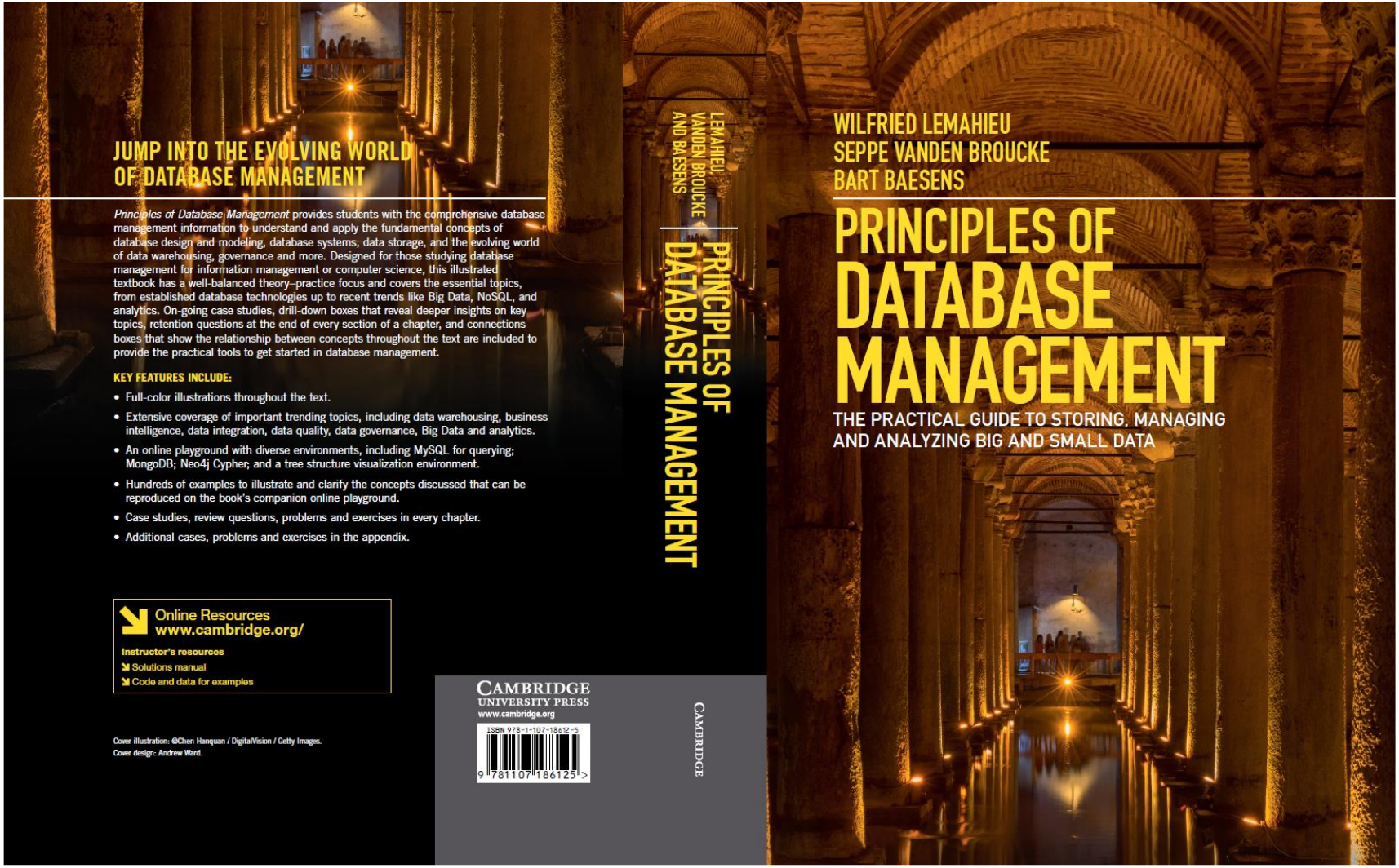
Analytics

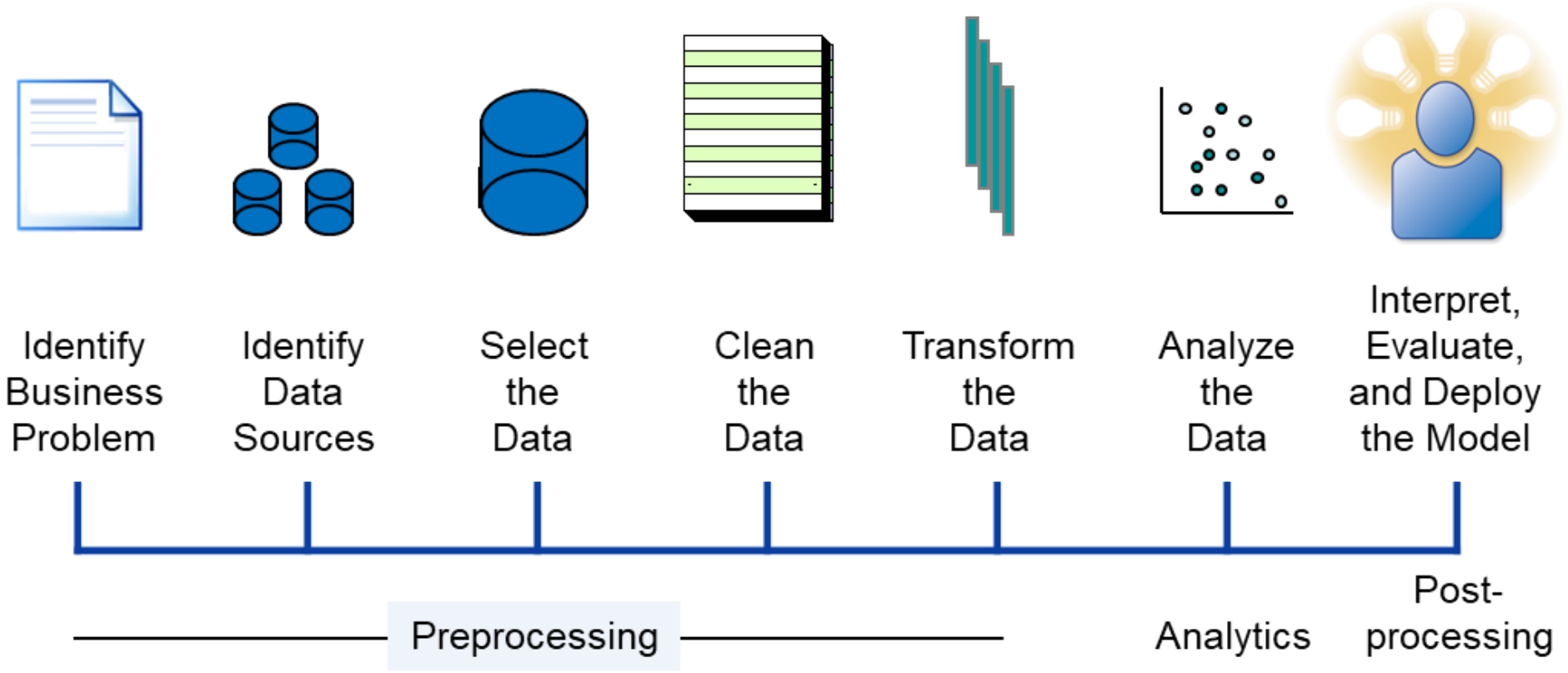


www.pdbmbook.com

Introduction

- Analytics Process Model
- Example Analytics Applications
- Data Scientist Job Profile
- Data Preprocessing
- Types of Analytics
- Post Processing of Analytical Models
- Critical Success Factors for Analytical Models
- Economic Perspective On Analytics
- Improving the ROI of Analytics
- Privacy and Security

Analytics Process Model



Example Analytics Applications

- Risk analytics
  - credit scoring
  - fraud detection
- Marketing analytics
  - churn prediction
  - response modeling
  - customer segmentation
- Recommender systems
- Texts analytics

Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net Lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business Process analytics
Retention modeling	Operational risk modeling	Money Laundering	Multivariate testing		HR analytics
Market basket analysis	Fraud detection	Terrorism detection			Healthcare analytics
Recommender systems					Learning analytics
Customer segmentation					

Data Scientist Job Profile

- Statistics, machine learning and/or quantitative modeling
- Programming
- Communication/Visualization
- Business Knowledge
- Creativity

Data Preprocessing

- Denormalizing data for analysis
- Sampling
- Exploratory Analysis
- Missing values
- Outlier Detection and Handing

Denormalizing data for analysis

Transactions		
ID	**Date**	**Amount**
XWV	2/01/2015	52 €
XWV	6/02/2015	21 €
XWV	3/03/2015	13 €
BBC	17/02/2015	45 €
BBC	1/03/2015	75 €
VVQ	2/03/2015	56 €

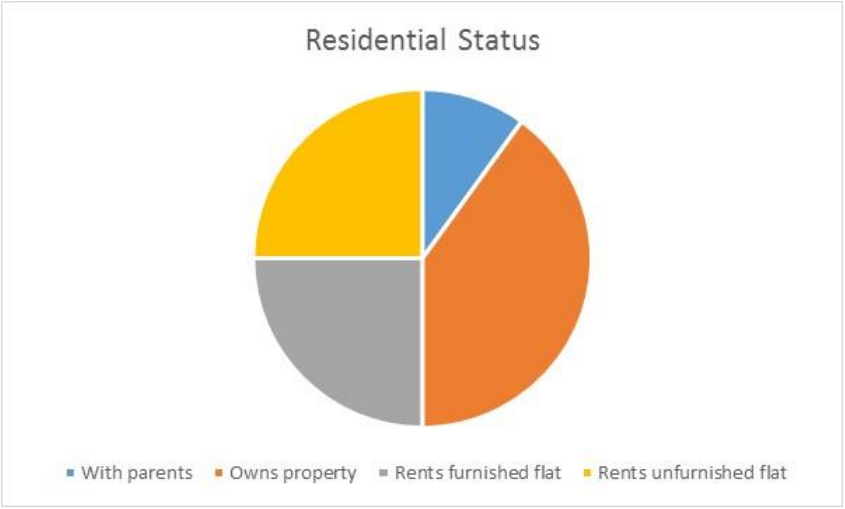
Customer data		
ID	**Age**	**Start date**
XWV	31	1/01/2015
BBC	49	10/02/2015
VVQ	21	15/02/2015

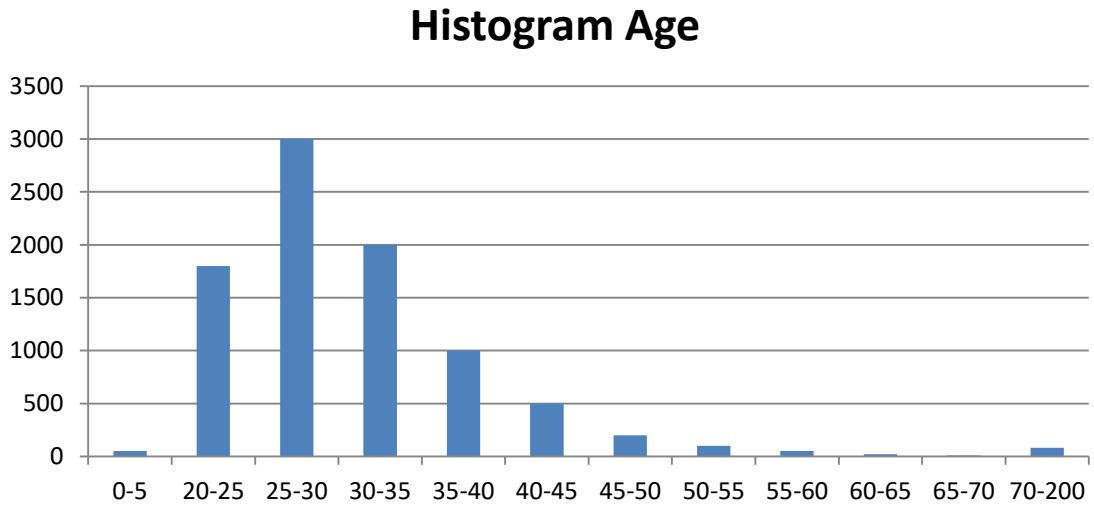
Non-normalized data table				
ID	**Date**	**Amount**	**Age**	**Start date**
XWV	2/01/2015	52 €	31	1/01/2015
XWV	6/02/2015	21 €	31	1/01/2015
XWV	3/03/2015	13 €	31	1/01/2015
BBC	17/02/2015	45 €	49	10/02/2015
BBC	1/03/2015	75 €	49	10/02/2015
VVQ	2/03/2015	56 €	21	15/02/2015

Sampling

- Take a subset of historical data to build analytical model
- Good sample should be representative for the future entities on which the analytical model will be run
- Choosing the optimal time window of the sample involves a trade-off between lots of data and recent data

Exploratory Analysis





Exploratory Analysis

- Descriptive statistics
  - Mean
  - Median
  - Mode
  - Standard deviation
  - Percentile values

Missing Values

ID	Age	Income	Marital status	Credit bureau score	Fraud
1	34	1800	?	620	Yes
2	28	1200	Single	?	No
3	22	1000	Single	?	No
4	60	2200	Widowed	700	Yes
5	58	2000	Married	?	No
6	44	?	?	?	No
7	22	1200	Single	?	No
8	26	1500	Married	350	No
9	34	?	Single	?	Yes
10	50	2100	Divorced	?	No

Missing Values

- Keep
- Delete (observation or variable)
- Replace (aka impute)

Outlier Detection and Handling

- Valid versus invalid observations
- Outlier detection
  - Minimum/Maximum
  - Histogram, box plot, scatter plot
- Outlier handling
  - Treat as missing value (invalid observation)
  - Capping (valid observation)

Types of Analytics

- Predictive Analytics
- Evaluating Predictive Models
- Descriptive Analytics
- Social Network Analytics

Predictive Analytics

- Linear Regression
- Logistic Regression
- Decision Trees
- Other predictive analytics techniques

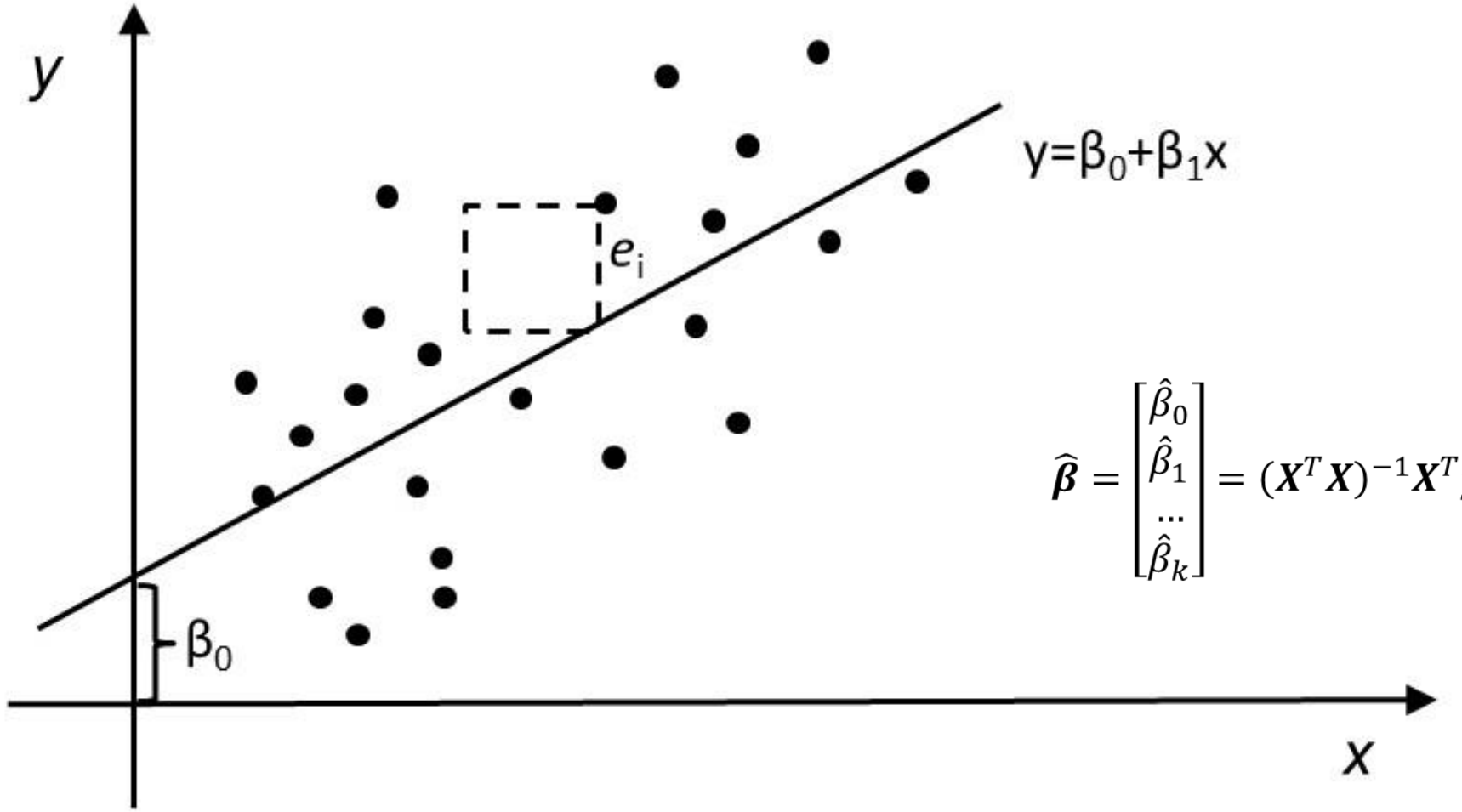
Linear Regression

- $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

	x_1	x_2	…	x_k	y
x_1	$x_1(^1)$	$x_1(^2)$	…	$x_1(^k)$	y_1
x_2	$x_2(^1)$	$x_2(^2)$		$x_2(^k)$	y_2
….	….	….	….	….	….
x_n	$x_n(^1)$	$x_n(^2)$		$x_n(^k)$	y_n

- $\frac{1}{2}\sum_{i=1}^{n} e_i^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - (\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i))^2$

Linear Regression



$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

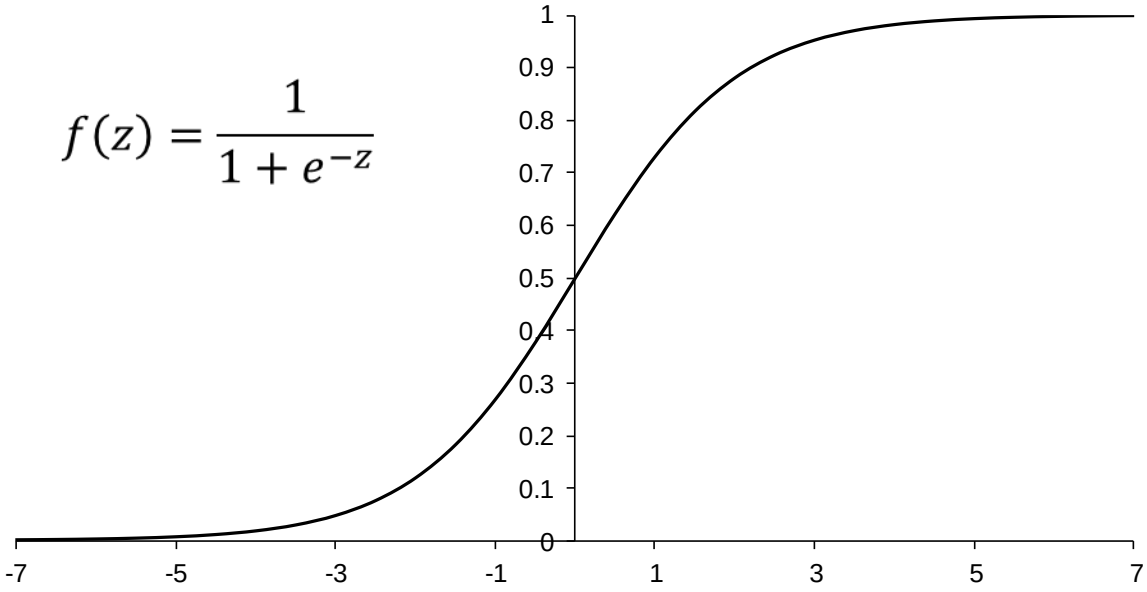
Logistic Regression

Customer	Age	Income	Gender	…	Response	y
John	30	1200	M		No	0
Sarah	25	800	F		Yes	1
Sophie	52	2200	F		Yes	1
David	48	2000	M		No	0
Peter	34	1800	M		Yes	1

$$y = \beta_0 + \beta_1 Age + \beta_2 Income + \beta_3 Gender$$

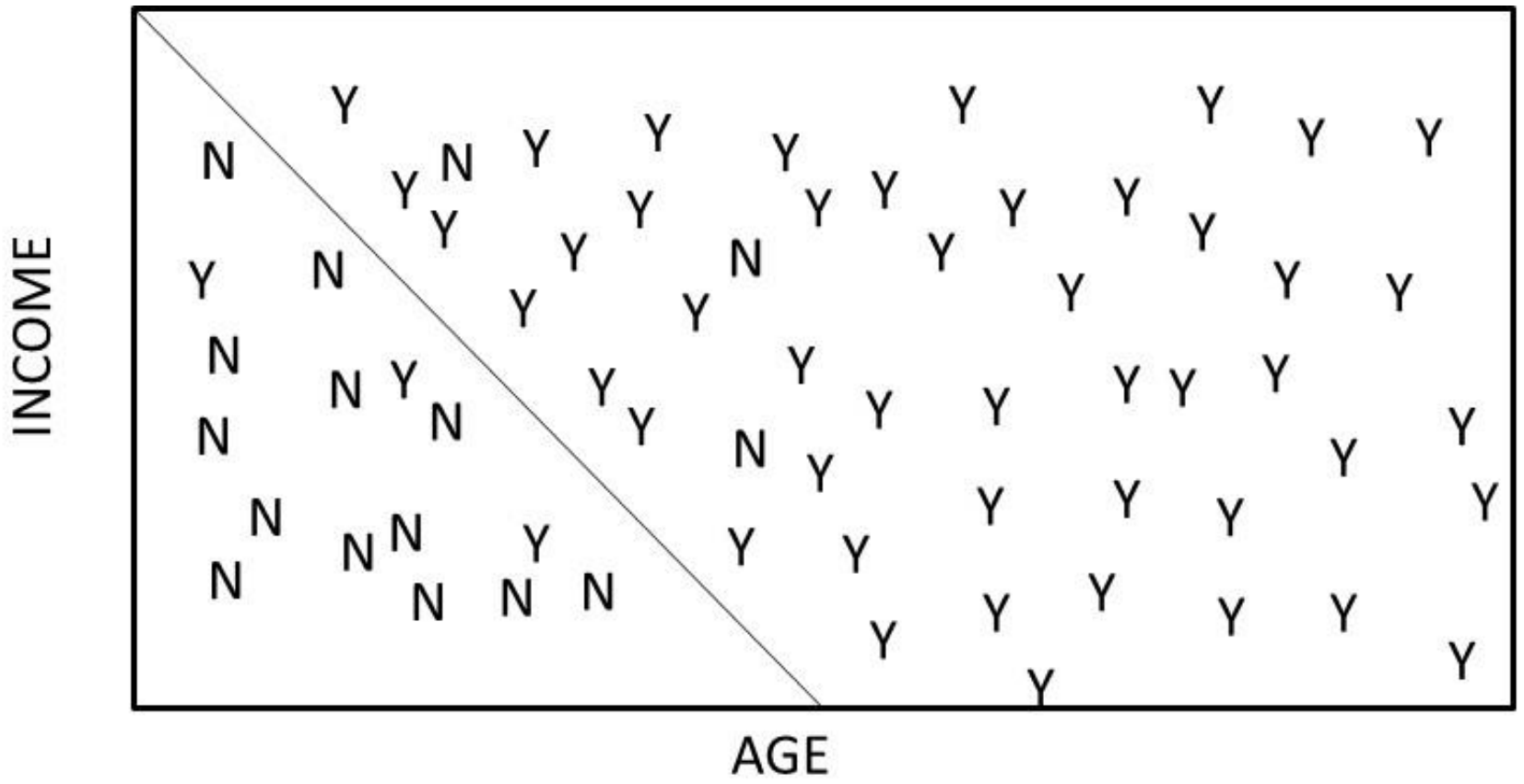
Logistic Regression

$$f(z) = \frac{1}{1 + e^{-z}}$$



$$p(response = yes|Age, Income, Gender) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Gender})}}$$

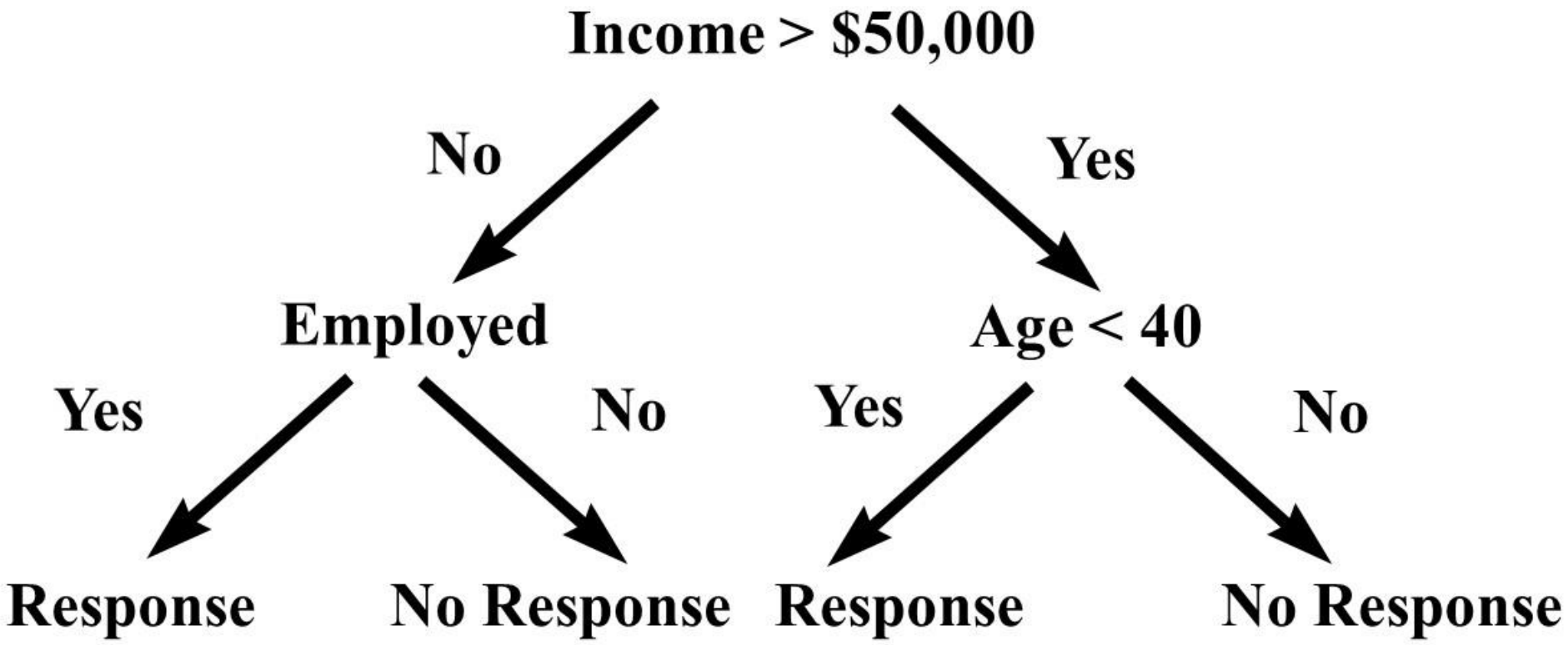
Logistic Regression



Logistic Regression

- Odds ratio
  - e^{β_i}
  - multiplicative increase in the odds when a variable increases by 1 (ceteris paribus)
- Doubling amount
  - $\log(2) / \beta_i$
  - amount of change required for doubling primary outcome odds

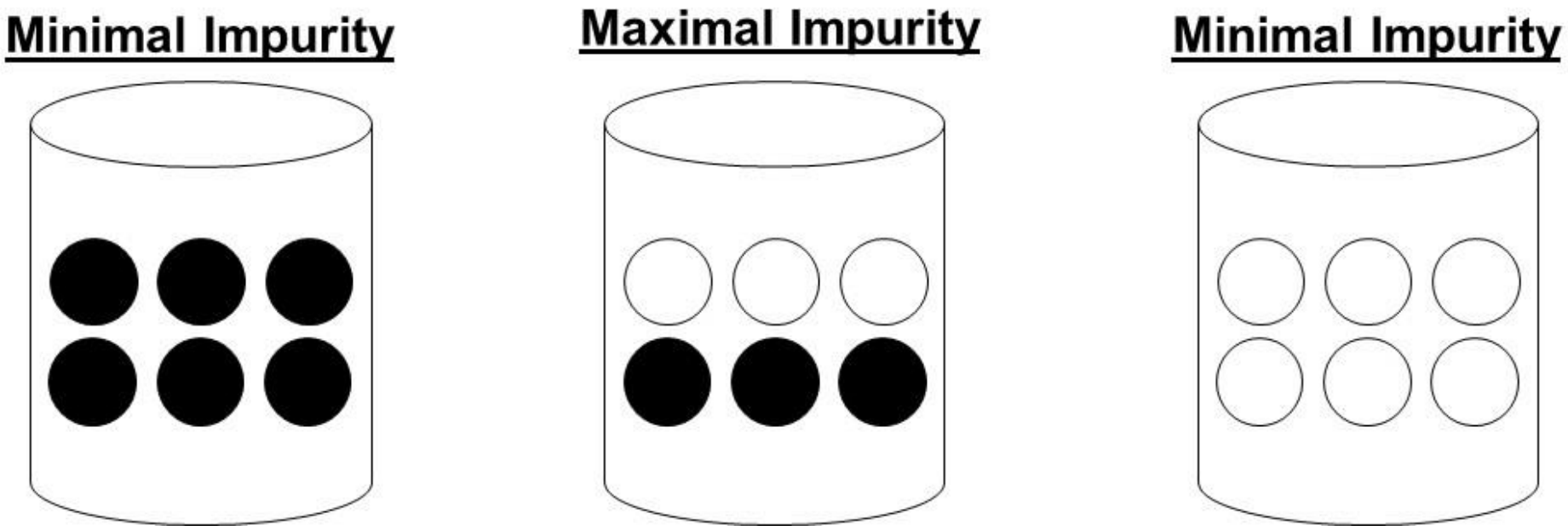
Decision Trees



Decision Trees

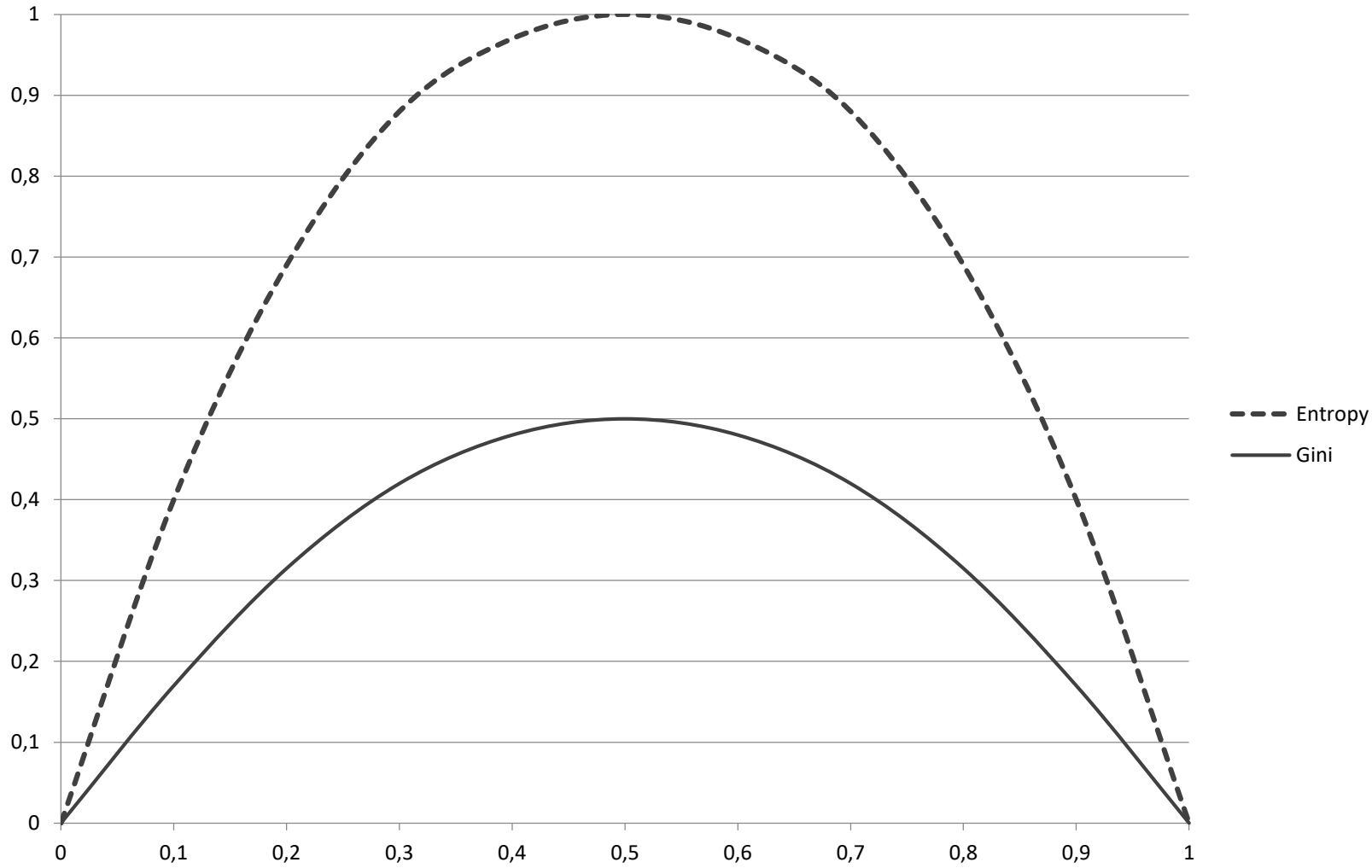
- Splitting decision
  - Which variable to split at what value
- Stopping decision
  - When to stop adding nodes to the tree?
- Assignment decision
  - What class to assign to a leaf node?

Decision Trees

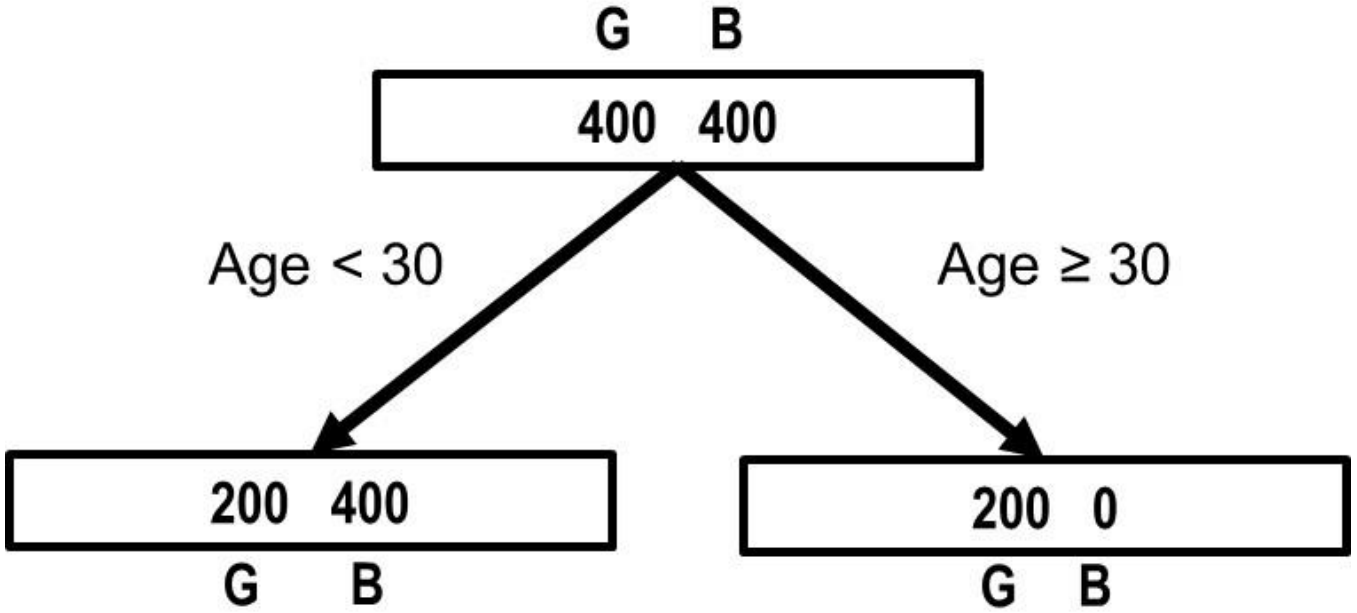


- Entropy: $E(S) = -p_G \log_2(p_G) - p_B \log_2(p_B)$ (C4.5/See5)
- Gini: $Gini(S) = 2 p_G p_B$ (CART)

Decision Trees

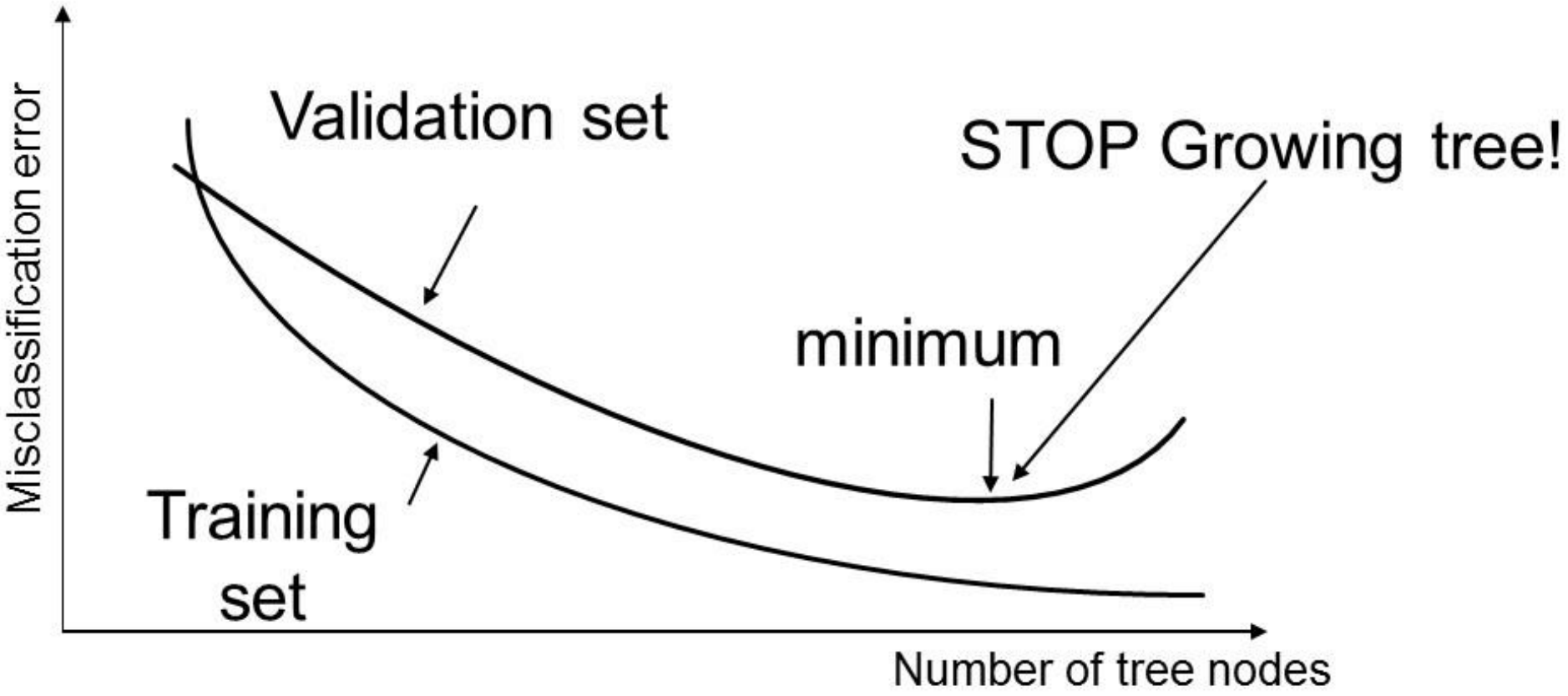


Decision Trees

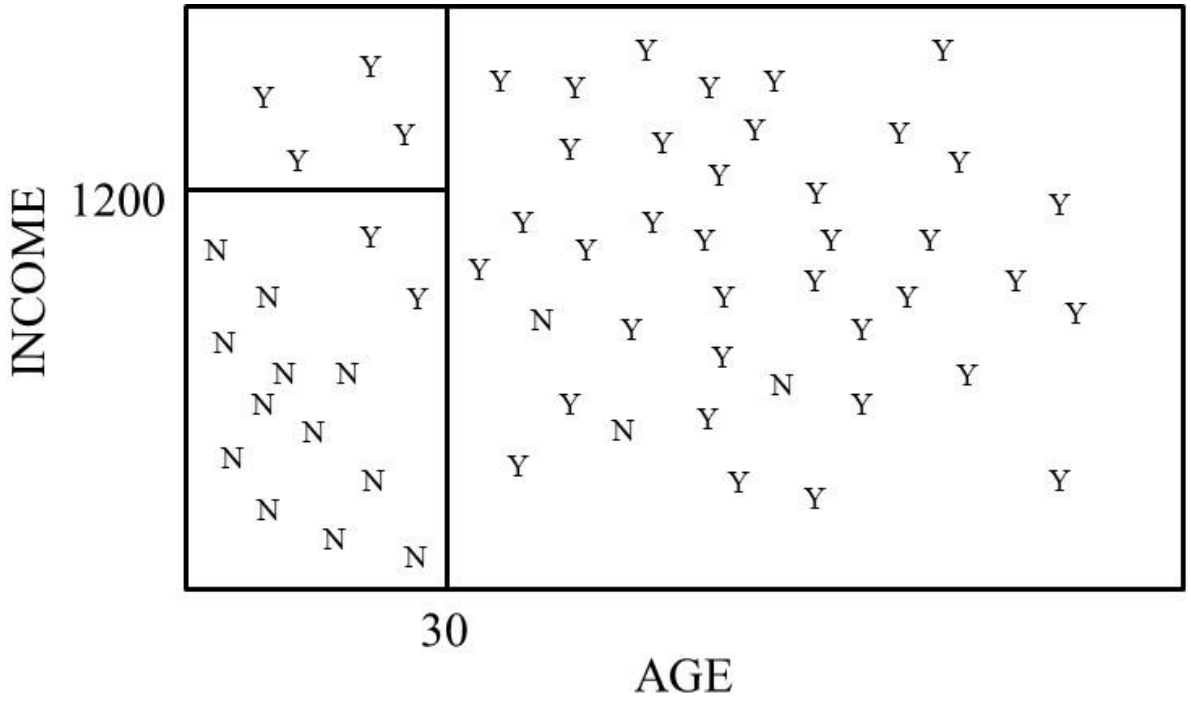


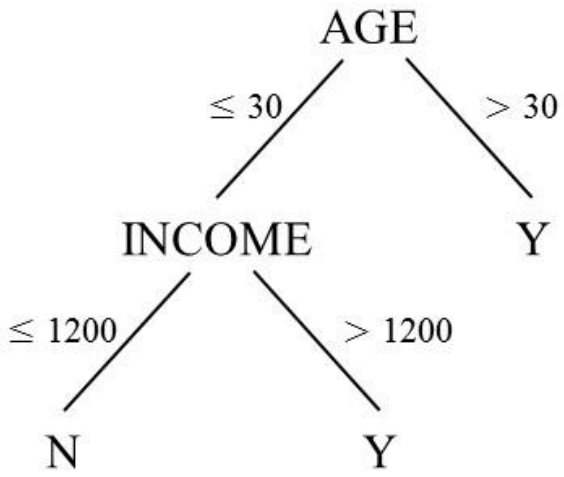
- Entropy top node = -1/2 x $\log_2(1/2)$ – 1/2 x $\log_2(1/2)$ = 1
- Entropy left node = -1/3 x $\log_2(1/3)$ – 2/3 x $\log_2(2/3)$ = 0.91
- Entropy right node = -1 x $\log_2(1)$ – 0 x $\log_2(0)$ = 0
- Gain = 1 – (600/800) x 0.91 – (200/800) x 0 = 0.32

Decision Trees



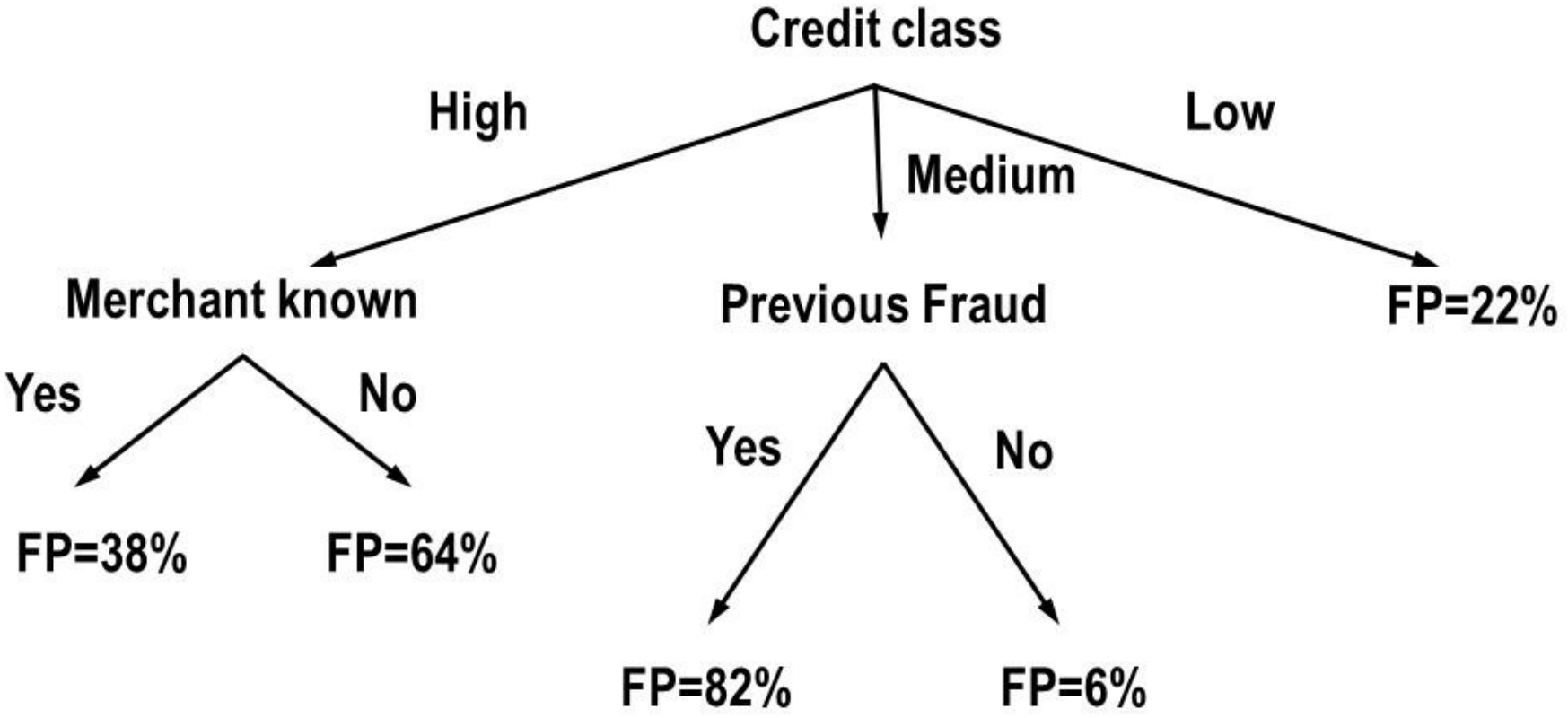
Decision Trees





Decision Trees

- Regression trees



Decision Trees

- Regression trees
  - $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$,
  - $F = \frac{SS_{between}/(B-1)}{SS_{within}/(n-B)} \sim F_{n-B,B-1}$
    - $SS_{between} = \sum_{b=1}^{B} n_b\,(\bar{y}_b - \bar{y})^2$
    - $SS_{within} = \sum_{b=1}^{B}\sum_{i=1}^{n_b}(y_{bi} - \bar{y}_b)^2$

Other predictive analytics techniques

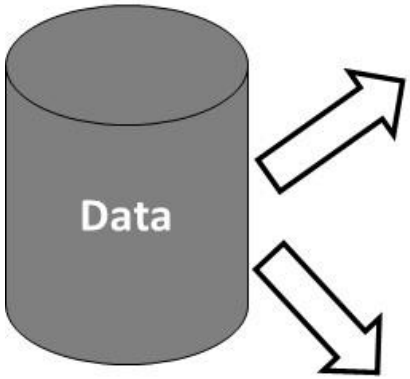
- Ensemble methods
  - Bagging, Boosting, Random Forests
- Neural Networks
- Support Vector Machines
- Deep Learning
- Trade-off between model performance and interpretability!

Evaluating Predictive Models

- Splitting up the data set
- Performance Measures for Classification Models
- Performance Measures for Regression Models
- Other Performance Measures for Predictive Analytical Models

Splitting up the data set

TRAIN/TEST DATA



Train data

Customer	Age	Income	Gender	Good/Bad	...	Target
John	30	1200	M	Bad		0
Sarah	25	800	F	Good		1
Sophie	52	2200	F	Good		1
David	48	2000	M	Bad		0
Peter	34	1800	M	Good		1

Build Model

$$p(Good \mid \text{age}, income, ...) = \frac{1}{1 + e^{-(0.10 + 0.50\, age + 0.0034\, income + ...)}}$$

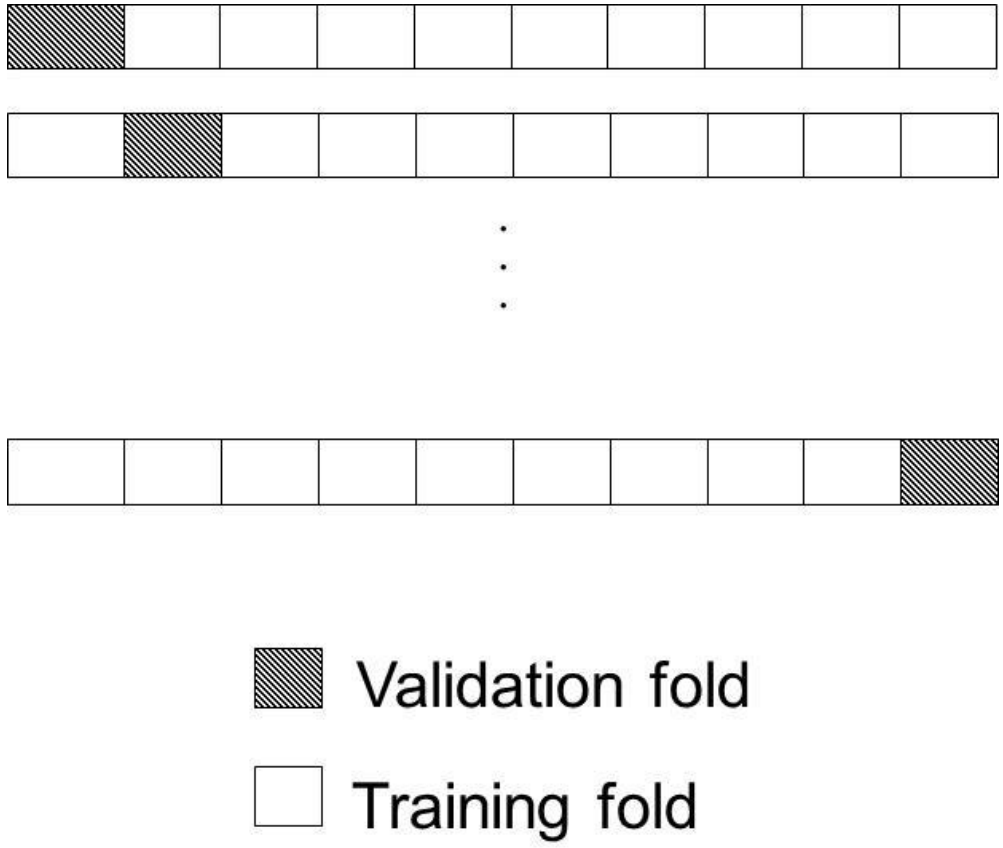
Apply Model

Test data

Customer	Age	Income	Gender	...	Good/Bad	Score
Emma	28	1000	F		Good	0,44
Will	44	1500	M		Bad	0,76
Dan	30	1200	M		Good	0,18
Bob	58	2400	M		Good	0,88

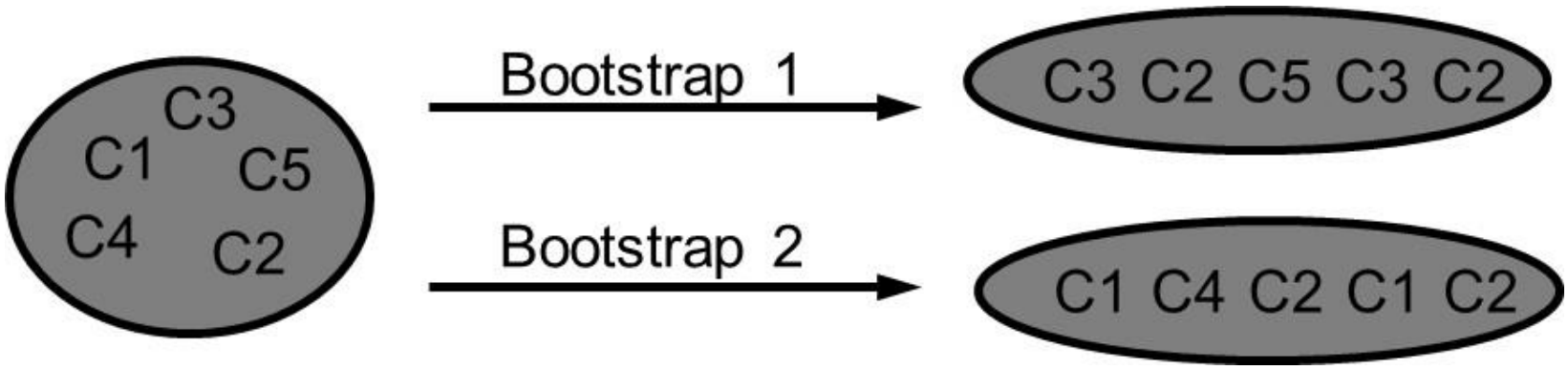
Splitting up the data set

CROSS-VALIDATION



Splitting up the data set

BOOTSTRAPPING



Performance Measures for Classification Models

	Churn	Score
John	Yes	0.72
Sophie	No	0.56
David	Yes	0.44
Emma	No	0.18
Bob	No	0.36

	Churn	Churn Score
John	Yes	0.72
Sophie	No	0.56
David	Yes	0.44
Emma	No	0.18
Bob	No	0.36

Cutoff=0.50

	Churn	Churn Score	Predicted
John	Yes	0.72	Yes
Sophie	No	0.56	Yes
David	Yes	0.44	No
Emma	No	0.18	No
Bob	No	0.36	No

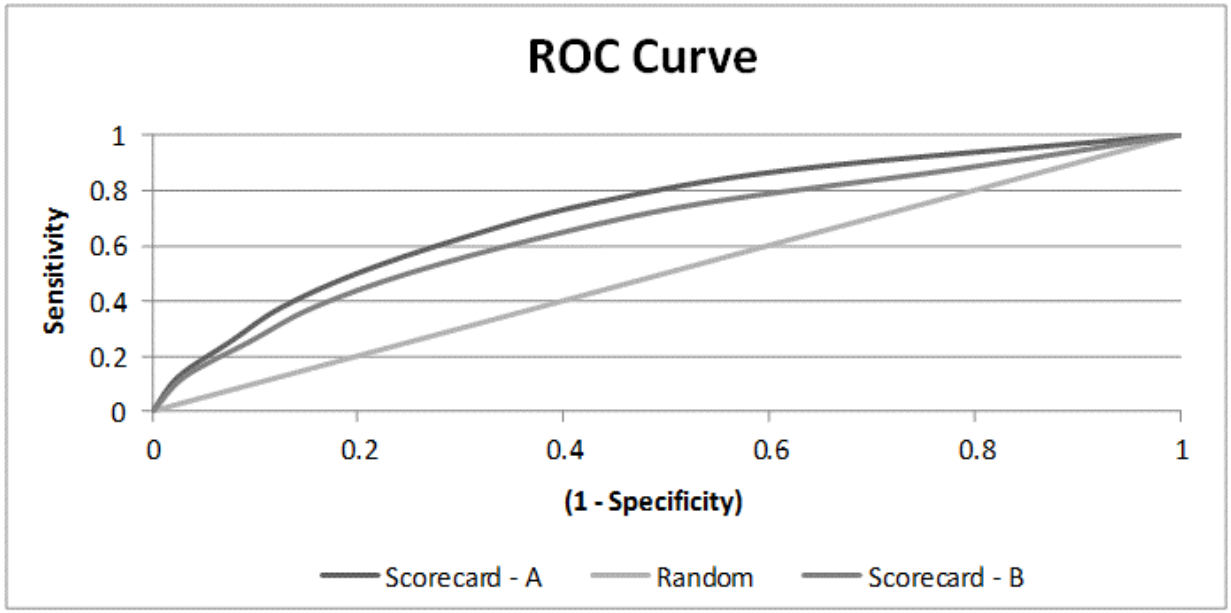
Performance Measures for Classification Models

		Actual status	
		Positive (churn)	Negative (no churn)
Predicted status	Positive (churn)	True Positive (John)	False Positive (Sophie)
	Negative (no churn)	False Negative (David)	True Negative (Emma, Bob)

- Classification accuracy = (TP+TN)/(TP+FP+FN+TN) = 3/5
- Classification error = (FP +FN)/(TP+FP+FN+TN) = 2/5
- Sensitivity = Recall = Hit rate = TP/(TP+FN) = 1/2
- Specificity = TN/(FP+TN) = 2/3
- Precision = TP/(TP+FP) = 1/2
- F-measure = 2 * (Precision * Recall)/(Precision +Recall) = 1/2

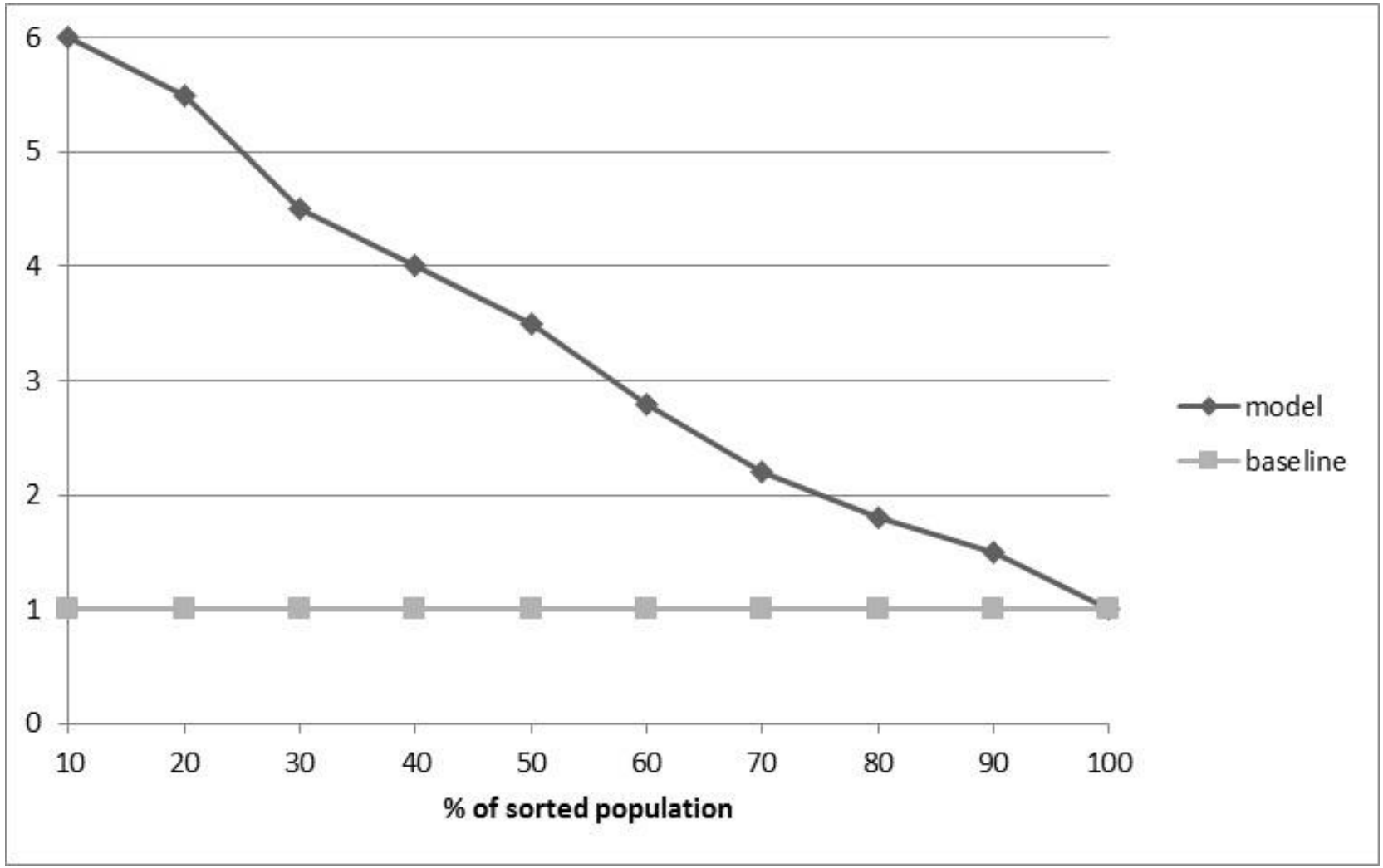
Performance Measures for Classification Models

Cut-off	sensitivity	specificity	1-specificity
0	1	0	1
0.01			
0.02			
….			
0.99			
1	0	1	0

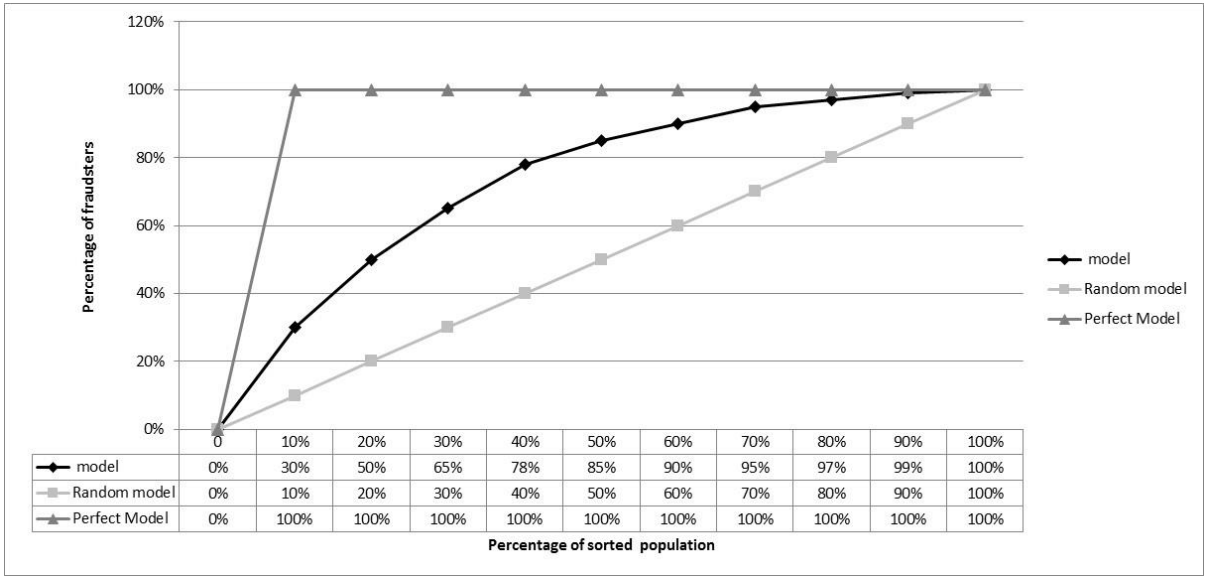


AUC represents probability that randomly chosen churner gets higher score than randomly chosen non-churner!

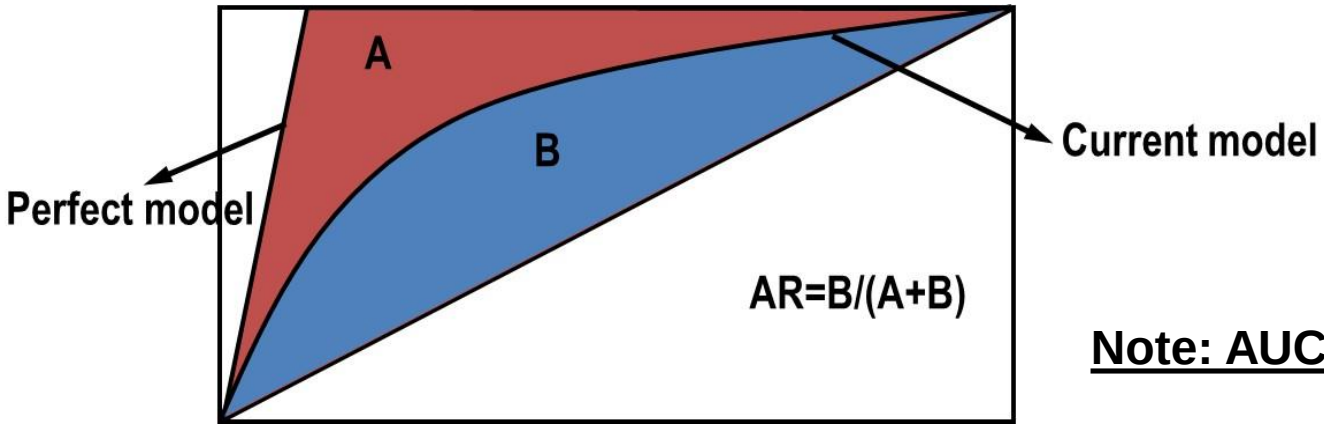
Performance Measures for Classification Models



Performance Measures for Classification Models



	0	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
model	0%	30%	50%	65%	78%	85%	90%	95%	97%	99%	100%
Random model	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Perfect Model	0%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%



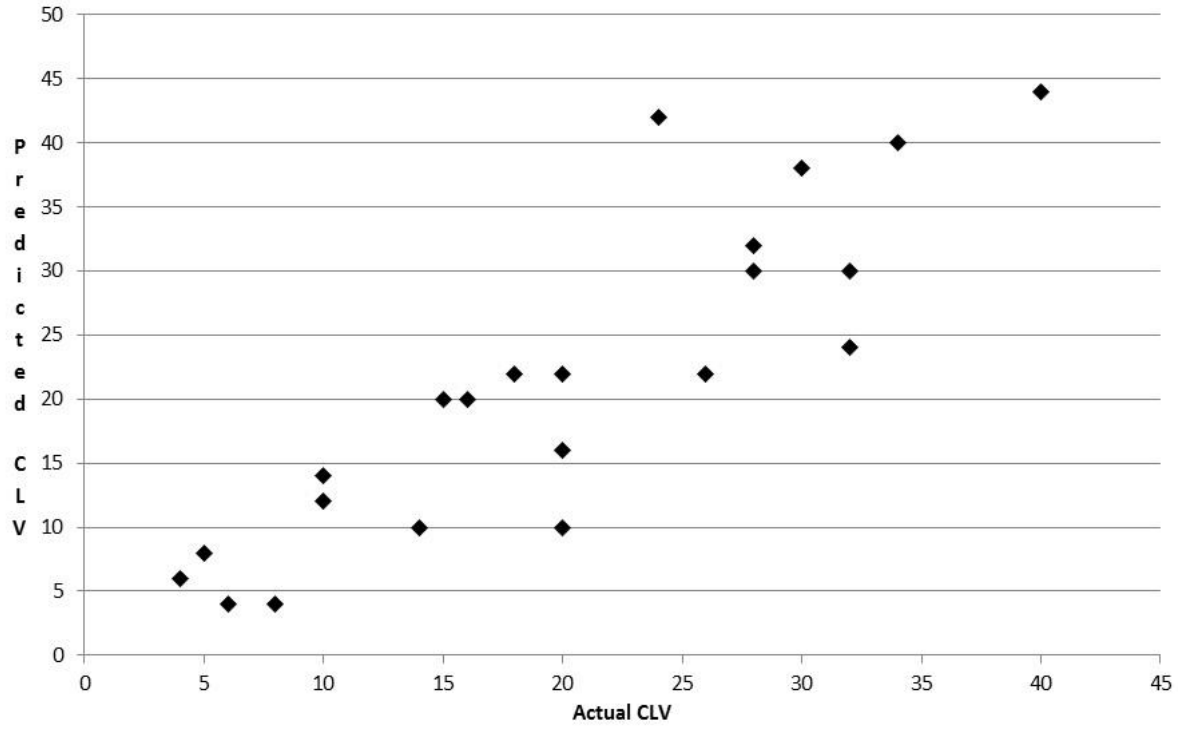
Note: AUC : AR=2*AUC-1

Performance Measures for Classification Models

Application	**Number of variables**	**AUC Ranges**
Credit Scoring	10–15	70%–85%
Churn Prediction (Telco)	6–10	70%–90%
Fraud Detection (Insurance)	10–15	70%–90%

Performance Measures for Regression Models

- $corr(\hat{y}, y) = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$,



Performance Measures for Regression Models

- $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$,
- $R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1 - R^2) = 1 - \frac{n-1}{n-k-1}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$
- $MAD = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

Other Performance Measures for Predictive Analytical Models

- Comprehensibility
- Justifiability
- Operational efficiency

Descriptive Analytics

- Association rules
- Sequence rules
- Clustering

Association Rules

Transaction identifier	Items
1	beer, milk, diapers, baby food
2	coke, beer, diapers
3	cigarettes, diapers, baby food
4	chocolates, diapers, milk, apples
5	tomatoes, water, apples, beer
6	spaghetti, diapers, baby food, beer
7	water, beer, baby food
8	diapers, baby food, spaghetti
9	baby food, beer, diapers, milk
10	apples, wine, baby food

Association rule is implication $X \Rightarrow Y$, whereby $X \subset I, Y \subset I$ and $X \cap Y = \emptyset$

Association Rules

- $support(X \cup Y) = \frac{number\ of\ transactions\ supporting\ (X \cup Y)}{total\ number\ of\ transactions}$
- $confidence(X \rightarrow Y) = p(Y|X) = \frac{support(X \cup Y)}{support(X)}$
- $lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X).support(Y)}$

Association Rules

- Post processing
  - Filter out trivial rules
  - Sensitivity analysis
  - Visualization
  - Measure economic impact

Sequence Rules

Session ID	Page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	1
5	A	1

Session 1: A, B, C

Session 2: B, C

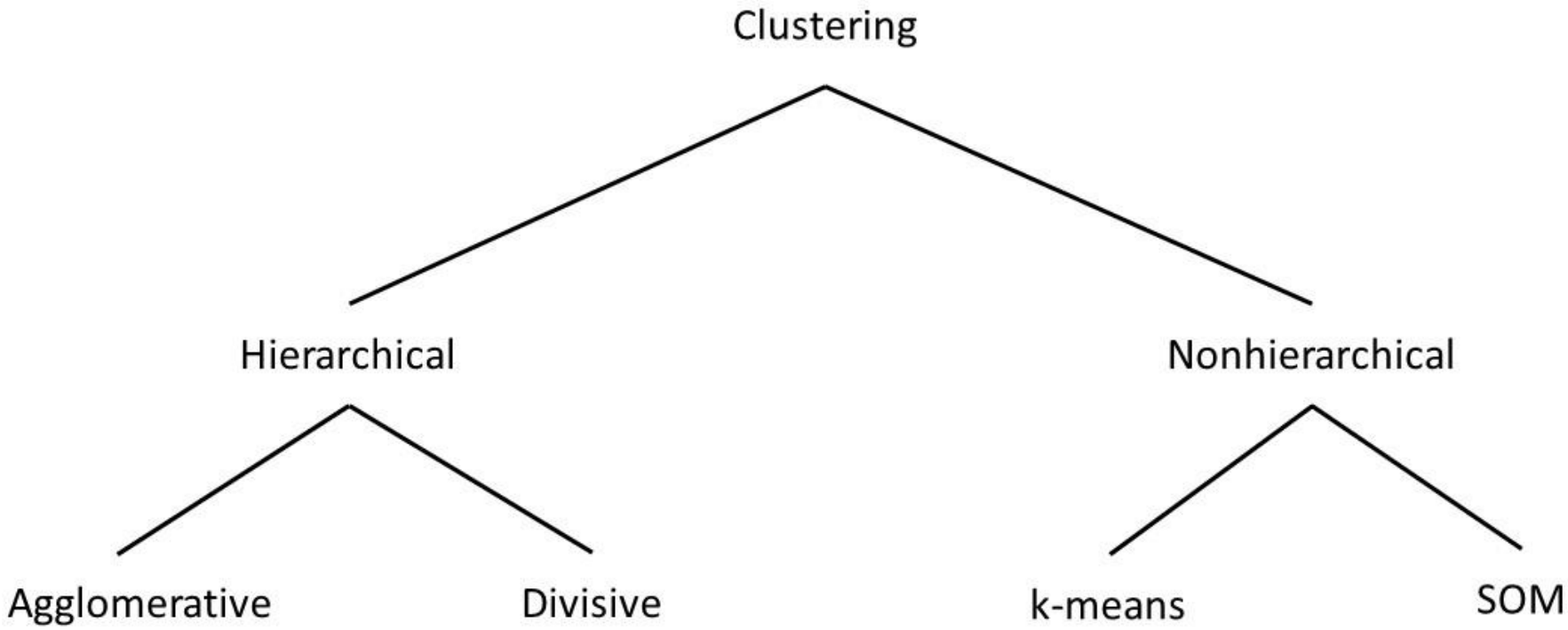
Session 3: A, C, D

Session 4: A, B, D

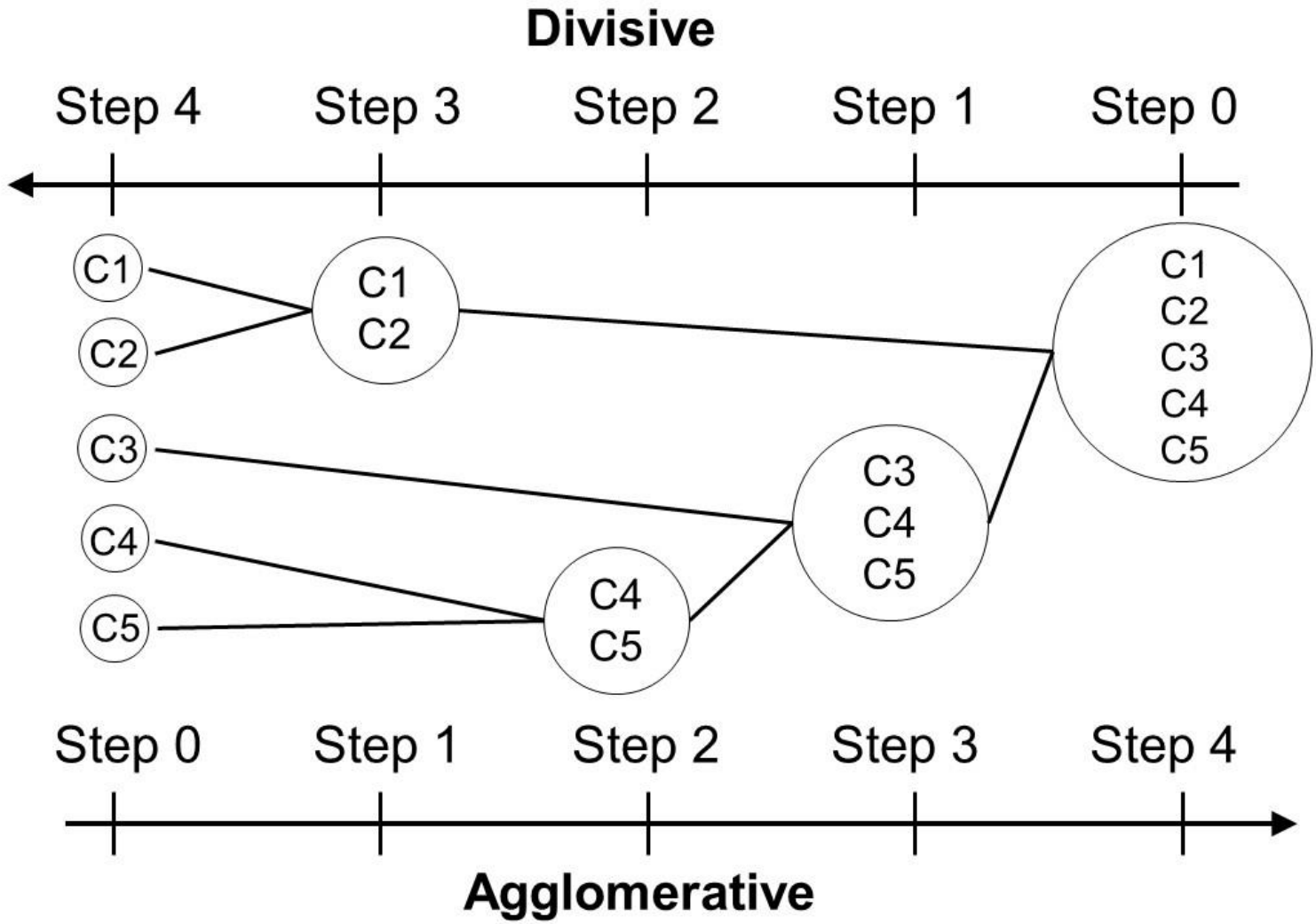
Session 5: D, C, A

Calculate confidence and support as with association rules!

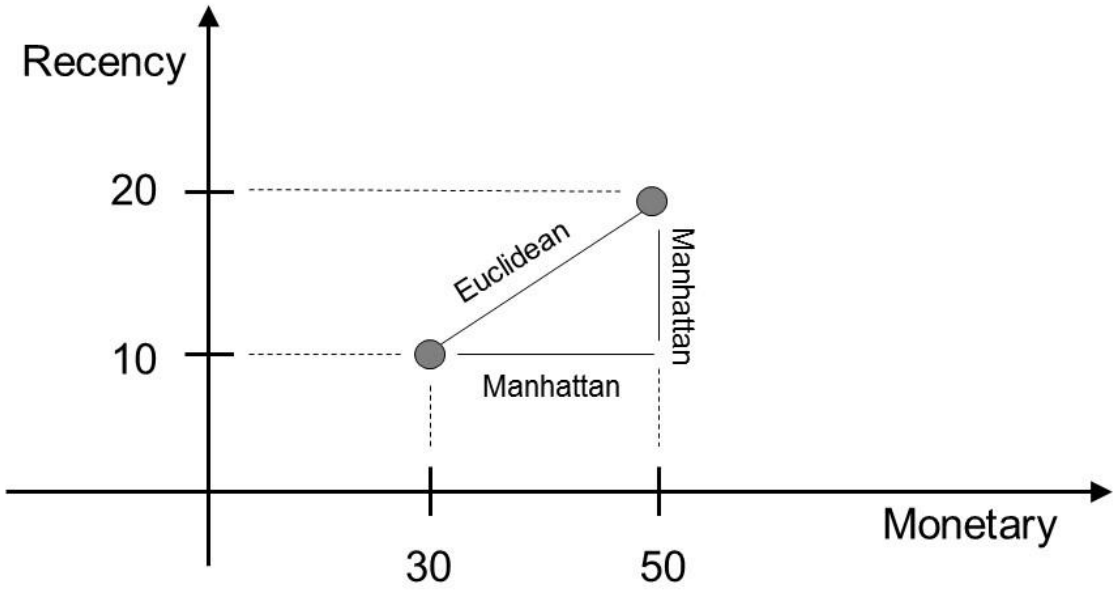
Clustering


Clustering
Hierarchical
Nonhierarchical
Agglomerative
Divisive
k-means
SOM

Clustering


Divisive
Step 4
Step 3
Step 2
Step 1
Step 0
C1
C2
C3
C4
C5
C1
C2
C4
C5
C3
C4
C5
C1
C2
C3
C4
C5
Step 0
Step 1
Step 2
Step 3
Step 4
Agglomerative

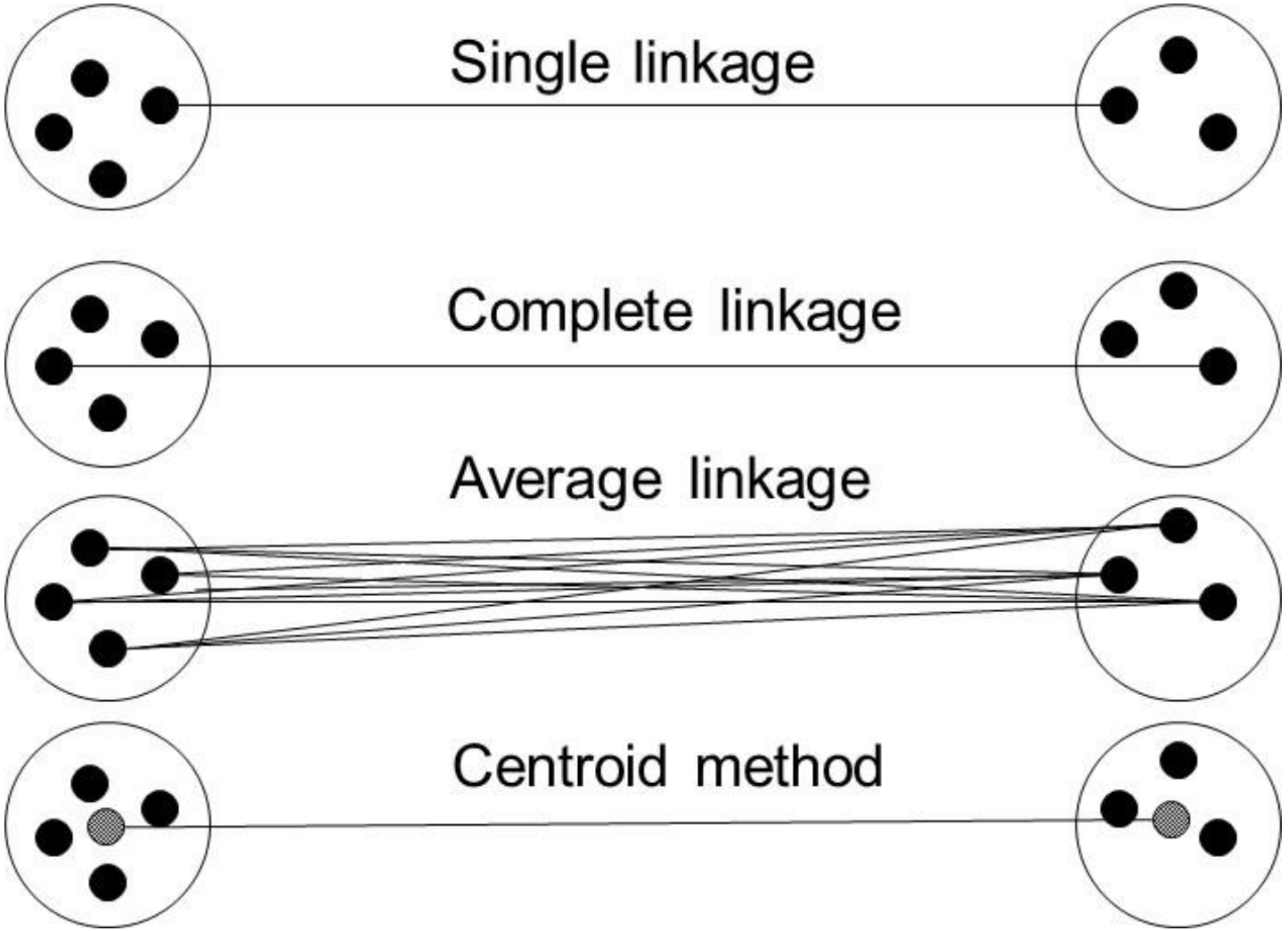
Clustering



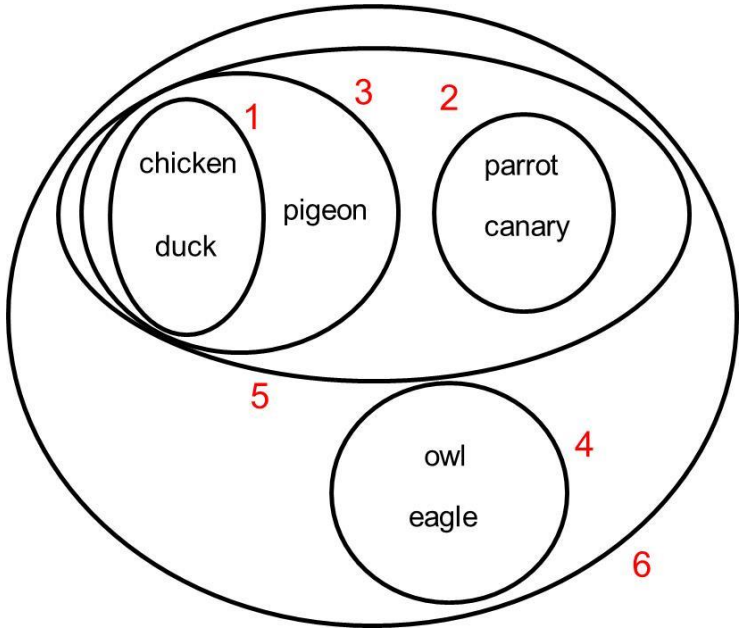
Euclidean: $\sqrt{(50-30)^2+(20-10)^2}$=22

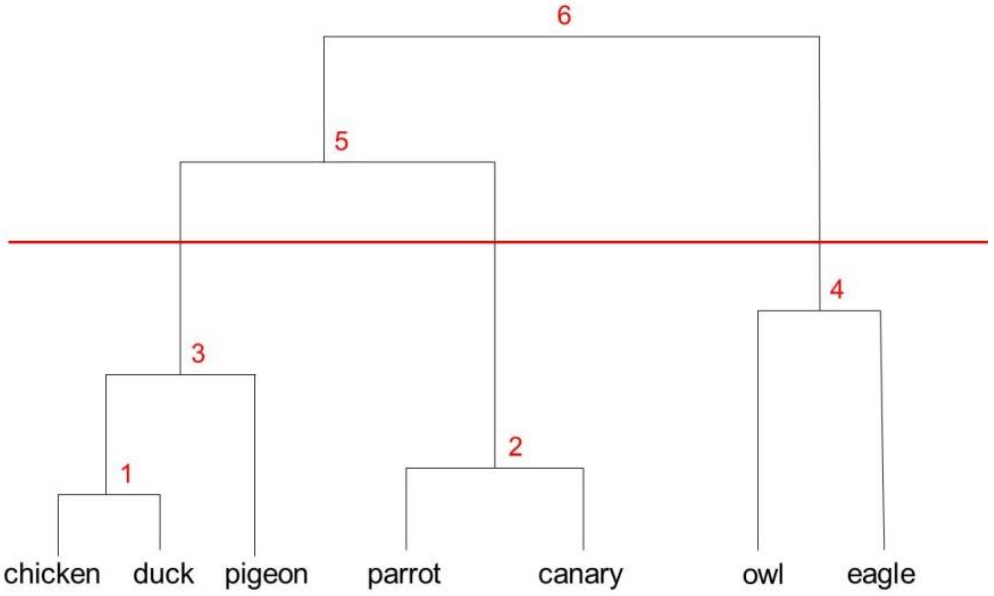
Manhattan: $|50-30| + |20-10| = 30$

Clustering


Single linkage
Complete linkage
Average linkage
Centroid method

Clustering


chicken
duck
pigeon
parrot
canary
owl
eagle
1
2
3
4
5
6


6
5
4
3
2
1
chicken
duck
pigeon
parrot
canary
owl
eagle

Clustering

- K-means clustering
  - Select K observations as initial cluster centroids (seeds)
  - Assign each observation to cluster that has closest centroid (for example, in Euclidean sense)
  - When all observations have been assigned, recalculate positions of *K* centroids
  - Repeat until cluster centroids no longer change

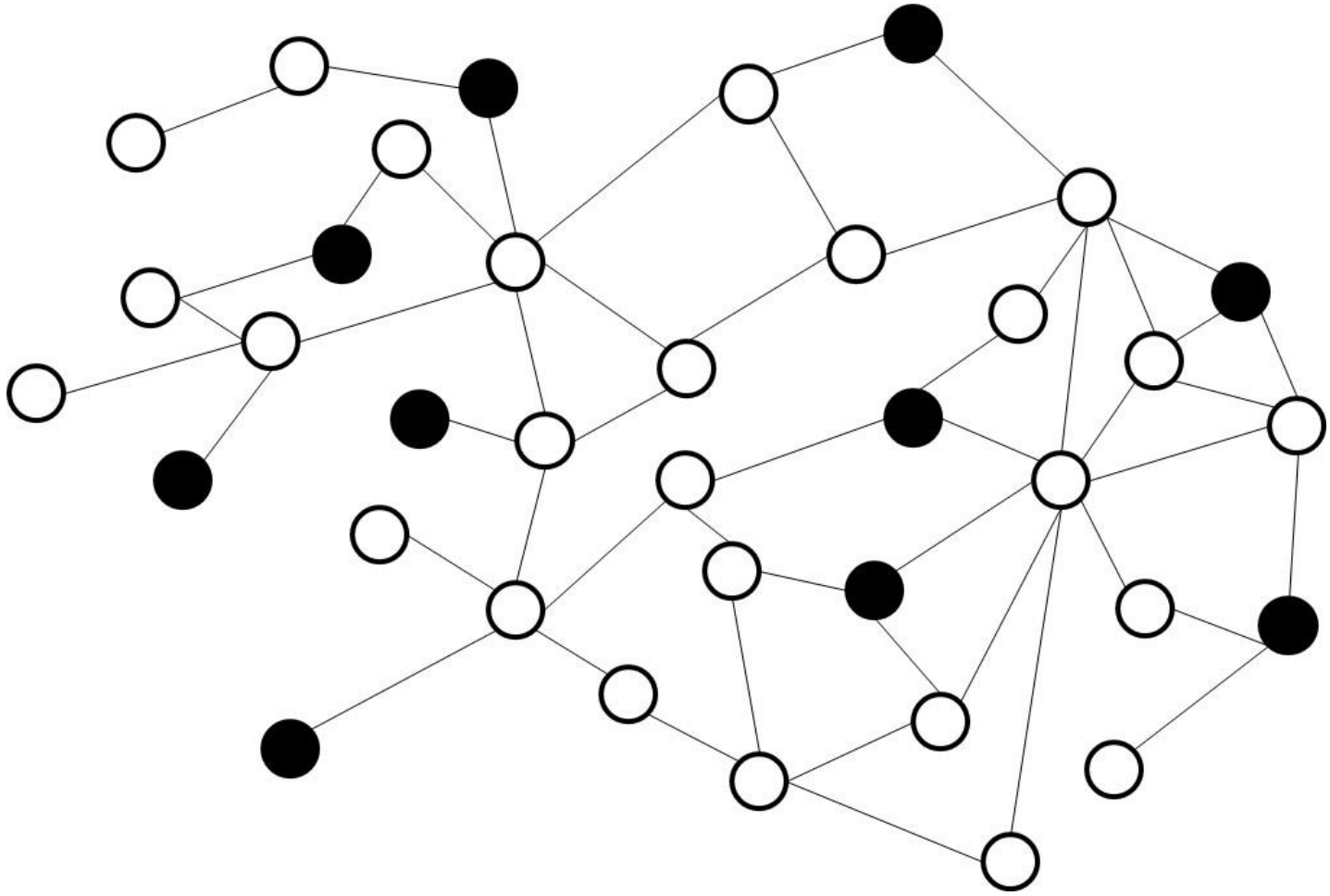
Social Network Analytics

- Social Network Definitions
- Social Network Metrics
- Social Network Learning

Social Network Definitions

- Social network consists of both nodes and edges
- Node could be defined as a customer (private/professional), household/family, patient, doctor, paper, author, terrorist, webpage, …
- Edge can be defined as a ‘friends’ relationship, a call, transmission of a disease, a ‘follows’ relationship, a reference, etc.

Social Network Definitions



Social Network Definitions

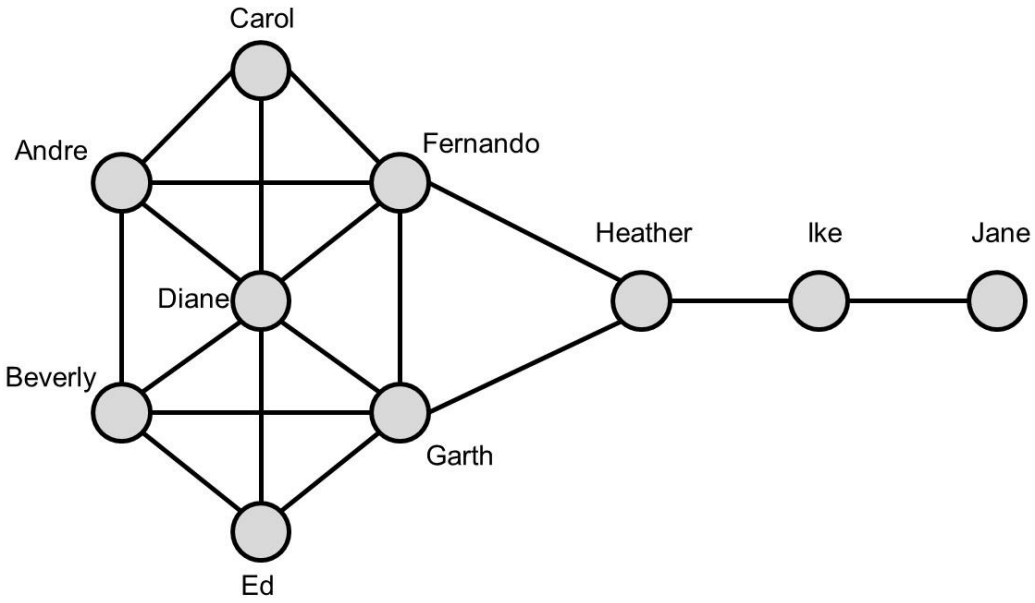
Sociogram

	C1	C2	C3	C4
C1	-	1	1	0
C2	1	-	0	1
C3	1	0	-	0
C4	0	1	0	-

Social Network Metrics

Geodesic	Shortest path between two nodes in the network.	
Degree	Number of connections of a node (in- versus out-degree if the connections are directed).	
Closeness	The average distance of a node to all other nodes in the network (reciprocal of farness).	$[\frac{\sum_{j=1}^{g} d(N_i, N_j)}{g}]^{-1}$
Betweenness	Counts the number of times a node or edge lies on the shortest path between any two nodes in the network.	$\sum_{j<k} \frac{g_{jk}(N_i)}{g_{jk}}$
Graph theoretic center	The node with the smallest maximum distance to all other nodes in the network.	

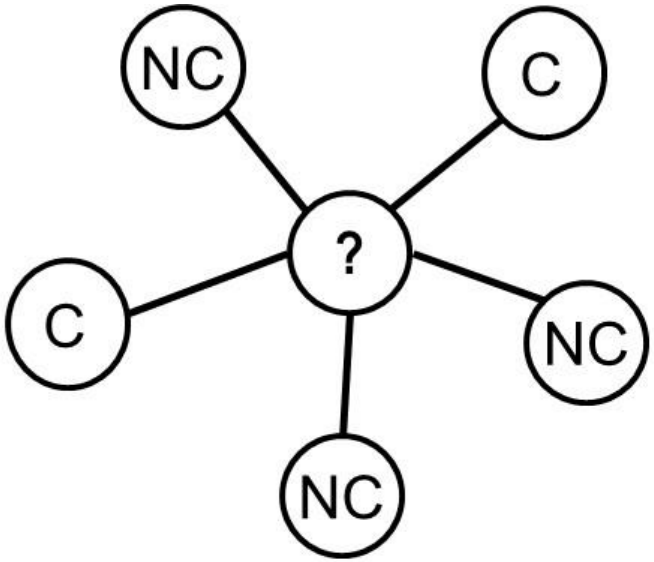
Social Network Metrics



Degree		Closeness		Betweenness	
6	Diane	0.64	Fernando	14	Heather
5	Fernando	0.64	Garth	8.33	Fernando
5	Garth	0.6	Diane	8.33	Garth
4	Andre	0.6	Heather	8	Ike
4	Beverly	0.53	Andre	3.67	Diane
3	Carol	0.53	Beverly	0.83	Andre
3	Ed	0.5	Carol	0.83	Beverly
3	Heather	0.5	Ed	0	Carol
2	Ike	0.43	Ike	0	Ed
1	Jane	0.31	Jane	0	Jane

Social Network Learning

Featurization



Customer	Age	Income	…	Mode link	Frequency no churn	Frequency churn	Binary no churn	Binary churn
Bart	38	2400		NC	3	2	1	1

Social Network Learning

Customer	Age	Recency	Number of contacts	Contacts with churners	Churn
John	35	5	18	3	Yes
Sophie	18	10	7	1	No
Victor	38	28	11	1	No
Laura	44	12	9	0	Yes

Customer	**Age**	**Avg duration**	**Avg revenue**	**Promo-tions**	**Avg age friends**	**Avg duration friends**	**Avg revenue friends**	**Promo-tions friends**	**Churn**
John	35	50	123	X	20	55	250	X	Yes
Sophie	18	65	55	Y	18	44	66	Y	No
Victor	38	12	85	None	50	33	50	X, Y	No
Laura	44	66	230	X	65	55	189	X	No

Post Processing of Analytical Models

- Interpretation and validation
- Sensitivity analysis
- Model deployment
- Backtesting

Critical Success Factors for Analytical Models

- Business relevance
- Statistical performance and validity
- Interpretability
- Justifiability
- Operational efficiency
- Economical cost
- Regulatory compliance

Economic Perspective on Analytics

- Total Cost of Ownership (TCO)
- Return on Investment (ROI)
- In- versus Outsourcing
- On-Premise versus Cloud Solutions
- Open Source versus Commercial Software

Total Cost of Ownership (TCO)

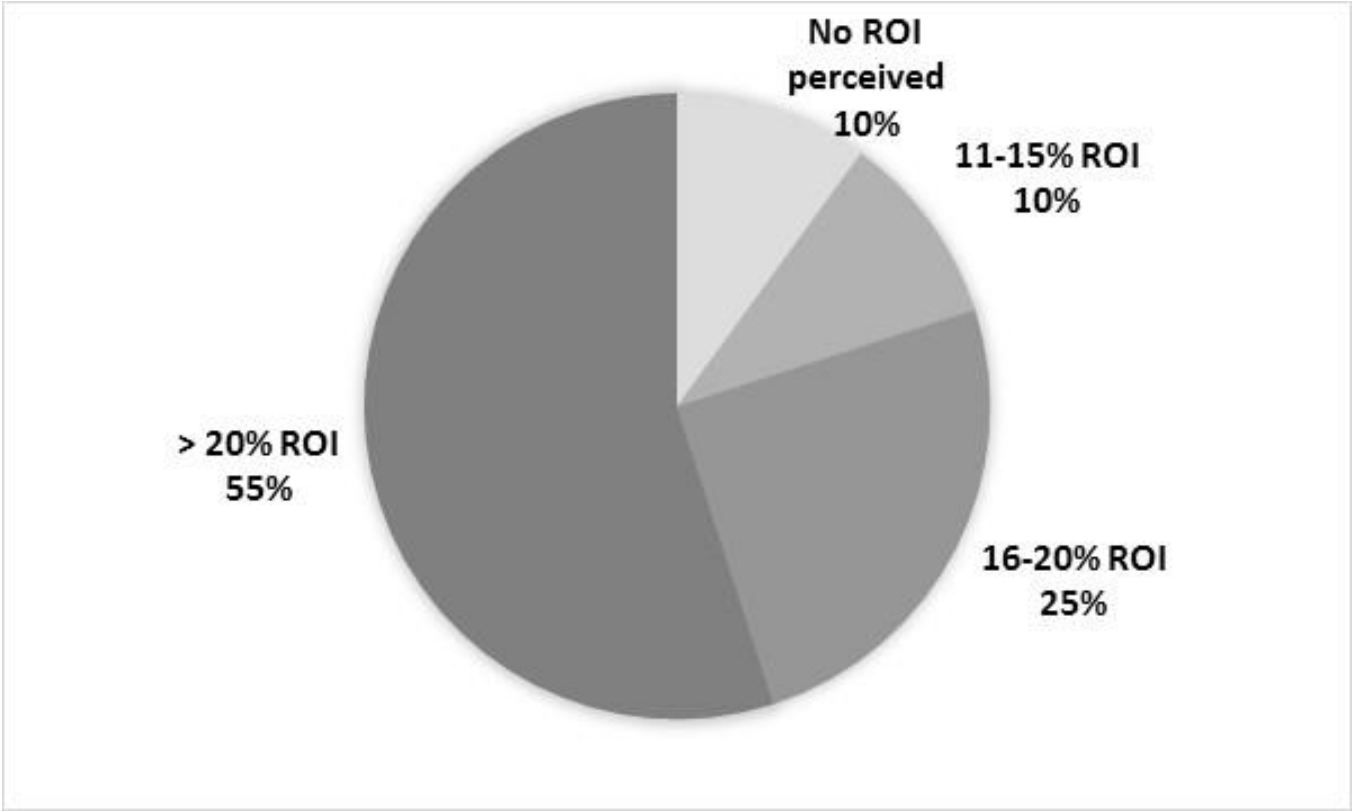
Acquisition costs	Ownership and operation costs	Post ownership costs
• Software costs including initial purchase, upgrade, intellectual property and licensing fees • Hardware costs including initial purchase price and maintenance • Network and security costs • Data costs including costs for purchasing external data • Model developer costs such as salaries and training	• Model migration and change management costs • Model setup costs • Model execution costs • Model monitoring costs • Support costs (troubleshooting, helpdesk, …) • Insurance costs • Model staffing costs such as salaries and training • Model upgrade costs • Model downtime costs	• De-installation and disposal costs • Replacement costs • Archiving costs

Return on Investment (ROI)

- Ratio of net benefits or net profits over investment of resources that generated this return
- Example benefits:
  - increase of sales
  - lower fraud losses
  - fewer credit defaults
  - identification of new customer needs/opportunities
  - automation of human decision making
  - development of new business models

Return on Investment (ROI)

- Results from PredictiveanalyticsToday.com poll from February 2015 to March 2015



In- versus Outsourcing

- Activities for outsourcing: data collection, cleaning and preprocessing, set up of platforms, training, model development, visualization, evaluation, monitoring and maintenance
- Risks
  - analytics concerns a company's frontend strategy
  - exchange of confidential information
  - continuity of the partnership
  - cultural mismatch
  - shortage of data scientists

On-Premise versus Cloud Solutions

- On-Premise analytics
  - Keep data in-house (full control)
  - Security risk
  - Expensive up- or downsizing
- Cloud solutions
  - Better security management
  - Scalability and economies of scale
  - Easy maintenance/upgrades
  - Improved collaboration across business departments
  - Risk of vendor lock in

Open Source versus Commercial Software

- Open source
  - Free
  - Less quality assurance
  - Full access to source code
- Commercial
  - Well-engineered business-focused solutions (end-to-end)
  - Extensive help facilities
  - Business continuity
  - Pre-packaged, black box routines

Improving the ROI of Analytics

- New sources of data
- Data quality
- Management support
- Organizational aspects
- Cross-Fertilization

New sources of data

- Network data (explicit versus implicit)
- Publicly available data
- Macro-economic data
- Textual data
- Audio, images, videos, fingerprint, location (GPS), geospatial, RFID data, …

Data quality

- GIGO: Garbage In, Garbage Out
- Causes of data quality issues often deeply rooted within core organizational processes and culture
- Data preprocessing activities are corrective measures for dealing with data quality issues
- Transparent and well-defined collaboration between data stewards and data owners key to improve data quality in sustainable manner

Management Support

- Either existing C-level executive takes responsibility or new CXO function is defined (e.g., Chief Analytics Officer or Chief Data Officer)
- Aim for top-down, data driven culture to catalyze trickledown effect
- Board of directors and senior management should be actively involved in analytical model building, implementation and monitoring processes

Organizational Aspects

- Well-articulated data governance program is a good starting point
- Approaches:
  - Centralized: central department of data scientists handles all analytics requests
  - Decentralized: all data scientists directly assigned to business units
  - Mixed: centrally coordinated center of analytical excellence with analytics organized at business unit level

Cross-Fertilization

- Most advanced analytical techniques in risk management
- Marketing analytics less mature
- HR analytics starting to kick-off
- Tremendous potential for cross-fertilization of model development and monitoring experiences across disciplines

Privacy and Security

- Overall considerations
- RACI Matrix
- Accessing Internal Data
- Privacy Regulation

Overall considerations

- Data security
  - set of policies and techniques to ensure confidentiality, availability and integrity of data
- Data privacy
  - parties accessing and using data can do so only in ways that comply with agreed upon purposes of data use in their role
- Security can be considered as necessary instruments to guarantee data privacy

Overall considerations

- Data security pertains to following concerns
  - Guaranteeing data integrity
  - Guaranteeing data availability
  - Authentication and access control
  - Guaranteeing confidentiality
  - Auditing
  - Mitigating vulnerabilities

RACI matrix

Accessing Internal Data

- Anonymization
- SQL views
- Label Based Access Control (LBAC)

Anonymization

- Techniques used:
  - Aggregation
  - Discretization
  - Value distortion
  - Generalization

Anonymization

Company's demographics

VAT	Name	Address	Size	Creation date	Revenue	Sector
532.581.34	Mony Bank	Main Street 1943, Brussels	592	09/05/1989	€ 9,900,000	banking
532.582.26	Villa Bella	Av. Elisa 66, Liege	6	12/08/1990	€ 25,000	cleaning
532.582.49	The Green Lawn	Lawnstreet 1, Ghent	63	24/02/2004	€ 185,000	agriculture
532.585.71	Salad Palace	Main Street 1472, Brussels	18	25/02/2007	€ 235,000	catering
532.586.52	Bart&Co.	Main Street 239, Brussels	37	04/03/2009	€ 1,700,000	transport
532.586.55	Elisa's Bar	Shortstreet 5, Antwerp	12	07/12/2011	€ 5,000	catering
532.590.00	Transport John	Av. Lovanias 31, Antwerp	104	18/12/2013	€ 34,000	transport
...	...	...	...	...	...	...

natural key

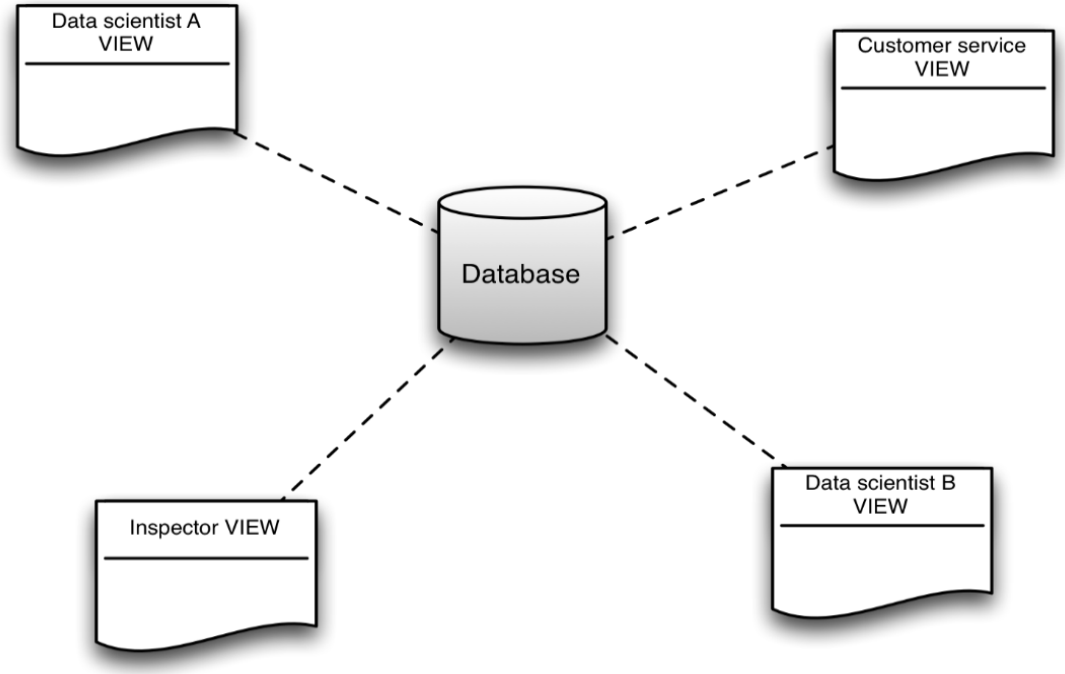
Personnel records

VAT	Name	Income	Recruitment	Resignation
532.586.52	Gerry Hill	€ 1,500	14/09/2012	-
532.586.52	Niel Tenson	€ 1,500	07/12/2009	-
532.586.52	Daisy Astalos	€ 1,800	26/03/2009	22/12/2009
532.586.52	William Wheately	€ 2,000	26/04/2014	-
532.586.52	Tom Book	€ 1,600	03/05/2010	14/01/2011
532.586.52	John Angeles	€ 1,750	17/05/2009	04/02/2015
...	...	...	...	...

Anonymised view

ID	Province	Size	Maturity	Revenue	Sector	Empl. Q1	Empl. Q2	Empl. Q3	Empl. Q4	Avg. wage
19649524	P7	3	A	€ 200,000	agriculture	2	4	0	0	€ 1,550
27499423	P2	4	Y	€ 30,000	transport	-5	-5	-3	-5	€ 1,650
31865139	P1	2	A	€ 2,000,000	transport	5	5	5	-5	€ 1,600
39174842	P1	2	A	€ 250,000	catering	-1	2	0	2	€ 1,500
59135796	P5	1	M	€ 30,000	cleaning	0	0	0	0	€ 1,400
73591064	P1	5	M	€ 10,000,000	banking	10	10	5	5	€ 1,800
91245975	P2	2	Y	€ 10,000	catering	0	-2	0	1	€ 1,350
...	...	...	...	...	...	...	...	...	...	...

SQL Views




```
CREATE VIEW FRAUD_INPUT
AS SELECT C.ANON_VAT, C.PROVINCE, C.ANON_SIZE,
C.ANON_REVENUE, C.SECTOR, C.ANON_AGE, AVG(P.WAGE), COUNT(*)
FROM COMPANIES C, PERSONNEL P
WHERE C.ANON_VAT = P.ANON_VAT
GROUP BY C.ANON_VAT;
```

Label-Based Access Control (LBAC)

- Control mechanism to protect data against unauthorized access

```
CREATE SECURITY LABEL COMPONENT my_sec_label_comp
ARRAY [CONFIDENTIAL, UNCLASSIFIED]

CREATE SECURITY POLICY my_sec_policy
COMPONENTS my_sec_label_comp
WITH DB2LBACRULES
```

Label-Based Access Control (LBAC)

```
CREATE SECURITY LABEL my_sec_policy.confidential
COMPONENT my_sec_label_comp CONFIDENTIAL

CREATE SECURITY LABEL my_sec_policy.unclassified
COMPONENT my_sec_label_comp UNCLASSIFIED

GRANT SECURITY LABEL my_sec_policy.unclassified TO USER
BartBaesens FOR ALL ACCESS
```

Label-Based Access Control (LBAC)

```
GRANT SECURITY LABEL my_sec_policy.unclassified TO USER
SeppevandenBroucke FOR READ ACCESS

GRANT SECURITY LABEL my_sec_policy.confidential TO USER
WilfriedLemahieu FOR ALL ACCESS

CREATE TABLE EMPLOYEE
       (SSN CHAR(6) NOT NULL PRIMARY KEY,
       NAME VARCHAR(40) NOT NULL,
       SALARY INT SECURED WITH confidential,
       …
SECURITY POLICY my_sec_policy)
```

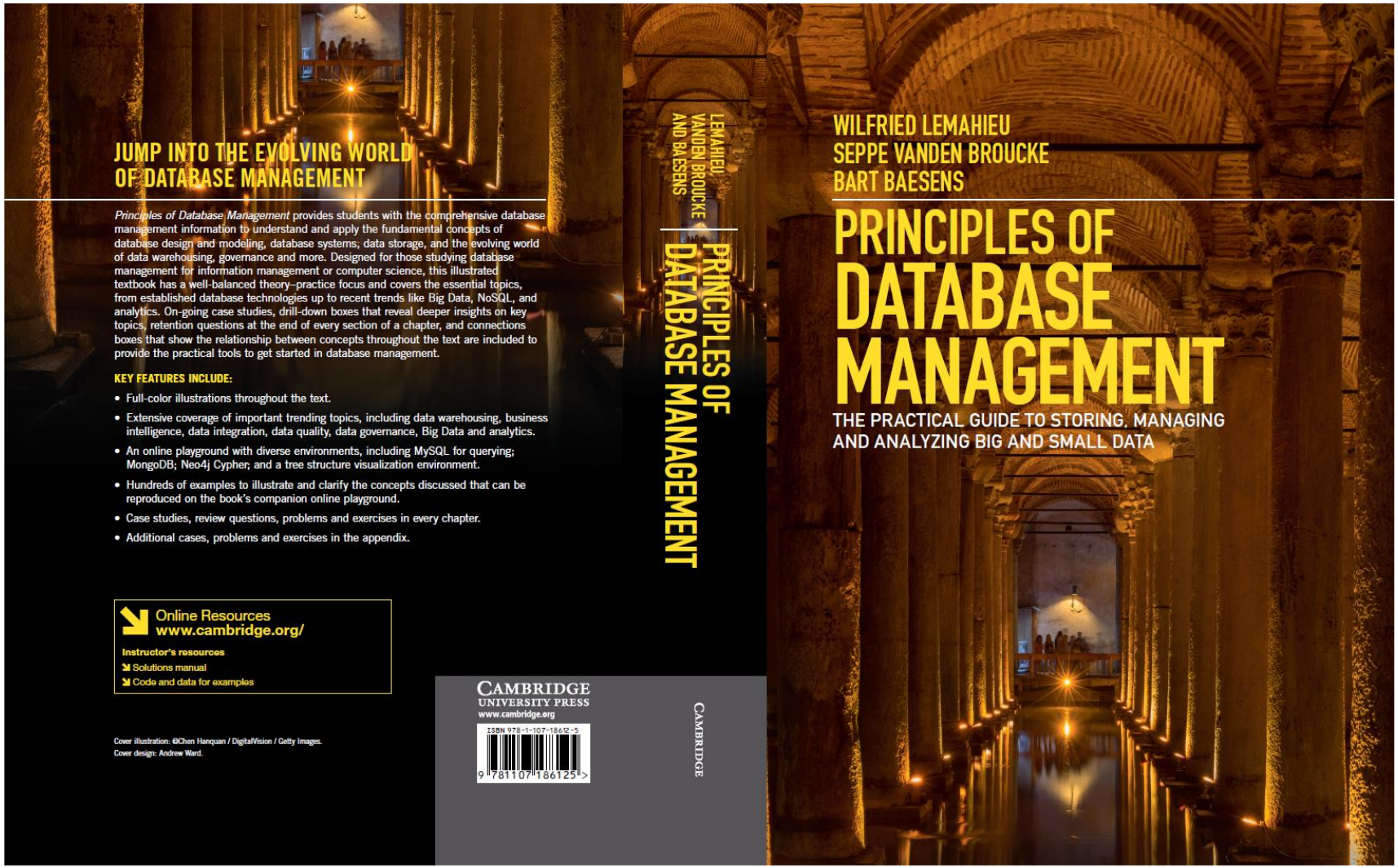
Privacy Regulation

- EU: GDPR
  - right to be informed about how your personal data will be used, right to access and rectify your personal data, right to erase your personal data and right for human intervention in automated decision models, such as analytical prediction models
- US: not highly-regulated (yet)
  - Privacy Act of 1974, Health Insurance Portability and Accountability Act of 1996, Electronic Communications Privacy Act (ECPA) of 1986
- EU-US Privacy Shield

Conclusions

- The Analytics Process Model
- Example Analytics Applications
- Data Scientist Job Profile
- Data Preprocessing
- Types of Analytics
- Post Processing of Analytical Models
- Critical Success Factors for Analytical Models
- Economic Perspective On Analytics
- Improving the ROI of Analytics
- Privacy and Security

More information?



www.pdbmbook.com