Integrated **D**ata **A**nalysis **P**ipelines for Large-scale Data Management, **H**PC, and Machi**ne** Learning; DAPHNE daughter of river god Peneus (fountains, streams), chased by Apollo

[Louvre, Paris]

DAPHNE

https://daphne-eu.eu/

The DAPHNE project is funded by the European Union's Horizon 2020 research and innovation program under grant agreement number 957407 for the time period from Dec/2020 through Nov/2024.

DAPHNE: An Open and Extensible System Infrastructure for Integrated Data Analysis Pipelines

Patrick Damme, Marius Birkenbach, Constantinos Bitsakos, Matthias Boehm, Philippe Bonnet, Florina Ciorba, **Mark Dokter**, Pawel Dowgiallo, Ahmed Eleliemy, Christian Faerber, Georgios Goumas, Dirk Habich, Niclas Hedam, Marlies Hofer, Wenjun Huang, Kevin Innerebner, Vasileios Karakostas, Roman Kern, Tomaž Kosar, Alexander Krause, Daniel Krems, Andreas Laber, Wolfgang Lehner, Eric Mier, Marcus Paradies, **Bernhard Peischl**, Gabrielle Poerwawinata, Stratos Psomadakis, Tilmann Rabl, Piotr Ratuszniak, Pedro Silva, Nikolai Skuppin, Andreas Starzacher, Benjamin Steinwender, Ilin Tolovski, Pınar Tözün, Wojciech Ulatowski, Yuanyuan Wang, Izajasz Wrosz, Aleš Zamuda, Ce Zhang, Xiao Xiang Zhu

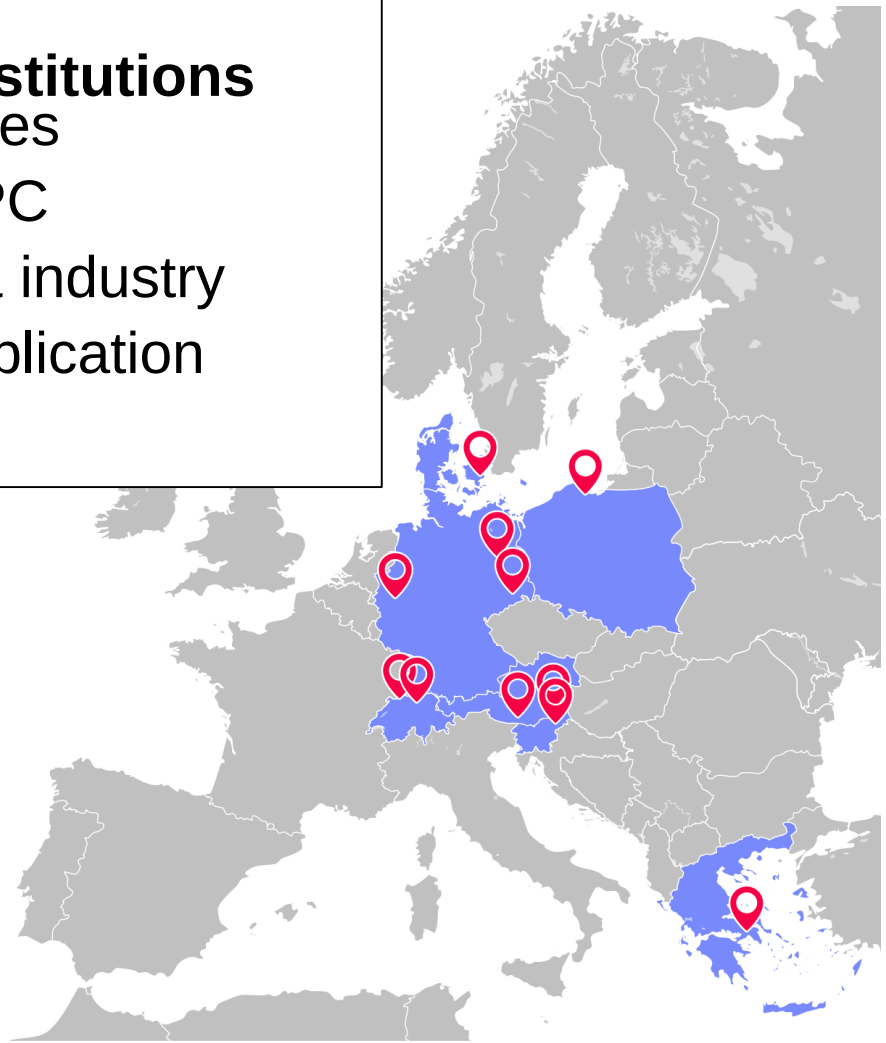
DATAWEEK 2023, June 13-15, Luleå, Sweden

Project Consortium

DAPHNE

14 partner institutions
from 7 countries

- DM, ML, HPC
- Academia & industry
- Different application domains



- Know-Center GmbH (**coordinator**), Austria
- AVL List GmbH, Austria
- Deutsches Zentrum fuer Luft- und Raumfahrt e.V., Germany
- Eidgenoessische Technische Hochschule Zuerich, Switzerland
- Hasso-Plattner-Institut for Digital Engineering gGmbH, Germany
- Institute of Communication and Computer Systems, Greece
- Infineon Technologies Austria AG, Austria
- , Intel Technology Poland sp. z o.o., Poland
- IT-Universitetet i København, Denmark
- Kompetenzzentrum Automobil- und Industrieelektronik GmbH, Austria
- Technische Universität Dresden, Germany
- Univerza v Mariboru, Slovenia
- Universitaet Basel, Switzerland
- Technische Universität Berlin

Integrated Data Analysis (IDA) Pipelines

DAPHNE

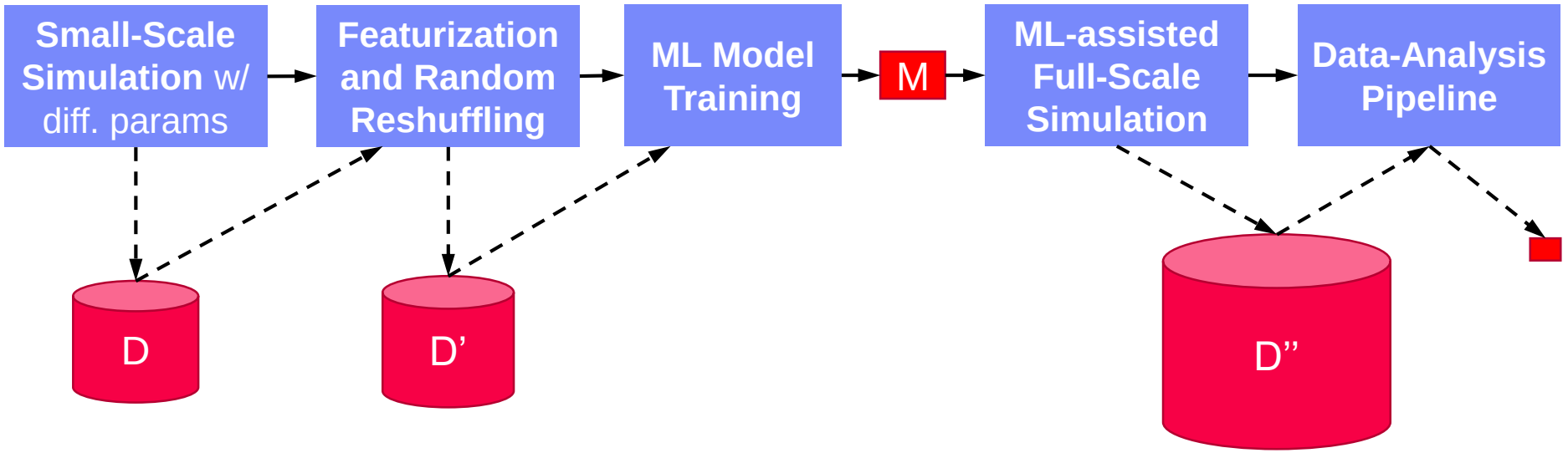
DM + **ML** + **HPC**

Data Management & query processing

Machine Learning training & scoring

High-Perf. Computing custom codes & simulations

Example: ML-assisted simulation



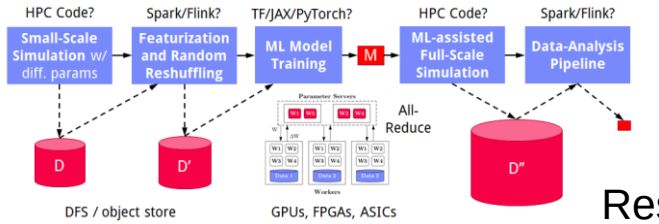
Challenges

DAPHNE

- **Deployment Challenges**

Dev Teams Programming Models

Different Systems/ Libraries



Cluster Under-utilization

Data/File Exchange

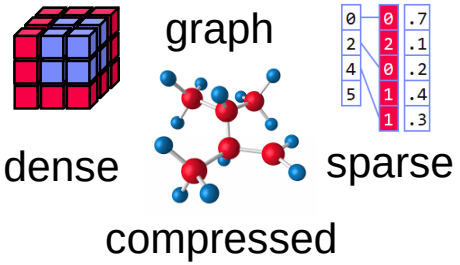
Resource Managers

➔ DAPHNE Overall Objective: Open and extensible system infrastructure

- **Hardware Challenges**
  - DM+ML+HPC share compilation and runtime techniques / converging cluster hardware
  - **End of Dennard scaling:** P = α CFV2 (power density 1)
  - **End of Moore's law**
  - **Amdahl's law:** sp = 1/s

➔ Increasing Specialization

#1 Data Representations



Sparsity Exploitation from Algorithms to HW

#2 Data Placement

Local vs **distributed**

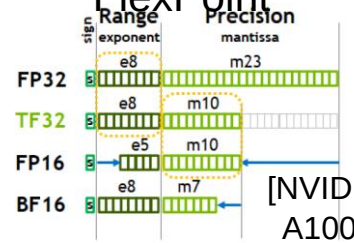
CPUs/ NUMA

GPUs

FPGAs/ ASICs

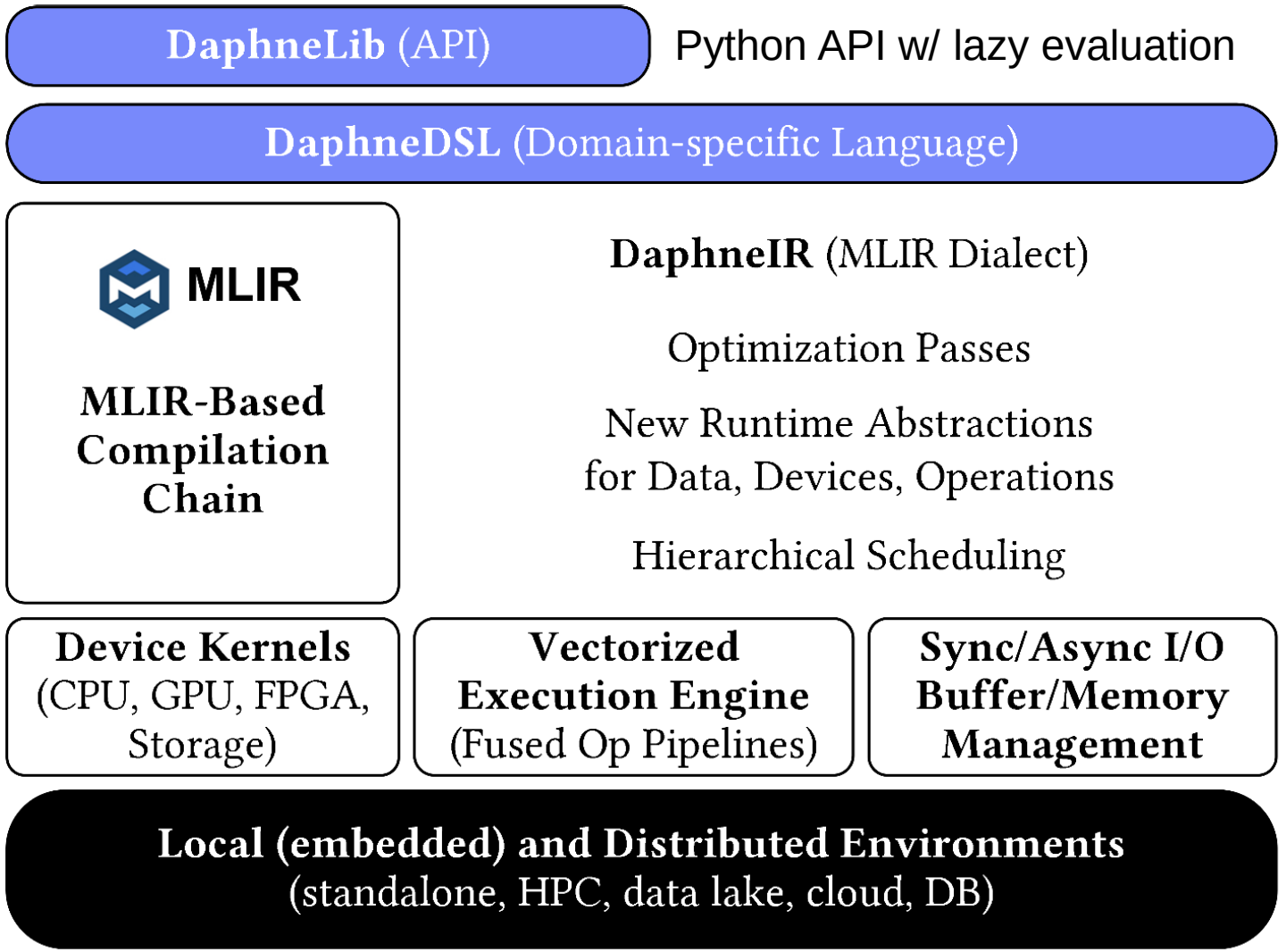
#3 Data (Value) Types

FP32, FP64, INT8, INT32, INT64, UINT8, BF16, TF32, FlexPoint



[NVIDIA A100]

System Architecture



Extensible Infrastructure

MLIR Dialects, Extension Catalog (new data types, kernels, scheduling algs)

Multi-level Compilation/ Runtime

Sideways Entry, DSL-level constraints (data formats & data/op placement)

Fine-grained Fusion and Parallelism

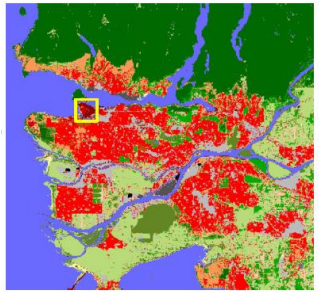
Integration w/ Resource Mgmt & Prog. Models

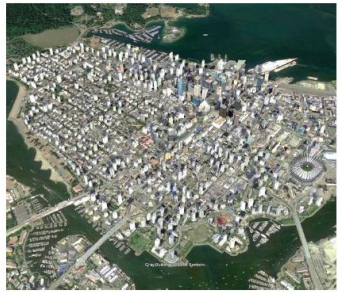
Example Use Cases

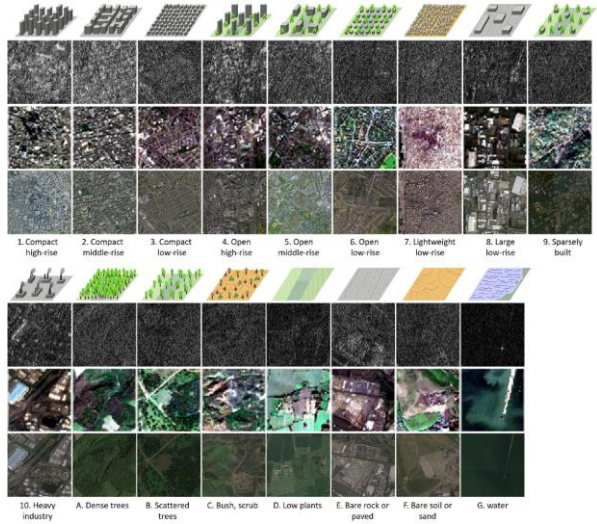
[Xiao Xiang Zhu et al: So2Sat LCZ42: A Benchmark Dataset for the Classification of Global Local Climate Zones. **GRSM 8(3) 2020**]

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

- **DLR Earth Observation**
  - **ESA Sentinel-1/2** datasets → 4PB/year
  - Training of local climate zone classifiers on **So2Sat LCZ42** (15 experts, 400K instances, 10 labels each, ~55GB HDF5)
  - **ML pipeline:** preprocessing, ResNet-20, climate models







- **IFAT Semiconductor Ion Beam Tuning**
- **KAI Semiconductor Material Degradation**
- **AVL Vehicle Development Process** (ejector geometries, KPIs)

- **ML-assisted simulations, data cleaning, augmentation**
- **Cleaning during exploratory query processing**

Automotive Use Cases

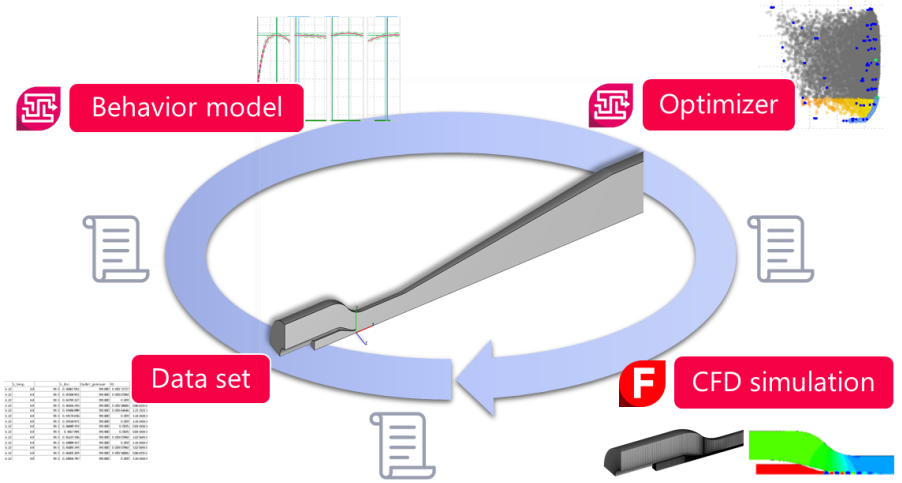
DAPHNE

ICT51 - extreme data

AVL

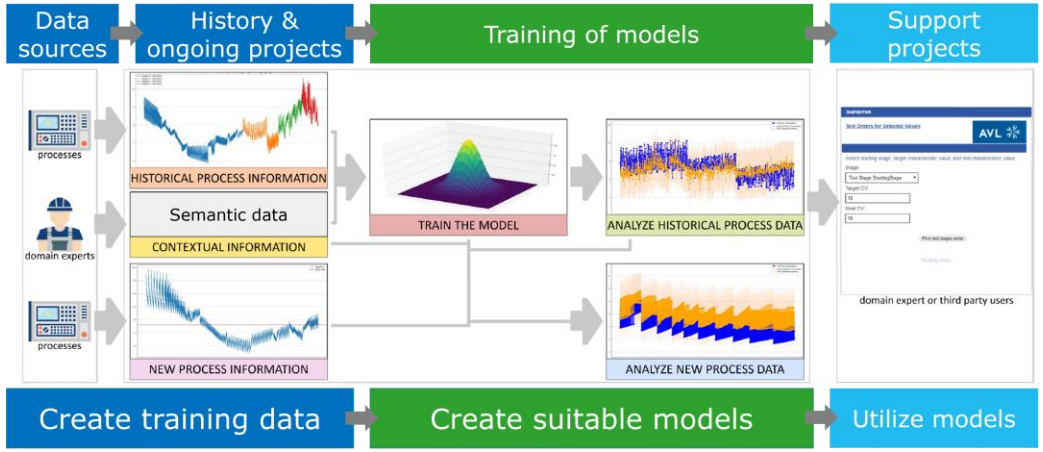
- Optimize ejector contour for target suction pressure
  - Geometric variables (lengths, diameters, radii) at different operating conditions
- Iterative process: searching param, set via CFD sim
- Volume: nightly optimization runs
- Value: combining physics-based (Active DoE) with Kernel Rich regression model (DAPHNE DSL)

Fuel cell: Ejector geometry optimization



ICT51 - extreme data

AVL

- Systematic data management
- Highly structured data
- Predict vehicle KPIs (e.g., fuel consumption)
- Volume: perform sim. to obtain values over time
- Value: align the values overt time across multiple dimensions to obtain synthetic time series (DAPHNE DSL & open API)

Synthetic data: automotive process mining