MÉTHODES

Astrophysique : les quatre défis du Big Data

Jean-Luc STARCK

Chef du laboratoire CosmoStat –
Service d'astrophysique du Commissariat à l'énergie atomique

Les nouveaux projets internationaux comme le télescope spatial Euclid font entrer les cosmologistes dans l'ère du Big Data. Nos interrogations sur la matière noire ou l'énergie sombre, qui composent à elles deux 95 % du contenu notre Univers, nous imposent de nouveaux défis algorithmiques, computationnels et théoriques. Un quatrième défi concerne la recherche reproductible, concept fondamental pour la vérification et la crédibilité des résultats publiés.

Le Big Data est considéré comme l'un des plus grands challenges et aussi comme une magnifique opportunité dans de nombreux domaines scientifiques, technologiques, et industriels. En cosmologie, il pourrait aider à résoudre les mystères de l'Univers voire mettre en défaut la théorie de la relativité d'Einstein.
Mais le volume des données acquises pose de sérieux problèmes de calibration, d'archivage et d'accès comme d'exploitation scientifique des produits obtenus (images, spectres, catalogues…). Les données archivées de la future mission spatiale Euclid[1] contiendront 150 pétaoctets[2] de données et le projet Square Kilometre Array (SKA)[3] générera 2 téraoctets de données par seconde, avec 1 pétaoctet par jour archivé.

Les défis algorithmiques et computationnels

Tout l'enjeu est d'analyser ces jeux de données avec des algorithmes capables de mettre en évidence des signaux à très faible rapport sur bruit et intégrant les méthodologies les plus avancées : techniques d'apprentissage, outils statistiques ou concepts provenant de l'analyse harmonique, récemment mise en honneur avec l'attribution du prix Abel à Yves Meyer (le père de la théorie des ondelettes).
Disposer de tels algorithmes est un véritable challenge pour les équipes dans les années à venir : leur capacité à y parvenir conditionne le retour scientifique de leur engagement dans les grandes missions internationales.

De nouveaux domaines scientifiques

Ces défis ont permis de faire émerger une communauté de scientifiques issus de différents

1. Le projet Euclid est un projet de lancement par l'Agence spatiale européenne d'un satellite destiné à améliorer la compréhension des origines de l'univers. Son lancement est prévu pour 2020 https://www.euclid-ec.org
2. Pétaoctet : 10^{15} octets ; téraoctet : 10^{12} octets
3. Le projet SKA est un projet international de radiotélescope de très grande taille, qui sera construit à partir de 2018 en Afrique du Sud et en Australie http://skatelescope.org.

domaines (astrophysique, statistique, informatique, traitement du signal etc.). Objectif : promouvoir des méthodologies, développer de nouveaux algorithmes, diffuser les codes, les utiliser pour l'exploitation scientifique des données et former de jeunes chercheurs à l'interface entre plusieurs disciplines. Deux organisations ont été récemment créées, l'IAA (International Astrostatistics Association) et la commission 5 de l'IAU (International Astronomical Union) pour promouvoir l'astro-statistique et l'astro-informatique. Des laboratoires d'astro-statistique ont vu le jour aux Etats-Unis, en Grande-Bretagne (à l'Imperial College à Londres) et en France au CEA (le laboratoire CosmoStat au sein du Service d'astrophysique), ainsi qu'un centre d'astrophysique computationnel en 2016 à New York[4].

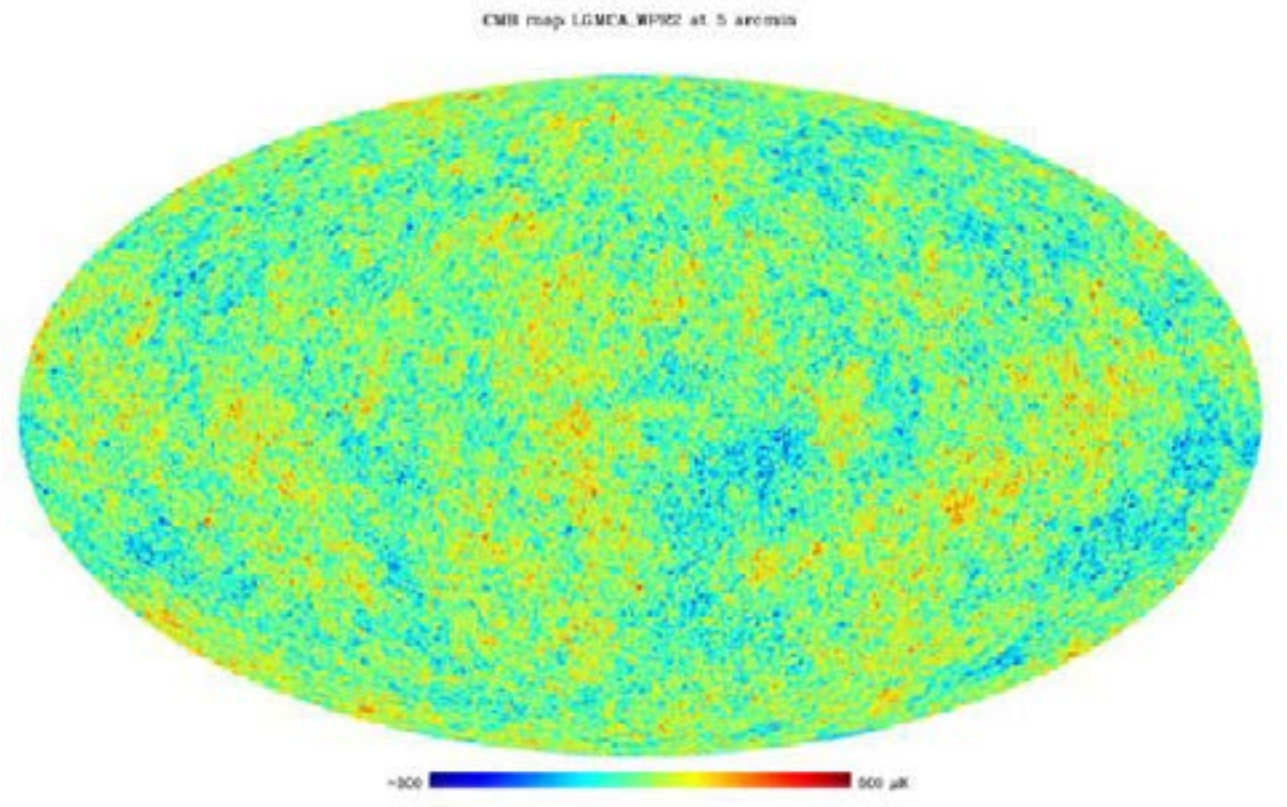

Figure : Image de la toute première lumière de l'univers, appelée le fond diffus cosmologique ou encore rayonnement à 3 kelvins, publiée par des chercheurs du Service d'astrophysique du CEA[5].

Le défi théorique

Pour comprendre la nature de l'énergie sombre et de la matière noire, et pour tester la relativité générale d'Einstein, il faut mesurer avec précision les paramètres du modèle standard de cosmologie, que l'on obtient à partir de données mesurées avec des télescopes spatiaux ou au sol.
Pendant longtemps, les erreurs sur l'estimation des paramètres cosmologiques provenaient d'effets stochastiques comme le bruit instrumental ou la variance cosmique, qui est liée au fait que l'on ne peut observer qu'une partie de l'univers à un instant donné, ce qui nous impose une incertitude supplémentaires sur nos mesures. D'où l'utilisation de détecteurs de plus en plus

4. Pour plus d'information : Astrostatistics and Astroinformatics Portal http://asaip.psu.edu
5. D'une précision exceptionnelle, cette image a été reconstruite à partir des données enregistrées par les télescopes spatiaux WMAP et Planck, à l'aide de méthodes mathématiques très poussées. Voir : http://www.cosmostat.org/research/cmb/planck_wpr2

sensibles et l'observation de champs du ciel de plus en plus grands. Ces erreurs stochastiques diminuant, les erreurs systématiques sont devenues de plus en plus importantes en valeur relative.
L'illustration la plus marquante de ce phénomène a certainement été l'annonce de la découverte des ondes gravitationnelles primordiales en mars 2014 par l'équipe américaine BICEP[6].
Il s'est avéré par la suite que le signal était bien réel, mais qu'il provenait en réalité de la poussière de notre galaxie. Une erreur de modélisation de l'émission de cette poussière avait laissé un signal résiduel dans les données.
En plus des erreurs stochastiques et systématiques, le Big Data génère un nouveau type d'erreurs, les erreurs d'approximations. L'estimation de certaines valeurs étant difficile avec la technologie actuelle, des approximations sont introduites dans les équations, pour accélérer le temps de calcul ou obtenir une solution analytique. Maîtriser ces erreurs est essentiel pour dériver des résultats corrects mais nécessite un effort théorique significatif.

Le défi de la recherche reproductible

Avec d'énormes volumes de données et des algorithmes très complexes, il devient souvent impossible pour un chercheur de reproduire les figures publiées dans un article. Or, la reproductibilité des résultats est au cœur de la démarche scientifique et constitue un des problèmes majeurs de la science moderne[7]. D'où le principe qui consiste à publier, en plus des résultats, les codes sources qui ont servi à analyser les données et les scripts utilisés pour traiter les données et générer les figures. Ce principe, désormais crucial, est rigoureusement appliqué par le laboratoire CosmoStat du CEA[8].

6. « BICEP » est une expérience de mesure de la polarisation du fond diffus cosmologique installée sur la base antarctique Amundsen-Scott, au pôle Sud
7. On pourra consulter à ce propos l'article du magazine en ligne Vox sur la recherche reproductible : http://www.vox.com/2016/7/14/12016710/science-challeges-research-funding-peer-review-process ; et aussi « A Manifesto for Reproductible Science » http://www.nature.com/articles/s41562-016-0021 ainsi que la charte de « reproducible science » www.nature.com/articles/s41562-016-0021/tables/1
8. Laboratoire CosmoStat : http://www.cosmostat.org