



ÉCOLE CENTRALE LYON

MOS 4.4 : NOUVELLES TECHNOLOGIES DE L'INFORMATION  
ET DE LA COMMUNICATION

---

# État de l'art : L'utilisation du Big Data dans la politique américaine

---

*Élève:*

Raphaël LECOUTRE

*Encadrant:*

Stéphane DERRODE

14 mars 2020

## Table des matières

<b>1</b>	<b>L'utilisation du Big Data dans les campagnes politiques</b>	<b>2</b>
1.1	L'arrivée du Big Data en politique : la campagne Obama . . . . .	3
1.2	Le phénomène "Cambridge Analytica" . . . . .	4
1.3	Le Gerrymandering . . . . .	5
<b>2</b>	<b>L'utilisation du Big Data par les politiques une fois au pouvoir</b>	<b>7</b>
2.1	Application aux services pour les citoyens . . . . .	7
2.2	Application aux renseignements secrets . . . . .	8
<b>3</b>	<b>État de l'art législatif</b>	<b>8</b>
3.1	Droit des citoyens américains . . . . .	8
3.2	Futures règles des campagnes par rapport au Big Data . . . . .	10
	<b>Veille</b>	<b>11</b>
	<b>Bibliographie</b>	<b>13</b>

# 1 L'utilisation du Big Data dans les campagnes politiques

L'essor d'Internet et de la donnée au fil de la dernière décennie a touché tous les secteurs d'activités, en particulier la politique. Cette nouvelle ressource est en effet devenue primordiale dans les élections politiques pour gagner l'avantage sur son adversaire, en connaissant notamment mieux les électeurs et leurs vies, pour ainsi prédire leur comportement aux futures élections.

Mais il ne s'agit pas seulement de prédiction car les candidats, surtout le camp démocrate a utilisé aux mieux les données pour mieux affiner leur stratégie. Car, rappelons-le, les élections américaines sont plutôt compliquées. C'est une élection indirecte dans laquelle il faut conquérir la majorité des grands électeurs et non celles des électeurs. En général, les deux vont de pair mais il est arrivé à plusieurs reprises – comme en 2000 – qu'un candidat soit élu alors qu'il est minoritaire en voix. Ce fut le cas de George Bush en 2000 face à Al Gore, après la décision, très contestée de la Cour Suprême, de lui donner la victoire de l'État de Floride et donc de l'élection. Les candidats en lice ont tendance à se concentrer sur les États qui font la différence et que l'on appelle les Swing States qui changent assez facilement de camp. Aujourd'hui, la TV et les canaux classiques ont perdu leur prééminence et le Web est un canal qui prend de plus en plus d'importance. Plus du tiers des électeurs américains ont déclaré avoir voté lors d'une élection à la suite d'une publicité électorale en ligne, et environ 60% des indécis choisissent un candidat moins d'un mois avant l'élection, alors que les plus décidés peuvent arrêter le choix de leur candidat plus de trois mois avant le jour du scrutin (pour 26%). L'utilisation de l'intelligence artificielle et du Big Data montre donc qu'il est préférable d'atteindre les électeurs indécis plus d'un mois avant une élection, et non au cours des derniers jours.

## Voters supporting a continued legalization of assault weapons were more likely to:

- Go on a family road trip (+66%)
- Be male (+61%)
- Be fiscally conservative (+55%)
- Buy tools for home improvement (+48%)
- Make home improvements during the summer (+37%)
- Buy Velveeta cheese, Rotel tomatoes, and canned green chiles to make a spicy cheese dip (35%)
- Enjoy CBS programming (+29%)
- Donate to a veterans' organization (+27%)
- Be online gamers (+26%)
- Be a dog owner (+26%)
- Buy furniture and accessories (+25%)
- Visit a theme park (+25%)
- Consider themselves to be patriotic Americans (+14%)

(Source : *Preparing for 2016*)

FIGURE 1 – Exemple de corrélations mises en valeur par le Big Data

## 1.1 L'arrivée du Big Data en politique : la campagne Obama

Aux États-Unis, c'est à travers la campagne de Barack Obama en 2008 que Internet (les réseaux sociaux en particulier) a commencé à être utilisé de façon méthodique et précise pour aider à la campagne.

En effet, après cette campagne, le camp Obama décide de quintupler le nombre d'analystes «maison», à presque 100, selon Mother Jones. Le patron, Jim Messina, très proche du président de Google, Eric Schmidt, recrute deux nerds. Rayid Ghani, un ancien d'Accenture, chercheur spécialisé en «data mining et machine learning», utilisait notamment ses talents pour maximiser l'efficacité des promotions des supermarchés. Spécialiste de l'analyse «en temps réel», Harper Reed, lui, s'est fait les dents sur le réseau des transports publics de Chicago. L'équipe de choc est réunie dans une pièce sans fenêtre baptisée «la cave». Ces « grosses têtes » de l'analyse statistique se sont alors mis en devoir d'exploiter toutes les données du web afin d'en extraire des tendances permettant, premièrement, de lever le maximum d'argent et, deuxièmement, de cibler au plus juste les thèmes intéressant l'opinion et tout particulièrement les électeurs indécis ou encore les communautés les plus délaissées.

Les Data Crunchers (= analystes) ont également (et peut-être surtout) servi à lever la plus importante somme d'argent jamais vue dans une campagne présidentielle : plus d'1 milliard de dollars. Pour ce faire, l'analyse statistique a été mise à contribution :

Pour quelles raisons certains soutiens de 2008 s'étaient désabonnés des listes et surtout comment les faire revenir ? Quelles étaient les campagnes d'emailing obtenant les meilleurs résultats ? Le message devait-il être signé par le candidat/président ou par une autre personne ? (le Président réélu a remercié une nouvelle fois sa femme pour son rôle, les messages signés par elle remportant effectivement le plus d'adhésion).

Afin de réaliser cette campagne numérique, l'équipe d'Obama a divisé la tâche en plusieurs étapes clés :

- Étape 1 : amasser les données. Les sources principales : les registres électoraux publics, les affiliations au parti démocrate ou républicain, les profils des supporters sur MyBarackObama et ObamaForAmerica et, grâce à Facebook Connect (utilisé pour se connecter sans créer un profil), les informations que Mark Zuckerberg accepte de partager avec des partis tiers.
- Étape 2 : interconnecter les données. En 2008, chaque pôle de la campagne (collecte de fonds, porte à porte, spots publicitaires) disposait de ses informations, parfois sous forme de registres papiers. Cette fois, une seule gigantesque base a été constituée, accessible à tous les niveaux, notamment sur smartphone et tablette, pour une action coordonnée.
- Étape 3 : utiliser les données. Obama a bombardé ses supporters d'emails ciblés les appelant à donner quelques dollars. Exemple type : un fan écrit sur le site officiel que les questions sociales sont particulièrement importantes pour lui, il reçoit un email sur les dérapages des républicains sur le viol et l'avortement. L'efficacité de chaque courrier était mesurée, notamment l'impact du nom de l'expéditeur (Michelle Obama, Joe Biden etc.). Publicité, budget alloué à chaque Etat... Chaque décision était basée sur un chiffre. L'élection était encore simulée 66.000 fois chaque soir avec les derniers chiffres. Le matin, la stratégie et les ressources étaient ajustées

en conséquence.

L'opposition en la personne de Romney a, de son côté tenté d'utiliser ce type d'outil. Le stratège Karl Rove a été l'un des pionniers d'une stratégie basée sur les nombres. Jusqu'aux primaires, la campagne républicaine a surtout sous-traité cette mission au lieu d'en faire le cœur de sa machine. De plus, les modèles républicains de la réalité sur le terrain se sont révélés incorrects, avec de nombreux États qu'Obama a largement remportés alors que Mitt Romney y avait activement fait campagne.

## 1.2 Le phénomène "Cambridge Analytica"

Les élections américaines de 2016 ont signifié l'arrivée d'une nouvelle ère à l'utilisation des données des électeurs pour les campagnes politiques, avec l'élection de Donald Trump en tant que président des États-Unis. Cette victoire a notamment était possible grâce à l'utilisation du Big Data, et à une entreprise en particulier : Cambridge Analytica.

Cambridge Analytica est une filiale du groupe de communication britannique Strategic Communication Laboratories (SCL), lui-même spécialisé dans l'analyse de données. . Au moment de la victoire de Trump, le patron de SLC s'était vanté d'avoir joué "un rôle déterminant dans l'identification des partisans [de Trump] et dans la persuasion des électeurs indécis pour les amener à participer au scrutin".

Aux Etats-Unis, la société s'est appuyée sur une application, **thisisyourdigitallife**, qui proposait un test à quelques 270 000 utilisateurs de Facebook. Ces volontaires étaient gratifiés de quelques dollars en échange d'informations sur leurs goûts, censés ensuite être analysés par des psychologues.

Cependant, l'utilisation des données de Facebook par Cambridge Analytica pose déjà problème sur la forme. La société s'est en effet servie des "Like" des usagers de Facebook pour profiler les utilisateurs. En faisant appel à un ensemble de statisticiens et de spécialistes de psychométrie, la société a puisé dans le réseau social pour identifier les habitudes et préférences culturelles, sociales, religieuses des utilisateurs.

Mais l'application développée par Aleksandr Kogan, psychologue à l'université de Cambridge, a profité de l'accès au profil de ces utilisateurs pour atteindre ceux de leurs amis Facebook. Au total, les données de 50 millions de comptes ont ainsi pu être collectées. Ils ont ainsi construit des modèles pour exploiter ces connaissances, et les profiler aussi finement que possible. Facebook a dès lors réclamé la suppression de ces données, et bien que Cambridge Analytica assure que ça soit fait, le doute subsiste encore.

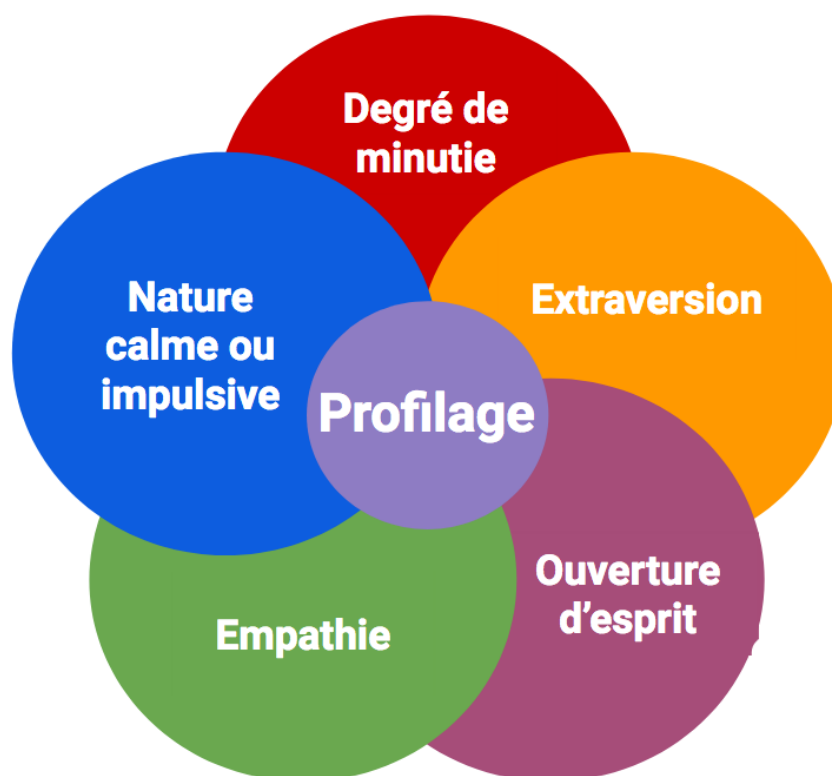


FIGURE 2 – Modèle de profilage des électeurs prenant en compte 5 critères

Une fois les utilisateurs classés selon les différents critères ci-dessus, appelés "Big 5", la société britannique va pouvoir cibler ses messages en fonction du résultat de la classification selon les objectifs de l'équipe de campagne (faire changer d'avis, conforter dans le vote, conquérir le vote). C'est une façon de s'adresser individuellement aux citoyens hésitants pour les persuader que le vote pour Trump est une nécessité.

L'entreprise utilise pour ceci de nombreuses informations dont elle dispose dans sa base sur chaque électeur : date et contenu du dernier vote, type de voiture, loisirs sur Internet, valeurs, etc.

### 1.3 Le Gerrymandering

Le *Gerrymandering*, qui signifie un **découpage électoral partisan**, est un terme qui désigne le découpage des circonscriptions électorales, ceci ayant pour but de donner l'avantage à un parti (ou un candidat ou un groupe donné).

La stratégie consiste principalement à découper les districts électoraux de manière à regrouper le nombre de votes de ceux qui sont perçus comme opposants, à l'intérieur d'un nombre restreint de districts où le parti au pouvoir va perdre dans une forte proportion, mais où il va gagner par de petites majorités ailleurs, dans un plus grand nombre de districts.

Considérons une région composée de 50 électeurs dont les votes se répartissent entre deux partis politiques, 20 votent pour le parti rouge, 30 votent pour le parti bleu. Considérons

une région composée de 50 électeurs dont les votes se répartissent entre deux partis politiques, 20 votent pour le parti rouge, 30 votent pour le parti bleu.

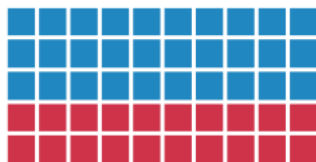


FIGURE 3 – La région à découper

Supposons qu'il va nous falloir scinder cet état en 5 circonscriptions, et que le parti qui a la majorité dans une circonscription gagne un siège au sénat. Nous pouvons proposer un premier découpage ci-dessous :

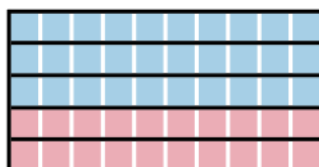


FIGURE 4 – Répartition juste

Ce découpage est juste : les rouges sont majoritaires dans 2 circonscriptions et les bleus dans 3. Cela respecte le nombre total de vote pour chaque parti. Cependant, on remarque que géographiquement, les découpages sont très étendus.

On peut modifier les circonscriptions et réaliser une séparation qui donnera des résultats non représentatifs :

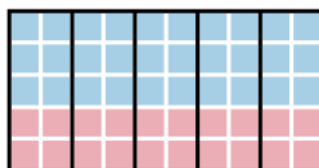


FIGURE 5 – Répartition non représentative

En utilisant cette répartition, on obtient 5 régions en faveur des bleus, ce qui est clairement non représentatif.

Finalement, il est possible de biaiser entièrement les résultats :

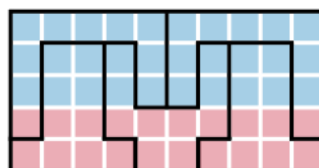


FIGURE 6 – Répartition biaisée

Le Gerrymandering n'est pas une pratique créée par le Big Data, comme nous avons pu le voir dans les sections précédentes. Cependant l'accès à une très grande quantité d'informations a permis au Gerrymandering d'agir sur une toute nouvelle échelle.

Les données utilisées proviennent du **United States Census Bureau** qui récolte tout types de données sur les citoyens, l'économie, la géographie, issus de différents sondages. Ces données sont alors disponibles en libreaccès suivant donc le principe de l'Open Data, au travers d'une plateforme présentant à la fois les données, les statistiques et certaines analyses.

Du côté des conséquences, on peut aboutir à des aberrations démocratiques où un parti peut ainsi obtenir une proportion de sièges plus importante que la proportion de voix récoltée. Ce fut notamment le cas pour les élections américaines de la Chambre des représentants de 2016 où le parti républicain s'est vu attribuer 241 sièges sur les 435 du Congrès, soit un taux d'occupation de 55,4% pour une proportion des voix de l'ordre de 49,1%.

Les récentes élections de mi-mandat de la Chambre ont d'autant plus représenté un levier stratégique cette année puisque les gagnants de cette élections auront le pouvoir en 2020 de redessiner à leur avantage les circonscriptions électorales, ce qui a lieu tous les dix ans et qui aura donc un impact majeur sur la situation politique américaine des dix prochaines années.

D'un point de vue légal, le *Voting right Act - section 2* stipule que le redistricting ne doit pas faire en sorte de diluer le vote des minorités, ce qui revient à insister sur la création de districts majoritairement minoritaires. En effet, les minorités ne doivent pas être noyées dans les votes majoritaires. Pour entrer en ligne de compte dans cette section de loi, les minorités doivent montrer un comportement politique cohésif et être relativement lésées par un vote majoritairement caucasien.

## 2 L'utilisation du Big Data par les politiques une fois au pouvoir

### 2.1 Application aux services pour les citoyens

Les données et leur traitement peuvent avoir d'importantes applications pour les citoyens. En effet, si on prend l'exemple de la santé, grâce à des croisements de données démographiques et médicales, on peut se rendre compte de différentes tendances comme le fait que la crise des opioïdes face rage (avec de nombreux morts) plus particulièrement dans les zone rurales où le revenu est faible et le taux de chômage est élevé.

D'autres applications existent, avec par exemple l'utilisation de statistiques liées aux données pour répondre rapidement et efficacement à des catastrophes naturelles comme le typhon Haiyan (dans le but de prédire des problèmes sanitaires et de coordonner les actions à effectuer).



## 2.2 Application aux renseignements secrets

Parmi ces autres applications, beaucoup sont liées à la sécurité du pays : les données massives sur les citoyens vont permettre de savoir et prévenir nombre de problèmes importants, tels que la fraude, le cyber-crime, ou encore le financement (illégal) de groupes terroristes et gouvernements étrangers.

# 3 État de l'art législatif

## 3.1 Droit des citoyens américains

Tout d'abord, on peut constater que la première couche de protection des données personnelles des citoyens américains se trouve au niveau de la constitution. L'amendement numéro 4 en particulier dont la cour suprême a elle-même déclaré que fonction était de "protéger la vie privée et la dignité contre l'intrusion de l'état". Cependant, ce dernier évoquant des "attentes raisonnables en matière de protection de la vie privée", on comprend qu'il soit grandement possible de jouer sur les mots pour éviter l'application de ce décret.

En juin 2015, le *Freedom act* a été voté par le Congrès Américain. Ce texte de loi qui est censé mettre un terme à la collecte des données massive par la NSA (Agence Nationale de Sécurité).

L'organisation des Etats-Unis en différents états fédéraux rend la législation en matière de vie privée très complexe. En effet, il s'agit d'une mosaïque de textes qui s'accumulent et peuvent parfois se contredire. Cette situation empêche souvent d'enquêter et de punir les abus. A titre informatif, la Californie est l'Etat le plus à la pointe en matière de vie privée, elle a en effet adopté récemment une loi très proche du RGPD qui entrera en vigueur en 2020.

En matière de loi emblématique, on peut citer l'*Online Privacy Act* qui est une loi californienne entrée en vigueur en 2003 ; c'est la première loi des Etats-Unis qui oblige les concepteurs d'un site web à afficher clairement la politique de confidentialité, et laisser la possibilité à l'utilisateur de consulter, modifier et supprimer les données qui le concernent. A l'heure actuelle, la FTC (Federal Trade Commission) est en charge de la gestion des litiges en matière de données personnelles mais son pouvoir est limité car une dizaine de personnes seulement est employée pour gérer ces litiges.



FIGURE 7 – Sceau de la Federal Trade Commission (FTC)

Le 10 novembre 2019, deux élues démocrates de la chambre des représentants ont déposé une proposition de loi exigeant l'instauration d'un organe de régulation de l'accès aux données personnelles, équivalent à la CNIL. Pour l'instant aucun organe de ce type n'existe, c'est la conséquence d'une philosophie ultra-libérale qui priorise la liberté sur le droit à la protection des données.

L'agence créée s'appellerait, si le texte est adopté, DPA (Digital Privacy Agency). Son rôle est de garantir aux Américains le contrôle de leurs données propres. Elle emploierait environ 1600 personnes. Le texte prévoit également de responsabiliser les entreprises sur l'exploitation de ces données et d'obtenir une surveillance du gouvernement stricte mais juste.

Le texte prévoit également d'élargir l'application de l'*Online Privacy Act* à toute entité qui collecte, traite ou conserve intentionnellement des informations personnelles et transmet des informations personnelles sur un réseau électronique, ce qui signifie donc qu'il pourrait s'appliquer à des collectivités publiques et ainsi, au cas du Gerrymandering. L'objectif le plus ambitieux de ce texte est enfin d'uniformiser la législation sur la vie privée à l'ensemble des Etats-Unis.

Cependant, il est nécessaire de rappeler que ce n'est pas la première fois qu'un tel texte a été soumis à l'adoption, pour l'instant ça a été un échec à chaque fois. On peut expliquer ceci par des conceptions ambiguës sur ce sujet. En effet, le droit à la protection des données personnelles n'est pas directement assimilable au droit à la vie privée qui fait partie de la Constitution Américaine. De plus, la tradition historique veut que la liberté prime presque toujours sur le droit à la vie privée, comme évoqué plus haut.

Dans le cas du Gerrymandering, d'après le *Voting right Act - section 2*, le redistricting ne doit pas faire en sorte de diluer le vote des minorités, ce qui revient à insister sur la création de districts majoritairement minoritaires. En effet, les minorités ne doivent pas être noyées dans les votes majoritaires. Pour entrer en ligne de compte dans cette section de loi, les minorités doivent montrer un comportement politique cohésif et être relativement lésées par un vote majoritairement caucasien.

La section 5 de cette loi va concerner les lieux plus spécifiques, qui ont un passé lourd en matière de discrimination raciale. Dans ce cas uniquement, une soumission du plan au gouvernement fédéral est requise pour validation.

Le *Voting Right Act* prévoit aussi la contiguïté des districts, la compacité, l'imbrication des plus petits quartiers dans les plus grands, et enfin un dernier critère plutôt nouveau qui concerne la préservation des communautés d'intérêt. Il s'agit de préserver

les "groupes de personnes" qui partagent des priorités et des intérêts similaires, culturels, sociaux, religieux ou politiques.

Ces critères sont nombreux et exigeants et ils sont actuellement largement relativisés les uns par rapport aux autres selon les états. En effet, l'organisation en états fédéraux implique une hiérarchisation différente selon l'état de ces critères. Le facteur commun à tous les états reste toutefois que le facteur prédominant est l'égalité de population entre les quartiers.

### 3.2 Futures règles des campagnes par rapport au Big Data

Fin 2018, l'autorité britannique de protection des données a bouclé une enquête de près de 18 mois sur la manière dont les campagnes politiques utilisent les données personnelles collectées sur Internet pour cibler les électeurs. Et l'implication de différents acteurs est mise en lumière.

Dans un rapport d'enquête de 112 pages, publié le 6 novembre 2018, le bureau du commissaire britannique à l'information a expliqué comment les campagnes politiques achetaient des listes marketing et d'autres informations sur le mode de vie des consommateurs à des courtiers en données "sans vérification préalable suffisante" et avec peu de la protection des consommateurs.

L'agence a identifié 172 organisations impliquées dans le commerce de données, dont 30 sont au centre de ce rapport. Le bureau du commissaire a également saisi 85 équipements, dont des serveurs, 22 documents et 700 téraoctets de données, ce qui représente environ 52,5 milliards de pages.

"La confiance dans l'intégrité de nos processus démocratiques risque d'être perturbée, car le citoyen moyen n'a aucune idée de ce qui se passe en coulisses" souligne Denham dans un article de blog. "Cela doit changer. Les gens ne peuvent faire de choix véritablement informés concernant leur vote que s'ils sont sûrs que ces décisions n'ont pas été indûment influencées."

Il est ainsi souligné que les électeurs ne sont pas des cibles publicitaires comme les autres et les entreprises de l'écosystème des données doivent prendre leurs responsabilités, au même titre que les partis politiques tentés d'exploiter ces technologies à des fins électorales.

Les gouvernements se disent ainsi prêts à légiférer pour introduire un code de pratique afin d'encadrer l'utilisation des données personnelles dans les campagnes.

## Veille

Cette veille technologique sur l'utilisation du Big Data dans la politique américaine a été réalisée à l'aide de plusieurs outils et d'une méthodologie bien précise, axée ainsi :

- Recherche et collecte de l'information
- Analyse de l'information
- Distribution/diffusion des résultats

### Recherche et collecte de l'information

Ainsi, la première étape a été d'identifier les mots clés de mon sujet et de préparer leur forme pour optimiser ma recherche initiale : **Big Data**, **Améri\***, **politique** ont alors été les premières recherches effectuées pour en apprendre un minimum sur le sujet.

De cette première recherche, ont pu découler plusieurs sous recherches liées à la chronologie de la dernière décennie sur le sujet (voir la première partie de ce rapport), et donc d'autres mots clés : **Obama**, **Trump**, **Cambridge Analytica**, **Gerrymandering** toujours liés aux premiers.

Ces nombreuses recherches ont ainsi pu être mises en oeuvres sur **Google Actualités** notamment, ainsi que **LinkedIn** et **Twitter** dans une moindre mesure. En effet, le sujet étudié est un sujet d'informations générales et d'actualité plutôt qu'un sujet de recherche très technique (ce qui exclue les sites du genre **Google Scholar** ou **Theses.fr** par exemple), et demande donc une recherche dans les outils cités plus haut qui sont idéals pour un tel sujet.

J'ai alors pu grâce à cette première itération de recherche regrouper une liste de sites web qui étaient liés de près ou de loin à mon sujet et ensuite passer à la deuxième étape de recherche d'informations, l'automatisation : Grâce à l'outil **Feedbro**, j'ai mis en place la collecte d'informations sur mon sujet via un flux RSS sur chaque site intéressant (LeBigData.fr, LExpress.fr, The Daily Beast, ...). J'ai également utilisé **Google Alerts** en parallèle pour être informé de façon plus globale sur les actualités du Big Data en politique en général.

### Analyse et distribution de l'information

L'analyse des informations récoltées grâce aux moyens mis en place précédemment a été faite manuellement de façon systématique. Les documents/pages web ainsi triées et choisies (selon leur *pertinence*, leur *degré de précision* et leur *unicité*) ont été ensuite regroupées sur l'outil de curation **Wakelet**, pour ainsi être détaillés, commentés et partagés au public intéressé.

Le processus complet de veille peut ainsi être résumé :

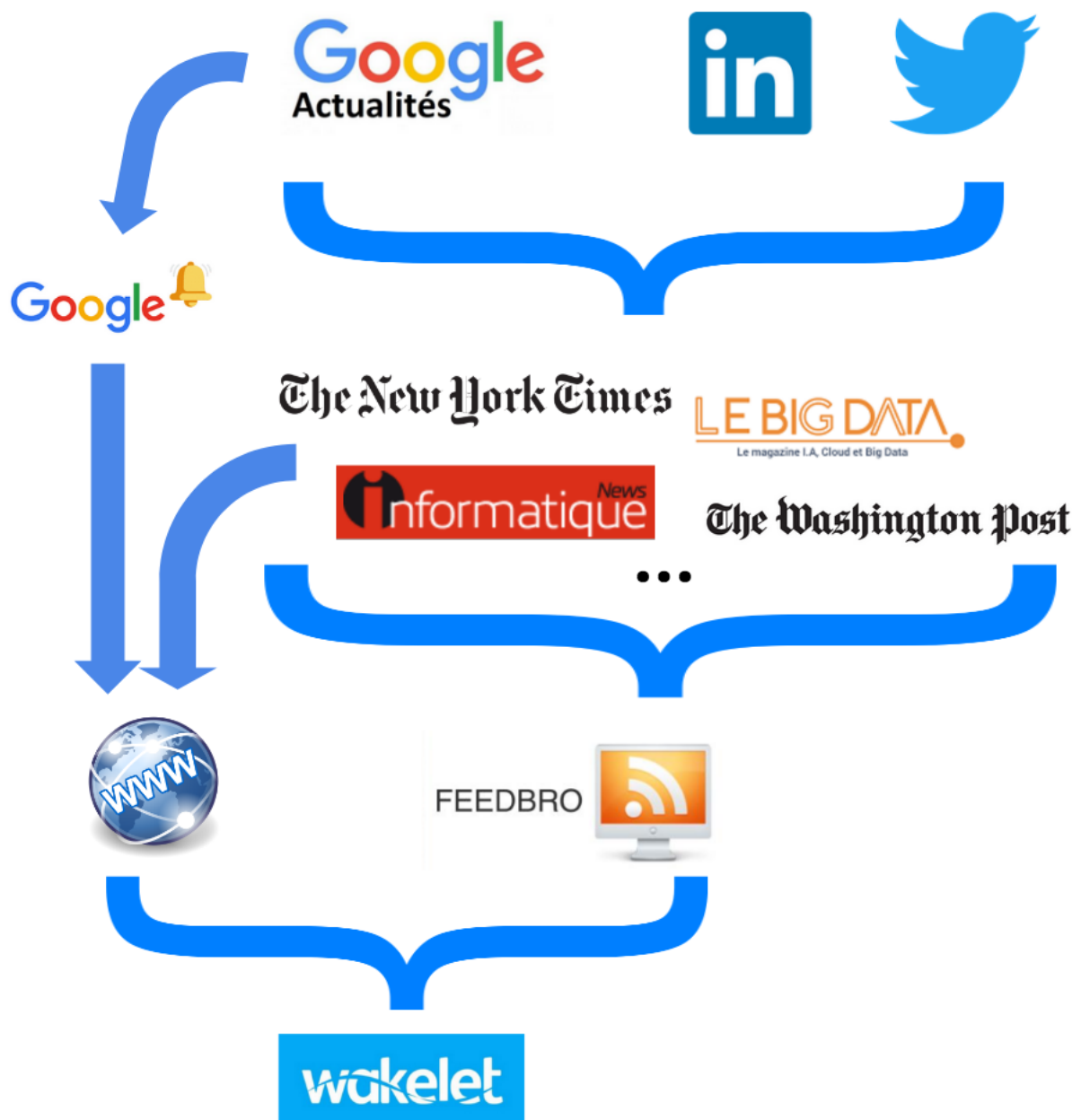


FIGURE 8 – Processus de veille

Pour rappel, la synthèse sous Wakelet peut être trouvée ici : <https://wke.lt/w/s/zNIJnb>

## Bibliographie

[https://www.sas.com/en\\_us/insights/articles/big-data/big-data-government.html](https://www.sas.com/en_us/insights/articles/big-data/big-data-government.html)  
<https://www.lexpress.fr/actualite/monde/amerique-nord/election-de-trump-le-hold-up-1993257.html>  
[https://www.lemonde.fr/pixels/article/2018/03/21/quelle-a-ete-l-importance-reelle-5274423\\_4408996.html](https://www.lemonde.fr/pixels/article/2018/03/21/quelle-a-ete-l-importance-reelle-5274423_4408996.html)  
<https://www.lebigdata.fr/trump-big-data-0911>  
<https://www.zdnet.fr/actualites/les-derives-du-couple-infernal-big-data-et-politique-htm>  
<https://www.thedailybeast.com/supreme-court-unleashes-new-era-of-extreme-gerrymandering>  
<https://www.developpez.com/actu/249347/Cloud-Act-la-loi-americaine-qui-donne-aux-US>  
<https://www.informatiquenews.fr/dossier-big-data-25-barack-obama-premier-president>