Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Occupational Fraud Detection through Agent-based Data Generation

Julian Tritscher, Alexander Roos, Daniel Schlör,
Andreas Hotho, and Anna Krause

CAIDAS Center of Artificial Intelligence and Data Science
University of Würzburg
Germany

data science

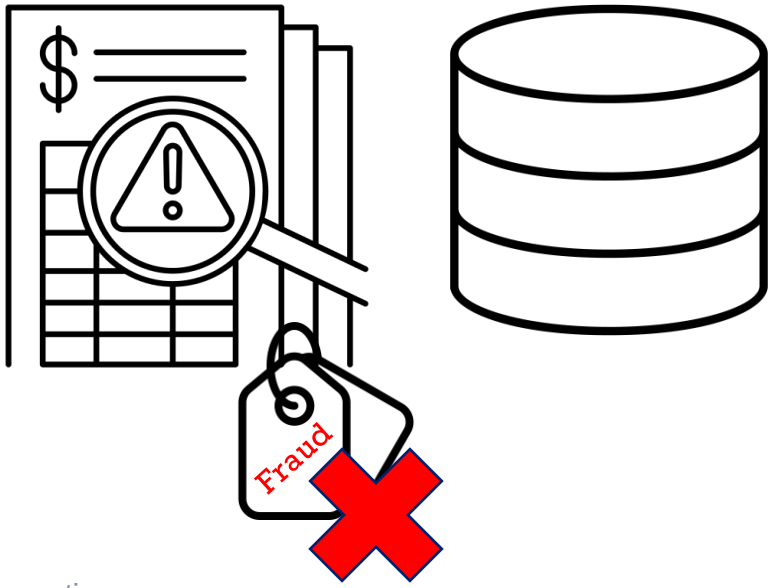
CAIDAS

Motivation

- Fraud detection: detecting anomalies in a large, unlabeled dataset
- Hyperparameters need to be optimized
- Solution to mitigate lack of labels for validation and HP optimization:

  Multi-Agent-based data generation including labels

https://legacy.acfe.com/report-to-the-nations/2022/



Agenda

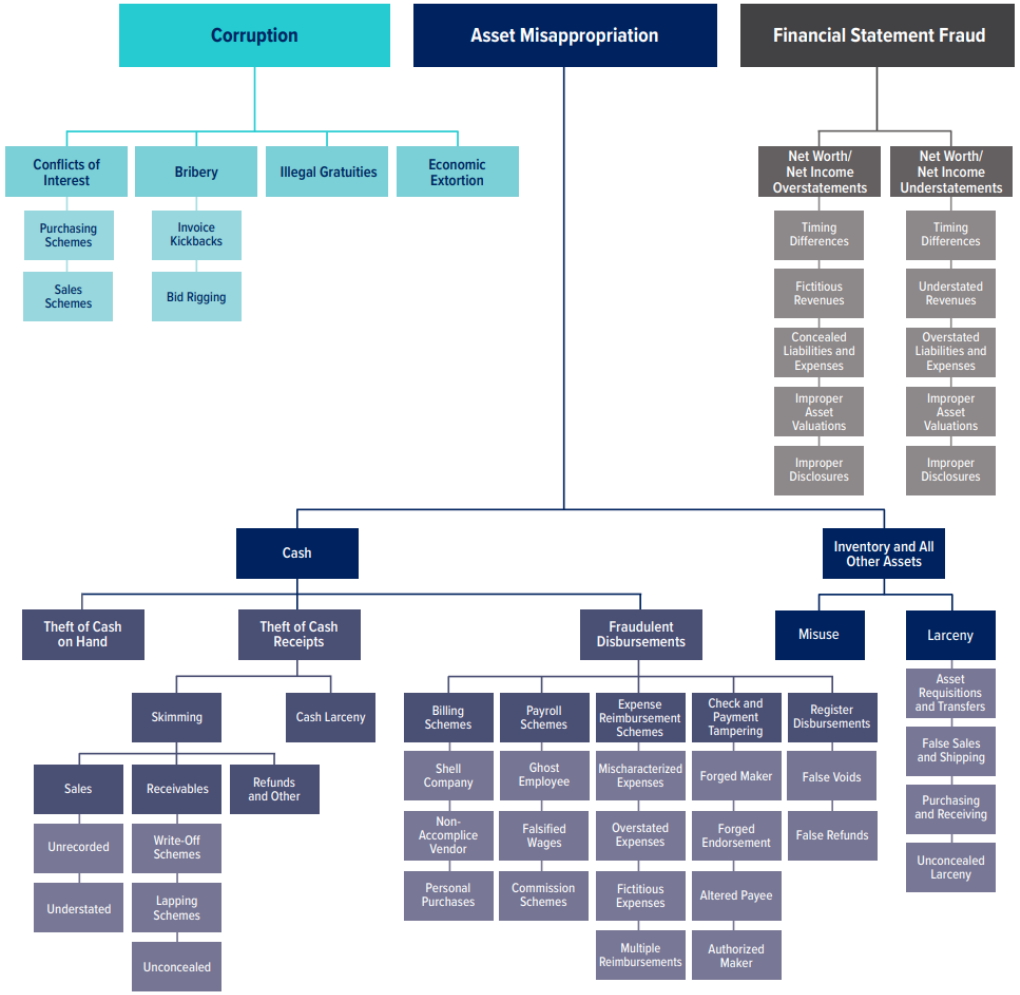
- Occupational Fraud Detection
- ERP Data simulation through Multi-Agent Systems (MAS)
- Fraud Detection with MAS Data
  - Data Generation
  - Experiments

Occupational Fraud Detection

Occupational fraud is formally defined as the use of one's occupation for personal enrichment through the **deliberate misuse or misapplication of the employing organization's resources** or assets.
– ACFE 2022

Occupational Fraud Detection

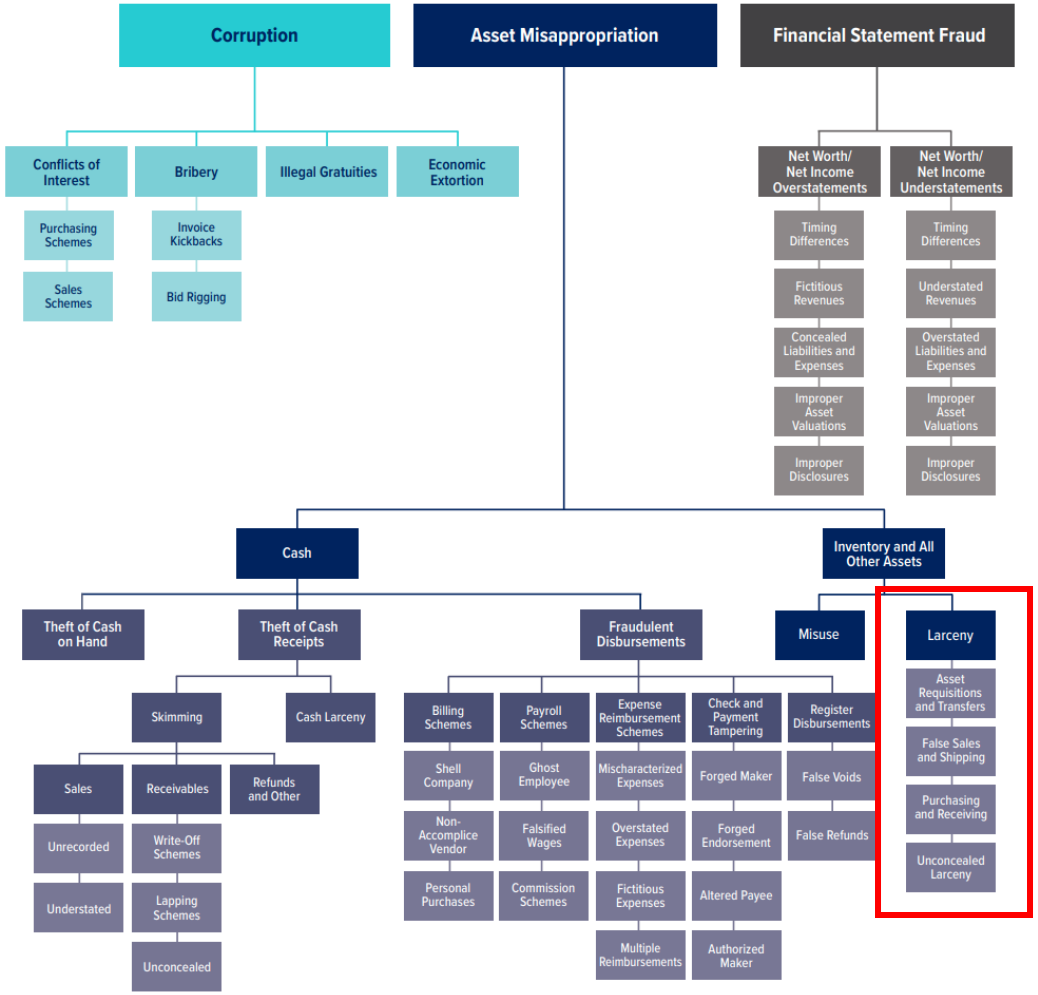
Occupational fraud is formally defined as the use of one's occupation for personal enrichment through the **deliberate misuse or misapplication of the employing organization's resources** or assets.
– ACFE 2022



https://legacy.acfe.com/report-to-the-nations/2022/

Occupational Fraud Detection

Occupational fraud is formally defined as the use of one's occupation for personal enrichment through the **deliberate misuse or misapplication of the employing organization's resources** or assets.
– ACFE 2022



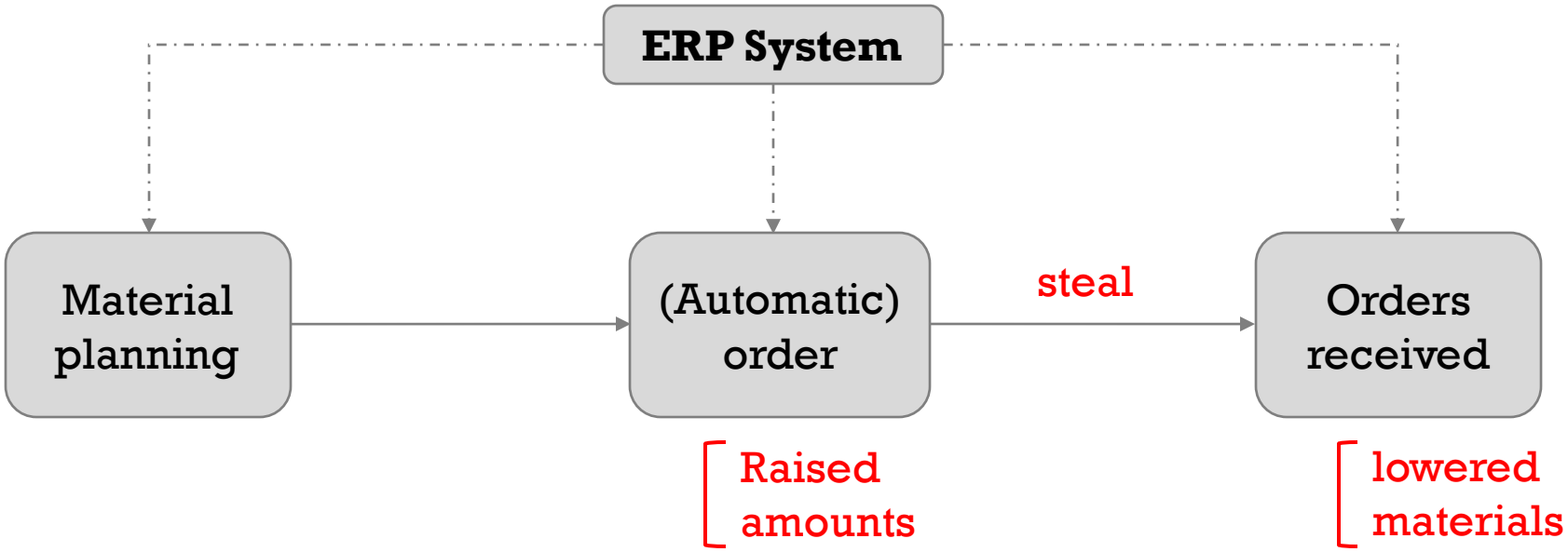
https://legacy.acfe.com/report-to-the-nations/2022/

Example case

- Asset missappropriation: Larceny 1

Example case

- Asset missappropriation: Larceny 1



Current Limitations for Fraud Detection

Data availability

- No real company datasets public
- Especially no real fraud cases

Expensive labeling

- No labeled real data available
- Requires auditing experts
- Fraud cases are rare
- No guarantees for representative datasets

Current Limitations for Fraud Detection

Data availability

- No real company datasets public
- Especially no real fraud cases

- Public simulated datasets [Tritscher2022]
  - Training simulation in a real ERP system
  - Simulated cereal production company run by research participants
  - Fraud cases committed directly in the ERP system
  - Machine-learning-ready datasets
    - Using financial accounting tables
    - Multiple datasets with 1 fiscal year each

Open ERP System Data For Occupational Fraud Detection

Julian Tritscher[1], Fabian Gwinner[2], Daniel Schlör[1], Anna Krause[1], and Andreas Hotho[1]

[1] University of Würzburg, Am Hubland, 97074 Würzburg, Germany
{tritscher, schloer, anna.krause, hotho}@informatik.uni-wuerzburg.de
[2] fabian.gwinner@uni-wuerzburg.de

Abstract. Recent estimates report that companies lose 5% of their revenue to occupational fraud. Since most medium-sized and large companies employ Enterprise Resource Planning (ERP) systems to track vast amounts of information regarding their business process, researchers have in the past shown interest in automatically detecting fraud through ERP system data. Current research in this area, however, is hindered by the fact that ERP system data is not publicly available for the development and comparison of fraud detection methods. We therefore endeavour to generate public ERP system data that includes both normal business operation and fraud. We propose a strategy for generating ERP system data through a serious game, model a variety of fraud scenarios in cooperation with auditing experts, and generate multiple years of data from a simulated make-to-stock production company. We aggregate the generated data into ready to used datasets for fraud detection in ERP systems, and supply both the raw and aggregated data to the general public to allow for open development and comparison of fraud detection approaches on ERP system data.

Keywords: Data generation · Fraud detection · SAP.

1 Introduction

Tritscher, Julian, et al. "Open ERP system data for occupational fraud detection." *arXiv preprint arXiv:2206.04460* (2022).

Synthetic Datasets

- Modeling business processes is complex
- User interaction is expensive
- Extension (e.g., by new fraud cases) requires reiteration of the whole process

Possible solution: Replace users in this "simulation" with agents

=> Simulation with Multi Agent Systems (MAS)

ERP & MAS

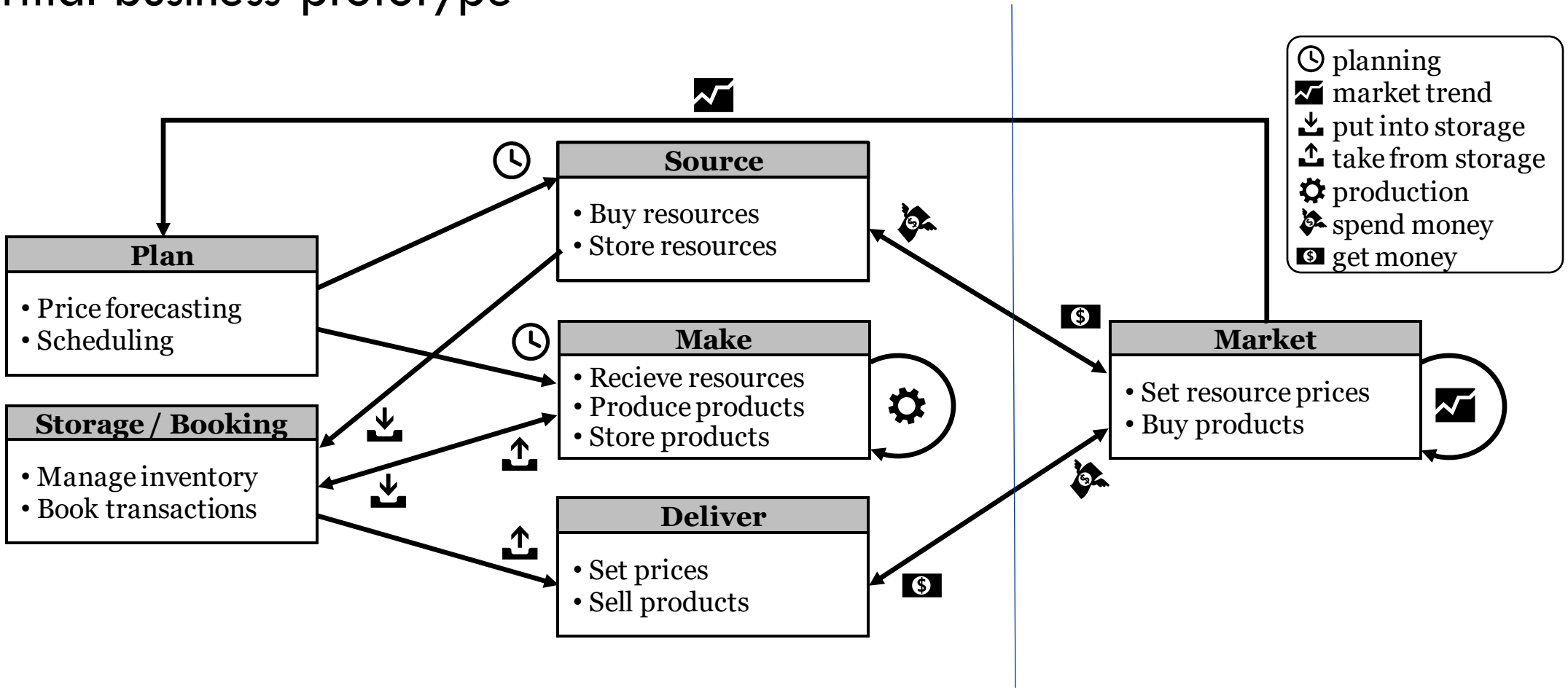
- Existing research that models companies in MAS
  - Can model business processes
  - Can be extended to include fraud
- This paper
  - Model fraud behavior
  - MAS for make-to-stock production company

Category	SCOR model	Planning Matrix (Stadtler, 2005)	Agents	Main Tasks
Planning Activities	Plan	Demand Fulfilment & ATP	*Demand Fulfilment Agent*	Demand management Communication with customers
		Purchasing & Material Requirements Planning	*MRP Agent*	Purchase management Communication with providers
		Demand Planning	*Demand Forecast Agent*	Demand forecast
		Master Planning	*Master Planning Agent*	Aggregate production planning
		Production Planning	*Production Planning Agent*	Disaggregate production planning
		Scheduling	*Scheduling Agent*	Jobs sequence
Physical Activities	Source	-	*Source Agent*	Reception and storage of raw materials
	Make	-	*Make Agent*	Manufacturing process (machines)
	Deliver	-	*Deliver Agent*	Storage of finished products and delivery to customers
	Return	-		

Domínguez, R., Cannella, S., Framinan, J.M.: SCOPE: A Multi-Agent system tool for supply chain network analysis. In: IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON). pp. 1–5 (Sep 2015)

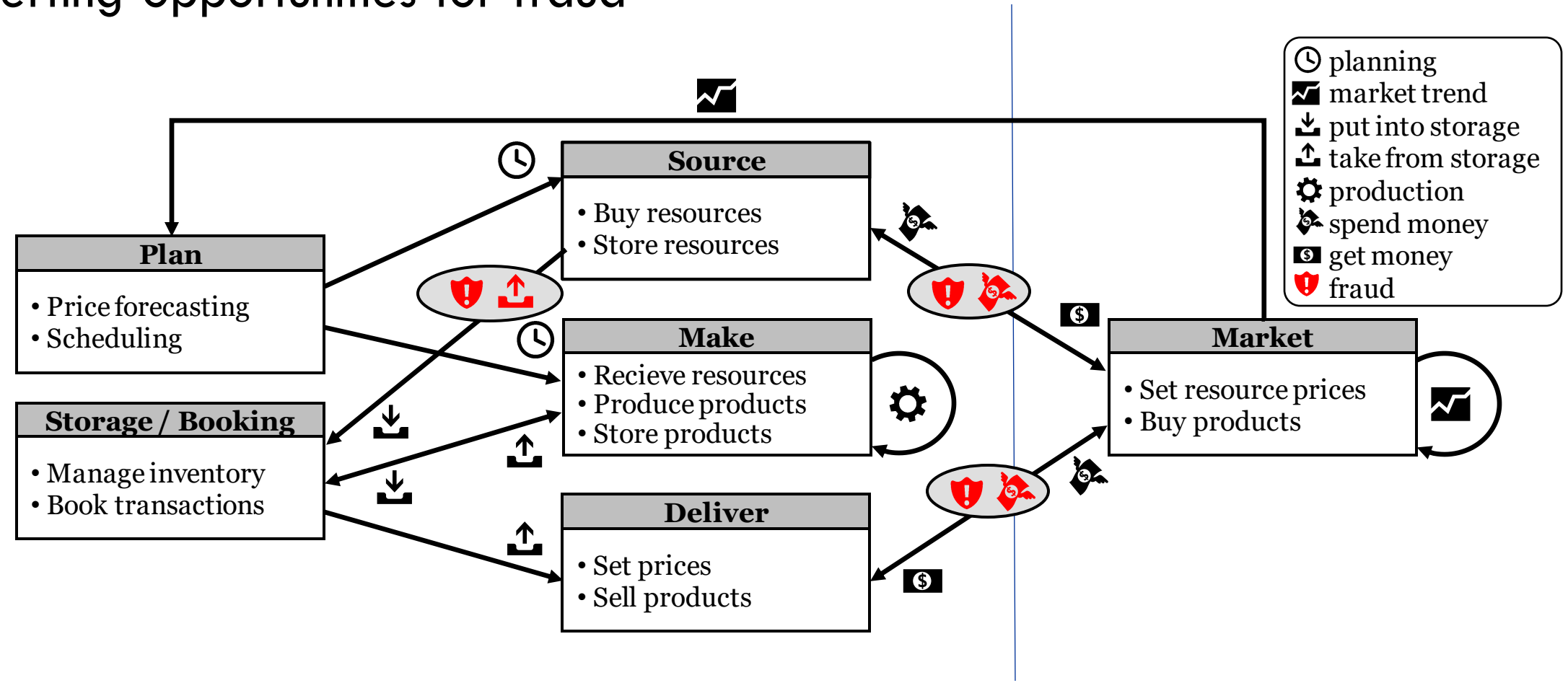
ERP & MAS

- Normal business prototype

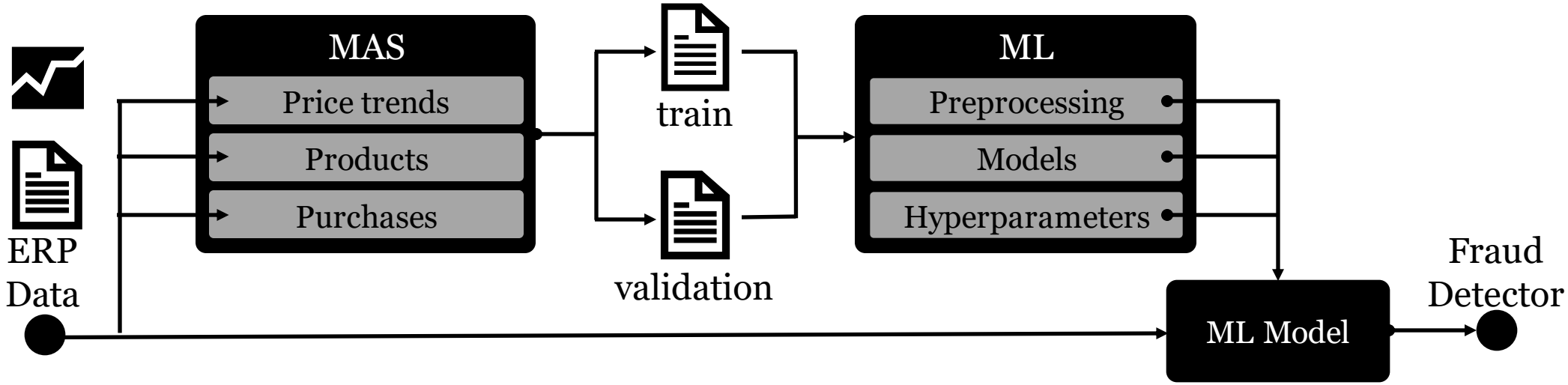


ERP & MAS

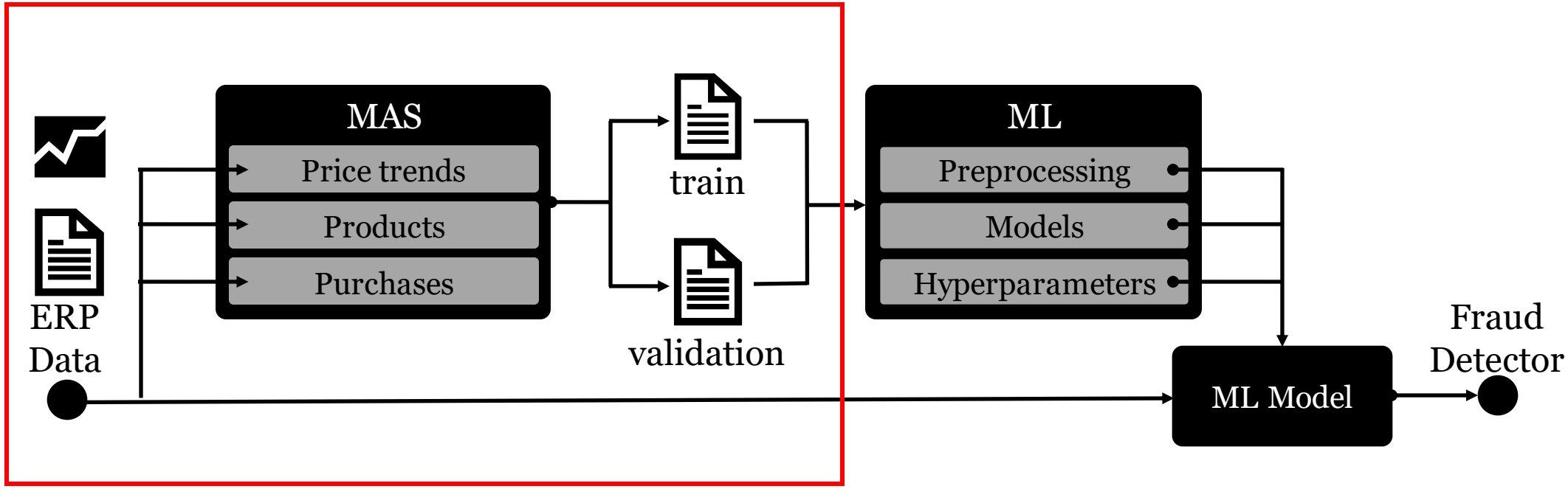
- Inserting opportunities for fraud



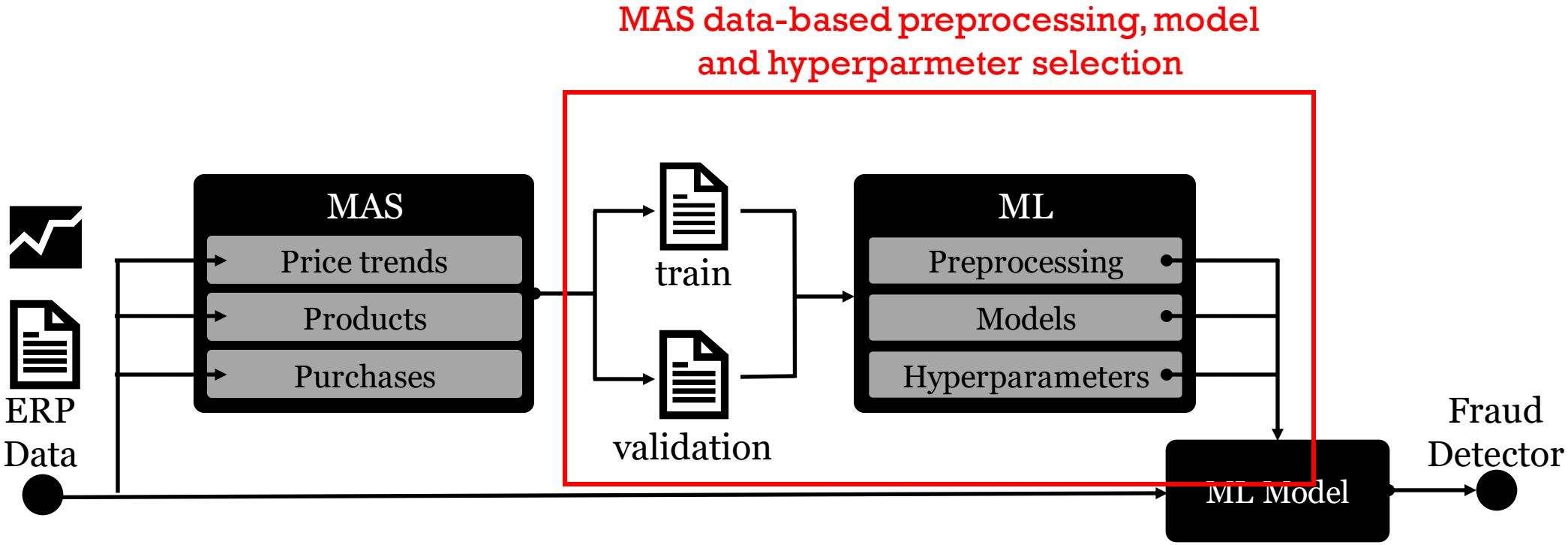
MAS Data for Fraud Detection



MAS Data for Fraud Detection

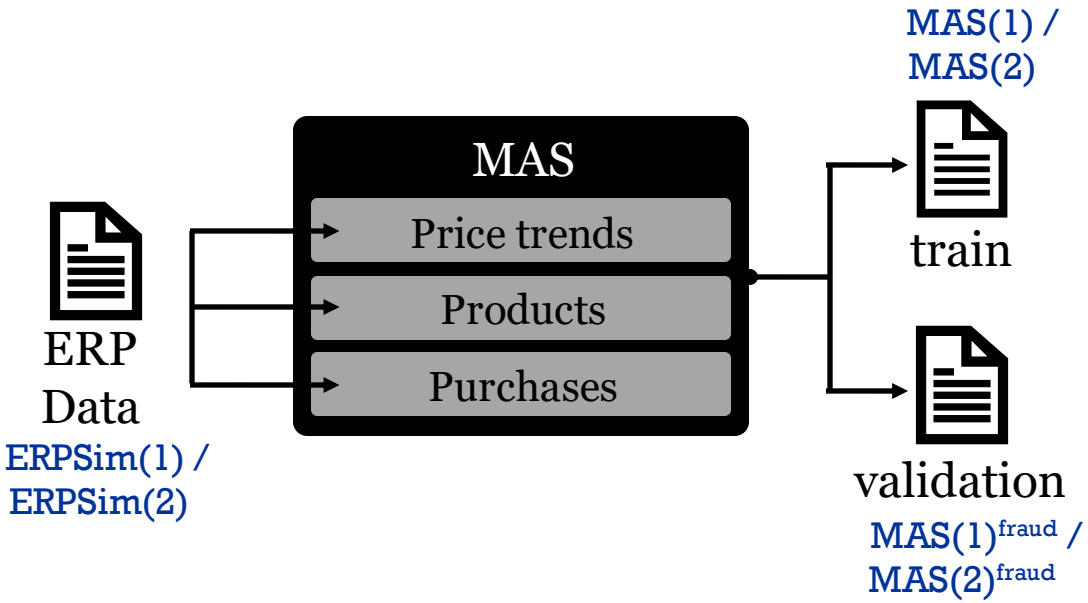


MAS Data for Fraud Detection



Data Generation

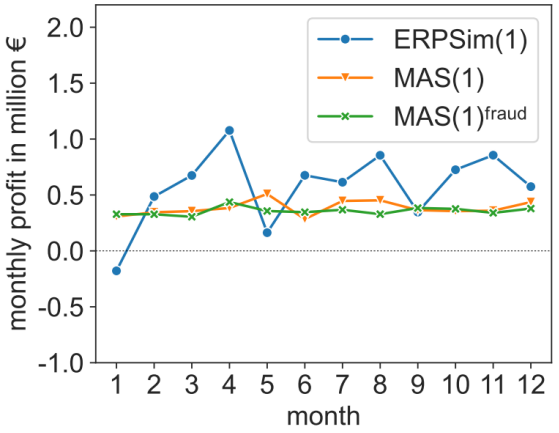
- Inserting known parameters
  - Price trends
    (public databases or interpolated from data)
  - Products
  - Number of purchases (upper bound)
- Market and business strategy assumptions
  - Trend forecasts through Holt-Winters

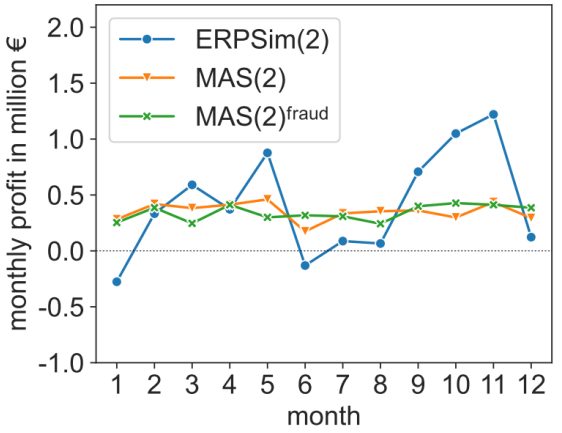


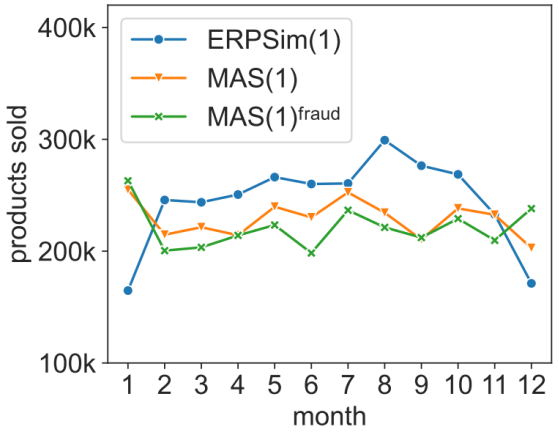
Dataset	Transactions	Frauds	Invoice Kickback	Selling Kickback	Larceny	Corporate Injury
ERPSim(1)	36778	50	24	0	22	4
MAS(1)	92985	0	0	0	0	0
MAS(1)^{fraud}	93356	223	51	104	66	2
ERPSim(2)	37407	86	30	0	48	8
MAS(2)	59378	0	0	0	0	0
MAS(2)^{fraud}	64858	187	51	102	34	0

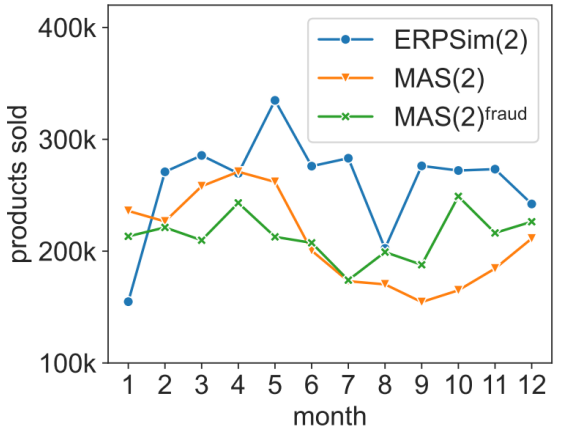
Data Generation

Dataset	Transactions	Frauds	Invoice Kickback	Selling Kickback	Larceny	Corporate Injury
ERPSim(1)	36778	50	24	0	22	4
MAS(1)	92985	0	0	0	0	0
MAS(1)^{fraud}	93356	223	51	104	66	2
ERPSim(2)	37407	86	30	0	48	8
MAS(2)	59378	0	0	0	0	0
MAS(2)^{fraud}	64858	187	51	102	34	0



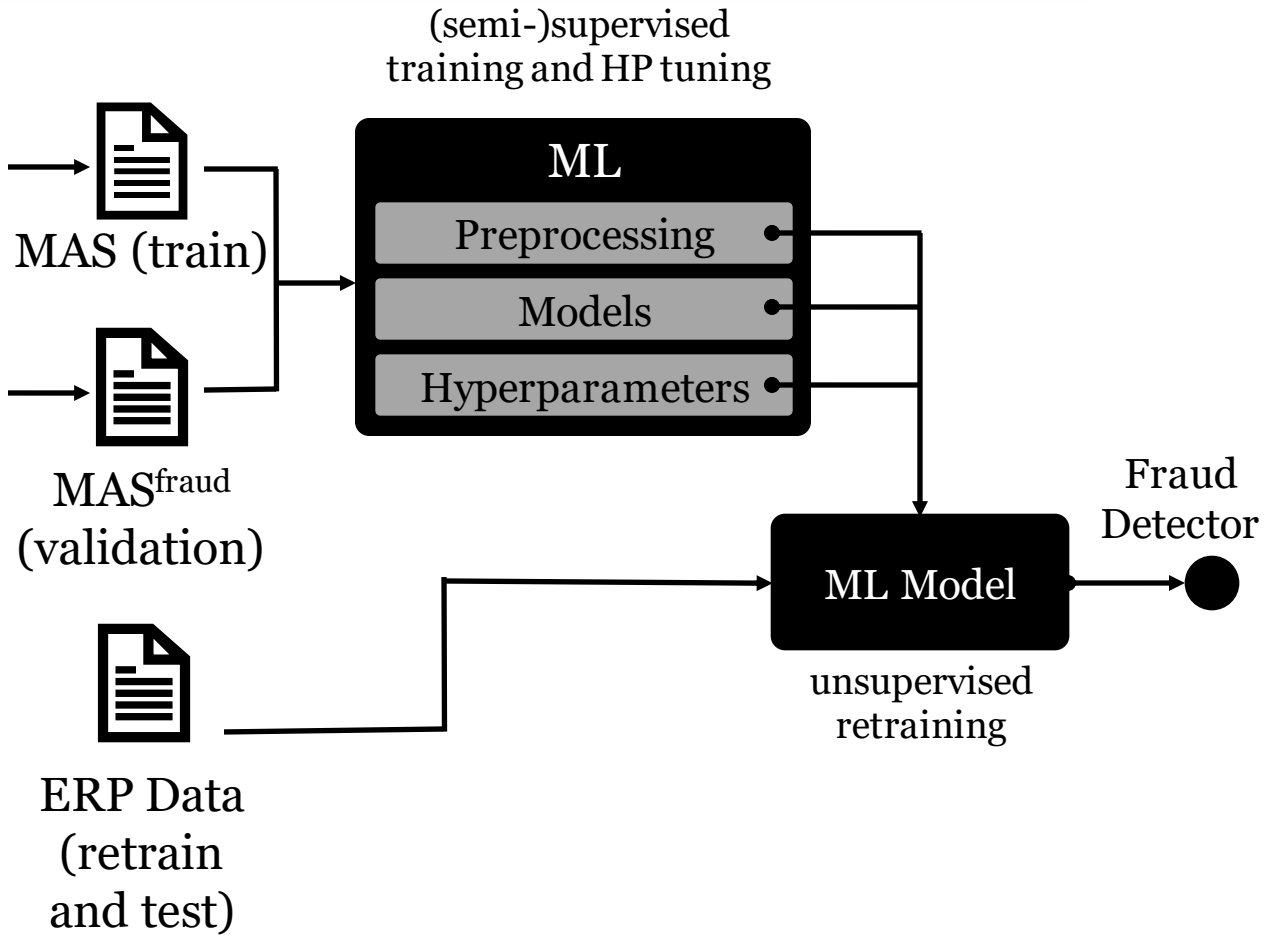






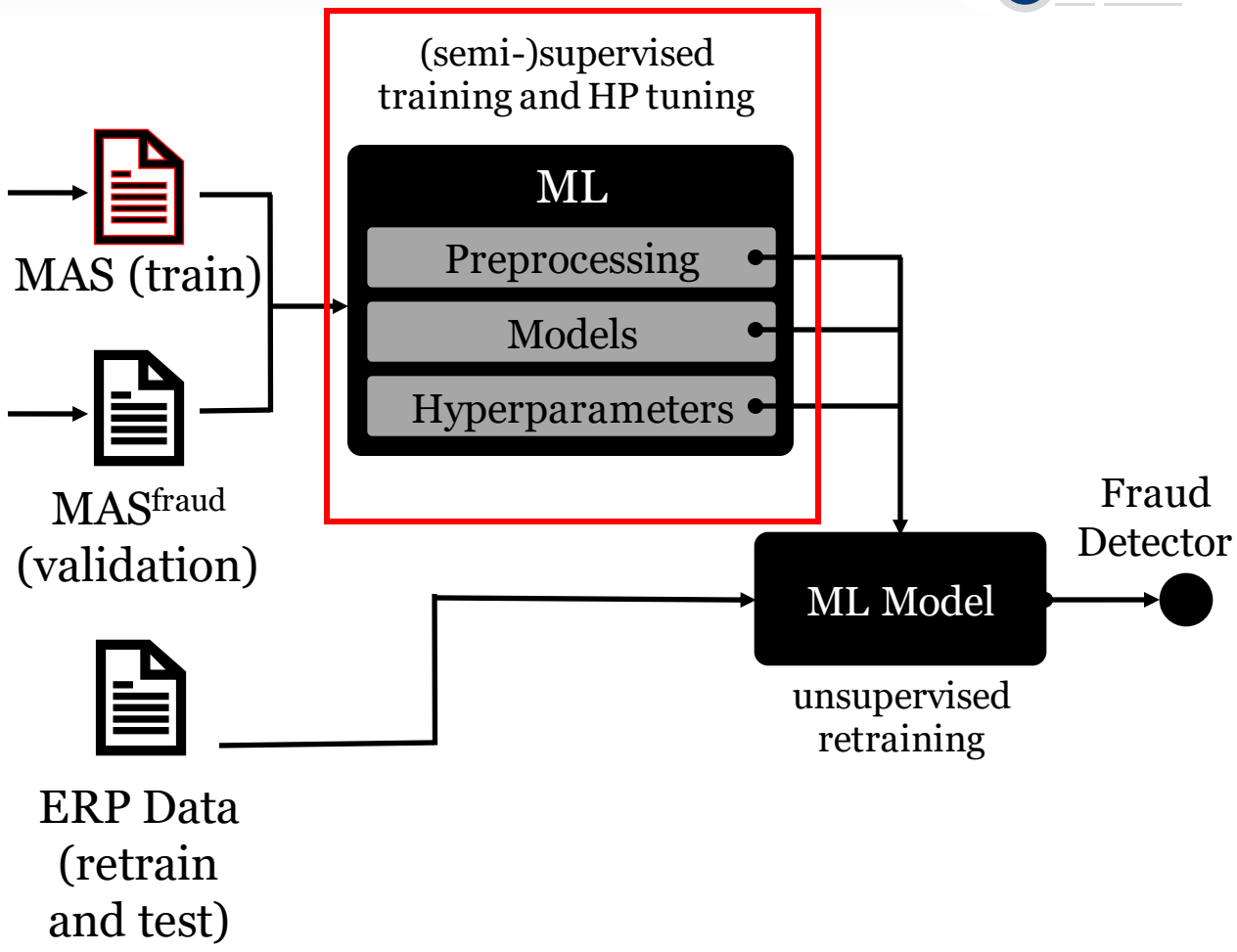
Fraud Detection

- Train-eval tuning loop on synthetic MAS data



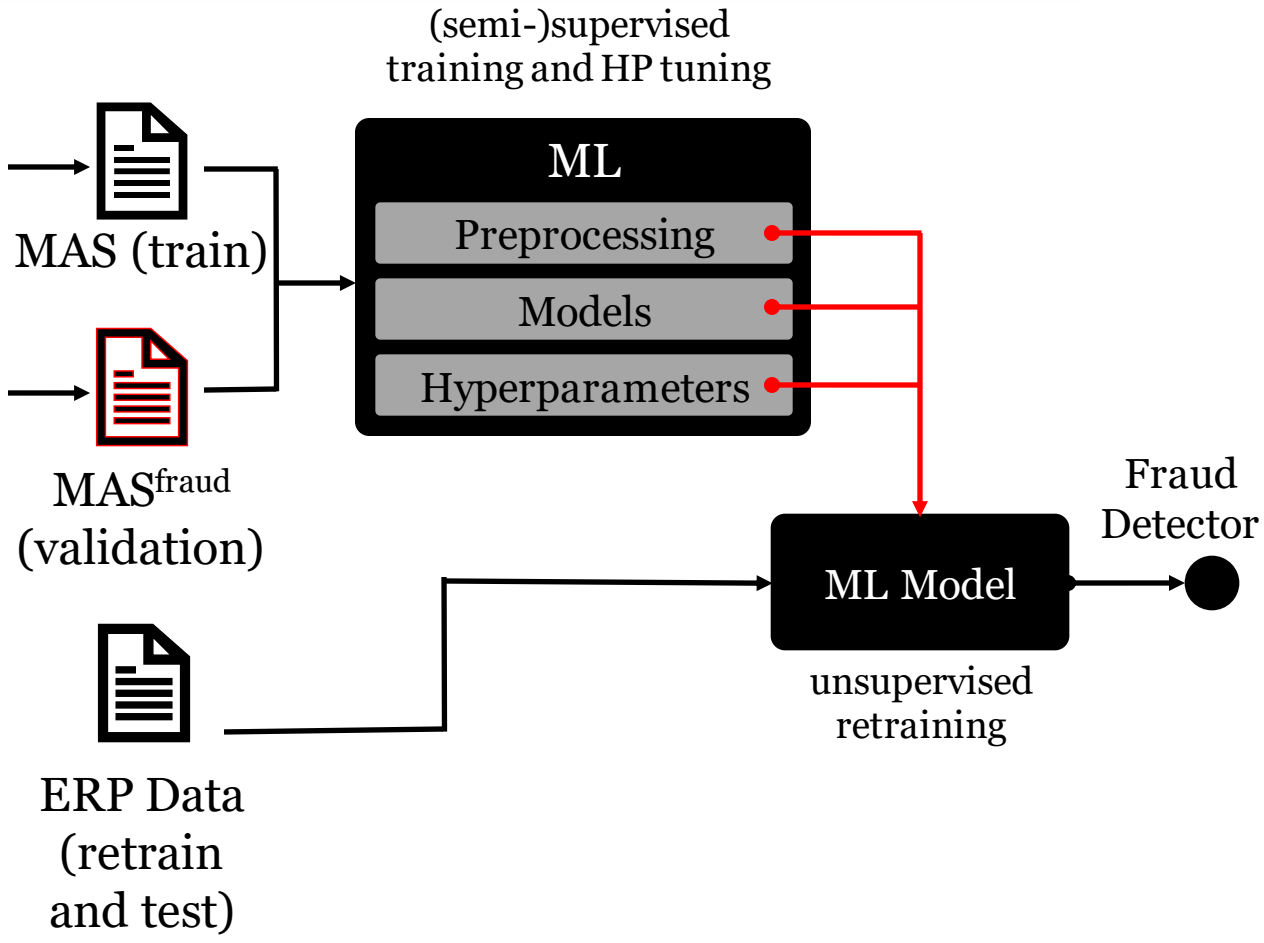
Fraud Detection

- Train-eval tuning loop on synthetic MAS data
- Train on clean MAS data
  - Different preprocessing choices
  - Different models
  - Different hyperparameters



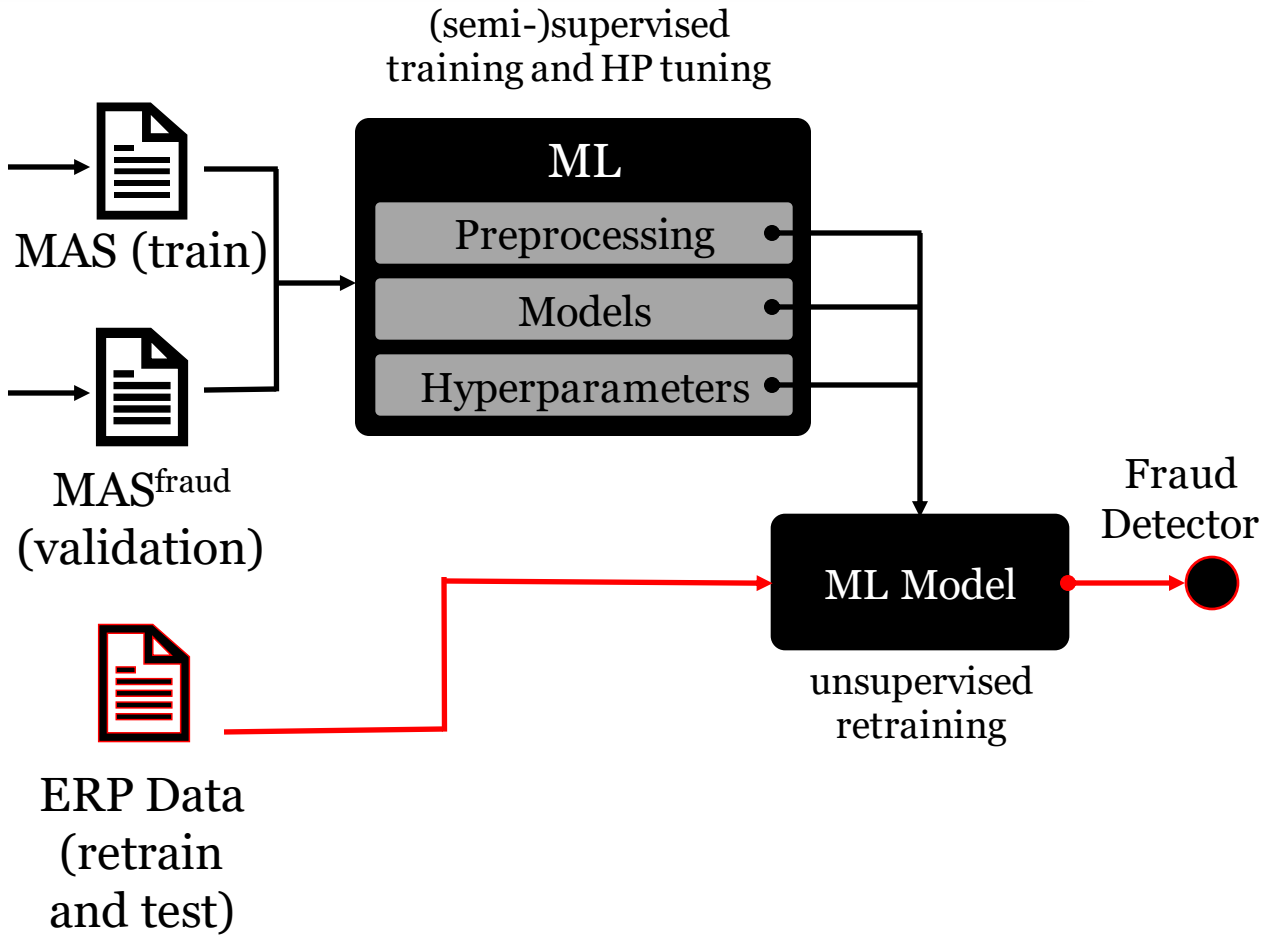
Fraud Detection

- Train-eval tuning loop on synthetic MAS data
- Train on clean MAS data
  - Different preprocessing choices
  - Different models
  - Different hyperparameters
- Select the best options according to MAS test set including simulated fraud

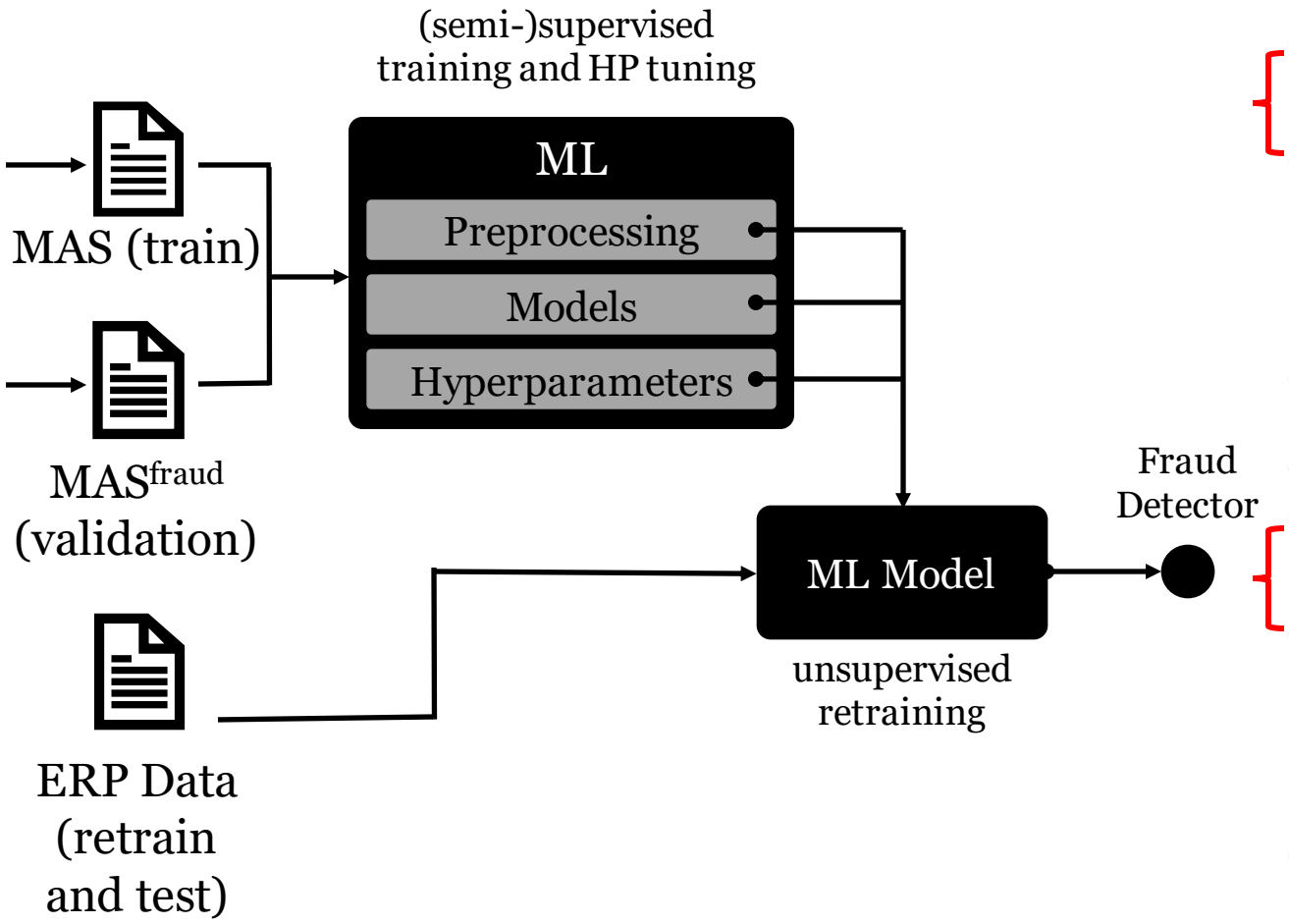


Fraud Detection

- Train-eval tuning loop on synthetic MAS data
- Train on clean MAS data
  - Different preprocessing choices
  - Different models
  - Different hyperparameters
- Select the best options according to MAS test set including simulated fraud
- Retrain best combinations on real ERP data in unsupervised regime
  - i.e., using contaminated data including fraud
- Evaluate on labels of ERP data



Fraud Detection Results



HP choice	model	preprocessing	PR_{synth}	$PR_{ERPSim(1)}$	$ROC_{ERPSim(1)}$
synthetic	AE	synthetic	$\underline{17.5 \pm 2.6}$	$\underline{26.5 \pm 4.3}$	$\mathbf{99.1 \pm 0.2}$
	IF	synthetic	3.8 ± 0.1	12.3 ± 0.8	97.5 ± 0.3
	OCSVM	synthetic	$\mathbf{21.5 \pm 0.0}$	$\mathbf{28.1 \pm 0.0}$	96.0 ± 0.0
default	AE	quantized	N/A	1.7 ± 0.3	73.0 ± 4.2
	AE	zscore	N/A	1.4 ± 0.2	72.6 ± 7.8
	IF	quantized	N/A	5.3 ± 2.3	98.2 ± 0.3
	IF	zscore	N/A	3.4 ± 0.6	97.4 ± 0.6
	OCSVM	quantized	N/A	8.6 ± 0.0	$\underline{98.8 \pm 0.0}$
	OCSVM	zscore	N/A	5.2 ± 0.0	97.2 ± 0.0

HP choice	model	preprocessing	PR_{synth}	$PR_{ERPSim(2)}$	$ROC_{ERPSim(2)}$
synthetic	AE	synthetic	$\underline{16.6 \pm 1.0}$	$\mathbf{53.7 \pm 5.7}$	$\mathbf{99.4 \pm 0.3}$
	IF	synthetic	10.9 ± 0.7	21.3 ± 2.6	98.2 ± 0.3
	OCSVM	synthetic	$\mathbf{17.4 \pm 0.0}$	$\underline{38.3 \pm 0.0}$	92.3 ± 0.0
default	AE	quantized	N/A	24.4 ± 13.6	99.2 ± 0.2
	AE	zscore	N/A	8.3 ± 2.3	98.3 ± 0.3
	IF	quantized	N/A	10.5 ± 2.0	98.9 ± 0.2
	IF	zscore	N/A	11.3 ± 4.0	98.7 ± 0.2
	OCSVM	quantized	N/A	23.7 ± 0.0	$\underline{99.3 \pm 0.0}$
	OCSVM	zscore	N/A	11.0 ± 0.0	98.2 ± 0.0

Limitations and Conclusion

- Limitations:
  - Proof of concept including only basic economic procedures and fraud cases
  - May be extended with new agents and strategies to cope with complexity of real-world data
- Conclusion
  - Framework for detecting occupational fraud in unlabeled ERP data using a multi-agent system
  - MAS-simulated data resembles data characteristics of real data or given economic trends
  - This approach allows preprocessing, model and hyperparameter selection without the need of labeled validation data

Links and Conclusion

https://professor-x.de/erp-fraud-data

ERPSim data

https://professor-x.de/erp-fraud-mas

MAS Code

Daniel Schlör

Data Science chair
University of Wuerzburg, Germany
schloer@informatik.uni-wuerzburg.de
https://www.dmir.org

Contact

- Conclusion
  - Framework for detecting occupational fraud in unlabeled ERP data using a multi-agent system
  - MAS-simulated data resembles data characteristics of real data or given economic trends
  - This approach allows preprocessing, model and hyperparameter selection without the need of labeled validation data