Les Rencontres Scientifiques Colas

“BIG DATA : comment gérer l’explosion des données numériques ?”

Mardi 28 octobre 2014

avec **Francis BACH**
Directeur de Recherche INRIA à l’École Normale Supérieure

et **François BANCILHON**
Président-Directeur Général de la société Data Publica

Conférence animée par un journaliste de La Recherche

Intervention de Francis BACH, Directeur de Recherche INRIA à l'École Normale Supérieure

Chaque minute dans le monde, on comptabilise quelque 204 milliards de mails envoyés, 2 millions de recherches sur Google, plus de 680 000 échanges de contenus sur Facebook, 100 000 tweets, 48 heures de vidéos téléchargées sur YouTube, 571 nouveaux sites web… ou bien encore plus de 272 000 dollars d'achats en ligne !

Tout ceci génère d'énormes masses de données numériques, certaines étant déjà exploitées pour créer des publicités et des recommandations d'achat personnalisées. Mais les scientifiques contribuent aussi à l'avalanche de données. C'est par exemple le cas en bioinformatique, avec l'étude des quelque deux millions de protéines chez l'Homme qui génère des données massives et complexes ; ou bien encore en astronomie : le projet de télescope Square Kilometer Array prévu pour 2024 engendrera par exemple 109 Go de données… par jour !

Les données numériques sont donc non seulement de plus en plus volumineuses mais aussi omniprésentes, hétérogènes, et à flux rapide. Pour pouvoir les exploiter, avoir des ordinateurs toujours plus puissants ne suffit plus. Il faut développer de nouveaux algorithmes, avec un défi de taille : que le temps de calcul augmente moins vite que la masse de données à traiter. Voilà pourquoi on s'inspire aujourd'hui d'algorithmes des années 1950, à l'époque où les ordinateurs peu puissants nécessitaient des algorithmes peu gourmands en temps de calcul. L'idée est aussi de développer des outils capables de faire des prédictions en prenant en compte des données observées dans le passé ; c'est ce que l'on nomme le "machine learning" ou "apprentissage statistique".

Intervention de François BANCILHON, PDG de la société Data Publica

Aucun secteur économique n'est à l'abri de la révolution du Big Data qui bouleverse des pans entiers de l'économie, avec de nouveaux acteurs intermédiaires possédant des données stratégiques. Dans la presse, Apple est par exemple la seule à détenir les informations sur les abonnés au monde.fr sur Ipad (profil, ce qu'ils lisent, comment ils naviguent…), de quoi influencer la ligne éditoriale. Dans le tourisme, la société de réservation d'hôtels booking.com accumule aussi des données stratégiques sur les clients, utilisables pour imposer certaines conditions aux chaînes d'hôtels. Dans la vidéo à la demande, NetFlix sait à quelle heure tel téléspectateur regarde une série, à quel moment il zappe… de quoi peser sur le scénario des prochaines saisons. Quant à Google - qui enregistre un nombre croissant de données sur nos comportements et nous suit à la trace -, il pourrait bien court-circuiter les assureurs.

Mais l'enjeu est aussi stratégique et géopolitique. Aujourd'hui, les données des utilisateurs européens des réseaux sociaux majeurs sont détenues par des plateformes basées aux États-Unis. Une erreur que n'ont pas commise la Chine et la Russie. Se pose aussi la question de l'équilibre entre transparence et données privées. En France, la CNIL empêche par exemple les entreprises françaises d'analyser les tweets : c'est un combat de type ligne Maginot car certains acteurs comme LinkedIn possèdent déjà ces données personnelles ! En revanche, la création de la plateforme ouverte de données publiques françaises Etalab doit être saluée car elle dénote une vraie prise de conscience des dirigeants du pays.