ENTRETIEN AVEC UN-E STATISTICIEN-NE

Dans cette chronique « Entretien avec un-e statisticien-ne », nous partons à la rencontre de celles et ceux qui font la statistique, tant des praticiens qui l'utilisent comme un outil essentiel dans le cadre de leur activité professionnelle que des universitaires qui la développent dans le cadre de leurs travaux de recherche et d'enseignement. Une ou plusieurs série(s) de questions mettent en particulier en lumière le rapport à l'enseignement. Les praticiens sont appelés à exprimer les besoins de formation qu'ils jugent prioritaires, les compétences qu'ils apprécient chez ceux qu'ils recrutent. Les universitaires sont interrogés sur leur vision personnelle de la statistique et de son importance, sur leur art de la transmission du savoir statistique, entre théorie et applications, idéalisation professorale et réalités étudiantes.

OLIVIER AULIARD : LES MAINS DANS LE CAMBOUIS

(Il faut mettre les mains dans le cambouis de la donnée.)

Olivier AULIARD[1] et Gilles STOLTZ[2]

Olivier Auliard est, depuis mai 2012, directeur scientifique de Capgemini Consulting, après avoir exercé dans les domaines des sondages et du géomarketing. A Capgemini Consulting, il supervise en particulier un centre de compétences en science des données. Il est par ailleurs membre du conseil de la Société française de statistique et dispense notamment des enseignements à l'IUT Paris Descartes. On peut le retrouver sur les réseaux sociaux et sur la toile, *via* son profil LinkedIn et sa page Capgemini Consulting, avec fil Twitter intégré.

L'entretien a été mené par des échanges de courriels entre le 4 janvier et le 29 avril 2016, ainsi que par un entretien en face à face dans les locaux de Capgemini Consulting le 22 avril 2016.

Parcours académique et professionnel

GS : *Olivier, merci infiniment d'être le premier statisticien en entreprise de cette chronique ! Je voudrais commencer par le commencement : où en étais-tu vers 20 ans, comment se déroulaient tes premières années d'études supérieures, quels étaient tes projets d'avenir à l'époque, quelle profession et dans quel univers te voyais-tu exercer ?*

OA : A 20 ans, je venais d'intégrer l'X[3], après deux années de classes préparatoires en province, en l'occurrence à Poitiers. Jusqu'à cette période, je ne m'étais pas vraiment posé de questions

[1]Capgemini Consulting, France, olivier.auliard@capgemini.com
[2]HEC Paris, CNRS, Jouy-en-Josas, France, stoltz@hec.fr
[3]GS : Nom familier donné à la très sélective et très réputée Ecole Polytechnique, située à Palaiseau, en région parisienne, et qui dispense une formation d'ingénieur généraliste, à compléter par une école d'application.

sur mon avenir. Les seules certitudes que j'avais alors, c'est que je ne voulais pas rentrer dans la fonction publique... ni faire de l'enseignement, en tout cas à plein temps, ma mère étant alors enseignante en histoire ancienne à l'université, et mon père, inspecteur d'académie. Je considérais également que la recherche ne correspondait pas à mon tempérament. Par ailleurs, malgré une formation scientifique, j'avais toujours eu un fort attrait pour les sciences sociales et les matières littéraires. J'ai donc cherché un métier qui pouvait marier ces deux aspects, et je me suis dit que les sondages et études de marché pouvaient assez bien correspondre à mes affinités. Dès ma formation à l'X, j'ai par conséquent pris contact avec les instituts en vogue à l'époque pour leur proposer de me prendre en pré-contrat tout en suivant ma scolarité à l'ENSAE[4] comme école d'application. Ce n'est d'ailleurs que lors de la formation à l'ENSAE, et en particulier lors des cours d'analyse des données, que j'ai découvert vraiment la statistique.

GS : *Excuse-moi, Olivier, mais pour moi, sondages et études de marché sonnent sciences de gestion, et me semblent assez loin des sciences sociales et humanités ! Comment voyais-tu le lien à l'époque ?*

OA : Pour moi, le lien, c'est que les problématiques marketing et politiques sont toutes relatives à l'étude des comportements et à l'étude des groupes sociaux. On essaie de comprendre comment les humains agissent, interagissent, réagissent... pour des problématiques *business*, certes, mais avec des ressorts qui tiennent de la nature humaine. C'est donc très différent pour moi de ce que je mets sous les étiquettes de gestion et de finance. De la même manière, j'adore la micro-économie parce que l'on va chercher à reconstituer des utilités individuelles alors que la macro-économie m'a toujours paru plus aride...

GS : *Je rebondis maintenant sur la dernière phrase de ta présentation à 20 ans. Tu sembles dire être venu à la statistique par les données (et non, par exemple, par les mathématiques !). Peux-tu nous en dire davantage ? Quel professeur à l'ENSAE t'a-t-il marqué, de quelle méthode de cours te souviens-tu tout particulièrement ?*

OA : Eh non ! je ne suis pas venu à la statistique par les données... Ce qui m'a attiré, c'est que l'on pouvait utiliser des objets mathématiques plus ou moins complexes et théoriques pour comprendre et interpréter des données réelles. En particulier, j'ai en effet adoré les cours d'analyse des données de Gilbert Saporta : utiliser la diagonalisation de matrices pour mettre en évidence des liens entre individus, cela était fascinant pour un étudiant baigné de concepts ultra-théoriques (que j'adorais aussi par ailleurs). J'ai beaucoup aimé aussi les cours de Jean-Marie Grosbras sur les sondages, mais là évidemment, je n'étais pas vraiment objectif.

GS : *J'allais justement te demander quels enseignants et quels cours t'ont marqué, tu m'as devancé ! Je pense donc qu'on a fait le tour de ta formation académique. Passons maintenant à ta carrière professionnelle !*

Je vois sur ton profil LinkedIn que (sans surprise) tu l'as commencée par TNS Sofres. Que penses-tu de cette première expérience et comment as-tu été amené à passer au poste suivant ?

OA : Globalement, j'en conserve plutôt de bons souvenirs, car j'ai occupé des postes variés et ai mené des projets intéressants ; la société se développait beaucoup, notamment à l'international. J'étais un peu déçu par la dimension statistique et analytique, qui était très embryonnaire à l'époque. En revanche, une dimension de veille et de R&D assez présente m'a permis de m'in-

[4] Ecole nationale de la statistique et de l'administration économique, désormais ENSAE ParisTech, sise à Malakoff, en région parisienne.

téresser très tôt au Web, avec des projets de développements importants autour de la mesure d'audience et des comportements des internautes dès la fin des années 90. Mais le rachat de Sofres par Taylor Nelson a un peu mis un frein à ces projets, et je suis alors parti, d'abord vers Démoscopie (une autre société de sondages, de taille moyenne, et qui n'existe plus, car elle a été rachetée et absorbée). J'ai ensuite tenté l'aventure de la start-up en 2000 avec deux autres fondateurs (juste avant l'éclatement de la bulle Internet, autant dire que j'ai eu beaucoup de flair...).

Au bout de deux années de succès relatif de ma start-up, il a fallu revenir à la vie réelle. J'ai alors failli repartir dans le monde des études, mais j'ai préféré accepter l'offre d'ASTEROP qui était alors une start-up (mais avec des actionnaires et investisseurs solides), dans le domaine du géomarketing.

GS : *Je vois sur LinkedIn toujours que ce poste suivant, que tu as rejoint en 2002, est catégorisé comme requérant des compétences de données massives (*big data*), le thème de ce numéro de la revue. En 2002, parlait-on déjà de* big data *? J'ai le sentiment d'avoir vu émerger le thème dans les 5 à 7 (disons) dernières années...*

OA : En 2002, personne ne parlait de *big data* ; en revanche, au moment où j'ai écrit un article pour la revue des anciens de l'ENSAE (Auliard, 2009), le thème commençait à percer. Il y avait en effet dans le géomarketing quelques amorces de ce qui pouvait être assimilé au *big data* : des données très variées, avec des sources internes et externes (en particulier les fichiers détail[5] de l'INSEE, qui ont été d'une certaine manière précurseurs de l'*open data*), des problématiques statistiques et analytiques sophistiquées avec des premières ébauches d'application du *machine learning* (apprentissage automatique). Les volumes n'étaient pas toujours au rendez-vous, mais parfois les données de transaction client issues des cartes de fidélité pouvaient atteindre des tailles non négligeables. Les outils de traitement de ces données restaient en revanche plutôt classiques.

Questionnaire à la Proust

GS : *Pour continuer de faire connaissance, Olivier, je te propose un moment de détente : un questionnaire de Proust (ou en tout cas, sa version statistique). La règle du jeu est qu'il faut donner une réponse courte, à tout le moins pour la plupart des réponses...*

Quel est ton résultat de statistique préféré (théorème ou application) ?

OA : On va dire le théorème de Bayes, même si le théorème limite central me plaît pas mal aussi.

GS : *Je précise pour nos lecteurs : le théorème de Bayes est la formule*

$$\mathbb{P}(A\,|\,B) = \frac{\mathbb{P}(B\,|\,A)\;\mathbb{P}(A)}{\mathbb{P}(B)}\,.$$

Je te dirai dans un instant ce que m'inspire ce résultat, mais toi, pourquoi as-tu cité cette formule ?

OA : Il y a plusieurs versions de ce théorème ; ce qui m'intéresse avant tout, c'est le principe des probabilités conditionnelles. Je trouve ce concept particulièrement riche, par exemple parce qu'il permet d'intégrer de la connaissance exogène dans les analyses, parce qu'il permet de mieux

[5] GS : L'INSEE précise sur son site web que « les fichiers détail [...] sont des bases de données électroniques comportant les enregistrements individuels, anonymisés, afférents à des enquêtes ou opérations statistiques réalisées par l'INSEE ».

rendre compte d'un environnement, d'un contexte. C'est un concept à la fois pragmatique et ouvert.

GS : *Comme promis, voici mes commentaires sur cette formule de Bayes : pour moi, elle semble plus probabiliste que statisticienne à première vue, sans doute parce que c'est généralement dans le cours de calcul des probabilités (qui précède de loin celui de statistique) que nous, futurs chercheurs en mathématiques, le découvrons. Mais en statistique, cette formule, comme toute la théorie de l'estimation bayésienne, c'est finalement la révision des connaissances au vu des données (au vu des expériences), ce qui est un bon concept de sciences expérimentales... et la statistique en est une.*

Reprenons le questionnaire. Quel est ton manuel de statistique préféré ?

OA : J'aimais bien *Sampling Techniques* de Cochran (1953), mais c'était avant... Sinon, maintenant les grands classiques type *Statistical Learning Theory* de Vapnik (1998) ou l'article *Random forests* de Breiman (2001).

GS : *Tu dis « avant »... mais avant quoi ?*

OA : Le « avant » est assez personnel : pour moi, il fait référence au monde des études et des sondages, dans lequel j'ai passé tout le début de ma carrière. Mais c'est aussi une activité qui est sinistrée aujourd'hui et qui n'a pas su, à mon sens, s'adapter à la révolution de la donnée, pour diverses raisons (je n'inclus évidemment pas mon départ parmi ces raisons...).

GS : *Fréquentiste, bayésien, ou autre ?*

OA : En premier lieu, je dirais « non paramétrique ». Je n'aime pas partir avec des *a priori*.

GS : *Statistique ou statistiques ?*

OA : Statistique !

GS : Data science *: science des données ou science de la donnée ?*

OA : Là en revanche, je préfère largement le pluriel...

GS : *J'aurais proposé deux fois le singulier, car j'aime voir une unité dans la démarche statistique, y compris lorsqu'il s'agit de traiter des jeux de données. « Science des données » sonnerait pour moi comme la reconnaissance d'une diversité incompressible dans les jeux de données et du fait que les traiter, c'est peut-être un art, avec des approches personnelles. Quels étaient les déterminants de ton couple de réponses singulier / pluriel ?*

OA : Tu as donné dans ta question mes éléments de réponse. D'une part, le singulier de statistique est effectivement lié à la démarche, à une approche génériques. En revanche, le pluriel de données vise à refléter leur diversité. Mais mettre un pluriel n'implique pas du tout – dans mon esprit – que cette diversité soit incompressible, bien au contraire : science des données, c'est une approche générique (science) permettant de rendre compte des diversités et des complexités (données au pluriel).

La donnée, je trouve cela trop unificateur, voire totalitaire. Malgré toutes les transformations que nous pouvons faire subir aux jeux de données, si nous détectons une erreur, une anomalie ou simplement des résultats contre-intuitifs, nous en revenons toujours aux données originelles, dans toute leur complexité. C'est en quelque sorte notre point d'ancrage dans le monde réel, et je préfère respecter leur diversité.

GS : *Quelle est la faiblesse principale de la statistique ?*

OA : Les limites des modèles prédictifs, la capacité à identifier des ruptures.

GS : *Quelle est ta vertu préférée en statistique ?*

OA : L'agnosticisme ! Je veux dire par là l'absence d'*a priori* par rapport aux données.

GS : *Quelle est notre principal défaut en tant que statisticiens ?*

OA : De ne pas assez intégrer l'informatique (comme science : la *computer science*) dans nos approches et de ne pas chercher à comprendre comment les données ont été recueillies (ou générées).

GS : *Quel est le rêve de tout statisticien ?*

OA : Pas de *missing values* dans nos *data sets*...

GS : *Et son cauchemar ?*

OA : L'*overfitting* évidemment...

GS : *Ce que j'aime dans tes deux réponses, Olivier, c'est que tant dans le fond que dans la forme, elles révèlent beaucoup ton univers professionnel. Je me permets toutefois, concernant leur forme et pour partager le souci de diffusion de l'enseignement de la statistique en français qui est la mission de notre journal, de traduire tes propos :* « Pas de valeurs manquantes dans nos jeux de données » *et* « Le sur-ajustement évidemment... ».

Le monde de la recherche académique en statistique est relativement mixte ; qu'en est-il du monde professionnel en science de la donnée (des données) ?

OA : Cela dépend des générations, c'est encore très masculin chez les seniors, cela évolue progressivement. Les générations actuelles sont presque à parité (presque seulement).

GS : *Je sais que tu connais la SFdS, puisque tu es membre de son conseil ! Que t'apporte-t-elle ?*

OA : Tout d'abord, je voudrais dire que j"aimerais avoir plus de temps encore à y consacrer. Mais pour moi, elle me permet de garder un contact essentiel avec le monde universitaire et de la recherche.

Mises en œuvre statistiques « classiques »

GS : *Il pourra te paraître surprenant que les universitaires (et les enseignants du secondaire) mettent rarement les pieds dans une entreprise. Pour nous, il est parfois difficile d'imaginer quel est le travail d'un statisticien au quotidien, par exemple chez TNS Sofres (il ne s'agit sûrement pas d'exhiber des estimateurs avec de meilleures propriétés mathématiques !). Peux-tu nous décrire en quelques mots ou exemples ce que tu faisais quotidiennement en début de carrière chez TNS Sofres, avant les responsabilités de management de la R&D ? J'imagine que tu ne collectais pas toi-même les données, mais supervisais-tu leur recueil ? Les traitais-tu toi-même ?*

OA : Mes interventions étaient très variées, d'un bout à l'autre de la chaîne de production des études : En appui commercial avant-vente, pour imaginer les méthodes et techniques permettant de répondre au mieux aux objectifs du client, et défendre ensuite les propositions dans le cadre d'appels d'offres par exemple.

Typiquement, ce type d'intervention permettait de mieux définir les échantillons (stratifications, plans de sondage, tailles d'échantillons) ou d'intervenir sur la méthodologie même de l'enquête ou sur le questionnaire (par exemple inclure une partie d'analyse conjointe[6]). Ensuite, les données étaient recueillies par les différents terrains (face à face, téléphone, postal), saisies par des opérateurs (sauf le téléphone). Il y avait ensuite un premier « dépouillement » des données de façon à produire des tableaux croisés. J'intervenais alors sur certaines études pour mener des analyses un peu plus sophistiquées (tests non paramétriques, calcul d'intervalles de confiance dans des plans de sondage sophistiqués, analyses factorielles, classifications, construction de quelques modèles prédictifs). Je produisais souvent moi-même ces éléments, du moins au début.

GS : *Peux-tu nous partager un ou deux exemple(s) de mises en œuvre de statistique « classique » que tu as réalisées ou supervisées et que tu aimes particulièrement partager ?*

OA : Il y a une intervention qui m'a laissé quelques souvenirs : Sofres avait dans les années 1990 un panel postal auprès des constructeurs automobiles européens, et ce panel avait été violemment critiqué par certains constructeurs européens parce que le recrutement s'appuyait sur la méthode des quotas et qu'il n'était pas possible de calculer de manière fiable des intervalles de confiance. J'avais été sollicité alors pour défendre le panel. A peu près au même moment, Deville à l'INSEE avait commencé à bâtir des premiers éléments de théorie sur cette méthode, et je m'étais appuyé sur ses travaux pour défendre le panel, et montré que l'on pouvait calculer des intervalles de confiance de manière solide. Il y avait eu une réunion à Rome notamment où j'avais en face de moi un expert mandaté par les constructeurs réticents pour démolir le panel. Je me souviens de débats assez enflammés sur le sujet – juste avant ma première intervention dans cette réunion, le responsable du projet chez Sofres m'avait glissé, à voix basse : « Débrouille-toi, il y a x millions de francs en jeu maintenant... » *In fine*, on avait convaincu un des deux constructeurs réticents de rester et le panel avait donc été renouvelé.

Parlons de « big data » (de données massives)

GS : *Je lis sur ta page Capgemini que tu es* « Director for Data Intelligence » *et dans la description de ta mission, tu écris :* « I help my clients to find out value from their own data and external data sources in order to improve customer experience [...] through innovative proof of concepts based on advanced analytics. »

Peux-tu nous expliquer en quelques lignes en quoi consiste, concrètement, ton poste ? Si tu vois des données et mets en œuvre des traitements ou visualisations statistiques, ou si tu ne fais que guider les clients vers une mise en œuvre par eux-mêmes de ces traitements ou visualisations ?

OA : Je pense qu'une des spécificités de Capgemini Consulting et de la *Data Farm* dont je m'occupe (« ferme de données », qui est un centre de compétences) est que justement, nous traitons vraiment des données, depuis l'acquisition jusqu'à la visualisation. Concrètement, la structure dont je m'occupe (la *Data Farm*, qui regroupe 15 *data scientists*) va intervenir sur toutes les missions qui ont une composante de traitement de la donnée (et il y en a de plus en plus !). Théoriquement, mon équipe devrait doubler d'ici à la fin de l'année. Si je devais te décrire

[6] GS : L'analyse conjointe, ou *trade-off* dans le jargon des sondeurs, est une modélisation des préférences individuelles (souvent par techniques de régression) en fonction des différents attributs d'un produit. Elle vise à mettre en lumière les compromis effectués par les consommateurs face à des situations de choix (d'où son nom anglais).

une mission typique (en général d'une durée de 3 à 6 mois), il y a systématiquement une première phase de cadrage qui va nous permettre de bien comprendre les besoins du client et ses attentes, et d'avoir une première vision des données disponibles (au moins de savoir où aller les chercher) ; puis une phase de récupération et d'acquisition des données internes et externes (mélange entre exploration des systèmes d'informations du client et *scraping* de données) ; puis la phase de préparation des données (souvent longue mais essentielle pour la suite). On passe alors à la modélisation (toujours trop courte !), puis à la restitution. Enfin bref, je pense un schéma assez classique...

Tout cela, nous le faisons nous-mêmes, sur nos outils si possible ou chez notre client. Quant à moi, je dois m'occuper de cette équipe et de nos clients, donc je suis de moins en moins les mains dans le cambouis. Je veille quand même à rester au contact de la donnée, à m'impliquer dans la phase de conception de nos méthodes, et à intervenir dès que cela commence à accrocher (ou à ne pas accrocher, justement). Je code encore, je manipule encore de la donnée, mais de moins en moins.

Sur d'autres missions, nous avons également un rôle plus « classique » de conseil, avec des recommandations sur les traitements, les outils, les problématiques à traiter. Mais le plus gros de notre temps, c'est bien avec la donnée.

GS : *Tu écris « la donnée » : c'est intéressant pour quelqu'un qui expliquait qu'il préférait dire science* des *données ! Mais je lis souvent les professionnels parler de la donnée (voire de la* data*), comme on parlait de l'or ou du minerai... c'est peut-être un filon de richesse à exploiter dans notre monde numérique.*

OA : Zut, je suis pris en flagrant délit d'incohérence ! J'utilise effectivement les deux, pour des motifs plus esthétiques que rationnels. Remarque tout de même que mon utilisation du singulier est en général (pas systématiquement) associée à l'article partitif « de », comme on dit du sel ou de l'eau. Sur le fond, je maintiens mon choix sur le pluriel.

GS : *Plus sérieusement, je voudrais revenir sur la phase de recueil et de préparation des données : pour nos lecteurs peu familiers avec les données, peux-tu expliquer ce qu'est le* scraping *d'une part (avec une traduction ?), et d'autre part, nous donner un bref aperçu de ce qu'il est nécessaire de faire pour préparer les données ?*

OA : Le *scraping*, c'est le « grattage » des données, soit l'aspiration des données en ligne sur des sites Web ou des réseaux sociaux, *via* diverses techniques, et notamment, l'automatisation de requêtes (requêtage). Un certain nombre de sites proposent aux utilisateurs du Web de tels outils de requêtage (des API, *application programming interfaces* ou interfaces de programmation applicative) permettant de récupérer des données très simplement, sous un format structuré, grâce à un appel dans un code (sous R ou un autre langage de haut niveau). Par exemple, Twitter permet de récupérer tous les *tweets* sur des mots-clés donnés (avec une limite de nombre de *tweets* et de temps). Je fournirai un autre exemple, de géo-codage d'adresses physiques par une API de Google, lorsque nous parlerons de mes enseignements.

Préparer les données, c'est le travail le plus long et le plus fastidieux dans le monde réel. Il y a les opérations de nettoyage : contrôles divers (de cohérence, de remplissage, etc.) ; puis le traitement des valeurs manquantes ; ensuite des problématiques de « mise à niveau » pour relier les différentes tables entre elles (par exemple individus *versus* familles, *versus* actes d'achat, *versus* navigation sur le Web) ; ensuite le calcul de variables dites utiles ou secondaires (une durée

à partir de deux dates, par exemple), tirées des données brutes (qui contiennent les variables dites primaires) ; enfin, la construction des échantillons de travail, avec notamment la construction de la variable cible dans les logiques supervisées.

La sélection des variables fait partie de la phase de modélisation à mon sens.

GS : *Tu as parlé plus haut de ton équipe de* data scientists *: dans ce numéro spécial de notre revue, plusieurs articles décrivent des formations au maniement des données massives. Ces articles décrivent souvent les qualités et compétences attendues des* data scientists*, sur lesquelles je te propose de revenir un peu plus tard dans cet entretien. Pour l'instant, je voudrais d'abord mieux cerner avec toi leur fiche de poste, leur travail quotidien. En particulier, comment situerais-tu le* data analyst *par rapport au* data scientist*, et au passage, aurais-tu des traductions agréables à nous proposer pour les titres de ces métiers ?*

OA : La traduction précise est complexe à mon avis. Je ne suis pas sûr d'avoir la définition la plus universellement reconnue de ces deux métiers, mais voici comment je vois les choses. Pour *data analyst*, je dirais préparateur ou analyste de données. C'est la première marche pour devenir *data scientist* : dans mon équipe, les débutants sont d'abord *analyst* avant d'être *scientist*.

Je n'ai pas de traduction satisfaisante à te proposer pour *data scientist* ; je récuse en tout cas « scientifique de la donnée » (ou des données). Peut-être pourrait-on suggérer quelque chose autour d'analyste des données massives ? Je définirai son profil par les trois compétences attendues. La première est statistique, avec en particulier la maîtrise des méthodes du *machine learning* et de la validation croisée. La deuxième est technologique, avec la maîtrise de tous les outils nécessaires au cycle de vie de la donnée (acquisition, stockage, préparation, modélisation, restitution et visualisation). Il faut savoir coder en R et Python, connaître le langage SQL et ses dérivés. La troisième est une compétence métier : parvenir à une compréhension des problématiques concrètes auxquelles les données doivent apporter des solutions. C'est pour cela qu'un débutant ne peut pas vraiment être *data scientist* selon moi, mais ce n'est pas forcément un point de vue partagé par tout le monde.

GS : *Je garde une question générale provocante pour la prochaine itération. Pour l'heure, je veux rester concret. Plus précisément, j'aimerais citer et discuter deux de tes publications Capgemini, que j'ai vues sur ton profil LinkedIn : « Big Data, Big Problems? Comment vous en sortir » (Auliard et Ferraris, 2013) et « Big data alchemy: How can banks maximize the value of their customer data? » (Coumaros* et al.*, 2014).*

A qui s'adressent ces publications (quel type de cadres, avec quelle formation) et quel est le message que tu veux y faire passer, en quelques mots très brefs ? Et de manière plus générale, quel est le sens pour Capgemini d'écrire de telles publications, accessibles librement : pour l'image de marque et la notoriété, afin d'attirer de nouveaux clients le moment venu ?

OA : Il s'agit d'articles de culture générale (*awareness* dans notre jargon) qui sont plutôt destinés à des dirigeants : membres du comité exécutif et cadres dirigeants, directeurs marketing, directeurs commerciaux (CXO, soit *chief experience officer*, dans notre jargon toujours).

L'objet de ces publications est d'être pédagogique, d'expliquer les possibilités et de rassurer sur la mise en œuvre, tout en donnant le type de résultats auxquels on peut s'attendre. Et en effet, l'objectif est de communiquer sur nos compétences et notre expérience, pour faire connaître au marché notre expertise et lui montrer que nos convictions sont fondées sur des expériences réelles. C'est ce qu'on appelle du rayonnement dans notre jargon.

GS : *J'en viens maintenant à la question qui me tient le plus à cœur dans cette partie : croise-t-on vraiment des données massives dans les entreprises « ordinaires » ? (Je veux exclure par là en particulier toutes les entreprises issues des réseaux sociaux.) N'est-ce pas plutôt plein de jeux de données de tailles individuellement raisonnables à croiser, plutôt qu'un seul immense jeu de données ?*

Bref, les mots de « big data » *ne sont-il pas plutôt de l'affichage et un nouveau mot à la mode* (buzz word) *ou une nouvelle posture, certes bien pratiques pour nous statisticiens qui avons toujours proclamé l'importance des données ?*

OA : C'est une très bonne question, mais pas si provocante que cela. La réponse n'est pas si simple en tout cas. Déjà, il faudrait définir ce qu'est le *big data*. La définition des 3V, voire 4V, 5V ou 6V est sans doute encore la meilleure. Volume, variété (mélanges de données structurées et non structurées) et vitesse (temps réel souhaité pour l'acquisition et le traitement, et rapidité de l'obsolescence des données) pour les 3 premiers V. On peut leur ajouter la validité (pas spécifique au *big data* à mon avis, mais on peut le prendre en disant que le travail de débruitage des données est vraiment crucial dans le *big data*) ; puis la valeur (pas vraiment spécifique non plus au *big data*, mais on se pose de plus en plus la question de la valorisation des données et de nouveaux *business models* liés aux données) ; et enfin, la visualisation (les outils d'exploration des données et de restitution prennent une place majeure).

Si l'on s'en tient au premier V, c'est-à-dire au volume, il faut constater que la réalité de l'explosion du volume des données a deux causes majeures. La première, c'est simplement que les images et surtout les vidéos sont effectivement beaucoup plus volumineuses que les données structurées. On peut dire que YouTube, les *selfies* et autres vidéos contribuent largement à l'explosion des volumes. Toutefois, l'intérêt *business* de ces nouvelles données reste assez limité, encore que... La deuxième cause, c'est la multiplication des sources de données, avec les objets connectés qui transmettent des données horodatées, localisées, avec une fréquence parfois très grande (on parle d'IoT, *Internet of Things*). Là, on est beaucoup plus proche du *big data* dans son acceptation classique.

Maintenant, qu'en est-il dans les entreprises ? Au niveau individuel, cela dépend largement du niveau de maturité vis-à-vis des données, mais si on raisonne de manière globale, il est clair que dans tous les projets que l'on a pu mener, la première marche à franchir, et celle qui donne de fait les meilleurs résultats pratiques, c'est avant tout le « désilotage » des données[7], et donc le croisement des données « ordinaires ». Malgré tout, ces données même dites « ordinaires » peuvent être parfois assez complexes, notamment quand on récupère tout ce qui relève des contacts de l'entreprise avec ses clients : on a souvent du texte, parfois de la voix, quelques fois des images, et ces données sont souvent très riches et utiles. (On va les traiter par OCR, *optical character recognition*, soit reconnaissance optique de caractères.) On utilise aussi assez régulièrement des données externes. Les données de l'INSEE sont typiquement très pertinentes pour enrichir des données de profil. On peut même recourir à des bases de données payantes sur les entreprises : données financières (Orbis, Altares) ou éléments descriptifs des grandes surfaces (Panorama, LSA). En cela, on n'est finalement pas si loin du *big data*, même si les volumes ne sont pas toujours exponentiels : récupérer des jeux de données qui dépassent les 10 To n'est pas si fréquent pour nos clients traditionnels, mais cela arrive, ... souvent d'ailleurs parce que les systèmes de collecte ou sto-

[7] GS : Le partage et le croisement de données issues de différents services d'une même entreprise ; c'est en un sens une expression de transversalité à l'aune des services internes.

ckage de nos clients sont mal organisés. En revanche, dès que l'on est sur des opérateurs du Web ou d'objets connectés, là les volumes deviennent importants. En ce qui concerne les données issues des réseaux sociaux, elles sont rarement exploitées à un niveau individuel aujourd'hui, mais plutôt à un niveau agrégé pour faire de la veille. Des applications commencent à émerger pour identifier des prospects, ou voir comment des idées de fraude se propagent par exemple ; mais ces applications se positionnent plus rarement sur du rebond commercial par exemple, même s'il y a des solutions techniques et pratiques qui arrivent, à croiser avec les éléments juridiques de protection de la vie privée...

En résumé, sur les données en elles-mêmes, on pourrait dire que le *big data* dans les entreprises est l'occasion de mieux exploiter les données déjà collectées, en les enrichissant ponctuellement de sources externes, et en y intégrant des analyses de données non structurées. Mais clairement, la tendance est à une plus grande maturité des entreprises : quand elles ont goûté aux données (et à condition de les accompagner de visualisations adéquates), elles en veulent davantage. En 2013–14, nos projets étaient avant tout des POC (*proofs of concept*), nous sommes aujourd'hui plutôt passés à des accompagnements de « *Data Lab* » internes à nos clients, ou à des problématiques de déploiement.

En revanche, si l'on parle des méthodes, là, les choses ont vraiment bougé très fort. Tout d'abord parce que, dans les problématiques supervisées en tout cas, le nombre de variables potentiellement explicatives a vraiment crû de manière très grande ; également parce que les risques de sur-ajustement sont de plus en plus grands ; et enfin parce que l'on s'intéresse maintenant à des événements de plus en plus rares. C'est ainsi que les techniques de *machine learning* sont en plein essor, la validation croisée est un standard pour calibrer les hyper-paramètres, et le *deep learning* émerge réellement...

Recrutements et besoins de formation

GS : *Commençons très concrètement :* http://info.dataiku.com/data-challenge/*, c'est toi ? Cette compétition est-elle l'occasion de bien recruter ?*

OA : Oui, c'est bien nous... On espère bien en profiter pour recruter effectivement.

GS : *C'est un peu court pour un format de recrutement aussi original ! Tu penses bien que j'aimerais en savoir davantage (et nos lecteurs sûrement aussi). Je reproduis ci-dessous le descriptif de l'événement. J'ai l'impression que ce genre d'événements est au recrutement de futurs* data scientists *ce que les télé-crochets sont au démarrage de la carrière de chanteurs désormais très reconnus, non ?*

> Extraits de la page http://info.dataiku.com/data-challenge/
>
> Pour tous les data scientists en devenir, Capgemini Consulting et Dataiku s'associent pour vous organiser un hackathon[8]. L'événement aura lieu le jeudi 7 avril de 12h30 à 19h30, dans les locaux de Télécom Paris [...]. Vous travaillerez sur un cas *business* réel et des données

[8]GS : Selon Wikipédia, consulté le 11 mai 2016, « un hackathon est un événement où des développeurs se réunissent pour faire de la programmation informatique collaborative, sur plusieurs jours. Le terme est un mot-valise constitué de hack et marathon. » Mais qu'est-ce que le « hack » ? C'est « une manipulation d'un système, de l'anglais *to hack*, tailler, couper quelque chose à l'aide d'un outil. »

fournies par Bouygues Telecom. Chaque groupe aura l'après-midi pour imaginer et prototyper une application en utilisant l'outil Dataiku Data Science Studio : modèle prédictif, tableau de bord analytique, visualisation interactive, etc.

L'objectif ?

- S'approprier le contexte *business* et les *datasets*
- Préparer les *datasets* et enrichir les données
- Créer et améliorer des modèles de *machine learning*
- Restituer les résultats

OA : C'est la première fois que nous, Capgemini Consulting, organisons un *challenge* avec Dataiku comme partenaire, mais en revanche, nous organisons des *data science games* plusieurs fois par an, généralement dans des écoles. Ces événements ont un caractère international.

En ce qui concerne le recrutement, nous avons des écoles cibles chez Capgemini Consulting (tant pour les *data scientists* que pour les consultants plus classiques). Ces *challenges* nous permettent alors de recruter différemment, hors du vivier formé par ces écoles cibles, mais au vu de performances concrètes en termes d'agilité à rentrer dans une problématique métier et à préparer les données, deux points absolument cruciaux dans le futur métier à exercer... et que l'on n'apprend pas dans les livres !

GS : *Justement, regardons plus en profondeur le profil de vos recrues. Tout d'abord, quels types de statisticiens (de quels niveaux) recrutes-tu ? Quelles notions, quelles compétences les statisticiens doivent-ils posséder pour être utiles à Capgemini Consulting (logiciels, méthodologies, contact ou pas avec la recherche, etc.) ?*

OA : Nous recrutons au niveau master minimum (pas au niveau licence). Le vivier traditionnel est formé par les écoles (ENSAE ParisTech et ENSAI, par exemple, et d'autres écoles d'ingénieur avec filière de formation en *data science*) ; nous recrutons également à la sortie de masters universitaires, comme les mentions MVA (mathématiques, vision, apprentissage) et *data science* du master de mathématiques et applications de l'Université Paris-Saclay[9], ou après une thèse.

En termes de compétences, il nous faut des connaissances et une pratique du *machine learning* (ce qui nécessite donc également la maîtrise des fondements de la statistique), et nous apprécions fortement que les candidats aient manipulé de vraies données. En termes de logiciels et langages, je pense à R et Python (il suffit à la rigueur d'un des deux), SQL, idéalement les outils du *big data* (Hive, Pig, Spark) et JavaScript. Les contacts avec la recherche sont un plus.

Je mentionne également que nous recrutons aussi des profils de même niveau en informatique.

GS : *Peux-tu nous dire (beaucoup plus brièvement) ce qu'il en était chez TNS Sofres ou ASTEROP ?*

OA : Il fallait plutôt maîtriser, côté connaissances statistiques, l'analyse des données, et SAS et SQL, côté outils.

[9]GS : Les masters MVA et *data science* étaient auparavant des masters indépendants, de plein exercice, portés respectivement par l'ENS Cachan et des écoles de ParisTech. A la rentrée universitaire de 2015–16, ils ont tous été intégrés à l'offre de formation de l'Université Paris-Saclay : à son unique master de mathématiques et applications, dont ils forment des mentions, c'est-à-dire des parcours de formation. En pratique, ces parcours de formation sont gérés indépendamment (en termes de recrutement d'étudiants par exemple), avec toutefois une forte mutualisation des cours et une grande souplesse pour faire valider dans un parcours les cours d'un autre parcours de ce master de mathématiques et applications.

GS : *Tu nous as dit quelles compétences tu aimais que les futurs ex-étudiants aient. Peux-tu nous donner en miroir ton point de vue critique sur la formation statistique à la française (bons points et mauvais points éventuels) ? En clair, que devrions-nous enseigner dans nos masters orientés formation professionnalisante, et comment devrions-nous l'enseigner ? Quelles éventuelles lacunes constates-tu dans les parcours de formation actuels ?*

OA : Dans les bons points, et c'est souvent le cas des formations à la française, les étudiants acquièrent des fondements théoriques solides et du recul sur l'ensemble des techniques disponibles.

En revanche, ce qui manque à mon sens, c'est une vision un peu plus complète du cycle de vie de la donnée. En particulier, la bonne compréhension des phases en amont (acquisition, stockage et préparation des données, choix et conception d'une méthode) est largement sous-estimée. En pratique, ces phases sont parfois à la fois longues et rébarbatives, mais souvent déterminantes dans les modèles qui suivent. Les cas traités (et certes, ils sont plus fréquents) sont souvent trop pré-digérés (données déjà assez propres et méthode cadrée).

Dans les ajouts éventuels, je trouverais pertinent d'utiliser une approche un peu plus orientée vers des problématiques classiques et d'offrir une ouverture sur les techniques de visualisation des données. La formation sur les outils de manipulation des données (SQL, Hive, Pig, Spark, MongoDB, NoSQL, etc.) serait intéressante.

GS : *Merci pour cette réponse très riche ! Je voudrais revenir sur plusieurs points. Tout d'abord, qu'appelles-tu « problématiques classiques » ?*

OA : Par « problématiques classiques », j'entendais des sujets auxquels nous, praticiens, sommes confrontés tous les jours, par exemple la détection de durées ou de moments de vie, la maintenance prédictive, le risque de défaut, la propension à acheter un produit ou à adhérer à un service, etc. Bref, je parlais de problématiques métier.

GS : *Pour les bons points des formations à la française, est-ce qu'il faut comprendre que ces futurs professionnels seront du coup bien armés pour la phase pour laquelle tu disais qu'on n'avait jamais assez de temps, à savoir la modélisation ? Pour les mauvais points, comment pouvons-nous les corriger ? Faudrait-il créer des bases de données brutes ouvertes* (open raw data) *?*

OA : Oui, cela me semble une solution viable. Je pense que pas mal d'entreprises sont prêtes à fournir leurs données : elles le font lors des compétitions, à destination des étudiants comme celles dont nous parlions précédemment ou celles auxquelles participent les professionnels, notamment sur Kaggle. Ces données ne sont pas toujours propres, et c'est tant mieux. Ma conviction est qu'il doit être possible de récupérer de tels jeux de données brutes et j'invite les universitaires à solliciter les entreprises qu'ils voient organiser activement de telles compétitions. Nous, Capgemini Consulting, ne pouvons pas fournir de données, car les données que nous exploitons ne sont pas nos données mais celles de nos clients. Toutefois, je pense que plusieurs de nos clients seraient d'accord pour procurer leurs données, après anonymisation (et éventuels changements d'échelle, et autres traitements n'affectant pas le sel des jeux de données en question mais évitant toute exploitation au profit de concurrents).

Une autre source possible de tels jeux de données brutes sont les sites de données publiques, peut-être pas celles de l'INSEE, souvent assez nettoyées, mais celles d'un site plus généraliste et plus fourre-tout, comme https://www.data.gouv.fr/.

GS : *Pour la formation sur les outils de manipulation des données, tu citais SQL, Hive, Pig, Spark, MongoDB, NoSQL, etc. Or, l'enseignement au moins de SQL et NoSQL me semble devenir un standard, en tout cas dans les parcours spécifiques* big data[10]...

OA : Je voulais en réalité souligner que cette formation à ces outils avait vocation à être donnée dans tout cursus statistique généraliste (en école d'ingénieur notamment), qu'il y ait ou non une spécialisation en *data science* ou *big data*. Bref, je disais mon souhait que cette formation devienne la règle et non une niche. En effet, comment faire de la statistique aujourd'hui sans comprendre ces outils ?

GS : *Sur un autre plan : tu n'as pas encore parlé des parcours STID*[11] *proposés par les IUT dans le cadre de leurs licences à visée professionnelle. Aucune des entreprises où tu as exercé ne semble intéressée par leurs profils, mais que peux-tu dire de ces formations et des besoins qu'elles couvrent dans l'industrie et les services ?*

OA : Pour ma défense, je ne connaissais pas ces formations quand j'étais chez TNS et ASTEROP, donc effectivement c'était plus difficile de les recruter. Cela aurait pu être pertinent dans les deux cas, en tout cas. Pour Capgemini Consulting, c'est un peu plus complexe, car comme je l'ai déjà expliqué, le niveau de recrutement est vraiment très élevé (aussi pour s'aligner sur le niveau de recrutement des consultants classiques) : en sortie d'école, ou *a minima* après un master, du moins pour les débutants. Je pense en revanche que pour les autres entités de Capgemini, ces recrutements devraient être envisageables. Par conséquent, je suis assez mal placé pour répondre... Toutefois, je suis assez persuadé que ces formations répondent à des besoins concrets des entreprises, typiquement pour des profils d'analystes. Cela implique qu'il faut vraiment insister dans la formation sur la préparation et le nettoyage des données.

Tes propres enseignements

GS : *Commençons par une question générale, pour m'aider à situer cette partie. Quels enseignements et formations donnes-tu ou as-tu donnés ces dernières années ? Je sais qu'il y a un module dans une formation continue de données massives à l'IUT Paris Descartes. (C'est d'ailleurs ainsi que j'ai pensé à toi, en lisant l'article correspondant de ce volume de notre revue). Y en a-t-il d'autres ? Que ce soit en formation initiale et continue à l'université ou dans des écoles, ou que ce soit en formation professionnelle en interne à Capgemini ou en externe auprès d'un organisme de formation professionnelle...*

OA : Les premiers enseignements que je citerai sont en effet ceux que je donne à l'IUT Paris Descartes, au titre des parcours STID ; l'intervention dans la formation continue *big data* que tu mentionnes mais également un cours de géomarketing en formation initiale. En complément de cela, j'effectue également des interventions ponctuelles de type séminaire dans des masters, comme le master spécialisé *big data* de Télécom ParisTech.

Dans un passé un peu plus lointain, j'ai donné des cours à l'ENSAI, en marketing quantitatif et géomarketing, et encore plus loin, au CNAM[12], un cours d'analyse des données dans ce qu'on appelait alors un DESS[13].

[10] GS : Voir à ce sujet divers articles de présentation de cursus *big data* dans ce numéro de la revue.
[11] STID – statistique et traitement informatique des données.
[12] Conservatoire national des arts et métiers, au siège de Paris.
[13] Diplôme d'études supérieures spécialisées.

Voilà pour la formation académique. En ce qui concerne la formation professionnelle, j'en assure bien évidemment en interne au sein de Capgemini. Mes interventions sont centrées en *analytics* (recherche analytique), c'est-à-dire que je présente les concepts et méthodes fondamentaux de l'analyse descriptive, de l'analyse prédictive, du *machine learning*, ou encore de la prévision de séries chronologiques (notamment, des processus auto-régressifs). Il peut également m'arriver d'animer des sessions d'*awareness*, où il s'agit de rendre conscients des cadres dirigeants des opportunités offertes par la sciences des données. Ces formations ont lieu en général dans nos locaux de La Défense, plus rarement au vert.

GS : *Suis-tu encore de manière régulière et prolongée des étudiants ou apprenants ?*

OA : Dans mes enseignements actuels en STID, les interventions sont regroupées : en une journée de 8h pour le diplôme d'université *big data* et en une journée de 8h par semaine sur un mois pour le cours de géomarketing. J'aurais donc tendance à dire non... si ce n'est peut-être l'encadrement de projets de recherche appliqués au titre de l'EN3S[14].

GS : *Comment définirais-tu ton style d'enseignement ? Qu'aimes-tu développer en tes étudiants ?*

OA : J'aime les faire réagir, stimuler un échange. Je tâche d'être à la fois concret et rigoureux ; par concret, j'entends l'illustration de mon propos par des exemples tirés de ma vie professionnelle. Par exemple, j'aime rendre actifs les étudiants en leur faisant utiliser une API de Google pour géo-coder un fichier.

GS : *Cet exemple, en tout cas, me fait réagir et va stimuler un échange : peux-tu me décoder cette phrase ?*

OA : L'objectif est de transformer une adresse physique (du type postal : « 11 rue Pierre et Marie Curie, 75 005 Paris ») en coordonnées géographiques (par exemple, latitude et longitude, mais pas uniquement) à insérer dans un référentiel. Il existe des outils, notamment une boîte à outils fournie par Google, pour géo-coder des adresses à la volée et les représenter sur une carte à partir d'un fichier d'adresses.

GS : *Je voudrais te poser maintenant une question commune à tous les entretiens. Quelle est la notion que tu trouves la plus difficile à enseigner (pour toi) ou à recevoir (pour les étudiant-e-s) ?*

OA : Ma première réponse concernerait la construction des variables et l'appréhension du jeu de données. Parfois, les étudiants font tous les mêmes erreurs. L'exemple suivant est frappant. J'utilise R dans mes cours de géomarketing. Les valeurs manquantes y sont codées par NA (pour *not available*). Je procure un fichier à géocoder, contenant de nombreuses telles valeurs manquantes. Les étudiants le traitent et surprise, obtiennent une carte centrée sur... la Namibie (de code pays international NA). Ils se demandent alors tous : « Pourquoi suis-je là ? » Maintenant, c'est presque un jeu un peu cruel pour moi que de leur fournir ce jeu de données, car je sais à l'avance qu'ils ne le regarderont pas avec attention et le traiteront de manière aveugle, en tombant sur un résultat surprenant en apparence !

GS : *Une question me vient à l'esprit : tu utilises R dans tes cours, bien entendu, c'est un langage académique en un sens, mais à Capgemini, au quotidien, l'utilisez-vous également ?*

OA : Oui ! Parce qu'on effectue essentiellement du prototypage (les « proofs of concept » dont je parlais ci-dessus), et assez peu de déploiement.

[14] Ecole nationale supérieure de sécurité sociale, sise à Saint-Etienne, qui recrute et forme les futurs cadres dirigeants du service public de sécurité sociale.

GS : *Pour en revenir aux défis et difficultés liés à l'enseignement, en vois-tu d'autres ?*

OA : Une tranche de vie de cours typique me frustre : quand je répète *n* fois la même chose (manière de procéder, point d'attention, mise en garde), avec *n* grand, et qu'au moment d'appliquer en pratique ce que j'ai tant répété, les questions que se mettent subitement à poser les apprenants portent pile sur ce point.

GS : *N'est-ce pas une question de* timing *et de plus grande expérience de la part de l'enseignant ? Les apprenants n'ont pas encore réalisé la difficulté au moment où elle est signalée, ils ont besoin d'y buter pour devenir réceptifs.*

OA : Oui, c'est exactement cela. Je reformulerais donc le défi posé à l'enseignant comme l'anticipation des difficultés rencontrées par les apprenants.

GS : *Faut-il vraiment anticiper ces difficultés ? La question est peut-être la bonne synchronisation sur le rythme de maturation des idées dans la tête des apprenants...*

OA : C'est vrai. Pour en revenir à la science des données, il est important de mettre les mains « dans le cambouis », c'est cela qui éveille et suscite les questions, et permet un dialogue fructueux entre l'enseignant et les étudiants. Les fameux échanges que j'aime stimuler, ... et pour lesquels toutes les parties doivent être mûres !

GS : *As-tu quelque chose à ajouter pour clore cette partie sur tes enseignements ?*

OA : Je voulais dire que malgré mes évolutions de carrière, j'ai toujours tenu à garder un pied, un contact avec l'enseignement.

GS : *Pourquoi ? Pour nous, universitaires, l'enseignement est un moyen de contact avec la réalité ; mais pour toi, qui es en contact permanent avec la réalité des données, qu'y trouves-tu ?*

OA : C'est une autre forme de contact avec la réalité, cela aère l'esprit !

Evolutions dans le monde de l'enseignement et du recrutement

GS : *Je voudrais conclure cet entretien par une question générale : comment vois-tu le futur ? Tu as déjà beaucoup parlé des évolutions en cours dans le monde de l'entreprise. Mais lesquelles attends-tu dans notre manière d'enseigner, par exemple ?*

OA : Est-ce que le métier d'enseignant va subsister ? Face à un MOOC[15] donné par un enseignant formidable (et des vedettes des MOOC vont nécessairement émerger), quelle est la place d'un enseignement traditionnel en face à face ? J'ai un ami universitaire (en école) qui se sent menacé. Quel est l'intérêt d'un cours magistral ?

GS : *Beaucoup de collègues passent en pédagogie inversée : les étudiants travaillent le cours par eux-mêmes, dans un polycopié ou écoutant des enregistrements vidéographiques, et les séances avec les enseignants sont interactives, et permettent de poser des questions, résoudre des exercices d'application, etc. Tout cela est fait dans l'optique de remplacer le format cours magistral en présentiel et exercices d'application à la maison, qui malgré tout, reste dominant...*

OA : C'est une évolution pertinente. Sur un autre plan, j'avoue ne pas craindre une « uberisation » du métier d'enseignant. Ce qui est et sera important dans l'exercice de ce métier, c'est la capacité

[15] *Massive open online course*, cours en ligne ouvert et massif.

à accompagner les apprenants, à réagir rapidement et de manière personnalisée sur les erreurs commises dans les mises en œuvre pratiques.

J'aimerais également soulever un dernier point, qui est la démographie des apprenants. Il y en aura de plus en plus de tous âges : au-delà de la formation initiale, nous observons et observerons toujours davantage des changements de carrière à des degrés divers, pouvant aller jusqu'à des reconversions. La formation continue se développe et se développera, les collaborateurs à recruter ne seront pas uniquement jugés sur leur formation initiale.

J'observe déjà une évolution dans le profil et le curriculum vitæ des postulants confirmés chez Gapgemini : ils ont suivi des MOOC (sur les données massives ou en *machine learning*) et ils précisent les certificats de participation ainsi obtenus. Ces derniers sont des marqueurs de motivations et reflètent l'envie et la capacité à se remettre en question, et à être pro-actif dans la trajectoire de formation et de conversion.

Je trouve le parcours de Nicolas Gaude inspirant : de formation initiale en physique, il s'est formé tout seul en science des données, est désormais *chief data scientist* à Bouygues Telecom, et obtient de très bonnes performances dans les compétitions Kaggle[16]. (Aujourd'hui, 22 avril 2016, il est dans les 150 meilleurs des plus de 520 000 compétiteurs de Kaggle.) Ces trajectoires de reconversion sont appelées à se multiplier.

Il y a peut-être un effet de mode pour la donnée, qui pourra un peu passer, mais je pense que malgré tout, une forte dépendance à la donnée (au sens de l'addictivité) est en train de se mettre en place.

GS : *Cette dépendance ne serait-elle pas malheureusement également associée à des illusions ? Les données ne peuvent pas aider à tout prévoir...*

OA : En effet, et à Capgemini Consulting, nous avons gagné des projets en tenant un discours de vérité : « Nous ne pouvons pas faire ce que vous nous demandez, mais nous vous proposons telle autre chose ».

GS : *Merci, Olivier, pour tout le temps que tu nous a consacré. J'ai pu me rendre compte en te rejoignant au siège de Capgemini Consulting à quel point tu étais sollicité, et je n'en suis que plus reconnaissant pour ces temps d'échange par courriels et en face à face.*

La troisième édition de cette chronique, dans le numéro de fin d'année de notre revue, recueillera à nouveau le témoignage d'un universitaire, pour continuer l'alternance praticien-universitaire.

Références

[1] Auliard, O. (2009), Le géomarketing décisionnel, une activité encore marginale mais en fort développement, *Variances, la revue des anciens de l'ENSAE*, **36**, consultable en ligne.

[2] Auliard, O. et P. Ferraris (2013), Big Data, Big Problems? Comment vous en sortir, *Journal of Marketing Revolution*, **1**, 16–19, site de cette revue interne à Capgemini Consulting.

[3] Breiman, L. (2001), Random forests, *Machine Learning*, **45**(1), 5–32, article librement téléchargeable.

[16] https://www.kaggle.com/

[4] Cochran, W. G. (1953), *Sampling Techniques*, John Wiley & Sons, New York.

[5] Coumaros, J., S. de Roys, L. C. Leroyer, J. Buvat, et O. Auliard (2014), Big data alchemy: How can banks maximize the value of their customer data?, rapport pour Capgemini Consulting, consultable en ligne.

[6] Vapnik, V. N. (1998), *Statistical Learning Theory*, John Wiley & Sons, New York.