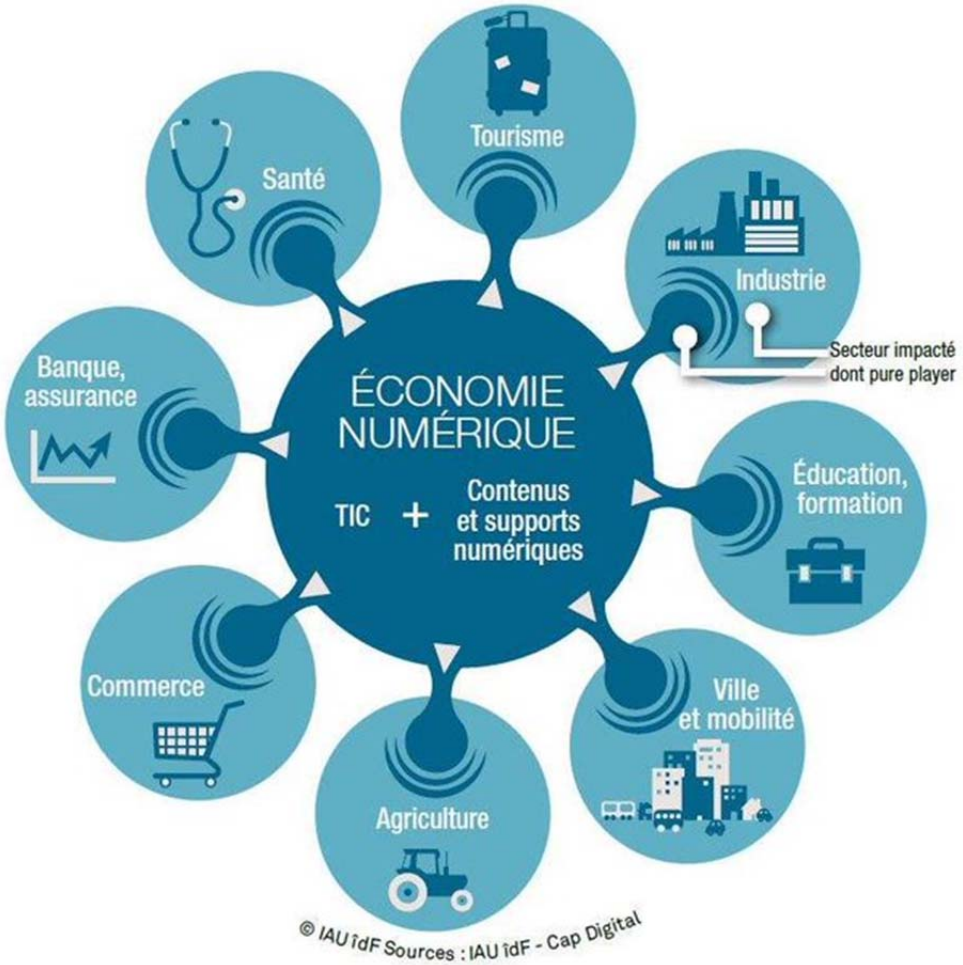
Big data et open data dans le même bateau ? Oui et non !

La notion de big data est abondamment utilisée et mise à toutes les sauces par les médias. Elle correspond essentiellement à la profusion de données numériques existantes et exploitables.

- Le big data évoque souvent **l'exploitation** de données à des fins commerciales, dans des applications marketing et publicitaires. Le champ d'application est beaucoup plus vaste comme l'indique le schéma suivant proposé par l'IAU IDF (Institut d'aménagement et d'urbanisme d'Île-de-France) pour décrire l'économie numérique.



- Celui d'open data correspond assez clairement à la **mise à disposition** de tous de données issues des services des États, des services publics et assimilés. Ce qui n'exclut pas qu'une exploitation commerciale ait lieu à partir de ces données.
  Bien entendu l'open data suppose de disposer de datas, plus ou moins big, structurées, plus ou moins également !

  Pour les puristes de la langue française, on peut proposer une traduction à partir du thème retenu pour les journées de la statistique SFdS 2014 « Données en grande dimension, Big data »

  L'affaire ne date pas d'hier puisque, selon le journal suisse *Le Temps* :

  « Le premier scientifique de données est John Tuckey, un statisticien (1915-2000) qui aurait fait de la science des données une éthique, celle du « Data first ». Plutôt que de partir d'une hypothèse et d'en vérifier la valeur, il aurait montré la nécessité de partir des données et d'en déduire des enseignements. »

Big data

Big data et marketing sont intimement liés ; cette émergence des datas résulte naturellement du développement des TIC (technologies de l'information et de la communication), de capacités de stockage quasi infinies et d'outils de traitement très puissants.
La définition communément admise par les spécialistes retient trois points de base :

- Grand volume de données
- Importante variété de données
- Vitesse de traitement

Auxquels s s'ajoutent 2 autres « V »

- Valeur ajoutée, création de valeur
- Véracité, données fiables (!)

Dans le domaine du marketing trois sources d'approvisionnement en données se complètent et se combinent. Il s'agit de gérer et exploiter les données…

- que l'ont possède,
- que l'on achète,
- que l'obtient.

Ces trois termes correspondent à un acronyme en vogue dans cet univers : le POE, ce qui signifie Paid, Owned, Earned. Quelques mots pour en donner la signification.
Le Paid correspond à des données achetées par l'entité concernée, par exemple des fichiers adresses. Le Owned correspond à ce qu'elle possède, son fichier clients/adhérents et les transactions effectuées avec les membres desdits fichiers. Le Earned est « gagné » sans bourse délier, ce sont par exemple les données récoltées sur les pages de cette entité sur les réseaux sociaux.

Les « traces » laissées par nos usages de l'internet constituent une source d'approvisionnement considérable en data ; parmi elles, rappelons les informations collectées par cookies, favoris, paramètres et préférences de navigation, passage sur les réseaux sociaux, consultation des sites des entreprises…Pour l'instant ces données sont reliées au numéro IP de la machine connectée, mais pas de soucis, Google cherche une solution et développe un programme permettant de probabiliser les caractéristiques de la personne au clavier!
Heureusement pour nous, actuellement les mobiles et les tablettes n'enregistrent pas de cookies, mais…. la parade sera bientôt trouvée avec un nouveau numéro d'identification!

Tous les passages de l'internaute peuvent être enregistrés et traités en temps réel. Le résultat de ces actions conduit à se voir proposer un produit ou un service dans une publicité en ligne correspondant à la recherche que l'on est en train de faire. De même et de façon particulièrement insistante, des bannières de publicité sur ce type de produit s'affichent systématiquement dans les heures et les jours qui suivent la recherche initiale, que l'on ait acheté ou non. Plus sophistiqués sont les algorithmes qui savent abandonner quand un achat a eu lieu!
Tout cela résulte des procédures du big data, comme la vente de la publicité qui arrive sur l'écran, vendue aux enchères en temps réel! Dans le jargon publicitaire, il s'agit de RTB (Real Time Binding).

Les nouveaux enjeux immédiats tiennent aux « objets connectés » :

- Les TV connectées. Elles font le plus souvent double emploi avec la tablette ou le mobile également connectés utilisés simultanément par le téléspectateur. D'autres engins ou dispositifs sont ou seront prochainement connectés, voitures, réfrigérateurs, compteurs et divers produits électroménagers.
- Les wearables, ou objets connectés personnels, telles les montres connectées et autres capteurs « portés » par les individus.

Les estimations les plus fantaisistes circulent quant au nombre d'objets connectés actuels et futurs.

Actuellement estimés entre 10 et 20 milliards dans le monde, les prévisions à 2020 ou 2030 vont jusqu'à 80 milliards. On notera que ces prévisions, d'origine nord américaine pour la plupart, sont toujours calculées à l'échelle mondiale.

L'exemple des compteurs électriques est intéressant. Aux USA les compagnies installent des compteurs connectés qui permettent au fournisseur et au client final de disposer en temps réel de l'information sur les consommations, on imagine aisément le profit que peuvent en tirer les deux parties. En France le compteur « intelligent » Linky d'EDF soulève quelques problèmes compte tenu du fait que la voie de retour vers le consommateur ne semble pas être prévue…
Il est en tout cas très probable que les applications liées au logement vont redonner à la domotique, trop tôt mise en avant, un nouveau contenu.

Lié aux objets connectés, Le Quantified Self est, selon wikipédia, un mouvement qui regroupe les outils, les principes et les méthodes permettant à chaque personne de mesurer ses données personnelles, de les analyser et de les partager. Les outils du Quantified Self peuvent être des capteurs, des applications mobiles ou des applications Web.
En mars 2013, le mouvement Quantified Self en France s'organise autour de MyDataLabs, association centrée sur la donnée personnelle. Les applications médicales sont en première ligne ; le coaching semble également prendre une place importante dans ce domaine. On peut supposer, ou même espérer que ce mouvement mondial deviendra un pare-feu face aux débordements du marketing et de la publicité.

Quels que soient l'origine et le donneur d'ordre, marques commerciales, ONG, associations caritatives… le big data implique l'utilisation de puissants algorithmes développés en interne ou proposés par des prestataires spécialisés.
Le terme d'algorithme est sans doute celui qui revient le plus souvent dans la littérature spécialisée.
Au passage, Il semble que ces activités nécessitant des développements informatiques très lourds, soient une voie de recyclage et/ou d'avenir pour les cerveaux français en rupture avec les activités financières de trading haute fréquence !
Notons par exemple qu'actuellement en Europe, plus de 50 sociétés spécialisées analysent en temps réel ce qui se passe sur les réseaux sociaux.

« En temps réel » est l'expression de plus en plus fréquemment associée au big data. Il s'agit là par exemple d'une réelle révolution pour les marketeurs et les pubicitaires, qui voient leur pré carré envahi par ces démarches quantitatives et les process informatiques qui les sous-tendent.

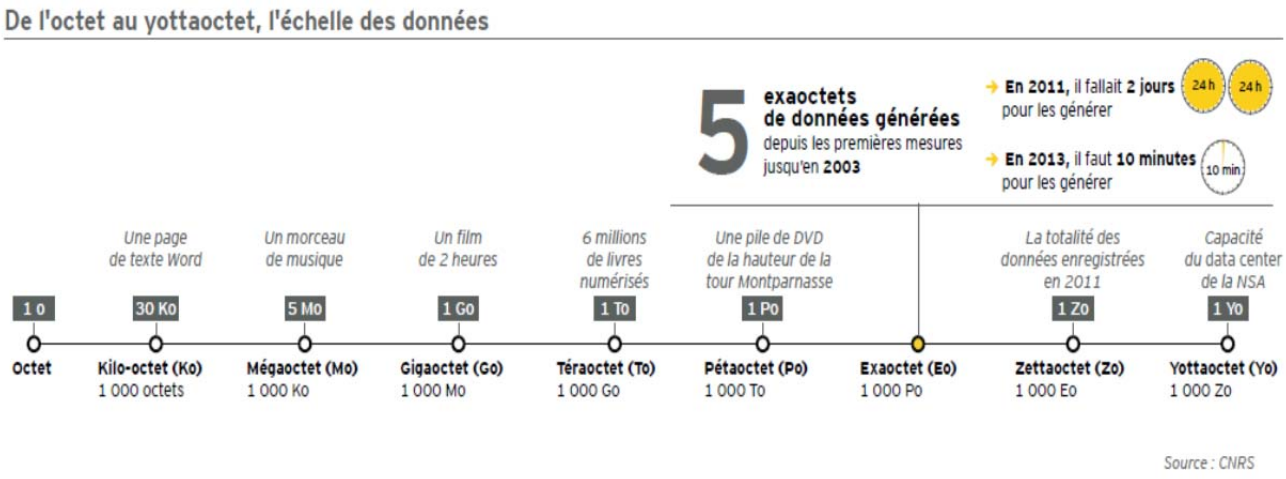
Pour bien comprendre la genèse du phénomène il faut remonter à une trentaine d'années en arrière.

Depuis le début des années 80, l'informatique réalise des progrès continus en matière de stockage et de capacité de traitement.
Au début de cette période les entreprises qui possédaient des datas, les VPCistes par exemple (La Redoute, et 3 Suisses notamment) eurent l'idée de confier à des prestataires extérieurs l'analyse d'extraits randomisés de leurs fichiers clients et des transactions correspondantes ; à cette époque, il était impensable de traiter l'intégralité du fichier.
Par ailleurs à l'époque, personne ne disposait des programmes de fidélisation qui permettent aujourd'hui d'ajouter du socio-démographique aux données brutes.
On savait tout sur les achats du client X, stockés sur longue période, mais au-delà de son nom et de son adresse, on ne savait rien d'autre. Pour pallier ces carences on probabilisait le genre des clients prénommés Dominique ou Claude, au moyen d'algorithmes travaillant à partir des achats effectués.
On commençait à réaliser des fusions avec des fichiers externes sur la base de variables communes
Cela aboutissait à des segmentations de clientèle, effectuées à partir de données passées, elles restaient opérationnelles un certain temps (plusieurs années le plus souvent) avant de remettre l'ouvrage sur le métier !

Le changement de paradigme opéré par le big data correspond à un mode de traitement de l'information qui tend vers le temps réel permettant un glissement rapide des segmentations statiques décrites ci-dessus vers des segmentations dynamiques mises à jour en continu.

Il est intéressant de souligner que les algorithmes fonctionnent essentiellement sur des calculs de corrélation sans recherche des causalités.

Quelques chiffres issus d'une publication d'Ernst et Young réalisée pour le forum culturel 2013 d'Avignon illustrent l'ampleur du phénomène. On pourra en retrouver le détail sur : http://www.forum-avignon.org/fr/etude-ernst-young-pour-le-forum-davignon



Un petit rappel des unités est sans doute utile :

1 kilo-octet (Ko)	= 10^3 octets	= 1 000 octets	
1 **mégaoctet** (Mo)	= 10^6 octets	= 1 000 ko	= 1 000 000 octets
1 **gigaoctet** (Go)	= 10^9 octets	= 1 000 Mo	= 1 000 000 000 octets
1 **téraoctet** (To)	= 10^{12} octets	= 1 000 Go	= 1 000 000 000 000 octets
1 **pétaoctet** (Po)	= 10^{15} octets	= 1 000 To	= 1 000 000 000 000 000 octets
1 **exaoctet** (Eo)	= 10^{18} octets	= 1 000 Po	= 1 000 000 000 000 000 000 octets
1 **zettaoctet** (Zo)	= 10^{21} octets	= 1 000 Eo	= 1 000 000 000 000 000 000 000 octets
1 **yottaoctet** (Yo)	= 10^{24} octets	= 1 000 Zo	= 1 000 000 000 000 000 000 000 000 octets

Soit un yottaoctet = 1 million de milliards de milliards d'octets, une friandise pour les pénombriens !

Le big data n'épargne pas la recherche. Marc Lipinski, conseiller régional d'Ile-de-France et directeur de recherche au CNRS a participé, le 6 décembre 2013, à une rencontre interdisciplinaire organisée autour de la question de l'ouverture des données massives scientifiques au CNRS. On retrouvera un entretien enregistré à cette occasion sur www.letudiant.fr. Cette rencontre fait le point sur les enjeux du big data. Elle a souligné que la recherche est aujourd'hui confrontée à un nombre croissant de données à traiter, analyser et stocker. Celles-ci ne proviennent pas seulement des chercheurs mais aussi d'un nombre important d'autres contributeurs non scientifiques qui ont la possibilité de participer à leur exploitation, si ces données leur sont ouvertes (open data).
Depuis, des dizaines de colloques plus ou moins sérieux ont eu lieu sur ce sujet….

L'Open data

Trois grands principes fondent l'open data :

- Un format ouvert
- La gratuité
- La liberté de réutilisation

De fait, l'univers public est présent dans le big data depuis le milieu des années 2000, avec les premiers pas de l'open data, à partir de données détenues par les administrations et services publics et mises à disposition du public dans un souci de transparence et d'une recherche d'une plus grande efficacité de l'action publique. Les compagnies de transport anglo-saxonnes et japonaises ont été les premières à mettre à disposition du public des données de fonctionnement de leurs services, permettant à des développeurs de proposer des applications mobiles utilisables au fil des déplacements.
Les transports publics et la circulation sont les domaines de prédilection de l'orientation temps réel dans l'open data.

L'extrait de la *Gazette des communes* (30/10/2013) ci-dessous montre que la France n'était pas très en avance en ce domaine. On consultera avec intérêt le site de la *Gazette* bien documenté à ce jour ainsi que le site d'Etalab pour plus d'information sur l'Open Data.

« Selon le classement Open Data Index de l'Open Knowledge Foundation, présenté lundi 28 octobre, la France n'arrive qu'en 16ème position sur 70 pays évalués. Cartographie, transports, dépenses publiques..., il reste des efforts à fournir pour se hisser au niveau de la Grande-Bretagne ou des États-Unis.
Henri Verdier, qui dirige la mission Etalab en France, réagit à ces résultats : 16ème sur 70, la France ne brille pas particulièrement dans le classement Open Data Indexhttps://index.okfn.org/ qui analyse l'ouverture des données, dévoilé lundi 28 octobre. Ce premier classement d'envergure a été établi de façon collaborative par l'Open Knowledge Foundation, http://okfn.org/ une association qui milite pour la culture libre et promeut à ce titre la gouvernance ouverte.
Il a été établi selon l'ouverture de dix sets de données majeurs : résultat des élections, dépenses publiques, émission de pollution, etc. Chaque item est lui-même noté selon plusieurs critères : la licence, le format, la gratuité… pour juger s'il respecte bien les principes fondamentaux de l'open data.http://sunlightfoundation.com/policy/documents/ten-open-data-principles/
« Si des progrès indéniables ont été réalisés ces dernières années, beaucoup reste à faire », regrette l'OKF. Sur les 700 jeux de données évalués, seuls 86 obtiennent le score maximal, soit 12%.
En tête de ce classement, on retrouve les pays anglo-saxons et les pays nordiques, pionniers sur l'open data, Grande-Bretagne, États-Unis, Danemark, Norvège. La France n'arrive donc qu'en 16ème place, derrière la Moldavie ou la Bulgarie, qui n'ont pas la réputation d'être particulièrement transparents. Bien que le pays ait signé la charte Open Data du G8 et soit engagé dans l'ouverture des données avec data.gouv.fr, trop de données publiques fondamentales restent indisponibles, déplore le chapitre français de l'OKF, ouvert cette année. »

La réutilisation et le partage des données sur les entreprises et sur les textes de lois restent soumis à redevance. Les cartes de faible résolution fournies par l'IGN en open data limitent les possibilités de réutilisation. Dans le secteur du transport, la SNCF ne publie toujours pas les horaires détaillés de ses trains à grande vitesse. Enfin, le détail des dépenses publiques reste hors de portée des citoyens. Les codes postaux ne sont également pas accessibles, alors qu'ils sont très utiles pour éviter des erreurs de géolocalisation dues à des communes homonymes et donc un nettoyage fastidieux des données.

« Décrire une réalité complexe avec des métriques unifiées » - Henri Verdier, qui dirige la mission Etalab en charge de l'ouverture des données en France, nuance ce résultat, quitte à tordre la définition de l'open data : « Un peu comme le classement de Shangaï pour les universités, l'Open Data Index a les inconvénients de ses avantages : il essaye de décrire une réalité complexe avec des métriques unifiées. Il adopte par exemple une définition très stricte de l'open data (format ouvert, gratuité et liberté de réutilisation). De ce fait, il récuse la qualité de « données ouvertes » à certaines données qui sont publiées par la France, gratuites et numériquement exploitables mais ne donnant pas lieu à droits de réutilisation. De même, il se concentre sur un cœur de jeux de données, et ne tient pas compte de toutes les autres données qui ont été partagées ».

Les cabinets de conseils américains rivalisent d'analyses sur les opportunités économiques offertes par l'open data ; on trouvera par exemple ci-dessous un rapport de Mc Kinsey : http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

Selon L'Atelier BNP PARIBAS :

« *Cette étude se concentre sur les gains en productivité grâce à l'Open Data dans 7 secteurs stratégiques : l'éducation, les transports, les biens de consommation, l'électricité, gaz et pétrole, la santé et enfin la finance personnelle. L'Open Data soit l'échange gratuit de données permettrait de corriger de nombreuses imprécisions inhérentes à certains secteurs, dans la production et la planification notamment. Par exemple dans le domaine des transports, près de 100 milliards de dollars pourraient être libérés grâce aux données issues de la maintenance des infrastructures, de la gestion des flottes et des comportements de consommation. En estimant avec précision les temps de trajet porte à porte les municipalités pourraient mieux ajuster les flux de transports en commun et économiser près de 35 heures de retards annuels. De même des mises à jour concernant l'état des véhicules pourraient résulter des économies de carburants et de maintenance représentant près de 370 milliards de dollars.»*

Enfin, la loi numérique, en discussion au parlement en janvier 2016 a donné lieu à de nombreux rapports préparatoires, notamment en date du 29 juillet 2015 :

ETUDE D'IMPACT
PROJET DE LOI
relatif à la gratuité et aux modalités de la réutilisation des informations
du secteur public
NOR : PRMX1515110L/Bleue-1

Et plus généralement le site http://www.cnnumerique.fr/ permet de recueillir une information dense et des liens utiles couvrant l'ensemble de la problématique.