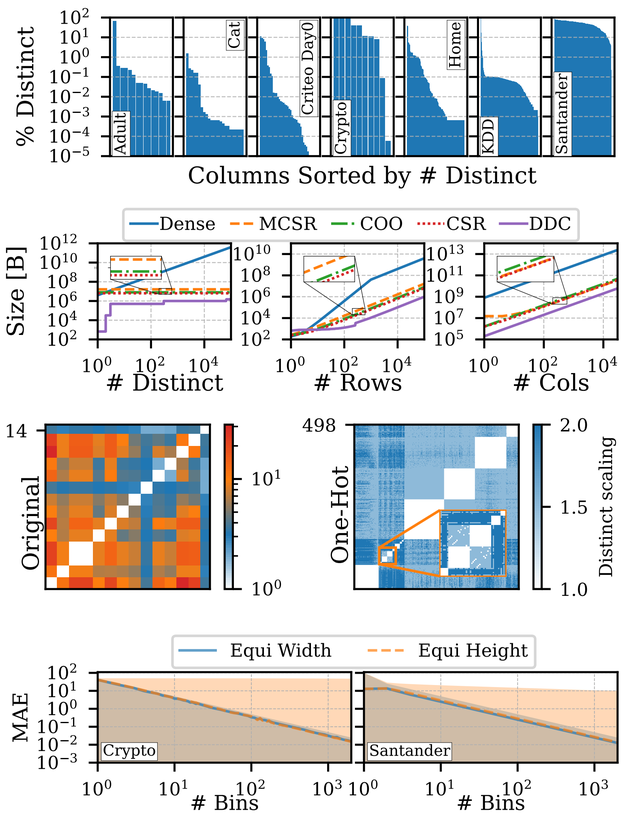
Morphing-based Compression for Data-centric ML Pipelines

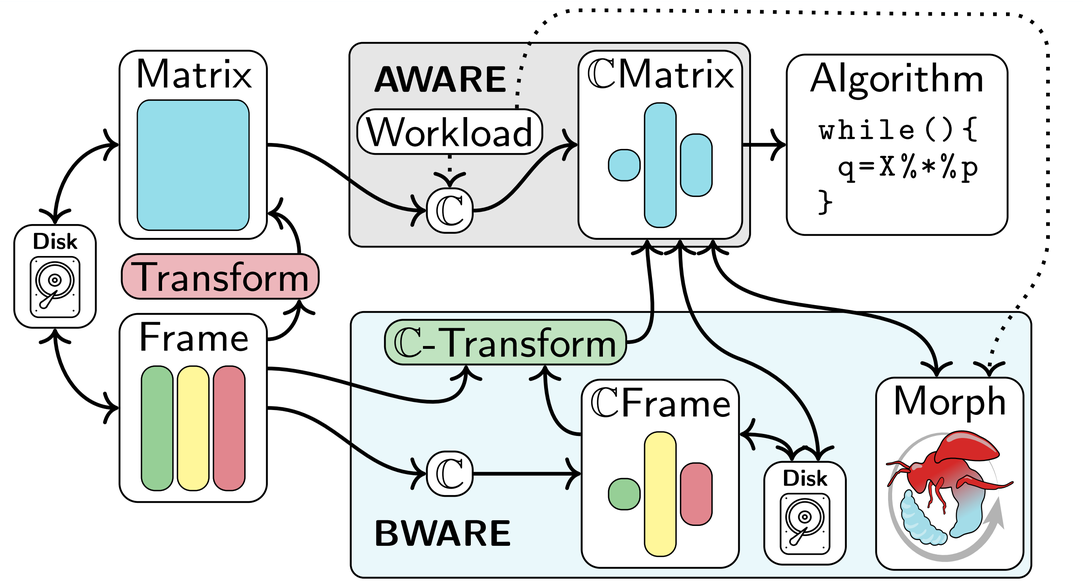
Sebastian Baunsgaard, Matthias Boehm

MOTIVATION — Exploitable Properties



- The relative number of **distinct values** is **low** in many columns of datasets.
- Compression schemes can use **less memory** than dense or sparse representations.
- **Column correlations** obvious before feature transformations can be **difficult and expensive to rediscover** after.
- The number of distinct values in datasets is **controllable** through techniques such as **quantization**.

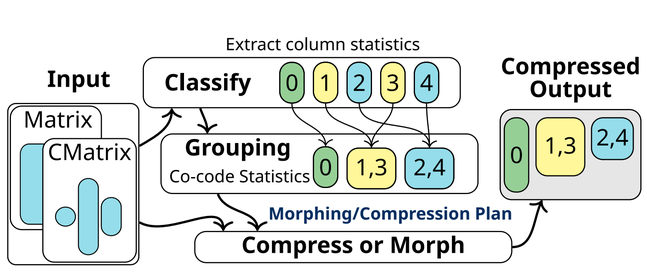
OVERVIEW — Workload-aware Compression



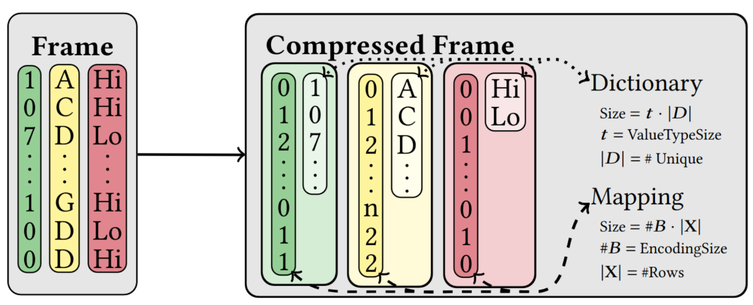
- **ULA** (uncompressed linear algebra): transform to a **dense matrix**, then train, redundancy **thrown away**.
- **AWARE** (prior work): compress the matrix **after** transformation, redundancy must be **rediscovered**.
- **BWARE** (this work): compress **at the source**, stay compressed, **morph** to fit the workload.

Avoid Rediscovering Compression Plans!

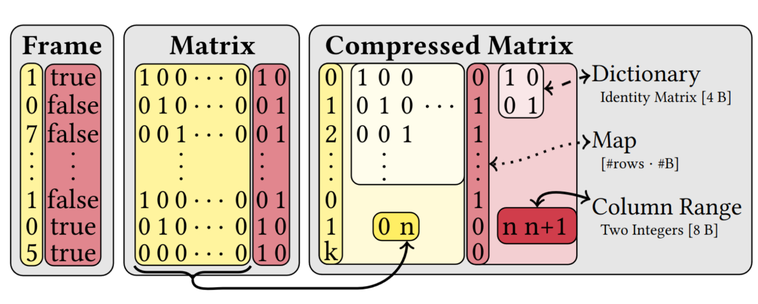
TECHNIQUES — Compression Throughout the Pipeline



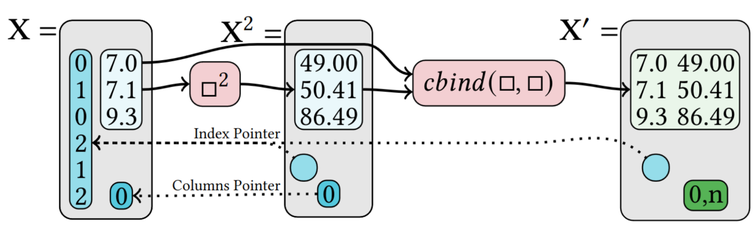
- **Online analysis** morphs formats without decompression.



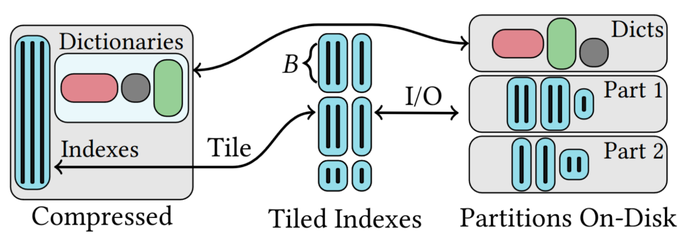
- **Frame compression** losslessly encodes heterogeneous columns.



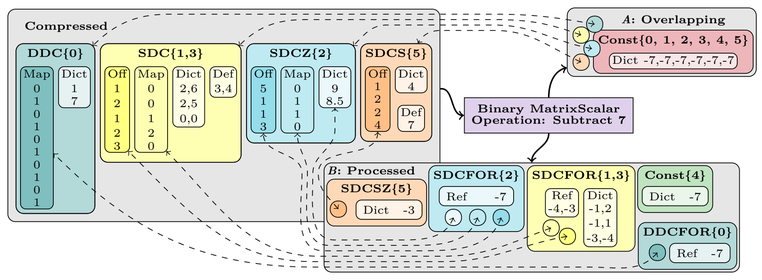
- **Compressed transformations** encode frames, e.g., one-hot.



- **Feature engineering** appends nonlinear features.

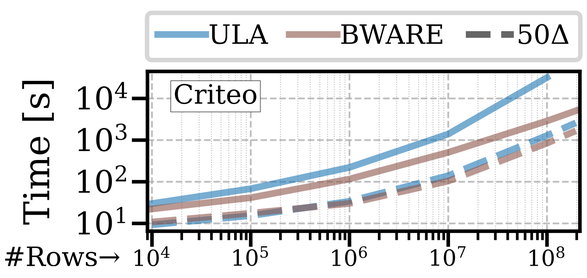


- **Compressed I/O** avoids decompression.

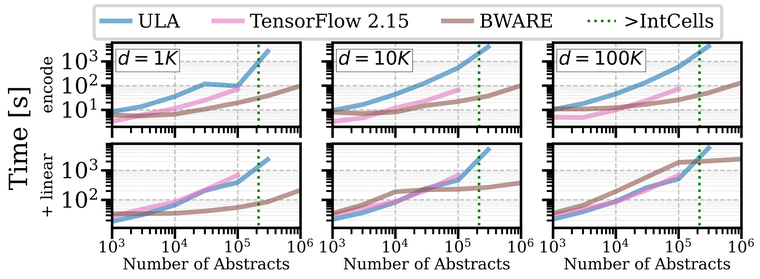


- **Compressed operations** run on compressed data.

RESULTS — Compressed Transformations and End-to-End Pipelines

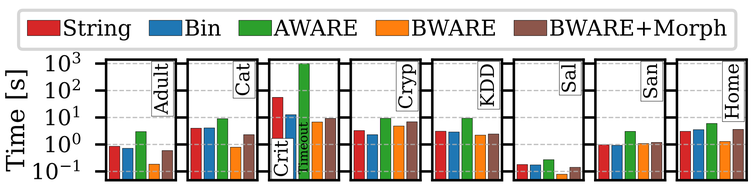


- **Linear model training** scaling in rows, 11x faster at 10^8 rows.

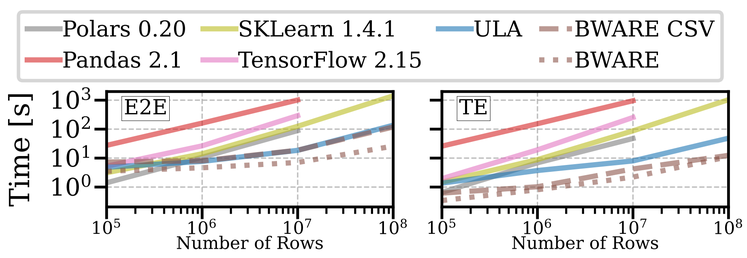


- **Word embeddings** against TensorFlow, with and without a dense layer.

		ULA	AWARE	BWARE
KDD98	Time	654s	452s	251s
	Instructions	$110_{\cdot 10^{12}}$	$100_{\cdot 10^{12}}$	$44_{\cdot 10^{12}}$
	Ins per Cycle	0.94	2.59	2.68
	L1-dcache-miss	$7{,}792_{\cdot 10^{9}}$	$1{,}740_{\cdot 10^{9}}$	$786_{\cdot 10^{9}}$
	Compress/Morph	—	148s	21.9s
	Transform-Encode	74.1s	59.9s	7.95s
	Energy Consumption	338kJ	185kJ	92kJ
Home	Time	431s	266s	160s
Adult	Time	19.9s	22.1s	16.6s
Santander	Time	489s	376s	374s
Cat	Time	467s	170s	63s



- **Transform-encode** times are better or equal, without rediscovering plans.



- **System comparisons** against Scikit-learn, Pandas, and Polars.

- **ML pipeline:** linear regression with compressed transformations and eight polynomials.

Contact Paper Code