ACTL-CPT Conference „Towards eXplainable Artificial Intelligence (XAI) in Taxation: The Future of Good Tax Governance”

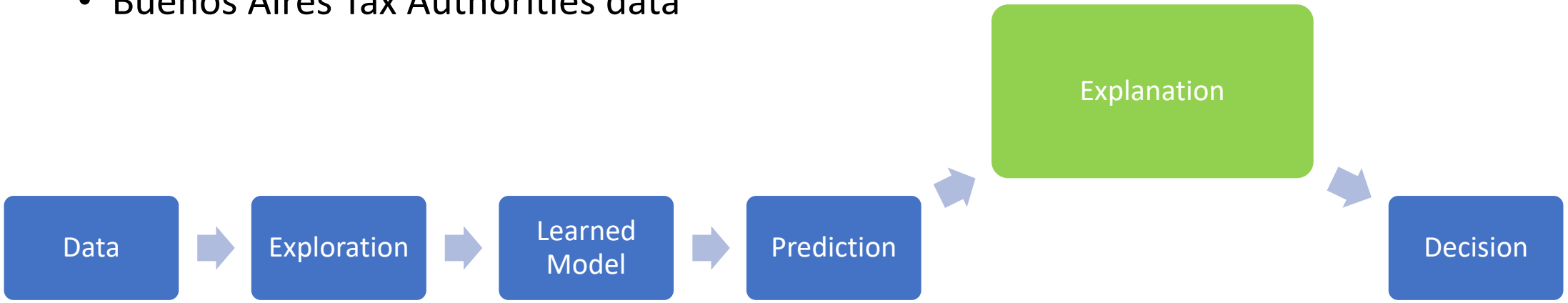
Exploring Explanation Methods in Tax AI: Case Study based on Synthetically Generated Taxpayer’s Data Provided by the Buenos Aires Tax Authorities

Ł. Górski

K. Tyliński

Motivation

- Legal analysis → Technical background
- Usability of current XAI techniques
- Proof-of-concept system
  - Buenos Aires Tax Authorities data



Dataset

- Bueons Aires Tax Authorities
  - Restaurant data for 2021
- Tax fraud risk prediction
  - GTT (ISIB) Tax
- 6465 cases (9% fraudelent)

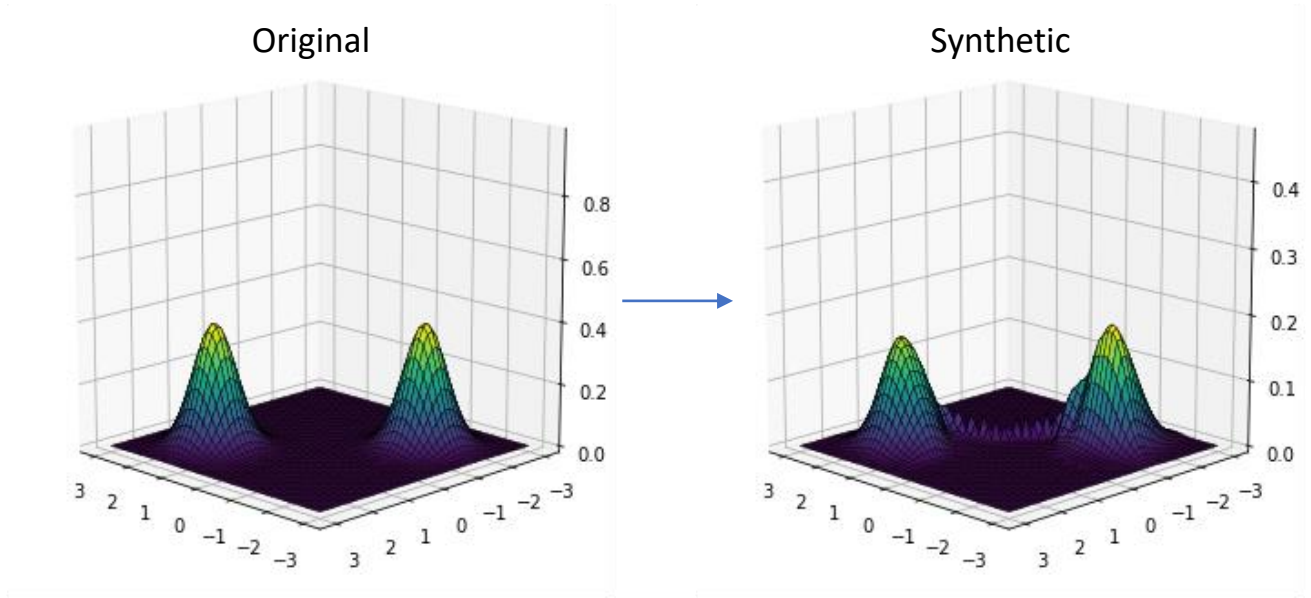
Pesos Sales	Labour Cost Sales	Labour Cost Net Sales	F931	Underreported Work Hours	Incorrect Rate	Excess Deductions	Fraud
(missing)	1	(missing)	1	0	0	0	0
1	1	1	1	1	1	0	0
0	1	1	1	0	1	0	0
1	1	1	1	0	1	0	1

Dataset

- Dataset imbalance
- Missing Data
- Dataset Noisy
  - Repeated data
  - Contradictory data
- Increasing Predictive Power
  - More Features
  - Longer Time Period

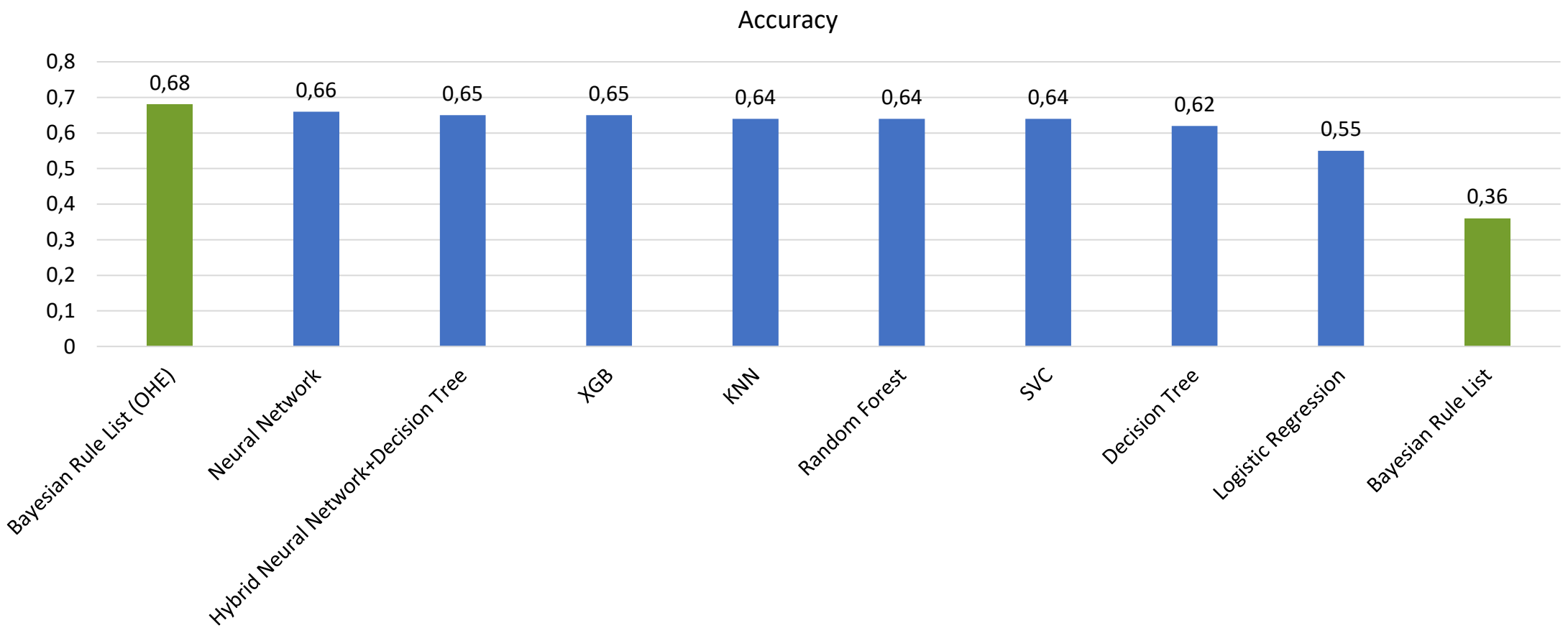
Dataset -> Synthetic Dataset

- Normalizing Flow Algorithm
- Synthetic data distribution ≈ Real life data
  - Privacy issues
  - Human subject research
- 1300 Samples
  - 999 Non-fraudulent
  - 301 Fraudulent



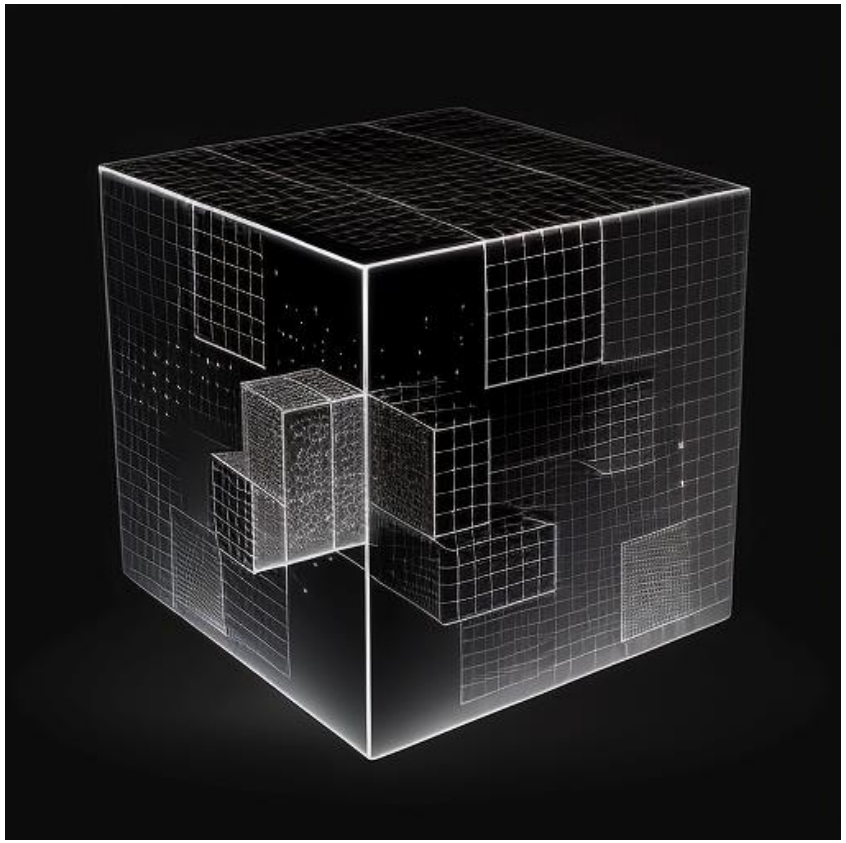
https://dfdazac.github.io/02-flows.html

Classifiers



Explainers

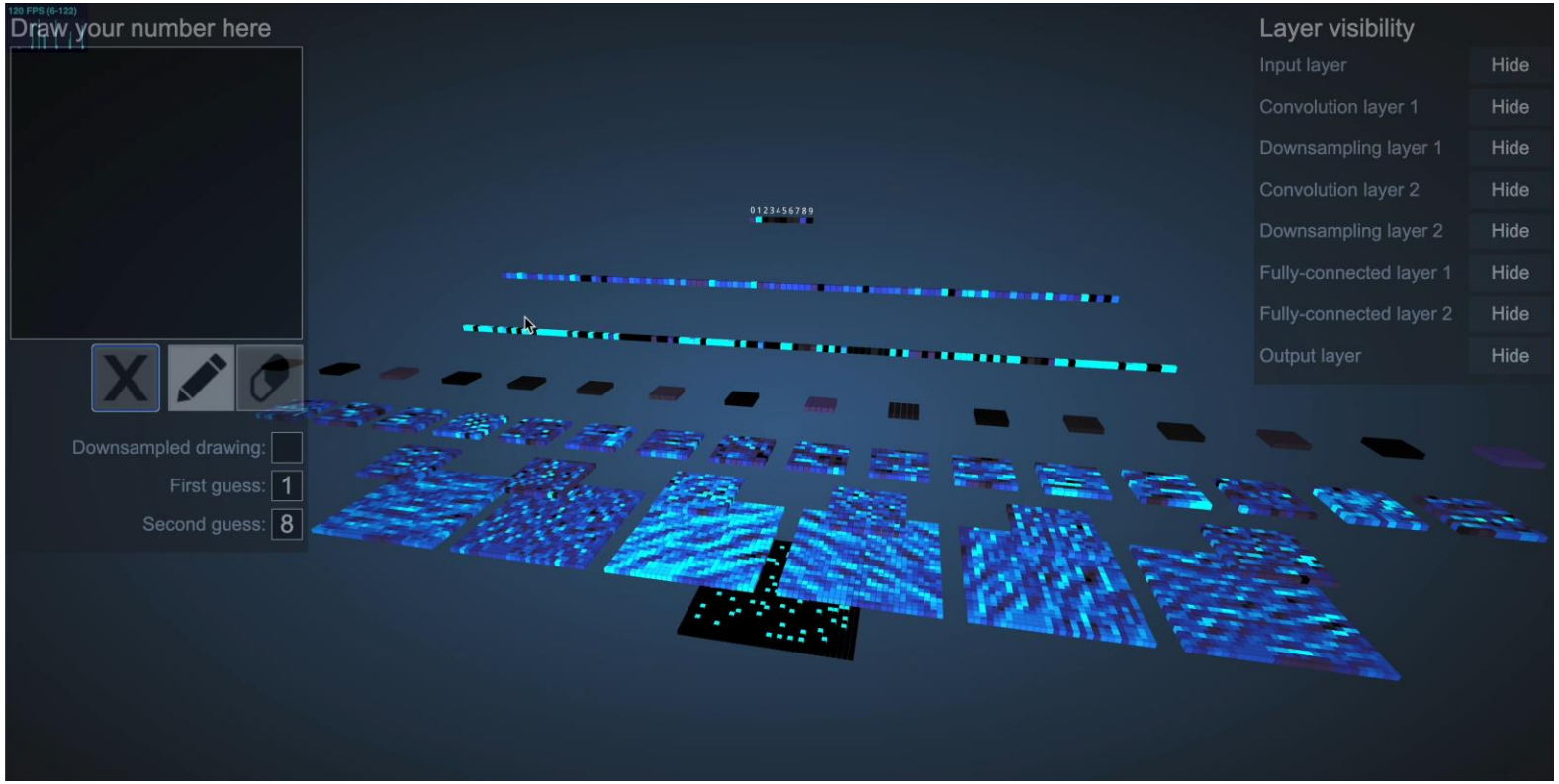
- White box
  - Linear Regression
  - Bayesian Rules Lists
  - Decision Trees
- Black box models
  - Neural network



Midjourney

Explainers

- Neural network
  - Complex Architectures
  - Huge models
  - Input <-> Prediction?
    - Black box
    - Unclear to experts



https://adamharley.com/nn_vis/cnn/3d.html

Sample white box model

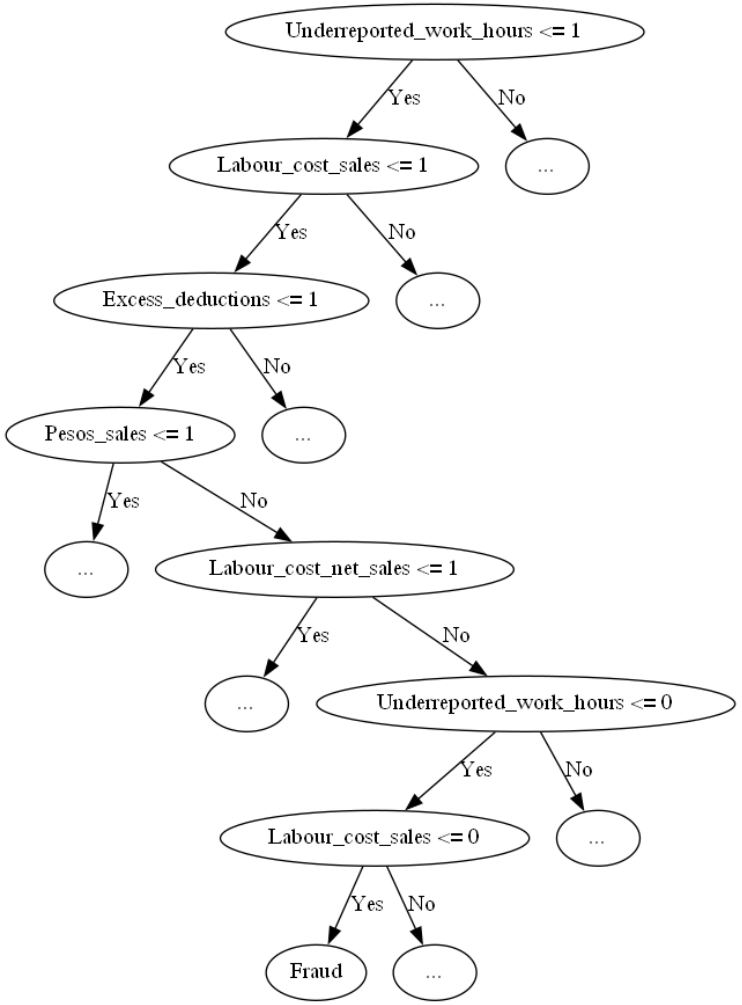
- Bayesian Rule Lists

```
RuleListClassifier Accuracy: 0.36153846153846153
Learned interpretable model: Trained RuleListClassifier for detecting Fraud
============================================
IF Underreported_Work_Hours > 0.5 THEN probability of Fraud: 57.2% (49.3%-65.0%)
ELSE IF Labour_Cost_Net_Sales > 0.5 and Pesos_Sales > 0.5 THEN probability of Fraud: 80.9% (70.8%-89.2%)
ELSE IF Labour_Cost_Sales > 0.5 THEN probability of Fraud: 6.2% (0.2%-21.8%)
ELSE IF Incorrect_Rate > 0.5 THEN probability of Fraud: 61.5% (34.9%-84.8%)
ELSE probability of Fraud: 14.3% (0.4%-45.9%)
============================================
```

```
RuleListClassifier Accuracy: 0.68
Learned interpretable model: Trained RuleListClassifier for detecting Fraud
============================================
IF Labour_Cost_Net_Sales_nan > 0.5 THEN probability of Fraud: 0.1% (0.0%-0.3%)
ELSE IF Incorrect_Rate_1 > 0.5 and Pesos_Sales_1.0 > 0.5 THEN probability of Fraud: 64.1% (62.1%-66.0%)
ELSE IF Excess_Deductions_0.0 > 0.5 THEN probability of Fraud: 55.5% (53.3%-57.7%)
ELSE IF Excess_Deductions_nan > 0.5 THEN probability of Fraud: 73.9% (69.1%-78.3%)
ELSE IF Pesos_Sales_nan > 0.5 THEN probability of Fraud: 6.7% (0.2%-23.2%)
ELSE IF Labour_Cost_Sales_1.0 > 0.5 THEN probability of Fraud: 48.9% (38.8%-59.1%)
ELSE probability of Fraud: 80.0% (66.5%-90.7%)
============================================
```

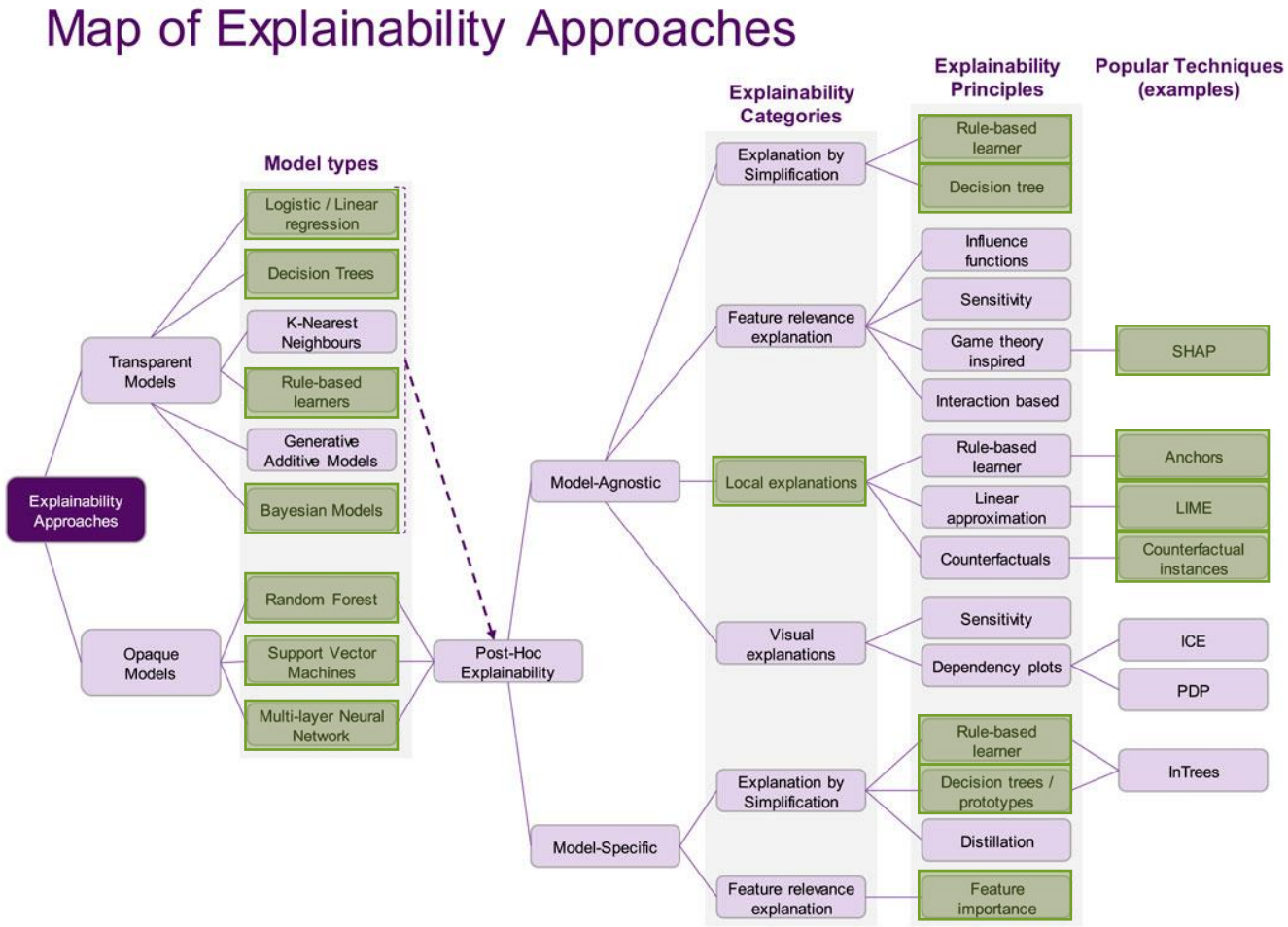
Sample white box model

- Decision tree



Exemplary explanations

- Most popular explanations generation methods:
  - Industrial & Research use-cases
  - Lime
  - SHAP
  - Counterfactuals
  - Anchors



https://www.frontiersin.org/articles/10.3389/fdata.2021.688969/full

SHAP force plot

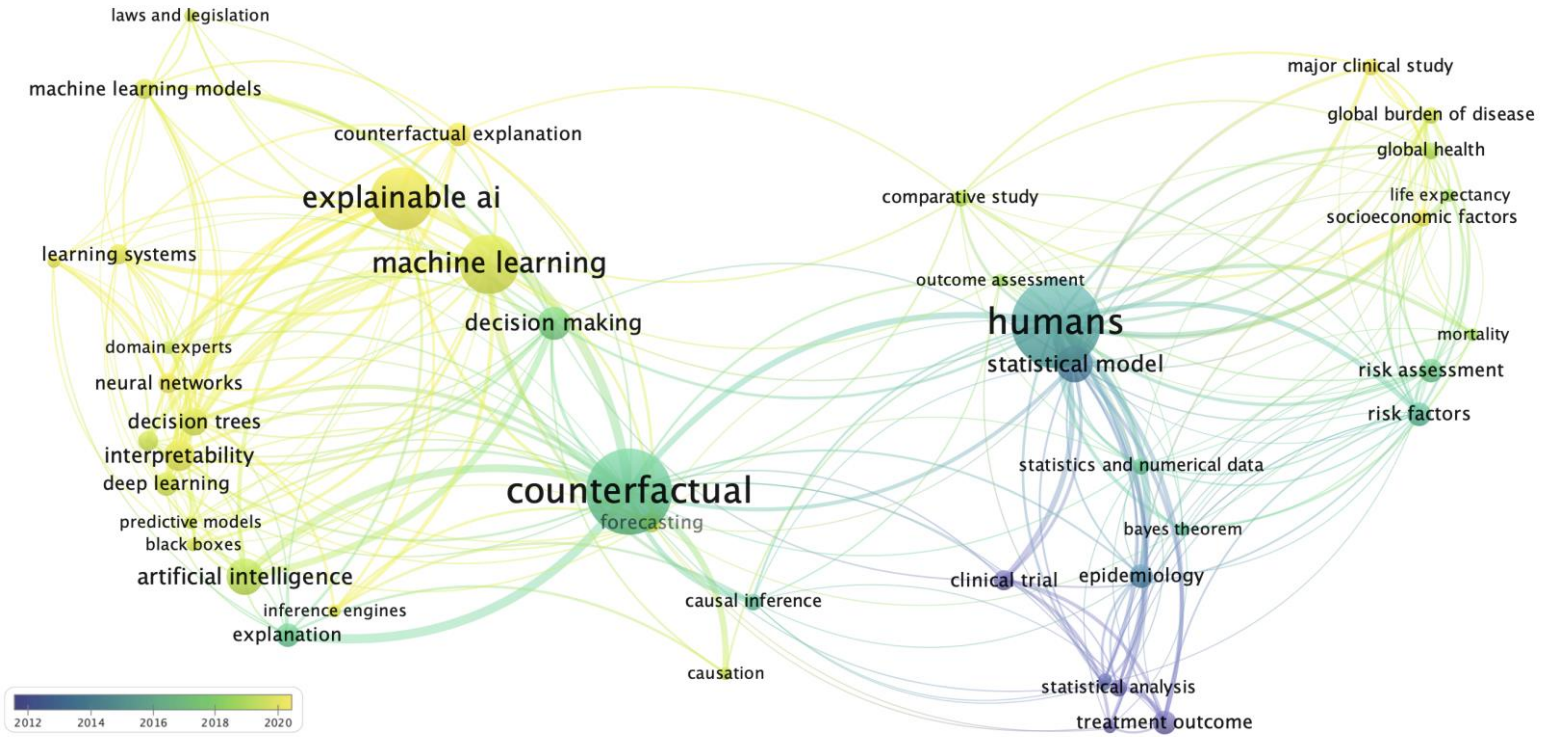
Counterfactuals

- Altering data -> Changing prediction

Pesos Sales	Labour Cost Sales	Labour Cost Net Sales	F931	Underreported Work Hours	Incorrect Rate	Excess Deductions	Fraud
(missing)	0	(missing)	1	0	1	1	1
(missing)	0	(missing)	**0**	**2**	1	1	**0**

Explanations assessment

- Knowledge of system's inner working
- Rule-based most understandable?
- Counterfactuals



https://arxiv.org/pdf/2103.04244.pdf

Conclusion & Future Work

- Dataset expansion
- Exploration of neurosymbolic computation and explainability
  - Knowledge graphs
  - Ontology
  - Structured Natural Languages
- Creation of new algorithms with explainability as a goal
- Explainability vs accuracy