Data-centric ML Pipelines in Various Application Domains

Prof. Dr. Matthias Boehm
Technische Universität Berlin
Berlin Institute for the Foundations of Learning and Data
Big Data Engineering (DAMS Lab)

About Me

- **Since 09/2022 TU Berlin**, Germany
  - University professor for Big Data Engineering (DAMS)
- **2018-2022 TU Graz**, Austria
  - BMK endowed chair for data management + research area manager
  - **Data management for data science** (DAMS), **SystemDS & DAPHNE**
- **2012-2018 IBM Research – Almaden**, CA, USA
  - Declarative large-scale machine learning
  - Optimizer and runtime of **Apache SystemML**
- **2007-2011 PhD TU Dresden**, Germany
  - Cost-based optimization of integration flows
  - Time series forecasting / in-memory indexing & query processing

DB group

Data-centric ML Pipelines

Key observation: SotA data engineering/cleaning based on ML

Data Engineering

Top-K Cleaning Pipelines
[**SIGMOD'24a**]

Alignment of Multi-modal Data

I/O for Custom Data Formats
[**SIGMOD'23c**]

Parallel Feature Transformations
[**PVLDB'22**]

Data Integration & Data Cleaning

Data Programming & Augmentation

Model and Feature Selection

Hyper-parameter Tuning + CV

Hierarchical Composition as Library Functions on top of ML systems

Data Preparation (e.g., one-hot, bins)

FX

`train()`

Model Training

X **y**

```
while(!converged) {
  … q = X %*% v …
}
```

`predict()`

Model Scoring

85% Accuracy

SliceLine
[**SIGMOD'21c**]

Validation & Debugging

Deployment & Scoring

Apache SystemDS [https://github.com/apache/systemds]

Apache SystemML™

07/2020 Renamed to **Apache SystemDS**
05/2017 Apache Top-Level Project
11/2015 Apache Incubator Project
08/2015 Open Source Release

APIs: Command line, JMLC, Python
Spark MLContext, Spark ML,
(Scalable Algorithms + Primitives)

DML Scripts

Language

Compiler

Runtime

Write Once,
Run Anywhere

[SIGMOD'15,'17,'19,'21abc,'23abc,'24a]
[PVLDB'14,'16ab,'18,'22]
[ICDE'11,'12,'15]
[CIDR'17,'20]
[VLDBJ'18]
[EDBT'25]
[CIKM'22]
[DEBull'14]
[PPoPP'15]

In-Memory Single Node
(scale-up)

Hadoop or Spark Cluster
(scale-out)

Federated
(LA progs, PS)

Others:
Netezza
Apache Flink

In-Progress:
GPU

since 2014/16

Java

since 2012

hadoop

since 2010/11

APACHE Spark

since 2015

exdra

since 2019

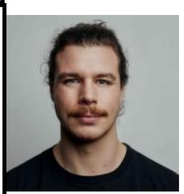
Federal Ministry
Republic of Austria
Climate Action, Environment,
Energy, Mobility,
Innovation and Technology

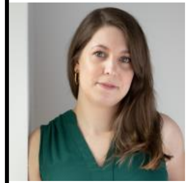
FG Big Data Engineering (DAMS Lab) – Selected Applications

- Inter-disciplinary Collaborations for Grounding and Preventing Artificial Problems & Solutions
- Positioning in BIFOLD

Earth Observation
• BigEarthNet / EO Reproducibility (TUB) • EO Citizen Science Platform and Eco System (TUB) • Interaction of Climate Turning Points (PIK) • Surface Cover Classification (DLR) • BS/MS Thesis on Data Augmentation

Health-care / Medical
• LungCAIRE multi-modal data representations and debugging (Charite) • Stomach cancer anomaly detection (Charite) • Time Series Alignment in NebulaStream (Charite) • MRI Scanner Artifacts Detection & Segmt. (TUB, Siemens Healthineers)



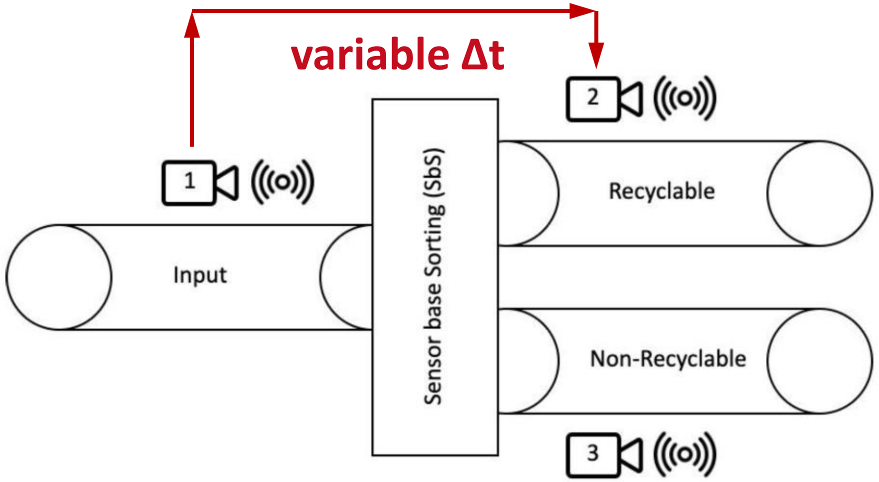


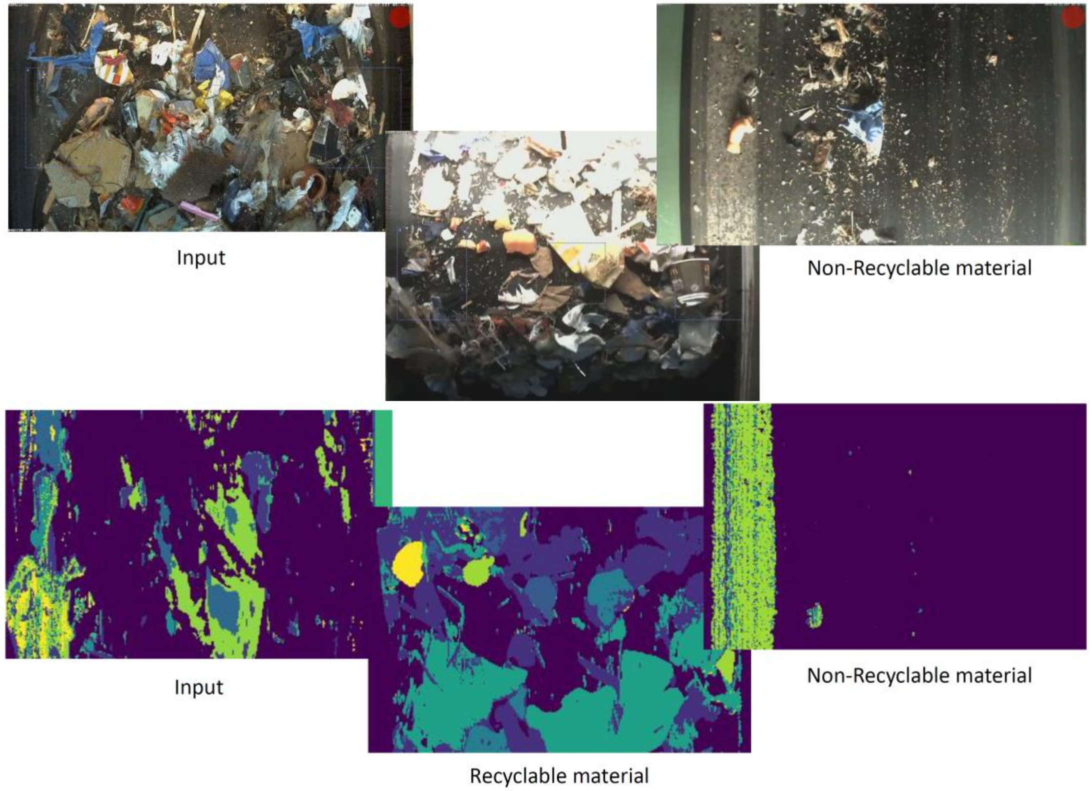
Others
• Automotive (MAGNA, AVL, Porsche, BMW, DB Services) • Process Industry / Recycling (Siemens, Bayer, VoestAlpine, REDWAVE, Andritz) • Energy (SAP, EnBW, AEE-Intec, DigSilent, Energiequelle) • Semiconductor Manufacturing (Infineon, KAI, Intel)

Data-centric ML in Example Applications

Spatial-temporal Alignment (in Recycling)

- **ReWaste F Project**
  - Digital Platform for Austrian Recycling Economy
  - 4 scientific and 14 industrial partners
- **Example Use Case: Prediction Model for Material Composition after Sorting**
  - 1-3: Video + NIR sensors

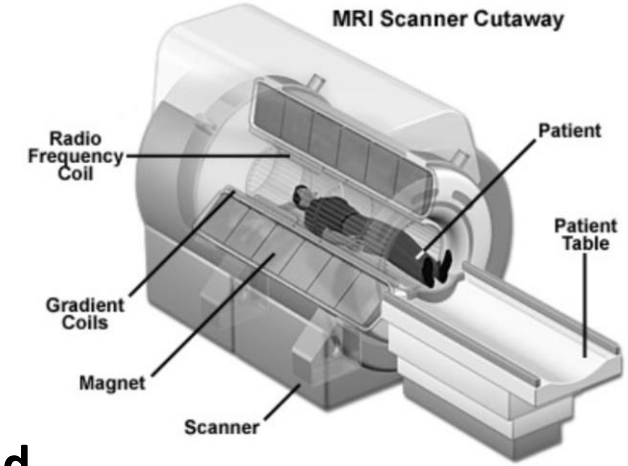




➔ **Alignment via Dynamic Time Warping, Anchor Objects, and Cross-correlation**

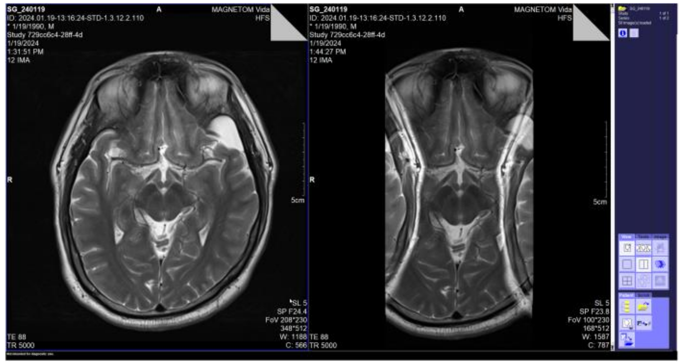
Data Augmentation (in Health-Care)

- **Magnetic Resonance Imaging (MRI) Scanners**
  - Widely used in various medical applications
  - **Various sources of image artifacts** (noise, movement, interference HW defects, transitions)

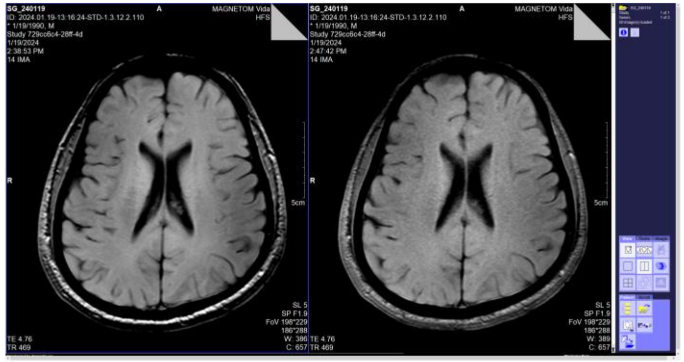


- **Example Use Case: ML-based Artifact Detection and Quantification**
  - **Physics-based data simulator for artifacts** (Turbo Spin Echo, HASTE, Flash2D, Beat, Space, Flash3D Vibe, Echo Planar 2D Diffusion, Gradient Echo GRE)
  - Highly accurate but slow, **combination w/ traditional data augmentation** (distortions and noise)
  - ResNet50 → **95+% accuracy**

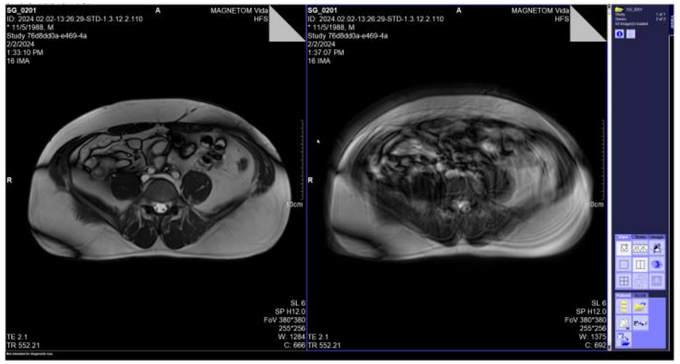
Infolding (brain)



Noise (brain)



Movement (abdomen)



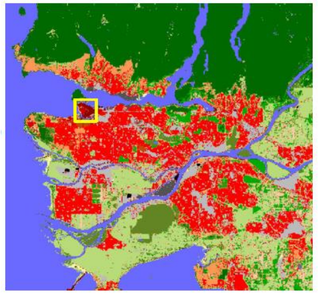
Data Preprocessing (in Earth Observation)

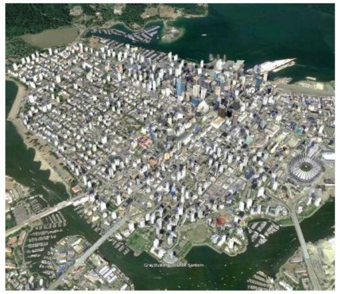
[Xiao Xiang Zhu et al: So2Sat LCZ42: A Benchmark Dataset for the Classification of Global Local Climate Zones. **GRSM 8(3) 2020**]

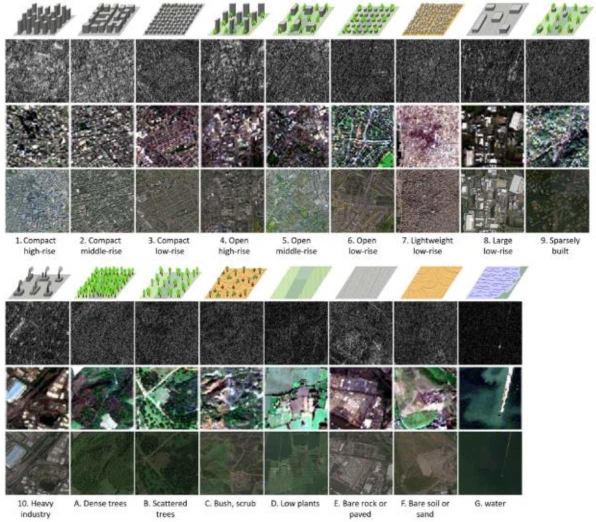
[So2Sat LC42: https://mediatum.ub.tum.de/1454690]

- **DLR Earth Observation Use Case**
  - **ESA Sentinel-1/2** datasets → 4PB/year
  - Training of local climate zone classifiers on **So2Sat LCZ42**
    (15 experts, 400K instances, 10 labels each, 85% confidence, ~55GB H5)
  - **ML Pipeline:** ResNet20, climate models

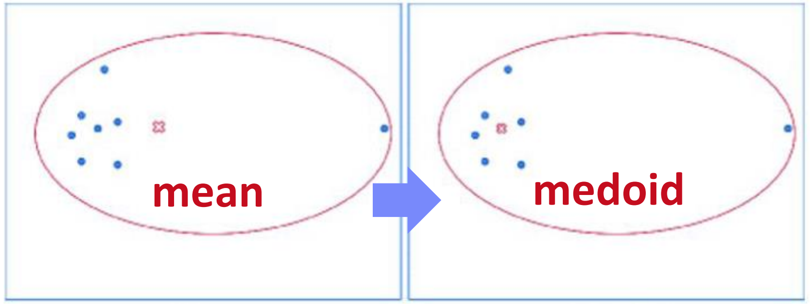
DAPHNE







- **Preprocessing**
  - **LTSM** for combing patches from **four seasons**
  - Time series of patches per location
    Combine cloud- and shadow-free pixels via cloud masks → select **medoid**



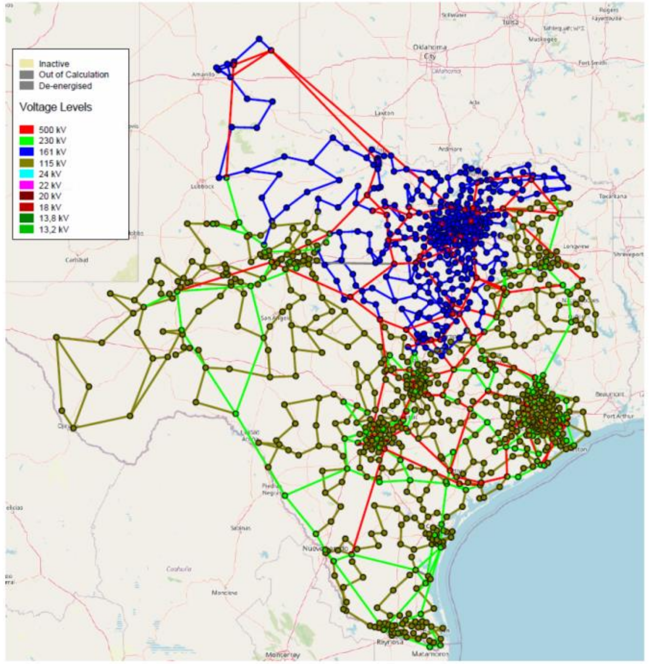
ML-based Simulations (in Energy & Automotive)

[T. Pfaff et al: Learning Mesh-Based Simulation with Graph Networks. **ICLR'21**]

- **Background ML-based Simulations**
  - Trend to replace traditional HPC simulation by cost-effective ML models
  - CFD simulation through ML (MLP enc/dec + MLP/GraphNet message passing)

- **Example #1: Dynamic Security Assessment of Energy Grids**
  - Traditional RMS simulation **too slow for online use**
  - **Train prediction model for critical fault clearing time** (CFCT), metric how close a system is to its stability limits
  - **Simulate fault conditions to create training dataset** (w/ careful composition of failure scenarios, and CFCT ranges)

[**Credit:** Ann-Sophie Messerschmid, Texas Transmission Grid]



- **Example #2: AVL Ejector Geometry Optimization (fuel cells)**
  - Currently mixed of data-driven ML pipeline and **3D CFD simulation**
  - **Towards cost-effective ML-based CFD simulation**

DAPHNE

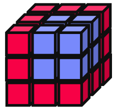
Selected Research Results & Directions for Data-Centric ML Pipelines

SAGA for Finding Data Cleaning Pipelines [SIGMOD'24a]

[best paper runner-up w/ Shafaq and Roman]

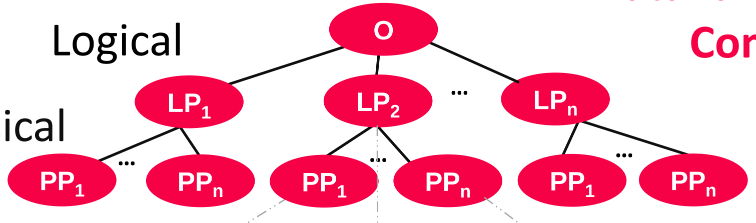
- **Automatic Generation of Cleaning Pipelines**
  - Library of robust, parameterized **data cleaning primitives**,
  - **Enumeration of DAGs** of primitives & **hyper-parameter optimization** (evolutionary, HB)

Data Samples



Target App

Rules/Objectives

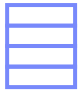


Data- and Task-parallel Computation

Outlier Detection → MVI → Deduplication → Resolve Mislabels

P_1: gmm → imputeFD → mergeDup → delML

P_n: outlierBySd → mice → delDup → voting



Top-k Pipelines

Debugging

University	Country
TU Graz	Austria
TU Graz	Austria
TU Graz	Germany
IIT	India
IIT	IIT
IIT	Pakistan
IIT	India
SIBA	Pakistan
SIBA	null
SIBA	null

Dirty Data

University	Country
TU Graz	Austria
TU Graz	Austria
TU Graz	Austria
IIT	India
IIT	India
IIT	India
IIT	India
SIBA	Pakistan
SIBA	Pakistan
SIBA	Pakistan

After **imputeFD(0.5)**

A	B	C	D
0.77	0.80	1	1
0.96	0.12	1	1
0.66	0.09	null	1
0.23	0.04	17	1
0.91	0.02	17	null
0.21	0.38	17	1
0.31	null	17	1
0.75	0.21	20	1
null	null	20	1
0.19	0.61	20	1
0.64	0.31	20	1

Dirty Data

A	B	C	D
0.77	0.80	1	1
0.96	0.12	1	1
0.66	0.09	17	1
0.23	0.04	17	1
0.91	0.02	17	1
0.21	0.38	17	1
0.31	0.29	17	1
0.75	0.21	20	1
0.41	0.24	20	1
0.19	0.61	20	1
0.64	0.31	20	1

After **MICE**

SliceLine for ML Model Debugging [SIGMOD'21b]

sliceline

[**Credit:** sliceline, Silicon Valley, HBO]

- **Problem Formulation**
  - **Intuitive slice scoring function**
  - Exact **top-k slice finding**
  - $|S| \geq \sigma \wedge sc(S) > 0,\ \alpha \in (0,1]$

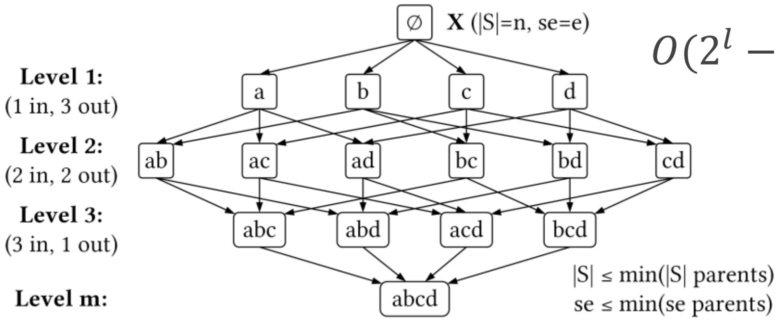
$$sc = \alpha\left(\frac{\bar{e}(S)}{\bar{e}(X)} - 1\right) - (1-\alpha)\left(\frac{|X|}{|S|} - 1\right)$$

$$= \alpha\left(\frac{|X|}{|S|} \cdot \frac{\sum_{i=1}^{|S|} es_i}{\sum_{i=1}^{|X|} e_i} - 1\right) - (1-\alpha)\left(\frac{|X|}{|S|} - 1\right)$$

slice error **slice size**

- **Properties & Pruning**
  - Monotonicity of slice sizes, errors
  - **Upper bound sizes/errors/scores**

    → pruning & termination



$$O\left(2^l - \sum_{j=1}^{m} 2^{d_j} + l + m\right)$$

- **Linear-Algebra-based Slice Finding**
  - Recoded/binned matrix **X**, error vector e
  - **Vectorized implementation in linear algebra** (join & eval via sparse-sparse matmult)
  - Local and distributed task/data-parallel execution

Candidate Slices

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Data

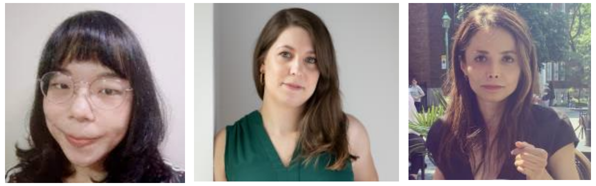
$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 2 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}$$

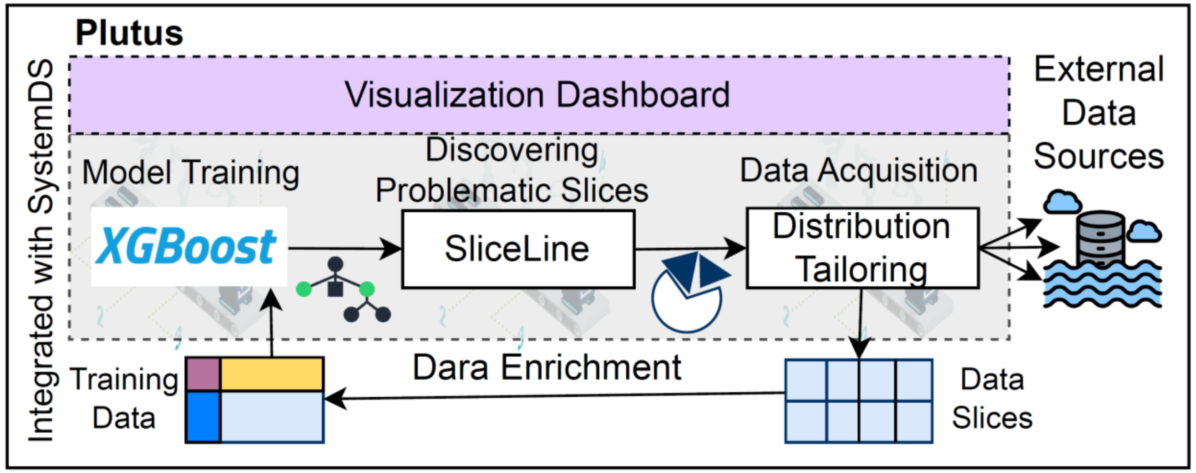
== Level

BIFOLD

SliceLine Extensions (Sampling, Incremental, Multi-modal)

- **#1 Distribution Tailoring for ML**
  - Model training → SliceLine → Sampling
  - Iterative procedure w/ debugging dashboard
  - **[SIGMOD'24 demo]**





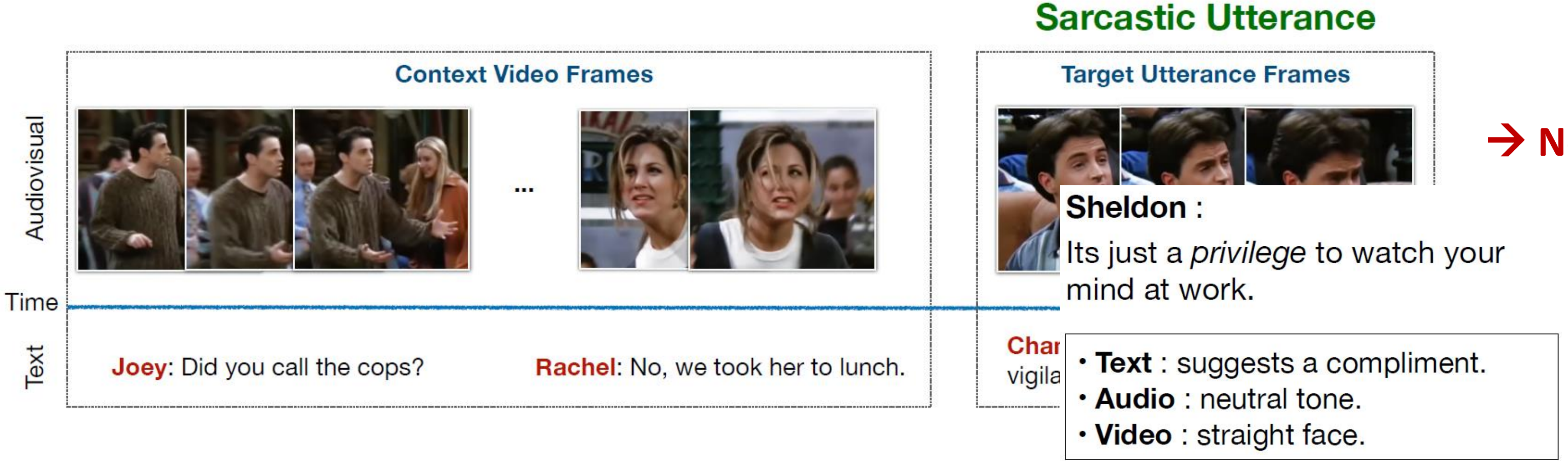
- **#2 Incremental SliceLine (under submission)**
  - Leverage collected state of previous SliceLine execution of modified dataset
  - **Pruning by previous top-K score, unchanged sizes, maximum reachable scores**

- **#3 SliceLine for Multi-modal Data (in progress)**
  - Modality-specific embeddings and combination
  - **Find high-level features for debugging** (e.g. distinct tokens, bounding boxes)

New Direction #1: Data Representation Search

- **Goal: Find effective data representations for multi-modal ML models**
- **Objectives: accuracy, runtime, and label-efficiency**
- **Example: MUStARD**

[**Credit:** S. Castro et al @ **ACL'19**]

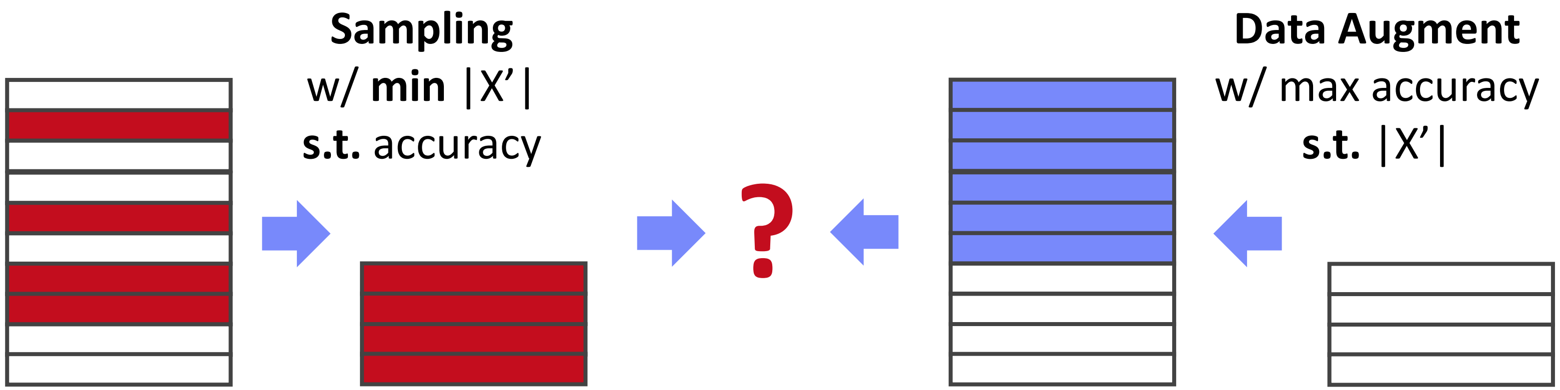


Challenges
→ Need for Context

- **Scuro Library** (part of SystemDS)
  - **Different modality representations** and **modality-specific features** (e.g., pitch, intonation)
  - **Spatial-temporal alignment** through different alignment strategies
  - **Algebra for composition** of representations + search for alternative plans

New Direction #2: Learned Sampling & Augmentation

- **Goal: Create small, high-quality datasets via learned sampling and augmentation**
- **Objectives: accuracy s.t. preserved data distribution**



- **Compositional Dataset Search**
  - DendroGrad: **dendrogram of gradients** of real and synthetic examples
  - Sampling while preserving **topological structure** (e.g., representation topology divergence)
  - **Kernel density estimation** and distribution sampling

Conclusions & QA

Thanks

- **#1 Data-centric ML Pipelines**
  - Increasingly complex, composite ML pipelines
  - State-of-the-art data engineering methods based on ML
  - Partial **resource, operational, and data redundancy**
- **#2 Data-centric ML in Applications**
  - Spatial-temporal Alignment (**in Recycling**)
  - Data Augmentation (**in Health-Care**)
  - ML-based Preprocessing (**in Earth Observation**)
  - ML-based Simulations (**in Energy & Automotive**)
- **#3 New Research Directions**
  - Data Representation Search
  - Learned Sampling & Augmentation

→ **Need for Abstractions and inter-disciplinary Collaborations**

→ **Optimizing Compiler and Runtime Infrastructure**

→ **Application-agnostic and Application/Domain-specific Primitives**

THE APACHE SOFTWARE FOUNDATION

DAPHNE

https://github.com/apache/systemds
https://github.com/daphne-eu/daphne

BIFOLD