

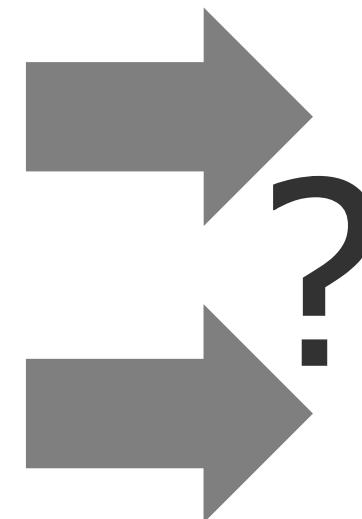
Data for Analytics

Domain knowledge and AutoML
in analytic processes

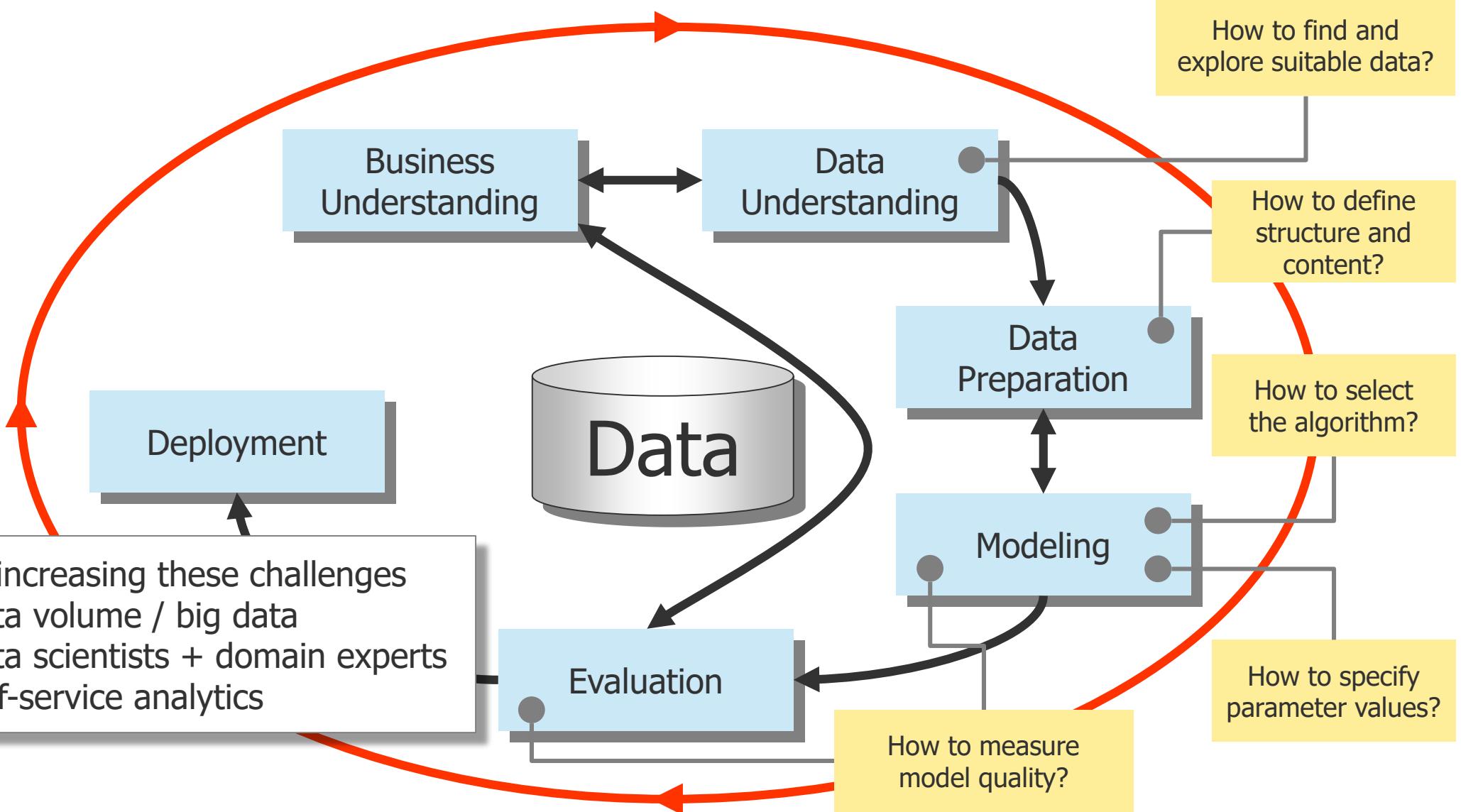
Holger
Schwarz

Data Shopping in Data Marketplaces is only one Step ...

The image shows two screenshots of a Data Marketplace interface. The top screenshot displays a search results page for 'Health' data. It includes a sidebar with filters for Category, Security Class, Owner, Type, Format, and a checkbox for 'located in Data Lake'. The results list three items: 'health dbms_db', 'sleep_study dbms_table', and 'hand_washing dbms_table', each with details like Data Owner, Security Class, and a 'Data Asset' button. A blue arrow points from this screen down to the second screenshot. The bottom screenshot shows a 'SHOPPING CART - 1 ITEM' page. It lists a single item: 'salaries hdf5_path' (Data with Salaries of Teachers (not anonymised)) with details: Data Owner: Erik Baumgartner, Created: 06.08.2021, Security Class: 3, Zone: raw zone. It includes sections for 'Intended Usage' (asking why the data is needed) and 'Data Provisioning' (asking how data will be used). A green 'Checkout' button is at the bottom right. A blue arrow points from this screen to the large question mark in the center.

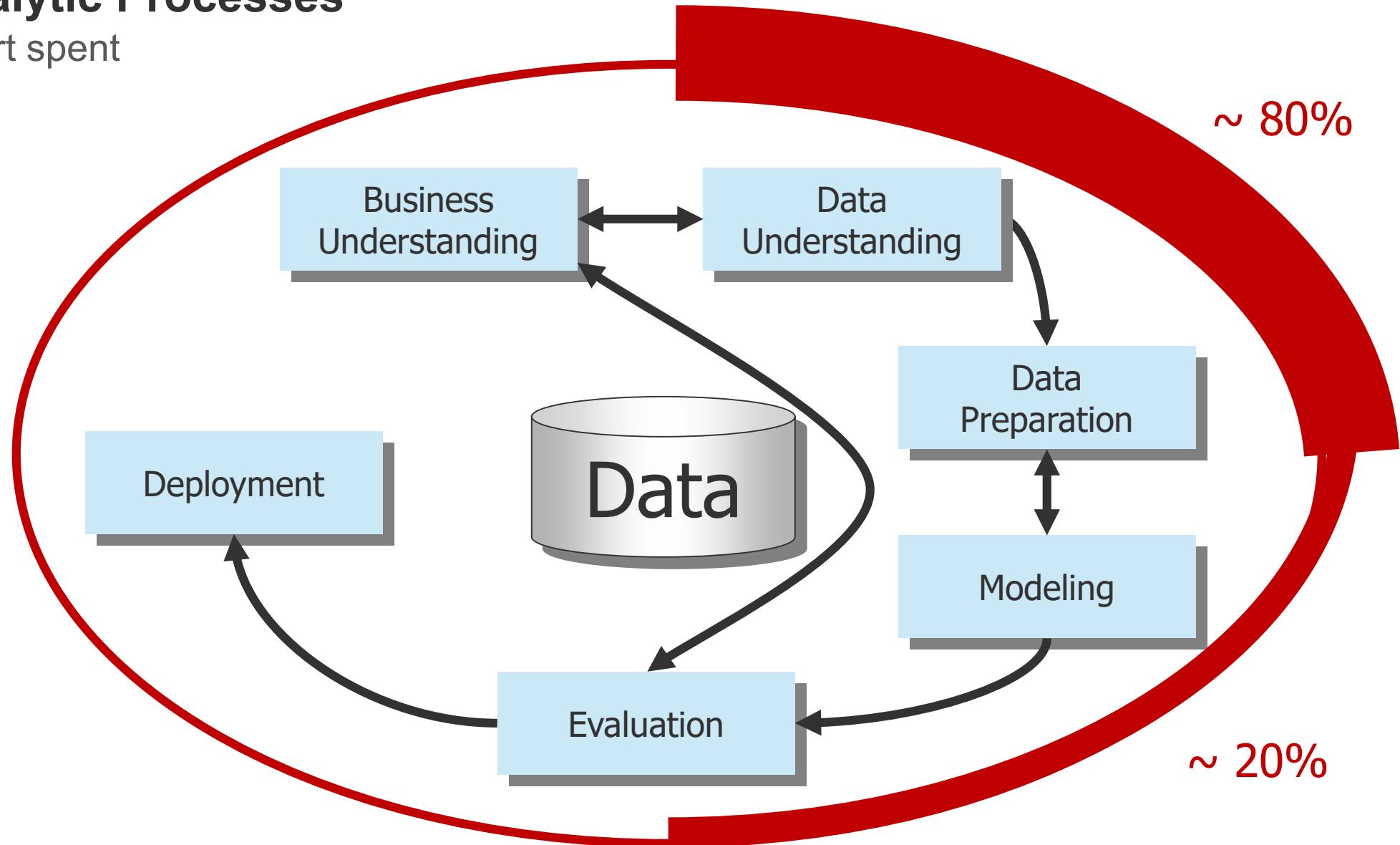


Challenges in Analytic Processes



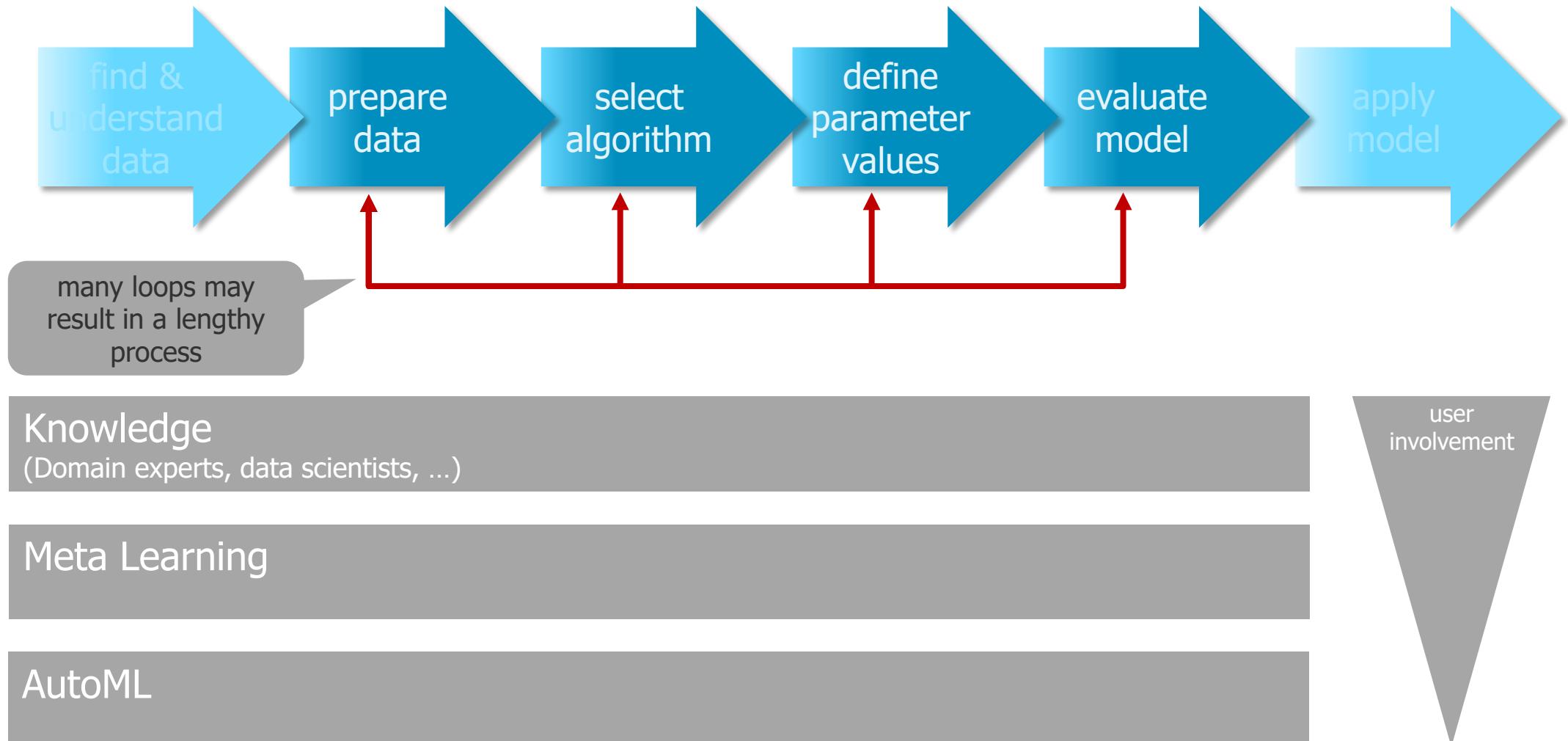
Analytic Processes

Effort spent

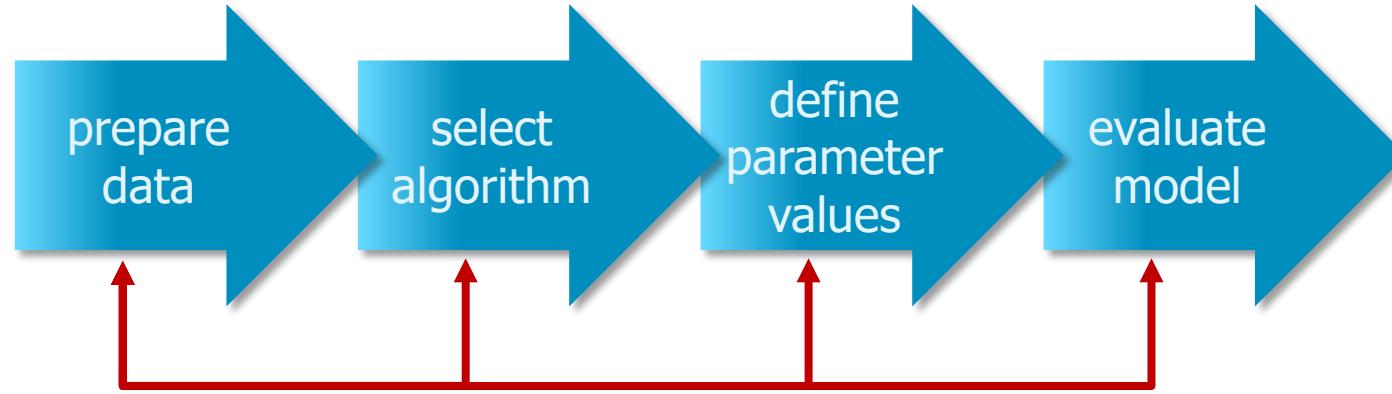


Analytic Processes

Focus and categories of approaches



Agenda



Knowledge
(Domain experts, data)

Engine [N: 1,050] [C: 84 S: 73]			
Series			
HDE [N: 277] [C: 60 S: 52]		MDE [N: 773] [C: 58 S: 59]	
Type	OM470 [N: 160] [C: 37 S: 43]	OM471 [N: 139] [C: 44 S: 52]	...
Model	471000 [N: 52] [C: 26 S: 48]	471002 [N: 12] [C: 9 S: 41]	...
	936680 [N: 309] [C: 43 S: 57]	936610 [N: 41] [C: 24 S: 44]	...

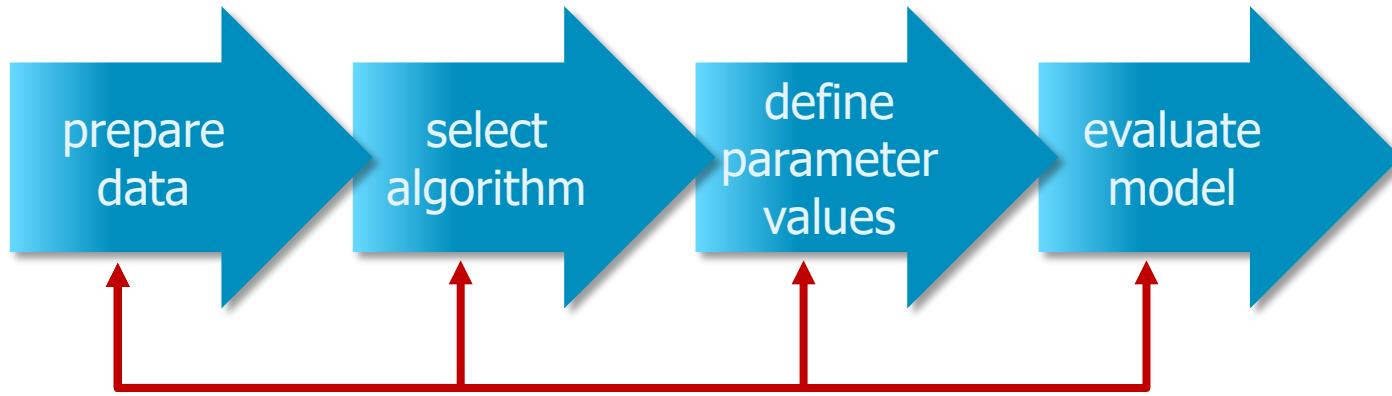
MDE: Medium-Duty Engine | HDE: Heavy-Duty Engine

Meta Learning

1 classification system to exploit domain knowledge for multi-class classification

AutoML

Agenda

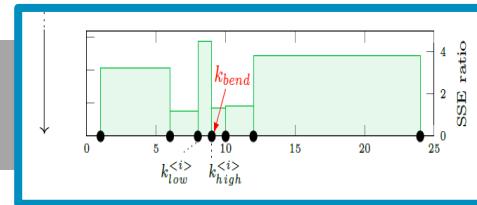


Knowledge

(Domain experts, data scientists, ...)

Meta Learning

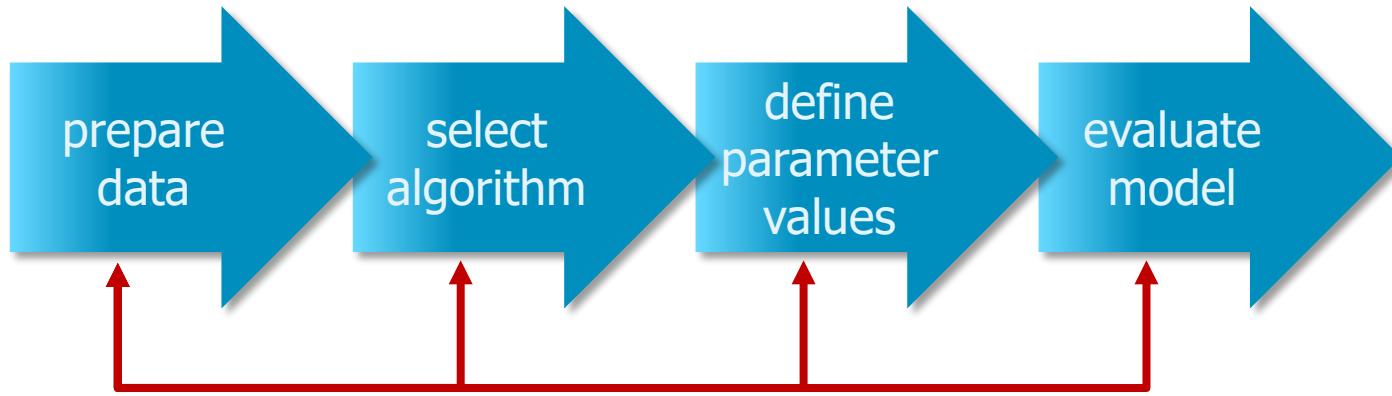
AutoML



2

method to efficiently determine the number of clusters in datasets

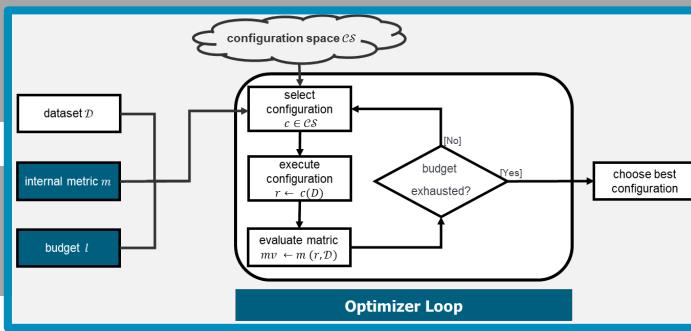
Agenda



Knowledge

(Domain experts, data scientists, ...)

Meta Learning

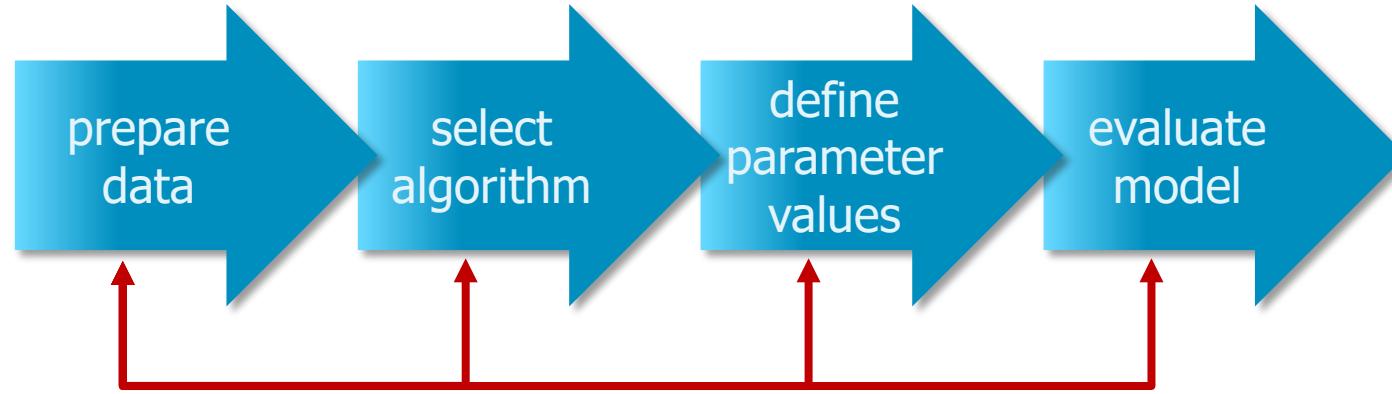


AutoML

3

AutoML
approach for
clustering

Agenda



Knowledge
(Domain experts, data)

Engine [N: 1,050] [C: 84 S: 73]		
HDE [N: 277] [C: 60 S: 52]		
MDE [N: 773] [C: 58 S: 59]		
Series		
OM470	OM471	...
[N: 160] [C: 37 S: 43]	[N: 139] [C: 44 S: 52]	[N: 200] [C: 43 S: 51]
Type		
471000	471002	...
[N: 52] [C: 26 S: 48]	[N: 12] [C: 9 S: 41]	[N: 309] [C: 43 S: 57]
Model		
471000	471002	...
[N: 52] [C: 26 S: 48]	[N: 12] [C: 9 S: 41]	[N: 309] [C: 43 S: 57]

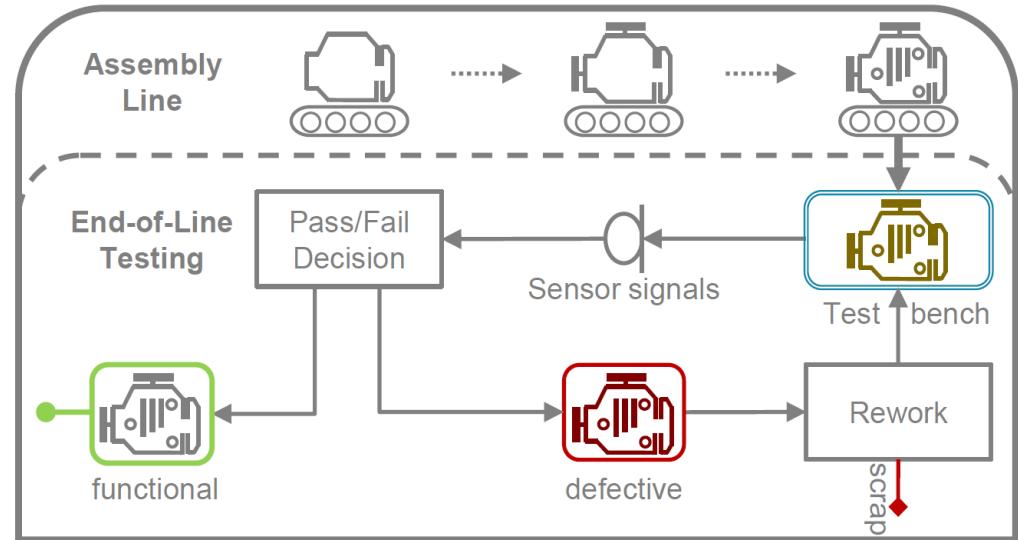
1 classification system to exploit domain knowledge for multi-class classification

Meta Learning

AutoML

Use Case: End-of-line Testing

- Final function test of complex products like engines of trucks
- Fault detection based on measurable sensor signals
- Rework includes
 - fault isolation
 - work to fix quality issue
- Rework based on engineers' subjective knowledge may lead to
 - wrong decisions
 - ineffective rework attempts / waste of time

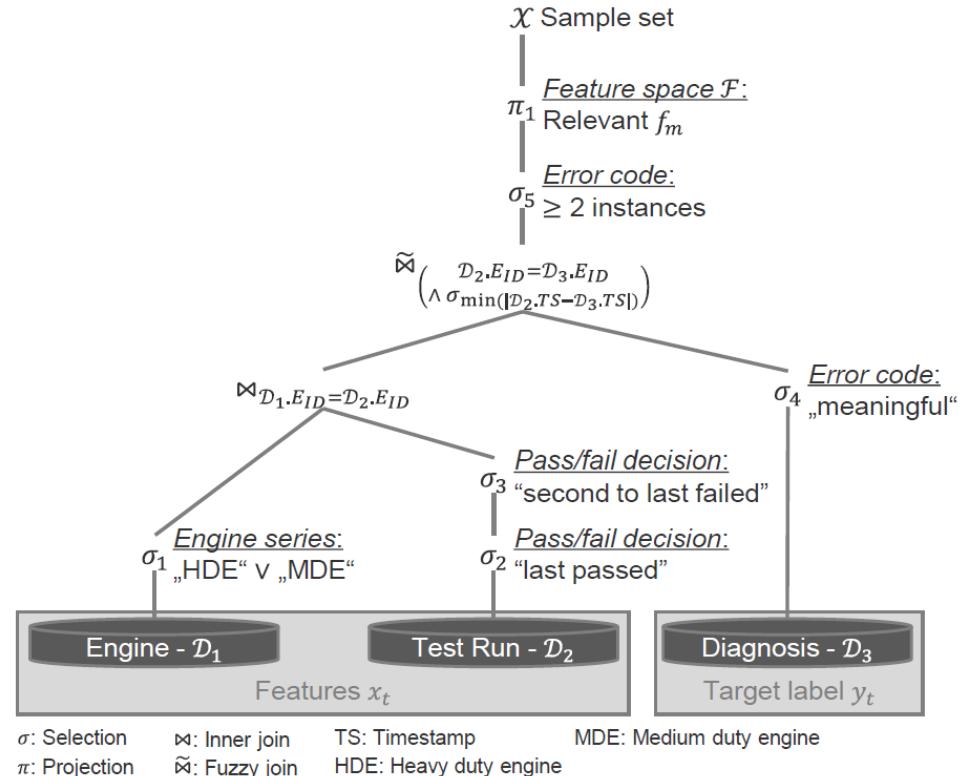


→ data driven approach desirable

Training Data on EoL Testing

- Available datasets
 - D_1 : engine data
 - D_2 : sensor signals with timestamps
 - D_3 : diagnostic data and rework attempts
- Data integration to obtain suitable training data with
 - number of samples $N = 1050$
 - number of features $M = 115$
 - number of classes: 84

Set	Name	Topic	Sample Feature	N	M
D_1	Engine	Spec sheet	-Engine type -Engine design	519,214	22
D_2	Test bench	Test runs	-Sensors s_k	621,689	383
D_3	Diagnostic	Faults	-Error code y_t	20,631	77

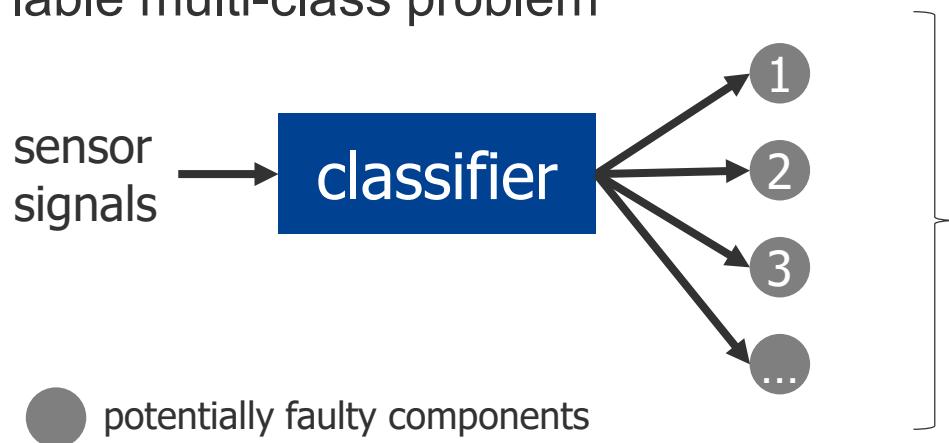


Data-driven Fault Isolation

- Fault detection based on binary classification



- Fault isolation considered as single-label multi-class problem

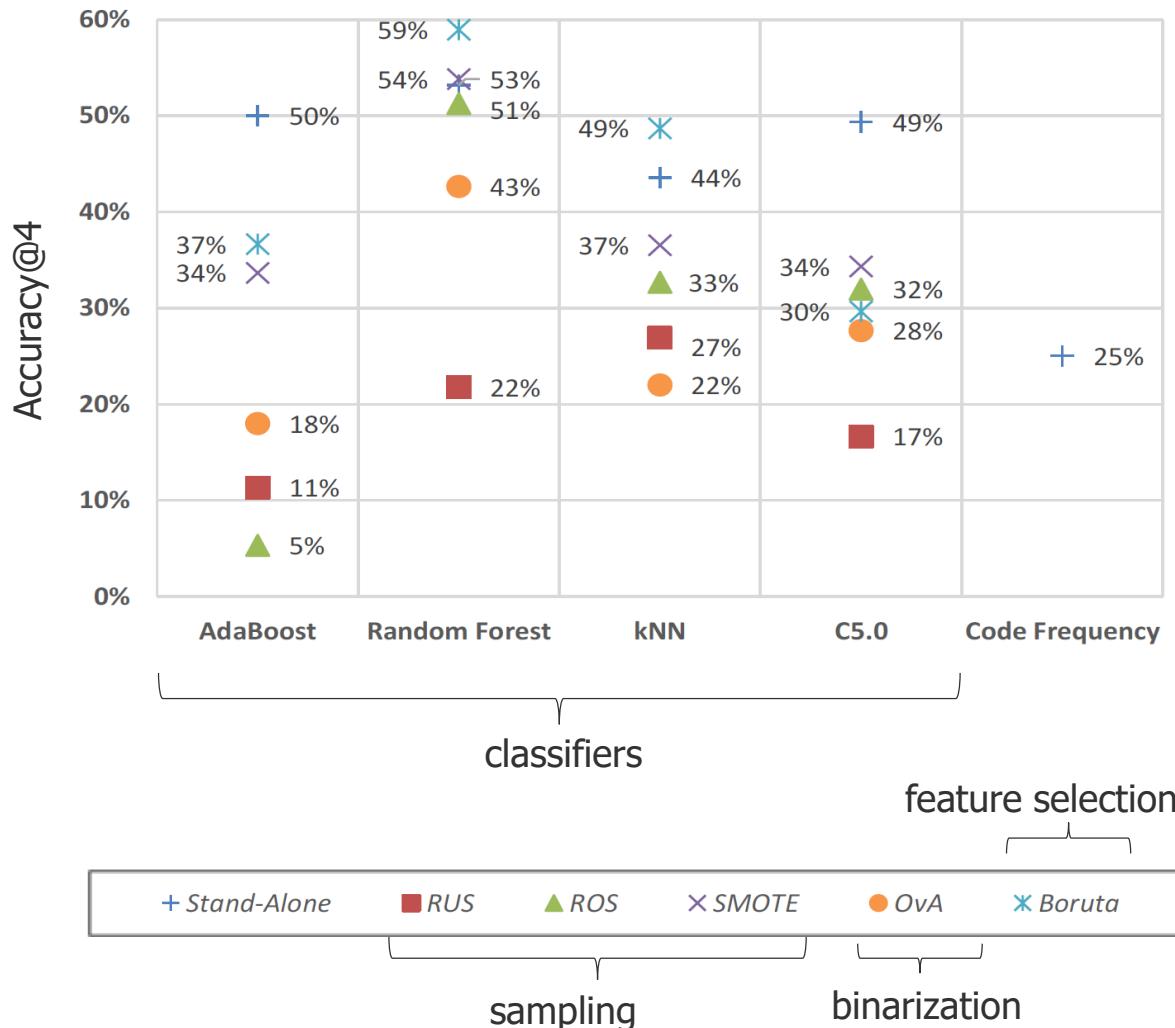


- Recommendation list of length p sorted by confidence values

[3 0.8; 1 0.65; 5 0.5; 2 0.45; ...]

rework attempts

Evaluation Results



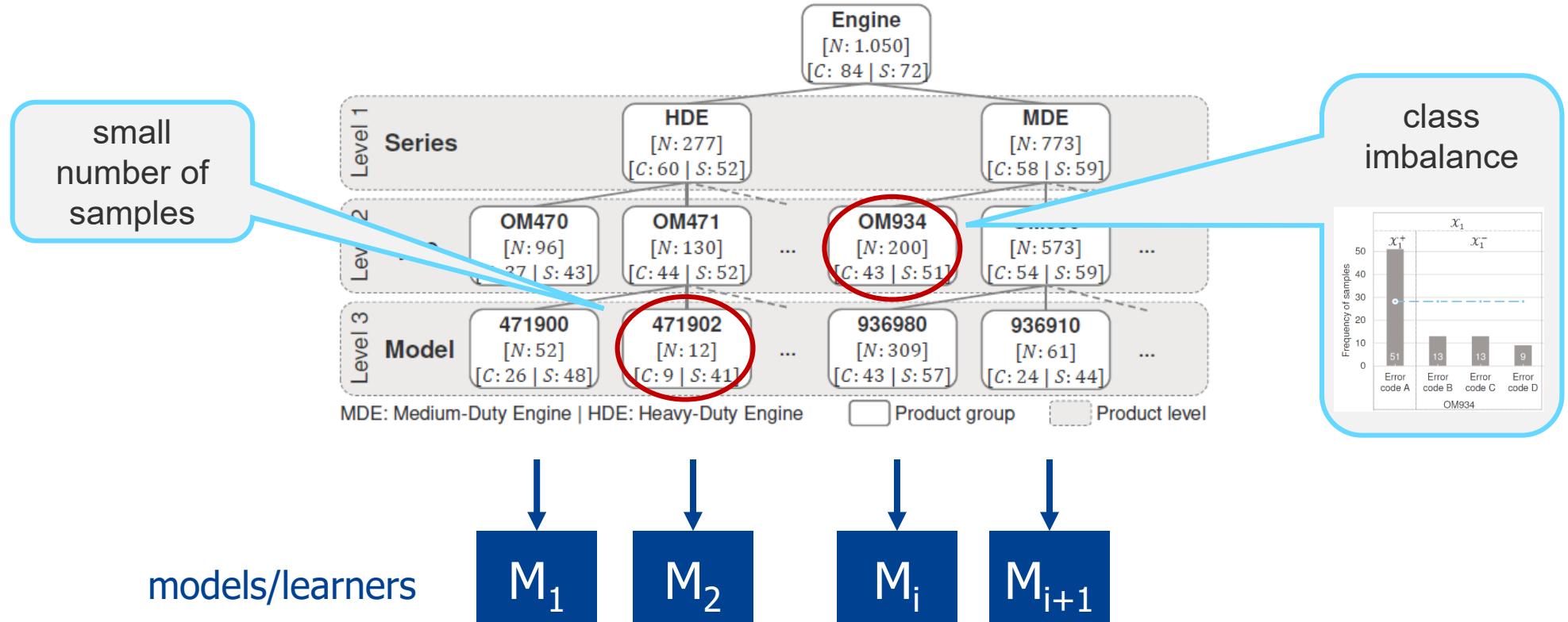
- Improvements of accuracy compared to baseline possible
- Leads to reduced number of rework attempts
- Achieved accuracy depends on
 - selected algorithm
 - type of data preparation

→ how to exploit domain knowledge?

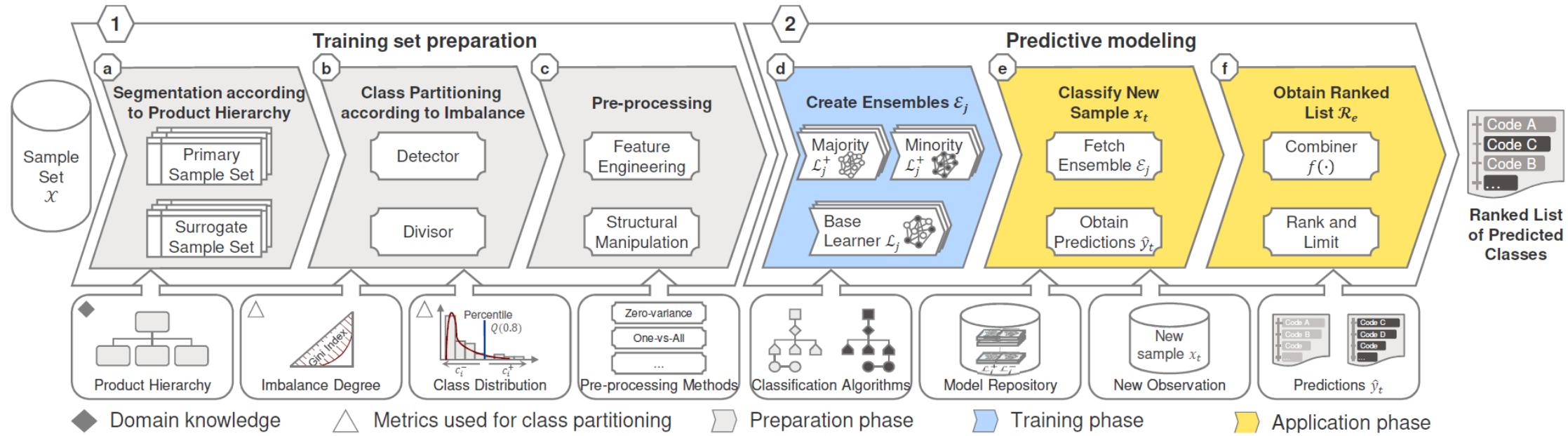
V. Hirsch, P. Reimann, B. Mitschang: Data-Driven Fault Diagnosis in End-of-Line Testing of Complex Products. DSAA 2019

Domain Knowledge: Product Hierarchy

- Engines are organized in groups according to series, type and model
- Approach to exploit this domain knowledge: **learn separate model for each group**
- Multiple challenges

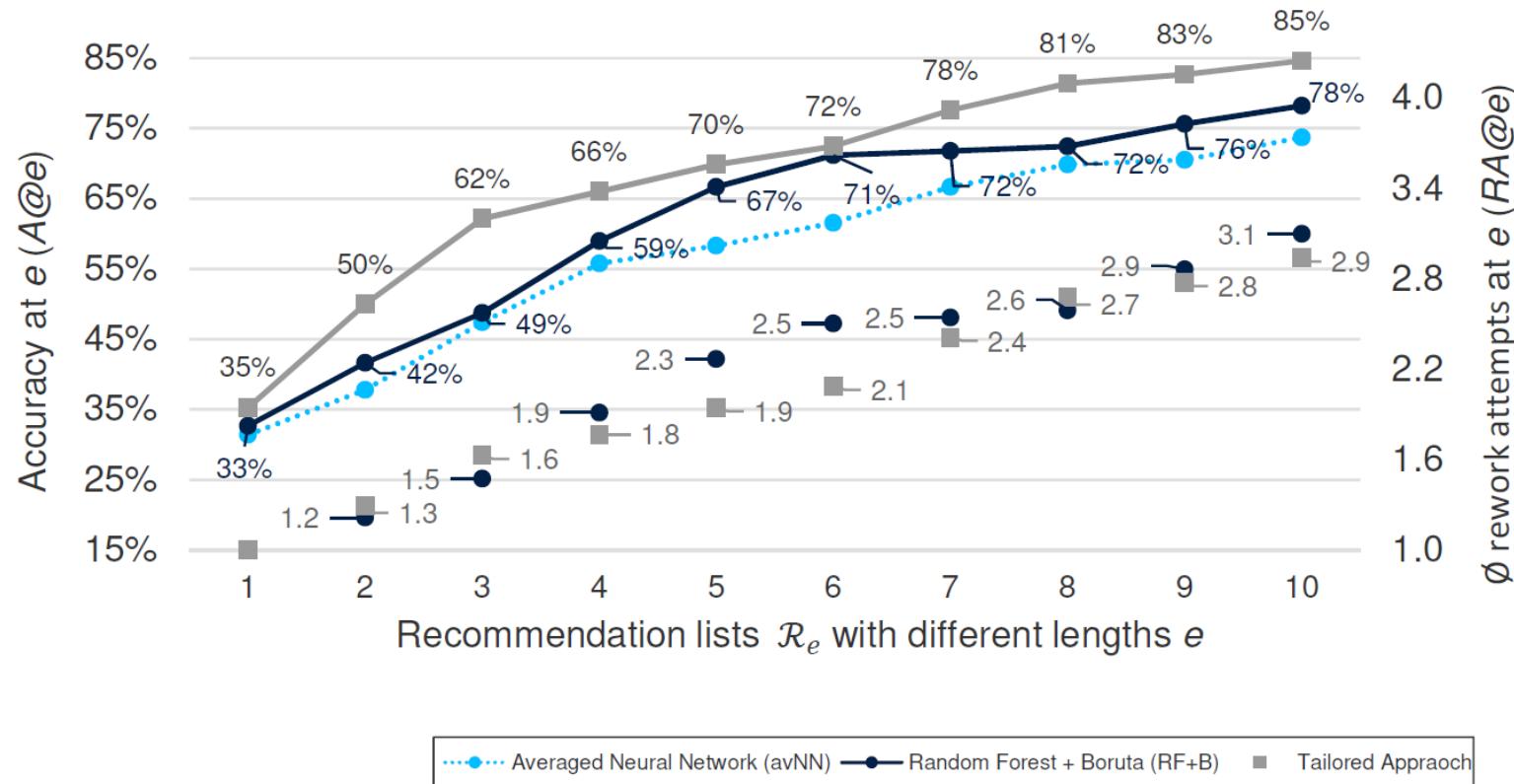


Classification System



V. Hirsch, P. Reimann, B. Mitschang: Exploiting Domain Knowledge to Address Multi-Class Imbalance and a Heterogeneous Feature Space in Classification Tasks for Manufacturing Data. VLDB 2020

Evaluation Results

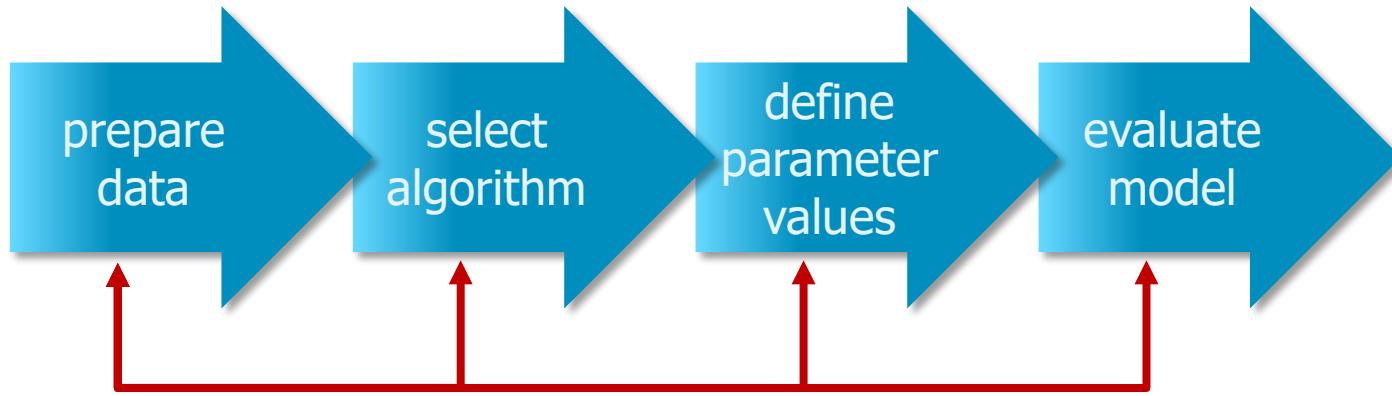


What's next?

- Generalization
 - are the main steps of the classification system independent from the use case?
 - what constraints have to hold on the domain knowledge to make the classification system work?
 - Use clustering in data preparation
 - can we achieve similar effects by automatically cluster the training data?
 - how can we exploit domain knowledge in constraint-based clustering?
 - Systematically exploit further domain knowledge
-
- | Source | Representation | Integration |
|---|--|--|
| Which source of knowledge is integrated? | How is the knowledge represented? | Where is the knowledge integrated in the machine learning pipeline? |
| Scientific Knowledge
(Natural Sciences, Engineering, etc.) | Algebraic Equations
Differential Equations
Simulation Results
Spatial Invariances | Training Data |
| World Knowledge
(Vision, Linguistics, Semantics, General K., etc.) | Logic Rules
Knowledge Graphs
Probabilistic Relations | Hypothesis Set
(Network Architecture, Model Structure, etc.) |
| Expert Knowledge
(Intuition, Less Formal) | Human Feedback | Learning Algorithm
(Regularization Terms, Constrained Opt., etc.) |
| | | Final Hypothesis |
-
- The diagram illustrates the integration of three types of knowledge into a machine learning pipeline:
- Source:** The knowledge is categorized into three main types: Scientific Knowledge (Natural Sciences, Engineering, etc.), World Knowledge (Vision, Linguistics, Semantics, General K., etc.), and Expert Knowledge (Intuition, Less Formal).
 - Representation:** Each type of knowledge is represented in different ways:
 - Scientific Knowledge is represented as Algebraic Equations, Differential Equations, Simulation Results, and Spatial Invariances.
 - World Knowledge is represented as Logic Rules, Knowledge Graphs, and Probabilistic Relations.
 - Expert Knowledge is represented as Human Feedback.
 - Integration:** The final output is a **Final Hypothesis**, which is generated by integrating the knowledge from the three sources through a process involving **Training Data**, a **Hypothesis Set** (containing Network Architecture and Model Structure), and a **Learning Algorithm** (which includes Regularization Terms and Constrained Opt.).

[Rue21]

Agenda

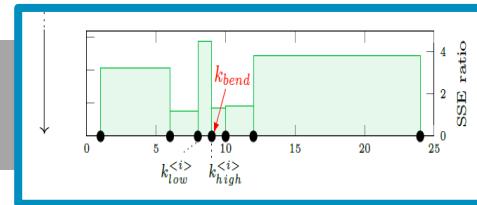


Knowledge

(Domain experts, data scientists, ...)

Meta Learning

AutoML

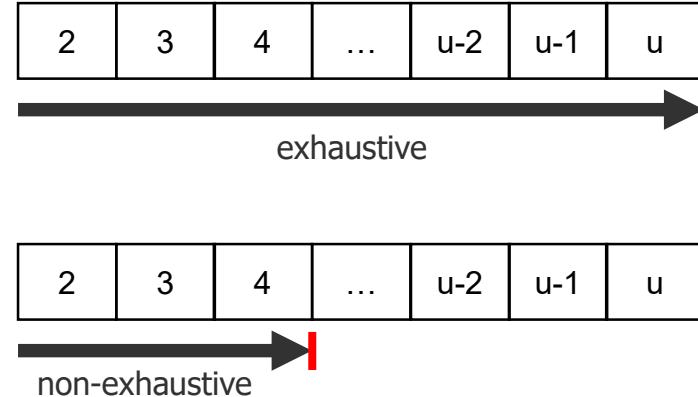


2

method to efficiently determine the number of clusters in datasets

Estimating Number of Clusters

- Search space $R = [2, n]$
- Exploit **domain knowledge** to reduce search space to $R = [2, u], u \ll n$
- Without domain knowledge
 - exhaustive search
 - non-exhaustive search
- Available **estimation methods**

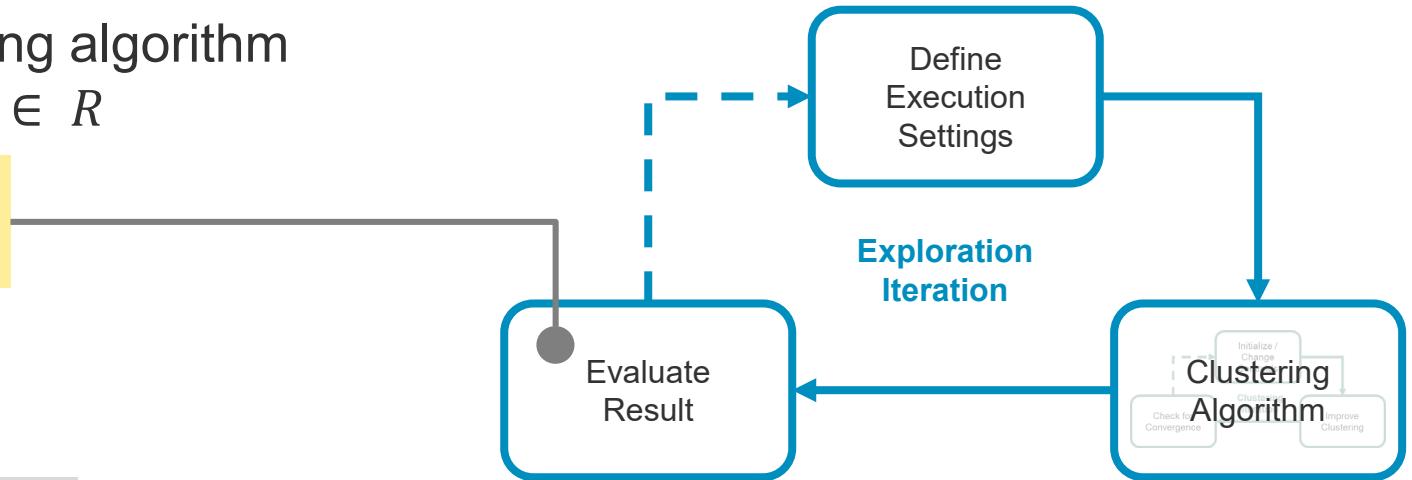


	automatic	semi-automatic
exhaustive	Akaike Information Criterion Bayesian Information Criterion Calinski-Harabasz Index Coggins-Jain Index	Davies-Bouldin Index Dunn Index Jump Method Silhouette Coefficient
non-exhaustive	Gap Statistics G-Means	X-Means (AIC) X-Means (BIC)

LOG-Means

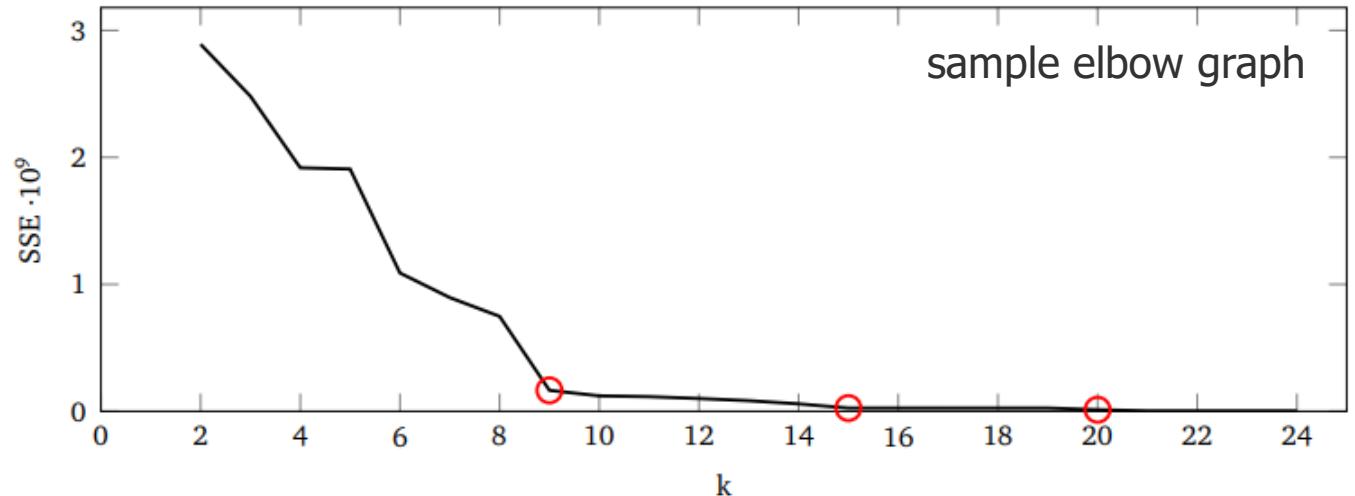
- Baseline: Execute clustering algorithm for all parameter values $k \in R$

How to limit search space?



Elbow method

1. Perform clustering for each parameter in R
2. Calculate SSE for each result
3. Create the elbow graph
4. Choose the *bend* ○ in the elbow graph (analyst)



LOG-Means

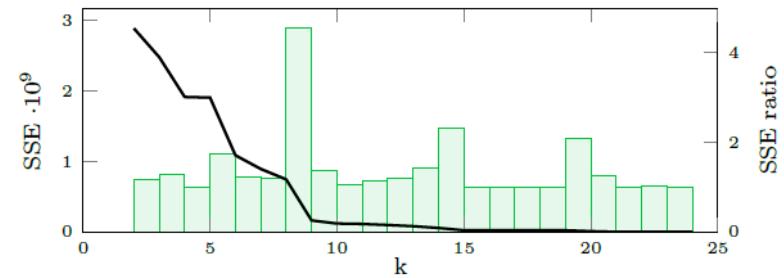
Intuition

- Elbow graph decreases in general with increasing values of $k \in R$
- The bend describes a “sudden drop” of the SSE

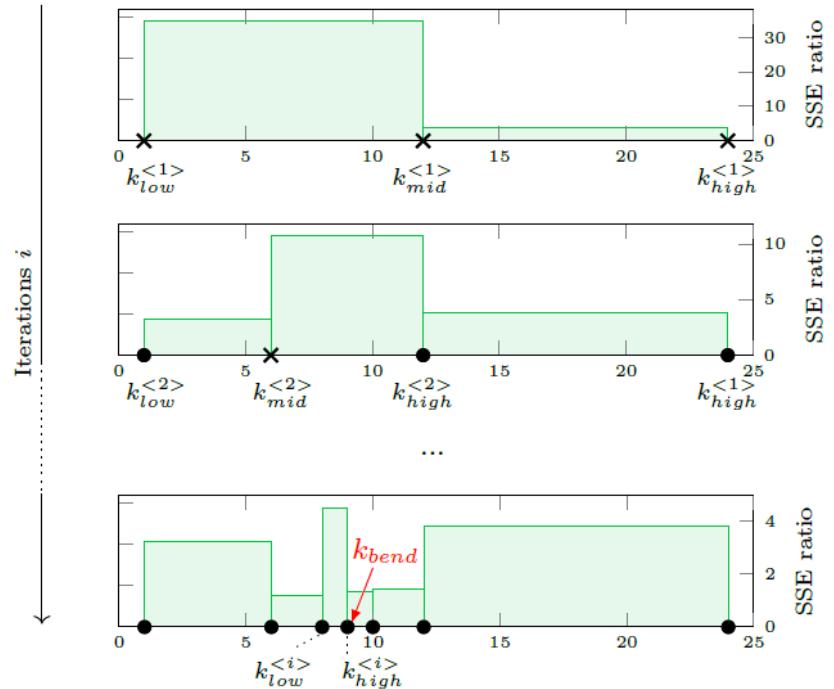
$$SSE\ ratio_k = \frac{SSE_{k-1}}{SSE_k}$$

- Use SSE ratio to compare areas between non-adjacent values for $k \in R$

 M. Fritz, M. Behringer, H. Schwarz: LOG-Means: Efficiently Estimating the Number of Clusters in Large Datasets. VLDB 2020



(a) Elbow graph with SSE ratio $\forall k \in \mathcal{R}$.



(b) Procedure of LOG-Means for $i = 1, 2$ and the last iteration.

LOG-Means

Intuition

- Elbow graph decreases in general with increasing values of $k \in R$
- The bend describes a “sudden drop” of the SSE

$$SSE\ ratio_k = \frac{SSE_{k-1}}{SSE_k}$$

- Use SSE ratio to compare areas between non-adjacent values for $k \in R$

 M. Fritz, M. Behringer, H. Schwarz: LOG-Means: Efficiently Estimating the Number of Clusters in Large Datasets. VLDB 2020

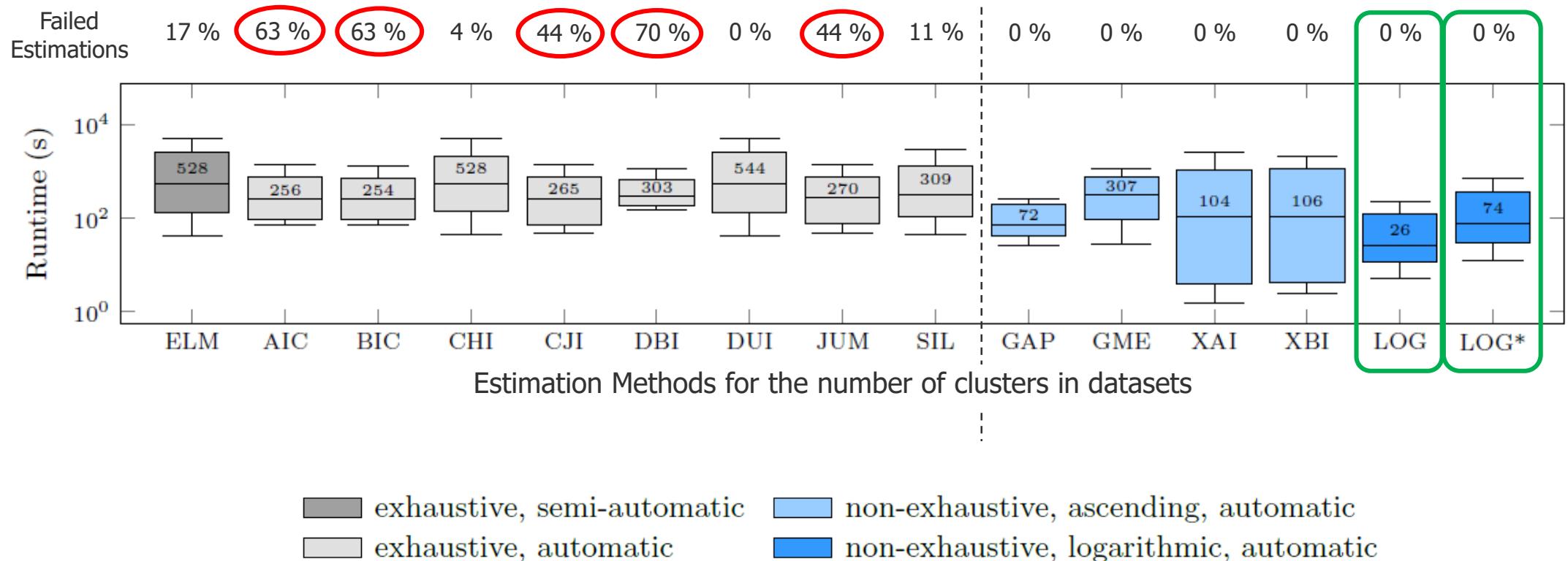
Input: \mathcal{X} - dataset, k_{low} - minimum number of desired clusters, k_{high} - maximum number of desired clusters, ε - number of neighbors to evaluate

Output: k_{est} - estimated number of clusters for \mathcal{X}

```
1  $k_{low} \leftarrow k_{low} - 1;$ 
2  $\mathcal{K} \leftarrow \emptyset;$ 
3  $\mathcal{M} \leftarrow \emptyset;$ 
4  $SSE_{low} \leftarrow$  SSE from  $k$ -Means with  $k_{low}$ ;
5  $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{low}, SSE_{low})\};$ 
6  $SSE_{high} \leftarrow$  SSE from  $k$ -Means with  $k_{high}$ ;
7  $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{high}, SSE_{high})\};$ 
8 while ( $k_{low}$  and  $k_{high}$  are not directly adjacent) {
9    $k_{mid} \leftarrow \lfloor(k_{high} + k_{low})/2\rfloor;$ 
10   $SSE_{mid} \leftarrow$  SSE from  $k$ -Means with  $k_{mid}$ ;
11   $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{mid}, SSE_{mid})\};$ 
12   $ratio_{left} \leftarrow SSE_{low}/SSE_{mid};$ 
13   $ratio_{right} \leftarrow SSE_{mid}/SSE_{high};$ 
14   $\mathcal{M} \leftarrow$  store or update  $\{(k_{mid}, ratio_{left})\}$ ;
15   $\mathcal{M} \leftarrow$  store or update  $\{(k_{high}, ratio_{right})\}$ ;
16   $k_{high} \leftarrow k$  with highest ratio from  $\mathcal{M}$ ;
17   $k_{low} \leftarrow$  left adjacent value of  $k_{high}$  from  $\mathcal{K}$ ;
18   $SSE_{high} \leftarrow$  SSE for  $k_{high}$  from  $\mathcal{K}$ ;
19   $SSE_{low} \leftarrow$  SSE for  $k_{low}$  from  $\mathcal{K}$ ;
20 }
21 if  $\varepsilon > 0$  then
22    $k_{bend} \leftarrow k \in [k_{low}, k_{high}]$  with highest ratio in  $\mathcal{M}$ ;
23    $k_{low} \leftarrow k_{bend} - \lfloor\varepsilon/2\rfloor;$ 
24    $k_{high} \leftarrow k_{bend} + \lfloor\varepsilon/2\rfloor;$ 
25   for ( $\forall k \in [k_{low}; k_{high}]$ ) {
26      $SSE_{k_{prev}} \leftarrow$  SSE of  $k_{prev}$  from  $\mathcal{K}$ ;
27     if  $k \in \mathcal{K}$  then
28        $SSE_k \leftarrow$  SSE for  $k$  from  $\mathcal{K}$ ;
29     else
30        $SSE_k \leftarrow$  SSE from  $k$ -Means with  $k$ ;
31        $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k, SSE_k)\};$ 
32     end
33      $ratio_k \leftarrow SSE_{k_{prev}}/SSE_k;$ 
34      $\mathcal{M} \leftarrow$  store or update  $\{(k, ratio_k)\}$ ;
35   }
36    $k_{est} \leftarrow k \in [k_{low}, k_{high}]$  with highest ratio in  $\mathcal{M}$ ;
37 return  $k_{est};$ 
```

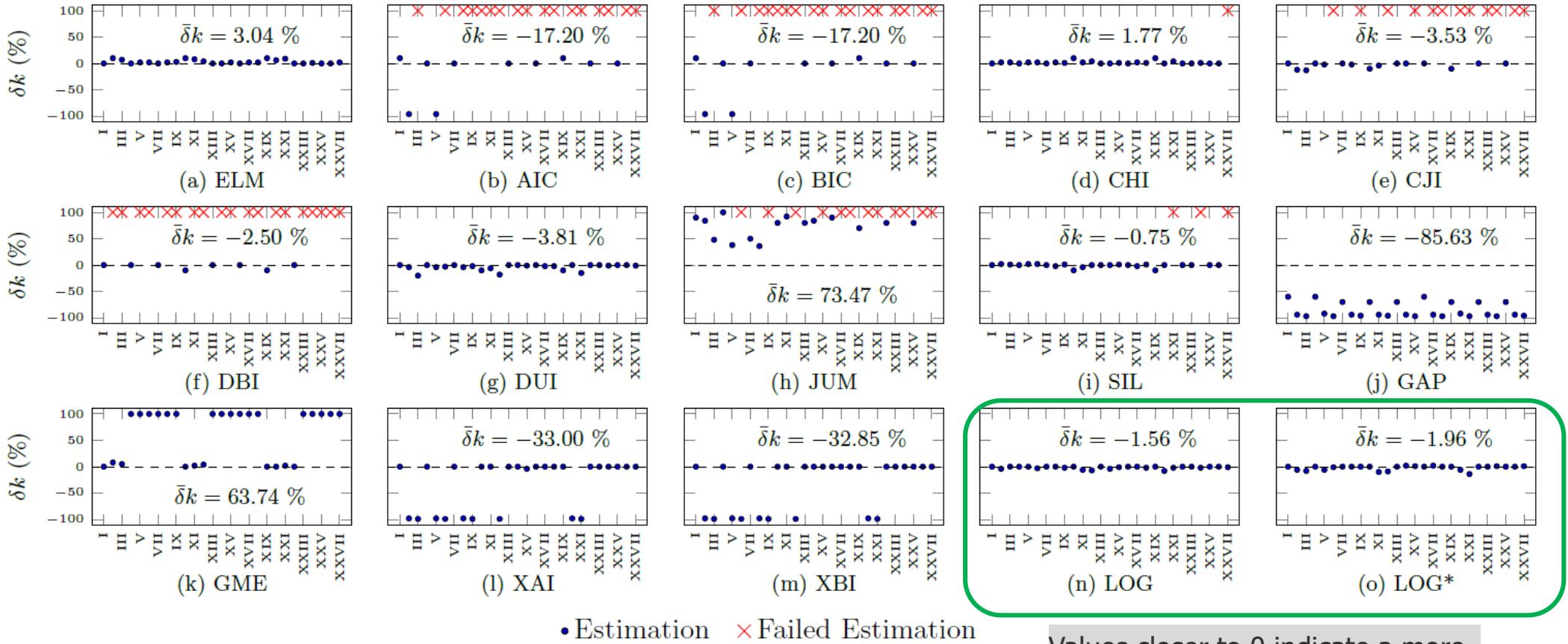
LOG-Means

Runtime Evaluation



LOG-Means

Accuracy Evaluation

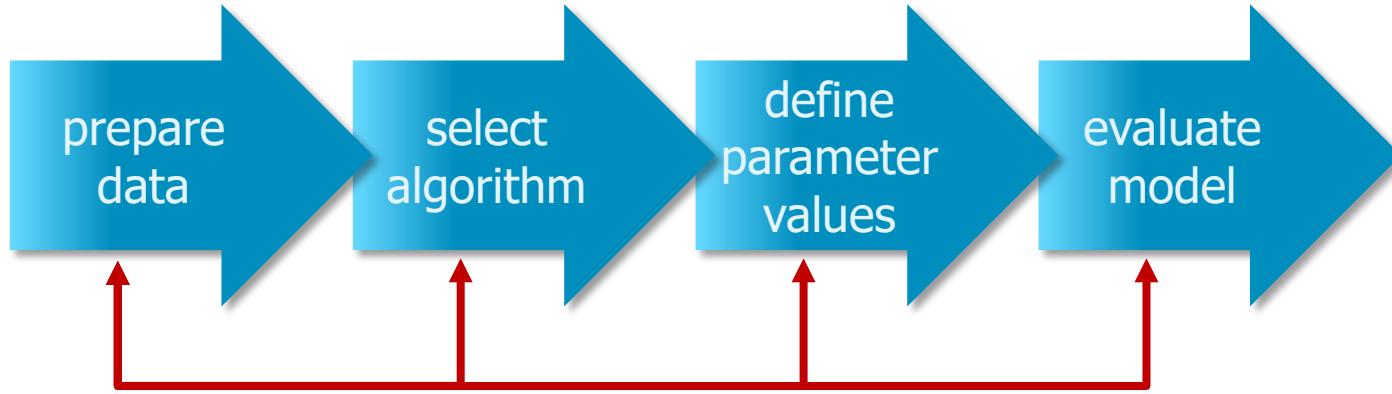


Values closer to 0 indicate a more accurate result and vice versa

Further improvements in:

M. Fritz, M. Behringer, D. Tschechlov, H. Schwarz: Efficient exploratory clustering analyses in large-scale exploration processes. VLDB Journal 2021

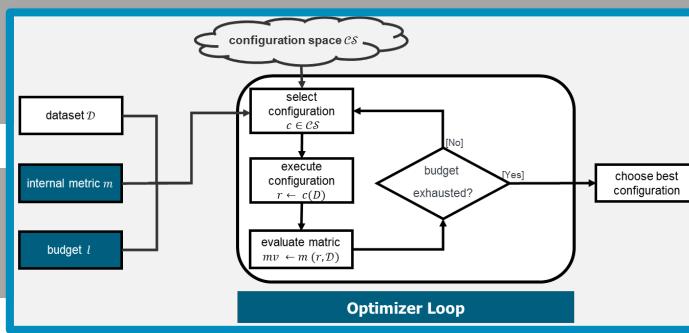
Agenda



Knowledge

(Domain experts, data scientists, ...)

Meta Learning



AutoML

3

AutoML
approach for
clustering

Categories of Clustering Techniques

hierarchical

- hierarchical decomposition of objects (top-down or bottom-up)
- spherical shape only
- Example algorithms: BIRCH, CURE



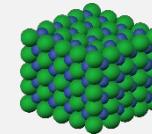
grid-based

- cluster on grid structure assigned to object space
- cluster shape aligned with grid structure
- Example alg.: STING, CLIQUE



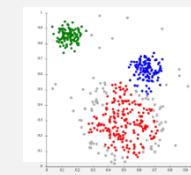
density-based

- grow cluster until a the density exceeds a threshold
- clusters of arbitrary shape
- allows to filter noise
- Example algorithms: DBSCAN, DENCLUE



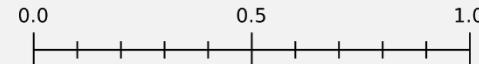
partitioning

- assign objects to k clusters
- stepwise improve cluster assignment
- spherical shape only
- Example alg.: **k-center clustering** like k-means



probabilistic model-based

- train a probabilistic model that assigns objects to clusters
- objects may belong to multiple clusters



AutoML4Clust

Motivation and Related Work



Goal: Address Combined algorithm selection and hyperparameter optimization for clustering

Hyperparameter Optimization for Clustering

- Exhaustive and non-exhaustive estimation methods

Algorithm Selection for Clustering

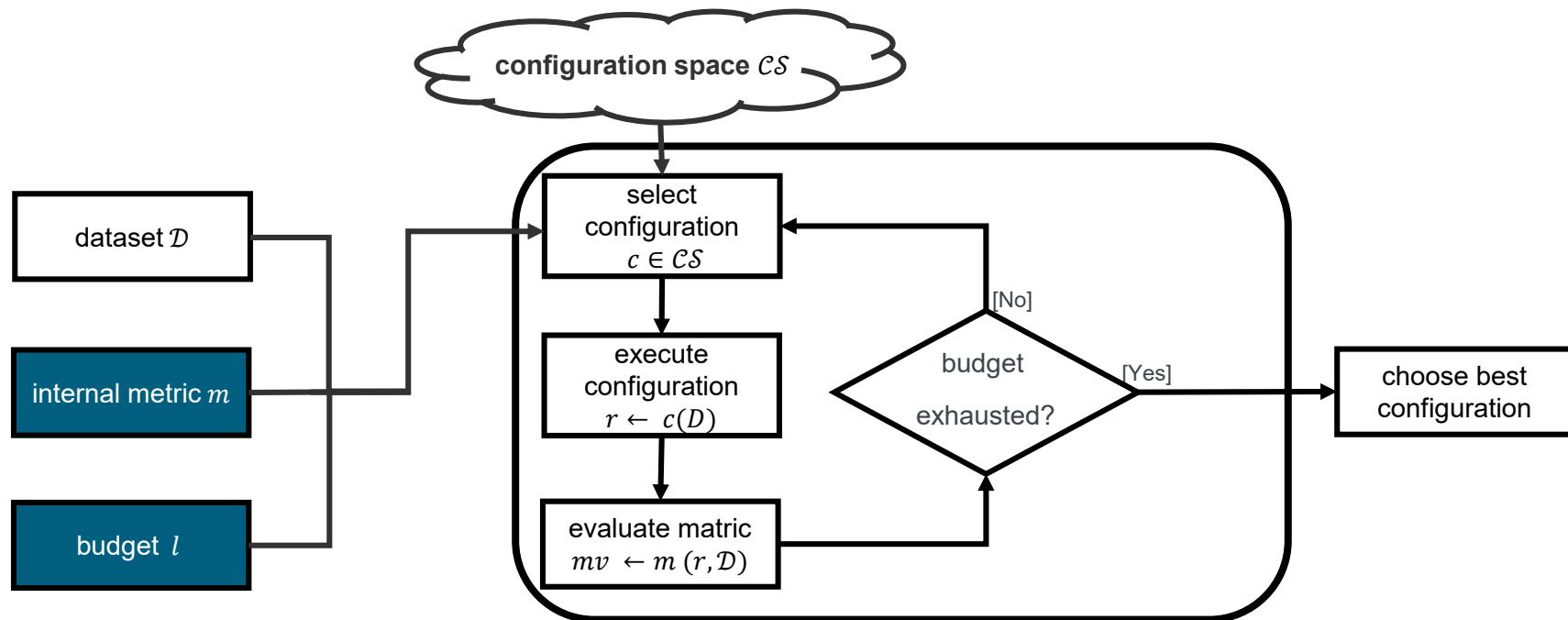
- Use meta-learning for selecting clustering algorithm [FC15, PC19]

AutoML Systems

- Auto-WEKA and Auto-sklearn
- Use hyperparameter optimization techniques [FKE15, THH13]



AutoML4Clust



\mathcal{CS} configuration space
(algorithm + hyperparameters)
 c configuration
 \mathcal{D} dataset
 m internal metric
 l budget
 r clustering result
 mv metric value

Optimizer Loop

D. Tschechlov, M. Fritz, H. Schwarz: AutoML4Clust: Efficient AutoML for Clustering Analyses. EDBT 2021

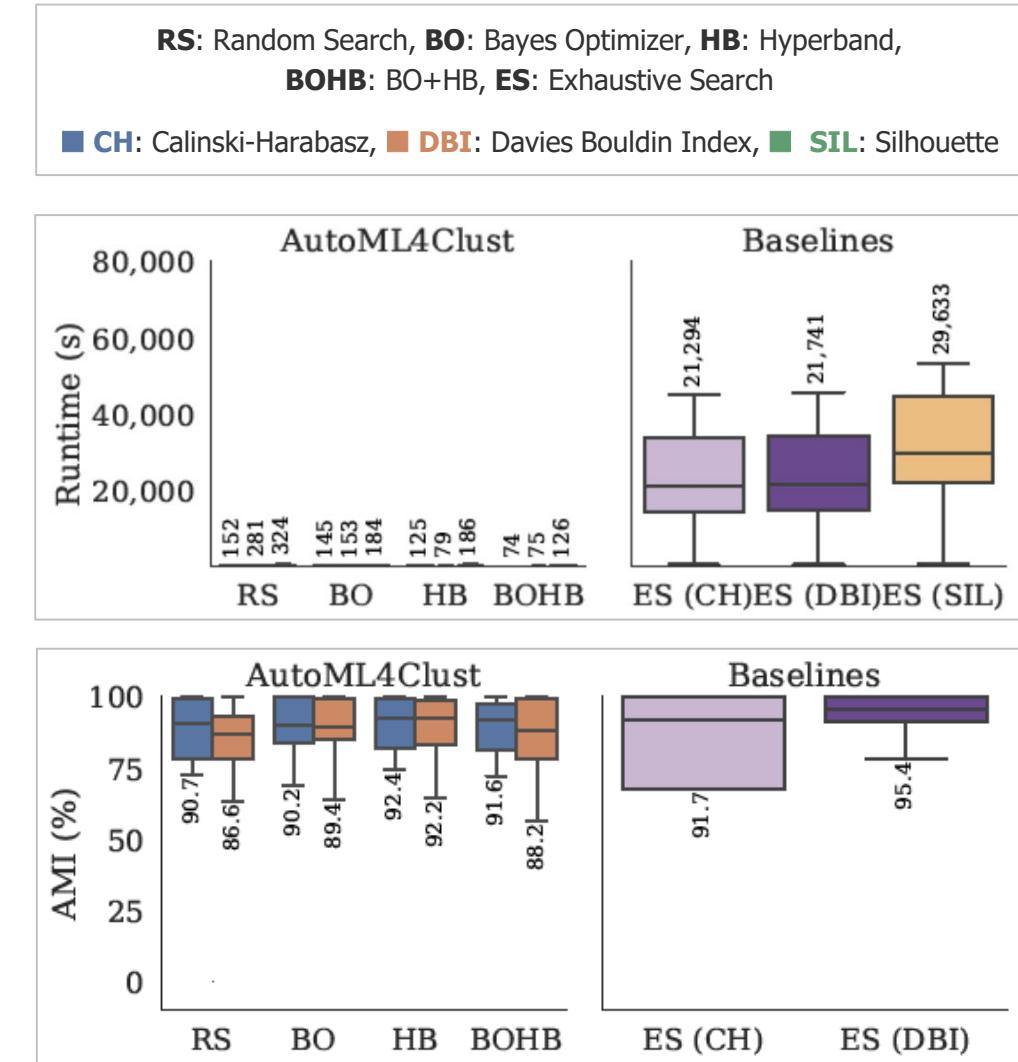
AutoML4Clust

Evaluation

- Setup
 - 4 different optimizers
 - 3 different internal metrics
 - accuracy measure: Adjusted Mutual Information (AMI)
- Results
 - efficient as it explores large configuration spaces greedily
 - generic regarding optimizer and metric instantiations
 - best instantiation with Hyperband optimizer and Calinski-Harabasz metric

speedups up to 437x

only 3% accuracy deviations



What's Next?

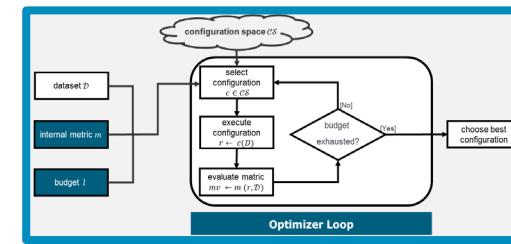
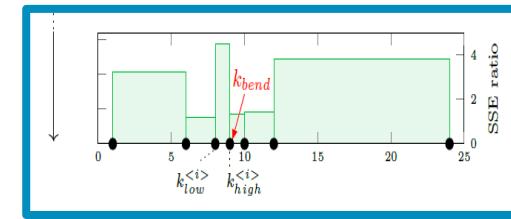
- Use **Meta Learning** to
 - ... decide on the most suitable internal metric depending on data characteristics
 - ... select promising configurations for the optimizer to start with
 - ... reduce the size of the configuration space

Conclusion

- Classification system for multi-class classification
 - exploiting domain knowledge helps to address issues with data characteristics (small training data set, class imbalance, ...)
- LogMeans
 - method to efficiently determine the number of clusters in datasets
- AutoML4Clust
 - applying AutoML approaches to clustering problems works

Series			
HDE	[N: 277] [C: 60 S: 52]	MDE	
OM470	[N: 96] [C: 37 S: 43]	OM471	[N: 130] [C: 44 S: 52]
471900	[N: 52] [C: 26 S: 48]	OM934	[N: 200] [C: 43 S: 51]
471902	[N: 12] [C: 9 S: 41]	OM936	[N: 573] [C: 54 S: 59]
936980	[N: 309] [C: 43 S: 57]	936910	[N: 61] [C: 24 S: 44]
936910	[N: 10] [C: 10 S: 10]

MDE: Medium-Duty Engine | HDE: Heavy-Duty Engine
■ Product group ■ Product level



Thank you for your attention!



- [CCK00] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth: CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS Inc., August, 2000.
- [EGH22] R. Eichler, C. Gröger, E. Hoos, H. Schwarz, B. Mitschang: Data Shopping - How an Enterprise Data Marketplace Supports Data Democratization in Companies. In: J. D. Weerdt, A. Polyvyanyy (Eds.): Intelligent Information Systems - CAiSE Forum 2022, Leuven, Belgium, June 6-10, 2022, Proceedings, Springer, Vol. 452, 2022.
- [FBS20] M. Fritz, M. Behringer, H. Schwarz: LOG-Means: Efficiently Estimating the Number of Clusters in Large Datasets. In: Proc. VLDB Endow., Vol. 13, No. 11, 2020.
- [FC15] Daniel G Ferrari and Leandro Nune de Castro. 2015. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. *Information Sciences* (2015).
- [FEK15] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in neural information processing systems* (2015).
- [HRM19] Hirsch, Vitali; Reimann, Peter; Mitschang, Bernhard: Data-Driven Fault Diagnosis in End-of-Line Testing of Complex Products. In: *Proceedings of the 6th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2019)*, Washington, D.C., USA.
- [HRM20] V. Hirsch, P. Reimann, B. Mitschang: Exploiting Domain Knowledge to Address Multi-Class Imbalance and a Heterogeneous Feature Space in Classification Tasks for Manufacturing Data. *VLDB 2020*



- [PC19] Bruno Almeida Pimentel and André C.P.L.F. de Carvalho. 2019. A new data characterization for selecting clustering algorithms using meta-learning. *Information Sciences* (3 2019).
- [Rue21] L. von Rueden *et al.*, "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," in *IEEE Transactions on Knowledge and Data Engineering* (2021)
- [TFS21] D. Tschechlov, M. Fritz, H. Schwarz: AutoML4Clust: Efficient AutoML for Clustering Analyses. EDBT 2021
- [THH13] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13. ACM Press.
- [Tho53] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267-276, 12 1953.