

Petasky

<http://com.isima.fr/Petasky>

Mastodons program of the Interdisciplinary Mission of CNRS



- INS2I



Fusion de **3 projets** Petasky + Amadeus + GAIA



- INS2I
 - ♦ LAL (UMR CNRS 8607, Paris)
 - ♦ Centre de Calcul de l'IN2P3/CNRS (CC-IN2P3)
- INSU
 - ♦ LAM (UMR CNRS 7326, Marseille)



Petasky: scientific challenges

- Management of scientific data in the fields of **cosmology and astrophysics**
 - ➔ Very large amount of data
 - ➔ Complex data (e.g., images, uncertainty, multi-scales...)
 - ➔ Heterogeneous formats
 - ➔ Various and complex processing (images analysis, reconstruction of trajectories, ad-hoc queries and processing, ...)
- Scientific challenges
 - ➔ Scalability
 - ➔ Data integration
 - ➔ Data analysis
 - ➔ Visualisation
- Application context : **LSST project**

From Astronomy to astroinformatics

- Modern digital detectors, CCDs,
- Early use of scientific computing, numeric simulations, ..
 - ➔ Antikythera mechanism, between 150 to 100 BC
 - ➔ Supernovae Cosmology Project, 1986
 - 1024x1024 CCD camera, 2 megabytes every five minutes
 - ➔ International Virtual Observatory
 - ➔ Web of astronomical data
- ➔ Sloan Digital Sky Survey (SDSS)
 - 2.5 m Telescope, 54 CCD imager
 - Operational since 2000
 - In 2010, a total archive of 140 TB
- ➔ GAIA, launched in 12/2010, operational since 2013/2014
- A culture of sharing data
 - ➔ Data with non-commercial value (more open than healthcare or biomedical science fields)

The LSST project

Large Synoptic Survey telescope

A new window over the sky: Telescope of 8.4 m



The New Sky



Polishing the Primary Mirror

WIDE

A large aperture, wide field survey telescope and 3200 Megapixel camera to image faint astronomical objects across the sky.

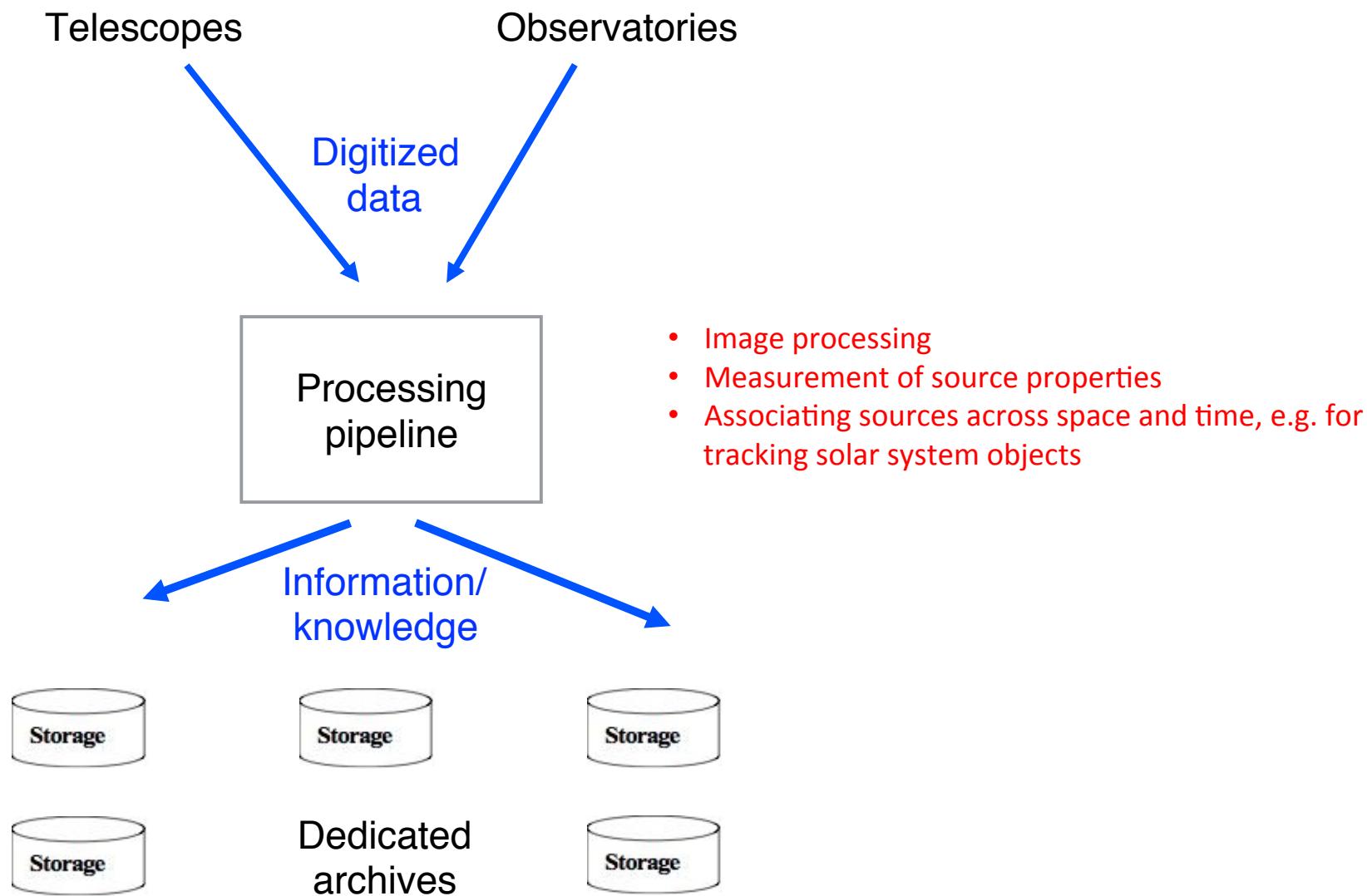
FAST

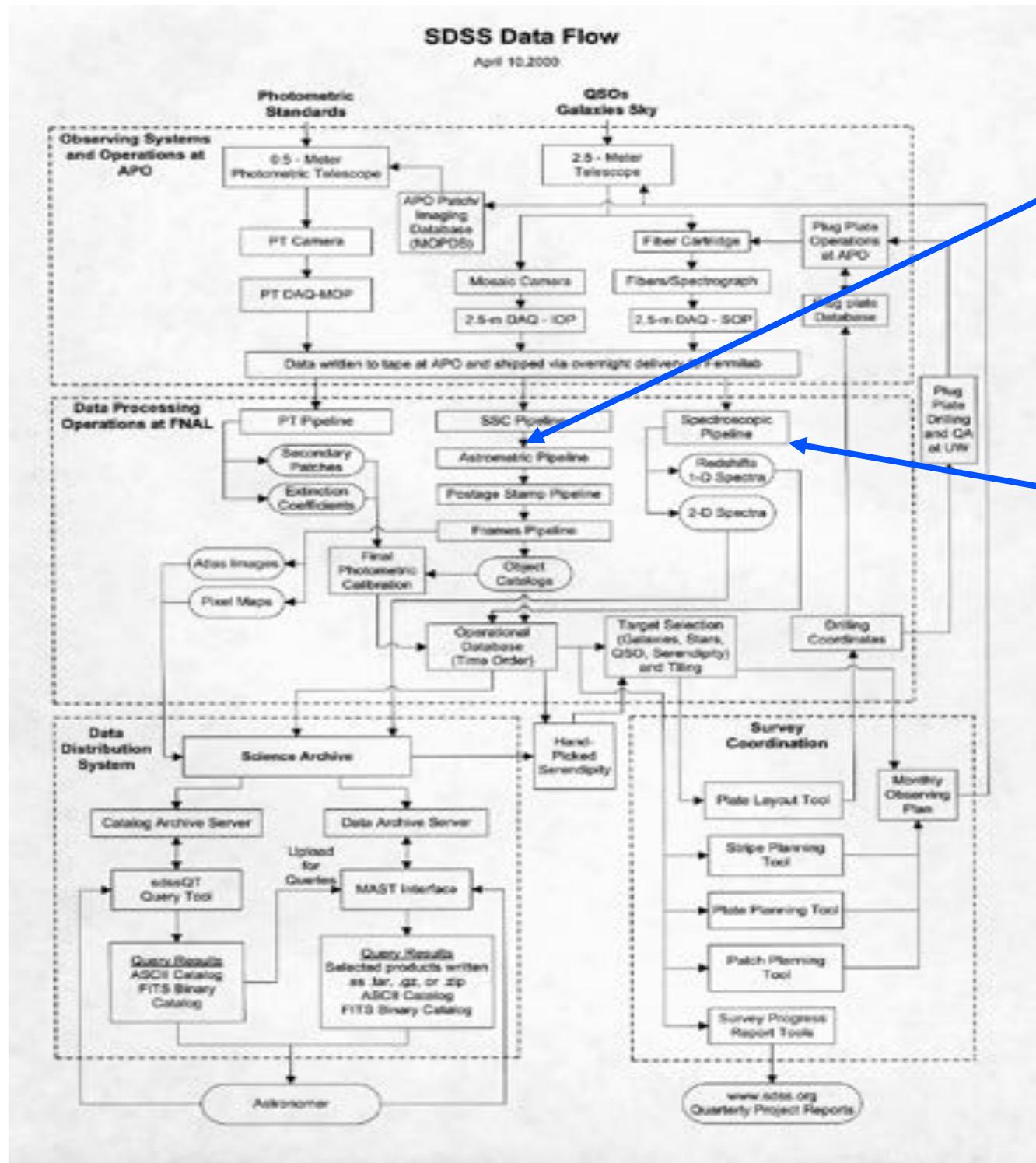
LSST will rapidly scan the sky, charting objects that change or move: from exploding supernovae to potentially hazardous near-Earth asteroids.

DEEP

LSST's images will trace billions of remote galaxies, providing multiple probes of the mysterious dark matter and dark energy.

Data-driven discovery in Astrophysics

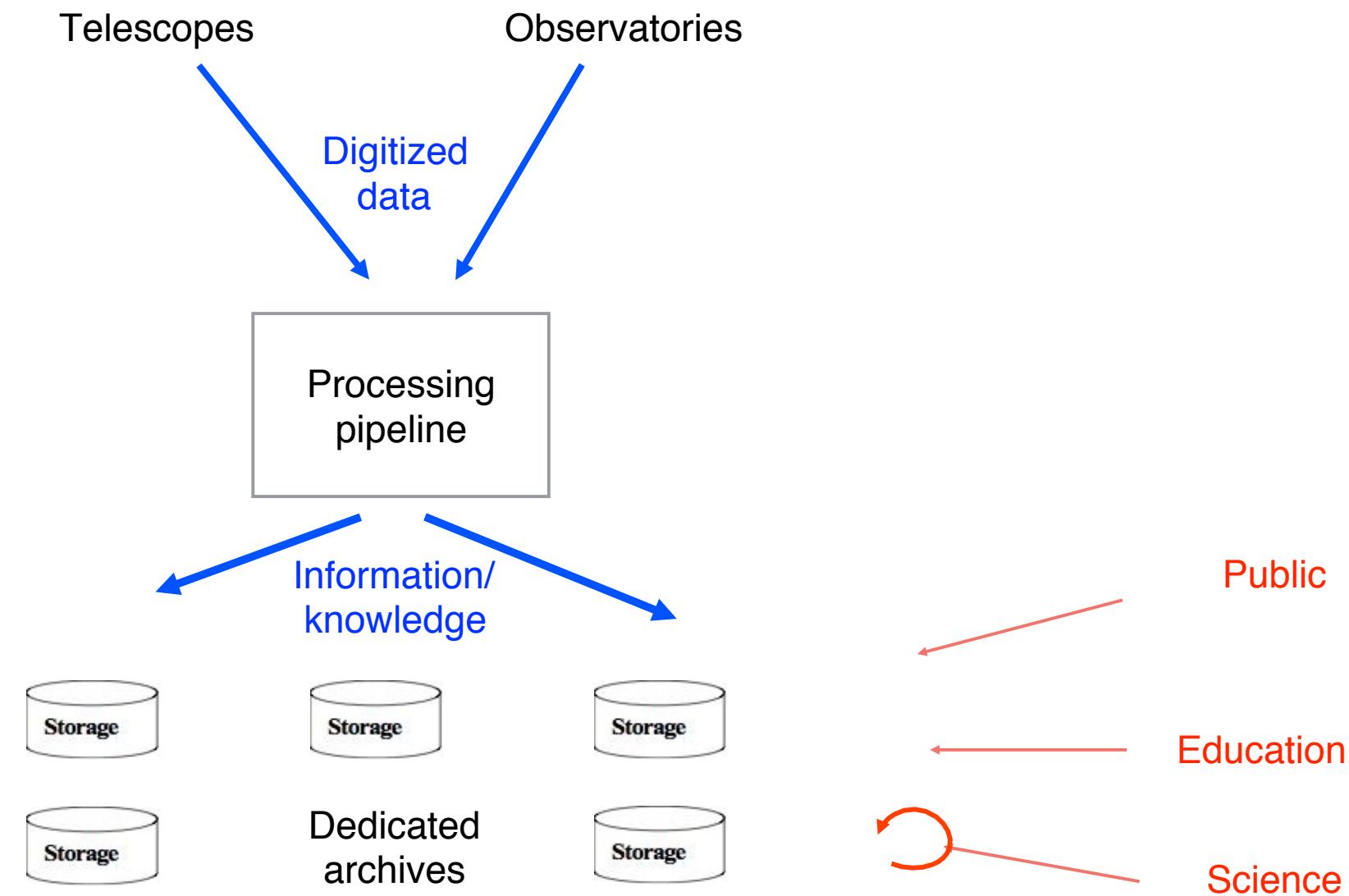




Astrometric pipeline
Compute object positions from raw images

Spectroscopic pipeline
redshifts, classification of objects as galaxies, stars, quasars

Data-driven discovery in Astrophysics



Data management challenges in LSST

“How much the (LSST) project will tell us about our solar system, the dark energy problem and more, will depend on how well we can process the information the telescope and its camera send back to us - an estimated sum of around ten petabytes of data per year.”

(Mari Silbey, Space: the big data frontier, <http://www.smartplanet.com/blog/thinking-tech/space-the-big-data-frontier/12180>)

“Plans for sharing the data from LSST with the public are as ambitious as the telescope itself”

Anyone with a computer will be able to fly through the Universe, zooming past objects a hundred million times fainter than can be observed with the unaided eye. The LSST project will provide analysis tools to enable both students and the public to participate in the process of scientific discovery.



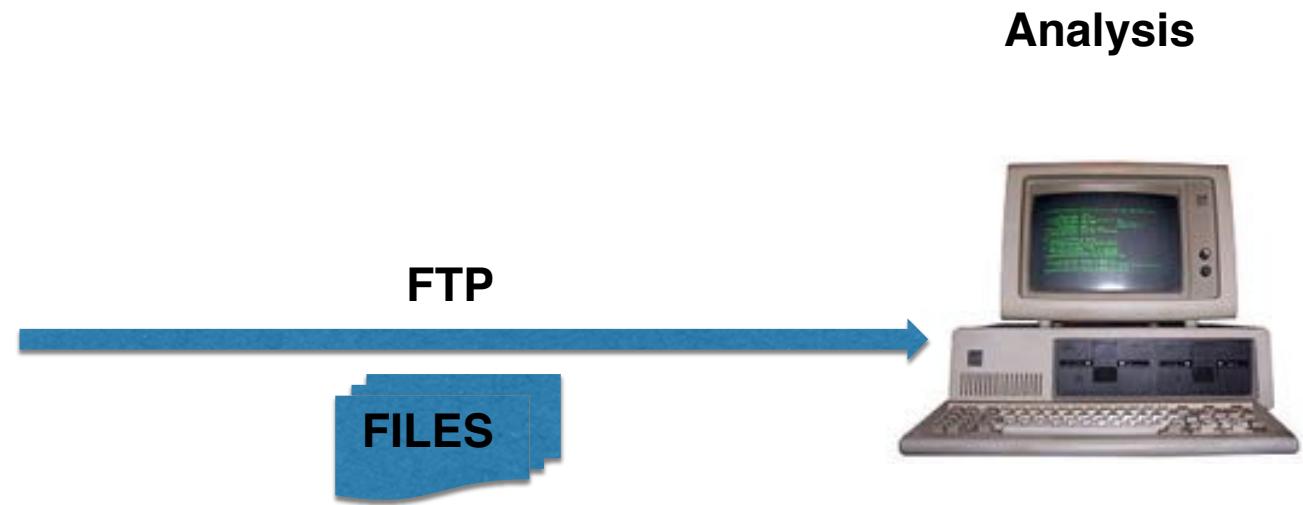
Jim Gray, Turing award 1998

FTP-GREP Model

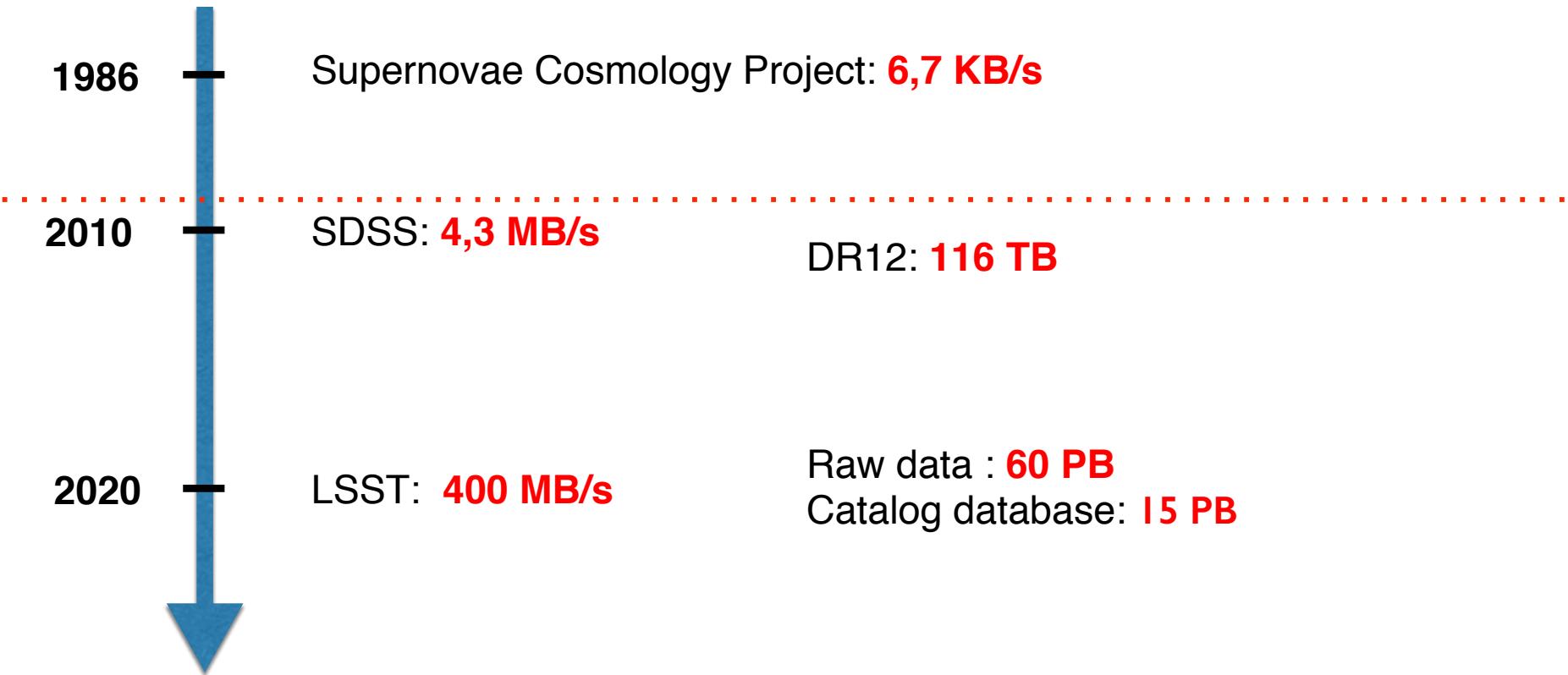
Remote Archive



Analysis



Evolution of data volumes



The LSST scientific database

Table	Size	#tuples	#attributes
Object	109 TB	37 B	470
Moving Object	5 GB	6 M	100
Source	3,6 PB	350 B	125
Forced Source	1,1 PB	32 T	7
Difference Image Source	71 TB	200 B	65
CCD exposure	0,6 TB	17 B	45

Data challenges

- Data Volume
 - Typical sky survey may detect 10^8 - 10^9 sources (stars, galaxies, ..)

Table scan : $\approx 3\text{h}$ to scan 1 TB

Parallelization

- < 2 minutes with 100 HD
- 1TB/sec : 10 000 HD (Google Dremel)
- 1 TB in less than 1 minute with Oracle DBMS : 2 RAC nodes + Multiples infiniBand adapters (Maklee)

- Data quality
 - Missing data
 - Lack of large training data sets

Modern data management technologies

- Massive parallelization
- Virtualization and cloud computing
- Data distribution
- Data storage
 - Row store vs. column store
 - (sophisticated) Indexes
- New computing paradigms
 - Failures resilience
 - Coordination
- Complexity theory and cost models

Main thesis

How to form a new generation of scientists capable to exploit the new technologies to pursue science goals at an unprecedent scale?

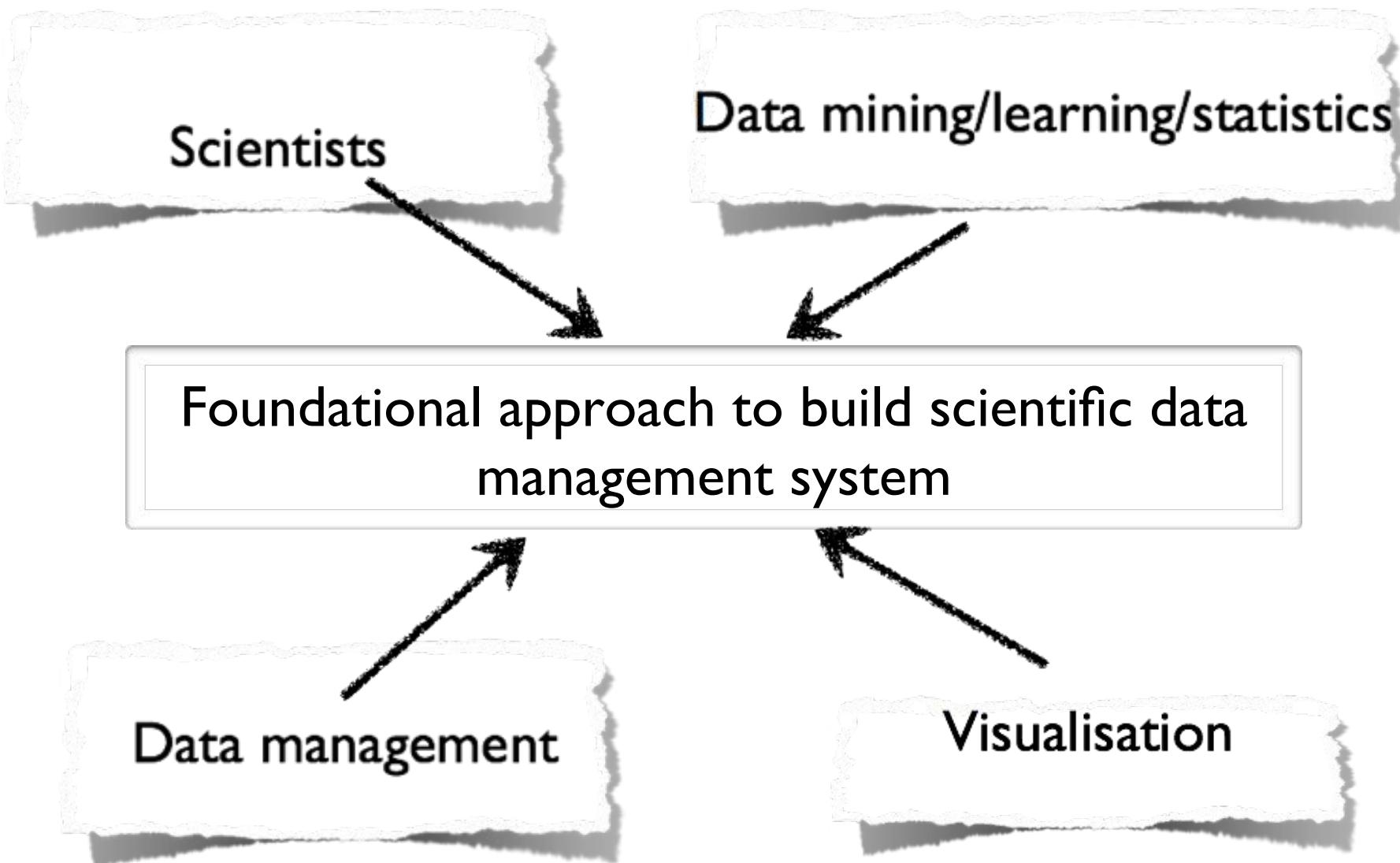
G.Longo¹

It should not be up to the scientists but to the technology (**data management system**) to overcome the computing barriers between them and the data

¹Talk, workshop « New challenges in astro- and environmental informatics in the Big Data era », May, 2014, Szombathely, Hungary

Working cross disciplines

... working cross cultures



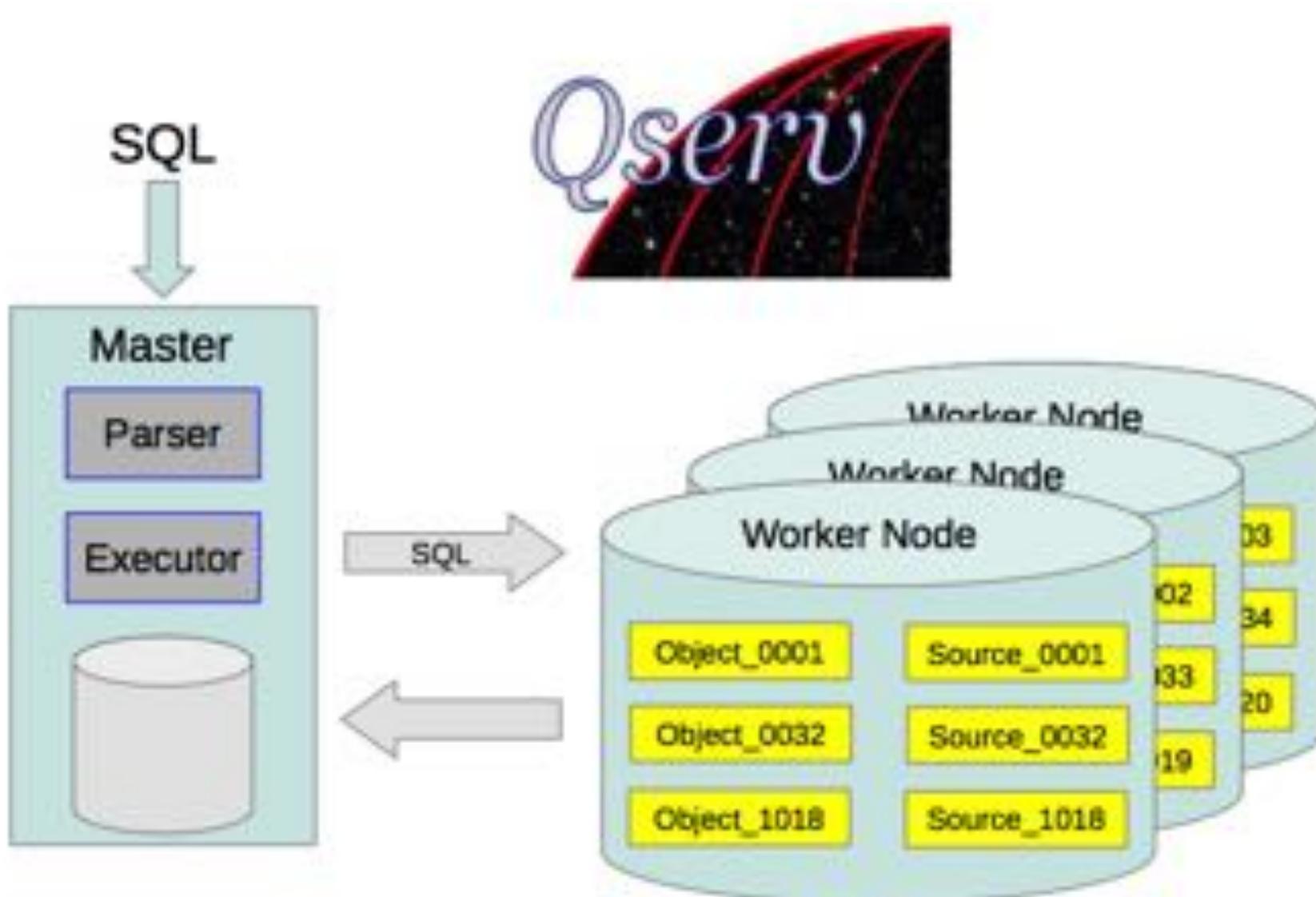
A multidisciplinary research program

- Astronomy and astrophysics
 - Computer science/applied mathematics
 - Machine learning and data mining
 - Bigdata management
 - Visualization
 - Distributed computing
 - Education

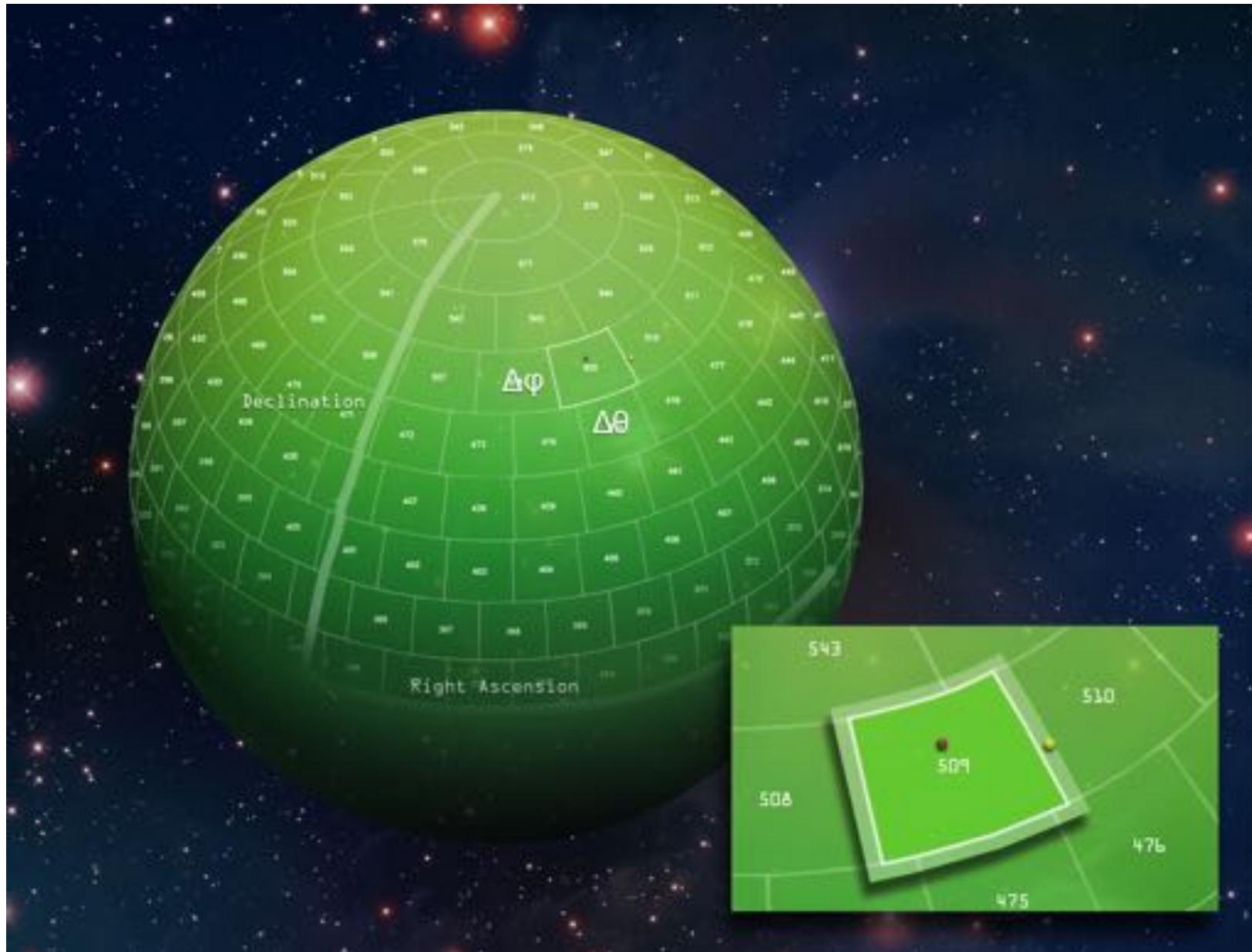
➤ Solutions could be exported to other big data science (biology, health, ...)

➤ Drive innovations in industry

QSERV Data Management System



Spatial partitioning



LSST queries per level of difficulty

Supported queries

- Retrieve any type of information about a single object (identified by a given objectId), including full time series.

SELECT * FROM Object JOIN Source USING (objectId) WHERE objectId = 293848594;

Few seconds

- Retrieve any type of information about a group of objects in a small area of sky, including neighborhood-type queries.

SELECT * FROM Object WHERE qserv_areaSpec_circle(1.0, 35.0, 5.0/60)

= 1 hour

- Analysing light curves across large area.

SELECT O.objectId, myFunction(S.taiMidPoint, S.psfFlux) FROM Object AS O JOIN Source AS S USING (objectId) WHERE O.varProb > 0.75 GROUP BY O.objectId;

= 1 day (24h)

- Analysing light curves of faint objects across large area.

SELECT O.objectId, myFunction(V.taiMidPoint, FS.flux) FROM Object AS O JOIN ForcedSource AS FS ON (O.objectId = FS.objectId) JOIN Visit AS V ON (FS.visitId = V.visitId);

= 1 week

LSST queries per level of difficulty

Expensive/impossible queries

- **Expensive queries**
 - Find objects far away from other objects (for a large number of objects).
Question: what is the largest distance we should plan to support for distance based queries involving (a) small number of objects, (b) all objects on the sky?
 - Sliding window queries: Find all 5 arcmin x 5 arcmin regions with an object density higher than rho
- **Impossible queries**
 - Large size results
 - Select all pairs of stars within 1 arc min of each other in the Milky Way region.
 - Expensive or hidden computation (e.g., Join)
 - Near neighbor query on the Source or ForcedSource table
 - Joining large tables between different LSST data releases
 - Time series analysis of every object
 - Cross-match with very large external catalog (e.g. LSST with SKA)
 - Any non-spatial join on the entire catalog (Object, Source, ForcedSource)
 - Join of Source with ForcedSource

Examples of LSST User Defined Functions

- `q3c_ang2ipix(ra, dec)` -- returns the ipix value at ra and dec
- `q3c_dist(ra1, dec1, ra2, dec2)` -- returns the distance in degrees between (ra1,dec1) and (ra2,dec2)
- `q3c_join(ra1, dec1, ra2, dec2, radius)` -- returns true if (ra1, dec1) is within radius spherical distance of (ra2, dec2). It should be used when the index on `q3c_ang2ipix(ra2,dec2)` is created.
- `q3c_ellipse_join(ra1, dec1, ra2, dec2, major, ratio, pa)` -- like `q3c_join`, except (ra1, dec1) have to be within an ellipse with major axis major, the axis ratio ratio and the position angle pa (from north through east)
- `q3c_radial_query(ra, dec, center_ra, center_dec, radius)` -- returns true if ra, dec is within radius degrees of center_ra, center_dec. This is the main function for cone searches. function should be used if when the index on `q3c_ang2ipix(ra,dec)` is created)
- `q3c_ellipse_query(ra, dec, center_ra, center_dec, maj_ax, axis_ratio, PA)` -- returns true if ra, dec is within the ellipse from center_ra, center_dec. The ellipse is specified by major axis, axis ratio and positional angle. function should be used if when the index on `q3c_ang2ipix(ra,dec)` is created)
- `q3c_poly_query(ra, dec, poly)` -- returns true if ra, dec is within the postgresql polygon poly.
- `q3c_ipix2ang(ipix)` -- returns a 2-array of (ra,dec) corresponding to ipix.
- `q3c_pixarea(ipix, bits)` -- returns the area corresponding to ipix at level bits (1 is smallest, 30 is the cube face) in steradians.
- `q3c_ipixcenter(ra, dec, bits)` -- the function returning the ipix value of the pixel center of certain depth covering the specified (ra,dec)

Petasky: data management challenge

Techniques to build an **efficient** and **easy to use** data access and analysis system at a **reasonable cost**

- | From | → | To |
|--|---|---|
| <ul style="list-style-type: none">• Specialized Hardware• Programming• Ad-hoc optimization | | <ul style="list-style-type: none">• Commodity machines• Querying• Generic system |

Need for more research on

- ✓ Abstraction adequate to the scientific domain
 - Array data model (SCiDB)?
- ✓ Support of user defined functions
- ✓ Optimization techniques embedded in the data management system
- ✓ Scalability of information integration framework and datamining techniques

Petasky: explored approaches

- Big data management approaches
 - ✓ Distributed and parallel systems
 - MapReduce-like approaches (shared nothing architecture)
 - Parallel DBMS (shared all thing architecture)
 - Spatial partitioning (QSERV for LSST)
 - ✓ Column store DBMSs (Vertica, MonetDB, ...)
 - ✓ Data integration to the rescue
 - Declarative approach
- Data mining and knowledge discovery
 - ✓ Interactive exploration of large datasets
 - ✓ Parallel mining of dependencies
 - ✓ New clustering algorithms
 - One-pass based algorithm
 - Incomplete enumeration
 - ✓ Using neural networks to estimate redshift distributions

Space of solutions and associated challenges

! Clearly beyond the capacities of centralized systems

- Distributed and parallel systems
 - ✓ Data distribution
 - ✓ Computation distribution
 - ✓ Failure resilience
- Storage model
 - ✓ row store vs. column store
 - ✓ (sophisticated) Indexes
- Benefit from modern hardware
- Scalable datamining and machine learning techniques
- Complexity theory and cost models
 - ✓ Standards measures: I/O, data transfer, ..
 - ✓ Cost of coordination

Exploration des données

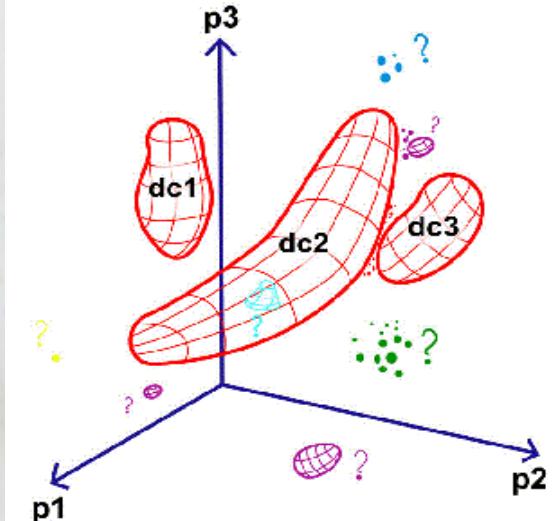
- **Le point de vue de l'astronome : analyse**

- Catégorisation des objets (pertinence, erreurs)
- Détection de rareté (découverte, anomalie)
- Réduction de dimensionnalité
- Mesure de paramètres (régression, statistique)
- Prise en compte des incertitudes

- **Le point de vue de l'informaticien : fouille**

- Données sous forme tabulaire (interface avec SGBD)
- Algorithmique efficace en dimension élevée (> 1000 , $> 10^{10}$ entrées)
- Recherche de relations (implications, dépendances fonctionnelles)
- Visualisation des données

Data Mapping and a Search for Outliers



S. G. Djorgovski,

Exploration interactive de masse de données basée sur les contre-exemples

- Formulation d'une **hypothèse** par un expert
- Sur les données, si l'hypothèse est fausse proposer des **contre-exemples**
 - Ils parlent à l'expert, sont puisés dans les données
 - Calcul de mesures de la qualité de l'hypothèse sur les données
 - **Itération**
 - Sélection automatique d'un sous-ensemble des données sur lesquelles l'hypothèse pourrait être vérifiée (apprentissage)
 - Reformulation de l'hypothèse
- Intérêt
 - **Complexité calcul réduite** : énumération vérification
 - L'expert est au centre du processus

Cas d'application

- Prototype iSQL
- Données ExoPlanet (Corot)
 - 97717 étoiles, 62 attributs
 - 50 planètes confirmées
 - 175 absences de planètes

- Hypothèse :

- Présence d'une planète

Q0 = SELECT * FROM EXOPL
WHERE object IS NOT NULL
AND object = 'p'

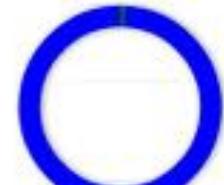
- Hypothèse réécrite :

Q'0 = SELECT * FROM EXOPL
WHERE AMP11 <=0.001717
AND MAG_B > 13.425

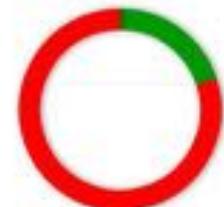
- Evaluation :

- Propose 1279 cibles potentielles
- Dont 20% des planètes connues
- Dont 0% de faux positifs

Type	Number of tuples	%
Examples found	10	1%
Counter-examples found	0	0%
Other tuples found	1279	99%



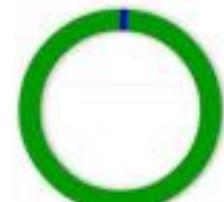
Type	Number of tuples	%
Examples found	10	20%
Examples not found	40	80%



Type	Number of tuples	%
Counter-Examples found	0	0%
Counter-Examples not found	175	100%



Type	Number of tuples	%
Other tuples found	1279	1%
Tuples not found	96213	99%



Emergence d'une communauté interdisciplinaire ..

- Projet européen COST BigSkyEarth
- Co-encadrements de thèses
- Plateforme Glactica
 - IR : Programme PlaSciDo, INS2I
 - Equipement : CPER région Auvergne
- Groupe de travail **Maestro**
 - MAsses de données En aSTROnomie et astrophysique du GDR Madics (<http://www.madics.fr/actions/actions-en-cours/maestro/>)



MERCI