



Big Data, corrélation et causalité

MIIA PARNAUDEAU

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

miia.parnaudeau@essca.fr

DATE RECEIVED:

June 10, 2015

DOI:

10.15200/winn.142960.01330

ARCHIVED:

April 21, 2015

CITATION:

Miia Parnaudeau, Big Data, corrélation et causalité, *The Winnower* 2:e142960.01330, 2015, DOI:

[10.15200/winn.142960.01330](https://doi.org/10.15200/winn.142960.01330)

© Parnaudeau This article is distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.



Meteo Protect invité au symposium international organisé par le CDRC, l'ESRC à la Royal Society, Londres 7-8 janvier 2015.

L'élaboration **d'un diagnostic de météo-sensibilité** passe par l'identification d'un lien statistique entre les flux de trésorerie de l'entreprise et des indicateurs de variabilité climatique adaptés au secteur concerné (température, pluviométrie, etc.). La **surabondance des données** qui caractérise désormais l'environnement de travail des ingénieurs météorologues, des data scientists et des évaluateurs de risques impose un renouvellement des méthodes.

Pourtant, **les problématiques liées au Big Data** ne sont pas nouvelles. La gestion des 4V, Volumes, Vitesse, Vérité et Variété des informations ne constitue pas un enjeu récent. C'est la manière dont on va traiter ces informations qui tend aujourd'hui à être revue.

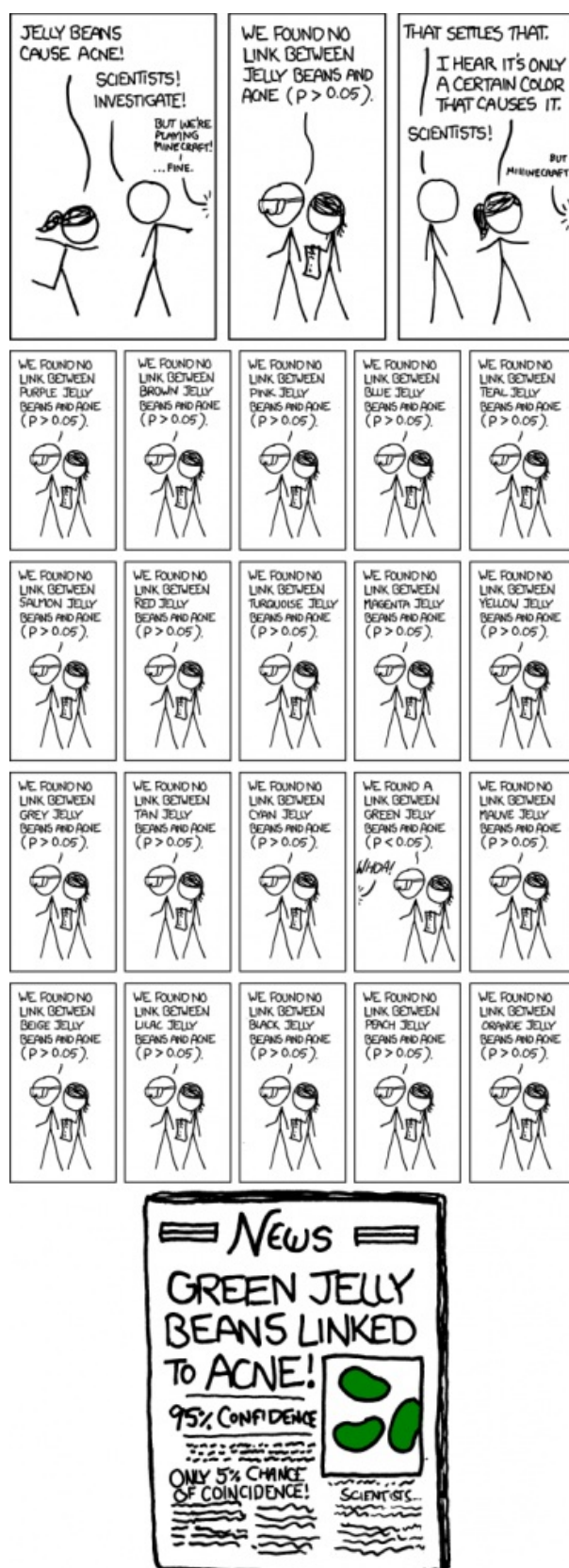
Le Pr. Sunil Gupta de la Harvard Business School, qui intervenait lors du big data symposium organisé à Londres par le CDRC et l'ESRC à la Royal Society, a souligné quelques changements majeurs dans la recherche quantitative et la modélisation des phénomènes économiques. Les approches portées notamment par les grands groupes tels que Google ont montré leurs limites. Dans un article paru en 2009, il apparaissait qu'il y avait une forte corrélation entre la recherche du mot « grippe » et la fréquentation des cabinets de médecins généraliste et que la recherche du mot « grippe » sur Google était le meilleur paramètre pour suivre voire prévoir l'évolution d'une épidémie grippale ¹. Pas besoin d'être un expert médical, il suffisait de compter le nombre de personnes qui recherchent le mot « grippe » pour devenir un expert en diffusion épidémiologique... Mais voilà, une étude plus récente montre que le modèle bâti sur le nombre de recherche du mot « grippe » surestime le nombre de visite chez le généraliste jusqu'à atteindre une erreur de plus de 100%².

Les techniques appliquées à l'heure actuelle pour traiter les Big Data sont-elles mal adaptées ? Les data scientists sont-ils trop éloignés du terrain pour comprendre ce qu'ils modélisent ? Plusieurs modèles simples valent-ils mieux qu'un seul modèle ultra-compiqué ? La significativité économique d'un modèle deviendrait-elle plus utile que la significativité statistique ?

Une première source de réponses à ces questions porte sur **les méthodes**. Si la première étape consiste à sélectionner précisément les objectifs d'analyse, il s'agit surtout d'extraire les bonnes données et d'appliquer les techniques statistiques de traitement appropriées. Cela est d'autant plus évident dans le cas de l'élaboration de diagnostics de météo sensibilité, où l'on cherche à quantifier de façon précise mais simple la corrélation entre les données météo et les données entreprises.

Post Hoc ergo Propter Hoc La question de la corrélation se trouve donc bien encore et toujours au cœur des préoccupations. **La corrélation n'est pas la causalité** : dans l'analyse de météo-sensibilité, il faut, au moment de la synthèse, apporter la preuve que les déterminants climatiques motivent réellement les flux de trésorerie de l'entreprise. Une des solutions peut ainsi reposer sur l'élaboration, en parallèle de tests de corrélation, de tests de causalité, afin de déterminer le sens de la relation qui s'exerce entre la météo et la trésorerie de l'entreprise.

Les tests d'hypothèses



En réalité, ce sont les **tests d'hypothèses** eux-mêmes qui sont remis en question à l'heure actuelle.

Est-ce à dire, comme le Pr. John Ioannidis⁴, que la plupart des travaux s'appuyant sur ces tests sont faux ? Faut-il alors, comme le soutiennent certains, abandonner le recours aux tests d'hypothèses ?

La réponse la plus clairvoyante à cette épineuse controverse se trouve dans les travaux du Pr. David Colquhoun⁵. Ce n'est pas parce qu'il y a de fortes chances de se tromper dans son diagnostic qu'il faut

abandonner la théorie des tests.

Il s'agit plutôt de déterminer combien de fois, lorsqu'un résultat est significatif, il l'est réellement, et combien de fois il ne l'est pas. On évite ainsi le problème des 'jelly beans'.

Les tests d'hypothèses ne constituent finalement qu'une étape dans l'élaboration d'un diagnostic qui a pour but premier d'éclairer les spécialistes sur les relations qui vont leur permettre de concevoir puis de proposer des solutions opérationnelles.

L'estimation de la météo-sensibilité d'une entreprise nécessite **une démarche sur-mesure**. Pour être efficace, cette démarche d'analyse devra confronter les expertises, mais aussi et surtout diversifier les approches, en croisant les compétences et en envisageant des alternatives parfois inédites.

Créer une 'data culture'



Le Big Data ne signera donc pas **la fin des data scientists**. Bien au contraire: le métier n'a jamais été aussi intéressant qu'aujourd'hui⁷. Toujours plus confrontés à la réalité, et à la surabondance des chiffres, un bon data scientist devra comprendre que tout ce qui compte ne peut pas forcément être mesuré, et que tout ce qui peut être mesuré ne compte pas forcément.

Réaliser un diagnostic de météo sensibilité, ce n'est pas uniquement se poser la question de savoir quelle base de données il faut utiliser, avec combien de personnes, et pour combien de temps. Il s'agit plutôt de mobiliser une équipe pluridisciplinaire, dont les interactions permettront d'aboutir aux résultats escomptés.

Créer les conditions favorable, dans une entreprise, d'une véritable 'data culture' implique de dépasser les approches projets classiques. S'il faut conserver une approche rigoureuse des bases de données, durant la phase d'étude, l'équipe se fait également une idée du degré de fiabilité des données et développera aussi des intuitions sur la façon dont elles peuvent être utilisées. **La créativité et l'interactivité sont les facteurs clés d'un environnement dit 'data driven'**.

Le colloque international big data a été l'occasion de débattre de l'ensemble de ces questions et les échanges ont été constructifs. Meteo Protect remercie à ce titre chaleureusement le Pr Kathryn Watson et le Pr Matthew Robson, de l'Université de Leeds, pour leur invitation.