Алгоритмы обработки больших данных. Big Data

Лектор к.т.н. Лобанов А.А. доцент кафедры ИППО

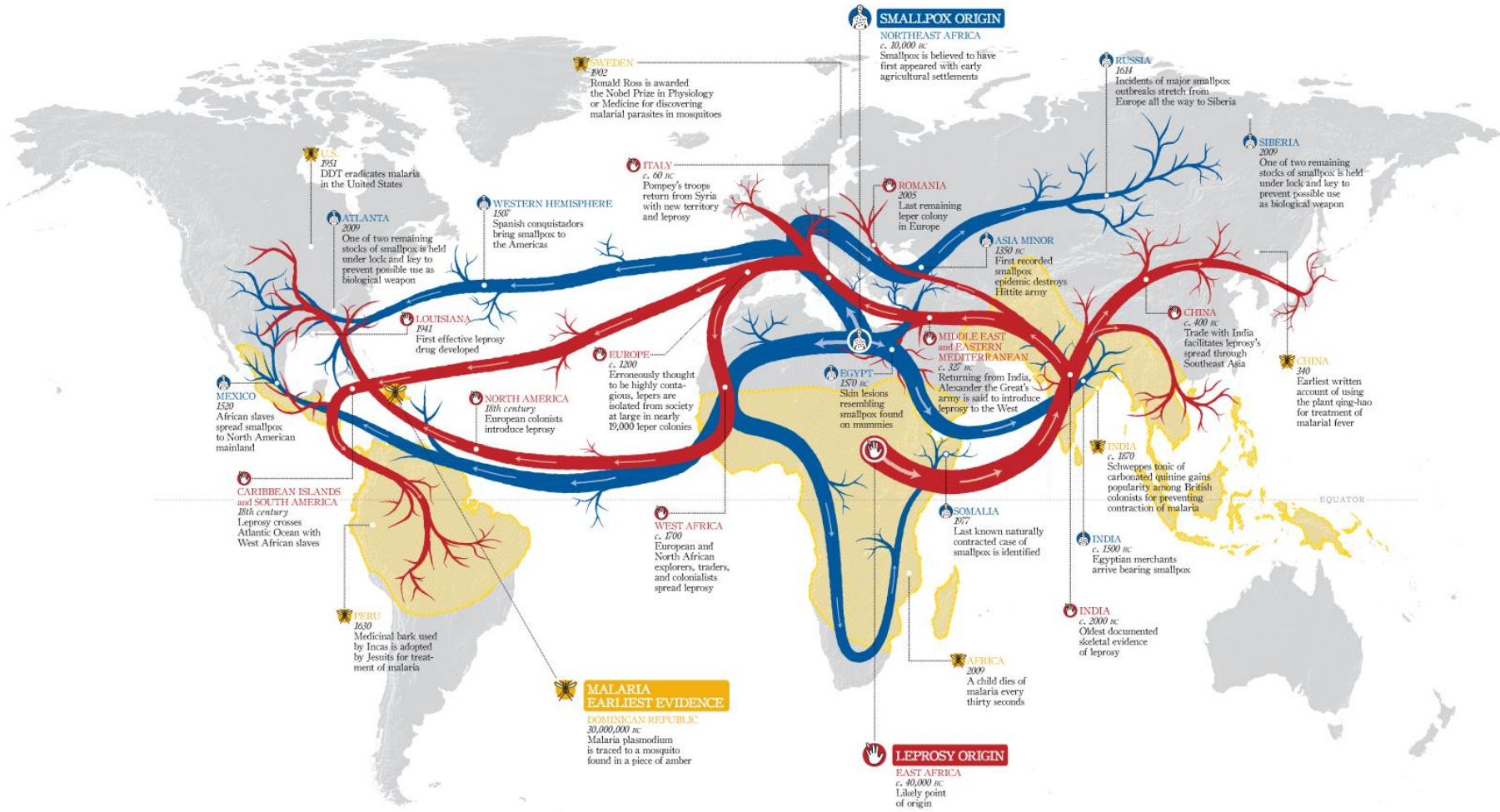
Что такое Big Data?

- Строго говоря, единого определения нет, но мы попробуем:
- Совокупность методов и алгоритмов обработки структурированных и неструктурированных данных с целью получения ценной информации.
- Как правило используется в прогнозировании

Несколько фактов для начала

Сеть Target и беременная девушка

Эпидемия гриппа и Google



Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим

Год: 2013

Автор: Viktor Mayer-Schönberger, Kenneth Cukie / Виктор Майер-Шенбергер, Кеннет Кукьер

Переводчик: Инна Гайдюк

Издательство: Манн, Иванов и Фербер

ISBN: 978-5-91657-936-9

Области применения

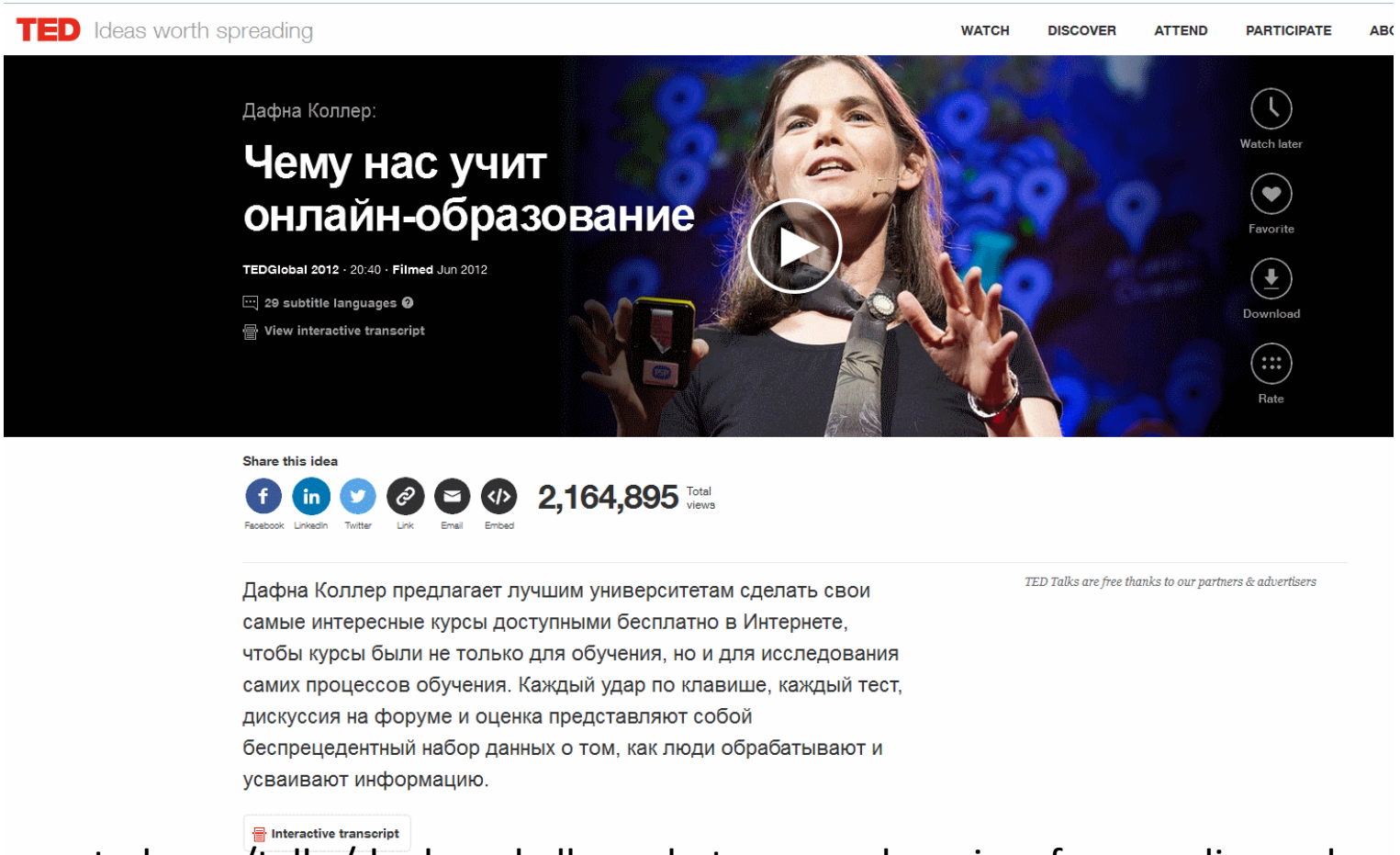
- Военные
- Прогноз погоды
- Бизнес
- Ученые
- Образование???

Как Big Data помогают образованию?

- Заочное образование
- Дистанционное образование
- E-learning
- Интеллектуальное образование

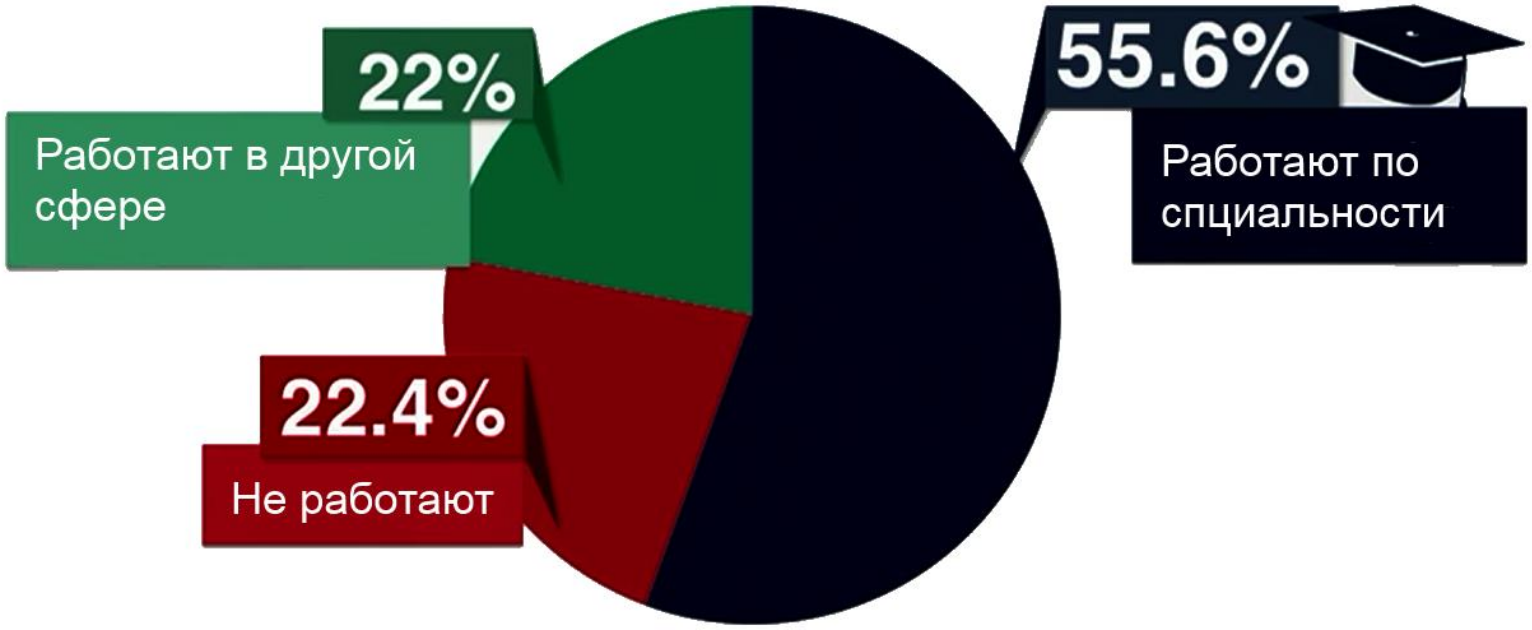
Интеллектуальное образование

Что можно извлечь из онлайн-образования



http://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education/transcript

Занятость выпускников в США



Source: "A College Degree, but Not a College Job." (A. Sum, *New York Times*, 19 May 2011)

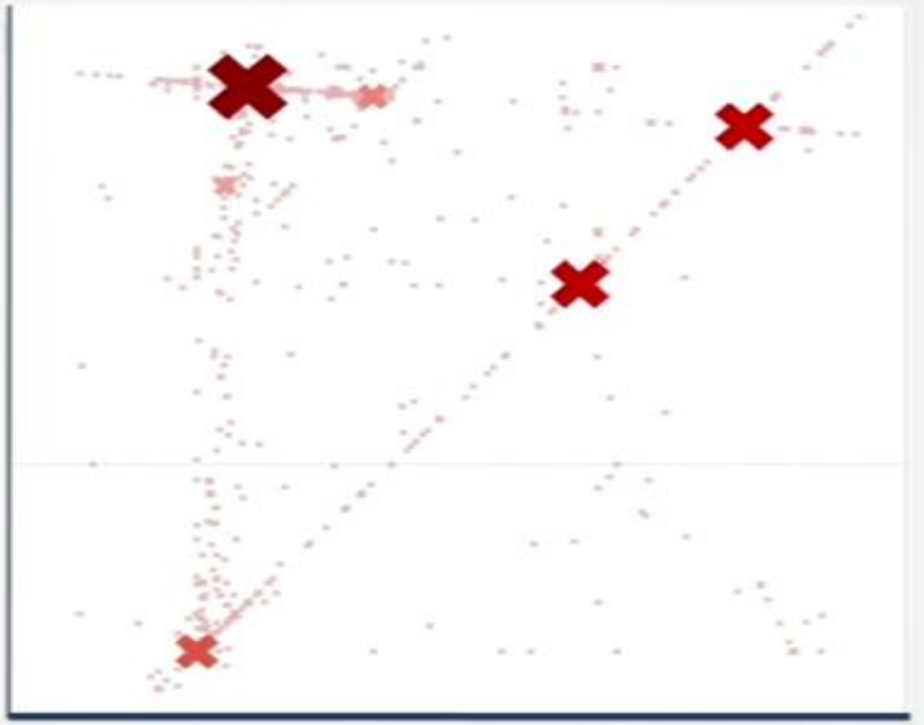
Востребованная форма обучения

- 100 000 студентов на одном предмете
- 190 стран
- 1 640 000 студентов за 2 года
- 6 000 000 контрольных работ

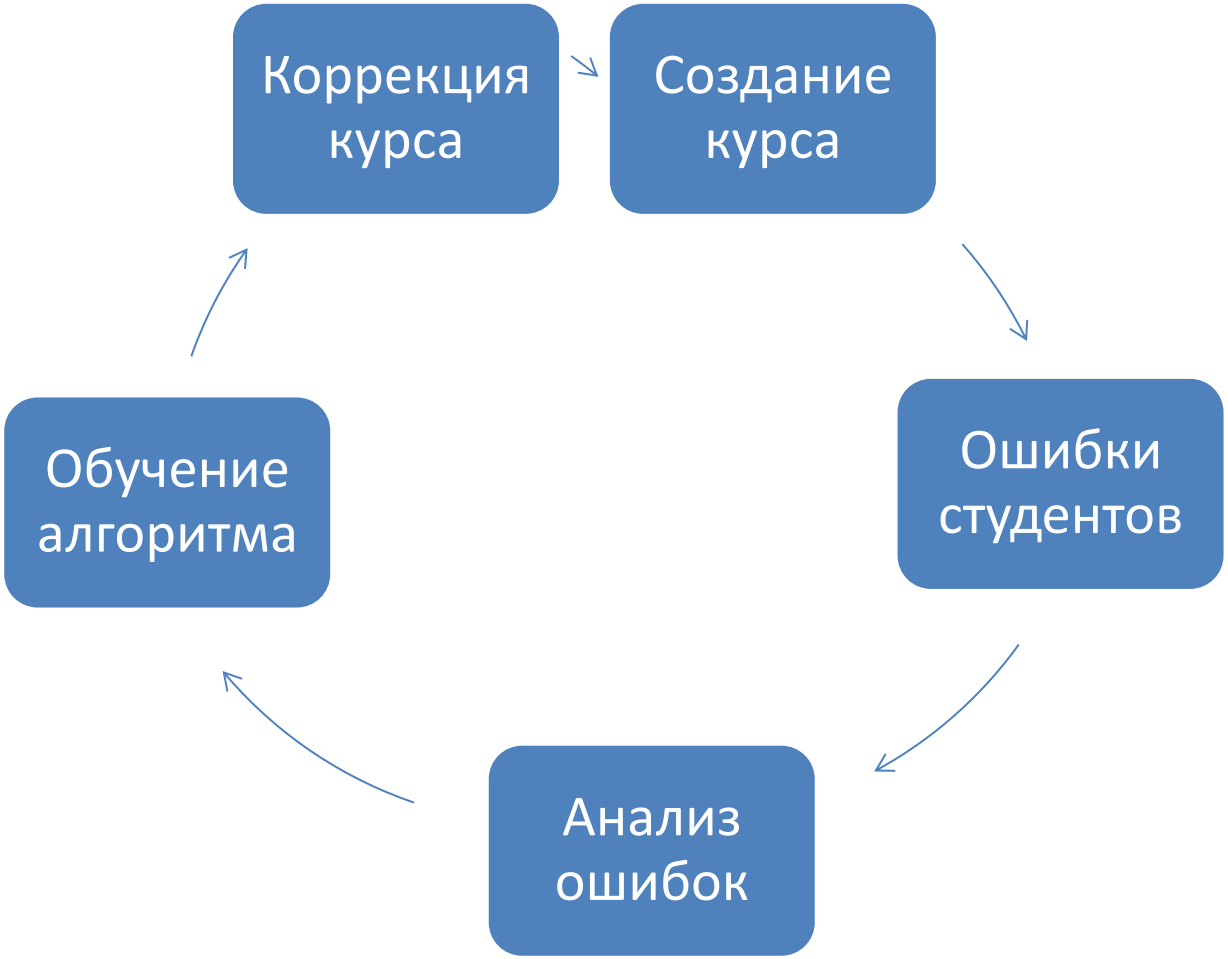
- *По данным Coursera.org

Особенности обучения большого числа студентов

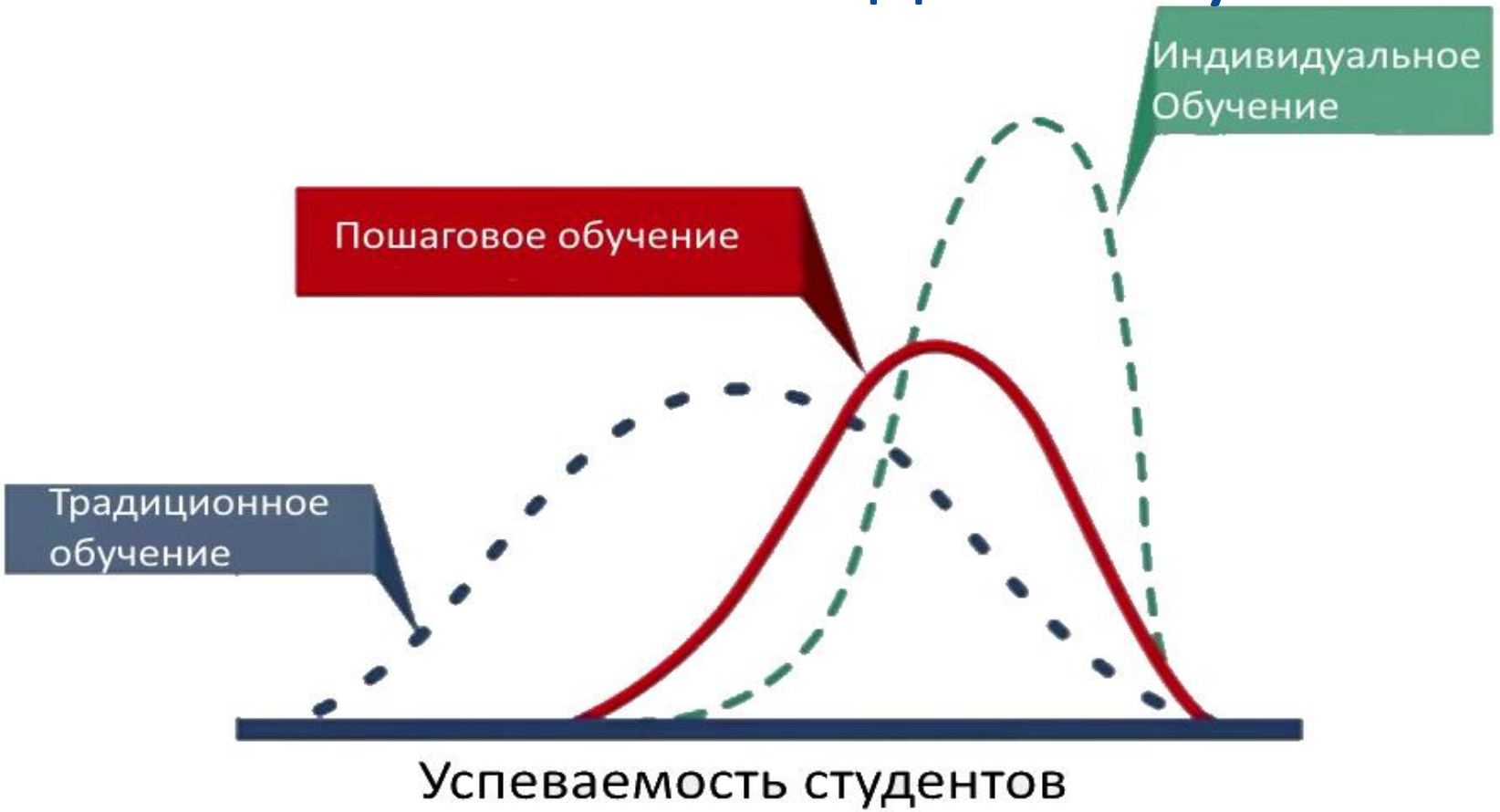
неправильные ответы студентов и их корреляция



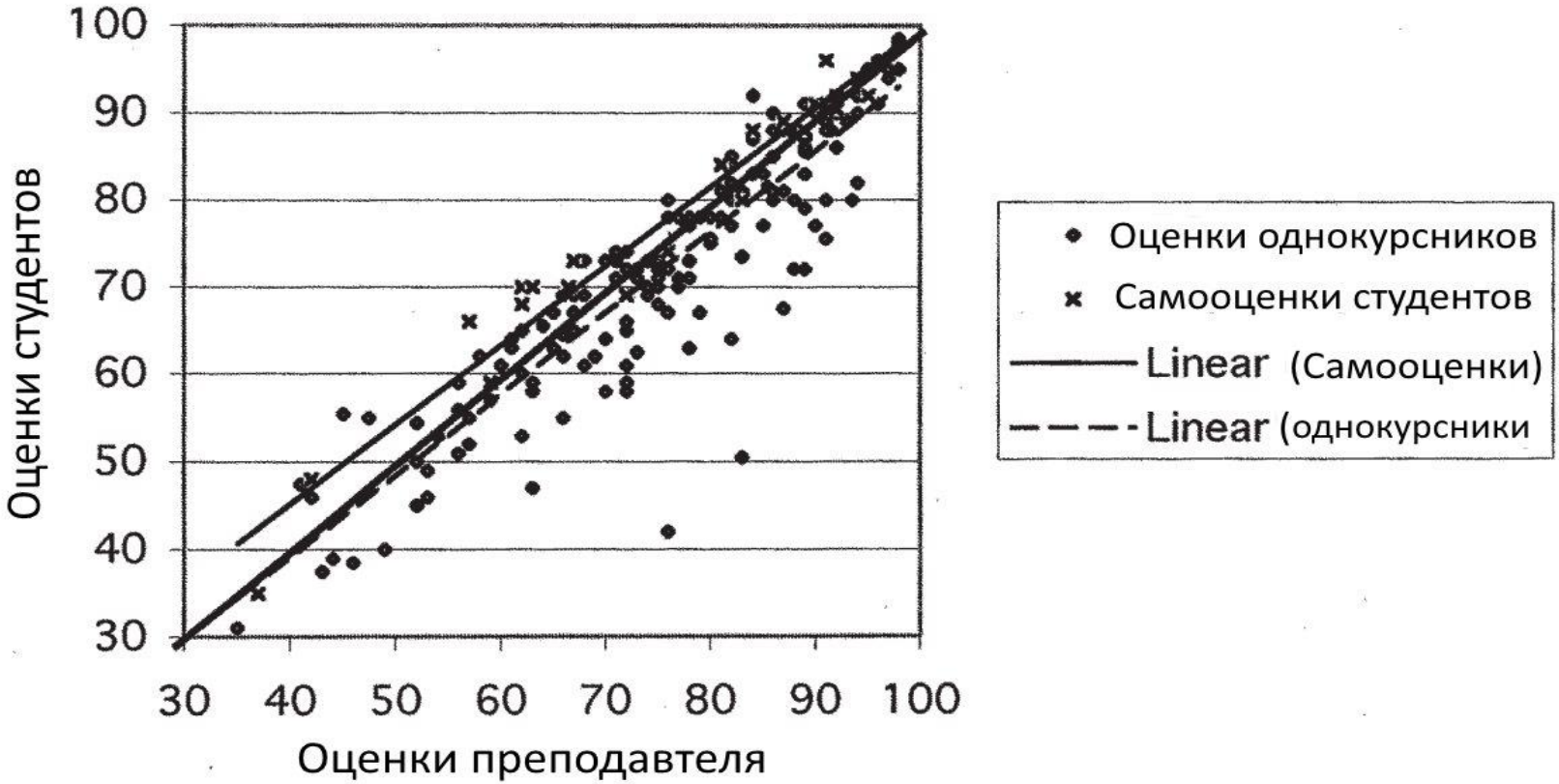
Интеллектуальная образовательная система



Успеваемость студентов в зависимости от методики обучения



Самооценка контрольных работ гуманитарного цикла



Что же такое машинное обучение?

- Это комплекс математических методов и алгоритмов, направленных на поиск скрытых закономерностей в данных.
- **Данные** – совокупность объектов (характеризующихся набором численных параметров – вектором переменнных, которые делятся на 2 группы: НАБЛЮДАЕМЫЕ и СКРЫТЫЕ (X и T))

Термины и определения

- Явные параметры – те, которые могут быть легко оценены численно.
- Скрытые или латентные параметры – те, которые трудно или невозможно оценить.
- Обучающая выборка – совокупность объектов, для которых известны как явные, так и скрытые параметры.

Задача машинного обучения

- Параметризовать явные и скрытые параметры, и с точностью до некоторого вектора весов W
- Находя параметры W можно восстановить (спрогнозировать) скрытые параметры
- Параметр W требуется отыскать на обучающей выборке.

Одна из простейших и довольно хорошо изученных задач машинного обучения это «деление на две группы», в которой:

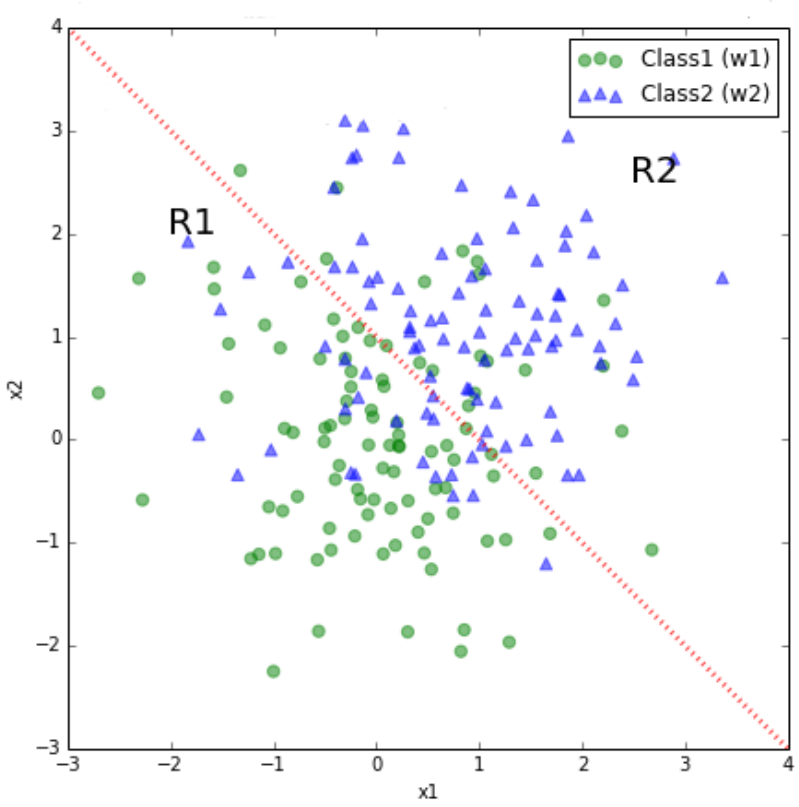


Рис. 1 классификация объектов на две группы

В качестве явных параметров выступают двумерные координаты объектов:

$$X = \{x_i\}_{i=1}^{n}, x_i \in R^2$$

скрытые параметры могут принимать значения только два +1 и -1

$$T = \{t_i\}_{i=1}^{n}, t_i \in \{-1,1\}$$

деление на две группы с использованием гиперплоскости

$$t(x) = \text{sign}(W^T x) + \omega_0$$

Современные задачи машинного обучения

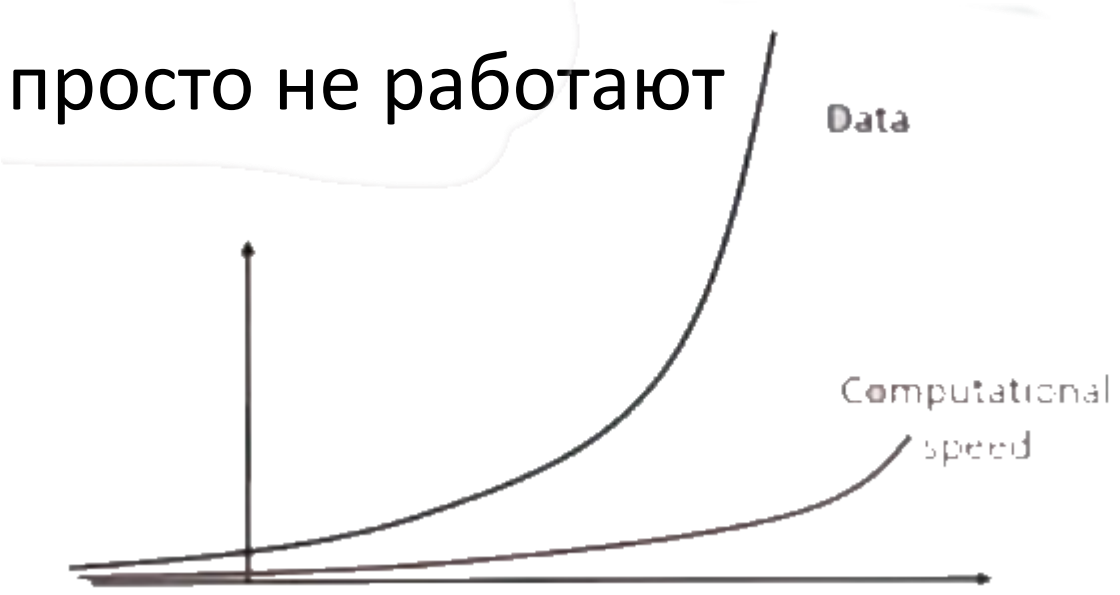
- Компьютерное зрение
- Кредитный скоринг
- Социологические исследования
- Распознавание (речи, образов, жестов …)
- Имитационное моделирование
- Диагностика (техническая и медицинская)
- Обнаружение (мошенничества и спама)

История развития методов

- 1990-ые Модель опорных векторов
- 1990-2000-ые Байесовские сети
- 2000-ые Графические модели (в основе Байесовские сети)
- 2010-ые Глубинные нейронные сети
- 2010-ые Big Data
- 2020-40- Настоящий искусственный интеллект

Большие данные

- Объемы и скорость накопления данных значительно (на несколько порядков) опережает вычислительные способности
- Сложно просто сохрянят, а не то, что обрабатывать
- Старые методы просто не работают



Байесовские сети

- Все случайные величины имеют закономерность, которую мы не знаем.
- Преобразует незнание в понятиях распределения.
- Воспользуемся теоремой Байеса

$$\text{апост вер} - \text{сть A при B} = \frac{(\text{вер} - \text{сть события A при B}) \times \text{априорная вер} - \text{сть A}}{\text{полная вероятность B}}$$

Формула Байеса

Формула Байеса:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

где

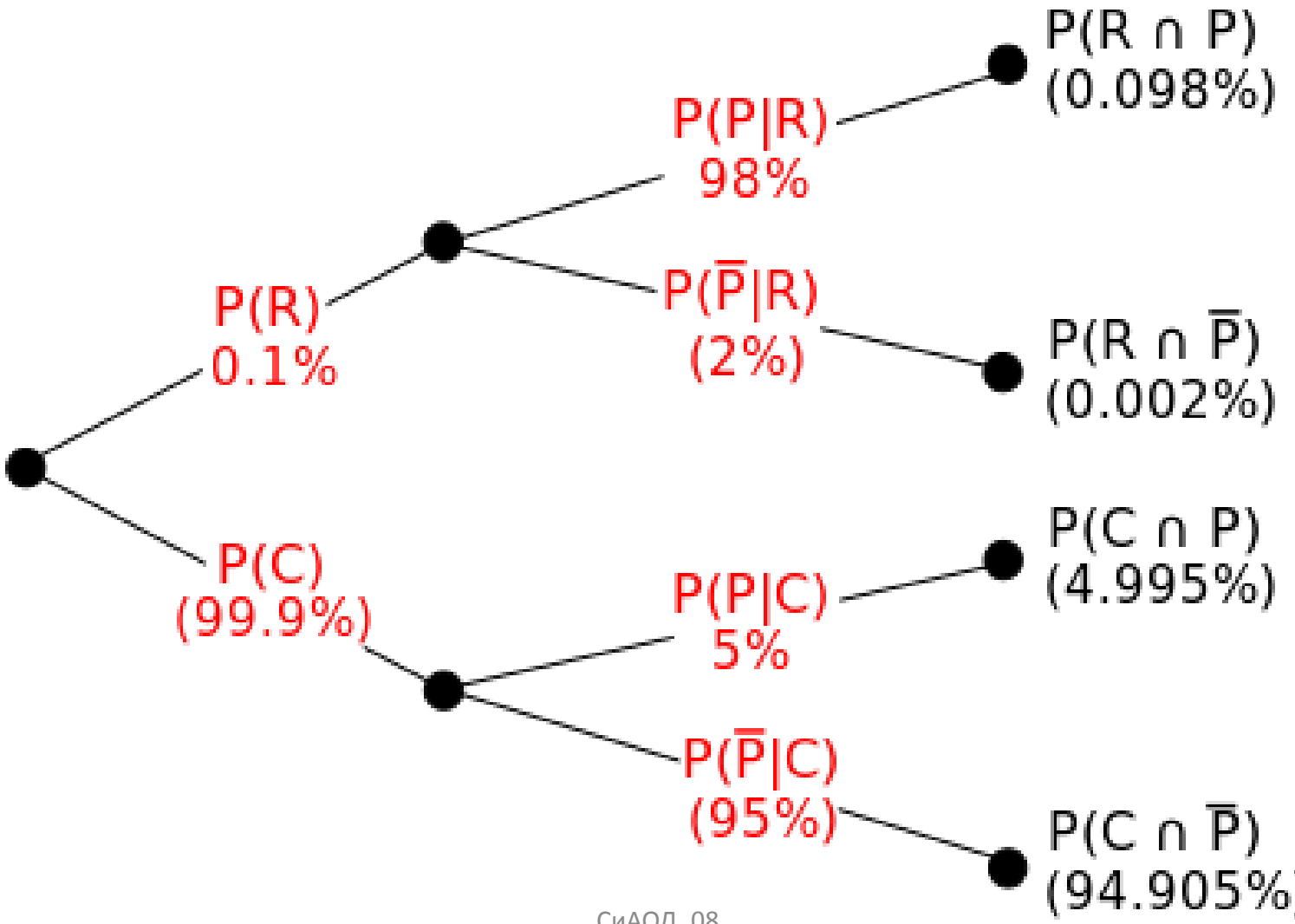
$P(A)$ — априорная вероятность гипотезы *A* (смысл такой терминологии см. ниже);

$P(A|B)$ — вероятность гипотезы *A* при наступлении события *B* (апостериорная вероятность);

$P(B|A)$ — вероятность наступления события *B* при истинности гипотезы *A*;

$P(B)$ — полная вероятность наступления события *B*.

Древовидная модель



Преимущество Байесовского вывода

- Рекурсия. Полученное апостериорного распределения в качестве априорного распределения в новом объекте исследования.
- Можно объединять несколько моделей апостериорного распределения
- Мгновенная обработка больших объемов данных (обработал и забыл)
- Создание графических моделей

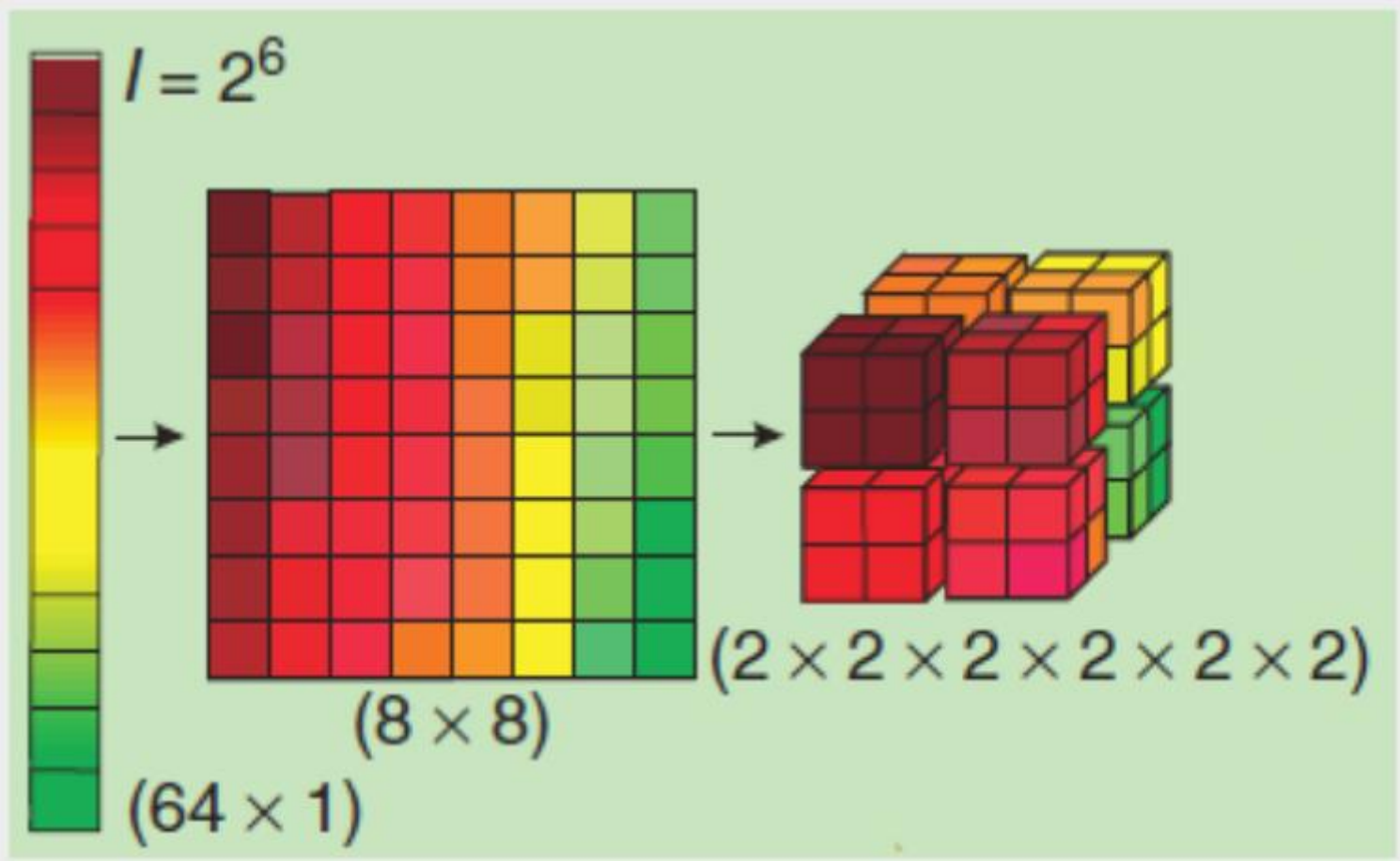
Граф Модели

- Сегментация изображений
- Анализ поведения и предпочтений
- Функциональная магнито-резонансная томография

Снова про рекомендации

	2			4	5	2.94*
	5		4			1
			5		2	2.48*
		1		5		4
			4			2
	4	5		1		1.12*

Разложение в тензорный поезд



Разложение в тензорный поезд

- Разложение в тензорный поезд имеет следующую форму:

произведение d 3-тензоров размером $n_i * r_{i-1} * r_i$

$$T(i_1, i_2, \ldots, i_d) = \sum_{\alpha_1, \ldots, \alpha_{d-1}} G_1(i_1, \alpha_1) G_2(\alpha_1, i_2, \alpha_2) \cdots G_{d-1}(\alpha_{d-2}, i_{d-1}, \alpha_{d-1}) G_d(\alpha_{d-1}, i_d)$$

Сокращение объемов

$$\mathbf{A}(x_1, x_2, x_3) = x_1 + x_2 + x_3,$$
$$x_1 \in \{1, 2, 3\},\ x_2 \in \{1, 2, 3, 4\},\ x_3 \in \{1, 2, 3, 4, 5\}.$$

$$\mathbf{A}(x_1, x_2, x_3) = G_1^{\mathbf{A}}[x_1] G_2^{\mathbf{A}}[x_2] G_3^{\mathbf{A}}[x_3],$$

$$G_1^{\mathbf{A}}[x_1] = \begin{bmatrix} x_1 & 1 \end{bmatrix} \quad G_2^{\mathbf{A}}[x_2] = \begin{bmatrix} 1 & 0 \\ x_2 & 1 \end{bmatrix} \quad G_3^{\mathbf{A}}[x_3] = \begin{bmatrix} 1 \\ x_3 \end{bmatrix}$$

$$G_1^{\mathbf{A}} = \left(\begin{bmatrix} 1 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 1 \end{bmatrix}, \begin{bmatrix} 3 & 1 \end{bmatrix} \right)$$
$$G_2^{\mathbf{A}} = \left(\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \right)$$
$$G_3^{\mathbf{A}} = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \end{bmatrix} \right)$$

Количество элементов в тензоре: $3 \cdot 4 \cdot 5 = 60$.
TT-формат требует хранения 32-х чисел.

Сокращение объемов

Опр. Тензор $\mathbf{A}$ представлен в *TT-формате*, если

$$\mathbf{A}(x_1, \ldots, x_n) = G_1^{\mathbf{A}}[x_1] G_2^{\mathbf{A}}[x_2] \ldots G_n^{\mathbf{A}}[x_n],$$

где $G_i^{\mathbf{A}}[x_i]$ — матрица размера $\mathrm{r}_{i-1}(\mathbf{A}) \times \mathrm{r}_i(\mathbf{A})$, $\mathrm{r}_0(\mathbf{A}) = \mathrm{r}_n(\mathbf{A}) = 1$.

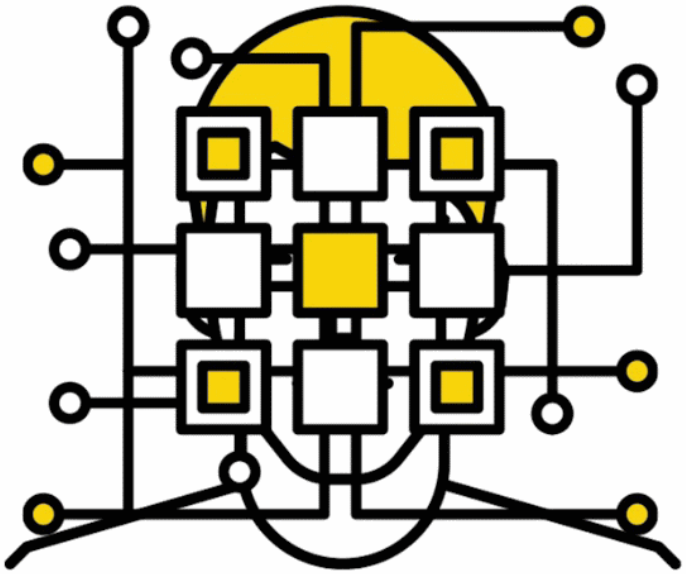
Терминология:

- $G_i^{\mathbf{A}}$ — *TT-ядра*;
- $\mathrm{r}_i(\mathbf{A})$ — *TT-ранги*;
- $r(\mathbf{A}) = \max\limits_{i=0,\ldots,n} r_i(\mathbf{A})$ — *максимальный TT-ранг*.

Замечание: TT-формат требует $O\left(ndr^2(\mathbf{A})\right)$ памяти для хранения $O\left(d^n\right)$ элементов ($d = \max\limits_{i=1,\ldots,n} d_i$).

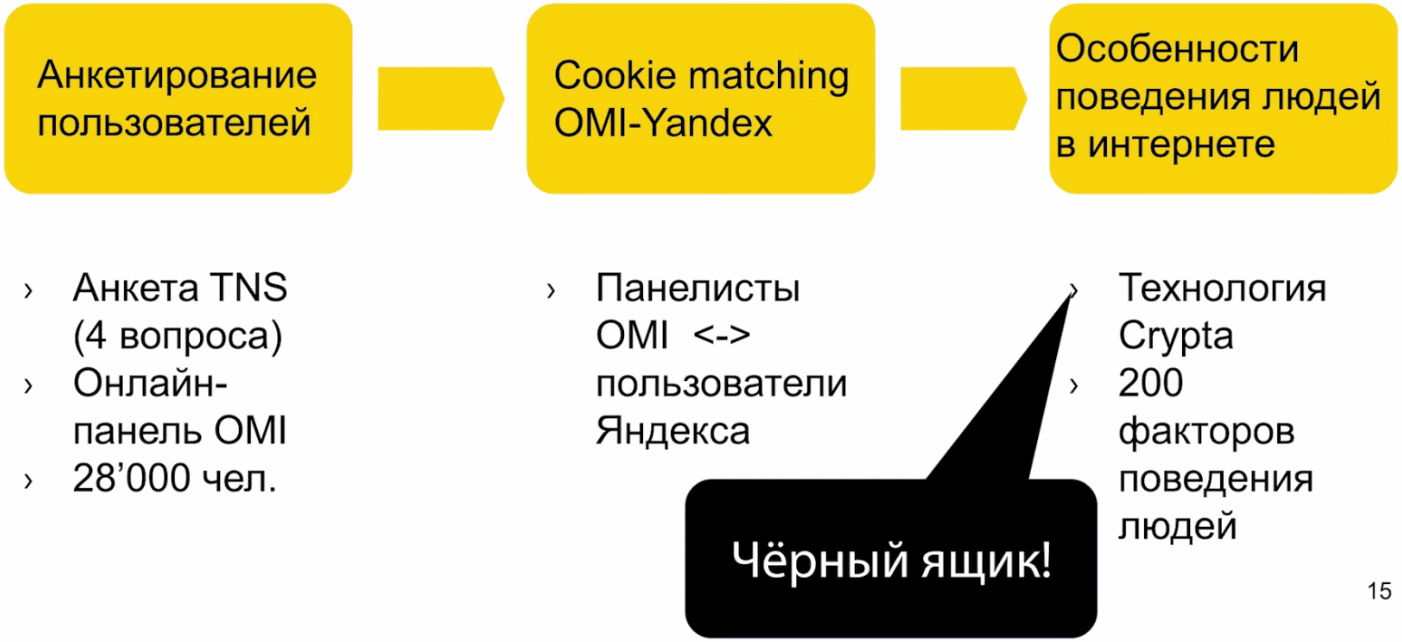
Где это уже работает

- Какую книгу (пост, новость) тебе стоит прочитать?
- Какую музыку ты хочешь послушать?
- Какой товар ты собираешься купить?
- С какой девушкой тебя познакомить?
- Готов ли ты сменить оператора?
- Можно ли тебе дать кредит?



Примеры Яandex

Light TV-viewers: методология



Примеры Яandex

Heavy TV viewers

СБЕРБАНК

«сбербанк», «коммунальный», «шарлотка», «выкройка», «биглион», «irr», «заработать»

Больше запросов кириллицей

Light TV viewers

«книга», «переводчик», «словарь», «формула», «японский», «французский», «немецкий», «такси»

Много запросов латиницей

Примеры Яandex

Heavy TV viewers

«тнт», «дом-2», «телепрограмма», «стс»

Light TV viewers

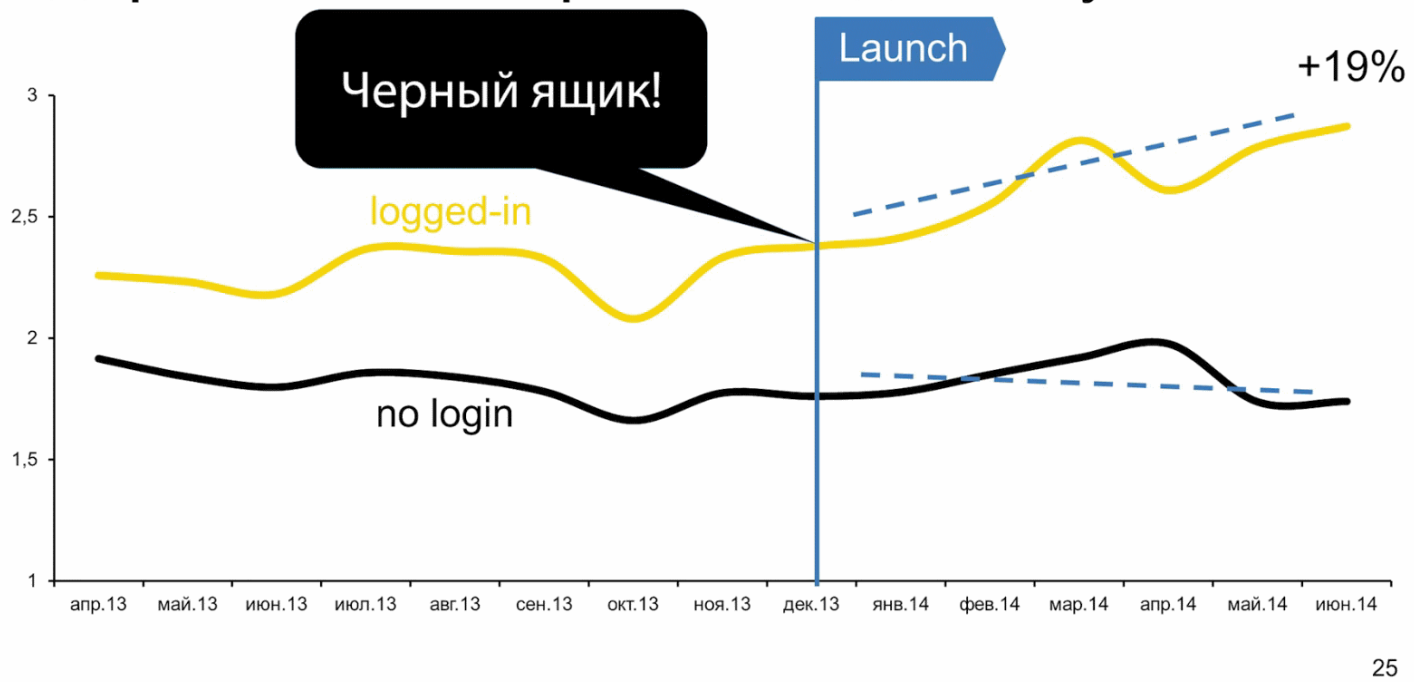
«C++», «wi-fi», «фотошоп», «torrent», «adobe»

Примеры Яandex

On-line gamers

[dota] [кпд] [массовка] [cheat] [эмулятор] [варфэйс] [капа][паркур] [гайд] [дрифт] [замужество] [партнерка] [прицел] [приворот]

Примеры Яandex

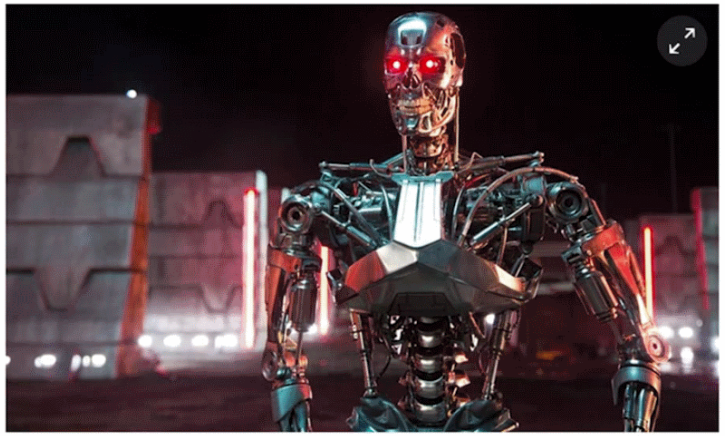


Примеры Яandex

Дальше - больше

Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons

More than 1,000 experts and leading robotics researchers sign open letter warning of military artificial intelligence arms race



“AI technology has reached a point where the deployment of [autonomous weapons] is – practically if not legally – feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms…

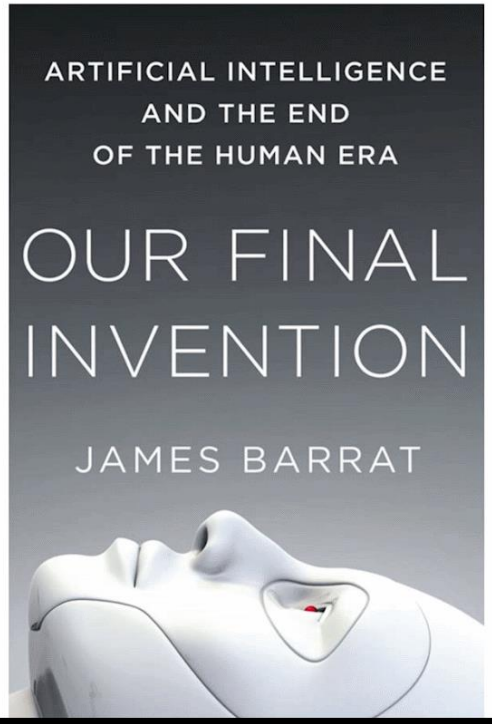
The endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow.”

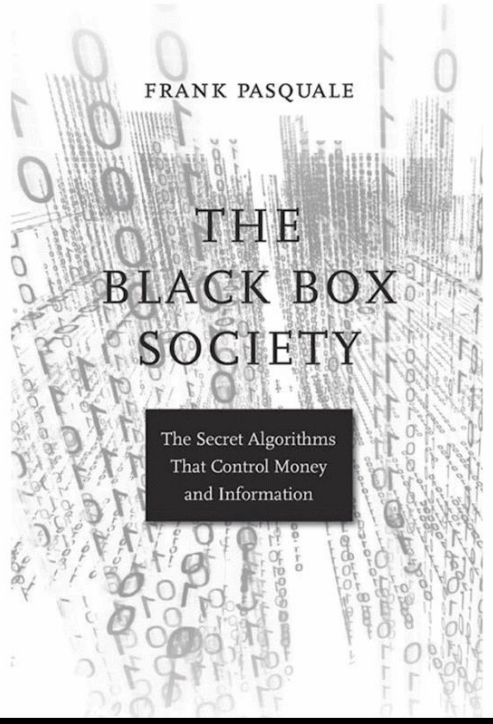
38

http://www.theguardian.com/technology/2015/jul/27/musk-wozniak-hawking-ban-ai-autonomous-weapons

Примеры Яandex

Активисты уже пишут книги

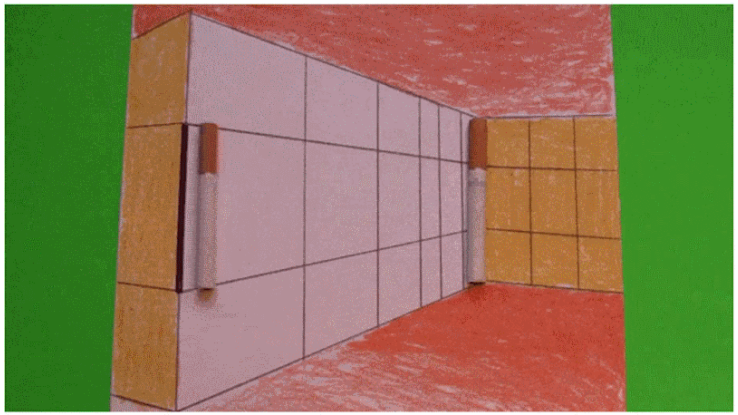




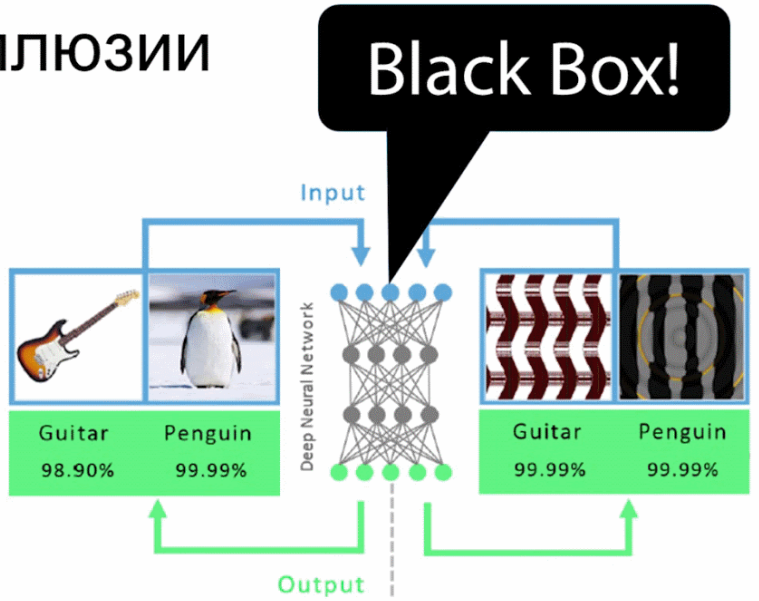
39

Примеры Яandex

У алгоритмов свои иллюзии



Эту классику мы все знаем



А вот это ученые только начинают исследовать ;)

http://arxiv.org/abs/1412.1897

Примеры Яandex

Термин и имя, которые надо знать

Термин:

Технологическая сингулярность

(важное свойство:
за этой точкой принципиально не работают
прогнозы и экстраполяции. 2025 – 2050?)

Имя:

Рэймонд Курцвейл (Raymond Kurzweil)

(отнюдь не просто футуролог)