AI for Audit Case Selection

Experiences, Caveats and Lessons Learned

Charles Blanco

Senior Global Public Sector Specialist

Economic Policy & Growth

Global Unit

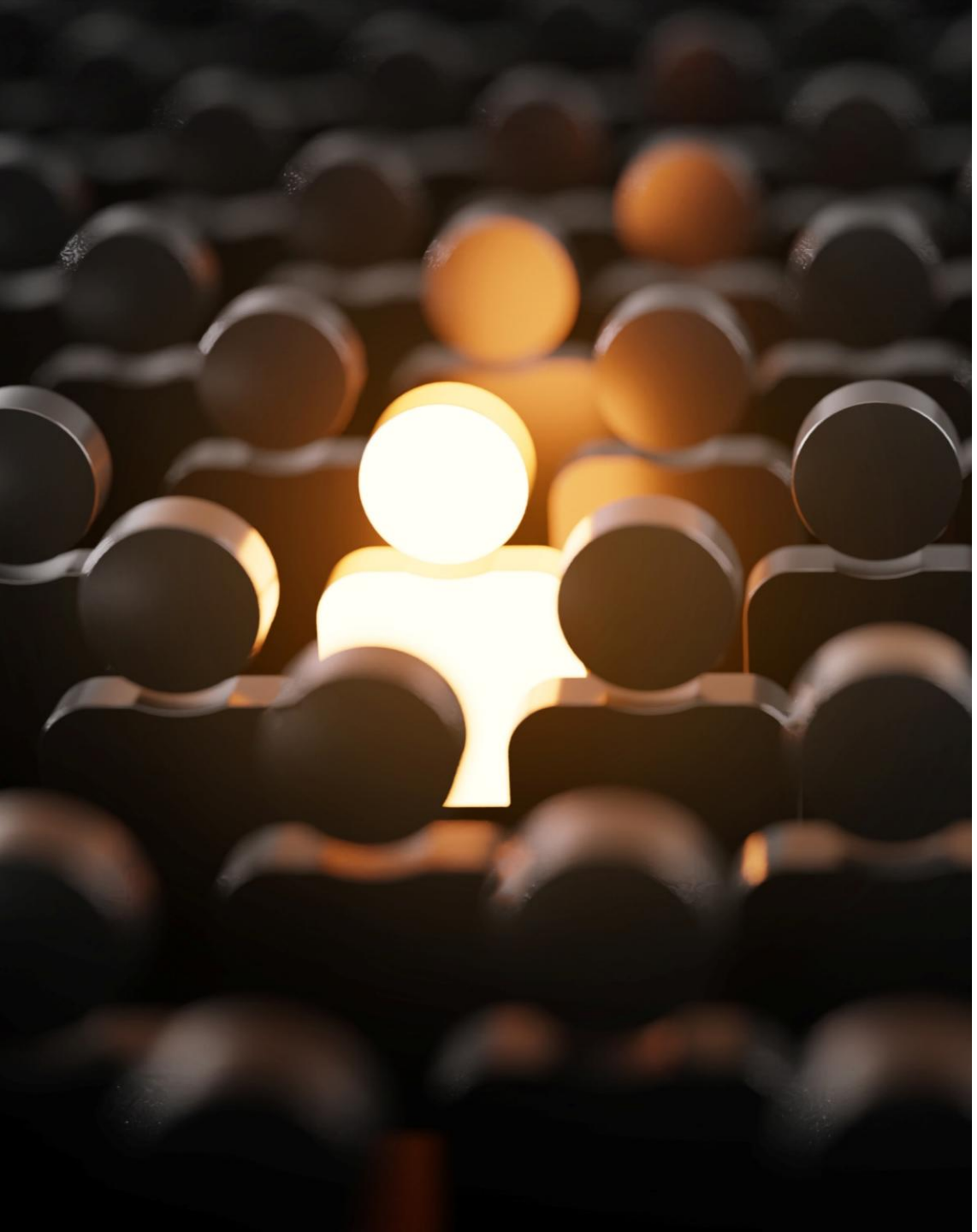
Ivan Krsul

Senior Governance Specialist

Economic Policy & Growth

Global Unit

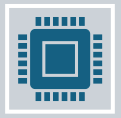
WORLD BANK GROUP



Current State

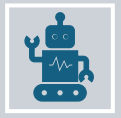
- Traditional audit selection methods (rule-based systems) struggle to detect sophisticated tax evasion tactics effectively.
- Limited collaboration, absence of standardized reference datasets, and privacy concerns hinder progress.
- Machine learning models can transform audit case selection, but real-world validation with robust datasets is crucial.

Problem Statement



Training AI systems for audit case selection is difficult, requires extensive prior data and requires collaboration.

There is no collaboration with external organizations to iteratively improve architectures because of privacy concerns.



Testing various AI architectures is time consuming.

Limited validation of AI solutions.

Solution

- Create a synthetic data repository that can be used by tax administrations to propose novel AI architectures for AI systems for audit case selection.
- Enable sharing architecture and training data with other administrations without violating privacy and confidentiality of data.

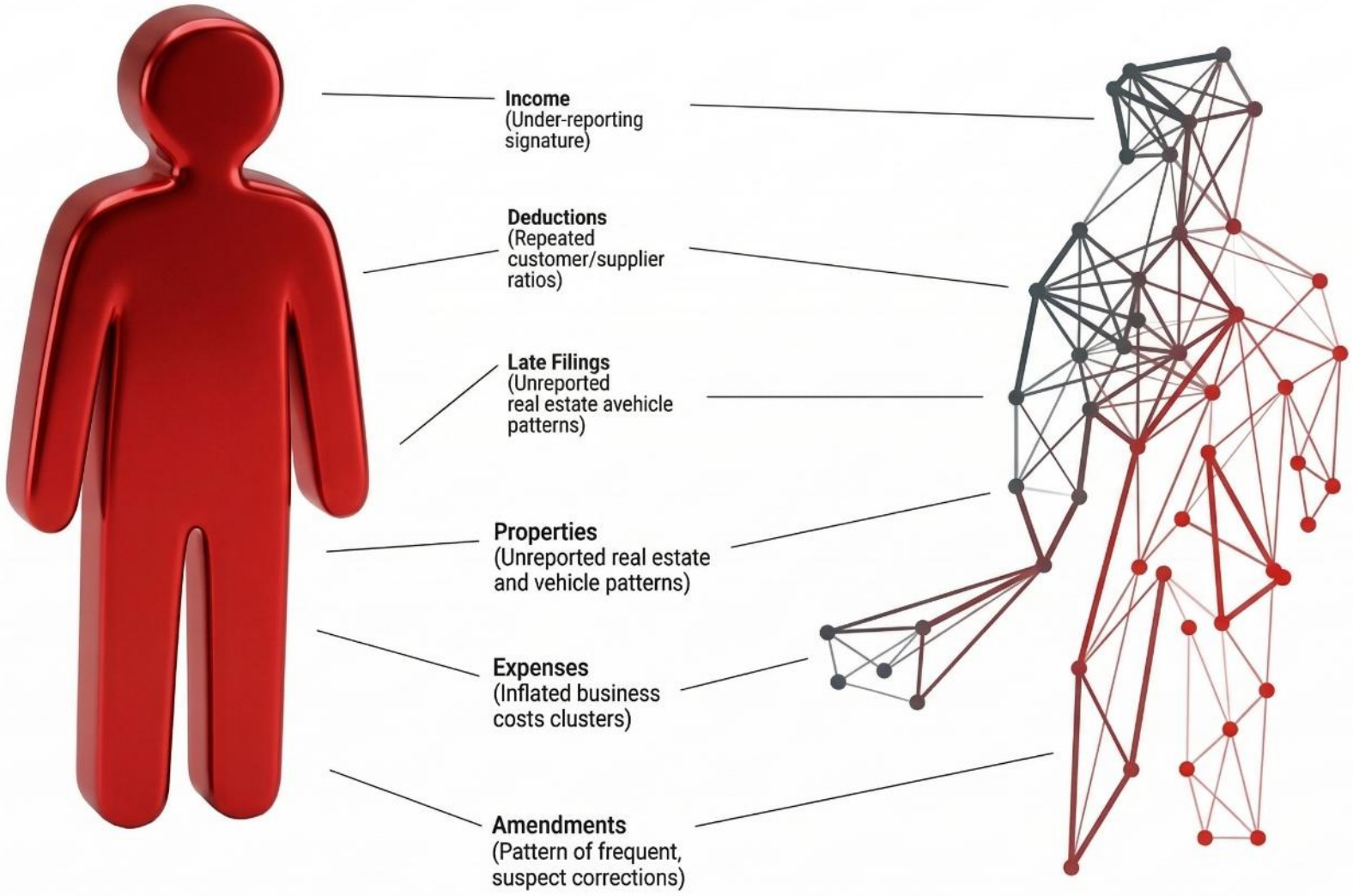
Objective

- Complement existing audit selection techniques with AI systems to enhance tax audit efficiency to boost tax revenues, increase risk perception and reduce administrative costs.

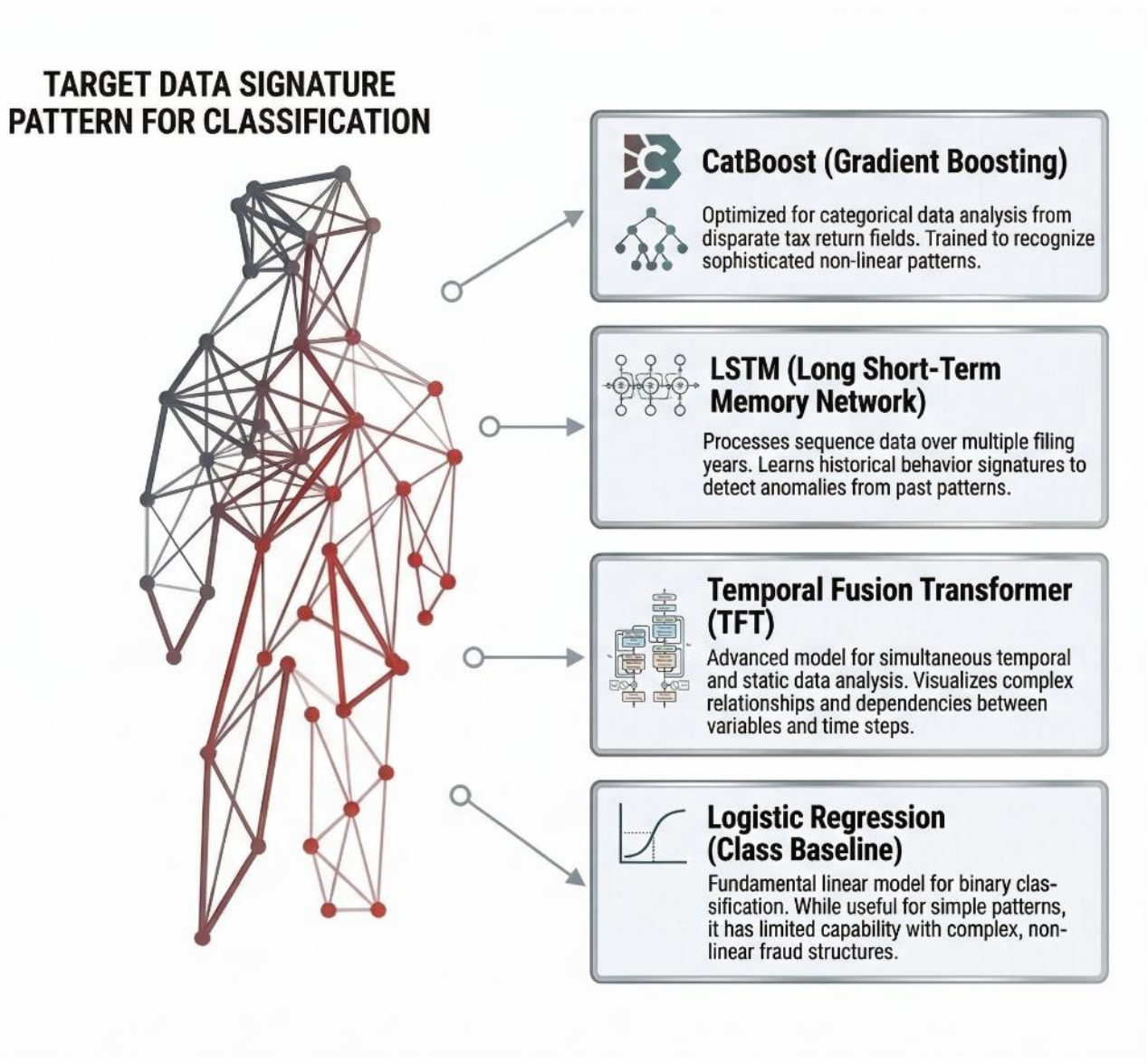
AI for Audit Selection Process

Create a Synthetic Data Repository (SDR)	**Based on International experiences and trends**
Customize the SDR using country data statistical distribution	**Based on internal & external experiences**
Developed a Machine Learning Algorithm (MLA) based on examples	**To identify irregularities or inconsistencies**
Obtain anonymized data on tax audits and other processes	**Strengthens the model**
Fine-tune the MLA using new SDR data and enhance to better identify irregularities	**To calibrate the model**
Test MLA with real anonymized data	**To test the model**
Add Data to the Model and embed it in existing systems	**To strengthen current Systems and methods**

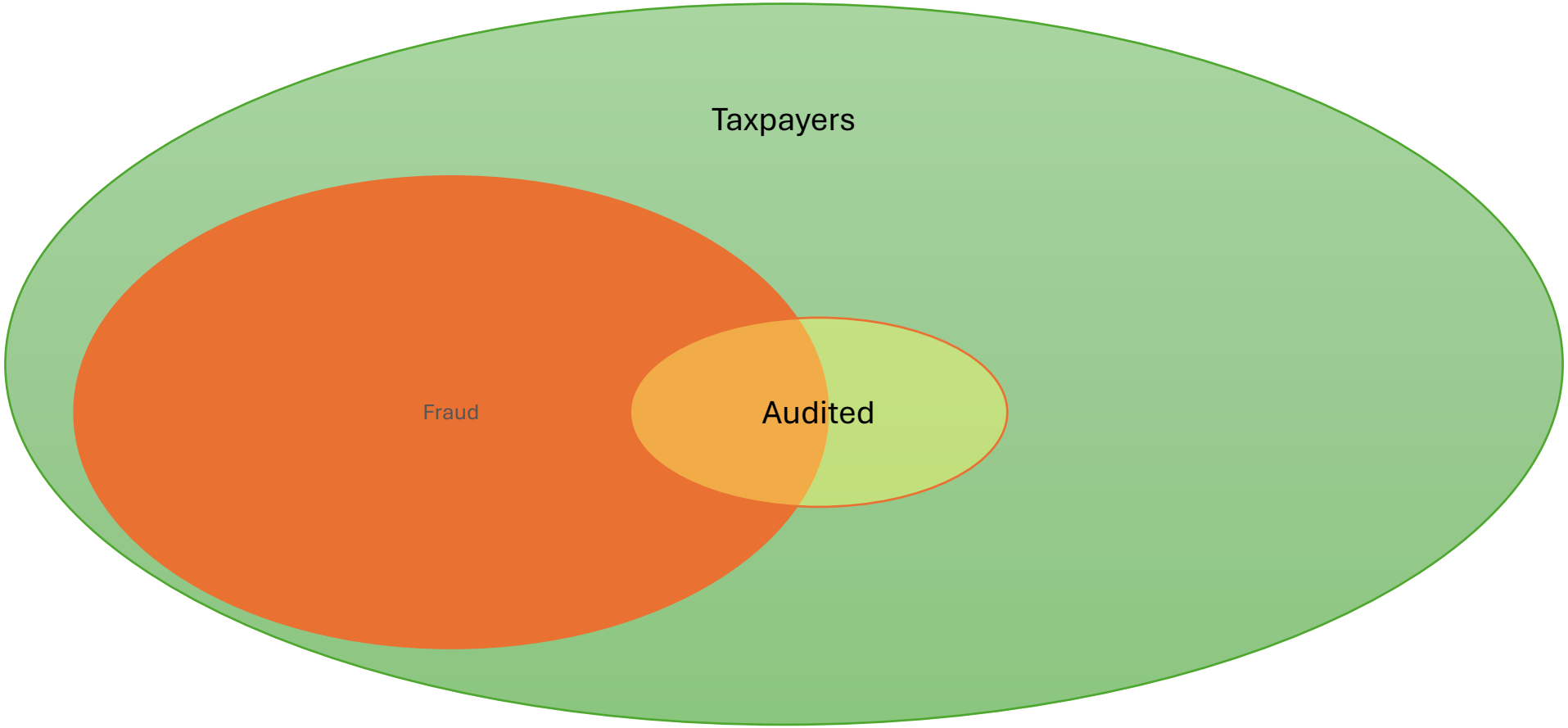
We assume that fraudulent taxpayers fit a pattern



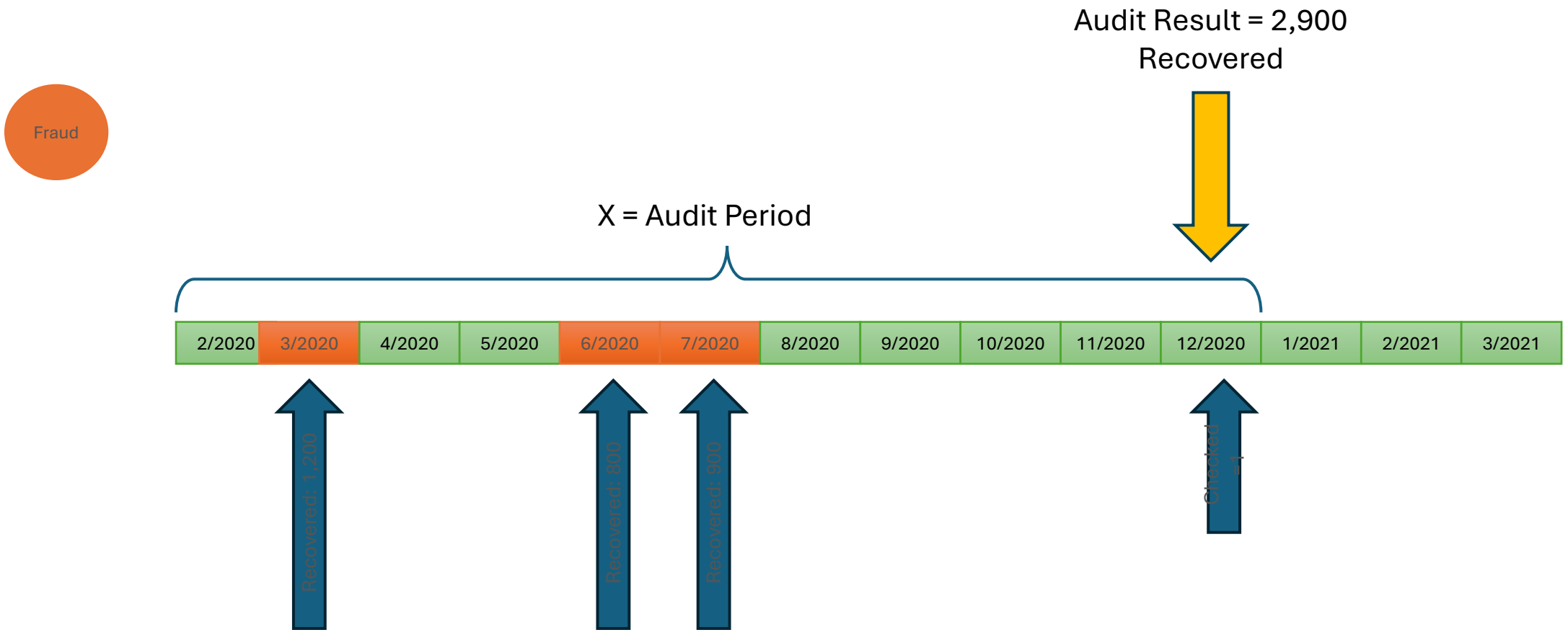
If a pattern exists, we can train an ML algorithm to detect it



Labels: What does fraud look like?



Fraud does not happen at the audit point



A Useful Analogy

Go to the doctor and tell him:

In the past 3 years I had high-blood pressure

In the past 3 years I had severe headaches

In the past 3 years I had pain in my stomach

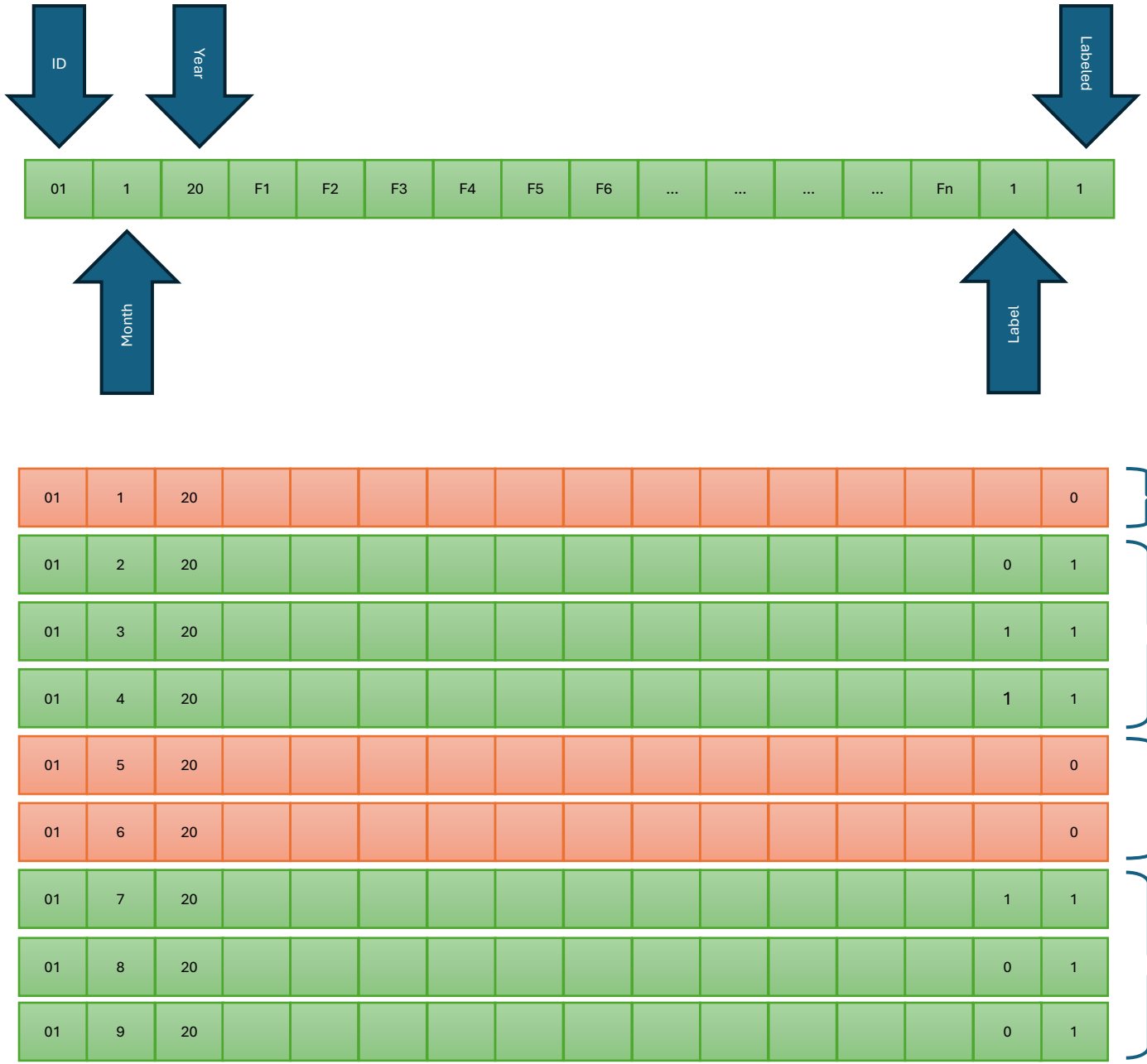
In the past 3 years I had a fever

In the past 3 years I had normal blood work

In the past 3 years I had an infection

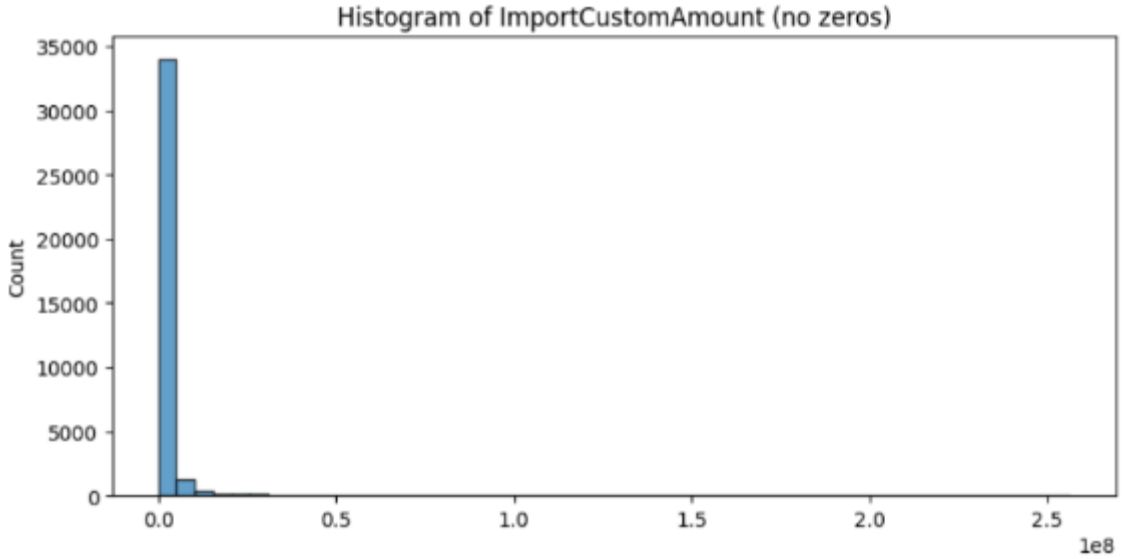
Audit reports are similar. When, and in what sequence matters.

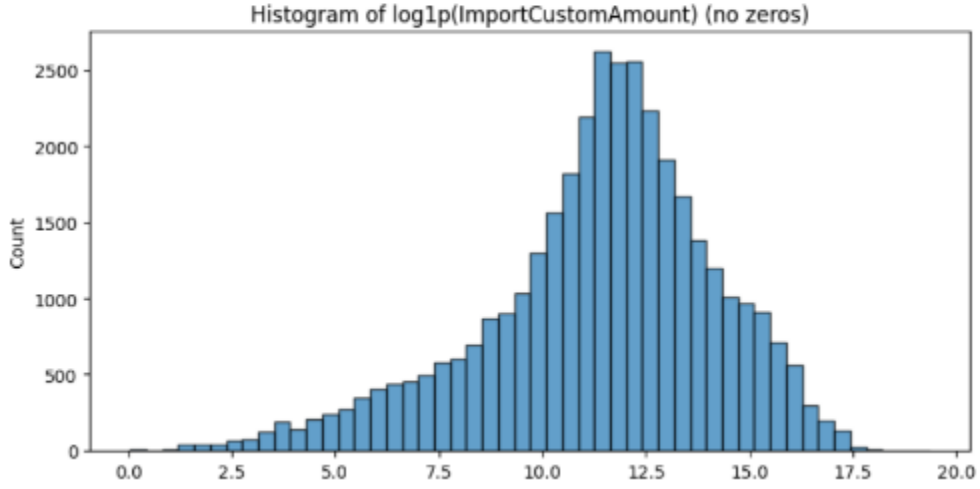
Observable and unobservable data



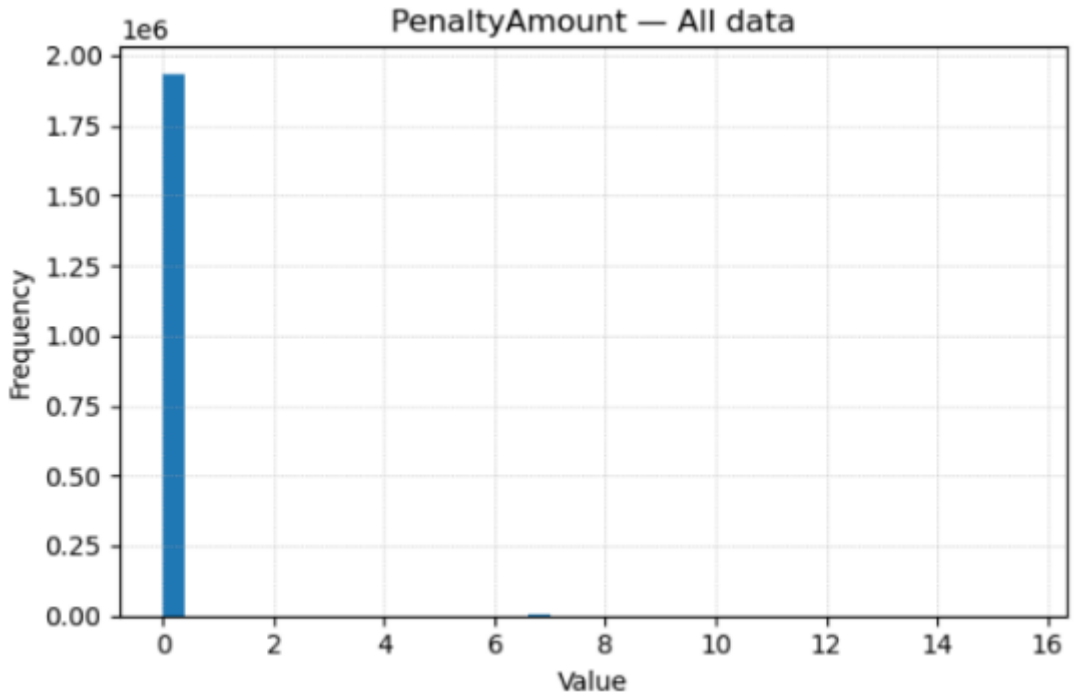
90+% of Effort is Data Cleaning

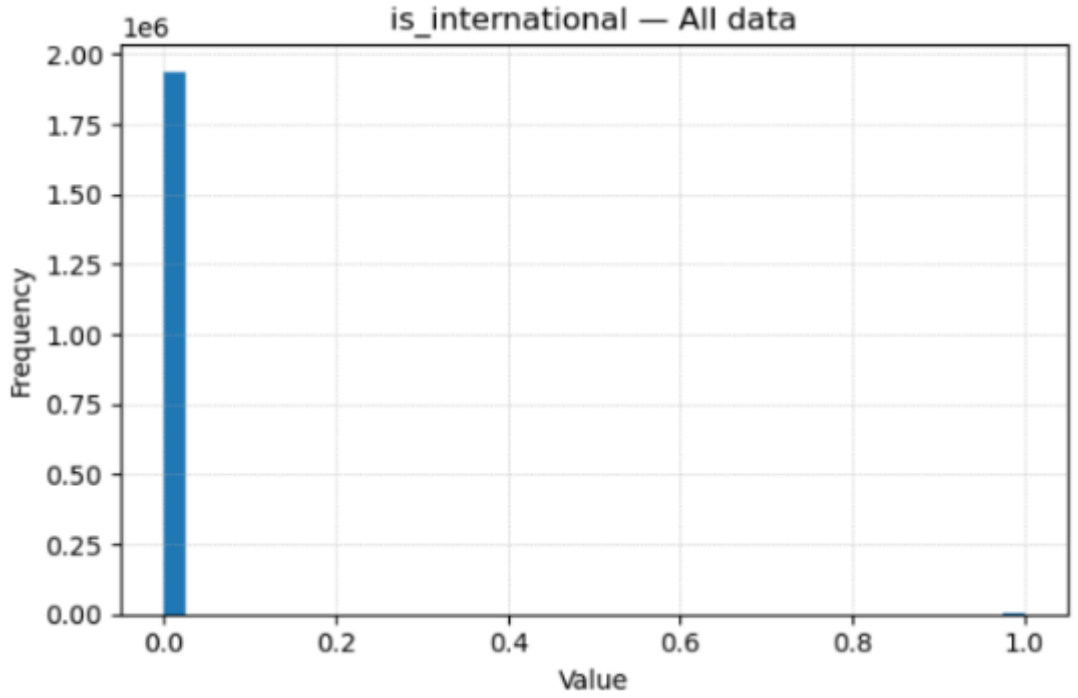
Data Disparity



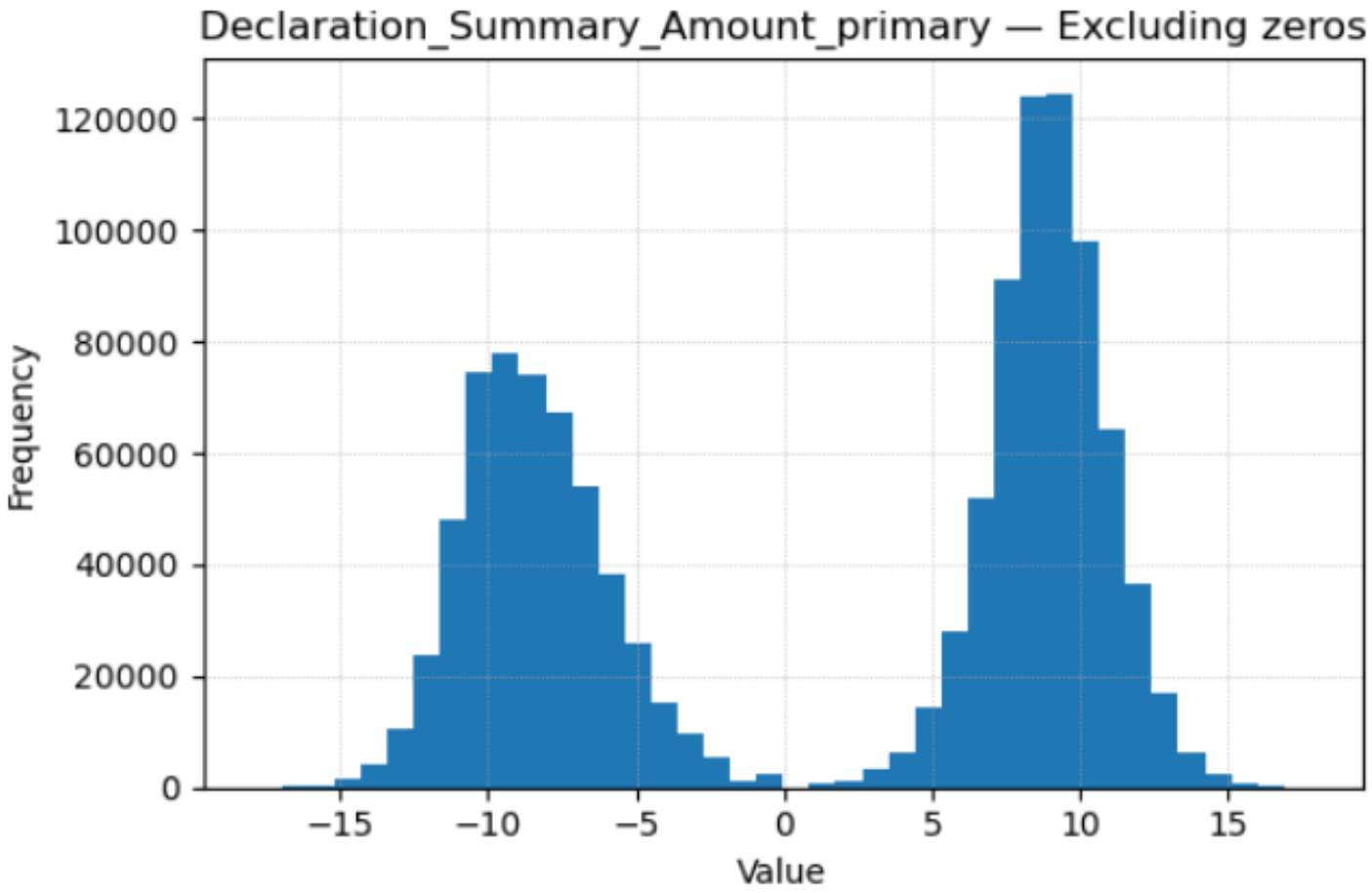


Low Data Variability



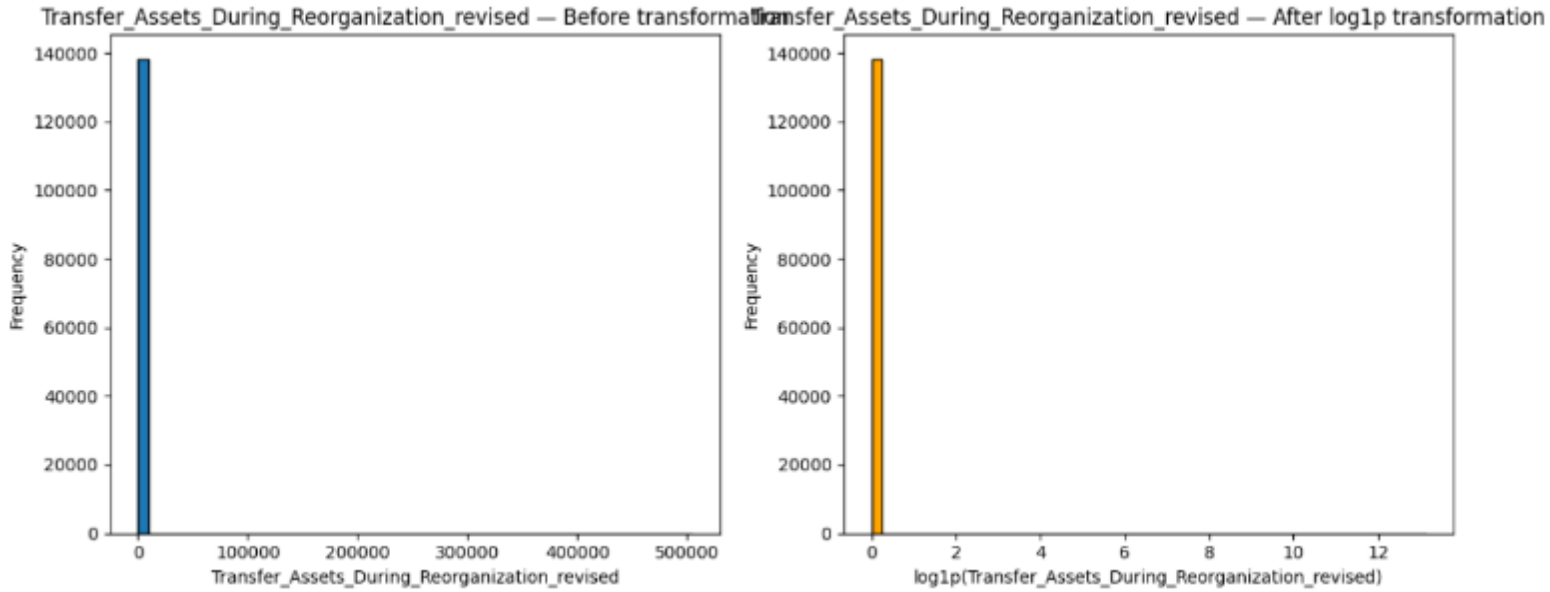


Strange Data

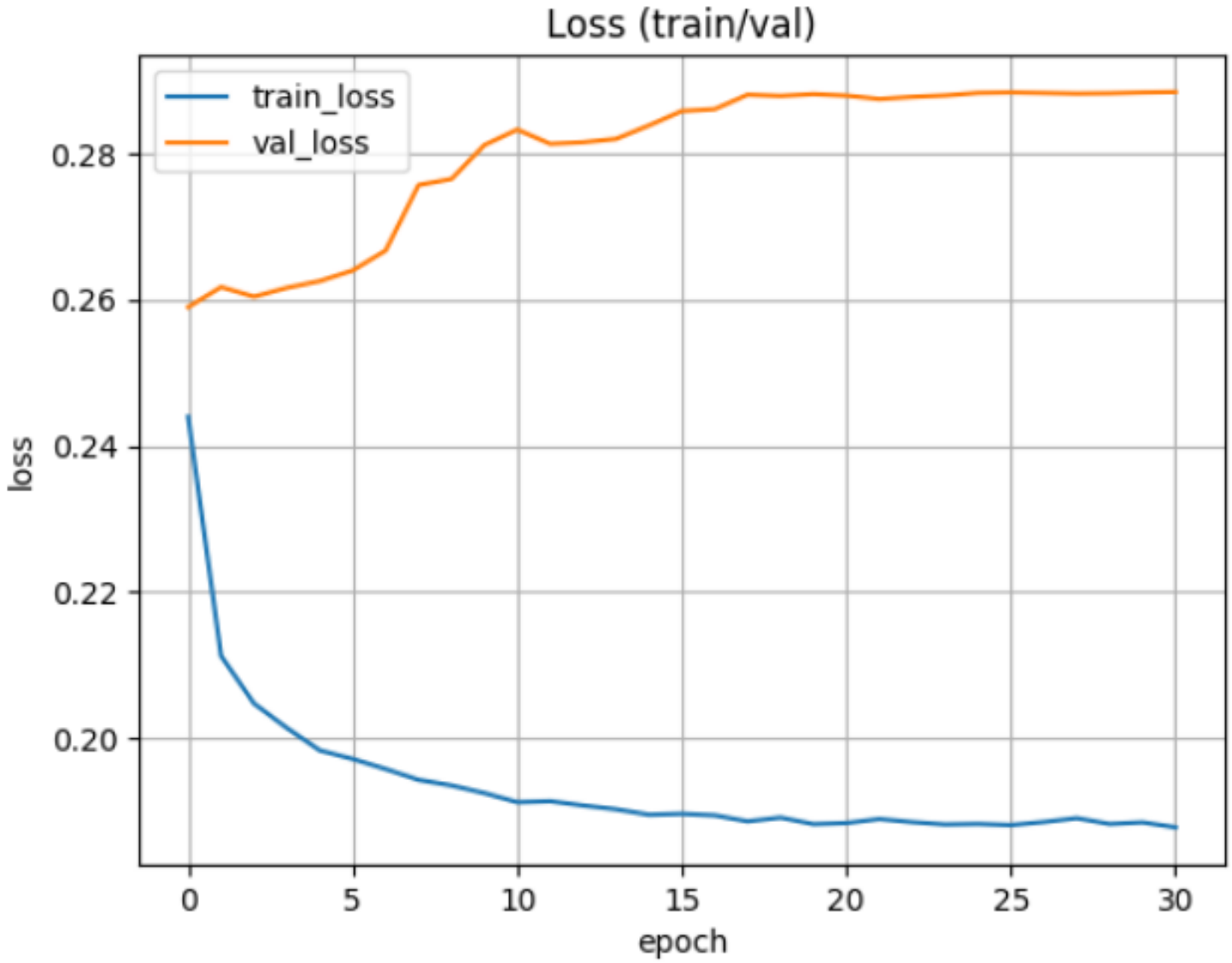


Low Variabilty

```
Low-variability boolean features (dominance > 0.98):
is_declared_paper               0.999964
owner_changed                   0.999639
is_international                0.994366
nb_board_foreigners_missing     0.992581
delayed_custom_declarations      0.99191
is_address_changed              0.985654
imported_goods_62               0.985394
imported_goods_61               0.985221
imported_goods_33               0.984874
imported_goods_27               0.983993
imported_goods_87               0.983371
imported_goods_68               0.982584
imported_goods_38               0.981277
NACE_code_short_10              0.980338
imported_goods_63               0.980258
dtype: Float64
```



Patterns are Company Specific (Not Generalizable)



Limited Resources

Training Advanced Models Requires Resources

Hardware:

GPUs and Powerful Workstations with lots of RAM

People:

Basic knowledge of Python Programming, Data Sciences, Data Analytics

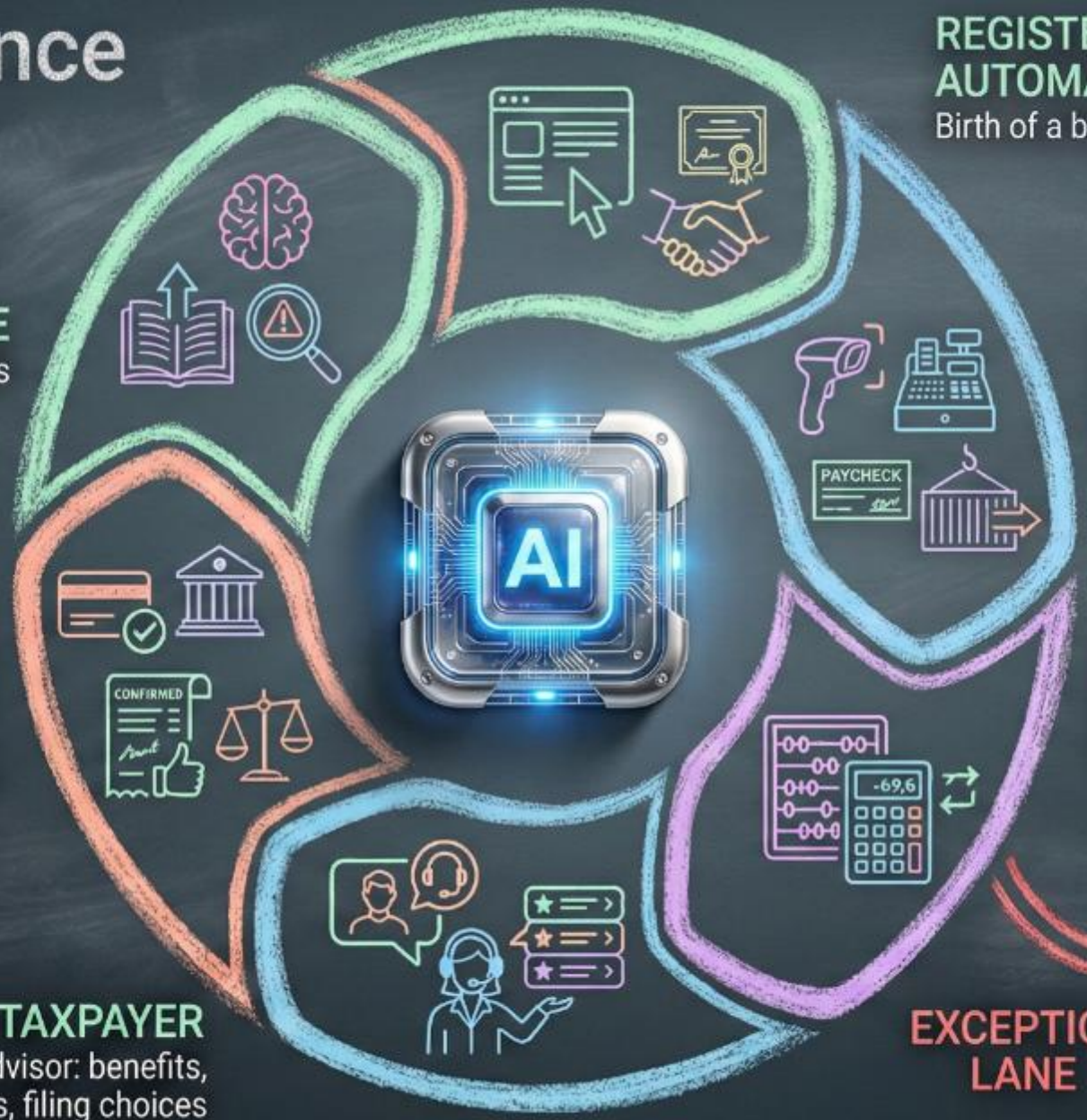
Time:

Experimentation, Iteration, Data Cleaning

Discipline:

Meticulous tracing of data transformations, data cleaning for future replicability

The Compliance Flywheel



REGISTER & UPDATE AUTOMATICALLY
Birth of a business, address change, VAT status

CAPTURE TRANSACTIONS
E-invoicing, POS, payroll, customs

COMPUTE OBLIGATIONS CONTINUOUSLY
Near-real-time liability estimates

EXCEPTION LANE

HUMAN
Audits, fraud, litigation

SETTLE & PAY SEAMLESSLY
With confirmations and dispute options

GUIDE THE TAXPAYER
Chat/agent advisor: benefits, credits, filing choices

LEARN & IMPROVE
Feedback loop + risk signals

Thank you!

Charles Blanco

cblanco@worldbank.org

Ivan Krsul

ikrsul@worldbank.org

WORLD BANK GROUP