

Data Science UA. Лекция 2.

Сбор данных

Январь 2017

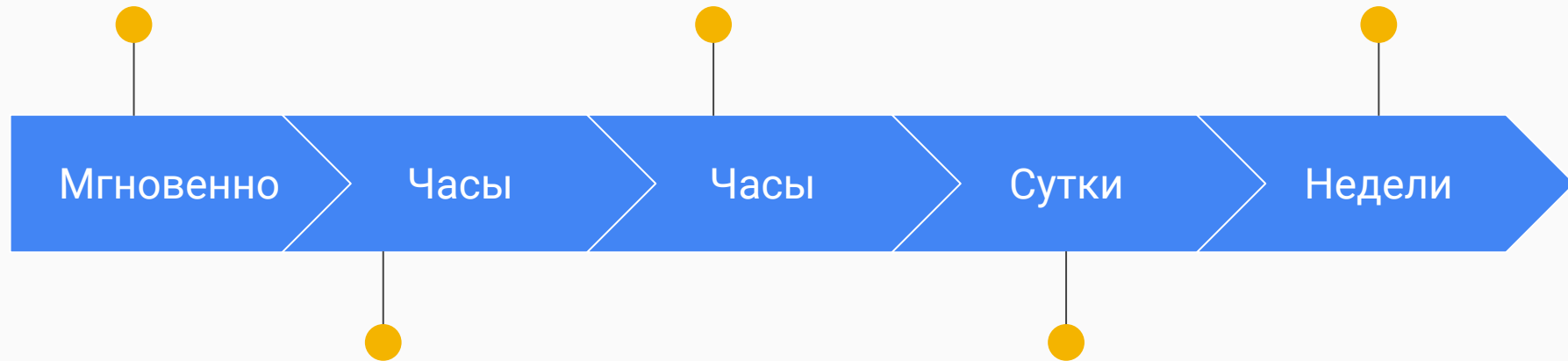
Data Mining



Открытые датасеты

API

Web scraping



Мгновенно

Часы

Часы

Сутки

Недели

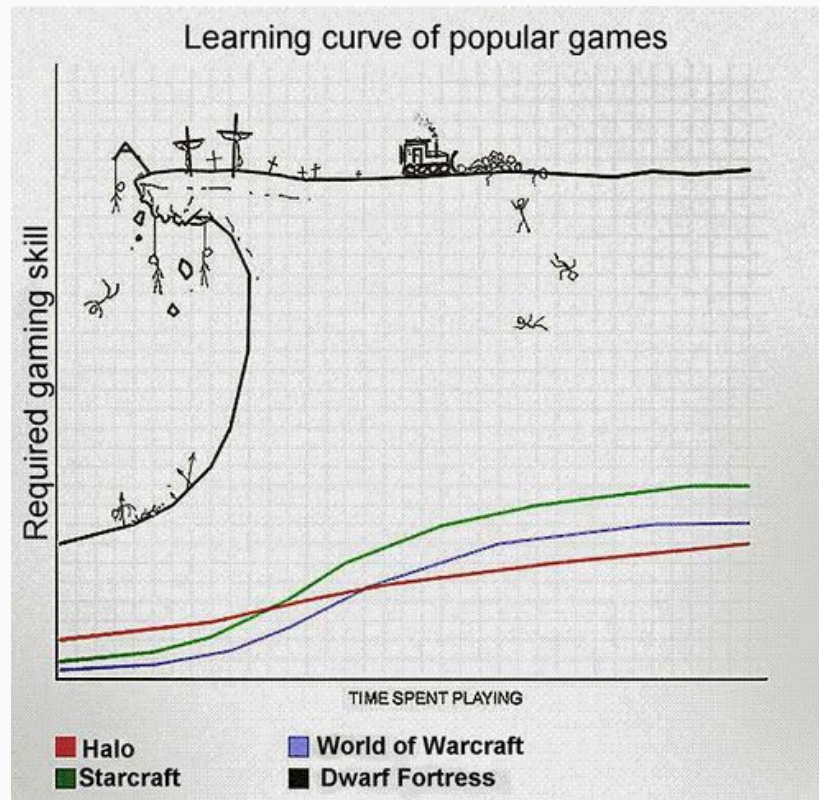
Доступные данные

Перехват
коммуникации
frontend - backend

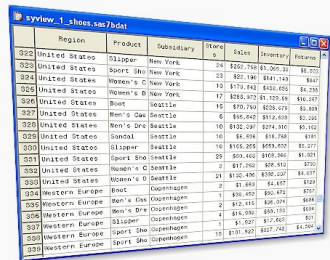
Сложность нахождения данных



Rank	Player	Score	Time	Score	Time
1	Player 1	1000	1000	1000	1000
2	Player 2	900	900	900	900
3	Player 3	800	800	800	800
4	Player 4	700	700	700	700
5	Player 5	600	600	600	600
6	Player 6	500	500	500	500
7	Player 7	400	400	400	400
8	Player 8	300	300	300	300
9	Player 9	200	200	200	200
10	Player 10	100	100	100	100



Открытые датасеты



	Region	Product	Subcategory	Item	Sales	Revenue
522	United States	Slippers	New York	24	\$20,740	\$1,209,321
523	United States	Sport Sho	New York	23	\$22,180	\$141,143
524	United States	Women's C	New York	13	\$179,760	\$10,653
525	United States	Women's D	New York	12	\$243,472	\$1,123,124
526	United States	Boat	Seattle	10	\$10,700	\$103,074
527	United States	Men's C	Seattle	6	\$10,540	\$112,634
528	United States	Men's D	Seattle	10	\$102,230	\$104,310
529	United States	Unlabeled	Seattle	10	\$1,400	\$10,700
530	United States	Slippers	Seattle	10	\$101,201	\$93,502
531	United States	Sport Sho	Seattle	23	\$10,400	\$100,300
532	United States	Women's C	Seattle	2	\$17,200	\$10,512
533	United States	Women's D	Seattle	24	\$122,400	\$100,220
534	Western Europe	Boat	Copenhagen	2	\$1,400	\$1,400
535	Western Europe	Men's C	Copenhagen	1	\$10,400	\$10,400
536	Western Europe	Men's D	Copenhagen	2	\$11,100	\$10,100
537	Western Europe	Slippers	Copenhagen	4	\$10,300	\$10,100
538	Western Europe	Sport Sho	Copenhagen	1	\$1,400	\$1,400
539	Western Europe	Sport Sho	Copenhagen	10	\$10,700	\$10,700
540	Western Europe	Sport Sho	Copenhagen	10	\$10,700	\$10,700

1. Великолепны для тренировки.
2. Чаще всего уже предварительно обработаны.
3. Множество ресурсов и большой выбор.
R-DIR Free Datasets.
4. Вероятность найти нужный в реальном проекте - ничтожно мала.



Доступные данные



1. Данные уже есть.
2. Ограниченный набор.
3. Чаще всего нуждаются в дополнении и расширении.
4. Иногда не в цифровом формате.
5. Если компания не задумывалась о Data Science заранее - скорее всего данные не те, что нужны :)
6. Вы можете “заказать” необходимое.





1. Удобный программный интерфейс.
2. (Обычно) высокое качество данных.
3. Поддержка поставщика данных.
4. Есть далеко не везде.
5. Может быть платным / неудобным / не иметь библиотеки на вашем языке.
6. Помните, что в итоге это всего лишь запросы и ответы.





1. Не всегда доступно.
2. Не всегда законно.
3. Формат данных может требовать экстрасенсорных способностей для понимания.
4. Не рекомендуется.
5. Все еще порой лучший из доступных путей. :(



Web scraping



1. В интернете есть всё.
2. Чтобы достать это всё - придется очень потрудиться.
3. Договориться с админами сайтов - проще, но не всегда доступно.
4. Копать придется долго.



Общие рекомендации

1. Когда только возможно - используйте несколько источников.
2. Если данные можно отфильтровать на этапе сбора - делайте это.
3. Логируйте всё, отмечайте прогресс сбора данных.
4. Оптимизируйте процесс сбора данных (в разумной мере).
5. Уважайте свои источники
 - a. Прочитайте лицензионное соглашение / соглашение пользователя / другие легальные документы. Я серьезно.
 - b. Не подвергайте ресурсы-источники неразумной нагрузке.
 - c. Если данные планируется использовать в коммерческих целях - спросите разрешение.
 - d. Robots.txt - обоюдоострый меч.



Полезность данных (для machine learning)



-2

0

1

2

Неправильные
данные

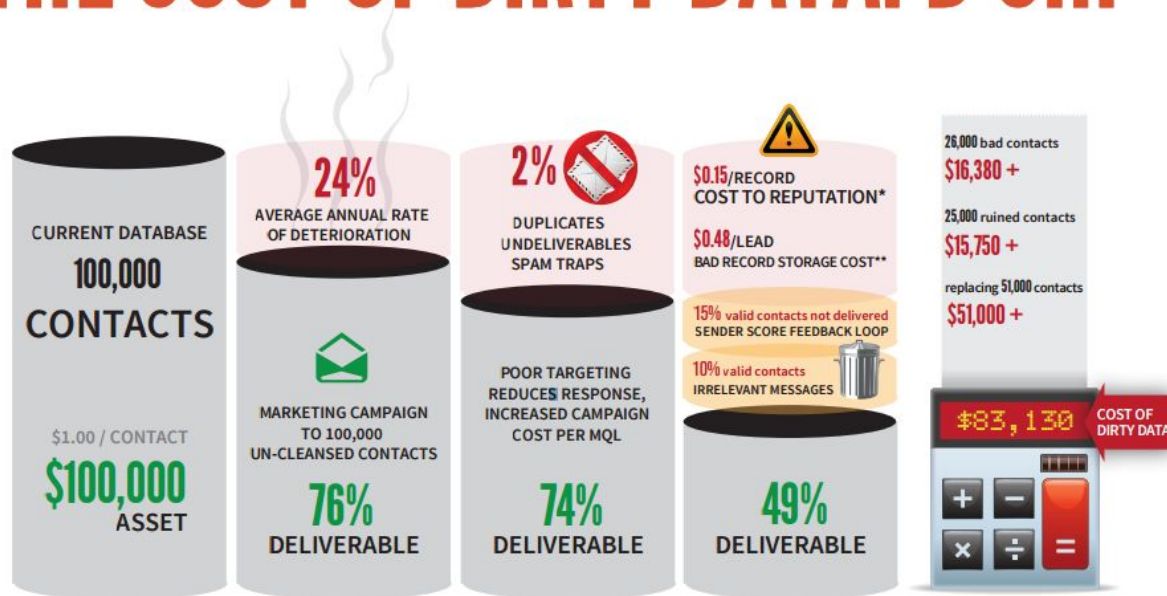
Отсутствующие
данные

Грязные
данные

Очищенные
данные

Пример “скрытой стоимости” плохих данных.

THE COST OF DIRTY DATA. D'OH!



a

REACHFORCE

ebook

Вопросы ?

При оформлении слайдов использованы изображения с следующих страниц.

<http://marketingincolor.com/garbage-in-garbage-out-the-real-cost-of-poor-customer-data/>

https://www.google.com.ua/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwjVhbn9os7RAhWBchQKHUbkBPgQjxwIAw&url=http%3A%2F%2Fwww.business2community.com%2Fmarketing%2Fmuch-dirty-data-costing-01241847&bvm=bv.144224172.d.ZGg&psig=AFQjCNHRV4sahsBbn1LOTPRs7Atwmr_qzA&ust=1484917314701977&cad=rja

<http://www.prosoftsolutions.net/blog/bid/146041/Dirty-Data-What-is-it-how-does-it-cause-problems-and-what-is-the-solution>

<http://measuredme.com/2013/01/personal-analytics-101-how-to-deal-with-holes-in-your-self-tracking-data/>

<http://blog.syncsort.com/2014/01/big-data/big-data-quality-gigo-lives-on/>

https://en.wikipedia.org/wiki/Oil_refinery

<http://glushko.com.ua/blog/2010/08/19/zabroshennye-shaxty-stebnika-chast1/>

<http://serptool.com/keyword-thief>

<https://todayilearned.dirty.ru/chto-s-mesta-v-karer-eto-ne-na-samom-dele-v-karer-566903/>

<http://cmeiinternational.com/metals-ingots.html>

<http://www.snkdiamondsdesign.com/2016/04/29/diamonds-rough/>

<http://gaming.stackexchange.com/questions/21664/dwarf-fortress-learning-curve>

http://meta-x.com/sy_view.html

<https://www.essentialsql.com/what-are-the-major-parts-of-a-database/>

<http://www.commgate.com/page/cloud-api>

<http://www.evolve-incorporated.com/services/ajax-programming/>

<http://www.sbp-romania.com/Blog/2011/01/11/web-vs-desktop-applications-friends-or-foes.aspx>

<http://www.xylem.com/treatment/fr/industries/mining>