# Risque de ré-identification : deux questions pratiques relatives au critère de la *l*-diversité

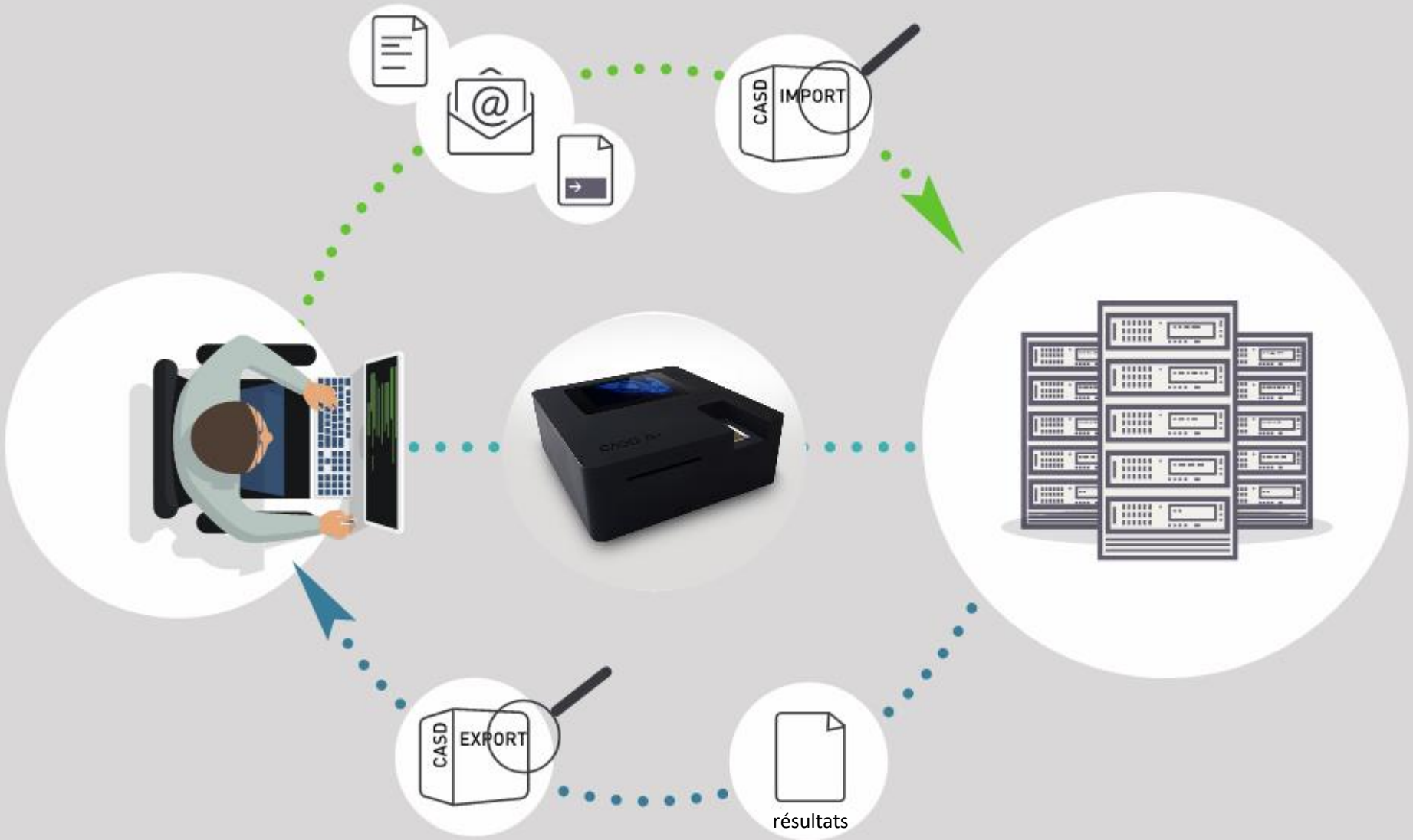
Séminaire de méthodologie statistique – INSEE – 24 juin 2019

**Kamel Gadouche**, CASD – **Dominique Blum**, CASD – **Caroline Boucly**, DESE Insee

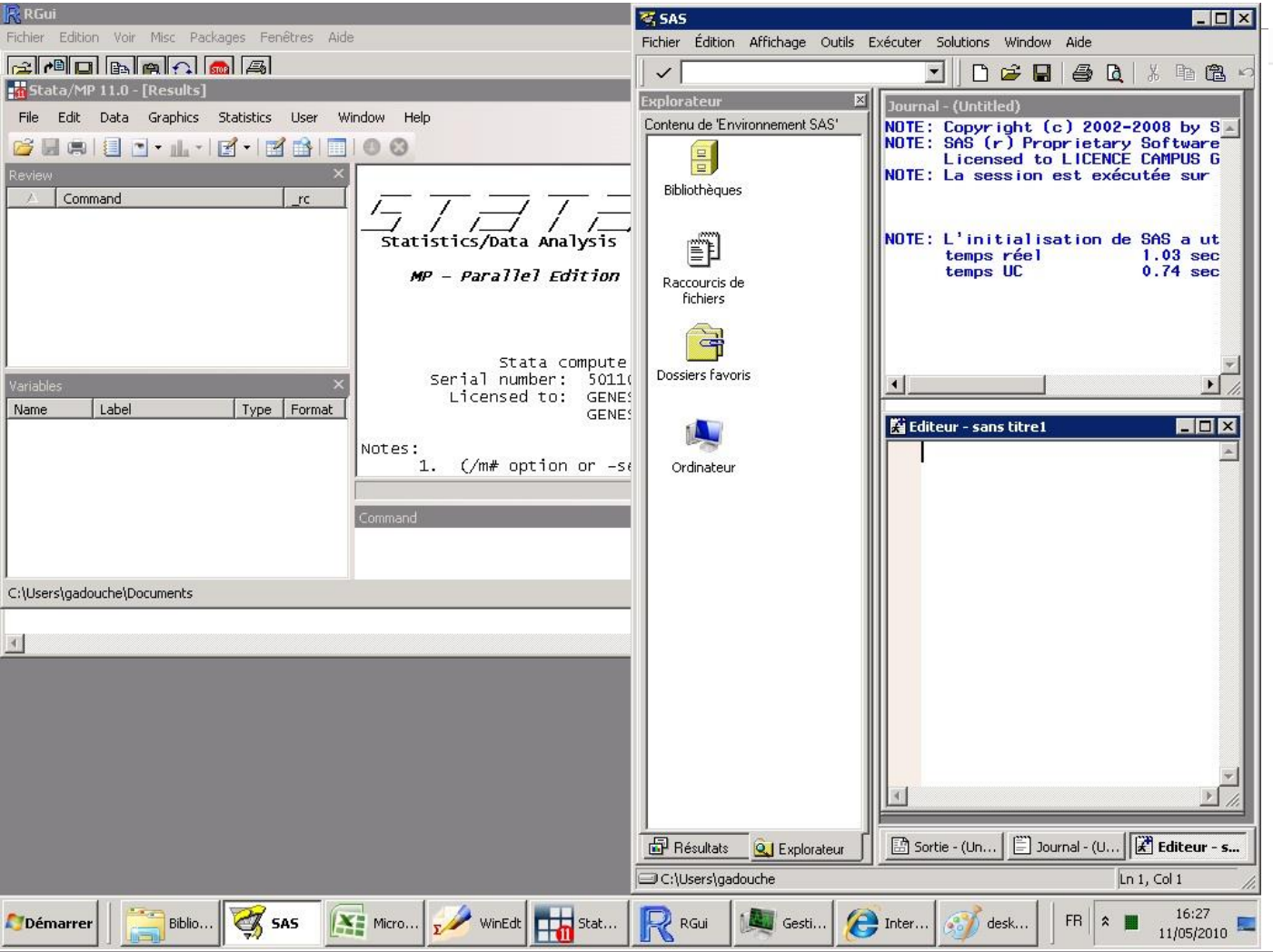
# Le contexte

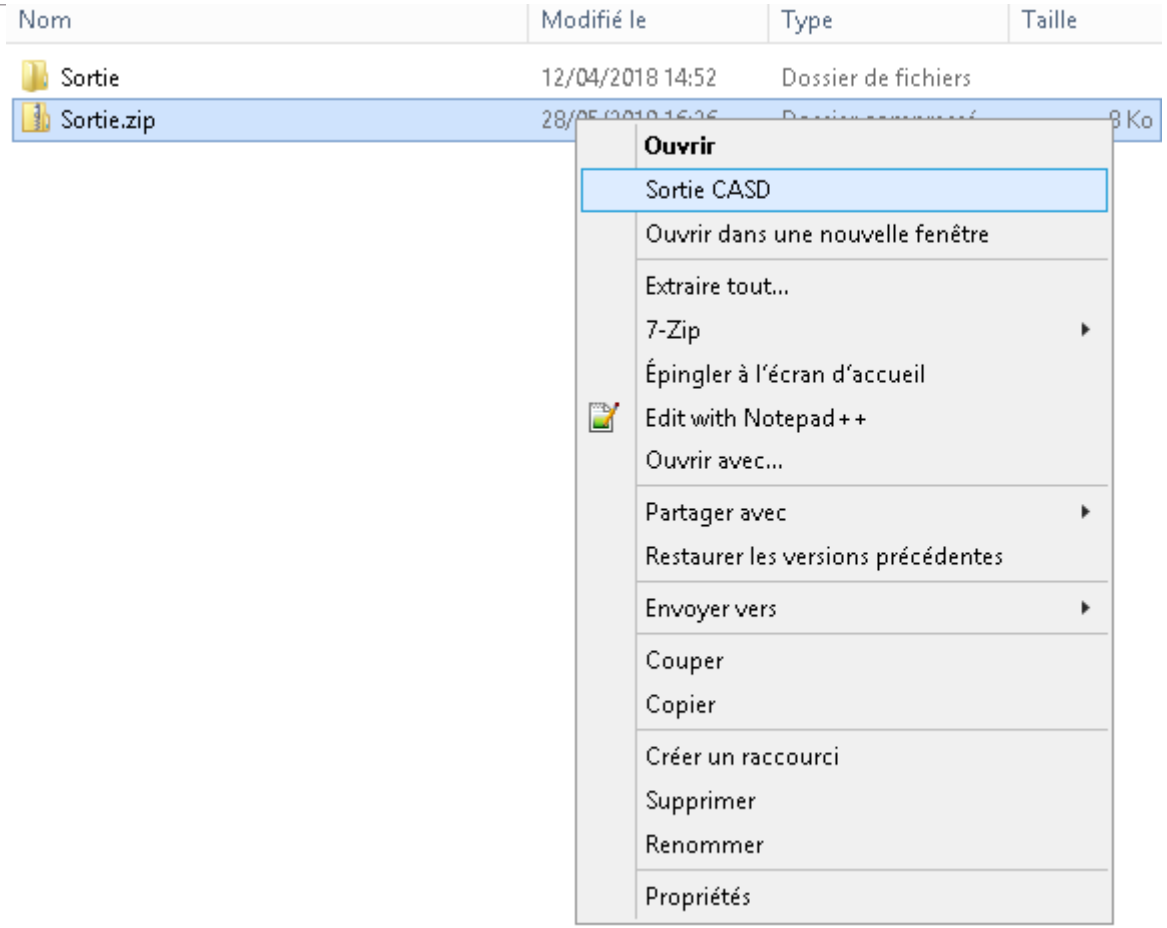
# Le CASD en quelques mots

- Le CASD est un équipement qui permet à des utilisateurs, principalement des chercheurs, de pouvoir travailler à distance sur des données très détaillées : qui ne respectent donc pas le secret statistique
- Avant de pouvoir travailler au CASD, les utilisateurs doivent soumettre un dossier au comité du secret statistique
    - afin de lever le secret statistique pour leur permettre d'avoir accès aux données détaillées tout en les soumettant à leur tour à l'obligation de respecter le secret statistique, comme c'est le cas pour les statisticiens du service statistique publique.
- Comment travaillent les chercheurs au sein du CASD ?


CASD
IMPORT
CASD
EXPORT
résultats

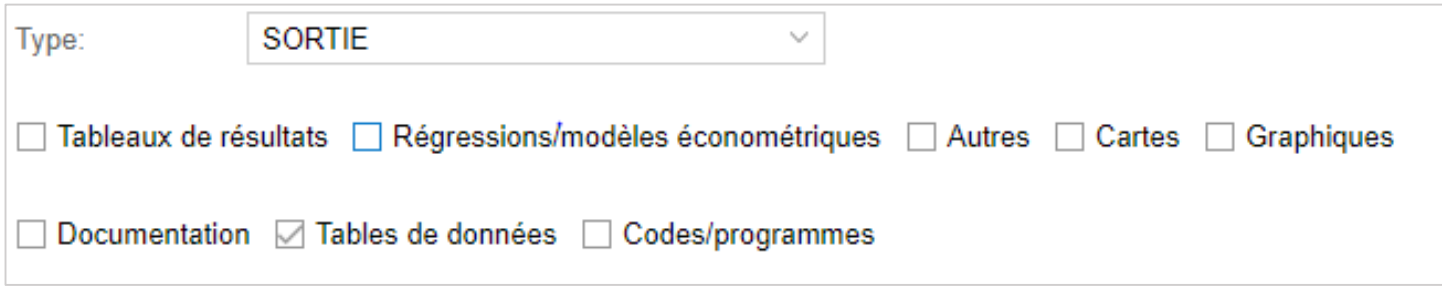
# Environnement de travail du CASD



# Les sorties de résultats

- Les chercheurs ont besoin de récupérer des fichiers de résultats mais ils doivent s'assurer qu'ils ne contiennent pas de données ré-identifiantes.
- Le personnel du service data management fait une vérification manuelle :
  - Qui demande du temps et donc un certain délai pour les chercheurs
  - De nombreux logiciels sont disponibles sur le CASD, de ce fait les sorties de résultats sont de formats et de formes très diverses : tableaux, graphiques (dans ce cas on demande les données sources), des régressions



# Les sorties de résultats

- On souhaite fournir un outil aux chercheurs pour automatiser le contrôle des sorties :
  - L'outil doit être maîtrisé par le CASD pour des raisons de sécurité et de traçabilité
  - Il doit parfaitement s'intégrer à l'architecture CASD
  - Compte tenu de la diversité des types de sorties, nous souhaitons nous concentrer sur les données individuelles
- Les logiciels existants (*mu-argus, arx par exemple*) sont complexes à intégrer dans l'architecture du CASD et de prise en main compliquée
- L'idée est de construire à partir de notre base de connaissance des sorties et en partenariat avec les chercheurs un dispositif d'anonymisation robuste pour les données individuelles (avec en entrée un format prédéfini)

# Le k-anonymat

# Estimation traditionnelle du risque de ré-identification

- Travaux de Latanya Sweeney – 1998 : $k$-anonymat
  - toute cellule (= classe d'équivalence) constituée par une combinaison déterminée de QIDs doit comporter au moins $k$ individus
  - même en connaissant les QIDs d'un individu déterminé, un attaquant ne pourra pas le distinguer des $k$-1 autres individus (au minimum) présentant les mêmes QIDs
- Seuil utilisé couramment
  - $k$-anonymat supérieur ou égal à 10

# Exemple avec trois QIDs : âge, sexe, département

| Nom | Age | Sexe | Région | Code postal | Diagnostic |
|---|---|---|---|---|---|
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Cancer |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Infection virale |
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Tuberculose |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Appendicite |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Cancer |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Tuberculose |
| * | Age ≤ 20 | Masculin | Île de France | * | Infection virale |

# Exemple avec trois QIDs : âge, sexe, département

## dans cet exemple, $k$-anonymat = 2

| Nom | Age | Sexe | Région | Code postal | Diagnostic |
|---|---|---|---|---|---|
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Cancer |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Infection virale |
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Tuberculose |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Appendicite |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Cancer |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Tuberculose |
| * | Age ≤ 20 | Masculin | Île de France | * | Infection virale |

# La *l*-diversité

# Il subsiste un risque : l'attaque par homogénéité

| Nom | Age | Sexe | Région | Code postal | Diagnostic |
|---|---|---|---|---|---|
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Cancer |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Infection virale |
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Tuberculose |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Appendicite |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Cancer |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Tuberculose |
| * | Age ≤ 20 | Masculin | Île de France | * | Infection virale |

# Il subsiste un risque : l'attaque par homogénéité
## car dans cet exemple, *l*-diversité = 1

| Nom | Age | Sexe | Région | Code postal | Diagnostic |
|---|---|---|---|---|---|
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Cancer |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Infection virale |
| * | 20 < Age ≤ 30 | Féminin | Normandie | * | Tuberculose |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | 20 < Age ≤ 30 | Féminin | Île de France | * | Appendicite |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Cancer |
| * | 20 < Age ≤ 30 | Masculin | Grand Est | * | Problème cardiaque |
| * | Age ≤ 20 | Masculin | Île de France | * | Tuberculose |
| * | Age ≤ 20 | Masculin | Île de France | * | Infection virale |

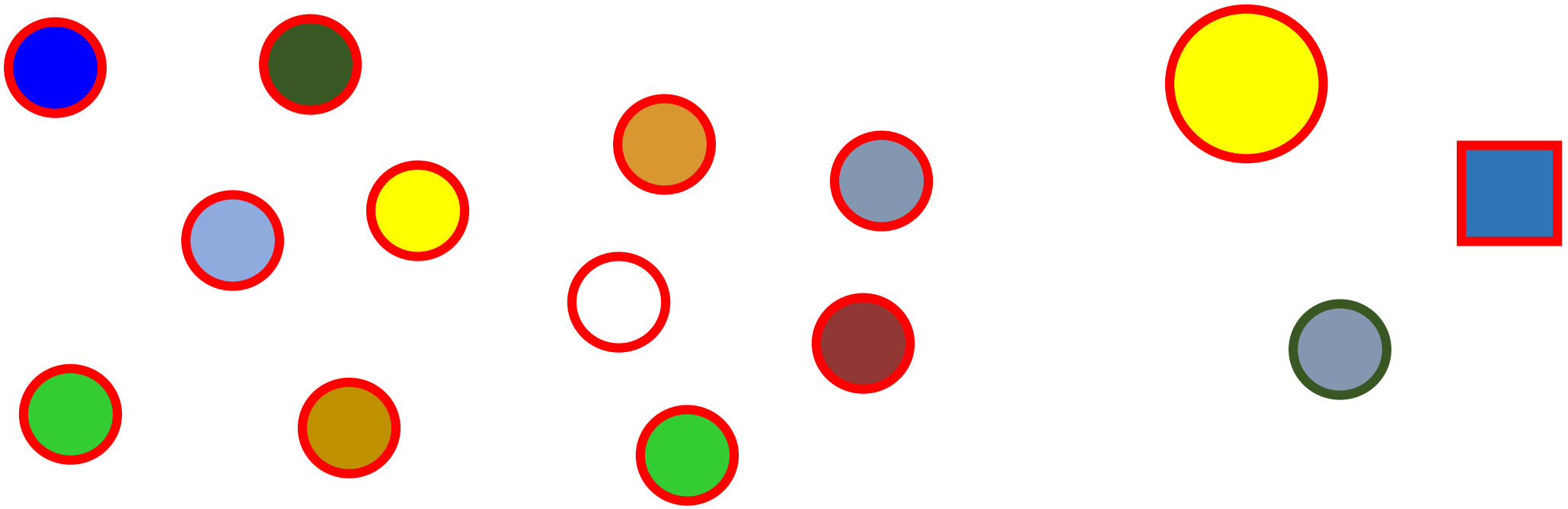
# Extension du modèle : la *l*-diversité

- au moins $l$ modalités de la « variable sensible » doivent être « suffisamment bien représentées » dans chaque cellule (= classe d'équivalence) constituée par une combinaison déterminée de QIDs
- même en connaissant les QIDs d'un individu déterminé, un attaquant ne pourra pas inférer la modalité qui lui correspond parmi les $l$ modalités (au minimum) de la variable sensible représentées par l'ensemble des individus présentant les mêmes QIDs

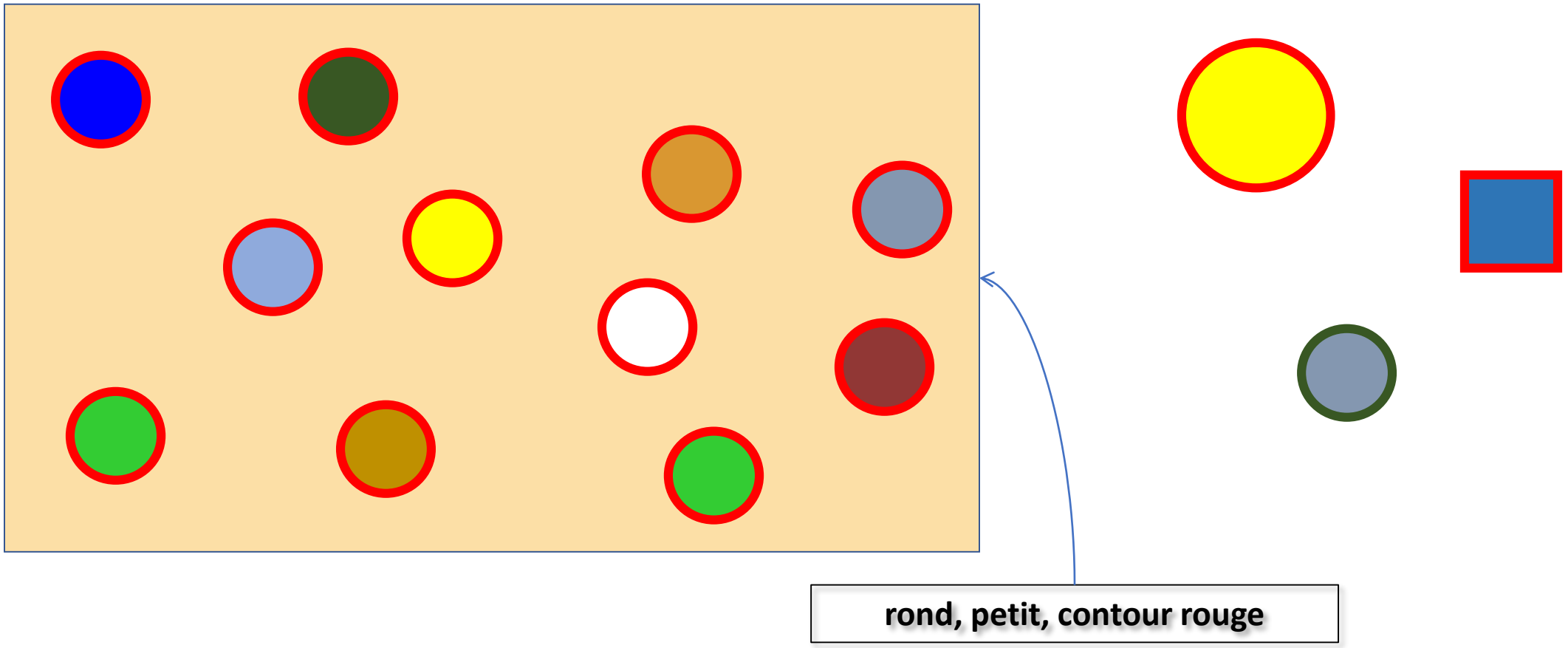
- seuil utilisé couramment
  - $l$-diversité égale ou supérieure à 3

# Impact artificiel de la granularité du critère de la $l$-diversité

# QIDs = forme, taille, contour - donnée sensible = couleur

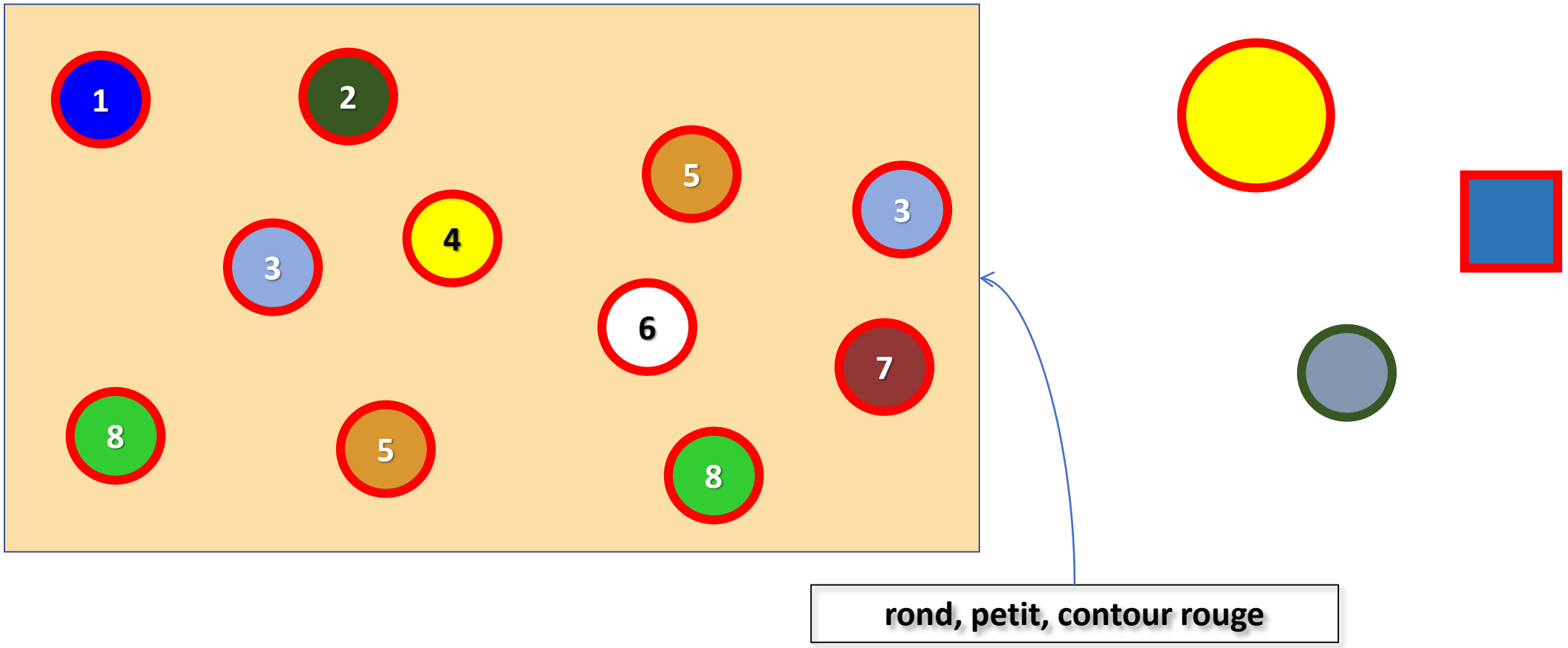


# QIDs = forme, taille, contour - donnée sensible = couleur



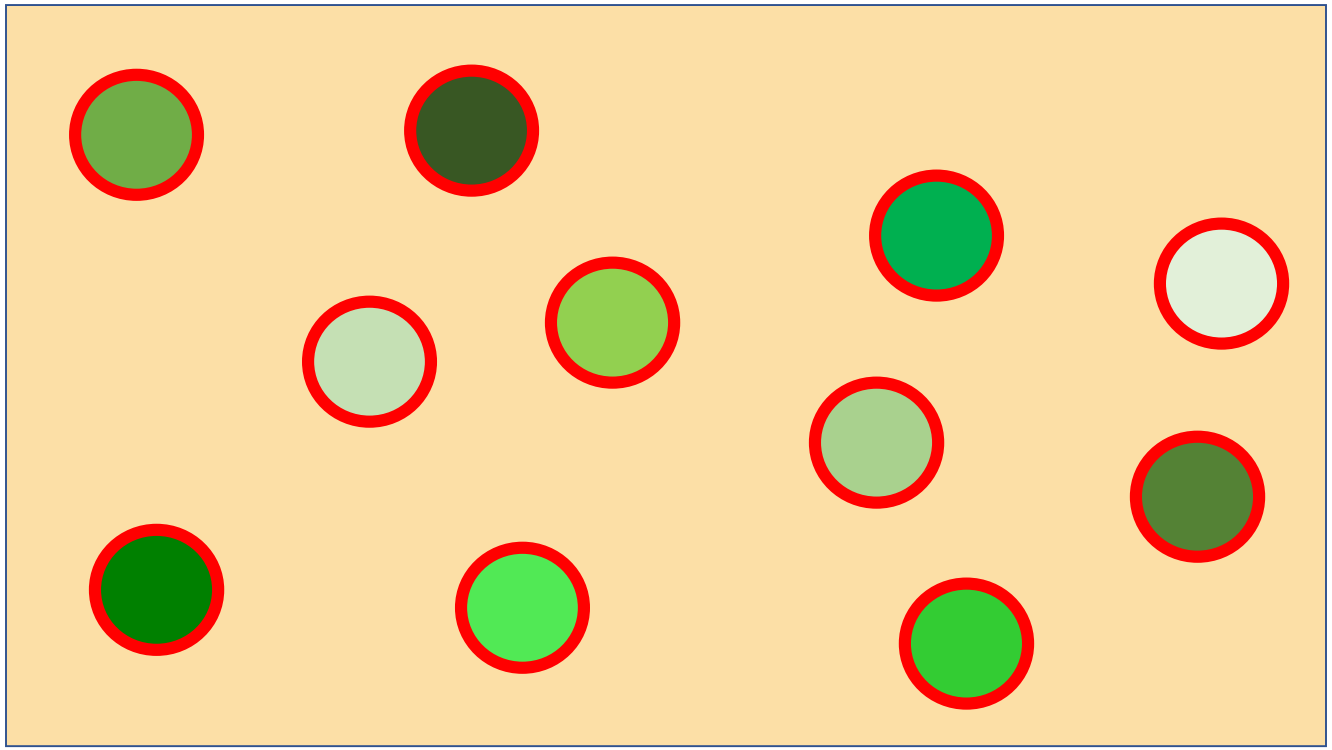
# QIDs = forme, taille, contour - donnée sensible = couleur

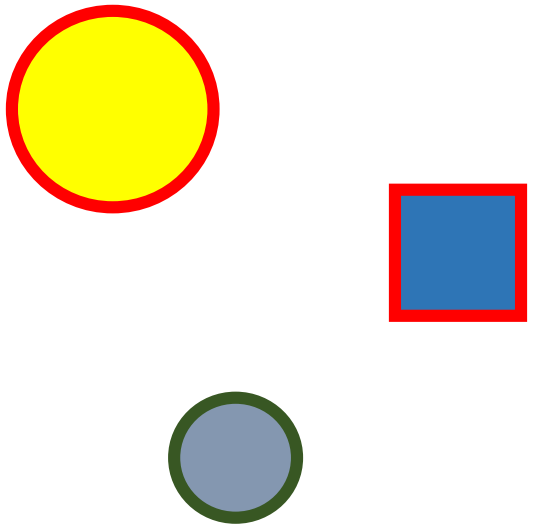
## dans cet exemple, *l*-diversité = 8



# Oui mais… la *l*-diversité vaut-elle 11 dans cet exemple ?

## sûrement pas : la diversité n'est qu'apparente, tous sont verts





**11 nuances de vert, mais vert quand même**

# … la *l*-diversité vaut-elle 11 dans cet exemple ?

## sûrement pas : la diversité n'est qu'apparente, tous sont diabétiques

| Âge | sexe | région | code | diagnostic |
|---|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | E10.7 | Diabète sucré de type 1, avec complications multiples |
| 45 à 60 ans | masculin | Hauts de France | E11.0 | Diabète sucré de type 2, avec coma |
| 45 à 60 ans | masculin | Hauts de France | E11.00 | Diabète sucré de type 2 insulinotraité, avec coma |
| 45 à 60 ans | masculin | Hauts de France | E10.3 | Diabète sucré de type 1, avec complications oculaires |
| 45 à 60 ans | masculin | Hauts de France | E11.2 | Diabète sucré de type 2, avec complications rénales |
| 45 à 60 ans | masculin | Hauts de France | E10.1 | Diabète sucré de type 1, avec acidocétose |
| 45 à 60 ans | masculin | Hauts de France | E11.1 | Diabète sucré de type 2, avec acidocétose |
| 45 à 60 ans | masculin | Hauts de France | E10.9 | Diabète sucré de type 1, sans complication |
| 45 à 60 ans | masculin | Hauts de France | E10.0 | Diabète sucré de type 1, avec coma |
| 45 à 60 ans | masculin | Hauts de France | E10.2 | Diabète sucré de type 1, avec complications rénales |
| 45 à 60 ans | masculin | Hauts de France | E11.9 | Diabète sucré de type 2, sans complication |

# Conséquences pratiques

- Une granularité inappropriée de la variable sensible sur laquelle se fonde le calcul de la *l*-diversité peut aboutir à un résultat faussement rassurant
- La granularité de ce critère doit être adaptée à la finesse descriptive que dévoilent les modalités de cette variable sensible
- En pratique on aura recours au « floutage » de la variable sensible
  - soit en remontant dans son arborescence
  - soit en établissant un mappage *ad hoc* (proxies)

# Difficultés du calcul dans le cas d’une variable sensible multivaluée

# Ce cas de figure est parfait pour la théorie…

| Age | Sexe | Région | Diagnostic |
|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | Cancer |
| 45 à 60 ans | masculin | Hauts de France | Infection virale |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque |
| 45 à 60 ans | masculin | Hauts de France | Appendicite |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque |
| 45 à 60 ans | masculin | Hauts de France | Cancer |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose |
| 45 à 60 ans | masculin | Hauts de France | Infection virale |

# … mais dans la vraie vie, voici ce qu'on observe

| Age | Sexe | Région | Diagnostics |
|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Diabète ; Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Pneumopathie ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Diabète ; Infection virale |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Appendicite ; Hypertension ; Parkinson |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Diabète ; Alzheimer |
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Hypertension ; Infection bactérienne |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Alzheimer ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Ostéoporose ; Diabète ; Hypertension ; Potomanie |

# … mais dans la vraie vie, voici ce qu'on observe

| Age | Sexe | Région | Diagnostics |
|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Diabète ; Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Pneumopathie ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Diabète ; Infection virale |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Appendicite ; Hypertension ; Parkinson |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Diabète ; Alzheimer |
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Hypertension ; Infection bactérienne |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Alzheimer ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Ostéoporose ; Diabète ; Hypertension ; Potomanie |

# … mais dans la vraie vie, voici ce qu'on observe

| Age | Sexe | Région | Diagnostics |
|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Diabète ; Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Pneumopathie ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Diabète ; Infection virale |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Appendicite ; Hypertension ; Parkinson |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Diabète ; Alzheimer |
| 45 à 60 ans | masculin | Hauts de France | Cancer ; Hypertension ; Infection bactérienne |
| 45 à 60 ans | masculin | Hauts de France | Problème cardiaque ; Alzheimer ; Diabète |
| 45 à 60 ans | masculin | Hauts de France | Tuberculose ; Hypertension |
| 45 à 60 ans | masculin | Hauts de France | Infection virale ; Ostéoporose ; Diabète ; Hypertension ; Potomanie |

# Tous ont (au moins) du diabète ou de l'hypertension

## dans cet exemple, *l*-diversité = 2

| Age | Sexe | Région | Diagnostics |
|---|---|---|---|
| 45 à 60 ans | masculin | Hauts de France | **Cancer** ; **Diabète** ; **Tuberculose** ; **Hypertension** |
| 45 à 60 ans | masculin | Hauts de France | **Infection virale** ; Pneumopathie ; **Diabète** |
| 45 à 60 ans | masculin | Hauts de France | **Tuberculose** ; **Diabète** ; **Infection virale** |
| 45 à 60 ans | masculin | Hauts de France | **Problème cardiaque** ; **Hypertension** |
| 45 à 60 ans | masculin | Hauts de France | Appendicite ; **Hypertension** ; Parkinson |
| 45 à 60 ans | masculin | Hauts de France | **Problème cardiaque** ; **Diabète** ; Alzheimer |
| 45 à 60 ans | masculin | Hauts de France | **Cancer** ; **Hypertension** ; Infection bactérienne |
| 45 à 60 ans | masculin | Hauts de France | **Problème cardiaque** ; Alzheimer ; **Diabète** |
| 45 à 60 ans | masculin | Hauts de France | **Tuberculose** ; **Hypertension** |
| 45 à 60 ans | masculin | Hauts de France | **Infection virale** ; Ostéoporose ; **Diabète** ; **Hypertension** ; Potomanie |

# En résumé, dans le cas d'une variable multivaluée

- Le calcul de la *l*-diversité s'apparente au problème connu en algorithmique sous le nom de « problème de couverture maximale » :
  - dans notre exemple : combien de diagnostics sont nécessaires pour couvrir la totalité des individus de la classe d'équivalence ?
- Ce problème est NP-difficile (théorie de la complexité)
- En raison de la durée des calculs pour le résoudre de manière exacte, on a recours habituellement à un algorithme d'approximation (dit « algorithme glouton ») qui fournit donc des valeurs approchées

# En pratique, avec l'exemple des diagnostics…

- Étant donné que le seuil utilisé couramment est
  - ➢ *l*-diversité égale ou supérieure à 3
- nous préconisons une solution alternative pragmatique, en deux étapes :
  1. **un seul diagnostic suffit-il à couvrir tous les individus de la classe ?**
     - l'algorithme est très simple et quasi-immédiat
     - si oui, la *l*-diversité de cette classe est égale à 1
     - si non, on passe à l'étape suivante
  2. **deux diagnostics suffisent-ils à couvrir tous les individus de la classe ?**
     - l'algorithme est relativement simple et très rapide
     - si oui, la *l*-diversité de cette classe est égale à 2
     - si non, la *l*-diversité de cette classe est égale ou supérieure à 3

# Travaux à venir

# Travaux à venir

- Finaliser l'implémentation expérimentale de la méthode
- Faire évaluer la pertinence des données résultantes anonymisées par des chercheurs
- Réaliser le développement de la méthode :
  - en la rendant générique (sans a priori sur les données),
  - en prenant en compte les contraintes de performance,
  - et en prenant en compte son intégration dans l'architecture sécurisée du CASD, notamment pour les questions de traçabilité.

# Merci pour votre attention

# Comme les chimistes le font avec des produits sensibles


CASD