Universität Stuttgart

IPVS Institut für Parallele und Verteilte Systeme

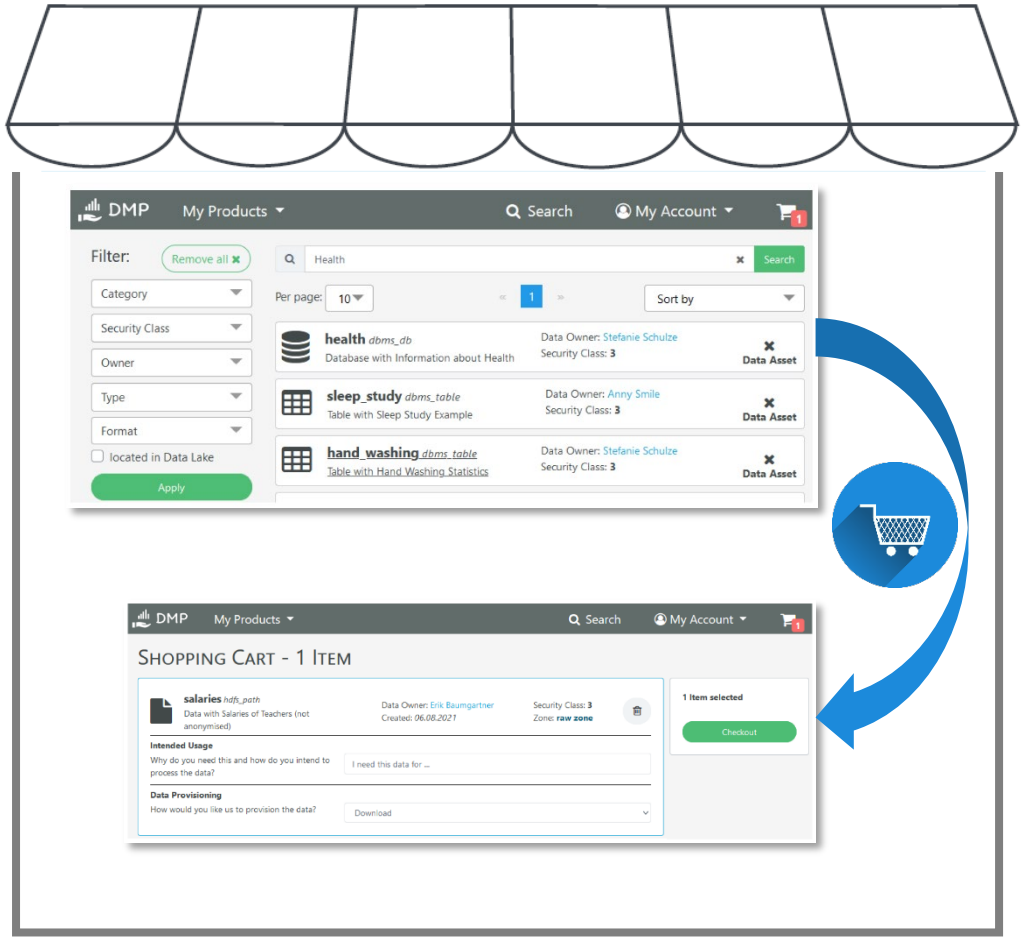
Data for Analytics

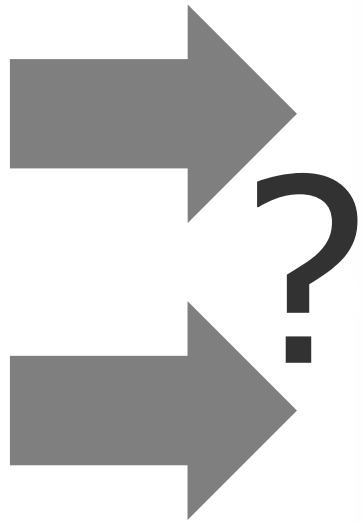
Domain knowledge and AutoML in analytic processes

Holger Schwarz

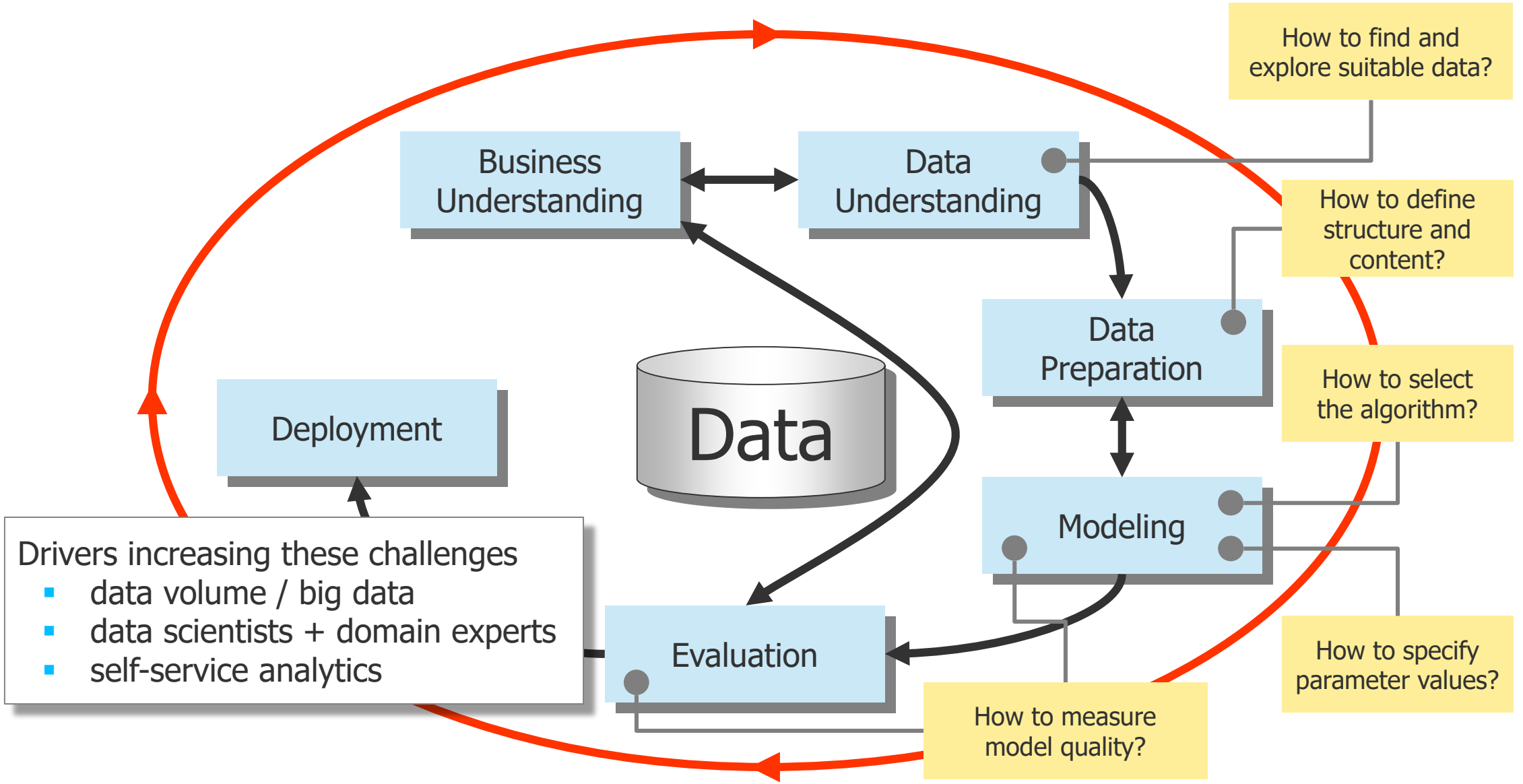
work of/with Manuel Fritz, Vitali Hirsch, Bernhard Mitschang, Peter Reimann, Dennis Tschechlov, …

Data Shopping in Data Marketplaces is only one Step …



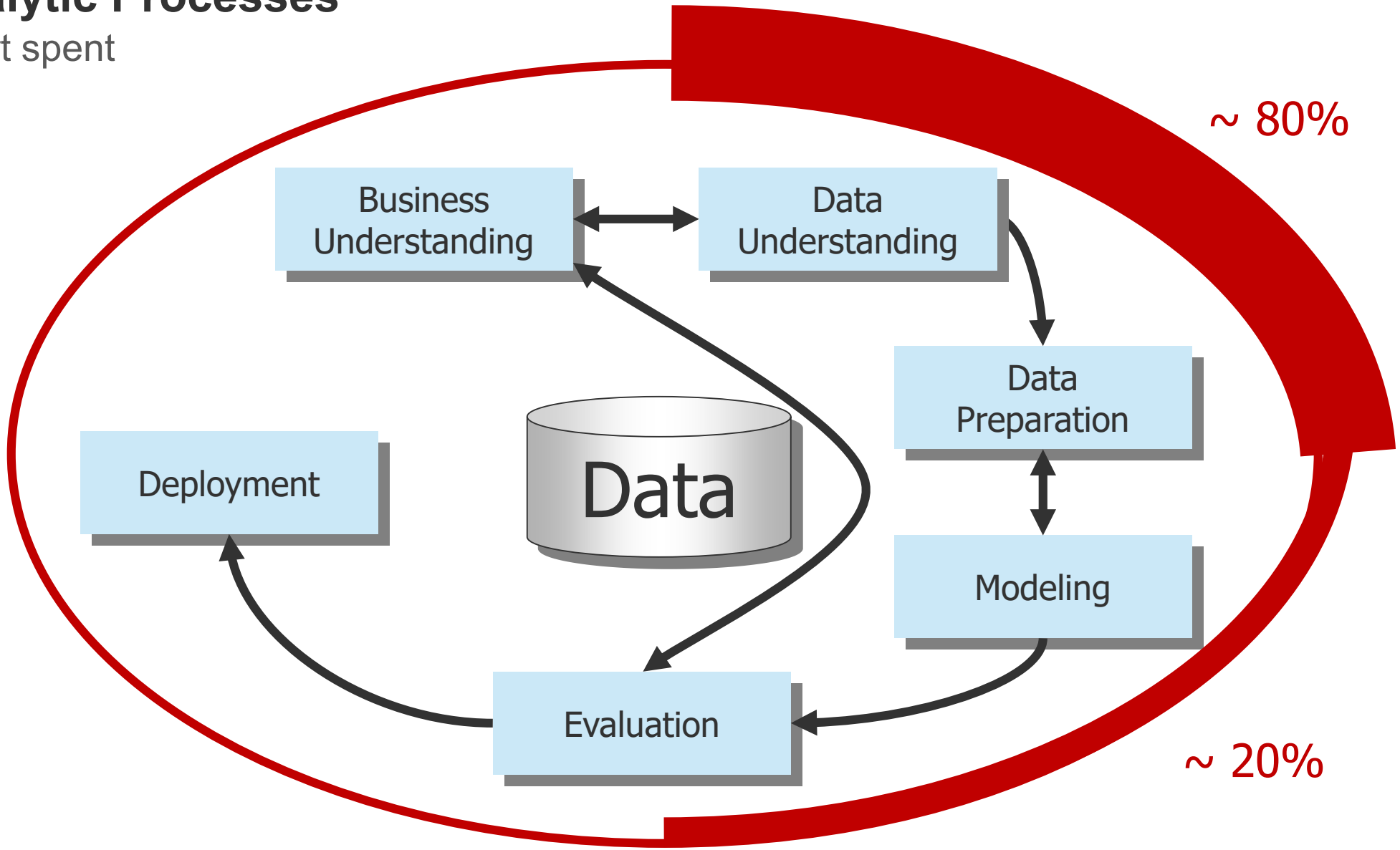


Challenges in Analytic Processes



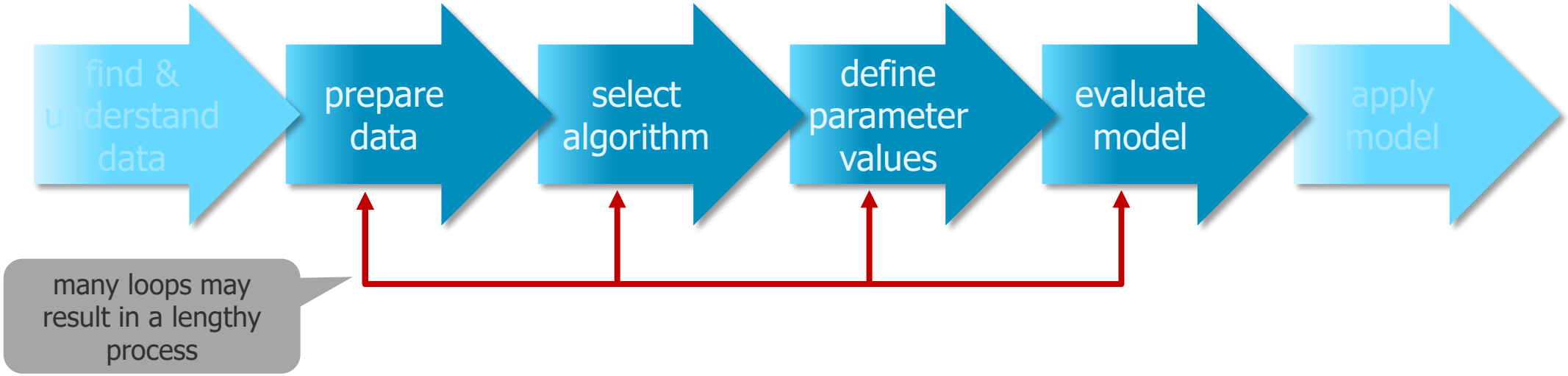
Analytic Processes

Effort spent



Analytic Processes

Focus and categories of approaches



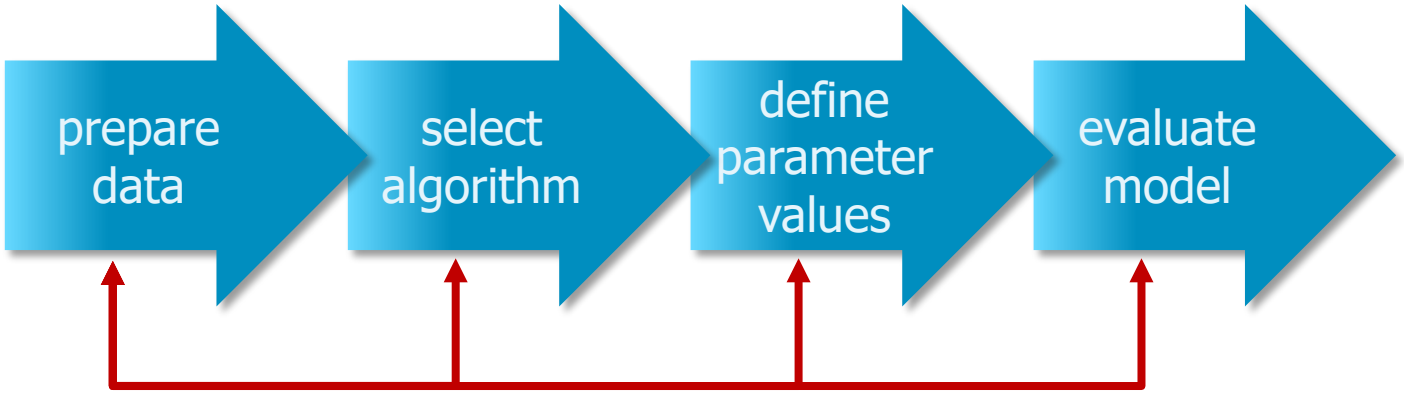
Knowledge
(Domain experts, data scientists, …)

Meta Learning

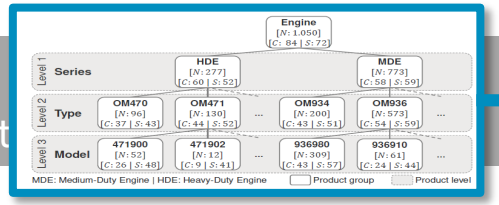
AutoML

user involvement

Agenda



Knowledge
(Domain experts, dat

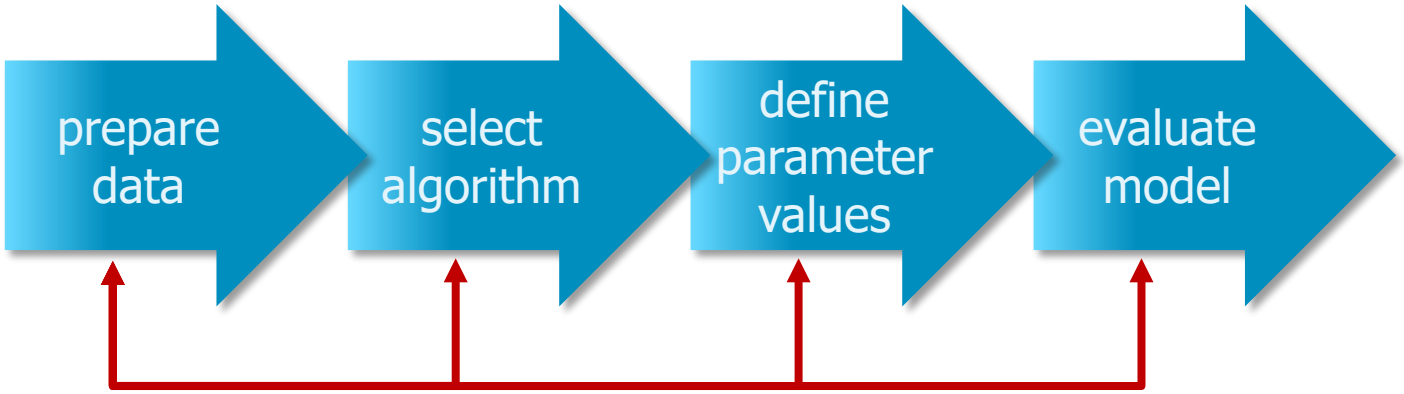


Meta Learning

AutoML

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

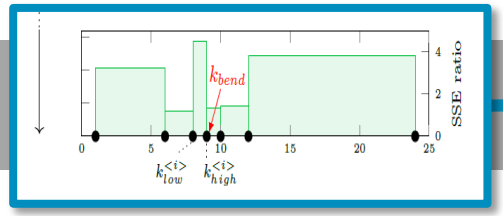
Agenda



Knowledge
(Domain experts, data scientists, …)

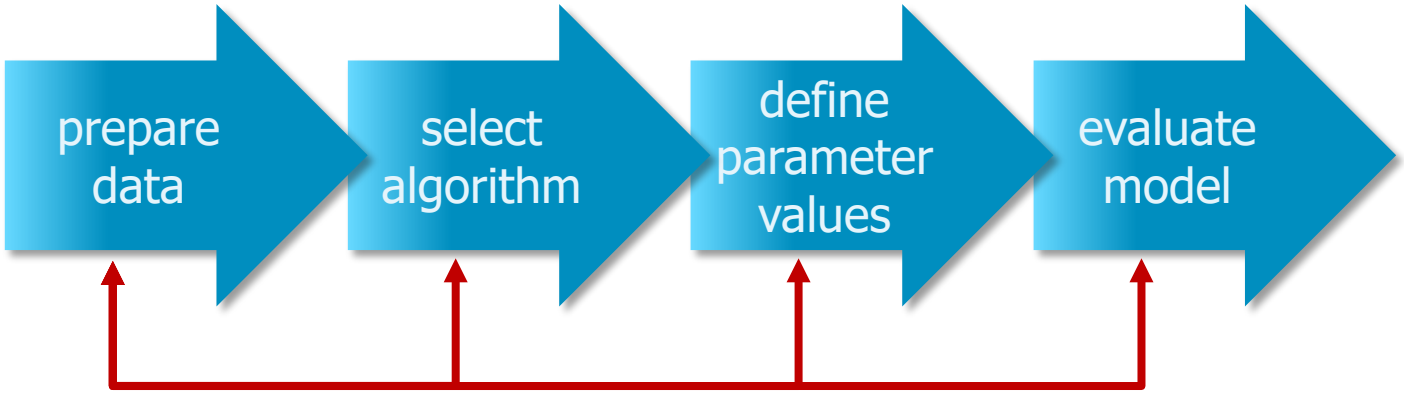
Meta Learning

AutoML



2 method to efficiently determine the number of clusters in datasets

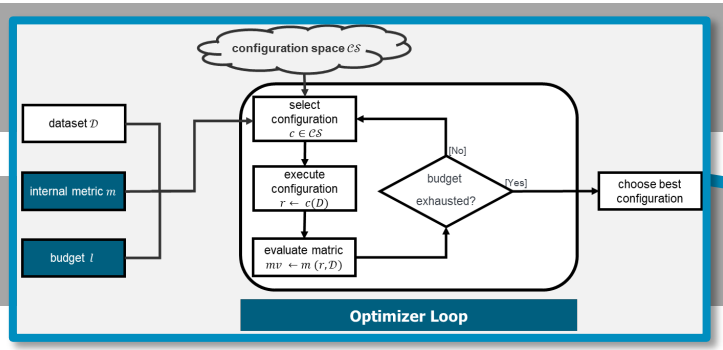
Agenda



Knowledge
(Domain experts, data scientists, …)

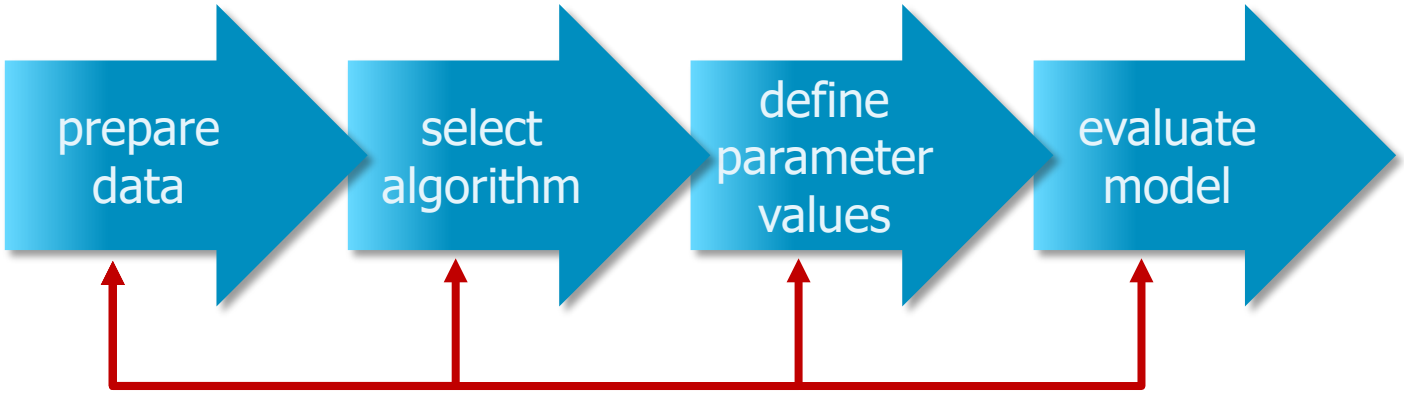
Meta Learning

AutoML

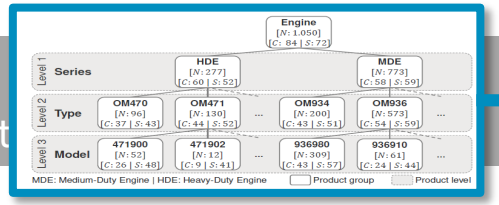


3 AutoML approach for clustering

Agenda



Knowledge
(Domain experts, dat



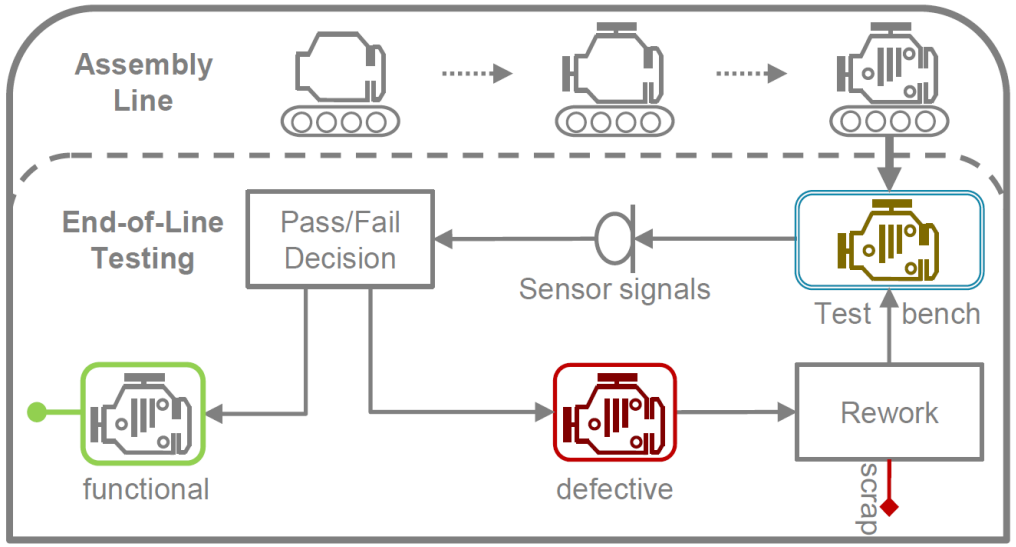
Meta Learning

AutoML

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

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

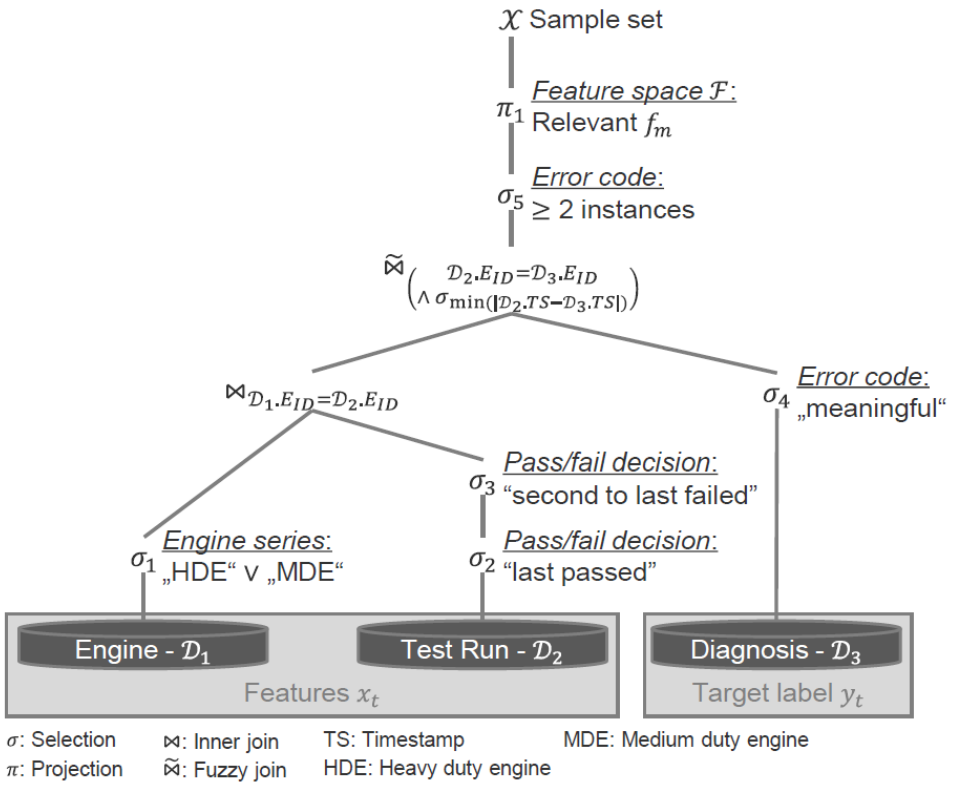


→ data driven approach desirable

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

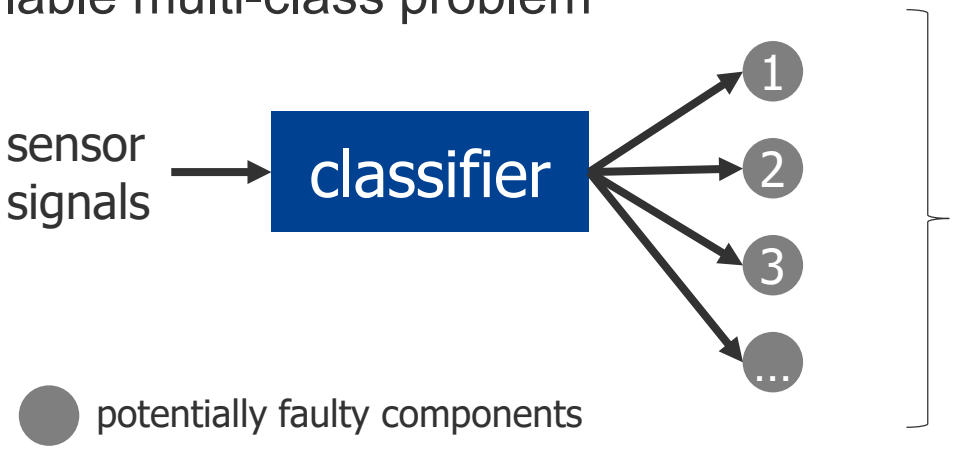
Set	Name	Topic	Sample Feature	N	M
$\mathcal{D}_1$	Engine	Spec sheet	-Engine type -Engine design	519,214	22
$\mathcal{D}_2$	Test bench	Test runs	-Sensors s_k	621,689	383
$\mathcal{D}_3$	Diagnostic	Faults	-Error code y_t	20,631	77



Data-driven Fault Isolation

- Fault detection based on binary classification

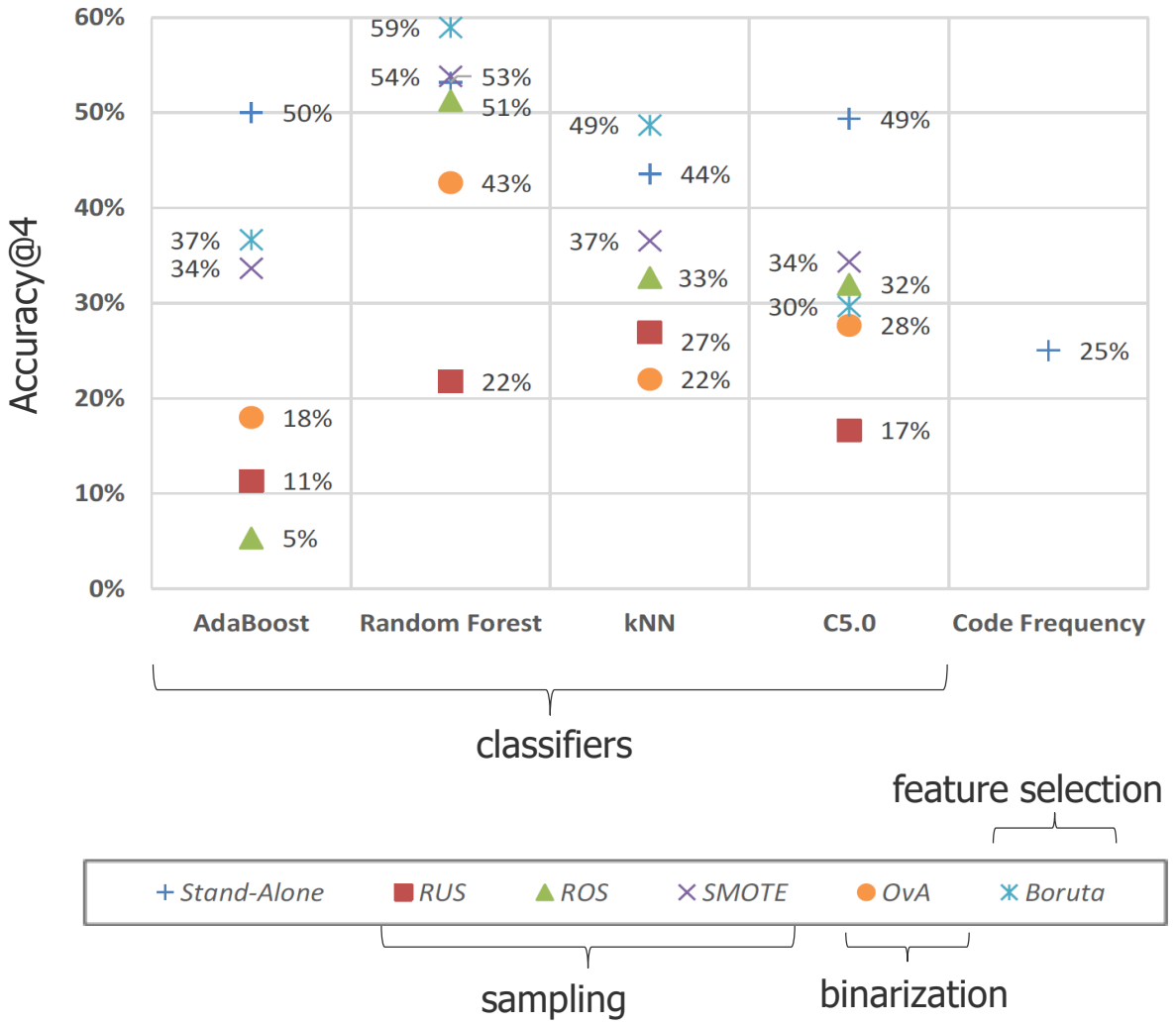
- Fault isolation considered as single-lable multi-class problem



- Recommendation list of length p sorted by confidence values

[3 0.8; 1 0.65; 5 0.5; 2 0.45; …]

Evaluation Results



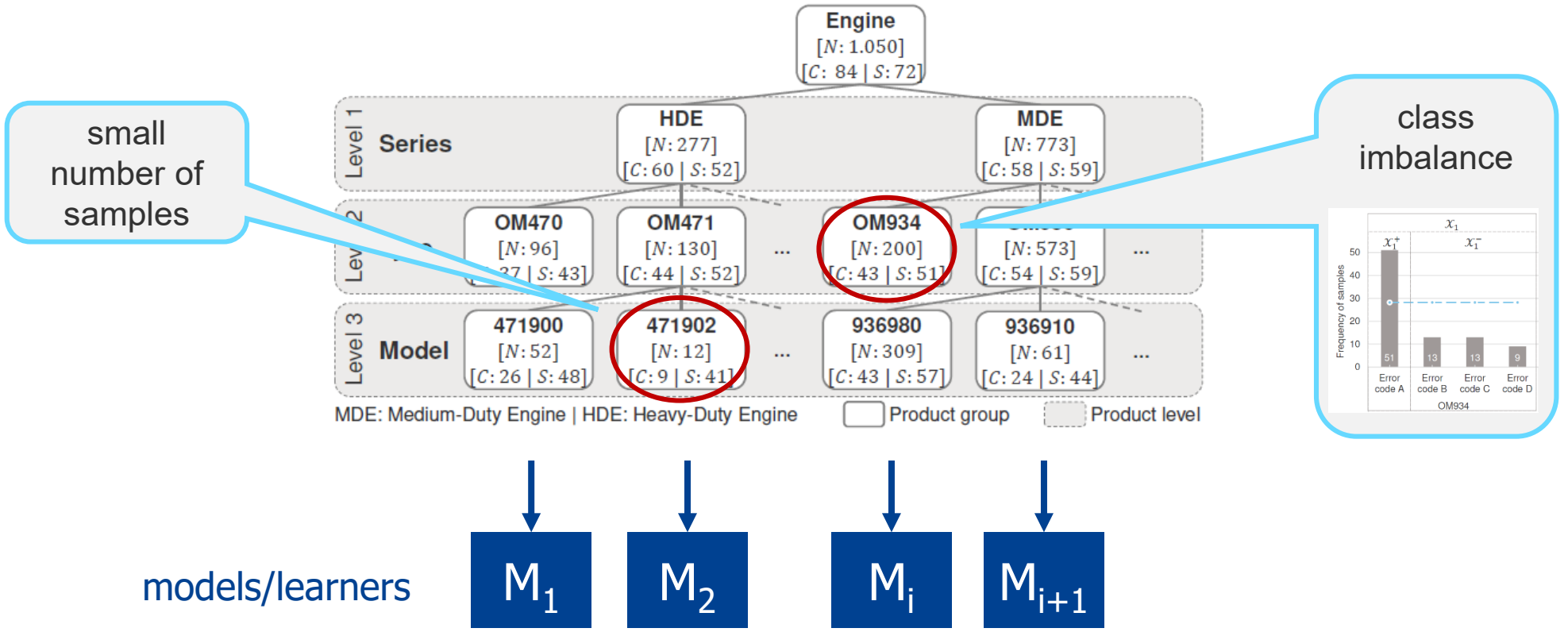
- Improvements of accuracy compared to baseline possible
- Leads to reduced number of rework attempts
- Achieved accuracy depends on
  - selected algorithm
  - type of data preparation

→ how to exploit domain knowledge?

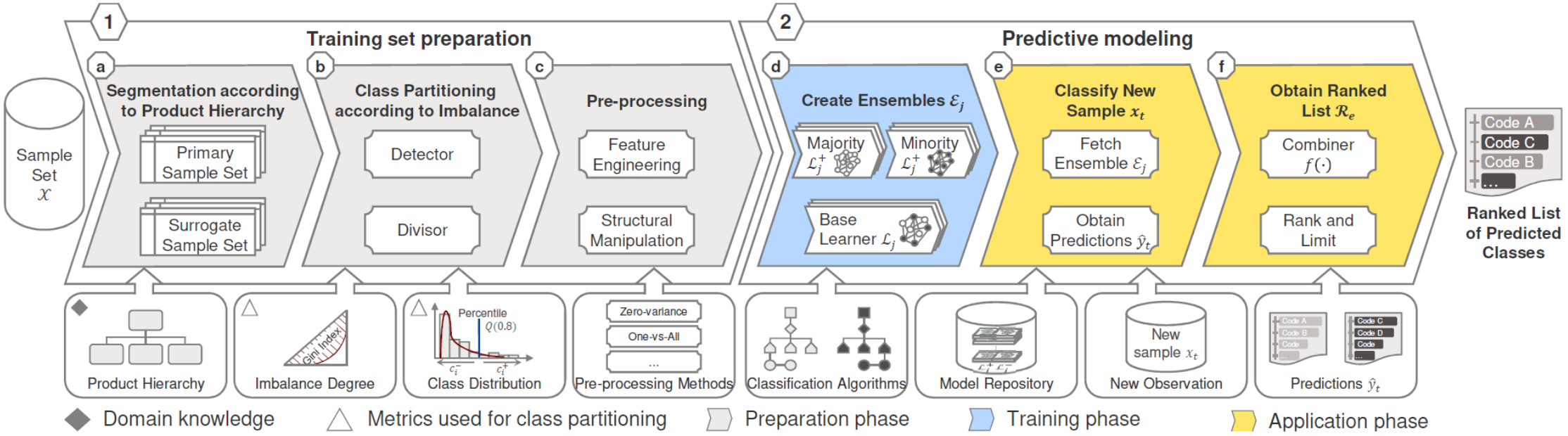
📖 V. Hirsch, P. Reimann, B. Mitschang: Data-Driven Fault Diagnosis in End-of-Line Testing of Complex Products. DSAA 2019

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

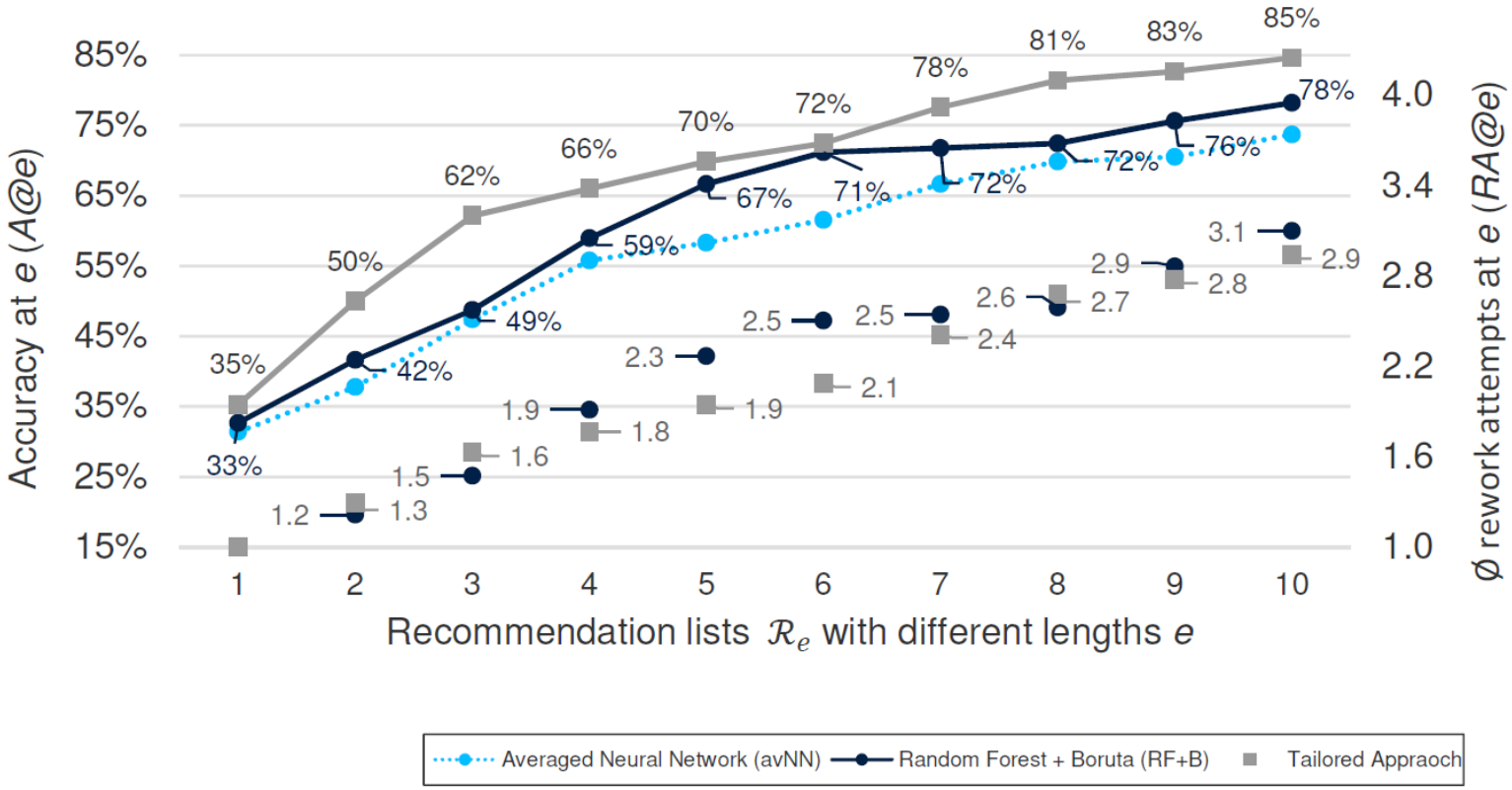


Classification System



📖 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

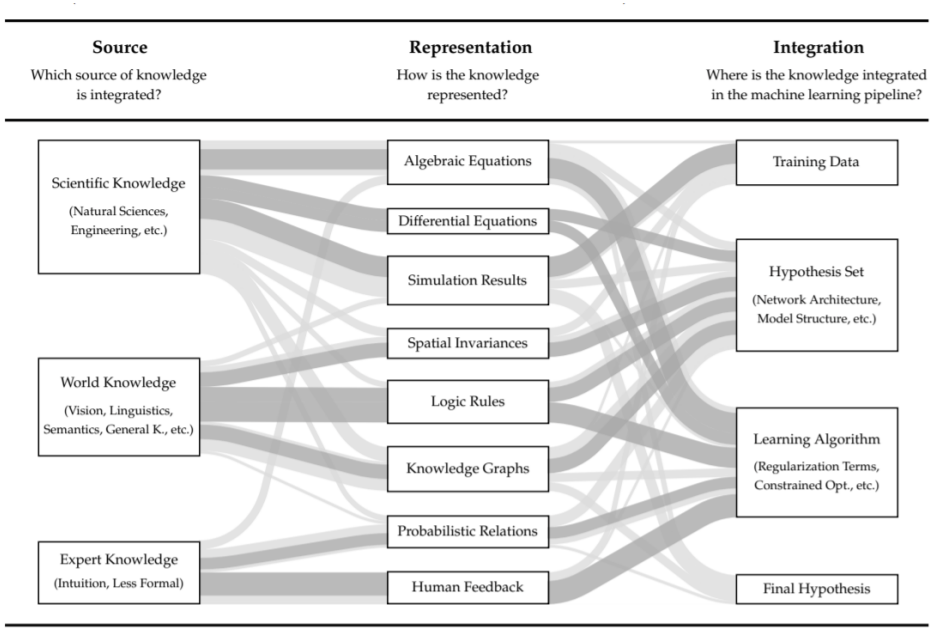


What‘s next?

- 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?

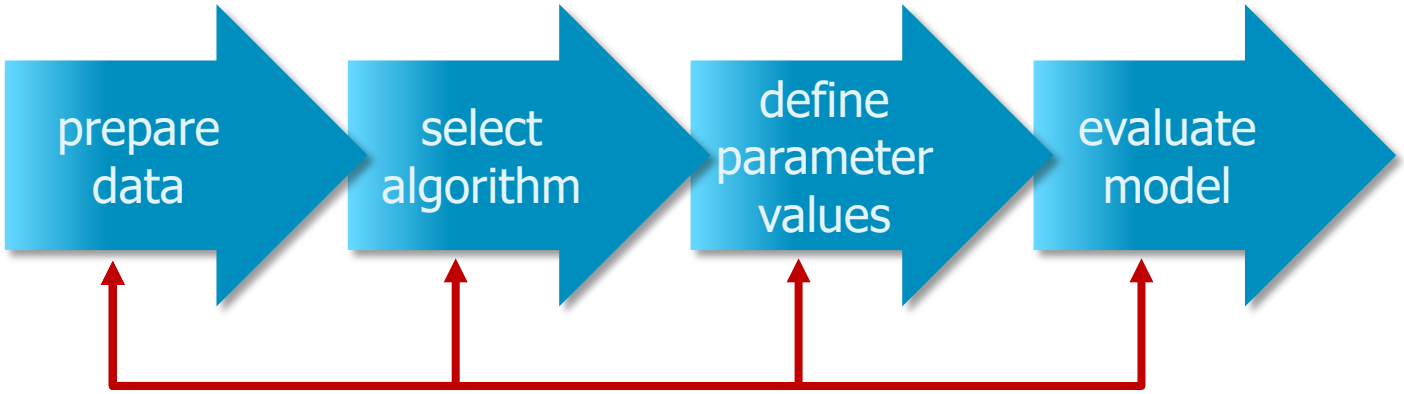
- Use clustering in data preparation
  - can we achieve similar effects by automatically cluster the training data?
  - how can we exploit domain knowledge in constraint-based clustering?

- Systematically exploit further domain knowledge



[Rue21]

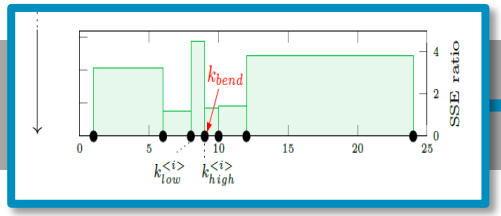
Agenda



Knowledge
(Domain experts, data scientists, …)

Meta Learning

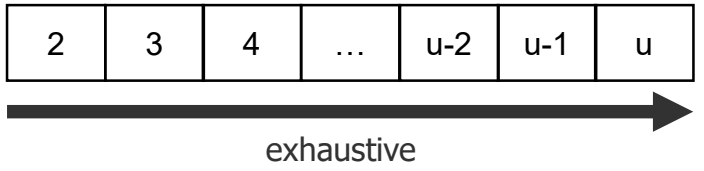
AutoML

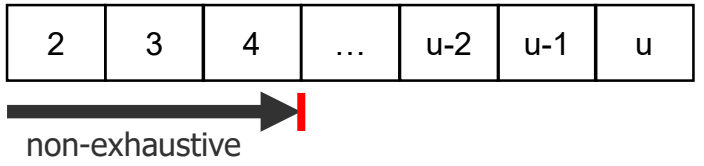


2 method to efficiently determine the number of clusters in datasets

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





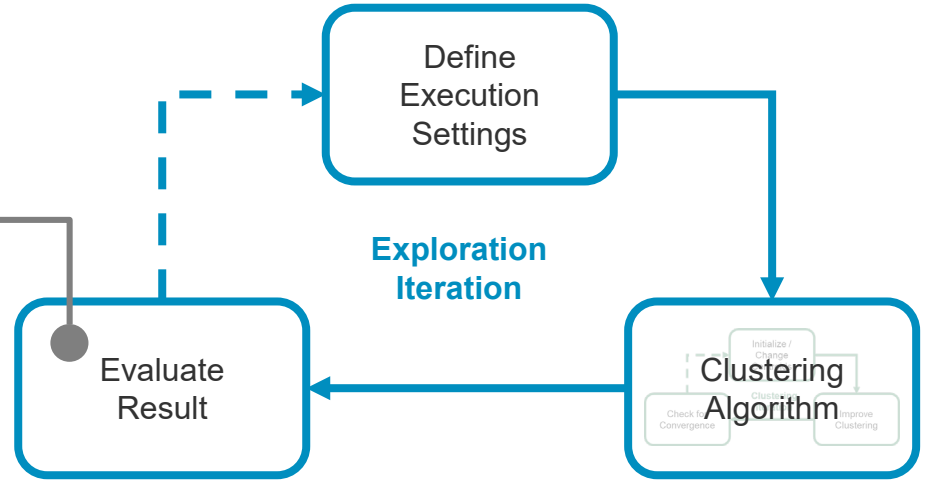
- Available estimation methods

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

LOG-Means

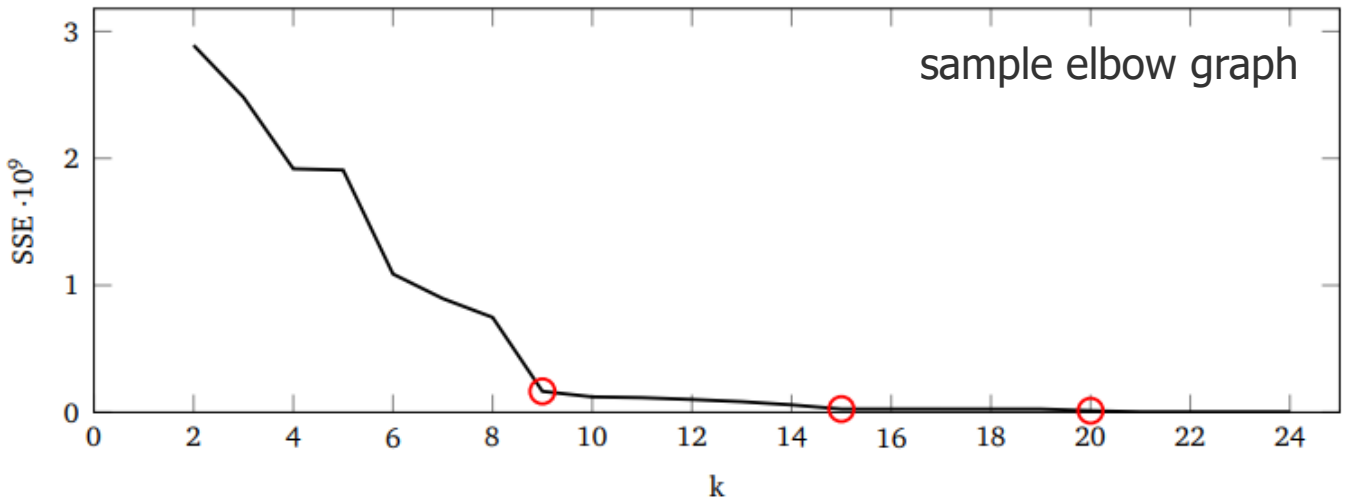
- Baseline: Execute clustering algorithm for all parameter values $k \in R$

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)



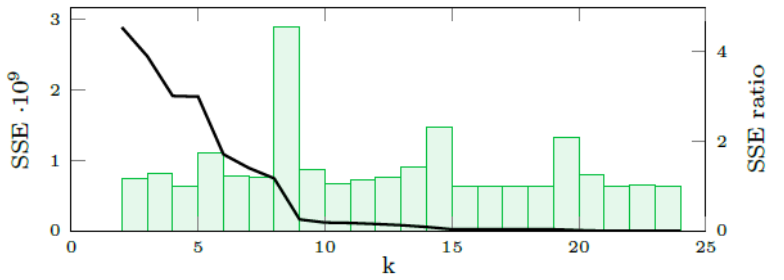
LOG-Means

Intuition

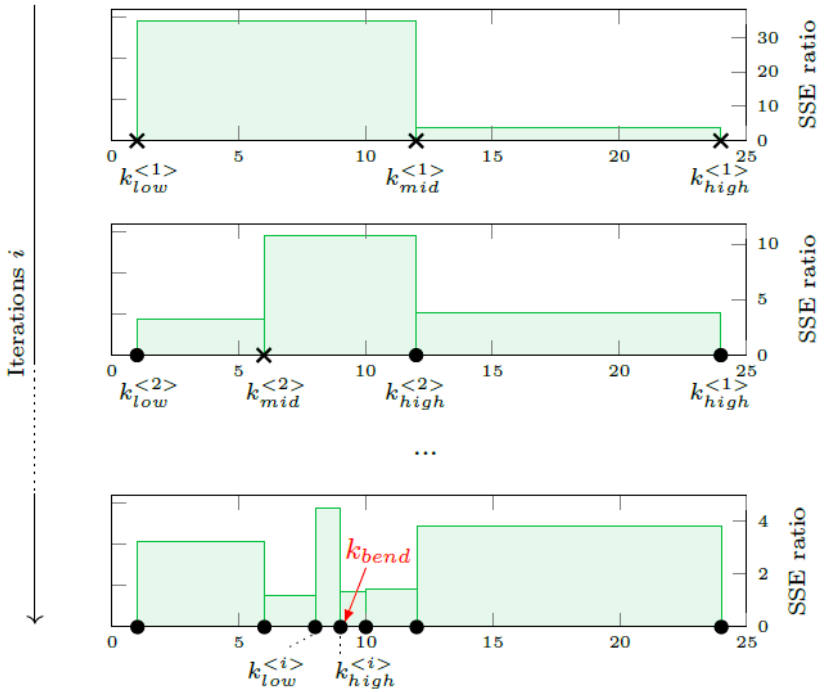
- Elbow graph decreases in general with increasing values of $k \in R$
- The bend describes a “sudden drop“ of the SSE

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

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



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



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

SSE SSE ratio × Executing ● Already executed

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

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 $k \in R$

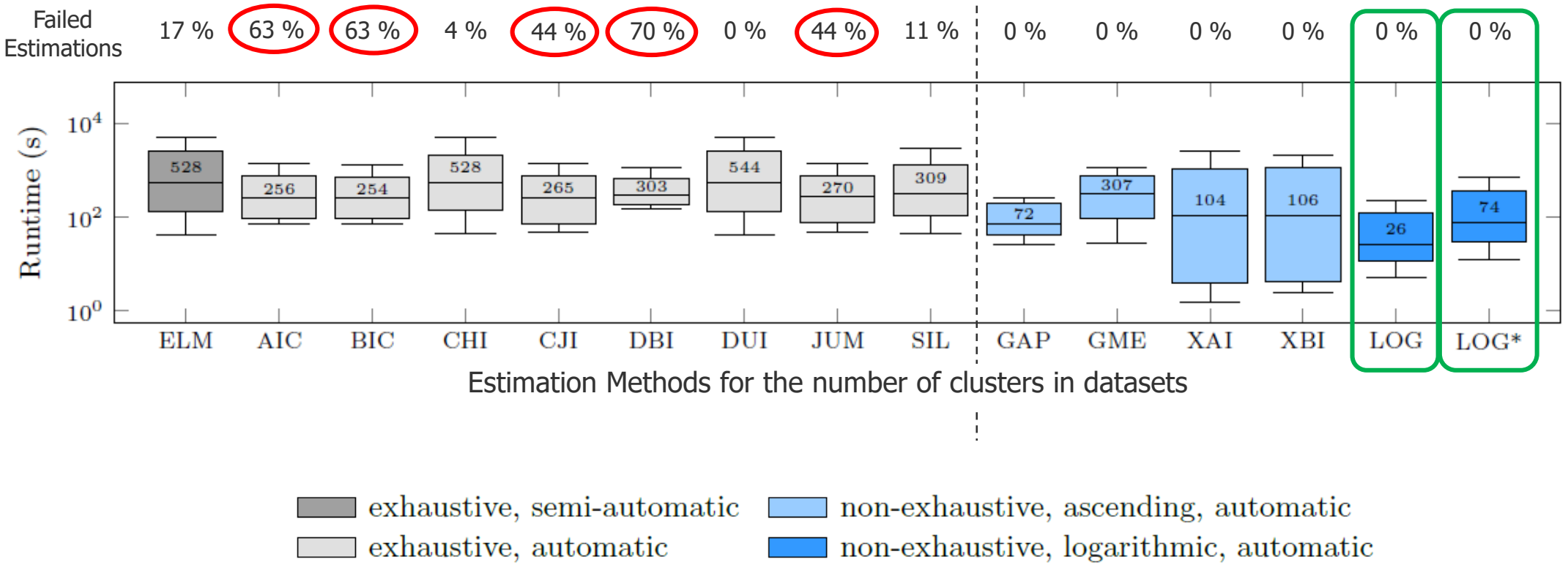
📖 M. Fritz, M. Behringer, H. Schwarz: LOG-Means: Efficiently Estimating the Number of Clusters in Large Datasets. VLDB 2020

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 $\mathcal{X}$

$k_{low} \leftarrow k_{low} - 1$;
$\mathcal{K} \leftarrow \emptyset$;
$\mathcal{M} \leftarrow \emptyset$;
$SSE_{low} \leftarrow$ SSE from k-Means with k_{low};
$\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{low}, SSE_{low})\}$;
$SSE_{high} \leftarrow$ SSE from k-Means with k_{high};
$\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{high}, SSE_{high})\}$;
while (k_{low} *and* k_{high} *are not directly adjacent*) {
  $k_{mid} \leftarrow \lfloor (k_{high} + k_{low})/2 \rfloor$;
  $SSE_{mid} \leftarrow$ SSE from k-Means with k_{mid};
  $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k_{mid}, SSE_{mid})\}$;
  $ratio_{left} \leftarrow SSE_{low}/SSE_{mid}$;
  $ratio_{right} \leftarrow SSE_{mid}/SSE_{high}$;
  $\mathcal{M} \leftarrow$ store or update $\{(k_{mid}, ratio_{left})\}$;
  $\mathcal{M} \leftarrow$ store or update $\{(k_{high}, ratio_{right})\}$;
  $k_{high} \leftarrow k$ with highest ratio from $\mathcal{M}$;
  $k_{low} \leftarrow$ left adjacent value of k_{high} from $\mathcal{K}$;
  $SSE_{high} \leftarrow$ SSE for k_{high} from $\mathcal{K}$;
  $SSE_{low} \leftarrow$ SSE for k_{low} from $\mathcal{K}$;
}
if $\varepsilon > 0$ **then**
  $k_{bend} \leftarrow k \in [k_{low}, k_{high}]$ with highest ratio in $\mathcal{M}$;
  $k_{low} \leftarrow k_{bend} - \lfloor \varepsilon/2 \rfloor$;
  $k_{high} \leftarrow k_{bend} + \lfloor \varepsilon/2 \rfloor$;
  for ($\forall k \in [k_{low}; k_{high}]$) {
    $SSE_{k_{prev}} \leftarrow$ SSE of k_{prev} from $\mathcal{K}$;
    if $k \in \mathcal{K}$ **then**
      $SSE_k \leftarrow$ SSE for k from $\mathcal{K}$;
    else
      $SSE_k \leftarrow$ SSE from k-Means with k;
      $\mathcal{K} \leftarrow \mathcal{K} \cup \{(k, SSE_k)\}$;
    end
    $ratio_k \leftarrow SSE_{k_{prev}}/SSE_k$;
    $\mathcal{M} \leftarrow$ store or update $\{(k, ratio_k)\}$;
  }
$k_{est} \leftarrow k \in [k_{low}, k_{high}]$ with highest ratio in $\mathcal{M}$;
return k_{est};

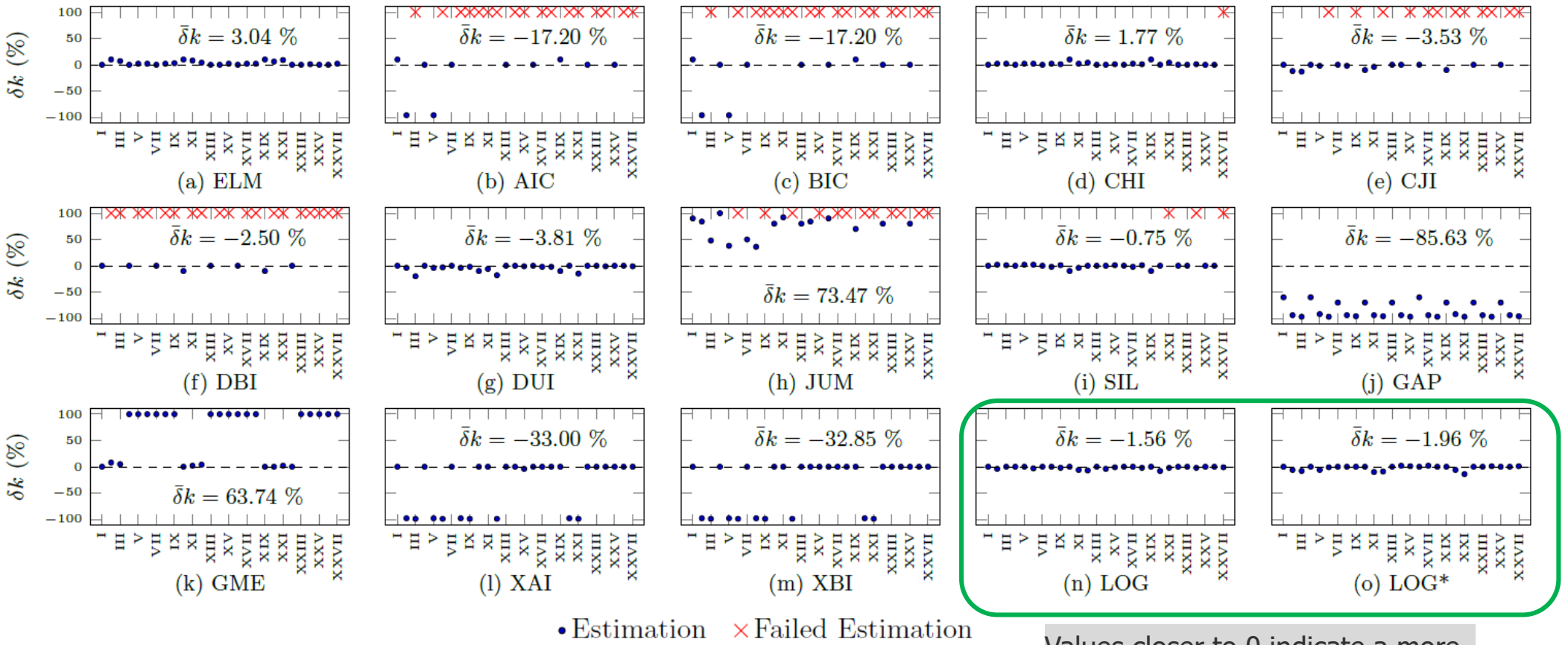
LOG-Means

Runtime Evaluation



LOG-Means

Accuracy Evaluation

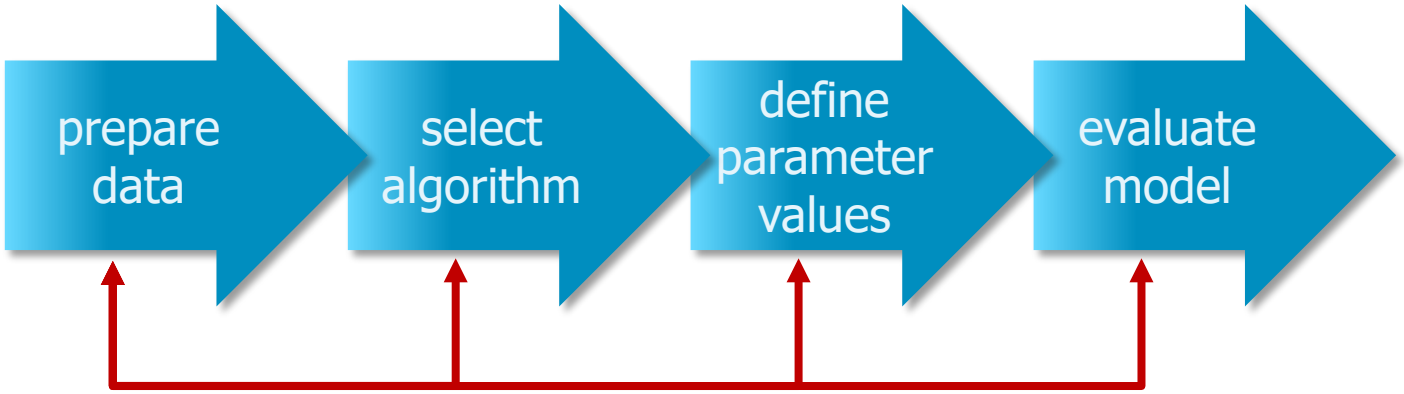


Values closer to 0 indicate a more accurate result and vice versa

Further improvements in:

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

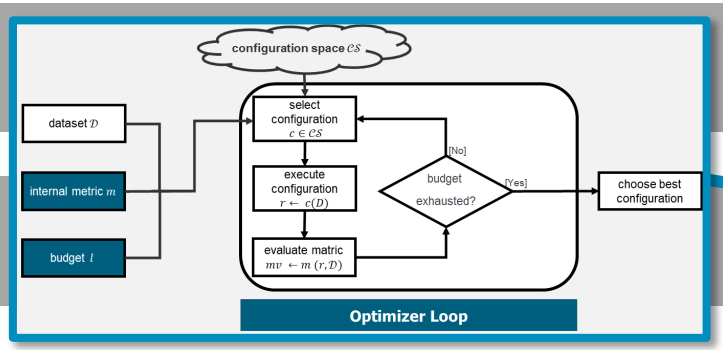
Agenda



Knowledge
(Domain experts, data scientists, …)

Meta Learning

AutoML

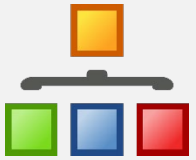


3 AutoML approach for clustering

Categories of Clustering Techniques

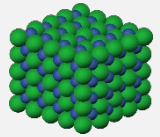
hierarchical

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



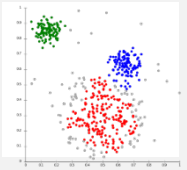
density-based

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



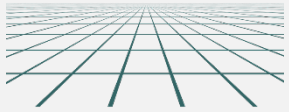
partitioning

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



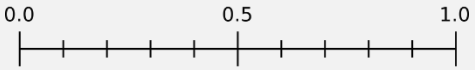
grid-based

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



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

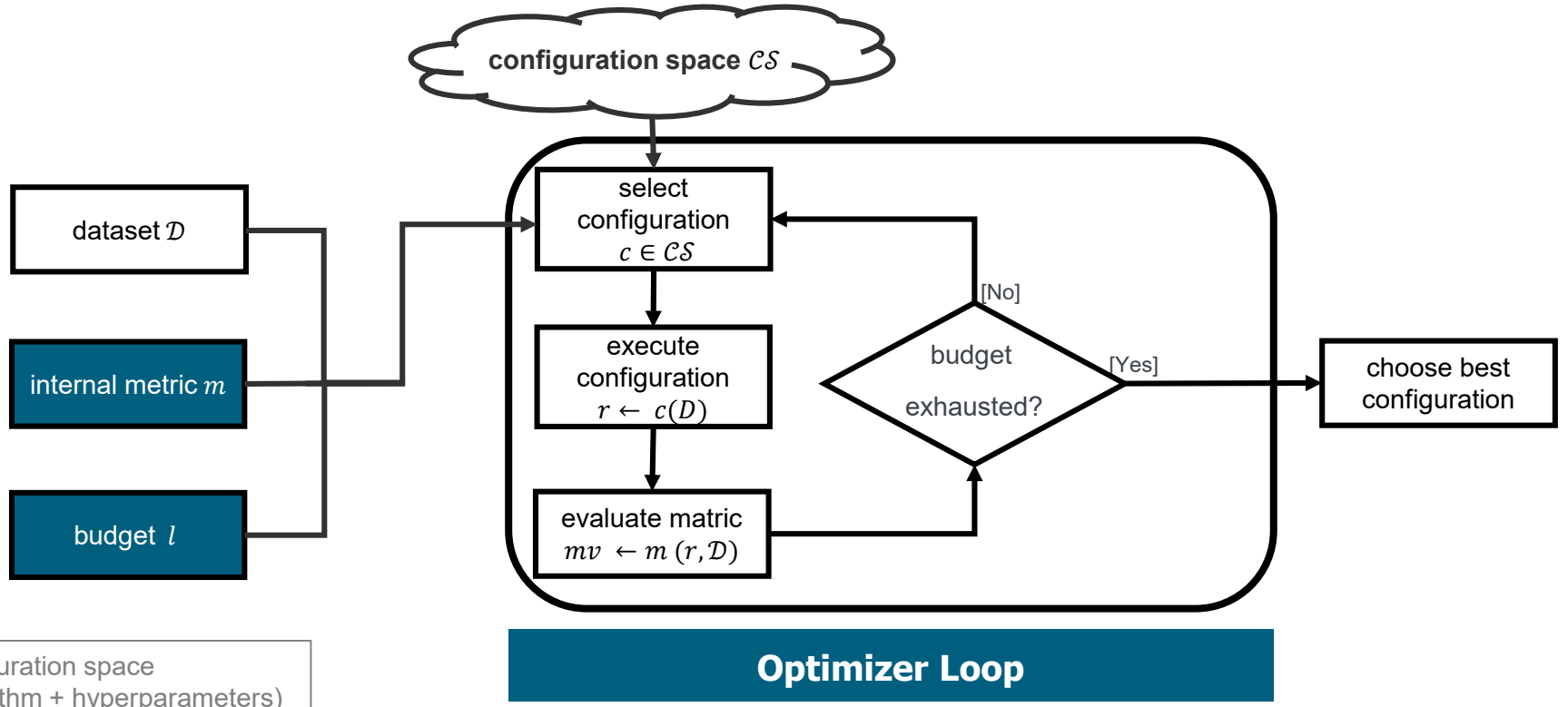
Algorithm Selection for Clustering

- 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

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

AutoML4Clust

Evaluation

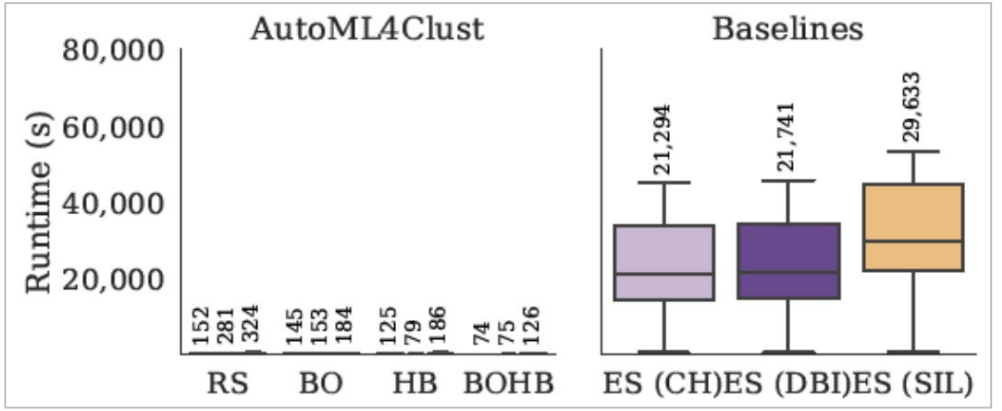
- Setup
  - 4 different optimizers
  - 3 different internal metrics
  - accuracy measure: Adjusted Mutual Information (AMI)
- Results
  - efficient as it explores large configuration spaces greedily
  - generic regarding optimizer and metric instantiations
  - best instantiation with Hyperband optimizer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

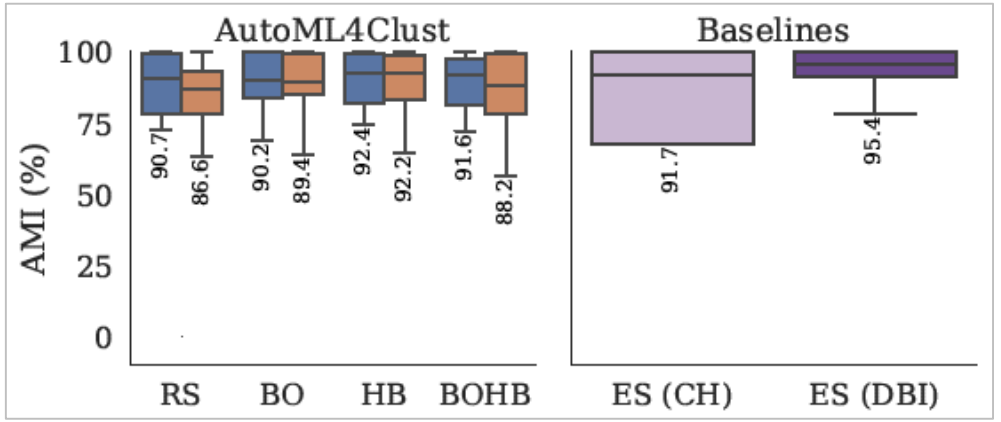
speedups up to 437x

only 3% accuracy deviations

RS: Random Search, **BO**: Bayes Optimizer, **HB**: Hyperband, **BOHB**: BO+HB, **ES**: Exhaustive Search

CH: Calinski-Harabasz, **DBI**: Davies Bouldin Index, **SIL**: Silhouette



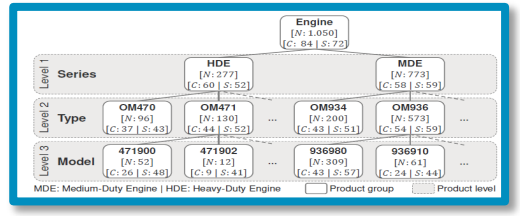


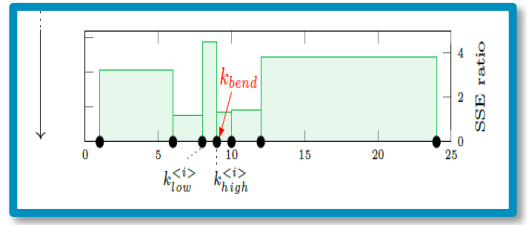
What‘s Next?

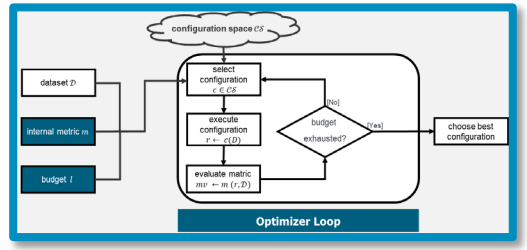
- Use Meta Learning to
  - … decide on the most suitable internal metric depending on data characteristics
  - … select promising configurations for the optimizer to start with
  - … reduce the size of the configuration space

Conclusion

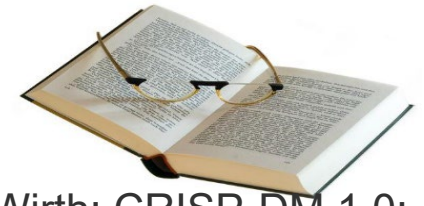
- Classification system for multi-class classification
  - exploiting domain knowledge helps to address issues with data characterisitcs (small training data set, class imbalance, …)
- LogMeans
  - method to efficiently determine the number of clusters in datasets
- AutoML4Clust
  - applying AutoML approaches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







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).

[FKE15] 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.