Неточные данные в эпоху Big Data

Олег Бартунов, ГАИШ-МГУ
Сергей Карпов, САО

- Экзафлопсы (10^18) будущего
  - Воеводин, утренний доклад
- Зетабайты (10^21) настоящего
  - IDC: 1.2 Zb — полное кол-во данных в цифровом виде

Big Data

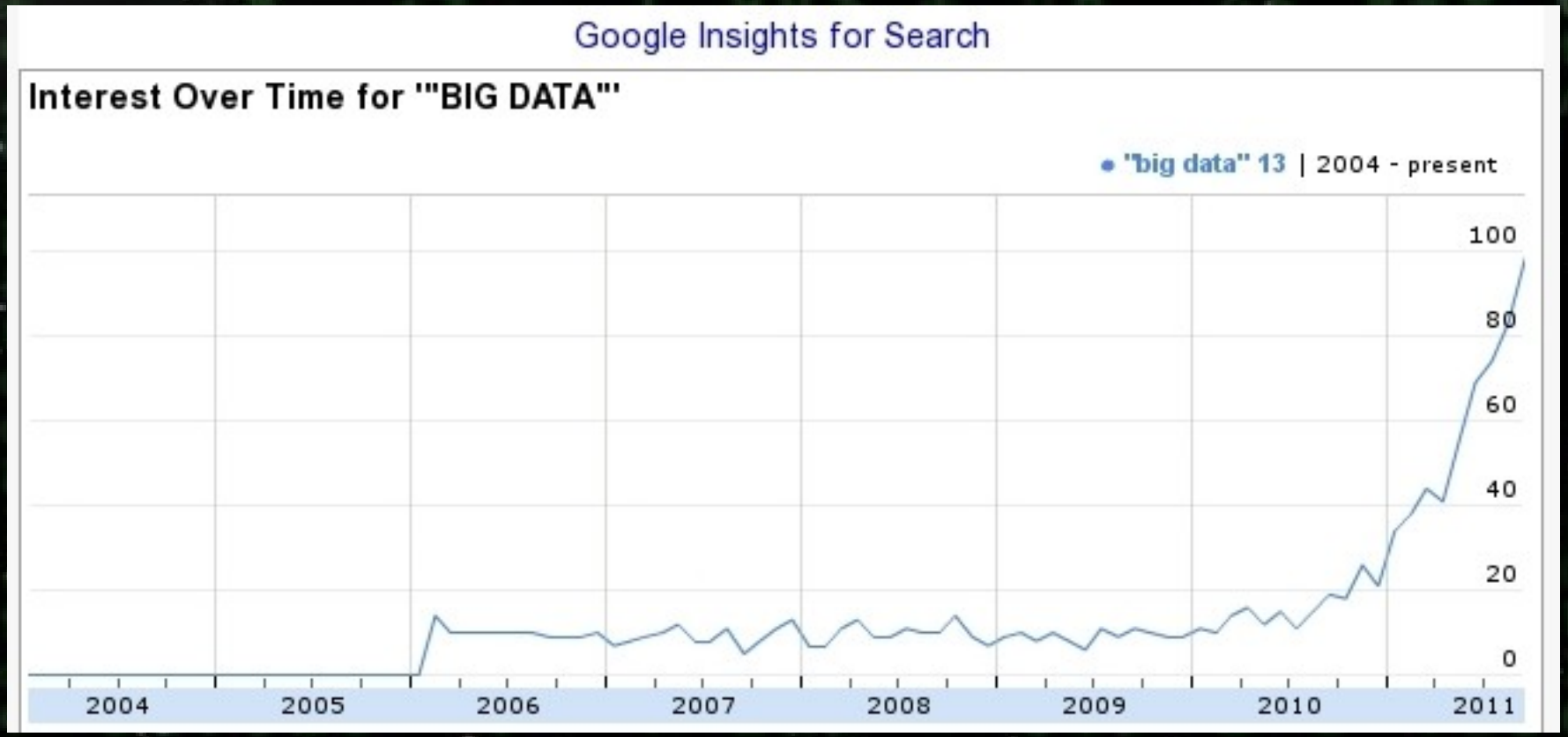
- Проблема [не]управления данными
  - Персональные архивы данных (xTb)
  - Институтские хранилища (xxxTb)
  - «Большая наука» (xPb)
    - ATLAS (LHC): петабайт/с !
    - LSST: 20 Tb/ночь, 200 Pb архив (Хаббл, 15 лет,25 Тб)
    - SKA (радиотелескоп, 1кв.км) — экзабайт сырых данных в сутки !
  - Гугл: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

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 Тб)
    - SKA (радиотелескоп, 1кв.км) — экзабайт сырых данных в сутки !
- Требуется *специальные* решения

Big Data

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



Big Data

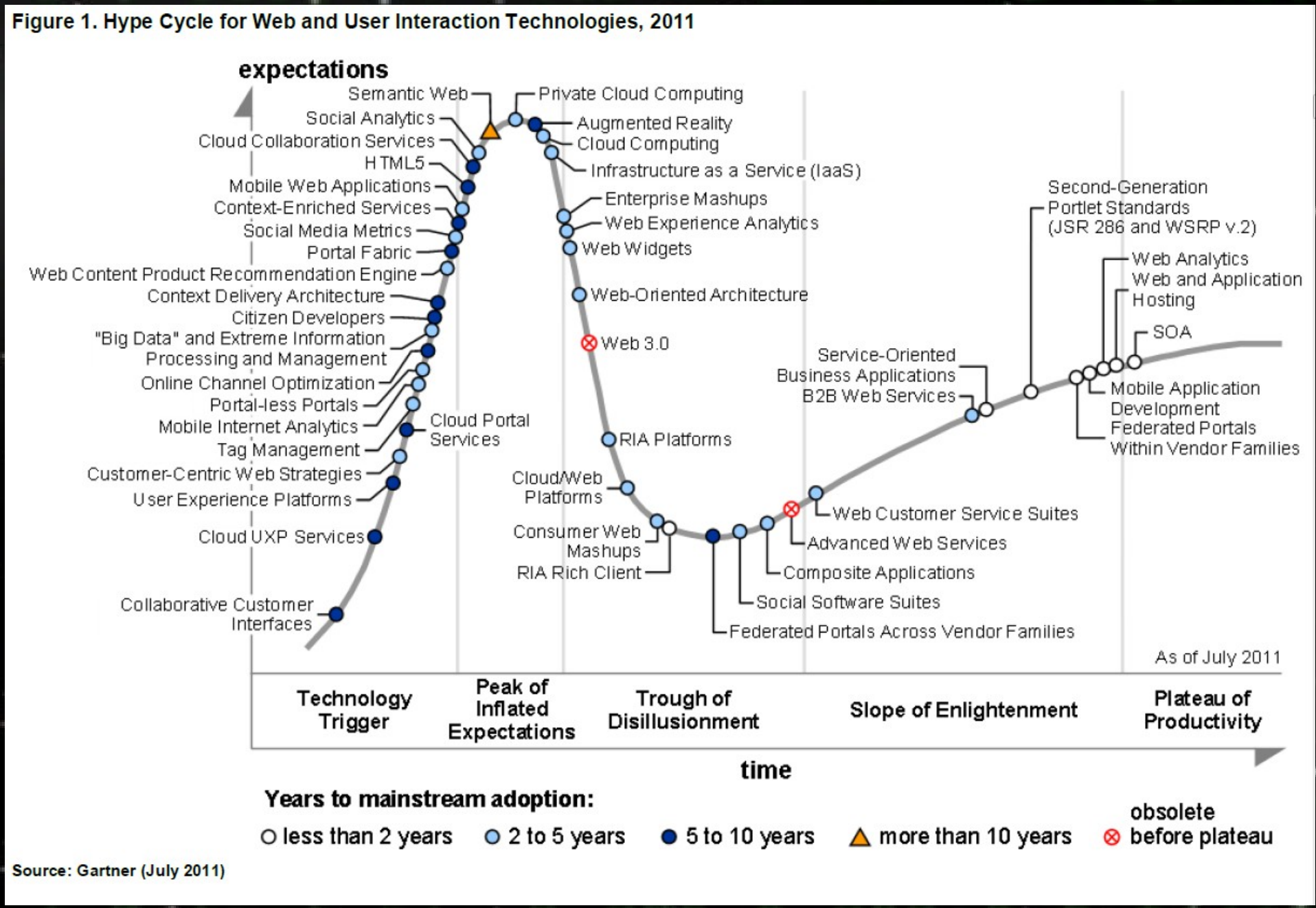


Figure 1. Hype Cycle for Web and User Interaction Technologies, 2011
expectations
Semantic Web
Private Cloud Computing
Social Analytics
Augmented Reality
Cloud Collaboration Services
Cloud Computing
HTML5
Infrastructure as a Service (IaaS)
Mobile Web Applications
Enterprise Mashups
Context-Enriched Services
Web Experience Analytics
Social Media Metrics
Web Widgets
Portal Fabric
Web Content Product Recommendation Engine
Web-Oriented Architecture
Context Delivery Architecture
Citizen Developers
Web 3.0
"Big Data" and Extreme Information Processing and Management
Online Channel Optimization
Portal-less Portals
Mobile Internet Analytics
Cloud Portal Services
RIA Platforms
Tag Management
Customer-Centric Web Strategies
Cloud/Web Platforms
User Experience Platforms
Consumer Web Mashups
Cloud UXP Services
RIA Rich Client
Collaborative Customer Interfaces
Social Software Suites
Federated Portals Across Vendor Families
Second-Generation Portlet Standards (JSR 286 and WSRP v.2)
Service-Oriented Business Applications
B2B Web Services
Web Analytics
Web and Application Hosting
SOA
Mobile Application Development
Federated Portals Within Vendor Families
Web Customer Service Suites
Advanced Web Services
Composite Applications
As of July 2011
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Years to mainstream adoption:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
Source: Gartner (July 2011)

Big Data

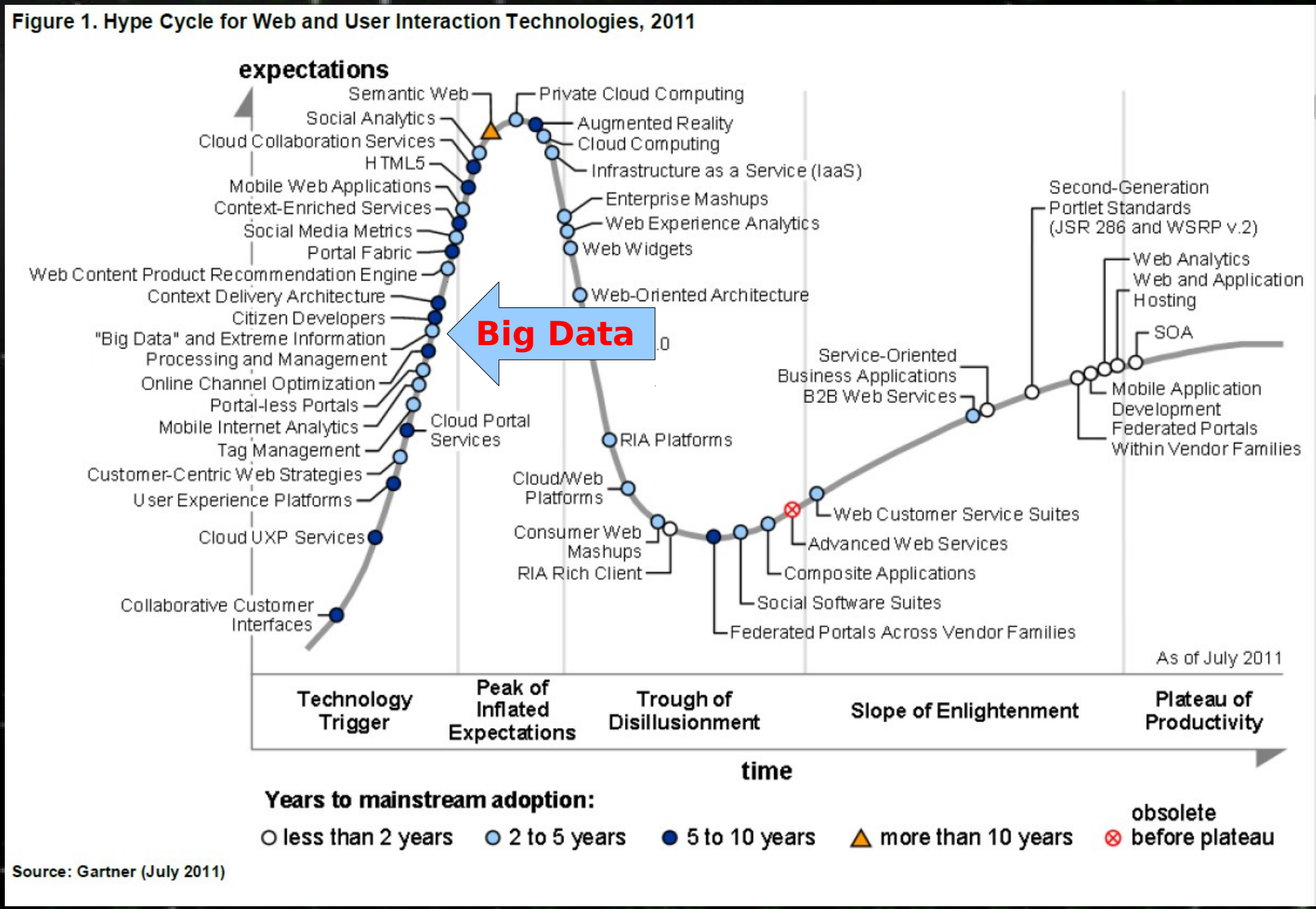


Figure 1. Hype Cycle for Web and User Interaction Technologies, 2011
expectations
Semantic Web
Private Cloud Computing
Social Analytics
Augmented Reality
Cloud Collaboration Services
Cloud Computing
HTML5
Infrastructure as a Service (IaaS)
Mobile Web Applications
Enterprise Mashups
Context-Enriched Services
Web Experience Analytics
Social Media Metrics
Web Widgets
Portal Fabric
Web Content Product Recommendation Engine
Web-Oriented Architecture
Context Delivery Architecture
Citizen Developers
Big Data
"Big Data" and Extreme Information Processing and Management
Online Channel Optimization
Portal-less Portals
Mobile Internet Analytics
Cloud Portal Services
Tag Management
RIA Platforms
Customer-Centric Web Strategies
Cloud/Web Platforms
User Experience Platforms
Consumer Web Mashups
Cloud UXP Services
RIA Rich Client
Collaborative Customer Interfaces
Social Software Suites
Federated Portals Across Vendor Families
Second-Generation Portlet Standards (JSR 286 and WSRP v.2)
Web Analytics
Web and Application Hosting
SOA
Service-Oriented Business Applications
B2B Web Services
Mobile Application Development
Federated Portals Within Vendor Families
Web Customer Service Suites
Advanced Web Services
Composite Applications
As of July 2011
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Years to mainstream adoption:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
Source: Gartner (July 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 милисекунды 1

Big Data

- Обработка петабайтных данных в HPC 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
  -

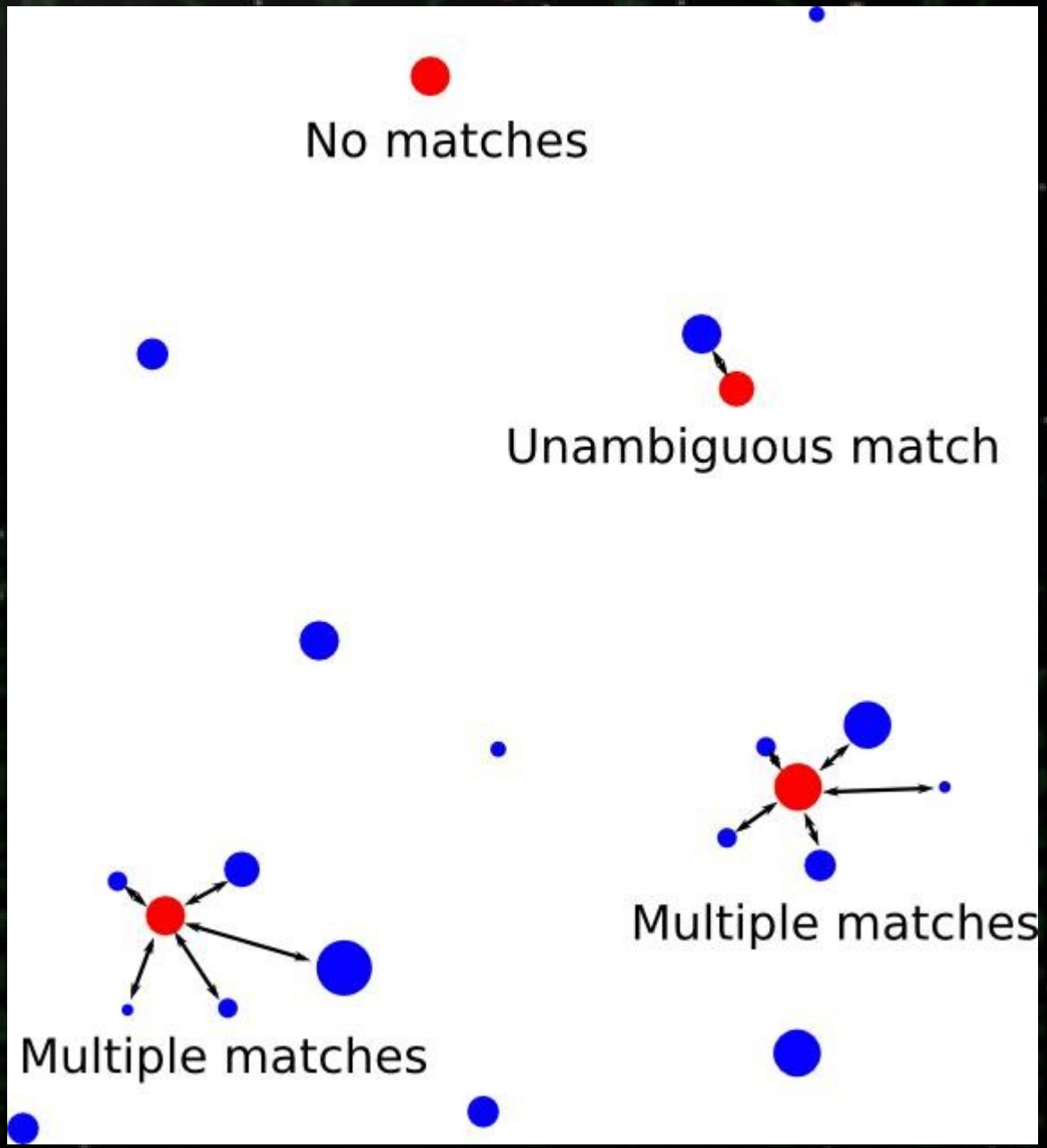
Uncertain Data

- 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

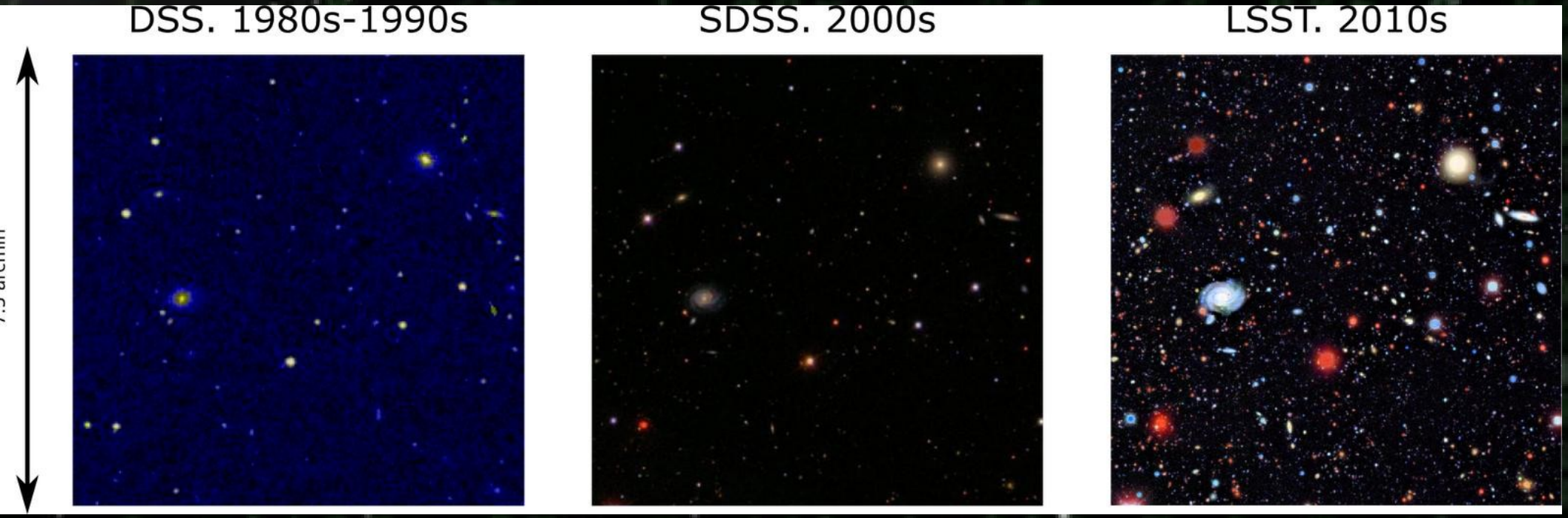
- Типичная задача астрономии — взаимное отождествление объектов из разных каталогов — вероятностная !
  - Координаты могут существенно отличаться.
    - Разные методики наблюдений, разные привязки систем координат, точности наблюдений...
    - Положение объектов на небе может меняться (иногда существенно)
  - 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



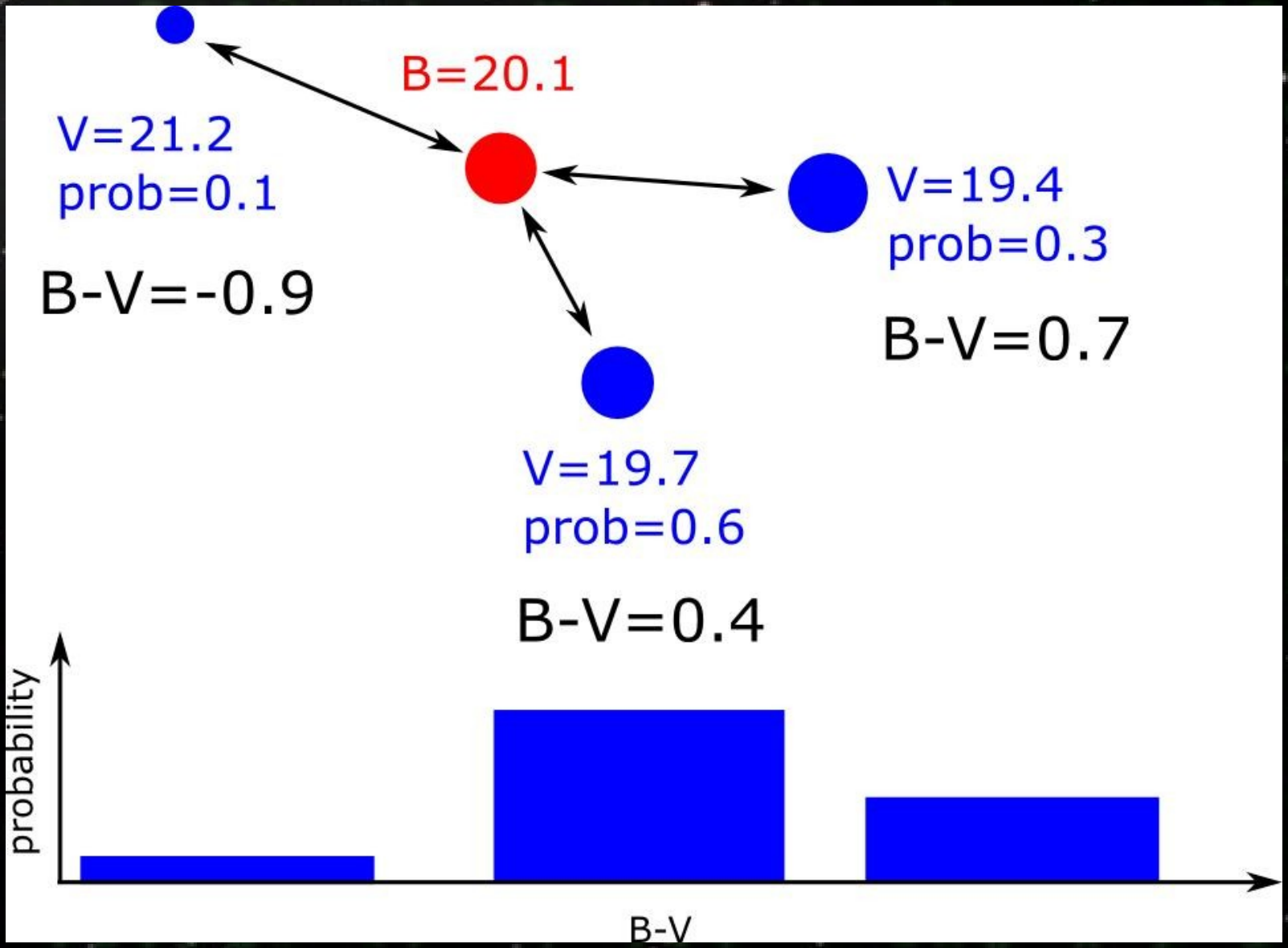
- LSST — 100 млн. звезд/минуту для выделения новых или движ. объектов

Uncertain Data

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

Uncertain Data



Неточности в координатах перешли в неточности в показатели цвета !

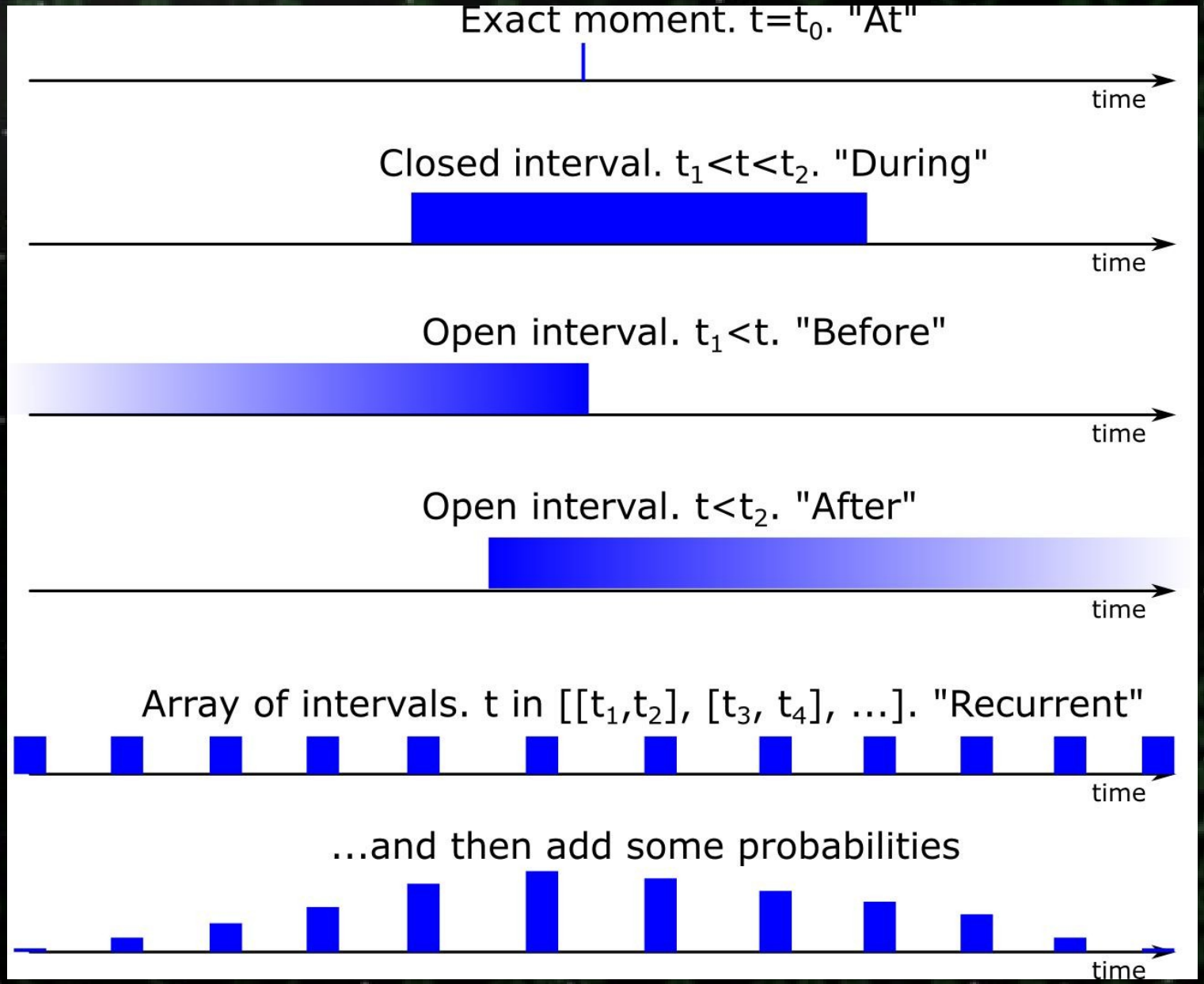
Uncertain Data

- Необходимо научить машины работать с неточными данными:
  - Хранение
  - Обработка
  - Анализ
  - Принятие решений !
- СУБД надо научить работать с неточными данными
  - Разработка с нуля - академическая работа
  - Расширение существующей — сразу в бой

Uncertain Data

- PostgreSQL — наиболее продвинутая, свободная, расширяемая, популярная в научном сообществе, СУБД
- тип данных hdate для исторических дат, представляемые интервалами
  - Интервалы открытые и закрытые
  - Поддерживается интервальная алгебра Аллена
  - Операции сравнения, пересечения, включения

Uncertain Data



Uncertain Data

```
CREATE TABLE supernovae (name TEXT, date hdate);
INSERT INTO supernovae (name, date) VALUES ('SN185', hdate('185'));
INSERT INTO supernovae (name, date) VALUES ('SN386', hdate('386'));
INSERT INTO supernovae (name, date) VALUES ('SN393', hdate('393'));
INSERT INTO supernovae (name, date) VALUES ('SN1006', hdate('1006'));
INSERT INTO supernovae (name, date) VALUES ('SN1054', hdate('1054'));
INSERT INTO supernovae (name, date) VALUES ('SN1181', hdate('1181'));
INSERT INTO supernovae (name, date) VALUES ('SN1572', hdate('1572'));
INSERT INTO supernovae (name, date) VALUES ('SN1604', hdate('1604'));
INSERT INTO supernovae (name, date) VALUES ('SN Cas', hdate('1680'));

CREATE TABLE astronomer(name TEXT, life hdate);
INSERT INTO astronomer (name, life) VALUES
    ('Tycho Brahe', hdate('1546.12.14', '1601.10.24'));
INSERT INTO astronomer (name, life) VALUES
    ('Johannes Kepler', hdate('1571.12.27', '1630.11.15'));

SELECT s.name, a.name FROM supernovae s, astronomer a
                          WHERE s.date = a.life;
```

SN1572 | Tycho Brahe
SN1572 | Johannes Kepler
SN1604 | Johannes Kepler

Uncertain Data

- Что планируется
  - Новый тип данных UNCERTAIN, задаваемый не значением, а функцией распределения вероятности
    - Неопределенность по значению
    - Неопределенность по существованию
  - Арифметические преобразование величин
  - Вероятностные аналоги обычных операций отношения двух величин (больше, меньше, равно, пересечение, включение,...)

Спасибо !