Les promesses du big data

Blog Laurent Gille du 1er novembre 2015
http://www.loggos.fr/2015/11/01/les-promesses-du-big-data/

Le big data comme promesse technologique

Toute technique génère, au moment de son apparition, un ensemble de promesses, généralement bienvenues, parfois malvenues comme nous le constatons mi-2015 avec les robots tueurs. Ces promesses sont rarement tenues, car il est extrêmement difficile d'anticiper correctement les usages qui seront fait d'une technique, mais d'autres bénéfices viennent généralement se substituer à ceux qui étaient originellement promis. Le laser, par exemple, a généré des anticipations heureusement non survenues, le rayon de la mort par exemple, d'autres plus difficiles à mettre en œuvre, la transmission d'énergie, et d'autres bien plus nombreuses et inattendues, puisqu'il est désormais utilisé massivement dans la vie courante.

Le big data n'échappe pas au jeu des promesses. Le big data est un nouveau nom pour désigner la capacité à traiter et analyser des données, appelée auparavant analytics. Mais ces données sont désormais de grandes masses de données, non ou faiblement structurées dont on cherche à tirer des connaissances utiles à l'action. Le big data traite donc de grands **volumes** de données, de très grande **variété**, avec une grande **vélocité**. A ces 3V (Volume, Variété, Vélocité), d'aucuns en ajoutent d'autres, notamment la valeur qu'elles recèlent, la véracité issue de leur traitement, la visualisation des résultats obtenus.

La promesse du big data, c'est donc d'une part cette capacité technique à traiter de grandes masses de données peu structurées, mais aussi la capacité à tirer de ces traitements des décisions ou actions, souvent automatiques, sans intermédiation ou validation humaine, et à travers ces actions, évidemment la prétention d'améliorer le bien-être des individus et le bien-être social. Derrière le big data, réside le fantasme d'un accès direct à la « réalité » ou la « vérité », sans médiation humaine débouchant sur des actions. Supprimer l'humain, qui devient inefficace, pour le bien-être de l'humain! Ceci a été fait dans le domaine de l'énergie au 19ième siècle, dans le domaine des automatismes au 20ième siècle, nous y voilà en matière de connaissances, avant d'y parvenir dans nos actions sur l'humain lui-même.

Mais à côté de succès indéniables du big data, des papiers font référence au « big data hubris », c'est-à-dire à la démesure, à l'orgueil que suscitent les attentes du big data. Tout le monde a expérimenté cette sorte de pensée magique associée aux chiffres, qui fait que le chiffre dit vrai, surtout s'il est issu d'une masse de calculs, chiffre qui génère des « fantasmes d'omnipotence et d'omniscience »[1] que connaît bien tout statisticien quand il présente des résultats d'enquête ou de sondage.

Il est clair pour tous que le traitement massif de données tout comme l'apprentissage automatique font des progrès inestimables grâce aux techniques numériques. Et il faut aller de l'avant dans ces

domaines, ne pas bloquer a priori les champs d'application, mais néanmoins conserver à l'esprit quelques réflexions basiques pour ne pas sombrer dans l'océan des promesses qui ne pourront être tenues.

La promesse actuelle du big data est sans doute surévaluée pour différentes raisons que nous voudrions évoquer ici. Nous en développerons trois:

- La pertinence des données
- La pertinence des algorithmes
- La responsabilité des acteurs du big data

Il y a sans doute en matière de big data comme de toute nouvelle technique, un excès du possible sur le probable, comme le souligne Antoinette Rouvroy, et vraisemblablement sur le souhaitable. Les promesses expriment le possible, il nous reste à détecter le probable et le souhaitable, démarche que connaissent bien tous les prospectivistes. Prenons nos trois points.

D'où viennent les données? Sont-elles les bonnes?

D'où viennent ces données non structurées? Principalement de deux sources que le monde numérique aujourd'hui génère en masse: d'une part les traces laissées par les navigations numériques que nous-même ou nos robots effectuent chaque jour, et d'autre part, les mesures faites par des capteurs ou senseurs dont nous équipons tous nos environnements, ce que l'on appelle communément l'internet des objets. L'ambition, la promesse du big data est de transformer ce déluge de données en connaissances utiles pour des actions génératrices de mieux-être.

Qui sont les grands producteurs de ces données:

- En matière de traces, les grandes plates-formes de médiation, e-commerce, réseaux sociaux, économie collaborative, etc. au sens très large
- En matière d'objets connectés, les filières verticales mettant en place ces objets (automobile, santé, sécurité, domotique…).

Les ouvrages sur le big data foisonnent d'exemples sur les usages potentiels du big data, en mettant en avant ceux qui portent les promesses les plus « utiles » ou les plus « agréables »; c'est donc avant tout dans les domaines de la santé et de la sécurité qu'on va trouver des illustrations d'applications big data, permettant de détecter des maladies ou des dangers avant qu'ils ne surviennent, et ainsi les prévenir ou en minorer les incidences. Dans le monde économique, on présentera surtout les promesses qui représentent des sources d'économie ou de productivité. La technique se présente ainsi sous ses plus beaux atours pour convaincre de ses bénéfices potentiels. Certes, pour faire bonne mesure, les risques associés sont fréquemment mentionnés, comme la question de la protection des données personnelles dans le cas du big data, pour souligner qu'il serait bon que l'usage de cette technique s'accompagne d'un peu d'éthique et de responsabilité, tout en soulignant qu'il ne faut surtout pas légiférer ou réguler en la matière sous peine de passer à côté des bénéfices attendus, qui, comme toujours, peuvent générer des effets secondaires indésirables, mais avec un « bilan global positif ».

Le déluge de données annoncé a beau être un déluge, la qualité, la représentativité et la pertinence de ces données ne sont absolument pas garanties. Les données accessibles seront celles que nous donnerons les systèmes mis en place. Que mesureront-ils? Qui définira les mesures qu'ils effectueront? Les mesures accessibles et leurs métriques vont soulever des questions infinies sur leur pertinence, leur qualité, leur représentativité voire leur idéologie (la question des ontologies – structurations de la connaissance)[2]. Nous allons être massivement dans ce que j'appelle le syndrome du lampadaire: la quête que nous menons, nous la menons sous un lampadaire, parce que c'est éclairé, même si nous savons pertinemment que son objet ne s'y trouve pas. Il est des mesures autorisées là et interdites ailleurs (pratiques religieuses, politiques, opinions, races…), il est des métriques sur nos activités qui risquent de devenir des « proxy » de comportements (nombre de followers, d'amis, de like adressés, de mots utilisés dans des requêtes…) qui n'ont sans doute rien à voir avec les comportements qu'elles modéliseront. L'excès de volume peut être dangereux en ce qu'il accroit la perception d'une vérité issue des traitements, alors que dans bien des cas, c'est l'inverse qui prévaut: les statisticiens savent qu'un échantillon est bien souvent meilleur qu'un recensement qui présente de nombreux biais. La donnée numérique est par nature biaisée, manipulée (cf. la manipulation des notoriétés sur le web), partielle, liée à un système et une métrique imparfaits: si les traitements ne prennent pas garde à ces insuffisances, le big data se discréditera très rapidement.

Notamment, les traces et requêtes enregistrées sur internet sont extrêmement sensibles aux changements de comportements des individus, mais ne renseignent que très imparfaitement sur ces changements de comportement. Les comportements d'insincérité, voire de tromperie, des navigations ou renseignements fournis, sont de plus en plus présentes. Les capteurs placés sur des objets sont moins sensibles à ces phénomènes. Le big data ne se construira pas sur des bad data, c'est un de ses grands écueils.

Les algorithmes mis en œuvre peuvent-ils dire le vrai? La question de la confiance

Les données seront traitées grâce à ce que nous appellerons des algorithmes. Le big data promet « presque » que tout sera prévisible ou prédictible, puisque le voir et le dire se confondent de plus en plus. Ce qui introduit la question de la vérité. Les algorithmes du big data sont fondamentalement assis sur des analyses de corrélation, et non sur des recherches de causalité, ce qui est à la fois un gage de robustesse (les causes sont fréquemment multiples et récursives, ce qui rend leur recherche très difficile) et un gage de fragilité (les corrélations peuvent ne présenter aucune stabilité[3]).

Le big data promet de nous révéler les structures de la complexité, sinon inaccessibles, en agriculture, en santé, en industrie et de prendre les bonnes décisions (de semer, de se soigner, de prévenir la panne…). Gilles Babinet parle de rupture, « de remise en cause du canon anthropologique même de notre civilisation », dès lors qu'on se fiche totalement des causalités

Quelque que soit le nom que l'on donne à cette capacité d'intervention des machines, algorithme, intelligence artificielle, apprentissage automatique, data science, analyse prédictive… que finalement résume l'expression big data, le rôle qui sera confié à ces machines « intervenantes » sera questionné.

Qui n'a pas subi les conséquences de mauvaises corrélations ? Qui n'a pas subi la mauvaise qualité des données stockées dans des bases clients ? Le big data ne peut être que fragile, au moins ses premières années, mais au lieu de rendre les citoyens responsables (critiques) face à cette fragilité intrinsèque, il risque de les rendre fragiles face aux croyances qu'il va engendrer.

Le big data va devoir trouver son champ de pertinence et ne pas faire de promesses trompeuses ailleurs. On pressent que dans le domaine de la maintenance prédictive (quand les données concernent des objets ou machines), dans le domaine de la navigation, dans certains champs de l'ergonomie, le big data peut apporter des avancées. Il est douteux que cela puisse se produire dans tous les champs d'intervention possibles.

Le big data a une ambition prédictive, tourné vers l'action. Nous savons tous les limites de la notion de vrai, nous savons que le hasard est un moteur puissant de nos univers, physiques et humains, nous savons que l'incomplétude est inhérente à tous nos formalismes. Et pourtant, nous continuons à penser comme si le monde de Hilbert restait la norme universelle. L'inadvenu (ce qui ne pourra pas se lire dans les données collectées) semble être la norme de l'évolution. Relisons l'ouvrage de Chaitin sur le hasard et la complexité pour nous en convaincre, si Darwin ne nous suffit pas. Que devient l'intention humaine si nos actions sont dictées par ce qu'il est advenu hier ou aujourd'hui? Guidé par le big data, le monde de demain risque de ressembler à l'identique au monde d'aujourd'hui. Pour Stiegler, l'important n'est pas de voir ou dire le monde, c'est de savoir quel monde nous voulons. Babinet illustre les avantages du big data en montrant que nous pourrons configurer chaque environnement nouveau comme ceux que nous avons configuré préalablement (retrouver dans sa chambre d'hôtel ses configurations domestiques, retrouver dans un avion les divertissements que nous aimons, etc.): sont-ce là promesses alléchantes? Parfois peut-être, parfois certainement pas.

Tous les scientifiques connaissent les vertus de la sérendipité, « le fait de réaliser une découverte scientifique ou une invention technique de façon inattendue à la suite d'un concours de circonstances fortuit et très souvent dans le cadre d'une recherche concernant un autre sujet. La *sérendipité* est le fait de « trouver autre chose que ce que l'on cherchait », comme Christophe Colomb cherchant la route de l'Ouest vers les Indes, et découvrant un continent inconnu des Européens. » (Wikipedia). Le big data nous fait courir un double risque: celui d'éliminer le hasard, celui de nous dire le vrai en toutes circonstances. Pas sûr que nous voulions d'une telle évolution, même si ponctuellement, nous pouvons la solliciter.

La fiabilité des algorithmes pour dire le vrai laisse également songeur, même si on peut espérer de grands progrès. L'un des secteurs les plus avancés en matière de traitement de données est la finance, qui manipule des masses considérables de données, à de très hautes fréquences, et qui sont plutôt des données structurées. Ses prévisions sont-elles plus fiables? Les krachs boursiers ont-ils disparu? Les traders eux-mêmes souhaiteraient-ils une prédictibilité absolue de leur portefeuille qui verrait disparaître leur profession? Malgré la somme de données accessibles, les sondeurs ont-ils réduit leurs erreurs de prévision? Malgré des modèles utilisant les plus grandes puissances de calcul, les météorologues y voient-ils plus clair sur le temps à venir? Sans citer la difficulté de Google Flu à tenir ses promesses. Les échantillons des centres de contrôle des épidémies sont infiniment plus fiables.

L'industrie saura-t-elle être responsable face au big data?

L'industrie sera-t-elle responsable de ce qu'elle va pouvoir manipuler? Laissons de côté la question des données personnelles abondement discutée à propos de big data, et la polémique lancée par Gilles Babinet sur la fermeture de la CNIL. L'anonymat ne peut être absolu, Patrick Tucker nous le rappelle, mais bon nombre d'acteurs continuent à penser le contraire. A delà de cette question, où se situe cette responsabilité?

L'erreur pourrait être congénitale du big data. Les acteurs sauront-ils mettre les garde-fous qui permettent d'éviter les plus grosses? Le premier est de rendre transparent les algorithmes utilisés, ce qui est rarement le cas, action au moins aussi urgente que la question de la neutralité du net. Le principe de précaution devrait trouver un nouvel espace de débat.

De nombreuses voix aujourd'hui s'élèvent sur les conséquences sur l'emploi du big data. A titre d'exemple, General Electric indique qu'elle collecte d'énormes quantités de données de senseurs présents dans ses turbines à gaz ou éoliennes, les plates-formes de forage, les moteurs d'avion et les locomotives qu'elle produit. En traitant ces données pour prévenir les problèmes et optimiser l'emploi de ces machines, GE estime qu'elle fera économiser 20 milliards de dollars à ses clients. La rhétorique sur les termes est intéressante: GE apporte 20 milliards de $ de valeur à ses clients, ce qui correspond environ à la destruction de 200 000 emplois. Il faut être sacrément schumpétérien pour penser la transformation de ces « économies » en « croissance future », mais nous le sommes, même si les économistes ont montré que les ruptures schumpétériennes n'allaient pas sans crises chaotiques importantes.

Un des enjeux industriels du big data sera l'appariement des sources de données pour en renforcer la pertinence: détruire les silos d'information qui auront tendance à se constituer, ouvrir les données (open data) de façon à renforcer l'efficacité des traitements, tirer parti des externalités très fortes que peuvent générer l'accès aux données. L'expérience nous montre que, dans de telles situations, les industriels tentent d'abord les solutions fermées, en silos, pour essayer d'acquérir des positions dominantes qui leur ouvrent de belles rentes. On ne leur en fera pas grief, mais ils risquent ainsi d'écorner fortement le potentiel du big data, ce qui se passe actuellement dans le domaine de la santé à travers la question sensible de l'ouverture de ces données.

La valeur du big data reste trop souvent pensée encore en termes de données. Et pas suffisamment en termes d'externalités. Ce sont les données qui porteraient la valeur, données fournies par les consommateurs/usagers des services. Internet nous apprend pourtant que la valeur est désormais principalement issue des externalités des marchés bifaces. Le modèle Google qui améliore son moteur grâce à l'interaction qui se joue entre les requêtes et les réponses en est une bonne illustration. Pour nous, le projet Watson d'IBM illustre le meilleur comme le pire de ce que nous pouvons attendre des big data sur le plan de la valeur. Le projet d'intelligence artificielle d'IBM, semble être le préfigurateur de ce que le big data pourrait produire. Le succès de Watson dans le jeu Jeopardy aux Etats-Unis en 2011 a mis sur le devant la capacité de la machine à draguer extensivement le web pour produire des réponses fiables à certains types de questions. L'enjeu est désormais d'introduire dans un tel écosystème des masses de données de plus en plus nombreuses pour affiner et étendre les réponses que le système peut apporter dans des domaines de plus en plus larges. IBM tisse aujourd'hui de nombreux accords d' »ingestion » de données, météorologiques (The Weather Company), médicales, de la relation client… de façon à se positionner sur de

nouvelles plates-formes bifaces à fortes externalités, entre les détenteurs de données et les « consommateurs de réponses »[4]. La requête appelle une réponse cognitive. Là résidera sans doute la valeur du big data.

Le risque

Enfin, il nous faut aborder la question du risque. Le big data promet de mieux évaluer les risques (de maladie, de danger de toute nature, de panne, d'accident…) afin d'en prévenir les conséquences. Réduire les niveaux globaux de risque est une ambition louable à laquelle chacun pourrait souscrire. Mais, nous ne sommes pas égaux devant les risques, du fait de nos gênes, du fait de notre éducation, du fait de nos professions, du fait de nos localisations, du fait de nos activités, de nos revenus, etc. : pour couvrir les risques individuels, ont été mis en place de nombreux mécanismes assurantiels qui traduisent la solidarité qui lie les membres des communautés et des sociétés qui font face à ces risques. L'évaluation individuelle des risques que promet le big data va-t-elle faire voler en éclat l'équité, la solidarité, la justice sociale qui émane de nos assurances? L'industrie saura-telle résister à une telle évolution, et les mécanismes de marché n'y conduisent-ils pas automatiquement?

Nous prenons certaines promesses du big data, nous doutons de certaines autres, nous en craignons aussi quelques-unes. Comme toujours face à une évolution technique, ceux qui la portent luttent farouchement contre toute tentative d'encadrement et de régulation de leur activité. On caresse l'idée de supprimer la CNIL, on songe à de nouveaux modèles d'affaires où l'on saura tout de ses clients sans qu'ils sachent comment on sait tout d'eux, où on fera payer le « juste » prix à chacun selon ses risques, on configurera automatiquement les environnements des uns et des autres, puisque c'est comme cela qu'ils aiment vivre, on poussera à encadrer les comportements des utilisateurs pour qu'ils réduisent le risque qu'ils font peser sur leurs fournisseurs, finalement, on saura de mieux en mieux ce qui est bon pour la société et chacun de ses membres. Est-ce là le monde que nous voulons?

Mais surtout, penser les évolutions comme cela, c'est tuer le big data dans l'oeuf: le big data, comme toute nouvelle technique, devra faire la preuve de son comportement éthique pour que l'éthique des utilisateurs se cristallise. Et en cela, quoiqu'en pensent les industriels et leurs soutiens, un peu de régulation peut y aider pour que les promesses consensuelles puissent voir le jour.

Le big data comme toute innovation soulève à nouveau la question du progrès: progrès pour qui? Ce faisant, il soulève à nouveau de vieilles ou de nouvelles croyances, de vieux mythes, de vieilles ou nouvelles peurs et espoirs.

Références:

Gilles Babinet, Big Data, penser l'homme et le monde autrement, Le passeur, 2015

Paul Hermelin, François Bourdoncle, Feuille de route Big Data, La nouvelle France Industrielle, 2014

[1] Julien Laugel (MFG Labs) in le Monde du 31 octobre, « le défi de tous les superlatifs », page 7 (Culture et Idées)

[2] Comme le souligne Daniel Kaplan, « une donnée n'a rien de « donné » : elle est construite comme une variable, avec une finalité précise, puis produite, saisie ou acquise par des mécanismes qui constituent autant de médiations plus ou moins masquées. Cette quête sans fin se condamne à la course aux armements : il faut des moyens sans cesse plus importants pour parvenir à des inférences un peu plus fines, avec des rendements de plus en plus décroissants. Et sans espoir, nous dit Mitchell, de parvenir au Graal : se connecter vraiment à l'unité, la subjectivité, la complexité des êtres humains et de leurs pratiques sociales. Pourquoi ? Parce que les objets du calcul, en l'occurrence les humains, n'y sont pas conviés. » InternetActu 11/04/2012

[3] « Haven't these guys been paying attention? It's easy to predict the past. Does anyone remember the University of Colorado professors who had a model that correctly predicted every election since 1980? In August 2012, they confidently announced that their model showed Mitt Romney winning in a landslide. » http://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/

[4] « IBM annonce la création d'une entité de conseil dédiée à l'informatique cognitive. Elle entend ainsi permettre aux entreprises de tirer de l'informatique cognitive pour développer leur business. Elles pourront ainsi bénéficier des capacités de Watson, mais aussi de l'expertise de big blue en matière d'analytique. Plus de 2 000 consultants spécialisés en Machine Learning, Analytics, Data Science et Développement seront mis à profit pour cette offre de conseil. Big blue précise qu'ils seront aidés par des spécialistes dans l'accompagnement au changement pour accélérer la transition des clients vers le Cognitive Business. Cette nouvelle entité proposera ses services aux entreprises actives dans les secteurs de la banque, du retail, de la supply chain ou encore de la santé. » 12-10-2015 http://www.lemondeinformatique.fr/actualites/lire-ibm-rassemble-une-entite-services-dediee-a-watson-62629.html