ITSA 2023 - ANOMALY DETECTION VIA ML - POTENTIAL FOR IT SECURITY

Computacenter

Splunk Data Scientist: Moussa El-Zaarah on date: 11th of October 2023

YOUR DATA SCIENCE CONSULTANT TODAY

MY JOURNEY TO COMPUTACENTER

- **Studies: Theoretical Physics:**
  Statistical Physics & Nonlinear Dynamics
  (= Stochastic Processes & Chaos theory)

- WS 2018:
  Big Data & Data Science
  Enthusiast/Evangelist

CoE: Data Science Lead 2021
Newcomer of the year 2020
Future Talent program 2019

Physics 2018

Humboldt-Universität zu Berlin

DESY

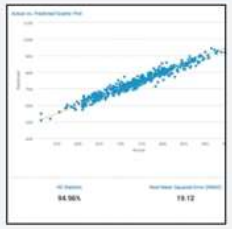
PHYSIK.BEGREIFEN
Schülerlabor in Zeuthen

Further Education:
Bioinformatics & Biostatistics

moussa.el-zaarah
@computacenter.com
+49 1736270711

MACHINE LEARNING METHODS IN SPLUNK WITH SHOWCASES

INCLUDES ALSO HYPOTHESIS-TESTING (NOT SHOWN)

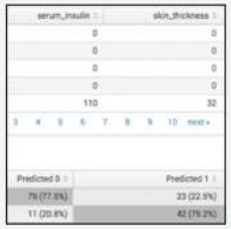


Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

Examples

- Predict Server Power Consumption
- Predict VPN Usage
- Predict Median House Value
- Predict Power Plant Energy Output
- Predict Future Logins
- Predict Future VPN Usage (sinusoidal time)
- Predict Future VPN Usage (categorical time)



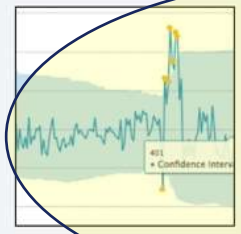
Answer binary Questions or Multiple Choice

Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

Examples

- Predict Hard Drive Failure
- Predict the Presence of Malware
- Predict Telecom Customer Churn
- Predict the Presence of Diabetes
- Predict Vehicle Make and Model
- Predict External Anomalies



Detect Numeric Outliers

Find values that differ significantly from previous values.

Examples

- Detect Outliers in Server Response Time
- Detect Outliers in Number of Logins (vs Predicted Value)
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Power Plant Humidity
- Detect Cyclical Outliers in Call Center Data
- Detect Cyclical Outliers in Logins

Probability Aspects

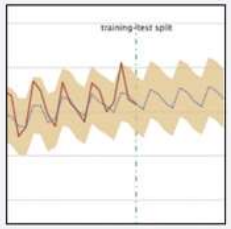
Detect Categorical Outliers

Find events that contain unusual combinations of values.

Examples

- Detect Outliers in Disk Failures
- Detect Outliers in Bitcoin Transactions
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Mortgage Contracts
- Detect Outliers in Diabetes Patient Records
- Detect Outliers in Mobile Phone Activity

Probability Aspects

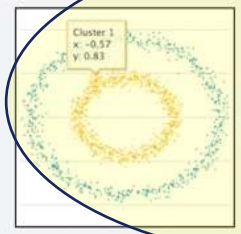


Forecast Time Series

Forecast future values given past values of a metric (numeric time series).

Examples

- Forecast Internet Traffic
- Forecast the Number of Employee Logins
- Forecast Monthly Sales
- Forecast the Number of Bluetooth Devices
- Forecast Exchange Rate TWI using ARIMA



Cluster Numeric Events

Partition events with multiple numeric fields into clusters.

Examples

- Cluster Hard Drives by SMART Metrics
- Cluster Behavior by App Usage
- Cluster Neighborhoods by Properties
- Cluster Vehicles by Onboard Metrics
- Cluster Power Plant Operating Regimes
- Cluster Business Anomalies to Reduce Noise

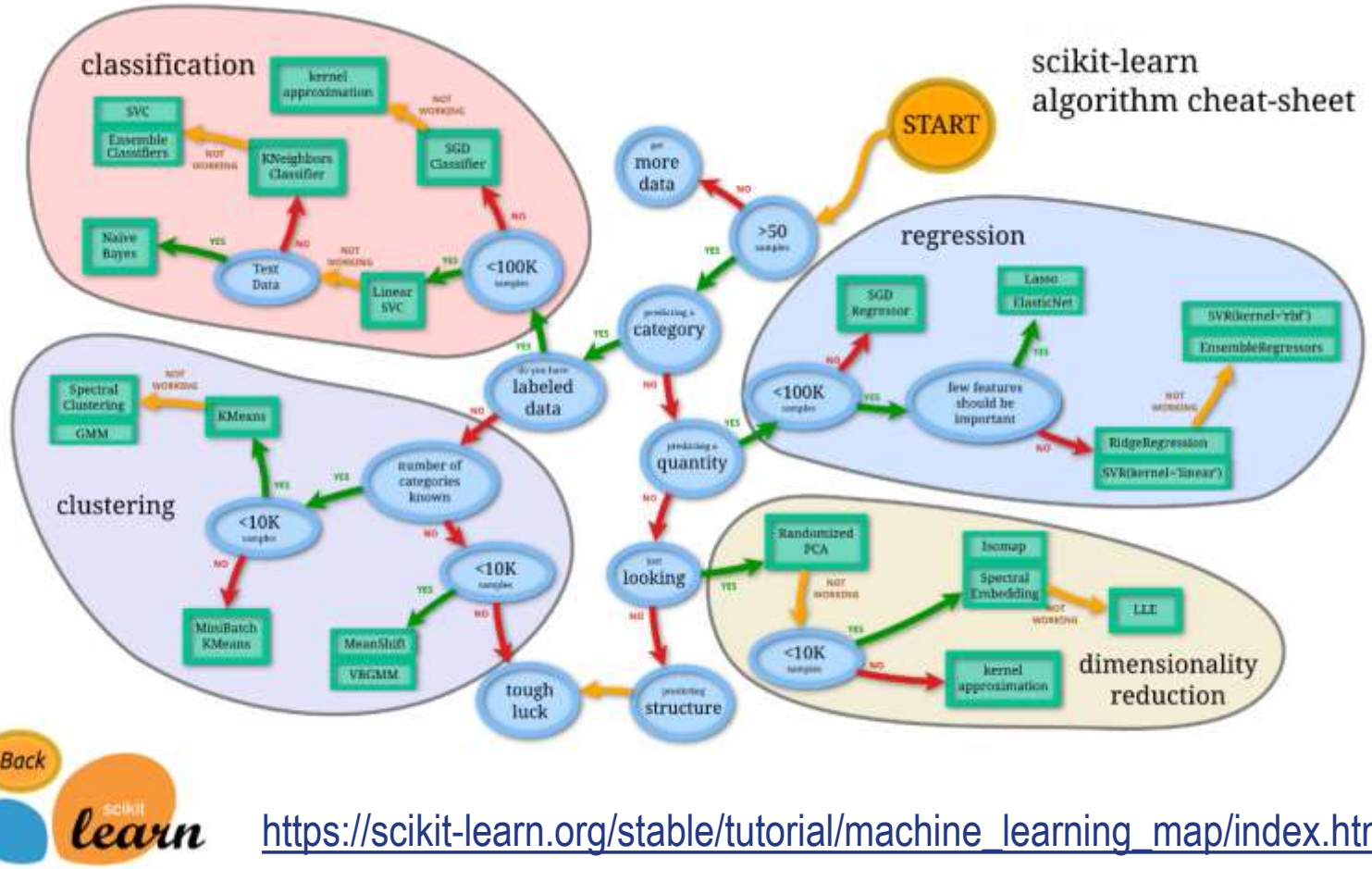
Automatically Group Data

MACHINE LEARNING TOOLKIT

splunk> Platform for Operational Intelligence

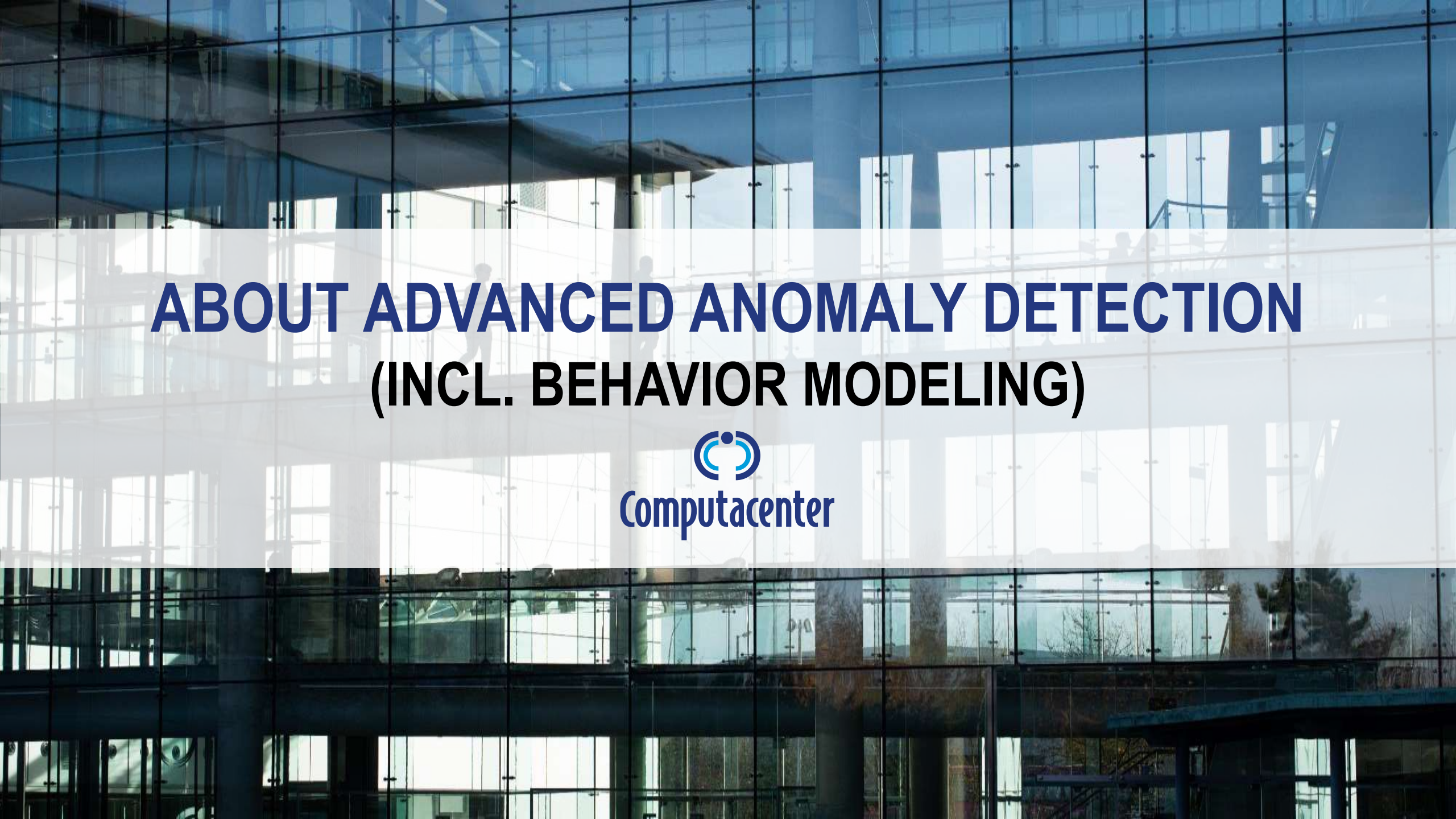
WHAT ALGORITHM TO CHOOSE FROM ML-LIBRARIES?

WE WILL HELP YOU NAVIGATE THROUGH THE DATA SCIENCE JUNGLE



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html#

Needs (at least) one data scientist to navigate towards customer goal!



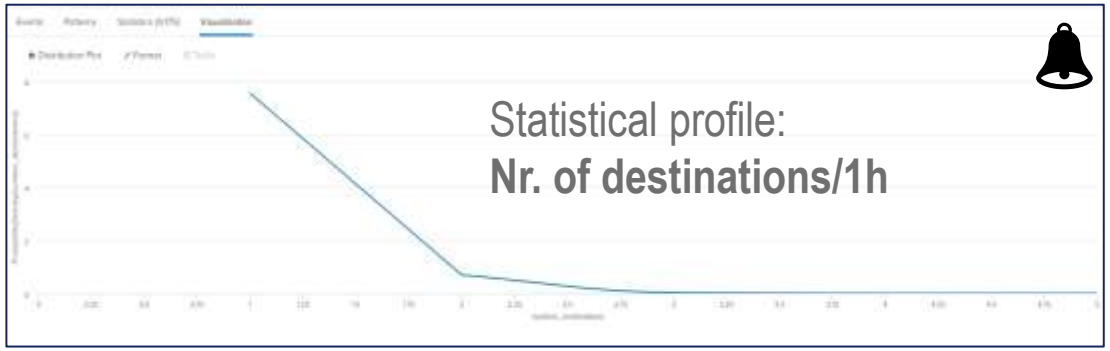
ABOUT ADVANCED ANOMALY DETECTION

(INCL. BEHAVIOR MODELING)

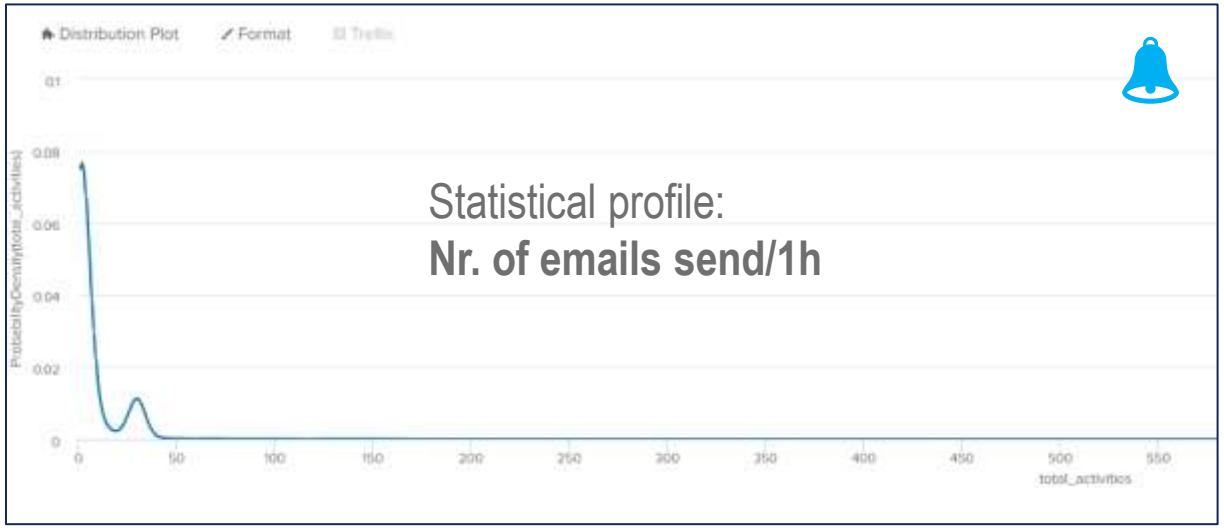
Computacenter

MULTIVARIATE BEHAVIOR BASELINES: EMAIL TRAFFIC BEHAVIOR

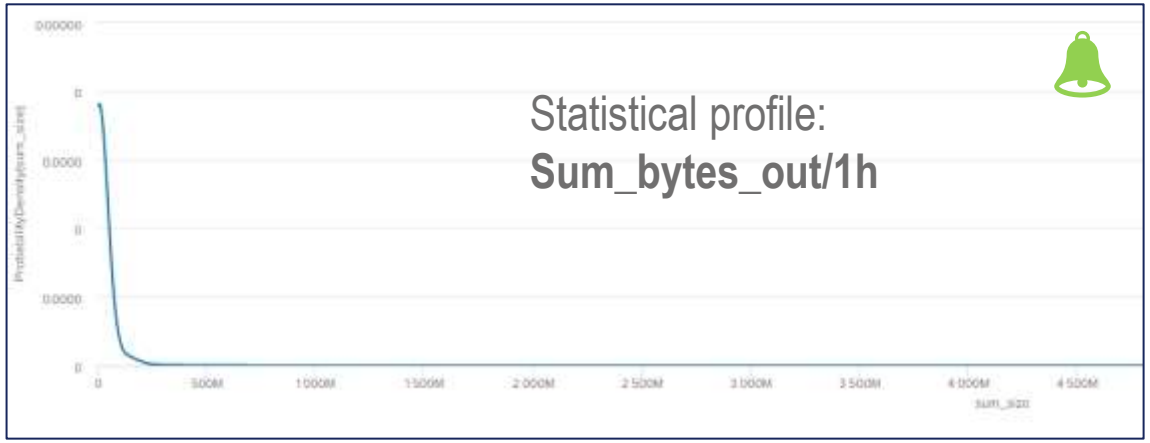
RANDOM DISTRIBUTION → AUTOMATED STATS MODEL VIA GAUSSIAN_KDE



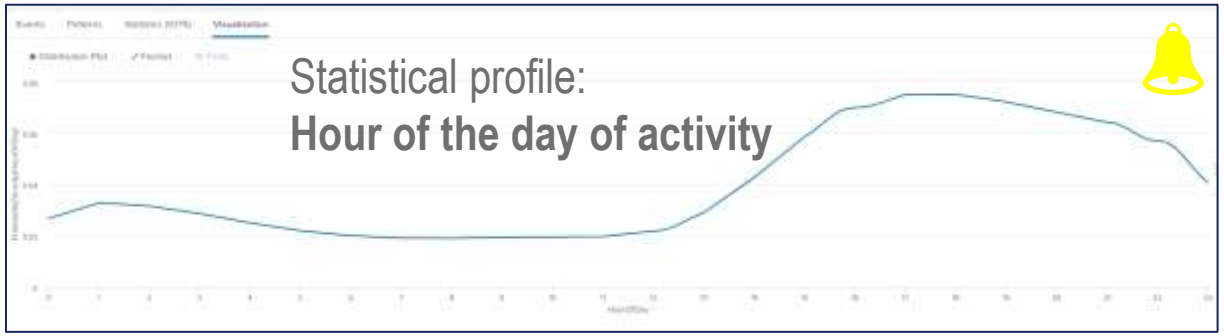
Statistical profile:
Nr. of destinations/1h



Statistical profile:
Nr. of emails send/1h



Statistical profile:
Sum_bytes_out/1h



Statistical profile:
Hour of the day of activity

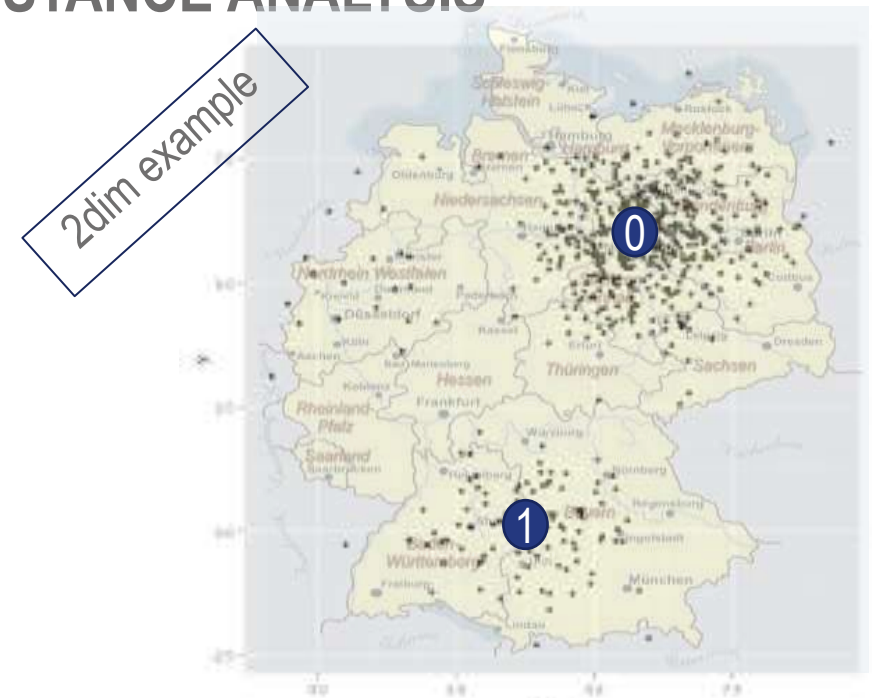
Multivariate alert triggers if statistical anomaly for the combination of aspects exist!

UNSUPERVISED ML: GEOGRAPHICALLY IMPROPABLE ACCESS

REAL SPACE CLUSTER & DISTANCE ANALYSIS

PART I:

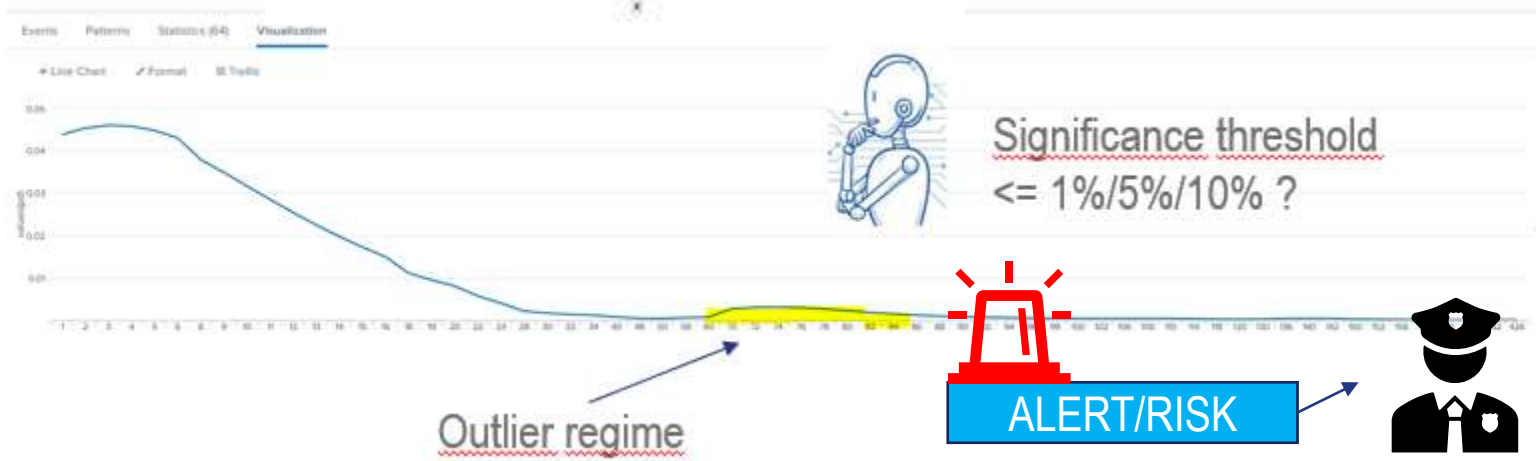
Specify nr. of clusters to find for KMeans -algorithm to find **within N numerical dimens./attributes**



- Use Cluster algorithms to **group your (un)labeled data** „spatially“ and detect centers.
- Each data point maps to a **login attempts ip-location** per identity/asset and gets assigned to a group

PART II:

Extract **cluster_distance statitsics model** and apply für new incoming measurements to find outliers or assign risk!



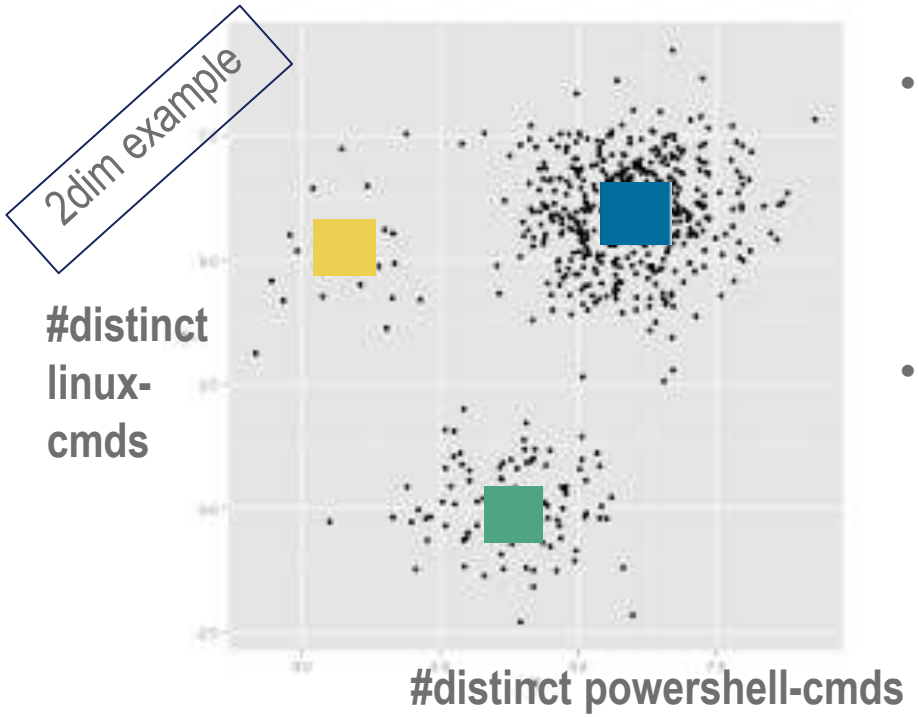
ANOMALIOUS GROUP MEMBERSHIP OVER TIME DETECTED

XMEANS DETECTS NR. OF CLUSTERS & SPECTRUM-CHANGES AUTOMATICALLY

PART I:

Use unsupervised learning via XMeans -algorithm to idenify hourly cluster membership **within N dimens./attributes**

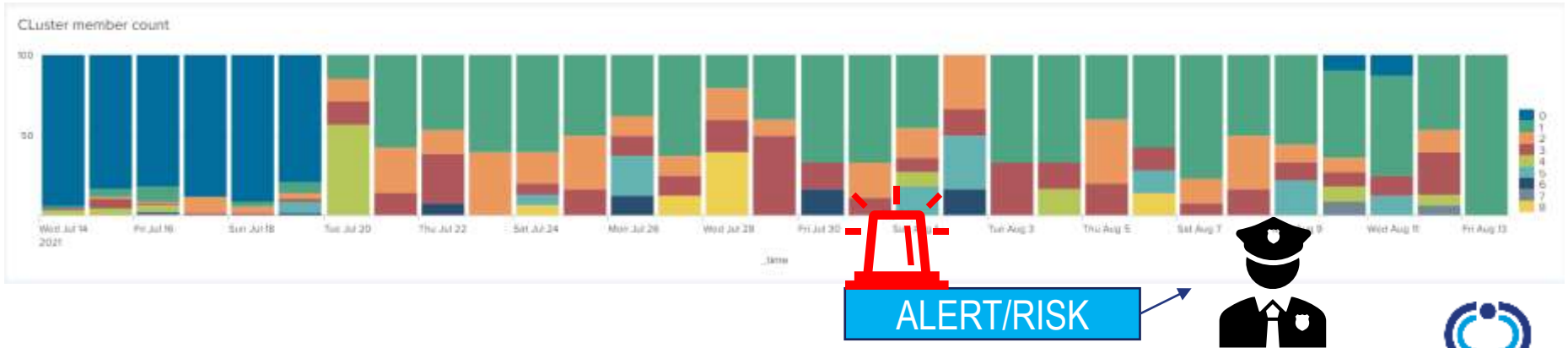
,



- Each **coordinate axis maps to the a behavior modeling attribute** (feature engineering).
- Each **data point describes an entities hourly behavior** & belongs to an (un)labeled data group (cluster)

PART II:

Monitor guessed cluster membership over the day and detect changes for alerting or assigning risk!

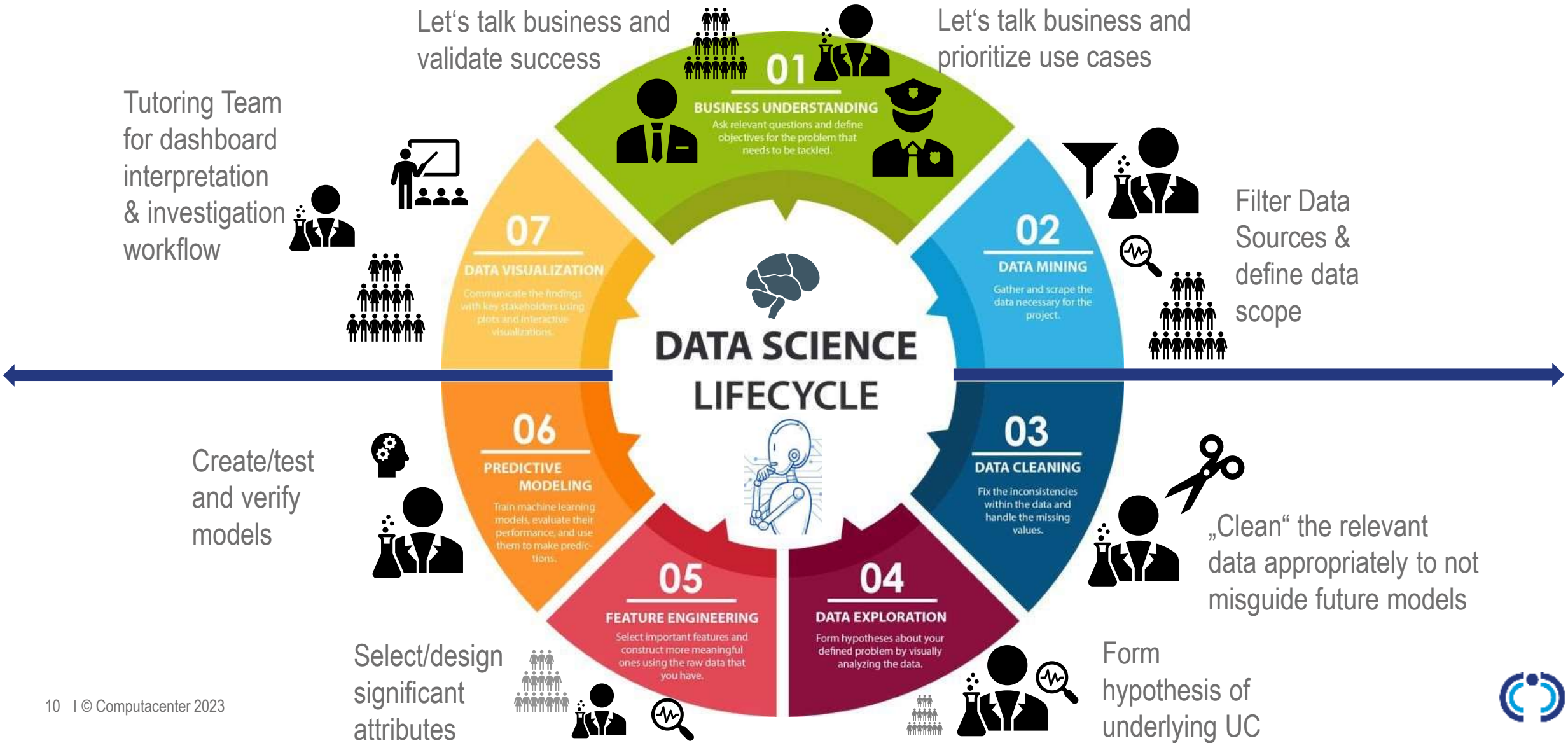


ML WORKBENCH VS OOTB SOLUTION OR BETTER HYBRID?

Computacenter

DATA SCIENCE LIFECYCLE PER USE CASE DEVELOPMENT

THIS IS AN ITERATIVE AND AGILE PROCESS TO GAIN A SUPPORTING ML/AI-AGENT



THANK YOU FOR YOUR ATTENTION!

Halle 7A
Stand 214

Computacenter

Halle 7A
Stand 214