NUMEA

La statistique : un atout inestimable dans de nombreux domaines

Yan Watts, M.Sc.

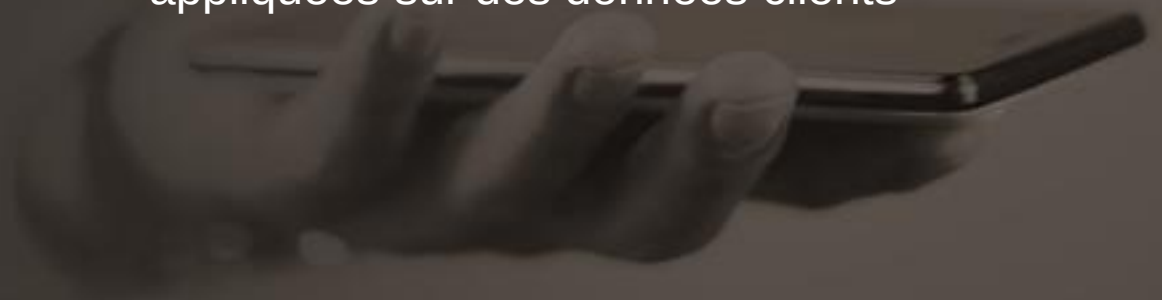
29 janvier 2025

Conférences en statistique pour étudiants du 1er cycle

Université de Montréal | Faculté des arts et des sciences
Département de **mathématiques et de statistique**

Plan

1. Présentation de mon parcours
2. C’est quoi la consultation et qu’est-ce qui m’a attiré chez Numea ?
3. Statistiques dans le domaine agricole : Présentation d’un projet sur les plantes
4. Segmentation client : des méthodes d’apprentissage appliquées sur des données clients



Mon parcours

Baccalauréat en mathématiques

Orientation statistique

- STT2400 : Régression linéaire
- STT2700 : Concepts et méthodes en statistique
- STT3260 : Modèles de survie
- STT3410 : Plans et analyses d’expériences
- STT3510 : Biostatistique
- STT3781 : Laboratoire de statistique
- STT3790 : Apprentissage statistique
- STT6115 : Théorie de la décision bayésienne

Maitrise en statistique, pourquoi ?

Maitrise en statistique

Option Générale : rédaction d'un mémoire sous la supervision de Prof. Mylène Bédard

- STT3795 : Fondements théorique en science des données
- STT6410 : Analyse de la variance
- STT6415 : Régression
- STT6220 : Méthodes de rééchantillonnage
- STT6516 : Données catégorielles
- STT6531 : Consultation statistique 1
- STT6532 : Consultation statistique 2
- **Sujet de mémoire** : Le lasso linéaire : une méthode pour des données de petites et grandes dimensions en régression linéaire

Maitrise, doctorat ou expérience professionnelle ?

Expériences diverses

Travail en plus des études !

- ❖ Assistant de recherche au CHU Sainte-Justine chez Dr. Philippe Bégin (2020 – 2022)
  - ✓ Rédaction d'un article scientifique publié dans Allergy[1]
  - ✓ Expérience dans le domaine d'allergie et immunologie
  - ✓ Expériences acquises sur plusieurs logiciels comme Stata, R, SPSS, etc.
- ❖ Auxiliaire d'enseignement au DMS : STT1700, STT2400, STT3260
  - ✓ Solidifier mes connaissances dans ces cours
  - ✓ Communication fluide en statistique

[1]Watts Y, Dufresne É, Samaan K, Graham F, Labrosse R, Paradis L, Des Roches A, Poder TG, Bégin P. Mapping the Food Allergy Quality of Life Questionnaire Parent Form onto the Short-Form Six-Dimensions version 2. Allergy. 2022 Jun;77(6):1815-1826

POURQUOI LA CONSULTATION?

DIVERSITÉ DES PROJETS ET DES SECTEURS

La consultation statistique offre l'opportunité de travailler sur des **projets variés**, allant de la santé publique à la finance, en passant par l'ingénierie ou les sciences sociales (**STT6531** et **STT6532)**. Chaque projet présente des **problématiques uniques**, ce qui permet d'acquérir des compétences polyvalentes et d'enrichir son expertise.

RÉSOLUTION DE PROBLÈMES COMPLEXES

Les consultants en statistique sont souvent sollicités pour résoudre des problèmes **complexes** à l'aide de données. **Différents types de logiciel** sont utilisés pour arriver à résoudre les problèmes chez le client.

IMPACT CONCRET

Un consultant a la possibilité d’aider des personnes ou des équipes qui **n’ont pas une expertise en statistiques**. Nous agissons en tant que **guide** pour tirer des conclusions pertinentes à partir des données.

APPRENTISSAGE CONTINU

La consultation statistique exige de **rester à jour** sur les nouvelles méthodes, outils et logiciels, ce qui constitue une excellente opportunité de **développement personnel et professionnel**.

Numea

Pourquoi j’ai décidé de joindre Numea ?

- Contacté par Jean-François (président de Numea) à travers Linkedin

Ce que Numea offre

- Équilibre entre le travail et la vie personnelle (surtout quand on est papa !)
- Horaires flexibles et télétravail
- Deux branches de consultations : service aux scientifiques et service aux affaires

Nos expertises

SERVICES D’AFFAIRES

SERVICES AUX SCIENTIFIQUES

PRODUITS

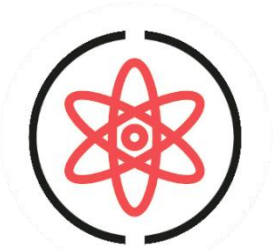
Nos services d’affaires

INTELLIGENCE D’AFFAIRES ET VISUALISATION

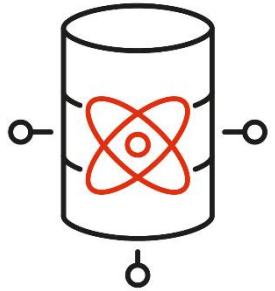
SEGMENTATION ET EXPÉRIENCE CLIENT

PROGRAMME DE FIDÉLISATION

PERFORMANCE MARKETING



INTELLIGENCE ARTIFICIELLE ET SCIENCE DES DONNÉES

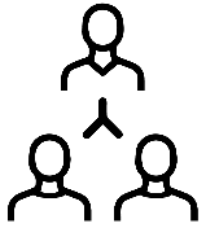


GESTION DE DONNÉES CLIENTS

Nos services aux scientifiques

SOUTIEN STATISTIQUE POUR LES CHERCHEURS

ANALYTIQUE AVANCÉE POUR LE SECTEUR INDUSTRIEL



FORMATIONS

Nos partenaires produits

IBM

ENVIRONICS ANALYTICS

MICROSOFT

Nos clients

Commerce de détail

Services

Médias

Financier

Telécommunications

Transport

Horticulture

OSBL

Gouvernement

Statistiques dans le domaine agricole

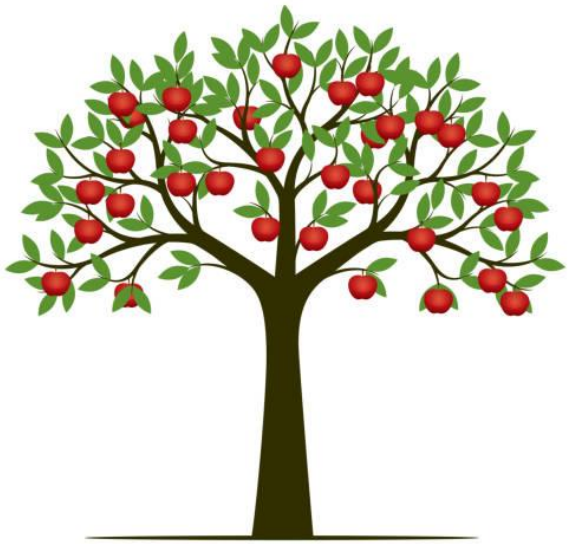
Mise en contexte

Domaine agricole

Nos clients sont des firmes qui commercialisent des produits **agricoles ou horticoles**.

Ces firmes doivent évaluer la **qualité de leurs produits** et développer des produits encore **plus performants**.

L'évaluation de la **qualité** d'un produit se fait par des expériences qui mettent en **compétition quelques traitements** dans un environnement contrôlé. Des protocoles scientifiques sont suivis pour mener ces expériences.



Notre mandat est d'analyser les données qui proviennent de ces expériences.

Mise en contexte

1. Imaginez que vous testez quatre types de fertilisants (A, B, C, D) sur des plants de tomates, et vous mesurez le poids des tomates produites. Vous voulez savoir si le type de fertilisant a un impact significatif sur la production.

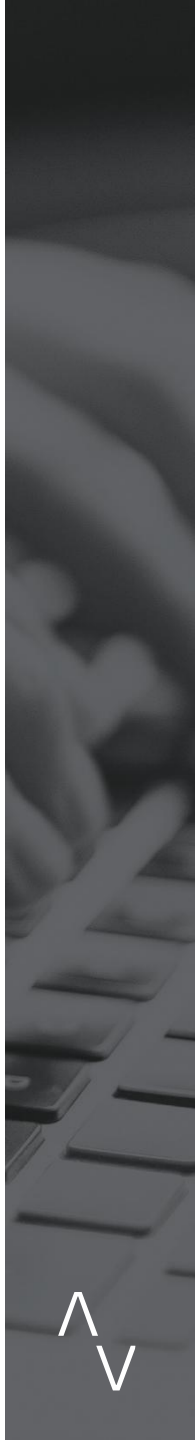
- Les différences observées entre les groupes (fertilisants) sont-elles dues au hasard ou à un véritable effet des fertilisants ?

2. Imaginons que vous ne mesurez pas seulement le poids des tomates à la fin de la saison, mais que vous les mesurez à **plusieurs moments** (par exemple, chaque semaine). Vous voulez savoir :

- Si les fertilisants ont un effet global.
- Si cet effet change au cours du temps (interaction entre fertilisants et temps).

3. Les plants ne poussent pas dans un environnement homogène. Il pourrait y avoir des différences de lumière, de sol ou d'humidité dans votre champ.

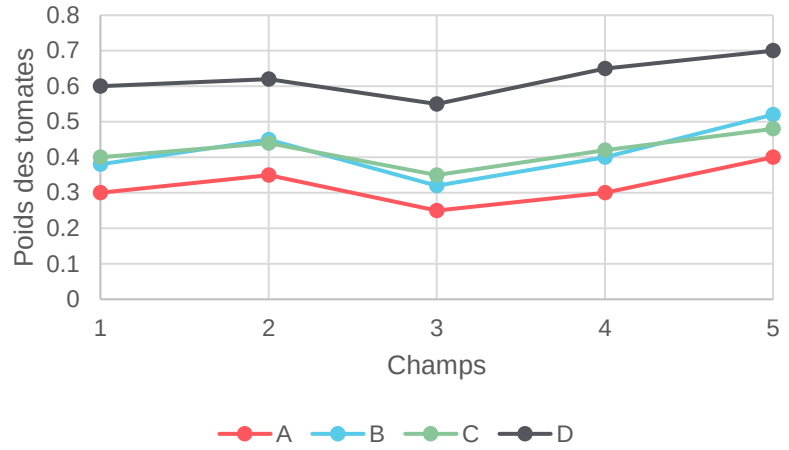
- Pour minimiser ces biais, vous divisez votre champ en **blocs** homogènes et appliquez chaque fertilisant à un sous-groupe (ou bloc) de plants.

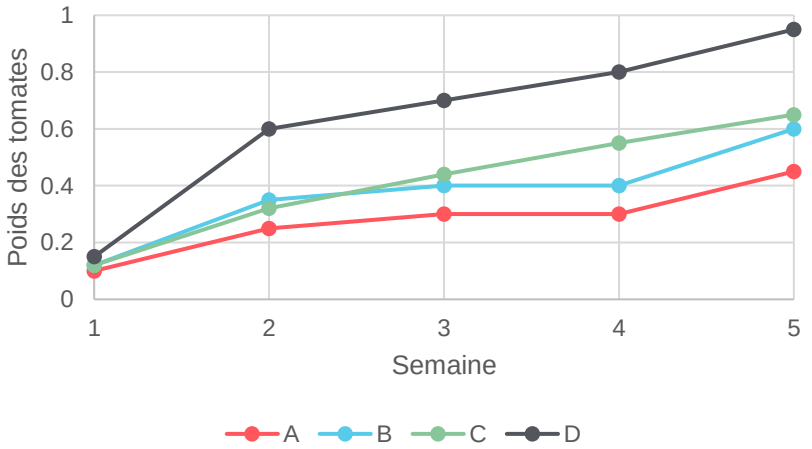


Mise en contexte

Solution pour les 3 situations décrites

1. Comparer fertilisants : Analyse de la variance (ANOVA)
   - ➢ STT3410
2. Comparer fertilisants dans le temps : ANOVA à mesures répétées
   - ➢ STT3781 et STT6531, STT6532
3. ANOVA à blocs randomisés
   - ➢ STT3410 et STT6410

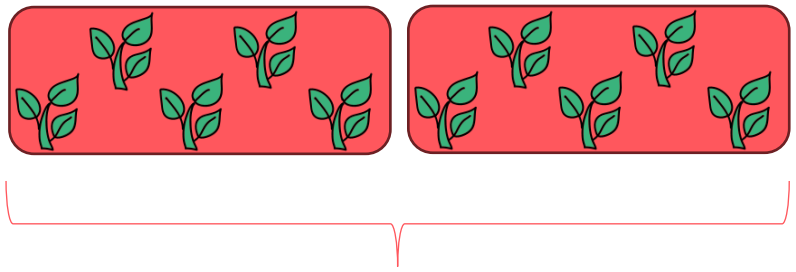




Champ 1	Champ 2	Champ 3	Champ 4	Champ 5
A C	C D	D C	A B	B A
D B	B A	A B	C D	D C

Un exemple sur des plantes

- Pour le projet sur les plantes, nous avions…
  - **7 traitements (T1 à T7)**
  - **4 blocs**
  - **2 pots de fleurs et 5 plantes par pot**
- Pour l'analyse longitudinale, des mesures sont prises tout au long du projet
- Pour les autres analyses, des mesures sont prises soit au début ou à la fin du projet



2 pots de fleurs (5 plantes)

Bloc 1	T2	T1	T4	T3	T6	T5	T7
Bloc 2	T3	T4	T6	T2	T7	T1	T5
Bloc 3	T4	T1	T3	T7	T5	T6	T2
Bloc 4	T1	T2	T6	T5	T4	T7	T3

Analyse : Extraits du sol

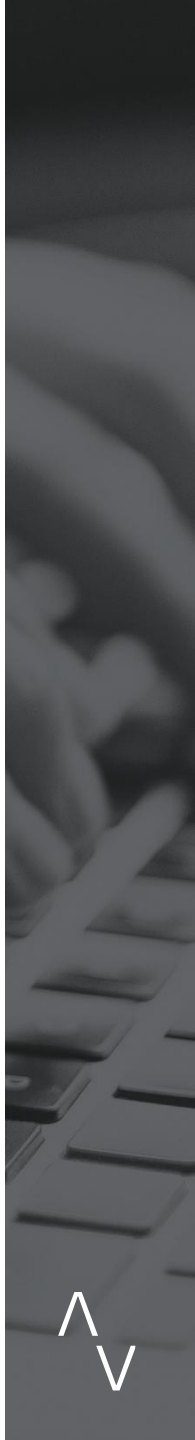
Mesures : Conductivité électrique, pH, calcium, potassium, phosphore, Total

Quand : À la fin du projet

But : Comparer les 7 traitements pour chacune de ces mesures

- Valeur-p < 0.05 → Il existe des différences entre les 7 traitements
- Valeur-p > 0.05 → Il n'y aucune différence entre les 7 traitements

Mesures	Effet	DDL num.	DDL den.	Valeur F	Valeur-p
Conductivité électrique	Traitement	6	16.3	5.65	0.0024
pH	Traitement	6	12.1	43.41	<.0001
Calcium	Traitement	6	13.2	4.13	0.0150
Potassium	Traitement	6	10.6	7.58	0.0024
Phosphore	Traitement	6	14.6	6.89	0.0013
Total	Traitement	6	11.2	4.56	0.0141



Analyse : Extraits du sol (pH)

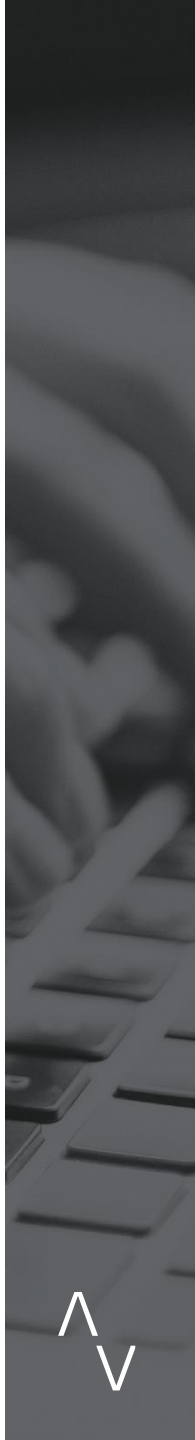
Ce tableau de comparaisons multiples permet au client d'observer…

- Quels traitements ont des différences ?
- Quelle est la valeur de ces différences ?
- Quel est le degré de ces différences (valeur-p très petite ou près de 0.05)

Remarque : Si l'effet du Traitement est **significatif**, ceci ne veut pas dire qu'il y aura une différence entre tous les traitements !

- 3 et 7 n'ont pas une différence significative, car valeur-p = 0.9978 > 0.05

Traitement (I)	Traitement (J)	Différence (I – J)	Valeur-p*	IC inf. *	IC sup. *
1	5	0.6887	<.0001	0.4069	0.9706
1	6	0.4125	0.0002	0.2120	0.6130
2	5	0.7288	0.0001	0.3861	1.0714
2	6	0.4525	0.0015	0.1728	0.7322
3	4	0.3500	0.0012	0.1381	0.5619
3	5	0.8675	<.0001	0.5993	1.1357
3	6	0.5913	<.0001	0.4104	0.7721
3	7	0.02250	0.9978	-0.1267	0.1717
4	5	0.5175	0.0007	0.2246	0.8104
4	6	0.2413	0.0251	0.02544	0.4571
4	7	-0.3275	0.0008	-0.5175	-0.1375
5	6	-0.2762	0.0450	-0.5475	-0.00496
5	7	-0.8450	<.0001	-1.0963	-0.5937
6	7	-0.5687	<.0001	-0.7234	-0.4141



Analyse longitudinale : Nombre de fruits

Mesures : Nombre de feuilles, **nombre de fruits**, nombre de fleurs, nombre de hampes

Quand : À chaque 2 – 3 semaines durant le projet

But : Observer une différence entre les 7 traitements **au fil du temps**

- Valeur-p < 0.05 → Il existe des différences entre les 7 traitements **au fil du temps**
- Valeur-p > 0.05 → Il n'y aucune différence entre les 7 traitements **au fil du temps**

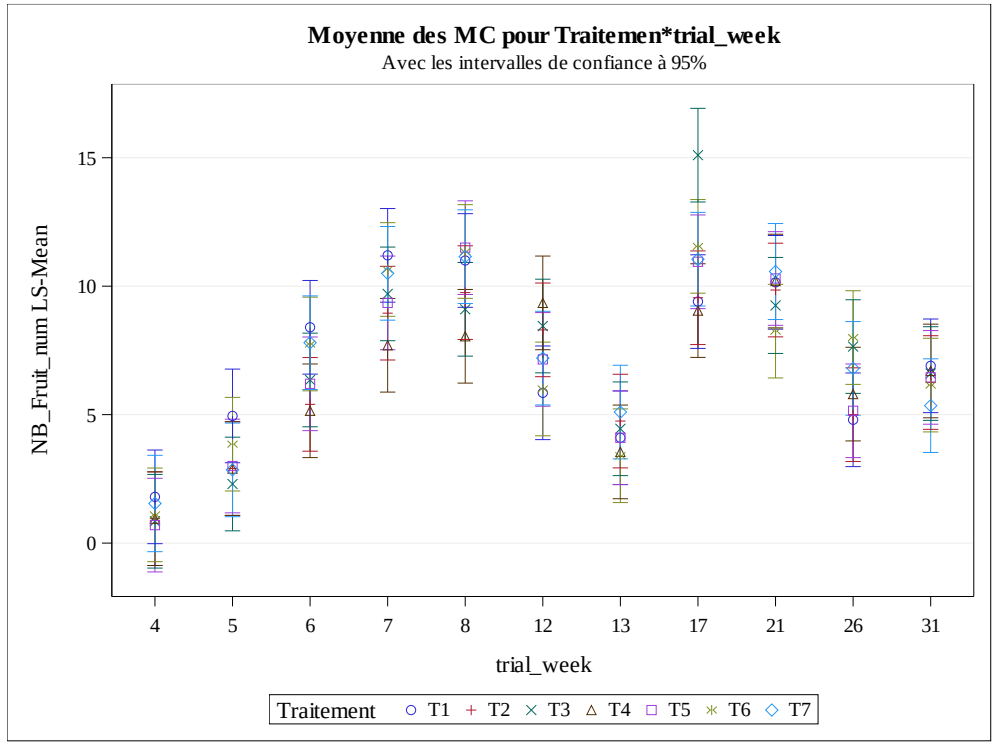
Remarque : Contrairement à l'analyse précédente, ici nous regardons la valeur-p de l'interaction entre l'effet Traitement et l'effet Semaine !

Mesure	Effet	DDL num.	DDL den.	Valeur F	Valeur-p
Nombre de fruits	Traitement	6	124	1.32	0.2521
	Trial week	10	346	81.32	<.0001
	Traitement* Trial week	60	792	1.53	0.0076

Analyse longitudinale : Nombre de fruits

Une analyse longitudinale est beaucoup plus intéressante, puisque nous avons 1537 données.

On peut confirmer avec le client qu'il y a effectivement deux récoltes de fruits durant ce projet, car il y a **2 pics** !



Trial week	Traitement (I)	Traitement (J)	Différence (I - J)	Valeur-p*	IC inf.*	IC sup.*
17	1	3	-5.70	0.0003	-9.5796	-1.8204
17	2	3	-5.55	0.0005	-9.4296	-1.6704
17	3	4	6.05	0.0001	2.1704	9.9296
17	3	5	4.15	0.027	0.2704	8.0296
17	3	7	4.05	0.0341	0.1704	7.9296

La segmentation

Segmentation

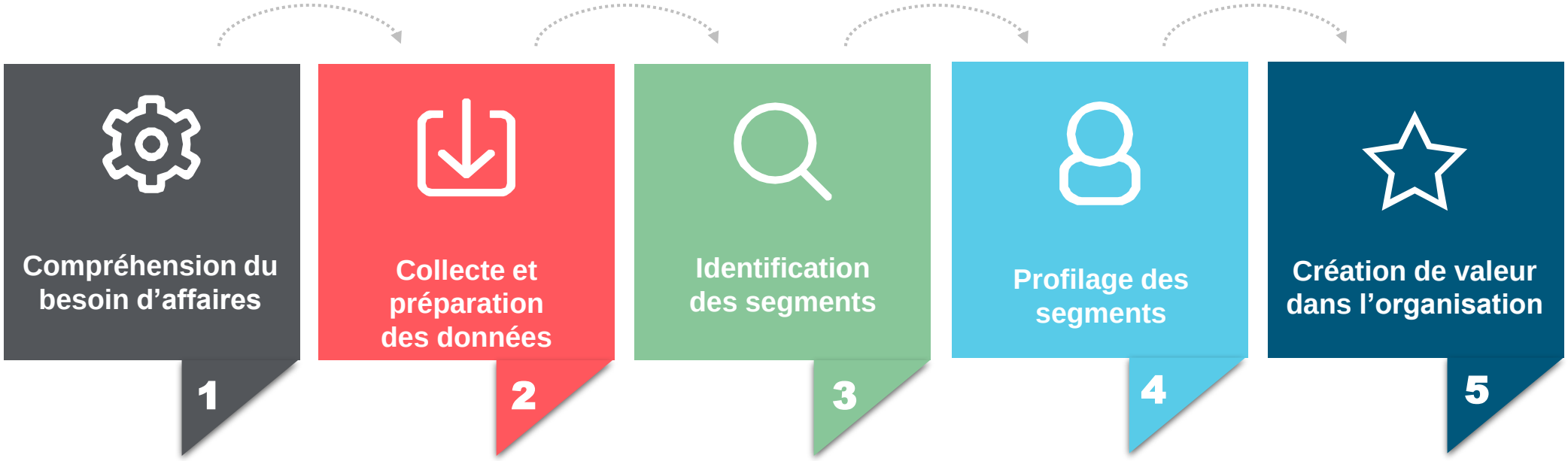
Le fait de diviser une population en sous-ensembles homogènes selon différents critères.

Segment

Regroupement d'individus qui partagent des caractéristiques communes.

Notre approche

Étapes pour une segmentation réussie



Étapes pour une segmentation réussie

1. **Compréhension du besoin d'affaires, des attentes et de l'utilisation potentielle**
2. **Collecte et préparation des données**
3. **Identification des segments**
4. **Profilage des segments**
5. **Création de valeur dans l'organisation**

L'étape la plus importante, souvent oubliée !

Débutons par le "pourquoi"?

Cette phase inclut :

- Une compréhension du contexte, du besoin d'affaires et les attentes de la segmentation
- Atelier avec les parties prenantes afin de livrer une segmentation utilisable
- Identifier la population à segmenter et les exclusions à appliquer
- Identifier et approuver les concepts jugés pertinents pour définir la segmentation

Étapes pour une segmentation réussie

1. Compréhension du besoin d'affaires, des attentes et de l'utilisation potentielle
2. **Collecte et préparation des données**
3. Identification des segments
4. Profilage des segments
5. Création de valeur dans l'organisation

Une bonne segmentation repose sur des données riches

1. Répertorier les sources de données pouvant être intégrées

2. Préparation et enrichissement des données pour l'analyse

a. Sélectionner et créer les variables

b. Nettoyer et valider les données manquantes

c. Transformer et enrichir les données en créant des indicateurs

Par contre, les données de sondage sont souvent catégorielles...

Étapes pour une segmentation réussie

1. Compréhension du besoin d'affaires
2. Collecte et préparation des données
3. **Identification des segments**
4. Profilage des segments
5. Création de valeur dans l'organisation

Un mélange de science et d'expérience

ACM (Analyse de correspondances multiples)

Pour décrire et comprendre les clients, on peut réaliser une série de tableaux de fréquences et de tableaux croisés. Bien que les statistiques descriptives révèlent de l'information intéressante, des analyses statistiques multivariées comme une ACM pourraient répondre aux questions suivantes :

- Y a-t-il des individus qui ont les mêmes modalités sur un ensemble de variables/caractéristiques?
- Quelles sont les modalités associées aux mêmes individus?

L'analyse des correspondances (ACM) permet d'analyser la relation entre plusieurs **variables catégorielles**.

Les avantages :

- Pas de postulats à respecter quant aux distributions.
- Ne présuppose pas la relation linéaire entre les paires de variables catégorielles.
- Les valeurs extrêmes sont moins problématiques.
- Permet d'inclure les catégories de non-réponse, non disponible, etc.
- Permet d'analyser des tableaux de réponses multiples.

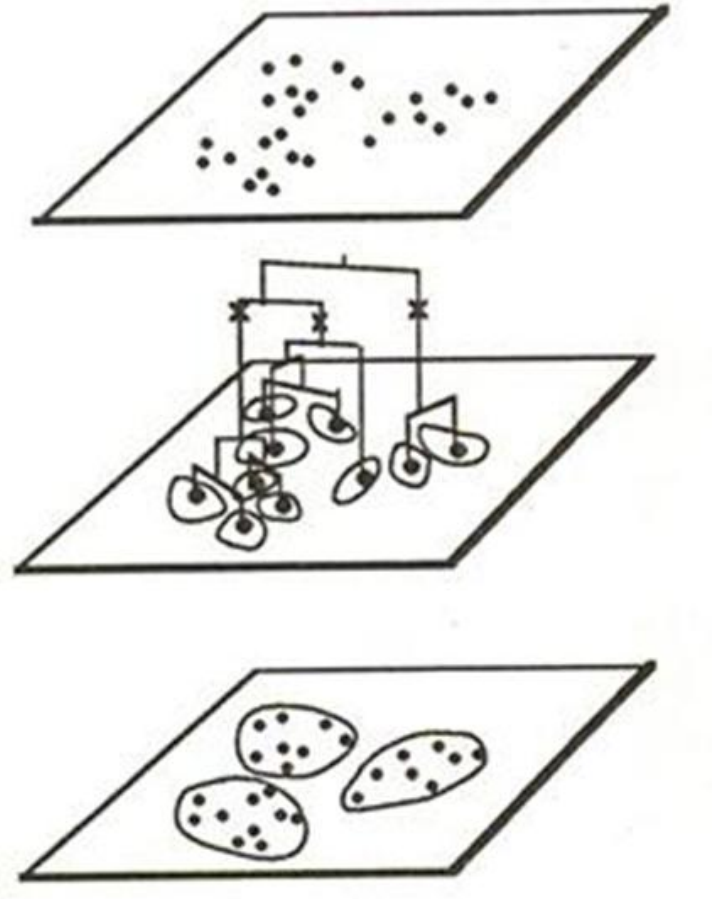
L'ACM est utilisée comme étape préliminaire à l'analyse de « cluster ».

Analyse de « clustering »

On complète l'ACM par une analyse de type « clustering » réalisée sur l'ensemble des facteurs (axes) ou bien sur les plus importants.

Différentes techniques de « clustering » peuvent s'appliquer:

- Méthode hiérarchique de Ward
- Méthodes itératives (K-Means)
- Méthodes mixtes (combine hiérarchique et K-Means)



Cours reliés à ces méthodes : STT3790, STT3795 !

Un exemple

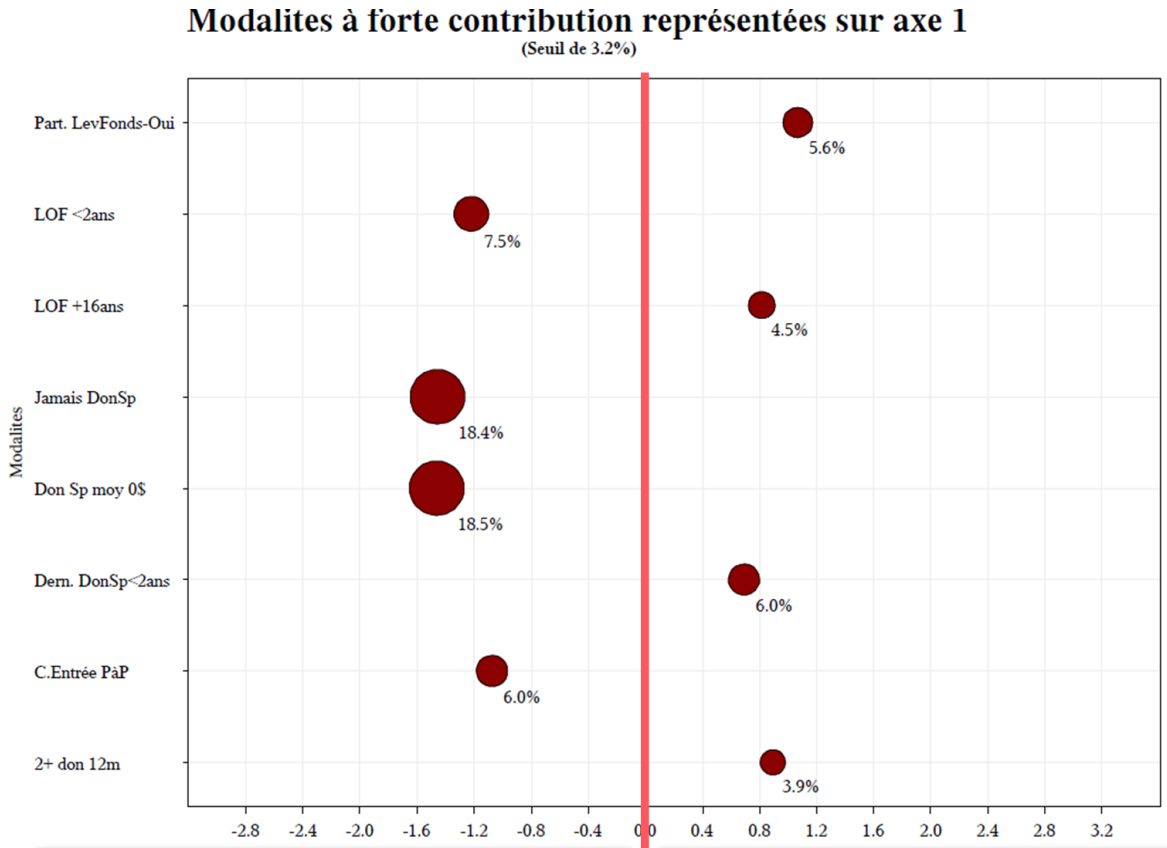
Une fondation cumule de l'information sur ses donateurs. Les donateurs font des dons mensuels et, à l'occasion, peuvent faire un don spontané. Nous avons une base de données qui contient l'information pour plusieurs milliers de donateurs.

Variables à l'étude	Modalités
Nb d'années depuis 1er don	1) <2 ans 2) 3-4 ans 3) 5-8 ans 4) 9-16 ans 5)>16 ans
Nb d'années depuis dernier don spontané	1) Jamais 2) <2 ans 3) >2 ans
Canal d'entrée	1) Web 2) TV 3) Pub 4) PàP 5) Sport 6) Soirée 7) Autres
Nb de dons spontanés 12 derniers mois	1) Jamais 2) 1 don 3) 2+ dons
Montant moyen par don spontané	1) 0$ 2) 1-29$ 3) 30-49 4) 50-99 5) 100$+
Commentaire au centre d'appels	1) Positif 2) Aucun commentaire
Aimerait participer à un événement de levée de fond	1) Oui 2) Non
Niveau d'engagement	1) Élevé 2) Moyen 3) Faible

Comment interpréter les dimensions retenues

- Chaque dimension permet de **regrouper** des modalités partagées par les mêmes individus et de **mettre en opposition** des modalités associées à différents individus.

- Le graphique à bulles permet **d'illustrer** les modalités qui contribuent davantage à un axe.

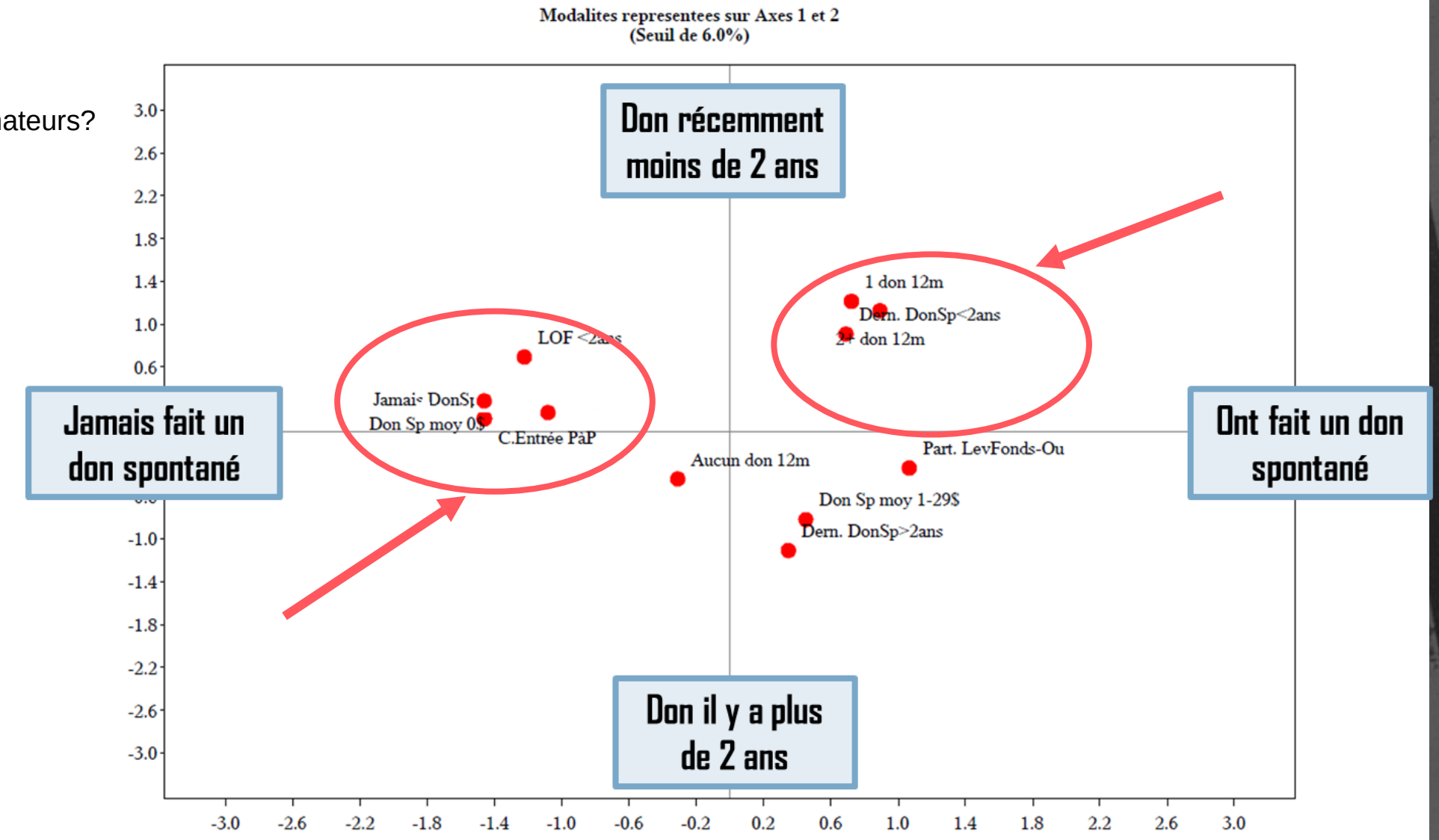


Récence : Jamais donné de dons spontanés
Montant don moyen : 0$
Nouveaux donateurs proviennent du porte à porte

Récence : Ont donné il y a moins de 2 ans
Fréquence : 2 dons et +
Anciens donateurs 16 ans +

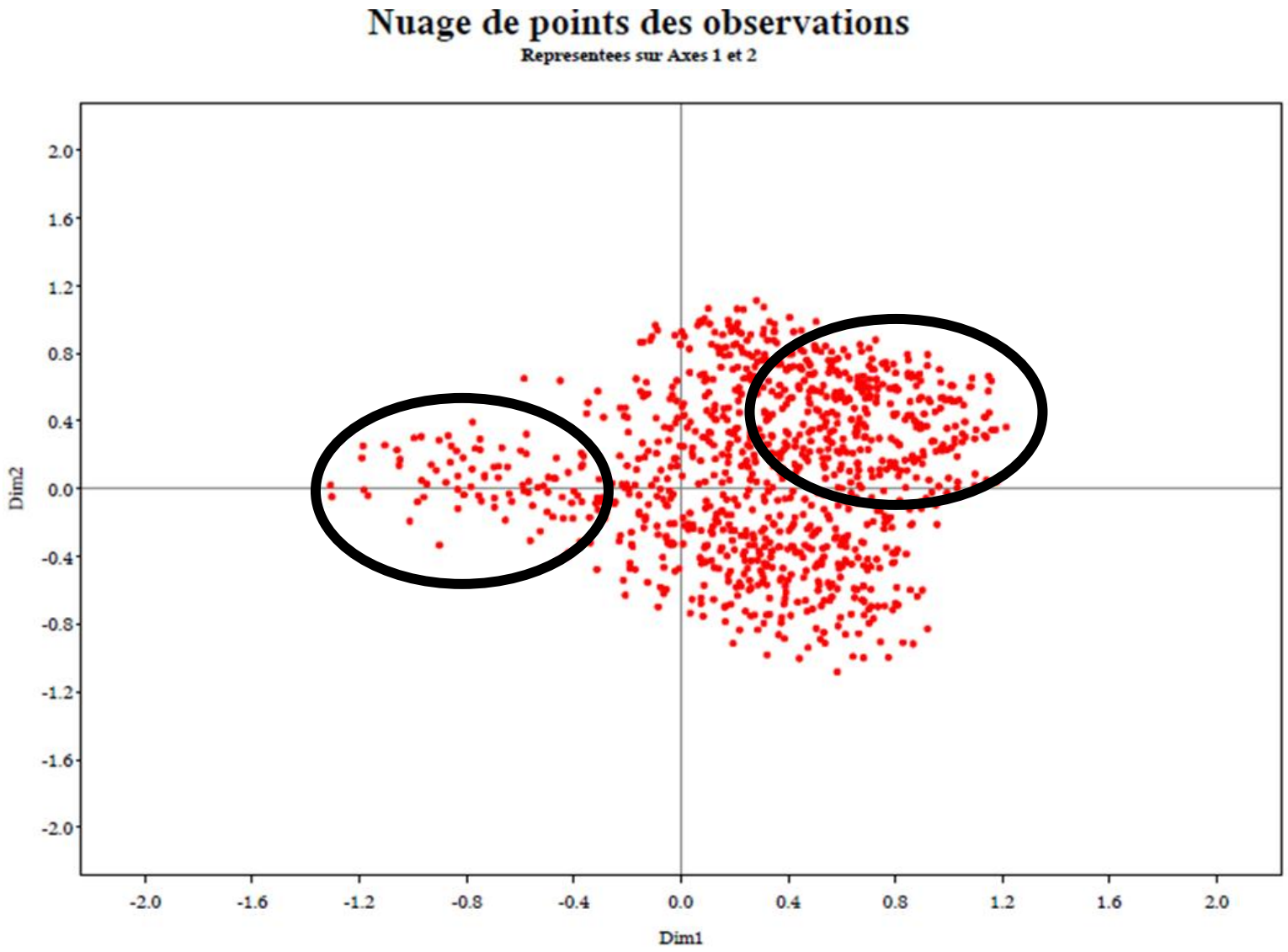
Représentation des modalités sur l'axe 1 et l'axe 2

- Qui sont ces donateurs?



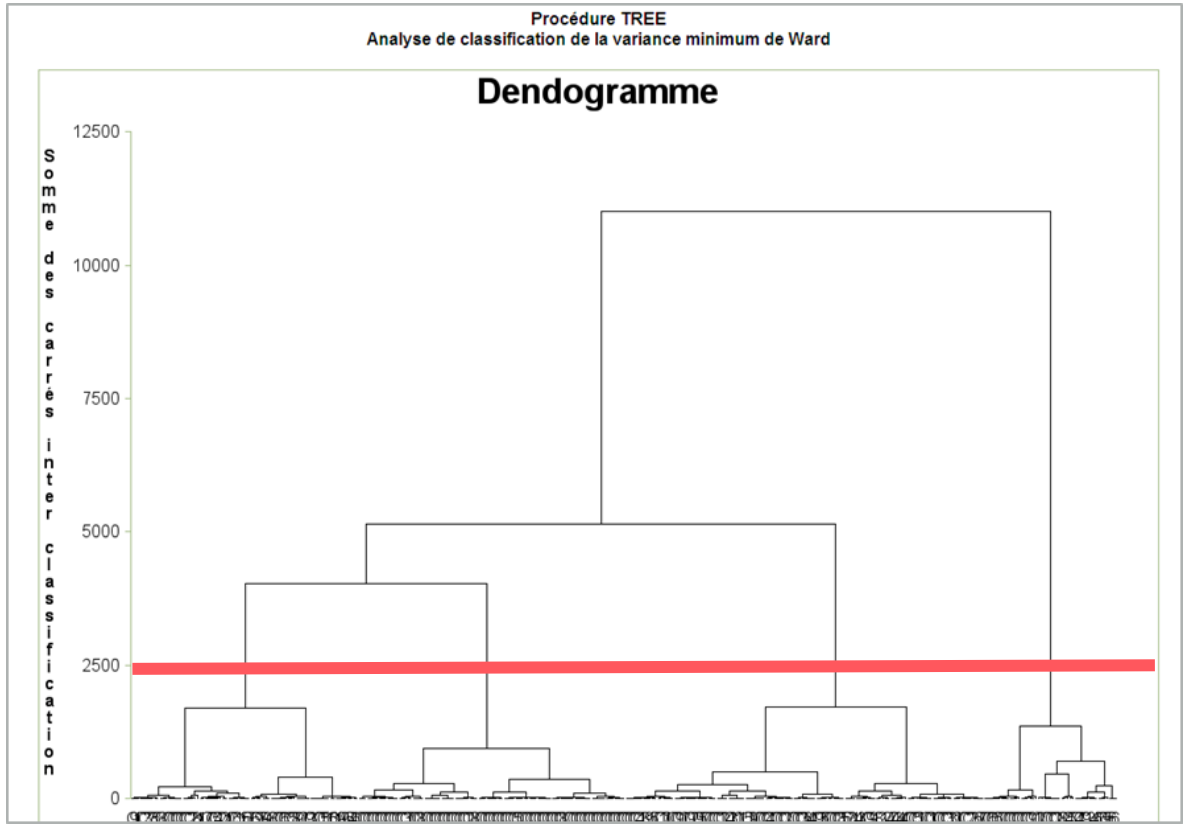
Nuages de points des donateurs

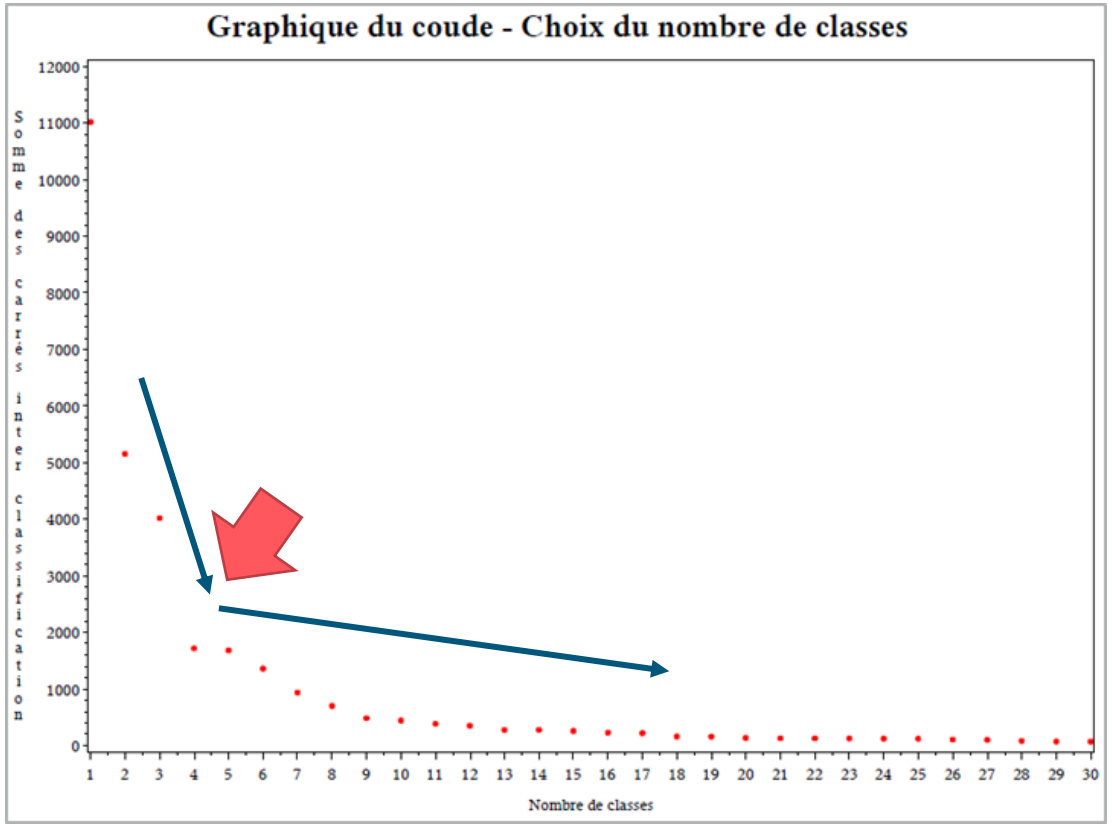
- À 2 dimensions, on peut visualiser les donateurs associés aux mêmes modalités.
- La présentation de l'axe 3 et l'axe 4 permettrait d'illustrer d'autres modalités fortement associées ou fortement éloignées.
- La visualisation de nos milliers d'individus sur nos 4 dimensions devient complexe à analyser.



1- Combien de segments retrouvons-nous? Dendogramme et graphique du coude

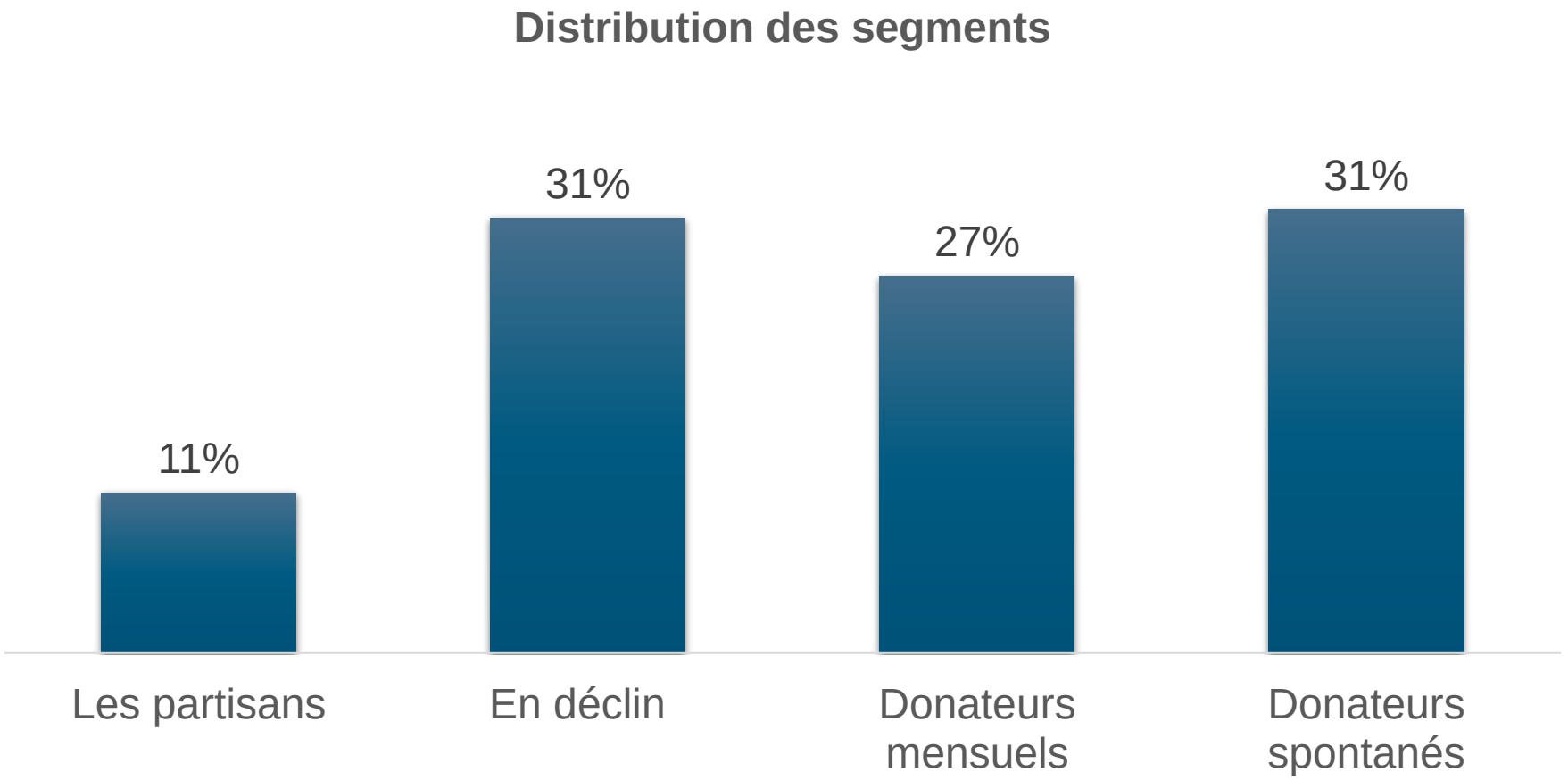
On visualise 4 segments!





Le mélange de science et d'expérience à l'œuvre !

2- Combien y a-t-il de donateurs par segment?



On se rappelle d’avoir au moins 5% des individus par segment

3- Quelles sont les caractéristiques de mes segments?

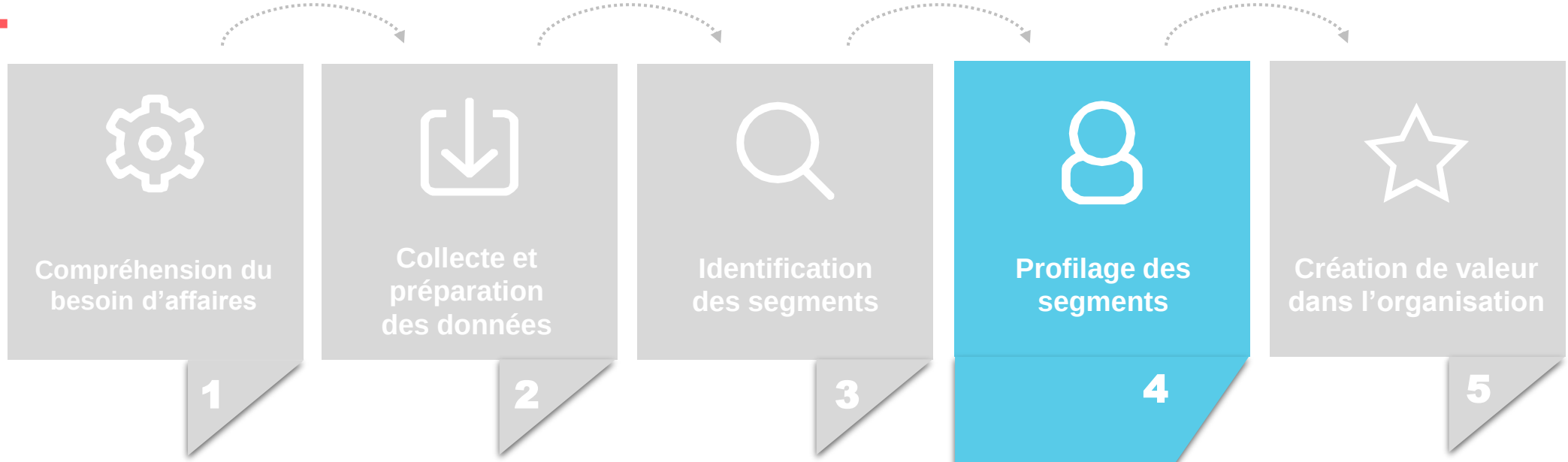
Une fois les donateurs regroupés, on fait un profilage sur toutes les variables. Cela permet de savoir en quoi les donateurs se ressemblent à l'intérieur de chaque segment et se distinguent entre les segments.

Par exemple, le segment 1 qui représente 31% des individus, est composé des donateurs qui n'ont pas fait de dons spontanés dans la dernière année.

Segment 1 : Les donateurs spontanés (31%)

Libellés	Modalités	% de gens avec cette modalité dans le segment (A)	% de gens avec cette modalité dans l'échantillon (B)	Index (A)/(B) x 100
Nb d'années depuis dernier don unique	**Dern. DonSp<2ans**	**99.75**	**38.65**	**258**
Nb de dons spontanés 12 derniers mois	2+ don 12m	43.24	15.18	285
Nb de dons spontanés 12 derniers mois	1 don 12m	34.87	12.29	284
Montant moyen par don spontanée	Don Sp moy 30-49$	33.89	20.67	164

Étapes pour une segmentation réussie

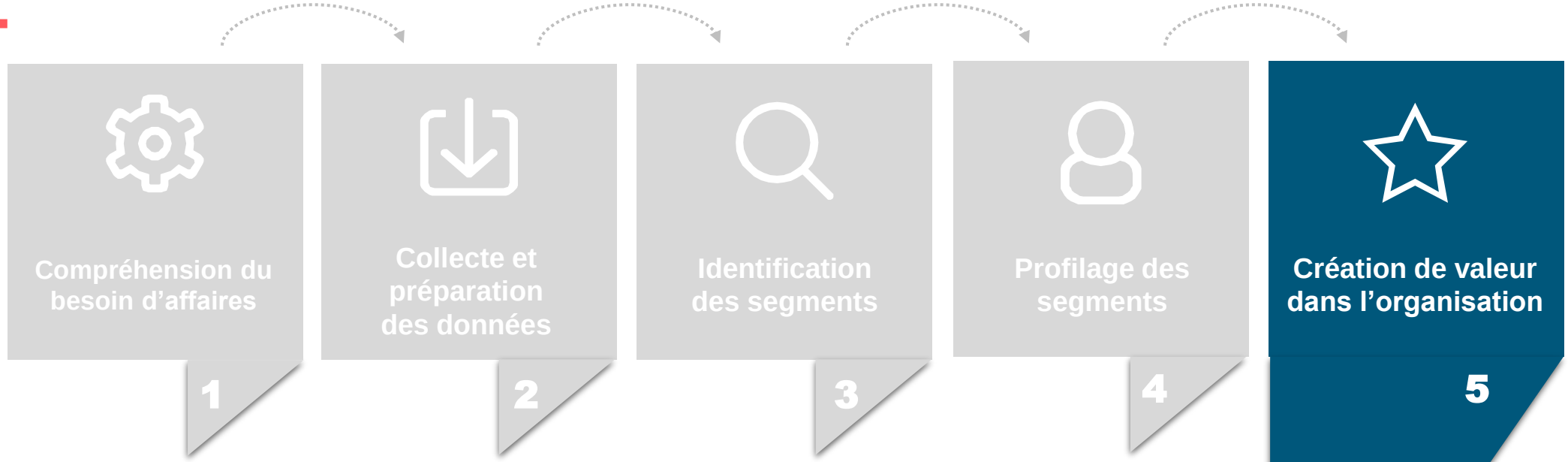


Les segments prennent vie!

Profilage des segments

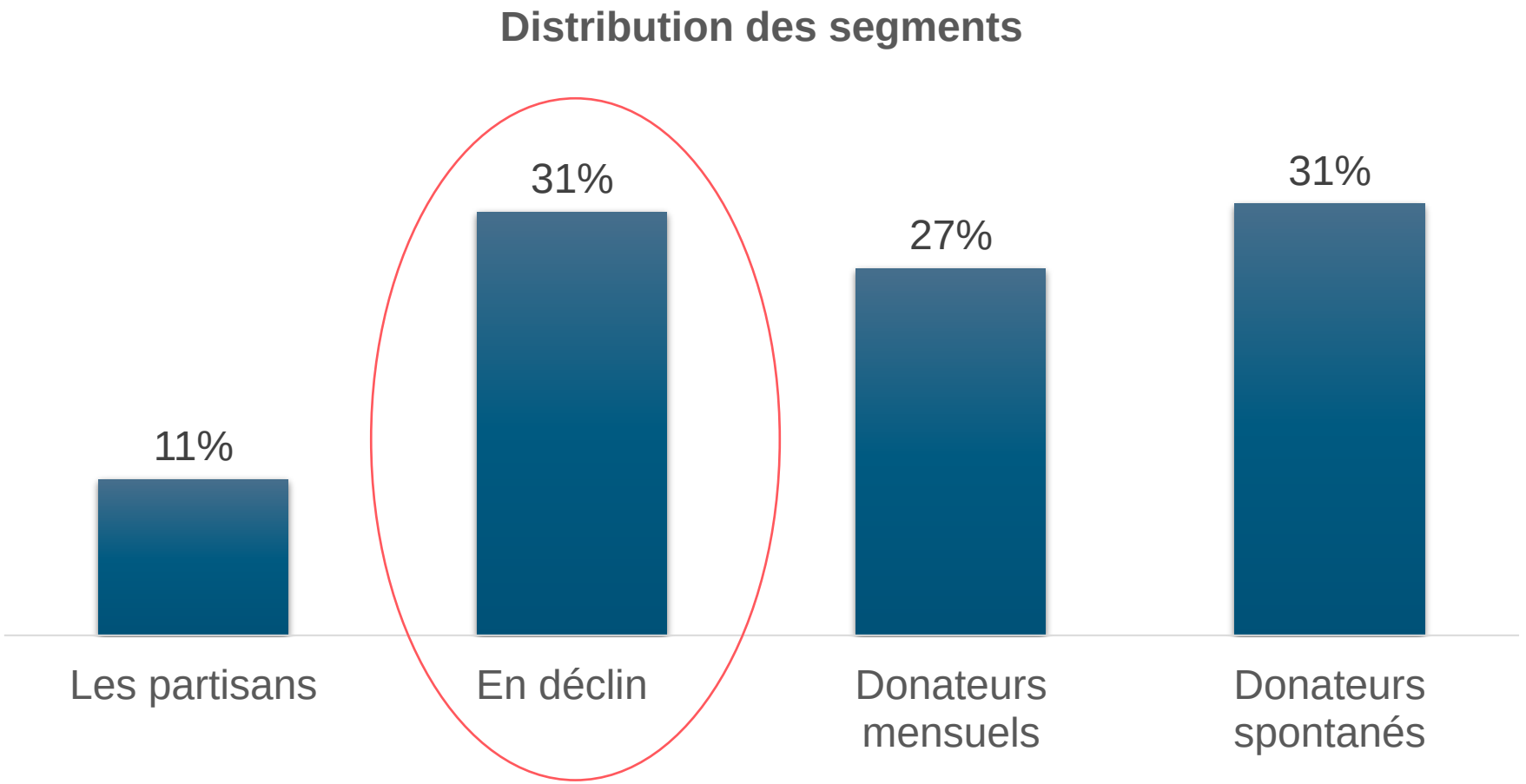
Des domaines variés !

Étapes pour une segmentation réussie



Une segmentation,
c'est bien plus que de jolis personas!

Les segments à prioriser, lesquels ?



Pourquoi ce segment ?

Création de personas (Exemple)

Jean-Louis, 32 ans, *Ingénieur mécanique*

Originaire de Québec, Jean-Louis a déménagé à Montréal avec sa femme est ses 3 enfants. Il est plus jeune que la moyenne et anxieux par sa situation financière. Voici ses caractéristiques :

- **Épargne moyenne** : 152 000 $
- **Portefeuille diversifié** : 35 % à 60 % en titres d'emprunt et le reste en titres de participation.
- **Tolérance aux pertes** : Peut supporter une baisse maximale de 15 % de la valeur de son portefeuille.
- **Objectif principal** : Économiser pour la retraite ou projet à court terme
- **Situation budgétaire** : Possède des surplus budgétaires et prévoit conserver son investissement pendant 15 ans ou plus.
- **Connaissances financières** : Solides connaissances des placements.
- **Préférences d'investissement** : Prêt à tolérer quelques fluctuations du rendement et des pertes modérées de capital.

NUMEA

10905, BOUL. HENRI-BOURASSA EST
MONTRÉAL, QUÉBEC. H1C 1H1

514.881.3700

INFO@NUMEA.CA / NUMEA.CA