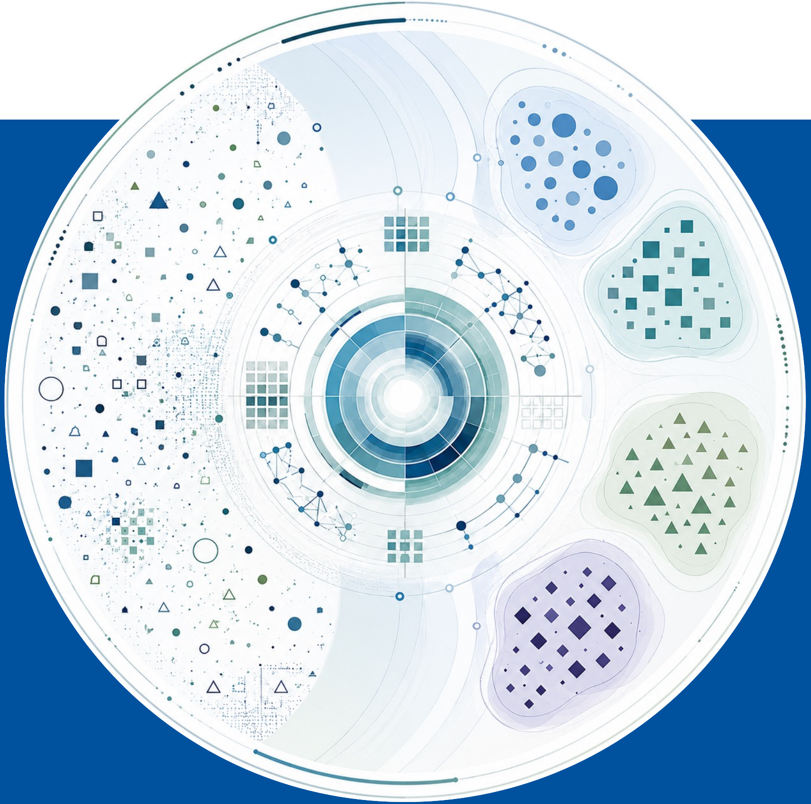
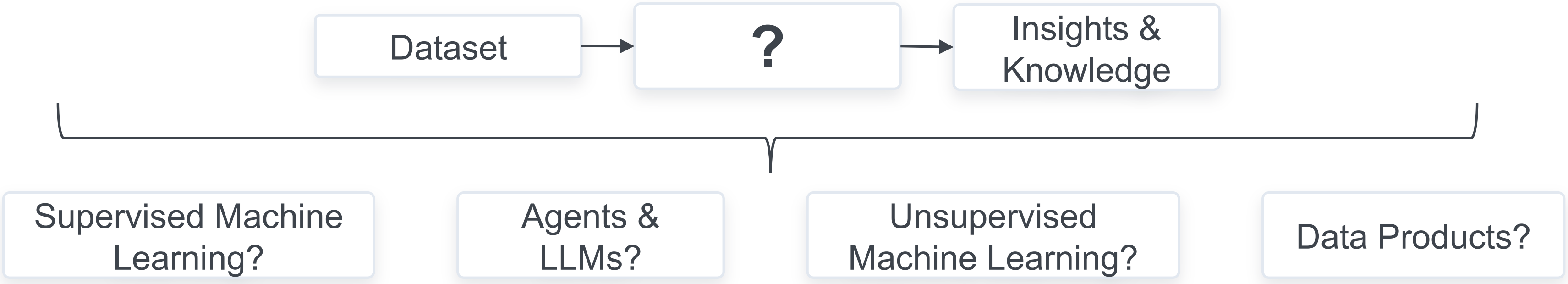
University of Stuttgart
Institute for Parallel and Distributed Systems (IPVS)
Applications of Parallel and Distributed Systems

Unsupervised Automated Preprocessing: the Role of Meta-Features and Data Characteristics

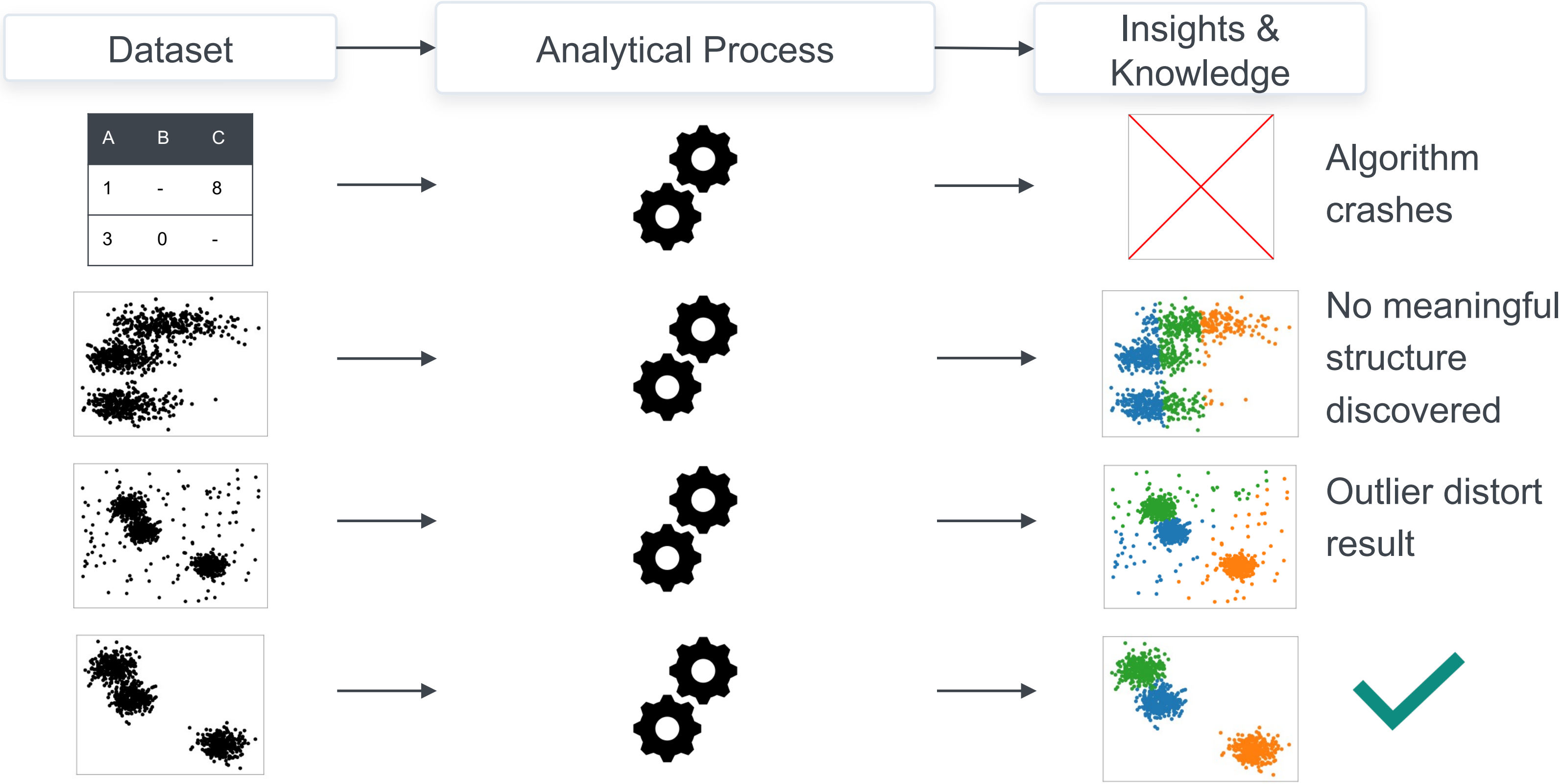
Leonard Labes, Dennis Treder-Tschechlov, Holger Schwarz



Motivation


Dataset
?
Insights & Knowledge
Supervised Machine Learning?
Agents & LLMs?
Unsupervised Machine Learning?
Data Products?

Motivation



Dataset
Analytical Process
Insights & Knowledge
A B C
1 - 8
3 0 -
Algorithm crashes
No meaningful structure discovered
Outlier distort result

Motivation

Dataset → Exploration → Preprocessing → Clustering → Insights & Knowledge

Dataset

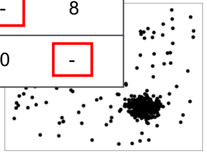
Tabular Data

A	B	C
1	-	8
3	0	-

Exploration

Find characteristics like missing values or outliers

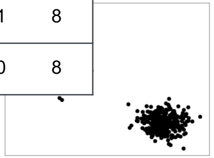
A	B	C
1	-	8
3	0	-



Preprocessing

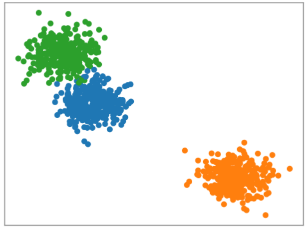
Remove outliers, impute missing values,…

A	B	C
1	1	8
3	0	8



Clustering

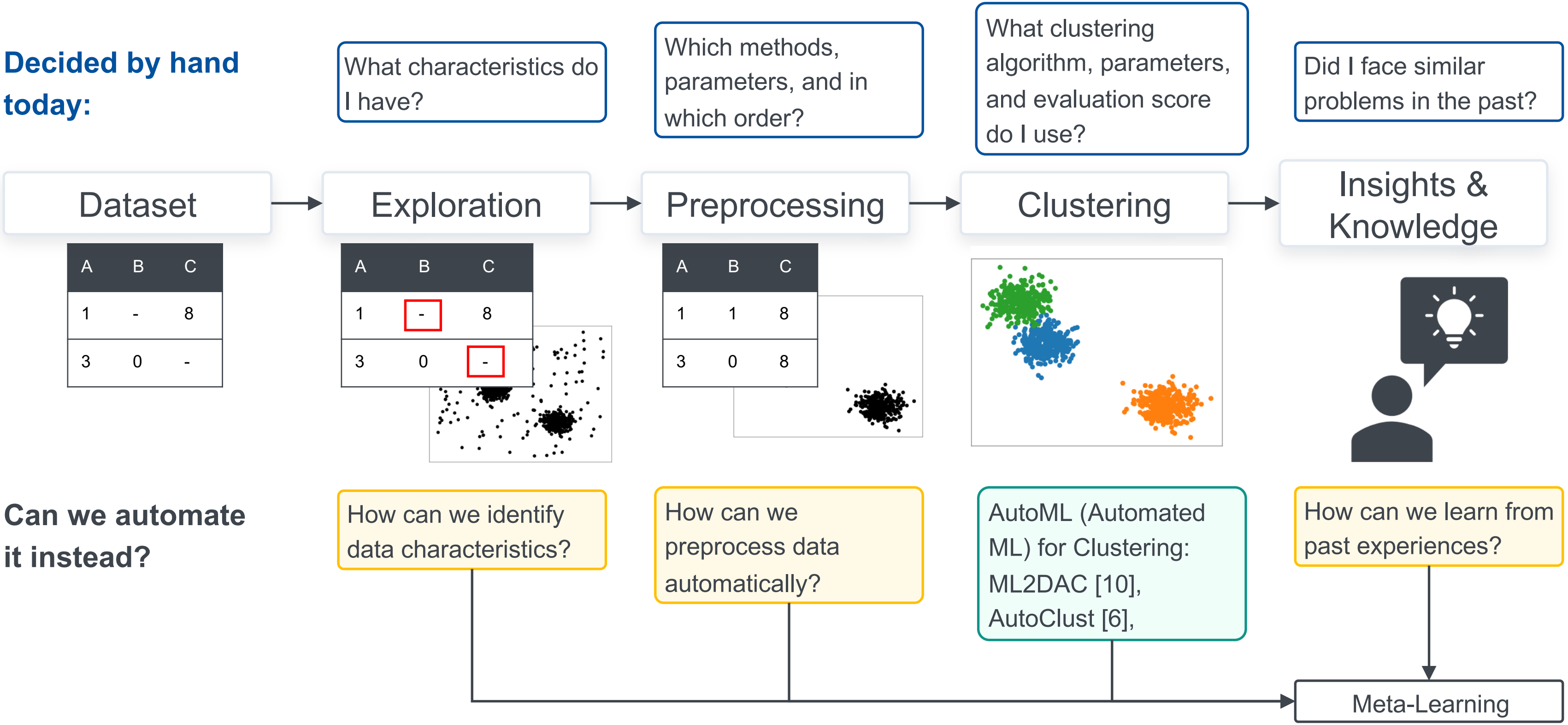
Apply machine learning, like clustering



Insights & Knowledge

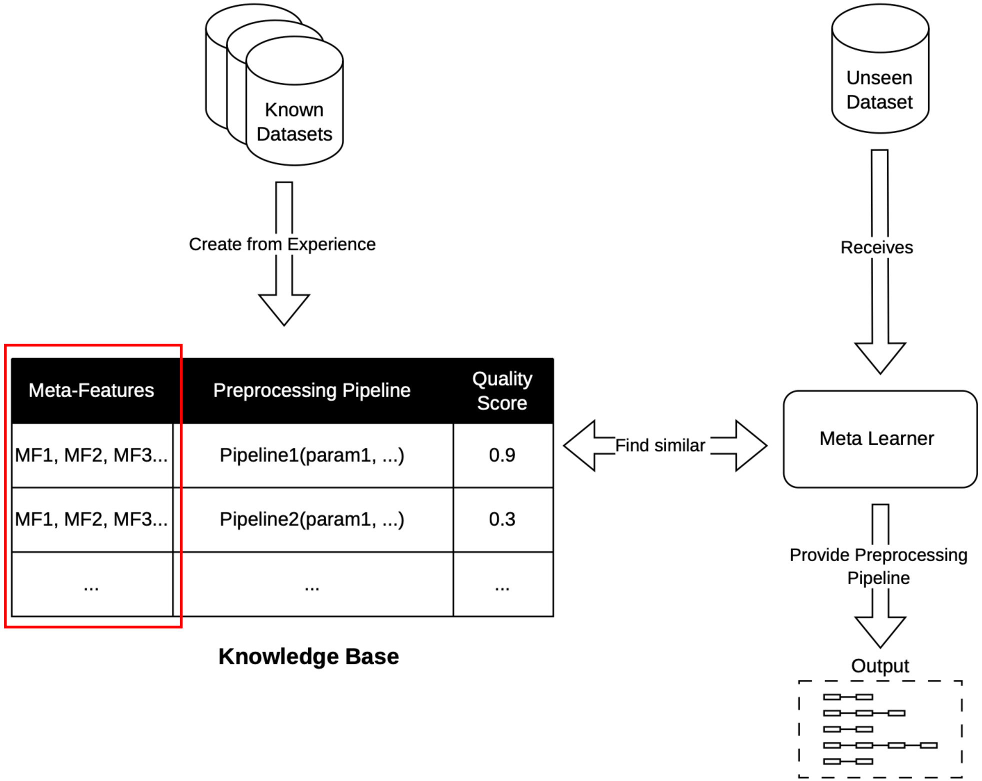
Get the answers to your questions

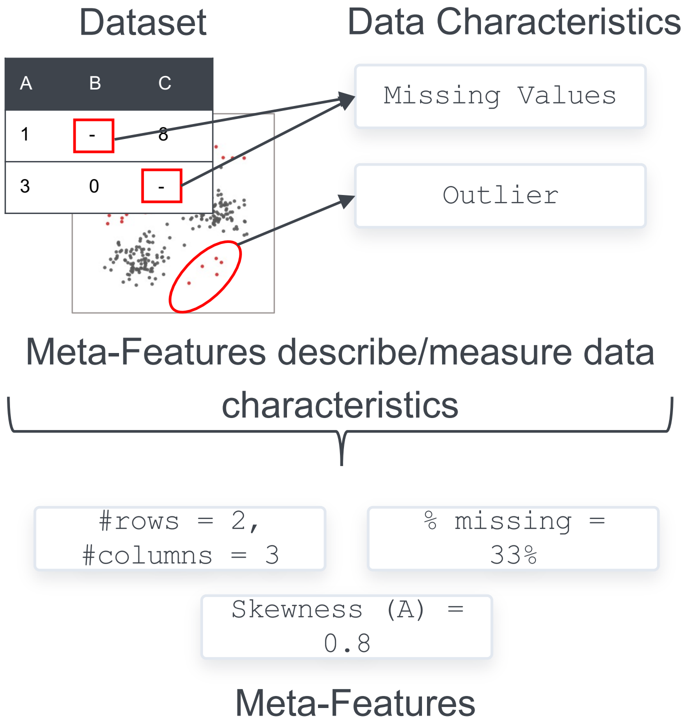
Motivation



Decided by hand today:
What characteristics do I have?
Which methods, parameters, and in which order?
What clustering algorithm, parameters, and evaluation score do I use?
Did I face similar problems in the past?
Dataset
Exploration
Preprocessing
Clustering
Insights & Knowledge
A B C
1 - 8
3 0 -
A B C
1 - 8
3 0 -
A B C
1 1 8
3 0 8
Can we automate it instead?
How can we identify data characteristics?
How can we preprocess data automatically?
AutoML (Automated ML) for Clustering: ML2DAC [10], AutoClust [6],
How can we learn from past experiences?
Meta-Learning

Meta-Learning and Meta-Features





Our Contribution

Data Characteristics, preprocessing [4], meta-features [1,11], and meta-learning [7,11] all exist.
Sorting and connecting them does not.

Dataset → Exploration → Preprocessing → Clustering → Insights & Knowledge

Can a system decide instead?

How can we identify data characteristics?
→ **THIS PAPER**

How can we preprocess data automatically?

AutoML for Clustering

How can we learn from past experiences?

1. **Collect and define** preprocessing-relevant data characteristics for clustering.
2. **Map** these characteristics to meta-features.
3. **Assess** how useful these meta-features are for unsupervised preprocessing decisions.

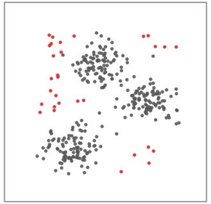
Data Characteristics

1

- **Preprocessing problems occur at different levels of the data.**
  - We identify 15 data characteristics

Dataset-level
6 characteristics

Outlier
distort distances & summary statistics



High #dimensions
High #features
Sparsity
Noise
Feature & dimension ratio

Feature-level
6 characteristics

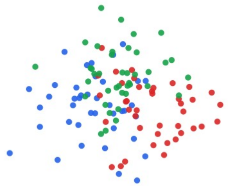
Missing Values
reduce information, bias estimates

A	B	C
1	-	8
3	0	-

Irrelevant features
Different scales
Skewed distributions
Duplicate and redundant features
Catgerical values

Cluster-level
3 characteristics

Overlapping Clusters
poor separation, unstable partition

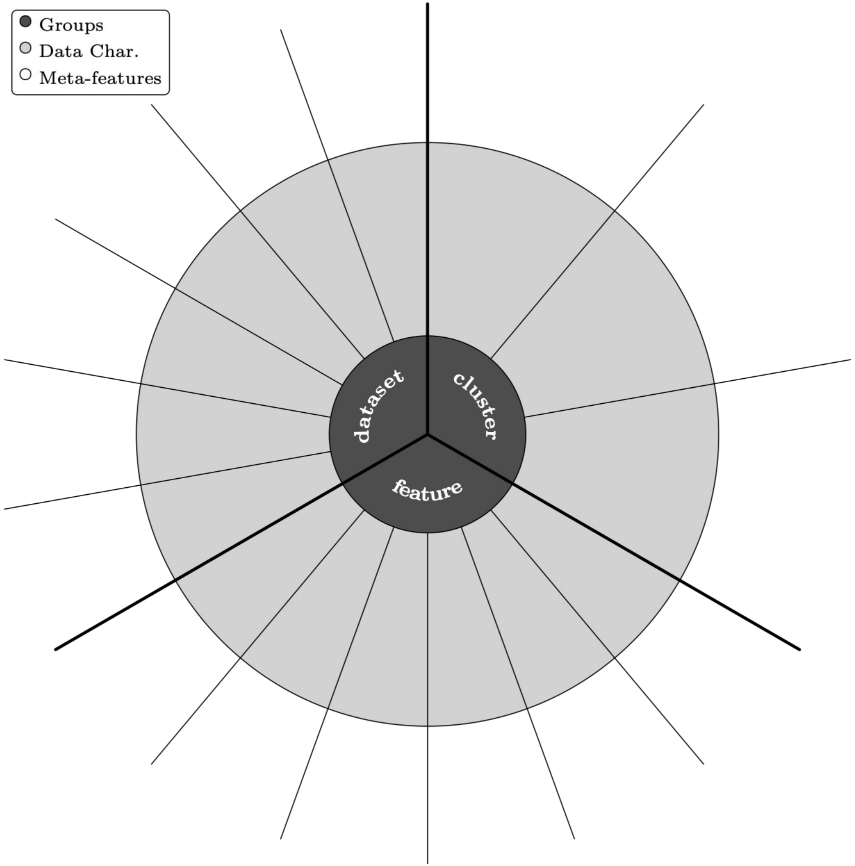


Convex and non-convex shapes
Imbalanced Clusters

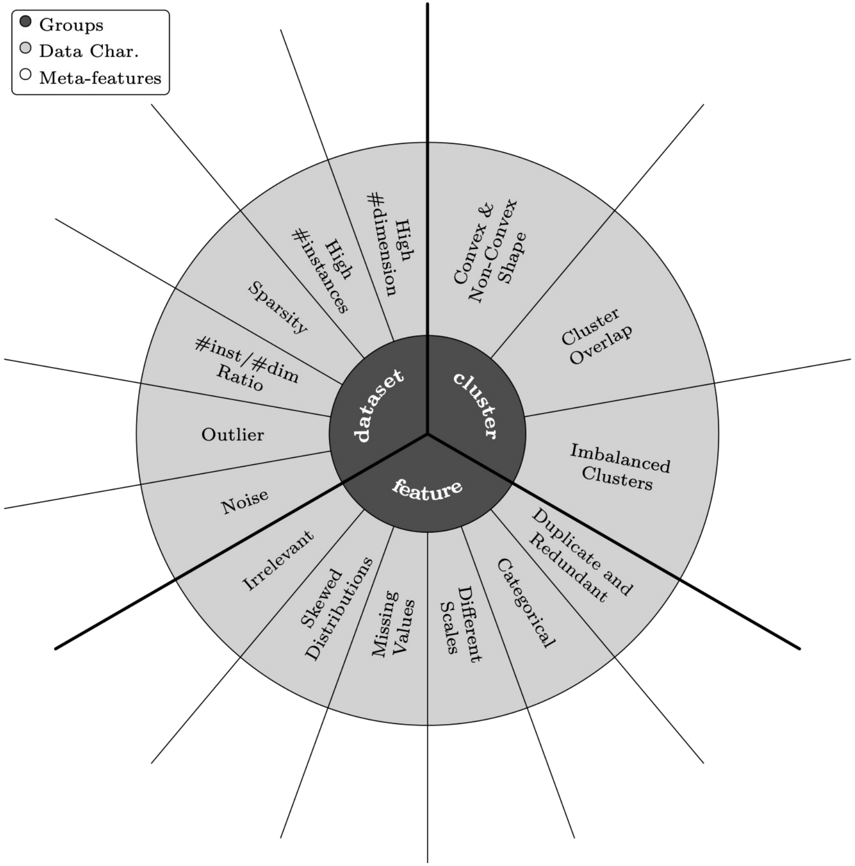
Our Contribution

1. **Collect and define** preprocessing-relevant data characteristics for clustering. ✓
2. **Map** these characteristics to meta-features.
3. **Assess** how useful these meta-features are for unsupervised preprocessing decisions.

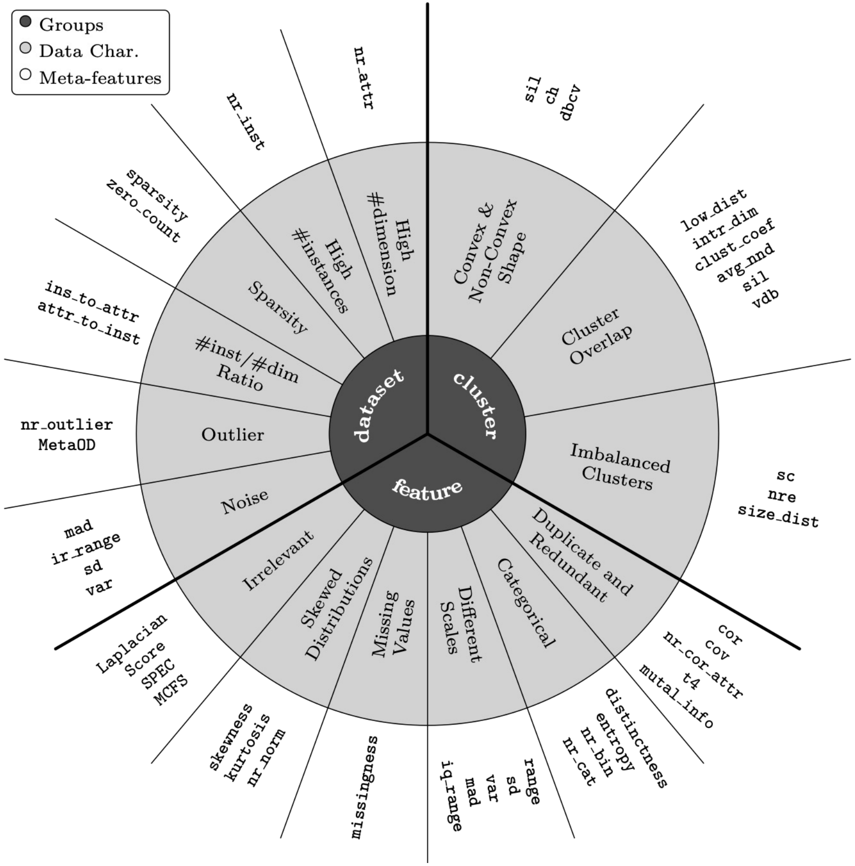
From Characteristics to Meta-Features



From Characteristics to Meta-Features



From Characteristics to Meta-Features

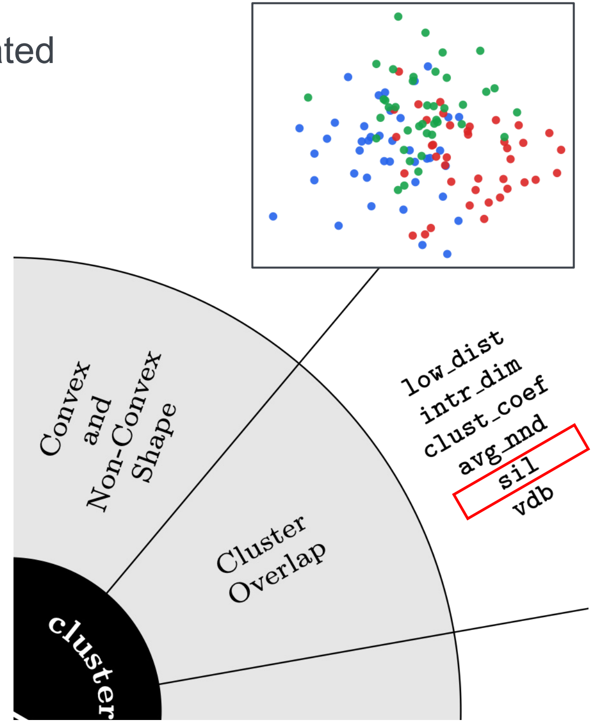


From Characteristics to Meta-Features

2

Cluster Level: Cluster Overlap: How strongly clusters are separated

- `sil`[9]: Silhouette-based proxy for cluster separation after provisional clustering
  - Compares the distance of points in their own cluster with the distance of points from other clusters
  - Low or near-zero values indicate weak separation



Our Contribution

1. **Collect and define** preprocessing-relevant data characteristics for clustering. ✓
2. **Map** these characteristics to meta-features. ✓
3. **Assess** how useful these meta-features are for unsupervised preprocessing decisions.

Assesing the Connection

Four Dimensions

1 Directness
To what extent does the meta-feature measure the characteristic

2 Grounding
Computable from raw data, does it need clustering first, or even more assumptions?

3 Actionability
Which preprocessing methods should I use?

4 Context robustness
How much does the value depend on thresholds and choices?

Levels of fulfilment per dimension

Overview of Assesment

Class	Meta-features of Data Characteristic	Directness	Grouding	Actionability	Context robustness
A	High number of instances	■■■	■■■	■■□	■■■
	High dimensionality	■■■	■■■	■■□	■■■
	Sample-to-dimension ratio	■■■	■■■	■■■	■■■
	Missing values	■■■	■■■	■■■	■■■
B	Sparsity	■□□	■■■	■■□	■■□
	Outliers	■■□	■■■	■■□	■■□
	Skewed distributions	■■□	■■■	■■■	■■■
	Different feature scales	■■□	■■■	■■■	■■■
	Categorical-valued features	■□□	■■■	■■□	■■□
	Duplicate / redundant feat.	■■□	■■■	■■■	■■■
C	Noise	■□□	■■■	■□□	■■□
	Imbalanced clusters	■□□	■■□	■■□	■■□
	Cluster overlap	■□□	■■□	■■□	■□□
	Convex / non-convex shapes	■□□	■■□	■□□	■□□
	Irrelevant features	■□□	■□□	■■□	■■□

Reading: **Directness:** *direct → near-direct → proxy* **Grounding:** *raw-data → partition-dependent / assumption-heavy* **Actionability:** *clear → partial → weak* **Context robustness:** *moderate → high → extreme*

Conclusion & Outlook

1. **Collect and define** preprocessing-relevant data characteristics for clustering. ✓
2. **Map** these characteristics to meta-features. ✓
3. **Assess** how useful these meta-features are for unsupervised preprocessing decisions. ✓

Outlook:

- Empirical evaluation of this assessment
- Embed it in a preprocessing-aware meta-learning system that recommends preprocessing pipelines

For more **details** about this **work** and the overall topic of **automated preprocessing**, feel free to check out my posters.

References

1. Alcobaça, E., et al.: Mfe: Towards reproducible meta-feature extraction. Journal of Machine Learning Research 21(111), 1–5 (2020)

2. Alcobaça, E., Siqueira, F.: PyMFE Documentation: Meta-feature Description Table (2021), https://pymfe.readthedocs.io/, version 0.4.2, accessed 2026-03-30

3. Fernandes, L.H.d.S., et al.: Towards understanding clustering problems and algorithms: an instance space analysis. Algorithms 14(3), 95 (2021)

4. Han, J., et al.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2022)

5. Pargent, F., et al.: Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. arXiv preprint arXiv:2104.00629 (2021)

6. Poulakis, Y., et al.: Autoclust: A framework for automated clustering based on cluster validity indices. In: 2020 IEEE Int. Conf. on Data Mining (ICDM), pp. 1220–1225. IEEE (2020)

7. Rice, J.R.: The algorithm selection problem. In: Advances in Computers, vol. 15, pp. 65–118. Elsevier (1976)

8. Rivolli, A., et al.: Meta-features for meta-learning. Knowledge-Based Systems 240, 108101 (2022)

9. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)

10. Treder-Tschechlov, D., et al.: Ml2dac: Meta-learning to democratize automl for clustering analysis. Proc. of the ACM on Management of Data 1(2), 1–26 (2023)

11. Vanschoren, J.: Meta-learning: A survey. arXiv preprint arXiv:1810.03548 (2018)

12. Zhao, Y., et al.: Automatic unsupervised outlier model selection. Advances in Neural Information Processing Systems 34, 4489–4502 (2021)

Thank you!

Leonard Labes

e-mail leonard.labes@ipvs.uni-stuttgart.de
phone +49 (0) 711 685-88295
www.ipvs.uni-stuttgart.de/de/institut/team/Labes/

University of Stuttgart
Institute for Parallel and Distributed Systems (IPVS)