

Введение в машинное обучение

Кирилл Святов

Ульяновский Государственный Технический Университет

k.svyatov@ulstu.ru

План

Краткая история развития технологий

Технологии ближайшего будущего

Определения

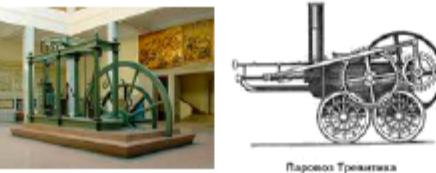
Формулировка проблем ML

Определения

Применение технологий ML

Методы ML

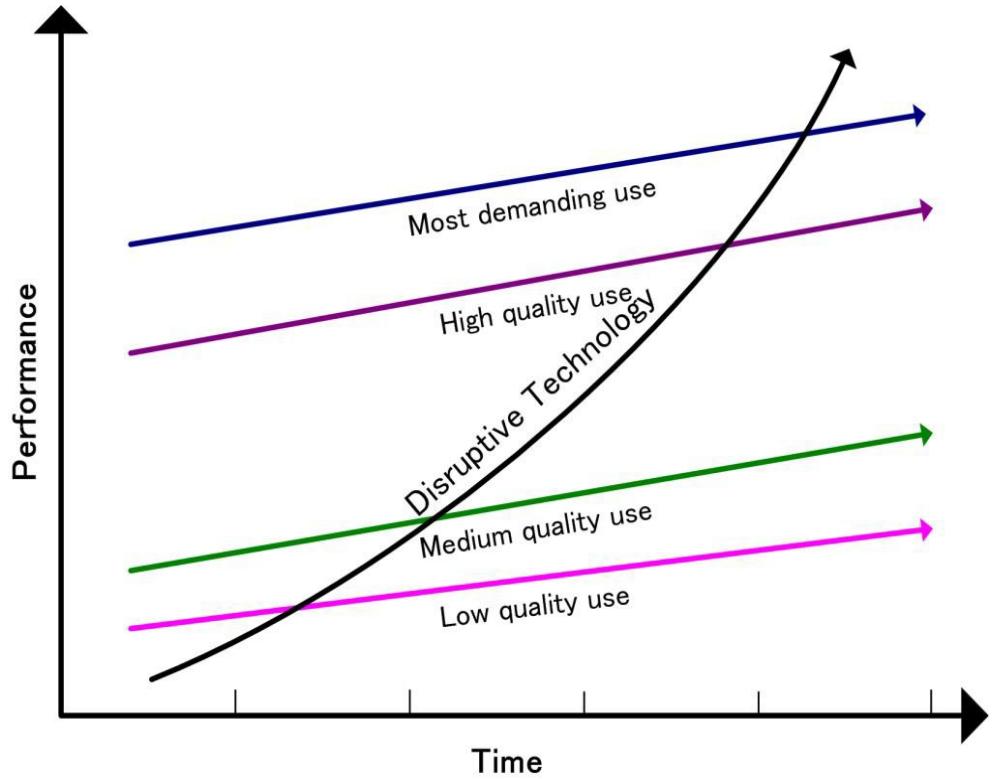
Промышленные революции

Technologies	«I» IR XVII century	II IR XVIII – XIX	III IR XIX – XX	IV IR XXI
	<ul style="list-style-type: none">▶ wood▶ peat burn▶ wind energy 	<ul style="list-style-type: none">▶ Iron, steel▶ coal burn▶ steam▶ surgery, anesthesia▶ agriculture machines 	<ul style="list-style-type: none">▶ plastic▶ oil, electricity▶ cars, airplains▶ satellites▶ antibiotics, early diagnosis▶ mineral fertilizers 	<ul style="list-style-type: none">▶ 3D printing & materials▶ thin films▶ clean energy▶ robots▶ small satellites▶ organic agriculture 

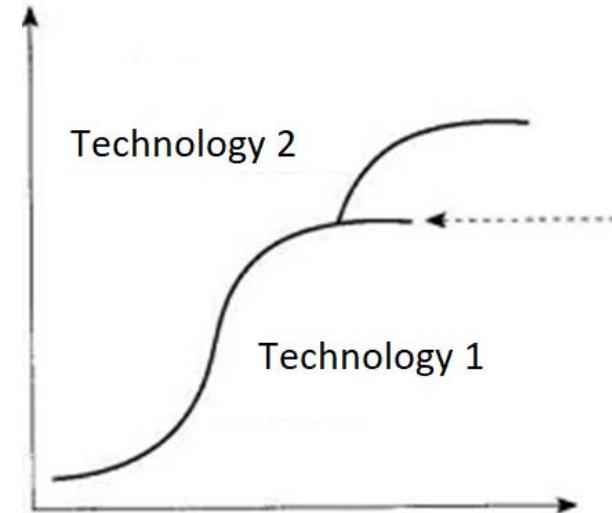


¹ <http://www.tv2.tomsk.ru/sites/www.tv2.tomsk.ru/files/mobil.jpg>

Подрывные инновации

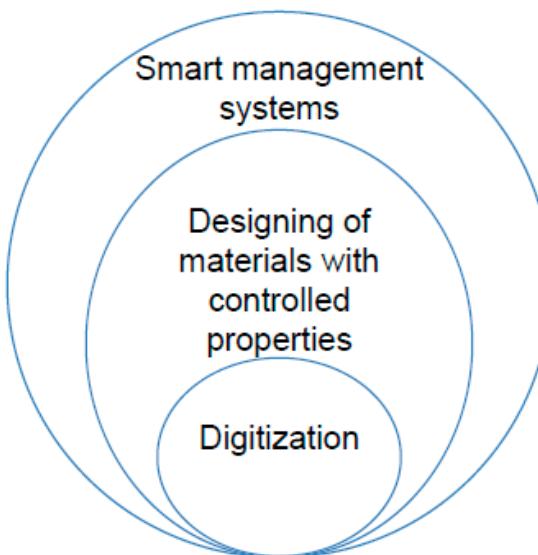


Clayton Cristensen



Индустрія 4.0

Corridors for the development of technology and activities:



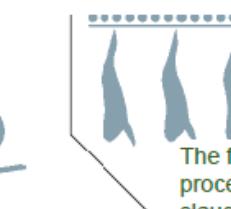
4 stages of industrial revolution

Industrie 1.0 follows the introduction of water and steam mechanical manufacturing enterprises



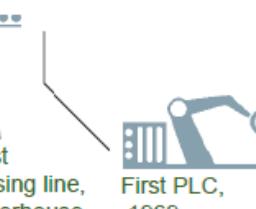
Mechanical loom, 1784

Industrie 2.0 comes as a result of the introduction of electrified mass production, based on the division of labor



The first processing line, slaughterhouse 1870

Industrie 3.0 uses electronics and information technology for further automation of production



First PLC, 1969

Industrie 4.0 based on cyber physical systems



Complexity

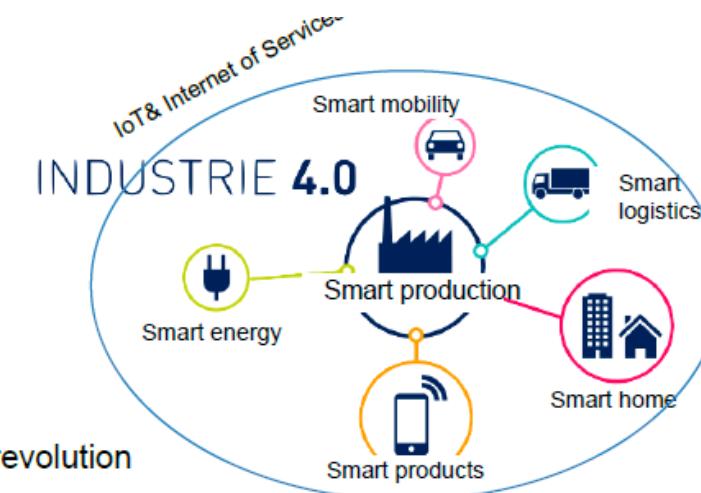
1800

1900

2000

heute

t



Производительность труда

Assessment of the growth of labor productivity from the development of new technologies in production

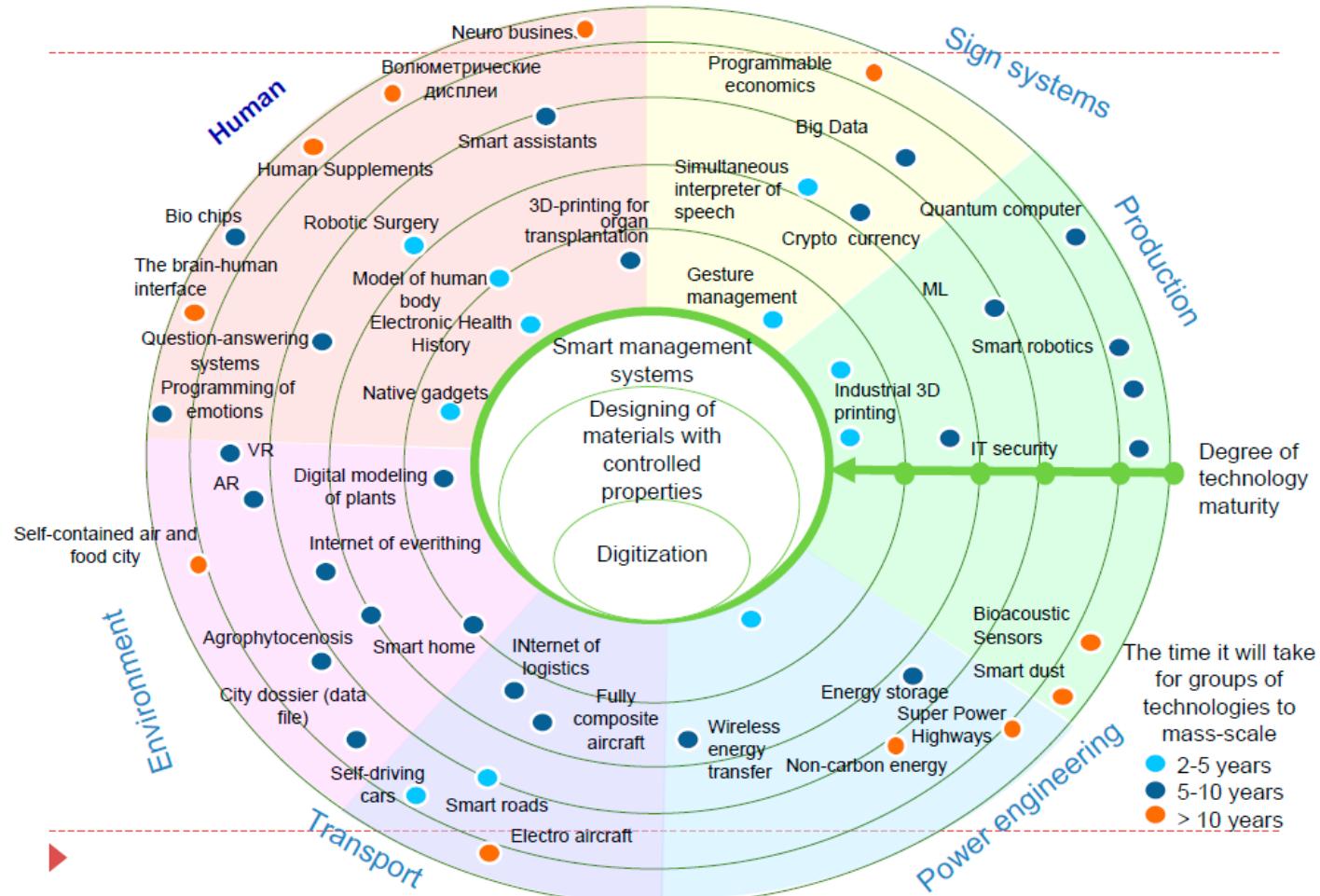


Steam engine	Early robotics	IT	Robots, AI, ML
1850-1910	1993-2007	1995-2005	2015-2065
Growth in labor productivity per year			
0,3%	0,4%	0,6%	0,8 - 1,4%

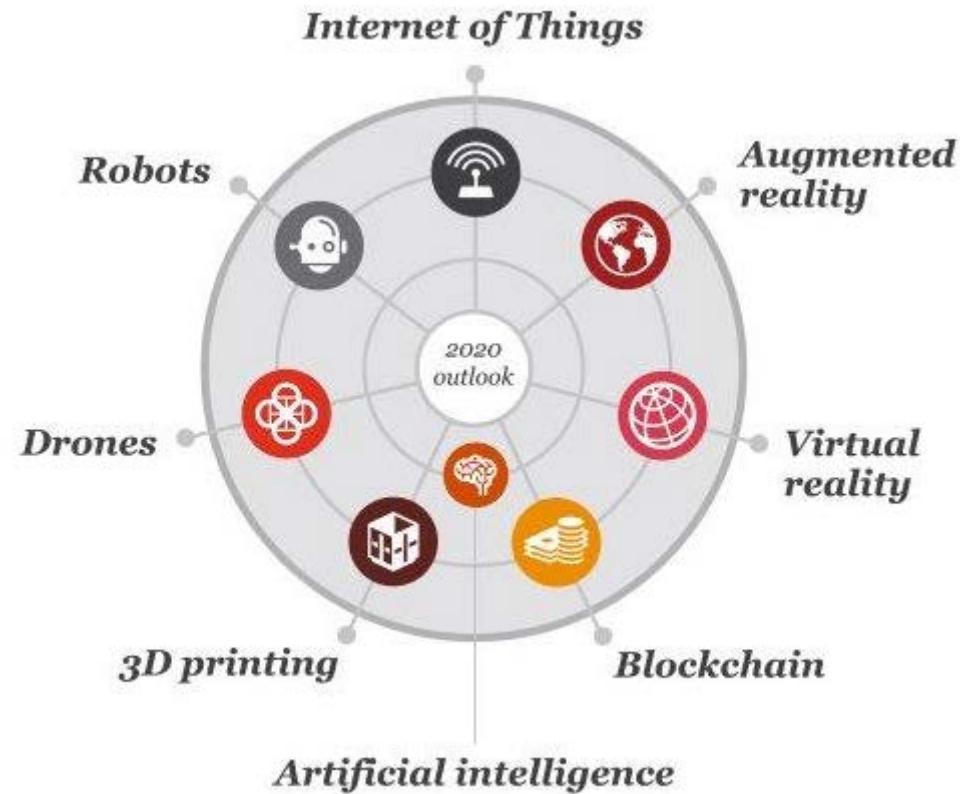
►
¹ McKinsey & Company, <http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works>

Современные технологии

Today there is a new platform of technologies of the New Industrial Revolution



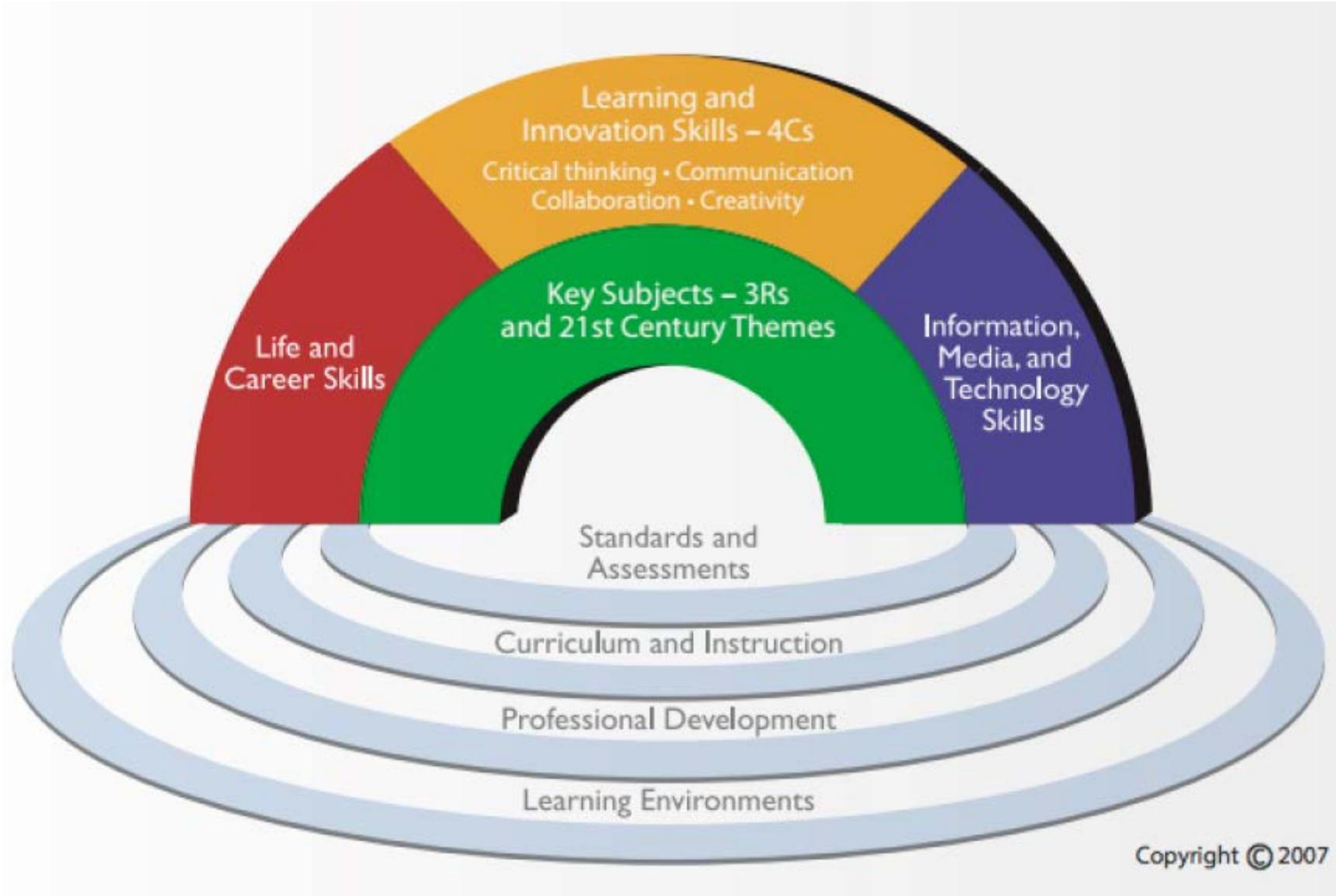
8 ключевых технологий

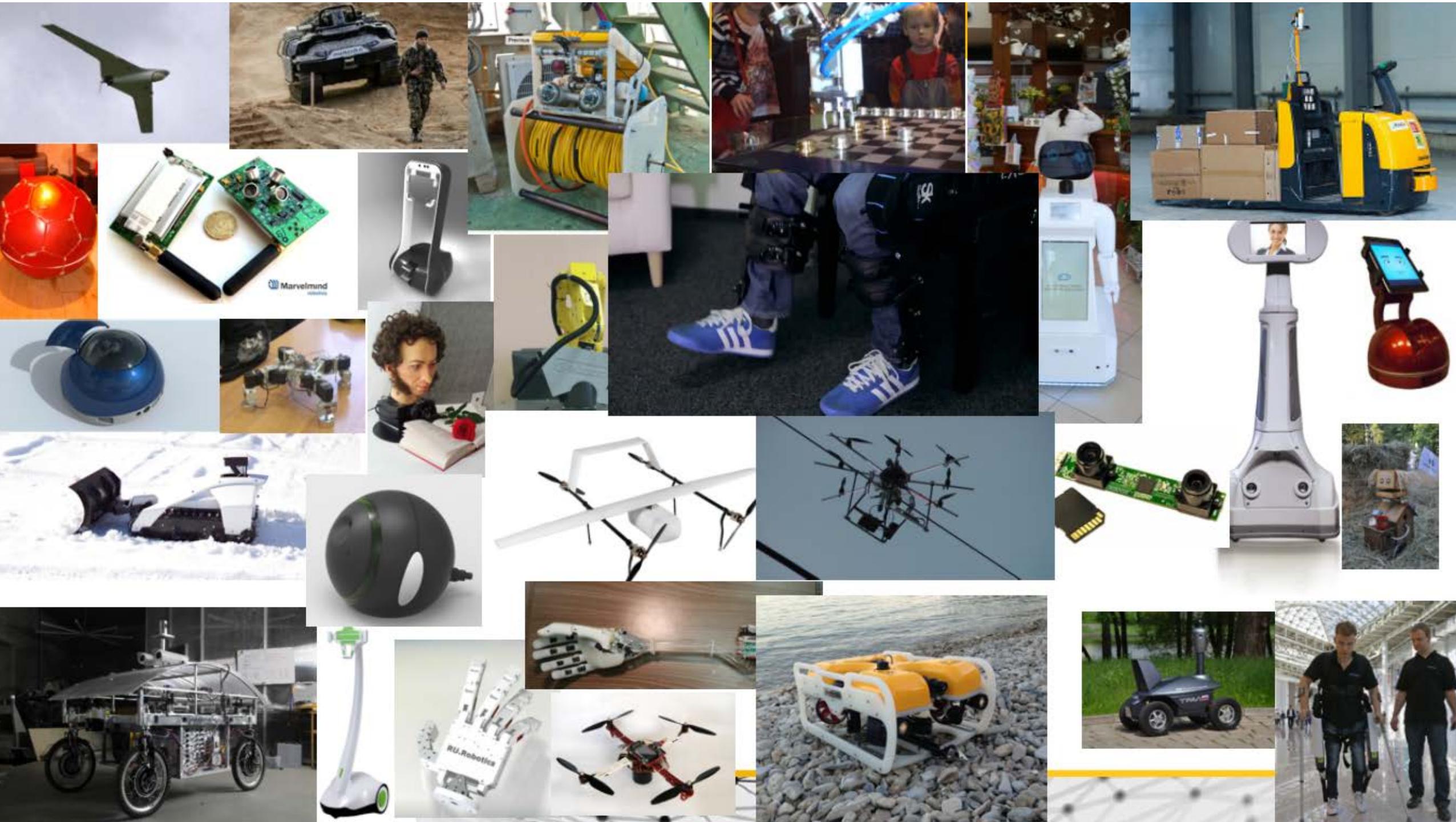


Проблемы трансформации

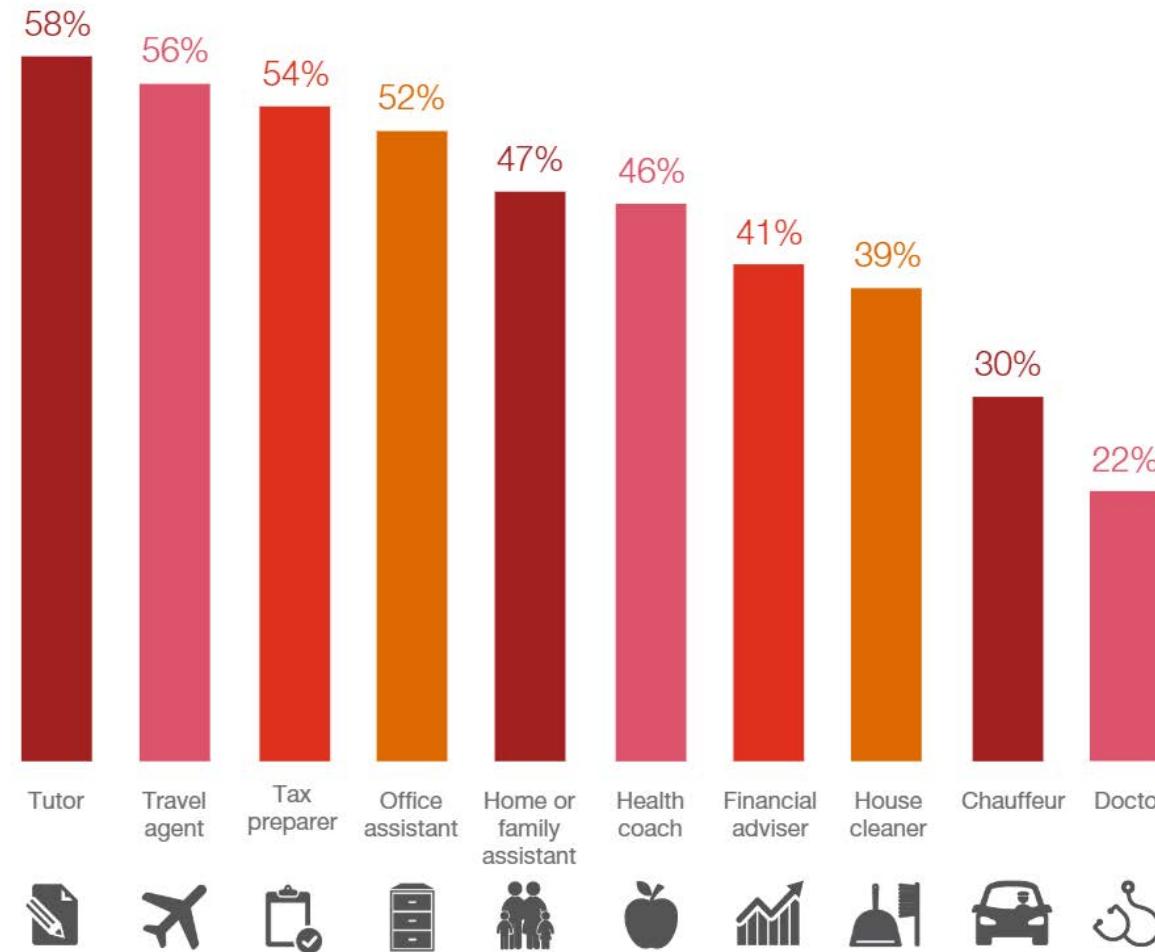


Важные навыки

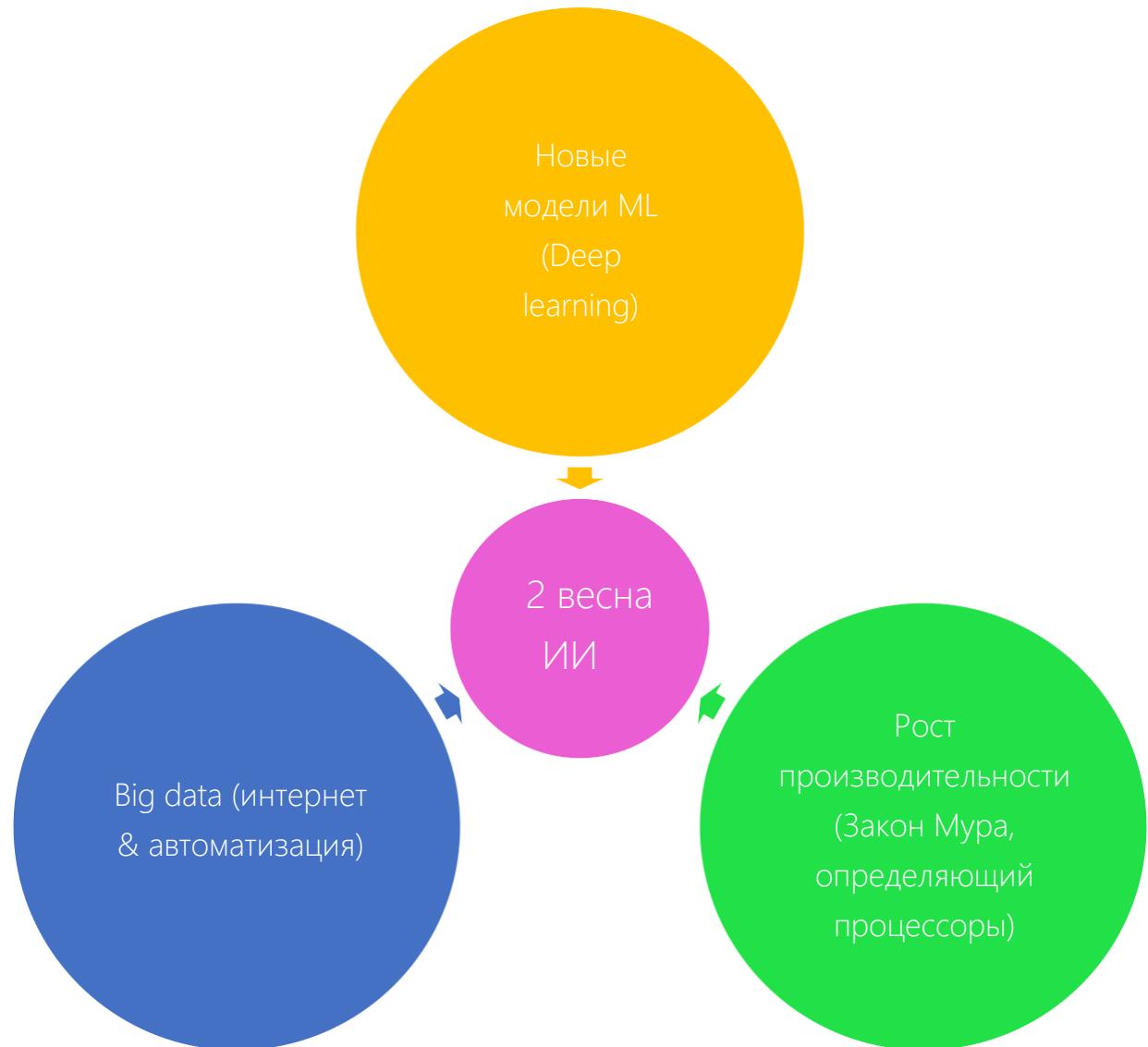




"Я вижу, что в течение следующих пяти лет ИИ заменит людей в области..."



Зима и весна ИИ



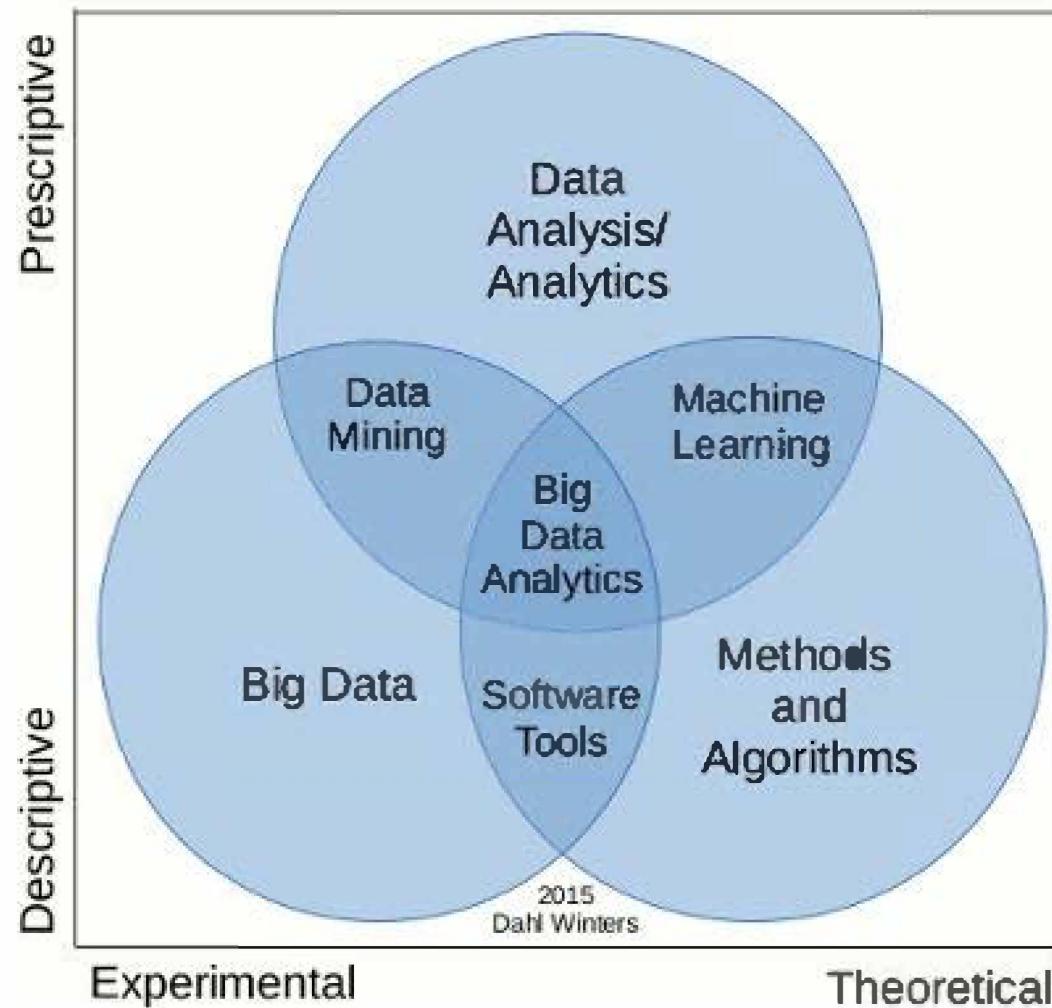
Определение

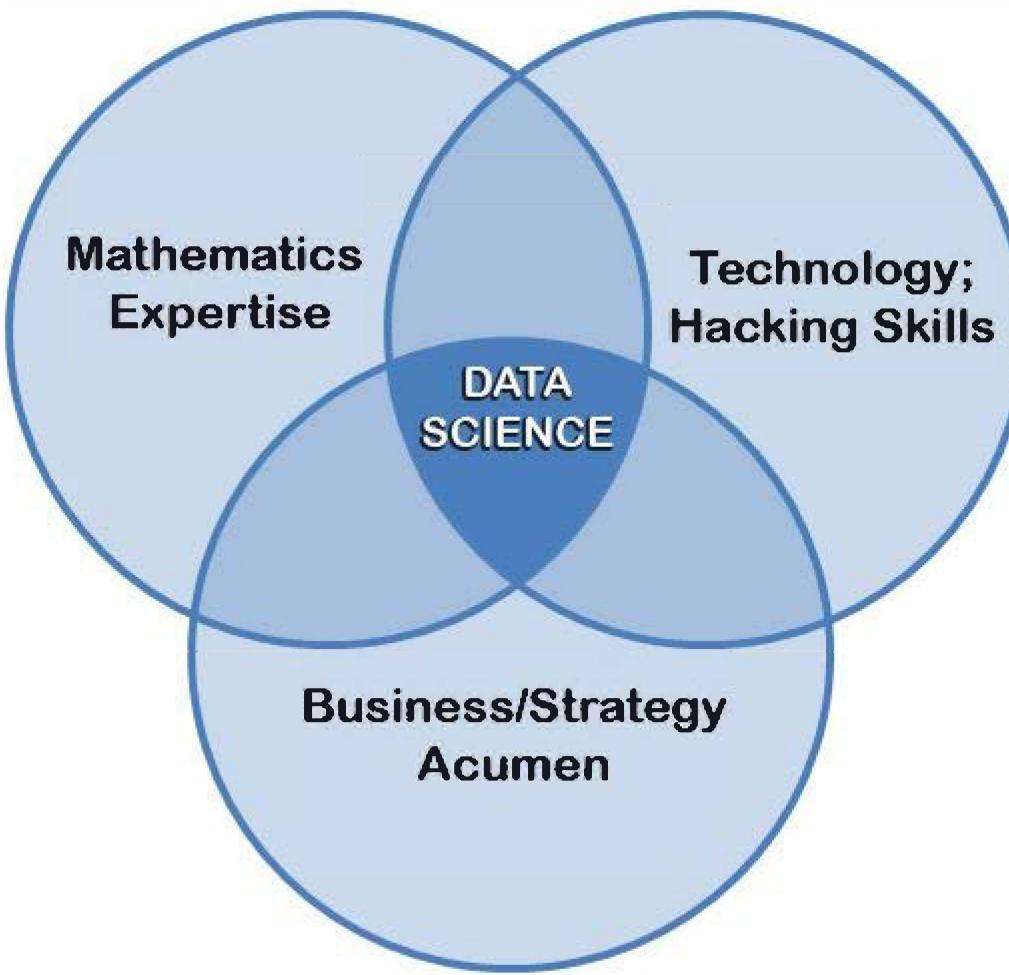
Machine Learning:

Machine learning это:

- Раздел искусственного интеллекта, математическая дисциплина, которая использует математическую статистику, численные методы оптимизации, теорию вероятности, дискретный анализ для извлечения сведений из данных.
- Область наук, которая предоставляет компьютерам возможность обучаться без предварительного явного программирования.

The Fields of Data Science





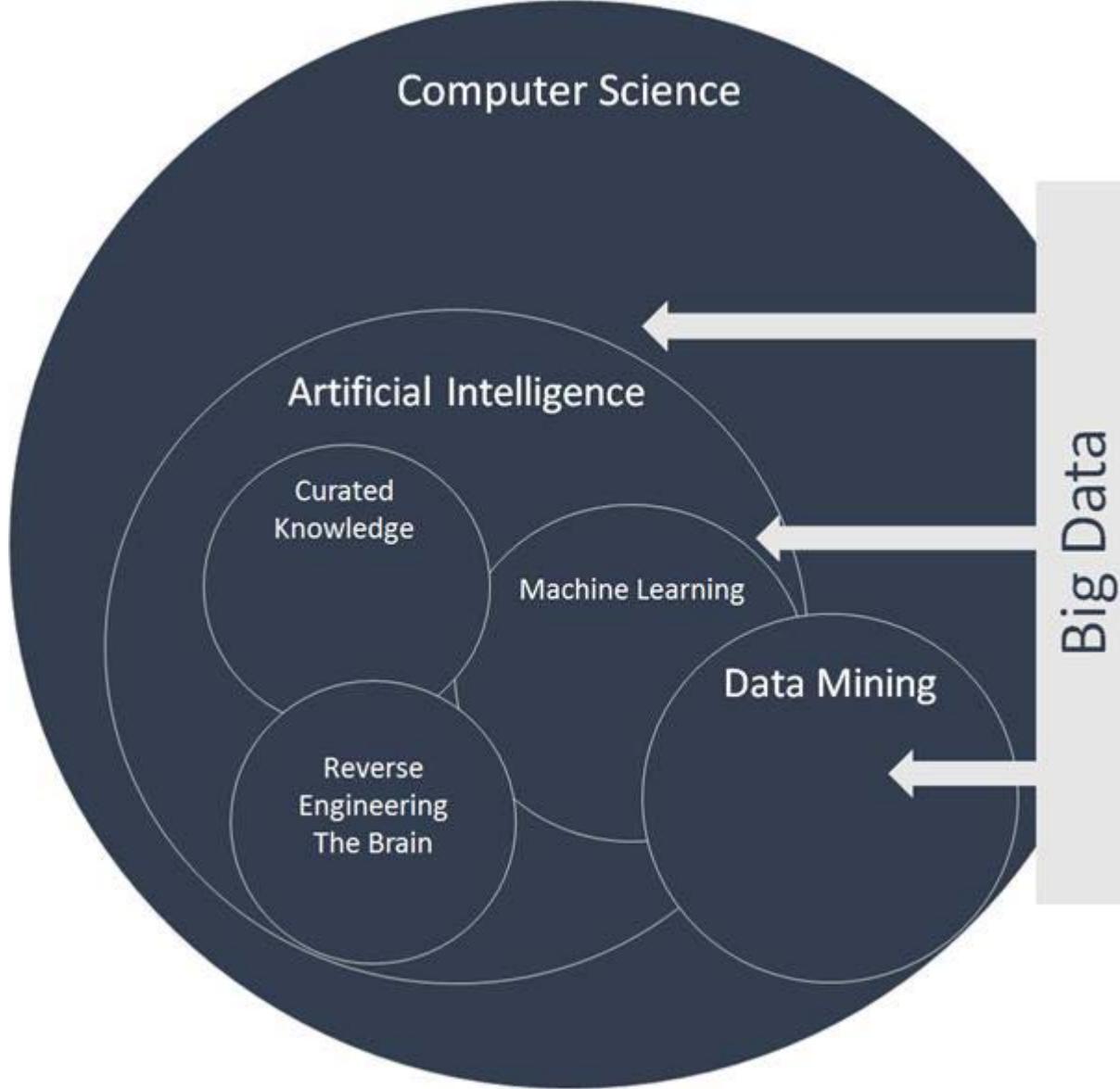
Ключевые особенности извлечения данных

- Данные могут быть неточными, неполными, гетерогенными, и в то же время иметь большие объемы;
- Алгоритмы анализа данных сами по себе могут иметь возможность обучаться через прецеденты;
- Процессы обработки "сырых" данных требуют нетривиальной автоматизации.

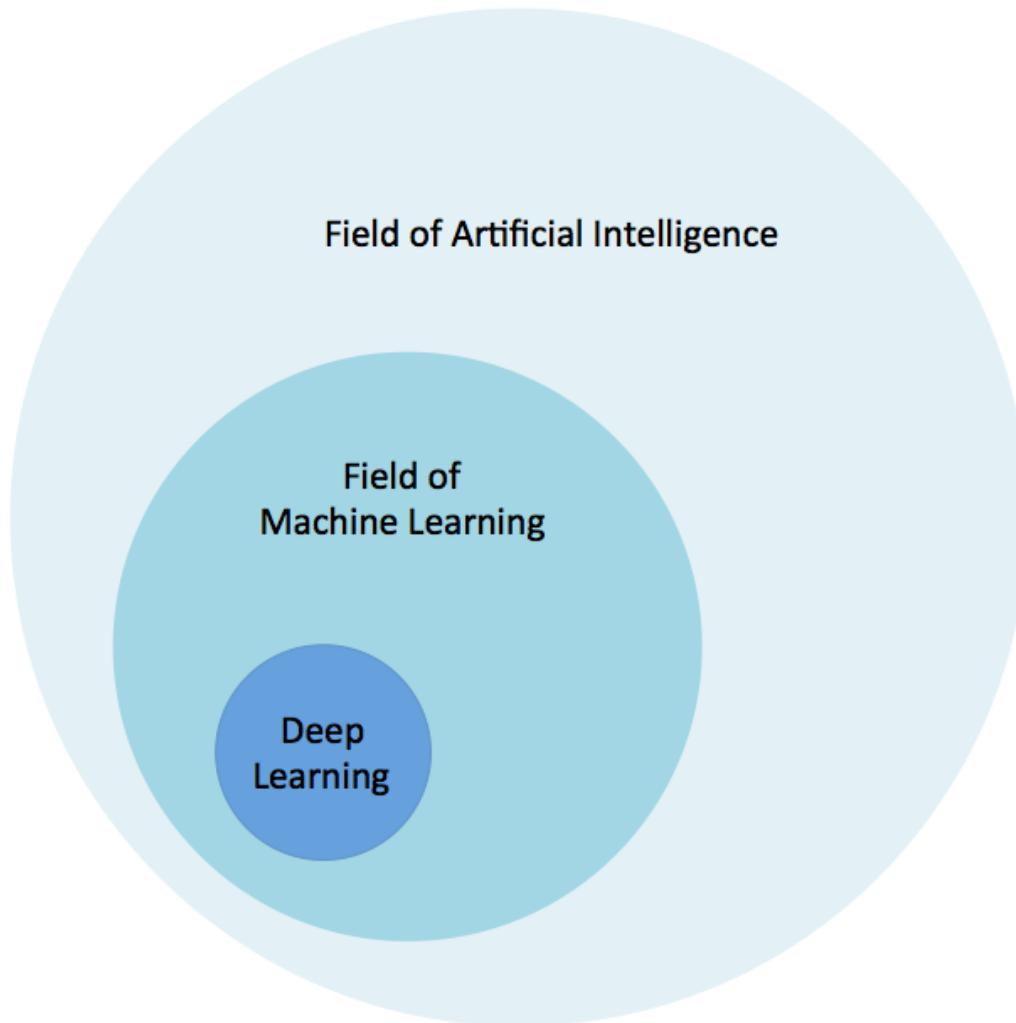
Ключевые особенности извлечения данных

В последние десятилетия значительные попытки в области извлечения данных сфокусировались на создании специальных алгоритмов, способных к выполнению схожих задач за линейное или даже логарифмическое время без серьезных потерь в точности.

Ключевые особенности извлечения данных



Ключевые особенности извлечения данных



Ключевые особенности извлечения данных

Компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E.

(Т.Митчел)

Ключевые особенности извлечения данных

Выделяют два типа обучения:

- Обучение по прецедентам
- Дедуктивное обучение

Ключевые особенности извлечения данных

Цель обучения по прецедентам заключается в обобщении или получении знания о "законе природы"

Ключевые особенности извлечения данных

Aa

ABCDEFGHIJKLMNOPQRSTUVWXYZ 01234
abcdefghijklmnopqrstuvwxyz 56789

Ключевые особенности извлечения данных

SPHINX OF BLACK QUARTZ JUDGE MY VOW

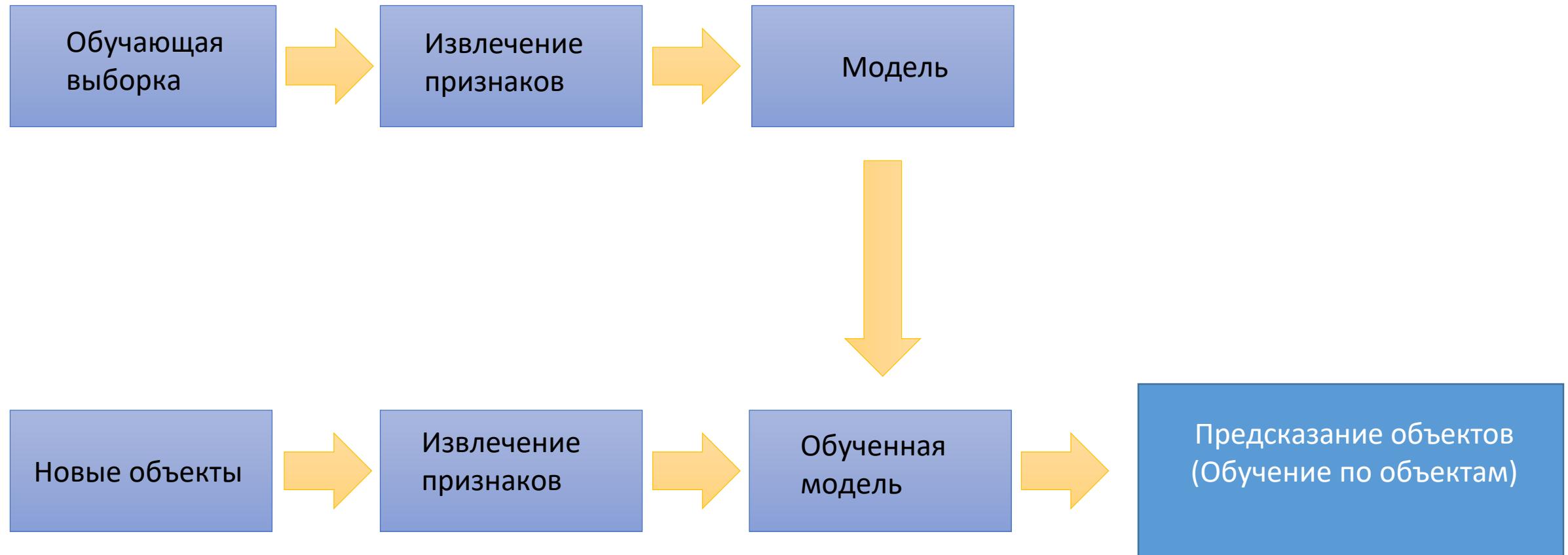
Sphinx of black quartz judge my vow

A B C D E F G H a b c d e f g h

0 1 2 3 4 5 6 7 8 9 / @ # \$ %

• глаза и оправа лучше. Но
ищем у глазок замысл
и они имеют особенности
личных традиций и вакансий
быть обновлена и купла
голосов. Помимо обычном
но законы природы и

Ключевые особенности извлечения данных



Определение задачи ML

Machine learning (машинное обучение):

- Что есть объект X (каковы признаки) и что есть ответ Y ?
- Как построить модель M ?
- Каким образом мы можем выполнить аппроксимацию (приближение) Y от X с M ?

Определение задачи ML

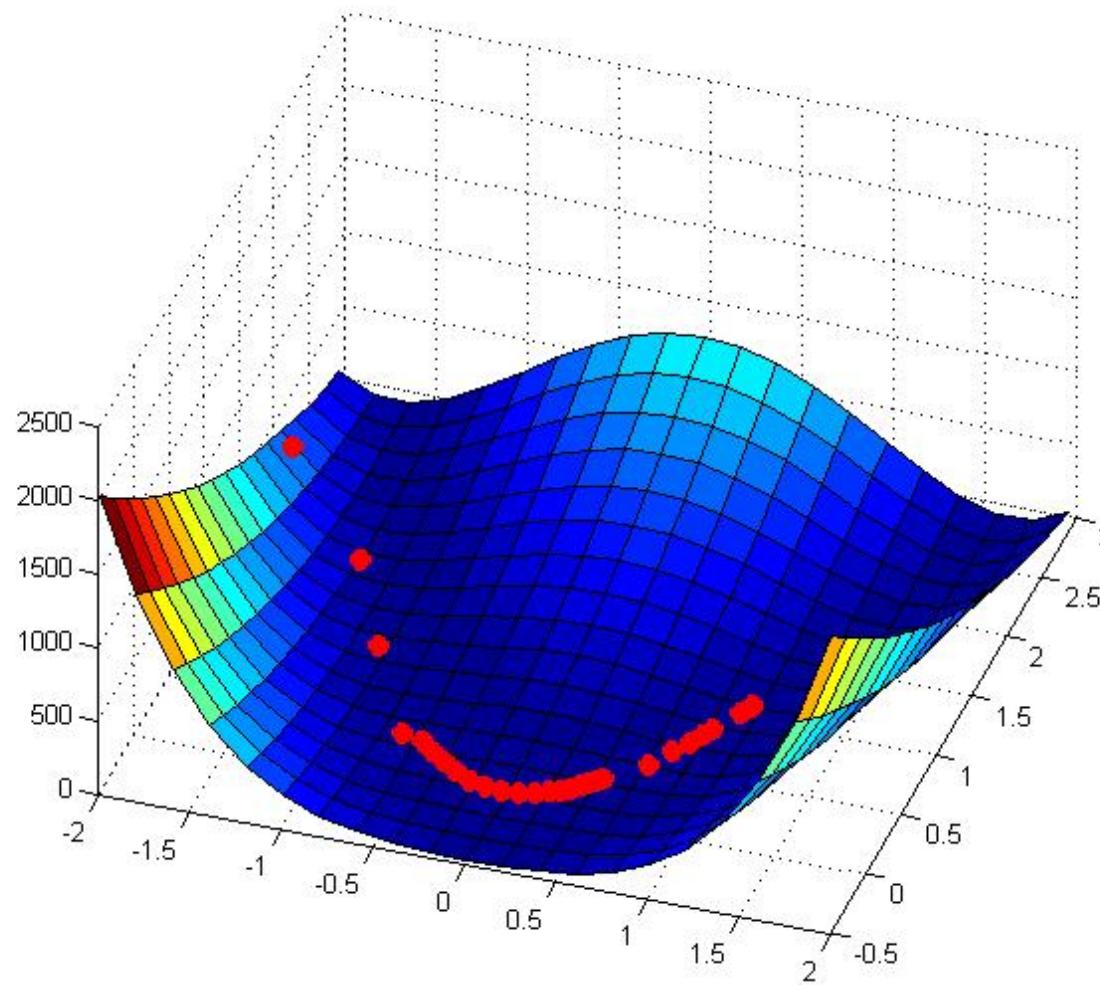
Типы входных данных:

- Изображение
- Текст
- Звук
- Геоданные
- Временные ряды
- Техническая диагностика, спецификации

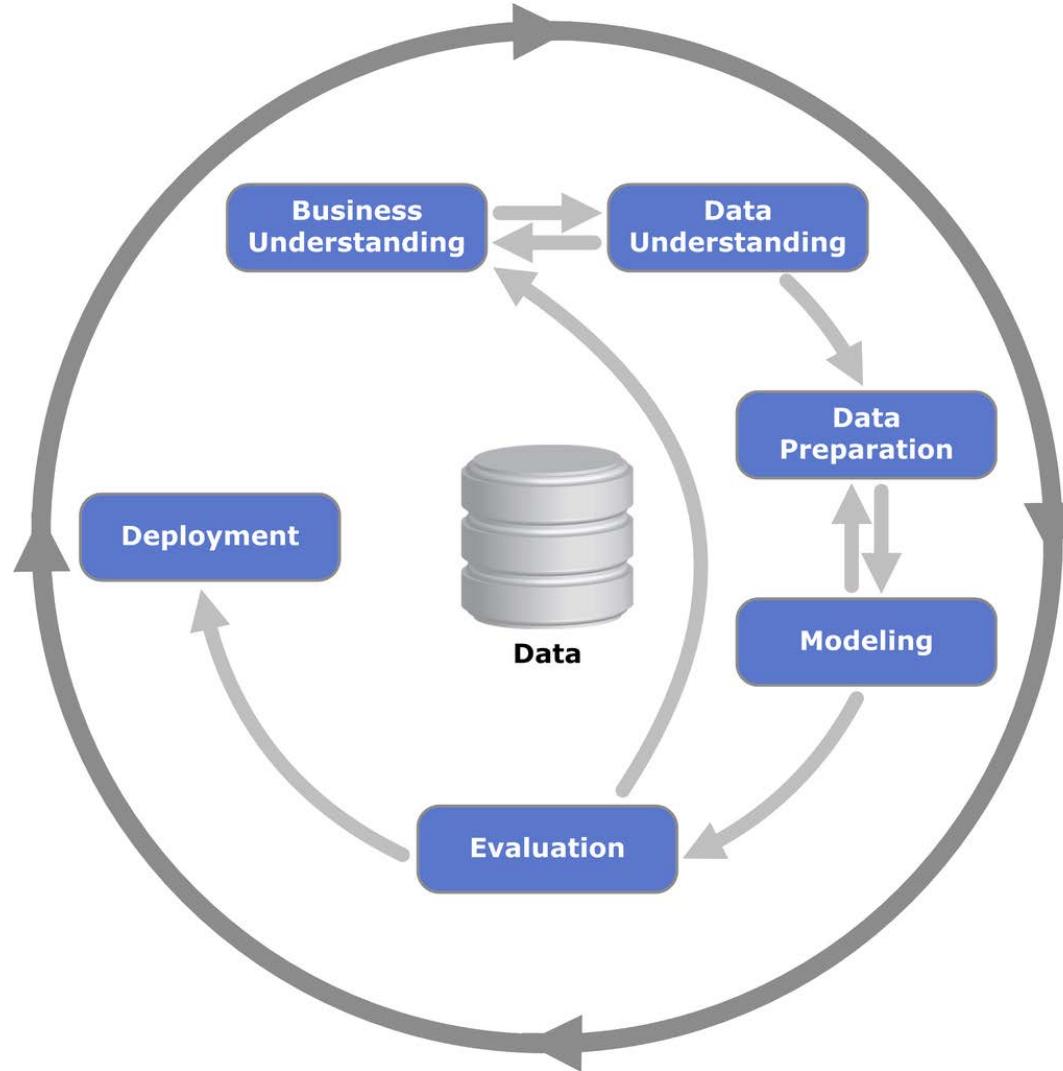
Обобщение

Обобщающая способность - это способность аналитической модели, построенной на основе машинного обучения (нейронной сети, дерева решений, карты Кохонена и др.) выдавать правильные результаты не только для примеров, участвовавших в процессе обучения, но и для любых новых, которые не участвовали в нем. Обобщающая способность является важнейшим свойством аналитической модели, приобретаемым в процессе обучения.

Метод градиентного спуска

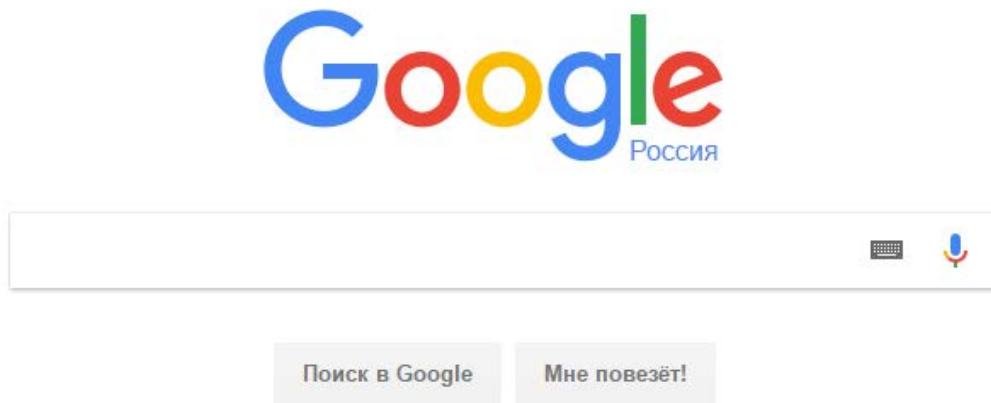


Методология анализа данных CRISP-DM

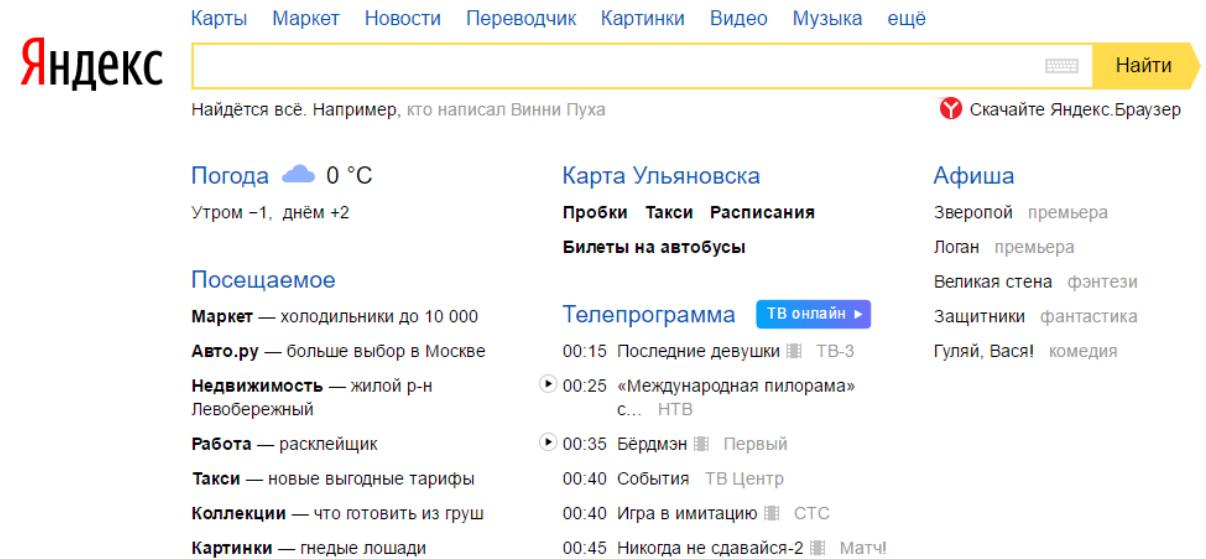


Сфера применения ML

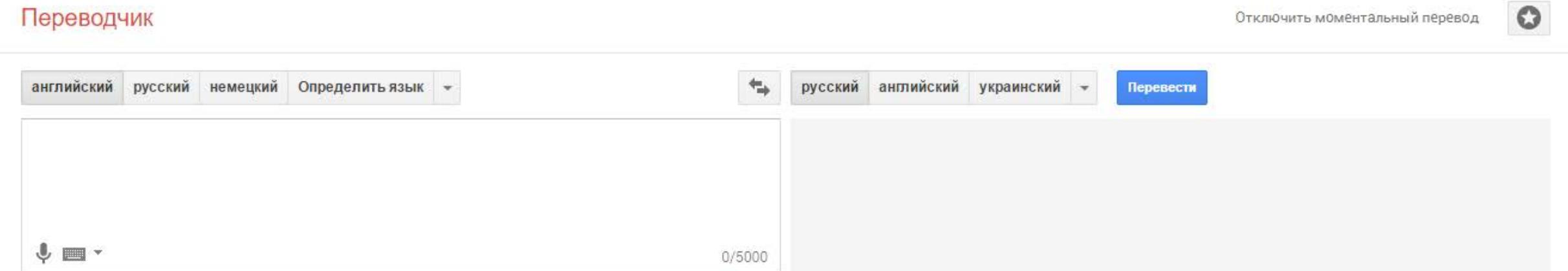
Автоматическое управление	Биоинформатика	Биржевой технический анализ	Генерация изображений	Генерация речи
Генерация текстов	Категоризация документов	Кредитный скоринг	Медицинская диагностика	Обнаружение мошенничества
Обнаружение спама	Обучение ранжированию в информационном поиске	Подбор потенциальных клиентов	Предсказание ухода клиентов	Принятие решений
Прогнозирование временных рядов	Распознавание жестов	Распознавание образов	Распознавание речи	Распознавание рукописного ввода
Распознавание физической активности	Техническая диагностика	Финансовый надзор	Хемоинформатика	...



The Google Russia search interface features the classic multi-colored 'Google' logo at the top left. Below it is the word 'Россия'. A search bar is centered, flanked by a keyboard icon and a microphone icon. At the bottom are two buttons: 'Поиск в Google' (Search in Google) and 'Мне повезёт!' (I'm feeling lucky!).

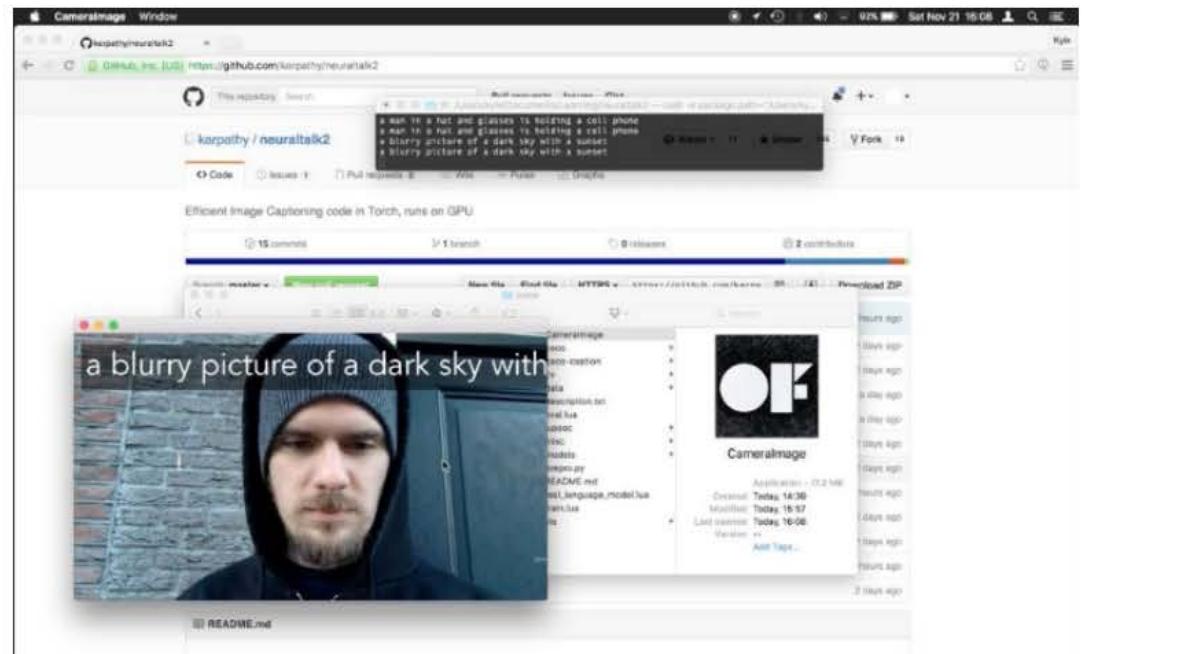


The Yandex search interface features the red 'Яндекс' logo at the top left. A search bar is centered with the placeholder text 'Найдётся всё. Например, кто написал Винни Пуха'. To the right are links for 'Карты', 'Маркет', 'Новости', 'Переводчик', 'Картинки', 'Видео', 'Музыка', and 'ещё'. Below the search bar is a weather widget showing 'Погода 0 °C' and 'Утром -1, днём +2'. To the right are sections for 'Карта Ульяновска', 'Пробки', 'Такси', 'Расписания', and 'Билеты на автобусы'. A 'Телепрограмма' section lists TV shows with their times and channels. On the far right, there's a 'Афиша' (Events) sidebar with movie and event listings.



The Yandex Translate interface is shown at the bottom. It includes language selection dropdowns for 'русский', 'английский', and 'украинский', with 'Перевести' (Translate) as the main button. Below these are two large input fields for text entry, each with a microphone and keyboard icon. The character count '0/5000' is displayed between them.

NeuralTalk



A man in a blue shirt and red shorts is playing tennis on a court. He is in the middle of a serve, with his right arm extended upwards holding a tennis racket. A ball is visible near his hand. In the background, there is a blue wall with the words "BOSTON" and "OPEN" partially visible. A woman in a green shirt and white pants stands on the right side of the court, possibly a referee or ball girl. The court has a green surface and a white baseline.

A photograph of a double-decker train with a yellow front and blue sides, traveling along tracks. In the background, there is a modern building with a distinctive red and white patterned facade.

A photograph of a slice of red velvet cake on a blue plate. The cake has a dark red base layer, a white cream cheese frosting layer, and a top layer with red strawberries.

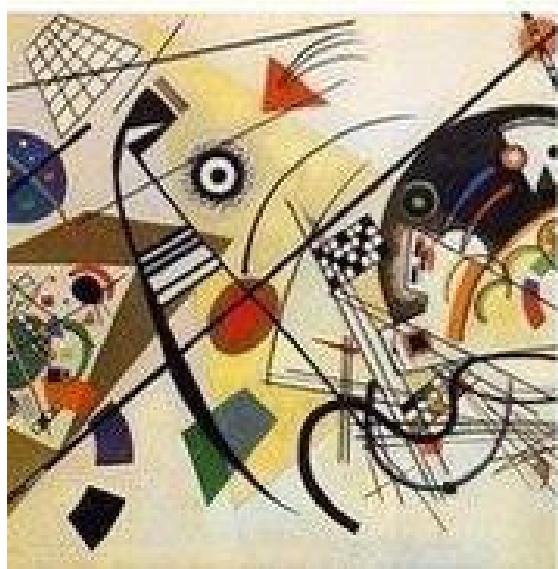
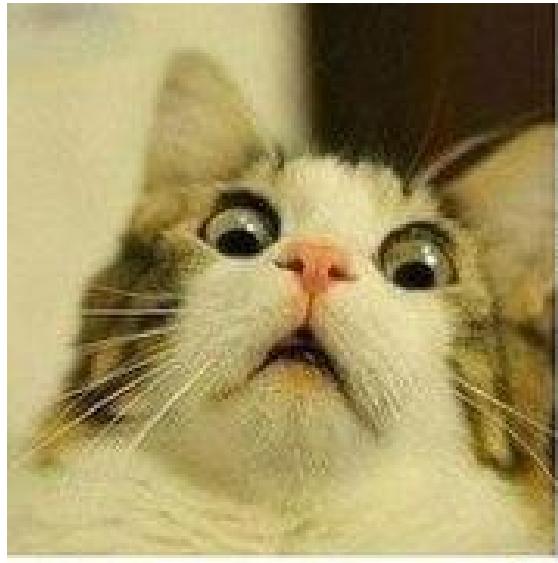
A photograph showing a long, green wooden park bench with white horizontal stripes. The bench is positioned on a small, rectangular patch of green grass. To its left is a paved sidewalk, and to its right is a dark, paved area. A few fallen leaves are scattered on the grass near the bench.

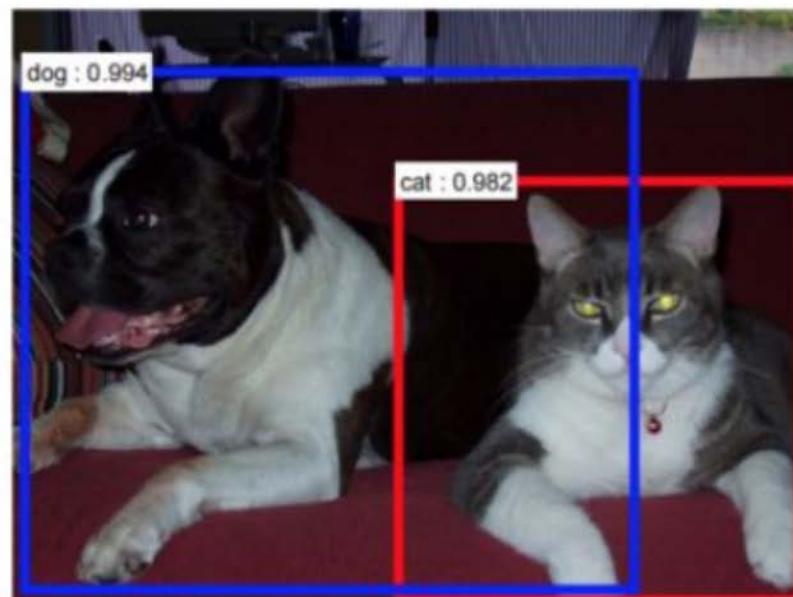
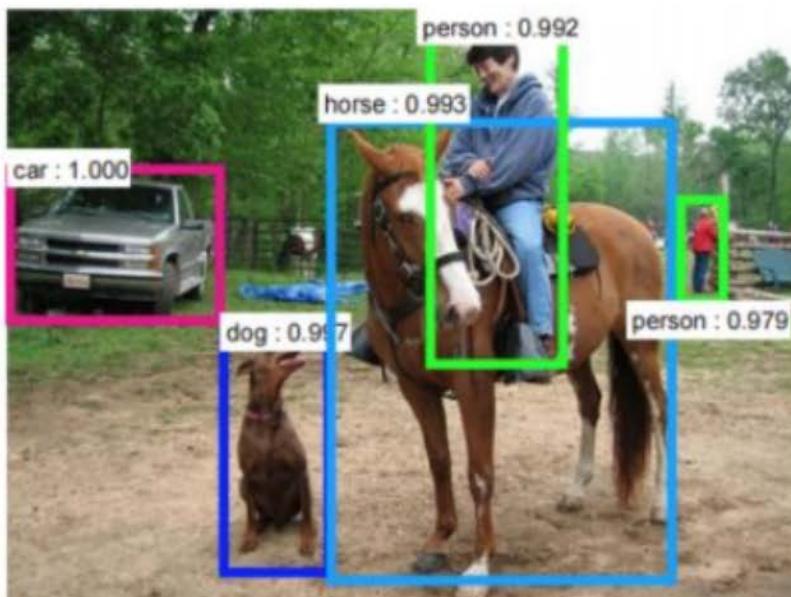
<https://github.com/karpathy/neuraltalk2>

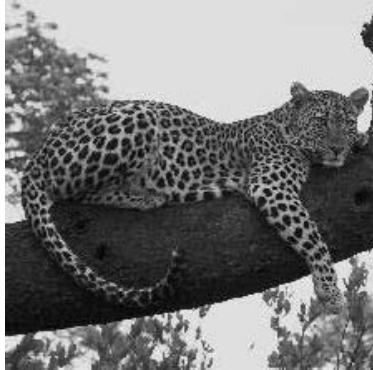
<https://vimeo.com/14649200>













"girl in pink dress is jumping **in** air."



"black and white dog **jumps** over bar."



"young girl **in** pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



"little girl is eating piece of cake."



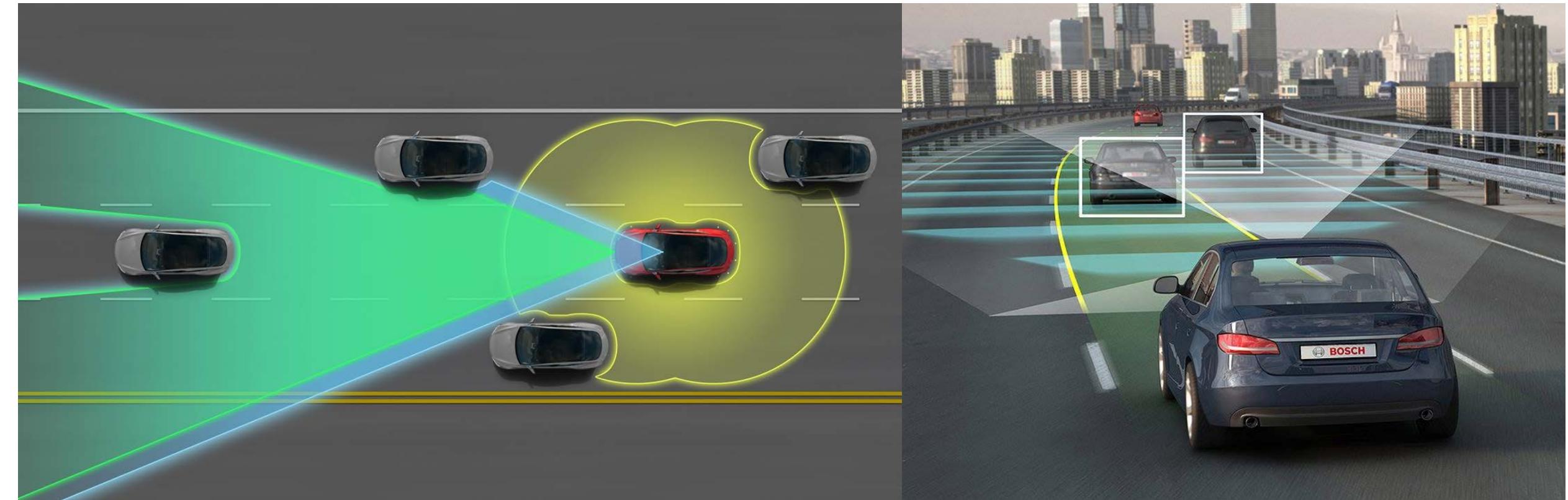
"baseball player is throwing ball in game."



"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."



Artificial Intelligence

855 companies

Machine Learning-Gen
(96 Companies)



Machine Learning-App
(201 Companies)



Computer Vision-Gen
(97 Companies)



Computer Vision-App
(73 Companies)



Smart Robots
(52 Companies)



Virtual Personal Assistants
(71 Companies)



NLP-Speech Recog.
(65 Companies)



NLP-Gen
(127 Companies)



Speech to Speech Trans.
(15 Companies)



Context Aware Comp.
(28 Companies)



Gesture Control
(30 Companies)



Recommendation Eng.
(54 Companies)



Video ACR
(14 Companies)



Типы задач ML

1. Обучение с "учителем":
 - Классификация
 - Регрессия
 - Ранжирование
 - Прогнозирование
2. Обучение без "учителя":
 - Кластеризация
 - Обучение на ассоциативных правилах (Association rules learning)
 - Обнаружение выбросов (outliers detection)
3. Обучение с подкреплением
4. Метаобучение (Meta-learning or learning-to-learn)
5. Формирование инвестиционного портфеля (Portfolio selection)
6. Коллаборативная фильтрация
7. ...

Подходы и методы

- Регрессия
- Байесовский вывод
- Деревья решений
- Нейронные сети
- Метод К-ближайших соседей
- Метод главных компонент (Principal component analysis)
- Метод опорных векторов (Support vector machines)
- Генетические алгоритмы

Инструменты

1. Python, R, Matlab, ...
2. IPython Notebook, PyCharm
3. Tensor Flow, Keras, Theano, Scikit-Learn
4. CNTK
5. Torch
6. Coffe
7. Apach Spark
8. Azure

Ссылки

- [ML course by Andrew Ng](<https://www.coursera.org/learn/machine-learning>)
- [ML course by Dmitry Efimov](https://github.com/diefimov/MTH594_MachineLearning)
- [ML course by OpenData Science](<https://github.com/Yorko/mlcourse.ai>)
- [MIT Deep learning](<https://github.com/lexfridman/mit-deep-learning>)
- <https://github.com/qati/DeepLearningCourse>
- https://github.com/roebius/deeplearning_keras2
- <https://github.com/enggen/Deep-Learning-Coursera>
- <https://github.com/fchollet/deep-learning-with-python-notebooks>