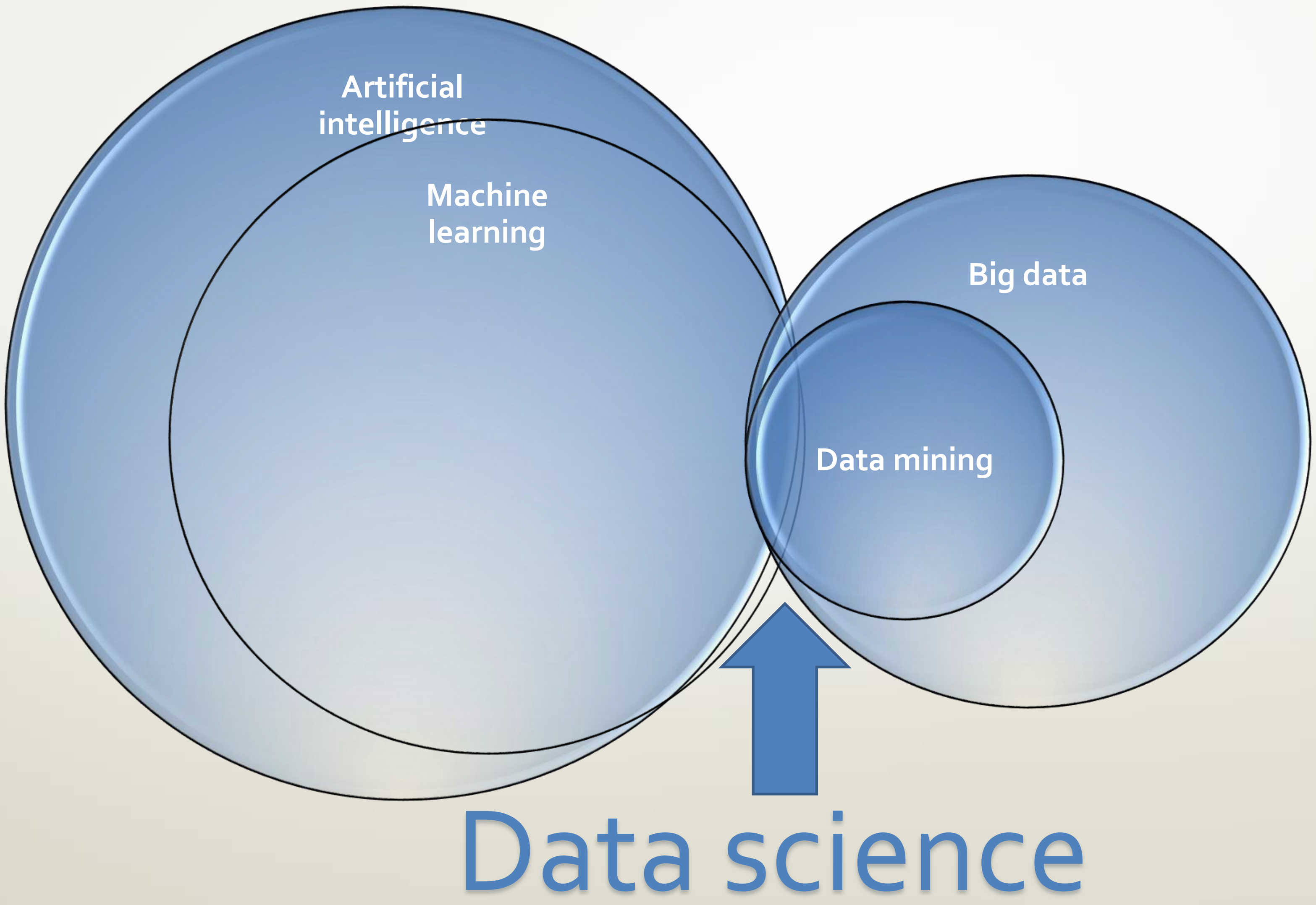
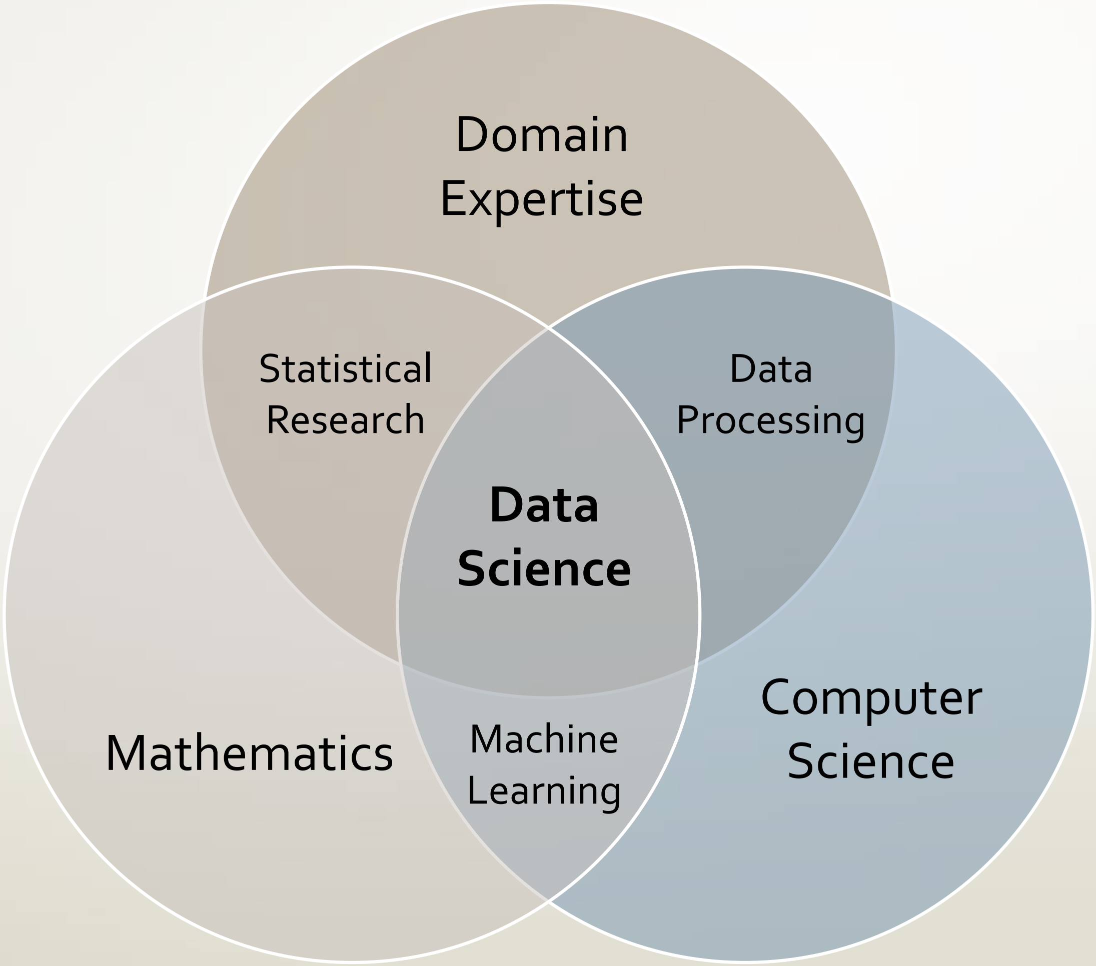
MACHINE LEARNING:
when big data is not enough

Filip Wójcik
Data scientist, senior .NET developer
Wroclaw University lecturer
filip.wojcik@outlook.com

What is machine learning? (1/4)

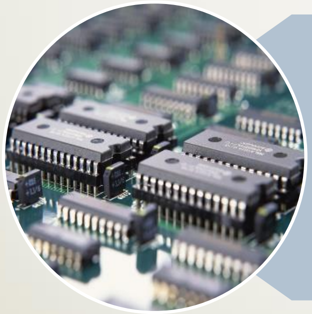


What is machine learning? (2/4)



What is machine learning? (3/4)

Data volumes are increasing



Need to process massive amounts of data

Data analysis processes automation

What is machine learning? (4/4)

Big data	Machine learning
• Large volumes of data storage & processing • Highly parallelized algorithms • Sophisticated architecture • Hardware-related (clusters, nodes, server machines)	• Smart data processing methods • Domain-agnostic • Technology-agnostic • Hardware-agnostic • Predictions and modelling • Strongly related to statistics

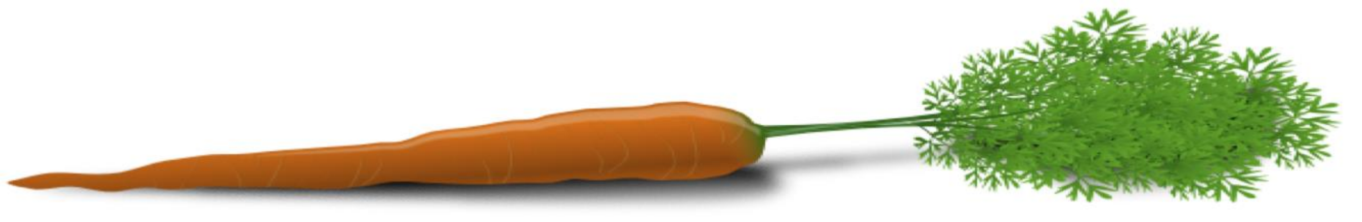
Machine learning tools

hadoop

mahout

Spark

WEKA
The University
of Waikato



H_2O.ai

scikit learn

pandas

STATISTICA

C++

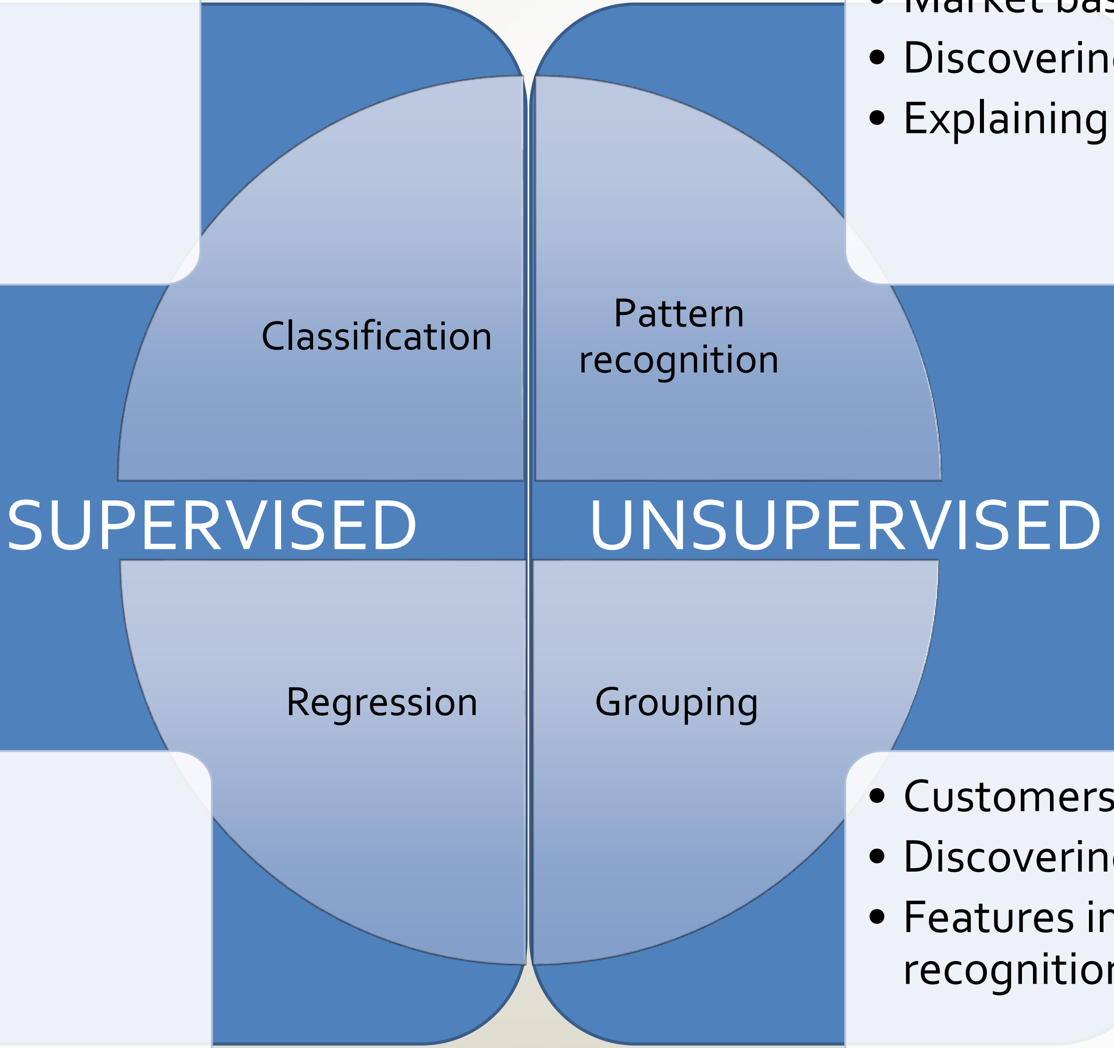
mlpack

lib C++ Library

Machine learning use cases (1/2)

- 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

- Customers grouping
- Discovering similarities
- Features importance recognition

Machine learning use cases (1/2)

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

- Market basket analysis
- Discovering preferences
- Explaining data

Reducing amount of data!!!

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

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

- Customers grouping
- Discovering similarities
- Features importance recognition

Machine learning use cases (2/2)

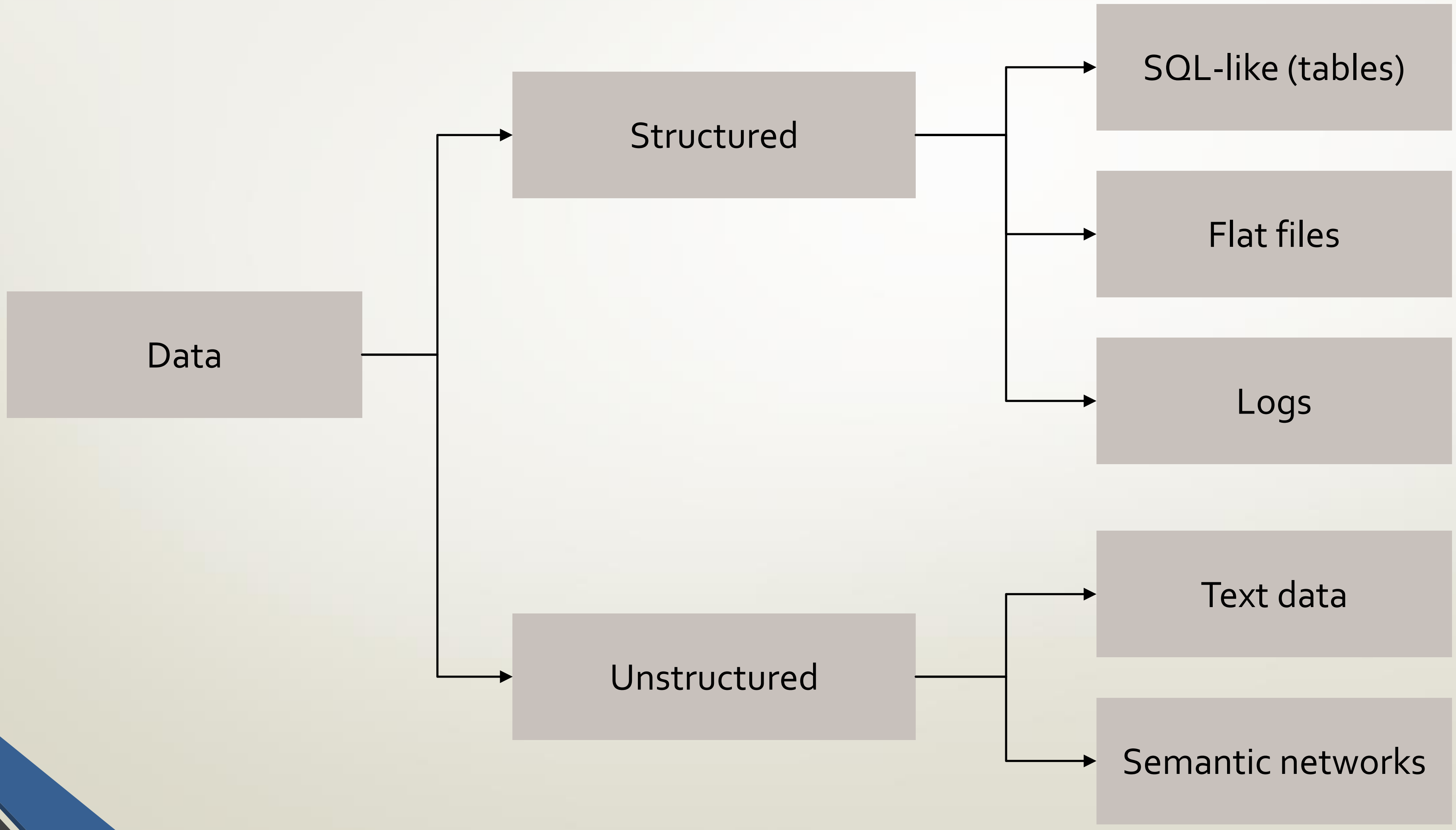
„Black box" methods

- Cannot be interpreter by humans
- Their internal structure is complicated and is hard to understand
- Mostly – very sophisticated mathematically
- „Justifications" of predictions are purely mathematical

„White box" methods

- Easily interpretable
- Can be translated to human-friendly form
- Not so sophisticated mathematically

Key data structures (1/3)



Data Frame

Key data structures (2/3)

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

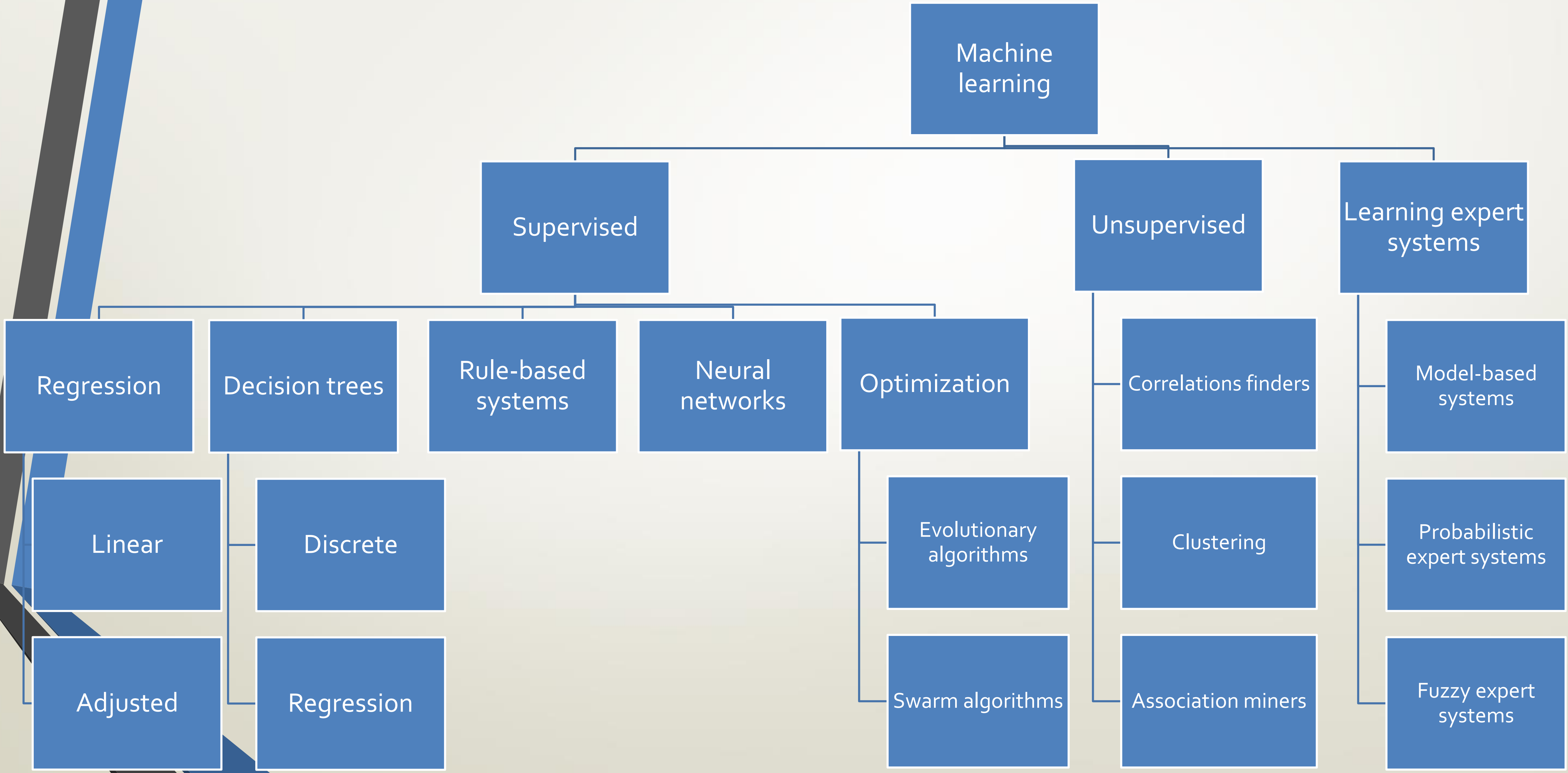
Records/objects

Data Frame
Key data structures (3/3)

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

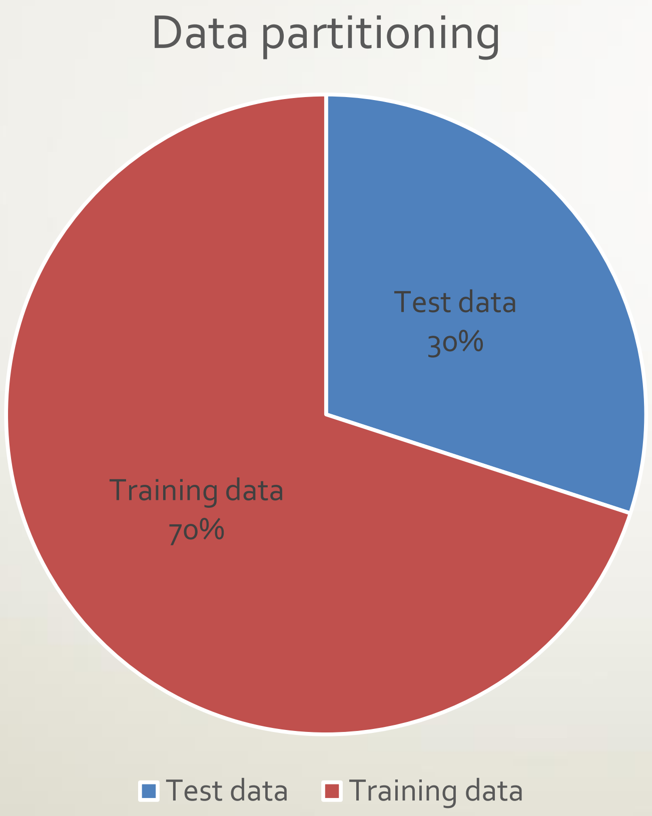


Supervised learning

Supervised learning (1/3)

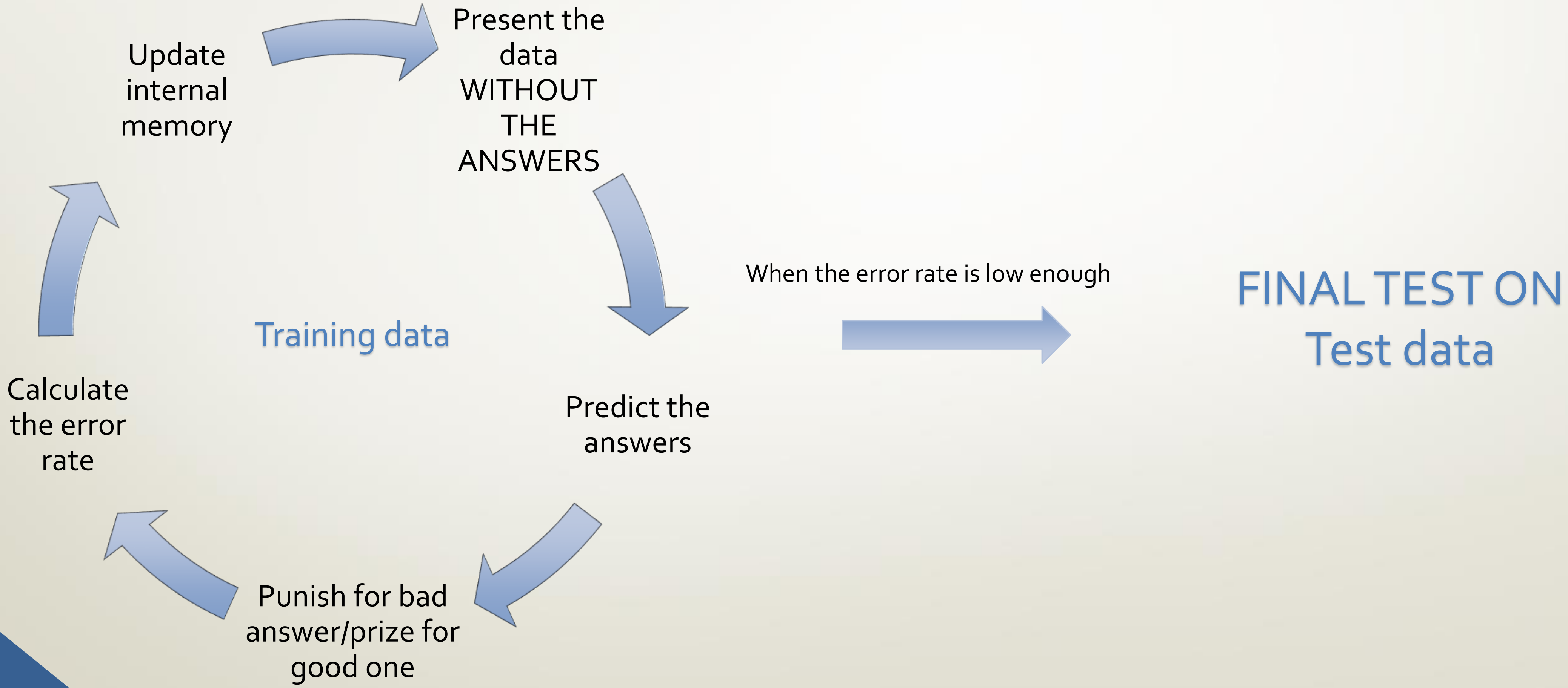
Two data sets	• Training– known „answers", given to algorithm • Test– known „answers", not given to algorithm
"Teacher/oracle"	• Objective rating function • Checks the algorithm progress
Learning based on the experience	• Application of teachers/oracle suggestions to improve score • Avoiding overfitting

Supervised learning (2/3)



- 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 (3/3)



Supervised learning – decision trees

Supervised learning
Decision trees (1/5)

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 (2/5)

1. Calculate the entropy/chaos of entry data
2. Divide data using the attributes that reduce the chaos mostly
3. Select attribute with biggest chaos reduction
4. Divide the data using selected attribute
5. Create decision node, and add child links. Process children recursively

Supervised learning

Decision trees (3/5)

client	hotel	addons	money_spent	offer
business	Hilton	trip	40,000	deluxe
business	Hilton	full board	38,000	deluxe
business	Hilton	trip	40,000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1,500	premium

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

Supervised learning

Decision trees (4/5)

client	hotel	addons	money_spent	offer
business	Hilton	trip	40,000	deluxe
business	Hilton	full board	38,000	deluxe
business	Hilton	trip	40,000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1,500	premium

Client == business?

True

hotel	addons	money_spent	offer
Hilton	trip	40,000	deluxe
Hilton	full board	38,000	deluxe
Hilton	trip	40,000	deluxe

False

hotel	addons	money_spent	offer
Meta	none	800	basic
Meta	meal	900	basic
Meta	spa	1,500	premium

Supervised learning
Decision trees (5/5)

Use cases

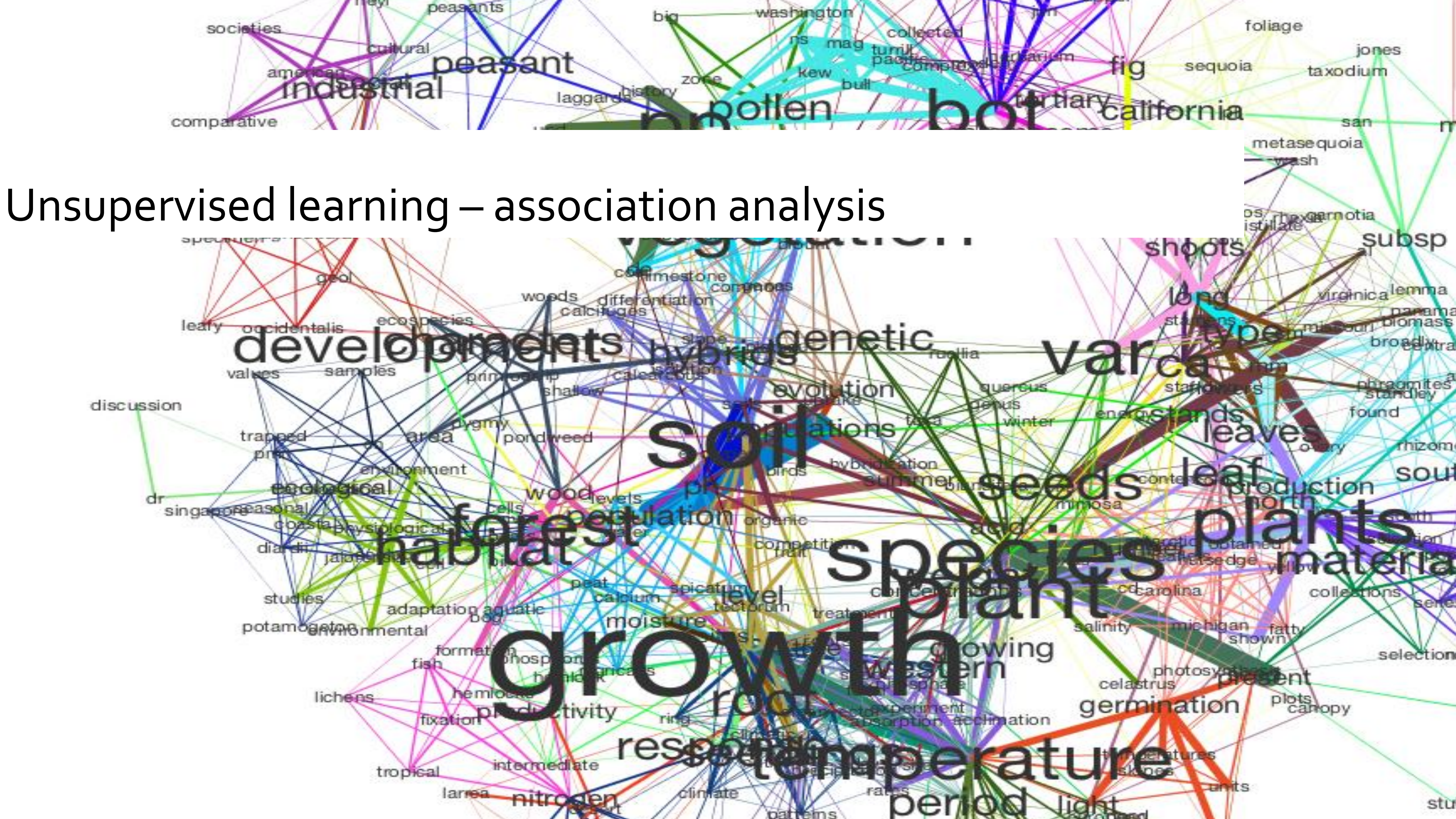
- Classification /regression tasks
- Explaining complicated data
- Detecting irrelevant features
- Clients profiling
- Data visualization
- Building rule systems

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 analysis



Unsupervised learning
Association analysis (1/3)

General approach

- Ordered data
- Searching for coincidences/correlations in data

Features

- Works only with nominal data or discretized (binned)/thresholded numeric data
- Easy to implement
- Flexible
- Easy to interpret by humans
- Can significantly reduce the amount of irrelevant features

Unsupervised learning
Association analysis (2/3)

Transaction number	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

Frequent items	support
Soya, salad	0.4
Soya, salad, walnuts	0.4
Salad	0.6

Implications	support
Soya => walnuts	0.4
Soya => salad	0.4
Soya, Walnuts, Wine => juice	0.4

Unsupervised learning
Association analysis (3/3)

Use cases of unsupervised learning algorithms

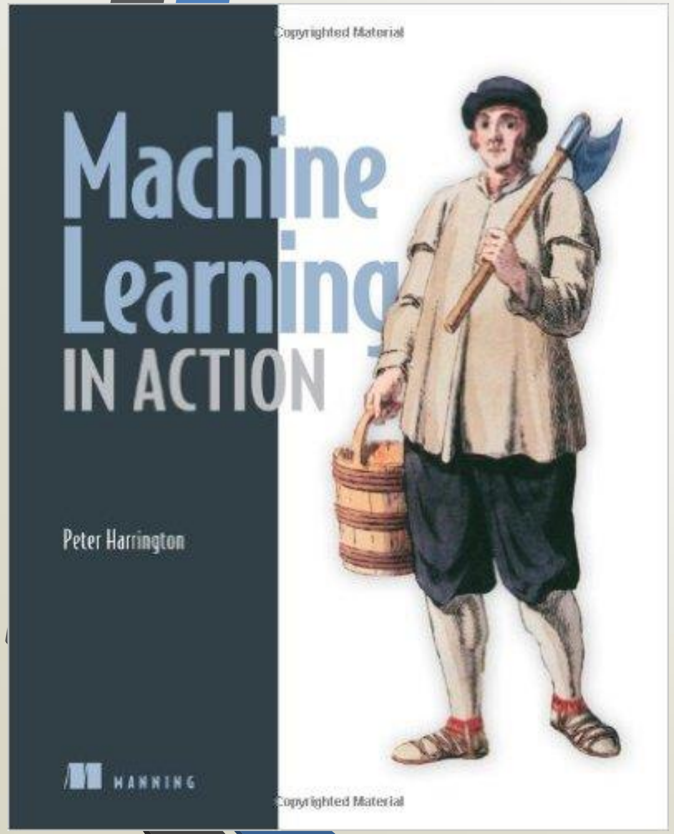
- Anomaly detection
- Searching for correlations
- Data explanation
- Pattern recognition
- Irrelevant features detection
- Clustering

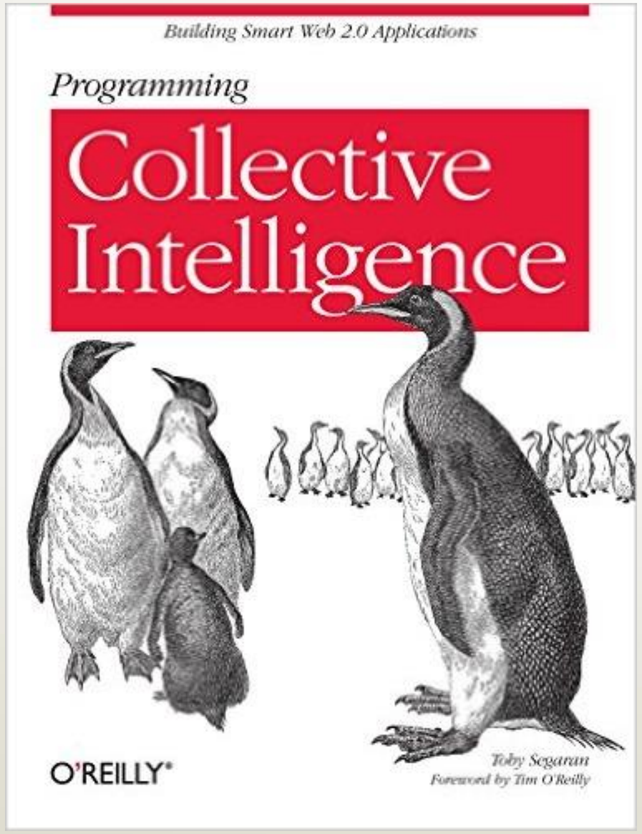
Must-reads

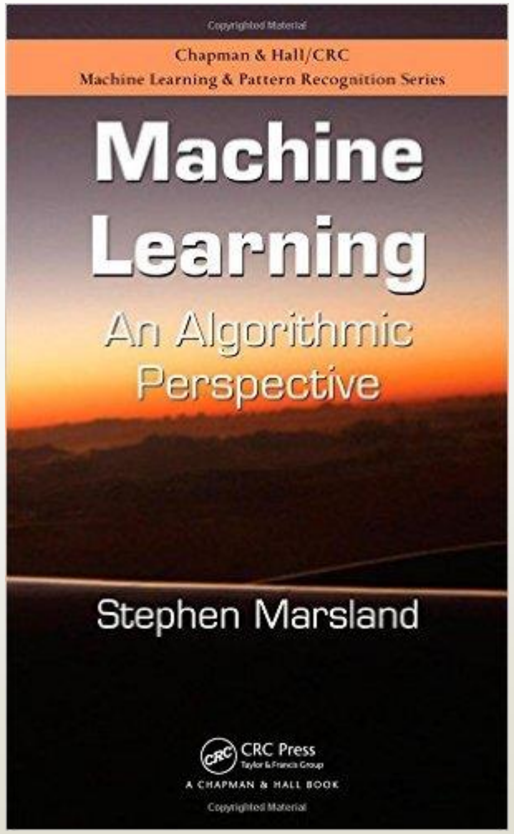
ML lecutures

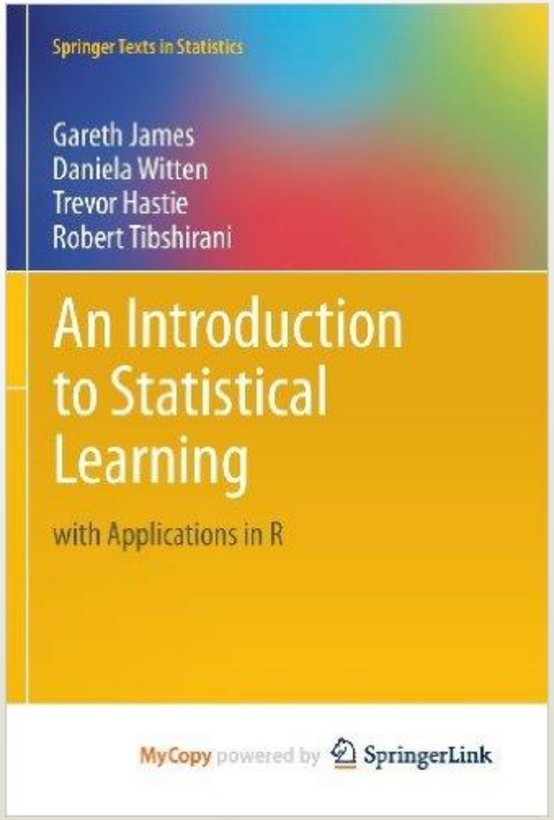
Pracical examples & code

Math & theory









THANK YOU!