
- Management must think data-analytically
- Management must create a culture to nurture data science and data scientists

The material is excerpted from
“Data Science for Business”
Foster Provost and Tom Fawcett