Systemy uczące (się) – kto jest uczniem, kto nauczycielem?

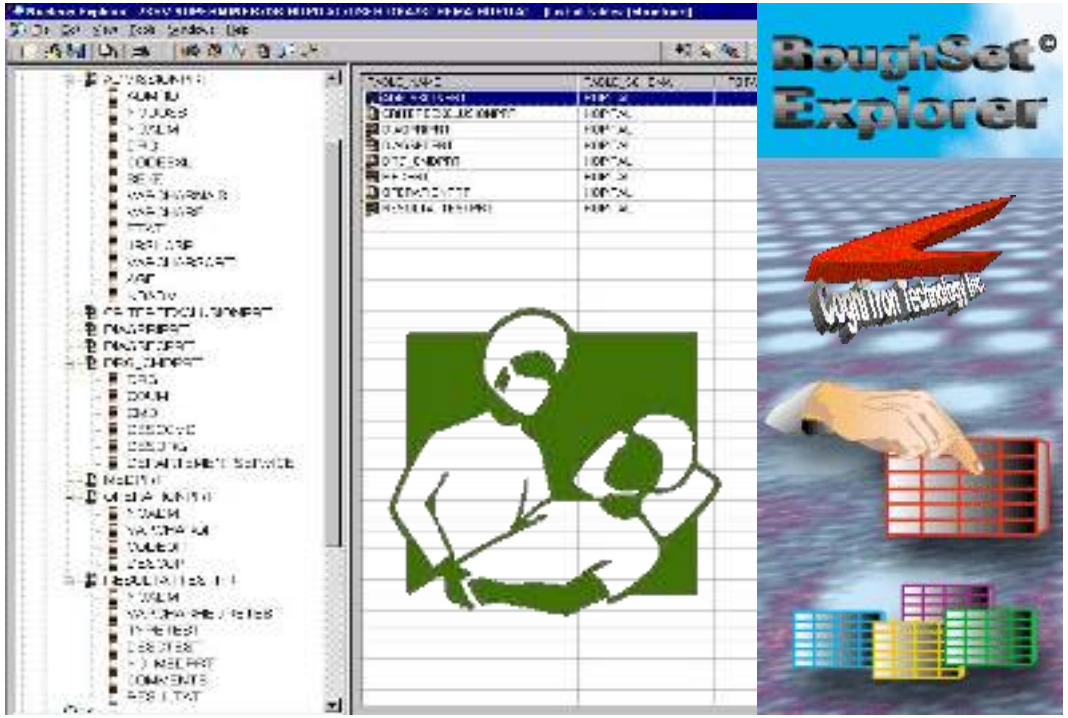
Dominik Ślęzak

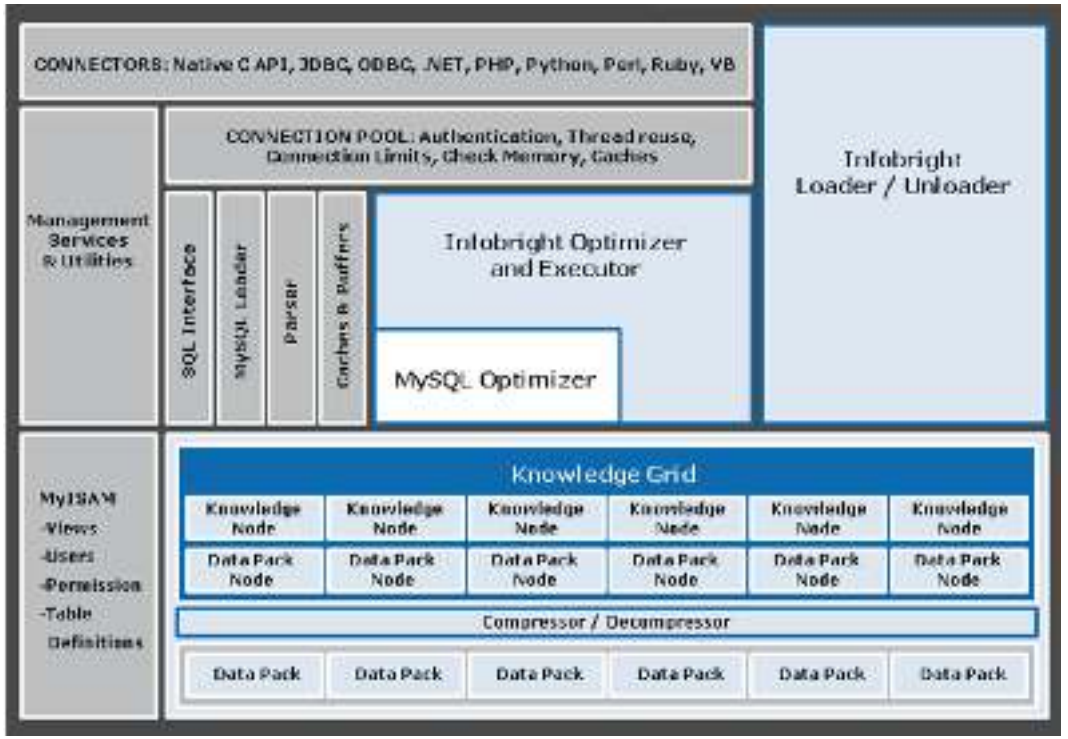
Wprowadzenie

- Obecne metody uczenia się z danych (ang. machine learning; popularny skrót myślowy to metody „AI/ML”) są bardzo skuteczne, lecz nie można ich pozostawiać samym sobie. Potrzebują zbiorów treningowych, sugestii co do priorytetów w procesie uczenia, weryfikacji, czy tworzone modele są zgodne ze zdrowym rozsądkiem. Z drugiej strony, możemy się od nich oczywiście wiele nauczyć.
- Poruszamy wybrane aspekty interakcji na linii „systemy AI/ML – ich użytkownicy”. Omawiamy, w jakim sensie działanie metod AI/ML może pogłębić naszą wiedzę o danych, a z drugiej strony – jak wykorzystywać naszą wiedzę do optymalizacji działania tych metod. Poruszamy kwestie ergonomii procesów wymiany informacji, w tym konieczność zapewnienia prędkości i intuicyjności obliczeń.
- Koncentrujemy się w szczególności na paru podejściach wywodzących się z teorii zbiorów przybliżonych i obliczeń granularnych, gdzie dużą wagę przywiązuje się do uproszczonych – a przez to zrozumiałych i efektywnych – reprezentacji wiedzy, danych, a także procesów uczenia się z danych.

Mój „background”

- Data Analytics 1995-2005
- Data Processing 2005-2015
- *Data Science* 2015-…





INFOBRIGHT

Information Builders

CISCO

SOD SECURITY ON-DEMAND

Label in the Loop

BRIGHTBOX

SMALL BIG DATA

CHALLENGES

QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

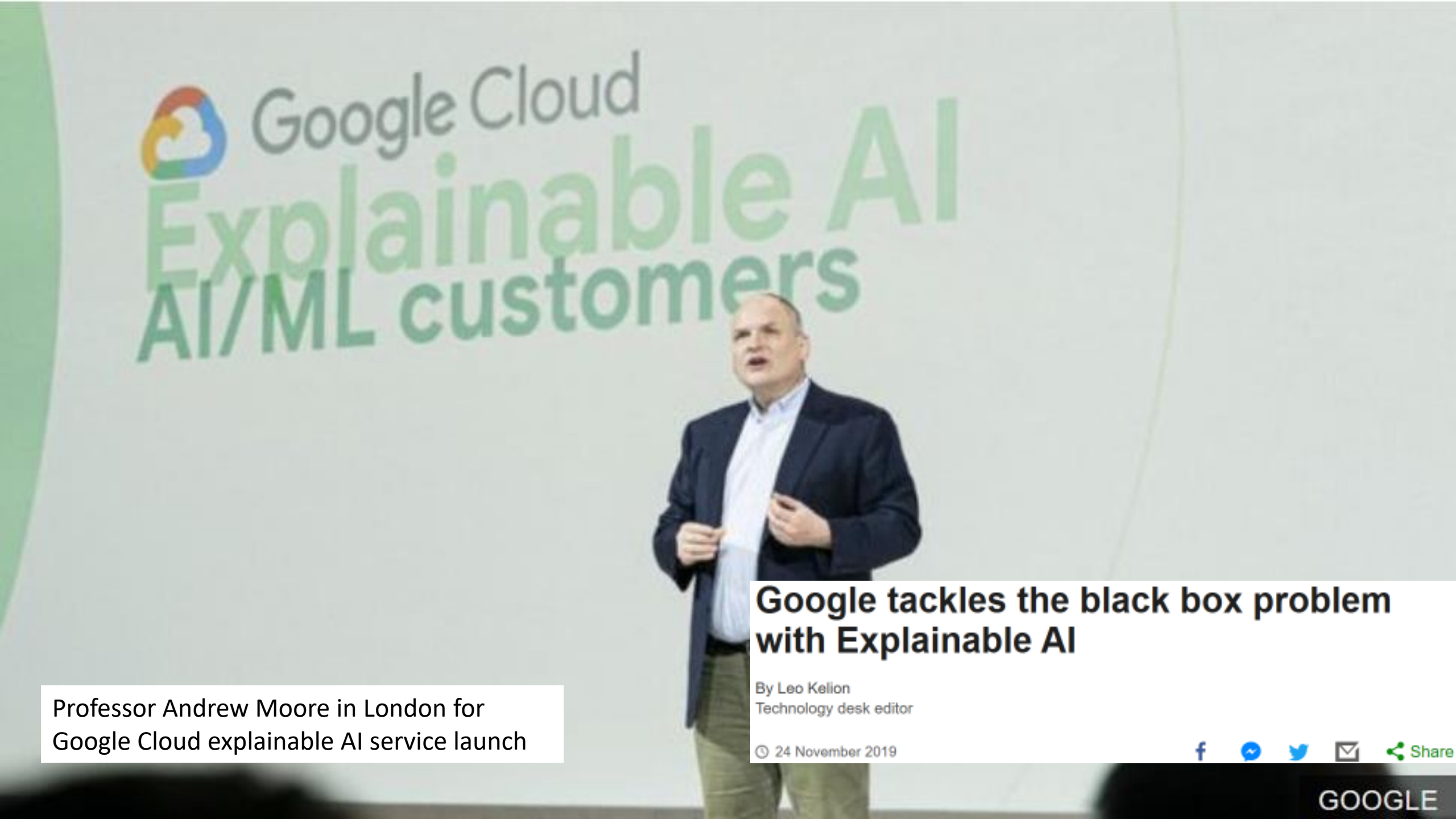
Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Professor Andrew Moore in London for Google Cloud explainable AI service launch

Google tackles the black box problem with Explainable AI

By Leo Kelion
Technology desk editor

24 November 2019

Share

There is still a lot to be done…

- Explaining to humans why AI/ML models… are not certain
- Explaining to humans why AI/ML models… make mistakes
- Explaining to AI/ML models what humans want from them

Explaining why AI/ML models make mistakes

(examples of so-called diagnostic rules)

- IF there were no similar objects in the training data set, THEN the mistake is most likely because the model is not ready for such cases
  - *(but maybe it doesn't need to?)*
- IF mistakes happen quite often for similar objects in the training data set, THEN the model is not sufficiently tuned for such cases
  - *(but maybe it doesn't need to?)*
- IF there was a single (or a few) similar object and there was no mistake, THEN maybe there was something wrong with that object?
  - *(maybe it was incorrectly labeled?)*

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the Documentation or learn about InClass competitions.

👋 New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 22801 Teams
Knowledge

All Competitions

Active | Completed | InClass

All Categories | Default Sort

OSIC Pulmonary Fibrosis Progression
Predict lung function decline
Featured • 3 months to go • Code Competition • 86 Teams
$55,000

SIIM-ISIC Melanoma Classification
Identify melanoma in lesion images
Featured • a month to go • 1886 Teams
$30,000

ALASKA2 Image Steganalysis
Detect secret data hidden within digital images
$25,000

kaggle

KnowledgePit
A Data Challenge Platform

Competitions Forum

There are all standard and public Competitions in this category. Any User can Take part in any of the Challenges within.

Ongoing Competitions | All Competitions | Student Competitions

IEEE BigData 2020 Cup: Predicting Escalations in Customer Support

Predicting Escalations in Customer Support is a data mining challenge organized in association with the IEEE BigData 2020 conference. The task is to predict which cases in Information Builders, Inc. (ibi) technical support ticketing system will be escalated in the nearest future by customers. The competition is organized jointly by ibi (https://www.ibi.com) and QED Software (http://www.qed.pl/).

7 weeks ago

Manager: Andrzej Janusz (andrzej) 255 teams

FedCSIS 2020 Challenge: Network Device Workload Prediction

FedCSIS 2020 Data Mining Challenge: Network Device Workload Prediction is the seventh data mining competition organized in association with Conference on Computer Science and Information Systems (https://fedcsis.org/). This time, the considered task is related to the monitoring of large IT infrastructures and the estimation of their resource allocation. The challenge is sponsored by FMCA Software and Polish Information Processing Society (PTI).

4 months ago

Manager: Andrzej Janusz (andrzej) 155 teams

IEEE BigData 2019 Cup: Suspicious Network Event Recognition

Suspicious Network Event Recognition is a data mining challenge organized in association with IEEE BigData 2019 conference. The task is to decide which alerts should be regarded as suspicious based on information extracted from network traffic logs. The competition is kindly sponsored by Security On-Demand (https://www.securityondemand.com/) and QED Software (http://qed.pl/).

1 year ago

Manager: Andrzej Janusz (andrzej) 290 teams

Log In

Login

Password

Log In

forgot password?

or Create an account

KnowledgePit
A Data Challenge Platform

Competitions Forum

There are all standard and public Competitions in this category. Any User can Take part in any of the Challenges within.

Ongoing Competitions | All Competitions | Student Competitions

IEEE BigData 2020 Cup: Predicting Escalations in Customer Support

Predicting Escalations in Customer Support is a data mining challenge organized in association with the IEEE BigData 2020 conference. The task is to predict which cases in Information Builders, Inc. (ibi) technical support ticketing system will be escalated in the nearest future by customers. The competition is organized jointly by ibi (https://www.ibi.com) and QED Software (http://www.qed.pl/).

Manager: Andrzej Janusz (andrzej)

255 teams

7 weeks ago

FedCSIS 2020 Challenge: Network Device Workload Prediction

FedCSIS 2020 Data Mining Challenge: Network Device Workload Prediction is the seventh data mining competition organized in association with Conference on Computer Science and Information Systems (https://fedcsis.org/). This time, the considered task is related to the monitoring of large IT infrastructures and the estimation of their resource allocation. The challenge is sponsored by FMCA Software and Polish Information Processing Society (PTI).

Manager: Andrzej Janusz (andrzej)

155 teams

4 months ago

IEEE BigData 2019 Cup: Suspicious Network Event Recognition

Suspicious Network Event Recognition is a data mining challenge organized in association with IEEE BigData 2019 conference. The task is to decide which alerts should be regarded as suspicious based on information extracted from network traffic logs. The competition is kindly sponsored by Security On-Demand (https://www.securityondemand.com/) and QED Software (http://qed.pl/).

Manager: Andrzej Janusz (andrzej)

290 teams

1 year ago

Log In

Login

Password

Log In

forgot password?

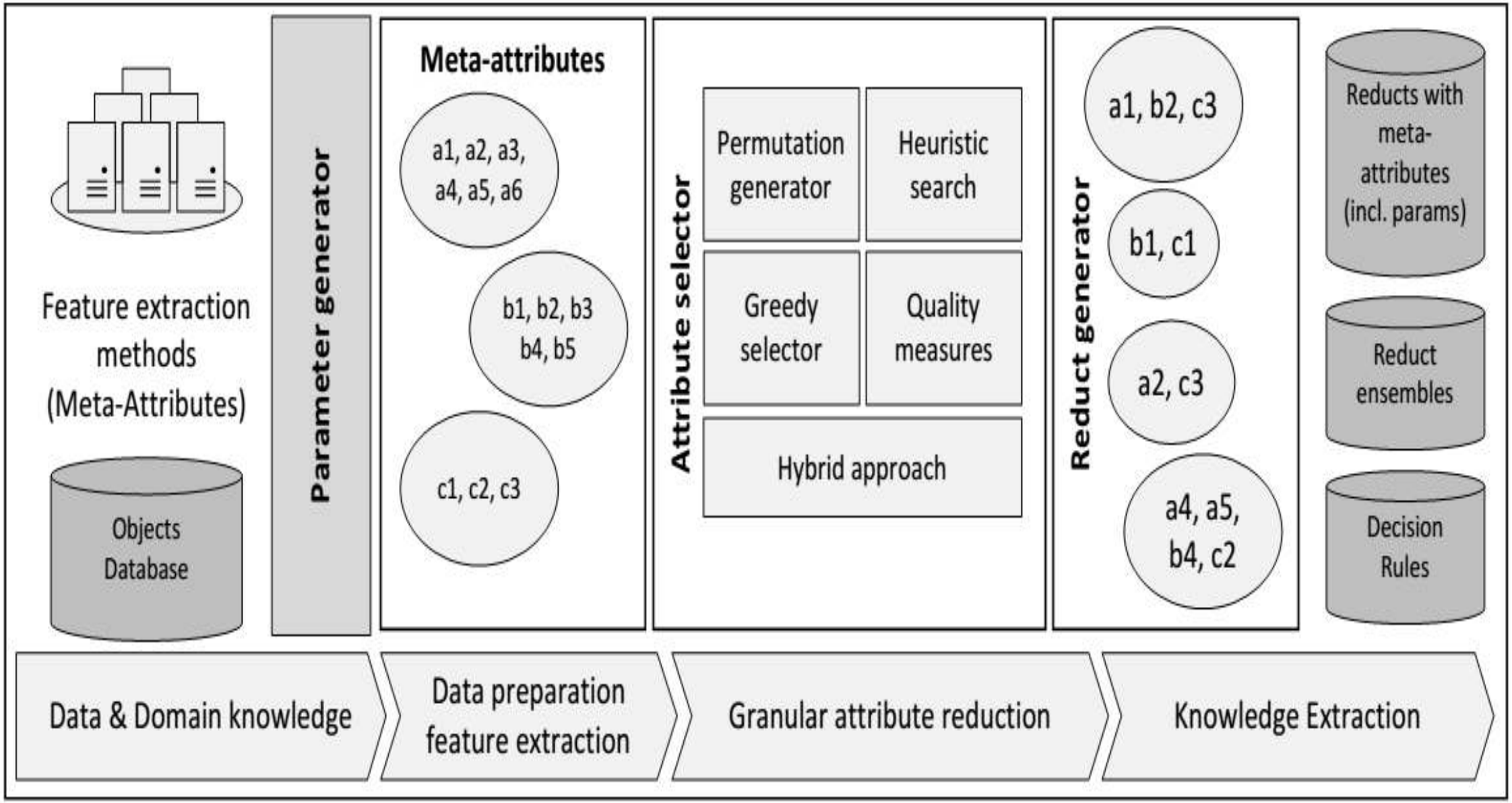
or Create an account

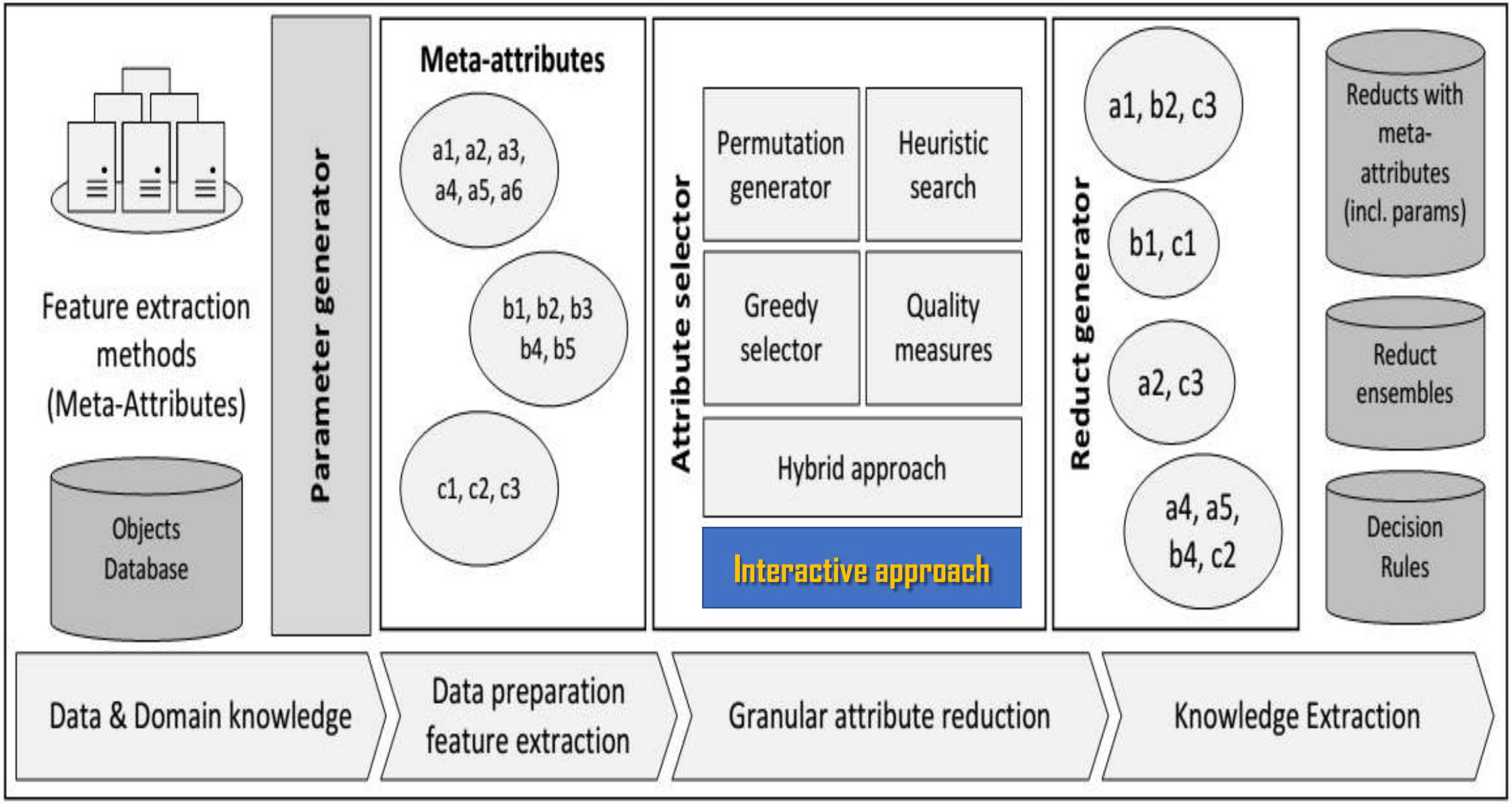
EXPLAIN

But how to select „similar objects”?

(one of possible answers: using ensembles of rough-set reducts)

- IF there were no similar objects in the training data set, THEN the mistake is most likely because the model is not ready for such cases
  - *(but maybe it doesn’t need to?)*
- IF mistakes happen quite often for similar objects in the training data set, THEN the model is not sufficiently tuned for such cases
  - *(but maybe it doesn’t need to?)*
- IF there was a single (or a few) similar object and there was no mistake, THEN maybe there was something wrong with that object?
  - *(maybe it was incorrectly labeled?)*


Feature extraction methods (Meta-Attributes)
Objects Database
Parameter generator
Meta-attributes
a1, a2, a3, a4, a5, a6
b1, b2, b3 b4, b5
c1, c2, c3
Attribute selector
Permutation generator
Heuristic search
Greedy selector
Quality measures
Hybrid approach
Reduct generator
a1, b2, c3
b1, c1
a2, c3
a4, a5, b4, c2
Reducts with meta-attributes (incl. params)
Reduct ensembles
Decision Rules
Data & Domain knowledge
Data preparation feature extraction
Granular attribute reduction
Knowledge Extraction


Feature extraction methods (Meta-Attributes)
Objects Database
Parameter generator
Meta-attributes
a1, a2, a3, a4, a5, a6
b1, b2, b3 b4, b5
c1, c2, c3
Attribute selector
Permutation generator
Heuristic search
Greedy selector
Quality measures
Hybrid approach
Interactive approach
Reduct generator
a1, b2, c3
b1, c1
a2, c3
a4, a5, b4, c2
Reducts with meta-attributes (incl. params)
Reduct ensembles
Decision Rules
Data & Domain knowledge
Data preparation feature extraction
Granular attribute reduction
Knowledge Extraction

Explaining why AI/ML models make mistakes

(models may make mistakes because of the data quality!)

- IF there were no similar objects in the training data set, THEN the mistake is most likely because the model is not ready for such cases
  - *(but maybe it doesn't need to?)*
- IF mistakes happen quite often for similar objects in the training data set, THEN the model is not sufficiently tuned for such cases
  - *(but maybe it doesn't need to?)*
- IF there was a single (or a few) similar object and there was no mistake, THEN maybe there was something wrong with that object?
  - *(maybe it was incorrectly labeled?)*

What if the data is not good enough?

- AI/ML algorithms require the training data
  - What if there are no (sufficient amount of) cases with appropriate labels?

- Could we ask „the crowd” for help?
  - Yes - unless… the labeling process requires highly specialized knowledge…

- **Active Learning**

  „is a special case of machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points.”

Label in the Loop

BRIGHTBOX

SMALL BIG DATA

CHALLENGES

QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

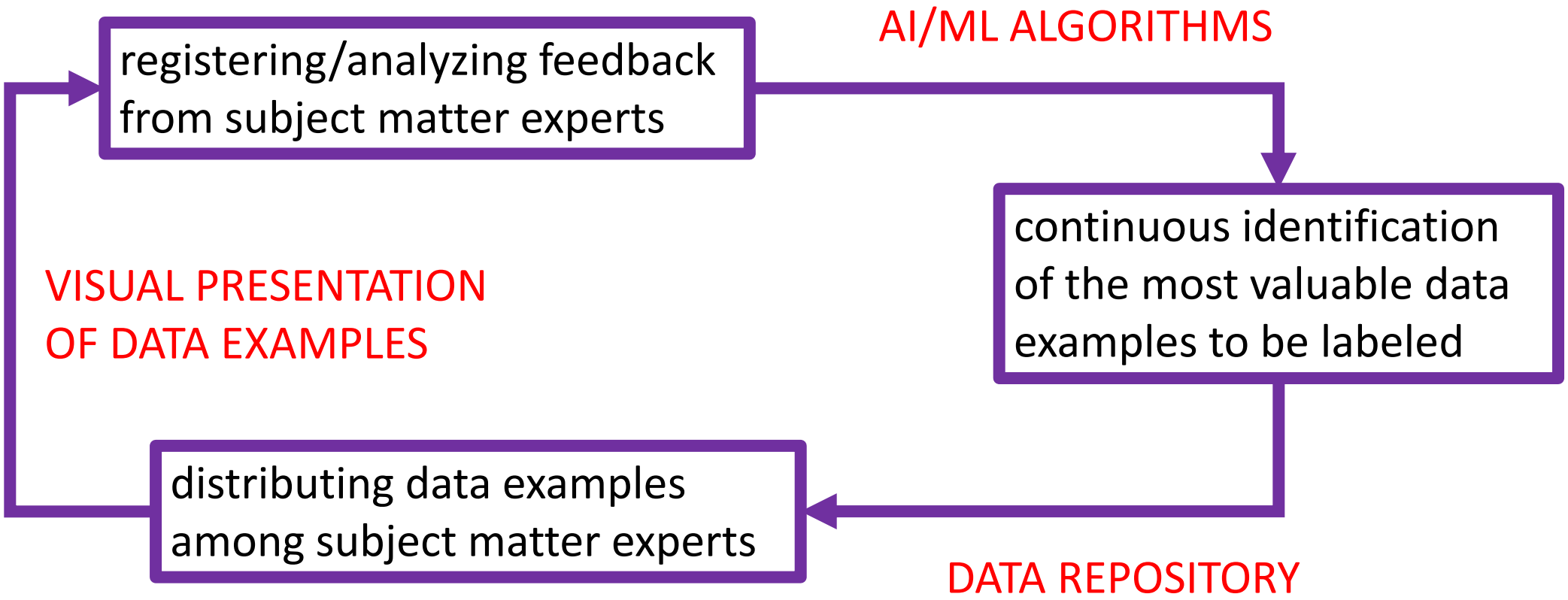
Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

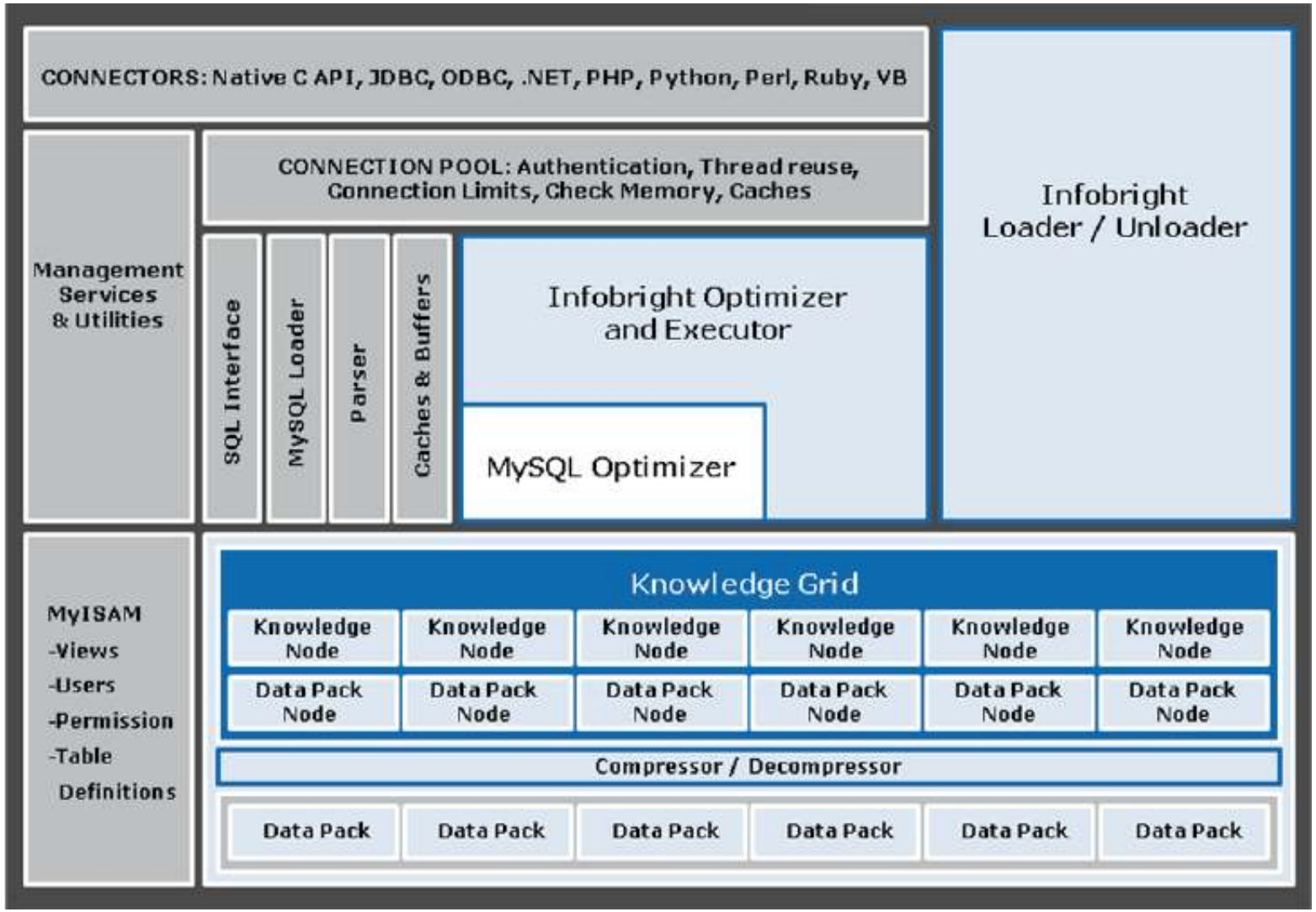
Enabling machine learning scalability for big data and/or big data flows

LITL (Label In The Loop)



But it needs to be fast!!

interactive data labeling
interactive feature selection
interactive …


CONNECTORS: Native C API, JDBC, ODBC, .NET, PHP, Python, Perl, Ruby, VB
CONNECTION POOL: Authentication, Thread reuse, Connection Limits, Check Memory, Caches
Management Services & Utilities
SQL Interface
MySQL Loader
Parser
Caches & Buffers
Infobright Optimizer and Executor
MySQL Optimizer
Infobright Loader / Unloader
MyISAM
-Views
-Users
-Permission
-Table Definitions
Knowledge Grid
Knowledge Node
Data Pack Node
Compressor / Decompressor
Data Pack

Database

Petabytes

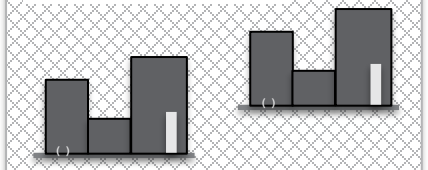
Traditional Query Execution:

- long time to do computations
- lots of disk/memory/processing resources required

- hard to manage in data lake / data cloud environments

Query Result

Summary Database



Gigabytes

Querying on Data Summaries:

- orders of magnitude faster (*original operations replaced by fast summary transformations*)

- far less resources consumed
- original data remaining in-place

Approximate Query Result (*accurate enough from business perspective*)

Total Logs Analyzed: 4,141,434 Logs

Last 24 Hours

IP Address + | Report Device + | Vendor/Type + | Port + | User + | Country + | Allow/Deny + | Direction +

Filters: Allow/Deny = Allow X | Destination Port = 443 X

Reset Filters | VIEW LOG DETAILS

Log Summary: Last 24 Hours

Destination Country*

Ranking Mutual Information Dependencies in a Summary-based Approximate Analytics Framework

Dominik Ślęzak*, Janusz Borkowski† and Agnieszka Chądzyńska-Krasowska‡
*Institute of Informatics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
Email: slezak@mimuw.edu.pl
†Security On-Demand
12121 Scripps Summit Dr 320, San Diego, CA 92131, USA
Email: janusz.borkowski@securityondemand.com
‡Polish-Japanese Academy of Information Technology
ul. Koszykowa 86, 02-008 Warsaw, Poland
Email: honzik@pjwstk.edu.pl

***Abstract*—We continue our research on utilizing histogram-based data summaries in approximate derivation of mutual information scores in large relational data sets. Our methodology of creating, storing and using summaries has been designed for the purpose of developing an approximate database engine that is currently deployed commercially in the area of cyber-security data analytics. However, a similar idea of approximate data processing operations can be considered also in other fields, including machine learning whereby heuristic calculations are a component of many methods. In this paper, we focus on investigation of one possible source of inaccuracy of our previously proposed approach to approximating mutual information – that is, neglecting a kind of column domain drift during distributed summary-based computations. We illustrate it using an artificially created benchmark data set and we discuss how to cope this particular challenge in the future.**

***Index Terms*—Approximate Data Processing, Granulated Data Summaries, Approximate Mutual Information**

The considered engine is designed to perform on petabytes of the summarized data. In [1], one can find empirical comparison of the speed of our style of approximate calculations versus state-of-the-art methods of scaling by means of adding computational resources. In [2], we reported the current major commercial deployment of our engine in the field of online cyber-security, whereby ad-hoc analytical queries need to be executed against data sets containing detailed event logs growing with intensity of over 300 billions of new rows per month. In both works, we emphasized that our ultimate goal is to integrate the proposed methodology with aforementioned state-of-the-art solutions, so it is possible to work with granulated summaries representing pairwise disjoint pieces of the data in a fully parallel/distributed environment.

Label in the Loop

BRIGHTBOX

SMALL BIG DATA

CHALLENGES

QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows

Dziękuję!

slezak@mimuw.edu.pl