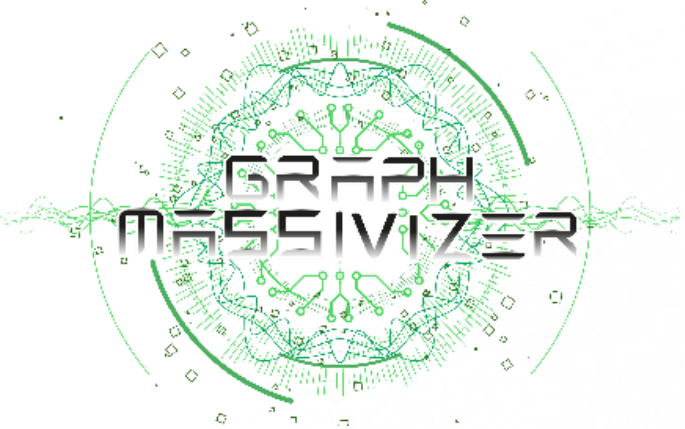


UNIVERSITÄT KLAGENFURT

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

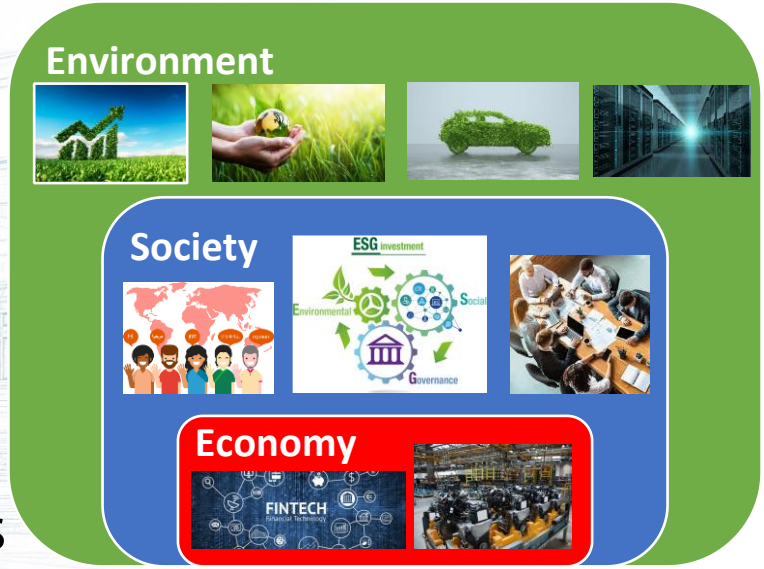
- HORIZON-CL4-2022-DATA-01-05
  - Extreme data mining, aggregation and analytics technologies and solutions (RIA)

- € 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

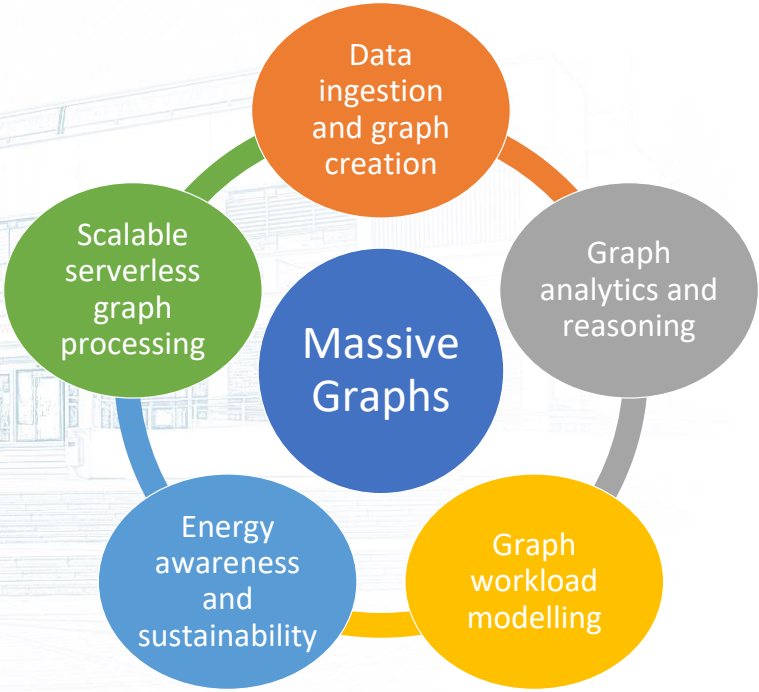


- Graphs
  - Universal mathematical abstractions
  - Capture, combine, model, analyze, and process knowledge about real and digital worlds

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

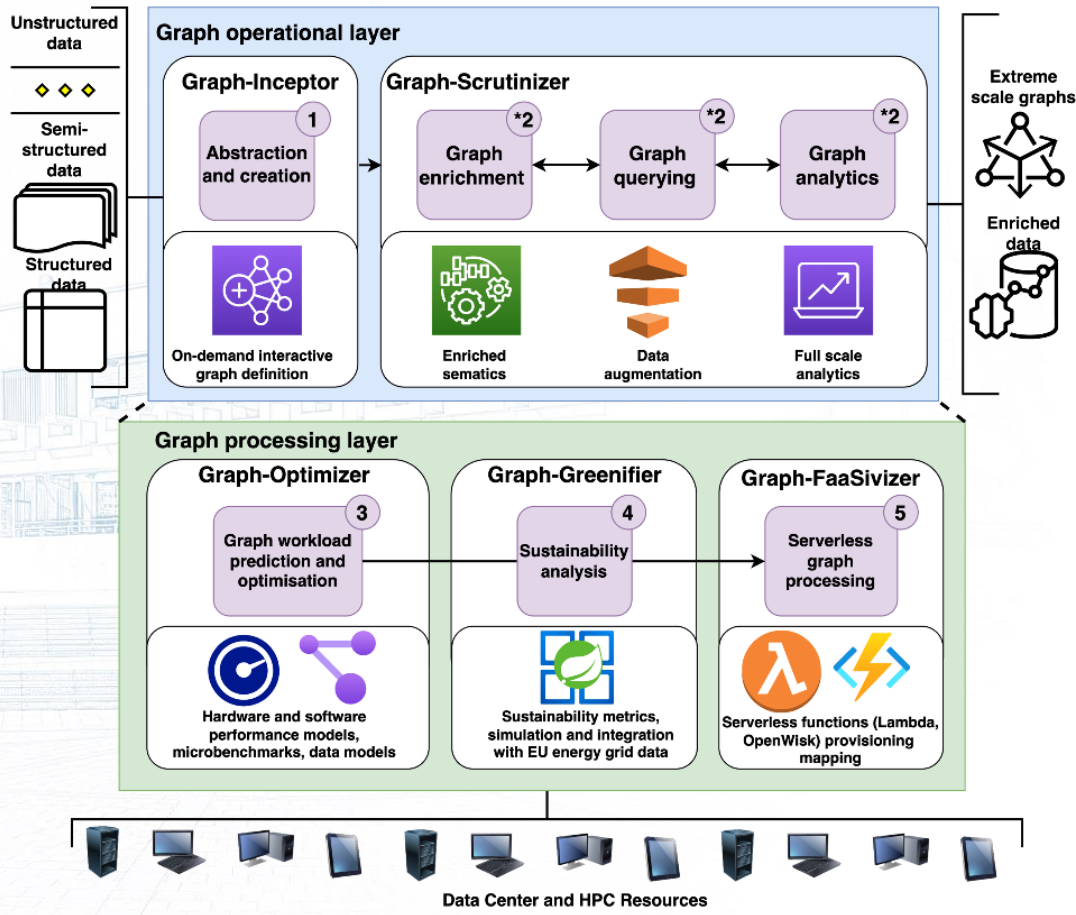
- **Extreme data challenge**, extending graph processing technology by orders of magnitude for several **"V"**-characteristics
  - **Volume** graph challenge by supporting up to billions of vertices and trillions of edges
  - **Velocity** graph challenge of dynamically changing topologies
  - **Viridescence**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



Graph-Massivizer Toolkit

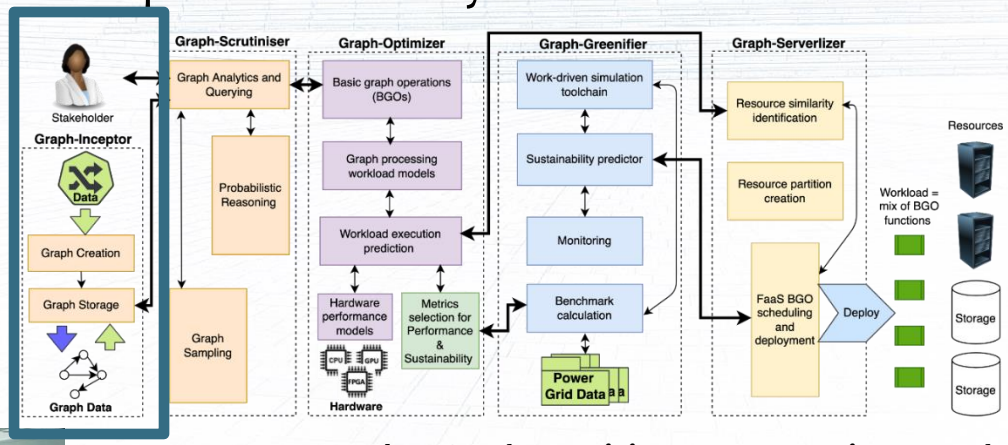
- Graph operational layer
  - **Graph-Inceptor**: extreme data ingestion, MG creation and storage
  - **Graph-Scrutinizer**: MG analytic and reasoning

- 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



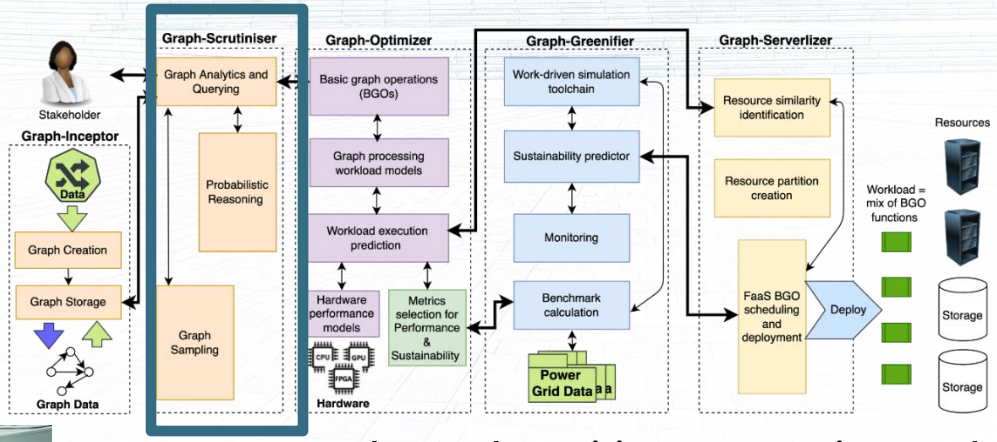
Graph-Inceptor

- 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

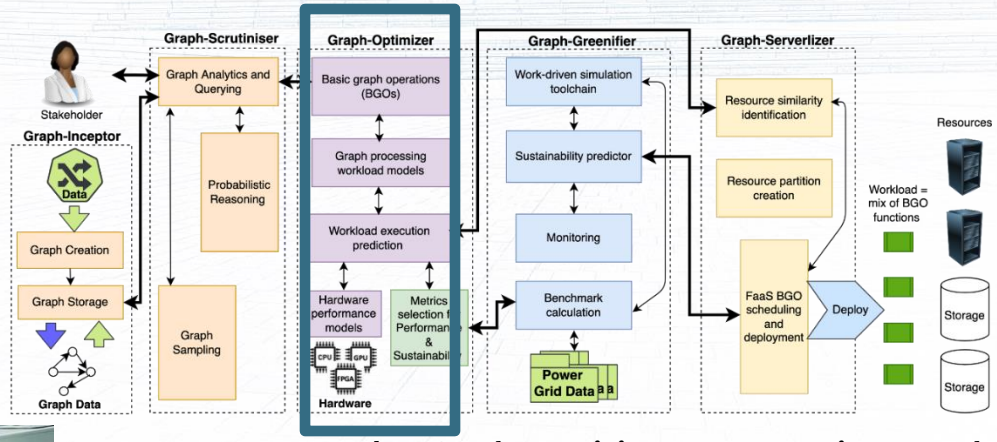


Graph-Optimizer

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

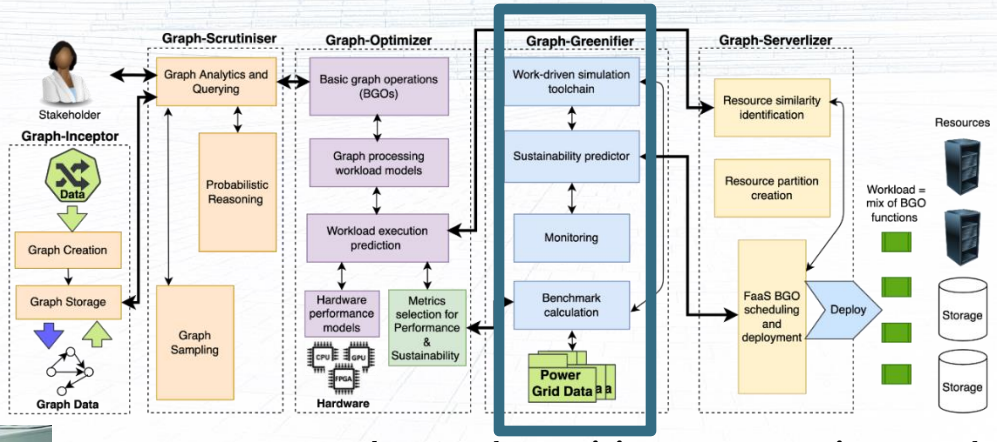


Graph-Greenifier

VU VRIJE UNIVERSITEIT AMSTERDAM

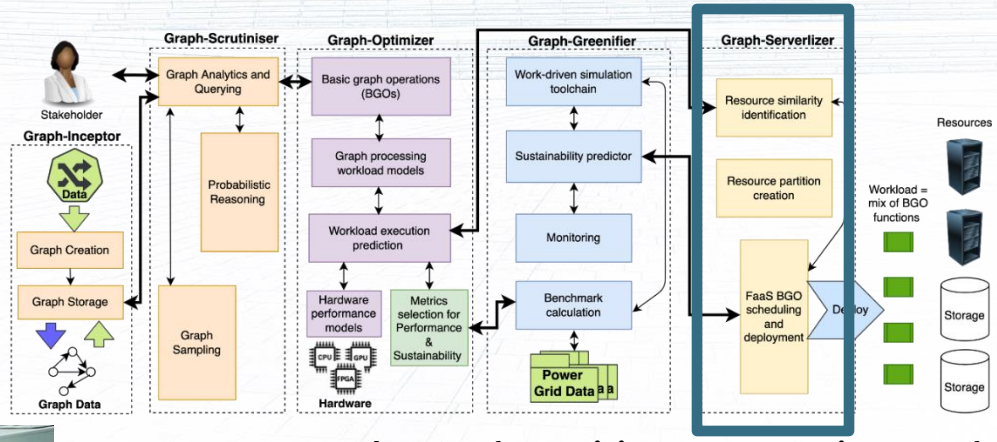
UNIVERSITÄT KLAGENFURT

- 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



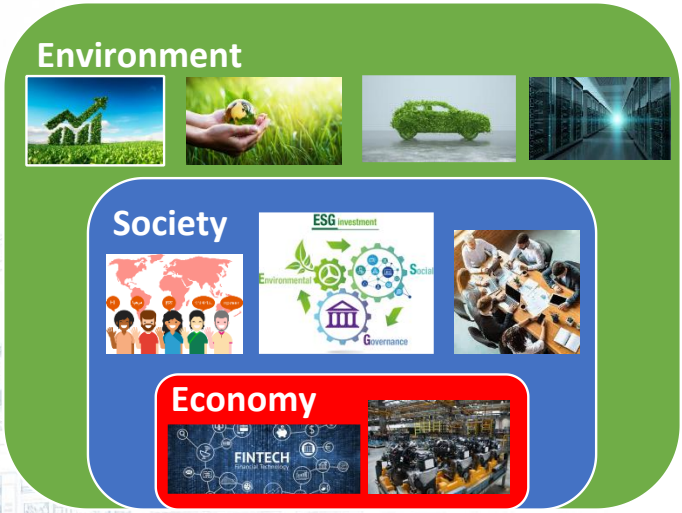
Graph-Serverlizer

- 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

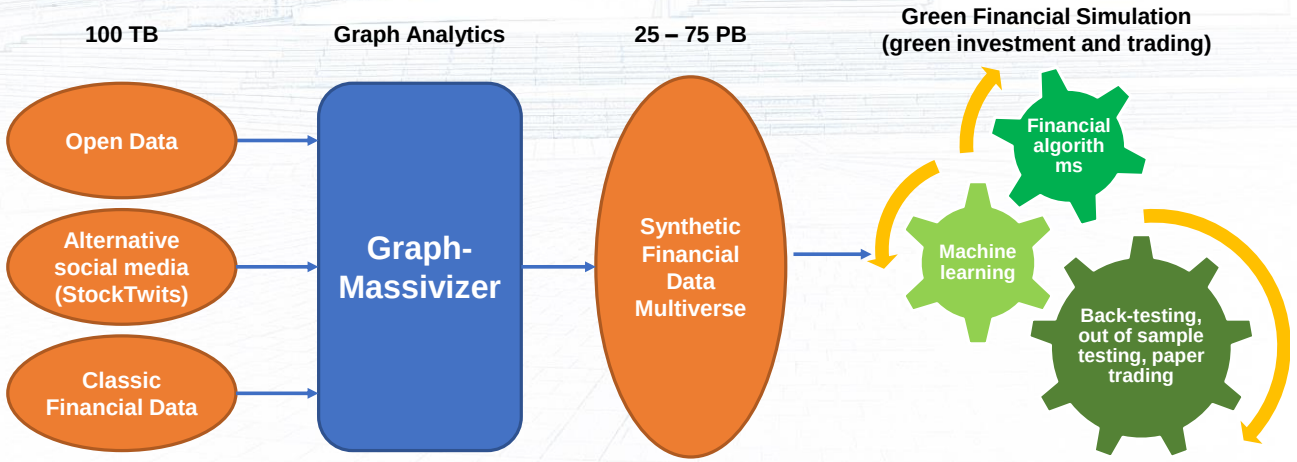
- UC-1: Green and sustainable finance
- UC-2: Global foresight for environment protection
- UC-3: Green AI for sustainable automotive industry
- UC-4: Data center digital twin for sustainable exascale computing



	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

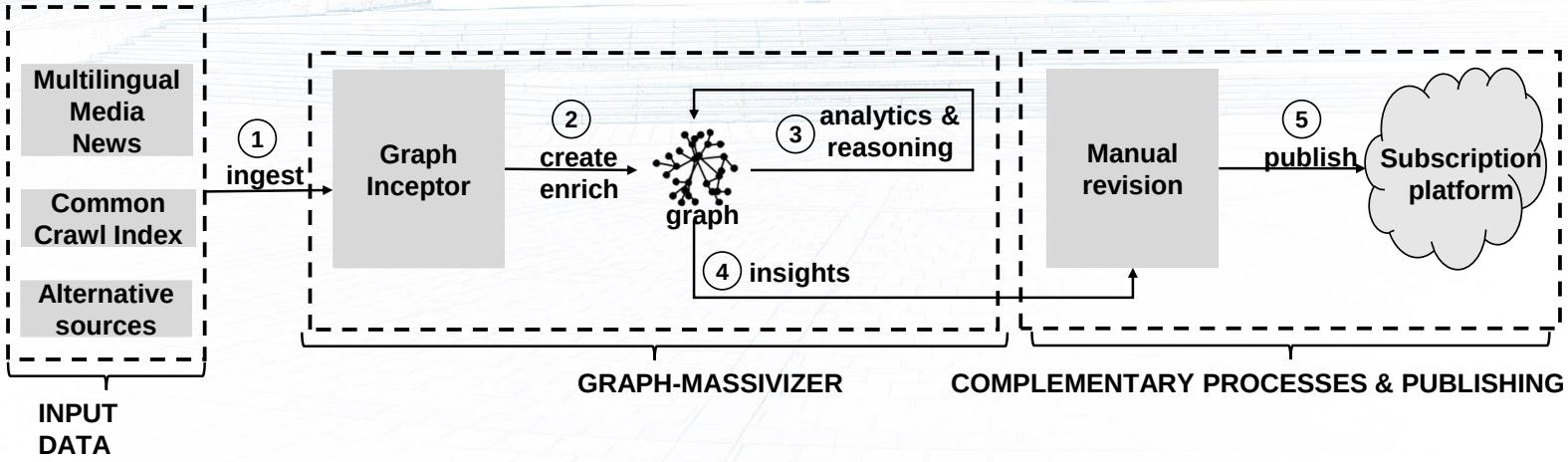


Global Foresight for Environment Prediction

EVENTREGISTRY

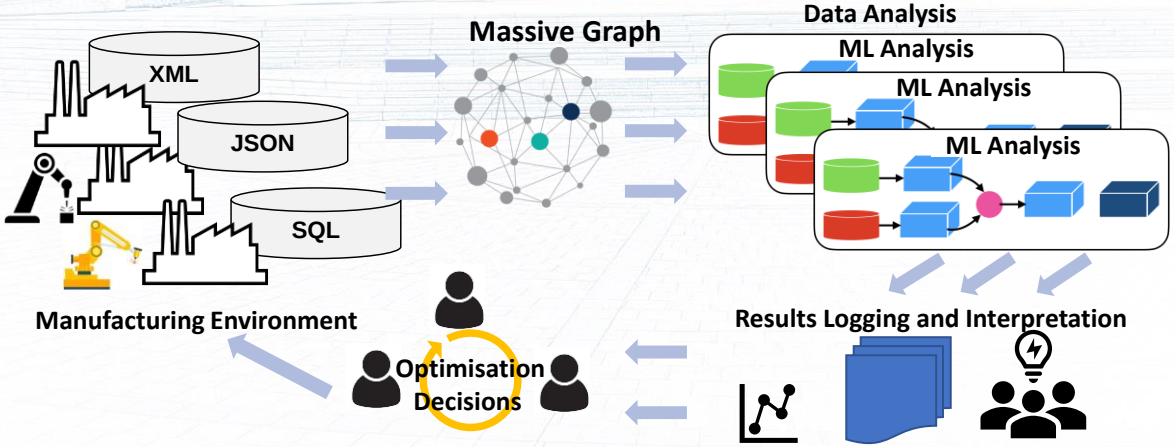
UNIVERSITÄT KLAGENFURT

- 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

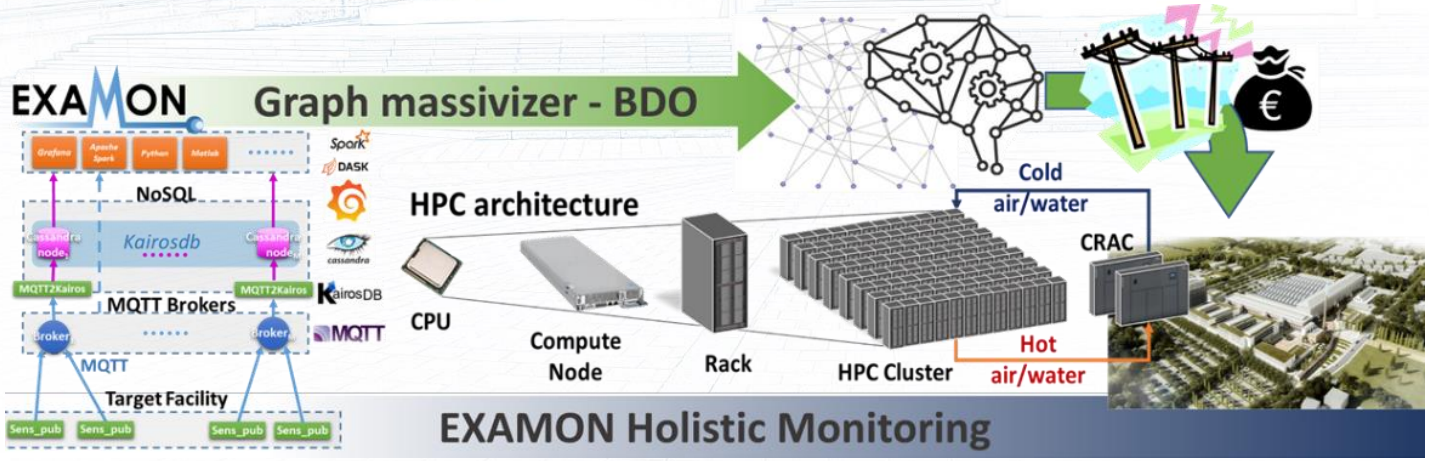


Data Center Digital Twin for Sustainable Computing

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



UNIVERSITÄT KLAGENFURT

IDC

Peracton

Jožef Stefan Institute

SINTEF

UNIVERSITY OF TWENTE.

metaphactory

VU VRIJE UNIVERSITEIT AMSTERDAM

CINECA

THANK YOU !

European Commission
Horizon Europe
2021-2027

EVENTREGISTRY

BOSCH