«Прикладные задачи анализа данных»

Термины

Александр Дьяконов

(ВМК МГУ имени М.В. Ломоносова)

7-8 ноября 2019 года

Ключевые слова

Наука о данных (Data Sciense)

Статистика (Statistics)

Искусственный интеллект (Artificial Intelligence)

Анализ данных (Data Mining)

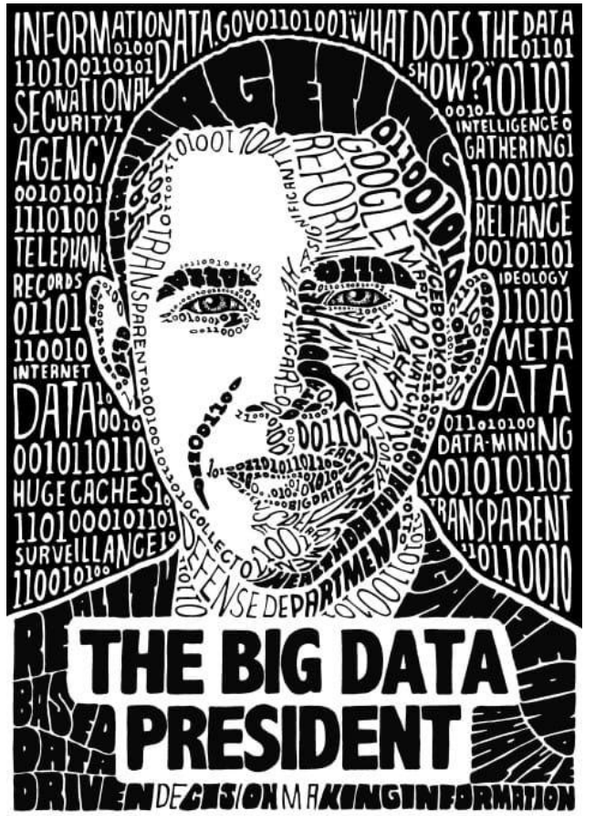
Машинное обучение (Machine learning)

Большие данные (Big Data)

Наука о данных (Data Sciense)

– направление науки и технологий представления, сбора, обработки, хранения, анализа и использования данных в цифровой форме

всё перечисленное выше – разделы DS



https://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html

Анализ данных (Data Mining)

– нахождение закономерностей и моделей, которые

- **валидны**
  (соответствуют действительности и есть в новых данных)
- **полезны**
  (экономят время, ресурсы, позволяют заработать $)
- **нетривиальны**
  (неочевидны до анализа)
- **понятны / интерпретируемы**
  (описываются, могут быть объяснены специалистам)

в широком смысле – область человеческой деятельности
(не наука! т.к. также искусство, ремесло, спорт)

Математическая статистика

– математическая дисциплина, разрабатывающая математические методы систематизации и использования статистических данных для научных и практических выводов

уже была в обязательных курсах…

Машинное обучение (Machine Learning)

Что такое обучение?

Машинное обучение (Machine Learning)

Обучение — приобретение необходимой функциональности посредством опыта

Обучение на примерах

Учимся ходить

Делаем шаг – получилось / нет

Учим названия животных

Показывают и называют

Обучение по определениям

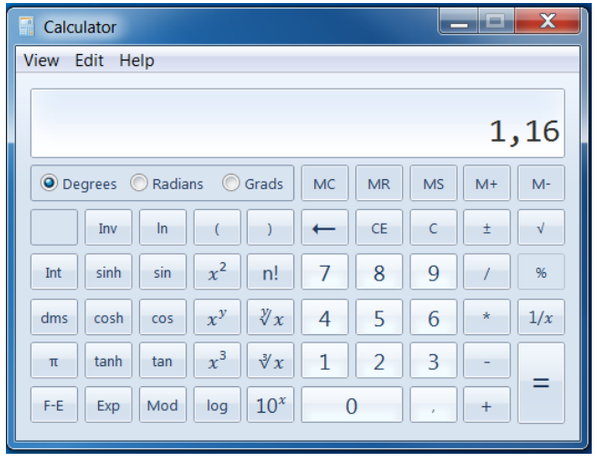
В школе – дают определения

Машинное обучение

Машинное обучение — процесс, в результате которого машина способна показывать поведение, которое в нее не было явно запрограммировано

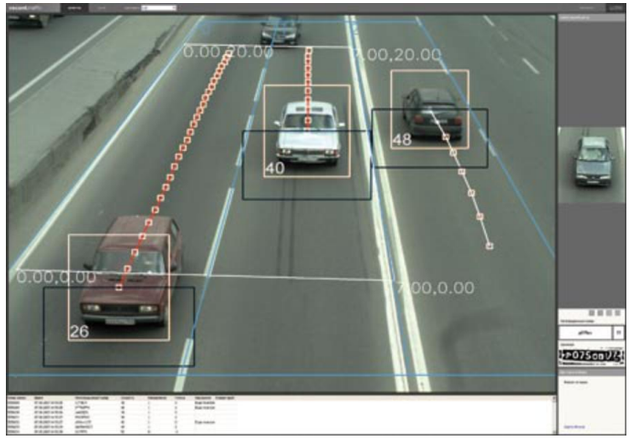
A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

Программирование



Программируем последовательность действий

Обучение

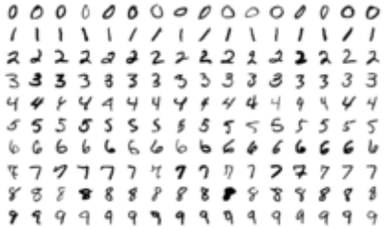


Программируем алгоритм анализа информации

«Машинное обучение» – наука!

Машинное обучение

«Компьютерная программа обучается из опыта E в классе задач T с мерой качества P, если качество измеренное с помощью P в классе задач T увеличивается по мере увеличения опыта E». Том Митчел



Задача: распознавание символов
Мера: процент правильно распознанных
Опыт: база, размеченных вручную, изображений символов

Задача: игра в шашки / шахматы / го
Мера: процент побед
Опыт: игра программы против себя

Задача: рекомендация товаров/услуг/видео
Мера: процент успешных рекомендаций
Опыт: список товаров, просмотренных/ купленных/оцененных пользователями

Примеры задач

- диагностика болезней, прогнозирование эффективности лекарства
- распознавание образов, символов (Character/ Handwriting Recognition)
- распознавание речи (Speech Recognition)
- распознавание лиц (Face detection)
- классификация спама (Spam filtering)
- идентификация (Person identification / Authentication) лица, отпечатков, радужка глаза и т.п.
- тональность текста (sentimental analysis)
- прогноз спроса / выручки (Demand Forecasting)
- скоринг (Credit scoring) – определение кредитоспособности
- определение суммы / пакета страхования
- психотип по профилю соцсети / фотографии
- предсказание оттока (ухода сотрудника / абонента)
- поиск кандидатов на вакансии
- рекомендации товаров
- ранжирование Web-страниц
- ожидание прибыли магазина (учитывая GPS) / рейтинга фильма / доходности сделки
- анализ форумов, поиск оскорблений, жалоб, автоматическая модерация
- предсказание поведения клиента / пользователя (ex: трат клиента)
- поиск похожих объектов, документов, событий (например, юридических дел)
- обнаружение нетипичных пользователей, фрода, инсайдеров
- нахождение зависимостей
- сегментация изображений
- тегирование/аннотирование документов (automatic summarization)

Пример задачи машинного обучения – классификация

Iris setosa

Iris virginica

Iris versicolor

Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Вид ириса
4.3	3.0	1.1	0.1	setosa
4.4	2.9	1.4	0.2	setosa
4.4	3.0	1.3	0.2	setosa
				…
4.9	2.5	4.5	1.7	virginica
5.6	2.8	4.9	2.0	virginica
				…
5.0	2.0	3.5	1.0	versicolor
5.1	2.5	3.3	1.1	versicolor

Пример задачи машинного обучения – скоринг

Id	статус	г.р.	Пол	офис	На счету	просрочки	возврат
43223	физ	1967	М	54	10000	0	Да
43224	физ	1970	Ж	33	2000	2	Нет
43225	юр	1954	М	54	23500	0	Да

Прогноз поведения пользователя с помощь описания (и кредитной истории)

Большие данные (Big Data)

– технологии сбора, хранения, обработки и анализа данных огромных объёмов и значительного многообразия

Характеристики:

VELOCITY
скорость поступления
VOLUME
объёмы
VARIETY
разнообразие

VERACITY
достоверность

Причины

- удешевление средств хранения
- ускорение средств обработки
- миниатюризация устройств (смартфоны, датчики и т.п.)
- новые форматы / неструктурированность
- новые технологии (GPS)
- интерес бизнеса
- успехи отдельных подходов в ML (например, DL)

коммерческий и технологический термин

Большие данные (Big Data)

Пример:

Google Flu Trends

https://www.google.org/flutrends/about/

- **анализ поисковых запросов**
- **корреляция с известными эпидемиями**
- **прогнозная модель**

Виктор Майер-Шенбергер и Кеннет Кукьер Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим

Искусственный интеллект (Artificial Intelligence)

– наука и технология создания интеллектуальных машин (в том числе, программ)

– свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека

- умные чат-боты
- автомобили-беспилотники
- умный дом

IBM построила Watson, который выиграл в Jeopardy

сейчас самый популярный термин

Искусственный интеллект (Artificial Intelligence)

Проблема «почти реализации»

Как только машина «учится новым способностям» выясняется, что за этим стоят простые вычисления. Можно ли считать это AI?

AI в сильном смысле

компьютеры могут приобрести способность мыслить и осознавать себя как отдельную личность (в частности, понимать собственные мысли)

AI в слабом смысле

Проблема сознания

- самоидентификация
- идентификация других и противопоставление
- борьба за ресурсы

Big Data-аналитика

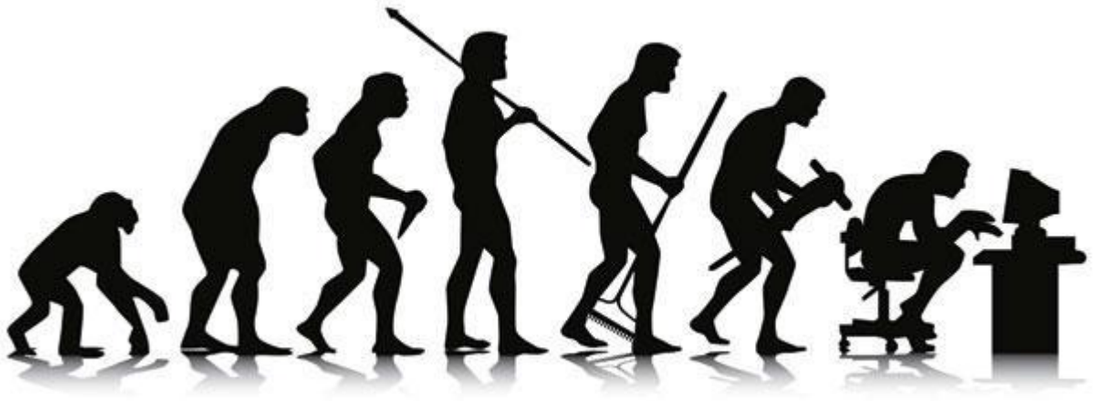
1. Использование ВСЕХ данных, а не случайных выборок

2. Меньшие требования к точности

3. Не ищем причины, а корреляции

4. Важна Датификация

Когда появился анализ данных



3000/6000 лет до н. э. – письменность

2000 лет до н. э. – протоматематика, протоастрономия

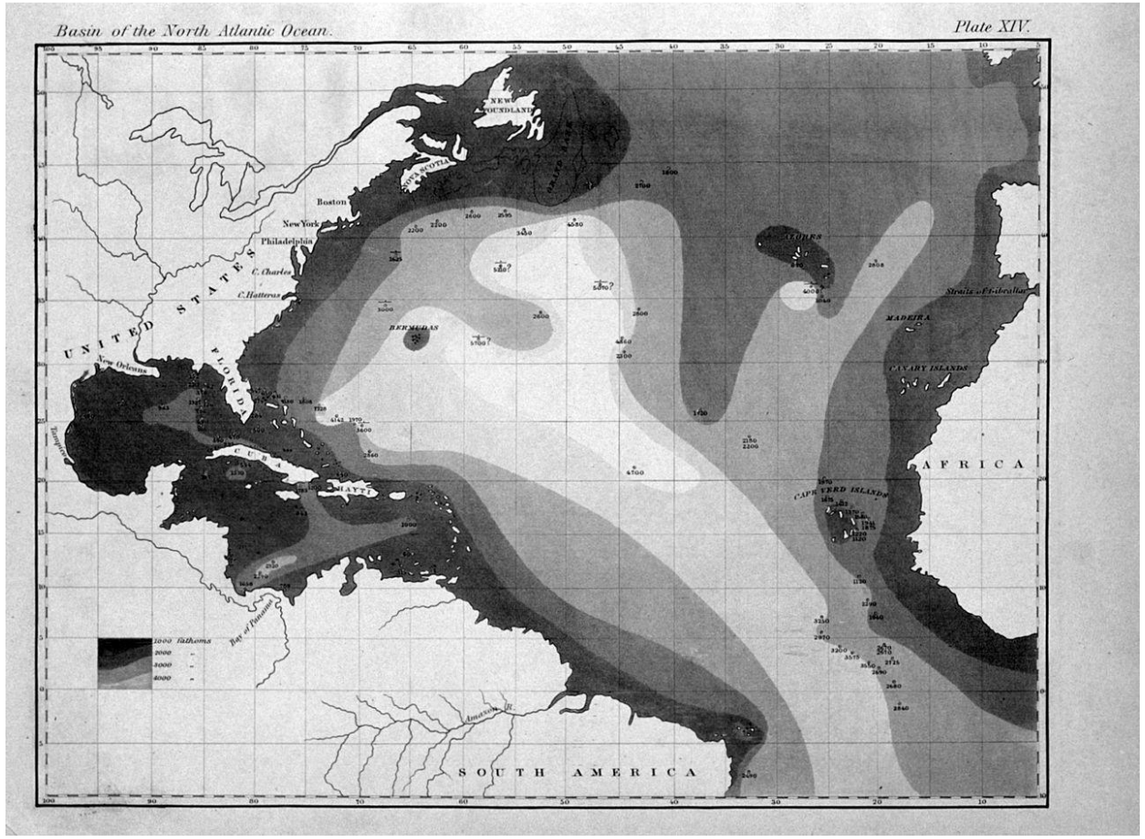
3000 лет до н. э. – протохимия (получали медь, серебро, свинец)

5–6 в. до н.э. – математика как наука

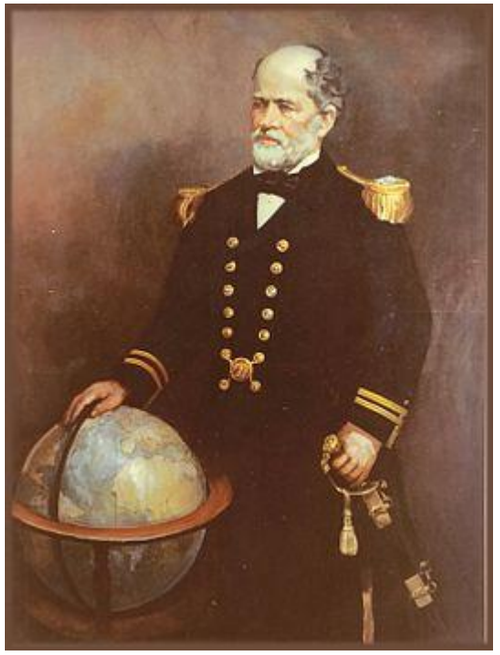
5–2 в. до н.э. – физика (Китай, Греция)

19 в. – протоанализ данных

Исследования начальника Архива морских карт в Вашингтоне



Сокращение времени плавания судов, пользуясь попутными ветрами и течениями

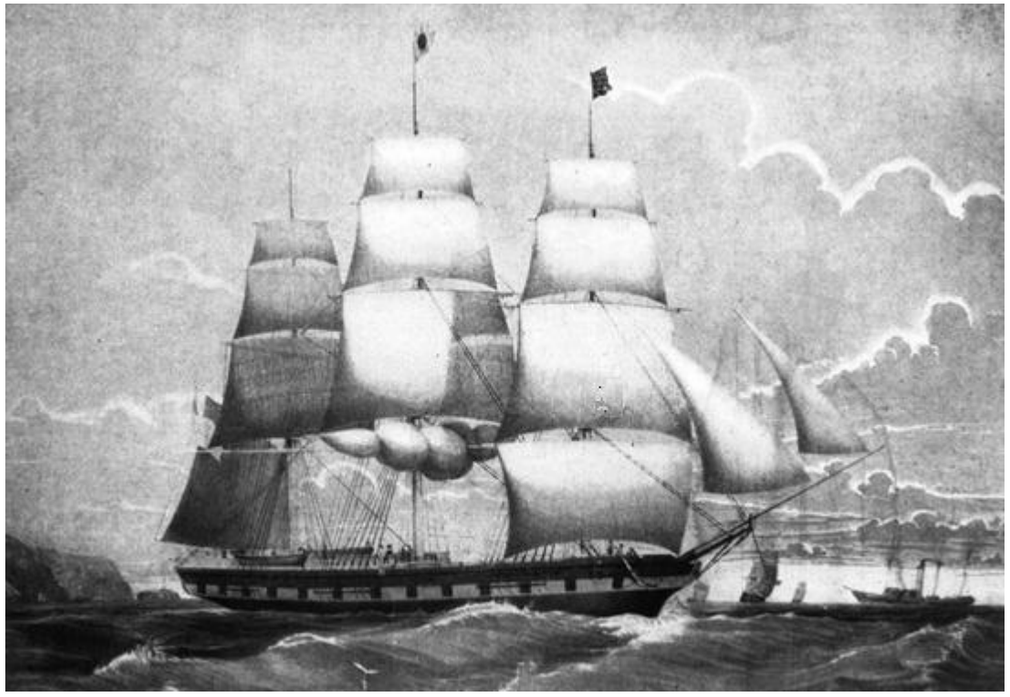


Мэтью-Фонтейн Мори

(14.01.1807 — 01.02.1873)

американский морской офицер, астроном, историк, океанограф, метеоролог, картограф, геолог

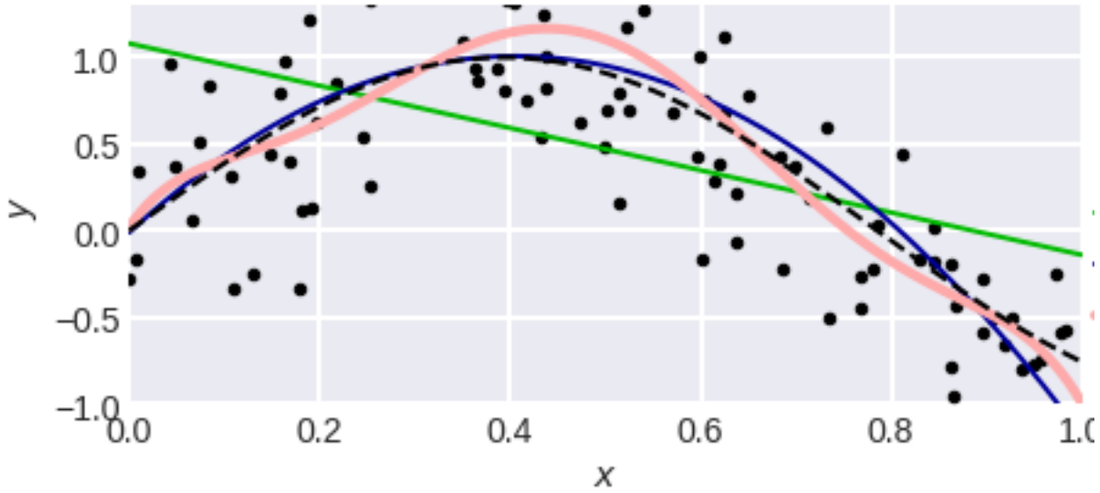
Исследования начальника Архива морских карт в Вашингтоне



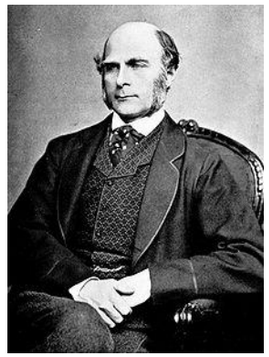
Первые «большие данные» в картографии – сбор сведений морских журналов

Первая профессиональная соцсеть – обмен информацией, сотрудничество в анализе течений (бутылочная почта)

Математический аппарат – первые работы



1886, Регрессия



Фрэнсис Га́льтон

(16.02.1822 – 17.01.1911)

1795 – 1805 Метод наименьших квадратов

Иога́нн Карл Фри́дрих Га́усс

(30.04.1977 – 23.02.1855)

Адриен Мари Лежа́ндр

(18.09.1752 – 10.01.1833)

Анализ поведения людей

Задача: оценка миграционных потоков, их изменение в зависимости от политики и административных решений

Анализ поведения людей

по данным городских служб

Задача: согласование данных разных источников

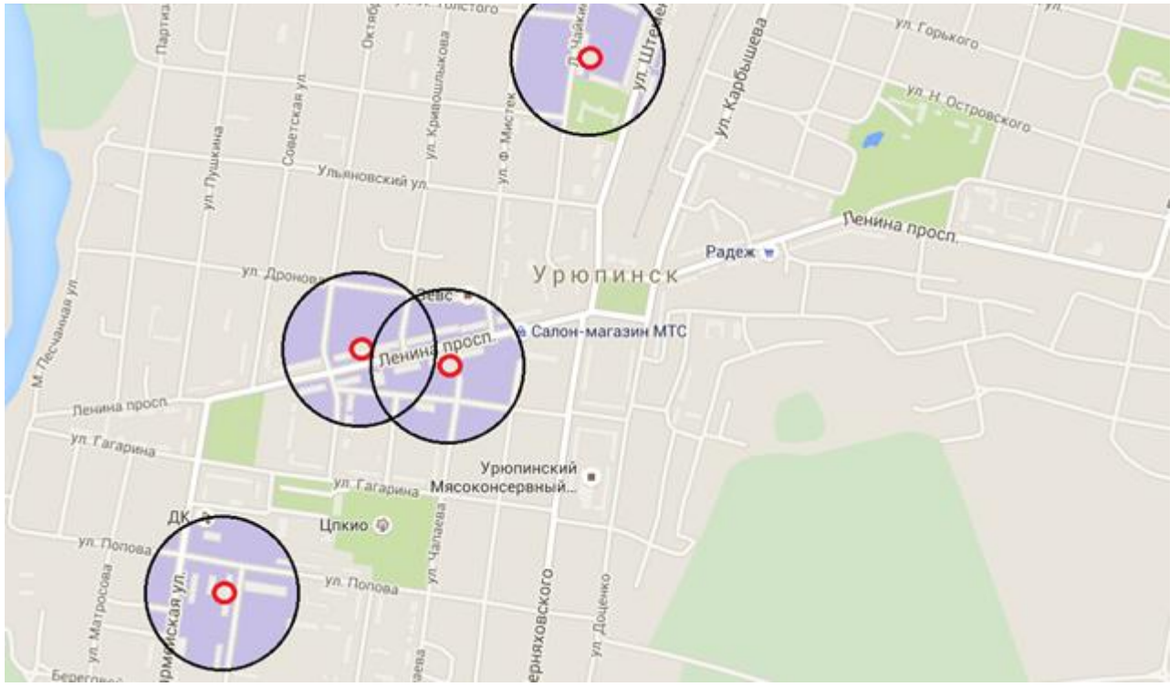
(иногда несоответствие очень большое)

Анализ поведения клиентов

Интересно: счётчики посещения есть даже в обычных магазинах

Задачи: анализ конверсии / трафика

Обнаружение аномалий: нетипичных точек продаж



Обнаружение аномалий

выявление нетипичного поведения

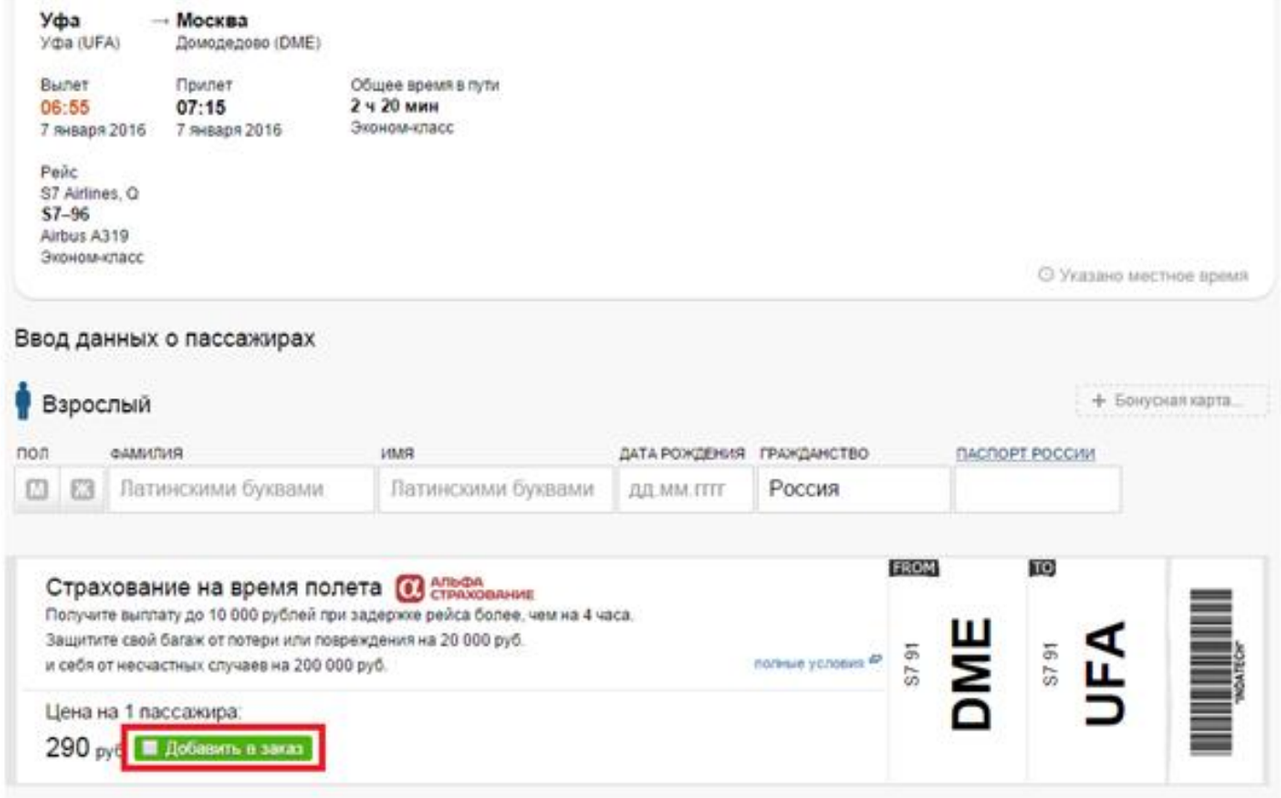
- **подозрительное поведение в толпе**
- **подозрительные финансовые операции**
- **выявления инсайдеров**

Анализ поведения клиентов

- нахождение целевой аудитории
- определение интересов клиента (рекомендательные системы)
- кросс-продажи
- дополнительные услуги
- прогнозирование спроса
- повышение конверсии, управление ценой
- оптимальный контент

(исследование – использование)

Предложение дополнительных услуг



Есть статистика – кто и когда покупал страховку, а кто – нет

Надо: сделать предложение таргетированным

Анализ поведения клиентов

ID 34377420

Компания Игра Игровой набор Шахматы и шашки 2 в 1

0 0 У меня это есть

Новинка

Цвет: черный, белый

Основные свойства:

Тип	Шахматы
Возраст ребенка	От 6 лет
Кол-во игроков	2
Вид настольной игры	В дорогу
Вид классической игры	Шахматы, Шашки

Рекомендуем также

Книгопечатная продукция (
Современное проектиро
Андрей Александре
821 ₽
В корзину

Настольная игра
Набор дорожный 2в1 "Шахматы, шашки".
1 155 ₽
В корзину

Настольная игра
Mack & Zack Магнитная игра Шахматы
358 ₽
В корзину

Шахматы
Настольная игра Tactic "Шахматы", 40218
1 690 ₽
В корзину

Шахматы
Игровой набор 3в1 "Игровые": нарды,
2 628 ₽
В корзину

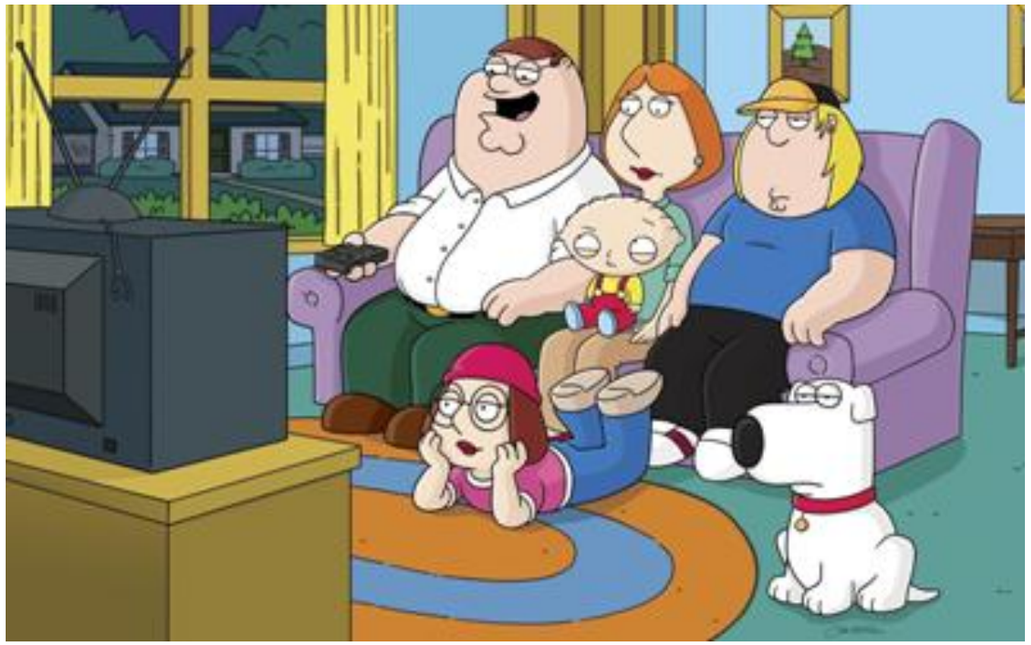
Шахматы
Уценённый товар. Шахматы "Сенатор".
3 466,80 ₽
В корзину

Настольная игра
Игровой набор 2в1 "Шахматы, шашки".
1 723 ₽
В корзину

Книгопечатная продукция (
Курс наглядной геометрии и топологии
707 ₽
В корзину

Рекомендации: статистика + контент

Анализ поведения клиентов

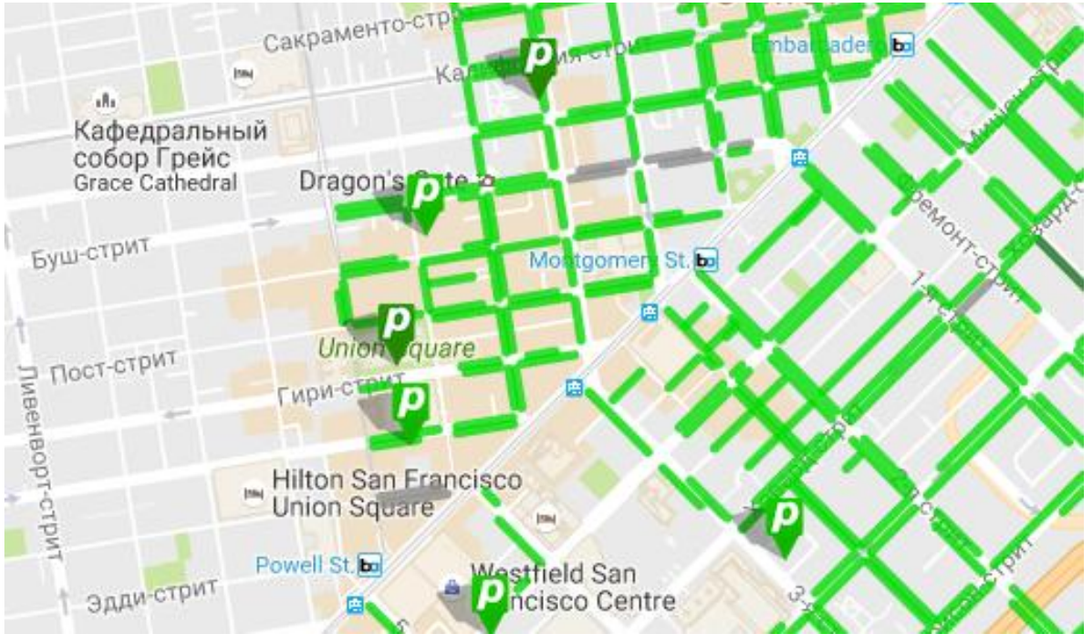


	МУЗ ТВ	НТВ	ПЕРВЫЙ КАНАЛ	ПЯТНИЦА	ПЯТЫЙ КАНАЛ	РЕН ТВ	РОССИЯ 1	РОССИЯ 24	РОССИЯ К	СТС
МУЗ ТВ	100	85	86	84	84	88	86	85	74	87
НТВ	85	100	94	89	91	90	94	88	81	90
ПЕРВЫЙ КАНАЛ	86	94	100	89	91	89	96	87	80	91
ПЯТНИЦА	84	89	89	100	87	86	88	84	78	87
ПЯТЫЙ КАНАЛ	84	91	91	87	100	88	90	86	81	87
РЕН ТВ	88	90	89	86	88	100	90	88	78	91
РОССИЯ 1	86	94	96	88	90	90	100	87	80	90
РОССИЯ 24	85	88	87	84	86	88	87	100	79	87
РОССИЯ К	74	81	80	78	81	78	80	79	100	77
СТС	87	90	91	87	87	91	90	87	77	100

Анализ аудитории каналов

Планирование рекламы

Анализ открытых данных



- **анализ данных счётчиков парковок, предложение маршрутов**
- **сервис по пробкам / прогноз пробок**
- **прогноз задержек транспорта и планирование маршрутов**

Задача: прогноз задержек общественного транспорта

Задача: прогноз криминальной активности

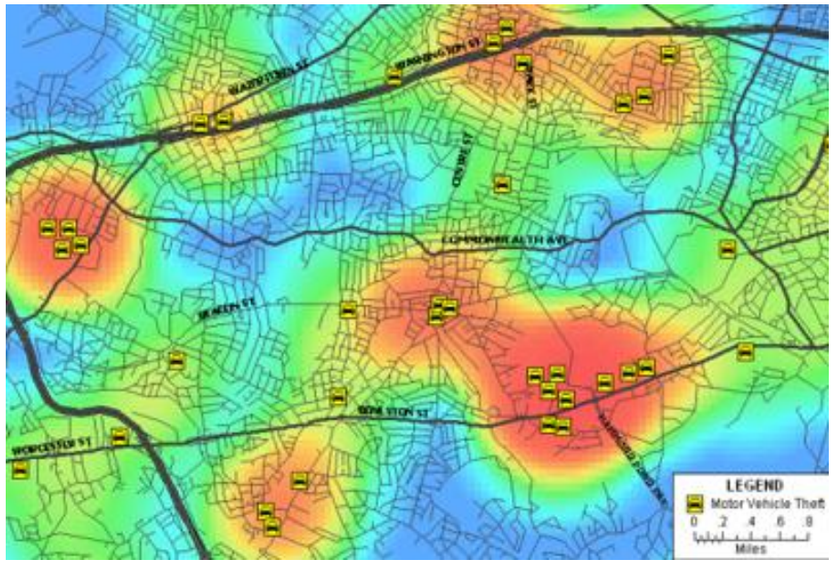
WIKIPEDIA
The Free Encyclopedia

Criminology and penology

Predictive policing

From Wikipedia, the free encyclopedia

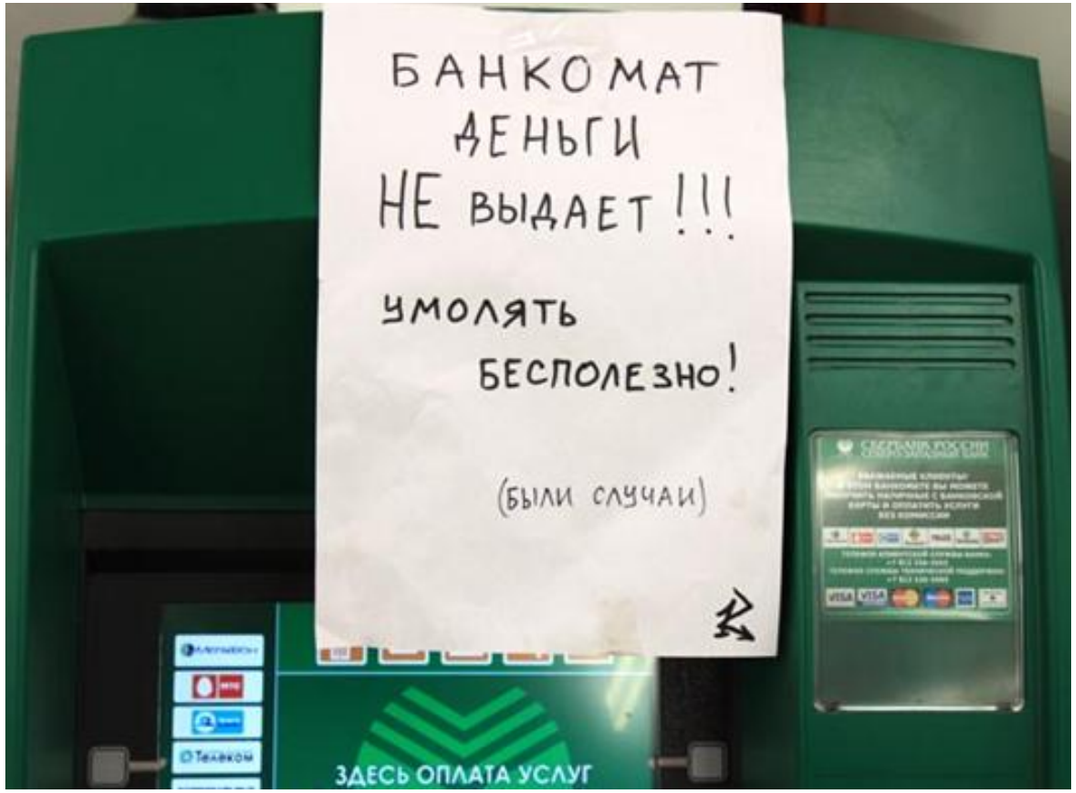
Predictive policing refers to the usage of mathematical, predictive and analytical techniques in law enforcement to identify potential criminal activity.[1] Predictive policing methods fall into four general categories: methods for predicting crimes, methods for predicting offenders, methods for predicting perpetrators' identities, and methods for



Интернет как источник данных

- Определение возраста по сообщениям в форуме
- Детектирование оскорблений
- Анализ отношения к бренду
- Анализ политической активности населения
- Рекомендации групп / новостей

Банковские задачи



- скоринг
- предсказание погашений кредитов
- предсказание сумм снятий с банкоматов

Автоматическая диагностика двигателей

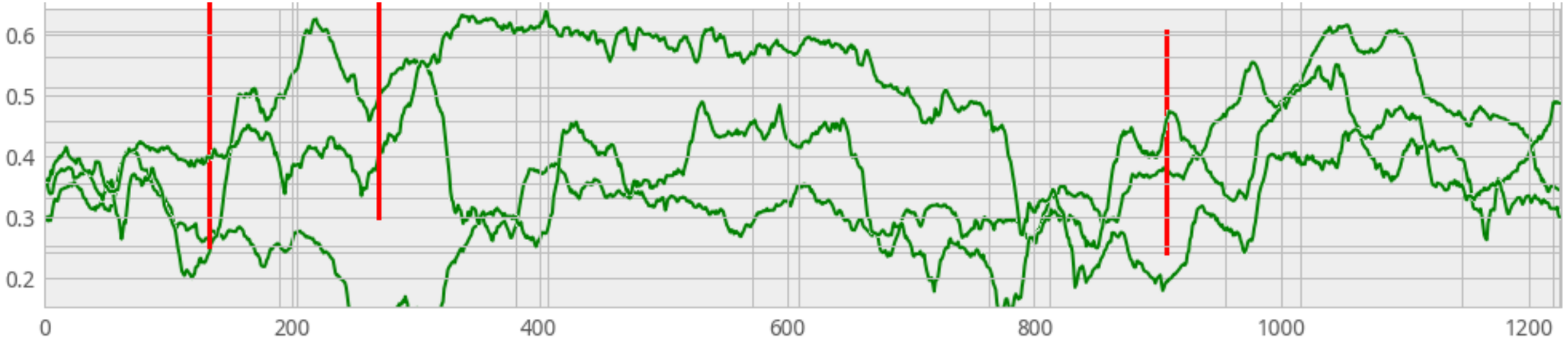
Автоматическая классификация и категоризация

Диагностика неисправностей оборудования

Детектирование поломок

Предсказание поломок

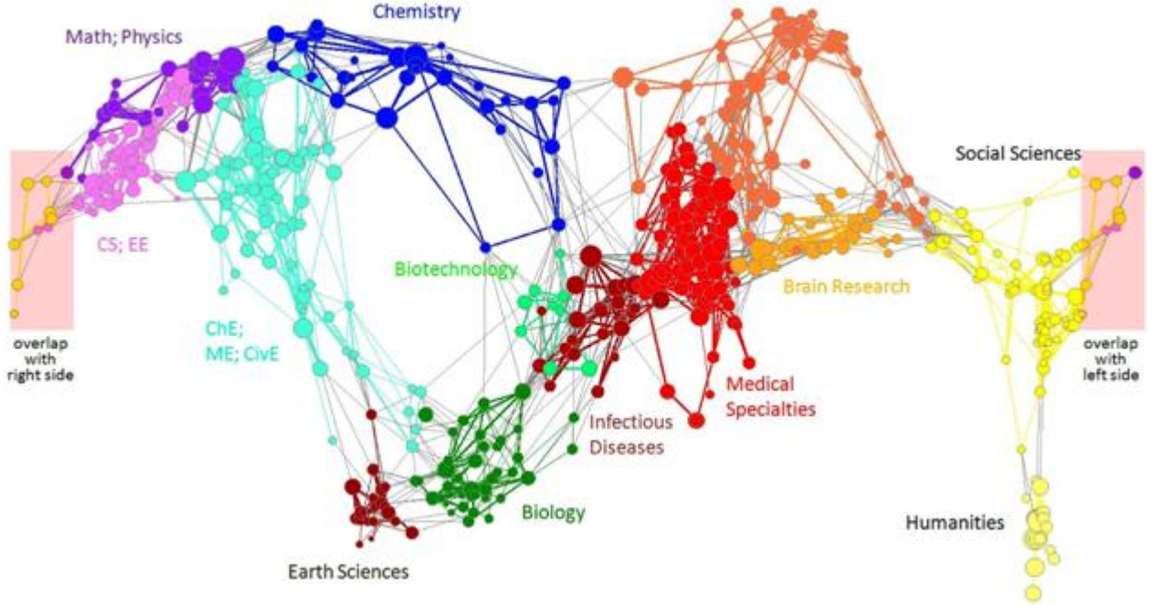
Анализ логов работ



Оценка персонала

- мониторинг качества обслуживания в колл-центрах
- оценка эффективности менеджеров
- система автоматического доступа к ресурсам

Анализ социальных сетей



- **Выявление сообществ в социальной сети**
- **Предсказание событий**
- **Рекомендации**

Граф цитирований Börner и др.

Валидация данных

→ да

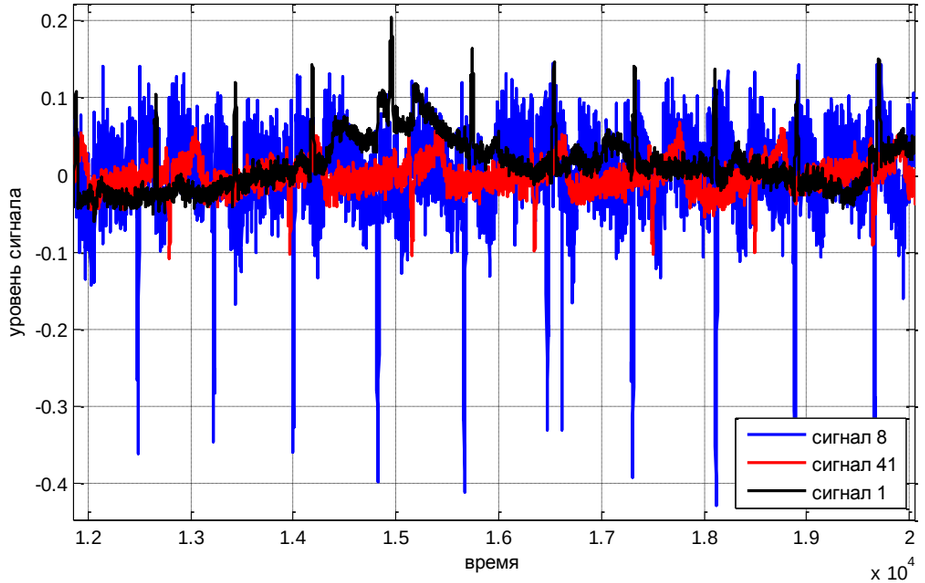
→ да

→ нет

Конкурс Avito:

Есть ли реклама на изображениях, выкладываемых на сайте

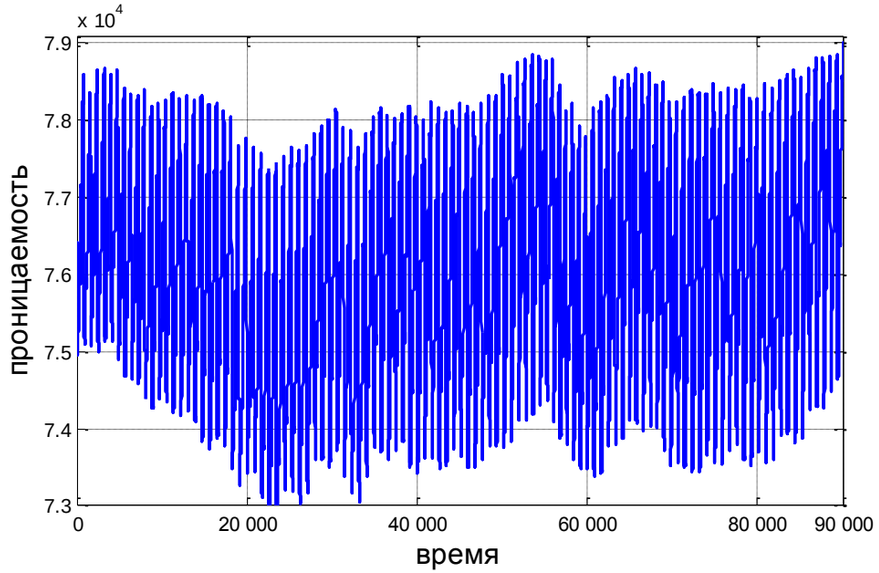
Анализ данных в медицине: Проект CardioQvark

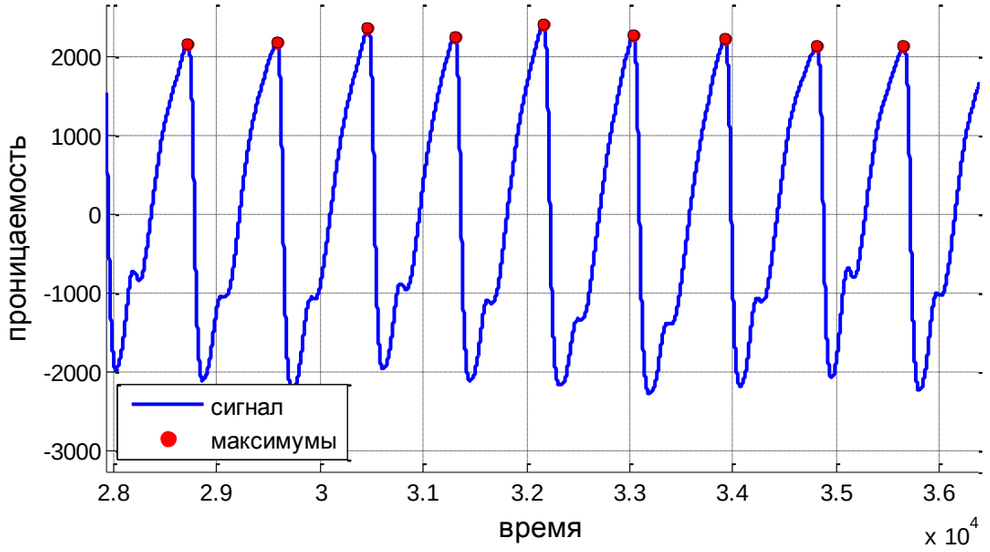


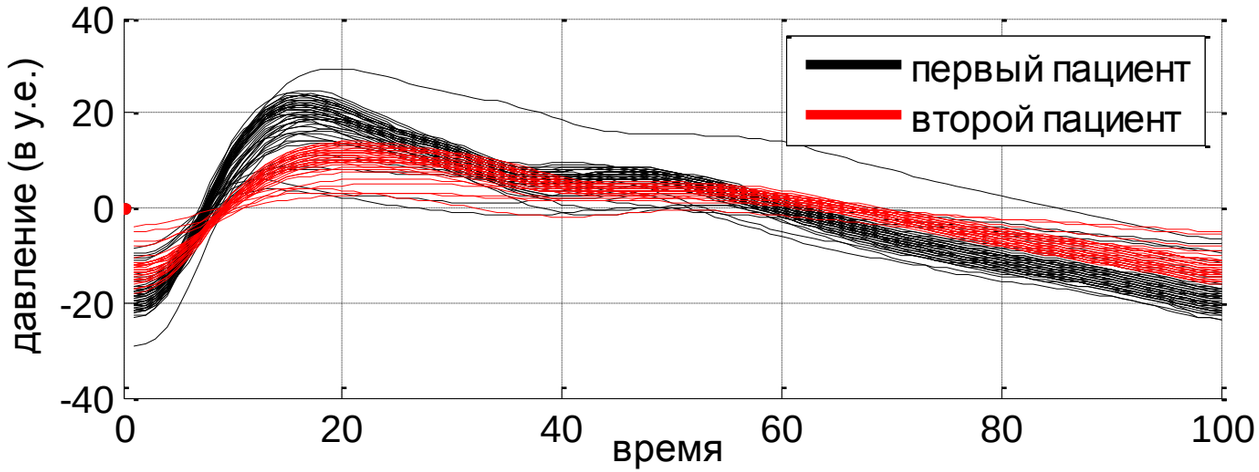
- Мониторинг состояния
- Предсказание осложнений
- Исследование ЭКГ
- (Big Data: постоянный поток данных от каждого пациента)
- Классификация (детекция курильщика по ЭКГ)

http://cardioqvark.ru

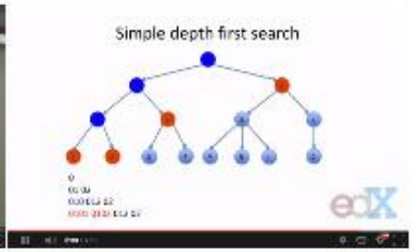
Анализ данных в медицине: анализ фотоплетизмограмм (Ангиоскан+АлгоМост)

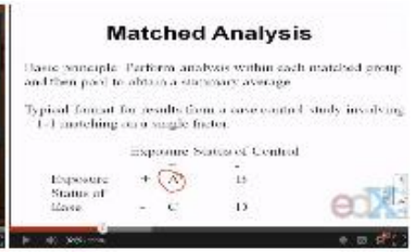


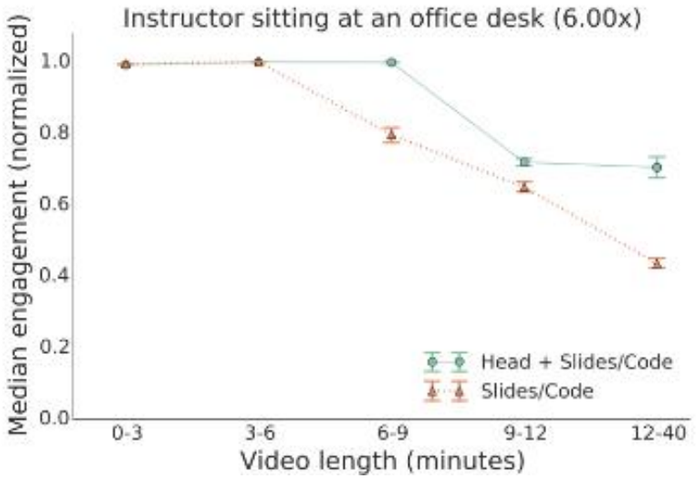


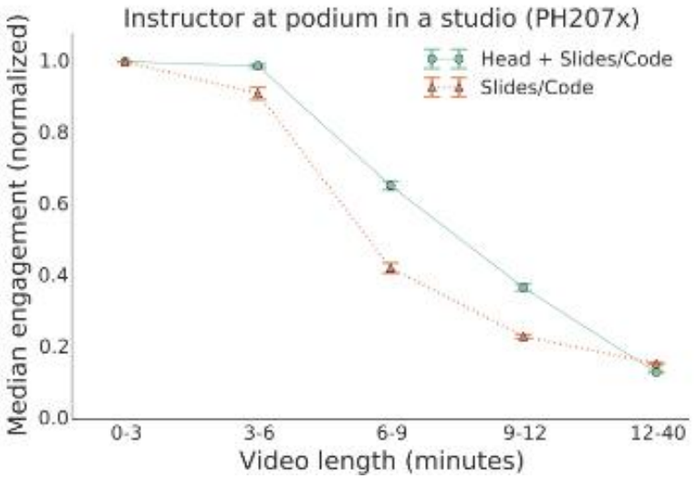


Анализ данных в образовании









Philip J. Guo, Juho Kim, Rob Rubin How video production affects student engagement: an empirical study of MOOC videos

короткие видео (<6 мин) эффективнее

Лучше лектор + слайды

Студийный видео менее привлекательней любительских

Рисование от руки более привлекательно, чем спецэффекты

Быстрый темп речи и энтузиазм более привлекательны

Анализ данных в образовании

Задача: предсказание ответов студентов на вопросы теста

для рекомендательной системы

(алгоритм решает за студента тест и сообщает ему «потенциально неприятные для него» вопросы)

Для изучения Машинного обучения

Лекции К.В. Воронцова

http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29

Derek Kane «Data Science»

https://www.youtube.com/channel/UC33qFpcu7eHFtpZ6dp3FFXw/videos

очень неплохой обзор всего-всего (включая применение в бизнесе + большие данные)

David S. Rosenberg «Foundations of Machine Learning»

https://bloomberg.github.io/foml/

очень хороший и продуманный курс,
обращают внимание на некоторые интересные детали

Roger Grosse, Amir-massoud Farahmand, Juan Carrasquill «Machine Learning and Data Mining»

http://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/

очень симпатичный курс, много тем, чётко изложены!

Книги

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer, 2009.
https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Charu C. Aggarwal Data Mining: The Textbook. Springer, 2015.

Mohammed J. Zaki, Wagner Meira Jr. «Data Mining And Analysis, Fundamental Concepts And Algorithms»

Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong «Mathematics for Machine Learning», 2019, https://mml-book.github.io

Книга по современному глубокому обучению
https://www.deeplearningbook.org

Обзорная книга

Виктор Майер-Шенбергер и Кеннет Кукьер
«Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим»

Соревнования по анализу данных

https://www.kaggle.com/