Hybrid AI for Trustworthy AI

Fredrik Heintz
Dept. of Computer Science, Linköping University
fredrik.heintz@liu.se
@FredrikHeintz

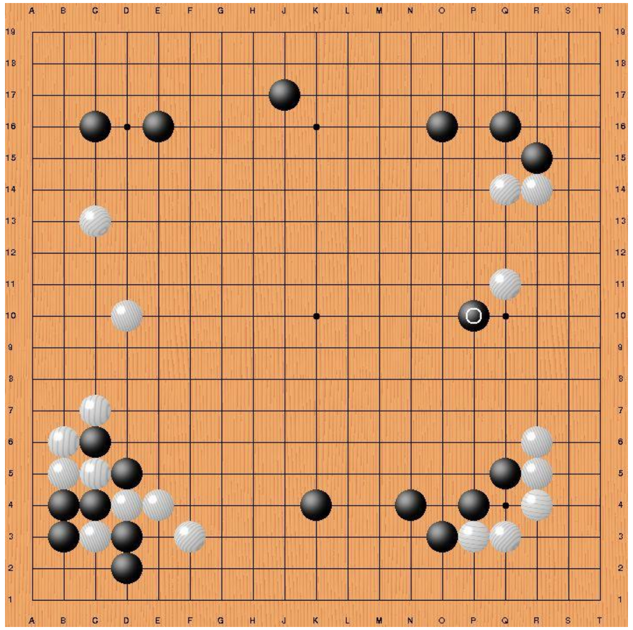
LINKÖPINGS UNIVERSITET

How to Evaluate AI Systems?

George Zarkadakis, Contributor
AI engineer and writer

Move 37, or how AI can change the world

11/26/2016 09:35 am ET



https://www.huffpost.com/entry/move-37-or-how-ai-can-change-the-world_b_58399703e4b0a79f7433b675

Ethics Guidelines for Trustworthy AI – Overview

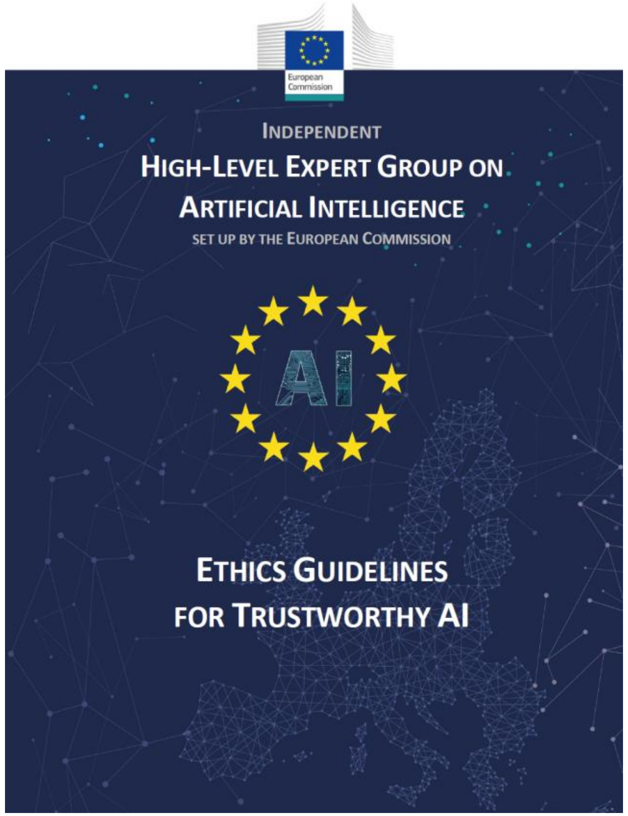
Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

Lawful AI	Ethical AI	Robust AI

Three levels of abstraction

from principles (Chapter I)	to requirements (Chapter II)	to assessment list (Chapter III)



Ethics Guidelines for Trustworthy AI – Principles

4 Ethical Principles based on fundamental rights

Respect for human autonomy

Augment, complement and empower humans

Prevention of harm

Safe and secure.
Protect physical and mental integrity.

Fairness

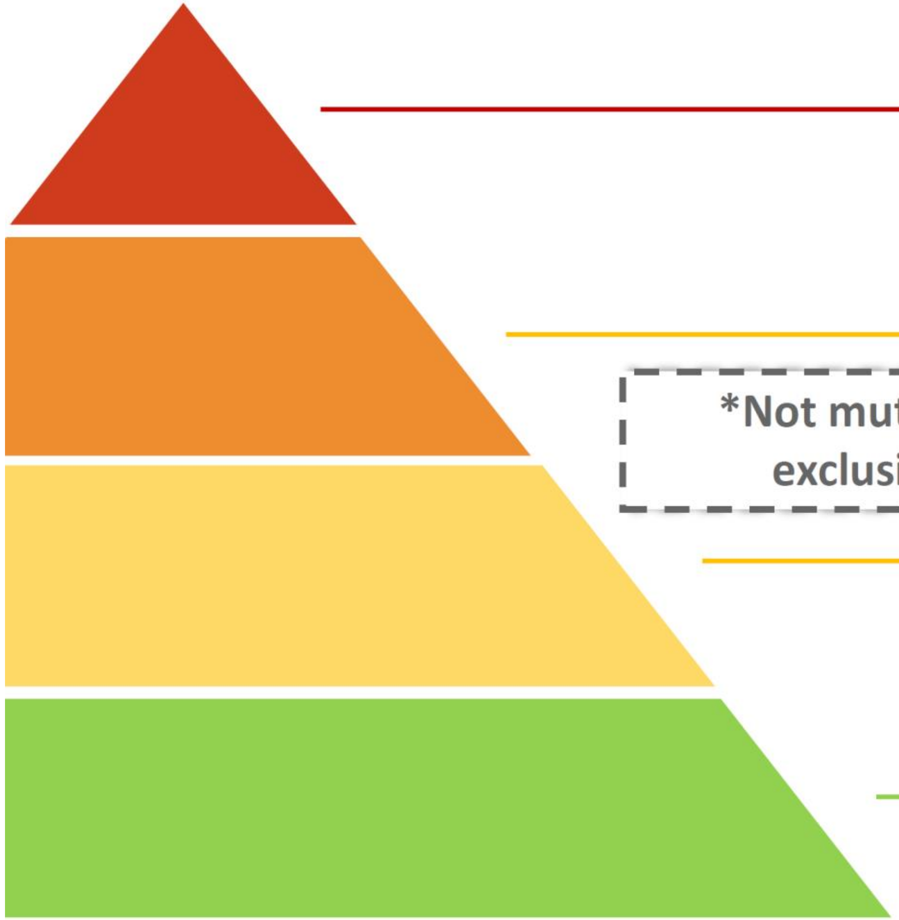
Equal and just distribution of benefits and costs.

Explicability

Transparent, open with capabilities and purposes, explanations

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

A risk-based approach



*Not mutually exclusive

Unacceptable risk
e.g. social scoring — **Prohibited**

High risk
e.g. recruitment, medical devices — **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

'Transparency' risk
'Impersonation' (bots) — **Permitted** but subject to information/transparency obligations

Minimal or no risk — **Permitted** with no restrictions

Requirements for high-risk AI systems (Title III, Chapter 2)

Establish and implement **risk management system**

&

in light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Draw up **technical documentation** & set up **logging capabilities** (traceability & auditability)
- Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Fredrik Heintz

Dept. of Computer Science, Linköping University

fredrik.heintz@liu.se, @FredrikHeintz

https://tailor-network.eu/

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme GA No 952215

LINKÖPINGS UNIVERSITET

TAILOR – Vision

Develop the scientific foundations for Trustworthy AI

- integrating learning, optimisation and reasoning
- realising the European Vision of human-centered trustworthy AI.

This project is funded by the EC under H2020 ICT-48

Boosting Capacity to Tackle Major Scientific Challenges

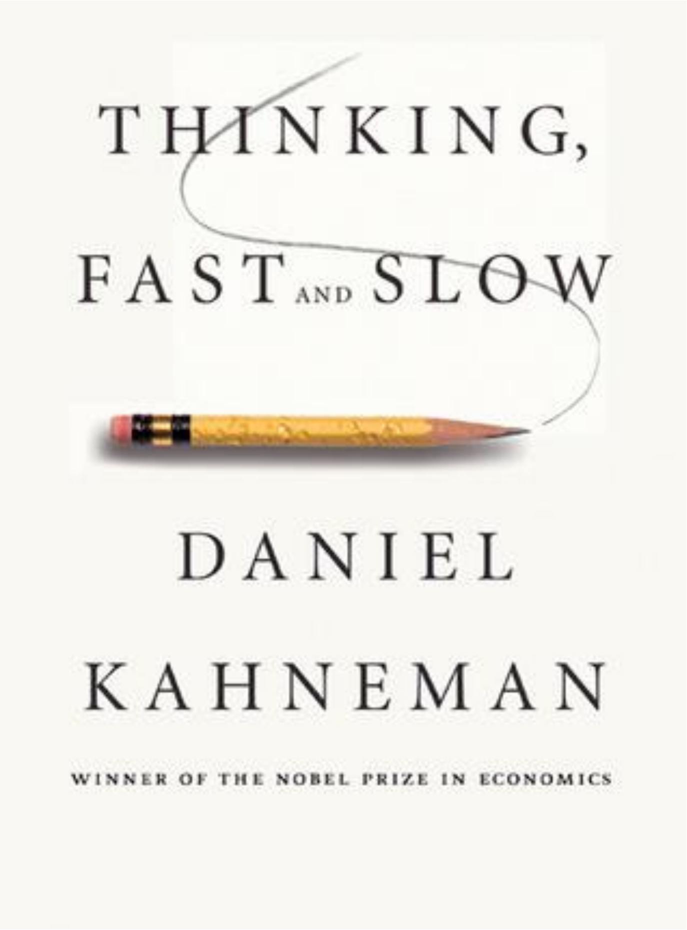
- A **core network** of outstanding AI research centres and major European companies (partners) plus **mechanisms for extending** the network (network members and connectivity fund) to be adaptive and inclusive.
- Five **virtual research environments** to address the **major scientific challenges** required to achieve Trustworthy AI supported by **AI-based** network **collaboration tools**.
- **Strategic** research and innovation **roadmap** to drive the long-term **scientific vision** combined with **bottom-up coordinated actions** collaboratively addressing specific research questions.

This project is funded by the EC under H2020 ICT-48

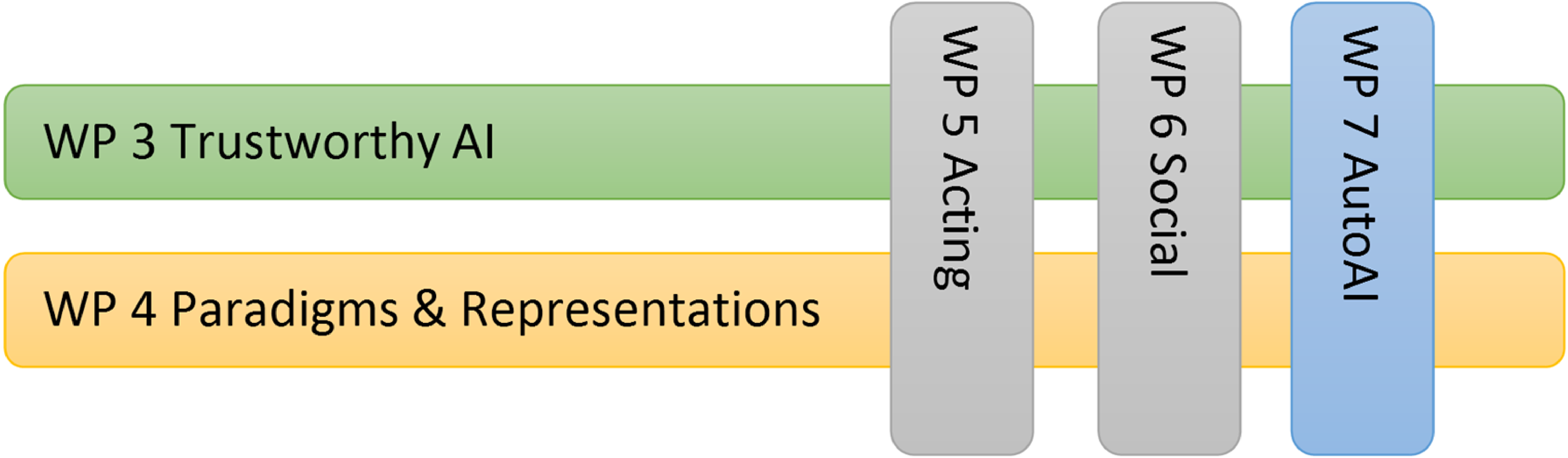
Human and Computational Thinking

Figure 1: A Comparison of System 1 and System 2 Thinking

System 1	System 2
"Fast"	"Slow"
DEFINING CHARACTERISTICS	DEFINING CHARACTERISTICS
Unconscious	Deliberate and conscious
Effortless	Effortful
Automatic	Controlled mental process
WITHOUT self-awareness or control	WITH self-awareness or control
"What you see is all there is."	Logical and skeptical
ROLE	ROLE
Assesses the situation	Seeks new/missing information
Delivers updates	Makes decisions



TAILOR – Basic Research Program

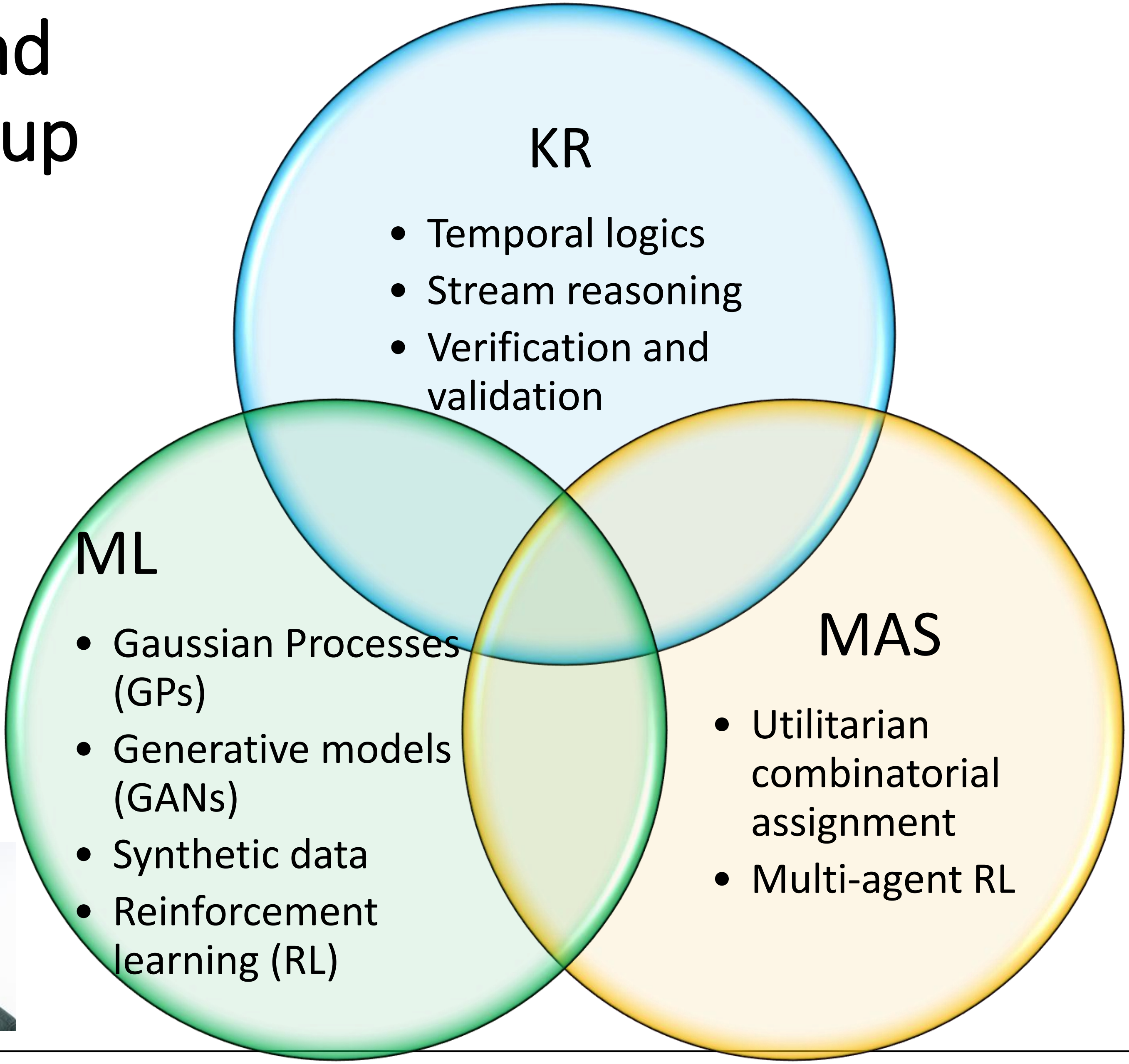


This project is funded by the EC
under H2020 ICT-48

Reasoning and Learning Group

6 PhD students

2 Postdocs
1 Research engineer



Knut and Alice Wallenberg Foundation

WASP

STIFTELSEN MARCUS OCH AMALIA WALLENBERGS MINNESFOND

VINNOVA

NFFP / UDI

ELLIIT

TAILOR

WASP—HS

LINKÖPINGS UNIVERSITET

Research Overview

Learning generative models based on trajectory data

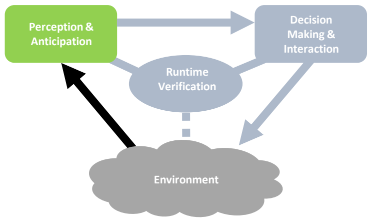
Probabilistic logical reasoning over observed and predicted trajectories

Utilitarian Combinatorial Assignment

Collaborative Unmanned Aircraft Systems

A principled approach to building collaborative intelligent autonomous systems for complex missions.

Motion Pattern Recognition

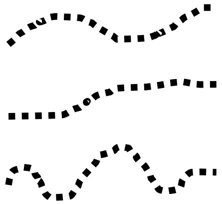


Generalized Motion Pattern Model

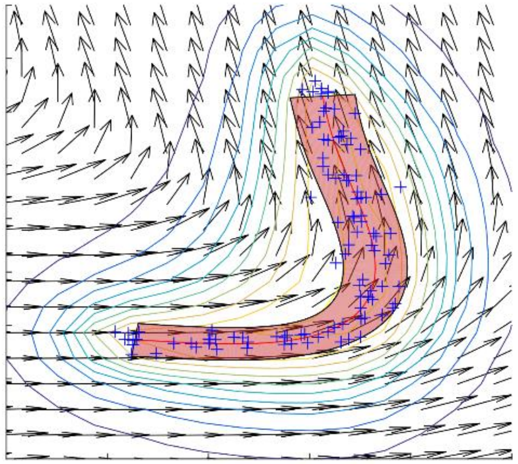
- Based on Gaussian processes
- Generative auto-encoder

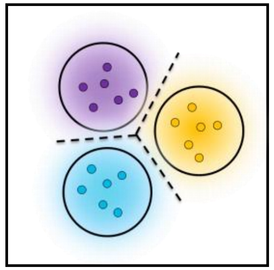
Tracker

Trajectories



Motion Pattern

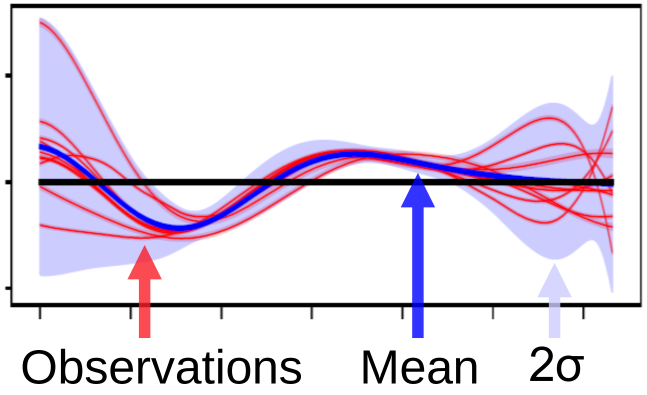




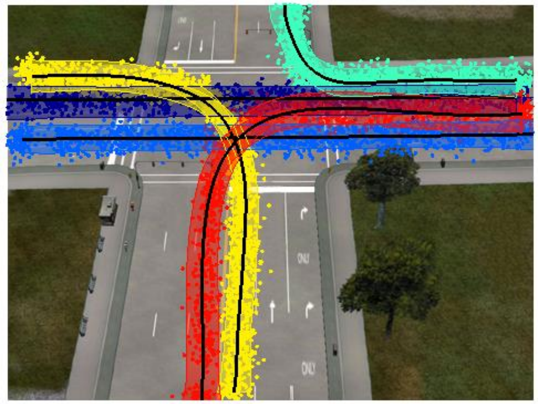
Multi-task

- One-class classification (anomaly detection)
- Multi-class classification
- Predict continuation
- Predict sequence
- Temporally *align* trajectories

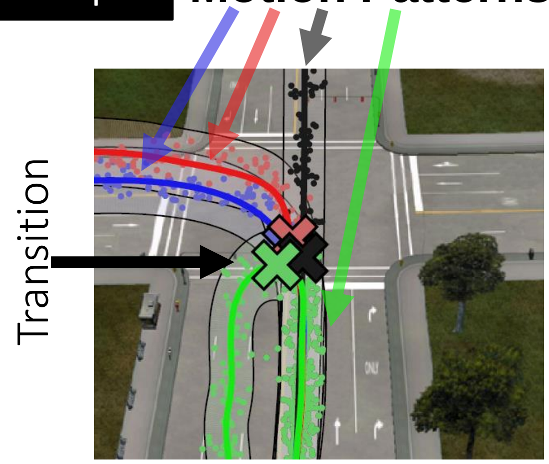
Example **Motion Pattern**



Example **Motion Patterns**

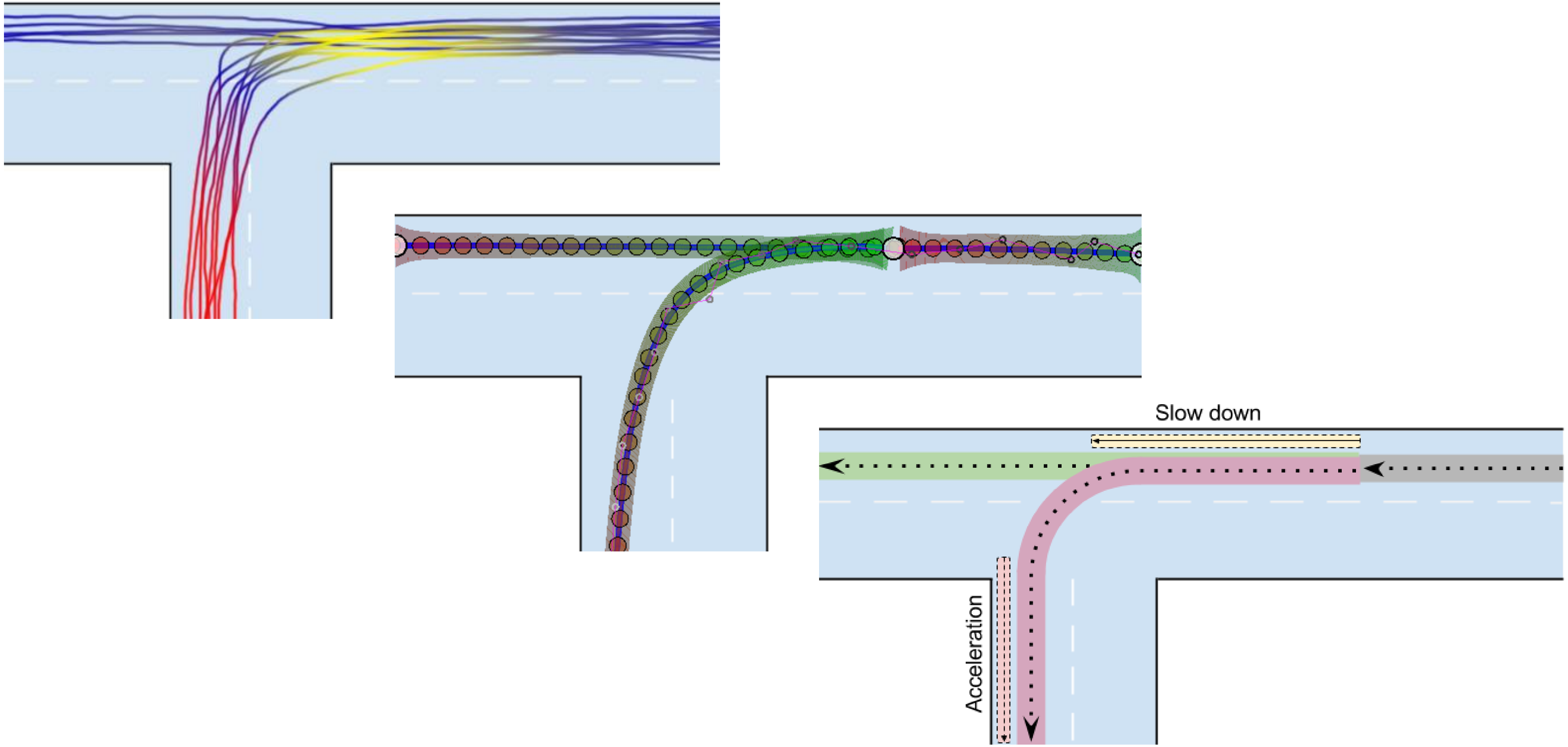


Example **Motion Patterns**



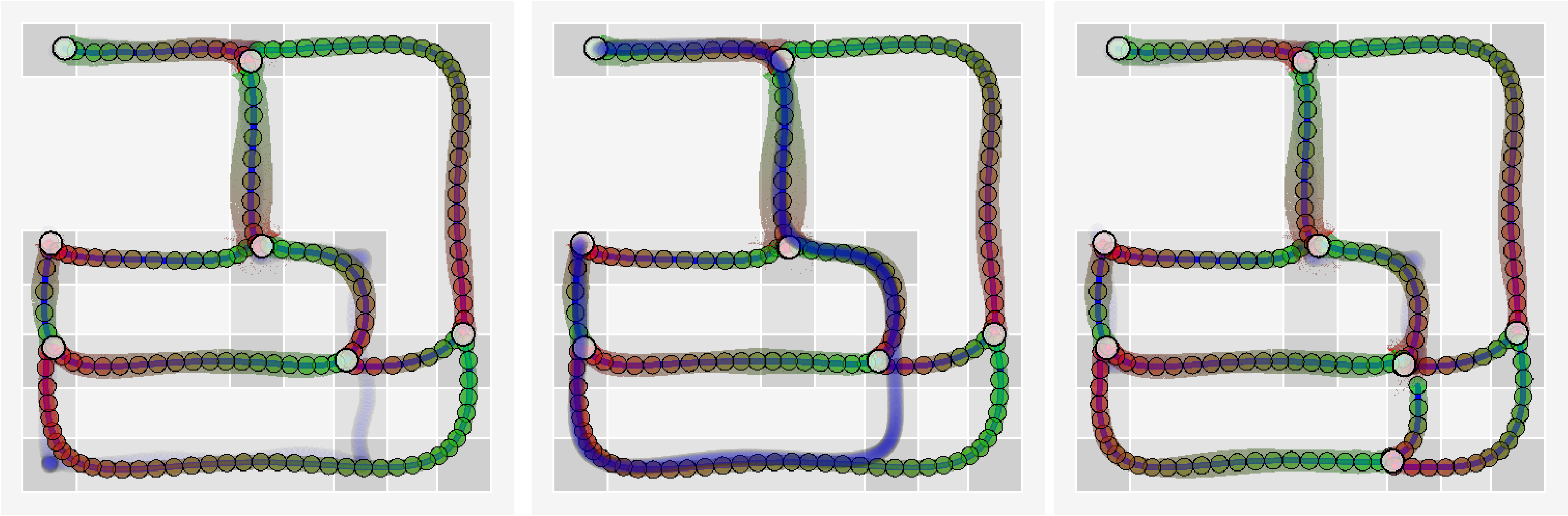
Learning generative models based on trajectory data

[Tiger and Heintz IV 2018, Tiger and Heintz FUSION 2015, Tiger and Heintz STAIRS 2014]



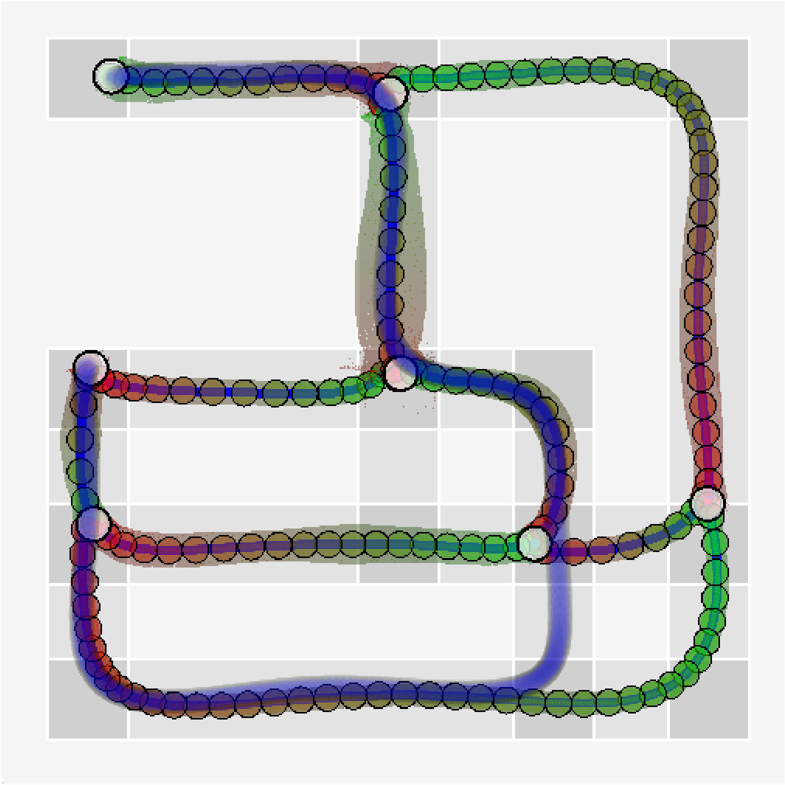
Learning generative models based on trajectory data

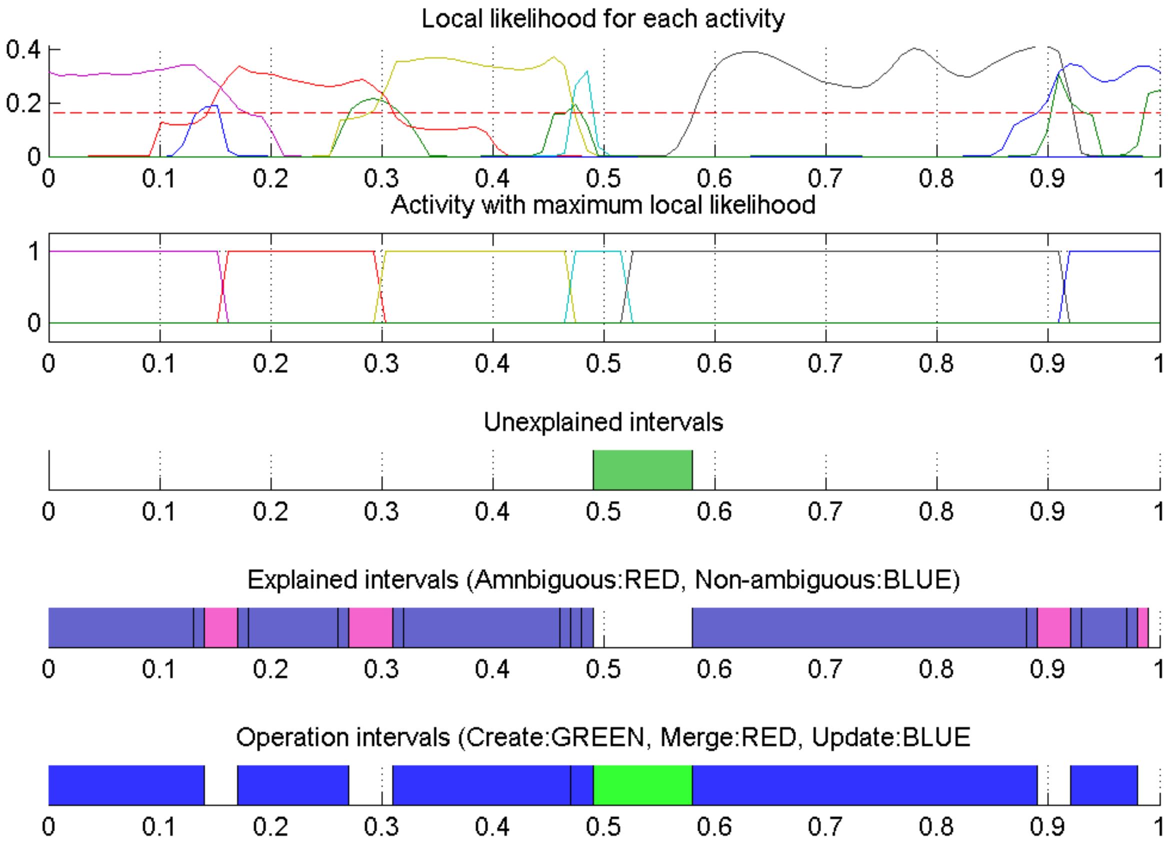
[Tiger and Heintz IV 2018, Tiger and Heintz FUSION 2015, Tiger and Heintz STAIRS 2014]



Learning generative models based on trajectory data

[Tiger and Heintz IV 2018, Tiger and Heintz FUSION 2015, Tiger and Heintz STAIRS 2014]





LINKÖPINGS UNIVERSITET

Learning generative models based on trajectory data

[Tiger and Heintz IV 2018, Tiger and Heintz FUSION 2015, Tiger and Heintz STAIRS 2014]

<u>State-of-the-art **Flow Field** approach</u>

$$(\boldsymbol{p_x}, \boldsymbol{p_y}) \rightarrow \boldsymbol{v_x}, \boldsymbol{v_y}$$

With two GP modelled latent functions:

$$[v_x \;\; v_y] = \left[f_{v_x}(p_x, p_y) \;\; f_{v_y}(p_x, p_y)\right]$$

<u>Proposed **Inverse Mapping** approach</u>

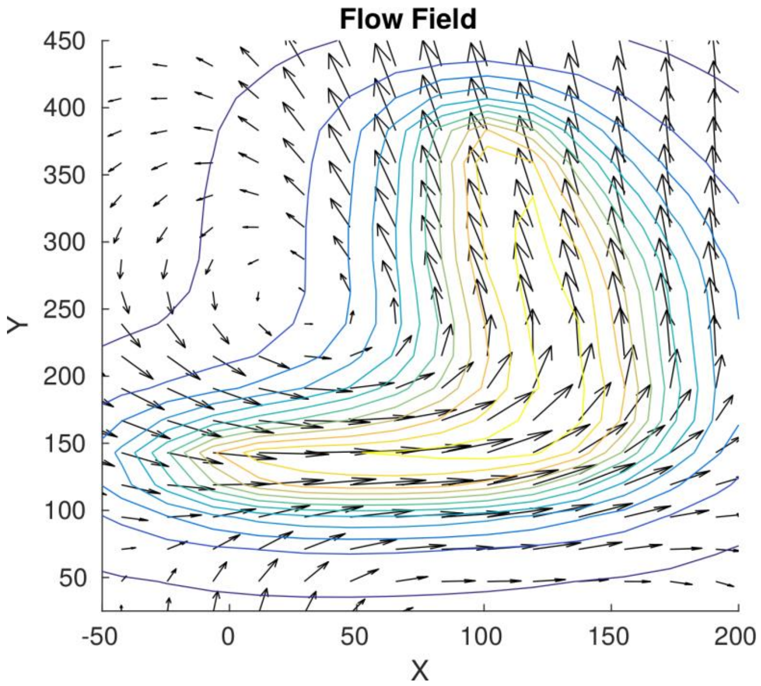
$$(\boldsymbol{p_x}, \boldsymbol{p_y}) \rightarrow \boldsymbol{\tau} \rightarrow \boldsymbol{p_x}, \boldsymbol{p_y}, \boldsymbol{v_x}, \boldsymbol{v_y}$$

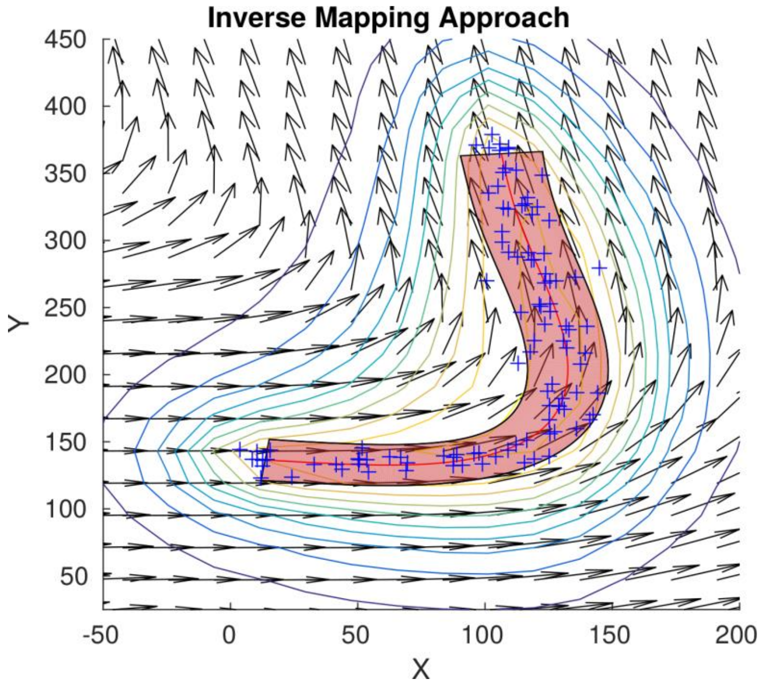
With five GP modelled latent functions:

$$[p_x \;\; p_y] = \left[f_{p_x}(\tau) \;\; f_{p_y}(\tau)\right]$$
$$[v_x \;\; v_y] = \left[f_{v_x}(\tau) \;\; f_{v_y}(\tau)\right]$$
$$\tau = f_\tau(p_x, p_y)$$

where $\tau \in [0\ 1]$ is parametrized time (motion pattern progression)

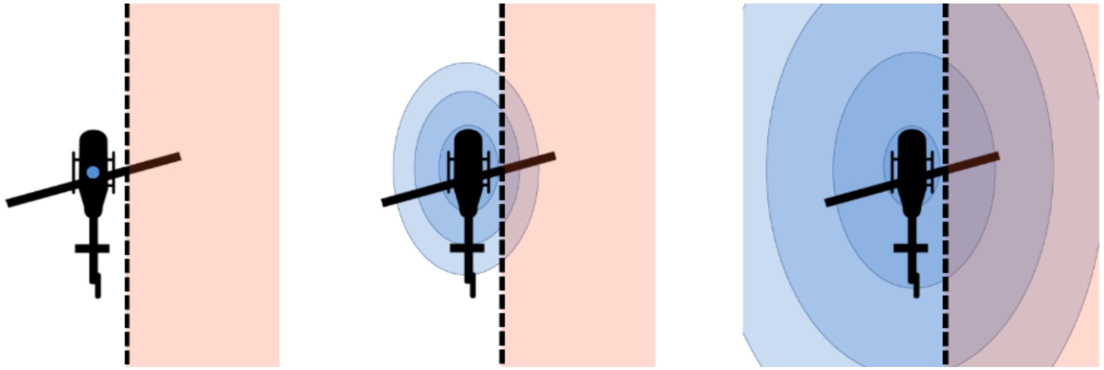
Models *flow*, *spatial extent*, *spatial locality* and *motion progression*.





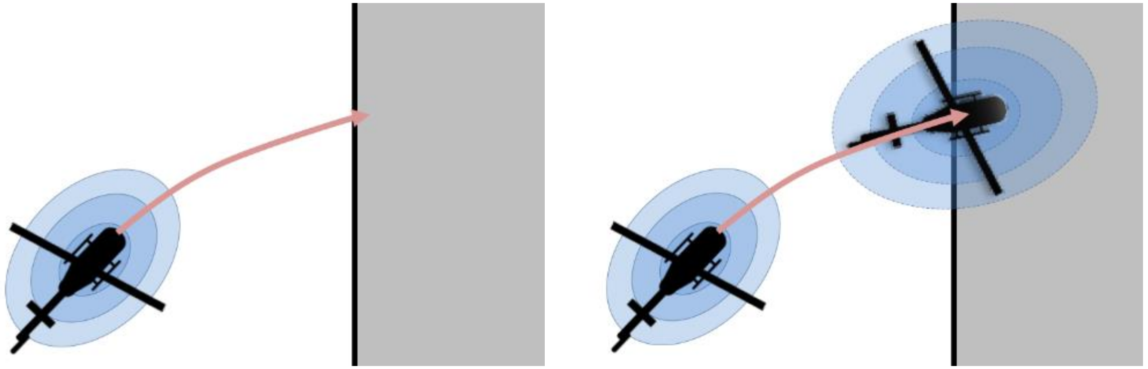
Probabilistic Predictive Stream Reasoning

[Tiger and Heintz TIME 2016, IJAR 2020]



collision: false   Pr(collision) = 0.1   Pr(collision) = 0.4

Reasoning over <u>Uncertainty</u>



Pr(collision now) = 0.0...   Pr(collision soon) = 0.5

Reasoning over <u>Predictions</u>

Mattias Tiger and Fredrik Heintz. 2020.
Incremental Reasoning in Probabilistic Signal Temporal Logic.
International Journal of Approximate Reasoning, **119**:325–352. Elsevier.

Probabilistic Predictive Stream Reasoning

[Tiger and Heintz TIME 2016, IJAR 2020]

always (altitude$_0$ > 3)

true

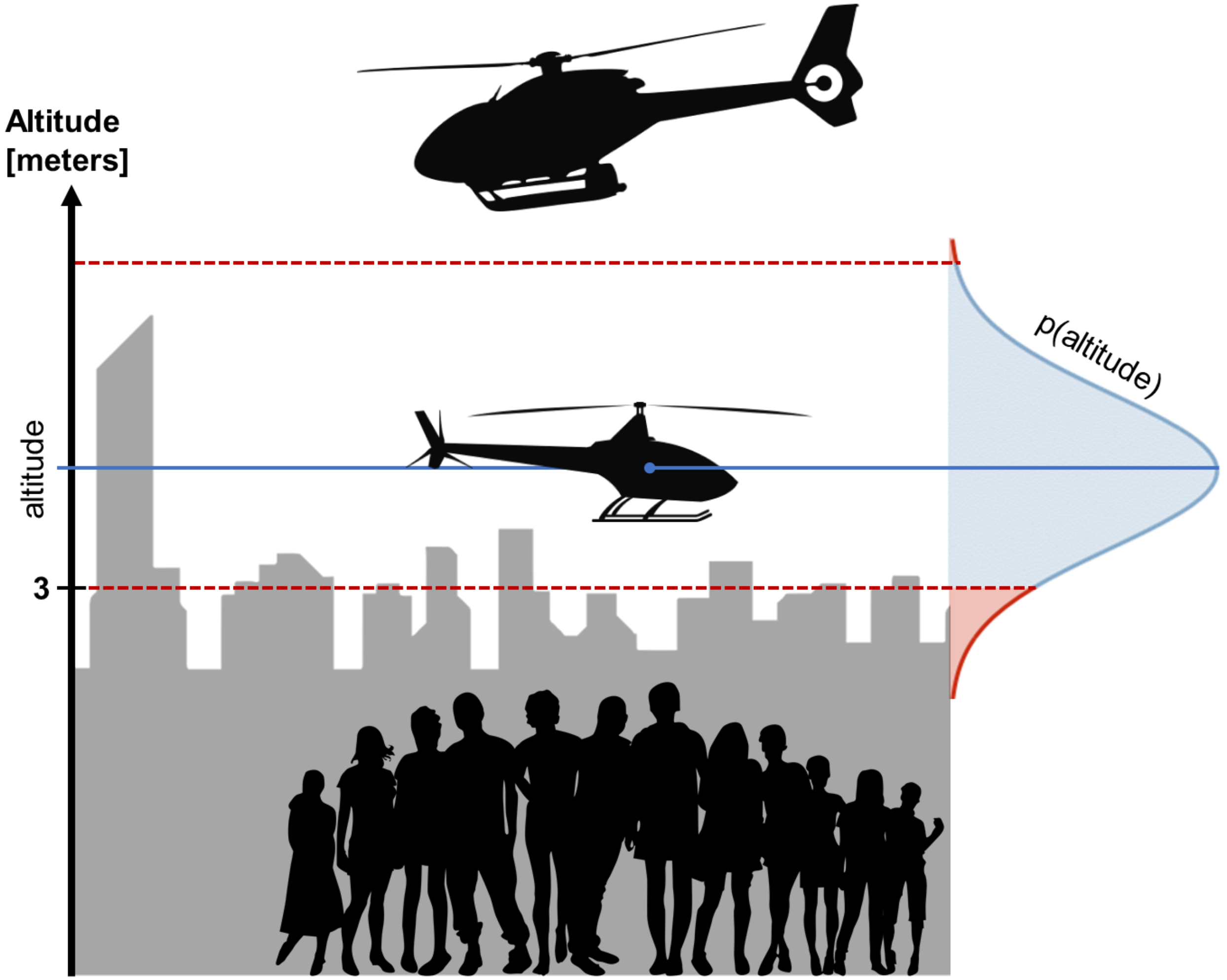
always (Pr(altitude$_{0|0}$ > 3) ≥ 0.99)

false

always (Pr(altitude$_{2|0}$ > 3) ≥ 0.99)

Relative time to estimate

Relative time to estimate from

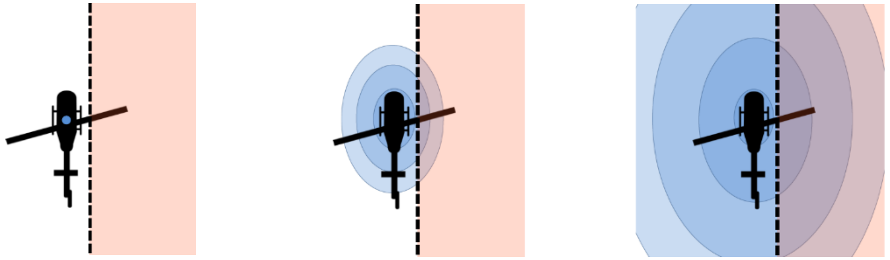


Probabilistic logical reasoning over observed and predicted trajectories

[Tiger and Heintz TIME 2016, IJAR 2020]

- Probabilistic
  - Is the UAV inside the no-fly-zone?

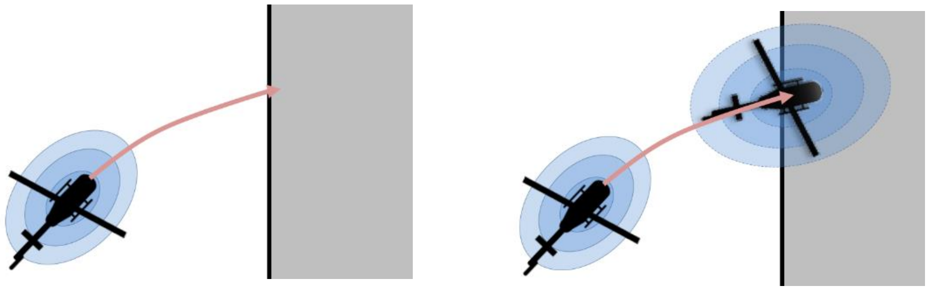
Reasoning over <u>Uncertainty</u>



collision: false | Pr(collision) = 0.1 | Pr(collision) = 0.4

- Anticipatory
  - Will the UAV be colliding in the near future?

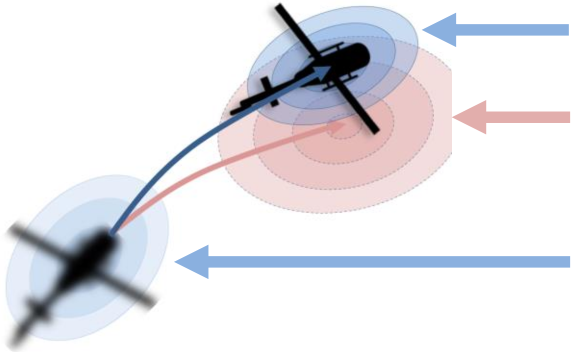
Reasoning over <u>Predictions</u>



Pr(collision now) = 0.0 | Pr(collision soon) = 0.5

- Introspective
  - Is the prediction similar to the realization?

Reasoning <u>about</u> Predictions

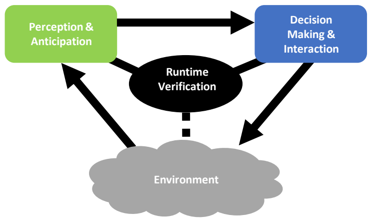


Estimated state (<u>now</u>)

Predicted state

Estimated state (when predicting)

Introspective Motion Planning and Control



Challenges

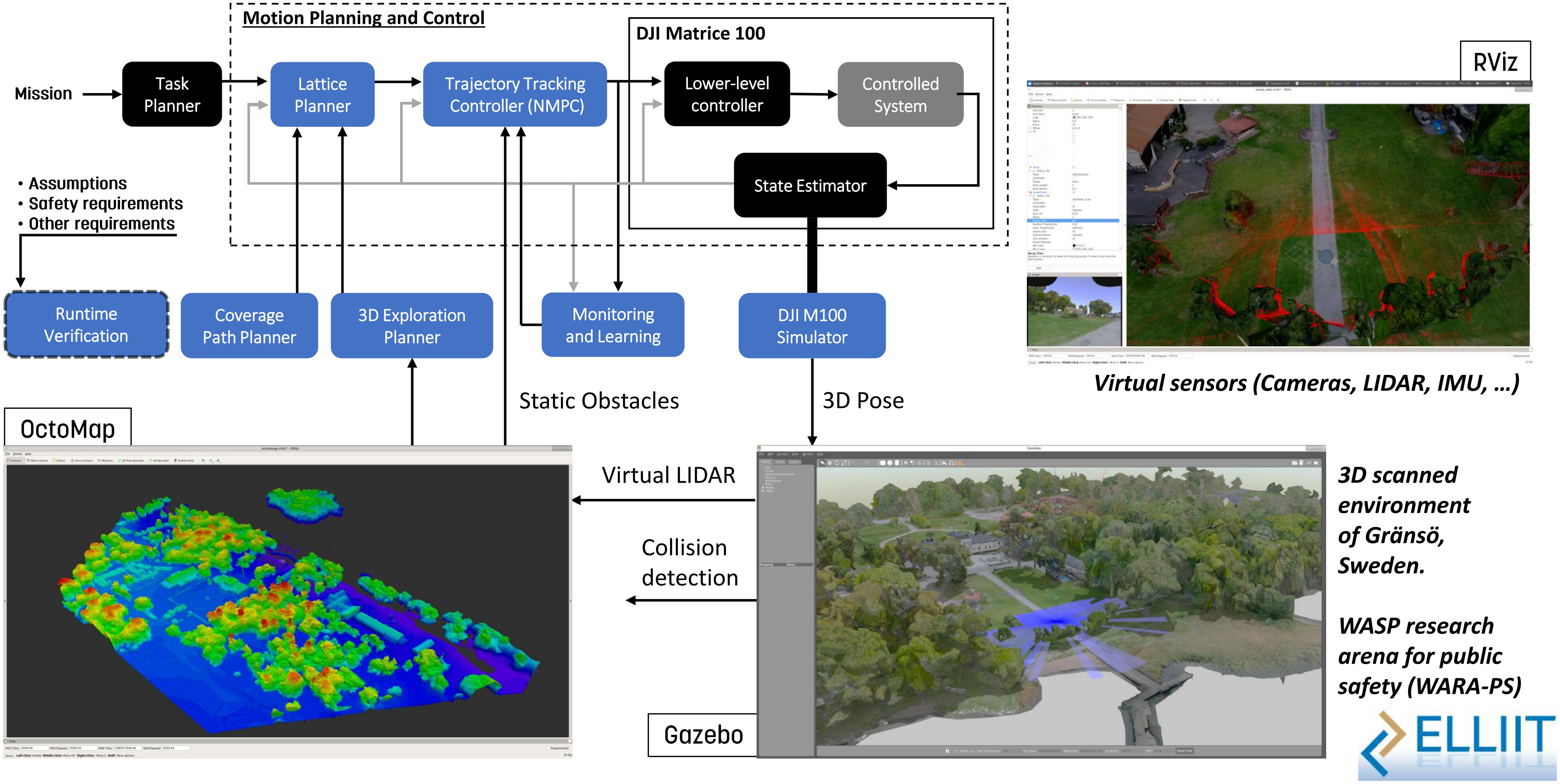
- What is **normal** behavior?
- Is the robot **behaving normally**?
- **Safe**, but not task effective?
- Are **learned models** safe to use?

Our Approach

- **Learn models** of execution variations from action sequences
- **Monitor executions** with respect to models during deployment
- Use **models with explicit uncertainty quantification** for tight safety-bounds
- **Monitor models** with respect to executions during deployment

[9] M. Tiger, et al. *Enhancing Lattice-Based Motion Planning With Introspective Learning and Reasoning.* IEEE Robotics and Automation Letters 6.3 (2021): 4385-4392.

Example AI-Robotics Stack and Simulation Environment

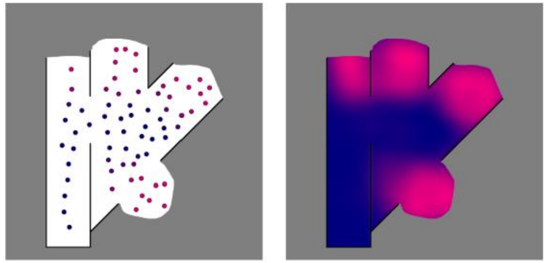


Motion Planning Applications

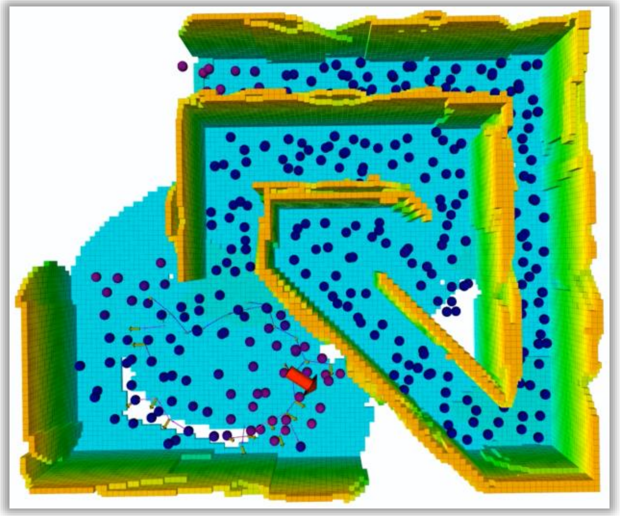
3D Exploration Planning | Coverage Path Planning

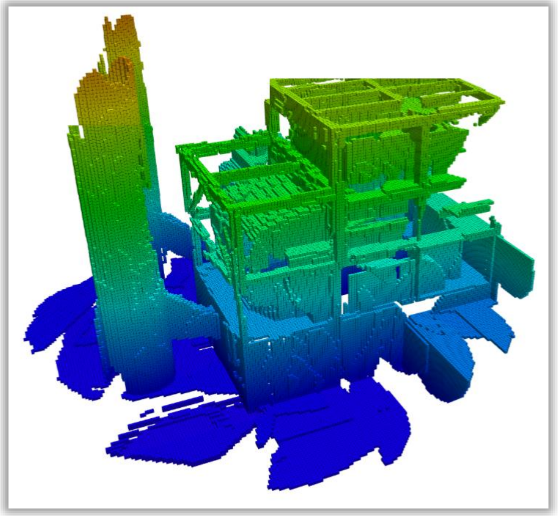
- Mapping
- Inspection
- Search for anomalies

Made efficient by Bayesian ML



Large-scale, complex geometry, maze-like structures

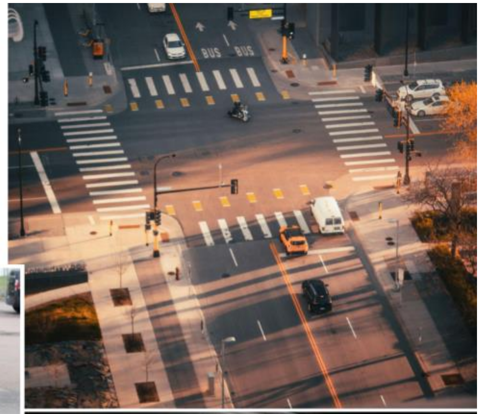


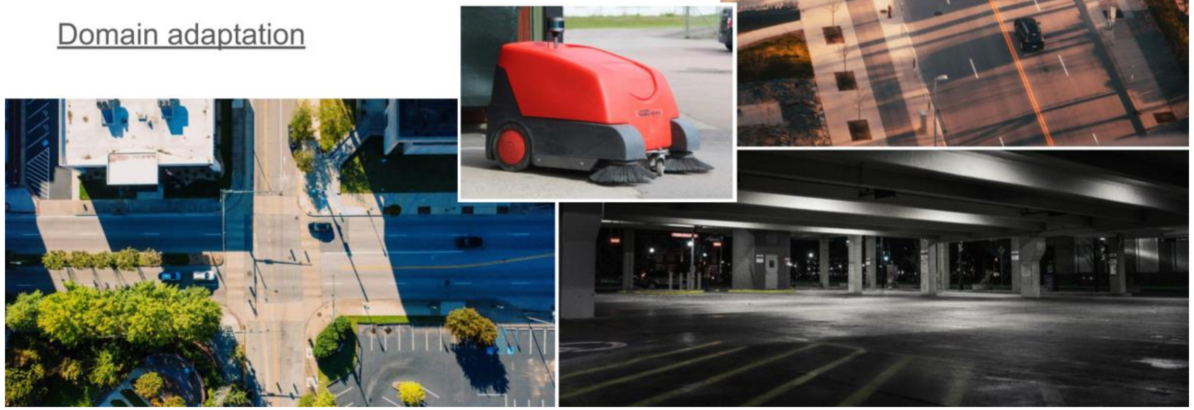


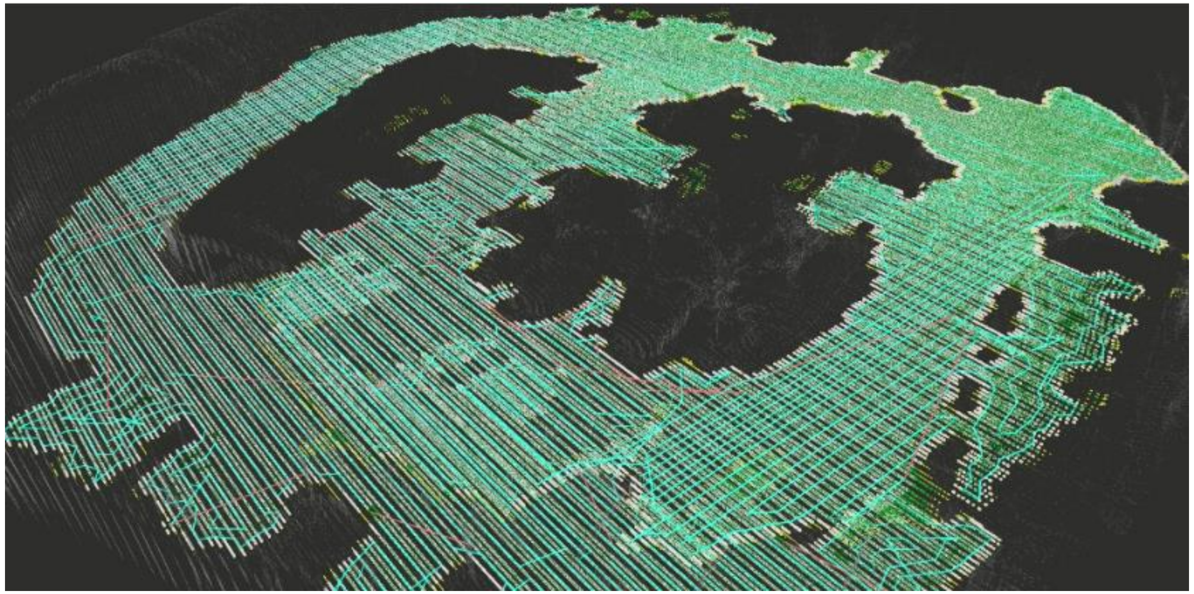
Coverage Path Planning for Road Sweeping in **Urban Environments**.

Large-scale, complex geometry, multi-floor, slopes, uneven ground, non-trivial space division

Domain adaptation



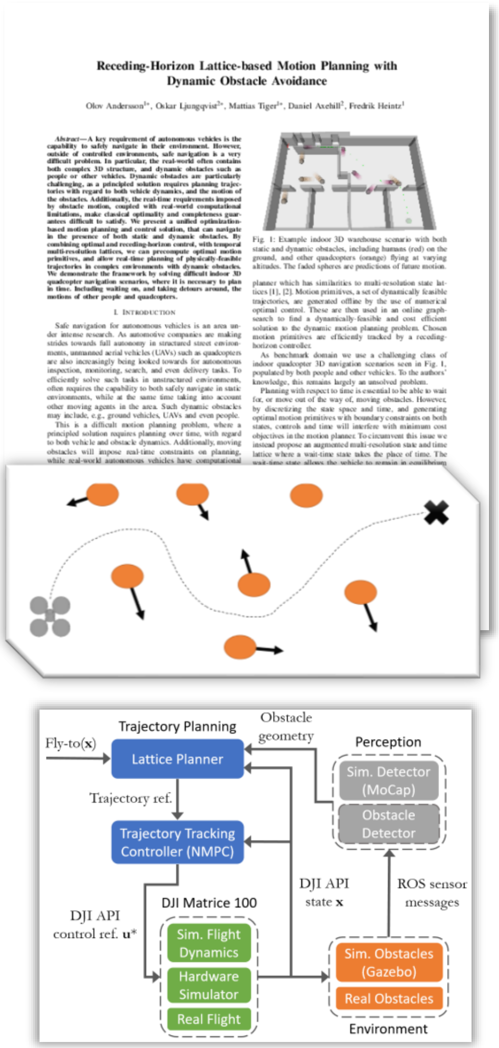




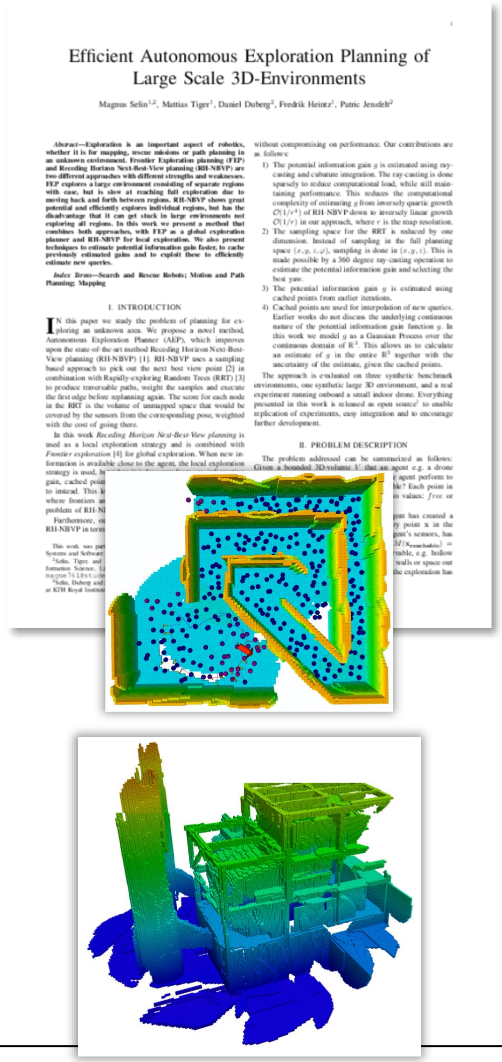
[6] M. Selin, M. Tiger, D. Duberg, F. Heintz and P. Jensfelt. *Efficient autonomous exploration planning of large-scale 3D environments*. IEEE Robotics and Automation Letters 4.2 (2019): 1699-1706.
[7] D. Engelson, M. Tiger and F. Heintz. *Coverage Path Planning in Large-scale Multi-floor Urban Environments with Applications to Autonomous Road Sweeping*. IEEE ICRA (2022). (Submitted)

Many Publications Related to Different Components

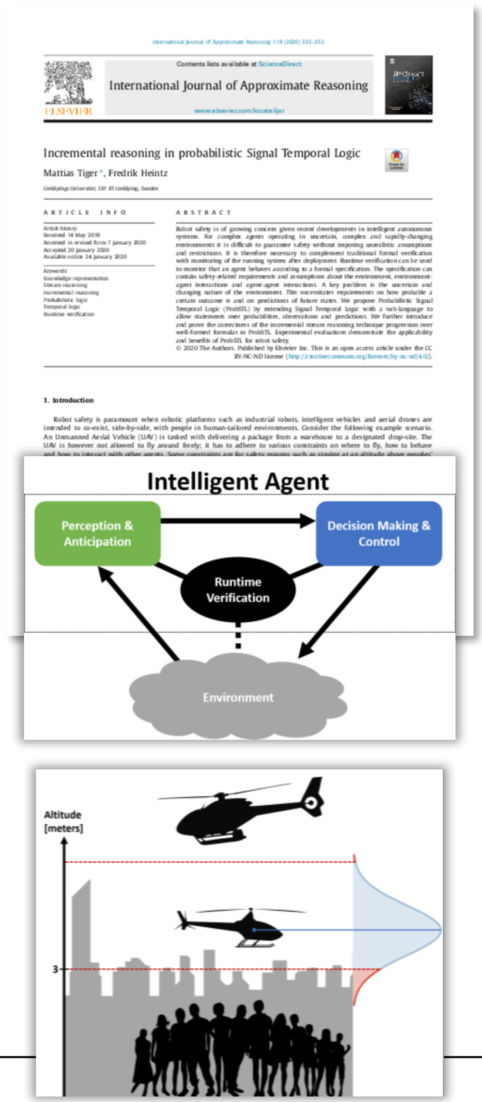
Motion Planning



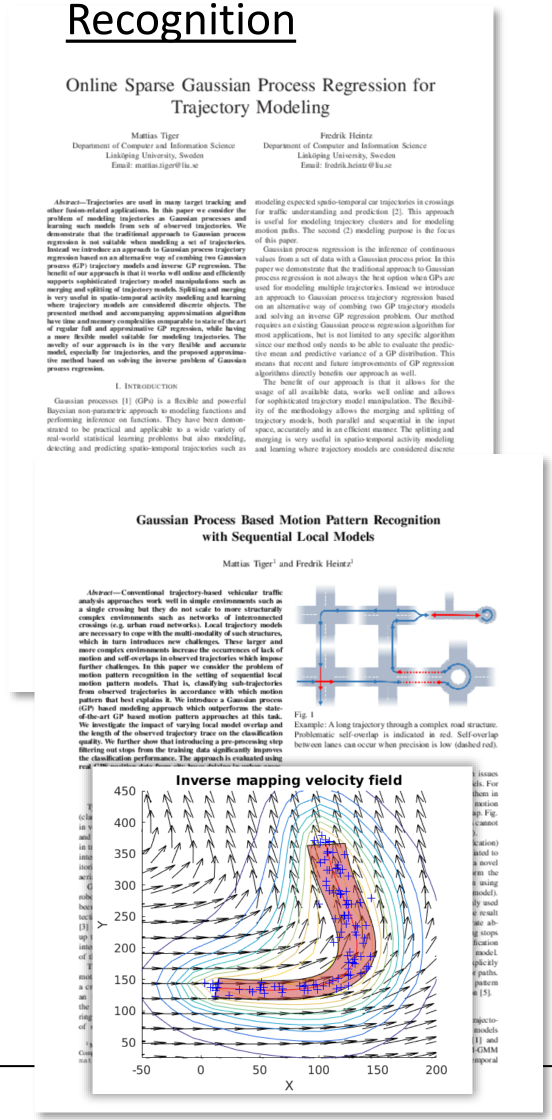
3D Exploration



Runtime Verification



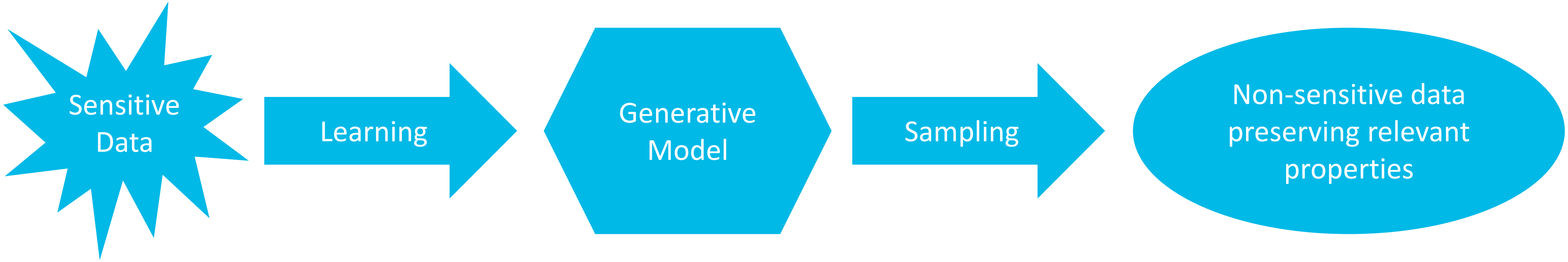
Motion Pattern Recognition



LINKÖPINGS UNIVERSITET

Privacy-preserving synthetic data generation

[D. Bergström, Md F. Sikder, R. Ramachandranpillai]



1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate enough
4. Methods for verifying that the synthetic data set is non-sensitive

Synthetic Healthdata – Existing Models, Measures, and Problems

Models

- MedGAN[1]
- HealthGAN[2]
- Synthea[3]

Measures

- Privacy
- Utility
- Resemblance

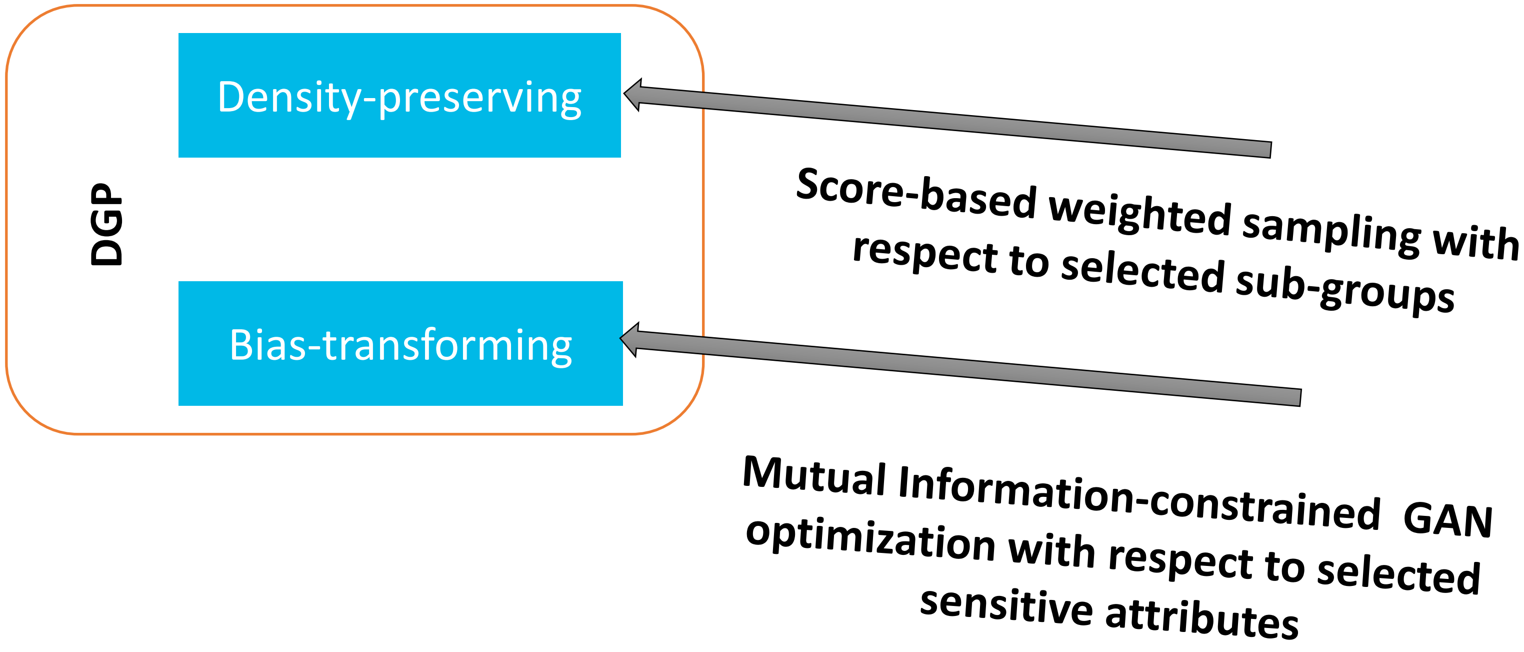
Research Gap

- Fairness? [4]

1. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S. and Yang, B., 2020. MedGAN: Medical image translation using GANs. *Computerized medical imaging and graphics*, *79*, p.101684.
2. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A. and Bennett, K.P., 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, pp.244-255.
3. Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T. and McLachlan, S., 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), pp.230-238.
4. Bhanot, K., Qi, M., Erickson, J.S., Guyon, I. and Bennett, K.P., 2021. The problem of fairness in synthetic healthcare data. *Entropy*, *23*(9), p.1165.

Our Proposed Solution : Generate fair health data from biased data

- We define a **Fair Data Generation Process** (FDGP) in **Generative Adversarial Networks** (GAN) :

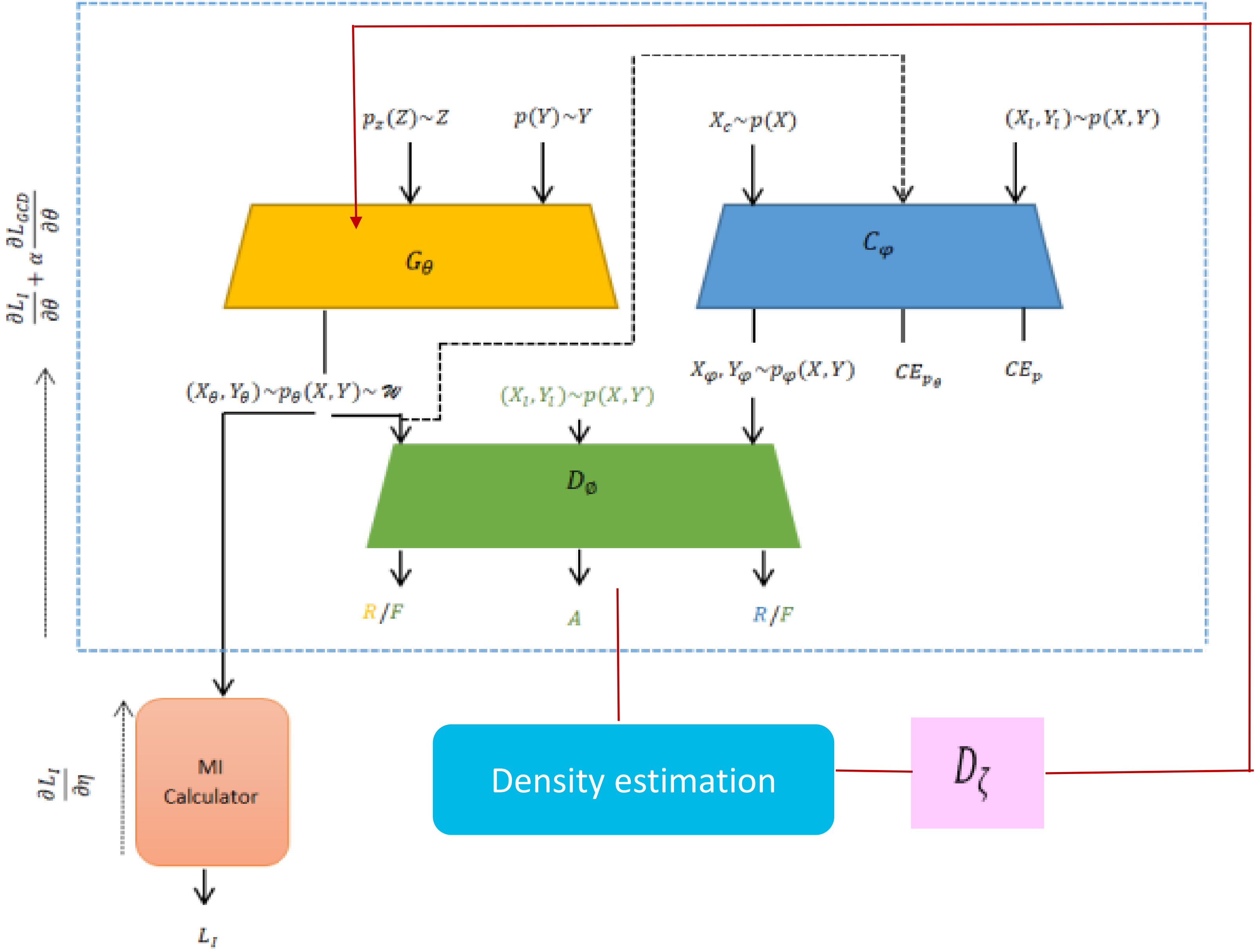


Goals:

- Formal equality with respect to some sub-groups
- Substantive equality with respect to some sensitive attributes
- Data utility

DeMISe

$$L_F = \underbrace{\underbrace{L_{GCD}}_{Semi-supervised-generation} + \underbrace{\alpha L_{MI}}_{MI\,de-biasing}}_{Fair-generation}$$



LINKÖPINGS UNIVERSITET

Resmi Ramachandranpillai, Fahim Sikder, David Bergström & Fredrik Heintz. (2022). Improving synthetic data fairness in healthcare settings (Under review).

ELLIIT

Results and Discussions

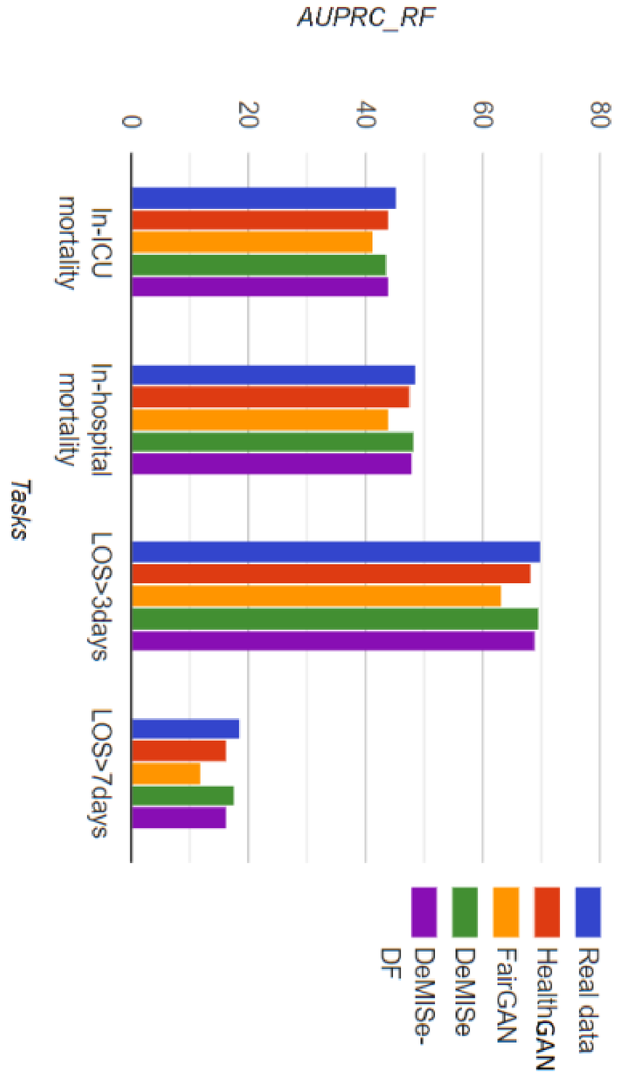
Dataset	Medical Information Mart Intensive Care , Version 3 [4]
Benchmarks	HealthGAN and FairGAN [5]
Evaluation Metric	Data utility - Accuracy and F1 score Downstream fairness (substantive equality) - AUROC gap and Demographic Parity gap Fair resemblance - density score

4. https://physionet.org/content/mimiciii/1.4/
5. Xu, D., Yuan, S., Zhang, L. and Wu, X., 2018, December. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 570-575). IEEE.

Data utility analysis

	Real data		Heal. GAN		Fair GAN		Ours	
	acc.	F1	acc.	F1	acc.	F1	acc.	F1
(a)	92.1	37.6	91.3	34.2	89.5	32.4	**91.5**	**34.7**
	91.8	12.1	**91.1**	**12.0**	88.3	11.6	90.9	11.7
(b)	71.2	59.9	**69.4**	**58.3**	67.1	56.3	68.1	57.3
	72.6	59	67.2	**59.1**	66.4	57.9	**68.9**	57.6
(c)	90.1	39.6	89.1	37	85.4	32.8	**89.6**	**39.9**
	89.3	17.9	88.3	15.8	86.3	14.3	**90**	**18.1**
(d)	89.9	7.0	87.9	**8.5**	86.1	4.3	**88.4**	6.8
	87.6	1.4	**88.4**	2.1	85.9	0.8	87.3	**2.4**

Accuracy and F1 on various prediction tasks with real data as reference point; (a) In-ICU mortality, (b) LOS > 3days,(c) In-hospital mortality, and (d)LOS > 7days. For each tasks, the first row denotes the predictions by LR and second row is the predictions by RF (higher is better for all the values).

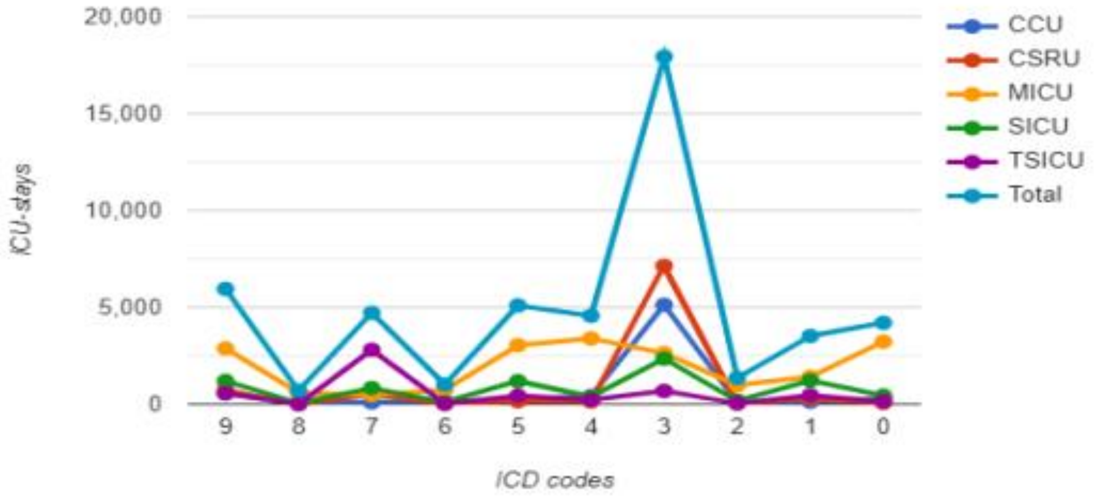


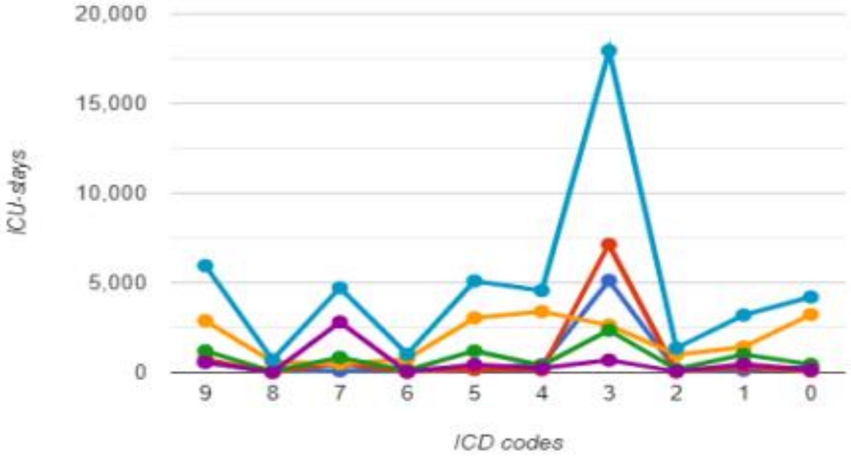
Downstream Fairness Analysis

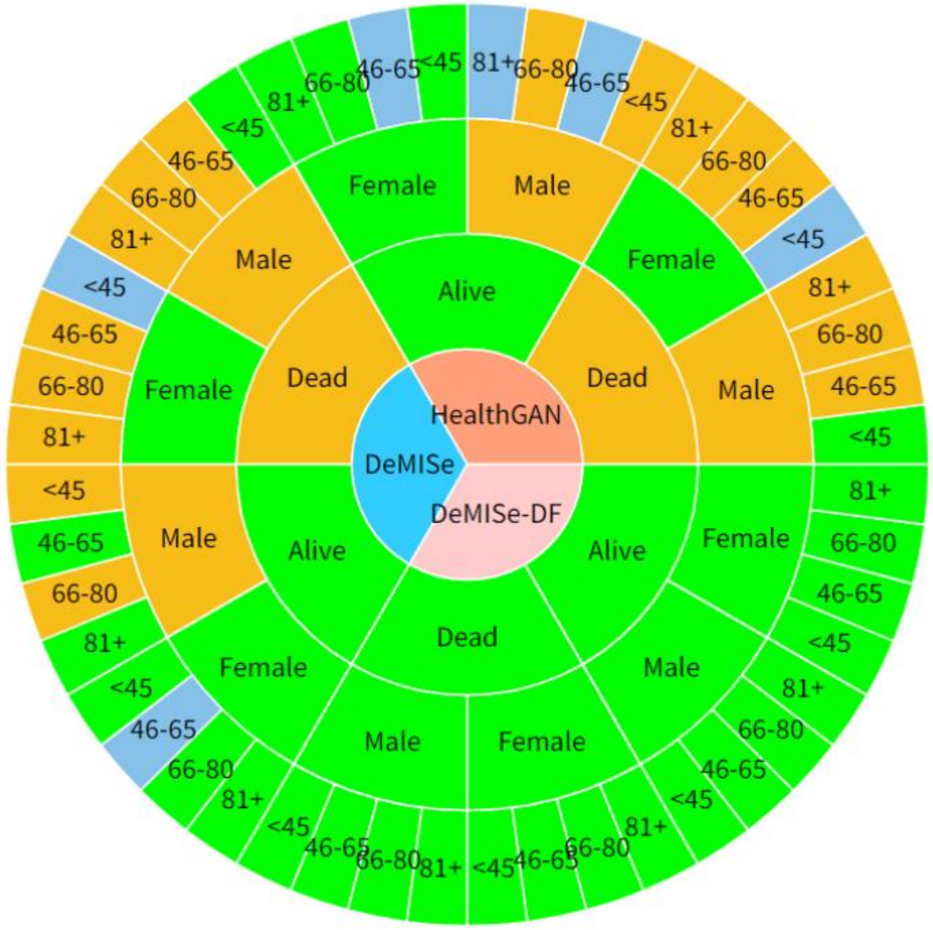
Metrics	Prediction	Real Data	HealthGAN	FairGAN	DeMISe
AUROC gap	In-hospital mortality	0.043 ± 0.001,	0.082 ± 0.002	0.021 ± 0.001	**0.001 ± 0.001**
	In-ICU mortality	0.03 ± 0.007	0.15 ± 0.035	0.023 ± 0.064	**0.012 ± 0.021**
	LOS>3days	-0.003 ± 0.002	-0.104 ± 0.001	-0.003 ± 0.001	**0.000 ± 0.001**
	LOS>7days	-0.005 ± 0.002	-0.076 ± 0.002	-0.061 ± .001	**-0.013 ± 0.001**
Parity gap	In-hospital mortality	-0.046 ± 0.018	-0.154 ± 0.010	-0.004 ± 0.014	**0.000 ± 0.001**
	In-ICU mortality	-0.031 ± 0.013	-0.331 ± 0.011	-0.005 ± 0.013	**0.000 ± 0.000**
	LOS>3days	0.022 ± 0.012	0.224 ± 0.012	0.022 ± 0.002	**0.000 ± 0.001**
	LOS>7days	-0.004 ± 0.002	-0.004 ± 0.002	**-0.002 ± 0.001**	-0.003 ± 0.001

The fairness gaps between White and Black patients across the different health care tasks, and models. Positive values represent a bias towards the white patients and negative values represent a bias towards the Black patients. The models are fair as the metric moves towards zero. The models are more unfair as the metric moves away from zero.

Fair Resemblance Analysis



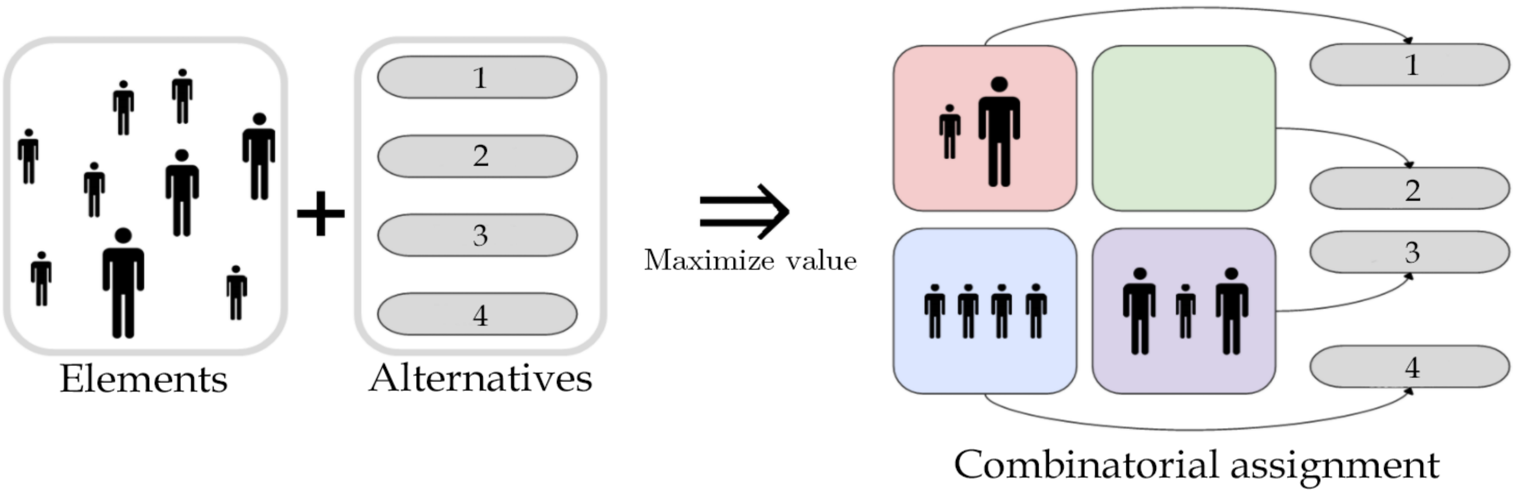




The colors orange, blue and green respectively indicate under-represented, over-represented and adequately represented sub-groups

Dividing the Indivisible to Maximize Value

We consider *combinatorial assignment*—the class of problems in which indivisible elements are partitioned into bundles among alternatives to maximize some notion of value (e.g., social welfare, expected utility).



The Combinatorial Assignment Problem

THE COMBINATORIAL ASSIGNMENT PROBLEM

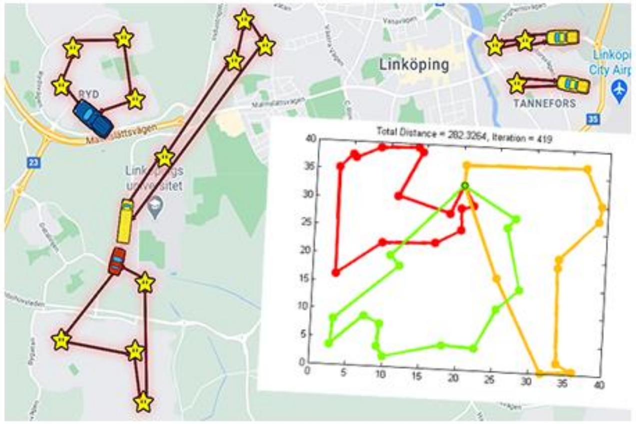
Input: A set of n items N, a set of m alternatives M, and a function (called the *social welfare function*) $\Phi : \Pi_N^m \to \mathbb{R}$.
Output: A combinatorial assignment $C = \langle B_1, \ldots, B_m \rangle$ over N that maximizes its *social welfare* $\Phi(C)$.

Important welfare function examples:

- *Utilitarian*: $\Phi(\langle B_1, \ldots, B_m \rangle) = \sum_{i \in M} v(B_i, i)$.
  Ex: Team formation; combinatorial auctions.
- *Egalitarian*: $\Phi(\langle B_1, \ldots, B_m \rangle) = \min_{i \in M} v(B_i, i)$.
  Ex: Minimize VRP makespan; "equitable" resource division.

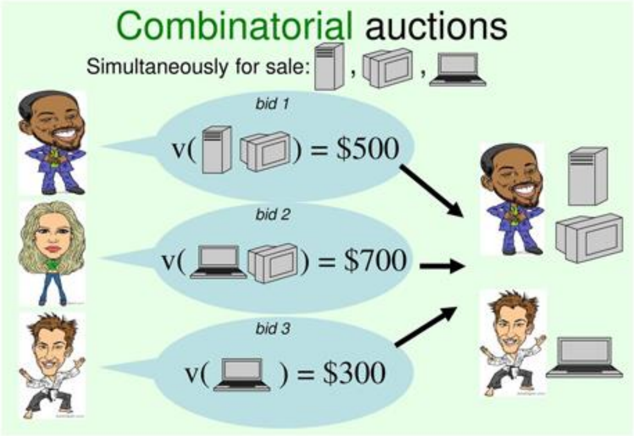
Both are unfortunately APX-hard (hard to approximate within a constant factor) and the input size is exponential with respect to n.

Assigning workers to jobs

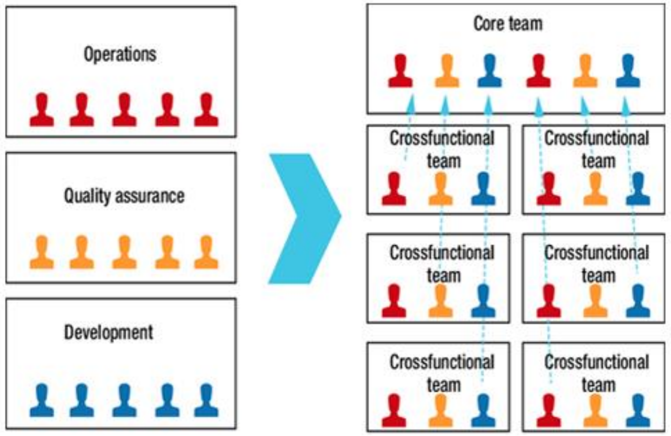


Multi-vehicle routing
(e.g., multiple TSP)

Multi-sensor
multi-target tracking



Combinatorial auctions



Team formation

Digit	Description
001-099	Service courses for nontechnical majors
100-199	Other service courses, basic undergraduate
200-299	Advanced undergraduate/beginning graduate
300-399	Advanced graduate
400-499	Experimental
500-599	Graduate seminars

Digit	Description
00-09	Introductory, miscellaneous
10-19	Hardware and Software Systems
20-39	Artificial Intelligence
40-49	Software Systems
50-59	Mathematical Foundations of Computing

Course allocation

Utilitarian Combinatorial Assignment - Conclusions

- Important UCA problems include coalition structure generation, the winner determination for combinatorial auctions, and generalized assignment.
- Utilitarian combinatorial assignment is APX-hard—but under certain restrictions, the problem is solvable and/or approximable in polynomial time.
  - Synergy hypergraphs is an expressive concise representation that admits faster algorithms for many cases (AAMAS2022).
- In all of the experiments so far, our best optimal algorithm (a hybrid) finds optimum in (worst-case) ≈ 1% of the time that the previous best method (IBM's CPLEX) needs.
  - It has been used successfully in a commercial setting (EU4).
- Machine learning can be used to generate heuristics that outperform conventional heuristics & Monte Carlo methods (AAMAS2022).

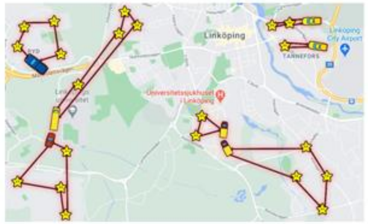
Select References

- Fredrik Präntare and Fredrik Heintz (2020). "An Anytime Algorithm for Optimal Simultaneous Coalition Structure Generation and Assignment". In: *JAAMAS*
- Fredrik Präntare and Fredrik Heintz (2020). "Hybrid Dynamic Programming for Optimal Simultaneous Coalition Structure Generation and Assignment". In: *PRIMA*
- Fredrik Präntare, Herman Appelgren, and Fredrik Heintz (2021). "Anytime Heuristic and Monte Carlo Methods for Large-Scale Simultaneous Coalition Structure Generation and Assignment". In: *AAAI*
- Fredrik Präntare, Mattias Tiger, David Bergström, Herman Appelgren, and Fredrik Heintz (2022). "Learning Heuristics for Combinatorial Assignment by Optimally Solving Subproblems". In: *AAMAS*
- Fredrik Präntare, Leif Eriksson, and George Osipov (2022). "Concise Representations and Complexity of Combinatorial Assignment Problems". In: *AAMAS*

1. Analyze hardness

2. Optimal algorithms

3. Non-exact algorithms



4. Real-world applications

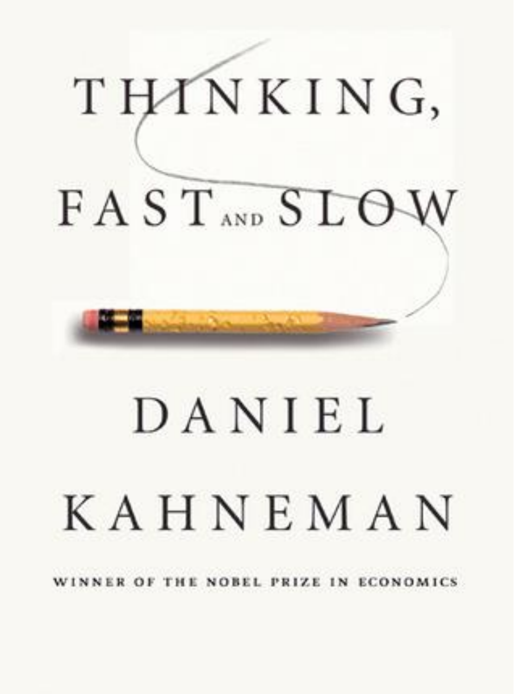
Research Overview

Learning generative models based on trajectory data

Probabilistic logical reasoning over observed and predicted trajectories

Utilitarian Combinatorial Assignment

The Way Forward



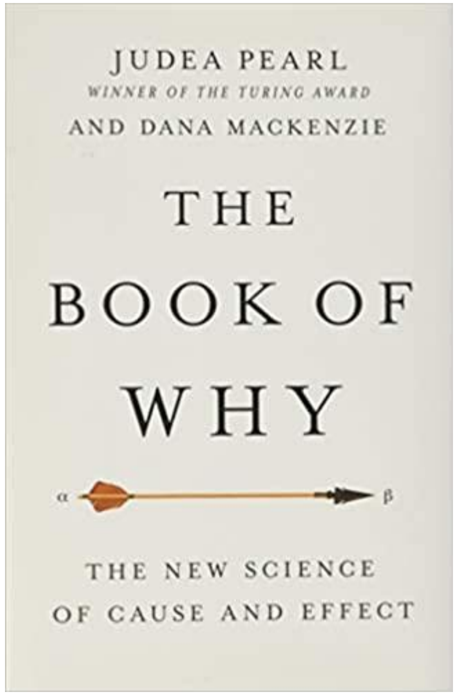
Data

Knowledge/
Assumptions

Causal Models
Combining reasoning
and learning

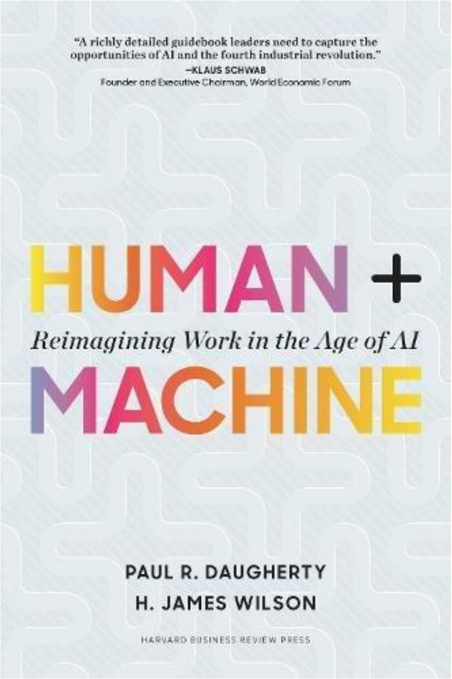
Explanations

Predictions



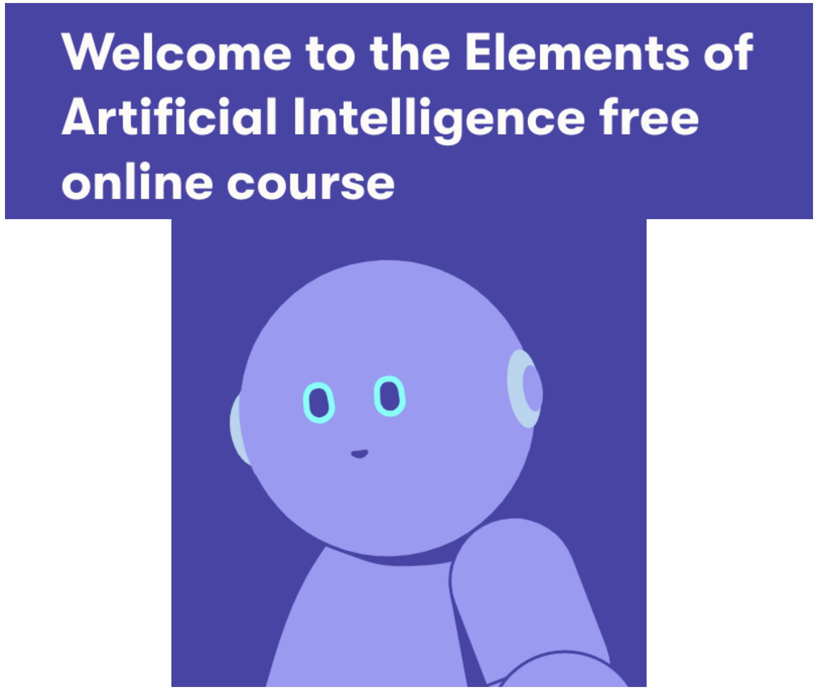
Other Components to Achieve Trustworthy AI

Humans + AI



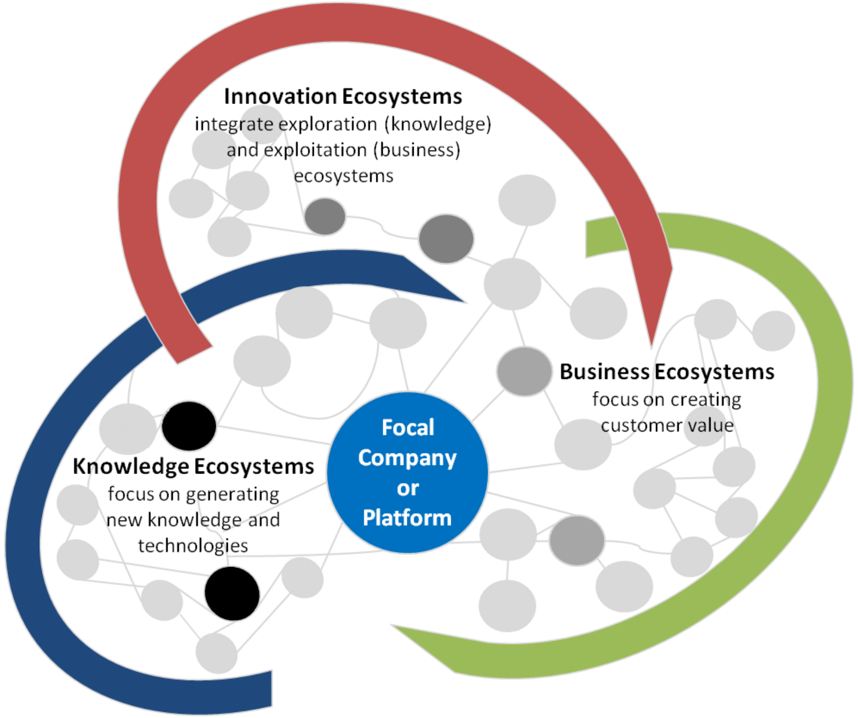
https://knowledge.wharton.upenn.edu/article/reimagining-work-age-ai/

Education



https://elementsofai.se

Ecosystems



https://timreview.ca/article/919

LINKÖPINGS UNIVERSITET

AI Innovation, Competence and Research Ecosystem

TAILOR

AI INNOVATION of Sweden

Elements of AI

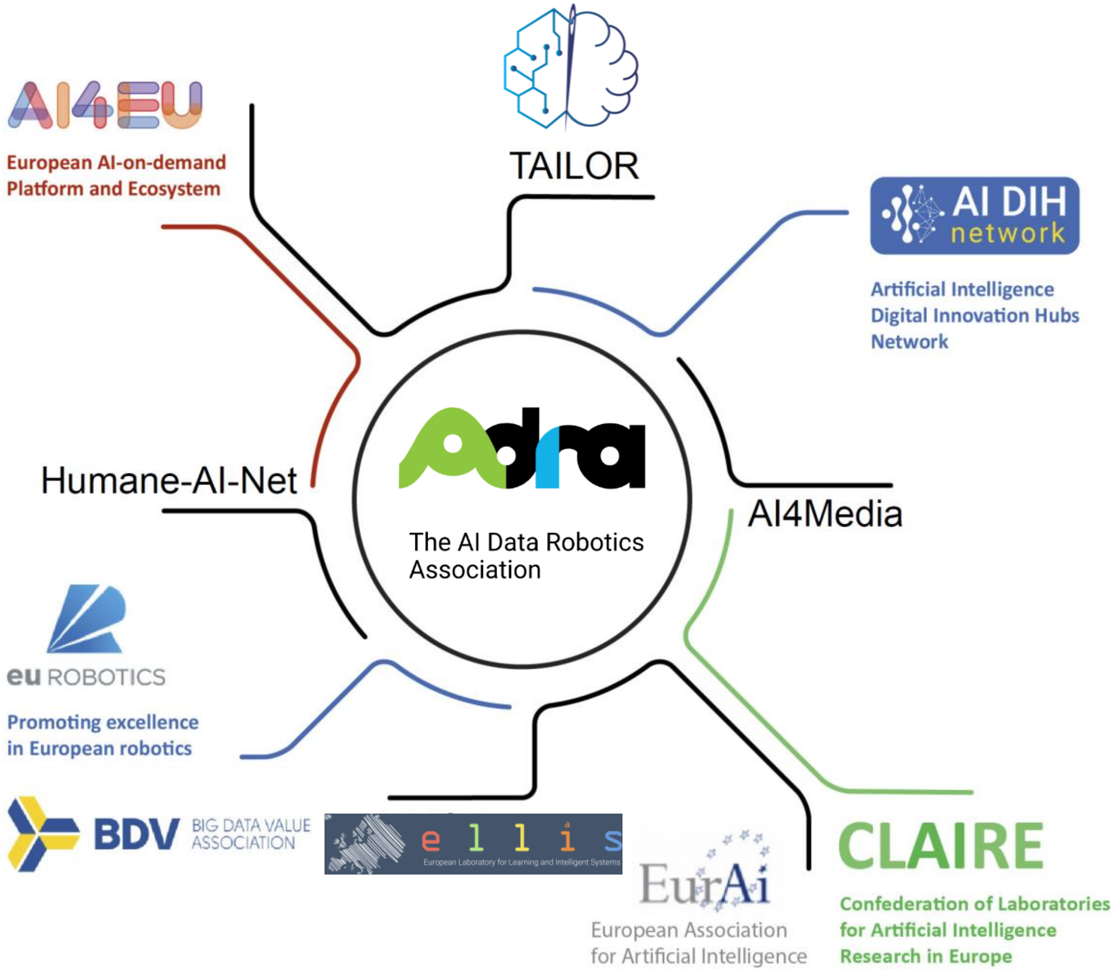
AI Competence of Sweden

WASP ED

WASP—HS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

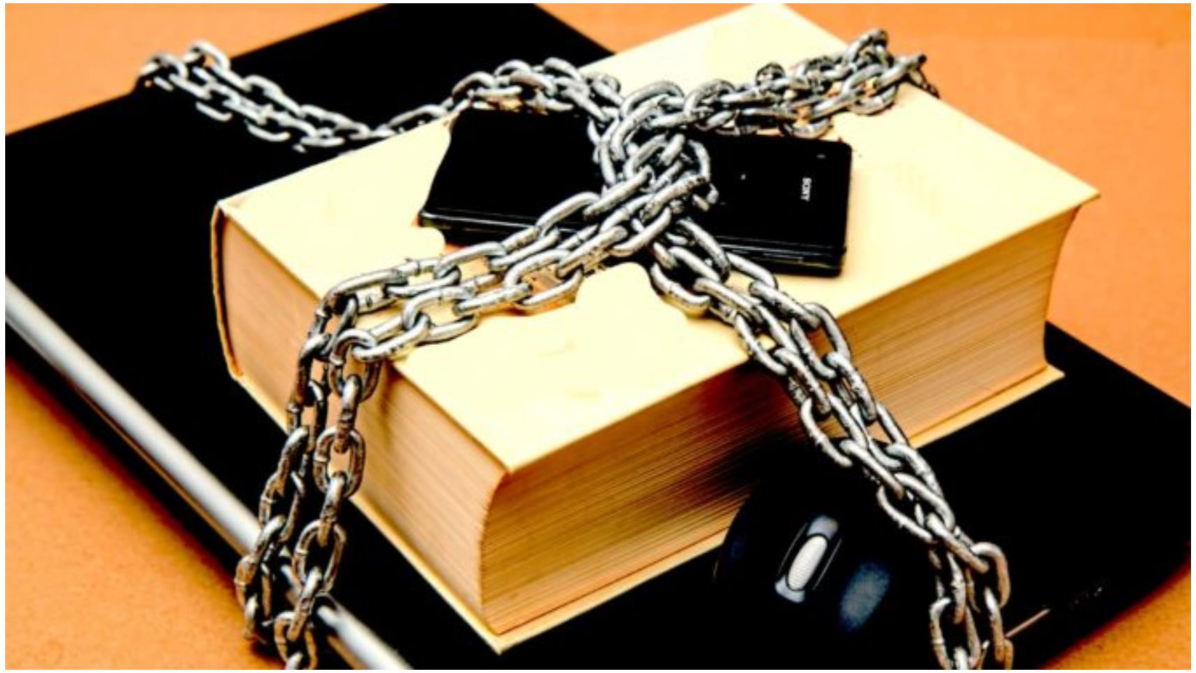
CHALMERS · KTH · LINKÖPINGS UNIVERSITET · LUNDS UNIVERSITET · UMEÅ UNIVERSITET · UPPSALA UNIVERSITET · ÖREBRO UNIVERSITY



External Analysis of Human Decision Making

France Bans Judge Analytics, 5 Years In Prison For Rule Breakers

4th June 2019 artificiallawyer Litigation Prediction 52



https://www.artificiallawyer.com/2019/06/04/france-bans-judge-analytics-5-years-in-prison-for-rule-breakers/

“Weak human + machine + superior process was greater than a strong computer and, remarkably, greater than a strong human + machine with inferior process.”

Garry Kasparov

AI and Humans – Together