Toward Approximate Intelligence

inspired by Rough Sets and Granular Computing

Dominik Ślęzak
University of Warsaw, Poland
slezak@mimuw.edu.pl

My Rough Set Background

DAAR calculation:

Input: a decision system $\mathbb{S}_d = (U, A \cup \{d\})$; quality measure $\phi_d : 2^A \to \mathbb{R}$; acceptable probability of adding irrelevant attribute $p_{irr} \in [0, 1)$; attribute sample size $mTry$;

begin

$AR := \emptyset$; $flag := FALSE$;

while $flag = FALSE$ **do**

randomly select a subset $A' \subseteq A \setminus AR$ consisting of $mTry$ attributes;

$a_{best} := \arg\max_{a \in A'} \phi_d(AR \cup \{a\})$;

if $P\left(\phi_d(AR \cup \{a_{best}\}) \leq \phi_d(AR \cup \{\hat{a}_{best}\})\right) < p_{irr}$ **then**

$AR := AR \cup \{a_{best}\}$;

end

else

$flag := TRUE$;

end

end

Eliminate unnecessary attributes from AR; **return** AR;

end

My Rough Set Background

DAAR calculation:

Input: a decision system $\mathbb{S}_d = (U, A \cup \{d\})$; quality measure $\phi_d : 2^A \to \mathbb{R}$; acceptable probability of adding irrelevant attribute $p_{irr} \in [0, 1)$; attribute sample size $mTry$;

begin

$AR := \emptyset$; $flag := FALSE$;

while $flag = FALSE$ **do**

randomly select a subset $A' \subseteq A \setminus AR$ consisting of $mTry$ attributes;

$a_{best} := \arg\max_{a \in A'} \phi_d(AR \cup \{a\})$;

if $P\left(\phi_d(AR \cup \{a_{best}\}) \leq \phi_d(AR \cup \{\hat{a}_{best}\})\right) < p_{irr}$ **then**

$AR := AR \cup \{a_{best}\}$;

end

else

$flag := TRUE$;

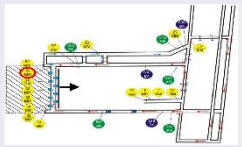
end

end

Eliminate unnecessary attributes from AR; **return** AR;

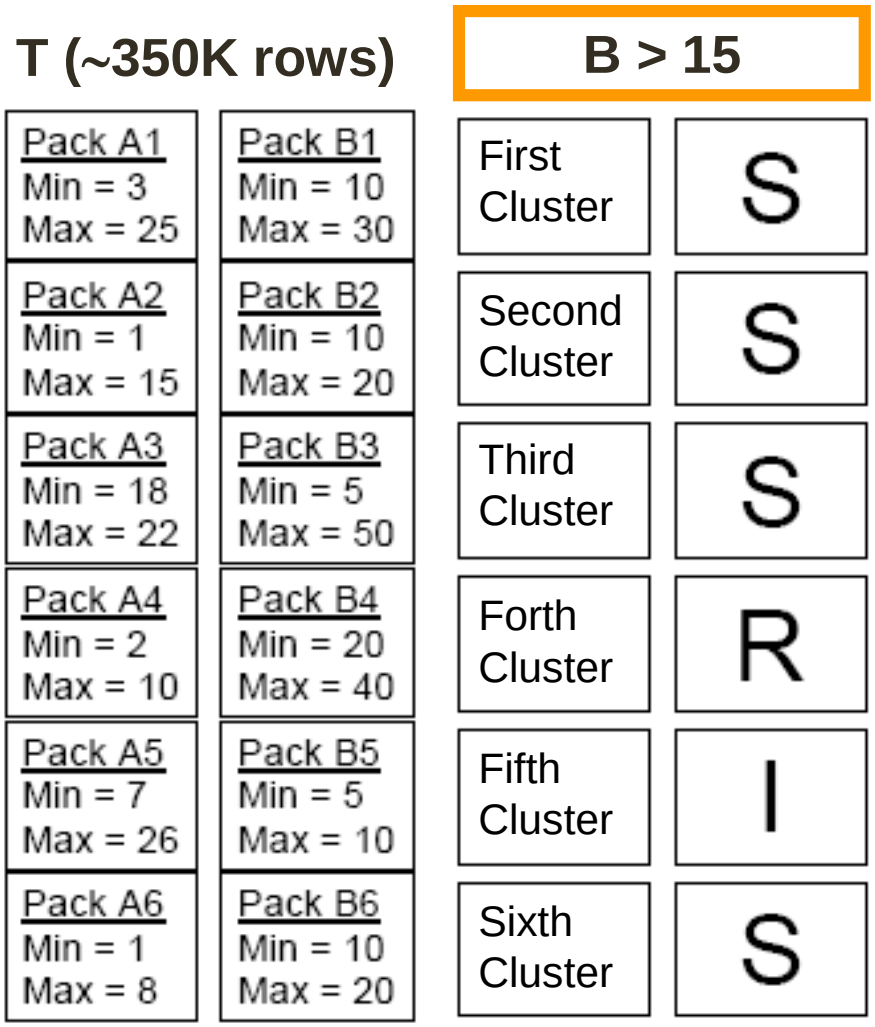
end

An ensemble of simple models based on different feature subsets (decision reducts)



PPBS2/B9/20/2013

SELECT MAX(A) FROM T WHERE B > 15



- **I**: Irrelevant Granules (*Negative Region*)
- **S**: Suspect Granules (*Boundary Region*)
- **R**: Relevant Granules (*Positive Region*)
- **E**: Exact Computation (necessary, if the final query result cannot be obtained only from the available summaries)

INFOBRIGHT

From Infobright (2005-2017) to Security On-Demand (2017-…)

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

ORIGINAL DATA

DATA LOAD

Outlook | Temp. | Humid. | Wind | Sport?

2^{16}

~ 2015

DATA PACKS
compressed collections of attribute values

QUERY
example related to filtering

rp[1] r[1] r[2] r[3] … r[2^{16}]

rp[2]

rp[3]

ROUGH VALUE USAGE

ROUGH ATTRIBUTES

ROUGH VALUE CALCULATION

	Outlook	Temp.	Humid.	Wind	Sport?
row pack 1	rough value	rough value	rough value	rough value	rough value
row pack 2	rough value	rough value	rough value	rough value	rough value
row pack 3	rough value	rough value	rough value	rough value	rough value

GRANULATED TABLE
a collection of rough values for each of rough attributes is stored as a separate knowledge node

identification of row packs and rows which satisfy query conditions

The Idea of Granular Analytics

Database

Petabytes

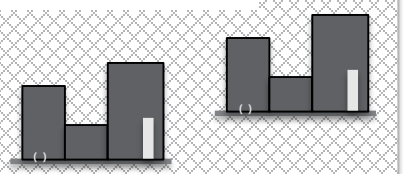
Traditional Query Execution:

- 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:

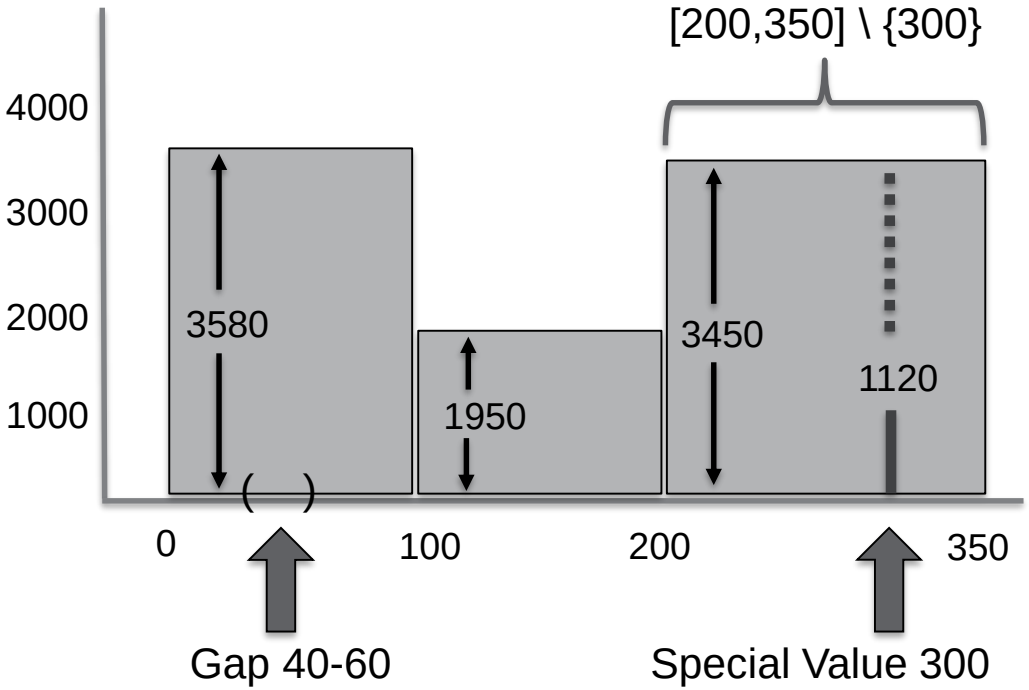
- orders of magnitude faster (*original operations replaced by fast summary transformations*)

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

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

Single-Data-Pack Summaries

- Value 300 occurred 1120 times
- There were 3450 occurrences of values between 200 and 350 *excluding value 300*
- There were no occurrences of values between 40 and 60
- Values 0, 40, 60, 100, 200, 350 occurred at least once



* Domain Granularity: How many ranges, special values and gaps?
* Data Granularity: How large each of single data packs should be?
* Summaries of correlations between columns are needed as well.

The New Era of Threat Protection. Driven by SuperScale Analytics

SECURITY ON-DEMAND

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*

The New Era of Threat Protection. Driven by SuperScale Analytics

SECURITY ON-DEMAND

Total Logs Analyzed: 4,141,434 Logs

Last 24 Hours

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

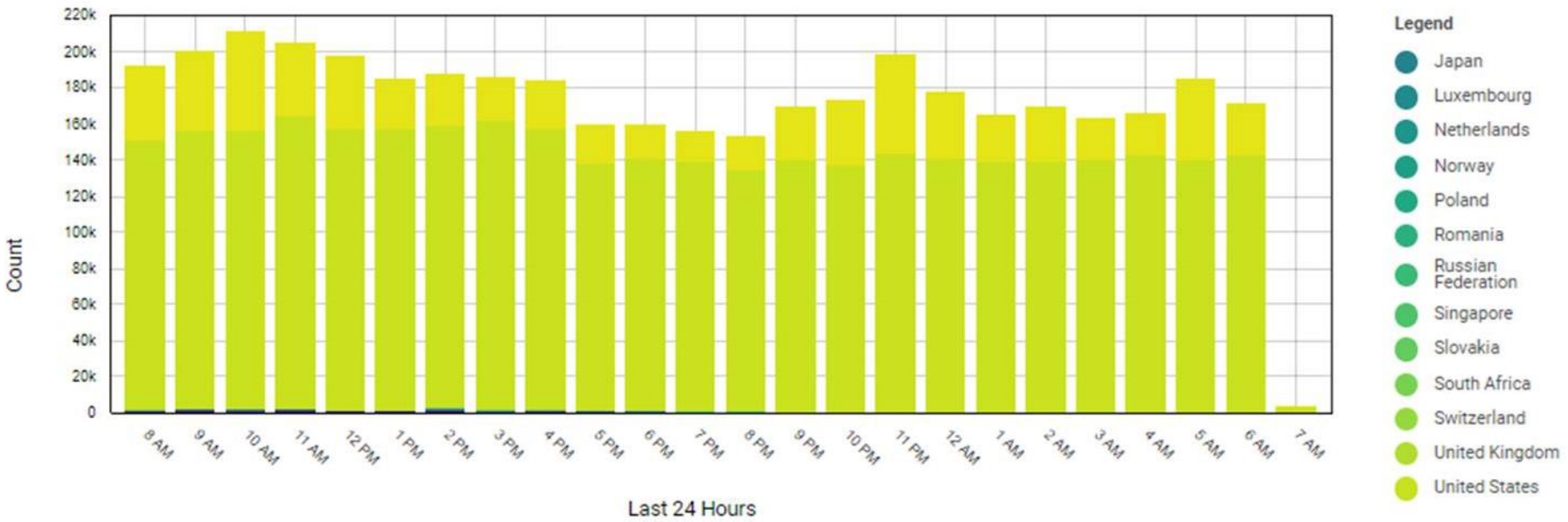
Filters: Allow/Deny = Allow X | Destination Port = 443 X

Exploration and Exploitation

Reset Filters | VIEW LOG DETAILS

Log Summary: Last 24 Hours

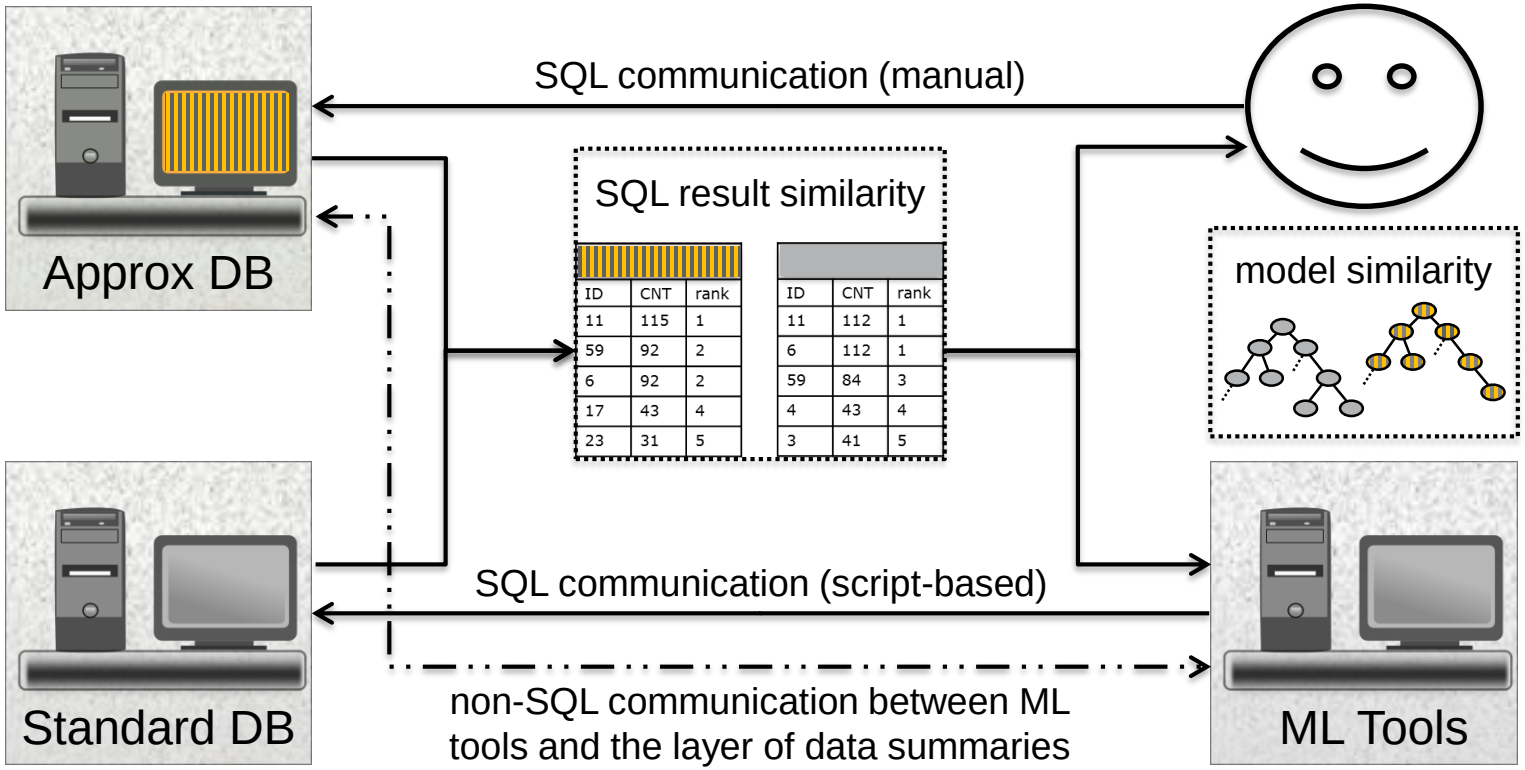
Destination Country*



Toward Approximate Intelligence

- Humans do not operate with fine-grained information and thus, it might be not needed to provide them with precise outcomes of reasoning, modeling or analytical processes.
- The question arises whether approximate querying (and learning!) processes can be sufficient for decision-making while being delivered faster than in standard scenarios.
- This leads us toward a discussion about the importance of approximations in BI, ML and, generally, about the meaning of approximate data representations and derivations in AI.

How Accurate Calculations Do We Need to Learn?



Approximate Computations

DAAR calculation:

Input: a decision system $\mathbb{S}_d = (U, A \cup \{d\})$; quality measure $\phi_d : 2^A \to \mathbb{R}$; acceptable probability of adding irrelevant attribute $p_{irr} \in [0, 1)$; attribute sample size $mTry$;

begin

$AR := \emptyset$; $flag := FALSE$;

while $flag = FALSE$ **do**

randomly select a subset $A' \subseteq A \setminus AR$ consisting of $mTry$ attributes;

$a_{best} := \arg\max_{a \in A'} \phi_d(AR \cup \{a\})$;

if $P\left(\phi_d(AR \cup \{a_{best}\}) \leq \phi_d(AR \cup \{\hat{a}_{best}\})\right) < p_{irr}$ **then**

$AR := AR \cup \{a_{best}\}$;

end

else

$flag := TRUE$;

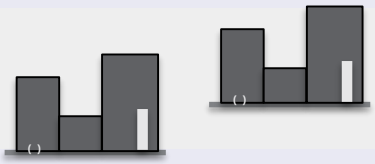
end

end

Eliminate unnecessary attributes from AR; **return** AR;

end

Do we need to compute values of this function exactly to run the heuristic process?



No! Those values can be computed approximately (and orders of magnitude faster) using summaries.

Learning on Granulated Data

- Data summaries at Infobright / SOD follow the idea of information granulation, with granules gathering groups of physically adjacent data items (*another example of granulation of data items refers to RS-based attribute reduction*).
- Besides, we can think, e.g., about granulation of domain (*quantization*) or resolution (*a kind of image granulation using neural networks*).
- But regardless of the specifics of granulation, operations related to approximations, belief propagation or even machine learning can be implemented over granular representations.

Granular Information Systems

DAAR calculation:

Input: a decision system $\mathbb{S}_d = (U, A \cup \{d\})$; quality measure $\phi_d : 2^A \to \mathbb{R}$; acceptable probability of adding irrelevant attribute $p_{irr} \in [0, 1)$; attribute sample size $mTry$;

begin

$AR := \emptyset$; $flag := FALSE$;

while $flag = FALSE$ **do**

randomly select a subset $A' \subseteq A \setminus AR$ consisting of $mTry$ attributes;

$a_{best} := \arg\max_{a \in A'} \phi_d(AR \cup \{a\})$;

if $P\left(\phi_d(AR \cup \{a_{best}\}) \leq \phi_d(AR \cup \{\hat{a}_{best}\})\right) < p_{irr}$ **then**

$AR := AR \cup \{a_{best}\}$;

end

else

$flag := TRUE$;

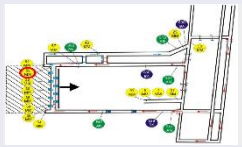
end

end

Eliminate unnecessary attributes from AR; **return** AR;

end

Approximations of features and their values or labels; Granular feature representations



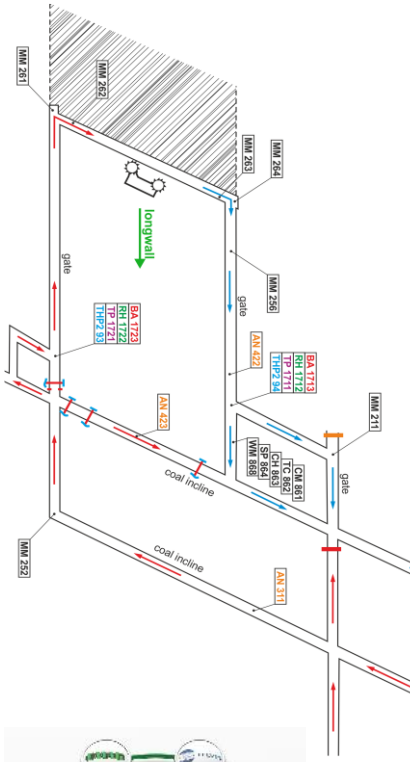
UNIVERSITY OF WARSAW

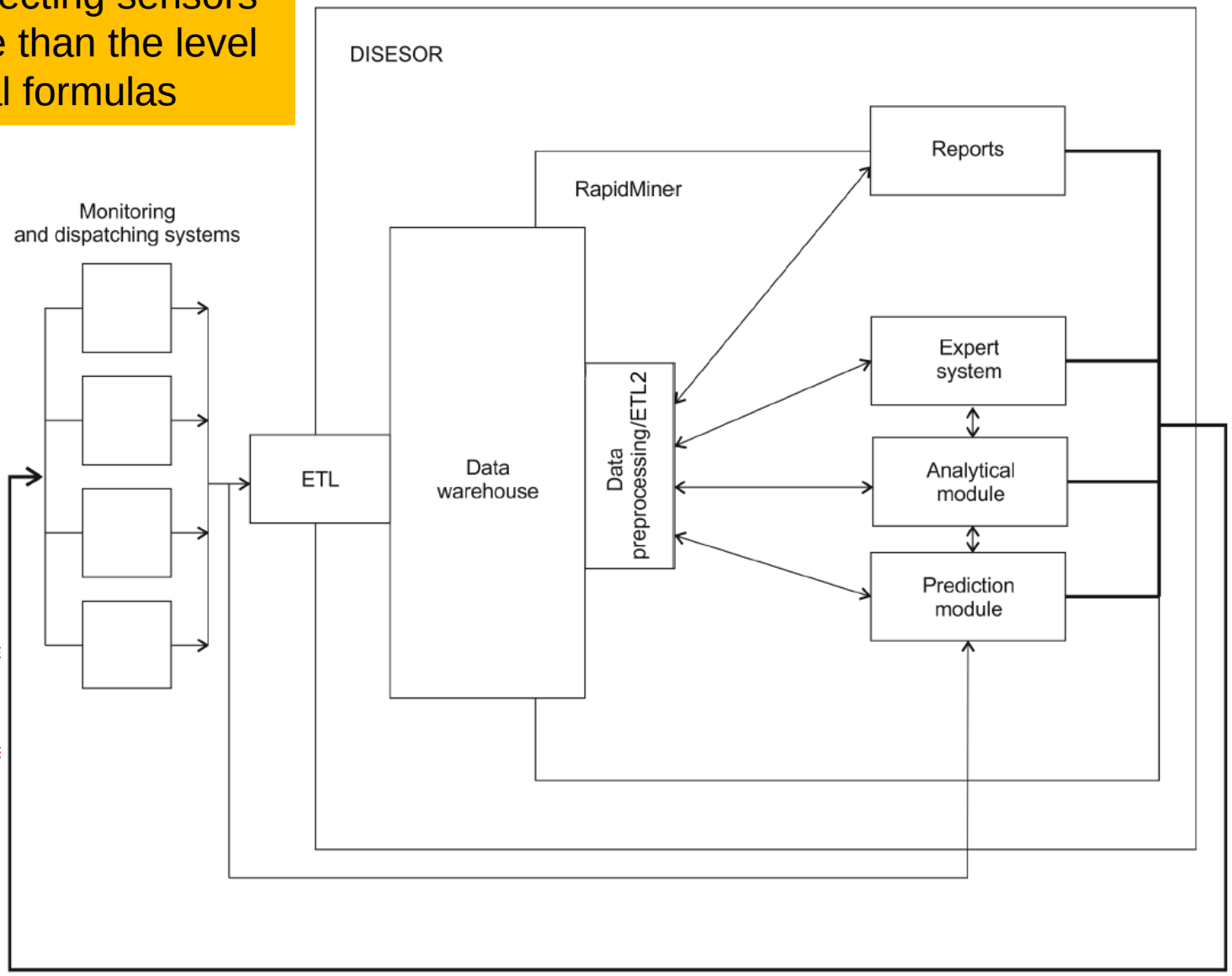
The National Centre for Research and Development

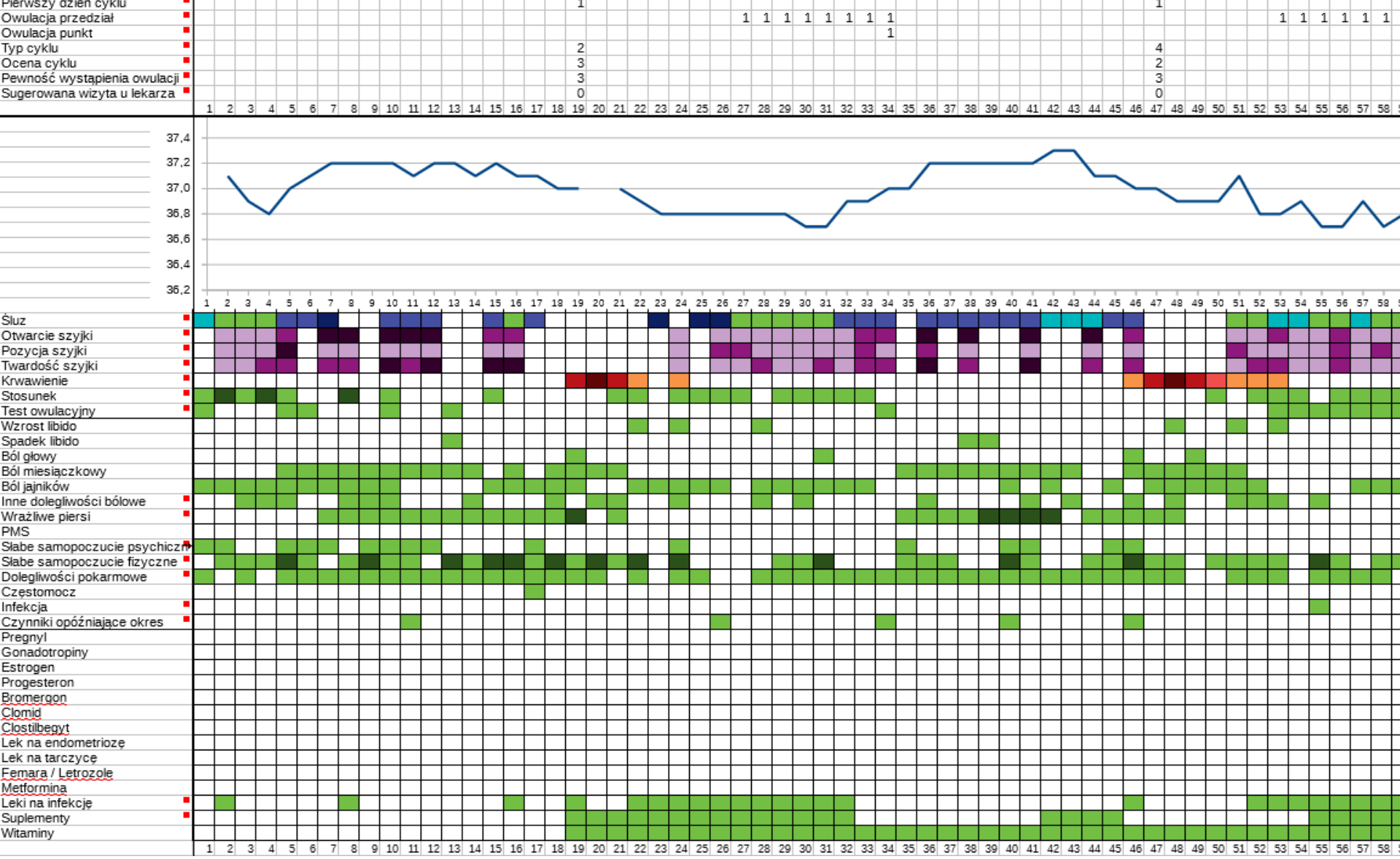
PPBS2/B9/20/2013

How to interact with the coal mining experts to select features?

- The level of selecting sensors is more intuitive than the level of mathematical formulas



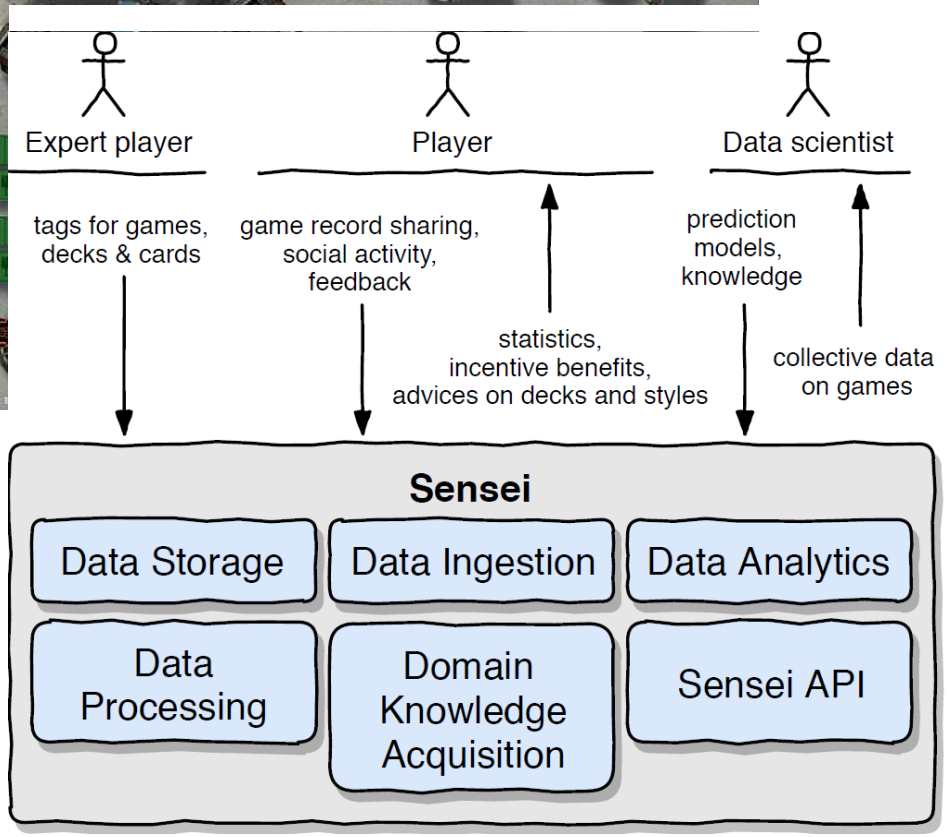




How to approximate, complement or replace temperature measurements in the process of prediction of fertile days?

OvuFriend.pl

grail esensei



How to create an intuitive layer of abstraction to work with real-time computer games?

- Granular level for game developers
- Granular level for game players

SECURITY ON-DEMAND

Thank You!

UNIVERSITY OF WARSAW

1. **ROUGH SET INSPIRATIONS IN INFOBRIGHT DB** D. Ślęzak, P. Synak, A. Wojna, J. Wróblewski: Two Database Related Interpretations of Rough Approximations: Data Organization and Query Execution. Fundamenta Informaticae 127(1-4) (2013) 445-459
2. **APPROXIMATE / GRANULAR QUERY ENGINE** D. Ślęzak, R. Glick, P. Betliński, P. Synak: A New Approximate Query Engine Based on Intelligent Capture and Fast Transformations of Granulated Data Summaries. Journal of Intelligent Information Systems 50(2) (2018) 385-414
3. **FEATURE SELECTION BASED ON SUMMARIES** D. Ślęzak, J. Borkowski, A. Chądzyńska-Krasowska: Ranking Mutual Information Dependencies in a Summary-based Approximate Analytics Framework. Proc. of HPCS 2018, 852-859
4. **ENSEMBLES OF REDUCTS IN COAL MINING** D. Ślęzak, M. Grzegorowski, A. Janusz, M. Kozielski, S.H. Nguyen, M. Sikora, S. Stawicki, Ł. Wróbel: A Framework for Learning and Embedding Multi-Sensor Forecasting Models into a Decision Support System: A Case Study of Methane Concentration in Coal Mines. Information Sciences 451-452 (2018) 112-133
5. **FEATURE SELECTION ON FEATURE GRANULES** M. Grzegorowski, A. Janusz, D. Ślęzak, M. Szczuka: On the Role of Feature Space Granulation in Feature Selection Processes. Proc. of IEEE Big Data 2017, 1806-1815
6. **GRANULAR MODELING IN REAL-TIME GAMES** M. Świechowski, D. Ślęzak: Granular Games in Real-Time Environment. Proc. of ICDM 2018

European Funds Intelligent development | European Union European regional development fund | The National Centre for Research and Development

EU Smart Growth Operational Programme 2014-2020: POIR.01.02.00-00-0150/16-00

EU Smart Growth Operational Programme 2014-2020: POIR.01.02.00-00-0184/17-00

EU Smart Growth Operational Programme 2014-2020: POIR.01.01.01-00-0831/17-00