Анализ данных - знакомство с одной из самых сексапильных IT-специальностей 21 века

© Kleiner Igor M.Sc. 2015
Школа обработки и анализа данных

Добро пожаловать

http://geekbrains.ru/events/118

Добро пожаловать

http://geekbrains.ru/events/118

Будем знакомы

Будем знакомы

Учеба		

Будем знакомы

Учеба		

Будем знакомы

Учеба		
Работа		

Будем знакомы

Учеба	МИРЭА, МИСиС, University of Haifa	CS, Ψ, Statistic
Работа	Microsoft, intel, skype, HUAWEI, University of Haifa	
Специализация	Image processing Big data \ machine learning Stochastic optimization	Psychology of perception

Цель нашей встречи

Цель нашей встречи

- Где я?
- Кто здесь?

- А что вы тут делаете?

Цель нашей встречи

- Где я?
- Кто здесь?
- А что вы тут делаете?

Цель нашей встречи

- Познакомиться с вселенной анализа данных

Цель нашей встречи

- Познакомиться с вселенной анализа данных в интересной и доступной форме

Цель нашей встречи

- Познакомиться с вселенной анализа данных
  в интересной и доступной форме

- Узнать как самостоятельно продолжить обучение в области работы с данными и их анализом

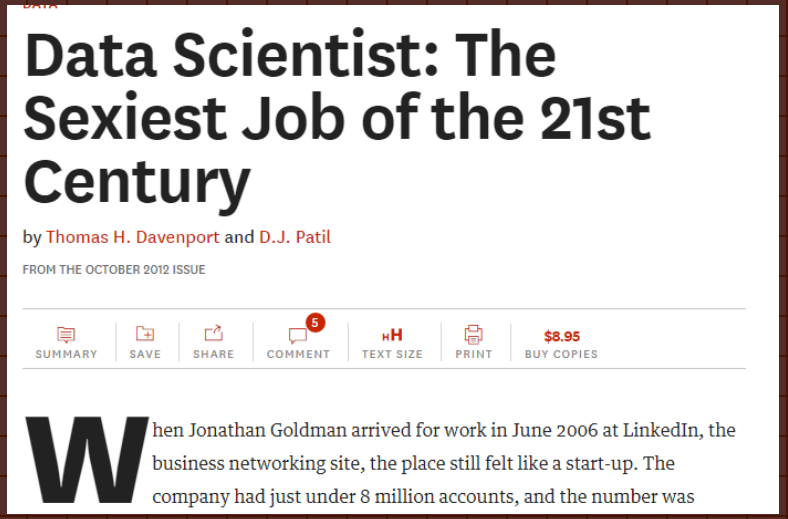
Цель нашей встречи

- Познакомиться с вселенной анализа данных
  в интересной и доступной форме
- Узнать как самостоятельно продолжить обучение в области работы с данными и их анализом
- Получить удовольствие и хорошо провести время

ГОТОВЫ?

Анализ данных – сексапильная специальность?

- hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT 5 | TEXT SIZE | PRINT | $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was

ARTICLE | HARVARD BUSINESS REVIEW | OCTOBER 2012

Data Scientist: The Sexiest Job of the 21st Century

The Sexiest Job of the 21st Century: Data Analyst

Chris Morris, Special to CNBC.com

Data Scientist
B.P.G.S Global Ltd
Jerusalem, IL • Sep 10, 2015 • From www.drushim.co.il
Similar
View

Data Scientist
Intel
Sep 16, 2015 • From intel.taleo.net
Similar
View

Data Scientist
Check Point Software Technologies, Ltd.
Jerusalem, IL • Sep 25, 2015 • From www.drushim.co.il
Similar
View

Data Scientist
Check Point Software Technologies, Ltd.
Tel-Aviv, Israel • Sep 16, 2015
24 connections to the poster • Similar
View

Data Scientist
Check Point Software Technologies, Ltd.
Tel-Aviv, Israel • Sep 7, 2015
24 connections to the poster • Similar
View

Data Scientist (Job 1066)
Viber
Israel • Aug 30, 2015
13 connections to the poster • Similar
View

дата дата дата

- Более чем 7.9 зетабайт электронной информации существует в мире сегодня


x 1024
x 1024
x 1024
Zettabyte
Exabyte
Petabyte
Terabyte

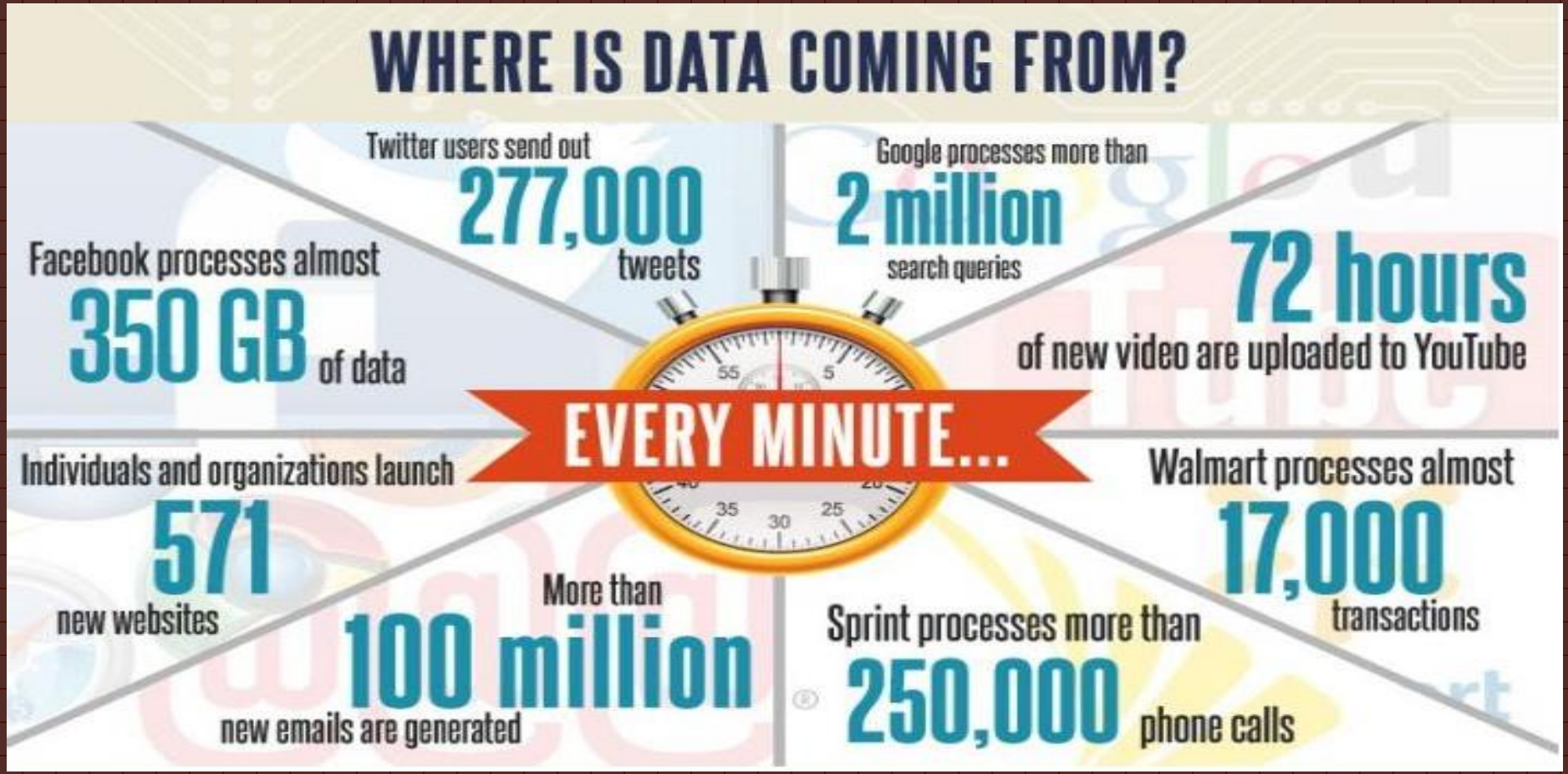
Данные Данные Данные

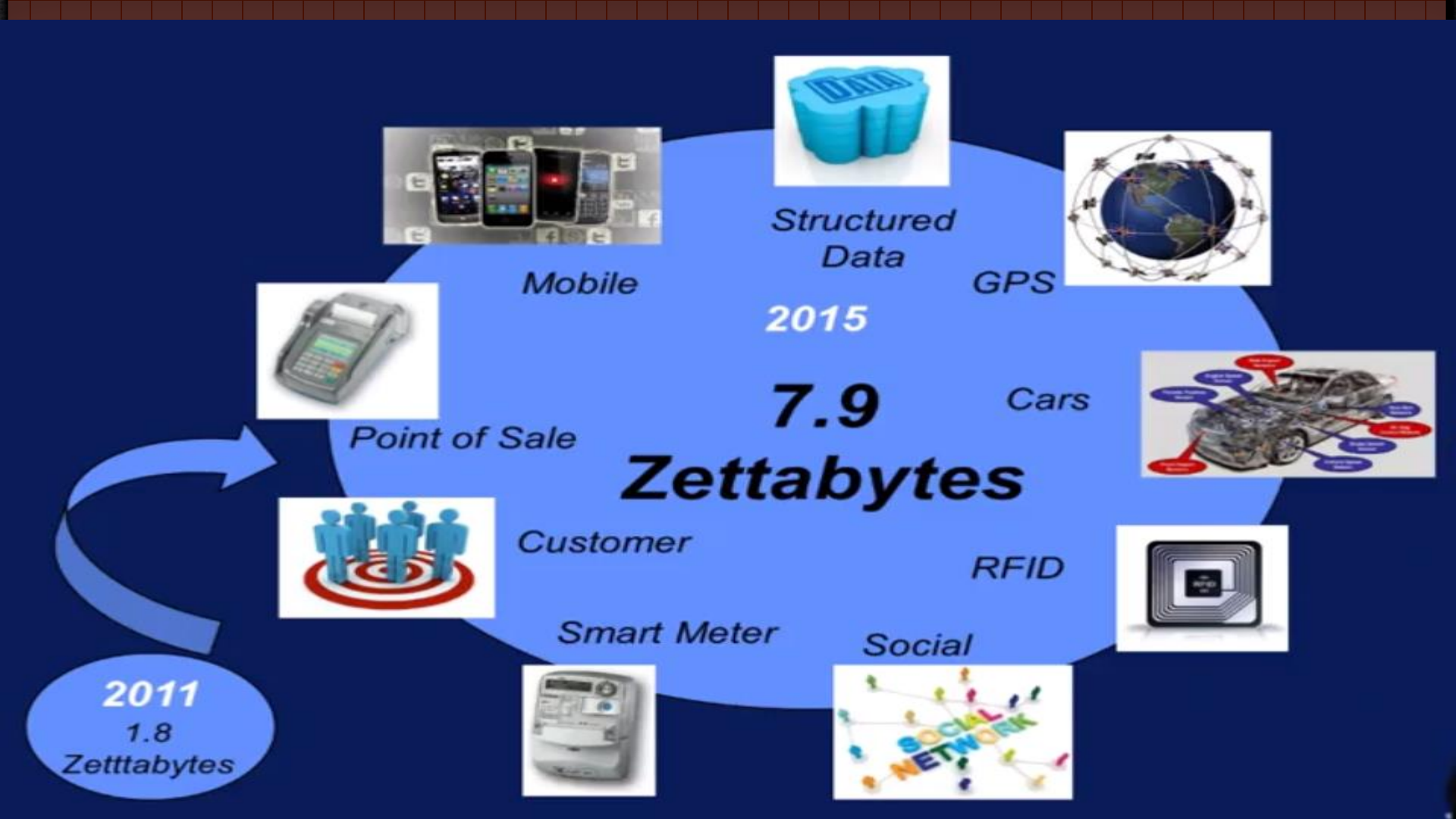
- Более чем 7.9 зетабайт электронной информации существует в мире сегодня
- 7 900 000 000 000 000 000 000 байт – это количество примерно эквивалентно информации, содержащейся в более чем 600 миллиардов фильмов в HD качестве

Данные Данные Данные

- Более чем 7.9 зетабайт электронной информации существует в мире сегодня
- 7 900 000 000 000 000 000 000 байт – это количество примерно эквивалентно информации, содержащейся в более чем 600 миллиардов фильмов в HD качестве
- Новые данные появляются с экспоненциальной скоростью

WHERE IS DATA COMING FROM?




Structured Data
Mobile
GPS
2015
7.9
Zettabytes
Point of Sale
Cars
Customer
RFID
Smart Meter
Social
2011
1.8
Zetttabytes

Данные и их анализ очень полезны

- Медицина
- Спорт
- Финансы
- Корпорации
- Государственные учреждения
- СМИ
- …

Что такое анализ данных?

- Наука (или искусство) об использовании данных, с целью строить модели, которые позволяют принимать лучшие решения и приносят пользу

Что такое анализ данных?

- Наука (или искусство) об использовании данных, для того чтобы строить модели, которые позволяют принимать лучшие решения и приносят пользу

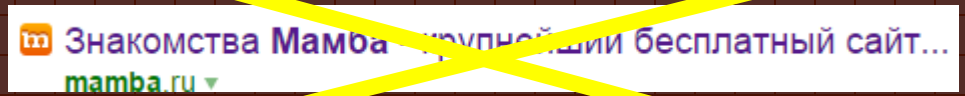
“Science is what we understand well enough to explain to a computer. Art is everything else we do”

Примеры успешного применения анализа данных

- eHarmony
- The Framingham Heart Study - фремингемское исследование сердца
- Выбор игроков в команду

eHarmony

- Сайт знакомств:
  - *модус операнди*: создание пар для долгосрочных отношений
  - *идея*: научный подход для поиска подходящих кандидатур
  - нет поиска по анкетам



eHarmony

- Сайт знакомств:
  - *модус операнди*: создание пар для долгосрочных отношений
  - *идея*: научный подход, для поиска подходящих кандидатур
  - нет поиска по анкетам

- Общая прибыль сайта превысила 1 миллиард долларов
- Около 4% браков в США это результат eHarmony

www.eharmony.com/press-release/31/

eHarmony

- Сайт знакомств:

  пользователь при регистрации заполняю длинную анкету

eHarmony

- Сайт знакомств:

  пользователь при регистрации заполняю длинную анкету

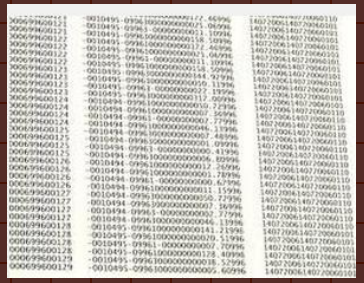
- проанализировав данные, сайт выдает подходящие анкеты пользователей, проживающих рядом

eHarmony

- Магия анализа данных

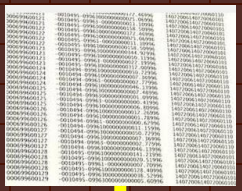
eHarmony

- Магия анализа данных



eHarmony

- Магия анализа данных

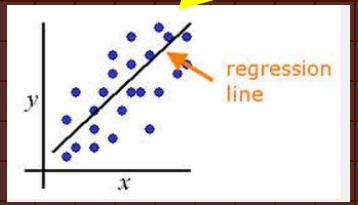


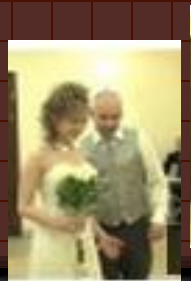
eHarmony

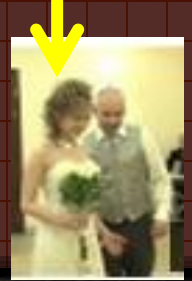
- Магия анализа данных

eHarmony

- Инструменты

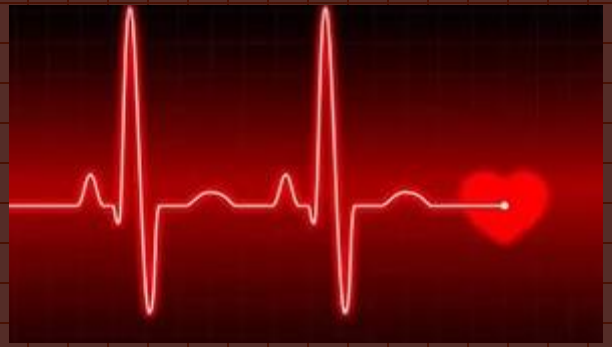






The Framingham Heart Study

- Исследование длиться более 65 лет и является одним из самых продолжительных эпидемиологических исследований в истории медицины (фремингемское исследование сердца)



фремингемское исследование сердца

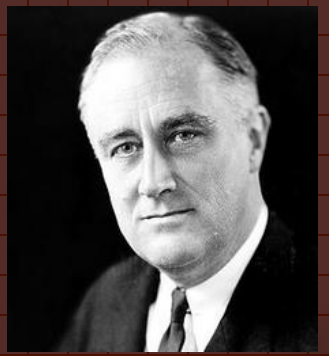
- Франклин Делано Рузвельт президент США 1933-1945
- Умер во время исполнения своих обязанностей в 1945

фремингемское исследование сердца

- Франклин Делано Рузвельт президент США 1933-1945
- Умер во время исполнения своих обязанностей в 1945
- Давление до 1933 года 140/100 - сегодня считается высоким давлением
- Давление за год до смерти 210/120 - сегодня считается опасным кризисом
- 260/150 давление за два месяца до смерти
- 300/190 в день смерти

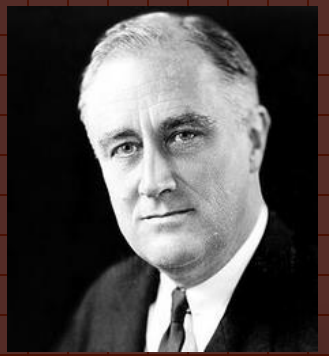
фремингемское исследование сердца

- Сегодня мы знаем об опасности высокого давления
- Откуда сегодня врачам известна эта информация?



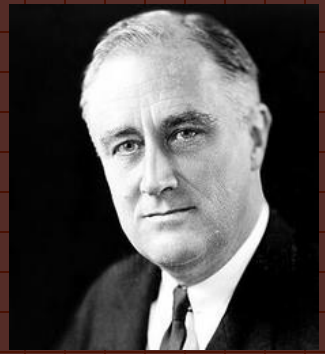
фремингемское исследование сердца

- Сегодня мы знаем об опасности высокого давления
- Откуда сегодня врачам известна эта информация?



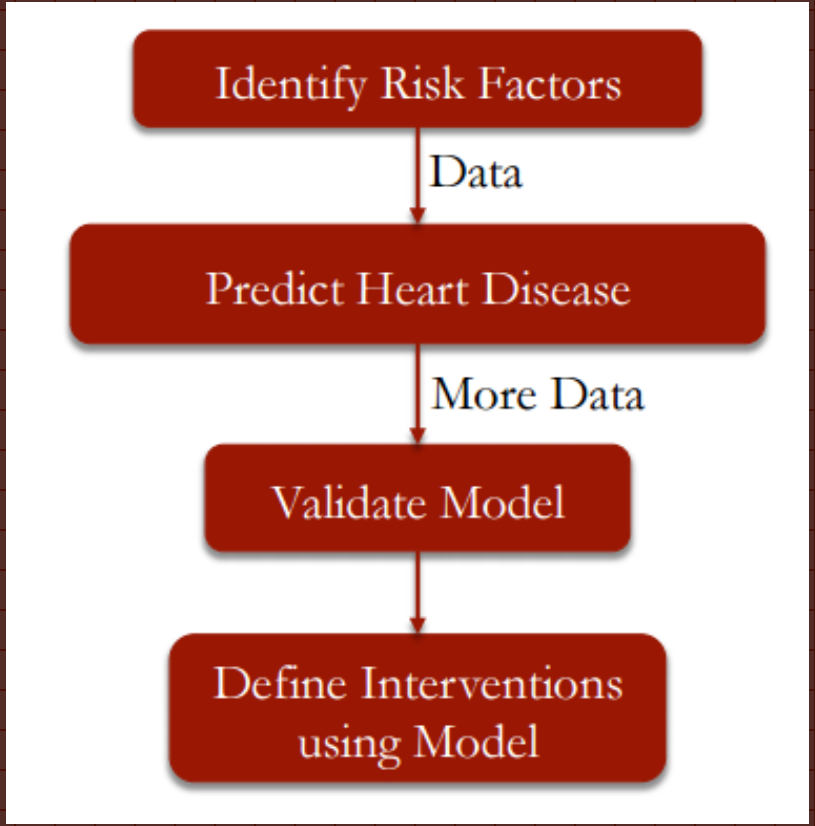
фремингемское исследование сердца

- 1948 год, город Фремингем
- 5209 участников
- участие в наблюдениях и тестах в течении длительного времени



- **Цель**: выявление факторов риска для болезней сердца

фремингемское исследование сердца

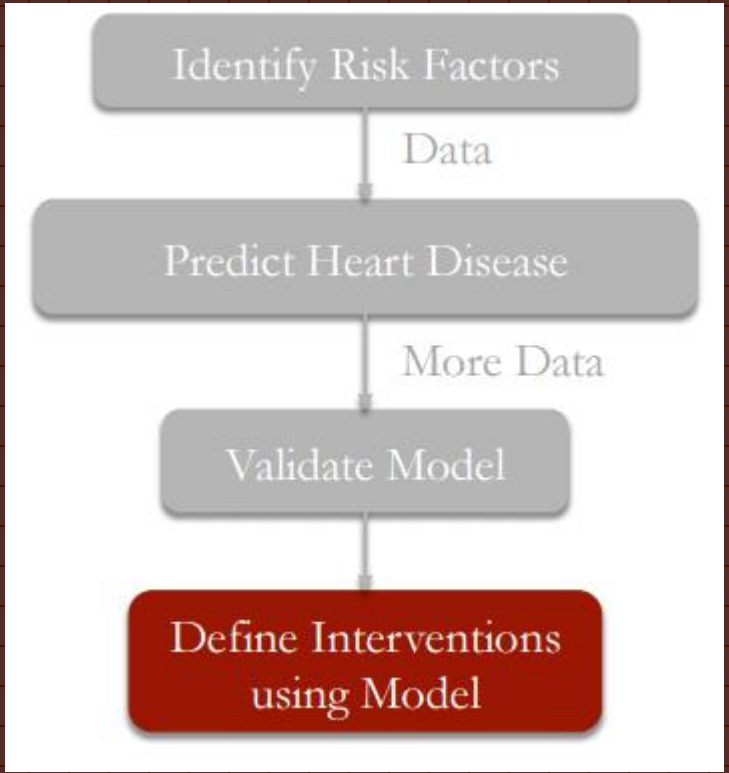


фремингемское исследование сердца

- Благодаря полученным данным и последующим исследованиям, учеными были обнаружены различный факторы риска:

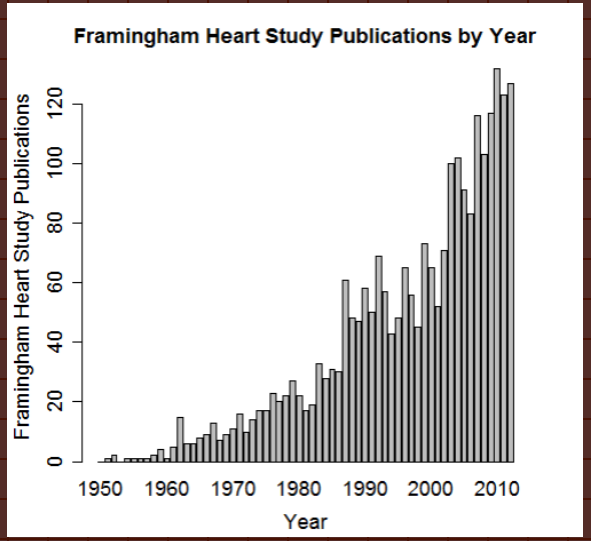
  курение
  уровень холестерина
  давление
  уровень сахара в крови
  …

фремингемское исследование сердца



фремингемское исследование сердца

- Более 2400 исследований на основе полученных данных
- Выявление множества факторов риска

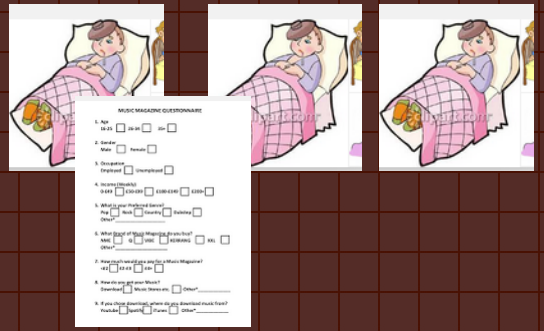


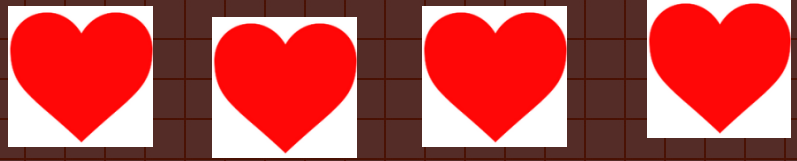
фремингемское исследование сердца

- Магия анализа данных

фремингемское исследование сердца

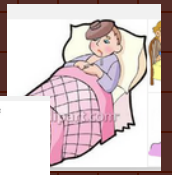
- Магия анализа данных

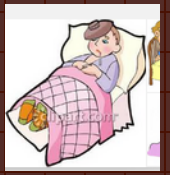




фремингемское исследование сердца

- Магия анализа данных





логистическая регрессия

Выбор лучшего игрока в команду

moneyball

Человек, который изменил всё

Moneyball

год	2011
страна	США
слоган	«А что в жизни сделал ты?»
режиссер	Беннетт Миллер
сценарий	Стивен Зеллиан, Аарон Соркин, Стэн Червин, ...
продюсер	Майкл Де Лука, Рэйчел Хоровиц, Брэд Питт, ...
оператор	Уолли Пфистер
композитор	Майкл Дэнна
художник	Джесс Гончор, Брэд Рикер, Дэвид Скотт, ...
монтаж	Кристофер Теллефсен
жанр	драма, биография, спорт, ... слова

В главных ролях:

Брэд Питт
Джона Хилл
Филип Сеймур Хоффман
Крис Пратт
Стефен Бишоп
Кэррис Дорси
Робин Райт
Рид Даймонд
Брент Дженнингс
Кен Медлок
...

Роли дублировали:

Всеволод Кузнецов
Диомид Виноградов
Александр Новиков

Бейсбол для чайников

- https://goo.gl/FS7uPg

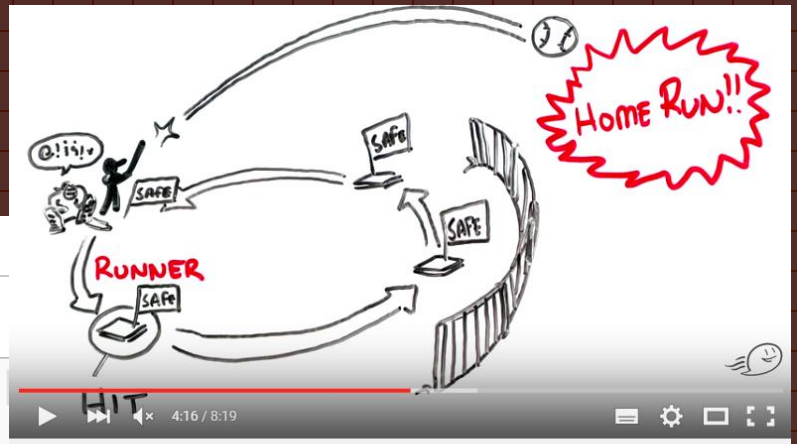
Бейсбол

Материал из Википедии — свободной энциклопедии

Бейсбо́л (англ. *baseball*, от *base* — «база, основание» и *ball* — «мяч») — командная спортивная игра с бейсбольным мячом и битой. В состязаниях участвуют две команды по девять (иногда десять) игроков каждая.

Бейсбол наиболее популярен на Кубе, в США, в Венесуэле, в Японии, Китае и Южной Корее. В США, Японии, Чехии и других странах распространён также софтбол — упрощённый вариант бейсбола — игра, которую можно проводить в помещении и на небольших полях. На данный момент в бейсбол играют более чем в 120 странах мира.

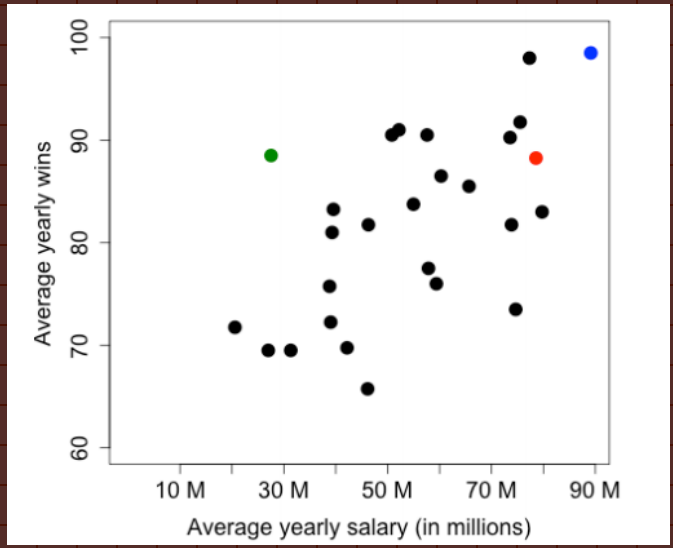
К родственным бейсболу видам спорта относятся крикет, песаполо в Финляндии, ойна в Румынии и лапта в России.



Baseball Rules Whiteboard Video Rules of Baseball

Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков



Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков

- Оклендская бедная команда после прихода нового менеджера стала показывать хорошие результаты

Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков

- Оклендская бедная команда после прихода нового менеджера стала показывать хорошие результаты
- Что произошло?

Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков

- Оклендская бедная команда после прихода нового менеджера стала показывать хорошие результаты
- Что произошло?

Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков

- Обработав множество параметров игроков, программа выявила тех игроков, которые были недооценены, т.е. качество игры которых было меньше заработка игроков их уровня

Выбор лучшего игрока в команду

- У богатых команд больше денег и они могут позволить купить лучших игроков
- Обработав множество параметров игроков, программа выявила тех игроков, которые были недооценены, т.е. качество игры которых было меньше заработка игроков их уровня
- Сегодня в любой команде высшей лиги есть свой статистик

Выбор лучшего игрока в команду

- Магия анализа данных




логистическая регрессия




Еще примеры

- предсказание решений высшего суда
- предсказание будущей цены вина
- предсказания цен на авиабилеты
- предсказания эпидемии гриппа на основе поисковых запросов
- …
- …

Еще примеры

- предсказание решений высшего суда
- предсказание будущей цены вина
- предсказания цен на авиабилеты
- предсказания эпидемии гриппа на основе поисковых запросов
- …
- …

предсказания эпидемии гриппа на основе поисковых запросов

Google предупредит о вспышке эпидемии гриппа

13.11.2008 12:31 Компания Google запустила новый интернет-сервис Google Flu Trends, который будет отслеживать вспышки эпидемии гриппа на территории Соединенных Штатов. Поисковик будет регистрировать все пользовательские запросы, связанные с гриппом, в частности вопро ...

По материалам: Радио Свобода: новости

предсказания эпидемии гриппа на основе поисковых запросов

Google предупредит о вспышке эпидемии гриппа

13.11.2008 12:31 Компания Google запустила новый интернет-сервис Google Flu Trends, который будет отслеживать вспышки эпидемии гриппа на территории Соединенных Штатов. Поисковик будет регистрировать все пользовательские запросы, связанные с гриппом, в частности вопро ...

По матери

When Google got flu wrong

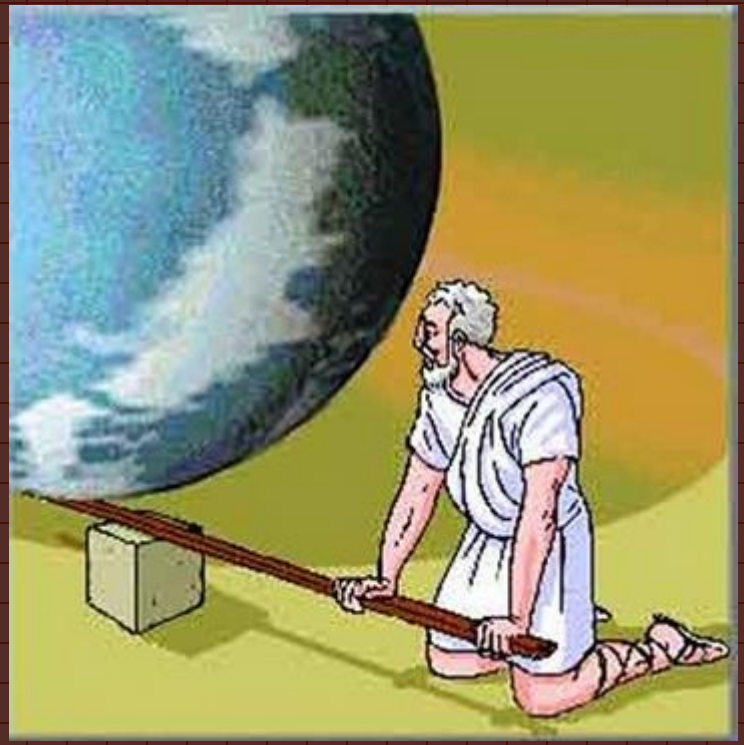
US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

13 February 2013

PDF | Rights & Permissions

Дайте мне точку опоры

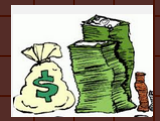


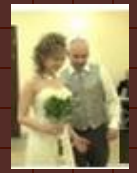
Дайте мне данные

Дайте мне данные, компьютер

Дайте мне данные, компьютер и специалиста по анализу данных

Дайте мне данные, компьютер и специалиста по анализу данных и я изменю жизнь людей





Магия анализа данных

данные

закономерности - предсказания

Магия анализа данных

данные

закономерности - предсказания

Магия анализа данных

данные

закономерности - предсказания

Детали дьявола анализа данных

- Поиск и сбор необходимых данных
- Приведение полученных данных в удобную для анализа форму
- Выбор подходящей модели для анализа данных
- Анализ данных
- Верификация полученных результатов
- Презентация полученных результатов и принятие решений

Детали дьявола анализа данных

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis

- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code
- Distribute results to other people

Краткие итоги

- Анализ данных это:

Краткие итоги

- Анализ данных это:
  - интересно
  - полезно
  - прибыльно

Вопросы

- Анализ данных
  - как изучить
  - что изучить
  - можно ли изучать самостоятельно
  - какие есть направление развития
  - что включает в себя анализ данных

Как изучать анализ данных

- Университет
- Онлайн курсы \ онлайн специализации
- Учебная литература \ интернет

Как изучать анализ данных

- Университет:
  - фундаментальные знания
  - долго
  - 1-4 релевантных курса за все время обучения

Как изучать анализ данных

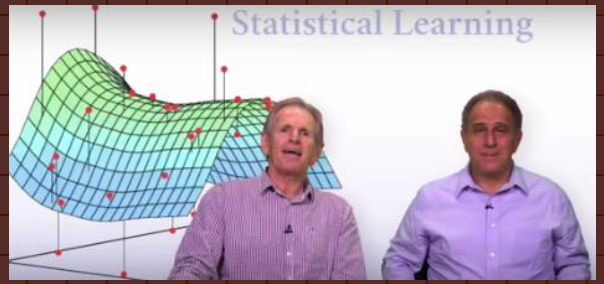
- Университет
- Онлайн курсы \ онлайн специализации
  - множество бесплатных курсов
  - не все курсы одинаково хороши и полезны
  - есть очень хорошие курсы и специализации

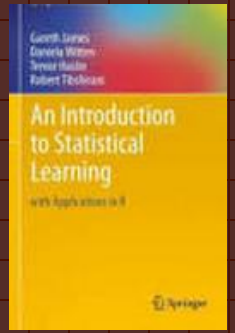
Онлайн курсы

- EDX – MIT – «Меч Аналитики»
- edx.org/course/analytics-edge-mitx-15-071x-0
- Достоинства курса:
  - множество интересных примеров
  - минимум теории,
  - максимум практики
- Язык программирования R

Онлайн курсы

- Stanford – Statistical Learning
- lagunita.stanford.edu/courses/HumanitiesandScience/StatLearning/Winter2015/about
- Достоинства курса:
  - немного теории
  - много практики
  - хороший учебник по курсу
- Язык программирования: R

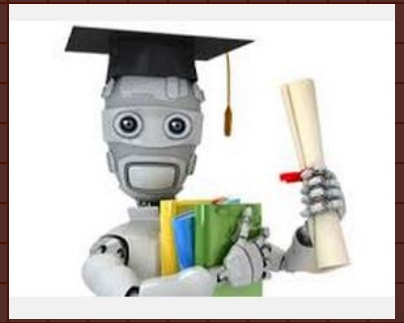




Онлайн курсы

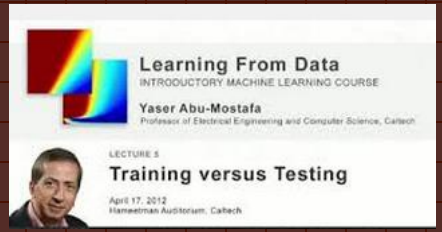
- Coursera, Stanford – Machine Learning
- *coursera.org/learn/machine-learning*
- Достоинства курса:
  - удачное сочетание практики и теории
- Язык программирования: Matlab, Octave

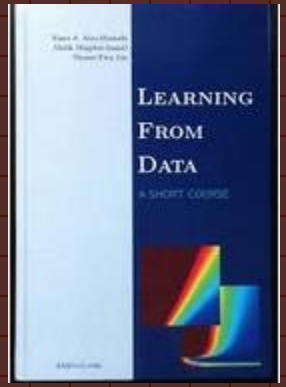
coursera



Онлайн курсы

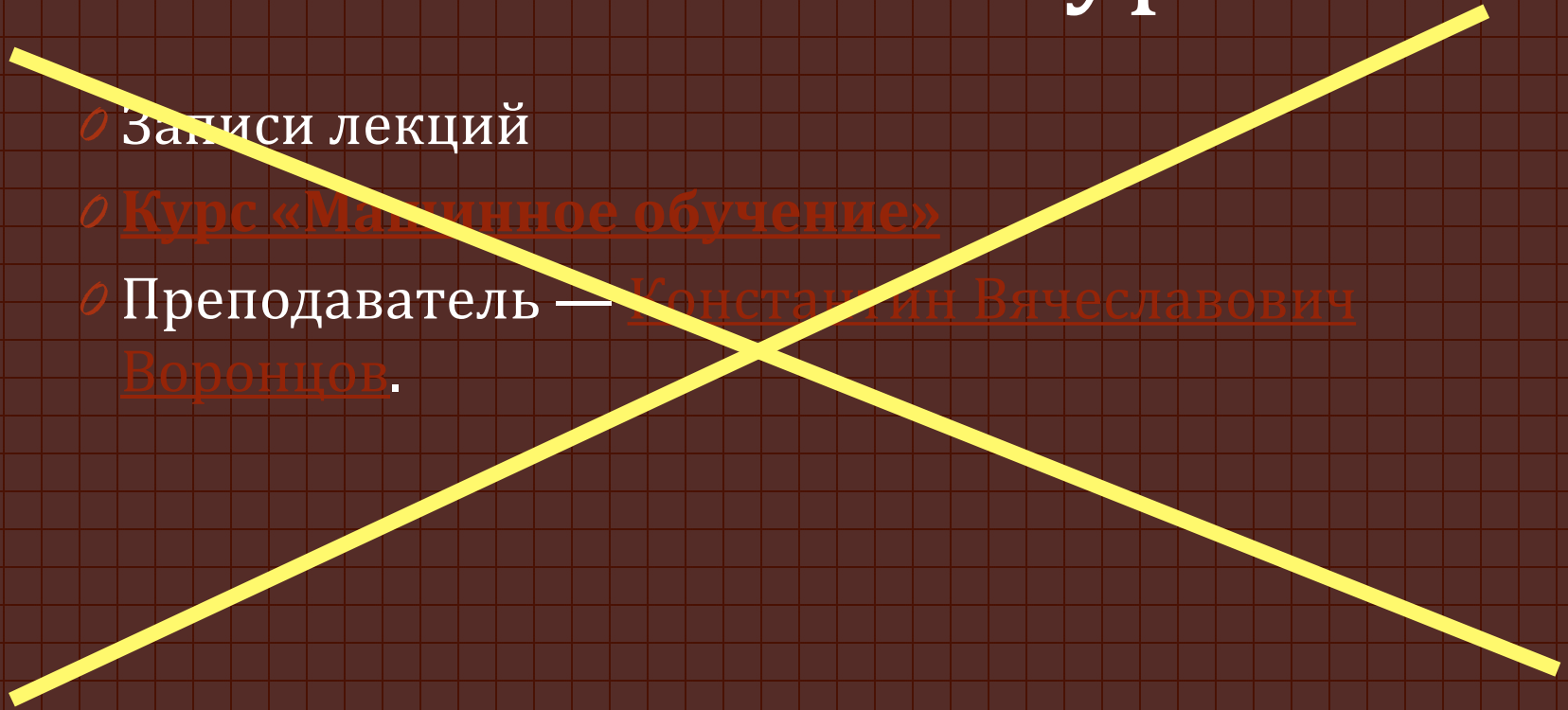
- Edx, Caltech, Learning from data
- https://work.caltech.edu/telecourse.html
- Достоинства курса:
  - твердый теоретический фундамент
  - основные теоретические моменты объяснены в интересной и доступной форме
  - хороший учебник сопровождающий курс





«Анти онлайн курсы»

- Записи лекций
- **Курс «Машинное обучение»**
- Преподаватель — Константин Вячеславович Воронцов.



Онлайн специализации

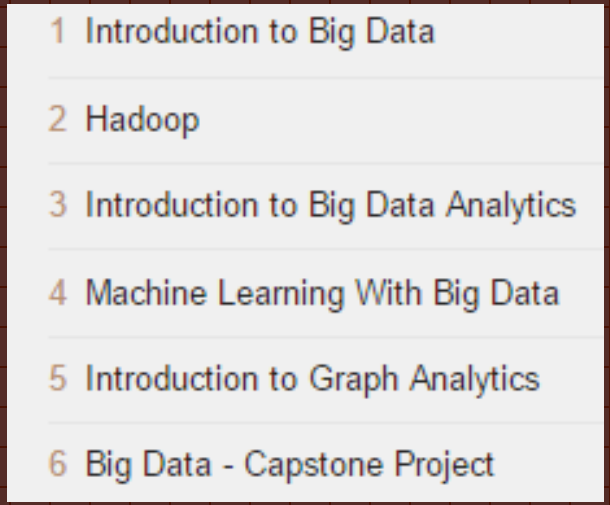
- Coursera, Machine Learning Specialization

  coursera.org/specializations/machine-learning

Язык программирования: *Питон*

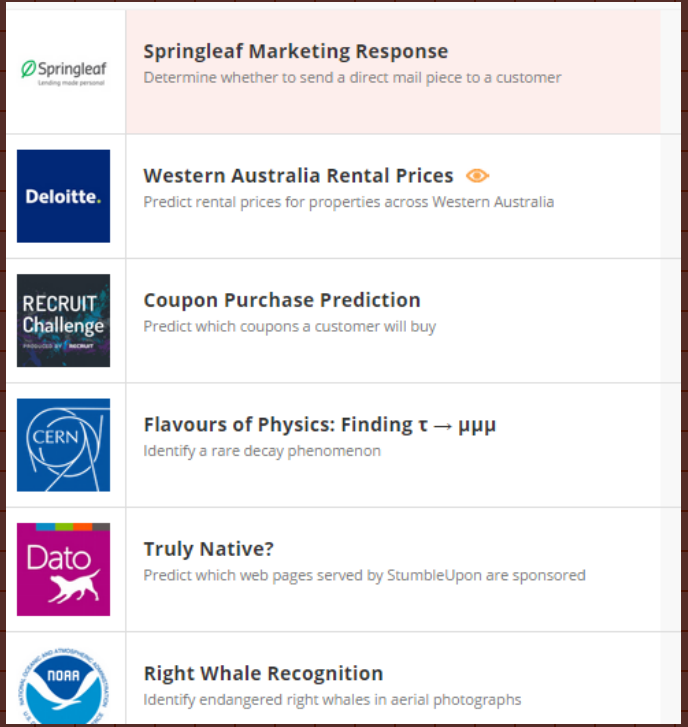
Онлайн специализации

- Coursera, Big Data Specialization
  coursera.org/specializations/big-data



Соревнования

- www.kaggle.com

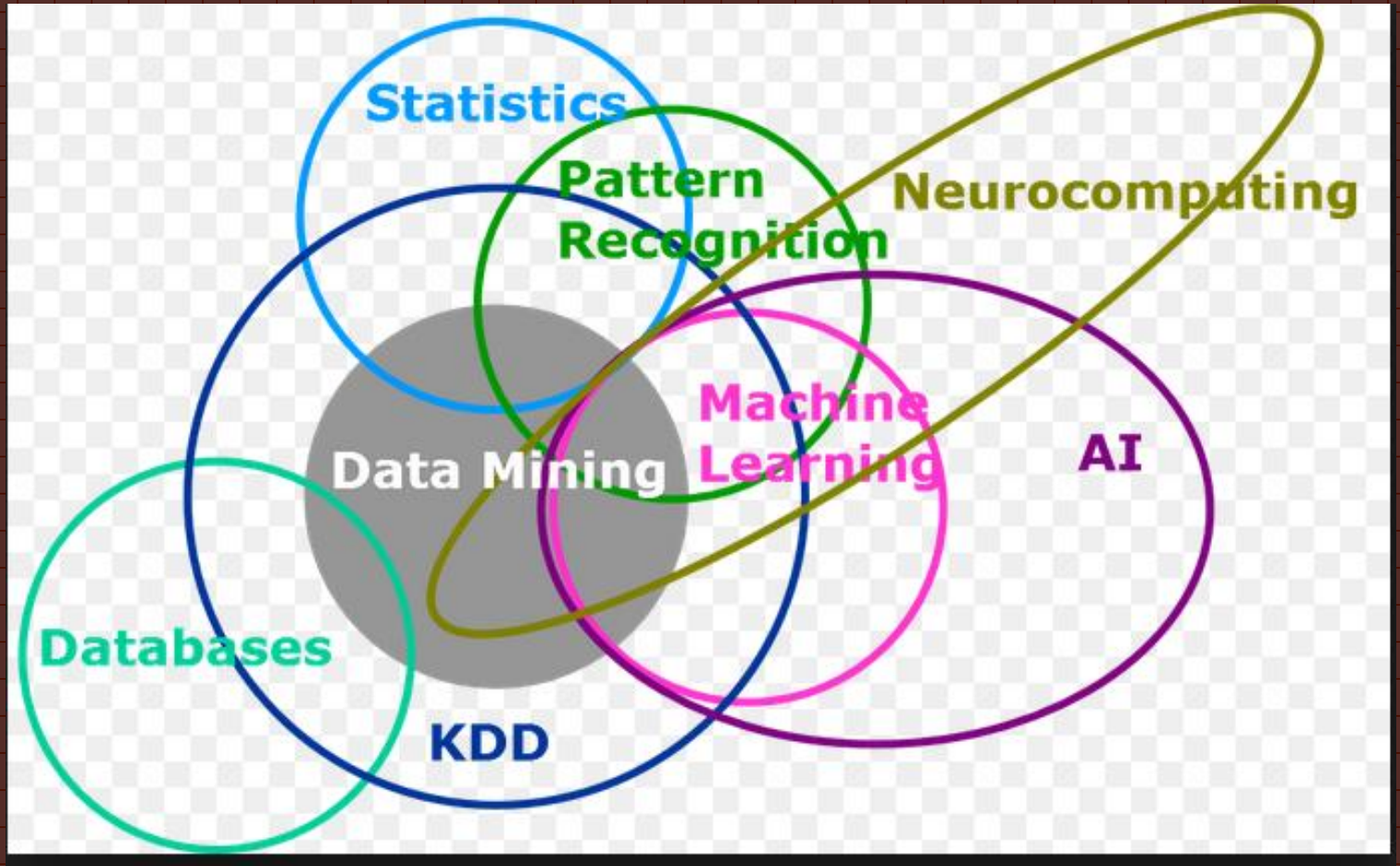


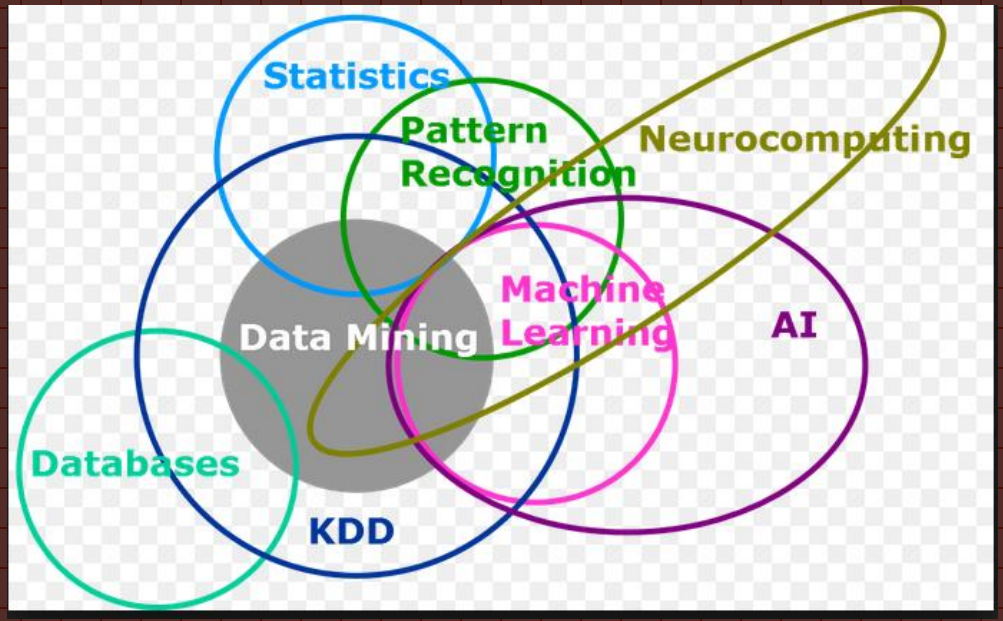
Вопросы

- Что в себя включает анализ данных?

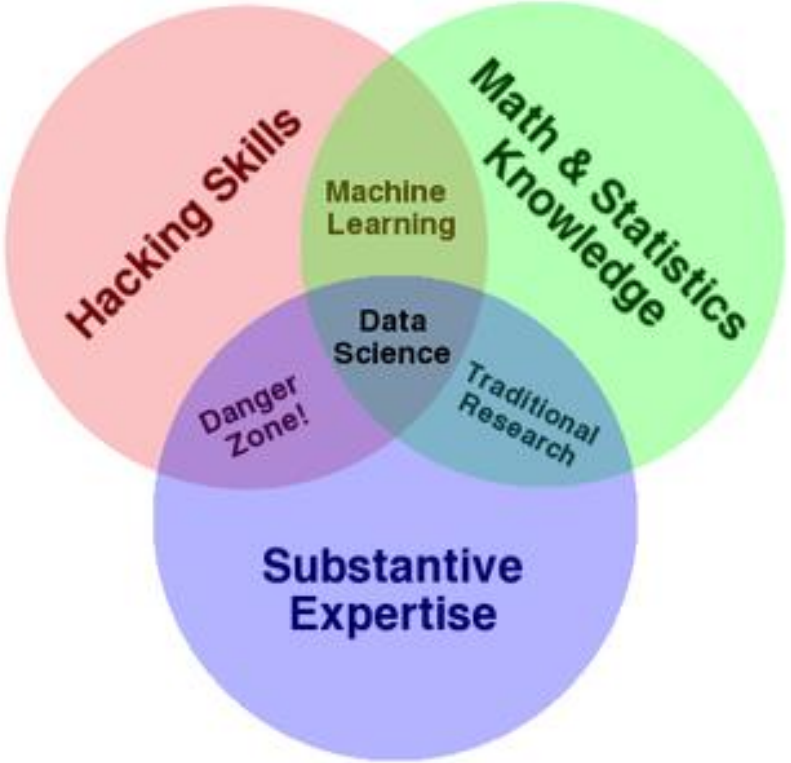
Вопросы

- Что в себя включает анализ данных?
- Анализ данных лежит на пересечениях множества областей наук


Statistics
Pattern Recognition
Neurocomputing
Machine Learning
AI
Data Mining
Databases
KDD



теория вероятностей
статистика
случайные процессы
структуры данных
ИИ
базы данных
параллельные вычисления
оптимизация
выпуклая оптимизация
линейное программирование
алгоритмы
структуры данных
финансовое моделирование
...
...


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Drew Conway

«Каждая кухарка может управлять государством»

Ленин

Может ли каждая кухарка научиться анализировать данные?

Вопрос

- На каких языках лучше заниматься анализом данных?

Вопрос

- На каких языках лучше заниматься анализом данных?
  - нет жестких правил
  - R
  - Python
  - Matlab \ Octave
  - C++
  - …

Вопрос

- Необходимы ли навыки программирования для анализа данных?

Вопрос

- Необходимы ли навыки программирования для анализа данных?

- нет, но желательны

Вопрос

- Необходимы ли навыки программирования для анализа данных?
- нет, но желательны
- Существуют специальные программы позволяющие анализировать данные без знаний программирования

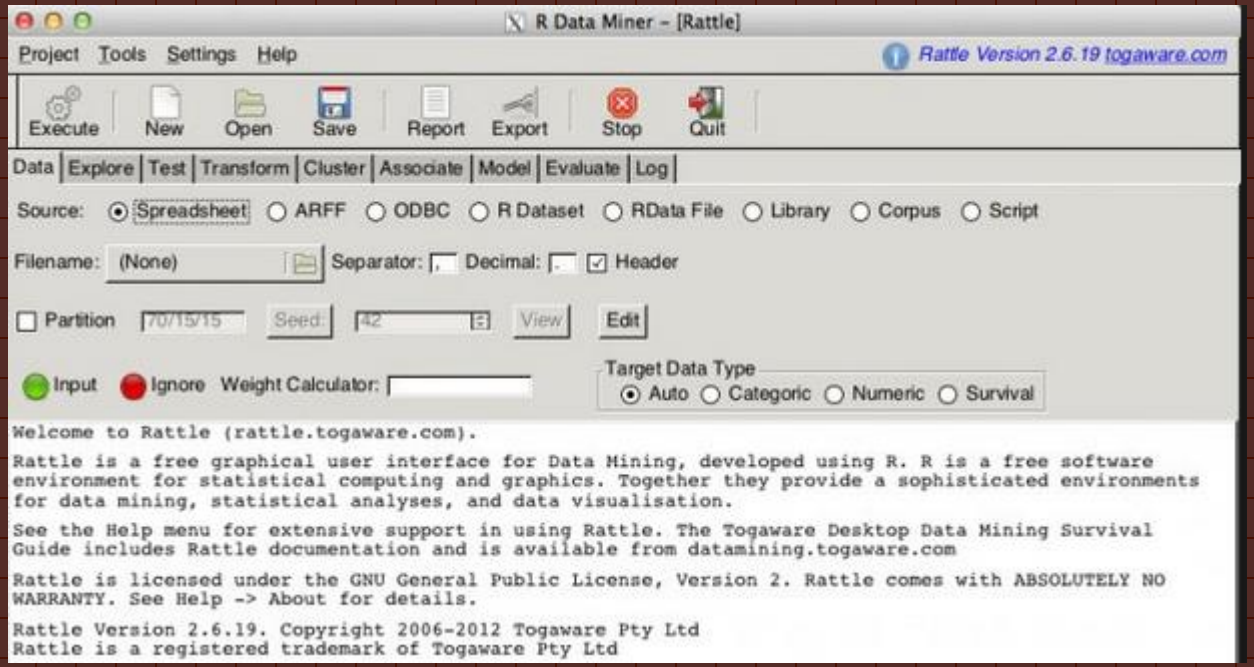
Вопрос

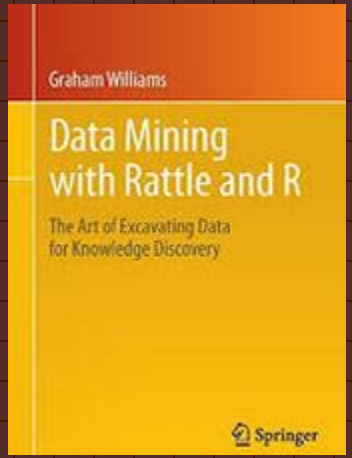
- Необходимы ли навыки программирования для анализа данных?

- Существуют специальные программы позволяющие анализировать данные без знаний программирования:
  - Excel, Rattle, RapidMiner и другие

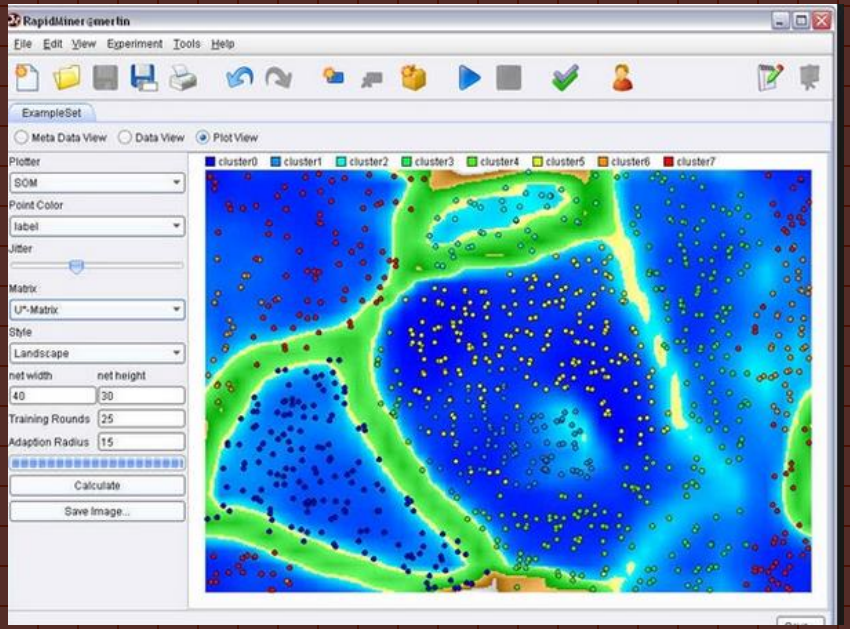
Анализ данных на Excel

Анализ данных на Rattle





Анализ данных RapidMiner

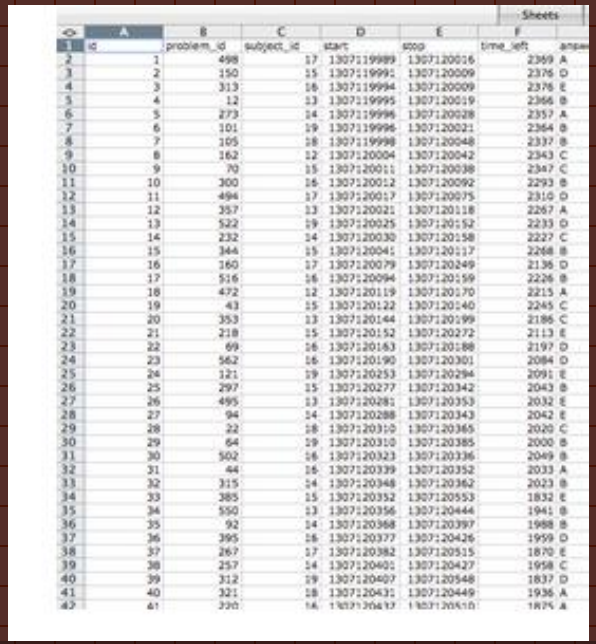


Анализ данных под микроскопом за 5 минут

Machine Learning

Анализ данных под микроскопом за 5 минут

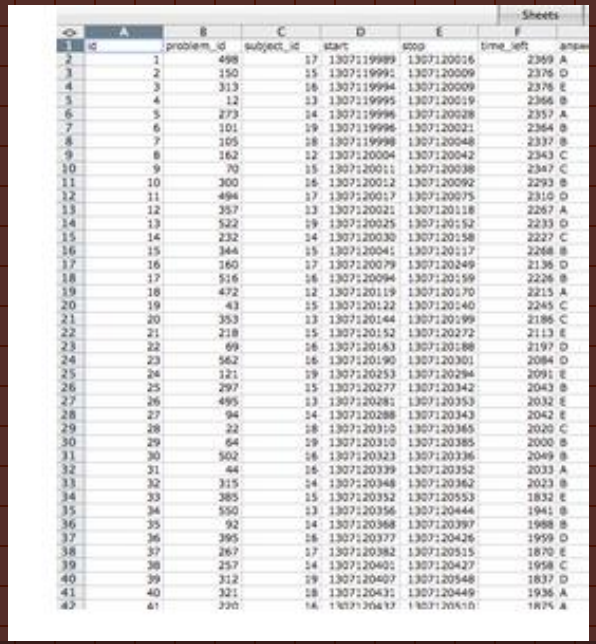
1. Данные:



	A	B	C	D	E	F	
1	id	problem_id	subject_id	start	stop	time_left	answ
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	16	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

Анализ данных под микроскопом за 5 минут

1. Данные:



	A	B	C	D	E	F	
1	id	problem_id	subject_id	start	stop	time_left	answ
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	18	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^\__`_```^a``a`^a_^__]a_]\]`a_______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabaabbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```
```

Анализ данных под микроскопом за 5 минут

1. Где взять данные?

Анализ данных под микроскопом за 5 минут

1. Где взять данные?

- United Nations http://data.un.org/
- U.S. http://www.data.gov/
  - List of cities/states with open data
- United Kingdom http://data.gov.uk/
- France http://www.data.gouv.fr/
- Ghana http://data.gov.gh/
- Australia http://data.gov.au/
- Germany https://www.govdata.de/
- Hong Kong http://www.gov.hk/en/theme/psi/datasets/
- Japan http://www.data.go.jp/

Анализ данных под микроскопом за 5 минут

1. Где взять данные?

- Stanford Large Newtork Data
- UCI Machine Learning
- KDD Nugets Datasets
- CMU Statlib
- Gene expression omnibus
- ArXiv Data
- Public Data Sets on Amazon Web Services

Анализ данных под микроскопом за 5 минут

- Перед анализом данные надо подготовить:

Анализ данных под микроскопом за 5 минут

- Перед анализом данные надо подготовить:
  - препроцессинг: нормализация, …
  - обработать отсутствующие значения
  - привести данные в удобный для анализа вид

- TIDY DATA

messy

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

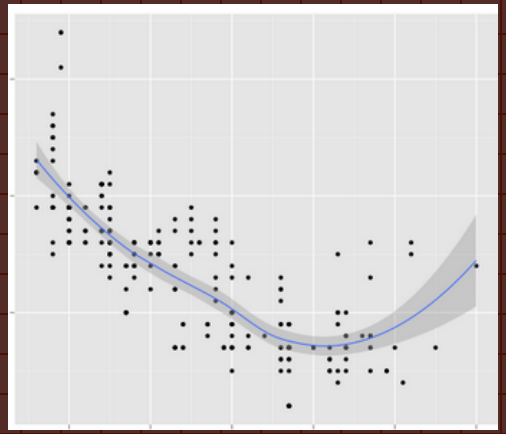
tidy

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Анализ данных под микроскопом за 5 минут

- Предварительное знакомство с данными с помощью графической информации может помочь в дальнейшем анализе

- Exploratory Data Analysis

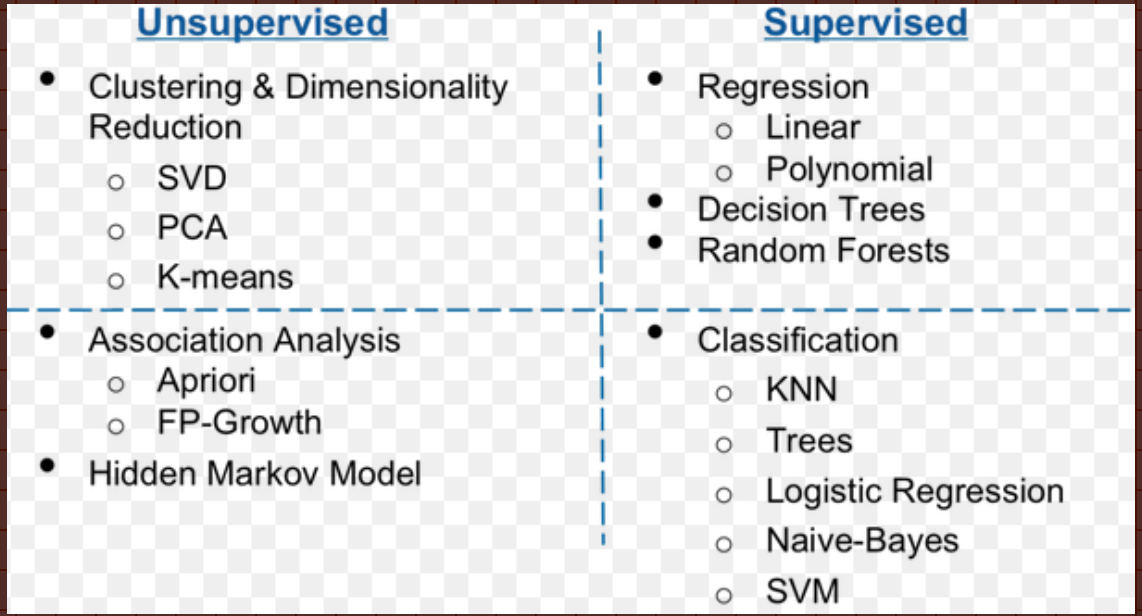


Анализ данных под микроскопом за 5 минут

- Анализ данных, выбор лучшей модели, тестирование качества выбранной модели, использование модели на новых данных

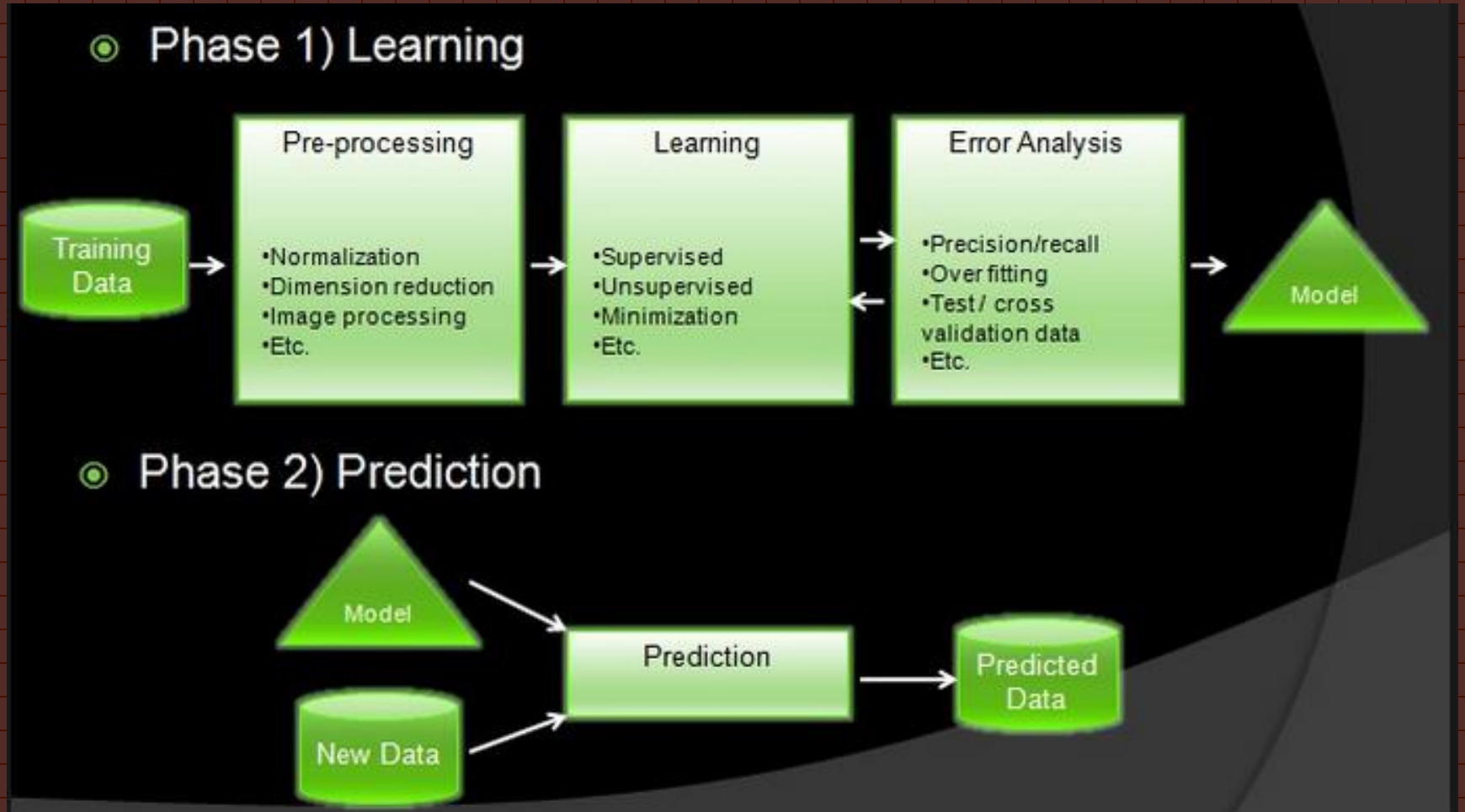
Анализ данных под микроскопом за 5 минут

- Анализ данных, выбор лучшей модели



Анализ данных под микроскопом за 5 минут

- Анализ данных, выбор лучшей модели
  - регрессия линейная
  - дискриминантный анализ
  - логистическая регрессия
  - сплайны
  - случайные деревья
  - случайные леса
  - PCP
  - метод опорных векторов
  - бустинг
  - метод ближайших соседей
  - ….
  - …


Phase 1) Learning
Training Data
Pre-processing
•Normalization
•Dimension reduction
•Image processing
•Etc.
Learning
•Supervised
•Unsupervised
•Minimization
•Etc.
Error Analysis
•Precision/recall
•Over fitting
•Test / cross validation data
•Etc.
Model
Phase 2) Prediction
Model
New Data
Prediction
Predicted Data

Вопрос

- Как вы думаете какой этап занимает больше всего времени? (как правило)
  - скачать данные
  - подготовить данные к анализу
  - выбор лучшей модели
  - представление результатов анализа

Вопрос

- Как вы думаете какой этап занимает больше всего времени? (как правило)
  - скачать данные
  - **подготовить данные к анализу**
  - выбор лучшей модели
  - представление результатов анализа

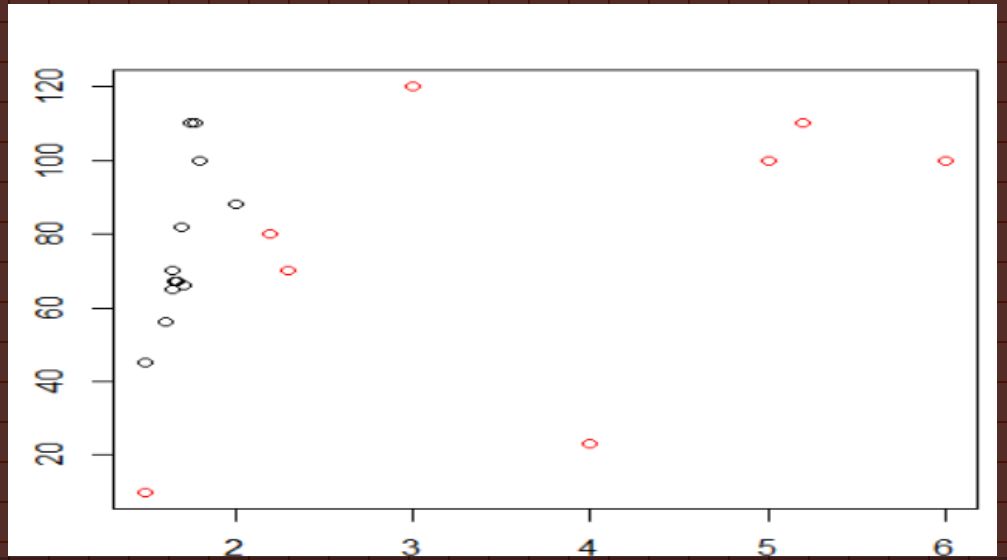
Пример модели для анализа данных

Задача классификации

Цель: научить систему различать людей и пришельцев с марса

Человек или пришелец

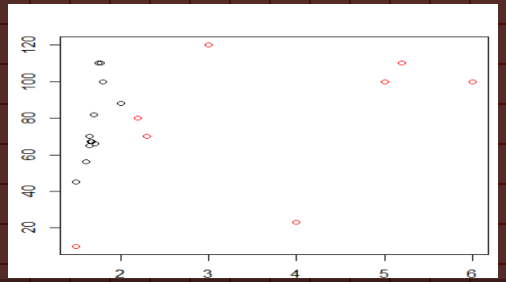
Данные для обучения: таблица содержащая рост и вес 20 кандидатов, 12 из которых люди и 8 из которых пришельцы с Марса

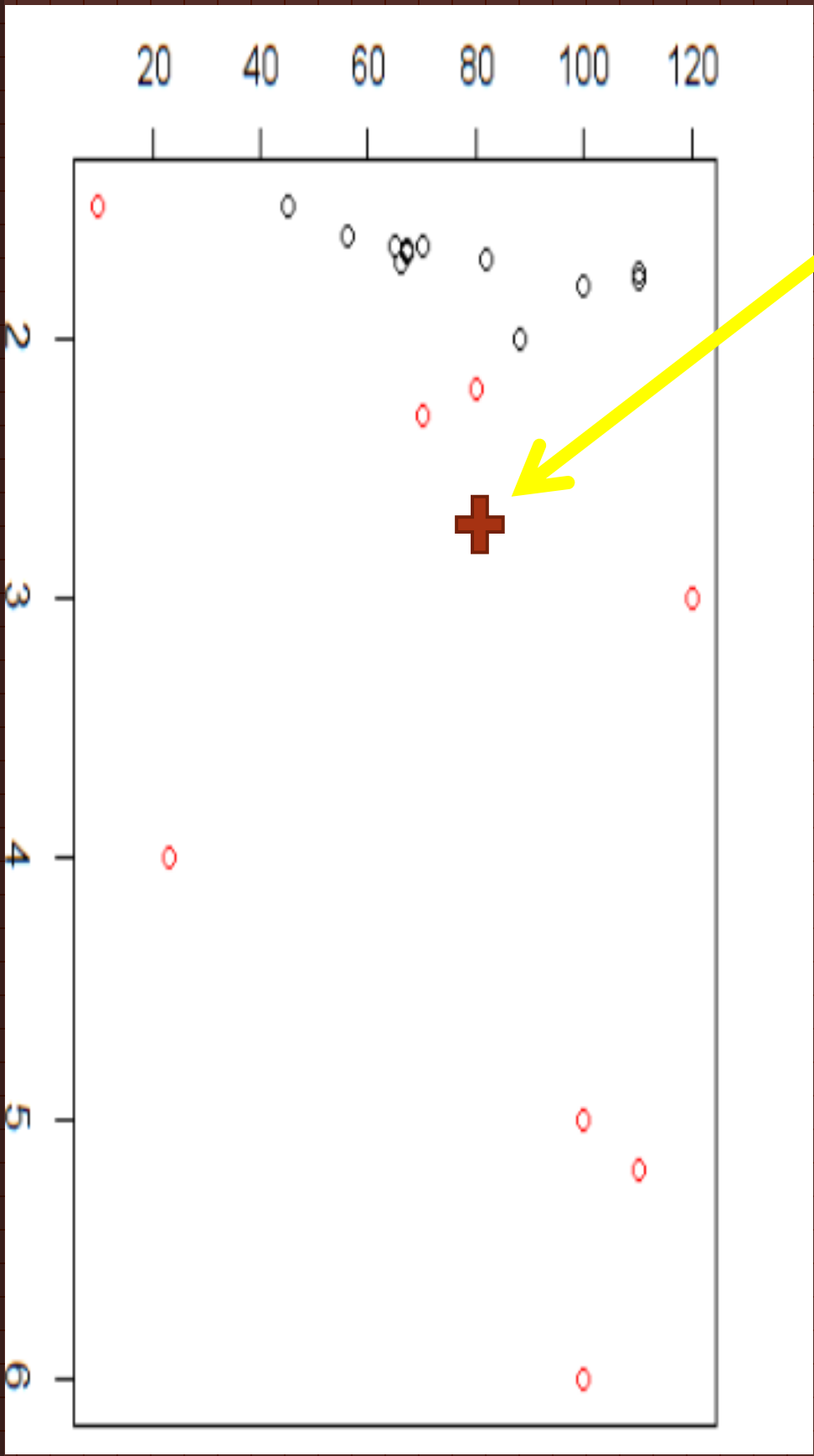


Человек или пришелец

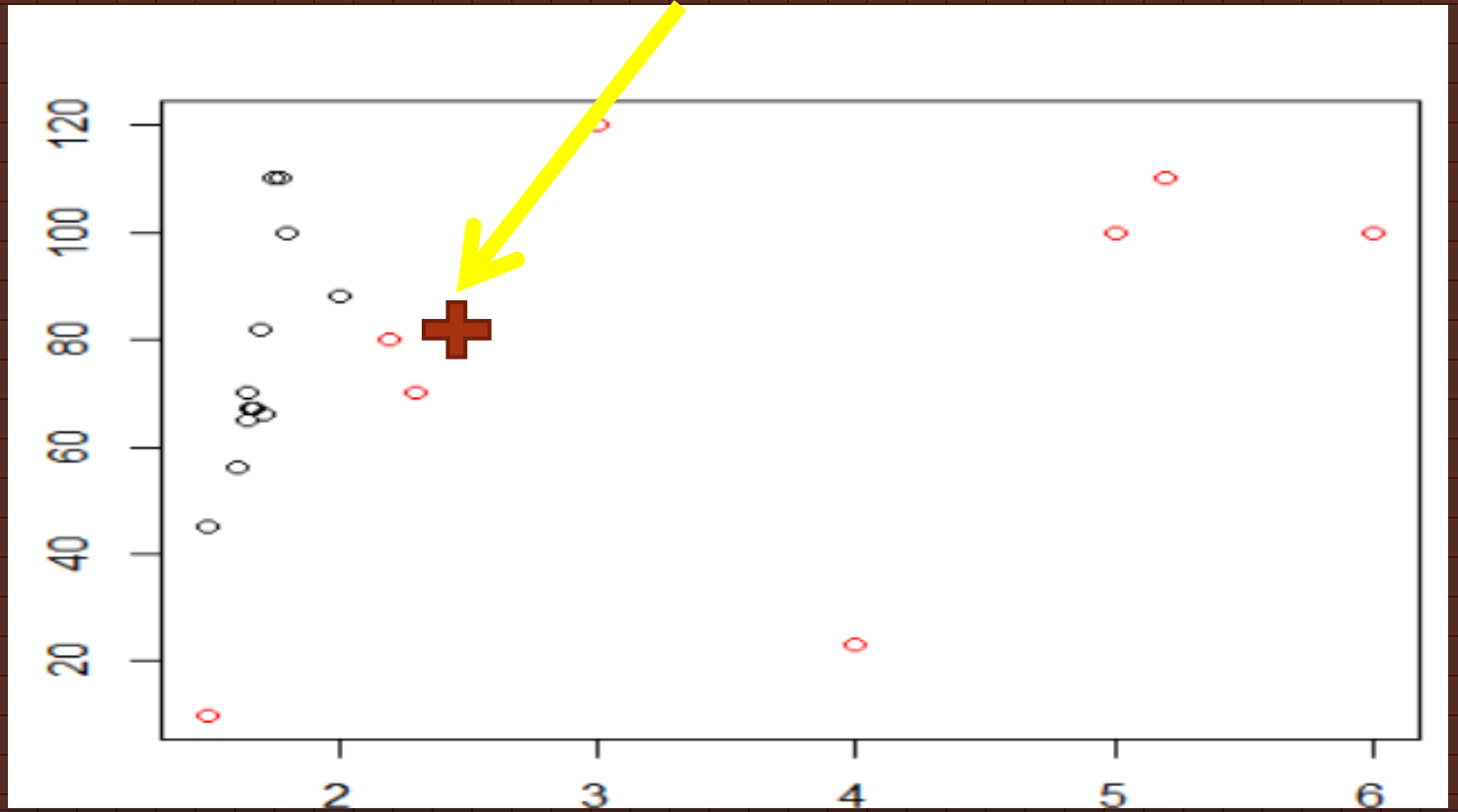
Данные для обучения: таблица содержащая рост и вес 20 кандидатов, 12 из которых люди и 8 из которых пришельцы с Марса

Перед нами новый персонаж, как понять человек это или пришелец?




20
40
60
80
100
120
2
3
4
5
6

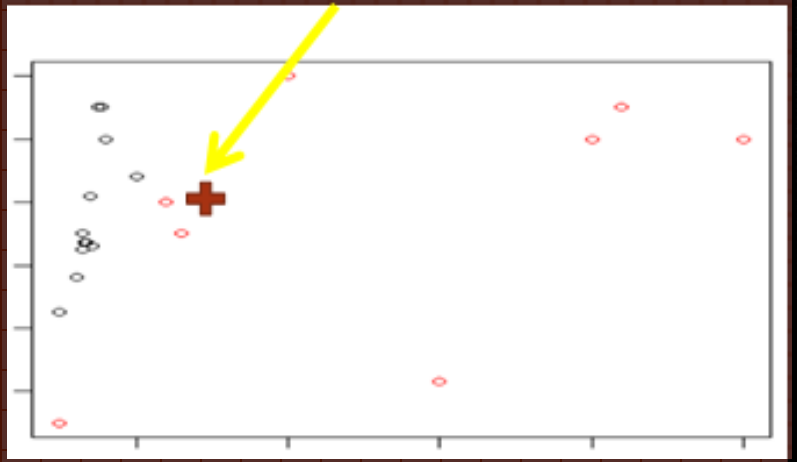
Кто это?

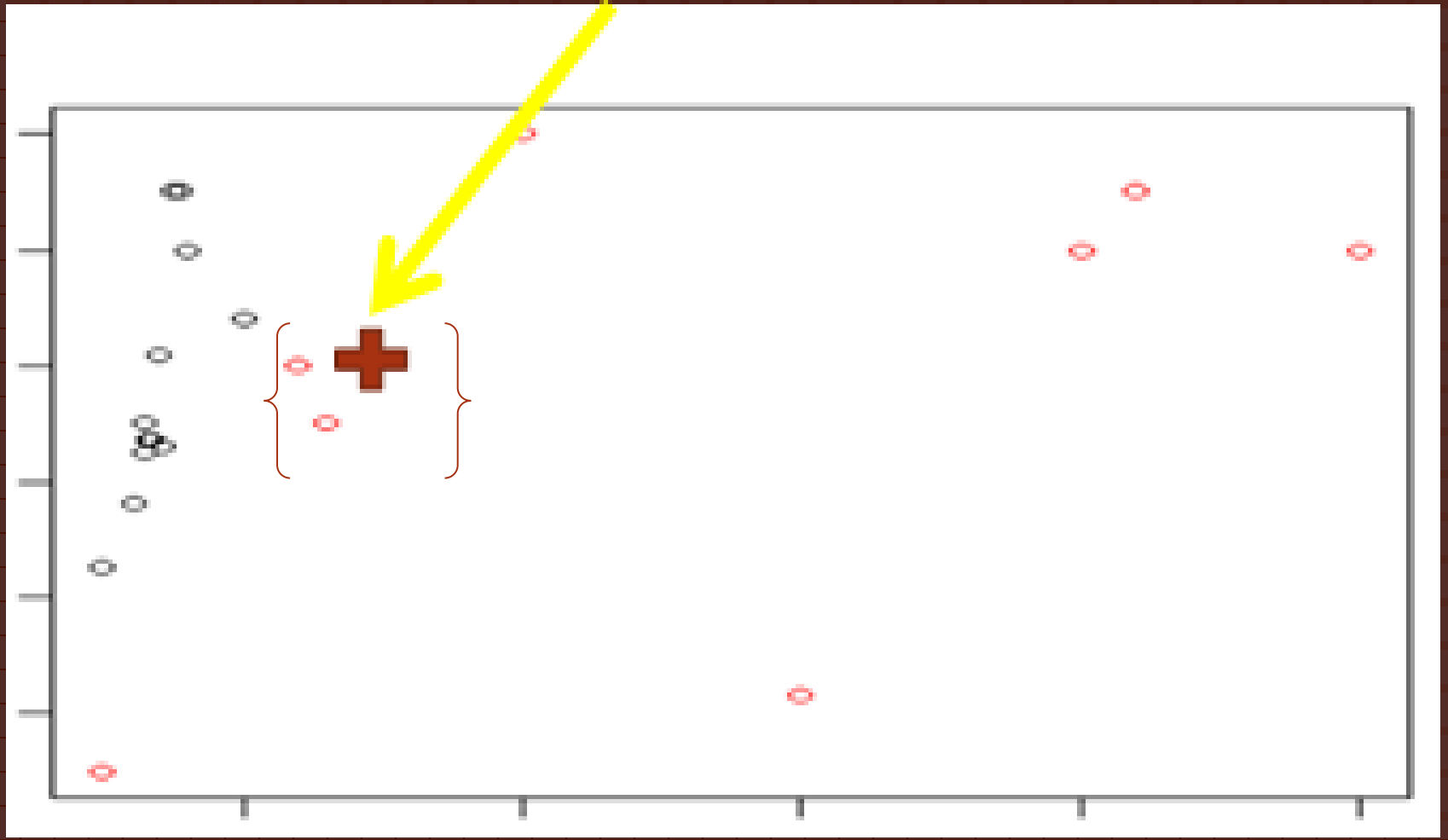


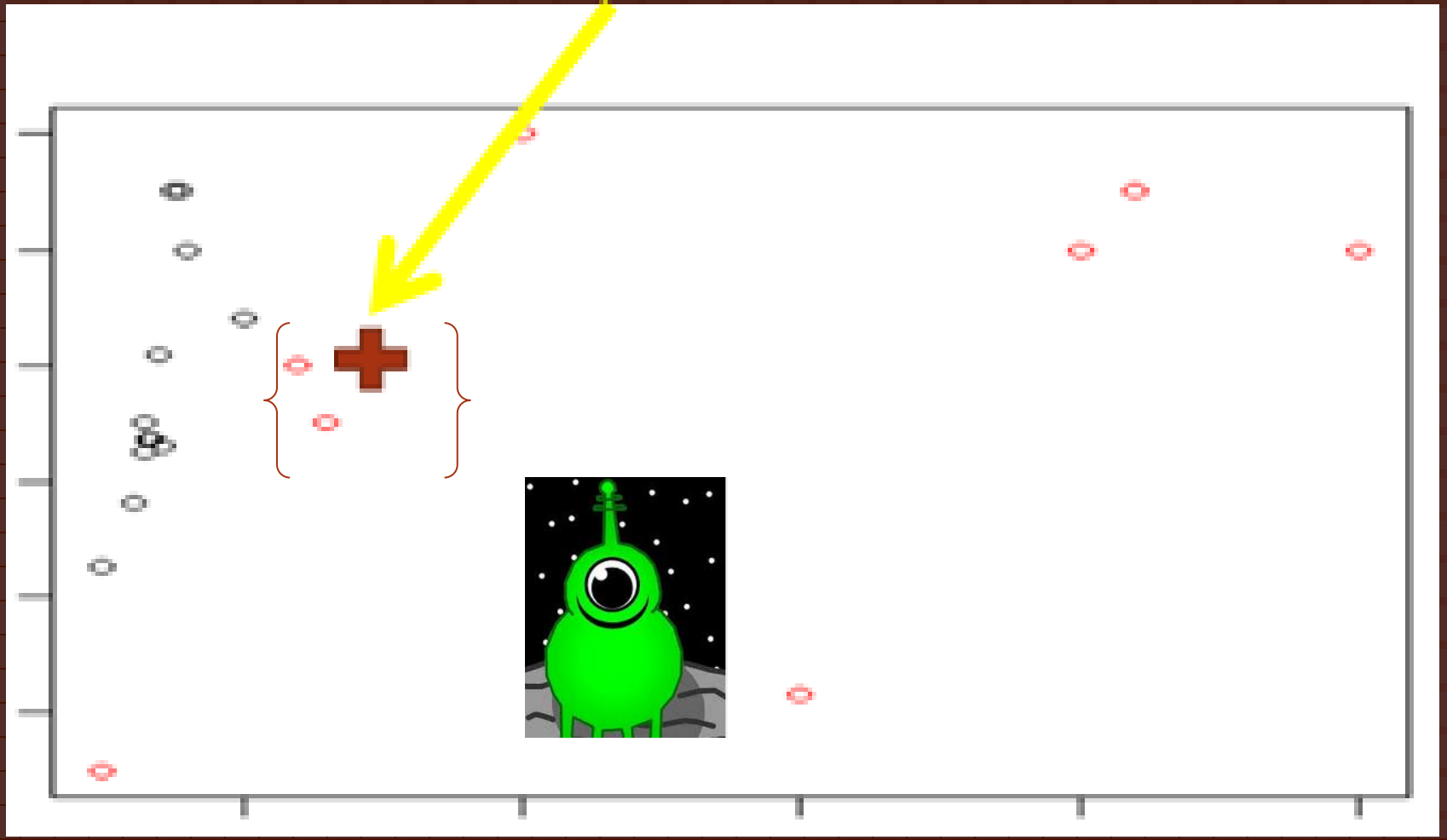
Метод ближайшего соседа

- «скажи мне кто твой друг и я скажу кто ты»
- «дурак дурака видит издалека»

- «Каковы соседи, таков и ты»







Метод ближайших соседей

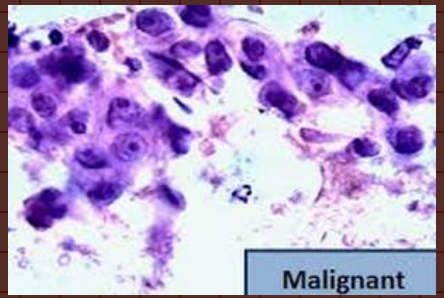
- Несмотря на свою простоту, метод достаточно эффективен при определенных условиях
- Метод хуже работает в пространствах высокой размерности (почему?)

Пример

- Анализ данных при диагностики рака груди
- Wisconsin breast cancer data

- Данные 569 примеров биопсий, каждая из которых характеризуется 30 параметрами

- 31 параметр - тип опухоли



Пример

- Загрузка данных

```
> wbcd <- read.csv("wisc_bc_data.csv",
stringsAsFactors = FALSE)
```

Пример

- Нормализация данных

```
> wbcd_n <- as.data.frame(lapply(wbcd
[2:31], normalize))
```

Пример

- Построение модели KNN

```
> wbcd_test_pred <- knn(train =
wbcd_train, test = wbcd_test,
                          cl =
wbcd_train_labels, k=21)
```

Пример

- Анализ результатов

wbcd_test_labels	wbcd_test_pred Benign	Malignant	Row Total
Benign	61 1.000 0.968 0.610	0 0.000 0.000 0.000	61 0.610
Malignant	2 0.051 0.032 0.020	37 0.949 1.000 0.370	39 0.390
Column Total	63 0.630	37 0.370	100

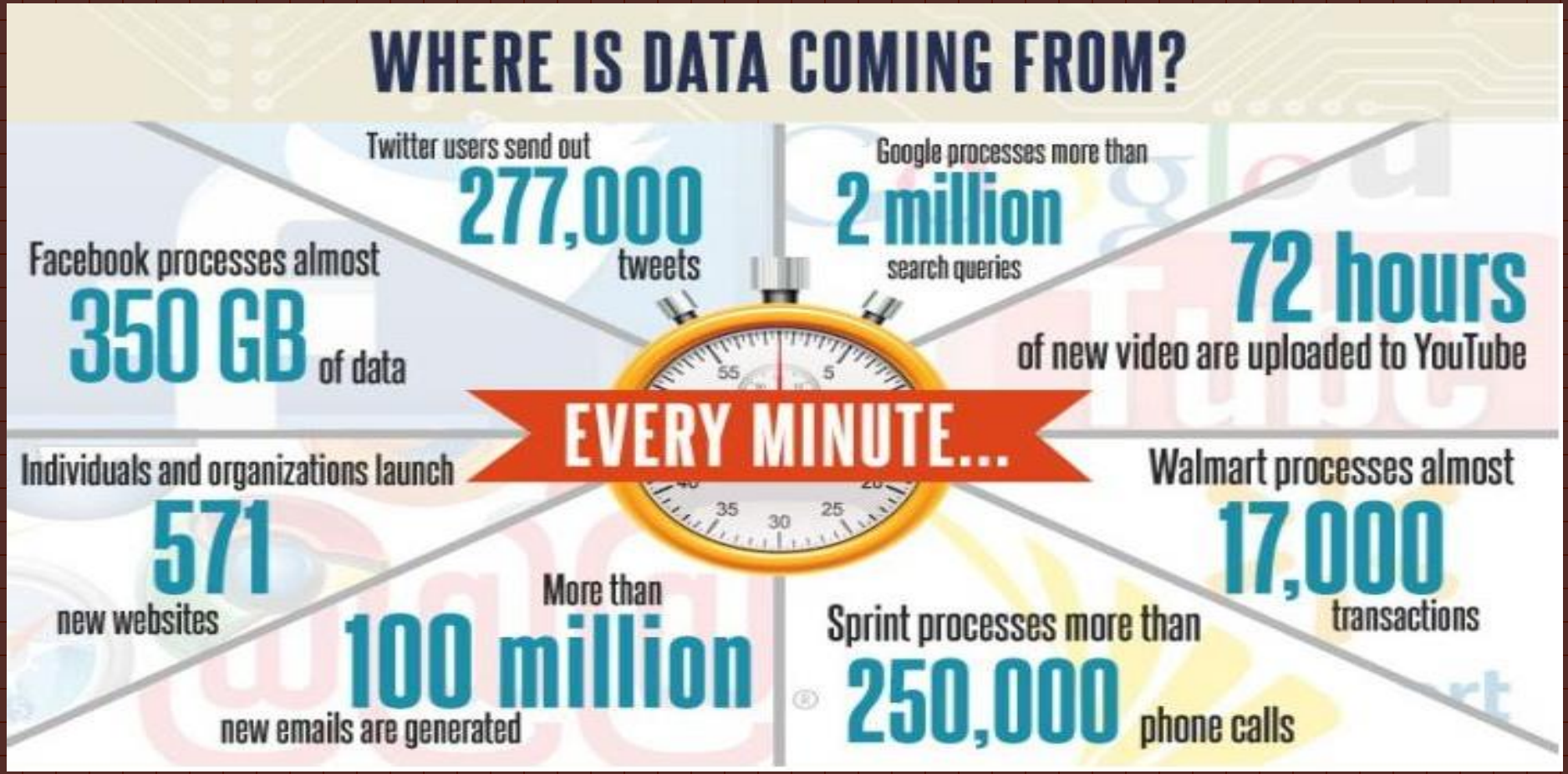
Краткие итоги

- Анализ данных это:
  - круто
  - интересно
  - доступно
  - выгодно

BIG DATA

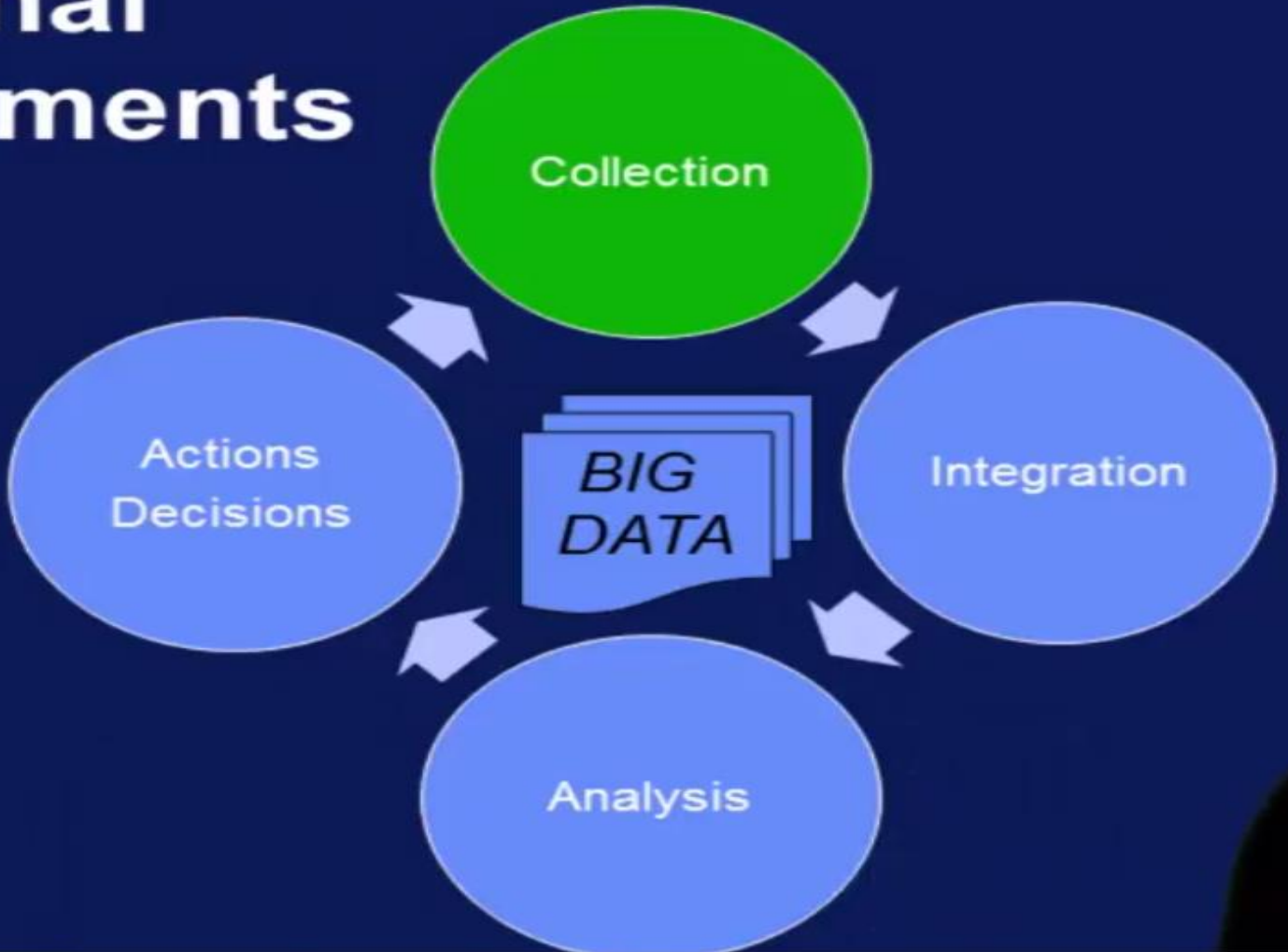

x 1024
x 1024
x 1024
Zettabyte
Exabyte
Petabyte
Terabyte

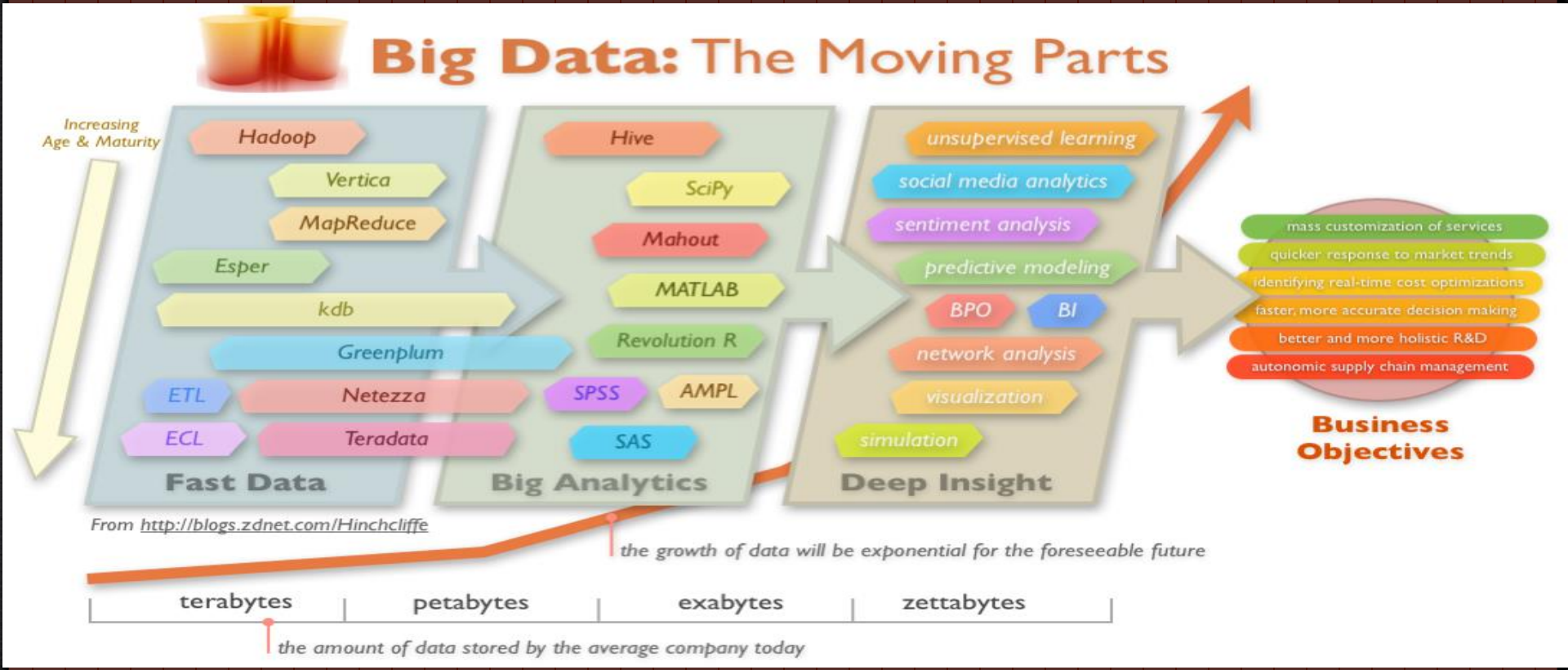
WHERE IS DATA COMING FROM?

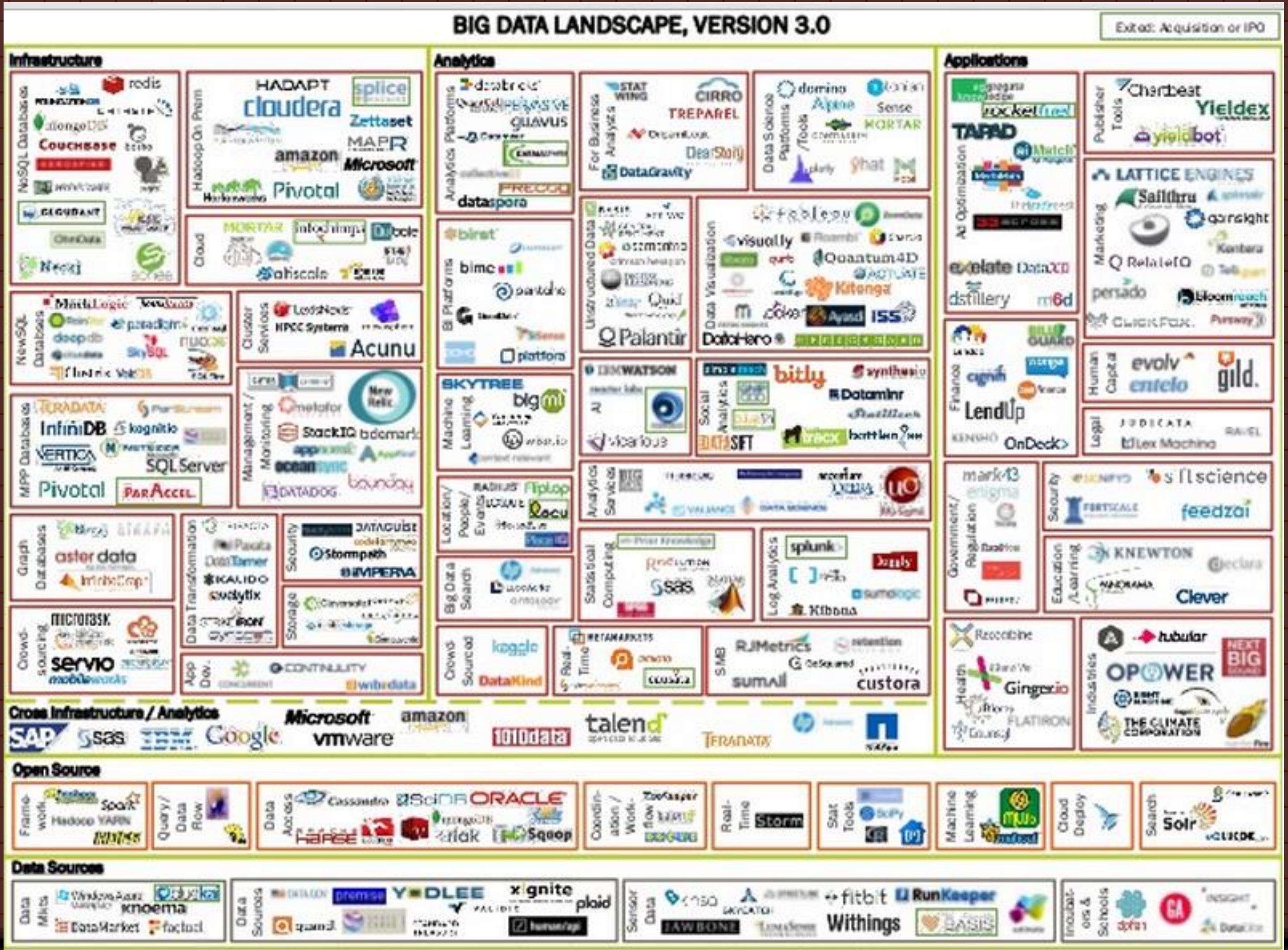


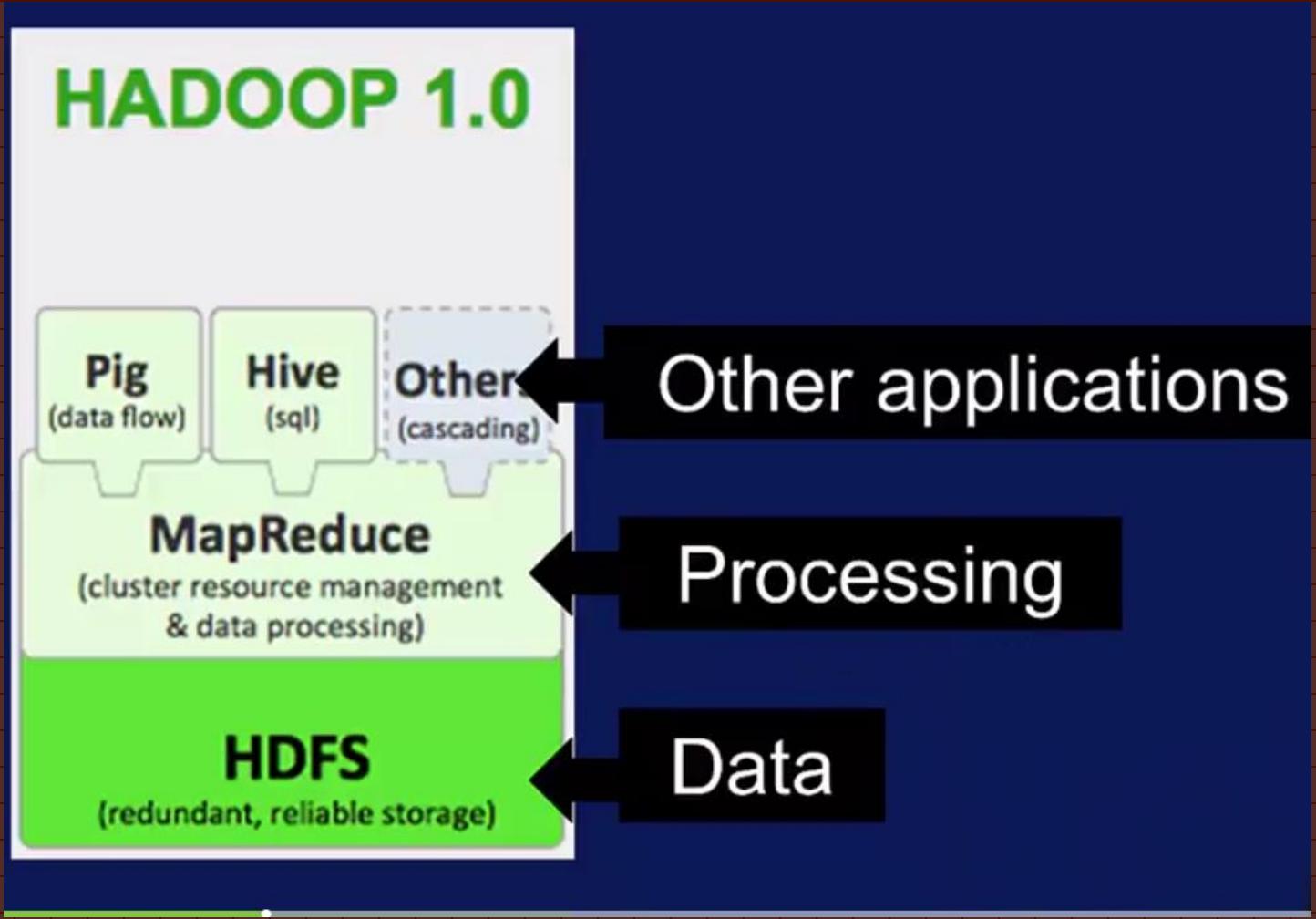

BIG
DATA

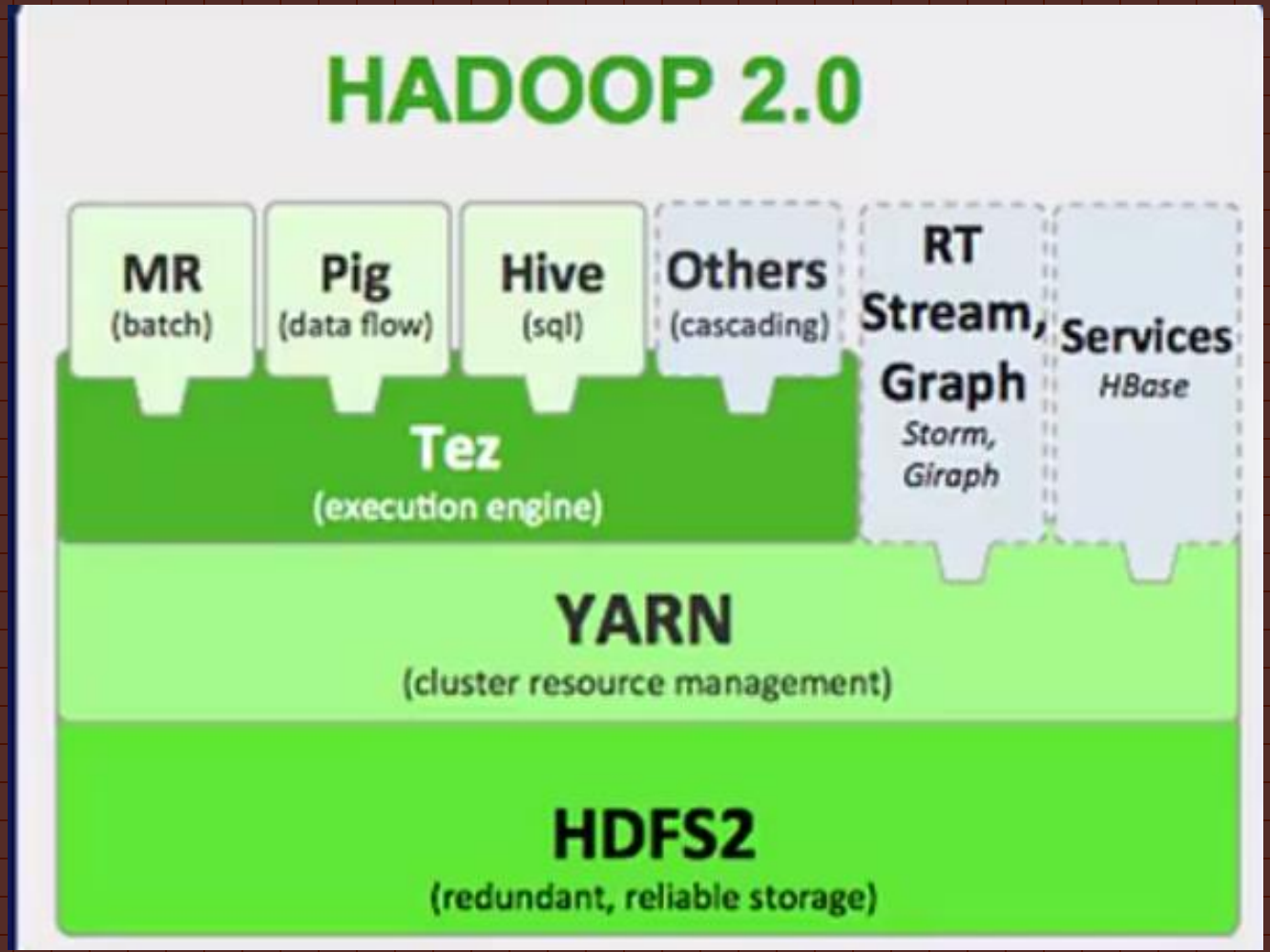
Functional Requirements




Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
From http://blogs.zdnet.com/Hinchcliffe
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today


BIG DATA LANDSCAPE, VERSION 3.0
Exited: Acquisition or IPO
Infrastructure
Analytics
Applications
NoSQL Databases
redis
Couchbase
Hadoop On Premise
HADAPT
cloudera
splice
Zettaset
MAPR
amazon
Microsoft
Pivotal
Hortonworks
Cloud
MORTAR
Qubole
altiscale
NewSQL Databases
MarkLogic
SkySQL
NuoDB
Cluster Services
HPCC Systems
Acunu
MPP Databases
TERADATA
InfiniDB
VERTICA
SQL Server
Pivotal
ParAccel
Management / Monitoring
New Relic
StackIQ
DATADOG
boundary
Graph Databases
aster data
Data Transformation
KALIDO
Security
DATAGUISE
Stormpath
IMPERVA
Storage
Crowd-sourcing
servio
App Dev.
CONTINUUITY
Analytics Platforms
databricks
guavus
PREDIO
dataspora
For Business Analysts
STATWING
CIRRO
TREPAREL
DataGravity
Data Science Platforms / Tools
domino
Alpine
Sense
MORTAR
yhat
Unstructured Data
Palantir
Data Visualization
tableau
visually
Quantum4D
ACTUATE
Kitenga
Ayasdi
DataHero
BI Platforms
birst
bime
pentaho
platfora
Machine Learning
SKYTREE
bigml
wise.io
AI
IBM WATSON
Vicarious
Social Analytics
bitly
Datasift
Dataminr
Location / People / Events
Placeiq
Big Data Search
Analytics Services
Statistical Computing
sas
Log Analytics
splunk
sumologic
Kibana
Crowd-Sourced
kaggle
DataKind
Real-Time
SMB
RJMetrics
custora
Ad Optimization
rocketfuel
TAPAD
eXelate
dstillery
m6d
Publisher Tools
Chartbeat
Yieldex
yieldbot
Marketing
LATTICE ENGINES
Sailthru
persado
bloomreach
ClickFox
Finance
LendUp
OnDeck
Human Capital
evolv
entelo
gild
Legal
JUDICATA
Lex Machina
Government / Regulation
mark43
Security
feedzai
Education / Learning
KNEWTON
Clever
Health
Gingerio
FLATIRON
Industries
OPOWER
THE CLIMATE CORPORATION
NEXT BIG SOUND
Cross Infrastructure / Analytics
SAP
sas
IBM
Google
Microsoft
vmware
amazon
1010data
talend
TERADATA
Open Source
Framework
Spark
Hadoop YARN
Query / Data Flow
Data Access
Cassandra
ORACLE
riak
Sqoop
Coordination / Workflow
ZooKeeper
Real-Time
Storm
Stat Tools
SciPy
Machine Learning
Cloud Deploy
Search
Solr
Data Sources
Data Markets
Windows Azure
knoema
DataMarket
factual
Data Sources
premise
YODLEE
xignite
plaid
quandl
Sensor Data
fitbit
RunKeeper
JAWBONE
Withings
BASIS
Incubators & Schools


HADOOP 1.0
Pig
(data flow)
Hive
(sql)
Other
(cascading)
MapReduce
(cluster resource management
& data processing)
HDFS
(redundant, reliable storage)
Other applications
Processing
Data


HADOOP 2.0
MR
(batch)
Pig
(data flow)
Hive
(sql)
Others
(cascading)
RT
Stream,
Graph
Storm,
Giraph
Services
HBase
Tez
(execution engine)
YARN
(cluster resource management)
HDFS2
(redundant, reliable storage)

Apache Hadoop Ecosystem

Flume
Log Collector

Sqoop
Data Exchange

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

Hive
SQL Query

Hbase
Columnar Store

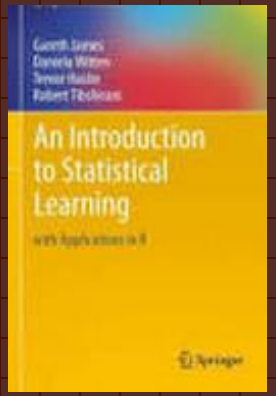
YARN Map Reduce v2
Distributed Processing Framework

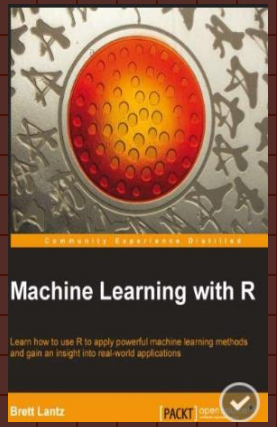
HDFS
Hadoop Distributed File System

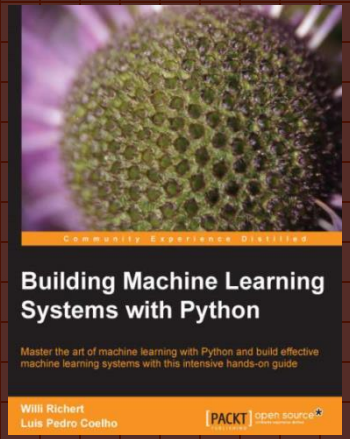
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

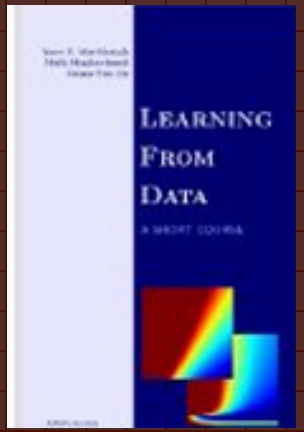

BIG
БОЛЬШИЕ ДАННЫЕ
DATA
РЕВОЛЮЦИЯ, КОТОРАЯ
ИЗМЕНИТ ТО, КАК МЫ ЖИВЕМ,
РАБОТАЕМ И МЫСЛИМ

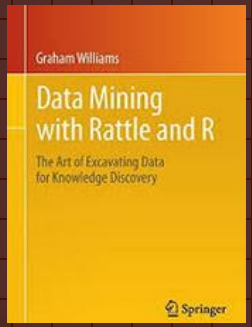
Уголок Библиофила

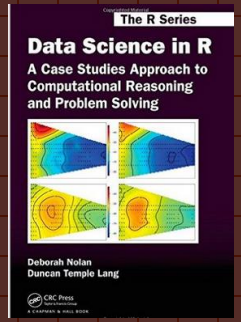


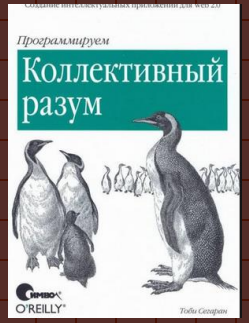












ОБРАТНАЯ СВЯЗЬ

- igkleiner@gmail.com

Ваши вопросы и обратная связь суть лучший источник мотивации

Благодарности

Благодарности

Образовательный IT-портал GeekBrains

Благодарности

- Клейнер Надежда
- Бородин Захар
- Гольцман Александр
- Дубинский Игаль
- Гликин Григорий

Ответы на вопросы слушателей