Modélisation prédictive avec R dans un contexte de production - De l'extraction au déploiement

R à Québec 2019

Bienvenue

- Mettre nos noms, experiences, etc (pourrait être funny un peu les descriptions)

Objectifs de l’atelier

Informations

- Expliquer le concept du livre (ressource pour eux)

- Lien vers le repo GitHub de set up

Une première fois !

- Grande première planétaire

- On est pas orgueilleux! On veut du feedback.

- Disclaimer : Jeu de données simple (mais gros!)

➔ Emphase sur le processus

Processus de modélisation

- Parler du concept de back-and-forth

- Pour atelier on va s’en tenir à une ligne toute tracée

BIXI

Problématique

- Pour des raisons inconnues, on nous demande :

1. Quels utilisateurs sont susceptibles de revenir à la même station?

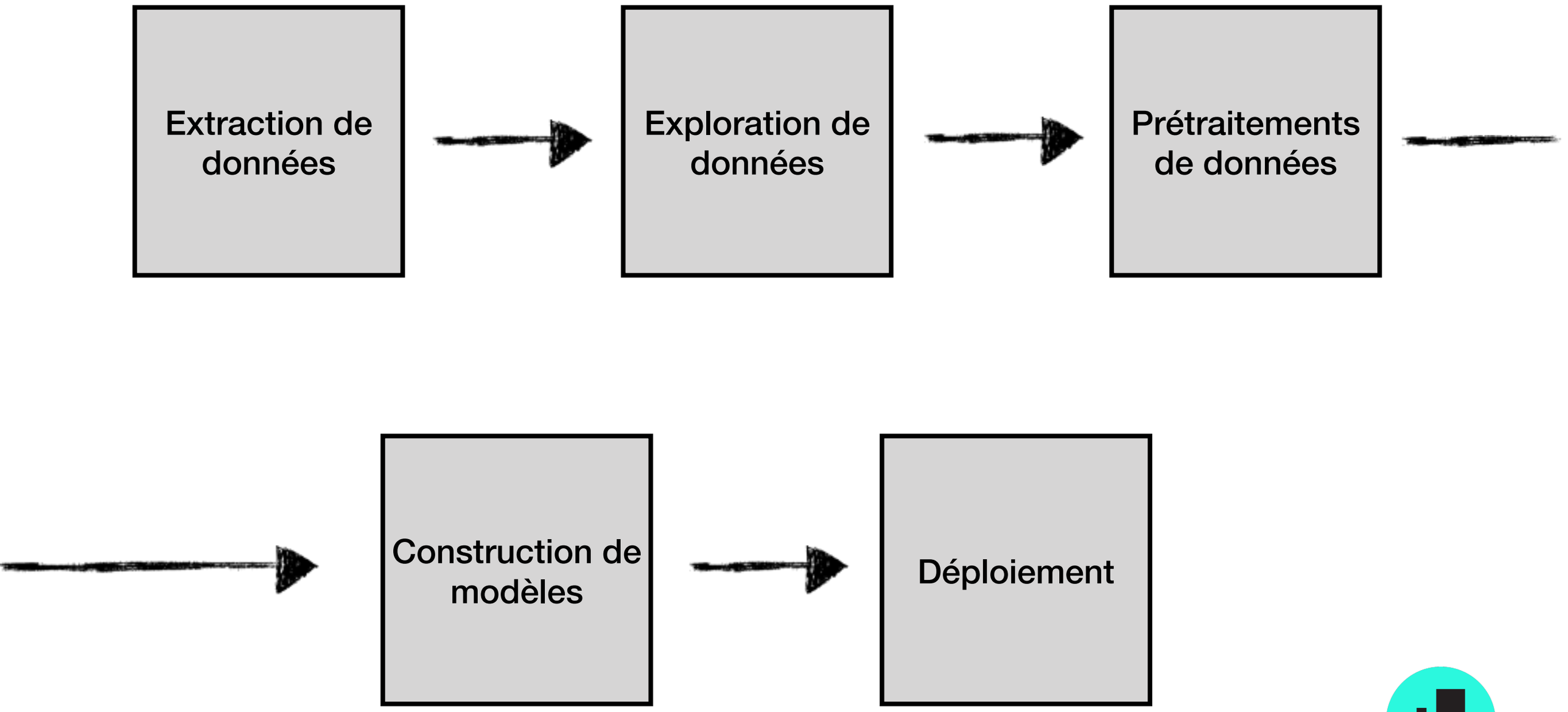
2. Quelle sera la durée du trajet d’un utilisateur?

Définition de la tâche

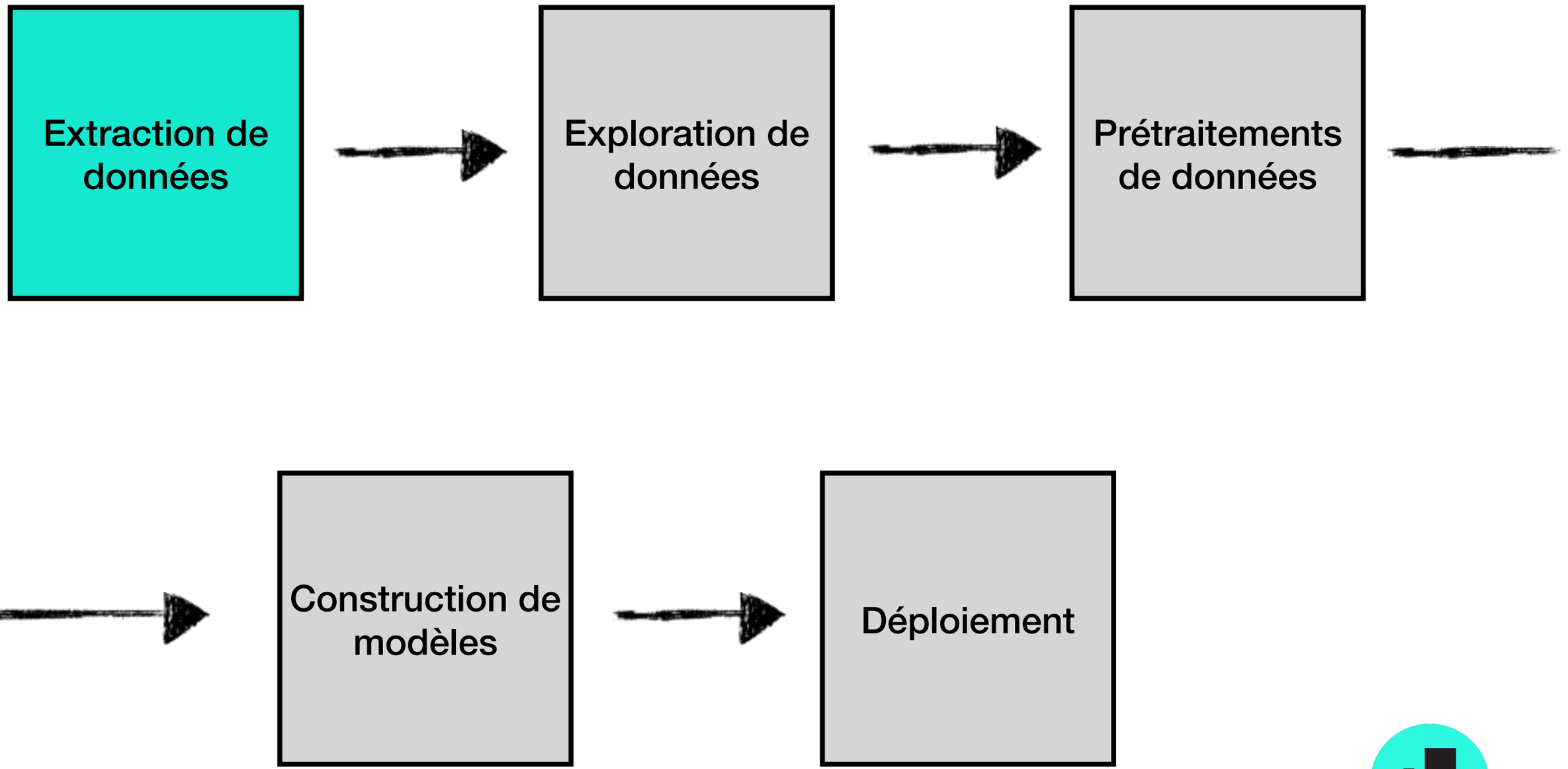
Régression vs Classification

$$y \approx f(x)$$

Étapes à accomplir



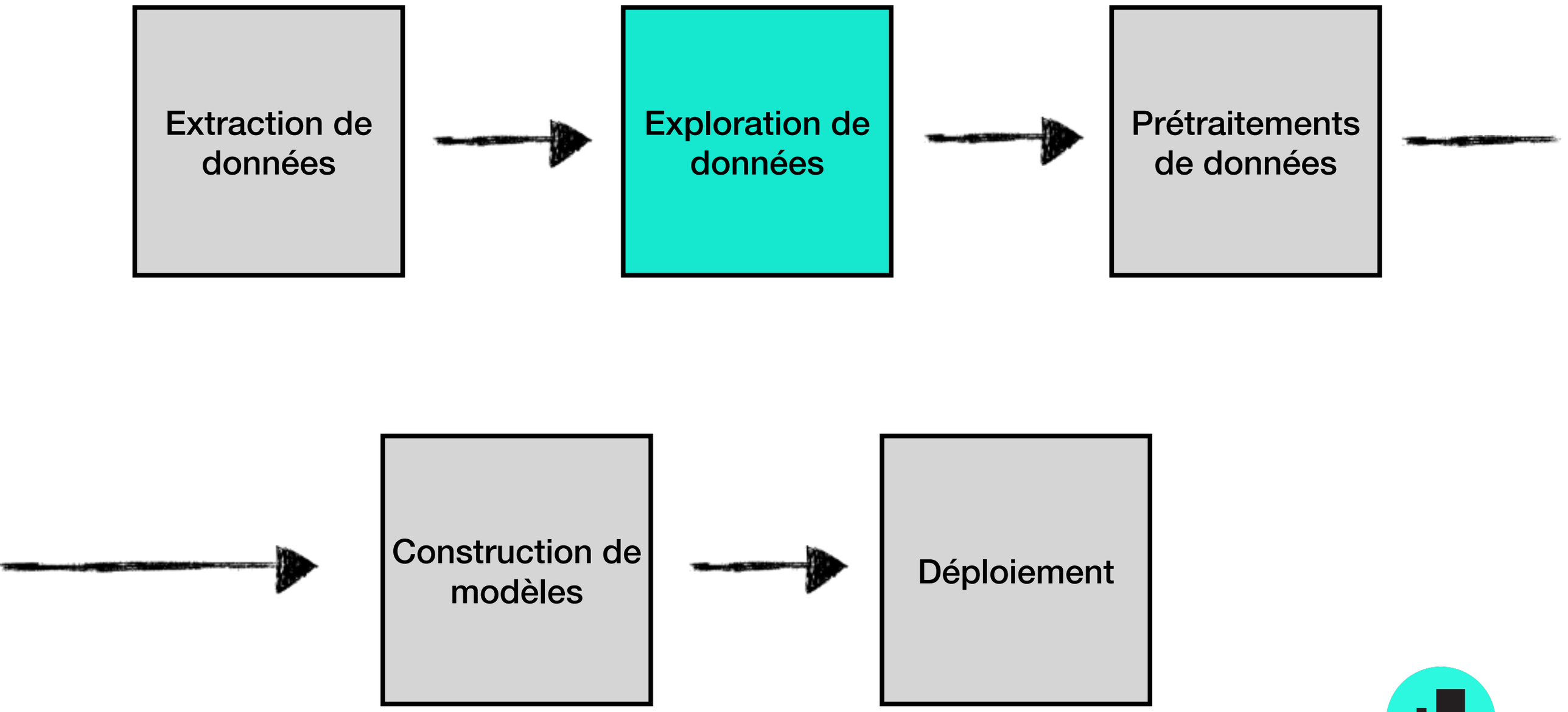
Collectons !


Extraction de données
Exploration de données
Prétraitements de données
Construction de modèles
Déploiement

Extraction de données

- Lau

Explorons !


Extraction de données
Exploration de données
Prétraitements de données
Construction de modèles
Déploiement

Exploration des données

Objectifs

1. S’approprier le jeu de données;

2. Suggérer des transformations pertinentes pour le prétraitement des données.

Appropriation

Pour chaque variable d’entrée…

1. Observer la distribution;
2. Observer la corrélation avec les autres variables d’entrée;
3. Observer l’effet unidimensionnel sur la variable réponse;
4. Observer les effets multidimensionnels avec les autres variables d’entrée sur la variable réponse.

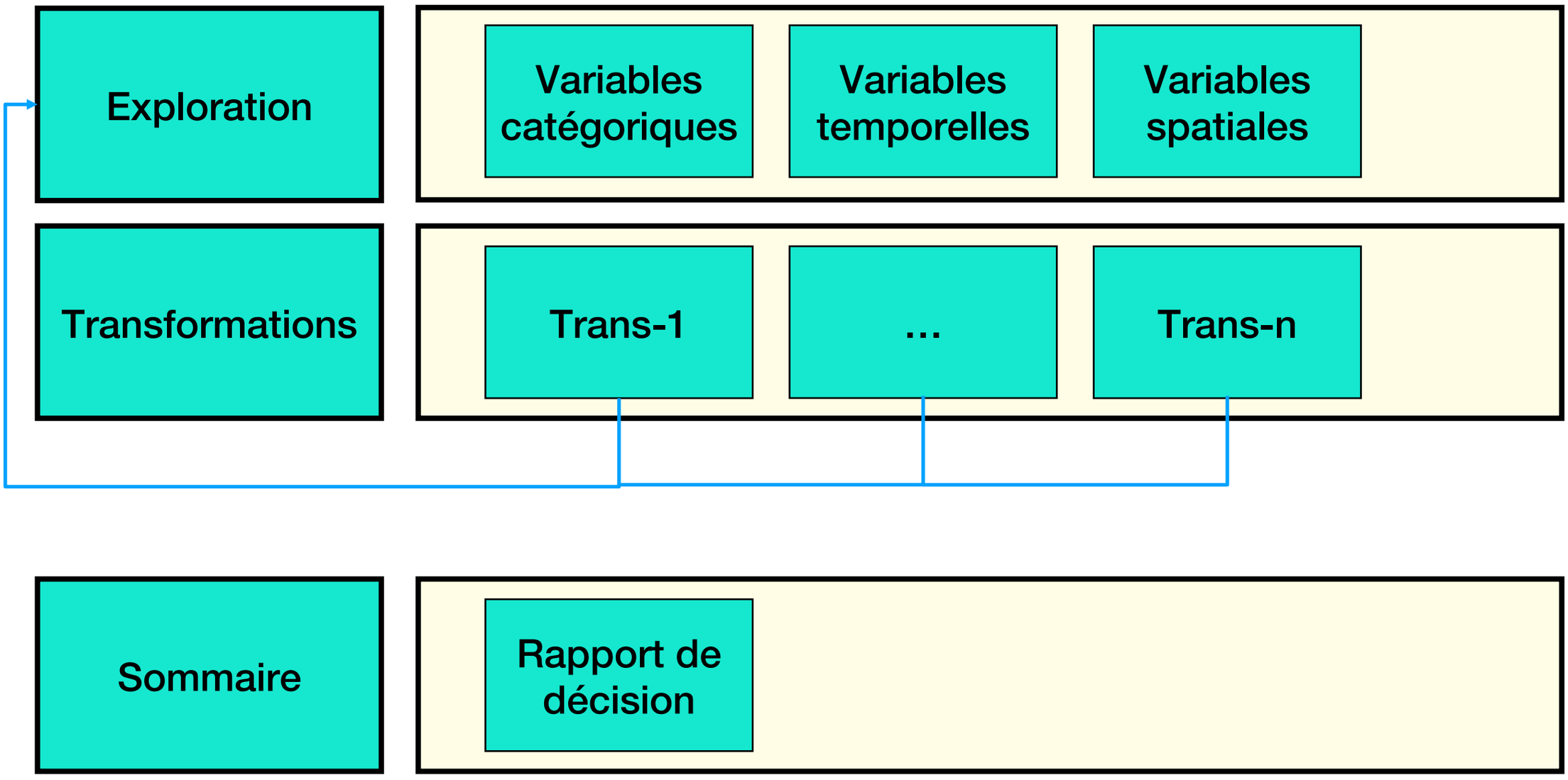
Transformations

- Exclusion;
- Identité;
- Regroupements;
- Tout autre fonction.

Variables d’entrée

- Numérique;
- Catégorique;
- Temporelle;
- Spatiale.

Récits à implémenter



Nom du récit	Variables catégoriques
ID	1
Section	Exploration des données

Description

Créer des visuels pour comprendre les distributions des variables d'entrée

Packages proposés

Agrégation des données :
- base
- dplyr
- data.table

Packages proposés pour les graphiques :
- base
- ggplot2
- plotly
- lattice

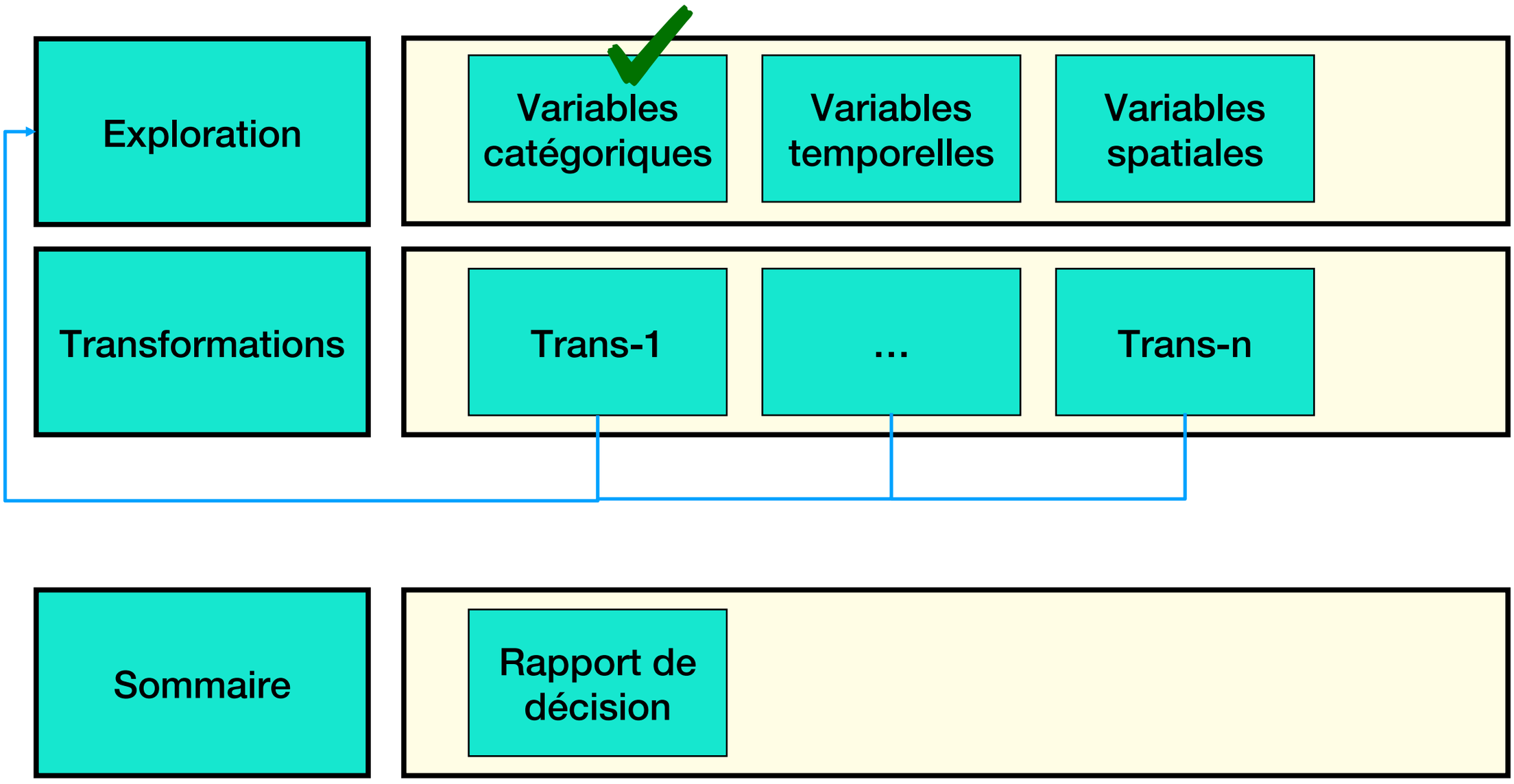
Entrée

1. Données brutes
2. Données utilisées pour le prétraitement
3. Données externes

Sortie

1. Certains des graphiques suivants :
   - Distribution
   - Corrélations
   - One-way
   - Interactions

Récits à implémenter



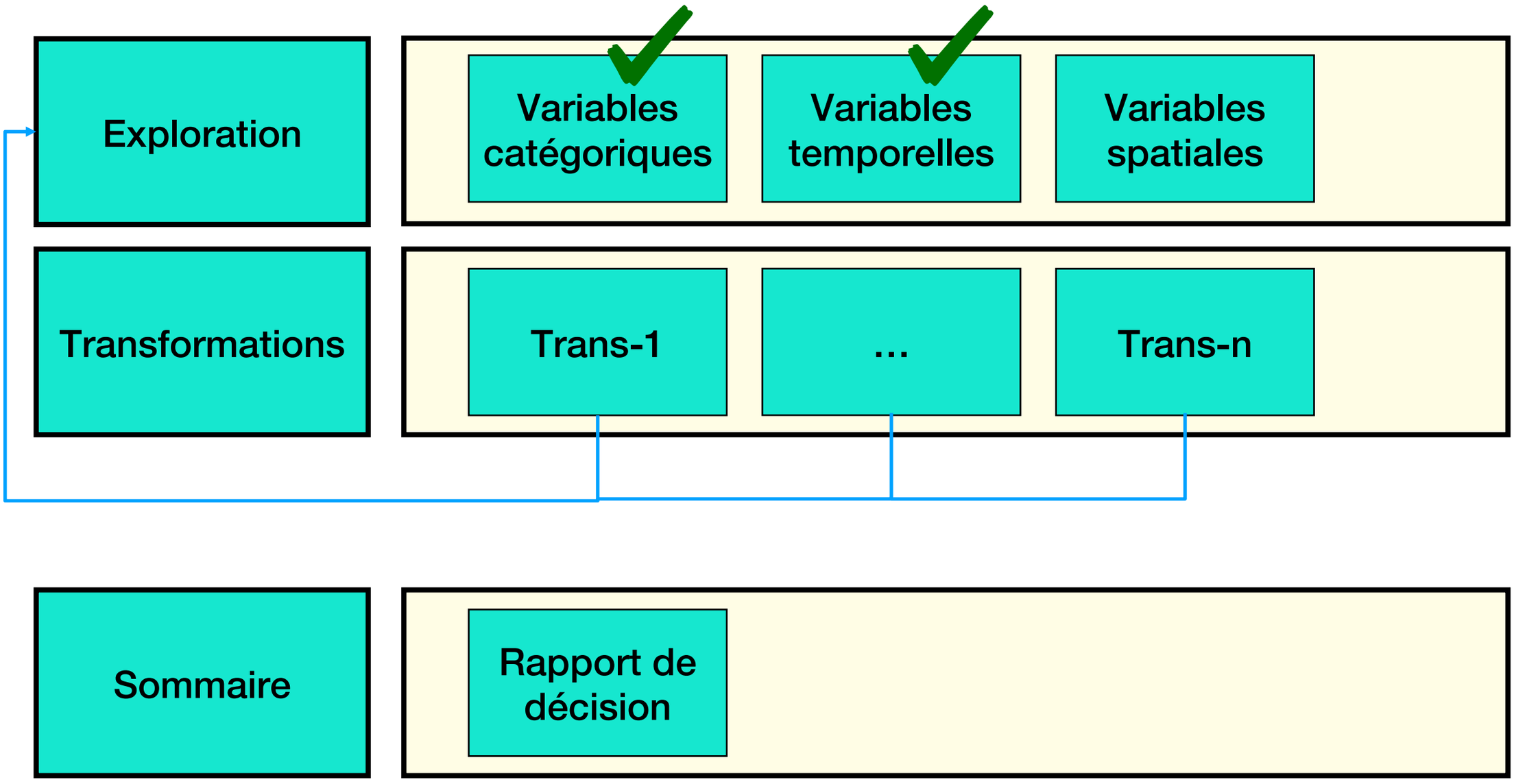
Nom du récit	Variables temporelles
ID	2
Section	Exploration des données

Description
Créer des visuels pour comprendre les distributions des variables d'entrée

Entrée
1. Données brutes 2. Données utilisées pour le prétraitement 3. Données externes

Sortie
1. Certains des graphiques suivants : - Distribution - Corrélations - One-way - Interactions

Récits à implémenter



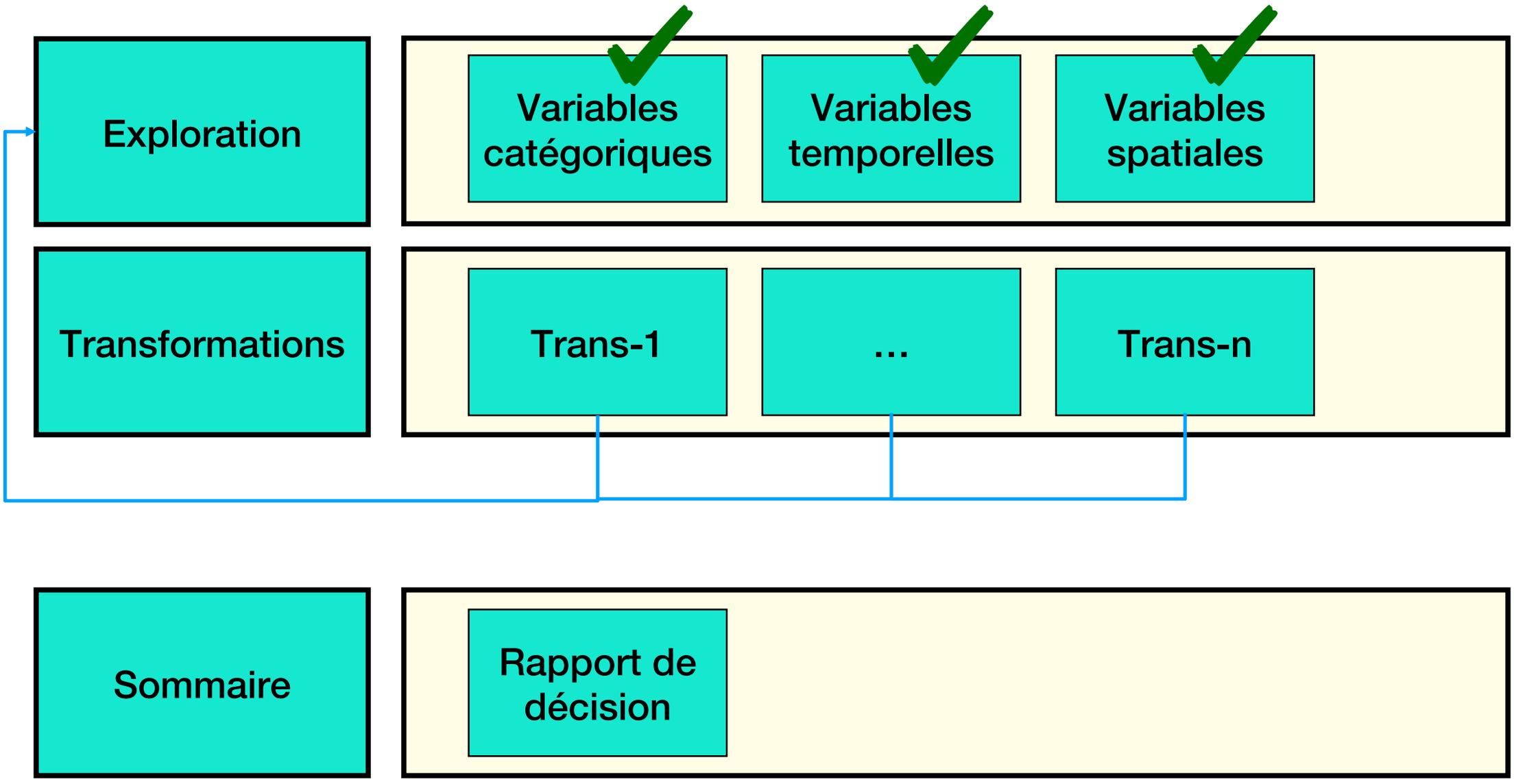
Nom du récit	Variables spatiales
ID	3
Section	Exploration des données

Description
Créer des visuels pour comprendre les distributions des variables d'entrée <u>Packages proposés</u> Création de cartes : • ggmap • plotly • leaflet • tmap

Entrée
1. Données brutes 2. Données utilisées pour le prétraitement 3. Données externes

Sortie
1. Certains des graphiques suivants : - Distribution - Corrélations - One-way - Interactions

Récits à implémenter



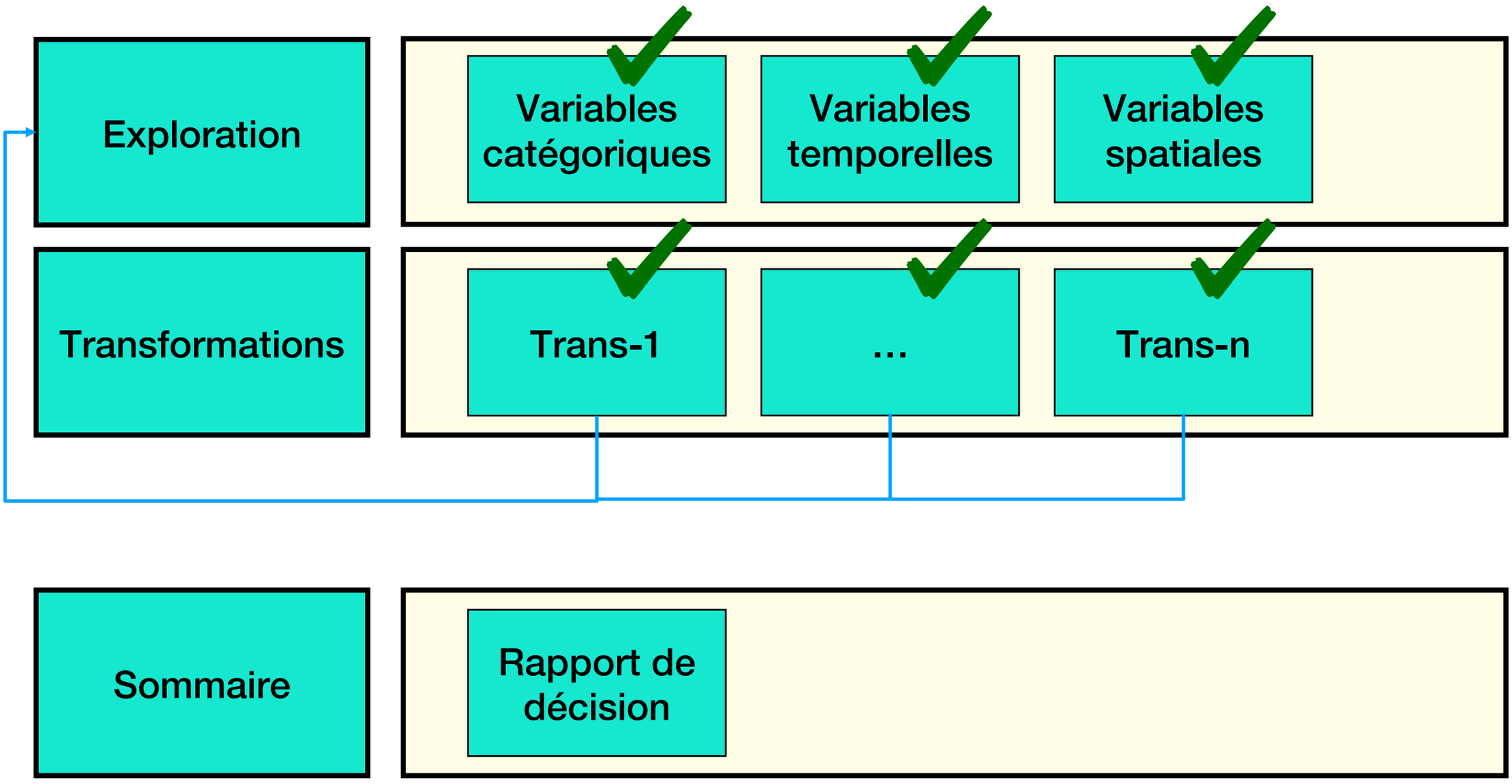
Nom du récit	Transformations
ID	4
Section	Exploration des données

Description
Analyser des transformations potentielles suite à l'analyse des variables disponibles dans la collecte des données

Entrée
1. Données brutes 2. Données utilisées pour le prétraitement 3. Données externes

Sortie
1. Certains des graphiques suivants : - Distribution - Corrélations - One-way - Interactions

Récits à implémenter



Nom du récit	Rapport de décision
ID	5
Section	Exploration des données

Description
Créer un rapport de décision concernant les transformations suggérées afin que l'étape du prétraitement des données puisse incorporer l'information

Entrée
1. Les graphiques créés depuis le début de la section

Sortie
1. Un rapport de recommandations

Recommandations (1/2)

- Moments de la journée
  - Matin : 6h à 11h
  - Journée : 11h à 16h
  - Soir : 16h à 23h
  - Nuit : 23h à 6h

- Semaine/Fin de semaine
  - Semaine : Lundi, Mardi, Mercredi, Jeudi, Vendredi
  - Fin de semaine : Samedi, Dimanche

Recommandations (2/2)

- Regroupement des quartiers
  - Groupe 1 : Plateau-Mont-Royal
  - Groupe 2 : Ville-Marie
  - Groupe 3 : Ahuntsic-Cartierville, Villeray-Saint-Michel-Parc-Extension, Rosemont-La Petite-Patrie, Mercier-Hochelaga-Maisonneuve
  - Groupe 4 : Outremont, Côte-des-Neiges-Notre-Dame-de-Grâce, Westmount, Le Sud-Ouest, Verdun, LaSalle
  - Groupe 5 : Autre

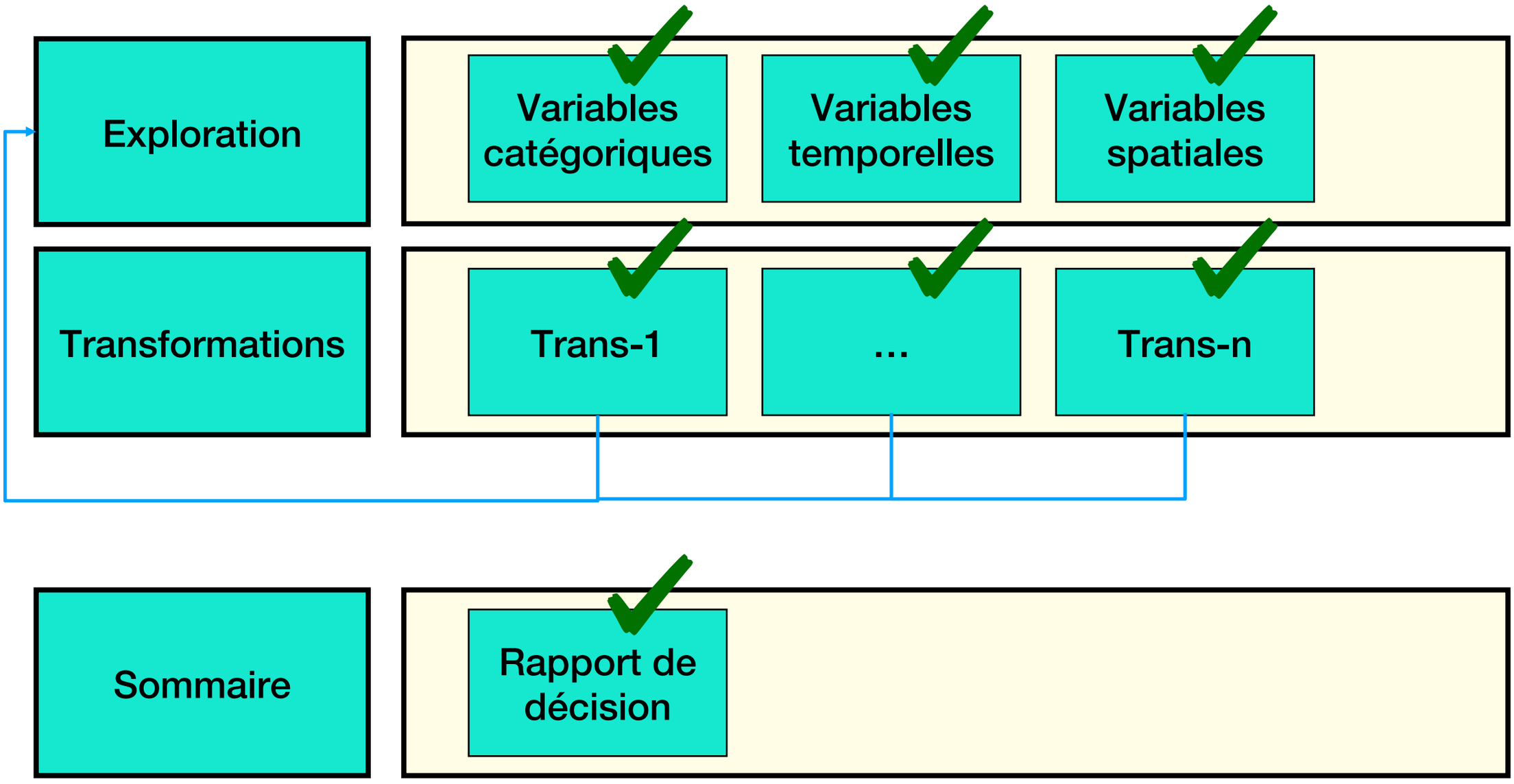
Données externes à explorer

- Météo
- FSAs au lieu de quartiers
- Réseau de Métro
- Jours fériés
- Dates de festivals

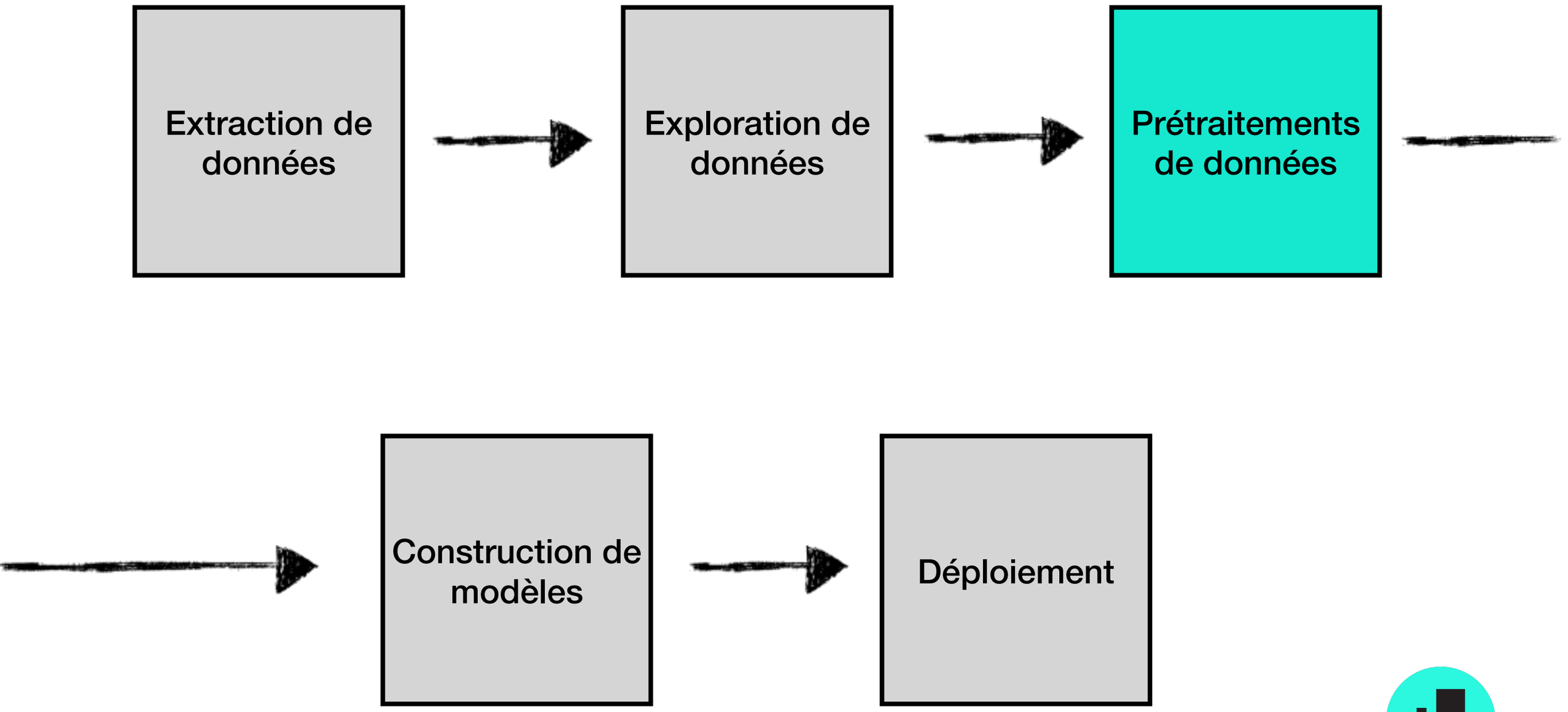
Transformations à explorer

- Meilleur regroupement des moments de la journée
- Distance avec les autres quartiers
- Distance avec le fleuve (≈distance avec la bordure de la carte des quartiers)

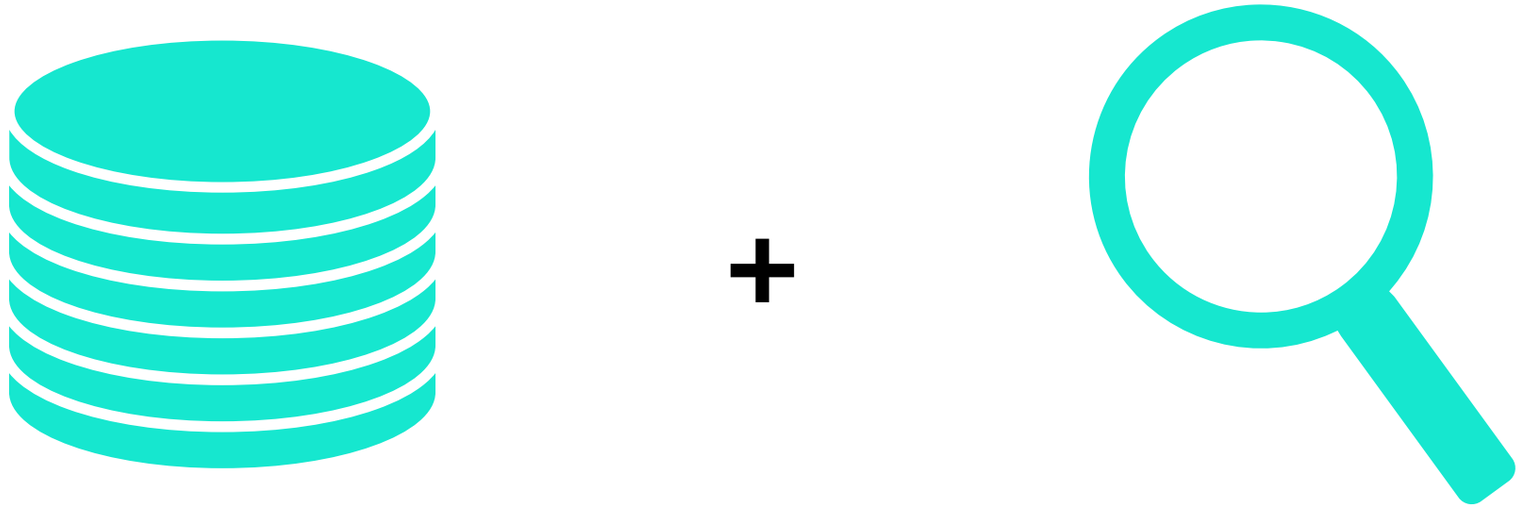
Récits à implémenter

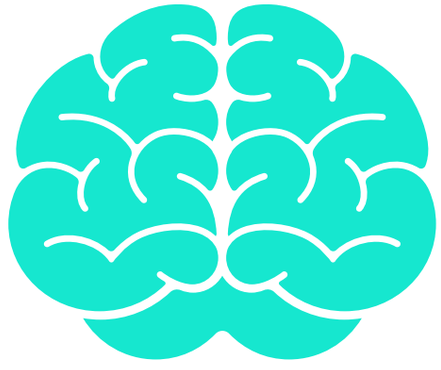


Nettoyons !


Extraction de données
Exploration de données
Prétraitements de données
Construction de modèles
Déploiement

Prétraitements de données





Objectifs

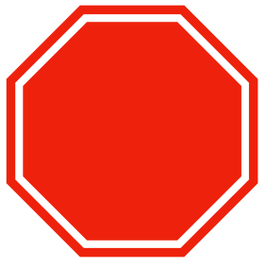
2 objectifs derrière le prétraitement de données:

1. Transformer les données dans un format compatible pour l'algorithme

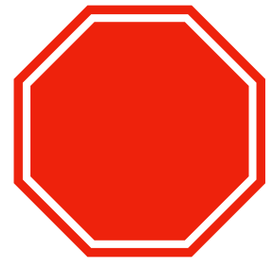
2. Transformer les données de manière à faciliter l'apprentissage

Prétraitements en 3 étapes

1. Nettoyage de données
2. Réduction de données
3. Transformations de données



Séparation

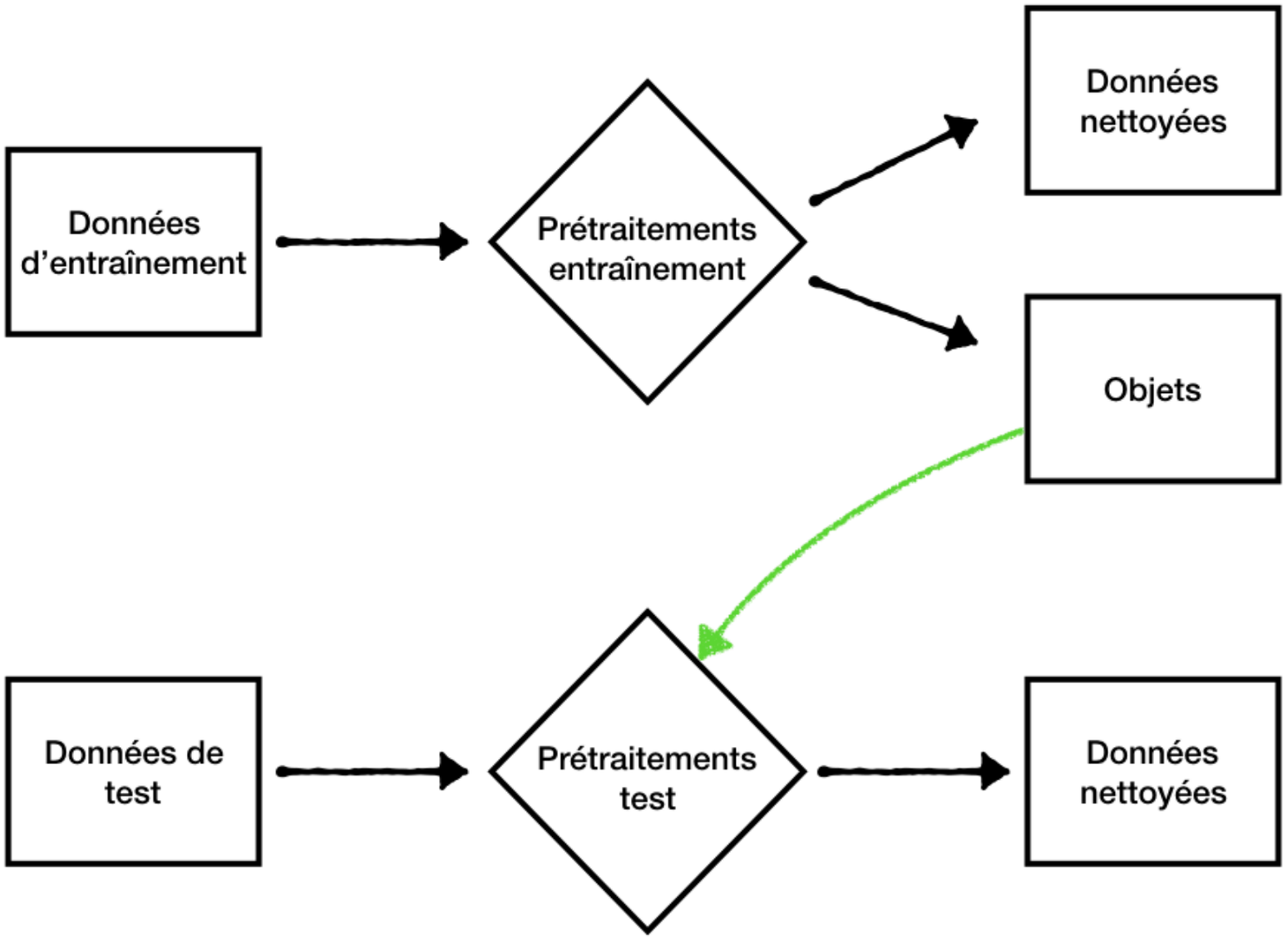


Tout d’abord, il faut séparer notre jeu de données …

Méthodes:

- Aléatoire
- Échantillonnage stratifié

Vu d’ensemble



Récits à implémenter

Séparation des données	Séparation			
Nettoyage des données	Implémenter l'imputation	Traitement données aberrantes	Traitement données catégoriques	
Transformation de données	Création de nouveaux attributs			

Nom du récit	Séparation du jeu de données
ID	1
Section	Prétraitements de données

Description

Implémenter une procédure pour séparer les données en 2 ensembles :

- Entraînement
- Test

Utiliser une séparation aléatoire (voir fonction `sample` de base R)

Entrée

1. Données brutes pour l'ensemble du jeu de données

Sortie

1. Un vecteur d'index pour les observations d'entraînement
2. Sauvergarder les données test

Récits à implémenter

Séparation des données	Séparation ✓
Nettoyage des données	Implémenter l'imputation Traitement données aberrantes Traitement données catégoriques
Transformation de données	Création de nouveaux attributs

Nettoyage des données

Imputations de données manquantes

1. Faire le **constat** sur la quantité de données manquante.
2. Identifier le mécanisme de non-réponse
3. Traiter les données manquantes

Constat sur les données manquantes

- Quelles variables ont des données manquantes?

- Quelle est la proportion de données manquantes par variable?

Mécanisme de non-réponse

- Données manquantes complètement au hasard (MCAR)
- Données manquantes au hasard (MAR)
- Données manquantes pas au hasard (NMAR)

Voir la section 5.1.1 du livre pour des exemples.

Faire l’imputation

- Analyse des cas complets: Conserver uniquement les observations pour lesquelles toutes les variables sont présentes.

- Imputation par une mesure de centralité: Utiliser la moyenne, la médiane ou le mode pour remplacer les données manquantes.

- Imputation par régression: Remplacer les données manquantes par la prévision de modèle de régression entraîné sur les observations pour lesquelles cette variable est présente.

- Imputation par régression stochastique: Même chose que la méthode par régression, mais on ajoute un résidu aléatoire à la prévision.

Nom du récit	Implémenter l'imputation
ID	2
Section	Prétraitements de données

Description
Implémenter l'imputation de données manqutes. Il faut divisier l'imputation en 2 sections: 1. Une qui aura comme but de définir les valeurs d'imputations. 2. Une qui aura comme but d'appliquer ces valeurs aux données manquants. Le 1. sera fait en entraînement seulement. Le 2. sera fait autant en entraînement qu'en inférence.

Entrée
1. Données brutes 2. Données brutes + liste de valeurs d'imputations

Sortie
1. Liste de valeurs d'imputaiton 2. Données imputées

Récits à implémenter

Séparation des données	Séparation ✓
Nettoyage des données	Implémenter l'imputation ✓ · Traitement données aberrantes · Traitement données catégoriques
Transformation de données	Création de nouveaux attributs

Traitement des données aberrantes

Pourquoi?: Effets importants dans le calcul des estimateurs

Comment?

- Plus ou moins 3 écarts-types de la moyenne
- Plus ou moins 1.5 EI (écarts interquartile)
- Partitionnement (*clustering*)

Données catégoriques

Identifier le type de données catégoriques:

- Attribut ordinal
- Attribut nominal

Le traitement n’est évidemment pas le même :

- Assigner une valeur numérique
- Encodage *un-chaud (one-hot encoding)*

Attribut ordinal

```
##    Id condition_station
## 1:  1             moyen
## 2:  2          mauvais
## 3:  3         excellent
## 4:  4               bon
```

```
##    Id condition_station
## 1:  1                 1
## 2:  2                 0
## 3:  3                 3
## 4:  4                 2
```

Super !

Attribut nominal

```
##    Id start_quartier
## 1:  1    Ville-Marie
## 2:  2        Verdun
## 3:  3     Westmount
## 4:  4      LaSalle
```

```
##    Id start_quartier
## 1:  1               1
## 2:  2               2
## 3:  3               3
## 4:  4               4
```

Issh ! ...

Attribut nominal

```
##    Id quartier_centre.ville quartier_plateau.mont.royal quartier_verdun
## 1:  1                     1                           0               0
## 2:  2                     0                           1               0
## 3:  3                     0                           0               1
## 4:  4                     0                           0               0
##    quartier_rosemont
## 1:                 0
## 2:                 0
## 3:                 0
## 4:                 1
```

Ahh c’est mieux …

Nom du récit	Traitement des données aberrantes
ID	3
Section	Prétraitements de données

Description
Implémenter une opération pour gérer les données abberantes. Choisir une méthode de détection de données aberrantes et appliquer un traitement pour ces données. Appliquer ce traitement dans le prétraitement des données d'entraînement seulement.

Entrée
1. Données brutes

Sortie
1. Données brutes sans les données abberantes 2. "Print" du nombre de données traitées (fins de documentation)

Nom du récit	Encodage des données catégoriques
ID	4
Section	Prétraitements de données

Description

Implémenter une opération pour gérer les données catégoriques.

1. Idenfitier les attributs catégoriques et déterminer leur type.
2. Faire l'assignation numérique pour les attributs ordinaux.
3. Lire la documentation de la fonction **caret::dummyVars**
4. Implémenter l'encodage *un-chaud*

Conseils: Il faut séparer l'encodage un-chaud pour l'entraînement et l'inférence. Prévoir un traitement pour les classes inconnus (défaut).

Entrée

1. Données brutes avec des catégories

Sortie

1. Données sans catégories
2. Objets nécessaire pour faire l'encodage *un-chaud*

Récits à implémenter

Séparation des données	Séparation ✓
Nettoyage des données	Implémenter l'imputation ✓ ; Traitement données aberrantes ✓ ; Traitement données catégoriques ✓
Transformation de données	Création de nouveaux attributs

Réduction de données

Fléau de la dimensionnalité

- Survient lorsque le nombre de dimensions est très élevée (p >> n)
- Éloigne les données les unes des autres
- Rend l’apprentissage plus difficile

Fléau de la dimensionnalité

On suppose qu'on a un jeu de données avec $p = 1$ attribut de n observations où $x_1, \ldots, x_n \overset{iid}{\sim} U(0, 1)$.
Combien d'observations en moyenne se trouveront dans l'intervalle $[0; 0.1]$?

La réponse : $\frac{n}{10}$ observations.

Maintenant, supposons que notre jeu de données est plus complexe et possède $p = 10$ attributs au lieu d'un seul attribut. Les observations suivent toujours une loi uniforme où $x_1, \ldots, x_n \overset{iid}{\sim} U([0, 1]^{10})$.
Combien d'observations en moyenne se trouveront dans l'intervalle $[0; 0.1]^{10}$?

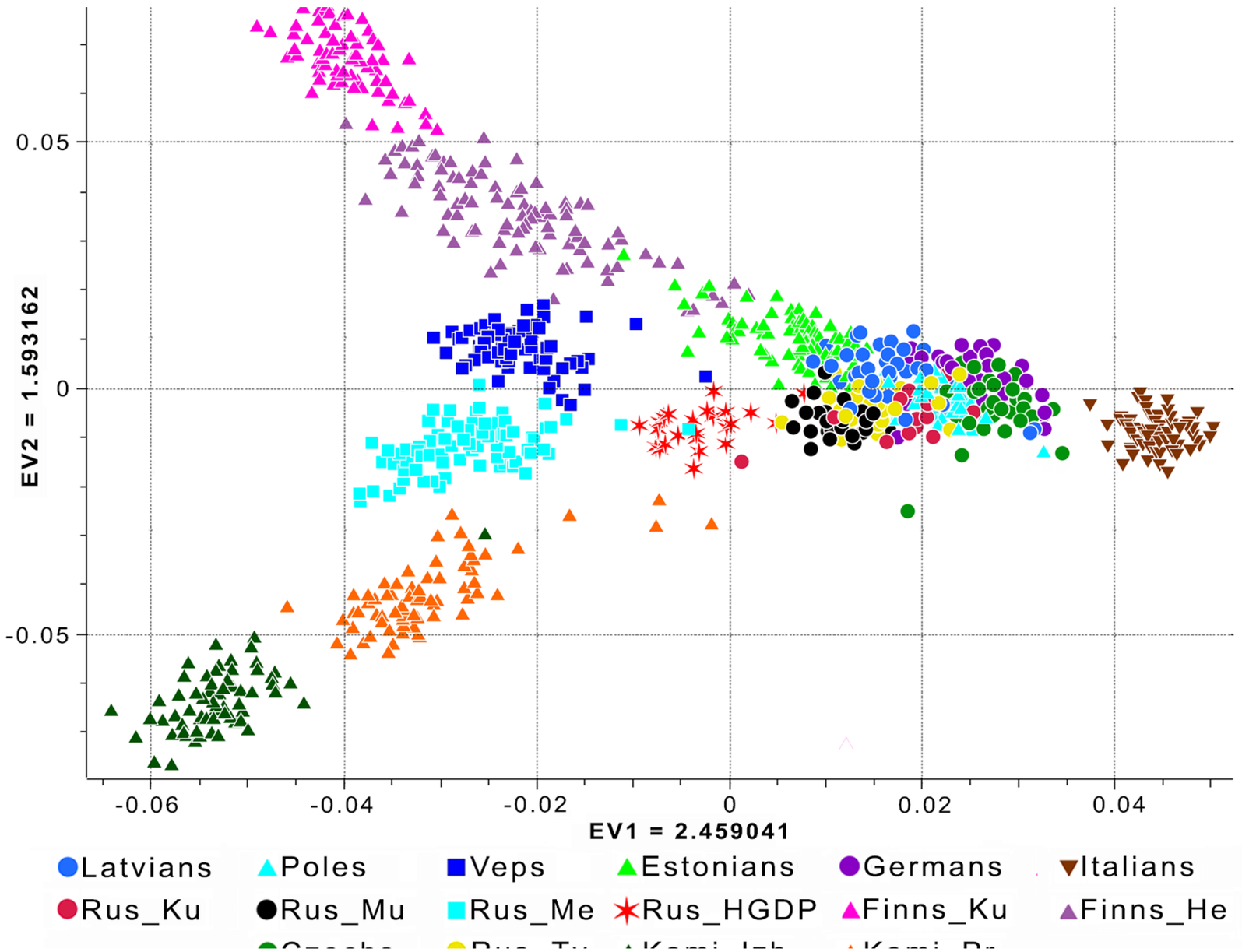
La réponse : $n(\frac{1}{10})^{10}$ observations.

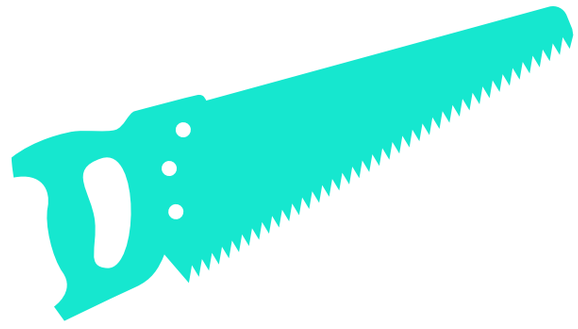
Source: Charest, Anne-Sophie. 2018b. "Données Manquantes."
In *STT-7330: Méthode d'analyse de Données*. Québec, Canada: Université Laval.

Méthodes de réduction de dimension

- Analyse en composantes principales
- Positionnement multidimensionnel
- Anlayse factorielle
- Calcul de scores

ACP


0.05
0
-0.05
EV2 = 1.593162
-0.06
-0.04
-0.02
0
0.02
0.04
EV1 = 2.459041
Latvians
Poles
Veps
Estonians
Germans
Italians
Rus_Ku
Rus_Mu
Rus_Me
Rus_HGDP
Finns_Ku
Finns_He



Transformation de données

Normalisation

- Les échelles des attributs peuvent être différents
- Peut avoir un impact important dans certains algorithmes basés sur des distances (k-PPV, *clustering*)
- Permet de ramener les données autour d’une distriubution plus “standard”

Méthodes de normalisation

- Normalisation centrée réduite
- Normalisation *min-max*
- Normalisation par décimation

Discrétisation et nouveaux attributs

- Communément appelé le *feature engineering*

- Grouper des observations à l'intérieur de *buckets* ou *bins*

- Transformer des attributs existants pour en créer des nouveaux

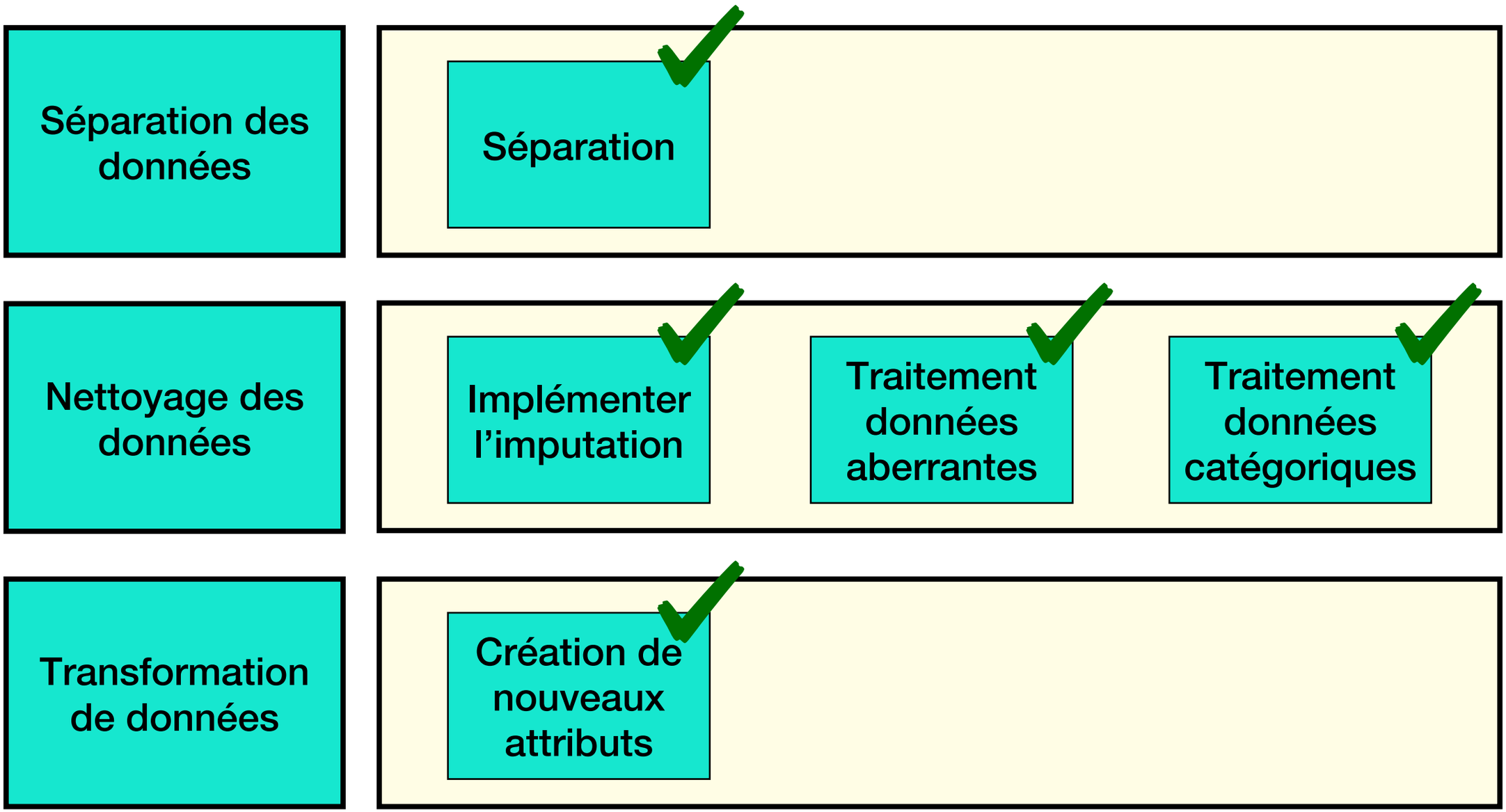
Nom du récit	Création de nouveaux attributs
ID	5
Section	Prétraitements de données

Description
Implémenter le code qui permet de créer les nouveaux attributs pour le modèle. Utiliser les connaissances acquises lors de l'exploration de données. Créer une liste de variables à conserver dans le modèle. Conseils d'attributs: - Grouper des quartiers - Moment de la journée - Fin de semaine versus semaine

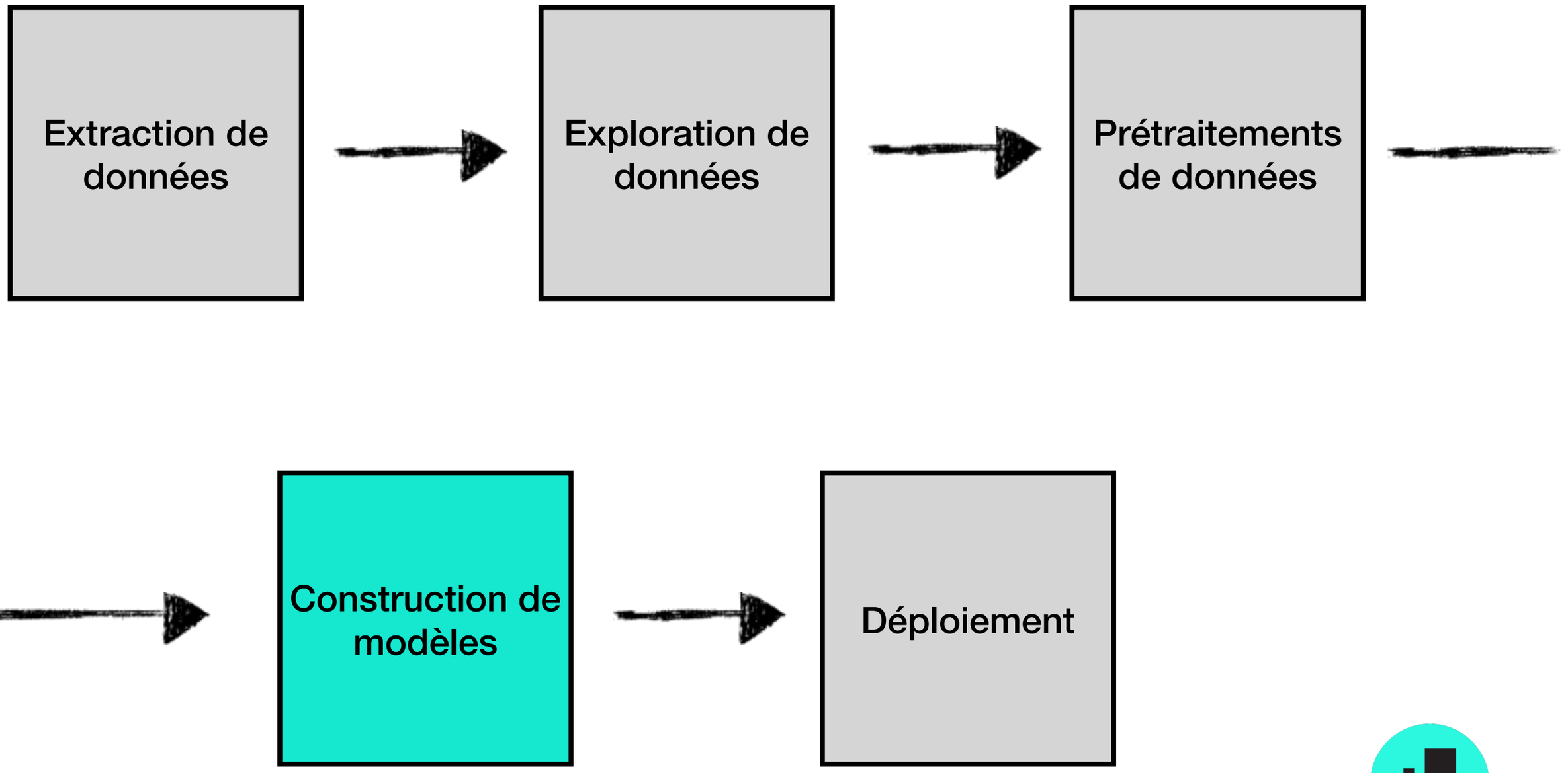
Entrée
1. Données brutes

Sortie
1. Données avec nouveaux attributs 2. Liste de variables à conserver dans le modèle lors de l'inférence

Récits à implémenter



Modélisons !


Extraction de données
Exploration de données
Prétraitements de données
Construction de modèles
Déploiement

Construction de modèles

Objectifs

1. Objectif de modélisation :

   Prédire la réponse y pour de **nouvelles données** x

2. Objectif global :

   Fournir un objet R à utiliser avec `predict`

Modélisation en 3 étapes

1. Choix des modèles (on y a déjà pensé!)

2. Estimation des paramètres

3. Sélection du modèle final

Concepts clefs

entraînement/validation/test fonction de perte régularisation*
hyper-paramètres biais/variance validation croisée …

Mais avant…

Un peu old school ça…

$$(\mathbf{X}|\mathbf{y}) = \left(\begin{array}{ccc|c} x_{11} & \dots & x_{1d} & y_1 \\ x_{21} & \dots & x_{2d} & y_2 \\ \vdots & & \vdots & \vdots \\ x_{n1} & \dots & x_{nd} & y_n \end{array}\right) \begin{array}{l} \} \overset{\approx 1/2}{\longrightarrow} (\mathbf{X}_{\text{train}}|\mathbf{y}_{\text{train}}) \\ \} \overset{\approx 1/4}{\longrightarrow} (\mathbf{X}_{\text{val}}|\mathbf{y}_{\text{val}}) \\ \} \overset{\approx 1/4}{\longrightarrow} (\mathbf{X}_{\text{test}}|\mathbf{y}_{\text{test}}) \end{array}$$

Optionnel

Entrée	Sortie
Données test	Données validation + Données test

Description d’un modèle

Un peu de statistique… ε

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

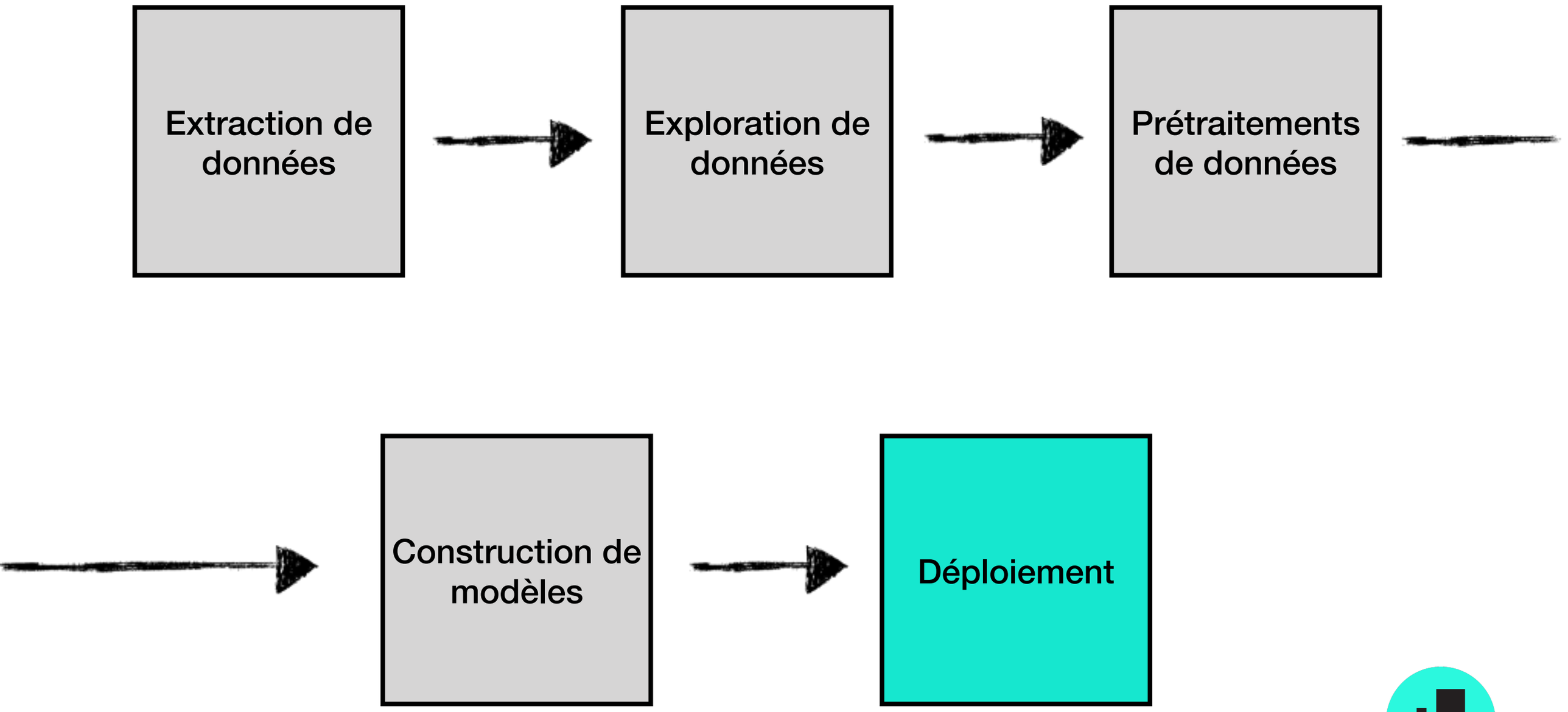
Formulation d’un modèle

Classification…

ε

Choix des modèles

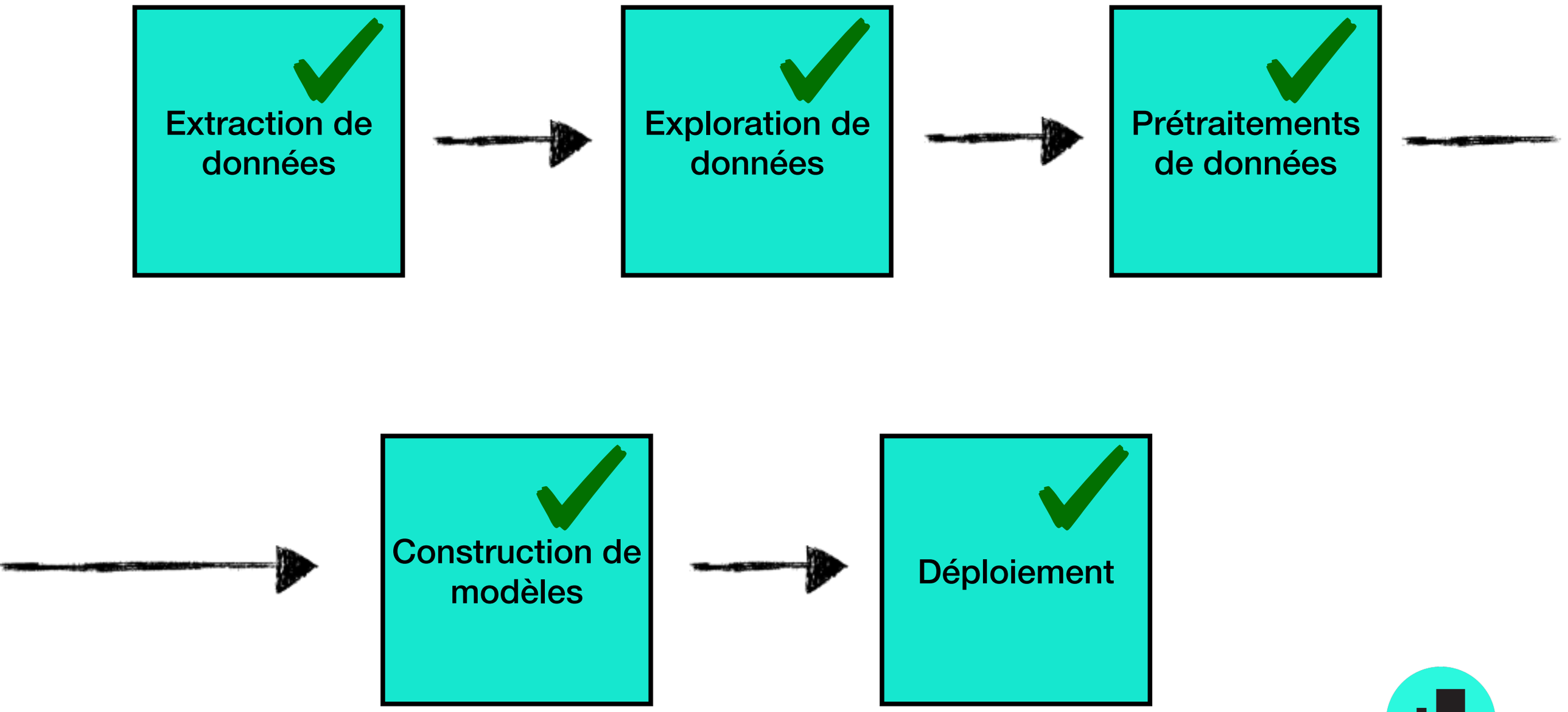
Déployons !


Extraction de données
Exploration de données
Prétraitements de données
Construction de modèles
Déploiement

Déploiement

- Jé the man !

Misson accomplie !



Mot de la fin