

Big Data

Panorama et outils

(Big Data Analytics)

Ricco RAKOTOMALALA
Université Lumière Lyon 2

Junior EWEBBI - Conférence Big Data - 11 octobre 2017



- Formation en économétrie (statistique, modélisation)
- Thèse de doctorat en Machine Learning ([Apprentissage statistique](#))
- Enseignant chercheur, en poste à l'Université Lumière Lyon 2
- Spécialité : statistique, data mining et ses applications, informatique - [Data Science](#)
- Responsable du Master [SISE](#) (Statistique et Informatique pour la Science des données)
- « Père » des logiciels libres [SIPINA v.3](#) et [TANAGRA](#)
- Auteur d'une dizaine d'[ouvrages libres](#)
- Auteur de plus de 500 [supports de cours](#) et tutoriels en [français](#) et en [anglais](#)
- [650 visites par jour](#) depuis le 1^{er} février 2008 (Compteur Google Analytics)



Plan

1. Big Data
2. Valorisation des Big Data – Big Data Analytics
3. Nouvelles applications des Big Data
4. R et Python – Etudes de cas, démos
5. Bibliographie



BIG DATA

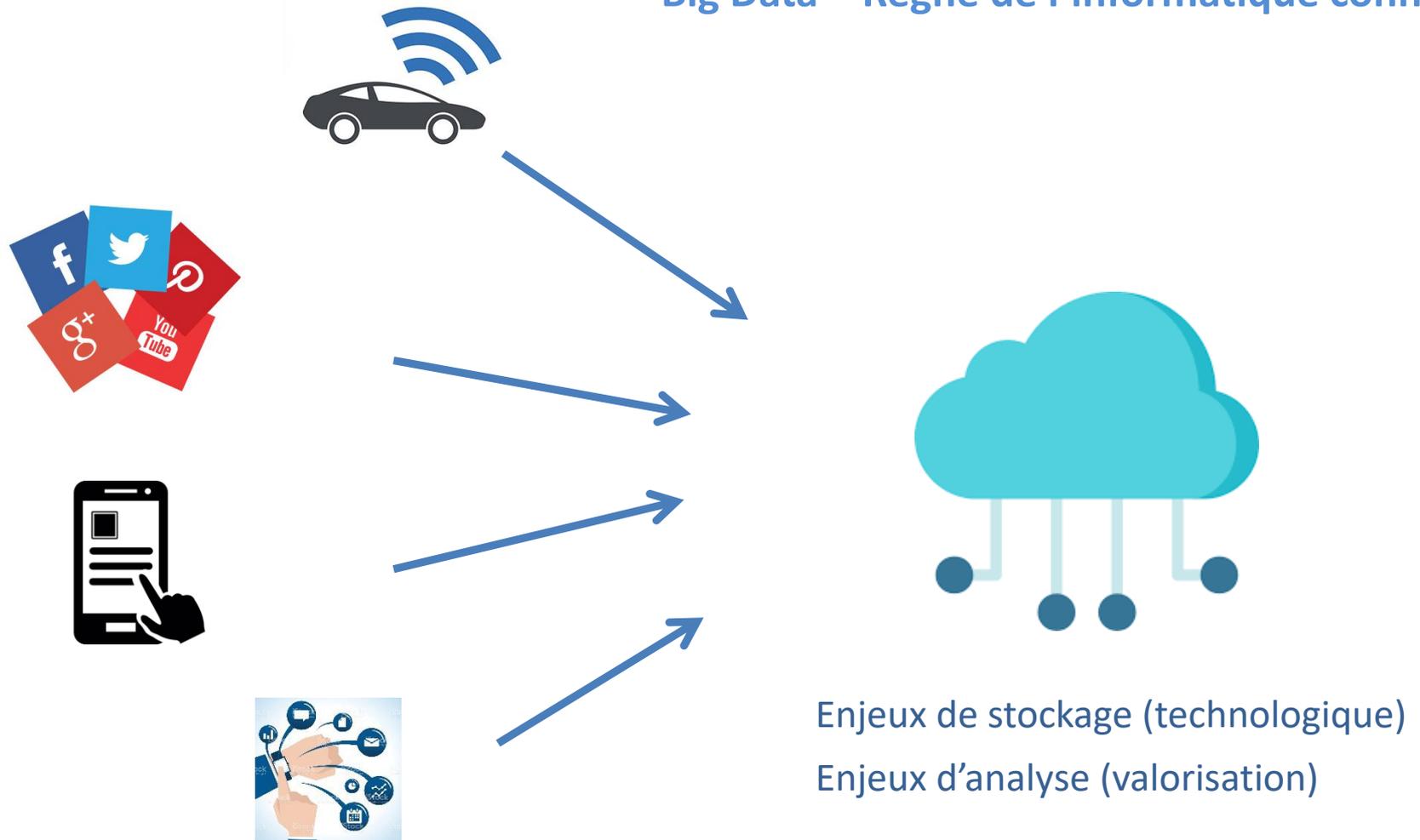
Tout le monde en parle... c'est le terme à la mode (cf. [Google Trends](#))

Tout le monde est persuadé que c'est très important

... mais de quoi il retourne exactement ?

... quel rapport avec la Data Science / Big Data Analytics ?





Variété des sources d'information, du type, des formats, fréquence des mises à jour, énorme volumétrie.



VOLUME

Outils de recueil de données de plus en plus présents, dans les installations scientifiques, mais aussi et surtout dans notre vie de tous les jours (ex. cookies, GPS, réseaux sociaux [ex. lien « like » - « profils »], cartes de fidélité, les simulations en ligne sur certains sites de prêts ou d'assurance, etc.).
Il faut pouvoir les stocker et pouvoir les traiter (rapidement, efficacement) !

VARIETE

Sources, formes et des formats très différents, structurées ou non-structurées : on parle également de données complexes (ex. texte en provenance du web, images, liste d'achats, données de géolocalisation, etc.).
Il faut les traiter conjointement !

VELOCITE

Mises à jour fréquentes, données arrivant en flux, obsolescence rapide de certaines données... nécessité d'analyses en quasi temps réel (ex. détection / prévention des défaillances, gestion de file d'attente)
Il faut les traiter fréquemment (et/ou tenir compte du facteur d'obsolescence) !



Cloud computing

Le cloud computing ... est l'exploitation de la puissance de calcul ou de stockage de serveurs informatiques distants par l'intermédiaire d'un réseau, généralement internet. Ces serveurs sont loués à la demande, le plus souvent par tranche d'utilisation selon des critères techniques (puissance, bande passante, etc.) mais également au forfait ([Wikipédia](#)). Ex. Amazon Web Services, Microsoft Azure,... [Azure Machine Learning](#).

Plateformes big data

L'architecture d'un environnement informatique ou d'un réseau est dite distribuée quand toutes les ressources ne se trouvent pas au même endroit ou sur la même machine.... Les architectures distribuées reposent sur la possibilité d'utiliser des objets qui s'exécutent sur des machines réparties sur le réseau et communiquent par messages au travers du réseau ([Wikipédia](#)). (Ex. Hadoop, Spark). Savoir programmer sous ces environnements devient un enjeu fort (cf. [tutoriels](#)).

Bases NOSQL

En informatique et en bases de données, NoSQL désigne une famille de systèmes de gestion de base de données (SGBD) qui s'écarte du paradigme classique des bases relationnelles. L'explicitation du terme la plus populaire de l'acronyme est Not Only SQL ([Wikipédia](#)). L'idée est d'acquiescer plus de souplesse pour gérer notamment la variété des données (ex. [MongoDB](#), orienté document ; [Neo4j](#), orienté graphe, etc.)



Big Data Analytics

Les Big Data Analytics désignent le processus de collecte, d'organisation et d'analyse de grands ensembles de données (Big Data) afin de découvrir de nouveaux modèles et en tirer des informations utiles. Les Big Data Analytics veulent fondamentalement découvrir la connaissance provenant de l'analyse des données ([Le Big Data](#)).

Aujourd'hui une priorité

Anne Lauvergeon et al., « Ambition 7 : La **valorisation** des données massives (Big Data) », in « [Un principe et sept ambitions pour l'innovation - Rapport de la commission Innovation 2030](#) », Octobre 2013 [[Rapport annoté](#)].



Big Data Analytics

De la Statistique à la Data Science

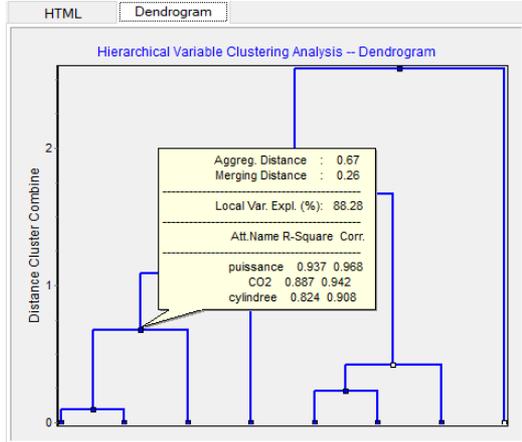
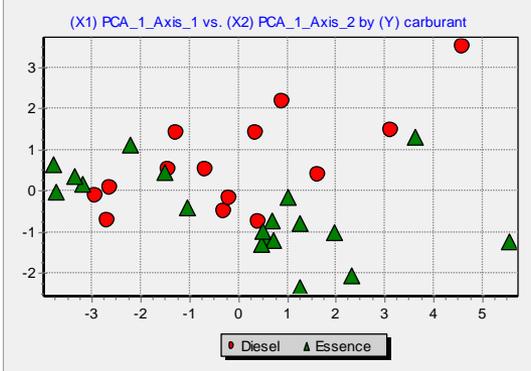
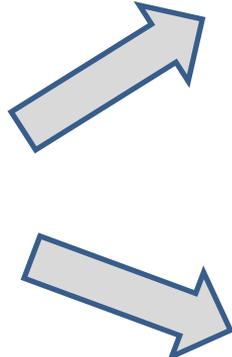


Application des techniques de modélisation et de statistique

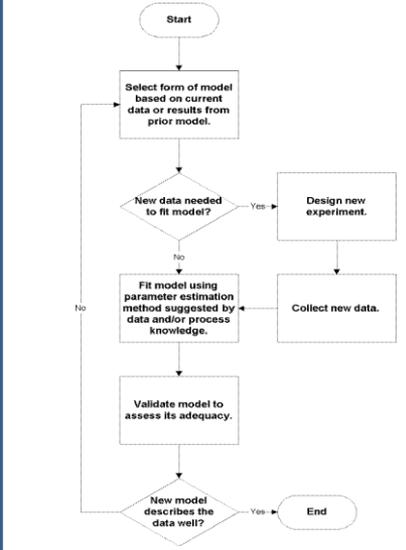


Les données sont spécifiquement recueillies à des fins d'étude (ex. enquête, etc.)

- Bonne qualité souvent
- Faible volumétrie



Modeling Steps (NIST – e-Handbook of Statistical Methods)

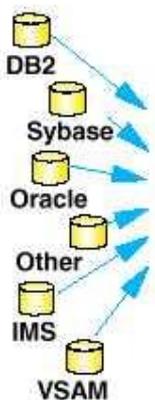


Evolution clé. Les données des entreprises sont organisées et stockées de manière à ce que nous puissions mener des analyses.

Construire une Infrastructure d'Information Intelligente pour l'Entreprise

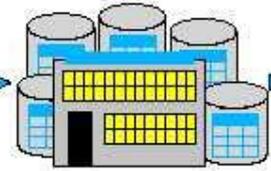
Bases décisionnelles

Données Opérationnelles (Operational Data)



Dessiner
Extraire
Nettoyer
Correler
Charger

Entrepôt de Données (Data Warehouse)



Filterer
Résumer
Distribuer

(Data Mart)



Quelles seront les tendances salariales la prochaine année?



Comment réduire les coûts de 20% ?



Quel est le meilleur canal de distribution pour ces produits ?

(Data Mining)

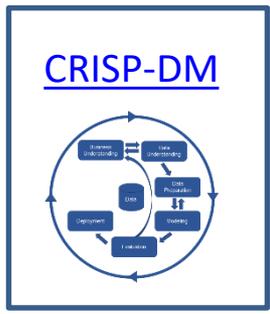
Production

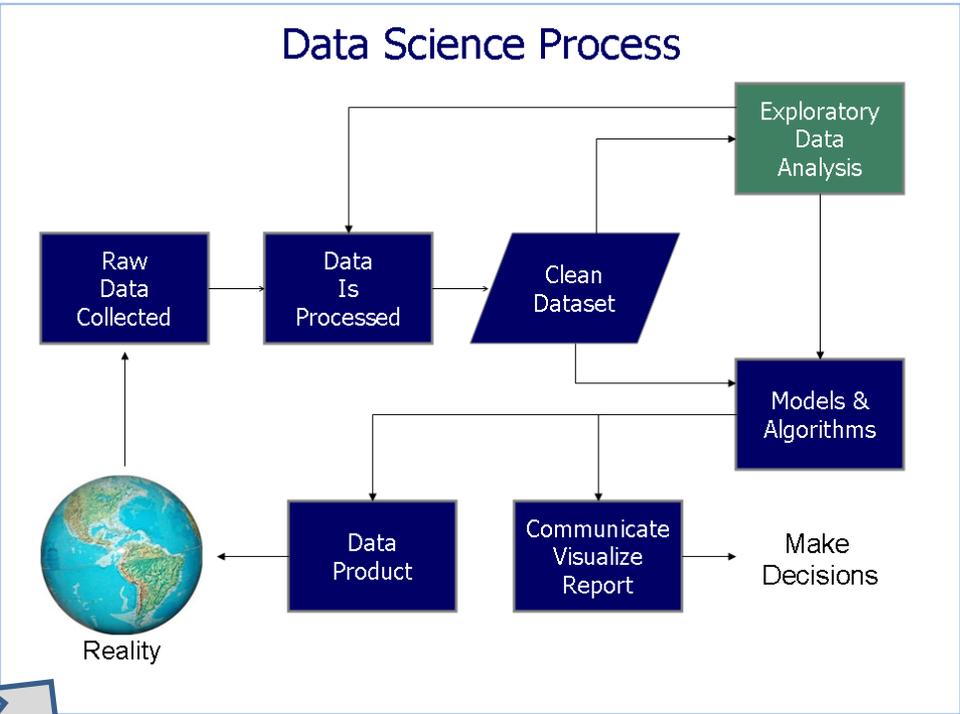
- orientation service (ventes, comptabilité, marketing...)
- volatiles

Stockage

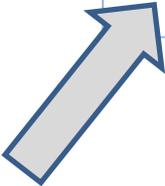
- orientation analyse
- historisées
- non-volatiles

La volumétrie devient (déjà) une problématique cruciale





(Wikipedia – [Data Science](#))



Démultiplication des sources d'information.

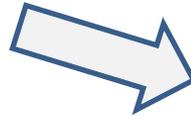
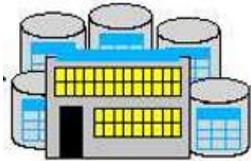
Maître mot : pluridisciplinarité

A central blue circle labeled 'Data Science' is surrounded by eight smaller blue circles, each representing a related field: Data Engineering, Scientific Method, Math, Statistics, Advanced Computing, Visualization, Interdisciplinary Mindset, and Domain Expertise. Arrows point from each of these surrounding circles towards the central 'Data Science' circle.



Big Data Analytics

Données internes à l'entreprise



Pour rendre les analyses plus performantes



Données externes à l'entreprise

La vague « **OPEN DATA** » va amplifier le déluge (des données)... et les attentes en termes d'analyse ([Enjeux de l'Open Data](#))



Nouvelles opportunités d'analyse

Text mining, Web mining, etc.



Services financiers

Scoring de l'emprunteur - <http://www.cbanque.com/credit/scoring-etude-dossier.php#>

« Crédit score » régit notre vie – Le « [diktat de la solvabilité](#) »

Y compris notre [vie amoureuse](#)

Grande distribution

Nous reste-t-il encore des [secrets](#) ?

Petite histoire du [père américain](#)

Cartes de fidélité - Renouvellement des informations au fil des années

Assurances

Scoring – Détermination des primes d'assurance (Amaguiz, Direct Assurances, etc.)

Assurance auto : les [conductrices](#) payeront plus cher

Sport

Dossier du Journal l'Equipe – La « data révolution » (<http://www.lequipe.fr/explore/la-data-revolution/>)

Tous les sports s'y mettent : le [foot](#), le [tennis](#), etc.

Autres

Les constructeurs automobiles s'y mettent ([Carburant de demain](#), [analyse prédictive](#), ...)

Fraude aux allocs ([cibler les contrôles...](#)), fraude à la carte bancaire ([transactions suspectes...](#))

Présidentielles USA (cibler les électeurs et les [donateurs...](#))

Recrutement et gestion des ressources humaines ([programmes informatiques](#), [drh](#), ...)



Nouveaux usages (1)

Filtrage collaboratif et systèmes de recommandation

The screenshot shows the Amazon.fr product page for 'Gil Jourdan : L'Intégrale 1' by Maurice Tillieux. The page includes the Amazon logo, search bar, navigation menu, and product details. A red box highlights the 'Produits fréquemment achetés ensemble' section, which shows three related items with a total price of EUR 72,00. A callout box points to this section with the text 'Recommandation basée sur les transactions.'

Recommandation basée sur les transactions.

Recommandation basée sur les utilisateurs (clients).

Les clients ayant acheté cet article ont également acheté

A grid of three recommended products: 'Gil Jourdan - L'Intégrale - tome 2 - Gil Jourdan 2 (intégrale) 1960 - 1963', 'Gil Jourdan : L'Intégrale 3', and 'Gil Jourdan - L'Intégrale - tome 4 - Gil Jourdan 4 (intégrale) 1970 - 1979'. Each item shows its cover, title, author, and price. A red box highlights the first two items, with a callout box pointing to them that says 'Evaluations des produits'.

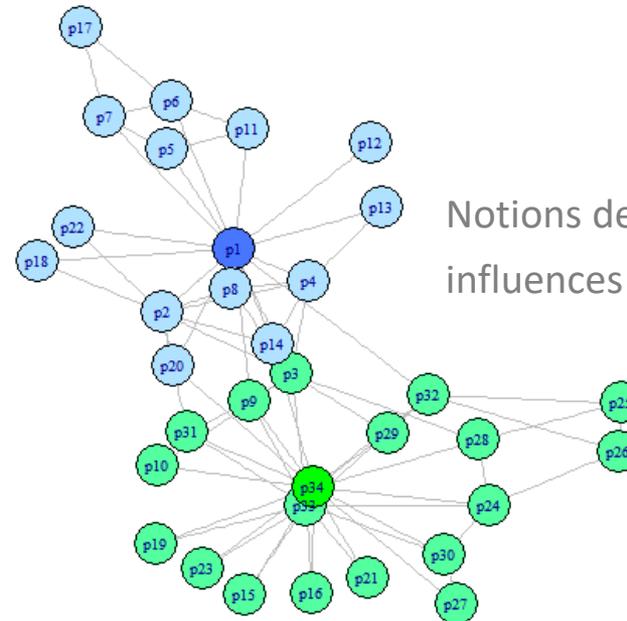
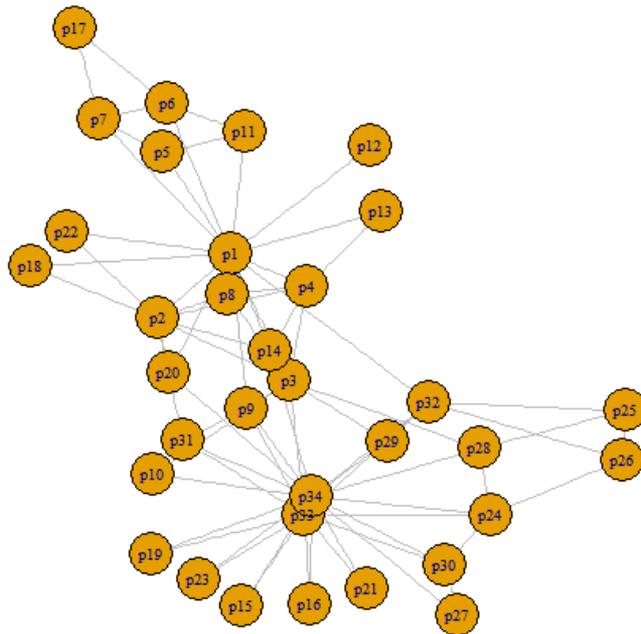
Evaluations des produits

Commentaires (text mining) + Evaluation des commentaires

A screenshot of a 5-star review by Jack Spads from July 17, 2009. The review text is: 'Que dire sinon que l'initiative de regrouper les aventures de Gil Jourdan par tome avec des archives est bonne; j'ai de nouveau 15 ans et si ela peut donner envie aux gamins actuels de lire ces polars BD et humoristiques, c'est très bien. J'avais déjà les albums en totalité depuis plus de 30 ans. Il y a une ambiance magique dans certains épisodes: mon préféré: " les cargos du crépuscule!" sublime! Tillieux est un maître: ces scenariis et découpage des vignettes, ça coule de source; et puis la maitrise du dessin: moins naïf qu'on ne croit: vérifiez les ombres, les perspectives: c'est d'une justesse diabolique. Un grand monsieur cela va s'en dire.' Below the review are buttons for 'Commentaire', 'Signaler un abus', and a 'Oui/Non' feedback box. A red box highlights the entire review section, with a callout box pointing to it that says 'Commentaires (text mining) + Evaluation des commentaires'.



Détection de communautés dans les réseaux sociaux



Notions de centralité,
influences et communautés

Les idées sont anciennes mais ont connu un regain d'intérêt extraordinaire avec l'apparition des médias sociaux ([Fergusson](#), [Paris Plage](#),...).

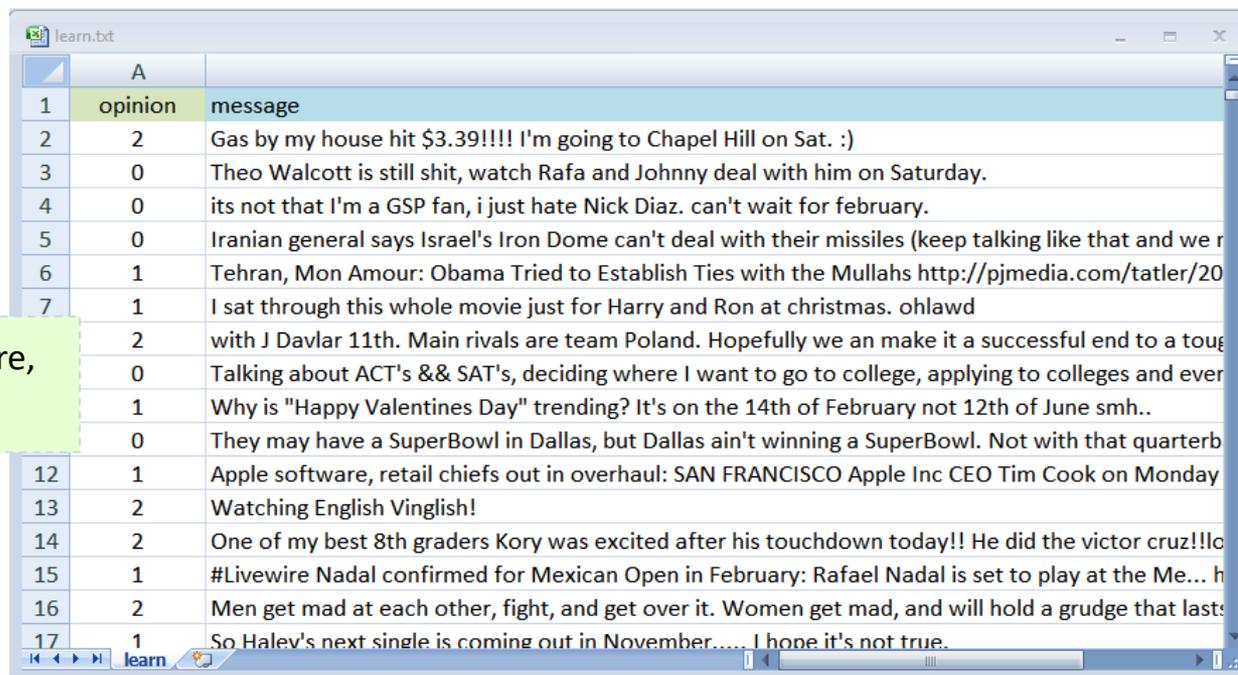


Nouveaux usages (3)

Analyse des opinions (sentiments, approbation, désapprobation, etc.). Ex. Twitter

<http://www.latribune.fr/opinions/tribunes/20140213trib000815265/les-cles-d-une-veritable-analyse-semantique-sur-twitter.html>

<http://www.france24.com/fr/20170428-france-presidentielle-big-data-erreur-fillon-le-pen-macron-predict-president-algorithme-twi>



	A	
1	opinion	message
2	2	Gas by my house hit \$3.39!!!! I'm going to Chapel Hill on Sat. :)
3	0	Theo Walcott is still shit, watch Rafa and Johnny deal with him on Saturday.
4	0	its not that I'm a GSP fan, i just hate Nick Diaz. can't wait for february.
5	0	Iranian general says Israel's Iron Dome can't deal with their missiles (keep talking like that and we r
6	1	Tehran, Mon Amour: Obama Tried to Establish Ties with the Mullahs http://pjmedia.com/tatler/20
7	1	I sat through this whole movie just for Harry and Ron at christmas. ohlawd
	2	with J Davlar 11th. Main rivals are team Poland. Hopefully we an make it a successful end to a toug
	0	Talking about ACT's && SAT's, deciding where I want to go to college, applying to colleges and ever
	1	Why is "Happy Valentines Day" trending? It's on the 14th of February not 12th of June smh..
	0	They may have a SuperBowl in Dallas, but Dallas ain't winning a SuperBowl. Not with that quarterb
12	1	Apple software, retail chiefs out in overhaul: SAN FRANCISCO Apple Inc CEO Tim Cook on Monday
13	2	Watching English Vinglish!
14	2	One of my best 8th graders Kory was excited after his touchdown today!! He did the victor cruz!!!o
15	1	#Livewire Nadal confirmed for Mexican Open in February: Rafael Nadal is set to play at the Me... h
16	2	Men get mad at each other, fight, and get over it. Women get mad, and will hold a grudge that lasts
17	1	So Halev's next single is coming out in November..... I hope it's not true.

(0 : négative, 1 : neutre,
2 : positive)



C'est du text mining avec un cadre et des finalités particulières !!!

(longueurs des textes contraintes et homogènes, mises à jour très fréquentes, etc.)



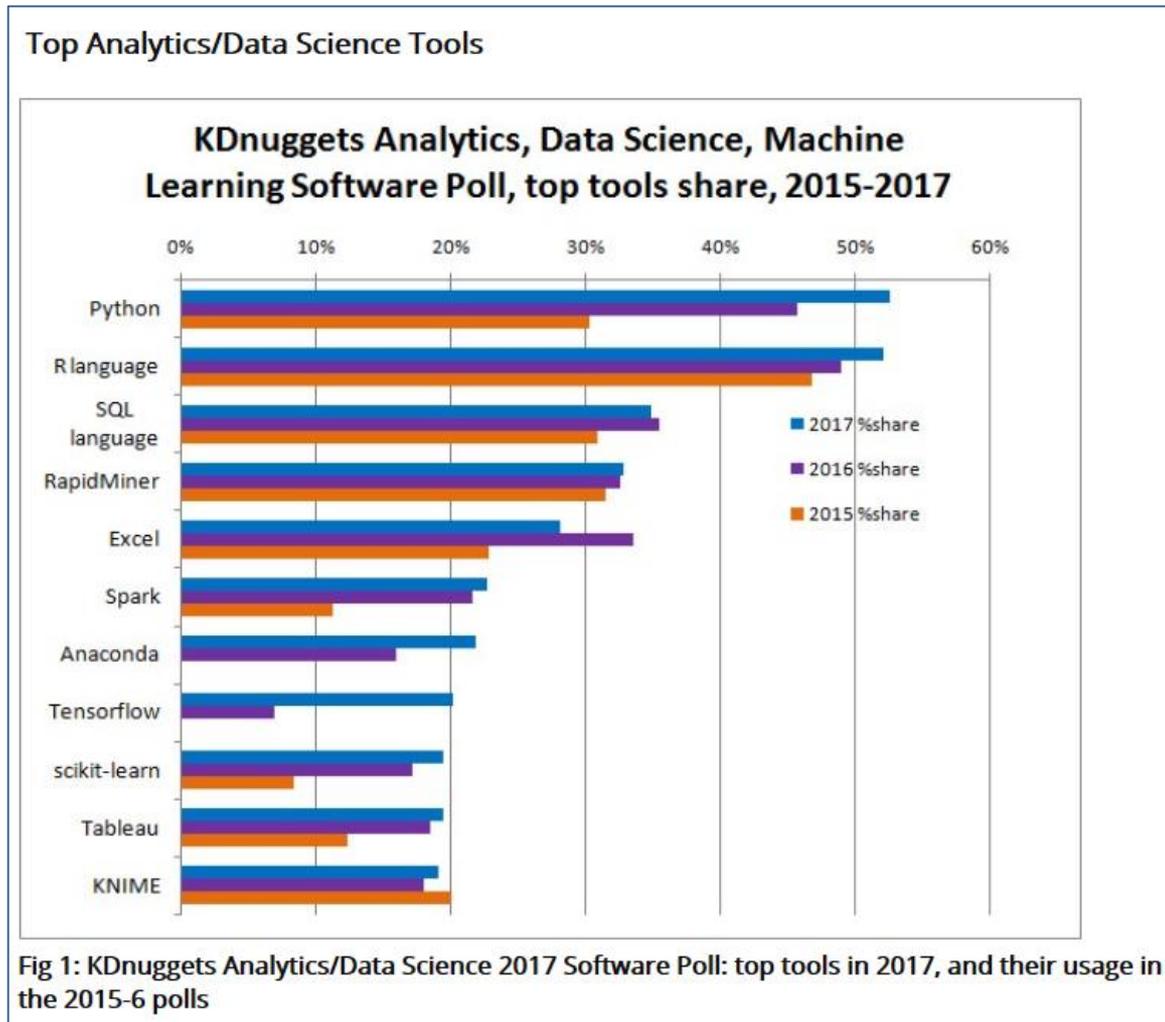
[Tweet Sentiment Visualization](#)



Démos

Avec R et Python





Et surtout, ces outils sont de plus en plus présents dans les offres d'emploi en France (<https://www.apec.fr/>)



Quelques projets étudiants

1. Analyse des offres d'emploi – Text mining et application Shiny
2. Reconnaissance faciale (+ [Google Image](#) , [Microsoft Age](#))
3. Reconnaissance musicale
4. Analyse des tweets sur la loi travail (cf. Article [Le Point](#))
5. Détection de communautés (le club de Karaté de Zachary)



Vidéographie



Stéphane Tufféry, « [Les Big Data : une révolution numérique](#) », Les Mardis de l'Espace des Sciences, Rennes, 19 nov. 2013 ; <https://www.youtube.com/watch?v=2EMlg7Voy3Y>

Nicolas Glady, « [Data science & Business Analytics](#) », Essec Business School, 12 mars 2015 (mise en ligne) ; <https://www.youtube.com/watch?v=1ubXgqIHhfw>

Gaëtan Constant (Datalyo), « [Table ronde – Critères de recrutement et métiers dans le domaine du big data](#) », Forum Entreprises 2016 : Université Claude Bernard Lyon 1, 17 nov. 2016 ; https://www.youtube.com/watch?v=x9OV9E_X2HI

