EUROPEAN BIG DATA VALUE FORUM

Fueling Industrial AI with Data Pipelines

Tuesday, Nov 22, 2022

16:00 - 17:30

DataCloud enRichMyData GRAPH MASSIVIZER

ORGANISED BY:

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH

European Commission

Plan4all

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

UNDER THE AUSPICES OF

EU2022.CZ

BDV BIG DATA VALUE ASSOCIATION

DataCloud

EUROPEAN BIG DATA VALUE FORUM

Fueling Industrial AI with Data Pipelines

Tuesday, Nov 22, 2022
16:00 – 17:30

Radu Prodan
University of Klagenfurt

Dumitru Roman
SINTEF AS

Fernando Perales
JOT Media

Irena Pavlova
Sofia University

Nikolay Nikolov
SINTEF AS

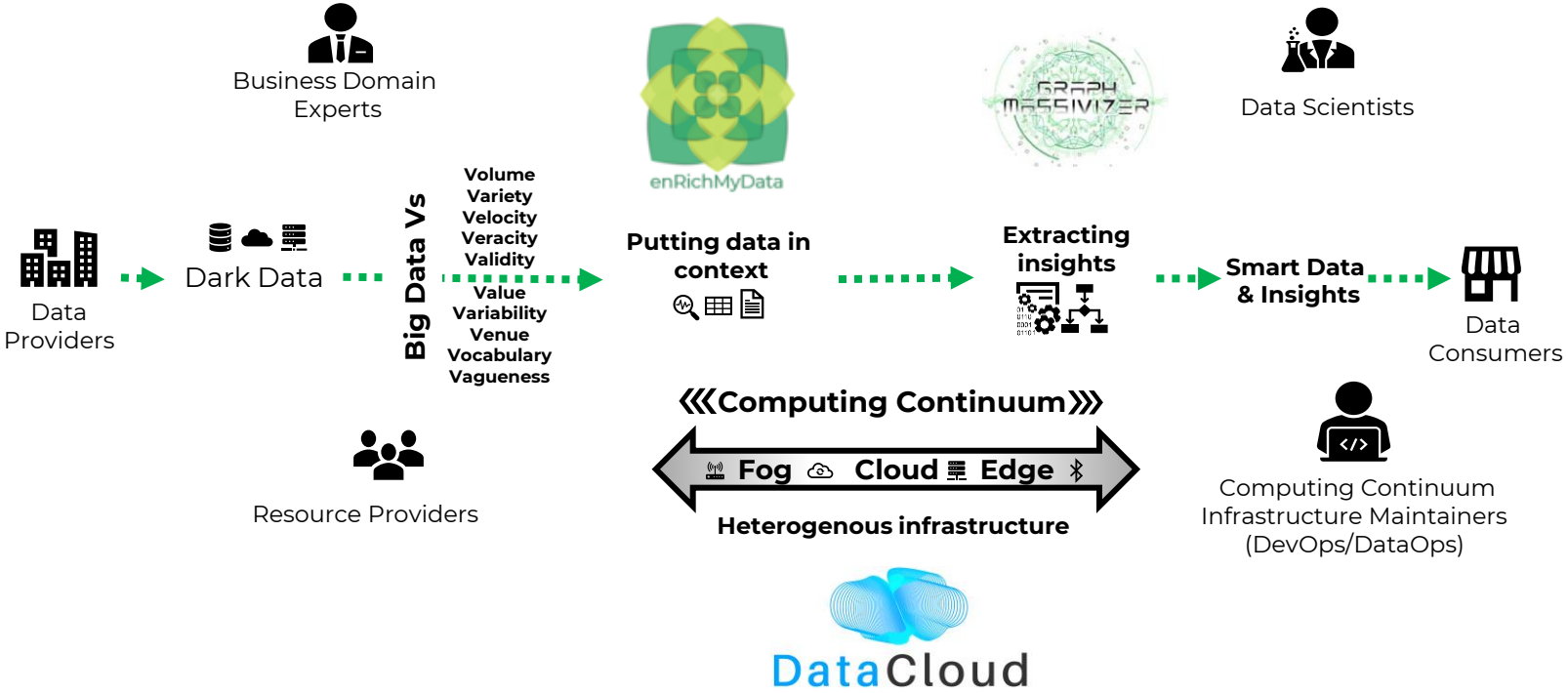
Silvia Castellvi
International Data Spaces Association

M Besher Massri
Jozef Stefan Institute

Rob Smeets
Philips

Till Christopher Lech
SINTEF AS

Overall context



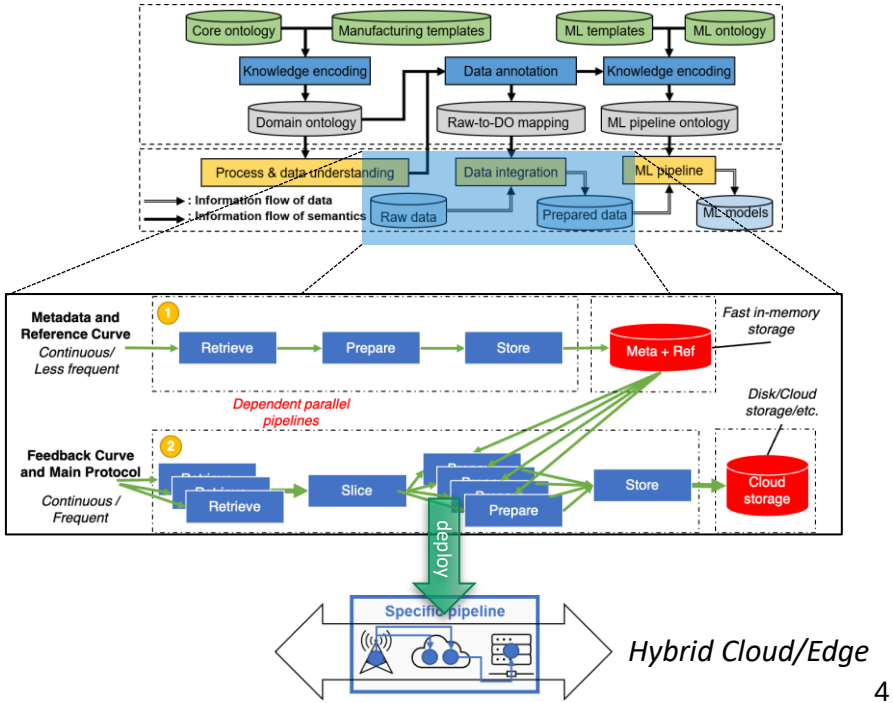
Example use case

Big Data pipelines are composite pipelines for processing data with non-trivial properties and characteristics, commonly referred to as the *Vs of Big Data*.

Industrial example: Data preprocessing for quality prediction of spot welding industrial process in car factories

Goal: To **ensure high quality** of welding spots, **predict failures** before they happen

- 1 assembly line = 500-1500 car bodies/day
- 1 x car body ~ 6 000 welding spots
  - 1 factory = ~6 million welding spots per month
  - Each welding spot = at least stored in four types of datasets
- Heterogeneous infrastructure for deployment of pipeline (Cloud + Edge)



Session agenda

1. Presentations
    1. Data enrichment and data enrichment pipelines (enRichMyData)
    2. Example use case (digital marketing)
    3. Big data pipelines on the Computing Continuum (DataCloud)
    4. Analyzing massive graphs (Graph-Massivizer)

2. Panel

Data enrichment and data enrichment pipelines (enRichMyData)

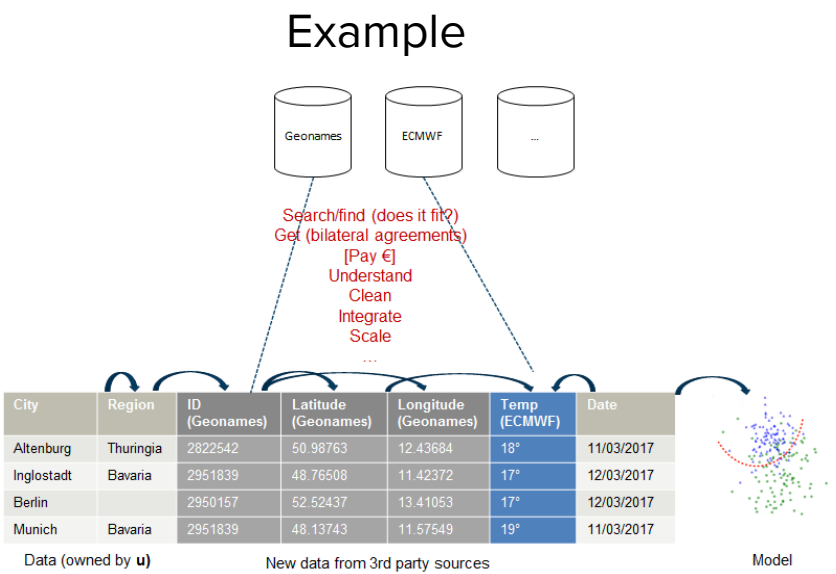
Data enrichment

Intuitively: **putting data in context, typically add more data to an input data set (structured / semi-structured / texts), improving data quality.**

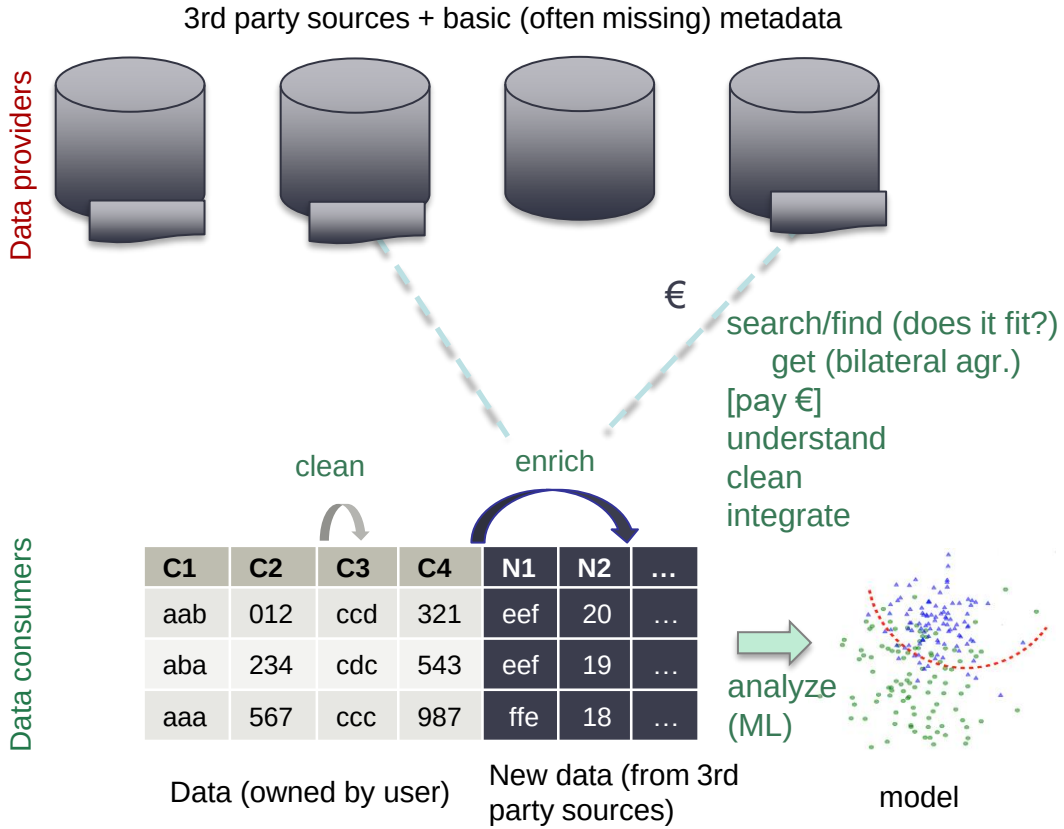
- Representative operations part of data enrichment
  - reconciliation/linking
  - extension/augmentation
  - cleaning and preparation
  - extraction of entities, keyphrases, etc
  - classification
  - profiling
  - ...

Compared to traditional data integration, the data to enrich *with* (the reference datasets) are **previously unknown to** and **not under the control of** the user who is enriching data (the source dataset)

- Introduces new challenges: search/find, understand, assess, access (e.g., due to limited control)



Data Enrichment in the Current Data Ecosystem

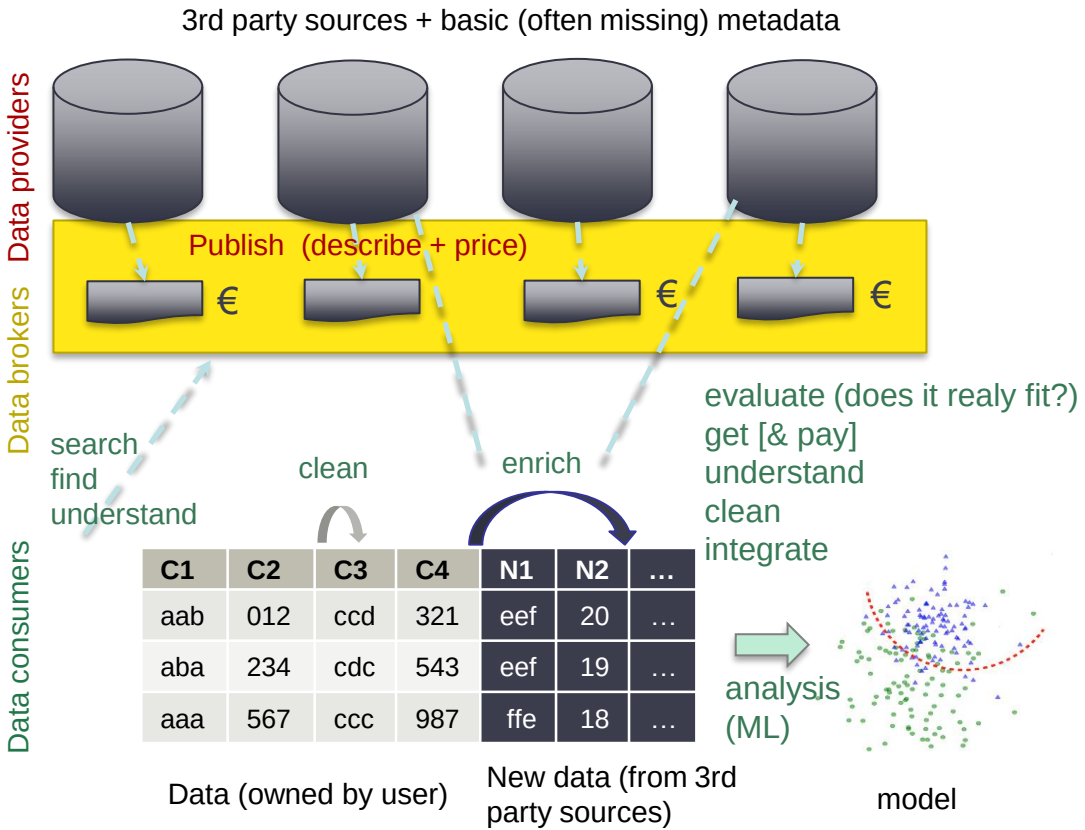


Consumers must:

- Search/find useful data themselves
- Pay with bilateral agreements
- Understand new data
- Clean / integrate the new data (they don't know well!)
- User their models to analyze the data (they don't know well!)

The full process is not supported…

Data Enrichment in the Current Data Ecosystem – with Data Marketplaces



Providers must

- Publish the data (describe + price)

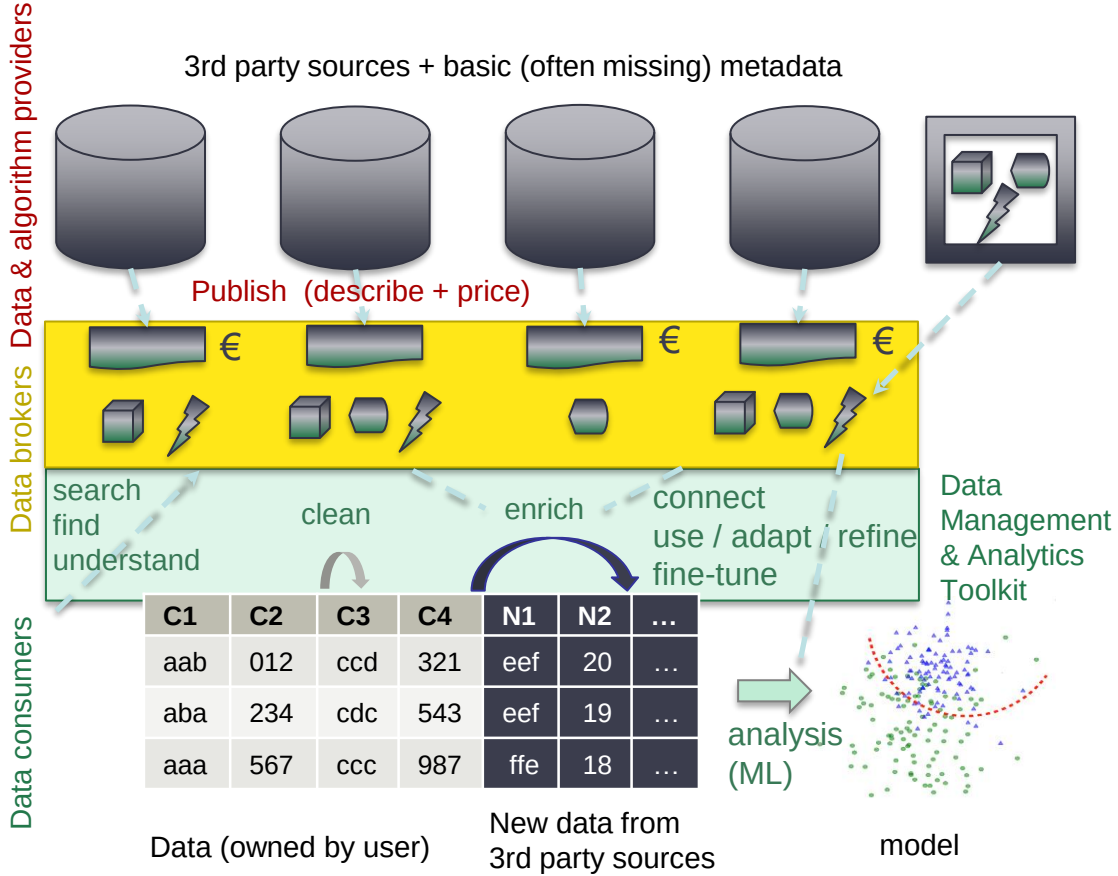
Data brokers support

- Search/find/understand (business-level)
- Payment

Users must

- Understand the new data
- Clean / integrate the new data (they don't know well!)
- User their models to analyze the data (they don't know well!)

Consumer-centered Data Enrichment in the Future Data Ecosystems



Providers must

- Publish the data (describe + price)
- Publish algorithms as a service

Data brokers support

- Search/find/understand (business and content-level via semantic profiling)
- Payment / access
- Recommend (best data for …)

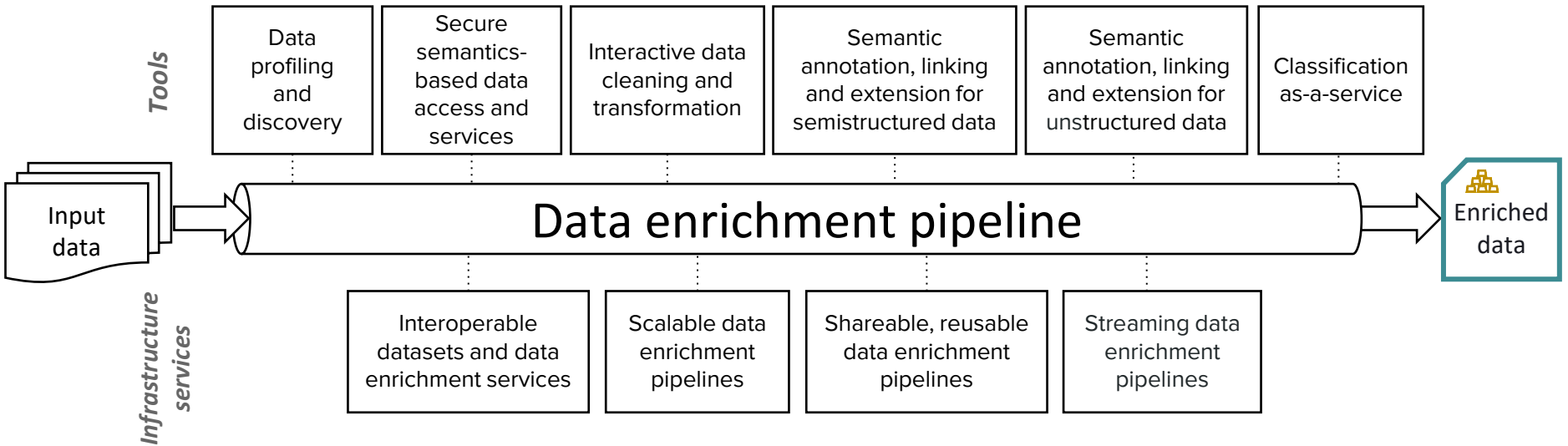
Tool-supported users must

- Connect
- Use / adapt / refine
- Fine-tune

Data Enrichment Pipelines

- Data enrichment pipeline: data enrichment typically requires support for a set of interconnected activities/steps to solve a particular data enrichment task
- Why is it challenging to support data enrichment pipelines?
  - Limited knowledge of reference datasets
  - Lack of holistic data enrichment lifecycle support (creation, validation, deployment, execution, reuse)
  - Difficulty to support enrichment of/with Big Data (scalability)
  - Lack of involvement of Humans-In-The-Loop (HITL) for addressing (assessing, tuning, correcting) uncertainty in data enrichment activities/steps
  - Difficulty in reusing and adapting data enrichment pipelines
  - Difficulty in making reference datasets and enrichment services FAIR
  - Lack of tools to support the complete data enrichment pipelines lifecycle for a wide range of shareholders from domain experts to data scientists and engineers

enRichMyData: Data Enrichment Pipelines (at scale)

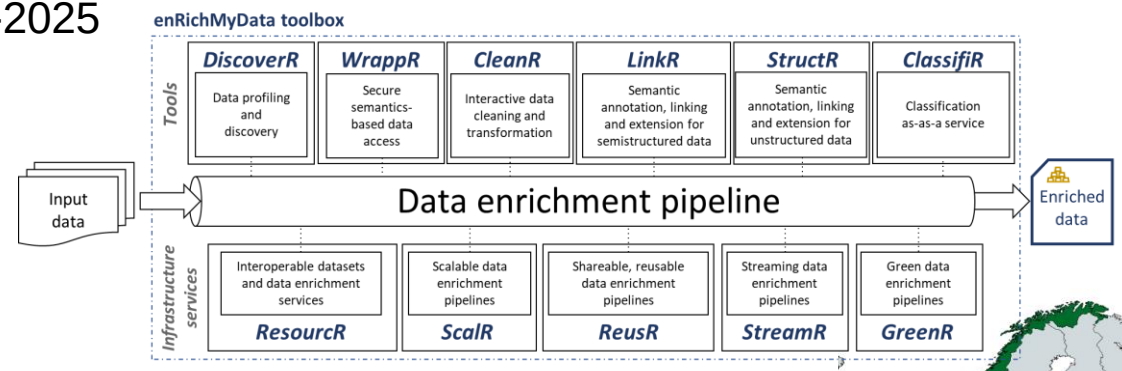


Wide variety of use cases

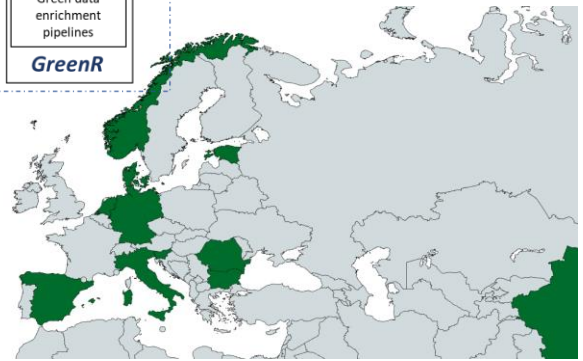
- (***Digital Marketing***) Marketing data enrichment for smart-bidding optimization
- (***Manufacturing***) AI-based Welding Analytics
- (***Predictive Maintenance***) Smart Maintenance of Medical Imaging Systems
- (***Public Procurement***) European Register of Entities from Known Actions
- (***Innovation Ecosystems***) Innovation Knowledge Graph
- (***Mineral Processing***) Mineral Processing Optimization

enRichMyData project

- Enabling Data Enrichment Pipelines for AI-driven Business Products and Services
- HORIZON-CL4-2021-DATA-01-03: Technologies for data management (IA)
- 2022-2025



Partners: SINTEF, Phillips, Bosch, Spend Network, JOT, CS Group, Expert AI, Sirma, University of Milano Bicocca, Jozef Stefan Institute, University of Copenhagen, Sofia University, BGRIMM Technology Group



Example use case (digital marketing)

enRichMyData

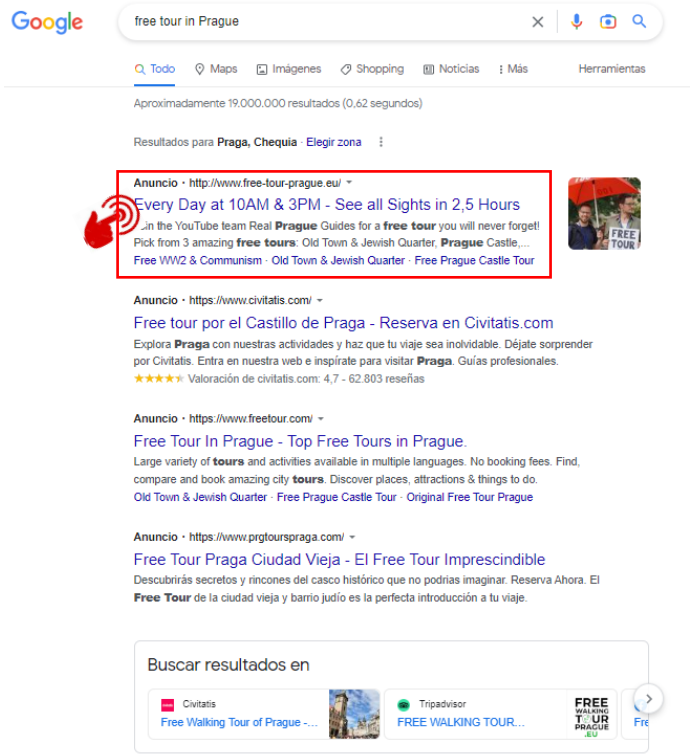
Marketing data **En**richment for smart-bidding optimization **(MarkEn)**

Dr. Fernando Perales
(fernando.perales@jot-im.com)

EUROPEAN BIG DATA VALUE FORUM
21-23 NOV | PRAGUE - CZECH REPUBLIC

JOT INTERNET MEDIA

The Digital Marketing Business



Performance Data

clicks
impressions
Conversion rate
Date
Location
Keyword
Category
COSTS
Revenue
Ad Text

JOT INTERNET MEDIA

The Digital Marketing Business

JOT INTERNET MEDIA

Value Proposition

Definition – Development – Implementation of data enrichment processes with marketing related data

Develop automated applications and algorithms in order to obtain a better Return of Investment

Provide full information to account manager to decide how to distribute the investments.

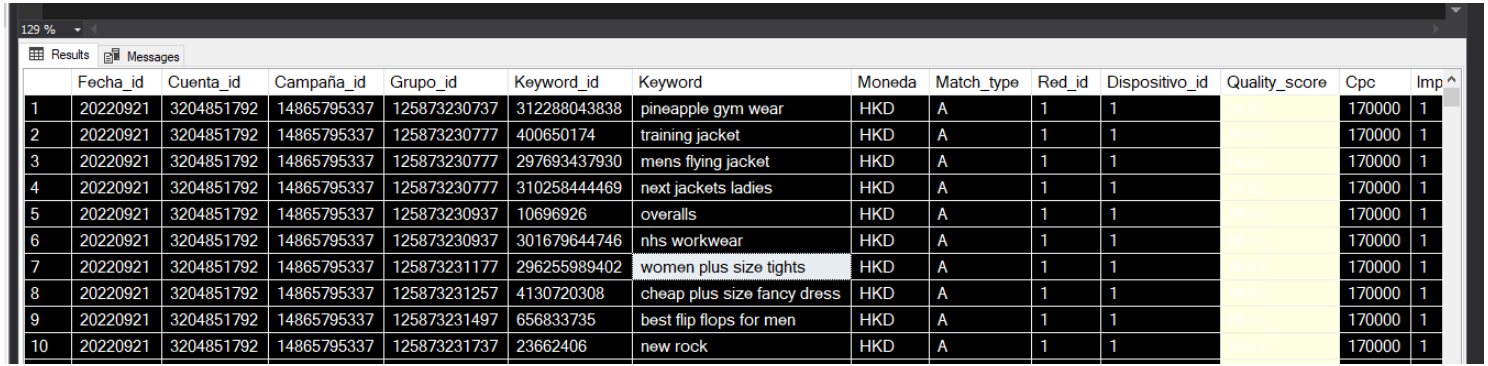
Generate new data monetization revenue stream

JOT INTERNET MEDIA

Data sets

- Data Volume
  - Keyword report: 1,13 Gb per day, 35 GB monthly
  - Since: 2022/01/01 more than 100M rows

50 variables per row



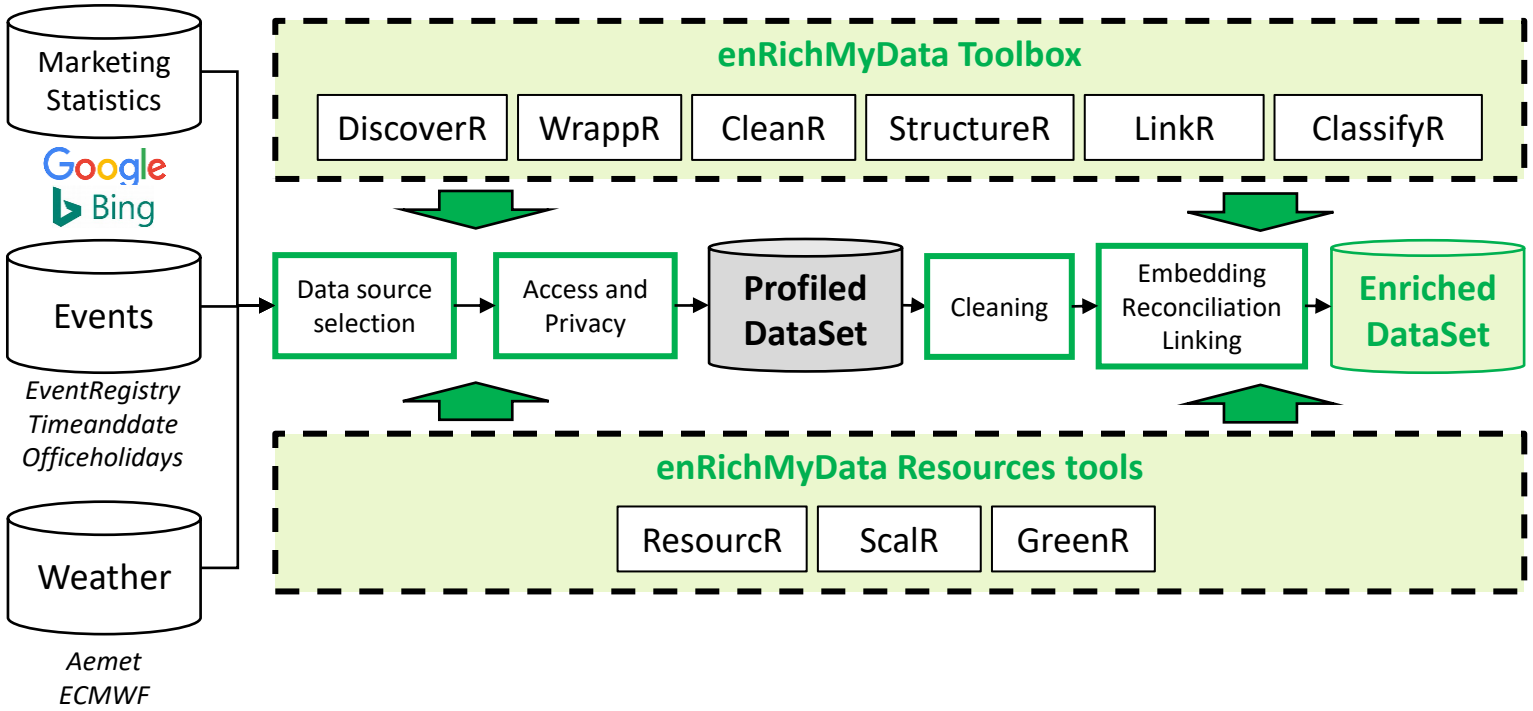
129 %

Results | Messages

	Fecha_id	Cuenta_id	Campaña_id	Grupo_id	Keyword_id	Keyword	Moneda	Match_type	Red_id	Dispositivo_id	Quality_score	Cpc	Imp
1	20220921	3204851792	14865795337	125873230737	312288043838	pineapple gym wear	HKD	A	1	1	NULL	170000	1
2	20220921	3204851792	14865795337	125873230777	400650174	training jacket	HKD	A	1	1	NULL	170000	1
3	20220921	3204851792	14865795337	125873230777	297693437930	mens flying jacket	HKD	A	1	1	NULL	170000	1
4	20220921	3204851792	14865795337	125873230777	310258444469	next jackets ladies	HKD	A	1	1	NULL	170000	1
5	20220921	3204851792	14865795337	125873230937	10696926	overalls	HKD	A	1	1	NULL	170000	1
6	20220921	3204851792	14865795337	125873230937	301679644746	nhs workwear	HKD	A	1	1	NULL	170000	1
7	20220921	3204851792	14865795337	125873231177	296255989402	women plus size tights	HKD	A	1	1	NULL	170000	1
8	20220921	3204851792	14865795337	125873231257	4130720308	cheap plus size fancy dress	HKD	A	1	1	NULL	170000	1
9	20220921	3204851792	14865795337	125873231497	656833735	best flip flops for men	HKD	A	1	1	NULL	170000	1
10	20220921	3204851792	14865795337	125873231737	23662406	new rock	HKD	A	1	1	NULL	170000	1

JOT INTERNET MEDIA

Data flow



JOT INTERNET MEDIA

Innovation

BUSINESS:

- To distribute investment in profitable keywords
- Improve investment quality

TECHNICAL:

- Being able to develop algorithms/methods to include the enriched pipeline in the current system
- Implement keyword list extension based on Semantic Similarity
- Multi criteria keyword classification for impact optimization

JOT INTERNET MEDIA

Impact on the business - KPIs

Classify information by a monthly or a weekly level, including seasonality.

Combine external and internal data sources with enrichment services

Extend keywords with 25 semantically related keywords.

Improve ROI from 10% and 25%

JOT INTERNET MEDIA

Main Challenges and Risks

Challenges

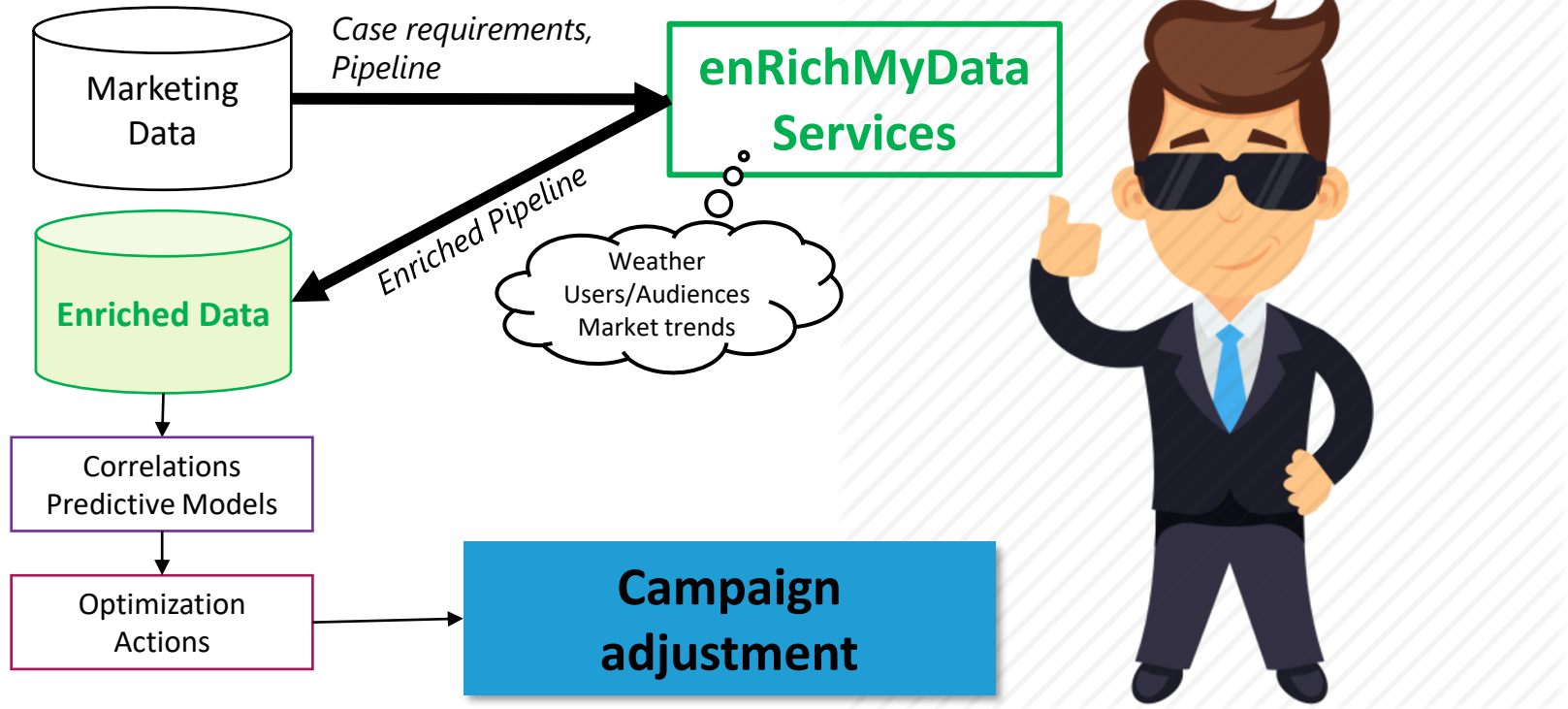
- Being able to provide the whole information in the way it is needed.
- Including the enriched pipeline provided in our system.

Risks

- Lack of knowledge generating data pipeline.
- The inability to include the pipeline and not manage to match the data with our database.

JOT INTERNET MEDIA

Minimum Viable Product (MVP)



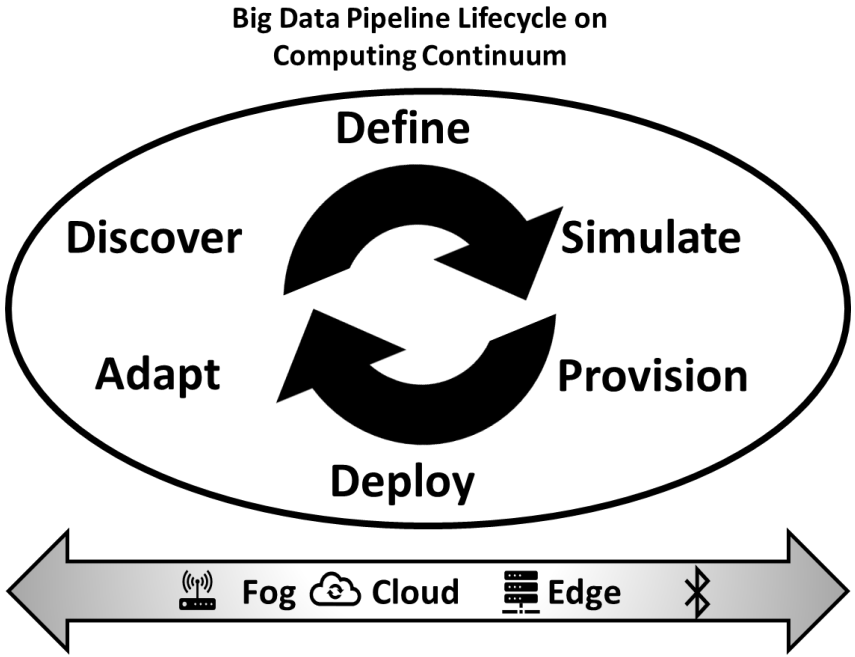
JOT INTERNET MEDIA

Big data pipelines on the Computing Continuum (DataCloud)

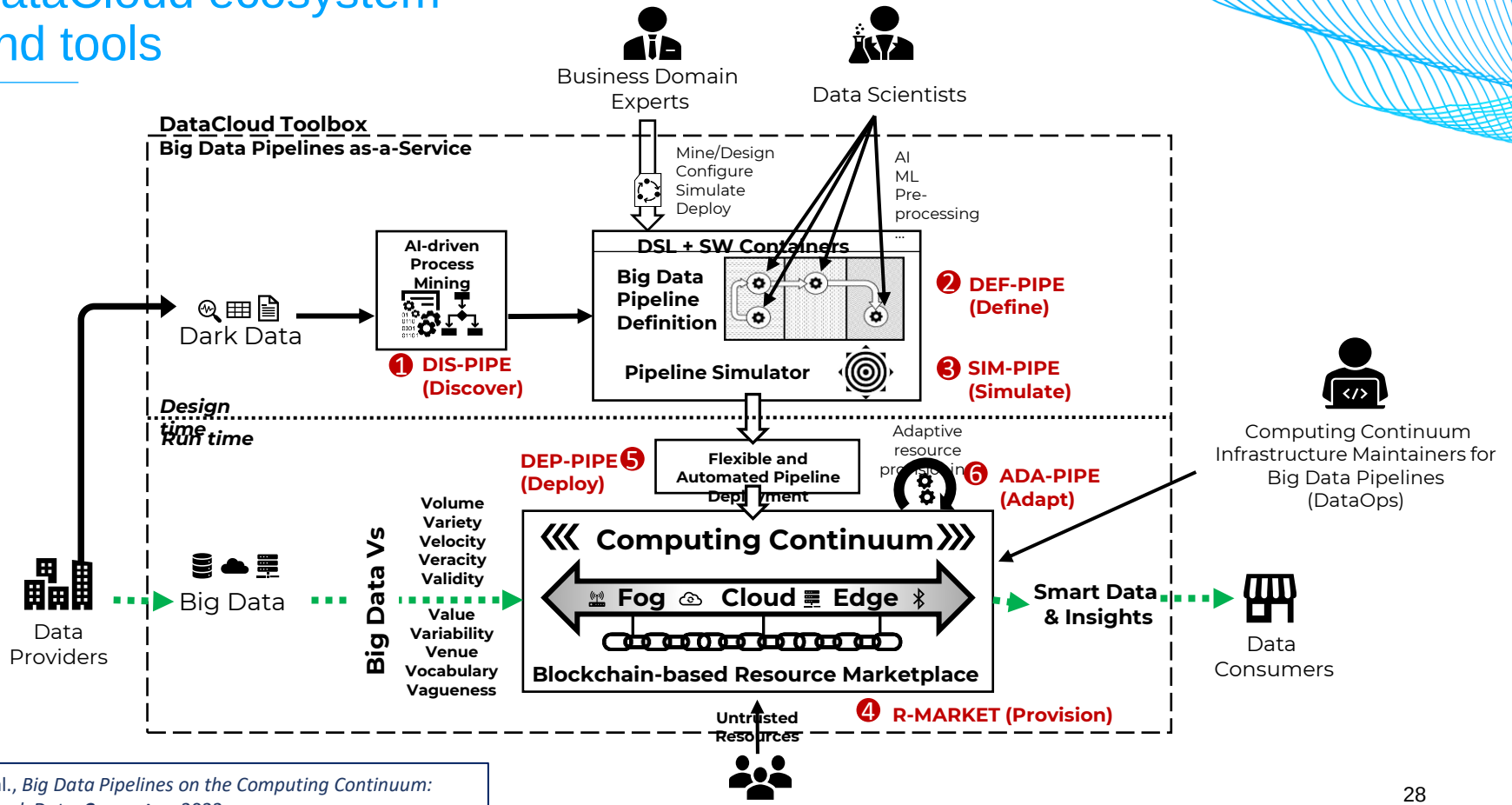
DataCloud Scope

Example: 1TB database with sensor measurements

Retrieve sensor data
Reformat data
Split data in chunks
Enrich split data
Perform ML training

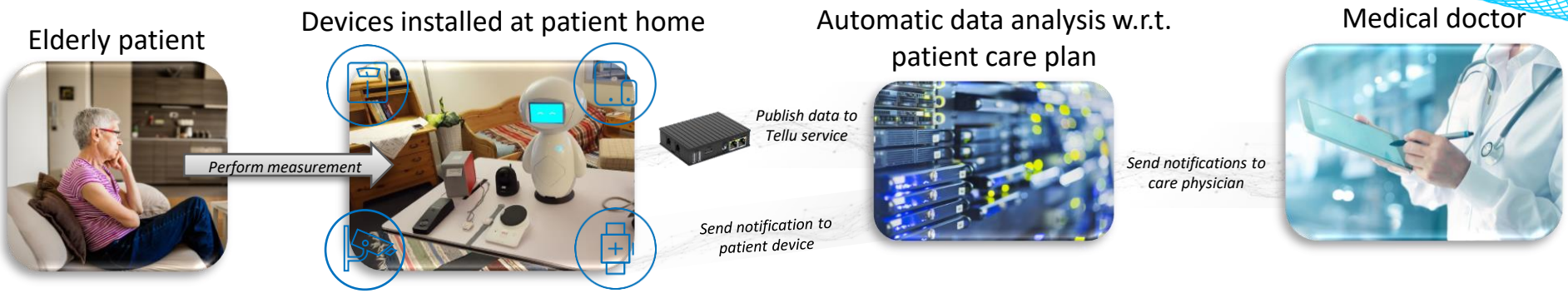


DataCloud ecosystem and tools



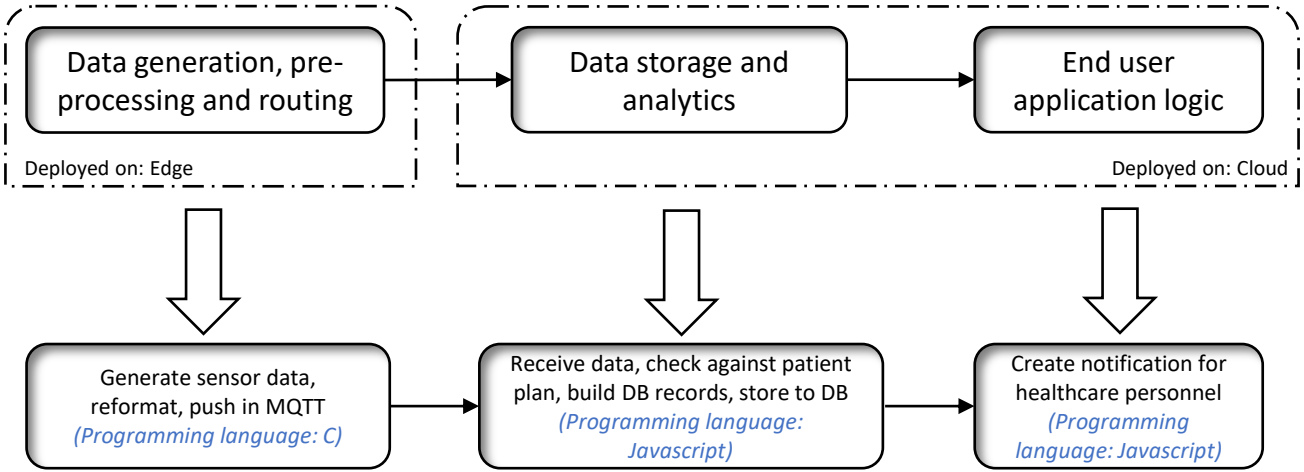
D. Roman et al., *Big Data Pipelines on the Computing Continuum: Tapping the Dark Data*, **Computer**, 2022

Application of pipelines for digital healthcare (PoC)



- Existing Tellu application/product is not designed as a data pipeline:
  - Difficult to scale individual steps/pipelines of the data processing
  - Cannot easily be replicated for new patients/devices
  - Difficult to find the optimal resource allocation for data processing steps

Proof-of-concept Data Pipeline



Design-time: Designing a data pipeline
- Configure pipeline description
- Validate pipeline
- Test configuration (before deployment)

Run-time: Distributed deployment of the data pipeline
- Provision heterogenous Edge/Cloud resources
- Automatically schedule and deploy
- Execute and monitor the pipeline

Application execution logs

1 Visually discover pipeline from logs using **DIS-PIPE**

- Import logs and manage complexity
- Display graph-based representation
- Generate DSL for discovered pipeline

Partial pipeline model

2 Visually design/edit pipeline using **DEF-PIPE**

- Edit steps of designed pipeline
- Add pipeline-specific parameters to steps
- Show DSL representation of pipeline

Configured pipeline model

3 Simulate executions of pipeline using **SIM-PIPE**

- Execute simulation of data pipeline
- Display information about runs

Configured pipeline model and run stats

4 Provision resource for pipeline steps to be executed using **R-MARKET**

- Display information about resources
- Add resources using SDK and wallet

Provisioned resources

5 Schedule execution of pipeline using **ADA-PIPE**

- Schedule resources according to defined requirements
- Generate deployment-ready schedule

Deployment schedule

6 Deployment of pipeline on Cloud/Edge using **DEP-PIPE**

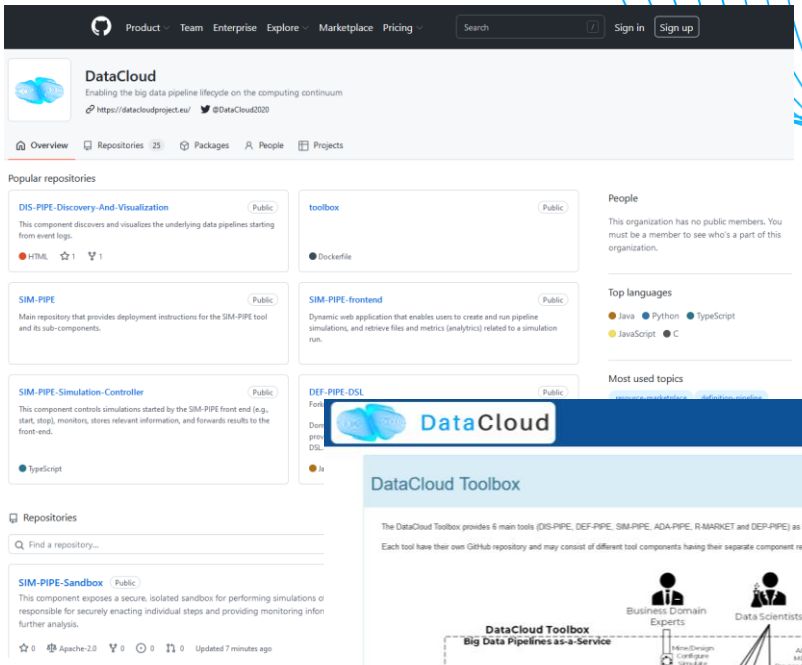
- Display available resources and providers
- Perform deployment of pipeline steps

7 **Pipeline execution**

- Sensor data simulation
- Notifications

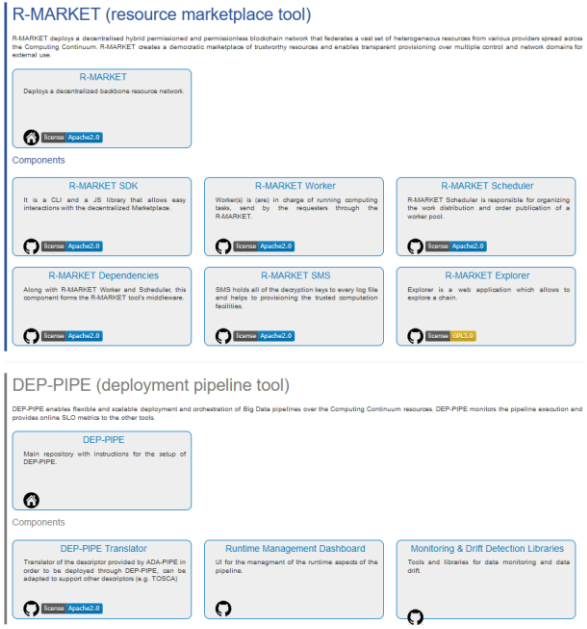
DataCloud Toolbox on GitHub

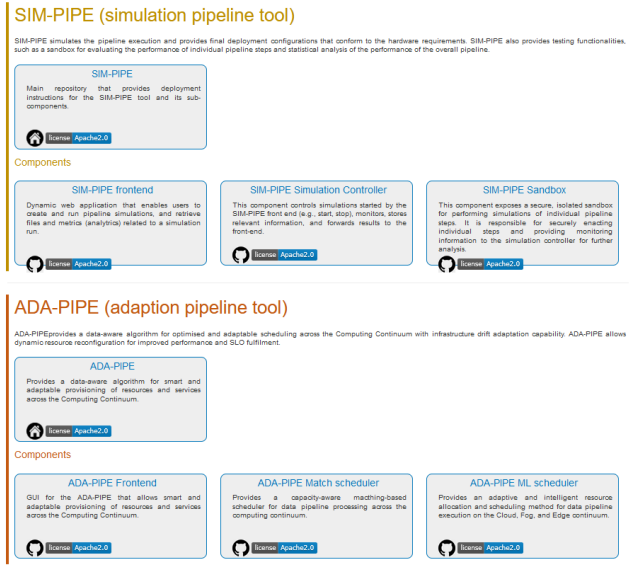
- DataCloud repository overview
  - https://github.com/DataCloud-project
  - DataCloud tool repositories
    - Documentation
    - License information
    - Source code
- DataCloud Toolbox overview
  - https://datacloud-project.github.io/toolbox
  - Section for each DataCloud tool
    - Main tool and components
    - Link to GitHub repositories
    - License icon



32

DataCloud Toolbox GitHub Overview



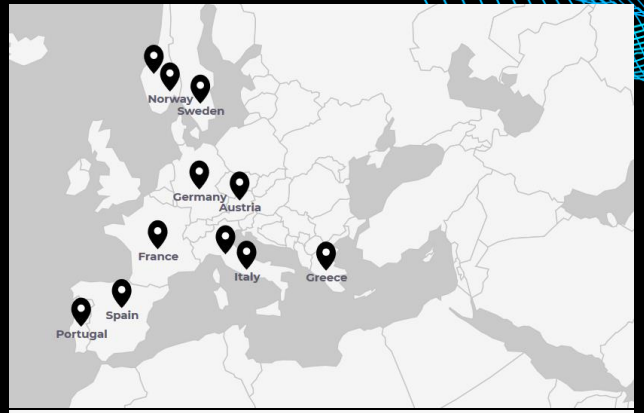


33

DataCloud project

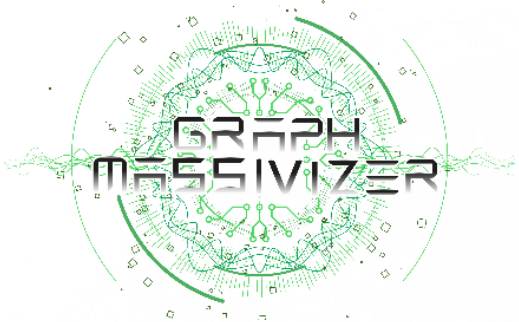
- DataCloud – Enabling the Big Data Pipeline Lifecycle on the Computing Continuum
- H2020-ICT-40-2020 Cloud Computing
- Research and Innovation Action, 2021-2023
- https://datacloudproject.eu

- Goals:
  - Software toolbox to support Big Data pipeline lifecycle on the Computing Continuum
  - Validation in five business cases (digital marketing, live media streaming, electronic healthcare, predictive maintenance, Industry 4.0)



SINTEF · Sapienza Università di Roma · Universitat Klagenfurt · Kungl Tekniska Högskolan · iExec · UBITECH · JOT Internet Media · MOG Digital Media · Catalano · tell.u · Bosch

Analyzing massive graphs (Graph-Massivizer)



UNIVERSITÄT KLAGENFURT

Graph-Massivizer

Radu Prodan

Andrea Bartolini, Marko Grobelnik, Peter Haase,
Alexandru Iosup, Evgeny Kharlamov, Gregor Leban,
Dumitru Roman, Roberta Turra,
Laurențiu Vasiliu, Ana Lucia Vărbănescu

Graph-Massivizer

- Extreme and Sustainable Graph Processing for Urgent Societal Challenges in Europe

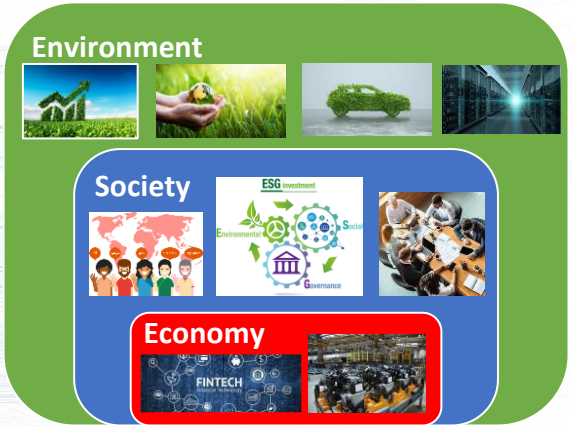
- HORIZON-CL4-2022-DATA-01-05
  - Extreme data mining, aggregation and analytics technologies and solutions (RIA)

- € 4,998,062.50

- 2023 – 2025

Goal

- High-performance and sustainable graph processing of extreme data
  - Holistic platform and integrated toolkit
  - Proper response for any need and any organization by 2030

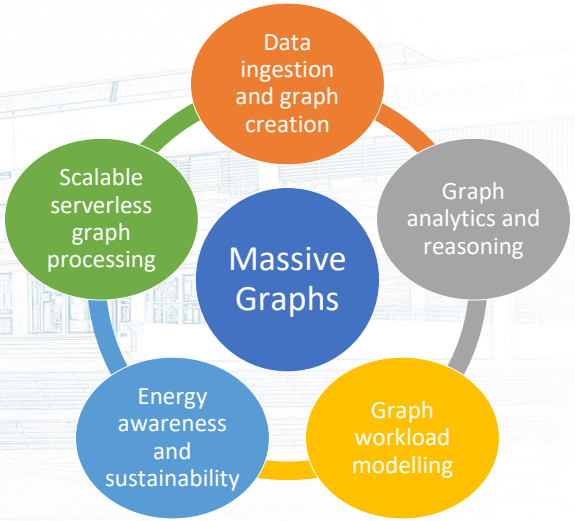


- Graphs
  - Universal mathematical abstractions
  - Capture, combine, model, analyze, and process knowledge about real and digital worlds

Massive Graph

- **Massive graph (MG)** representation of **extreme data**, integrating patterns and storing interlinked descriptions of objects, events, situations, concepts, and semantics
  - General graphs
  - Knowledge graphs
  - Property graphs

- **Extreme data challenge**, extending graph processing technology by orders of magnitude for several **"V"**-characteristics
  - **Volume** graph challenge by supporting up to billions of vertices and trillions of edges
  - **Velocity** graph challenge of dynamically changing topologies
  - **Viridescence** graph challenge for sustainable processing at scale

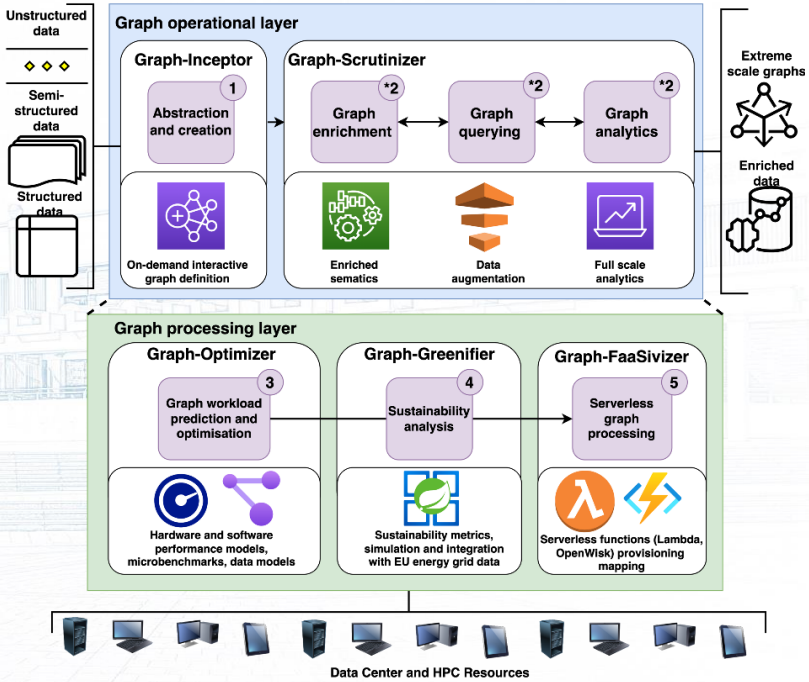


Graph-Massivizer Toolkit

metaphactory

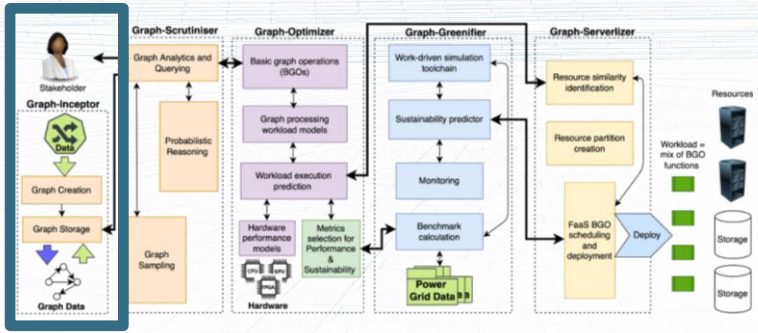
UNIVERSITÄT KLAGENFURT

- Graph operational layer
  - **Graph-Inceptor**: extreme data ingestion, MG creation and storage
  - **Graph-Scrutinizer**: MG analytic and reasoning
- Graph processing layer
  - **Graph-Optimizer**: workload modelling with performance and energy guarantees
  - **Graph-Greenifier**: sustainable and energy-aware MG processing
  - **Graph-Serverlizer**: scalable serverless MG analytics over a codesigned computing continuum



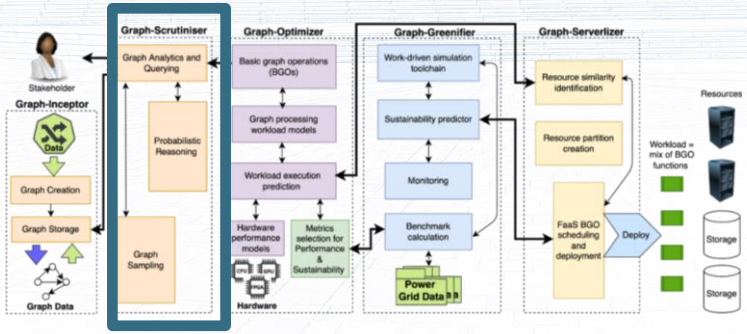
Graph-Inceptor

- Graph creation
  - ETL procedures to persist extreme data into MG of 10 billion vertices and 100 billion edges
  - Batch and stream data processing, compatible with Apache Spark
  - Streaming ingestion latency below 500 milliseconds for 95% of the data at 1,000s of new edges per second
- Graph storage
  - RDF graphs modelling semantic aspects
  - Property graphs modelled with edge properties
  - RDF-star graphs logically unifying the two views
  - Distributed for computation locality



Graph-Scrutinizer

- Graphs analytics and querying
  - Sampling and analytics throughput of 500 giga edges per second (Apache GraphX)
  - Streaming-graph reasoning speed of 3 million triples per second (PyTorch Geometric)
  - Retain 90% performance for the last 10 million edges compared to batch algorithms
- Graph sampling
  - Optimised analysis and query, retaining critical information
- Graph enrichment with expanding data sets
  - Probabilistic reasoning
  - Graph discovery, low footprint graph generation, real-time, error-bounded queries

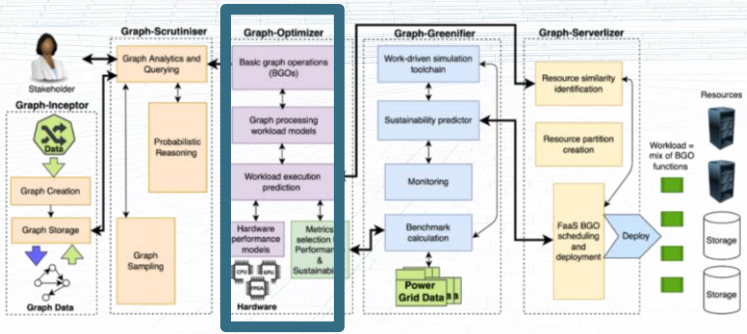


Graph-Optimizer

UNIVERSITY OF TWENTE.

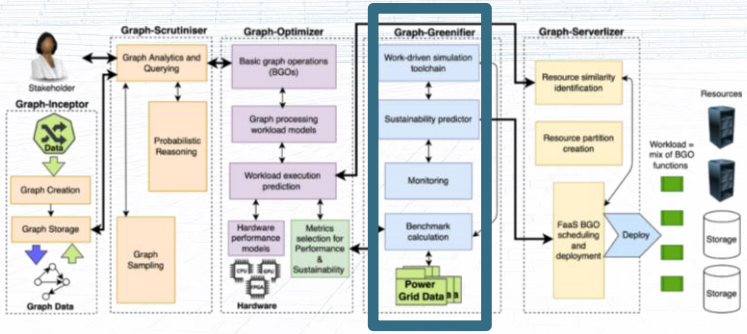
UNIVERSITÄT KLAGENFURT

- Hardware operation (H-op) models relevant to BGO
  - Runtime performance and energy costs (CPU, GPU, FPGA)
  - Memory latency, bandwidth, BGO throughput, I/O bandwidth
- Graph processing and workload models
  - Model BGO as compositions of H-ops (Graph-BLAS, GBTL-CUDA)
  - Generate workload models using design space exploration
- Workload execution prediction
  - 80% accurate performance and energy consumption models for BGOs
  - Various heterogeneous multiprocessors, GPUs and FPGAs



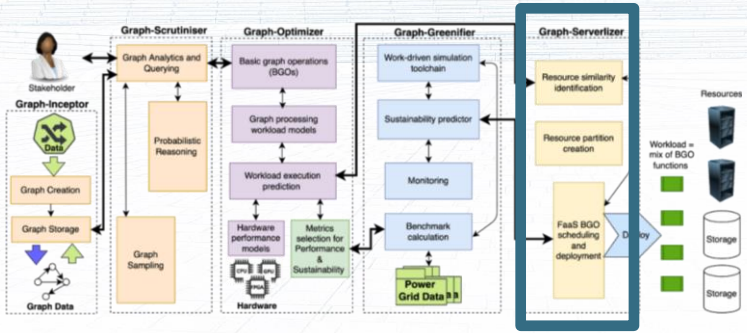
Graph-Greenifier

- Sustainability data collection
  - Monitoring infrastructure (hardware and software)
  - Power grid data interface (open market data)
- Work-driven graph processing simulation toolchain
  - Carbon footprint, CO2 and methane emissions
  - Open DC and SimLess simulators
- Benchmarking, prediction and recommendations for sustainability-performance tradeoffs at scale
  - Sustainability labelling for graph analytics of 100 individual nodes and devices
  - Two-fold improvement in data center energy efficiency and over 25% lower GHG emissions



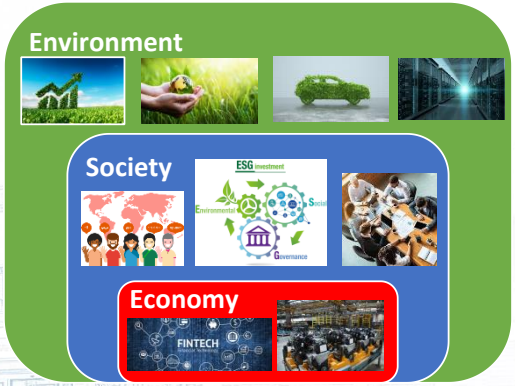
Graph-Serverlizer

- Workload-infrastructure codesign
  - Multilayer infrastructure facilitation
  - Resource partitioning based on similarity relationships (HPC, cloud, edge)
- Sustainable BGO function operation and scheduling
  - OpenWhisk, Lambda, …
  - 70% faster graph analytics over commercial AliGraph or other OSS solutions
  - 30% less energy for ETL storage than Amazon Redshift and OSS
- Cybersecure deployment
  - 40% faster runtime deployment compared to AWS CloudFormation
  - State-of-the-art security and privacy mechanisms



Use Cases

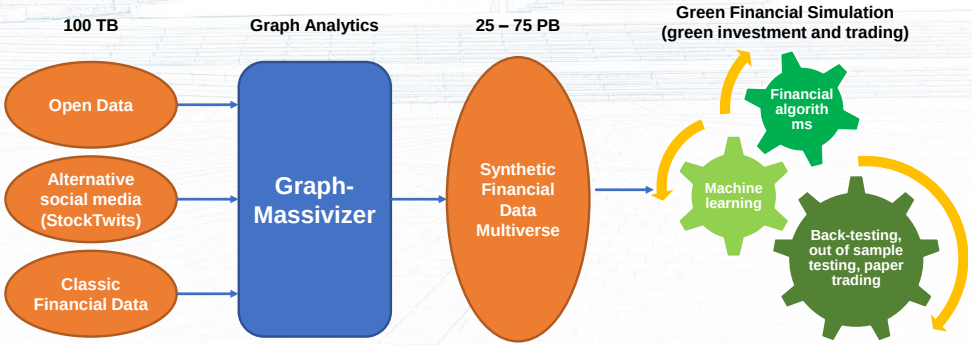
- UC-1: Green and sustainable finance
- UC-2: Global foresight for environment protection
- UC-3: Green AI for sustainable automotive industry
- UC-4: Data center digital twin for sustainable exascale computing



	Volume	Velocity	Value	Veracity	Variety	Viscosity	Viridescence
UC-1	✓		✓	✓			✓
UC-2	✓				✓	✓	✓
UC-3		✓		✓	✓		✓
UC-4	✓	✓			✓		✓

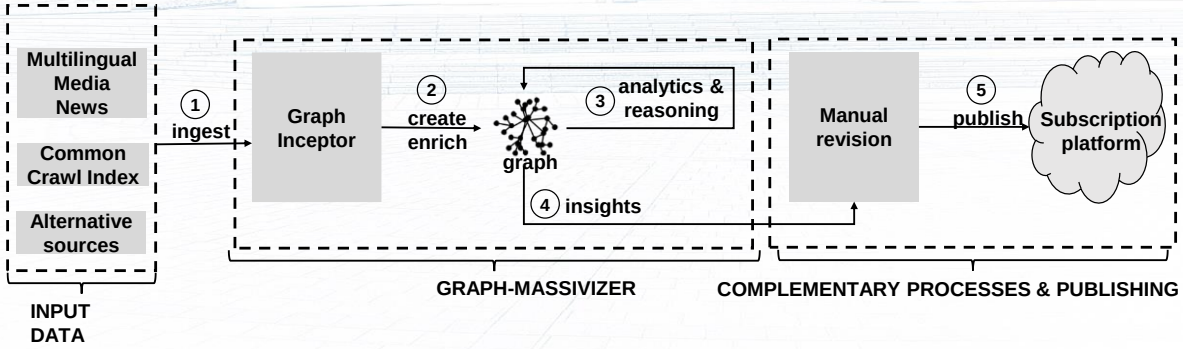
Green and Sustainable Finance

- Financial products, investments, and services that channel investments into green-focused companies
  - Green bonds, green lending, green equity investments
  - Limited financial data volume, reduced accessibility, high costs
- **Synthetic financial data multiverse**
  - Semi-automated creation of realistic and affordable synthetic extreme financial data sets, unlimited in size and accessibility
  - Improved and optimized green investments and trading



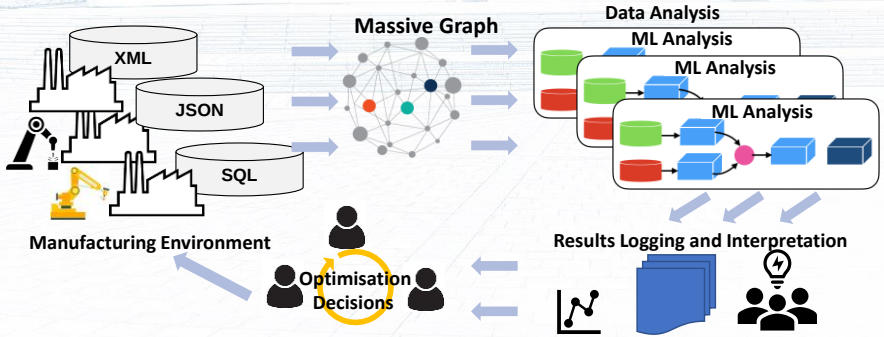
Global Foresight for Environment Prediction

- Geopolitical and business aspects of United Nation's sustainable environmental goals
  - Climate action, responsible production and consumption, clean water and sanitation, and clean and affordable energy
  - Foresights on future trends and scenarios guiding better policies limited to expert surveys
- **Environment protection foresight**
  - MG encoding Common Crawl, Linked Open Data Cloud, global media news
  - **Horizontal scanning** identifies threats, risks, issues and opportunities
  - **Megatrend analysis** explores relation of MG structures to shifts in behavior



Green AI for Sustainable Automotive Industry

- Value chains involving heterogeneous, distributed, fast-growing, disconnected and incompatible extreme data
  - Welding, production robots, car bodies-in-white (BiW), etc.
- **Green manufacturing line diagnosis**
  - MG integrating a collection of digital twins
  - OWL 2 ontologies and SHACL constraints
  - Green ML methods for MG alignment, summarization, search, pattern mining, quality assurance and trustworthiness analyses
  - Quality prediction of welding spots produced by welding machines, optimized configuration, remote diagnostics, repair schedules, accessory replacements

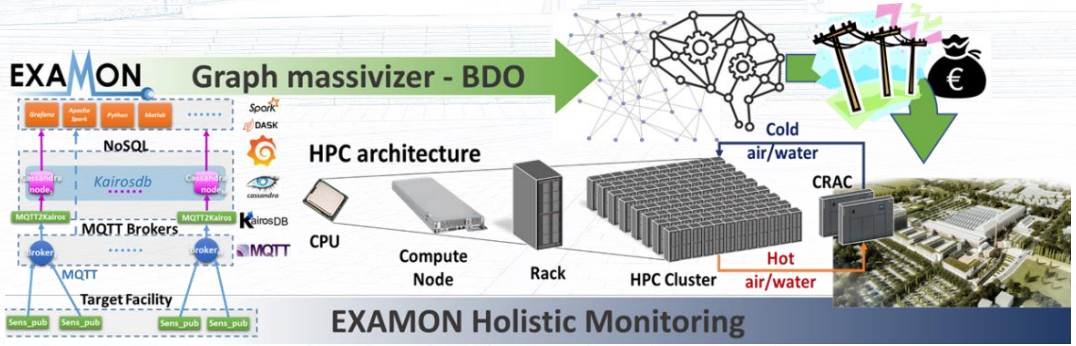


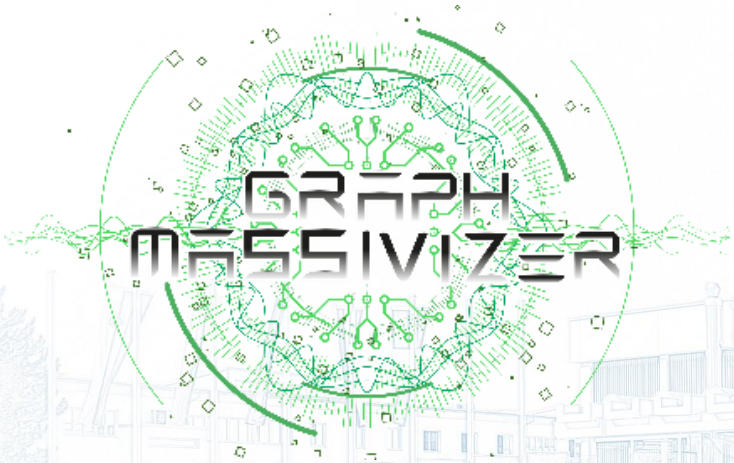
Data Center Digital Twin for Sustainable Computing

UNIVERSITÄT KLAGENFURT

CINECA

- Sustainable HPC science throughput limitations
  - Energy consumption for lowering TCO and GHG emissions
  - System utilization for improving science throughput
- **Data Centre Digital Twin**
  - MG capturing spatial, semantic, and temporal relationships between monitoring metrics, hardware nodes, cooling equipment and workloads
  - Predictive ML-based analytics and what-if simulations
  - Improved power use effectiveness (PUE) and GHG emissions
  - Improved resource utilization





THANK YOU !

Panel

BDV BIG DATA VALUE ASSOCIATION

DataCloud

EUROPEAN BIG DATA VALUE FORUM

Fueling Industrial AI with Data Pipelines

Tuesday, Nov 22, 2022
16:00 – 17:30

Radu Prodan
University of Klagenfurt

Dumitru Roman
SINTEF AS

Fernando Perales
JOT Media

Irena Pavlova
Sofia University

Nikolay Nikolov
SINTEF AS

Silvia Castellvi
International Data Spaces Association

M Besher Massri
Jozef Stefan Institute

Rob Smeets
Philips

Till Christopher Lech
SINTEF AS