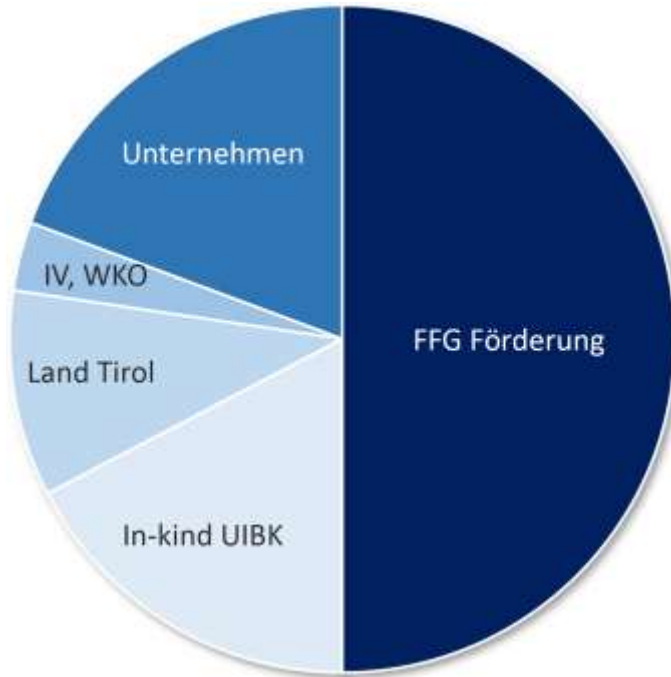


Prof. Radu Prodan

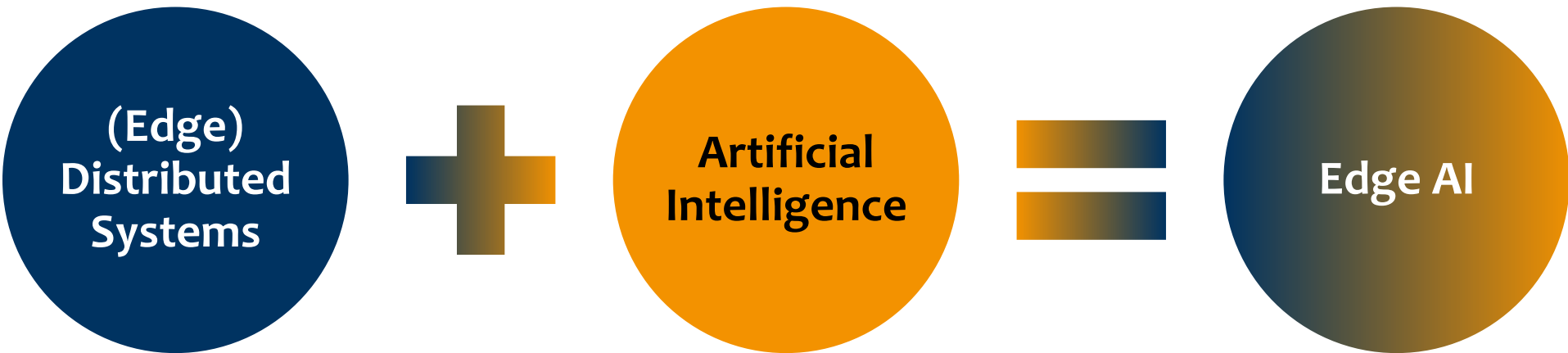
DISTRIBUTE AND LEARN 1 BILLION

Endowed Professorship Edge AI

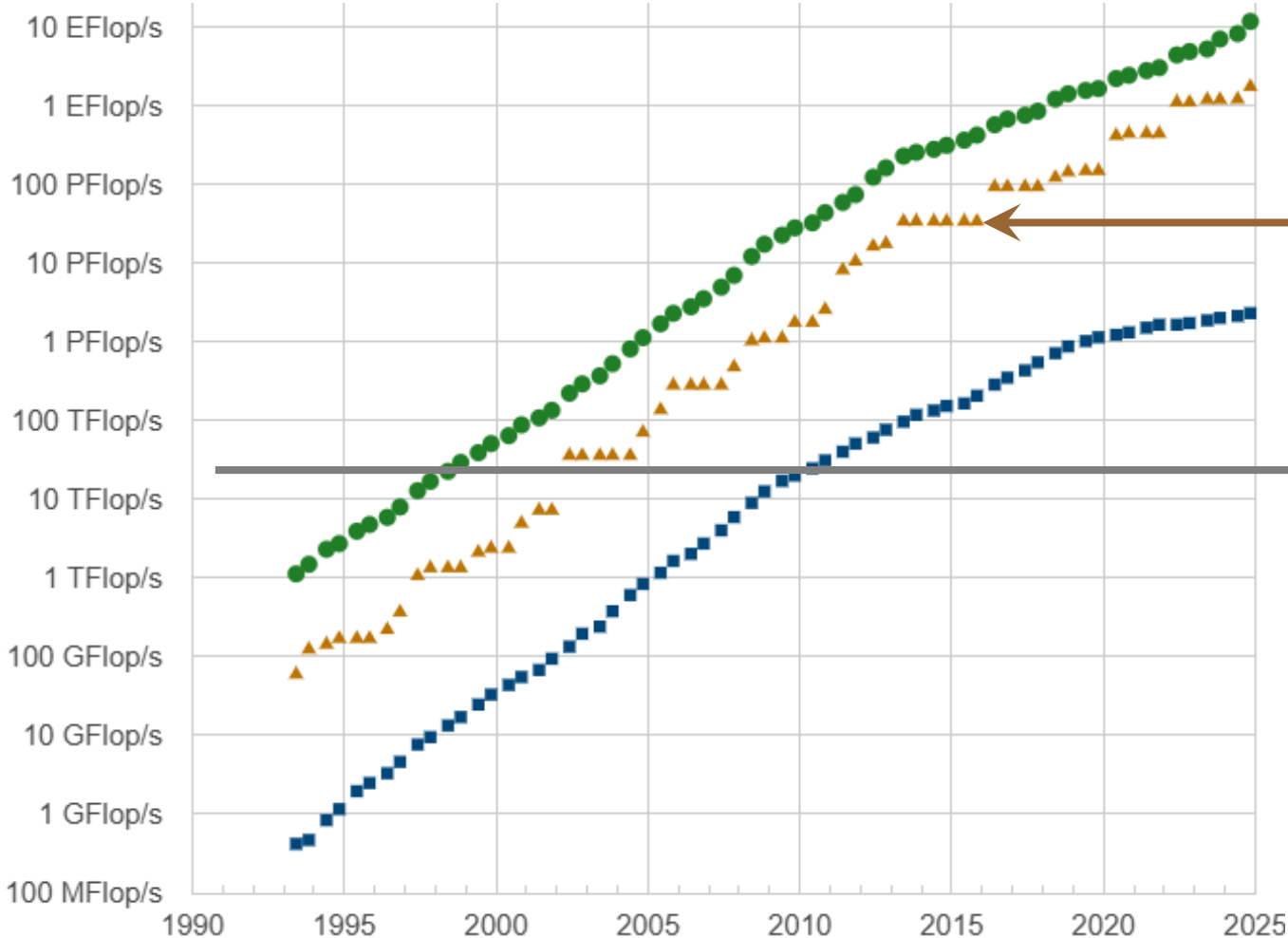
- AI Mission Austria (AIM AT)
- Department of Computer Science,
University of Innsbruck
 - 2025 – 20230



What is Edge AI ?



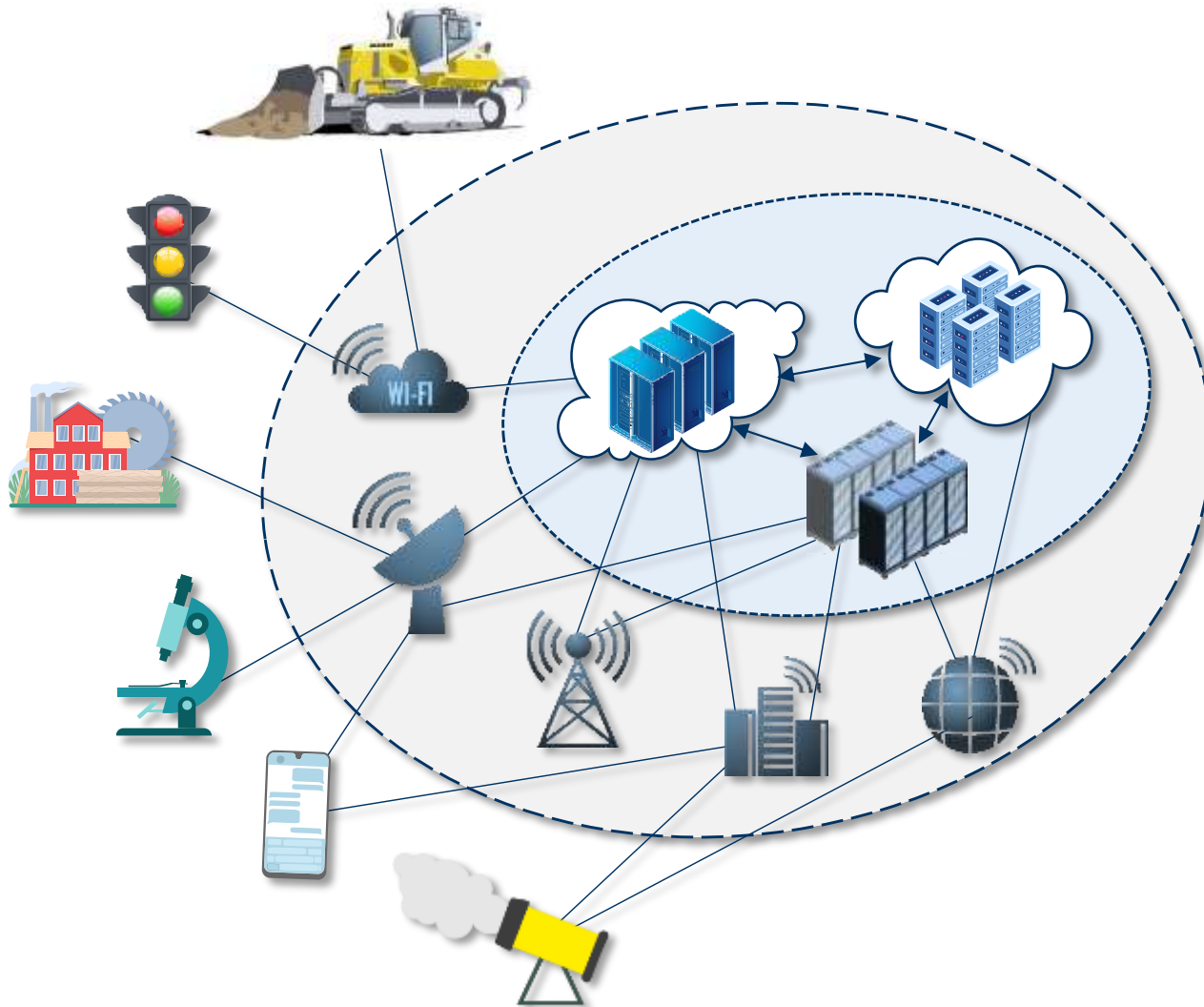
Edge Computing



Fastest
Supercomputer



Edge AI Applications



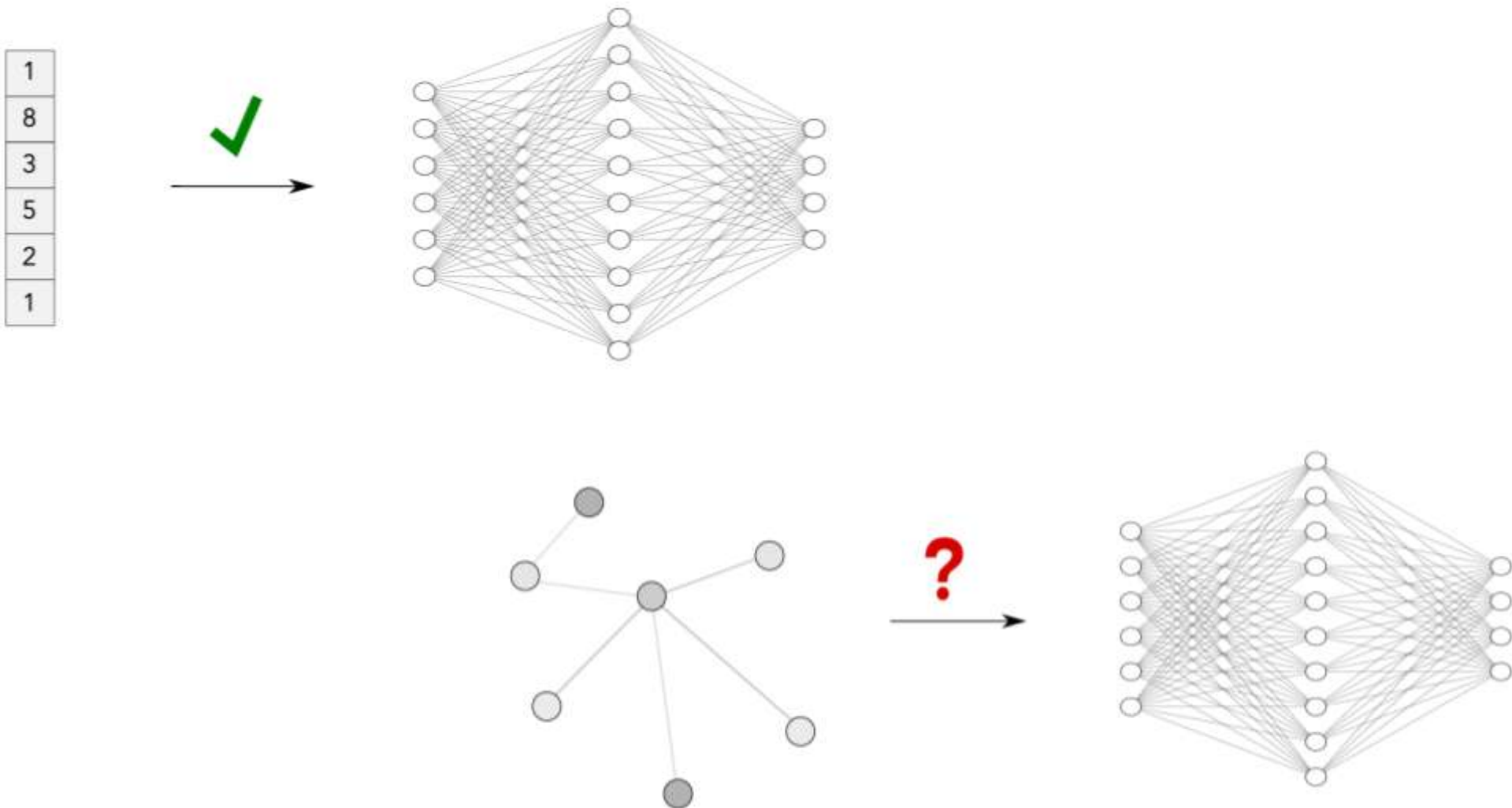
Graph-Massivizer



- Towards Extreme and Sustainable Graph Processing for Urgent Societal Challenges in Europe
 - € 4,998,062.50
 - 2023 – 2025
- HORIZON-CL4-2022-DATA-01-05
 - Extreme data mining, aggregation and analytics technologies and solutions (RIA)
- 12 partners
 - 4 universities, 2 research centers, 1 HPC center, 5 companies



Non-Euclidian AI



Real World Graph Structures

- Molecule graph
 - Thousand of nodes and edges
- Transportation networks
 - Tens of thousands of nodes and edges
- Social gaming network
 - Billion vertices
 - 100 billion connections
- Word Wide Web
 - 50 billion pages
 - 1 trillion hyperlinks
- Human brain
 - 100 billion neurons
 - 100 trillion connections



Image courtesy of shutterstock.com



Image courtesy of shutterstock.com

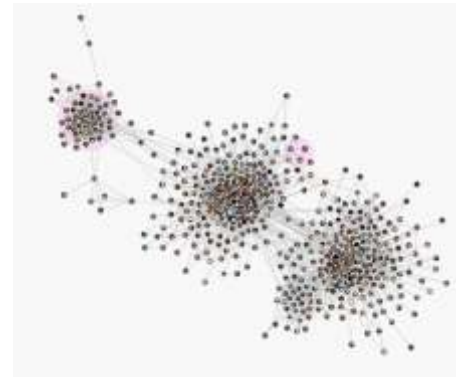
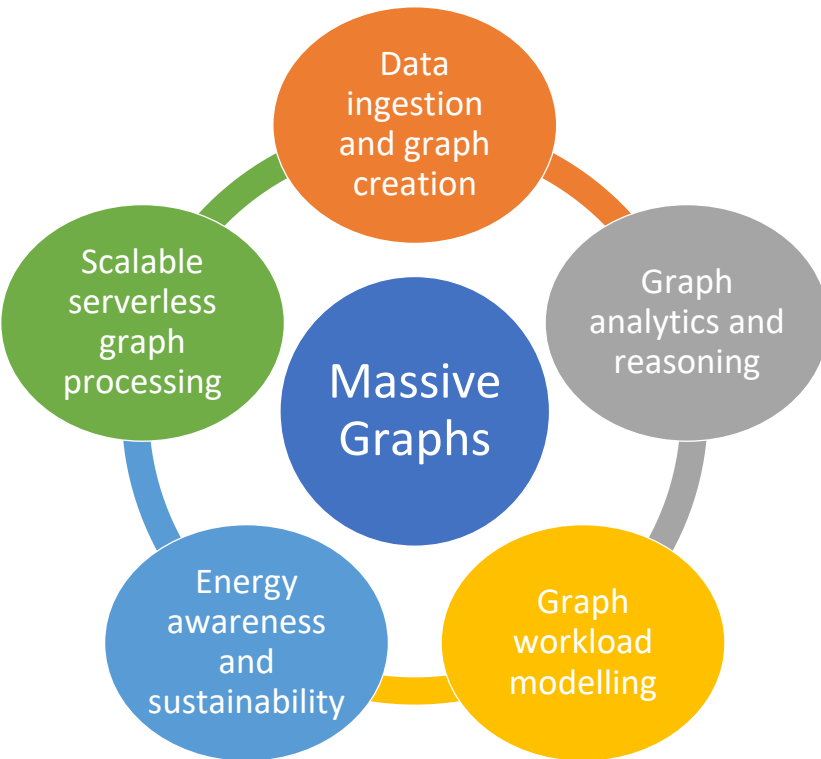


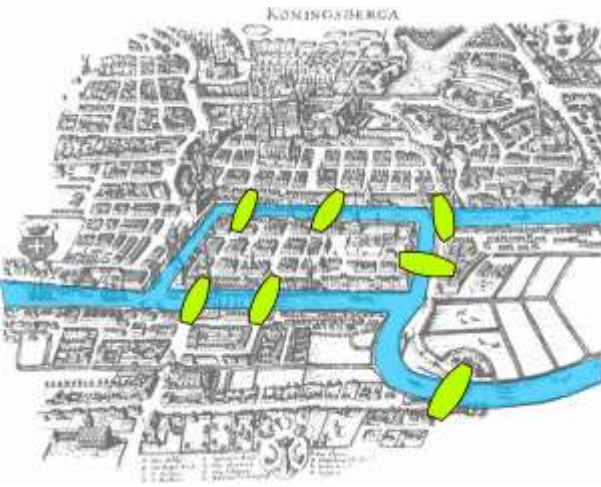
Image courtesy of shutterstock.com

Holistic Graph Processing

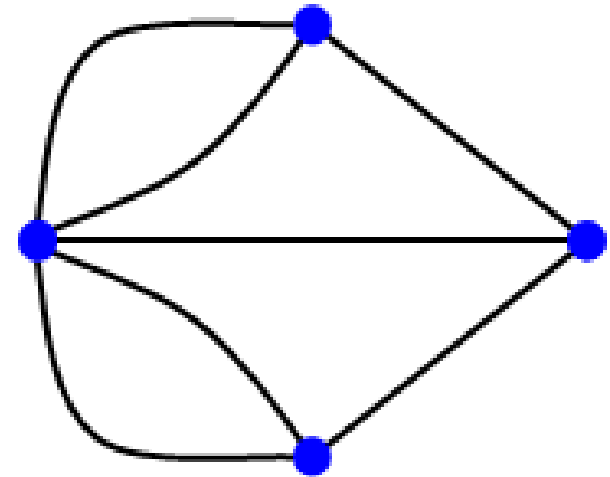
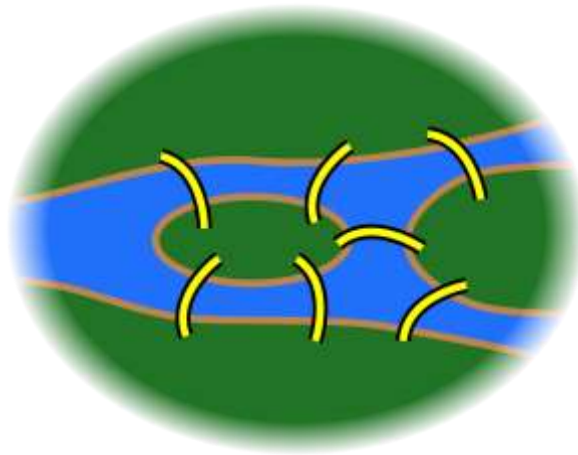


	Volume	Velocity	Value	Veracity	Variety	Viscosity	Viridescence
Finance	✓		✓	✓	✓		✓
Event forecast	✓				✓	✓	✓
Manufacturing		✓		✓	✓		✓
Supercomputing	✓	✓			✓		✓

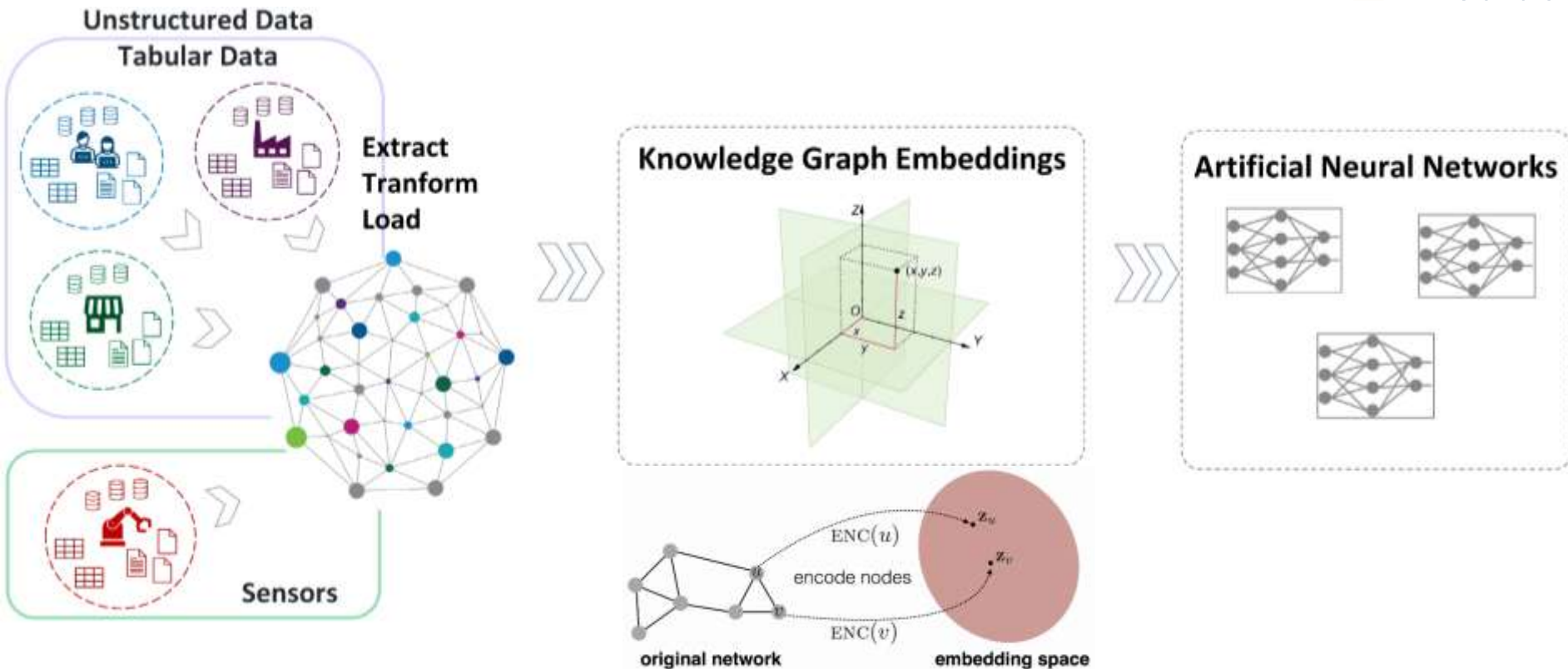
First Edge Graph AI Problem



https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg



Neuro-Symbolic Graph Operations

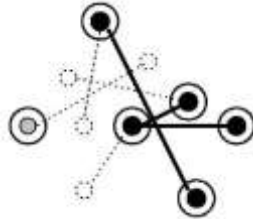


- Basic graph operations
 - Creation, load, query, traversal, node/graph properties, sampling, summarization, link prediction, ...

Graph Sampling Algorithms

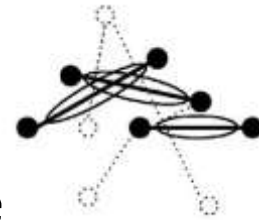
■ Node sampling

- Random node
- Random degree node



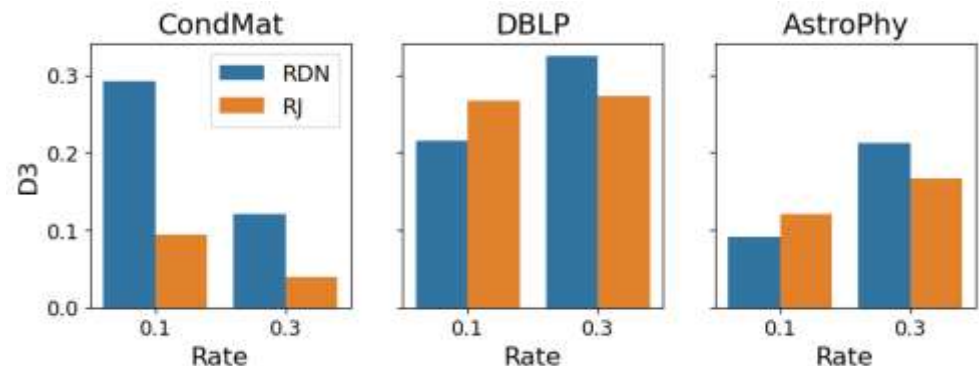
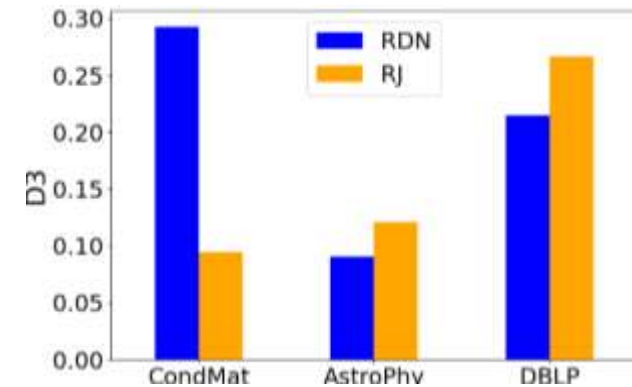
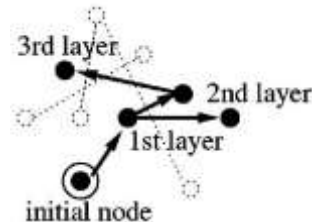
■ Edge sampling

- Random edge
- Random node-edge
- Induced random edge

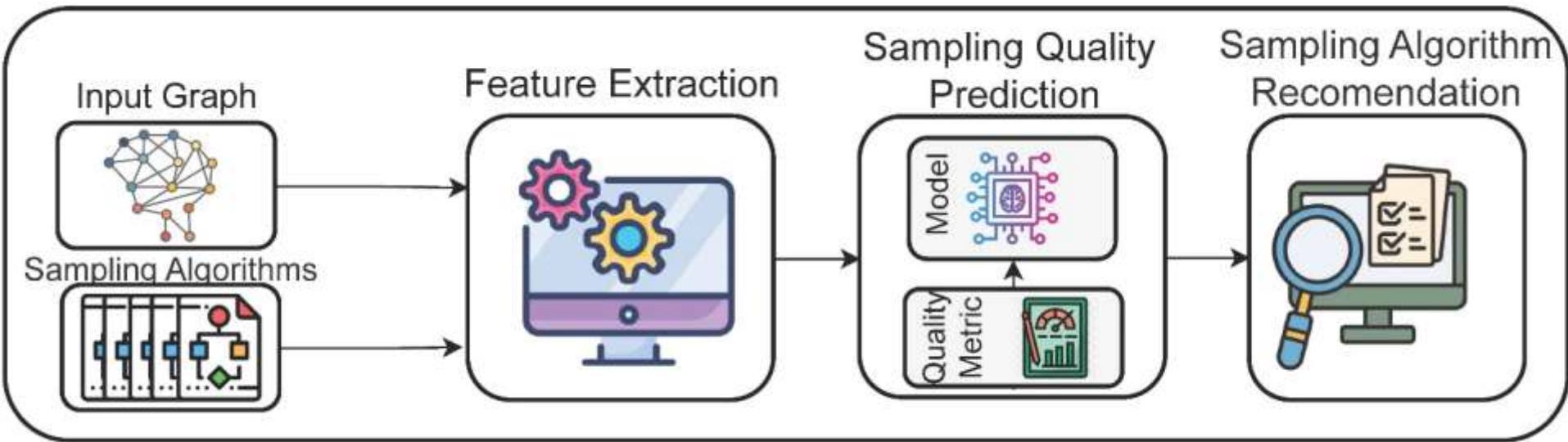


■ Traversal sampling

- Random walk
- Random jump
- Snowball
- Forest fire
- Frontier sampling
- Rank degree



Graph Sampling Prediction

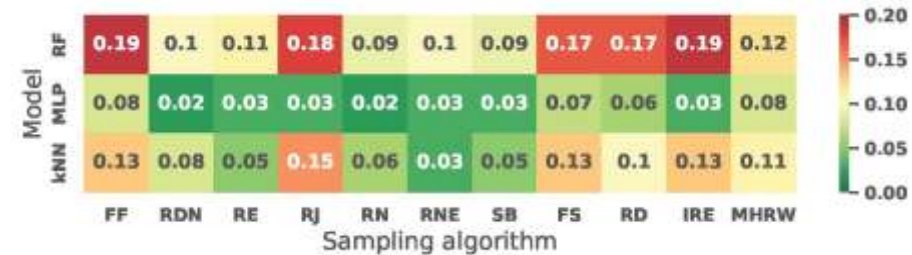


- **Sampling quality metrics**
 - Degree distribution
 - Clustering coefficient distribution
 - Hop-plot distribution
 - Algorithm execution time
 - ...

Sampling Divergence Accuracy

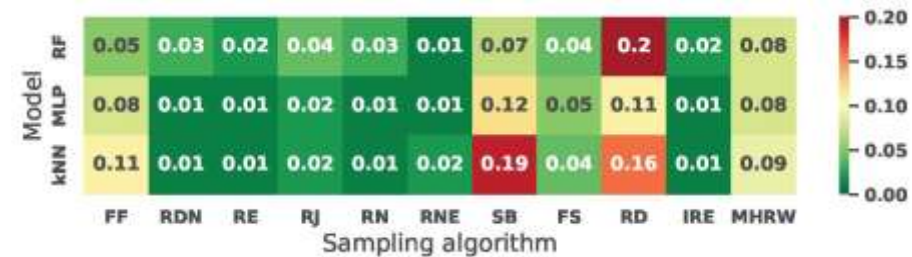
■ Degree distribution divergence (D3)

- MLP: $\overline{RMSE} \approx 4\%$



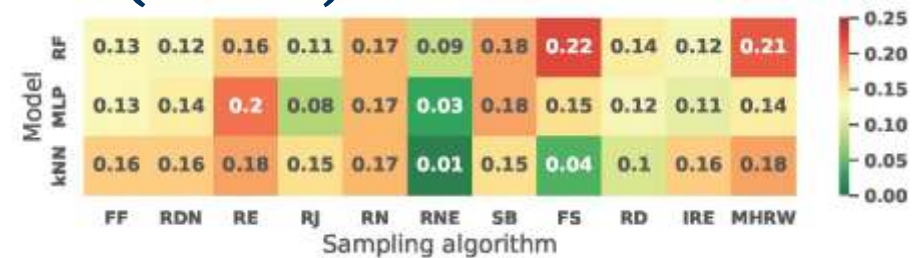
■ Clustering coefficient distribution divergence (C2D2)

- MLP: $\overline{RMSE} \approx 5\%$



■ Hop-plot distribution divergence (HPD2)

- kNN: $\overline{RMSE} \approx 13\%$



Sampling Recommendation Map

	Graph	Degree						Clustering Coefficient					
		Alg.	ML	QoS	Err.	Hits@1	Err.	Alg.	ML	QoS	Err.	Hits@1	Err.
Topology	Scale-free	FF	kNN	0.43	0.13	94	0.12	RN	RF	0.05	0.04	91	0.04
	Small world	FF	kNN	0.62	0.16	70	0.17	RN	RF	0.01	0.01	60	0.00
	Random	FF	kNN	0.51	0.15	85	0.14	RN	kNN	0.01	0.05	30	0.00
Category	Authorship	RJ	MLP	0.29	0.16	41	0.21	RN	RF	0.06	0.05	100	0.05
	Citation	FF	kNN	0.30	0.22	25	0.18	RN	RF	0.06	0.04	100	0.04
	Technology	FF	kNN	0.45	0.23	0	0.18	RN	RF	0.06	0.05	100	0.05

Graph Neural Networks

■ Input

- Node feature vectors: $X[N, D]$
- **Adjacency matrix**: $A[N, N]$

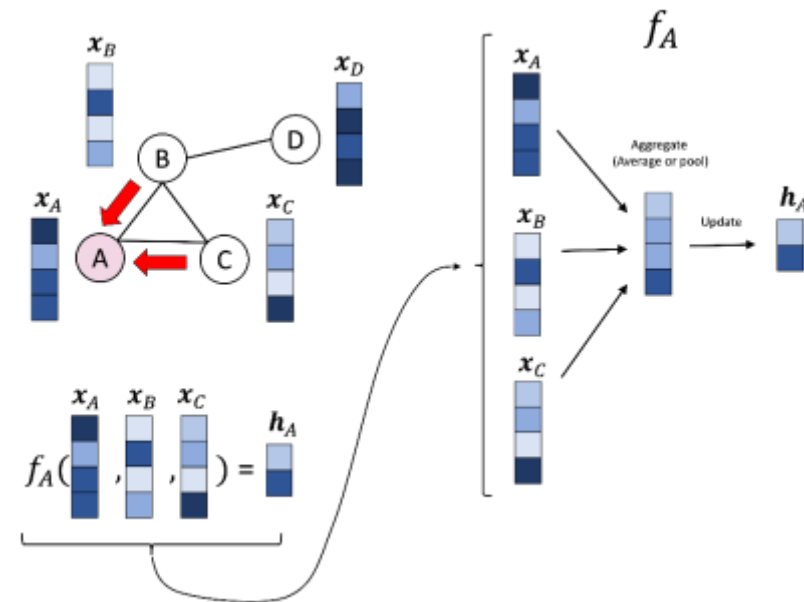
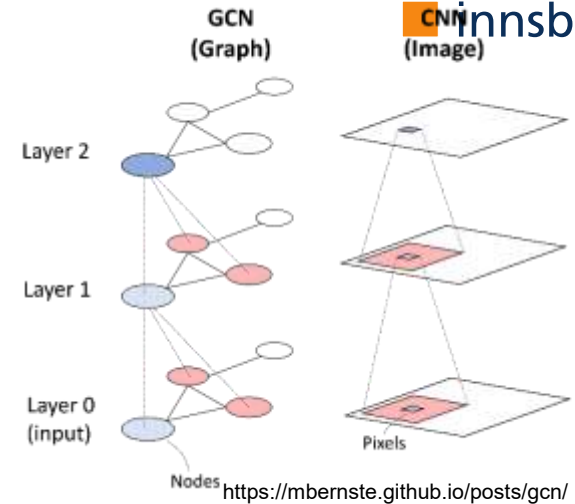
■ Graph convolutional layers

- **Aggregation** of neighboring node feature vectors:

$$\overline{A[i]} \cdot X$$

- **Update** of node feature vectors:

$$f_w(A, X) = \sigma(\overline{A[i]} \cdot X \cdot W)$$



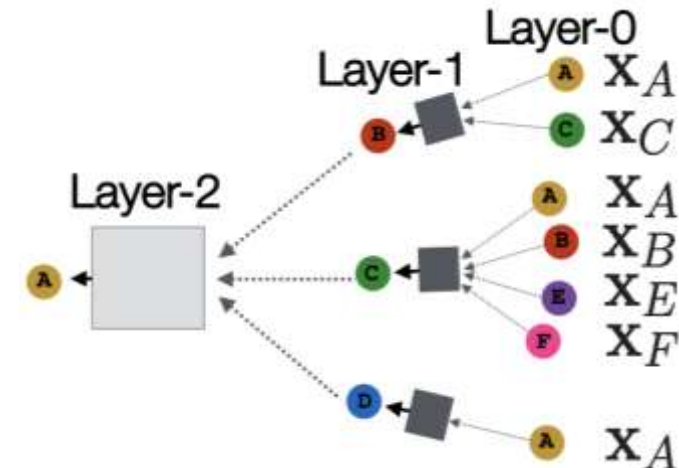
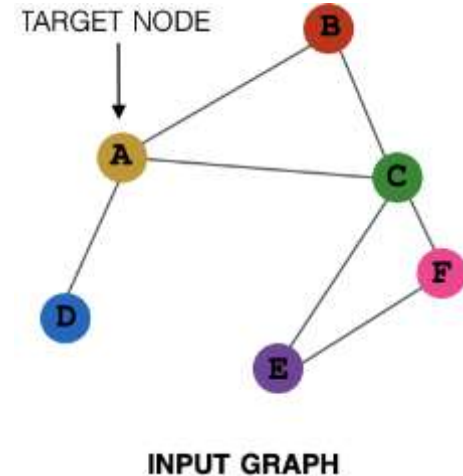
GNN Message Passing Architecture

- n stacked message passing convolutions

- $\mathbf{Y}_1 = f_{W_1}(X, \bar{A})$
- $\mathbf{Y}_2 = f_{W_2}(\mathbf{Y}_1, \bar{A})$
- ...
- $\mathbf{Y}_{n-1} = f_{W_{n-1}}(\mathbf{Y}_{n-2}, \bar{A})$
- $\mathbf{Y}_n = f_{W_n}(\mathbf{Y}_{n-1}, \bar{A})$
- $J = -(\hat{Y} \cdot \log(\mathbf{Y}_n) + (1 - \hat{Y}) \cdot \log(1 - \mathbf{Y}_n))$

- Backward propagation: $i = n \rightarrow 1$

- $\nabla \vec{J} = \left(\frac{\delta J}{\delta W_n} \dots \frac{\delta J}{\delta W_1} \right), \quad \frac{\partial J}{\partial W_i} = \frac{\delta J}{\delta Y_n} \cdot \dots \cdot \frac{\delta Y_i}{\delta W_i}$
- $\vec{W} = \vec{W} - \alpha \cdot \nabla \vec{J},$

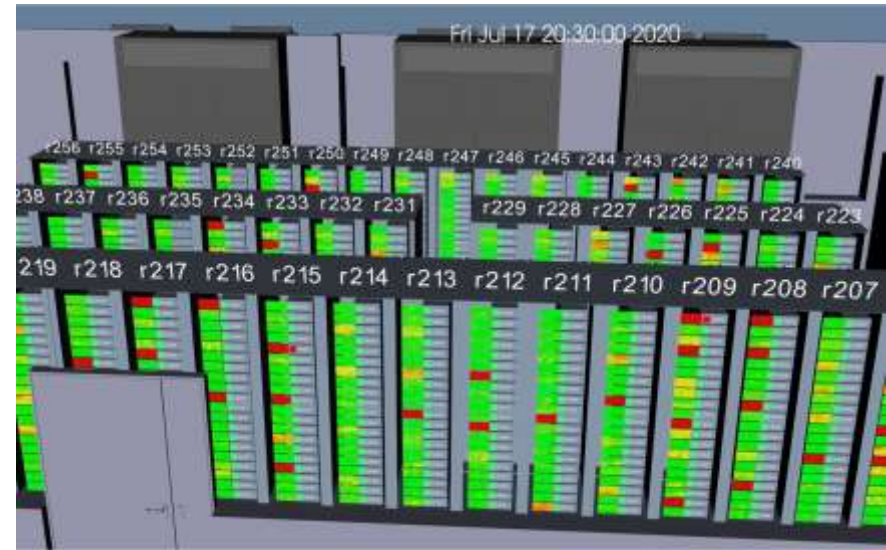


Anomaly Prediction in AI Factories

- Significant deviation from normal operation
 - Hardware: component failure (crash, degradation, overheat)
 - Software: drained node (overload, memory leak, configuration drift, system update)
- Reduce IT downtimes and maintenance
 - Avoid unplanned downtime
 - \$5,600 per minute (Gartner, 2022)
 - Improved planning, utilization, operation, and sustainability
- State-of-the-art in anomaly detection
 - Individual node-level telemetry metrics
 - No spatial interconnection (network topology)

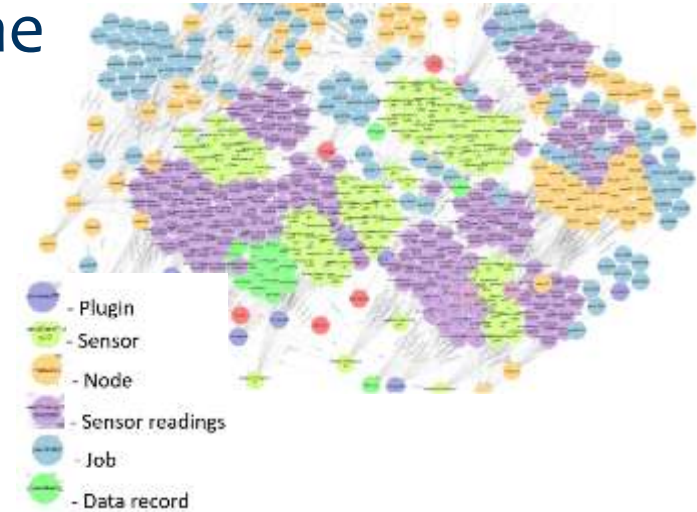
Marconi 100 Supercomputer

- Architecture
 - 49 racks, 980 computing nodes
 - 2x16 cores IBM POWER9 AC922
 - 4 Nvidia Volta V100
- ExaMon monitoring system
 - Slurm, Nagios, Ganglia, IPMI, ...
 - 573 metrics (Avg, Max, Min, SD)
 - 31-month operation (15-minute interval)
 - KairoDB, Cassandra NoSQL database
 - ExaData: 50 TB open data (Parquet, Zenodo)
- Operational data analytics (ODA) ontology
 - Rack, node, device, position, job, sensor, metrics, ...



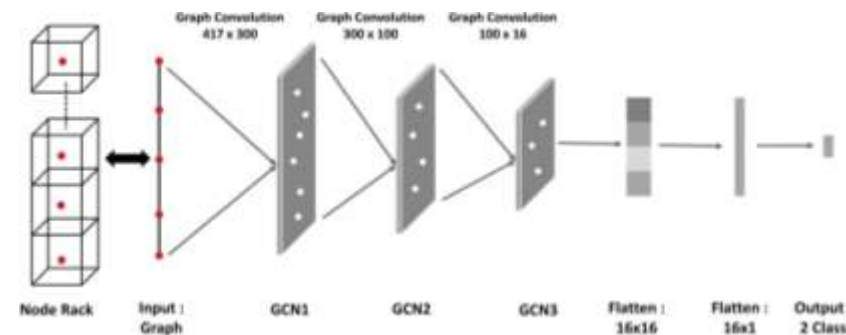
GNN-based Anomaly Prediction

- ETL monitoring data ingestion pipeline
 - Mapping time series to knowledge graph
 - Network architecture topology
 - GraphDB RDF database
 - 11.6 billion nodes, 68.6 billion edges



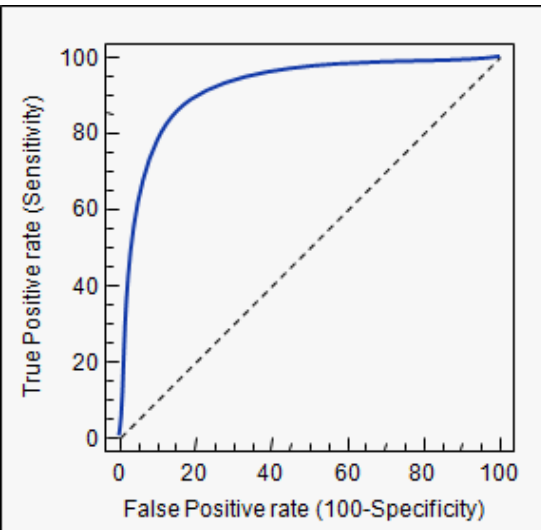
- GNN with 3 convolution layers
 - Room, rack, node
 - 573 semantic node features

- GNN-based node classification
 - Operational or anomalous
 - 25 months of training
 - 6 months of testing

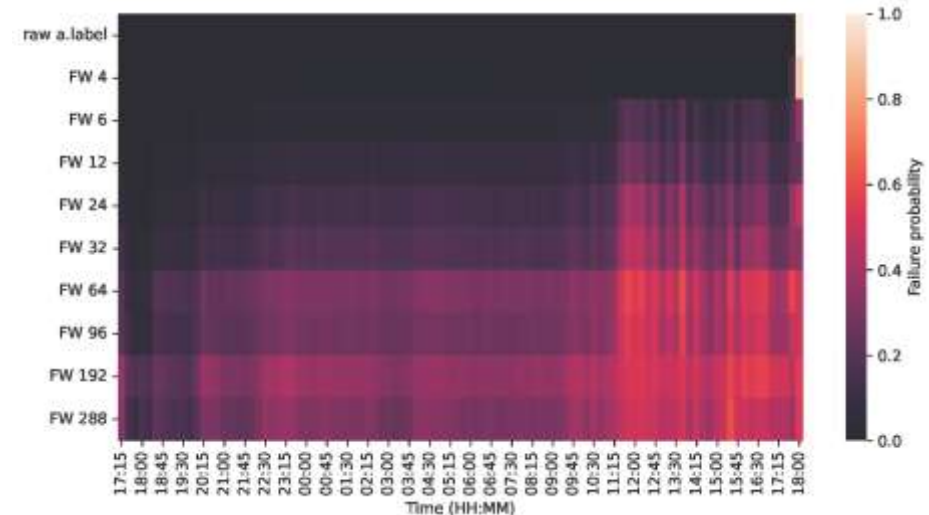


Anomaly Prediction Results

Area Under the Curve



Anomaly: Node 240, 7.05.24, 6:45



FW	rack GNN	room GNN	DNN	GB	RF	DT
4	0,91	0,59	0,64	0,63	0,61	0,51
6	0,89	0,58	0,66	0,64	0,59	0,5
12	0,84	0,47	0,65	0,63	0,59	0,5
24	0,78	0,58	0,62	0,6	0,55	0,5
32	0,75	0,55	0,59	0,58	0,55	0,5
64	0,66	0,42	0,5	0,48	0,49	0,49
96	0,62	0,58	0,55	0,51	0,58	0,51
192	0,55	0,52	0,47	0,48	0,52	0,5
288	0,53	0,50	0,52	0,51	0,52	0,49

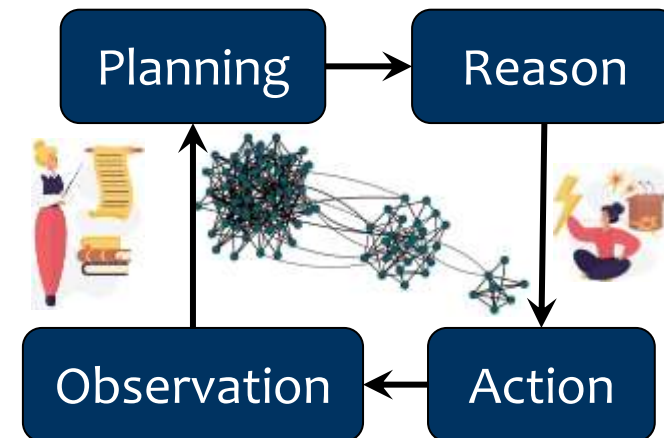
CoAgent

- Cognitive Computing Continuum for Large-Scale Distributed GenAI Agents
 - Cognitive Computing Continuum for Large-Scale Distributed GenAI Agents
 - 2026 – 2029
 - €8 million
- HORIZON-CL4-2025-04
 - Empowering AI/generative AI along the Cognitive Computing continuum (RIA)
- Partners
 - 17 partners, 9 countries
 - 5 universities, 1 research center, 1 public body, 10 companies



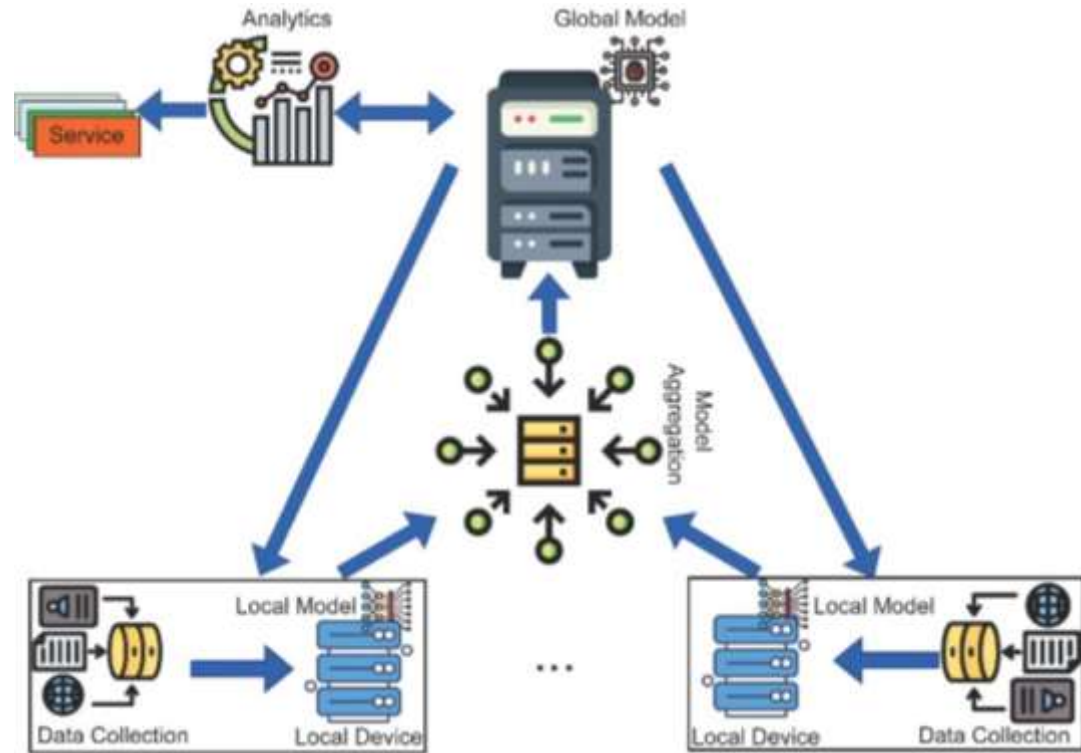
Agentic Edge AI

- Large foundational models **do not cut the edge**
 - Oversized, costly, unsustainable, unfocused, unreliable, unexplainable, centralized, unscalable, ...
 - **Monolithic, reactive, unstructured**
- **Proactive autonomous small AI agents**
 - Intent, fine-tuning, federated learning, knowledge distillation
 - Telecommunication, drone mission control, smart mobility, healthcare
- **Compiler 4.0 for neuro-symbolic agentic AI pipelines**
 - **ReAct loop** grounded in **structured data**
 - **Inductive**, probabilistic reasoning and acting
 - **Deductive**, deterministic observation and planning
 - Declarative **regulatory compliance guardrails**

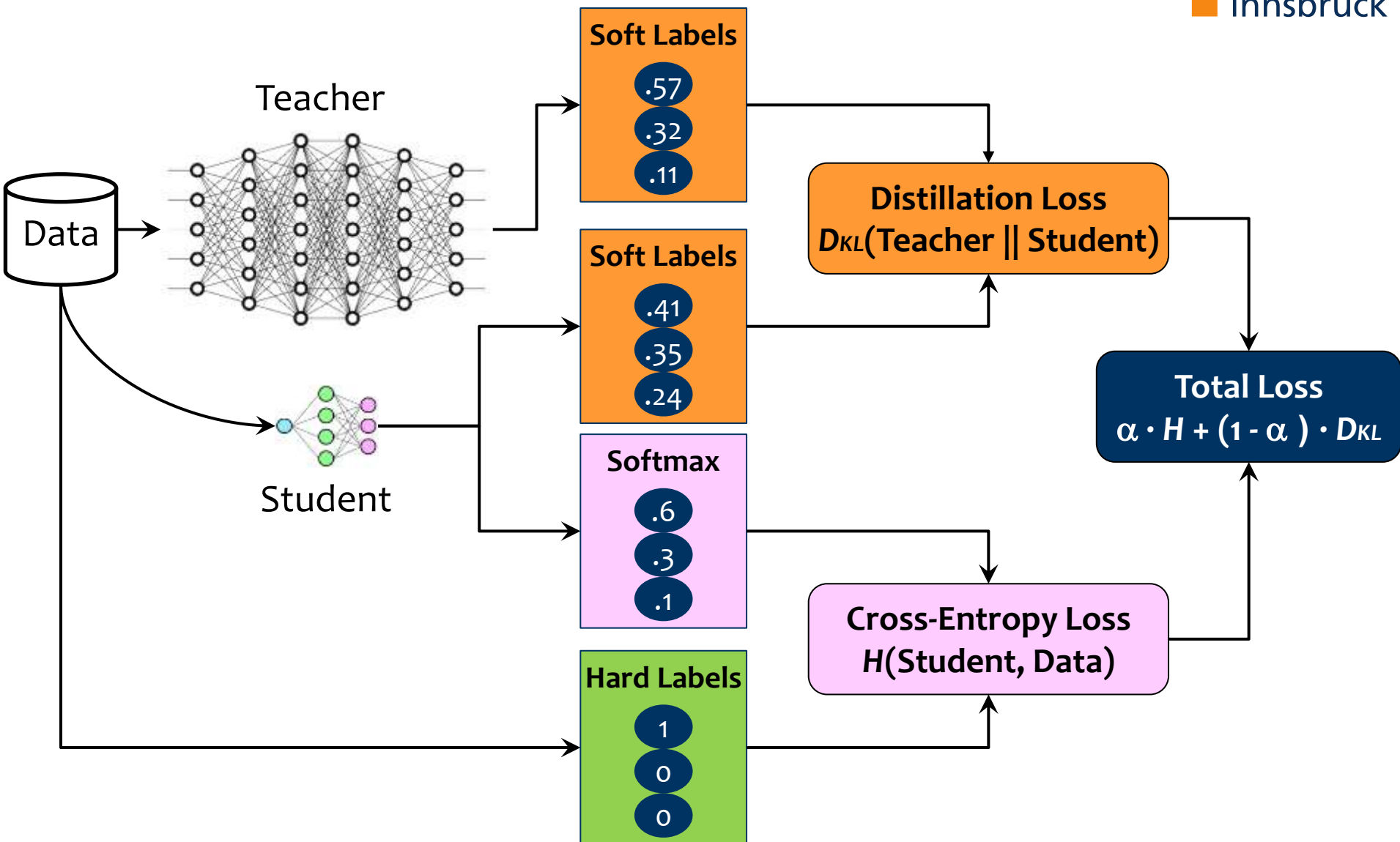


Federated Learning

- Share global knowledge, keep local data
 - Train AI models locally on private data
 - Share the model for global aggregation (e.g., federated averaging)
- Disadvantages
 - Homogeneous model
 - Large gradient communication
 - Model inversion
 - Membership inference
 - Non-IID data inaccuracies

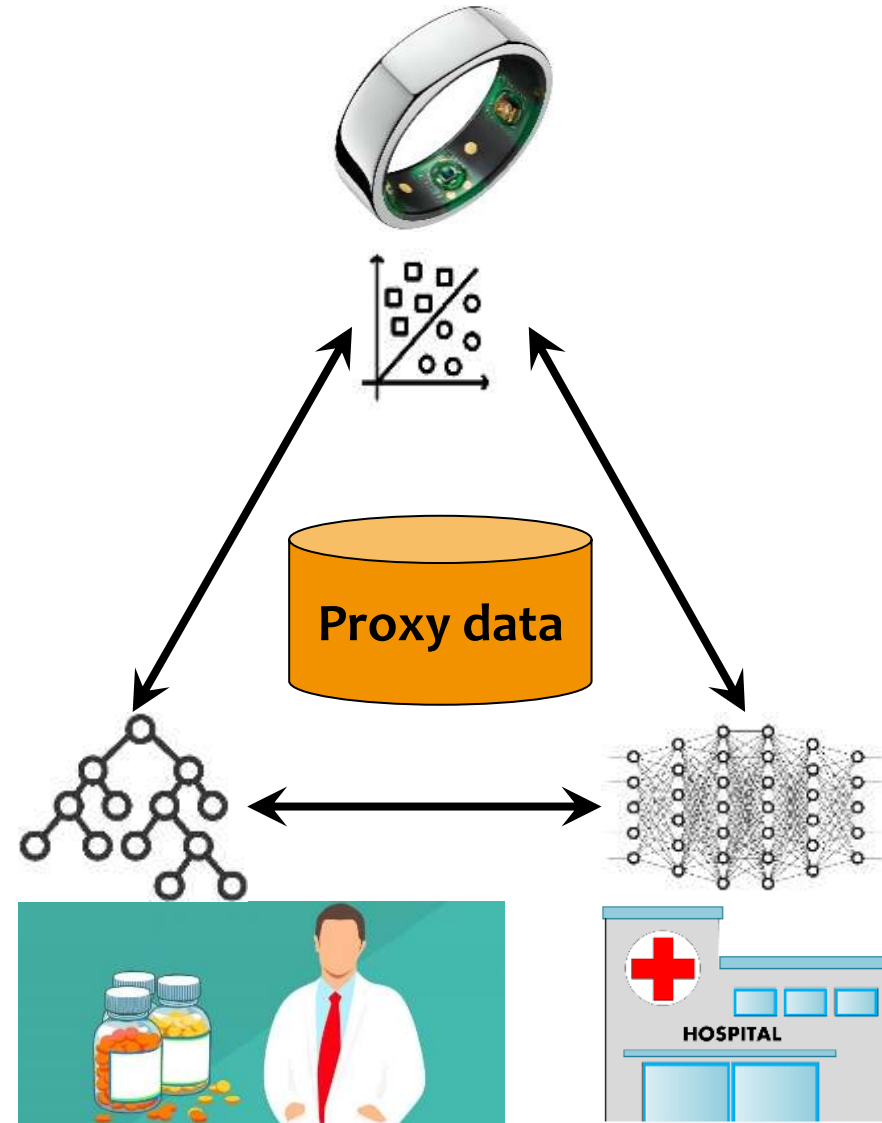


Knowledge Distillation

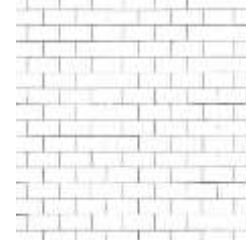


Federated Distillation

- Participants
 - Different **model architectures**
 - Different **private data**
- Federated learning
 - Knowledge exchange (no teacher)
 - Low logits communication overhead
- Challenges
 - Common proxy dataset for distillation (**synthetic data**)
 - Possibly different feature space
 - Large, non-IID-data
 - Slow, synchronous convergence on edge devices (e.g., stragglers)
- *3rd International Conference on Federated Learning Technologies and Applications (FLTA 2025)*



The Wall



- Technological walls hinder AI at scale
 - **Memory**: model size versus VRAM
 - **Power**: Gigawatt-scale AI factories
 - **Economic**: TSMC, Samsung, Intel
 - **Data**: GIGO, model collapse
- AI as a refinery for high-quality data
 - **Data laundromats** under regulatory sandboxes (corporate, deep web, dark data)
 - High-fidelity and statistically-accurate **synthetic data generation**
- *Scaling data like computing, Communications of ACM, To appear*



Thanks 1.000.000.000 !

