Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DES ARMÉES

SGA
Secrétariat général pour l'administration

Innovations et statistiques officielles

Rencontres Économiques de la Défense n°8

CALZADA C., GAUTHIER L.
DAF/QEFI/OED

22 juin 2017

Plan

- Le paysage de la production statistique officielle
- Le nouvel Age des data
- Des modèles pour prévoir
- Quels impacts sur la production statistique
- Quels impacts sur la statistique publique
- La statistique publique bouge

Paysage de la production statistique officielle

Gouvernance

- SSE (Système Statistique Européen) : programme statistique européen 2013-2017
- Eurostat (CE) + INS (Instituts Nationaux de Statistiques) + Autorités nationales … SSM Défense (OED)

Les statistiques officielles sont des biens publics

Code des bonnes pratiques de la statistique européenne
(Règlement (CE) nº 223/2009)

- Indépendance - Mandat pour la collecte - Adéquation des ressources
- Engagement sur la qualité - Secret statistique - Impartialité et objectivité
- Méthodologie solide - Procédures statistiques adaptées - Charge non excessive pour les déclarants
- Rapport coût-efficacité - Pertinence - Exactitude et fiabilité
- Actualité et ponctualité - Cohérence et comparabilité - Accessibilité et clarté

Le nouvel Age des data

- Trends technologiques : Electronic publishing (1990s) ➔ e-Business (2000s) ➔ d-Business (2010s) : Uber, Amazon Echo, web sémantique, …
- Trends sur l'information et les données : transferts de propriété légale, monétisation des données, désinformation / communication politique, Analytics vs Statistics, …
- Un écosystème compétitif : Géants (Gafa), groupements d'intérêts acquis, nouveaux entrants, données comme générateurs de marques, …
- Passage d'une logique de stock à une logique de flux

Le nouvel Age des data

Définitions

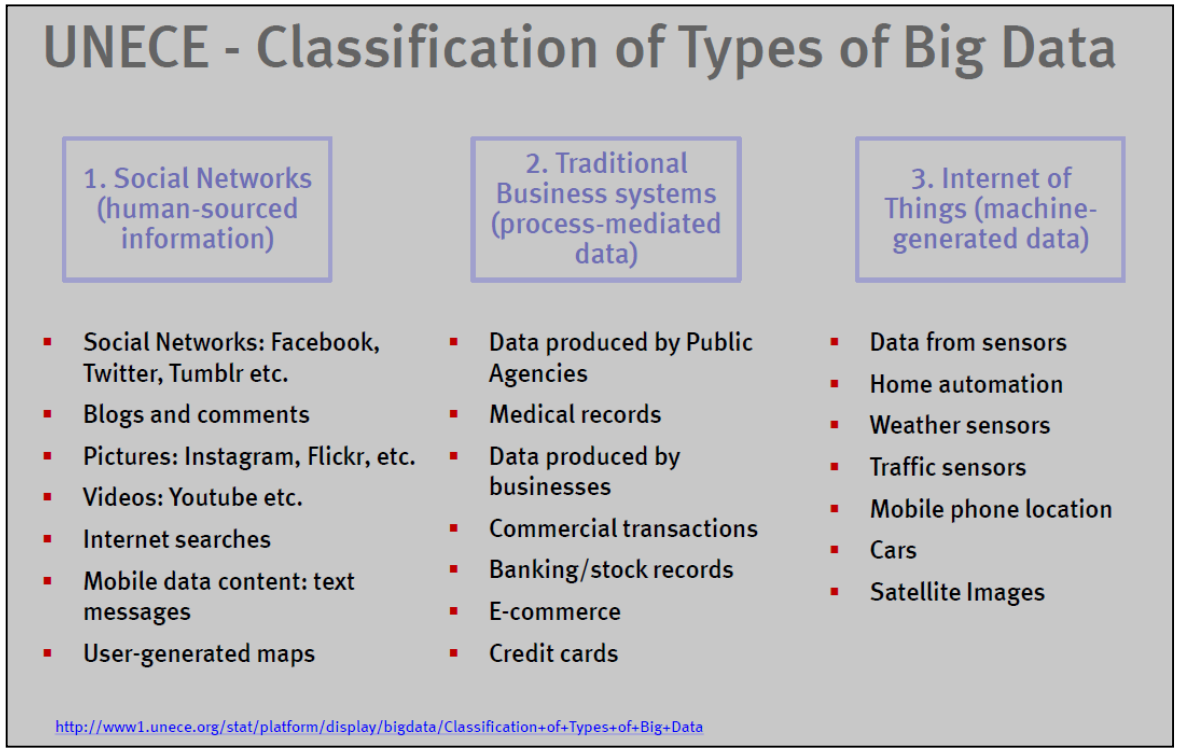
- Gartner's (2012) definition of Big Data (3Vs) :

" *Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making* "

- Machine Learning (Wikipedia) :

" *Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders* ".

Le nouvel Age des data



Des modèles pour prévoir

Vision classique

- Des modèles pour comprendre

Vision « Big Data Analytics »

- Modèles pour prévoir
- Capacité prédictive sur de nouvelles observations : « généralisation »
- Modèles issus des données (3 types d'échantillons)
- Le dilemme biais-variance

Quels impacts sur la production statistique

Innovations et changements dans le cycle de production statistique

- Utilisation de sources multiples : extension des données administratives et BigData
- « *Mashup* » de données : assembler et réassembler des données issues de multiples sources
- « Usines » à données
- Services d'analyses de données pour « *prosumers* »
- Nécessité de proposer de nouveaux produits statistiques pour répondre à de nouveaux besoins :
  - produits statistiques sur-mesure
  - enquêtes conduites par les données
  - modélisation économique
  - prévisions et projections
- Moderniser les processus de production

Quels impacts sur la production statistique

Bénéfices attendus

- Des résultats rapides
- Coûts moindres
- Plus de précision : petites populations, zonages fins, …
- Exhaustivité
- Une charge statistique moindre pour les répondants

Les questions qui se posent au statisticien

- Absence de contrôle sur la production des données
- Manque de vérité de terrain
- Qualité et précision des variables (capteurs, caméras, téléphonie / réseaux sociaux,
  e-commerce)
- Garantir un accès permanent à l'information
- Concurrence avec le privé
- Risque de dégradation de l'image des INS
- Nécessité de protéger la confidentialité : risque de ré-identification, questions éthiques

Quels impacts sur la production statistique

- Transition de méthodologies d'échantillonnage sur des populations finies à la modélisation statistique et au machine learning
- De concepteurs de processus de collecte de données à des concepteurs de produits statistiques
- Accréditation et certification risquent de devenir des activités majeures des INS
- Concurrence avec le privé
- Compétition pour les data scientists

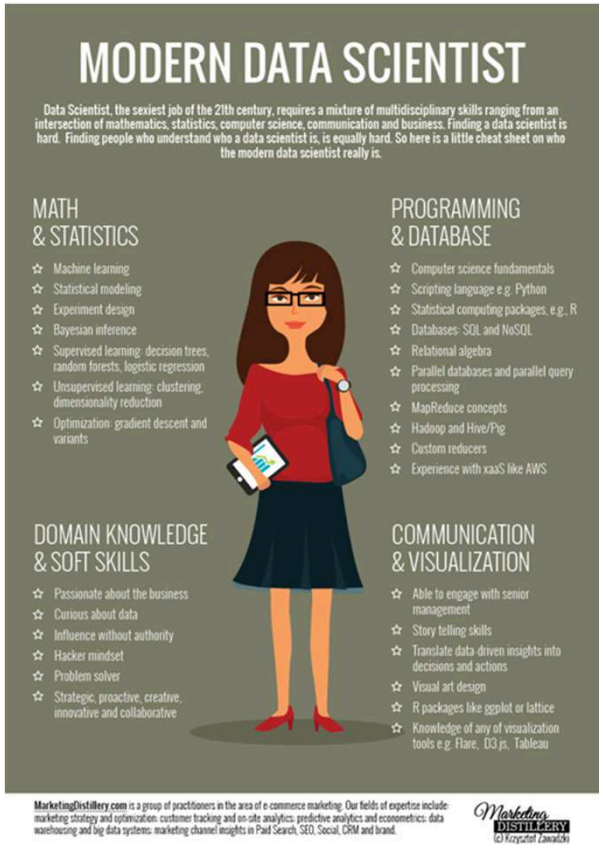
La statistique publique bouge

La statistique publique bouge

- « *Passer d'un monde d'enquêtes à un monde de données multisources et multimode* » (W. Radermacher - ancien DG Eurostat)
- « Change the factory »
- Dimensions des données : de relations 1 à 1 à des relations m à k
- Entrepôts de données de statistiques officielles :
  - générés par des sources digitales et d'enquêtes
  - dont l'objectif initial n'est pas statistique
  - la modélisation statistique devenant l'activité principale
- ESS Big Data task Force (Eurostat)
- ESS Vision 2020 et Insee 2025

iStatisticien

- Du Statisticien au iStatisticien (data scientist)
  - Compétences en : statistique, mathématiques, informatique, qualité, éthique, management de processus, communication, visualisation, etc…



La statistique publique bouge

Quatre exemples parmi d'autres au sein du SSP …

- Des données de caisse pour le calcul de l'indice des prix à la consommation
- Les statistiques sur l'emploi :
  - Le marché numérique du travail
  - Comment prévoir l'emploi en lisant le journal
- Identifier les cas de violences intra familiales dans les dépôts de plainte

La statistique publique bouge

Des données de caisse pour le calcul de l'indice des prix à la consommation

- Le projet « Données de caisse » de l'Insee
- Les relevés de prix collectés par les enquêteurs dans les magasins remplacés par des données enregistrées, par les enseignes de la grande distribution lors du paiement en caisse
- Volume de l'ordre de 40 Giga-Octets par semaine

- Périmètre limité aux articles alimentaires industriels, aux produits d'hygiène-beauté et aux produits d'entretien de la maison (17% de la consommation des ménages)
- Question de l'effet qualité dans le cadre de remplacement de produit
- Norvège (2001), Pays-Bas (2002), Suisse (2008), Suède (2012), Belgique et Danemark (2016), …
- Méthodologie du calcul de l'indice non harmonisée
- L'article 12 de la Loi pour une république numérique modifie la loi du 7 juin 1951 pour introduire un accès aux données privées

- Le Billion Prices Project (Alberto Cavallo, Roberto, MIT Sloan School of Management - Cambridge, 2006). 15 millions de prix journaliers, 1000 enseignes dans plus de 60 pays (http://www.thebillionpricesproject.com/)

La statistique publique bouge

L'utilisation des agrégateurs d'offres d'emploi

- WP1 Webscraping job vacancies (ESSnet Big Data)

- Agrégateurs d'offres d'emploi : indexation des offres d'emplois publiées sur le web pour offrir aux candidats la possibilité d'effectuer une recherche globale au travers d'une interface unique (Keljob, option carrière, Indeed.fr ...)

- La question de l'identification des agrégateurs : plusieurs centaines en France, spécialisées sur une profession, une région, … leur nombre évolue régulièrement
- Champ : se limiter aux sites les plus importants, ou développer des partenariats avec plusieurs sites (Pôle emploi et ses « sites partenaires »)
- L'absence de langage commun pour décrire les offres : hétérogénéité syntaxique (formats) et de catégorisation (métiers, compétences)
- Doublons, les employeurs peuvent poster la même offre sur plusieurs sites, avec un niveau de description qui peut varier selon le site utilisé

- Dévoilée fin 2016, la Cloud Jobs API (Google) propose aux recruteurs et aux candidats un « langage commun », dispositif cherchant à créer un langage commun pour rassembler toutes les offres et demandes d'emploi sous une même nomenclature via les services de diffusion d'offres d'emploi et de CV.…
- Google Hire, offrant aux entreprises des fonctionnalités de gestion de leurs recrutements en ligne

La statistique publique bouge

Comment prévoir l'emploi en lisant le journal

- Les médias annoncent régulièrement des décisions d'entreprises qui affectent directement le marché du travail
- Le journal Le Monde :
  - Plus d'un million d'articles publiés depuis 1990
  - Combiner modèles statistiques et d'analyse textuelle sur un échantillon de 200.000 textes environ.
  - « *En classant ces articles selon leur tonalité, positive ou négative, à partir d'une liste de mots, il est possible de calculer un indicateur mensuel de sentiment médiatique relatif à l'emploi ou à la situation économique de manière plus générale* »
  - Signal rapide, pertinent et lisible sur les fluctuations de court terme de l'économie
  - Disponible rapidement, presque en temps réel

  - « *Lorsqu'il est introduit dans un modèle de prévision à très court terme d'emploi salarié, l'indicateur de sentiment médiatique apporte en général une réelle information : à partir du deuxième mois du trimestre, il améliore significativement la prévision par rapport à un modèle simple incluant uniquement les variations passées de l'emploi et de l'activité* »

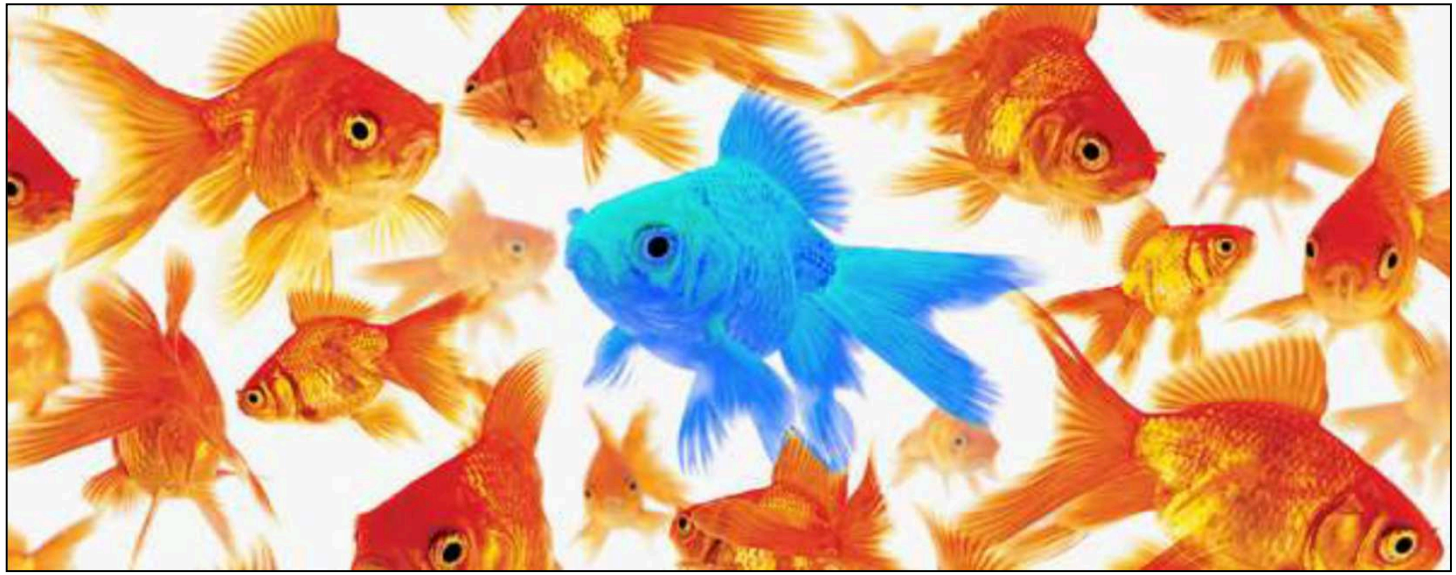
La statistique publique bouge

Identifier les cas de violences intra familiales dans les dépôts de plainte

- Identifier pour chaque plainte déposée, si elle peut être considérée comme une VIF
- Dépôts de plainte petite couronne (92, 93, 94), 3 millions d'enregistrements, plusieurs années
- Informations sur la procédure, le fait, la personne, « champ libre »
- « champ libre » : informations permettant d'identifier un cas de VIF, par exemple les termes explicites "violences conjugales", mais également d'autres formulations inconnues a priori
- Tenter de déterminer les cas de violences intra familiales à partir de l'information textuelle du « champ libre », on cherche à construire des règles de classement à partir de l'apparition de mots clé
- Techniques d'analyse textuelle et de machine learning
- Evolution de la proportion de cas de VIF au cours du temps
- Limites : textes très courts, mal orthographiés et utilisation de mots métiers

Conclusion

Produire des statistiques à forte valeur ajoutée …



… afin de faire la différence dans un océan d'informations

Merci de votre attention

Vos questions …