Enseignement de la sécurité des données personnelles en STID

Paul-Marie Grollemund, Clément Jacq, **Kevin Thiry-Atighehchi**

RESSI 2022, Chambon-sur-Lac
3 juin 2022

Quelques infos

1er cycle : bac+2 (DUT), bac+3 (BUT)
Type de bac :

- 30% technos (STI2D) en 1A, 50% bac généraux
- env. 20% de technos en 2A

Programme en sécu des bases de données :

- Contrôle d'inférence
- Contrôle d'accès
- Pseudonymisation et anonymisation

Ressources :

- William Stallings (Computer Security : Principles and Practice)
- Ramez Elmasri et Shamkant B. Navathe (Fundamentals of Database Systems)
- Database System Concepts (Abraham Silberschatz, Henry Korth et S. Sudarshan)
- Ressources de collègues E-C
- 1 publication du G29
- Ouvrages de méthodologie d'enquête stat

Données personnelles et données anonymes

Deux types de donnée :

- Donnée à caractère personnel : "Dès lors qu'elles concernent des personnes physiques identifiées directement ou indirectement." (CNIL)
- Donnée anonyme : "Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne." (Article 2 de la loi Informatique et Libertés)

Qualité d'une anonymisation

Selon le G29, trois critères pour évaluer la qualité d'une anonymisation :

- **L'individualisation** : est-il possible d'isoler une partie ou la totalité des enregistrements identifiant un individu dans l'ensemble des données ?
- **La corrélation** : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ?
- **L'inférence** : est-il possible de déduire de l'information sur un individu à partir des données disponibles ?

Bris de vie privée avec des données pseudonymisées

Netflix, Latanya Sweeney et le GIC

2010 : Netflix prize, 1M$

Désanonymisation de la base grâce à des goûts cinématographiques

2 notes IMDb $\rightarrow$ identification de 68 % des utilisateurs

Milieu des années 90 : le Group Insurance Commission du Massachusetts décide de rendre publiques des données "anonymisées" concernant les hospitalisations des employés de l'état.

Désanonymisation de la base grâce à des registres publics

L. Sweeney, étudiante à l'Université Carnegie Mellon, recoupe ces données avec les listes électorales :
genre + ddn + CP $\rightarrow$ identification de 87 % des citoyens

Bases de données statistiques

- BD statistiques : fournit des données de nature statistique
- Deux types :
  - **BD purement statistiques** : données ne accessibles que par des fonctions d'aggrégation. On parle de **requêtes statistiques**.
    Exemple : BD de recensement de la population.
  - **BD ordinaire** avec accès statistique : certains utilisateurs ont un accès normal, d'autres ont accès qu'aux fonctions d'aggrégat

- Risque de sécurité : Fuite d'une information spécifique au sujet d'un individu particulier
  Exemple : La moyenne de Pierre Dupond à l'UE X

- Meta-données : Connaissances supplémentaires (non stockées dans la BD)

Terminologie en BD statistiques

- **Population** : ensemble d'enregistrements d'une table (relation) qui satisfait une condition de restriction
- Soit l'exemple de relation

  $personne(nom, \underline{num\_secu}, revenu, adresse, ville, pays, genre, dernier\_diplome)$

  Chaque condition de sélection sur la relation *personne* spécifie une population particulière des enregistrements qu'elle contient.
  Exemples :
  - La condition `genre = 'M'` spécifie la population masculine.
  - La conditon `((genre = 'F') AND (dernier_diplome = 'MSc' OR dernier_diplome = 'Ph.D'))` spécifie la population féminine qui ont un master ou un doctorat.

Bris de vie privée par inférence

- Protection de la privacy **vs** utilité statistique
- Une suite de requêtes statistiques peut mener à une atteinte de vie privée :

  Q1 : `SELECT COUNT(*) FROM personne WHERE <condition>;`

  Q2 : `SELECT AVG(revenu) FROM personne WHERE <condition>;`

  Cas intéressants : la première requête retourne une petite valeur.

  Si Q1 retourne une valeur > 1, besoin de métadonnées ou d'une requête statistique supplémentaire (p. ex. MAX) pour fournir des bornes sur les salaires.

Autre exemple : structuration d'une BDD

La BDD contient des infos personnelles : noms, adresses et salaires d'employés.

2 rôles (subalterne, admin) menant à $\neq$ droits d'accès :

- Individuellement, le nom, l'adresse et le salaire sont accessibles par un subalterne.
- L'association des noms et salaires n'est permise qu'à l'administrateur.

Solution : 3 tables

`employe(emp#,nom,adresse) - - accessible au rôle subalterne`
`salaire(s#,salaire) - - accessible au subalterne`
`emp_salaire(emp#,s#) - - accessible à l'administrateur`

Quid de l'ajout d'un attribut `date_debut` dans la table `salaire` ?

Gestion du risque de divulgation par inférence

- **Détection d'inférence durant la conception de la BD.** La détection peut être faite en analysant la BD et les contraintes de sécurité. Des solutions incluent :
  - Enlever les dépendances de données en modifiant la structure de la BD, p. ex. fractionner une table en plusieurs tables.
  - Changer le régime de contrôle d'accès pour empêcher l'inférence, p.ex. utiliser des rôles de contrôle d'accès très précis avec un système RBAC.

  Les problèmes d'inférence ne sont pas tous indiqués par la BD !

- **Détection d'inférence au moment de la requête.** On élimine une violation par canal d'inférence durant une requête ou une série de requêtes. Par exemple, si un canal d'inférence est détecté, la requête est refusée ou altérée.
  Des détections peuvent être manquées !

Contrôle d'accès

Dans le cadre des systèmes de gestion de base de données :

- Contrôle d'accès discrétionnaire
- Contrôle d'accès basé sur les rôles

Objectifs : Connaissances pratiques en SQL (Structured Query Language) :

- Octroyer des privilèges ou des rôles
- Révoquer des privilèges ou des rôles

Exemple de BD

employe

prenom	nom	code_insee	date_naissance	adresse	sexe	salaire	tel	code_insee_resp	no_dpt

departement

nom_dpt	numero_dpt	code_insee_directeur	date_prise_fct

dpt_emplacement

numero_dpt	lieu

projet

nom_projet	numero_projet	lieu_projet	num_dpt

travaille_sur

code_insee_emp	no_projet	heures

personne_a_charge

code_insee_emp	nom_proche	sexe	date_naissance	lien_parente

RBAC

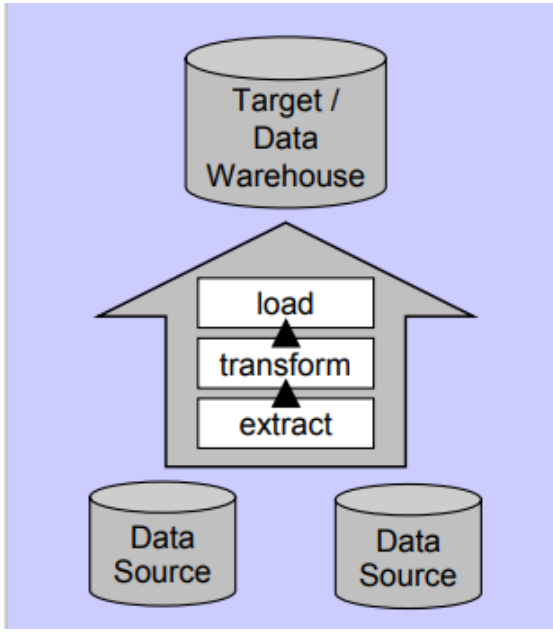
Énoncé : L'administrateur X de la base de données décide de mettre en place un contrôle d'accès basé sur des rôles. Écrivez les expressions en SQL que taperait X pour :

1. Créer 4 rôles : personnel, personnel_DRH, personnel_DAF, et personnel_technique.
2. Hiérarchiser ces rôles par généralisation/spécialisation.
3. Assigner ces rôles à des utilisateurs.
4. Octroyer des privilèges aux rôles, en utilisant lorsque c'est approprié des vues :
   - Le rôle personnel permet de récupérer les noms et prénoms de la table *employe*.
   - Le rôle personnel_DRH permet d'interroger ou de modifier les tables *employe* (à l'exception de l'attribut salaire), *travaille_sur*, *dpt_emplacement* et *personne_a_charge*.
   - Le rôle personnel_DAF permet de récupérer les attributs *code_insee*, *prenom*, *nom* et *salaire* de la table *employe*. Etre personnel de la DAF permet également de modifier le *salaire*.
   - Le rôle personnel_technique permet d'interroger les tables *departement* et *dpt_emplacement*.

Anonymiser des données avec un ETL

Un ETL (Extract - Load - Transform) permet l'intégration et la consolidation des données à l'aide des trois opérations suivantes :

- Extraction : Identifier et extraire les données de sources ayant subi une modification depuis la dernière exécution
- Transformation : Appliquer diverses transformations aux données pour les nettoyer, les intégrer et les agréger
- Chargement : Insérer les donnés transformées dans l'entrepôt et gérer les changements aux données existantes.



Types de transformation

- Révision de format (changer le type ou la longueur de certains items)
- Décodage de champs ('homme' vs 'M')
- Pré-calcul de valeurs dérivées (profit = ventes-coûts)
- Découpage de champs complexes (extraire prénom et nom à partir d'une chaine nomComplet)
- Fusion de plusieurs champs
- Conversion des format de caractères, des unités de mesure, des dates
- Pré-calcul des agrégations (ventes par produit par semaine par region)
- Déduplication (plusieurs enregistrements pour un même client)

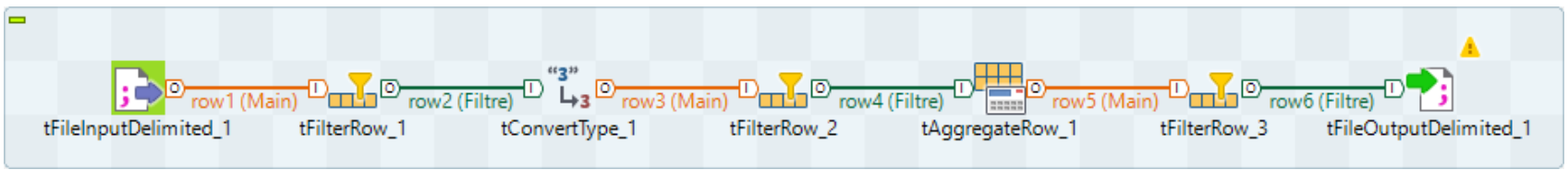
Un job Talend

Énoncé : donner la liste des prénoms populaires choisis plus de 500 fois ces 50 dernières années.

sexe	prenom	annee	nombre
1	FLORENCIO	2008	303
1	FLORENCIO	2014	320
1	FLORENCIO	XXXX	75
1	FLORENT	1972	502
1	FLORENT	1901	320

$\Rightarrow$

sexe	prenom	nombre
1	FLORENCIO	623
1	FLORENT	502



Traitement simple d'un étudiant pour le nettoyage des données :

- Élimination des valeurs exotiques ou aberrantes
- Consolidation des doublons

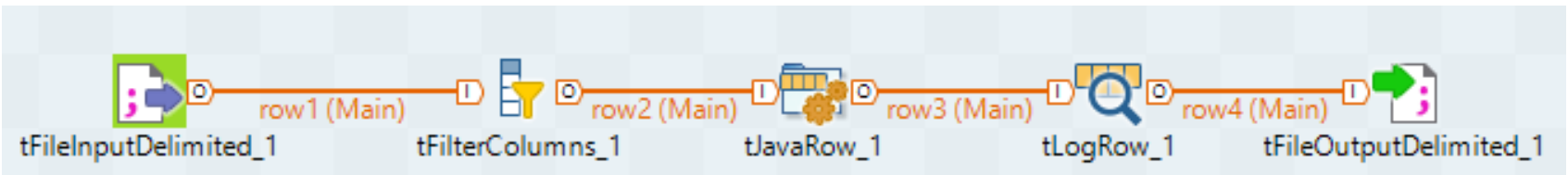
Utilisation d'un ETL pour la généralisation des données

Énoncé : Appliquer les techniques de généralisation/suppression pour obtenir un degré de k-anonymité.

Nom	Age	Genre	Ville	CP	Maladie
Geraldine Brassard	70	F	MONTLUÇON	03100	Cancer
Emilie Laroque	61	F	ESPINASSE-VOZELLE	03110	Cancer
Justin Pitt	62	M	POUZOL	63340	Cancer
Christophe Huron	65	M	AUZON	43390	Cancer
Mariane Guerrero	26	F	ANDON	06750	Rhume
Lauriane Ponce	38	F	ANTIBES	06160	Rhume

⇓

Nom	Age	Genre	Ville	Region	Maladie
	61+	F		AURA	Cancer
	61+	F		AURA	Cancer
	61+	M		AURA	Cancer
	61+	M		AURA	Cancer
	21-40	F		PACA	Rhume
	21-40	F		PACA	Rhume



Chaîne de traitements :

- FilterColumn/Projection : supprimer le nom et la ville
- JavaRow : généralisation par classe d'âge 0-20, 21-40, 41-60, 61+, ainsi que sur le CP
- Log : affiche le résultat pour vérifier que ça marche
- Output : on remet les deux colonnes vides

Étude de cas sur le secret médical dans un contexte épidémique

Rappel RGPD : Ne pas demander plus d'informations que celles nécessaires au traitement des données ou à la fourniture du service.

Gestion d'une crise sanitaire d'un point de vue décisionnel :

- Doit-on vacciner ?
- Quand ?
- À quelle intensité ?

Leviers :

- Levée du secret médical
- Utilisation d'autres leviers préservant la confidentialité des données personnelles

Étapes de travail avec les étudiants

1. Mise en place et compréhension de la confidentialité différentielle.
2. Modélisation d'une épidémie avec le modèle SIR.
3. Décision autour d'une campagne de vaccination.
4. Modélisation de l'incertitude dur à la confidentialité différentielle.

Le modèle SIR

Notations :

- S : la population des personnes saines
- I : la population des personnes infectées
- R : la population des personnes guéries.

On note N l'ensemble de la population. À chaque instant : $N = S + I + R$

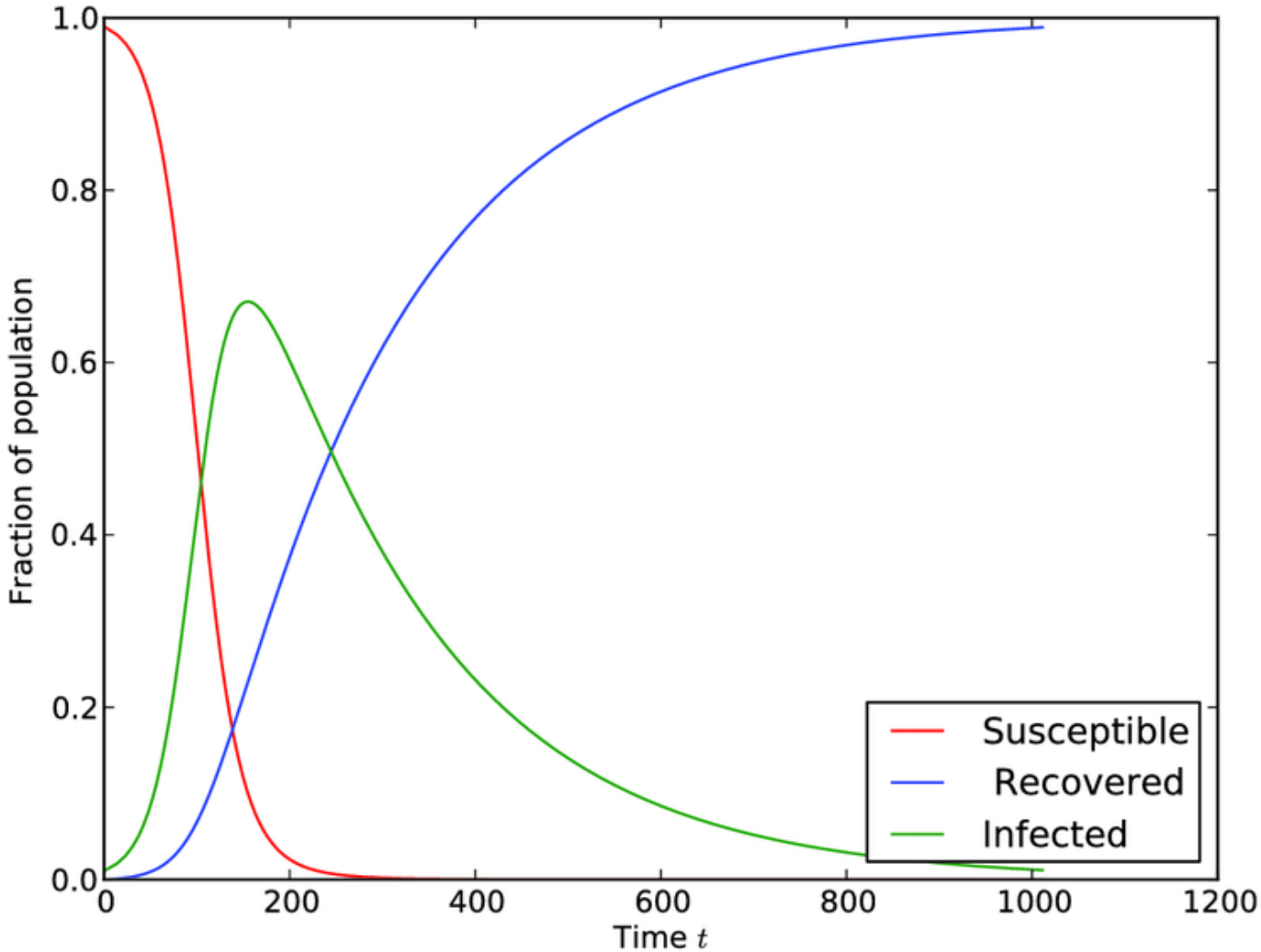
Objectif : Déterminer un système d'équations différentielles pour modéliser l'évolution de l'épidémie

Les paramètres à modéliser sont relatifs aux transitions :

- S $\rightarrow$ I (infection)
- I $\rightarrow$ R (guérison)

Remarque : S, I et R sont des fonctions de t.

Exemple d'évolution de l'épidémie



Technique de réponse aléatoire en enquête

Objectif : Réduire les biais de non-réponse et de désirabilité sociale

L'utilisation d'une technique de réponse aléatoire permettra au répondant de répondre honnêtement sans que l'interrogateur sache à quelle question il répond.

Exemple de la réponse forcée (fixe)

On veut poser la question sensible : "Avez-vous été infecté ?".

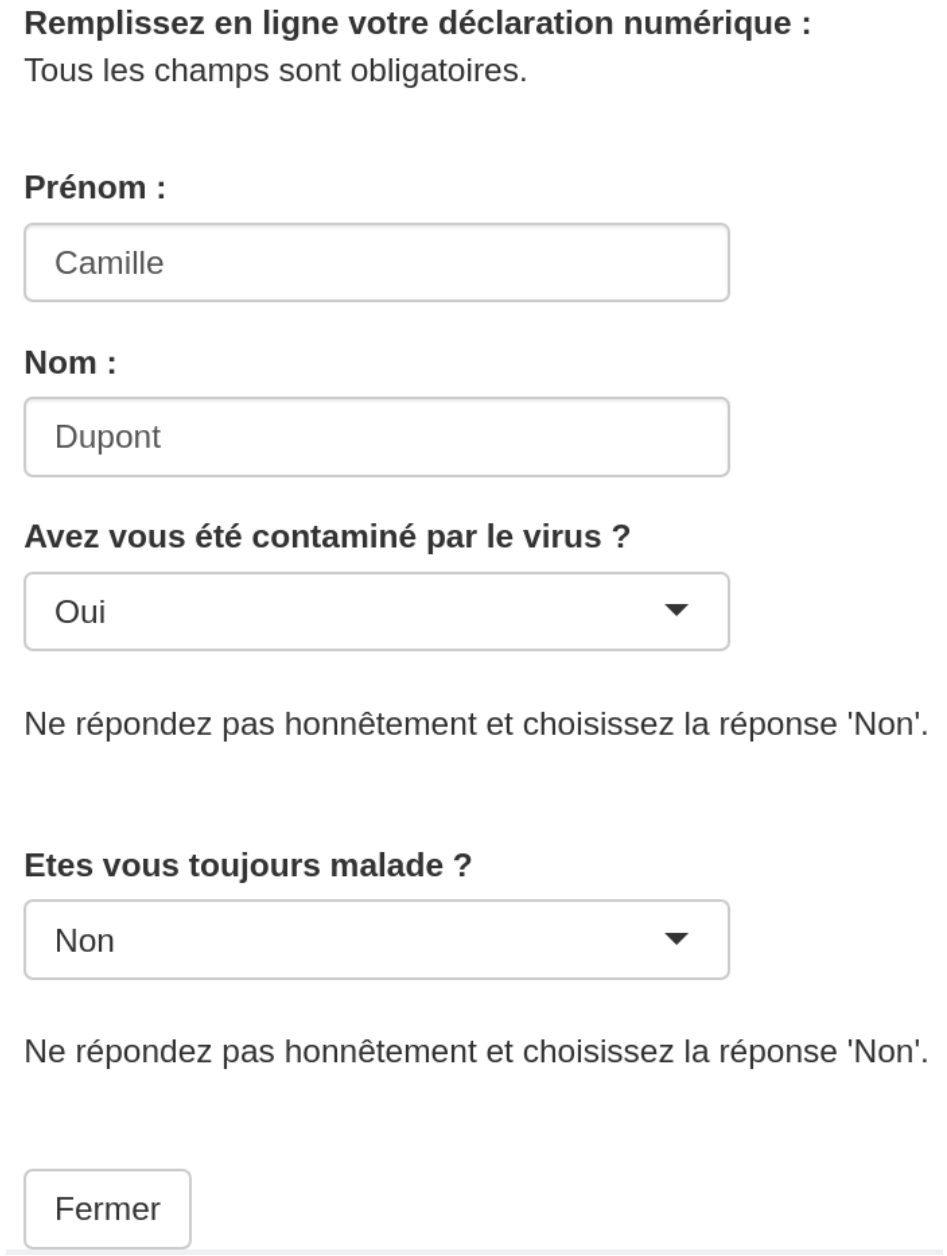
Avant que le sujet ne réponde, on lui demande de tirer à pile ou face hors du champ de vue de l'enquêteur, et de procéder comme suit :

- Si la pièce tombe sur face, répondez honnêtement à la question.
- Si la pièce tombe sur pile, répondez "oui".

Besoin de privacy pour les deux modalités (oui ou non) :

- Technique de la réponse aléatoire.

Codage du formulaire R Shiny



Exemple de la réponse aléatoire

On demande au sujet de tirer à pile ou face hors du champ de vue de l'enquêteur, et de procéder comme suit :

- Si la pièce tombe sur face, il répond honnêtement aux deux questions ("infecté" (oui puis oui), "saint" (non), ou "guéri" (oui puis non).
- Si la pièce tombe sur pile, il lance un dès à 3 faces :
  - Si c'est 1, il répond "infecté" (oui puis oui).
  - Si c'est 2, il répond "saint" (non).
  - Si c'est 3, il répond "guéri" (oui puis non).

Débiaisement de l'estimateur

Principales sources d'erreur d'une enquête

- Erreur de couverture
- Erreur d'échantillonnage
- Erreur de non-réponse
- Erreur de mesure, d'observation
- Privacy différentielle

Notations :

- $\widetilde{p}_S$ la probabilité de répondre S
- p_S la probabilité d'avoir effectivement l'état S
- p_h est la probabilité de répondre honnêtement à la question.

Estimateur biaisé : $\widetilde{p}_S = p_S p_h + \frac{(1-p_h)}{3}$.

1. On demande un estimateur de la vraie proportion, en isolant p_S.
2. La formule pour $\mathbb{V}(\hat{p}_S)$ est donnée, et on demande de calculer un intervalle de confiance.

Modèle SIR stochastique

Contexte :
On ne connaît pas avec précision S, I et R
Conséquence d'utilisation d'anonymisation (*differential privacy*)

Prise en compte dans le modèle :
Rajout d'un terme dans les condtions initiales
On connaît les approximations $\tilde{S}$, $\tilde{I}$ et $\tilde{R}$ pour l'instant $t = 0$.
On supposera que :

- $S(0) = \tilde{S} + \varepsilon_S$ où $\varepsilon_S \sim N(0, \sigma_S^2)$
- $I(0) = \tilde{I} + \varepsilon_I$ où $\varepsilon_I \sim N(0, \sigma_I^2)$
- $R(0) = \tilde{R} + \varepsilon_R$ où $\varepsilon_R \sim N(0, \sigma_R^2)$

Approche numérique :
Simuler n trajectoires :

1. Simuler les nombres $\varepsilon_.$.
2. Déterminer les conditions initiales S_0, I_0 et R_0.
3. Calculer la trajectoires avec un schéma numérique.

L'ensemble des trajectoires obtenues représentent l'évolution potentielle de l'épidémie au regard de l'incertitude sur les conditions initiales.