

# Le *Big Data*, au fond, qu'est-ce que ça change ? Après le centième Café de la statistique



Jean-François ROYER

SFds<sup>1</sup>

Le 27 mars 2017 s'est tenu le centième « Café de la statistique ». Il était consacré au thème omniprésent aujourd'hui : « Big Data : Big Science ? Big Brother ? ». Il a rassemblé 130 participants autour de l'invitée, Valérie Peugeot, sociologue et commissaire de la Cnil. J'étais l'un d'eux : voici un témoignage personnel.

En observatrice de la mutation numérique, Valérie Peugeot nous a d'abord appelés à reconnaître la nouveauté que le *Big Data* apporte dans la société. Certains, devant des promesses mirobolantes, parfois suivies de déceptions, peuvent se demander s'il ne s'agit pas d'une « bulle » médiatique. Non – des pratiques réellement nouvelles, réellement liées à l'exploitation des données massives sont attestées dans de nombreux secteurs, et pas seulement le marketing commercial ou politique. La médecine, l'agriculture, la finance et l'assurance, l'urbanisme ont été cités comme exemples. Dans le champ des connaissances, astronomie, écologie, géographie et même analyse littéraire obtiennent des résultats grâce aux données massives. Et des listes complètes seraient certainement plus longues.

S'agit-il d'une « révolution technologique » ? Valérie Peugeot est très méfiante devant le « déterminisme technologique ». Les nouvelles possibilités de la communication et du calcul ne contiennent pas dès le début leurs capacités « révolutionnaires » qui se déploieraient indépendamment des influences sociales. Il faut récuser toute « prédestination » et constater les allers-retours permanents entre technique et économie, entre conception et appropriation sociale. Les points de contact, ou de friction, entre techniques nouvelles et pratiques sociales méritent d'être observés pour eux-mêmes, et aussi pour agir dans la société nouvelle. L'invitée est impliquée dans une action « citoyenne » pour promouvoir les usages du numérique qui respectent les droits des personnes et pour combattre les menaces que les nouvelles technologies font courir aux libertés publiques<sup>2</sup>.

Le débat qui a suivi l'exposé initial de Valérie Peugeot a pendant deux heures permis de développer ces différents points. Beaucoup d'interventions prolongeaient dans le sens de l'inquiétude les analyses de l'invitée quant à l'usage des données individuelles dans le monde qui vient. Les craintes pour soi-même sont vives ; les craintes pour la société le sont encore plus.

1. Les Cafés de la statistique sont organisés à Paris depuis 2005 par le groupe « Statistique et enjeux publics » de la SFds, pour examiner « si la statistique éclaire les questions de société ». Chaque séance est introduite par un invité ou une invitée, et se poursuit par un débat. Des comptes rendus écrits sont rédigés ; depuis 2012, ces comptes rendus sont complétés par des vidéos. Textes et vidéos sont téléchargeables à partir du site de la SFds, rubrique Cafés de la statistique.

2. Outre ses fonctions à la Cnil, elle préside l'association « Vecam – Citoyenneté dans la société numérique »

Les réglementations protectrices des individus ne cessent de se renforcer<sup>3</sup> ; et il est possible de donner à chacun de bons conseils pour assurer la sécurité de ses propres données<sup>4</sup>. Mais sur le plan collectif, la perspective d'une individualisation de plus en plus poussée des « traitements » appliqués aux personnes par des institutions de plus en plus gigantesques (les « GAFAs<sup>5</sup> », les grands États) laisse craindre un fonctionnement de plus en plus atomisé de la société, voire oppressif, face auquel on discerne mal comment des « corps intermédiaires » pourraient émerger et devenir des recours. C'est sans doute là qu'on touche de plus près la menace de « révolution » contenue dans le *Big Data*.

Le public du Café comportant beaucoup d'étudiants et de professionnels de la statistique, la question épistémologique n'a pas été absente du débat. Rupture, ou pas rupture ? La tonalité dominante était là rassurante : les anciennes méthodes ne seraient pas supplantées par celles qui traitent des données massives, mais seulement appelées à se renouveler ; le discours selon lequel le *Big Data* signerait la fin des hypothèses et des théories serait déjà dépassé. Pourtant, dans bien des contextes scientifiques, la recherche des meilleures prédictions, facilitée par l'abondance des données, n'est-elle pas privilégiée par rapport à la recherche sur les mécanismes sous-jacents aux observations, recherche qui est plus aléatoire ? On peut s'en inquiéter. Mais le débat a surtout porté sur l'accompagnement nécessaire du développement des traitements de données massives. Le souci de la qualité des données doit rester présent même si elles sont recueillies « au vol ». La formation des étudiants en statistique se transforme presque partout, pour faire place aux nouvelles compétences théoriques et pratiques requises pour traiter les données massives. Et l'invitée a décrit de façon positive deux innovations dans le domaine de la connaissance scientifique : le recours à des collaborations ouvertes nombreuses (« science participative ») et les droits nouveaux accordés à la fouille des gisements de données pour la recherche, fût-ce au détriment des droits de propriété (droits d'auteurs, brevets).

La statistique publique peut profiter du *Big Data*. Elle est une possible « consommatrice » de gros fichiers produits par des acteurs privés, qui lui permettraient d'économiser sur ses modes de collecte traditionnels<sup>6</sup>. Encore faut-il qu'elle sache s'en emparer, et que ces nouvelles sources garantissent une production régulière et de qualité. Ce point de vue a été largement exprimé au cours du Café. Un autre angle a été moins développé : et si l'ère du *Big Data* venait délégitimer beaucoup de productions de la statistique publique, en proposant des informations concurrentes plus rapides, plus détaillées, moins « étatiques » ? La menace<sup>7</sup> d'une élimination par la concurrence ne semble guère ressentie : les statisticiens publics restent dans l'ensemble confiants dans la protection que leur donne la réputation actuelle de leurs productions. En revanche, beaucoup se préoccupent du trouble que de telles informations, si elles étaient de qualité insuffisante, pourraient jeter dans les débats publics. Certains participants n'hésitent pas à suggérer une « charte universelle » de la diffusion d'informations publiques issues de traitements de données. Une sorte de « label rouge » des traitements de données massives destinés à être versés à un débat public et irréprochables sur le plan méthodologique, voire éthique...

Oui, le *Big Data*, « ça change beaucoup de choses » ! Mais caractériser ce changement reste difficile, et on est loin d'en pouvoir mesurer « statistiquement » les effets. Le centième Café de la statistique a parcouru ce vaste domaine : puisse-t-il être suivi de beaucoup d'autres pour débattre des mesures et des solutions.

- 
3. Voir dans ce numéro l'analyse par Judith Rochfeld des dispositions de la « loi numérique » française et du nouveau règlement européen sur la protection des données
  4. A été cité l'ouvrage de Tristan Nitot « Surveillance:// Les libertés au défi du numérique » C&F Editions 2016 – 19€
  5. Google – Amazon – Facebook – Apple
  6. Voir dans ce numéro les exemples donnés par Dominique Bureau et par Gilbert Saporta
  7. Voir dans ce numéro l'article de William Davies pour « The Guardian »