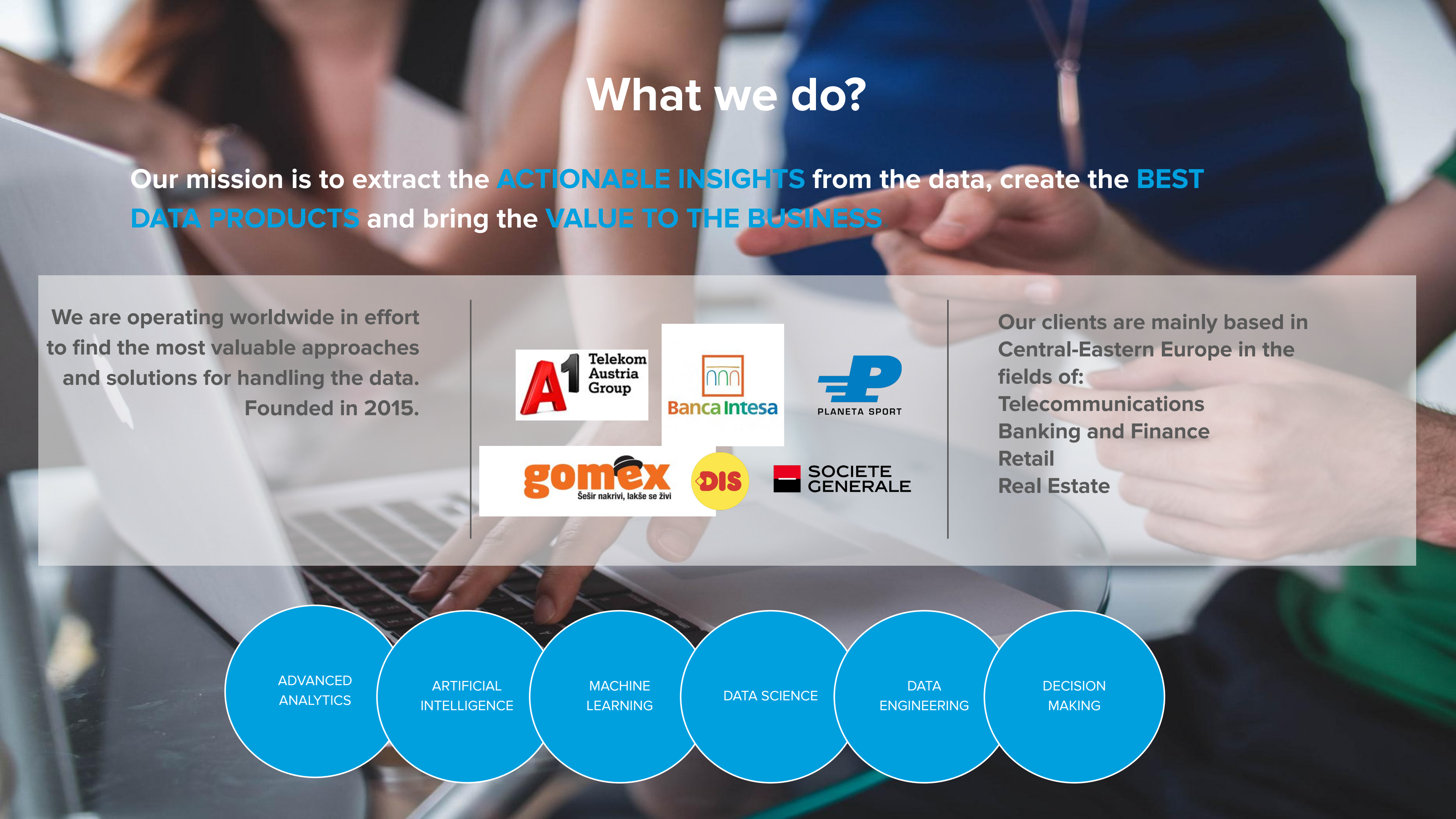
Machine learning techniques for anomaly detection

Valentina Djordjevic
valentina@thingsolver.com



What we do?

Our mission is to extract the **ACTIONABLE INSIGHTS** from the data, create the **BEST DATA PRODUCTS** and bring the **VALUE TO THE BUSINESS**.

We are operating worldwide in effort to find the most valuable approaches and solutions for handling the data. Founded in 2015.

A1 Telekom Austria Group

Banca Intesa

PLANETA SPORT

gomex Šešir nakrivi, lakše se živi

DIS

SOCIETE GENERALE

Our clients are mainly based in Central-Eastern Europe in the fields of:

Telecommunications
Banking and Finance
Retail
Real Estate

ADVANCED ANALYTICS

ARTIFICIAL INTELLIGENCE

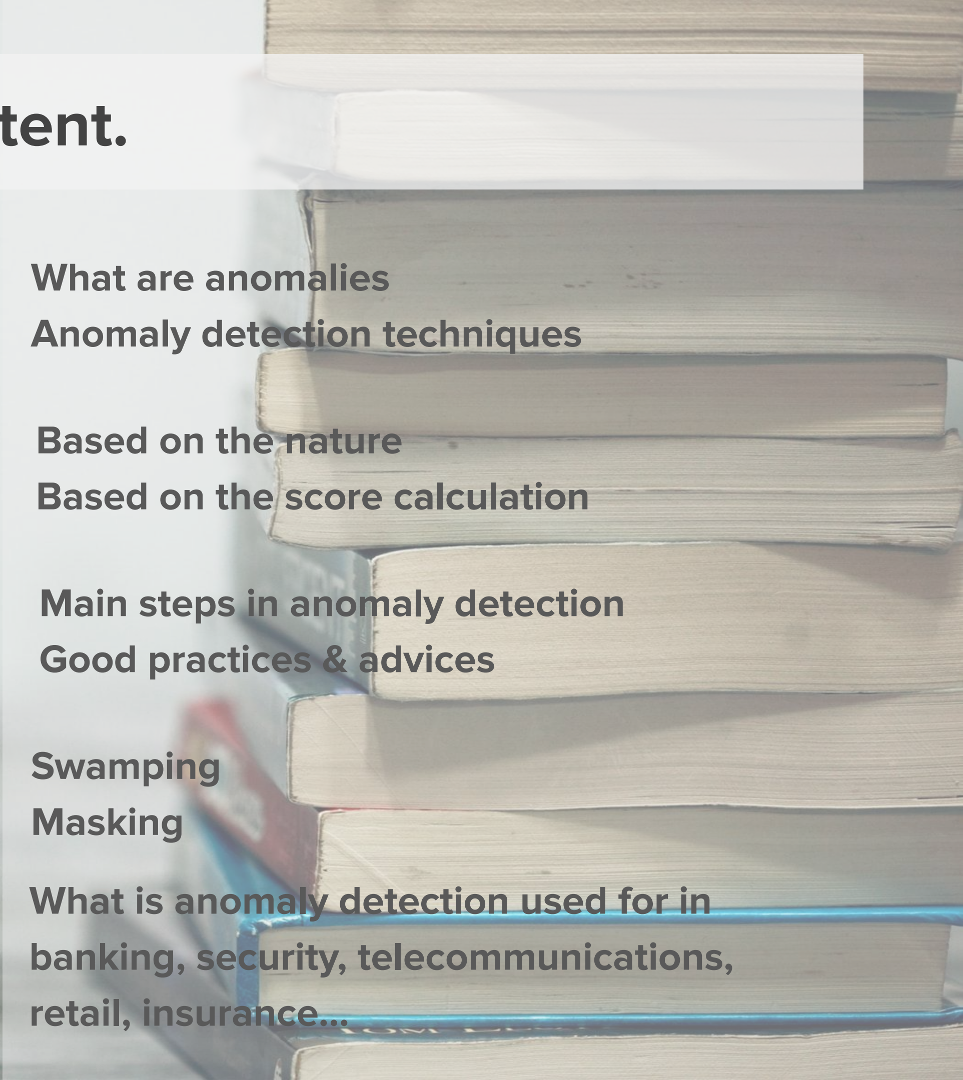
MACHINE LEARNING

DATA SCIENCE

DATA ENGINEERING

DECISION MAKING

Content.



01 Defining the term "anomaly"

- What are anomalies
- Anomaly detection techniques

02 Anomaly detection techniques

- Based on the nature
- Based on the score calculation

03 Methodology

- Main steps in anomaly detection
- Good practices & advices

04 Bottlenecks

- Swamping
- Masking

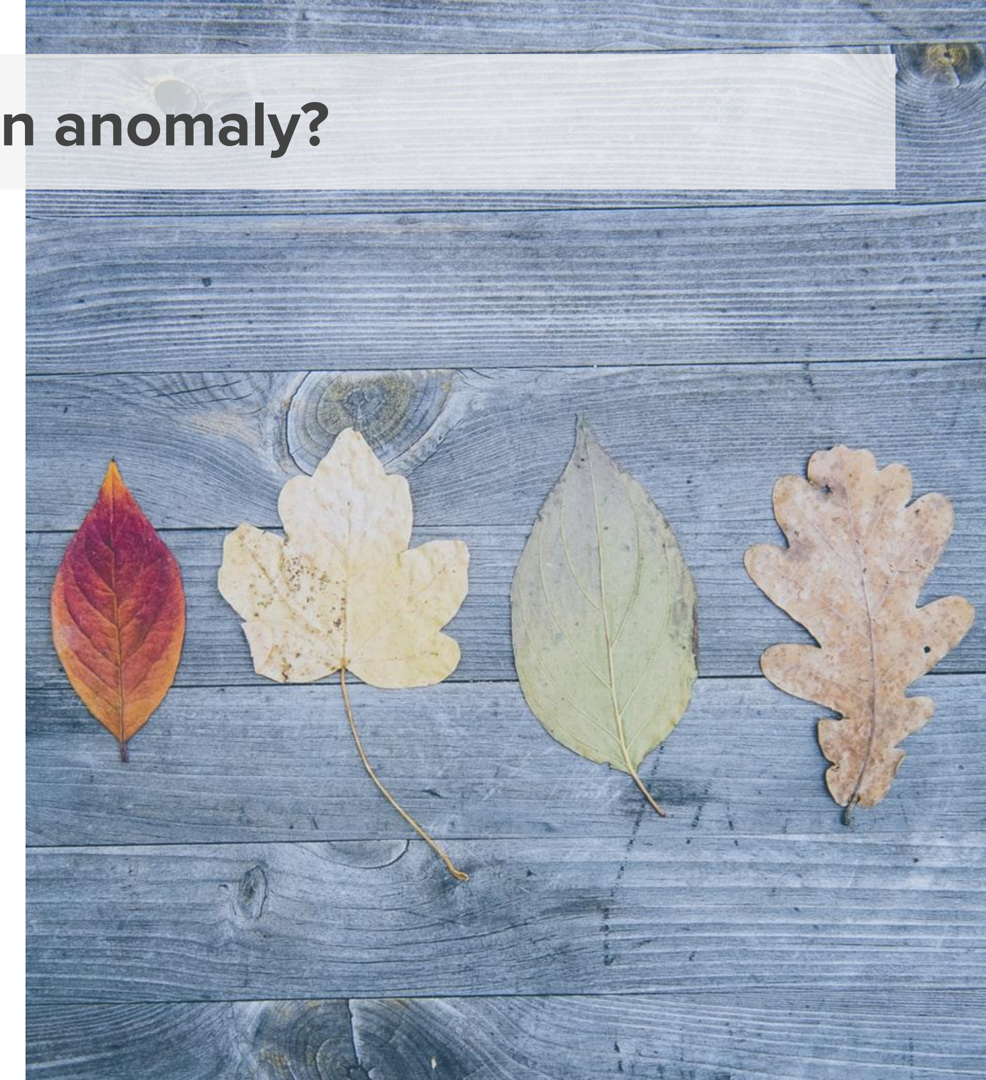
05 Use cases

- What is anomaly detection used for in banking, security, telecommunications, retail, insurance...

What is an anomaly?

- Anomaly represents the type of behaviour in the data that differs significantly from some expected behaviour.
- Anomaly != Outlier != Novelty
- Types of anomalies:

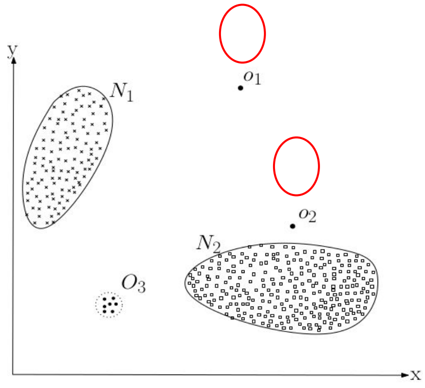
1. Point anomalies
2. Contextual anomalies
3. Collective anomalies

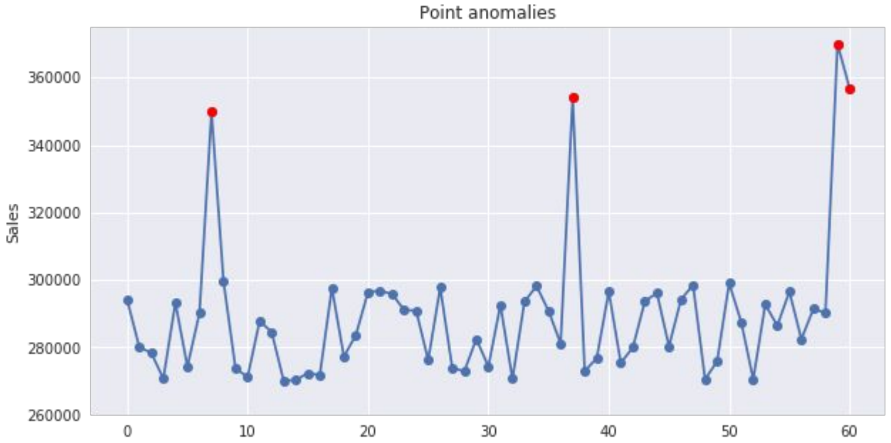


Point anomaly.

Point anomaly is an instance that could be considered as anomalous among other instances in the dataset.

Point anomalies often represent some extremum, irregularity or deviation that happens randomly and have no particular meaning.

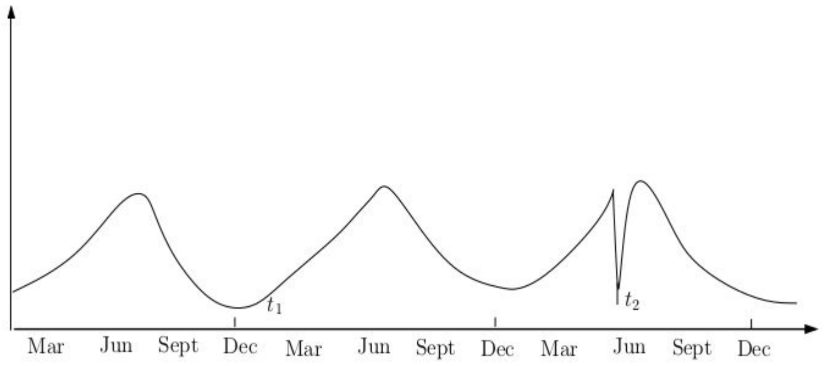


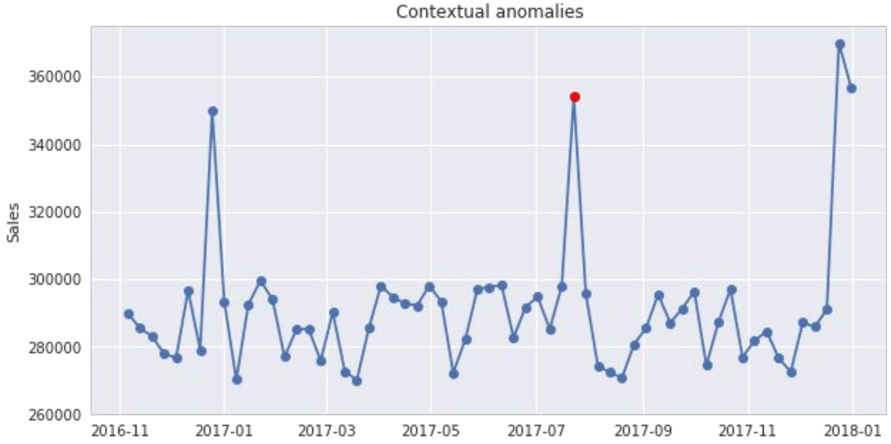


Contextual anomaly.

Contextual anomaly is an instance that could be considered as anomalous in some specific context.

The contextual anomaly is determined by combining contextual and behavioural features, like space and/or time with some quantitative measurement (total money spent, average temperature, average end user throughput,...)

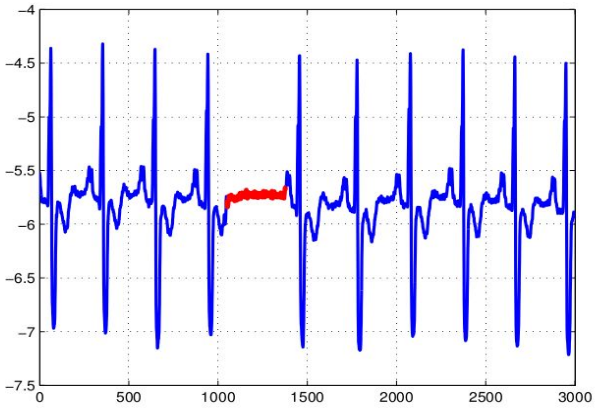


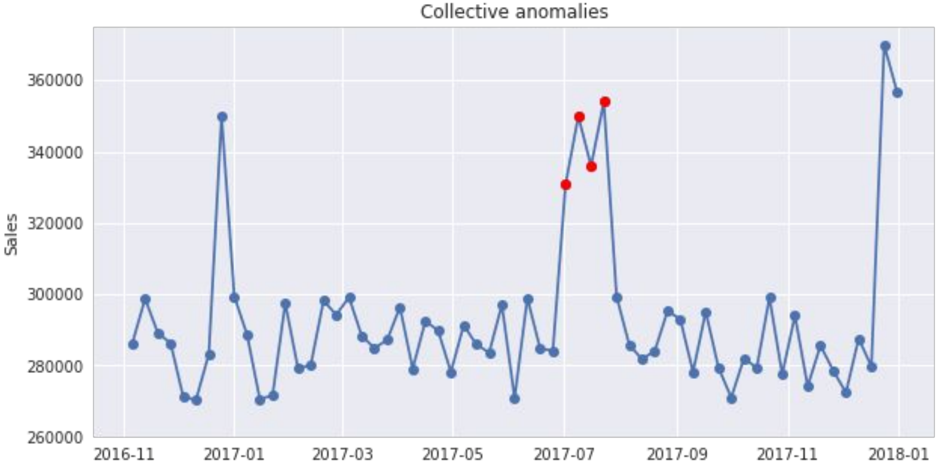


Collective anomaly.

Collective anomaly is often represented as a group of correlated, interconnected or sequential instances.

While each particular instance of this group doesn’t have to be anomalous itself, their collective occurrence is anomalous.





Techniques.

Supervised anomaly detection.

Unsupervised anomaly detection.

Semi-supervised anomaly detection.

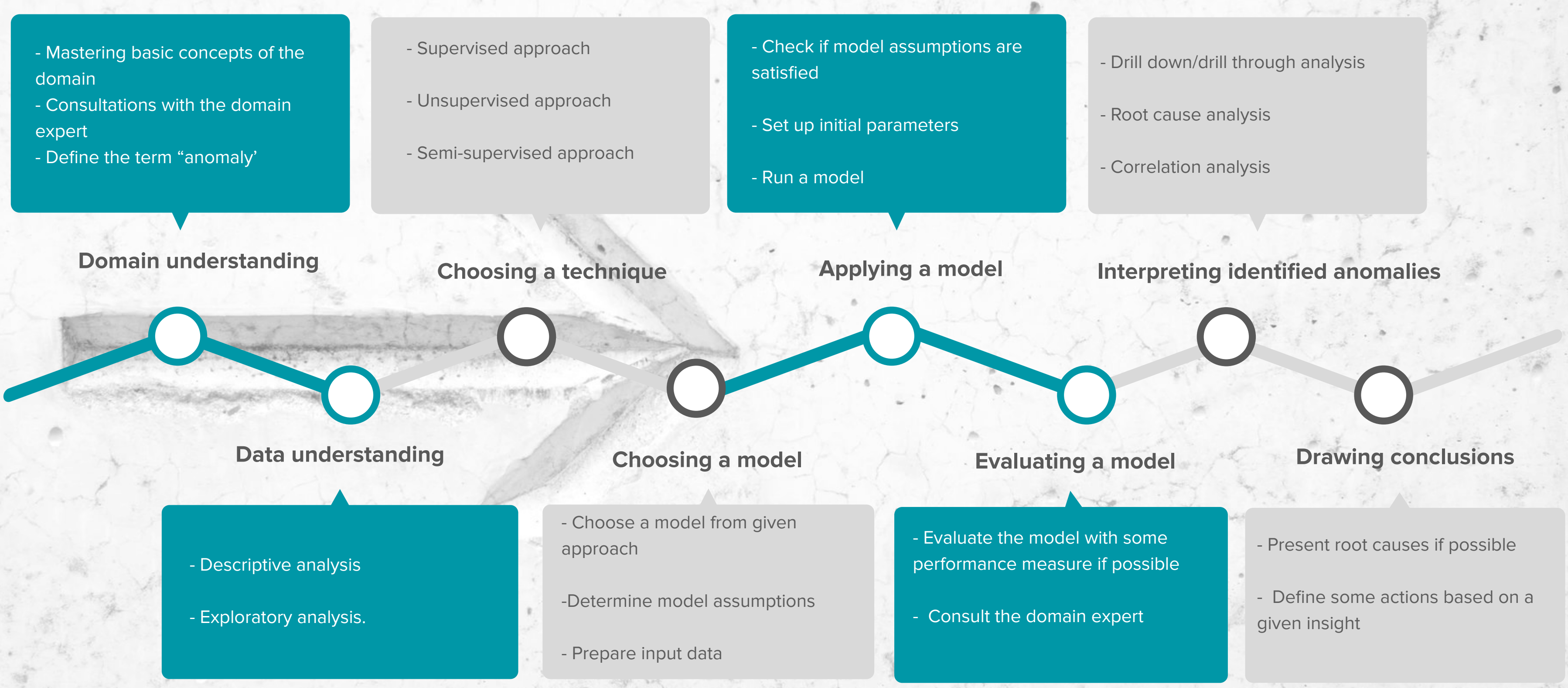
Techniques

- Classification
- Clustering
- Neighbour-based
- Statistical methods
- Information theory
- Spectrum theory

Anomaly score

- Probability-based
- Distance-based
- Density-based
- Path-length based
- Entropy-based

Methodology.



Domain understanding

- Mastering basic concepts of the domain
- Consultations with the domain expert
- Define the term “anomaly’

Data understanding

- Descriptive analysis
- Exploratory analysis.

Choosing a technique

- Supervised approach
- Unsupervised approach
- Semi-supervised approach

Choosing a model

- Choose a model from given approach

-Determine model assumptions

- Prepare input data

Applying a model

- Check if model assumptions are satisfied
- Set up initial parameters
- Run a model

Evaluating a model

- Evaluate the model with some performance measure if possible
- Consult the domain expert

Interpreting identified anomalies

- Drill down/drill through analysis
- Root cause analysis
- Correlation analysis

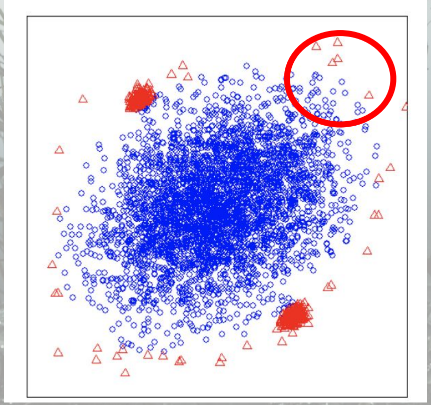
Drawing conclusions

- Present root causes if possible
- Define some actions based on a given insight

Bottlenecks.

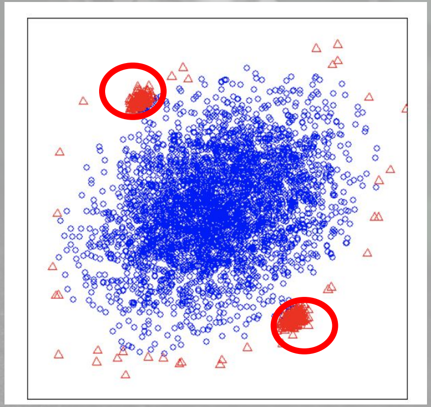
Swamping

Swamping refers to wrongly identifying normal instances as anomalies. This can happen when normal instances are too close to anomalies.



Masking

Masking is the existence of too many anomalies concealing their own presence. This can happen when an anomaly cluster is large and dense



Applications.

Banking

Security

Telecommunication

Computer vision

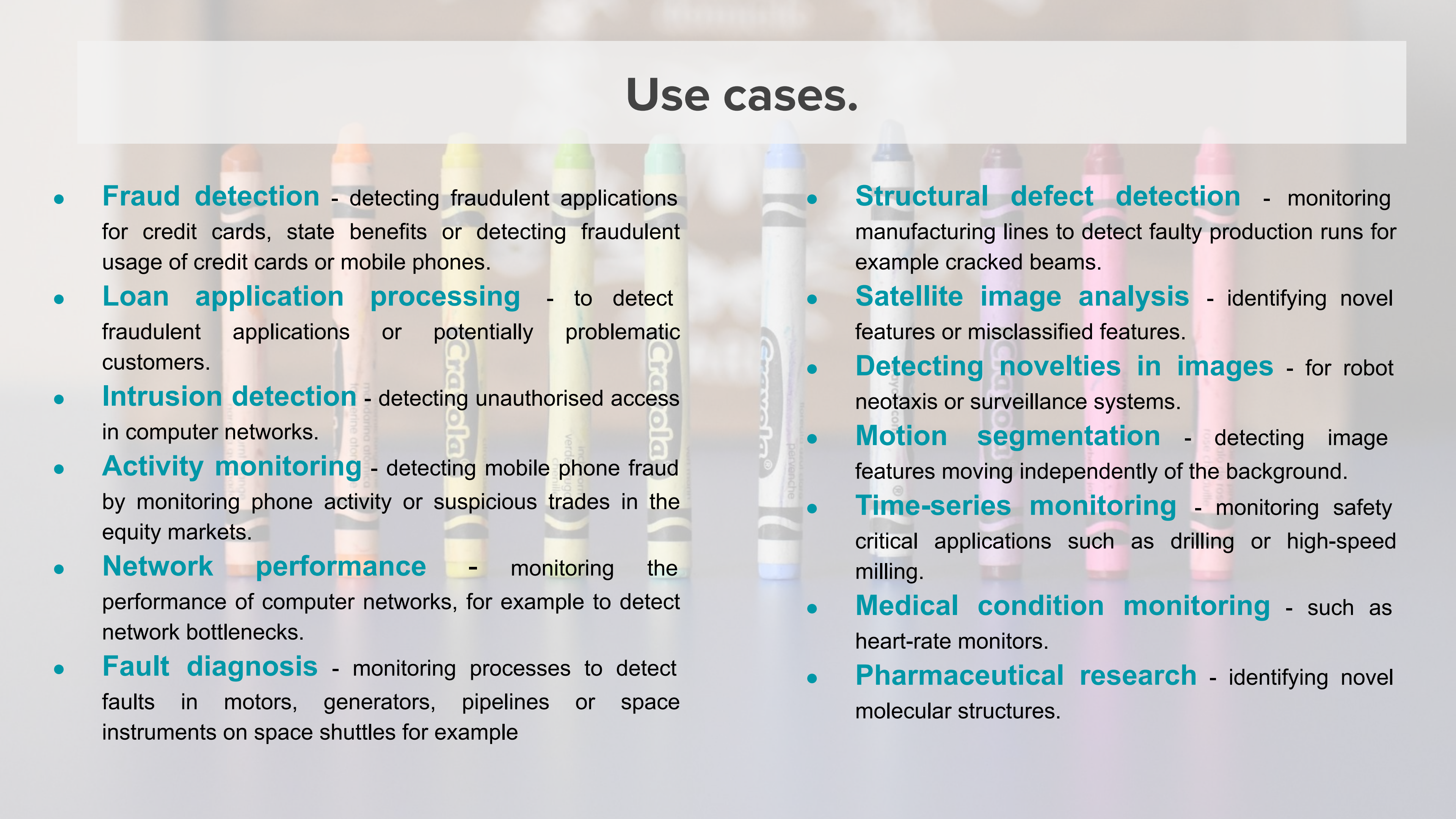
Retail

Insurance

Medicine

Molecular biology

...



Use cases.

- **Fraud detection** - detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones.
- **Loan application processing** - to detect fraudulent applications or potentially problematic customers.
- **Intrusion detection** - detecting unauthorised access in computer networks.
- **Activity monitoring** - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- **Network performance** - monitoring the performance of computer networks, for example to detect network bottlenecks.
- **Fault diagnosis** - monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for example
- **Structural defect detection** - monitoring manufacturing lines to detect faulty production runs for example cracked beams.
- **Satellite image analysis** - identifying novel features or misclassified features.
- **Detecting novelties in images** - for robot neotaxis or surveillance systems.
- **Motion segmentation** - detecting image features moving independently of the background.
- **Time-series monitoring** - monitoring safety critical applications such as drilling or high-speed milling.
- **Medical condition monitoring** - such as heart-rate monitors.
- **Pharmaceutical research** - identifying novel molecular structures.

THINGS SOLVER

ENLIGHTEN YOUR DATA

November, 2019
www.thingsolver.com

Valentina Djordjevic
valentina@thingsolver.com