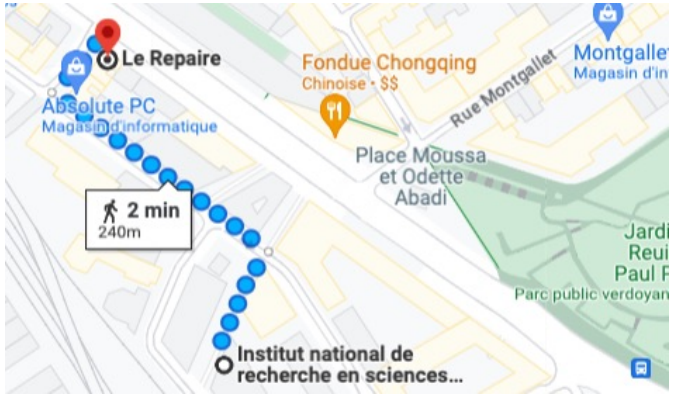
HyAIAI
Final Evaluation
4 years
27/06/2023

Défi HyAIAI
Hybrid Approaches for Interpretable AI

Elisa Fromont (Lacodam)

Lacodam – Magnet – Multispeech – Orpailleur – Scool – TAU

Today's program



- 9h40 → 10h00 Arrivals

(Morning in English, Afternoon in French)

- 10:00 → 10h10 Introduction from **Jean-Frédéric Gerbeau**
- 10h10 → 10h30: Overview of the project **by Elisa Fromont (LACODAM)**
- 10h30 → 11h00 discussion with the INRIA instances
- 11h00 → 11h25: Talk by **Debabrota Basu (SCOOL)** on "Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback"
  11h25 → 11h35 Q&A
- 11h35 → 12h00: Talk by **Jan Ramon (MAGNET)** "Hybrid Approaches for Interpretable Private AI"
  12h00 → 12h15 Q&A
- **12h20 —> 14h00 LUNCH BREAK (Le Repaire)**

- 14h20 → 14h50 Talk by **Miguel Couceiro (ORPAILLEUR)** "A (de)tour through bias mitigation and analogy based ML"
  14h50 → 15h05 Q&A
- 15h00 → 15h30 Talk by **Emmanuel Vincent (MULTISPEECH)** "Speech anonymization"
  15h30 → 15h45 Q&A
- **15h35 → 16h10 Break**
- 16h10 —> 16h40 Talk by **Michèle Sebag (TAU)** "Cut the Black Box"
  16h40 → 16h55 Q&A
- 16h55 —> 17h20 Talk by **Luis Galarraga (LACODAM)** "Rule-based explanations in knowledge graphs"
  17h20 --> 17h35 Q&A
- 17h35 —> 18h00 Debriefing with committee

Reminder: project motivation

- Huge current interest for AI – mostly ML
- ML: learn task from examples

2 families of models:

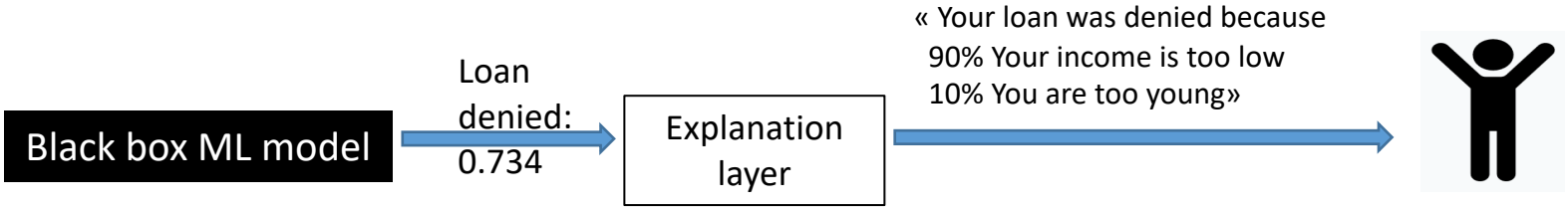
- Subsymbolic – ex: SVM, (Deep) neural networks
- Symbolic – ex: Decision trees, rules

- Predominance of numerical models
  - Better capture the complexity of many tasks
  - More audience

But….numerical model's decisions are hard to understand

- Important issue for many applications. Ex: medicine, justice, ,…
- RGPD: citizens should have the possibility to get explanation

- Recent research trend: explain decisions of « black box » numerical models
  - Numerical model is (mostly) untouched (post-hoc)
  - An upper layer interacts with it to output an understandable (symbolic) explanation



HyAIAI objectives and teams

- **Towards 2-way interactions ML system / human user**

Provide understandable explanations of ML answers to human users

Allow human user to steer the ML system in an understandable way (ex: constraints)

- **Method: hybrid symbolic / numerical approaches**

Complement strengths of both approaches

Combine skills of involved Inria teams

- **Inria teams involved:**

Lacodam (coordination)

Orpailleur

Magnet

mostly symbolic

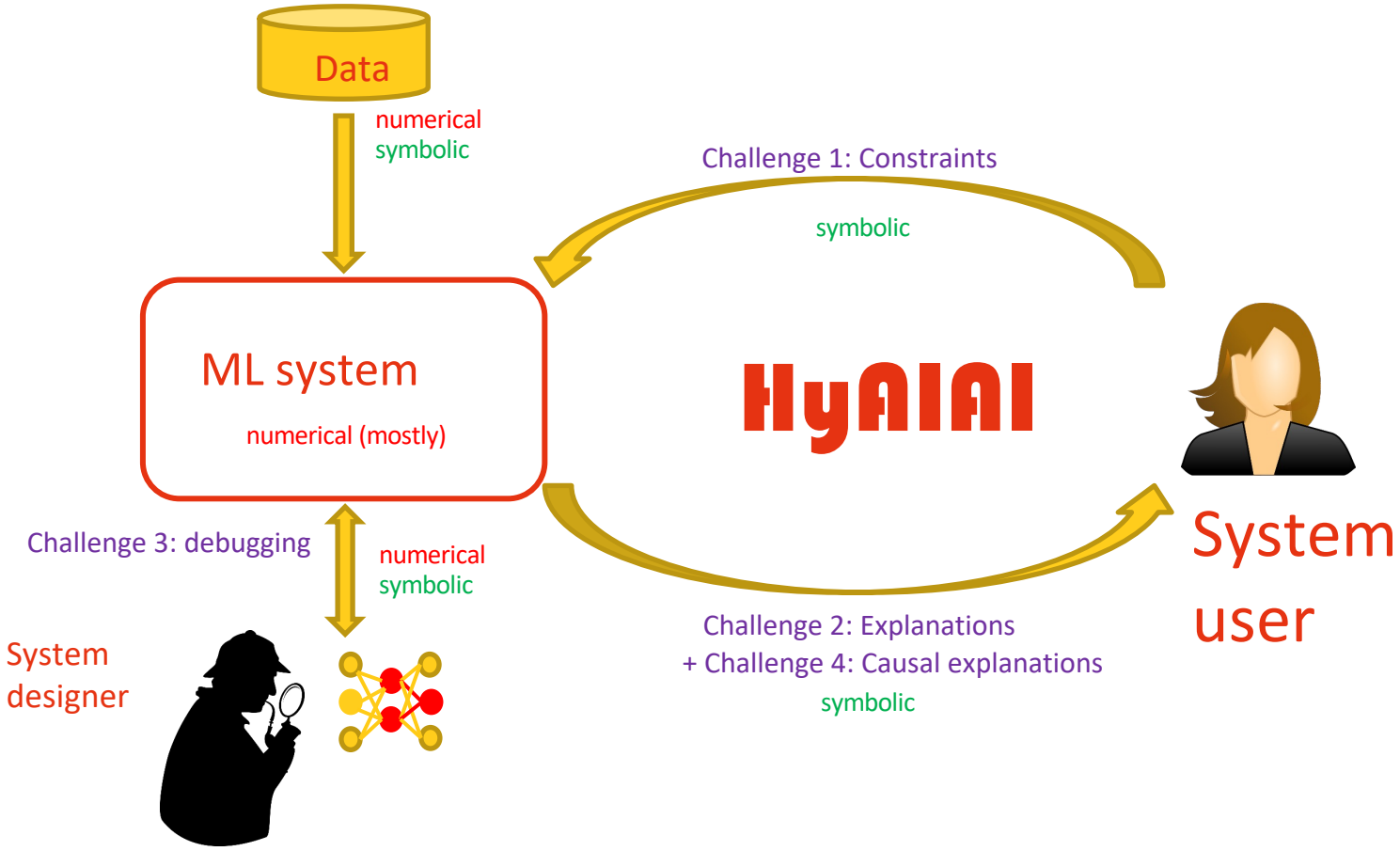
TAU

Multispeech

Scool

mostly numerical

In a Nutshell



Outline

1) HyAIAI cocoon
2) HyAIAI results
3) HyAIAI legacy

Collaborations between INRIA teams

- Orpailleur + Multispeech + Tau: challenge 1 (PhD Georgios Zervakis)
- Multispeech + Lacodam: challenge 3 (post doc Neetu Kushwaha + interns)
- Scool + Lacodam: challenge 3 (post doc Mohit Mittal)
- Magnet + Lacodam: challenge 1 (post doc Carlos Cotrini)
- Orpailleur + Lacodam: challenge 1 & 2 (interns)

- Fail : hiring on **challenge 4**

Meetings of HyAIAI members (4 years)

2023

- 06/27: **HyAIAI Final Evaluation**
- 09/22 ECMLPKDD workshop on Advances in Interpretable Machine Learning and Artificial Intelligence (co-organized by LACODAM members, HyAIAI members in PC)

2022

- 10/17 CIKM workshop on Advances in Interpretable Machine Learning and Artificial Intelligence (co-organized by LACODAM members, HyAIAI members in PC)
- 10/06: Scientific meeting (remote)
- 06/08: Scientific meeting
- 04/20: IDA 2022. A special day on “Explainable AI” with a keynote from Michèle Sebag about “Causal Modeling”.

2021

- 09/24: Scientific meeting and mid-term evaluation
- 09/13: ECML PKDD Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence (co-organized by LACODAM members) (remote)
- 04/23: Scientific meeting (remote)
- 01/25: Scientific meeting (remote)

2020

- 10/19: CIKM workshop on Advances in Interpretable Machine Learning and Artificial Intelligence (co-organized by LACODAM members)
- 09/28: Scientific meeting (remote)
- 05/07: Scientific meeting (remote)
- 01/13: Scientific meeting

2019

- 09/20: ECML PKDD Joint International Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence & eXplainable Knowledge Discovery in Data Mining (co-organized by LACODAM members)
- 09/30: **HyAIAI Kick Off Meeting**

Project Nurturing

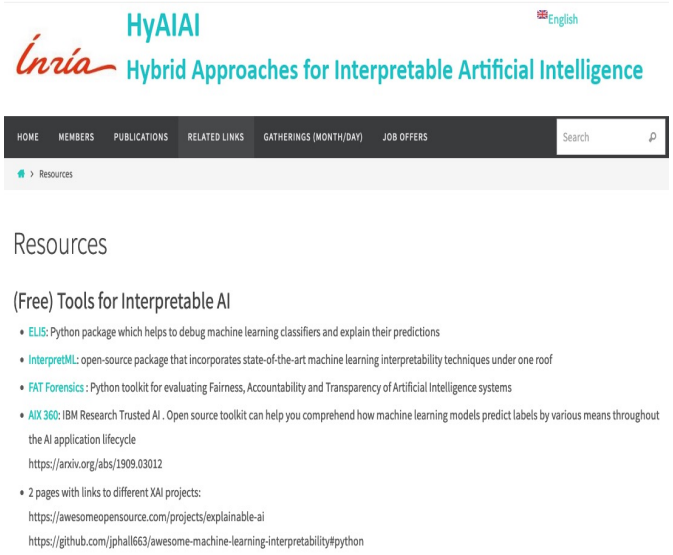
- (challenge 2, Lacodam) IUF, Elisa Fromont
- (challenge 2, Lacodam) ANR FABLE “Framework for Automatic Interpretability in Machine Learning », Luis Galarraga
  - When is a linear attribution explanation more accurate and unambiguous than a rule-based explanation? Is it possible to automate this selection?
  - Can we model the user’s background to achieve this?
- (challenge 2 &4, Scool) EU Chist-era CausalXRL - Causal eXplanations in Reinforcement Learning (with Sheffield et Vienne) ➔ **TALK TODAY**
- (challenge 2, Tau) EU Pathfinder (FET Proactive) project TrustAI “Transparent, Reliable and Unbiased Smart Tool for AI”
- (challenge 1) DFKI- France IMPRESS
- (challenge 1 and 2) EU TAILOR **https://tailor-network.eu/**

Dissemination

- https://project.inria.fr/hyaiai/

- Workshop AIMLAI on “Interpretaple AI” organized **every year** (Lacodam): https://project.inria.fr/aimlai/
- Special Session "Fair and Explainable Models" at EURO 2021 (Orpailleur/Lacodam)
- Feature issue EURO-Journal on **decision process** (Orpailleur/Lacodam)
- Inria-DFKI **Summer school** (Multispeech)
- Within (ICT48) **TAILOR** (Tau, Lacodam, Multispeech, Orpailleur)

- **Scikit-explain** ? ➔ NO but: https://project.inria.fr/hyaiai/related-links/ + FixOUT
- **Challenge** ? ➔ NO did not work out despite the efforts



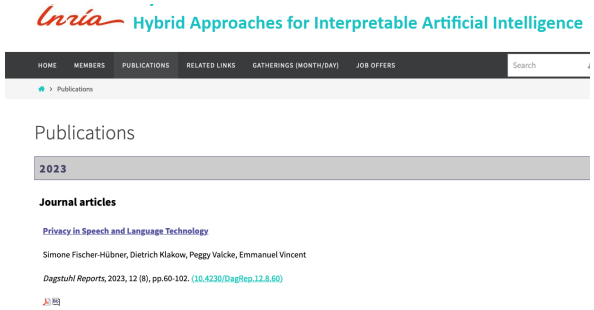
Outline

1) HyAIAI cocoon
2) HYAIAI results
3) HyAIAI legacy

Publications

https://project.inria.fr/hyaiai/publications-and-softwares/

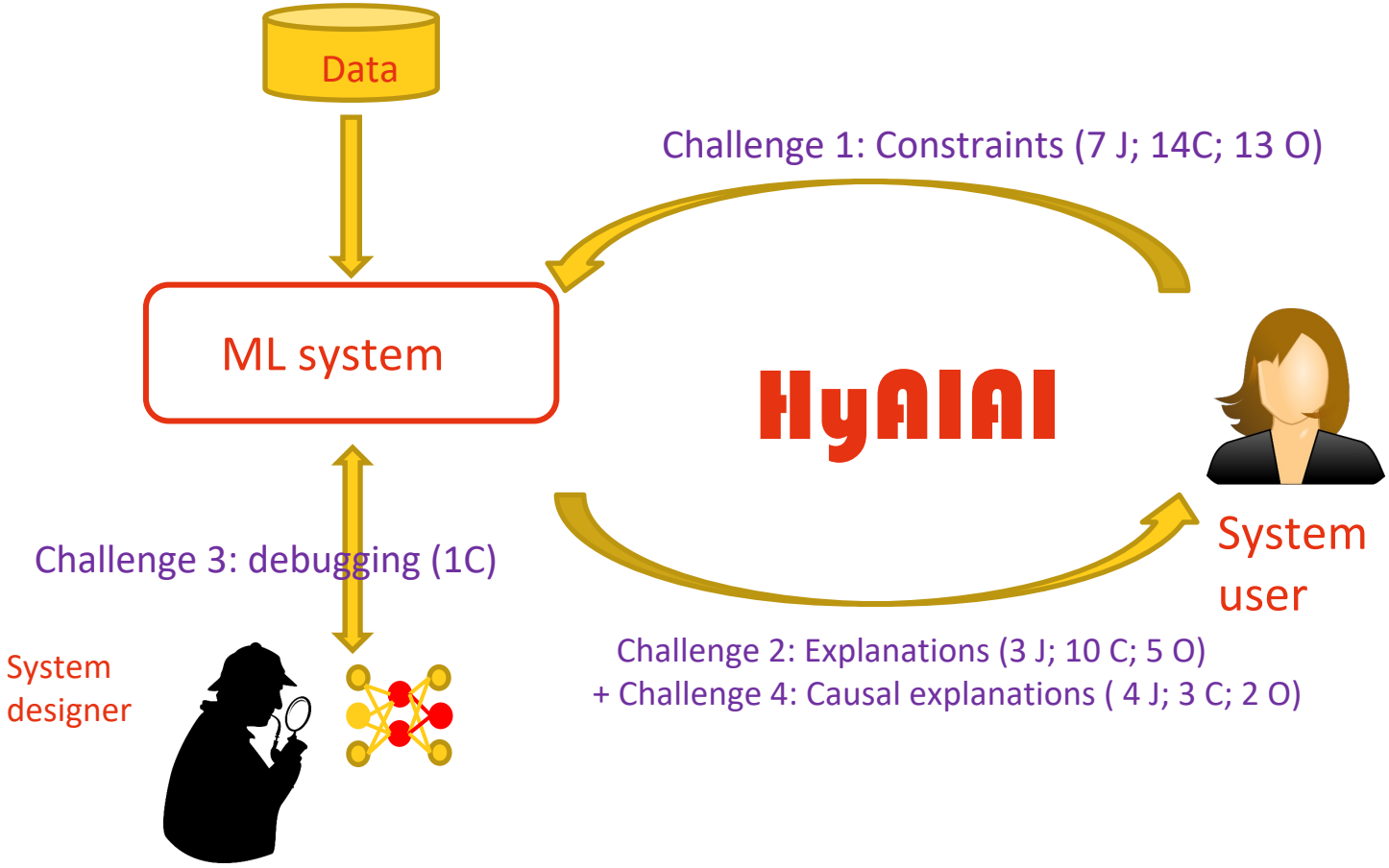
	Around the project	Paid by the project
Journals	14	1
Conferences	28	7
Others	20	5



Selected publications

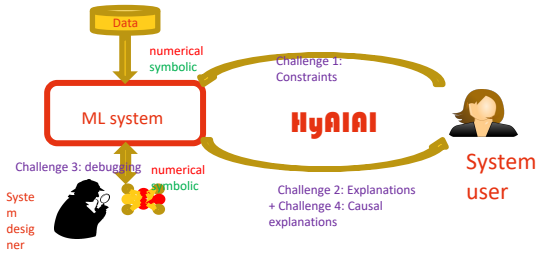
- Steps Towards Causal Formal Concept Analysis Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli ***International Journal of Approximate Reasoning***, 2022
- XEM: An explainable-by-design ensemble method for multivariate time series classification Kevin Fauvel, Elisa Fromont, Véronique Masson, Philippe Faverdin, Alexandre Termier ***Data Mining and Knowledge Discovery***, 2022, 36 (3), pp.917-957.
- An analogy based approach for solving target sense verification **Georgios Zervakis**, Emmanuel Vincent, Miguel Couceiro, Marc Schoenauer, Esteban Marquer *NLPIR 2022 – 6th* ***International Conference on Natural Language Processing and Information Retrieval*****,** Dec 2022, Bangkok, Thailand
- When Should We Use Linear Explanations? Julien Delaunay, Luis Galárraga, Christine Largouët *CIKM 2022 – 31st ACM* ***International Conference on Information and Knowledge Management***, ACM, Oct 2022, Atlanta, United States. pp.355-364,
- VCNet: A self-explaining model for realistic counterfactual generation Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, Alexandre Termier *ECML PKDD 2022 –* ***European Conference on Machine Learning and Knowledge Discovery in Databases****.*, Sep 2022, Grenoble, France. pp.1-16
- s-LIME: Reconciling Locality and Fidelity in Linear Explanations Romaric Gaudel, Luis Galárraga, Julien Delaunay, Laurence Rozé, Vaishnavi Bhargava *IDA 2022 – Symposium on Intelligent Data Analysis*, Apr 2022, Rennes, France. pp.1-13
- Combination of explicit segmentation with Seq2Seq recognition for fine analysis of children handwriting Omar Krichen, Simon Corbillé, Eric Anquetil, Nathalie Girard, Elisa Fromont, Pauline Nerdeux **International Journal on Document Analysis and Recognition**, 2022

Distributions within the challenges



Scientific talks today

- (challenge 1) **Debabrota Basu (SCOOL)** on “Online Instrumental Variable Regression: Regret Analysis and Bandit Feedback”
- (challenge 1) **Jan Ramon (MAGNET)** “Hybrid Approaches for Interpretable Private AI”
- (challenge 1& 2) **Miguel Couceiro (ORPAILLEUR)** “A (de)tour through bias mitigation and analogy based ML”
- (challenge 1) **Emmanuel Vincent (MULTISPEECH)** “Speech anonymization”
- (challenge 2) **Michèle Sebag (TAU)** “Cut the Black Box”
- (challenge 2) **Luis Galarraga (LACODAM)** “Rule-based explanations in knowledge graphs”



Outline

1) HyAIAI cocoon
2) HyAIAI results
3) HyAIAI legacy

Theses (Challenge 2) in Lacodam

1. Yichang Wang "Interpretable Time Series Classification" supervised by Elisa Fromont (defended Sept 2021)
2. Kevin Fauvel "Enhancing Performance and Explainability of Multivariate Time Series Machine Learning" supervised by Alexandre Termier (defended in Dec 2020)
3. Maël Guillemé "Extraction of Interpretable Knowledge from Time Series" supervised by Véronique Masson, Laurence Rozé and Alexandre Termier (defended in Dec 2019)
4. Victor Guyomard "Explainability of decisions taken by Machine Learning algorithms" supervised by Thomas Guyet, Tassadit Bouadi, Françoise Fessant (Orange) and Alexandre Termier (ongoing, **CIFRE Orange**)
5. Julien Delaunay "Automatic Construction of Explanations for AI models" supervised by Luis Galarraga (ongoing, **ANR FABLE**)

Other related theses

- Challenge 1: (Orpailleur) Guilherme Alves "Meta-mining through decision theory for exploratory knowledge discovery" supervised by Miguel Couceiro & Amedeo Napoli
- Challenge 2: (Magnet) Moitree Basu “Integrated privacy-preserving AI”
  - contribution to "utility optimization subject to privacy constraints”
- Challenge 1: (Scool) Matheus Medeiros Centa “Bridging symbolic reasoning and induction”
- Challenge 1: (Scool) Hector Kohler, “Semantic Representations for Interpretable Reinforcement Learning”
- Challenge 2: (Multispeech) Sunit Sivasankaran "Localization Guided Speech Separation"
- Challenge 1: (Magnet + Multispeech) Brij Mohan Lal Srivastava "Speaker Anonymization: Representation, Evaluation and Formal Guarantees »
- Challenge 1: (Magnet + Multispeech) Cennet Oguz "Integrating Lexical and Semantic Knowledge in Multimodal Embeddings for Language-Vision Processing Tasks"

+ Postdocs on related topics

Related (M2) Internships

- (challenge 4) Maturin Videau (Tau) “Discovering Interpretable Reinforcement Learning Policies via Genetic Programming”
- (challenge 4) Alex Westbrook (Tau) ”Black-box model explanation using multi-objective counterfactuals”
- (challenge 2) Rameez Qureshi (Orpailleur/Lacodam) “Tackling unintended bias through reinforcement” (v1)
- (challenge 2) Cindy Pereira (Orpailleur/Lacodam) “Tackling unintended bias through reinforcement” (v2)
- (challenge 3) Christian Bile (Multispeech/Lacodam) “Lie Detector - Can we detect the wrong predictions of Deep Neural Networks?”
- (challenge 1) Esteban Marquer (ORPAILLEUR) “Generating concept lattices using Variational Autoencoders”

Related (M2) Internships

- (challenge 2) Vaishnavi Bhargava (LACODAM + ORPAILLEUR *2) “ Automatic Neighborhood Design for Localized Model-interpretation" ; “ LimeOut: An Ensemble Approach To Improve Process Fairness”
- (challenge 1&2) Mayssaa Zeaiter (ORPAILLEUR) “ A Study about Explainability in Machine Learning and Knowledge Discovery ”
- (challenge 2) Arthur Katossky (M1, Lacodam) “ Local explanation of learned models “
- (challenge 2) Théo Velletaz (M1, Lacodam) “ Interpretation of continuous models “
- (challenge 2) Emielin Visentini (M2, Multispeech) "Extended study of contrastive learning for hate speech detection »
- (challenge 1) Soklong Him (M2, Multispeech + Magnet) “Disentanglement in Speech Data for Privacy Needs”

Future

- Collaboration with LACODAM/MAGNET ongoing (about the post-doc of Carlos Cotrini)
- **CIFRE** PHD with Stellantis in LACODAM “Counterfactual Explanations on Multivariate Time Series”
- Project with **DGA AID** on “explanations for time series” in LACODAM
- **Submitted ANR project** PANDORA (challenge 3) in LACODAM (Phase 2)
- HORIZON EU projects FLUTE & TRUMPET (WP3) on understandable privacy metrics compatible with GDPR and verification strategies for constrained ML (MAGNET)
- **PEPR** Causality in TAU
- **PEPR** IA (projet ADAPTING) in LACODAM
- **PEPR** Cybersécurité (projet iPOP) in Magnet + Multispeech

Merci !

Suivez-nous sur https://project.inria.fr/hyaiai/