Crafting ML components in safety critical systems

Andrea Bondavalli

DIMAI - UNIFI
bondavalli@unifi.it

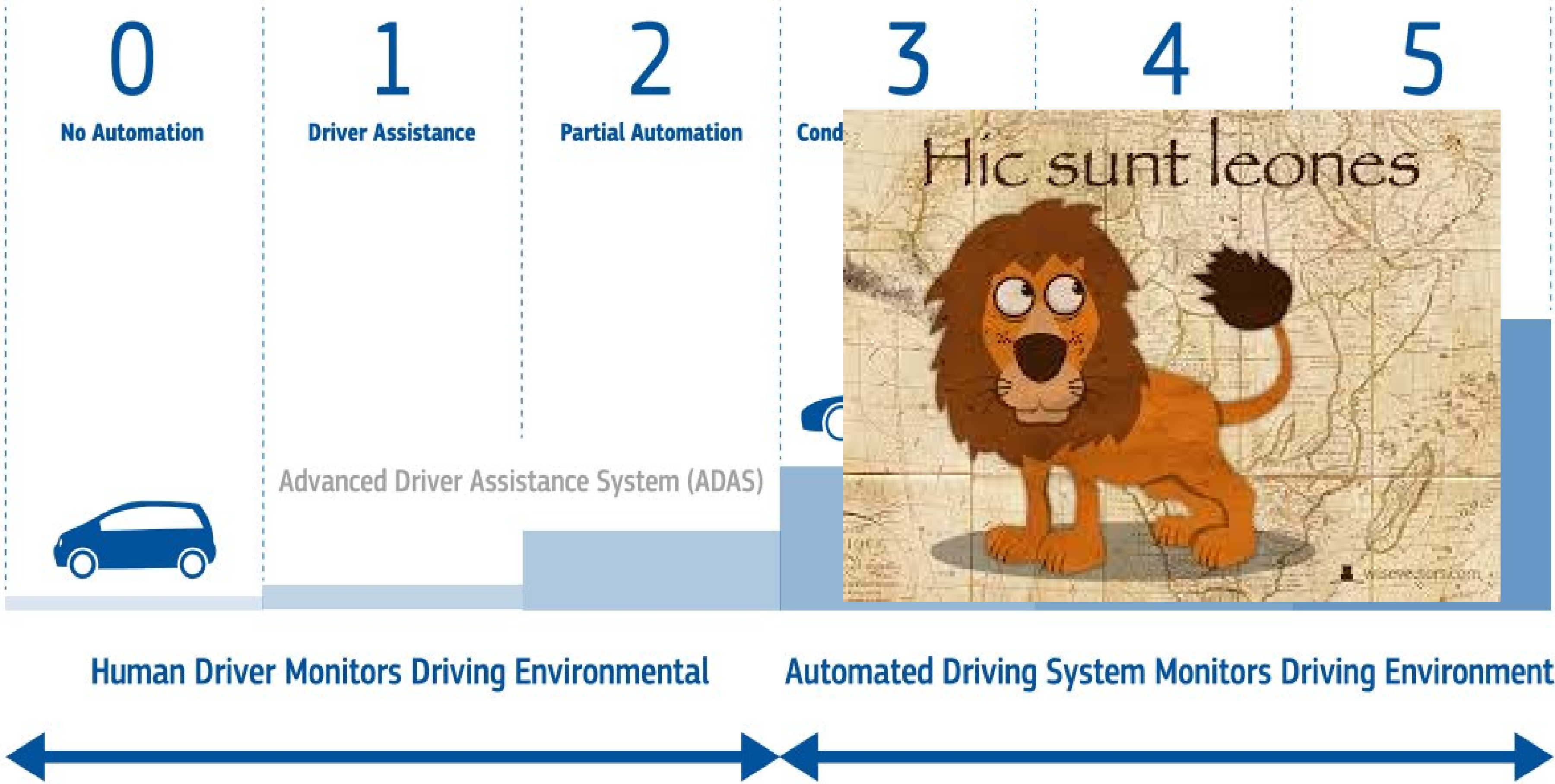
UNIVERSITÀ DEGLI STUDI FIRENZE
DIMAI
DIPARTIMENTO DI MATEMATICA E INFORMATICA "ULISSE DINI"

RCL
RESILIENT COMPUTING LAB

What is missing in mission-critical systems?

- So many things……..
- What I see in current mission critical (Cyber physical) systems is …
- more and more sophisticated functions
- in more and more unknown and unpredicatble environments….
- using technologies we do not master properly

An example: Autonomous driving...



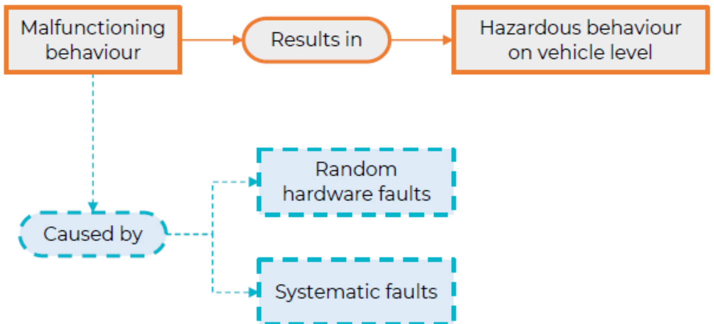
The challenge

- more and more sophisticated functions :
- Eg. AUTOMATED DRIVING

- in more and more unknown and unpredicatble environments….
- Automated driving system MONITORS environment

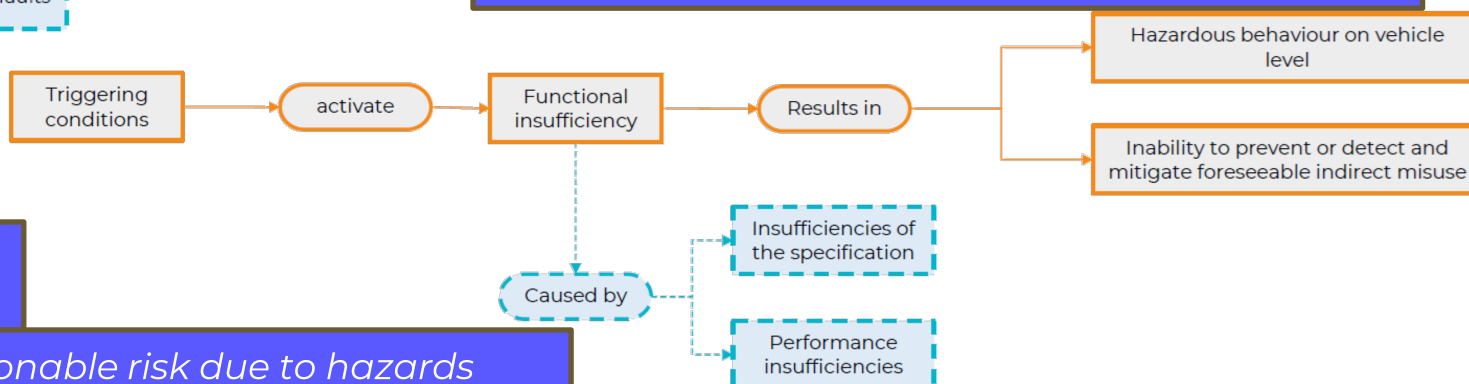
- using technologies we do not master properly (especially wrt safety and security)
- AI and ML primarily

An eye on Standards….
ISO 26262 and ISO 21448 Sotif



ISO 26262

*“Absence of unreasonable risk due to hazards caused by **malfunctioning behaviour** of the electrical and/or electronic systems»*



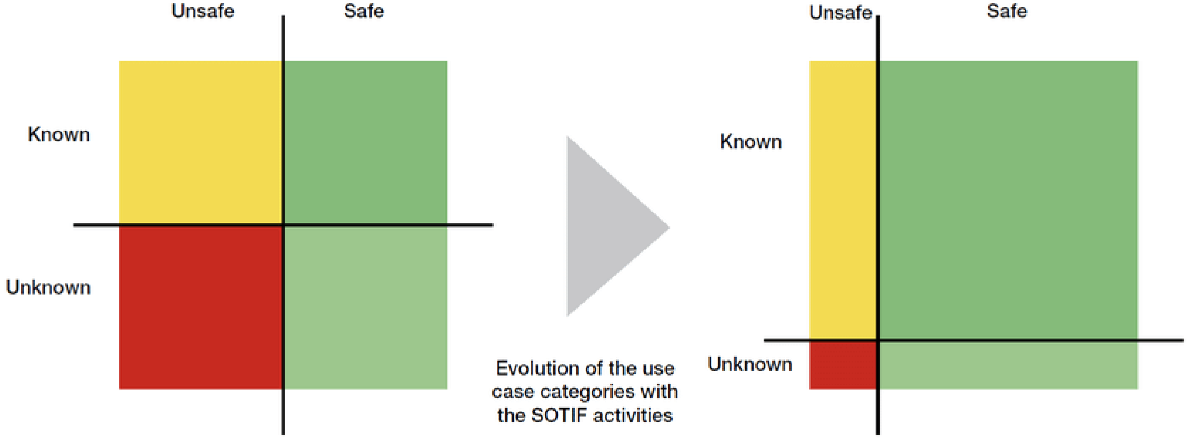
ISO 21448 - SOTIF

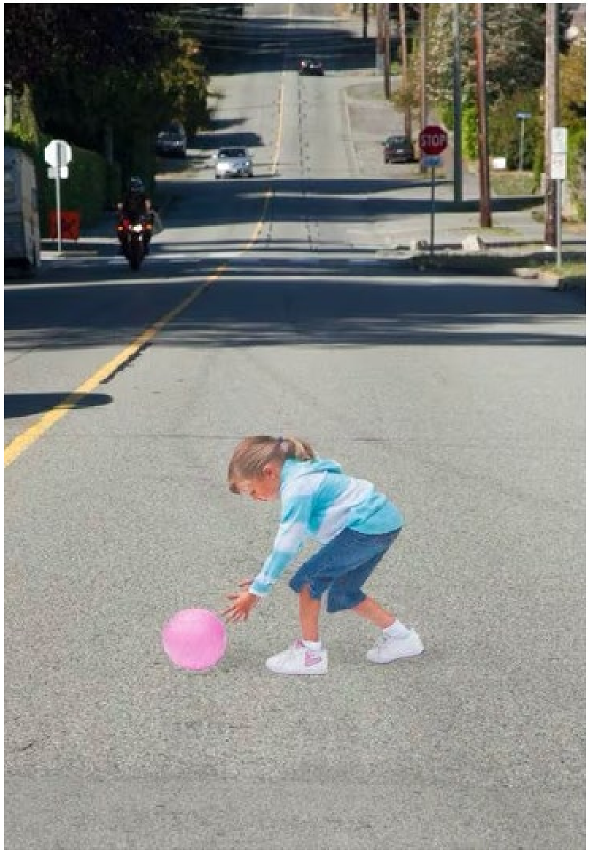
*“Absence of unreasonable risk due to hazards resulting from **functional insufficiencies** of the intended functionality or by reasonably **foreseeable misuse** by road users”*

Source: CFAA – University of York – Prof. Burton

SOTIF VIEW: Insufficiencies of Specification and Performance Insufficienceis

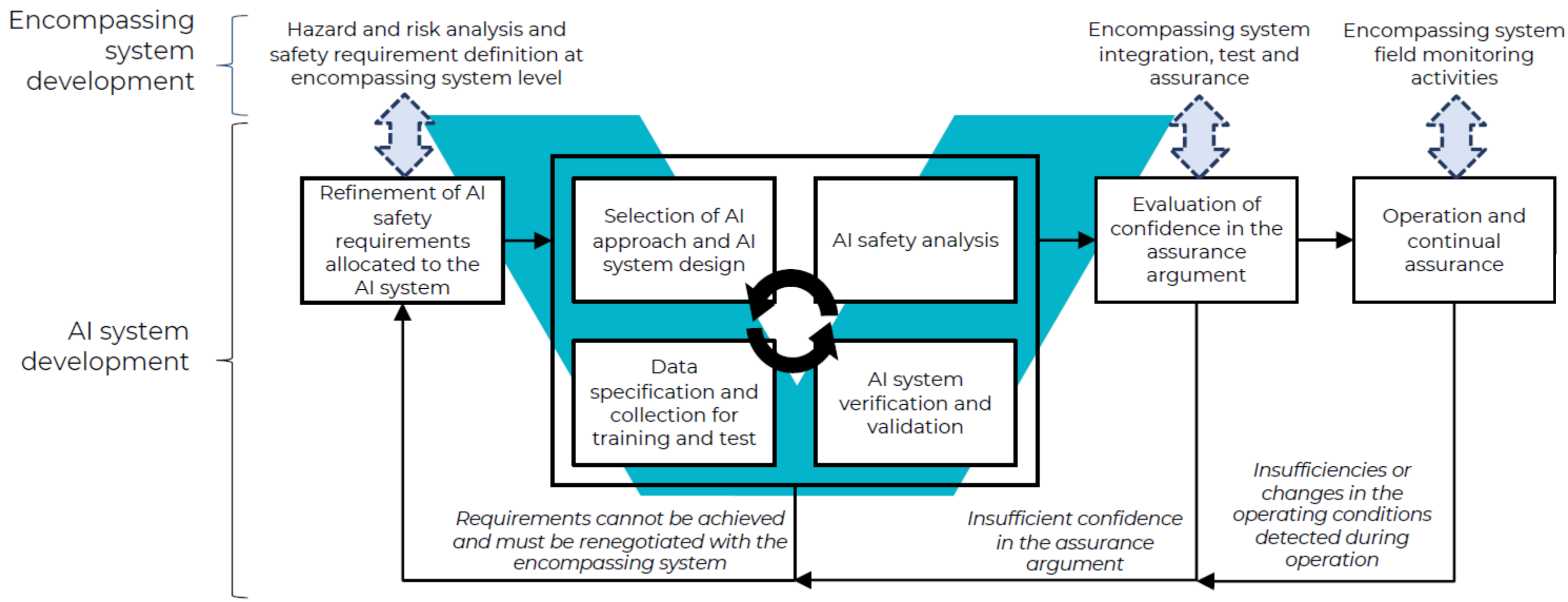
- How to define a "complete" specification:
  - Dealing with rare but critical events
  - Distributional shift / changes in the environment over time
- Performance Insufficiencies -> Model uncertainty:
  - Residual errors:
    - due to bias and lack of generalization and robustness: outputs sensitive to small changes in the inputs and insufficiencies in training data
  - Prediction uncertainty:
    - Confidence scores not necessarily indication of probability of correctness





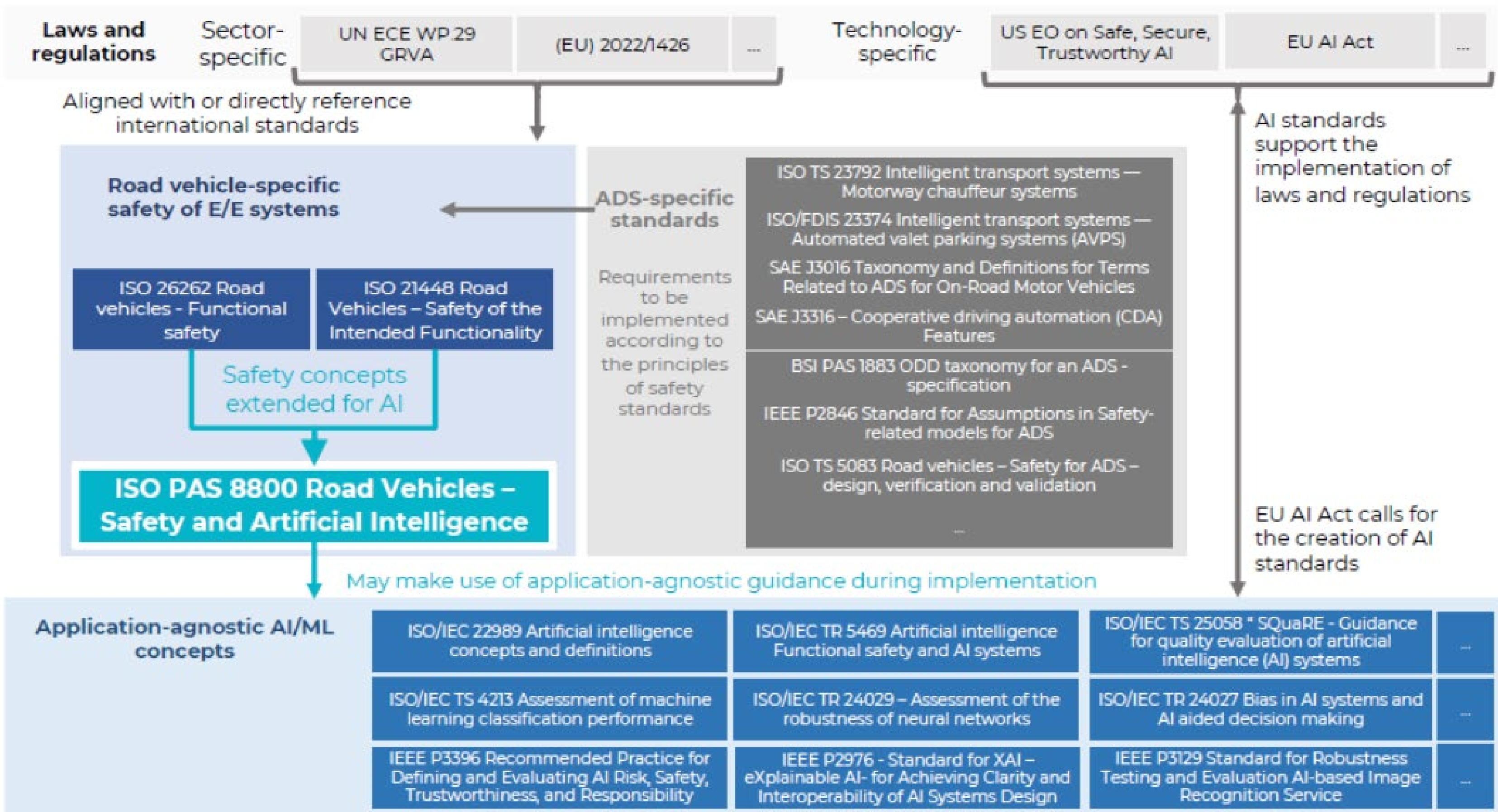
ISO PAS 8800 Safety and AI

- How to reduce Impact of AI Errors?
- How much do I have to reduce Uncertainty ?
- Which Safety Metrics Shall be quantitatively measured?
- What quantitative acceptance criteria for AI Safety metrics?



Source: CFAA – University of York – Prof. Burton

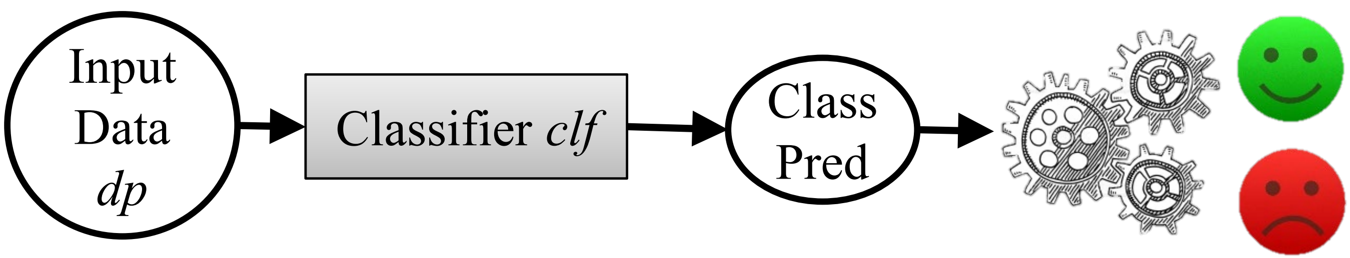
Complex Standards and Regulation Landscape



Source: CFAA – University of York – Prof. Burton

ML classifiers

- Machine Learning (ML) classifiers are increasingly used in critical systems.
- Classifiers, despite effective training, are prone to misclassifications → harmful in critical systems.
- Unclear Decision Boundaries: Difficulty in defining (perfect) decision boundaries in complex environments.

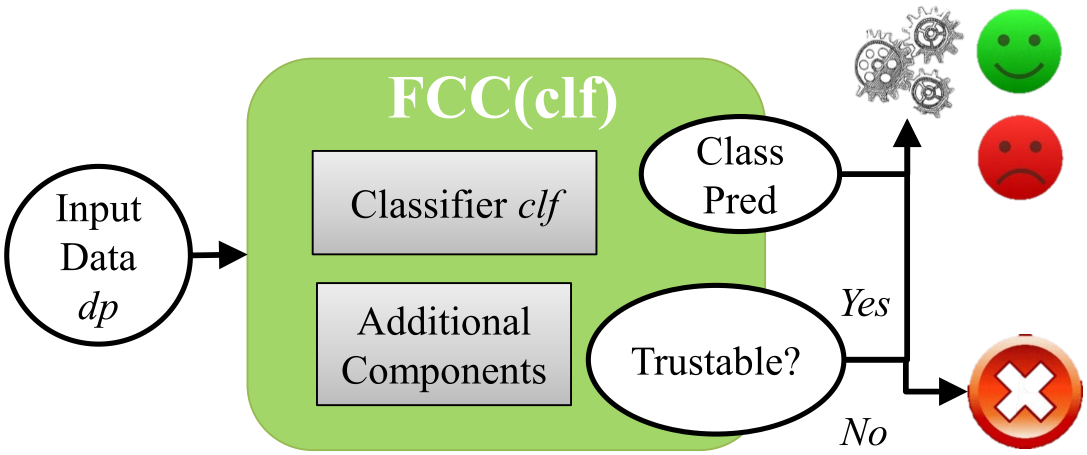


Dealing with ml classifiers

- Reducing misclassifications in critical systems where incorrect outputs can lead to severe consequences.
- Insight: Rather than striving for perfect accuracy, focus on integrating fail-controlled mechanism
- Classifiers as system components → flexible error handling of their failures.
- We look @Classifiers which **can reject** uncertain predictions.

Fail-Controlled Classifier (FCC)

- FCCs are designed to provide a correct prediction and reject uncertain ones.



- Advantages:
  - Reduces likelihood of incorrect decisions.
  - Shift from uncontrolled failures to controlled ones (omissions).

Evaluation Metrics

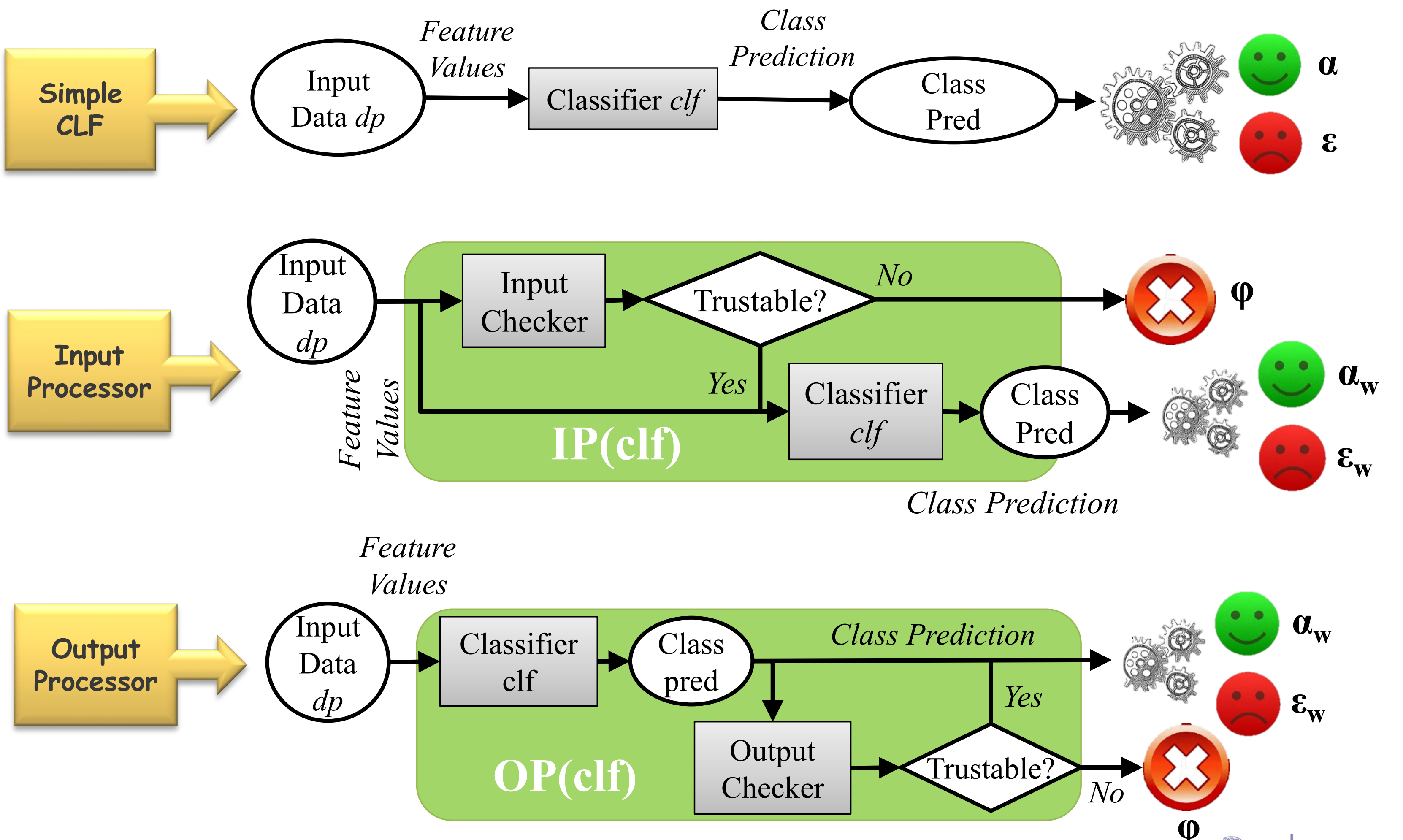
New evaluation metrics needed to account for rejection.

classifier behavior → FCC(clf) behavior ↓	Correct Prediction	Mis-classification	Sum
clf behavior	α	ε	1
Omitted	φ_c	φ_m	φ
Not omitted	α_w	ε_w	$1-\varphi$

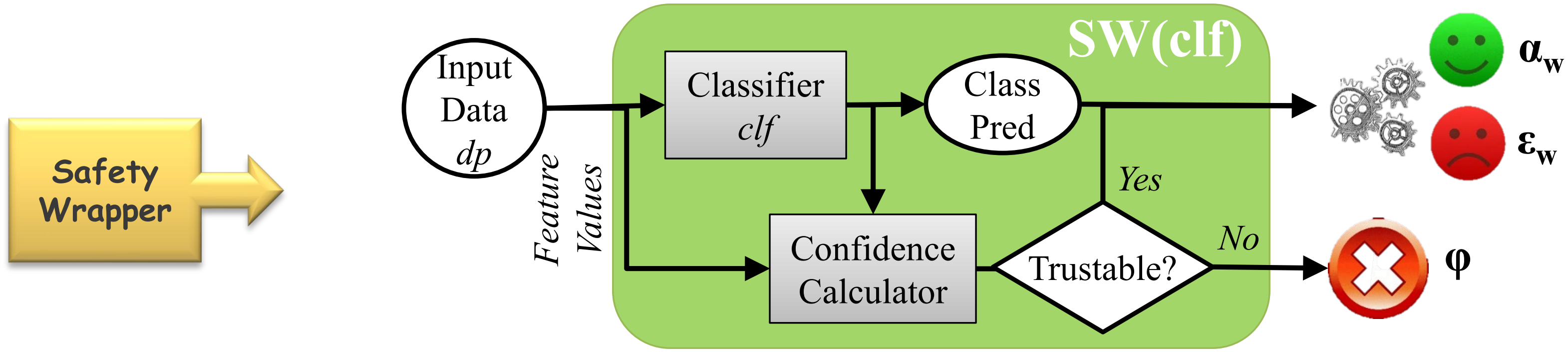
φ_m ratio = φ_m / φ, the ratio of omitted misclassifications over all omissions, to be maximized. (ideally one would like to omit misclassifications only)

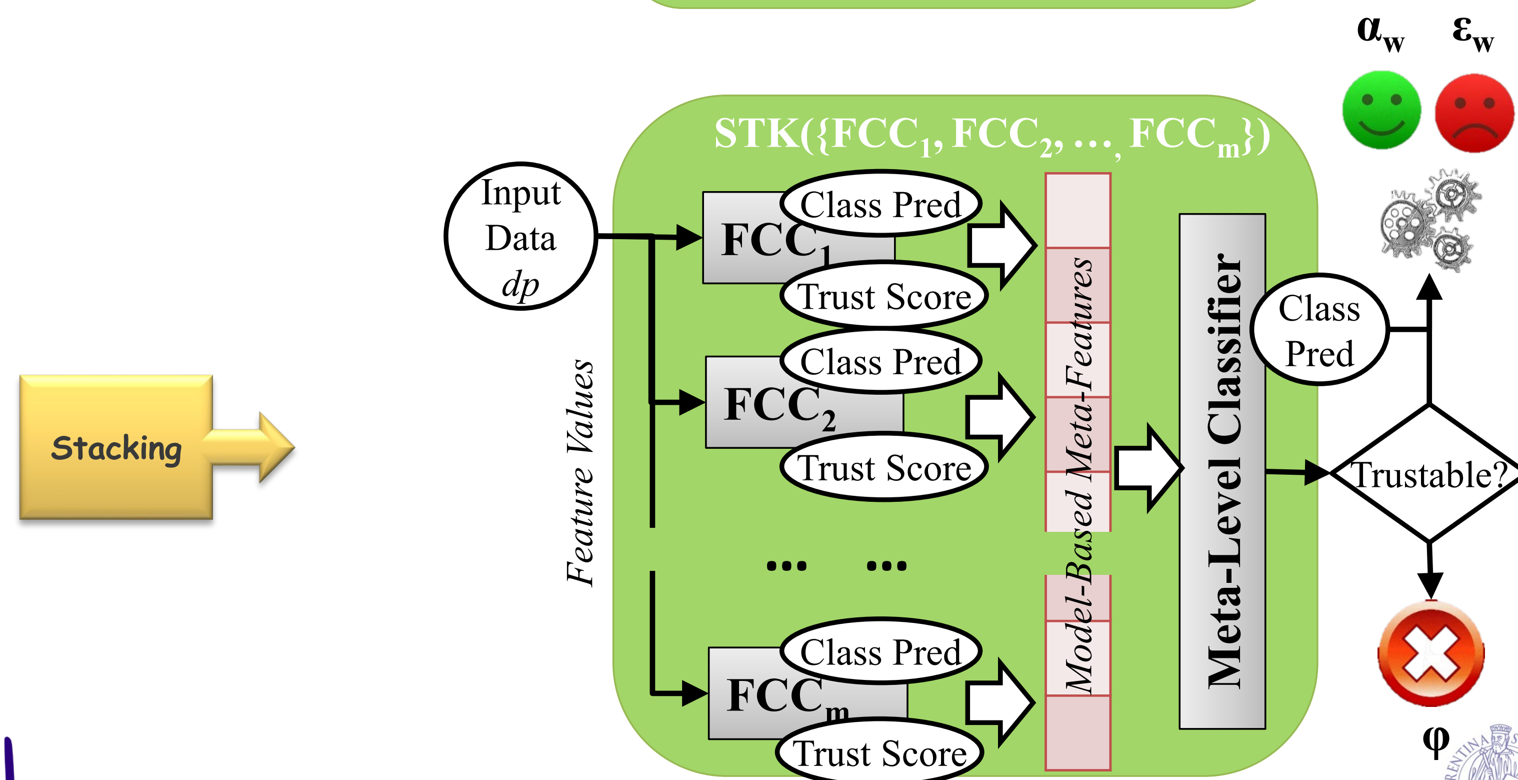
$\varepsilon_{drop} = (\varepsilon - \varepsilon_w)/\varepsilon$ the drop in misclassifications, to be maximized. (ideally ε_w should go to 0)

Software Architectures for FCCs

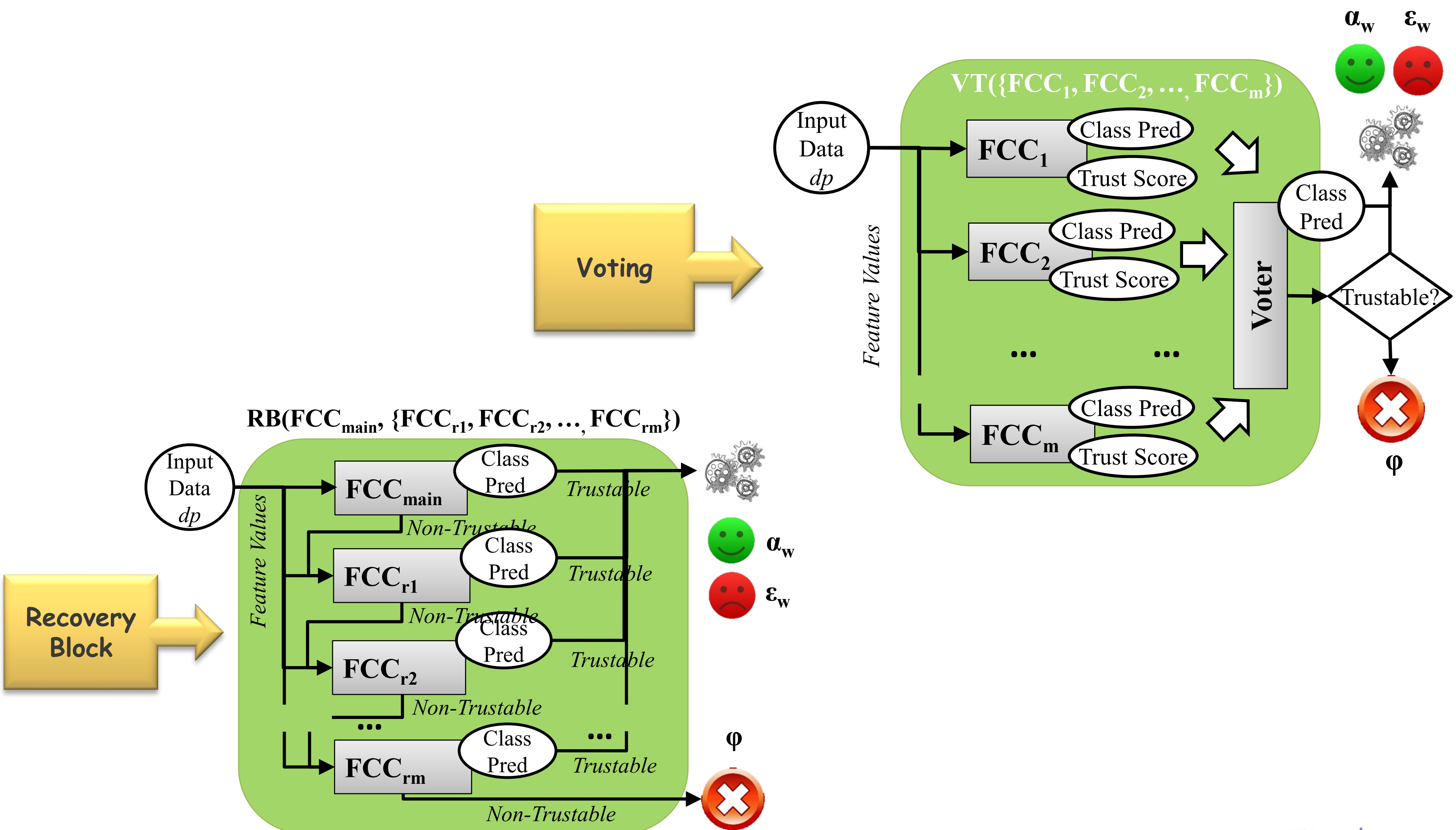


Software Architectures for FCCs-2



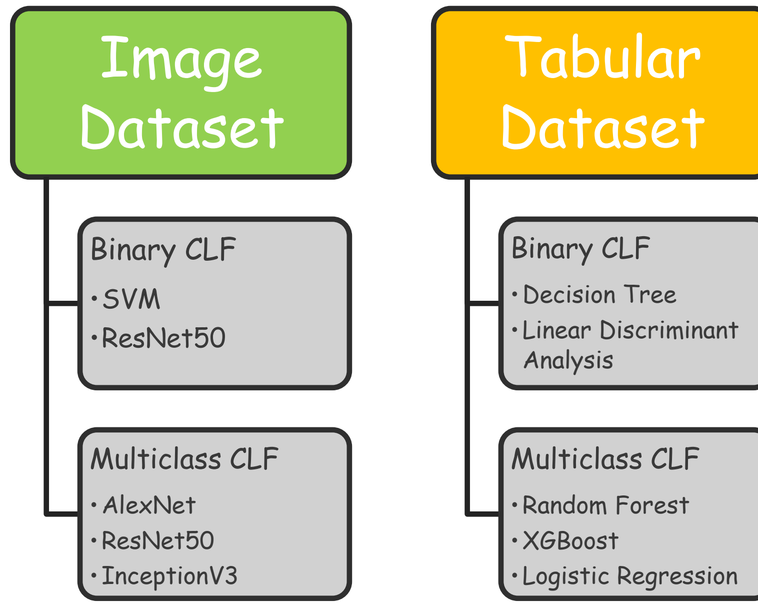


Software Architectures for FCCs -3



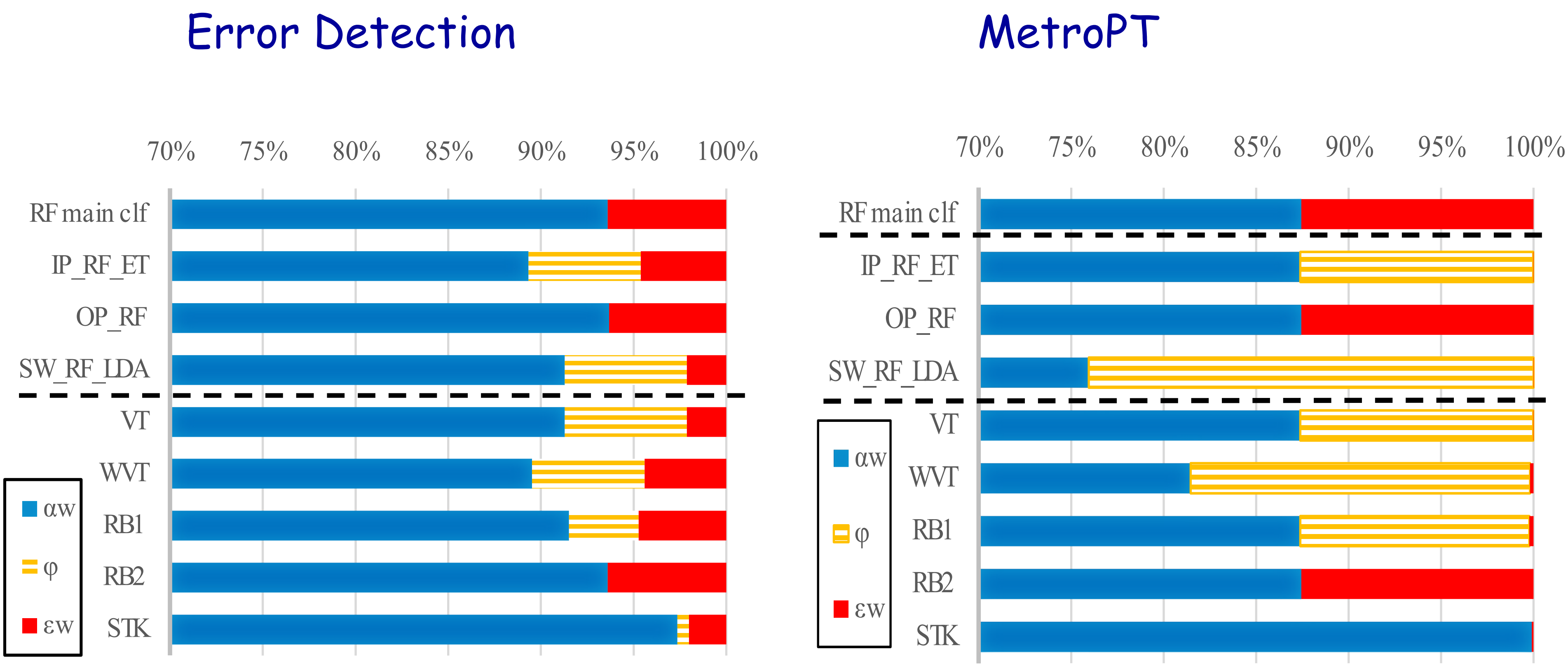
Some Experiments

- Two Types of Classifiers
  - Input Checker (Binary CLF)
    - Enables to detect either Normal or Anamolous input data.
  - Main Classifier (Multiclass CLF)
    - Enables to classify the class of the input data
- Dataset Used:
  - Tabular datasets:
    - CICIDS18 (Intrusion Detection), ARANCINO (Error Detection), MetroPT (Control Systems).
  - Image datasets:
    - FER-10, Food.



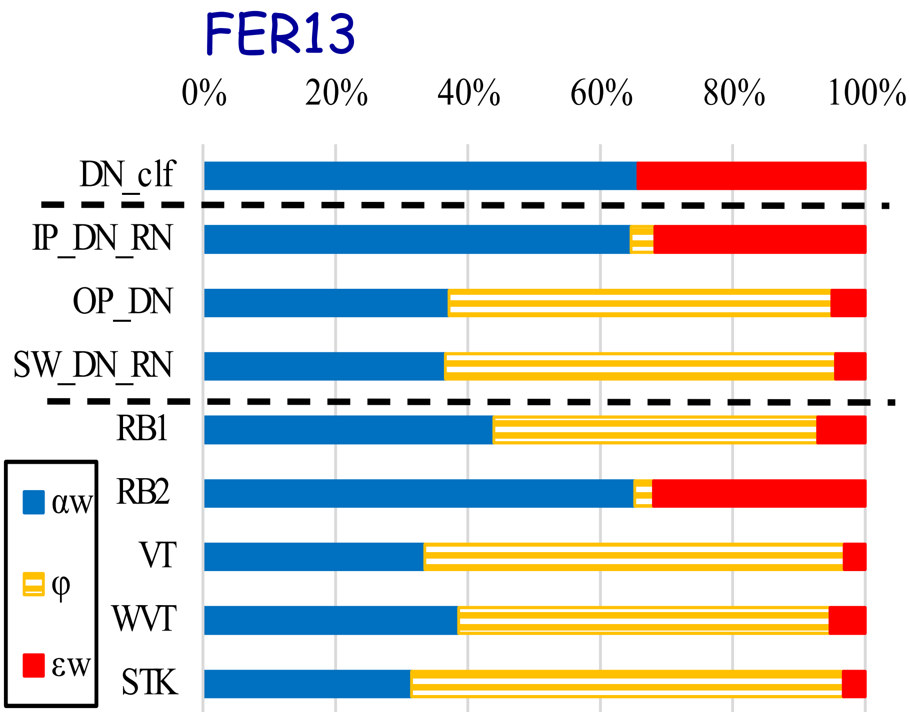
Results (Tabular Dataset)

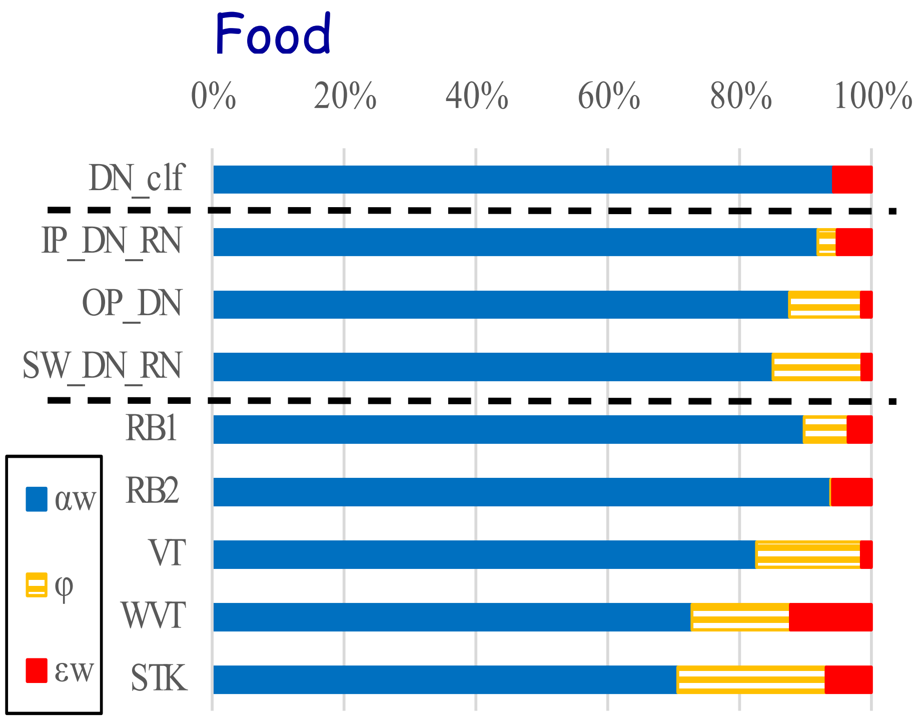
Comparison of FCCs using RF as the main classifier on tabular datasets.



Results (Image Dataset)

Comparison of FCCs using DN as the main classifier on image datasets.





Results (Unknown Inputs)

Rejection probability φ of unknown inputs for different FCCs (ideal 1.00)

Tabular Datasets	NIDS	Error Detection	MetroPT	Image Datasets	FER13	Flower	Food
IP_RF_ET	0.83	0.82	1.00	IP_DN_RN	0.99	0.83	0.86
OP_RF	0.00	0.04	0.00	OP_DN	0.90	0.32	0.26
SW_RF_LDA	0.78	0.50	0.98	SW_DN_RN	1.00	0.87	0.89
RB1	0.73	0.48	0.98	RB1	0.82	0.19	0.12
RB2	0.00	0.00	0.00	RB2	0.82	0.17	0.11
VT	0.83	0.82	1.00	VT	0.90	0.37	0.25
WVT	0.78	0.49	0.98	WVT	1.00	0.80	0.86
STK	0.00	0.00	0.00	STK	0.90	0.22	0.20

Diversity (Tabular Dataset)

- Classification performance, DISagreement, Double Fault DF, double reject DR of FCCs used for building RB, VT, WVT, STK tabular classifiers.
- Results are averaged across the three tabular datasets

				DISagreement DIS (best if high)						Double Fault (DF) (best if low)						Double reject DR (best if low)					
FCC	φ	α_w	ε_w	IP_DT_ET	IP_RF_LDA	OP_DT	OP_LR	SW_LR_ET	SW_XGB_LDA	IP_DT_ET	IP_RF_LDA	OP_DT	OP_LR	SW_LR_ET	SW_XGB_LDA	IP_DT_ET	IP_RF_LDA	OP_DT	OP_LR	SW_LR_ET	SW_XGB_LDA
IP_DT_ET	0.138	0.846	0.016	-	0.05	0.05	0.27	0.23	0.06	-	0.01	0.02	0.01	0.01	0.01	-	0.11	0.00	0.05	0.14	0.12
IP_RF_LDA	0.159	0.814	0.027	0.05	-	0.09	0.29	0.26	0.01	0.01	-	0.03	0.02	0.01	0.03	0.11	-	0.00	0.04	0.12	0.16
OP_DT	0.000	0.896	0.104	0.05	0.09	-	0.25	0.28	0.09	0.02	0.03	-	0.06	0.01	0.02	0.00	0.00	-	0.00	0.00	0.00
OP_LR	0.262	0.654	0.084	0.27	0.29	0.25	-	0.03	0.29	0.01	0.02	0.06	-	0.03	0.02	0.05	0.04	0.00	-	0.26	0.05
SW_LR_ET	0.352	0.619	0.029	0.23	0.26	0.28	0.03	-	0.26	0.01	0.01	0.01	0.03	-	0.01	0.14	0.12	0.00	0.26	-	0.12
SW_XGB_LDA	0.169	0.805	0.026	0.06	0.01	0.09	0.29	0.26	-	0.01	0.03	0.02	0.02	0.01	-	0.12	0.16	0.00	0.05	0.12	-

Diversity (Image Dataset)

- Classification performance, DISagreement, Double Fault DF, double reject DR of FCCs used for building RB, VT, WVT, STK image classifiers.
- Results are averaged across the three Image datasets.

FCC	φ	α_w	ε_w	DISagreement DIS (best if high)						Double Fault (DF) (best if low)						Double reject DR (best if low)					
				IP_DN_RN	OP_DN	OP_IC	SW_AN_GN	SW_IC_GN	SW_VGG_RN	IP_DN_RN	OP_DN	OP_IC	SW_AN_GN	SW_IC_GN	SW_VGG_RN	IP_DN_RN	OP_DN	OP_IC	SW_AN_GN	SW_IC_GN	SW_VGG_RN
IP_DN_RN	0.032	0.787	0.181	-	0.19	0.18	0.30	0.17	0.20	-	0.05	0.06	0.05	0.06	0.06	-	0.02	0.01	0.03	0.03	0.03
OP_DN	0.327	0.622	0.051	0.19	-	0.12	0.20	0.13	0.13	0.05	-	0.04	0.04	0.04	0.04	0.02	-	0.23	0.28	0.23	0.25
OP_IC	0.279	0.656	0.065	0.18	0.12	-	0.23	0.02	0.15	0.06	0.04	-	0.04	0.06	0.05	0.01	0.23	-	0.24	0.28	0.21
SW_AN_GN	0.449	0.494	0.056	0.30	0.20	0.23	-	0.21	0.15	0.05	0.04	0.04	-	0.04	0.04	0.03	0.28	0.24	-	0.25	0.30
SW_IC_GN	0.297	0.638	0.064	0.17	0.13	0.02	0.21	-	0.14	0.06	0.04	0.06	0.04	-	0.05	0.03	0.23	0.28	0.25	-	0.23
SW_VGG_RN	0.336	0.594	0.070	0.20	0.13	0.15	0.15	0.14	-	0.06	0.04	0.05	0.04	0.05	-	0.03	0.25	0.21	0.30	0.23	-

Concluding

- Machine learning classifiers are one of the must-have for critical systems designers despite the difficulties in properly integrating and operating them.
- Instead of dreaming and striving for perfect accuracy, focus on **reducing misclassifications by** integrating fail-controlled mechanism
- FCCs provide a safer alternative to traditional classifiers in critical systems.
- Emphasize system-level design to manage uncertainty and failures.

My roadmap

- Further research on uncertainty quantification and rejection mechanisms.
- Structures to minimize rejections effectively.
- Design different software architectures using FCC's.
- Integrating the Design of FCCs with the industry-specific standards
  - See e.g. the ISO PAS 4000.