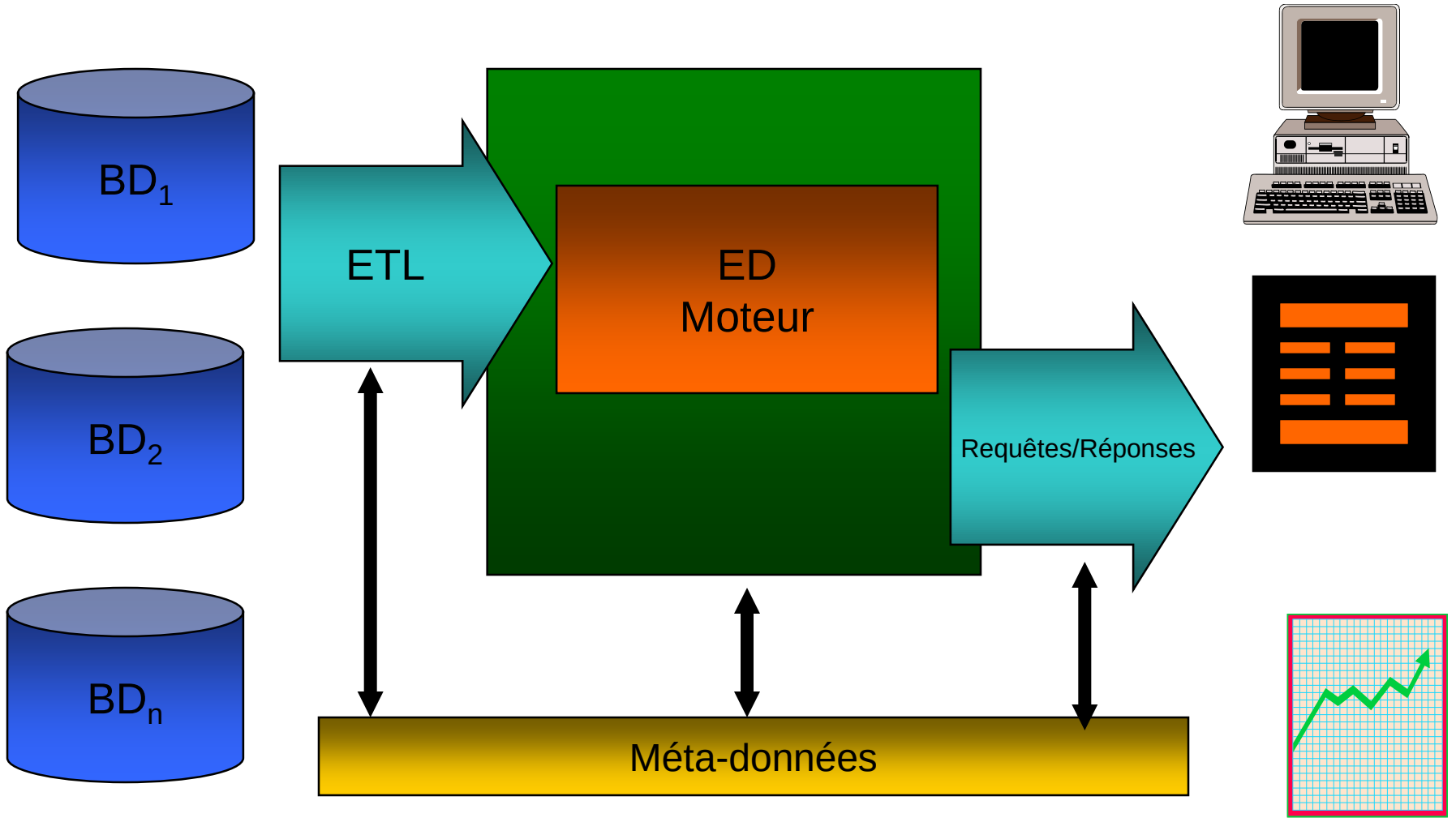
# Qualité des données

## *Introduction*

*Mohand-Saïd Hacid, UCBL, LIRIS CNRS*

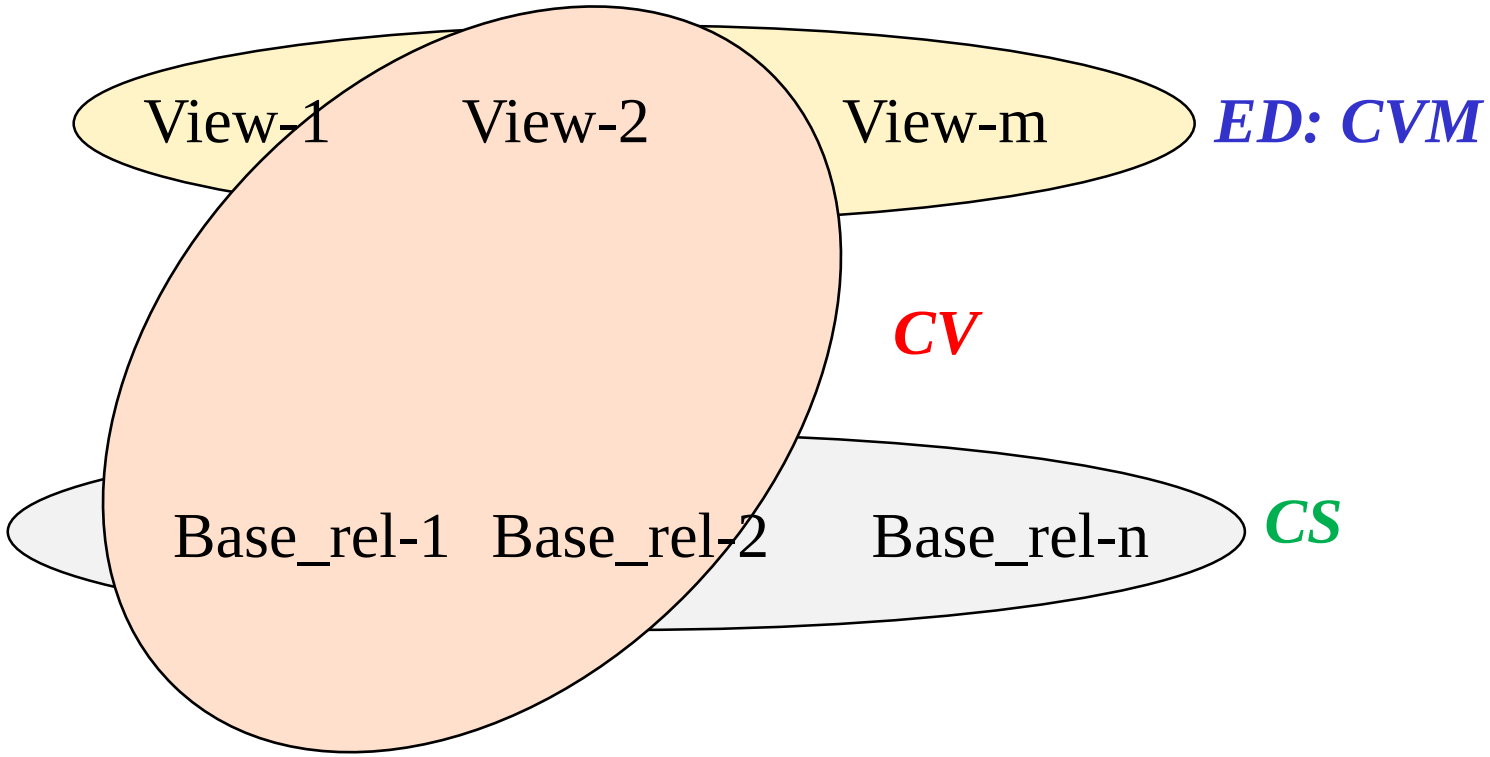
# Architecture ED



## *Consistance*

Trois niveaux :

- Source
- Vue unique
- Vues multiples



# Le problème

**Pour traiter efficacement la question de la qualité des données, nous devons être en mesure de gérer :**

- ❑ L'identification des attentes des clients en matière de qualité des données
- ❑ La définition de mesures contextuelles
- ❑ L'évaluation des niveaux de qualité des données (EQD)
- ❑ Suivi des problèmes pour la gestion des processus
- ❑ Détermination des meilleures opportunités d'amélioration
- ❑ Élimination des sources de problèmes
- ❑ Mesure continue pour l'amélioration par rapport à la base de référence

# Cadre pour la qualité des données

## Qualité des données

**La qualité des données comporte six dimensions, à savoir :**

- ❑ l'exactitude
- ❑ l'exhaustivité
- ❑ la cohérence
- ❑ disponibilité
- ❑ la validité
- ❑ caractère unique (unicité)

**Chacune de ces dimensions de la qualité des données évalue un aspect unique des données, en testant la capacité des données à répondre à l'objectif visé.**

- ❑ Exactitude : dans quelle mesure les données reflètent-elles la réalité ?
- ❑ Exhaustivité : y a-t-il des informations clés manquantes dans les données ?
- ❑ Cohérence : toutes les valeurs des données suivent-elles un formatage cohérent et correspondent-elles correctement à ce formatage ?
- ❑ Disponibilité : les données sont-elles disponibles au moment où elles sont nécessaires et attendues ? Validité : les données sont-elles présentées dans le format attendu ? Par exemple, les courriels comportent-ils le symbole "@" ?
- ❑ Unicité : y a-t-il des doublons dans l'ensemble de données ?

# Politiques de qualité des données

Orienter les activités de gestion des données vers la gestion des aspects de conformité aux directives de l'entreprise, tels que :

- ❑ La certification des données
- ❑ La gestion de la confidentialité
- ❑ Lignée des données
- ❑ Limitation de l'utilisation
- ❑ Source de référence unifiée

*Data Lineage en français "lignée des données" est un* ***processus qui vise à fournir une cartographie du système d'information****. Il permet une visualisation du cycle de vie de la donnée en vue de répondre aux questions suivantes : de quelle source provient cette donnée, et quelles transformations a-t-elle subies.*

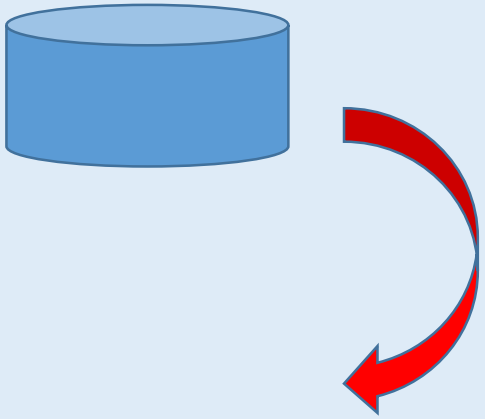
# Procédures relatives à la qualité des données

Les processus de gestion de la qualité des données supposent le respect des politiques de qualité des données. Exemples :

- ❑ Modèles normalisés d'inspection des données
- ❑ Qualité opérationnelle des données
- ❑ Suivi des problèmes et remédiation
- ❑ Intervention manuelle si nécessaire
- ❑ l'intégrité de l'échange de données
- ❑ Plan d'urgence
- ❑ Validation des données

# Attentes des entreprises et qualité des données

**Règles de qualité des données**

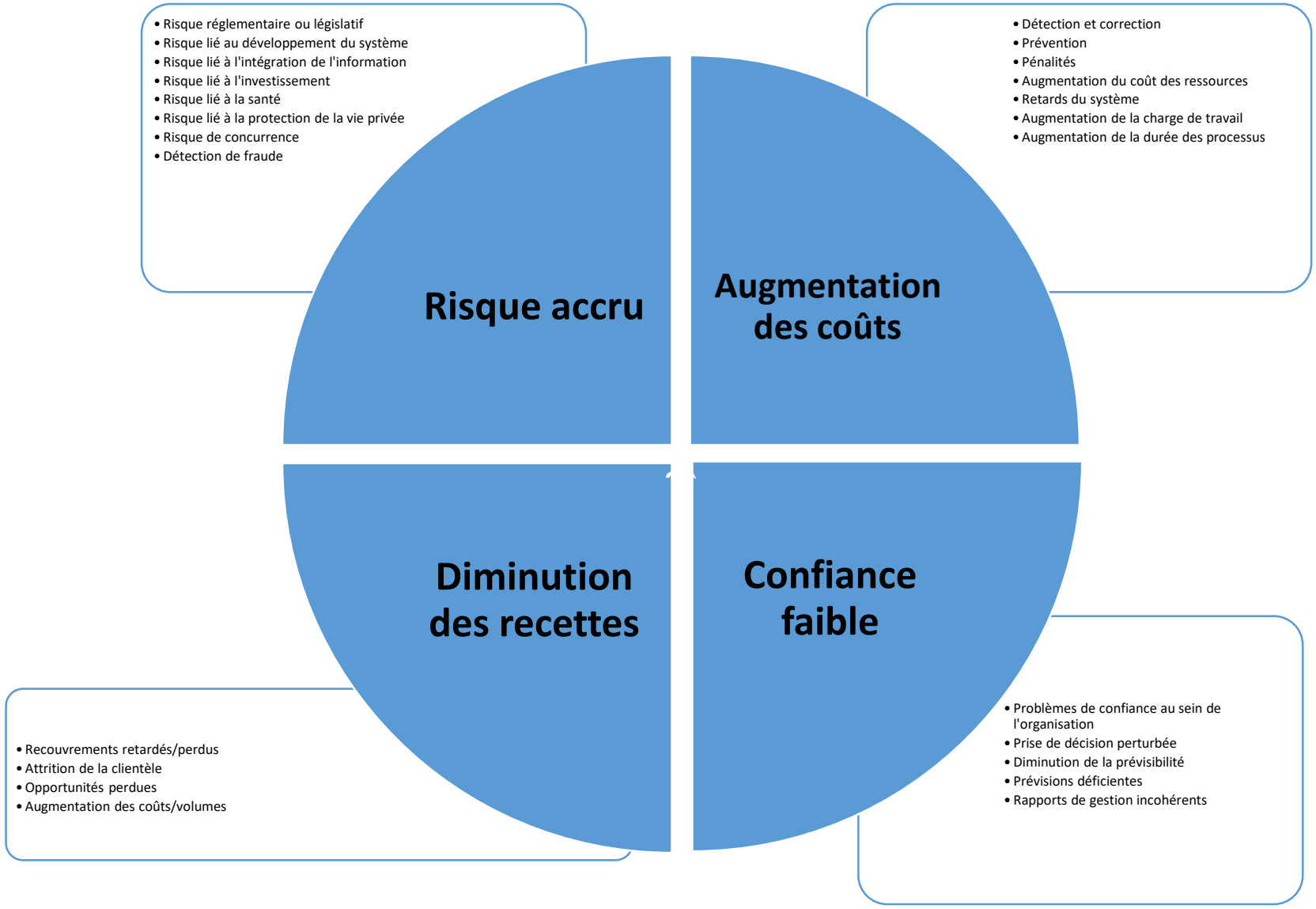


Doublons
Incohérences
Valeurs manquantes
Données inutilisables
…

**Attentes des entreprises**

Rendement
Échec des transactions
Réponse aux opportunités
…

# Impact sur les entreprises



# Exemples - Augmentation des coûts

- Grande entreprise de fournitures de bureau :
  - ✓ Les données redondantes et les données inutilisées représentaient un pourcentage important de l'utilisation de l'espace de stockage.
  - ✓ L'élimination des données inutilisées ou redondantes entraînerait une réduction significative (20 %) des coûts de SDAD (et de la gestion des données correspondantes).

- Lignes directrices du ministère de la défense sur la qualité des données (USA) :
  - ✓ "... l'incapacité à faire correspondre les registres de paie avec le dossier d'emploi officiel a couté cher en trop-perçus …",
  - ✓ "... l'incapacité à établir une corrélation entre les bons de commandes et les factures constitue un problème majeur en ce qui concerne les décaissements non concordants."

- Entreprise manufacturière :
  - ✓ L'incapacité à déterminer si des composants similaires ont déjà été conçus et construits a entraîné la duplication des coûts de conception et de développement dépassant 70 000 dollars par article

# Exemples - Baisse des revenus

- Entreprise de télécommunications :
  - ✓ Application de la garantie des recettes pour détecter la sous-facturation conduisent à une perte de revenus d'un peu plus de 3 % des revenus totaux due à une mauvaise qualité des données
  - ✓ Identification de 49 circuits à haut débit mal configurés (mais supposés inutilisables) qui pourraient être remis en service (au lieu de les remplacer/produire à nouveau).
- Agence fédérale :
  - ✓ "55 % des enregistrements d'une base de données de bâtiments étaient erronés, ce qui a entraîné la sous-facturation de loyers de 12 millions de dollars".
- Une autre agence fédérale :
  - ✓ Des adresses de contact périmées ralentissent le processus de recouvrement des impayés.

# Exemples - Baisse de confiance

- Entreprise pharmaceutique :
  - ✓ Investissement important dans la création d'une application de vente frontale alimentée par une base de données dorsale.
  - ✓ Les clients de l'application ont refusé d'utiliser la nouvelle application en raison de leur méfiance à l'égard de la base de données.

- Entreprise agricole :
  - ✓ Plusieurs bases de données de vente étaient en conflit avec les bases de données comptables
  - ✓ Les vendeurs n'étaient pas convaincus que leurs commissions étaient calculées correctement

# Exemples - Risque accru

- Entreprise de produits pharmaceutiques/appareils médicaux
  - ✓ Base de données utilisée pour gérer les bénéficiaires de subventions
  - ✓ Les bénéficiaires peuvent également être des fournisseurs
  - ✓ L'incapacité à réaliser un meilleur suivi des bénéficiaires expose l'entreprise au risque de violation des lois anti-corruption

- Secteur bancaire, risque de crédit :
  - ✓ "PWC (https://www.pwc.fr/) estime que 90 % des 100 premières banques mondiales sont déficientes dans la gestion des données relatives au risque de crédit…"

# Le rôle de l'évaluation de la qualité des données dans un projet

## Ce qu'il faudrait comprendre/retenir :

- ❑ Pourquoi l'EQD (Evaluation de la Qualité des Données) est importante et comment elle peut être appliquée à des projets
- ❑ Les étapes de l'EQD et l'objectif de chacune d'entre elles
- ❑ L'application de l'EQD à un ensemble de données
- ❑ L'interprétation des statistiques de base et des graphiques simples
- ❑ *Les différents outils logiciels et autres ressources permettant de réaliser l'EQD*

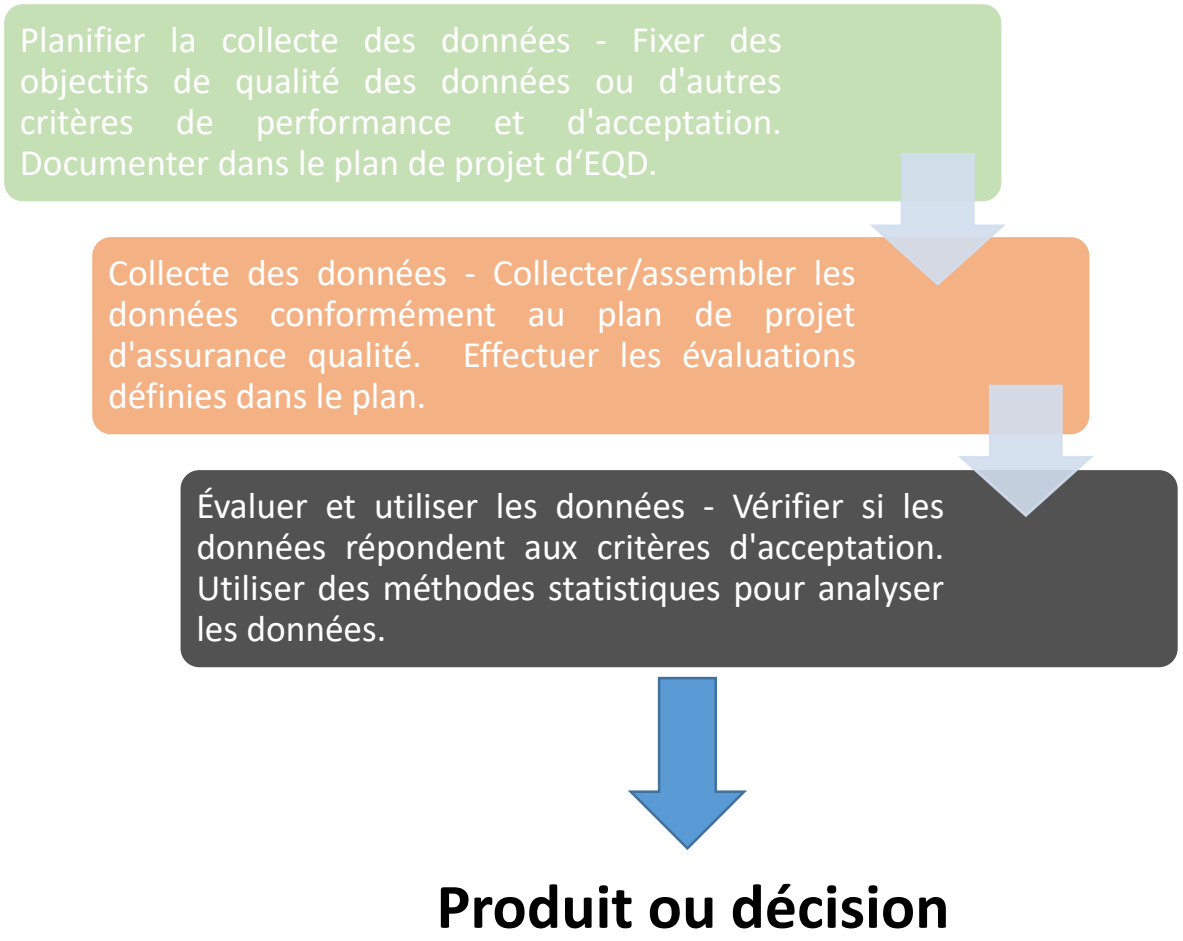
# Qualité des données

- Significative uniquement lorsque la "qualité des données" est liée à l'utilisation prévue des données.
- Certaines données sont bonnes ("*haute qualité*") pour certains usages, mais mauvaises ("*basse qualité*") pour d'autres.

# Evaluation de la Qualité des Données (EQD)

- Une évaluation/approche scientifique et statistique visant à déterminer si les données sont pertinentes pour l'usage auquel elles sont destinées.
- L'**EQD** est décrite, p.e., dans le document *Guidance for Data Quality Assessment : Practical Methods for Data Analysis*

# Le cycle de vie d’un projet

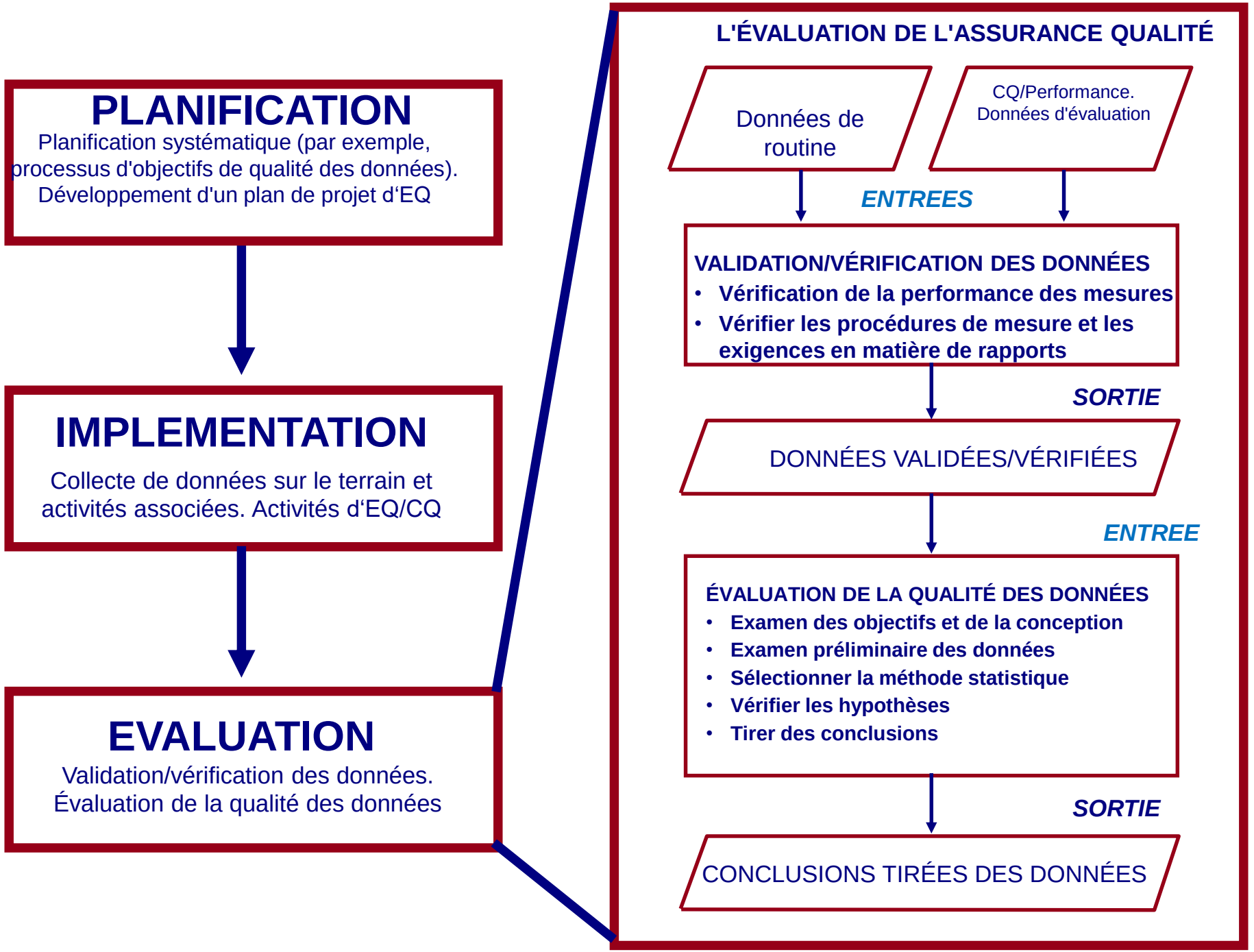


# L’évaluation de la qualité des données est réalisée

Chaque fois que les données sont utilisées pour prendre une décision, pour une estimation ou à des fins de recherche.

Ceci s'applique à :

- Nouvelles données à collecter
- Données collectées par quelqu'un d'autre
- Données collectées par vous dans le cadre d'un autre projet


PLANIFICATION
Planification systématique (par exemple, processus d'objectifs de qualité des données).
Développement d'un plan de projet d'EQ
IMPLEMENTATION
Collecte de données sur le terrain et activités associées. Activités d'EQ/CQ
EVALUATION
Validation/vérification des données.
Évaluation de la qualité des données
L'ÉVALUATION DE L'ASSURANCE QUALITÉ
Données de routine
CQ/Performance. Données d'évaluation
ENTREES
VALIDATION/VÉRIFICATION DES DONNÉES
• Vérification de la performance des mesures
• Vérifier les procédures de mesure et les exigences en matière de rapports
SORTIE
DONNÉES VALIDÉES/VÉRIFIÉES
ENTREE
ÉVALUATION DE LA QUALITÉ DES DONNÉES
• Examen des objectifs et de la conception
• Examen préliminaire des données
• Sélectionner la méthode statistique
• Vérifier les hypothèses
• Tirer des conclusions
SORTIE
CONCLUSIONS TIRÉES DES DONNÉES

# Vérification *vs.* Validation *vs.* Evaluation

- **Vérification des données** - processus d'évaluation de l'exhaustivité, de l'exactitude et de la conformité d'un ensemble de données spécifique par rapport à la méthode, aux procédures ou aux exigences contractuelles.

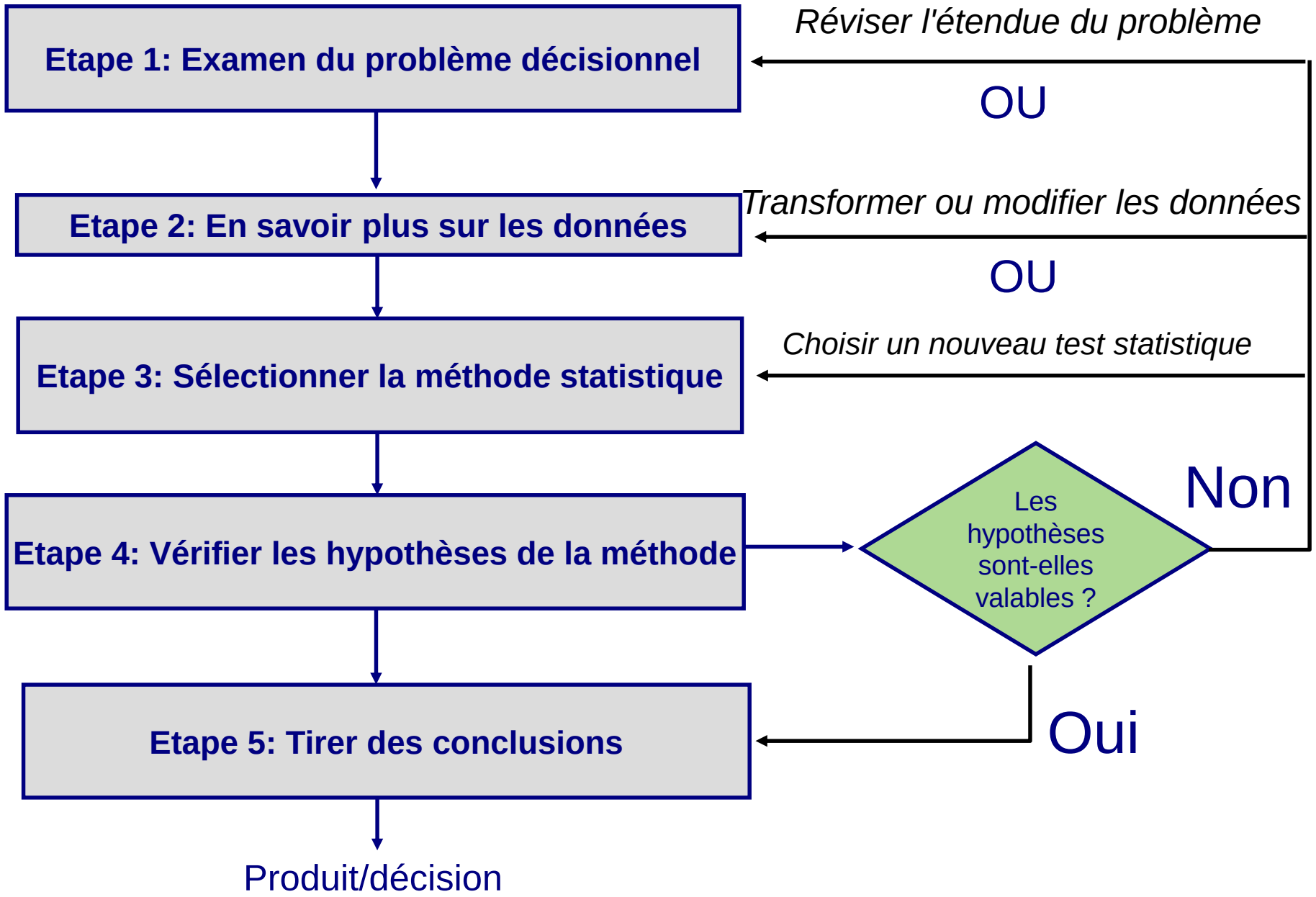
- **Validation des données** - processus spécifique à l'analyste et à l'échantillon qui étend l'évaluation des données au-delà de la méthode, de la procédure ou de la conformité contractuelle (c'est-à-dire la vérification des données) afin de déterminer la qualité analytique d'un ensemble de données spécifique.

- **Évaluation de la qualité des données** - processus visant à déterminer si les données sont adaptées à une utilisation spécifique.

# Les cinq étapes de l'évaluation de la qualité des donnée

1. Examiner les objectifs et le plan d'échantillonnage
2. Procéder à un examen préliminaire des données
3. Choisir la méthode statistique
4. Vérifier les hypothèses de la méthode statistique
5. Tirer des conclusions à partir des données

# Evaluation de la Qualité des Données



# EQD : Deux points de vues

## Point de vue des décideurs

1. Définir la règle de décision et les erreurs de décision
2. Spécifier les limites acceptables des erreurs de décision
3. Identifier la méthode d'application de la règle de décision
4. S'assurer que la méthode est défendable
5. Appliquer la règle de décision

## Point de vue des analystes de données

1. Définir les hypothèses statistiques
2. Déterminer les taux d'erreurs (de type I et de type II ?) acceptables
3. Identifier le test ou la méthode statistique et les hypothèses
4. Évaluer la validité du test ou de la méthode statistique
5. Réaliser le test statistique et évaluer la conception

# EQD – Table de synthèse

| Objectif du projet & conception de la collecte des données (**Etape 1**) | Observations - rapports sur l'évaluation de la qualité, rapport, syntheses, Statistiques et graphiques (**Etape 2**) | Méthodes statistiques et hypotheses (**Etape 3**) | Vérification des hypothèses (**Etape 4**) | Résultats issus de la méthode statistique (**Etape 5**) |
|---|---|---|---|---|
| LISTE:<br>- Objectif<br>- Paramètres d'intérêt<br>- Type d'analyse dont on a besoin<br>- Type de collection de données<br>- Information sur les deviations (par rapport à la conception) dans l'implémentation | LISTE :<br>- Non-détectés<br>- Distribution probable<br>-Valeurs aberrantes potentielles<br>- Anomalies | LISTE :<br>- Méthode d'analyse<br>- Hypothèses à vérifier<br>- Niveaux de signification | LISTE :<br>- Hypothèses, si elles ont été observées et comment elles ont été vérifiés (y compris les niveaux de signification) | LISTE :<br>- Résultats finaux issus de l'analyse de données<br>- Autres facteurs affectant le produit final ou la décision |
| [Cette colonne contiendra une vue d'ensemble du projet et des informations de base permettant de déterminer la "qualité"]. | [Cette colonne contiendra des informations qui permettront de déterminer quelles hypothèses pourraient être satisfaites]. | [Cette colonne contient des informations sur la méthode statistique et ses hypothèses]. | [Cette colonne décrira les hypothèses vérifiées, la manière dont elles ont été vérifiées et les résultats obtenus]. | [Cette colonne résume les résultats finaux du test statistique et les autres facteurs à prendre en compte dans le produit ou la décision finale]. |

# Les cinq étapes de l'évaluation de la qualité des donnée

1. Examiner les objectifs et le plan d'échantillonnage
2. Procéder à un examen préliminaire des données
3. Choisir la méthode statistique
4. Vérifier les hypothèses de la méthode statistique
5. Tirer des conclusions à partir des données

# Étape 1 : Examen des objectifs et conception de la collecte des données

- ❑ Traduire les objectifs de l'utilisateur des données en un énoncé de l'hypothèse statistique primaire ou de l'objectif de l'estimation.
- ❑ Traduire les objectifs de l'utilisateur des données en limites tolérables de la probabilité de commettre des erreurs de décision.
- ❑ Examiner le plan d'échantillonnage et noter toute particularité ou tout écart par rapport au plan d'échantillonnage.

## Etape 1 : Entrée

- ❑ Plan de projet d'EQ ou tout autre document de planification qui contient :
  - ✓ l'objectif du projet ou la question à laquelle il faut répondre
  - ✓ les critères de décision et de performance (CDP) ou d'autres critères de performance et d'acceptation
- ❑ Plan d'échantillonnage sur le terrain et tout rapport sur la mise en œuvre effective du plan d'échantillonnage

# Si une planification systématique a été effectuée...

Utiliser les rapports documentant la planification pour répondre aux questions suivantes :

- ❑ Quel est l'objectif du projet ?
- ❑ Quels sont les critères de performance ou d'acceptation du produit ou de la décision ?

# Si la planification systématique n'a PAS été effectuée...

**Prise de décision :** Appliquer le processus relatif aux objectifs de qualité des données ou un autre processus de pour :

- ❑ développer des hypothèses
- ❑ définir les erreurs de décision potentielles
- ❑ spécifier les limites tolérables des erreurs de décision

**Estimation :** Utiliser un processus de planification systématique pour

- ❑ sélectionner des paramètres
- ❑ développer des critères de performance ou d'acceptation

# Etape 1 : sortie

- ❑ Objectifs et critères du projet bien définis
- ❑ Vérifier que l'hypothèse choisie est cohérente avec l'objectif et les critères
- ❑ Une liste de tous les écarts par rapport au plan d'échantillonnage prévu et les effets de ces écarts.

# Étape 2 de l’EQD : Examen préliminaire des données

- ❑ Examiner les rapports d'assurance qualité pour détecter les anomalies
- ❑ Calculer des quantités statistiques standard
- ❑ Afficher les données à l'aide de représentations graphiques

# Etape 2 : entrée

- ❑ Données vérifiées et validées
- ❑ Rapports EQ, CQ données
- ❑ Résultats des audits des systèmes techniques :
  - ✓ Évaluations des performances
  - ✓ Rapports sur les mesures correctives
  - ✓ Rapports de vérification et de validation des données
- ❑ Plan de projet d'EQ, plan d'échantillonnage et d'analyse ou autres documents de planification

## ❑ **Rechercher** :

- ❑ Non-respect des critères d'acceptation/violations évidentes du contrôle de la qualité :
  - ✓ limites de détection variables
  - ✓ méthodes d'analyse non équivalentes

- ❑ Anomalies dans la mise en œuvre du plan de projet d‘EQ :
  - ✓ taux d'émission négatifs
  - ✓ valeurs de pH (potentiel hydrogène) supérieures à 14,0
  - ✓ Valeurs exprimées dans des unités de déclaration erronées

# ❑ Calculer des quantités statistiques standard

- Les quantités statistiques comprennent les mesures de :
  - ✓ la tendance centrale (moyenne, médiane, etc.)
  - ✓ la position relative
  - ✓ Dispersion (étendue, variance)
  - ✓ Association (corrélation)

- Examinez ces quantités pour déterminer :
  - ✓ Les données semblent-elles raisonnables - les valeurs ont-elles un sens ?
  - ✓ Y a-t-il des anomalies évidentes ?
  - ✓ Existe-t-il des tendances ou des modèles/motifs ?

# ❑ Afficher les données à l'aide de graphiques

- Graphiques courants :
  - ✓ Histogramme
  - ✓ Diagramme de dispersion
  - ✓ Diagramme du temps

- Examinez les graphiques pour déterminer :
  - ✓ Les données semblent-elles raisonnables ?
  - ✓ À quoi ressemble la distribution ?
  - ✓ Est-elle symétrique, bimodale ?
  - ✓ Y a-t-il des valeurs extrêmement élevées ou extrêmement basses ?
  - ✓ Y a-t-il des tendances évidentes ?

# Etape 2 :  Sortie

- Quantités statistiques et graphiques qui vous permettent d'avoir une première compréhension des données et de tout problème potentiel, y compris :
  - ✓ la distribution des données
  - ✓ Les valeurs aberrantes potentielles
  - ✓ les « *non-detects* »

# Étape 3 de l’EQD : Sélection de la méthode statistique

- ❑ Sélectionner la méthode statistique en fonction des objectifs de l'utilisateur des données et de l'examen préliminaire des données.

- ❑ Identifier les hypothèses sous-jacentes à la méthode statistique

# Étape 3 : Entrée

- ❑ Objectifs du projet, hypothèses et méthode statistique préliminaire si elle a été identifiée
- ❑ Contexte des méthodes statistiques

# Sélectionner la méthode

- Si elle a été identifiée lors de la planification, déterminer si ce choix semble raisonnable sur la base de l'examen préliminaire des données.

- Sinon, sélectionner la méthode statistique en fonction des objectifs de l'utilisateur des données et de l'examen préliminaire des données.

- Exemples de méthodes :
  - Tests : Test t à un échantillon, test t à deux échantillons, test pour une proportion unique, test de rang signé de Wilcoxon.
  - Estimation
  - Analyse de régression
  - Analyse des séries temporelles

# Identifier les hypothèses

- Chaque méthode repose sur des hypothèses.

- Hypothèses courantes :
  - Forme de distribution
  - Indépendance
  - Caractéristiques de dispersion
  - Homogénéité
  - Base de la randomisation

- **Exemple - Test t à un échantillon** : échantillon aléatoire, indépendance des données ; la moyenne de l'échantillon est normalement distribuée ; pas de valeurs aberrantes ; peu de "non-détections".

# Étape 3 : Sortie

- ❑ Méthode statistique proposée qui semble appropriée pour les données et les objectifs du projet
- ❑ Liste des hypothèses pour la méthode statistique

# Étape 4 de l’EQD : vérifier les hypothèses de la méthode statistique

- ❑ Déterminer l'approche pour la vérification des hypothèses
- ❑ Effectuer des tests d'hypothèses
- ❑ Si nécessaire, déterminer les mesures correctives à prendre

# Étape 4 : Entrée

- ❑ Données
- ❑ Hypothèses identifiées pour la méthode statistique
- ❑ Méthodes de vérification de ces hypothèses, avec leurs formules

## Déterminer la méthode de vérification des hypothèses et effectuer le test

- ❑ Évaluer l'étape 3 pour déterminer les hypothèses à vérifier
- ❑ Déterminer les tests disponibles pour vérifier les hypothèses pour cet ensemble de données
- ❑ Sélectionner le test et le niveau de signification approprié

# Déterminer les actions correctives

Si cet ensemble de données ne répond pas aux hypothèses requises, déterminer les prochaines étapes à suivre :

- ❑ Répéter l'étape 3 et sélectionner une autre méthode d'analyse des données
- ❑ Transformer les données
- ❑ Réduire le «niveau de signification»
- ❑ Collecter des données supplémentaires
- ❑ Modifier l'objectif

. . . Mais cela doit être fait avec prudence

## Etape 4 : Sortie

- ❑ Documentation de la méthode utilisée pour vérifier chaque hypothèse et les résultats de ces méthodes

- ❑ Mesures correctives (si nécessaire)

# Étape 5 de l’EQD : Tirer des conclusions à partir des données

- ❑ Effectuer les calculs pour la méthode statistique
- ❑ Évaluer les résultats et tirer des conclusions

# Étape 5 : Entrée

- ❑ Données
- ❑ Objectif, hypothèses (le cas échéant) et critères de performance ou d'acceptation
- ❑ Formules pour la méthode statistique
- ❑ Facteurs non statistiques à intégrer dans la décision ou le produit final

# Exécuter la méthode statistique

- Utiliser les formules et les procédures des manuels standard
- Utiliser un logiciel pour effectuer les calculs :
  - SAS ou Splus
  - DataQUEST
  - …

- ***Remarque*** : le logiciel peut être utilisé sur n'importe quelles données, que les hypothèses aient été vérifiées ou non. Mais lorsque les hypothèses ne sont pas vérifiées, les résultats sont souvent suspects.

# Évaluer les résultats

Les résultats statistiques ne sont pas nécessairement la réponse.  Tenir compte d'éléments tels que

- Importance pratique
- Facteurs politiques/sociaux
- Importance contextuelle

## Etape 5 : Sortie

- ❑ Résultats statistiques avec un niveau de signification spécifié
- ❑ Produit final ou décision

| |
|---|
| Exhaustivité |
| Cohérence logique |
| Précision de position |
| Précision thématique ou précision sémantique |
| Qualité temporelle |
| Utilisabilité |

| |
|---|
| C'est le degré de cohérence avec un ensemble spécifique d'exigences de qualité des données **(1)** |
| C'est la précision de positionnement des données sur la Terre **(2)** |
| C'est la conformité des valeurs des éléments du jeu de données avec les valeurs de leurs homologues dans le terrain nominal **(3)** |
| C'est la qualité des attributs temporels et des relations temporelles entre objets **(4)** |
| C'est le degré de cohérence interne des données selon des règles de modélisation et les règles inhérentes à la spécification de produit du jeu de données **(5)** |
| C'est la conformité de la présence ou de l'absence des éléments du jeu de données par rapport au terrain nominal **(6)** |

| | |
|---|---|
| Exhaustivité **(6)** | C'est le degré de cohérence avec un ensemble spécifique d'exigences de qualité des données **(1)** |
| Cohérence logique **(5)** | C'est la précision de positionnement des données sur la Terre **(2)** |
| Précision de position **(2)** | C'est la conformité des valeurs des éléments du jeu de données avec les valeurs de leurs homologues dans le terrain nominal **(3)** |
| Précision thématique ou précision Sémantique **(3)** | C'est la qualité des attributs temporels et des relations temporelles entre objets **(4)** |
| Qualité temporelle **(4)** | C'est le degré de cohérence interne des données selon des règles de modélisation et les règles inhérentes à la spécification de produit du jeu de données **(5)** |
| Utilisabilité **(1)** | C'est la conformité de la présence ou de l'absence des éléments du jeu de données par rapport au terrain nominal **(6)** |

| | |
|---|---|
| Le champ d'application de la mesure | C'est l'expression du résultat de la mesure de qualité **(1)** |
| Le type de mesure effectuée | C'est la date à laquelle est effectuée le contrôle **(2)** |
| La procédure utilisée pour la mesure | C'est la description de la formule utilisée **(3)** |
| Le type de valeur et l'unité | C'est ce que l'on s'engage à mesurer **(4)** |
| La date du contrôle | C'est la méthode employée pour détecter les anomalies éventuelles **(5)** |

| | |
|---|---|
| Le champ d'application de la mesure **(4)** | C'est l'expression du résultat de la mesure de qualité **(1)** |
| Le type de mesure effectuée **(3)** | C'est la date à laquelle est effectuée le contrôle **(2)** |
| La procédure utilisée pour la mesure **(5)** | C'est la description de la formule utilisée **(3)** |
| Le type de valeur et l'unité **(1)** | C'est ce que l'on s'engage à mesurer **(4)** |
| La date du contrôle **(2)** | C'est la méthode employée pour détecter les anomalies éventuelles **(5)** |


Fin