

BIG DATA: ОСНОВЫ РАБОТЫ С БОЛЬШИМИ МАССИВАМИ ДАННЫХ

Денис Афанасьев



ЦЕЛИ И ЗАДАЧИ КУРСА

Ключевые роли, необходимые для интеграции функции передовой аналитики в организационную структуру

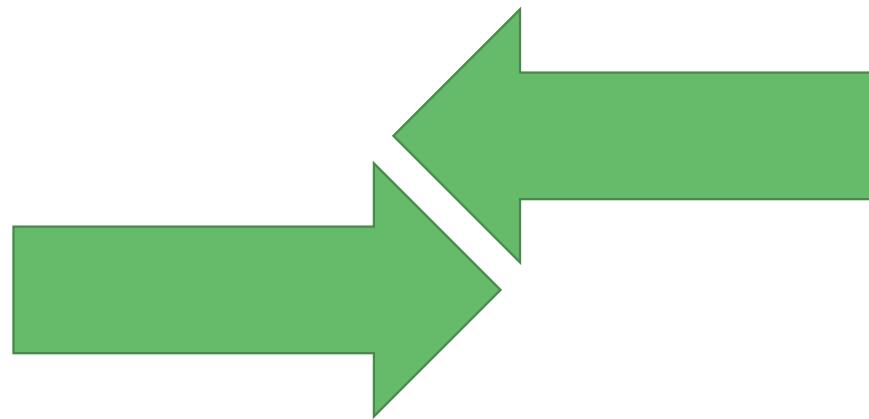




ЦЕЛИ И ЗАДАЧИ КУРСА

Аналитик данных

Менеджер



"Successful machine learning is a two-way thing. Data scientists need to know something about your organization and what it does, and you need to understand a little bit about machine learning. Without this joint understanding it's unlikely that you or your organization will be able to realize the full benefits that machine learning has to offer."

Finlay, Steven. **Artificial Intelligence and Machine Learning for Business: A No-Nonsense Guide to Data Driven Technologies** (p. 3). Relativistic. Kindle Edition.



ЧТО ТАКОЕ BIG DATA

Динамика популярности



В среднем

Artificial Intelligence

Machine Learning

Big Data

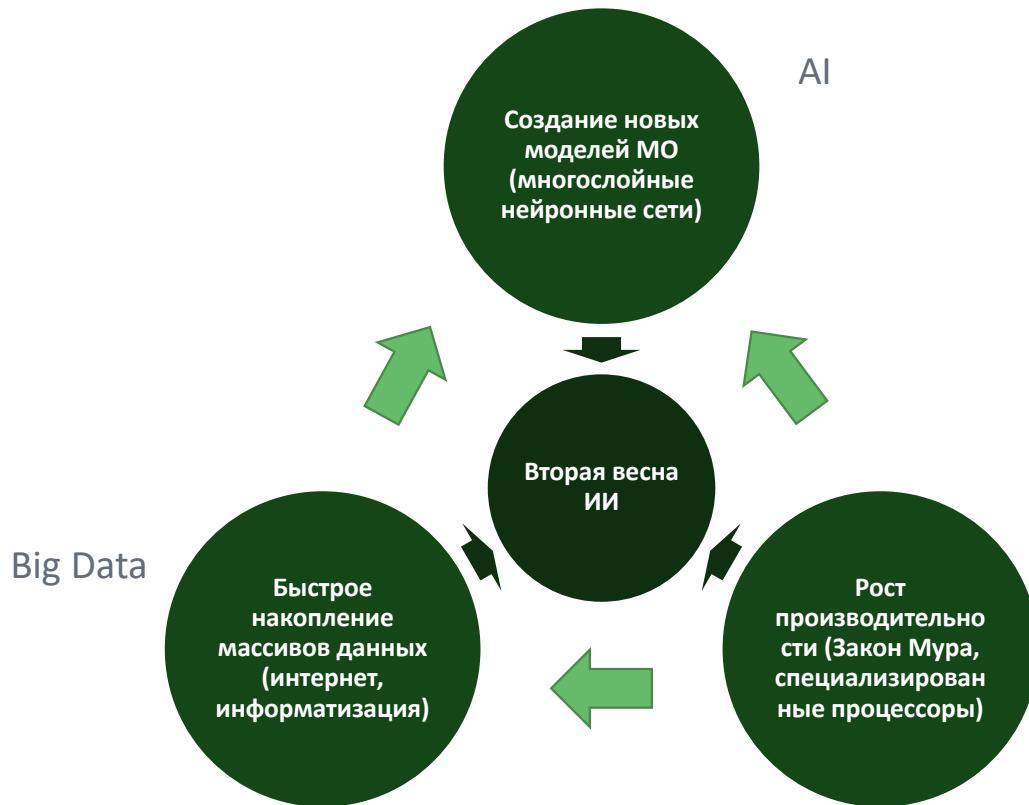
В среднем

1 янв. 200... 1 янв. 2009 г. 1 янв. 2014 г. 2012 2016



<https://trends.google.com/trends/>

ВТОРАЯ ВЕСНА AI



1940—1960: создание первых нейронных сетей

1970—1990: «зима ИИ»

1980—2000: создание алгоритмов, лежащих в основе современной «весны ИИ»

2010+ — ряд зрелищных демонстраций возможностей ИИ

2016+ — инвестиционный и медиа-бум

СКОРОСТЬ ВЫЧИСЛЕНИЙ

Закон Мура (1965) - эмпирическое наблюдение, изначально сделанное Гордоном Муром, согласно которому (в современной формулировке) количество транзисторов, размещаемых на кристалле интегральной схемы, удваивается каждые 24 месяца.

Часто цитируемый интервал в 18 месяцев связан с прогнозом Давида Хауса из Intel, по мнению которого, **производительность** процессоров должна удваиваться каждые 18 месяцев из-за сочетания роста количества транзисторов и увеличения тактовых частот процессоров^[1].

Флоп - количество операций с плавающей точкой в секунду. От английского FLoat OPerations

Компьютер:

- Процессор
(скорость вычислений)
- Память (объем)

1 The accelerating pace of change ...

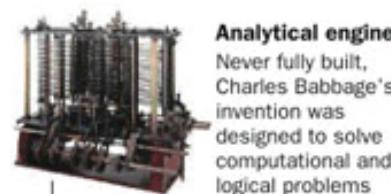


2 ... and exponential growth in computing power ...

Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

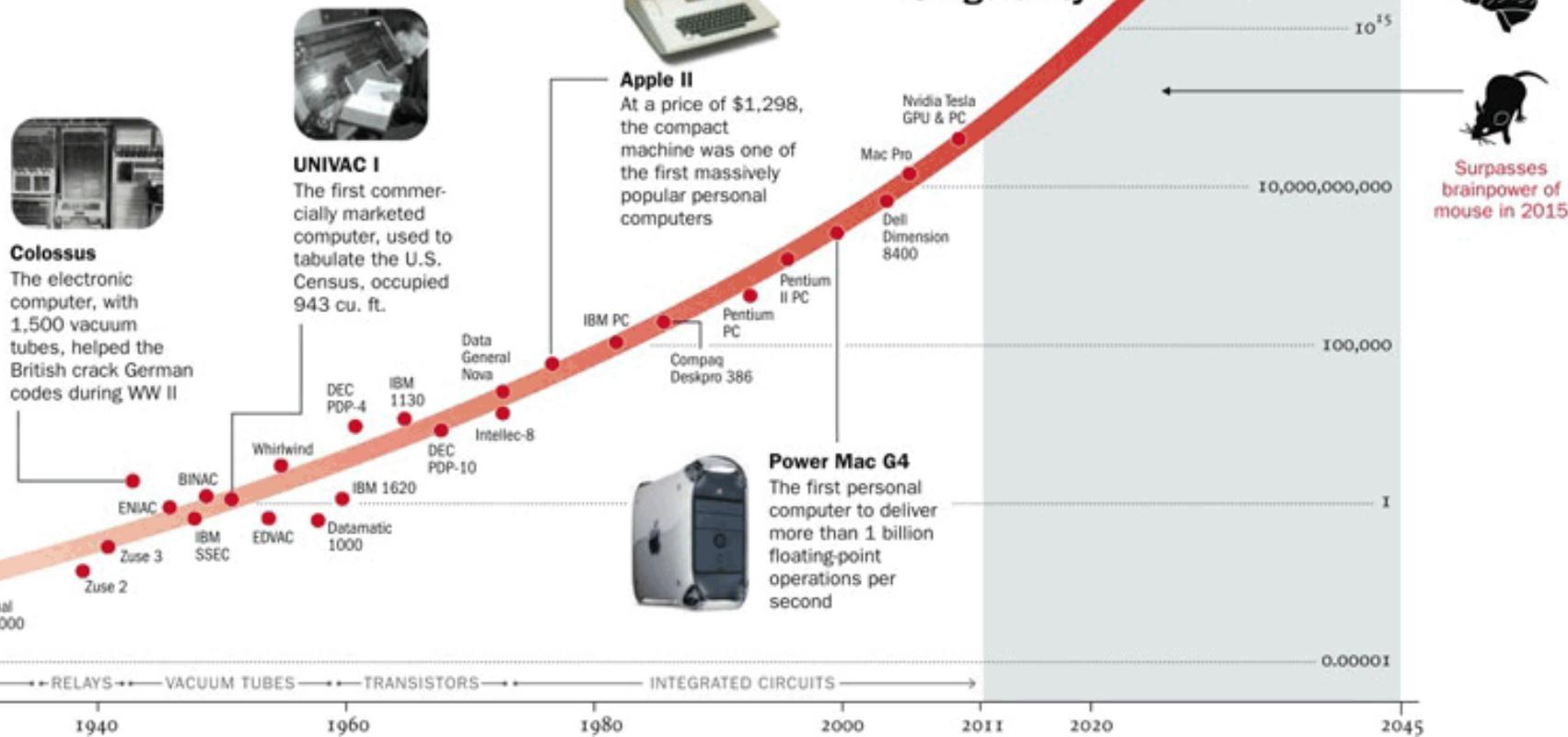
COMPUTER RANKINGS

By calculations per second per \$1,000



Analytical engine

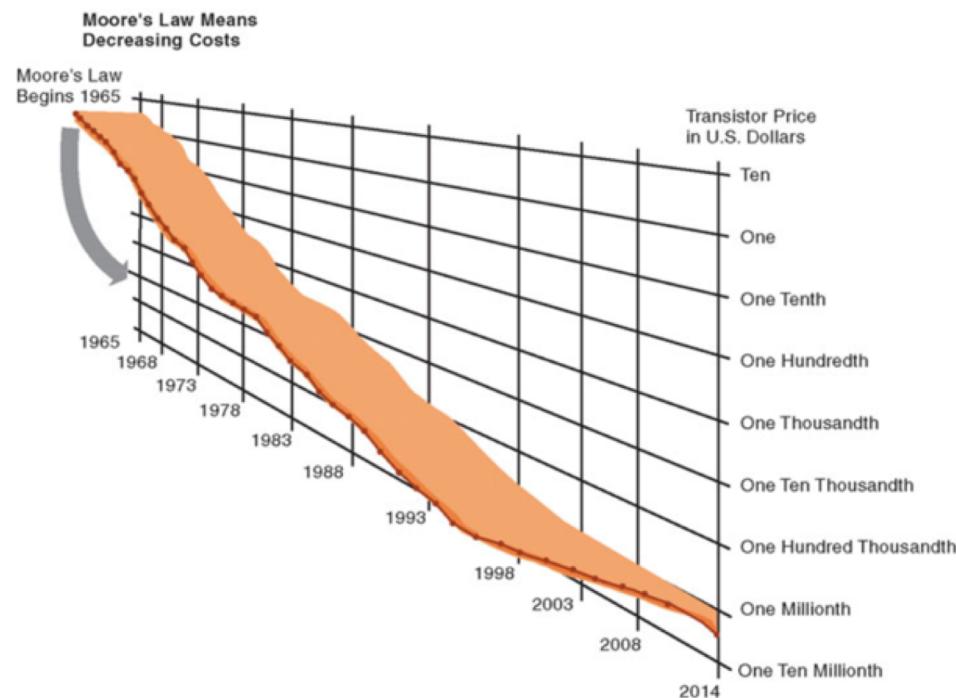
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems



3 ... will lead to the Singularity



СЛЕДСТВИЯ ЗАКОНА МУРА



Examples of rate of change

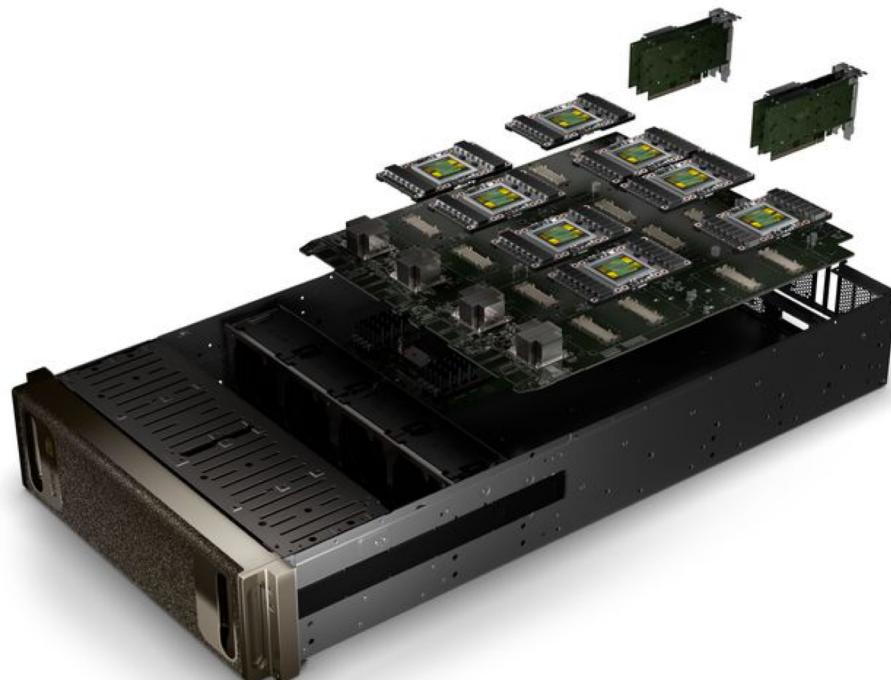
Technology	Average cost for equivalent functionality	Scale
3D printing	\$40,000 (2007) to \$100 (2014)	400x in 7 years
Industrial robots	\$500,000 (2008) to \$22,000 (2013)	23x in 5 years
Drones	\$100,000 (2007) to \$700 (2013)	142x in 6 years
Solar energy	\$30 per kWh (1984) to \$0.16 per kWh (2014)	200x in 20 years
3D LIDAR Sensors	\$20,000 (2009) to \$79 (2014)	250x in 5 years
DNA genome seq	\$10,000,000 (2007) to \$1,000 (2014)	10,000x in 7 years
BCI neuro devices	\$4,000 (2006) to \$90 (2011)	44x in 5 years
Full body med scan	\$10,000 (2000) to \$500 (2014)	20x in 14 years



Source: "Exponential Organizations"
<http://www.slideshare.net/vangeest/exponential-organizations-h>

@dw2

GPU



Вычислительная
мощность: более 110
TFLOPS

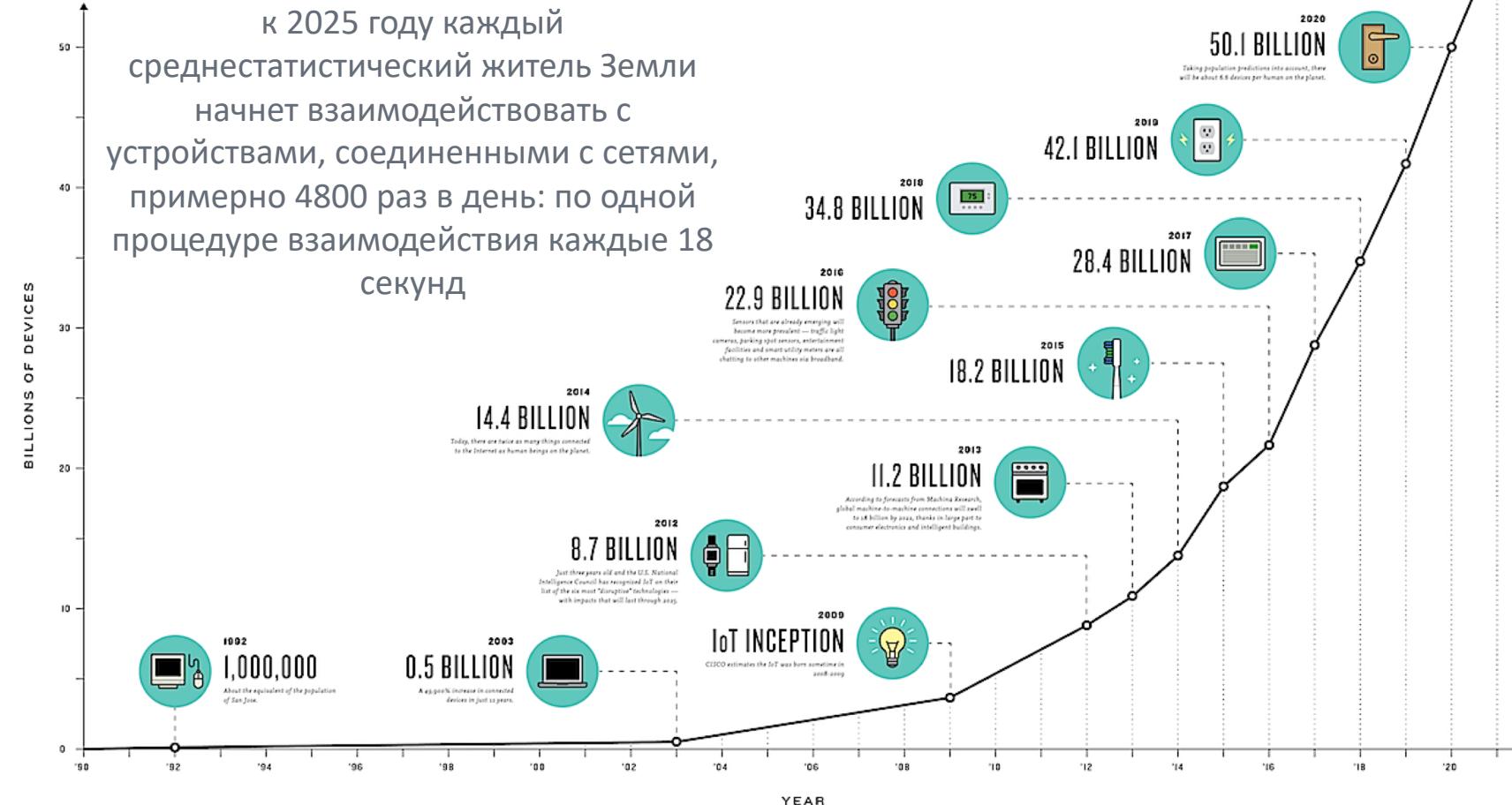
Стоимость: \$3-4K

Сравнение:
суперкомпьютер DGX-1
на 170 TFLOPS стоит
\$129K





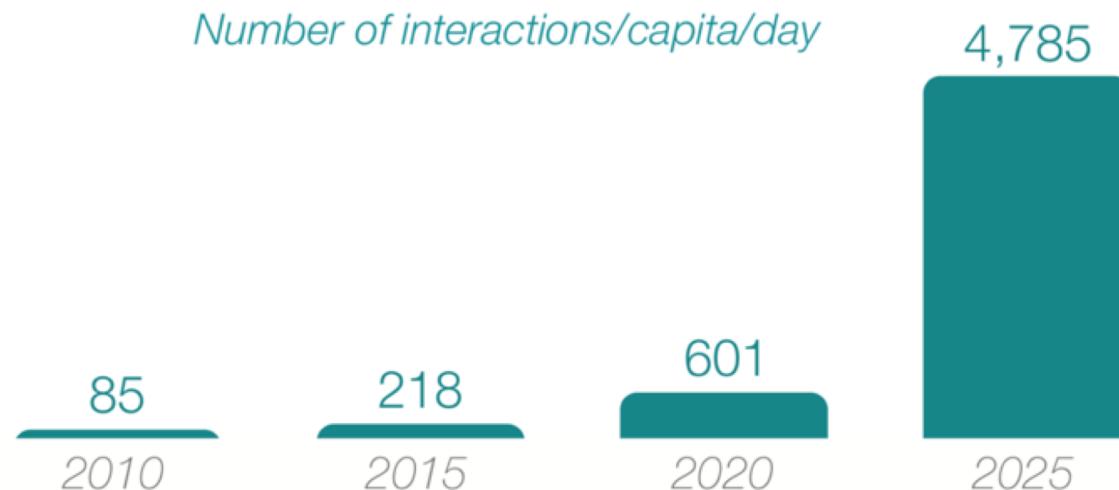
ЦИФРОВИЗАЦИЯ





ЧТО ТАКОЕ BIG DATA

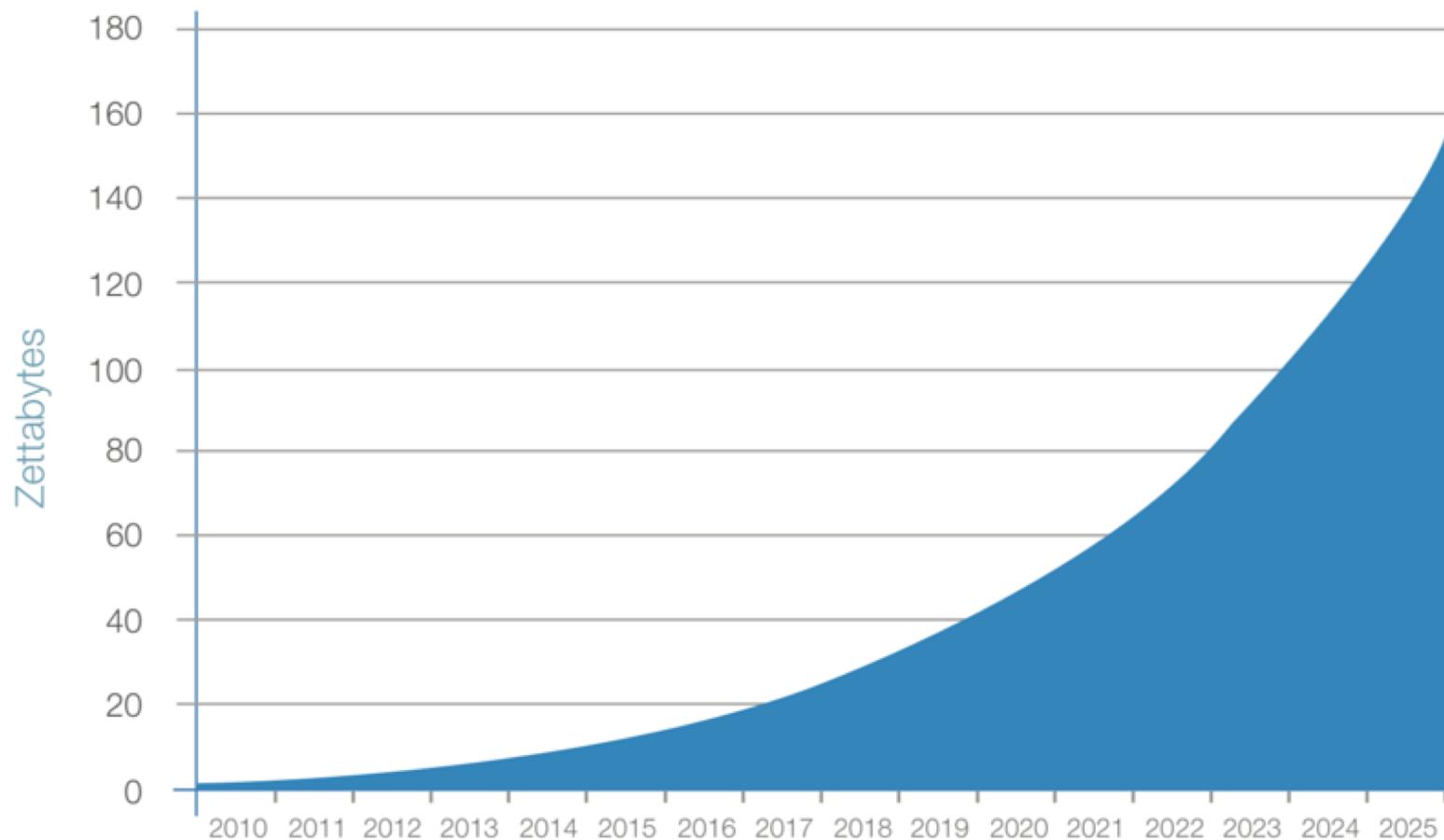
Figure 8. Interactions per Connected Person per Day



Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017



ПРОГНОЗ РОСТА ОБЪЕМА ДАННЫХ



Согласно прогнозу IDC, доля глобальной информационной сферы, подвергаемой анализу, к 2025 году вырастет по сравнению с нынешней в 50 раз, достигнув 5,2 Збайт; а объем данных, анализируемых при участии когнитивных систем, вырастет в 100 раз, составив 1,4 Збайт. Когнитивные системы позволят чаще и более гибко анализировать данные во многих отраслях и во многих ситуациях.

Data created

ЧТО ТАКОЕ «BIG DATA»

1880 год - обработка информации и представление данных переписи населения в Америке в таблице заняло 8 лет. При этом по прогнозам обработка данных переписи 1890 года заняла бы еще больше времени, и результаты не были бы готовы даже до проведения новой переписи. Тогда проблему решила табулирующая машина, изобретенная Германом Холлеритом (Herman Hollerith) в 1881 году.

1997 год - термин **Big Data** был впервые (по данным электронной библиотеки Association for Computing Machinery) введен в 1997 году Майклом Коксом (Michael Cox) и Дэвидом Эллsworthом (David Ellsworth) на 8-й конференции IEEE по визуализации. Они назвали проблемой больших данных нехватку емкости основной памяти, локального и удаленного диска для выполнения виртуализации. А в 1998 году руководитель исследовательских работ в SGI Джон Мэши (John R. Mashey) на конференции USENIX использовал термин Big Data в его современном виде.

Большие данные – данные, обработка которых в заданных условиях/ограничениях требует применения новых технических подходов

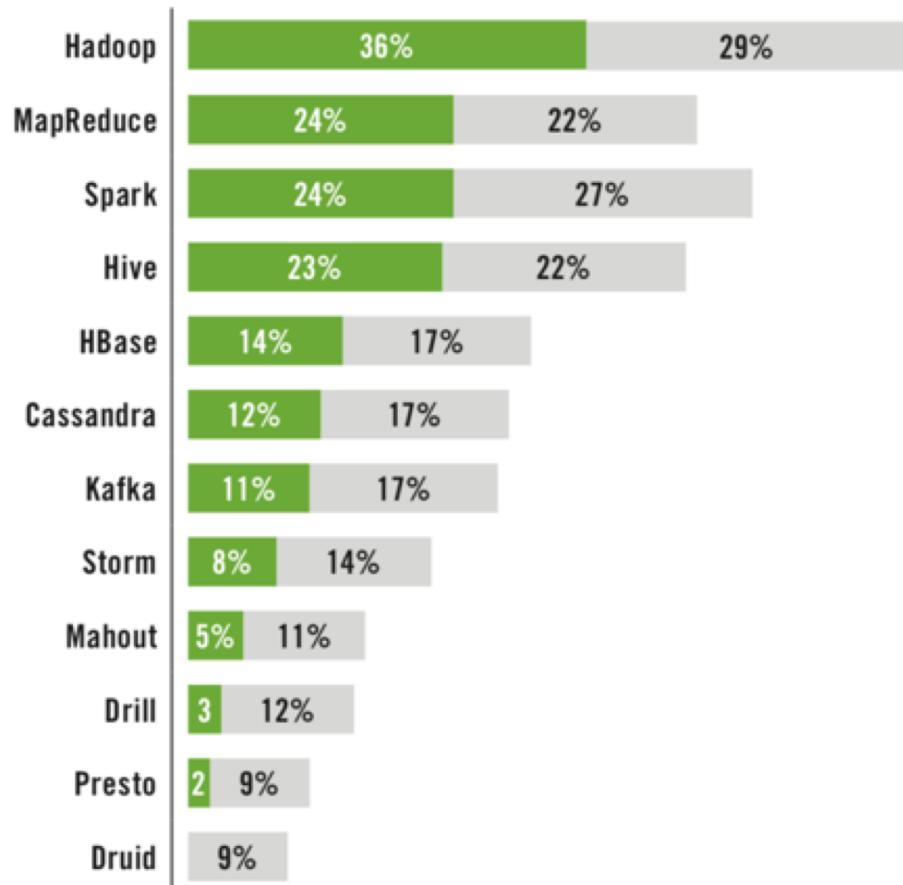
ЧТО ТАКОЕ «BIG DATA»

2003 год

И хотя проблема хранения большого объема данных осознавалась давно и усилилась после появления интернета, переломным моментом стал 2003 год, за который было создано информации больше чем за все предыдущее время.

Примерно в это же время выходит публикация Google File System о вычислительной концепции MapReduce, которая легла в основу Hadoop. Над этим инструментом в течении нескольких лет работал Дуг Каттинг (Doug Cutting) в рамках проекта Nutch, а в 2006 году Каттинг присоединился к Yahoo и Hadoop стал отдельным полноценным решением.

Google YAHOO!





НЕМНОГО ИСТОРИИ - АНАЛИТИКА

1950 Аллан Тьюринг (Alan Turing) создает Тьюринг тест для оценки интеллекта компьютера

1952 Артур Самуэль, пионер в области искусственного интеллекта создает первую шашечную программу для IBM 701. В 1955 году Самуэль добавляет в программу способность к самообучению.

1958 Фрэнк Розенблattt (Frank Rosenblatt) придумал **Персептрон** — первую искусственную нейронную сеть и создал первый нейрокомпьютер «Марк-1» .

1967 Написан метрический алгоритм классификации (Метод k ближайших соседей). Алгоритм позволил компьютерам использовать простые шаблоны распознавания.

1997 Компьютер Deep Blue обыграл чемпиона мира по шахматам Гарри Каспарова.

2006 Джекфри Хинтон (Geoffrey Hinton), ученый в области искусственных нейросетей, ввел в обиход термин «Глубинное обучение» (Deep learning).

2011 Суперкомпьютер IBM Watson, оснащенный системой искусственного интеллекта, одержал победу в телевикторине Jeopardy!. Его соперниками были маститые игроки Брэд Раттер (Bred Ratter) и Кен Дженнингс (Ken Jennings).

2016 Программа AlphaGo, разработанная гугловской компанией DeepMind, выиграла в четырех играх из пяти у чемпиона мира по игре в го корейца Ли Седоля (Lee Se-dol).

ОСНОВНЫЕ ПРЕДПОСЫЛКИ

Стоимость вычислений постоянно падает

Стоимость снижается экспоненциально, согласно закону
Мура

Увеличивается количество цифровых устройств

В мире задействованы почти 5 миллиардов
смартфонов, что оказывает влияние на все виды
коммерции и вовлечения клиентов/поставщиков.

**Постоянно растет объем генерируемых и
накапливаемых цифровых данных**

К 2020 году 40–44 Збайт информации будет накоплено в мире
(суммарно на всех системах хранения данных), по разным
оценкам.

Согласно прогнозу IDC, доля глобальной информационной сферы, подвергаемой анализу, к
2025 году вырастет по сравнению с нынешней в 50 раз, достигнув 5,2 Збайт; а объем данных,
анализируемых при участии когнитивных систем, вырастет в 100 раз, составив 1,4 Збайт.

Когнитивные системы позволяют чаще и более гибко анализировать данные во многих отраслях
и во многих ситуациях.

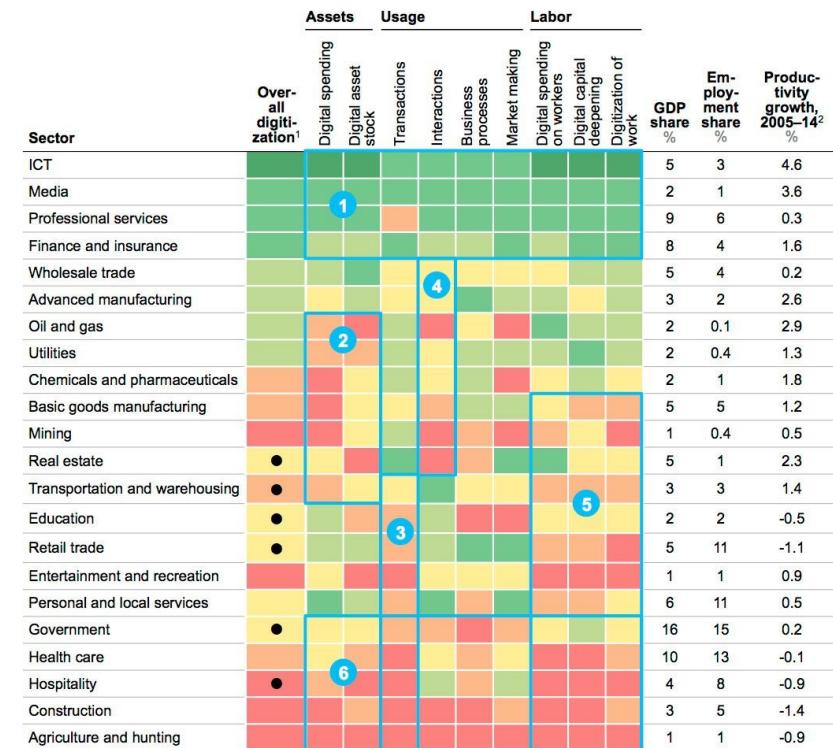
Exhibit E1

The MGI Industry Digitization Index

2015 or latest available data

Relatively low digitization Relatively high digitization

● Digital leaders within relatively undigitized sectors



① Knowledge-intensive sectors that are highly digitized across most dimensions

② Capital-intensive sectors with the potential to further digitize their physical assets

③ Service sectors with long tail of small firms having room to digitize customer transactions

④ B2B sectors with the potential to digitally engage and interact with their customers

⑤ Labor-intensive sectors with the potential to provide digital tools to their workforce

⑥ Quasi-public and/or highly localized sectors that lag across most dimensions

1 Based on a set of metrics to assess digitization of assets (8 metrics), usage (11 metrics), and labor (8 metrics); see technical appendix for full list of metrics and explanation of methodology.

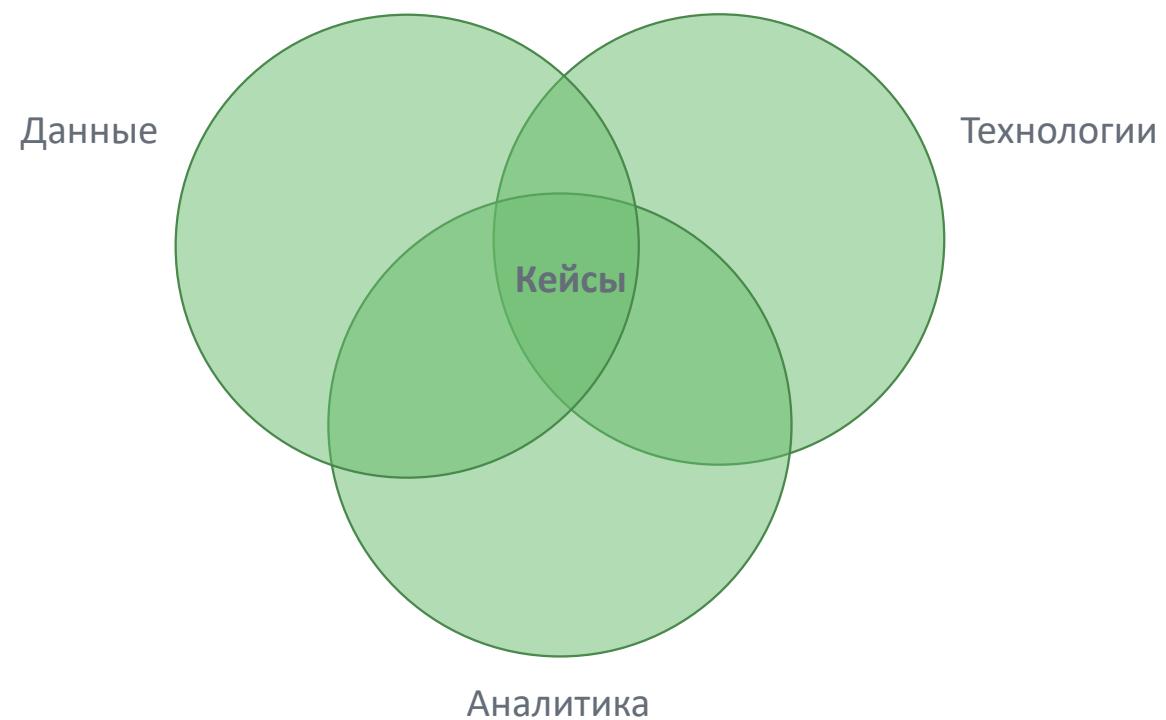
2 Compound annual growth rate.

SOURCE: BEA; BLS; US Census; IDC; Gartner; McKinsey social technology survey; McKinsey Payments Map; LiveChat customer satisfaction report; Appbrain; US contact center decision-makers guide; eMarketer; Bluewolf; Computer Economics; industry expert interviews; McKinsey Global Institute analysis



ЧТО ТАКОЕ BIG DATA

Большие данные — совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence





ДАННЫЕ



ДАННЫЕ – ХАРАКТЕРИСТИКИ, ИСТОЧНИКИ



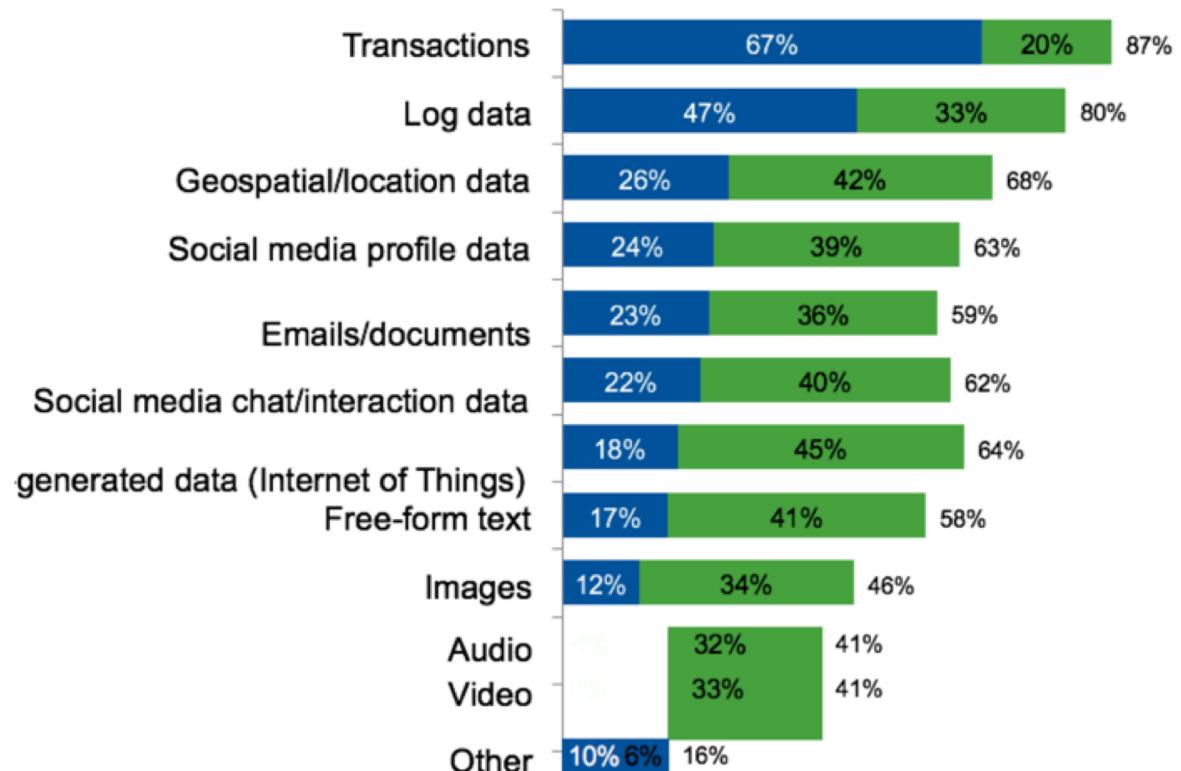
Источник: *The Next Generation of Big Data Analytics*, Hortonworks, Teradata Aster, 22 августа 2012 г.



ДАННЫЕ – ХАРАКТЕРИСТИКИ, ИСТОЧНИКИ

Характеристики данных

- 1 • ВЫСОКАЯ СКОРОСТЬ
• БОЛЬШОЙ ОБЪЕМ
ЗНАЧИТЕЛЬНАЯ ВАРИАТИВНОСТЬ
- 2 • Дата-актив
• Dark-data
- 3 • Собственные
• Внешние
- 4 • Простые
• Структурированные
• Не структурированные
• Полу структурированные
- 5 • Сырые
• Производные

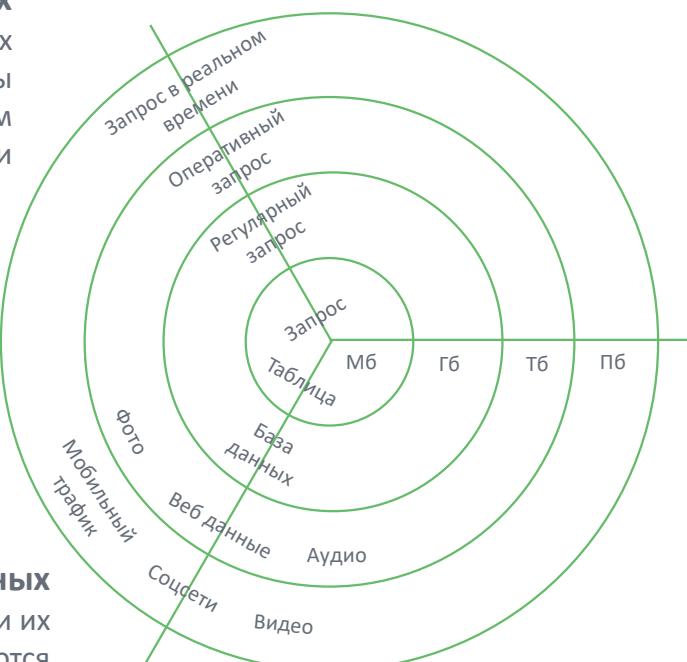




ТЕХНОЛОГИИ

Скорость передачи данных

Скорость передачи данных настолько возросла, что запросы можно формулировать в реальном времени



Объем данных

Объем доступных данных непрестанно растёт

Разнообразие данных

Разнообразие данных и их форматов возрастает, появляются неструктурированные данные

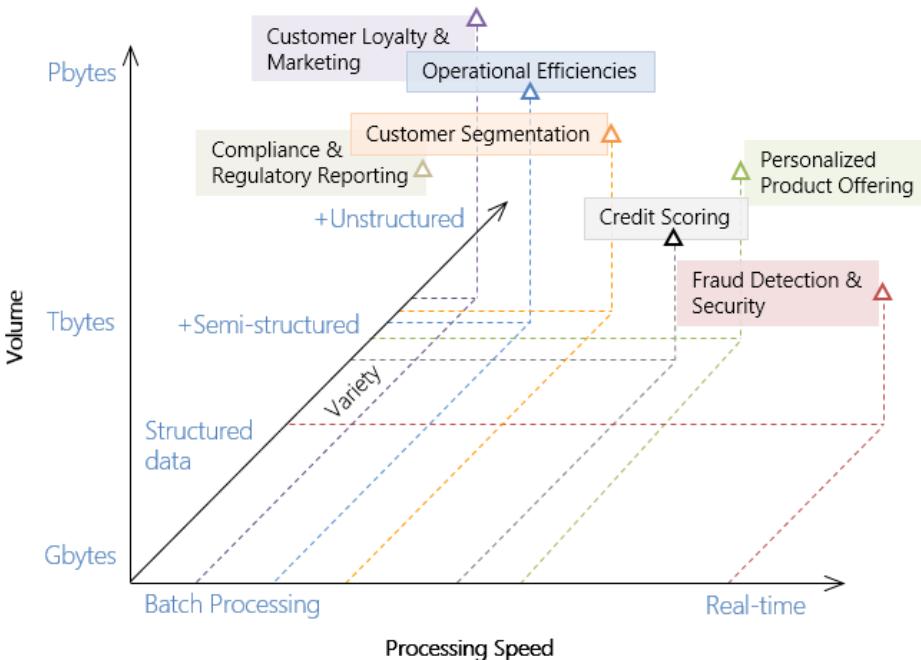
Источник: *Data-centric decision making: the eBay approach* (март 2016 г.), Давиде Чевеллин (Davide Cervellin), руководитель европейского отдела аналитики eBay



ТЕХНОЛОГИИ

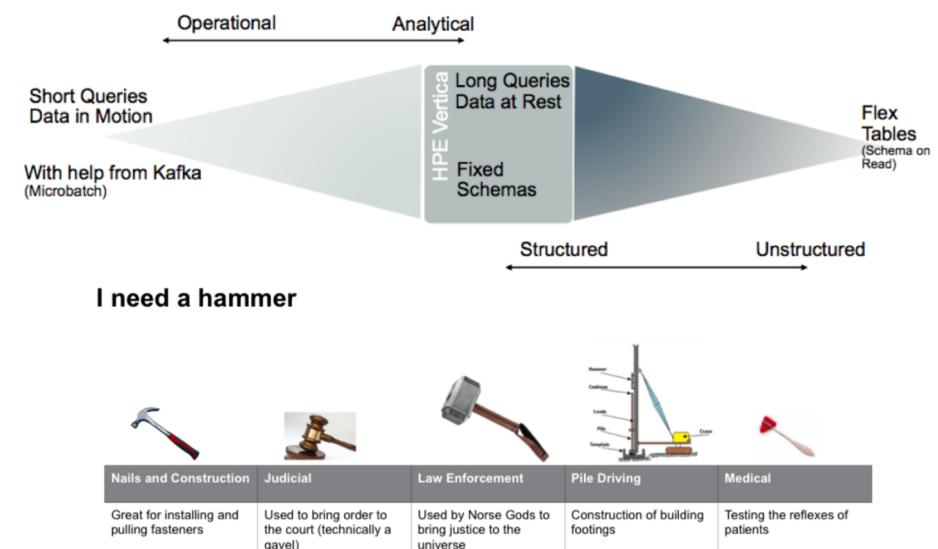


Big Data in
Retail Banking

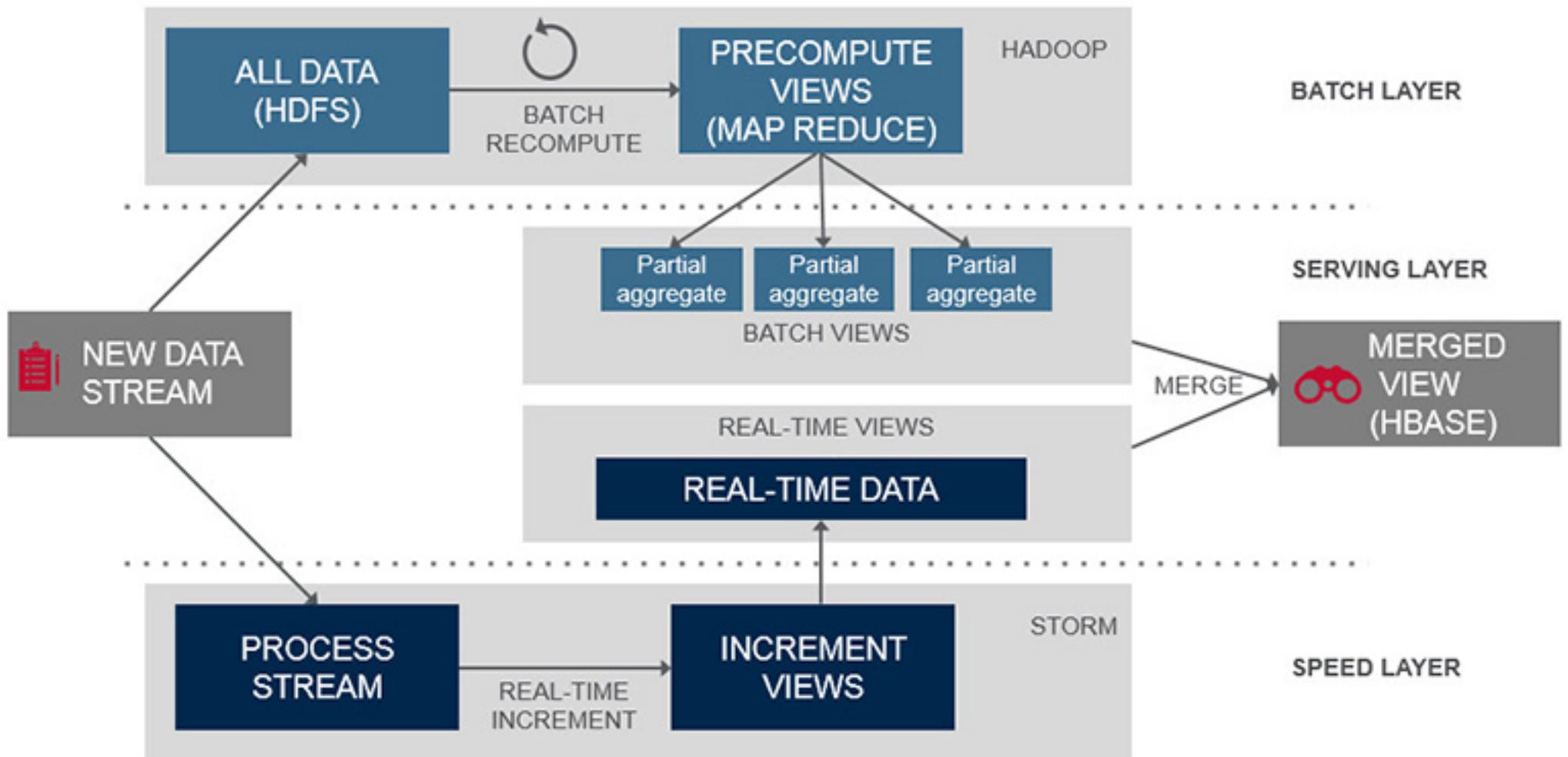


Source: <http://0xCode.in/big-data-in-banking>
This work is licensed under a Creative Commons Attribution 4.0 International License

К 2025 году почти 20% генерируемых данных станут информацией, получаемой в режиме реального времени. При этом более 95% составят данные, поступающие от устройств Интернета вещей. В связи с этим данные должны быть мгновенно доступными для пользователей и предприятий в любое время и в любом месте.

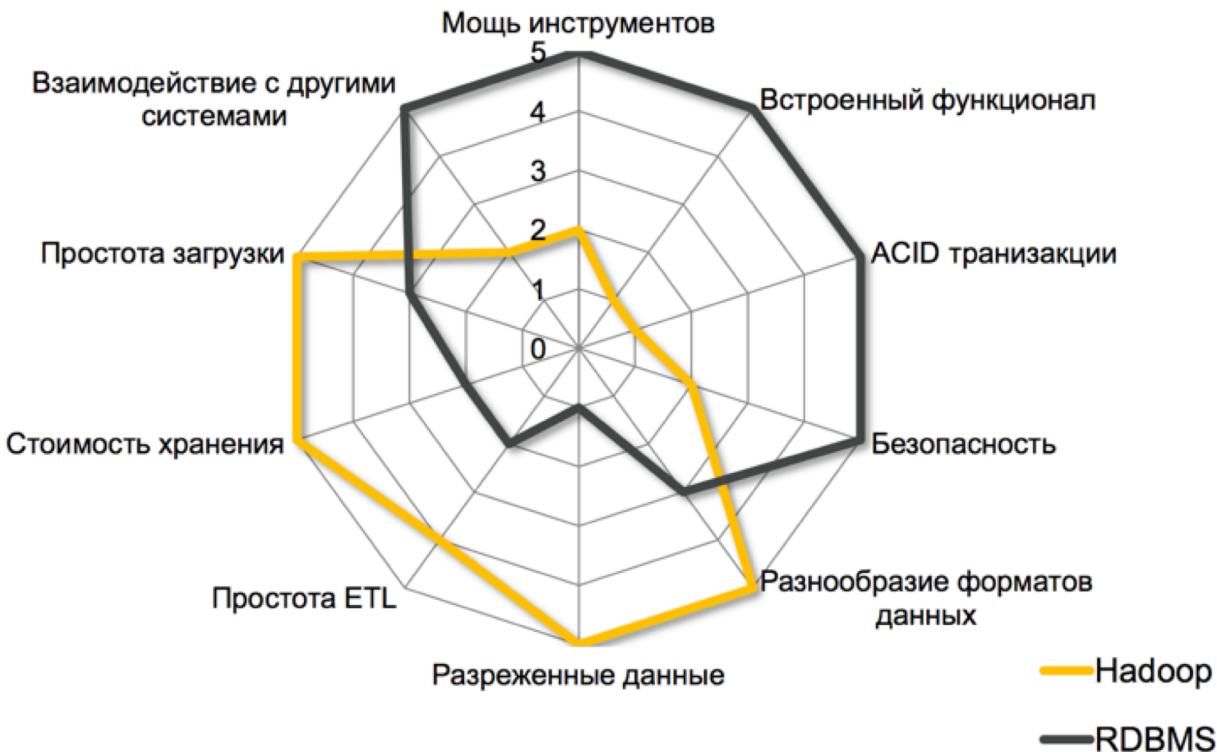


Lambda Architecture





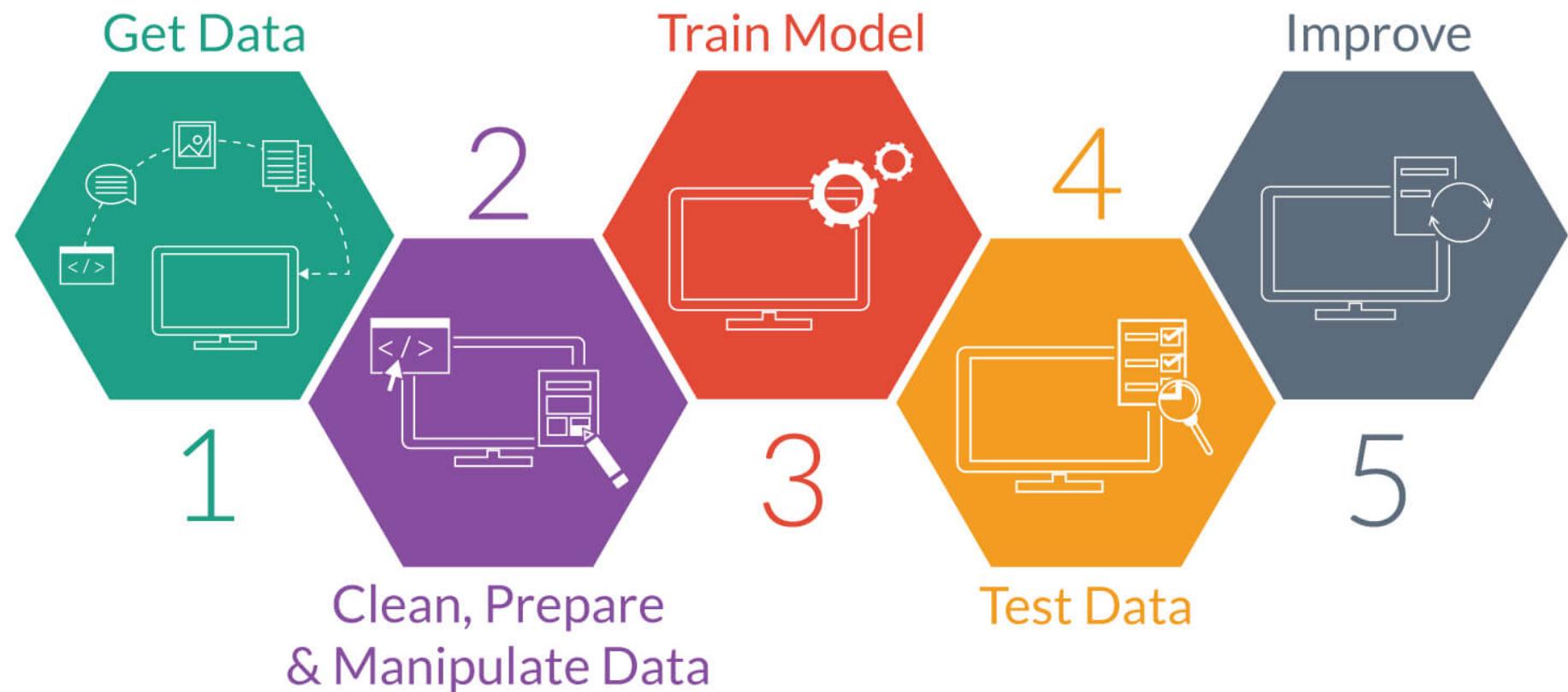
ТЕХНОЛОГИИ



- У Hadoop свои плюсы
- У СУБД свои



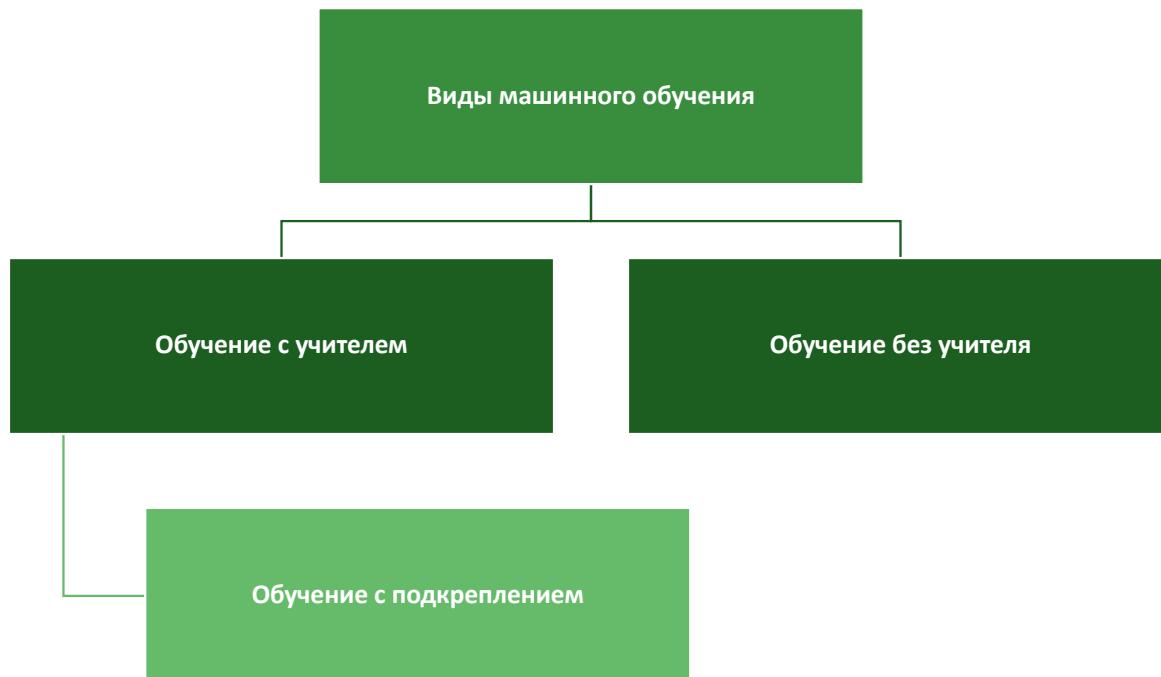
ДАННЫЕ – ПРОЦЕСС ОБРАБОТКИ



МАШИННОЕ ОБУЧЕНИЕ

Машинное обучение (англ. Machine Learning) — обширный подраздел искусственного интеллекта, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа, и извлекающая знания из данных.

Машинное обучение занимается построением прикладных систем ИИ, в которых параметры моделей вычисляются в ходе автоматического или автоматизированного процесса обучения.



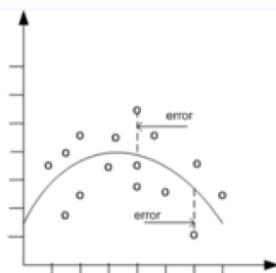


МАШИННОЕ ОБУЧЕНИЕ

Machine learning models can answer 4 basic questions:

How many?

Regression

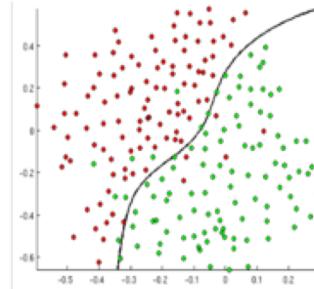


What will revenue be in Q3 in Latin America for product X?

How many salespeople will we have at the end of Q4?

What category?

Classification

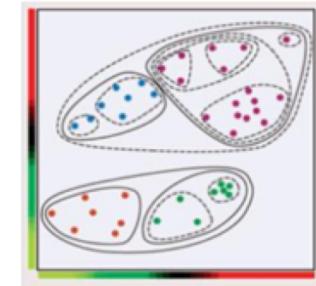


What is probability of a customer to purchase more expensive product SKU?

What is probability the customer will churn to a competitive product?

In what group?

Clustering

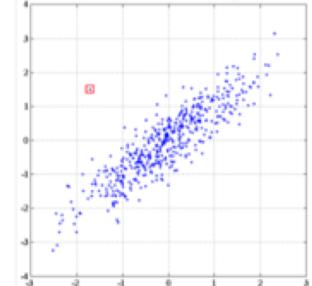


What are 5 groups of APAC customers with similar purchasing?

What bundles of products sell well together?

Is it weird?

Anomaly Detection



What expense reports are potentially fraudulent?

Which customer is likely to default on its payment?

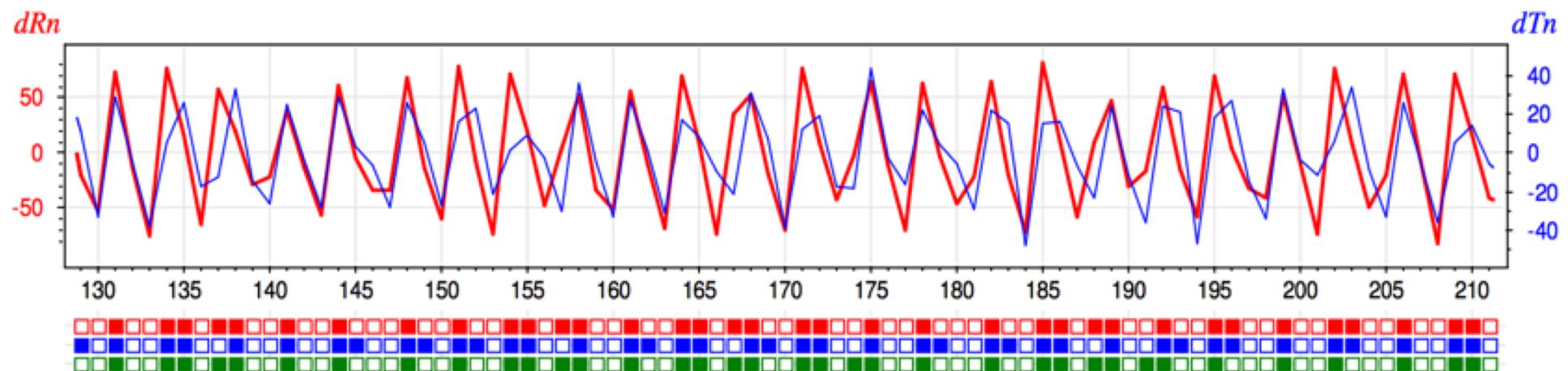
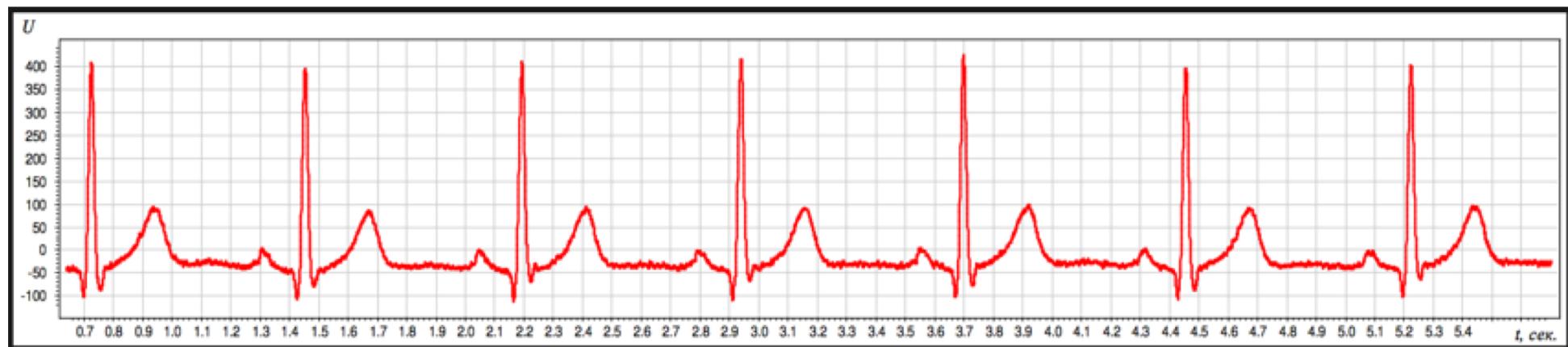
Извлечение, отбор и преобразование признаков

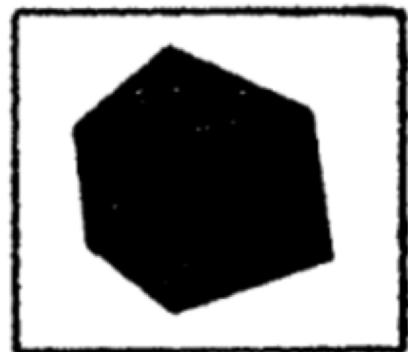
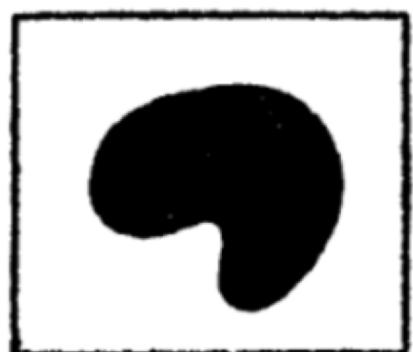
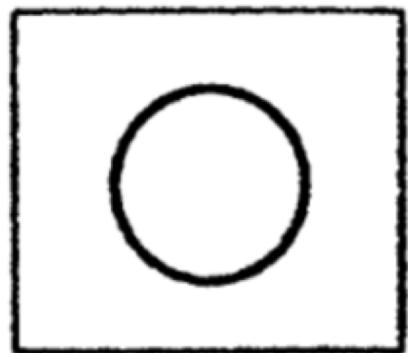


- Машина – не человек:
 - Если в качестве признака есть дата, то машина не понимает время суток
 - Если дано имя – машина не понимает, что оно женское
 - Если дан числовой признак – машина не понимает, много это или мало
 - Машина не может группировать признаки
 - Машина не различает «много» или «мало»
- Примеры преобразования признаков:
 - При прогнозировании спроса на вело прокат дату можно преобразовать в признаки - «утро», «день», «вечер»
 - При прогнозировании цены квартиры «длину» и «ширину» нужно преобразовать в площадь

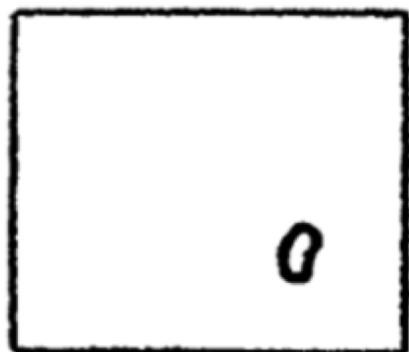
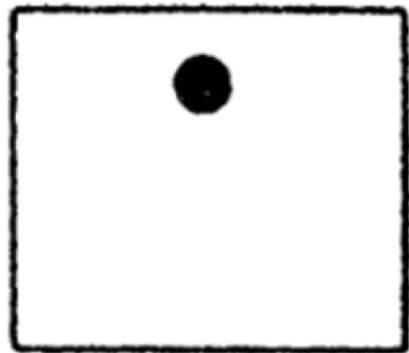
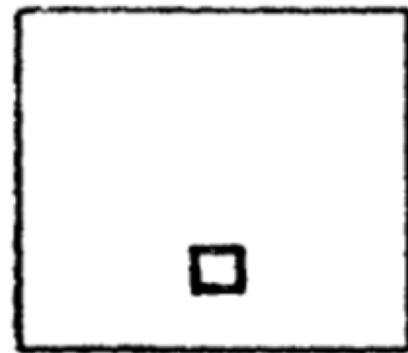


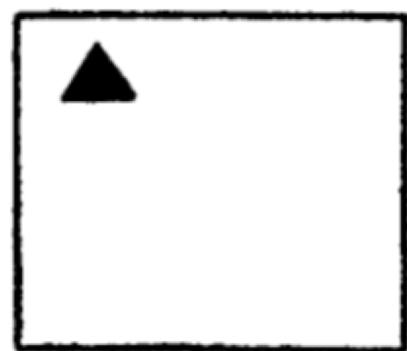
ВЫБОР ПЕРЕМЕННЫХ



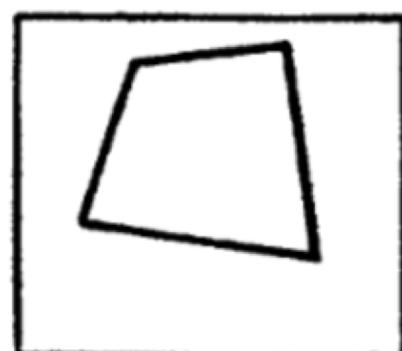
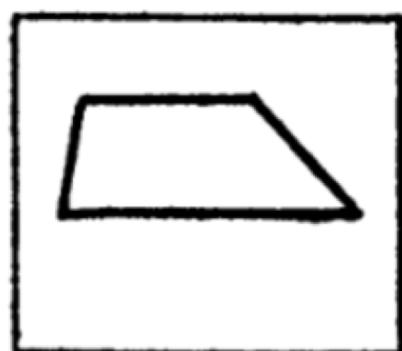
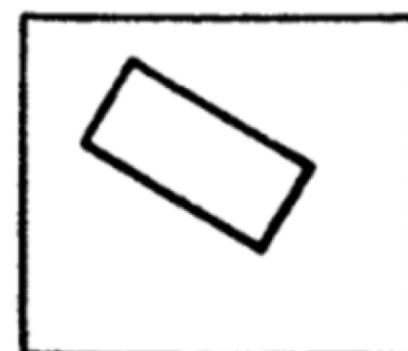
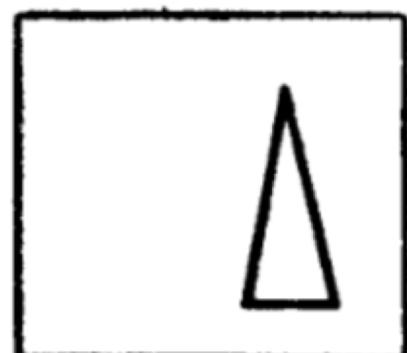


2





6





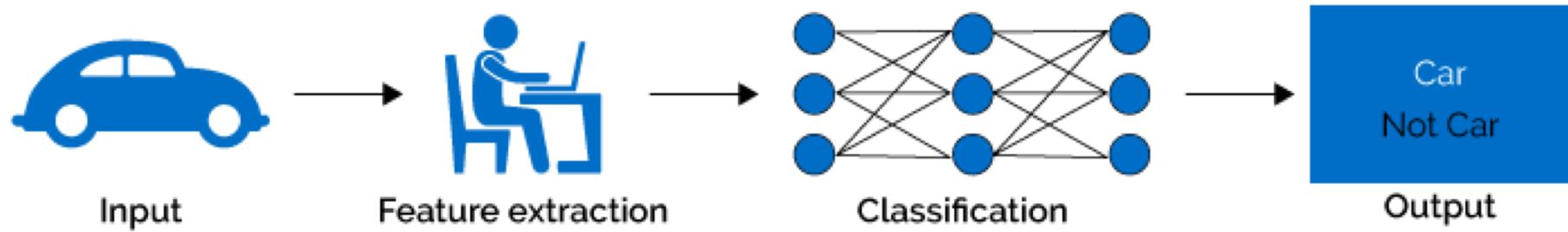
16



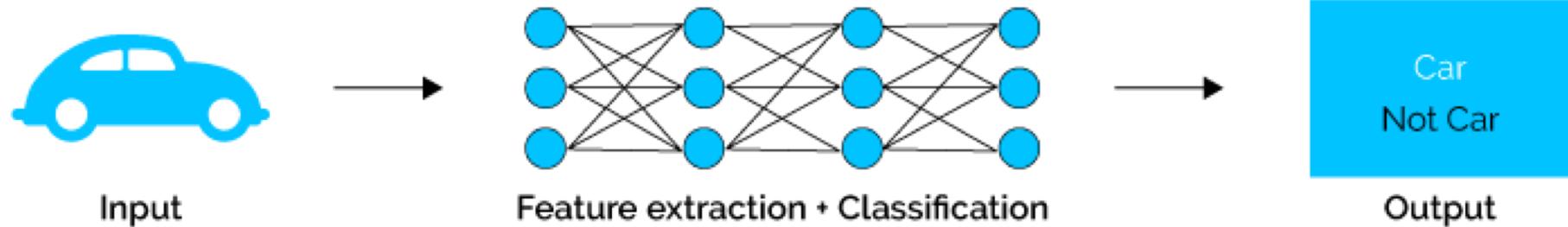


DEEP LEARNING

Machine Learning



Deep Learning

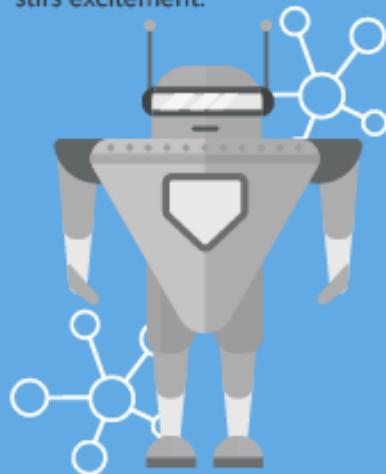




АНАЛИТИКА ДАННЫХ

ARTIFICIAL INTELLIGENCE

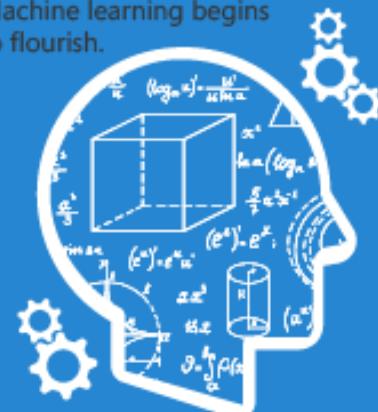
Early artificial intelligence stirs excitement.



1950's 1960's 1970's 1980's 1990's 2000's 2010's

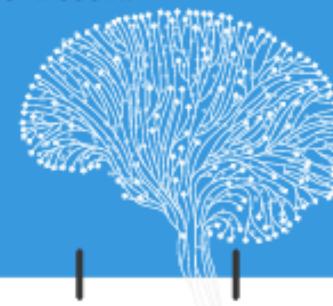
MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



Правила принятия решений задаются человеком

Правила принятия решений формируются автоматически
Человек определяет метод формирования правил и признаки, участвующих в правилах

Правила принятия решений формируются автоматически

Признаки формируются автоматически
Человек определяет метод формирования правил

Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

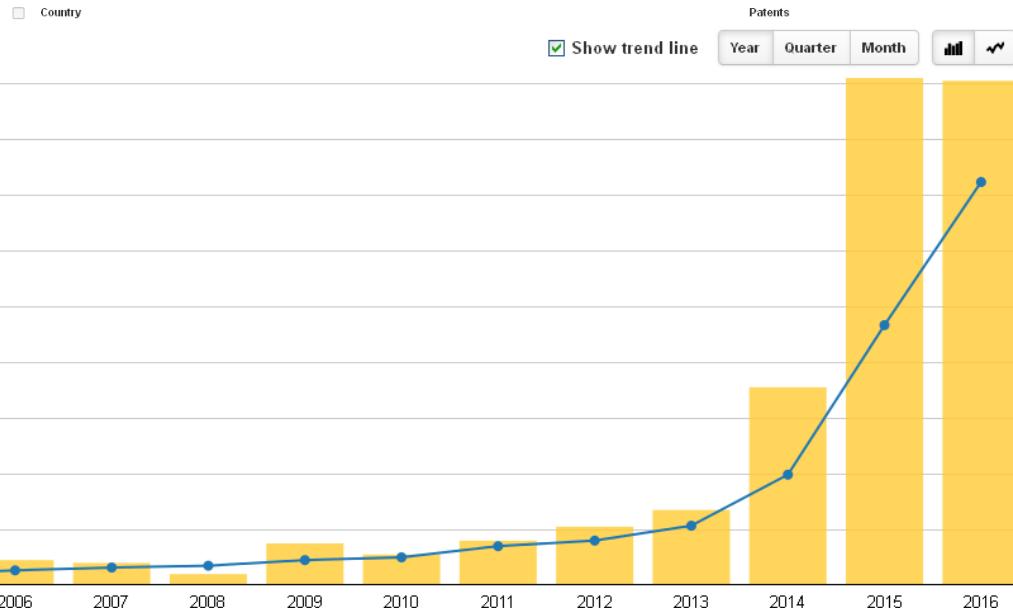


DEEP LEARNING

ABRAMOVITZ ASSI · ADVANCED COMPUTING INC Z · ANHUI COSWIT INFORMATION TECHNOLOGY CO LTD · BALIC JOZE
BEIJING KUANGSHI TECHNOLOGY CO LTD · CHINESE ACAD INST AUTOMATION · CISCO TECH INC
DWAVE SYS INC · EDELMAN GERALD M · FUJI XEROX CO LTD · GOOGLE INC · HILLHOUSE JOHN THOMAS · IBM · INSPUR BEIJING ELECT INF IND · JEKUSU KK · KRICHMAR JEFFREY L
KRIMPOTICH ROBERT · LUCAS PAUL D · MEICHUN YI · MICROSOFT CORP
MICROSOFT TECHNOLOGY LICENSING LLC · MICS INSTR INC · NAT INST INF & COMM TECH
NEC LAB AMERICA INC · NEUROSCIENCES RES FOUND · ROM RAMI · SHANGHAI NO 1 HIGH SCHOOL · SIEMENS AG · SORIN CRM SAS
SULLIAN PHILIP W · TAPIANO CARLOS C · THERANOUR LLC · UNIV BEI HANG · UNIV CHONGQING · UNIV ELECTRONIC SCIENCE & TECH
UNIV NANJING POSTS & TELECOMM · UNIV NORTH CHINA ELEC POWER · UNIV NORTHWESTERN POLYTECHNICAL · UNIV SHANGHAI JIAOTONG
UNIV SHENZHEN · UNIV SOUTH CHINA TECH · UNIV SUN YAT SEN · UNIV TIANJIN · UNIV TSINGHUA · UNIV WUHAN · UNIV XI AN JIAOTONG
UNIV XIDIAN · UNIV ZHEJIANG · WEILLOFER · ZADEH LOTFI A



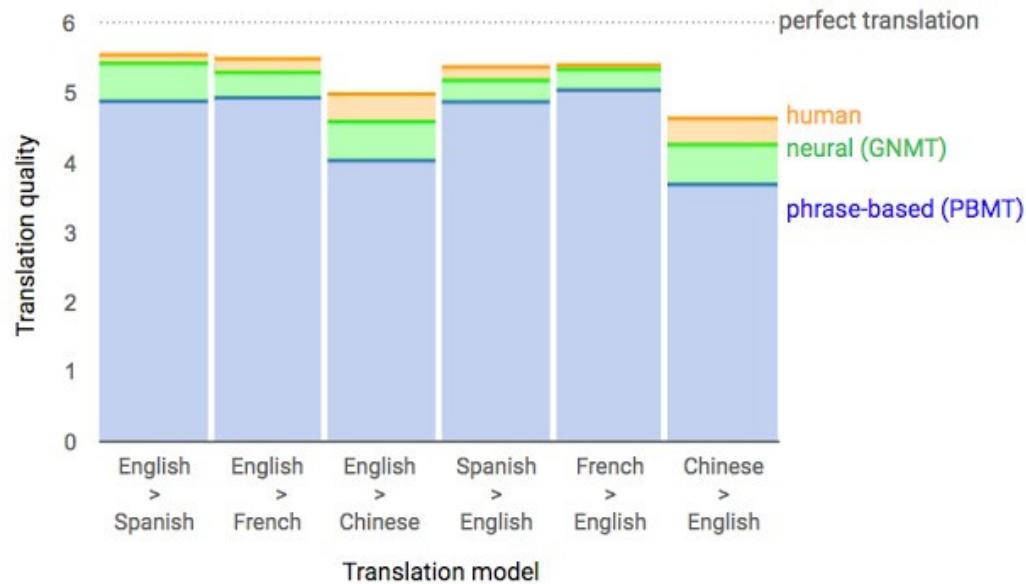
Country





ТЕКУЩИЕ ДОСТИЖЕНИЯ AI

Google Neural Machine Translation



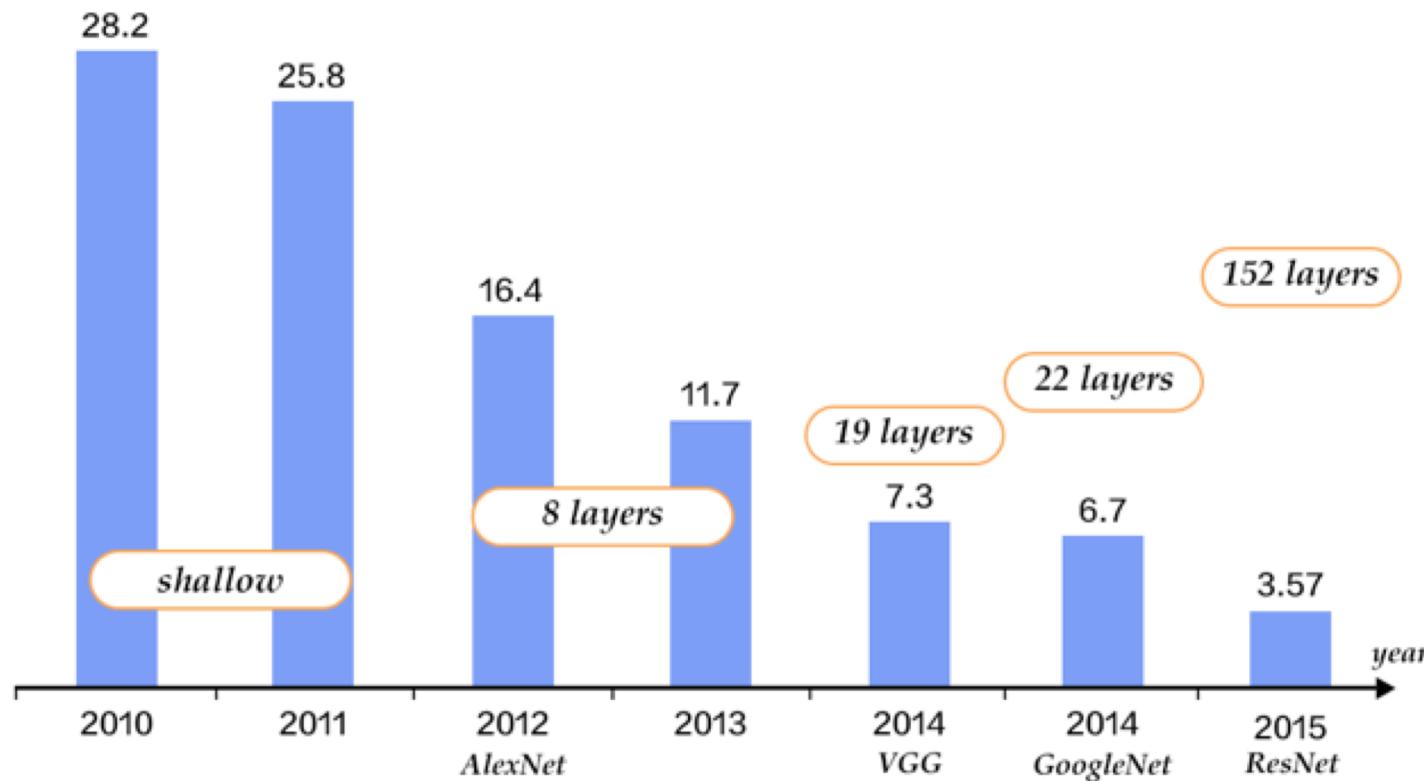
сокращение отставания от человека по точности перевода на 55—85 % (оценивали люди по 6-балльной шкале).

<https://habrahabr.ru/company/mailru/blog/338248/>



ТЕКУЩИЕ ДОСТИЖЕНИЯ АІ

Компьютерное зрение

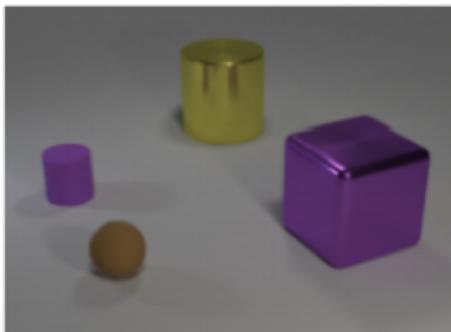




ТЕКУЩИЕ ДОСТИЖЕНИЯ AI

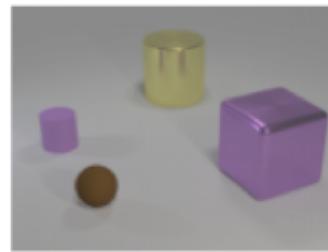
Visual Reasoning

Original Image:



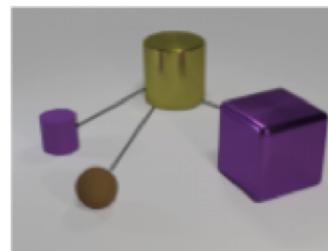
Non-relational question:

What is the size of the brown sphere?



Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?



Нейросеть должна по фотографии ответить на какой-то вопрос. Например: «Есть ли на картинке резиновые вещи того же размера, что и желтый металлический цилиндр?»

Вопрос и правда нетривиальный, и до недавнего времени задача решалась с точностью всего лишь 68,5 %.

И вновь прорыва добилась команда из Deepmind: на датасете CLEVR они достигли super-human точности в 95,5 %.

<https://habrahabr.ru/company/mailru/blog/338248/>



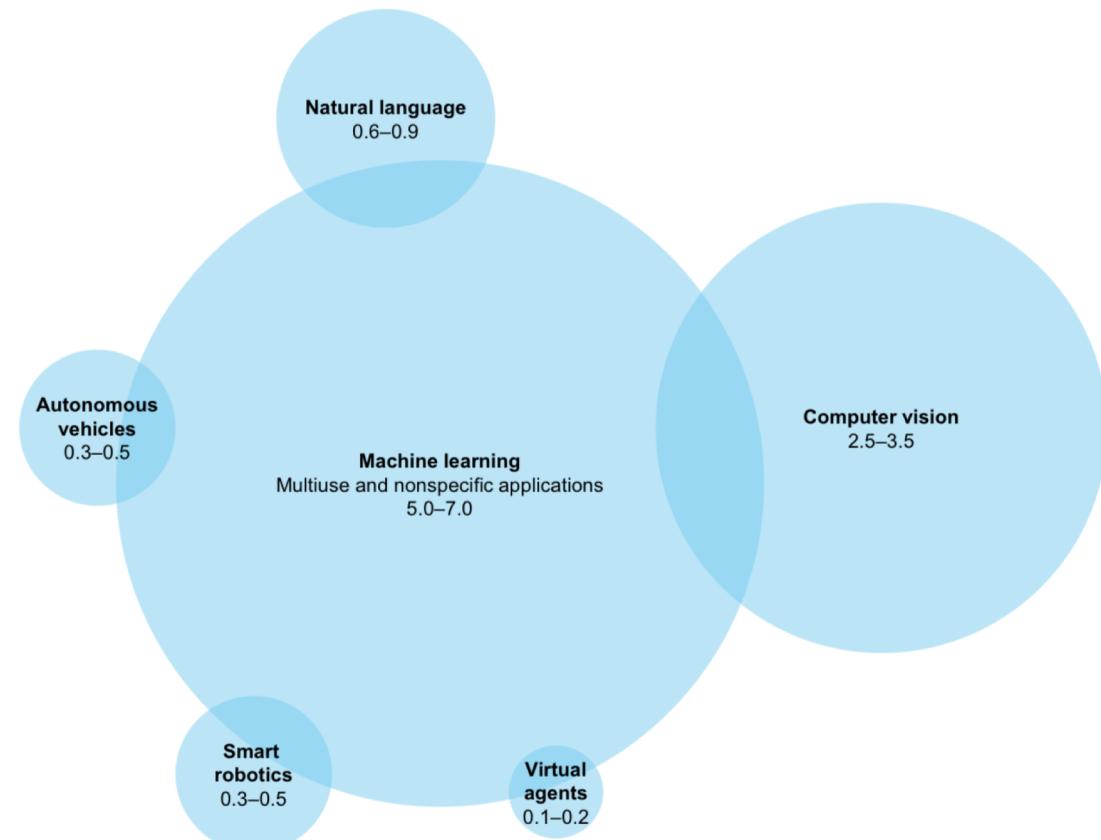
КЕЙСЫ И ПРИМЕНЕНИЕ В БИЗНЕСЕ



КАПИТАЛ

McKinsey Global Institute **оценивает, что в 2016 ИТ гиганты инвестировали \$20-30 млрд в AI, в то время как небольшие компании в совокупности инвестировали около \$6-9 млрд.**

External investment in AI-focused companies by technology category, 2016¹
\$ billion





BIG DATA CASES



“Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don’t think AI will transform in the next several years. “

Andrew Ng



ВЛИЯНИЕ AI

Меняет индустрии



Индустрия

Индустрия

Меняет бизнес компаний



Компания

Компания

Компания

Меняет жизнь людей



Человек

Человек

Человек

Человек

Человек

Человек



ВЛИЯНИЕ АИ НА ЖИЗНЬ ЛЮДЕЙ

- **Нет больше спама:** благодаря компьютерному обучению и годам данных электронной почты Google говорит, что Gmail теперь точно до 99,9 процента при определении (и карантин) спама или фишинговых писем.
- **Автоматические ответы:** специфичные для контекста. Gmail анализирует полученное сообщение («Встречайте во вторник или среду?») И автоматически генерирует ответы на выбор («Let's do Tuesday»).
- **Быстрее Netflix:** когда-нибудь поток фильм по плохой Wi-Fi? Видео становится пиксельным, поскольку приложение пытается использовать меньше данных и продолжать играть. Чтобы предотвратить это, Netflix построил программное обеспечение, которое идентифицирует визуальную сцену, а затем решает, как разделить пропускную способность. Поэтому, когда вы смотрите кульминационную битву в Marvel, Netflix будет использовать всю имеющуюся силу сигнала для сцены.
- **Более длительный срок службы батареи:** ваш iPhone будет анализировать использование приложения и данные датчика движения, чтобы предсказать, что вы делаете в течение дня. Для обновления программного обеспечения требуется скачать большое количество данных. Телефон определяет, когда вы, вероятно, будете ездить на работу. Поэтому вместо того, чтобы налагать заряд аккумулятора вашего телефона на поиск сотовых вышек, чтобы вытащить все эти данные, iOS ждет, пока вы, скорее всего, спите, а телефон подключен к приличному Wi-Fi.

Что дает:

- Снижение рутинной когнитивной нагрузки



АВТО

- Целевая аудитория модели и ее «типовые» привычки
- Used-based страхование и кредитование
- Электронный штурман, рекомендации ТО
- Аномалии поведения и экстренные вызовы
- Рекомендации автодилеру
- Предсказание аварийных ситуаций
- Машины без водителя



Что дает:

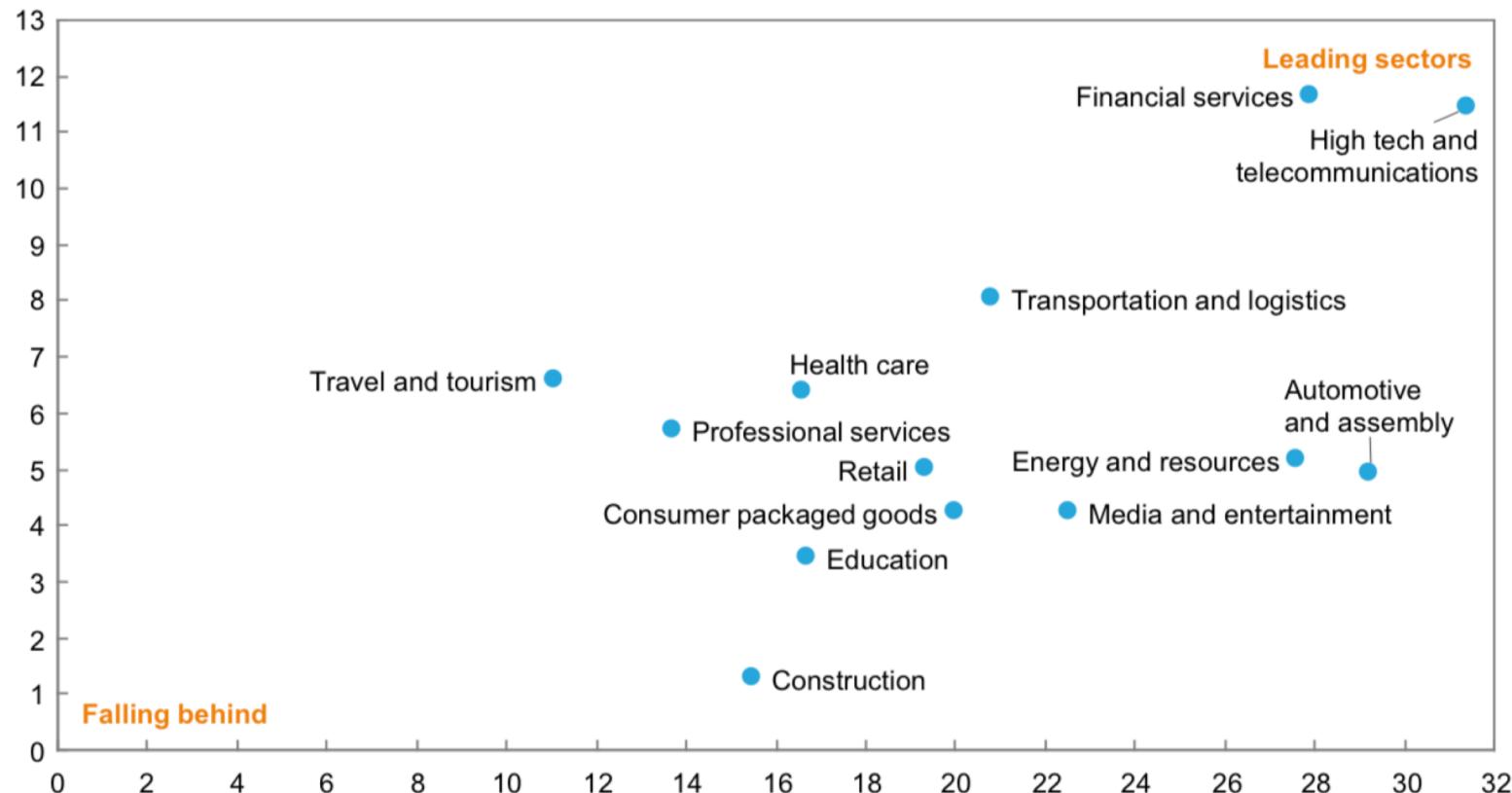
- Снимает с человека рутинную нагрузку



ВЛИЯНИЕ АИ НА ИНДУСТРИИ

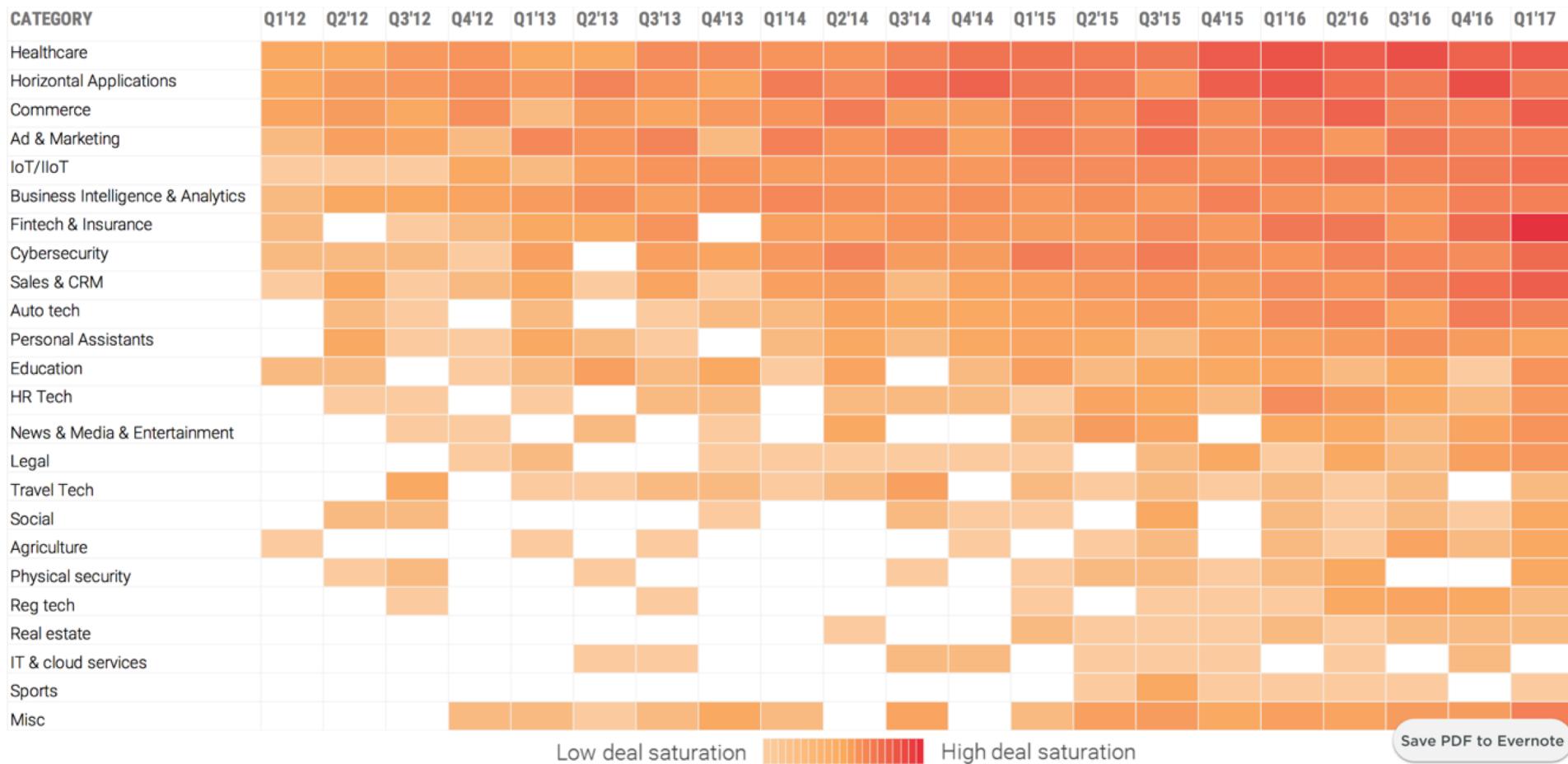
Future AI demand trajectory¹

Average estimated % change in AI spending, next 3 years, weighted by firm size²





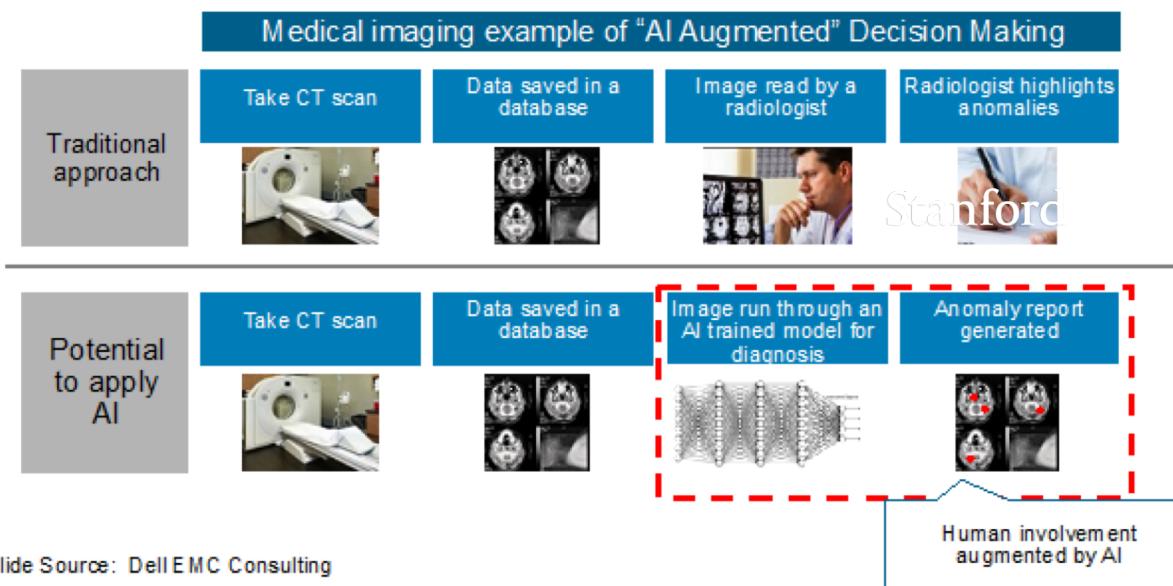
РОСТ КОЛИЧЕСТВА STARTUP-ОВ





AI В МЕДИЦИНЕ

AI | ML | DL Augments Human Decision-making in Healthcare



Что дает:

- доступ к сервису для широкого круга потребителей
- Разгрузку врачей на этапе предварительной диагностики
- Повышение точности диагноза

Алгоритм, созданный Себастьяном Труном, Андре Эстева и Бреттом Купрелом (Sebastian Thrun, Andre Esteva, Brett Kuprel), может обнаружить кератиноцитовую карциному (тип рака кожи), глядя на различные кожные проявления (акне, сыпь, родинку и т.п.). В июне 2015 года он давал правильные ответы в 72% случаев, в то время как два дипломированных врача-дерматолога отвечали верно относительно тех же изображений в 66% случаев

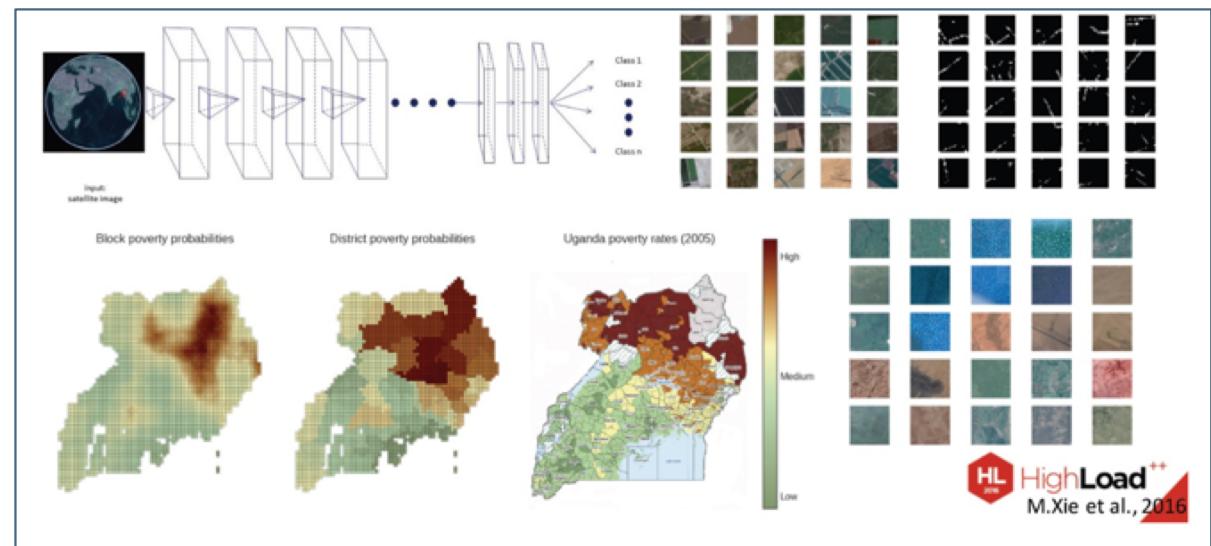
Первые алгоритмы искусственного интеллекта, призванные помочь врачам с диагностикой заболеваний сердца и лёгких, начнут работать в государственных клиниках Великобритании уже летом 2018 года. Услугами ИИ каждый пациент сможет воспользоваться бесплатно.





BIG DATA CASES

- Ученые Стэнфорда недавно придумали очень необычное применение нейронной сети CNN для предсказания бедности



Что дает:

- Повысить скорость реакции на изменение
- Снизить стоимость функции



BIG DATA CASES

- Специалисты земельного кадастра в 2014 году провели аэрофотосъемку Московской области. И сопоставили данные аэрофотосъемки с данными Росреестра
- На сегодня мы выявили уже порядка 250 тысяч объектов недвижимости неоформленных неверно.
- Специалисты земельного кадастра связали ее с земельным реестром. Выявлены незарегистрированные объекты. Объем платежей по ним в бюджет должен составить не менее 56 млн. руб. сразу и затем примерно 10% этой суммы составят ежегодные платежи за строения, ранее не зарегистрированные, за землю, официально не оформленную.

Что дает:

- Повысить скорость реакции на изменение
- Снизить стоимость функции





СПОРТ

- В 2002 году генеральный менеджер бейсбольной команды Oakland Athletics Билли Бин решил разрушить парадигму того, как нужно искать себе спортсменов — он выбрал и обучил игроков «по цифрам»



Что дает:

- Замена «опыта», независимость от экспертов
- Дифференциация от других команд и возможность достигнуть того-же результата меньшими ресурсами



СЕЛЬСКОЕ ХОЗЯЙСТВО

Results: Automatic hyper parameter optimization resulted in a 98.9% accurate model



Slide Source: Dell EMC Consulting

Running a final model on this test image of a tomato leaf suffering from late blight disease yields (score determines how confident the model is in the label):

- Tomato late blight (score = 0.81282)
- Apple cedar apple rust (score = 0.09649)
- Grape healthy (score = 0.04684)
- Apple apple scab (score = 0.00848)
- Grape black rot (score = 0.00651)

Что дает:

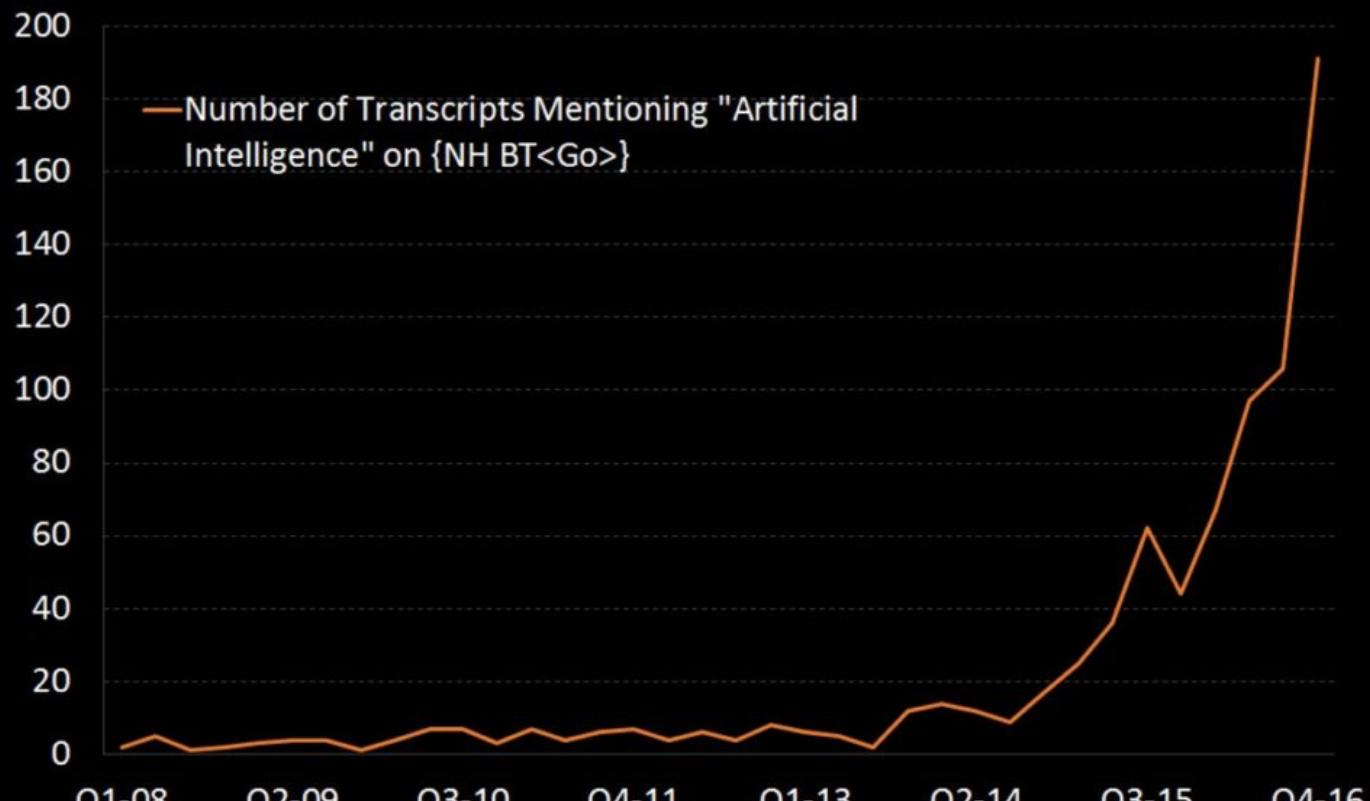
- Повысить скорость реакции на изменение
- Снизить стоимость функции

- Патрулирование дронами
- Контроль уровня удобрений и полива
- Предсказание уровня урожая
- Оптимизация цепочки поставок
- Предсказание погоды

К 2050 году человечество должно увеличить объем пищи на 70%



Companies Mentioning 'Artificial Intelligence' Rising Rapidly



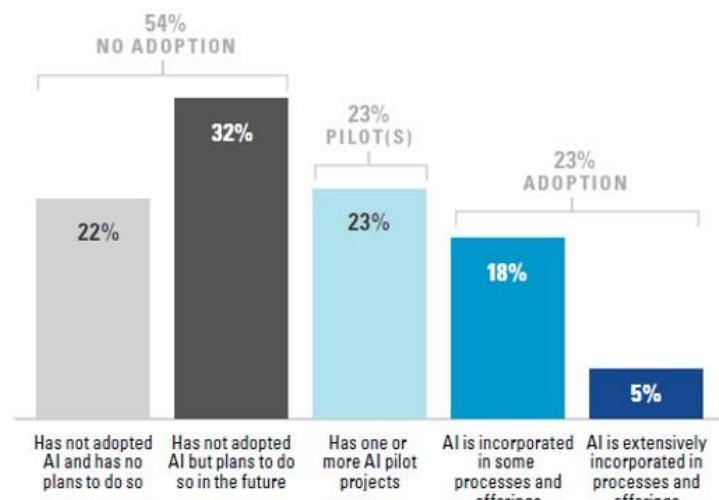
Source: Bloomberg



BIG DATA CASES

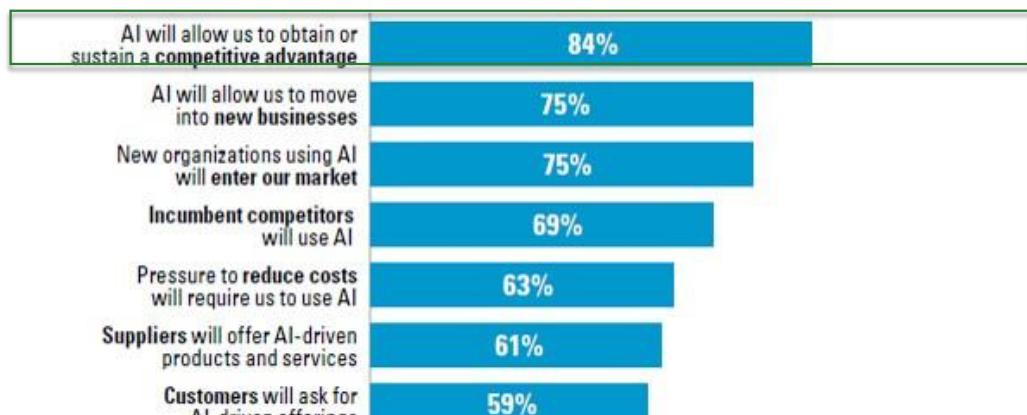
Adoption level of AI

What is the level of AI adoption in your organization?



Reasons for adopting AI

Why is your organization interested in AI?

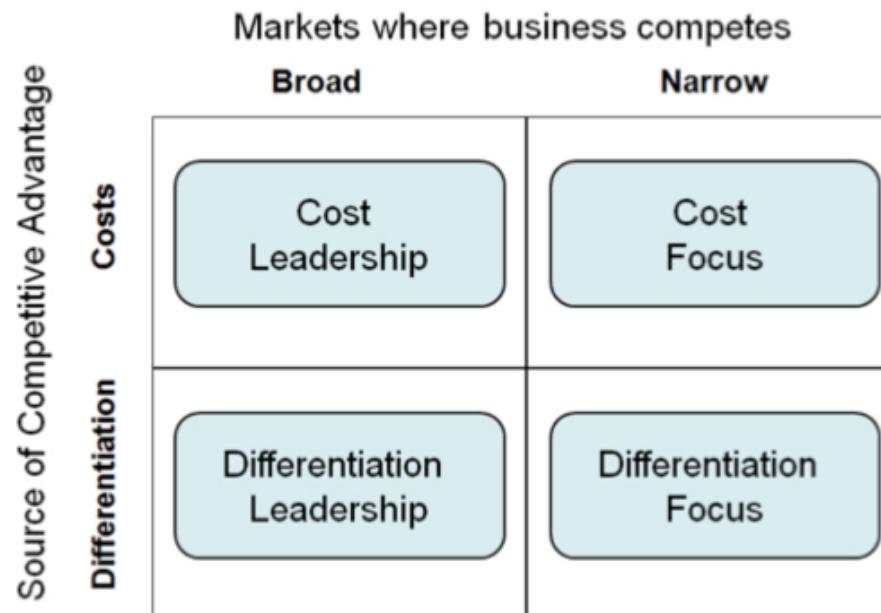


Percentage of respondents who somewhat or strongly agree with each statement



PORTER'S STRATEGIES

Porter's Generic Strategies

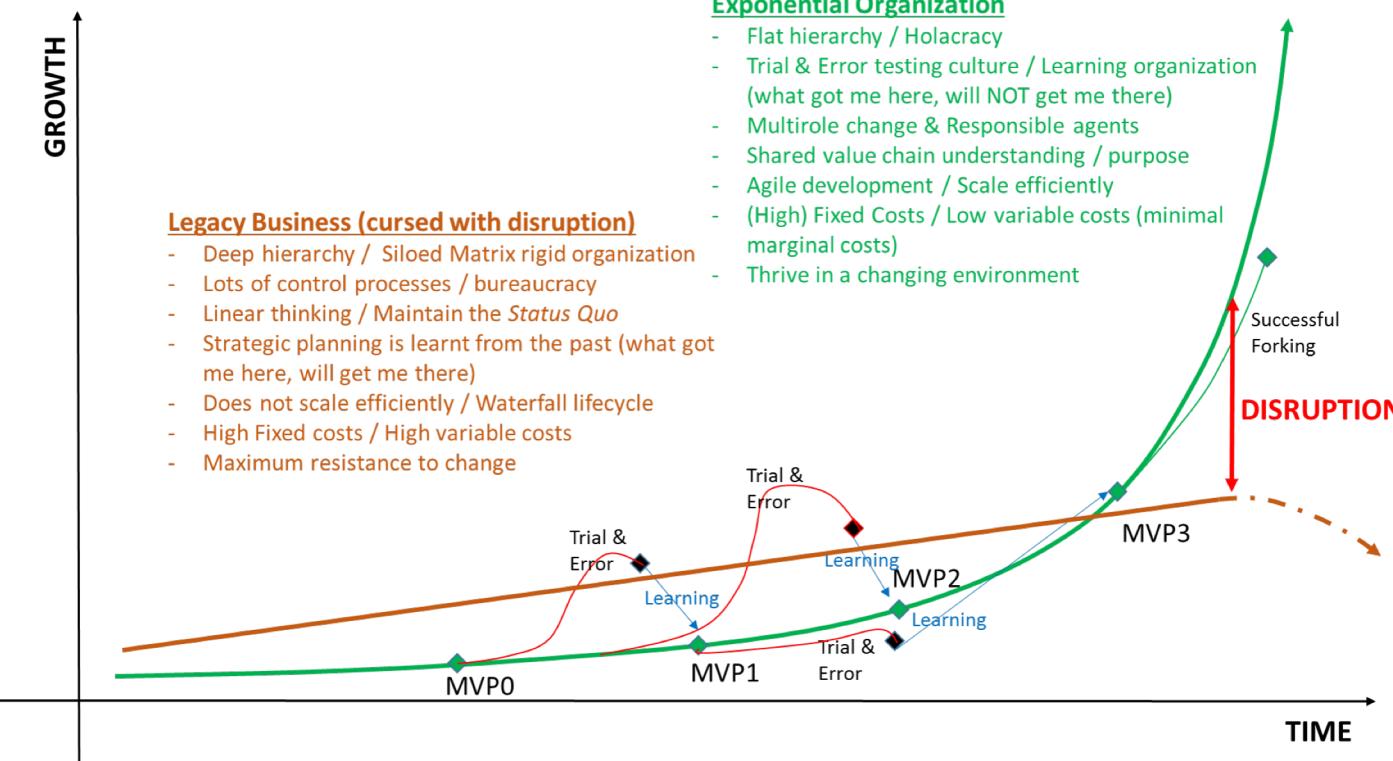


tutor2u

- Какие цели и задачи стоят перед компаний?
- Какую стратегию она старается имплементировать?
- Какие препятствия стоят на пути?

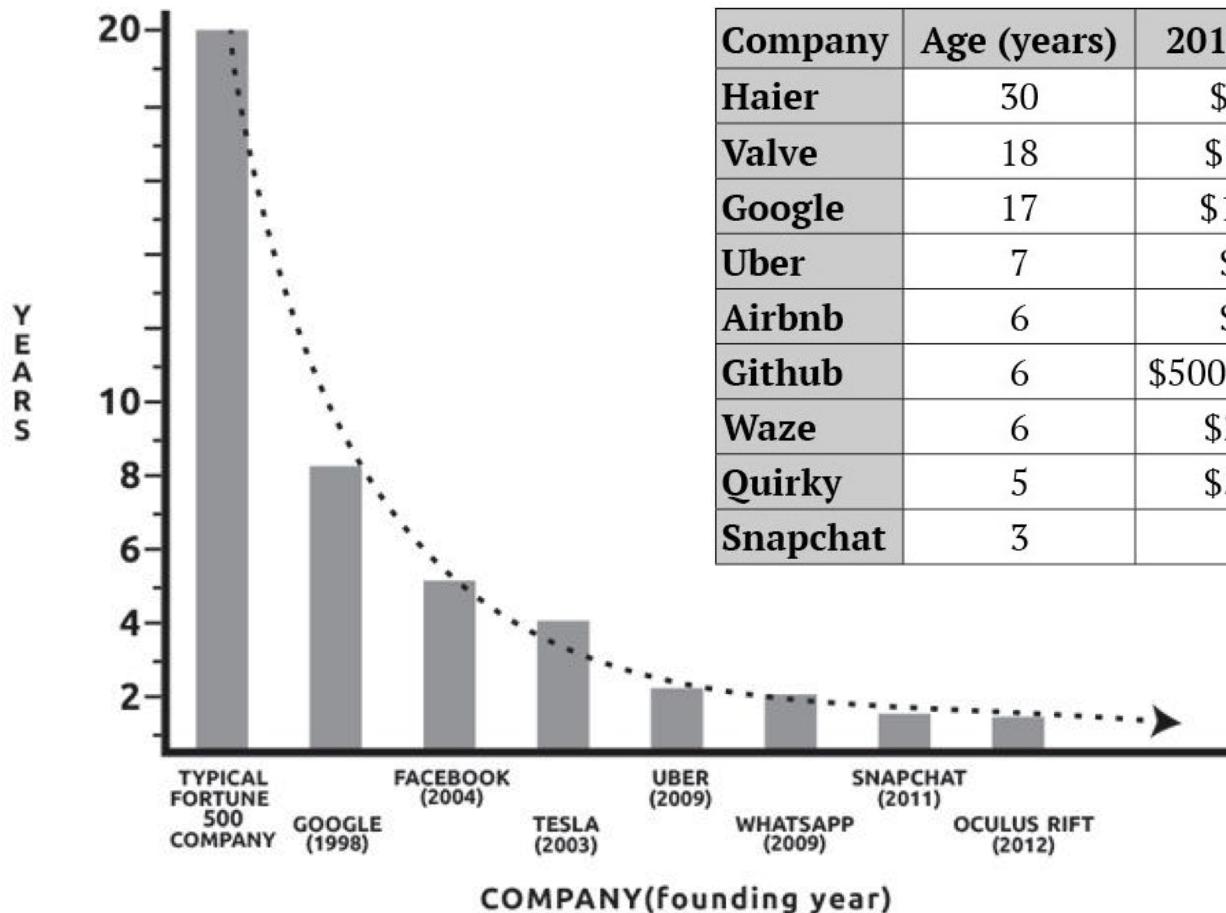


ЭКСПОНЕНЦИАЛЬНЫЕ ОРГАНИЗАЦИИ





ЭКСПОНЕНЦИАЛЬНЫЕ ОРГАНИЗАЦИИ



Brian Chesky
@bchesky

Follow

Marriott wants to add 30,000 rooms this year. We will add that in the next 2 weeks.

Reply Retweet Favorite More HootSuite

RETWEETS 286

FAVORITES 228



11:11 PM - 10 Jan 2014



ЭКСПОНЕНЦИАЛЬНЫЕ ОРГАНИЗАЦИИ

Интерфейсы

Self-service

AdWords от Google, App Store

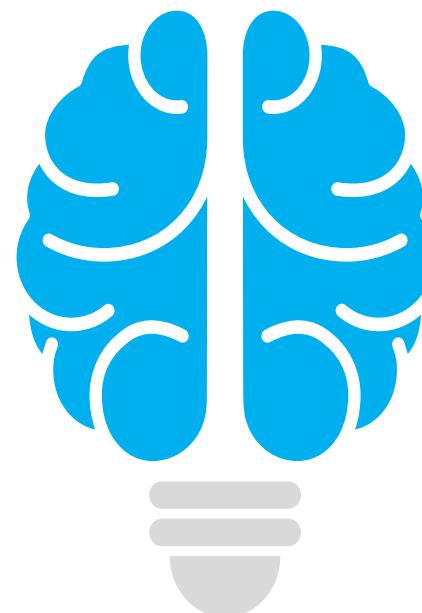
Дашборды

IoT, Real-time

Walmart

Автономность

Outsourcing, Scrum, Valve, Holacracy,
Zappos



Персонал *by demand*

oDesk, Roamler, Elance,
TaskRabbit, Kagle

Алгоритмы

Big Data, AI

Amazon, Google, Netflix

Сторонние активы

Shared economy
Uber, Airbnb, TechShop

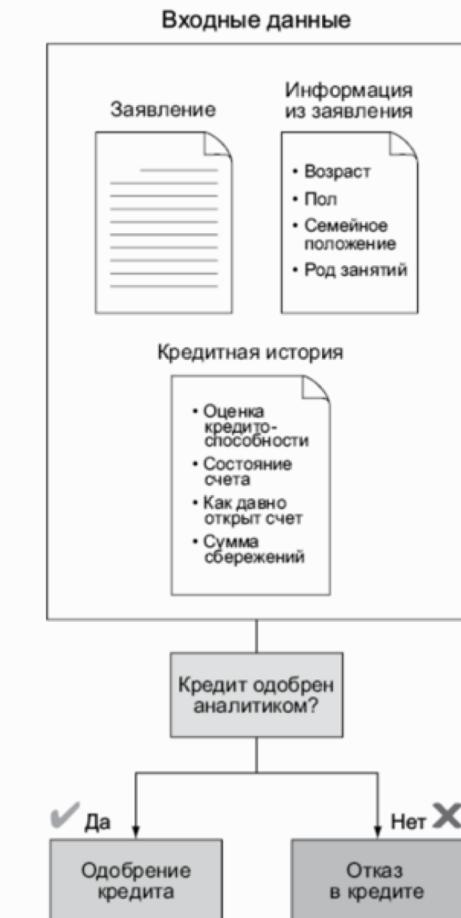
Адаптируемость

Масштабируемость



ПРИМЕР – КРЕДИТНЫЙ СКОРИНГ

- Ручной труд
- Ручная обработка одним компетентным сотрудником
 - По мере роста количества заявок – надо нанимать больше экспертов
 - Падает качество анализа, растут расходы





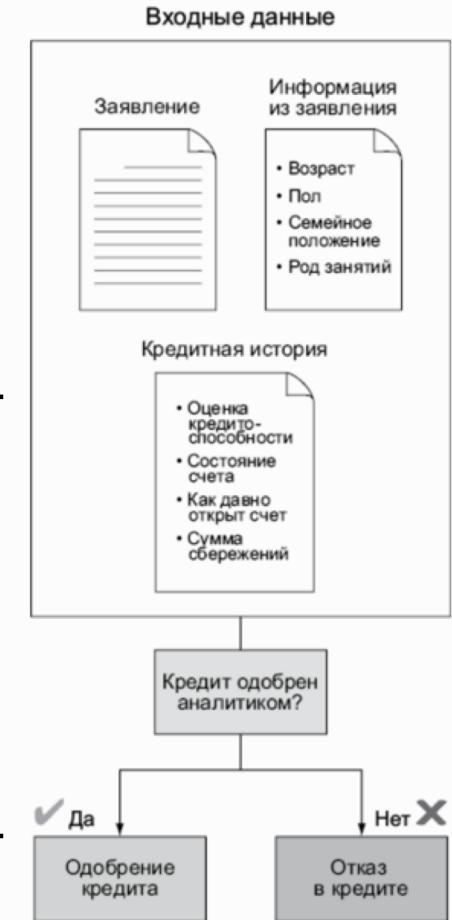
ПРИМЕР – КРЕДИТНЫЙ СКОРИНГ

Ручной труд

- Ручная обработка одним компетентным сотрудником
- По мере роста количества заявок – надо нанимать больше экспертов
- Падает качество анализа, растут расходы

Автоматизация
На базе правил

- Для сохранения качества мы вводим правила (стандартная экспертная система)
 - Правила сложные
 - Меняется бизнес-ситуация
 - Мы не можем проанализировать все ситуации
 - Система не учится



КОГНИТИВНЫЕ ИСКАЖЕНИЯ



Когда запоминаем и вспоминаем

Когда много информации

Когда быстро реагируем

Чтобы действовать, должны
быть уверены в способности
что-то изменить
и чувствовать важность
своих поступков

designblocks.co | Категоризация Buster Benson
Алгоритм и дизайн оригинал John Manoogian III
Данные wikiedia.org

 attribution · share-alike



ПРИМЕР – КРЕДИТНЫЙ СКОРИНГ

Ручной труд

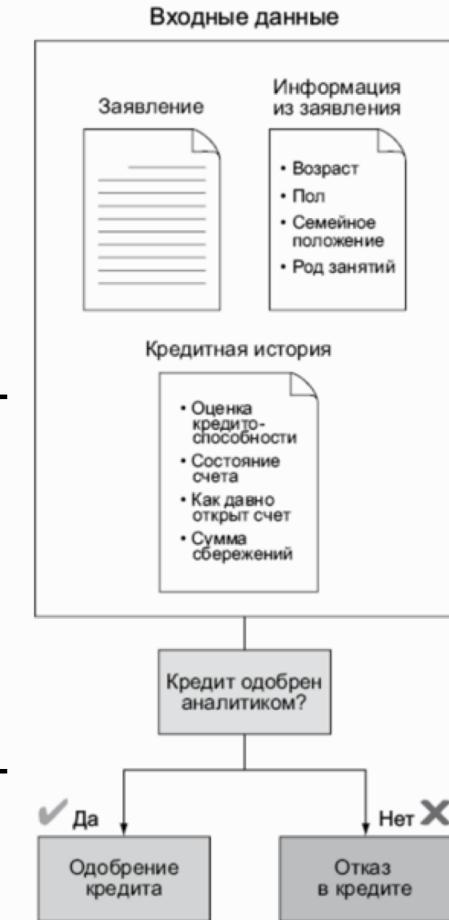
- Ручная обработка одним компетентным сотрудником
- По мере роста количества заявок – надо нанимать больше экспертов
 - Падает качество анализа, растут расходы

Автоматизация

- Для сохранения качества мы вводим правила (стандартная экспертная система)
 - Правила сложные
 - Меняется бизнес-ситуация
 - Мы не можем проанализировать все ситуации
 - Система не учится

AI

- Внедрение систем предиктивной аналитики
- Возможность роста бизнеса без дополнительных затрат





ПРЕИМУЩЕСТВА МАШИННОГО ОБУЧЕНИЯ

- **Точность.** Машинное обучение использует данные для создания принимающей решение программы, оптимизированной под поставленную задачу. По мере накопления данных автоматически возрастает точность прогнозов.
- **Скорость.** Машинное обучение дает ответы за доли секунды после поступления новой информации, позволяя системам реагировать в реальном времени.
- **Автоматизация.** По мере подтверждения и отбрасывания ответов ML-модель может автоматически обнаруживать новые шаблоны. Это позволяет встраивать машинное обучение непосредственно в автоматизированные рабочие процессы.
- **Масштабируемость.** При росте бизнеса ML-модель легко приспосабливается к увеличивающимся объемам данных. Некоторые алгоритмы можно использовать для обработки множества данных на разных вычислительных машинах в облаке.



ОСНОВНЫЕ ЗАДАЧИ AI

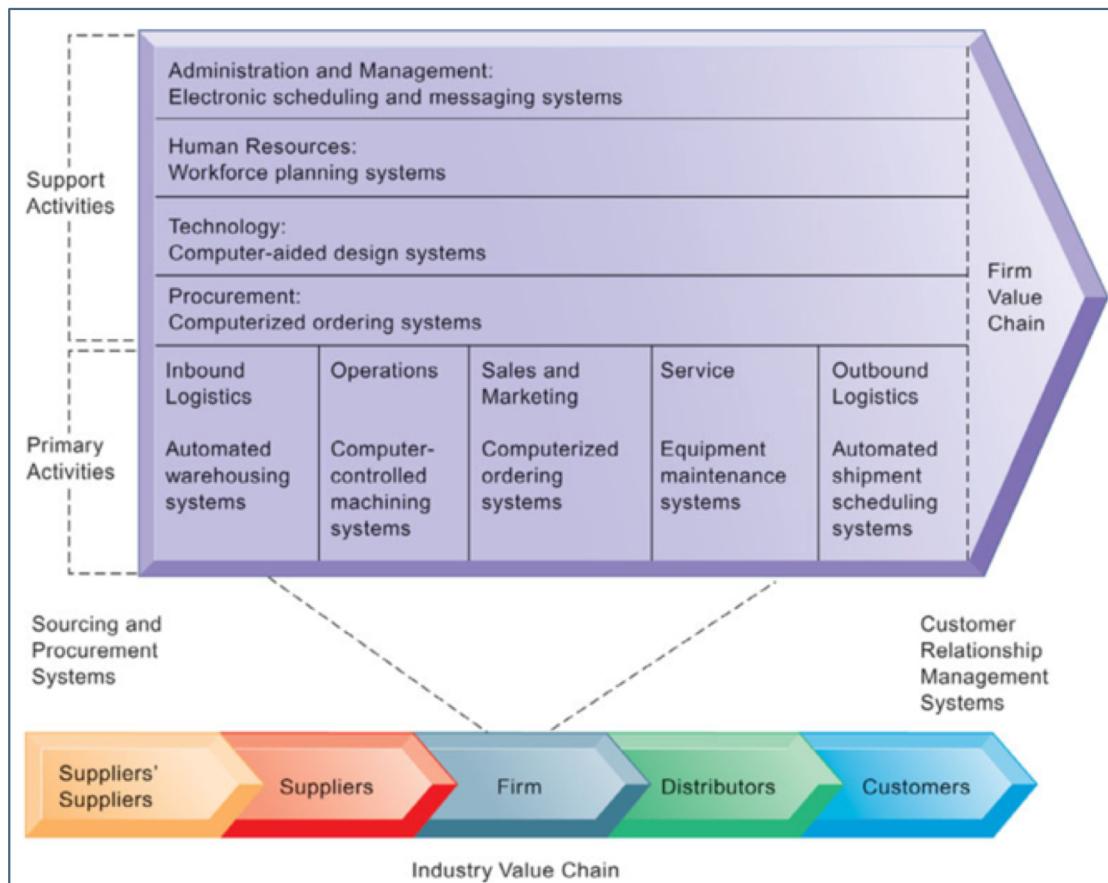
Возможность точно отработать большее количество гипотез, чем может сделать человек

- Insights
- Optimization
- Decision making

На текущий момент AI находится на самой ранней стадии своего развития и обладает способностью автоматизировать (масштабировать) рутинные когнитивные операции



БИЗНЕС ПРОЦЕССЫ ОРГАНИЗАЦИИ



1. Управляющие — бизнес-процессы, которые управляют функционированием системы.

2. Основные (увеличение дохода) — бизнес-процессы, которые составляют основной бизнес компании и создают основной поток доходов.

3. Поддерживающие (снижение издержек) — бизнес-процессы, которые обслуживают основной бизнес.



ЛОГИСТИКА

В столице Бурятии Улан-Удэ 2 апреля состоялся тестовый запуск первого в российской истории почтового дрона. Описав дугу в воздухе, беспилотник на полной скорости врезался в стену соседнего дома, чудом не разбив чье-то окно. Обломки аппарата, как видно на многочисленных видео с места событий, упали в считанных сантиметрах от случайно проходивших мимо людей.

Что дает:

- Снизить стоимость функции
- Качественно изменить уровень сервиса





HR

"I no longer look at somebody's CV to determine if we will interview them or not,"

Teri Morse, XEROX

Компания Xerox осуществляет набор сотрудников call-centre (штат 55 000 человек через предварительный скринг-тест, который предсказывает, какое время человек проработает в компании:

Люди, имеющие 1-2 аккаунта в соц сетях работают дольше чем те, кто имеет 3-4

Факт работы на аналогичной позиции не влияет на успех кандидата



"I don't know why this works," admits Ms Morse, "I just know it works."

86% компаний из списка Fortune 1000 собираются внедрить эти подходы в ежедневную практику

Что дает:

- Снизить стоимость функции
- Масштабировать функцию
- Повысить качество



HR – ЧАТ БОТЫ

- Банк «Открытие» завершил пилотный проект по автоматизации найма совместно с платформой Skillaz. В первую неделю работы робот нашел больше кандидатов, чем 5 ресечеров за месяц, причем эти сотрудники максимально четко попадают в критерии подбора.
- Пропускная способность рекрутеров увеличилась в два раза за счет замены одного из этапов очного интервью онлайн оценкой роботом.
- При дальнейшем использовании платформы ожидается сокращение времени на подбор персонала в 2 раза, а также сокращение стоимости найма на 42%, говорится в пресс-релизе компании.

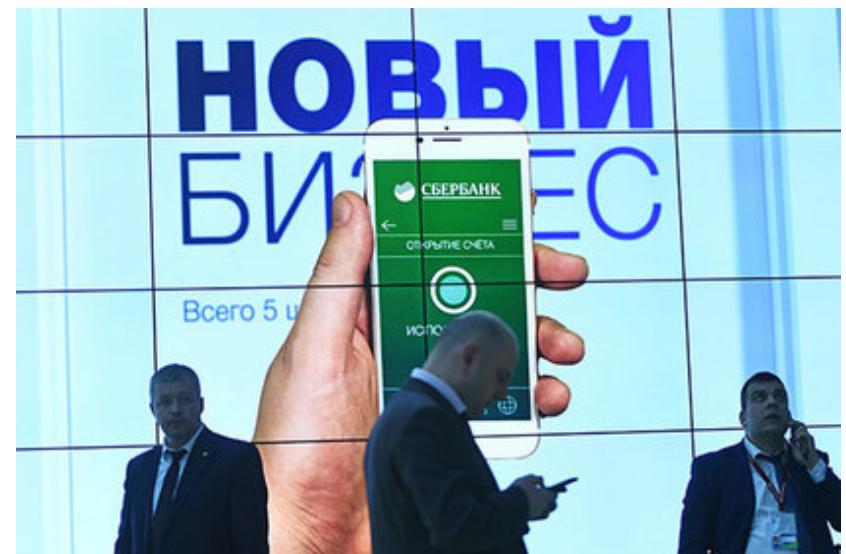


Что дает:

- Снизить стоимость функции
- Масштабирование функции

КЕЙС СБЕРБАНКА

- Сбербанк приступил к роботизации контактного центра, теперь на вопросы корпоративных клиентов будет отвечать робот по имени Анна. Об этом в понедельник, 26 февраля, сообщает пресс-служба кредитной организации.
- В январе стало известно, что Сбербанк намерен уволить три тысячи сотрудников и заменить их роботами. Еще в конце 2016 года заработал робот-юрист, составляющий исковые заявления по физическим лицам.



Что дает:

- Снижение стоимости функции
- Масштабирование функции



ТЕЛЕКОМ

«Мегафон» запустил Федеральную систему управления радиосетью (ФСУР) во всех своих филиалах, которая в автоматическом режиме осуществляет задачу коррекции и сверки параметров базовых станций сотовой связи.

Зачем нужна такая система

Необходимость во ФСУР связана с тем, что в сети «Мегафона» работает большое количество оборудования различных производителей: Nokia, Huawei, Ericsson, Siemens, NEC. Для корректной работы сети необходимо, чтобы параметры оборудования находились в правильном и согласованном состоянии, причем эта работа должна проводится в автоматическом режиме.

ФСУР еженедельно передает в сеть около 700 тыс. команд на коррекцию. «Для подачи такого количества команд вручную потребовался бы ежедневный труд двух тысяч инженеров, - отмечают в пресс-службе «Мегафона». - Результат внедрения автоматических сверок заметен на ключевых показателях функционирования сети».



Что дает:

- Снизить стоимость функции
- Качественно изменить уровень сервиса

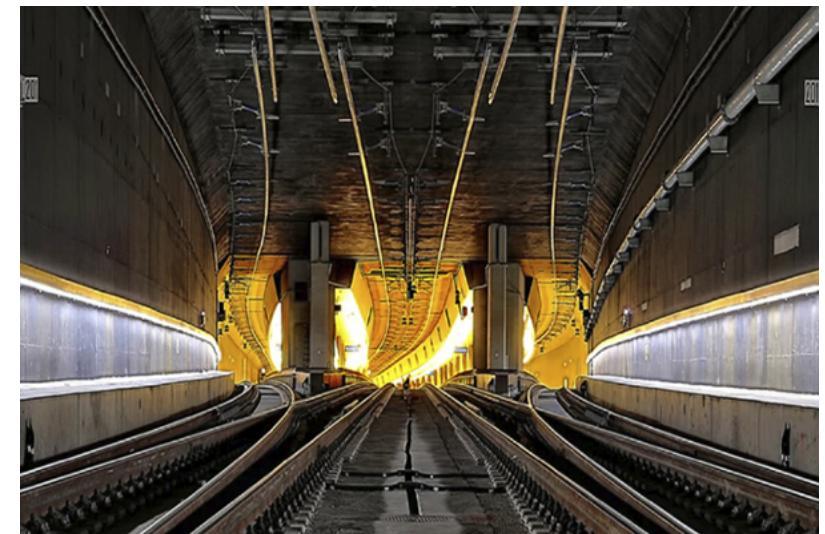


BIG DATA НА ТРАНСПОРТЕ

Предиктивное обслуживание

- Нивелирование человеческого фактора
- Учет условий эксплуатации и нагрузки на оборудование
- Минимизация простоев техники

Десятки тысяч локомотивов под брендом Trenitalia сегодня работают безотказно, компания сократила расходы по статье «техобслуживание» на 8–10%, а клиенты получили возможность доехать от Милана до Рима — без опозданий и задержек — за 2,5 часа и всего 40 евро.





РИТЕЙЛ – ПЛАНИРОВАНИЕ МАГАЗИНОВ

У сети «Пятерочка» в 2014 году насчитывалось 4789 магазинов, в 2016-м их число практически удвоилось — до 8363. Внедрив геоинформационную систему, компания стала открывать по 5—6 точек в день. Для разработки этой системы «Пятерочка» закупила базы данных по домохозяйствам (количество жителей конкретных населенных пунктов, их доходы, проходимость торговых точек), наличию конкурирующих магазинов, собрала информацию по всем объектам недвижимости, которые она когда-либо оценивала, а также учла расходы и доходы собственных магазинов.



Что дает:

- Оптимизация операционной деятельности



РИТЕЙЛ



- Увеличение пропускной способности магазина
- Увеличение среднего чека
- Экономия на зарплате
- Экономия на обслуживании касс



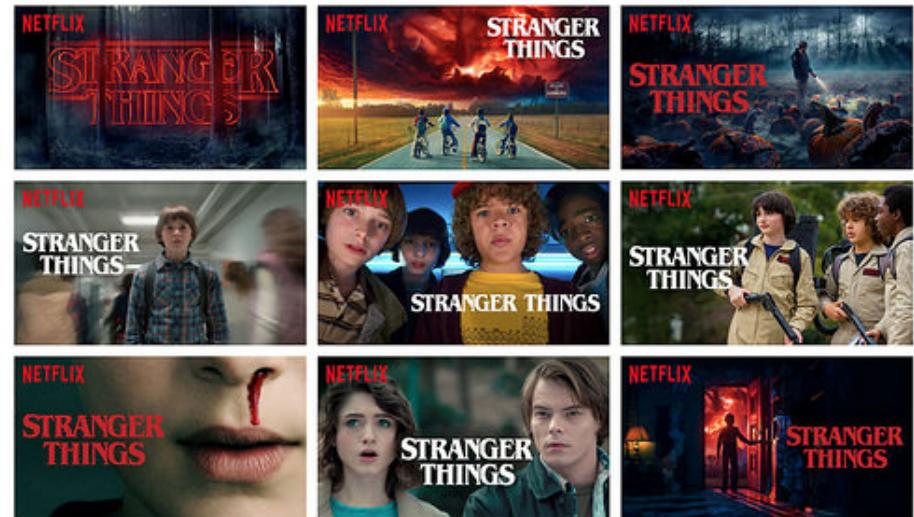
ПЕРСОНАЛИЗАЦИЯ В NETFLIX

Формирование обложки в соответствии с предпочтениями пользователями, которое определяется путем массового А/Б тестирования.

Выбор картинки для обложки, в соответствии с предпочтениями пользователя, такими как любимый актер, жанр.

Формирование витрины контента на основе рекомендательных моделей.

Повышение вероятности просмотра

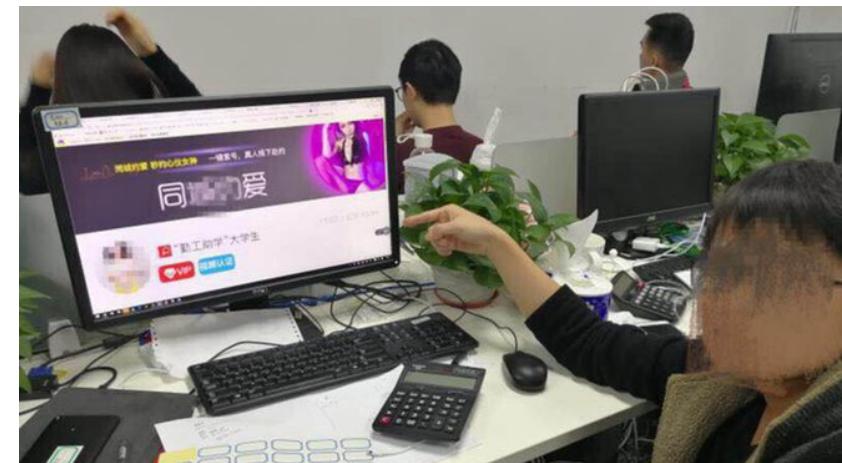


БОТЫ

Правоохранительные органы Китая раскрыли масштабную сеть из компьютерных программ, которые под видом девушек общались с пользователями дейтинговых приложений. Полиция закрыла около 20 компаний и задержала более 600 подозреваемых из 13 провинций страны.

«Они [программы] выпрашивали подарки и отправляли другие сообщения, чтобы пользователи тратили больше денег. Это незаконное получение прибыли».

Полиция считает, что алгоритмы заставили десятки тысяч пользователей потратить около 154 миллионов долларов



Что дает:

- Масштабировать функцию, снизить стоимость

СИНЕРГИЯ КОМПЬЮТЕРА И ЧЕЛОВЕКА

Table 1 - Accuracy and ex post Sharpe ratio results by type of prediction markets

	Accuracy		Sharpe Ratio	
	MSE	LSR	AMSE	ALSR
			(Benchmark: 0.75)	(Benchmark: 1.70)
Humans-only Markets	0.19	0.25	0.41	0.41
Agents-only Markets	0.17	0.23	0.39	0.37
Hybrid Markets	0.15	0.21	0.74	0.72

Table 2: Area under the ROC curves – all three conditions

	Area under ROC Curve	SE ¹
Humans-only Markets	0.76	0.03
Agents-only Markets	0.81	0.03
Hybrid Markets	0.90	0.02

Figure 1: ROC Plots for Study 1

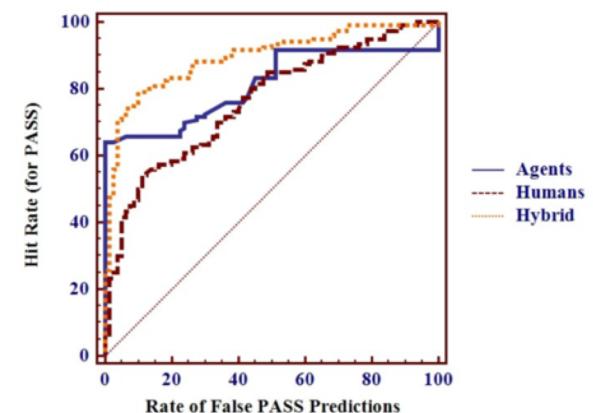
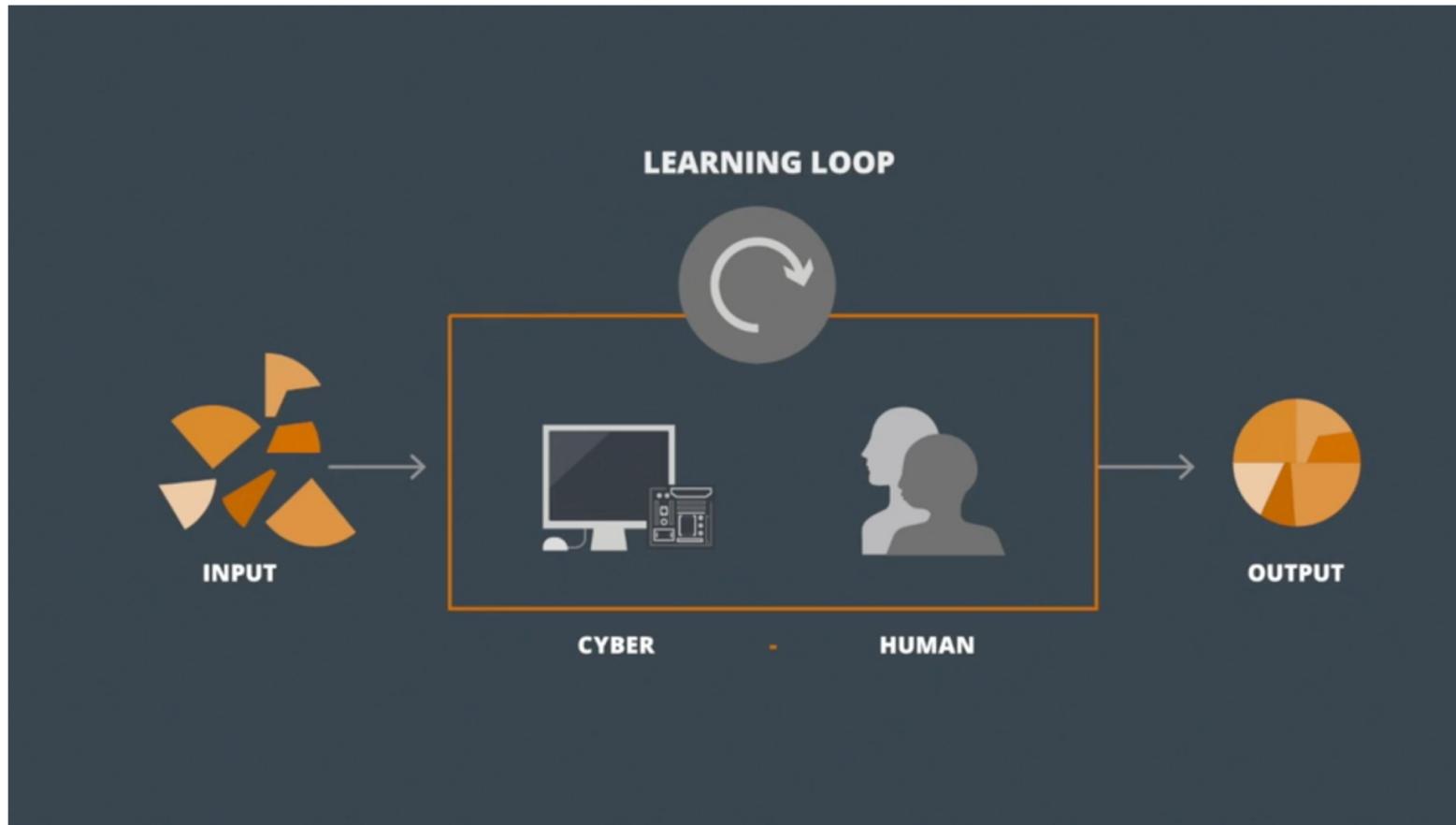


Table 2: Area under the ROC curves – all three conditions

Center for Collective Intelligence Massachusetts Institute of
Technology Cambridge, MA 02142

CYBER/HUMAN LEARNING LOOP

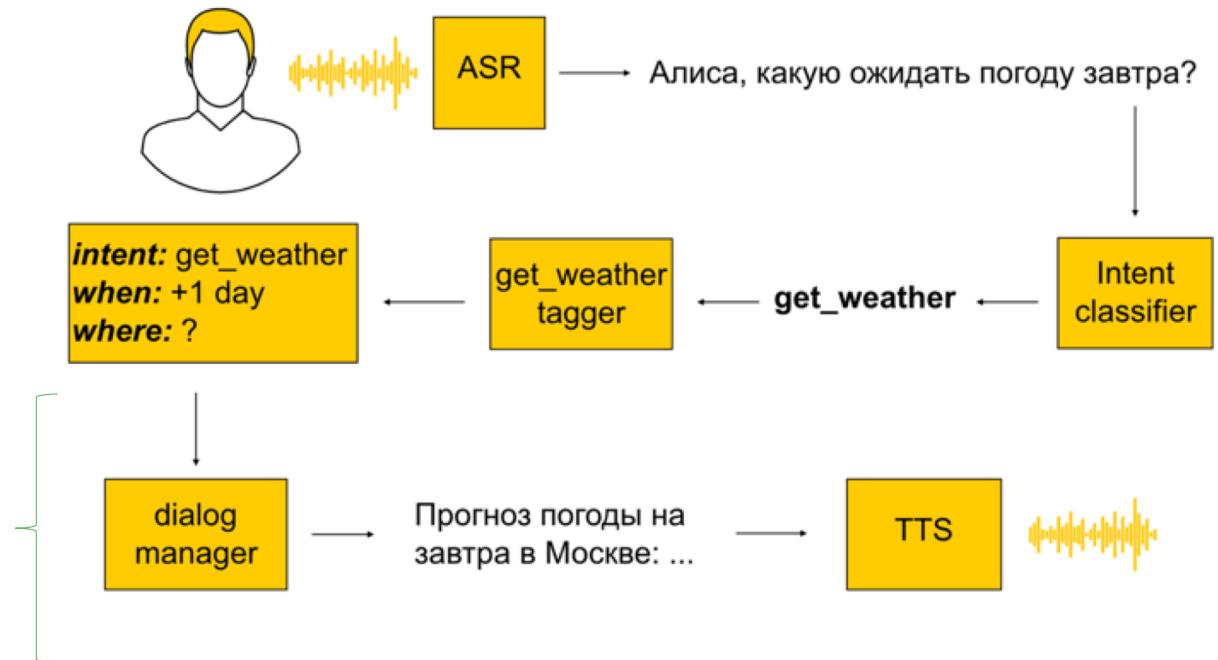


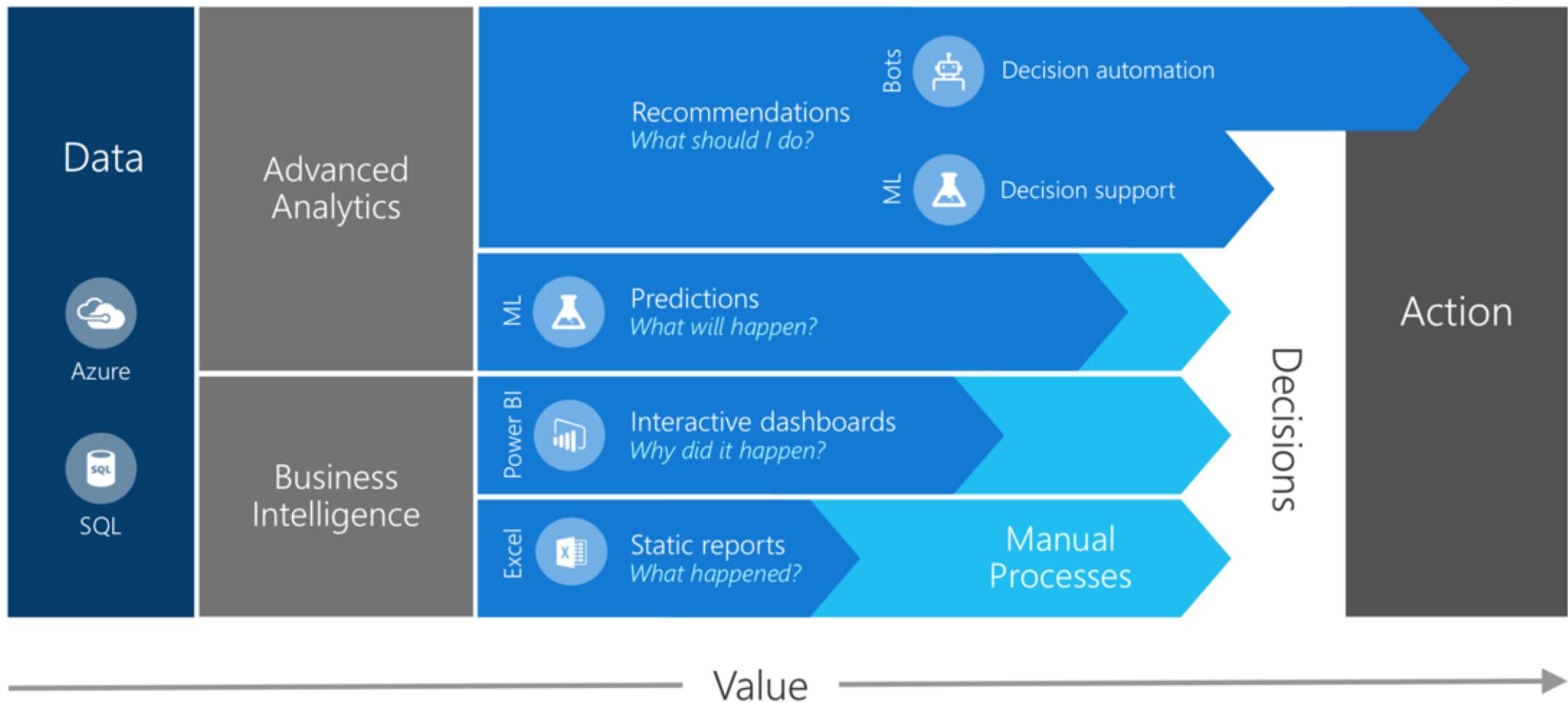
CYBER/HUMAN LEARNING LOOP

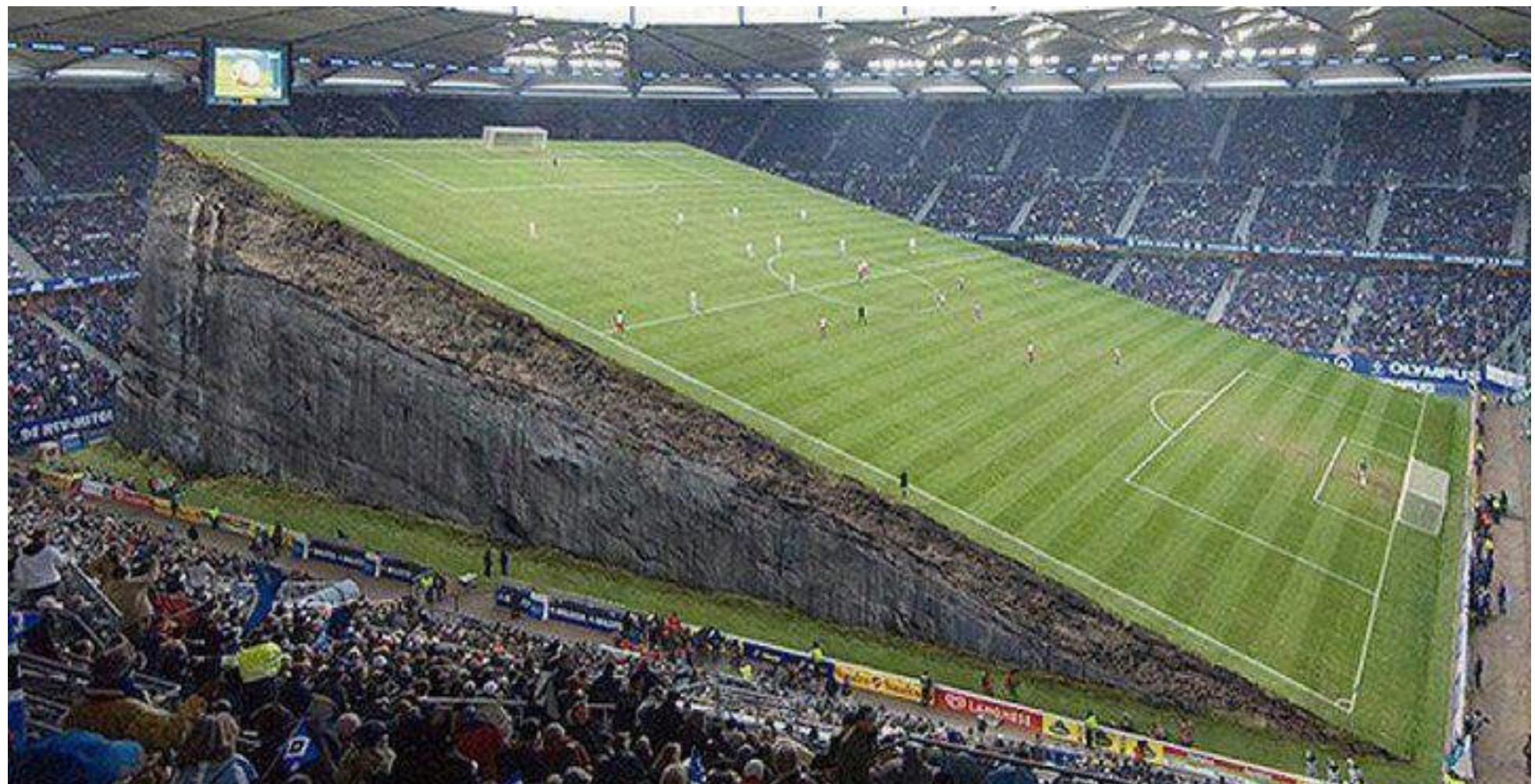
Алиса, вид сверху

Язык шаблонов для генерации текстов

```
1  [% nlginclude "common/error.nlg" %]
2  [% nlginclude "common/suggests.nlg" %]
3  [% nlginclude "scenarios/intents/cards/call.nlg" %]
4
5  [% phrase render_result %]
6  [% if form.recipient_info %]
7  [% if context.attention and context.attention.attention_type == "calls_not_supported_on_device" %]
8  {%- chooseline %}
9    Телефон {{ form.recipient_info.title | inflect('gen') }} - {{ form.recipient_info.phone }}, звоните скорее со с
10   {{ form.recipient_info.title | capitalize_first }} - номер {{ form.recipient_info.phone }}, наберите на телефон
11   {{ form.recipient_info.phone }} - номер {{ form.recipient_info.title | inflect('gen') }}, наберите скорее со с
12   Я вспомнил за вас, {{ form.recipient_info.title }} - номер {{ form.recipient_info.phone }}, звоните скорее с та
13   Звоните с телефона я {{ form.recipient_info.title | inflect('acc') }} по номеру {{ form.recipient_info.phone }}
14  {%- endchooseline %}
15  [% else %]
16  {%- chooseline %}
17    Вызываю {{ form.recipient_info.title | inflect('acc') }}...
18    Звоню я {{ form.recipient_info.title | inflect('acc') }}...
19  {%- endchooseline %}
20  [% endif %]
21  [% else %]
22  {%- chooseline %}
23    Пока что я умею звонить только в экстренные службы.
24  {%- endchooseline %}
25  [% endif %]
26  {%- endphrase %}
```









ОСНОВНЫЕ ДРАЙВЕРЫ

- Данные
- Вычислительная мощность
- Капитал (применение)
- Специалисты

The Beginning

