
Gaia
Validation des
données
F. Arenou, P. Di Matteo

Publication unit: DPAC CU9

- CU9 est la seule responsable de la diffusion des données (sauf les alertes scientifiques)
- ESA aide pour l'infrastructure et le support
- Les tâches CU9:

	Release 1+2		Release 3		Release 4		Final			FTE
	2014	2015	2016	2017	2018	2019	2020	2021	2022	**Total**
Management	1.7	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	**11.3**
Documentation	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	**2.25**
Architecture	7.45	6.75	6.75	5.9	5.7	5.2	5.2	5.2	5.2	**53.35**
Validation	13.28	13.65	13.8	13.8	13.8	13.8	13.8	13.8	13.8	**123.53**
Operations	5.2	5.3	5.3	5.2	5.17	5.17	5.17	5.17	5.17	**46.85**
Outreach	1.6	3	3	2.9	2.9	2.9	2.9	2.9	2.9	**25**
Sci. enab. apps.	11.55	12.25	11.75	10.3	10.3	10.3	10.3	10.3	10.3	**97.35**
Visualization	7.2	7.2	7.2	6.3	6	5.5	5.5	5.5	5.5	**55.9**
Total	**48.23**	**49.6**	**49.25**	**45.85**	**45.32**	**44.32**	**44.32**	**44.32**	**44.32**	**415.53**

Pourquoi une validation?

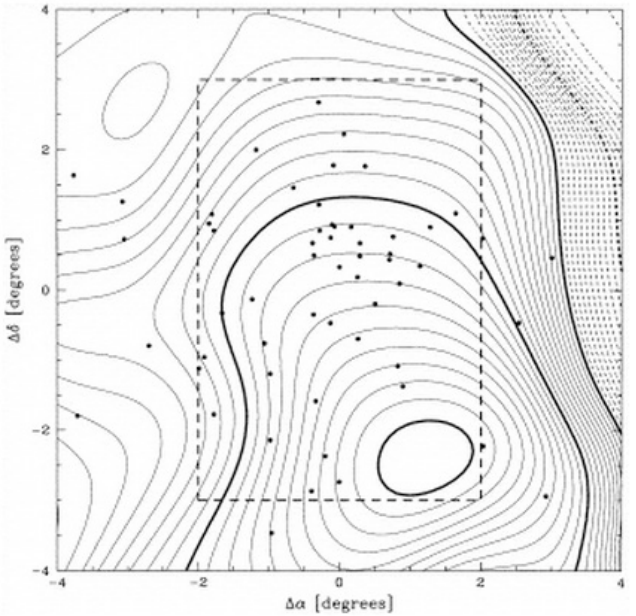
- Gaia est une mission difficile
  - Un satellite complexe pour mesurer un ciel complexe !
  - Des milliards de paramètres avec des algorithmes complexes
    - Beaucoup d'occasions d'erreurs systématiques
- DPAC est responsable de la qualité du Catalogue
  - 450+ scientifiques/ingénieurs… > 1000 années-homme
  - Le Catalogue ne peut être publié sans validation
  - Il y a une vérification du traitement, pas des résultats globaux
- Expérience d'Hipparcos
  - Les utilisateurs se méprennent aisément sur la nature (statistique) des données
  - Déjà, un effort avait porté sur la validation (1 thèse et 2 articles sur l'astrométrie, 3 chapitres de documentation)

THE ASTRONOMICAL JOURNAL, 129:1616–1624, 2005 March

CONFIRMATION OF ERRORS IN *HIPPARCOS* PARALLAXES FROM *HUBBLE SPACE TELESCOPE* FINE GUIDANCE SENSOR ASTROMETRY OF THE PLEIADES[1]

DAVID R. SODERBLOM AND ED NELAN
Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218; sode

G. FRITZ BENEDICT, BARBARA MCARTHUR, IVAN RAMIREZ, AND
McDonald Observatory, University of Texas, Austin, TX 78712; fritz@a
mca@astro.as.utexas.edu, ivan@astro.as.utexas.edu, spies@astro.a



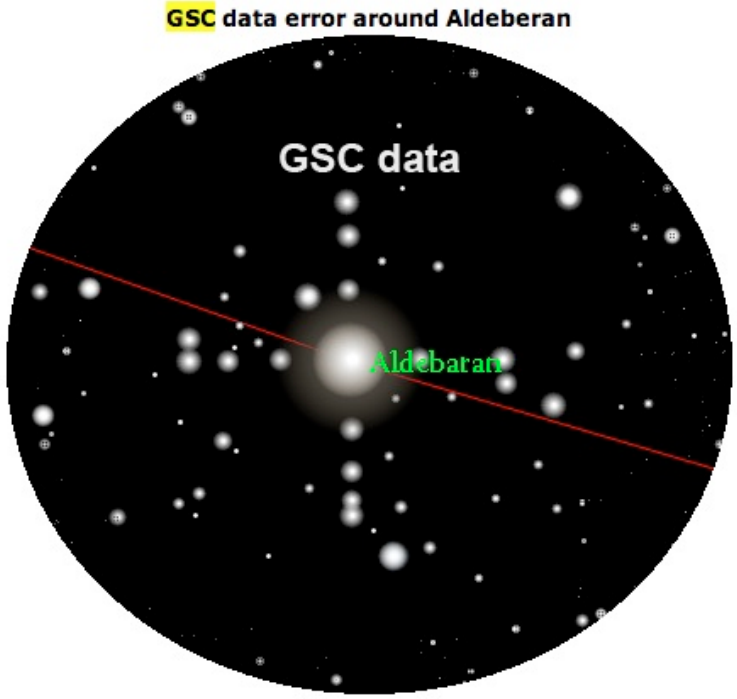
A&A 439, 805–822 (2005)
DOI: 10.1051/0004-6361:20053192

Rights and wrongs of the Hipparcos data

A critical quality assessment of the Hipparcos catalogue

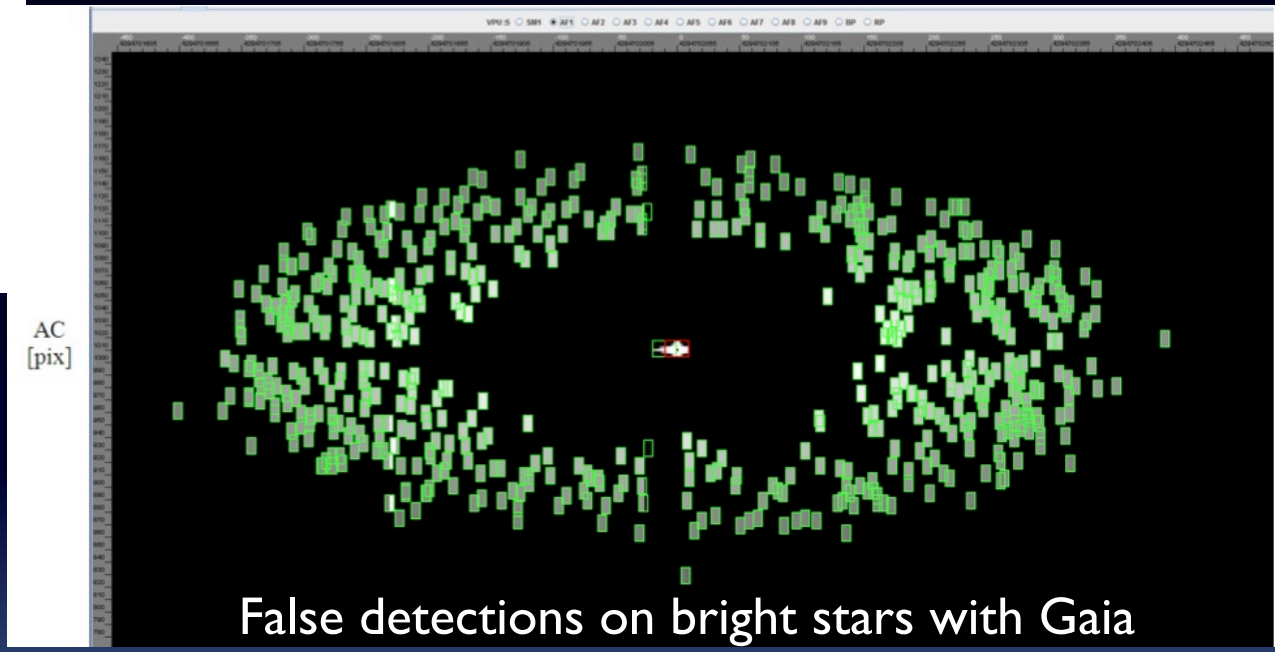
F. van Leeuwen

Artefacts…



GSC 1.0 problems

- Many possible problems to expect, e.g.
  - Scanning anomalies
  - Basic angle variations
  - Thermal control anomalies



False detections on bright stars with Gaia

Motivation

- Contexte: absence de période de données propriétaires
- La meilleure forme de préparation à l'arrivée des données
  - Consiste à travailler sur leur contenu
  - La validation du futur Catalogue en est l'occasion
- Permet également de fournir des outils pour la communauté
  - Pour exploiter scientifiquement les résultats de la mission.
- L'Observatoire a de nombreuses compétences
  - GEPI (stellaire/galactique), IMCCE (SSO), SYRTE (quasars)
  - Historique de la validation des données d'Hipparcos

What validation is

- Cross-CU check of the quality of the Catalogue
  - Have a critical look at the output
  - Do not leave gross errors undetected before publication
  - And correct problems — if any — as soon as possible
    - Feedback to CUs between intermediate Catalogue releases
- Assess the statistical properties
  - Hopefully
    - Unbiased parameters (look for systematic errors)
    - Unbiased parameter standard errors (assess precision, correlations)
    - Find outliers
  - Explain main features in the data
  - Validation results are an integral part of the documentation
    - Documenting the Catalogue properties

What validation is not

- Not a verification of the DPAC workflow
  - Already (and much better) done within DPAC
  - Though an indirect verification of CU9 tools for the archive access
- Not infallible
  - Minor problems in such a large set of data may remain
- It is not scientific research in our field of expertise
  - But it requires scientific expertises to understand the tests
  - CU9 is part of DPAC and is bound by the Science Management Plan which states that there will not be proprietary data rights for Gaia.

W.P. typiques de validation

- Tests de cohérence
  - Interne entre instruments (astrométrie/photométrie/spectroscopie)
  - Basés sur les conséquences d'erreurs de calibration
    - C. Fabricius (U. Barcelone)
- Comparaison avec un modèle de Galaxie
  - A. Robin (CNRS/Utinam)
- Comparaison avec des (grands) catalogues externes
  - C. Babusiaux (OPM)
- Outils statistiques & graphiques (fouille de données)
  - M. Manteiga (OAC) - A.Helmi (Groningen)
- Objets spéciaux:
  - Variabilité, L. Eyer (U. Genève)
  - Amas, A. Vallenari (INAF)
  - Objets du système solaire, F. Mignard (OCA)

Organisation

- Responsabilité GEPI
  - Groupe de 65 personnes
  - ~14 FTE/an, France (75%), Espagne, Italie, Suisse, Pays-bas
- Financement FP7-Space-2013-1
  - Programme Genius, 2ème nœud
  - GEPI (3 ans post-doc)+IMCCE (2ans)
- Soutien CNRS
  - Programme Mastodons « Grandes masses de données »
  - Mission Interdisciplinarité

Défi Mastodons, Grandes masses de données scientifiques
Mission Interdisciplinarité du CNRS

Gaia, l'origine et l'évolution de notre Galaxie : validation des données

http://wwwhip.obspm.fr/mastodons

Astro
- P. Di Matteo, F. Arenou, C. Babusiaux, CNRS UMR 8111 / GÉPI, Observatoire de Paris/Meudon
- D. Hestroffer, J. Berthier, CNRS UMR 8111 / IMCCE, Observatoire de Paris
- A. Robin, C. Reylé, CNRS UMR 6213 / Institut Utinam, Observatoire de Besançon

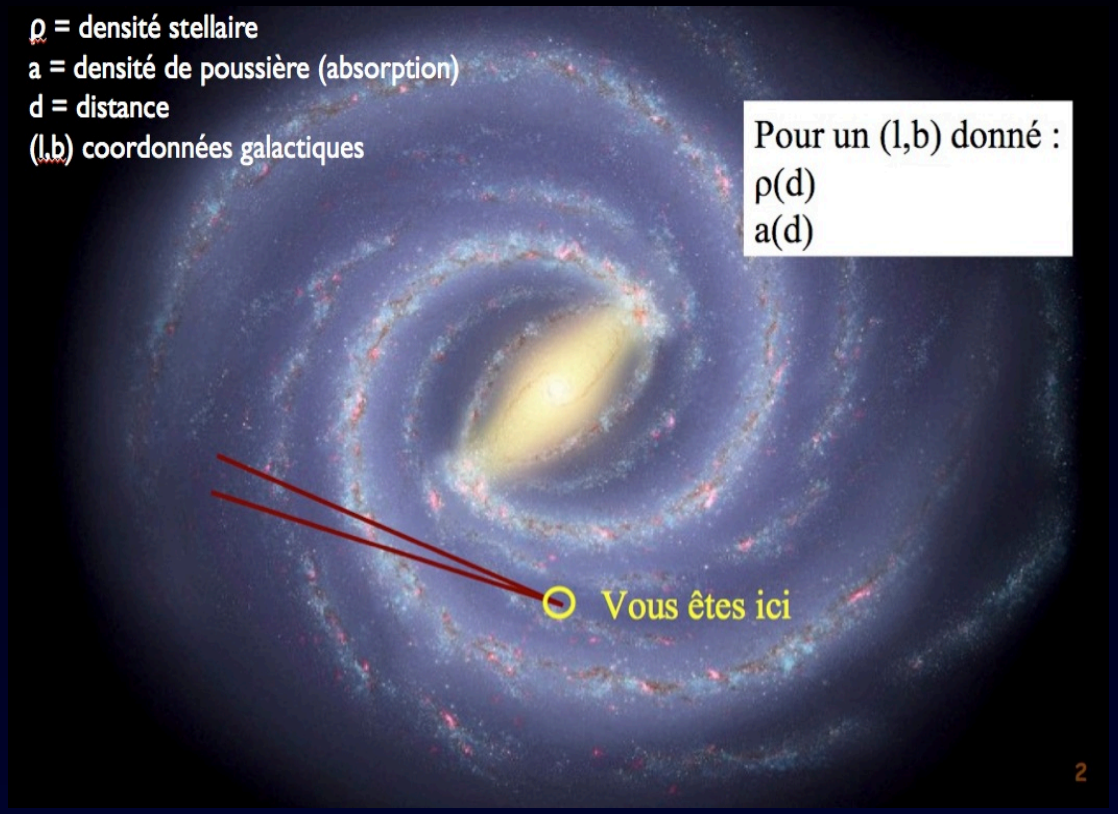
Info
- K. Zeitouni, L. Yeh, CNRS UMR 8144 / PRiSM, Université de Versailles St Quentin

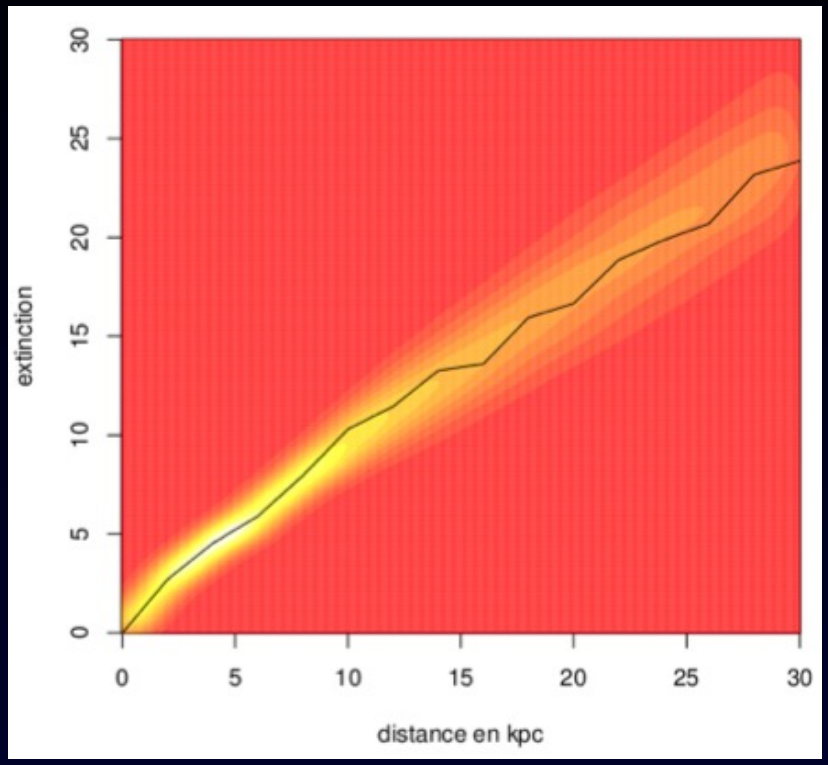
Stat
- C. Robert, J. Rousseau, CNRS UMR 7534 / Ceremade, Université Paris-Dauphine
- Radu Stoica, CNRS UMR 8524 / Laboratoire Paul Painlevé, Université Lille I
- Jean-Marie Cornuet, DR ex-INRA Montpellier

Programme Mastodons

- Mission Interdisciplinarité CNRS (depuis 2012, 5 ans ?)
  - GEPI + IMCCE + Obs. Besançon (astro)
  - Univ. Paris-Dauphine + Lille I (stat)
  - UVSQ (BD)
- Basé sur la validation, et l'exploitation des données Gaia
- Problématiques statistiques
  - Détermination simultanée de distances et extinctions galactiques
  - Détermination d'orbites et masses astéroïdes
  - Méthodes ABC, réduction de coût calcul
- Problématiques bases de données
  - Montrer les limites des serveurs de données traditionnels
  - Proposer des méthodes d'indexation et des algorithmes efficaces d'analyse et de fouille pour le "spatial Big Data"

Ex: estimation distance+ extinction



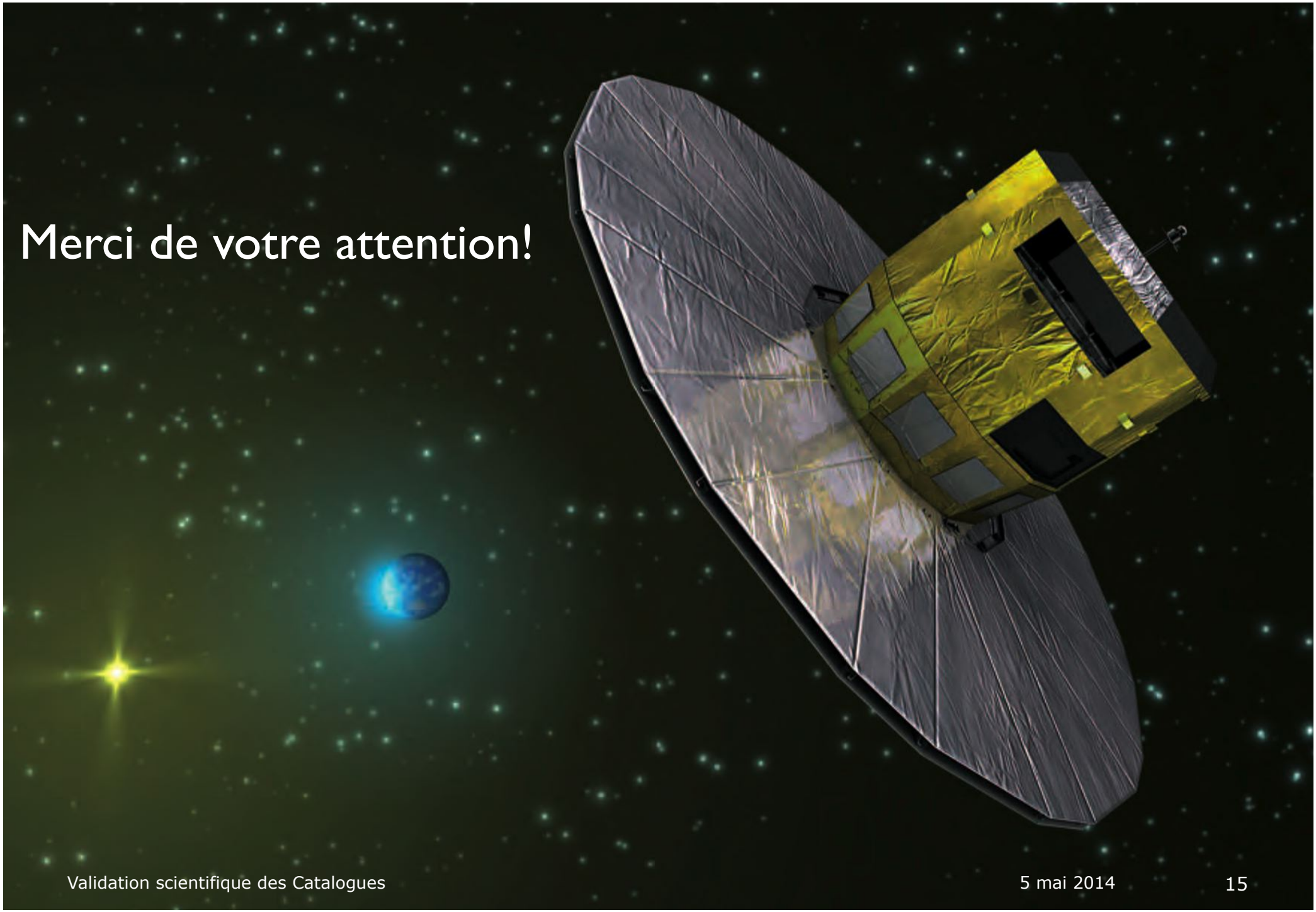


- C. Babusiaux (GEPI) + Statisticiens (Montpellier+Dauphine)
- Modélisation conjointe: analyse directe vs ABC
- Cette modélisation peut aider à valider les données de Gaia (comparaison des distances), puis exploitées scientifiquement grâce aux données Gaia quand disponibles.

Bilan Mastodons 2013

- Analyses statistiques
  - Ajustement de modèles de la Galaxie sur des données de Grands Relevés,
  - Analyse de la distance et de l'extinction dans le disque de la Galaxie,
  - Algorithme MCMC pour l'inversion des données d'astéroïdes binaires
- Archive Gaia et infrastructure « Big Data ».
  - Besoins et caractéristiques du système d'archivage pour tests de validation
  - Expérimentation de requêtes sur 8 to 77 millions d'objets simulés sur 48 VM
  - Début d'un travail commun Amadeus (données Corot)-Gaia-Petasky (LSST)
- Un objectif a d'ores et déjà été atteint: faire travailler ensemble
  - Astronomes + statisticiens + spécialistes BD
  - Ce qui n'aurait pas été entrepris sinon !
- Le résultat
  - … après la mise en place d'un vocabulaire commun
  - Montée très rapide de la courbe d'apprentissage (méthodes statistiques)
  - Recrutement (participation bénévole) d'un statisticien
  - Expertise pour (ré)orienter le travail

Merci de votre attention!



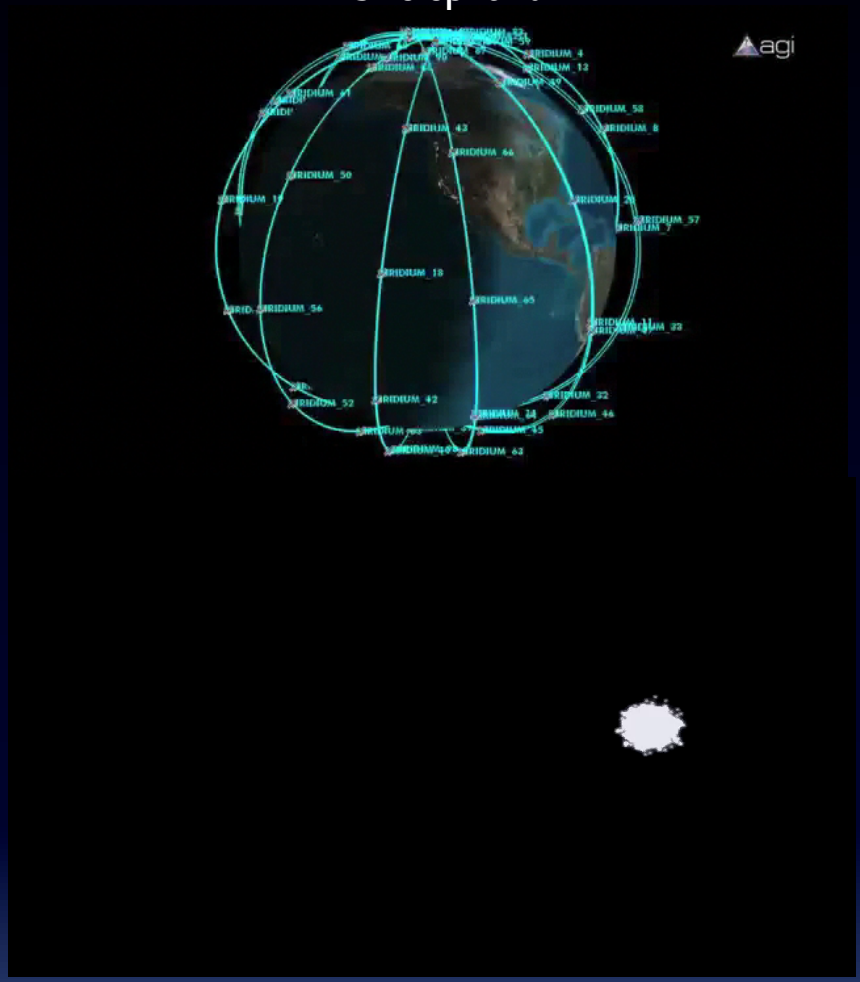
Publications Mastodons 2013

- Articles :
  - C. Babusiaux et al., *Metallicity and kinematics of the bar in-situ*, accepté dans A&A, arXiv:1401.1925
  - Oszkiewicz, D., Hestroffer, D., David P. 2013 *MCMC computation for binary orbits*. Proc. of SF2A conf.
  - Robin, A.C., Reylé, C., Fliri, J., Czekaj, M., Robert, C., Martins A.M.M. *Characterization of the thick disc and the halo of the Milky Way*, A&A submitted.
  - Yeh L., Zeitouni, K., *Distance Self-Join with Minimizing Replicated Objects in MapReduce*, submitted, 19th International Conference on Database Systems for Advanced Applications (DASFAA'14)
- Posters et contributions aux conférences :
  - « *Constraining the Milky Way formation and evolution with APOGEE and population synthesis: The Chemical gradients* » Martins, A.M.M., Robert, C., Robin, A.C. 5th IMS-ISBA joint meeting. MCMSki IV.
  - « *MCMC computation for binary orbits*. » Poster EPSC conf. 2013, London UK.
  - « *MCMC computation for binary orbits*. » Poster CelMec conf 2013, Viterbo, Italie
- Une partie des travaux communs non encore publiés

Objectifs 2014

- Arrivée des premières données
  - Premier Catalogue prévu fin 2015
  - Mise en place 1ères requêtes validation
    - beaucoup de données, peu d'attributs
- BD
  - Compléter les requêtes
  - Optimisation de l'implémentation
- Poursuite travaux statistiques
  - Méthodes ABC, réduction coût calcul
  - Nouveau: recherche de surdensités

Débris spatiaux



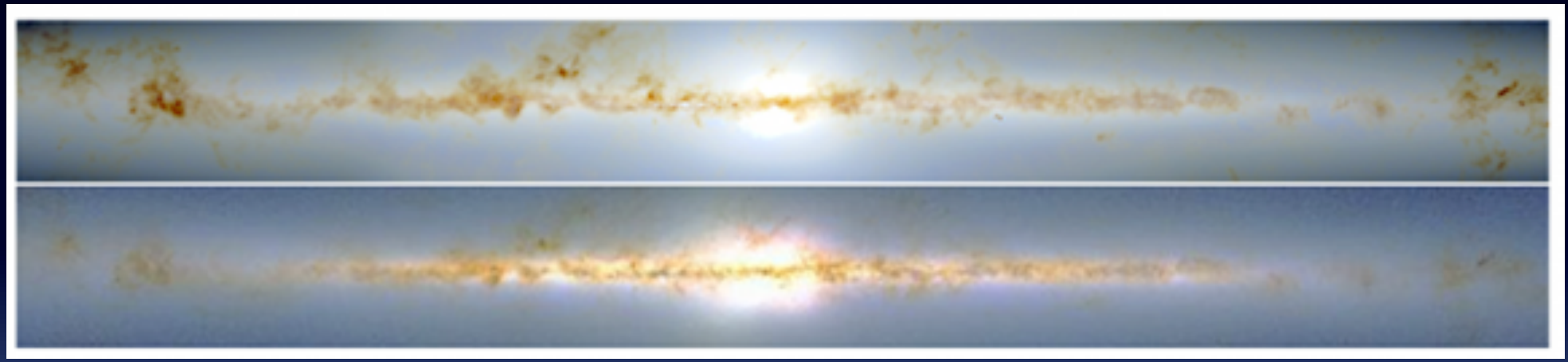
Structure du halo via accrétion galaxies

Mastodons

Astrostat

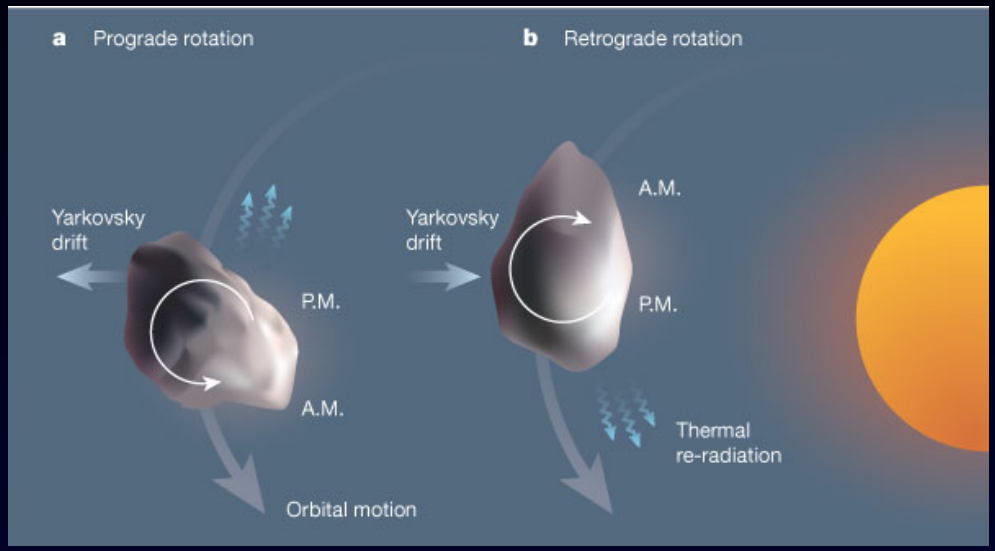
Ajustement modèles de la Galaxie

- Ajustement sur des données de *Grands* Relevés (SDSS, 2MASS)
  - Paramètres du disque épais de notre galaxie
  - MCMC + ABC (Approximate Bayesian computation)
  - Utinam + Dauphine
- Validation ---→ exploitation
  - Le modèle servira d'abord à valider les résultats de Gaia (par ex. retrouver les grandes structures)
  - Dans le futur, les données Gaia amélioreront (considérablement) le modèle



Système solaire

- Statistiques
  - Orbital inversion avec MCMC
  - Modèle dynamiques
  - Astéroïdes binaires
  - MLE Bayes et distributions multivariées,
- extraction d'information ou structure sur grosse BD
  - taxonomie, familles, traceurs galactique etc.



Mastodons

Bases de données

Objectif général

- The Gaia Universe Model Snapshot - GUMS
  - Simulation du Catalogue
  - Comprend 1 milliard d'objets(équivalent de 20 TB)
- Méthodologie
  - Identifier quelques spécificités des données et des requêtes dans GAIA
  - Etudier les solutions existantes
  - Proposer des approches adaptées aux besoins dans GAIA
- 2 alternatives :
  1. Utiliser un SGBD Relationnel / spatial & SQL (Parallèle ou non)
  2. Utiliser le modèle Key-value store et le paradigme map-reduce (Parallel)

Premier problème étudié

- Cône search (généralement très sélectif)
  - Utilise une sélection sur intervalles (ex. B-tree)
  - A priori ne pose pas de problème

- Jointure distance (cross-match)
  - Pour relier les objets de GAIA à des catalogues existants (1 fois)
  - Pour trouver des objets voisins du même catalogue ou pour le clustering (self-join)
  - Requête très couteuse !!

  ➢ Focus sur la jointure distance !
  avec Map/reduce

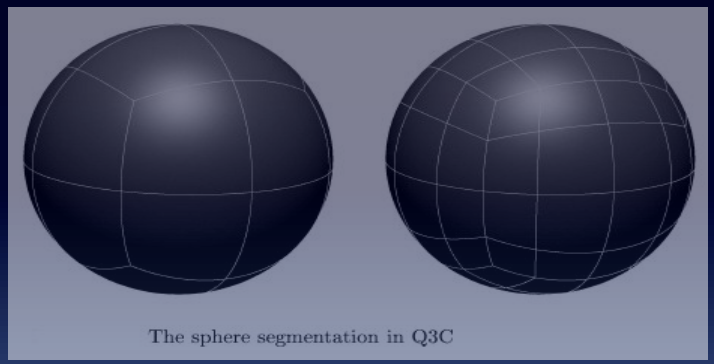
Premier problème étudié (suite)

- Observation 1 :

  Beaucoup de requêtes utilisent la *distance* 2D

  - Cone search (spatial range query)
  - Cross-match (distance join)

- Les données astrométriques sont indexées
  - Par numéro de cellule d'une grille (ex. HTM, HealPix)
  - Des librairies spécifiques manipulent ces index (ex. Q3C)
  - Les données de numéros proches sont plus susceptibles d'être accédées ensemble


The sphere segmentation in Q3C

Jointure distance avec MR

- Observation 2 :

La répartition des objets dans l'espace est très hétérogène

- Problème 1:
  - Le mapping basique risque de déséquilibrer fortement la charge des reducers ou dépasser la capacité mémoire
- Problème 2:
  - Les objets de 2 parties peuvent joindre
  - Comment éviter le transfer des objets entre reducers ?

Jointure distance avec MR (suite)

- Idée – estimer les densités locales à la volée
  - Génère des métadonnées statistiques dans un arbre
  - Dérive une répartition optimale théorique
  - En déduit la fonction *map*
- Tâche du Reducer
  - Applique localement un algorithme connu de jointure distance (**données en mémoire**)

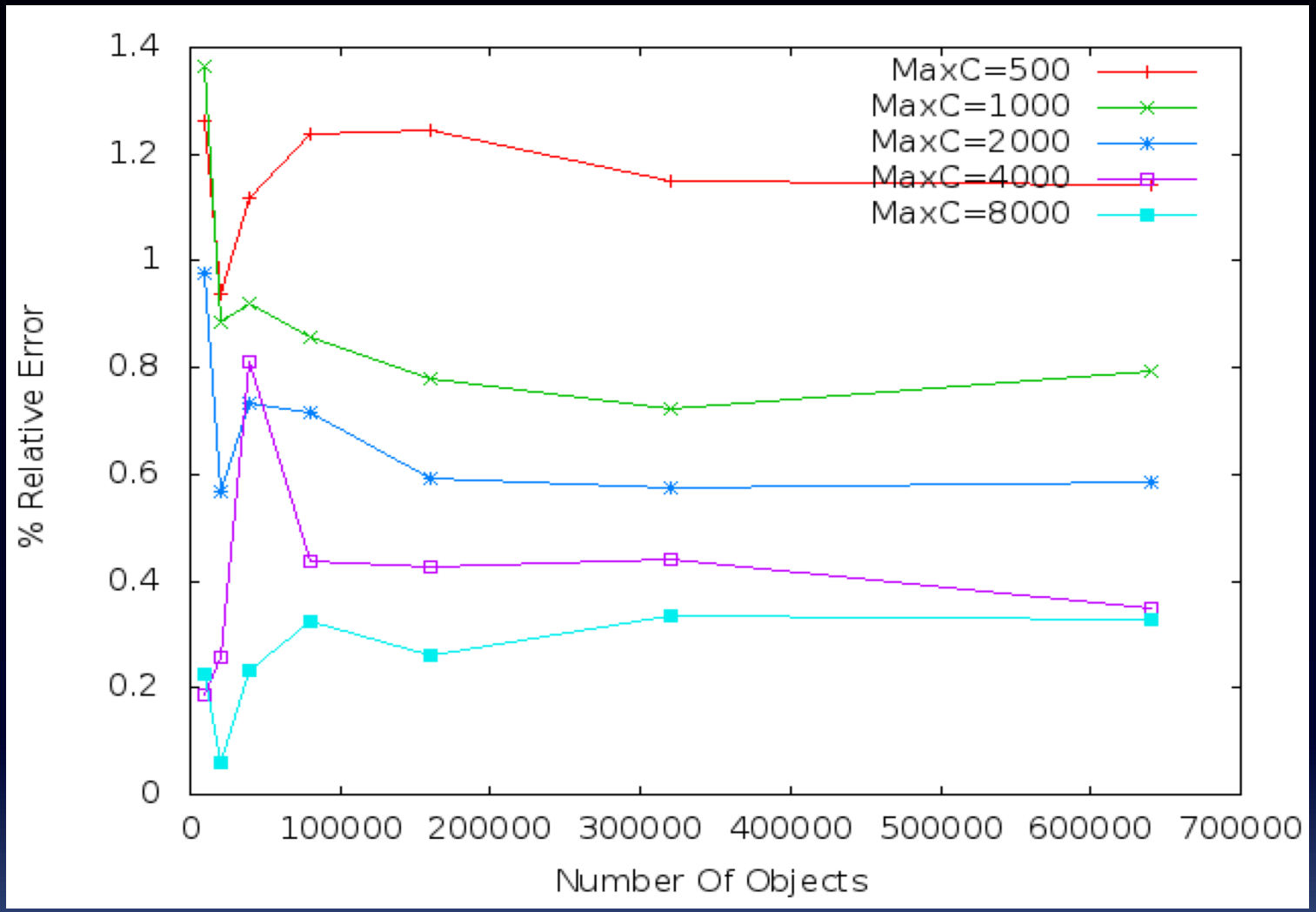
- Avantages :
  - Equilibre la charge des reducers
  - S'adapte à leur capacités (mémoire)
  - Evite les transferts et minimise des réplicats sur les bords

Expérimentation

- Données : 8 to 77 millions d'objets simulés
- 6 serveurs * 8 cores (48 VM)

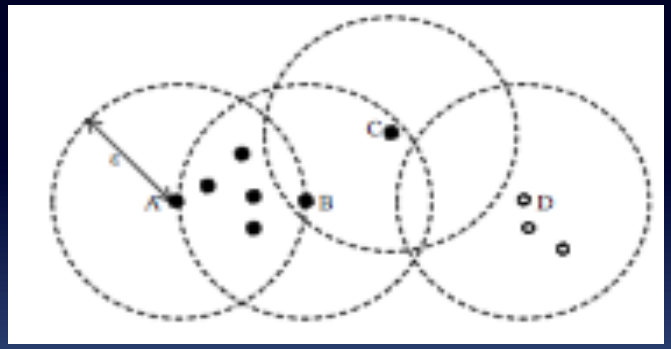
- Mesures
  - Surcout de construction de la structure statistique
  - Erreurs d'estimation de densités
  - Taux de réplications en fonction des paramètres
  - Temps de réponse
  - Scalabilité

Erreurs d'estimation de densité



Perspectives côté BD

- Finaliser / se comparer et compléter :
  - Plus de tests et comparaison avec d'autres travaux / techniques BD
  - Elargir le spectre des requêtes
  - Intégrer la 3ème dimension observée par GAI et le problème d'incertitude lié
  - Appliquer à l'analyse de corrélations et à la fouille de données
  - Et d'autres analyses statistiques ?



Merci de votre attention!