ASSOCIATION ARISTOTE

JOURNÉE « CONFIANCE NUMÉRIQUE »

# PRIVACY ENHANCING TECHNOLOGIES

## APPLIQUÉES AUX DONNÉES ELECTRIQUES

BENOIT.GROSSIN@EDF.FR

# NOS DONNÉES DE CONSOMMATIONS D'ÉLECTRICITÉ SONT DES DONNÉES À CARACTÈRE PERSONNEL

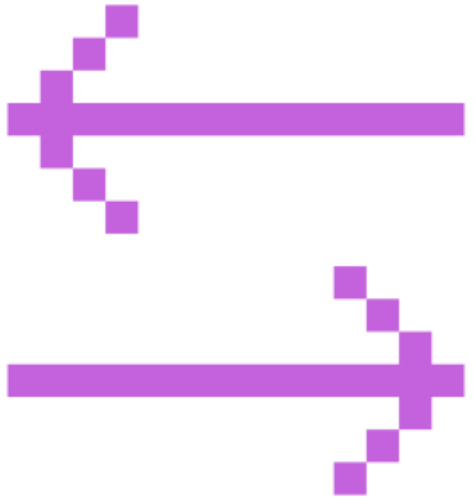
- **Pouvoir identifiant** : une série de 5 index quotidiens suffit à constituer un identifiant unique pour la quasi-totalité de la population*

- **Proxy de l'activité d'un foyer** : les courbes de charge «*peuvent révéler des informations sur la vie privée*» (CNIL) : périodes d'absence, heures de lever et de coucher, … même si dans les faits ce n'est pas si simple que ça !

* Thèse "Privacy risk analysis of large-scale temporal data : application to electricity consumption data", A. Voyez, 2023

# CONTEXTE

Les questions de Privacy des données énergétiques sont devenues incontournables. Les mesures de nos compteurs électriques sont des DCP (Données à Caractère Personnel) au même titre que nos noms et adresses. Dans le même temps, l'urgence climatique et la transition énergétique doivent mobiliser toutes les ressources disponibles, a fortiori les données énergétiques. Cela est indispensable pour concevoir de meilleurs algorithmes de pilotage offre/demande, mieux cibler les actions de rénovation ou bien encore lutter plus efficacement contre la précarité énergétique.

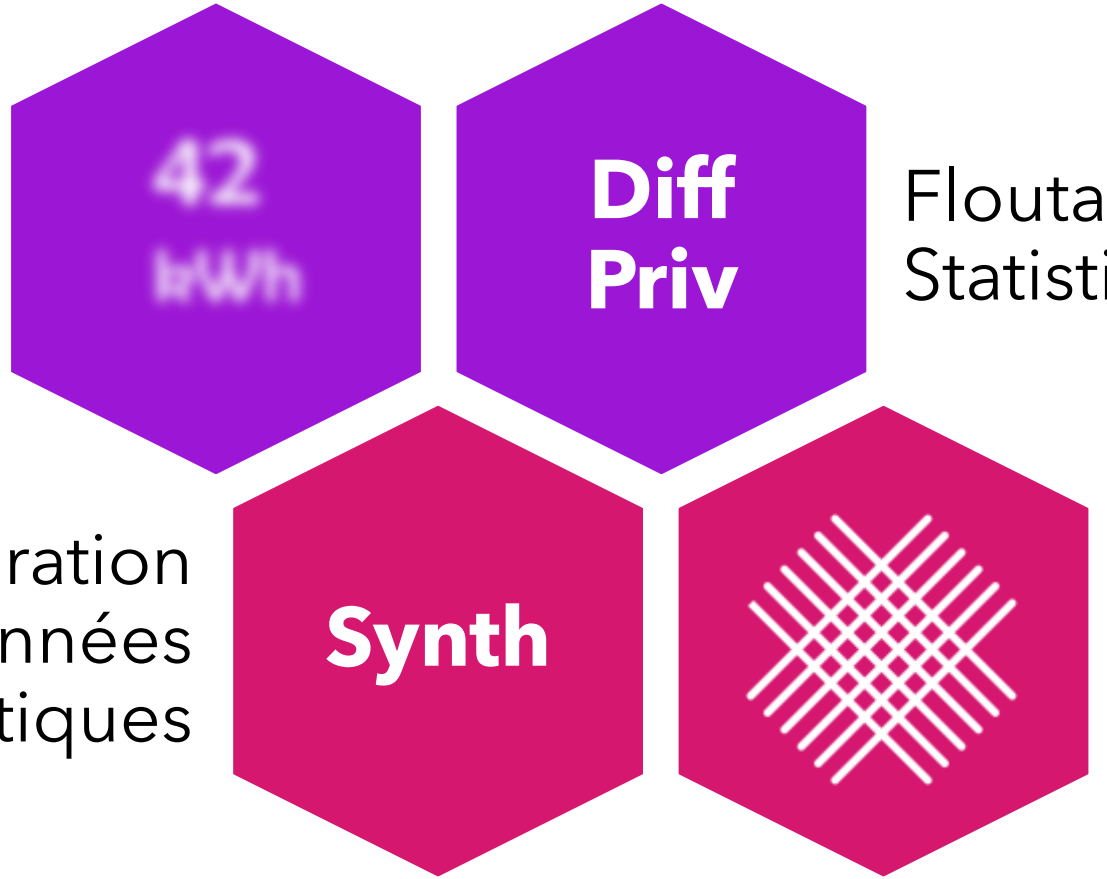


**Comment continuer à concilier besoins croissants d'utilisation des données énergétiques et protection de leur confidentialité ?**

# PRIVACY ENHANCING TECHNOLOGIES

- **PETs** = *"**digital solutions** that allow information to be collected, processed, analysed, and shared while **protecting data confidentiality and privacy"*** (OCDE, 2023)

Emerging privacy-enhancing technologies : Current regulatory and policy approaches. OECD, March 2023




Floutage
Statistique

Génération
de données
synthétiques

- Données agrégées
- Données individuelles

Diff Priv

# DIFFERENTIAL PRIVACY

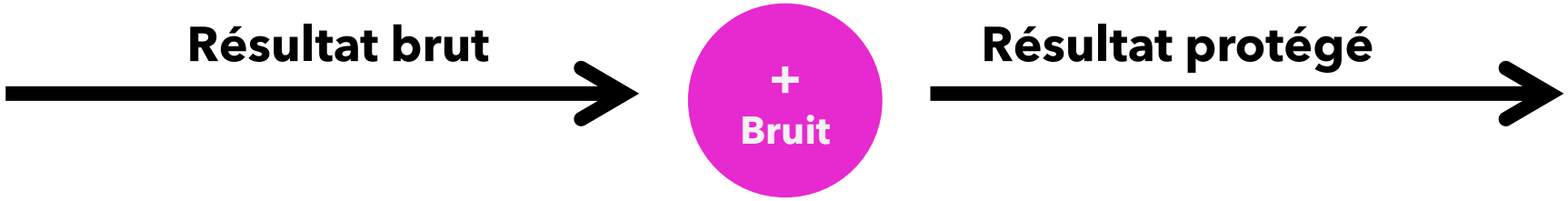
- Principe : Ajouter un « bruit statistique » proportionné dans des données agrégées pour que les informations spécifiques d'un individu soient cachées

- Solution très populaire chez les GAFAM, considérée comme la « Cutting Edge Privacy Protection »
  - Pilotée par un **le contenu des données individuelles** à agréger
  - Garantie forte de protection de la confidentialité, avec preuves mathématiques [Dwork, 2006]
  - Paramétrage non intuitif, solution non prévue pour les séries temporelles

Dwork, Cynthia. "Differential Privacy." International Colloquium on Automata, Languages and Programming (2006).

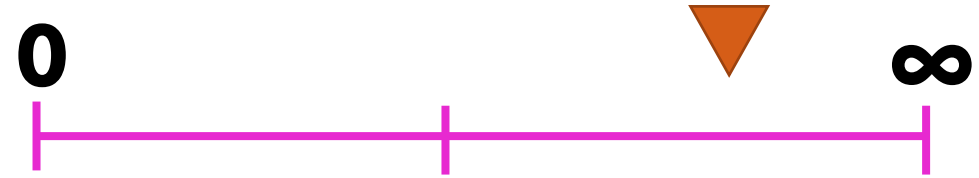
Diff Priv

# DIFFERENTIAL PRIVACY



Diff Priv

# COMPROMIS ENTRE PROTECTION ET UTILITE




Protection parfaite,
Utilité nulle
(trop de bruit)

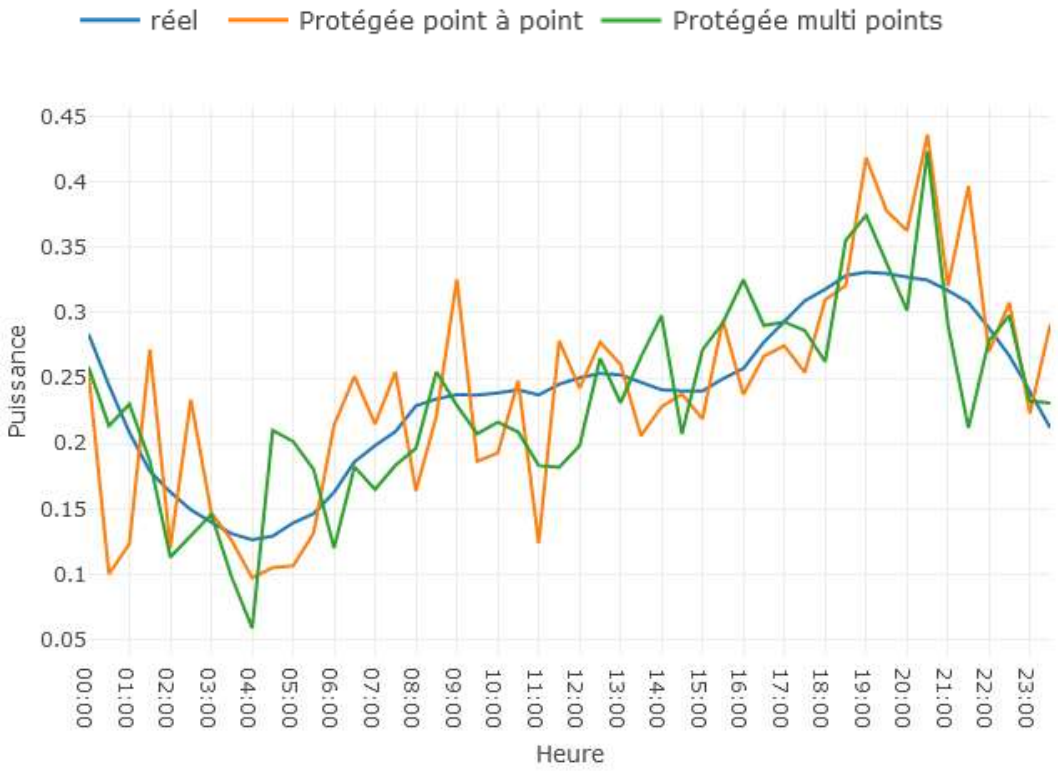
Utilité maximale
(aucun bruit),
Pas de protection

$\varepsilon$

- Epsilon, **paramètre central** de la Diff Priv
- Définit le **niveau de protection**
- Appelé aussi « **privacy-loss budget** »

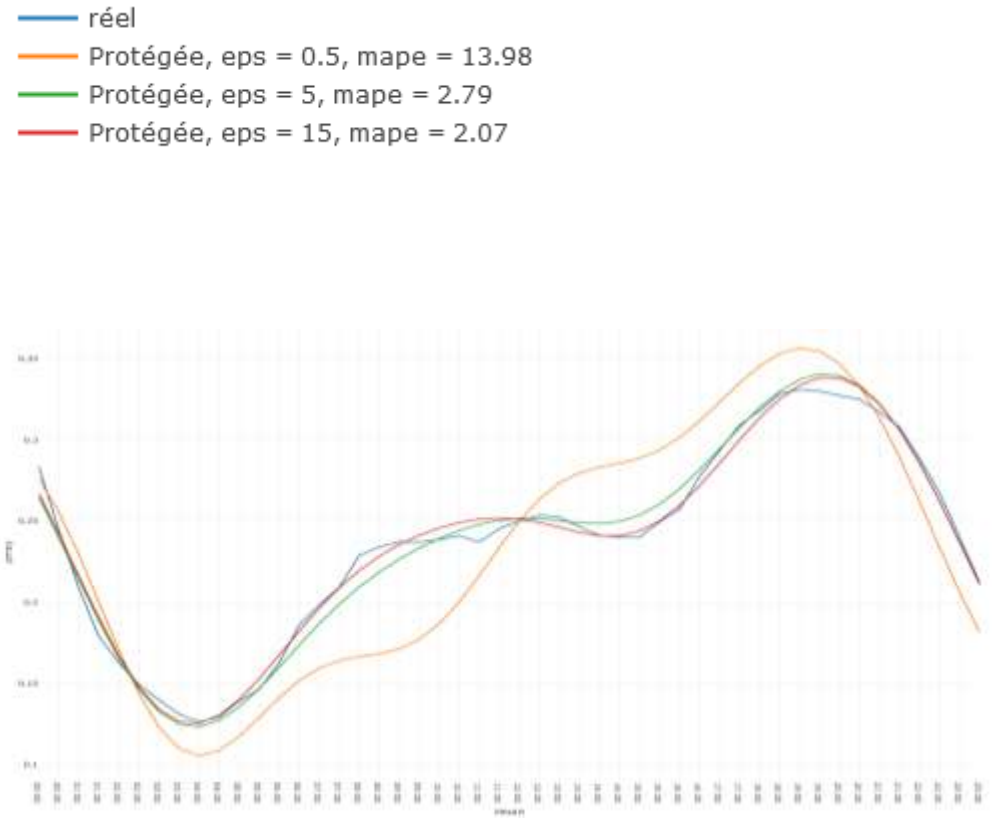
Diff Priv

# IMPLÉMENTATION OPTIMISEE POUR LES DONNEES DE CONSOMMATION D'ÉLECTRICITÉ




**Implémentations naïves de la DP**

3 ans plus tard

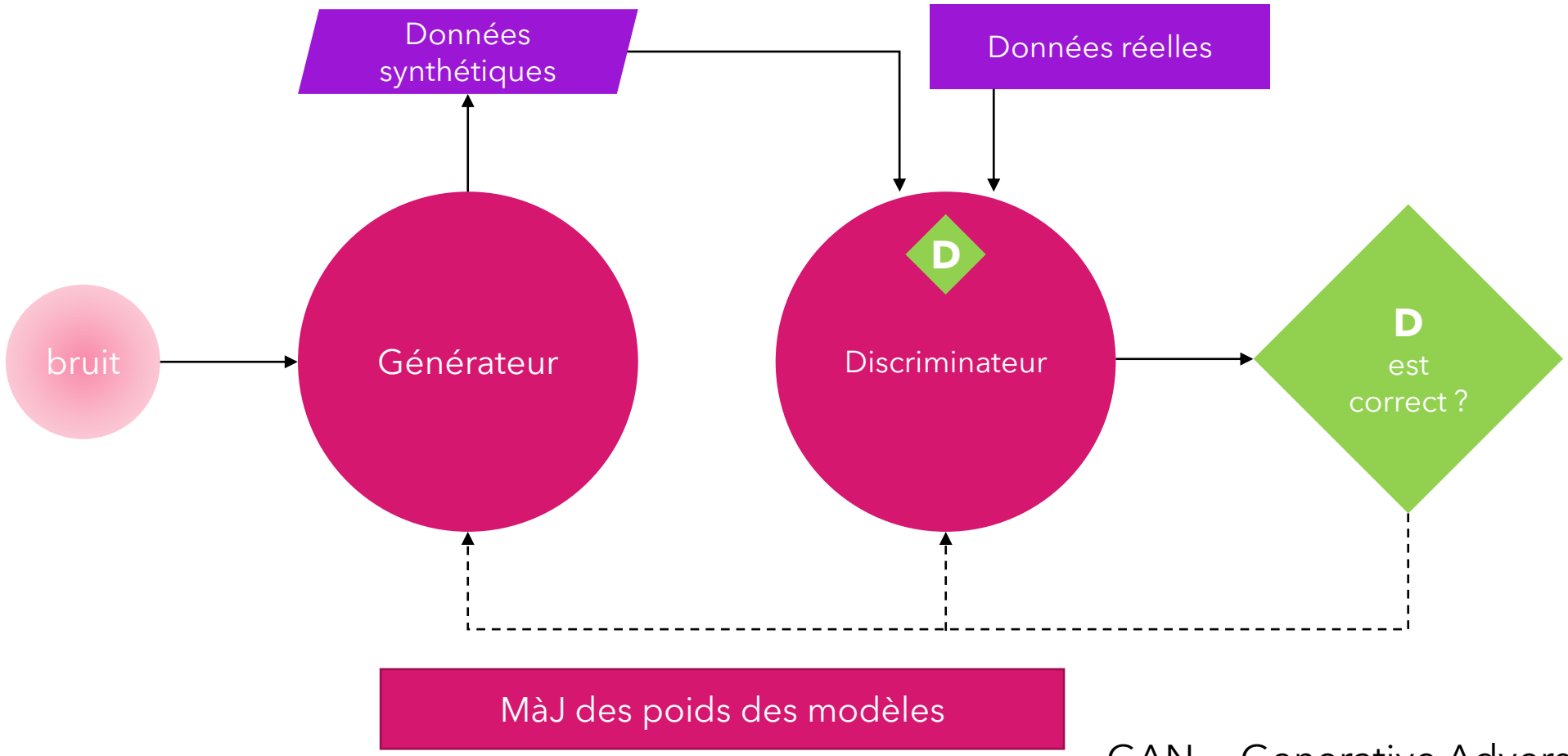



**Implémentations optimisées de la DP**

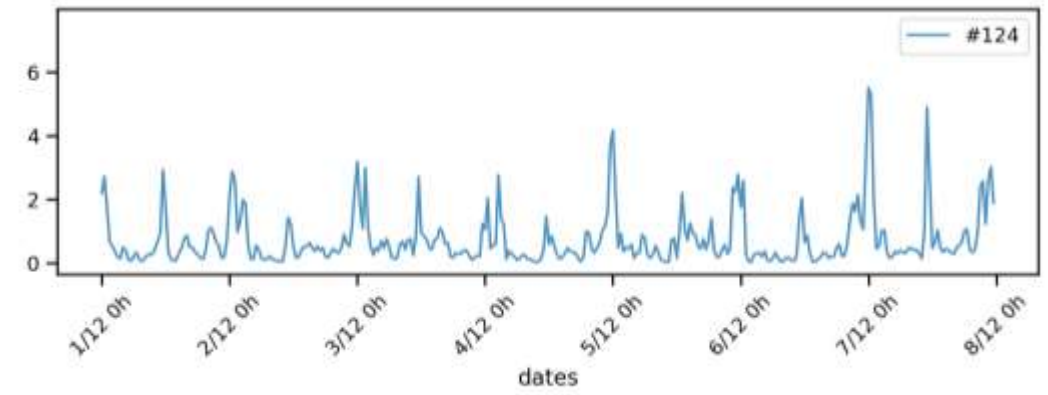
Pour en savoir plus, article CIRED « Differential Privacy For Energy Data Publication » Agoua et al., 2021

Synth

# UN EXEMPLE DE GÉNÉRATEUR DE DONNÉES SYNTHÉTIQUES




GAN = Generative Adversarial Network


*Ceci n'est pas une donnée Linky*

**« Le distributeur […]** Enedis **mène un complexe projet de machine learning. Le but est ainsi de** générer, grâce à un GAN, des courbes de consommation fictives, mais cependant réalistes**. Une solution conciliant conformité et capacité d'innovation. »**

**« […] dans le cadre d'un partenariat avec la** R&D d'EDF**. »**

https://www.alliancy.fr/enedis-objectif-industrialisation-generation-donnees-consommation

Synth

# AVANTAGES DES DONNÉES SYNTHÉTIQUES

- **Configurables** : données à la demande, avec un générateur paramétrable

- **Economiques** : les données synthétiques sont en général peu coûteuses par rapport à des données réelles

- **Disponibles** : les générateurs de données synthétiques sont capables de produire rapidement des masses importantes de données

- **Préservent la Privacy** : … allégation à vérifier !

Synth

# C'EST QUOI UNE « BONNE DONNÉE SYNTHÉTIQUE » ?

Bon chasseur et mauvais chasseur

- 3 dimensions clefs, à évaluer avec des métriques
  - **Fidélité**

    e.g. Histogram similarity score
  - **Utilité**

    e.g. Train Synthetic Test Real (TSTR) score vs. Train Real Test Real (TRTR)
  - **Privacy**

    e.g. Membership inference score


nnexes

# TRACES
# PROJET EDF R&D SUR LA VIE PRIVÉE

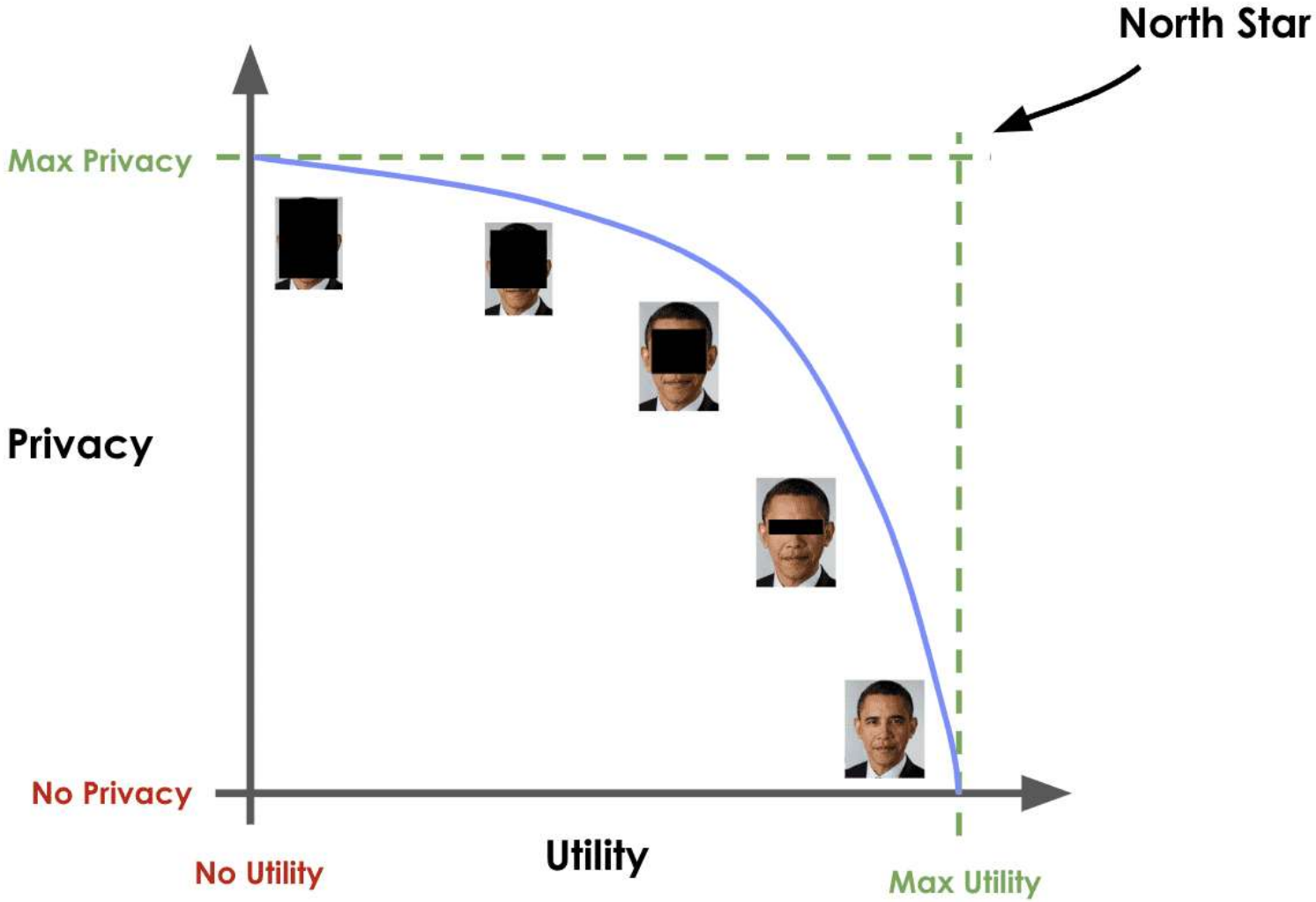
- Analyser les liens entre **société numérique et vie privée**
- Elaborer et diffuser des **méthodes de protection de la vie privée pour le secteur de l'énergie**

✓ Identifier, sélectionner et étudier des PETs depuis 2018

✓Anticiper des déclinaisons pour les métiers d'EDF

✓Partage des bonnes pratiques

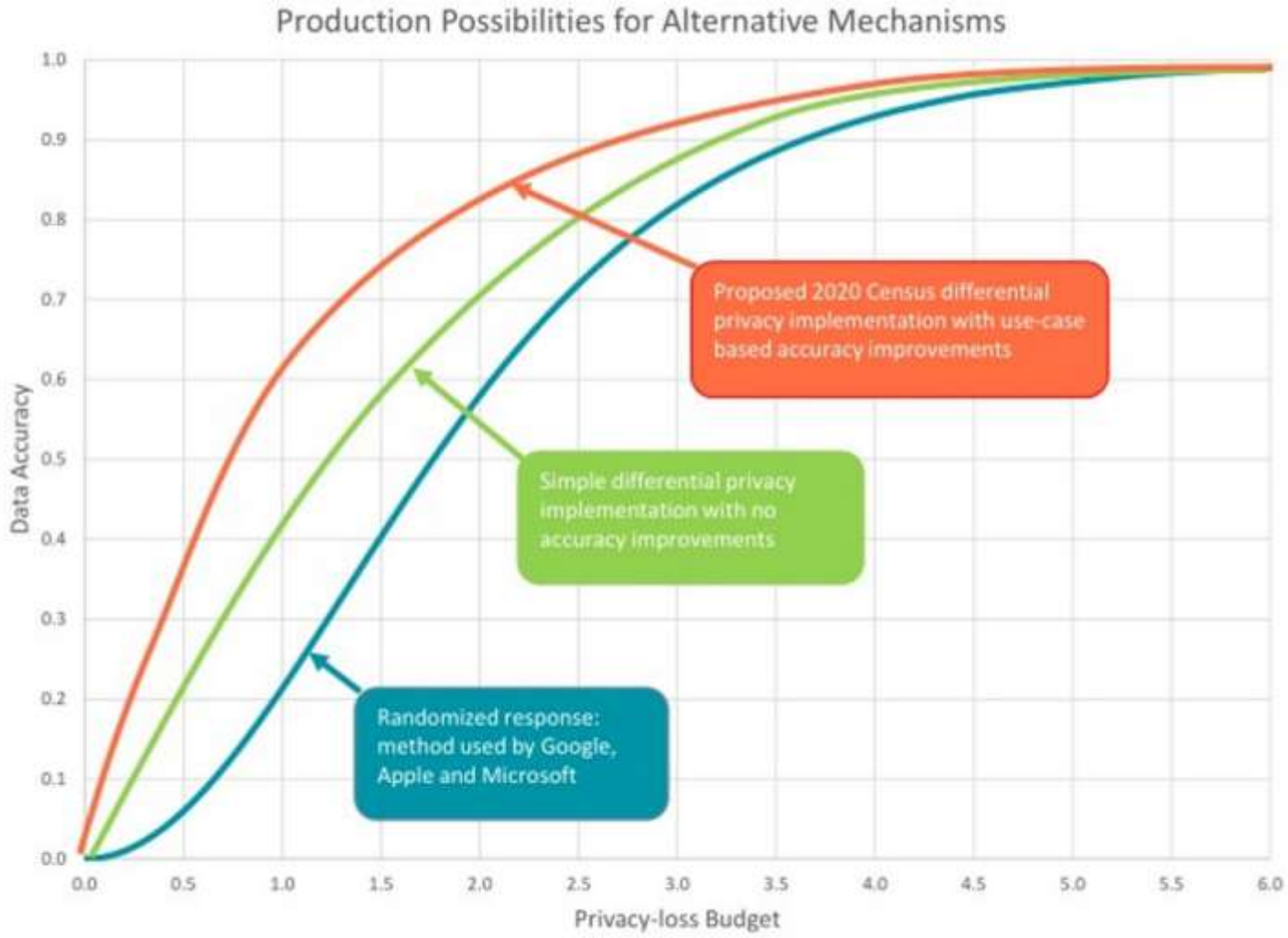
Cheffe de projet : Cécile CARON


North Star
Max Privacy
Privacy
No Privacy
No Utility
Utility
Max Utility

Diff Priv

# IMPLÉMENTATION OPTIMISEE PAR USE-CASE

## US CENSUS BUREAU




United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov