

Анализ данных Большие данные Машинное обучение

Дьяконов А.Г.

**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**

Первая лекция курса «Введение в машинное обучение» для студентов 1-2 курсов ВМК

Сразу пример



2009 г. – новый штамм вируса гриппа H1N1

Необходима локализация, но мед. статистика опаздывает на 10 дней

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

Google Flu Trends

Раньше

Анализ отчётов поликлиник

Теперь

**Анализ поисковых запросов
+ прогнозная модель**

Какие запросы?

«высокая температура»

«что делать при насморке»

...

Корреляция с распространением уже известных заболеваний

Google Flu Trends

Признаки того, что потом назовут «Big Data-аналитикой»

- не строим модель
- используем все данные
- ищем закономерности для аналитики

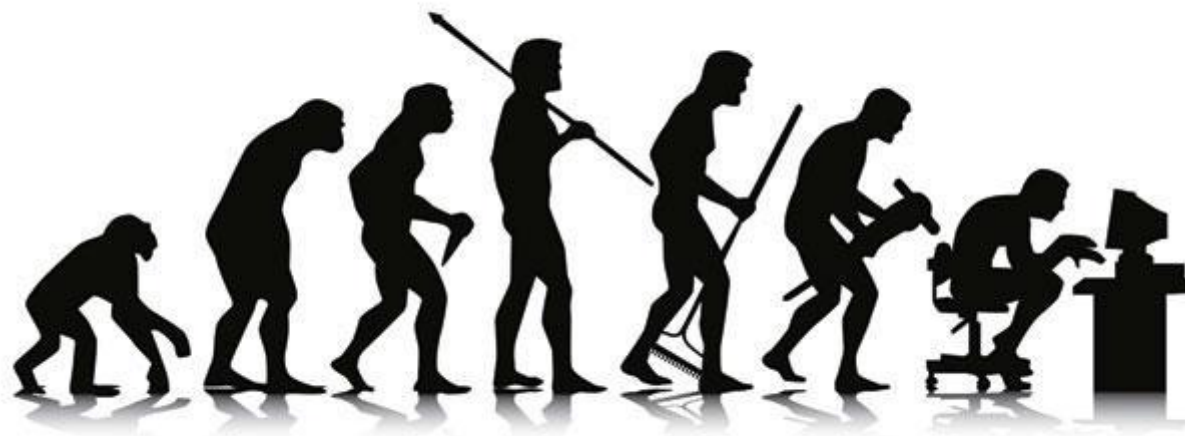
Поучительно

2009–2011 – удивительно точный прогноз

2012–2013 – огромные ошибки при прогнозе (**проект закрыт**)

Возможная причина: «отменяющиеся прогнозы»

Когда появился анализ данных



3000/6000 лет до н. э. – письменность

2000 лет до н. э. – протоматематика, протоастрономия

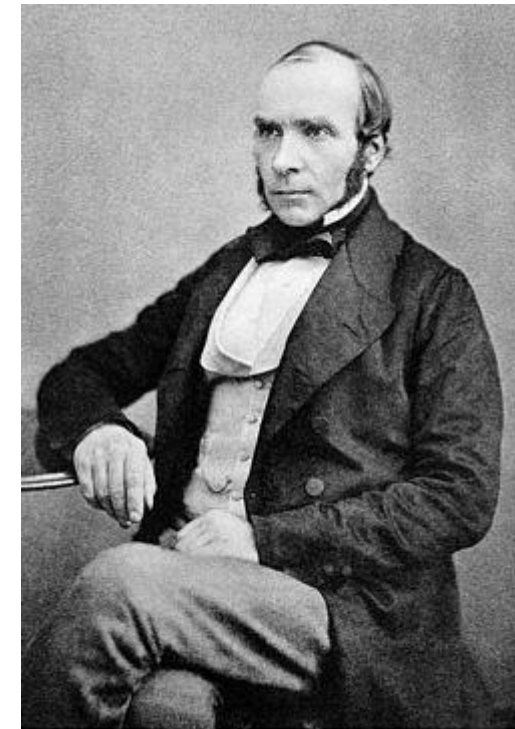
3000 лет до н. э. – протохимия (получали медь, серебро, свинец)

5–6 в. до н.э. – математика как наука

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

19 в. – протоанализ данных

Вспышка холеры на Брод-стрит в 1854 году



John Snow

Джон Сноу

(15.03.1813 — 16.06.1858)

британский врач, один из пионеров
массового внедрения анестезии и
медицинской гигиены

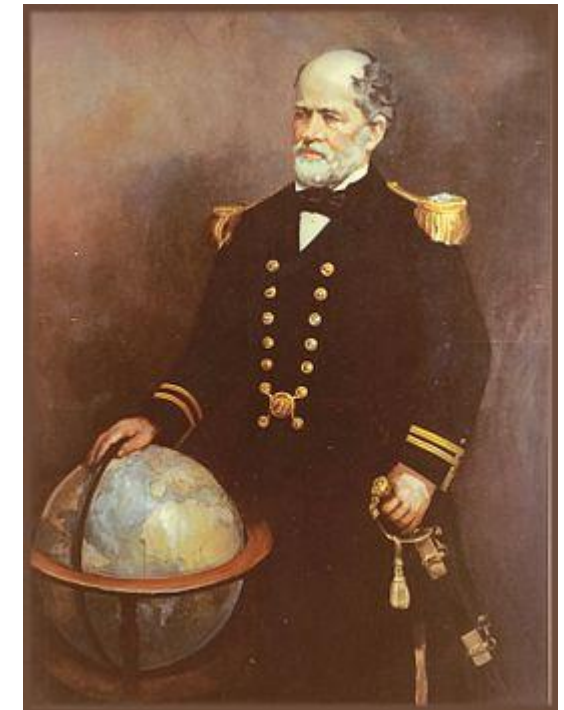
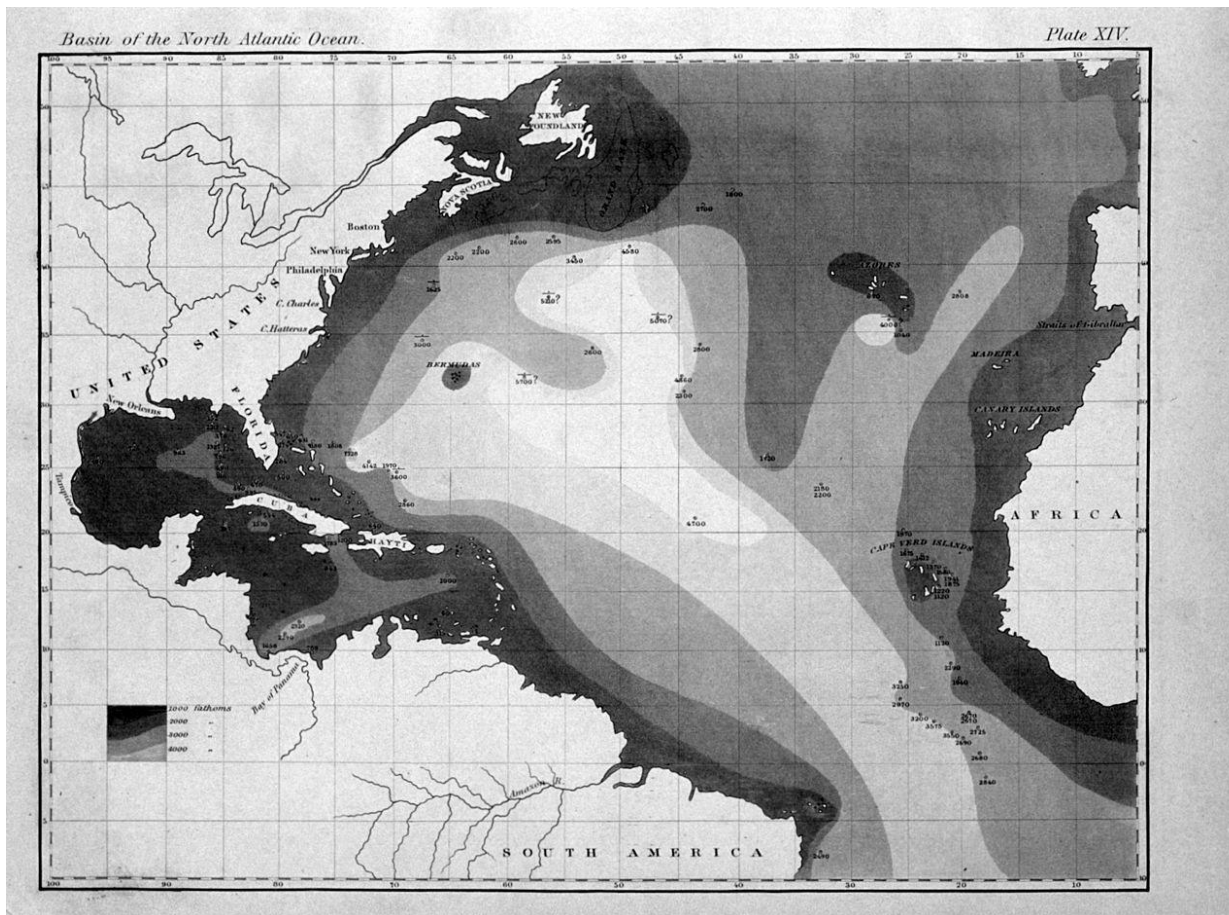
Вспышка холеры на Брод-стрит в 1854 году

Решение медицинской проблемы немедицинскими методами

Простое решение (нет сложной математики)

Не первое решение в АД (даже в медицине)

Исследования начальника Архива морских карт в Вашингтоне



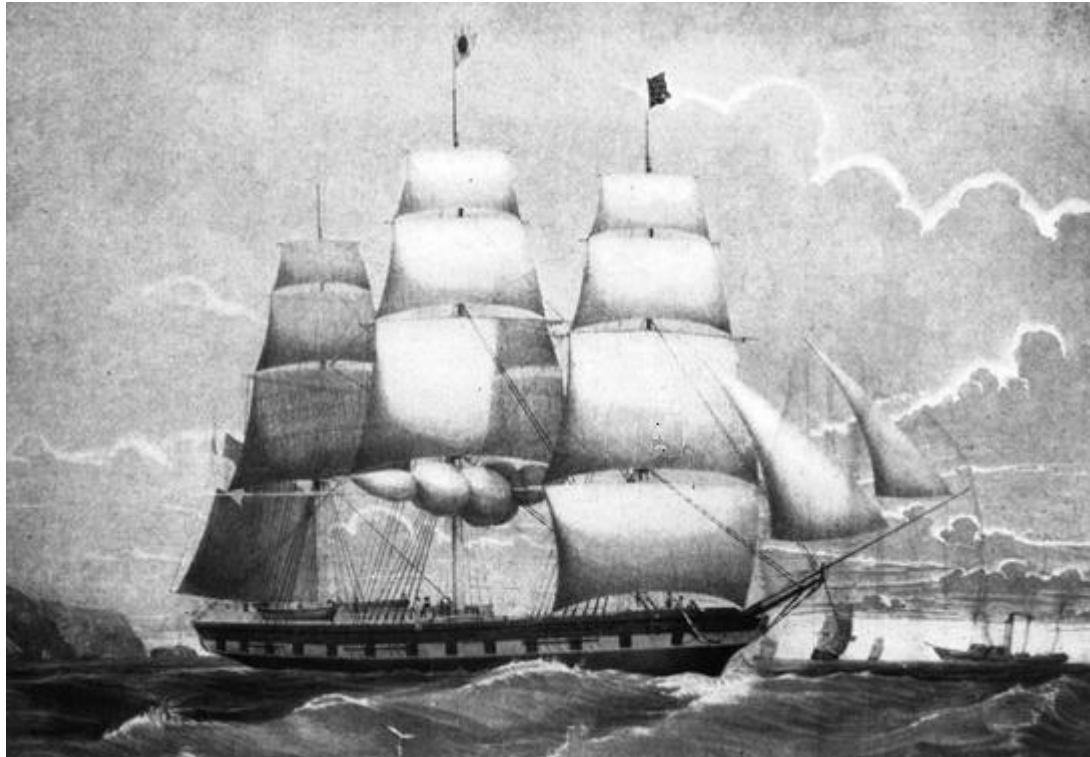
Мэтью-Фонтейн Мори

(14.01.1807 — 01.02.1873)

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

**Сокращение времени плавания судов,
пользуясь попутными ветрами и
течениями**

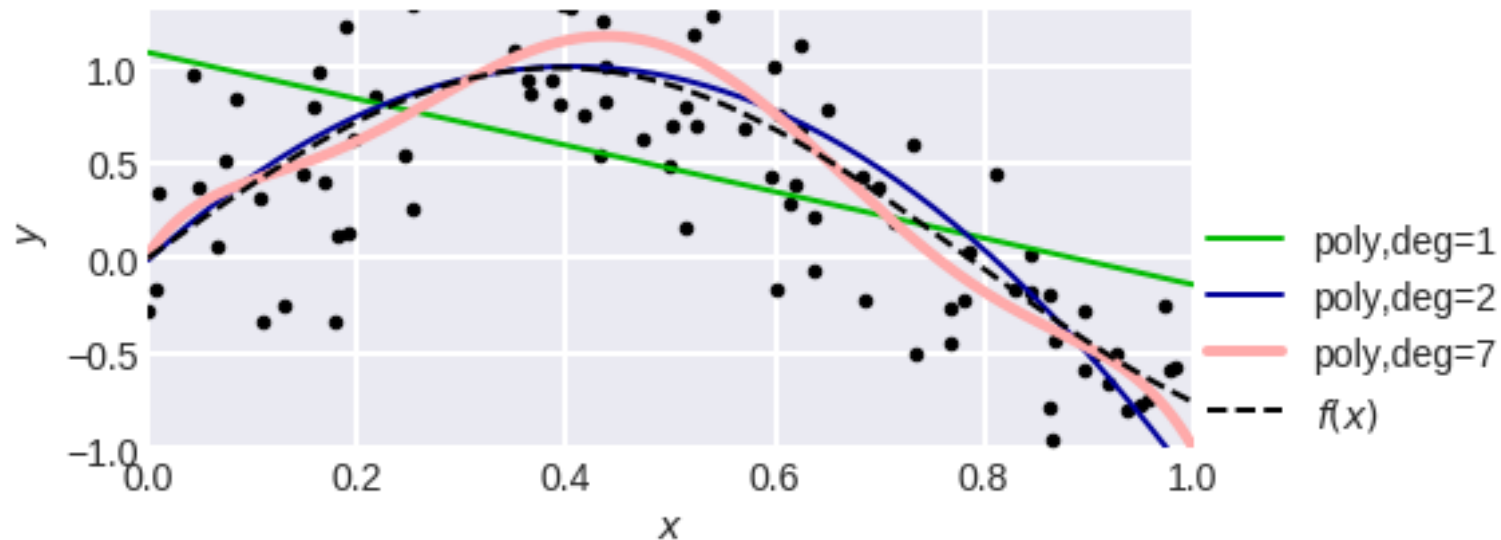
Исследования начальника Архива морских карт в Вашингтоне



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

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

Математический аппарат – первые работы



1795– 1805 Метод наименьших квадратов

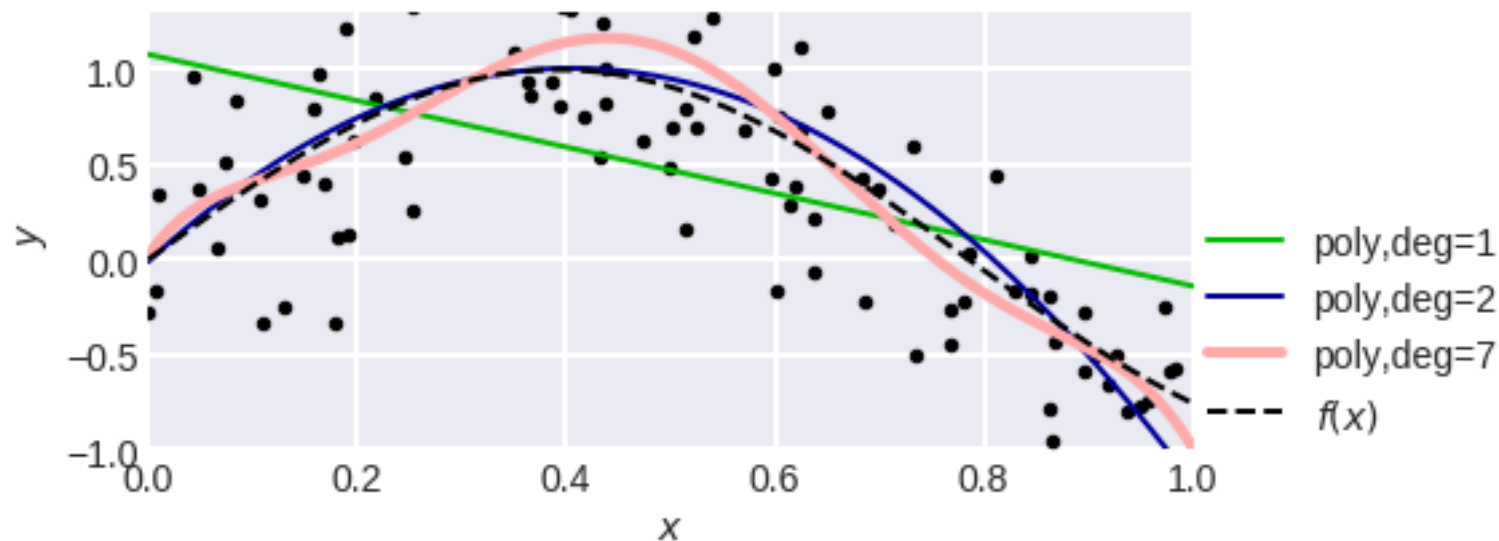


Иоганн Карл Фридрих Гаусс
(30.04.1777 – 23.02.1855)

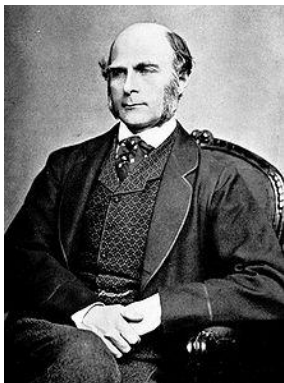


Адриен Мари Лежандр
(18.09.1752 – 10.01.1833)

Математический аппарат – первые работы



1886, Регрессия



Фрэнсис Гальтон
(16.02.1822 – 17.01.1911)

205 пар родителей и 930 их взрослых детей
«закон регрессии к среднему»

Для многих непрерывных признаков
(рост, интеллект и т.п.) взрослое потомство данного
родителя отклоняется в меньшей степени от среднего
значения для данной популяции, чем родитель

Причины появления «Больших данных»



VELOCITY

скорость поступления

VOLUME

объёмы

VARIETY

разнообразие

VERACITY

достоверность

- 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)

Особенности Big Data

1. Использование ВСЕХ данных, а не случайных выборок

2. Меньшие требования к точности

3. Не ищем причины, а корреляции

4. Важна Датификация

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



Особенности Big Data

Big Data – больше коммерческий и технологический термин

Visa: с помощью Hadoop сокращение времени обработки тестовых записей за 2 года с 1 месяца до 13 минут

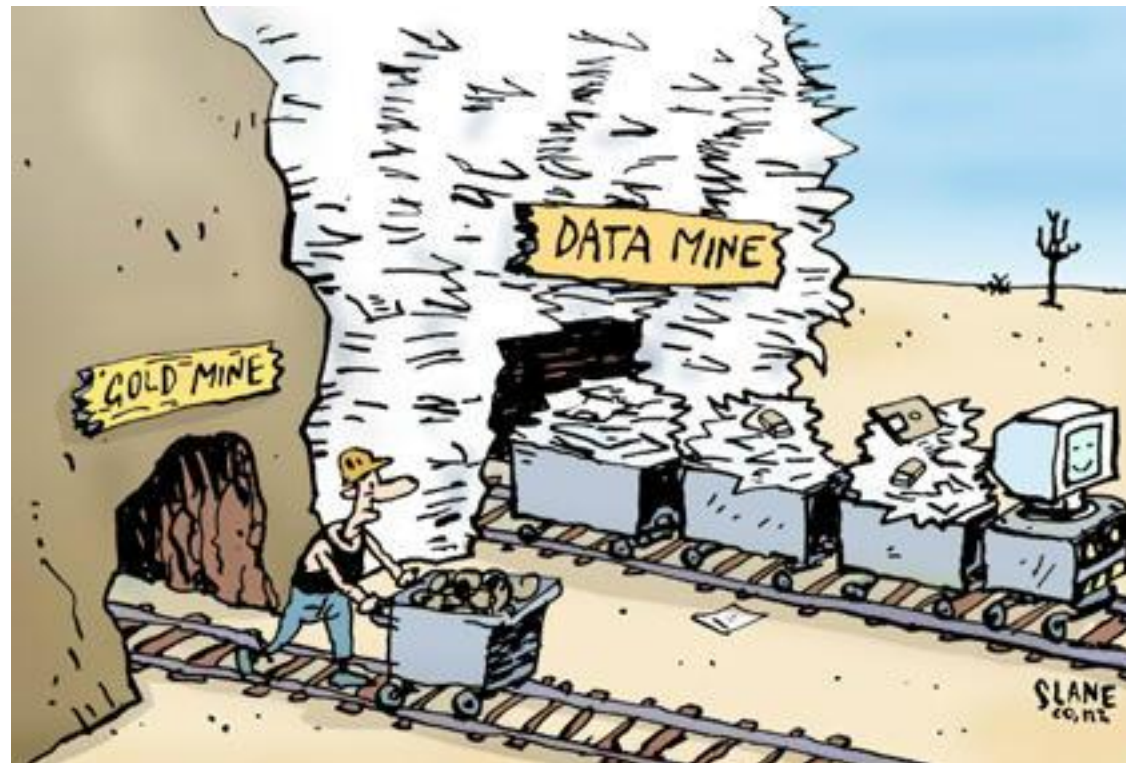
В основе:

- **обновлённая математическая статистика**
 - **анализ данных (Data Mining)**
- **машинное обучение (Machine Learning)**

всё в рамках «науки о данных» (Data Science)

Дальше...

Задачи анализа данных из опыта автора доклада



@Slanecartoons

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



	Площадь	Персонал	Трафик	Остановка	Конкурент	Магазин продукты
Красноармейская, 10	40	4	5000	2	0	1
Просп. Ленина, 10	32	3	4000	3	1	1
Просп. Ленина, 15	30	3	15000	3	1	1
Ул. Л. Чайкиной	35	4	4000	2	2	2

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)

Задача: предсказание визита клиента и суммы покупки

<http://www.kaggle.com/c/dunnhumbychallenge/>

Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. №1 (27) С.68-77.

Предложение дополнительных услуг



Уфа → Москва
Уфа (UFA) Домодедово (DME)

Вылет: 06:55, 7 января 2016
Прилет: 07:15, 7 января 2016
Общее время в пути: 2 ч 20 мин
Эконом-класс

Рейс: S7 Airlines, Q
S7-96
Airbus A319
Эконом-класс

© Указано местное время

Ввод данных о пассажирах

Взрослый

+ Бонусная карта...

пол: М, Ж; фамилия: Латинскими буквами; имя: Латинскими буквами; дата рождения: дд.мм.гггг; гражданство: Россия; паспорт России:

Страхование на время полета Альфа СТРАХОВАНИЕ
Получите выплату до 10 000 рублей при задержке рейса более, чем на 4 часа.
Защитите свой багаж от потери или повреждения на 20 000 руб.
и себя от несчастных случаев на 200 000 руб.

полные условия

Цена на 1 пассажира:
290 руб.

FROM: S7 91 DME TO: S7 91 UFA

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

У меня это есть







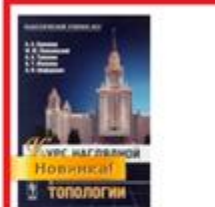
Новинка

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

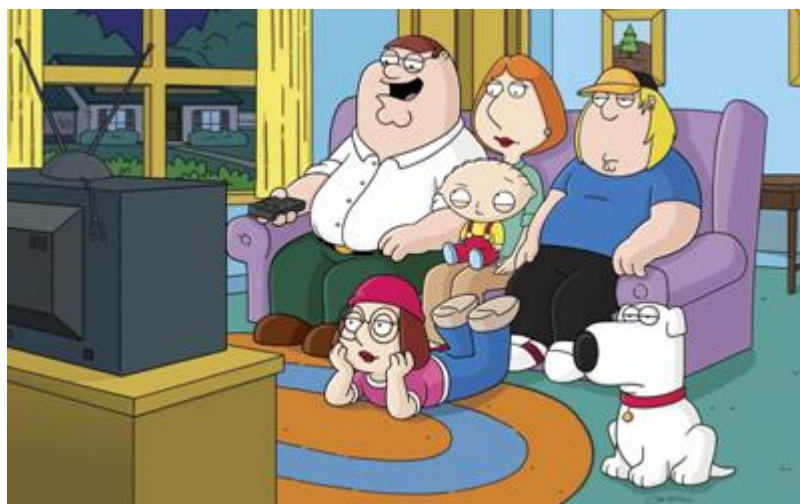
Рекомендуем также

 <p>Книгопечатная продукция (С) Современное проектирование Андрей Александрес 821 Р В корзину</p>	 <p>Настольная игра Набор дорожный 2в1 "Шахматы, шашки". 1 155 Р В корзину</p>	 <p>Новинка!</p> <p>Настольная игра Mask & Zask Магнитная игра Шахматы 358 Р В корзину</p>	 <p>Шахматы Настольная игра Tactic "Шахматы". 40218 1 690 Р В корзину</p>	 <p>Шахматы Игровой набор 3в1 "Игровые": нарды, 2 628 Р В корзину</p>	 <p>Шахматы Уценённый товар. Шахматы "Сенатор". 3 466,80 Р В корзину</p>	 <p>Настольная игра Игровой набор 2в1 "Шахматы, шашки". 1 723 Р В корзину</p>	 <p>Новинка!</p> <p>Книгопечатная продукция (С) Курс наглядной геометрии и топологии 707 Р В корзину</p>
--	---	--	---	--	---	--	--

Рекомендации: статистика + контент

Алгоритмы для рекомендательной системы: технология LENKOR // Бизнес-Информатика, 2012, №1(19), С. 32–39.

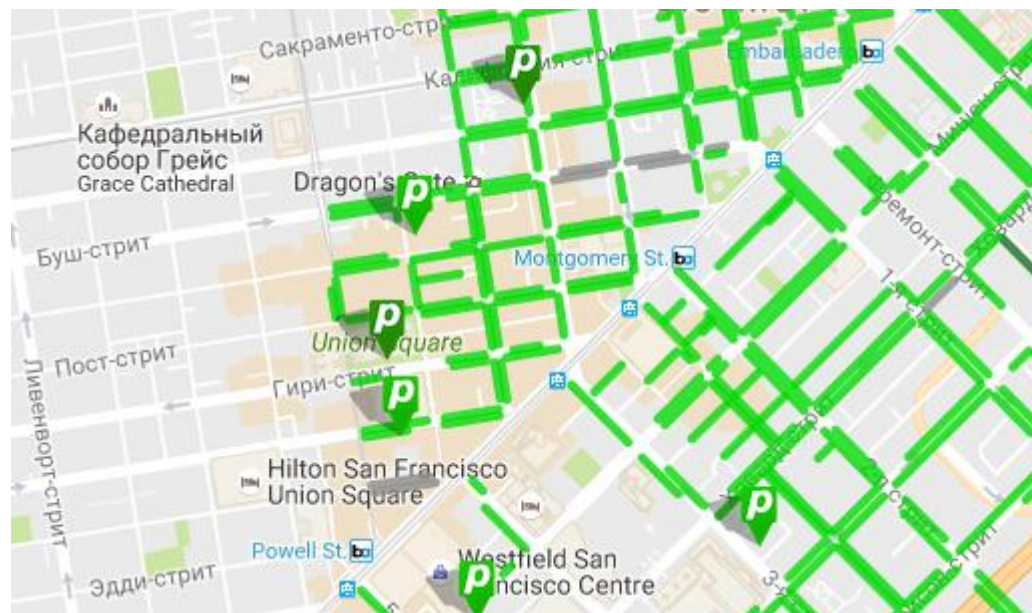
Анализ поведения клиентов



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
	МУЗ ТВ	НТВ	ПЕРВЫЙ КАНАЛ	ПЯТНИЦА	ПЯТЫЙ КАНАЛ	РЕН ТВ	РОССИЯ 1	РОССИЯ 24	РОССИЯ К	СТС

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**

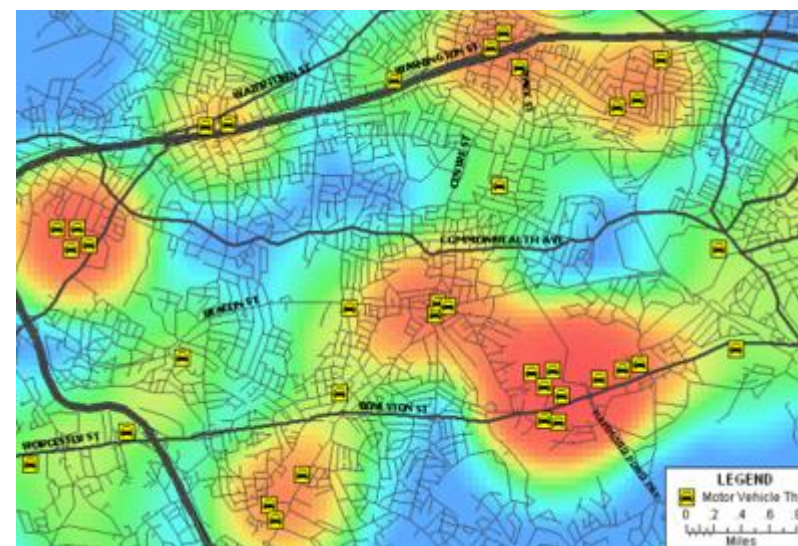


Predictive policing

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**

<https://alexanderdyakonov.files.wordpress.com/2015/07/dyakonovfunnydm.pdf>

Автоматическая диагностика двигателей



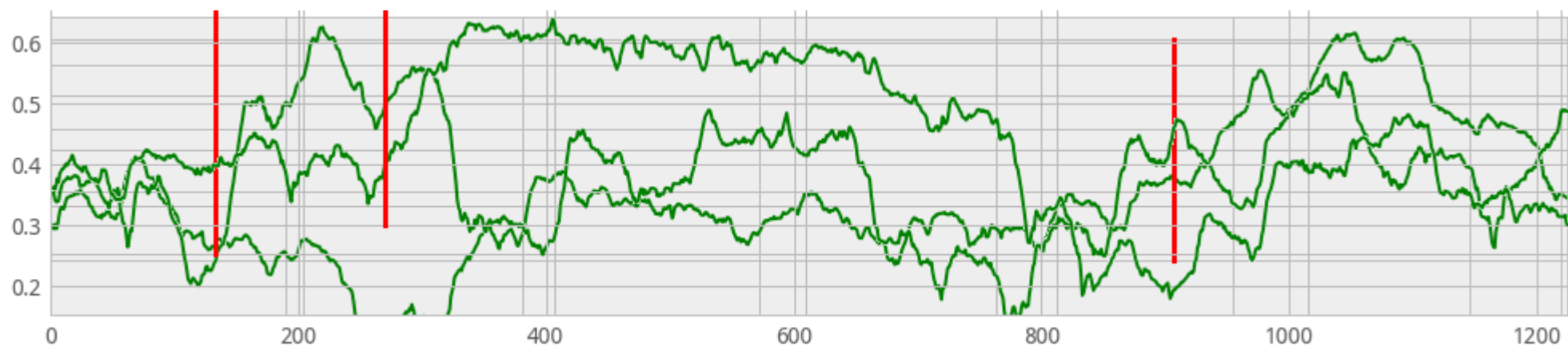
Автоматическая классификация и категоризация

WISE 2014 Challenge: Multi-label Classification of Print Media Articles to Topics // Lecture Notes in Computer Science, том 8787, с. 541-548.

Диагностика неисправностей оборудования



Детектирование поломок
Предсказание поломок
Анализ логов работ



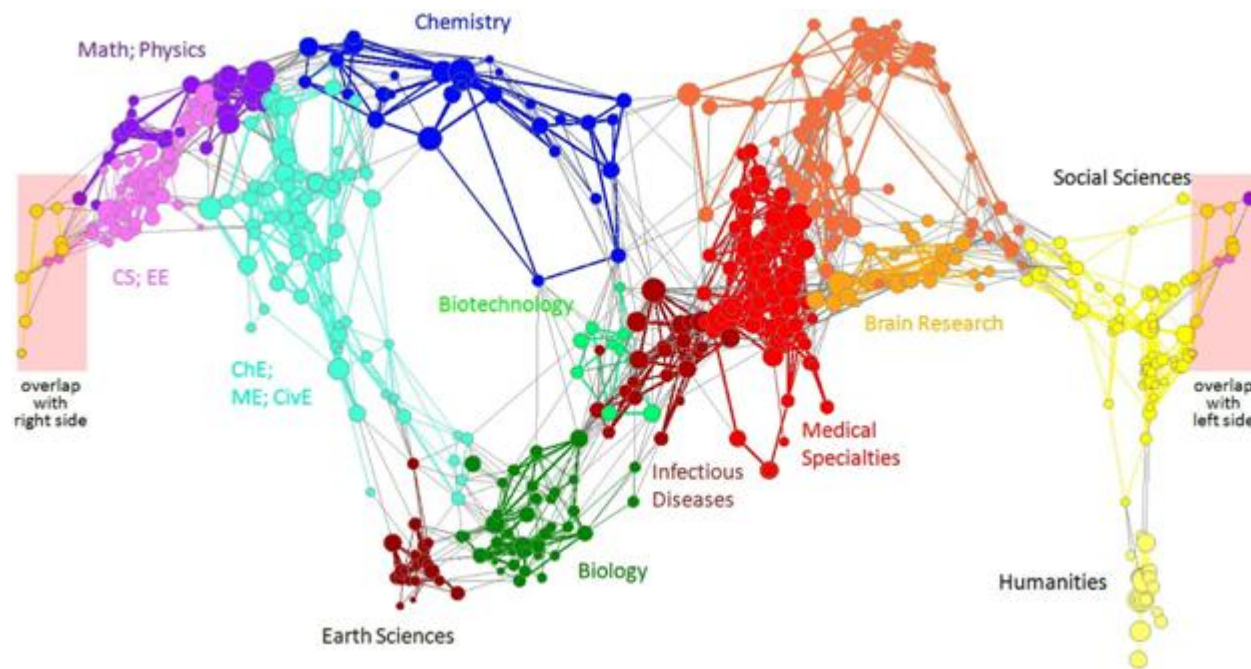
Оценка персонала



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

Методы решения задач классификации с категориальными признаками // Прикладная математика и информатика. Труды факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, № 46, с. 103-127

Анализ социальных сетей



Граф цитирований Börner и др.

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

Прогнозирование связности графа // Математические методы распознавания образов, 2011

<http://alexanderdyakonov.narod.ru/graph-dyakonov-2011.pdf>

Валидация данных



→ да



→ да



→ нет

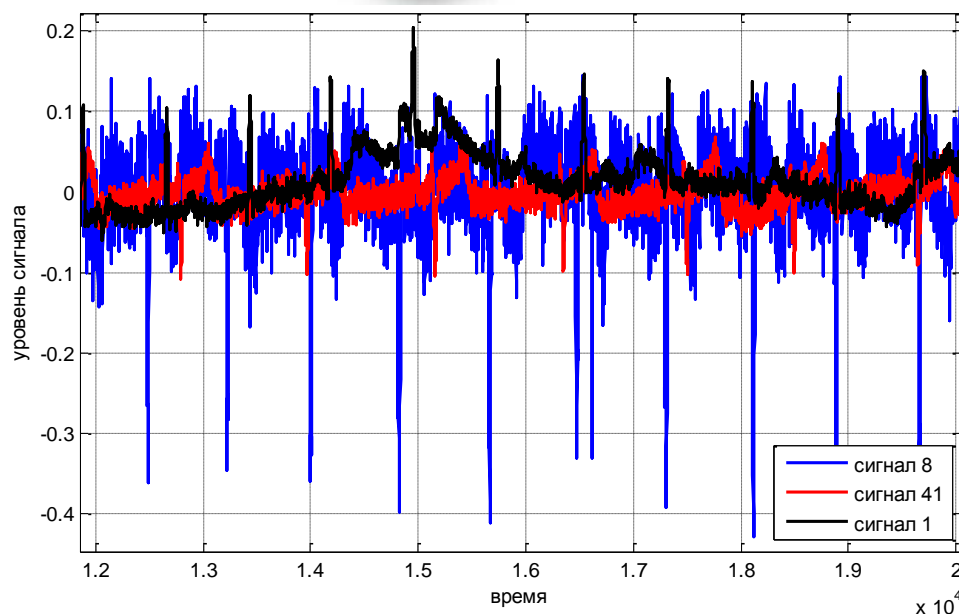
Конкурс Avito:

Есть ли реклама на изображениях, выкладываемых на сайте

Анализ данных в медицине: Brain Computer Interface



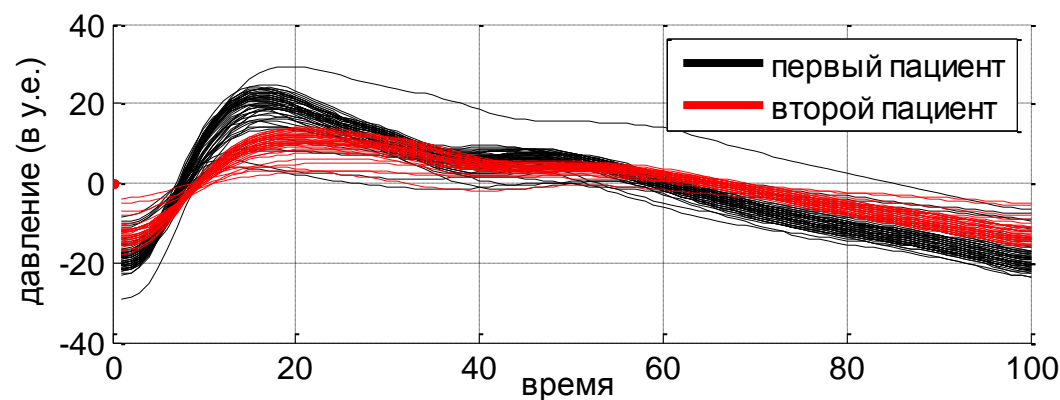
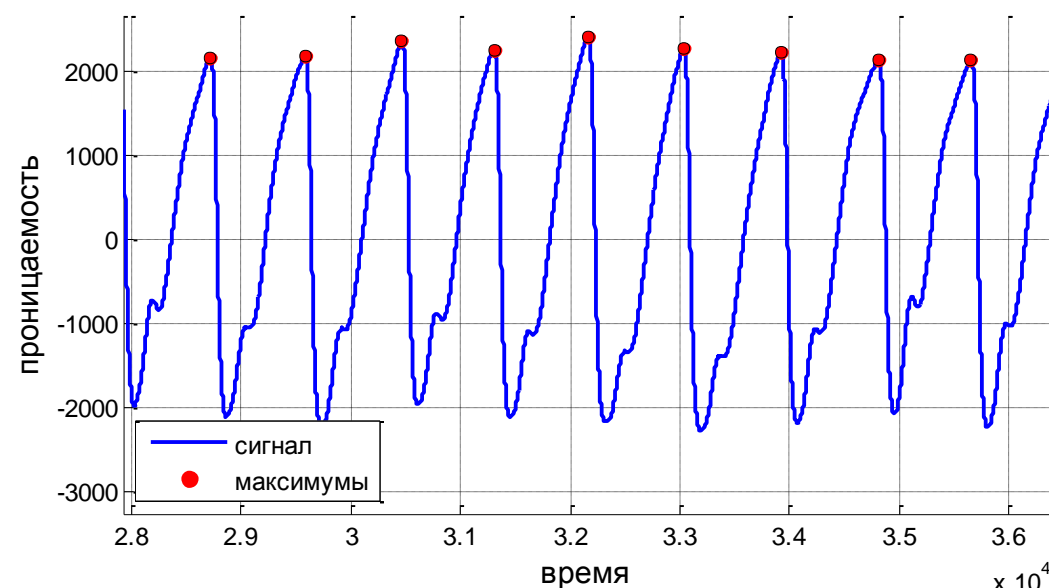
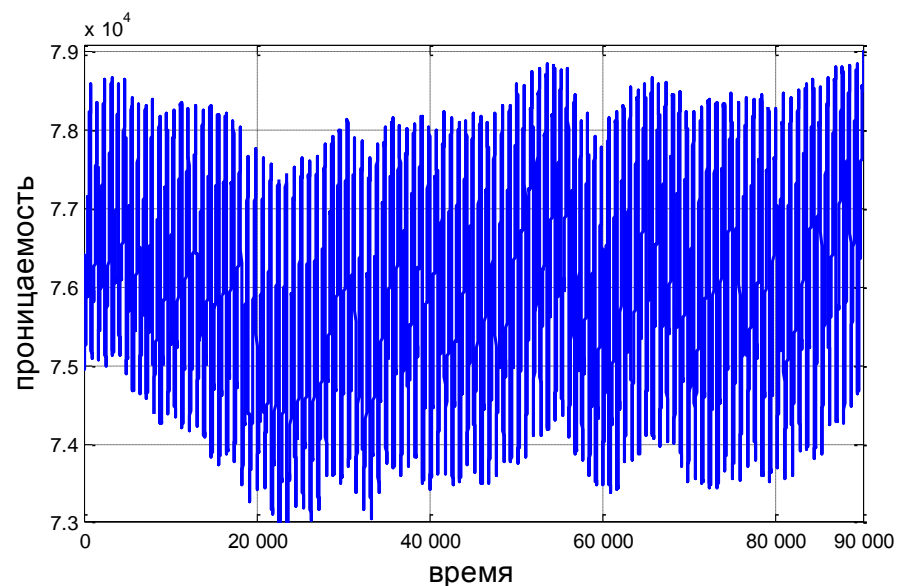
Анализ данных в медицине: Проект CardioQvark



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

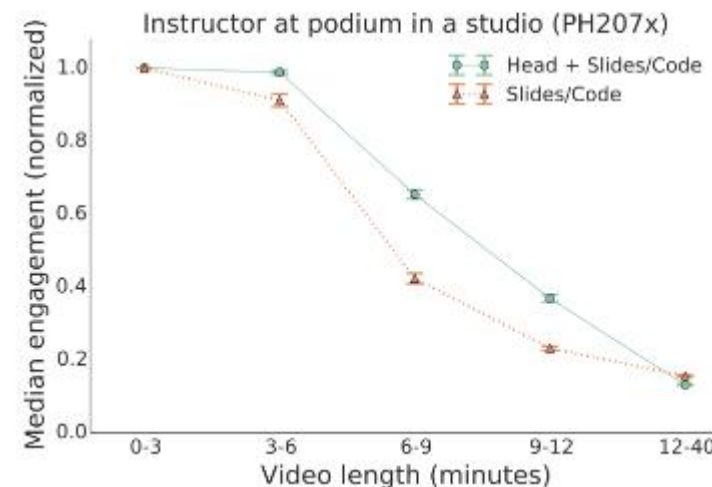
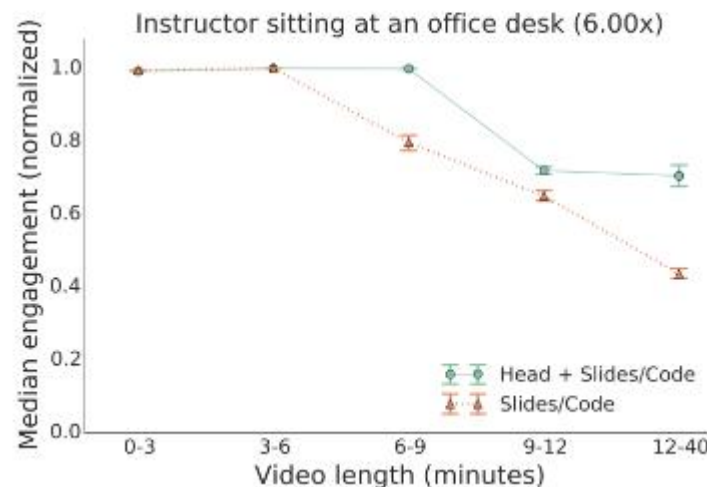
<http://cardioqvark.ru>

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



Анализ схожести пульсовых волн в фотоплетизмограммах // Прикладная математика и информатика, № 53, с. 46-58

Анализ данных в образовании



Philip J. Guo, Juho Kim, Rob Rubin How video production affects student engagement: an empirical study of MOOC videos // L@S '14 Proceedings of the first ACM conference on Learning @ scale conference

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).**

<http://www.kaggle.com/c/WhatDoYouKnow>

Как решаются задачи анализа данных

Инструменты:

- теория вероятностей и математическая статистика
 - машинное обучение
 - программирование

Что такое машинное 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

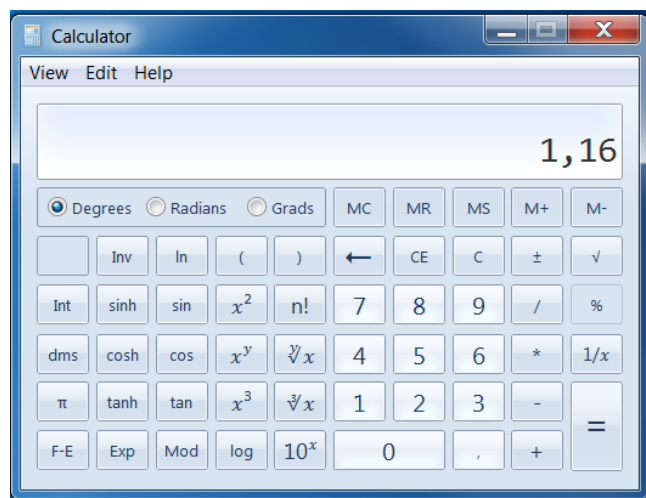
В школе – дают определения

Машинное обучение

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

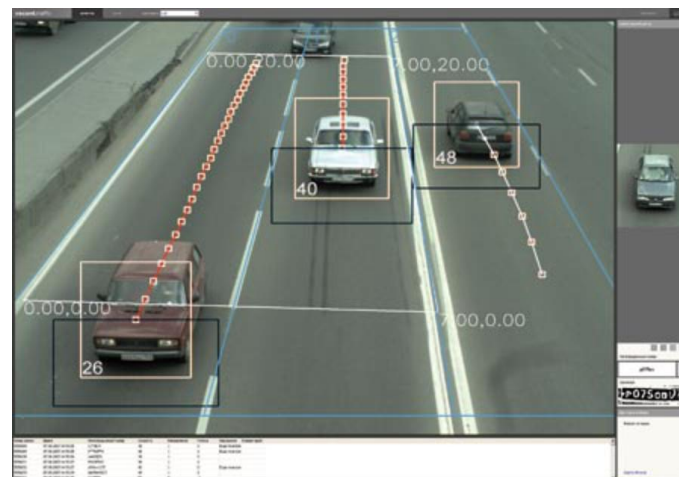
A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

Программирование



**Программируем
последовательность действий**

Обучение



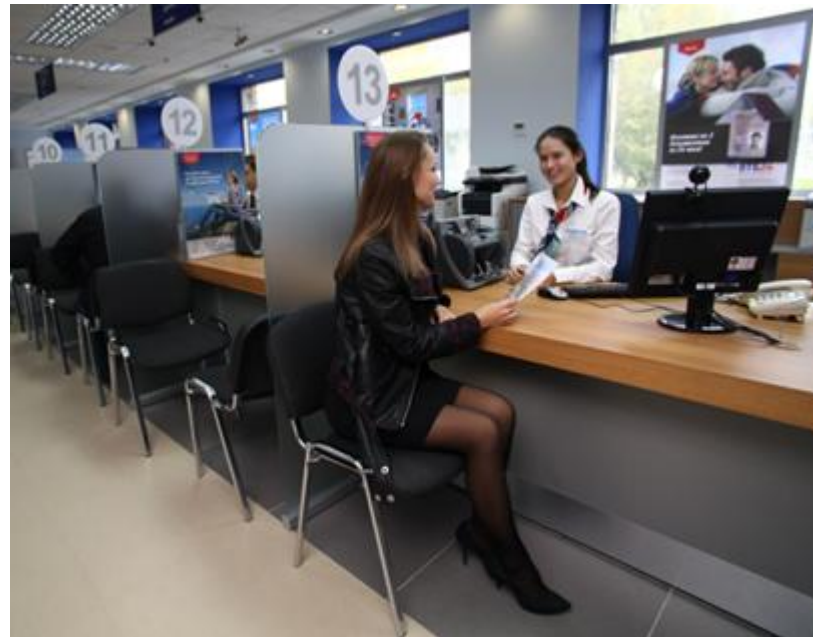
**Программируем алгоритм
анализа информации**

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 помощью описания
(и кредитной истории)**

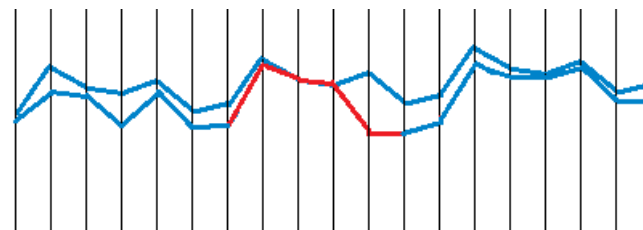
Примеры задачи машинного обучения – классификация / детекция



Классификация / идентификация



Примеры задачи машинного обучения – прогнозирование



Примеры задачи машинного обучения – ранжирование

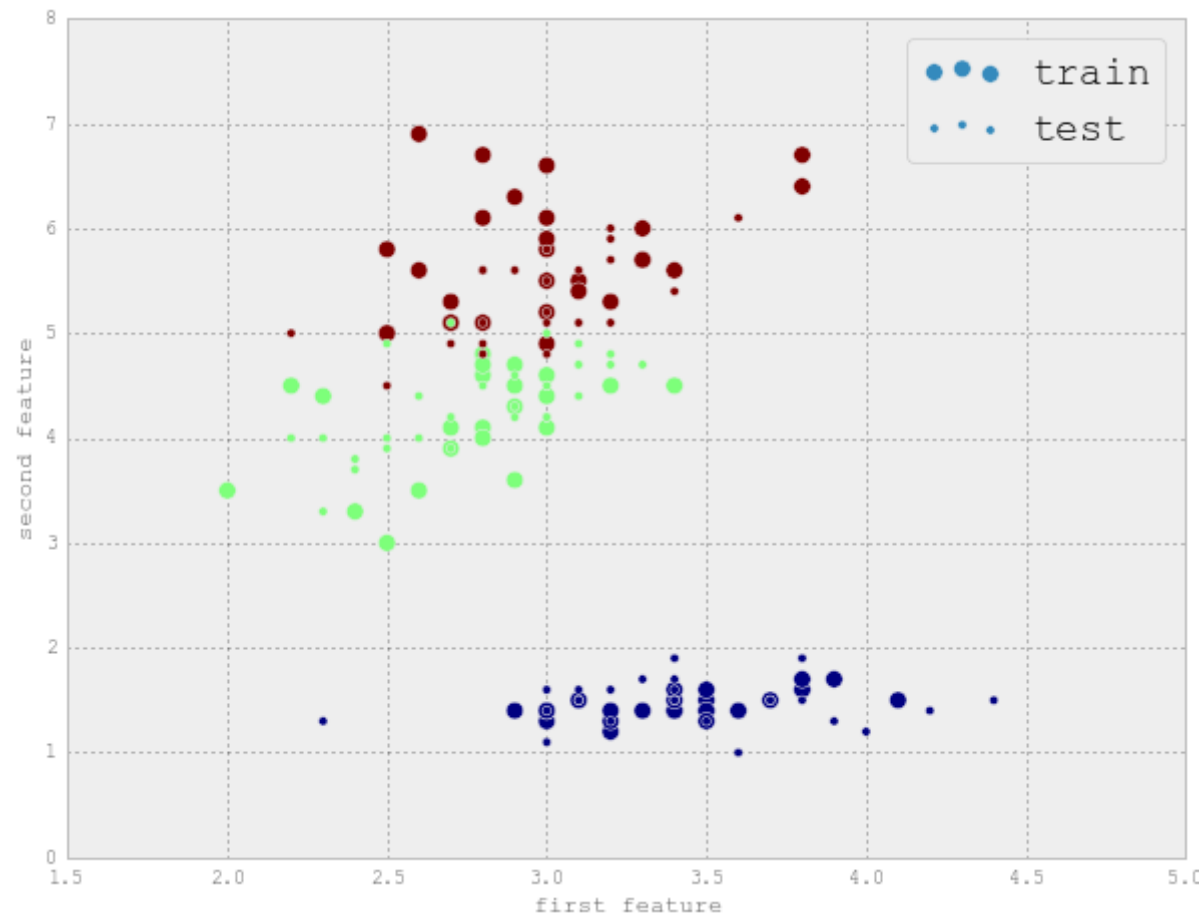
ОЗОН

Интернет магазин OZON.ru **5**
Реклама www.ozon.ru/
Всегда новые скидки и акции. Бесплатная доставка от 3 000 руб.!

OZON.travel - бронирование гостиниц, билетов на ... **5**
www.ozon.travel/
Продажа авиа и ж/д билетов. Бронирование отелей. Страховые полисы для выезжающих за границу.

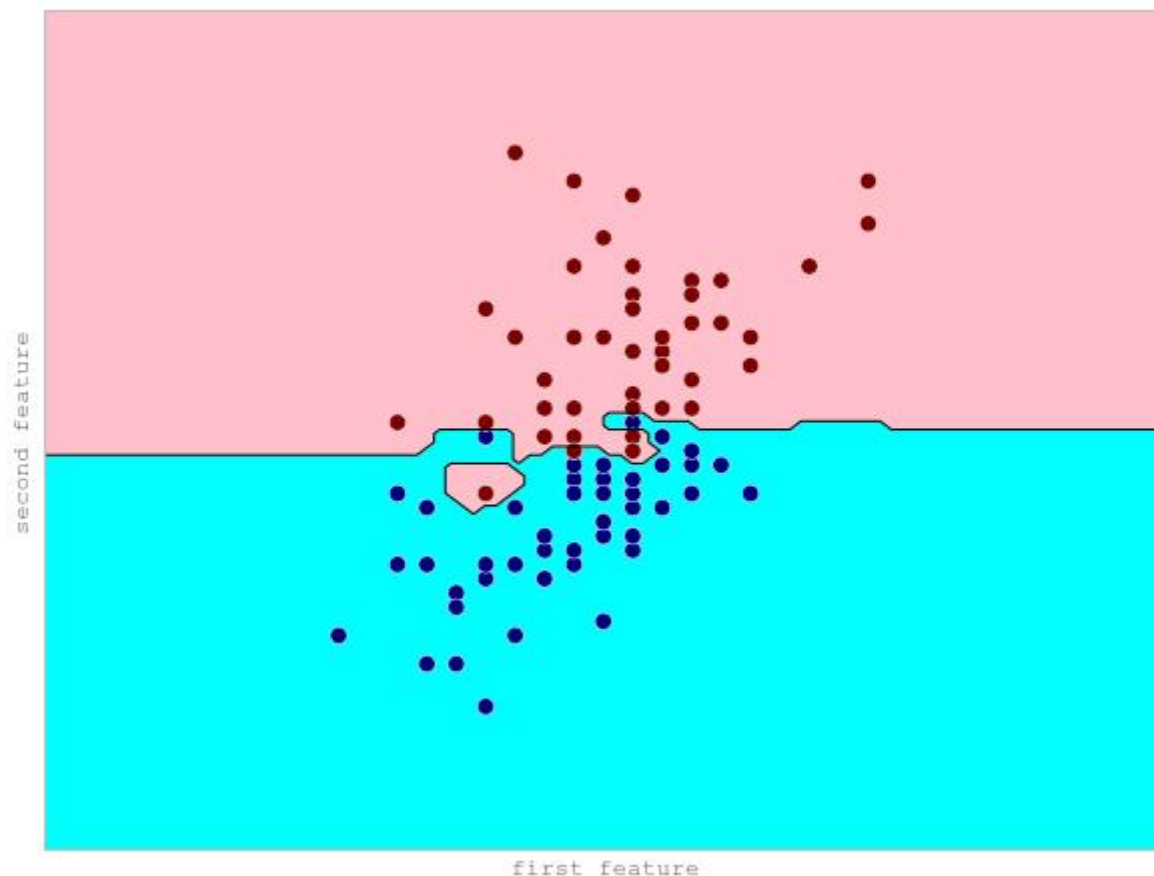
Дешевые авиабилеты онлайн на OZON.travel **4**
www.ozon.travel/flight/
Автоподбор минимальной цены билета — лучшие цены автоматически!
Спецпредложения от авиакомпаний и гостиниц всегда доступны на OZON.travel ...

Как выглядит задача классификации



Как решается? Метод ближайшего соседа

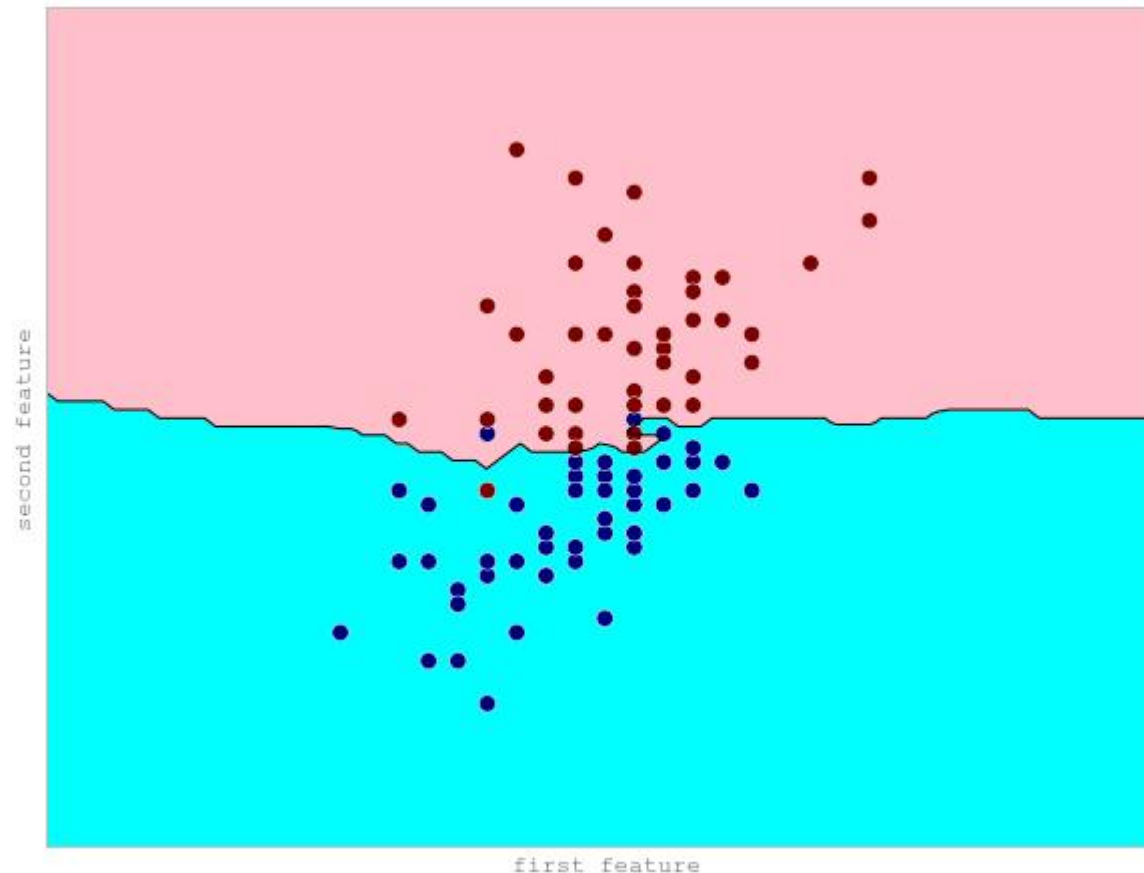
К какому классу ближе – к тому и принадлежит



Переобучение – слишком точная настройка на обучающую выборку, при этом алгоритм показывает плохое качество на контрольной

Метод 3х ближайших соседей

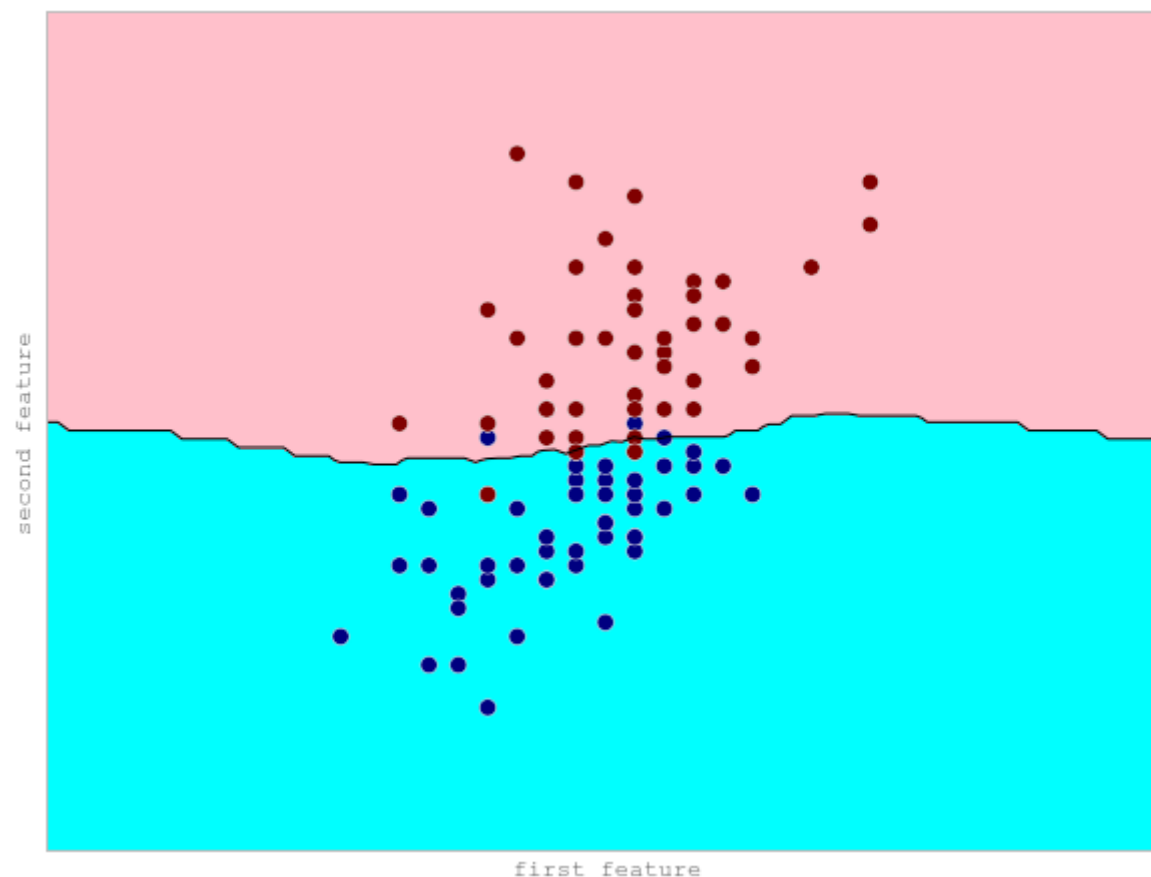
Найти 3 ближайших соседа, по большинству определяем класс



Изменение параметра: k – число соседей
У алгоритма много параметров: например, метрика

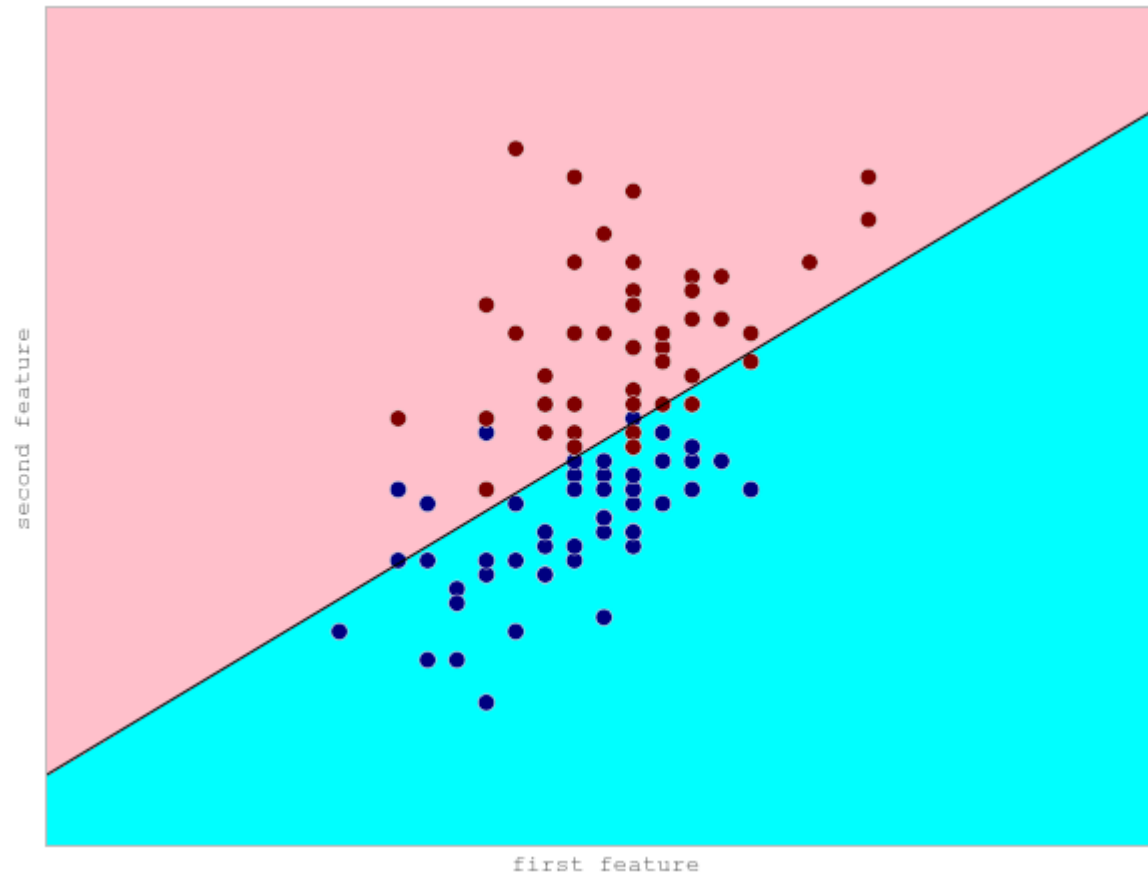
Метод 13ти ближайших соседей

Качество падает, но нет переобучения...



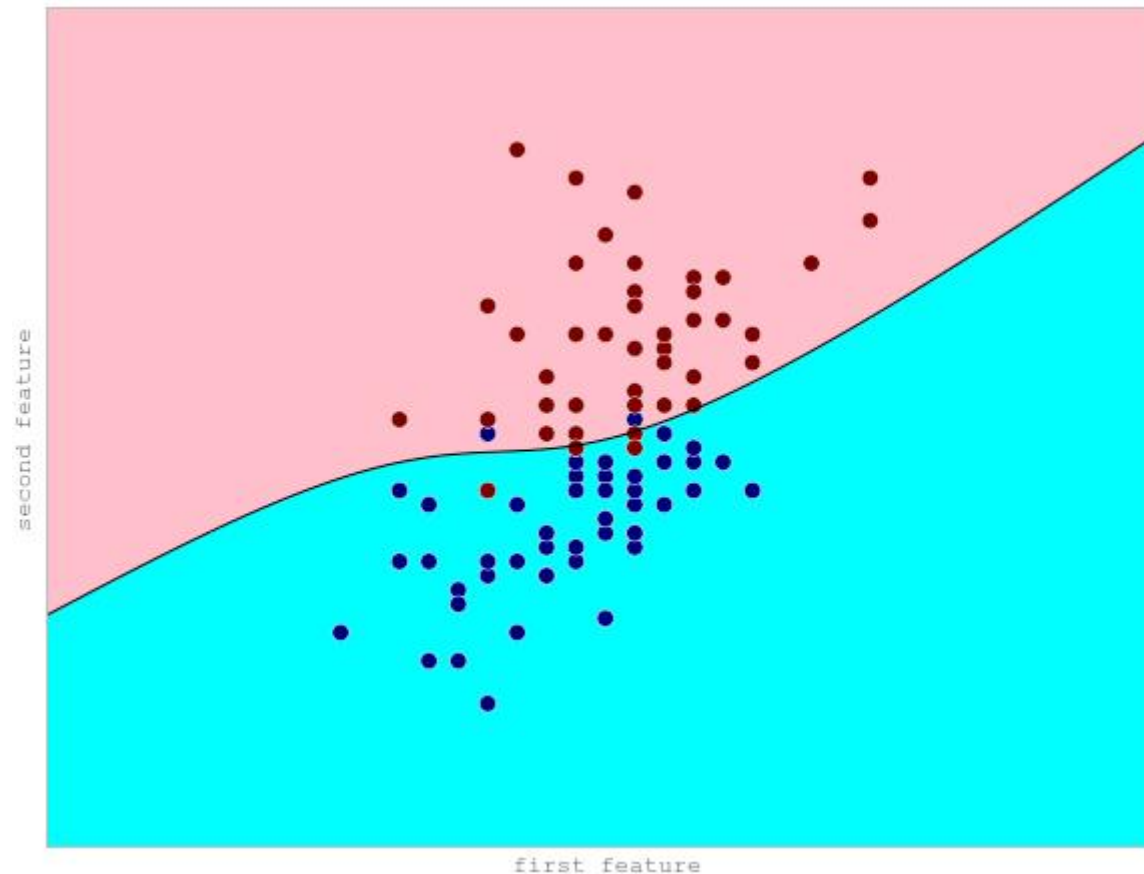
Второй способ – разделение прямой

Ищем прямую, которая разделяет объекты разных классов

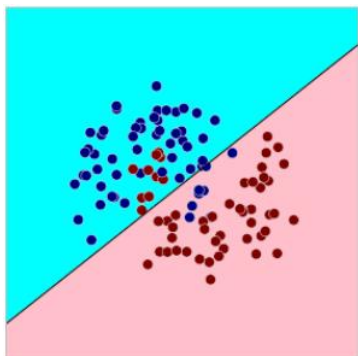


Почему прямой?

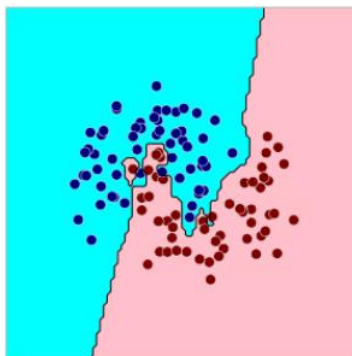
Можно параболой... поверхностью 3го порядка и т.д.



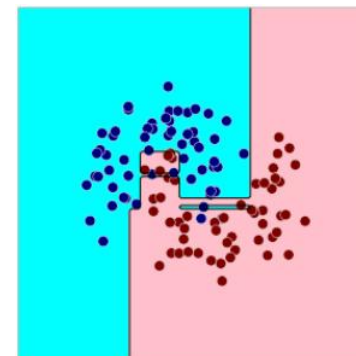
Методов много...



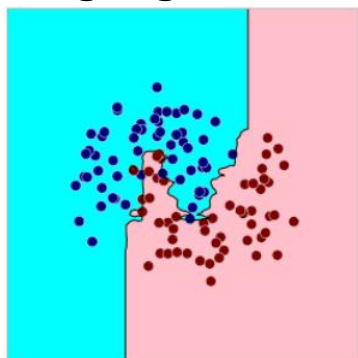
log regression



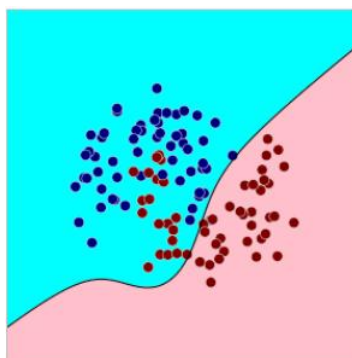
1NN



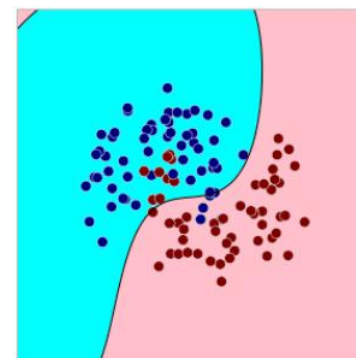
tree



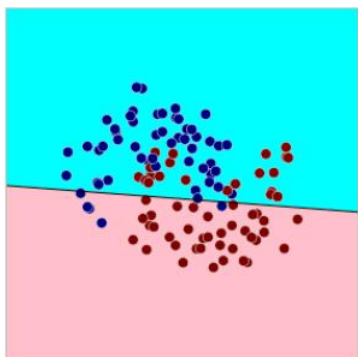
rf



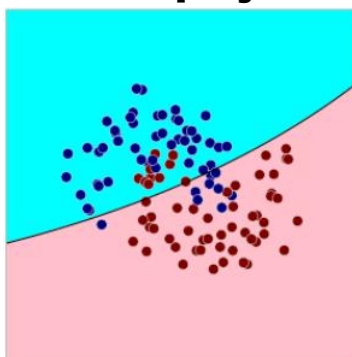
svm poly3



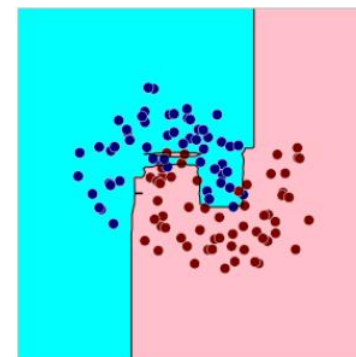
svm rbf



SGD



naive bayes



grad boosting

Как записываются наши решения «в компьютере»...

Линейный метод

$$2.1 * \text{Длина чашелистника} - 0.7 * \text{Длина лепестка} - 0.2 * \text{Ширина лепестка} > 0.3$$

Метод второго порядка

$$0.1 * \text{Длина чашелистника} * \text{Ширина чашелистника} - 0.5 * \text{Длина лепестка}^2 + 0.9 * \text{Ширина лепестка} > 0.3$$

Дерево

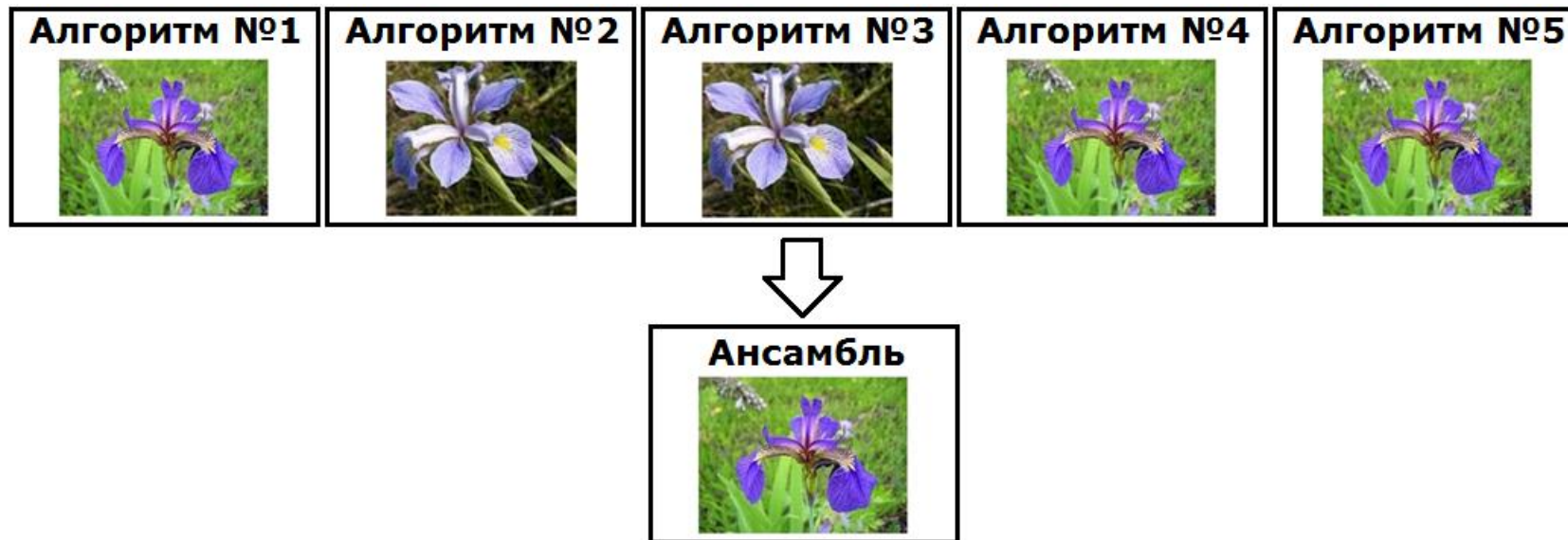


Как записываются наши решения «в компьютере»...

**«Ближайший сосед» не записывается в виде формулы,
но зато интерпретируется**



Что такое сложные алгоритмы?



Голосование – принятие решений по большинству

Алгоритмы могут быть совершенно разные!

Что такое сложные алгоритмы?



Ансамбли могут быть разные!

Бустинг – построение ансамбля,
в котором каждый следующий алгоритм исправляет ошибки
предыдущих.

Основные виды машинного обучения

Обучение с учителем (supervised learning)

классификация

регрессия / прогнозирование

порядковая регрессия

Обучение без учителя (unsupervised learning)

кластеризация

поиск аномалий

уменьшение размерности/описание данных

Рекомендации (Recommender Systems)

Обучение с подкреплением (reinforcement learning)

Пример: что надо знать

Рекомендации (персональные по статистике)

товары							
пользователи							
				5	4	2	
		3				2	
			1	3			5
			-1	1			

Рекомендации (неперсональные по контенту)



Samsung EF-PG360
Protective Cover чехол
для Galaxy Core Prime
Red



Luxcase защитная
пленка для Samsung
Galaxy Core Prime
антибликовая



Математика

Матричные разложения
Методы оптимизации

Программирование

Парсинг
Регулярные выражения
Анализ текстов

Где учат

наш курс «Введение в машинное обучение»

кафедра ММП ВМК МГУ



<http://mmp.cs.msu.ru/>