

# Большие данные

Технологии обработки сверхбольшого  
объёма данных («больших данных»)

*Инженерия данных*

613x-010402D      x={1,2,3}

*осень 2024*

Сергей Борисович Попов  
[sepo@ssau.ru](mailto:sepo@ssau.ru)

**Материалы лекций:**

[https://1drv.ms/f/s!ApFj4iOLPNegvEPfMZL\\_5jjkXsqfQ](https://1drv.ms/f/s!ApFj4iOLPNegvEPfMZL_5jjkXsqfQ)

# КНИГИ

Майер-Шенбергер В., Кукъер К.  
**Большие данные.**  
**Революция, которая изменит то,**  
**как мы живем, работаем и мыслим.**  
– Манн, Иванов и Фербер, 2014



**Data-Intensive Text  
Processing with MapReduce**  
Jimmy Lin and Chris Dyer  
*Draft of January 27, 2013*





## Дейтел Пол, Дейтел Харви

Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»).

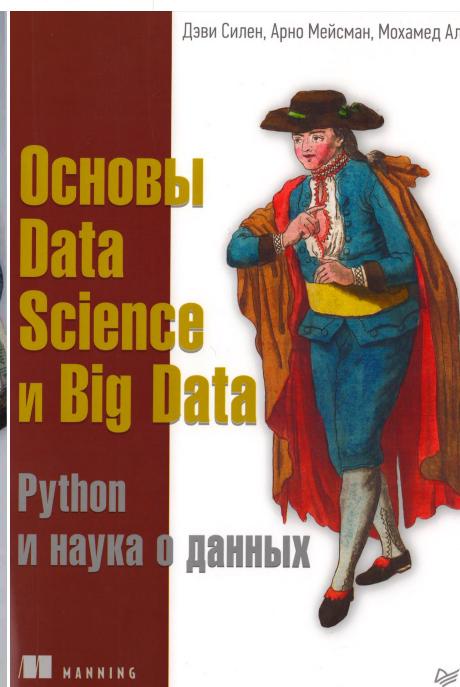
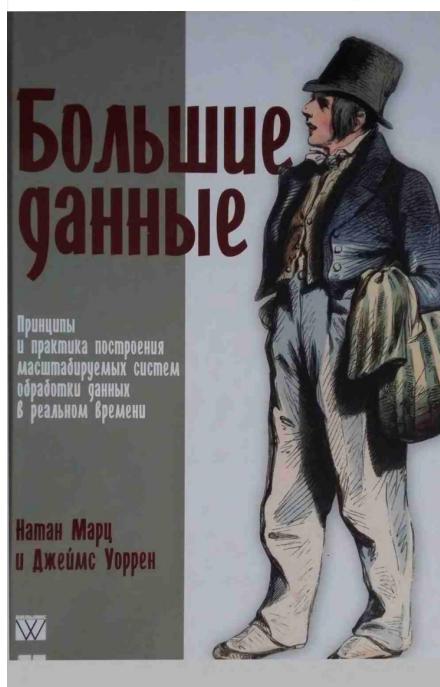
ISBN 978-5-4461-1432-0

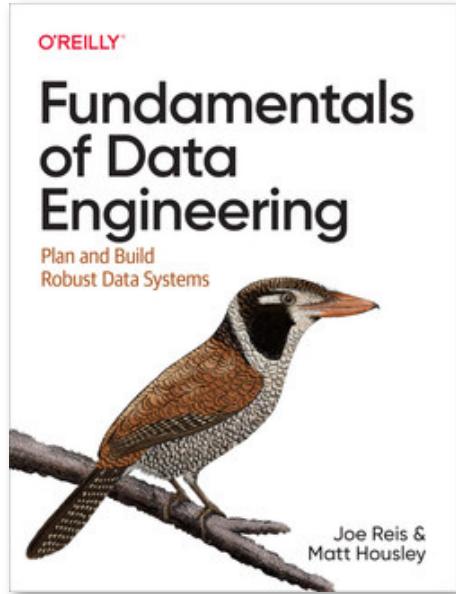
16

## Большие данные: Hadoop, Spark, NoSQL и IoT

В этой главе:

- Концепция больших данных и темпы роста.
- Работа с реляционной базой данных SQLite на языке SQL (Structured Query Language).
- Четыре основные разновидности баз данных NoSQL.
- Сохранение твитов в документной базе данных MongoDB и их визуализация на карте Folium.
- Технология Apache Hadoop и ее применение в приложениях пакетной обработки больших данных.
- Построение MapReduce-приложения на базе Hadoop в облачном сервисе Microsoft Azure HDInsight.
- Применение Apache Spark в высокопроизводительных приложениях, работающих с большими данными в реальном времени.
- Использование потоковой передачи Spark для обработки данных в формате мини-пакетов.
- «Интернет вещей» (IoT) и модель публикации/подписки.
- Публикация сообщений с моделируемого устройства, подключенного к интернету, и их визуализация на информационной панели.
- Подписка на «живой» Twitter PubPub и IoT-потоки, и визуализация данных.





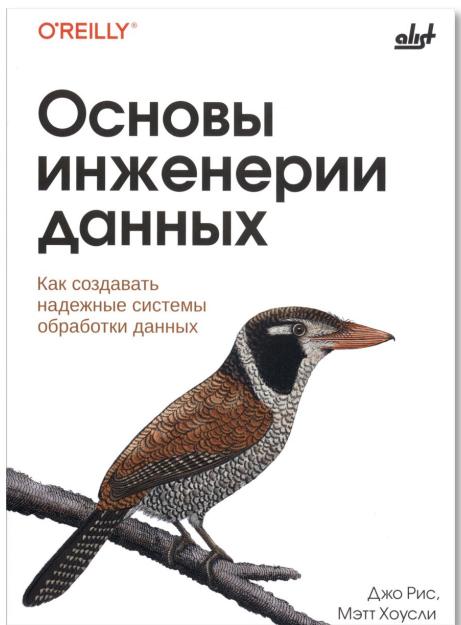
## Fundamentals of Data Engineering

by [Joe Reis, Matt Housley](#)

Released June 2022

Publisher(s): O'Reilly Media, Inc.

ISBN: 9781098108304



## Рис Дж.

Основы инженерии данных: Пер. с англ.

/ Дж. Рис, М. Хоусли – Астана: АЛИСТ, 2024. - 464 с.: ил.

ISBN 978-601-08-4116-1

FIRSTMARK

# MAD

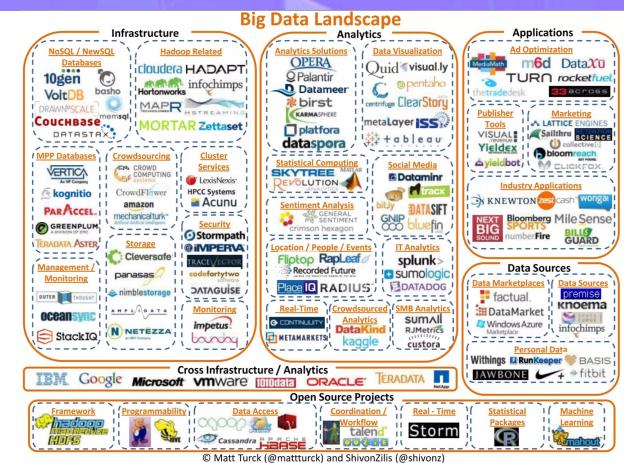
## LANDSCAPE

MACHINE LEARNING, AI & DATA

# 2024

<https://mattturck.com/category/big-data/>

2012, 2014, 2016, 2017, 2018, 2019 (Part I and Part II),  
2020, 2021 and 2023 (Part I, Part II, Part III, Part IV).



# Как идти в ногу со временем в быстро меняющейся сфере деятельности

В момент когда новая технология накатывает на вас  
как дорожный каток, если вы не часть этого катка,  
то вы становитесь частью дороги.

- Стюарт Брэнд

## Keeping Pace in a Fast-Moving Field

Once a new technology rolls over you, if you're not part of the steamroller, you're part of the road.

—Stewart Brand

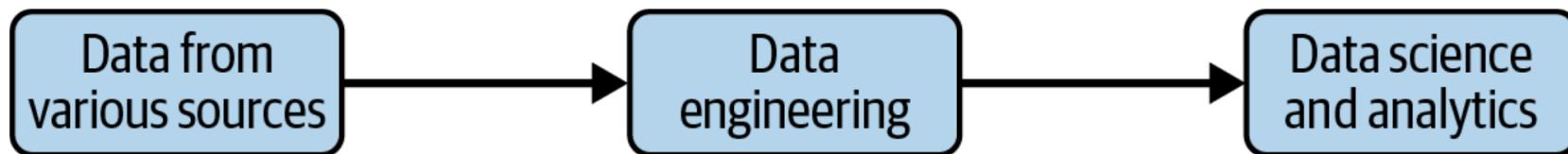
How do you keep your skills sharp in a rapidly changing field like data engineering? Should you focus on the latest tools or deep dive into fundamentals? Here's our advice: focus on the fundamentals to understand what's not going to change; pay attention to ongoing developments to know where the field is going. New paradigms and practices are introduced all the time, and it's incumbent on you to stay current. Strive to understand how new technologies will be helpful in the lifecycle.

# Определение понятия «инженерия данных»

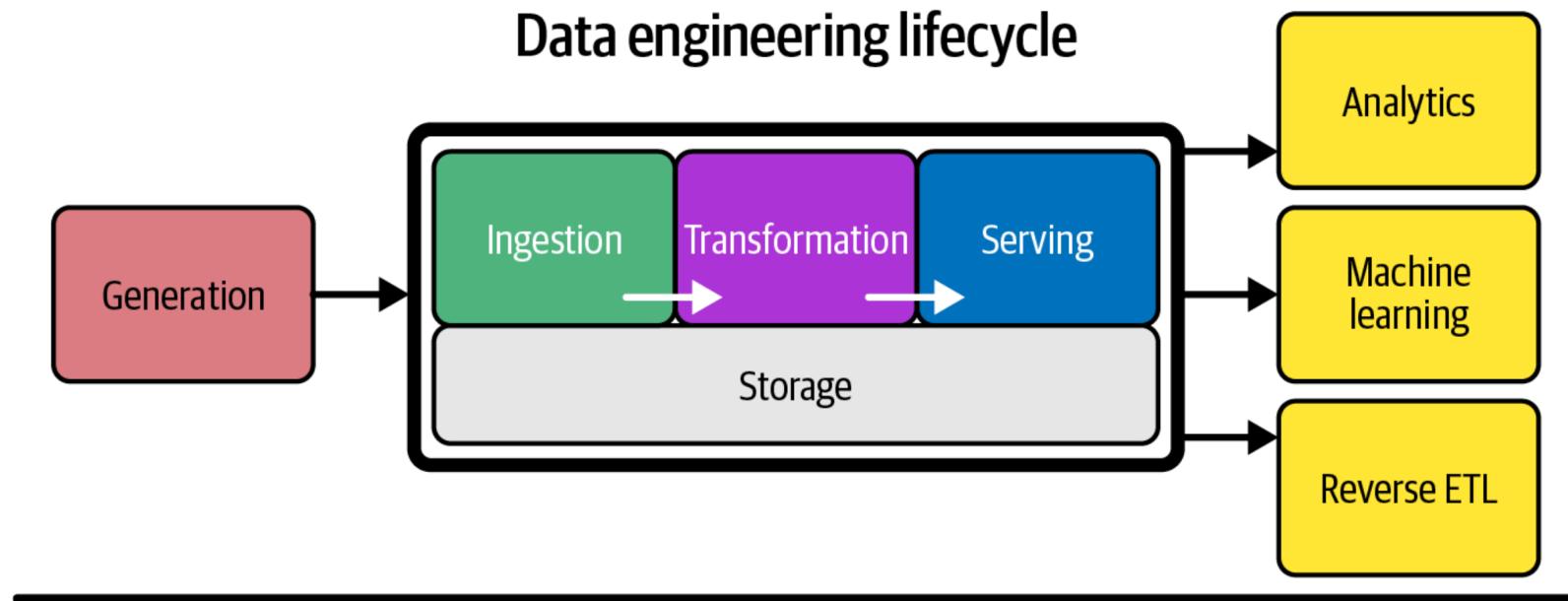
**Инженерия данных** – это *разработка, внедрение и сопровождение* систем и процессов, которые принимают необработанные данные и производят высококачественную, непротиворечивую информацию, поддерживающую последующие сценарии использования, такие как анализ и машинное обучение.

**Инженерия данных** – это пересечение безопасности, управления данными, DataOps, архитектуры данных, оркестровки и программной инженерии.

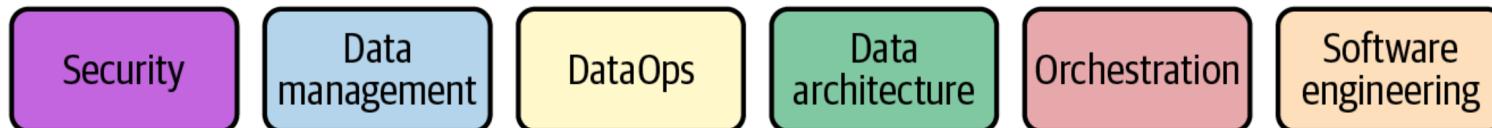
Инженер данных управляет *жизненным циклом инженерии данных*, начиная с получения данных из исходных систем и заканчивая предоставлением данных для использования в прикладных задачах, таких как анализ или машинное обучение.



# Жизненный цикл инженерии данных



## Undercurrents:



- Генерация
- Хранение
- Приём
- Преобразование
- Предоставление

- Аналитика
- Машинное обучение
- Обратное ETL

## Фоновые процессы:

- Безопасность
- Управление данными
- DataOps
- Архитектура данных
- Оркестровка
- Программная инженерия

# Иерархия потребностей в науке о данных

## THE DATA SCIENCE **HIERARCHY OF NEEDS**

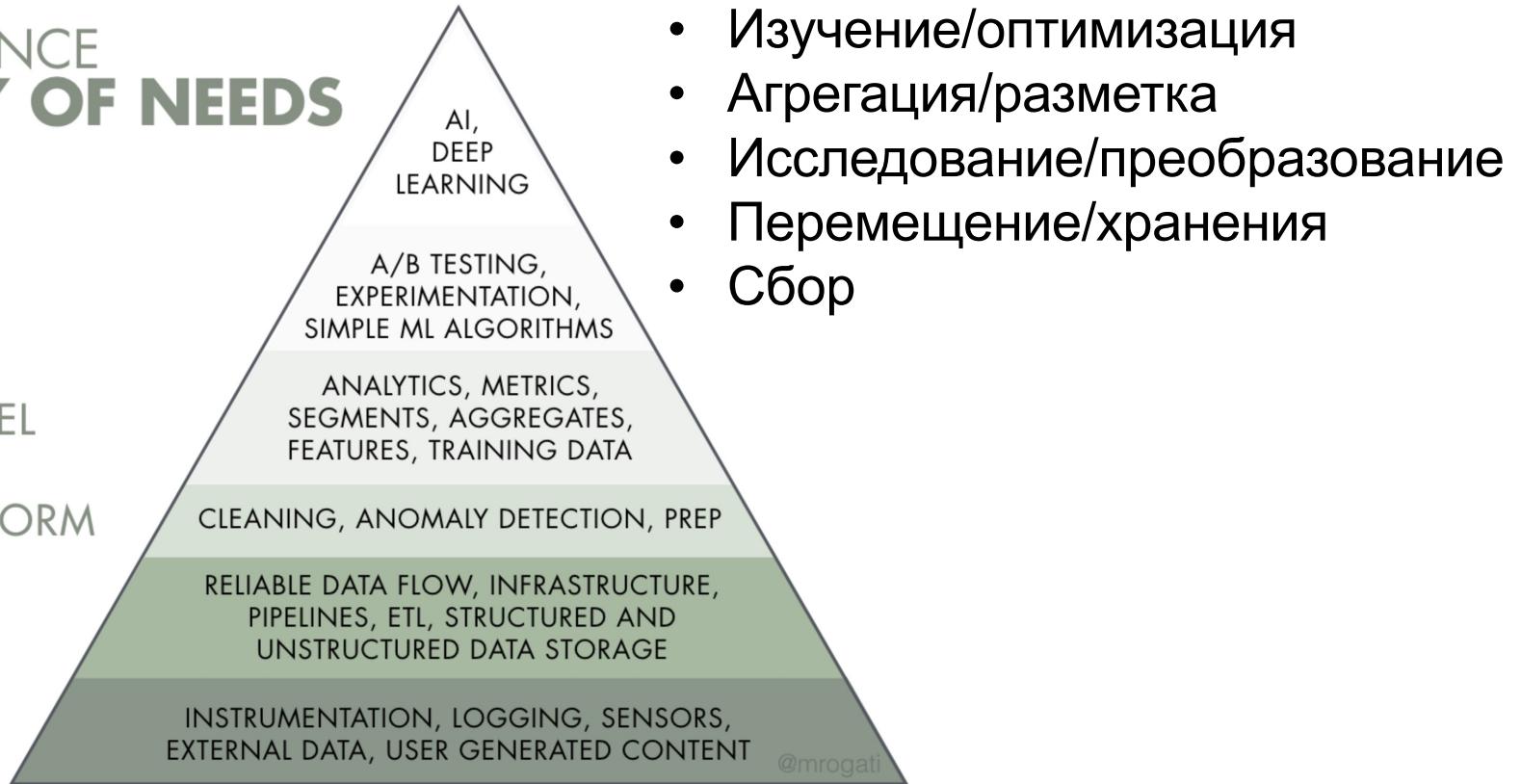
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

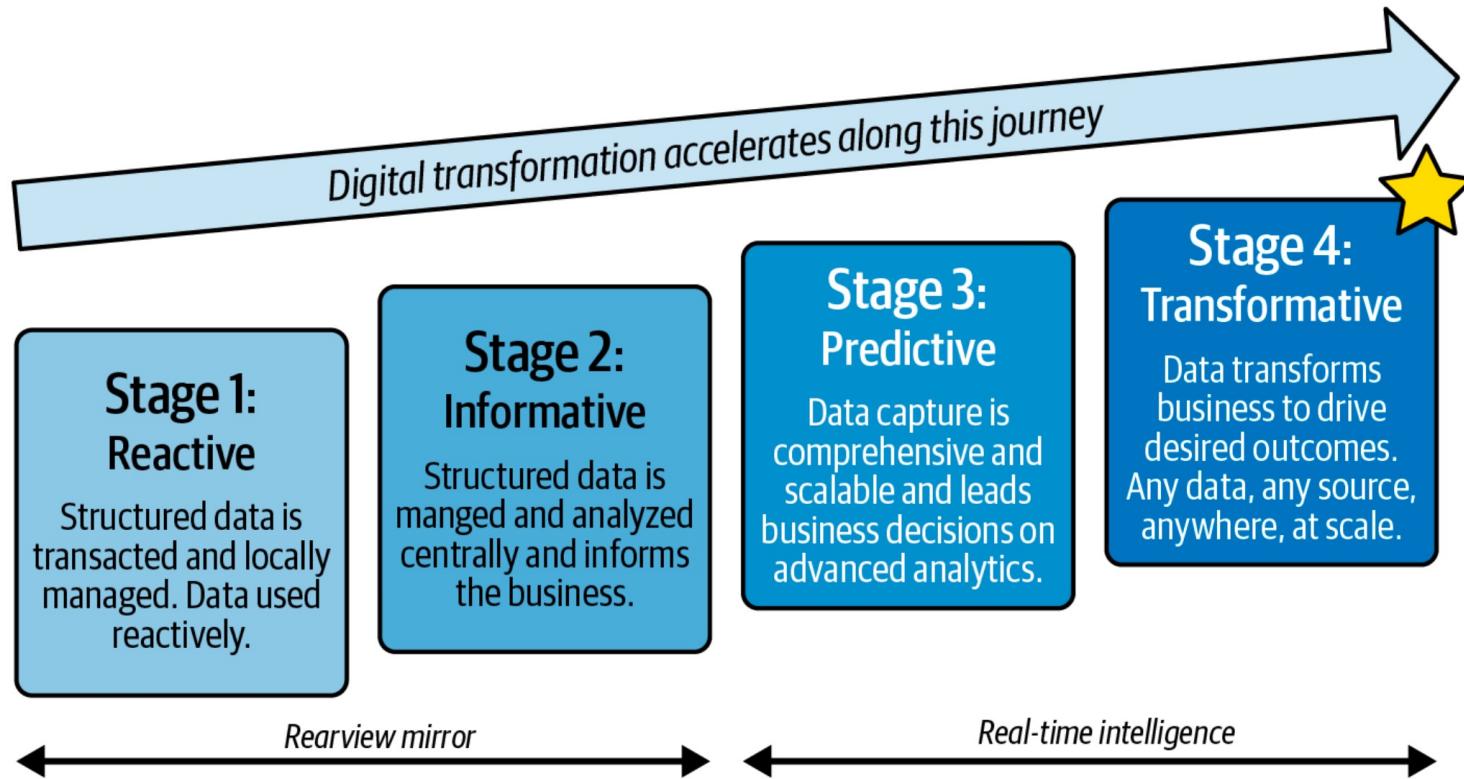
COLLECT



- Изучение/оптимизация
- Агрегация/разметка
- Исследование/преобразование
- Перемещение/хранения
- Сбор

- Искусственный интеллект, глубокое обучение
- Эксперименты A/B-тестирования, простые алгоритмы машинного обучения
- Аналитика, метрики, сегменты, агрегаты, характеристики, обучающие данные
- Очистка, обнаружение аномалий, подготовка
- Надёжная организация процесса обработки данных, инфраструктура, конвейеры, ETL, хранение структурированных и неструктурных данных
- Инструментарий, журналирование, датчики, внешние данные, содержимое от пользователя

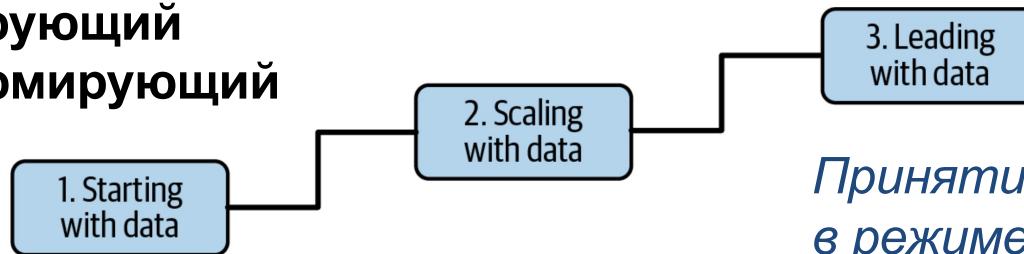
# Этапы зрелости корпоративных данных



## *Реагирование на ситуацию в прошлом:*

# Этап 1: Реагирующий

## Этап 2: Информирующий



## *Принятие решений в режиме реального времени:*

## **Этап 3: Предсказательный**

## **Этап 4: Преобразующий**

# Закон Паркинсона для данных

"Data expands to fill the  
space available for storage."

- I.A. Tjomsland в своём докладе "Where Do We Go From Here?" сформулировал Закон Паркинсона применительно к индустрии устройств хранения данных (Апрель 1980)

# The Problem of Big Data (Июль 1997)



The term "big data" was used for the first time in an article by NASA researchers Michael Cox and David Ellsworth. The pair claimed that the rise of data was becoming an issue for current computer systems. This was also known as the "problem of big data".

# Doug Laney: Проблемы технологий обработки данных (Февраль 2001)

Application Delivery Strategies



Date: 6 February 2001

File: 949

Author: Doug Laney

**3D Data Management: Controlling Data Volume, Velocity, and Variety.** Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

*META Trend: During 2001/02, leading enterprises will increasingly use a centralized data warehouse to define a common business vocabulary that improves internal and external collaboration. Through 2003/04, data quality and integration woes will be tempered by data profiling technologies (for generating metadata, consolidated schemas, and integration logic) and information logistics agents. By 2005/06, data, document, and knowledge management will coalesce, driven by schema-agnostic indexing strategies and portal maturity.*

**Figure 1 — Data Management Solutions**

▲ **Volume**

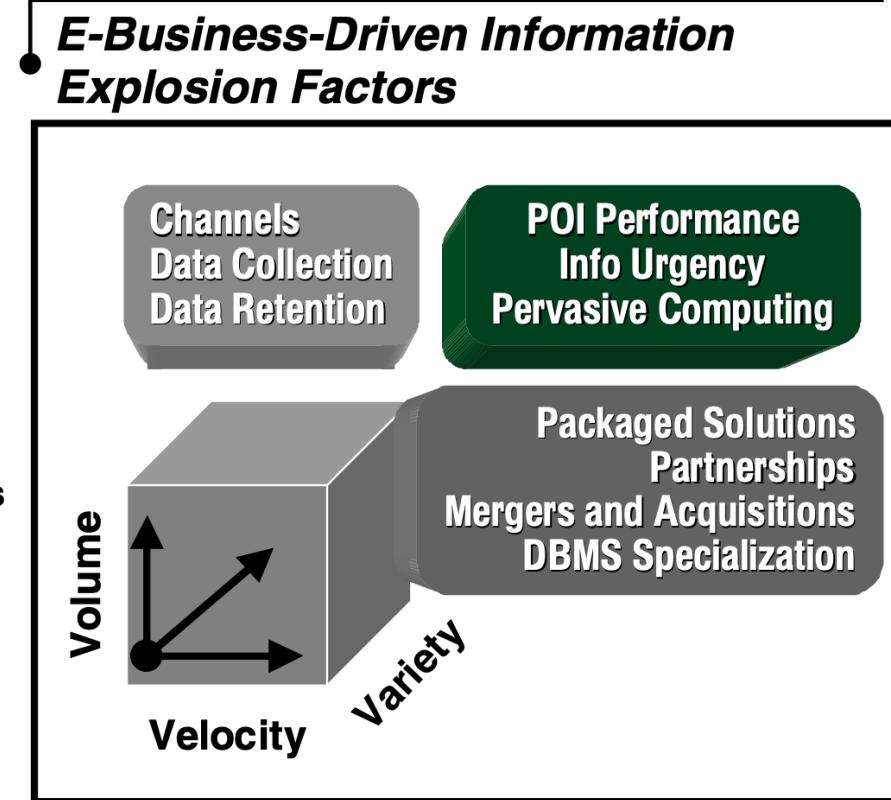
- Tiered storage/hub and spoke
- Selective data retention
- Statistical sampling
- Redundancy elimination
- Offload “cold” data
- Outsourcing

▲ **Velocity**

- Operational data stores
- Data caches
- Point-to-point data routing
- Balance data latency with decision cycles

▲ **Variety**

- Inconsistency resolution
- XML-based “universal” translation
- Application-aware EAI adapters
- Data access middleware and ETLM
- Distributed query management
- Metadata management



***Extending data management options enables greater returns on information assets***

Source: META Group

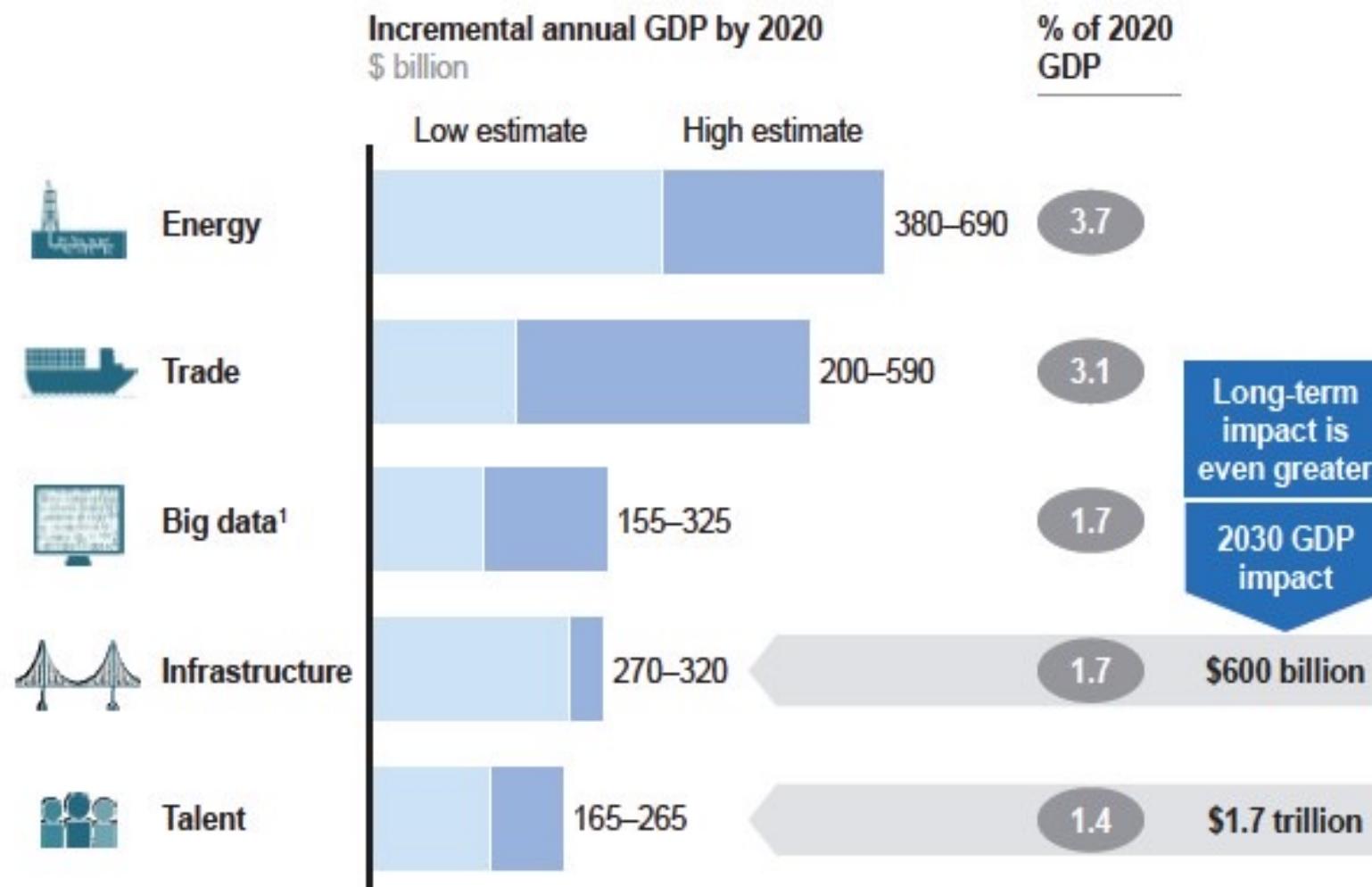
# Термин Big Data (Сентябрь 2008)



- Клиффорд Линч, редактор журнала Nature, 3 сентября 2008 года.
- Специальный выпуск по теме “что могут значить для современной науки наборы больших данных”.
- Ассоциации с качественной оценкой: «Большая руда», «Большая нефть»

# Big Data как источник прибавочной стоимости

## Главные источники прироста ВВП в США

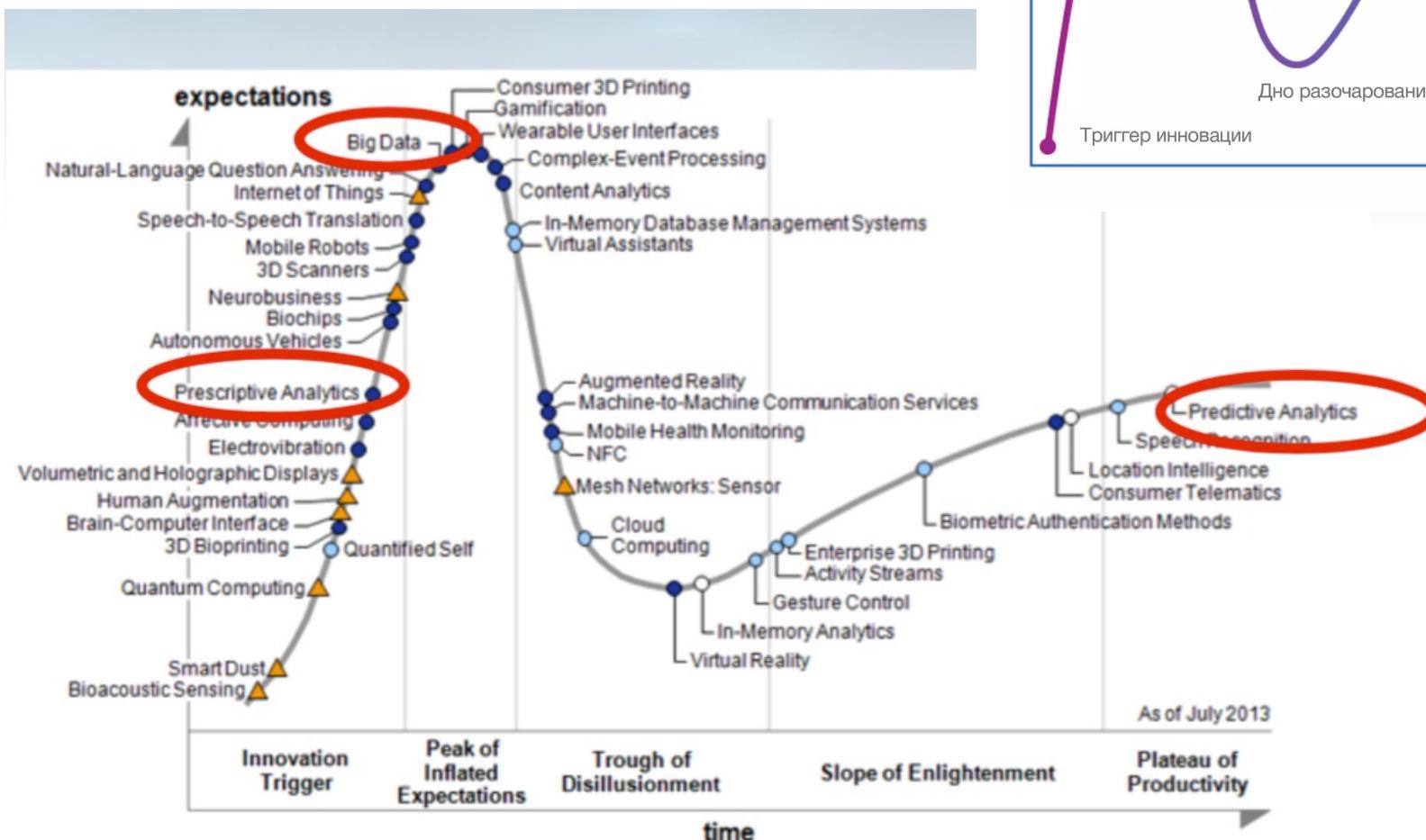


# Большие данные – феномен (2012)

- Данах Байд и Кэт Крауфорд сформулировали определение Big Data как технологического, научного и культурного феномена, включающего в себя :
  - (1) **Технология**: максимизация вычислительной мощности и сложности алгоритмов для сбора, анализа, связывания и сравнения огромных наборов данных.
  - (2) **Анализ**: представление и понимание огромных наборов данных, чтобы идентифицировать паттерны для формирования экономические, социальных, технических и юридических утверждений.
  - (3) **Мифология**: всеобщая уверенность, что огромные наборы данных представляют более высокую форму знаний и сведений, которые могут генерировать озарения, которые ранее были невозможны, с ореолом верности, объективности и точности.

Gartner's «Hype cycle»

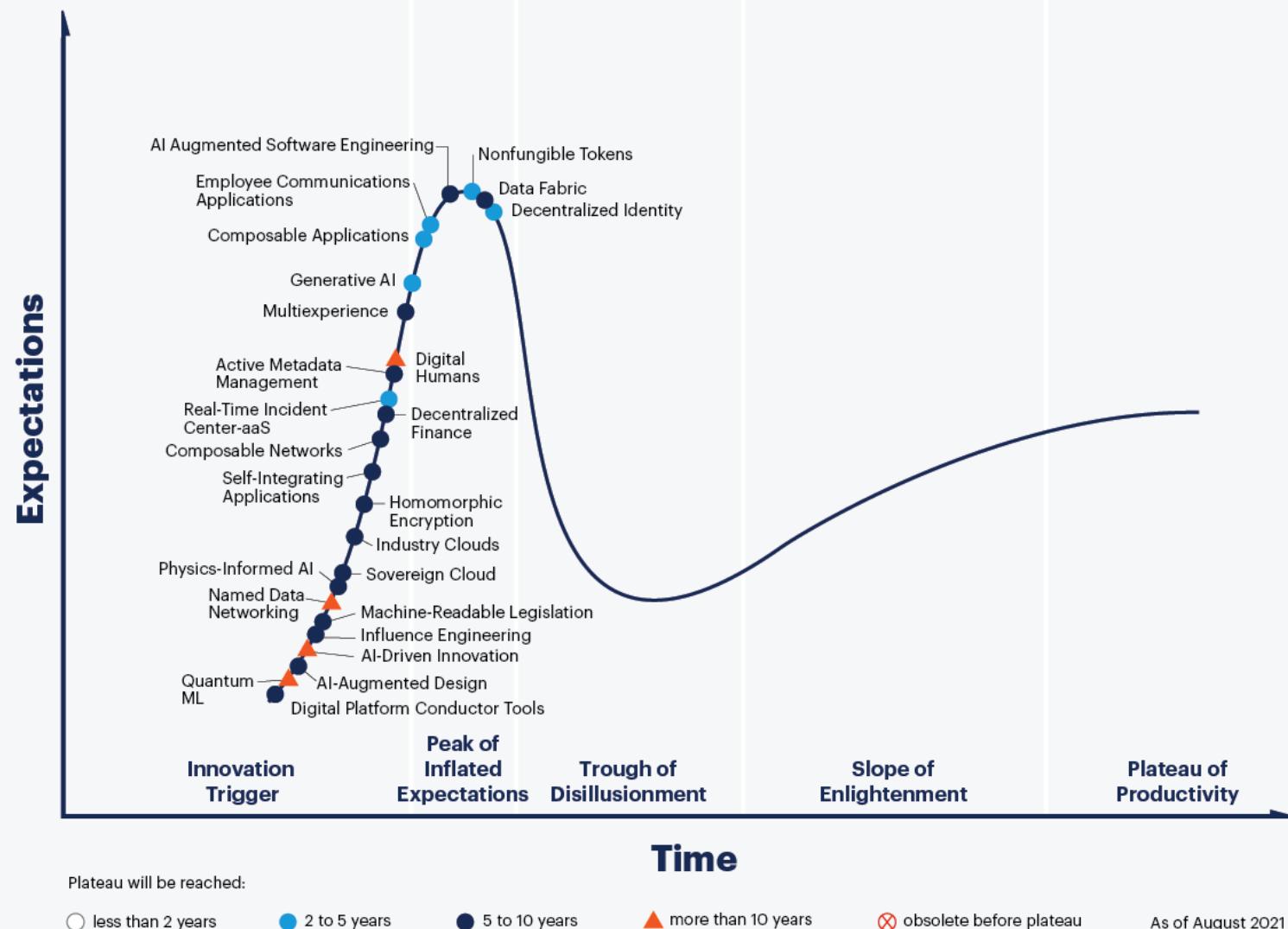
# Gartner Hype cycle 2013



Plateau will be reached in:

- less than 2 years
- 2 to 5 years
- 5 to 10 years
- more than 10 years
- obsolete before plateau

# Hype Cycle for Emerging Technologies, 2021

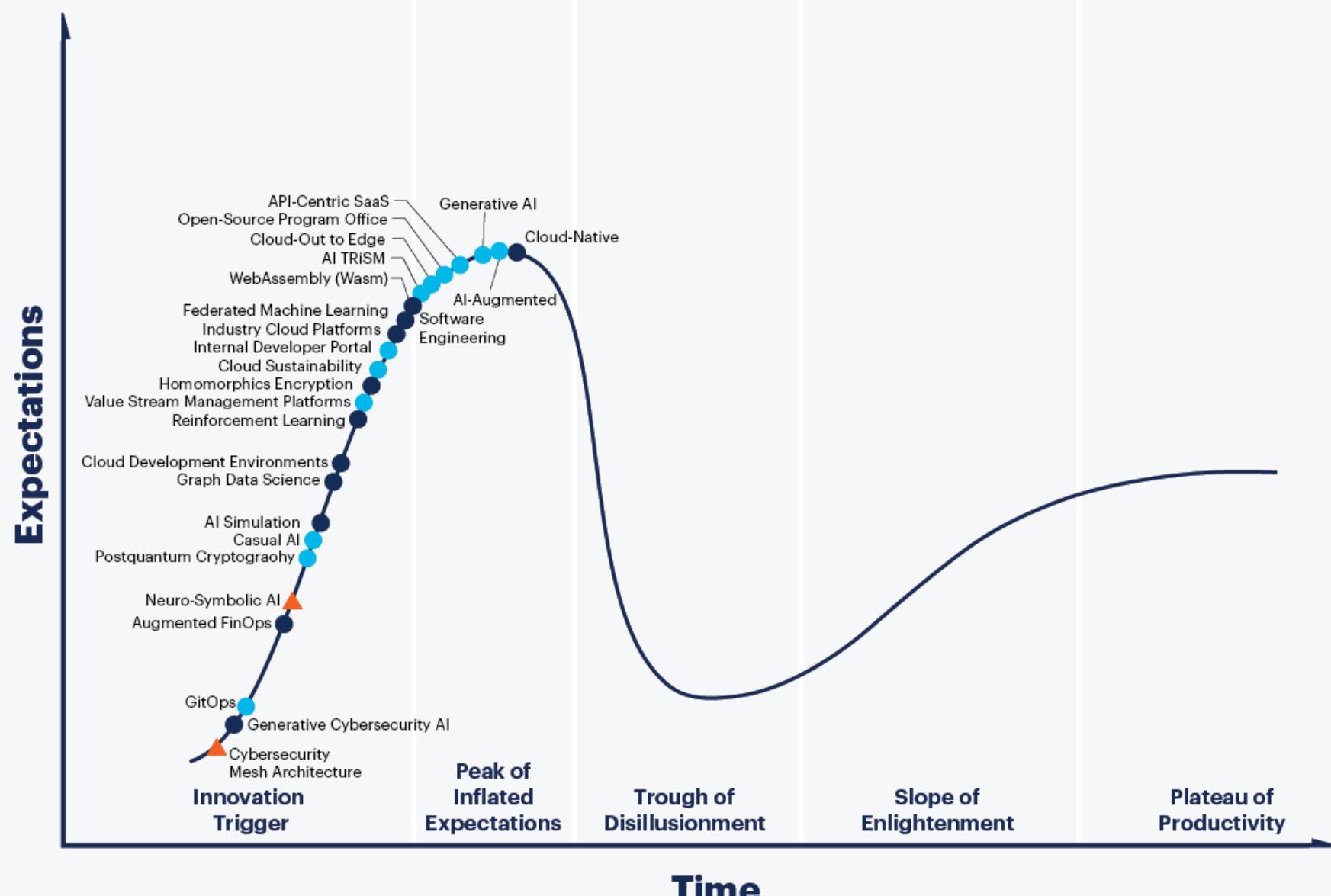


[gartner.com](http://gartner.com)

Source: Gartner  
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1448000

Gartner®

# Hype Cycle for Emerging Technologies, 2023



Plateau will be reached:

○ less than 2 years

● 2 to 5 years

● 5 to 10 years

▲ more than 10 years

✗ obsolete before plateau

As of August 2023

# Big Data Landscape

## Log Data Apps



## Vertical Apps



## Business Intelligence



## Analytics and Visualization



## Data Providers



## Analytics Infrastructure



## Operational Infrastructure



## Infrastructure As A Service



## Structured Databases



## Technologies



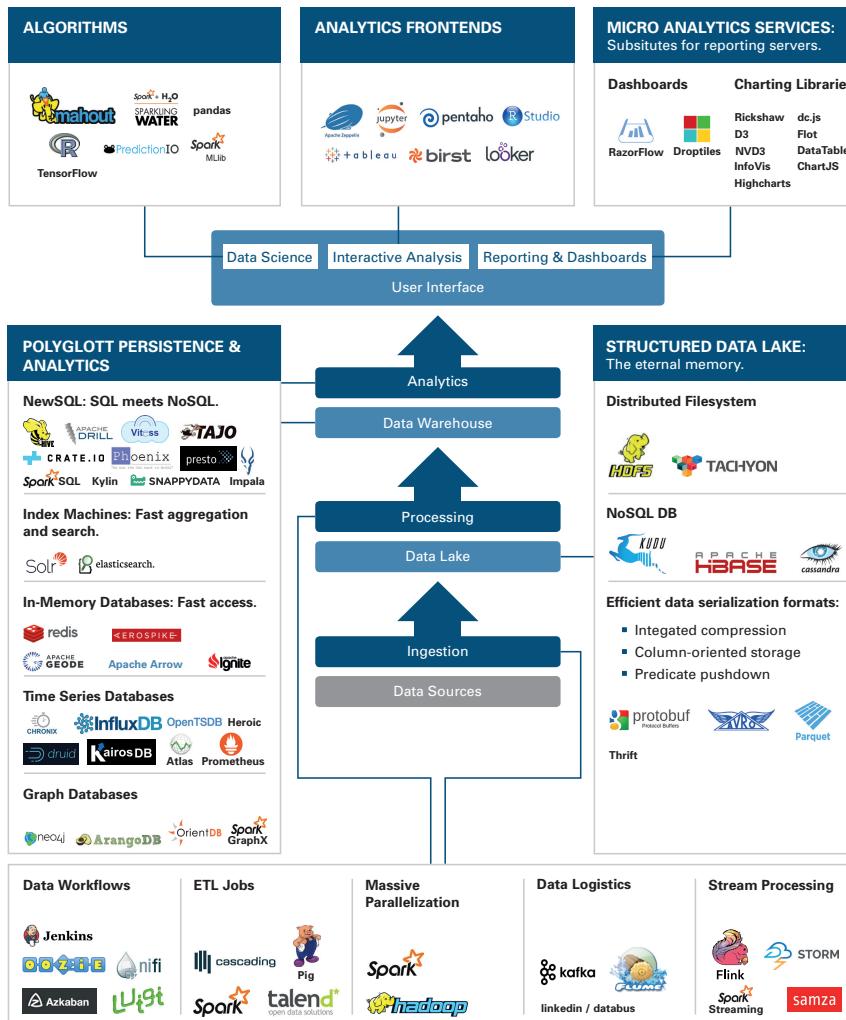
[www.bigdatalandscape.com](http://www.bigdatalandscape.com)

# Big Data Landscape (Version 2.0)



## Big Data Landscape 2016

For more big data know-how see:  
[qaware.de/news/big-data-landscape](http://qaware.de/news/big-data-landscape)

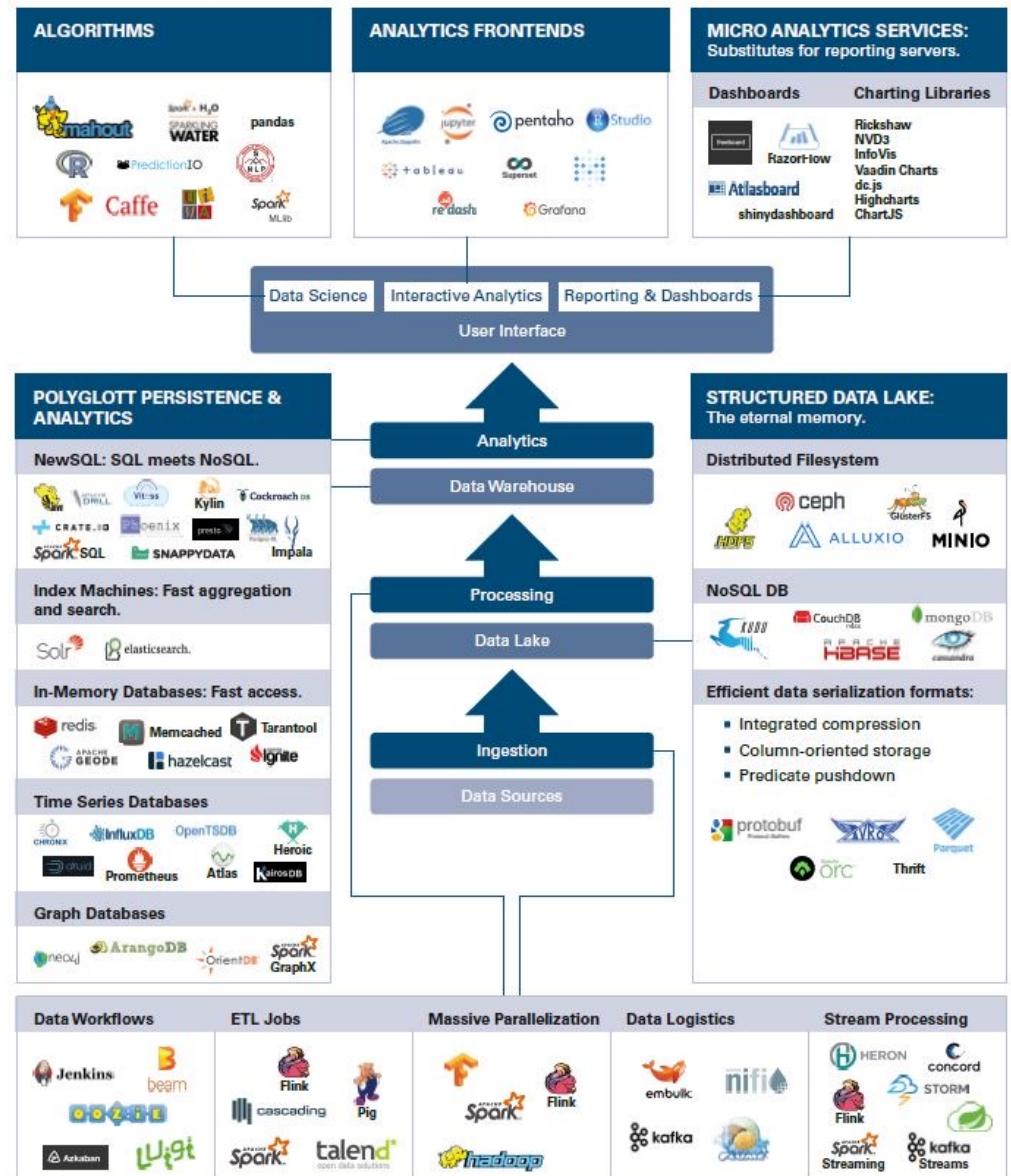


QAware GmbH info@qaware.de  
+49 (0) 89 23 23 15 - 0 qaware.de



## Big Data Landscape 2019

For more big data know-how see:  
[qaware.de/news/big-data-landscape](http://qaware.de/news/big-data-landscape)

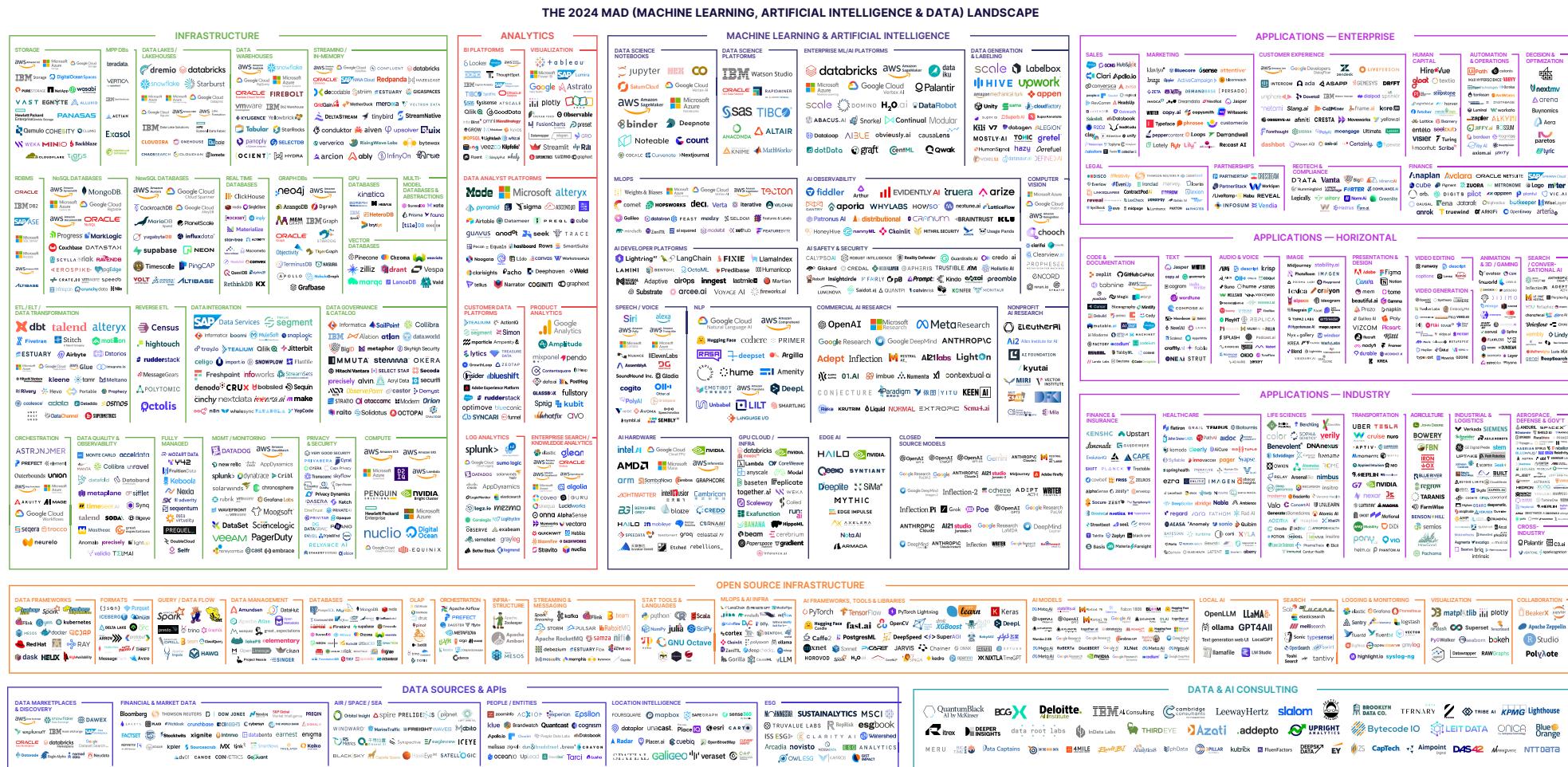


QAware GmbH  
+49 89 232315-0  
info@qaware.de  
qaware.de

twitter.com/qaware  
xing.com/companies/qwaregmbh  
linkedin.com/company/qaware-gmbh  
slideshare.net/qaware  
github.com/qware  
youtube.com/qawaregmbh



# Ландшафт 2024 экосистемы данных&ML&AI



Version 1.0 - March 2024

© Matt Turek (@mettturek), Amen Keheer (@AmenKeheer11), & F

Blog post: metteturk.com/MAP2034

Interactive version: MAD.firetmerkeen.com      Comments? Email MAD2024@firetmerkeen.com

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

# Supercomputing vs. Big Data

- Суперкомпьютеры – решение уникальных задач с экстремально высоким объёмом вычислений
  - Компьютер – «супер», если его вычислительная производительность **значительно превосходит** большинство компьютеров
  - Граница пересматривается раз в полгода – **Top500**
- 
- Суперкомпьютер – это любой компьютер, который весит более тонны
  - Большие данные – неприемлемо долгая обработка обычными программными средствами
  - Размер (в 2012 году) – от десятков ТБ до ПБ ( $2^{50}$ ), на настоящий момент счёт идет на ZB ( $2^{70}$ ) и на пороге уже YB ( $2^{80}$ ).
  - Данные – «большие», если они измеряются в мегаваттах (МВт)

# Определения «Больших данных» – Big Data

Как характеристика  
данных ( 3V [Gartner,  
2001]):

- **Volume** (объём)
- **Velocity** (скорость поступления/наращивания)
- **Variety** (разнообразие)

позднейшие дополнения  
( + 2V):

- **Veracity** (достоверность)
- **Value** (ценность или смысл)

Как характеристика технологии:

- Технологии Big Data – это технологии обработки информации, которые применяются тогда, **когда традиционные технологии** обработки на базе реляционных баз данных **не применимы** для решения стоящих задач

- Большие данные объединяют техники и технологии, которые **извлекают смысл** из данных на экстремальном пределе практичности. [Forrester]

- Технологии Big Data – это технологии преобразования информации в знание

# Big Data – технологии получения знаний

Знание = информация +

- связи, зависимости, контекст
  - история изменения (во времени)
  - модель (получения и использования)
- 
- Главное отличие знаний от данных состоит в их структурности и активности, появление в базе новых фактов или установление новых связей может стать источником формирования новых знаний, а, следовательно, изменений в принятии решений.
  - Для принятия решения необходимы знания, а не информация.

# **Технологии, «породившие» большие данные**

- Internet 2.0 (т.е. интернет с обратной связью) и, как развитие, социальные сети (пользователь стал источником данных!)
- мобильные телекоммуникации с использованием билинга (фиксация времени и места)
- «интернет вещей»
- технологии геопозиционирования (GPS)

**Персонализация информации** – стало возможным «привязать» данные к конкретному человеку (источнику), причём с указанием времени генерации данных и местоположения источника.

- Каждая транзакция (с параметрами её генерации!) стала самоценной.
- Человек (или объект, см. интернет вещей) постоянно генерирует поток данных, которые информируют о процессе его жизнедеятельности.

# Стандартизация в области Больших данных

- NIST Big Data Interoperability Framework

Цель: эталонная архитектура  
больших данных NIST (NBDRA)

<https://www.nist.gov/itl/big-data-nist>



- ISO/IEC JTC 1, Big Data



- ГОСТ Р ИСО/МЭК 20546-2019.  
Информационные технологии.  
Большие данные. Обзор и словарь.



# NIST Big Data Interoperability Framework

**Цель:** эталонная архитектура больших данных NIST  
(NBDRA – NIST Big Data Reference Architecture)

Разработка проходила в три этапа:

Этап 3: проверка NBDRA путем создания общих приложений для больших данных через общие интерфейсы;

Этап 2: определение общих интерфейсов между компонентами NBDRA;

и

Этап 1: определение ключевых компонентов эталонной архитектуры больших данных высокого уровня, которые не зависят от технологий, инфраструктуры и поставщиков.

# NIST Big Data Interoperability Framework

Том 1, Определения

Том 2, Таксономии

Том 3, Примеры использования  
и общие требования

Том 4, Безопасность и  
конфиденциальность

Том 5, Обзор официального  
документа по  
архитектурам

Том 6, Эталонная архитектура

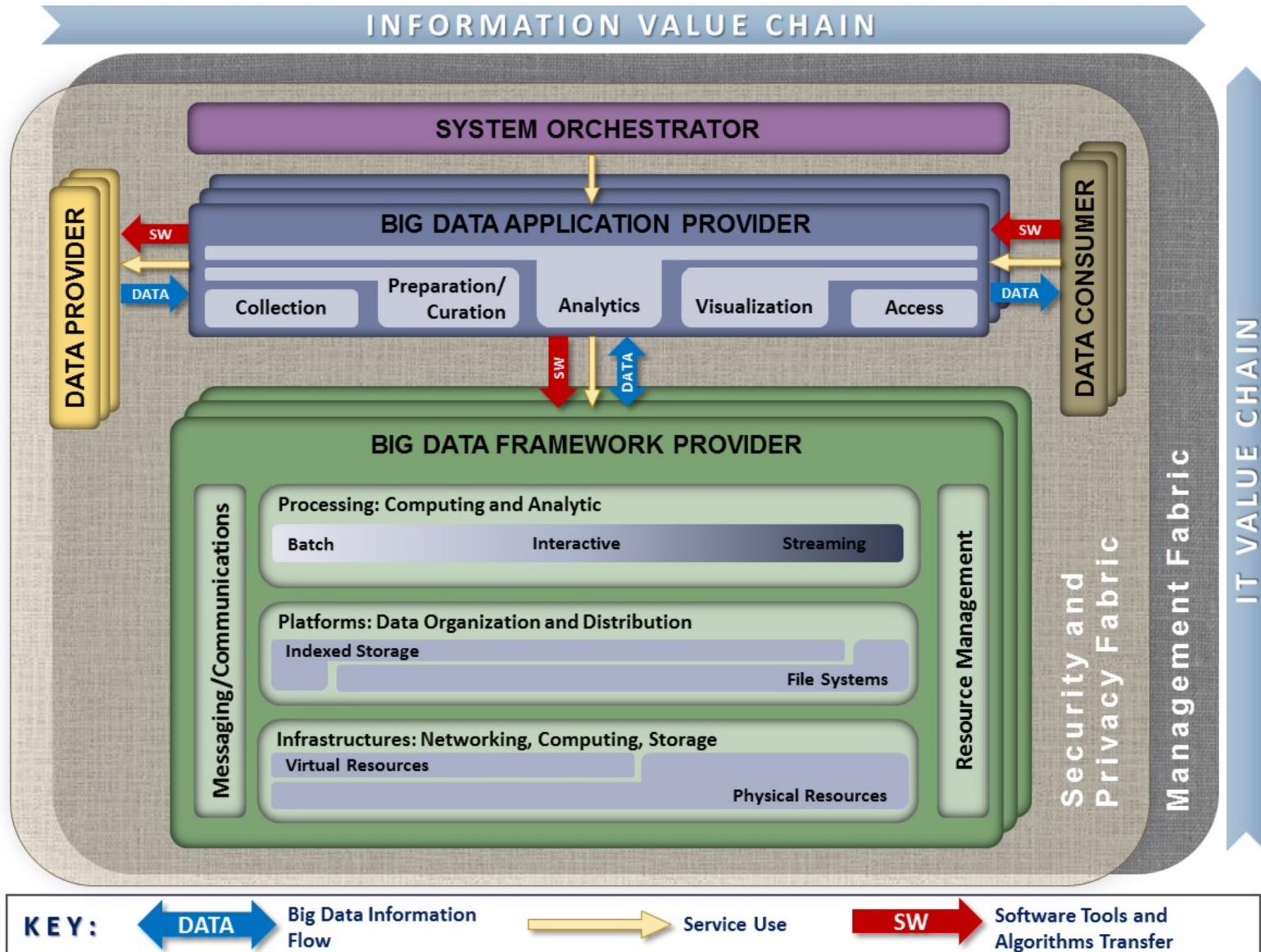
Том 7, Дорожная карта  
стандартов

Том 8, Интерфейсы эталонной  
архитектуры

Том 9, Внедрение и  
модернизация



# NIST Big Data Reference Architecture (NBDRA) Эталонная архитектура



# NIST-определения Больших данных

## 2 TERMS AND DEFINITIONS

---

***Big Data*** consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis.

***Big Data engineering*** is the discipline for engineering scalable systems for data-intensive processing.

The ***Big Data Paradigm*** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

## 3 BIG DATA CHARACTERISTICS

---

### 3.1 BIG DATA DEFINITIONS

Big Data refers to the need to parallelize the data handling in data-intensive applications. The characteristics of Big Data that force new architectures are as follows:

- ***Volume*** (i.e., the size of the dataset);
- ***Velocity*** (i.e., rate of flow);
- ***Variety*** (i.e., data from multiple repositories, domains, or types); and
- ***Variability*** (i.e., the change in velocity or structure).

# NIST-определения Больших данных

**Большие данные** состоят из больших (громадных) наборов данных – в первую очередь по характеристикам объема, разнообразия, скорости и / или изменчивости – которые требуют **масштабируемой** архитектуры для эффективного хранения, обработки и анализа.

**Инжиниринг больших данных** – это дисциплина разработки **масштабуемых** систем для обработки больших объемов данных.

**Парадигма больших данных** заключается в распределении систем данных по горизонтально связанным независимым ресурсам для достижения **масштабуемости**, необходимой для эффективной обработки обширных наборов данных.

Под **большими данными** понимается необходимость распараллелить обработку данных в приложениях с интенсивным использованием данных. Следующие характеристики больших данных вынуждают использовать новые архитектуры:

- Объем (*Volume*), т.е. размер набора данных;
- Скорость (*Velocity*), то есть скорость потока;
- Разнообразие (*Variety*), т.е. данные из нескольких репозиториев, доменов или типов; и
- Изменчивость (*Verifiability*), т.е. изменение скорости или структуры.

# Определения: Слова на «V»

- **Validity** – обоснованность (достоверность, подтверждаемость, законность) относится к пригодности данных для предполагаемого использования.
- **Value** – ценность относится к заложенной в любой набор данных присущей ему ценности, экономической и социальной (полезность, наличие знаний, использование которых приносит выгоду, обеспечивает преимущество в достижении целей).
- **Variability** – под вариативностью понимаются изменения в наборе данных, будь то скорость потока данных, формат / структура, семантика и / или качество, которые влияют на приложение аналитики.
- **Variety** – разнообразие относится к данным из нескольких репозиториев, доменов или типов.
- **Velocity** – под скоростью понимается скорость потока данных.
- **Veracity** – под достоверностью понимается точность данных.
- **Volatility** – волатильность указывает на наличие тенденции структур данных изменяться с течением времени.
- **Volume** – под объёмом понимается размер набора данных.

# Характеристики больших данных: *Volume*

Наиболее распространенной характеристикой больших данных является наличие громадных наборов данных, представляющих большой объем данных, доступных для анализа с целью извлечения ценной информации. Здесь неявно предполагается, что более значимая информация будет получена в результате обработки большего количества данных. Есть много примеров этого, известного как сетевой эффект, когда модели данных улучшаются с увеличением объема данных. Большая часть достижений машинного обучения связана с методами, которые обрабатывают больше данных. Например, распознавание объектов на изображениях значительно улучшилось, когда количество изображений, которые можно было проанализировать, увеличилось с тысяч до миллионов за счет использования масштабируемых методов. Время и затраты, необходимые для обработки массивных наборов данных, были одними из исходных драйверов распределенной обработки. Объем требует параллелизма обработки и хранения, а также управления им во время обработки больших наборов данных.

# Volume: Объём (размер) данных

Название	Размер по ГОСТ 8.417-2002 (приставки по СИ)	Символ	Примечание: размер по стандартам МЭК
байт	8 бит	B	
килобайт	$10^3$ B	KB	$2^{10} = 1024$ байт
мегабайт	$10^6$ B	MB	$2^{20}$ байт
гигабайт	$10^9$ B	GB	$2^{30}$ байт
терабайт	$10^{12}$ B	TB	$2^{40}$ байт
петабайт	$10^{15}$ B	PB	$2^{50}$ байт
эксабайт	$10^{18}$ B	EB	$2^{60}$ байт
зеттабайт	$10^{21}$ B	ZB	$2^{70}$ байт
йоттабайт	$10^{24}$ B	YB	$2^{80}$ байт

# Характеристики больших данных: Velocity

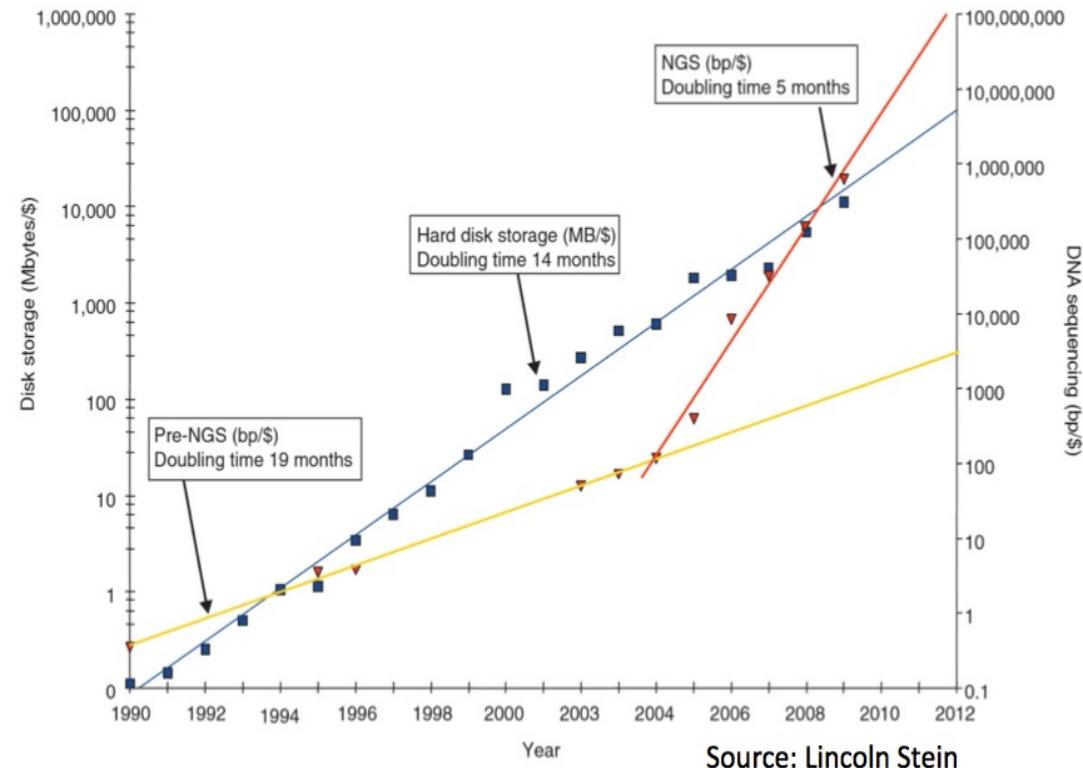
Скорость – это мера скорости потока данных. Традиционно высокоскоростные системы описываются как потоковые данные. Хотя эти аспекты являются новыми для некоторых отраслей, другие отрасли (например, телекоммуникации и транзакции с кредитными картами) обрабатывают большие объемы данных за короткие промежутки времени в течение многих лет. Потоковые данные обрабатываются и анализируются в реальном времени или почти в реальном времени, и их следует обрабатывать совершенно иначе, чем данные в состоянии покоя (т.е. постоянные данные). При обработке потоковых данных есть тенденция использовать архитектуру систем обработки событий и ориентироваться на приложения реального времени или оперативного анализа. Потребность в обработке данных в реальном времени, даже при наличии больших объемов данных, приводит к другому типу архитектуры, в которой данные не хранятся, а обычно обрабатываются в памяти. Обратите внимание, что временные ограничения для обработки в реальном времени могут создать потребность в распределенной обработке, даже если наборы данных относительно небольшие – сценарий, часто присутствующий в Интернете вещей (IoT).

# Velocity: скорость поступления/роста

Закон Мура применим к технологиям, которые производят данные.

Новые парадигмы:

- Наука интенсивной обработки данных – “data intensive science”
- Вычисления, интенсивно работающие с данными – “data intensive computing”

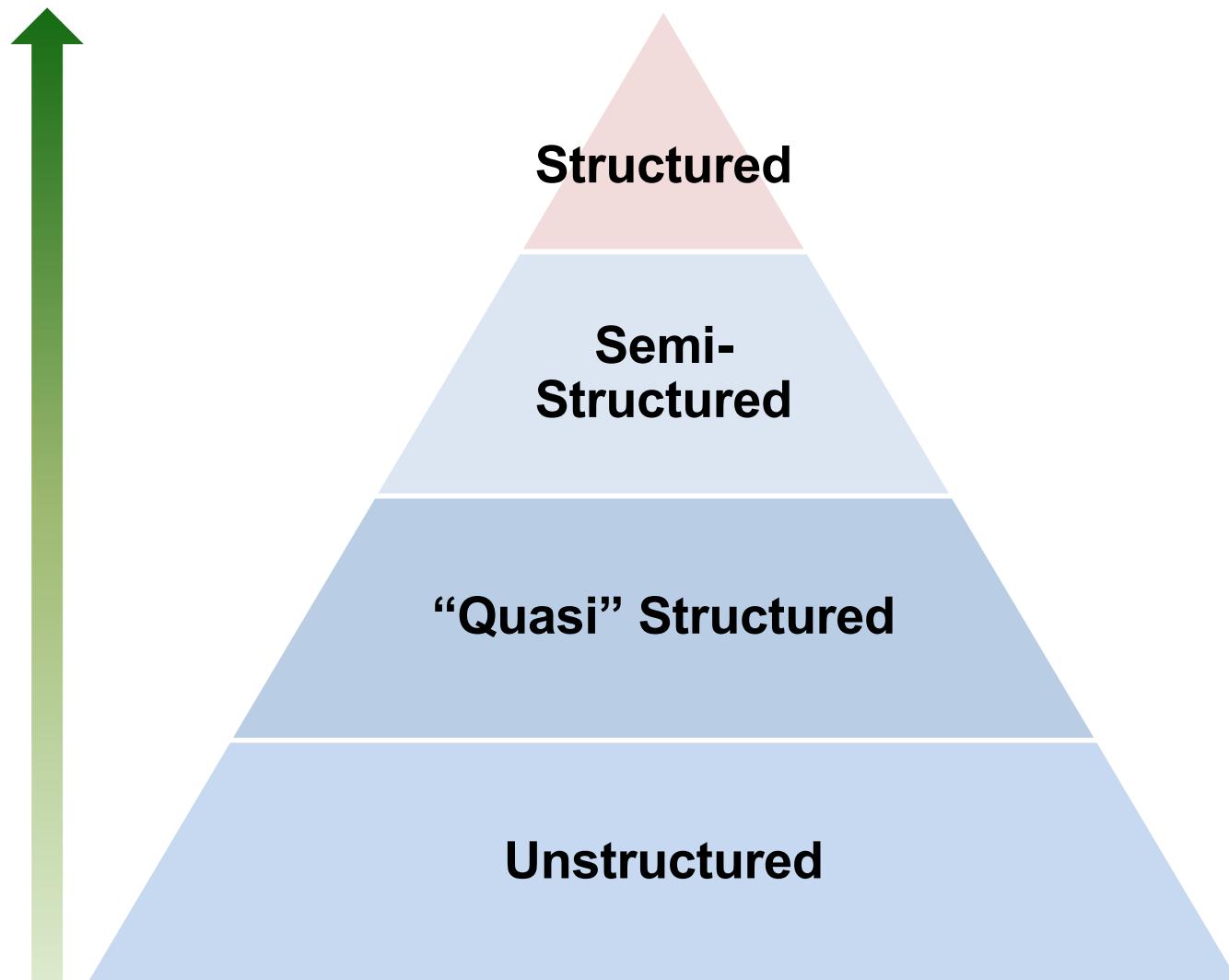


Next Generation Sequencing (NGS) – технологии секвенирования нового поколения

# Характеристики больших данных: Variety

Характеристика **разнообразия** означает необходимость анализа данных из нескольких депозитариев, доменов (областей, источников) или **типов**. Разнообразие данных из нескольких доменов ранее обрабатывалось посредством идентификации функций, которые позволяли бы согласовывать наборы данных и их объединение в хранилище данных. Автоматическое слияние данных основывается на семантических метаданных, когда понимание данных через метаданные позволяет их интегрировать. Разнообразие типов данных, доменов, логических моделей, шкал времени и семантики усложняет разработку аналитики, которая может охватывать такое разнообразие данных. Распределенная обработка позволяет проводить индивидуальную предварительную аналитику различных типов данных, за которой следует различная аналитика для охвата этих промежуточных результатов. Обратите внимание, что, хотя объем и скорость позволяют проводить более быструю и экономичную аналитику, именно разнообразие данных позволяет получать аналитические результаты, которые раньше были невозможны. «Бизнес-выгоды часто выше при рассмотрении разнообразных данных, чем при использовании больших объемов данных [M. A. Beyer and D. Laney, “The Importance of Big Data: A Definition,” Jun. 2012]».

# Variety: Разнообразие



# Variety: Разнообразие

## Structured Data

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2)
	-----Thousands-----	--Mil--	---Million \$---	
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3)	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1

## Semi-Structured Data

View → Source

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
    <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
    <META name="y_key" content="859b4020e1c9acec">
        <link rel="canonical" href="http://www.emc.com/index.htm" />
            <META NAME="verify-v1" CONTENT="vitzgVOP4eVOjFdiPeVtFtP3g4qtwFE0I2UvTmfsU"/>
            <title>EMC - Data Recovery, Cloud Computing, and Storage Hardware</title>
            <META NAME="description" CONTENT="EMC is a leading provider of storage hardware solutions that data recovery and improve cloud computing." />
            <META NAME="keywords" CONTENT="emc, network storage, data recovery, information management, software, has storage, information protection, information management" />
                <!-- Start :stylebaseet includes -->
                <link rel="stylesheet" href="/_admin/css/styles.css" />
                <link rel="stylesheet" href="/_admin/css/styles_nav.css" />
            <!--[if IE]>
```

## Quasi-Structured Data

data scientist - Google Search - Windows Internet Explorer

data scientist

data scientist

data scientist salary

data scientist skills

data scientist jobs

What is data science? - O'Reilly Radar

radar.oreilly.com/2010/06/what-is-data-science.html

Jun 2, 2010 - Editor's Pick: From the Archives — The future belongs to the companies who figure out how to collect and use data successfully

Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...

jobs.dal.com/Articles&News

Aug 10, 2011 - Data scientists are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

Data Science & The Role of the Data Scientist « Wikiblog

Mar 23, 2011 - The Role of the Data Scientist. While the concept of data science has been around for decades, the notion of a data scientist has become an ...

Career Advice: How do I become a data scientist? - Quora

www.quora.com/Career-Advice/how-do-I-become-a-data-scientist

Answer 1 of 13: Strictly speaking, there is no such thing as "data science" (see What is data science?). See also: Vardi: Science has only two legs: ...

Data Scientist Summit 2011

This May, data scientists and leaders from industry and academia came together to ... For two days this May, the data scientist community came together from ...

[http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs\\_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs\\_sm=&gs\\_upl=&bav=on.2,or.r\\_gc.r\\_pw,cf.osb&fp=d566e0fb09c8604&biw=1382&bih=651](http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw,cf.osb&fp=d566e0fb09c8604&biw=1382&bih=651)

## Unstructured Data

The Red Wheelbarrow, by  
William Carlos Williams

so much depends  
upon  
a red wheel  
barrow  
glazed with rain  
water  
beside the white  
chickens.

GREENPLUM. A DIVISION OF EMC

About Us | News & Events | Contact Us | Downloads

Products & Solutions | Media Center | Greenplum Community

LOG IN / REGISTER | Login | Register

COMMUNITY | Software Downloads | Forums | Data Scientists | Developers

©MIKETWEETS ALMANAC OF HUMAN EMOTIONS ON DISPLAY THIS MORNING, LITERALLY, JUST AMAZING. #DATASCI

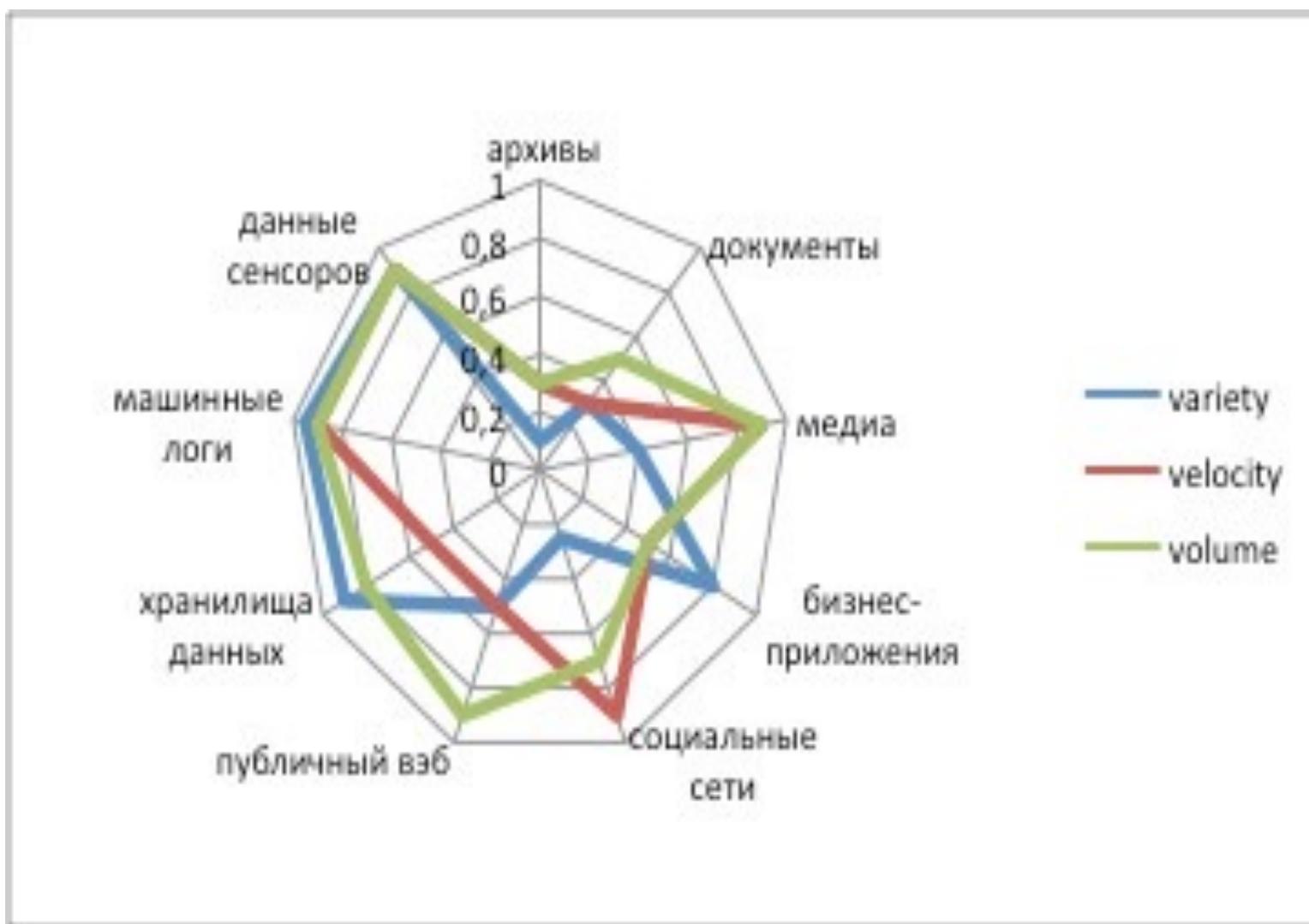
Watch an excerpt delivered by Jonathan Harris at the Data Scientist Summit: The Art and Science of Storytelling

Data Scientist Summit 2011

These 12 video presentations are from the Data Scientist Summit, a unique gathering of data pioneers, business leaders, entrepreneurs, technologists and scientists. This summer, the Data Scientist Summit will bring together the most innovative people, techniques and trends that are driving the big data opportunity for organizations across the globe. Follow Greenplum on Twitter.

DATA SCIENTIST SUMMIT 2011 Stay informed of Data Science events and news

# Большие данные от различных источников имеют различные характеристики



# Характеристики больших данных: **Variability**

Изменчивость, вариативность, вариабельность – это характеристика, немного отличающаяся от объема, скорости и разнообразия, поскольку она относится к изменению характеристик набора данных, а не к изменениям в самом наборе данных или в их потоке. Под вариативностью (вариабельностью) понимаются такие изменения в наборе данных как скорость потока данных, формат/структура и/или объем, которые влияют на его обработку. Последствия могут включать необходимость рефакторинга архитектур, интерфейсов, технологий/алгоритмов обработки, интеграции/слияния или хранения. Вариабельность объемов данных подразумевает необходимость масштабирования как в сторону увеличения, так и в сторону уменьшения виртуализированных ресурсов, чтобы эффективно справляться с дополнительной вычислительной нагрузкой, что является одним из преимуществ облачных вычислений.

Следует отметить, что обсуждаемая изменчивость относится к изменениям характеристик набора данных, тогда как термин изменчивость (*volatility*, в смысле волатильность) (раздел 5.4.3) относится к изменяющимся значениям отдельных элементов данных. Поскольку последнее не влияет на архитектуру - только на аналитику, - только изменчивость в смысле вариабельности (*variability*) влияет на архитектурный дизайн.

# **Value: Big Data Are the Next Great Natural Resource**

- Компании, которые принимают решения на "миллиарды долларов" руководствуясь "инстинктами" (gut instincts, «Нутром чувствую!»), а не "интеллектуальным анализом" больших данных, станут «проигравшими» (losers) в постоянно растущей глобальной экономике, основанной на информации.

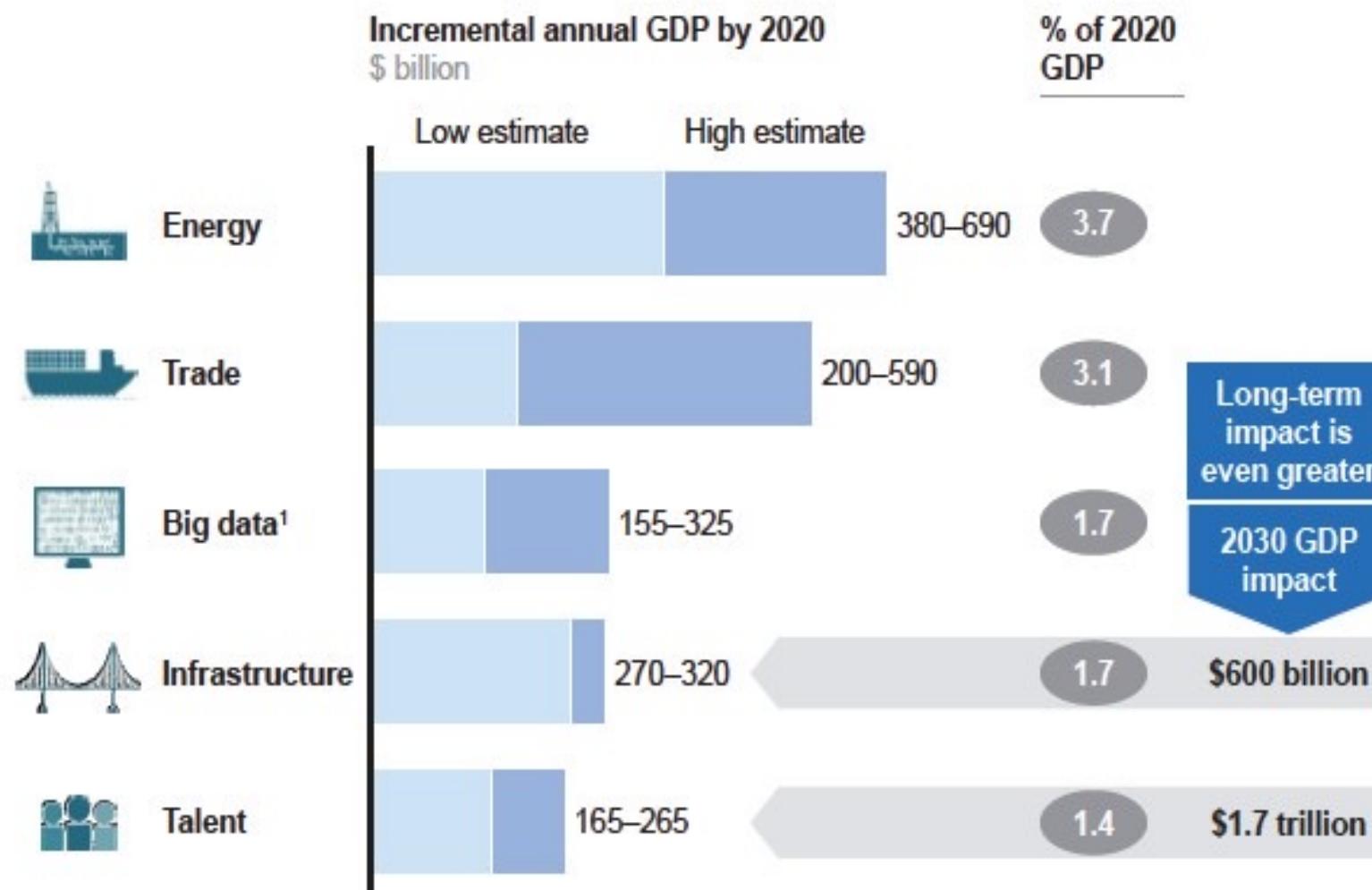
(генеральный директор IBM  
Вирджиния Рометти ("Ginni" Rometty)

<http://www.cfr.org/world/conversation-ginni-rometty/p35497>

Изменить принципы организации компании (правительства, города )

1. Как вы принимаете решения.
2. Как вы на самом деле создаёте прибавочную стоимость.
3. Как вы поставляете результаты прибавочной стоимости.
  - The first one – it will change how you make decisions.
  - It will change how you in fact create value.
  - And the third is, it will change how you deliver value

# Value: Главные источники прироста ВВП в США



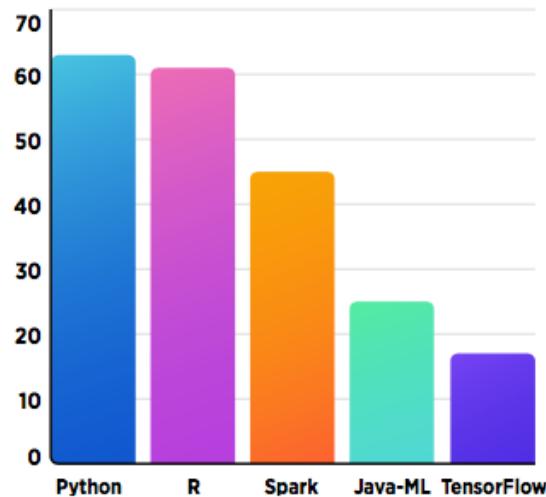
# Техники и технологии больших данных

- Техника (чего-либо) – способ или процедура выполнения какой-либо задачи
- Технология – приложение результатов науки, чаще всего к промышленным или коммерческим целям

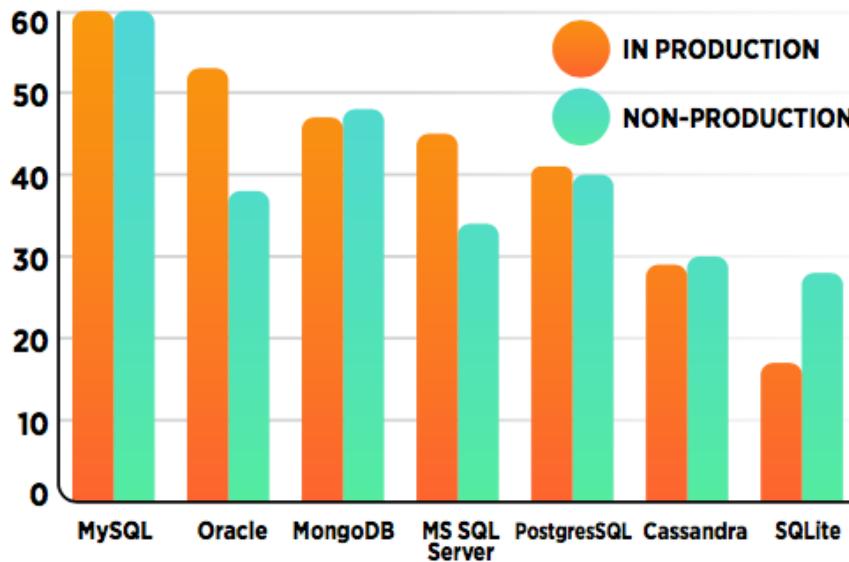
Большие данные, как феномен, включают в себя не только собственно данные как объект операций, но и комплекс особых операций над этим объектом, поскольку объект оказался специфическим и неподвластным известным ранее методам.

Операции над объектом всегда включают в себя ответ на вопрос “Как?” – и это и есть собственно выбор той или иной техники, и ответ на вопрос “Чем?” – а это выбор технологии, инструмента и методики его применения.

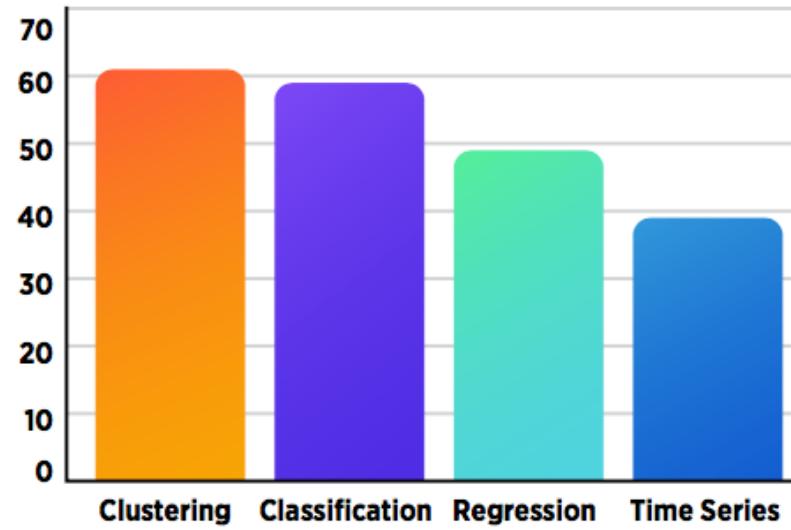
- What languages/libraries/frameworks do you use for data science and machine learning?



- What database management systems do you use?



- What data mining algorithms do you use at work?



# Традиционные технологии vs. Big Data

## Традиционные технологии:

- Не создают знания
- Накопление и интерпретация информации в рамках существующего знания
- Заранее определяем структуры хранения
- Заранее знаем способы использования
- Технология обработки определена заранее
- Изменения сопряжены с доработкой (или даже переработкой) ПО, переформатированием данных

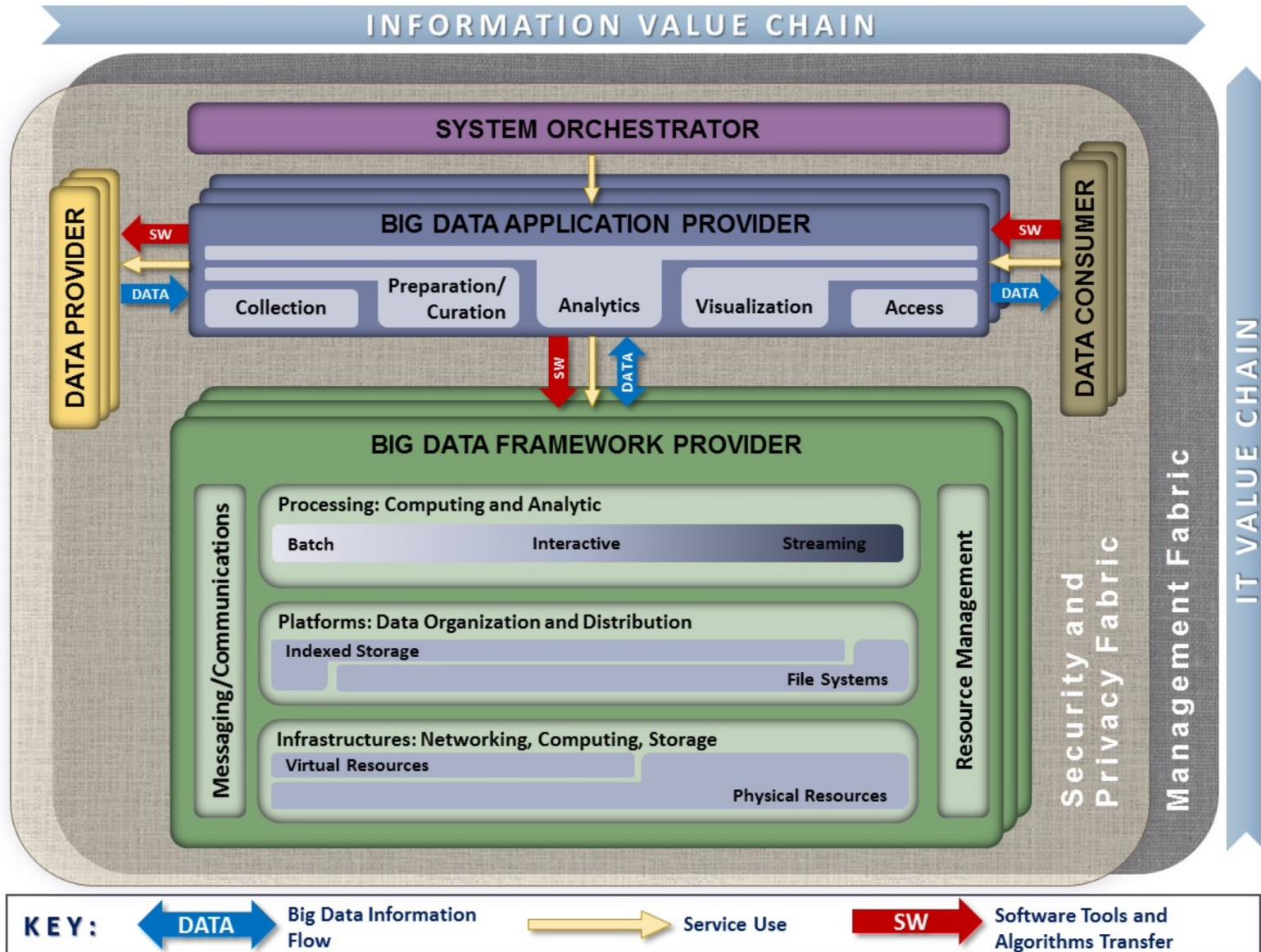


## Технологии Big Data:

- Хранение и обработка/анализ неструктурированных (сырых, «как есть») данных
- Простой инструмент для создания широкого спектра процедур обработки/анализа
- Итерационный исследовательский подход к анализу хранимых данных:
  - отбор, структурирование и агрегация данных
  - формирование модели (или набора)
  - проверка модели (моделей) на существующих данных
  - сравнительная оценка моделей на вновь поступающих данных
  - обновление моделей



# NIST Big Data Reference Architecture (NBDRA) Эталонная архитектура



# Бизнес данных

- Большие данные:
  - Хранение данных
  - Анализ данных
- Data driven companies:
  - Принятие стратегических решений на основе данных
  - Создания продуктов основанные на данных
  - Предсказательная аналитика



# Прикладные задачи

- Маркетинг:

- Сегментация рынка
- Моделирование приобретения и оттока клиентов
- Рекомендательные системы
- Анализ социальных медиа



- Здравоохранение и Фармакология:

- Генетический анализ
- Анализ клинических испытаний
- Клинические системы принятия решений



# Основные направления проектов по аналитике больших данных

- Рекомендательные системы
- Анализ «чувств» (sentiment analysis)
- Моделирование рисков
- Детектирование хищений
- Анализ маркетинговых кампаний
- Анализ оттока клиентов
- Анализ социальных графов
- Аналитика пользовательского опыта
- Мониторинг сетей

# Amazon

## Рекомендательная система

Customers Who Bought This Item Also Bought

The screenshot shows a grid of recommended books based on item similarity. Each book entry includes the title, author, price, and a 'Look Inside' button.

Book Title	Author	Rating	Reviews
Predictive Analytics: The Power to Predict ...	Eric Siegel	★★★★★	(116)
Data Science for Business: What you ...	Foster Provost	★★★★★	(24)
The Big Data Revolution	Jason Kolb	★★★★★	(25)
The Signal and the Noise: Why So Many ...	Nate Silver	★★★★★	(647)
Hadoop: The Engine That Drives Big Data (New ...	Lars Nielsen	★★★★★	(3)
Big Data DUMMIES	Judith Hurwitz	★★★★★	(12)

**Diagram illustrating the recommendation process:**

The diagram illustrates the collaborative filtering process. On the left, a large group of users is shown clustered into subgroups by dashed lines. Arrows point from these subgroups to three specific products: Product A (red), Product B (blue), and Product C (green). Each product is associated with a set of recommended items represented by colored circles (B, D, H, K, P, J, L, H, P).

**Amazon Fulfillment Center:**

An aerial photograph of a massive Amazon fulfillment center, showing rows upon rows of shelving units filled with packages and a complex network of conveyor belts and sorting equipment.

# LinkedIn

Leonid Zhukov  
Director of Data Science  
San Francisco Bay Area - Research

Done editing

500+ connections

Activity

Background

Summary

Proficient researcher and R&D leader with extensive experience conducting research, leading and managing research teams and projects and entrepreneurship. Expert in data science, machine learning, information-retrieval and scientific visualization. Frequent speaker at conferences and professional meetings.

Currently Director of Data Science at Ancestry.com and Professor at the Department of Data Analysis at National Research University Higher School of Economics. Co-founder of information security startup Trefica.

LinkedIn

## Люди, которых вы можете знать

### People You May Know

See people from different parts of your professional life

Московск...  
Государств...  
Университе...  
University of  
Utah  
Stanford  
University

University of  
California,  
Berkeley

Ancestry

Yandex

Brigham  
Young  
University

Belaruskii  
Dzharzhauny  
Universitet

Государств...  
Университе...  
- Высшая  
школа

<p>Erin Takeuchi <small>(rnd)</small> educator at Los Angeles Unified School District Greater Los Angeles Area</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>2 shared connections</small></p>	<p>Andrey Skopenko <small>(rnd)</small> technical director at MCHOST.RU Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>6 shared connections</small></p>
<p>Valeriy Podlesnyi <small>(rnd)</small> CEO &amp; Founder - LAMsystem Russia Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>2 shared connections</small></p>	<p>Tatiana Zolotova <small>(rnd)</small> корреспондент, комикрайтер, редактор Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>3 shared connections</small></p>
<p>Sergey Telegin <small>(rnd)</small> Коммерческий директор (NAG LLC) Sverdlovsk Region, Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>6 shared connections</small></p>	<p>Victor Volgin <small>(rnd)</small> Интернет-маркетолог, ООО "ВИНКОД" Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>2 shared connections</small></p>
<p>Vit Goncharuk <small>(rnd)</small> Entrepreneur - Augmented Reality for Retail, Advertising and Marketing. Ukraine</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>4 shared connections</small></p>	<p>Roman Sokolov <small>(rnd)</small> Software engineer at undev.ru Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a> <small>2 shared connections</small></p>
<p>Андрей Калинин <small>(rnd)</small> CIO - iHomeNet.ru Russian Federation</p> <p><a href="#">Connect</a> <a href="#">Message</a></p>	<p>Alex Toh Kian Hong <small>(rnd)</small> Entrepreneur   Managing Director at Top Image Systems (Asia Pacific)   Angel Investor   Mentor Singapore</p> <p><a href="#">Connect</a> <a href="#">Message</a></p>

LinkedIn

A diagram illustrating a professional network. A central red human-shaped icon is connected by lines to several surrounding blue human-shaped icons, representing a professional network or connections on LinkedIn.

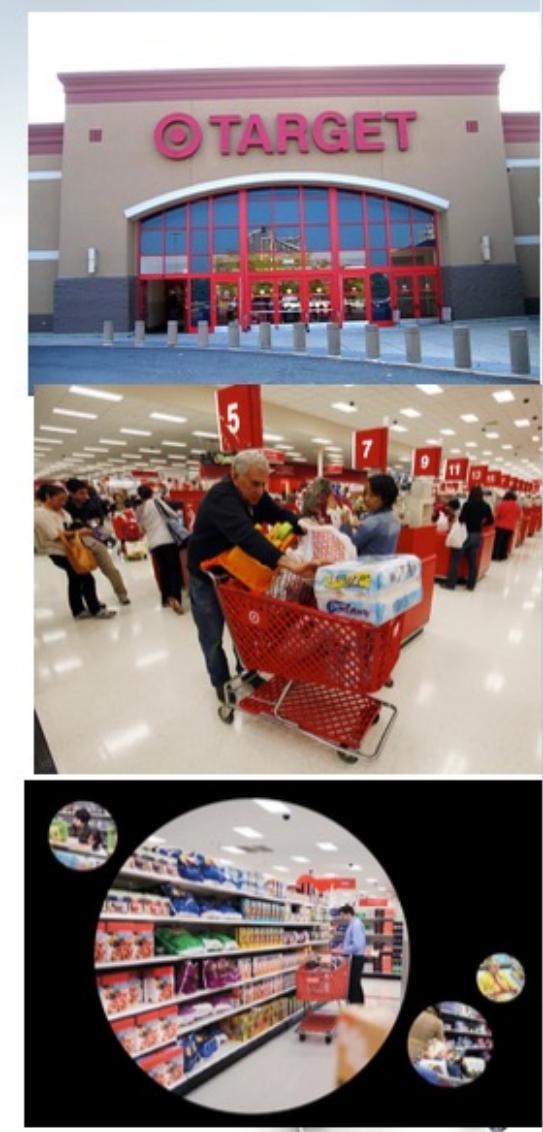
238 mln members

# Target

“How companies learn your secrets”



- Уникальный Guest ID
- Транзакции по кредитной карте
- Примеры факторов (сигналов):
  - Покупка крема без запаха
  - Пищевые добавки кальций, цинк, магний
  - Мыло без запаха
- Предсказательный «индекс» беременности и ожидаемая дата рождения

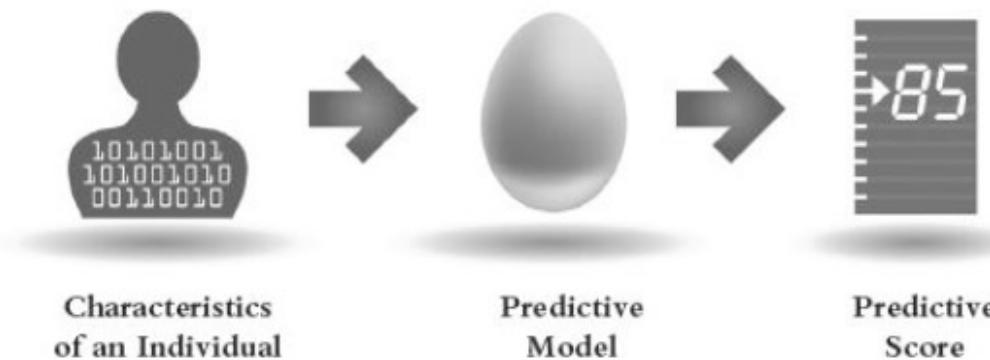


# Netflix

- Обучение модели



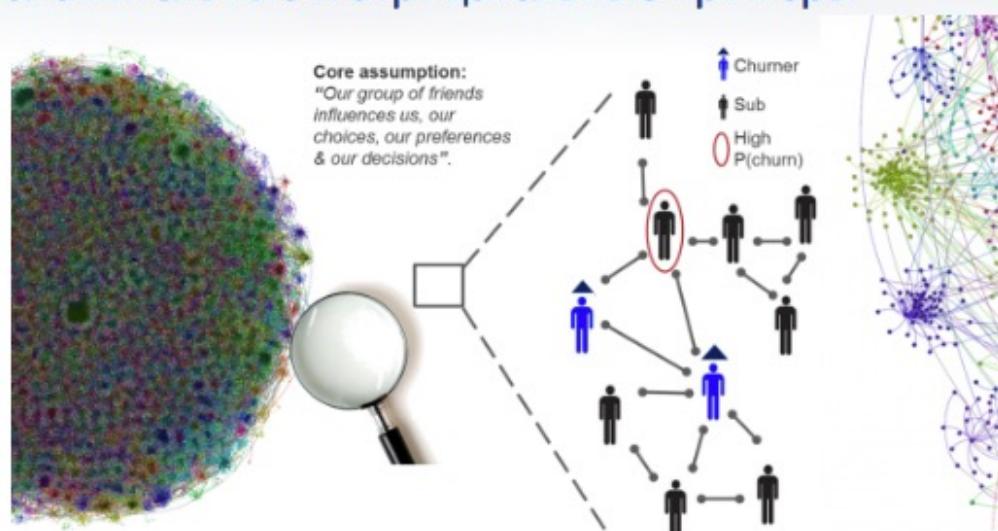
- Применение модели



from Eric Siegel, "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die"

# Операторы мобильной связи

- Churn prediction: моделирование оттока клиентов
- Закономерности поведения подписчиков с течением времени
- Положительные и отрицательные примеры



Факторы модели:

- История пользования сервисом (число звонков, смс)
- История платежей за сервис
- История обращений в службу поддержки
- История изменений в контракте
- Граф звонков (поведение друзей)

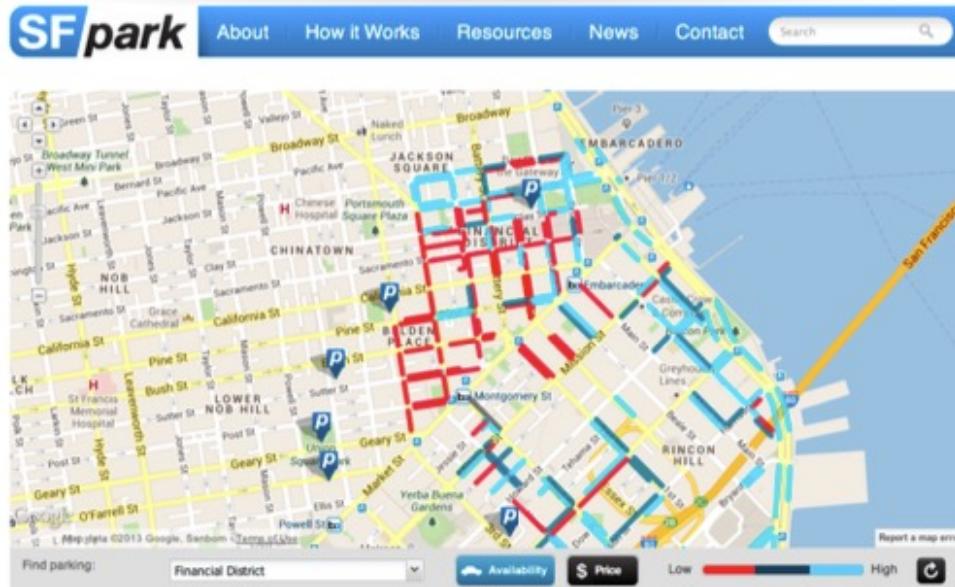


T-Mobile



# Big Data в городе: удобный город

**Сан-Франциско:** датчики парковки, датчики скорости транспортных потоков, GPS в общественном транспорте



## Использование:

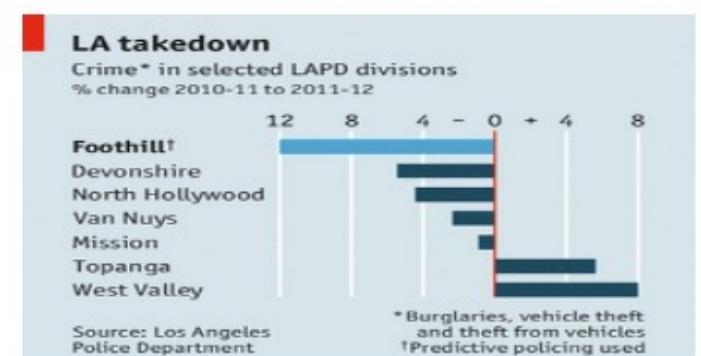
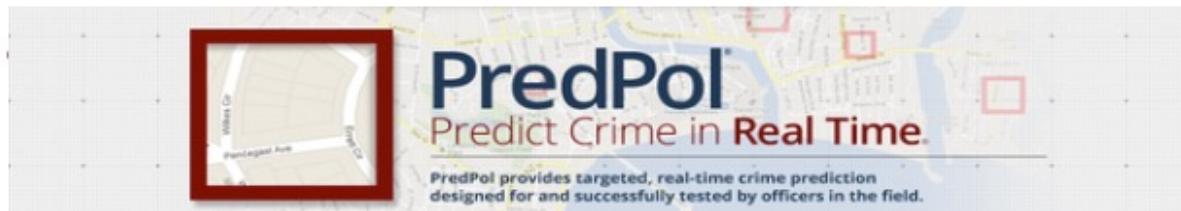
- нахождение свободных парковок
- точное время прибытия общественного транспорта

# Big Data в городе: безопасный город

## Лос-Анджелес: предсказания преступлений

LAPD (**predictive policing**, 2011-2013):

- Определение мест и времени с повышенной вероятностью совершения преступлений
- Исторические данные о преступности в городе (80 месяцев), демографические и др. данные, социологические модели
- Преступления против собственности снизились за год на 12%



# Big Data и здоровье

## Платформа мониторинга астмы и других респираторных заболеваний

- Помогать пациентам и врачам лучше справляется с заболеванием
- Ингаляторы с встроенным сенсорами, мобильные приложения
- Дневник пациента, доступен врачу онлайн
- Неотложная помощь
- Глобальная аналитика по заболеванию



## Системы self-мониторинга (quantify-self):

- Fitbit, Jawbone: физические нагрузки
- Wahoo: сердечный ритм
- Zeo: ночной сон (EEG)
- AliveCor: одноканальная кардиограмма (ECG)



# Применения Big Data

- Анализируя большие данные интернет-запросов, исследователи обнаружили странный феномен. Уже несколько лет всплеск поисковых запросов Google по таким терминам, как лечение гриппа, симптомы гриппа и т.п. на несколько недель предваряет начало стремительного нарастания эпидемии гриппа. Эта закономерность уже сегодня используется для проведения превентивных мер по предотвращению во многих штатах эпидемии гриппа, подготовке врачей, освобождению лечебных коек и т.п. Следует отметить, что используемая до этого информация, поступающая от участковых врачей и пунктов неотложной помощи, как правило, отставала от реальной картины.

# Разоблачение Google и Big Data

Свежая статья в Science [указывает](#) на существенные неточности в прогнозах Google Flu Trends. Сервис более чем на 50% преувеличил размах эпидемии гриппа в сезоны 2012-2013 и 2011-2012 годов. Согласно оценке Google Flu Trends, в разгар прошлогодней эпидемии около 11% жителей США заразились гриппом. Это почти вдвое выше цифр Центра по контролю и профилактике заболеваний США, который не оценивает количество больных по косвенным признакам, а просто пересчитывает их. Кроме того, алгоритмы Google совершенно прозевали вспышку эпидемии вируса H1N1-А («свиной грипп») в 2009-м.

*Нам пришлось обновить модель и опубликовать обзор нашего анализа и сопутствующих изменений после эпидемии вируса H1N1 в 2009 году. То же самое произошло, когда во время сезона 2012-2013 годов оценки нашей модели недостаточно точно соответствовали реальной распространённости гриппа в США. Мы обновили её в августе 2013 года».*



[Skip to content](#)

## Thank you for stopping by.

Google Flu Trends and Google Dengue Trends are [no longer publishing](#) current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we're excited to see what comes next. Academic research groups interested in working with us should fill out this [form](#).

Sincerely,

The Google Flu and Dengue Trends Team.

## Google Flu Trends Data:

You can also see this data in [Public Data Explorer](#)

- [World](#)
- [Argentina](#)
- [Australia](#)
- [Austria](#)
- [Belgium](#)
- [Bolivia](#)
- [Brazil](#)
- [Bulgaria](#)
- [Canada](#)
- [Chile](#)

- Специалисты Федеральной резервной системы выяснили, что статистика поисковых запросов Google относительно покупки домов является более надежным источником для определения тенденций в увеличении или уменьшении объемов продаж недвижимости и динамики жилищного строительства, чем прогнозы наиболее известных экономистов
- По мнению участников Всемирного экономического форума 2012 года в Давосе, те, кто оседлает тему интеллектуального анализа больших данных, станут хозяевами информационного пространства. Этой теме был посвящен специальный доклад на Форуме «Большие данные – большое влияние». Ключевой вывод доклада – цифровые активы становятся не менее значимым экономическим активом, чем золото или валюта.