

# PetaSky

- Réponse à l'appel d'offre du défi de la MI du CNRS sur les BigData 2012 : Mastodons

La disponibilité de très **grandes masses de données** et la **capacité de les traiter de manière efficace** est en train de modifier la manière dont nous faisons de la science

- Consortium pluridisciplinaire :
  - IN2P3 : LPC, APC, LAL, CC
  - INS2I : LIMOS (Clermont-Ferrand), LIRIS (Lyon)

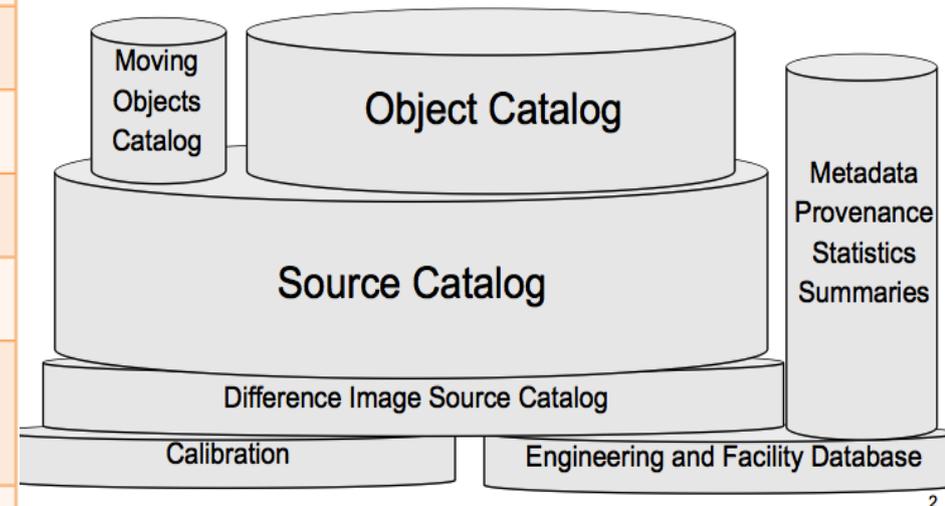
# *Position du projet*

- Gestion de données astrophysiques de type LSST
  - Volume (~100 Pb en fin de projet, ~6 Pb de catalogues)
  - Complexité relationnelle (Objets, Sources, ...)
  - Hétérogénéité des formats (Images, catalogues)
- **Problématiques**
  - Passage à l'échelle et intégration de données (Axe 1)
  - Visualisation (Axe 2)
  - Analyse et fouille de données (Axe 3)

# Axe 1 : stockage et gestion de données

(Mohand Said ; Emmanuel Gangler)

| Table                   | Taille | #enregistrements | #colonnes (arité) |
|-------------------------|--------|------------------|-------------------|
| Object                  | 109 TB | 38 B             | 470               |
| Moving Object Source    | 5 GB   | 6 M              | 100               |
| Forced Source           | 3.6 PB | 5 T              | 125               |
| Difference Image Source | 1.1 PB | 32 T             | 7                 |
| CCD Exposure            | 71 TB  | 200 B            | 65                |
|                         | 0.6 TB | 17 B             | 45                |



+ ~80 tables

- Quel modèle logique : relationnel, multidimensionnel, graphe ?
- Quel modèle physique : ligne, colonne, arbre, multidimensionnel ?
- Quelle architecture matérielle : partitionnement, multi-threading

# Axe 1 : état des lieux

- Plateformes matérielles :
  - Au LPC : 5 machines, 1 maître + 4 nœuds, 10 TB/nœud
  - Au CC : 250 machines, 100 GB/nœud
- Logiciel :
  - **Qserv** ( baseline LSST / MySQL) installé, data en cours de déploiement (PT1.1 puis W13)
    - Objectif = déploiement au CC pour FDR ; benchmarking
  - Benchmarks de Hive et HadoopDB (LIRIS)
    - Objectif = exploration des solutions alternatives
  - En discussion : comparaison avec solutions commerciales (Orange, Amazon)

# Axe 2 : visualisation

(Florent Dupont, Guy Barrand)

- Visualisation d'images (et de données) de très grande taille
- Axes explorés :
  - Représentation des données ;
  - exploitation de la temporalité ;
  - interface aux catalogues.
- Exemple : découpage multi-résolution  
<http://liris.cnrs.fr/petasky/images/>
- Plan de travail :
  - Solutions web
  - Mur d'images

# Axe 3 : fouille de données

(Engelbert Mephu, Eric Aubourg)

- **Objectifs**

- Classification des objets astrophysiques
  - Supervisée et non-supervisée
  - Caractérisation a priori ou non
- Emergence de propriétés à partir des données

- **Verrous**

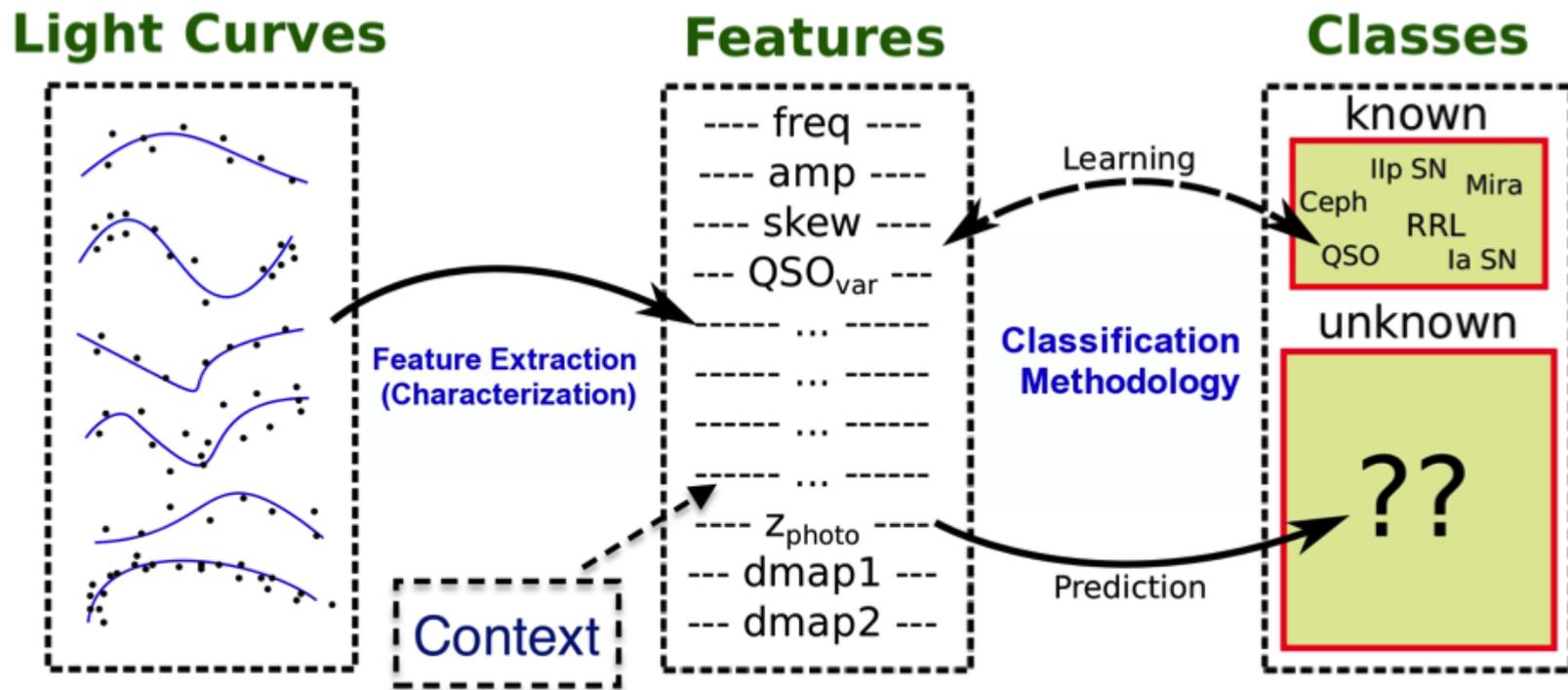
- Passage à l'échelle
- Caractérisation couplée à l'apprentissage
- Incertitudes
- Accès aux données (interface avec Axe 1)

- **Feuille de route pour 2013**

- Séparation étoile-galaxie
- Photo-z

# Axe 3 : fouille de données

A road map for ML light curve classification:



**See:** Richards et al. (2011) ApJ, 733, 1; arXiv:1101.1959

Bloom & Richards (2011) arXiv:1104.3142



# Points pour discussion

# Discussion

- Intégration de PetaSky et de l'effort LSST
  - Davantage d'intégration serait un plus :
    - Développements à l'intérieur de Qserv de solutions alternatives
    - Travail en connexion avec l'U. Wasington sur la visualisation
    - Projet d'analyse commun sur fouille de données
  - Difficultés : méthodes de travail différentes
    - INS2I : Manpower sur financement projet
  - Le contour du projet n'est pas fermé

# Feuille de route

- Axe 1 : (Base de données)
  - Déployer Qserv standard : PT1.1 (100 G avec sources), puis W13 (2TB)
  - Prioritisation à faire (LPC) entre
    - Validation Qserv 250 nœuds,
    - Benchmark de la solution,
    - Etudes de performances mono-noeuds.
    - Support aux utilisateurs externes (PetaSky) sur Qserv
  - Etude des performances Hive et Hadoop (LIRIS)
  - Amazon ?

# Feuille de route

- Axe 2 : Visualisation
  - Contour et use-case à éclaircir.
    - Inputs externes bienvenus
- Axe 3 : fouille de données
  - Test-case des photo-z bien défini
    - Premier pas : séparation étoile-galaxie
    - Implication souhaitée des physiciens IN2P3.
  - Autres idées à décanter
    - Fouille != classification