Основные сквозные технологии цифровой экономики

Большие данные (Big Data)

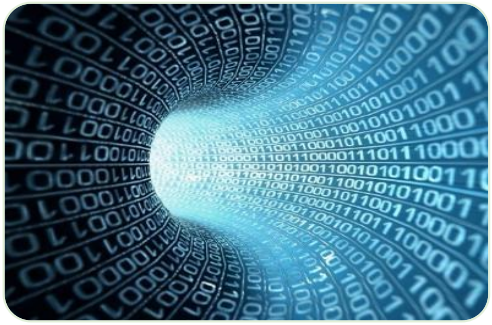
Липатова С.В.,
к.т.н., доцент кафедры ТТС, УлГУ

2019

Информация, море информации

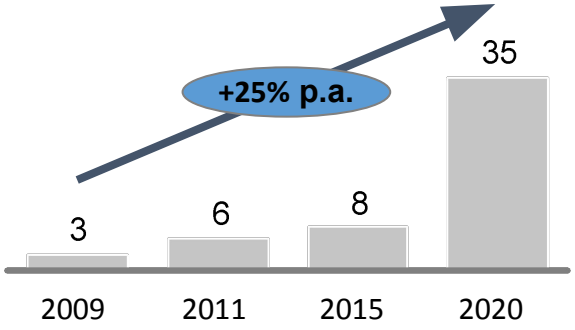
Почему это важно

Более 90 % всех данных было создано в последние 2 года



Объем данных будет **удваиваться каждые 2 года**

Зетабиты



К 2020 году, **1,7 мегабайт новых данных** будет создаваться **каждую секунду для каждого человека** на планете

К 2020 году количество устройств, подключенных к Интернету, **достигнет 50 миллиардов**

Что говорят эксперты

"Данные становятся новым видом сырья для бизнеса"

Крейг Манди, Старший советник директора Microsoft

"Без "больших данных" руководитель подобен глухому и слепому человеку, стоящему посреди автострады"

Джеффри Мур, автор книг и консультант

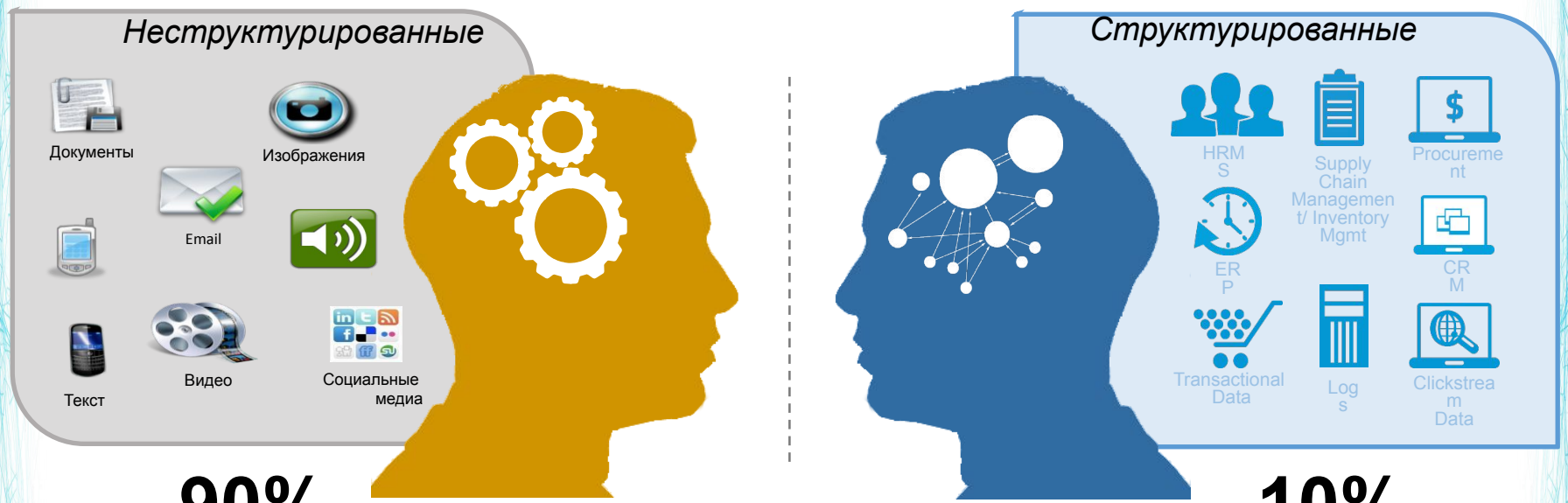
"Десять из 75 человек, задержанных в этом году по подозрению в терроризме, были арестованы благодаря мониторингу соц. сетей"

Высокопоставленный офицер спецслужб США

ИСТОЧНИК: Reuters; Gartner; IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East

Ежегодный рост данных

62% **22%**



90% **10%**

Физика информации

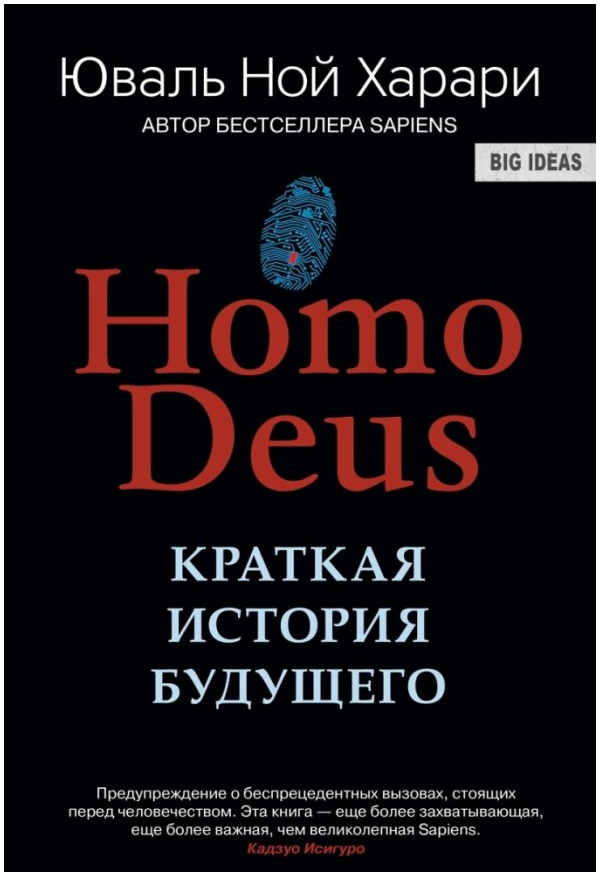
- **Информация** сама по себе является объективной физической величиной в ряду других величин, таких как масса, энергия, импульс и т.д.

- «Все больше теоретиков считают, что ключевой идеей, ведущей к „великому объединению“ гравитации и квантовой теории, может стать переформулирование взглядов на природу не в терминах материи и энергии, а в терминах информации»

Источник: http://quantmag.ppole.ru/QuantumMagic/Doronin1/34.html

Датаизм

— концепция, согласно которой большие данные и алгоритмы обработки этих данных являются высшей ценностью



- 2013 Дэвид Брукс газета The New York Times:
  «Если бы вы попросили меня описать восходящую философию дня, я бы сказал, что это data-ism.»
  -мышление или философия, созданной новым значением больших данных.

- 2016 Юваль Ной Харари книга «Homo Deus»
  «Датаизм идеологией или даже новой формой религии, в которой информационный поток является высшей ценностью».

«В наше время мы страдаем не столько из-за недостатка информации, сколько от избытка ненужной, бесполезной информации, не имеющей никакого отношения к выходу из кризисных ситуаций. Найдите возможности отделить бесполезное от важного, и вы почувствуете, что владеете ситуацией.»

Джефф О`Лири

«Информационный поток, в котором человечество пребывает, смыл все вопросы о смысле жизни.»

Анатолий Канашкин

Негативные последствия перегруженности информацией

Перегруженность информацией приводит к снижению когнитивных (познавательных) функций мозга

ИЛИ

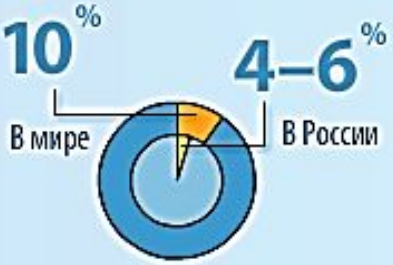
Возникает феномен информационно-коммуникативной зависимости

Феномен интернет-зависимости

Интернет-зависимость – навязчивое желание войти в Интернет, находясь в оффлайне, и неспособность выйти из Интернета, будучи онлайн

18 часов способны проводить люди за компьютером, страдающие интернет-зависимостью

По некоторым данным, в мире интернет-зависимыми являются 10% пользователей ПК



Информационный взрыв

- Впервые об угрозе "информационного взрыва" ученые заговорили в 60-х годах XX века.
- Мозг обычного человека способен воспринимать и безошибочно обрабатывать информацию со скоростью **не более 25 бит в секунду** (в одном слове средней длины содержится как раз 25 бит).
- **Макулатурный фактор** - 90 процентов литературы пользуется нулевым спросом.
- **Период полураспада актуальных знании** в области высшего образования составляет примерно семь - десять лет, в компьютерных технологиях сократился до года.

Источник: http://www.mirprognozov.ru/prognosis/society/informatsionnyiy-vzryiv-ugroza-buduschemu-tsivilizatsii/

В мире проанализировано менее 1% всей информации, защищено менее 20%

- Основные прогнозы
  - ❑ Объемы информации будут удваиваться каждые два года еще восемь лет. Один из основных факторов роста - увеличение доли автоматически генерируемых данных с 11% от общего объема (2005 г.) до более 40% в 2020 г.
  - ❑ Большие объемы полезных данных теряются. На сегодня используется менее 3% из 23% потенциально полезных данных, которые могли бы найти применение с технологиями Big Data.

- Большая часть информации плохо защищена
  - ❑ В 2010 г. в защите нуждалось менее трети информации, к 2020 г. ее доля может превысить 40%.
  - ❑ Уровень защиты варьируется по регионам — у развивающихся рынков он гораздо ниже.
  - ❑ У развивающихся рынков на 2010 г. доля цифровой вселенной была 23%, к 2012 г. она составила 36%, а к 2020 г. , согласно прогнозам IDC, дойдет до 62%.

сточник: http://bit.samag.ru/uart/more/23

Определения и концепция больших данных

(Big Data)

Термин и тренд Big Date

1998 Джон Мэши:

ввёл в обиход термин Big Data.

2005 издание компании O'Reilly media:

первое упоминание данных, с которыми традиционные технологии управления и обработки данных не справлялись в силу их сложности и большого объёма.

2008 Клиффорд Линч, специальный номер журнала Nature:

введение термина «большие данные» в современном понимании.

2011 компания Gartner:

прогноз, что Big Date окажет влияние на подходы в области информационных технологий в производстве, здравоохранении, торговле и государственном управлении;

большие данные - тренд номер два в информационно-технологической инфраструктуре (после виртуализации).

2015 компания Gartner:

исключил Big Data из числа прорывных технологий (emerging technologies): "чтобы перевести дискуссию о Больших Данных из области спекуляций в практическую плоскость".

Клиффорд Линч

Большие данные сравнивали с:

- минеральными ресурсами —
  - the new oil (новая нефть),
  - goldrush (золотая лихорадка),
  - data mining (разработка данных), чем подчеркивается роль данных как источника скрытой информации;
- с природными катаклизмами —
  - data tornado (ураган данных),
  - data deluge (наводнение данных),
  - data tidal wave (половодье данных), видя в них угрозу;
- с промышленным производством —
  - data exhaust (выброс данных),
  - firehose (шланг данных),
  - Industrial Revolution (промышленная революция).

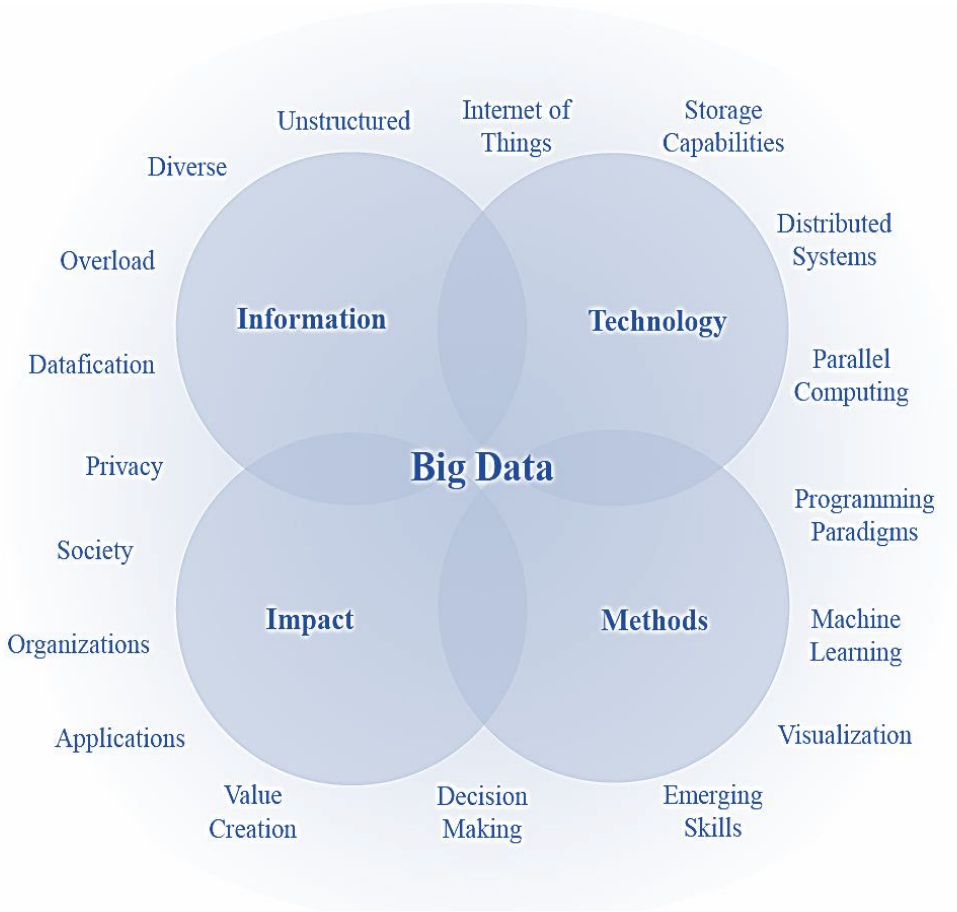
Существует ли проблема Больших Данных ?

- Большие Данные - red herring (букв. «копченая селедка» — ложный след, отвлекающий маневр.
- Большие Данные - прежде всего маркетинговый ход разработчиков, продвигающих свою продукцию.

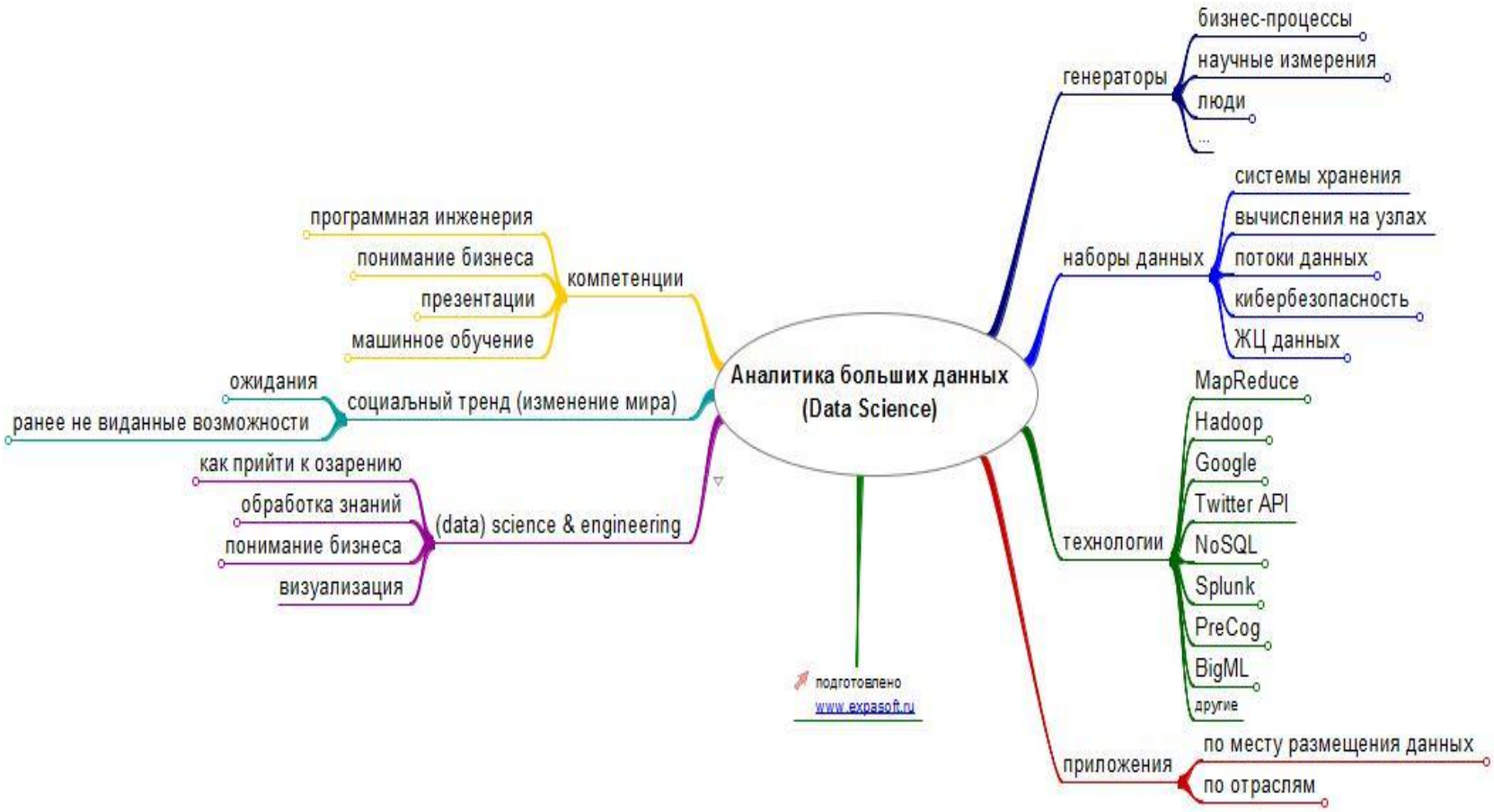
Возможно, Большие Данные есть что-то качественно иное, чем то, к чему подталкивает обыденное сознание.

Большие данные (Big Data)

– это «зонтичный» термин, объединяющий группу понятий, технологий и методов производительной обработки очень больших объёмов данных, в том числе неструктурированных, в распределённых информационных системах, обеспечивающих организацию качественно новой полезной информации (знаний).



Разные взгляды на применение больших данных



Определение больших данных как технологии

Большие данные – это:

- серия подходов, инструментов и методов
  - *обработки* структурированных и неструктурированных данных огромных объёмов и значительного многообразия
    - *для получения* воспринимаемых человеком результатов,
      - эффективных в условиях *непрерывного прироста* и *распределения* по многочисленным узлам вычислительной сети,
        - *альтернативных традиционным системам* управления базами данных.

Cloud of tags

Вот как Интернет представляет понятие BigData

Точнее – как Большие Данные представляют сами себя.

Характеристики больших данных

Объем (Volume)

- 10% организаций обрабатывают 1+ Пб данных
- Социальные сети – миллионы транзакций в минуту

Скорость (Velocity)

- 30% организаций имеют 100+ Гб/день
- Данные обновляются и нужны раз в день, час

Разнообразие (Variety)

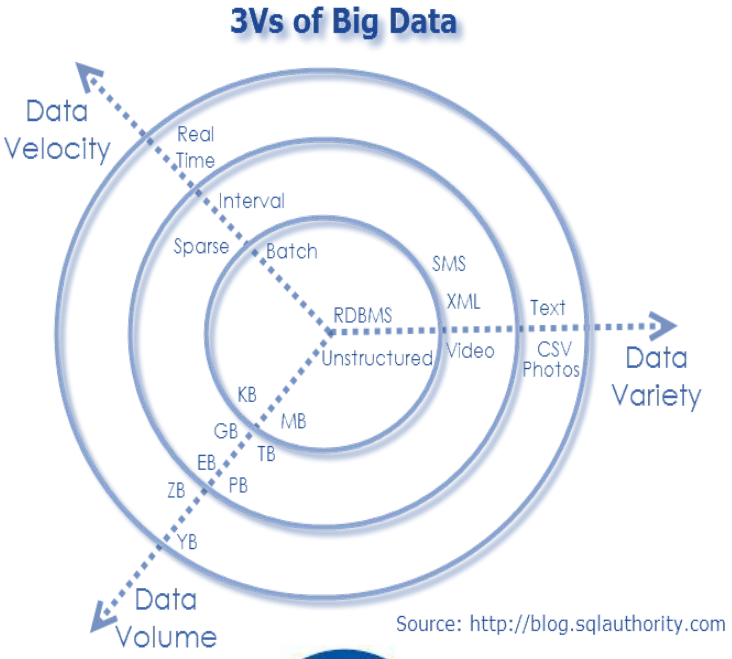
- Тексты, Аудио и видео файлы
- Блоги, сообщения в сетях – для изучения клиентов
- Внутренние источники данных

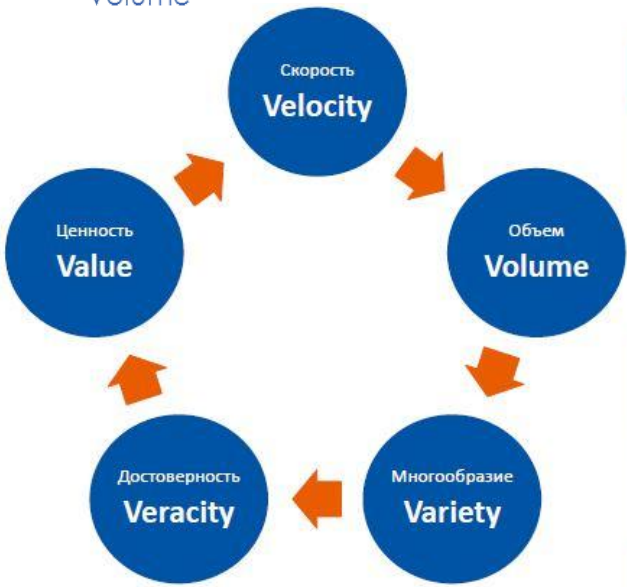
Достоверность (Veracity)

- Осмысленные связи
- Преобразование
- Очистка

Значимость (Value)

- ценность
- накопленной информации



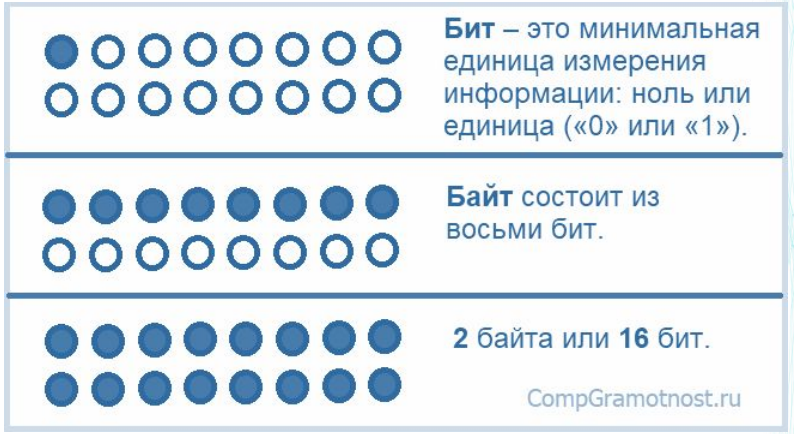


Отличия данных от больших данных

Характеристика	Традиционная база данных	База Больших Данных
Объем информации	От гигабайт (10^9 байт) до терабайт (10^12 байт)	От петабайт (10^15 байт) до эксабайт (10^18 байт)
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурирована	Полуструктурирована и неструктурирована
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Таблица байтов:

- 1 байт = 8 бит
- 1 Кб (1 **Килобайт**) = 2^{10} байт = 2*2*2*2*2*2*2*2*2*2 байт = = 1024 байт (примерно 1 тысяча байт – 10^3 байт)
- 1 Мб (1 **Мегабайт**) = 2^{20} байт = 1024 килобайт (примерно 1 миллион байт – 10^6 байт)
- 1 Гб (1 **Гигабайт**) = 2^{30} байт = 1024 мегабайт (примерно 1 миллиард байт – 10^9 байт)
- 1 Тб (1 **Терабайт**) = 2^{40} байт = 1024 гигабайт (примерно 10^{12} байт). Терабайт иногда называют *тонна*.
- 1 Пб (1 **Петабайт**) = 2^{50} байт = 1024 терабайт (примерно 10^{15} байт).
- 1 **Эксабайт** = 2^{60} байт = 1024 петабайт (примерно 10^{18} байт).
- 1 **Зеттабайт** = 2^{70} байт = 1024 эксабайт (примерно 10^{21} байт).
- 1 **Йоттабайт** = 2^{80} байт = 1024 зеттабайт (примерно 10^{24} байт).



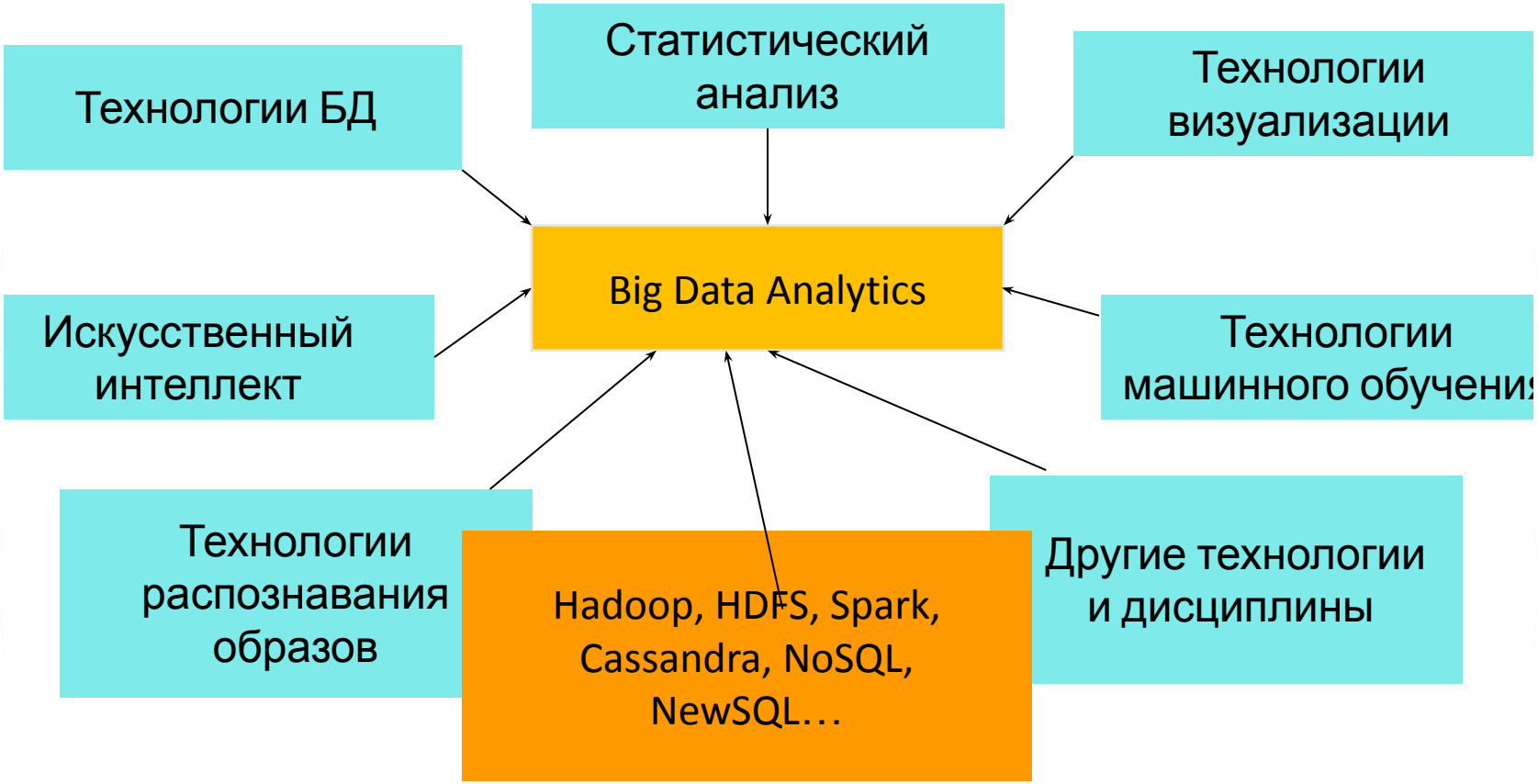
Классификация Больших Данных

Дайон Хинчклиф, редактора журнала Web 2.0 Journal делит Большие данные на 3 группы:

- Быстрые Данные (Fast Data), их объем измеряется терабайтами;
- Большая Аналитика (Big Analytics) — петабайтные данные
- Глубокое Проникновение (Deep Insight) — экзабайты, зеттабайты.

Группы различаются между собой не только оперируемыми объёмами данных, но и качеством решения по их обработки.

Взаимосвязь между технологиями



Источники данных

Внутренние основные

- Данные с камер, сенсоров, и пр.
- Данные с GPS общественного транспорта
- Данные по платежам за проезд
- Данные с турникетов (транзакционные данные по проездам)
- [Прочие источники]

Внутренние дополнительные

- Специфичные данные пассажира (карта москвича)
- Активность на web-сайтах (посещения, комментарии и пр.)
- Записи звонков
- Реестр e-mail
- Данные по резуль-татам обратной связи с клиентами
- [Прочие источники]

Внешние основные

- Данные использования услуг (Wi-Fi в метро)
- Данные мобильных операторов
- Данные торговых сетей
- WEB-browsing data
- Данные бюро кредитных историй
- [Прочие источники]

Внешние дополнительные

- Данные в открытом доступе (новости, блоки, wiki, и. пр.
- Социальные сети (VK, Facebook Одноклассники, и пр.)
- [Прочие источники]

Сложность получения данных

Источник: McKinsey & Company

Методы анализа больших данных

Группа аналитических методов		Постановка задачи
Описательные методы	**1**	**Описать** взаимосвязи или составить выводы на основе ваших данных, например ▪ Какие сегменты возможно выделить на основе потребностей клиентов? ▪ Каким образом осуществляется информационное взаимодействие между сотрудниками вашей компании?
Прогнозные методы	**2**	**Спрогнозировать** результаты и/или влияющие на них факторы, например ▪ Какие транзакции являются мошенническими? ▪ Каким будет объем продаж в следующем квартале?
Директивные методы	**3**	**Оптимизировать** работу системы с учетом определенных ограничений, например ▪ Какой уровень запасов позволяет минимизировать затраты и обеспечить поставку продукции менее чем за неделю?

Источник: McKinsey & Company

Эволюция статистических алгоритмов

Статистика больших данных (машинное обучение)

- Gradient Boosting (GBM)
- Машина Больцмана
- Learning Vector
- Квантование (LVQ)
- Random Forest
- FP-growth
- ApprioriDP
- OPUS
- CBPNARM
- Snipers
- Multi-armed bandits
- ….

Классическая статистика (эра до больших данных)

Регрессии

Байесовская статистика

Цепь Маркова

Нейронные сети

Монте Карло

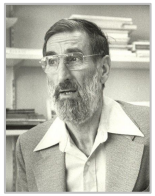
Clustering (k-means)

Деревья решений

1795 1810 1906 1943 1946 1957 1973 **2000** … 2014

Гаусс

Фон Нейман



Морган

Брейман

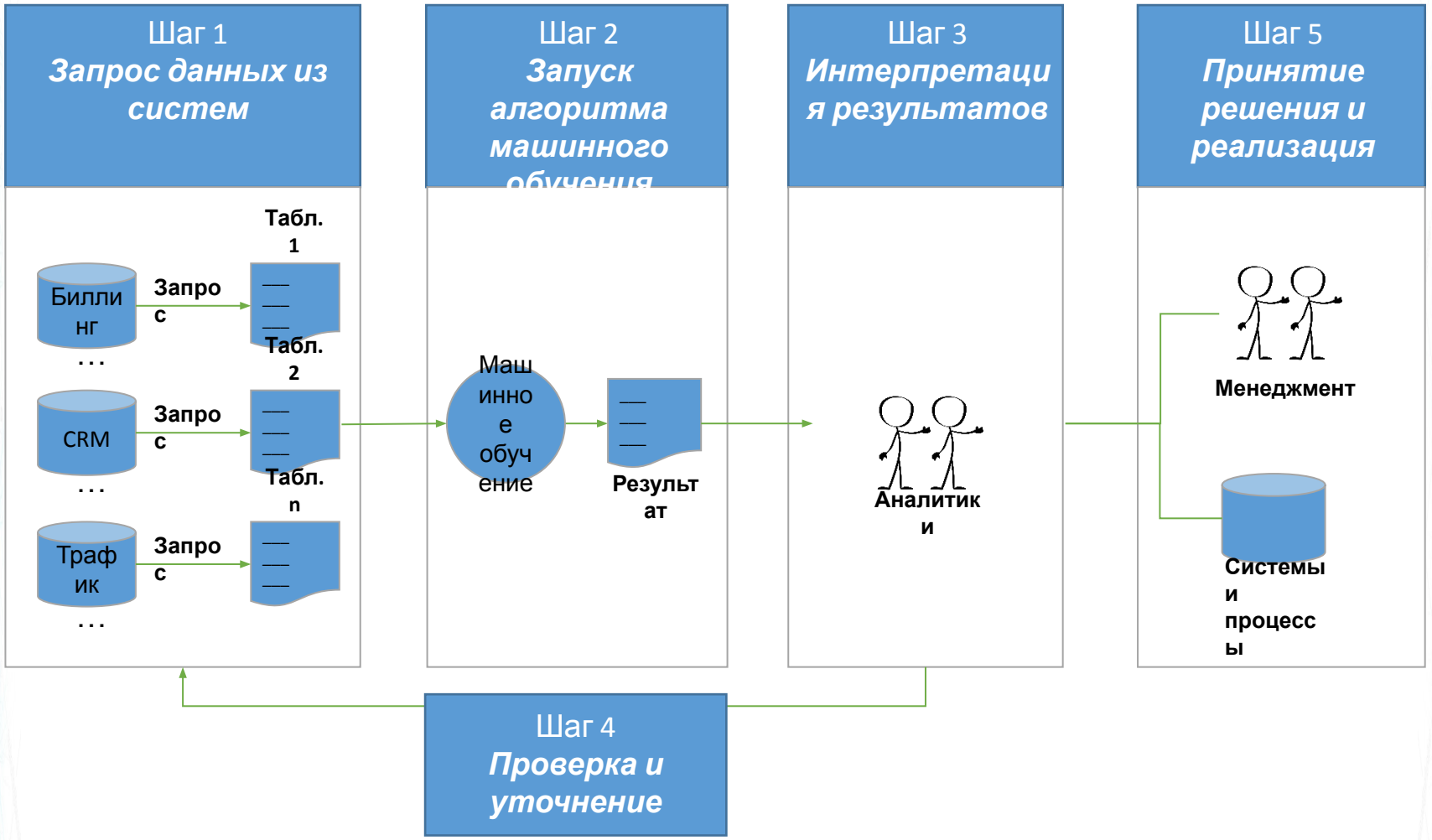
Source: McKinsey Analysis

В результате этих преобразований подходы к анализу данных радикально изменились

От…	…К
Структурированные и централизованные массивы данных	**Многообразие**: неструктурированные и рассредоточенные данные – **Сбор Данных**
Анализ – вспомогательный вид деятельности	Анализ – **стратегический инструмент создания стоимости**
Описание произошедших событий	**Прогнозирование** будущих событий
Предположение → Анализ (подтверждение или опровержение) → Действие	Данные → Решения и Закономерности → Действие
Асинхронный режим	**Режим реального времени**
Традиционный последовательный процесс	Адаптивный метод **постоянного апробирования**

К: McKinsey Analytics

Упрощённая схема работы с большими массивами данных



Источник: McKinsey & Company

Комплексный подход к большим данным

1. ДАННЫЕ

- **Источники** (внутренних и внешних/ структурированных и неструктурированных), **методы сбора, трансформации** и **обогащения** данных
- **Модели создания хранилища данных** ("озеро" обогащенных данных для всех use-cases или система взаимосвязанных "прудов" под различные use-cases)
- **Места хранения данных** (собственные сервера или сервера партнеров) и **владельцы данных**

2. АНАЛИТИКА

- **Стратегия** по работе с большими данных и глубокой аналитикой (включая определение объема работ выполняемых своими силами и силами партнеров и стратегии работы с партнерами)
- **Модели и алгоритмы** для работы с большими данных и глубокой аналитикой

3. ОРГАНИЗАЦИОННАЯ СТРУКТУРА И БИЗНЕС-ПРОЦЕССЫ

- **Целевая организационная модель и бизнес-процессы** для внедрения больших данных
- **Процессы и процедуры** (выполняемые собственными предприятиями и партнерами); **распределение процессов на уровне подведомственных предприятий** Департамента Транспорта
- **Необходимые компетенции** и их источники (внутренние или внешние ресурсы)

4. ИНФРАСТРУКТУРА

Определение **стратегии по инфраструктурным решениям** для работы с большими данными и глубокой аналитикой

- Программных решений для сбора, трансформации и интеграции данных
- Способов предоставления общего доступа для внутренних пользователей компании и ее партнеров
- Систем для предоставления доступа пользователей к большим данным
- Интеграции с системами управления взаимоотношениями с клиентами (CRM)

Источник: McKinsey & Company

Джозеп Курто, управляющий независимой консалтинговой компанией Delfos Research, ассоциированный профессор IE Schoolof Social, Behavioral & Data Sciences:

«Внедрение Big Data—это не просто привлечение одного специалиста, это изменение мышления всех сотрудников.... Очень важно развеять миф о том, что Big Data—это просто какая-то часть IT-департамента.»

http://future.theoryandpractice.ru/12109

Профессии Big Date

- исследователь данных
- консультант в области больших данных
- инженер по большим данным
- архитектор больших данных
- специалист по управлению большими данными

https://www.bigdataschool.ru/bigdata/data-professional-agile-team-big-data.html

Игроки на рынке Big Date

- **Поставщики инфраструктуры** — решают задачи хранения и предобработки данных.

  Например: IBM, Microsoft, Oracle, Sap и другие.

- **Датамайнеры** — разработчики алгоритмов, которые помогают заказчикам извлекать ценные сведения.

  Среди них: Yandex Data Factory, «Алгомост», Glowbyte Consulting, CleverData и др.

- **Системные интеграторы** — компании, которые внедряют системы анализа больших данных на стороне клиента.

  К примеру: «Форс», «Крок» и др.

- **Потребители** — компании, которые покупают программно-аппаратные комплексы и заказывают алгоритмы у консультантов.

  Это «Сбербанк», «Газпром», «МТС», «Мегафон» и другие компании из отраслей финансов, телекоммуникаций, ритейла.

- **Разработчики готовых сервисов** — предлагают готовые решения на основе доступа к большим данным. Они открывают возможности Big Data для широкого круга пользователей.

https://www.uplab.ru/blog/big-data-technologies/

Направления Big Date

- Сбор и обработка больших данных
- Аналитика
- Инженерия больших данных
- Архитектура больших данных и системная интеграция
- Разработка продуктов и услуг на основе больших данных
- Управление большими данными и системами на основе больших данных
- Проведение исследований с целью получения новых математических и технических решений для работы с большими данными

Приоритетные направления для компаний

Источник: Economist Intelligence Unit

Цели применения:

- ❑ Эффективность
- ❑ Удовлетворение клиентов
- ❑ Снижение риска
- ❑ Расширение бизнеса

Big Data не нужны, если

- сотрудники в состоянии обработать и автоматизировать данные по клиентам с помощью обычных CRM-систем;
- планирование, учёт и контроль бизнес-процессов вполне реализуем с помощью ERP-систем;
- раньше объединяли данные из различных источников информации, обрабатывали их, оценивали полученный результат с помощью BI-систем и не испытывали со всем вышеперечисленным никаких трудностей.

https://netology.ru/blog/6-mif-bigdata

Мэтт Слокум из **O'Reilly Radar** считает, что хотя **большие данные** и **бизнес-аналитика** имеют одинаковую цель (поиск ответов на вопрос), они отличаются друг от друга:

- Большие данные предназначены для обработки более значительных объёмов информации, чем бизнес-аналитика, и это, конечно, соответствует традиционному определению больших данных.

- Большие данные предназначены для обработки более быстро получаемых и меняющихся сведений, что означает глубокое исследование и интерактивность. В некоторых случаях результаты формируются быстрее, чем загружается веб-страница.

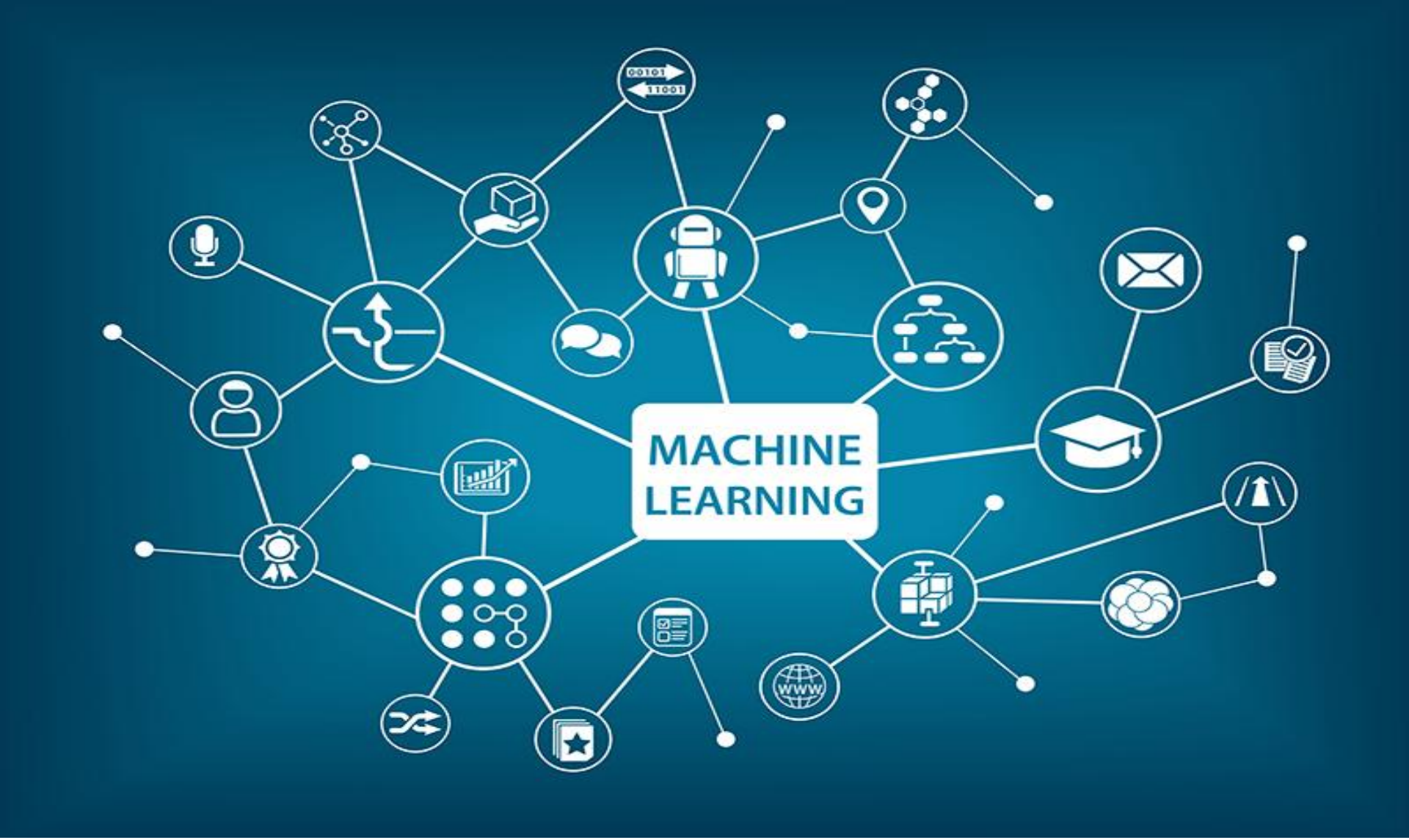
- Большие данные предназначены для обработки неструктурированных данных, способы использования которых мы только начинаем изучать после того, как смогли наладить их сбор и хранение, и нам требуются алгоритмы и возможность диалога для облегчения поиска тенденций, содержащихся внутри этих массивов.

http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data)

Факторы развития технологии

Драйверы	Ограничители
Высокий спрос на Big Data для повышения конкурентоспособности с помощью возможностей технологий	Необходимость обеспечивать безопасность и конфиденциальность данных
Развитие методов обработки медиафайлов на мировом уровне	Нехватка квалифицированных кадров
Реализация отраслевого плана по импортозамещению программного обеспечения	В большинстве российских компаний объем накопленных информационных ресурсов не достигает уровня Big Data
Тренд на использование услуг российских провайдеров и системных интеграторов	Новые технологии сложно внедрять в устоявшиеся информационные системы компаний
Создание технопарков, которые способствуют развитию информационных технологий	Высокая стоимость технологий
Государственная программа по внедрению грид-систем — виртуальных суперкомпьютеров, которые распространяются по кластерам и связываются сетью	Заморозка инвестиционных проектов в России и отток зарубежного капитала
Перенос на территорию России серверов, которые обрабатывают персональную информацию	Рост цен на импортную продукцию

https://netology.ru/blog/6-mif-bigdata



Машинное обучение и большие данные

Место машинного обучения среди других технологий

- Pattern Recognition (**распознавание образов**)

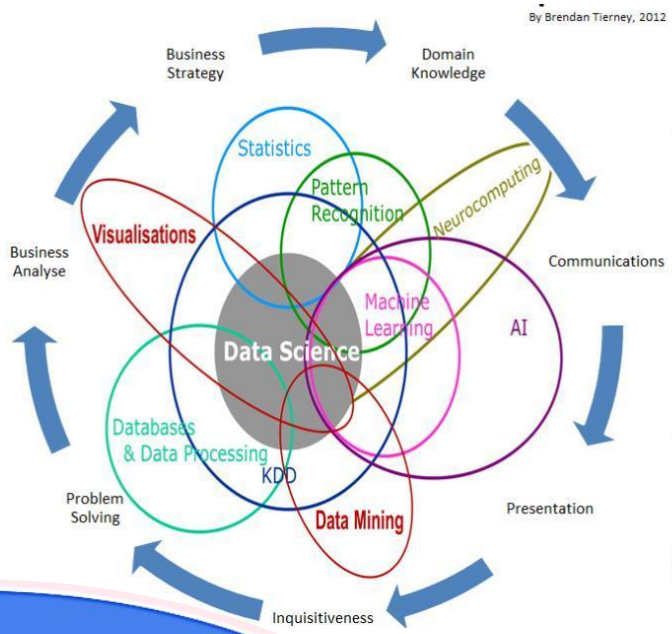
  Pattern Recognition ≈ Machine Learning

- Data Mining (**интеллектуальный анализ данных**) (включая Big Data)

  Data Mining ∩ Machine Learning <>0

- Artificial Intelligence (**искусственный интеллект**)

  Machine Learning ⊂ Artificial Intelligence



https://vc.ru/future/59364-aimath

https://studylib.ru/doc/1708570/mashinnoe-obuchenie-i-analiz-dannyh

Типы обучения

- Дедуктивное или аналитическое обучение (экспертные системы).

  Имеются знания, сформулированные экспертом и как-то формализованные.

  Программа выводит из этих правил конкретные факты и новые правила.

- Индуктивное обучение (≈статистическое обучение).

  На основе эмпирических данных программа строит общее правило.

  Эмпирические данные могут быть получены самой программой в предыдущие сеансы ее работы или просто предъявлены ей.

- Комбинированное обучение.

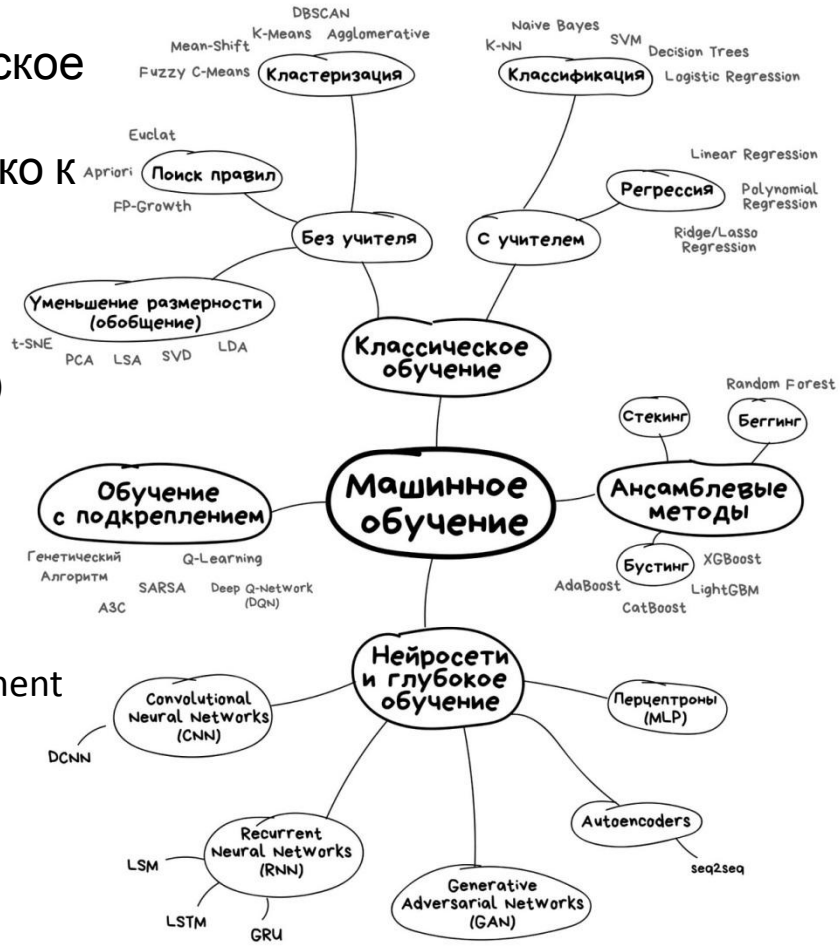
https://studylib.ru/doc/1708570/mashinnoe-obuchenie-i-analiz-dannyh

Классификация задач машинного обучения

- Дедуктивное обучение (экспертные системы)
- Индуктивное обучение (≈ статистическое обучение)

(определение Митчелла относится только к такому обучению)

- Обучение с учителем:
  - классификация
  - восстановление регрессии
  - структурное обучение (structured learning)
  - . . .
- Обучение без учителя:
  - кластеризация
  - визуализация данных
  - понижение размерности
  - . . .
- Обучение с подкреплением (reinforcement learning)
- Активное обучение
- –. . .



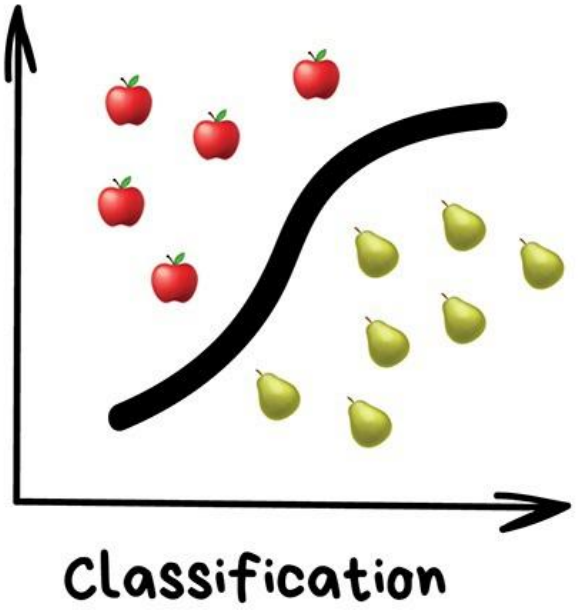
https://studylib.ru/doc/1708570/mashinnoe-obuchenie-i-analiz-dannyh

Обучение с учителем: Классификация

– это зависимость входных данных от дискретных выходных.

Популярные алгоритмы:

- ❑ Наивный Байес,
- ❑ Деревья Решений,
- ❑ Логистическая Регрессия,
- ❑ К-ближайших соседей,
- ❑ Машины Опорных Векторов

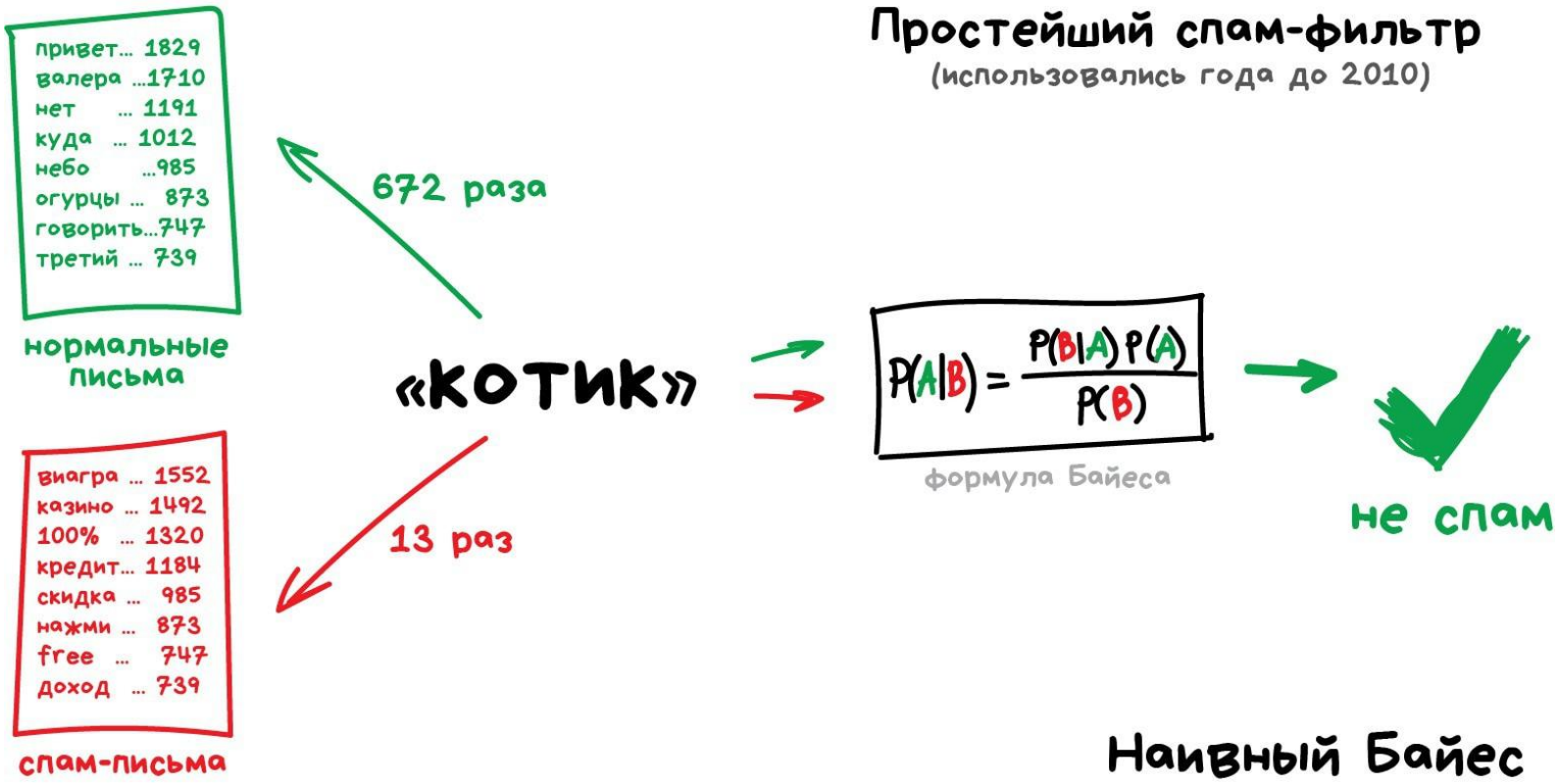


Используют для:

- ❑ Спам-фильтры
- ❑ Определение языка
- ❑ Поиск похожих документов
- ❑ Анализ тональности
- ❑ Распознавание рукописных букв и цифр
- ❑ Определение подозрительных транзакций

https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение с учителем: классификация: Наивный Байес



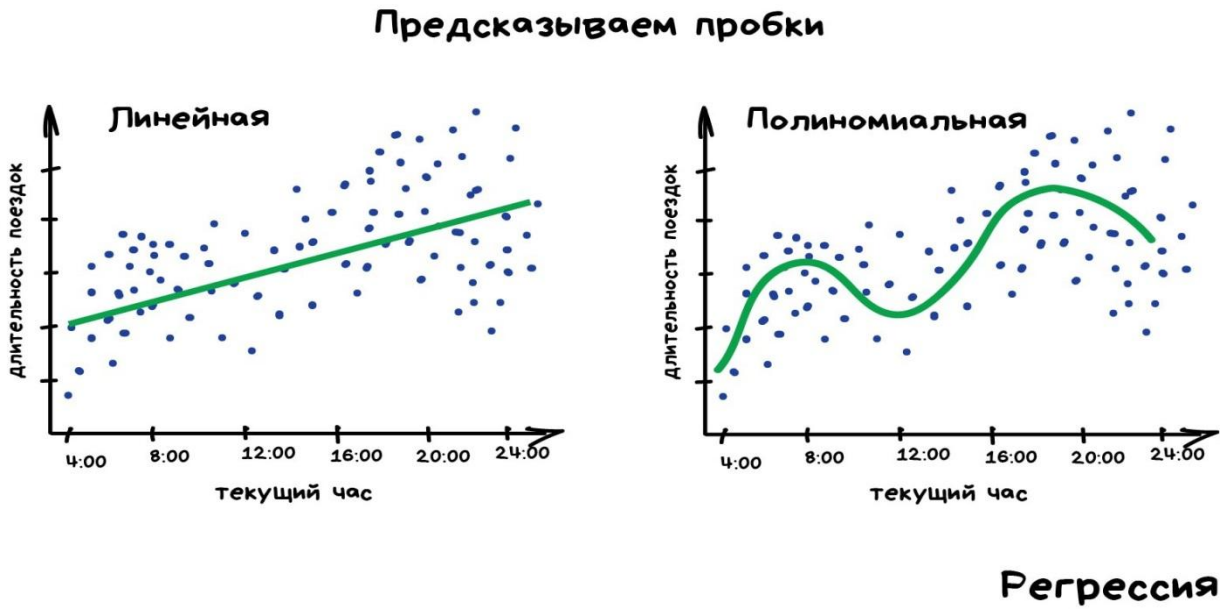
https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение с учителем: классификация: дерево решений

https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение с учителем: Регрессия

- это зависимость между входными данными и непрерывными выходными.



Регрессия

Используют для:

- ❑ Прогноз стоимости ценных бумаг
- ❑ Анализ спроса, объёма продаж
- ❑ Медицинские диагнозы
- ❑ Любые зависимости числа от времени

Популярные алгоритмы:

- ❑ Линейная Регрессия
- ❑ Полиномиальная Регрессия

https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение без учителя: Кластеризация

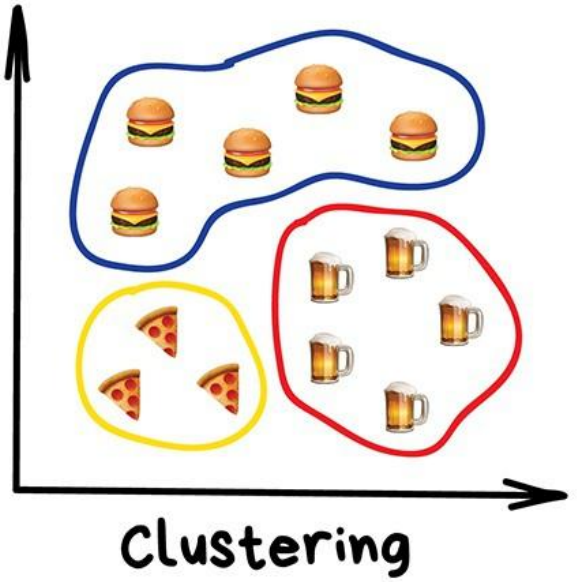
- это группировка данных руководствуясь свойствами этих данных. Данные внутри кластера должны иметь одинаковые свойства и отличаться от свойств данных других кластеров.

Популярные алгоритмы:

- ❑ Метод K-средних,
- ❑ Mean-Shift,
- ❑ DBSCAN

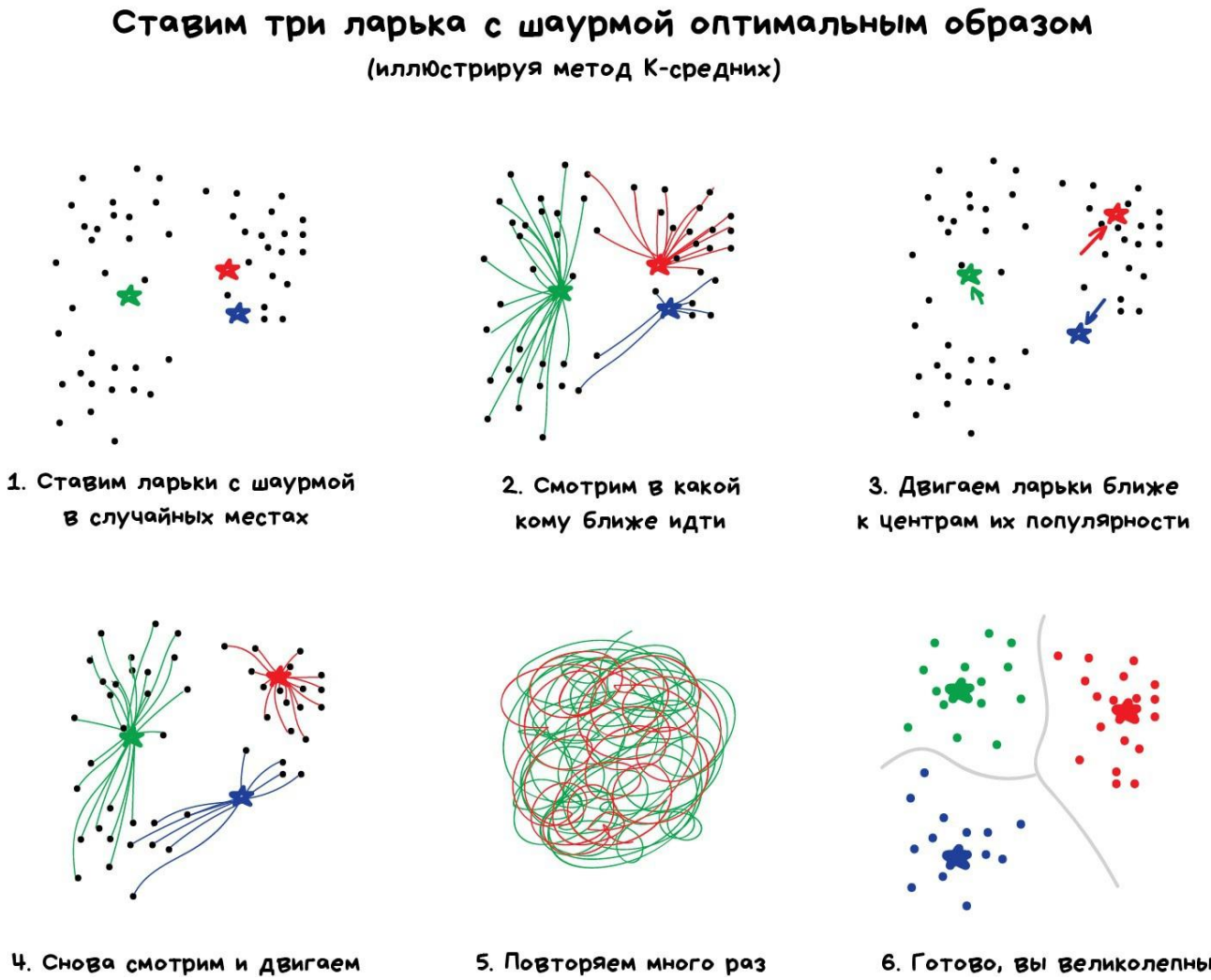
Используют для:

- ❑ Сегментация рынка (типов покупателей, лояльности)
- ❑ Объединение близких точек на карте
- ❑ Сжатие изображений
- ❑ Анализ и разметки новых данных
- ❑ Детекторы аномального поведения



https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение без учителя: Кластеризация: метод k-средних



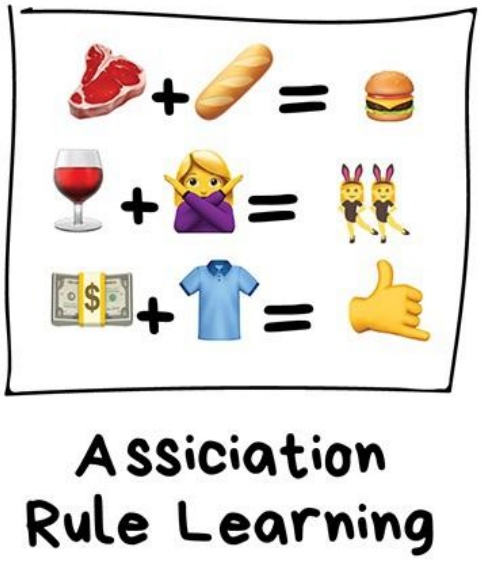
https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение без учителя: Ассоциация

– поиск закономерностей между связанными событиями. К примеру, можно привести следующее правило, что из события X следует событие Y. Такие правила называются ассоциативными.

Популярные алгоритмы:

- ❑ Apriori,
- ❑ Euclat,
- ❑ FP-growth



Используют для:

- ❑ Прогноз акций и распродаж
- ❑ Анализ товаров, покупаемых вместе
- ❑ Расстановка товаров на полках
- ❑ Анализ паттернов поведения на веб-сайтах

https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Последовательные шаблоны

– установление закономерностей между связанными во времени событиями, т.е. обнаружение зависимости, что если произойдёт событие X, то спустя заданное время произойдёт событие Y.

Аналогичен ассоциации, но с учётом временной составляющей.

Популярные алгоритмы:

- AprioriAll
- AprioriSome
- DynamicSome

Используют для:

- Прогноз цепочек событий
- Поиск причинно-следственных связей

https://basegroup.ru/community/articles/sequential-patterns-1

Обучение без учителя: Уменьшение размерности

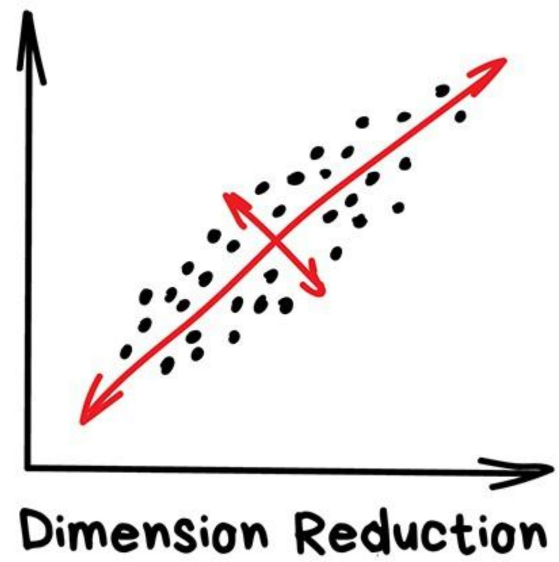
– собирает конкретные признаки в абстракции более высокого уровня.

Популярные алгоритмы:

- ❑ Метод главных компонент (PCA),
- ❑ Сингулярное разложение (SVD),
- ❑ Латентное размещение Дирихле (LDA),
- ❑ Латентно-семантический анализ (LSA, pLSA, GLSA),
- ❑ t-SNE (для визуализации)

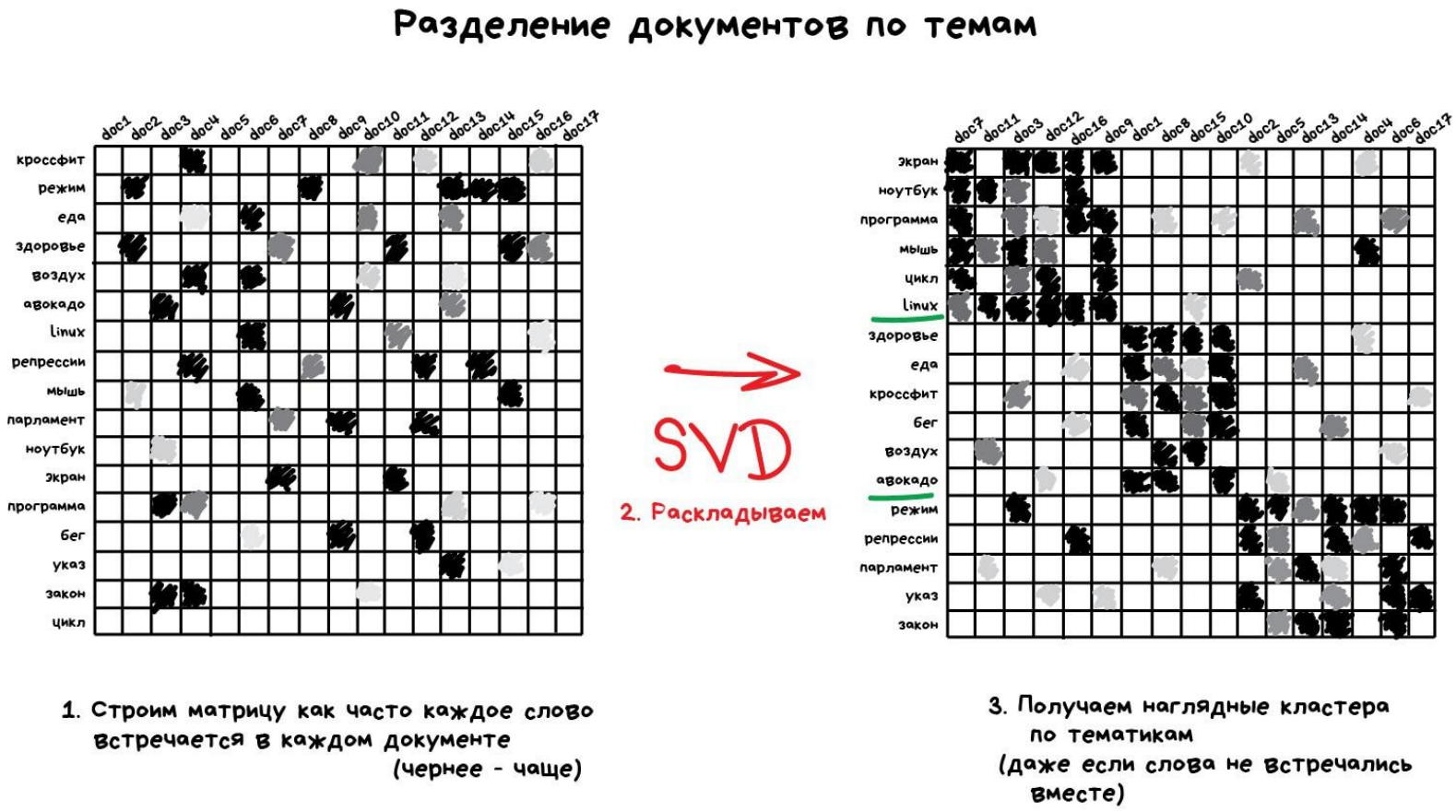
Используют для:

- ❑ Рекомендательные Системы
- ❑ Красивые визуализации
- ❑ Определение тематики и поиска похожих документов
- ❑ Анализ фейковых изображений
- ❑ Риск-менеджмент



https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Обучение без учителя: Уменьшение размерности: LSA



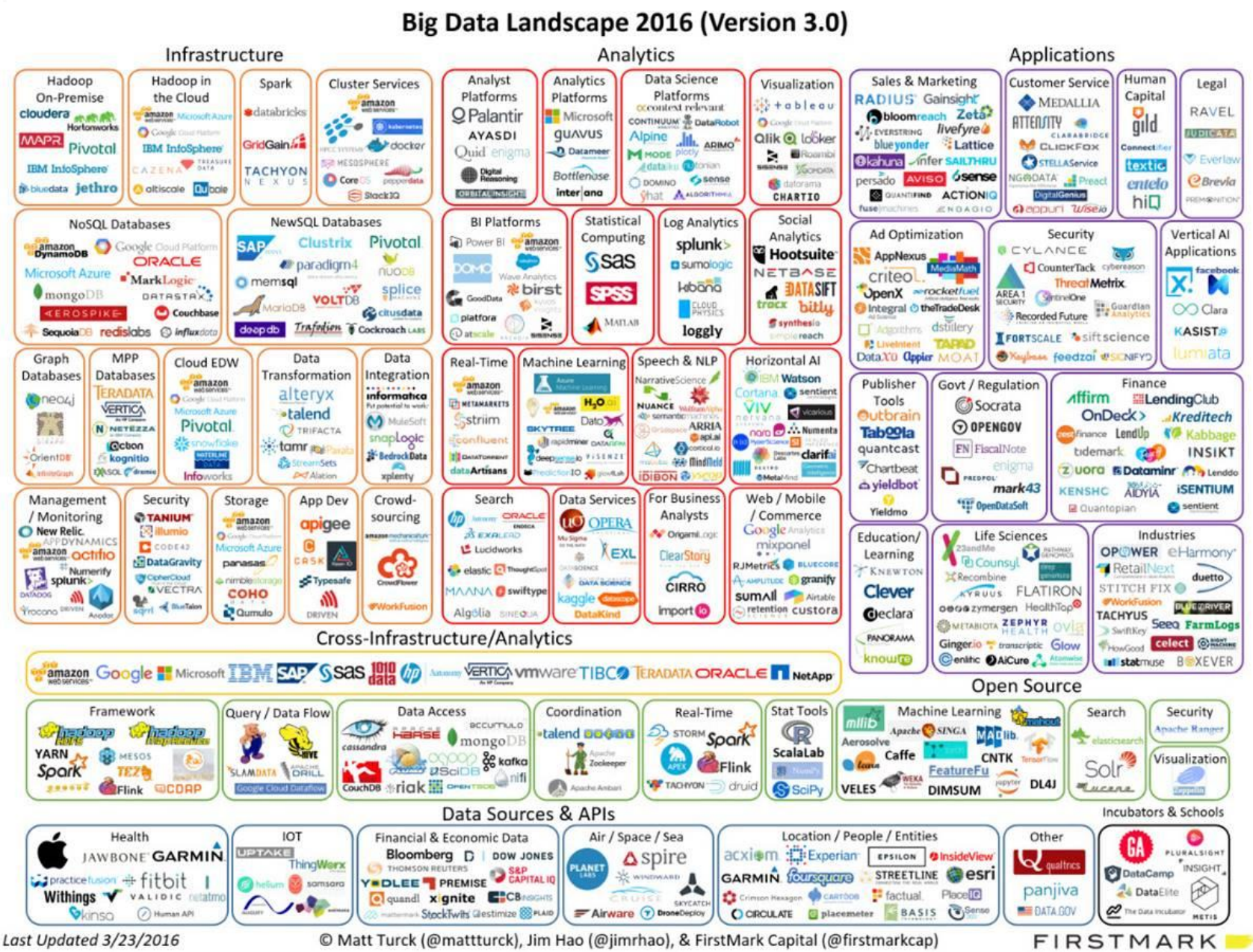
https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

https://vas3k.ru/blog/machine_learning/?fbclid=IwAR3FBfyKACQAmvFhpZRbW3hhiGvN4ghz1fiOxz4GPZZM_3mEAr6FqofUl84

Связь задач, методов ML с бизнес-задачами

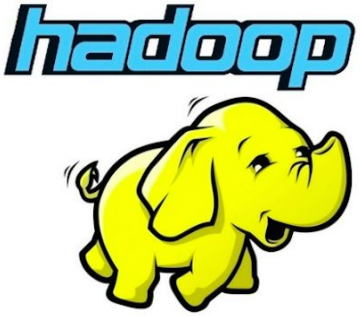
https://blog.zverit.com/machine-learning/2017/11/11/ml-bd-mining-business/

Инструменты больших данных



http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data)

Платформа Hadoop

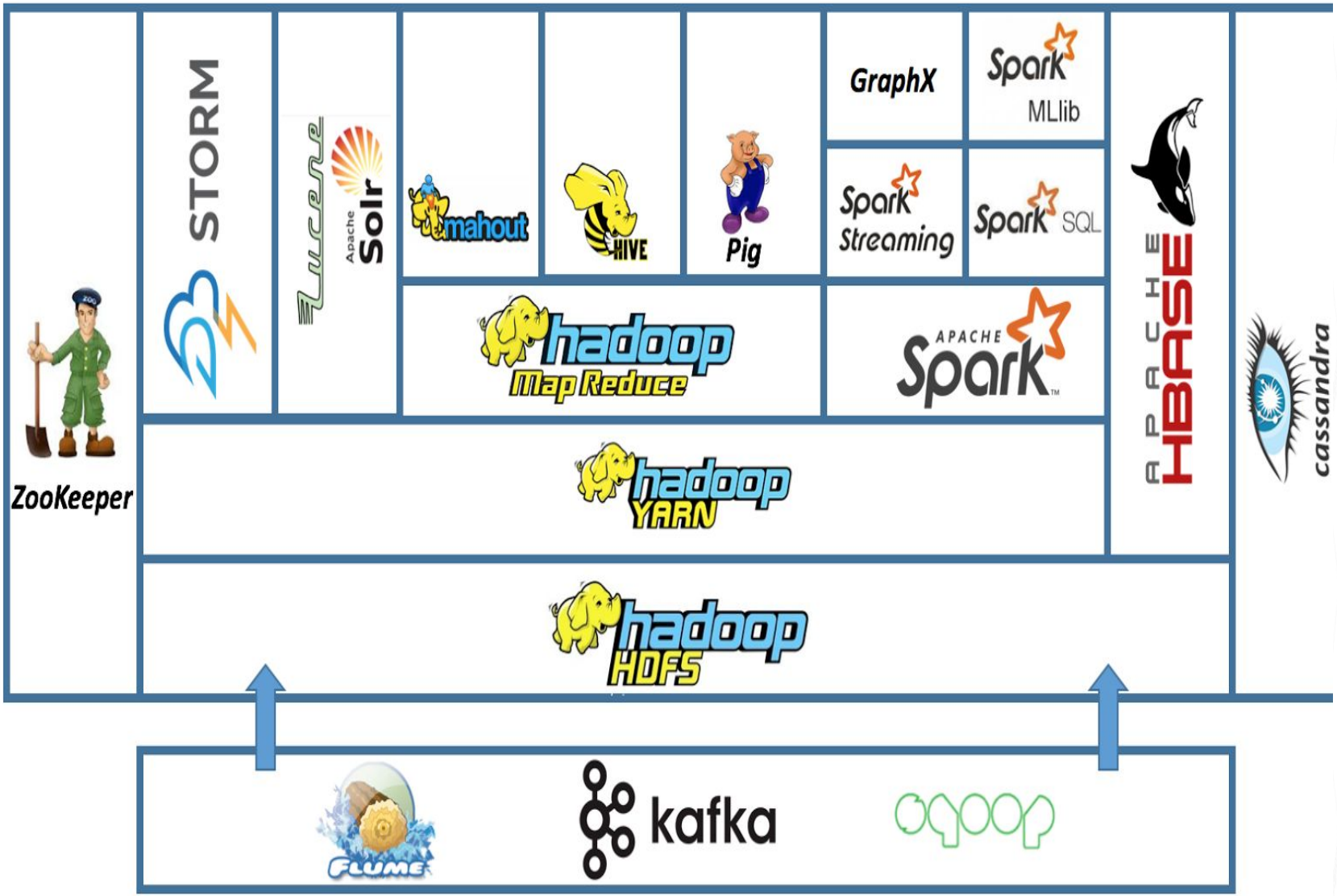


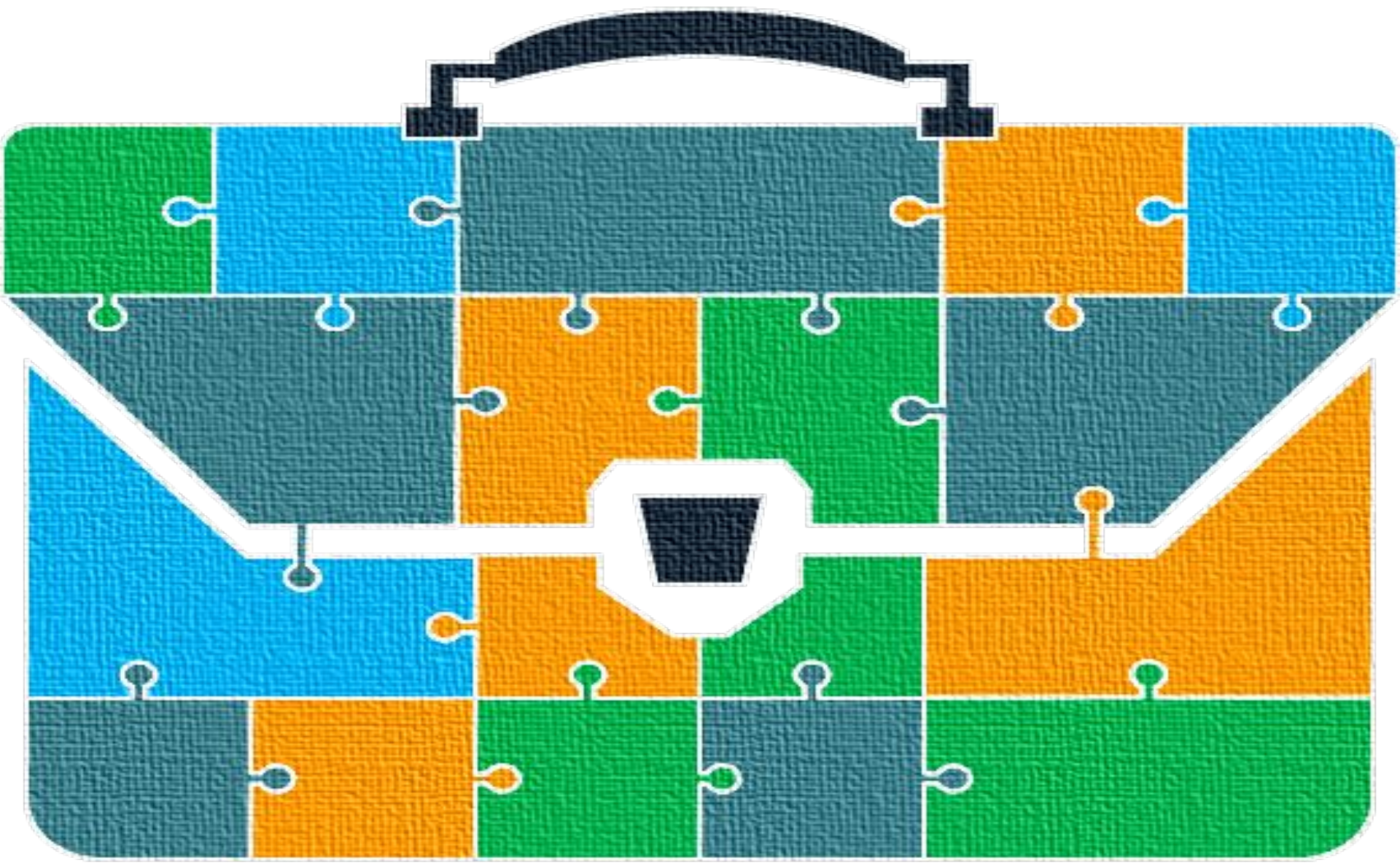
- **Hadoop** – это это *свободно распространяемый* набор программных средств (Software Framework) для разработки и выполнения *распределённых приложений*, предназначенных для *массивно-параллельной обработки* (*Massive Parallel Processing*, MPP) данных.

- Hadoop наиболее эффективен при работе чрезвычайно *большими* объемами данных, но фактически система может применяться и при обработке массивов.

- Термин big data появился несколько *позже* развития концепции платформы Hadoop (2008 vs 2004-2006)

Экосистема Hadoop





Примеры кейсов с большими данными

Командообразование

2002, **бейсбольная команда «Oakland Athletics»** генеральный менеджер Билли Бин, выпускник экономического факультета Йельского университета Питер Брэнд.

- **Задача:** подбор команды с помощью статистического анализа индивидуальных характеристик игроков.
- **Результат:** команда выиграла двадцать матчей подряд.

история легла в основу фильма

https://ланит.рф/press/smi/dmitriy-izmestev-kto-i-kak-zarabatyvaet-na-vashikh-dannykh/

Таргетированный маркетинг

ВОЗМОЖНОСТИ

- Воздействие на клиента в нужное время
- В нужном месте (определение локаций)
- Распознавание интересов и типов пользователей
- Разнообразие каналов оповещения

НАРУЖНАЯ РЕКЛАМА

- как передвигается в течении дня нужная им аудитория,
- куда едет или идёт,
- где стоит на светофорах,
- куда скорее всего смотрит в этот момент

По данным eMarketer, в США уже две трети цифровых рекламных бюджетов закупается по технологии аукционов рекламы в реальном времени, так как они предполагают более точное таргетирование аудитории.

https://www.forbes.ru/tehnologii/336239-put-v-bolshie-dannye-zachem-megafonu-kontrolnaya-dolya-v-mailru-group

Американская сеть магазинов Target и беременная девочка

Торговая сеть в 2012 году узнала о беременности девушки раньше, чем её отец.

Проанализировав покупательские привычки беременных женщин, аналитиками была разработана система прогнозирования беременности.

Ситуация: молодая женщина заходит в магазин и покупает лосьон с кокосовым маслом, сумку для прогулок и ярко-голубой плед, купила витаминов больше, чем обычно, забила в поисковике «самый эффективный способ бросить курить» и т.д.

Вывод: вероятность беременности этой покупательницы — 87%.

Действие: выслать купон со скидкой на детскую кроватку, присыпку, детские бутылочки и т. д. (скидки на товары для детей помещаются среди скидок на другие товары).

За один день акции компании подорожали на 19,16%. Это наибольший внутридневной прирост для Target по меньшей мере с 2000 года.

Target ежегодно тратит около 4 миллионов долларо на содержание аналитического отдела из 50 челове базирующихся в США и Индии.

https://www.retail.ru/cases/na-chto-sposobny-big-data-ili-super-keys-seti-target/

Классификация пользователя

«Перед строкой запроса поисковой системы все честны»

Психологи **Кембриджского университета** изучали, как ставят «лайки» 58 тысяч пользователей фейсбука и на основании этого можно:

- с точностью в 95% установить национальность человека;
- с точностью в 82% отличать пользователя христианина от мусульманина;
- 100% определить сексуальную ориентацию (по информации о предпочтительных ф[illegible], брендах одежды и кулинарных блюдах).

https://www.kp.ru/daily/26692/3716817/

Данные для таргетирования – все действия в сети

Google, Amazon, Apple и Facebook

Google и Facebook: корпорации сегодня контролируют 85% рынка диджитал-рекламы

Компания Bombora (США) отслеживает

- ❑ поисковые запросы,
- ❑ загрузки документов,
- ❑ вебинары,
- ❑ регистрации на выставках,
- ❑ просмотры статей и блогов,
- ❑ потребление видео, лайки в соцсетях и другие свидетельства активности предпринимателей, которые ищут те или иные товары и продукты.

- В этом проекте задействованы Forbes, Aberdeen Group и около 2500 других сайтов, которые предоставляют данные о более чем миллиарде ежемесячных взаимодействий со своими посетителями. 3
- Затем рекламодатели и агентства используют эту информацию для маркетинга и продаж, предлагая ее на основе таргетирования заинтересованным бизнес-компаниям.

Попытки конкуренции:

8 из 10 крупнейших издательских домов Германии работают над созданием единой базы данных о своих читателях. Параллельно данные о пользователях объединяют The Guardian, CNN, Financial Times, Reuters и The Economist.

https://www.sostav.ru/publication/kak-bolshie-dannye-menyayut-rynok-reklamy-v-smi-22969.html

https://www.forbes.ru/tehnologii/336239-put-v-bolshie-dannye-zachem-megafonu-kontrolnaya-dolya-v-mailru-group

Дискриминация в ценообразование или…

- «Некоторые онлайн-ресурсы показывают разные цены на товары в зависимости от того, с какого именно устройства вы зашли — для владельцев iPhone цена часто выше, чем для владельцев смартфона на базе Android»

Анатолий Сморгонский, сооснователь ИТ-холдинга Ambite

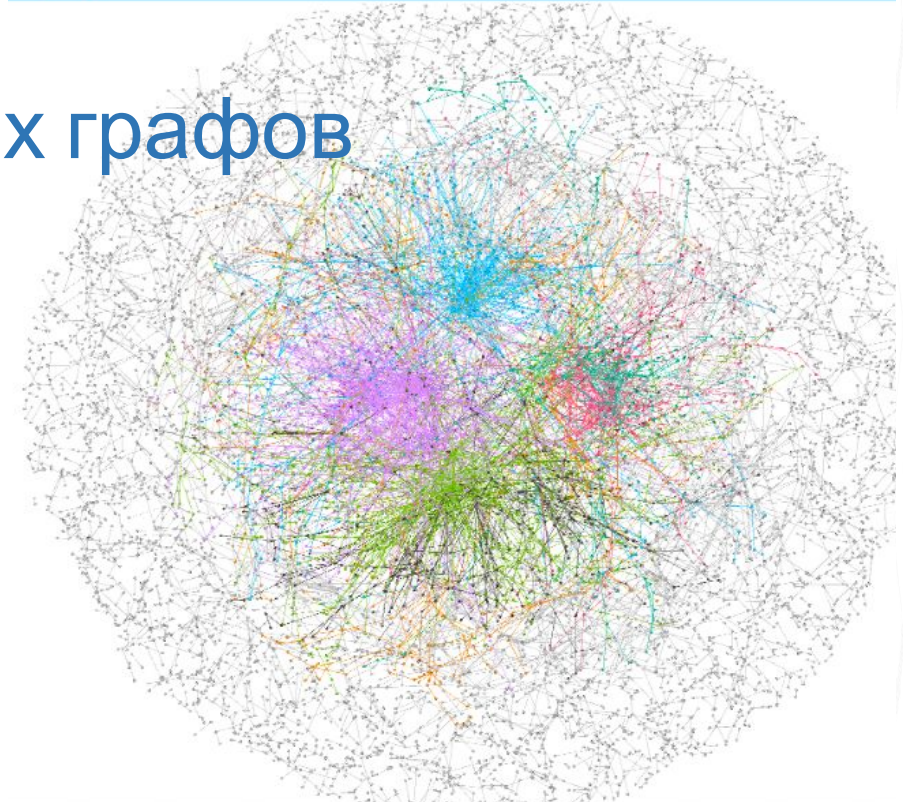
- «Если вы покупаете билет в Питер в один конец, то при покупке обратного билета агрегатор добавит вам 30–40 рублей к стоимости»
- «Tesco анализирует более 30 тысяч категорий — рыбалка, охота, книги и другие хобби и вычисляет, что вы купите в следующий раз.»

Алексей Филатов, руководитель направления профайлинга компании «СёрчИнформ»

- «Сотовый оператор продаёт данные о местонахождении ритейлерам, которые предлагают промоакции, когда клиент находится рядом»

Николай Добровольский, вице-президент компании Parallels

Анализ социальных графов



Задачи:

- выделение лидеров мнений, влияние в группах и из вне;
- выделение сообществ;
- определения членства в группах;
- идентификация пользователя (разных аккаунтов);
- выявление истинных связей между пользователями;
- и т.д.

Количество американских патентных заявок связанных с социальными сетями последние 5 лет росло на **250% каждый год**. Например, метод ценообразования который учитывает положение покупателя в социальном графе (новые телефоны влиятельным узлам социального графа за $0, а остальным за $530).

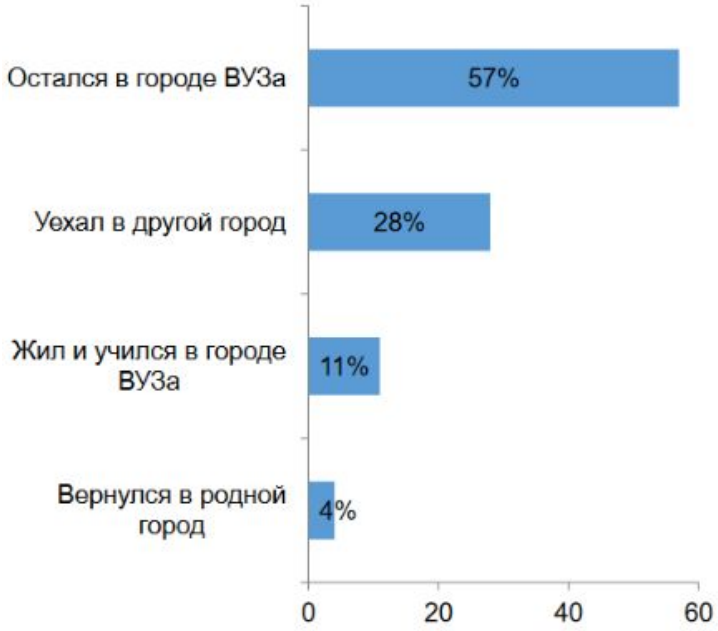
https://habr.com/ru/post/81225/

Национальный исследовательский Томский государственный университет

Университетский консорциум исследователей больших данных

Миграционные модели

Соотношение миграционного статуса выпускников по университетам (обобщенное)

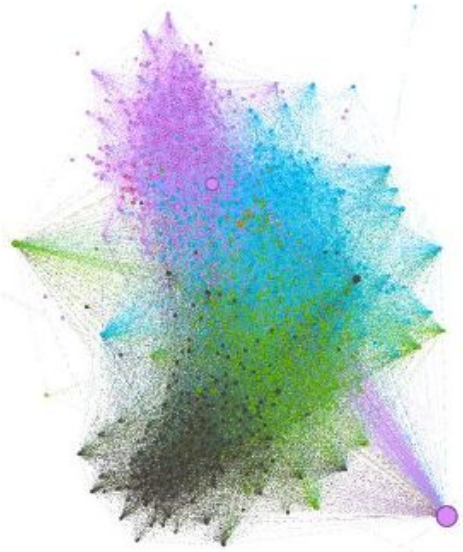


Распределение миграционный моделей поведения выпускников по университетам

Национальный исследовательский Томский государственный университет

Университетский консорциум исследователей больших данных

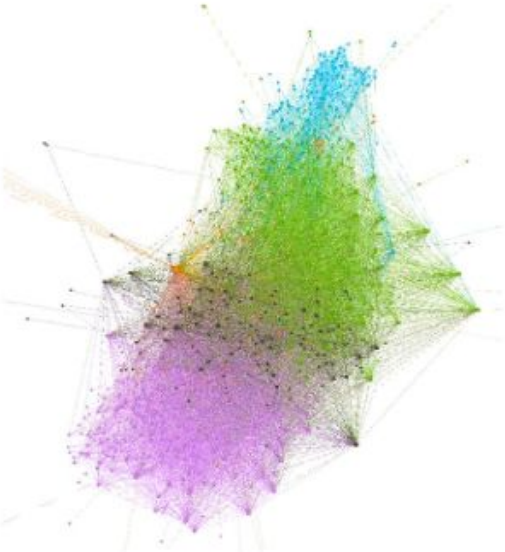
Фонды при образовательных учреждениях



Эндаумент фонд НИТУ «МИСиС»

- 1911 участников
- 12% изолянтов
- Ориентация на внутренние связи
- Отсутствие ярко выраженных лидеров

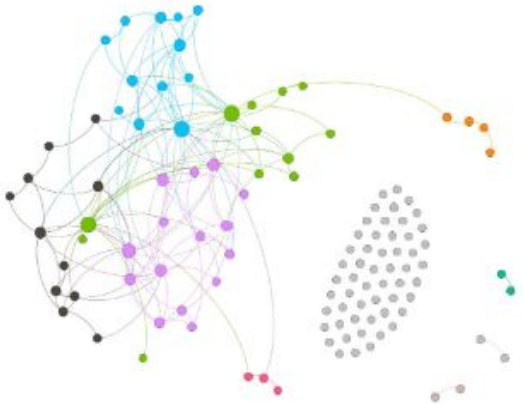
Плотная коммуникативная сеть



Эндаумент МГИМО

- 1166 участников
- 19% изолянтов
- Сильные не пересекающиеся кластеры
- Ориентация на внутреннюю организацию
- Отсутствие ярко выраженных лидеров

Плотная коммуникативная сеть



Эндаумент выдающихся выпускников НИУ ВШЭ

- 117 участников
- 51% изолянтов
- Сильные разделенные кластеры
- Ориентация на внутреннюю организацию

Система коммуникации практически отсутствует

Методы NLP, Text Mining

Обработка естественного языка (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста.

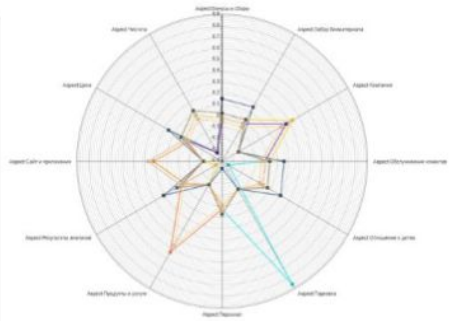
Интеллектуальный анализ текстов (text mining) — направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка.

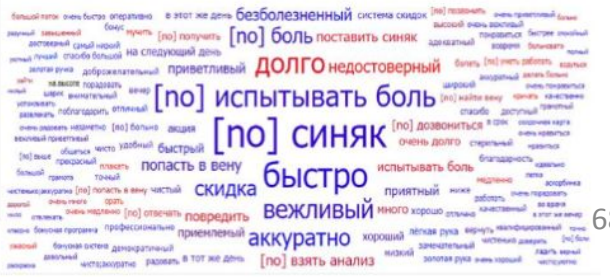
- О чем говорят?
- Кто говорит?
- Как говорят?

ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

АНАЛИЗ ОТЗЫВОВ О КОМПАНИИ

Megaputer
Your Knowledge Partner



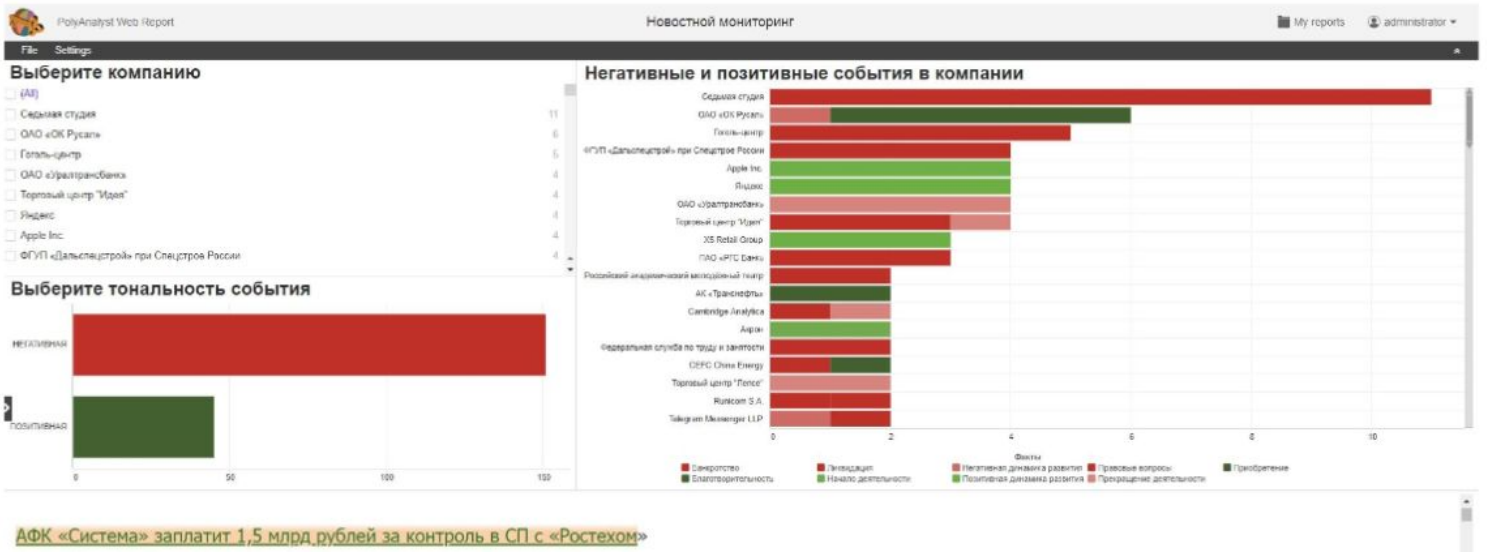


PolyAnalyst: Примеры Аналитических Решений

- Структуризация и контроль за содержанием учебных программ
- Автоматизация извлечения информации о контрагентах
- Сверка счетов и накладных
- Извлечение информации из юридических документов
- Управление номенклатурой материально-технических ресурсов
- Анализ новостного потока
- Роботизация информационных процессов в медицинском страховании
- Анализ отзывов клиентов
- Анализ данных по ремонту и обслуживанию оборудования

ВИЗУАЛИЗАЦИЯ ТОНАЛЬНОСТИ СОБЫТИЙ, ПРОИСХОДЯЩИХ С КОМПАНИЕЙ

АНАЛИЗ НОВОСТНОГО ПОТОКА



Примеры (зарубежные кейсы)

- HSBC повышает безопасность клиентов пластиковых карт. Компания утверждает, что в 10 раз улучшила распознавание мошеннических операций и в 3 раза – защиту от мошенничества в целом.

- Суперкомпьютер Watson, разработанный IBM, анализирует финансовые транзакции в режиме реального времени. Это позволяет сократить частоту ложных срабатываний системы безопасности на 50% и выявить на 15% больше мошеннических действий.

- Procter&Gamble проводит с использованием Big Data маркетинговые исследования, более точно прогнозируя желания клиентов и спрос новых продуктов.

- Министерство труда Германии добивается целевого расхода средств, анализируя большие данные при обработке заявок на пособия. Это помогает направить деньги тем, кто действительно в них нуждается (оказалось, что 20% пособий выплачивались нецелесообразно). Министерство утверждает, что инструменты Big Data сокращают затраты на €10 млрд.

https://invlab.ru/texnologii/bolshie-dannye/

Примеры (отечественные кейсы)

- **Яндекс.** Это корпорация, которая управляет одним из самых популярных поисковиков и делает цифровые продукты едва ли не для каждой сферы жизни. Для Яндекс Big Data – не инновация, а обязанность, продиктованная собственными нуждами. В компании работают алгоритмы таргетинга рекламы, прогноза пробок, оптимизации поисковой выдачи, музыкальных рекомендаций, фильтрации спама.

- **Мегафон.** Телекоммуникационный гигант обратил внимание на большие данные примерно пять лет назад. Работа над геоаналитикой привела к созданию готовых решений анализа пассажироперевозок. В этой области у Мегафон есть сотрудничество с РЖД.

- **Билайн.** Этот мобильный оператор анализирует массивы информации для борьбы со спамом и мошенничеством, оптимизации линейки продуктов, прогнозирования проблем у клиентов. Известно, что корпорация сотрудничает с банками – оператор помогает анонимно оценивать кредитоспособность абонентов.

- **Сбербанк.** В крупнейшем банке России супермассивы анализируются для оптимизации затрат, грамотного управления рисками, борьбы с мошенничеством, а также расчёта премий и бонусов для сотрудников. Похожие задачи с помощью Big Data решают конкуренты: Альфа-банк, ВТБ24, Тинькофф-банк, Газпромбанк.

https://invlab.ru/texnologii/bolshie-dannye/

«Анонимности в сети нет»

один из руководителей SocialDataHub Артур Хачуян

Российская компания SocialDataHub

- в считанные часы смогла опознать террориста-смертника, подорвавшего в апреле поезд в питерской подземке по фотографии головы предполагаемого преступника нашли шесть аккаунтов в социальных сетях и обнаружили связь с другим террористом, который расстрелял приёмную ФСБ в Хабаровске.

- за три дня до протестных акций оппозиции в Москве выложила исследование «Сколько человек придут на митинг 12 июня и кто они».

- на сайтах «для взрослых» они собрали фотографии 27 856 женщин и 1387 мужчин, которые предлагают любовь за деньги, отыскали реальные аккаунты в социальных сетях и составили своеобразный рейтинг вузов.

«…один из российских банков проанализировал свои данные о клиентах и убедился, что на 99% может определить, есть ли у человека любовница, на основании его покупок»

Сергей Сошников, заместитель генерального директора информационно-финансового блока компании «Актив»

https://www.kp.ru/daily/26692/3716817/

https://www.kommersant.ru/doc/3579697

Насколько законно собирать данные о людях?

- Федеральный закон "О персональных данных" эксперты оценивают как "размытый".
- Например, получение и использование телефона клиента с помощью технологий трактуется по-разному: в одном суде могут посчитать эту информацию персональной, в другом — нет.
- Нельзя использовать данные из переписки пользователей, данные из кредитных историй, из медицинских карт.
- Все, что в социальных сетях можно увидеть своими глазами, — можно использовать, но были и прецеденты по этому вопросу.

https://tass.ru/ekonomika/5138017

О чем спорят «ВКонтакте» и Double Data

- **2017** ВКонтакте судится с ООО «Дабл», требуя запрета для этой компании на использование сведений о пользователях соцсети (фамилий, имен, мест работы и учебы и другой открытой информации). Компания «Дабл» использовала их в коммерческих целях, например, для оценки кредитоспособности заемщиков для банков.

  октябрь 2017 Московский арбитражный суд отклонил все требования ВКонтакте к Double Data.
  январь 2018 Девятый арбитражный апелляционный суд удовлетворил требования ВКонтакте, обязав компанию Double Data прекратить использовать данные пользователей социальной сети.
  лето 2018 суд по интеллектуальным правам отменил решение о запрете использования открытых данных из социальной сети ВКонтакте и направил дело на новое рассмотрение.
  август 2019 «Сегодня это ходатайство было отклонено по причине того, что требования пользователей подсудны суду общей юрисдикции, имеют иную природу и должны рассматриваться в отдельном судебном процессе, — продолжает он. — Определения пока нет. Полагаю, мы будем обжаловать его. И я также не исключаю вероятности того, что мы подадим самостоятельный групповой иск к «Дабл».

- Судебное разбирательство в США, где стартап HiQ Labs (использовал открытые данные пользователей LinkedIn, чтобы прогнозировать поведение наемных работников) выиграл у LinkedIn, «отстояв право стартапов использовать публичные данные социальных сетей».

- Российский рекрутинговый сервис HeadHunter подал иск против сервиса автоматизации рекрутинга «Робот Вера» (компания «Стафори») за использование базы данных hh.ru без ведома кадрового портала. Всего на портале находится более 34 млн резюме. По собственным данным «Робота Веры», сервис за 15 месяцев провел 1,5 млн телефонных, 4000 видеоинтервью и обработал 1 млн резюме.

  Московский городской суд отклонил иск из-за отсутствия доказательства использования ответчиком данных HeadHunter.
  В 2018 году у HeadHunter уже был опыт судебного разбирательства по подобному делу — в январе портал выиграл иск к сервису для автоматизации рекрутинга FriendWork Recruiter, который также использовал резюме с hh.ru.

tps://www.rbc.ru/technology_and_media/04/09/2019/5d6e5f6d9a794784dea7c2e2

Кто владеет информацией, тот владеет миром.

Натан Ротшильд

Именно то, как вы собираете, организуете и используете информацию, определяет, победите вы или проиграете.

Билл Гейтс

До девяноста пяти процентов всей информации, которую воспринимают твои глаза и уши каждый день, заранее отобраны по чьей-то воле и оплачены из чьего-то кармана.

Харуки Мураками

Не забывай: информация не есть знание, знание не есть мудрость, мудрость не есть истина, истина не есть красота, красота не есть любовь, любовь не есть музыка, музыка лучше всего, что есть на свете.

Фрэнк Заппа

Спасибо за внимание!