

---

# Datamining

Gabriel Bacq

CNAMTS



# In a few words

DCCRF uses two ways to detect fraud cases: one which is fully implemented and another one which is experimented:

## 1. Database queries (fully implemented)

Example: extraction from the database of practitioners who earn more than x €.

## 2. Dataming (experimented at a regional level for daily allowances, and at a national level for Additional universal sickness coverage - CMUC)

How it works:

briefly: the machine « makes the job » for us!

We make a group of non defrauders and a group of defrauders by injecting some defrauders and some non defrauders into the system.

Then, when a new person is entered, the machine is supposed to classify him/her in the appropriate group.

Problem:

As we do not have enough examples of defrauders, the system is not relevant yet.

# Fraud detection

- Fundamental step that has to be done before setting up of a litigation control plan.
- Gives the possibility of identifying some atypic stakeholders



# Data

- Health insurance reimbursement basis**

N° of the prescriber healthcare professional	N° of the executive healthcare professional	N° of the beneficiary	Prescription date	Execution date	Drug name	Number of boxes	Paid amount
751XXXX	752XXX	P49123456	2nd June 2011	4th June 2011	Doliprane	2	3.60 €
541XXXX	542XXX	M25495868	3rd August 2011	7th August 2011	Gaviscon	3	9.12 €
561XXXX	222XXX	J987654321	5th May 2011	6th May 2011	Celestène	4	9.6 €

- Are but healthcare consumption**



# Methods

- Reason-based targeting queries:
  - Database queries
- Datamining :
  - Mathematical or statistical method with an automatic learning



# Reason-based targeting queries

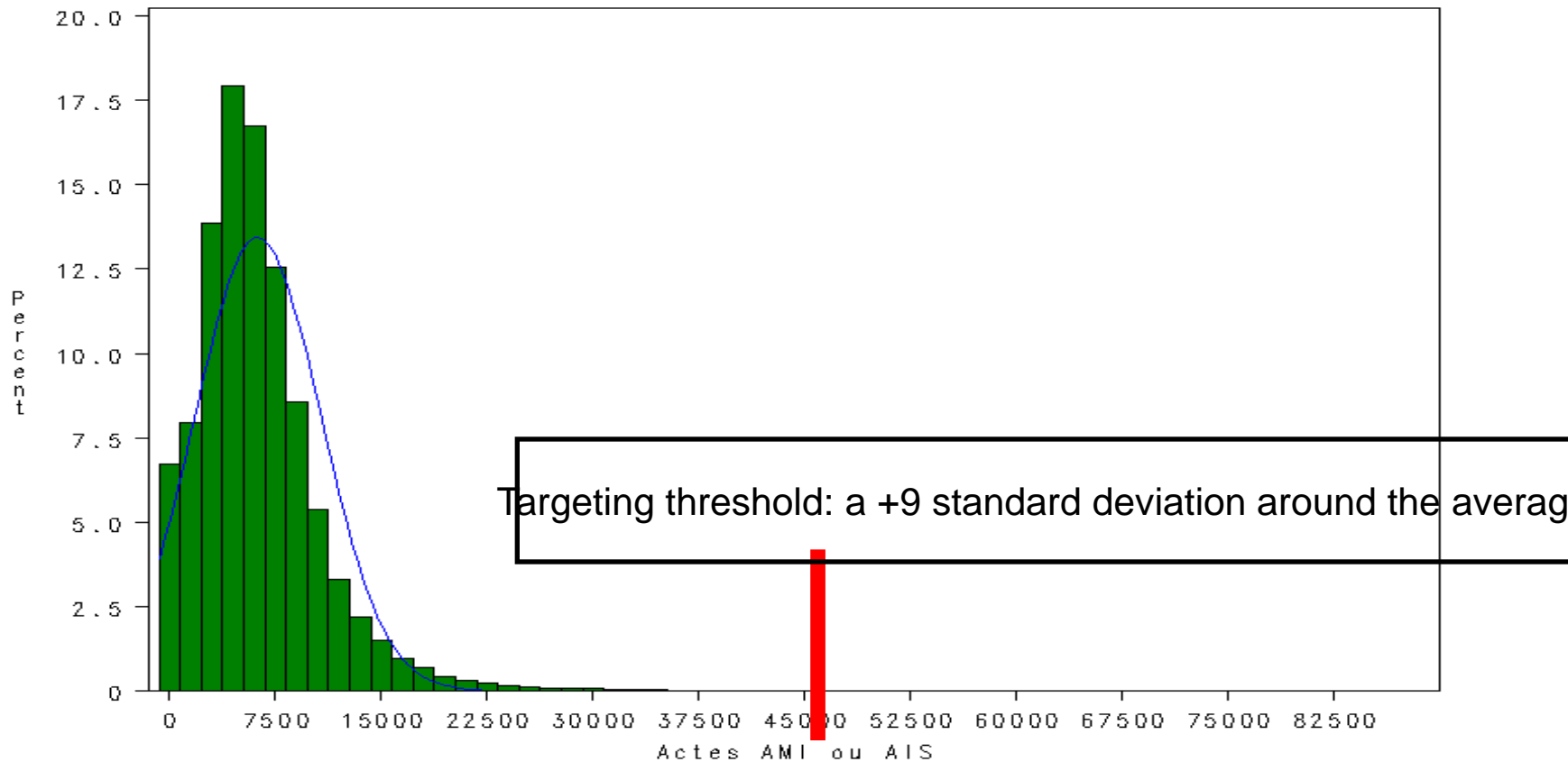
## A few examples

- Targeting **general practitioners** whose reimbursements exceed a certain amount
- Targeting **nurses** who have made an aberrant number of acts
- Targeting **carriers** with a very important number of daily transportations by vehicle
- Targeting **beneficiaries** with an unusual reimbursed amount



# Threshold determination

Example of targeting related to the number of acts performed by nurses



# Datamining

- Experiments of new detection methods
- Identification of new fraud procedures
- Important quantity of data
- Automatic or semi automatic methods

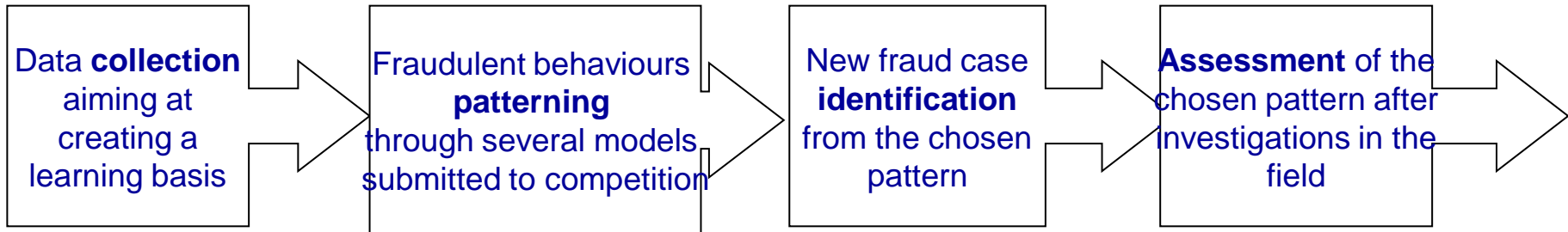


# Supervised learning:

- Learning basis made up of :
  - Proven fraud cases
  - Non fraud cases
- Fraud patterning from predictors
- Use of the pattern to identify potential fraudsters
- Validation of potential fraudsters who are detected through investigations in the field



# The major steps of a **DATAMINING** project



# Learning basis

- **Data gathering to create a learning basis:**
  - **Recognized fraud cases**
    - ↪ Feedback related to fraud cases referred to a court and locally identified with the whole defrauders' characteristics.
  - **Reference cases of non defrauders**
    - ↪ **Ideally:** local identification of recognized non defrauders, after investigations in the field.
    - ↪ Or random drawing from a database of potential non defrauders

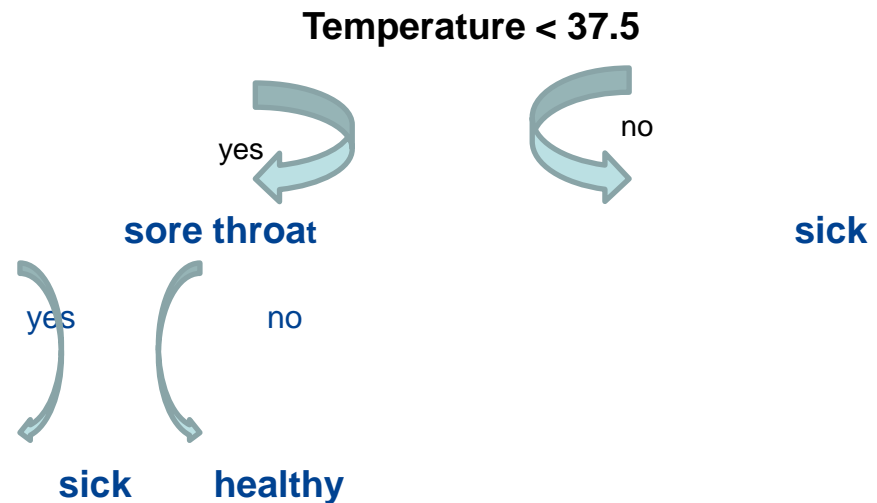
# Fraudulent behaviours

## patterning

Patterning through several learning patterns set up by a specific software.

- **Decision tree**

– *Example:*



- **Logistic regression** (statistical pattern that gives the possibility of assigning every new person who is entered in the system a probability of frauding > we then pay attention to people whose probability of frauding is high)



- **Measure of error or misclassification rate**

*Error*

$$\text{rate} = \frac{FP + FN}{TP + FN + FP + TN}$$

- **Specificity and sensitivity measure**

Tests pattern performance on:

Defrauders

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Non defrauders

$$\text{Specificity} = \frac{T}{FP + TN}$$

- **Measure of negative and positive predictive values**

Tests predictive power of the pattern among:

Defrauders

$$TPP = \frac{TP}{TP + FP}$$

Non defrauders

$$TPN = \frac{FN}{TN + FN}$$