Особенности Big Data

Максим Губин

Томск

econophysica

Два типа больших данных

Это Big Data:	Это тоже Big Data:
Поток данных телескопа VLA	Сделки биржи NYSE
Поток данных LHC	Действия игроков Eve Online
GPS-треки с общественного транспорта	Facebook
Покупки в супермаркете	YouTube
Wayback Machine	E-Bay
	Транзакции Visa и MasterCard (и Мир)

Copyright © Econophysica 2019. All Rights Reserved

«Научные» Big Data

- Обычно выход датчиков какого-то рода;
- Может быть ограничение на количество данных, получаемых в единицу времени, но обычно это довольно большое значение;
- Обогащение данных может привести к тому, что они вырастут до произвольного размера, одновременно увеличивая их полезность.

При работе с «научным» видом больших данных необходимо помнить одну важную концепцию:

Copyright © Econophysica 2019. All Rights Reserved

«Научные» Big Data

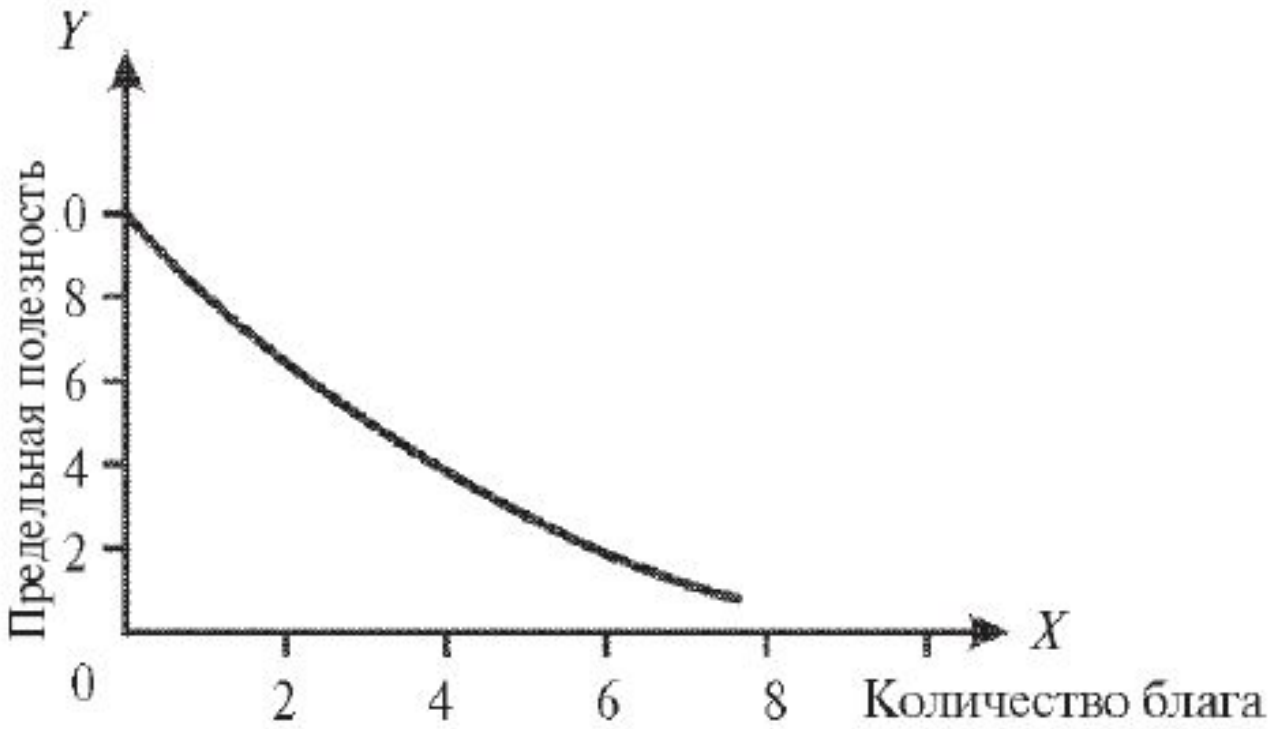


Рис. 2. Кривая предельной полезности

Copyright © Econophysica 2019. All Rights Reserved

«Научные» Big Data

Вопросы, на которые стоит ответить:

- Насколько растёт точность наших моделей с каждой следующей записью?
- Сколько стоит получить следующую запись?
- Являются ли некоторые данные более ценными, чем другие?
- Можем ли мы пожертвовать некоторыми записями, при этом получая пользу?
- Когда мы начнем терять полезность из-за накладных расходов?
- Выйдем ли мы в плюс, если продолжим наращивать объёмы данных?

Copyright © Econophysica 2019. All Rights Reserved

«Научные» Big Data

Важные особенности:

- Цена данных обычно низкая;
- Стабильное хранение данных обычно не требуется;
- Сами данные не ценны и не полезны, ценны и полезны результаты их исследования;
- Согласованность данных важна, но не критична;
- Потеря даже 100% данных во многих случаях является лишь незначительной проблемой.*

Copyright © Econophysica 2019. All Rights Reserved

«Бизнес» Big Data

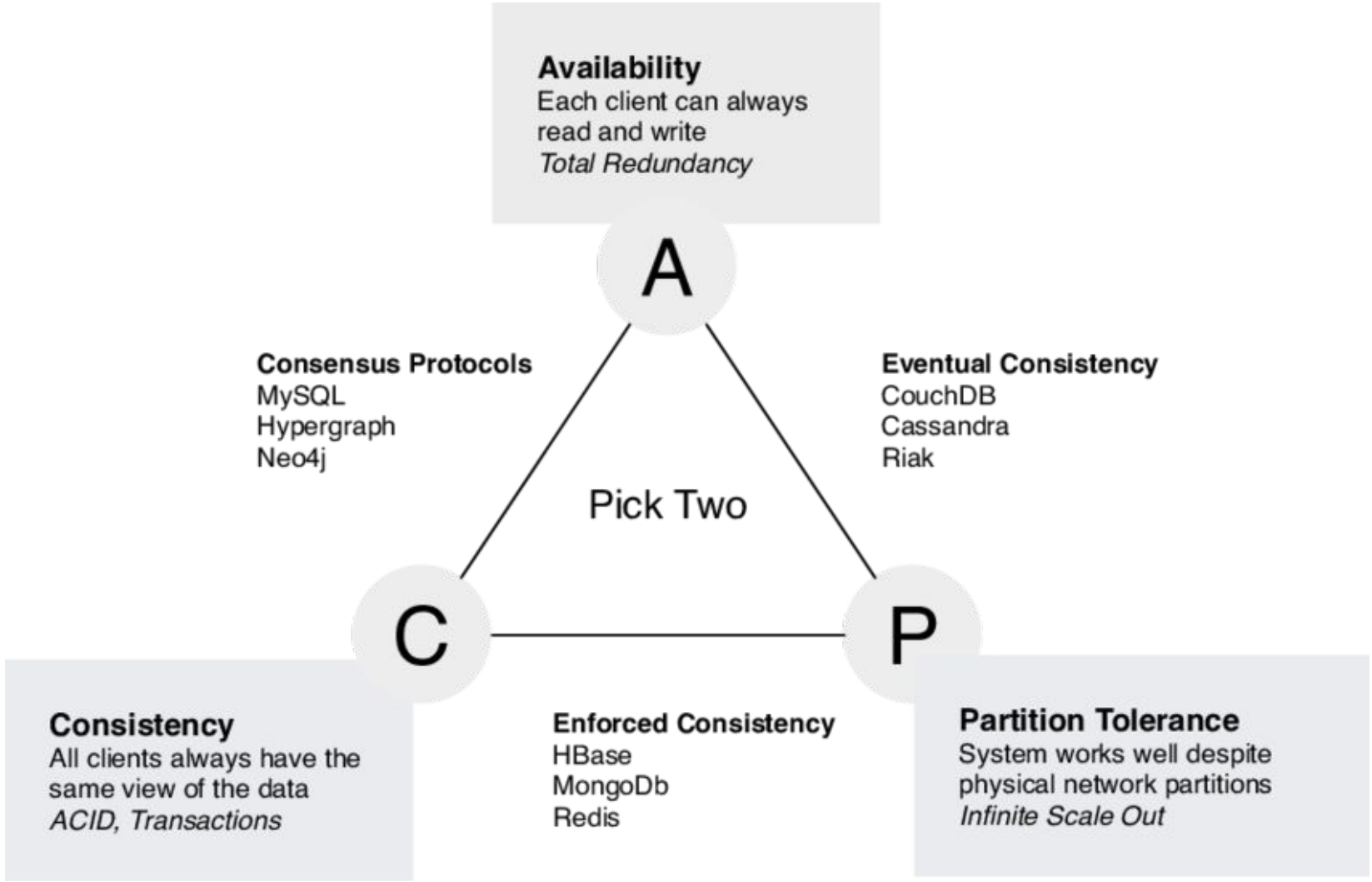
Важные особенности:

- Бизнес-данные, выросшие настолько, что с ними уже нельзя работать традиционными подходами.
- Такие данные обычно важные, дорогие, требуют длительного хранения, и потеря даже малой их части может быть катастрофической.

Здесь компромиссы гораздо менее выражены, потому что такие данные очень плохо переносят деградацию объёма.

Copyright © Econophysica 2019. All Rights Reserved

Теорема CAP



Copyright © Econophysica 2019. All Rights Reserved

Теорема CAP

Consistency:

Каждое чтение возвращает самые свежие записанные данные либо ошибку.

Copyright © Econophysica 2019. All Rights Reserved

Теорема CAP

Availability:

Каждый запрос вернет ответ без гарантий, что в ответе содержатся самые свежие данные.

Copyright © Econophysica 2019. All Rights Reserved

Теорема CAP

Partition Tolerance:

Система продолжает работать, несмотря на произвольный уровень потери связности её узлов.

Copyright © Econophysica 2019. All Rights Reserved

Теорема CAP, 3 варианта

Consistency:

Система вернет **ошибку или таймаут**, если не может гарантировать актуальность данных из-за проблем с сетью.

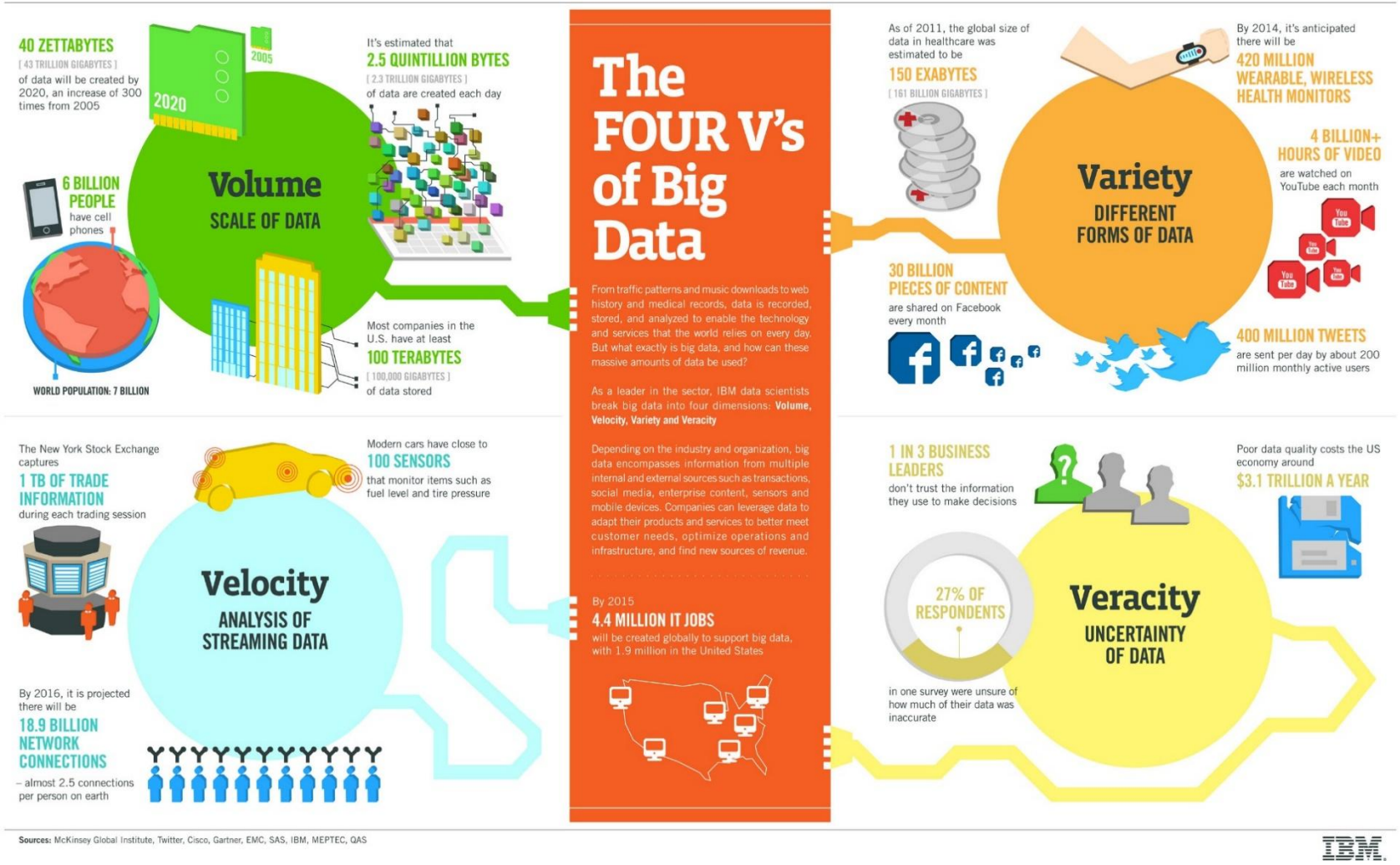
Availability:

Система всегда ответит на запрос **самой новой доступной версией** данных, даже если она не может гарантировать актуальности информации из-за проблем с сетью.

Третьего варианта нет. Если БД не фрагментирована, у вас есть все 3 полезных свойства.

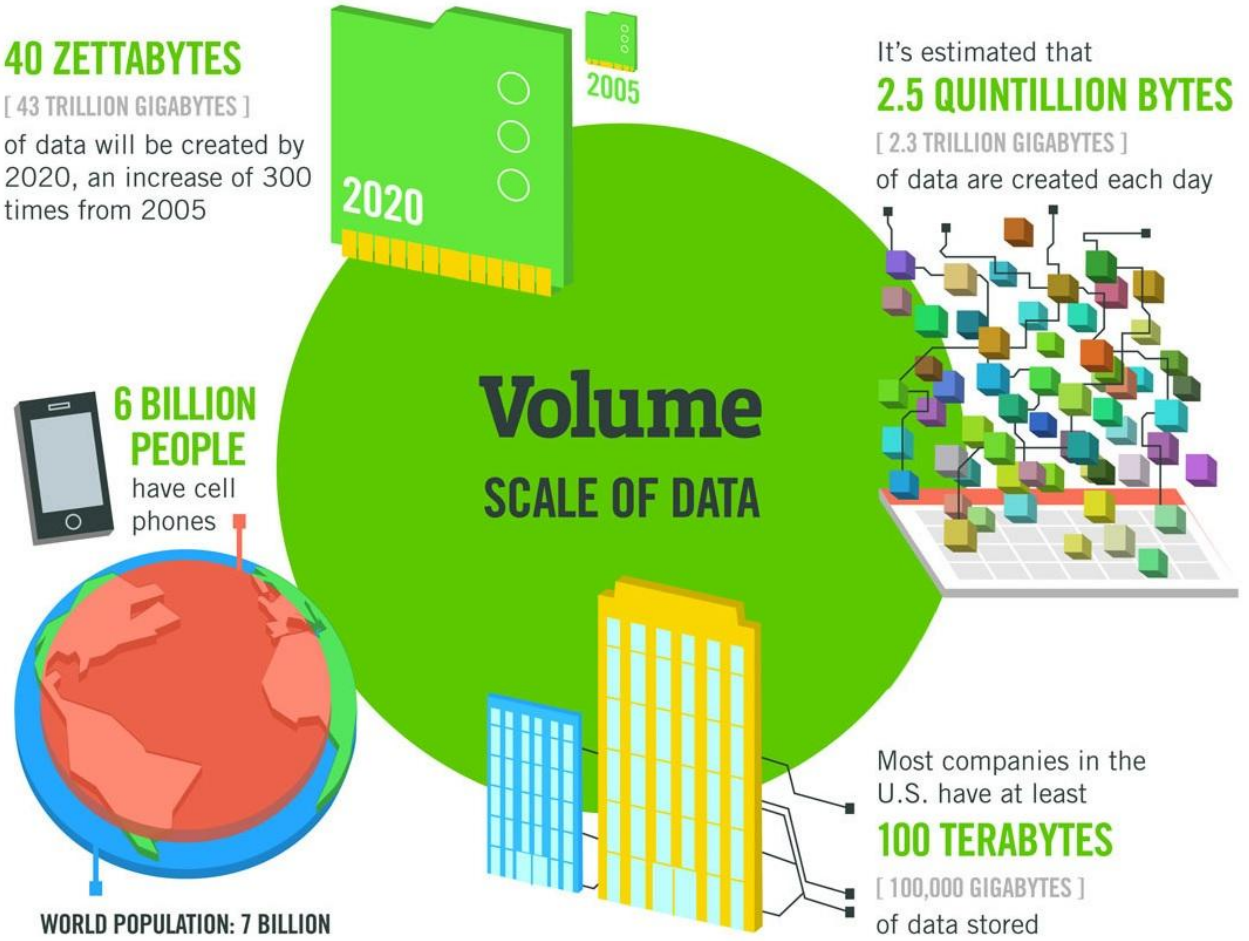
Copyright © Econophysica 2019. All Rights Reserved

Volume, Variety, Veracity, Velocity



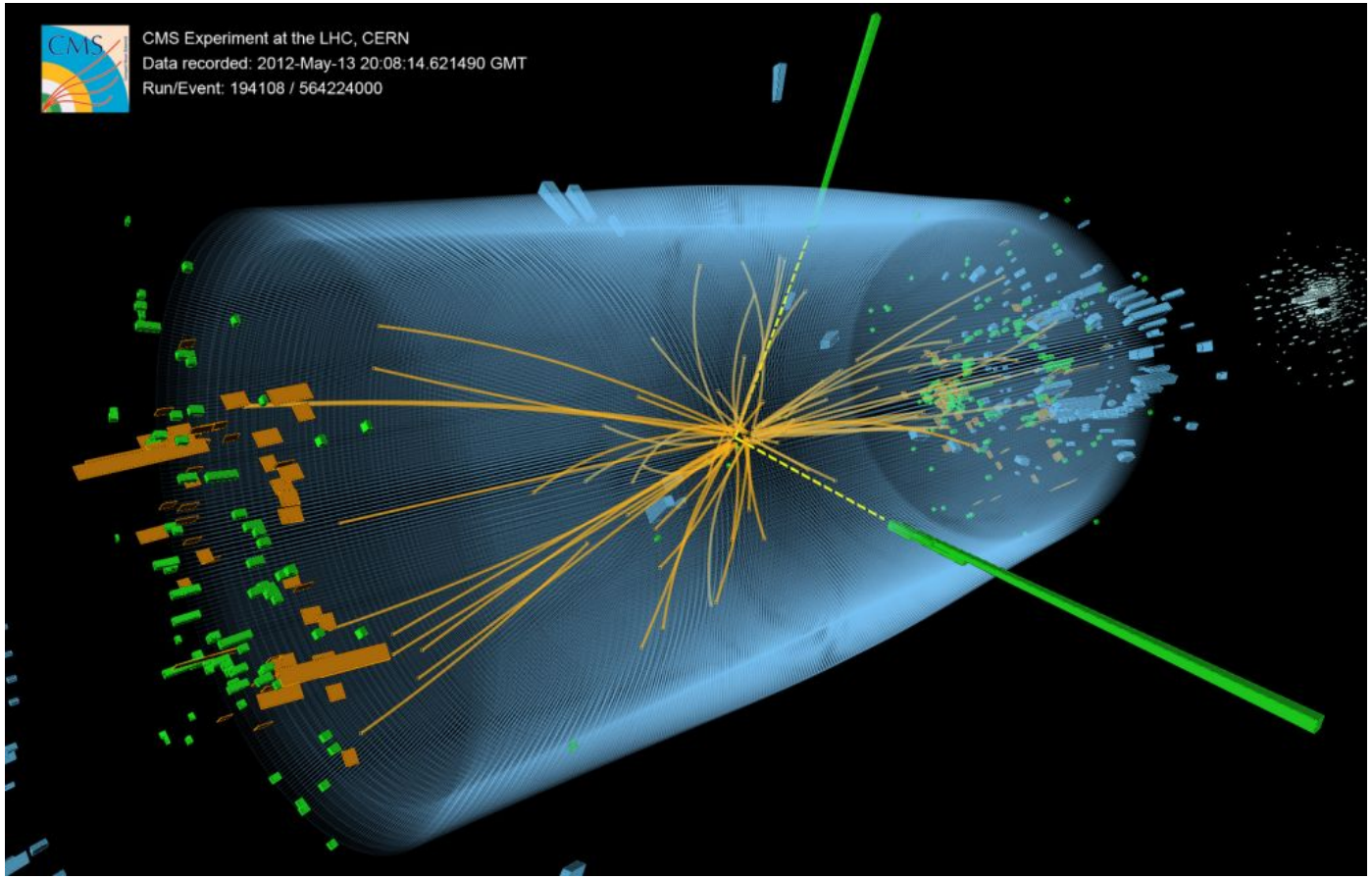
Copyright © Econophysica 2019. All Rights Reserved

Volume



Copyright © Econophysica 2019. All Rights Reserved

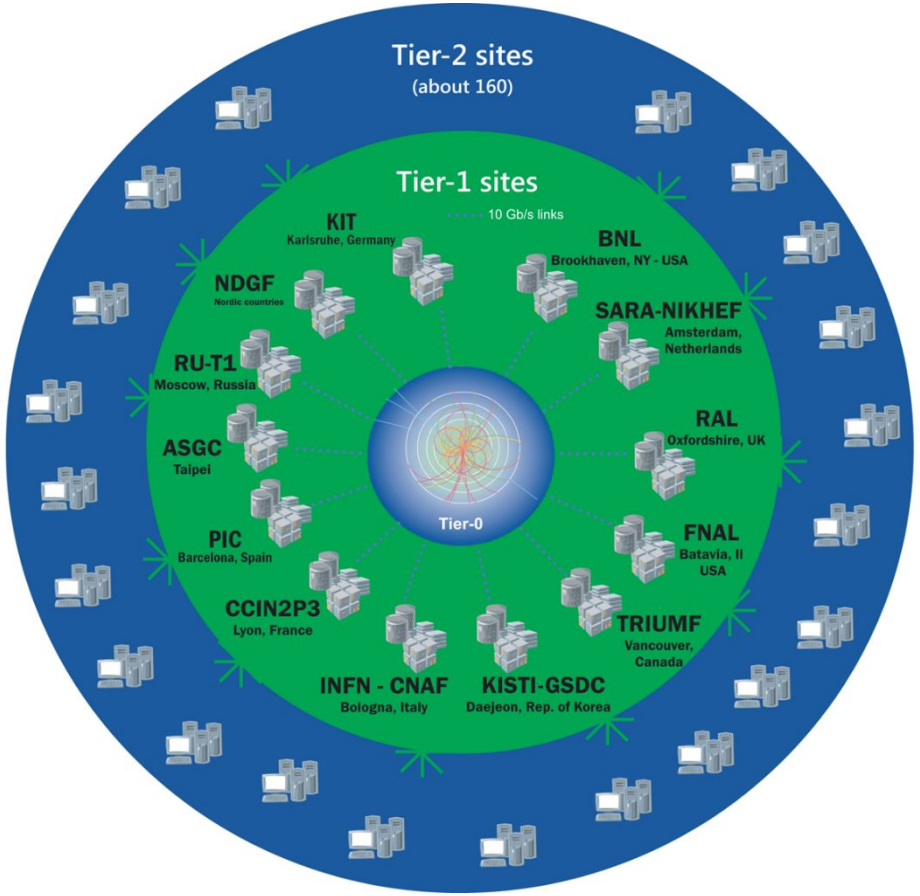
Volume



CERN

Copyright © Econophysica 2019. All Rights Reserved

Volume



CERN — Rucio

Более350 PB данных, миллиарды файлов, в более чем 120 ЦОД по всему миру.

Три копии файла на разных континентах и одна на плёнке? Ok.

Стирать непопулярные файлы автоматически? Ok.

Copyright © Econophysica 2019. All Rights Reserved

Volume

CERN

Долговременное хранение на плёнке, библиотечные стримерные роботы.

Copyright © Econophysica 2019. All Rights Reserved

Volume

Wayback Machine

http://archive.org

>15 петабайт данных,

объём растёт на 20 Тб в неделю.

Copyright © Econophysica 2019. All Rights Reserved

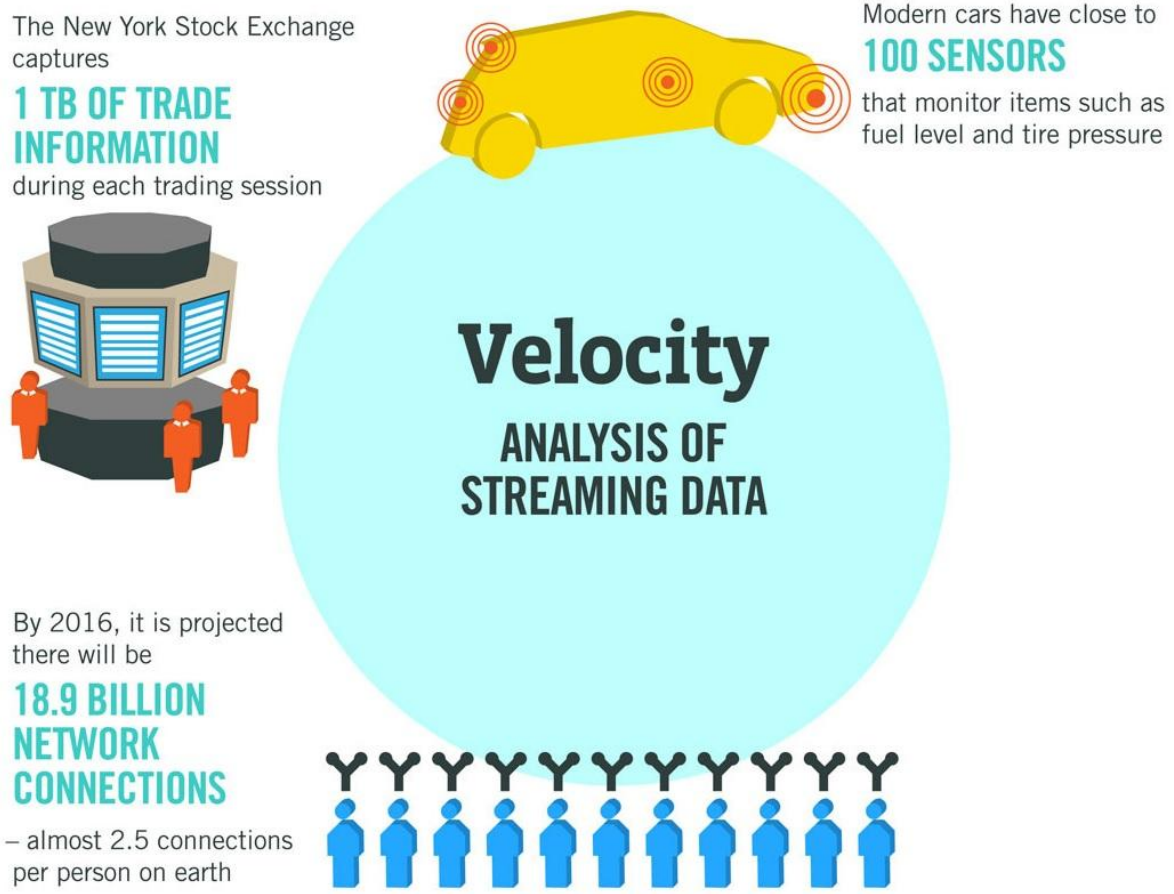
Volume

Подходы к проблеме:

- **Управление объёмом данных;**
- Вложения в хранилища;
- Разработка специализированного ПО и АО;
- **Использование специализированного ПО от других разработчиков.**

Copyright © Econophysica 2019. All Rights Reserved

Velocity



Copyright © Econophysica 2019. All Rights Reserved

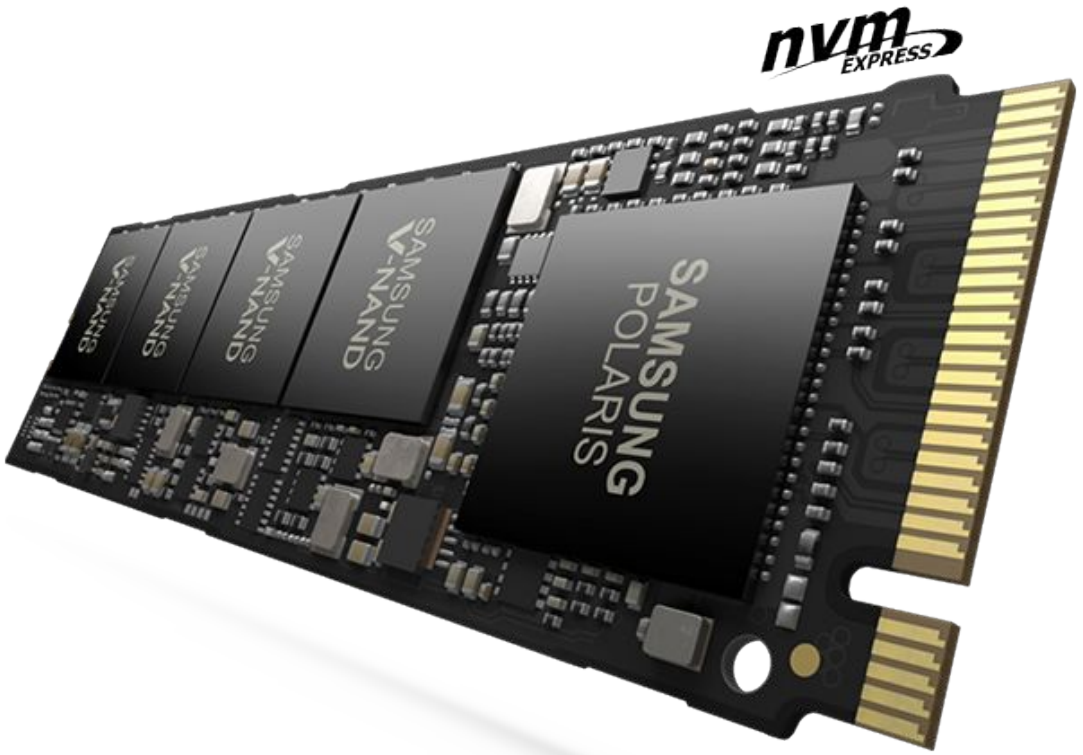
Velocity

Возможное решение проблемы:

Copyright © Econophysica 2019. All Rights Reserved

Velocity

Причина рождения «больших данных»:



Copyright © Econophysica 2019. All Rights Reserved

econophysica

Velocity

CERN:

ATLAS выдаёт ~100 TB/c со своих датчиков.

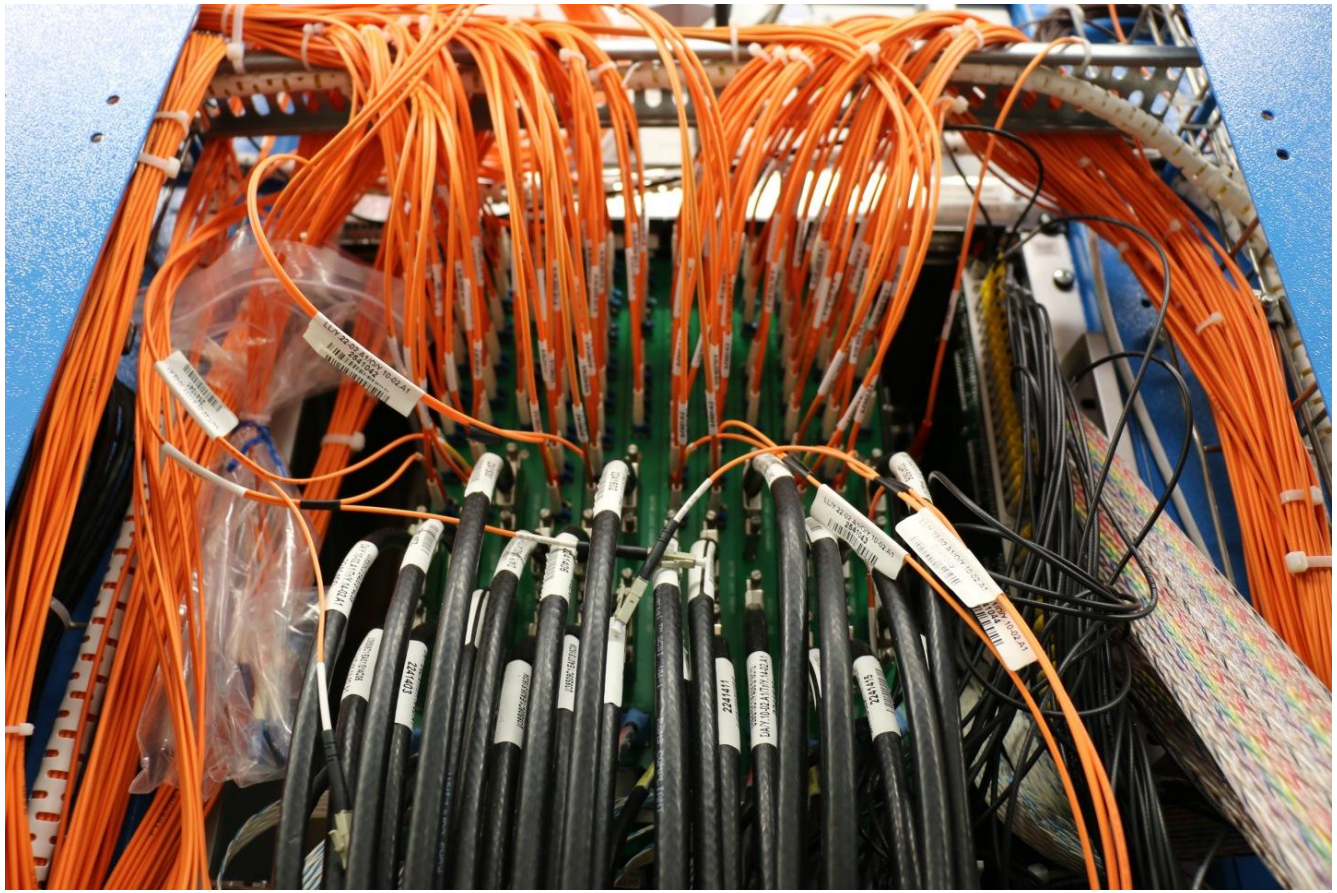
Лишь ~1 GB/c сохраняется для дальнейшей обработки.

Trigger:

The data reduction is carried out in two stages: first, **custom electronics** performs an initial level of data rejection for each bunch crossing based on partial and localized information. Only data corresponding to collisions passing this stage of selection will be actually read-out from the on-detector electronics. Then, a large computer farm **(~17 k cores)** analyses these data in real-time and decides which ones are worth being stored for Physics analysis.

Copyright © Econophysica 2019. All Rights Reserved

Velocity



CERN Trigger hardware

Copyright © Econophysica 2019. All Rights Reserved

Velocity

NYSE:

2Gbps link to allow for latency reduction, to allow for even faster high-speed trading.

Передача данных по лазеру

Copyright © Econophysica 2019. All Rights Reserved

Velocity

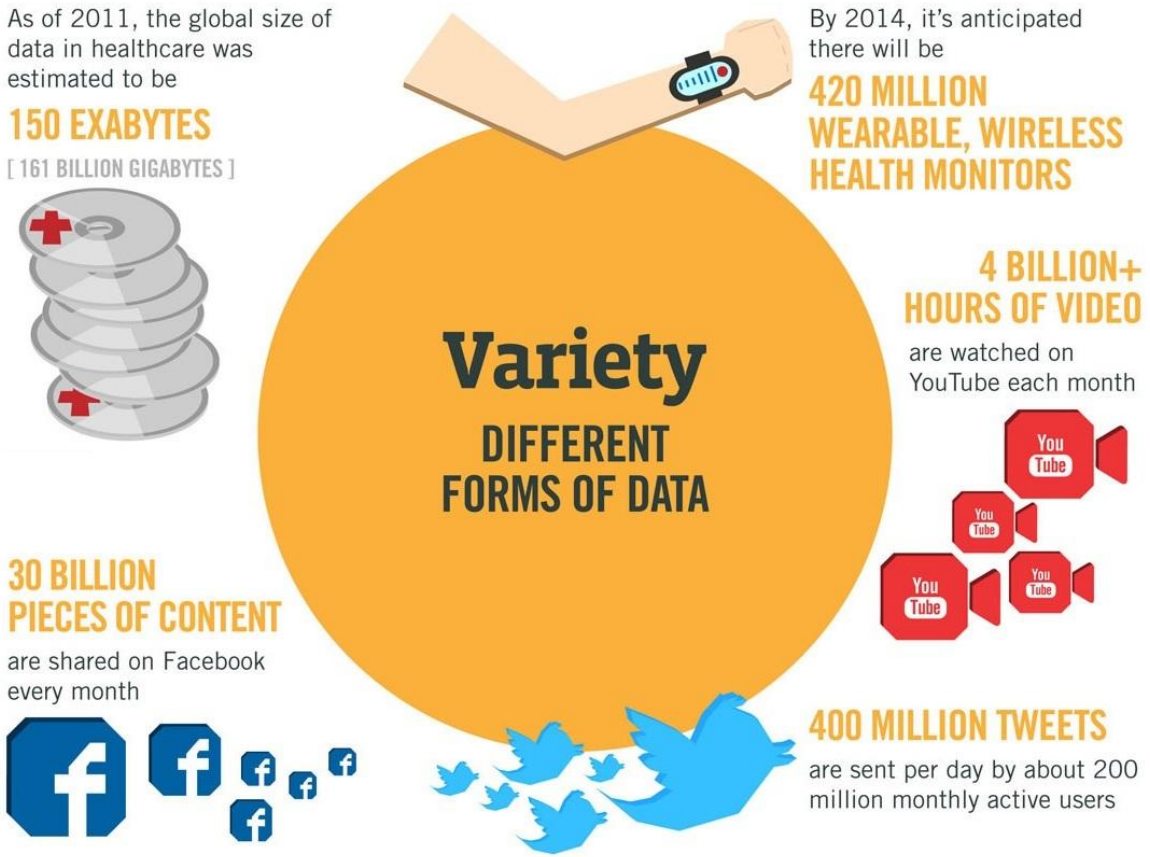
Пути решения проблемы:

Kafka, **Flume** и **Logstash** дают возможность потокового сбора данных и совместимы с множеством разных источников и приёмников данных.

Copyright © Econophysica 2019. All Rights Reserved

econophysica

Variety



Copyright © Econophysica 2019. All Rights Reserved

Variety

Около 1500 единиц общественного транспорта, оборудованных GPS	GSM
Данные о местоположении абонентов от сотовых операторов	IR; CCTV
Карты, данные аэрофотосъёмки и спутниковой съёмки	Видеорегистраторы
Поток видео с веб камер	Дорожные камеры
GPS	Спутниковое видео

Copyright © Econophysica 2019. All Rights Reserved

Variety

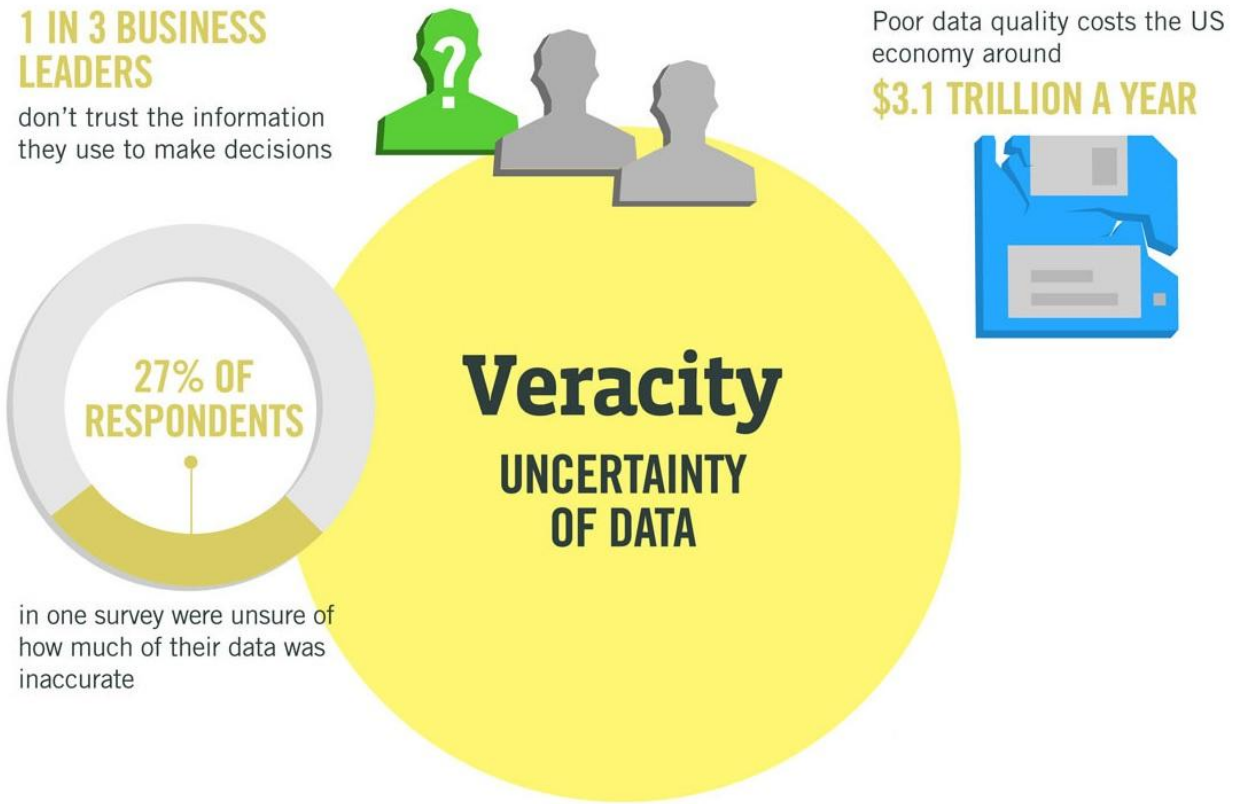
Подходы к проблеме:

- Обработка естественного языка;
- Текст-в-речь;
- Классификация изображений;
- Machine Learning-классификация всего;
- Семантические технологии, Web 2.0;
- NoSQL.

Универсального решения не существует, каждая предметная область требует своего подхода.

Copyright © Econophysica 2019. All Rights Reserved

Veracity



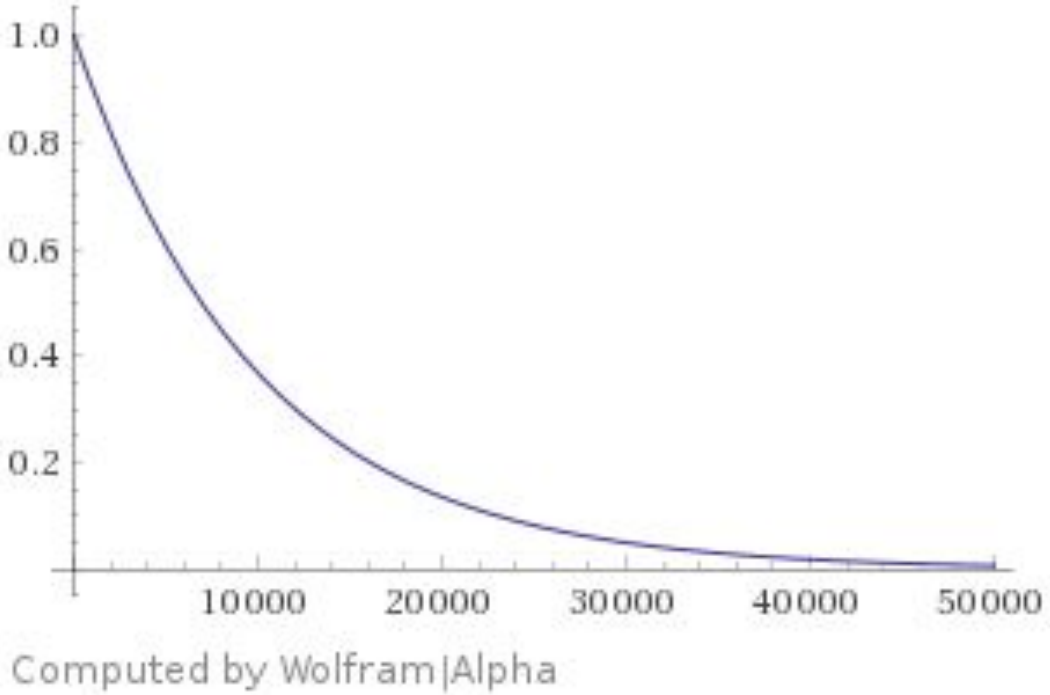
Copyright © Econophysica 2019. All Rights Reserved

Veracity

Следствие из теории вероятностей

Когда объём выборки стремится к бесконечности, вероятность ошибки в данных возрастает до определённости.

P(A and B) = P(A) * P(B)



0.9999**x

Copyright © Econophysica 2019. All Rights Reserved

Veracity: AAA

Anyone can say anything about anything at any moment.

Разрешение неопределённости:

- Проверка, либо
- Реификация.

.

Copyright © Econophysica 2019. All Rights Reserved

Спасибо за внимание!

mgubin@tpu.ru

econophysica