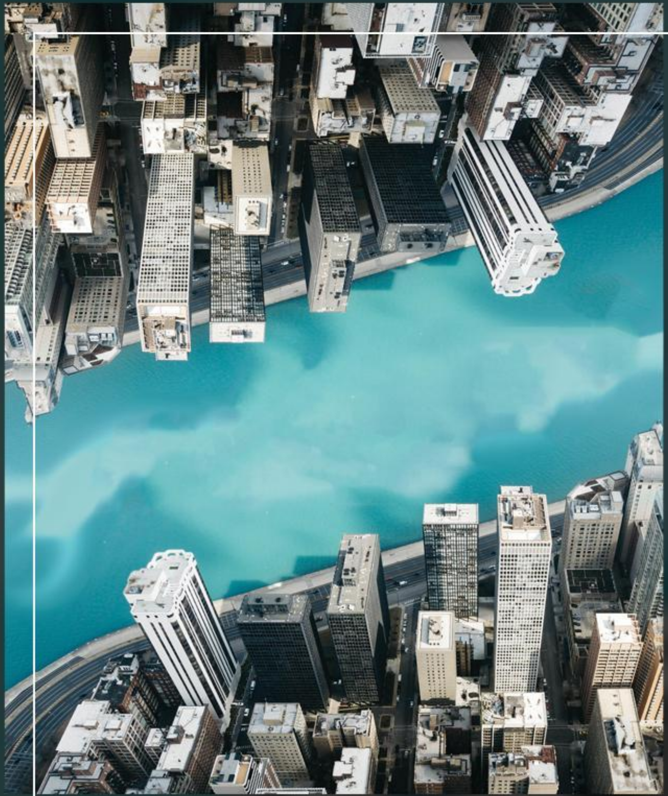
feedzai

UNIVERSIDADE NOVA DE LISBOA

Concept-based Explainability:
Challenges & Applications to Fraud Detection

Vladimir Balayan, Catarina Belém, Pedro Saleiro, Ludwig Krippahl, Pedro Bizarro

Deep Learning Sessions Lisbon - Meetup
2021, July 21st



Problem & Motivation

Problem

feedzai

- The field of eXplainable AI (XAI) aims to tackle the lack of interpretability in ML.

Problem

feedzai

- The field of eXplainable AI (XAI) aims to tackle the lack of interpretability in ML.
- State-of-the-art methods in explainable AI (XAI) either:

Problem

feedzai

- The field of eXplainable AI (XAI) aims to tackle the lack of interpretability in ML.
- State-of-the-art methods in explainable AI (XAI) either:

  1. produce low-level feature attributions explanations that are not suited for non-ML experts (e.g. fraud analyst).

Problem

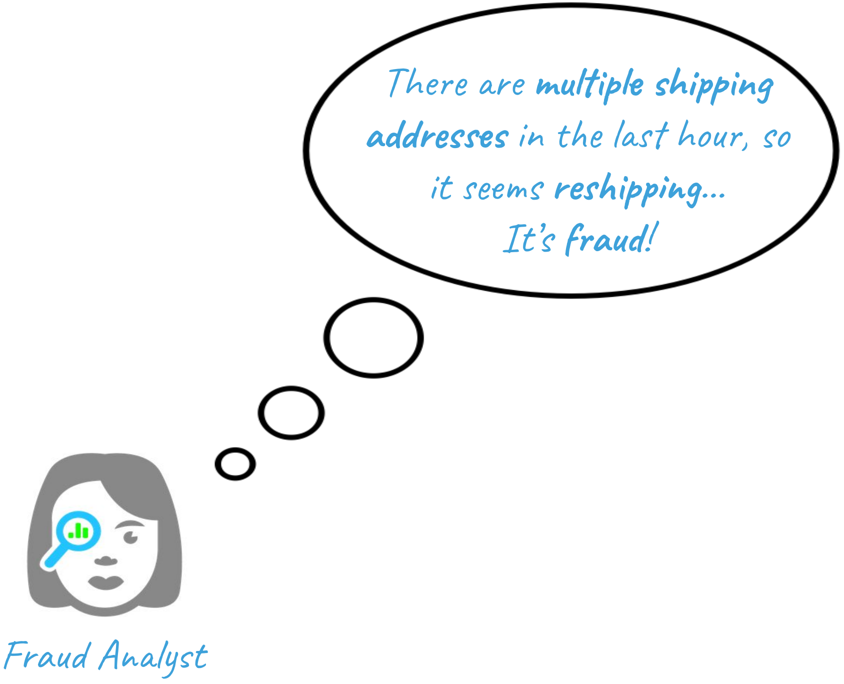
- The field of eXplainable AI (XAI) aims to tackle the lack of interpretability in ML.
- State-of-the-art methods in explainable AI (XAI) either:

  1. produce low-level feature attributions explanations that are not suited for non-ML experts (e.g. fraud analyst).

  Or

  2. produce concept-based explanations that do not work for tabular data.

Domain expert reasoning example



Domain expert reasoning example

feedzai

The ***ideal*** human-interpretable explanation for domain experts provide the ***high-level*** insights about the models' predictions.

Goals

- Develop a self-explainable neural network that jointly learn a predictive task and also associated domain knowledge explanations.

Goals

- Develop a self-explainable neural network that jointly learn a predictive task and also associated domain knowledge explanations.

- Develop a taxonomy of fraud concepts to be used as explanations.

Goals

feedzai

- Develop a self-explainable neural network that jointly learn a predictive task and also associated domain knowledge explanations.

- Develop a taxonomy of fraud concepts to be used as explanations.

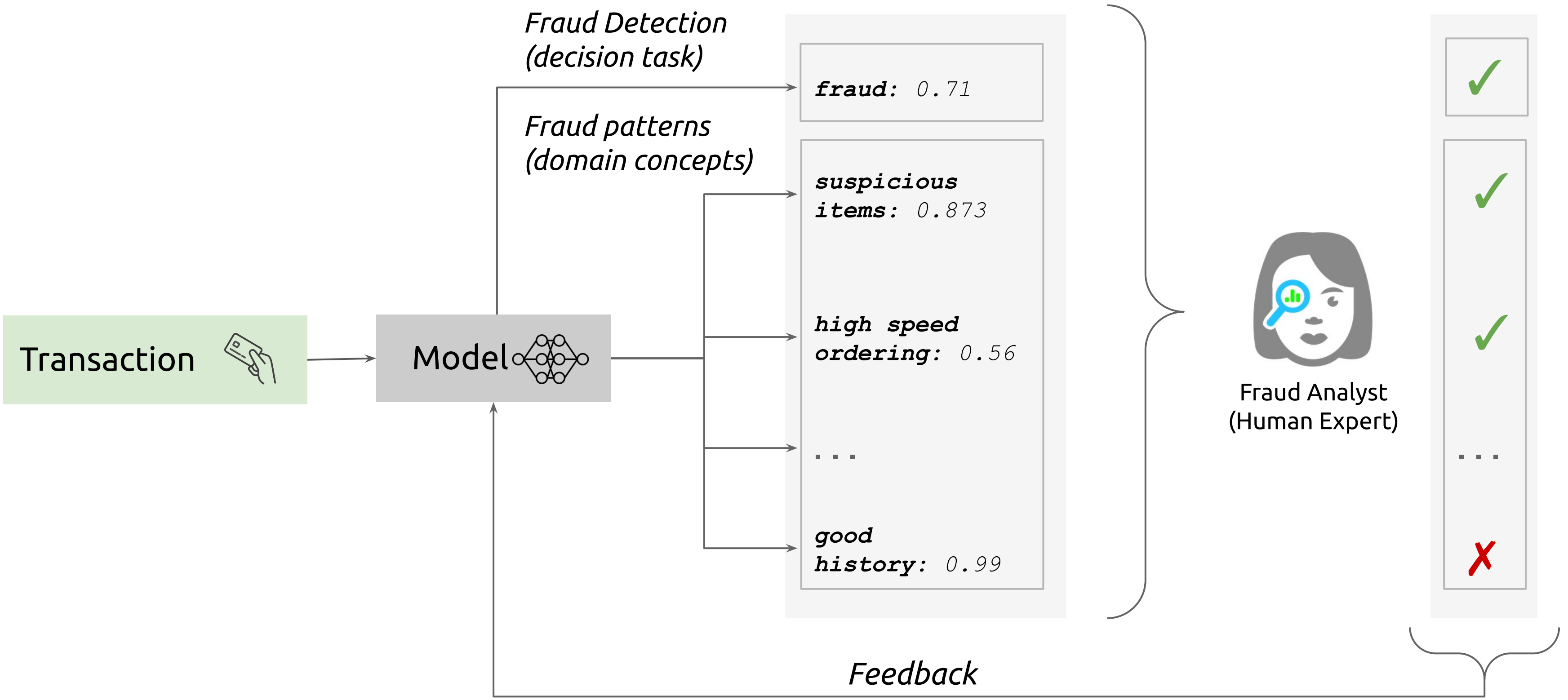
- Leverage the human-in-the-loop feedback to continuously improving both predictive accuracy and explainability.

Goals

- Develop a self-explainable neural network that jointly learn a predictive task and also associated domain knowledge explanations.
- Develop a taxonomy of fraud concepts to be used as explanations.
- Leverage the human-in-the-loop feedback to continuously improving both predictive accuracy and explainability.
- Create a semantic mapping bootstrapping strategy for automatically labeling concept-based explanations dataset.

Proposed solution in a real world fraud detection setting

feedzai



Background & Related Work

XAI Personas - Different personas, different XAI needs...

feedzai

Fields of expertise:

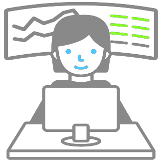
- Domain knowledge
- No ML knowledge

- Data Science + ML

- Know what they want
- No domain nor ML knowledge

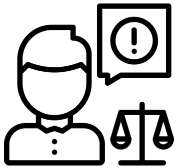
- Regulations/law
- Limited domain & ML knowledge

Human-in-the-loop



Data Scientist

Decision Subject



Regulator

Goals:

- Efficiency
- Better & faster decisions

- Efficiency
- Iterate and debug models

- Improve outcomes
- Reduce friction while feeling safe

- Audit and Assess
- if the system is compliant.

XAI Personas - Different personas, different XAI needs...

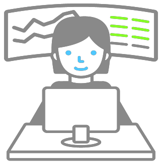
feedzai

Fields of expertise:

Human-in-the-loop

Fields of expertise:
- Domain knowledge
- No ML knowledge

Goals:
- Efficiency
- Better & faster decisions



Data Scientist

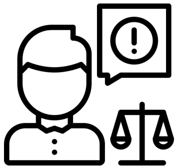
Fields of expertise:
- Data Science + ML

Goals:
- Efficiency
- Iterate and debug models

Decision Subject

- Know what they want
- No domain nor ML knowledge

- Improve outcomes
- Reduce friction while feeling safe



Regulator

- Regulations/law
- Limited domain & ML knowledge

- Audit and Assess
- if the system is compliant.

*The transaction is **suspicious** because MCC = 7801.*

*The transaction is **suspicious** because it contains Suspicious Items.*

XAI Personas - Different personas, different XAI needs...

feedzai

Fields of expertise:

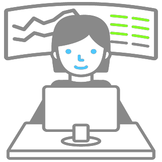
Human-in-the-loop
- Domain knowledge
- No ML knowledge

Data Scientist
- Data Science + ML

Decision Subject
- Know what they want
- No domain nor ML knowledge

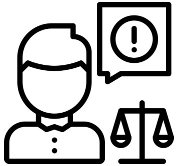
Regulator
- Regulations/law
- Limited domain & ML knowledge

Human-in-the-loop



Data Scientist

Decision Subject



Regulator

Goals:

Human-in-the-loop
- Efficiency
- Better & faster decisions

Data Scientist
- Efficiency
- Iterate and debug models

Decision Subject
- Improve outcomes
- Reduce friction while feeling safe

Regulator
- Audit and Assess
- if the system is compliant.

The optimal choice of explanation depends on the end persona!

Interpretability: When to build

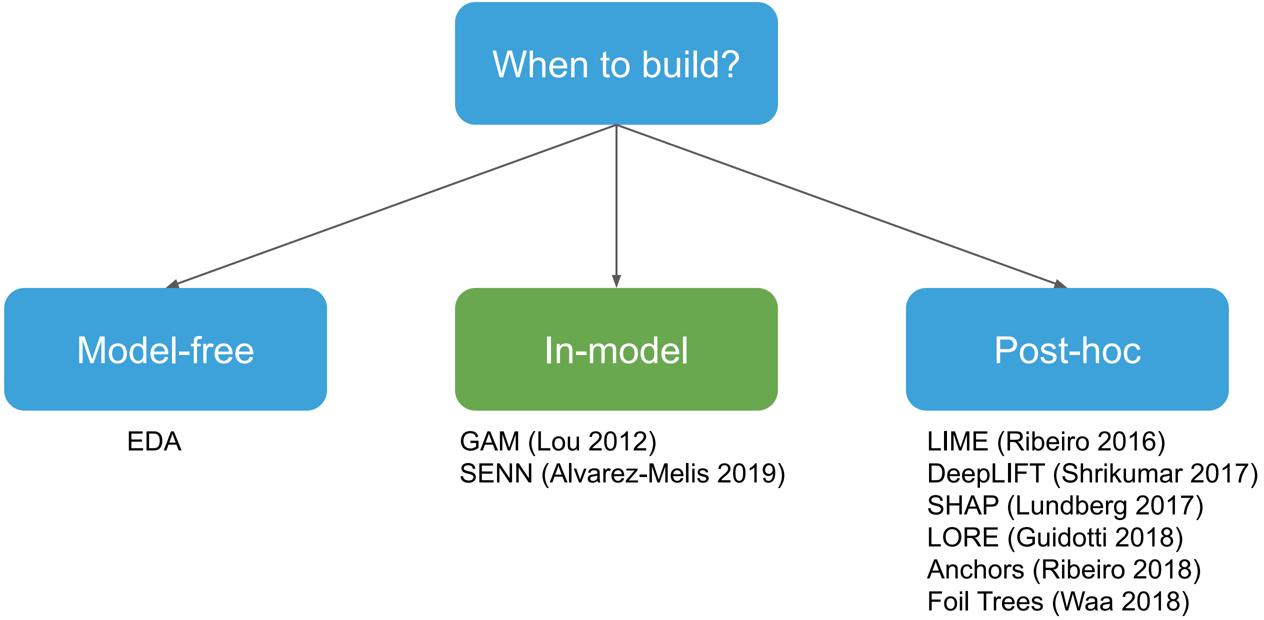
feedzai

When to build?

- **Model-free**
  - EDA
- **In-model**
  - GAM (Lou 2012)
  - SENN (Alvarez-Melis 2019)
- **Post-hoc**
  - LIME (Ribeiro 2016)
  - DeepLIFT (Shrikumar 2017)
  - SHAP (Lundberg 2017)
  - LORE (Guidotti 2018)
  - Anchors (Ribeiro 2018)
  - Foil Trees (Waa 2018)

Interpretability: When to build

feedzai



Taxonomy of Explanations’ Output

feedzai

Explanation output

- Feature-based
  (Greenwell 2017)
  (Ribeiro 2018)
  (Molnar 2019)
- Example-based
  (Dhurandhar 2018)
  (Plumb 2018)
  (Waa 2019)
- Model internals
  (Guo 2018)
- Concept-based
  (Kim 2018)
  (Ghobani 2019)

Taxonomy of Explanations' Output

feedzai

Explanation output

- Feature-based
  (Greenwell 2017)
  (Ribeiro 2018)
  (Molnar 2019)
- Example-based
  (Dhurandhar 2018)
  (Plumb 2018)
  (Waa 2019)
- Model internals
  (Guo 2018)
- Concept-based
  (Kim 2018)
  (Ghobani 2019)

There are many different methods and libraries available

Alibi

TreeInterpreter

DeepLift

LORE

TCAV

SHAP

DiCE

SENN

LIME

Attention

ACE

Anchors

Saliency

GradCAM

GuidedBackprop

Integrated Gradients

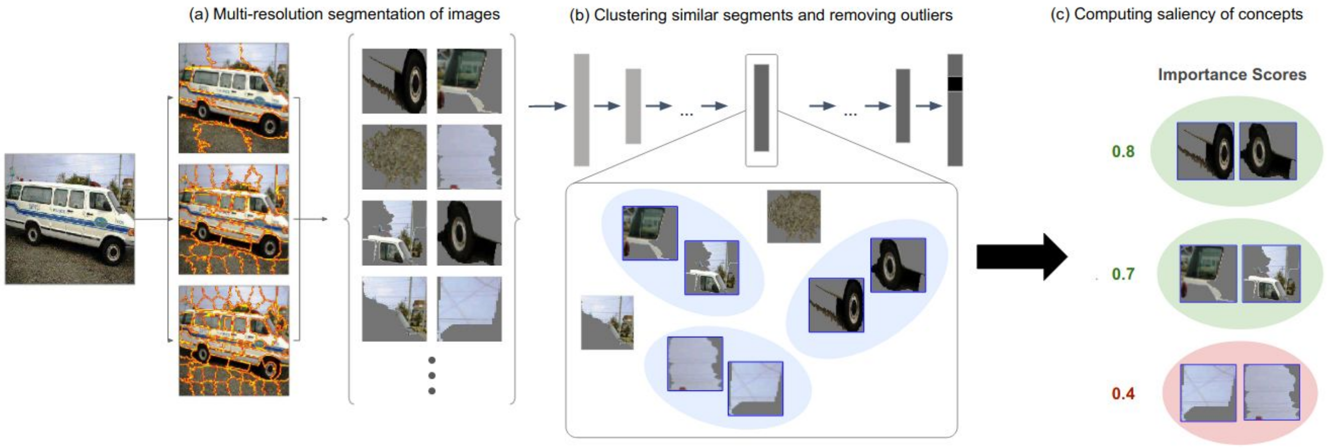
And others…

Automated Concept-based Explanation

Ghorbani, A, NeurIPS 2019, Towards Automatic Concept-based Explanations

- ACE is a global (and local), model-specific and concept-based explanation method that automatically groups input features into high-level concepts.

- The concepts are represented by groups of pixels (segments).



Our Explainability Requirements

To best to our knowledge, there is state-of-the-art XAI method that satisfies our explainability requirements.

In-Model

Local

Concept-based

Tabular data

Proposed Solution

Jointly learned cOncept-based ExpLanations (JOEL)

- JOEL, a NN-based framework to jointly learn a decision-making task and associated domain knowledge explanations.

Jointly learned cOncept-based ExpLanations (JOEL)

feedzai

- JOEL, a NN-based framework to jointly learn a decision-making task and associated domain knowledge explanations.

- JOEL is a self-explainable model, i.e., it incorporates the interpretability architecturally, allowing to produce the decision and also the explanations related to its decision.

Jointly learned cOncept-based ExpLanations (JOEL)

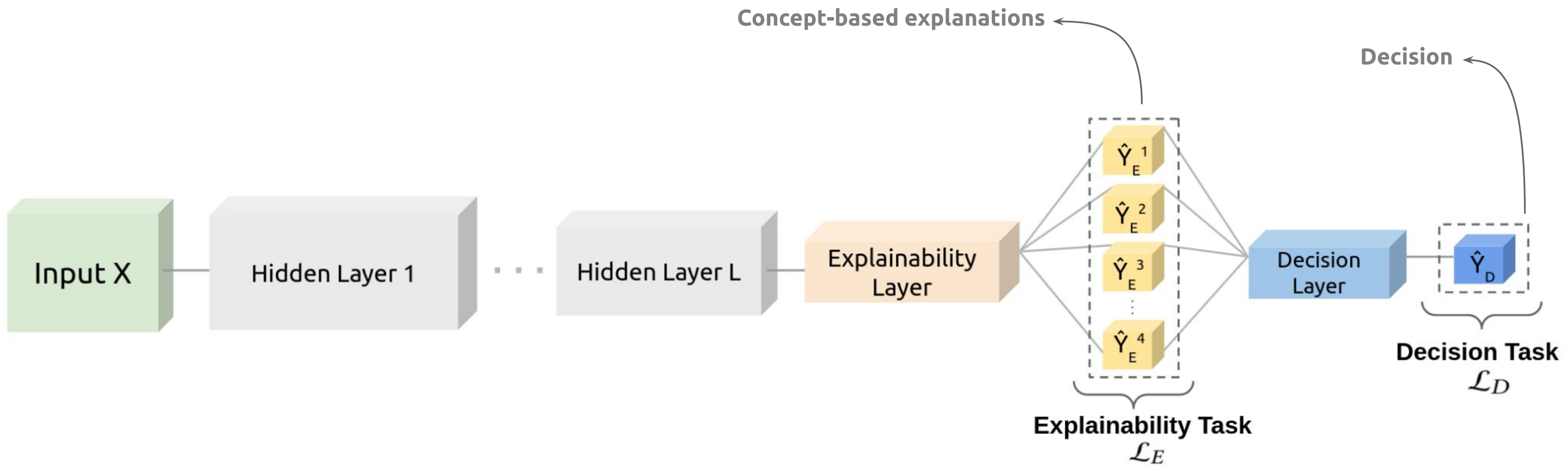
feedzai

- JOEL, a NN-based framework to jointly learn a decision-making task and associated domain knowledge explanations.

- JOEL is a self-explainable model, i.e., it incorporates the interpretability architecturally, allowing to produce the decision and also the explanations related to its decision.

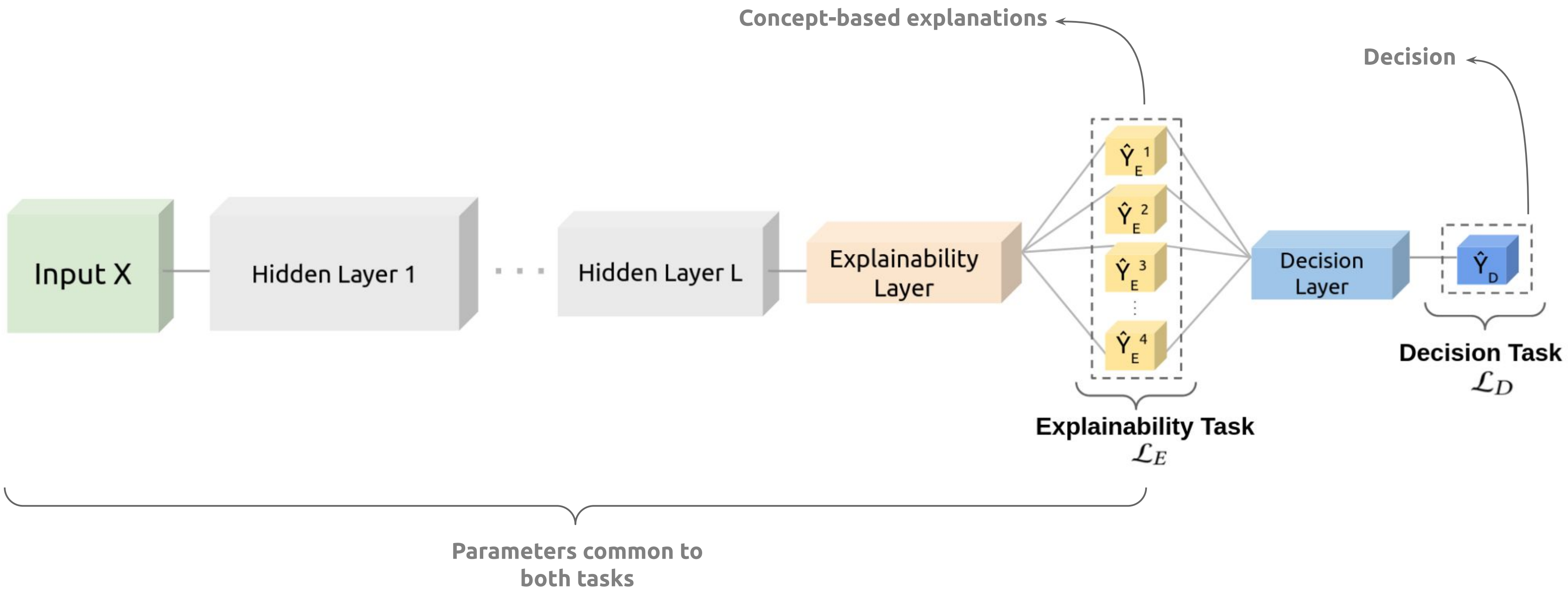
- JOEL provides high-level insights about the model's predictions that very much resemble the domain experts' own reasoning.

JOEL architecture



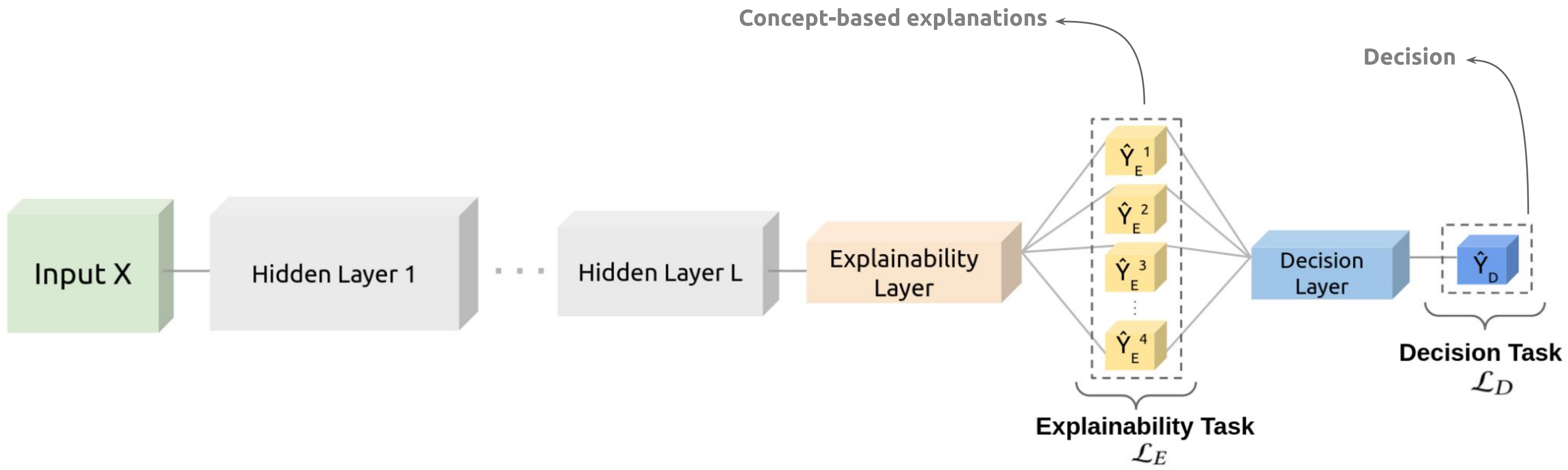
JOEL architecture

feedzai



JOEL architecture

feedzai



$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_E, \boldsymbol{y}_D) = \mathcal{L}_D(\hat{\boldsymbol{y}}_D, \boldsymbol{y}_D) + \mathcal{L}_E(\hat{\boldsymbol{y}}_E, \boldsymbol{y}_E)$$

Concept-Based Explainability and its limitations

Our problem is characterized by having:

- High-resources for the decision task **but** low-resources for the explainability task;
- Out-of-the-shelf domain knowledge (with no added cost).

Research Question:

*Can we **do better** than the **fully supervised** and **low-resources** baseline?*

Challenges

feedzai

Label Scarcity
(insufficient concept labels)

Multi-task Learning
(how to explain the model's predictions?)

- Good DL generalization requires massive datasets;
- Labeling campaigns are arduous and expensive to carry;

- Explainability task should explain the decision task;
- Explanations must reflect the human-in-the-loop's reasoning.

Possible solutions

1. Explore **Weak Supervision** techniques;
   - Can we leverage domain expertise and already existing components in Human-AI system?

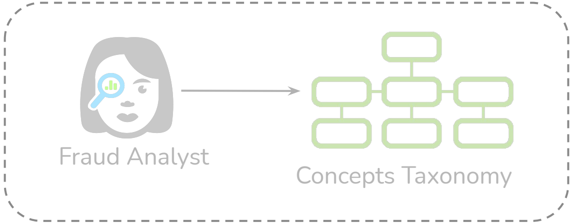
Possible solutions

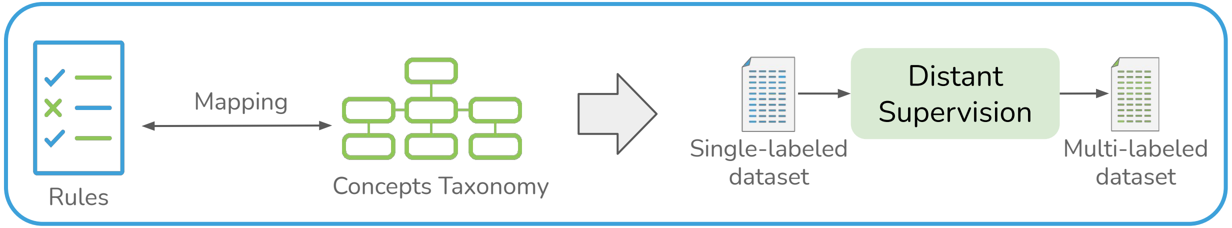
feedzai

1. Explore **Weak Supervision** techniques;
   - Can we leverage domain expertise and already existing components in Human-AI system?

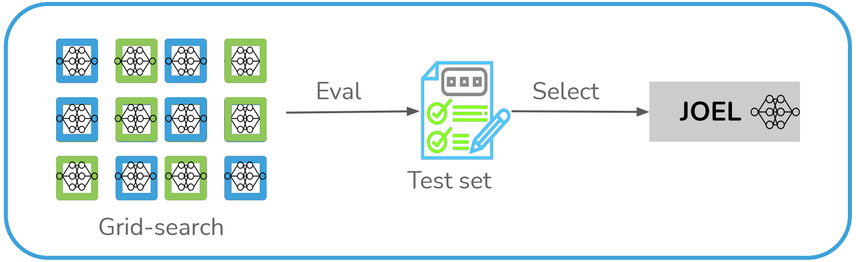
2. Explore different **Learning Strategies**;
   - Use noisy labels only?
   - Use noisy labels and then fine-tune using golden labels?
   - Mix both labels?

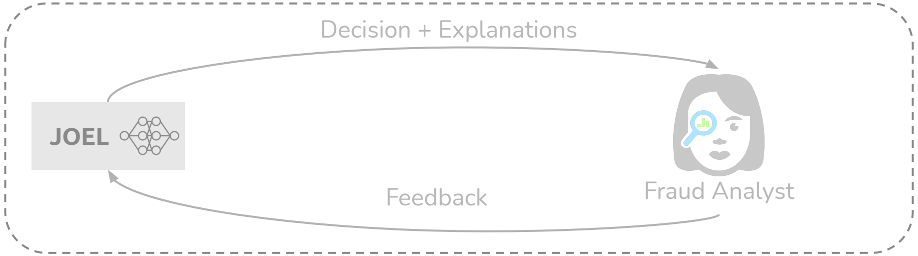
Implementation Workflow





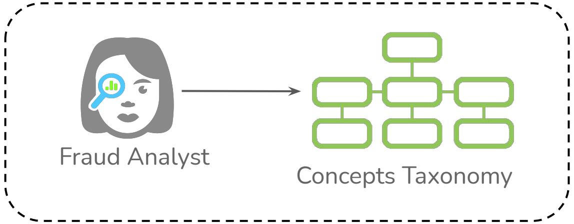
Label Scarcity



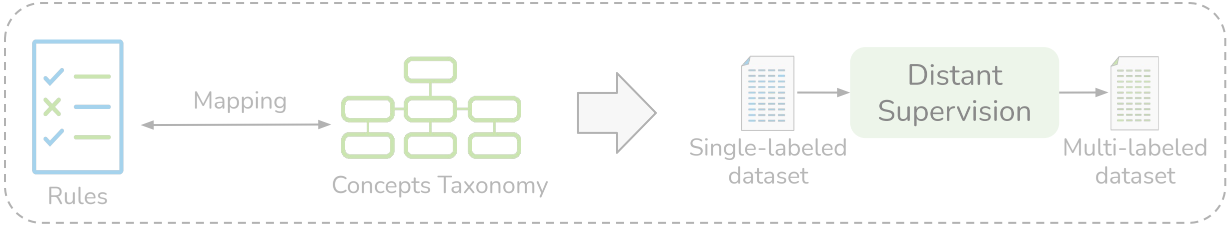


Multi-task learning

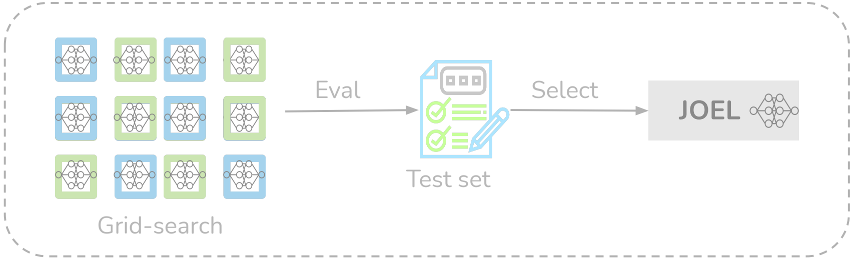
Implementation Workflow



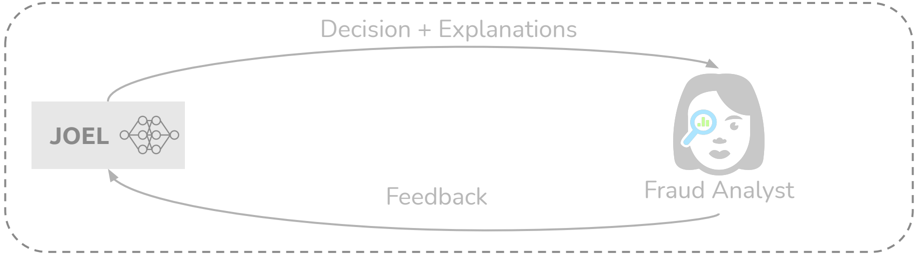
1 **Domain Expert defines concepts that will be used as explanations.**



2 **Apply Distant Supervision.**

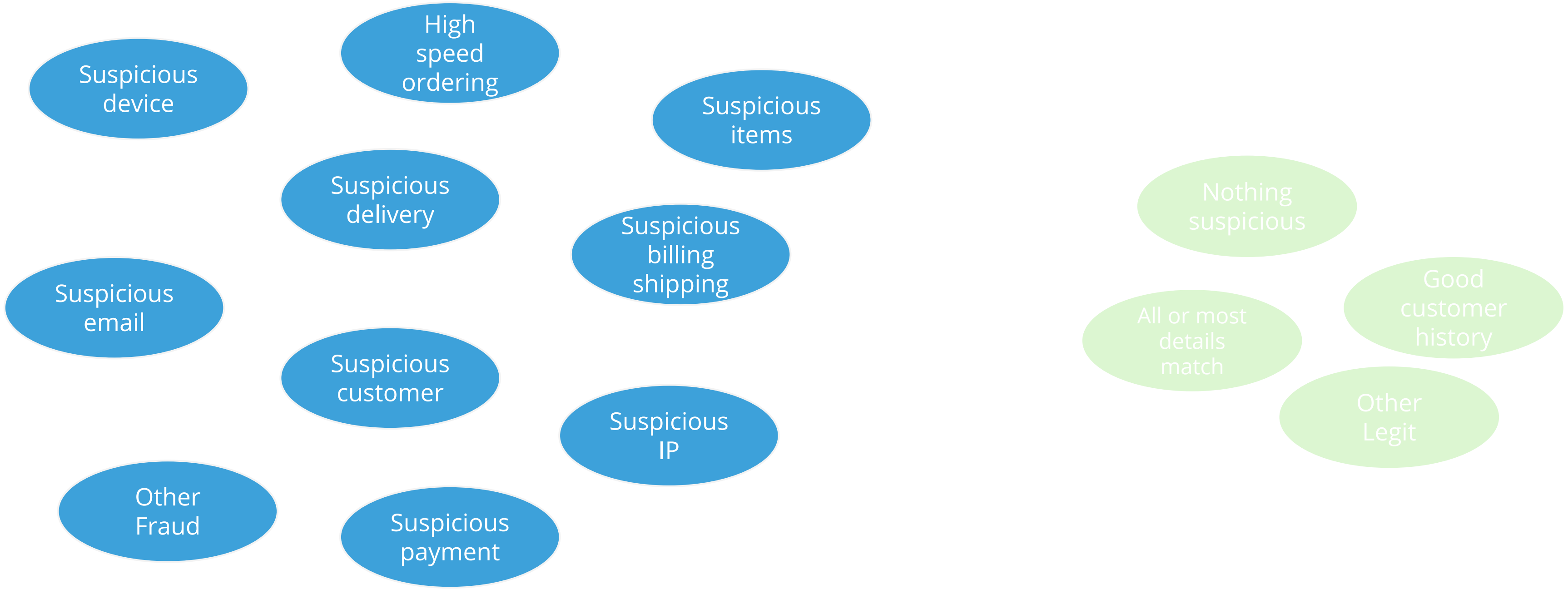


3 **Run a grid search to find the best hyperparameters for Distant Supervision and Supervised Learning.**



4 **JOEL's online loop generates concept-based explanations and collects human feedback in a fraud detection setting.**

Fraud Taxonomy: Fraudulent Concepts

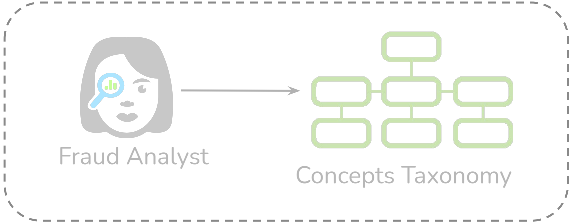


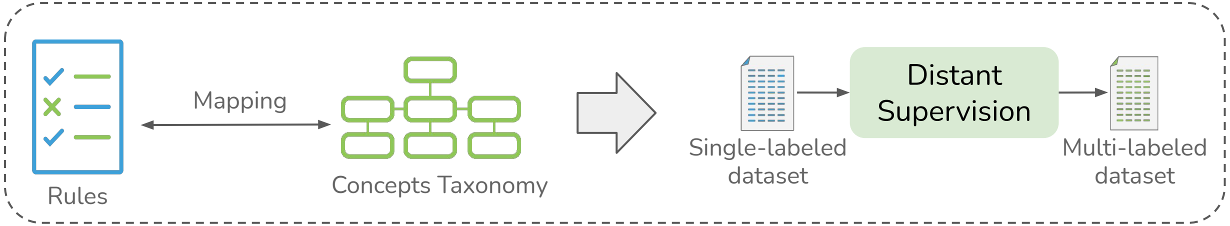
9 Fraudulent concepts (+ **1** Other)

Fraud Taxonomy: Legitimate Concepts

3 Legitimate concepts (+ **1** Other)

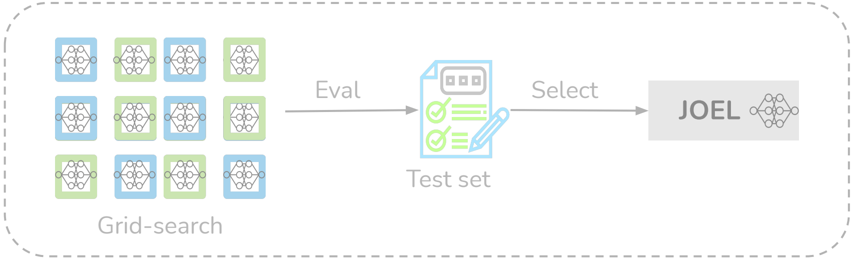
Implementation Workflow

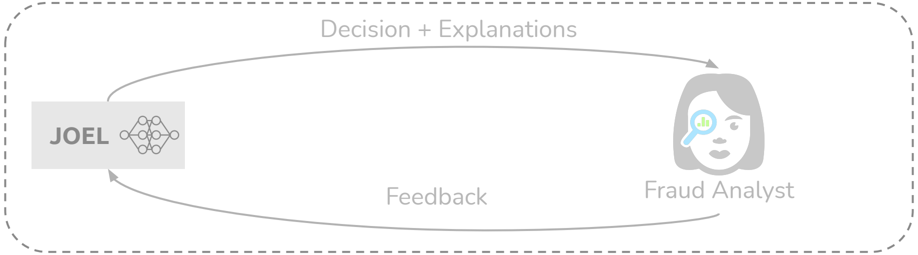




1 Domain Expert defines concepts that will be used as explanations.

2 **Apply Distant Supervision.**





3 Run a grid search to find the best hyperparameters for Distant Supervision and Supervised Learning.

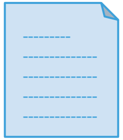
4 JOEL's online loop generates concept-based explanations and collects human feedback in a fraud detection setting.

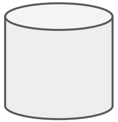
Tackle Label Scarcity through Weak Supervision

feedzai

1. Apply a **Distant Supervision** technique using available domain knowledge.

Rules

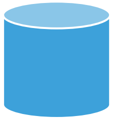




Unlabeled Data
(no concept labels)

Rule-to-Concept Mapping

Rule description	Mapped concepts
Order contains risky product styles.	Suspicious Items
User tried n different cards last week.	Suspicious Customer, Suspicious Payment



Noisy Labeled Data
(weak concept labels)

Concepts Taxonomy

Rules-Concepts Mapping for Distant Supervision

feedzai

- One rule can be associated with one or more fraud concepts.

Rules-Concepts Mapping for Distant Supervision

- One rule can be associated with one or more fraud concepts.

- By having a Rules-Concepts mapping, we can use “noisy labels” (mapped concepts) for each transaction using the rules that were triggered.

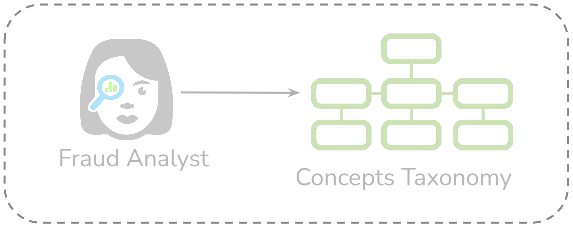
Rules-Concepts Mapping for Distant Supervision

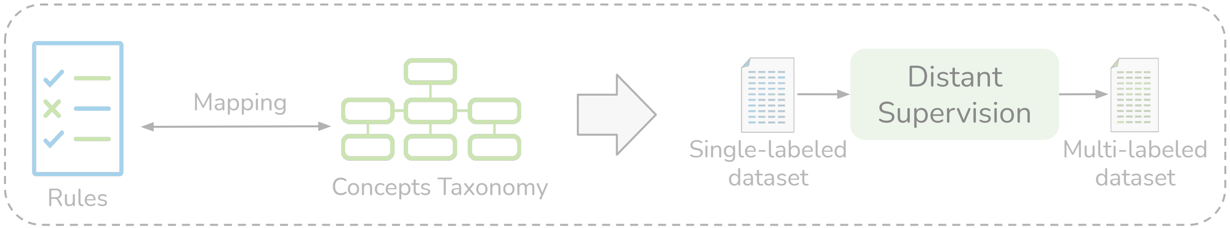
- One rule can be associated with one or more fraud concepts.

- By having a Rules-Concepts mapping, we can use “noisy labels” (mapped concepts) for each transaction using the rules that were triggered.

Canonical rule name	Description	Associated concepts
amount_rule	Transaction amount larger than $N	Suspicious Items
expl_shippingip_mismatch	Shipping and IP mismatch	Suspicious IP, Suspicious billing shipping
jp__card_id_count_N__K_m	The payment card id was detected more than N times in less than K minutes.	Suspicious Payment, High speed ordering

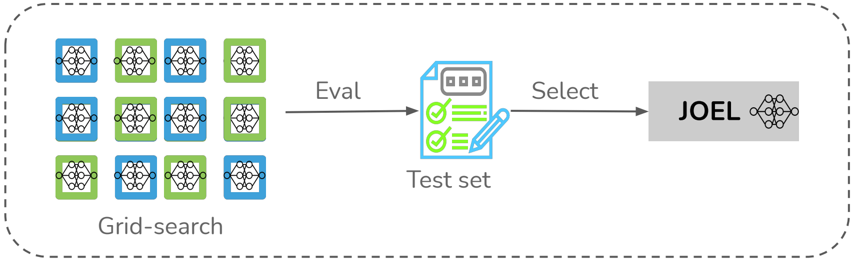
Implementation Workflow





1 **Domain Expert defines concepts that will be used as explanations.**

2 **Apply Distant Supervision.**



3 **Run a grid search to find the best hyperparameters for Distant Supervision and Supervised Learning.**

4 **JOEL's online loop generates concept-based explanations and collects human feedback in a fraud detection setting.**

Jointly learn to explain and decide: Learning Strategies

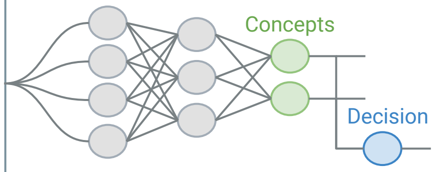
feedzai

Data → Learning Strategies → Concept-based Explainable Model

Golden Labels

Noisy Labels

A Fully Supervised (Baseline)

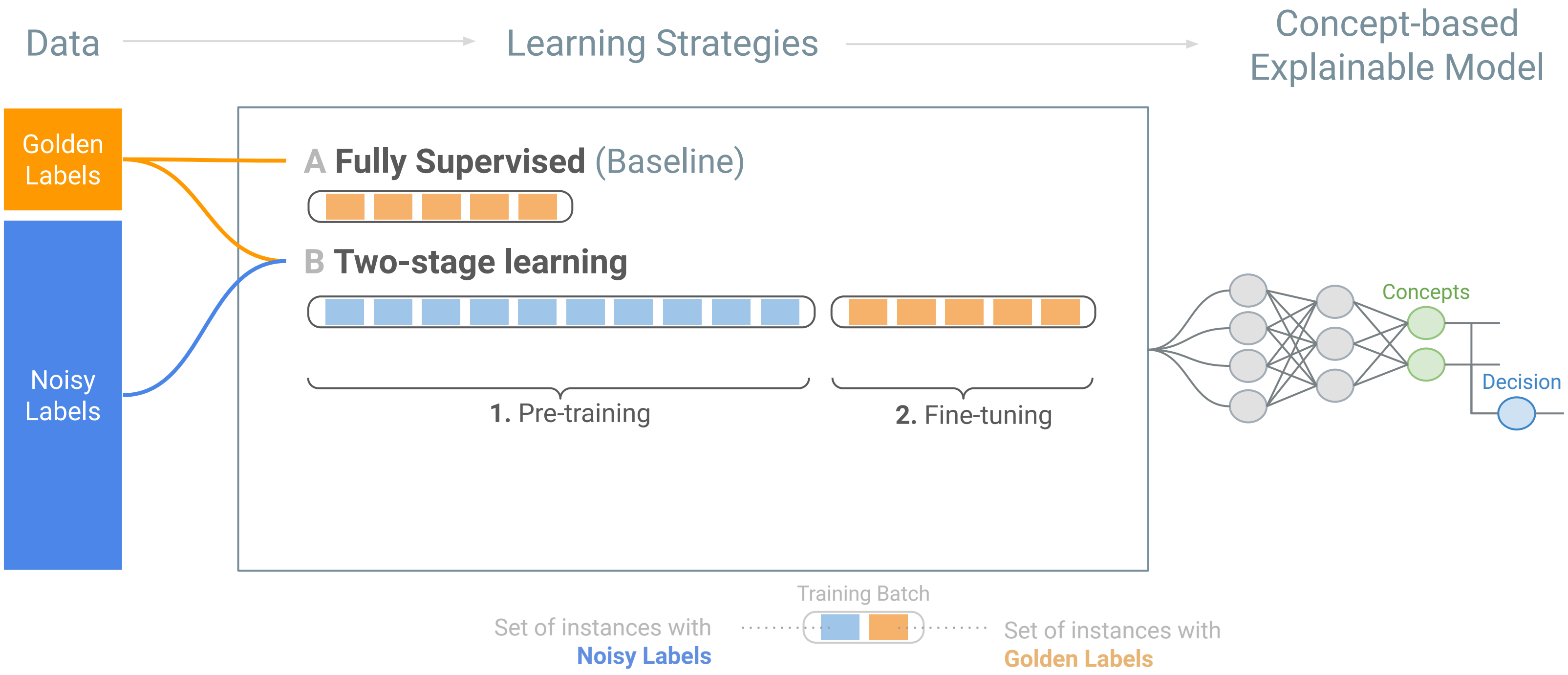


Training Batch

Set of instances with **Noisy Labels**

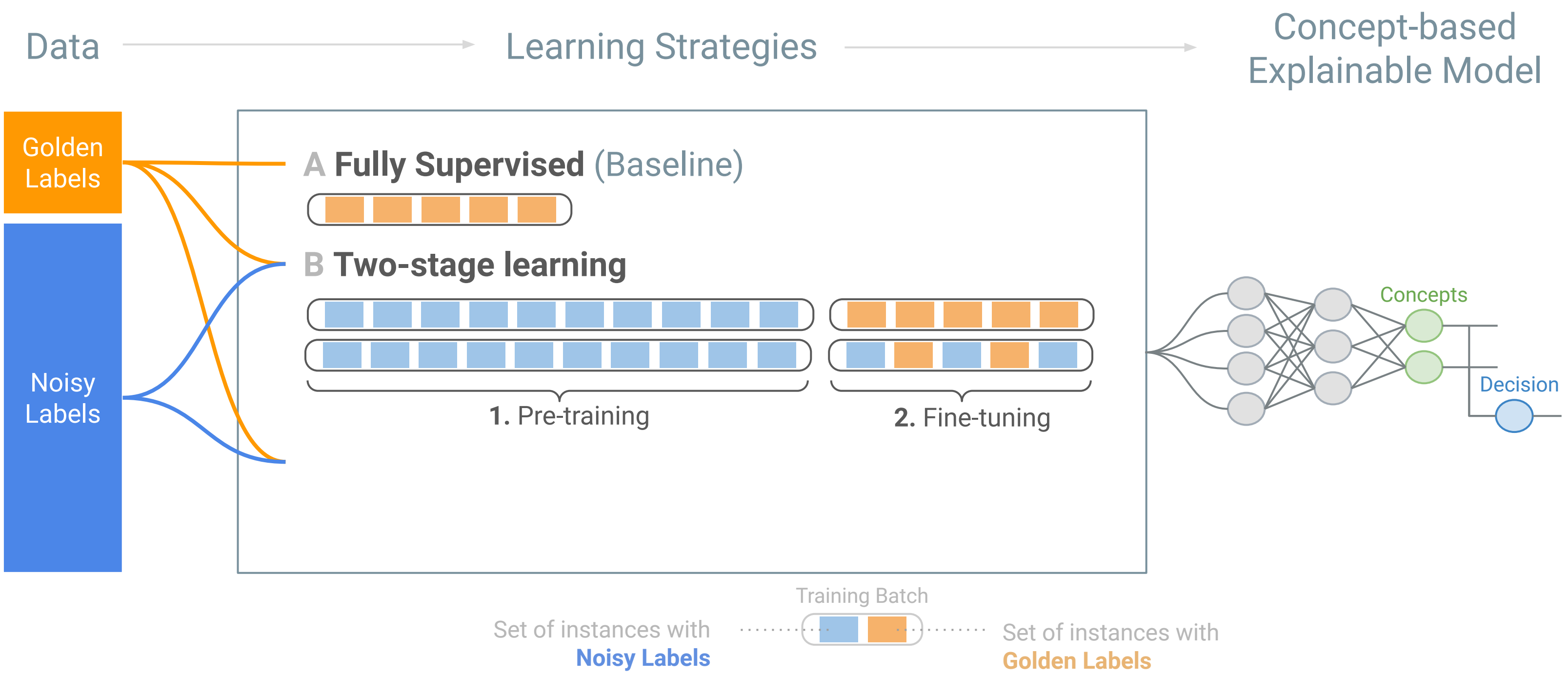
Set of instances with **Golden Labels**

Jointly learn to explain and decide: Learning Strategies



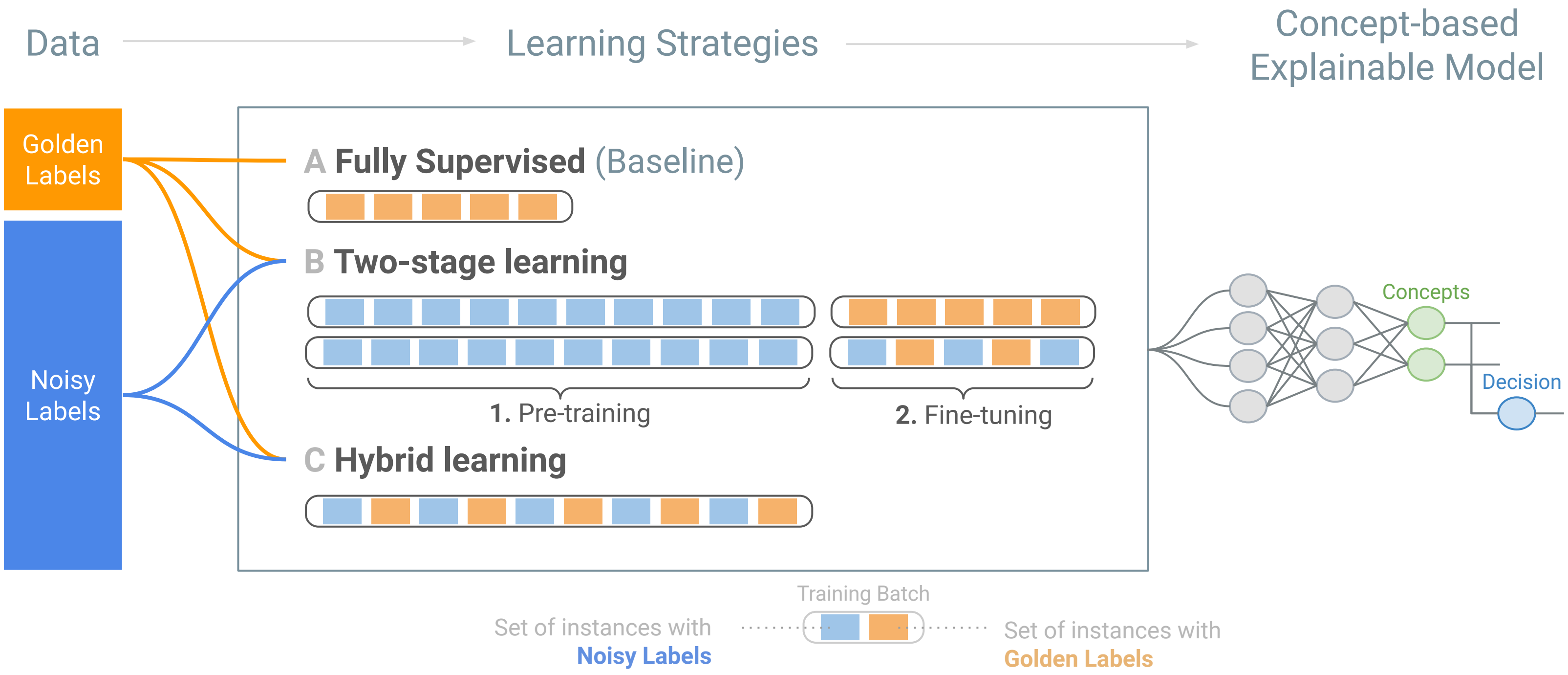
Jointly learn to explain and decide: Learning Strategies

feedzai

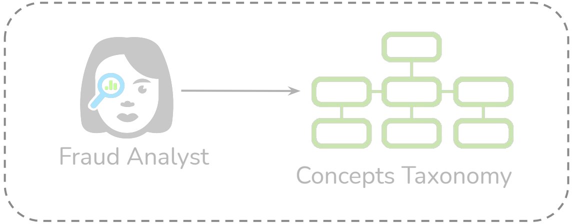


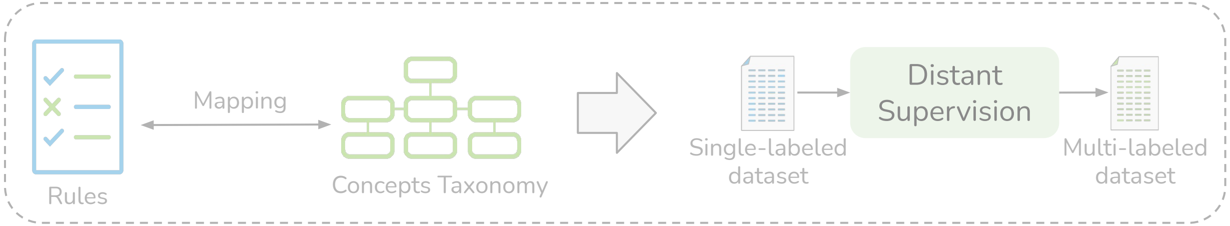
Jointly learn to explain and decide: Learning Strategies

feedzai



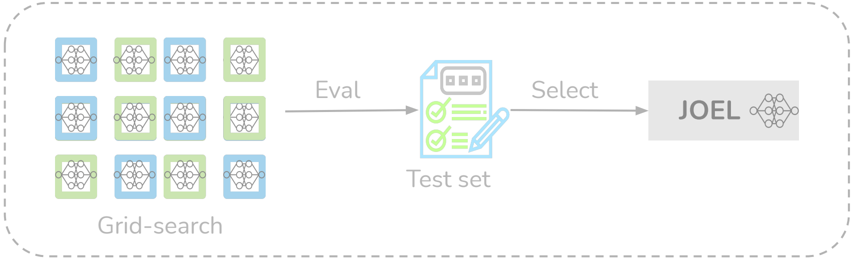
Implementation Workflow

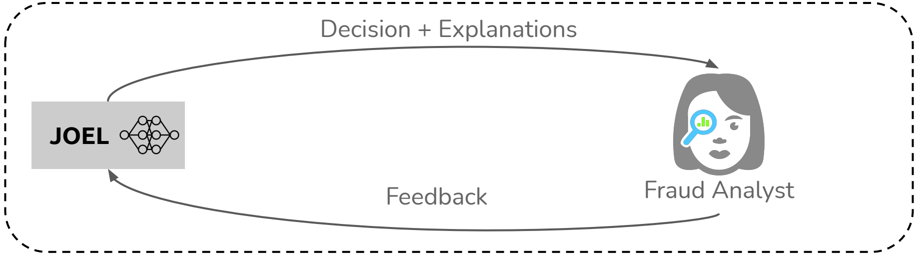




1 **Domain Expert defines concepts that will be used as explanations.**

2 **Apply Distant Supervision.**





3 **Run a grid search to find the best hyperparameters for Distant Supervision and Supervised Learning.**

4 **JOEL's online loop generates concept-based explanations and collects human feedback in a fraud detection setting.**

Human Feedback

- JOEL's explanations and prediction are shown to analyst through Web UI;
- A human-in-the-loop (*e.g.*, fraud analyst) makes a decision based on this information;
- When submitting its review, it also gives feedback about the concepts that led to his decision (and also about the decision task);
- This feedback can be used to continuously improve predictive accuracy, and also explainability of the model.

Experiments

Experimental Setup

feedzai

Payment retailer aims to deploy multi-task approach for concept-based explanations.

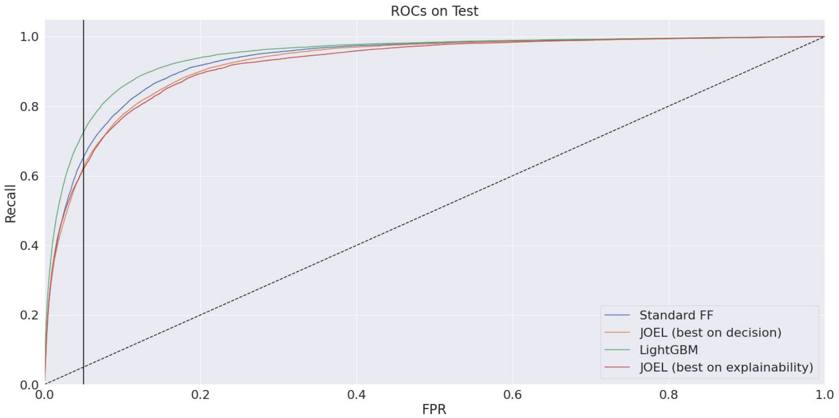
- **Binary decision task**: detect fraudulent transactions;
- **Multi-Label explainability task** (with 14 concepts): output high-level explanation about the model's prediction.

Labels	Availability
Golden decision	High (~6M, ~2% fraud rate)
Noisy Explainability	High (~6M, ~2% fraud rate)
Golden* explainability	Low (~1.3k, ~37% fraud rate)

Evaluation Metrics

Decision Task

Fraud recall @ 5% FPR



Explainability Task

Mean Average Precision (mAP)

$$AP = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{number\ of\ relevant\ items}$$

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q)$$

Hyperparameter & Model Selection

feedzai

*Can we **do better** than the **fully supervised** and **low-resources** baseline?*

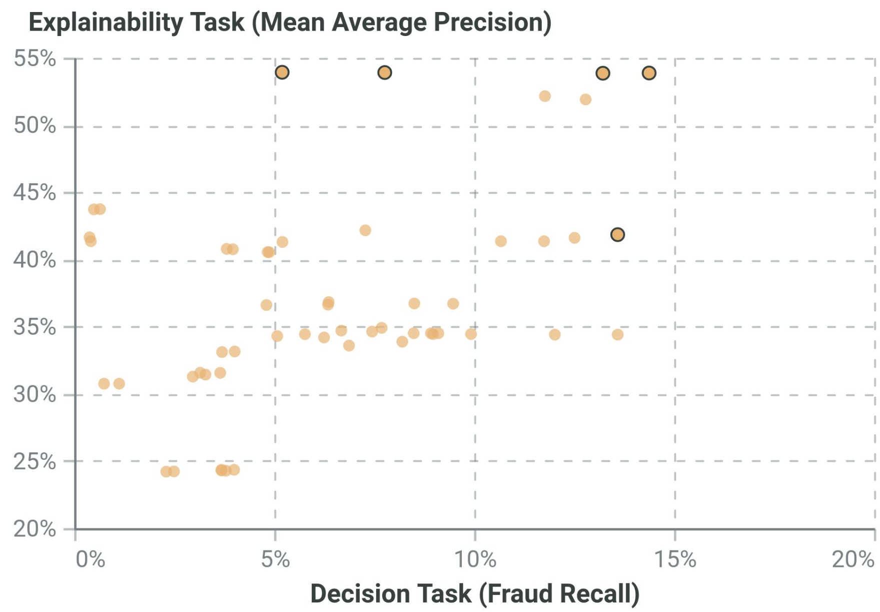
Model development:

- Run the same hyperparameter grid for each variant;
  (*i.e.*, number and dimension of hidden layers, learning rate, explainability task importance)

- Run 2 random seeds.

Fully Supervised (baseline) - Test set

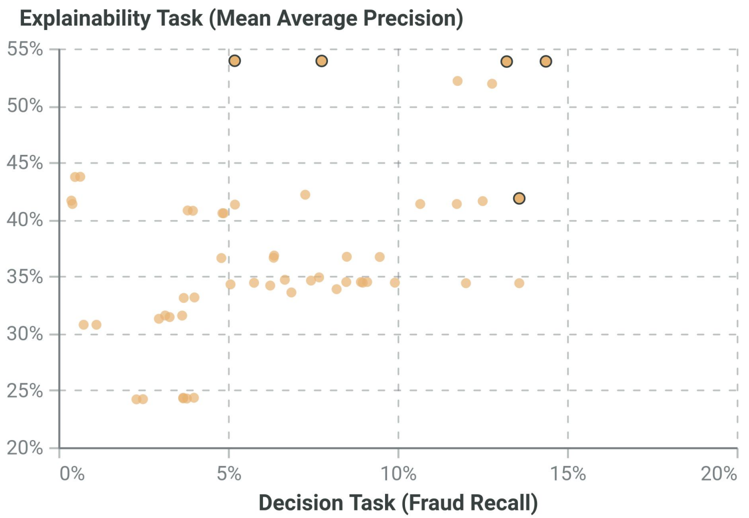
feedzai



Full supervised learning on a small dataset yields **poor results** in the decision task.

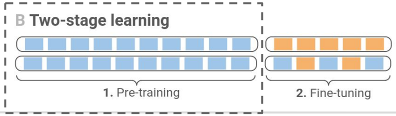
Fully Supervised (baseline) - Test set

A **Fully Supervised** (Baseline)



	Full Supervision
Recall	[0.04; 0.15]
mAP	[0.52; 0.54]

Two-stage learning - Test set



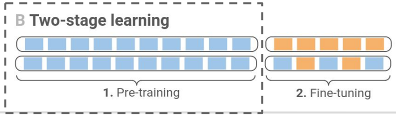
First stage:
Train base models using
Distant Supervision (Noisy Labels).

Second stage:
Fine-tune base models with
Golden Labels.

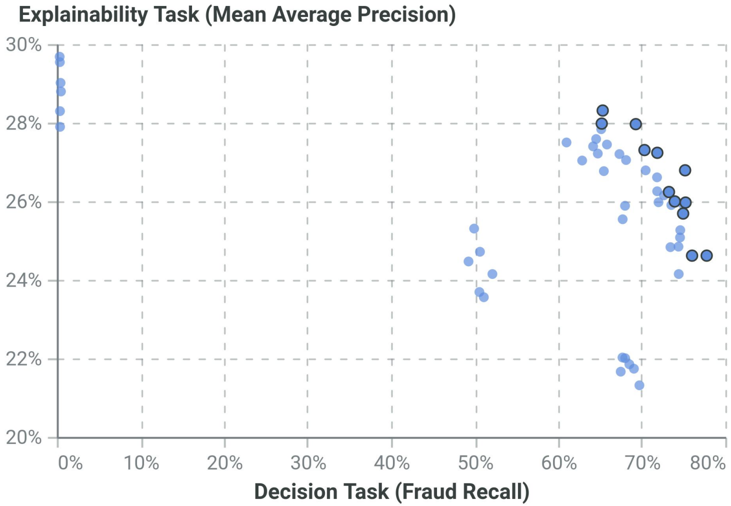
Main goal:
improve **explainability** task
without hurting the **decision** task.

	Full Supervision	Two-stage (base models)	Two-stage (w/ fine tuning)
Recall	[0.04; 0.15]		
mAP	[0.52; 0.54]		

Two-stage learning - Test set



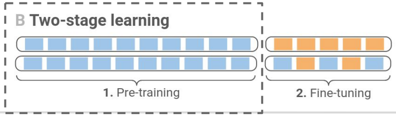
Base models (random seeds 10 and 42)



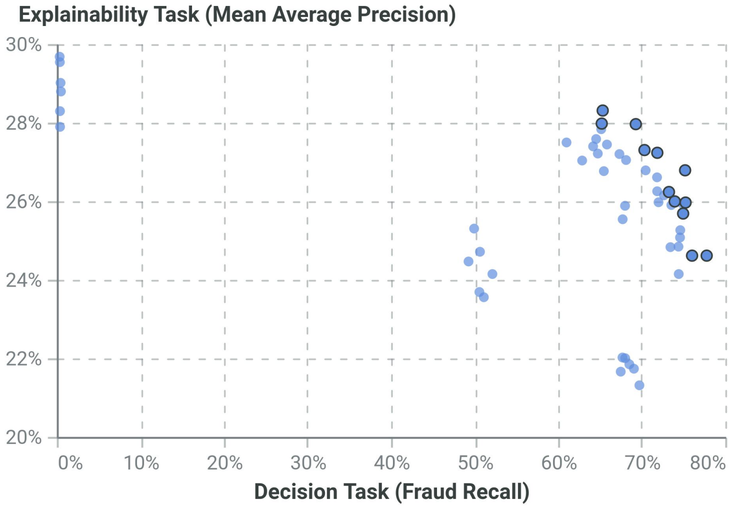
	Full Supervision
Recall	[0.04; 0.15]
mAP	[0.52; 0.54]

- Explainability performance degrades when using the larger but noisy explainability dataset.
- Training with larger dataset improves decision task;

Two-stage learning - Test set

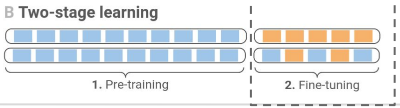


Base models (random seeds 10 and 42)



	Full Supervision	Two-stage (base models)
Recall	[0.04; 0.15]	[0.6; 0.78]
mAP	[0.52; 0.54]	[0.24; 0.29]

Two-stage learning - Test set

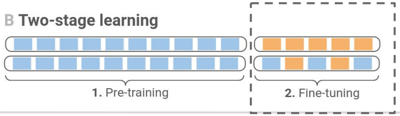


Second stage:
Fine-tune base models with Golden Labels.

Main goal:
improve **explainability** task without hurting the **decision** task.

	Full Supervision	Two-stage (base models)	Two-stage (w/ fine tuning)
Recall	[0.04; 0.15]	[0.6; 0.78]	
mAP	[0.52; 0.54]	[0.24; 0.29]	

Two-stage learning - Test set



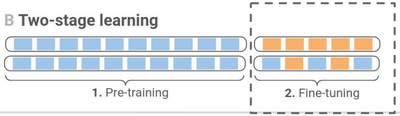
First stage:
Train base models using Distant Supervision (Noisy Labels).

SELECT PARETO MODELS

Second stage:
Fine-tune base models with Golden Labels.

Main goal:
improve **explainability** task without hurting the **decision** task.

Two-stage learning - Test set



Pick **Pareto Optimal** models from Stage 1.

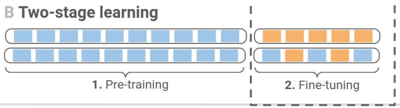
Fine-tune them with:

- Hyperparameters;
  (*e.g.*, learning rate, epochs, batch size)
- Loss scalers;
  (*e.g.*, 0.75, 0.5, 0.25)
- Freeze and unfreeze of layers;
- Different batch techniques.
  (*e.g.*, hybrid batching - add % of noisy labels in each batch)

Second stage:
Fine-tune base models with Golden Labels.

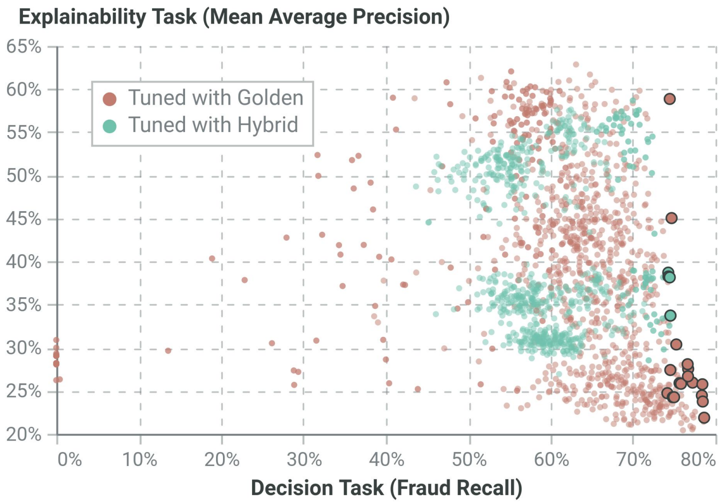
Main goal:
improve **explainability** task without hurting the **decision** task.

Two-stage learning - Test set



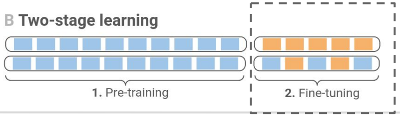
Base models (random seeds 10 and 42)

Fine-tuned models



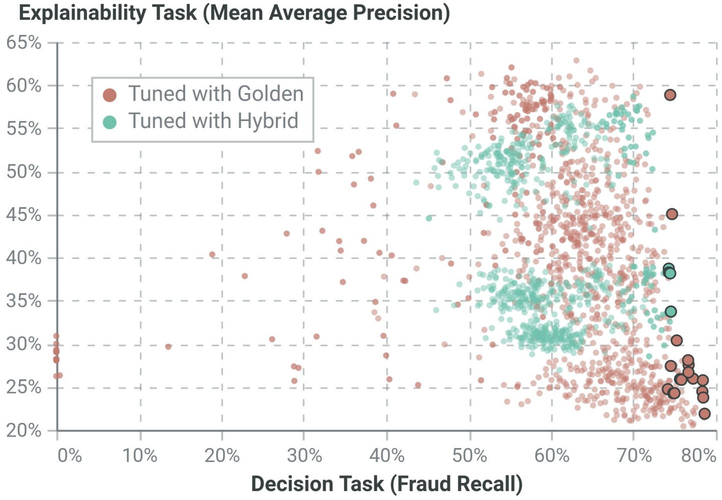
For same decision task performance, we can improve explainability task.

Two-stage learning - Test set



Base models (random seeds 10 and 42)

Fine-tuned models



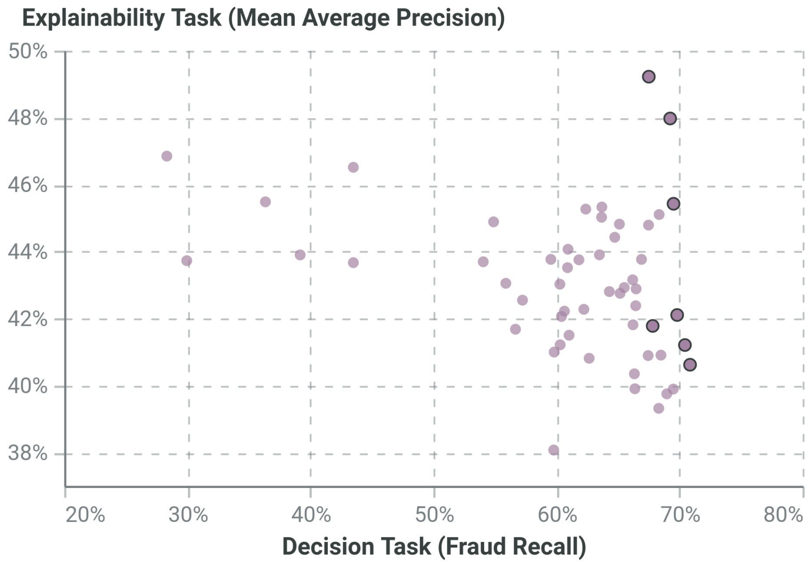
	Full Supervision	Two-stage (base models)	Two-stage (w/ fine tuning)
Recall	[0.04; 0.15]	[0.6; 0.78]	[0.74; 0.78]
mAP	[0.52; 0.54]	[0.24; 0.29]	[0.24; 0.63]

Hybrid learning - Test set

Each training batch contains a fraction of Explainability Golden Labels (Our experiments use 10%).

Hybrid learning - Test set

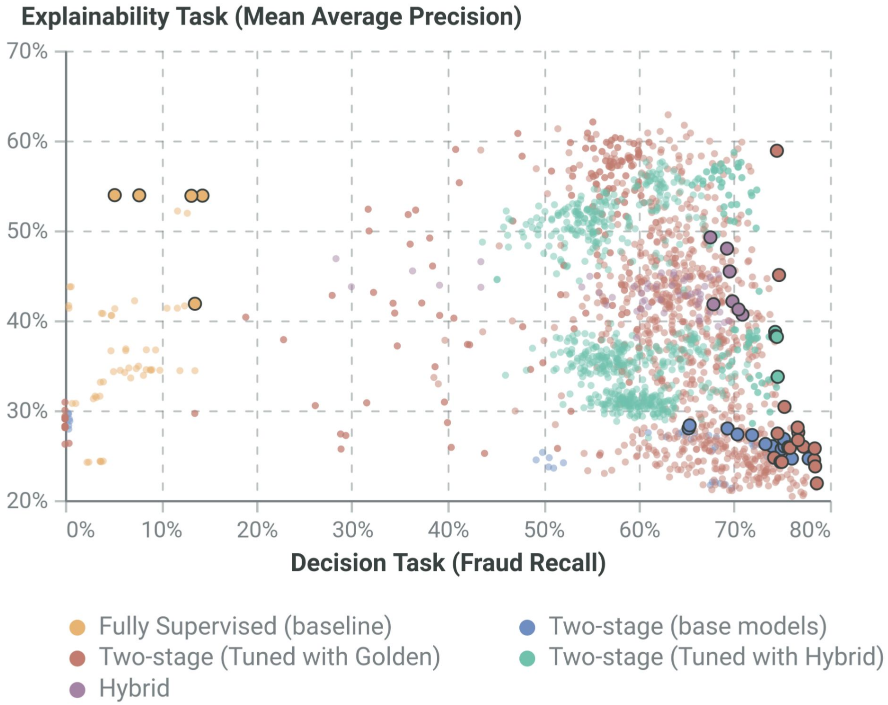
feedzai



	Full Supervision	Two-stage (base models)	Two-stage (w/ fine tuning)	Hybrid
Recall	[0.04; 0.15]	[0.6; 0.78]	[0.74; 0.78]	**[0.65; 0.71]**
mAP	[0.52; 0.54]	[0.24; 0.29]	[0.24; 0.63]	**[0.4; 0.49]**

Better than **fully supervised** in decision and **two-stage base** models in explainability, but seems to be worse than **two-stage fine-tuned**.

All Pareto from learning strategies - test set



Preliminary results seem to be **promising** but further experiments (more seeds and more runs) to gain statistical confidence.

Conclusions

Recap

- Concept-based explainability through multi-task learning poses challenges:
    - Label scarcity;
    - Joint learning of decision and associated explanations.

- This work proposes to:
    - Use Distant Supervision and exploit the available off-the-shelf domain knowledge;
    - Use different Learning Strategies and combine label qualities to improve performance at both tasks.

Conclusions

feedzai

- The explanations should be tailored to the persona's knowledge and task performed;
  - Concept-based explanations are suitable to domain experts that make ML-informed decisions but lack ML knowledge!

- Experiment in a real-world e-commerce fraud detection dataset show:
  1. JOEL is able to learn both domain concept explanations and fraud decisions;
  2. Distant supervision allows us to overcome the label scarcity problem;
  3. There is no clear winner learning strategy.
     (it might depend on business requirements)

Thank you!

feedzai

Questions?

vladimir.balayan@feedzai.com
catarina.belem@feedzai.com
a4338@fct.unl.pt
pedro.saleiro@feedzai.com
pedro.bizarro@feedzai.com