Software Center

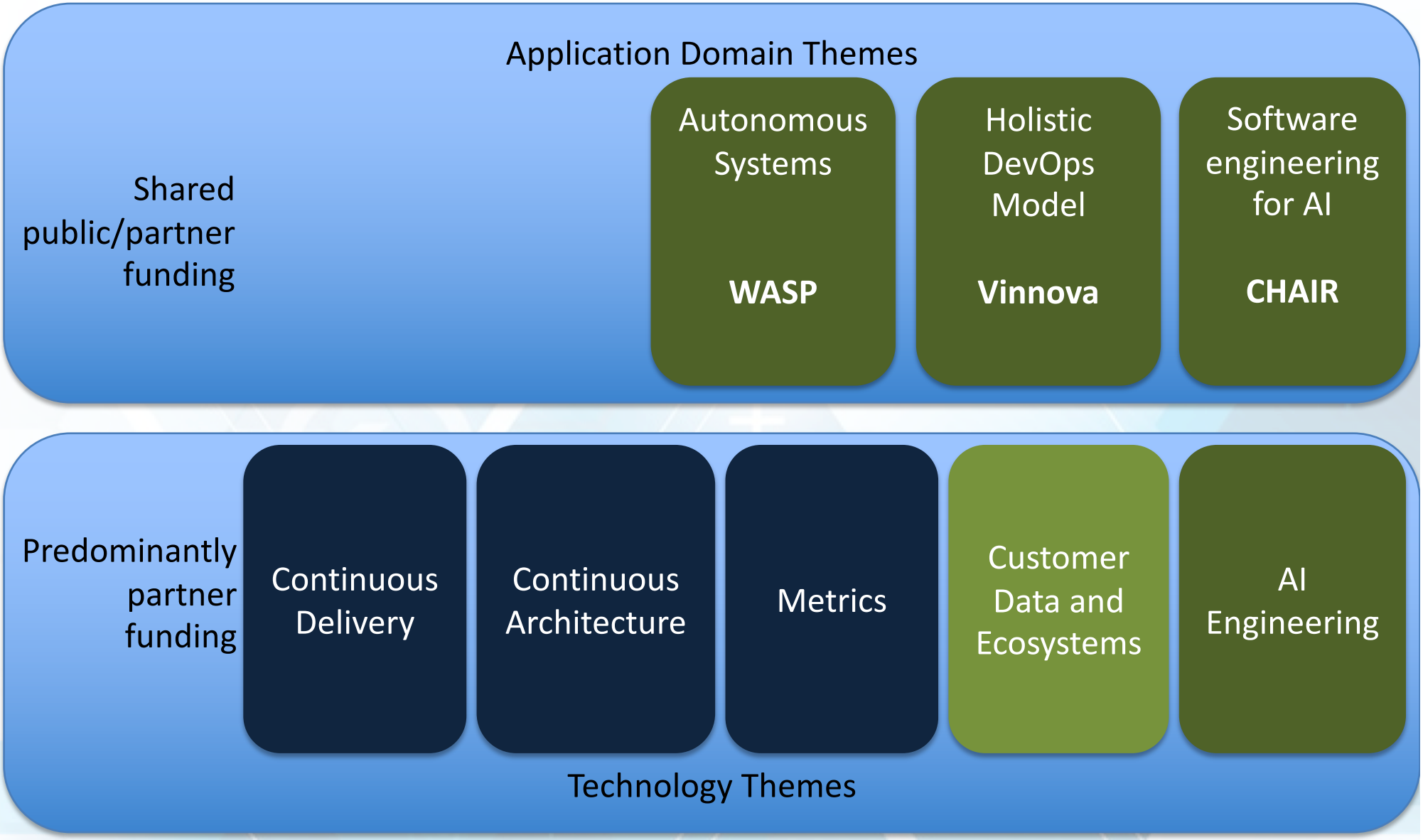
AI and data driven development

Overview of on-going research topics and activities

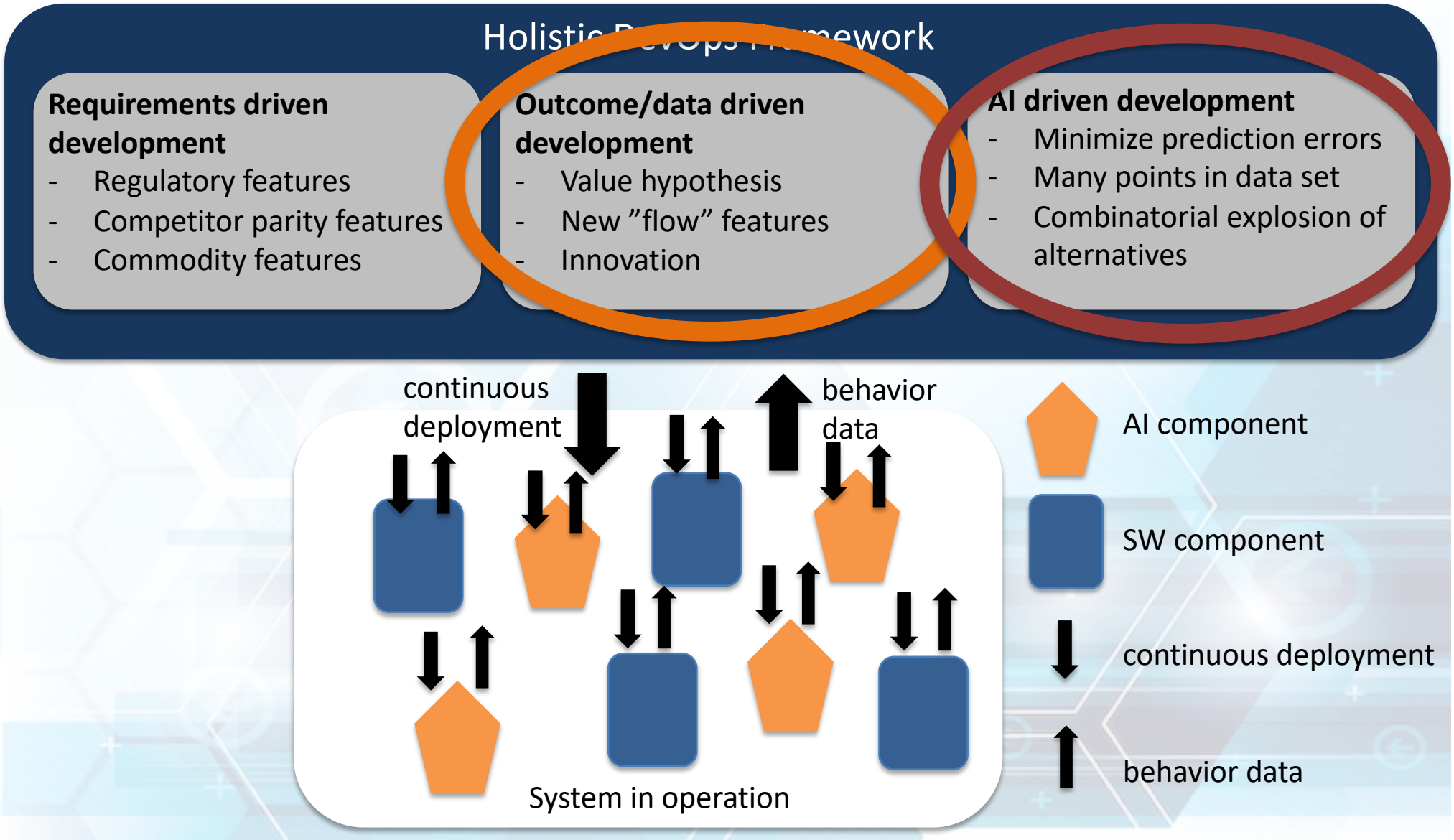
Helena H. Olsson

Software Center Reporting workshop, Gothenburg, December 5th, 2019

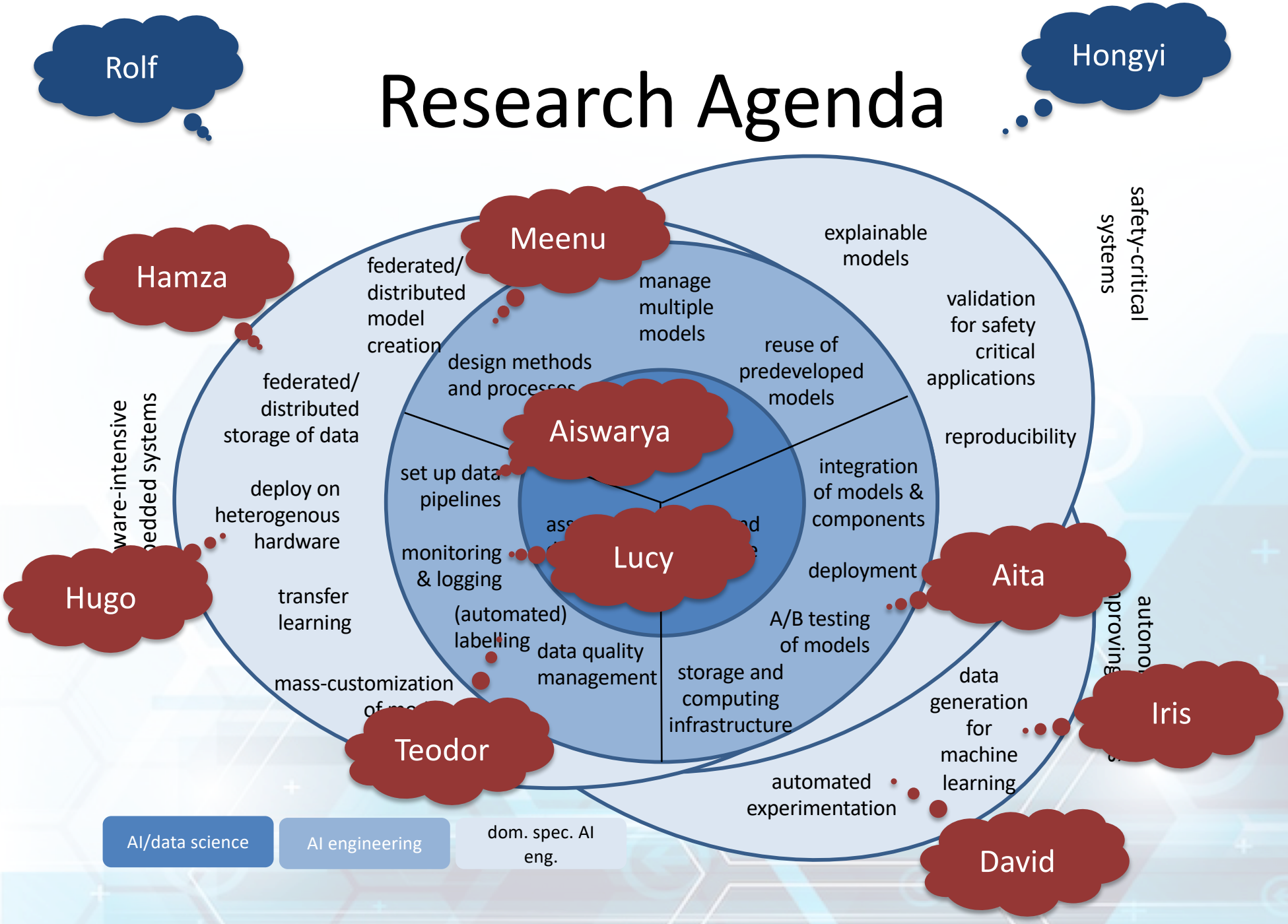
Research Themes



Holistic DevOps Framework



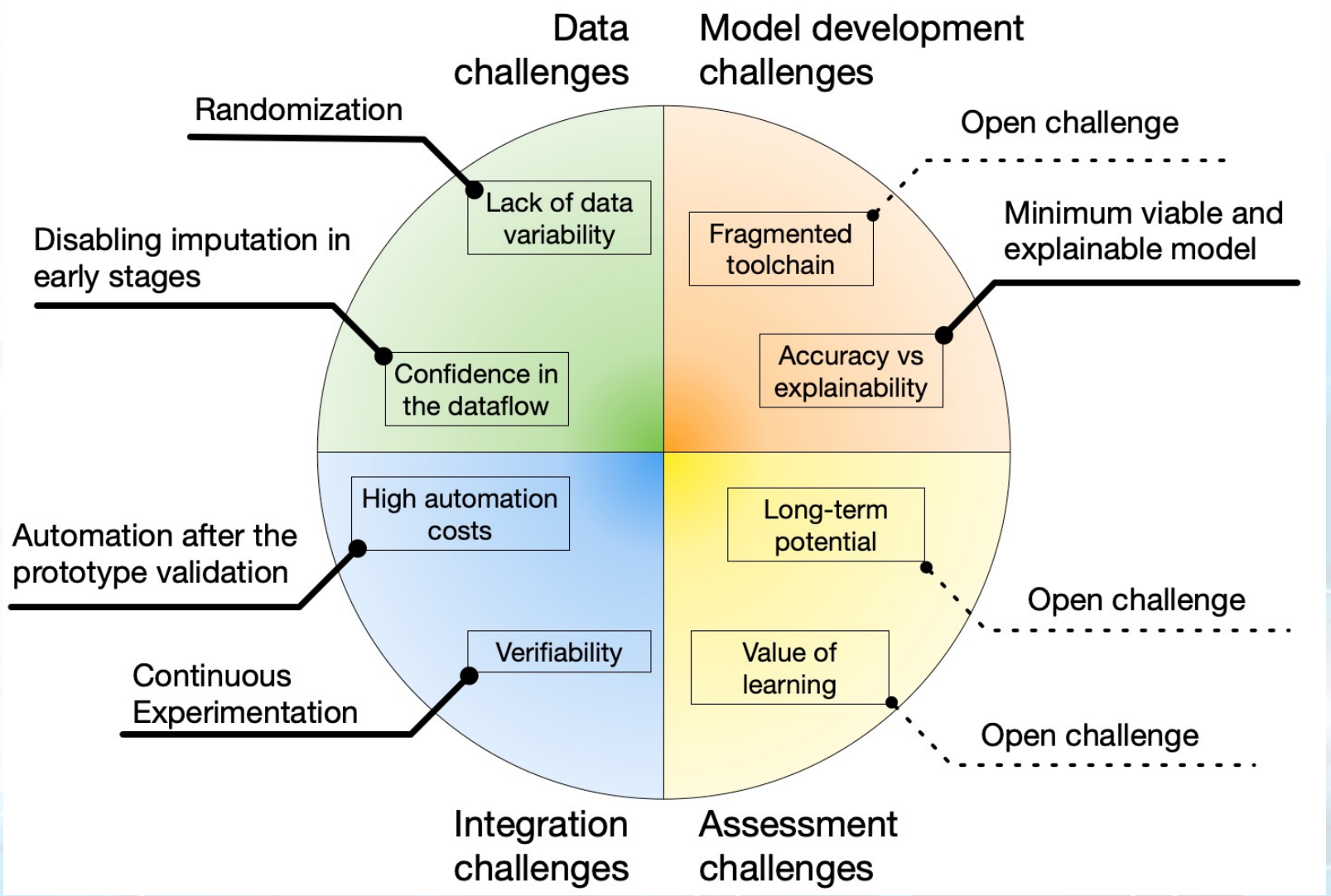
Research Agenda



Statistical comparison of black box algorithms for optimization of expensive functions

- Large scale comparison between algorithms
  - Over 25 algorithms
  - Over 170 benchmark functions
  - Different measurements metrics
  - Different conditions (noise, budget etc…)
- Applications
  - Field optimization
  - Hyperparameter tuning of ML models
  - AutoML

Business transformations through ML experiments



Data Management Challenges

MAPPING BETWEEN DATA MANAGEMENT CHALLENGES AND USE CASES

Phase	Challenge	Use cases of DL components					
		RS[1]	WPP[2]	HPP[3]	MD[4]	FFD[5]	MS[6]
Data Collection	Lack of metadata	X	X	X	X	X	X
	Data Granularity		X	X			
	Shortage of diverse samples		X	X	X	X	
	Need for sharing and tracking techniques	X	X	X	X	X	
	Data Storage	X					
Data Exploration	Statistical Understanding		X			X	X
	Deduplication Complexity	X	X	X	X	X	X
	Heterogeneity in data	X	X	X	X	X	
Data Preprocessing	Dirty data	X	X	X	X	X	X
	Managing sequences in data					X	X
	Managing categorical data				X	X	
Dataset Preparation	Data Dependency	X	X	X	X	X	X
	Data Quality	X	X	X	X	X	X
Data Testing	Tooling	X	X	X	X	X	X
	Expensive Testing	X			X		X
Deployment	Data Extraction Methods	X	X	X	X	X	X
	Overfitting				X	X	
Post Deployment	Data sources and Distribution	X	X	X			
	Data drifts	X	X	X			
	Feedback loops	X					

Data pipeline robustness

- A **robust data pipeline** can be an effective solution to solve some of the data management problems
- **RQ:** How to keep a data pipeline healthy?
  - Automation of the pipeline
  - Monitoring of each step of the pipeline
  - Investigation of the potential errors that might occur at each step and implement detection mechanism
  - Implementation of automated alerting mechanisms and mitigation actions

DataOps

- DevOps manages code, infrastructure and tools
- DataOps adds a fourth dimension – **data**
  - **DataOps** automates the sequence of steps taken to deliver value to the customer where possible, minimize waste and redundancy, and stimulate continuous improvement
  - **DataOps** can support data teams evolve from environments of data silos, backlogs, and quality control issues to an agile, automated, and accelerated data cycle that continuously improves and delivers value to the customers

Dynamic Data Management Challenges*

Focus:

- Challenges involved in storing and preparing data for AI applications in embedded systems
- Collection, storing, pre-processing of data and training of AI applications in distributed architectures
- Data storage and analysis (edge – server - cloud)

Challenges:

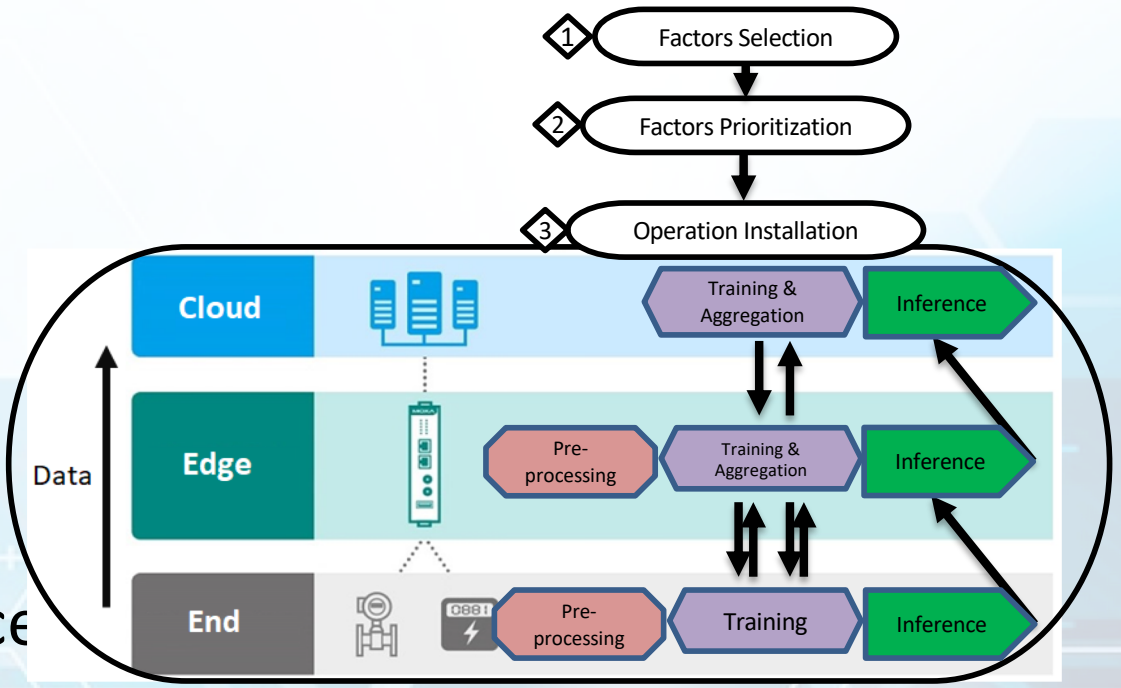
- Expensive and error-prone collection of sensor data
- Difficulties in maintaining semantics of continuously evolving data
- Unclean and noisy data
- Restrictive security constraints
- Difficulties in interpreting heterogeneous and dynamic data sets
- Lack of well-defined purpose and goals

*Ouhaichi H., Olsson H.H., Bosch J. (2019) Dynamic Data Management for Machine Learning in Embedded Systems: A Case Study. *In: Hyrynsalmi S., Suoranta M., Nguyen-Duc A., Tyrväinen P., Abrahamsson P. (eds) Software Business. ICSOB 2019. Lecture Notes in Business Information Processing, vol 370. Springer, Cham*

Architectural distribution framework

Key factors:

- Latency
- Privacy
- Convergence
- Accuracy
- Communcation cost
- Computational resource
- Data quality
- …



Developing ML/DL Models: A Design Framework

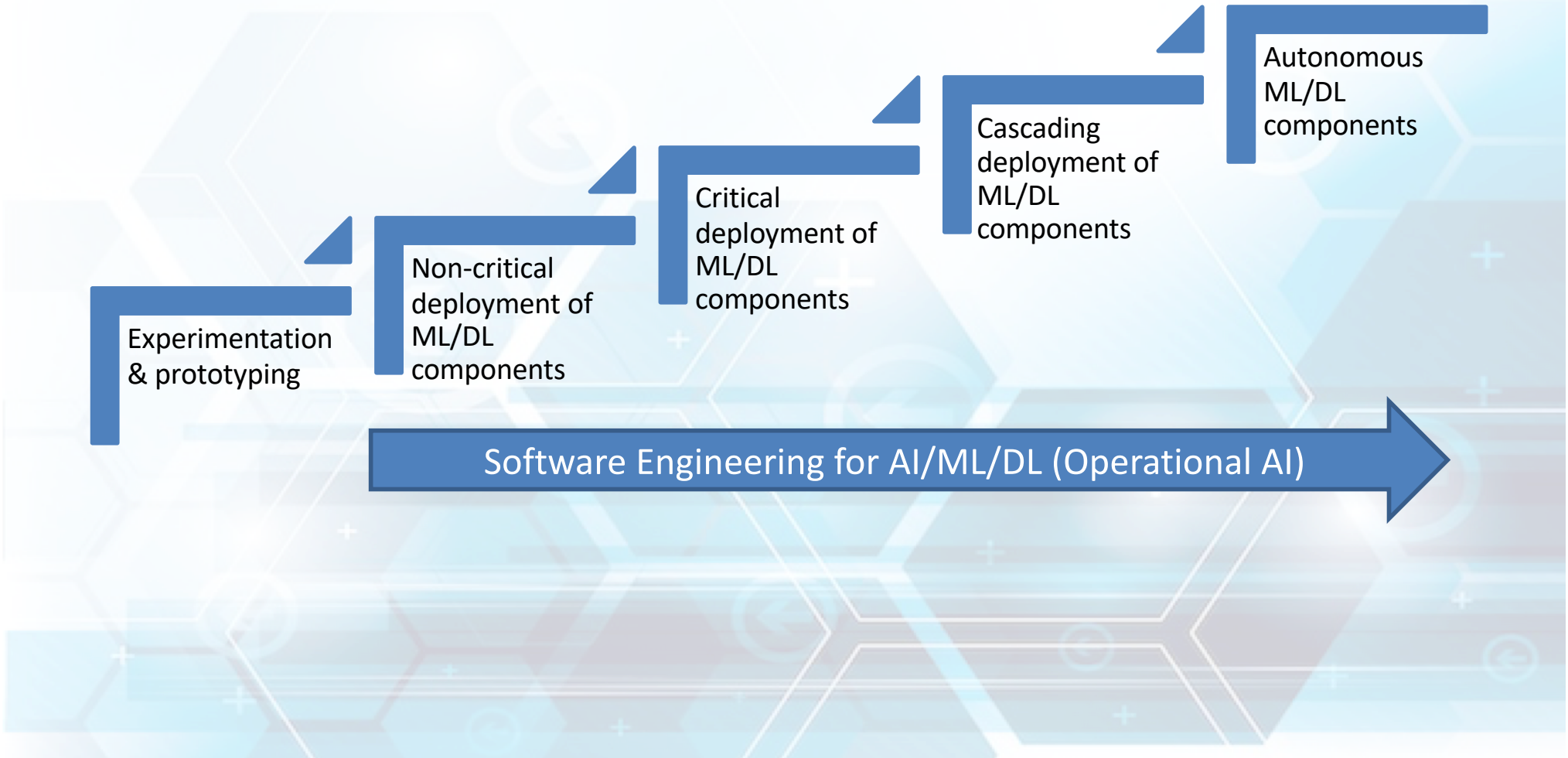
- Organizations integrate ML/DL technologies in software-intensive systems to increase **value delivery**
- Despite the level of expertise, developing high-performing and accurate ML/DL is a **challenging** task
- Need for a structured and systematic **design process** to build and operate ML/DL models
- Design process that identifies the phases involved in development of ML/DL models, the iterations that occur in between these phases and the challenges associated with each phase

Design process of ML/DL models

Phases, activities and challenges when developing ML/DL models

Phases	Activities	Challenges
Business case Specification	• Spend 1-2 days running simple models to check business case • Discussion with Product owners • Verify accessibility and availability of data	• High setup costs • Communication gap with Stakeholders • High AI demands • Data scientists shortage • Dataset availability
Data Exploration	• Meetings with domain experts • Build and test hypothesis around data • Using visualization techniques	• Privacy concerns • Noisy data • Domain Experts accessibilty • Labelling
Feature Engineering	• Start with high-dimensional features and scale down to few-dimensions • Add relevant features scaling from small to large feature set • Add features that are not directly part of the dataset	• Increasing complexity • Improper feature selection
Experimentation	• State-of-the-art approach to find current best algorithm • Obtain basic performance and experimenting with other algorithms • Automate experimentation using tools	• Introduce bias • Selection of algorithms • High DL complexity • DL knowledge need • Validity of related work
Development	• Hyperparameter Tuning • Final model selection depends on end goal requirements - Accuracy, Prediction time, Computational resources, Explainability, etc.	• Difficulty in determining final model • Model execution environment • Complex hyperparameter settings • Verification and Validation
Deployment	• Place requirements on models before deployment • Prepare code ready to put in docker container • Plan for integration with internal systems • Use API service reusable functionality to wrap the model	• Less DL deployment • Integration Problems • Internal deployment
Operational	• A/B testing • Model Execution environment with built-in A/B support, Canary selection, etc. • Services deployed have training interface, inference and evaluation interface • Continuous retraining by adding functionality to existing models	• End user interaction • Model drifts • Training-Serving Skew

Artificial Intelligence



Challenges ML/DL Evolution

	Experiment Prototyping	**Non-critical deployment**	**Critical deployment**	**Cascading deployment**
assemble dataset	Issues with problem formulation and specifying desired outcome	Data silos, scarcity of labelled data, imbalanced training set	Limitations in techniques for gathering training data from large-scale, non-stationary data streams	Complex and effects of data dependencies
create model	Use of non-representative dataset, data drifts	No critical analysis of training data	Difficulties in building highly scalable ML/DL pipeline	Entanglements causing difficulties in isolating improvements
train and evaluate model	Lack of well-established ground truth	No evaluation of models with business-centric measures	Difficulties in reproducing models, results and debugging DL models	Need of techniques for sliced analysis in final model
deploy model	No deployment mechanism	Training-serving skew	Adhering to stringent serving requirements e.g., of latency, throughput	Hidden feedback-loops and undeclared consumers of the models

Table 2. Challenges in the evolution of use of ML/DL components in software-intensive systems

Selected publications

Olsson, H.H., and Bosch, J. (2019). Data Driven Development Adoption Process. *In Proceedings of the 20th International Conference on Product-Focused Software Process Improvement (PROFES), November 27 – 29, Barcelona, Catalunya, Spain.*

Raj, A., Bosch, J., Olsson, H.H., Arpteg, A. and Brinne, B. (2019). Data Management Challenges for Deep Learning. *In Proceedings of the EUROMICRO Conference on Software Engineering for Advanced Applications, August 28 - 30, Kallithea, Chalkidiki, Greece.*

Lwakatare, L. E., Raj, A., Bosch, J., Olsson, H. H., & Crnkovic, I. (2019). A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation. *In International Conference on Agile Software Development (pp. 227-243). Springer, Cham.*

Mattos, D.I., Bosch, J., Olsson, H.H., Dakkak, A., and Bergh, K. (2019). Automated Optimization of Software Parameters in a Long-Term Evolution Radio Base Station. *In Proceedings of the IEEE Annual Systems Conference, April 8 – 11, Orlando, Florida, USA.*

Mattos, D.I., Bosch, J., and Olsson, H.H (2019). Multi-armed bandits in the Wild: Common Pitfalls in Online Experiments. *Information and Software Technology, pp.1-14.*

Mattos, D.I., Bosch, J., and Olsson, H.H. (2019). ACE : Easy Deployment of Field Optimization Experiments. *In the Proceedings of the European Conference on Software Architecture (ECSA), September 9 – 13, Paris, France.*

Mattos, D.I., Bosch, J., and Olsson, H.H. (2019). Leveraging Business Transformations with ML Experiments. *In Proceedings of the International Conference on Software Business (ICSOB), November 18 – 20, Jyväskylä, Finland.*

Figalist I., Elsner C., Bosch J., Olsson H.H. (2019) Scaling Agile Beyond Organizational Boundaries: Coordination Challenges in Software Ecosystems. *In: Agile Processes in Software Engineering and Extreme Programming. XP 2019. Lecture Notes in Business Information Processing, vol 355. Springer, Cham.*

Figalist I., Elsner C., Bosch J., Olsson H.H. (2019) Business as Unusual: A Model for Continuous Real-Time Business Insights Based on Low Level Metrics. *In: Proceedings of the 45th EUROMICRO Conference on Software Engineering and Advanced Applications, Kallithea, Greece.*

Figalist I., Elsner C., Bosch J., Olsson H.H. (2019) Customer Churn Prediction in B2B Contexts. *In Proceedings of the International Conference on Software Business (ICSOB), November 18 – 20, Jyväskylä, Finland.*

Ouhaichi, H., Olsson, H.H., and Bosch, J. (2019). Dynamic Data Management for Machine Learning in Embedded Systems: A Case Study. *In Proceedings of the International Conference on Software Business (ICSOB), November 18 – 20, Jyväskylä, Finland.*