What and How of Machine Learning Transparency

Building Bespoke Explainability Tools with Interoperable Algorithmic Components

Welcome!

ECML-PKDD 2020

- Hands-on Tutorial.
- 2.00--6.00pm CEST.
- Events.fat-forensics.org
- FATForensicsEvents.slack.com (Registration via separate URL given in webinar.)
- Recordings published after the event.

Peter Flach

Instructors

- **Kacper Sokol**
  - Researcher at Bristol University
  - Working on Explainable AI
  - Lead developer of FAT Forensics
- **Raul Santos-Rodriguez**
  - Senior Lecturer at Bristol University
  - Working on data science and intelligent systems with applications in healthcare
- **Alexander Hepburn**
  - Researcher at Bristol University
  - Working on cost-sensitive deep learning
  - Core developer of FAT Forensics
- **Peter Flach**
  - Professor at Bristol University
  - Working on human-centred and interactive AI as well as evaluation and calibration of ML models

Schedule

Part 1:
Identifying Modules of Black-box Explainers

2.00--2.15pm CEST (15 minutes)	Background and motivation of research on modular explainers. • Human-centred and interactive artificial intelligence. • Robust and trustworthy machine learning.	Peter Flach
2.15--3.15pm CEST (60 minutes)	Modular interpretability by dissection. • Bespoke surrogate explainers for tabular data and beyond. • The “What?”, “Why?” and “How?” of algorithmic transparency.	Kacper Sokol

Part 2:
Getting to Know FAT Forensics

3.15--3.30pm CEST (15 minutes)	Introduction to open source interpretability with FAT Forensics. • Promises and perils of modular research software. • FAT Forensics -- reproducibility by design.	Alex Hepburn
3.30--3.45pm CEST (15 minutes)	Hands-on session preparation. • Setting up the environment -- Binder, Colab, local installation. • FAT Forensics' documentation -- tutorials, how-to guides, API.	Alex Hepburn
3.45--4.15pm CEST (30 minutes)	Break. • Opportunity to resolve issues with the environment setup. • Sign up for the Slack channel; find a data set; get a black box.	Kacper Sokol & Alex Hepburn

Part 3 (Hands-on): Building Bespoke Surrogate Explainers

4.15--4.30pm CEST (15 minutes)	Introduction to the hands-on resources. • Overview of the Jupyter Notebooks -- building modular surrogates. • Interoperable algorithmic components for ML explainability.	Alex Hepburn
4.30--5.50pm CEST (80 minutes)	Active participation facilitated by the instructors (no setup needed). • Building bespoke surrogate explainers of tabular data. • Bring your own data and explain away.	Kacper Sokol & Alex Hepburn
5.50--6.00pm CEST (10 minutes)	Summary and farewell. • Revisiting modular interpretability with surrogate explainers. • Recap of interoperable transparency software -- FAT Forensics.	Raul Santos-Rodriguez

Background

Where are we coming from?

- AI research in the Intelligent Systems Lab at Bristol combines
  - Data-driven AI (machine learning and data science)
  - Knowledge-intensive AI (reasoning, uncertainty, measurement)
  - Human-centred AI (explainability, human-AI interaction)
- Some examples:
  - Classifier calibration: tutorial last Monday, recording available soon
  - Measurement theory (project funded by the Alan Turing Institute)
    Performance Evaluation in Machine Learning: The Good, The Bad, The Ugly and The Way Forward

The Alan Turing Institute

INTELLIGENT SYSTEMS LAB

Interactive and Human-Centred AI

- As Artificial Intelligence is deployed across an expanding range of scenarios, getting the **interaction** between humans and intelligent machines right is critical.
  - Human agents can play many roles in a data-processing pipeline.
- To achieve **trustworthiness**, Fairness, Accountability and Transparency (FAT) are of paramount importance.
- In Bristol we are particularly interested in informing the AI perspective from other human-centred disciplines
  - Cognitive & social science, philosophy, law, humanities, …

INTERACTIVEAI
Centre for Doctoral Training

Trustworthy AI in Europe

Bristol is a core partner in the TAILOR network of European centres of excellence in AI (https://liu.se/en/research/tailor/), funded by H2020 ICT-48.

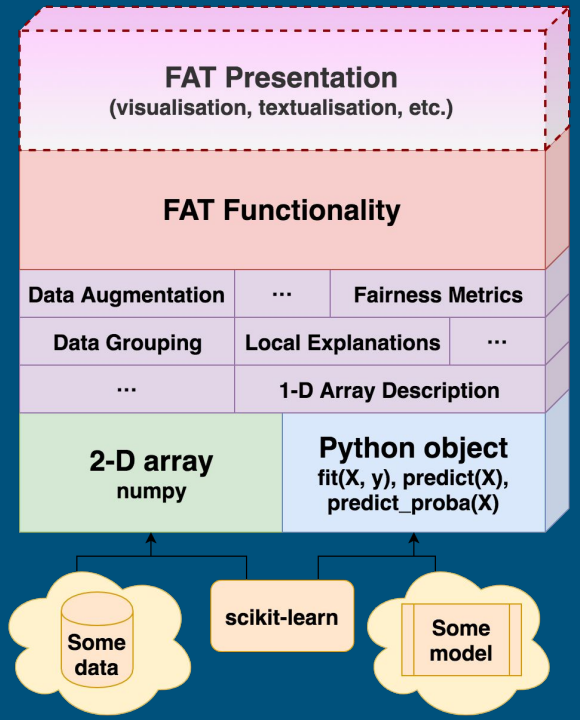
- Trustworthy AI through integrating Learning, Optimisation and Reasoning
- Fundamental and applied research on combining AI paradigms
- PhD curriculum, summer schools, educational events
- Training materials and resources on trustworthy AI

TAILOR has kicked off in September 2020 and is funded for 3 years, so watch this space!

FAT Forensics

FAT Forensics <https://fat-forensics.org>

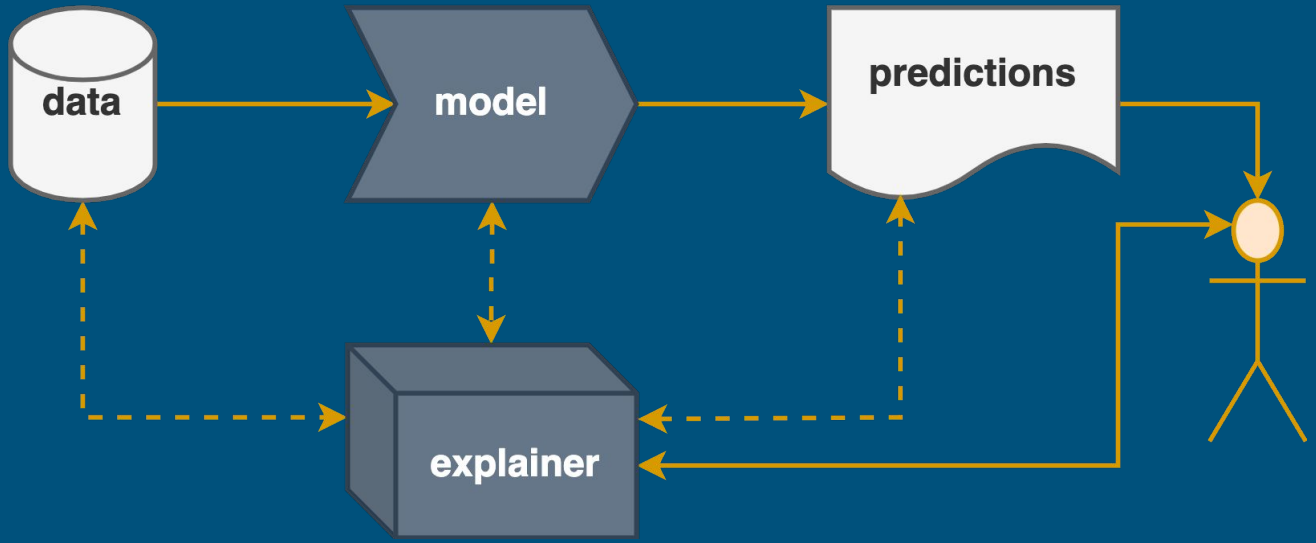
- A modular Python toolkit for algorithmic Fairness, Accountability and Transparency.
- Aimed at both end-users and domain experts.
- Built for research and deployment.
- Originally developed in collaboration with Thales UK.



Motivation

Black-box Explainability

- Explainers can be **black boxes** as well.
- We should be aware of their *algorithmic* assumptions and caveats.



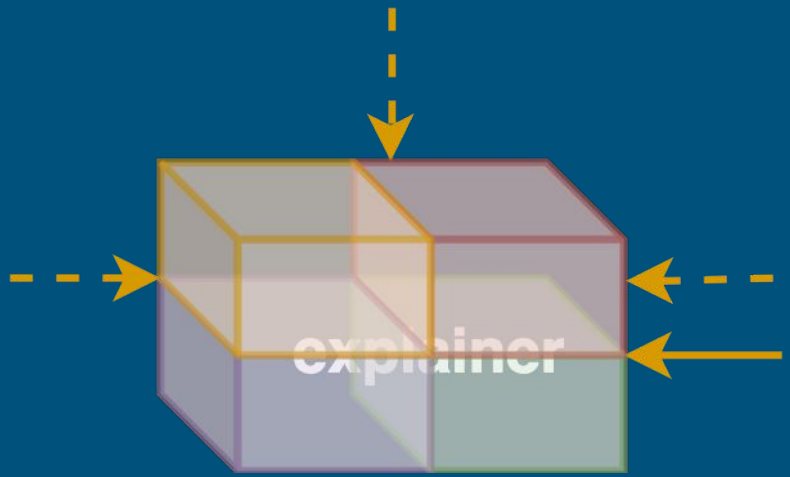
One Explainer Does Not Fit All -- Desiderata

No free lunch (theorem) ⟶ No universal explainer.

- (Perceived) explainability depends on **explainees** and ***use cases***.
- Instead of **end-to-end** explainers, offer explainability **modules**.
- Humans may be the recipients -- they may expect an interactive “dialogue”.
- Additionally, consider: explanation **breadth** and **scope**, explanation **family**, explanatory **medium**, explanation **domain** and explanation **audience** (prior knowledge), among many others.

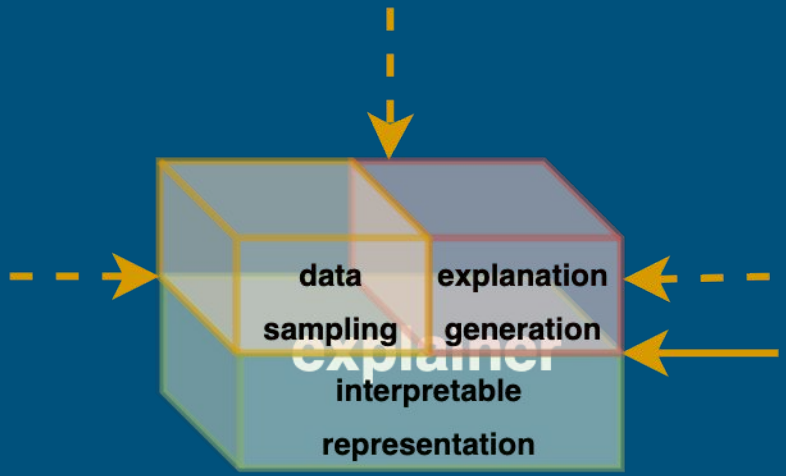
Modular Explainability

- Identify core (algorithmic) building blocks.
- Determine their influence on the resulting explanations -- configure away.



Modular Surrogate Explainers

- We show this process for (local) surrogates of image, text and tabular data.
- The hands-on part focuses on tabular data.



Learning Outcomes

- Understanding explainers, and not only their explanations.
- In-depth, operational appreciation of (local) surrogates.
- Hands-on experience with building and evaluating (local) surrogate explainers for tabular data.

Next Up

What and How of Modular Interpretability

(Kacper Sokol)