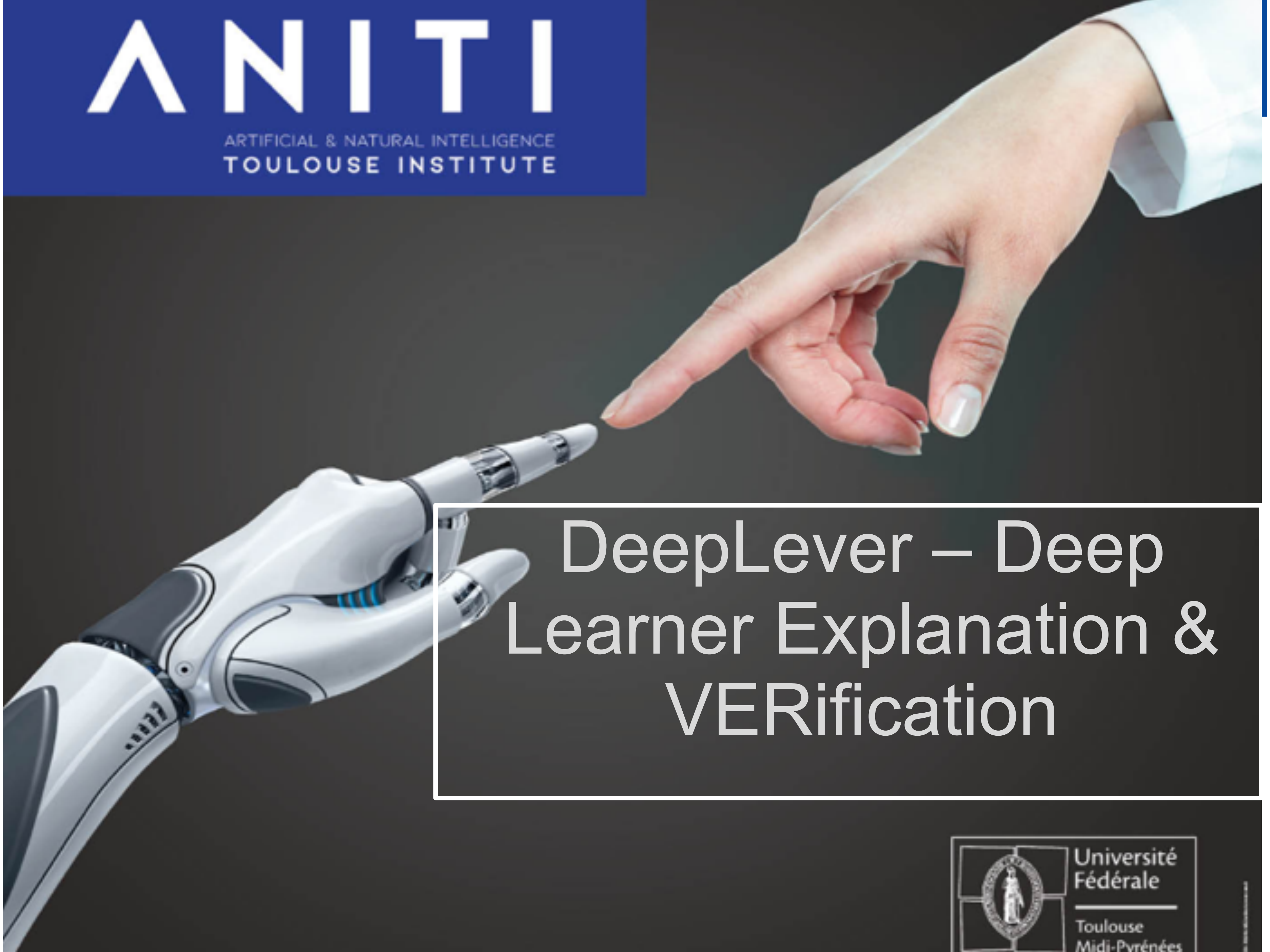
ANITI

ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE



DeepLever – Deep Learner Explanation & VERification

Université Fédérale
Toulouse Midi-Pyrénées

Outline

- **General presentation**
- **Some results**
  - Members
  - Scientific results
  - Related works
  - (Planned PhD / post doc proposals)
- **(Interaction with other chairs / industrial)**

Chair members

Joao Marques-Silva *Université de Lisbonne* Logic and Satisfiability (SAT, MaxSAT, QBF)	**Martin Cooper** *Université Paul Sabatier* Complexity of Constraint Satisfaction Prob. (CSP)	**Emmanuel Hebrard** *CNRS* Cycling

Automated Reasoning

Context

- **"New" vs. "Old" AI**
  - Do we still need logic and models? (debate at AAAI'19)

- **Some arguments for Hybrid AI among the 7 invited talks at IJCAI'19:**
  - *Adnan Darwiche*: Explanation of AI systems via OBDD
  - *Michela Milano*: Merging model-based and data-driven models
  - *Zhi-Hua Zhou*: Deep ML is not necessarily deep NN

- **Verification & Explanation**
  - Using automated reasoning and logic to reason about ML

- **Model Synthesis**
  - Learning via automated reasoning and combinatorial optimization

Explainability

Consider a Machine Learning model as a Boolean function M

- Explanations are *prime implicants* of **M(x) = π**
  - Minimal subset of features that entail the prediction

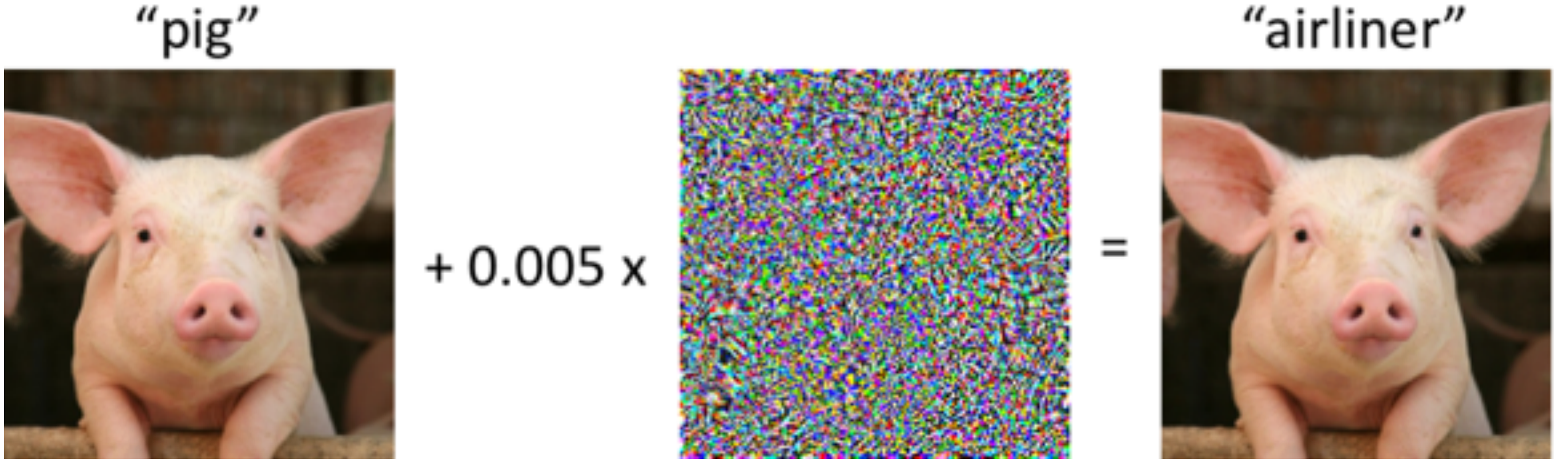
Weekend ∧ **¬(Price = $$$)** ∧ **¬(Estim ≥ 60)** then **Wait**

- Encode the model within a framework and find prime implicants
  - OBDD [Shih, Choi, and Darwiche. IJCAI'2018] polytime (size of OBDD)
  - CNF [Ignatiev, Narodytska, and Marques-Silva. AAAI'2019] SAT
- Verify and repair heuristic explanations (LIME, Anchor)
  - Model counting to verify and explain LIME and Anchor [Naroditska, Shrotri, Meel, Ignatiev and Marques-Silva. SAT'19], which are mostly wrong [Ignatiev, Narodytska and Marques-Silva ArXiv preprint]

Verification

Robustness of learned model M

- Small perturbations do not change the prediction



Find an adversarial example, or prove that none exist

- **Instance I’ such that M(I’) ≠ π at distance ε from I such that M(I) = π**
  - Encode ***M(x) ≠ π AND dist(x,I) ≤ ε*** and query Satisfiability
    - [Katz, Barrett, Dill, Julian and Kochenderfer, CAV’17]
    - [Narodytska, Kasiviswanathan, Ryzhyk, Sagiv and Walsh. AAAI’18]

Duality explanations / counter-examples

[Ignatiev, Narodytska and Marques-Silva NeurIPS'19]

- **Explanation: minimal and entail the prediction π**
- **Counter-example: minimal and entail not π**
  - Any counter-example contradicts every explanation (& vice-versa)

Explanation: **Weekend** ∧ **¬(Price = $$$)** ∧ **¬(Estim ≥ 60)** then **Wait**

  - Counter-example must include either **¬Weekend**, **(Price = $$$)** or **(Estim ≥ 60)**

- **Same hitting-set duality as diagnoses and conflicts** [Reiter 80]
  - Same hitting-set duality used in modern MaxSAT solvers

- **Adversarial examples are counter-examples at distance ≤ ε**

Model Synthesis

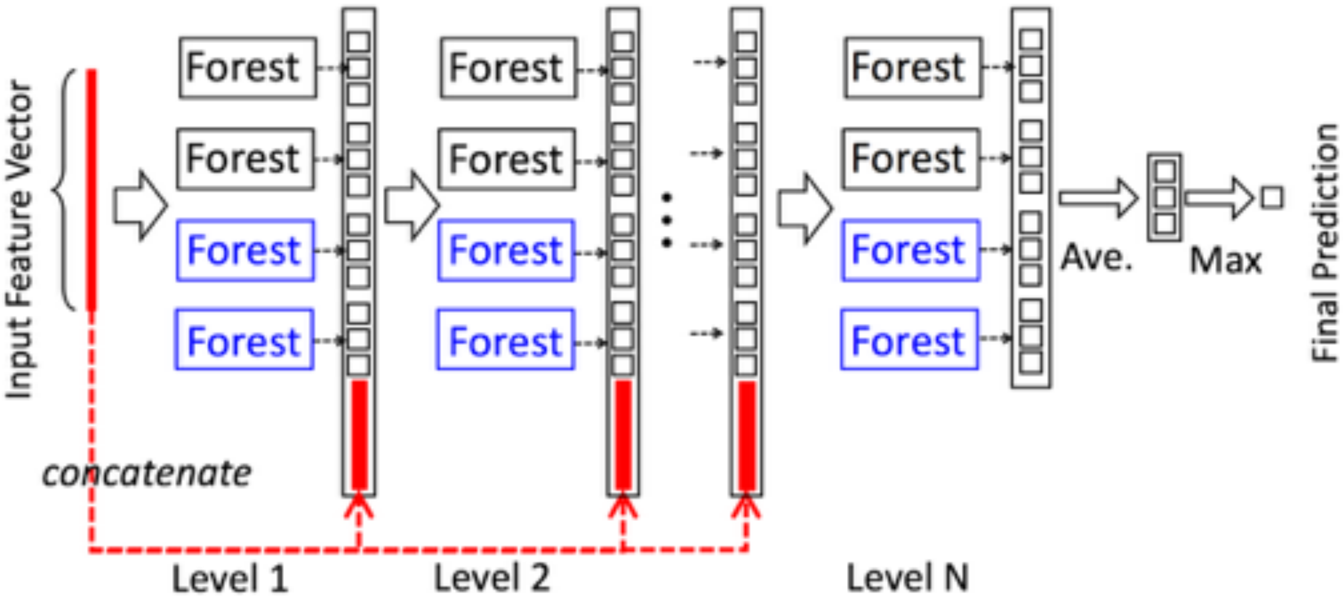
Models again, but to solve the learning problem

- **Binarized neural network via CP and MIP [Icarte, Illanes, Castro, Cire, McIlraith and Beck. CP'19]**
- **Practical? Not really so far (only for tiny networks)**
- **Why using SAT, CP or MIP?**
  - Can potentially provably optimize some objective
  - Can easily add extra constraints, such as fairness [Aïvodji, Ferry, Gambs, Huguet and Siala, soumis à FAT'20]
  - Efficient dedicated methods can be designed
- **State of the art for decision trees**
  - [Bessiere, Hebrard and O'Sullivan. CP'09], [Narodytska, Ignatiev, Peirera and Marques-Silva. IJCAI'18], [Verhaeghe, Nijssen, Pesant, Quimper and Schaus. CP'19]

Model Synthesis

Deep random forest are competitive with deep neural networks

- Deep models and efficient algorithms [Zhou and Feng. IJCAI'17]



Likely that state of the art can be attained or improved via (Max)SAT

- Generate (a lot of) small trees, can encode complex objectives

Model Synthesis

- **Learning logic models**
    - Learning combinatorial structure can be challenging for standard neural networks, e.g. Sudoku [Palm et al. ArXiv preprint]
    - Embed a combinatorial structure (e.g. SAT formula) as a layer of a neural network [Wang, Dolti, Wilder and Kolter ICML'19]
        - Input: subset **I** of variables of the formula
        - Output: complement **V \ I**, *consistent* with the formula
    - Dedicated algorithms for the forward pass **and for the backward pass**
- **Constraint Acquisition [Bessiere, Coletta, Hebrard, Katsirelos, Lazaar, Narodytska, Katsirelos and Walsh, IJCAI'13]: learn a CSP**

Conclusions

- **(Ambitious) Objectives**
    - The problems of explaining, verifying, and learning ML models
    - Complexity, Algorithms, Solvers

- **Possible Interactions**
    - Thomas Schiex
    - Louise Travé-Massuyès
    - Leila Amgoud
    - Jean-Michel Loubes
    - …