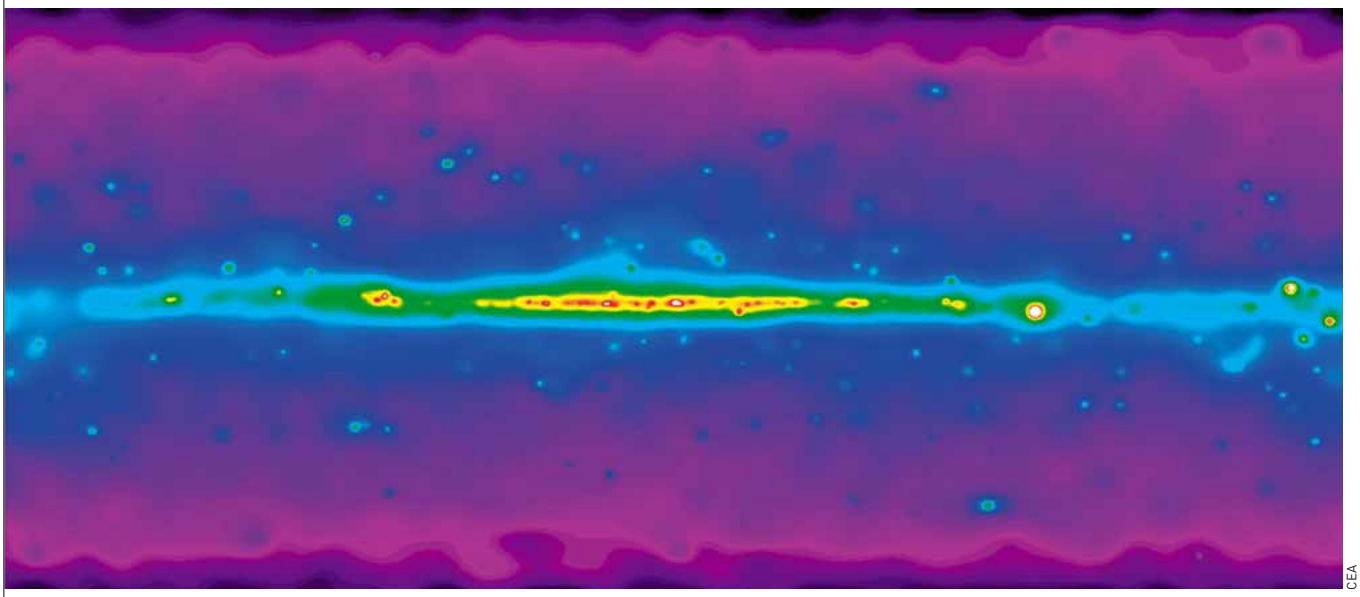
L'analyse des données en astrophysique

Pour un astrophysicien, le signal traduit l'information émise par une source et qu'il doit interpréter. Grâce à l'informatique, le traitement du signal connaît depuis une dizaine d'années des développements spectaculaires qui permettent aux astrophysiciens de valider, affiner ou remettre en question la compréhension de l'Univers.



CEA

Simulation de six jours de données du satellite GLAST (pour *Gamma-ray Large Area Space Telescope*), rebaptisé *Fermi Gamma-Ray Space Telescope*. Bande d'énergie comprise entre 0,1 et 1 GeV, filtrée par l'algorithme MR-filter à base d'ondelettes.

Par suite de l'évolution des détecteurs qui touche toutes les longueurs d'onde, l'analyse des données occupe une place de plus en plus prépondérante en astronomie.

Le flux de données

Si, en 1980, les *Charge Coupled Device* (CCD ou dispositifs à transfert de charge) affichaient une taille de 320 x 512 **pixels**, les astronomes disposent aujourd'hui de véritables mosaïques de CCD équivalant à 16 000 x 16 000 pixels. Les méthodes ont progressé au point que l'engagement humain et financier pour traiter les données d'un instrument peut atteindre l'ordre de grandeur de la réalisation de l'instrument lui-même. Par exemple, la caméra ISOCAM équipant l'**Observatoire spatial infrarouge** (**ISO**) a nécessité l'élaboration de logiciels de commande, d'analyse en temps réel et en temps différé : soit 70 hommes.an alors que 200 hommes.an suffisaient pour la construction de la caméra. L'effort consenti pour le projet Planck s'avère encore plus important. De plus, la quantité de résultats – parfois plusieurs centaines de **téraoctets** – fait appel à des bases de données et au développement d'outils sophistiqués (figure 1).

Les connaissances donnent lieu à de nouvelles questions dont la résolution s'appuie sur l'observation d'un objet ou d'une région du ciel. L'analyse de données intervient lors de la calibration de ces données, de l'extraction de l'information ou de la manipulation des bases de données. Des études statistiques permettent aussi d'enrichir la connaissance : c'est le cas pour les études sur le nombre de **galaxies** d'une **luminosité** donnée par unité de volume. On appelle « connaissance » l'ensemble des théories relatives à l'astronomie (formation d'**étoiles**, de galaxies, cosmologie...), les bases de données d'objets, d'images et les catalogues, autrement dit la liste des objets détectés avec un instrument dans une région du ciel. Cette connaissance, qu'il s'agisse des articles ou des bases de données, se trouve le plus souvent disponible sur l'Internet.

Selon l'instrument, les données se présentent sous forme d'images, de spectres, de mesures **photométriques**. En général, les astronomes disposent d'un ensemble de données d'une région ou d'un objet d'étude – plusieurs images à différentes **longueurs d'onde** par exemple. Mettre en orbite des instruments (Hubble, Rosat, ISO, Soho...) présente l'avantage d'éviter les contraintes atmosphériques. Le processus nécessite plusieurs étapes.

En premier lieu, vient la phase de calibration indispensable pour corriger les effets instrumentaux des données grâce à plusieurs opérations :

- la correction du zéro (*Dark*) : en l'absence de **lumière**, les valeurs renvoyées par le détecteur ne sont jamais nulles à cause des effets électroniques et il faut donc soustraire le niveau « zéro » aux données ;

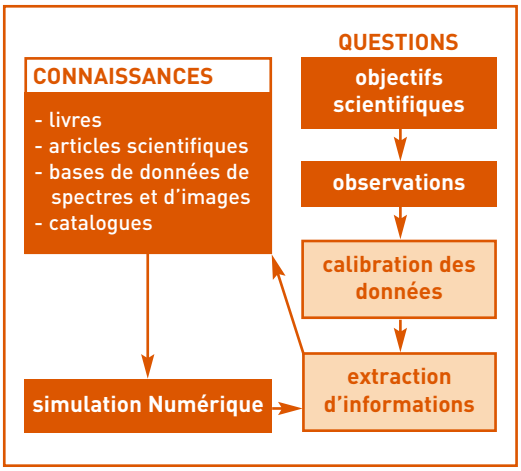


Figure 1.
Le flux d'information situe les niveaux où intervient l'analyse de données. Il permet de mieux comprendre à quels niveaux intervient cette analyse de données. À partir des connaissances, de nouvelles questions se posent et il est nécessaire de procéder à des observations d'un objet ou d'une région du ciel. Les données acquises vont devoir préalablement être calibrées, puis l'information utile à la réalisation de l'objectif scientifique doit être extraite. L'analyse de données intervient lors de la calibration, de l'extraction de l'information, et de la manipulation des bases de données. Il faut noter aussi que des études statistiques des bases de données permettent d'enrichir la connaissance (comme par exemple, les études du nombre de galaxies d'une luminosité donnée par unité de volume). On définit par connaissance l'ensemble des théories relatives à l'astronomie (formation d'étoiles, de galaxies, cosmologie...), les bases de données sur les objets, les bases de données d'images, et les catalogues (liste d'objets détectés avec un instrument sur une région du ciel). Cette connaissance est maintenant le plus souvent accessible par le réseau Internet (qu'il s'agisse des articles ou des bases de données).

• la correction des variations de réponse (*Flat*) : à éclairement égal, le détecteur ne répond pas de la même manière en chaque pixel ; aussi faut-il normaliser les données en les divisant par la réponse du détecteur ; les paramètres du détecteur doivent être bien connus, sinon les erreurs de précision se répercutent sur les mesures.

D'autres effets peuvent également être corrigés lors de cette première phase, comme la suppression des rayons cosmiques ou les effets de rémanence du détecteur ; toutes ces tâches étant plus ou moins ardues (figure 2).

Une fois les données calibrées, la phase d'analyse peut alors commencer. Suivant les objectifs, plusieurs mesures peuvent alors se faire, par exemple, sur la détection des étoiles et des galaxies ou sur la mesure de leur intensité, de leur position, de différents paramètres morphologiques : des résultats à comparer ensuite aux catalogues existants. Il s'avère impossible de citer toutes les opérations réalisables sur une image astronomique et nous n'avons donc cité ici que les plus courantes. Mener à bien cette extraction d'informations, suppose de surmonter des obstacles comme le bruit[1] et « la réponse instrumentale »[2]. Lorsque les informations utiles ont été extraites, elles sont confrontées à la connaissance en l'état. Cette étape valide, affine ou remet en question la compréhension de l'Univers. Le résultat final de cette réflexion se matérialise par la publication d'articles scientifiques paraissant dans des revues spécialisées.

Le volume de connaissance croît donc rapidement et nécessite des outils performants pour l'exploiter. Ces outils font office de moteurs de recherche pour aider à rassembler les derniers articles parus, des méthodes pointues d'imagerie ou encore des algorithmes établissant une correspondance entre des objets détectés dans une image et une base de données (sachant qu'il y a des choix à faire quand plusieurs « candidats » se présentent). La difficulté réside, en particulier, dans le volume des bases de données. SIMBAD, un ensemble d'identificateurs, de mesures et de bibliographie pour les données astronomiques, développé par le **Centre de données astronomiques de Strasbourg**, regroupe des informations sur plusieurs millions d'objets, correspondant à des millions de mesures observationnelles, avec des millions de références bibliographiques. La taille des images fournies par les instruments de nouvelle génération rend quasiment impossible l'accès aux archives d'images par le réseau lorsque celles-ci n'ont pas été préalablement comprimées.

La maîtrise et l'exploitation des bases de données représentent donc un enjeu important pour le futur. L'analyse statistique de catalogues amène, en outre, à contraindre les paramètres des modèles cosmologiques et donc à augmenter la connaissance. Par exemple, l'étude statistique des galaxies, individuellement différentes, révèle la présence de trois grandes familles (**spirales**, **elliptiques** et **irrégulières**), un peu comme les grains de sable d'une plage, tous différents, forment une zone homogène.

La théorie du *Compressed Sensing* et le projet spatial Herschel

Dans le cadre de certains projets spatiaux, il s'avère parfois impossible de transférer sur terre un volume important de données sans procéder à une compression de celles-ci. C'est le cas pour l'instrument

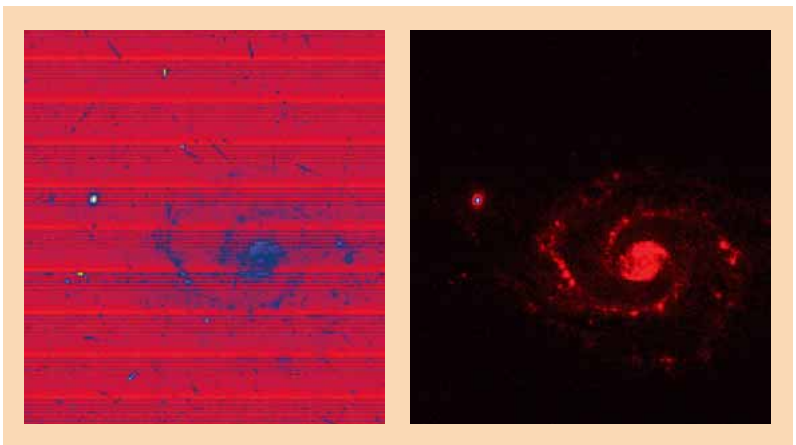
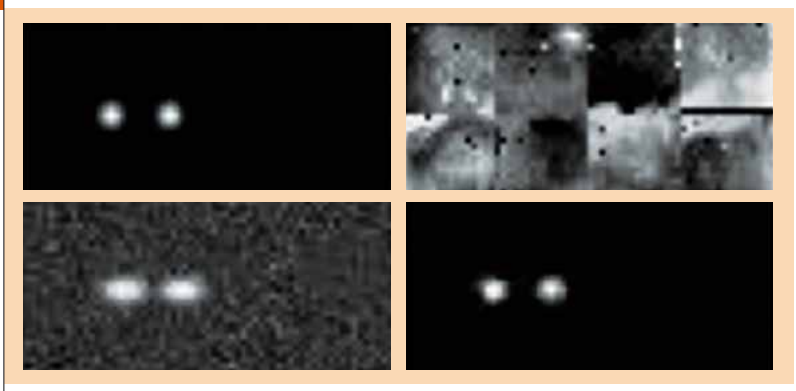

DR

Figure 2.
La galaxie M51, vue par ISO, avant et après calibration.

(1) Le bruit est une fluctuation d'intensité aléatoire qui se superpose aux données ; il provient en partie du détecteur et, en partie, des données. En plus des erreurs que le bruit peut amener sur les mesures, il gêne énormément lors de la détection d'objet et peut être responsable de nombreuses fausses détections.

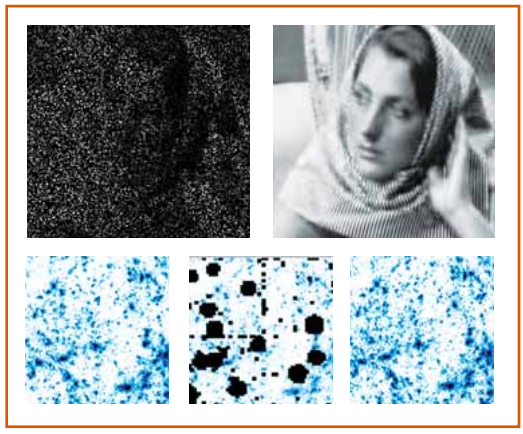
(2) L'image d'une étoile n'est pas un point, mais une tache. Cet étalement lié à l'instrument est appelé «réponse instrumentale». Son principal effet consiste en une perte de résolution car les objets proches se mélangent.



DR

Figure 3.
En haut, la simulation d'une image contenant deux sources et la même image vue par Herschel. En bas, à gauche, l'image a été calibrée après une compression classique et, en bas, à droite, la solution obtenue par la technique du *Compressed Sensing*.

PACS Herschel (*Photodetector Array Camera and Spectrometer*) qui nécessite une compression d'un facteur huit avec une puissance de calcul extrêmement faible. Les méthodes standard de compression d'image comme JPEG ne conviennent pas. Heureusement, depuis une dizaine d'années, d'importants développements en analyse harmonique permettent de représenter des images dans des bases de fonctions convenant à certains types d'objets. Par exemple, les ondelettes sont idéales pour détecter des structures, la **transformée ridgelet** est optimale pour la recherche de ligne et les curvelets représentent bien les contours ou les filaments contenus dans une image. Plus généralement, une représentation « parcimonieuse » des données conduit à de meilleures performances pour des applications aussi variées que la compression des données, la restauration d'images ou la détection d'objets. Une nouvelle théorie, *Compressed Sensing*, lie désormais formellement le nombre de coefficients non nuls dans une base donnée et l'échantillonnage nécessaire à une reconstruction exacte du signal. Ce récent concept montre que la contrainte sur le pas d'échantillonnage, fixée par le **théorème de Shannon**, peut être largement dépassée si le signal observé vérifie un « critère de parcimonie », c'est-à-dire s'il existe une base dans laquelle le signal présente peu de coefficients différents de zéro. Une étude préliminaire a montré que cette approche présenterait une excellente alternative aux systèmes de transfert de données actuellement en vigueur pour le satellite Herschel et, qu'à taux de compression constant, un gain de 30 % en résolution s'obtiendrait sur les images décomprimées (figure 3).



DR

Figure 4.
En haut et à gauche, l'image comporte 80 % de pixels manquants ; à droite, elle est restaurée.
En bas et à gauche, on trouve la simulation d'une carte de masse de matière noire, au milieu l'image est la même mais avec des zones manquantes et, à droite, l'image est reconstruite.
On a montré que l'erreur sur le spectre de puissance et le bispectre de la carte restaurée par *inpainting* est de l'ordre de quelques unités pour-cent. Cette méthode originale dépasse les problèmes d'astrophysique et a été transférée vers l'industrie dans le cadre d'un contrat CIFRE[3] avec la Sagem.

L'*inpainting* au secours des données manquantes

Les données manquantes constituent un problème récurrent en astronomie, dû à des pixels défectueux ou à des zones éventuellement polluées par d'autres émissions et que l'on souhaite masquer lors de l'analyse de l'image. Ces zones masquées occasionnent des difficultés lors de traitements ultérieurs, en particulier pour extraire des informations statistiques comme le spectre de puissance ou le bispectre. L'*inpainting* est la procédure qui va venir combler ces zones. Des travaux récents montrent que l'on peut reconstruire les zones manquantes en recherchant une solution parcimonieuse dans un dictionnaire de formes prédéfinies. Avec un choix judicieux de dictionnaire, on obtient des résultats fantastiques (figure 4).

Planck et l'extraction du fond diffus cosmologique

La mission spatiale Planck, qui a été lancée, le 14 mai 2009, par l'**Agence spatiale européenne** (**Esa**), en même temps que la mission Herschel, a pour objectif de cartographier des fluctuations spatiales d'intensité et de **polarisation** de l'émission du ciel millimétrique, en vue notamment de caractériser les propriétés statistiques des **anisotropies** du fond de **rayonnement** cosmologique fossile. Ces mesures permettront de contraindre fortement les **modèles** cosmologiques et, entre autres, de tester le modèle standard du big bang ainsi que de déterminer, avec une précision inégalée, les paramètres cosmologiques décrivant l'ensemble de l'Univers. Planck offre les meilleures perspectives pour comprendre ce modèle, de l'Univers primordial (inflation) à l'astrophysique des émissions galactiques, en passant par la formation des structures, les **amas de galaxies**, la **matière noire** et l'**énergie noire**, ou encore la topologie de l'Univers. Deux instruments sont opérationnels, le *Low Frequency Instrument* (LFI) et le *High Frequency Instrument* (HFI), pour obtenir neuf cartes de tout le ciel entre 30 et 1 000 **GHz**. Ces cartes contiendront le fond cosmologique, mais aussi d'autres composantes liées à des émissions galactiques (poussière) ou intergalactiques (galaxies, amas...). Celles-ci sont également d'un grand intérêt. Chaque carte présentant un mélange des différentes composantes (**fond diffus cosmologique**, poussière galactique...), la difficulté

(3) Conventions industrielles de formation par la recherche (CIFRE) : instruites et gérées par l'Association nationale de la recherche technique (ANRT) pour le compte du ministère de l'Enseignement supérieur et de la Recherche, elles permettent à une entreprise de bénéficier d'une subvention annuelle forfaitaire en contrepartie des coûts qu'elle engage pour employer le jeune doctorant qu'elle a embauché pour trois ans.

consiste donc à retrouver les composantes « ciels » à partir des cartes. On appelle cette opération « séparation de sources » (figure 5).
En pratique, il faut de surcroît tenir compte des effets instrumentaux (le bruit...) venant compliquer encore cette séparation. Un problème de restauration de données se superpose donc à celui de la séparation de sources. En utilisant une méthode appelée *Generalized Morphological Component Analysis* (GMCA), basée sur la **transformée en ondelettes**, il devient possible de reconstruire le fond diffus. Le principe repose sur le fait que mélanger les composantes rend les images plus complexes et que, si on utilise un critère de régularisation basé sur le principe de la « simplicité de la solution » dans le problème de séparation, on peut retrouver des composantes recherchées. Dans cette approche, une image « simple » est représentable sur peu de coefficients en ondelettes ; il s'agit d'une solution dite parcimonieuse (figure 6).

Tests statistiques sur le fond diffus cosmologique

Certaines applications nécessitent des outils statistiques élaborés afin de mettre en évidence des signaux extrêmement faibles, noyés dans du bruit. Parmi les exemples intéressants figure celui de la détection de **sources non gaussiennes** dans le fond diffus cosmologique **micro-onde** (FDCM). Celui-ci résulte d'un découplage de la matière et de la **radiation** à un décalage cosmologique de 1 000. Il est une « relique » des premiers instants de l'Univers et aide à comprendre la formation et l'évolution des structures provenant de l'amplification des fluctuations initiales. Les propriétés statistiques des anisotropies de température du FDCM nous informent donc sur la physique de l'**Univers primordial**. En effet, si leur distribution est gaussienne, elles sont produites par des modèles simples d'inflation. Sinon, elles sont issues de défauts topologiques comme des cordes cosmiques. Des anisotropies peuvent également provenir de l'interaction des photons du FDCM avec des **électrons** libres du gaz chaud intra-amas : c'est l'**effet Sunyaev-Zel'dovich** (figure 7).
Pour trouver des signatures non gaussiennes très faibles, des tests statistiques très sensibles s'imposent. Ils pourraient dériver de l'étude statistique de la distribution des coefficients obtenus par des méthodes multi-échelles. Les ondelettes sont bien adaptées à l'analyse des structures spatialement **isotropes** et contribuent à détecter l'effet Sunyaev-Zel'dovich, tandis que les fonctions curvelets sont optimales pour la recherche de structures spatialement anisotropes. La combinaison de ces deux transformées multi-échelles permet non seulement de détecter au mieux les anisotropies dans le FDCM, mais aussi de déterminer leur origine, éventualité impossible avec les méthodes traditionnelles.

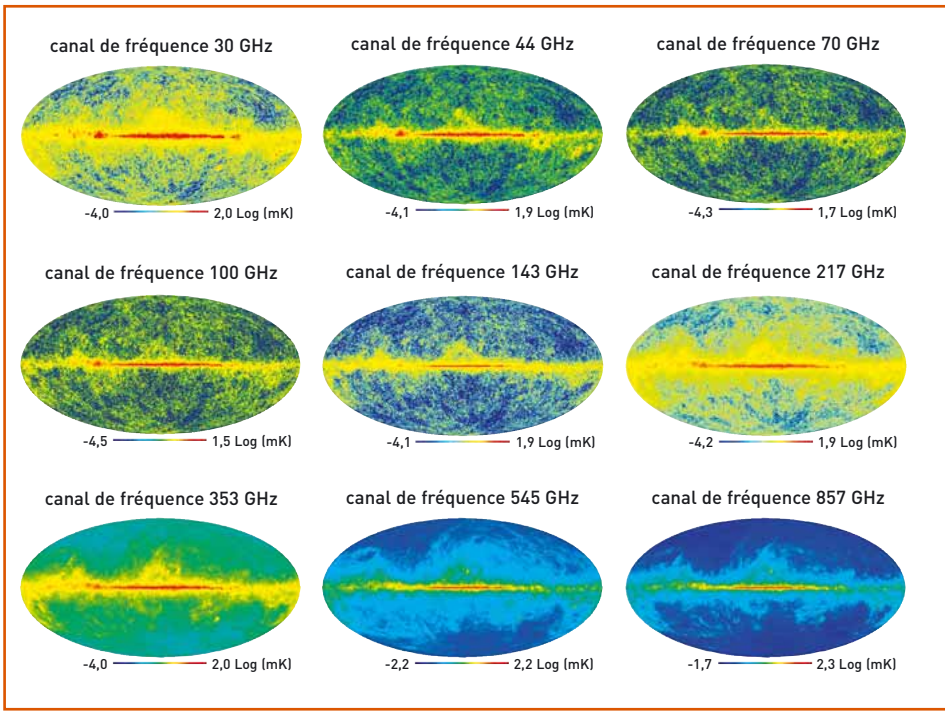


Figure 5.
Simulation des neuf cartes de Planck. Par exemple, pour obtenir le fond diffus cosmologique, il faut retrancher la contribution des autres composantes aux observations.
Échelle des valeurs : il s'agit de cartes de températures en mK, pour pouvoir avoir un contraste suffisant ; le logarithme des cartes est affiché.

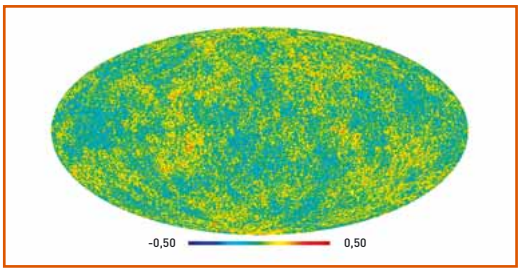


Figure 6.
Carte du fond diffus cosmologique obtenue par GMCA à partir des neuf cartes de la figure 5.

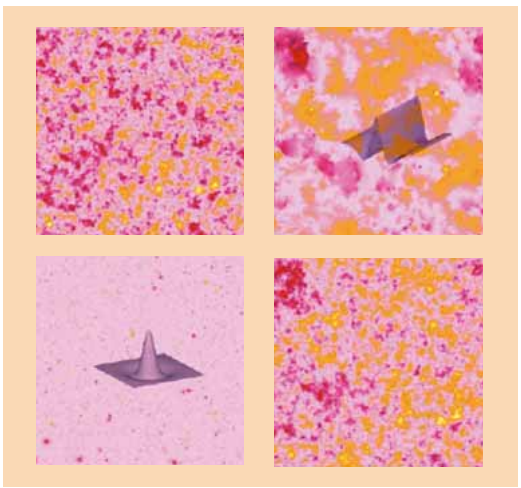


Figure 7.
Parmi ces quatre cartes, les trois premières sont des simulations du FDCM, de l'effet Sunyaev-Zel'dovich (en haut, à droite) et des cordes cosmiques (en bas, à gauche).
La quatrième représente le mélange de ces trois composantes (en bas, à droite), un type de données que pourrait fournir la mission PLANCK. La fonction ondelette est surimprimée en haut, à droite ainsi que la fonction curvelet est surimprimée en bas, à gauche.

> Jean-Luc Starck
Service d'électronique des détecteurs et d'informatique (Sedi)
Institut de recherche sur les lois fondamentales de l'Univers (Irfu)
Direction des sciences de la matière (DSM)
Unité mixte de recherche astrophysique interactions multi-échelles
(CEA-Université Paris 7-CNRS)
CEA Centre de Saclay (Orme des Merisiers)