

Сто часов астрономии – 2014

ГАИШ МГУ



Научные вызовы технологиям СУБД

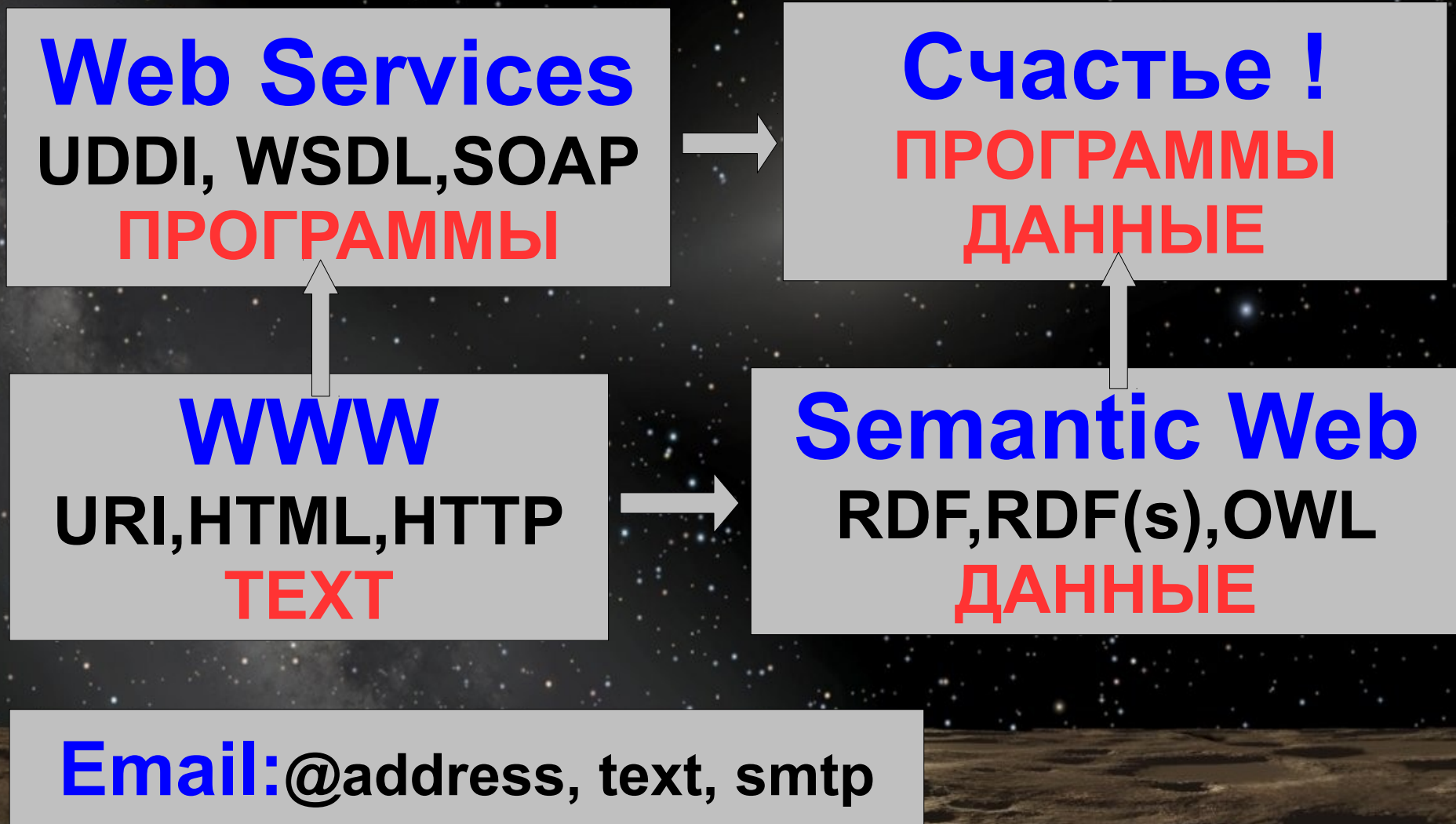
Олег Бартунов, ГАИШ-МГУ

**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-Knowledge management**
 - Цифровое поколение
 - NSF- CDI, изучение базовых элементов инфраструктуры киберобщества.

ВЕБ — Универсальная платформа обмена информацией



**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
- Коллаборация людей и ресурсов, необходимых для решения новых задач науки и промышленности
- 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

Новый «шаблон» работы в науке

- Коллективность
- Специализация
- Расширенный кругозор

Выписка из правил оценки научных достижений ИТЭФ:

18.5. Категории статей:

- 1 категория — до 100 авторов;
- 2 категория — от 100 до 300 авторов;
- 3 категория — от 300 до 500 авторов;
- 4 категория — от 500 авторов.

Новый «шаблон» работы в науке

- Коллективность
- Узкая специализация
- Замкнутый кругозор

Например, программисты:

- Системные администраторы (стандартное железо и поддержка систем);
- Программирование устройств;
- Программирование интерфейсов;
- Программирование БД;
- Численные методы;
- Научное моделирование;
- Обработка данных; ... и так далее

Новый «шаблон» работы в науке

- Коллективность
- Узкая специализация
- Специализированный кругозор

Другой пример — ученые:

- обработка сырых данных (человек - прибор)
- создатели приборов (установок ...)
- экспериментаторы
- специалисты по статистике
- специалисты по статистике
- по написанию статей ... и так далее

Новый «шаблон» работы в науке

- Коллективность
- Узкая специализация
- Ограниченный кругозор

Следствие предыдущего:

60 лет назад — читали все вышедшие статьи;
40 лет назад — все статьи по теме + абстр. РЖ
20 лет назад — все абстр. по теме + нек. статьи;

А что и как читают сегодня?

Административная и финансовая политика «Early Science»



BIG DATA — BIG SCIENCE!

- Эксафлопсы (10^{18}) будущего
 - Воеводин, абрау-дюрсо, 2011
- Зетабайты (10^{21}) настоящего
 - IDC: 1.2 Zb — полное кол-во цифр. данных
- Гартнер, 2011 — 3 Vs
 - Volume — много данных
 - Variety — разнообразие форматов
 - Velocity — скорость in/out

BIG DATA!

- Гартнер, 2012
"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."
- Veracity — достоверность данных

Big Data

- Проблема [не]управления данными
 - Персональные архивы данных (xTb)
 - Институтские хранилища (xxxTb)
 - «Большая наука» (xPb)
 - ATLAS (LHC): петабайт/с !
 - LSST: 20 Tb/ночь, 200 Pb архив (Хаббл, 15 лет, 25 Tb)
 - SKA (радиотелескоп, 1 кв. км) — эксабайт сырых данных в сутки !
 - Гугл: обрабатывается 24 Pb/сутки, социальные сети : Перешагнули Pb-порог

Big Data

- Информационный шум в обществе
 - Nature(2008), CACM(2008), Economist(2010), Science(2011)



Big Data

- Проблема [не]управления данными «подручными» инструментами
 - Персональные архивы данных (xTb)
 - Институтские хранилища (xxxTb)
 - «Большая наука» (xPb)
 - ATLAS (LHC): петабайт/с !
 - LSST: 20 Tb/ночь, 200 Pb архив (Хаббл, 15 лет, 25 Tb)
 - SKA (радиотелескоп, 1 кв. км) — эксабайт сырых данных в сутки !
- Требуется *специальные* решения

COMMUNICATIONS

CACM.ACM.ORG OF THE

ACM

12/08 VOL.51 NO.12

Surviving the Data Deluge

Open Information
Extraction
from the Web

CTOs on
Virtualization

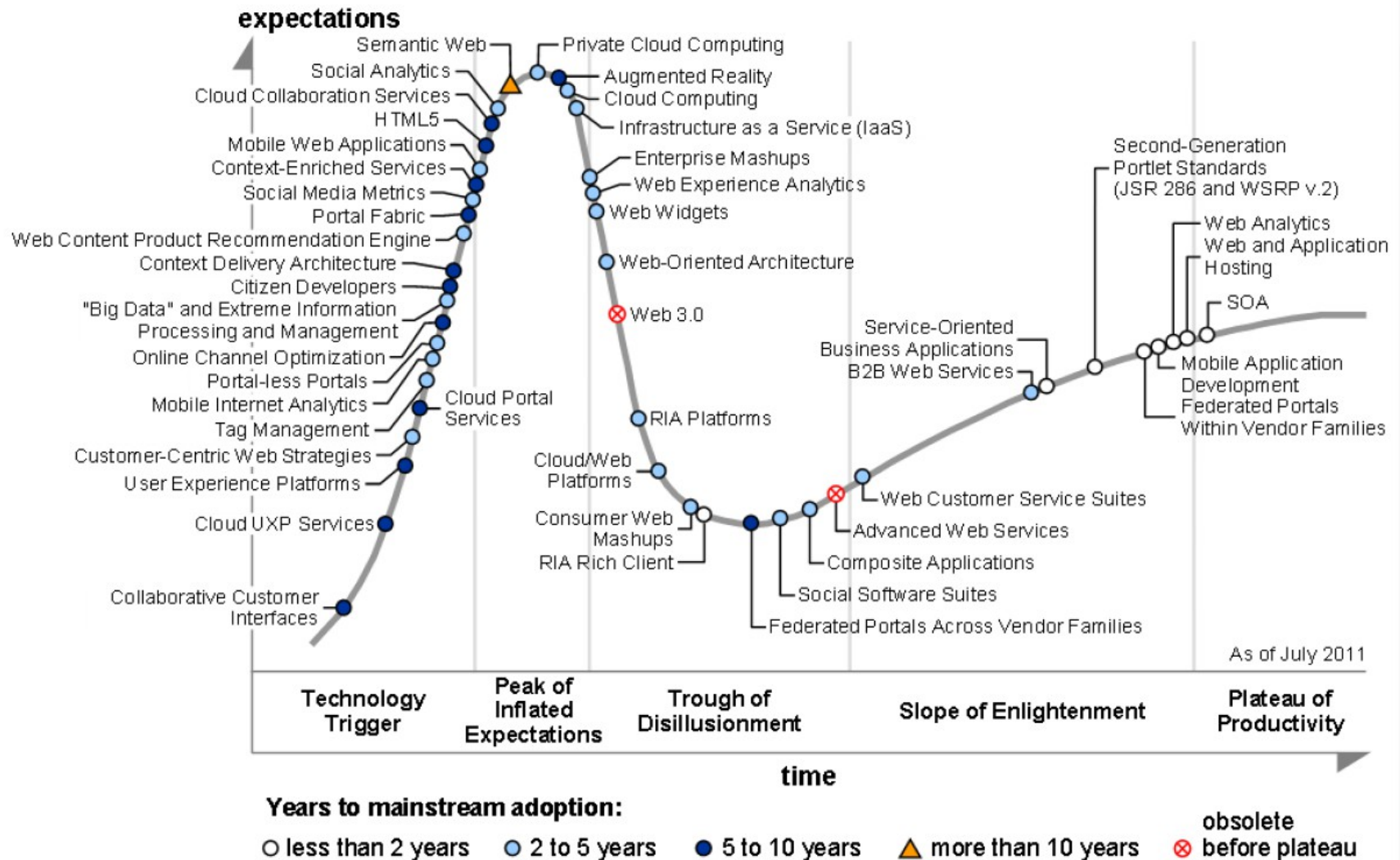
Living Machines

High-Performance
Web Sites

- **IDC 2008 report**
- **2007 - crossover point:**
 - Total amount of digital data est. 264 exabytes (264×10^{18})
 - 1000,000 x US Library of Congress
- **2011 year**
 - 1.8 zettabytes (1.8×10^{21})
 - 2x of available storage (800+ EB)

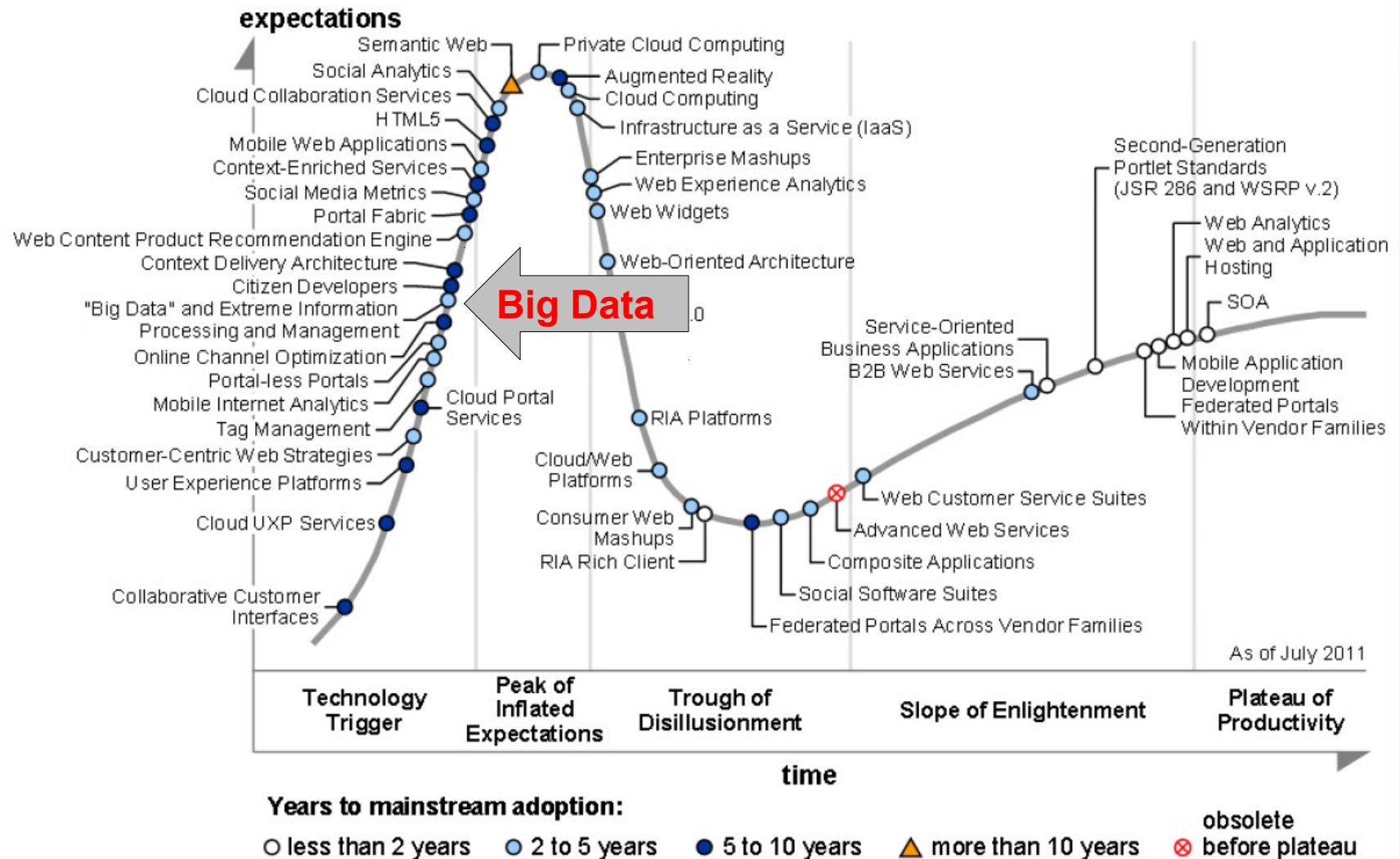
Big Data

Figure 1. Hype Cycle for Web and User Interaction Technologies, 2011



Big Data

Figure 1. Hype Cycle for Web and User Interaction Technologies, 2011



Big Data

- Проблемы хранения Big Data нет
 - Основной производитель данных — сенсоры, сенсорные сети
 - Емкость дисков растет ~ как и сенсоры
 - 120 Pb хранилище уже скоро (IBM, 200,000 SAS-дисков) для предсказания ураганов)
- Трудности с обработкой и анализом
 - Разные форматы данных — табличные, слабо-структурированные, иерархические, видео, аудио, изображения....
 - Требование «быстрых» результатов, в бизнесе счет идет уже на миллисекунды 1

Big Data

- Обработка петабайтных данных в НРС невыгодна — дорогой трафик.
 - Новая архитектура - СУБД на кластере машин как распределенное, масштабируемое хранилище, плюс обработка и аналитика (SciDB)
- Межмашинное взаимодействие — масштабируемый путь к управлению данными.
 - За 10 лет население выросло в 1.2 раза, кол-во транзисторов (емкость дисков) — в 200 раз

Uncertain Data

- Интерактивная работа (человек) с данными не масштабируется
- Помимо роста данных, повышается требование к их качеству и уровню анализа
 - Все данные экспериментов, численного моделирования — неточные.
 - Все результаты и выводы несут отпечаток этой неточности

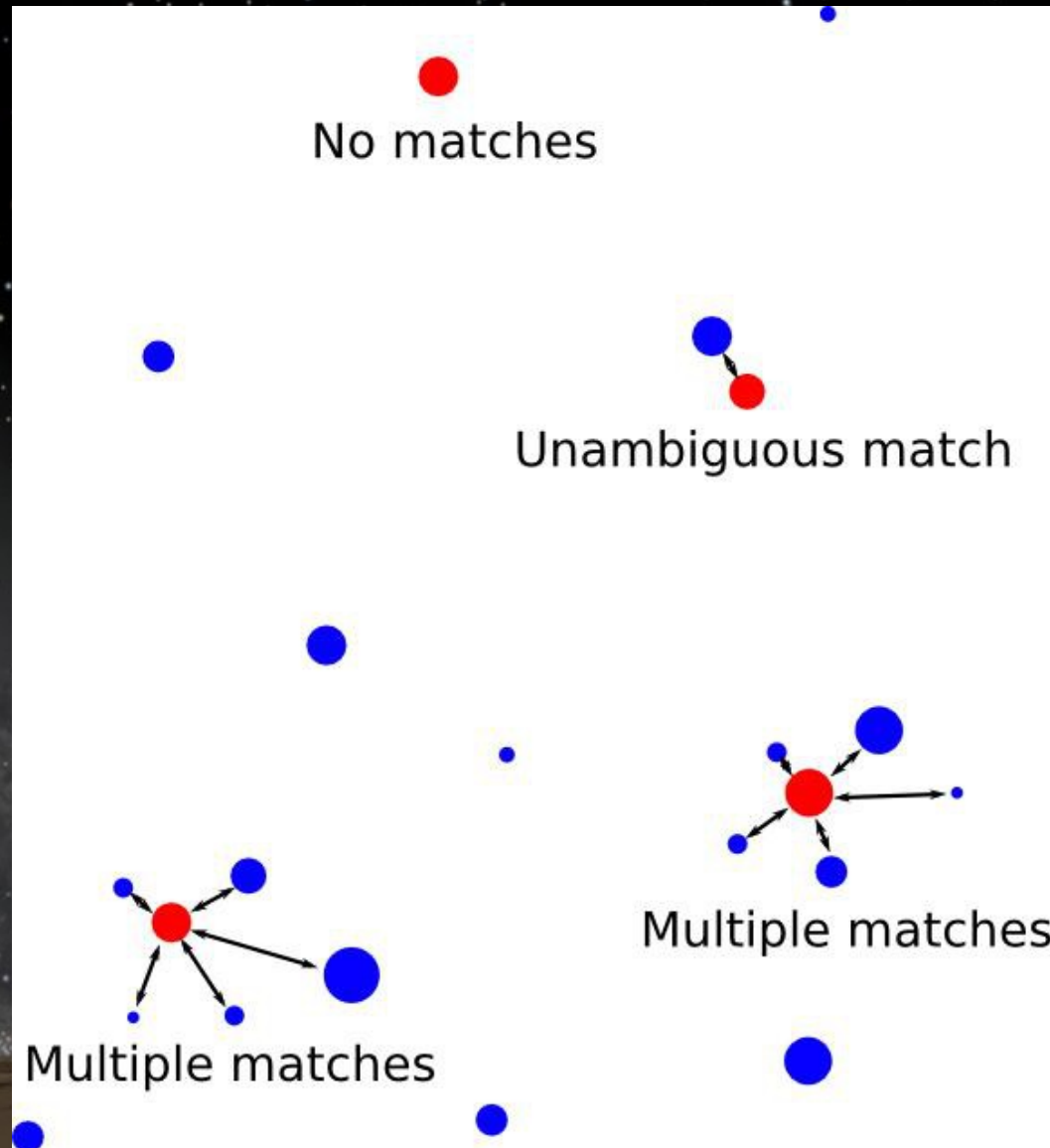
Uncertain Data

- Неточные данные
 - Погрешность измерений
 - Пропущенные события
- Исторические даты:
 - неточно - в 13 веке
 - Интервал — во времена Реформации
 - Неравенства — до нашей эры
 - Массивы - в понедельник в январе
- Астрономия — пропущенные точки, неуверенное отождествление

Uncertain Data

- Типичная задача астрономии — взаимное отождествление объектов из разных каталогов — вероятностная !
 - Координаты могут существенно отличаться.
 - Разные методики наблюдений, разные привязки систем координат, точности наблюдений...
 - Положение объектов на небе может меняться (иногда существенно)
 - Одной звезде могут соответствовать несколько звезд или ни одной. Требуется привлекать внешние соображения.

Uncertain Data



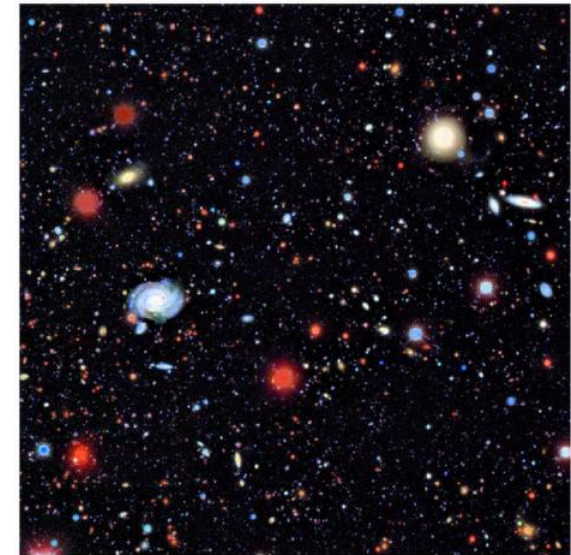
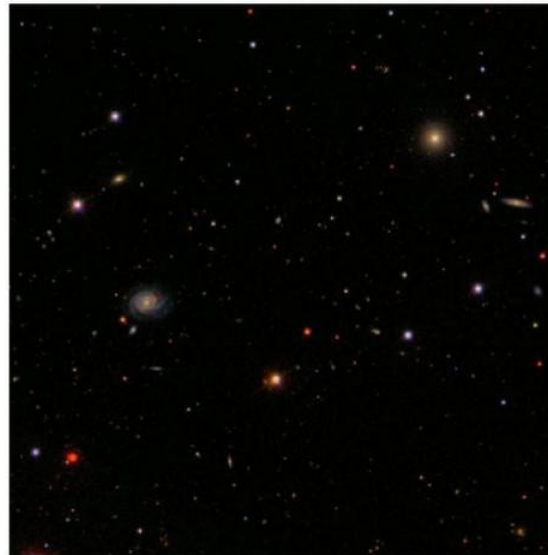
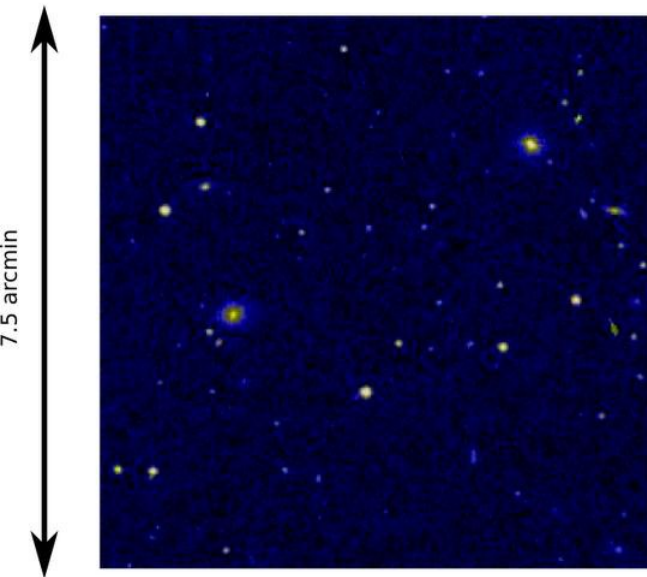
Uncertain Data

- Отождествление звезд становится важнейшей задачей с вводом новых инструментов

DSS. 1980s-1990s

SDSS. 2000s

LSST. 2010s



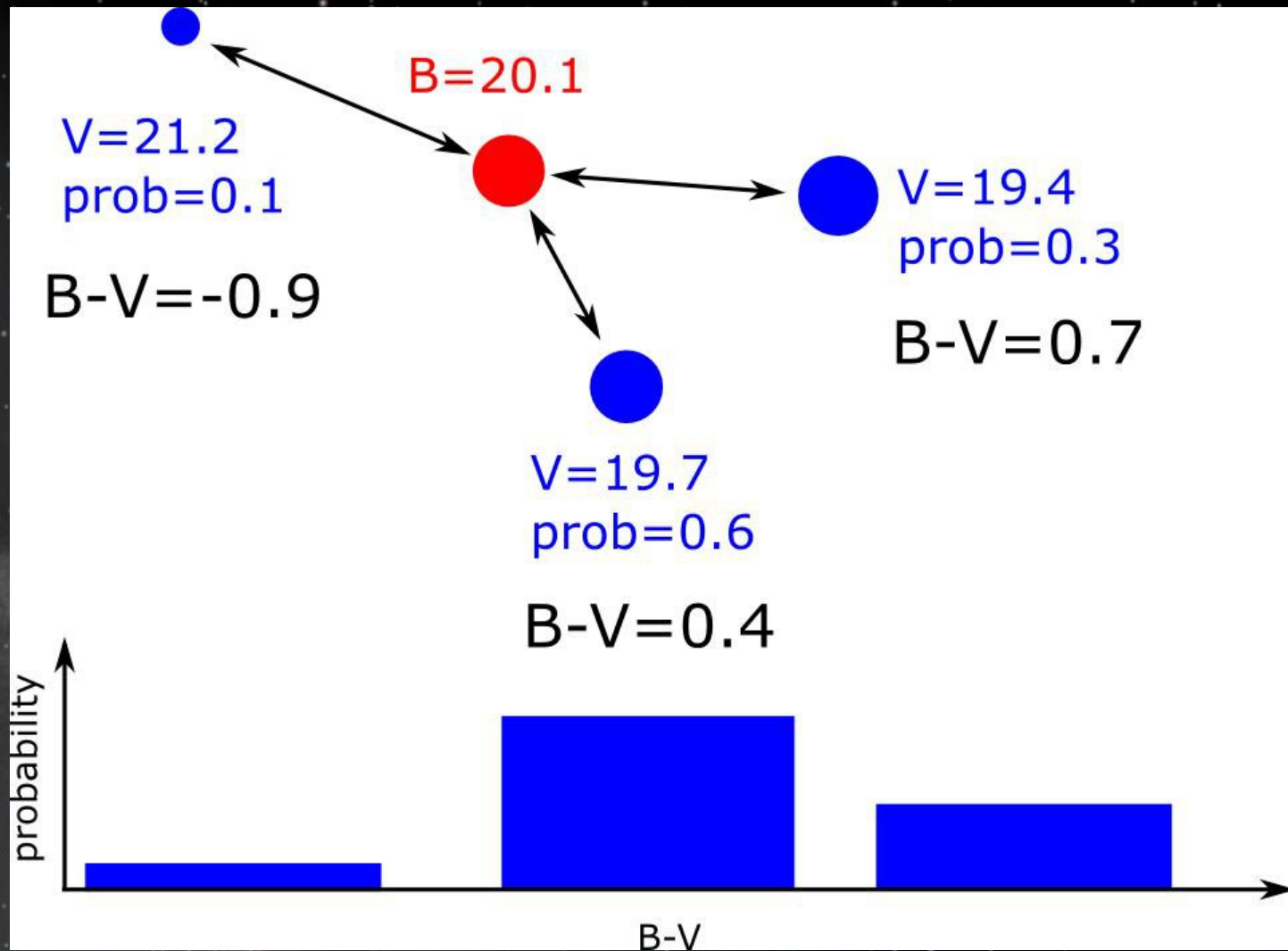
- LSST — 100 млн. звезд/минуту для выделения новых или движ. объектов

Uncertain Data

Звезда	Первый каталог, В	Второй каталог, V	Вероятность
1	20.0	21.2 22.1	0.1 0.9
2	19.3 20.1	21.0	0.4 0.6
3	20.5	22.0	1

Мир	Звезда	В величина	V величина	(B-V)	Вероятность
1	1	20.0	21.2	-1.2	0.04
	2	19.3	21.0	-1.7	
	3	20.5	22.0	-1.5	
2	1	20.0	21.2	-1.2	0.06
	2	20.1	21.0	-0.9	
	3	20.5	22.0	-1.5	
3	1	20.0	22.1	-2.1	0.36
	2	19.3	21.0	-1.7	
	3	20.5	22.0	-1.5	
4	1	20.0	22.1	-2.1	0.54
	2	20.1	21.0	-0.9	
	3	20.5	22.0	-1.5	

Uncertain Data



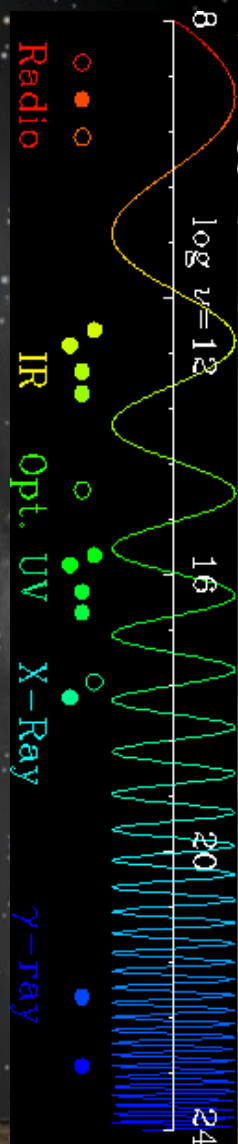
Неточности в координатах перешли в неточности в показатели цвета !



**Очень много
информации!**

(на примере астрономии)

Астрономия стала всеволновой



Дипольные антенны

Параболические антенны

Болометры

Телескопы-рефлекторы

Зеркала косого падения

Кодирующие маски

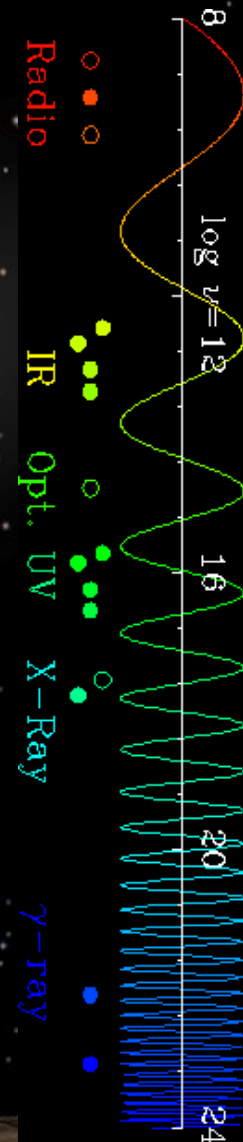
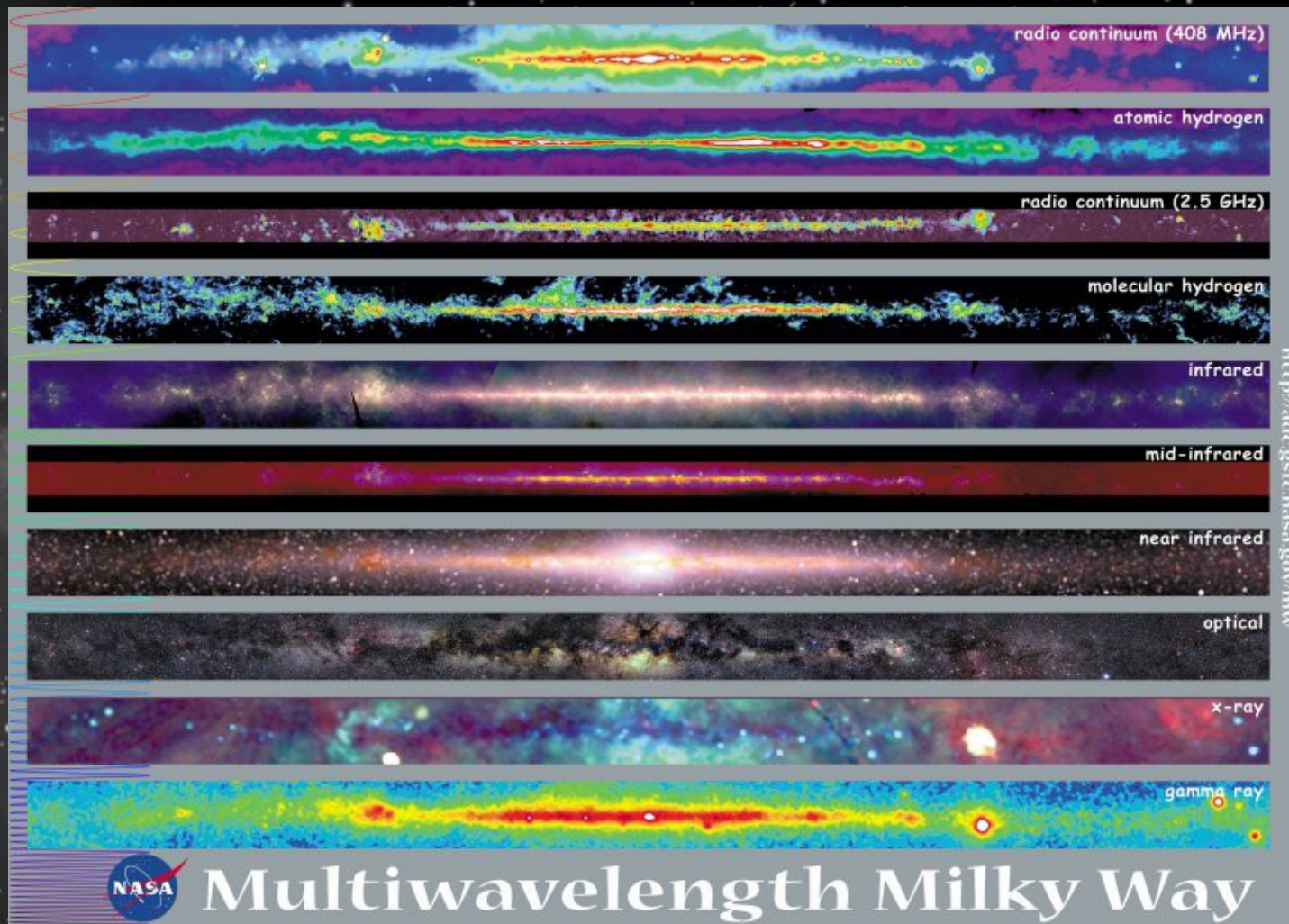
Атмосферные черенковские телескопы,
Широкие атмосферные ливни

Нейтринные телескопы:
(Солнце, SN 1987A)

Гравитационные антенны
(начинают работать)

Космические лучи

Астрономия стала всеволновой



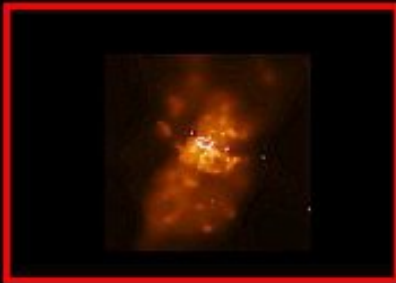
Астрономия стала всеволновой

M82 – Peculiar Starburst Galaxy

Distance: 12,000,000 light-years (3.7 Mpc)

Image Size = 10 x 7 arcmin

Visual Magnitude = 8.4



X-Ray: Chandra



Ultraviolet: ASTRO-1 UIT



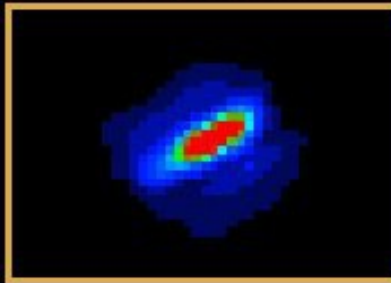
Visible: DSS



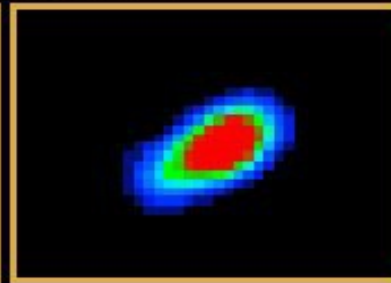
Visible: Color - R. Gendler



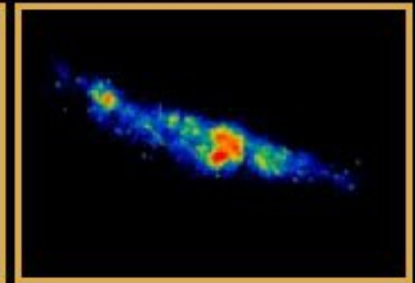
Near-Infrared: 2MASS



Mid-Infrared: IRAS



Far-Infrared: IRAS



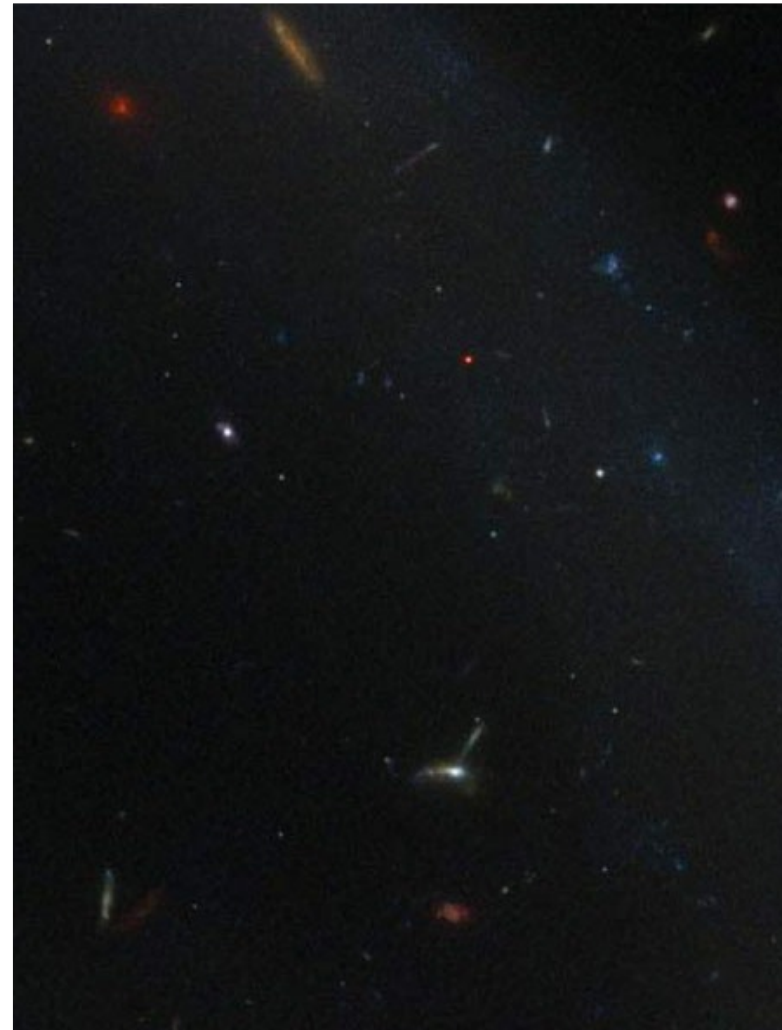
Radio: VLA+Merlin

Астрономия обзавелась историей

Ряды наблюдений:

- двойной пульсар PSR1913+16 (“Нобелевский лауреат”) — 33 года
- некоторые переменные звезды — до 150 лет (фототека ГАИШ — первые с 1872, регулярные — с 1895)
- остальные — неоднородные ряды в 10-15-20 лет в разных диапазонах

Астрономия стала широкопольной (много данных с одного снимка)



Стало очень много данных !

VLDB -> XLDB

Very Large → Extremely Large
XXX Tb XXX Pb

Ежегодные встречи XLDB-сообщества:

Большая наука - астрономия, биология, науки о земле, физика высоких энергий + разработчики СУБД (MIT, Yahoo, Microsoft, IBM, BEA) + eBay

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 Tflor обработка
 - 10 лет: 60 Pb raw data, Каталог — 30 Pb
- LHC — Large Hadron Collider
 - 15 Pb ежегодно, 100K CPU
 - 200 центров в ~ 40 странах

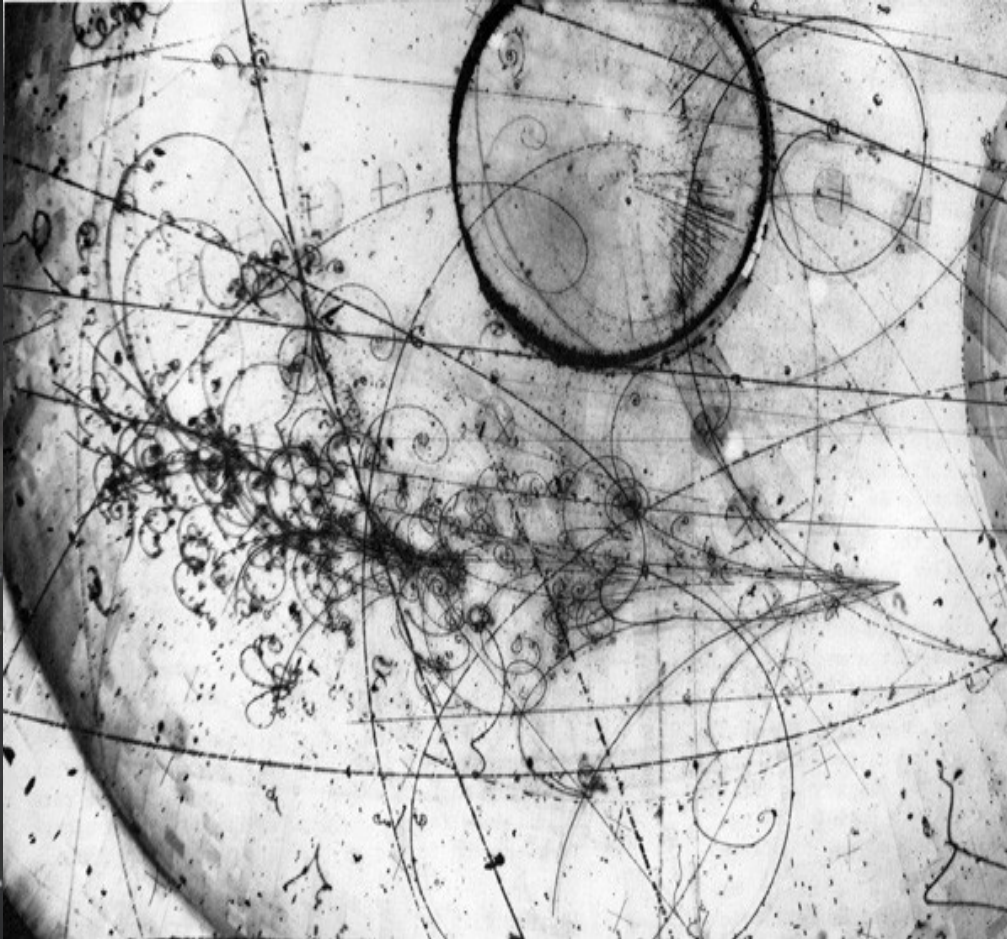
Что такое научные данные ?

- «Сырые» (raw) данные — данные непосредственно из сенсора **Хранить вечно !**
 - Бытовые камеры - raw файлы 12/14 bit
- Мы все дальше абстрагируемся от объекта изучения

Что такое научные данные ?

- Обработка (cooking) «сырых» данных — сложная процедура.
 - 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 ?



Взаимодействие
нейтрино с тяжелой
неон-водородной жидкой
смесью в пузырьковой
Камере Fermilab, 1976

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 !!!

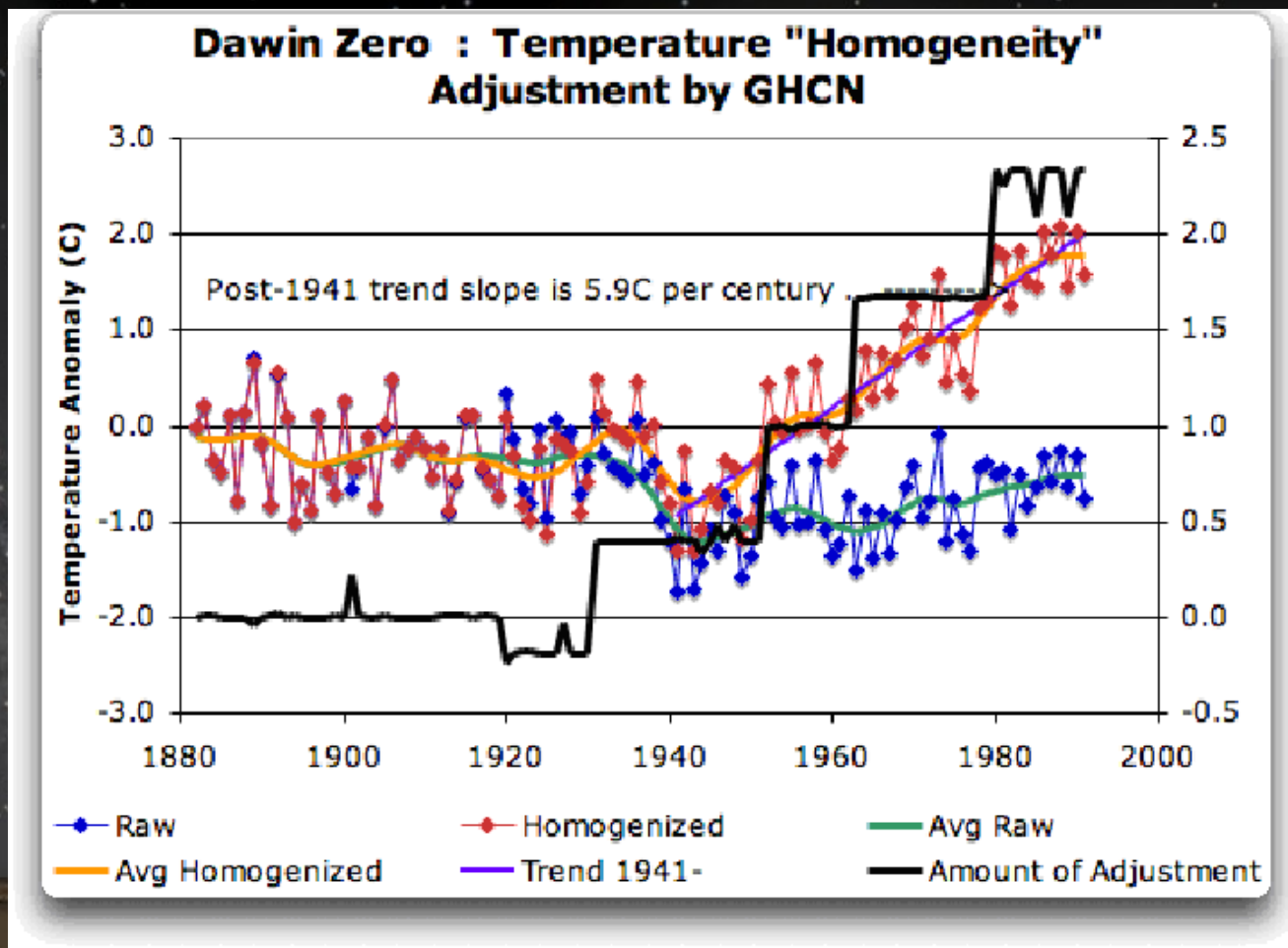
Принцип науки

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 ?

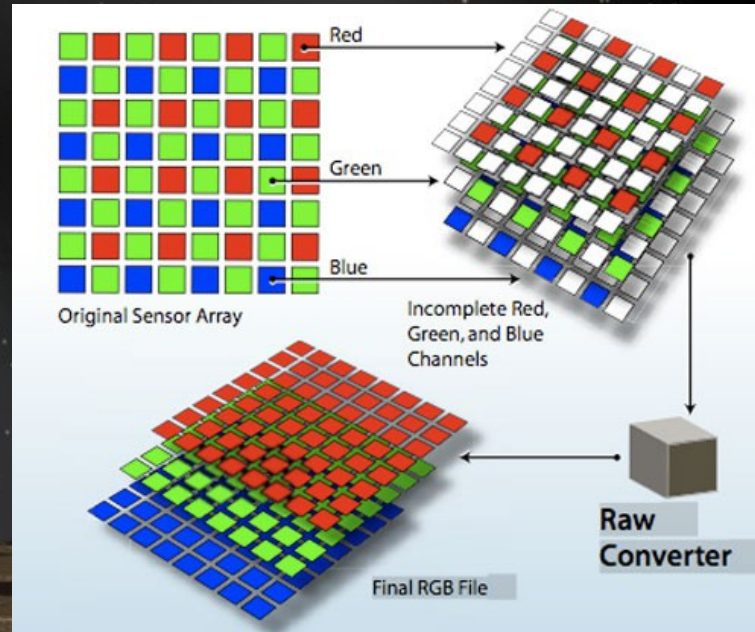
- Маленькая мелочь может привести к эффекту бабочки
- Климатгейт — подрыв доверия
- Модель CPU, hostname - скорее всего, не важны
- Информации о provenance должно хватить для воспроизведения анализа независимым исследователем и подтверждения результатов и выводов

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

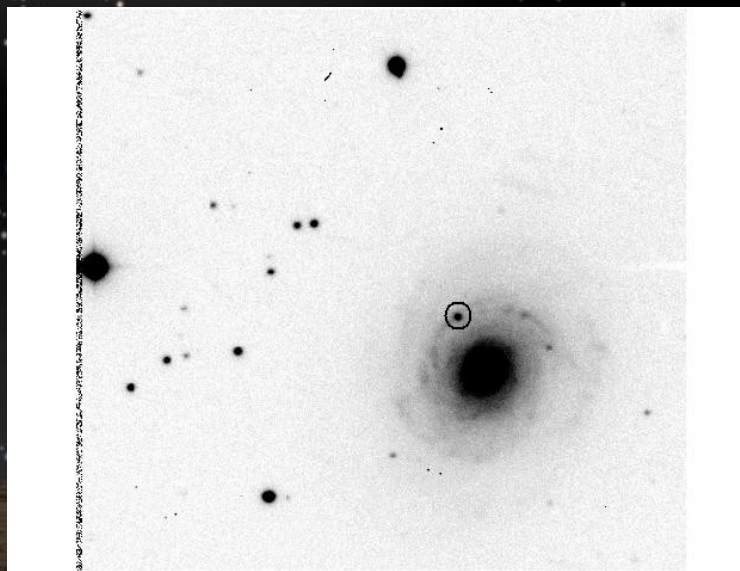
Специфика научных данных

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
(Дмитрий Цветков, ГАИШ) - 2-d массив пикселей



Стало очень много данных (Научная политика)

Данные всех астрономических экспериментов NASA, ESA и (частично) ESO и станут полностью открытыми через 1 год (максимум через 2) после их получения

Аналогичная политика для данных больших наземных оптических телескопов России проводится КТБТ (бывший КТШТ).

Данные становятся распределенными

Причины:

- Нельзя всё хранить в одном месте (слишком большой объем)
- Резервирование
- Скорость доступа
- «Физические» причины — передача данных с орбиты

Что делать ?

- Старый подход «скачать все к себе на компьютер и обработать» **не работает !**
 - трудно собрать (много разных архивов)
 - трудно хранить (много места, сохранность данных)
 - трудно обработать (разные форматы , производительность)
 - трудно актуализировать (данные меняются)
 - Очень дорого (трафик, стоимость носителей)

Ограничения со стороны компьютеров

Законы «Мура»

Ограничения видны!

- Число транзисторов в микросхемах:
удваивается за 2 года (~30 раз за 10 лет)
- Объемы памяти (RAM):
удваиваются за 2 года (~30 раз за 10 лет)
- Объемы дисков (HDD):
удваиваются за 1.5 года (~100 за 10 лет)
- ~~Скорость передачи данных:~~
~~удваиваются за 2 года (~30 раз за 10 лет)~~

Нет ограничений

Ограничения далеко

Архитектура программ

- Монолитные программы,
Многопользовательские программы
- Клиент-Сервер
- Распределенные системы
 - RPC (Unix) — удаленный вызов процедур
 - RMI (Java, SUN) — удаленный вызов методов
 - COM/DCOM (Windows, Microsoft)
 - CORBA — (OMG)
- Проблемы со стандартизацией интерфейсов !

Общение и обмен данными — 70-е годы прошлого века

Email:@address, text, smtp

Общение и обмен данными — 90-е годы прошлого века

WWW

URI, HTML, HTTP

TEXT

- **URI** - Universal Resource Identifier
- **HTML** - Hypertext Markup Language
- **HTTP** — Hypertext Transfer Protocol

Email:@address, text, smtp

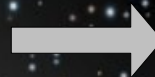
Общение и обмен данными — начало 21 века

- **RDF** — Resource Description Framework
- **RDF(s)** — RDF Schema
- **OWL** — Web Ontology Language

WWW

URI, HTML, HTTP

TEXT



Semantic Web

RDF, RDF(s), OWL

ДАННЫЕ

Email:@address, text, smtp

Общение и обработка данных — начало 21 века

Web Services
UDDI, WSDL, SOAP
ПРОГРАММЫ

- **UDDI** - Universal Description, Discovery and Intergration
- **WSDL** — WS Description Language
- **SOAP** — Simple Object Access Protocol

WWW
URI, HTML, HTTP
TEXT

Semantic Web
RDF, RDF(s), OWL
ДАННЫЕ

Email:@address, text, smtp

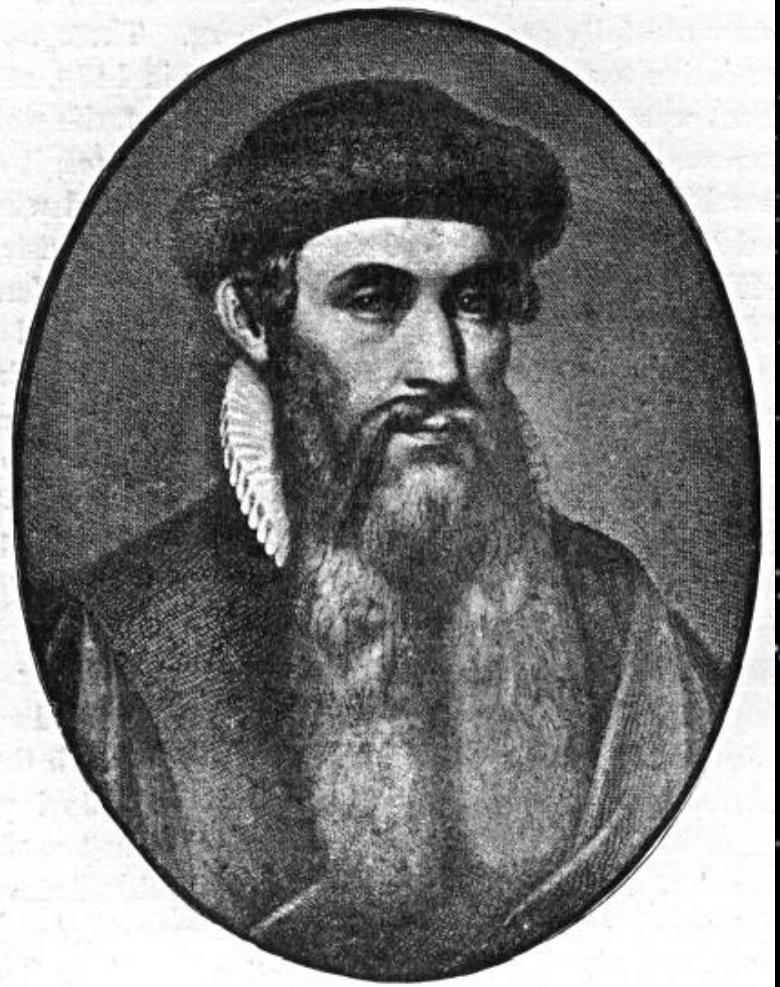
Web Services

- Это программа, приложение
- Запускается через интернет
- Его описание доступно через интернет
- Как ее запускать описывается в WSDL файле, внутреннее устройство скрыто
- Результаты доступны как SOAP сообщение
- Обеспечивает межмашинное взаимодействие через сеть

Web Services

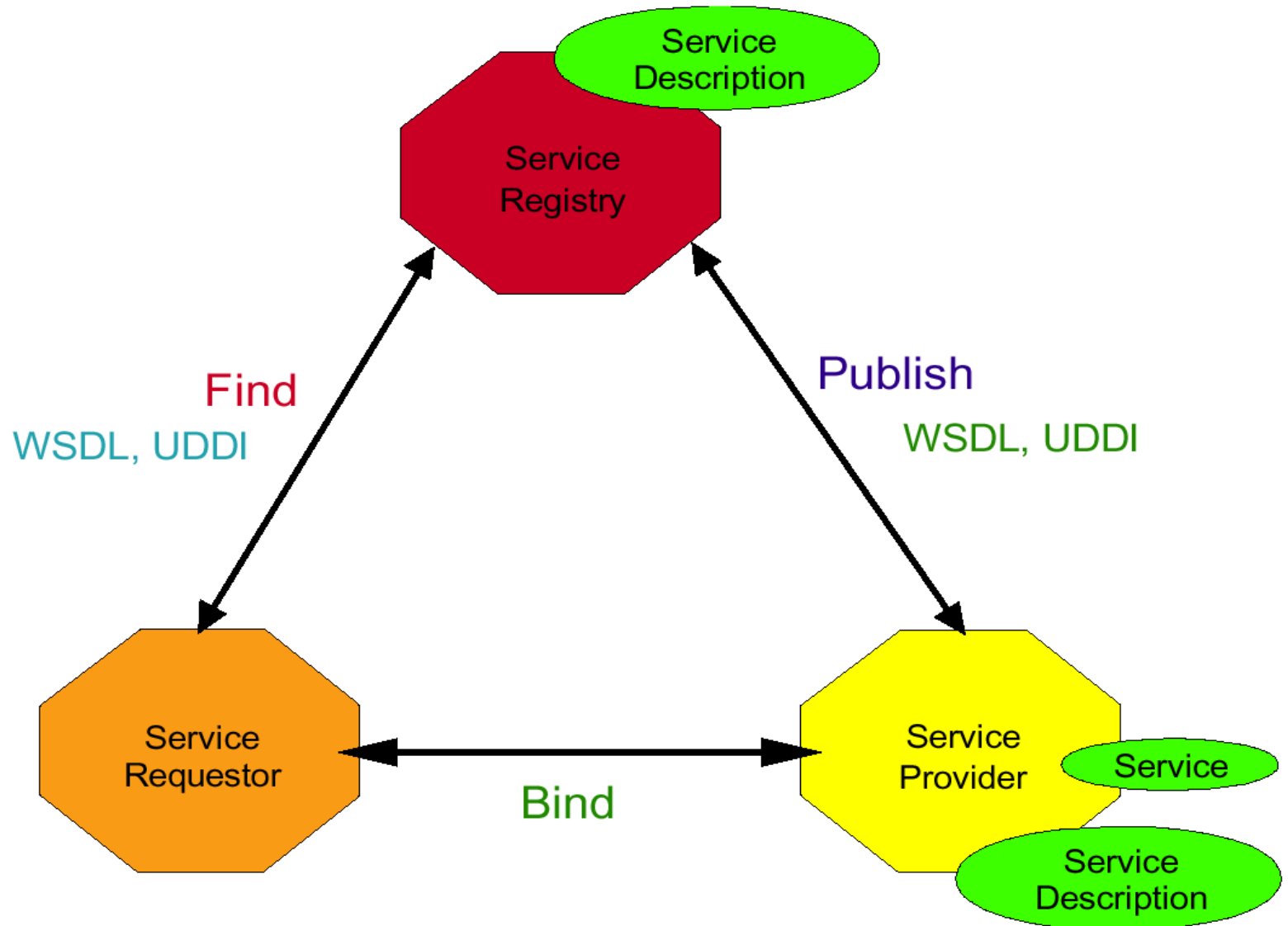
- Веб-сервисы можно рассматривать как подпрограммы. Отличия:
 - веб-сервис— **исполняемый** код
 - веб-сервис — **распределенный** код
 - унифицированный интерфейс
 - унифицированные протоколы
- Веб-сервисы - «кубики лего» для создания сложных веб-сервисов

Movable types — Gutenberg !



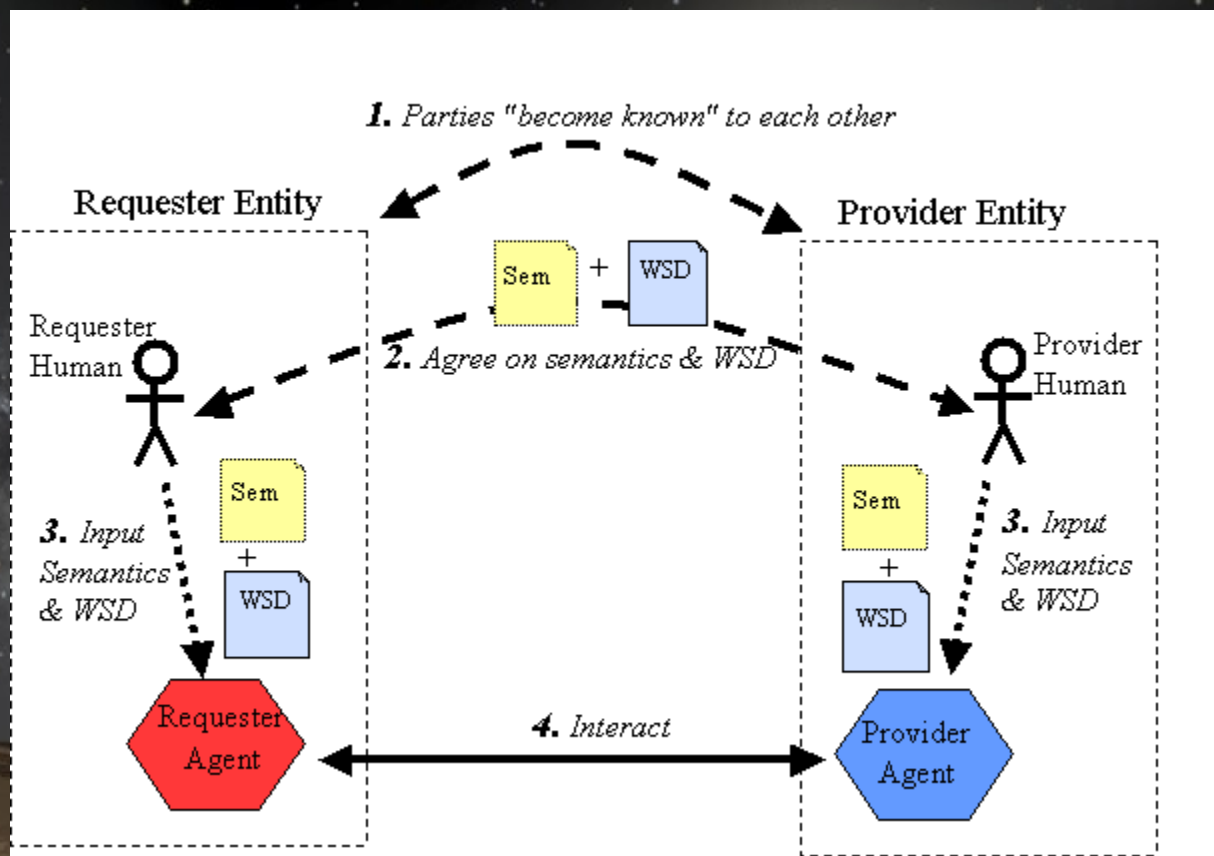
Модульность + стандартизация

Web Services Architecture



Использование WS

Человек нужен для постановки задачи и принятия решения использовать сервис. Счастье наступит когда программа **сама** сможет это сделать. Нужна семантика !



WEB – universal medium for data,
information, and knowledge exchange.

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

Virtual Observatory

Счастье для астронома

- Технология, которая призвана доставить астроному любые данные, независимо от их расположения и методов хранения
- Базируется на SOA + астрономические стандарты и соглашения
- Позволяет автоматизировать рутинные процедуры — от технологии «клик» к взаимодействию программных агентов
- WWT — Word Wide Telescope, телескоп, который не зависит от погоды !

Виртуальная Обсерватория

проблемы и потребности

- Где и какие серверы есть?
- Создание сложных сервисов
- Создание сложных запросов (заданий)
- Учет особенностей и потребностей астрономии
- Описание ресурсов (каталогов, содержания БД)
- Авторские права
- Платные ресурсы
- Политика доступа
- Real-time web-сервисы
- Мультиверсионность
- Происхождение данных

Нужен реестр
Библиотеки сервисов
Язык заданий: BPEL
(Business Process Execution Lang.)

Стандарты

UCD (Unified Content Descriptors)

Ещё не решенные проблемы

▪
▪
▪

Виртуальная Обсерватория: UCD

Зачем нужен UCD?

μ (собственное движение)

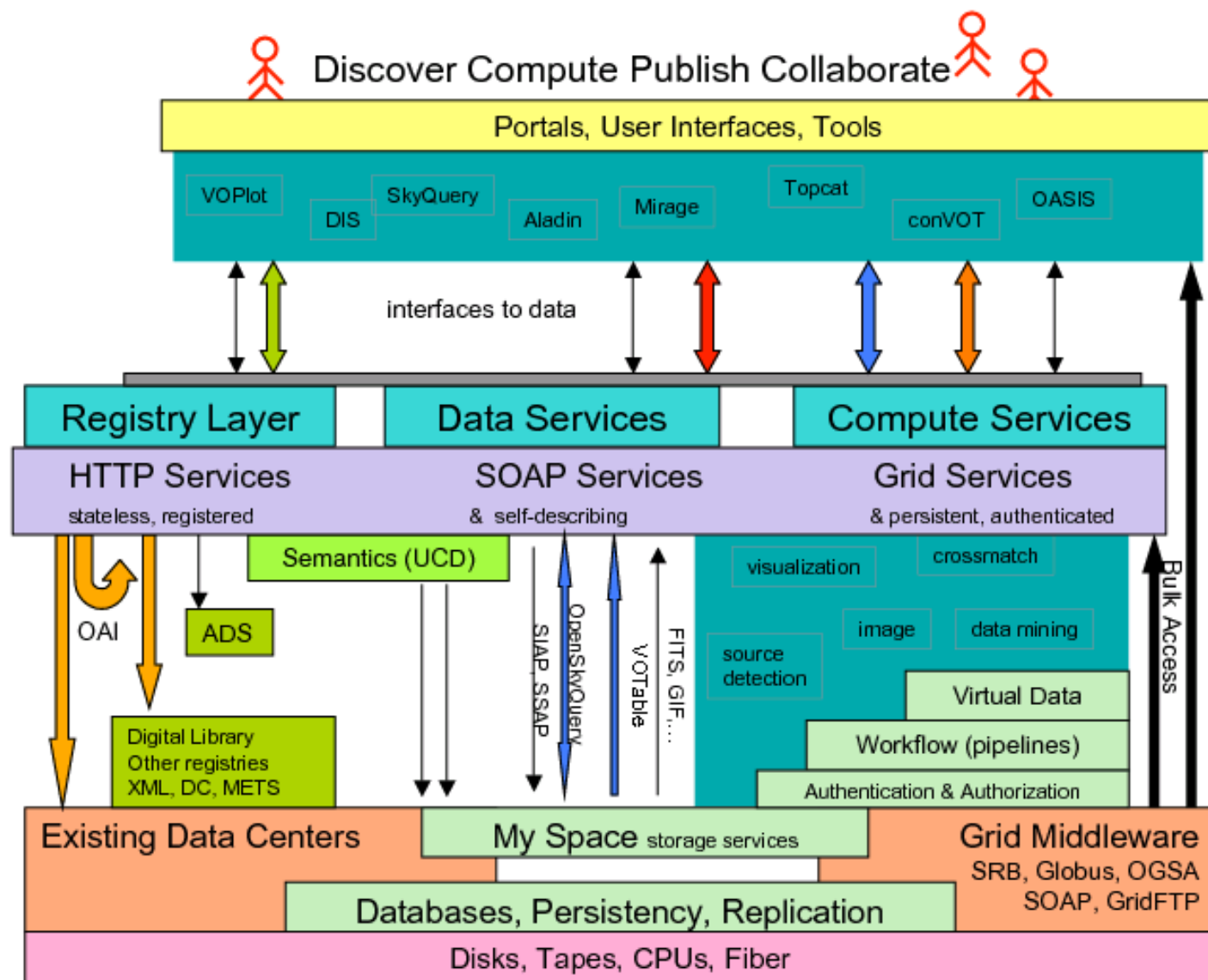
Сколько у него
различных
обозначений?

Более 300 !

В каких единицах
измеряют расстояние?

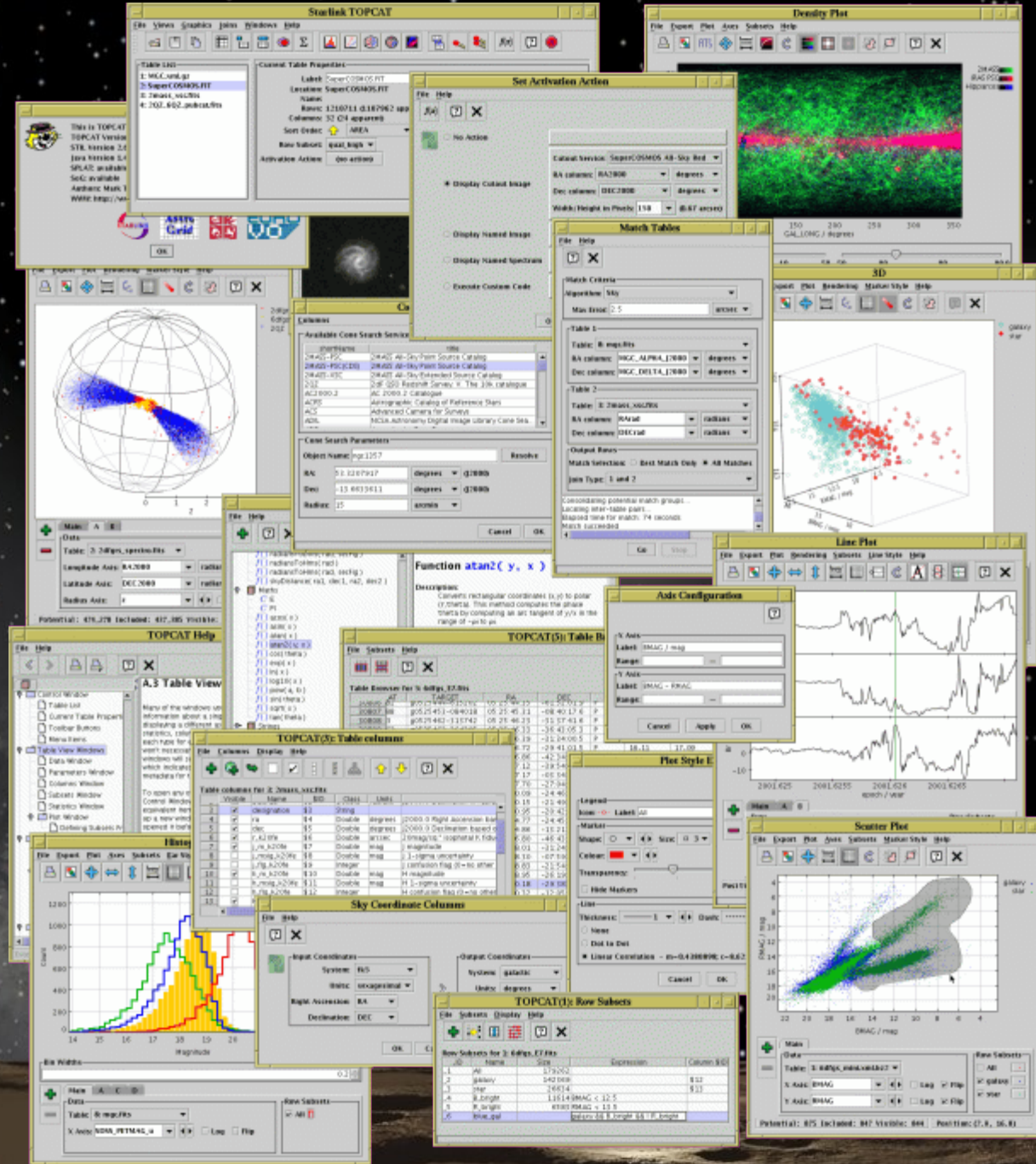
- сантиметры
- метры
- километры
- мили
- километры в секунду
- А.Е.
- световые секунды
- световые годы
- парсеки
- килопарсеки
- мегапарсеки
- . . .

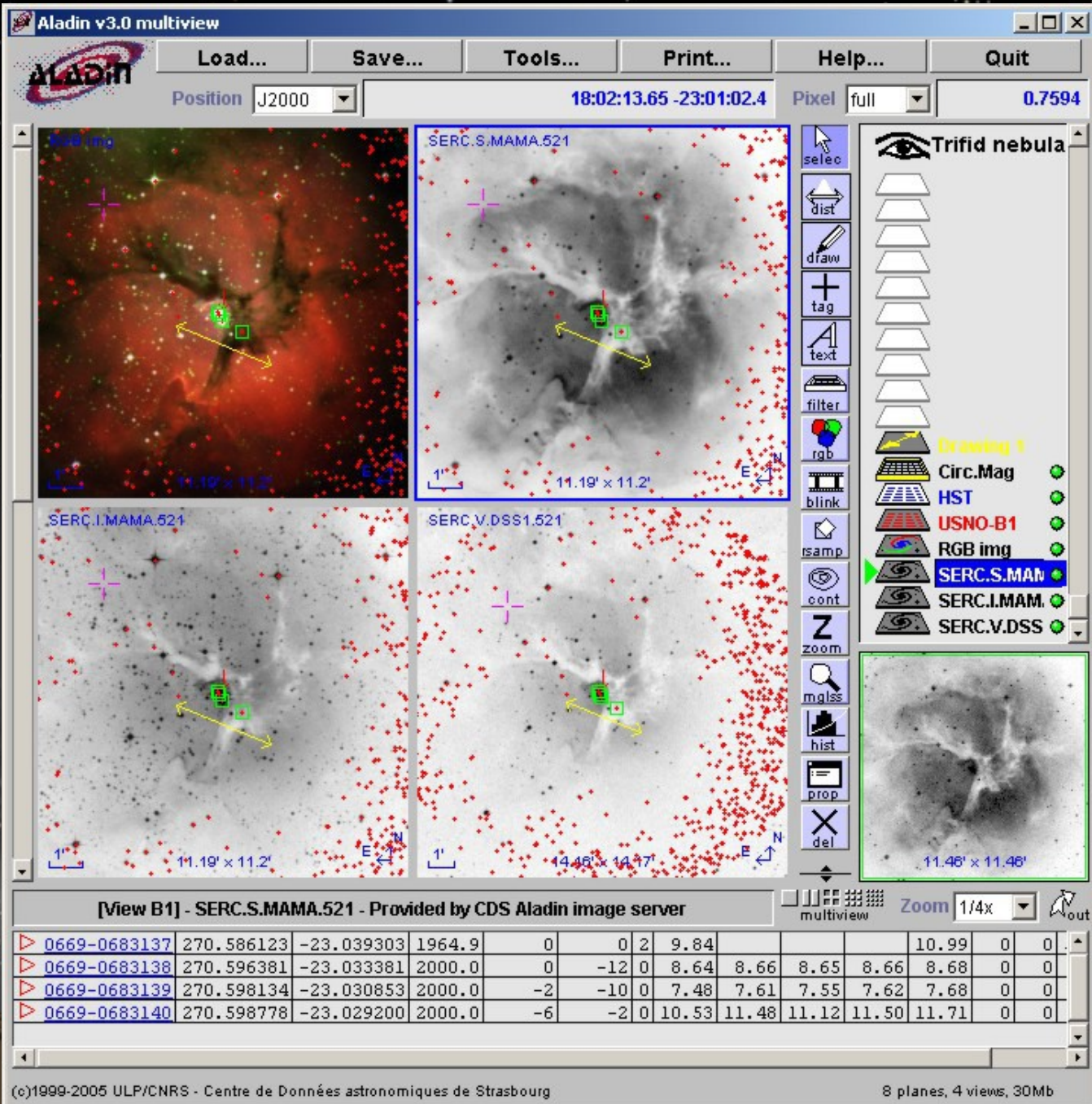
Архитектура Виртуальной Обсерватории



Обучение VO

Introduction to SQL	Resource Metadata
Introduction to XML	Publication and Registry Clients
XML Technologies	Publication and Registry Clients: Exercise
Web Service Technologies	Data and Service Discovery
Grid Computing	Dealing with Invalid Data
Java Introduction	VOEvent
Java Web Libraries	Exercise 1 - Building a VO client
Language Mini-tutorial: C#	Exercise 2 - Building a VO Service
Language Mini-tutorial: IDL	Exercise 3 - Combining VO elements
Language Mini-tutorial: IRAF	What's on tap in the VO
Language Mini-tutorial: Perl	Using the Virtual Observatory
Language Mini-tutorial: PHP	
Language Mini-tutorial: Python	Virtual Observatory Tools
	Using the VO to Study the Relationship of Galaxies and Clusters
Protocols Overview	Using the VO for Cross-correlations
Science Applications	Object Classification Using the VO
VOTables	Using the Grid for Astronomical Data
Data Access Protocols	Data quality in the VO
SkyNodes	Student Projects
ADQL	
Managing VO data and Process Flows	HI Archives-Data Access and Tools
UCDs	Spectrophotometry Client and Web Service
Integrating Existing Tools into the VO: IRAF	





Какие ресурсы доступны в VO ?



DataScope Results for sn 1992bo [01 22 02.00, -34 11 48.0]/ 0.25°

New Query - DataScope Help - [Feedback](#) - Cache date: 2005-12-09 08:40 UTC

National Virtual Observatory

Hosted at NASA/HEASARC

Request Status: Request scanning completed. Data found for 323 of 431 resources.

Major Multiwavelength Services

☐ ADS(100) ?

☐ NED(images)(0/8) ?

☐ SkyView(0/18) ?

Images (Data in one or more FITS files)

Multi

☐ CADC/HST(0/4) ?

☐ HST/SIAP/PREVIEW(1) ?

☐ MAST Scrapbook(0/2) ?

☐ MAST-Scrapbook(0/2) ?

Optical

☐ DSS1(1) ?

☐ H-alpha(1) ?

☐ HST Previews(0/51) ?

Radio

☐ NVSS(1) ?

☐ SUMSS(1) ?

Infrared

☐ 2MASS QL(0/18) ?

☐ IRAS(0/4) ?

☐ ISSA(0/12) ?

☐ SFD IR(0/2) ?

No data found in 108 resources:

2MASS-PSC , 2MASS-XSC , 2QZ , 2cmVLBA , ADIL , Abell CofG , Aladin , CADC/CFHT , CADC/GEMINI , CADC/JCMT , CEDAG , CHIANTI , CLASS , CNOC1 - IM , COSMOS , Chandra , DPOSS , DSS1 , DSS1B , DSS1R , DSS1Y , DSS2 , DSS2B , DSS2IR , DSS2R , Downes CVs , EGRET , EIGA , EUVE , Edge Energies , FIRST , FIRST , FLS_ELAIN1_R , FLS_MAIN_R , FLS_VLA , GB6 , GCVS , GOODS-HST , GOODS-HST , GRB/IPN , GSC1 , GSC2.2 , HD , HIP , HST.HDF , HST.HDF_SOUTH , HST.HELIX , HST.TNO , HST.UDF , HST.maoz_atlas , Hewitt , IGA , IRTS , ISO SIAP AIO , LCRS , LGA , LH , LineEnergy , MCG , MIGA , MITVL , MSX , NBG , NEAT , NED(sources) , NAO , NOT PROVIDED , NOT PROVIDED , NVSS , ROSAT/HRI , SAO , SDSS-DR1 , SDSS-DR3 , SDSS-DR4 , SDSS-EDR , SDSSDR1-G , SDSSDR1-I , SDSSDR1-JPG , SDSSDR1-R , SDSSDR1-U , SDSSDR2-G , SDSSDR2-I , SDSSDR2-JPG , SDSSDR2-R , SDSSDR2-U , SDSSDR2-Z , SDSSDR3-Cut , SDSSDR3-G , SDSSDR3-I , SDSSDR3-R , SDSSDR3-U , SDSSDR3-Z , SDSSDR4-Colour , SDSSDR4 , SWIRE , Shakhbazya CGG , Simbad , Sky2000 , SubaruSXDS , UGC , UIT , VCS , VLA-FIRST , WDS , WENSS , XMM-Newton , sia-sxds , sxds_sia



SAI VO

[About Us](#)
[Contact Us](#)
[SAI VO Wiki](#)

SAI CAS

[ConeSearch](#)
[CrossMatch](#)
[Program Access](#)
[SAI CAS Wiki](#)

SAI Web Services

[WCSFix](#)

VO Links

[IVOA](#)
[RVO](#)

SAI Virtual Observatory Project

This site is about the Virtual Observatory project at the Sternberg Astronomical Institute, Moscow University ([SAI MSU](#)). We provide a public access to the growing set of astronomical services developed by the [SAI RVO development team](#) in accordance with the Virtual Observatory standards. We also provide an interactive access to our services.

We provide access (programmatic and interactive) to the all major and biggest astronomical catalogs (about 4 billions objects total) as well as other web-services.

The technology behind this project is based on the original algorithms and open-source software.

The Virtual Observatory (VO) is the main streamline of the modern technology development to facilitate the solution of science problems by providing *programmatic access* to the huge amount of astronomical data stored in the distributed and heterogeneous archives. Access of this kind requires software architecture, which was developed and adopted recently on the base of open standards Service Oriented Architecture ([SOA](#)) and astronomical agreements, coordinated by [IVOA](#) (International Virtual Observatory Alliance).



Catalogs

- ☐ denis
- ☐ usnoa2
- ☐ twomass
- ☐ gsc1_2
- ☐ tycho2
- ☐ nomad
- ☐ ucac2
- ☒ sdssdr5
- ☐ xmm2p
- ☐ usnob1
- ☐ gsc2_3_2

Object name or coordinates (degrees or HH MM SS or HH:MM:SS):

Center: Radius: arcmin ▾

N13, or 16 41 41.44 +36 27 36.9, or 250.42267 +36.46025

No tables selected. Click on the table name to select/unselect it.

Output format: ☒ VOTable (XML) ☐ CSV (Comma-Separated Values)

Cone Search:

CrossMatch of selected tables in a given area:

Match radius: arcsec ▾

Must be less than 10 arcsec

How to use ConeSearch

- ◊ Click *Catalog name* to view the catalog **Info**
- ◊ Click *Table name* to **select / unselect** it for the search
- ◊ Click [?] to view the table **Structure**
- ◊ Click [[New Table](#)] to upload **New Table** into your personal catalog
- ◊ Click [[Query](#)] to **Download** your query results
- ◊ Click [[Show in Aladin!](#)] to **View** your query results in Aladin applet
- ◊ Click [[Match](#)] to **Cross-Match** several catalogue tables in a given region

SAI CAS Catalogs

- vo.astronet.ru/cas
- ConeSearch Service (VOTable, CSV):
 - USNO B1 – 1 миллиард
 - 2MASS xsc - 1.6 миллионов
 - 2MASS psc - 250 миллионов
 - Tycho2 - 2.5 миллионов
 - UCAC2 - 50 миллионов
 - NOMAD - 1.2 миллиарда
 - DENIS ~ 100 миллионов
 - others
- SDSS DR5 - 300 миллионов (2Tb),
DR6,DR7

Астрономические данные: каталоги в реляционных СУБД

Каталог = неупорядоченная таблица

Астрономические данные: каталоги в реляционных СУБД

- SDSS DR7

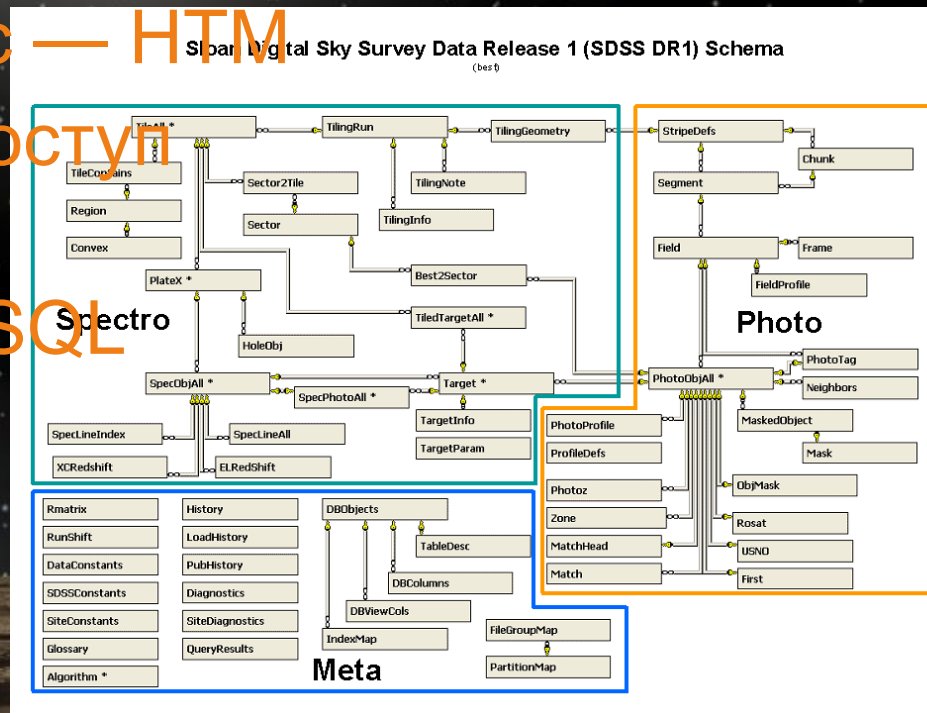
- 357 млн объектов, 400 атрибутов на каждый
- 76 млн спектральных линий для 1.6 млн объектов

- Сферический индекс — HTM

- Пользовательский доступ

- casjobs.sdss.org

- ObjectivityDB → MS SQL



Астрономические данные: каталоги в реляционных СУБД

- SAI CAS (Бартунов, Копосов, Белинский, Карпов)
 - SDSS DR6/7, USNO A2/B1, GSC 2.2, 2MASS...
 - Хранение произвольных каталогов

- таблицы
- атрибуты
- типы данных и UCD

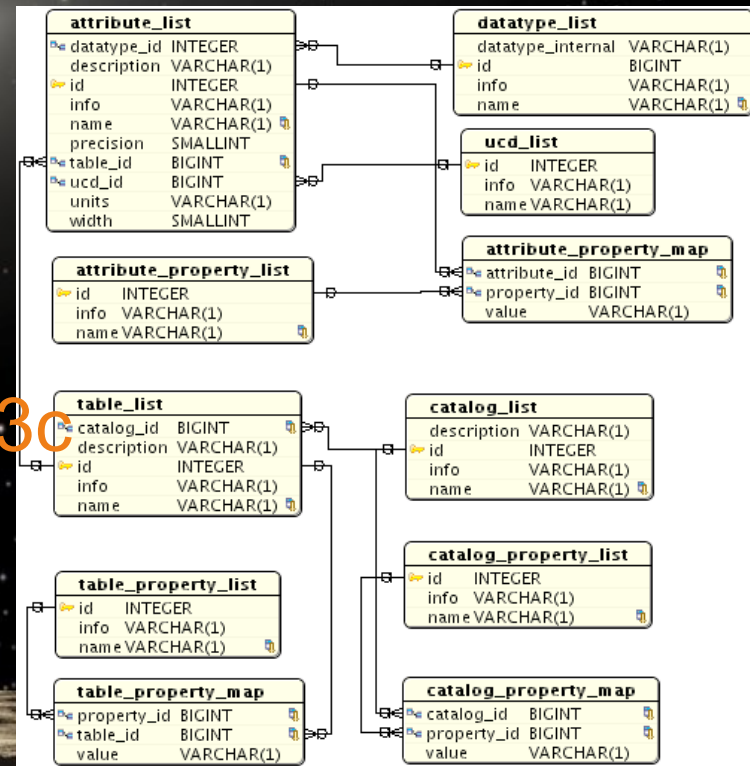
– Единый интерфейс

– Сферический индекс — q_3c

- радиальный поиск
- кросс-отождествление

– PostgreSQL

– ~47 Тб на данный момент

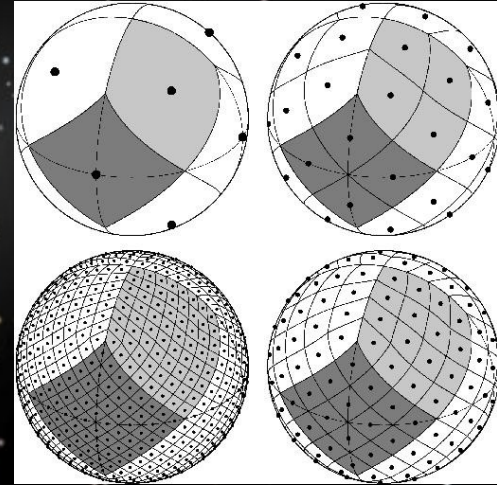


Каталоги в реляционных СУБД: сферические координаты

Однозначная проекция сферы на числовой ряд

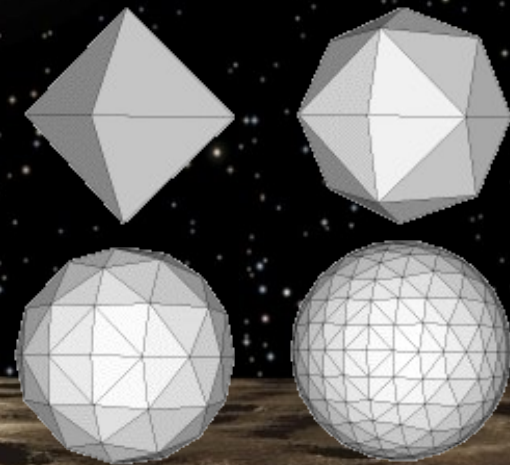
- HEALPIX

- Hierarchical Equal Area iso-Latitude
PIXelization
- Равные площади
- Сложный расчет



- HTM (SDSS)

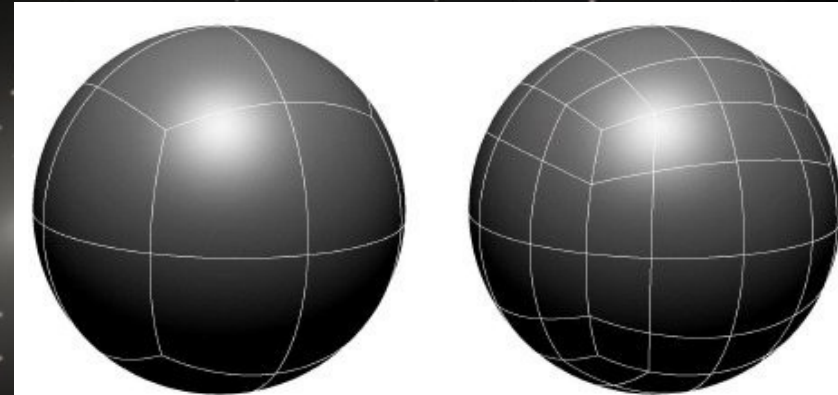
- Hierarchical Triangular Mesh
- Октаэдрон + треугольники
- Рекурсивный алгоритм



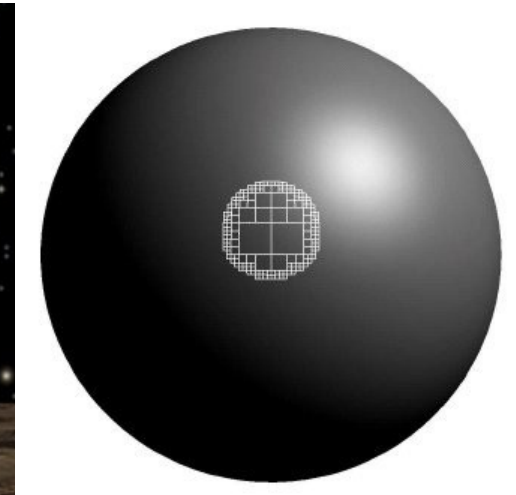
Каталоги в реляционных СУБД: сферические координаты

Однозначная проекция сферы на числовой ряд

- Q3C (Копосов и Бартунов)
 - Quad Tree Cube
 - Куб + квадраты
 - Разная форма и площадь
 - Для индексации это не важно!
 - Очень быстрый алгоритм
 - Модуль для PostgreSQL
 - Функциональный индекс
 - Не требует модификации таблиц
 - Радиальные и сложные запросы
 - Кросс-отождествление таблиц

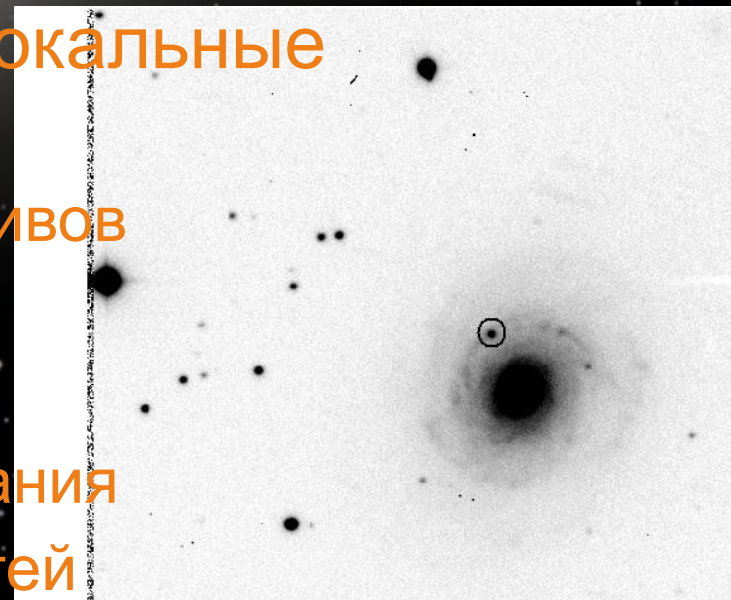


The sphere segmentation in Q3C



Астрономические данные: другие типы данных

- Изображения, спектры, временные ряды, ...
 - Внутренняя упорядоченность данных
 - Оптимальное представление — многомерный массив
 - Анализ предполагает нелокальные операции
 - Сложение/вычитание массивов
 - Поиск соседних точек
 - Фильтрация
 - Интегральные преобразования
 - Выделение связных областей
 - Погрешности! **плохо ложатся на реляционную модель**



Астрономические данные: недостатки реляционной модели

- Нет внутренней упорядоченности
 - Подходит только для каталогов
- Нет концепции погрешностей
 - Интервальные операции, перенос ошибок
- Нет концепции версионности
 - «История» для каждого элемента базы
- Сложность хранения исходных данных
 - Полный цикл жизни данных внутри СУБД
- Сложность масштабирования СУБД
 - Петабайтные базы данных = кластеры

Необходима новая СУБД, построенная на иных принципах

Современные базы данных: база данных для науки

- SciDB (LSST, LHC, ...)



- Координатор — М.Стоунбрейкер

- Ориентация на научные данные

- Write once, read many

- Основа — многомерные вложенные массивы

- Сложная форма

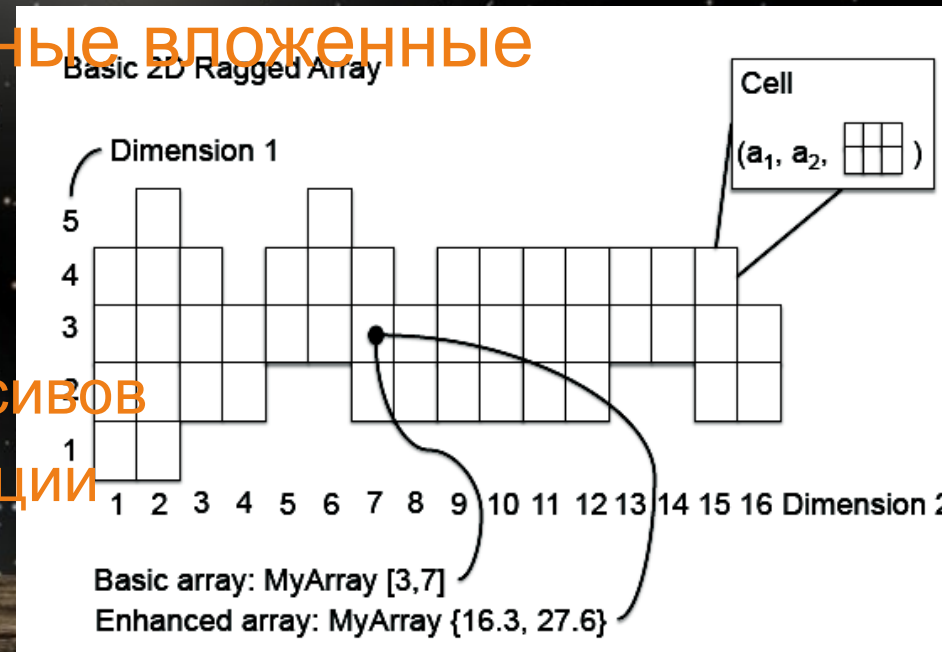
- Разреженность

- Массивы внутри массивов

- Интервальные операции

- Переразбиения

- Системы координат



Современные базы данных: база данных для науки

- SciDB (LSST, LHC, ...)
 - Встроенная поддержка версионности
 - Данные не стираются — добавляется новая версия
 - Встроенная поддержка погрешностей
 - Модель нормальных (гауссовых) ошибок
 - Расширяемость
 - Пользовательские модули и процедуры на C++
 - Масштабируемость
 - Свободная лицензия

Астрономические данные: Заключение

- Данных становится все больше
- Типы данных существенно различны
- Хранить данные надо в СУБД
- Существующие СУБД удовлетворительно справляются с современными каталогами
- ... но не с другими видами данных
- ... и они не позволяют обеспечить полный цикл жизни научных данных
(получение — хранение — анализ — обмен)
- Будущее — за специализированными научными системами управления базами данных

RDBMS

XMLDB

GRID

GoogleApp

ACID

ColumnarDB

NoSQL

EC2

S3

BASE

StreamDB

NEWSQL

CAP

(key,value)

SciDB

UtilityComputing

COA

IaaS

BPELWS

URI

CloudDB

SaaS

MapReduce

SOAP

Web2.0

OWL

UDDI

Science 2.0

REST

RDF

WS

WSDL

HTTP

RDFS

SOA

WOA

Semantic Web

SMTP

XML

VO, WWT

Что делать ?

Учиться самим и учить студентов !

Биоинформатика, геоинформатика

Астроинформатика ?

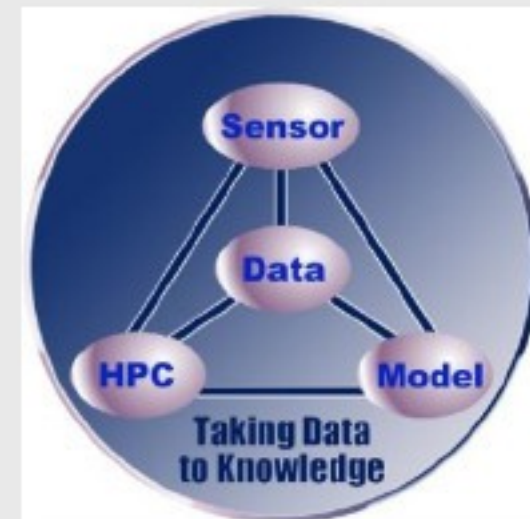
Трудности:

Нет готовых курсов, загруженность студентов, мало практических приложений

Example Course Syllabus:

CSI 710 – Scientific Databases

- CSI 710 Scientific Databases (taught by K.B.) – lectures include:
 - Relational Databases: Modeling, Schemas, Normalization, SQL
 - Scientific Databases, Big Data in Science, The 4th Paradigm
 - E-Science, Ontologies, Semantic E-Science, X-Informatics
 - Distributed Data, Federated Data, Virtual Observatories
 - Citizen Science with Big Data
 - Scientific Data Mining I
 - Scientific Data Mining II
 - Astroinformatics and Astro databases
 - Bioinformatics and Bio databases
 - Geoinformatics and Geo databases
 - Health Informatics
 - Online Science (Jim Gray's KDD-2003 lecture)
 - Intelligent Archives of the Future



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

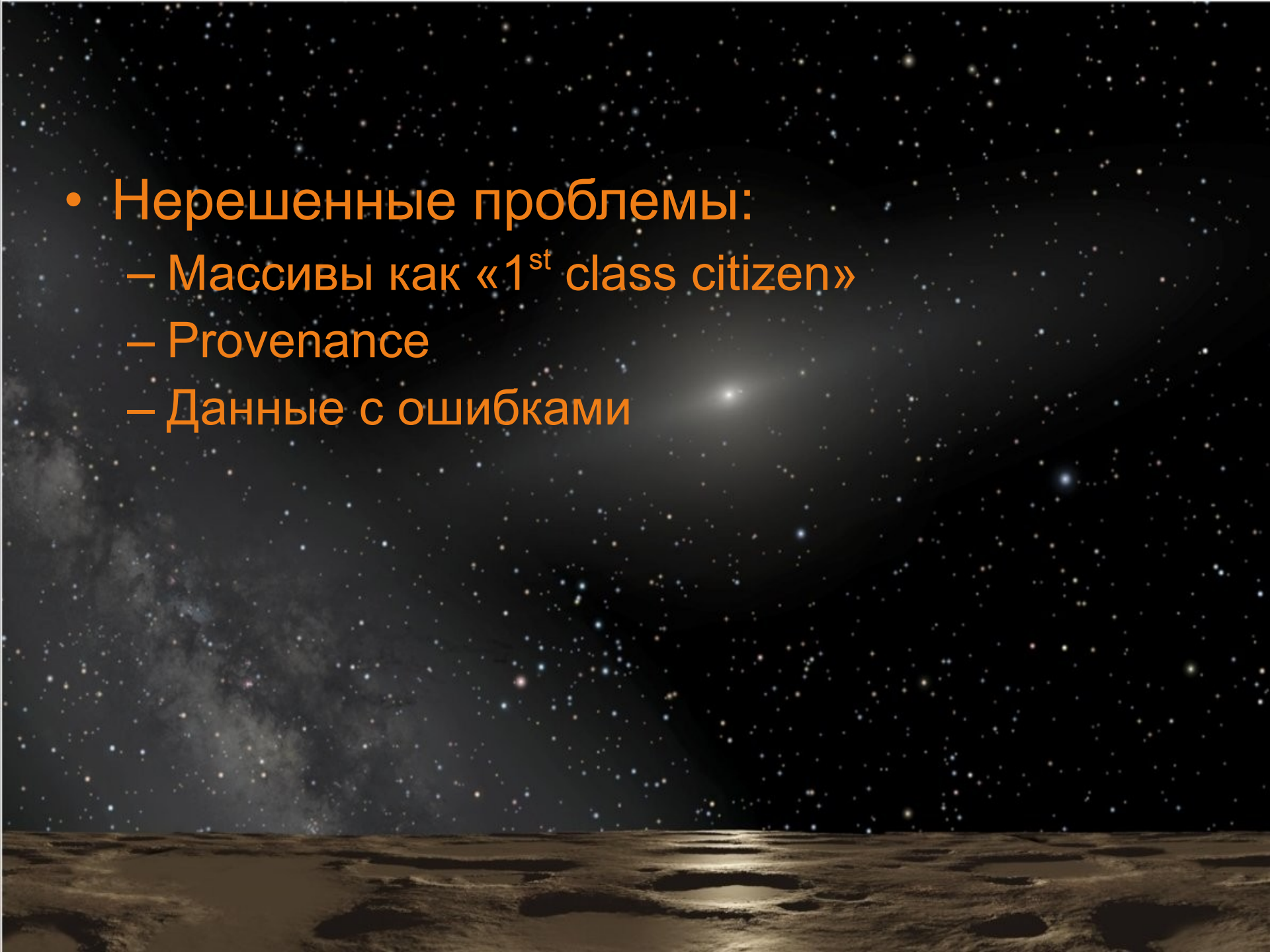
- Тренды
 - Shared-nothing parallel database
 - Дешевые сервера
 - Агрессивная компрессия
- Попытки продвинуть технологии
 - Aster Data, Vertica, ParAccel, Greenplum, Neteza, Teradata

- 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

Специфика научных БД

- Сверхбольшие БД — сотни петабайт в ближайшие 10 лет
- Метаданные в БД и данные в файловой системе — поддерживать целостность
- Редко меняются, особенно «сырые» данные
- OLAP запросы преобладают над OLTP
- Версионность данных — воспроизводимость результатов
- Распределенность данных

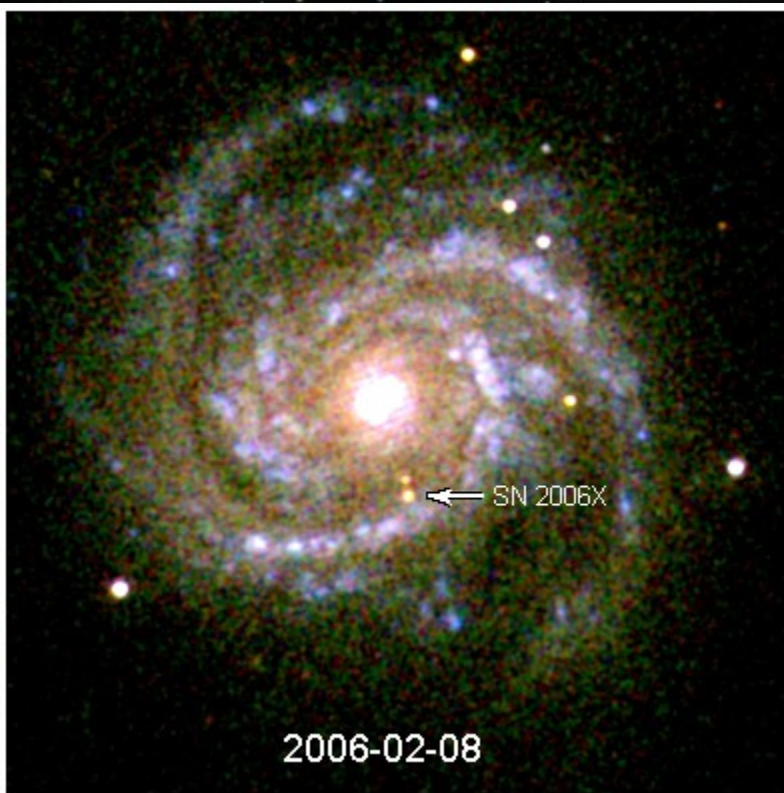
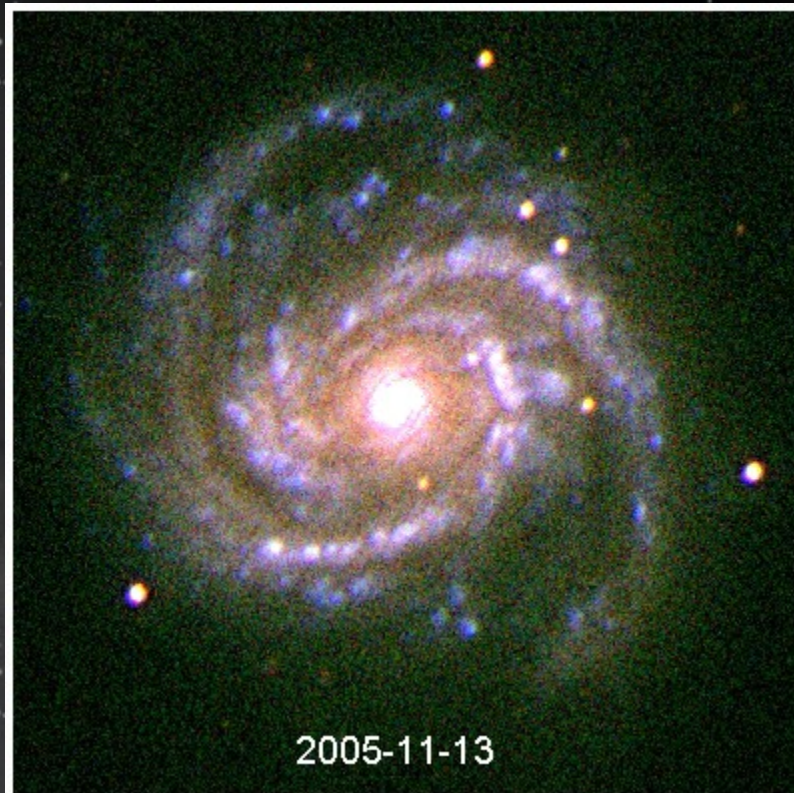


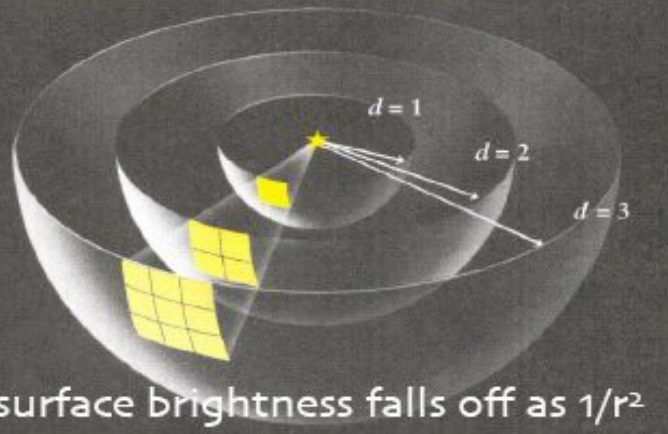
Очень важная научная задача

SN Ia 1994D in NGC 4526

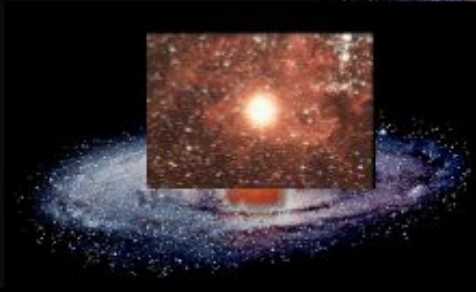
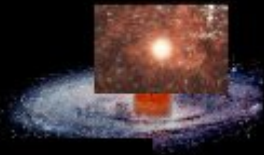


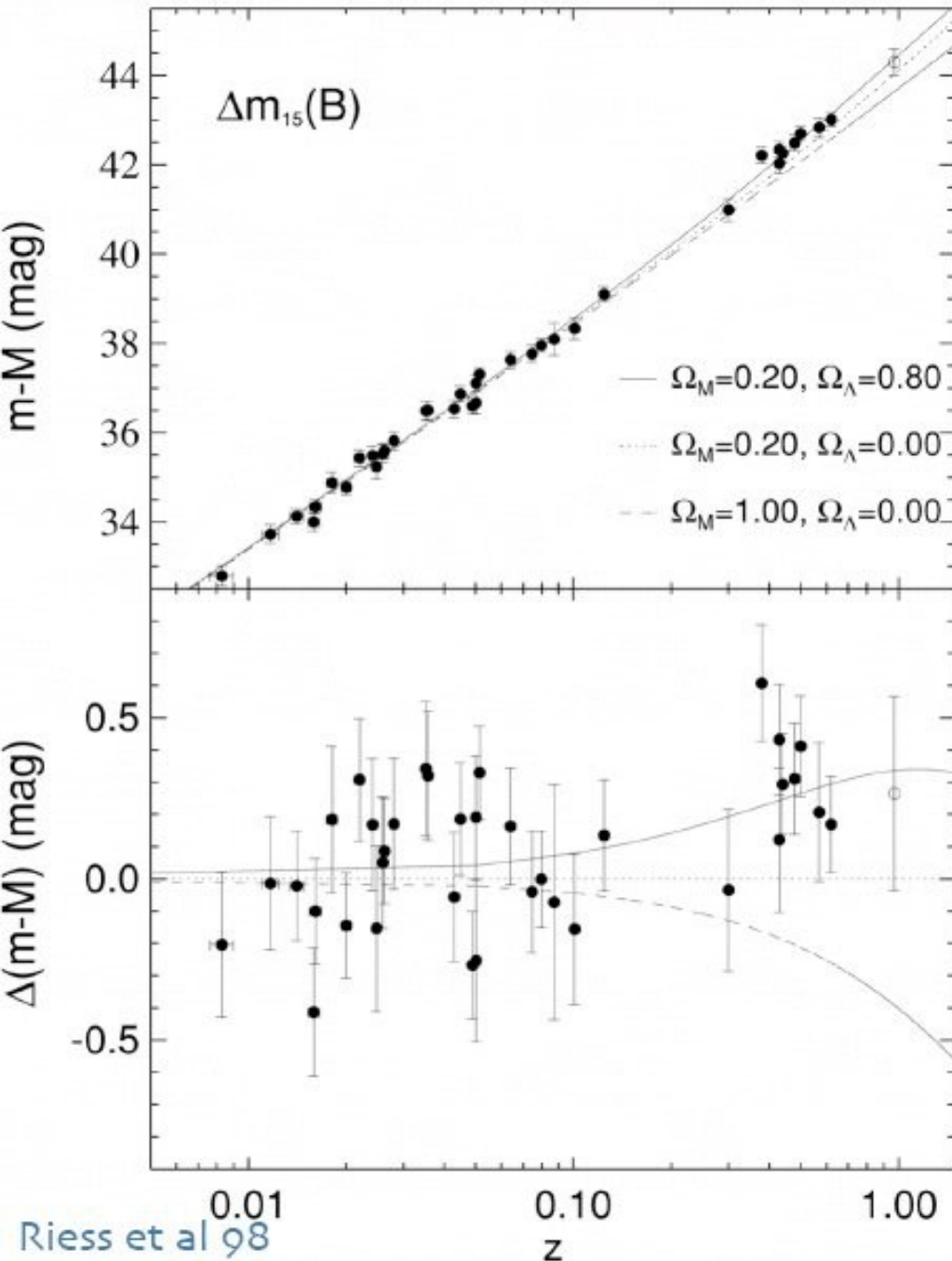
SN Ia 2006X in M100





SN Ia Standard Candle

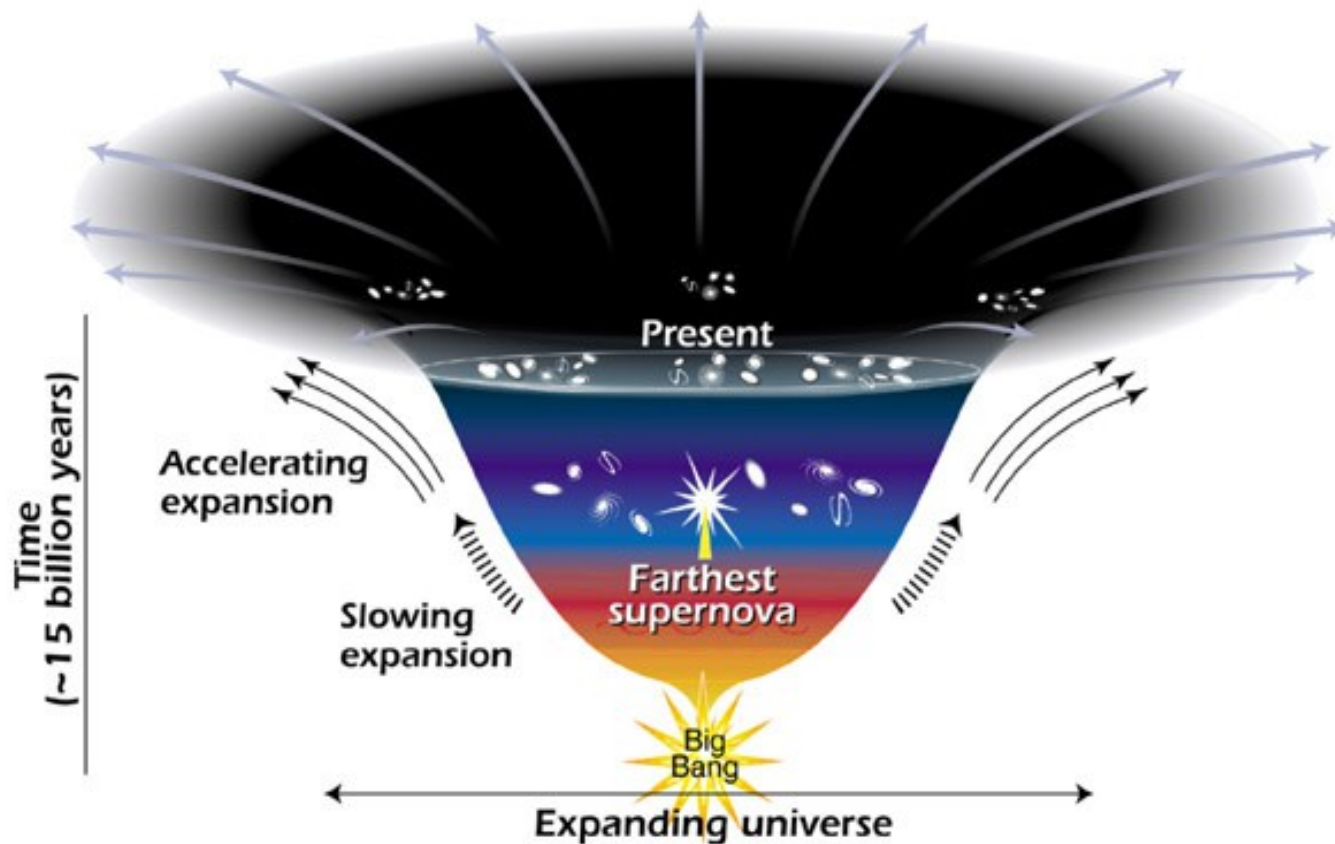




Surprise!

The Universe is
speeding up, not
slowing down!

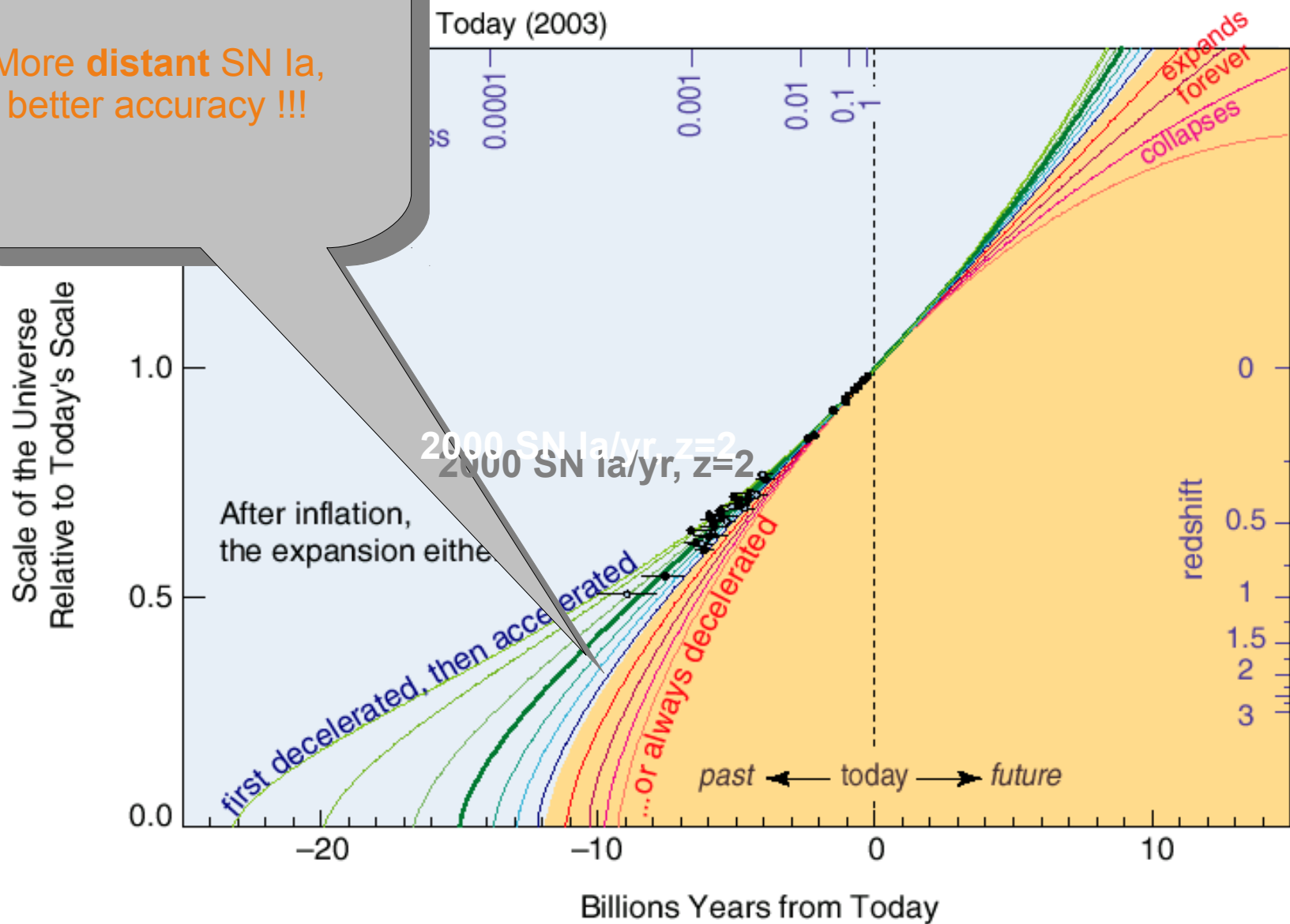
Galaxies are
actually
accelerating away
from us!



This diagram reveals changes in the rate of expansion since the universe's birth 15 billion years ago. The more shallow the curve, the faster the rate of expansion. The curve changes noticeably about 7.5 billion years ago, when objects in the universe began flying apart at a faster rate. Astronomers theorize that the faster expansion rate is due to a mysterious, dark force that is pushing galaxies apart.

Expansion of Universe

More distant SN Ia,
better accuracy !!!



How do we discover a supernovae ?

- Make a picture of the galaxy
 - usually, thousands of galaxies
 - extract all stars from the image
- Find a *new* star
 - compare positions of all stars extracted from the picture with catalog data
 - exclude defects, NEO objects (asteroids, comets,...)
 - find stars-candidates
 - validate stars-candidates

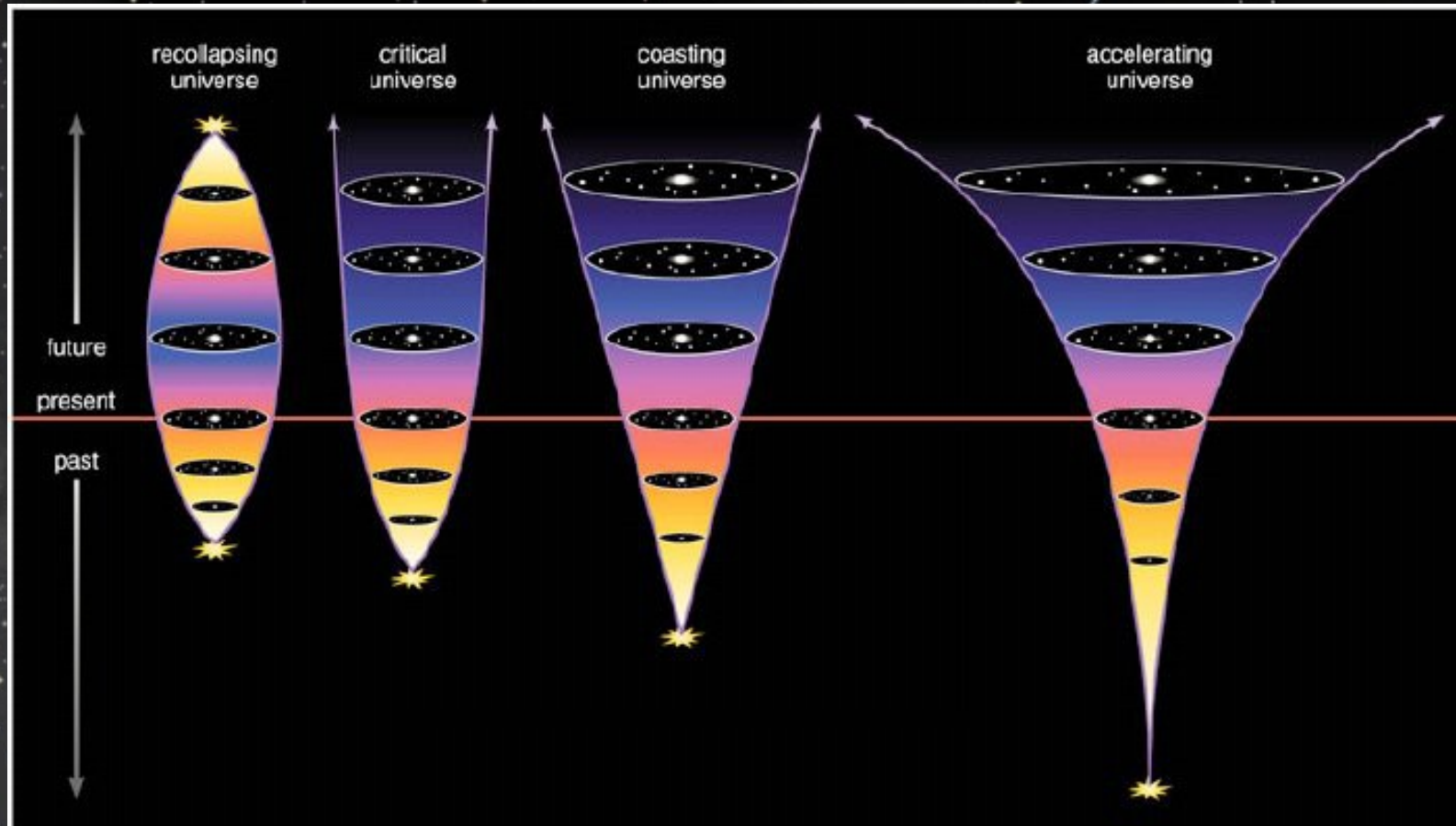
Basic Astronomical Queries

- ConeSearch (radial) query
 - find all stars in a circle on sky of a given radius and position of the center
 - spatial query – spherical coordinates !
- CrossMatch query
 - identify all stars in one set with another with given accuracy
 - fuzzy join of two tables

What do we need ?

- Petabyte databases
 - several seconds for ConeSearch queries
 - several minutes for CrossMatch queries
 - data mining algorithms – N^2 , N^3 are no way for Petabyte databases
 - Flexible usage policy to the system resources
 - Version management

We want to know our FUTURE !



**Спасибо за
Внимание !**