Séance 1.2: Web-Scraping

Données digitales, forces et faiblesses

Visseho Adjiwanou, PhD.

Plan de présentation

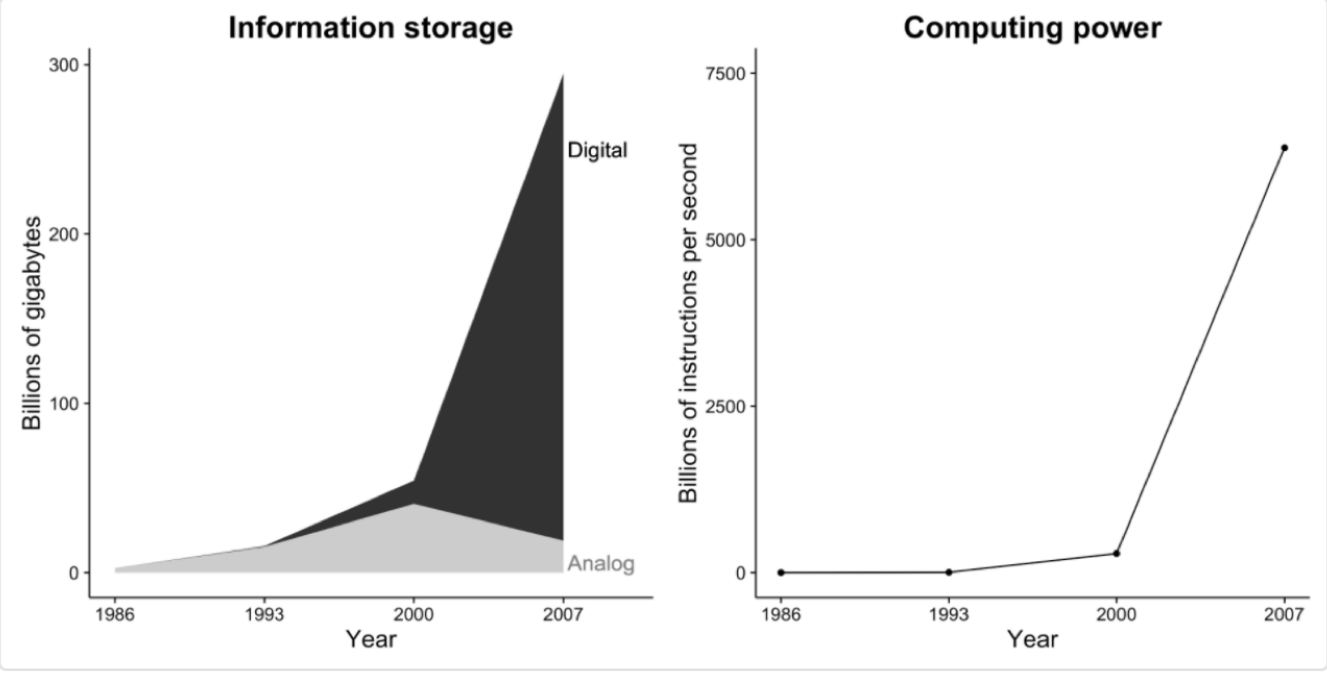
Données digitales

Introduction

- La dernière décennie a été témoin d'une quantité de plus en plus volumineuse de données numériques produites sur Internet qui décrivent le comportement humain et d'autres objets d'investigation scientifique.
- A cela s'ajoutent des volumes de numérisation de texte,
- et des données administratives de plus en plus volumineuses et accessibles

Introduction

- Comme le montre la figure ci-dessous, les dernières décennies ont non seulement vu une augmentation de la quantité de données textuelles, mais également une augmentation de la puissance de calcul qui est de plus en plus nécessaire pour les analyser.



Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik

Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik
- Cependant, l'une des définitions les plus courantes du Big Data se concentre sur les «3 V»:

Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik
- Cependant, l'une des définitions les plus courantes du Big Data se concentre sur les «3 V»:
- Volume,

Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik

- Cependant, l'une des définitions les plus courantes du Big Data se concentre sur les «3 V»:

- Volume,
- Variété et

Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik

- Cependant, l'une des définitions les plus courantes du Big Data se concentre sur les «3 V»:

- Volume,
- Variété et
- Vitesse.

Qu'est-ce que les données de traces numériques ou digitales?

- "Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L'utilisation de ces données pour la recherche nécessite donc une réutilisation." Matthew Salganik

- Cependant, l'une des définitions les plus courantes du Big Data se concentre sur les «3 V»:

- Volume,
- Variété et
- Vitesse.

- En gros, il y a beaucoup de données, dans une variété de formats, et elles sont constamment créées.

Qu'est-ce que les données de traces numériques ou digitales?

- Pour des fins de recherche sociale, Salganik pense qu'un meilleur endroit pour commencer est les 5 «W»:

- Who: Qui,
- What: Quoi,
- Where: Où,
- When: Quand ,
- Why: pourquoi.

Qu'est-ce que les données de traces numériques ou digitales?

- Pour des fins de recherche sociale, Salganik pense qu'un meilleur endroit pour commencer est les 5 «W»:

- Who: Qui,
- What: Quoi,
- Where: Où,
- When: Quand ,
- Why: pourquoi.

- De nombreux défis et opportunités créés par les sources de données volumineuses découlent d'un seul «W»: pourquoi?

Qu'est-ce que les données de traces numériques ou digitales?

"Tout comme l'invention du télescope a révolutionné l'étude du ciel, de même qu'en rendant l'immesurable mesurable, la révolution technologique dans les communications mobiles, sur le Web et sur Internet pourrait révolutionner notre compréhension de nous-mêmes et de nos interactions... Trois cents ans après qu'Alexandre Pope ait fait valoir que l'étude appropriée de l'humanité ne devrait pas se trouver dans les cieux mais en nous-mêmes, nous avons enfin trouvé notre télescope. Que la révolution commence."

— Duncan Watts (2011, p. 266)

Qu'est-ce que les données de traces numériques ou digitales?

> [J]ust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact . . . . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin.
>
> — Duncan Watts (2011, p. 266)

Qu'est-ce que les données de traces numériques ou digitales?

"... Les chercheurs peuvent maintenant observer le comportement, poser des questions, mener des expériences et collaborer de manière tout à fait impossible dans un passé récent. Ces nouvelles opportunités s'accompagnent de nouveaux risques: les chercheurs peuvent maintenant nuire aux gens d'une manière qui était impossible dans un passé récent."

- (Salganik, 2017)

Qu'est-ce que les données de traces numériques ou digitales?

1. Données en ligne créées et collectées par les entreprises

   - Sites de média sociaux
   - Données de recherche Web
   - Blogs / Autres forums Internet
   - Internet Archive
   - Données audio-visuelles

Qu'est-ce que les données de traces numériques ou digitales?

2. Données des appareils numériques dans le monde physique
   - données de caisse de supermarché ((Mas and Moretti 2009))
   - Données de téléphonie mobile (Blumenstock, Cadamuro and On, 2015)

Qu'est-ce que les données de traces numériques ou digitales?

3. Données créées par les gouvernements
   - Données administratives sur les sites Web: dossiers fiscaux, les dossiers scolaires et les dossiers de l'état civil
   - Numérisation de textes historiques / archives
   - Exemple:
     - données des compteurs de taxis numériques du gouvernement de la ville de New York (Farber 2015)
     - dossiers de vote recueillis par le gouvernement ont été utilisés dans une enquête (Ansolabehere and Hersh 2012) et d'une expérience (Bond et al. 2012)

Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :
- Même si, du point de vue des chercheurs, les grandes sources de données sont «trouvées», elles ne tombent pas du ciel.

Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :
- Même si, du point de vue des chercheurs, les grandes sources de données sont «trouvées», elles ne tombent pas du ciel.
- Au lieu de cela, les sources de données qui sont «trouvées» par les chercheurs sont conçues par quelqu'un dans un but précis.

Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :

Puisque les données "trouvées" sont conçues par quelqu'un, je vous recommande toujours :

1. d'essayer de comprendre autant que possible les personnes et les processus qui ont créé vos données.

Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :

Puisque les données "trouvées" sont conçues par quelqu'un, je vous recommande toujours :

1. d'essayer de comprendre autant que possible les personnes et les processus qui ont créé vos données.

2. Deuxièmement, lorsque vous réutilisez des données, il est souvent extrêmement utile d'imaginer l'ensemble de données idéal pour votre problème, puis de comparer cet ensemble de données idéal avec celui que vous utilisez.

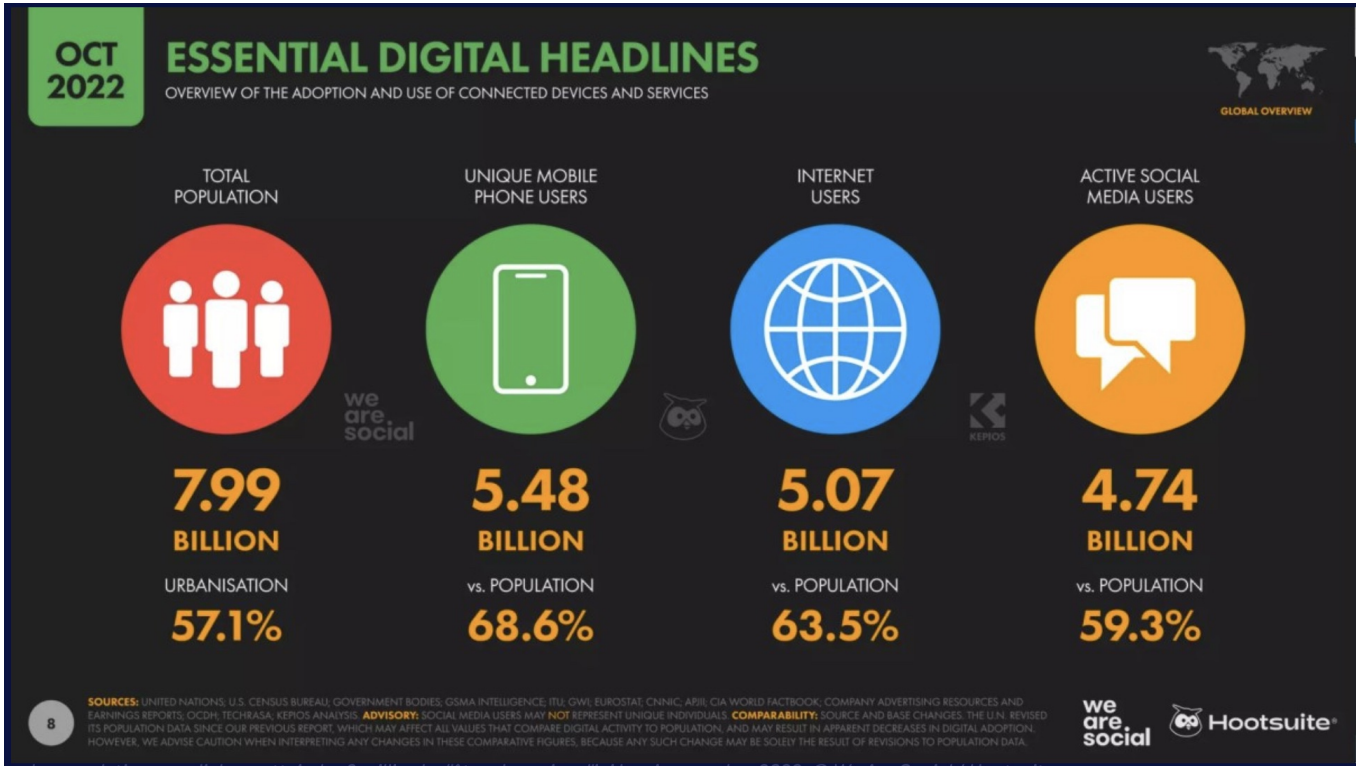
Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :
- Si vous n'avez pas recueilli vos données vous-même, il y a probablement des différences importantes entre ce que vous voulez et ce que vous avez.

Qu'est-ce que les données de traces numériques ou digitales?

- Deux précautions à prendre :
- Si vous n'avez pas recueilli vos données vous-même, il y a probablement des différences importantes entre ce que vous voulez et ce que vous avez.
- En notant ces différences, vous pourrez clarifier ce que vous pouvez et ne pouvez pas apprendre des données que vous avez et suggérer de nouvelles données que vous devriez collecter.

Quelques chiffres



Quelques chiffres

- 171 millions de nouveaux internautes sur les 12 derniers mois (+3,5 %)
- 190 millions de nouveaux utilisateurs des réseaux sociaux au cours des 12 derniers mois (+4,2 %)
- 96,1 % des internautes utilisent leur mobile pour surfer sur le web (+0,3 %),
- 6h37 passées en moyenne sur Internet par jour (en recul de 20 minutes),
- 2h28 passées en moyenne sur les réseaux sociaux par jour (+1 minute),
- 7,2 plateformes utilisées en moyenne par les socionautes chaque mois.

(Source: https://www.blogdumoderateur.com/chiffres-cles-internet-reseaux-sociaux-monde-octobre-2022/)

Caractéristiques des données digitales

Caractéristiques

1. Bon pour la recherche:

   - Volumineuse
   - Toujours active
   - Non réactive
   - Capture les relations sociales

Caractéristiques

2 Probématique pour la recherche

- Incomplète
- Inaccessible
- Non représentative
- Dérivante
- Algorithmiquement confondue
- Sale
- Sensible

1. Volumineuses

- Elles sont volumineuses

"[Notre] corpus contient plus de 500 milliards de mots, en anglais (361 milliards), français (45 milliards), espagnol (45 milliards), allemand (37 milliards), chinois (13 milliards), russe (35 milliards) et hébreu (2 milliards). Les œuvres les plus anciennes ont été publiées dans les années 1500. Les premières décennies ne sont représentées que par quelques livres par an, comprenant plusieurs centaines de milliers de mots. En 1800, le corpus atteint 98 millions de mots par an; en 1900, 1,8 milliard; et en 2000, 11 milliards. Le corpus ne peut pas être lu par un humain...

1. Volumineuses

- Elles sont volumineuses

. . . Si vous avez essayé de lire seulement les entrées en anglais de l'an 2000 seulement, au rythme raisonnable de 200 mots / min, sans interruptions pour la nourriture ou le sommeil, cela prendrait 80 ans. La séquence de lettres est 1000 fois plus longue que le génome humain: si vous l'écrivez en ligne droite, elle atteindrait la Lune 10 fois plus vite. " (Michel et al. 2011)

1. Volumineuses

- Pourquoi est-ce important?
  - Étude des phénomènes rares
  - Hétérogénéité: études de Raj Chetty et ses collègues sur la mobilité sociale.

1. Volumineuses

- Pourquoi est-ce important?
  - Étude des phénomènes rares
  - Hétérogénéité: études de Raj Chetty et ses collègues sur la mobilité sociale.
  - Utilisation des données fiscales de 40 millions de personnes pour estimer l'hétérogénéité de la mobilité intergénérationnelle entre les régions des États-Unis

1. Volumineuses

- Pourquoi est-ce important?
  - Étude des phénomènes rares
  - Hétérogénéité: études de Raj Chetty et ses collègues sur la mobilité sociale.
  - Utilisation des données fiscales de 40 millions de personnes pour estimer l'hétérogénéité de la mobilité intergénérationnelle entre les régions des États-Unis
- Détecter des petites différences: Par exemple, s'il y a deux interventions de santé publique et que l'une est légèrement plus efficace que l'autre, choisir une intervention plus efficace pourrait permettre de sauver des milliers de vies supplémentaires.

1. Volumineuses

- Pourquoi est-ce important?
  - Étude des phénomènes rares
  - Hétérogénéité: études de Raj Chetty et ses collègues sur la mobilité sociale.
  - Utilisation des données fiscales de 40 millions de personnes pour estimer l'hétérogénéité de la mobilité intergénérationnelle entre les régions des États-Unis
- Détecter des petites différences: Par exemple, s'il y a deux interventions de santé publique et que l'une est légèrement plus efficace que l'autre, choisir une intervention plus efficace pourrait permettre de sauver des milliers de vies supplémentaires.
- Faire de la randomisation (expérience dans les magasins en ligne par exemple)

2. Continues

- L'une des caractéristiques les plus attrayantes des données de traces numériques est leur collecte continue, contrairement aux enquêtes qui ne fournissent généralement qu'un bref instantané du monde social.

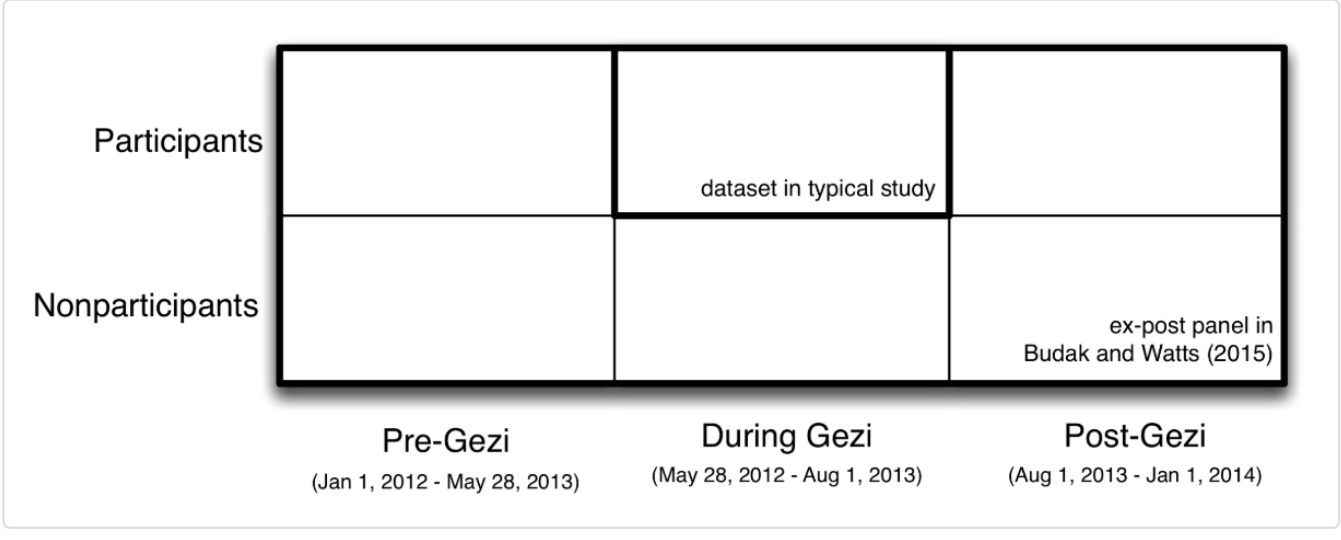
2. Continues

- L'une des caractéristiques les plus attrayantes des données de traces numériques est leur collecte continue, contrairement aux enquêtes qui ne fournissent généralement qu'un bref instantané du monde social.
- Permet d'étudier des évènements inattendus

2. Continues

- L'une des caractéristiques les plus attrayantes des données de traces numériques est leur collecte continue, contrairement aux enquêtes qui ne fournissent généralement qu'un bref instantané du monde social.
- Permet d'étudier des évènements inattendus
- Exemple: Ceren Budak et Duncan Watts (2015) dans l'étude sur les manifestations en Turquie: ils ont été en mesure d'estimer quels types de personnes étaient plus susceptibles de participer aux manifestations de Gezi et d'estimer les changements d'attitudes de participants et non-participants, à la fois à court terme (comparant les pré-Gezi à Gezi) et à long terme (comparant les pré-Gezi aux post-Gezi).

2. Continues


Participants
Nonparticipants
dataset in typical study
ex-post panel in
Budak and Watts (2015)
Pre-Gezi
(Jan 1, 2012 - May 28, 2013)
During Gezi
(May 28, 2012 - Aug 1, 2013)
Post-Gezi
(Aug 1, 2013 - Jan 1, 2014)

2. Continues

Table 2.1: Studies of unexpected events using always-on big data sources.

Unexpected event	Always-on data source	Citation
Occupy Gezi movement in Turkey	Twitter	Budak and Watts (2015)
Umbrella protests in Hong Kong	Weibo	Zhang (2016)
Shootings of police in New York City	Stop-and-frisk reports	Legewie (2016)
Person joining ISIS	Twitter	Magdy, Darwish, and Weber (2016)
September 11, 2001 attack	livejournal.com	Cohn, Mehl, and Pennebaker (2004)
September 11, 2001 attack	pager messages	Back, Küfner, and Egloff (2010), Pury (2011), Back, Küfner, and Egloff (2011)

3. Non réactives

- Un autre avantage important des données de traces numériques est qu'elles sont non réactives ou ne sont pas produites via une interaction entre les chercheurs et ceux qu'ils étudient.
- Dans certains cas, cela peut entraîner une réduction significative du biais de désirabilité sociale ou d'autres formes d'effet de l'intervieweur.

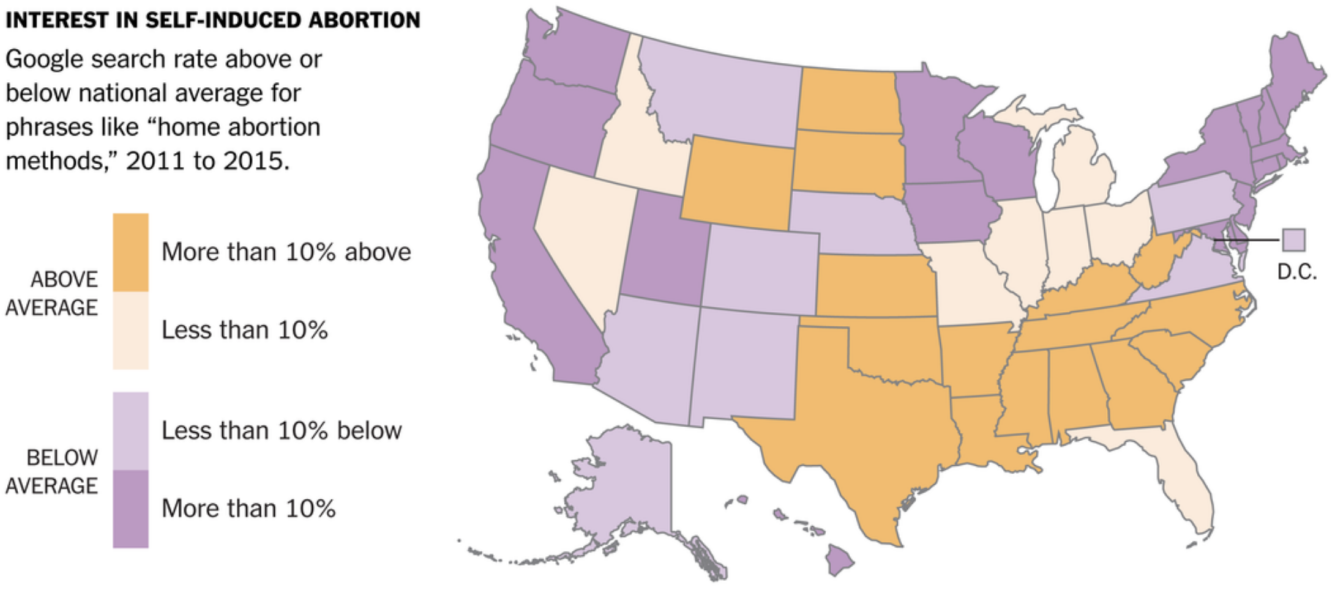
3. Non réactives

- Un autre avantage important des données de traces numériques est qu'elles sont non réactives ou ne sont pas produites via une interaction entre les chercheurs et ceux qu'ils étudient.
- Dans certains cas, cela peut entraîner une réduction significative du biais de désirabilité sociale ou d'autres formes d'effet de l'intervieweur.
- Exemple : Stephens-Davidowitz (2014) utilisé la prévalence des termes racistes dans les requêtes des moteurs de recherche pour mesurer l'animosité raciale dans différentes régions des États-Unis

3. Non réactives

- Un autre avantage important des données de traces numériques est qu'elles sont non réactives ou ne sont pas produites via une interaction entre les chercheurs et ceux qu'ils étudient.
- Dans certains cas, cela peut entraîner une réduction significative du biais de désirabilité sociale ou d'autres formes d'effet de l'intervieweur.
- Exemple : Stephens-Davidowitz (2014) utilisé la prévalence des termes racistes dans les requêtes des moteurs de recherche pour mesurer l'animosité raciale dans différentes régions des États-Unis
- Considérons, par exemple, l'utilisation des données de recherche Google pour étudier l'avortement volontaire (voir la figure ci-dessous).

3. Non réactives



3. Non réactives

- Ne reflète pas d'une manière ou d'une autre le comportement ou les attitudes des gens : «ce n'est pas que je n'ai pas de problèmes, je ne les mets tout simplement pas sur Facebook» (Newman et al. 2011).

- Les gens ont tendance à se présenter de la meilleure manière possible.

3. Non réactives

- Ne reflète pas d'une manière ou d'une autre le comportement ou les attitudes des gens : «ce n'est pas que je n'ai pas de problèmes, je ne les mets tout simplement pas sur Facebook» (Newman et al. 2011).
- Les gens ont tendance à se présenter de la meilleure manière possible.
- Affecté par les objectifs des objectifs des propriétaires de plate-forme

3. Non réactives

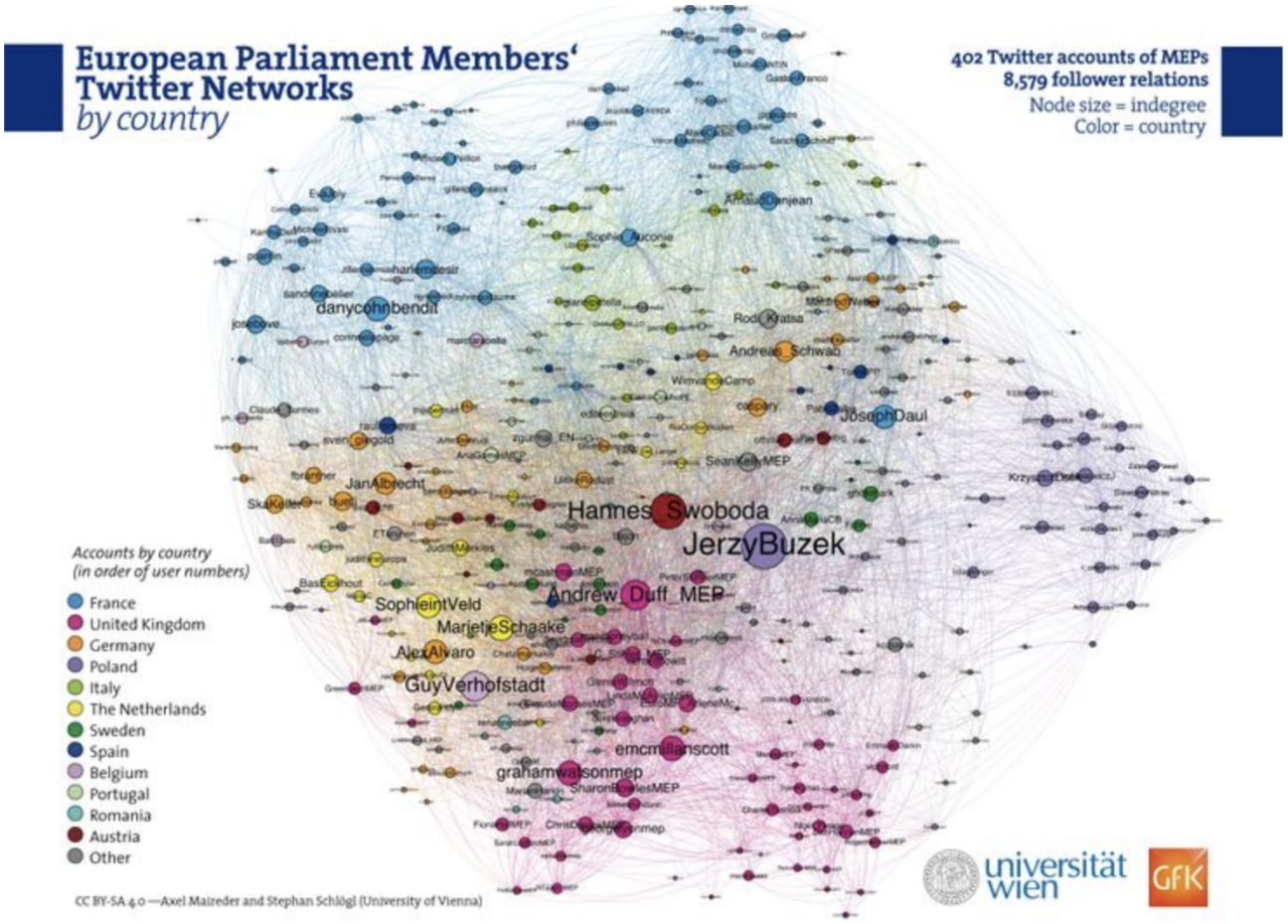
- Ne reflète pas d'une manière ou d'une autre le comportement ou les attitudes des gens : «ce n'est pas que je n'ai pas de problèmes, je ne les mets tout simplement pas sur Facebook» (Newman et al. 2011).
- Les gens ont tendance à se présenter de la meilleure manière possible.
- Affecté par les objectifs des objectifs des propriétaires de plate-forme
- Enfin, bien que la non-réactivité soit avantageuse pour la recherche, le suivi du comportement des personnes sans leur consentement et leur sensibilisation soulève des préoccupations éthiques.

4. Capture les relations sociales

- Les données de traces numériques sont également quelque peu inhabituelles dans la mesure où elles décrivent souvent des relations sociales.

- Alors que les techniques d'enquête classiques ne mesurent généralement que les caractéristiques de sujets individuels, par exemple, les données de trace numériques peuvent souvent être utilisées pour **mesurer des relations sociales** telles que le réseau de politiciens européens illustré ci-dessous.

4. Capture les relations sociales



5. Incomplètes

- Bien que la taille et l'échelle des données de traces numériques pouvant être collectées soient souvent considérées comme un gage, les nouveaux arrivants sur le terrain sont souvent surpris par la quantité de données qui manque souvent ou est incomplète.
- **Incomplète naturellement**: La plupart des grandes sources de données sont incomplètes, dans le sens où elles ne disposent pas de l'information que vous voulez pour votre recherche. Elles manquent trois types d'information:
  - les informations démographiques sur les participants,
  - le comportement sur d'autres plateformes et
  - les données permettant d'opérationnaliser les concepts théoriques: comment mesurer par exemple l'intéligence à partir de ces données?
- **Incomplète par suppression**: prenons, par exemple, une étude sur les comportements d'intimidation sur les réseaux

6. Inaccessibles

- Un défi encore plus redoutable est que les données sont souvent inaccessibles. Bien que Twitter fournisse une quantité énorme de données accessibles au public, la grande majorité des données générées sur Facebook sont privées.
- Bien que certaines pages Facebook telles que les “pages de fans” aient des paramètres publics par défaut, la grande majorité des utilisateurs de Facebook définissent leurs paramètres de confidentialité par défaut de manière à ce que les utilisateurs ne puissent accéder à leurs données que s’ils sont affiliés les uns aux autres en tant qu’ “amis”.

7. Non-représentatives

- Ceux qui souhaitent utiliser des données de traces numériques doivent également faire face à un autre défi majeur: un échantillon aléatoire d'utilisateurs de Facebook ou de Twitter n'est pas représentatif de la population plus large des États-Unis ou de la plupart des autres pays.

- Les données du Wall Street Journal sur les données démographiques des utilisateurs de plusieurs sites de médias sociaux démontrent des différences significatives entre plates-formes en fonction de la race.

7. Non-représentatives

- Ceux qui souhaitent utiliser des données de traces numériques doivent également faire face à un autre défi majeur: un échantillon aléatoire d'utilisateurs de Facebook ou de Twitter n'est pas représentatif de la population plus large des États-Unis ou de la plupart des autres pays.

- Les données du Wall Street Journal sur les données démographiques des utilisateurs de plusieurs sites de médias sociaux démontrent des différences significatives entre plates-formes en fonction de la race.

- D'autre part, l'utilisation de Facebook est devenue tellement répandue que certains lecteurs pourraient être surpris de voir à quel point il est devenu plus représentatif du public américain au cours des dernières années.

7. Non-représentatives

- Faire la différence entre transportabilité (comparaison intra-échantillon)
  - Étude de John Snow sur l'épidémie de choléra de 1853-54 à Londres
  - Étude de Richard Doll et A. Bradford Hill (1954) qui ont suivi environ 25 000 hommes médecins pendant plusieurs années et ont comparé leurs taux de mortalité en fonction du nombre de cigarettes fumé. Ce résultat est transportable sur d'autres populations, pas à cause de la représentativité, mais du mécanisme liant tabac et cancer
- Et représentativité (généralisation – problème statistique)

8. Dérive (Drift)

- Selon certains analystes, MySpace était autrefois le plus grand site de média social au monde. C'est maintenant qu'il réside dans le cimetière de l'histoire d'Internet, comme tant d'autres sites. Cela augmente le risque de «dérive» dans les données de traces numériques.
- Les plateformes numériques ne changent pas seulement de popularité (ce qui a bien sûr des implications importantes pour leur représentativité), mais aussi selon qui les utilise et pourquoi.

8. Dérive (Drift)

- Selon certains analystes, MySpace était autrefois le plus grand site de média social au monde. C'est maintenant qu'il réside dans le cimetière de l'histoire d'Internet, comme tant d'autres sites. Cela augmente le risque de «dérive» dans les données de traces numériques.
- Les plateformes numériques ne changent pas seulement de popularité (ce qui a bien sûr des implications importantes pour leur représentativité), mais aussi selon qui les utilise et pourquoi.
- Bien que Facebook fût autrefois la plateforme la plus populaire pour les étudiants américains de premier cycle, beaucoup d'entre eux se sont tournés vers Instagram ou Snapchat, peut-être en réaction à la hausse de l'utilisation de Facebook par la génération de leurs parents :)

9. Algorithmiquement confondant

- Parfois, les données de traces numériques qui semblent décrire le comportement humain reflètent en réalité des changements dans la façon dont les humains interagissent avec des algorithmes.
- Un exemple relativement simple de confusion algorithmique est le fait que sur Facebook, il y a un nombre anormalement élevé d'utilisateurs avec environ 20 amis, comme l'ont découvert Johan Ugander et ses collègues (2011)
- La «parabole de Google Flu» en est un exemple populaire. À l'origine, Google Flu était un outil populaire permettant aux utilisateurs d'estimer la prévalence de la grippe à l'aide des données de recherche Google.

9. Algorithmiquement confondant

- Parfois, les données de traces numériques qui semblent décrire le comportement humain reflètent en réalité des changements dans la façon dont les humains interagissent avec des algorithmes.
- Un exemple relativement simple de confusion algorithmique est le fait que sur Facebook, il y a un nombre anormalement élevé d'utilisateurs avec environ 20 amis, comme l'ont découvert Johan Ugander et ses collègues (2011)
- La «parabole de Google Flu» en est un exemple populaire. À l'origine, Google Flu était un outil populaire permettant aux utilisateurs d'estimer la prévalence de la grippe à l'aide des données de recherche Google.
- L'outil était si précis que certains ont suggéré de déplacer les enquêtes officielles des Centers for Disease Control (CDC).

9. Algorithmiquement confondant

- Au début de 2013, les estimations de Google étaient bien supérieures à celles de la CDC.
- Des chercheurs ont par la suite découvert que les liens hypertextes liés à la grippe sur laquelle les internautes avaient cliqué apparaissaient dans leur navigateur Web après avoir recherché des informations sur les symptômes du rhume.

9. Algorithmiquement confondant

- Au début de 2013, les estimations de Google étaient bien supérieures à celles de la CDC.
- Des chercheurs ont par la suite découvert que les liens hypertextes liés à la grippe sur laquelle les internautes avaient cliqué apparaissaient dans leur navigateur Web après avoir recherché des informations sur les symptômes du rhume.
- Contrairement à d'autres problèmes avec les traces numériques, la confusion algorithmique est largement invisible

9. Algorithmiquement confondant

"Plutôt que de penser aux grandes sources de données comme observant les gens dans un cadre naturel, une métaphore plus appropriée est d'observer les gens dans un casino. Les casinos sont des environnements hautement conçus pour induire certains comportements, et un chercheur ne s'attendrait jamais à ce que le comportement dans un casino fournisse une fenêtre ouverte sur le comportement humain. Bien sûr, vous pourriez apprendre quelque chose sur le comportement humain en étudiant les gens dans les casinos, mais si vous ignorez le fait que les données ont été créées dans un casino, vous pourriez tirer de mauvaises conclusions." (Salganik, 2017)

10. Non-structurées

- Les données de traces numériques sont également souvent très désordonnées.
- Les nouveaux venus sur le terrain pensent souvent que les données générées sous forme numériques sont bien structurées, faciles à rechercher et rapidement transposables dans différents formats: ce n'est généralement pas vrai.

10. Non-structurées

- Les données de traces numériques sont également souvent très désordonnées.
- Les nouveaux venus sur le terrain pensent souvent que les données générées sous forme numériques sont bien structurées, faciles à rechercher et rapidement transposables dans différents formats: ce n'est généralement pas vrai.
- Un article récent du New York Times indiquait que les scientifiques du traitement des données consacraient plus de 80% de leur temps à nettoyer les données!

11. Sensibles

- Les données de traces numériques sont également souvent très sensibles.
- Les récents événements impliquant Facebook et le cabinet de conseil politique Cambridge Analytica soulignent les dangers d'un accès illimité à de grandes quantités de données de traces numériques.

11. Sensibles

- Les données de traces numériques sont également souvent très sensibles.
- Les récents événements impliquant Facebook et le cabinet de conseil politique Cambridge Analytica soulignent les dangers d'un accès illimité à de grandes quantités de données de traces numériques.
- Un de ces incidents, illustré ci-dessous, impliquait des chercheurs européens qui avaient extrait des données du site de rencontres Internet OK Cupid, puis avaient rendu publiques leurs données en ligne.

12. Biais positif

- Les données de traces numériques ont souvent des dimensions performatives.
- De nombreuses personnes ne signalent pas d'informations négatives à leur sujet en ligne précisément parce qu'elles savent que leurs amis, leurs collègues ou d'autres personnes qu'elles ne connaissent pas peuvent les surveiller.
- Cela crée une autre forme commune de biais dans la recherche sur les médias sociaux.

Les stratégies de recherche

Les stratégies de recherche:

1. Compter les choses
2. Prédirer les choses et
3. Approximer les expériences

1. Compter les choses

"Souvent, les étudiants motivent leur recherche en disant: Je vais compter quelque chose que personne n'a jamais compté auparavant. Par exemple, un étudiant pourrait dire que beaucoup de gens ont étudié des migrants et beaucoup de gens ont étudié des jumeaux, mais personne n'a étudié les jumeaux migrants. D'après mon expérience, cette stratégie, que j'appelle la motivation par l'absence, ne mène généralement pas à de bonnes recherches. La motivation par l'absence est un peu comme dire qu'il y a un trou là-bas, et je vais travailler très dur pour le remplir. Mais pas tous les trous doivent être remplis." (Salganik, 2017)

1. Compter les choses

- Rechercher plutôt des questions de recherche qui sont :
  - Importantes: a un impact mesurable ou qu'elle alimente une décision importante des décideurs
  - ou Intéressantes: Étude de Henry Farber (2015) sur le comportement des chauffeurs de taxis de New York

1. Compter les choses

Intéressantes: Étude de Henry Farber (2015) sur le comportement des chauffeurs de taxis de New York

- Les modèles néoclassiques en économie prédisent que les chauffeurs de taxi travailleront davantage les jours où ils ont des salaires horaires plus élevés.
- Alternativement, les modèles de l'économie comportementale prédisent exactement le contraire.
- Données utilisées: chaque trajet en taxi effectué par les taxis de New York de 2009 à 2013
  - Informations sur chaque voyage: heure de départ, lieu de départ, heure de fin, lieu de fin, tarif et pourboire (si le pourboire a été payé avec une carte de crédit)

2. Prédire les choses

- Nowcasting: Étude de Ginsberg et ses collègues sur la grippe (2009)

3. Approximer les expériences

- Quel est l'effet d'un programme de formation professionnelle sur les salaires?
- Expérimentation naturelle: augmentation du salaire minimum
- Ajustement statistique

Les stratégies de recherche: conclusion

“Les données volumineuses sont créées et collectées par les entreprises et les gouvernements à des fins autres que la recherche. L’utilisation de ces données pour la recherche nécessite donc une réutilisation.”

Qualité (des données massives)

Qualité des données

- Les problèmes sur les données digitales tels que nous venons de le voir ne sont pas limités uniquement à ces données. Ils sont inhérents à toutes les données.
- Par exemple, une enquête très bien réfléchie et exécutée qui comporte beaucoup de données manquantes a les mêmes problèmes dont nous venons de parler.

Qualité des données

- Les problèmes sur les données digitales tels que nous venons de le voir ne sont pas limités uniquement à ces données. Ils sont inhérents à toutes les données.
- Par exemple, une enquête très bien réfléchie et exécutée qui comporte beaucoup de données manquantes a les mêmes problèmes dont nous venons de parler.
- Même si la collecte des données est moins coûteuse, le coût pour en faire des données de qualité est exorbitant.

Qualité des données

La qualité des données peut être caractérisée de différentes manières :

- **Précision**: quelle est la précision des valeurs d'attribut dans les données?

Qualité des données

La qualité des données peut être caractérisée de différentes manières :

- **Précision**: quelle est la précision des valeurs d'attribut dans les données?
- **Exhaustivité**: les données sont-elles complètes?

Qualité des données

La qualité des données peut être caractérisée de différentes manières :

- **Précision**: quelle est la précision des valeurs d'attribut dans les données?
- **Exhaustivité**: les données sont-elles complètes?
- **Cohérence**: Dans quelle mesure les valeurs sont-elles cohérentes dans et entre les bases de données?

Qualité des données

La qualité des données peut être caractérisée de différentes manières :

- **Précision**: quelle est la précision des valeurs d'attribut dans les données?
- **Exhaustivité**: les données sont-elles complètes?
- **Cohérence**: Dans quelle mesure les valeurs sont-elles cohérentes dans et entre les bases de données?
- **Actualisée**: dans quelle mesure les données sont-elles actualisées?

Qualité des données

La qualité des données peut être caractérisée de différentes manières :

- **Précision**: quelle est la précision des valeurs d'attribut dans les données?
- **Exhaustivité**: les données sont-elles complètes?
- **Cohérence**: Dans quelle mesure les valeurs sont-elles cohérentes dans et entre les bases de données?
- **Actualisée**: dans quelle mesure les données sont-elles actualisées?
- **Accessibilité**: toutes les variables sont-elles disponibles pour l'analyse?

Qualité des données

- Pour avoir des données de qualité, plusieurs étapes sont importantes:

1. Analyse (parsing)
2. Standardisation
3. Dé-duplication
4. Normalisation

1. Analyse

- Processus de décomposition d'une variable complexe en ces éléments constitutifs
- Exemple: La variable adresse "1245 Jean-Talon Est" peut être décomposée à:
  - Numéro
  - Nom
  - Direction

1. Analyse

Les étapes typiques d'une procédure d'analyse comprennent:

- Fractionnement des champs en jetons (mots) sur la base de délimiteurs,
- Standardisation des jetons par tables de correspondance et substitution par un formulaire standard,

1. Analyse

Les étapes typiques d'une procédure d'analyse comprennent:

- Fractionnement des champs en jetons (mots) sur la base de délimiteurs,
- Standardisation des jetons par tables de correspondance et substitution par un formulaire standard,
- Catégorisation des jetons,

1. Analyse

Les étapes typiques d'une procédure d'analyse comprennent:

- Fractionnement des champs en jetons (mots) sur la base de délimiteurs,

- Standardisation des jetons par tables de correspondance et substitution par un formulaire standard,
- Catégorisation des jetons,
- Identification d'un motif d'ancres, de jetons et de délimiteurs,

1. Analyse

Les étapes typiques d'une procédure d'analyse comprennent:

- Fractionnement des champs en jetons (mots) sur la base de délimiteurs,

- Standardisation des jetons par tables de correspondance et substitution par un formulaire standard,
- Catégorisation des jetons,
- Identification d'un motif d'ancres, de jetons et de délimiteurs,
- Appel de sous-programmes selon le modèle identifié, mappage des jetons vers les composants prédéfinis

2. Standardisation

- La standardisation fait référence au processus de simplification des données en remplaçant les variantes de représentation de la même observation sous-jacente par une valeur par défaut afin d'améliorer la précision des comparaisons de terrain.
- Exemple: av. et avenue désigne la même manière décrire les adresses.

2. Standardisation

- La standardisation fait référence au processus de simplification des données en remplaçant les variantes de représentation de la même observation sous-jacente par une valeur par défaut afin d'améliorer la précision des comparaisons de terrain.
- Exemple: av. et avenue désigne la même manière décrire les adresses.
- Si vous cherchez tous les avenues dans votre base de données, vous allez manquer celles qui sont écrites avec **av**.

2. Standardisation

- La standardisation fait référence au processus de simplification des données en remplaçant les variantes de représentation de la même observation sous-jacente par une valeur par défaut afin d'améliorer la précision des comparaisons de terrain.
- Exemple: av. et avenue désigne la même manière décrire les adresses.
- Si vous cherchez tous les avenues dans votre base de données, vous allez manquer celles qui sont écrites avec **av**.
- La standardisation permet de résoudre ce problème.

2. Standardisation

Des exemples communs de standardisations sont:

- Standardisation des différentes orthographes de mots fréquemment rencontrés: par exemple, remplacer les abréviations courantes dans les noms de rue (Ave, St, etc.) ou les titres (Ms, Dr, etc.) par une forme commune. Ces types de règles sont très spécifiques aux pays et aux langues.
- Standardisation générale, y compris la conversion des champs de caractères en majuscules et la suppression de la ponctuation et des chiffres

3. De-duplication

- La de-duplication consiste à supprimer les enregistrements redondants d'une seule liste, c'est-à-dire plusieurs enregistrements de la même liste qui font référence à la même entité sous-jacente.

4. Normalisation

La normalisation consiste à garantir que les champs comparés entre les fichiers sont aussi similaires que possible dans le sens où ils auraient pu être générés par le même processus.

- Par exemple, considérez un champ de salaire dans une enquête. Le salaire peut être enregistré de différentes manières: il peut être tronqué comme mesure de protection de la vie privée ou arrondi au millier le plus proche, et les valeurs manquantes peuvent être imputées avec la moyenne ou avec zéro. Lors de la normalisation, nous notons exactement comment les champs sont enregistrés.

Ressources

1. Données

- https://labs.jstor.org/projects/text-mining/
- https://iris.isr.umich.edu/

2. Méthodes
3. Sites web

- github.com/BigDataSocialScience