

Bases de données : un peu de culture...

Karën Fort

karen.fort@sorbonne-universite.fr

Sources d'inspiration

- ▶ (Premier) Cours d'informatique de première année à l'école des Mines de Nancy, par P. Vaxivière, avec son accord
- ▶ <https://www.cite-telecoms.com/accueil/cite-des-telecoms/les-peres-fondateurs/hollerith/>
- ▶ PBS Early Computing: Crash Course Computer Science #1
- ▶ Rapport d'activité de la CNIL (1999)
- ▶ Wikipédia

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

"All your data are belong to us"

Pour finir

L'informatique et les données : un peu d'histoire

Le dénombrement (de la population)

La mécanisation du comptage

Les dérives du comptage : le fichage

"All your data are belong to us"

Pour finir

Contexte de l'informatique

Aucune technique ne surgit du néant

- ▶ Nécessité du stockage et du traitement de l'information
- ▶ Nécessité de la transmission d'informations
- ▶ Nécessité de calculs rapides et exacts

Pour qui ? Pour quoi ?

Contexte de l'informatique

Nécessité de l'information

Tout **gouvernement** a besoin de données sur ses peuples :

- ▶ Données quantitatives
- ▶ Données qualitatives

L'informatique et les données : un peu d'histoire

Le dénombrement (de la population)

La mécanisation du comptage

Les dérives du comptage : le fichage

"All your data are belong to us"

Pour finir

Les dénombrements de population

Jean Bodin : « La République » (1576)

Or les utilités qui reviennent au public du dénombrement qui se faisoit, estoyent infinies. Car premièrement quant aux personnes on savoit et le nombre et l'âge et la qualité ; combien on en pourrait tirer, fust pour aller en guerre, ou pour demeurer, soit pour envoyer en colonies, fust pour employer aux labeurs et corvées des réparations, et fortifications publiques, fust pour savoir les provisions ordinaires et les vivres estoyent nécessaires aux habitants de chacune ville : et principalement quand il fallait soutenir le siège des ennemis : à quoi il est impossible de remédier, si on ne sait le nombre des sujets.

Les dénombrements de population

1328 (Philippe VI – Guerre de 100 ans) :

Estat des paroisses et feux des bailliages et sénéchaussées de France
(Chambre des comptes)

- ▶ 23 671 paroisses
- ▶ 2 469 987 feux

(Sans la Bourgogne, la Flandre, la Gascogne, la Bretagne, l'Anjou...)

Les dénombrements de population

1397 tailles en Bourgogne

1492 Charles VIII

~ 1503 Louis XII

~ 1525 François Ier

~ 1560 Charles IX décide d'un nouveau recensement de la population.

Début des grandes enquêtes statistiques

- 1664 Colbert, premier dénombrement avec une méthodologie
- 1686 Vauban propose sa Méthode générale et facile pour le dénombrement des peuples, avec une méthode de tableaux donnant la répartition par lieux, maisons, professions. . .
⇒ Création des agents recenseurs
- 1693 les enquêtes contiennent des informations sur les biens, grâce à des tableaux pré-imprimés transmis aux intendants

Début des grandes enquêtes statistiques

1720 Claude Saugrain publie *Le Nouveau Dénombrement du royaume par généralités, élections, paroisses et feux*.

Sa méthode de calcul permet de recenser le nombre de contribuables en fonction d'un nombre d'habitants par feu, variable selon les régions.

1744-45 dénombrement du contrôleur des finances, Orry

1760-62 dénombrement du contrôleur général Bertin

1766 Louis Messance propose une méthode avec des échantillons pour **vérifier la validité** des estimations

Les enquêtes statistiques

- Apparition d'une méthodologie d'acquisition et de validation des données
- Apparition d'un modèle

Mais :

- ▶ problème de la **quantité de données** croissante. . .
quand une loi électorale s'en mêle. . .

L'informatique et les données : un peu d'histoire

Le dénombrement (de la population)

La mécanisation du comptage

Les dérives du comptage : le fichage

"All your data are belong to us"

Pour finir

Les enquêtes statistiques

Obligation légale aux États-Unis du recensement décennal
(Article I, Section 2 de la Constitution) :

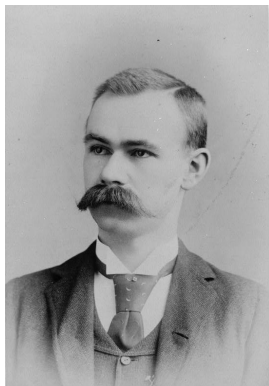
Representatives and direct Taxes shall be apportioned among the several States [...] according to their respective Numbers [...]. The actual Enumeration shall be made within three Years after the first meeting of the Congress of the United States, and within every subsequent Term of ten Years.

Le dépouillement de 1880 a pris **7 ans** !

→ Appel d'offre pour mécaniser la saisie et le dépouillement de celui de 1890

Les enquêtes statistiques

Herman Hollerith remporte le marché de mécanisation de la saisie et fonde la Tabulating Machine Co., qui deviendra l'International Business Machine (IBM)



Adam Schuster - Flickr: Proto IBM, CC BY 2.0

Système électro-mécanique de comptage

Les données sont transcrites sur une **carte perforée** :

1	1	3	0	2	4	10	On	S	A	C	E	a	c	e	g		EB	SB	Ch	Sy	U	Sh	Hk	Br	Rm
2	2	4	1	3	E	15	Off	IS	B	D	F	b	d	f	h		SY	X	Fp	Cn	R	X	Al	Cg	Kg
3	0	0	0	0	W	20		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	1	1	1	1	0	25	A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B	2	2	2	2	5	30	B	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
C	3	3	3	3	0	3	C	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
D	4	4	4	4	1	4	D	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
E	5	5	5	5	2	C	E	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
F	6	6	6	6	A	D	F	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
G	7	7	7	7	B	E	G	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
H	8	8	8	8	a	F	H	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
I	9	9	9	9	b	c	I	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Celle-ci est introduite dans la machine, munie de tiges reliées à un compteur. Lorsqu'une tige est face à un trou, il y a contact électrique et l'information est comptabilisée.

Résultats ?

L'énorme masse de données est traitée en 2,5 ans au lieu des 13 années initialement prévues !

Quelles données ?



Quelles sont les données traitées ici ?

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

La mécanographie au service du pire

Un peu plus sur le numéro de sécurité sociale

"All your data are belong to us"

Pour finir

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

La mécanographie au service du pire

Un peu plus sur le numéro de sécurité sociale

"All your data are belong to us"

Pour finir

Mécanographie en 1939-1945

Démographie - Fichage - Logistique

- ▶ Dehomag (Deutsche Hollerith-Maschinen Gesellschaft) en 1945
- ▶ 6000 machines : 39 Millions de cartes/mois
 - ▶ État des wagons : nombres et position, toutes les 48 h
 - ▶ Recensement de la Tchécoslovaquie en 3 mois
 - ▶ *Arbeitsstatistic*, fichier de tous les professionnels sur tous les territoires du Reich

Le rôle de René Carmille, père de l'Insee

Met en place la mécanographie dans l'administration française

Pendant la guerre (contrôleur général des armées) :

- ▶ crée le **numéro de sécurité sociale**
- ▶ recense les mobilisables
- ▶ transmet le fichier à Alger
- ▶ repéré par les nazis, déporté, meurt à Dachau



Julien Demade, CC BY-SA 3.0

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

La mécanographie au service du pire

Un peu plus sur le numéro de sécurité sociale

"All your data are belong to us"

Pour finir

Le « numéro de Français »

René Carmille avait prévu un numéro matricule à 12 chiffres :

- ▶ deux pour l'année de naissance
- ▶ deux pour le mois de naissance
- ▶ deux pour le département de naissance
- ▶ trois pour la commune de naissance
- ▶ trois pour un numéro d'ordre dans le mois de naissance

puis est ajouté un 13e chiffre en première colonne, pour le sexe

Un identifiant fiable et stable

= numéro d'inscription au répertoire national d'identification des personnes physiques

Cas	Positions	Signification
Tous	1	sexe : 1 pour les hommes, 2 pour les femmes, 3 ou 7 pour les personnes étrangères de sexe masculin en cours d'immatriculation en France ^{9,A} , 4 ou 8 pour les personnes étrangères de sexe féminin en cours d'immatriculation en France ^{9,A}
	2 et 3	deux derniers chiffres de l'année de naissance (ce qui donne l'année à un siècle près)
	4 et 5	mois de naissance ^B
A	6 et 7	département de naissance métropolitain (2A ou 2B pour la Corse) ^C
	8, 9 et 10	code officiel de la commune de naissance ^{C'D}
B	6, 7 et 8	département de naissance en outre-mer ^C
	9 et 10	deux chiffres du code commune de naissance ^{C'D}
C	6 et 7	naissance hors de France ^C
	8, 9 et 10	identifiant du pays de naissance ^C
Tous	11, 12 et 13	numéro d'ordre de la naissance dans le mois et la commune (ou le pays) ^{D'E}
	14 et 15	clé de contrôle = complément à 97 du nombre formé par les 13 premiers chiffres du NIR modulo 97 ^F (complément au NIR pour la Sécurité sociale)

https://fr.wikipedia.org/wiki/Num%C3%A9ro_de_s%C3%A9curit%C3%A9_sociale_en_France

Un numéro pas comme les autres


« Je ne suis pas un numéro ! »

Ce numéro est [...] porté sur la carte d'assuré social, sur les feuilles de soins, sur les décomptes de prestations. En revanche, il n'apparaît ni sur la carte d'identité, ni sur le passeport, ni sur le permis de conduire, ni sur la déclaration de revenus, ni sur les relevés bancaires, ni encore sur le livret scolaire.

Pourquoi ? Parce que ce n'est pas un numéro comme un autre. [CNIL, 1999]



Des trous ?

 Pourquoi certains chiffres de la première colonne ne sont pas (et n'ont sans doute jamais été) utilisés (5 et 6, notamment) ?

Instruction du 30 mai 1941

la première composante est ainsi définie : 1 et 2 [selon le sexe] désignent les citoyens français y compris les Juifs, 3 et 4 les "Indigènes d'Algérie et de toutes colonies sujets français, à l'exception des Juifs", 5 et 6 les Juifs indigènes sujets français, 7 et 8 les étrangers y compris les Juifs

→ la constitution numérique du NIR est significative !

Un numéro pas comme les autres

utilisation soumise à accord de la CNIL

Loi informatique et libertés du 6 janvier 1978 :

l'utilisation du répertoire national d'identification des personnes physiques en vue d'effectuer des traitements nominatifs est autorisé par décret en Conseil d'État pris après avis de la Commission.

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

"All your data are belong to us"

Définition

Types de données

Pour finir

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

"All your data are belong to us"

Définition

Types de données

Pour finir

Définition

<http://www.cnrtl.fr/definition/donn%C3%A9e>

■ **DONNÉE**, subst. fém.

DONNER, verbe.

DONNÉ, ÉE, part. passé, adj. et subst.

A. – MATH. Quantité connue dans l'énoncé d'un problème et qui sert à trouver la solution. *Les données d'un problème* (Ac.1932) :

- 1. ... elle [la langue de l'algèbre] fournit les moyens de soumettre les grandeurs aux mêmes opérations de calcul, sans distinction de **données** et d'*inconnues*. COURNOT, *Essai sur les fondements de nos connaissances*, 1851, p. 391.

B. – P. ext.

1. Ce qui est connu et admis, et qui sert de base, à un raisonnement, à un examen ou à une recherche. *Toute question de politique intérieure doit être vidée d'après les données de la statistique départementale* (PROUDHON, *Propriété*, 1840, p. 340). *Les données actuelles de l'embryologie* (BERGSON, *Évol. créatr.*, 1907, p. 25) :

- 2. ... cette seule constatation doit nous inciter à chercher, pour les phénomènes de régénération, une interprétation moins philosophique et plus conforme aux **données** de l'expérience. J. ROSTAND, *La Vie et ses probl.*, 1939, p. 72.

– Spéc. „Ensemble des indications enregistrées en machine pour permettre l'analyse et/ou la recherche automatique des informations” (CROS-GARDIN 1964). *Banque de données; données documentaires, données lexicales.*

Exemples

Créer un compte

C'est gratuit (et ça le restera toujours).

Date de naissance

25 avr 1994

[Pourquoi indiquer ma date de naissance ?](#)

☐ Femme ☐ Homme

En cliquant sur Inscription, vous acceptez nos [Conditions générales](#). Découvrez comment nous recueillons, utilisons et partageons vos données en lisant notre [Politique d'utilisation des données](#) et comment nous utilisons les cookies et autres technologies similaires en consultant notre [Politique d'utilisation des cookies](#). Vous recevrez peut-être des notifications par texto de notre part et vous pouvez à tout moment vous désabonner.

<https://www.facebook.com>

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

"All your data are belong to us"

Définition

Types de données

Pour finir



Article 4 - Définitions

Aux fins du présent règlement, on entend par :

1. «données à caractère personnel», toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée») ; est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale;

<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1>

Données et données

Art. 4 du RGPD

13. «données génétiques», les données à caractère personnel relatives aux caractéristiques génétiques héréditaires ou acquises d'une personne physique qui donnent des informations uniques sur la physiologie ou l'état de santé de cette personne physique et qui résultent, notamment, d'une analyse d'un échantillon biologique de la personne physique en question;
14. «données biométriques», les données à caractère personnel résultant d'un traitement technique spécifique, relatives aux caractéristiques physiques, physiologiques ou comportementales d'une personne physique, qui permettent ou confirment son identification unique, telles que des images faciales ou des données dactyloscopiques;
15. «données concernant la santé», les données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne;


<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1>

Protection des données sensibles

Art. 9 du RGPD

Article 9 - Traitement portant sur des catégories particulières de données à caractère personnel


1. Le traitement des données à caractère personnel qui révèle l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique sont interdits.

<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2>

Protection des données sensibles ?

Art. 9 du RGPD : suites

2. Le paragraphe 1 ne s'applique pas si l'une des conditions suivantes est remplie :

- a) la personne concernée a donné son consentement explicite au traitement de ces données à caractère personnel pour une ou plusieurs finalités spécifiques, sauf lorsque le droit de l'Union ou le droit de l'État membre prévoit que l'interdiction visée au paragraphe 1 ne peut pas être levée par la personne concernée;
- b) le traitement est nécessaire aux fins de l'exécution des obligations et de l'exercice des droits propres au responsable du traitement ou à la personne concernée en matière de droit du travail, de la sécurité sociale et de la protection sociale, dans la mesure où ce traitement est autorisé par le droit de l'Union, par le droit d'un État membre ou par une convention collective conclue en vertu du droit d'un État membre qui prévoit des garanties appropriées pour les droits fondamentaux et les intérêts de la personne concernée;
- c) le traitement est nécessaire à la sauvegarde des intérêts vitaux de la personne concernée ou d'une autre personne physique, dans le cas où la personne concernée se trouve dans l'incapacité physique ou juridique de donner  consentement;

<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2>

Protection des données sensibles ?

Art. 9 du RGPD : suites

- d) le traitement est effectué, dans le cadre de leurs activités légitimes et moyennant les garanties appropriées, par une fondation, une association ou tout autre organisme à but non lucratif et poursuivant une finalité politique, philosophique, religieuse ou syndicale, à condition que ledit traitement se rapporte exclusivement aux membres ou aux anciens membres dudit organisme ou aux personnes entretenant avec celui-ci des contacts réguliers en liaison avec ses finalités et que les données à caractère personnel ne soient pas communiquées en dehors de cet organisme sans le consentement des personnes concernées;
- e) le traitement porte sur des données à caractère personnel qui sont manifestement rendues publiques par la personne concernée;
- f) le traitement est nécessaire à la constatation, à l'exercice ou à la défense d'un droit en justice ou chaque fois que des juridictions agissent dans le cadre de leur fonction juridictionnelle;
- g) le traitement est nécessaire pour des motifs d'intérêt public important, sur la base du droit de l'Union ou du droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et des intérêts de la personne concernée.

<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2>

Protection des données sensibles ?

Art. 9 du RGPD : suites

h) le traitement est nécessaire aux fins de la médecine préventive ou de la médecine du travail, de l'appréciation de la capacité de travail du travailleur, de diagnostics médicaux, de la prise en charge sanitaire ou sociale, ou de la gestion des systèmes et des services de soins de santé ou de protection sociale sur la base du droit de l'Union, du droit d'un État membre ou en vertu d'un contrat conclu avec un professionnel de la santé et soumis aux conditions et garanties visées au paragraphe 3;

i) le traitement est nécessaire pour des motifs d'intérêt public dans le domaine de la santé publique, tels que la protection contre les menaces transfrontalières graves pesant sur la santé, ou aux fins de garantir des normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux, sur la base du droit de l'Union ou du droit de l'État membre qui prévoit des mesures appropriées et spécifiques pour la sauvegarde des droits et libertés de la personne concernée, notamment le secret professionnel;

j) le traitement est nécessaire à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques, conformément à l'article 89, paragraphe 1, sur la base du droit de l'Union  et droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et

<https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2>

Production de données

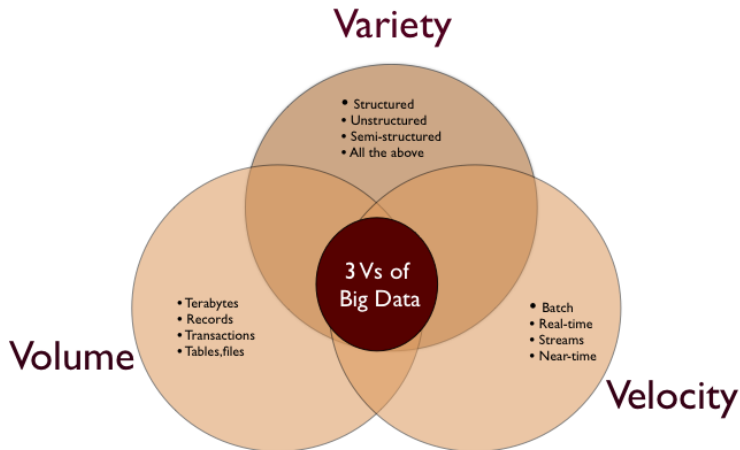
www.planetoscope.com/Internet-/1523-informations-publiees-dans-le-monde-sur-le-net-en-gigaoctets.html

En 2018, on estime que :

- ▶ 3,8 milliards d'humains utilisent Internet dans le monde (sur 7,7 milliards d'humains)
- ▶ 29 000 gigaoctets de données sont produites par seconde
- ▶ 90 % des données disponibles dans le monde ont été créées dans les deux dernières années

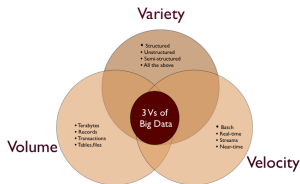
Données massives (*Big Data*)

Les 3 V : Variété, Volume, Vitesse



Données massives (*Big Data*)

Les 3 V : Variété, Volume, Vitesse



Évolutions technologiques :

- ▶ stockage : *cloud computing*
- ▶ temps de traitement : bases de données (géantes) NoSql

Conclusion de l'introduction

Base de données

- ▶ ensemble de données stockées dans un système informatique
- ▶ organisé selon un modèle de données pré-déterminé en fonction de la nature des informations qui y seront stockées
- ▶ les fichiers comportent des index destinés à accélérer les opérations de recherche et de tri

L'informatique et les données : un peu d'histoire

Les dérives du comptage : le fichage

"All your data are belong to us"

Pour finir

CQFR : Ce Qu'il Faut Retenir
TD



- ▶ nécessité de la mécanisation pour la gestion des données
- ▶ caractéristiques du NIR
- ▶ dangers et questions éthiques liés aux bases de données

Exercice 1

Allez sur le site de vente en ligne suivant :

<https://enventelibre.org/>

- ▶ listez dans un tableur les données nécessaires à une vente en ligne avec livraison :
 - ▶ prenez un exemple, avec un acheteur (vous) et un produit précis
- ▶ ajoutez au moins 10 autres ventes, avec notamment :
 - ▶ le même acheteur, avec un autre produit
 - ▶ le même produit, avec un acheteur différent

Exercice 1 : suite


- ▶ vous (premier acheteur) changez d'adresse et vous achetez un produit : ajoutez cet achat dans le tableur
- ▶ ajoutez la gestion des stocks : X produits sont disponibles à l'origine, il faut en déduire les achats pour prévenir les utilisateurs quand il n'y a plus de produits



Quels problèmes rencontrez-vous ? Comment les résoudre ?


Exercice 2

Une entreprise de cours particuliers de grande envergure (Acamodia) doit recruter des enseignants pour son activité.

 Comment peut-elle identifier ses employés de manière non ambiguë dans sa base de données (en évitant les problèmes d'homonymies) ?

Exercice 3

Lire ceci : <https://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html>

 Quelle différence faites-vous entre le NIR et le numéro de sécurité sociale américain ?