SAIF

Safe AI through Formal Methods

Kick-off Meeting

November 23rd, 2023

SAIF

Safe AI through Formal Methods

- **Project Coordinator**: **Caterina Urban** (+ Zakaria Chihani)
- **Web Site**: https://project.inria.fr/saif/
- **Project Members**
  - **ANTIQUE**, Inria Paris
  - **LaBRI**, Université de Bordeaux
  - **LIX**, Institut Polytechnique de Paris
  - **LMF**, Université Paris-Saclay
  - **LSL**, CEA-List
  - **SuMo**, Inria Rennes
  - **TAU**, Inria Saclay

Project Goal

Harness and *rethink* decades of work in FORMAL METHODS to tackle the modern challenges of MACHINE LEARNING

Objective 1

Specifying ML-based Systems

Objective 2

Validating ML-based Systems

Objective 3

Guiding the Design of ML-based Systems

Objective 1

Specifying ML-based Systems

- **Pushing the boundaries in the landscape of formal specification techniques for ML-based systems**
- Motivated by the following CHALLENGES
  - ML-based systems are **difficult to specify**
    (e.g., “what is a human in an image?”)
  - ML-based systems are **built from large example bases**,
    rather than from well-structured specifications
  - ML-based systems are **monolithic**:
    unlike human-written software, ML models are rarely decomposable
    in smaller components, each with its own specification

Objective 2

Validating ML-based Systems

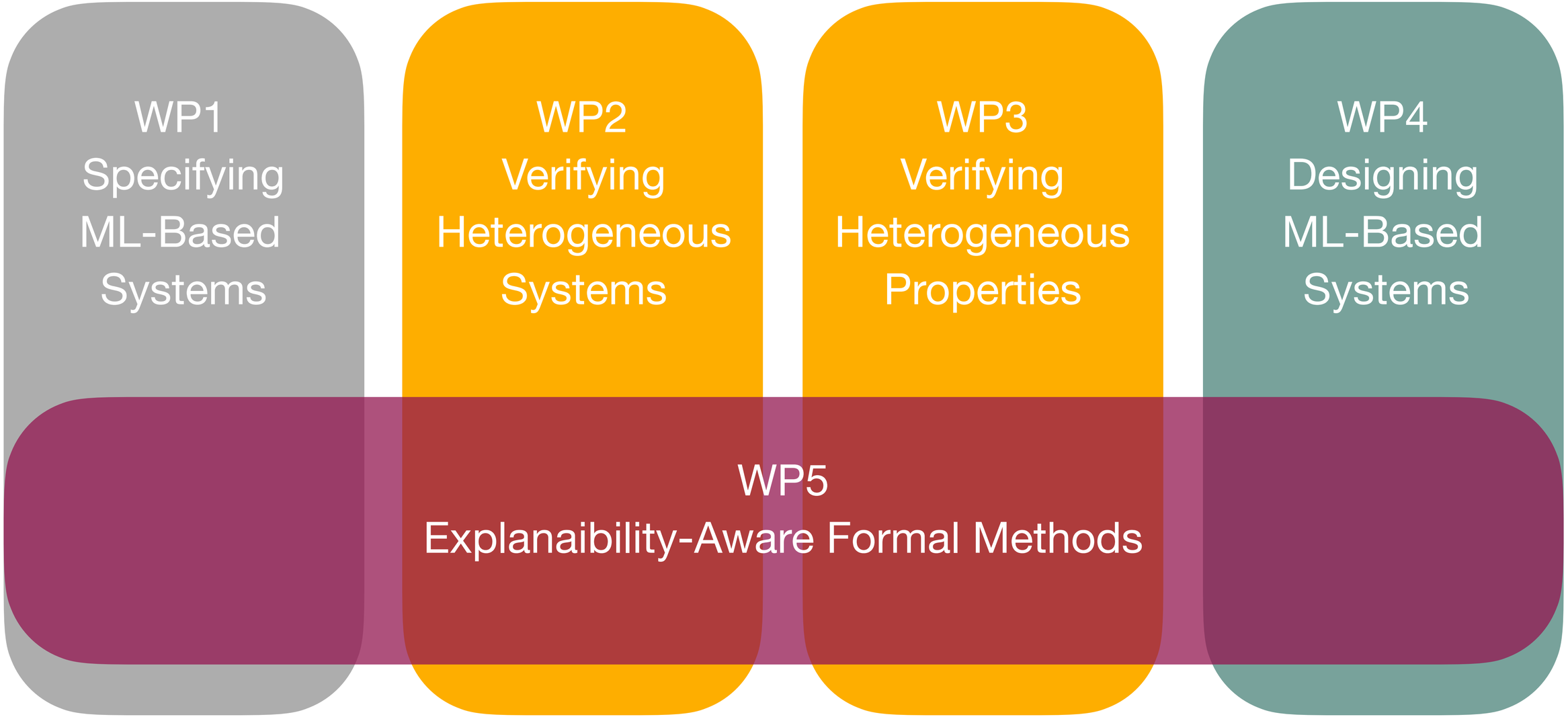
- **Tackling the verification of a much broader spectrum of properties for a much more comprehensive spectrum of ML-based systems**
- Motivated by the following CHALLENGES
  - ML-based systems are **large**:
    ML models can have hundreds of millions of parameters
  - ML-based systems are **heterogeneous**
    (i.e., different activation functions, different architectures, etc.)

Objective 3

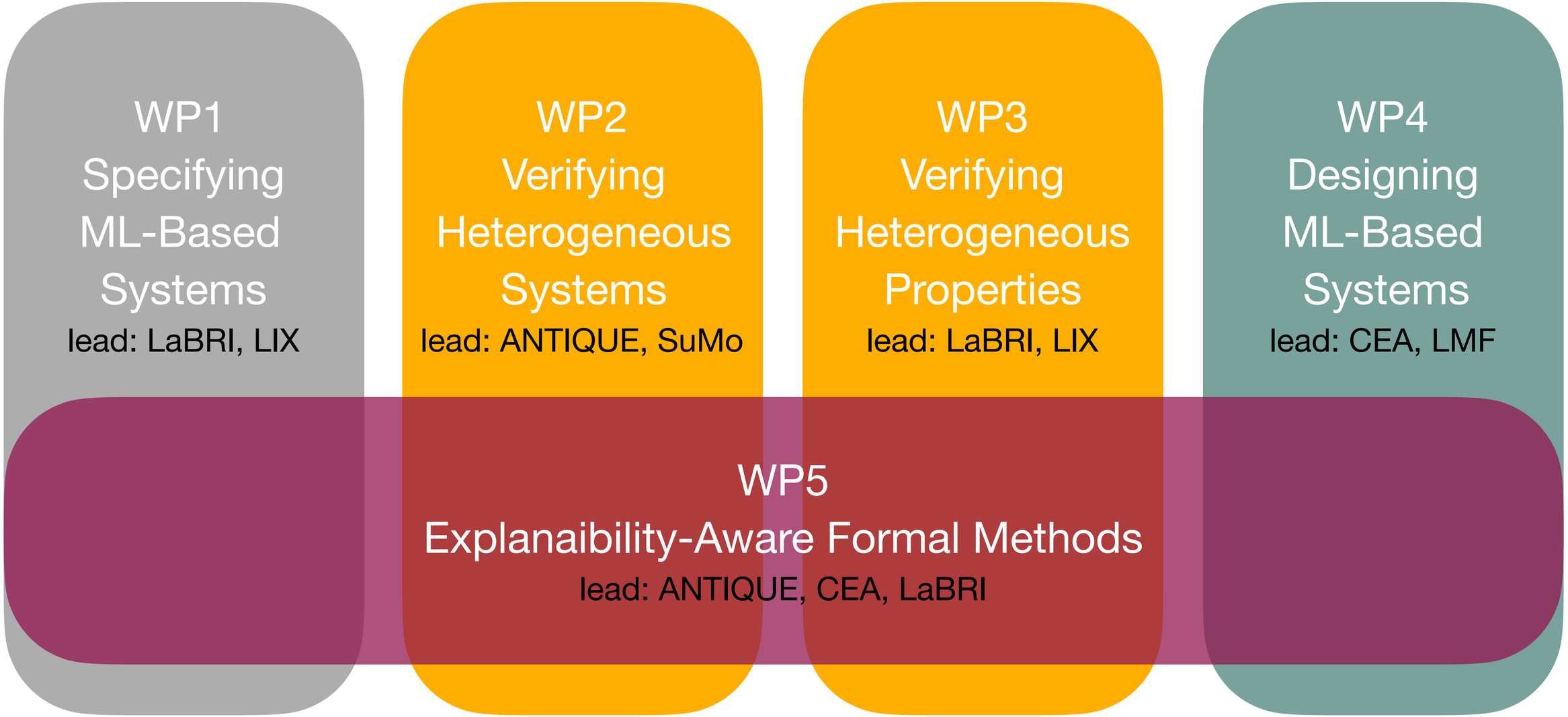
Guiding the Design of ML-based Systems ML-based Systems

- **Making formal methods an asset in the design of more trustworthy and easier to verify ML-based systems**
- Motivated by the following CHALLENGES
  - ML-based systems are essentially **stochastic / uncertain objects**
  - ML-based systems are **built from large example bases**, rather than from well-structured specifications
  - ML-based systems are **monolithic**: unlike human-written software, ML models are rarely decomposable in smaller components, each with its own specification
  - ML-based systems are **opaque** (i.e., danger or bias, lack of interpretability, etc.)

Action Plan


WP1
Specifying
ML-Based
Systems
WP2
Verifying
Heterogeneous
Systems
WP3
Verifying
Heterogeneous
Properties
WP4
Designing
ML-Based
Systems
WP5
Explanaibility-Aware Formal Methods

Management


WP1
Specifying
ML-Based
Systems
lead: LaBRI, LIX
WP2
Verifying
Heterogeneous
Systems
lead: ANTIQUE, SuMo
WP3
Verifying
Heterogeneous
Properties
lead: LaBRI, LIX
WP4
Designing
ML-Based
Systems
lead: CEA, LMF
WP5
Explanaibility-Aware Formal Methods
lead: ANTIQUE, CEA, LaBRI

Management

6.2 Modalités de pilotage et engagements de collaboration

L'Établissement coordinateur élabore, avec l'appui du Responsable du projet et des Etablissements partenaires les comptes-rendus annuels d'avancement et de fin du Projet pour l'ensemble des travaux menés en collaboration avec les Établissements partenaires. Il assure la centralisation des relevés de dépenses et des éléments de suivi établis notamment par les Etablissements partenaires et leur bonne transmission à l'ANR.

Expected Outcomes

- **Scientific Outcomes**
  - scaling up of formal methods for ML-based systems
  - design of new dedicated abstractions and approaches
  - release of open-source tools, libraries, and benchmarks

- **Societal Impact**
  - new frameworks for the specification of trustworthiness properties
  - guidelines for the design of ML-based systems amenable to verification
  - improvement of the quality and reliability of ML-based systems

Other Matters

Project Duration

Article 4 : DURÉE DU PROJET

La date de commencement du Projet et de prise en compte des dépenses est fixée au 01/10/2023.

La durée de réalisation du Projet est fixée à **48 mois**, soit un achèvement prévu au 30/09/2027 qui correspond à la date de fin de prise en compte des dépenses.

L'ANR doit être informée de l'achèvement du Projet si celui-ci intervient avant la date prévue ci-dessus.

Other Matters

Communication

Article 9 : COMMUNICATION

L'Établissement coordinateur et les Établissements partenaires s'engagent à mentionner le soutien apporté par l'ANR au titre de France 2030, en indiquant le numéro du Contrat, dans leurs propres actions de communication sur le Projet « SAIF » (ANR-23-PEIA-0006) et dans leurs publications (par exemple : « Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence « ANR-23-PEIA-0006 »). Les supports de communication orale, les communications par voie d'affiche, les sites internet doivent également afficher les logos « France 2030 ».

Other Matters

Project Seminar

- **Hybrid vs Virtual**
  - Where?
  - Zoom?
- **Bi-Weekly vs Monthly**
- Which **day**?
- What **time**?

Other Matters

(Summer) School

- **FoPPS Schools?**
  - Foundations of Programming and Software Systems https://etaps.org/about/fopss-schools/
  - biannual, in 2026 the idea is to have a pre-FLoC edition (Lisbon)
  - supported by ETAPS, EATCS, ACM SIGLOG, ACM SIGPLAN
- **Plans for 2024 and 2025?**

Other Matters

Next Meeting

7.2 Réunions de suivi du Projet

Le coordinateur de la stratégie nationale d'accélération et le(s) pilote(s) scientifique(s) sont conviés aux réunions prévues aux articles suivants.

7.2.1. Réunion de lancement

Le Responsable du projet organise une réunion de lancement du Projet avec les Établissements partenaires dans un délai de quatre mois suivant la date de signature du Contrat. L'ANR est consultée sur la date de cette réunion au moins un (1) mois à l'avance afin de pouvoir y participer.

Other Matters?

Agenda

- 9h00-9h30 Project Overview and Coordination
- 9h30-9h45 ANTIQUE
- 9h45-10h05 LaBRI
- 10h05-10h25 LIX
- 10h25-10h40 Break
- 10h40-11h00 LMF
- 11h00-11h20 LSL
- 11h20-11h40 SuMo
- 11h40-12h00 TAU
- 12h00 Lunch @ Brass&Co