Научные вызовы Технологиям баз данных

Олег Бартунов, ГАИШ-МГУ

Общество стало другим

Общество стало другим

- Информационное общество
  - главные продукты производства — это информация и знание
  - Основа капитализация компании — это знание. Основной вопрос — как накопить знание и обмениваться знанием
  - CEO, CTO, **CIO-senior information officer**
  - NSF-CDI (Cyber-enabled Discovery and Innovation)
    - **От данных к знанию**
    - **изучение базовых элементов инфраструктуры киберобщества.**

ВЕБ — Универсальная платформа обмена информацией

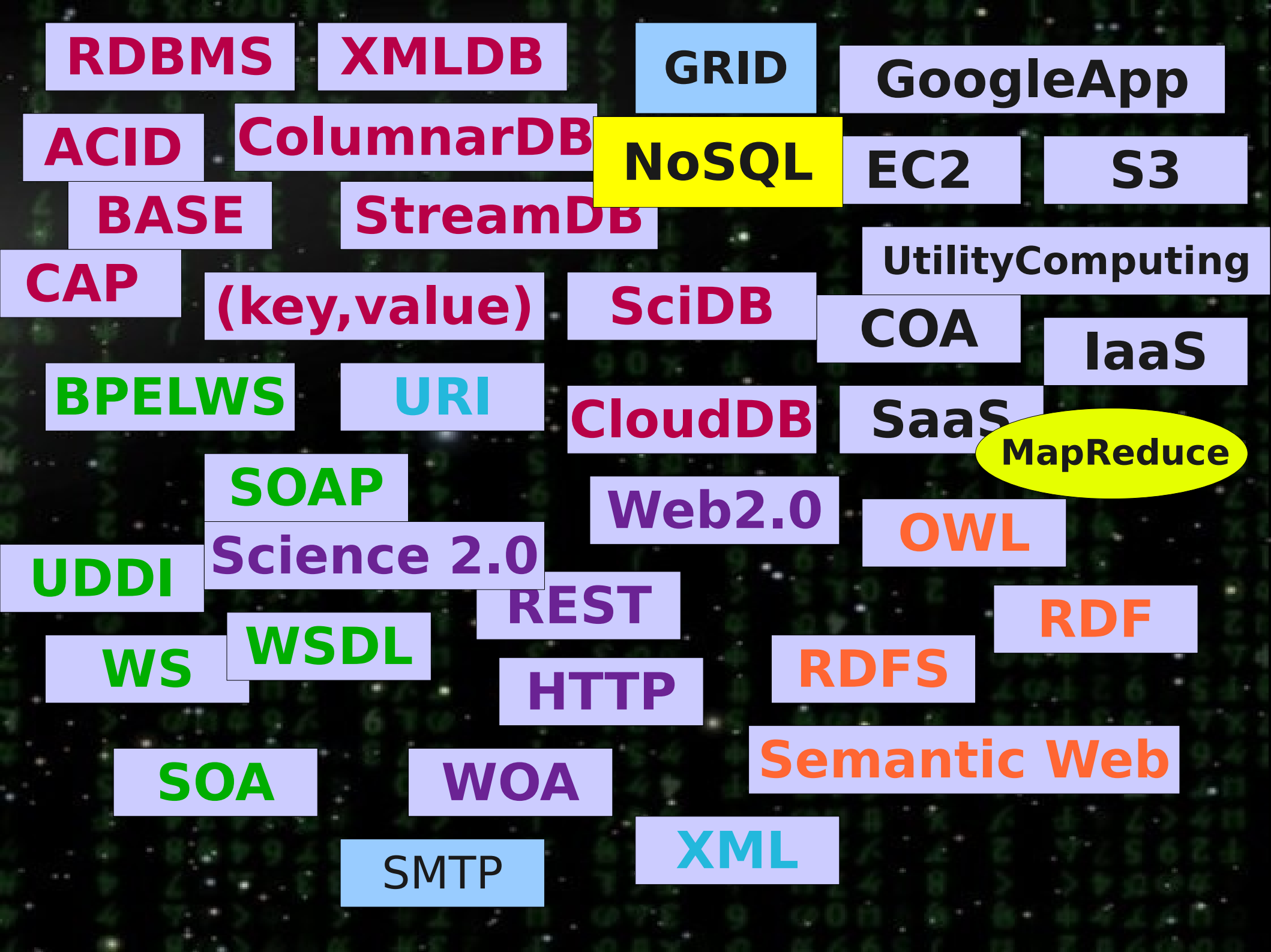
Web Services
UDDI, WSDL,SOAP
ПРОГРАММЫ

Счастье !
ПРОГРАММЫ
ДАННЫЕ

WWW
URI,HTML,HTTP
TEXT

Semantic Web
RDF,RDF(s),OWL
ДАННЫЕ

Email: @address, text, smtp


RDBMS
XMLDB
GRID
GoogleApp
ACID
ColumnarDB
NoSQL
EC2
S3
BASE
StreamDB
UtilityComputing
CAP
(key,value)
SciDB
COA
IaaS
BPELWS
URI
CloudDB
SaaS
MapReduce
SOAP
Web2.0
OWL
Science 2.0
UDDI
REST
RDF
WSDL
WS
RDFS
HTTP
Semantic Web
SOA
WOA
XML
SMTP

Наука стала другой

Наука стала другой

- eScience — составная часть информационного общества - синтез науки и информатики
  - роль информации и ее обработка становится доминирующей
- Переход на е-стадию — реальная ситуация, которая затронула много наук, оперирующих громадными объемами информации
  - физика (эл. частицы и высоких энергий, науки о земле, погода, астрономия, социология, медицина, биология

Наука стала другой

- eScience — глобальная коллаборация людей и ресурсов для решения новых задач науки и промышленности
  LHC: 50+ стран, 200+ институтов
- Это технология, инфраструктура
  - физика — Grid (Open Grid)
  - астрономия — VO (Virtual Observatory)
  - биология — биоинформатика

Наука стала другой

- Другой «шаблон» работы в науке:
  - коллективность, узкая специализация ...
- Административная и финансовая научная политика: финансирование ожидает быстрых результатов ! «Early Science»
- Очень много информации/данных:
  - Распределенные, разнородные

Очень много данных !

VLDB -> XLDB

Very Large → Extremely Large

XXX Tb XXX Pb

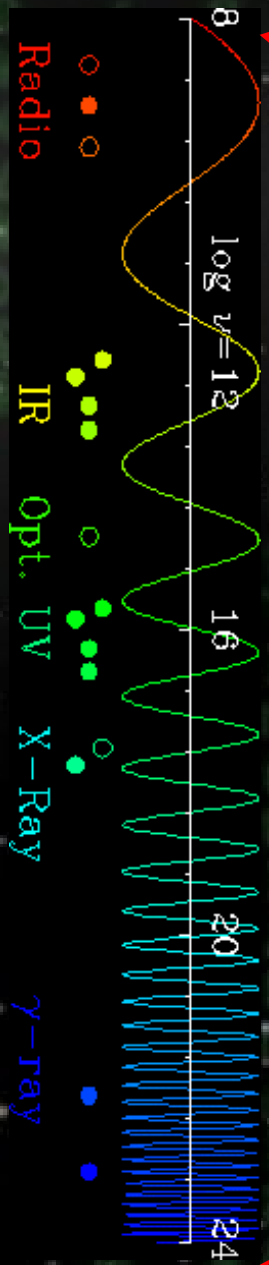
Стало очень много данных !

-
- Успехи в технологии сенсоров
  - Большие размеры
  - Качественные (малошумящие)
  - Доступные
  - Все диапазоны спектра
- «Sensor-centric» science !
- Мощные Машины — основные производители и потребители данных

СверхБольшие научные БД

- **Тихо Браге - (1570-1601) ~ 500Kb**
- **SDSS — 2007 год 3 Tb (метаданные)**
- **Библиотека конгресса — 15 Tb**
- **LSST** — большой обзор неба
  - 8.4 м зеркало, 3.2 Gpx CCD
  - 49 млрд. объектов, 2.8 млрд источников
  - 30 Tb/night, 100 Tflop обработка
  - 10 лет: 60 Pb raw data, Каталог — 30 Pb
- **LHC** — Large Hadron Collider
  - 15 Pb ежегодно, 100K CPU
  - 200 центров в ~ 40 странах

Астрономия стала всеволновой



Дипольные антенны

Параболические антенны

Болометры

Телескопы-рефлекторы

Зеркала косого падения

Кодирующие маски

Атмосферные черенковские телескопы, Широкие атмосферные ливни

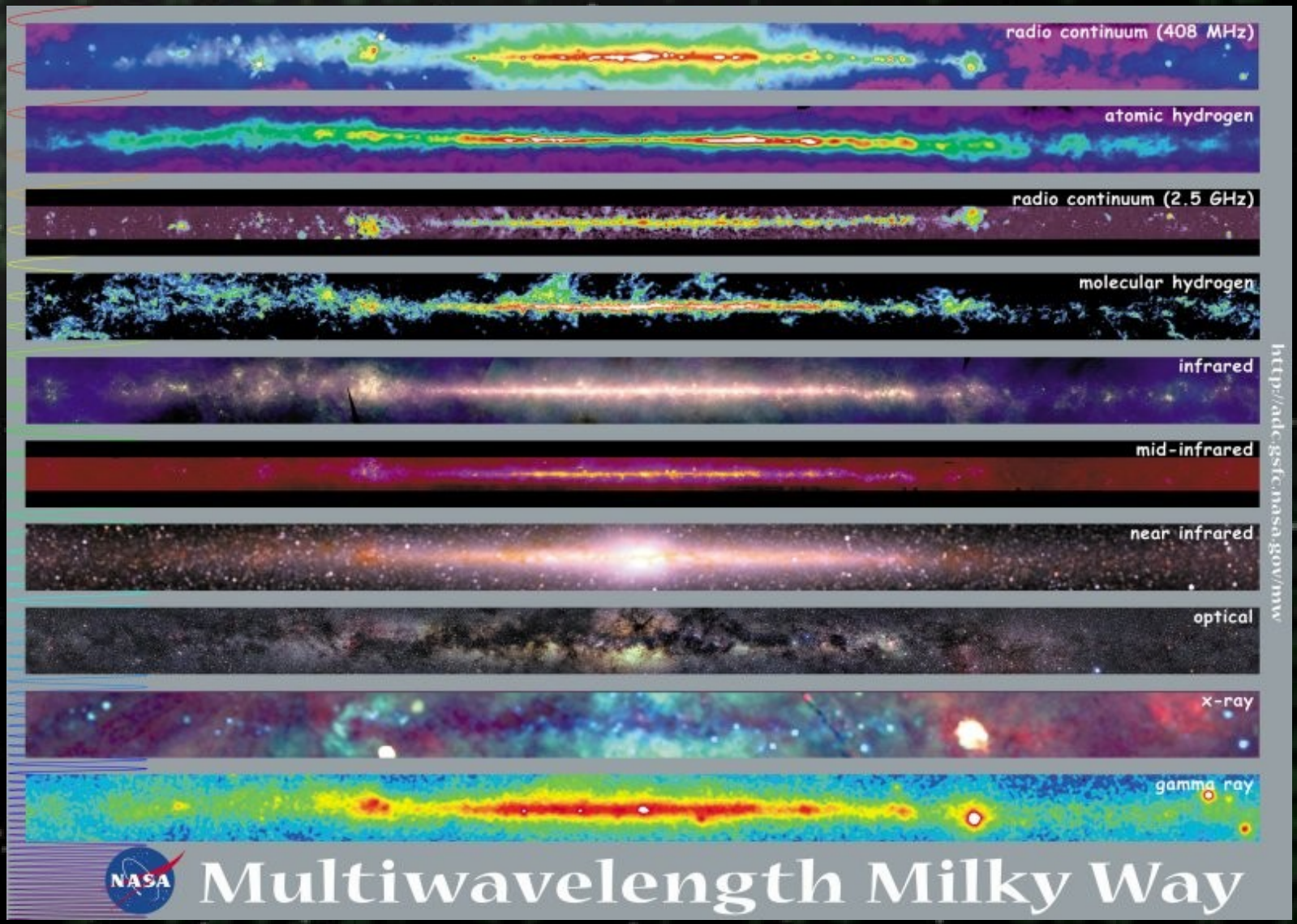
Нейтринные телескопы: (Солнце, SN 1987A)

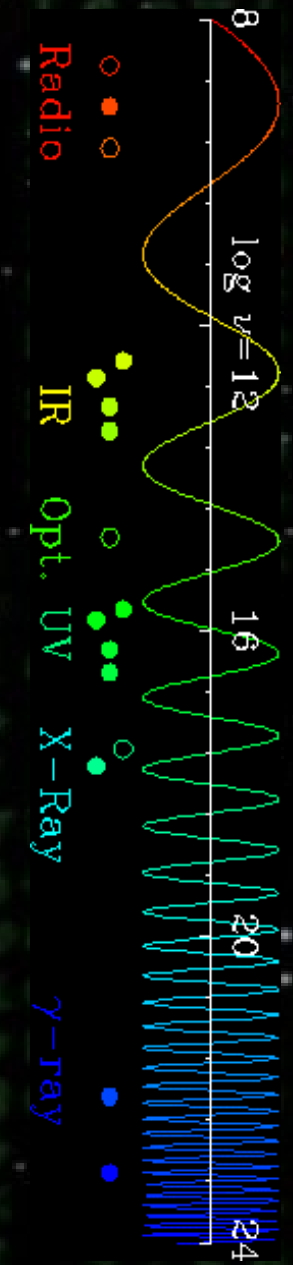
Гравитационные антенны (начинают работать)

Космические лучи

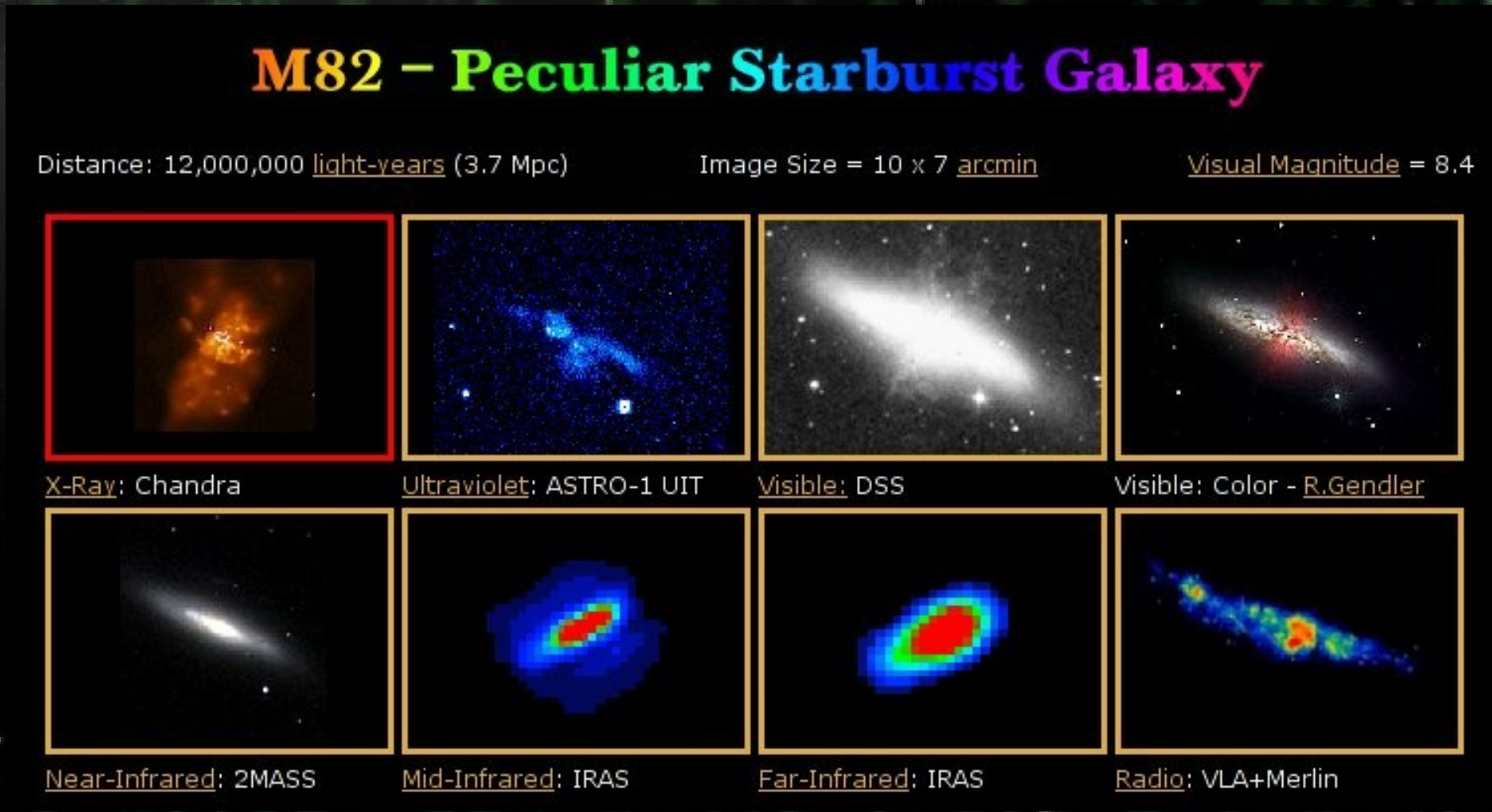
Разная процедура обработки сырых данных !

Астрономия стала всеволновой





Астрономия стала всеволновой



Что такое научные данные ?

- «Сырые» (raw) данные — данные непосредственно из сенсора (Nikon's raw 12/14 bit). Хранятся вечно !
- Обработка (cooking) «сырых» данных — сложная процедура (Capture NX) , изменяется. Мы все больше абстрагируемся от объекта изучения
  - Изучение звезд: **глаз-**фото-эоп-ccd
  - Открытие частиц: треки в камере Вильсона, сейчас CCD в LHC
  - Raw converter: Capture NX, NX2

Что такое научные данные ?

- Результат обработки - Научные данные
  - Астрономы используют каталожные данные — таблицы атрибутов различных источников (звезды, галактики,...)
  - Сжатие данных — несколько % от картинок
  - Удобно представляются в rdbms — индексируются по координатам на небе
  - Но иногда сырые данные (изображения, спектры, списки событий) бывают нужны.
  - Jpg (8-bit) → www.flickr.com

Что вовлечено в процесс ?

- Сенсоры, инструменты, Данные
- Алгоритмы, программы, конфиги, параметры
- Компьютерные системы (железо, OS, software)
- Документация (design, обработка)
- Люди, Организации
- Статьи
- Все может иметь версии !!!

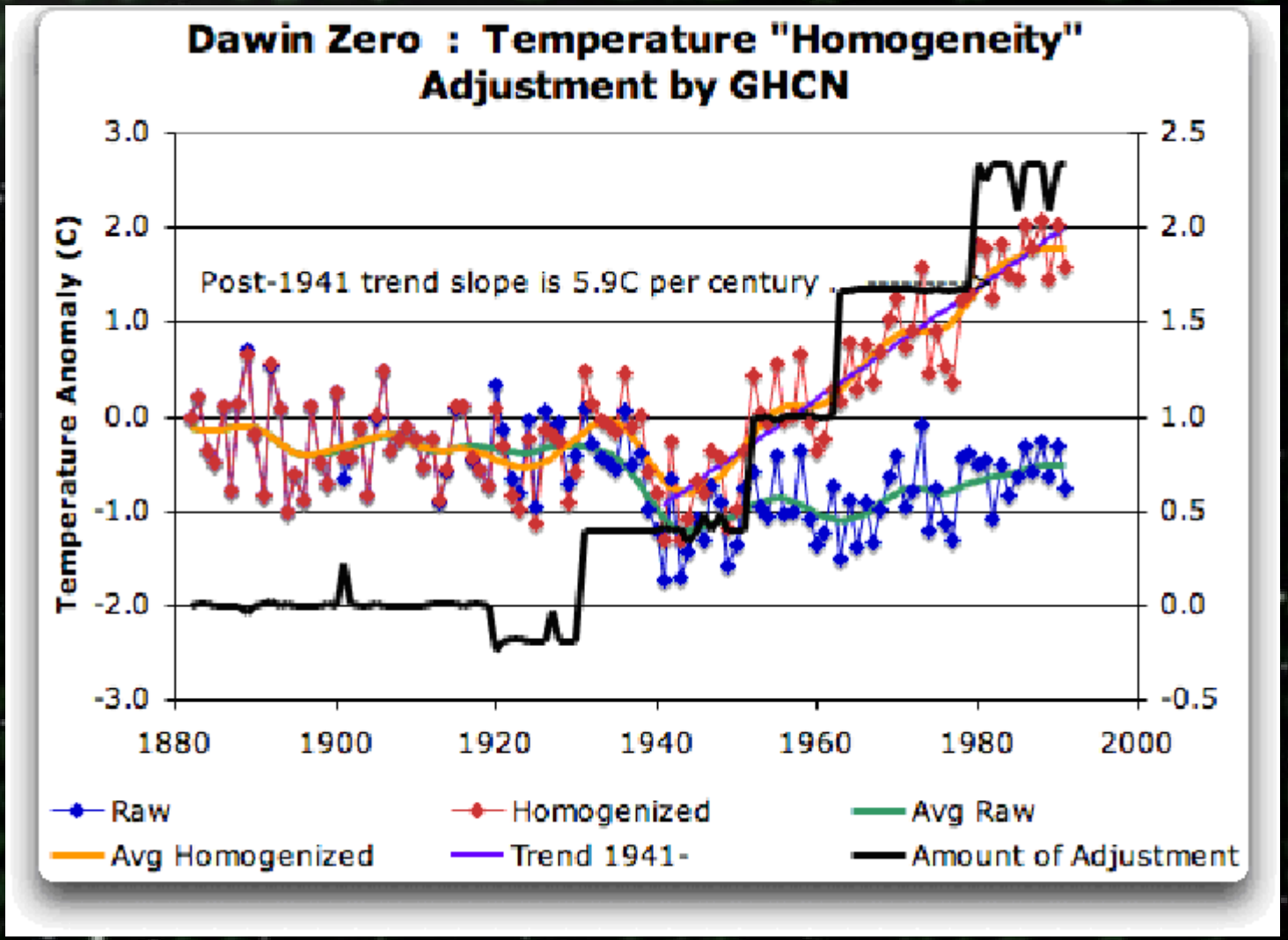
Принцип науки

Воспроизводимость Научных Результатов

под угрозой !

Climategate !

http://wattsupwiththat.com/2009/12/08/the-smoking-gun-at-darwin-zero/



Воспроизводимость научных результатов

- Как ссылаться на данные (purl) ?
  - Название журнала, том, страница, год
  - Oops! This link appears to be broken.
  - Данные меняются в архивах
  - «Бегущая ссылка» ....?page=237
- Как обеспечить доступность данных ?
  - Свободный обмен данными
  - Независимость от одного вендора
  - Вопросы лицензии

Воспроизводимость научных результатов

- Как обеспечить сохранность данных ?
  - «Сырые» данные хранить вечно !
- Как обеспечить целостность данных ?
  - Большинство проектов хранят метаданные в БД, а объекты - вне.
- Как проследить происхождение данных (data provenance, lineage)
  - Качество данных, Источники данных
  - Какие операции привели к появлению или изменению данных ?

Что вовлечено в процесс ?

- Сенсоры, инструменты, Данные
- Алгоритмы, программы, конфиги, параметры
- Компьютерные системы (железо, OS, software)
- Документация (design, обработка)
- Люди, Организации
- Статьи

Что важно ?

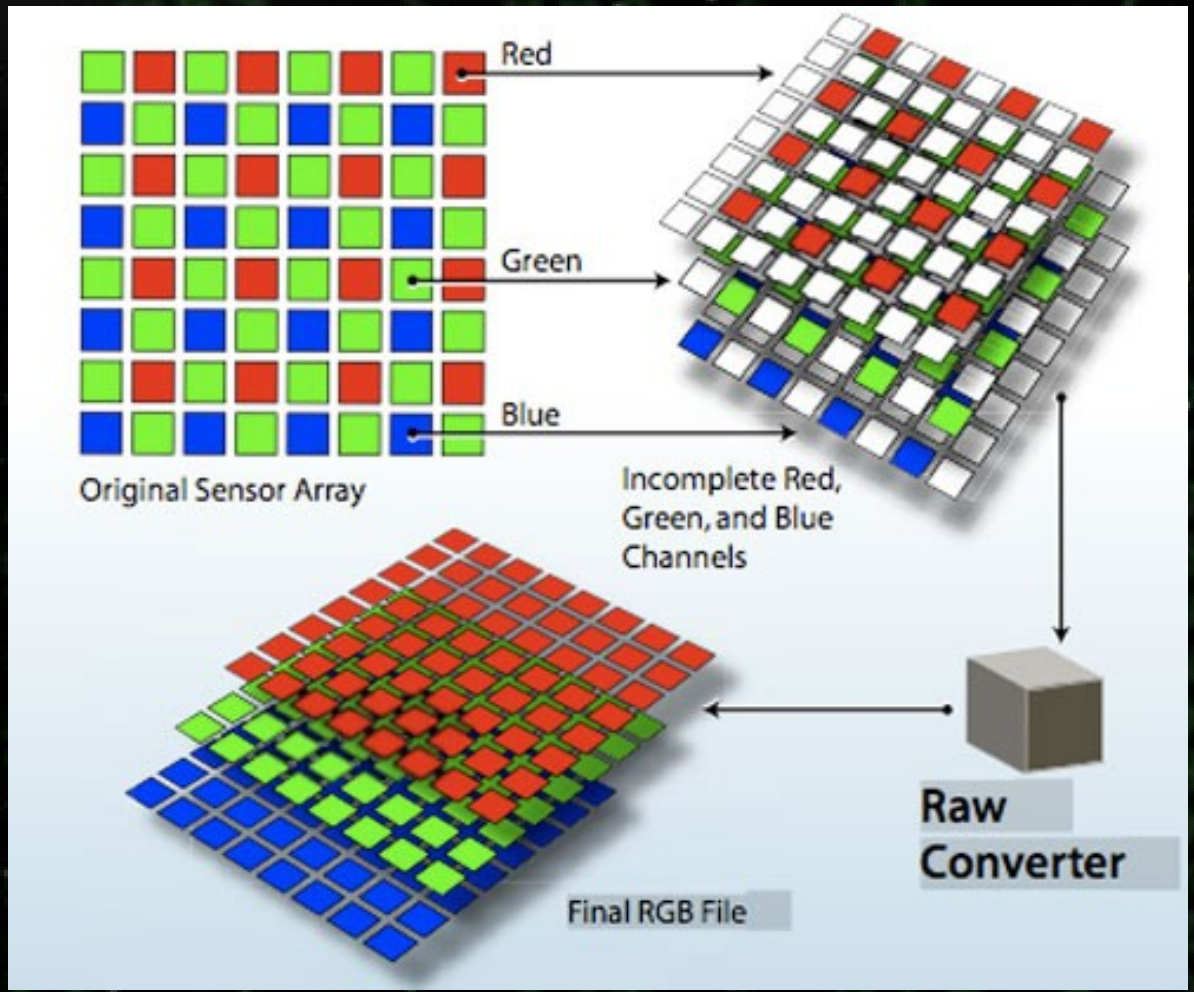
- Маленькая мелочь может привести к эффекту бабочки
- Климатгейт — подрыв доверия
- Модель CPU, hostname - скорее всего, не важны
- Информации о provenance должно хватить для воспроизведения анализа независимым исследователем и подтверждения результатов и выводов

Специфика научных данных

- Данные только добавляются, WORM. Изменение данных приводит к появлению новой версии.
- Научные данные - это результаты экспериментов, вычислений
- Данные в науке как правило имеют погрешности измерений (error bar)
- Цензурированные данные используются в астрономии, медицине, биологии

Специфика научных данных

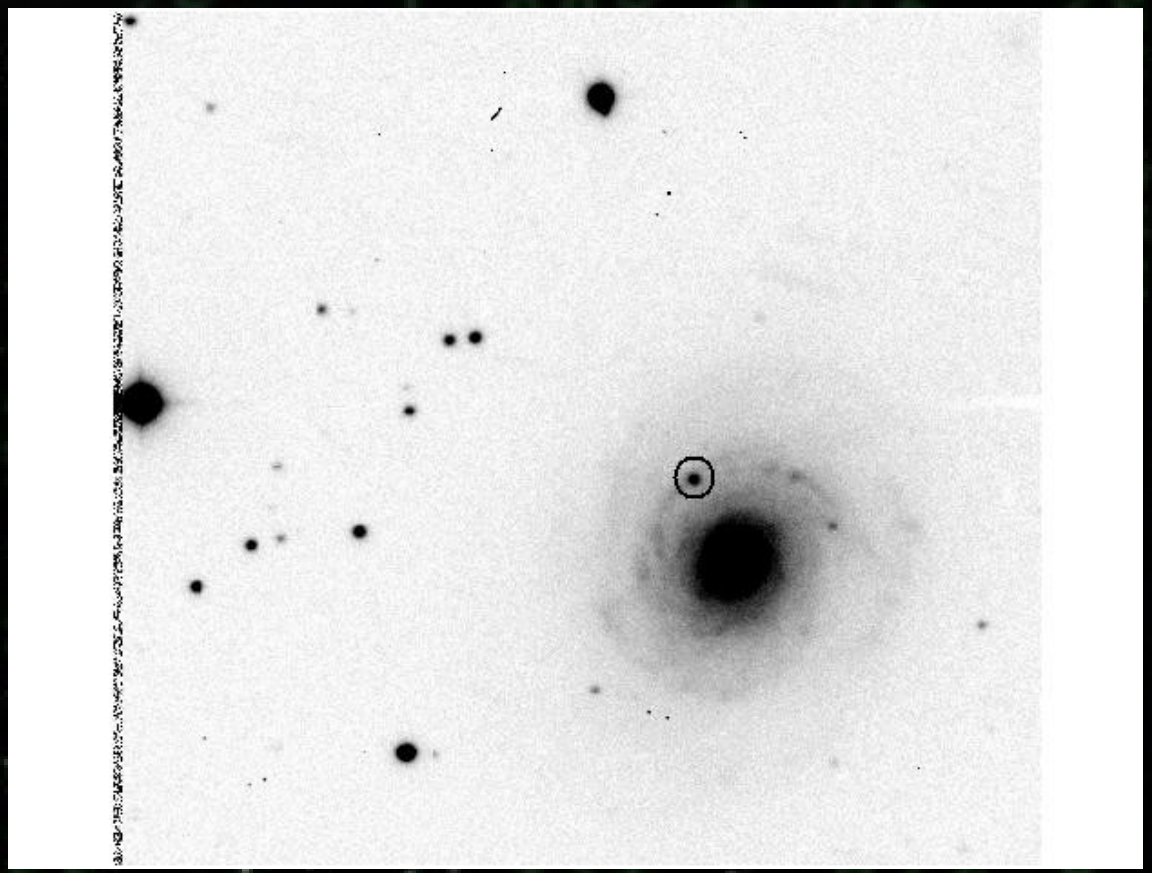
Современная наука построена вокруг сенсоров - «sensor-centric science»



Специфика научных данных

Сенсорные данные естественно хранить в массивах

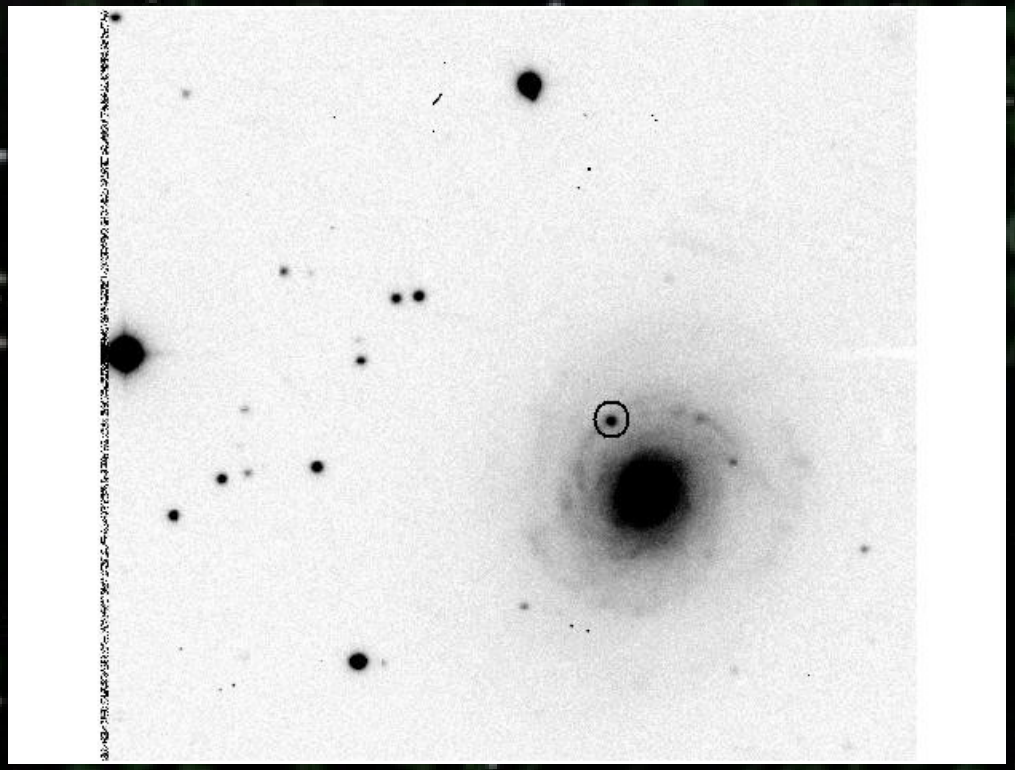
- SN 2008fv в галактике NGC 3147 в Драконе — 2d массив пикселей (Дмитрий Цветков, ГАИШ)



Специфика научных данных

Основные операции работают с 'окрестностями'

RDBMS неэффективны, так как не имеют внутренней упорядоченности как у массивов !



Специфика научных данных

I	J	V
0	0	0.02
0	1	0.01
0	2	0.002
...	...	...
1	2	0.5
...	...	...
3	3	0.02

Декларативный язык запросов и логическая модель хороши !

Но на практике проблема состоит в необходимости для каждой ячейки массива перечитать массив, чтобы найти записи из ее окрестности.

```
SELECT A1.I, A1.J, AVG(A2.V)
  FROM Observation A1, Observation A2
 WHERE A2.I BETWEEN A1.I - 1 AND A1.I + 1
   AND A2.J BETWEEN A1.J - 1 AND A1.J + 1
 GROUP BY A1.I, A1.J;
```

Компьютерные технологии изменились

- Классические RDBMS <- system R
  - 1 большой $$$$ компьютер (mainframe)
  - БД не помещалась в оперативную память - дисковые структуры, индексы
  - Треды как квазипаралелльность
- Компьютерный мир изменился
  - «Железо» дешевое, много RAM, много ядер
  - Распределенность: кластеры-датацентры
  - Быстрые и широкополосные сети

Разнообразие видов СУБД

- Классические RDBMS модернизируются
  - Расширения на основе встроенных средств (например, GiST,GIN в PostgreSQL)
  - Кластеризация (shared-nothing, middleware)
  - Cloud Computing адаптеры — PostgreSQL Plus + Elastra -> Amazon WS
  - Специализированные сервера с СУБД в FPGA для сверхбыстрых OLAP
  - Параллельное выполнение запросов
    - GreenPlum, AsterData — технология MapReduce

Yahoo Everest

- Порог петабайтных БД преодолен !
  - Yahoo Everest (2008) -
  - 10Pb в 2010
  - PostgreSQL +
    - Распределенное атрибутное хранилище (компрессия)
    - Параллелизм и Асинхронность
    - .....
- Не хватает - уходят в сторону Hadoop

Разнообразие видов СУБД

- Классических RDBMS или не хватает, или они избыточны (ACID → A . . D)
- Появились специализированные СУБД,
  - Потоковые СУБД (StreamDB) — данные не важны ! Только агрегаты
- XML базы данных — слабоструктурированные данные
- NoSQL (~40 !) - простая структуры данных (key,value), no join. ACID->BASE

NoSQL databases (wikipedia)

Document store

* Lotus Notes
* CouchDB
* MongoDB
* Apache Jackrabbit
* Colayer
* XML databases
  - o MarkLogic Server
  - o eXist

Graph

* Neo4j
* AllegroGraph

Tabular

* BigTable
* Mnesia
* Hbase
* Hypertable

Key/value store on disk

* Tuple space
* Memcachedb
* Redis
* SimpleDB
* flare
* Tokyo Cabinet
* BigTable

Key/value cache in RAM

* memcached
* Velocity
* Redis

Eventually-consistent key-value store

* Dynamo
* Cassandra
* Project Voldemort

Ordered key-value store

* NMDB
* Luxio
* Memcachedb
* Berkeley DB

Object database

* Db4o
* InterSystems Caché
* Objectivity/DB
* ZODB

Разнообразие видов СУБД

- Вертикально-ориентированные СУБД (C-store, Vertica, BigTable, Cassandra....)
  - Хранение по колонкам, а не по записям
  - Лучшая компрессия данных
  - Сильно меньше размер метаданных
  - Оптимизация для колонок фикс. длины
  - Оптимизированы для чтения, OLAP
  - «Легкие» транзакции — undo из лога(*), redo из реплик
  - Прямой доступ к упорядоченным данным (в обычных RDBMS нужен Btree индекс)

Разнообразие видов СУБД

- Вертикально-ориентированные СУБД
  - Оптимизированы для GRID
  - Shared-nothing горизонтальная сегментация
  - Параллельное выполнение запросов
  - Большие блоки, нет выравнивания данных
  - TimeTravel — простая изоляция транзакций, версионность,
  - Распределенные транзакции без 2PC

Используемые решения

- Метаданные в СУБД, куча данных в файлах (xPb)
  - BaBar, LHC, LCLS, PanSTARRS
- Все в СУБД
  - Ebay(xPb, Terradata, GreenPlum), WalMart (xPb, Oracle), SDSS (xTb MS SQL), Genome
- Home-grown системы (xPb)
  - ATT, Google, Yahoo, Amazon, Facebook
- Все это нетиражируемые решения !

Что делается

- Тренды
  - Shared-nothing parallel database
  - Дешевые сервера
  - Агрессивная компрессия
- Попытки продвинуть технологии
  - Aster Data, Vertica, ParAccel, Greenplum, Neteza, Teradata

- Классические RDBMS
  - + функциональность, расширяемость
  - + декларативный язык
  - - лишний overhead (ACID)
  - - слабая масштабируемость (middleware)
- Масштабируемые (key, value)
  - + native распределенность
  - + масштабируемость
  - - бедная модель данных

- Нерешенные проблемы:
  - Массивы как «1st class citizen»
  - Provenance
  - Данные с ошибками

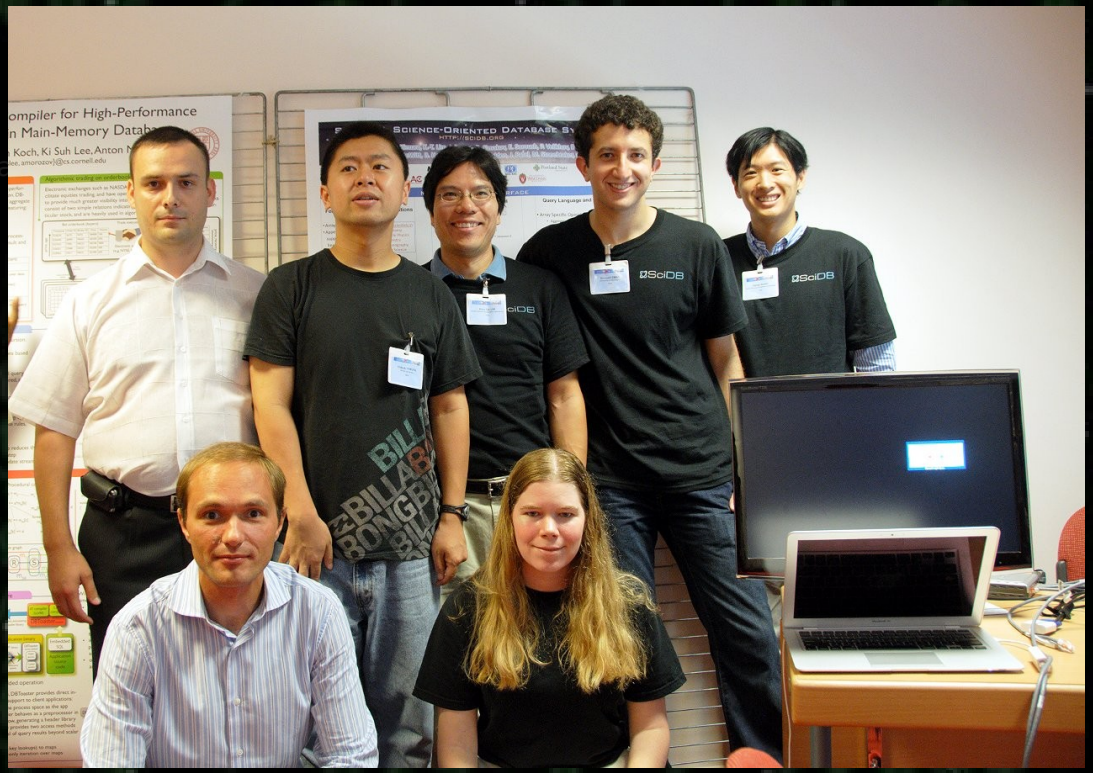
XLDB

- XLDB — сверхбольшие базы данных
- Большая наука - астрономия, биология, науки о земле, физика высоких энергий + разработчики СУБД (MIT, Yahoo, Microsoft, IBM, BEA) + EBay
  - 3 встречи: 2007,2008,2009
- Традиционные СУБД бесполезно «штопать» - они разрабатывались для других машин, других данных, других требований

SciDB.org

- SciDB - новая СУБД с нуля !
- LSST — тестовая площадка
- Майк Стоунбрейкер — лидер проекта

- Ingres,Postgres,
- Illustra,StreamDB,
- Vertica, VoltDB



Полный цикл анализа данных в БД !

- Data cleaning (cooking)
- Feature extraction
- Data mining
- Data sharing

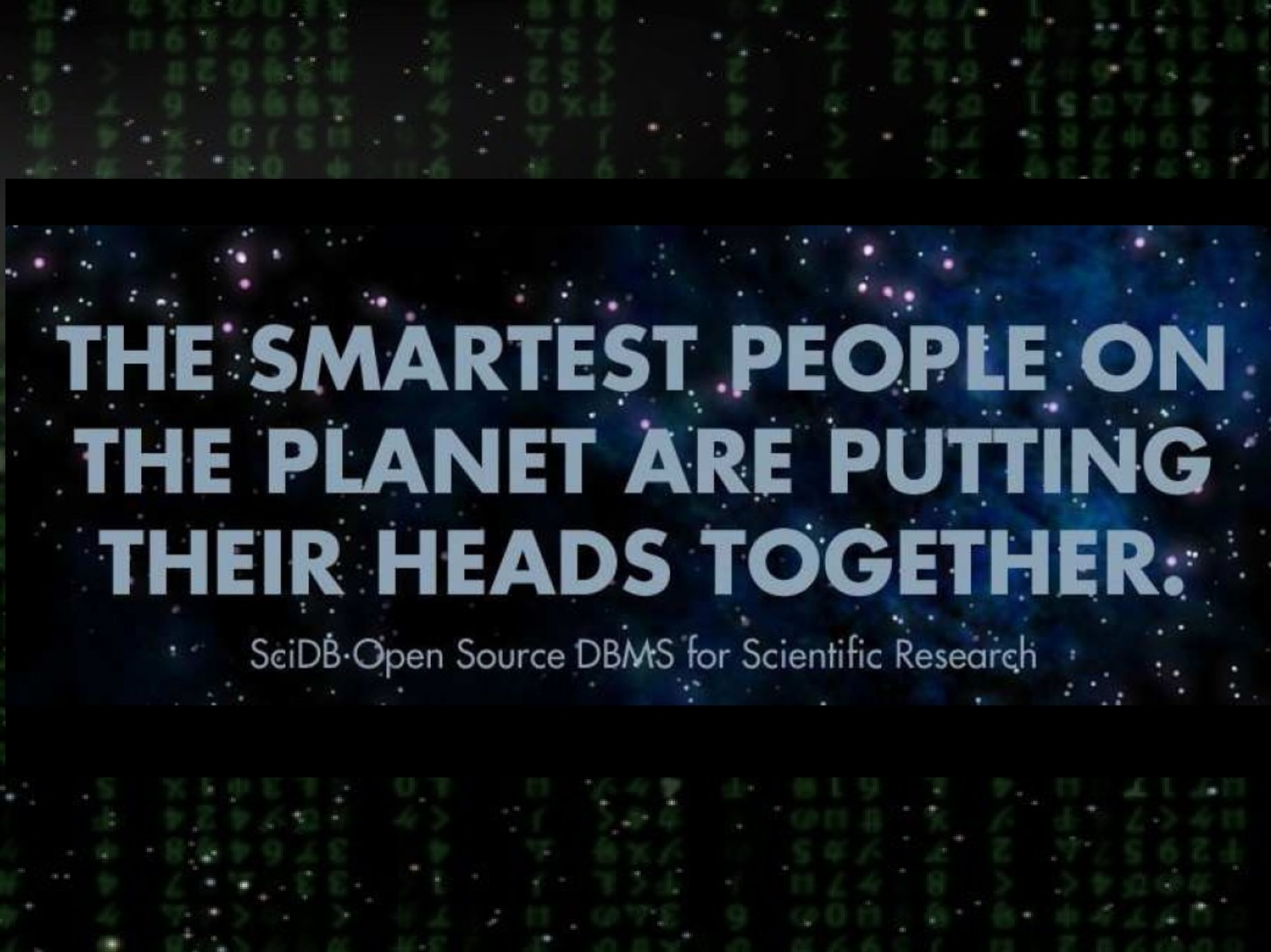
Scidb Требования

- Открытая модель развития, open-source
- Ориентирована на науку
- оптимизация для аналитики, не OLTP, ACID → BASE
- многомерные вложенные массивы — основная структура данных, декларативный язык. Разреженные массивы с неровными краями
- Масштабируемость на сотни петабайт, от ноутбука до кластера, минимизация административных затрат

Scidb Требования

Интерфейсы к научным приложениям (R, Matlab, IDL), к языкам (C++, Python)

- No overwrite storage, поддержка версионности данных (named versions), дельта-компрессия
- Происхождения данных, аннотирования данных
- Поддержка данных с ошибками


THE SMARTEST PEOPLE ON THE PLANET ARE PUTTING THEIR HEADS TOGETHER.
SciDB·Open Source DBMS for Scientific Research