Google finance une intelligence artificielle pour le Big Data

Les inquiétudes du physicien Stephen Hawking sur le potentiel de l'Intelligence Artificielle – il y voit une menace pour l'Humanité – ne ralentissent pas Google. Le géant californien finance un projet baptisé Automatic Statistician qui vise à **développer une forme d'intelligence artificielle pour la data science**, soit l'analyse des Big Data. En résumé, le projet, qui en est encore à ses stades de recherche préliminaire au sein de l'université de Cambridge, vise à **automatiser les tâches essentielles du statisticien** : le choix du modèle statistique le plus pertinent pour un jeu de données, ainsi que celui des variables permettant de dégager une analyse de la masse d'information.

Dans un billet de blog annonçant le financement de Google, Kevin Murphy, scientifique travaillant pour la première régie publicitaire online, explique les freins que cette approche vise à lever : « *Le premier problème est que les méthodes actuelles de Machine Learning requièrent toujours une expertise humaine considérable dans la conception des fonctions et modèles. Le second problème est que le résultat des méthodes actuelles, bien qu'exact, est souvent difficile à comprendre, donc à croire.* » L'objectif du projet Automatic Statistician est précisément de s'attaquer à ces deux difficultés « *en utilisant des stratégies de sélection de modèles Bayésiens pour choisir les bons modèles et fonctions et interpréter le résultat de façon compréhensible, dans des* ***rapports générés automatiquement mais lisibles par un être humain*** ».

Une première version pour tester

Le site du projet offre quelques exemples de ces statistiques robotisées. Mais sur des jeux de données relativement simples, des séries temporelles comme l'évolution du nombre de passagers aériens ou celle des rayonnements du soleil. Kevin Murphy explique que cette approche est en passe d'être généralisée « *pour trouver des modèles dans d'autres types de données, comme des problèmes de régression multidimensionnelle ou des bases de données relationnelles* ». Une **version simplifiée du système est disponible** depuis août dernier, signale Google. Elle permet d'importer un set de données et de recevoir une analyse produite automatiquement quelques minutes plus tard. Une version plus riche de ce service est attendue début 2015.

The raw data and full model posterior with extrapolations are shown in figure 1.

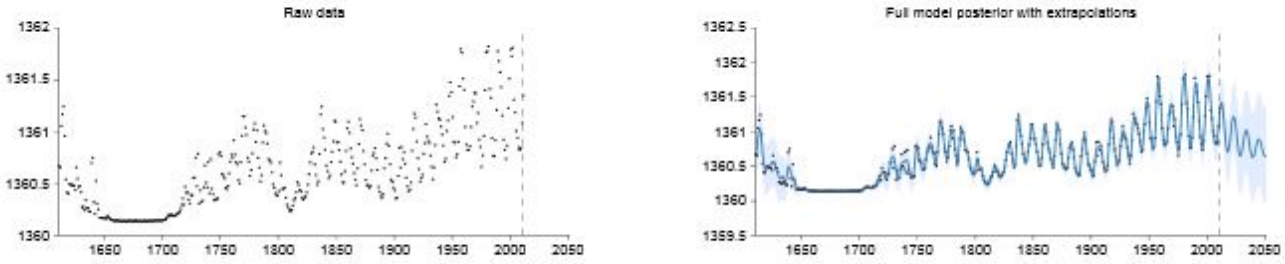

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified eight additive components in the data. The first 4 additive components explain 92.3% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 6 additive components explain 99.7% of the variation in the data. After the first 5 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

Comme le signalent nos confrères de GigaOm, Google n'est pas la seule société à s'essayer à ce type d'approches visant à dépasser le manque de compétences en statistiques sur le marché. Une start-up, Skytree, a bâti un outil similaire (Adviser), aujourd'hui abandonné. Et d'autres jeunes pousses s'intéressent également à des fonctionnalités similaires.

A lire aussi :
Le Machine Learning vient au secours du Big Data
Recruter un data scientist ? Bienvenue au Far-West

Crédit photo : agsandrew / Shutterstock