
INSTITUT DES
ACTUAIRES

La Data Science et son apport pour mieux détecter les besoins de prévention des assurés en santé

Alexandra BARRAL, Romain GAUCHON, Nabil RACHDI
ACTUARIS

INSTITUT DES ACTUAIRES

Les origines

1943 — A. Turing
1956 — Shannon, McCarthy, Minsky et al.
1970
1980 — V. Vapnik & A. Chervonenkis
1990
1995
2001
Auj.

INTERNET

- Premiers travaux d'« IA »
- Intelligence Artificielle !
- Moins d'intérêt pour l'IA
- Regain d'intérêt avec d'autres applications
- Machine Learning !
- Statistical Learning !
- Data Science !

comportement humain

comportement humain & systèmes complexes

chimie, médicale

robotiques, logiciels

gafa, web services, marketing

finance
Industrie
assurances
etc.

- 1er réseaux de neurones

+

- Modèles Markoviens
- Simulation de Monte-Carlo
- Réseaux Bayésiens

+

- Arbres de décisions
- SVM
- Reg. Proc. Gaussien
- Reg. à noyaux

+

- Deep Learning
- Active Learning
- Etc.

Première apparition du terme « data science »

Data Science !

International Statistical Review (2001)

Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

INSTITUT DES ACTUAIRES

Quelle(s) différence(s) ?

Principalement culturelle !

- **Glossaire** de R. Tibshirani (Pr. à Stanford)
  co-auteur de « The Elements of Statistical Learning », 2008

- MACHINE Learning
  = approche algorithmique

- STATISTICAL Learning
  = approche par modélisation de la donnée

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

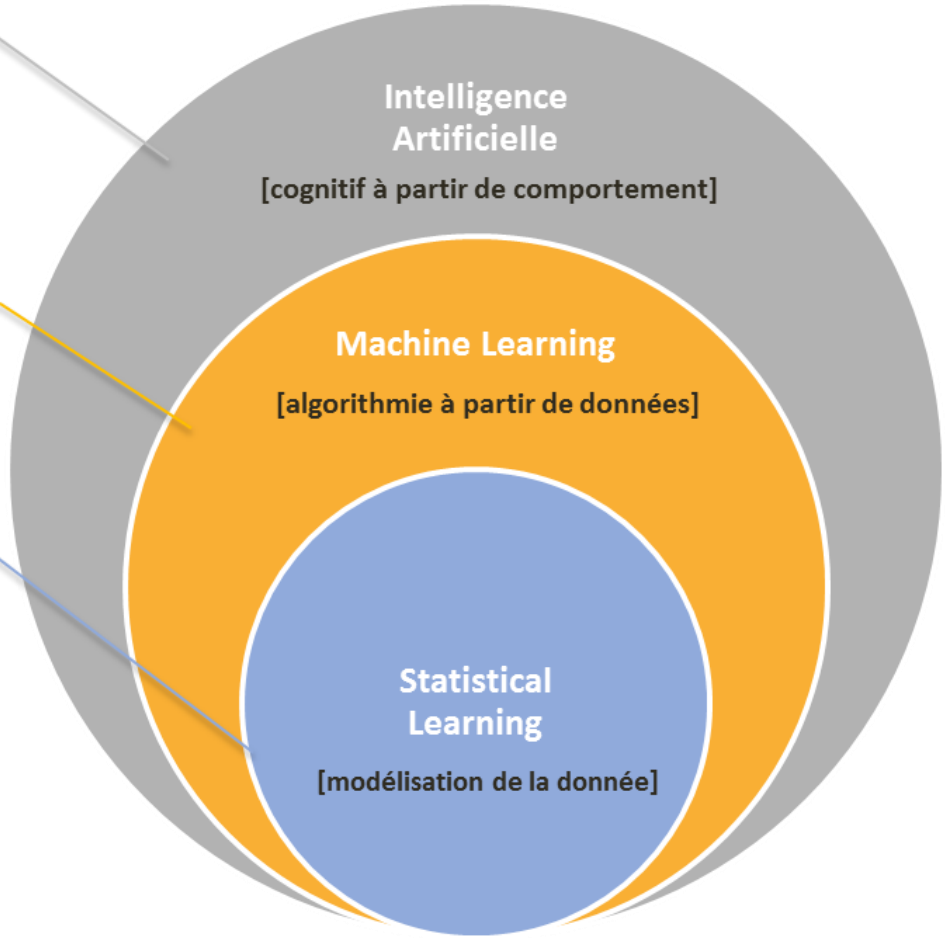
INSTITUT DES ACTUAIRES

En résumé

« Imitation »
comportement humain

« Reconnaissance »
image, texte, son, etc.

« Stochastique »
risque, fiabilité, etc.



INSTITUT DES ACTUAIRES

Exemple : Les GLM revisitées

Vue « classique »

Y_i = fréquences
$\mathbf{X}_i$ = contrats

Vue « Statistical Learning »

1/ Modélisation des données

- $Y \sim \mathcal{P}(\lambda)$
- $\log(\mathbb{E}(Y|\mathbf{X}=\mathbf{x})) = \sum_{i=1}^{d} \beta_i x_i = \mathbf{x}^T\boldsymbol{\beta}$

Soit

$$(Y|\mathbf{X}=\mathbf{x}) \approx \mathcal{P}[e^{\mathbf{x}^T\boldsymbol{\beta}}]$$

2/ Maximum de Vraisemblance des données observées

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmax}} \prod_{i=1}^{n} \mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)$$

Maximum de Vraisemblance

INSTITUT DES ACTUAIRES

Exemple : Les GLM revisitées

Y_i = fréquences
$\mathbf{X}_i$ = contrats

Vue « classique »

1/ Modélisation des données

- $Y \sim \mathcal{P}(\lambda)$
- $\log(\mathbb{E}(Y|\mathbf{X}=\mathbf{x})) = \sum_{i=1}^{d} \beta_i x_i = \mathbf{x}^T\boldsymbol{\beta}$

Soit

$$(Y|\mathbf{X}=\mathbf{x}) \approx \mathcal{P}[e^{\mathbf{x}^T\boldsymbol{\beta}}]$$

2/ Maximum de Vraisemblance des données observées

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmax}} \prod_{i=1}^{n} \mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)$$

Maximum de Vraisemblance

Vue « Statistical Learning »

1/ Quantité d'intérêt cible (*« target »*)

$$\text{loi jointe de } (\mathbf{X}, Y) \rightarrow f^*(\mathbf{x}, y)$$

2/ Modélisation de la cible (*« model »*)

$$\text{modèle : } f_{\mathbf{X}}(\mathbf{x})\,\mathcal{P}[e^{\mathbf{x}^T\boldsymbol{\beta}}](y) \rightarrow f_{\boldsymbol{\beta}}(\mathbf{x}, y)$$

3/ Fonction de perte associée à la cible (*« loss function »*)

$$\text{métrique considérée} \rightarrow \text{Entropie Relative}$$

aussi appelée divergence de Kullback-Leibler $KL(f_1, f_2) = \int \log\left(\frac{f_1}{f_2}\right) f_1$

4/ Minimisation Empirique de la fonction de perte (*« ERM »*)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}}\, KL(\widehat{f^*}, f_{\boldsymbol{\beta}})$$
$$= \underset{\boldsymbol{\beta}}{\text{Argmax}} \sum_{i=1}^{n} \log\left(\mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)\right)$$

INSTITUT DES ACTUAIRES

Exemple : Les GLM revisitées

Y_i = fréquences
$\mathbf{X}_i$ = contrats

Vue « classique »

1/ Modélisation des données

- $Y \sim \mathcal{P}(\lambda)$
- $\log(\mathbb{E}(Y|\mathbf{X}=\mathbf{x})) = \sum_{i=1}^{d} \beta_i x_i = \mathbf{x}^T\boldsymbol{\beta}$

Soit

$$(Y|\mathbf{X}=\mathbf{x}) \approx \mathcal{P}[e^{\mathbf{x}^T\boldsymbol{\beta}}]$$

2/ Maximum de Vraisemblance des données observées

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmax}} \prod_{i=1}^{n} \mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)$$

Vue « Statistical Learning »

1/ Quantité d'intérêt cible (*« target »*)

$$\text{loi jointe de } (\mathbf{X}, Y) \rightarrow f^*(\mathbf{x}, y)$$

2/ Modélisation de la cible (*« model »*)

$$\text{modèle : } f_{\mathbf{X}}(\mathbf{x})\,\mathcal{P}[e^{\mathbf{x}^T\boldsymbol{\beta}}](y) \rightarrow f_{\boldsymbol{\beta}}(\mathbf{x}, y)$$

3/ Fonction de perte associée à la cible (*« loss function »*)

$$\text{métrique considérée} \rightarrow \text{Entropie Relative}$$

aussi appelée divergence de Kullback-Leibler $KL(f_1, f_2) = \int \log\left(\frac{f_1}{f_2}\right) f_1$

4/ Minimisation Empirique de la fonction de perte (*« ERM »*)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}}\, KL(\widehat{f^*}, f_{\boldsymbol{\beta}})$$
$$= \underset{\boldsymbol{\beta}}{\text{Argmax}} \sum_{i=1}^{n} \log\left(\mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)\right)$$

Maximum de Vraisemblance = Minimisation de l'Entropie Relative

INSTITUT DES ACTUAIRES

Exemple : Les GLM revisitées

Y_i = fréquences
$\mathbf{X}_i$ = contrats

Vue « classique »

1/ **Modélisation des données**

- $Y \sim \mathcal{P}(\lambda)$
- $\log(\mathbb{E}(Y|\mathbf{X}=\mathbf{x})) = \sum_{i=1}^{d} \beta_i x_i = \mathbf{x}^T\boldsymbol{\beta}$

Soit

On peut **changer** de « target »
et/ou de « model »
et/ou de « loss function » !

2/ **Maximum de Vraisemblance des données observées**

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmax}} \prod_{i=1}^{n} \mathcal{P}[e^{\mathbf{x}_i^T\boldsymbol{\beta}}](y_i)$$

Maximum de Vraisemblance

Vue « Statistical Learning »

1/ **Quantité d'intérêt cible (*« target »*)**

2/ **Modélisation de la cible (*« model »*)**

3/ **Fonction de perte associée à la cible (*« loss function »*)**

4/ **Minimisation Empirique de la fonction de perte (*« ERM »*)**

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}}\ \text{loss}(\widehat{target}, target_{\boldsymbol{\beta}})$$

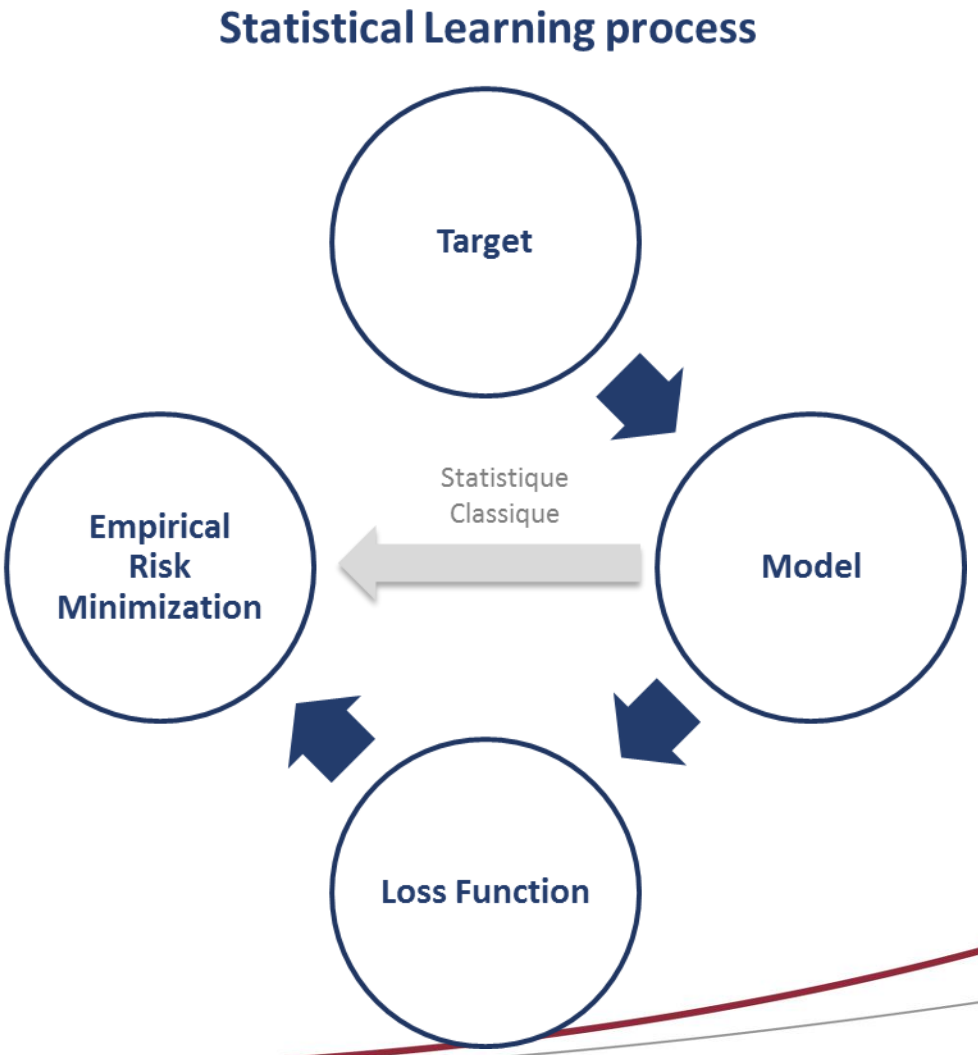
≠

Autre(s) algorithme(s)

INSTITUT DES ACTUAIRES

En résumé

- Le Statistical Learning **étend les méthodes de prédictions statistiques classiques** à un cadre d'analyse beaucoup plus large
- donnant **plus de flexibilité** à la modélisation
- permettant l'utilisation d'une **grande variété d'algorithmes d'apprentissage**
- Il est **important de savoir ce qui se cache derrière les algorithmes**
  (quelle target ? quelle métrique ?)



ML & SL : Un arsenal d'algorithmes

CONTEXTE

TACHE	Supervisé $(\mathbf{X}_i, Y_i)_{i=1:n}$	Non-supervisé $(\mathbf{X}_i)_{i=1:n}$	Semi-supervisé mix
Régression	▪ Régression linéaire (Lasso, Ridge, etc.) ▪ Régression non-linéaire (noyaux, etc.) ▪ Arbres (CART, Random Forest, etc.) ▪ kNN ▪ Réseaux de Neurones, SVR ▪ GLM ▪ Etc.	X	• SVM • Graphs
Classification	▪ Régression logistique ▪ Réseaux Bayésiens ▪ Arbres ▪ Réseaux de Neurones, SVM ▪ Etc.	X	• SVM • Graphs
Clustering	X	• K-Means • Densités • Cartes Kohonen	X
Réduction Dim.	• ANOVA (Analyse de la variance) • Algorithmes génétiques • Etc.	• Analyse des corrélations • ACP	X

méthodes traditionnelles en assurance

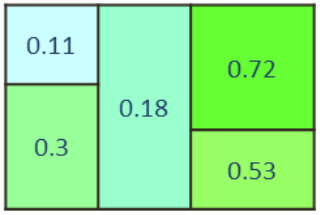
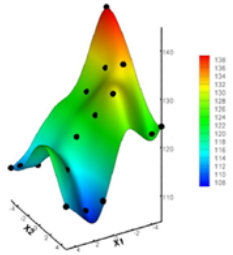
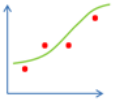
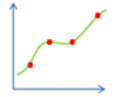
Pour un couple (Tâche, Contexte) donné, l'algorithmie dépendra de la **NATURE des données**

- **Quantitatives** (discrètes, continues)
- **Qualitatives** (ordinales, nominales)

ALGO = TACHE + CONTEXTE + NATURE DONNEES

INSTITUT DES ACTUAIRES

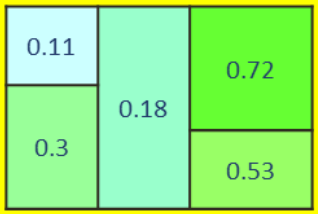
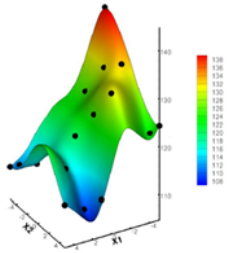
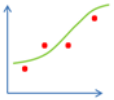
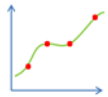
ALGO = TACHE + CONTEXTE + NATURE DONNEES

Nature des Inputs	Nature des algorithmes		
	Familles de méthodes	Méthodes	Algorithmes
Catégorielles et/ou Continues	« Ensemblistes » *Approximation en partitionnant de l'espace des entrées* 	• Arbres • Forêts	• CART • AdaBoost • RandomForest • XGBoost • Gradient Boosting
Continues et « peu » Catégorielles	« Fonctionnelles » *Approximation par représentation fonctionnelle* 	• Polynômes • Radial Basis Functions • Processus Gaussien	**Extrapolation** • Ordinary Least Squares • Polynomial Chaos • Tensorization • MARS   **Interpolation** • Kriging (process. gaussien) • Sparse Kriging

INSTITUT DES ACTUAIRES

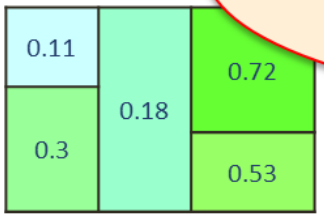
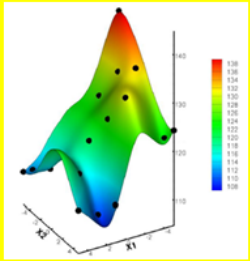
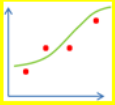
ALGO = TACHE + CONTEXTE + NATURE DONN

Les **données assurantielles** sont très souvent un **mix catégorielles/continues**. D'où l'exploration croissante des **méthodes ensemblistes** (arbres)

Nature des Inputs	Famille méthodes		algorithmes
Catégorielles et/ou Continues	« Ensemblistes » *Approximation en partitionnant de l'espace des entrées* 	• Arbres • Forêts	• CART • AdaBoost • RandomForest • XGBoost • Gradient Boosting
Continues et « peu » Catégorielles	« Fonctionnelles » *Approximation par représentation fonctionnelle* 	• Polynômes • Radial Basis Functions • Processus Gaussien	**Extrapolation** • Ordinary Least Squares • Polynomial Chaos • Tensorization • MARS   **Interpolation** • Kriging (process. gaussien) • Sparse Kriging

INSTITUT DES ACTUAIRES

ALGO = TACHE + CONTEXTE + NATURE DONNEES

Nature des Inputs	Nature des algorithmes		
	Familles de méthodes	Méthodes	Algorithmes
Catégorielles et/ou Continues	« Ensemblistes » *Approximation en part* *l'espace des en* 		Gradient Boosting
Continues et « peu » Catégorielles	« Fonctionnelles » *Approximation par représentation fonctionnelle* 	• Polynômes • Radial Basis Functions • Processus Gaussien	**Extrapolation** • Ordinary Least Squares • Polynomial Chaos • Tensorization • MARS   **Interpolation** • Kriging (process. gaussien) • Sparse Kriging

L'approche « fonctionnelle » peut se révéler **plus performante** dans les cas avec **peu de variables catégorielles**

INSTITUT DES ACTUAIRES

« NO FREE LUNCH » THEOREM

Aucun algorithme ne domine tous les autres pour chaque type de problème et de caractéristiques attendues

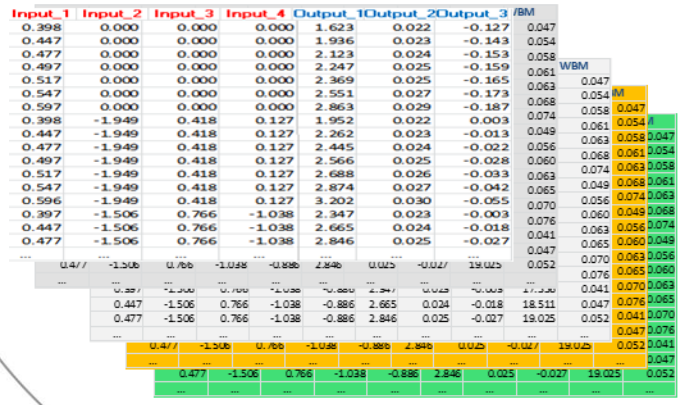
Caractéristiques du prédicteur	Algorithmes			
	GLM/Lasso	Kriging/Chaos	Deep Learning	Random Forest
Données « mixtes »	=	−	−	+
Automatique	+	+	−	+
Interprétation	+	=	−	=
Prédictivité	=	+	+	+
Facteurs d'importances	+	+	=	=
Facteurs Interactifs	−	+	+	−

INSTITUT DES ACTUAIRES

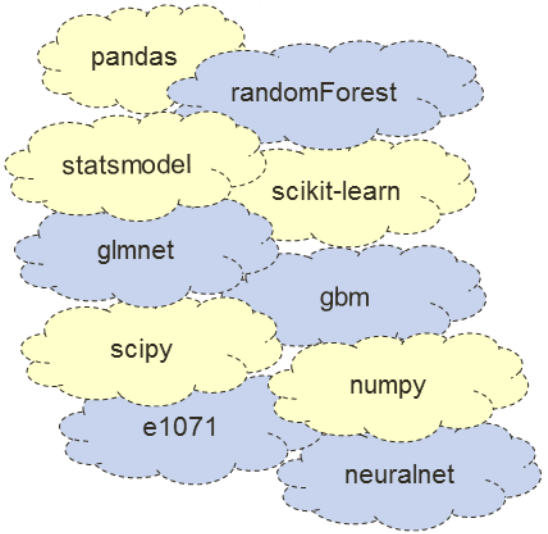
Méthodologie générale

Pourquoi une méthodologie ?

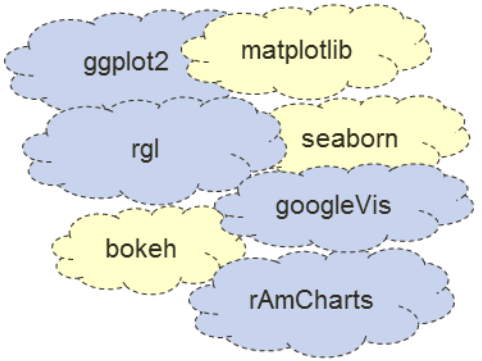
Datasets



Algorithmes



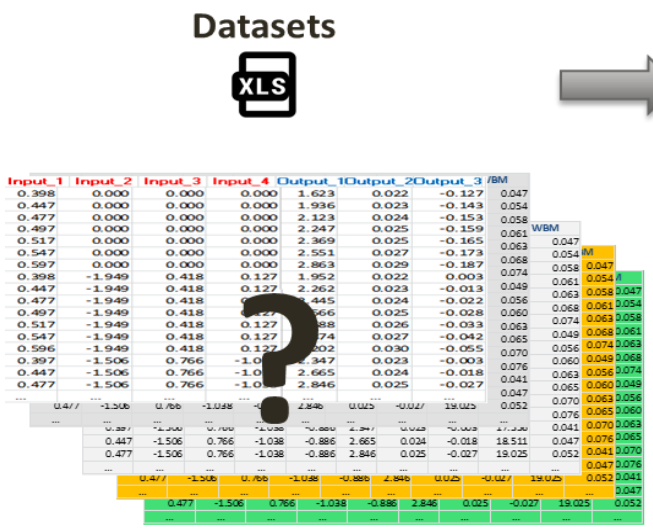
Visualisation

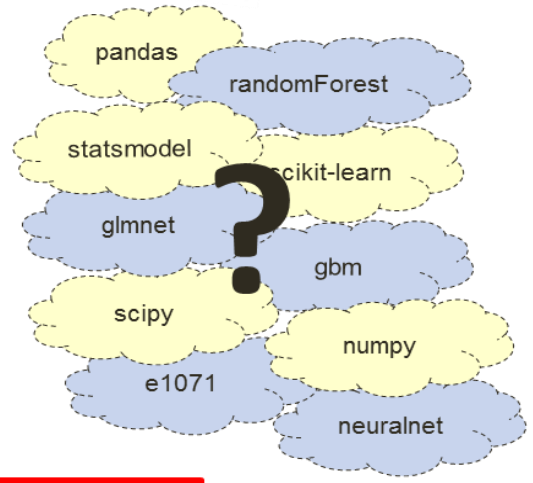


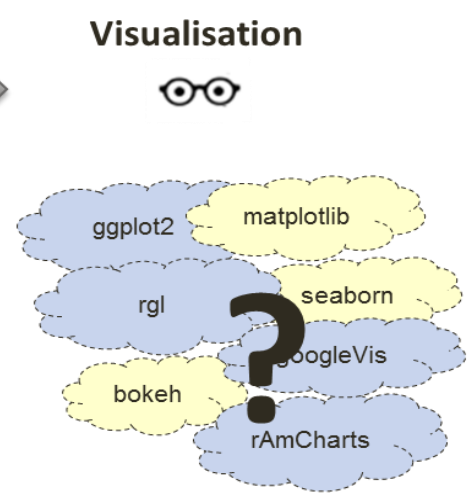
INSTITUT DES ACTUAIRES

Vers une approche méthodologique

- **Pourquoi une méthodologie ?**







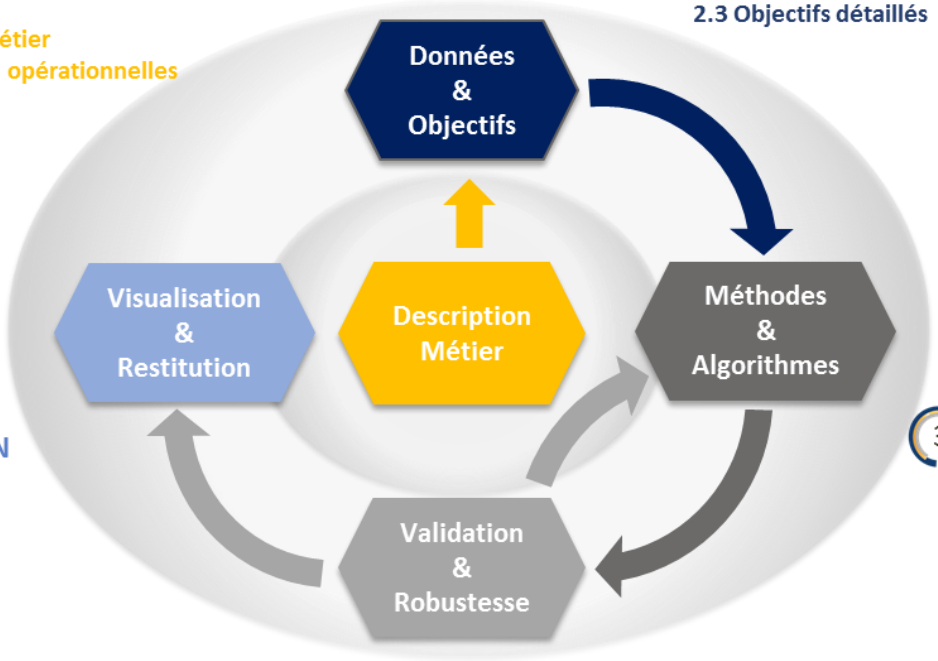
- Comment structurer la base de données ?
- pour quel(s) algorithme(s) ?
- pour quelle fin ?

Besoin d'une méthodologie structurante

INSTITUT DES ACTUAIRES

1. DESCRIPTION METIER
- 1.1 Contexte
- 1.2 Objectifs métier
- 1.3 Contraintes opérationnelles

2. DONNEES ET OBJECTIFS
- 2.1 Sources des données
- 2.2 Description de la base de données
- 2.3 Objectifs détaillés

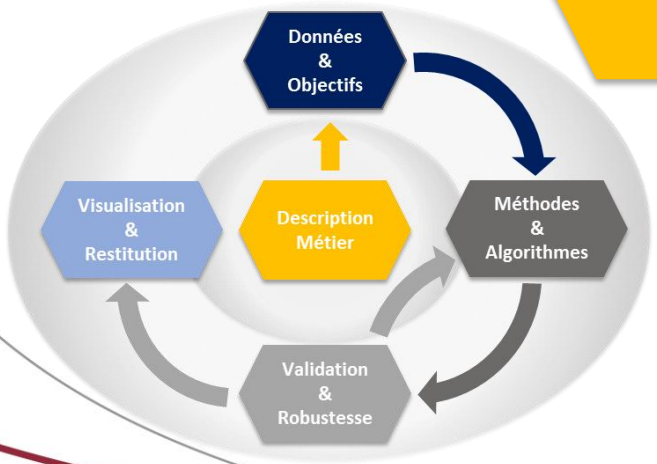


3. METHODES ET ALGORITHMES
- 3.1 Méthodes utilisées
- 3.2 Algorithmes

4. VALIDATION ET ROBUSTESSE
- 4.1 Validation intrinsèque et par recoupement
- 4.2 Robustesse des algorithmes
- 4.3 Feedback avec les algorithmes

5. VISUALISATION & RESTITUTION
- 5.1 Communication des résultats
- 5.2 Visualisation
- 5.3 Intégration et Déploiement

DESCRIPTION METIER



INSTITUT DES ACTUAIRES

Définition de la prévention

Prévention Primaire

- Agir en amont du risque pour empêcher qu'il ne se réalise
- Définition naturelle de la prévention

Prévention Secondaire

- Détection précoce du risque
- Repérer la réalisation du risque le plus tôt possible afin de le traiter

Prévention Tertiaire

- Eviter l'aggravation du sinistre - curatif
- Prévenir les rechutes

Enjeux et problématiques

Assurés

- Services personnalisés
- Au-delà du paiement des soins
- Meilleure connaissance des assurés

Commercial

- Amélioration de l'image de l'entreprise
- Créer une relation de confiance
- Distinction des concurrents

Risques

- Projection de l'amélioration de notre risque (moins de sinistres)
- Baisse des provisions

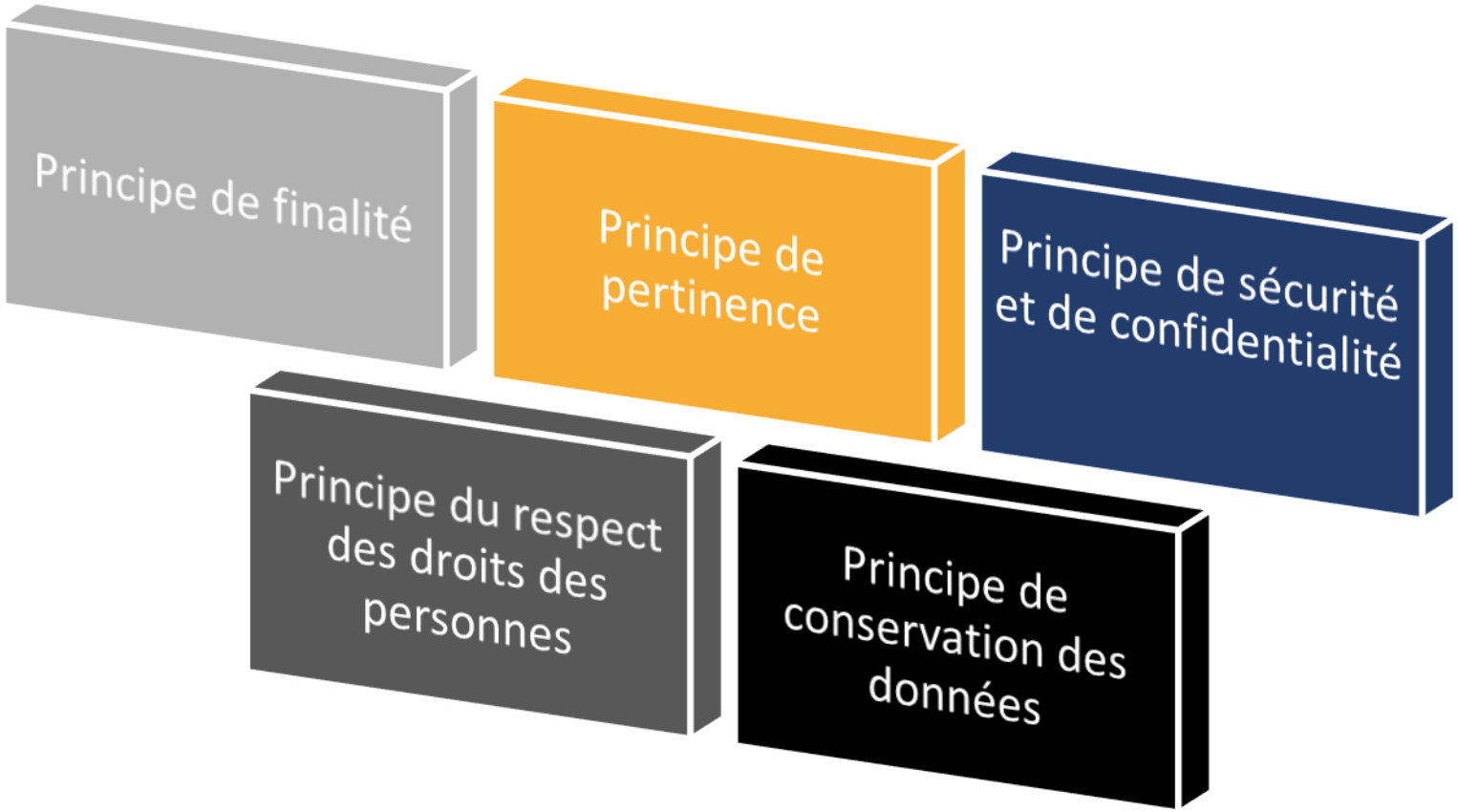
Les prérequis à identifier

La prévention

- Quel risque veut-on diminuer ?
- Quelle action mettre en place ?
- Pour quelles personnes ?
- Les assurés vont-ils participer ?
- Comment évaluer le programme de prévention ?

INSTITUT DES ACTUAIRES

Le contexte RGPD



RGPD : règlement européen pour la protection des données

INSTITUT DES ACTUAIRES

Base de données BENEFICIAIRES

Informations relatives aux bénéficiaires :

- Identifiant du bénéficiaire
- Identifiant de l'assuré associé
- Période de couverture
- Type de bénéficiaires : Assuré, Conjoint, Enfant
- Année de naissance du bénéficiaire
- Sexe du bénéficiaire
- Identifiant de la zone géographique
- Niveau de la couverture santé
- Eventuellement : Secteur d'activité, CSP, régime spécifique, situation familiale et nombre d'enfants à charge de l'assuré…

Des variables permettant d'affiner la segmentation des bénéficiaires et d'intégrer **d'autres critères complémentaires à l'âge et au sexe**

INSTITUT DES ACTUAIRES

Base de données PRESTATIONS

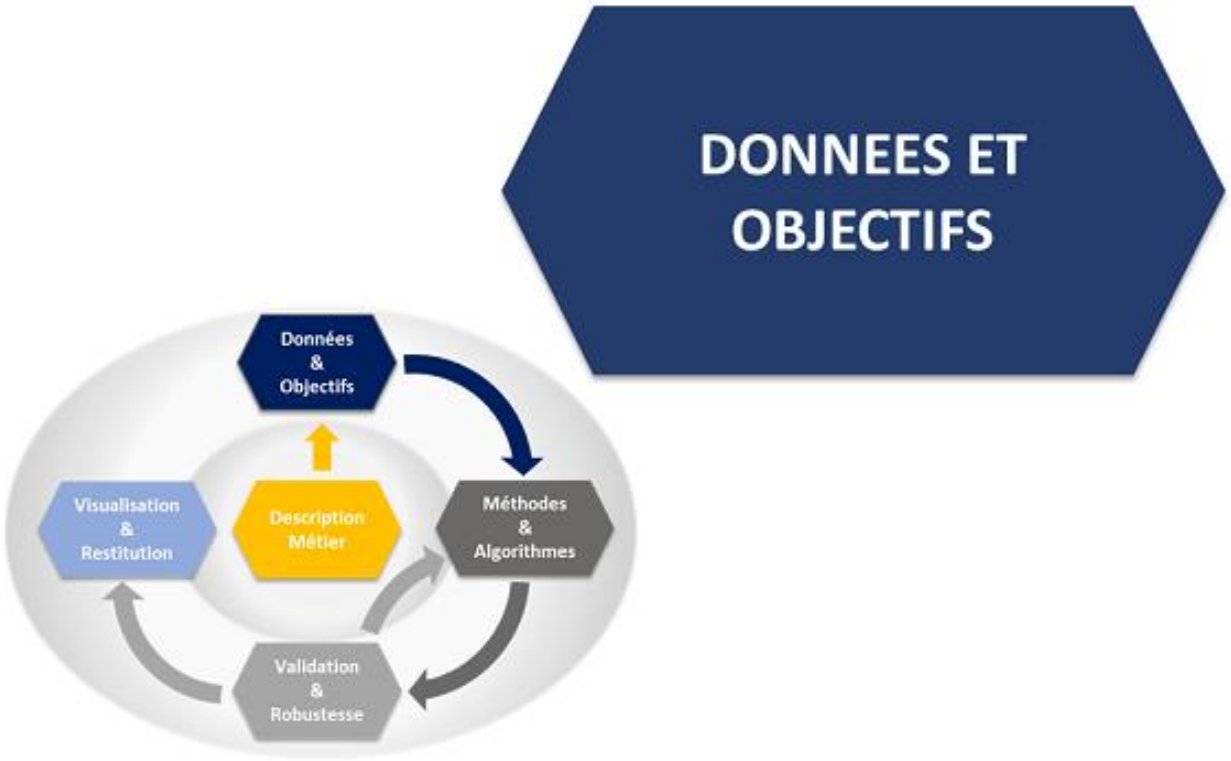
Informations relatives à la consommation des bénéficiaires :

- Identifiant du bénéficiaire
- Date de soins
- Acte concerné

(en fonction d'une segmentation + ou – fine des actes)

- Dépense réelle
- Base de remboursement
- Remboursement obligatoire et taux de prise en charge
- Remboursement complémentaire, Reste à charge
- Eventuellement : indicateur d'une hospitalisation, d'un parcours de soins spécifique, spécialité consultée ou prescripteur…

Des informations plus ou moins agrégées, plus ou moins appauvries **dont la finesse des résultats va dépendre**


DONNEES ET OBJECTIFS
Données & Objectifs
Description Métier
Méthodes & Algorithmes
Validation & Robustesse
Visualisation & Restitution

La base de données prestations

➢ **Base de prestation santé utilisée pour l'étude :**

- Contrat d'assurance santé à adhésion individuelle
- Un seul niveau de garantie
- 80 000 personnes observées sur une année
- 80 libellés d'actes

FEMMES			HOMMES		
Moins de 45 ans	45-61 ans	Plus de 62 ans	Moins de 45 ans	45-61 ans	Plus de 62 ans

20 000 personnes

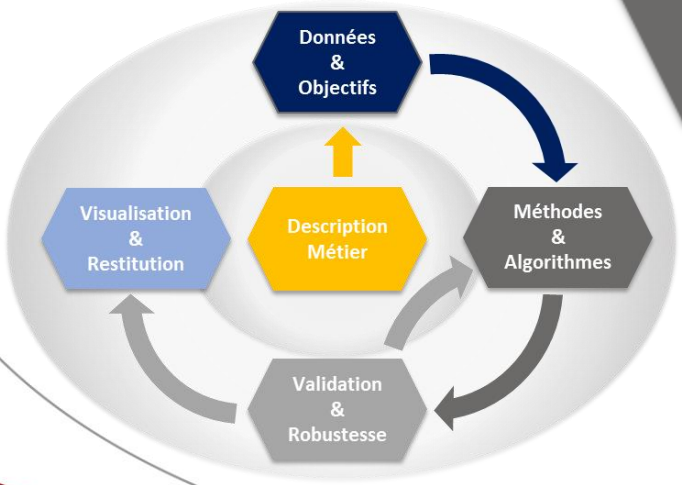
Le ciblage optimisé d'actions de prévention

Classifier les assurés de manière non supervisée pour proposer des actions de prévention

Identifier des comportements de consommation

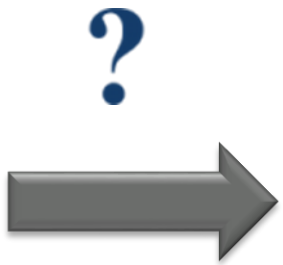
Constituer des groupes d'assurés qui ont les mêmes comportements de consommation

MODELISATION ET ALGORITHMES



INSTITUT DES ACTUAIRES

Quelle approche non supervisée ?



MODELISATION

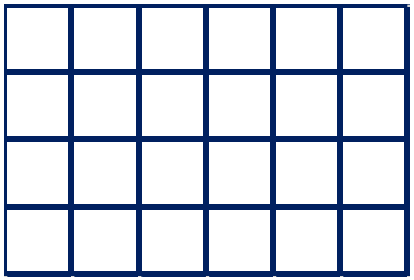
PRE-TRAITEMENT

REDUCTION DE DIMENSION

CLUSTERING

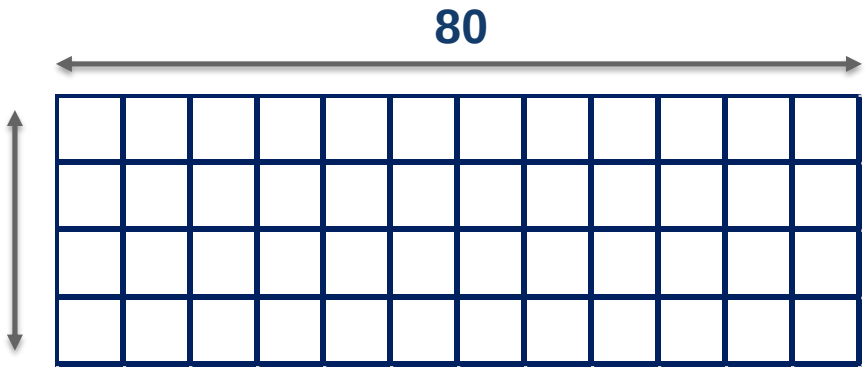
Passage en fréquence

Base de données de prestations santé



Toutes les prestations de chaque assuré

Base de fréquences
Avec une colonne par acte

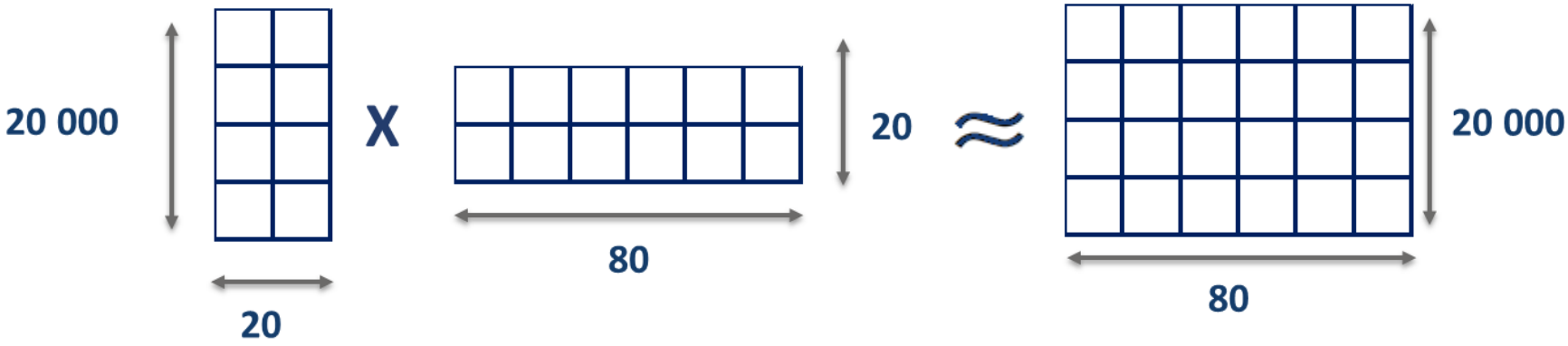


Pour chaque assuré, fréquence pour chaque acte consommé

INSTITUT DES ACTUAIRES

Réduction de dimension

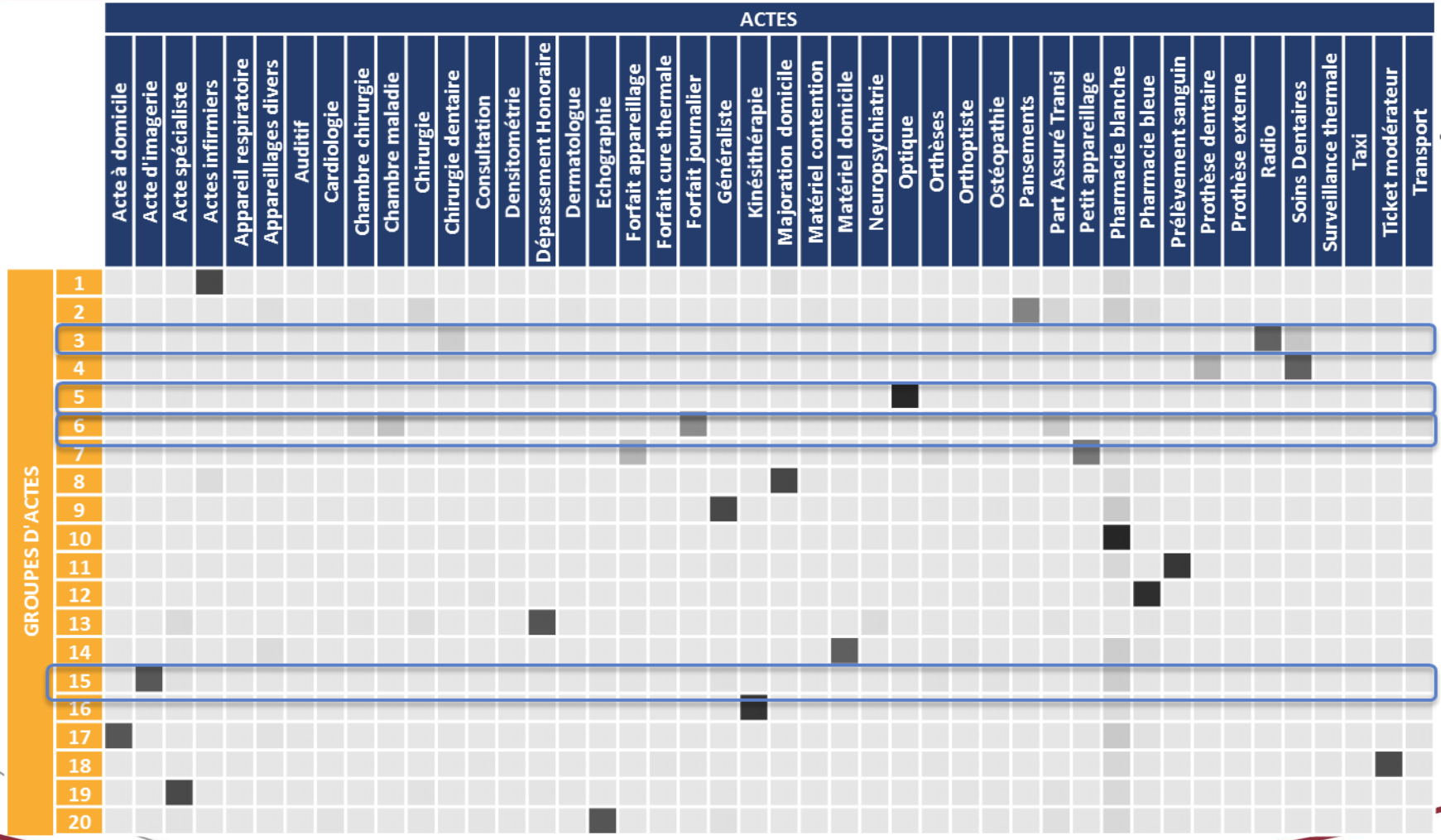
Matrice assurés X **Matrice variables groupe d'actes** = **Base de données de prestations santé en fréquence**



REDUCTION DE DIMENSION | INTERPRETATION | DATA

INSTITUT DES ACTUAIRES

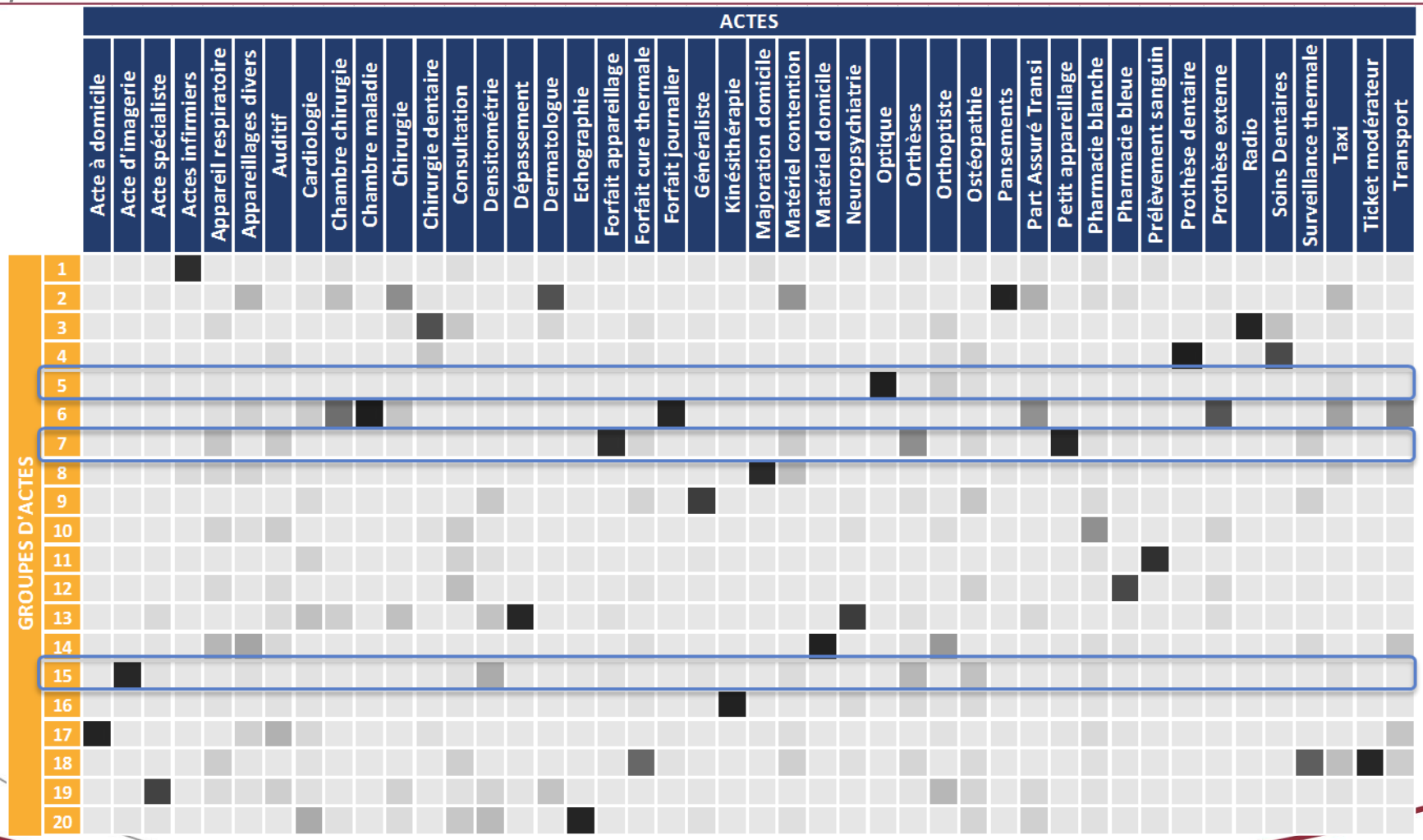
Matrice groupe d'actes – Normalisation horizontale



4, rue Chauveau-Lagarde - 75008 Paris - tél. 01 44 51 72 72 - fax 01 44 51 72 73
www.institutdesactuaires.com

INSTITUT DES ACTUAIRES

Matrice groupe d'actes – Normalisation verticale

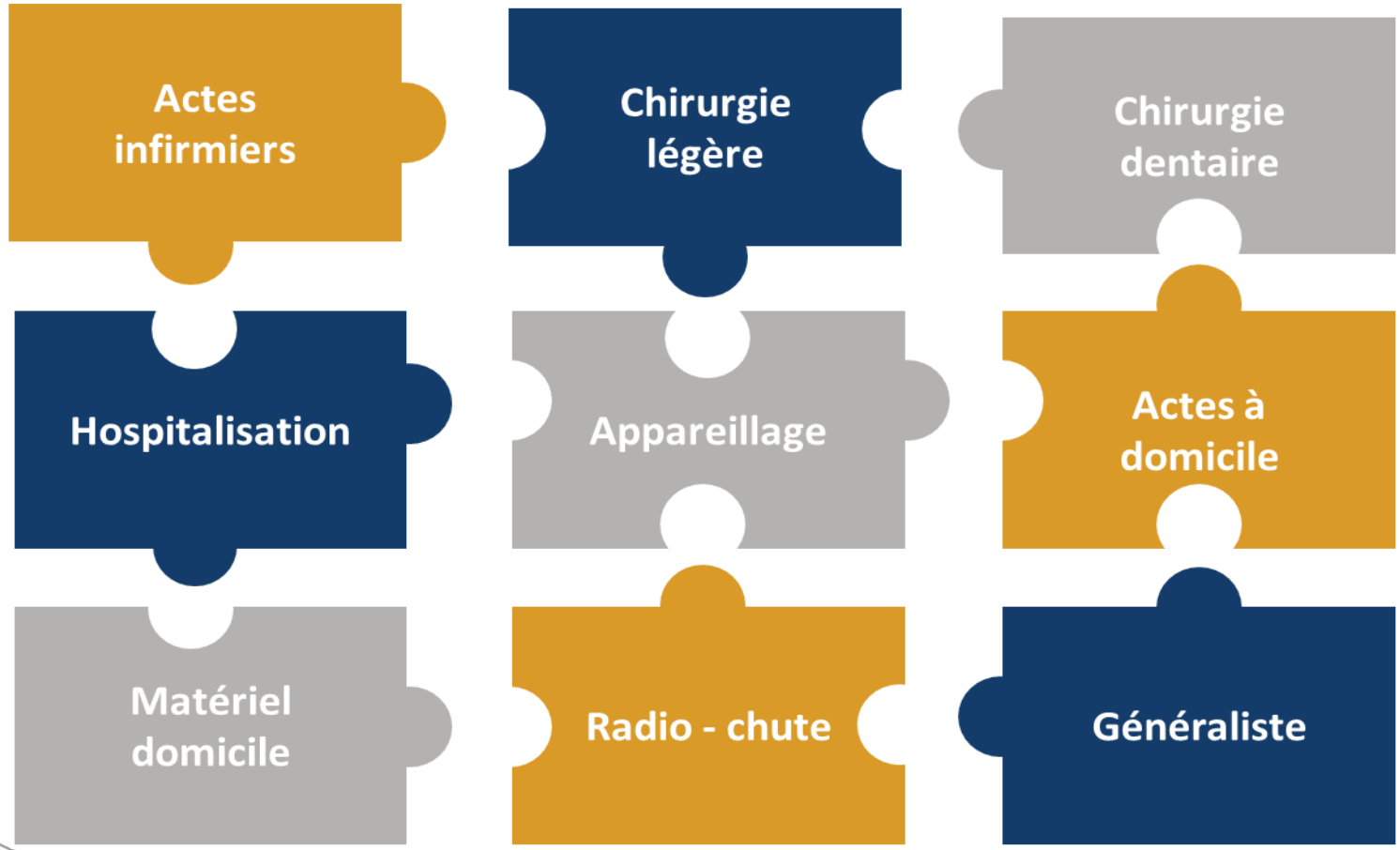


4, rue Chauveau-Lagarde - 75008 Paris - tél. 01 44 51 72 72 - fax 01 44 51 72 73
www.institutdesactuaires.com

INSTITUT DES ACTUAIRES

Interprétation des groupes d'actes – Comportement de consommation

Réduction de la dimension : 80 actes → 20 groupes d'actes



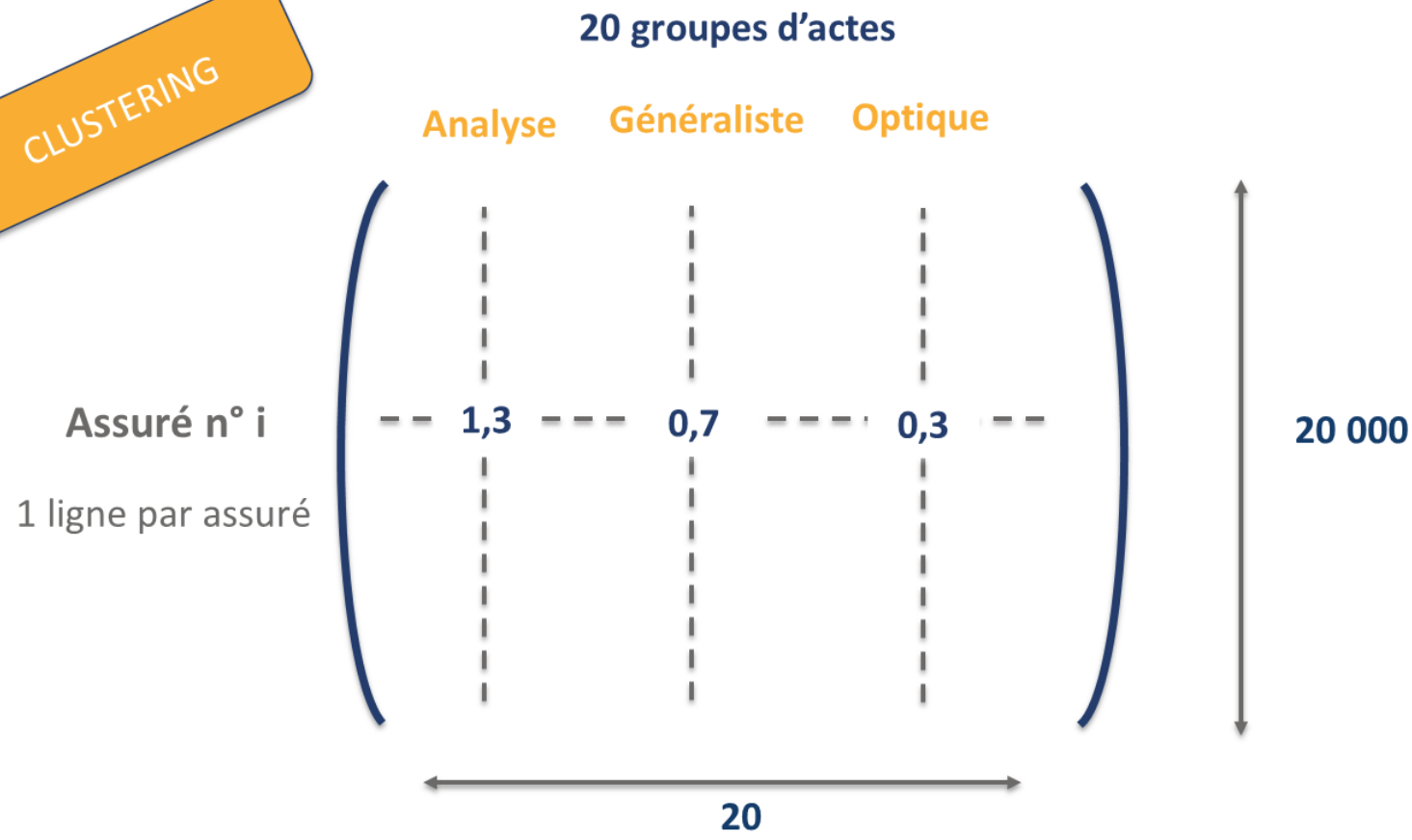
Interprétation des groupes d'actes selon les tranches d'âges

	FEMMES	
EXEMPLES	**45-61 ans**	**Plus de 62 ans**
Groupes d'actes communs	Pharmacie Optique Dentaire Analyse	
Groupes d'actes spécifiques	Orthoptie Psychiatrie Ostéopathie	Soins à domicile Matériel domicile Hospitalisation

INSTITUT DES ACTUAIRES

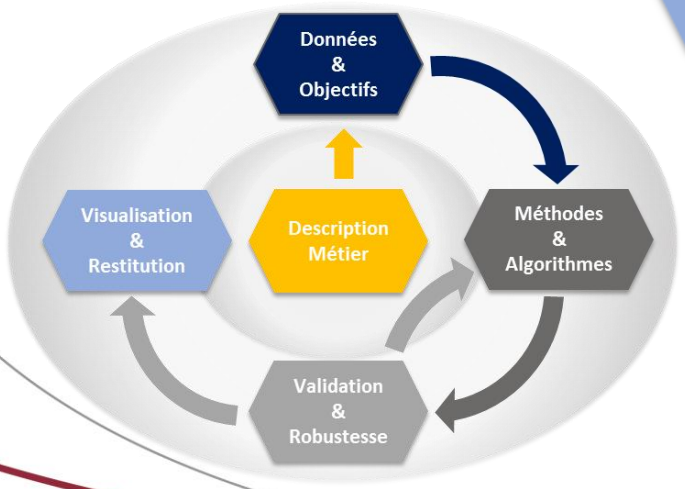
Matrice assurés consommant ces groupes d'actes

CLUSTERING



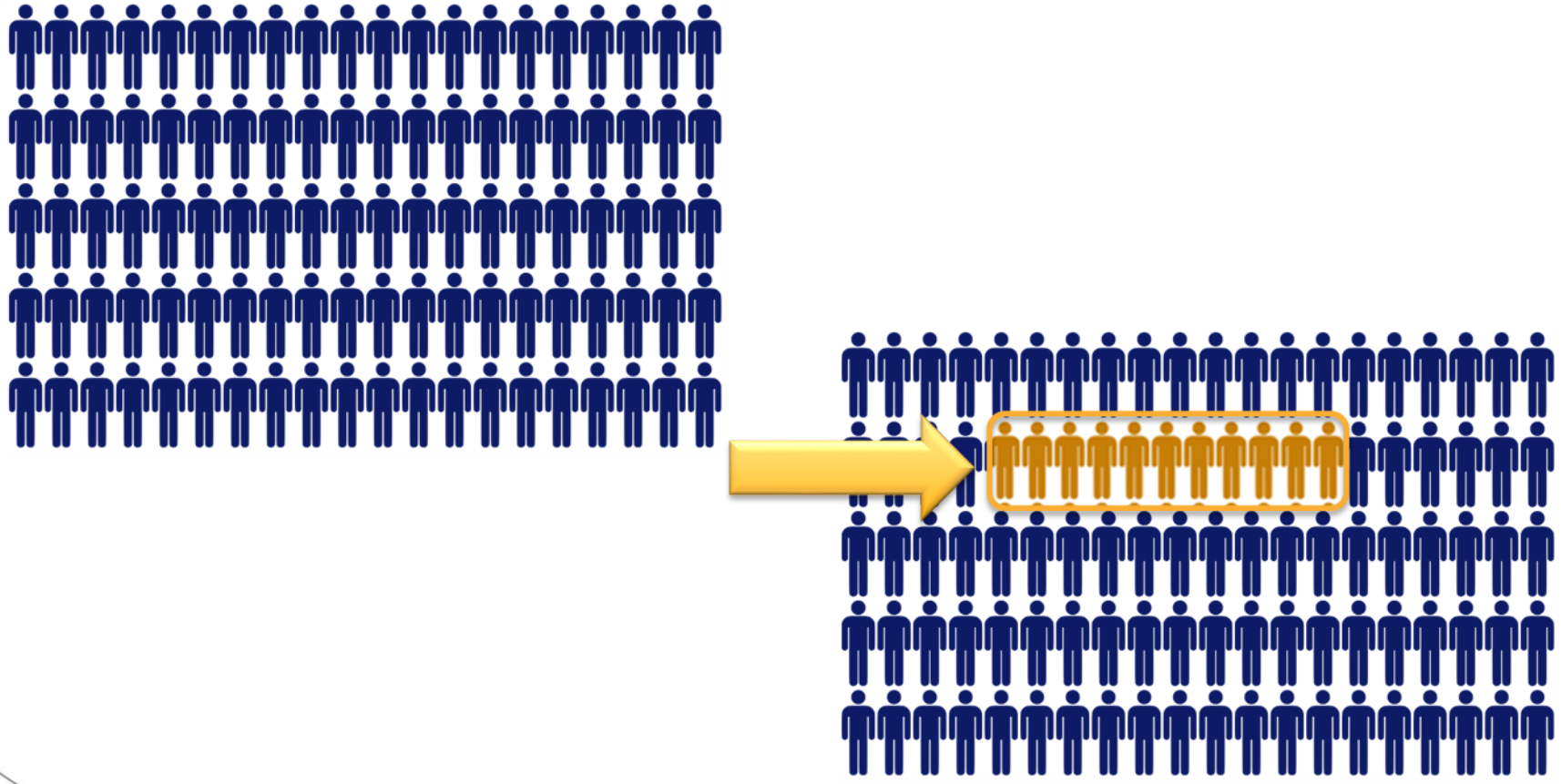
- La dimension réduite permet de classifier nos assurés

VISUALISATION ET RESTITUTIONS



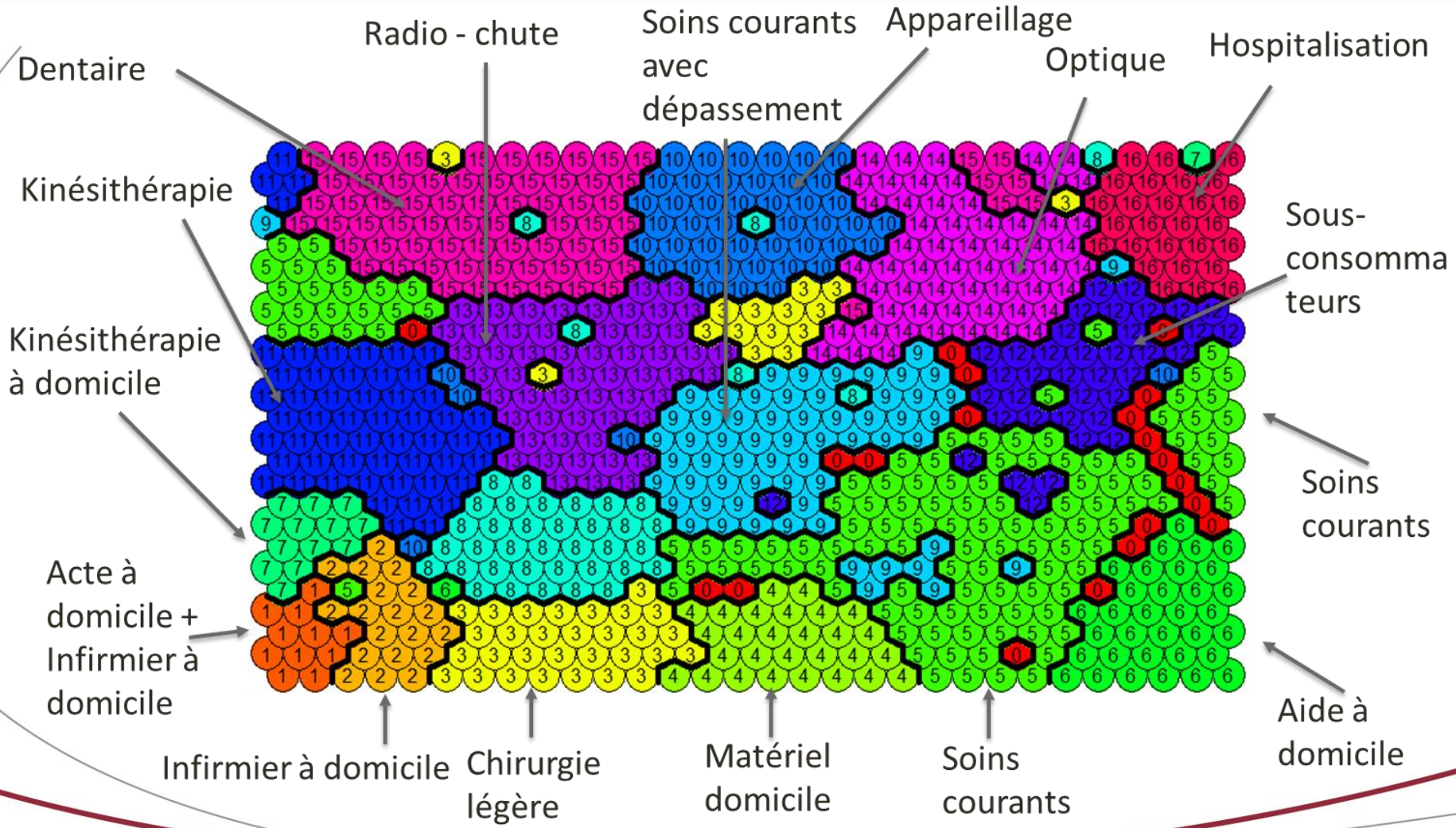
INSTITUT DES ACTUAIRES

Le ciblage de segments d'assurés



INSTITUT DES ACTUAIRES

Quels segments ?



INSTITUT DES ACTUAIRES

Identification d'un parcours de fragilité

69-70 ans

Optique

Dentaire

Sous-consommateur

Soins courants avec ou sans dépassement

Radio – chute

73-74 ans

Appareillage

Chirurgie légère

Infirmier à domicile

77-82 ans

Hospitalisation

Acte à domicile

Acte à domicile + infirmier à domicile

Kinésithérapie à domicile

INSTITUT DES ACTUAIRES

Quel programme de prévention ciblée ?

70 ans

✓ **Activités communes** au senior pour éviter l'isolement
✓ **Visite d'étudiant en psychologie** (signe de dépression)
✓…

74 ans

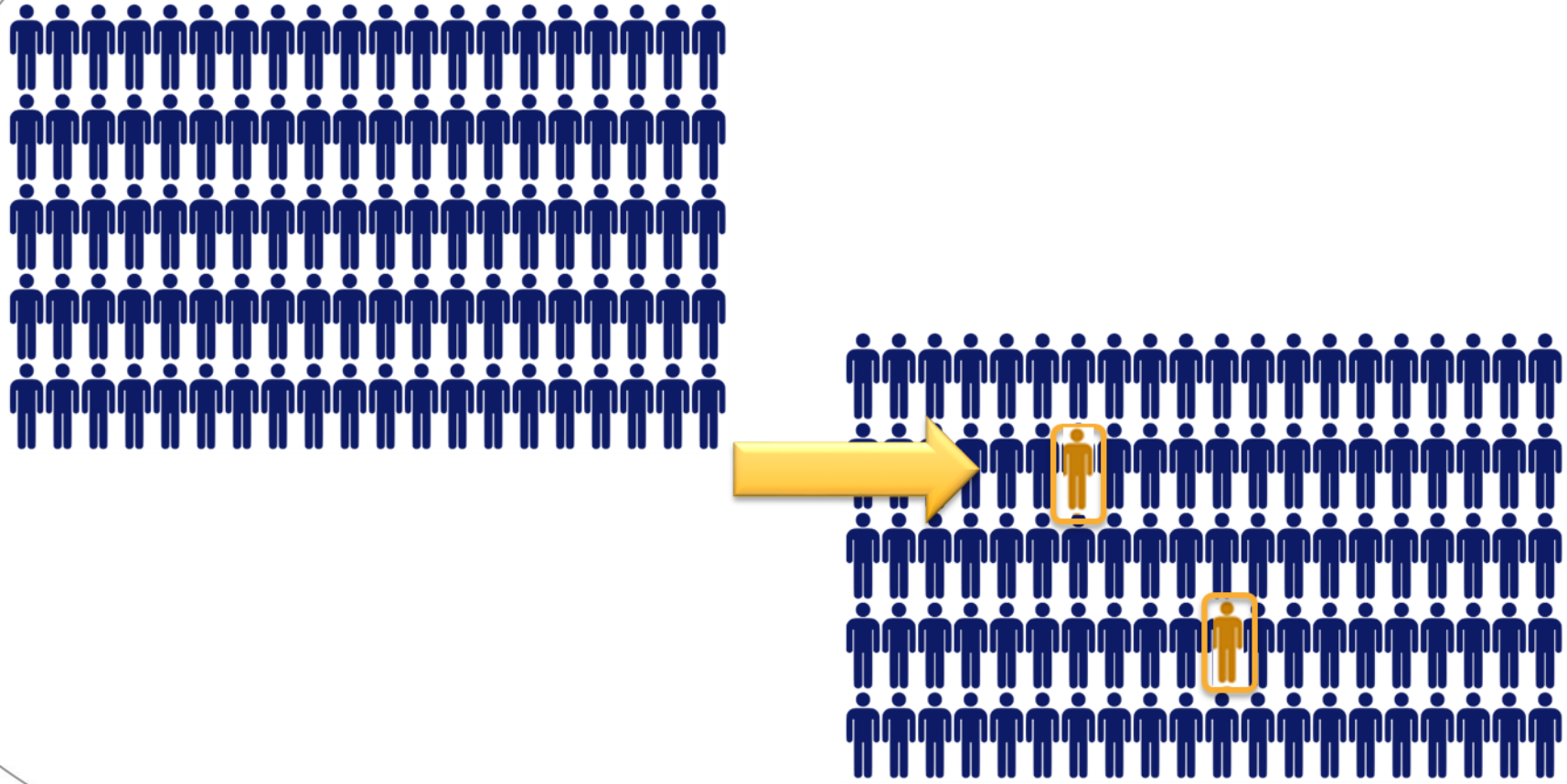
✓ **Sport adapté** pour éviter l'entrée en dépendance lourde
✓ **Formation de la famille** à la détection des signes de dépendance
✓…

80 ans

✓ **Télémédecine** pour favoriser le suivi à domicile
✓ **Objets connectés** pour favoriser le suivi quotidien des personnes dépendantes et alerter si accident ou anomalie mesurée
✓…

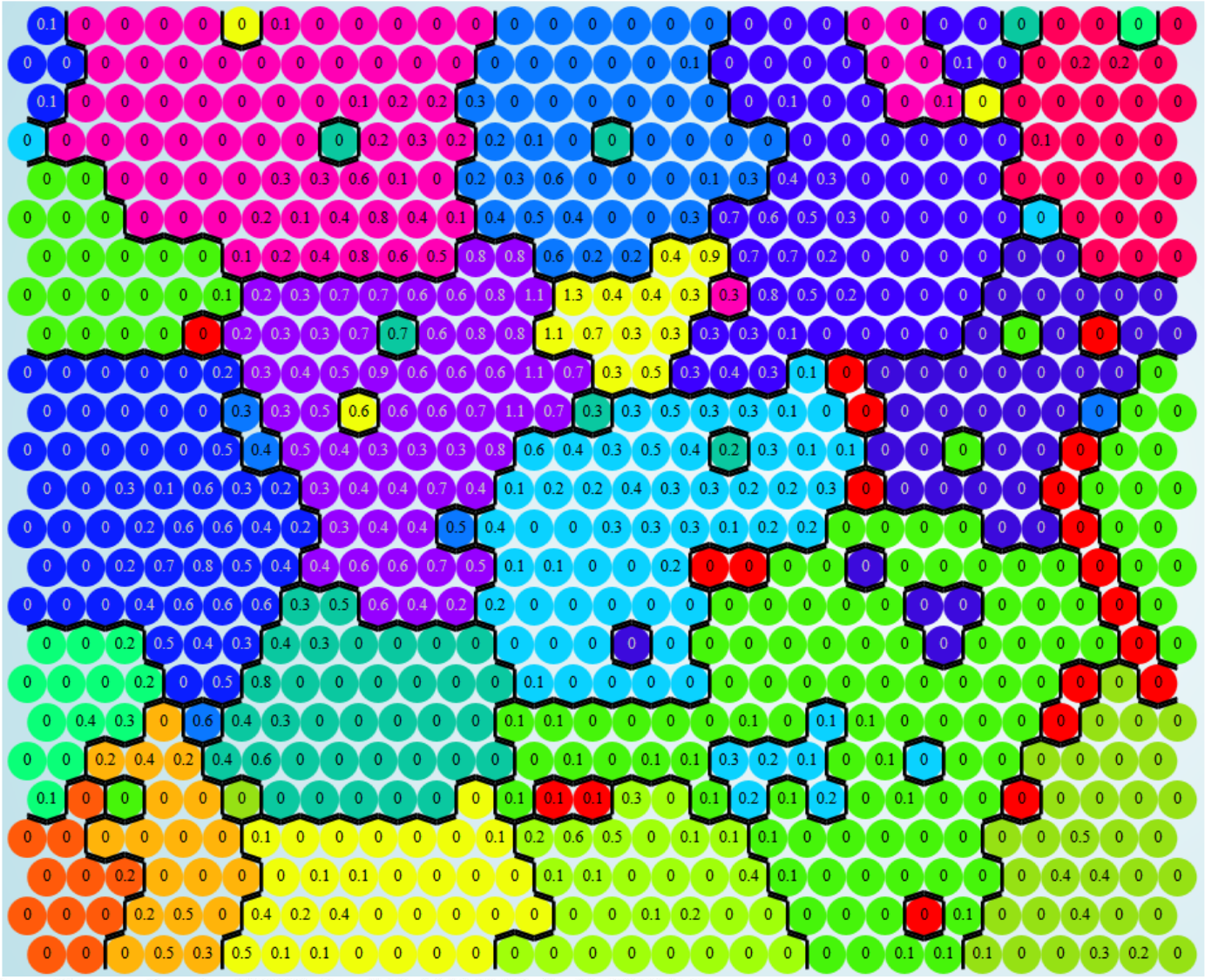
INSTITUT DES ACTUAIRES

Le ciblage d'assurés



INSTITUT DES ACTUAIRES

Quel besoin de prévention pour un assuré ?



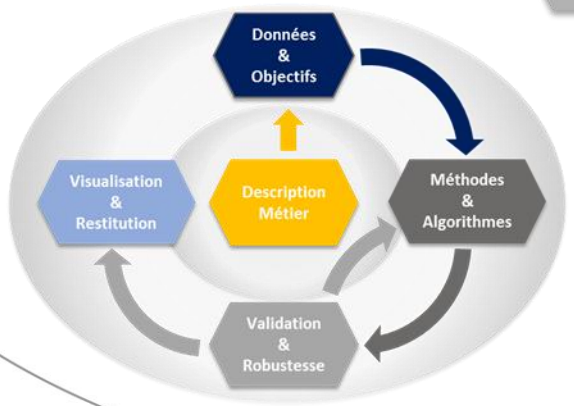
Probabilité d'appartenance à une classe de risque

L'assuré n°96 appartient à la classe :

- « Radio-chute » avec **30%** de chances
- à la classe « kinésithérapeute » avec **11%** de chances
- à la classe « Chirurgie légère » avec **10%** de chance

Les autres classes n'étant pas significatives

VALIDATION ET ROBUSTESSE



INSTITUT DES ACTUAIRES

Comment valider notre méthode de classification non supervisée ?

Robustesse

- Classe fictive d'assurés
- Fort consommateur
- Autre base de données prestation

Validation

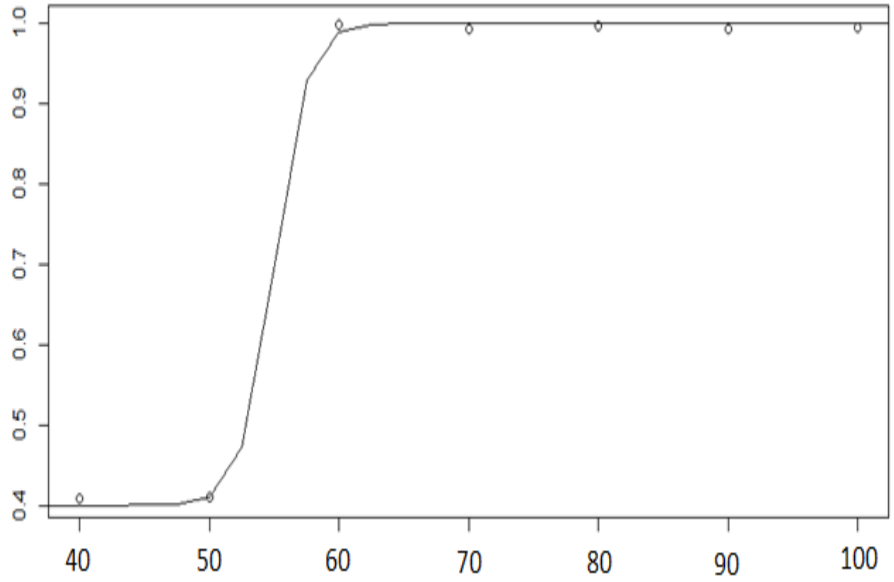
- Text mining
- Comparaison méthodes
- Classification seule

Classe fictive d'assurés

- **Tester la robustesse**
  - ✓ Création d'une classe fictive d'assurés
  - ✓ A partir de combien d'assurés fictifs cette classe est-elle détectée ?

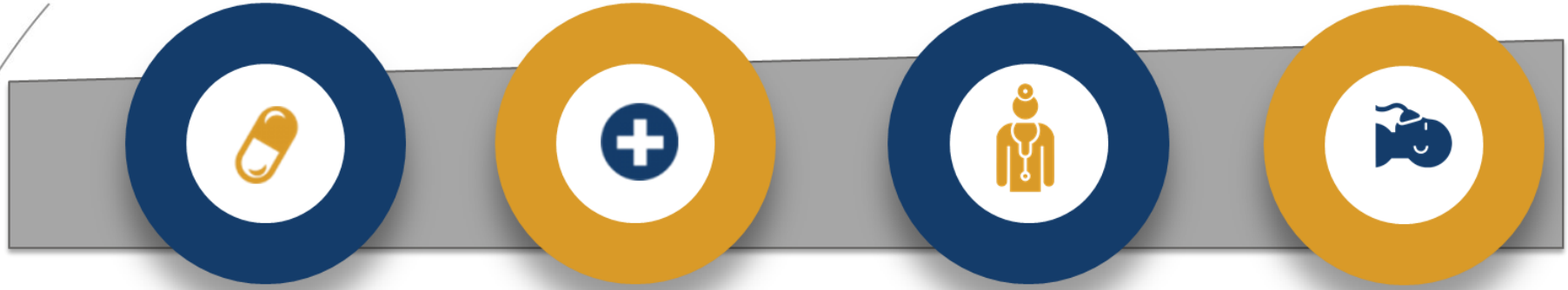
- Exemple :

  ✓ Création de corrélation entre libellés



INSTITUT DES ACTUAIRES

Quels segments de prévention pour les forts consommateurs ?



48 ans

Psychiatrie et dépassement d'honoraires

54 ans

Soins dentaires

Optique

Echographie

64 ans

Matériel domicile

Chirurgie légère et spécialiste

Pharmacie

71 ans

Hospitalisation

INSTITUT DES ACTUAIRES

Autre base de prestations santé

	FEMMES	
EXEMPLES	**Base 1**	**Base 2**
Groupes d'actes communs	Acte à domicile Chirurgie Analyse	
Différences dues à la construction de la base	Ticket modérateur Radio / Chute	Actes biologie Kiné orthopédique Radio
Autres différences	Appareillage	Neuropsychiatrie

- **Tester sur un jeu de données déjà classifié**
  - ✓ Ancien jeu de données classifié
  - ✓ Jeu de données issu du text mining
  - ✓ Permet de comparer les techniques entre elles

Mail 1	La data science, c'est quoi ?
Mail 2	C'est la science des données. Elle regroupe la collecte et l'exploitation de la donnée afin mieux comprendre la source dont elles proviennent.
Mail 3	Donc si j'ai bien compris la data-science, c'est apprendre à connaitre le monde grâce à l'informatique !

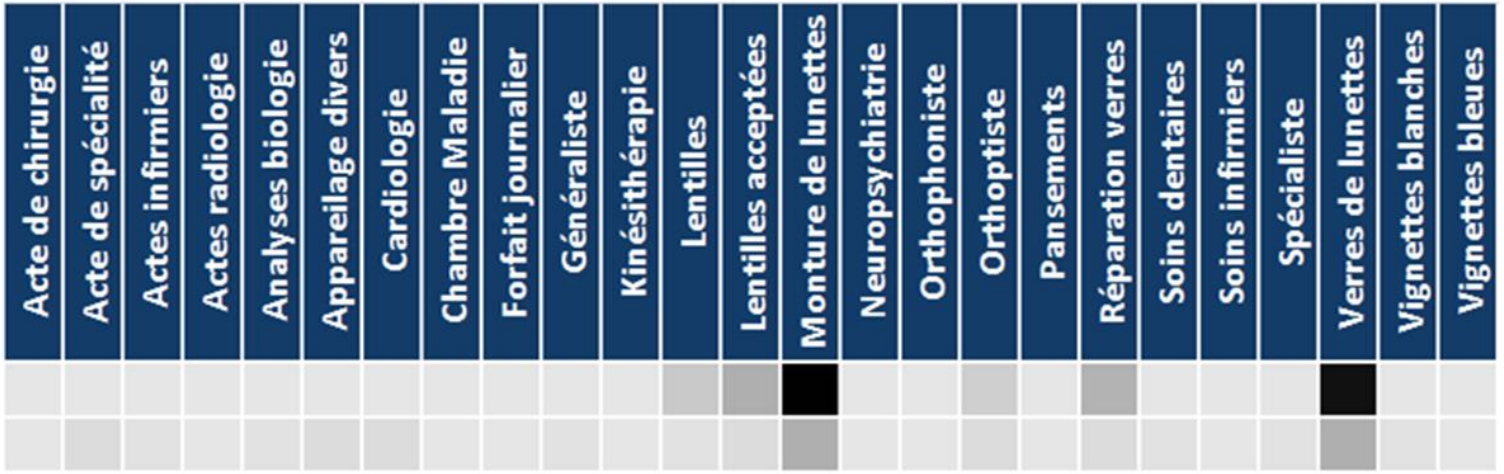
	Data-science	Comprendre	Donnée	la
Mail 1	1	0	0	1
Mail 2	0	1	2	4
Mail 3	1	1	0	1

INSTITUT DES ACTUAIRES

Comparaison avec ACP

- **Comparer différentes méthodes entre elles**
  - ✓ Avec un jeu de données externe
  - ✓ Avec notre jeu de données
  - ✓ Quelle méthode donne les classes les plus cohérentes ?

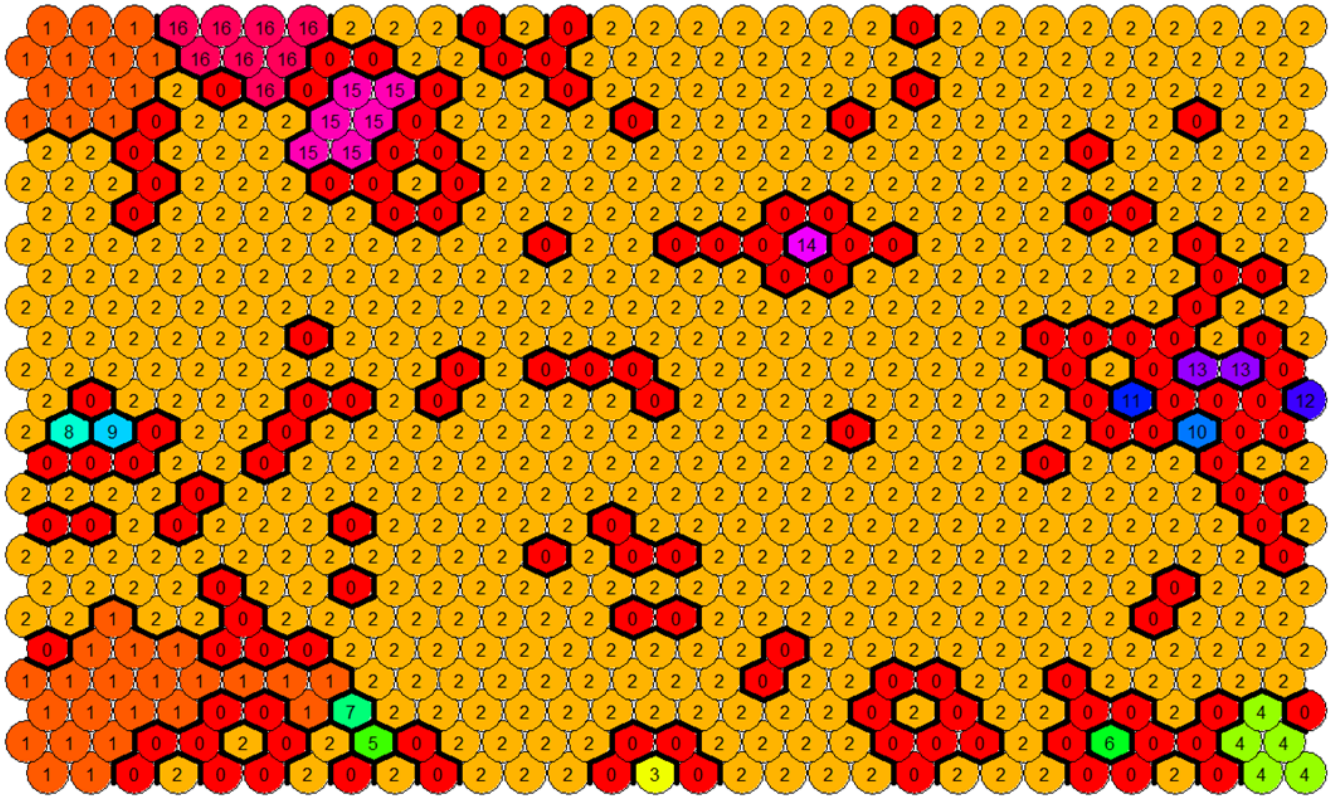
- Exemple : Comparaison de deux méthodes
  - ✓ Quelle méthode regroupe le mieux les actes d'optique ?



INSTITUT DES ACTUAIRES

Suppression de la réduction de dimension

- **Comparer différents process entre eux**
- Pourquoi réduire la dimension ?
  - ✓ Fléau de la dimension



CONCLUSION

INSTITUT DES ACTUAIRES

Le ciblage optimisé d'actions de prévention en 3 étapes

DATA

Assurés et Prestations

MODELISATION

Groupe d'actes consommés

Classes d'assurés

REDUCTION DE DIMENSION

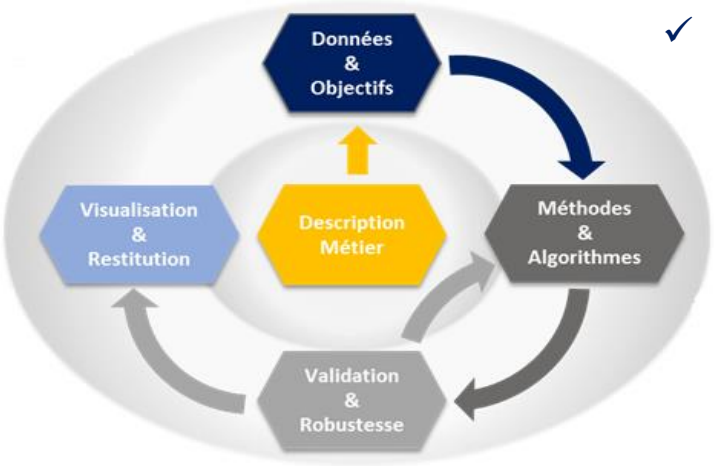
CLUSTERING

INTERPRETATION

Profils de consommation (chute, TMS, cardio-vasculaire…) **Variables caractérisant ces profils**

Prévention

Une méthodologie structurée pour designer les méthodes Data Science



- ✓ La méthodologie permet de **structurer** les projets Data Science
- ✓ Selon un **même process**, quel que soit le domaine d'application
- ✓ Permet une **traçabilité** des hypothèses et des méthodes
- ✓ Capitalise et **mutualise** les développements
- ✓ **Facilite les échanges** avec les experts métier
- ✓ **Favorise la réussite** des projets de Data Science