cdefi
Conférence des Directeurs
des Écoles Françaises
d'Ingénieurs

ENAC

INP TOULOUSE ENSAT

isae
Institut Supérieur de l'Aéronautique et de l'Espace
SUPAERO

06 et 07 juin 2019

Ethique et données de la recherche

Introduction et questions

Gilles ADDA, Charlotte SICRE

COMETS

Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

06 et 07 juin 2019

Données de la recherche

Selon l'OCDE, les « **données de la recherche** » sont définies comme « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme **sources principales pour la recherche scientifique** et sont généralement reconnus par la communauté scientifique comme
nécessaires pour valider des résultats de recherche.

Données de la recherche

Le **contexte** est important : quand, quelle question scientifique, pour quoi, etc… ? Exemples (Fayet, 2013) :

- les images d'une ville préhistorique deviennent des données pour un chercheur qui étudie l'histoire de cette ville.
- les « données » d'un linguiste peuvent être des écrits ou des discours, des enregistrements de locuteurs ;
- les « données » d'un médiéviste sont des sources archivistiques, archéologiques, épigraphiques, iconographiques, littéraires ;
- les « données » d'un géologue rassemblent des coupes et observations de terrain consignées sur un carnet, des résultats de carottage, des analyses d'échantillons, des données sismographiques…

Production des données

Des questions éthiques variées, déontologique

- des outils performants, mais pas autant qu'on pourrait le croire
- desidentification (anonymisation, pseudonymisation) forcément imparfaite

Une maîtrise imparfaite (impossible ?) de l'apprentissage automatique :

- biais à l'entraînement
- utilisation de données du Web, plus ou moins légalement (exemple, Twitter , Doctissimo, etc.)
- « uberisation » de la production de données, plateformes à la Amazon Mechanical Turk

Désidentification (anonymisation)

Facile :

Mme X... a eu connaissance de ce que l'arrêt de la cour d'appel de Douai avait été publié sur Internet sans être anonymisé

Moins facile :

Le maire d'Agnos, président de la Fédération des œuvres laïques (FOL) de 1999 à 2003, a été condamné par la cour d'appel de Pau à 2 ans de prison avec sursis

Utilisation de ressources extérieures ➔ réidentification

Pseudonymisation (RGPD) idem anonymisation, sans perte mais en plus problème de gestion des fichiers d'identité

Nécessité de grandes masses de données annotées pour entraîner les systèmes

- besoin d'annotateurs humains, ce qui coûte cher
- Utilisation de plateformes de crowdsourcing type Amazon Mechanical Turk
- Chercheurs employeurs

Amazon Mechanical Turk

amazon mechanical turk

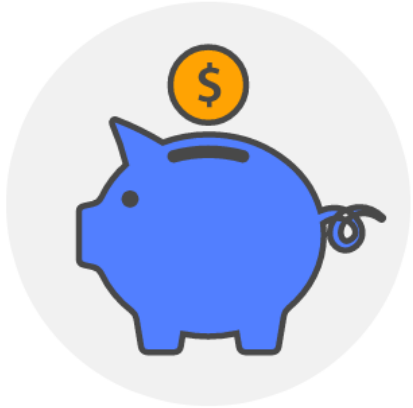
Get Started with Amazon Mechanical Turk

Create Tasks

Human intelligence through an API. Access a global, on-demand, 24/7 workforce.

Create a Requester account

or



Make Money

Make money in your spare time. Get paid for completing simple tasks.

Create a Worker account

Amazon Mechanical Turk

AMT est une plate-forme de crowdsourcing (myriadisation), et de microworking (travail parcellisé) : les tâches sont découpées en pièces (HITs) et leur exécution est rémunérée par les *Requesters*

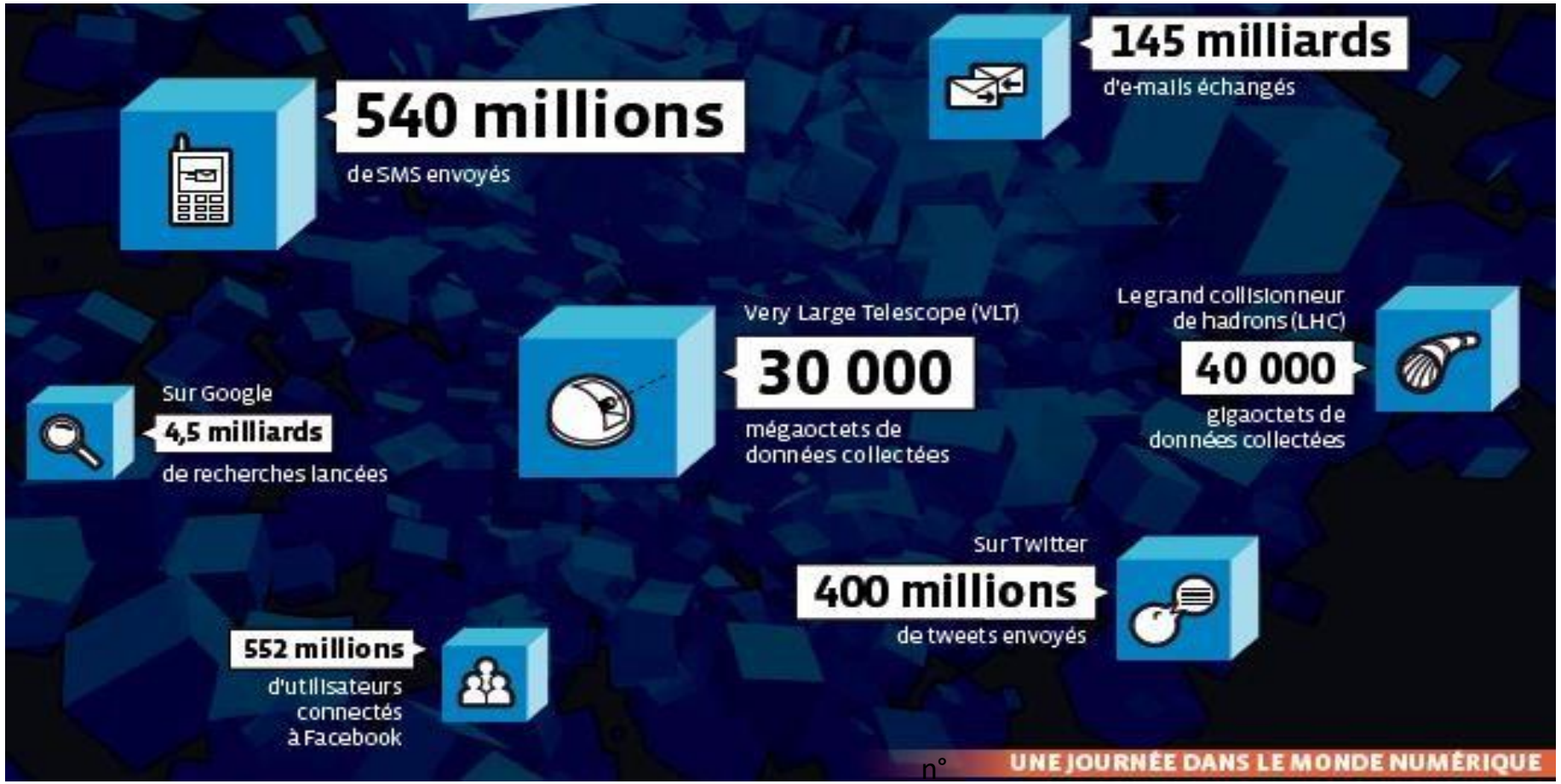
- Aucune identification : aucune relation entre les *Requesters* et les *Turkers* et entre *Turkers*
- Aucune possibilité de créer un syndicat, pour protester ou ester en justice.
- Pas de salaire minimum (< 2 $/h en moyenne)
- Possibilité de refuser de payer les *Turkers*

Conséquence d'annotation AMT

- Qualité des données vs Quantité
- Protection des données sensibles
- Traçabilité des données si problème
- Problème éthique d'utilisation de travail dissimulé

Utilisation des données

La déferlante des données



Données Massives : définition

Que sont les données massives ?

- Quelle nature spécifique aux données massives ?
- Volume, Variété, Vélocité, Véracité, Visibilité. . .
- Données comme image du monde

Effet Lampadaire

Données Massives : définition

Quel traitement spécifique sur les données massives ?

- Approche inductive
- Machine learning comme instruments d'exploration

Données massives, révolution scientifique ?

La fin de la théorie ?

- Jim Gray (Microsoft, 2007). 1. Empirisme 2. Théorie 3. Informatique 4e paradigme : données massives ➔ eScience

  Publication du livre "The Fourth paradigm" (2009)

- "Petabytes allow us to say : "Correlation is enough." We can stop looking for models." (Chris Anderson,"The End of Theory : The Data Deluge Makes the Scientific Method Obsolete“

  Erronée car données incomplètes.

  Tout ensemble de données est fondé sur des hypothèses.

Exemple : déplacements de la population française avec données téléphone mobile

Graphique 17 – Taux d'équipement en téléphone mobile selon l'âge, le diplôme, la taille du foyer et le niveau de revenu

- Champ : ensemble de la population de 12 ans et plus, en % -

Source : CREDOC, Enquêtes sur les « Conditions de vie et les Aspirations ».

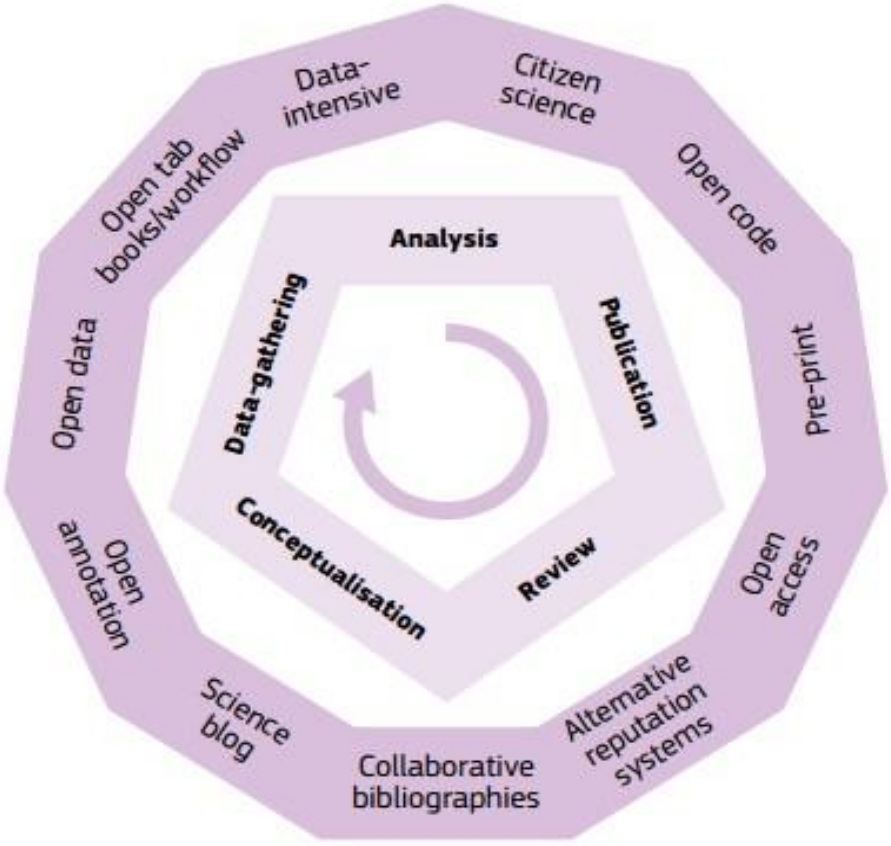
Science ouverte

L'Open Science

Eléments de contexte :

- Big Data,
- Technologies numériques,
- Mondialisation des communautés de recherche ,
- Exigence publique croissante face aux défis sociétaux (nutrition, environnement, liberté…)

L'Open Science a un impact sur l'ensemble du processus de recherche



Plan national pour la science ouverte (Juillet 2018)

- Rendre obligatoire la **diffusion ouverte des données** de recherche issues de programmes financés par appels à projets sur fonds publics.
- Créer la fonction d'**administrateur des données** et le réseau associé au sein des établissements.
- Contribuer activement à la structuration européenne au sein du *European Open Science Cloud* et par la participation à **GO FAIR.**
- Généraliser la mise en place de **plans de gestion des données** dans les appels à projets de recherche

Plan S

DEUXIÈME AXE : STRUCTURER ET OUVRIR LES DONNÉES DE LA RECHERCHE

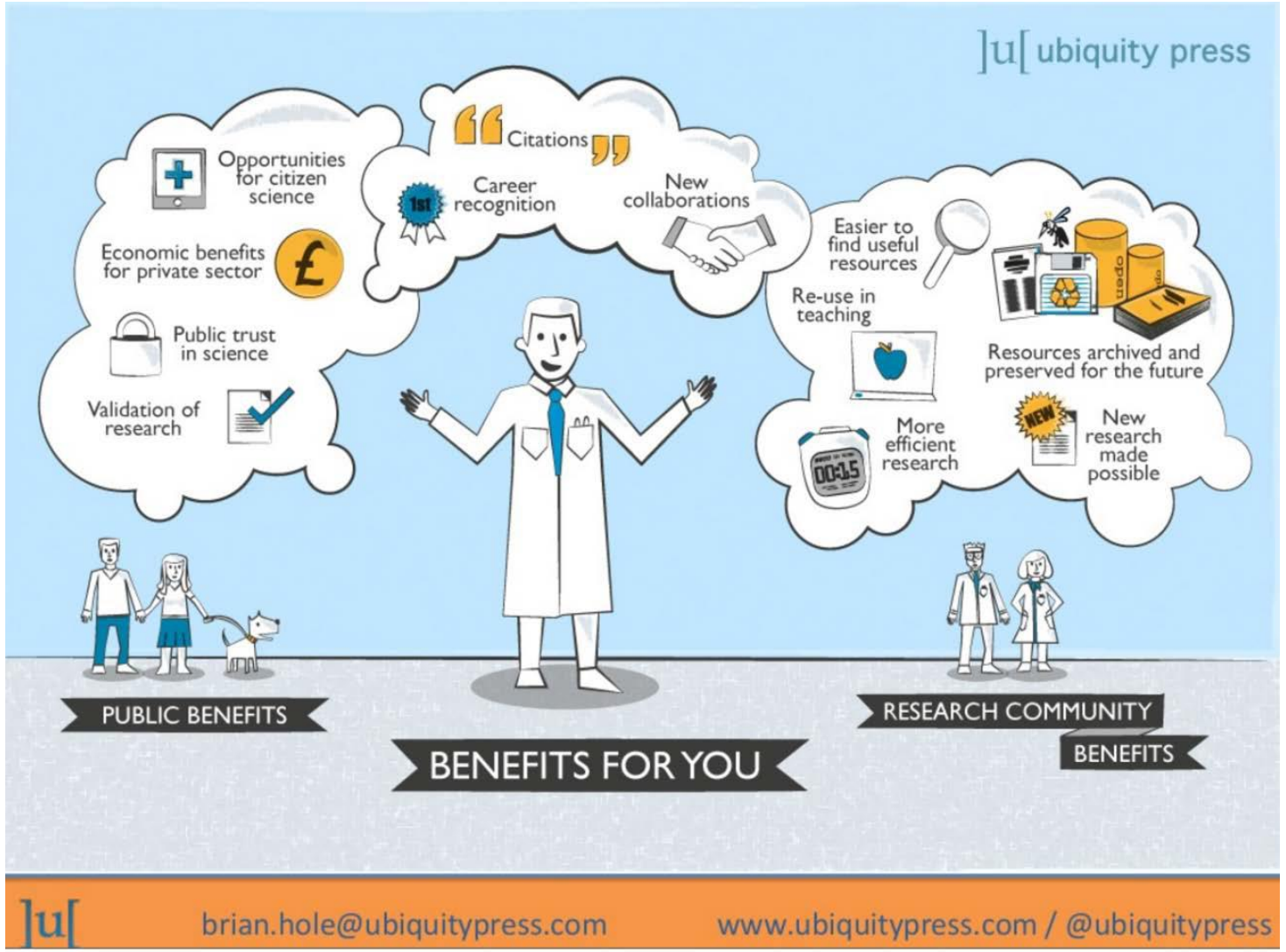
●●●

Notre ambition est de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec **les principes FAIR** (Facile à trouver, Accessible, Interopérable, Réutilisable), préservées et, quand cela est possible, ouvertes. (…)

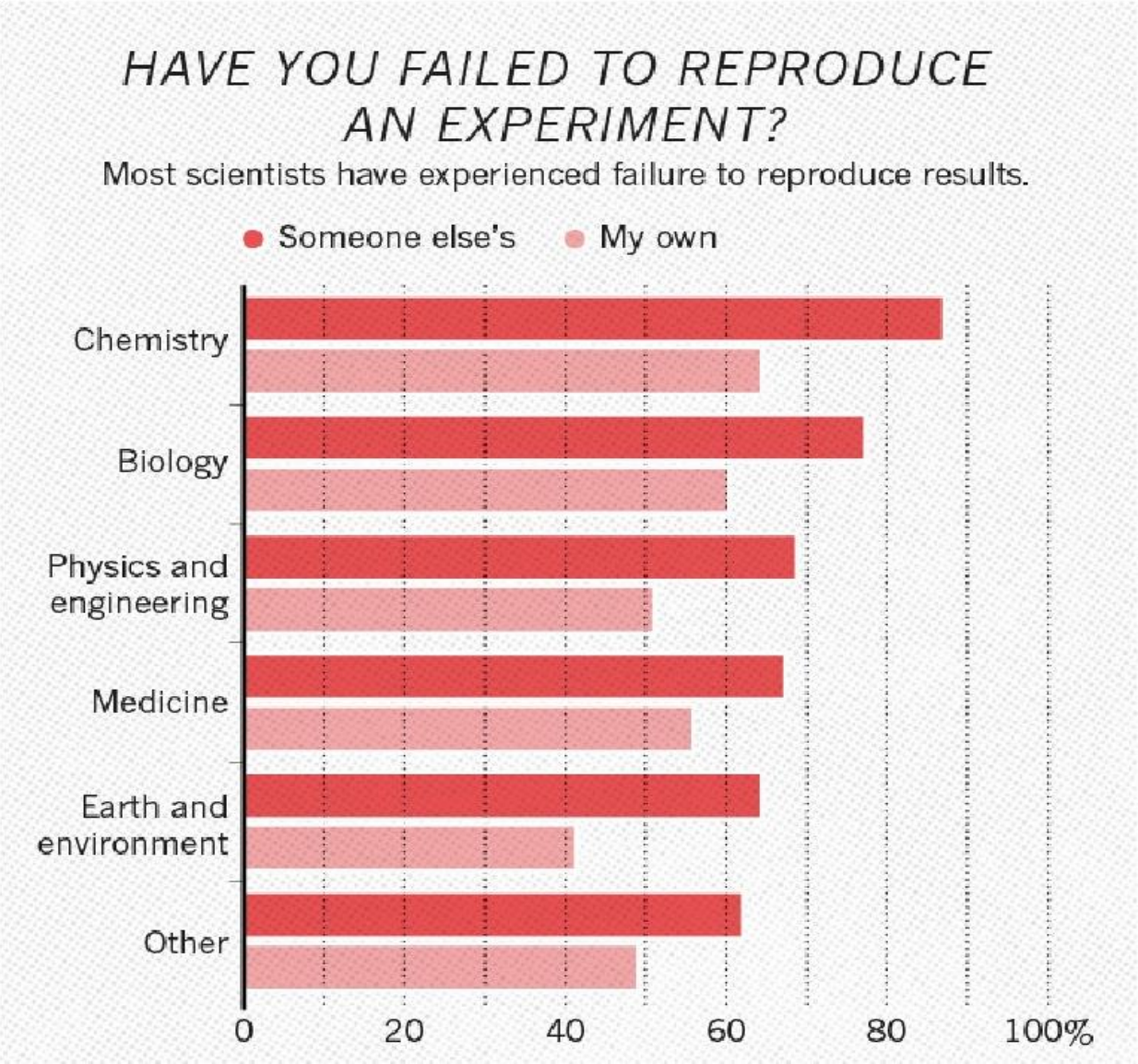
Prendre des précautions avant de partager

- Données relatives au potentiel scientifique et technique de la nation
- Données protégées par le droit d'auteur et réalisées dans le cadre de l'établissement
- Données protégées par un droit de propriété intellectuelle autre que le droit d'auteur
- Données personnelles
- Données de santé
- Données provenant d'un tiers

Avantages du partage des données

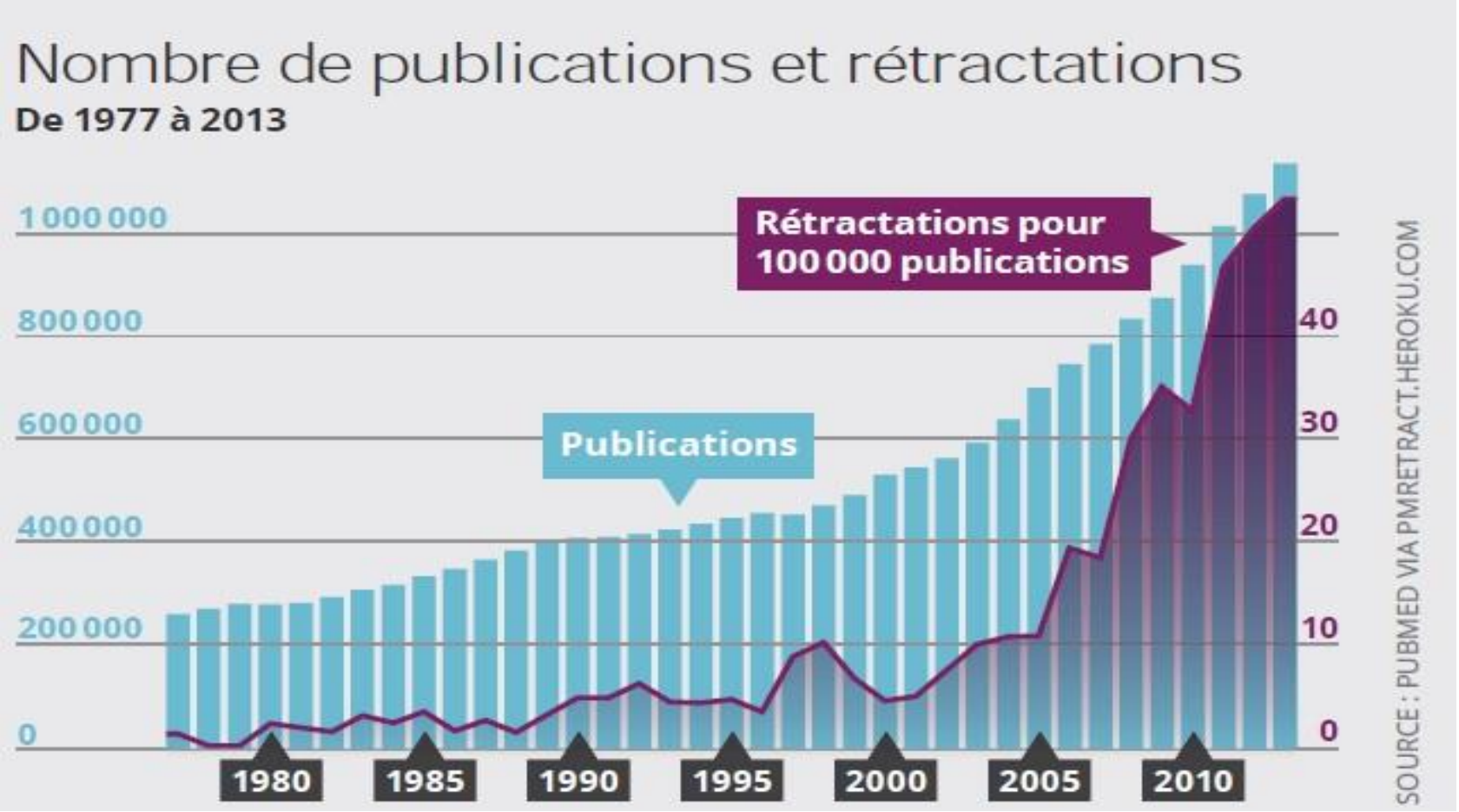


Crise de la reproductibilité?



(M. Baker, Nature, vol. 533, 2016)

Problème d'intégrité scientifique



Paysage complexe de l'éthique de la recherche

- Comités d'éthique des grands instituts (CNRS, INSERM, INRIA, INRA …)
- Comités d'éthique nationaux (CCNE, CERNA, Académies des Sciences)
- Comités de Protection des Personnes (CPP)
- Comités d'Ethique pour les Recherches Non Interventionnelles (CERNI)
- Institutional Review Boards (IRB)
- Référent Intégrité Scientifique + Office Français de l'Intégrité Scientifique (OFIS)
- Référent Déontologue

MERCI DE VOTRE PARTICIPATION