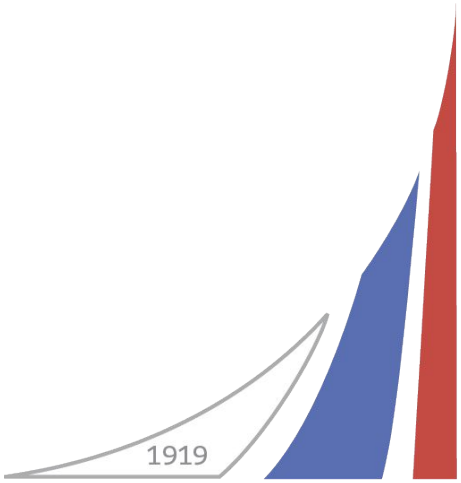
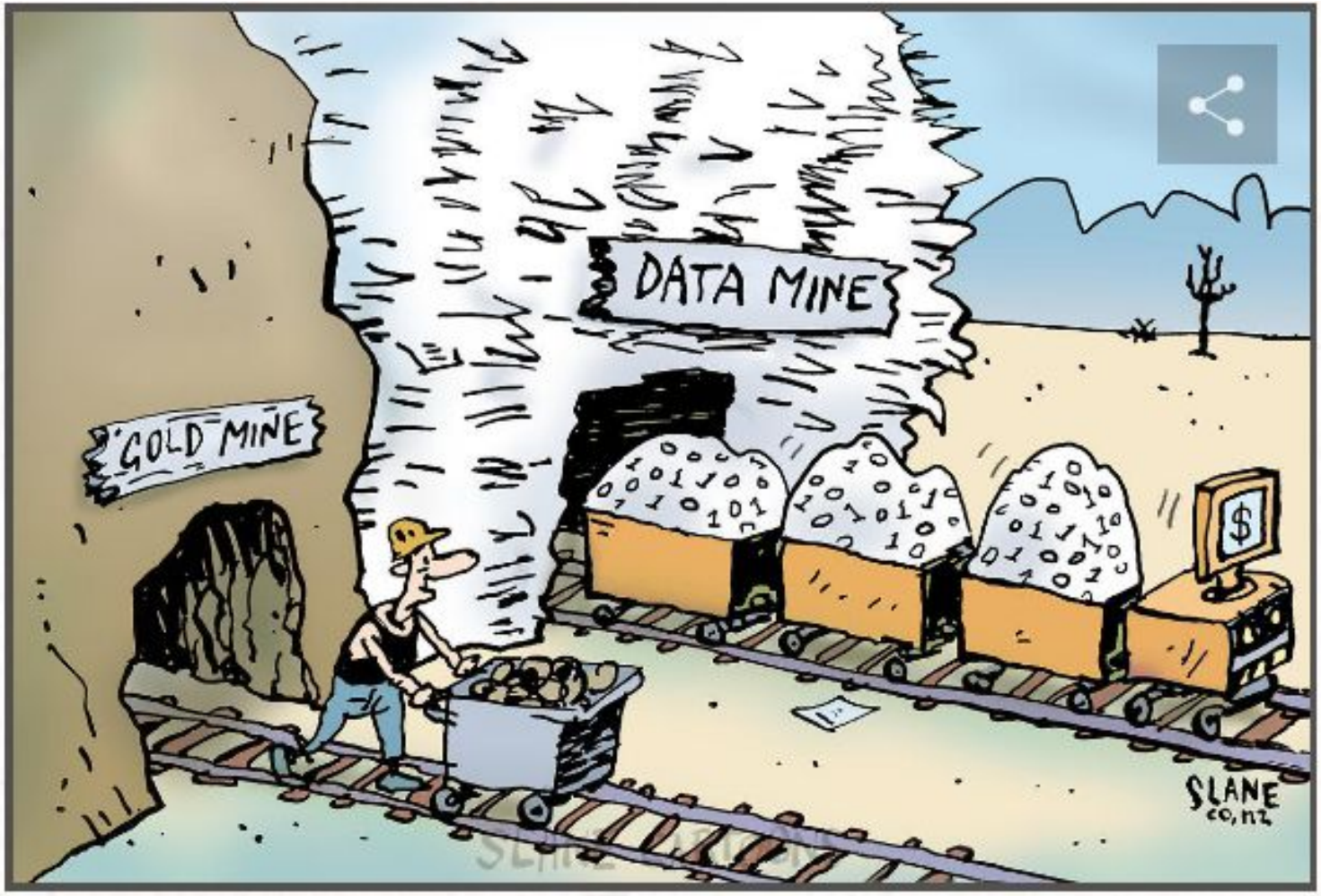


ФИНАНСОВЫЙ УНИВЕРСИТЕТ

ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Введение в большие данные

Корчагин Сергей Алексеевич, кандидат физико-математических наук,
Заместитель руководителя Департамента анализа данных и машинного обучения
SAKorchagin@fa.ru
Москва, 2021



https://www.slanecartoon.com/-/galleries/privacy-1/-/medias/f882b7b9-5893-4572-99f1-1c17c7ebbe81-data-mine/share

Большие данные (Big Data, биг дата) — это структурированные и неструктурированные данные огромных объемов и разнообразия, а также методы их обработки, которые позволяют распределено анализировать информацию.

Клиффорд Линч, 2008

Business Intelligence = BI = Бизнес – аналитика (rus) — это набор IT-технологий для сбора, хранения и анализа данных, позволяющих предоставлять пользователям достоверную аналитику в удобном формате, на основе которой можно принимать эффективные решения для управления бизнес-

Входные данные: разнородные отчеты

Отчет первого магазина

Товар	Количество	Примечания
Бум Тул	100 шт.	Договор 100500
Кока-кола литровая	84 шт.	Закупить ещё
Кока-кола полу-литровая	12 шт.	-
Шетки зуб	145 шт.	-
Прочее	134 шт.	-

Отчет второго магазина

Продукция	Кол-во	Комментарии
Бумага туалетная	105	Обсудить скидки
Coca-cola 1л	14	Нет
Coca-cola 0.5л	12	Нет
Зубные щётки	54	Нет
Другое	42	Разобрать

Отчет третьего магазина

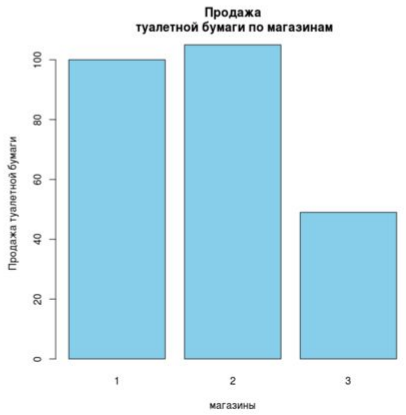
Изделие	Продано	Примечание
Туалетная бумага	49 рулонов	Закупить двуслойные
Кола литровая	14 бутылок	Обсудить опт
Кола поллитра	12 бутылок	
Щётка зубная	54 упаковок	
Остальное	42 штук	

Выходные данные: общая таблица и визуализация

Сводная таблица

Товар	Магазин 1	Магазин 2	Магазин 3
Туалетная бумага	100	105	49
Кола 1л	84	14	14
Кола 0.5 л	12	12	12
Зубные щётки	145	54	54
Прочее	134	42	42

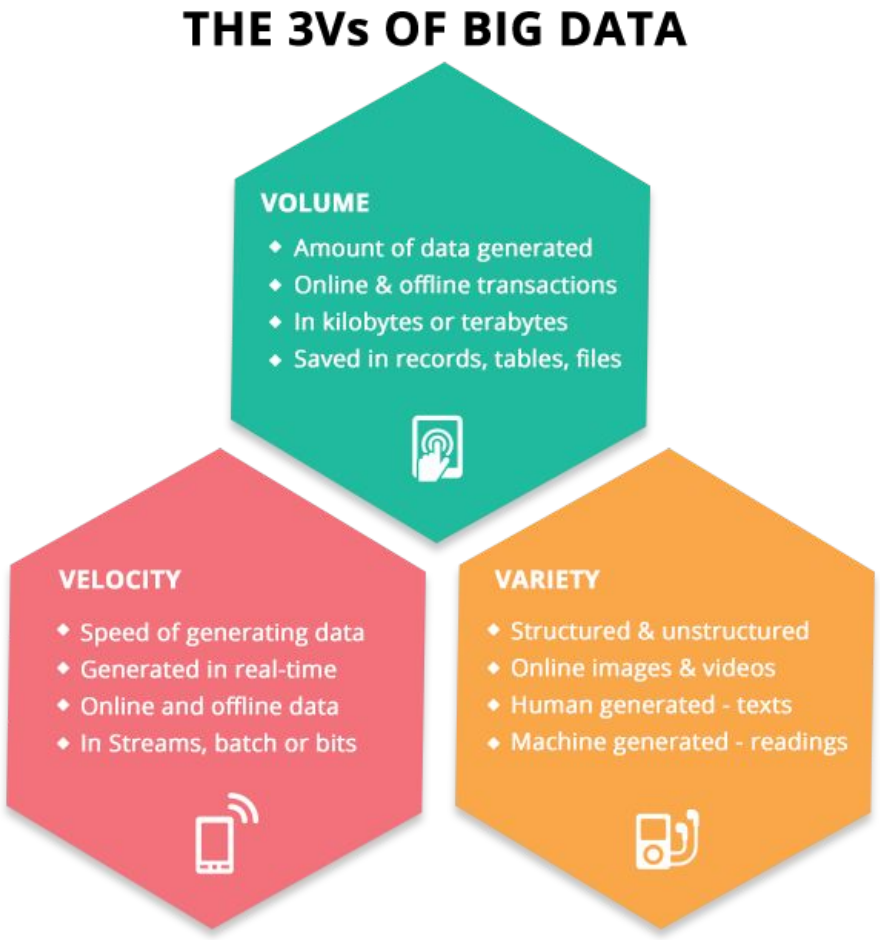
Аналитика по сводным данным



Традиционная аналитика	Big data аналитика
Постепенный анализ небольших пакетов данных	Обработка сразу всего массива доступных данных
Редакция и сортировка данных перед обработкой	Данные обрабатываются в их исходном виде
Старт с гипотезы и ее тестирования относительно данных	Поиск корреляций по всем данным до получения искомой информации
Данные собираются, обрабатываются, хранятся и лишь затем анализируются	Анализ и обработка больших данных в реальном времени, по мере поступления

Правило VVV — три признака или свойства, которыми большие данные должны обладать:

1. Volume – объем
2. Velocity – скорость
3. Variety - многообразие



Функции и задачи больших данных

Функция	Задача
Big Data — собственно массивы необработанных данных	Хранение и управление большими объемами постоянно обновляющейся информации
Data mining — процесс обработки и структуризации данных, этап аналитики для выявления закономерностей	Структурирование разнообразных сведений, поиск скрытых и неочевидных связей для приведения к единому знаменателю
Machine learning — процесс машинного обучения на основе обнаруженных связей в процессе анализа	Аналитика и прогнозирование на основе обработанной и структурированной информации

Источники больших данных

- **Интернет** — соцсети, блоги, СМИ, форумы, сайты, интернет вещей (IoT).

- **Корпоративные данные** — транзакционная деловая информация, архивы, базы данных.

- **Показания устройств** — датчиков, приборов, а также метеорологические данные, данные сотовой связи и т.д. -

- Горизонтальная масштабируемость
- Отказоустойчивость
- Локальность данных

Предсказать победителя Оскара!

Найти военную базу НАТО

Диагностировать беременность

Анализ данных опросов:

- Мониторинг общественного мнения и анализ социально-экономической ситуации
- Определение проблем, формирующих кризисную ситуацию
- Анализ реакции населения на внедрение различных федеральных и региональных программ
- Анализ экономического положения и уровня жизни населения

Предвыборные исследования

- Анализ эффективности политической рекламы
- Анализ средств массовой информации
- Выявление наиболее эффективных средств влияния на мнения различных групп избирателей
- Диагностика предвыборной ситуации
- Анализ основных проблем избирателей

Общественная безопасность

- Анализ преступности
- Отслеживание уровня рецидивизма

Образование

- Планирование школьных округов
- Отслеживание успеваемости учащихся, выявление факторов
способствующих повышению успеваемости
- Администрирование - контроль за уровнем выполнения обязательны

Трудоустройство

- Анализ рынка труда - понимание состава и структуры рабочей силы
- Анализ заявлений о приеме на работу - разработка профилей претендентов.

Анализ прибыли

- Оценка соответствия размеров уплаченных налогов и имущества
- Анализ мошенничеств

Здравоохранение

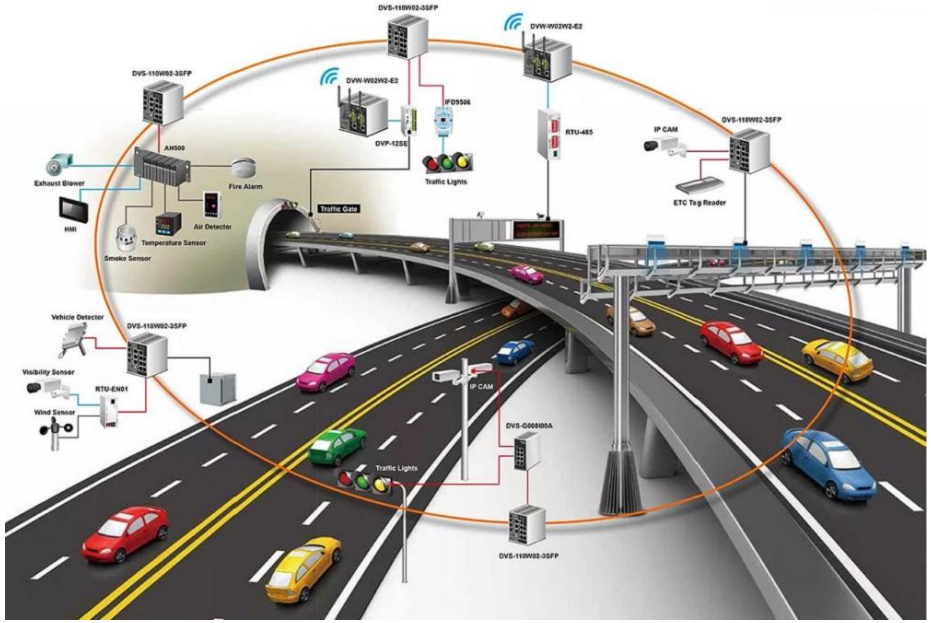
- Отслеживание болезней и создание отчетов о случаях заболеваний
- Эпидемиология - выявление причин заболеваний и территории их
распространения, а также контроль заболеваемости
- Медицинская помощь - определение профилей тех, кому часто требуется медицинская помощь
- Профилактика
медицинского вм

Окружающая среда

- Анализ экосистем - выяснение факторов, влияющих на здоровье
экосистемы
- Оценка качества воды/воздуха - осуществление контроля за
соотв
станд

Транспорт

- Планирование наиболее эффективных маршрутов для лучшей организации транспортных и пассажирских потоков
- Создание отчетов о дорожных происшествиях для выяснения факторов, влияющих на происшествия
- Моделирование программ поддержания надлежащего состояния дорожного покрытия, прогнозирование возможного ремонта дорог.



Стратегическое планирование

- Анализ удовлетворенности клиентов и изучения изменений потребностей общественности
- Оценка программ - понимание факторов успешной реализации программы
- Профилирование населения - более эффективное направление действия программы на определенные слои населения
- Анализ затрат - выявления наиболее эффективных программ
- Анализ результатов выполнения программ

- Создание точных портретов целевых потребителей.
- Предсказание реакции потребителей на маркетинговые сообщения.
- Максимальная персонализация рекламных сообщений.
- Увеличение кросс-продаж, повторных продаж, ремаркетинга.
- Поиск и определение причин популярности востребованных товаров и продуктов.
- Совершенствование продуктов и услуг, повышение лояльности клиентов.
- Повышение качества обслуживания.
- Предупреждение мо
- Снижение издержек циками и клиентами.

Поставщики инфраструктуры — решают задачи хранения и предобработки данных.

IBM Microsoft ORACLE

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

Yandex Data Factory CleverDATA glowbyte

Системные интеграторы — компании, которые внедряют системы анализа больших данных.

ФОРС КРОК

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

СБЕРБАНК ГАЗПРОМ МТС

Google

Yandex
Data Factory

Google
BigQuery

Cloud
Bigtable

@ mail.ru
group

Рамблер/

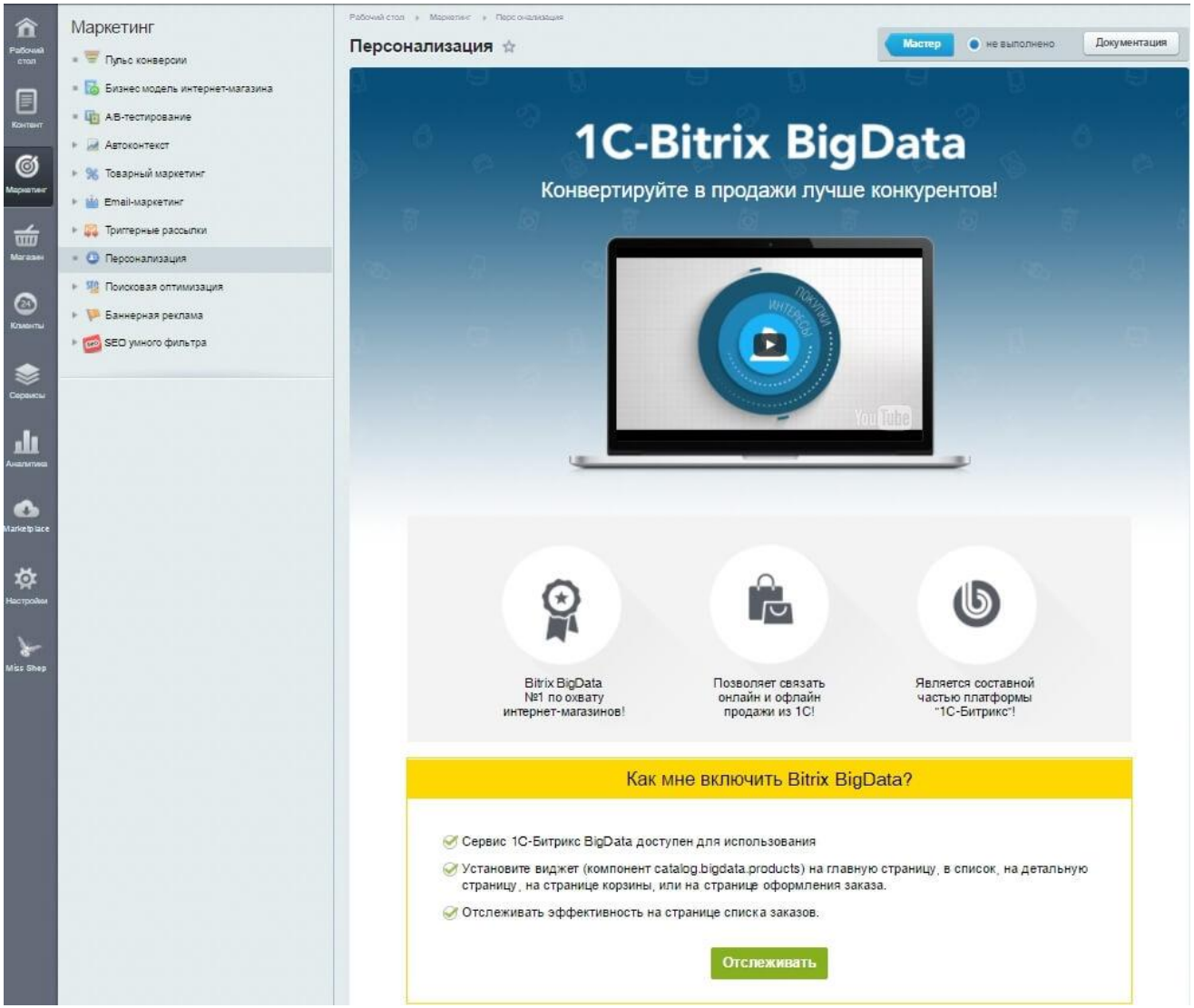
Выгоды использования Big Data в бизнесе

- Упрощается планирование.
- Увеличивается скорость запуска новых проектов.
- Повышаются шансы проекта на востребованность.
- Можно оценить степень удовлетворенности пользователей.
- Проще найти и привлечь целевую аудиторию.
- Ускоряется взаимодействие с клиентами и контрагентами.
- Оптимизируются интеграции в цепи поставок.
- Повышается качество клиентского сервиса, скорость взаимодействия.
- Повышается лояльность текущих клиентов.

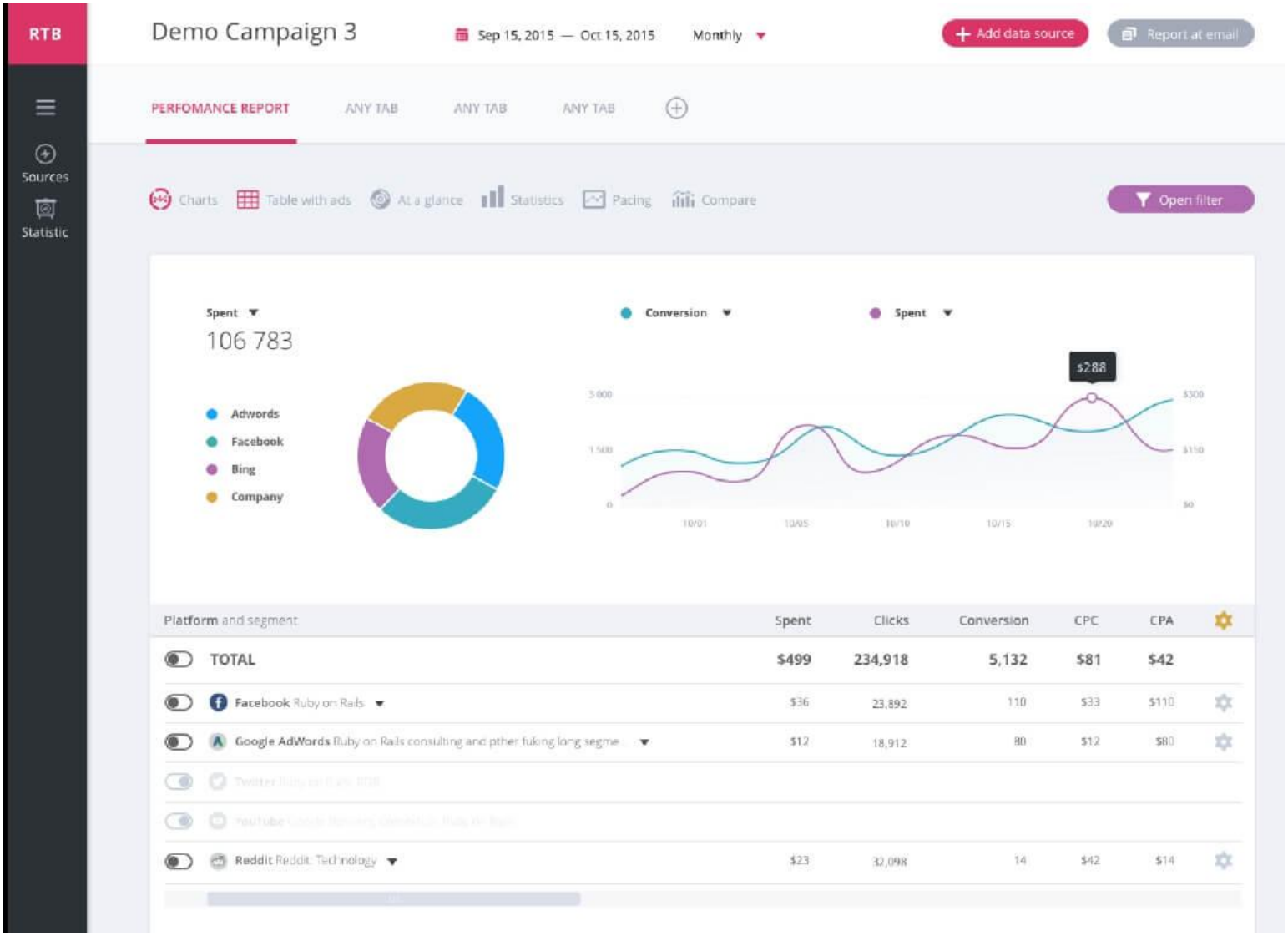
Драйверы и ограничители Big Data в России

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,	Рост цен на импортную продукцию

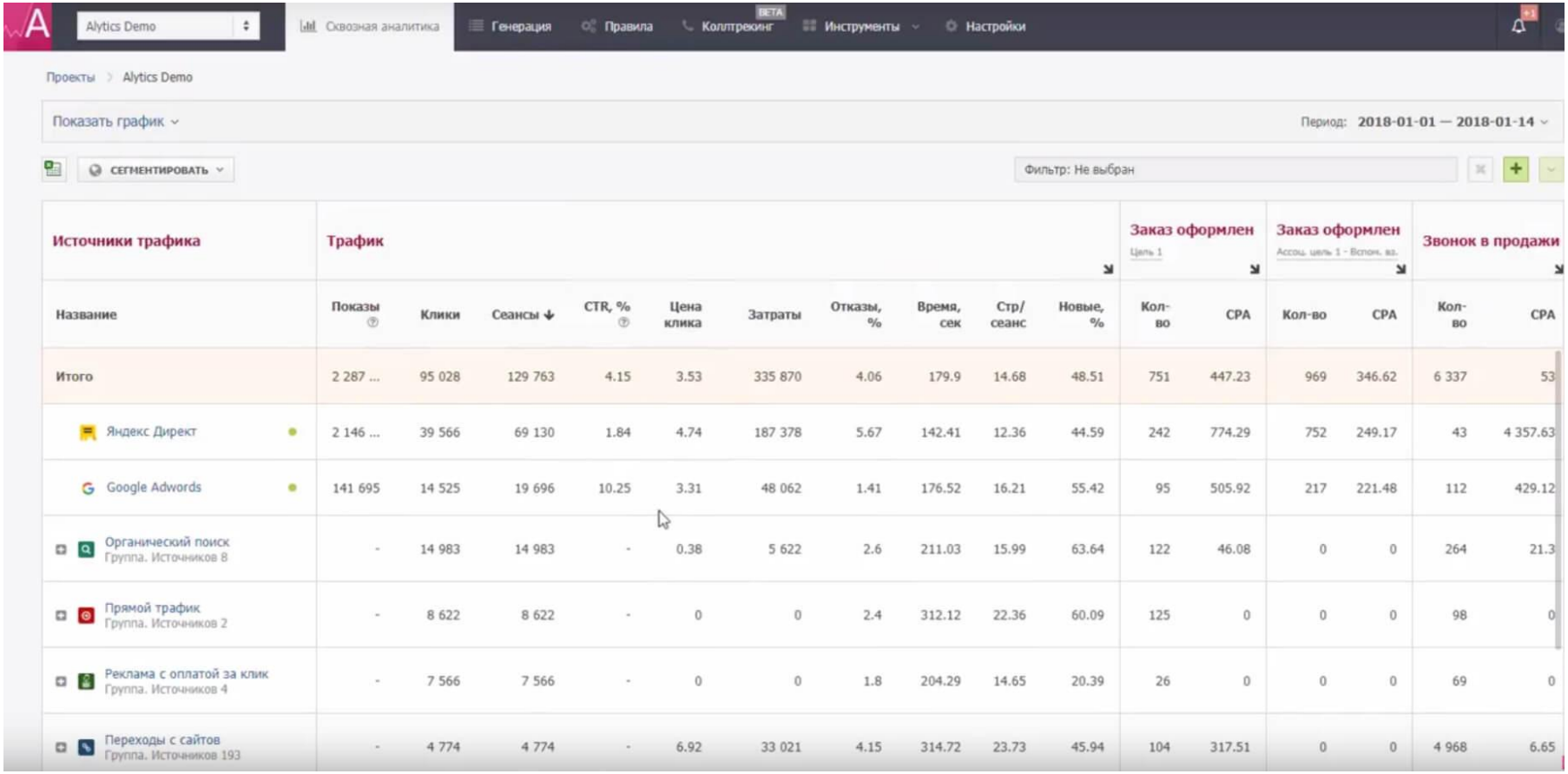
«1С-Битрикс BigData»



RTB Media

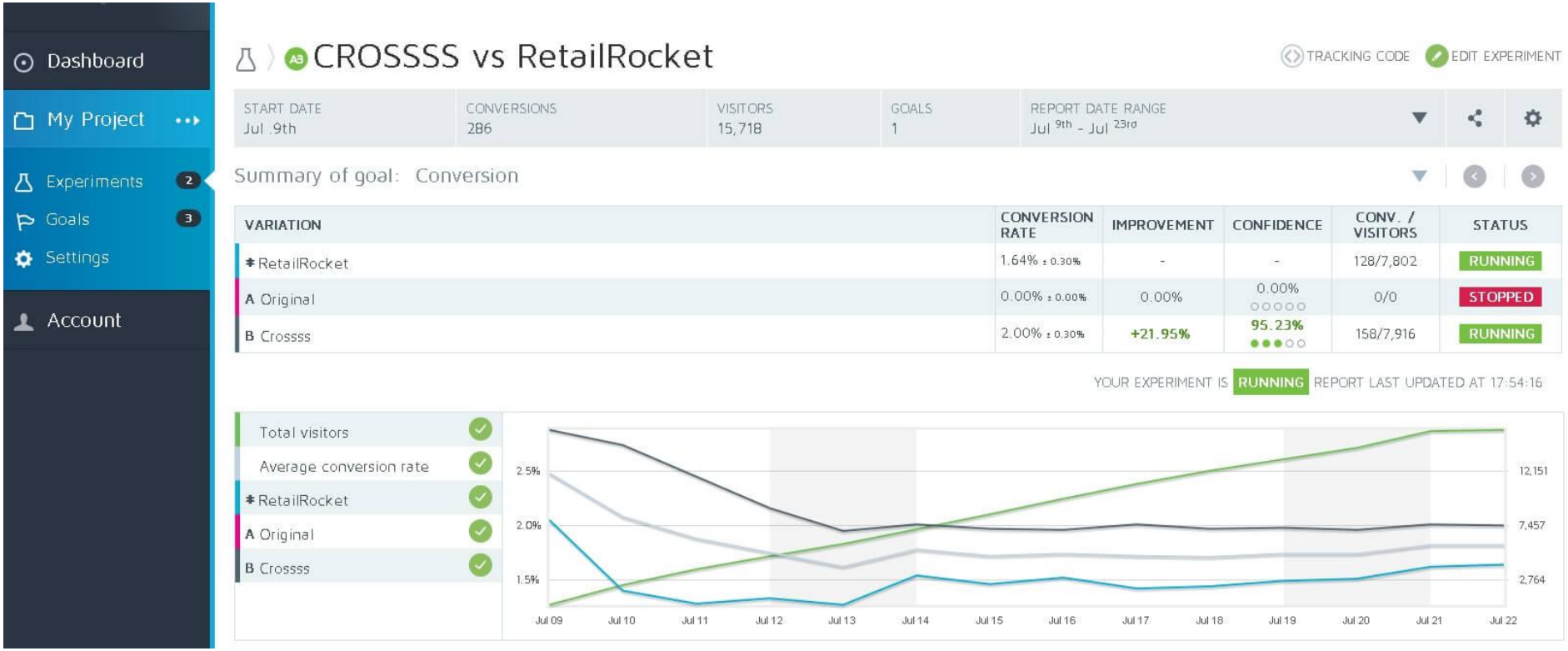

RTB
Demo Campaign 3
Sep 15, 2015 — Oct 15, 2015
Monthly
Add data source
Report at email
PERFOMANCE REPORT
ANY TAB
ANY TAB
ANY TAB
Sources
Statistic
Charts
Table with ads
At a glance
Statistics
Pacing
Compare
Open filter
Spent
106 783
Adwords
Facebook
Bing
Company
Conversion
Spent
$288
Platform and segment
Spent
Clicks
Conversion
CPC
CPA
TOTAL
$499
234,918
5,132
$81
$42
Facebook Ruby on Rails
$36
23,892
110
$33
$110
Google AdWords Ruby on Rails consulting and other fuking long segme
$12
18,912
80
$12
$80
Reddit Reddit Technology
$23
32,098
14
$42
$14

Alytics



Источники трафика	Трафик										Заказ оформлен (Цель 1)		Заказ оформлен (Ассоц. цель 1 - Вспом. вз.)		Звонок в продажи	
Название	Показы	Клики	Сеансы ↓	CTR, %	Цена клика	Затраты	Отказы, %	Время, сек	Стр/сеанс	Новые, %	Кол-во	CPA	Кол-во	CPA	Кол-во	CPA
Итого	2 287 ...	95 028	129 763	4.15	3.53	335 870	4.06	179.9	14.68	48.51	751	447.23	969	346.62	6 337	53
Яндекс Директ	2 146 ...	39 566	69 130	1.84	4.74	187 378	5.67	142.41	12.36	44.59	242	774.29	752	249.17	43	4 357.63
Google Adwords	141 695	14 525	19 696	10.25	3.31	48 062	1.41	176.52	16.21	55.42	95	505.92	217	221.48	112	429.12
Органический поиск Группа. Источников 8	-	14 983	14 983	-	0.38	5 622	2.6	211.03	15.99	63.64	122	46.08	0	0	264	21.3
Прямой трафик Группа. Источников 2	-	8 622	8 622	-	0	0	2.4	312.12	22.36	60.09	125	0	0	0	98	0
Реклама с оплатой за клик Группа. Источников 4	-	7 566	7 566	-	0	0	1.8	204.29	14.65	20.39	26	0	0	0	69	0
Переходы с сайтов Группа. Источников 193	-	4 774	4 774	-	6.92	33 021	4.15	314.72	23.73	45.94	104	317.51	0	0	4 968	6.65

Crossss


Dashboard
My Project
Experiments 2
Goals 3
Settings
Account
CROSSSS vs RetailRocket
TRACKING CODE
EDIT EXPERIMENT
START DATE Jul 9th
CONVERSIONS 286
VISITORS 15,718
GOALS 1
REPORT DATE RANGE Jul 9th - Jul 23rd
Summary of goal: Conversion
VARIATION
CONVERSION RATE
IMPROVEMENT
CONFIDENCE
CONV. / VISITORS
STATUS
RetailRocket 1.64% ± 0.30% - - 128/7,802 RUNNING
A Original 0.00% ± 0.00% 0.00% 0.00% 0/0 STOPPED
B Crossss 2.00% ± 0.30% +21.95% 95.23% 158/7,916 RUNNING
YOUR EXPERIMENT IS RUNNING REPORT LAST UPDATED AT 17:54:16
Total visitors
Average conversion rate
RetailRocket
A Original
B Crossss
2.5%
2.0%
1.5%
12,151
7,457
2,764
Jul 09
Jul 10
Jul 11
Jul 12
Jul 13
Jul 14
Jul 15
Jul 16
Jul 17
Jul 18
Jul 19
Jul 20
Jul 21
Jul 22

- NoSQL;
- MapReduce;
- Hadoop;
- R;
- Python;
- Аппаратные решения.

Основными методами и техниками анализа, применимыми к Большим данным, являются следующие:

- Методы класса Data Mining
- Краудсорсинг
- Смешение и интеграция данных
- Машинное обучение
- Визуализация аналитических данных

Data Mining


1
01
100
110
001
0110110010
0110011010101
0101010010111
00110101010110100
110
1101100

Data Mining – это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий

Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Термин введён
Григорием Пятецким-Шапиро
в 1989 году.

Data Mining – мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная *статистика*, *распознавание образов*, *искусственный интеллект*, теория баз *данных*

1	• ассоциативные правила
2	• деревья решений
3	• кластеры
3	• математические функции

Обзор существующих методов Data Mining

- искусственные нейронные сети
- деревья решений, символьные правила
- методы ближайшего соседа и k-ближайшего соседа
- метод опорных векторов
- байесовские сети
- линейная регрессия
- корреляционно-регрессионный анализ
- иерархические методы кластерного анализа

- неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы

- эволюционное программирование и генетические алгоритмы
- метод ограниченного перебора
- эволюционное программирование и генетические алгоритмы

- разнообразные методы визуализации данных и множество других методов.

- точность
- масштабируемость
- интерпретируемость
- проверяемость
- трудоемкость
- гибкость
- быстрота
- популярность

АЛГОРИТМ	ТОЧНОСТЬ	МАСШТАБИРУЕМОСТЬ	ИНТЕРПРЕТИРУЕМОСТЬ	ПРИГОДНОСТЬ К ИСП.
линейная регрессия	нейтральная	высокая	высокая / нейтральная	высокая
нейронные сети	высокая	низкая	низкая	низкая
методы визуализации	высокая	очень низкая	высокая	высокая
деревья решений	низкая	высокая	высокая	высокая / нейтральная
нейронные сети	высокая	нейтральная	низкая	высокая / нейтральная
k-ближайшего соседа	низкая	очень низкая	высокая / нейтральная	нейтральная

АЛГОРИТМ	ТРУДОЕМКОСТЬ	РАЗНОСТОРОННОСТЬ	БЫСТРОТА	ПОПУЛЯРНОСТЬ
линейная регрессия	нейтральная	нейтральная	высокая	низкая
нейронные сети	нейтральная	низкая	очень низкая	низкая
методы визуализации	очень высокая	низкая	чрезвычайно низкая	высокая / нейтральная
деревья решений	высокая	высокая	высокая / нейтральная	высокая / нейтральная
нейронные сети	низкая / нейтральная	нейтральная	низкая / нейтральная	нейтральная
k-ближайшего соседа	нейтральная низкая	низкая	высокая	низкая

Таблица 1 Сравнительная характеристика методов Data Mining

Классификация методов Data Mining

Методы группы:

- кластерный анализ
- метод ближайшего соседа
- метод k-ближайшего соседа
- рассуждение по аналогии

Методы группы:

- логические методы
  - нечеткие запросы и анализы
  - символьные правила
  - деревья решений
  - генетические алгоритмы.
- методы визуализации
- методы кросс-табуляции
- методы, основанные на уравнениях

Статистические методы Data Mining

Методы группы:

- *дескриптивный анализ и описание исходных данных*
- *анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ)*
- *многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.)*
- *анализ временных рядов (динамические модели и прогнозирование)*

Кибернетические методы Data Mining

Методы группы:

- *искусственные нейронные сети (распознавание, кластеризация, прогноз);*
- *эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);*
- *генетические алгоритмы (оптимизация);*
- *ассоциативная память (поиск аналогов, прототипов);*
- *нечеткая логика;*
- *деревья решений;*
- *системы обработки экспертных знаний.*

Описательные методы

Методы группы:

- *алгоритм k-средних*
- *k-медианы*
- *иерархические методы кластерного анализа*
- *самоорганизующиеся карты Кохонена*
- *методы кросс-табличной визуализации*

Прогнозирующие методы

Методы группы:

- *нейронные сети*
- *деревья решений,*
- *линейная регрессия*
- *метод ближайшего соседа*
- *метод опорных векторов*

Aberdeen Group: " *Data Mining* - технология добычи полезной информации из баз *данных*. Однако в связи с существенными различиями между инструментами, опытом и финансовым состоянием поставщиков продуктов, предприятиям необходимо тщательно оценивать предполагаемых разработчиков *Data Mining* и партнеров.

ABERDEEN GROUP

TWO CROWS CONSULTING

Herb Edelstein: «Недавнее исследование компании Two Crows показало, что Data Mining находится все еще на ранней стадии развития. Многие организации интересуются этой технологией, но лишь некоторые активно внедряют такие проекты. Удалось выяснить еще один важный момент: процесс реализации Data Mining на практике оказывается более сложным, чем ожидается».

1. *Data Mining* не может заменить аналитика
2. Сложность разработки и эксплуатации приложения **Data Mining**
3. Квалификация пользователя
4. Извлечение полезных сведений невозможно без хорошего понимания сути данных
5. Сложность подготовки данных
6. Большой процент ложных, недостоверных или бессмысленных результатов
7. Высокая стоимость
8. Наличие достаточного количества репрезентативных данных

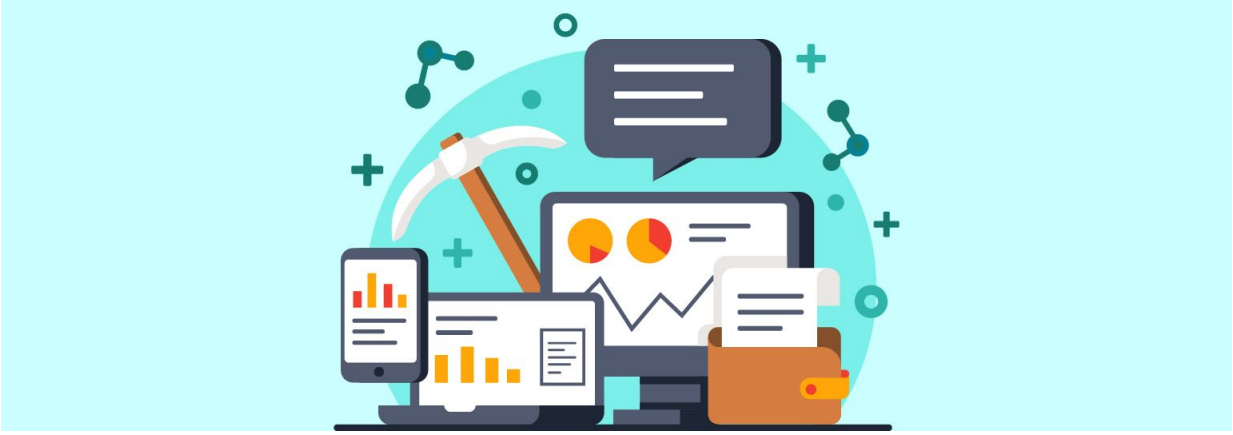
Перспективы технологии Data Mining

• выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям;

• создание формальных языков и логических средств, с помощью которых будут формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях;

• создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории, опирающиеся на эмпирические данные ;

• преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области



Области, где применения технологии *Data Mining*, скорее всего, будут успешными, имеют такие особенности:

- требуют решений, основанных на *знаниях* ;
- имеют изменяющуюся окружающую среду;
- имеют доступные, достаточные и значимые *данные* ;
- обеспечивают высокие дивиденды от правильных решений.

Международная конференция по Knowledge Discovery Data Mining (**International Conferences on Knowledge Discovery and Data Mining**).

Среди наиболее известных WWW-источников - сайт www.kdnuggets.com , который ведет один из основателей Data Mining Григорий Пиатецкий-Шапиро.

Периодические издания по Data Mining: Data Mining and Knowledge Discovery, KDD Explorations, ACM-TODS, IEEE-TKDE, JIIS, J. ACM, Machine Learning, Artificial Intelligence.

Материалы конференций: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, Machine learning (ICML), AAAI, IJCAI, COLT (Learning Theory).

Краудсорсинг

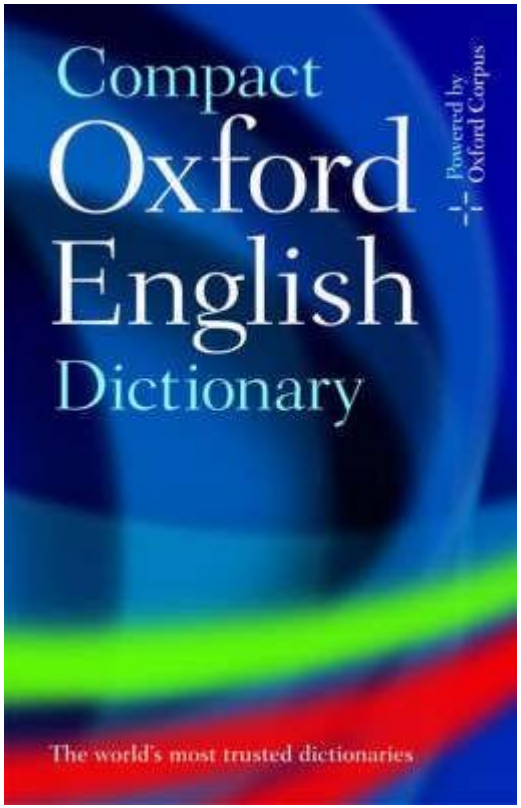
Краудсорсинг — привлечение к решению какой-либо проблемы большой группы людей

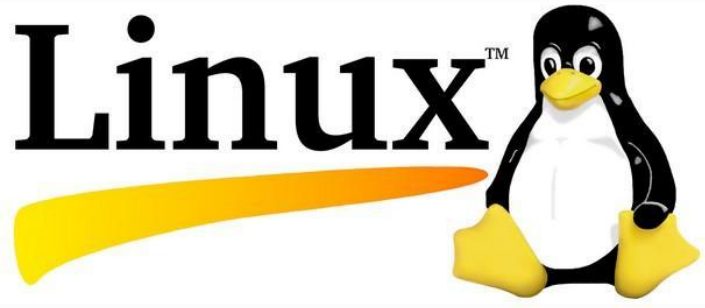
В 2003 году **Луис фон Ах (Luis von Ahn)** вместе со своими коллегами впервые предложил понятие **"человеческих вычислений**

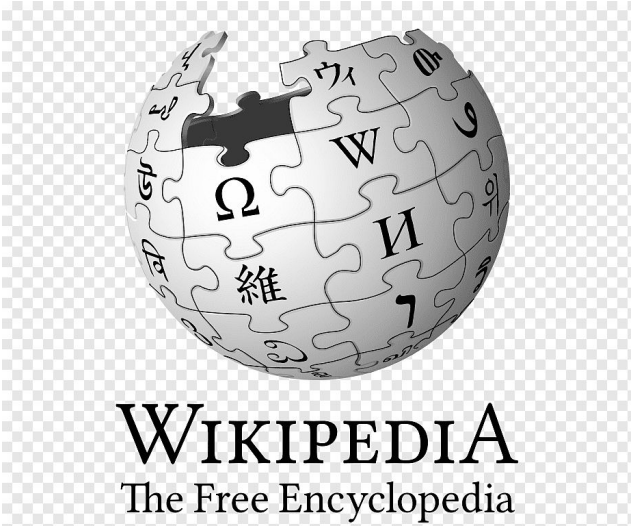
в 2006 году термин "Краудсорсинг" (crowdsourcing) был сформулирован редактором журнала "Wired" **Джеффом Хау (Jeff Howe)**

WIRED

Краудсорсинг - это мобилизация ресурсов людей посредством информационных технологий с целью решения задач, стоящих перед бизнесом, государством и обществом в целом.







I. По сфере жизни (бизнес, социальный, политический)

II. По типу решаемых задач (создание продукта (контента), голосование, поиск решения, поиск людей, сбор информации, сбор мнений, тестирование, служба поддержки, сбор средств - Краудфандинг).

*http://crowdsourcing.ru/article/what_is_the_crowdsourcing

Классификация Краудсорсинга

По сфере жизни:

1) **Бизнес**

2) **Социальный или общественный**

3) **Политический или государственный**

Классификация Краудсорсинга

По типу решаемых задач:

1) Создание продукта (контента)
99designs, TopCoder, Witmart, Tongal, Audiodraft

2) Голосование

3) Поиск решения
Kaggle, CrowdFlower, InnoCentive, Academy of Ideas, Wazoku

4) Поиск людей, например - Liza Alert

5) Сбор информации, например - Zooniverse

6) Сбор мнений, например - Chaordix, Innopinion и AnswerTap

7) Тестирование, например uTest,

8) Сбор средств - Краудфандинг

1) Большой охват
2) Вовлечение пользователей
3) Разнообразие выбора
4) Единственно возможный вариант
5) Фиксированные сроки
6) Экономия финансовых ресурсов


witology


LizaAlert
ПОИСКОВО-СПАСАТЕЛЬНЫЙ ОТРЯД


CITYCELEBRITY
CROWDSOURCING PLATFORM


Go


WORK
DONE

СПАСИБО ЗА ВНИМАНИЕ!

Сергей Алексеевич Корчагин
SAKorchagin@fa.ru
2021