BANQUE DE FRANCE
EUROSYSTÈME

French-German Summer-school

Self organizing maps for anomaly detection in an operational context

Matthias LAPORTE – June 24th 2020

AGENDA

Introduction

Control and risk management at Banque de France

Control and risk management mobilize **significant human resources** with a **high level of expertise**

Why ?

- Compliance with regulation at Banque de France and in supervised institutions (banks and insurance companies)
- Financial security, fraud, cybersecurity, operational risk …

What ?

- Transactions and other financial operations
- User behavior on an IT infrastructure

How ?

- Monitoring activity, analyzing alerts, sampling operations
- Expert systems “old school AI”

How can AI help ?

Machine learning techniques and anomaly detection algorithms can help business experts to focus on higher value-added tasks and gain a more complete view of risks

- **Scalability**
  - Scalibility to the input : tackle **big datasets** and leverage **complex information**
  - Scalibility of the output : **control the workload** induced on the operational process
- **Interpretability**
  - Granular level : accelerate decision making with useful and **interpretable insights**
  - Global level : provide a better appreciation of **risk drivers** and **risk profiles**
- **Adaptability**
  - Integration in an already existing process : make clever use of **prior knowledge**
  - Anticipation : keep up to date with **risk evolution**

Development frameworks

Development of **"in-house" Machine Learning platforms**, based on Python and open source libraries, to **capitalize** on projects and **accelerate** experiments

- **MARIA** (Reference Methods and Algorithms for Artificial Intelligence)
  - Based on Python's popular libraries ***numpy***, ***pandas*** and ***sklearn***
  - Enriched with custom preprocessing functions for **tailor-made features' engineering**
  - Enables **advanced learning strategies** implementation for artificial neural networks
  - Industrializes model prototyping with a **pipeline approach**
  - Provides a model **explainability module**
- **LUCIA** (Software for Use in Control with Artificial Intelligence)
  - Associates **unsupervised learning** techniques with **advanced data visualisation** tools
- **GAIA** (Graph Analysis by Artificial Intelligence)
  - Integrates **graph analysis** techniques into **machine learning** models

AGENDA

General principles

Systematically **characterize behaviors** and **detect suspicious ones** thanks to self organizing maps and proper metrics

- The three subjects are use cases for **anomaly detection**
  - The aim is to **identify atypical behaviors** within a population
  - Capitalization thanks to **LUCIA**, **GAIA** and **MARIA**
- **Generic** approach
  - Identification of the **scope of investigation**
  - Determination of the **level of granularity** of the analysis
  - Construction of **relevant indicators**

  } **Data consolidation and feature engineering**

  - Modeling by **unsupervised machine learning**
  - Analyzing the results with adapted **data visualization** and **risk metrics**

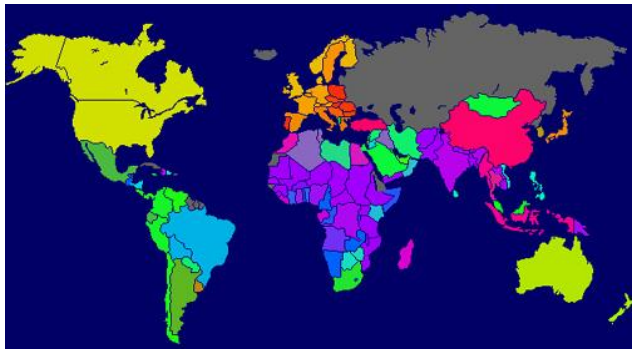
Modelling

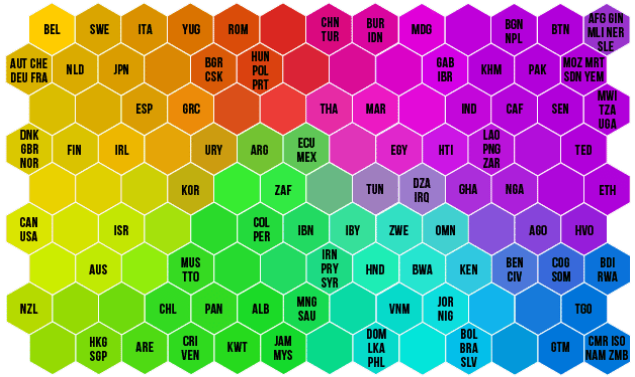
A **self-organizing map** (SOM) is a type of artificial neural network which aims at producing a low-dimensional, discretized representation of the input data

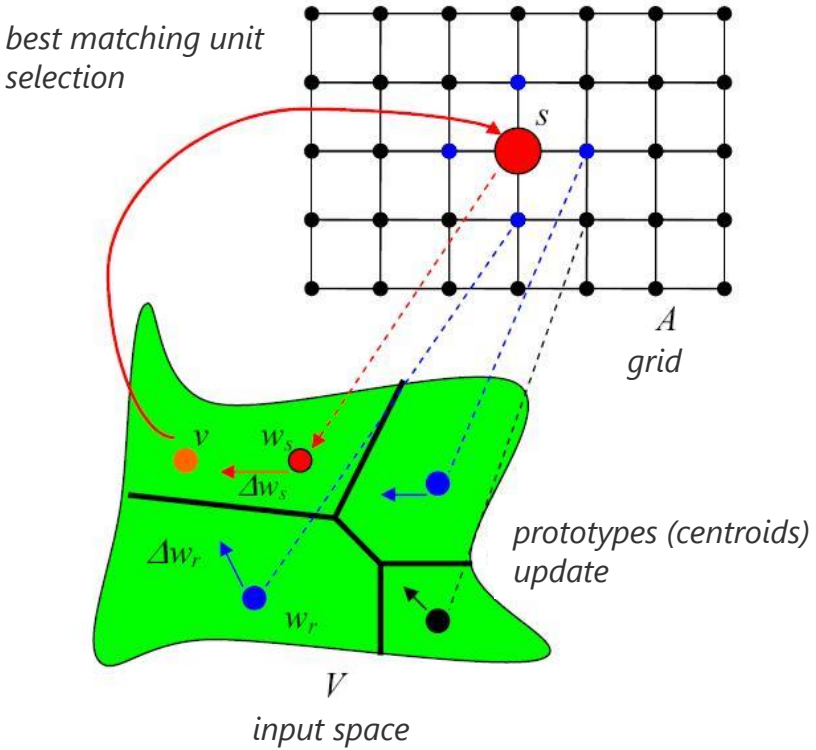
- **Reducing the dimensionality** of the dataset
  - Projects a dataset with many variables on a two-dimensional map
  - Makes the dataset intelligible by automatically establishing an underlying typology
- **Discretization** of the analysis space
  - Projects many observations on a single analysis grid with fewer typical cases (centroids)
  - Allows mass processing of a set of similar cases
- **Preservation of the topological properties** of the input space
  - Two close points in the input space will be close on the map as well
  - Gaps in the map highlight discontinuities in the underlying distribution

Methodological approach

Modelling







Visualization and metrics

Deliver the full power of self organizing maps to business experts without requiring prior data science knowledge

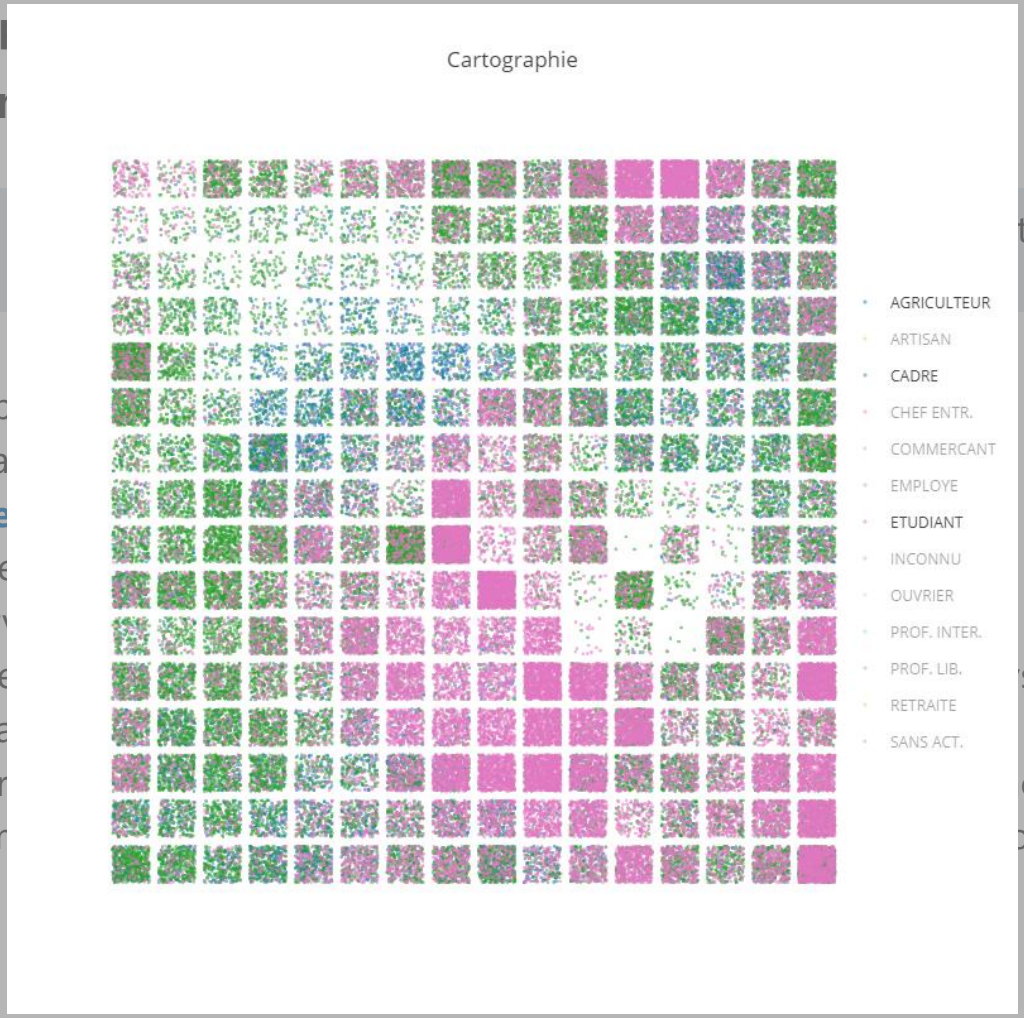
- Enriching the map with **qualitative information**
  - **Project** a categorical variable and analyze its distribution
  - **Extrapolate** from known anomalies (semi-supervised learning)
- Deep diving in the map
  - Get an overview of the **typology's dependency to a specific risk factor**
  - Visualize the **variance inside the cluster** with Principal Component Analysis
- Measure abnormality
  - Score abnormal observations with the **reconstruction error** (intrinsic abnormality)
  - Tackle regime changes with **dynamics analysis and volatility** (deviation over time)

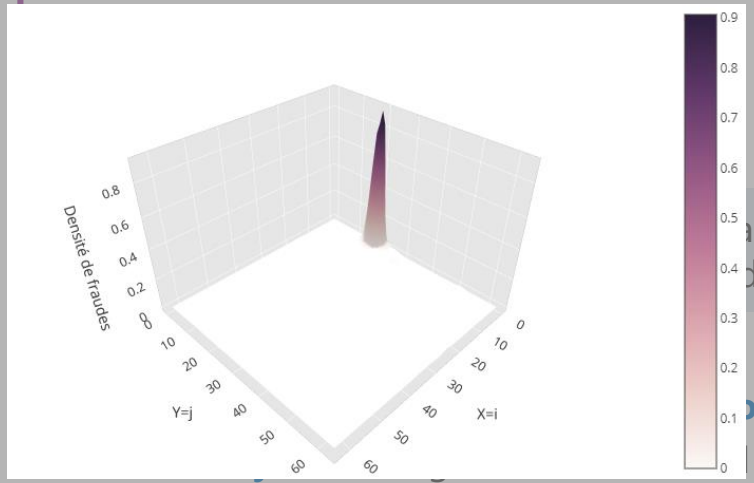
Methodological ap

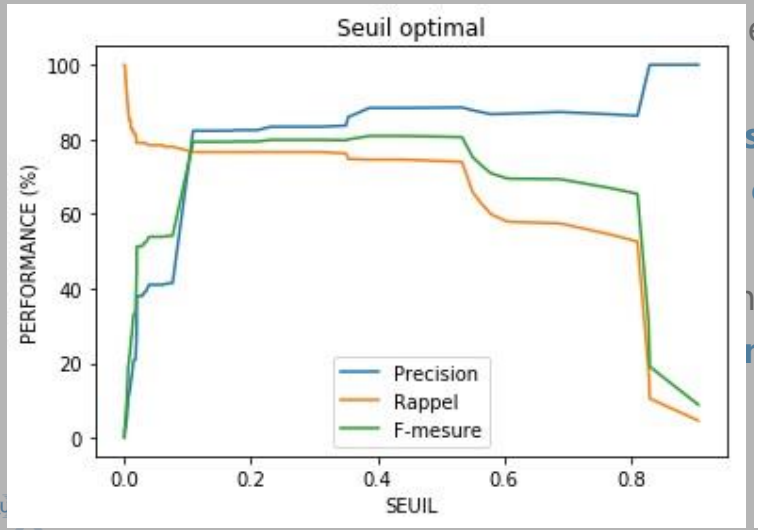
Visualization ar

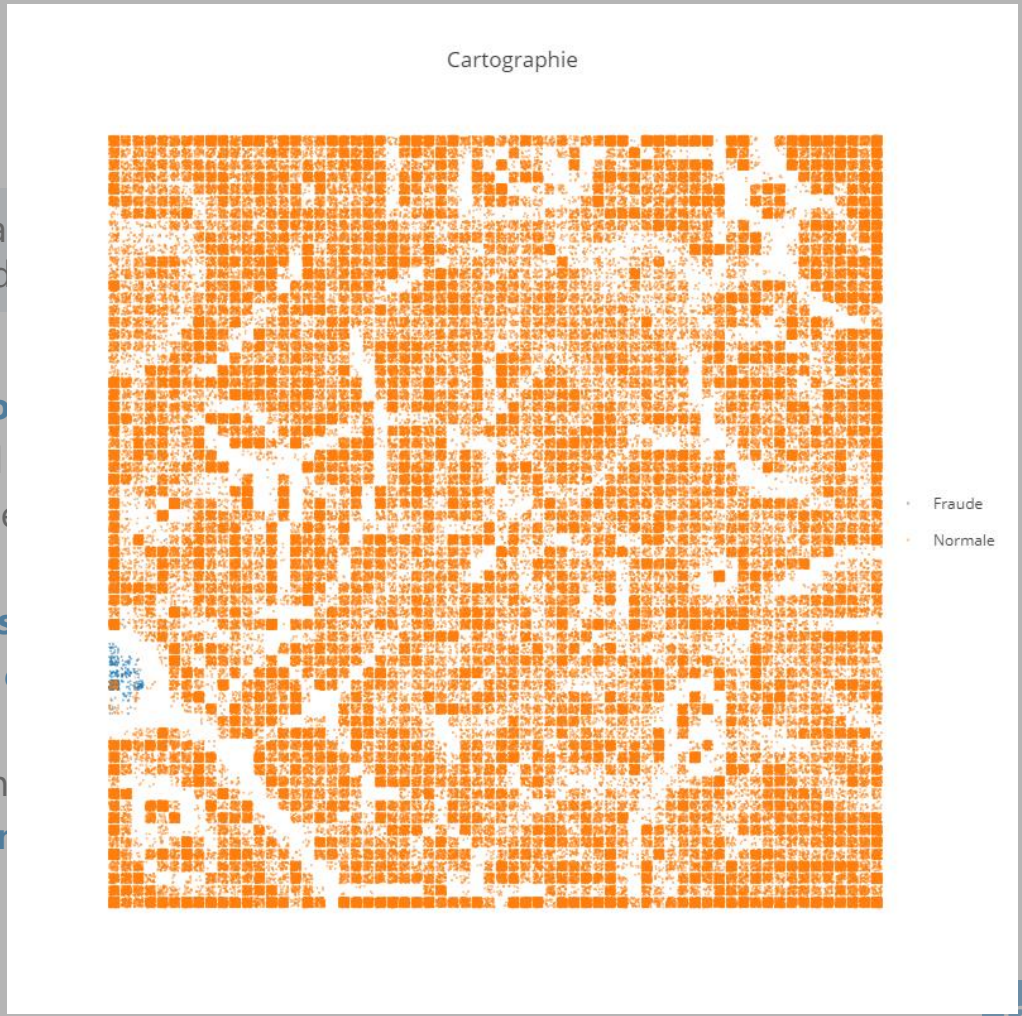
Deliver the t requiring

- Enriching the map
  - **Project** a ca
  - **Extrapolate**
- Deep diving in the
  - Get an overv
  - Visualize the sis
- Measure abnorma
  - Score abnor ormality)
  - Tackle regim over time)




Densité de fraudes
Y=j
X=i


Seuil optimal
PERFORMANCE (%)
SEUIL
Precision
Rappel
F-mesure


Cartographie
Fraude
Normale

Visualization and metrics

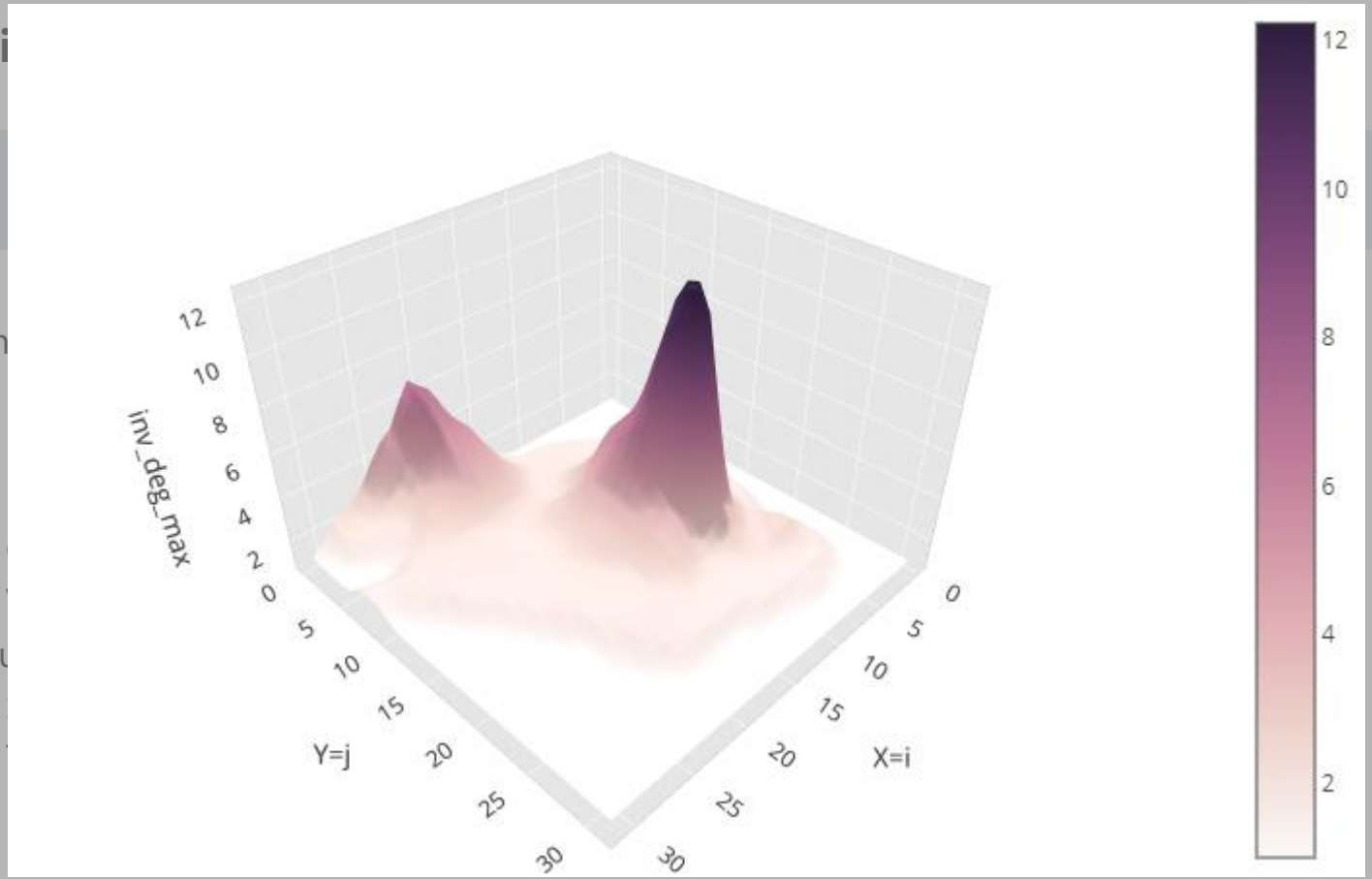
Deliver the full power of self organizing maps to business experts without requiring prior data science knowledge

- Enriching the map with **qualitative information**
  - **Project** a categorical variable and analyze its distribution
  - **Extrapolate** from known anomalies (semi-supervised learning)
- Deep diving in the map
  - Get an overview of the **typology's dependency to a specific risk factor**
  - Visualize the **variance inside the cluster** with Principal Component Analysis
- Measure abnormality
  - Score abnormal observations with the **reconstruction error** (intrinsic abnormality)
  - Tackle regime changes with **dynamics analysis and volatility** (deviation over time)

Methodological approach

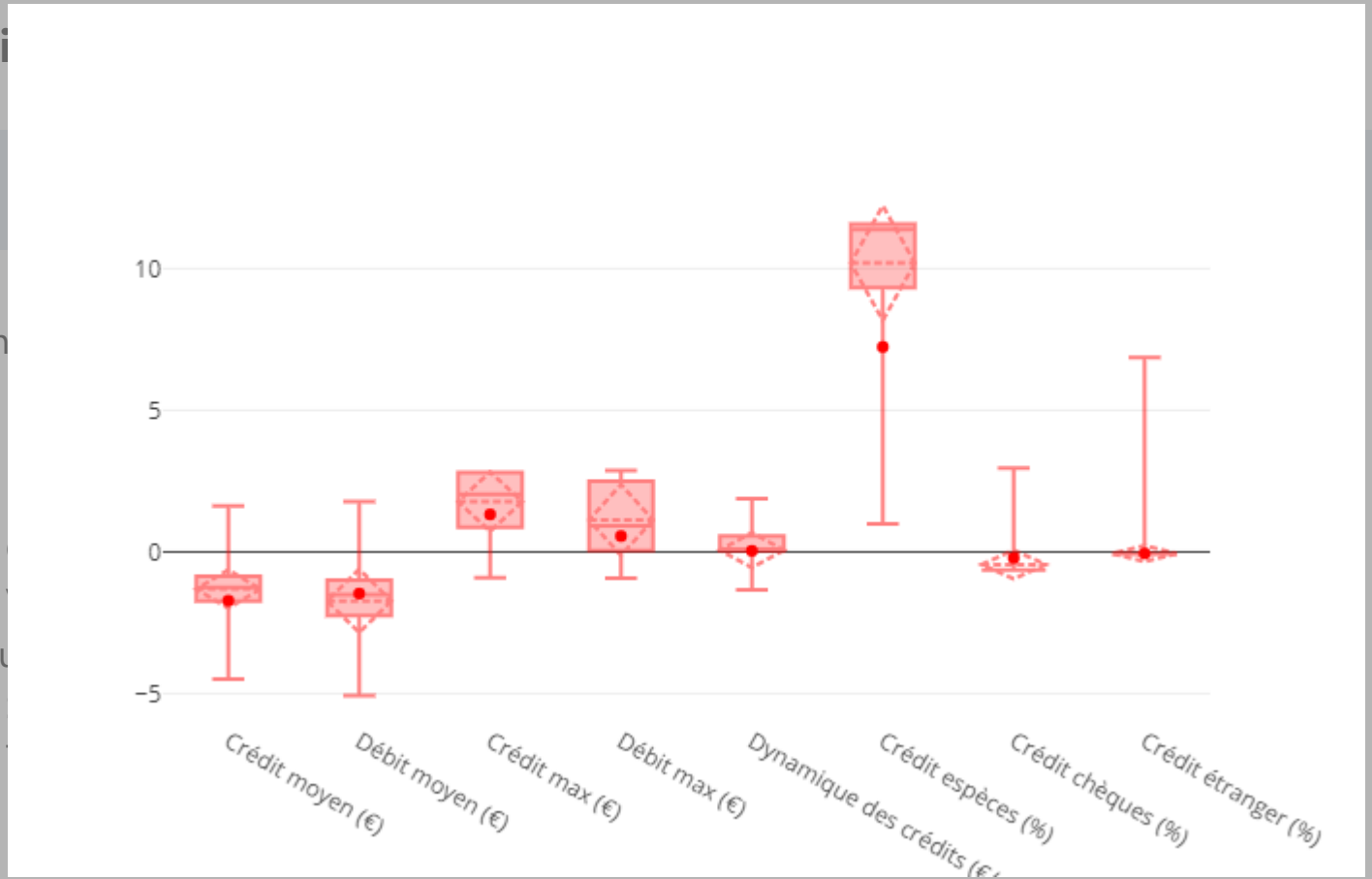
Visuali

- Enrich
  -
  -
- Deep
  -
  -
- Measu
  -
  -



Visuali

- Enrich
  -
  -
- Deep
  -
  -
- Measu
  -
  -



Me
Vis

- E
- D
- M


Segment 11|0
10
5
0
−5
−10
−5
0
5
10
15
0.01
0.005
0
AGRICULTEUR
ARTISAN
CADRE
CHEF ENTR.
COMMERCANT
EMPLOYE
ETUDIANT
OUVRIER
PROF. INTER.
PROF. LIB.
RETRAITE
SANS ACT.
Centroïde

Visualization and metrics

Deliver the full power of self organizing maps to business experts without requiring prior data science knowledge

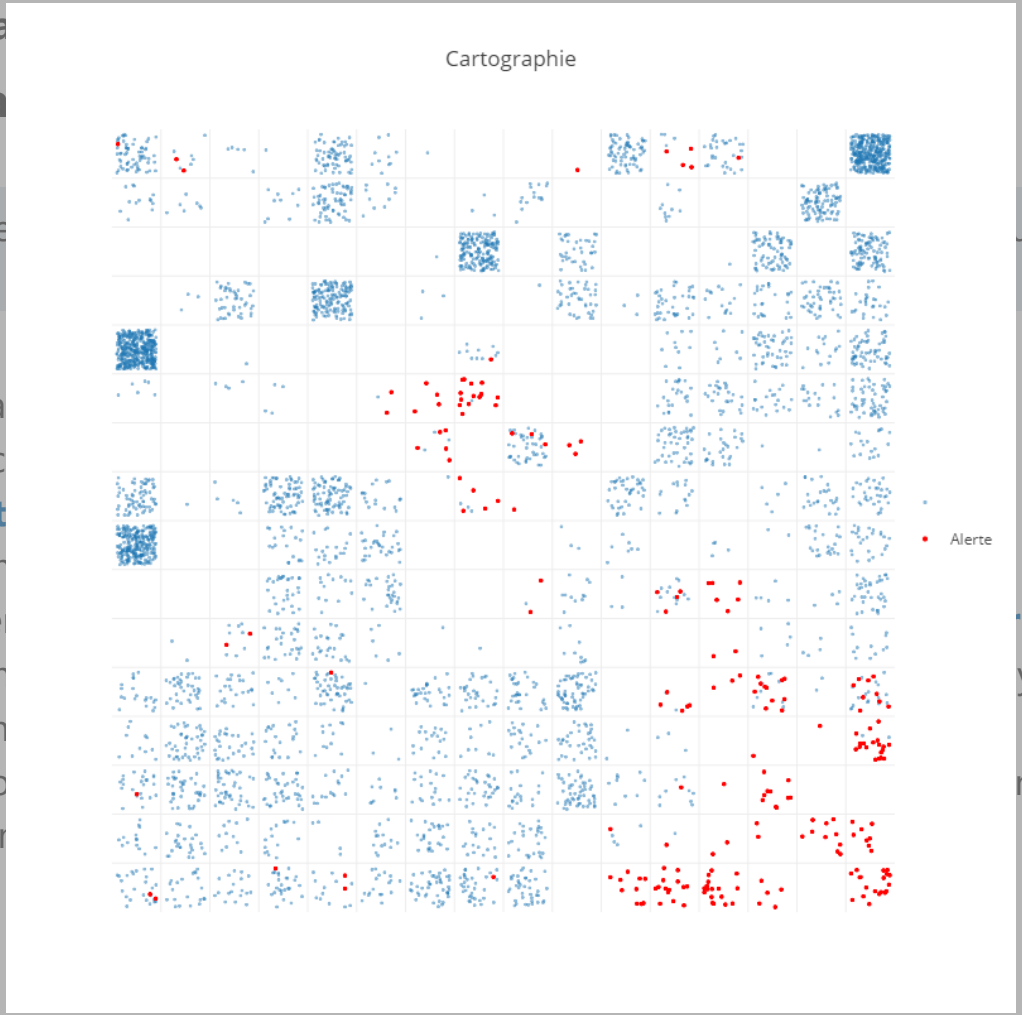
- Enriching the map with **qualitative information**
  - **Project** a categorical variable and analyze its distribution
  - **Extrapolate** from known anomalies (semi-supervised learning)
- Deep diving in the map
  - Get an overview of the **typology's dependency to a specific risk factor**
  - Visualize the **variance inside the cluster** with Principal Component Analysis
- Measure abnormality
  - Score abnormal observations with the **reconstruction error** (intrinsic abnormality)
  - Tackle regime changes with **dynamics analysis and volatility** (deviation over time)

Methodological a

Visualization a

Deliver the ut requiring

- Enriching the ma
  - **Project** a c
  - **Extrapolat**
- Deep diving in th
  - Get an ove
  - Visualize th ysis
- Measure abnorm
  - Score abno normality)
  - Tackle regi over time)

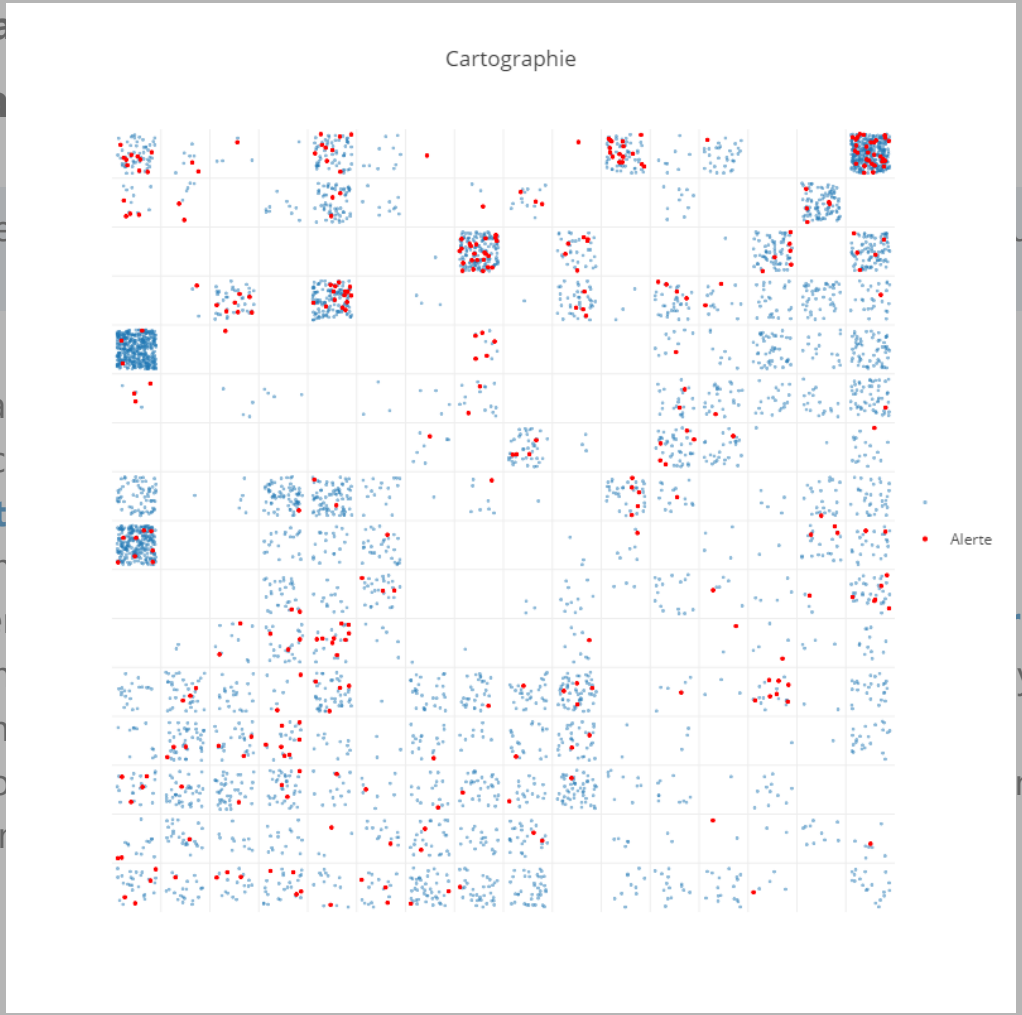


Methodological a

Visualization a

Deliver the ut requiring

- Enriching the ma
  - **Project** a c
  - **Extrapolat**
- Deep diving in th
  - Get an ove
  - Visualize th ysis
- Measure abnorm
  - Score abno normality)
  - Tackle regi over time)



AGENDA

Use cases : implementation and results

On-site supervision and Anti Money Laundering (AML)

Each year, several AML dedicated inspection missions are performed by the ACPR in supervised institutions

Why ?

- To ensure of the compliance of the anti-money laundering system implemented in the institutions

What ?

- A risk based system leading to an adequate case processing (reinforced review and suspicious transaction report to TRACFIN)

How ?

- Challenging the system both formally and empirically by analysing cases that were not properly processed by the institution

On-site supervision and Anti Money Laundering (AML)

The **evaluation of the anti-money laundering measures** as they are implemented in the supervised institutions requires the **examination of individual files**

- Limitations of the traditional approach
  - Extensive scopes (up to **several millions customers/accounts**) but rough sampling methods
  - Tedious manipulation of data to compute aggregates and basic descriptive statistics
  - Simple criteria filtering based on expert knowledge and **a priori scenarios**
- Improvements brought by a comprehensive and advanced anomaly detection tool
  - Intelligent filtering based on **hypothetical profiles**
  - Cross-referencing with qualitative data to **visualize inconsistencies**
  - Exploration of **atypical patterns** and identification of **risk factors**

On-site supervision and Anti Money Laundering (AML)

The prototype is currently being tested **in situ** by on-site inspection missions related to AML led by the ACPR

- Adapt the preprocessing methodology to different contexts
  - **Design new features and indicators** for other types of missions (legal entities, payment institutions, correspondence banking …)
- Build a **user-friendly interface**
  - Improve the accessibility of the tools for non programmers
- Extrapolate from high value added data with **semi-supervised learning**
  - Gather feedback data from TRACFIN
- Benchmark supervised institutions
  - Embed collective knowledge in a generic map

Use cases : implementation and results

Fraud detection on French Treasury payments

The Banque de France keeps the accounts and carries out the operations of the French Treasury

Why ?

- To protect its customer from fraud and provide a high value added service by enforcing state of the arts techniques

What ?

- A real-time alerting system helping analysts to prevent a fraudulent operation from being successfully carried out within 24 hours

How ?

- Detecting anomalies in the behaviours of receiving accounts (potential fraudulent accounts)

Fraud detection on French Treasury payments

An ongoing experimentation about fraud detection on **STEP2 transactions** has led to the launch of a first alert engine based on "expert metrics"

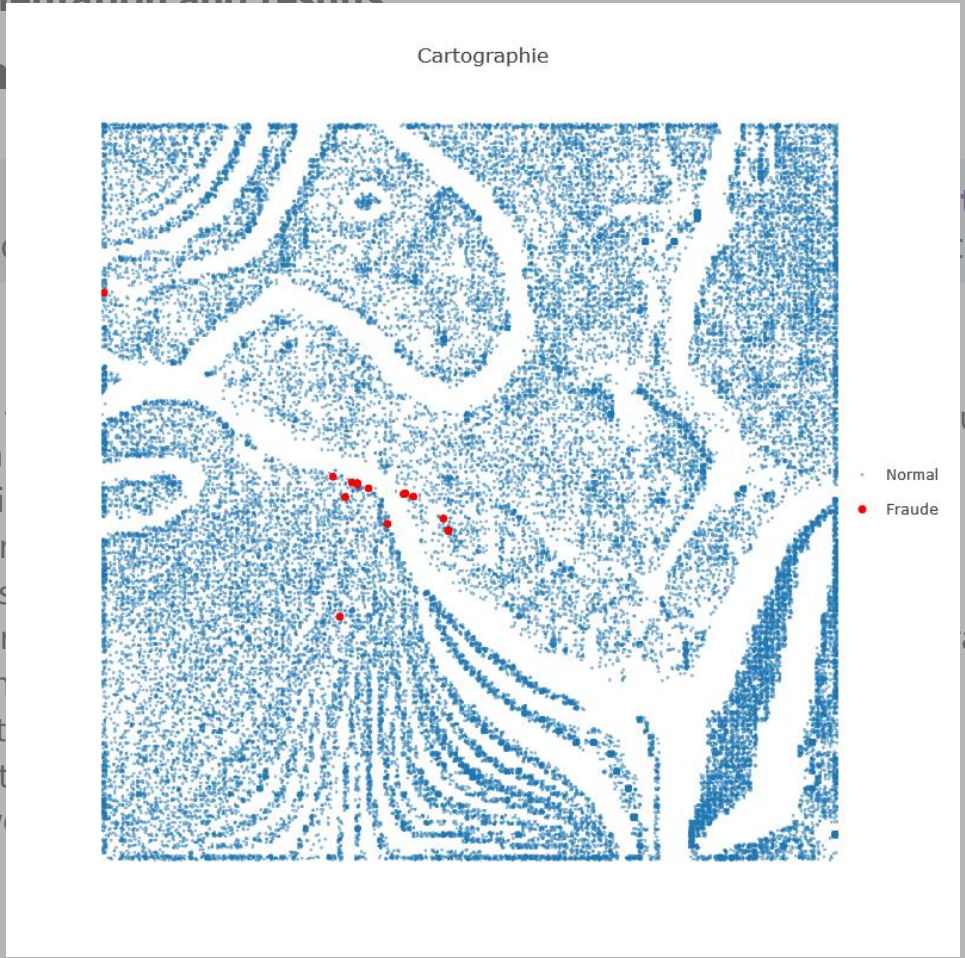
- The score based on expert metrics **proved to be effective**
  - Over the first week of implementation, 5 proven frauds identified for an amount of more than EUR 1.2 million on French Treasury operations
- **Implement LUCIA** in addition to the "expert metrics"
  - Detect non-predefined risk patterns
  - Protect against "innovation" in fraud strategies
- Adaptation to the problem of **transactional data** by exploiting the formalism of graphs
  - Use the information contained in the graph structure
  - Compute metrics efficiently in order to process transactions in real-time
- Operational integration
  - Control the workload by adjusting the alert level

Use cases : implementation and results

Fraud detection

An ongoing … **tions** has led … t… …s"

- The score based on …
  - Over the first … 1.2 million on … …unt of more than EUR
- **Implement LUCIA** i…
  - Detect non-pr…
  - Protect agains…
- Adaptation to the pr… …aphs
  - Use the inform…
  - Compute met…
- Operational integrat…
  - Control the w…

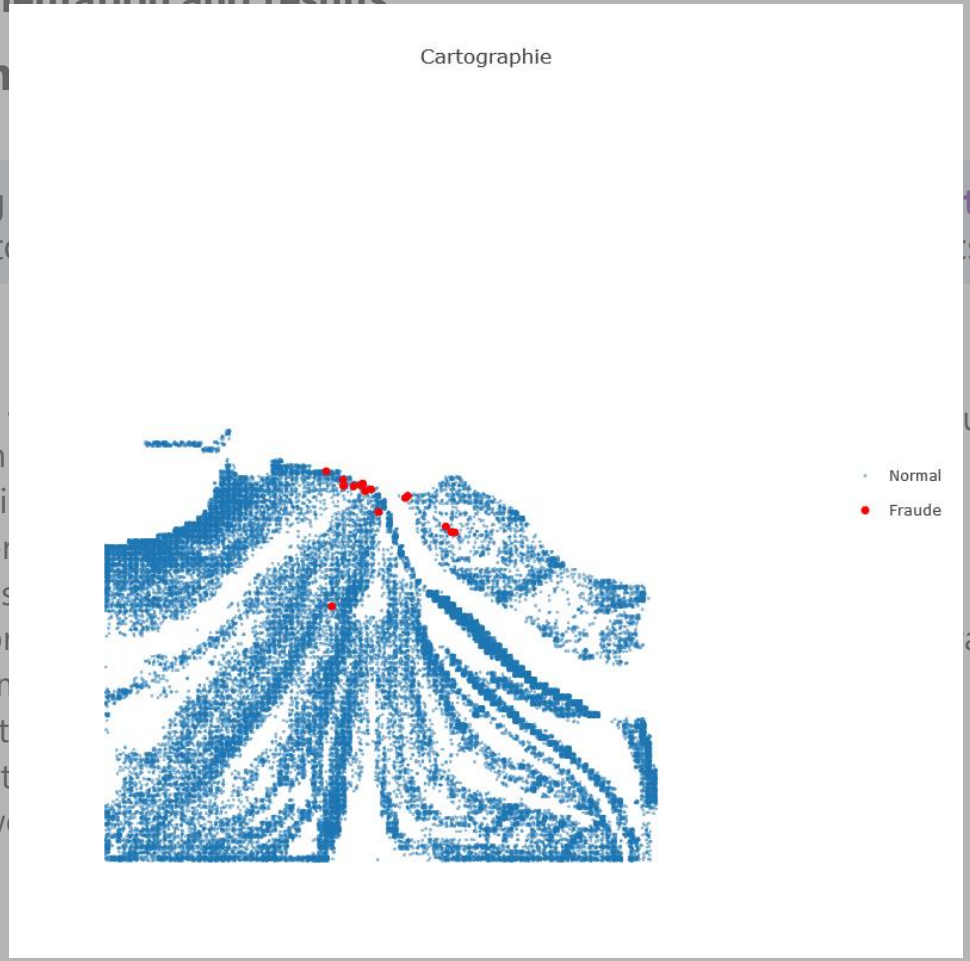


Use cases : implementation and results

Fraud detection

An ongoing … **tions** has led … t… …s"

- The score based on …
  - Over the first … …unt of more than EUR 1.2 million on …
- **Implement LUCIA** i…
  - Detect non-pr…
  - Protect agains…
- Adaptation to the pr… …aphs
  - Use the inform…
  - Compute met…
- Operational integrat…
  - Control the w…



Use cases : implementation and results

Cybersecurity and protection against insider threat

The CERT (computer emergency response team) is responsible for monitoring the security of Banque de France information system

Why ?

- As a central bank and a vital operator for the nation, Banque de France is exposed to cyber risk from both outsider and insider threat

What ?

- A real-time alerting system helping analysts to prevent cyber-attacks and deliver an effective and diligent response

How ?

- Detecting anomalies in the behaviours of users by exploiting system and applicative logs

Cybersecurity and protection against insider threat

A first experimentation about **user behavior analysis** on the IT infrastructure of Banque de France has been led on the scope of **domain administration**

- Use every single **event** logged in Splunk that can relate directly to a user
  - Users can be either the **subject** or the **target** of an event
  - Over four months of logs, almost **2 million events** are linked to a user (subject and/or target)
- **Aggregate** events by user and time period
  - **Behavior descriptors** are defined as an **event distribution** over a given period (a week) and for a given user, taking into account its status as a subject or a target
  - Distinguish daytime and nighttime, as well as week days and week ends
- Analysis of **intrinsic and temporal anomalies** as well as **profile coherence**
  - Using both anomaly scoring metrics, and visual analysis with qualitative data

Use cases : imple

Cybersecurity

A first ex
Banq

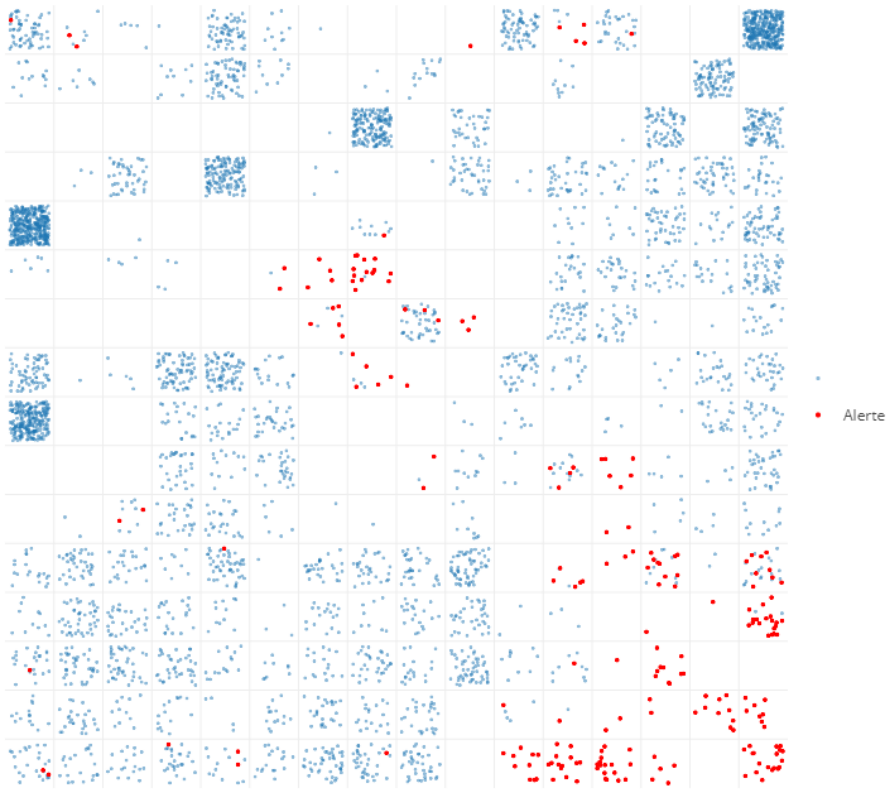
ructure of
ration

- Use every single
  - Users can b
  - Over four
- **Aggregate** even
  - **Behavior**
    for a given
  - Distinguish
- Analysis of **intri**
  - Using both

ubject and/or target)

period (a week) and

data



Use cases : imple

Cybersecurity

A first ex
Banq

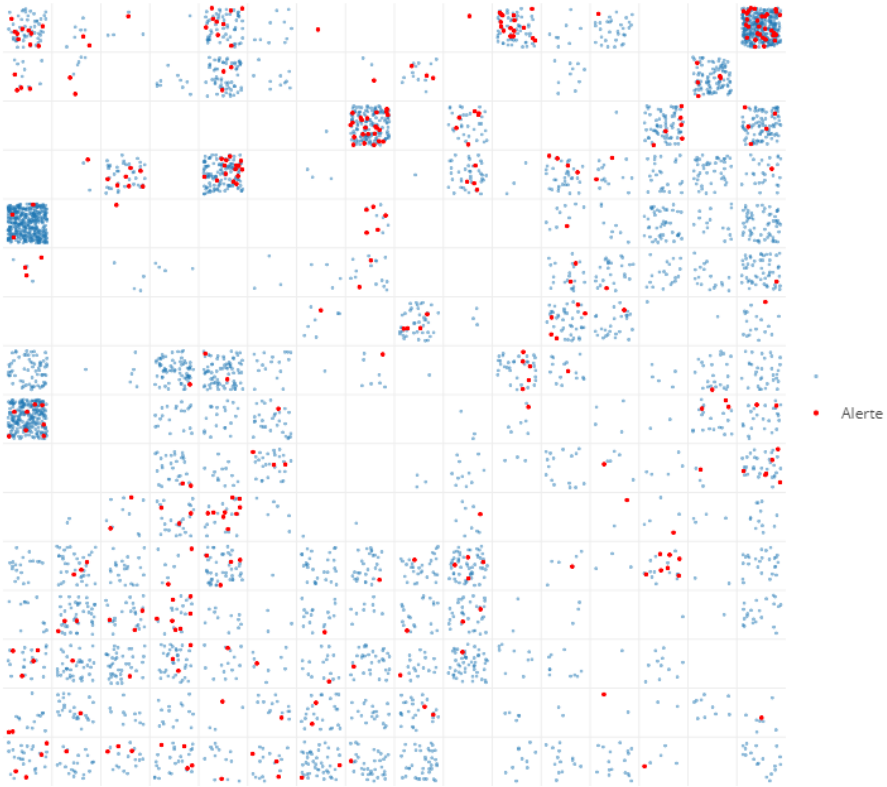
ructure of
ration

- Use every single
  - Users can b
  - Over four
- **Aggregate** even
  - **Behavior**
    for a given
  - Distinguish
- Analysis of **intrin**
  - Using both

ubject and/or target)

period (a week) and

data

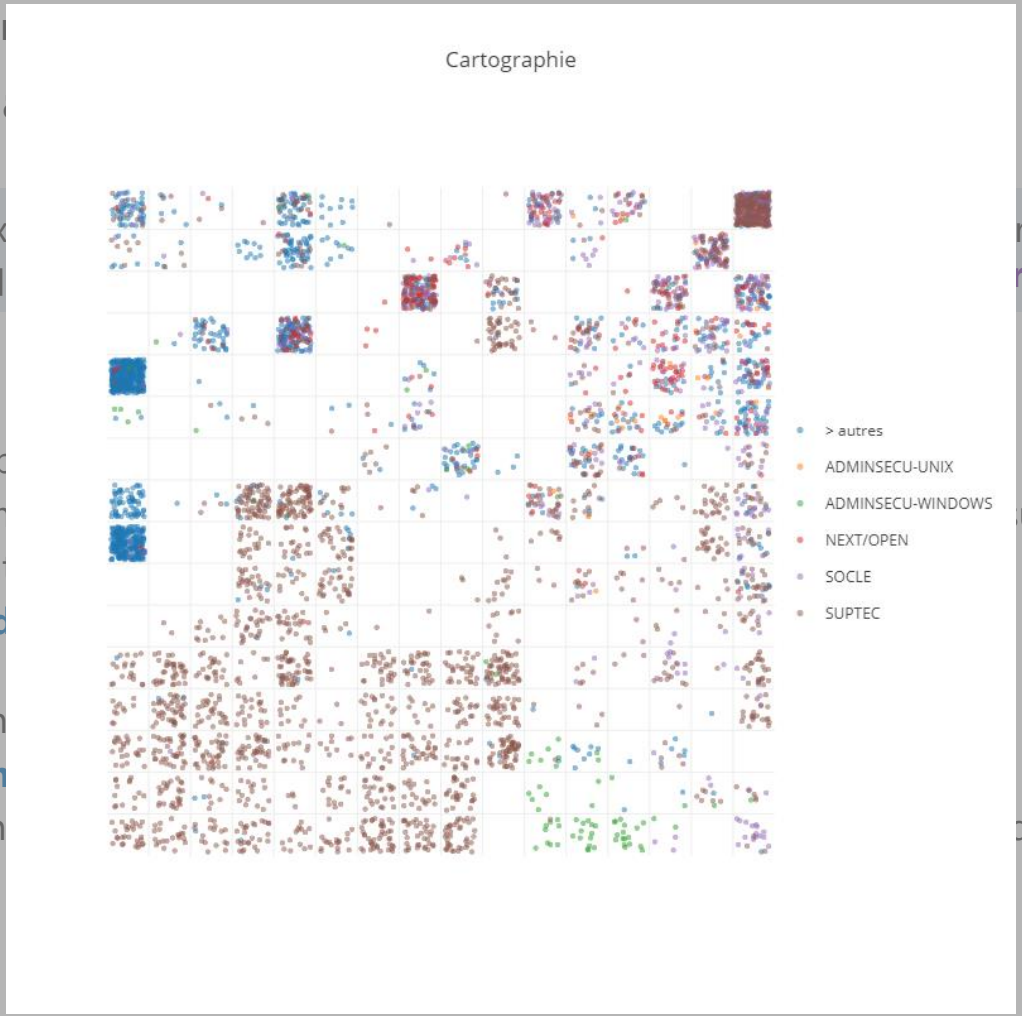


Use cases : imple

Cybersecurity

A first ex … ructure of
Banq … **ration**

- Use every single …
  - Users can b…
  - Over four m… …ubject and/or target)
- **Aggregate** event…
  - **Behavior d**… … period (a week) and for a given …
  - Distinguish…
- Analysis of **intrin**…
  - Using both … …data



Thanks for your attention !

Any question ?

matthias.laporte@banque-france.fr