Safety of AI Systems with Executable Causal Models and Statistical Data Science

Professor Yiannis Papadopoulos

University of Hull

y.i.papadopoulos@hull.ac.uk

University webpage

Personal Webpage

Video summary of work

Dependable Intelligent Systems Group Webpage

Circa 90 AD

St John writes the "Apocalypse" or book of "Revelation" in a cave in the island of Patmos.

In the book, the opening of four seals of a divine scroll releases the "4 Horsemen" bringing conquest, civil war, famine and death upon earth

AI as an existential Risk

Nick Bostrom pictures humanity as extracting balls (technologies) from a giant urn.

One may devastate humanity. Could it be AI? Superintelligence?

There is a grand technology challenge for Dependable AI and Intelligent Systems

Implications for industry and society are enormous

Dependable Intelligent Systems (DEIS) Group

- World class research on complex software and systems, including intelligent systems targeting Dependability, including of AI
- Pioneering novel techniques and state-of-the-art tools with academic and commercial impact
- In the UK our Research impact is officially ranked as internationally excellent. See recent REF2021 public impact case BIOLOGIC

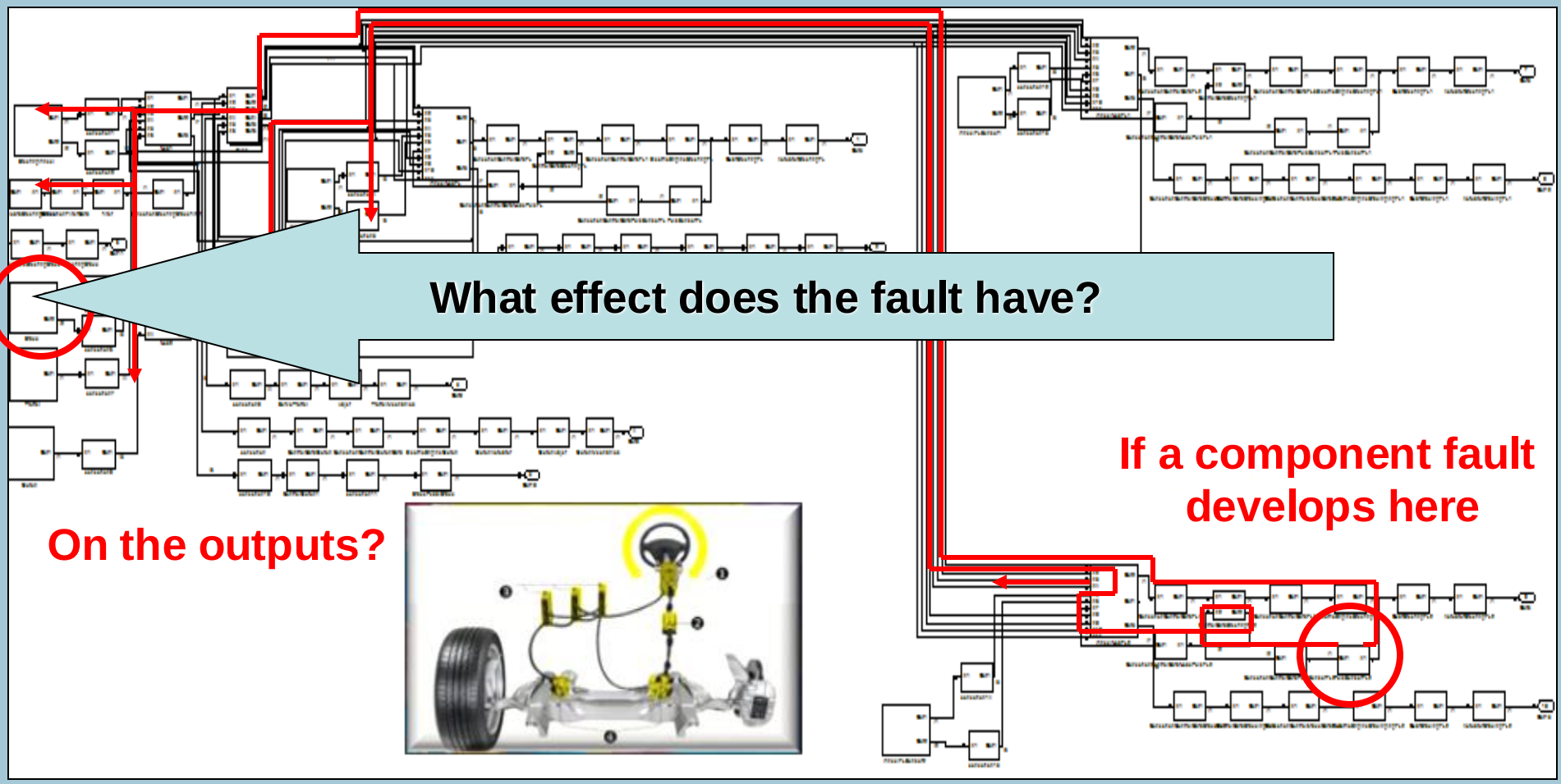
Context and Motivation of Research

- **Dependability: Safety, Reliability, Availability, Maintainability, Data Integrity, Security, Privacy**
- **Increasing concerns about new systems**
- **The 4 horsemen of the "Apocalypse"**
  - ***C*omplexity**
  - ***I*ntelligence**
  - ***A*utonomy**
  - ***Open Systems of Systems***
- ***CIAO* affects emerging systems, including multi-robot systems, cooperative swarms, machine learning in transport, health, manufacturing**

University of Hull

Challenge	Technology
Complexity	HiP-HOPS & Dymodia: • Model-Based Methods & Tools for automating dependability analysis and design of systems
Intelligence	SafeML, SMILE, SafeLLM: • SafeML: Dynamic estimation of confidence on accuracy of Machine Learning (ML), • SMILE: Explainability of ML • SafeLLM: Detection of Hallucinations, Filtering of Unsafe Response
Autonomy Openness	EDDIs (Executable Digital Dependability Identities) • Executable Model-Based Safety Monitors for runtime safety assurance & adaptation of SoS

CIAO: The Challenge of Complexity

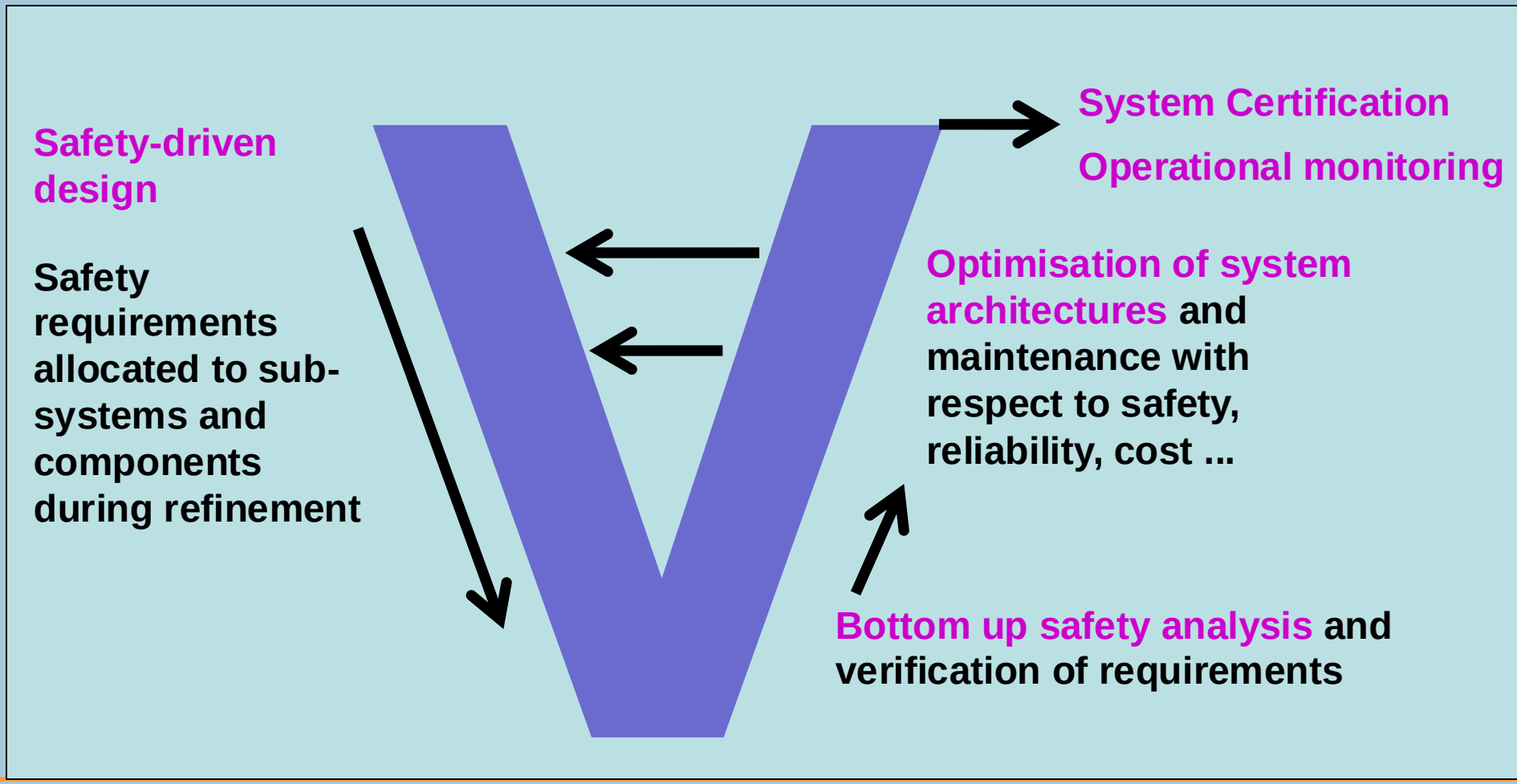


We develop:

- A Model-Based method and tools that simplify dependability analysis and optimisation of systems by partly automating the process
- Known as Hierarchically Performed - Hazard Origin and Propagation Studies (HiP-HOPS)

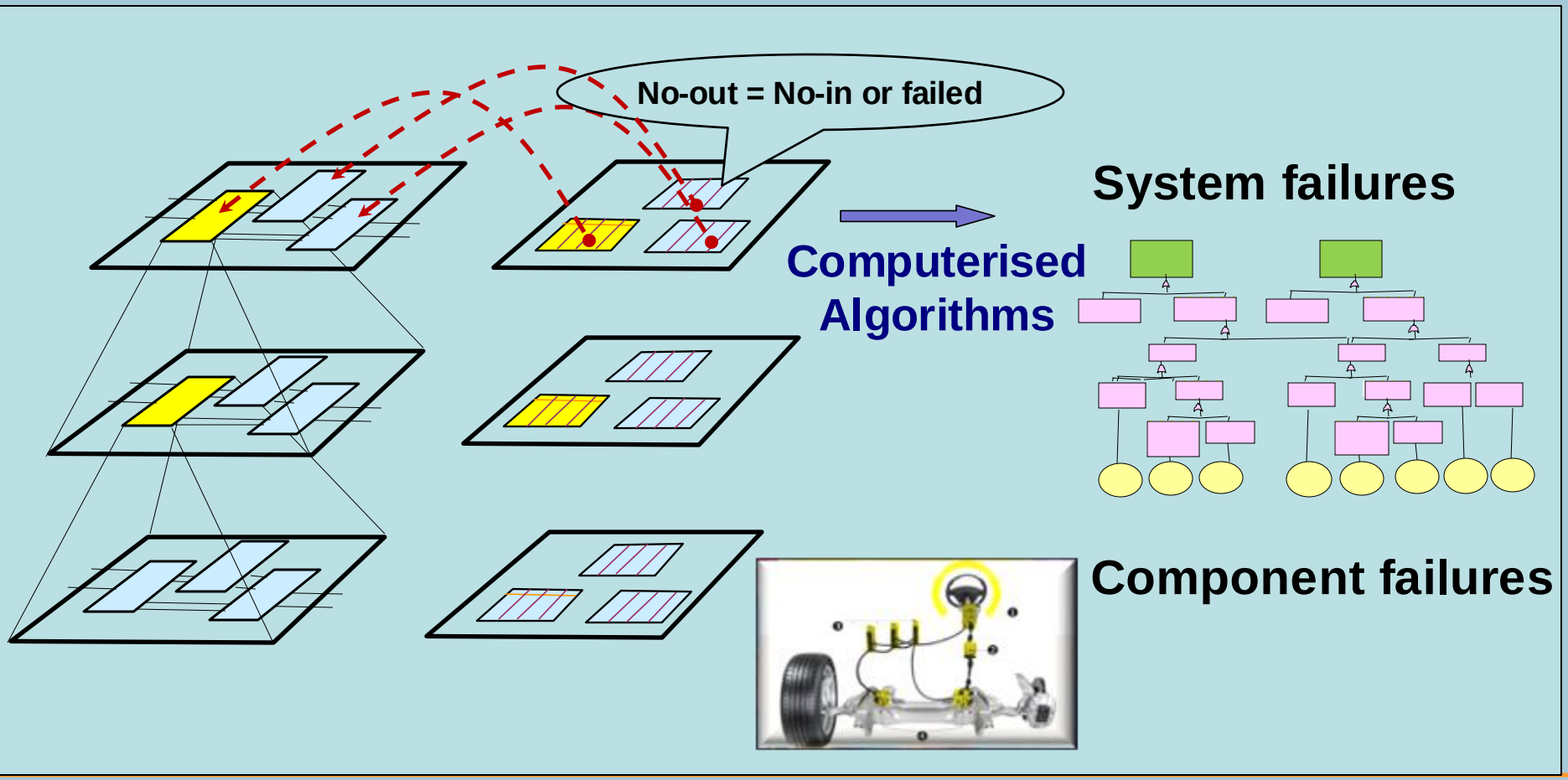
Scope and achievements of HiP-HOPS

Span the lifecycle of Systems Engineering



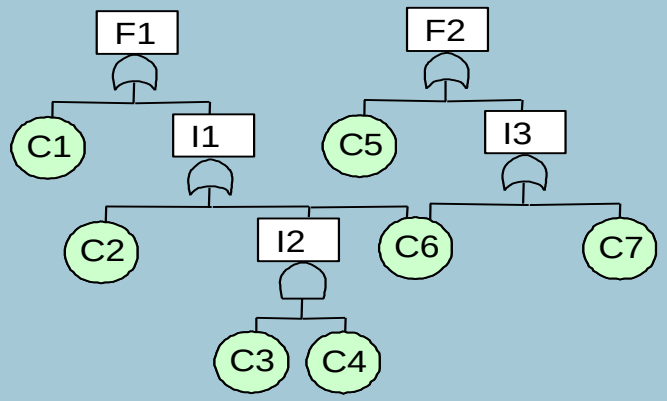
HiP-HOPS: Dependability Analysis

System Model + Failure annotations => Global view of failure:



Auto-synthesis of Fault-Trees and then FMEAs

Set of Interconnected System Fault Trees



Logical reduction

Equivalent FMEA

	Component failure	Direct effects on the system	Effects caused in conjunction with (other events)
A	C1	F1	-
	C2	F1	-
B	C3	-	F1 (C4)
	C4	-	F1 (C3)
	C5	F2	-
C	C6	F1, F2	-
	C7	F2	…

Components A, B and C – System failures F1 & F2

The Directed-Graph of Fault Trees can be reduced into an FMEA, a table of direct relationships between component and system failures

Moving beyond analysis to automatic improvement of dependability

- What if a design is found not safe enough?

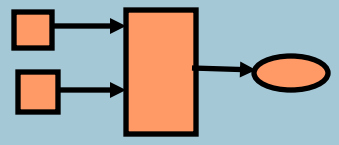
  How can it be improved?

  Substitute components & sub-systems, replicate increase frequency of maintenance

- And which solution is cost-optimal?

- Hard design problems that can only be addressed effectively with automation

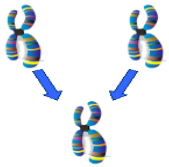
Evolutionary Design Optimisation Algorithms

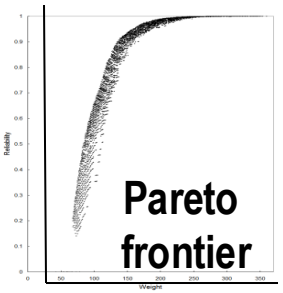


Model, Variants Cost, Weight Failure data

HiP-HOPS

Genetic Algorithm

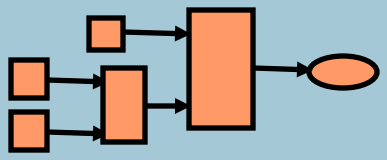




Cost analysis

Dependability analysis

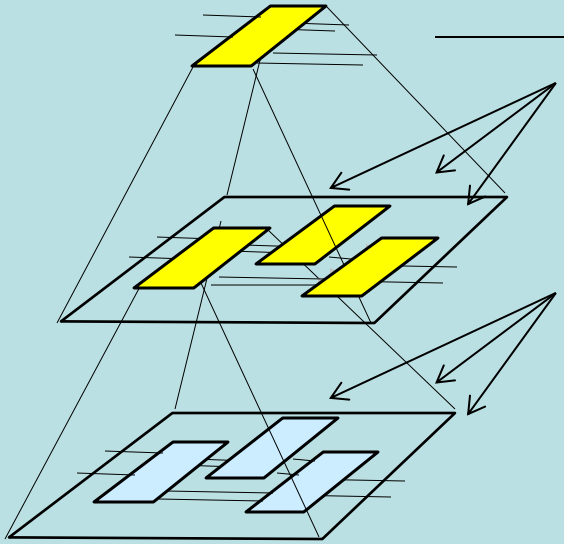
Weight analysis



Set of Optimised models

HiP-HOPS enables Allocation of Requirements

Cost-optimal automatic allocation of System Safety Requirements (SILs) is done using model-based analyses and AI metaheuristics



System Requirements
(derived via risk analysis)

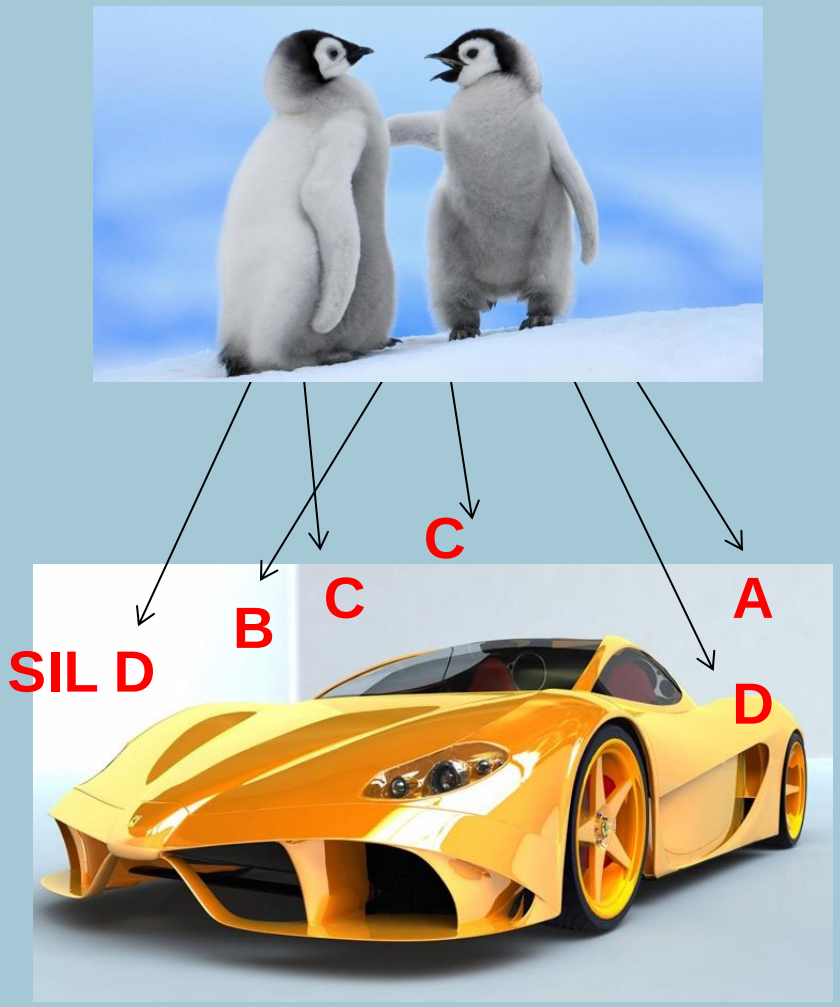
allocated/decomposed

Subsystem Requirements

allocated/decomposed

Component Requirements

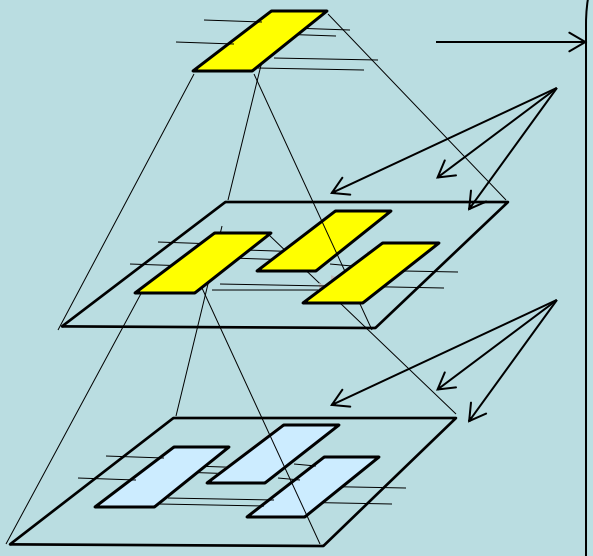
Penguins and safety of connected systems



PeSOA algorithm has been used together with HIP-HOPS for safety requirement allocation in cars.

[BBC ARTICLE]
[DAILY MAIL]
[EE Journal]
[Automotive IQ]
[BBC RADIO INTERVIEW]

Andromeda: a development of HiP-HOPS that creates Safety Cases for certification (experimental)



SC

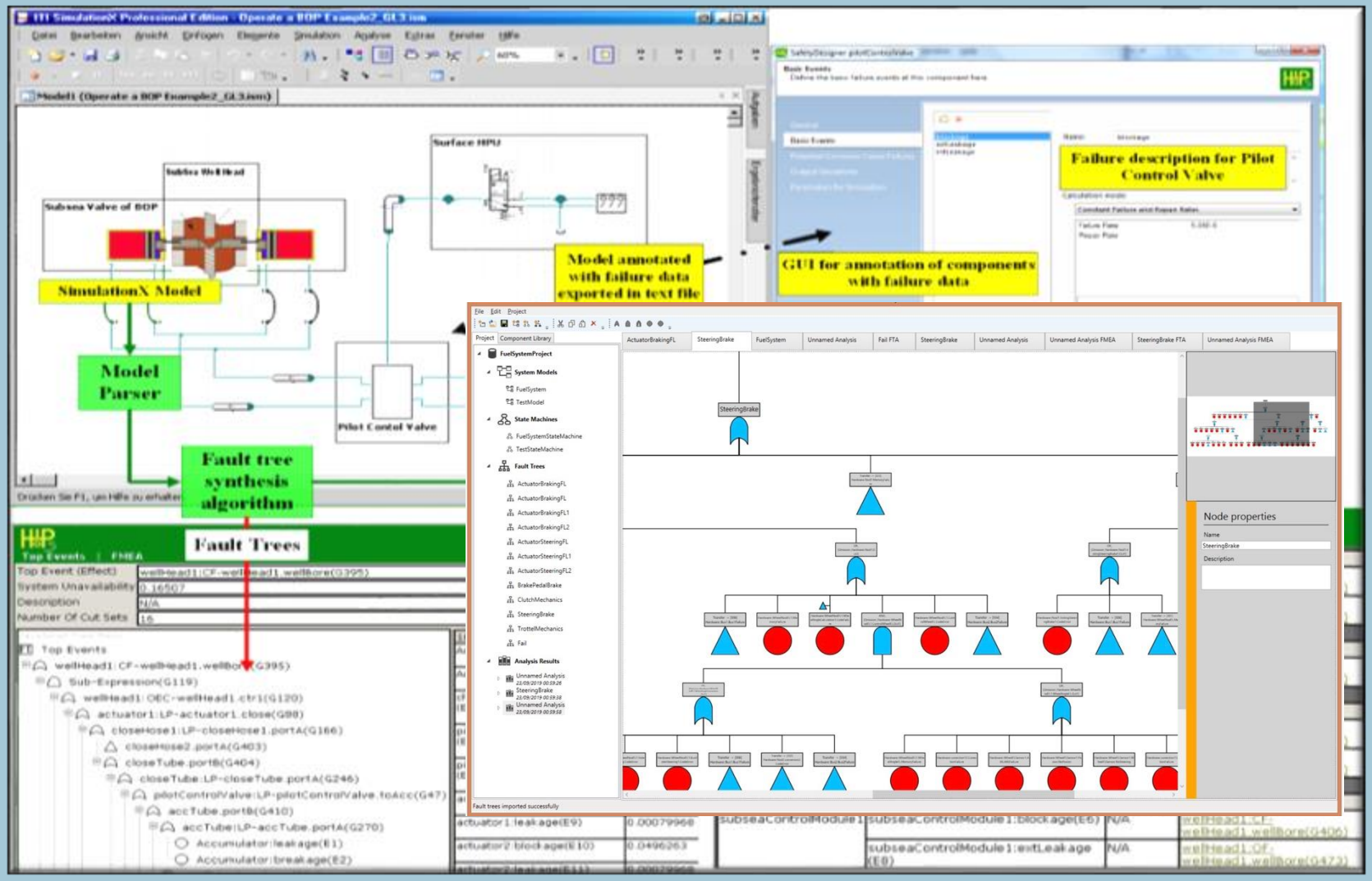
System Meets Safety Reqs derived through exhaustive risk analysis

Subsystems meet correctly allocated Integrity Requirements and all model assumptions are met.

Components meet correctly allocated Integrity Requirements and all model assumptions are met

Evidence for the above

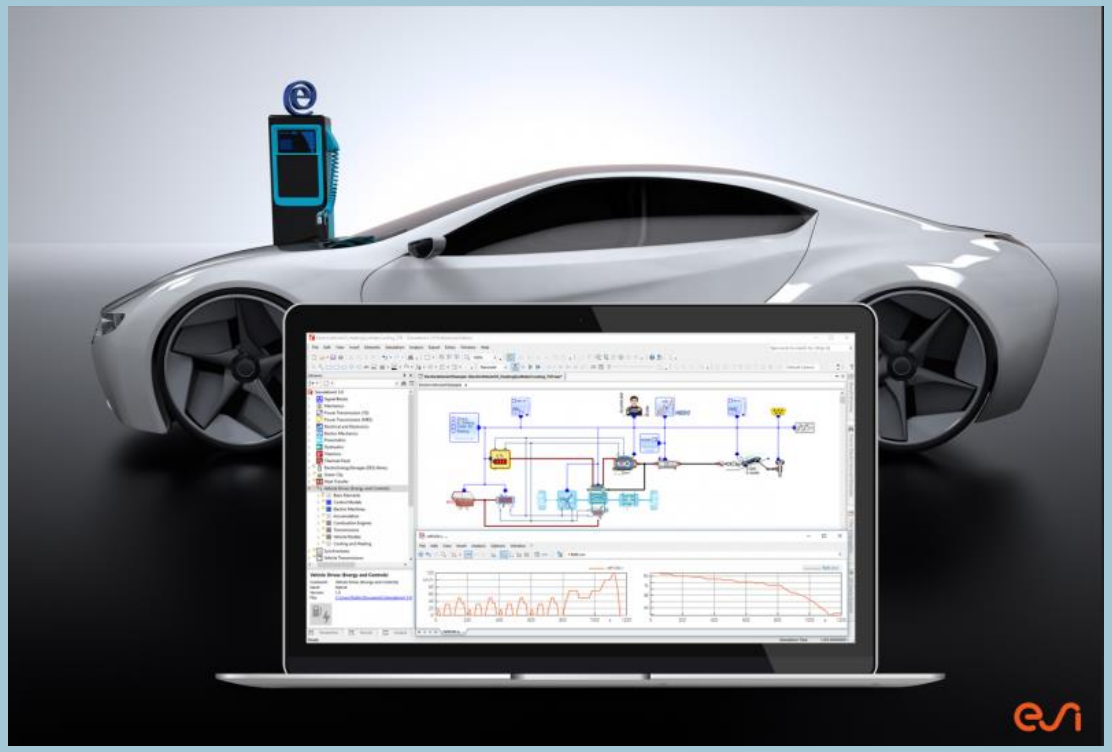
HIP-HOPS Tool is Commercial



Tools for Automating Analysis and Design of Dependable Systems

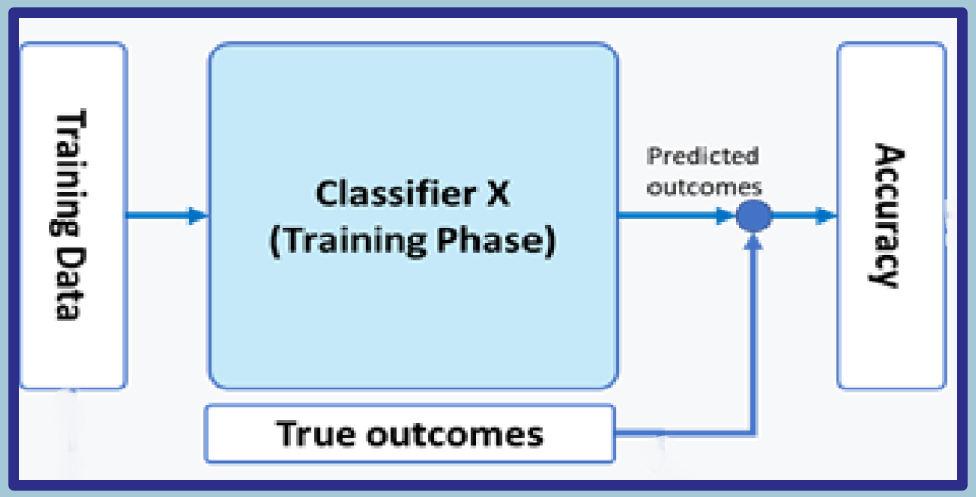
- HiP-HOPS by University of Hull
- Simulation X with ESI GmbH (Germany)
- EAST-ADL for automotive safety with Metacase (Finland)
- Dymodia, dynamic dependability analysis (non-commercial)

Technology transfer with global reach



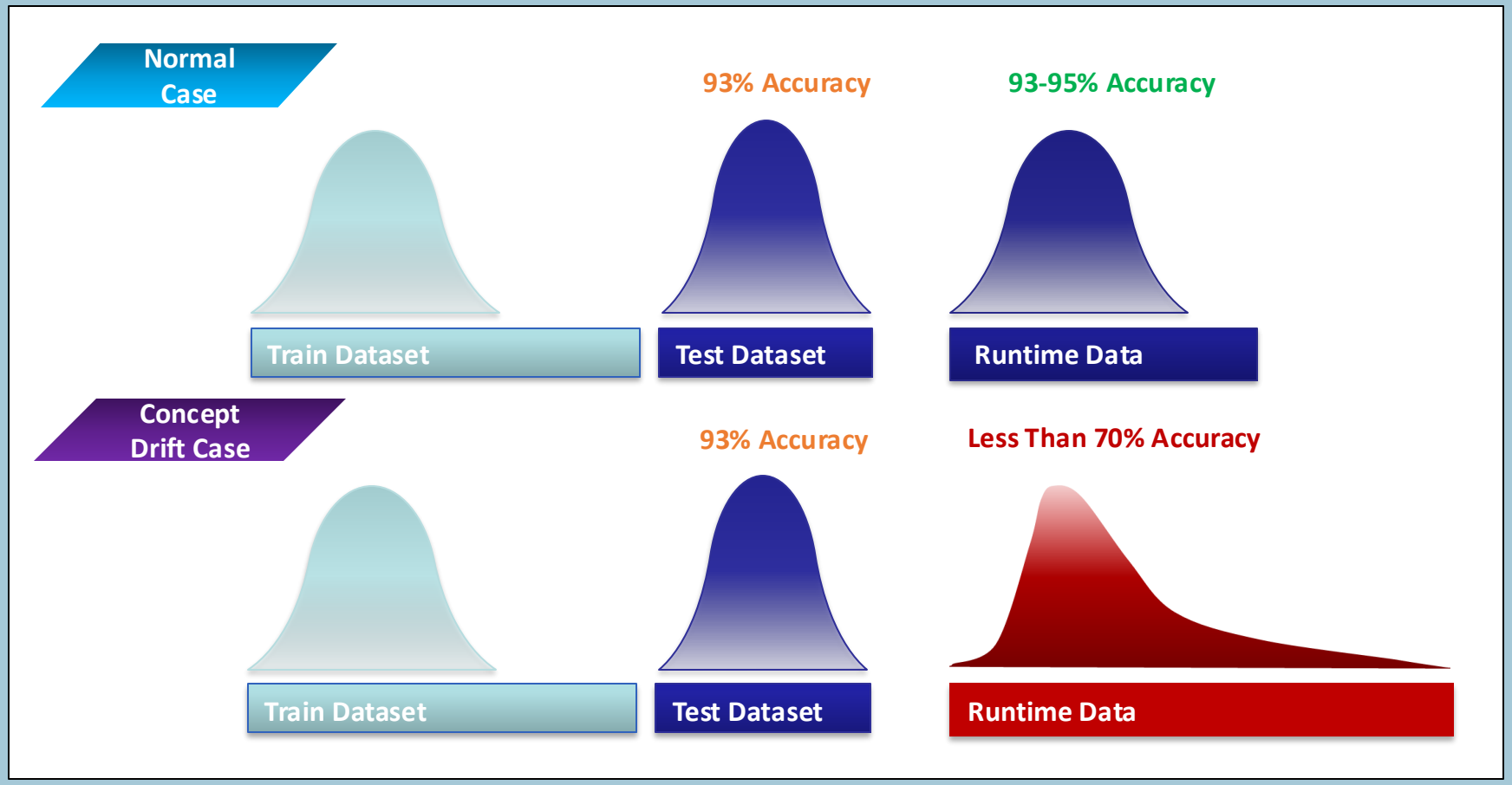
Clients include Honda, Toyota, Huawei, Honeywell Volvo, Continental, Fiat, Embraer

CIAO: The challenge of Intelligence



- Machine learning (ML) classifiers are often trained to detect objects (traffic signs) or events (process failure).
- Errors (undetected, wrongly detected) may have safety implications.
- ML has a predicted accuracy, we can calculate it if we know true classifications in verified datasets
- But this accuracy is a function of the training dataset and may drift for data outside this set

Machine Learning accuracy may drift

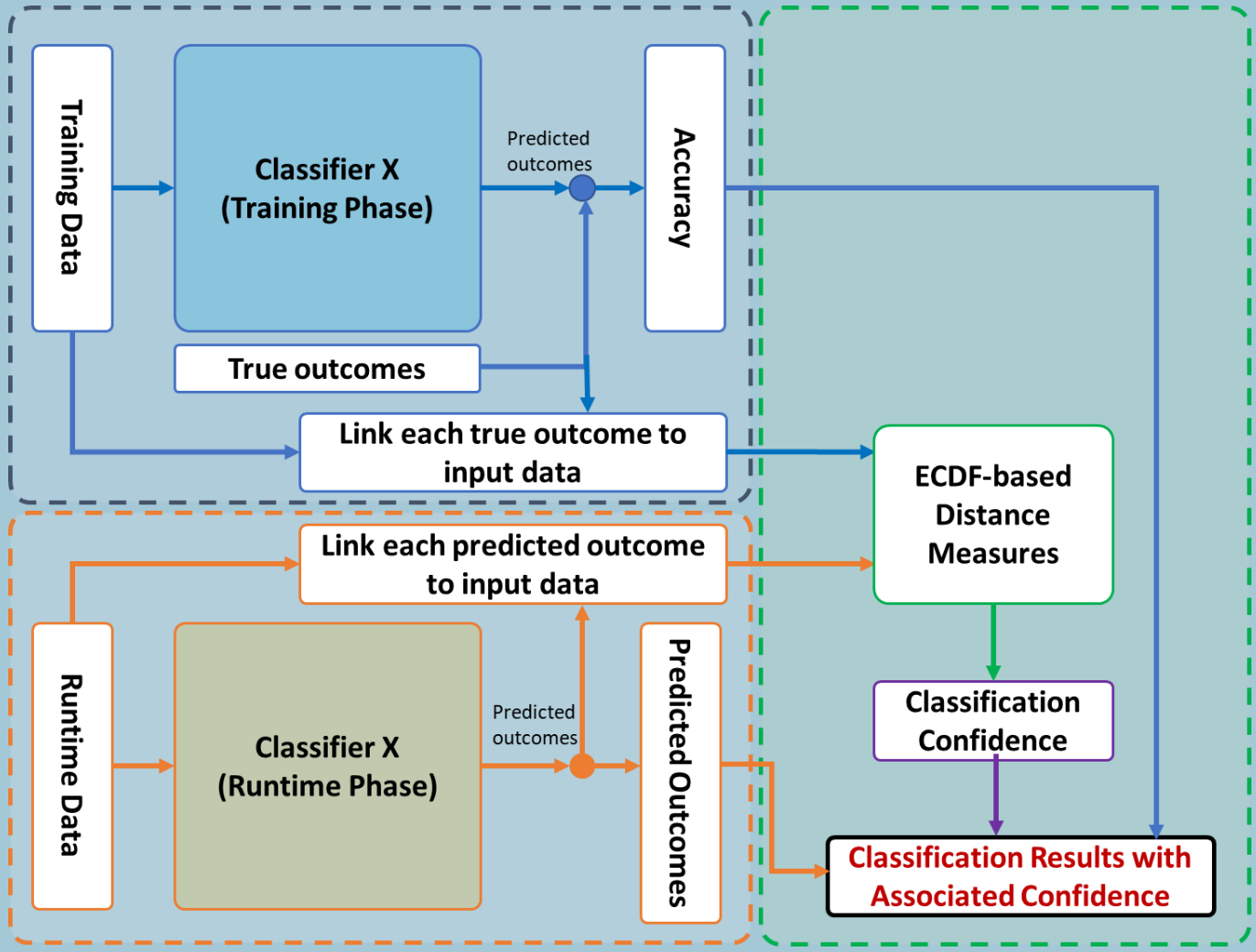


In operation accuracy may be worse if the data is statistically different from training dataset. Problem is we can't measure this shifting accuracy.

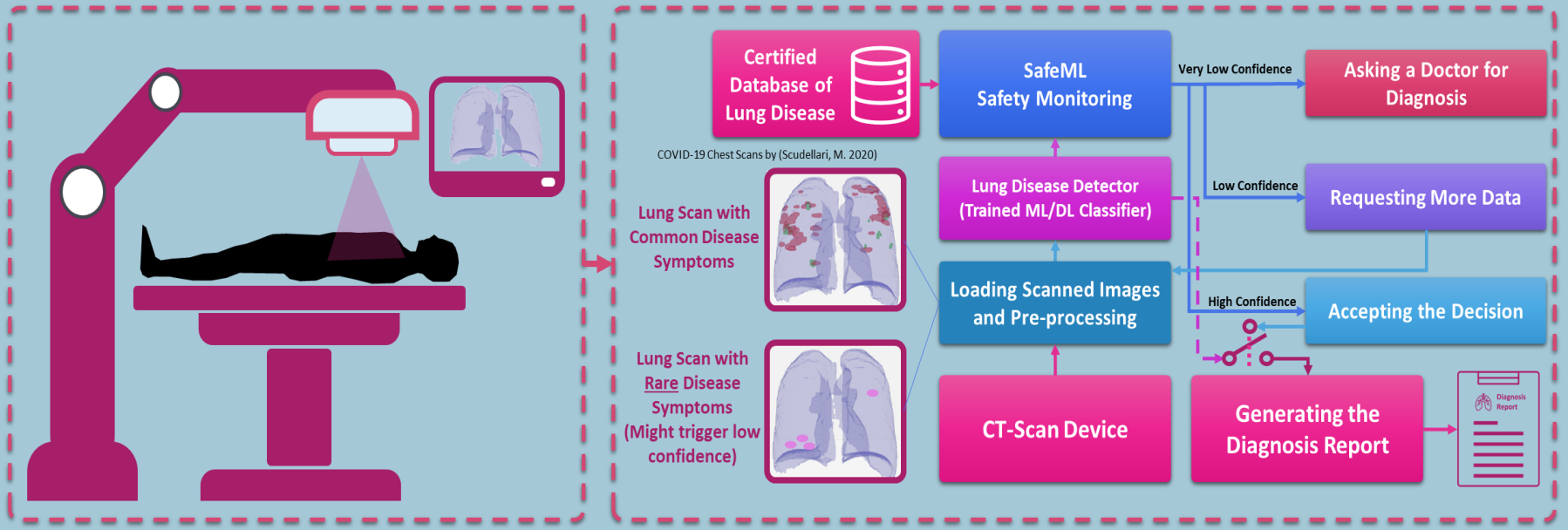
SafeML: Safety of Machine Learning

SafeML uses statistical techniques to measure the drift in input data linked to each reasoning outcome

It establishes a measure of confidence in the accuracy of ML results



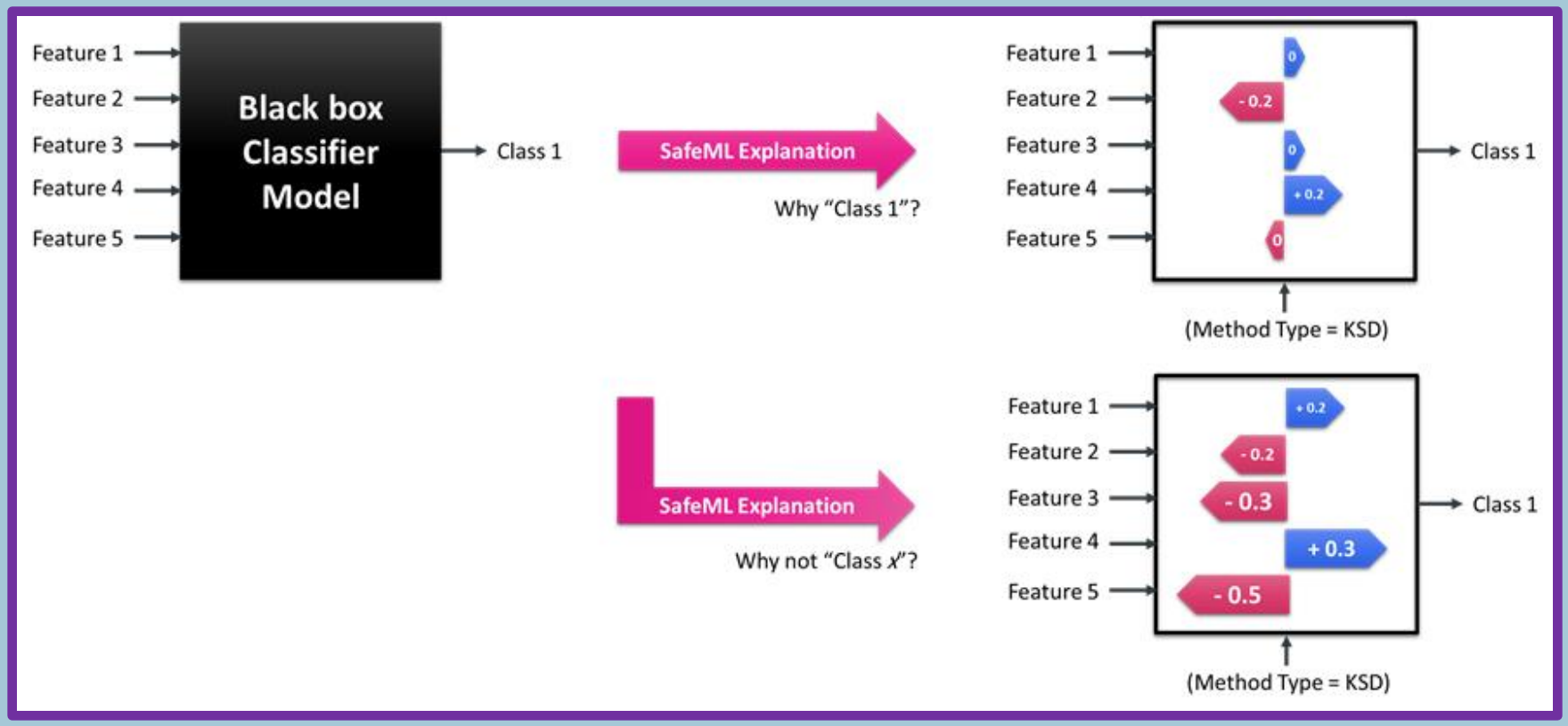
SafeML:
Example application on Lung Cancer Diagnosis



- SafeML cited in new German Industry Standard for Machine Learning Uncertainty Quantification (DIN SPEC 92005)

- Kuniko Paxton, one of our PhD students has received an Alan Turing Institute award for her work on Safer Cancer Detection

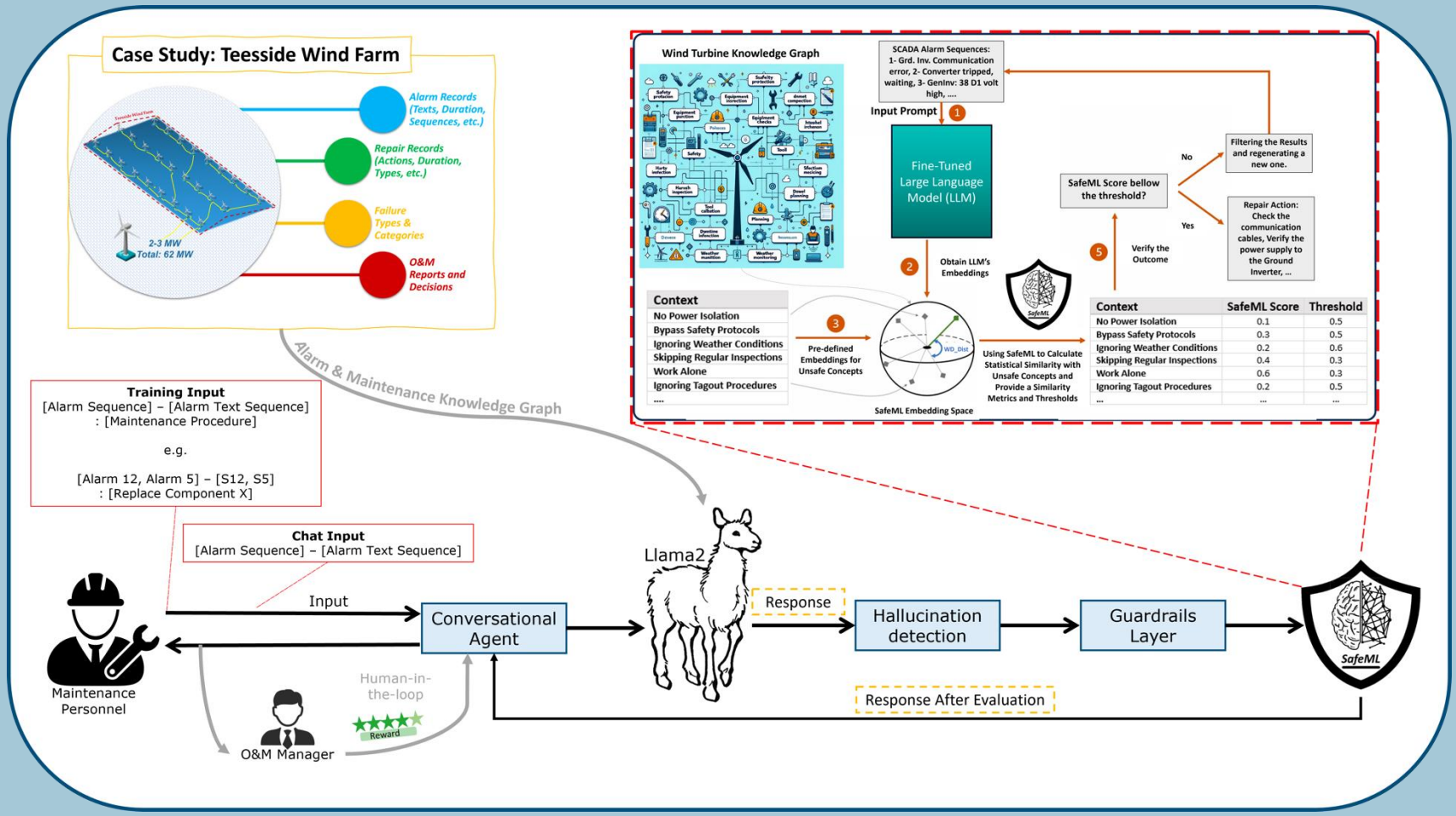
SMILE: Statistical Model-agnostic Interpretability with Local Explanations



- Helps to explain why ML reached the decision it did - what attributes of the input were most important
- Uses ECDF distance measures

SafeLLM: Safety of LLM (Generative AI)

Example application on Wind Turbine Maintenance with EDF



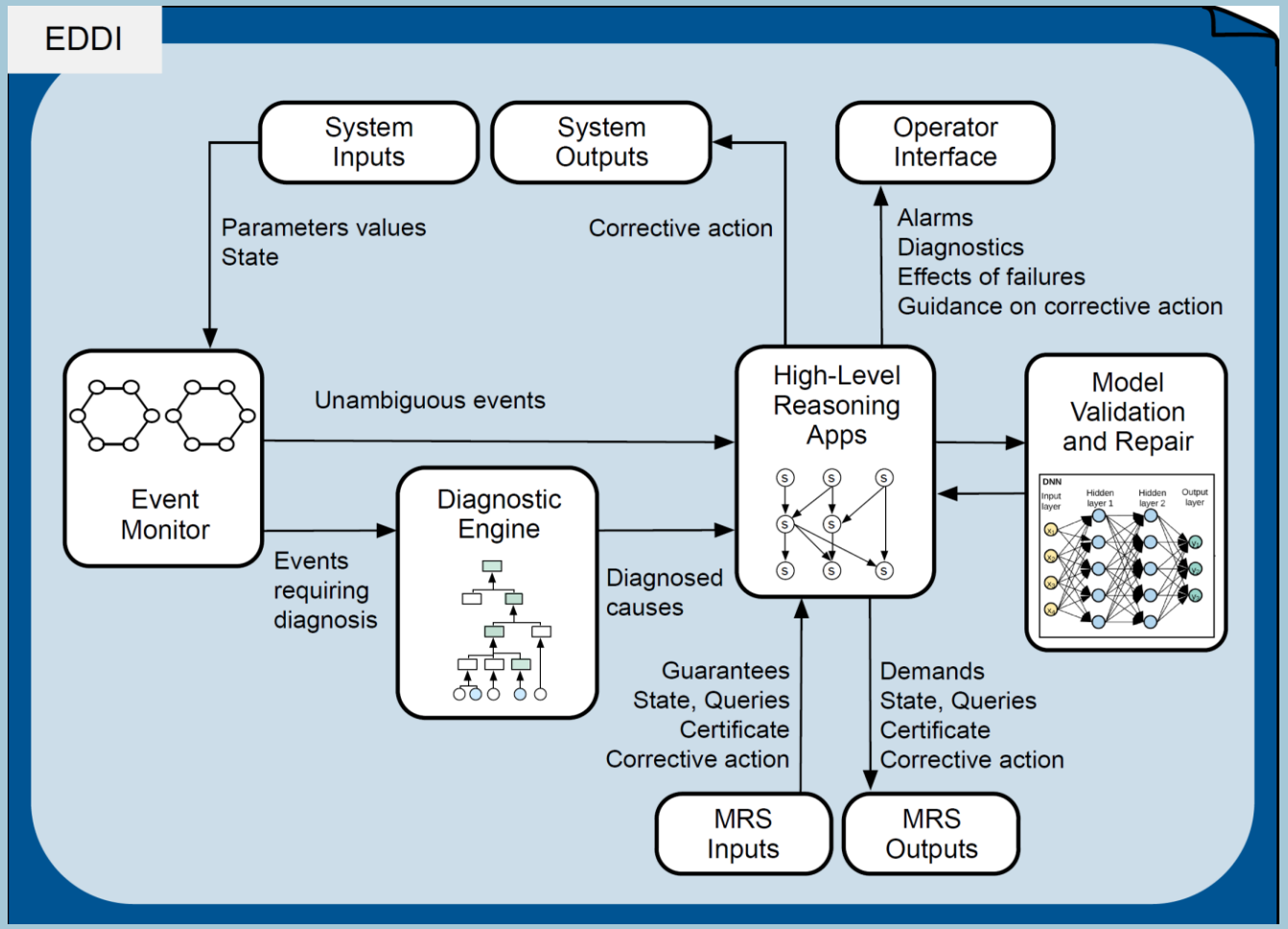
CIAO: The challenges of Autonomy and Open SoS

- **Dimensions:**
  - No operators when Autonomy fails
  - Openness -> infinity of configurations
  - Emergent behaviours & uncertainties
  - Higher security threats and implications for safety
  - Heterarchies and absence of hierarchy of control
- **Implication:** At least part of dependability assessment and certification needs to be moved at run-time

Executable Digital Dependability Identities

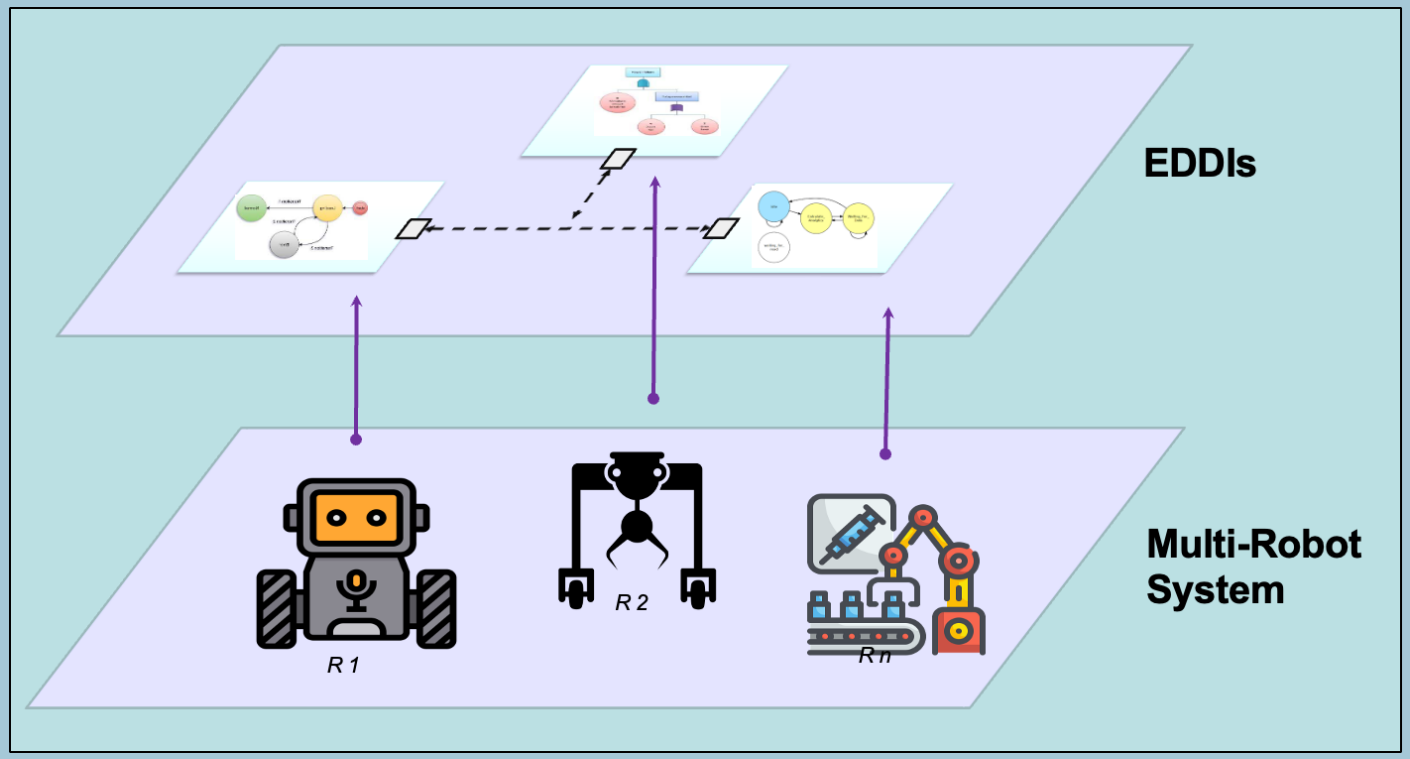
- **Modular, composable and executable specifications of dependability for components and systems in open systems of systems.**
- **EDDIs are models derived from safety assessment, Bayesian Nets, Fault Trees, Markov Models, Hierarchical State-Machines, State-sensitive Fault Trees, ConSerts**
- **Models expressed in ODE new Open Dependability Exchange metamodel developed in SESAME EU project**

Executable Digital Dependability Identities



- **High-level reasoning via executable Markov Model, Bayesian Nets, Hierarchical SMs, ConSerts …**

Executable Digital Dependability Identities



- **Safety Monitors** that sit on Systems and cooperate to dynamically guarantee Safety & Security of SoS
- Currently developing executable **Bayesian Nets** exploiting **belief values** returned **by SafeML**

Other related work

Challenge	Technology discussed today	Other significant work in the area
Complexity	HiP-HOPS: automated analysis and design	XSAP, COMPASS, Altarica, MBSA, various automated analyses
Intelligence	SafeML/SMILE/SafeLLM: confidence on accuracy of Machine Learning (ML)	Uncertainty wrappers, rigorous/formal ML development
Autonomy & Openness	EDDIs: Digital Dependability Identities (with Fraunhofer IESE)	ConSerts, Contracts, models@runtime, run-time safety cases

Possible Future of Technology

From intelligent design to things that learn and evolve.
This move is both exciting and problematic

(Source – Daniel Dennett in New Scientist)

AI as Pandora's Box

In the ancient Greek myth, Pandora opens a box gifted to her by Gods. All evils escape but she closes the box in time to keep "Hope"

A few papers

- **HIP-HOPS & MBSA latest:** Andromeda: A model-connected framework for safety assessment & assurance", Journal of Systems and Software, https://doi.org/10.1016/j.jss.2024.112256
- **SafeML:** Toward Improving Confidence in Autonomous Vehicle Software: A Study on Traffic Sign Recognition Systems, IEEE Computer, https://hull-repository.worktribe.com/preview/3757643/SafeML_II_Author_Version.pdf
- **SMILE:** Explaining Black Boxes With a SMILE: Statistical Model-Agnostic Interpretability With Local Explanations, IEEE Software, https://arxiv.org/abs/2311.07286
- **SafeLLM:** SafeLLM: Domain-Specific Safety Monitoring for Large Language Models: A Case Study of Offshore Wind Maintenance, Journal of Physics, https://arxiv.org/abs/2410.10852
- **EDDIs:** Computational intelligence for safety assurance of cooperative systems of systems, IEEE Computer, https://hull-repository.worktribe.com/preview/3728869/IEEE_2020_Kabir_Papadopoulos_DDIs_BN.pdf

Acknowledgements

Current and former members of the DEIS group and contributors to the ideas presented today:
Martin Walker, Koorosh Aslansefat, David Parker, Ioannis Sorokos, Luis Azevedo, Sohag Kabir, Connor Walker, Kuniko Paxton

Thank you for attending!

Questions and comments, post-talk communications most welcome