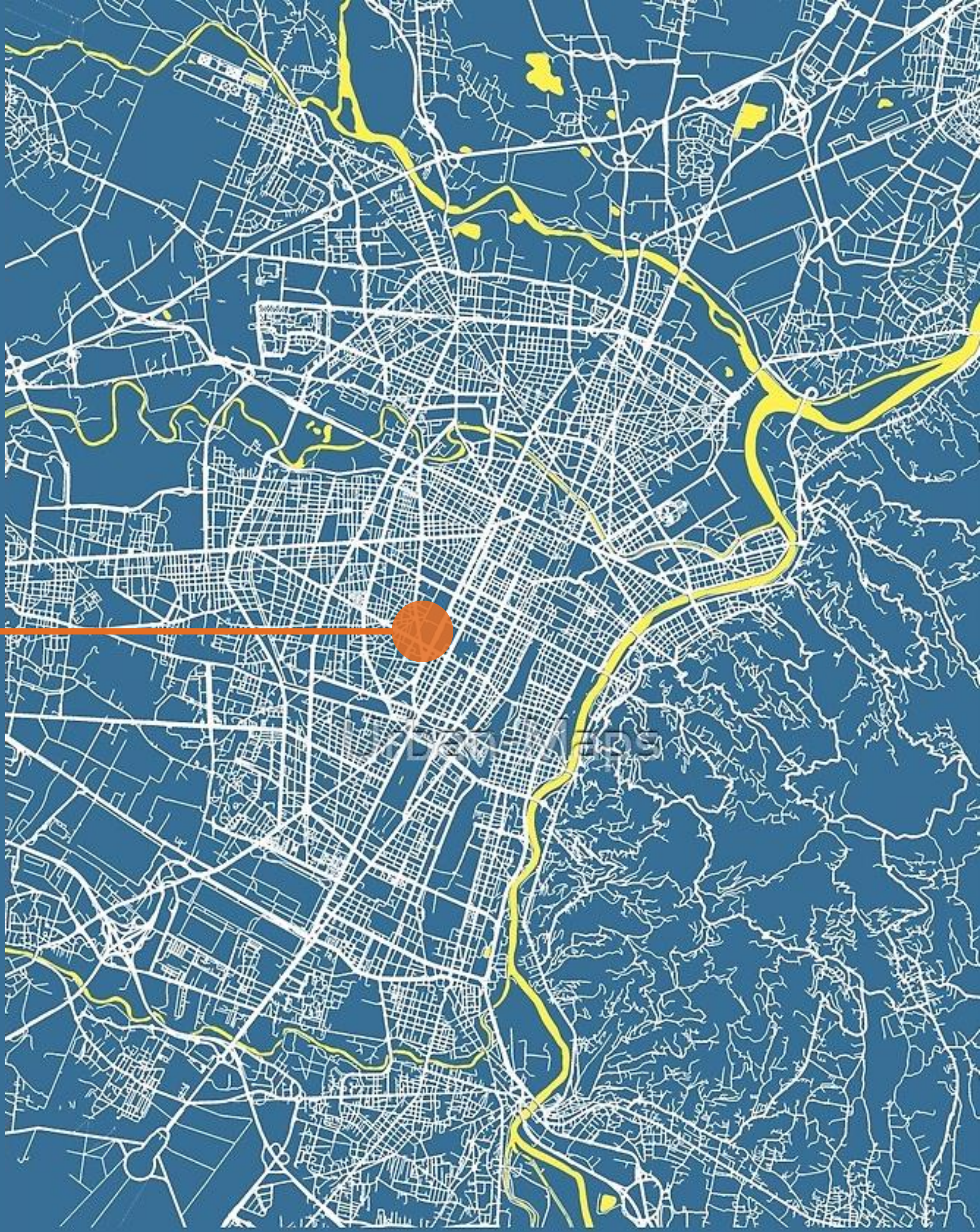
Spoke 3

DS&AI Office

Intesa Sanpaolo



Roma, 28.10.2022

Use Cases

WP 3.2

1. Fraud Detection on Transactions
2. Corporate credit risk
3. Churn analysis

Challenges on Time Series:

- Sampling strategy
- XAI on TS
- Covariate shift

FDoT - data

Data info:

- Client id
- Transaction timestamp
- Transaction type
- Other info
  - device
  - beneficiary
  - GPS coordinate
  - …

not all the information is available for all types of transactions (NaNs)

FDoT – As Is

x_1	x_2	x_3	x_4	y
A	0	0,2	500	0
B	1	0	600	0
A	1	0	300	0
B	1	0,3	0	0
C	1	0,1	0	0
C	0	0,2	0	1

FDoT – Time Series algorithm

The idea is to **train a network** that, **given a client's previous actions** (different length per each client), learns to **predict the next action** in the sequence (or a set of subsequent actions with a certain probability distribution). And then, when the customer performs the real action, **compare** (how?) The **actual action with the predicted one** and return a (dis) similarity score.

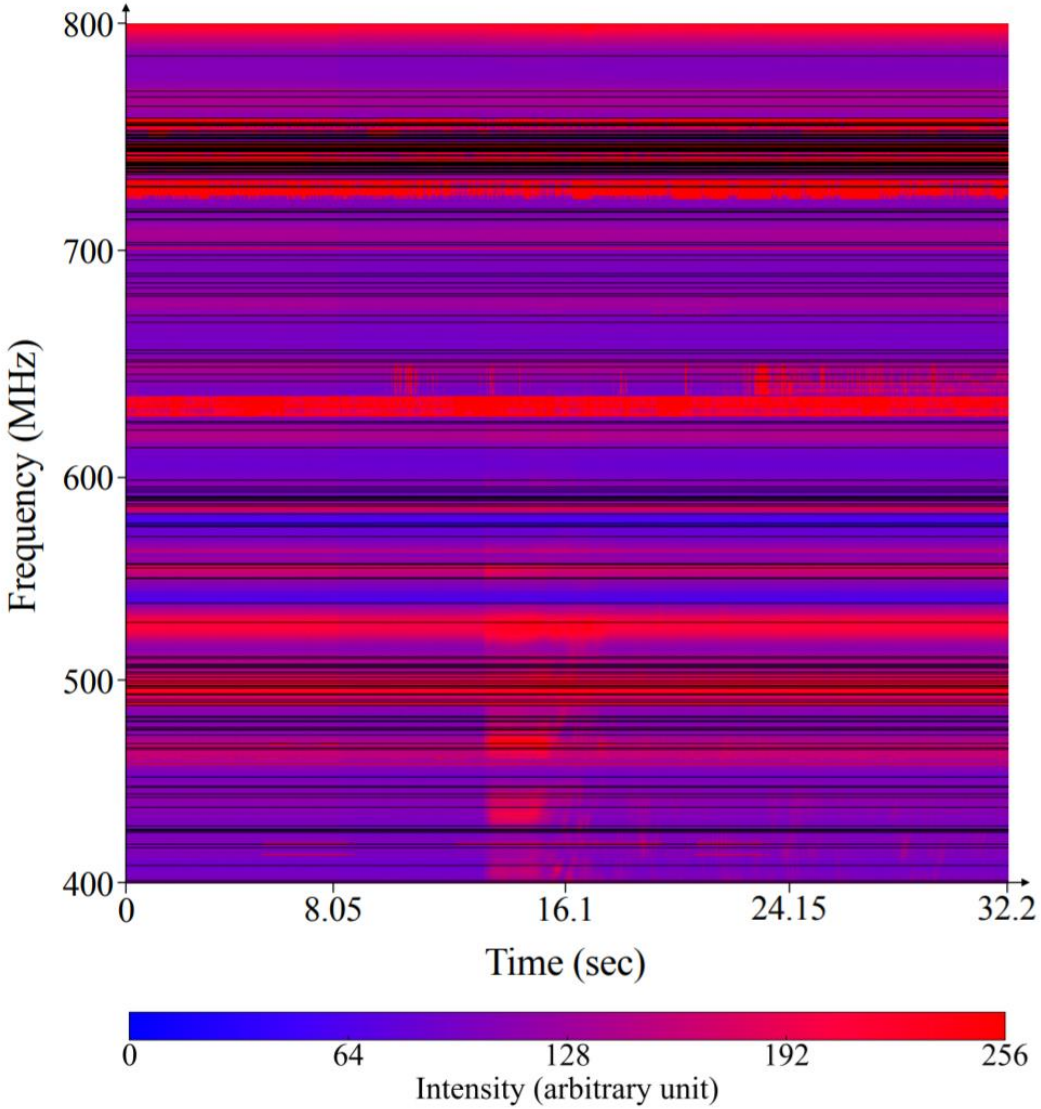
Table example

Timestamp	1	10	12	20	40	100	130
T-type	AA	CC	BB	BB	BB	AA	BB
Beneficiario	1235	NaN	1235	NaN	NaN	1111	1235
Importo	10,00	100,54	1,00	30,82	455,40	1000,00	15,00
…							
Fraud	0	0	0	0	0	1	0

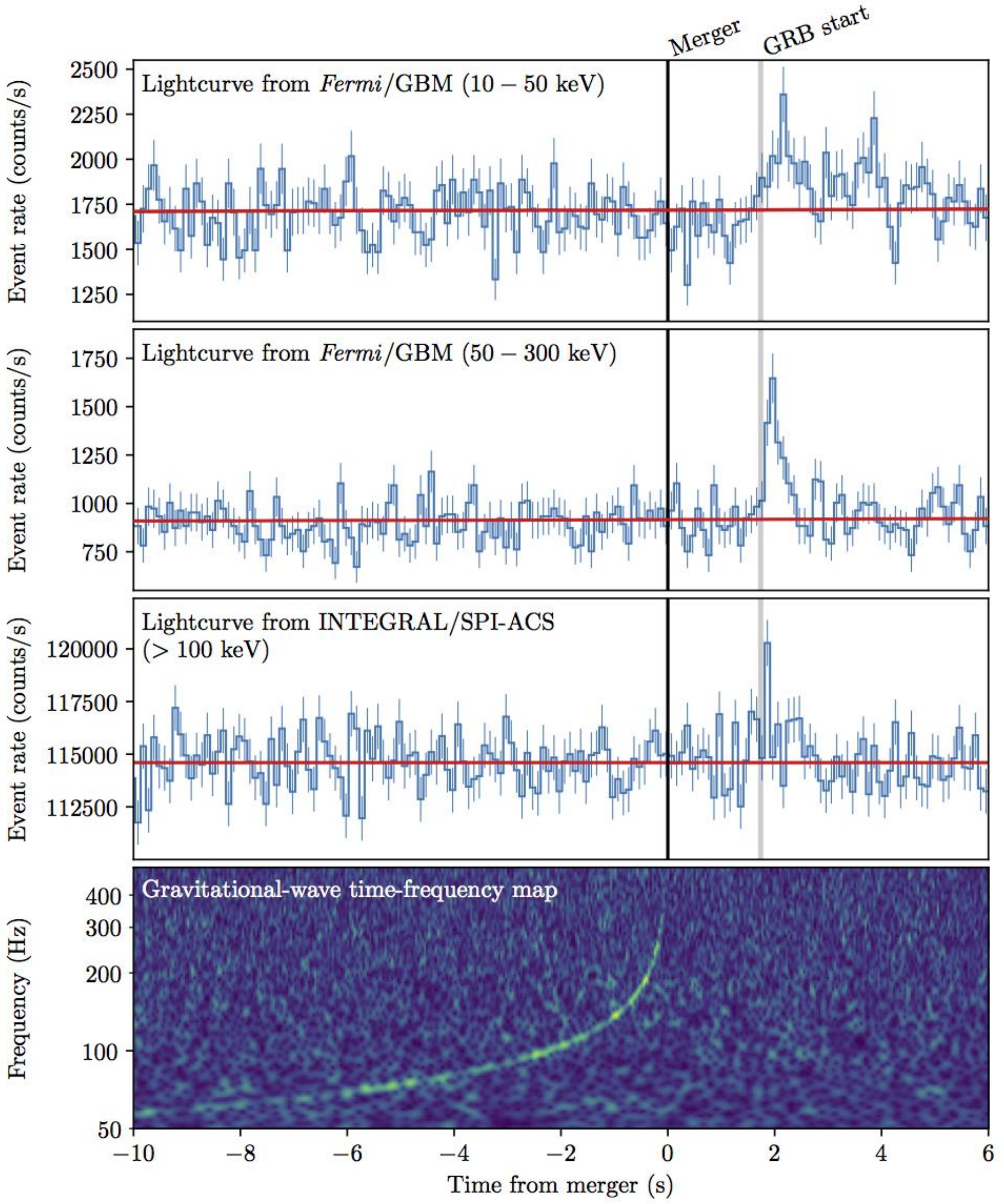
Table example

Timestamp	1	10	12	20	40	100	130
T-type	0.9	0.8	0.99	0.87	0.76	0.1	0.75
Beneficiario	0.6	0.8	0.4	0.7	0.8	0.2	0.6
Importo	0.98	0.89	1,0	0.5	0.4	0.5	0.7
…							
Fraud	0	0	0	0	0	1	0

FDOT – similarity with FRB



FDOT – similarity with GRB



https://indico.in2p3.fr/event/26887/contributions/111841/attachments/71385/101602/Crupi_Poster.pdf

https://arxiv.org/abs/2105.08841

https://pos.sissa.it/358/775/pdf

FDoT – sampling strategy

Problem definition: In Fraud Detection on Transactions there is a huge amount of transactions that are no fraud. It is required an undersample of this set to exploit the eterogeneity of the non-fraud transactions. The data are tabular, the target is binary

Solution now: plain subsampling

Possible solutions: Clustering on clients (graph on clients or in a latent space) and subsample on cluster

XAI on timeseries

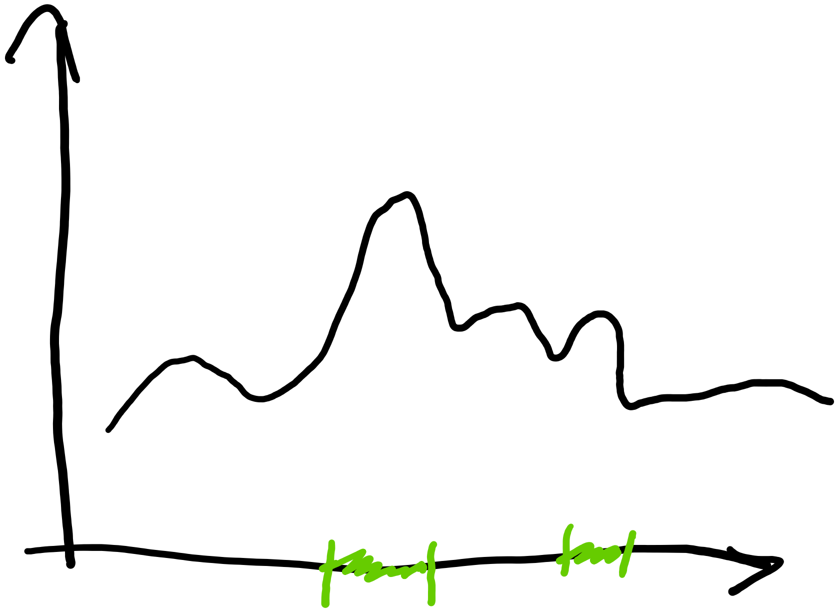
Problem definition: In the time series machine learning, es. Algorithm of NLP (sentiment analysis or question/answer), we need methods to highlight the token important for the prediction. Time Series data format

Solution now: no explainability

Possibile solution: Assess library open source for time series explainability or exploit the transformer architeture

XAI on timeseries

The dog broke everything.



Corporate credit risk

Problem definition:

Predict the *status* of a «**performing**» client after 6 months from the reference time of observation (i.e. predict if the client will be able to repay his loans/debts).

Challanges:

- **Default rate** of «performing» clients is low ranging from **1 to 3%**.
- **Feature distribution can vary significantly** in test set if compared to training set due to unexpected major events (e.g. covid crisis, bank merge and acquisitions, …)
- The nature of clients in **perimeter is heterogeneous** (natural persons, sole proprietorships, shared accounts, ..)
- The output of the model must be **explainable**

Current approach: classification model (XGBoost), binary target, tabular data

Possible alternative approaches :

- Use a generative model to assess **covariate shift**
- **Clustering** on clients (graph on clients or in a latent space) and subsample on cluster
- **Explainable** Time series
- …

Table example

Features\Timestamp	1	10	12	20	40	100	130
T-type	AA	CC	BB	BB	BB	AA	BB
sconfino	0,00	0,54	1,00	30,82	455,40	1000,00	15,00
Beneficiario	1235	NaN	1235	NaN	NaN	1111	1235
Importo bon	10,00	100,54	1,00	30,82	455,40	1000,00	15,00
…							
Default	0	0	0	0	0	1	3

We are working to share this dataset

Tables in CCR

- **ANAG** = This is basically a table version of a customer registry. This will enable us to do an initial analysis on the customers based on their origin(ITALY/FOREIGN), industrial activity (Trade, construction, hotel etc), province and many more. Hence we could understand the portfolio of customers sector.
- **BONIFICI** = The details of bank transactions extracted which is an essential data to build the network of customer and suppliers
- **SCONFINI** = This contains the information about the overdrafts of companies. The companies withdraw money from the bank up to certain limit to fund their short term cash requirements. This data is considered separately as a predictor variable since this aspect was not considered when determining MOCRED.
- **Stato Amministrativo (target)** = This contains the data of customers who obtained a classification code by the bank based on their financial distress. The COD_CPSR of raw data is transformed to a code of 1 or 0 in such a way that the real defaulted(liquidated) companies, the customers who haven’t paid the bank for 90 to 180 days and the customers who are in the risk of being default (as identified by the bank) are given a code 1 while the other Codes a zero.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250115

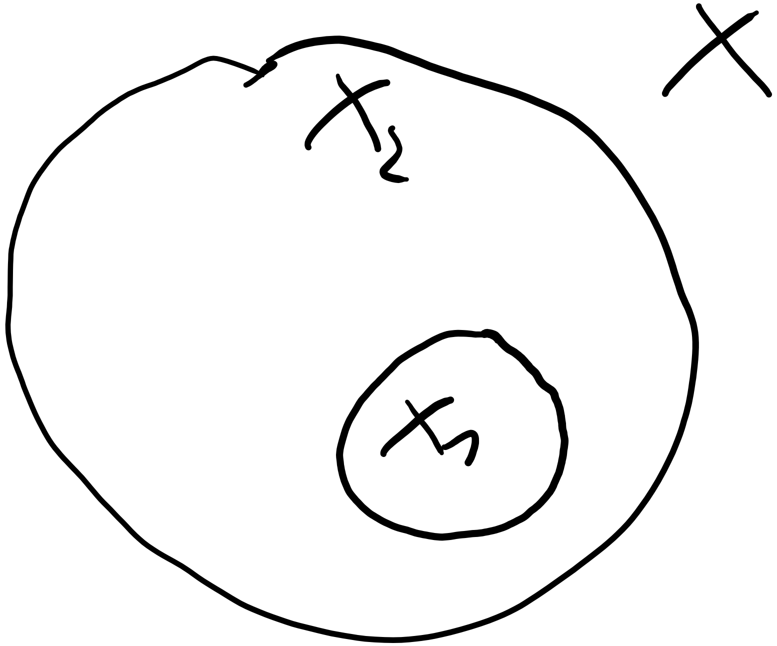
covariate shift

Problem definition: the **distribution** of variables in test phase could vary significantly if compared to **training** set because of change in time of clients behaviour or **unavaiability of distribution** of test sample.
This problem is stated as covariate shift.
The data are tabular, the target is multiclass .

Solution now: none

Possible solution: Implement a generative model (e.g. GAN or VAE) to trasform the clients in train set more similar to test set.
Stratified Learning: a general-purpose statistical method for improved learning under Covariate Shift

covariate shift



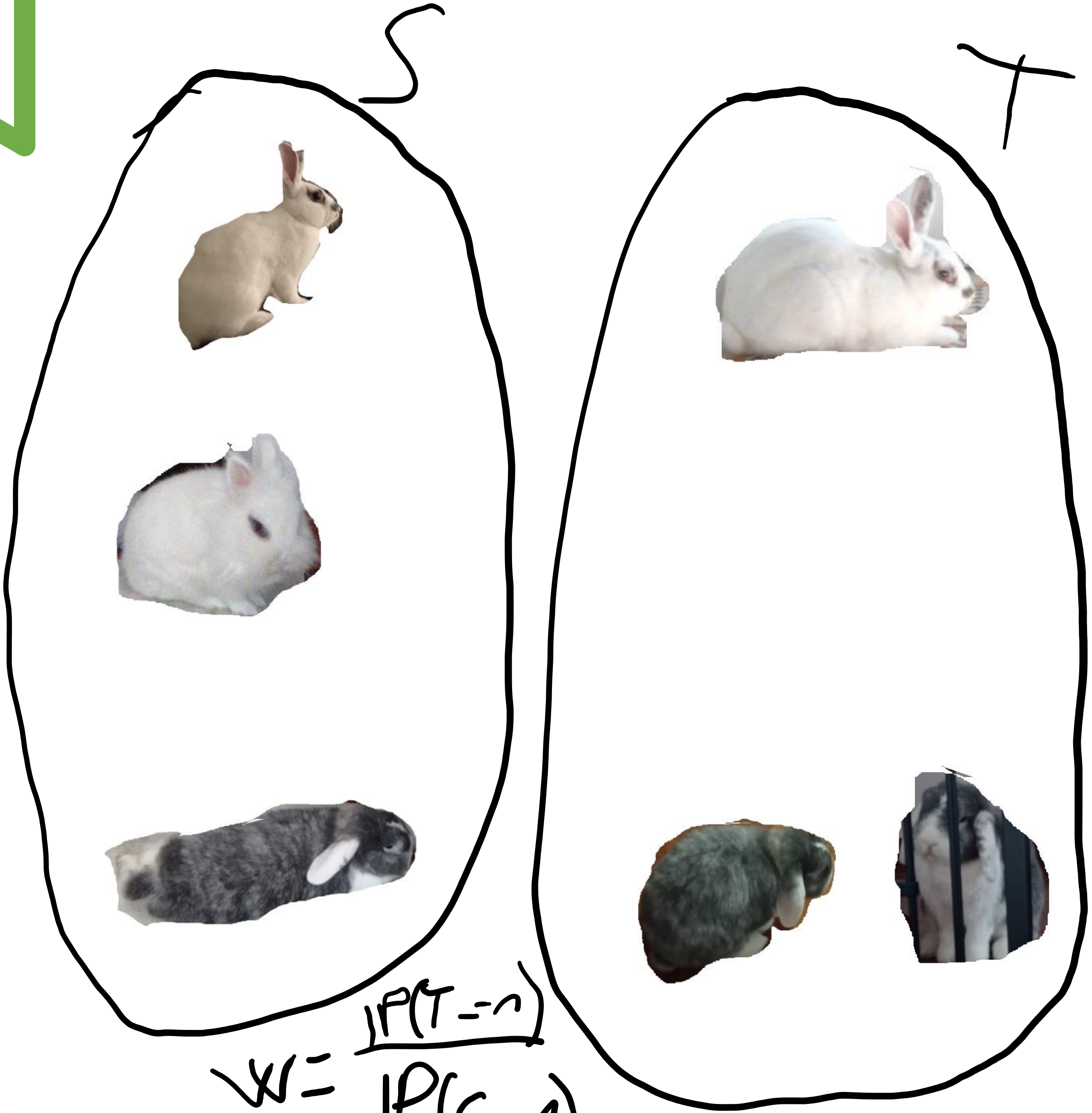
$X = X_1 \cup X_2$ $f: X_1 \to Y$

X_1 : clients that vote

X_2 : // // don't //

$M: X_1 \to X_2$

covariate shift



$$w = \frac{\mathbb{P}(T=1)}{\mathbb{P}(S=1)}$$

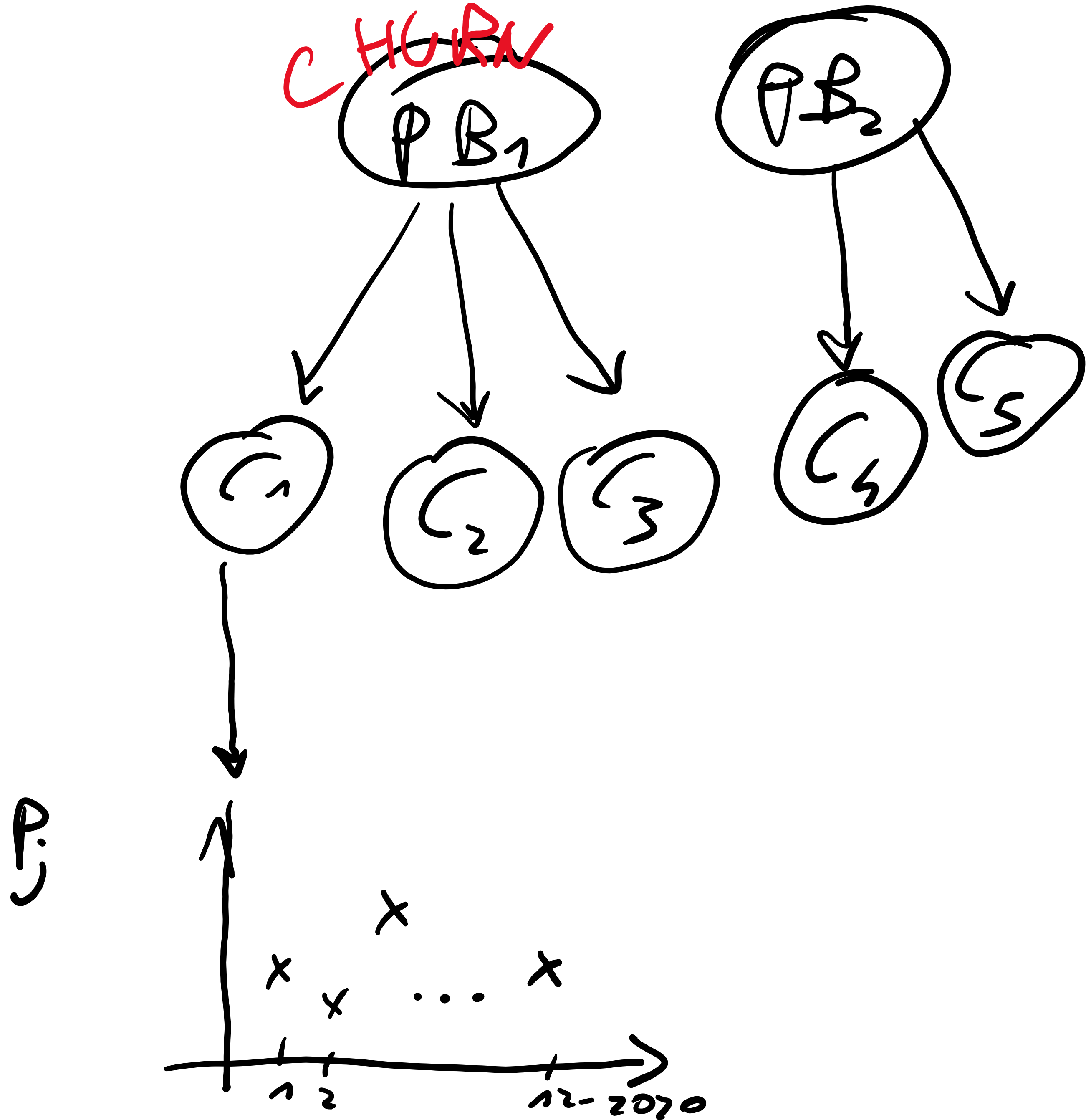
Churn analysis

Problem definition: The **Private Bankers** (PB) are the professionals that **manage** the **portfoglio** of several **clients**. Rarely happen that the **PB quits** to take an opportinuty in another bank. The **aim** is to **indetify** in the operations of clients (time series) the pattern to anticipate **this behaviour**.

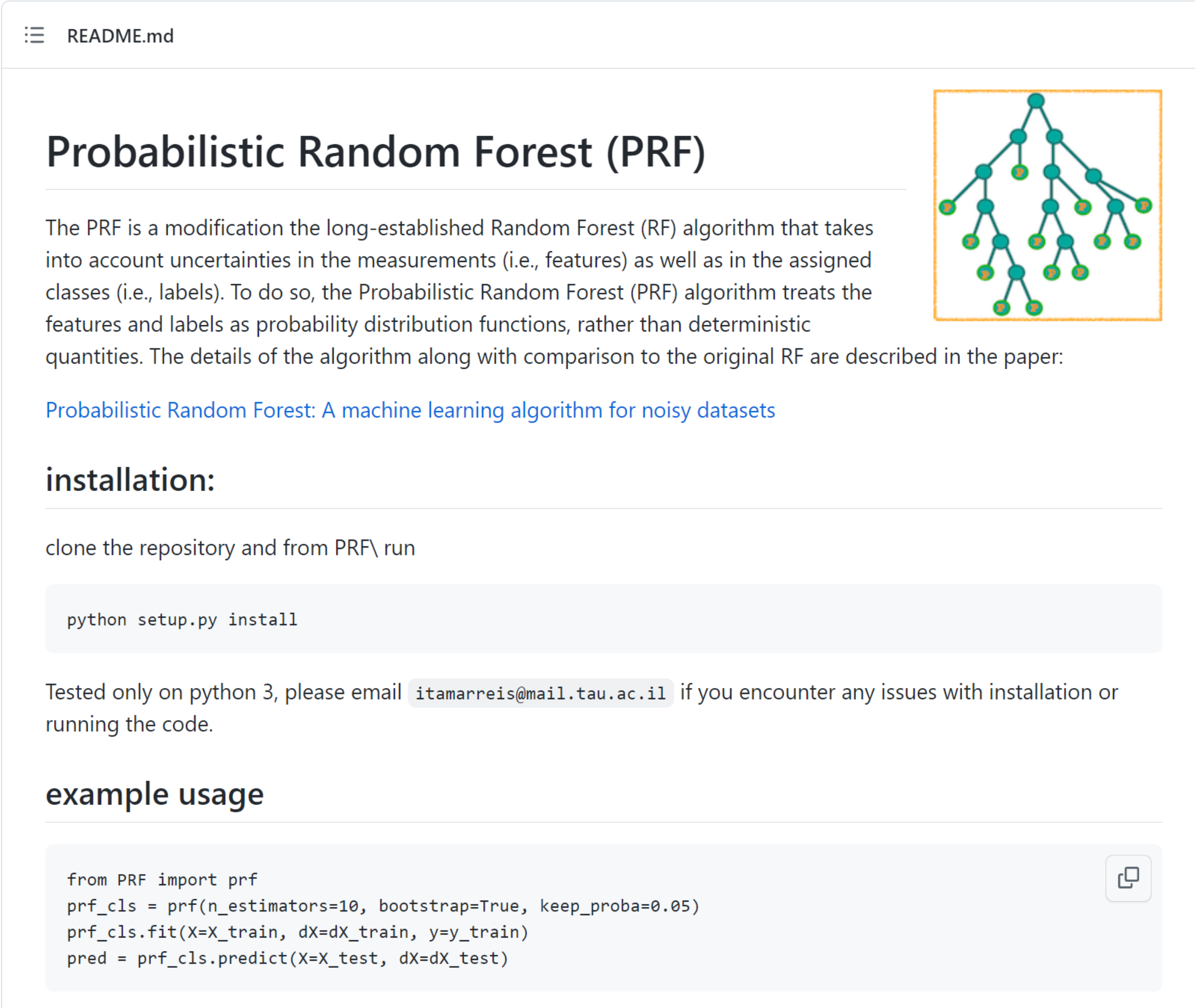
Solution now: classifier on tabular data based on cumulated montly operation per each client with binary target (1 if the PB of that client quit)

Possible solution: Exploit the data directly in time series. Use method for rare events: Anomaly detection? E.g. CVAE

Churn analysis



Result of the research



Probabilistic Random Forest (PRF)

The PRF is a modification the long-established Random Forest (RF) algorithm that takes into account uncertainties in the measurements (i.e., features) as well as in the assigned classes (i.e., labels). To do so, the Probabilistic Random Forest (PRF) algorithm treats the features and labels as probability distribution functions, rather than deterministic quantities. The details of the algorithm along with comparison to the original RF are described in the paper:

Probabilistic Random Forest: A machine learning algorithm for noisy datasets

installation:

clone the repository and from PRF\ run

```
python setup.py install
```

Tested only on python 3, please email `itamarreis@mail.tau.ac.il` if you encounter any issues with installation or running the code.

example usage

```
from PRF import prf
prf_cls = prf(n_estimators=10, bootstrap=True, keep_proba=0.05)
prf_cls.fit(X=X_train, dX=dX_train, y=y_train)
pred = prf_cls.predict(X=X_test, dX=dX_test)
```

Activities and effort

- Data Exploration
- TS Data Preprocessing (minimal)
- Forecast architecture and prediction confidence interval
- XAI on predictions
- Covariate shift
- Sampling strategies
- GitHub library

any question?

~~keep it for yourself~~

please ask!

riccardo.crupi@intesasanpaolo.com