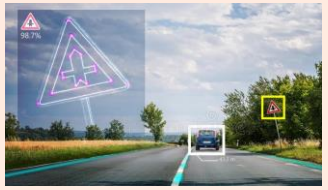
Methods and tools for trusted AI: an urgent challenge for industry

Juliette Mattioli, Thales

Rodolphe Gelin, Renault

Trust relies on many aspects

AI challenges plenty of them



AI is an unescapable technology

But is has fundamental weaknesses

Bias in learning data set

Lack of robustness

Unknow unknow

Data obsolescence

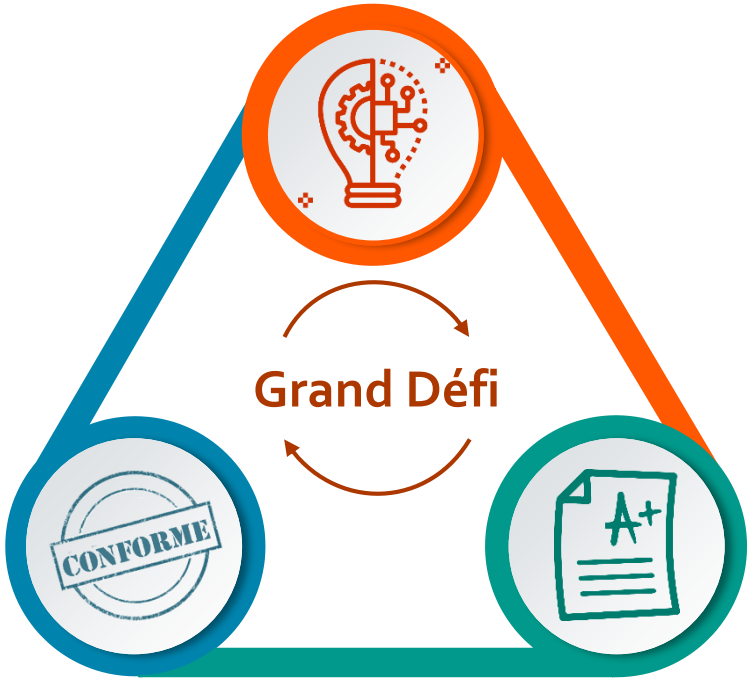
Trustworthy & certification AI: from data/algo to AI SW & Systems Engineering

How to design, deploy, maintain, certify AI based critical systems ?

Technological pillar

DATAS, AI ALGO, SW, SYSTEMS engineering to design, deploy and maintain AI based critical system

… industry strongly involved in programs, especially AI Manifesto members



… Cooperation with French basic research Initiatives, such as ANITI or DataIA, and academic research

Norms pillar

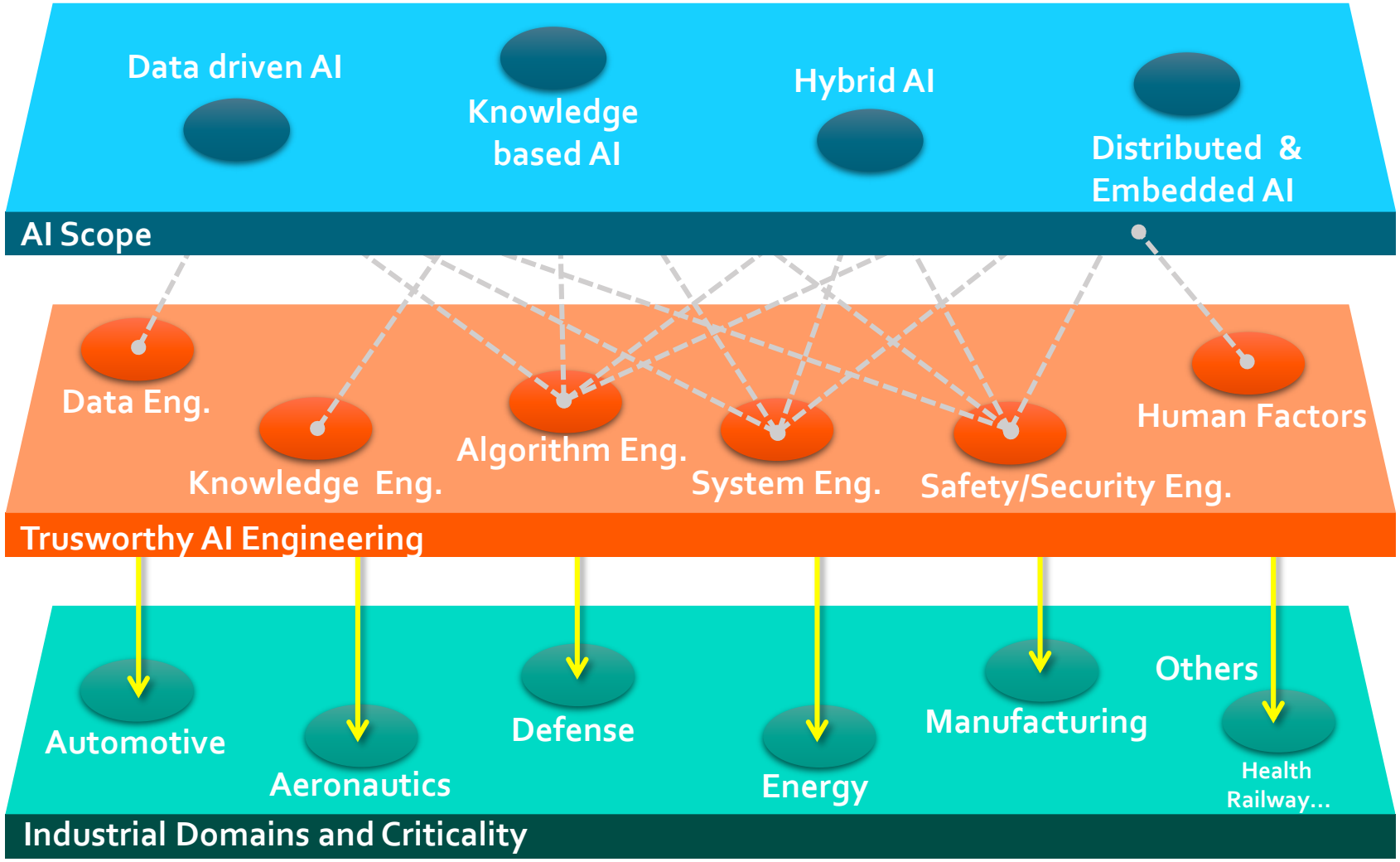
Norm, standard and regulation environment toward certification

Applications Evaluation Pillar

Ensure the right operational exploitation

Toward global strategy with coordinated programs and funding (Private, Public)

Confiance.ai program
(Global budget: 45M€, Duration: 4 years)



Confiance.ai structural view

Transversal functions

- Manage Requirements
- Manage Explainability
- Justify changes in methods
- Ensure the data source traceability
- Manage datasets (including egal compliance)
- Fuel the certification process
- Revisit V&V strategy
- Collaborate
- Manage the building blocs
- Manage in configuration
- Cyber-secure the workbench

Tooled methods & guidelines

- Define design methods
- Define safety analysis methods
- Define modalities & methods for human interaction
- Train final users
- Evangelize workbench stakeholders

Design, deploy and maintain AI modules and AI-based critical systems

(Human factors are always considered)

Design, deploy and maintain AI modules (including explainability)

	Specification	Design	Implement	Unit Verification
Data-driven AI — Training data	*Specification (inc. quality, cybersecurity & evaluation KPI)*	*Design (inc. collect & retrieval)*	*Implement (inc. annotate, complete, organize, anonymize)*	*Verify (correctness, coverage, quality, bias, ...)*
Data-driven AI — Test/V&V Data	*Specification (inc. quality, cybersecurity & evaluation KPI)*	*Design (inc. collect & retrieval)*	*Implement (inc. annotate, complete, organize, anonymize)*	*Verify (correctness, coverage, quality, bias, ...)*
Data-driven AI — Learning methods and algos	*Specification*	*Design (inc. Architectural Design)*	*Implement (inc. on the execution targeted HW)*	*Verify (correctness, coverage, quality, bias, ...)*
Knowledge-based AI — Knowledge bases	*Specification (inc. quality, cybersecurity & evaluation KPI)*	*Design (inc. collect & retrieval)*	*Implement*	*Verify (correctness, coverage, quality, bias, ...)*
Knowledge-based AI — Test/V&V Data	*Specification (inc. quality, cybersecurity & evaluation KPI)*	*Design (inc. collect & retrieval)*	*Implement*	*Verify (correctness, coverage, quality, bias, ...)*
Knowledge-based AI — Symbolic AI component	*Specification*	*Design (inc. Architectural Design)*	*Implement (inc. on the execution targeted HW)*	*Verify (correctness, coverage, quality, bias, ...)*
Hybrid AI (considered as an AI based system)	*Specification*	*Design*	*Implement*	*Verify (correctness, coverage, quality, bias, ...)*
Distributed AI (considered as an AI based system)	*Specification*	*Design*	*Implement*	*Verify (correctness, coverage, quality, bias, ...)*

Design, deploy and maintain AI-based critical systems

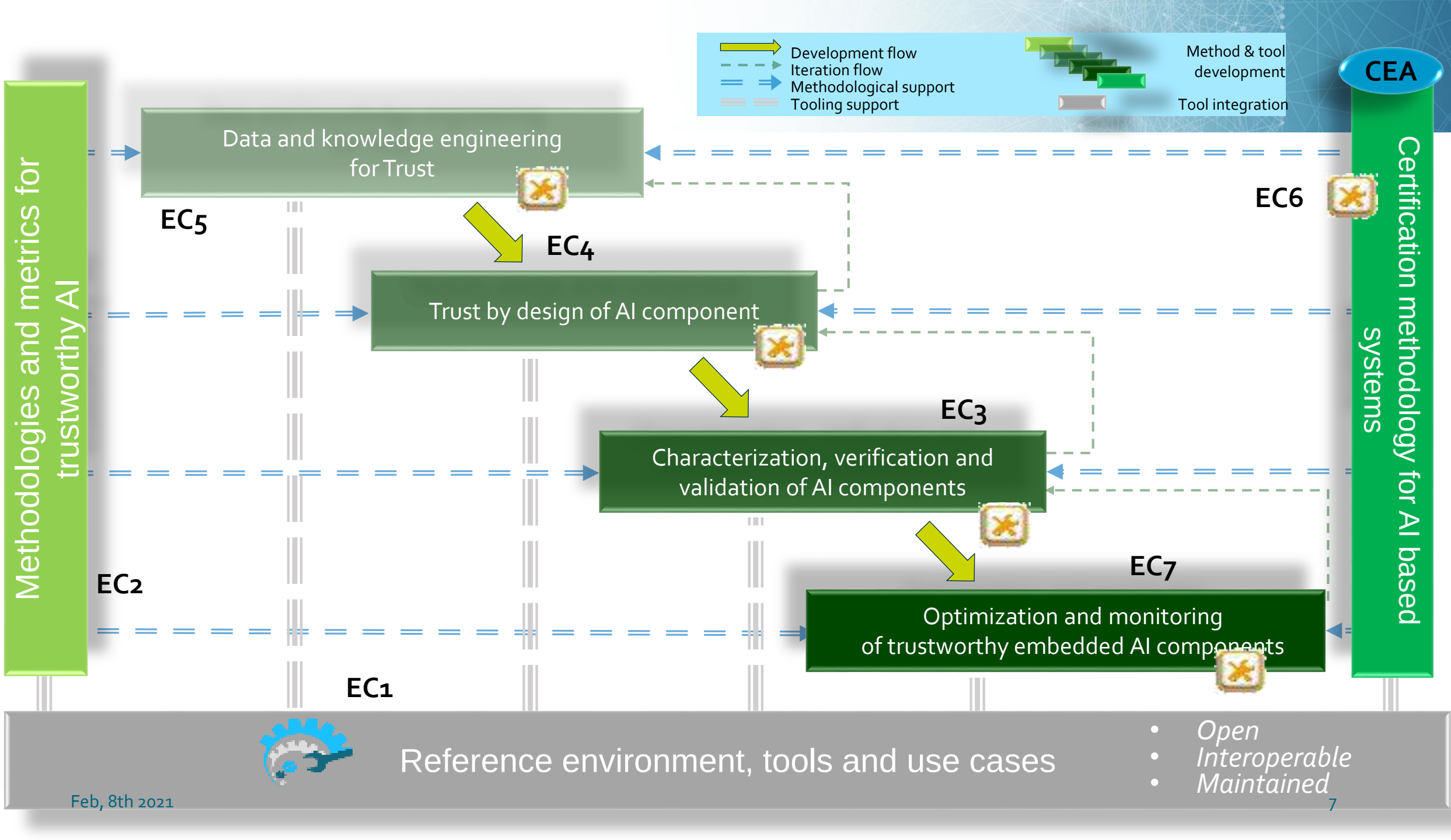
	Characterize operational domain	System Specification	Design	Pre-integration / integration	System Verification & Validation	System Monitoring	Operational Domain Monitoring
Data-driven AI	*Characterize operational domain*	*System Specification*	-	Pre-integration / integration (impact on safety, dependability...)	*System Verification & Validation*	System Monitoring (inc. Assess discrepancy from the model)	*Operational Domain Monitoring (inc. obsolescence management of datasets)*
Knowledge-based AI	*Model operational domain*	*System Specification*	-	Pre-integration / integration (impact on safety, dependability...)	*System Verification & Validation*	System Monitoring (inc. Assess discrepancy from the model)	*Operational Domain Monitoring (inc. obsolescence management of data & knowledge bases)*
Hybrid AI (considered as an AI based system)	*Characterize/model operational domain*	*System Specification*	*AI based system Design*	Pre-integration / integration (impact on safety, dependability...)	*System Verification & Validation*	System Monitoring (inc. Assess discrepancy from the model)	*Operational Domain Monitoring (inc. obsolescence mgt of data & knowledge bases)*
Distributed AI (considered as an AI based system)	*Characterize/model operational domain*	*System Specification & emerging behaviors*	*AI based system Design (via models)*	Pre-integration / integration (AI impact on safety, dependability...)	*System Verification & Validation*	System Monitoring (inc. Assess discrepancy from the model)	*Operational Domain Monitoring (inc. obsolescence mgt of data & knowledge bases)*

Interface

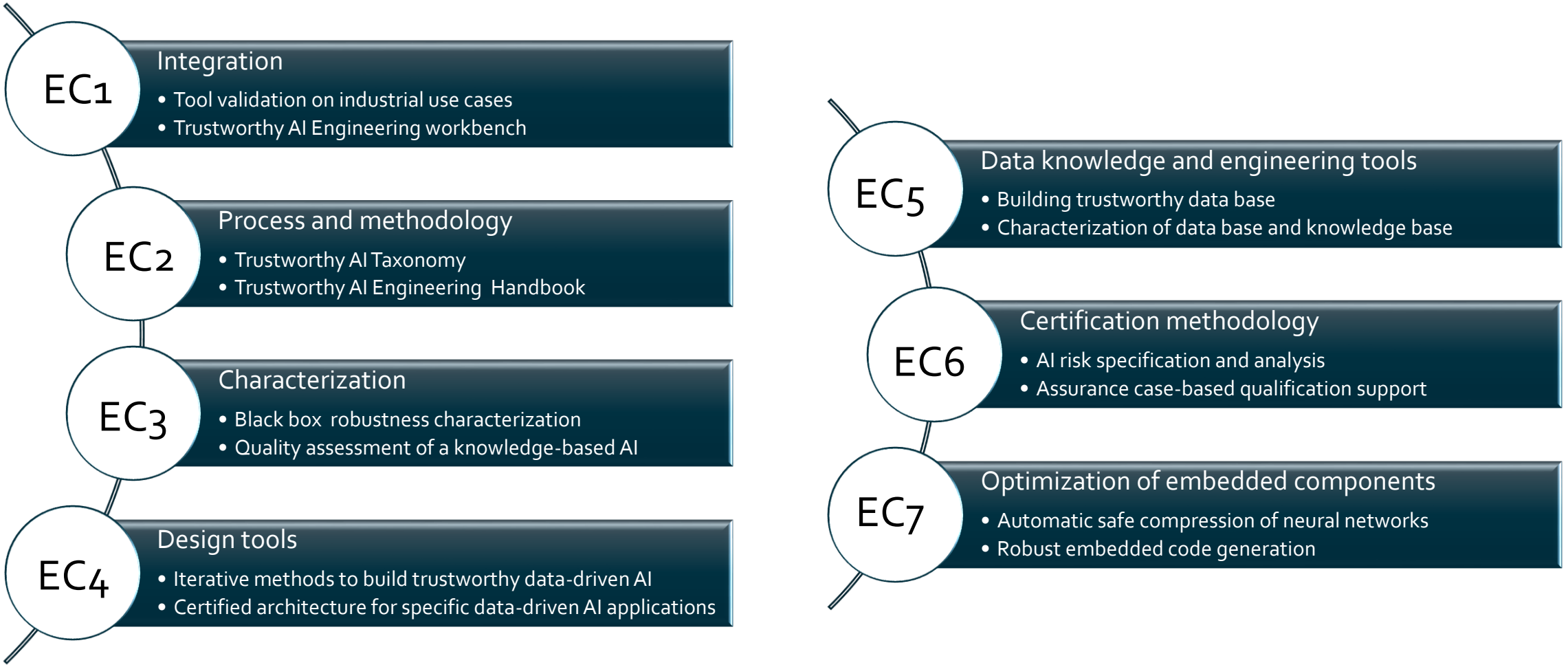
- Interfacing functionalities
- Interfacing with modeling, simulation & operational domain
- **Interfacing with existing industrial workbench**
- **Interfacing with existing AI SDKs and tools**
- **Interfacing with Static frameworks** *(dataset, ...)*
- Interfacing with Dynamical frameworks (HIL, MIL, SIL, ...)
- Interfacing with existing model libraries
- Interfacing with existing tooled methods
- Interfacing with legal and regulation frames

Feb, 8th 2021

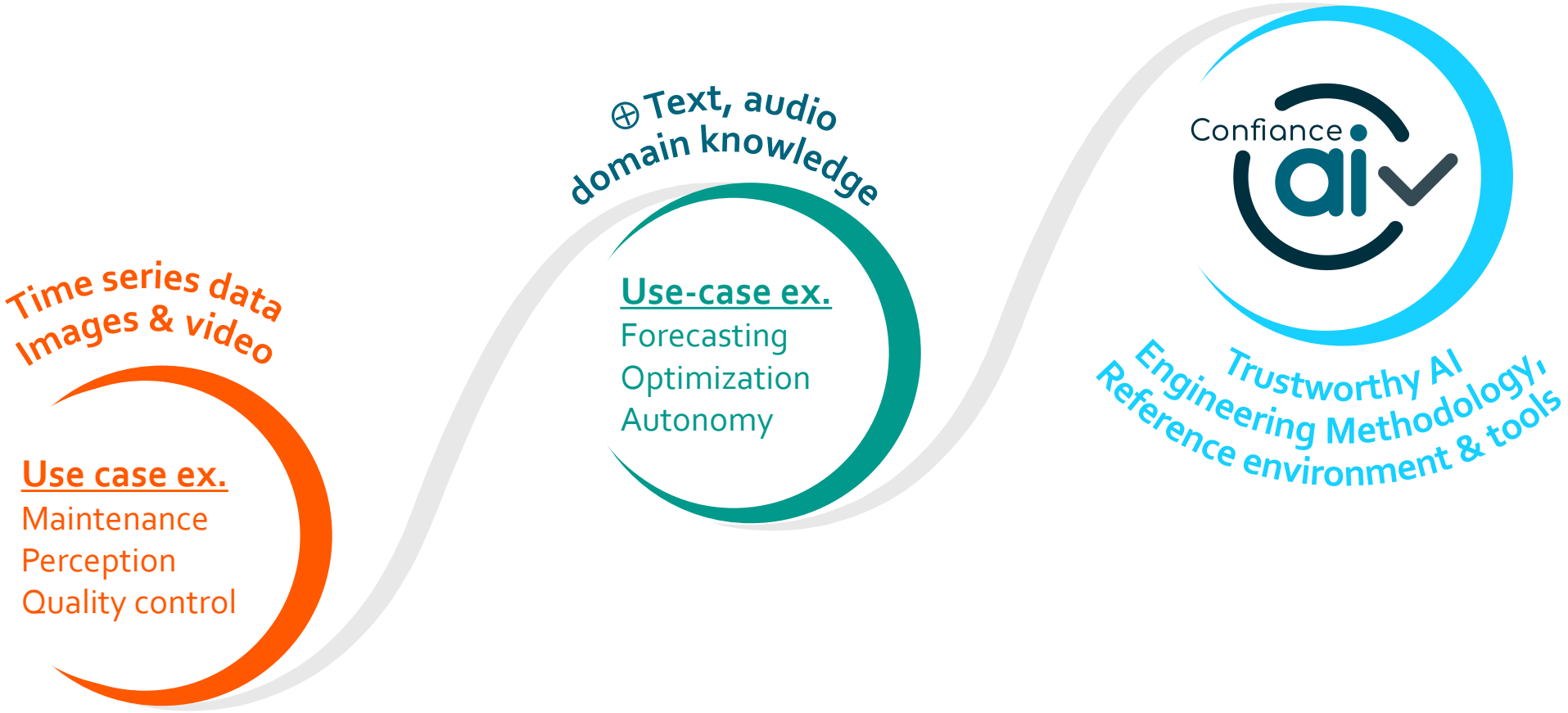
Taxonomy, methodology and guidelines


Development flow
Iteration flow
Methodological support
Tooling support
Method & tool development
Tool integration
CEA
Methodologies and metrics for trustworthy AI
Certification methodology for AI based systems
Data and knowledge engineering for Trust
EC5
EC4
EC6
Trust by design of AI component
EC3
Characterization, verification and validation of AI components
EC7
EC2
Optimization and monitoring of trustworthy embedded AI components
EC1
Reference environment, tools and use cases
Open
Interoperable
Maintained

Projects outputs (samples)



An incremental roadmap validated by various use-cases



2021 — Data driven AI — Low criticality

2022 — Data driven AI & Knowledge based AI — Medium criticality

2023 — Overall AI including hybrid AI — High criticality

2024

A non exhaustive view of the Thrustworthy AI Ecosystem

www.confiance.ai
contact@irt-systemx.fr