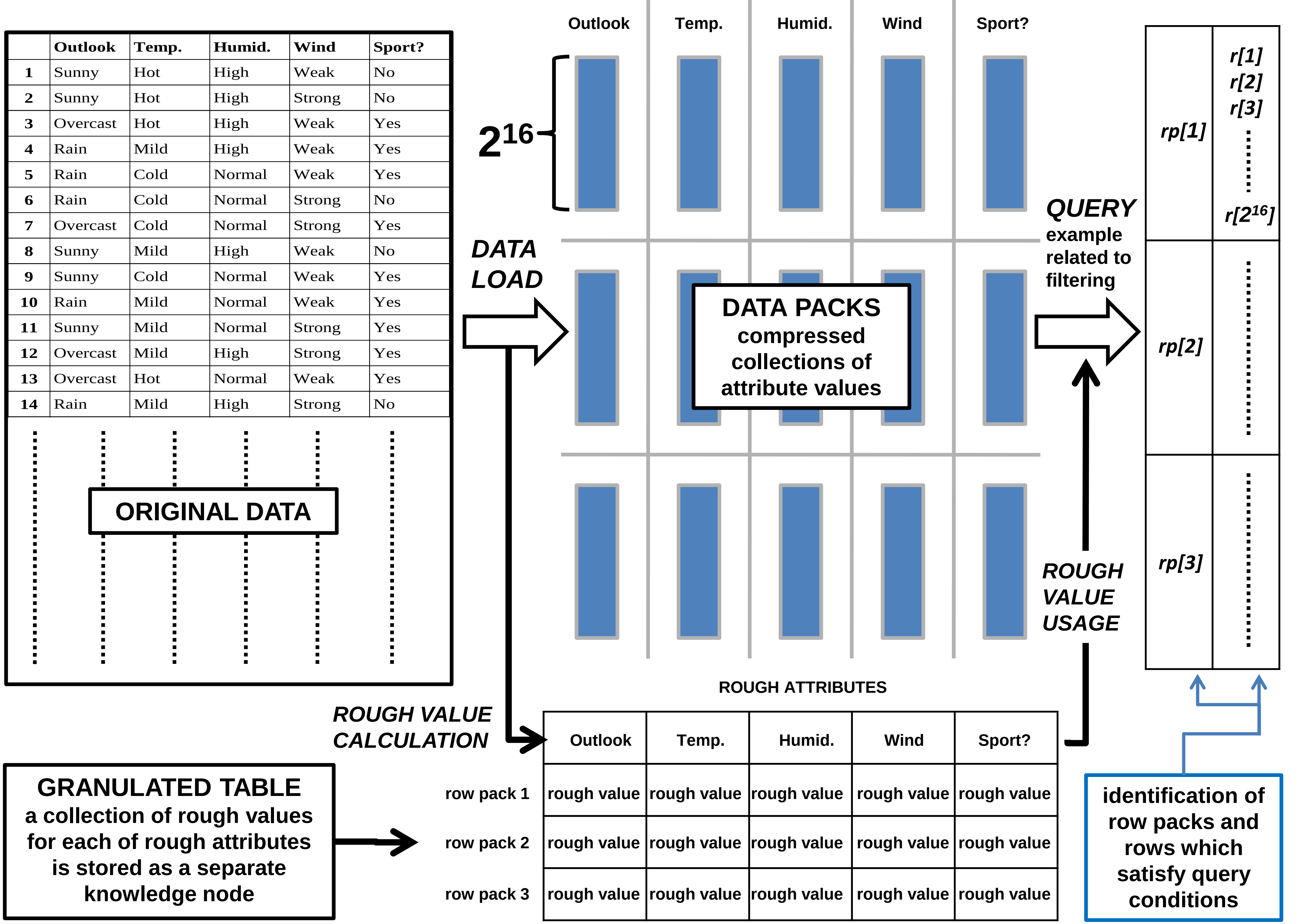
Przybliżony SQL jako jeden z aspektów skalowalności obliczeń

Dominik Ślęzak

Big Data: Przetwarzanie i Eksploracja

Poznań, 2016-04-22



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
2^16
Outlook
Temp.
Humid.
Wind
Sport?
DATA PACKS
compressed collections of attribute values
QUERY
example related to filtering
rp[1]
rp[2]
rp[3]
r[1]
r[2]
r[3]
r[2^16]
ROUGH VALUE USAGE
ROUGH VALUE CALCULATION
ROUGH ATTRIBUTES
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

SELECT MAX(A) FROM T WHERE B > 15;

Data Table T

			B > 15
Pack A1 Min = 3 Max = 25	Pack B1 Min = 10 Max = 30		S
Pack A2 Min = 1 Max = 15	Pack B2 Min = 10 Max = 20		S
Pack A3 Min = 18 Max = 22	Pack B3 Min = 5 Max = 50		S
Pack A4 Min = 2 Max = 10	Pack B4 Min = 20 Max = 40		R
Pack A5 Min = 7 Max = 26	Pack B5 Min = 5 Max = 10		I
Pack A6 Min = 1 Max = 8	Pack B6 Min = 10 Max = 20		S

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

[18,25] → [18,22-25] after Exact Computation on A1/B1

INFOBRIGHT

How Accurate Calculations do we Need?

Trade Settlement

Pharmaceutical Effectiveness

Network Intrusion Detection

Demand Forecasting

Financial Risk, Scoring and Analysis

Audience Profiling

Operational Reporting

Technical Troubleshooting and Remediation

Logistics and Demand Forecasting

Agricultural Yield Management

Test Scoring

Price Sensitivity Analysis and Optimization

Cyber Security

Trading Algorithms Development

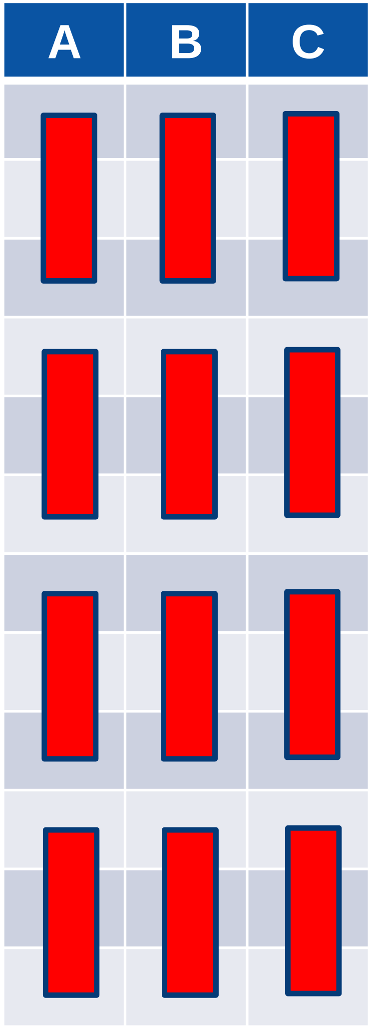
Tools for Machine Learning and Data Mining

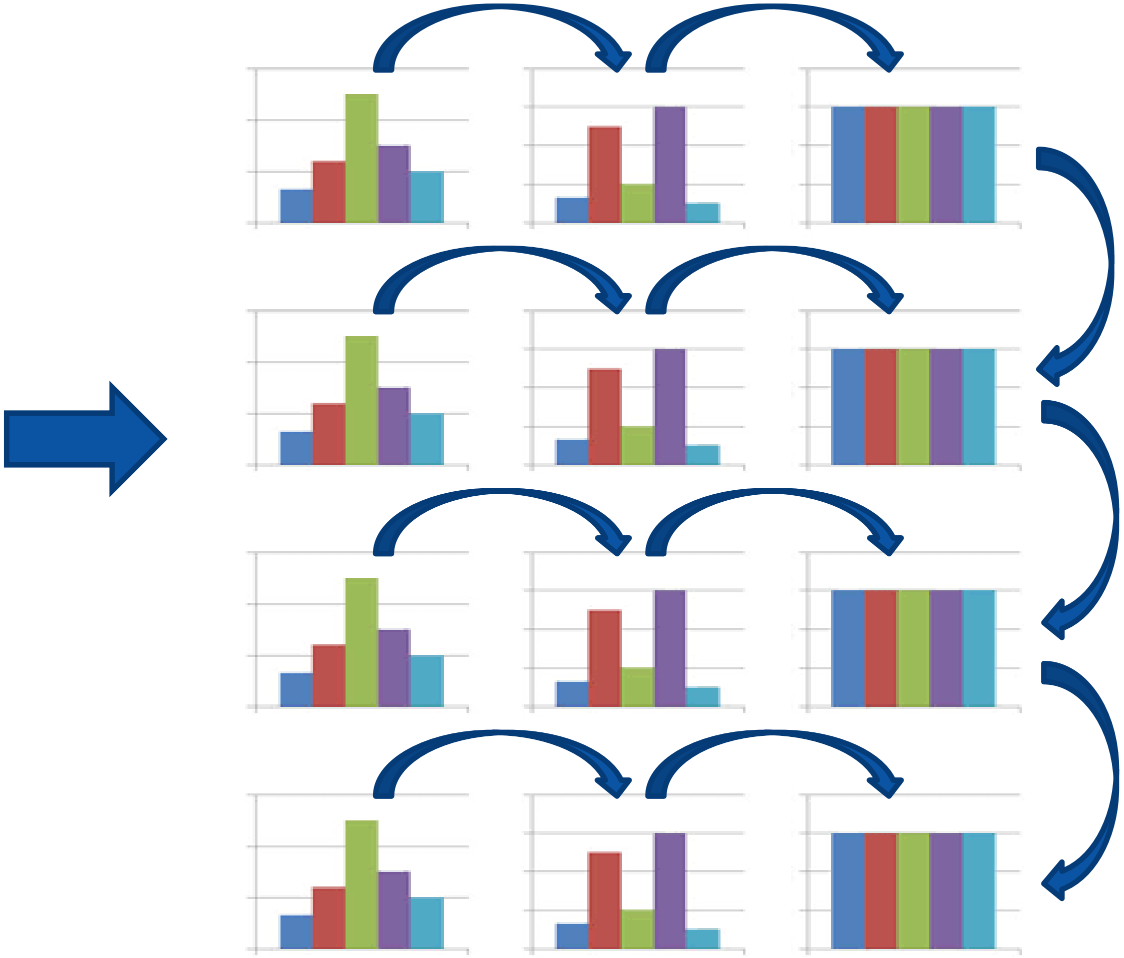
Exact Accuracy

Eventual Accuracy

Approximate Only

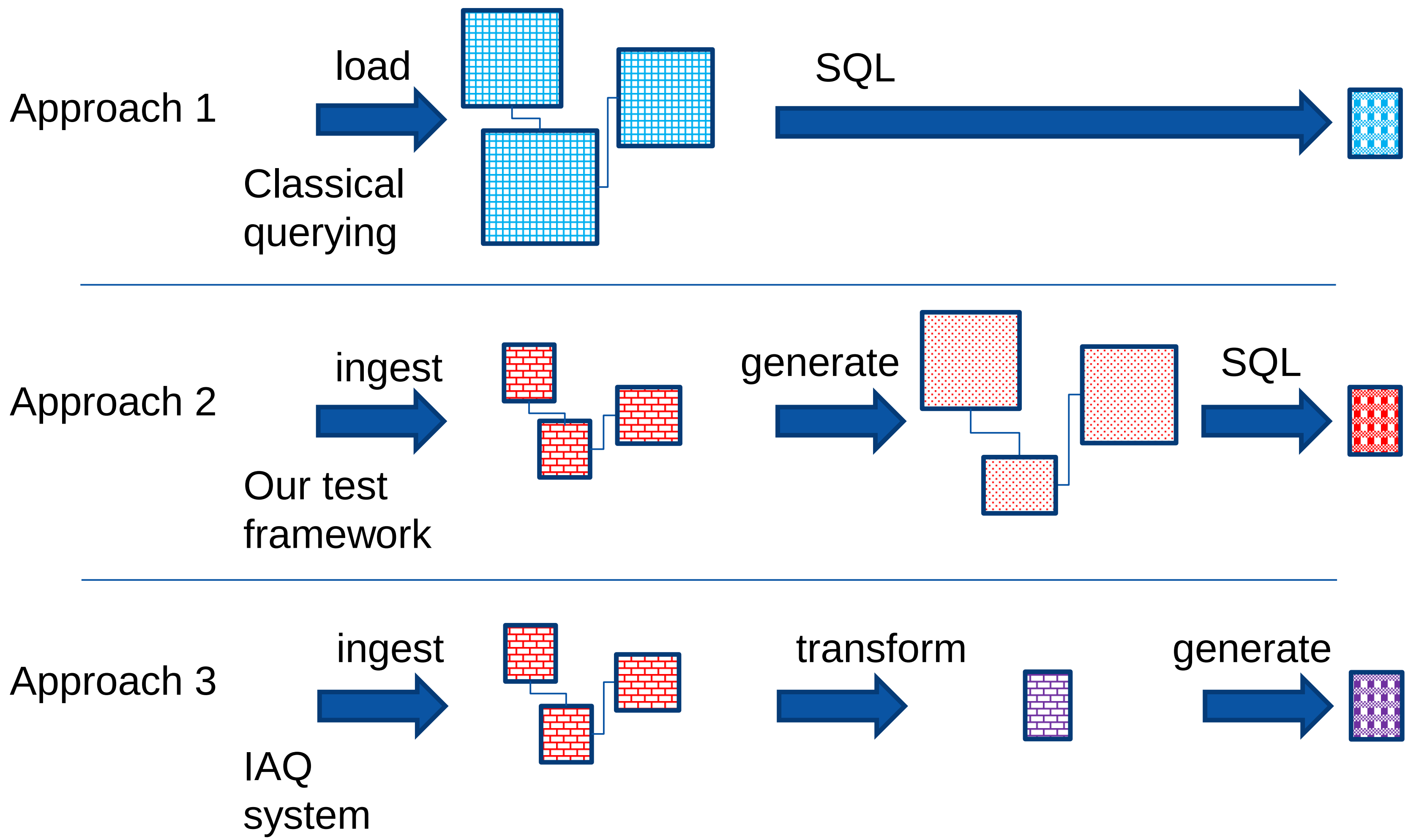
„Statistics is Our New Data"





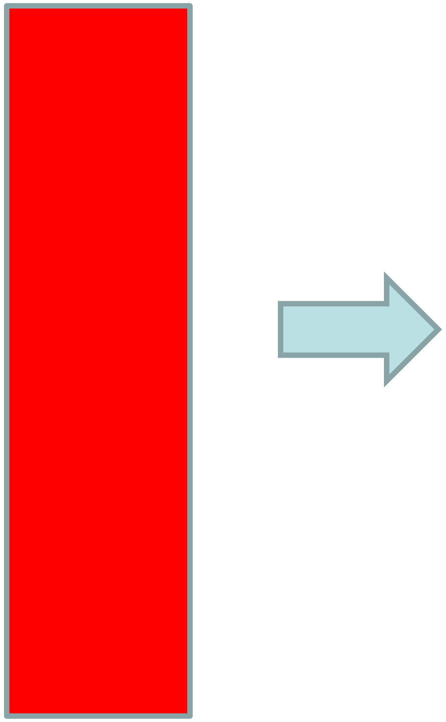
INFOBRIGHT

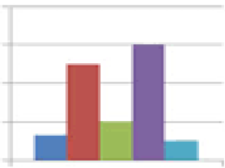
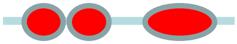
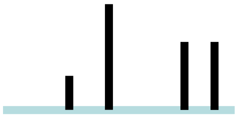
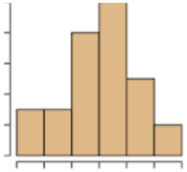
Approaches to Data Operations



INFOBRIGHT

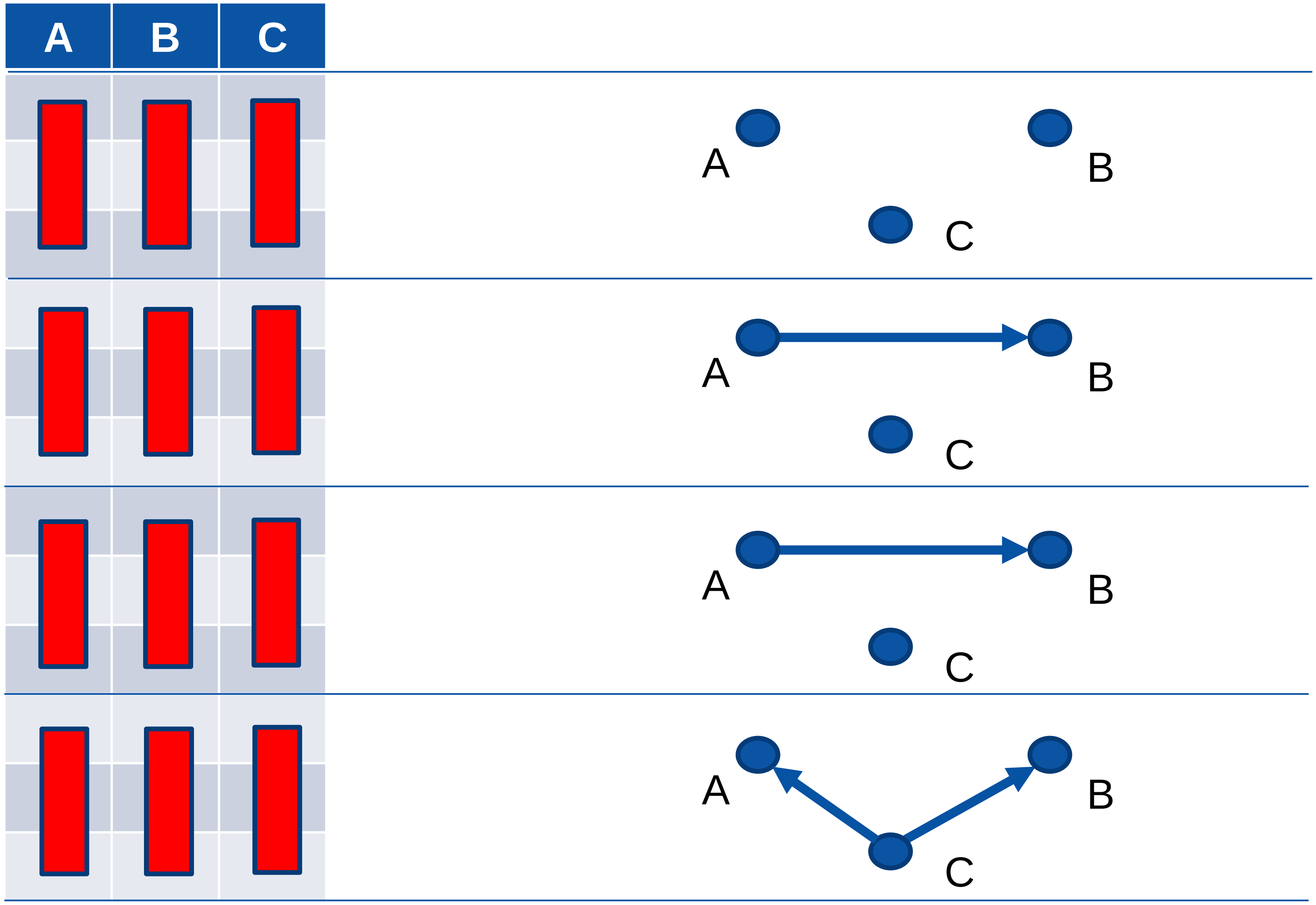
Single-Column Descriptions



- Histogram 
- Domain 
- Specials 
- Densities 

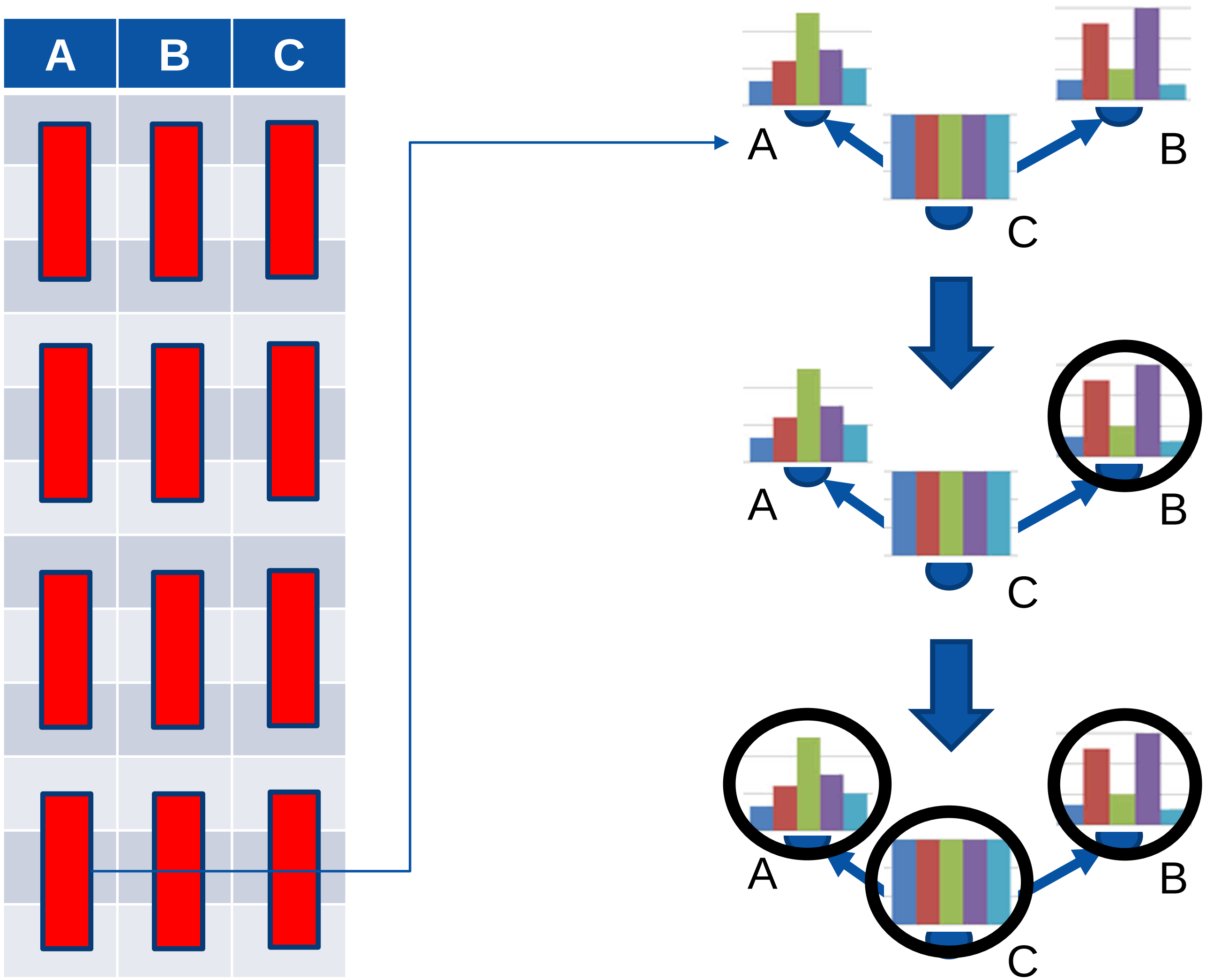
Do not think only about numeric columns

Multi-Column Descriptions



INFOBRIGHT

How to Transform Data Descriptions?

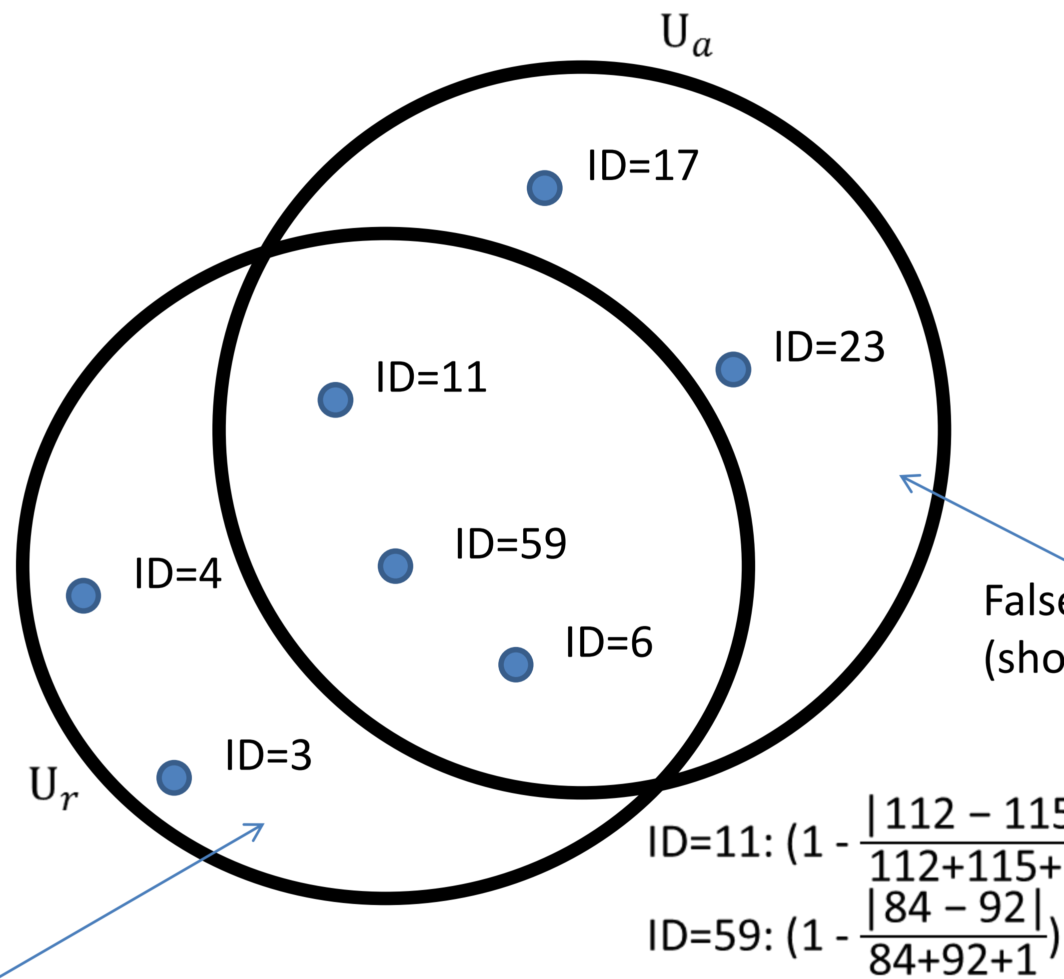


SELECT SUM(A) FROM T WHERE B > 7;

SELECT ID, COUNT(*) AS CNT FROM VISITS GROUP BY 1 ORDER BY 2 DESC LIMIT 5;

RESULT $R_r = (U_r, C)$		
ID	CNT	rank
11	112	1
6	112	1
59	84	3
4	43	4
3	41	5

RESULT $R_a = (U_a, C)$		
ID	CNT	rank
11	115	1
59	92	2
6	92	2
17	43	4
23	31	5



False Positives
(shouldn't occur but did)

True Negatives
(should occur but didn't)

$card(U_r) = 5$
$card(U_a) = 5$
$card(U_r \cap U_a) = 3$

$$\text{ID=11: } \left(1 - \frac{|112 - 115|}{112+115+1}\right)\left(1 - \frac{|1 - 1|}{1 + 1}\right) = 0.99$$

$$\text{ID=59: } \left(1 - \frac{|84 - 92|}{84+92+1}\right)\left(1 - \frac{|3 - 2|}{3 + 2}\right) = 0.76$$

$$\text{ID=6: } \left(1 - \frac{|112 - 92|}{112+92+1}\right)\left(1 - \frac{|1 - 2|}{1 + 2}\right) = 0.60$$

$$\text{TotSim}(R_r, R_a) = (0.99 + 0.76 + 0.60) / 7 = \mathbf{0.34}$$

Summary

- How to generate and evaluate descriptions, so they can reflect „perceptual" similarity of exact and approximate results?
- How to explain the expected similarities to the users?
- Is there analogy to „classical" data mining?
- Is it only about SQL?
- Where data descriptions are coming from?
- What is the trade-off between speed and accuracy?

DZIĘKUJĘ!!!

slezak@mimuw.edu.pl
slezak@infobright.com
www.dominikslezak.org