EUROPEAN CENTRAL BANK

EUROSYSTEM

Machine Learning and Data Science at ECB

Presented by

Dr. Redouane Boumghar

Eurofiling Conference 2019, Frankfurt

ECB and its statistics

Macro-level statistics

- Balance sheet statistics
- Monetary aggregates (M1 – M3)
- Securities issues
- Banks interest rates
- Government finance
- Euro area financial accounts

Micro-level statistics

- Security-by-security statistics
- Holdings of individual securities
- Money market statistics reporting
- Loans by loans register (Ana Credit)
- Register of Financial Institutions
- Individual bank supervisory data

Statistics and Analytics

Fostering Innovation across the data value chain

Collection → Production → Dissemination

For example innovation can help to …

1. Reduce reporting burden; automated reporting and corrections
2. Produce more data and statistics faster at fit for purpose quality
3. Enable elaborated use of data for evidence based decision making

Machine Learning use cases within DG-S

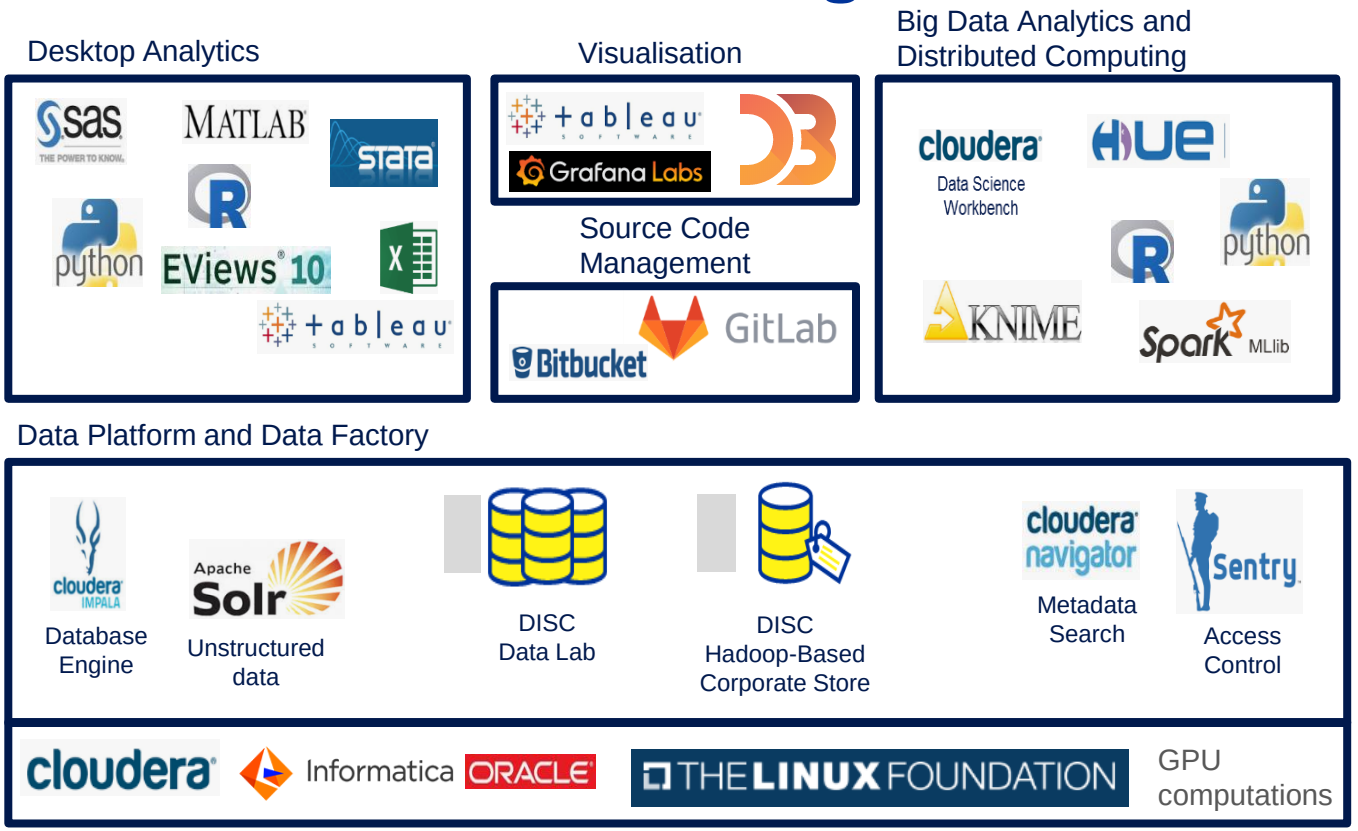
- **Anomaly/outlier detection** algorithms, data validation, plausibility checks especially for large datasets where standard statistical techniques could not be used (AnaCredit)
- **Data classification**: assessing, matching or pairing duplicate records (sometimes containing errors)(EMIR, MMSR)
- Evaluating **data credibility and data quality** (expert systems for validating data)
- **Forecasting, backcasting, interpolating**, estimating missing data using ML algorithms (Balancing the Financial Accounts Accounting matrix)
- **Record linkage** to link records that represent the same entity in different databases, calibrating missing data by data integration (linking RIAD with MMSR)

Data Science Infrastructure and integration

- → Variety of data
- → Volume of data
- → Velocity
- → Know how
- → Search for data

DevOps

CI/CD



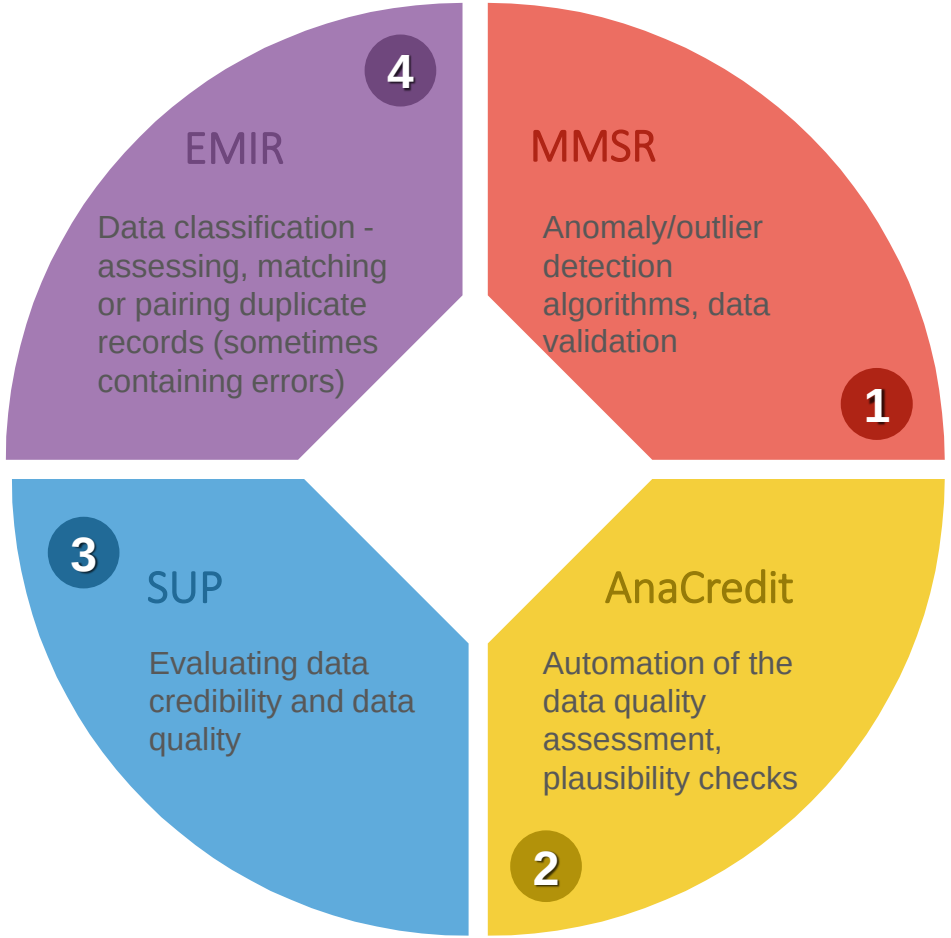
ML Business Cases at ECB

1. **MMSR data** - Daily information on all money market transactions conducted in euro conducted by the 52 largest banks in the euro area. 45,000 transactions per day with 30 to 60 attributes per transaction, collected since April 2016

2. **AnaCredit data** contains detailed information on individual bank loans in the euro area, harmonised across all Member States. The AnaCredit dataset includes detailed information on about 60 million individual bank loans in the euro area granted to legal entities.

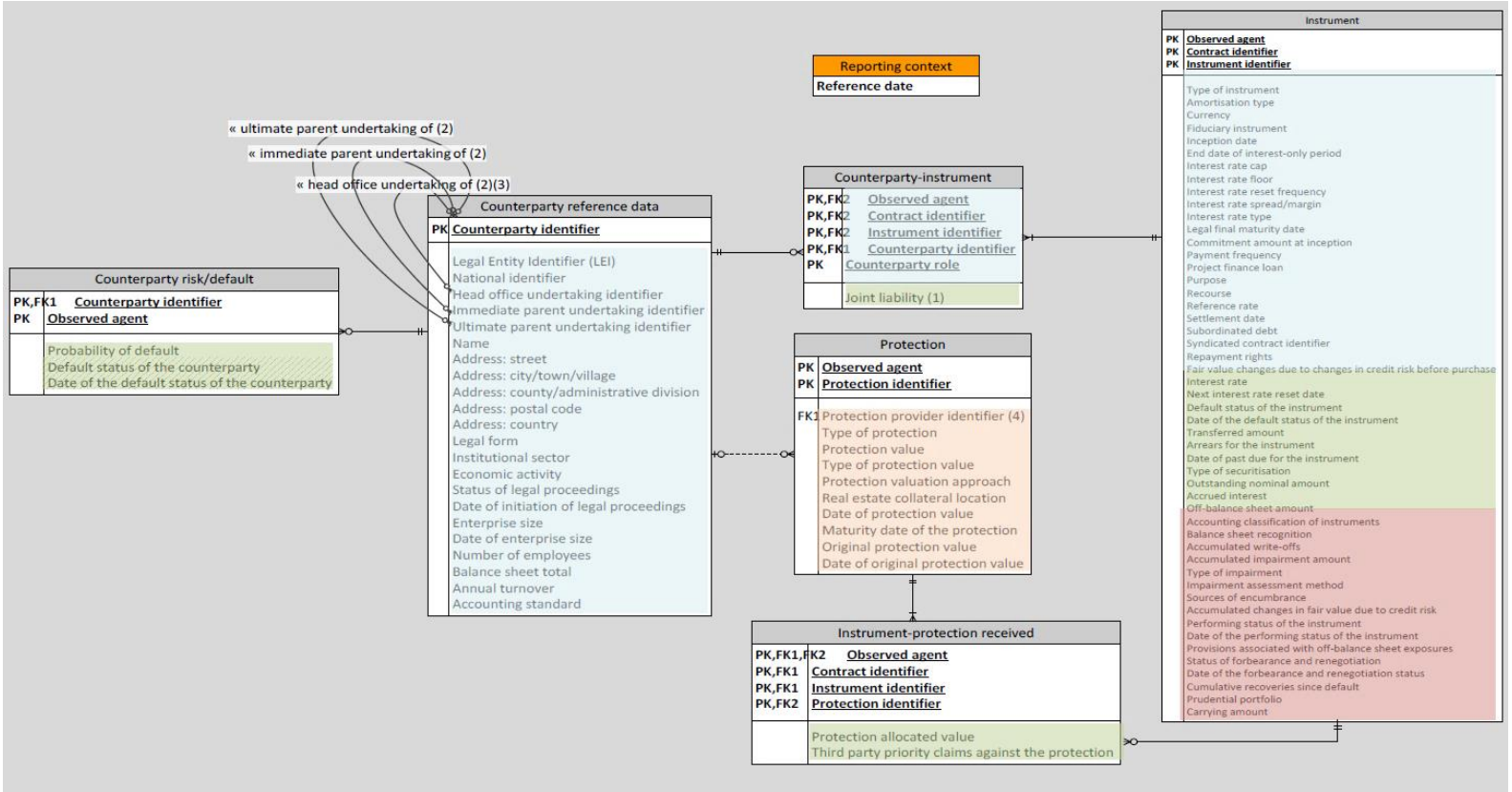
3. **SUP Data -** The data comes from two frameworks -The ITS (Implementing Technical Standards) and the STE (Short Term Exercises), which are a number of detailed reports, describing the financial, risk, liquidity and leverage position of the banks for a total of approx. 5660 banks.

4. **EMIR data** - Daily information on all derivatives trades outstanding and all transactions in the EU, with double-sided reporting; between 20 to 100 million transactions *per day* collected since 2014, with 80 to 120 data attributes per transaction.



AnaCredit - Machine Learning Use Case

AnaCredit Dataset



Data pre-processing and feature engineering

incorporation of domain knowledge

- Amounts
- Rates
- Dates
- Categorical attributes

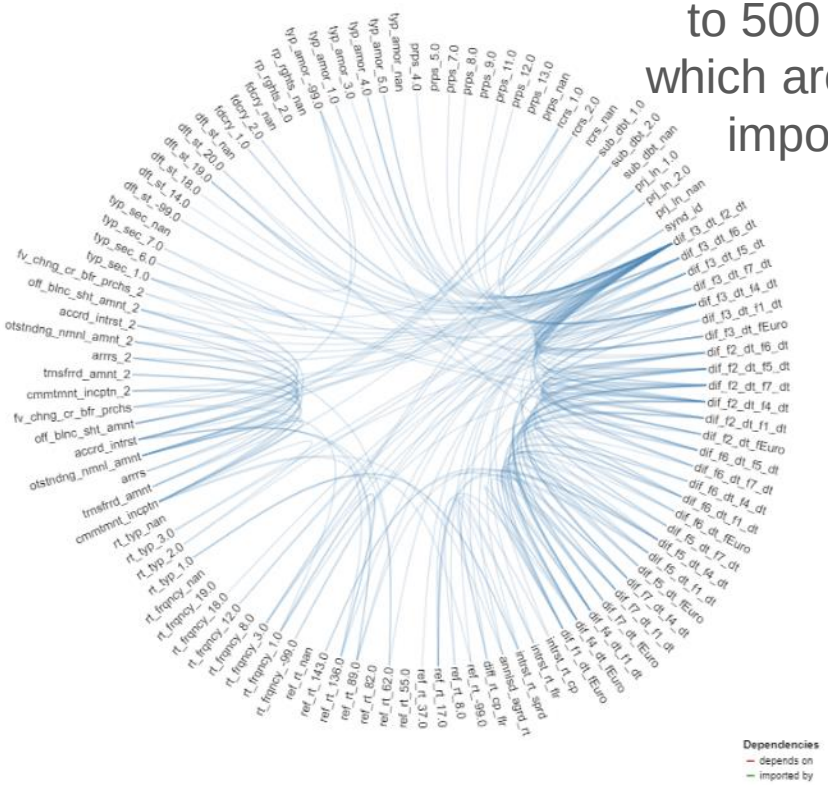
- **Differences**
- **Squares**
- **Root Squares**
- **Exponential/Log**
- **Seasonal Info**
- **…**

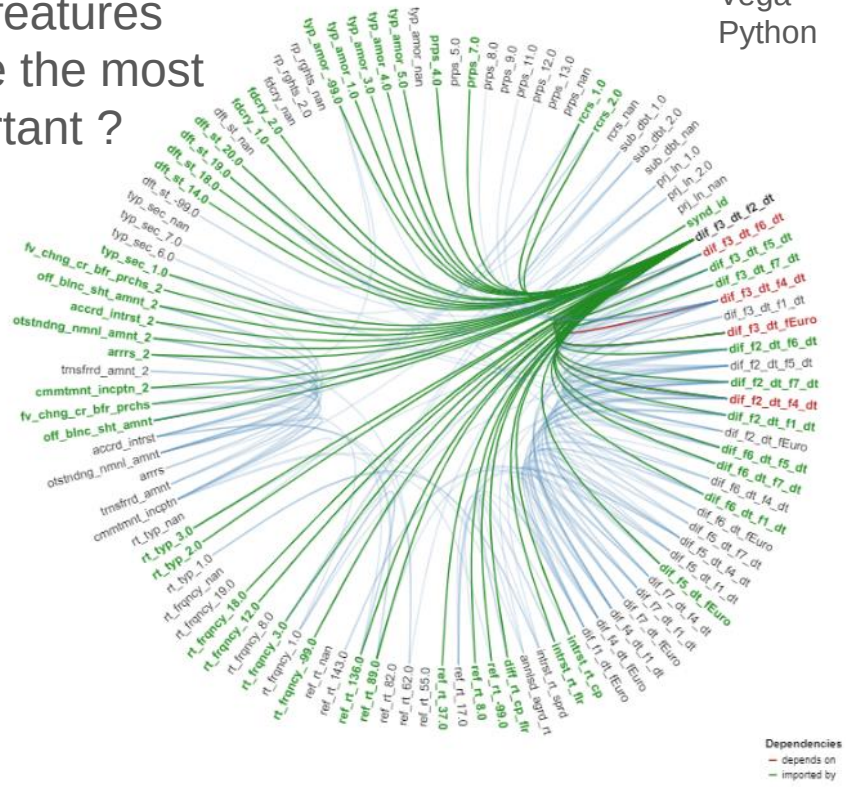
- Periods
- Non linear expressions
- Day of the month
- Day of the year
- Relation to calendar events

Feature Selection

from 80 attributes
to 500 features
which are the most
important ?

Tools used
XGBoost
Vega
Python

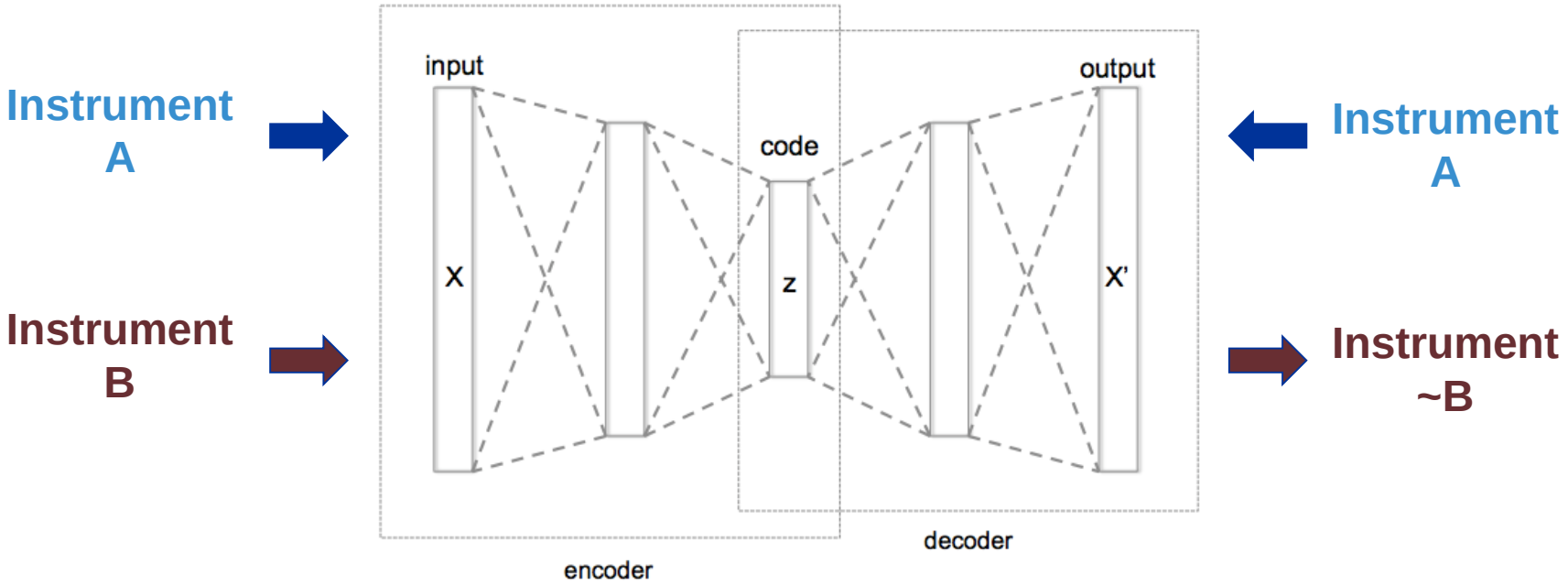




Machine Learning algorithm: Isolation Forest

Machine Learning algorithm: Isolation Forest

Machine Learning algorithm: Auto-Encoders



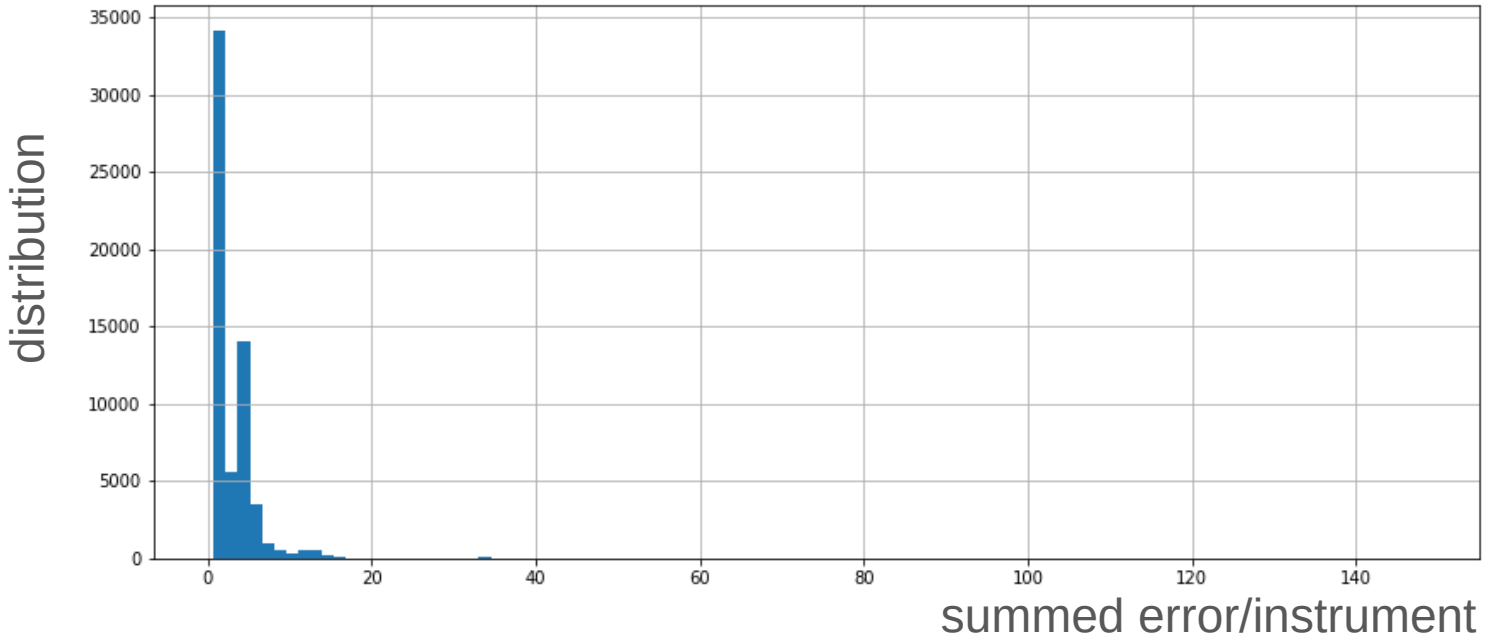
Auto-encoder neural network architecture

Image credits: Chervinskii CC BY-SA 4.0

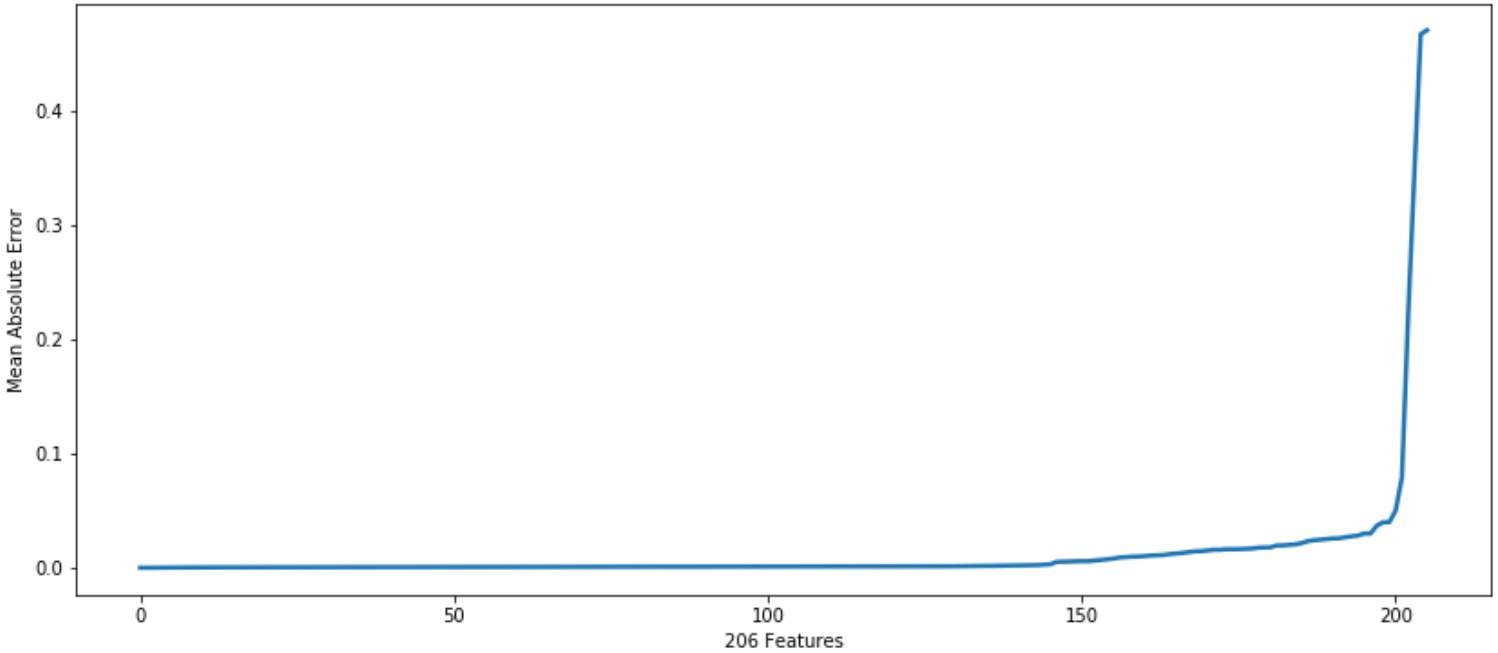
Machine Learning algorithm: Auto-Encoders

- loss of 0.7 with scaled dummies: DO NOT SCALE dummies ☺
- loss of 0.027 with non scaled dummies

Tools used
Keras
Tensorflow
Python



Machine Learning algorithm: Auto-Encoders



Machine Learning use case: next steps

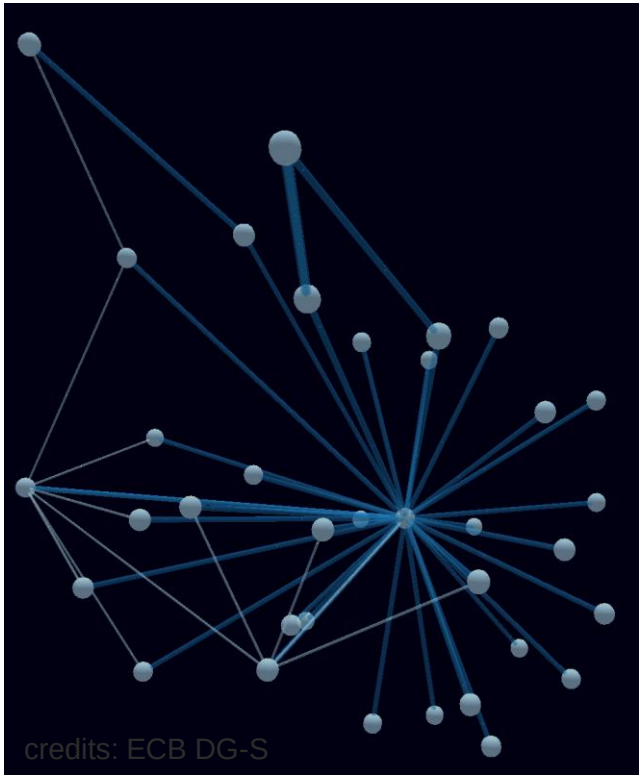
- Lots of shared tools
- Interactions between ML output and end users
- Interactive process, continual learning

Explainability

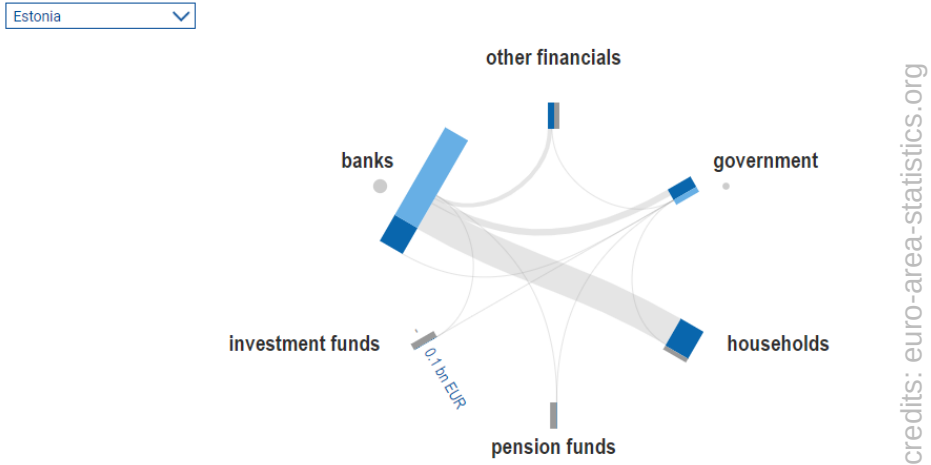
Data Visualization is Key

Awareness
and
Understanding

Dynamic and interactive data visualization



Results of a cyclic dependence analysis between timeseries (XGBoost) viewed in a 3D dynamic graph



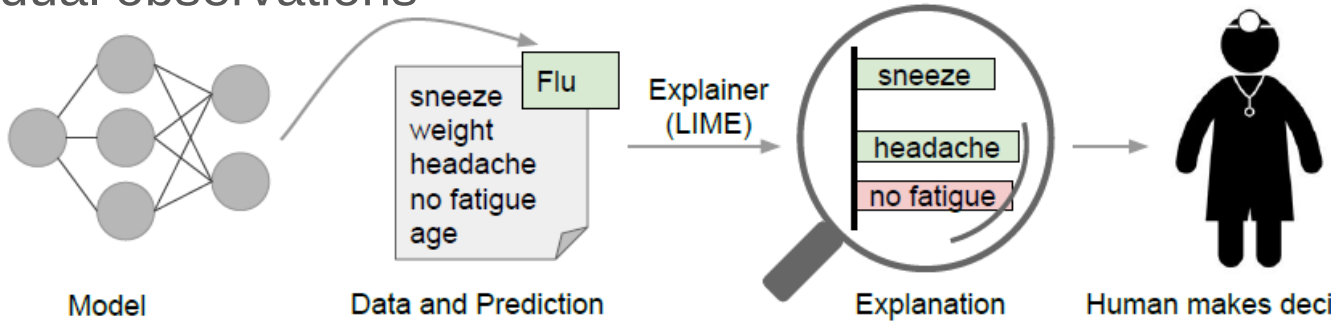
Financing and Investment Dynamics

Visualization is key for understanding and exploration

Explaining ML output to central bankers

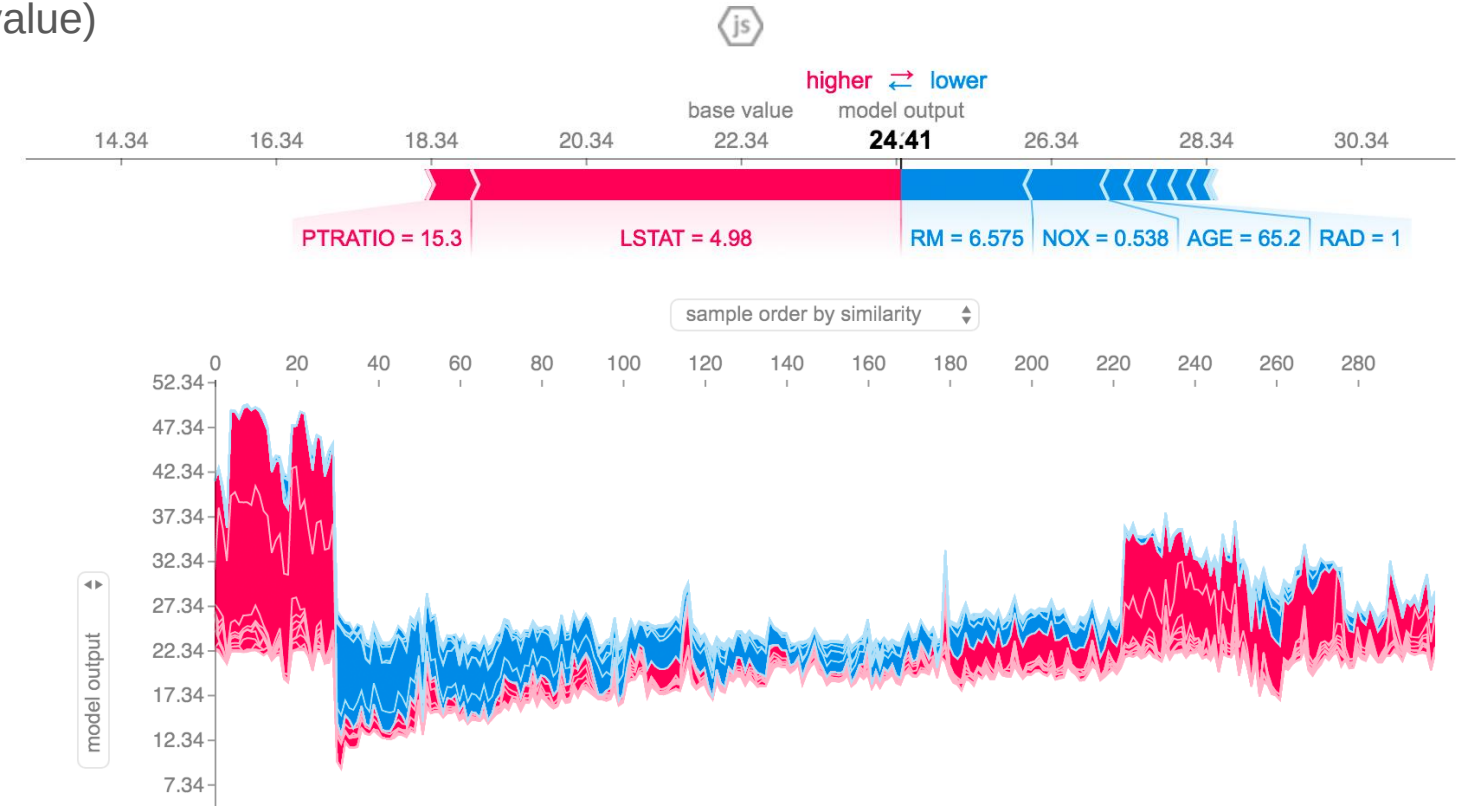
If the users do not trust a model or a prediction, they will not use it.

- **LIME** - Local Interpretable Model-Agnostic Explanations
- an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- The LIME algorithm opens the way to interpret the results of complex statistical models.
- LIME augments the black-box model results with interpretability of individual observations



Other tools for explainability

- **SHAP** (Shapley value)
- DeepLIFT
- TreeInterpreter
- ELI5
- DeepVis
- …



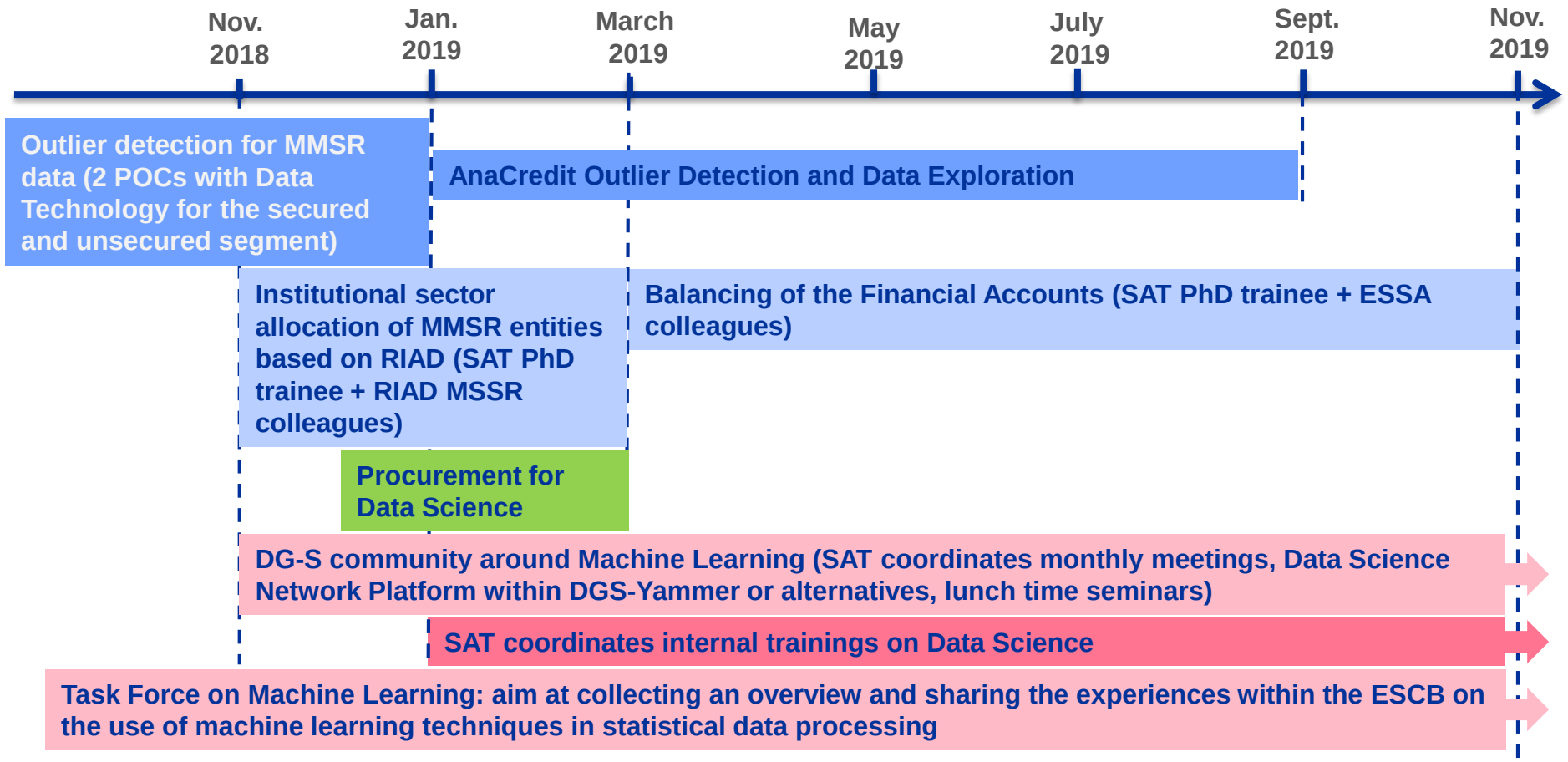
Machine Learning Community

Fluid knowledge exchange to foster innovation

- Information
- Tips and Tricks
- Trainings/Tutorials
- Impactful Initiatives
- Projects and discussions

Progress / Catalyst / Inspirational

Conclusion


Nov. 2018
Jan. 2019
March 2019
May 2019
July 2019
Sept. 2019
Nov. 2019
Outlier detection for MMSR data (2 POCs with Data Technology for the secured and unsecured segment)
AnaCredit Outlier Detection and Data Exploration
Institutional sector allocation of MMSR entities based on RIAD (SAT PhD trainee + RIAD MSSR colleagues)
Balancing of the Financial Accounts (SAT PhD trainee + ESSA colleagues)
Procurement for Data Science
DG-S community around Machine Learning (SAT coordinates monthly meetings, Data Science Network Platform within DGS-Yammer or alternatives, lunch time seminars)
SAT coordinates internal trainings on Data Science
Task Force on Machine Learning: aim at collecting an overview and sharing the experiences within the ESCB on the use of machine learning techniques in statistical data processing

Be aware of technologies and communities

https://matrix.org/

https://about.gitlab.com

https://unit.nginx.org/

Nothing is so embarrassing as watching someone do something that you said couldn't be done.

Sam Ewing