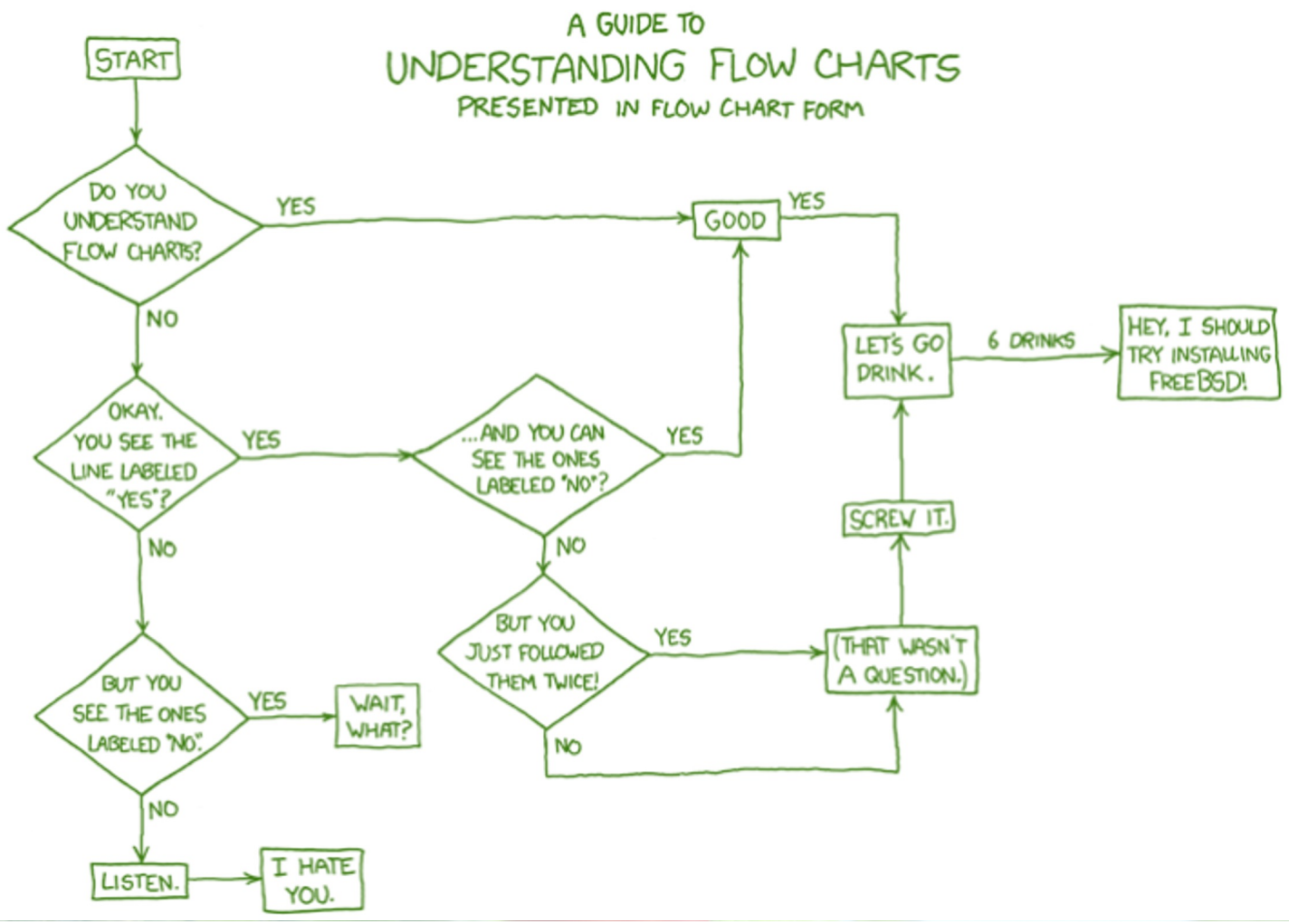


5. Le flux de travail analytique

Le flux de travail analytique

Vous en avez probablement assez des **discussions sur le contexte** et préféreriz passer à l'analyse des données proprement dite.

Une dernière chose : le **contexte du projet**.

La science des données ne se résume pas à l'analyse des données ; cela apparaît clairement lorsque l'on examine les étapes typiques d'un **projet de science des données**.

L'analyse a lieu dans un contexte de projet plus large, ainsi que dans le contexte d'une plus grande **infrastructure technique** ou d'un **système pré-existant**.

La méthode “analytique”

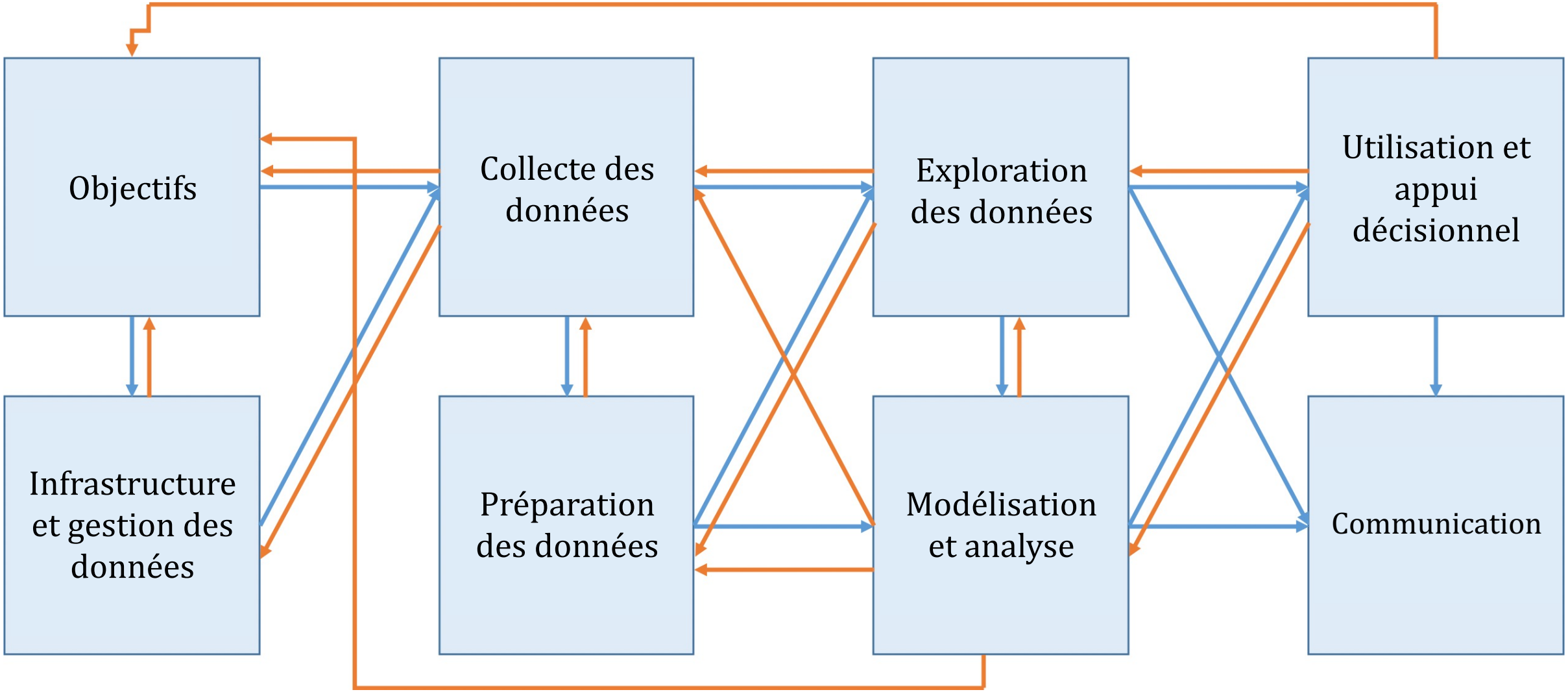
Comme c’est le cas pour la **méthode scientifique**, il existe un guide "étape par étape" pour l'analyse des données

- déclaration d'objectif
- collecte de données
- nettoyage des données
- analyse des données/analytique
- dissémination
- documentation

Notez que l'**analyse des données** ne constitue qu'un petit segment de l'ensemble du flux.

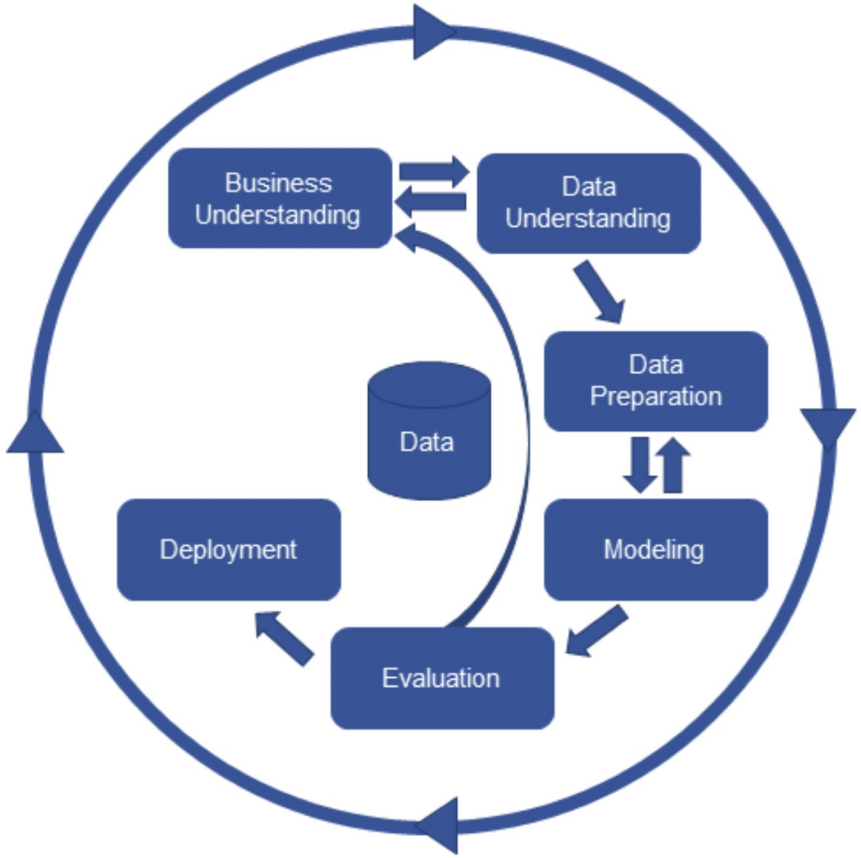
En pratique, le processus est souvent **désordonné** ; étapes ajoutées et retirées de la séquence, répétitions, reprises, etc.

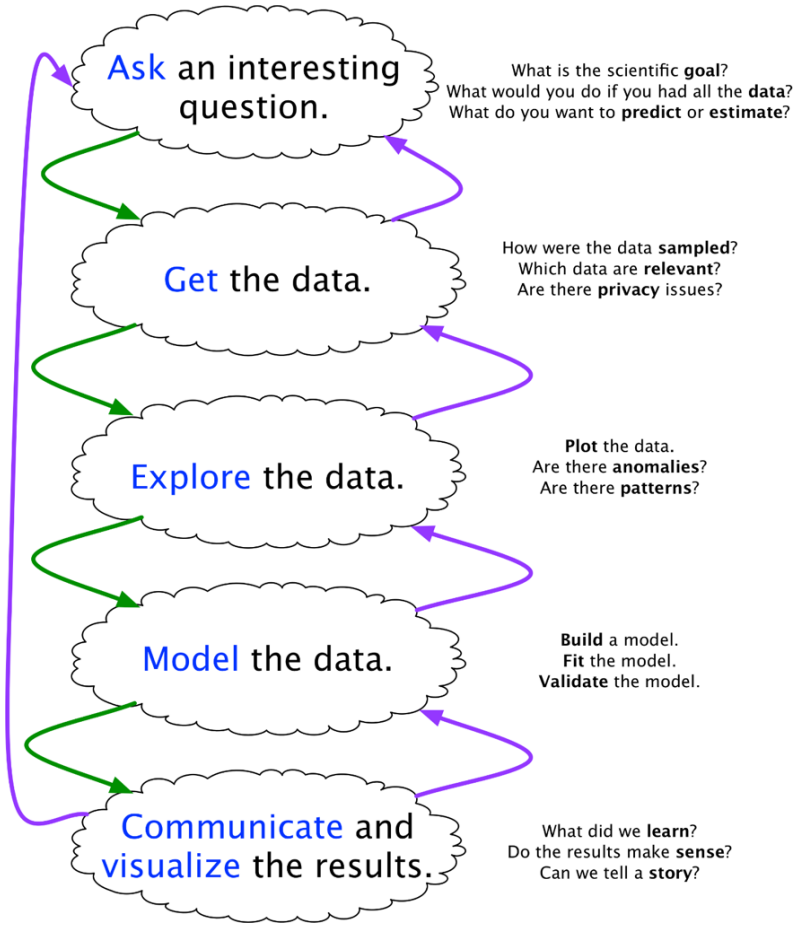
Cela a tendance à fonctionner…
quand **c'est mené correctement**.


Objectifs
Collecte des données
Exploration des données
Utilisation et appui décisionnel
Infrastructure et gestion des données
Préparation des données
Modélisation et analyse
Communication

[Cadre CRISP-DM (à gauche) ; flux de travail de science des données de Blitzstein et Pfister (à droite)].

La méthode “analytique”





La méthode “analytique”

En pratique, l'analyse des données est souvent corrompue par :

- le manque de clarté
- remaniement et travail inutile
- transfert aveugle vers TI
- pas d'itération

Les approches ont un noyau commun

- les projets sont **itératifs**
- (souvent) **non séquentiel**.

En aidant les parties prenantes à reconnaître cette **vérité centrale**, il est plus facile pour les scientifiques des données :

- d'obtenir des **informations utiles**

À retenir : il y a beaucoup de choses à prendre en compte avant la modélisation et l'analyse.

- **l'analyse des données ne se limite pas à l'*analyse* des données**

La collecte de données

Les données entrent dans le **pipeline de la science des données** en étant **collectées**.

Il existe plusieurs façons de procéder :

- les données peuvent être collectées en **un seul passage**
- elle peut être collectée par **lots** (“batches”)
- elle peut être collectée **en continu**

Le **mode d'entrée** peut avoir un impact sur les étapes suivantes, notamment sur la fréquence de **mise à jour des** modèles, des métriques, etc.

Le stockage des données

Une fois recueillies, les données doivent être **stockées**.

Les choix relatifs au stockage (et au **traitement**) doivent refléter :

- la manière dont les données sont recueillies (**mode d'entrée**)
- la quantité de données à stocker et à traiter (**petite ou grande**)
- le type d'accès/de traitement nécessaire (**quelle rapidité**, **quelle quantité**, **par qui**)

Les données stockées peuvent devenir **périmées** (*aux sens figuré* et *littéral*) ; il est recommandé de procéder à des audits réguliers des données.

Le traitement des données

Les données doivent être **traitées** avant de pouvoir être analysées.

Principalement, les **données brutes doivent** être converties dans un format qui **se prête à l'analyse**, en :

- identifiant les entrées **non valides**, **non fondées**, et **anormales**
- traitant les **valeurs manquantes**
- **transformant** les variables afin qu'elles répondent aux exigences des algorithmes choisis

L'**analyse** elle-même est presque anti-climatique : il suffit tout simplement d'exécuter les méthodes ou algorithmes sélectionnés sur les données traitées.

La modélisation

Les équipe de SD doivent connaître :

- le nettoyage les données
- les statistiques descriptives et la corrélation
- La probabilité et les statistiques inférentielles
- l'analyse de régression
- la classification et apprentissage supervisé
- le regroupement et appr. non supervisé
- la détection des anomalies et l'analyse des valeurs aberrantes
- les données massives/de hautes dimensions
- la modélisation stochastique, etc.

Cela ne représente qu'une **petite part** de l'analyse (cf. diapo précédente).

Aucun analyste ou scientifique des données ne peut tous les maîtriser (ou même une majorité d'entre eux) ; c'est l'une des raisons pour lesquelles la science des données est une **activité de groupe**.

Évaluation du modèle

Avant d'appliquer les résultats, nous devons d'abord confirmer que le modèle aboutit à des conclusions valables sur le système qui nous intéresse.

Les processus analytiques sont **réducteurs :** les données brutes sont transformées en **résumé numérique**, que nous espérons **lié** au système.

Les méthodologies de SD comprennent une **phase d'évaluation**

- contrôle “d'hygiène analytique” : y a-t-il quelque chose **qui cloche ?**

Méfiez-vous de la **tyrannie des succès précédents :** même si une approche a donné des réponses utiles par le passé, elle peut ne pas toujours le faire.

[http://dailycatdrawings.tumblr.com ; #309]

Le monde réel

Théorie

Identification des détails pertinents pour la **description** et la **traduction** des objets du monde réel en variables de modèle

Modèle

L’analyse de la vie après le modèle

Lorsqu'une analyse ou un modèle est "lâché dans la nature", il prend souvent une vie qui lui est propre. Lorsqu'il cesse inévitablement d'être **actuel**, les SD ne peuvent pas toujours faire grand-chose pour remédier à la situation.

Comment déterminer si le modèle de données actuel est :

- **démodé** ?
- n'est plus **utile** ?
- combien de temps faut-il à un modèle pour réagir à un **changement conceptuel** ?

Des audits réguliers peuvent être utilisés pour répondre à ces questions.

L’analyse de la vie après le modèle

Les SD ont rarement le contrôle total de la **diffusion des modèles**.

- les résultats peuvent être détournés, mal compris, mis de côté, ou ne pas être mis à jour
- les analystes consciencieux peuvent-ils faire quelque chose pour l’empêcher ?

Il n'y a pas de réponse facile : on ne doit pas seulement se concentrer sur l'analyse, mais aussi reconnaître les opportunités qui se présentent pour **éduquer les parties prenantes** sur l'importance des étapes auxiliaires.

En raison de la **déclin analytique**, la dernière étape du processus analytique n'est pas une **impasse**, mais une invitation à retourner au début du processus.

Pipelines de données

Dans le **contexte de la prestation de services**, le processus d'analyse des données est mis en œuvre sous forme de **pipeline de données automatisé** pour permettre des exécutions automatiques.

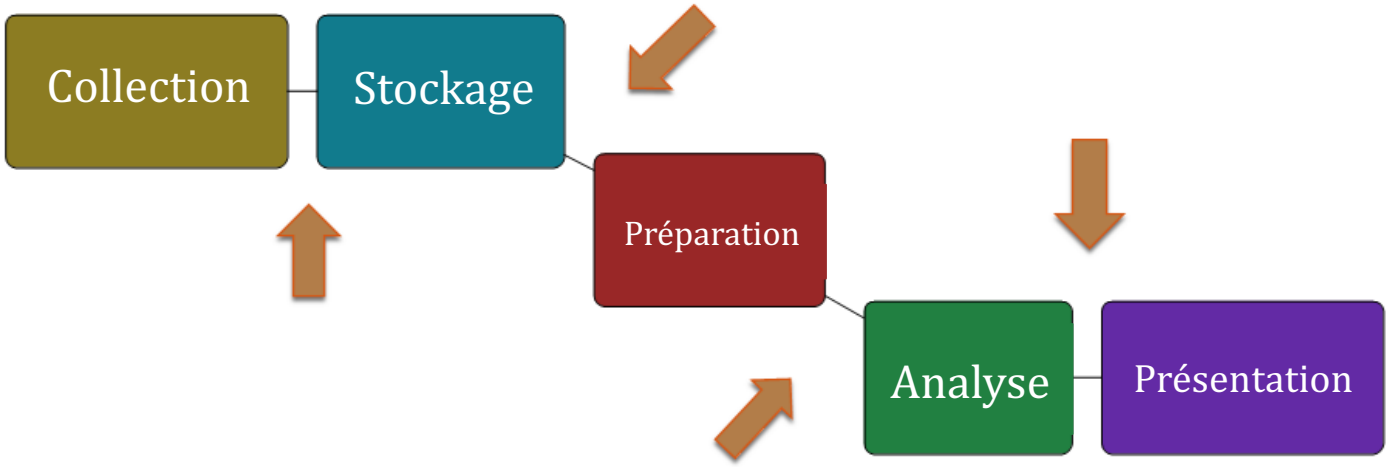
Les pipelines de données se composent généralement de 9 éléments (5 **étapes** et 4 **transitions**) :

- collecte de données
- stockage de données
- préparation des données
- analyse des données
- présentation des données

Pipelines de données

Chaque composant doit être **conçu** et ensuite **mis en œuvre**.

Généralement, au moins une passe d'analyse des données doit être effectué **manuelle-ment** avant que l'implementation ne soit terminée.



Lectures suggérées

Le flux de travail analytique

Data Understanding, Data Analysis, Data Science
Data Science Basics

Analytics Workflows

- The "Analytical" Methods
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercices

Le flux de travail analytique

1. Installez R / RStudio (Posit), et les librairies de la liste fournie par l'instructeur.

2. Testez l'installation à l'aide des exemples du Programming Primer (sections 2 - 4) pour vous assurer que le logiciel fonctionne comme prévu.