MACHINE LEARNING: why should you care?

Filip Wójcik

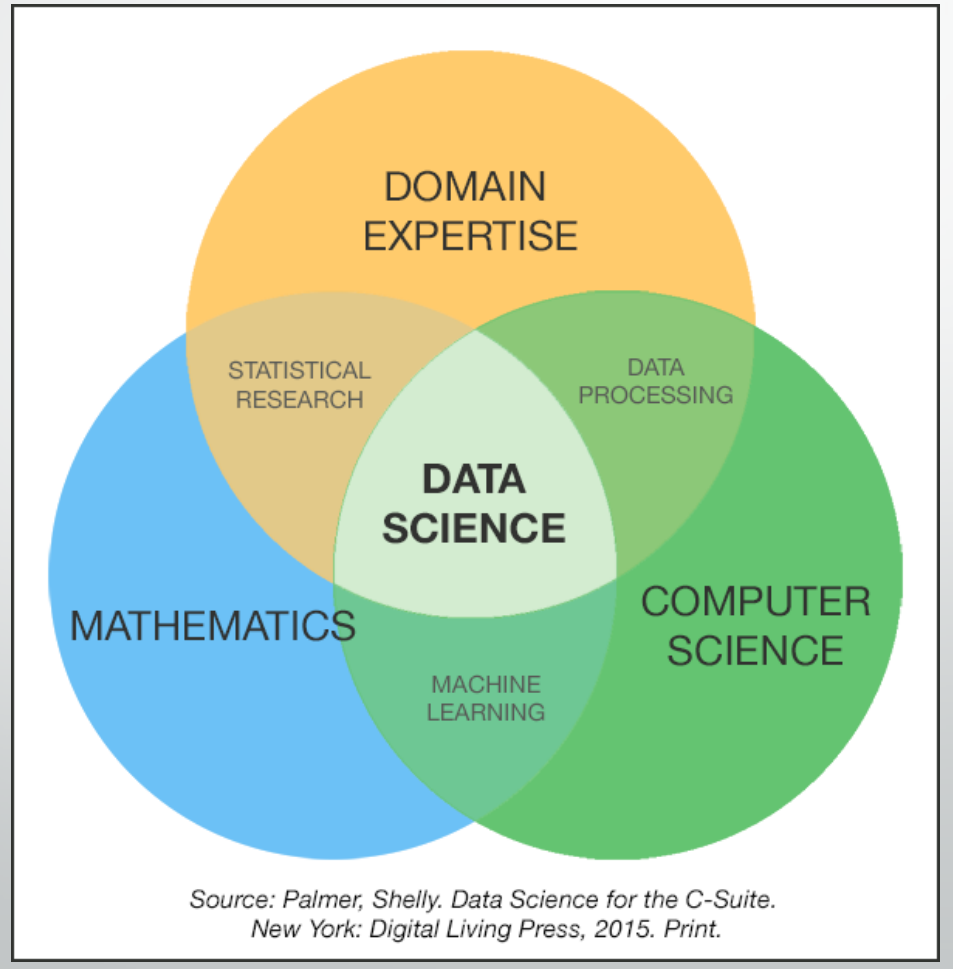
Credit Suisse

Fx Sales Analytics VICE PRESIDENT

University lecturer

filip.wojcik@outlook.com

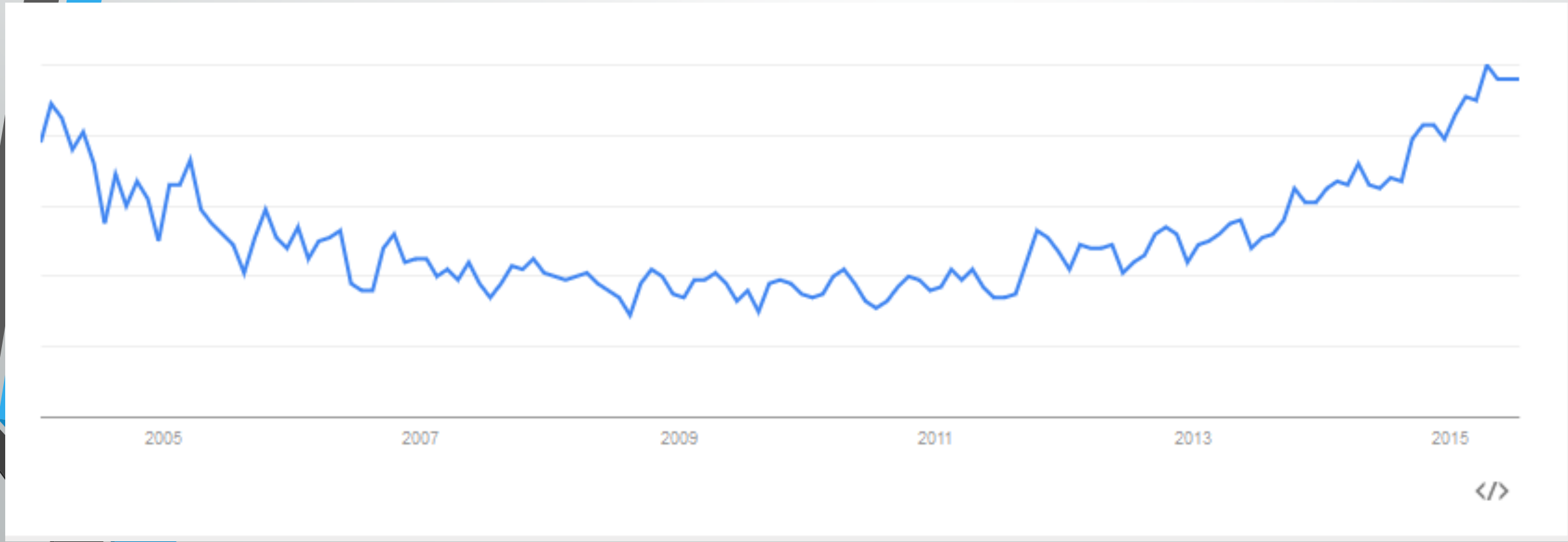
WHAT IS MACHINE LEARNING?



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

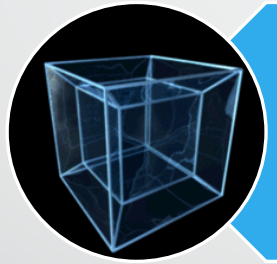
WHAT IS MACHINE LEARNING?

Google trends analysis for phrase „Machine learning”



WHAT IS MACHINE LEARNING?

More and more data produced by corporations



Need to processs more complex data

Automation of data analysis processes

WHAT IS MACHINE LEARNING?

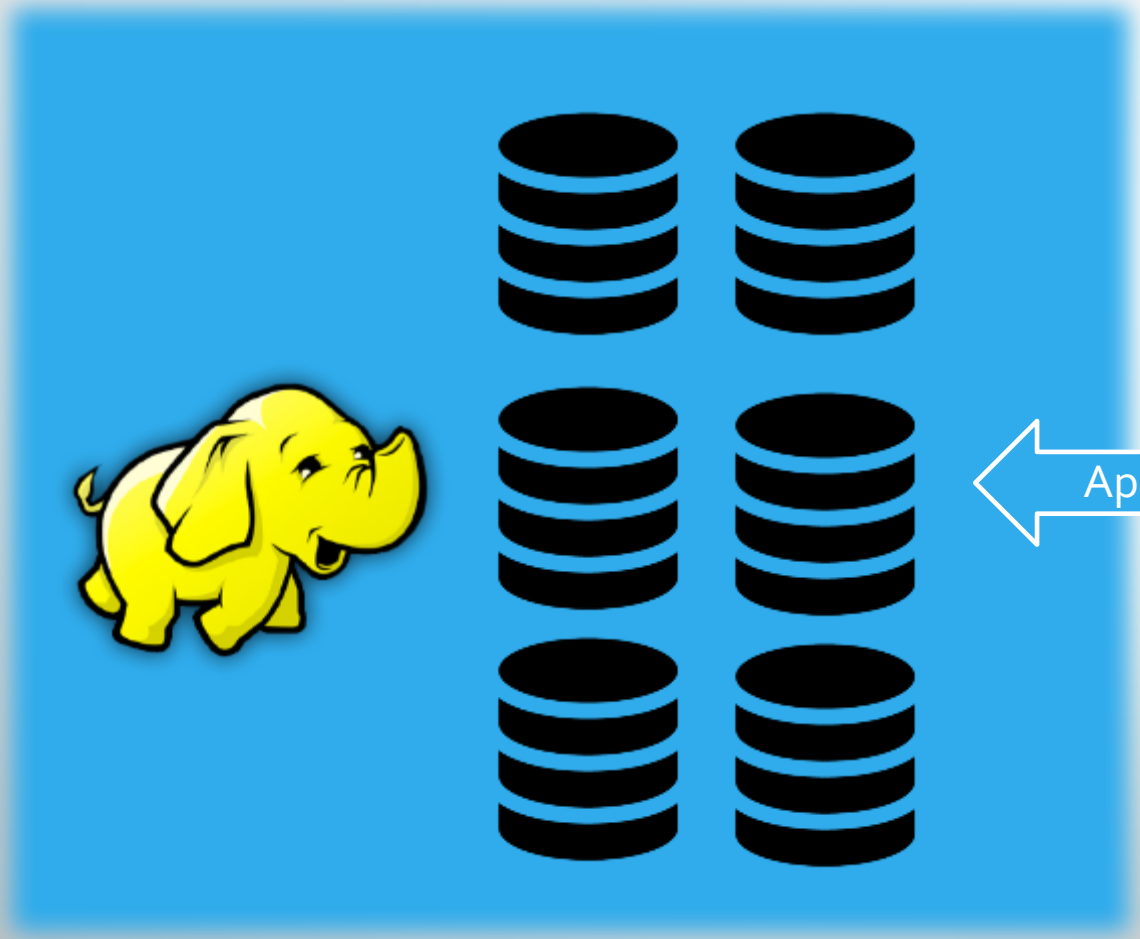
Big data

- Large volumes of data storage & processing
- Highly parallelized algorithms
- Sophisticated architecture
- Hardware-related (clusters, nodes, server machines)

Machine learning

- Smart data processing methods
- Domain-agnostic
- Technology-agnostic
- Hardware-agnostic
- Predictions and modelling
- Strongly related to statistics

WHAT IS MACHINE LEARNING?

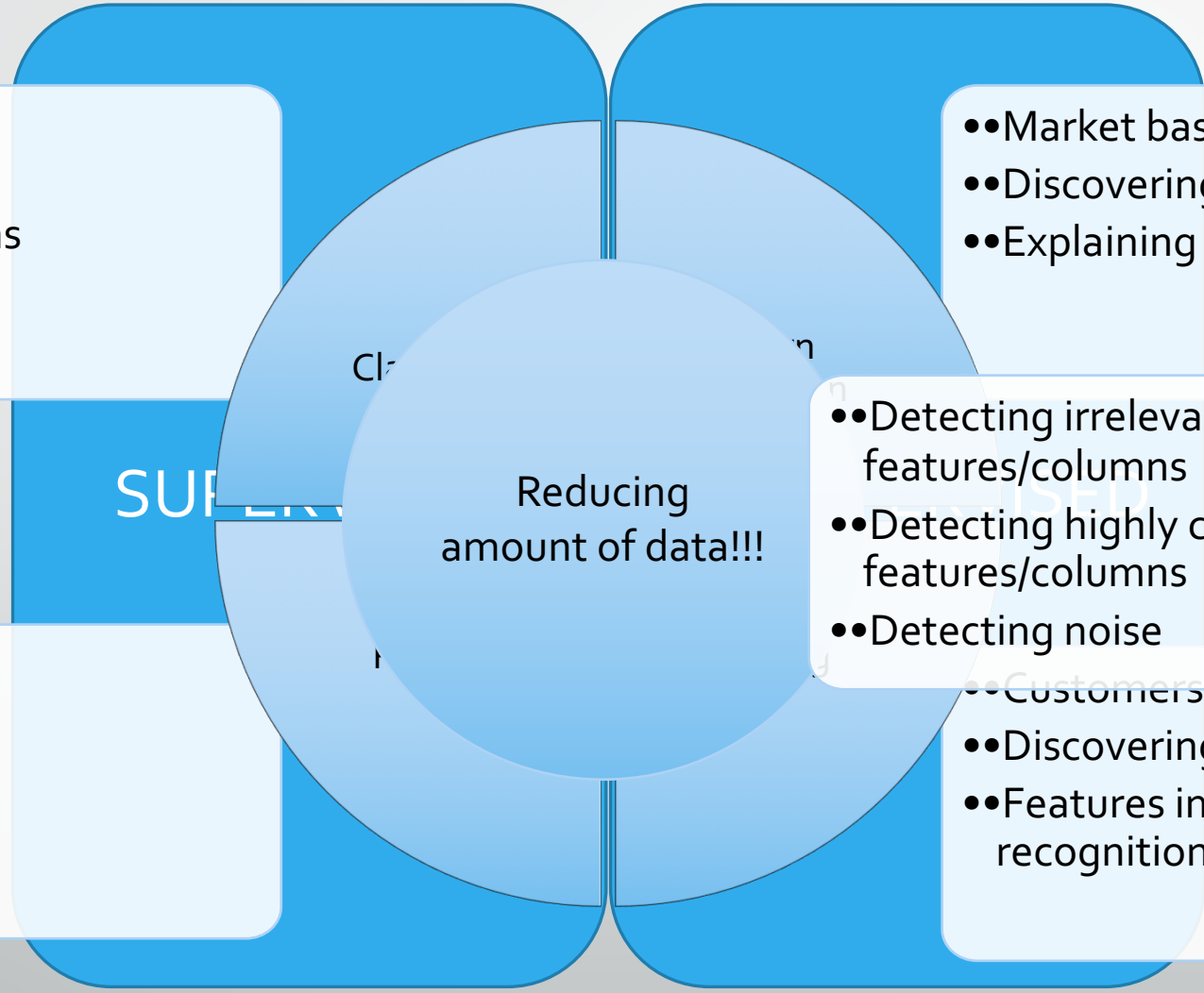


Apply to

Hadoop cluster
(data storage)

Machine learning algorithms
(data processing)

MACHINE LEARNING USE CASES



- Customer preferences discovery
- Automated expert systems construction
- Assigning new data to groups

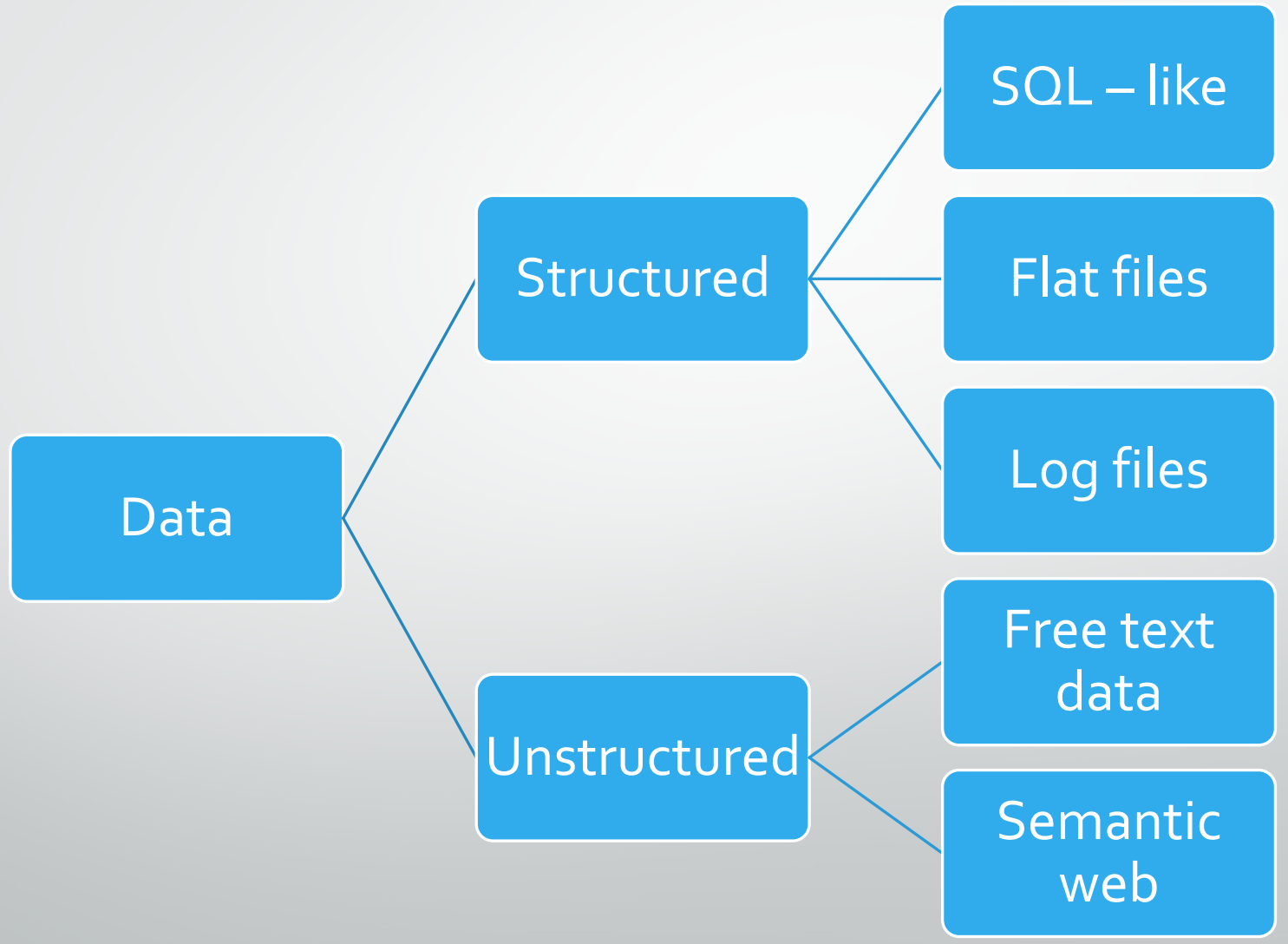
- Market basket analysis
- Discovering preferences
- Explaining data

- Financial trends discovery
- Statistical analysis
- Prediction of numerical values/outcomes

- Detecting irrelevant features/columns
- Detecting highly correlated features/columns
- Detecting noise

- Customers grouping
- Discovering similarities
- Features importance recognition

DATA FOR MACHINE LEARNIG



DATA FRAME – BASIC DATA STRUCTURE OVERVIEW

Features/attributes

Discrete features

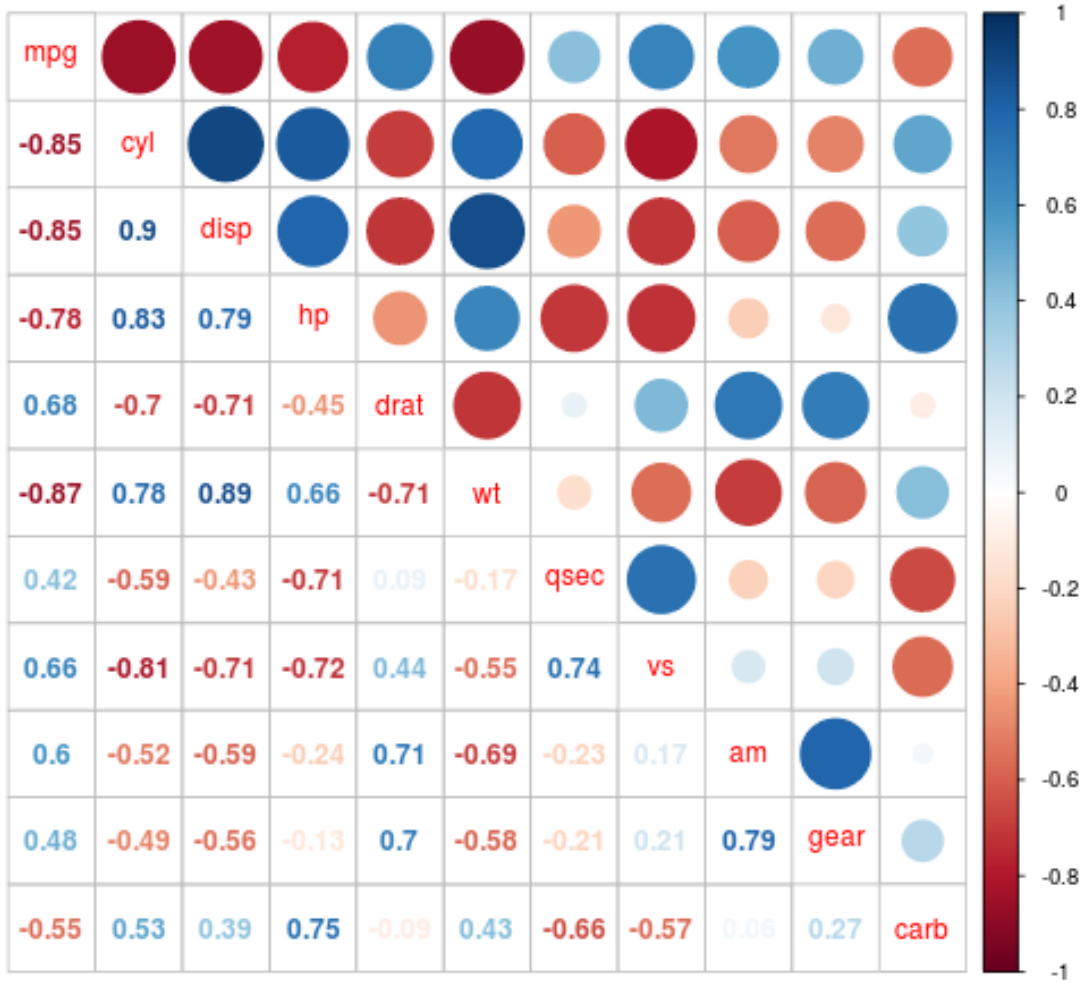
Boolean feature

Numerical feature

Company	**Financial instruments**	**Status**	**Revenue**
Company X	Equities	Open	0.6
Company Y	Corporate Bonds	Open	0.03
Company Z	Structure hybrid	Closed	0.02

Instances

DATA FRAME – BASIC DATA STRUCTURE
FEATURES INVESTIGATION



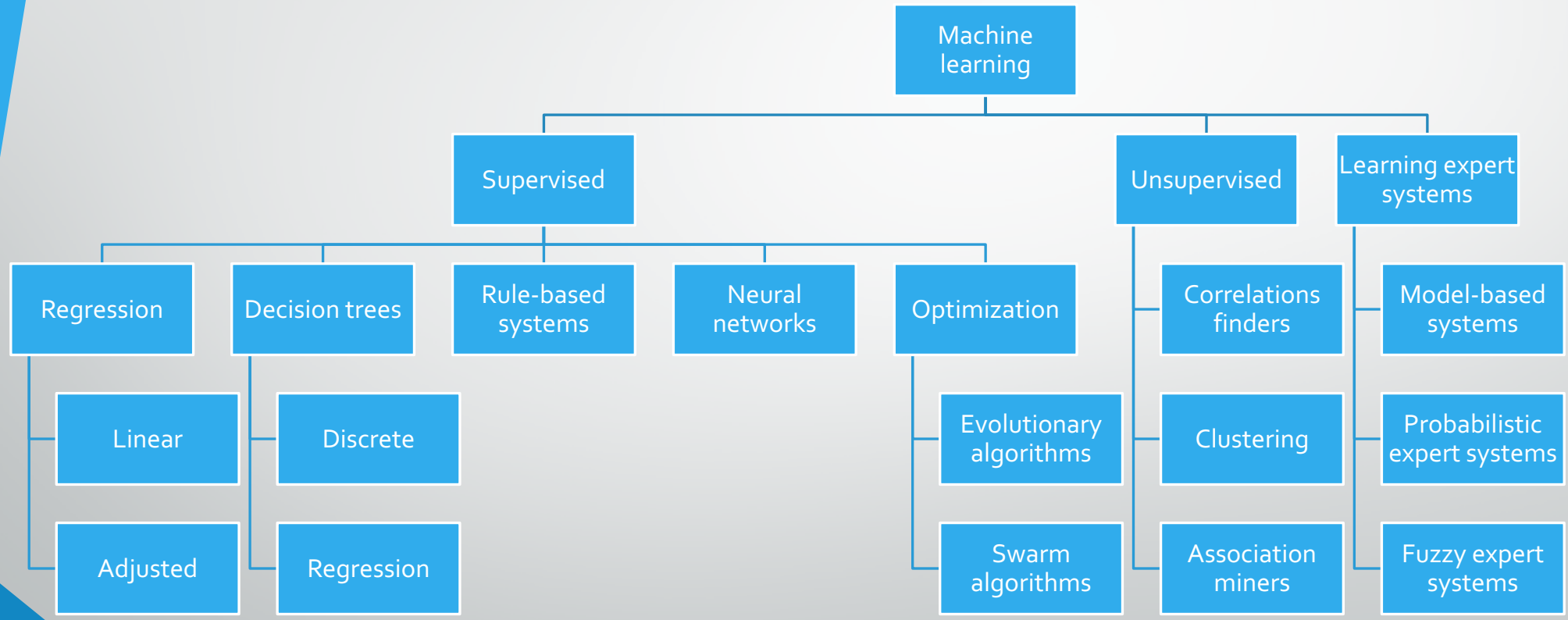
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

DATA FRAME – BASIC DATA STRUCTURE
FEATURES ENCODING

Company	Financial instruments	Status	Revenue
Company X	Equities	Open	0.6
Company Y	Corporate Bonds	Open	0.03
Company Z	Structure hybrid	Closed	0.02

Company	Financial instruments	Status	Revenue
001	001	1	0.6
010	010	1	0.03
100	100	0	0.02

ALGORITHMS OVERVIEW

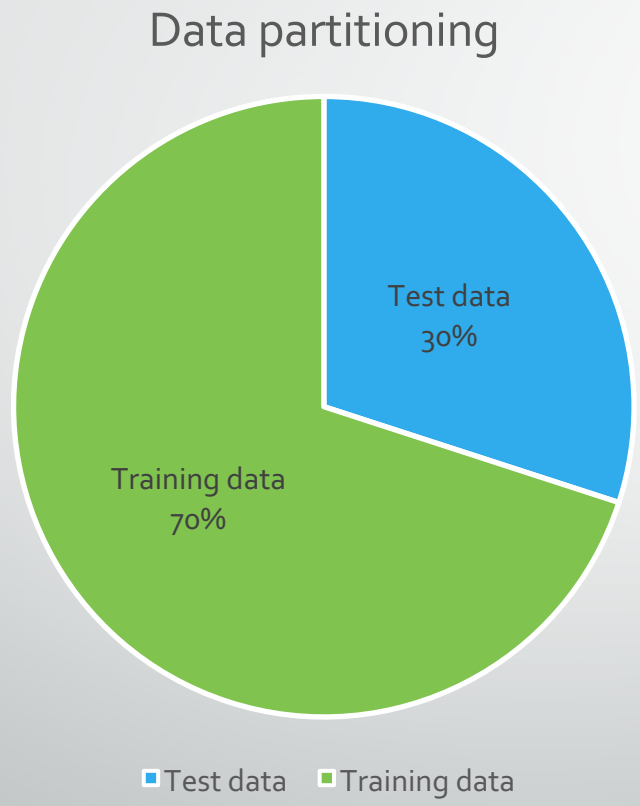


SUPERVISED LEARNING

SUPERVISED LEARNING

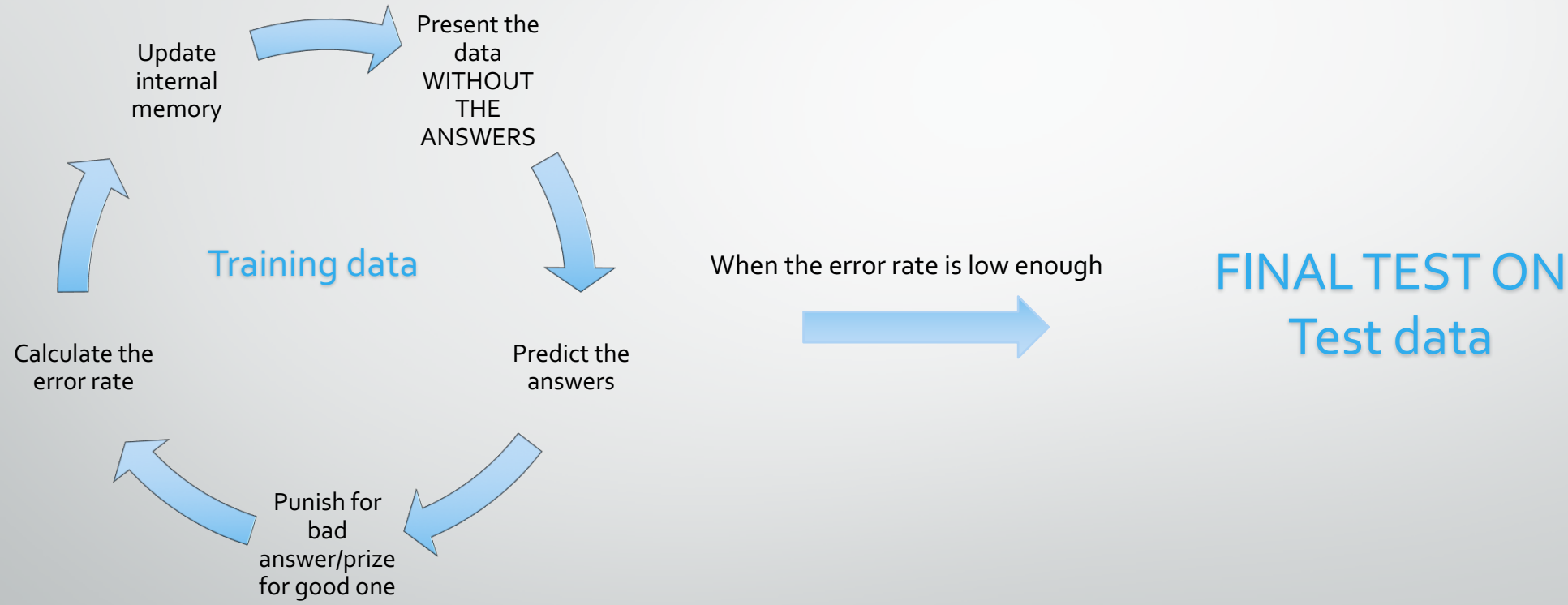
Two data sets	••Training– known „answers", given to algorithm ••Test– known „answers", not given to algorithm
"Teacher/oracle"	••Objective rating function ••Checks the algorithm progress
Learning based on the experience	••Application of teachers/oracle suggestions to improve score ••Avoiding overfitting

SUPERVISED LEARNING



- Sometimes the amount of data with known „answers” is limited
- Data division helps in better controlling the learning process
- Improving the effectiveness of data usage

SUPERVISED LEARNING



SUPERVISED LEARNING
Decision trees

General approach

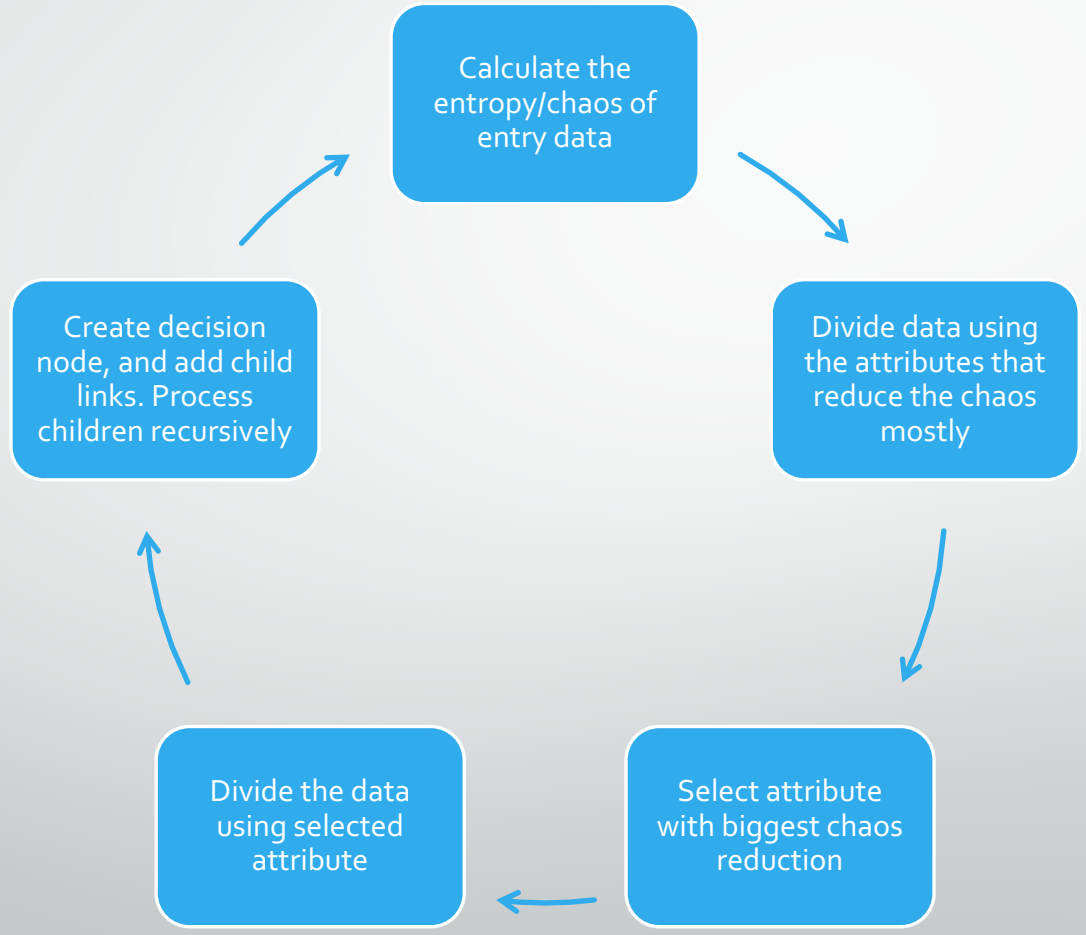
- Uses structured data
- Recursive top-down approach: divide and conquer, based on the best-promising attributes
- Can use numerical and discrete data as well

Pros

- Very flexible
- Easy to implement
- Easy to interpret by humans
- Can be translated to easy-to-read rules and included in reports/documentations

SUPERVISED LEARNING

Decision trees



SUPERVISED LEARNING
Decision trees

client	hotel	addons	money_spent	offer
business	Hilton	trip	40000	deluxe
business	Hilton	full board	38000	deluxe
business	Hilton	trip	40000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1500	premium

Value	Count	%
Deluxe	3	0.5
Basic	2	0.333
Premium	1	0.16666

client	hotel	addons	money_spent	offer
business	Hilton	trip	40000	deluxe
business	Hilton	full board	38000	deluxe
business	Hilton	trip	40000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1500	premium

Split on: „client"
Value to split: „business"

Client == business ?

True

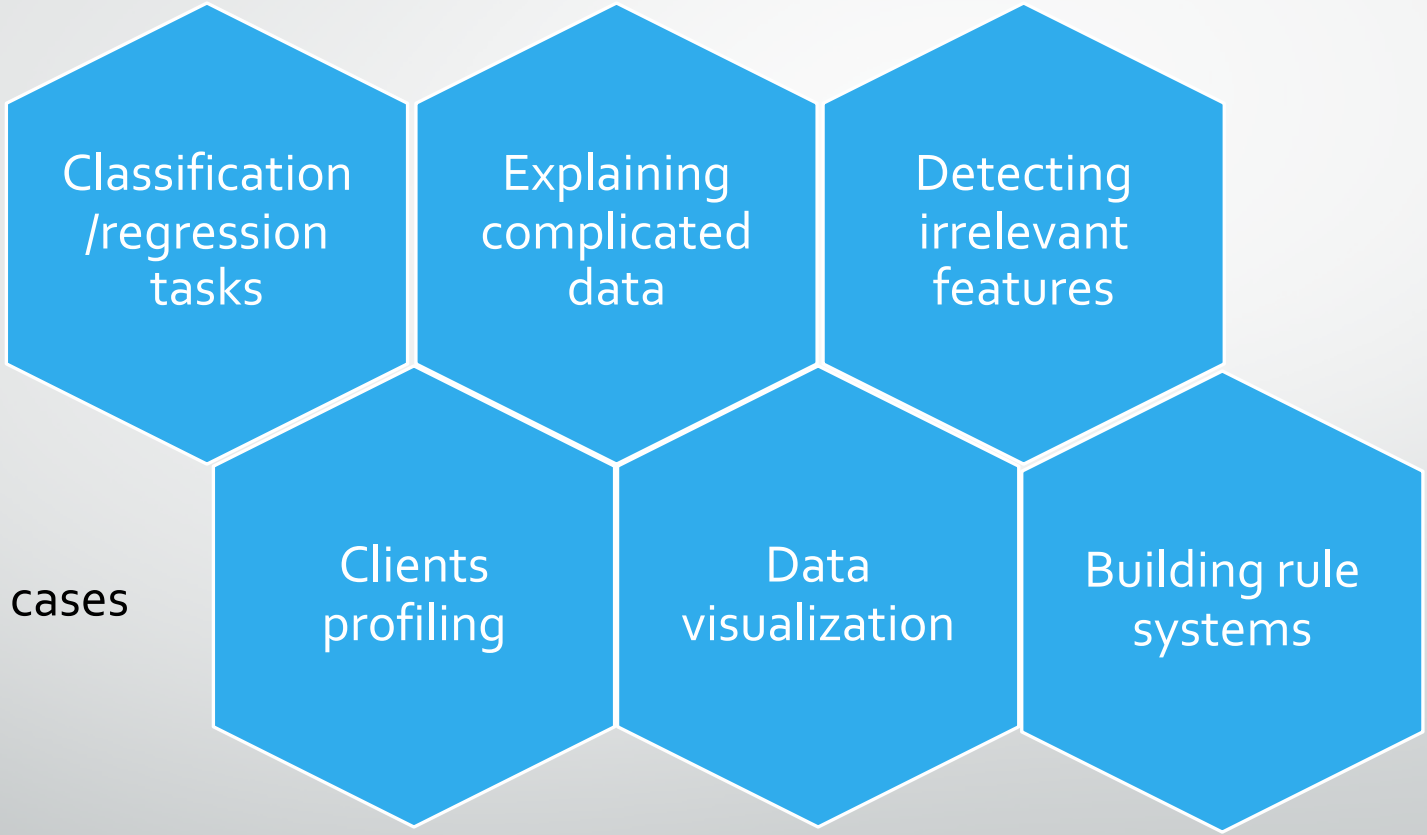
False

hotel	addons	money_spent	offer
Hilton	trip	40000	deluxe
Hilton	full board	38000	deluxe
Hilton	trip	40000	deluxe

hotel	addons	money_spent	offer
Meta	none	800	basic
Meta	meal	900	basic
Meta	spa	1500	premium

SUPERVISED LEARNING
Decision trees

Use cases



UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

One data set	••Single set of data ••No „good answers" provided (in most cases)
No teacher/oracle	••No option to evaluate prediction against „correct answers" ••Algorithm evaluation based on similarity measures/chaos measures/etc.
Algorithm operates on data on its own	••Algorithm explores the possible data partitioning ••Algorithm maintains its internal error measures

UNSUPERVISED LEARNING
Association rules learning

General approach

- Ordered data
- Searching for coincidences/correlations in data

Features

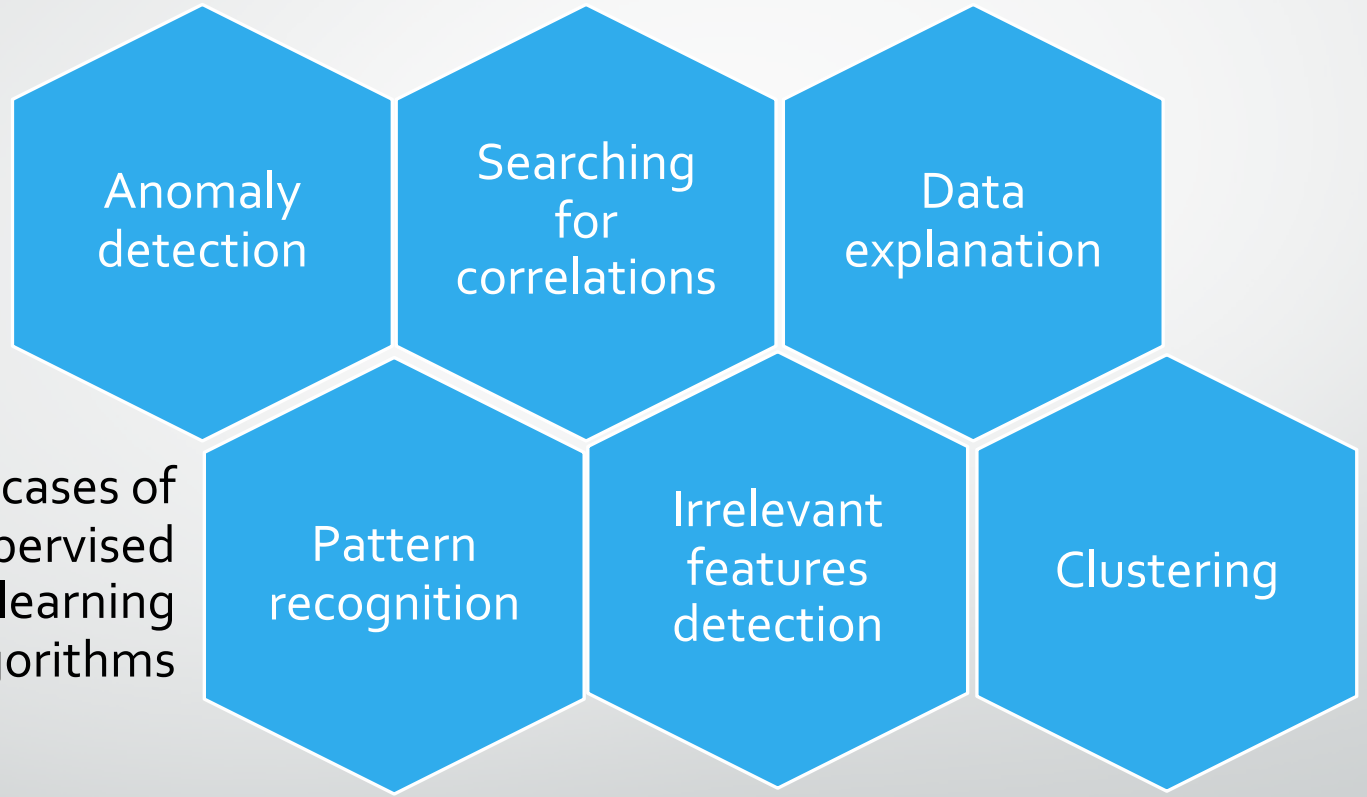
- Easy to implement
- Flexible
- Easy to interpret by humans
- Can significantly reduce the amount of irrelevant features

UNSUPERVISED LEARNING
Association rules learning

Transaction no.	Products
1.	1. Soya milk 2. Salad
2.	1. Salad 2. Walnuts 3. Wine 4. Bread
3.	1. Soya milk 2. Walnuts 3. Wine 4. Juice
4.	1. Salad 2. Soya milk 3. Walnuts 4. Wine
5.	1. Salad 2. Soya milk 3. Walnuts 4. Juice

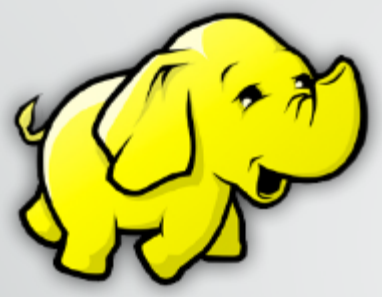
UNSUPERVISED LEARNING

Use cases of unsupervised learning algorithms



TOOLS OF THE TRADE

MACHINE LEARNING TOOLS



Spark

R

Accord
Extensions .NET

scikit learn

WEKA
The University of Waikato

Encog