Научные вызовы Анализу 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 кибероббщества.**

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

Наука стала другой

Наука стала другой

- eScience — составная часть информационного общества - синтез науки и информатики
  - роль информации и ее обработка становится доминирующей
- X-informatics - науки, оперирующие 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 !

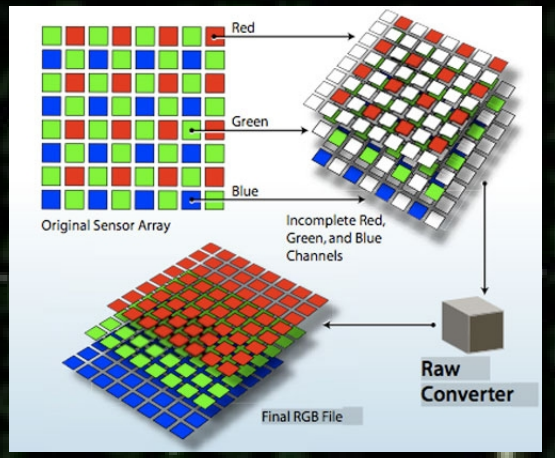
VLDB -> XLDB

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: 100 Pb raw data, Каталог — 60 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

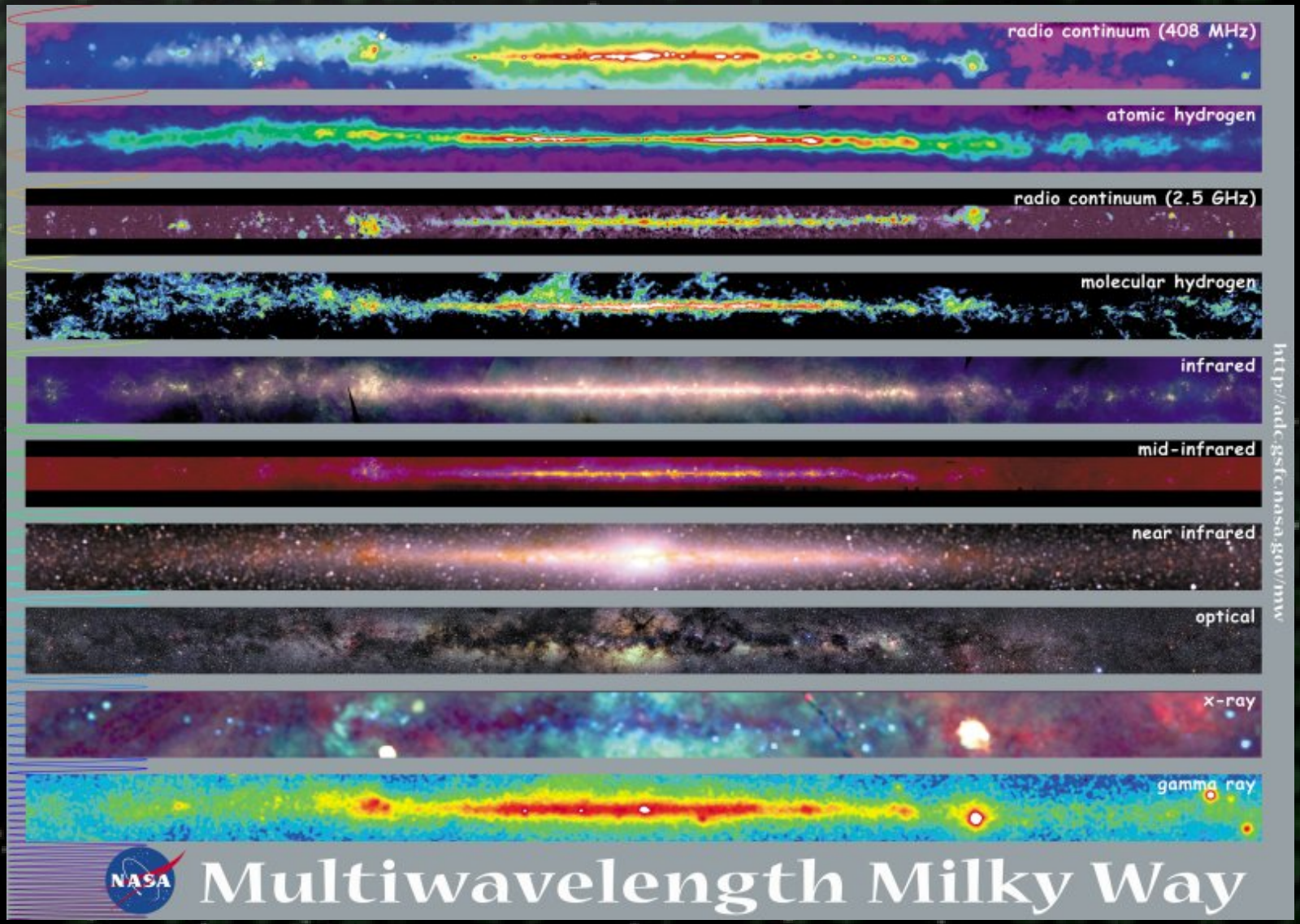
Нейтринные телескопы: (Солнце, SN 1987A)

Гравитационные антенны (начинают работать)

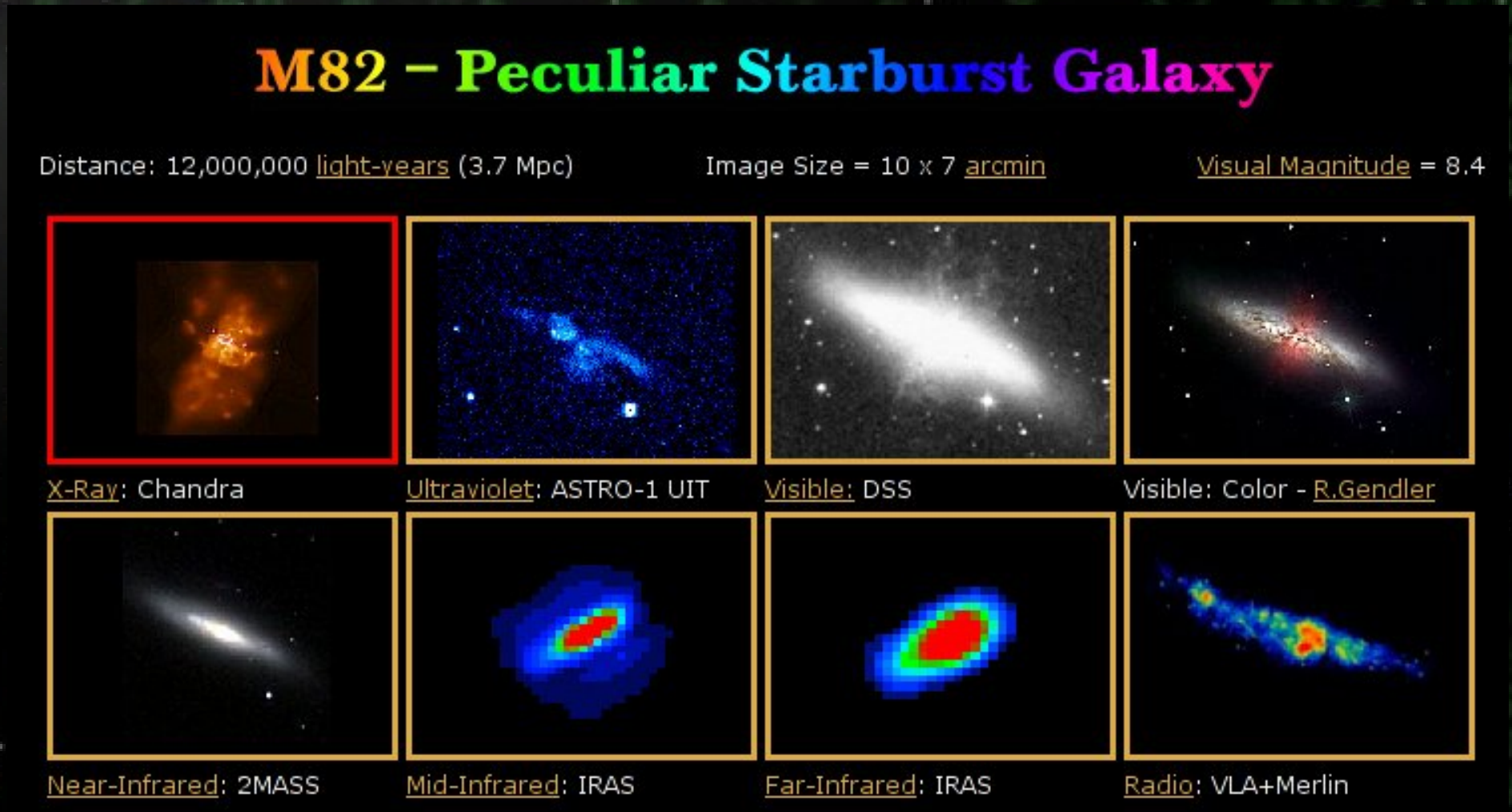
Космические лучи

Разная процедура обработки сырых данных !

Астрономия стала всеволновой



Астрономия стала всеволновой



- Долгое исследование одиночных объектов → гигантские обзоры
- Рутинные работы астрономии → машины
- Интерактивная работа с машиной → межмашинное взаимодействие
- Методы: Computer Science — Data Science — Citizen Science

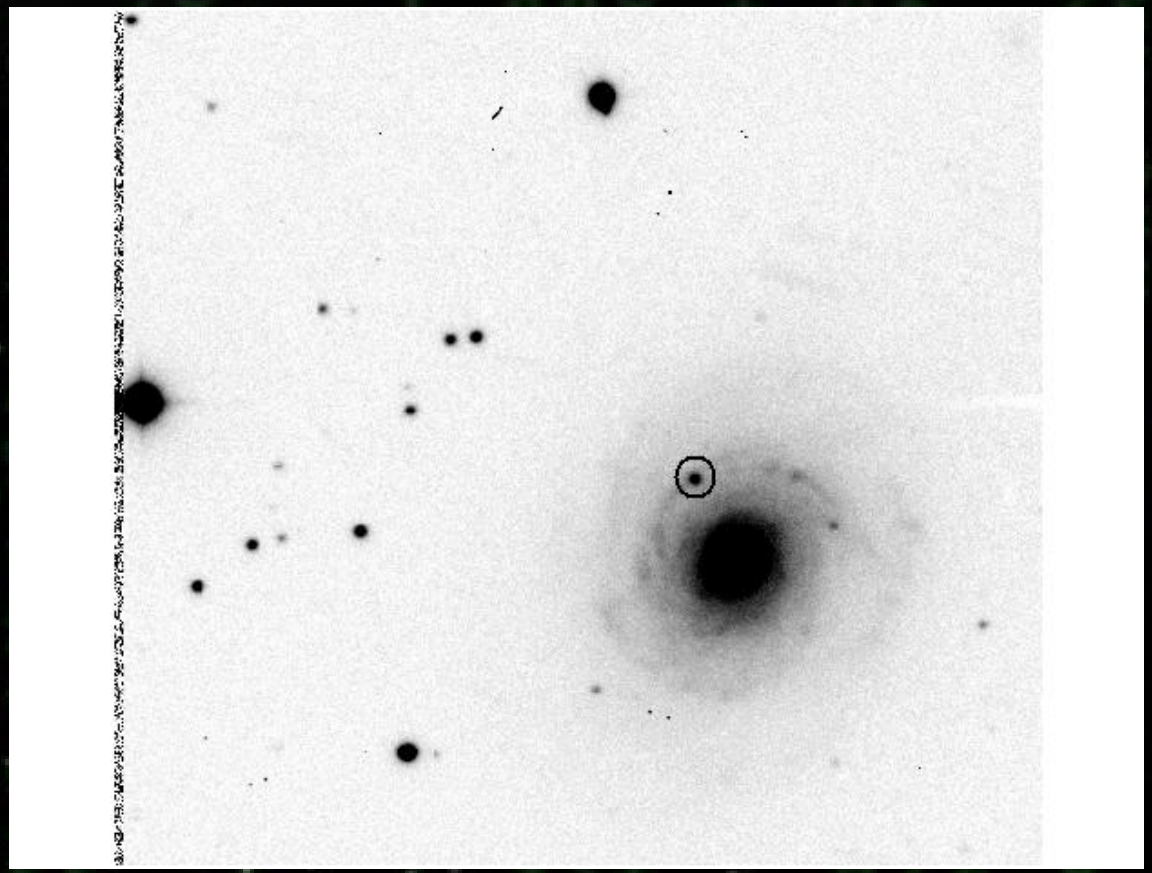
Что такое научные данные ?

- «Сырые» (raw) данные — данные непосредственно из сенсора . Хранятся вечно !
- Обработка (cooking) «сырых» данных — сложная процедура. Мы все больше абстрагируемся от объекта изучения
  - Изучение звезд: **глаз-**фото-эоп-ccd
  - Открытие частиц: треки в камере Вильсона, сейчас CCD в LHC

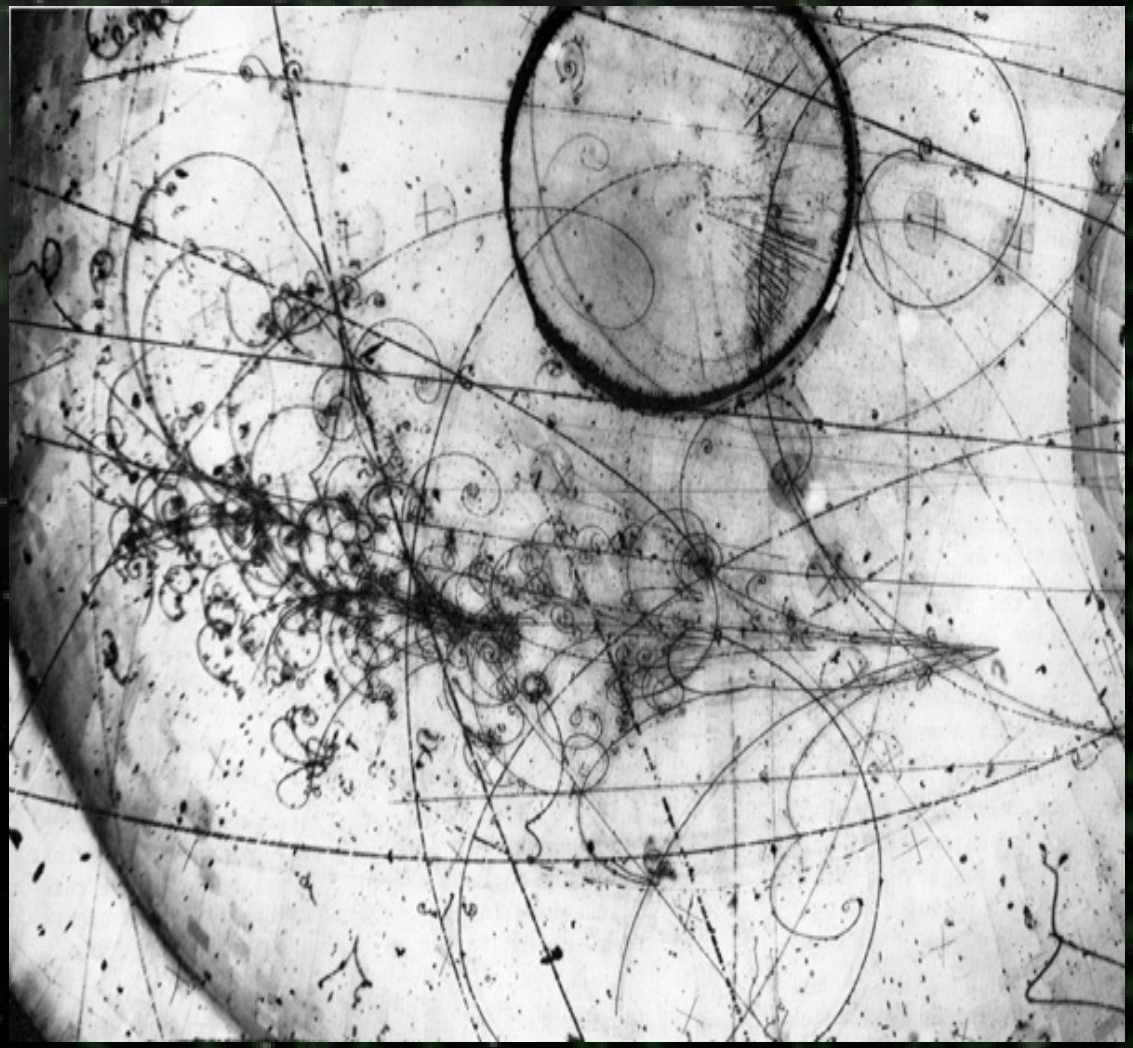
«Сырые» данные

Сенсорные данные естественно хранить в массивах

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



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

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

Uncertain Data

- 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)
  - Ошибки в координатах → ошибки в отождествлении → ошибки в показателях цвета

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 !!!

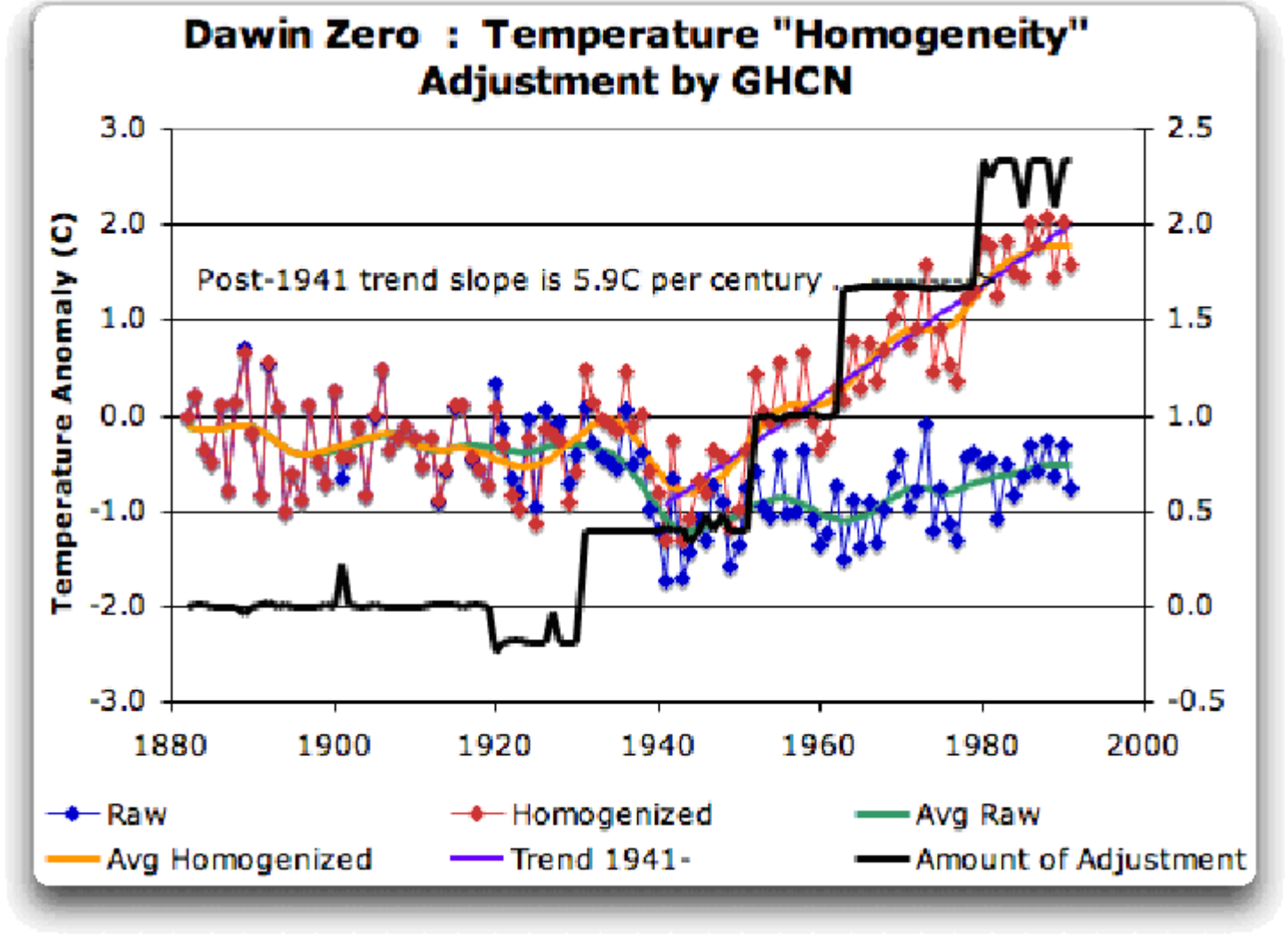
Принцип науки

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

Заключение

- «Sensor-centric» science - много данных
- Воспроизводимость научных результатов
  - Хранение «сырых данных» - вечное, реальная возможность извлечения
  - Хранение всех метаданных процесса обработки «сырых» данных

Data Science

- DataScience (научные данные) → Знания
- Астрономия — сеть роботизированных телескопов:
  - Автоматическое вытаскивание фич
  - Классификация
  - Поиск аномалий
  - Формирование заданий для крупных и специализированных инструментов

http://mygalaxies.co.uk