COLLECTE DES DONNÉES

« Les gens résistent à un recensement, mais présentez-leur une page de profil et ils passeront la journée à vous raconter qui ils sont. »

Max Berry, Lexicon

APERÇU

1. Caractéristiques des données à recueillir : Théorie de l'échantillonnage et plan d'étude
2. Collecte de données moderne : Interfaces de programmation d'applications (API) et moissonnage du Web

THÉORIE DE L'ÉCHANTILLONNAGE ET PLAN D'ÉTUDE

COLLECTE DES DONNÉES

« La dernière enquête indique que trois personnes sur quatre représentent 75 % de la population »

D. Letterman

L'OBJECTIF D'UN PLAN D'ÉTUDE ET D'ÉCHANTILLONNAGE EFFICACE

Nous recherchons des données de nature à :

- donner un aperçu légitime de notre système d'intérêt
- fournir des réponses correctes et précises aux questions pertinentes
- soutenir la formulation de conclusions légitimes et valides, en permettant de nuancer ces conclusions en matière de portée et de précision

Un tel processus commence par le **plan d'étude** – quelles données recueillir et comment les recueillir.

[G. Smith, The Exaggerated Promise of So-Called Unbiased Data Mining]

« À l'aide d'un appareil d'imagerie par résonance magnétique (IRM), un diplômé de Dartmouth a étudié l'activité cérébrale d'un saumon lorsqu'on lui montrait des photographies et qu'on lui posait des questions. L'aspect le plus intéressant de cette recherche, ce n'est pas qu'on ait étudié un saumon, mais que ce saumon était mort. Hé oui! On a acheté un saumon mort au marché local, on l'a placé dans un appareil d'IRM, et on a observé certains schémas. Il y avait inévitablement des schémas, mais ils étaient invariablement dénués de sens. »

ÉCHANTILLONNAGE NON PROBABILISTE ET « PÊCHE » AUX TENDANCES

Deux situations distinctes peuvent s'associer pour causer dans **problèmes** d'analyse des données :

- la formulation de conclusions (inférences) à partir d'un échantillon de population qui ne se justifient pas par la méthode de collecte de l'échantillon (symptomatique d'un échantillonnage non probabiliste)
- la recherche d'un quelconque schéma dans les données, puis la formulation d'explications a posteriori concernant ces schémas

Seules ou combinées, ces deux situations conduisent à des conclusions médiocres (et **potentiellement nuisibles**).

ÉTUDES ET ENQUÊTES

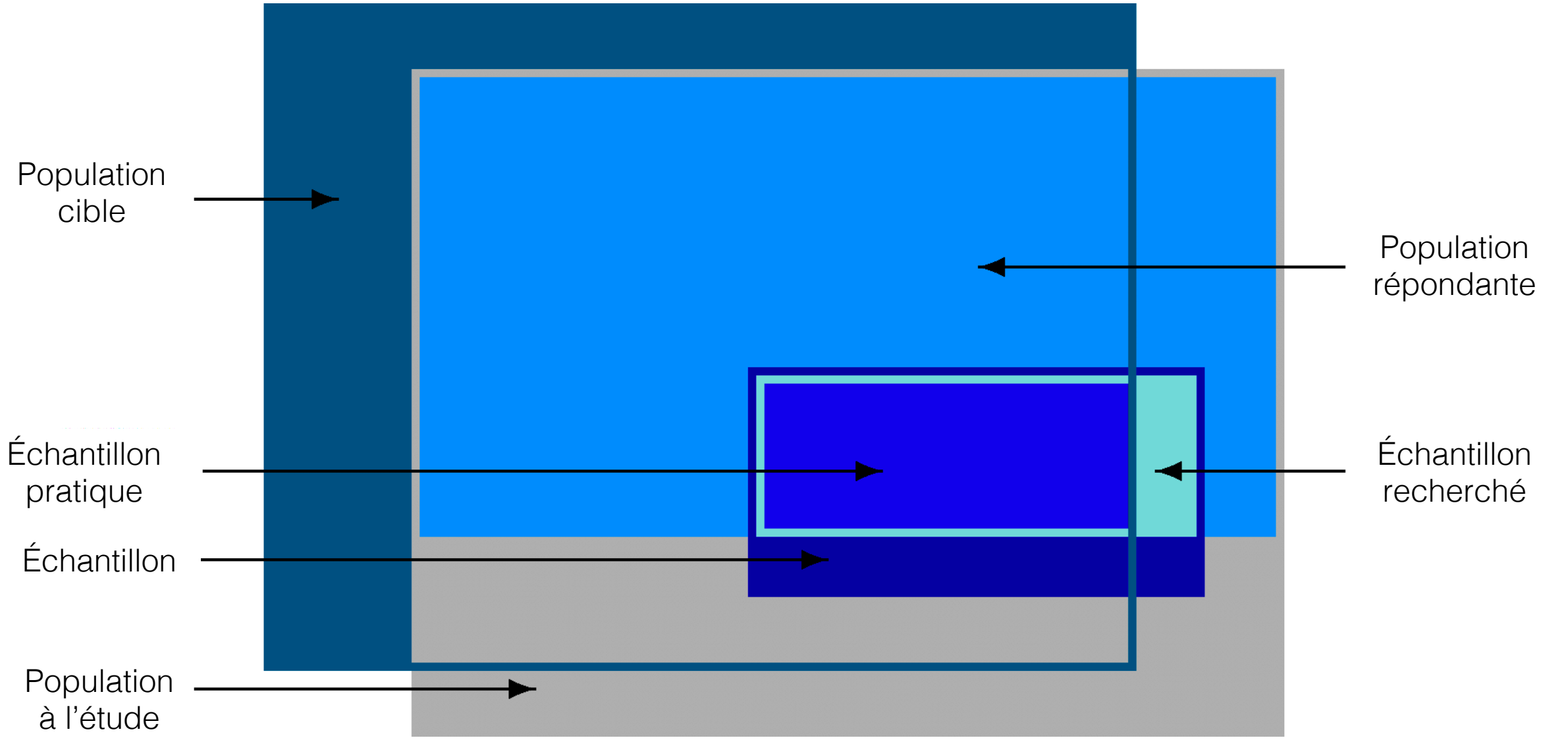
Une **enquête** est une activité qui consiste à recueillir de l'information sur des caractéristiques d'intérêt :

- de manière **organisée** et **méthodique**;
- sur une partie ou la totalité des **unités** d'une population;
- à l'aide de concepts, de méthodes et de procédures **bien définis**;
- grâce à la compilation de renseignements sous forme d'un résumé **significatif**.

MODÈLES D'ÉCHANTILLONNAGE

Un **recensement** est une collecte de données sur toutes les unités d'une population, alors qu'une **enquête sur échantillon** n'utilise qu'une fraction des unités.

Lorsque l'échantillonnage de l'enquête est effectué correctement, il est possible de recourir à diverses **méthodes statistiques** pour faire des **inférences** sur la **population cible** en échantillonnant un (comparativement) petit nombre d'unités dans la **population étudiée**.


Population cible
Population répondante
Échantillon pratique
Échantillon recherché
Échantillon
Population à l'étude

BASES D'ENQUÊTE

La base idéale contient les données d'identification, les données de contact, les données de classification, les données de maintenance et les données de couplage, et doit réduire au minimum le risque de **sous-dénombrement** ou de **surdénombrement**, ainsi que le nombre de dédoublements et d'erreurs de classification (même si certains problèmes éventuels peuvent être réglés à l'étape du traitement des données).

Une approche d'échantillonnage statistique est contre-indiquée à moins que la base d'enquête choisie ne soit :

- **pertinente** (autrement dit qu'elle corresponde et permette l'accessibilité à la population cible);
- **exacte** (l'information qu'elle contient est valide);
- **opportune** (elle est à jour);
- **offerte à un prix compétitif**.

ERREUR D'ENQUÊTE

Erreur totale = erreur d'échantillonnage + erreur de mesure + erreur de non-réponse + erreur de couverture

- erreur d'échantillonnage : enquête, pas recensement
- erreur de mesure : manque d'exactitude dans la mesure des observations
- erreur de non-réponse : non-répondants présentant des différences d'observation systématiques
- erreur de couverture : dégradation ou corruption de la base

L'échantillonnage statistique permet de fournir des estimations, mais, surtout, il permet aussi de contrôler dans une certaine mesure l'**erreur totale** (ET) dans les estimations.

Idéalement, ET= 0. Dans la pratique, deux principaux éléments contribuent à l'ET : les **erreurs d'échantillonnage** (attribuable au choix du plan d'échantillonnage) et les **erreurs non attribuables à l'échantillonnage** (tout le reste).

ERREUR NON ATTRIBUABLE À L'ÉCHANTILLONNAGE

Dans une certaine mesure, il est possible de contrôler une erreur non attribuable à l'échantillonnage :

- l'**erreur de couverture** peut être réduite au minimum en choisissant des bases d'enquête à jour et de grande qualité;
- l'**erreur de non-réponse** peut être atténuée en choisissant soigneusement le mode de collecte des données et le plan du questionnaire, et au moyen de « rappels » et de « suivis »;
- l'**erreur de mesure** peut être grandement diminuée par une conception minutieuse du questionnaire, un essai préliminaire de l'appareil de mesure et une validation croisée des réponses.

Dans les faits, ces suggestions ne s'avèrent pas d'une grande utilité à l'heure actuelle (les bases d'enquête fondées sur la téléphonie filaire perdent de leur pertinence en raison de la démographie, les taux de réponse aux enquêtes non obligatoires en vertu de la loi sont faibles, etc.). Cela explique, en partie, la trop large utilisation du **moissonnage du Web** et de l'**échantillonnage non probabiliste**.

ÉCHANTILLONNAGE NON PROBABILISTE

Les méthodes d'**échantillonnage non probabiliste** sélectionnent les unités d'échantillonnage de la population cible à l'aide d'approches subjectives et non aléatoires.

- L'échantillonnage non probabiliste a le mérite d'être rapide, relativement peu coûteux et pratique (aucune base d'enquête requise).
- Les méthodes d'échantillonnage non probabiliste sont idéales pour l'analyse exploratoire et l'élaboration des enquêtes.

Malheureusement, on a souvent recours aux échantillonnages non probabilistes au lieu des échantillonnages probabilistes (ce qui est problématique).

- Le biais de sélection qui y est associé rend les méthodes d'échantillonnage non probabiliste peu sûres lorsqu'il s'agit d'inférences (elles ne peuvent être utilisées pour fournir des estimations fiables de l'erreur d'échantillonnage, la seule composante de l'ET sur laquelle l'analyste a une emprise directe).
- La collecte automatisée des données tombe souvent dans le champ des échantillonnages non probabilistes – il est toujours possible d'analyser les données recueillies selon cette méthode, mais pas de généraliser les résultats à la population cible.

ÉCHANTILLONNAGE PROBABILISTE

Les plans d'échantillonnage probabiliste sont généralement plus **difficiles** et plus **coûteux** à mettre en place (car ils requièrent une base d'enquête de qualité), et ils prennent plus de temps à réaliser.

Ils fournissent des **estimations fiables** de la caractéristique d'intérêt et de l'**erreur d'échantillonnage**, ouvrant la voie à l'utilisation de petits échantillons pour tirer des inférences sur des populations cibles plus vastes (en théorie, du moins, les composantes de l'erreur non attribuable à l'échantillonnage peuvent tout de même jouer sur les résultats et la généralisation).

INTERVALLES DE CONFIANCE

Si l'estimation $\hat{\beta}$ est non biaisée, $\mathrm{E}\left(\hat{\beta} - \beta\right) = 0$, **un intervalle de confiance au seuil d'environ 95 %** (IC à 95 %) pour β est alors donné approximativement par

$$\hat{\beta} \pm 2\sqrt{\widehat{\mathrm{V}}\left(\hat{\beta}\right)},$$

où $\widehat{\mathrm{V}}\left(\hat{\beta}\right)$ est une estimation **propre au plan d'échantillonnage** de $\mathrm{V}\left(\hat{\beta}\right)$.

Mais à quoi correspond exactement un IC à 95 %?

PLAN D'ÉCHANTILLONNAGE

Les différents **plans d'échantillonnage** présentent des avantages et des inconvénients distincts.

Ils peuvent servir à calculer des estimations

- pour diverses caractéristiques de la population : moyenne, total, proportion, ratio, différence, etc.
- pour l'IC à 95 % correspondant.

On pourrait aussi vouloir calculer la taille des échantillons pour une **limite d'erreur** donnée (une limite supérieure à l'intérieur de l'IC à 95 % désiré), et déterminer la **répartition de l'échantillon** (combien d'unités à échantillonner dans divers groupes de sous-population).

PLAN D'ÉCHANTILLONNAGE – L'UNIVERS DU DISCOURS

Population cible :

- N unités et mesures $\mathcal{U} = \{u_1, \dots, u_N\}$

Caractéristiques réelles de la population :

- moyenne μ, variance σ^2, total τ, proportion p

Échantillon de population :

- n unités et mesures $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Caractéristiques de l'échantillon de population :

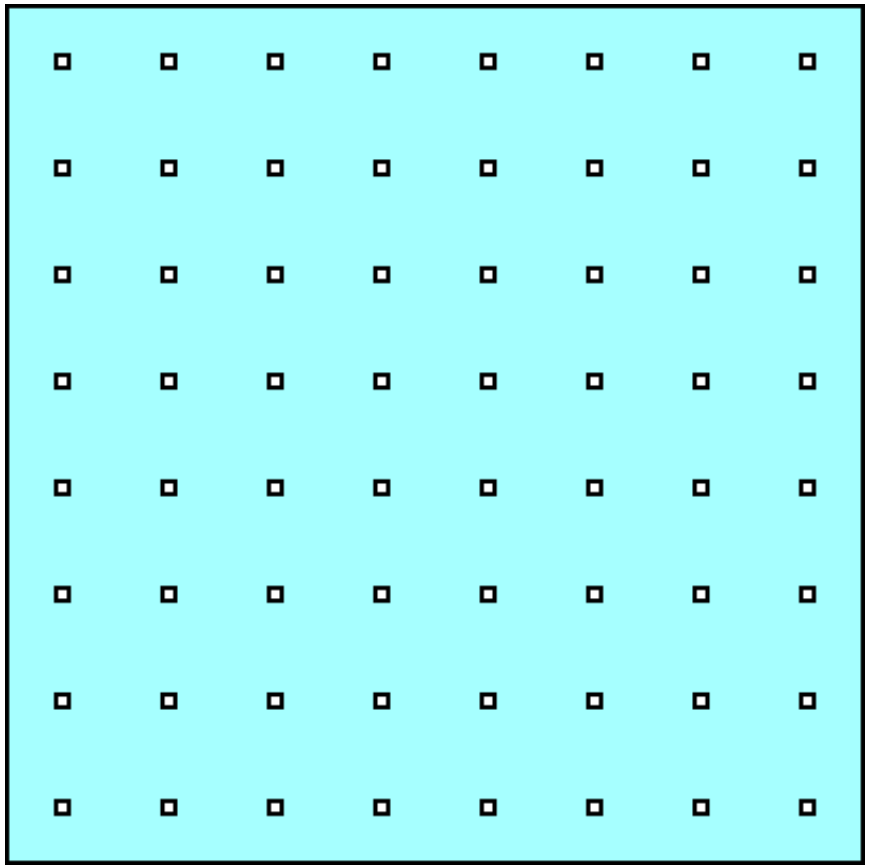
- moyenne de l'échantillon $\bar{y}$, variance de l'échantillon s^2, total de l'échantillon $\hat{\tau}$, proportion de l'échantillon $\hat{p}$

PLAN D'ÉCHANTILLONNAGE – L'UNIVERS DU DISCOURS

Objectif : faire l'estimation des véritables caractéristiques de la population μ, σ^2, τ, p grâce aux caractéristiques de l'échantillon de population $\bar{y}$, s^2, $\hat{\tau}$, $\hat{p}$, n, et à la taille N de la population cible.

Pour une caractéristique donnée, on définit δ_i comme prenant la valeur 1 ou 0 selon que l'unité échantillon y_i possède ou non la caractéristique en question.

On utilise la limite d'erreur $B = 2\sqrt{\hat{\mathrm{V}}}$.



ÉCHANTILLONNAGE ALÉATOIRE SIMPLE (EAS)

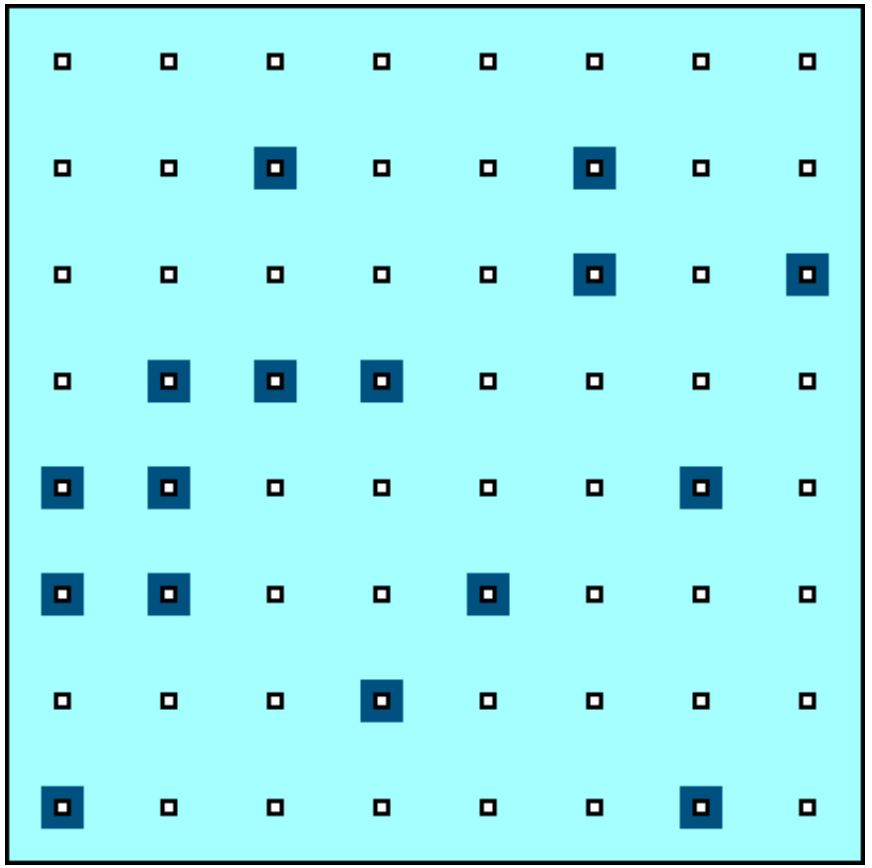
Dans l'EAS, n unités sont sélectionnées au hasard dans la base.

Avantages :

- Plan d'échantillonnage le plus facile à mettre en place
- Erreurs d'échantillonnage bien connues et faciles à estimer
- Pas nécessaire d'avoir de données auxiliaires

Inconvénients :

- Ne fait aucunement appel aux données auxiliaires
- Ne fournit aucune garantie quant à la représentativité de l'échantillon
- Coûteux si l'échantillon est largement réparti géographiquement



ESTIMATEURS DE L'EAS

Estimateurs :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \qquad \hat{\tau} = N\bar{y}, \qquad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

Estimations des variances propres au plan d'échantillonnage :

$$\widehat{\mathrm{V}}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right), \qquad \widehat{\mathrm{V}}(\hat{\tau}) = N^2\widehat{\mathrm{V}}(\bar{y}), \qquad \widehat{\mathrm{V}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}\left(1 - \frac{n}{N}\right)$$

Répartition de l'échantillon :

$$n_{\bar{y}} = \frac{4N\tilde{\sigma}^2}{(N-1)B^2+4\tilde{\sigma}^2}, \qquad n_{\hat{\tau}} = \frac{4N^3\tilde{\sigma}^2}{(N-1)B^2+4N^2\tilde{\sigma}^2}, \qquad n_{\hat{p}} = \frac{4\tilde{p}\,(1-\tilde{p})}{(N-1)B^2+4\tilde{p}\,(1-\tilde{p})}$$

ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

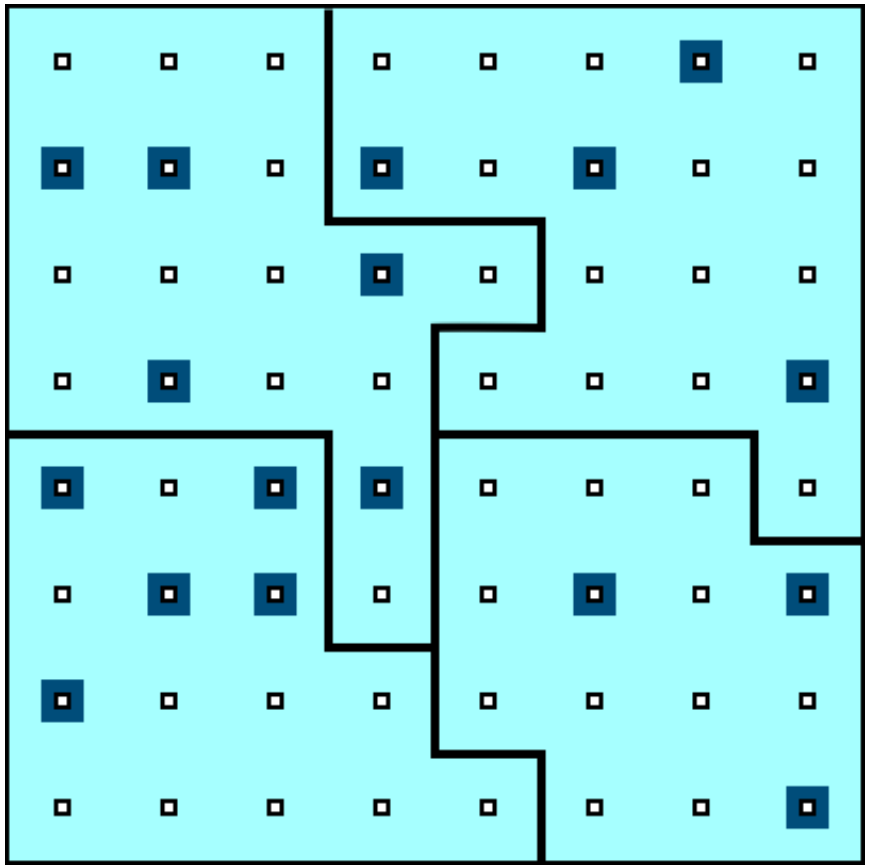
Dans l'échantillonnage aléatoire stratifié, $n = n_1 + \cdots + n_k$ unités sont sélectionnées de manière aléatoire à partir de k **strates** de la base.

Avantages :

- Peut produire une limite d'erreur inférieure sur l'estimation, en comparaison de l'EAS;
- Peut être moins coûteux à condition de stratifier adéquatement les éléments;
- Peut fournir des estimations pour des sous-populations.

Inconvénients :

- Aucun inconvénient majeur
- S'il n'existe aucun moyen naturel de stratifier la base d'enquête en groupes homogènes, l'échantillonnage aléatoire stratifié devient à peu près équivalent à l'EAS



ESTIMATEURS DE L'ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

Estimateurs :

$$\bar{y}_{st} = \sum_{j=1}^{k} \frac{N_j}{N} \bar{y}_j \,, \qquad \hat{\tau}_{st} = N\bar{y}_{st}, \qquad \hat{p}_{st} = \sum_{j=1}^{k} \frac{N_j}{N} \hat{p}_j$$

Estimations des variances propres au plan d'échantillonnage :

$$\widehat{\mathrm{V}}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 \widehat{\mathrm{V}}(\bar{y}_j) \,, \qquad \widehat{\mathrm{V}}(\hat{\tau}_{st}) = N^2 \widehat{\mathrm{V}}(\bar{y}_{st}), \qquad \widehat{\mathrm{V}}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 \widehat{\mathrm{V}}(\hat{p}_j)$$

EXERCICES

On vous demande de faire une estimation du salaire annuel des scientifiques des données au Canada.

Cernez les éléments possibles suivants :

- populations (cible, étude, répondant, bases d'enquête);
- échantillons (prévus, obtenus);
- informations sur l'unité (unité, variable de réponse, caractéristique de population);
- sources de biais (couverture, non-réponse, échantillonnage, mesure) et de variabilité (échantillonnage, mesure).

EXERCICES

Le fichier `cities.txt` contient des informations sur la population urbaine d'un pays. On définit qu'une ville est « petite » si sa population est inférieure à 75 000 habitants, « moyenne » si elle se situe entre 75 000 et 1 million d'habitants, et « grande » au-delà.

1. Trouvez le fichier et téléchargez-le dans l'espace de travail de votre choix. Combien y a-t-il de villes? Combien dans chaque groupe?
2. Affichez les statistiques démographiques sommaires des villes, à la fois globalement et par groupe.
3. Calculez un IC à 95 % pour la moyenne de la population en 1999 en utilisant un EAS de taille $n = 10$.
4. Calculez un IC à 95 % pour la moyenne de la population en 1999 en utilisant un échantillonnage aléatoire stratifié de taille $(n_s, n_m, n_l) = (5,3,2)$.
5. Comparez les estimations avec la valeur réelle. Les résultats sont-ils étonnants? Sinon, auraient-ils pu l'être?

Matériel supplémentaire

FACTEURS DÉCISIFS

Dans certains cas, il faut disposer de l'information sur l'**ensemble** de la population pour répondre aux questions, alors que dans d'autres, ce n'est pas nécessaire. Le **type d'enquête** dépend de multiples facteurs, notamment :

- le type de questions auxquelles il faut répondre
- la précision requise
- le coût du sondage d'une unité
- le temps requis pour sonder une unité
- la taille de la population faisant l'objet de l'enquête
- la prévalence des caractéristiques d'intérêt

ÉTAPES DE L'ÉTUDE OU DE L'ENQUÊTE

Les études ou enquêtes suivent les mêmes étapes générales :

1. énoncé de l'objectif
2. sélection de la base d'enquête
3. plan d'échantillonnage
4. plan du questionnaire
5. collecte des données
6. saisie et codage des données
7. traitement des données et imputation
8. estimation
9. analyse des données
10. diffusion
11. documentation

Le processus n'est pas toujours linéaire, mais il existe un cheminement clair depuis l'objectif jusqu'à la diffusion.

BASES D'ENQUÊTE

La **base** d'enquête permet de **sélectionner** et de **contacter** les unités de population visées par l'enquête. Sa création et sa maintenance sont en général coûteux (en fait, il existe des organisations et des entreprises spécialisées dans la constitution ou la vente de telles bases).

Les bases utiles contiennent :

- les données d'identification
- les données de contact
- les données de classification
- les données de maintenance
- les données de couplage

MODES DE COLLECTE DES DONNÉES

Sur support papier ou assisté par ordinateur

- Les **questionnaires auto-administrés** sont utilisés lorsque l'enquête nécessite des renseignements détaillés pour permettre aux unités de consulter les dossiers personnels; associés à un taux de non-réponse élevé.
- Les **questionnaires assistés par un intervieweur** adéquatement formé sont utilisés pour augmenter le taux de réponse et la qualité globale des données; en personne ou au téléphone.
- Les **entrevues assistées par ordinateur** associent la collecte et la saisie des données, ce qui fait gagner du temps.
- Observation directe discrète.
- Journaux à remplir (format papier ou électronique).
- Enquêtes omnibus.
- Courriel, Internet et médias sociaux.

MÉTHODES D'ÉCHANTILLONNAGE NON PROBABILISTE

Au hasard

- Un passant; dépend de la disponibilité des unités et du biais lié à l'intervieweur.

Volontaire

- Biais d'autosélection.

A priori

- Biaisé par des idées préconçues inexactes concernant la population cible.

Par quotas

- Sondage fait à la sortie de l'isoloir, ignore le biais de non-réponse.

MÉTHODES D'ÉCHANTILLONNAGE NON PROBABILISTE

Modifié

- D'abord probabiliste, puis par quotas en réaction à des taux de non-réponse élevés

En boule de neige

- Plan « pyramidal »

Dans certains contextes, les méthodes d'échantillonnage non probabiliste pourraient répondre aux besoins d'un client ou d'une organisation (et c'est à eux qu'il appartient de prendre la décision en dernier lieu), mais on doit l'informer des inconvénients et lui proposer des solutions probabilistes.

CONCEPTS MATHÉMATIQUES DE BASE

Posons une population finie $\mathcal{U}$, avec N unités et mesures $\{u_1, \dots, u_N\}$.

La **moyenne** et la **variance** de la population pour la variable d'intérêt sont données par

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j\,, \qquad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(u_j - \mu\right)^2.$$

Si $\mathcal{Y} \subseteq \mathcal{U}$ représente un **échantillon** de la population avec n unités et mesures $\{y_1, \dots, y_n\}$, alors la **moyenne** et la **variance de l'échantillon** sont données par

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

CONCEPTS MATHÉMATIQUES DE BASE

Soit $X_1, \ldots, X_n$ des **variables aléatoires**, $b_1, \ldots, b_n \in \mathbb{R}$, et E, V, Cov l'**espérance**, la **variance** et la **covariance**, respectivement, c.-à-d. :

- $\mathrm{E}(X_i) = \mu_i$
- $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$
- $\mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}(X_i^2) - \mathrm{E}^2(X_i) = \mathrm{E}(X_i^2) - \mu_i^2 = \sigma_i^2$ and

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \mathrm{E}(X_i) = \sum_{i=1}^{n} b_i \mu_i$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{V}(X_i) + \sum_{i \neq j} b_i b_j \,\mathrm{Cov}(X_i, X_j)$$

CONCEPTS MATHÉMATIQUES DE BASE

Le **biais** d'une composante d'erreur est la moyenne de cette composante d'erreur si l'enquête est répétée plusieurs fois indépendamment et dans les mêmes conditions. Une estimation **sans biais** est une estimation pour laquelle le biais est nul.

La **variabilité** d'une composante d'erreur est la mesure dans laquelle cette composante varierait par rapport à sa valeur moyenne dans le scénario idéal décrit ci-dessus.

L'**erreur quadratique moyenne** d'une composante d'erreur est une mesure de sa taille :

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta}),$$

Où $\hat{\beta}$ est un estimateur de β.

PLANS D'ÉCHANTILLONNAGE PROBABILISTE

Échantillonnage aléatoire simple (EAS)

Échantillonnage aléatoire stratifié

Échantillonnage systématique

Échantillonnage en grappes

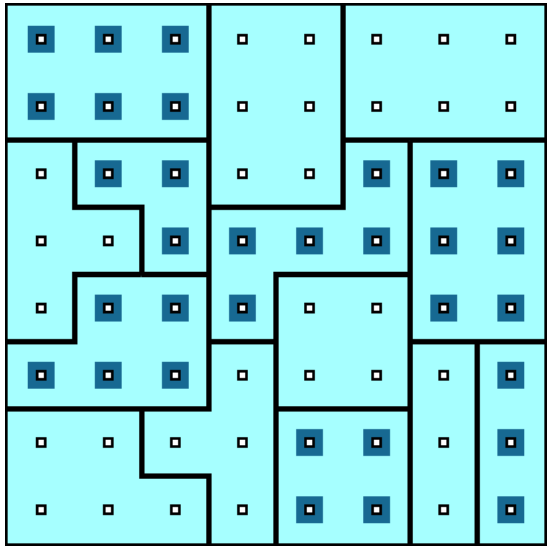
Échantillonnage avec probabilité proportionnelle à la taille (PPT)

Échantillonnage répété

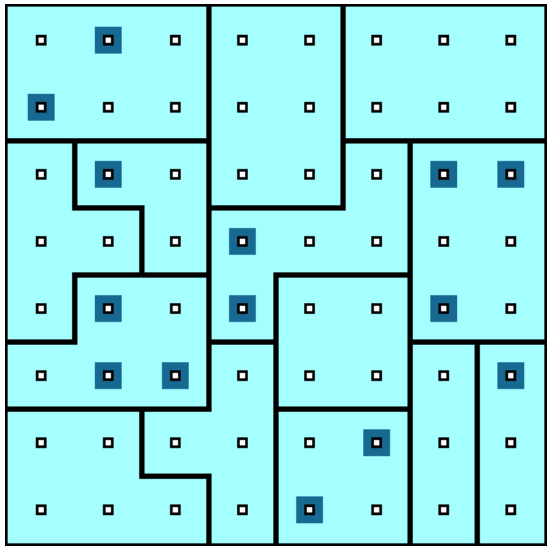
Échantillonnage à plusieurs degrés

Échantillonnage à plusieurs phases

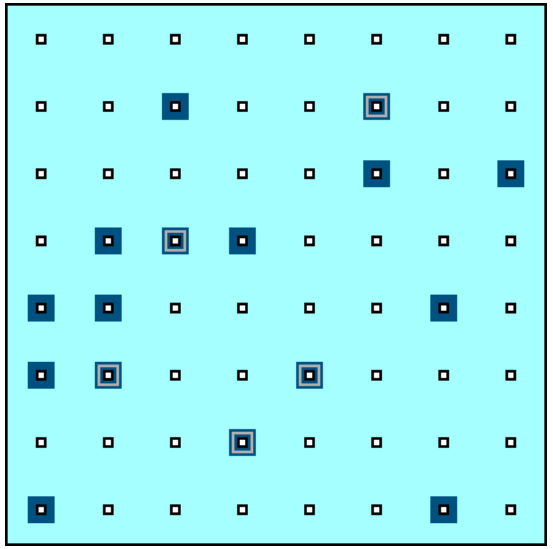
AUTRES EXEMPLES DE PLANS D'ÉCHANTILLONNAGE



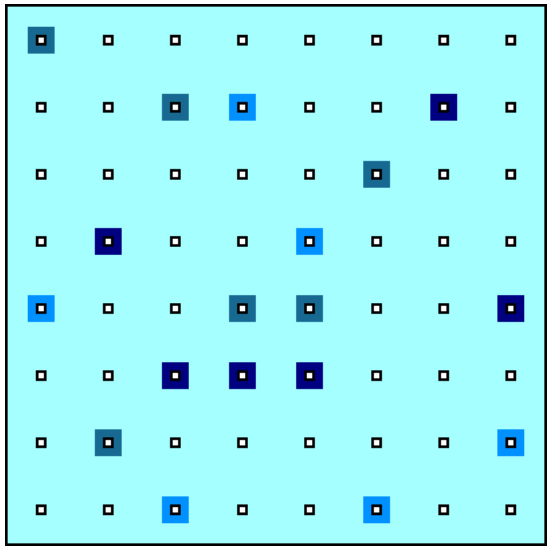
Échantillonnage en grappes



Échantillonnage à plusieurs degrés



Échantillonnage à plusieurs phases



Échantillonnage répété

API ET MOISSONNAGE DU WEB

COLLECTE DES DONNÉES

« Les rues du Web sont pavées de données qui n'attendent que d'être recueillies. »

Munzart, Rubba, Meissner, Nyhuis, Automated Data Collection with R

WORLD WIDE WEB

Il fut un temps, assez récent, où tant la rareté des données que leur inaccessibilité constituaient un problème pour les chercheurs et les décideurs. Tel n'est **manifestement** plus le cas désormais.

L'abondance des données présente son propre lot de problèmes particuliers :

- des masses de données enchevêtrées
- les méthodes classiques de collecte des données et les techniques usuelles d'analyse des données (en petites quantités) peuvent ne plus suffire aujourd'hui

EXEMPLE DE MOISSONNAGE DU WEB – NOUVEAU TÉLÉPHONE

Supposons que vous aimeriez savoir ce que la population pense d'un nouveau téléphone. Approche standard : étude de marché (p. ex. sondage téléphonique, système de récompenses, etc.).

Pièges :

- échantillon non représentatif : il se pourrait que l'échantillon sélectionné ne représente pas la population visée
- non-réponse systématique : les personnes qui n'aiment pas les sondages téléphoniques pourraient être moins (ou plus) portées à ne pas aimer le nouveau téléphone
- erreur de couverture : à titre d'exemple, il serait impossible de joindre les personnes qui ne disposeraient pas d'un téléphone filaire
- erreur de mesure : les questions du sondage fournissent-elles des renseignements convenant au problème posé?

QUALITÉ DES DONNÉES DU WEB – NOUVEAU TÉLÉPHONE

Ces solutions peuvent être **onéreuses**, **chronophages**, **inefficaces**.

Variables de substitution – indicateurs qui sont étroitement reliés à la popularité du produit, sans mesurer directement celle-ci pour autant.

Si la notion de **popularité** renvoie au fait que de grands groupes de gens préfèrent un produit par rapport à un produit concurrent, les statistiques de vente que l'on retrouve sur un site Web commercial pourraient constituer un substitut de la popularité.

Les classements sur Amazon pourraient offrir une idée plus **complète** du marché des téléphones par rapport à ce que permettrait d'obtenir un sondage classique.

PROBLÈMES POTENTIELS – NOUVEAU TÉLÉPHONE

Représentativité des **produits répertoriés**

- Tous les téléphones sont-ils répertoriés?
- Si tel n'est pas le cas, est-ce parce que le site Web ne les vend pas?
- Y a-t-il une autre raison?

Représentativité des **clients**

- Certains groupes spécifiques achètent-ils/n'achètent-ils pas de produits en ligne?
- Certains groupes spécifiques achètent-ils sur des sites spécifiques?
- Certains groupes spécifiques laissent-ils ou non des commentaires?

Honnêteté des clients et **crédibilité** des commentaires.

LE MOISSONNAGE DU WEB EST-IL LÉGAL?

Qu'est-ce qu'une araignée?

- Il s'agit d'un programme qui parcourt ou arpente le Web pour en extraire de l'information rapidement
- L'araignée, ou programme collecteur, saute d'une page à l'autre, en en extrayant l'intégralité du contenu

Le **moissonnage** consiste à extraire de l'information spécifique de sites Web spécifiques (c'est le but) : en quoi ces méthodes sont-elles **différentes**?

« Comme, fondamentalement, le moissonnage consiste à **copier** de l'information, l'une des revendications les plus évidentes à l'encontre des dispositifs de récupération de données tient à la violation du droit d'auteur. »

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

eBay c. Bidder's Edge (BE)

- BE a eu recours à des programmes automatisés pour extraire de l'information de différents sites de vente aux enchères.
- Les utilisateurs pouvaient consulter les listes sur la page Web de BE, plutôt que d'avoir à se rendre sur les différents sites de vente aux enchères.
- En 1999, BE a accédé aux sites d'eBay environ 100 000 fois par jour (1,53 % du nombre de requêtes, 1,1 % de l'ensemble des données transférées par eBay).
- eBay a réclamé des dommages-intérêts allant de 45 000 $ et 62 000 $, sur une période de 10 mois.
- BE n'a volé aucune information qui n'était pas déjà publique, mais l'augmentation du trafic a imposé une charge additionnelle aux serveurs d'eBay.
- **Votre verdict?**

COOPÉRATION AMICALE AVEC LES API

Qu'est-ce qu'une API? L'acronyme « API » signifie *application program interface*, ou interface de programmation d'applications, soit un ensemble de routines, de protocoles et d'outils pour la construction d'applications logicielles.

Plusieurs API restreignent l'utilisateur à un certain nombre d'appels d'API par jour (ou à d'autres formes de limites).

Il importe de respecter ces limites.

Matériel supplémentaire

POURQUOI PROCÈDE-T-ON À LA COLLECTE AUTOMATISÉE DES DONNÉES?

En ce qui concerne les données scientifiques sociales :

- caractère limité des ressources financières
- peu de temps ou de désir de recueillir les données manuellement
- désir de travailler avec des sources riches en données à jour et de grande qualité
- documenter le processus du début (collecte des données) à la fin (publication) de sorte qu'il puisse être reproduit

Problèmes que pose la collecte manuelle :

- processus non reproductible
- présente des risques d'erreur en plus d'être lourd
- présente un risque plus élevé de « mourir d'ennui »

POURQUOI PROCÉDER À LA COLLECTE AUTOMATISÉE DES DONNÉES?

Avantages des solutions fondées sur un programme :

- fiabilité
- reproductibilité
- rapidité
- groupe d'ensembles de données de meilleure qualité

LISTE DE VÉRIFICATION APPLICABLE À LA COLLECTE AUTOMATISÉE

Le **moissonnage du Web** ou **le traitement de texte statistique** (collecte automatisée ou semi-automatisée des données) est-il absolument nécessaire?

Critères :

- Prévoyez-vous répéter l'opération de temps à autre, p. ex. pour mettre à jour votre base de données?
- Désirez-vous que d'autres puissent reproduire votre processus de collecte des données?
- Traitez-vous fréquemment avec des sources de données en ligne?
- La tâche est-elle non négligeable en termes de portée et de complexité?

LISTE DE VÉRIFICATION APPLICABLE À LA COLLECTE AUTOMATISÉE

Critères : (suite)

- Si la tâche peut être réalisée manuellement, n'avez-vous pas à votre disposition les ressources nécessaires pour laisser autrui faire le travail?
- Êtes-vous disposé à automatiser le processus au moyen de la programmation?

Si la plupart des réponses sont données par l'affirmative, une méthode automatisée pourrait être la voie à suivre.

WORLD WIDE WEB

La façon dont nous **partageons**, **recueillons** et **publions** les données a changé au cours des dernières années, du fait de l'omniprésence du *World Wide Web* (WWW).

Les **entreprises privées**, les **gouvernements** et les **utilisateurs individuels** publient et partagent toutes sortes de données et d'information.

À tout moment, de nouveaux canaux génèrent de vastes quantités de données sur le comportement humain.

LOGICIEL LIBRE

Une autre tendance :

- la croissance, ainsi que la popularité et la puissance sans cesse plus grandes des **logiciels libres** (le code source peut être inspecté, modifié et amélioré par quiconque)

Aspect communautaire → évolution continue et amélioration constante

Les logiciels **R** et **Python** sont des logiciels libres qui peuvent servir à des fins d'analyse de données dans le domaine des sciences sociales et dans d'autres domaines.

Ils intègrent des **interfaces** avec d'autres langages de programmation et **solutions** logicielles.

NETTOYAGE ET TRAITEMENT DES DONNÉES

La collecte des données, en tant que telle, ne constitue que la pointe de l'iceberg.

Le nettoyage ainsi que le traitement des données sont **essentiels** (en plus de nécessiter du temps).

Tâches :

- Sélection des colonnes (variables) présentant de l'intérêt
- Réétiquetage de ces colonnes
- Modification du type de données des colonnes de sorte que les données puissent être utilisées comme nous le souhaitons

NETTOYAGE ET TRAITEMENT DES DONNÉES

Tâches : (suite)

- Édition et/ou extraction des données d'une colonne
- Décider comment gérer les données manquantes (ce qui peut être délicat)
- De multiples autres tâches, selon les données et leurs utilisations

Certaines tâches peuvent être automatisées, d'autres non.

QUESTIONS AU SUJET DE LA QUALITÉ DES DONNÉES

1. Quel type de données est le plus approprié pour répondre à vos questions?
2. La qualité des données est-elle suffisamment élevée pour répondre à votre question?
3. L'information est-elle systématiquement déficiente?

Pouvez-vous parvenir à éviter la redoutée formule : « Eh bien, ce sont les meilleures données dont nous disposions… »?

QUALITÉ DES DONNÉES

Information de première main : à titre d'exemple, un gazouillis ou un article de nouvelles.

Données de deuxième main : données qui ont été copiées d'une source hors ligne ou extraites d'ailleurs.

Parfois, vous ne pouvez vous souvenir de la source des données ou retrouver celle-ci, lorsqu'il s'agit de données de deuxième main.

Convient-il tout de même de s'en servir? Cela dépend.

La **validation croisée** constitue une procédure standard liée à l'utilisation de toute donnée secondaire.

QUALITÉ DES DONNÉES ET OBJECTIF DE L'UTILISATEUR

La qualité des données est fonction de l'**utilisation**.

Par exemple :

- Un échantillon de gazouillis recueillis au cours d'une journée aléatoire pourrait servir à analyser l'utilisation qui est faite d'un mot-clic ou l'utilisation de termes selon le sexe.
- Pas aussi utiles si elles sont recueillies le jour de l'élection pour prédire les résultats de celle-ci (**biais associé à la collecte**).

SOURCES DE DONNÉES (COMPROMIS)

Automatisée c. classique

Exactitude c. exhaustivité

Couverture c. validité

Vitesse c. coût

etc.

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

1. Savoir exactement de quel type d'information vous avez besoin

- Spécifique : PIB de tous les pays membres de l'OCDE au cours des dix dernières années; ventes des dix principales marques de chaussures en 2017
- Vague : l'opinion des gens sur la marque de chaussures X

2. Déterminer s'il existe des sources de données sur le Web qui pourraient fournir de l'information directe ou indirecte sur votre problème

- Plus facile dans le cas de faits spécifiques : la page Web d'un magasin de chaussures fournira de l'information sur les chaussures qui sont actuellement prisées, c.-à-d. sandales, bottes, etc.
- Les gazouillis peuvent permettre de dégager des tendances en matière d'opinion sur *tout et n'importe quoi*
- Les plateformes commerciales peuvent fournir de l'information sur le niveau de satisfaction à l'égard d'un produit

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

3. Élaborer une théorie quant au processus de production des données lorsque l'on se penche sur des sources éventuelles

- Quand les données ont-elles été générées?
- Quand ont-elles été téléchargées sur le Web?
- Qui a téléchargé les données?
- Y a-t-il d'autres aspects qui pourraient ne pas avoir été couverts? Cohérence? Précision?
- À quelle fréquence les données sont-elles mises à jour?

PROCESSUS DE COLLECTE DES DONNÉES (5 ÉTAPES)

4. Trouver un équilibre entre les avantages et les inconvénients des sources de données potentielles

- Valider la qualité des données utilisées
- Existe-t-il d'autres sources indépendantes qui fournissent de l'information similaire, et par rapport auxquelles il serait possible de procéder à une vérification croisée?
- Pouvez-vous identifier la source originale des données secondaires?

5. Prendre une décision

- Choisir la source de données qui semble la plus appropriée
- Documenter les raisons de cette décision
- Recueillir des données de plusieurs sources afin de valider les sources de données

LE MOISSONNAGE DU WEB EST-IL LÉGAL?

Lignes directrices en matière d'éthique :

- Travailler de manière aussi transparente que possible
- Documenter les sources de données en tout temps
- Accorder le crédit à ceux qui, les premiers, ont recueilli et publié les données
- Si vous n'avez pas recueilli l'information, vous aurez vraisemblablement besoin d'une permission pour la reproduire
- Ne faites rien d'illégal

L'extraction d'information d'une autre entreprise en vue de la traiter et de la revendre constitue un motif de plainte courant.

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

Associated Press (AP) c. Meltwater

- Meltwater offre un logiciel qui permet de récupérer ou d'extraire des nouvelles au moyen de mots clés spécifiques.
- Les clients commandent des résumés portant sur certains thèmes en particulier dans lesquels figurent des extraits d'articles de nouvelles.
- AP affirmait que son contenu avait été volé et que Meltwater avait besoin d'une licence pour distribuer l'information extraite.
- Le juge a rendu une décision en faveur d'AP, faisant valoir que Meltwater était un concurrent.
- **Votre verdict?**

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

Facebook c. Pete Warden

- Pete Warden extrayait des renseignements de base des profils d'utilisateurs de Facebook, afin d'offrir des services de gestion des communications et des réseaux.
- Selon lui, son processus allait dans le même sens que robots.txt.
- Après qu'il a eu publié son premier billet de blogue faisant mention des données extraites de Facebook, il a été invité à effacer celles-ci.
- Facebook a fait valoir que robots.txt n'avait aucune valeur juridique et qu'elle pouvait poursuivre quiconque accédait à son site, même si cette personne ou ce groupe se conformait aux instructions en matière de moissonnage, et que la seule façon légale d'accéder à quelque site Web que ce soit au moyen d'un robot Web, c'était en obtenant une autorisation écrite préalable.
- **Votre verdict?**

ACTIONS EN JUSTICE – MOISSONNAGE DU WEB

États-Unis c. Aaron Swartz

- Swartz a participé à la création de RSS, markdown, et Infogami.
- Il a été arrêté en 2011 pour avoir téléchargé illégalement des millions d'articles des archives de JSTOR.
- L'affaire a été classée après qu'il se fut suicidé, en janvier 2013.
- **Votre verdict?**

LEÇONS APPRISES

On ne peut établir clairement quelles mesures de moissonnage sont illégales et lesquelles sont légales.

On estime que le fait de publier de nouveau du contenu à des fins commerciales est plus grave que ne l'est celui qui consiste à télécharger des pages à des fins de recherche ou d'analyse.

Robots.txt : le *protocole d'exclusion des robots* est un fichier qui indique au logiciel de récupération quelle information peut être recueillie sur le site.

Soyez gentil! Il n'est pas nécessaire de récupérer tout ce qui peut être récupéré. Les programmes de récupération devraient se comporter « gentiment », fournir les données que vous recherchez en plus d'être efficients, dans cet ordre de priorité.

cqads.carleton.ca/robots.txt

```
#
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
Disallow: /UPGRADE.txt
Disallow: /xmlrpc.php
```

```
User-agent: Twitterbot
Allow: /

User-agent: *
Disallow: /esi/
Disallow: /webview
Disallow: /vueweb
Disallow: /news/sponsored
Disallow: /search
Disallow: /19849159/
```

theweathernetwork.com/robots.txt

```
User-agent: *
Disallow:
Crawl-delay: 10
```

cfl.ca/robots.txt

COMMUNIQUER AVEC LES FOURNISSEURS DE DONNÉES

Toutes les données auxquelles il est possible d'accéder par le truchement d'un formulaire HTTP sont stockées dans une base de données quelconque.

Demandez tout d'abord aux propriétaires des données s'ils peuvent concéder l'accès à la base de données ou aux fichiers.

Plus la quantité de données qui vous intéressent est importante, **plus il est préférable, pour les deux parties, de communiquer avant le lancement de la collecte de données**.

Pour des petites quantités de données, cela a moins d'importance.

CE QU'IL FAUT ET CE QU'IL NE FAUT PAS FAIRE EN MATIÈRE DE MOISSONNAGE

1. Demeurer identifiable

2. Réduire le trafic

- Accepter les fichiers comprimés
- En cas de moissonnage des mêmes ressources à plusieurs reprises, vérifiez tout d'abord si celles-ci ont changé avant d'y accéder de nouveau
- Ne récupérer que des parties de fichier

CE QU'IL FAUT ET CE QU'IL NE FAUT PAS FAIRE EN MATIÈRE DE MOISSONNAGE

3. Ne pas soumettre de demandes multiples au serveur

- Le fait de soumettre de nombreuses demandes par seconde peut entraîner la mise hors service des serveurs peu puissants
- Les webmaîtres peuvent vous bloquer si votre logiciel de récupération de données se comporte de cette façon
- On considère qu'une ou deux demandes par seconde est un rythme acceptable

4. Concevoir un logiciel de récupération de données modeste (efficient et poli)

- Il est inutile de récupérer des pages quotidiennement ou de répéter la même tâche sans cesse; faites en sorte que votre programme de récupération de données soit aussi efficient que possible
- Ne pas soumettre des pages à un trop grand nombre de demandes récupération
- Sélectionner les ressources que vous souhaitez utiliser et laisser le reste intact

OUTILS DE DÉVELOPPEMENT

Les outils de développement nous permettent (notamment) d'observer la correspondance entre le code HTML d'une page et la version rendue que nous retrouvons dans le navigateur.

Contrairement à la fonction « Afficher la source », les outils de développement présentent la version *dynamique* du code HTML (c. à d. que les instructions HTML apparaissent sans les modifications apportées par JavaScript depuis que la page a été reçue, la première fois).

Il est essentiel, pour récupérer des données des sites Web de façon efficiente, d'inspecter les divers éléments qui composent une page et de déterminer où ils se trouvent dans le fichier HTML.

OUTILS DE DÉVELOPPEMENT

Firefox

- clic du bouton droit de la souris dans la page → Inspecter l'élément

Safari

- Safari → Préférences → Avancées → Afficher le menu Développer dans la barre de menus
- Développer → Afficher inspecteur Web

Chrome

- clic du bouton droit de la souris dans la page → Inspecter

Nice Peter - ERB

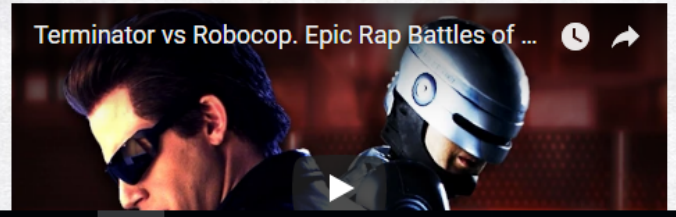
Secure | https://nicepeter.com/erb

HOME | LIVE SHOWS | ERB | MUSIC | VIDEOS | GALLERY | SHOP | PRESS | ARCHIVE | CONTACT

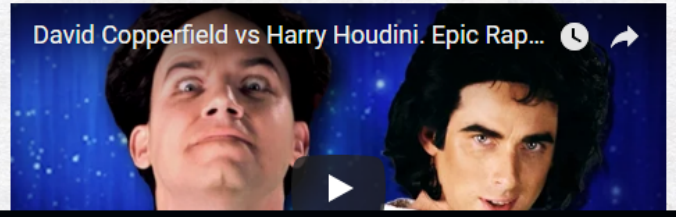
Shaka Zulu vs Julius Caesar

Eastern Philosophers vs Western Philosophers

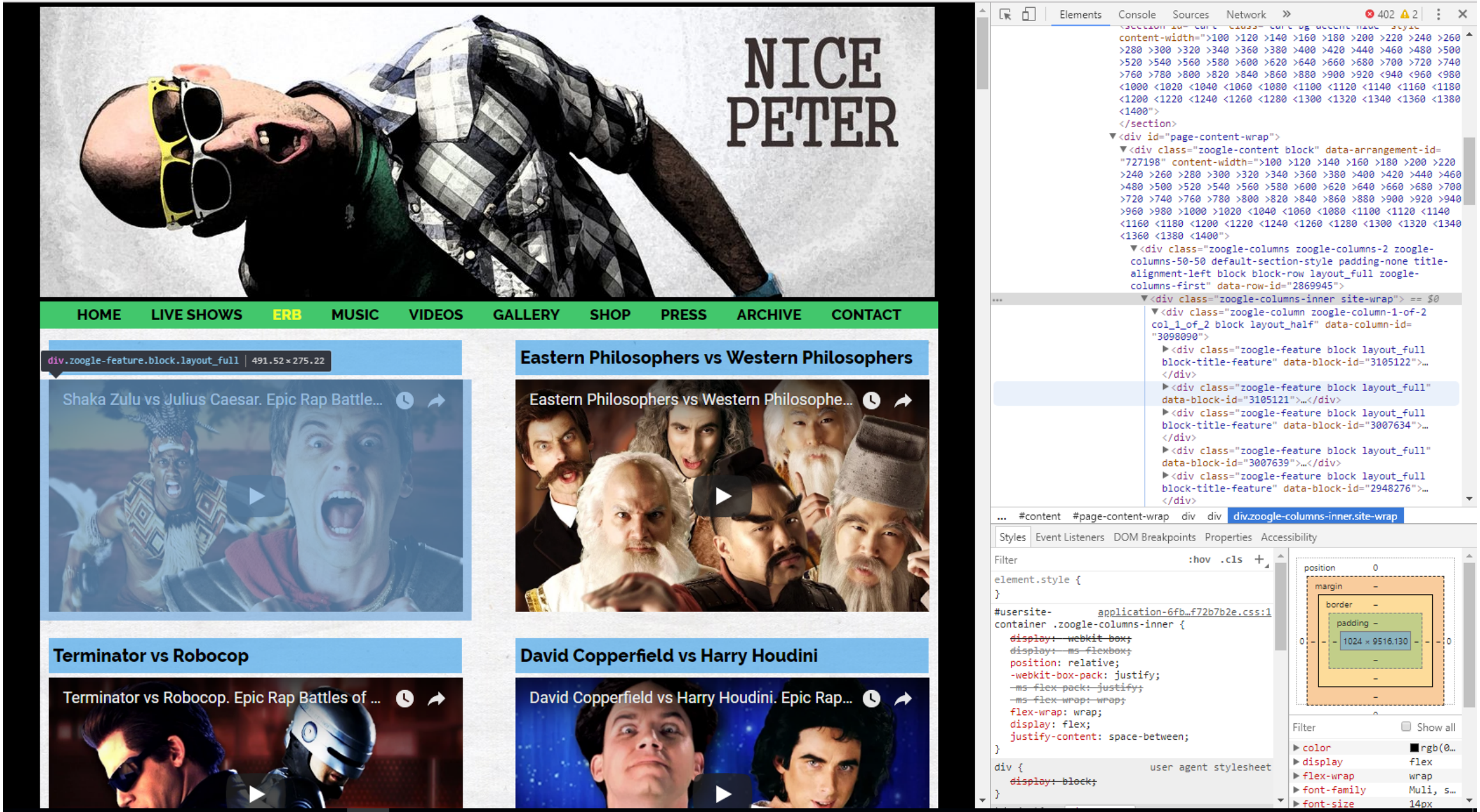
Terminator vs Robocop



David Copperfield vs Harry Houdini



12:10 PM

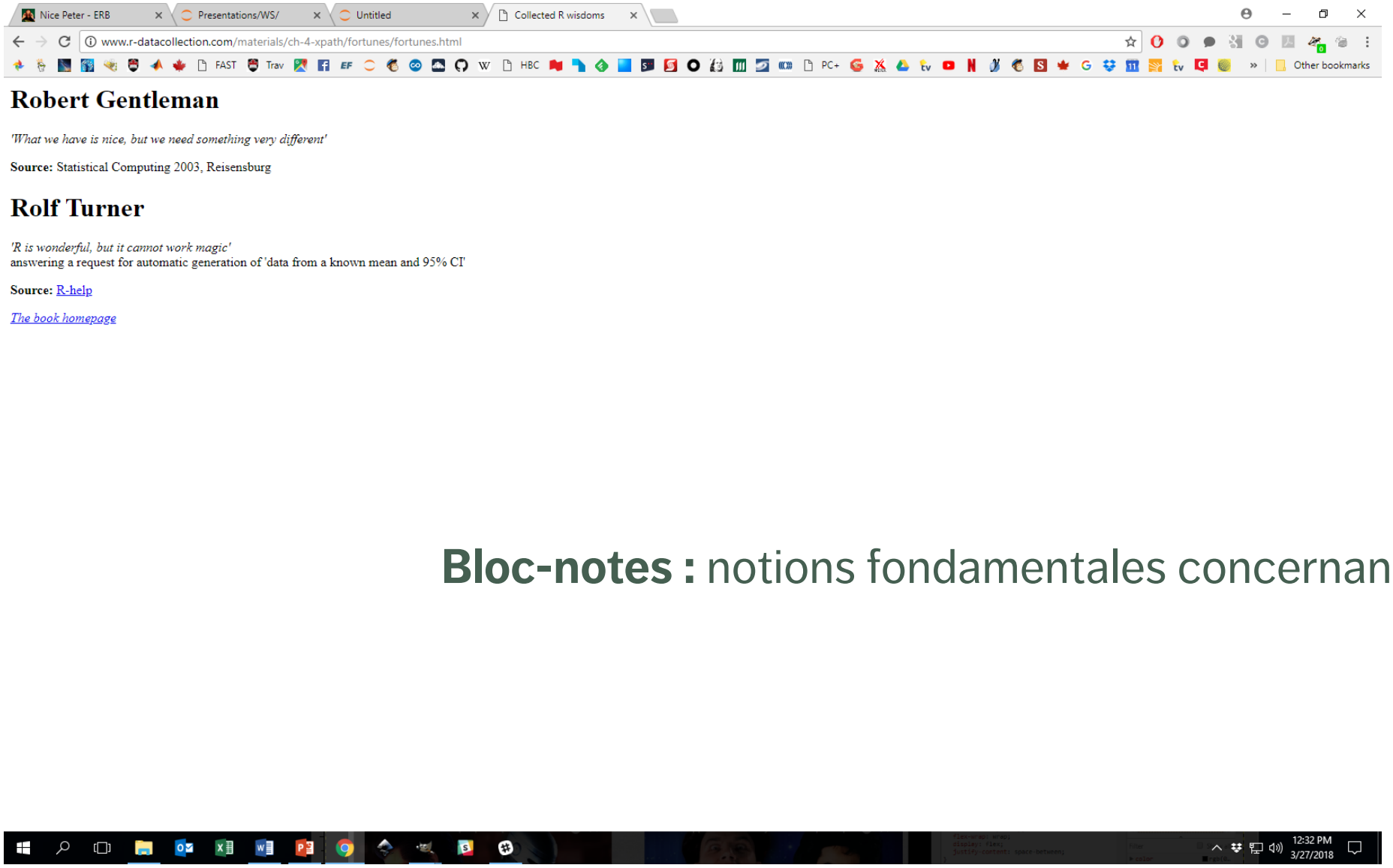

Nice Peter - ERB
Secure | https://nicepeter.com/erb
NICE PETER
HOME LIVE SHOWS ERB MUSIC VIDEOS GALLERY SHOP PRESS ARCHIVE CONTACT
div.zoogle-feature.block.layout_full | 491.52 × 275.22
Shaka Zulu vs Julius Caesar. Epic Rap Battle...
Eastern Philosophers vs Western Philosophers
Eastern Philosophers vs Western Philosophe...
Terminator vs Robocop
Terminator vs Robocop. Epic Rap Battles of ...
David Copperfield vs Harry Houdini
David Copperfield vs Harry Houdini. Epic Rap...
Elements Console Sources Network
#content #page-content-wrap div div div.zoogle-columns-inner.site-wrap
Styles Event Listeners DOM Breakpoints Properties Accessibility
element.style {
#usersite-container .zoogle-columns-inner {
display: flex;
flex-wrap: wrap;
justify-content: space-between;
div { user agent stylesheet
display: block;
position 0
margin
border
padding
1024 × 9516.130
12:11 PM
3/27/2018

XPATH

XPath est un langage d'interrogation (propre à un domaine)

- Il est utilisé pour sélectionner des éléments d'information spécifiques dans des documents balisés, comme HTML, XML ou des variantes telles que SVG et RSS
- L'information stockée dans les documents balisés doit être convertie dans des formats qui se prêtent au traitement et à l'analyse statistique
- Mise en œuvre dans le `XML` du progiciel R
- Étapes du processus :

  1. Préciser les données présentant de l'intérêt
  2. Les situer dans un document spécifique
  3. Adapter une interrogation au document en vue d'extraire les renseignements souhaités



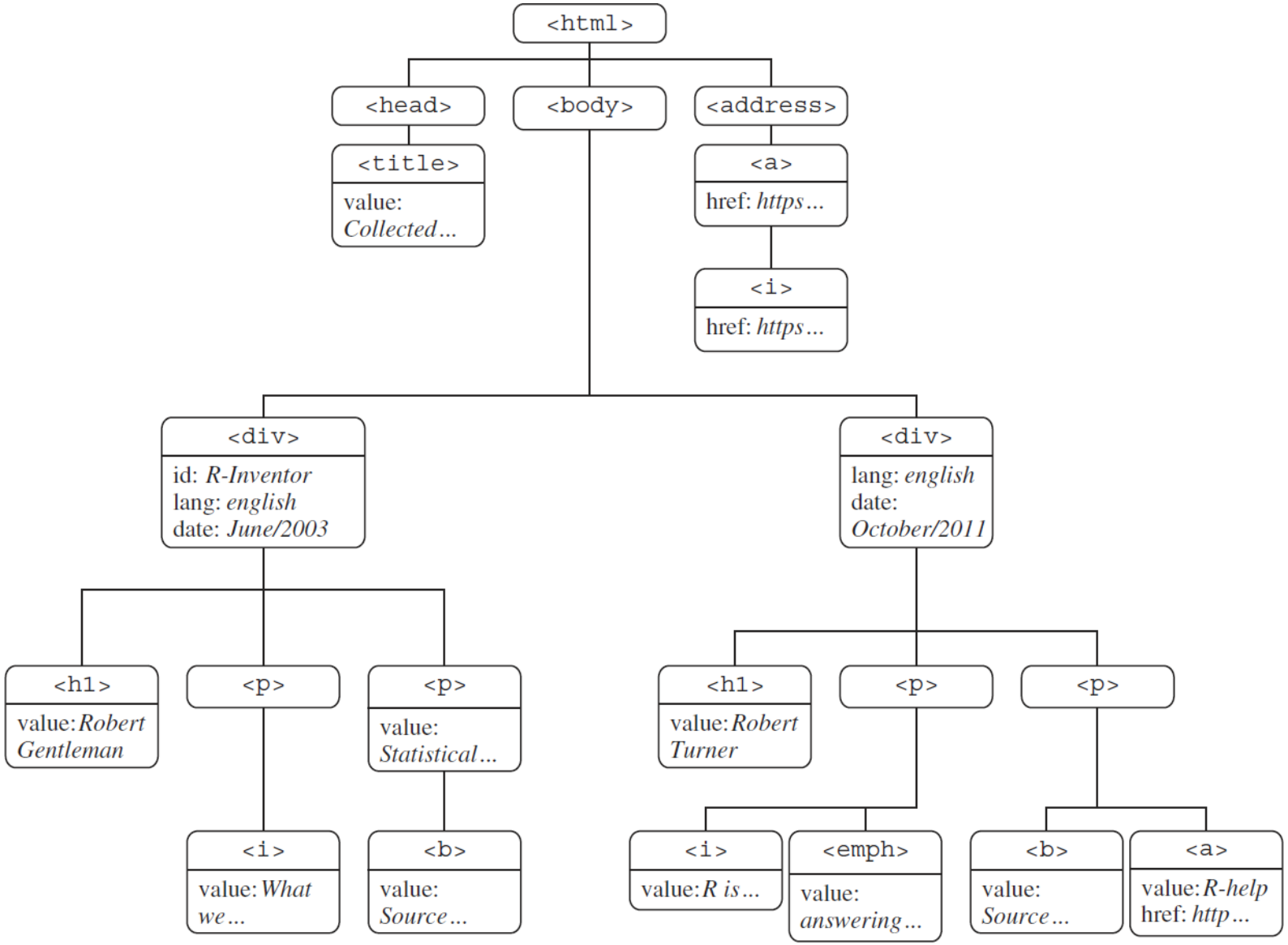
Bloc-notes : notions fondamentales concernant XPath

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head><title>Collected R wisdoms</title></head>
<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i></p>
  <p><b>Source: </b>Statistical Computing 2003, Reisensburg</p>
</div>

<div lang="english" date="October/2011">
  <h1>Rolf Turner</h1>
  <p><i>'R is wonderful, but it cannot work magic'</i> <br><emph>answering a request
for automatic generation of 'data from a known mean and 95% CI'</emph></p>
  <p><b>Source: </b><a href="https://stat.ethz.ch/mailman/listinfo/r-help">R-help</a>
</p>
</div>

<address>
<a href="http://www.r-datacollectionbook.com"><i>The book homepage</i></a><a></a>
</address>

</body>
</html>
```


<html>
<head>
<body>
<address>
<title>
value: Collected...
<a>
href: https...
<i>
href: https...

id: R-Inventor
lang: english
date: June/2003

lang: english
date: October/2011
<h1>
value: Robert Gentleman
<p>
<p>
value: Statistical...
<i>
value: What we...
<b>
value: Source...
<h1>
value: Robert Turner
<p>
<p>
<i>
value: R is...
<emph>
value: answering...
<b>
value: Source...
<a>
value: R-help
href: http...

XPATH – STRUCTURE DE BASE

Les balises HTML/XML présentent des **attributs** et des **valeurs**.

Les fichiers HTML doivent être analysés avant qu'ils puissent faire l'objet d'une interrogation par XPath.

Les interrogations XPath ont besoin d'un chemin d'accès et d'un document visé par la recherche.

- les chemins d'accès consistent en un mécanisme d'adressage hiérarchique (succession de nœuds, séparés par des barres obliques [« / »])
- les interrogations se présentent selon le format `xpathSApply(doc,path)` :
  - `xpathSApply(doc_analysé,"/html/body/div/p/i")`

  cette instruction permet d'extraire toutes les balises `<i>` qui se trouvent à l'intérieur d'une balise `<p>` à l'intérieur d'une balise `<div>` dans le `corps` du fichier `html`.

XPATH – RELATIONS DES NŒUDS

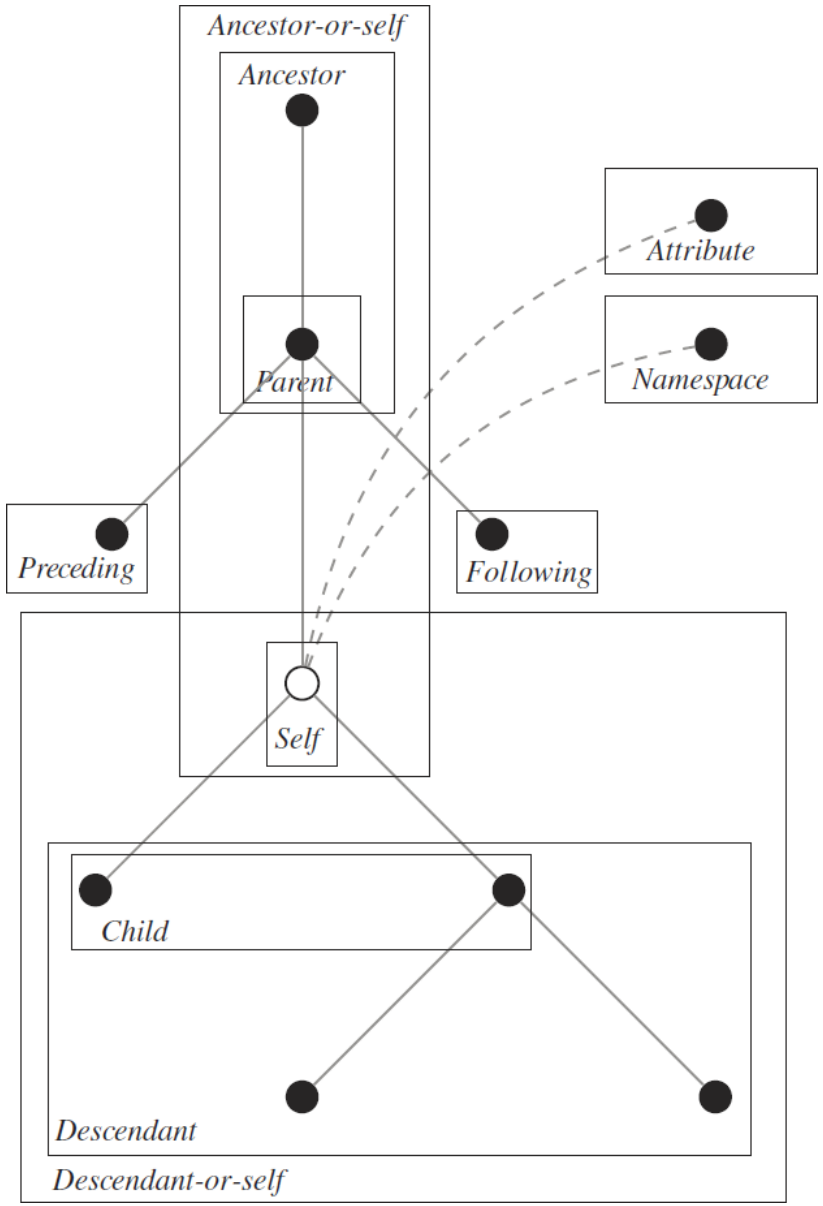
Les chemins d'accès absolus (voire relatifs) ne peuvent pas toujours sélectionner de manière succincte des nœuds dans de gros fichiers ou dans des fichiers complexes.

Analogie de l'arborescence familiale : la place du nœud à l'intérieur de l'arborescence analysée se rapproche fréquemment des relations qu'entretiennent les familles élargies.

Les relations sont désignées par rapport à `node1/relation::node2`.

Exemples :

- « `//a/ancestor::div` » permet d'extraire tous les nœuds `<div>` qui précèdent le nœud `<a>`.
- « `//a/ancestor::div//i` » permet d'extraire tous les nœuds `<i>` qui se trouvent à l'intérieur d'un nœud `<div>` qui précède un nœud `<a>`, etc.



Axis name	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
namespace	Selects all namespace nodes of the current node
parent	Selects the parent of the current node
preceding	Selects all nodes that appear before the current node in the document except ancestors, attribute nodes, and namespace nodes
preceding-sibling	Selects all siblings before the current node
self	Selects the current node

XPATH – PRÉDICATS

Un prédicat est une fonction qui s'applique au nom, à la valeur ou aux attributs d'un nœud et qui produit une réponse logique *VRAI (TRUE)* ou *FAUX (FALSE)*.

Les prédicats modifient le chemin d'entrée d'une interrogation XPath. Les nœuds pour lesquels la relation s'avère exacte sont sélectionnés par l'interrogation.

Les prédicats sont présentés entre crochets, et suivent un nœud.

Exemples :

- « `//p[position()=1]` » permet d'extraire le premier nœud <p> par rapport à son nœud parent `<p>`.
- « `//p[last()]` » permet d'extraire le dernier nœud <p> par rapport à son nœud parent `<p>`.
- « `//div[count(./@*)>2]` » permet d'extraire tous les nœuds `<div>` avec deux attributs ou plus, etc.

Function	Description	Example
`name(<node>)`	Returns the name of `<node>` or the first node in a node set	`//*[name()='title']`; Returns: `<title>`
`text(<node>)`	Returns the value of `<node>` or the first node in a node set	`//*[text()='The book homepage']`; Returns: `<i>` with value *The book homepage*
`@attribute`	Returns the value of a node's *attribute* or of the first node in a node set	`//div[@id='R Inventor']`; Returns: `<div>` with attribute *id* value *R Inventor*
`string-length(str1)`	Returns the length of `str1`. If there is no string argument, it returns the length of the string value of the current node	`//h1[string-length()>11]`; Returns: `<h1>` with value *Robert Gentleman*
`translate(str1, str2, str3)`	Converts `str1` by replacing the characters in `str2` with the characters in `str3`	`//div[translate(./@date, '2003', '2005')='June/2005']`; Returns: first `<div>` node with date attribute value *June/2003*
`contains(str1,str2)`	Returns `TRUE` if `str1` contains `str2`, otherwise `FALSE`	`//div[contains(@id, 'Inventor')]`; Returns: first `<div>` node with id attribute value *R Inventor*
`starts-with(str1,str2)`	Returns `TRUE` if `str1` starts with `str2`, otherwise `FALSE`	`//i[starts-with(text(), 'The')]`; Returns: `<i>` with value *The book homepage*
`substring-before (str1,str2)`	Returns the start of `str1` before `str2` occurs in it	`//div[substring-before(@date, '/')='June']`; Returns: `<div>` with date attribute value *June/2003*
`substring-after (str1,str2)`	Returns the remainder of `str1` after `str2` occurs in it	`//div[substring-after(@date, '/')=2003]`; Returns: `<div>` with date attribute value *June/2003*
`not(arg)`	Returns `TRUE` if the boolean value is `FALSE`, and `FALSE` if the boolean value is `TRUE`	`//div[not(contains(@id, 'Inventor'))]`; Returns: the `<div>` node that does not contain the string *Inventor* in its id attribute value
`local-name(<node>)`	Returns the name of the current `<node>` or the first node in a node set—without the namespace prefix	`//*[local-name()='address']`; Returns: `<address>`
`count(<node>)`	Returns the count of a nodeset `<node>`	`//div[count(.//a)=0]`; Result: The second `<div>` with one `<a>` child
`position(<node>)`	Returns the index position of `<node>` that is currently being processed	`//div/p[position()=1]`; Result: The first `<p>` node in each `<div>` node
`last()`	Returns the number of items in the processed node list `<node>`	`//div/p[last()]`; Result: The last `<p>` node in each `<div>` node

COMMUNIQUÉS DE PRESSE DU GOUVERNEMENT DU ROYAUME-UNI – CONTEXTE

Le gouvernement du Royaume-Uni publie tous ses communiqués de presse en ligne, à l'adresse gov.uk/government/announcements.

Le 29 mars 2018, on dénombrait plus de 65 000 communiqués de presse sur le site.

Questions :

- Pouvons-nous prédire quel organisme a fait une annonce en se fiant uniquement au contenu textuel de cette dernière?
- Y a-t-il des thèmes qui semblent sans cesse revenir à l'avant-plan?

GOV.UK

Search

Departments Worldwide How government works Get involved
Policies Publications Consultations Statistics Announcements

Announcements

You can use the filters to show only results that match your interests

Contains

keywords

Announcement type

All announcement types

Policy area

All policy areas

Department

All departments

Person

All people

World locations

All locations

65,716 announcements

Get updates to this list email feed

Welsh innovation is key to Britain's future export success
24 March 2018 WO Speech

Preventing Hunger as a Weapon of War
23 March 2018 FCO Speech

"Our vote today against this resolution is a vote against the politicization of the Commission on the Status of Women."
23 March 2018 FCO Speech

Lord Ahmad welcomes conclusions of the 37th Session of the UN Human Rights Council
23 March 2018 FCO Speech

Rt Hon Mark Field MP speech at Global FinTech Investor Forum
23 March 2018 FCO Speech

Press release

Foreign Secretary statement on Iran

The Foreign Secretary has made the following statement on the protests in Iran.

Published 1 January 2018

From: Foreign & Commonwealth Office and The Rt Hon Boris Johnson MP

The Foreign Secretary Boris Johnson said:

" The UK is watching events in Iran closely. We believe that there should be meaningful debate about the legitimate and important issues the protesters are raising and we look to the Iranian authorities to permit this."

" We also believe that, particularly as we enter the 70th anniversary year of the Universal Declaration on Human Rights, people should be able to have freedom of expression and to demonstrate peacefully within the law."

" We regret the loss of life that has occurred in the protests in Iran, and call on all concerned to refrain from violence and for international obligations on human rights to be observed."

Chaque communiqué de presse contient ce qui suit :

- titre
- date de publication
- organismes/personnes l'ayant publié
- texte du communiqué de presse

Les communiqués de presse portent principalement sur 2017 et proviennent des bureaux ou organismes suivants :

- Bureau du pays de Galles
- Ministère des Affaires étrangères
- Ministère des Sciences et de la Technologie
- Ministère de l'Environnement, de l'Alimentation et des Affaires rurales

---

Bloc-notes : communiqués de presse du gouvernement du Royaume-Uni

EXPRESSIONS RÉGULIÈRES

L'objectif principal du moissonnage du Web consiste à recueillir de l'information **utile** pour le problème faisant l'objet de travaux de recherche, à partir d'une quantité considérable de données textuelles.

Nous nous intéressons aux éléments systématiques des données textuelles, tout particulièrement si des méthodes quantitatives seront éventuellement appliquées.

Les structures systématiques peuvent prendre les formes suivantes :

- nombres
- noms (pays, etc.)
- adresses (postales, courriel, URL, etc.)
- chaînes de caractères spécifiques, etc.

EXPRESSIONS RÉGULIÈRES

Les expressions régulières (« regexps ») permettent l'extraction systématique des composantes d'information.

Les **expressions régulières** sont des séquences abstraites de chaînes qui correspondent à des modèles concrets récurrents qui se retrouvent dans le texte.

Elles peuvent servir à extraire de l'information de fichiers en texte brut, voire de type HTML et XML.

Utiles lorsque l'information est dissimulée à l'intérieur de valeurs *atomiques*.

Bloc-notes : expressions régulières en Python et plus encore

BEAUTIFUL SOUP

Pour qu'elles permettent d'extraire une page et le contenu HTML, les demandes Web élémentaires doivent être assorties d'instructions réseau.

Les navigateurs accomplissent énormément de travail pour analyser de manière intelligente une syntaxe HTML absolument inappropriée, comme dans le cas suivant :

```
<a href="crummy.com> <b>link text<a> </b>
```

Beautiful Soup est une bibliothèque Python qui facilite l'extraction de données de fichiers HTML et XML. Cette bibliothèque analyse les fichiers HTML, même s'ils sont brisés.

BEAUTIFUL SOUP

Beautiful Soup (BS) ne se limite pas à convertir des instructions HTML qui laissent à désirer en instructions XHTML, de sorte que celles-ci puissent être analysées au moyen d'un logiciel d'analyse XML.

BS permet à un utilisateur d'inspecter intégralement la structure HTML (appropriée) qu'elle produit, grâce à un programme.

Une fois que BS a terminé son travail portant sur un fichier HTML, il en résulte une API qui permet de soumettre les éléments du document à un survol, à une recherche et à une lecture.

BEAUTIFUL SOUP

Les éléments HTML qui sont généralement extraits/lus se présentent selon divers formats :

- texte
- tableaux
- valeurs de champs
- images
- vidéos
- etc.

Beautiful Soup offre des façons **idiomatiques** de soumettre l'arborescence d'analyse du fichier HTML à des opérations de navigation, de recherche et de modification (ce qui fait gagner énormément de temps).

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>

<p class="story">...</p>
"""
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
#  <head>
#   <title>
#    The Dormouse's story
#   </title>
#  </head>
#  <body>
#   <p class="title">
#    <b>
#     The Dormouse's story
#    </b>
#   </p>
#   <p class="story">
#    Once upon a time there were three little sisters; and their names were
#    <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#    </a>
#    ,
#    <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#    </a>
#    and
#    <a class="sister" href="http://example.com/tillie" id="link2">
#     Tillie
#    </a>
#    ; and they lived at the bottom of a well.
#   </p>
#   <p class="story">
#    ...
#   </p>
#  </body>
# </html>
```

```
soup.title
# <title>The Dormouse's story</title>

soup.title.name
# u'title'

soup.title.string
# u'The Dormouse's story'

soup.title.parent.name
# u'head'

soup.p
# <p class="title"><b>The Dormouse's story</b></p>

soup.p['class']
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```

```
for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

```
print(soup.get_text())
# The Dormouse's story
#
# The Dormouse's story
#
# Once upon a time there were three little sisters; and their names were
# Elsie,
# Lacie and
# Tillie;
# and they lived at the bottom of a well.
#
# ...
```

SELENIUM

Selenium est un outil qui permet d'automatiser les interactions avec les navigateurs Web (en Python). S'il sert généralement à automatiser les applications Web à des fins de mise à l'épreuve, il offre également d'autres applications.

Il permet principalement à un utilisateur d'ouvrir un navigateur et d'effectuer des tâches analogues à celles qu'exécuterait un humain, comme les suivantes :

- cliquer sur un bouton
- introduire des données dans un formulaire
- rechercher de l'information particulière dans des pages Web
- etc.

SELENIUM

Selenium a besoin d'un pilote pour établir une interface avec le navigateur choisi. À titre d'exemple, Firefox a besoin de *geckodriver*.

Les autres navigateurs pris en charge disposeront de leurs propres pilotes :

Chrome : https://sites.google.com/a/chromium.org/chromedriver/downloads
Edge : https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/
Firefox : https://github.com/mozilla/geckodriver/releases
Safari : https://webkit.org/blog/6900/webdriver-support-in-safari-10/

SIMULATION D'UN NAVIGATEUR WEB

Selenium contrôle automatiquement l'ensemble du navigateur, y compris en ce qui concerne le rendu des documents Web et l'exécution de JavaScript.

Ceci est utile pour les pages qui intègrent un fort contenu dynamique qui ne se retrouve pas dans le fichier HTML de base.

Selenium peut programmer des actions telles que « cliquer sur ce bouton » ou « taper ce texte », et vous avez en tout temps accès au fichier HTML dynamique qui correspond à l'état actuel de la page, comme ce que l'on retrouve dans les outils de développement.

UTILISATION DES API

Une API constitue la façon, pour un site Web, d'offrir l'accès à ses données à un programme, sans qu'il soit nécessaire d'en récupérer le contenu.

Ainsi donc, une API offre un **accès structuré** à des **données structurées**.

À titre d'exemple, un site à caractère financier pourrait offrir une API assortie de données financières, tout comme le *New York Times* pourrait offrir une API adaptée aux articles de nouvelles.

Dans l'un ou l'autre des cas, les données se présenteraient dans un format prédéfini, structuré (lequel est fréquemment JSON).

UTILISATION DES API

Les API que nous examinerons intègrent des bibliothèques R/Python qui prennent en charge l'ensemble des opérations de réseau et de codage requises.

Cela signifie qu'il suffit de lire la documentation relative à la bibliothèque pour savoir quoi faire.

Exercice : servez-vous de Zomato pour trouver quelle ville canadienne propose les meilleurs restaurants de sushi (https://github.com/fatihsucu/pyzomato).

API DE YOUTUBE – KHAN ACADEMY

Des millions de vidéos sont accessibles par **YouTube**.

Il n'est pas évident de déterminer comment l'on s'y prendrait pour extraire, de manière générale, du contenu vidéo du Web (si ce n'est par l'entremise des URL); à certaines vidéos correspond du contenu sous forme de texte (**transcriptions**).

Nous avons recours à l'API de YouTube pour extraire ce contenu.

Bloc-notes : transcriptions YouTube

[https://www.youtube.com/playlist?list=PLjAmOqZwbYJ1kwdT6kn_YaulXK_ZyB4Ww]

YouTube CA

Search

SIGN IN

- Home
- Trending
- History

BEST OF YOUTUBE

- Music
- Sports
- Gaming
- Movies
- TV Shows
- News
- Live
- 360° Video

Browse channels

Sign in now to see your channels and recommendations!

SIGN IN

Statistics

68 videos • 3,290,303 views • Last updated on Jul 2, 2014

Khan Academy

SUBSCRIBE

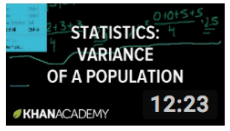
Introduction to statistics. Will eventually cover all of the major topics in a first-year statistics course (not there yet!)

1 

Statistics: The average | Descriptive statistics | Probability and Statistics | Khan Academy
Khan Academy

2 

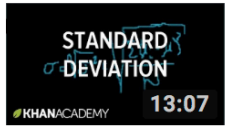
Statistics: Sample vs. Population Mean
Khan Academy

3 

Statistics: Variance of a population | Probability and Statistics | Khan Academy
Khan Academy

4 

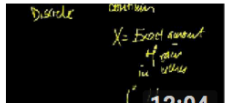
Statistics: Sample variance | Descriptive statistics | Probability and Statistics | Khan Academy
Khan Academy

5 

Statistics: Standard deviation | Descriptive statistics | Probability and Statistics | Khan Academy
Khan Academy

6 

Statistics: Alternate variance formulas | Probability and Statistics | Khan Academy
Khan Academy

7 
Introduction to Random Variables
Khan Academy

RÉFÉRENCES

COLLECTE DES DONNÉES

RÉFÉRENCES

http://www.roymfrancis.com/scraping-instagram-choosing-hashtags/

Munzert, S., C. Rubba, P. Meissner et D. Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley, 2015.

Mitchell, R. *Web Scraping with Python: Collecting Data From the Modern Web*, O'Reilly, 2015.

https://www.w3schools.com/xml/xpath_intro.asp

https://www.w3schools.com/

https://fr.wikipedia.org/wiki/Extensible_Hypertext_Markup_Language

https://medium.com/the-andela-way/introduction-to-web-scraping-using-selenium-7ec377a8cf72

https://pypi.python.org/pypi/selenium

RÉFÉRENCES

Lessler, J. et Kalsbeek, W. *Nonsampling Errors in Surveys*, New York, Wiley, 1992.

Oppenheim, N. *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's, 1992.

Hidiroglou, M., J. Drew et G. Gray. « A Framework for Measuring and Reducing non-response in Surveys », *Survey Methodology*, vol. 19, nº 1, 1993, p. 81 à 94.

Gower, A. « Questionnaire Design for Business Surveys », *Survey Methodology*, vol. 20, nº 2, 1994.

Méthodes et pratiques d'enquête, Statistique Canada. No 12-587-X au catalogue.

Boily, P., J. Schellinck, S. Hagiwara *et coll. Introduction to Quantitative Consulting*. En cours d'élaboration.

Buttrey, S.E. *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Wiley, 2017.