
# The main works and challenges of an insurance Data'Lab

September 2017

# Data'Lab's mission and structure example

**Data'Lab's composition :**

- 4 Actuaries
- 1 Computer Scientist
- 1 Data Scientist
- 4 Statisticians
- 1 Engineer

**Example of work :**

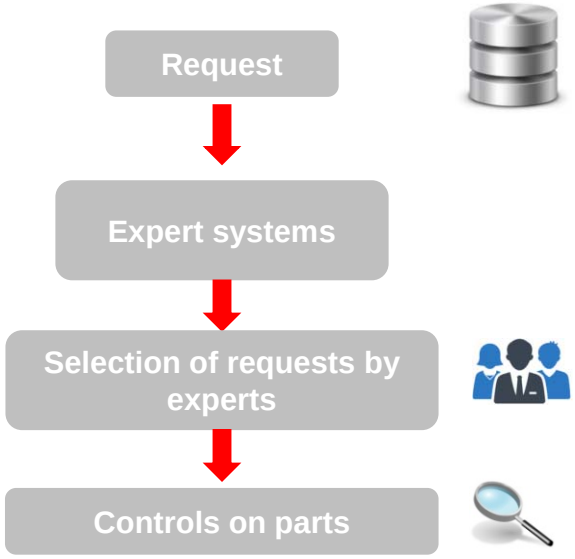
1. **Réaliser des études statistiques « traditionnelles » mais en présence de problématiques de grande dimension (aspect informatique) ou d'évènement extrêmement rare.**
   **Conduct "traditional" statistical studies, with either large-scale and computer-related issues or extremely rare events.**
2. **IOT : Telematics**
3. **Using new data sources : TANL / Image processing to extend existing models or to supply RPA**
4. **IT Project**

# Detection of rare events : fraud

⇨ **Aim** : Increase the sensing capacity of the model by improving its accuracy

⇨ **Development of two algorithms:**
- ➢ A **supervised predictive model** based on proven fraud to capitalize on the history of the checks carried out
- ➢ An **unsupervised clustering model** which relies on the entire data history (controlled or not)

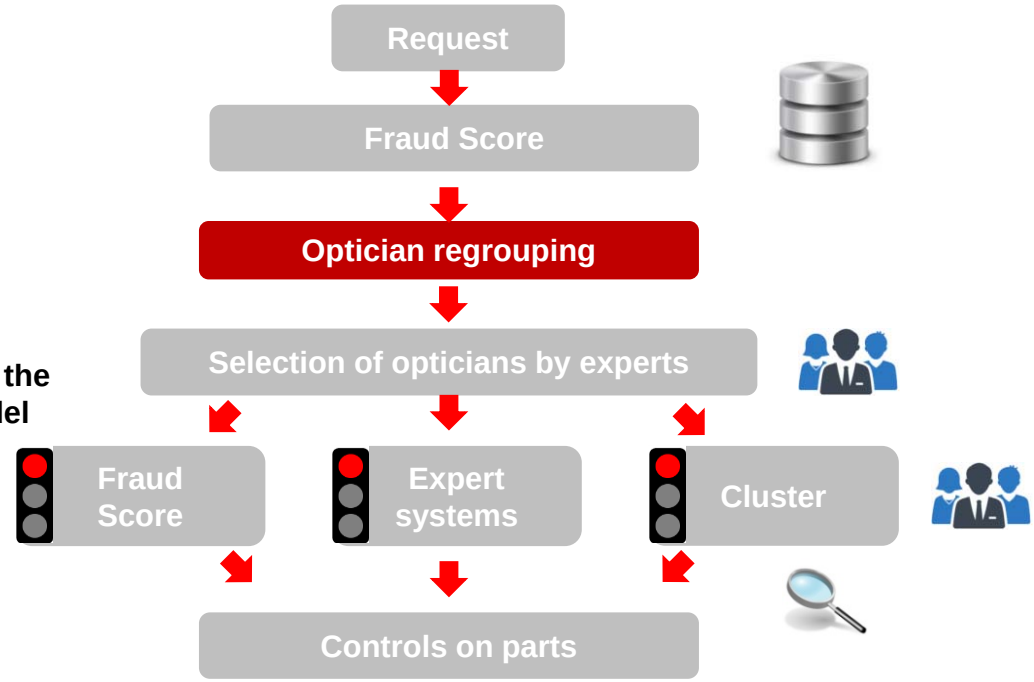
**CURRENT SYSTEM**




**10,000 cases are considered fraudulent with a positive true rate of 40%**

**Automatic readjustment of the predictive model**

**NEW SYSTEM**




**A probability of fraud for each file with a positive true rate of 70%**

# Detection of rare events : fraud

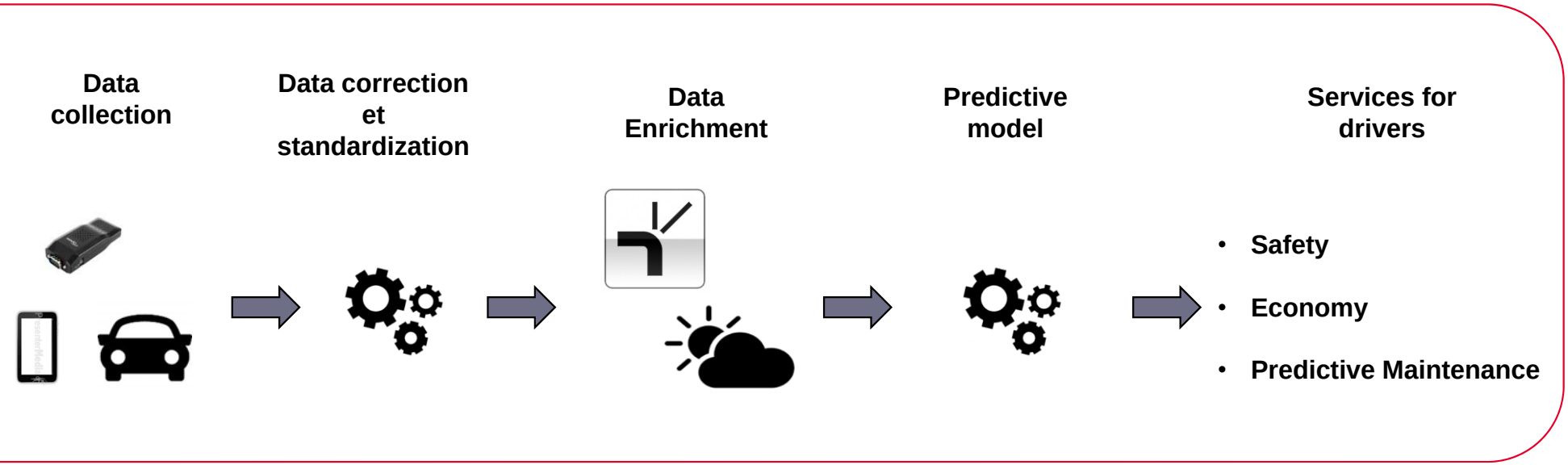
| Step | Role | Comments |
|---|---|---|
| Supervised model | Corrects the expert system weaknesses :<br>• without preconditions<br>• self-learning<br>Provides a probability of fraud for each optician and case. | Given the low number of flagged cases, the main risk is over-learning related:<br>– Select a very low learning rate<br>– Select a low subsample/colsample<br>– Mix several models to gain stability |
| Clustering model | • Secures the predictive model<br>• Plays on expert system thresholds<br>• Detects new fraudulent behavior, regardless of controls already performed<br><br>But<br><br>• Search for similarities in applications instead of fraud : requires the validation of experts<br>• Extremely sensitive to model inputs | The main difficulty lies in the volumetric (O(k x n) complexity)<br>– PCA to reduce the database (risk for the industrialization phase).<br><br>– Calculation of clusters on a sub-base :<br>▪ Calculation of the distance from each point to the center of gravity (via the Euclidean distance).<br>▪ Selection by stratified sampling of 80 000 claim cases on the basis of their distance to the center of gravity<br>▪ Taking into account the claim cases already appraised by the experts.<br><br>– Measurement of the stability of the intra-class inertia by boostraping<br>– The synthesis of each cluster is carried out using variables with the highest deviation test statistic (between the cluster and the rest of the population), to the mean for the numerical variables, to the proportion observed for the qualitative variables. |

**Even if many points can be improved (choice of distance measurement, automatic calculation of the optimal number of clusters), actuaries are naturally able to address this type of work**

# IOT and new insurance products

1. **Aim**
   Take advantage of the data collected by the connected objects in order to propose new insurance produtcs. Among the existing insurance products, automobile insurance is one of the most IOT-interested through the telematics : take advantage of connected cars to improve our knowledge of automobile claims and provide an innovative insurance offer.

2. **Process**




| *Business skills* | *Data Mining*<br>*Business skills* | *Business skills*<br>*Computer skills* | *Machine learning* | *Computer skills* |
|---|---|---|---|---|

**3. Data**

The enrichment part remains the key point of the mission. It requires knowledge of automobile risk, but also best practices in terms of telematic models. These data can be divided into 4 main categories:

- Telematic Data
- Contract Data
- Open Data
- Enrichment based on vehicle physics

| Data | Source | Main concern |
|---|---|---|
| Acceleration | GPS / accelerometer | GPS Frequency |
| Speed | GPS / accelerometer | GPS Frequency |
| Curvature | GPS - compass | GPS Frequency |
| Car Dimensions | Contract, Market Data,… | |
| Car Adherence,… | Start-up | |
| Weather | API | On-the-fly Enrichment |
| Road Type | API | On-the-fly Enrichment |
| Traffic | API | On-the-fly Enrichment |

## 4. The main issues and problems to be solved

- Large-scale problem: 300 GB for 50 000 vehicles observed over 1 year
  - In the POC phase : need to integrate upstream this constraint during the pre-treatment phase
  - In the industrialization phase: migrate algorithms in languages dedicated to Big Data (SPARK).

- Enrichment of data:
  - API Mastery
  - Stock enrichment issue (assign a weather report to each GPS point of more than 100 million journeys).

- Rare events issue (*under*/*over samplling*, SMOTE) :
  - About 1/4000 in terms of claims per trip.
  - About 1,5% in terms of hazardous events per trip but a low dependency between claims and dangerousness.

- The communication with the driver and the link with his fare.

Les solutions :

**A relatively classic project for actuaries with an essential business line**

# TNL & GDPR

## Context

In some situations, the presence of prohibited terms is noted in the memos of the client advisors. This can be detrimental to the company.

*"Only objective, directly relevant and strictly necessary information can be entered in this field (specify the purpose of the field, eg management of the file, collection), excluding any subjective assessment or any direct or indirect reference. indirectly to racial origins, to political, philosophical or religious opinions, trade union affiliation or the morals of the persons concerned or, in general, to the privacy of his private life. They must be communicable in plain language to the person concerned."*

*(Source : Société Général Group Portal – Informatique et libertés)*

## Aim

Being able to decrease the occurrence of verbatim with:

- Sensitive data (religious opinion, sexual orientation, etc.)
- Judgment of value on the personal financial situation of the person concerned
- Assessment of a client's social difficulties
- Any behavioral, social or physical remarks
- Data of offenses or convictions

# TNL & GDPR

1 **Prohibited words Detection in the verbatim**

- Creating a dictionary of forbidden words (scraping, business, etc.)
- Automatic text processing (typing errors, spelling, rooting, etc.)
- Detection of problematic verbatim

*Quickly gives a first score and supplies the process of constructing the supervised model*

2 **Supervised Machine Learning**

- Flag "by hand" of verbatim (business)
- Automatic text processing
- Creating indicators from verbatims (N-gram, etc.) and calibrating a Machine Learning model

*Provides a pourcentage of problematic verbatims and identifies main verbatims*

3 **Sampling**

- Verbatim sampling allowing to know the % of problematic verbatims in the global corpus.
- Comparison between % of problematic verbatims and % of verbatims detected by the supervised algorithm.
- Updating the supervised model if the two percentages are not consistent.

*Monitors the quality of prediction of the supervised model, and allows to trigger an update of the model and / or the dictionary*

**Deliverable :**

In addition to the score, the model shows the items that were mostly in alert. Allows to focus the training on the detected problems and to customize it if necessary.

# TNL & GDPR

The main stages of the project :

| | |
|---|---|
| Construction of a PNL toolbox in French | Tokenization :<br>Stemming<br>Typing erros : |
| Preprocessing of verbatim | Stop words,<br>Pondération des mots |
| Descriptive statistics | Descriptive statistics<br>Words frequency display<br>Co-occurrence analysis<br>Central word analysis<br>t-sne |
| Modeling | Logit, Xgboost,…<br>ngrams, tfidf<br>Explanation of the model (LIME). |

**A new project for the actuary, but addressed by the algorithms club**

# Data'Lab and Institute of Actuaries

**The missions of a Data'Lab naturally find an echo within the subjects addressed by the IA Innovation Commission**

- Data Science School : allows actuaries to acquire or consolidate the key knowledge needed to carry out machine learning tasks.
- Pricing Game : competition allowing actuaries to test their machine learning methods on concrete cases
- Algo Club : presentation of use cases of innovative algorithms (last example: breakfast of September 20, 2017: "Deep learning applied to auto claims").
- Beyond the statistical and data processing aspects, Data Science raises many ethical and legal issues (GDPR). These points are addressed by the working group "Ethics and Standards"
- IOT : working group on the impact of connected objects on insurance products.
- Round tables, conferences, …

Through the Innovation Commission, the Institute of Actuaries has set up an ecosystem enabling actuaries to consolidate their knowledge and to acquire the new skills introduced by Data Science.