



Providing assurance with "Data Driven Assurance"

Audience: KNIME Talks 2021

Thanks for having me...

Who am I?



Michiel Krol

Head of Audit Data Excellence bij
Rabobank Groep



Rabobank



Rabobank



Digital world drives need for NextGen Audit



Rabobank



The digital world is changing rapidly and is becoming more and more data driven



This also has a profound impact the Audit profession, on Audit services and on auditors



EFFICIENT

-10/20% HOURS

EFFECTIVE

ASSURANCE LEVELS UP IN
TWO IN THREE AUDITS

COMPLIANT

DATASETS READILY
AVAILABLE FOR Q&A

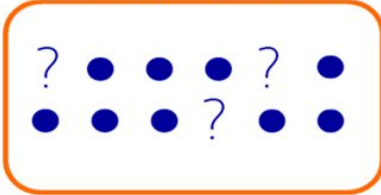
Audit Rabobank
has set three
ambitious data
driven targets

DDA is 90% smart business intelligence and 10% advanced analytics applied in the audit cycle

How often? How much?
Descriptives & Dashboarding



Is this complete and correct?
Data quality profiling & Reconciliation



How does this compare?
Benchmarking



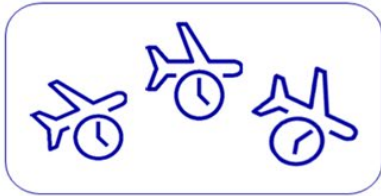
Is this logical?
Business rules & Smart Stitching



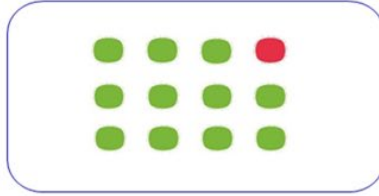
How does this change?
Trend analysis



What steps, who, how long?
Process mining



Is this weird?
Anomaly detection



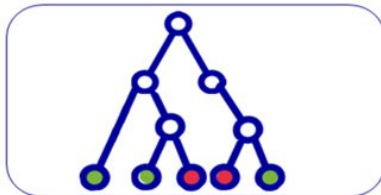
What do the words reveal?
Text analysis



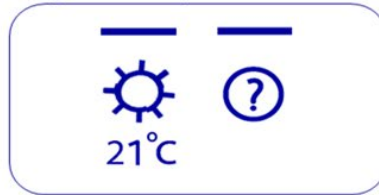
How is this organized? (1)
Clustering analysis



How is this organized? (2)
Decision tree



How much? How many?
Regression analysis

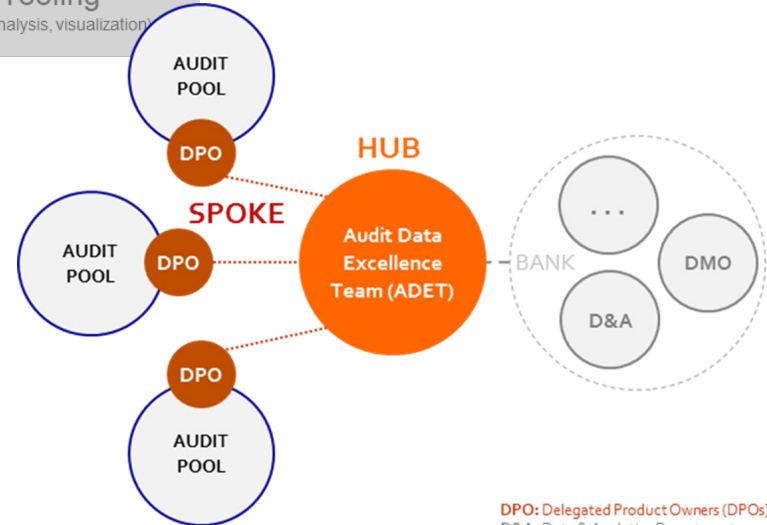


Is this A or B?
Classification

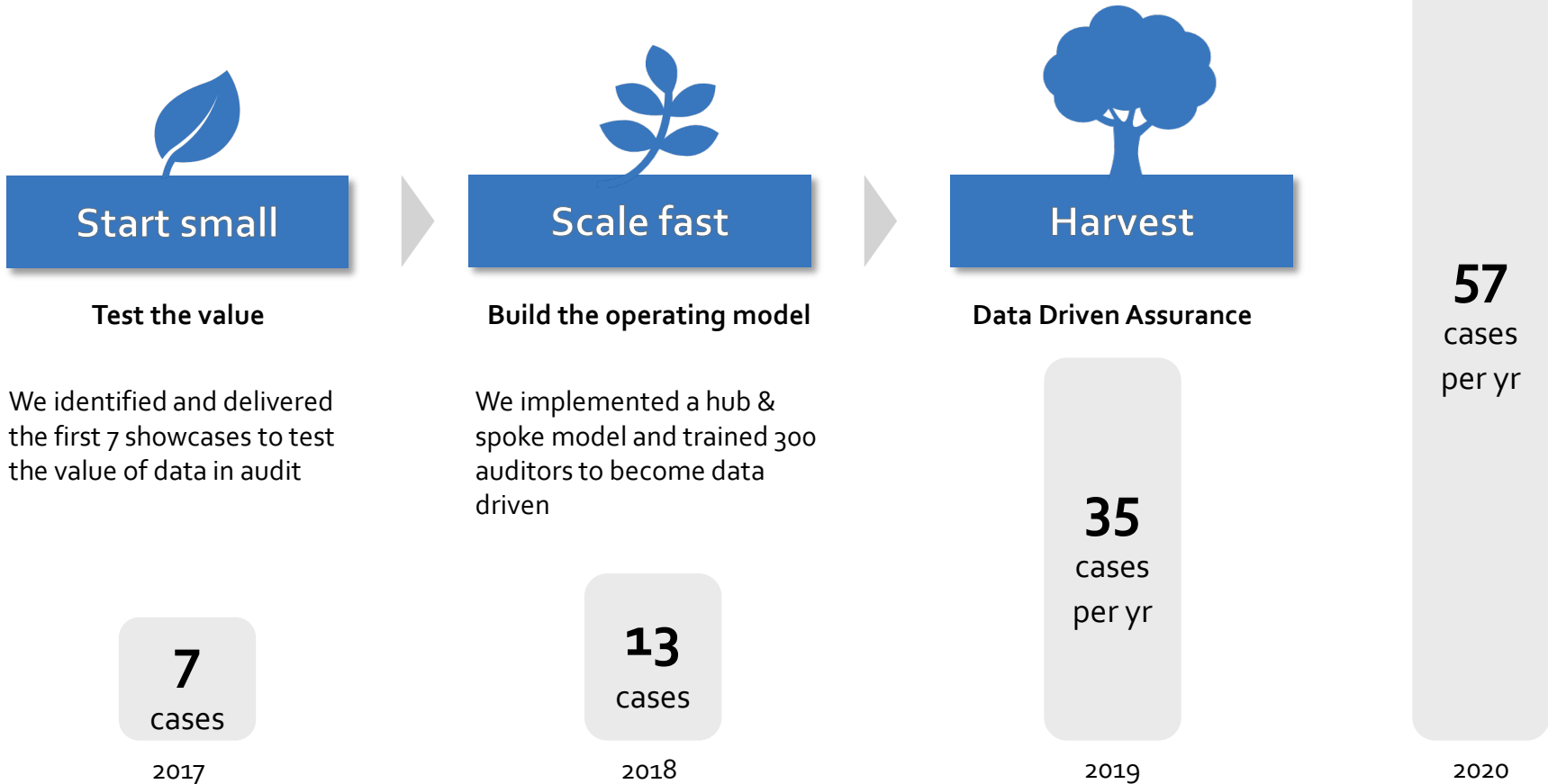


It is a Target Operating Model rather than only a team of specialists

ORGANISATION	PEOPLE	DATA	TECHNOLOGY
Operating Model	Staffing	Data Acquisition	Architecture <i>(e.g. Data Lake, Data Lab, Data Factory)</i>
Innovation Cycle	Learning & Development	Data Governance & Operations	Tooling <i>(e.g. analysis, visualization)</i>
Ways of working	Data Driven Culture		



Start small... scale fast



No change without challenges



LEARNINGS:

- Tone at the top, strategic change
- Auditor analysts drive scale up
- Carrot and the stick
- Integrated in Audit Procedures
- Benefit realization
- Data approval process

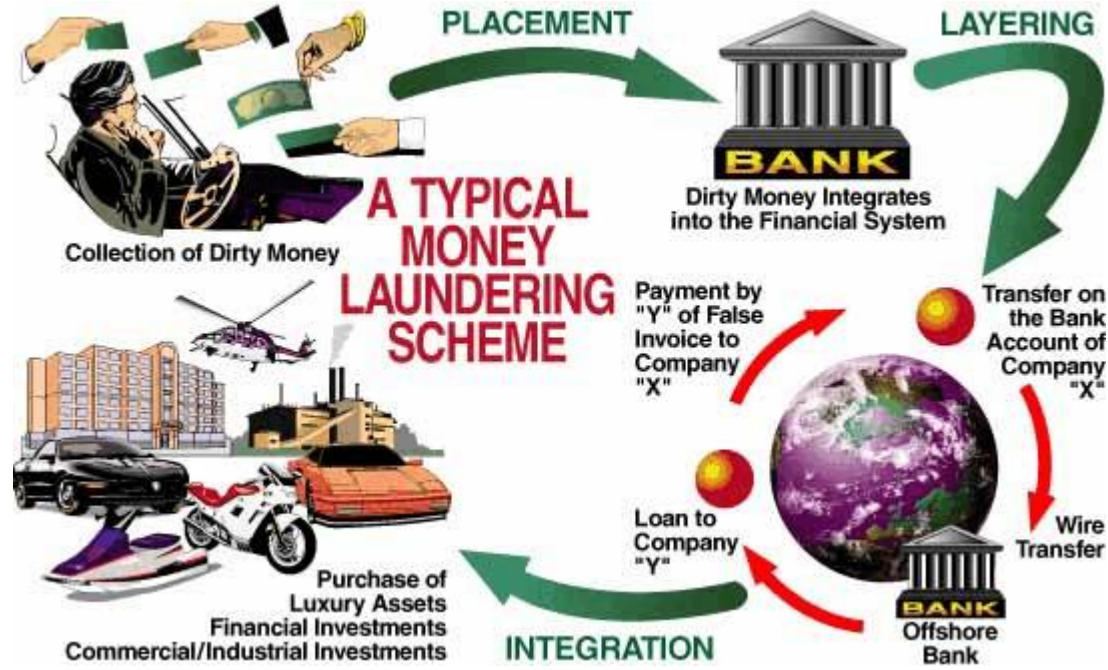
CHALLENGES:

- Conservatives
- Top-down hypotheses
- Next maturity Lab environment
- More positive assurance

USE CASE

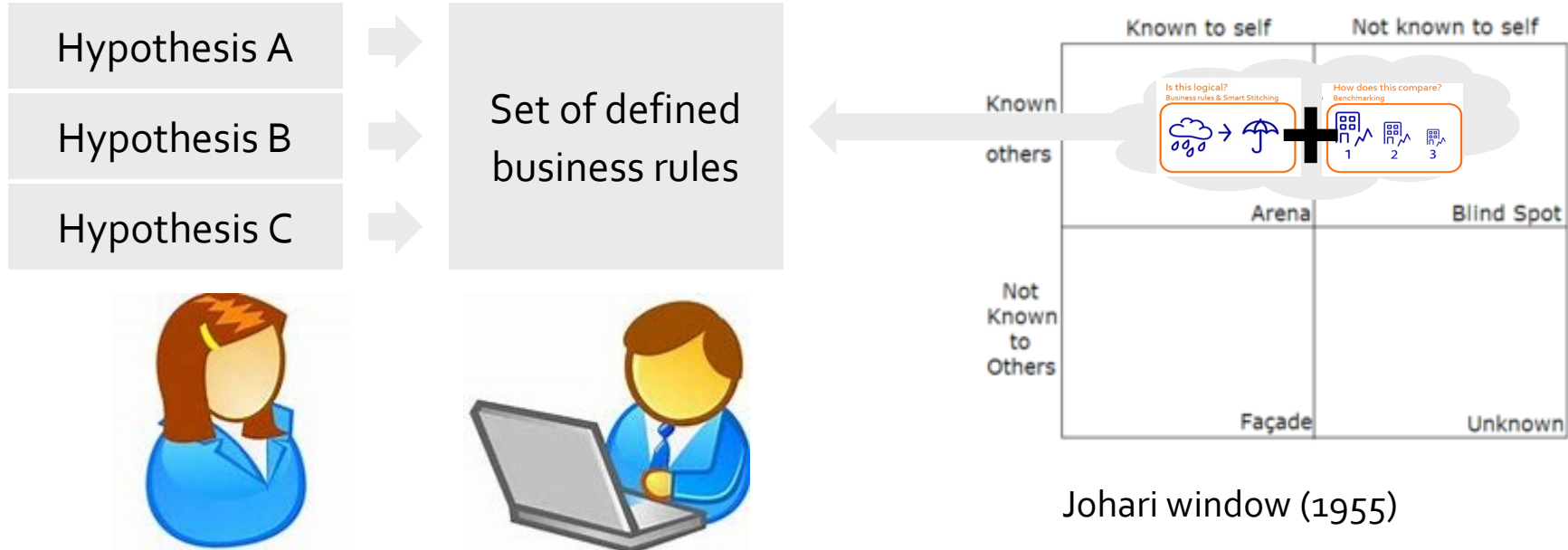
Anomaly Detection: Unknown Patterns in Anti-Money Laundering

What is AML?

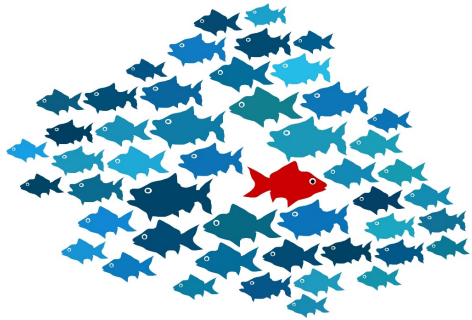


Source: UNODC

Known versus unknown patterns



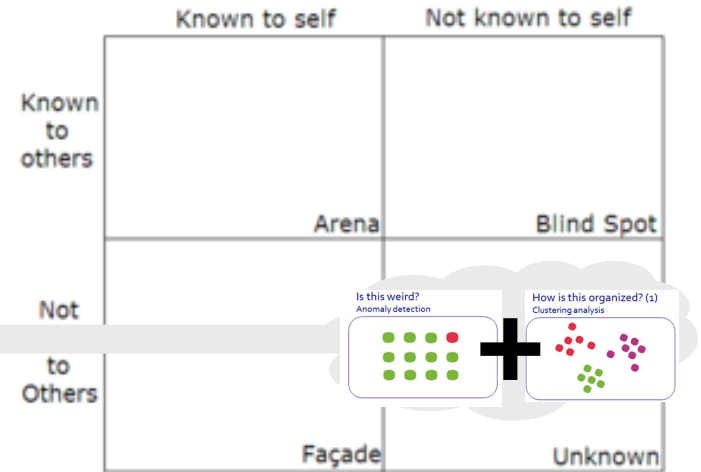
Known versus unknown patterns



What are the clients that show different behavior than others?



What are clusters of clients that show similar behavior?



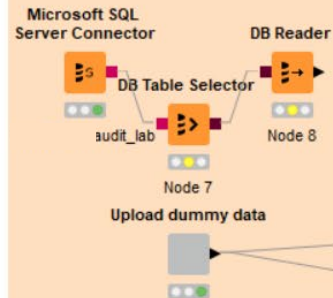
Johari window (1955)

KNIME for non-techies



Rabobank

DATA ACCESS,
Select an input file.
Prepare your data before using this workflow
Data preparation, e.g., cleaning the data, is not part of this workflow.



INSPECT AND EXCLUDE COLUMNS
Inspect the nominal and categorical columns.
Choose what columns to exclude in further analyses.
- Columns with many missing values are likely to be less valuable for isolation forest
- Columns with many categories are deemed less valuable for isolation forest. Try to group categories together, e.g., group countries in logical categories.
- Excluding columns from large datasets takes long. Alternatively, filter by using a columns filter before applying this component, or use smaller dataset as input.

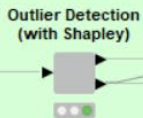


ANOMALY DETECTION
Runs the anomaly detector. Each case is assigned an anomaly score between 0 and 1.
The H2O anomaly detector is applied: this is an isolation forest method that does not need python to be installed.

Optional: Adapt the number of models and/or the number of levels (Tree depth) that the anomaly detector uses.

The Number of models parameter allows you to configure the number of decision trees composing the ensemble of trees (hence, the forest size). This must be a number up to 256; Higher numbers tend to give better results although they take longer to process. By default the number of models is 100.

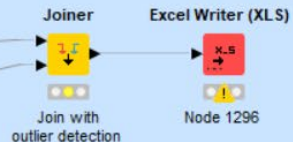
The Number of levels parameter allows you to configure the number of levels of a tree (hence, the tree depth). By default the number of models is 10.



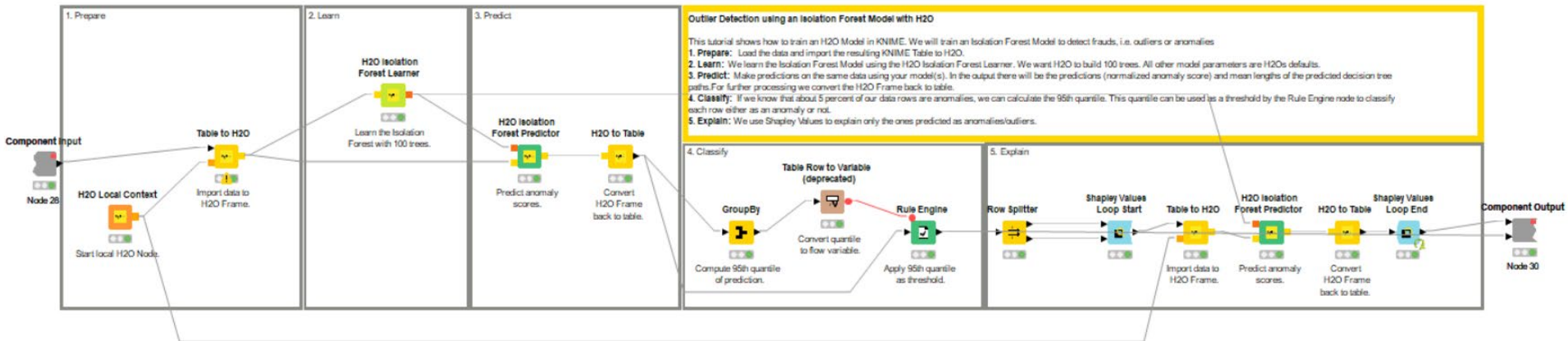
VISUALIZATION
Calculation of the feature importance for the anomaly scores of each case/row.
To inspect all relevant features of an anomalous row, select this row, click 'close and apply' and run 'visualize single entry'.



OUTPUT
Download the results including the anomaly score per case and feature importances.
- Do you want to inspect a specific anomaly (hence, row)? You could use a rule-based row filter and select a row with rownumber/id of original file. Example statement: \$\$ROWID\$\$ = "Row7778" => TRUE.



Under-the-hood...



The art of anomaly detection



Points to take into account

- Start small (<n variables)
- Understand the detected anomalies
- Garbage in, garbage out! Quality of the input...
- Very very important which features you import in the model
- You can have groups with the same anomaly scores → identify clusters
- Spot data quality issues

Thanks for listening, be involved!

