

Analityka Big Data / Data Science w instytucji finansowej

3 marca 2019, Warszawa

O mnie...

Kamil Żbikowski

- adiunkt @Instytut Informatyki Politechniki Warszawskiej
- lead data scientist @mBank
- CTO @ inhire.io

wcześniej:

Bazaar Blockchain Technologies,

Turbine Analytics,

Accenture

kamil.zbikowski@ii.pw.edu.pl

Czym jest Big Data?

Definicja „kanoniczna” – dane zbyt duże do bezpośredniego przetwarzania dostępnymi narzędziami (vs. „PC Data”, „Server Data” itd.)

Definicja „marketingowa” – 4xV (*Doug Laney*)

- Volume **B**ardzo duże ilości danych
- Velocity **B**ardzo szybko przyrastające dane
- Variety **B**ardzo zróżnicowane typy danych
- Variability **B**ardzo zmienne dane

Bullshit...

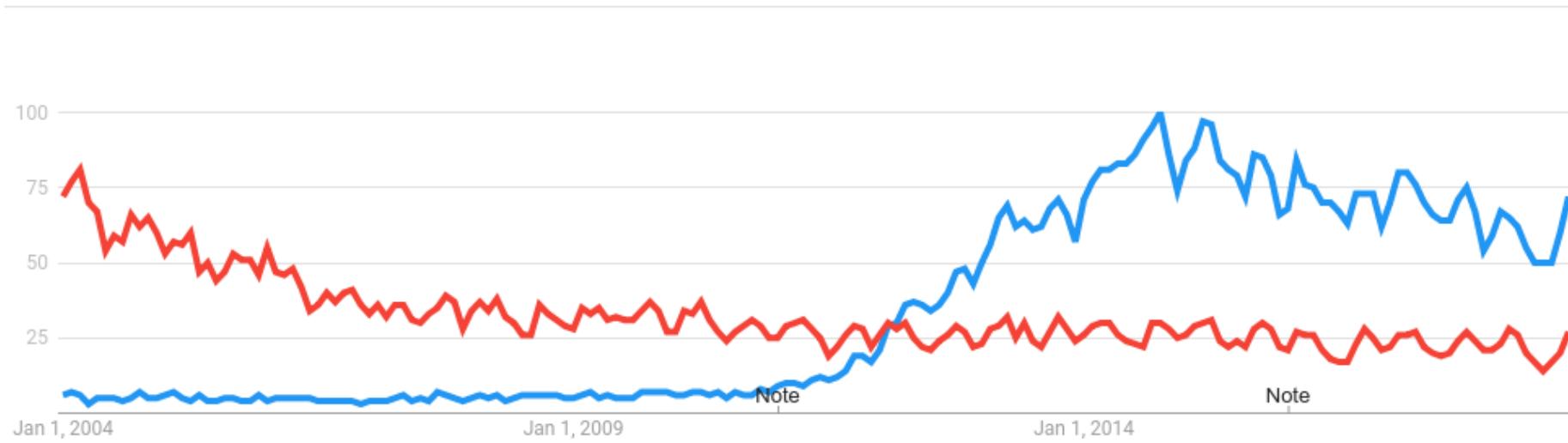
Ta definicja *ceteris paribus* tak naprawdę niewiele mówi o charakterze danych...

Kwestie terminologiczne

Google Trends
dysponuje
danymi dopiero
od 2004 roku



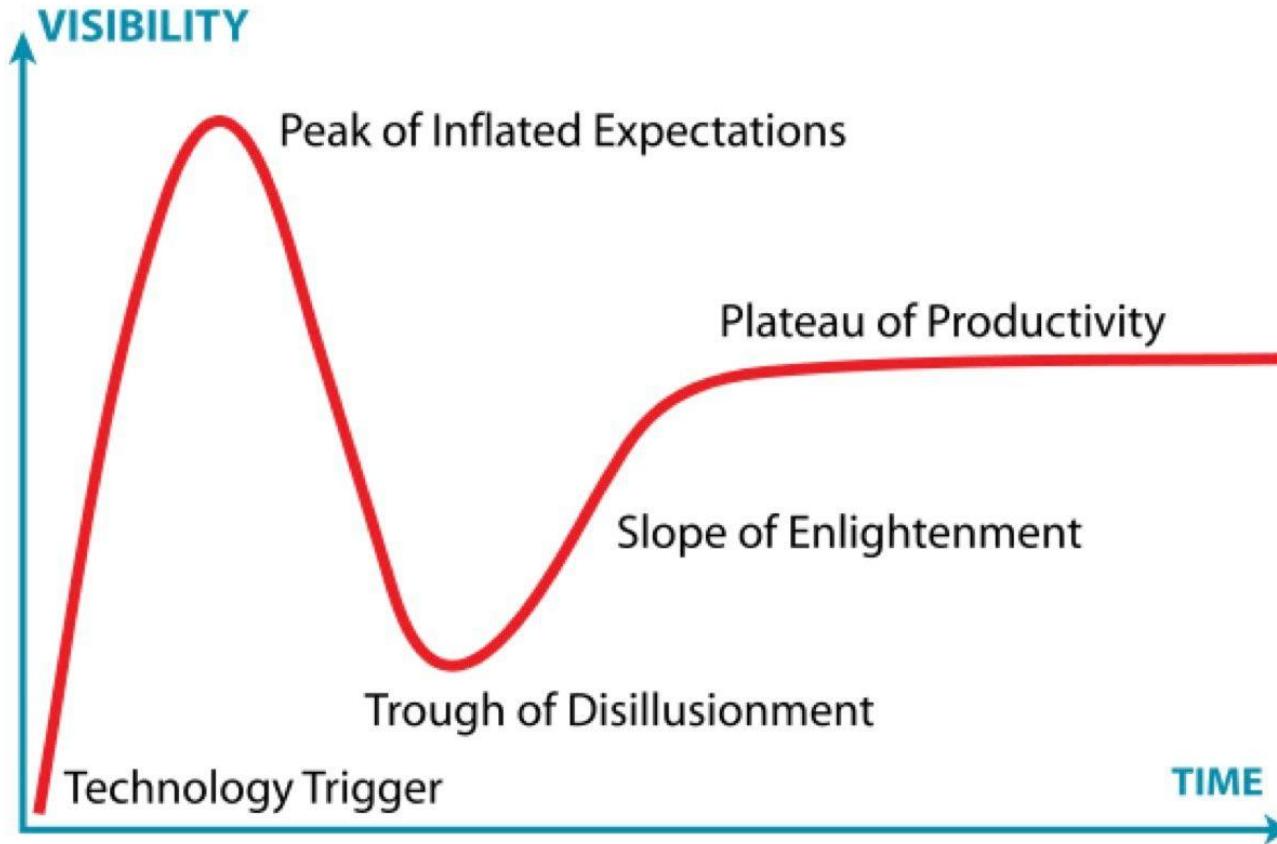
To nie rozmiar danych jest istotny, a podejście do ich analizy
A to już jest koncepcja znacznie starsza niż ostatni „hype”...



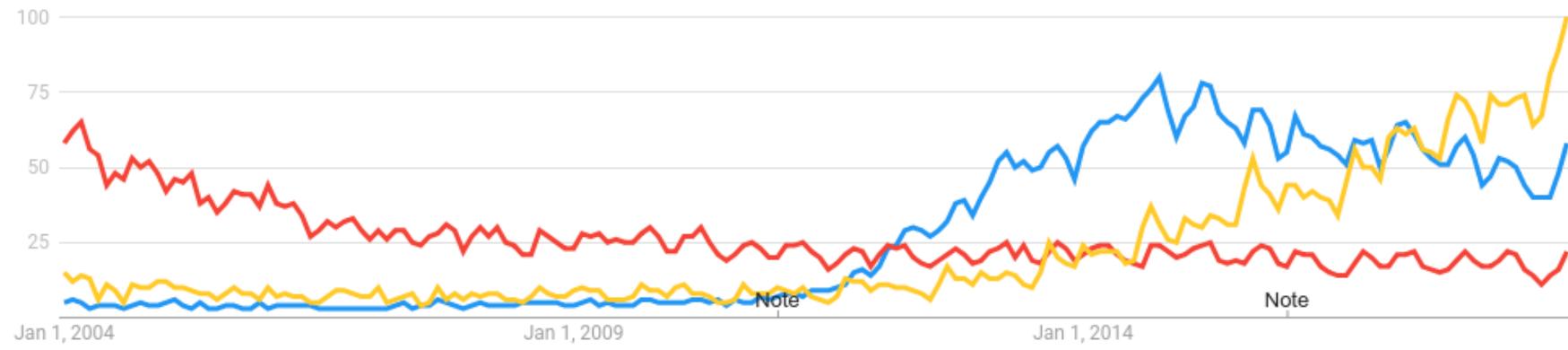
Big Data vs. Data Mining

Kwestie terminologiczne

Gartner Hype Cycle

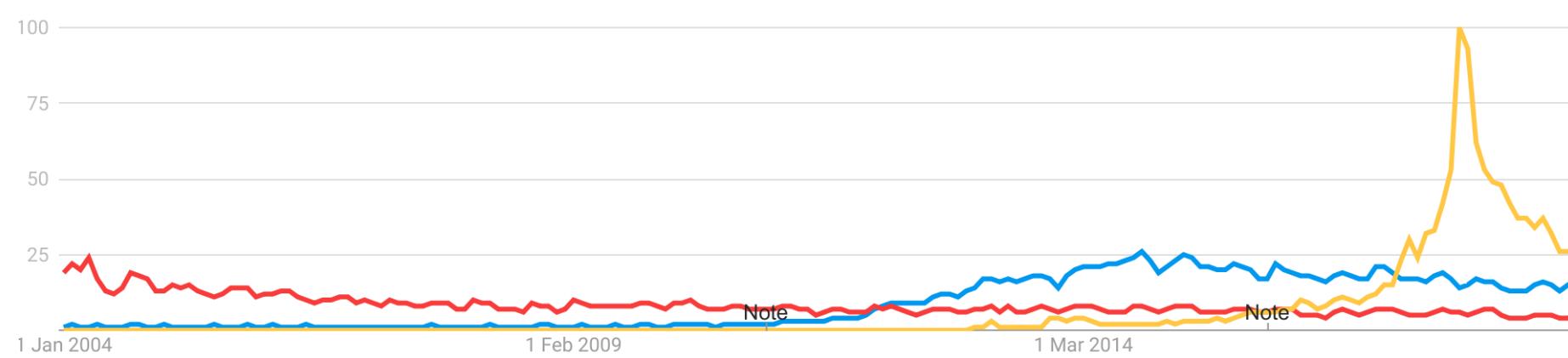


QUIZ1 – czym jest żółta linia?



Big Data vs. Data Mining vs. ?

QUIZZ – czym jest żółta linia?



Big Data vs. Data Mining vs. ?

OK, zatem skąd hype (teraz)?

Istotne zmiany technologii i procesów w ostatnich latach
także w sektorze finansowym:

chronologicznie:

- Zmiany w podejściu do retencji danych
- Wzrost możliwości przetwarzania danych
- Nowe metody / paradygmaty analizy danych
- Dostępność danych zewnętrznych (w stosunku do organizacji)

Retencja danych

Od późnych lat 90-tych wdrażane systemy hurtowni danych...

- W większości systemy oparte o bazy relacyjne
- Przede wszystkim systemy raportowania
- Prosta analityka (OLAP, data cubes itp.) lub jej brak
- Założenie – raz stworzone dane nie powinny ginąć

Współczesne lingo – „data lakes”

Kiedy to już jest „Big Data”?

- Big Data
- Data Science
- Data Mining

Acha! Czyli cały czas rozmawiamy o danych!

Czyli o... ???

*The quantities, characters, or symbols **on which operations are performed** by a computer, which may **be stored and transmitted** in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media*

Faktycznie dane mogą się różnić jeśli chodzi o te dwie podstawowe właściwości tzn. o sposób składowania oraz ich transmisję

Czyli chodzi nam o obliczenia komputerowe, dla których dane są podstawą

Pewnie również obraz, dźwięk itp
Jakby nie patrzeć to
000110100100010011111100

Czyli z jakiegoś powodu zmienia nam się sposób składowania i transmisji danych. Z jakiegóż to?

- a) tak, to prawda, dane są BIG ale..
- b) następuje rewolucyjny przeskok w podejściu do wykorzystania danych na potrzeby **analityczne i predykcyjne!!!**

Krótki (i niekompletny) rys historyczny z rynku internetowych reklam

- 27.10.1994 at&t kupuje pierwszy baner



- 1998 – pierwsza aukcja słów kluczowych używanych w czasie wyszukiwania
- 2000 – startuje Google AdWords
- 2004 – startuje Facebook

Krótki (i niekompletny) rys historyczny z rynku internetowych reklam

- 2007 – Facebook zaczyna targetowanie na podstawie **danych demograficznych** (.. i nie tylko. Również w 2007 powstaje Like)

Who should see this Flyer?

You are targeting about 100 women between 20 and 40 years old who are engaged in Salem, OR.

Location: United States choose cities

Salem, OR

Sex: Male Female

Age: 20 to 40

Keywords:

(interests, favorite music, movies, etc.)

Political Views: Liberal Moderate Conservative

Relationship Status: Single In a Relationship Engaged Married

Current Education Status: All High School College Alumni

Workplace:

How much are you willing to spend?

Max price per click: \$ / per click (min \$0.01)

Max daily budget: \$ / per day (min \$1.00)

Duration: Run my Flyer continuously starting today

Run my Flyer only during specified dates

A higher max price per click increases the chance your ad will be shown. We discount clicks on your behalf, so you may pay less than your max price depending on the current demand for your ad's audience. We will never bill you more in a day than your budget.

Krótki (i niekompletny) rys historyczny z rynku internetowych reklam

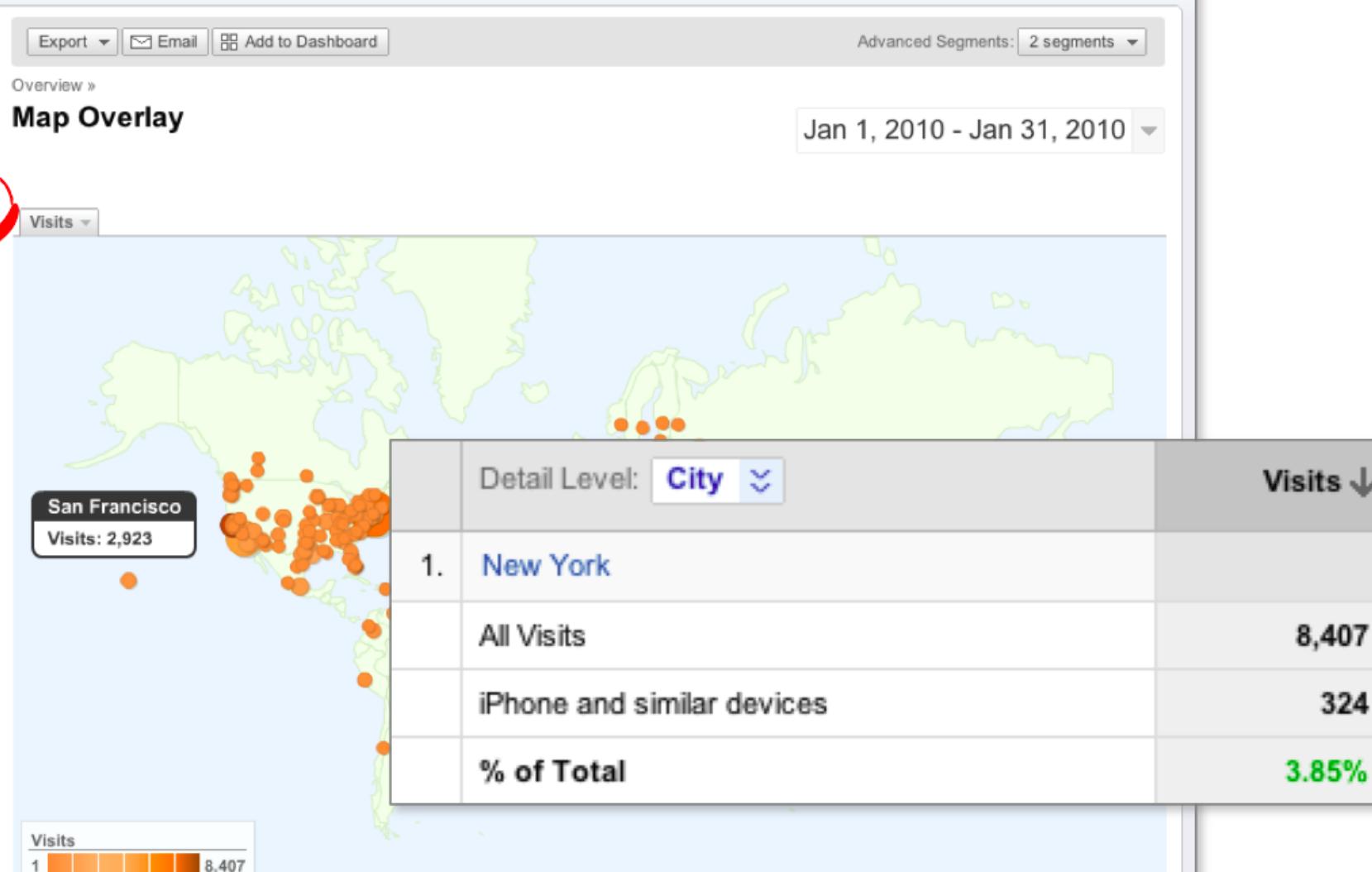
- ~2010 - Google rozpoczyna zwiększa efektywność targetowania na podstawie **danych geolokalizacyjnych**

[Analytics Settings](#) | [View Reports:](#) www.googlestore.comMy Analytics Accounts: [Google Store](#)

- [Dashboard](#)
- [Intelligence Beta](#)
- [Visitors](#)
 - [Overview](#)
 - [Benchmarking](#)
 - [Map Overlay](#)
 - [New vs. Returning](#)
 - [Languages](#)
 - [Visitor Trending](#)
 - [Visitor Loyalty](#)
 - [Browser Capabilities](#)
 - [Network Properties](#)
 - [Mobile](#)
 - [User Defined](#)
 - [Custom Variables](#)
- [Traffic Sources](#)
- [Content](#)
- [Goals](#)
- [Ecommerce](#)
- [Custom Reporting](#)

My Customizations

- [Custom Reports](#)



See all

About this Facebook ad



New Balance

Sponsored ·

Idealne do startu w r

LEPSZY
KIEDYKO

FRESH
FOAM
1080v5

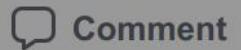
WYPRÓ

NBSKLEP.PL/1080

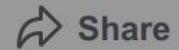
New Balance Fresh Foam 1080



Like



Comment



Share



Why am I seeing this ad?

Options ▾

One reason why you're seeing this ad is that **New Balance** wants to reach people interested in **Running**, based on activity such as liking Pages or clicking on ads.

There may be other reasons why you're seeing this ad, including that New Balance wants to reach **men aged 18 to 45 who live or have recently been in Poland**. This is information based on your Facebook profile and where you've connected to the Internet.

Let us know if this topic interests you

Running



Manage Your ad preferences

Tell us what you think

Was this explanation useful? Yes No

Learn more about Facebook Ads

English (UK) · English (US) · Polski ·
šlónsko gôdka · Español

Google, Facebook i wielu innych wie, że:

Kowalski jest mężczyzną, ma tyle lat ile ma i ma określone poglądy polityczne i mieszka w konkretnej lokalizacji.

Co z tego wynika?

W szerokiej skali zaczęto zauważać, że:

- Kowalski np. będąc klientem podmiotu A ma pewne preferencje zakupowe
- Jeśli połączymy dane z A, B, C, D to wiemy więcej o Kowalskim i jego preferencjach
- Jeśli popatrzymy co Kowalski robi w wolnym czasie (np. co czyta) to wiemy jeszcze więcej
- Jeśli połączymy dane o Kowalskim z danymi o jego znajomym Nowaku to dowiemy się więcej o nich obu
- Jeśli zobaczymy co kupują sąsiedzi Kowalskiego to dowiemy się więcej zarówno o nim jak i o samych sąsiadach
- .. i wiele, wiele więcej

I tak rozpoczęła się nowa era zbieractwa danych, których, tak to prawda, jest dużo.

A tak na marginesie.. zaczyna się robić tyleż ciekawie co groźnie

Metody:

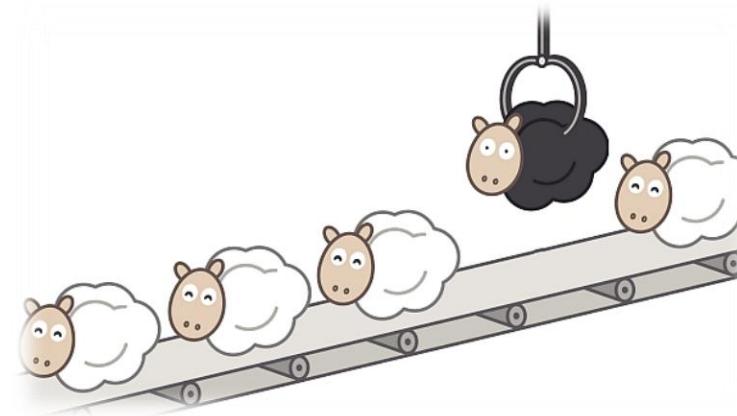
- Jeśli chcemy poznać preferencje stosujemy **algorytmy wykrywania wzorców częstych**.
- Jeśli chcemy “przełożyć” tekst na język komputera (czyli np. “co czyta Kowalski”) to stosujemy **metody NLP** (ang. natural language processing).
- Jeśli szukamy podobieństw w ramach danej populacji stosujemy **algorytmy klasteryzacji**.
- Jeśli szukamy jakieś cechy Kowalskiego a znamy tę cechę dla szerszej populacji i posiadamy inne cechy zarówno populacji jak i Kowalskiego to stosujemy **algorymy klasyfikacji**.

A to przecież algorytmy z dziedziny uczenia maszynowego/data mining!

Kilka definicji

“Uczenie maszynowe (ang. machine learning) jest analizą procesów uczenia się oraz tworzeniem systemów, które doskonalą swoje działanie na podstawie doświadczeń z przeszłości.” [Stefanowski 2009]

Kilka definicji



Uczenie z nadzorem (ang. supervised learning) – odkrywanie zależności (funkcyjnej) z danych oznaczonych (ang. labeled data). Na podstawie wektora wejściowego *zmiennych objaśniających* próbujemy przewidzieć wyjściowy wektor *zmiennych objaśnianych*.

Przykładowe zadania klasyfikacja/regresja.

Kilka definicji

Przykład:

Zmienne objaśniające:

- dane transakcyjne klientów korporacyjnych
- dane zewnętrzne z KRS, wywiadowni gospodarczych
- dane pozyskane z serwisów społecznościowych (np. #Microsoft, #Comarch)

Zmienna objaśniana:

- zdarzenie default'u

Kilka definicji

Uczenie bez nadzoru (ang. unsupervised learning) – odkrywanie zależności z danych nieoznaczonych (ang. unlabeled data).

Przykładowe zadanie: klasteryzacja

Uczenie online i offline

Kilka definicji

Przykład: chcemy poznać grupy podobnych klientów korporacyjnych w celu lepszego targetowania oferty

Dane wejściowe:

- Zagregowane dane dotyczące cashflowu (również w zależności od źródeł, okresów w ciągu roku itp)
- Linie kredytowe, spłacałość
- Płatności wynagrodzeń dla pracowników
- Opłaty za media, aktywa trwałe (?) itd.

Big Data vs Data Mining

- „*So in a sense we are talking about **distributed data mining or machine learning**,*”
- Big data ≈ dane, które są większe niż pojemność pojedynczego hosta
- Ok – gdzie zatem leży problem?
- Jednak najpierw – jakie są potencjalne korzyści:
 - Równoległe ładowanie danych oraz zrównoleglone obliczenia
 - Fault tolerance wpisany w architekturę systemów
- Z drugiej jednak strony synchronizacja i komunikacja nie jest tak prosta.

Multiplying matrices in R

- > z <- matrix(1:9000000, ncol = 3000, nrow = 3000)
- > system.time(z %*% z)
- użytkownik system upłynęło
- 20.43 0.06 20.64

sekund

... whereas in Java/Scala

```
val t0 = System.nanoTime()

for(i<-0 to matrix.length-1){
    for(j<-0 to matrix(i).length-1){
        for(k<-0 to matrix(i).length-1){
            matrix0(i)(j) = matrix(i)(k) * matrix(k)(j)
        }
    }
}
val t1 = System.nanoTime()
println("Elapsed time: " + (t1 - t0) + "ns")
}
```

```
/usr/lib/jvm/java-8-oracle/bin/java ...
```

```
Elapsed time: 334855288657ns
```

~ 5 min 30s

```
Process finished with exit code 0
```

Computer architecture meets mathematics

execute typical instruction	1/1,000,000,000 sec = 1 nanosec
fetch from L1 cache memory	0.5 nanosec
branch misprediction	5 nanosec
fetch from L2 cache memory	7 nanosec
Mutex lock/unlock	25 nanosec
fetch from main memory	100 nanosec
send 2K bytes over 1Gbps network	20,000 nanosec
read 1MB sequentially from memory	250,000 nanosec
fetch from new disk location (seek)	8,000,000 nanosec
read 1MB sequentially from disk	20,000,000 nanosec
send packet US to Europe and back	150 milliseconds = 150,000,000 nanosec

- BLAS: basic linear algebra subprograms

- Level 1

- Level 2

- Level 3

From: <http://norvig.com/21-days.html#answers>

Spark Issues

Spark / SPARK-6442

MLlib Local Linear Algebra Package

Agile Board

Details

Type:	 New Feature	Status:	OPEN
Priority:	 Critical	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	MLlib		
Labels:	None		

Description

MLlib's local linear algebra package doesn't have any support for any type of matrix operations. With 1.5, we wish to add support to a complete package of optimized linear algebra operations for Scala/Java users.

The main goal is to support lazy operations so that element-wise can be implemented in a single for-loop, and complex operations can be interfaced through BLAS.

The design doc: <http://goo.gl/sf5LCE>

... rok później

 [Spark](#) / [SPARK-15882](#)
Discuss distributed linear algebra in spark.ml package

Agile Board

Details

Type:	<input checked="" type="checkbox"/> Brainstorming	Status:	OPEN
Priority:	↑ Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	ML		
Labels:	None		

Description

This JIRA is for discussing how `org.apache.spark.mllib.linalg.distributed.*` should be migrated to `org.apache.spark.ml`.

Initial questions:

- Should we use Datasets or RDDs underneath?
- If Datasets, are there missing features needed for the migration?
- Do we want to redesign any aspects of the distributed matrices during this move?

Problemy leżą nie tylko w warstwie rozproszenia obliczeń dotyczących podstawowych operacji algebraicznych.

Większość algorytmów ML nie została zaprojektowana pod kątem optymalizacji w środowiskach rozproszonych!

W szczególności algorytmy ML nie były optymalizowane do korzystania z procesorów połączonych ze sobą relatywnie wolnym dostępem sieciowym.

Podobnie jak w problemie mnożenia macierzy,
musimy przeanalizować dwie kwestie:

- Matematyczne założenia algorytmu
- Architekturę systemową

Chcemy minimalizować dostęp do danych i
komunikację między hostami.

Stawiamy na **near data processing**

Spark Issues

 [Spark](#) / [SPARK-3717](#)
DecisionTree, RandomForest: Partition by feature

Agile Board

Details

Type:	 Improvement	Status:	OPEN
Priority:	 Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	MLlib		
Labels:	None		

Description

Summary

Currently, data are partitioned by row/instance for DecisionTree and RandomForest. This JIRA argues for partitioning by feature for training deep trees. This is especially relevant for random forests, which are often trained to be deeper than single decision trees.

Chwilę o technologiach

- SQL on cluster



- Hive
- Impala
- Spark SQL



- Other tools / frameworks / programming languages:
 - Spark
 - R (shiny, sparkr, dplyr, randomforest, deepnet, **h2o** etc.)
 - Play Framework
 - Java Script presentation libraries: Leaflet, d3js
 - Scala

Storage tylko pozornie jest tani



VS



Storage tylko pozornie jest tani

“Managing all this information is no longer a storage problem – it’s about how well we can manage, harness, and govern that information”



Kiedy stosować Big Data?

“I have seen people insisting on using Hadoop for datasets that could easily fit on a flash drive and could easily be processed on a laptop.”

Yann LeCun (ten pan od deep learningu)

“Never Underestimate the Power of a Single Node”
“Distributed ≠ Fast”

Karthik Ramasamy, Uber

Kiedy stosować Big Data?

Tak.. czyli kiedy?

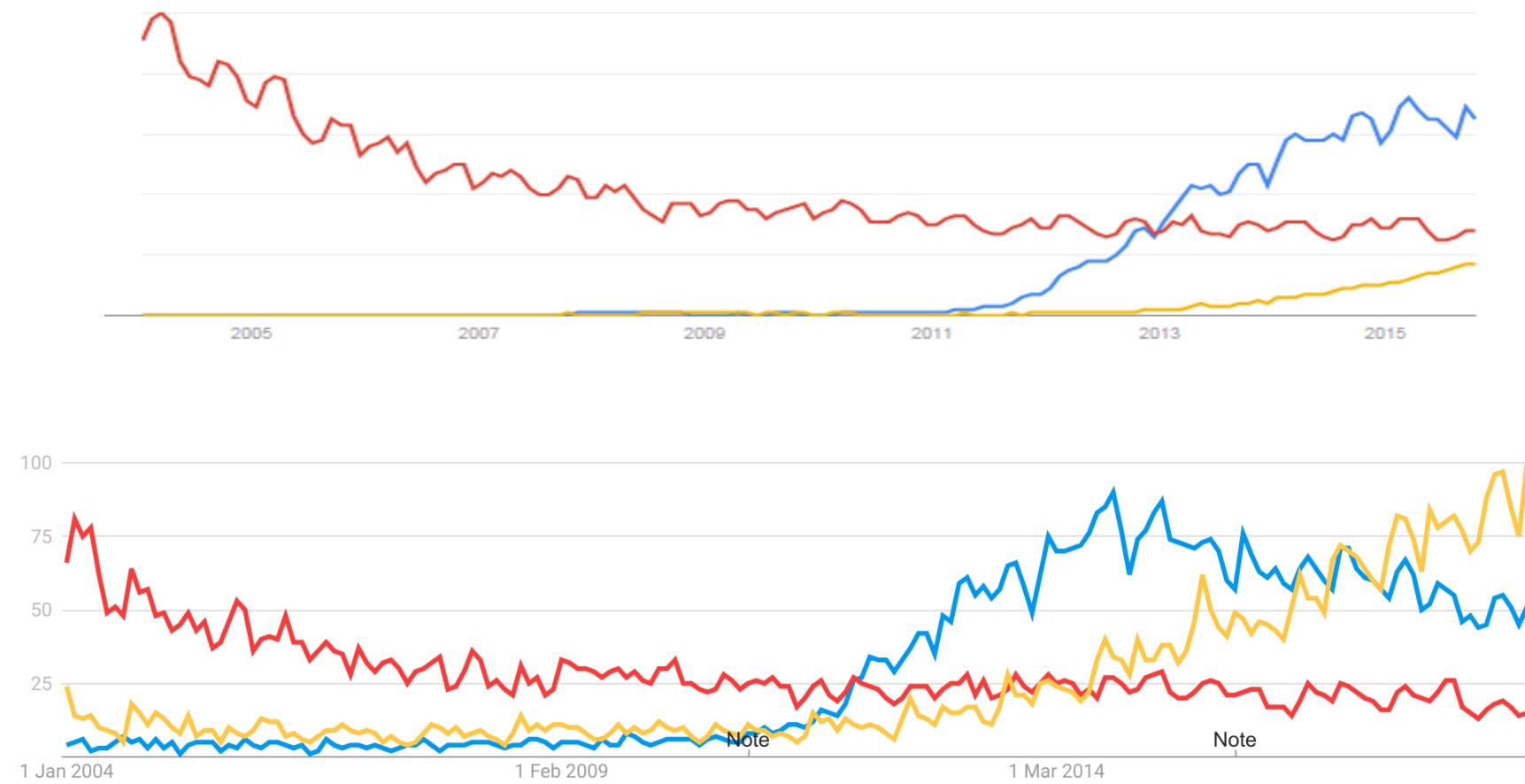
Wtedy, gdy **zawiódły**_(by) inne metody analityczne.

Niewielki rozmiar danych jest tylko i wyłącznie
sygnałem!

Kiedy dla zbioru danych wielkości pendrive'a
chcielibyśmy tworzyć klaster?

Data Science!

Data Science – „eksploracyjne podejście do analizy danych”



Big Data vs. Data Mining vs. Data Science

Data Science

„Why and how” is the question, not „how much”

Big Data → „how much”?

Data Science → „why and how”?

Zadawanie pytań „dlaczego” oraz „jak” nie jest łatwe w banku!

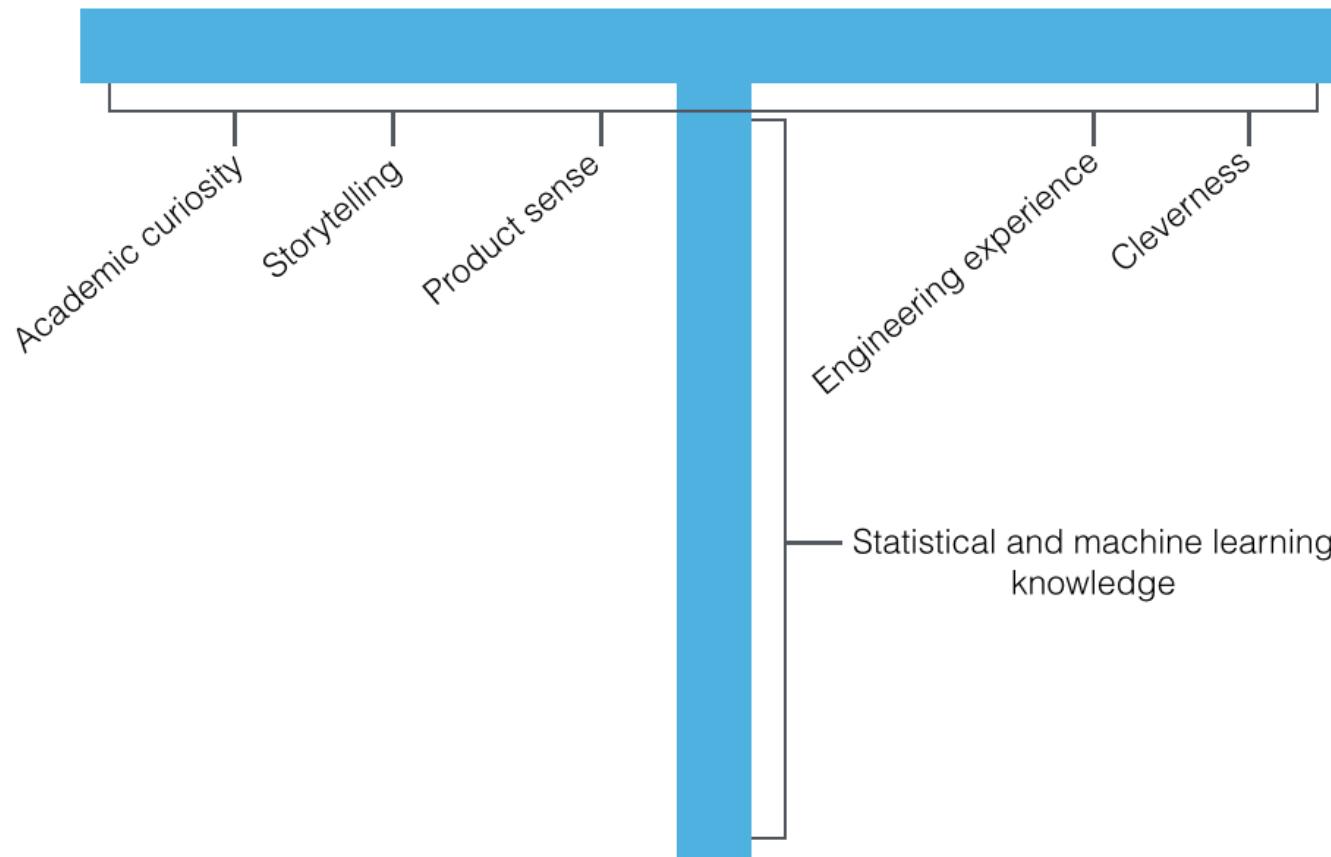
Może to wymagać sięgnięcia po dane zewnętrzne oraz niestandardowe narzędzia analityczne.

Nawet najprostsza analiza wymaga:

- czyszczenia i uzupełniania danych (np. adresy vs kody pocztowe vs ...)
- pozyskania informacji dotyczących właściwości administracyjnych źródła
- przedstawienie danych geo w postaci tabeli jest nietrywialne ...

Data Scientist - ... a kóż to taki?

„The key word in Data Science is not Data, it is Science”



Data Scientist - ... a ktoś to taki?



Michael E. Driscoll
@medriscoll

 Follow

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu

[@peteskomoroch](#)

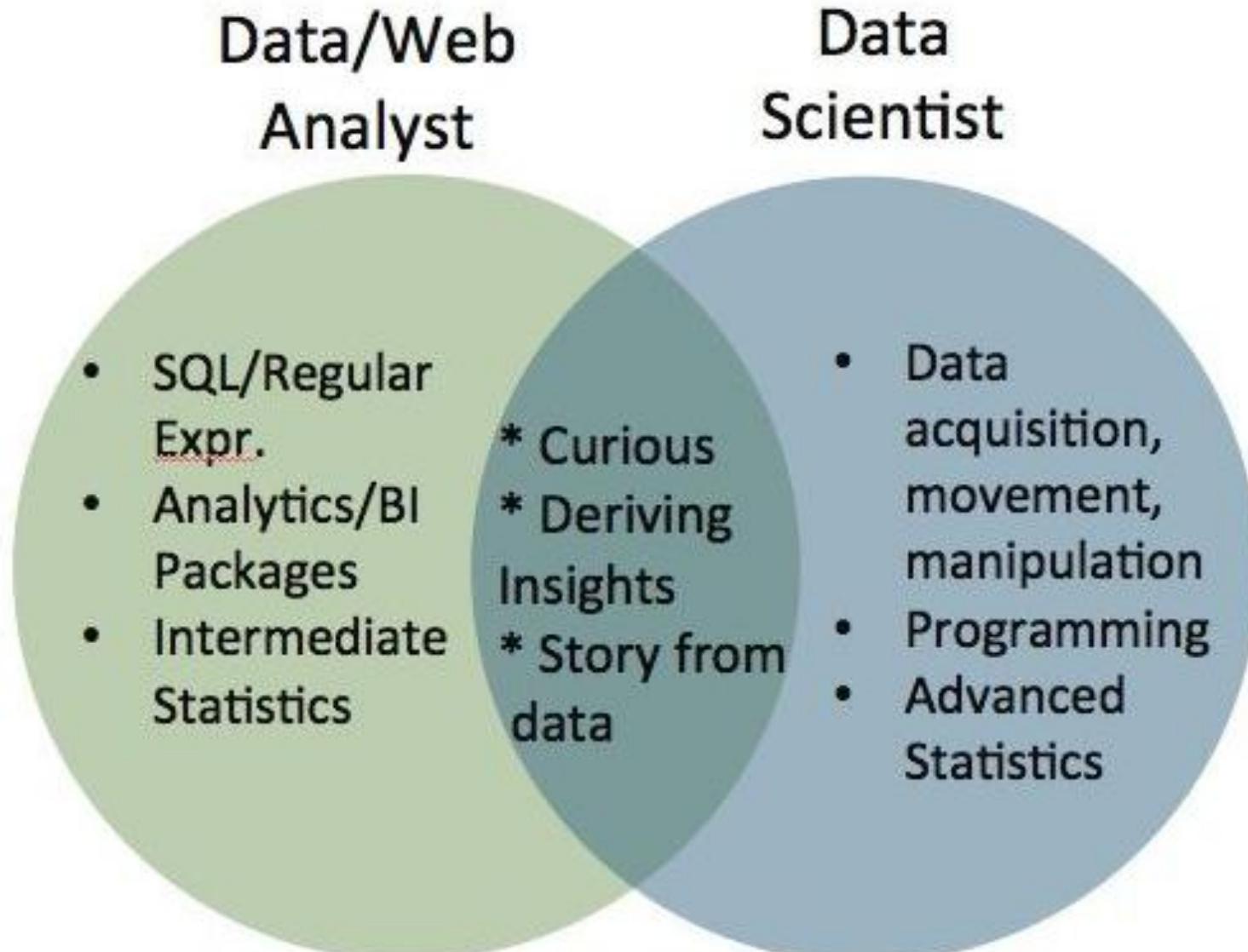
10:57 PM - 17 Jul 2012



 44

 43

Data Scientist - ... a kóż to taki?





A co na to prawo?

- Dane osobowe i dane wrażliwe (pochodzenie rasowe, etniczne; poglądy; religia; życie seksualne; nałogi; skazania i orzeczenia sądowe)
- Przechowywanie danych danych wrażliwych jest niedozwolone (poza wyjątkowymi przypadkami za pisemną zgodą)

no tak ale.. p. Kowalski realizuje transakcje w sklepie monopolowym średnio 6 razy w tygodniu → czy można się zastanowić co z tego wynika??

NIE



A co na to prawo?

Jako administrator danych musimy zapewnić aby dane były:

- przetwarzane zgodnie z prawem
- zbierane dla oznaczonych, zgodnych z prawem celów i niepoddawane dalszemu przetwarzaniu niezgodnemu z tymi celami (z zastrzeżeniem)
- merytorycznie poprawne i **adekwatne** w stosunku do celów, w jakich są przetwarzane
- przechowywane w postaci umożliwiającej identyfikację osób, których dotyczą, nie dłużej niż jest to niezbędne do osiągnięcia celu przetwarzania



A co na to prawo?

Zastrzeżenie:

w celach badań naukowych, dydaktycznych, historycznych lub statystycznych

możemy dalej przetwarzać jeśli zastosujemy np. anonimizację.



O anonimizacji słów kilka...

- 87% populacji USA można zidentyfikować wyłącznie po {kod pocztowy, płeć, data urodzenia} $\leftarrow \sim\{43k, 2, \sim30 \times 365\} \sim 941$ mln możliwości
- 53% populacji USA można zidentyfikować po {miasto, płeć, data urodzenia}
- 18% populacji USA można zidentyfikować po {hrabstwo, data urodzenia, płeć}



O anonimizacji słów kilka...

Casus Netflix'a:

- Mamy dane użytkowników. Chcielibyśmy bardziej adekwatnie rekommendować filmy.
- Zróbmy otwarty konkurs!
- Dane wejściowe <user, movie, date of grade, grade>
- Oczywiście *user* został zastąpiony unikalnym ID
- Dla każdego użytkownika zmieńmy pewien % ocen (wprowadźmy szum) ... ale przecież za dużo nie możemy, prawda?

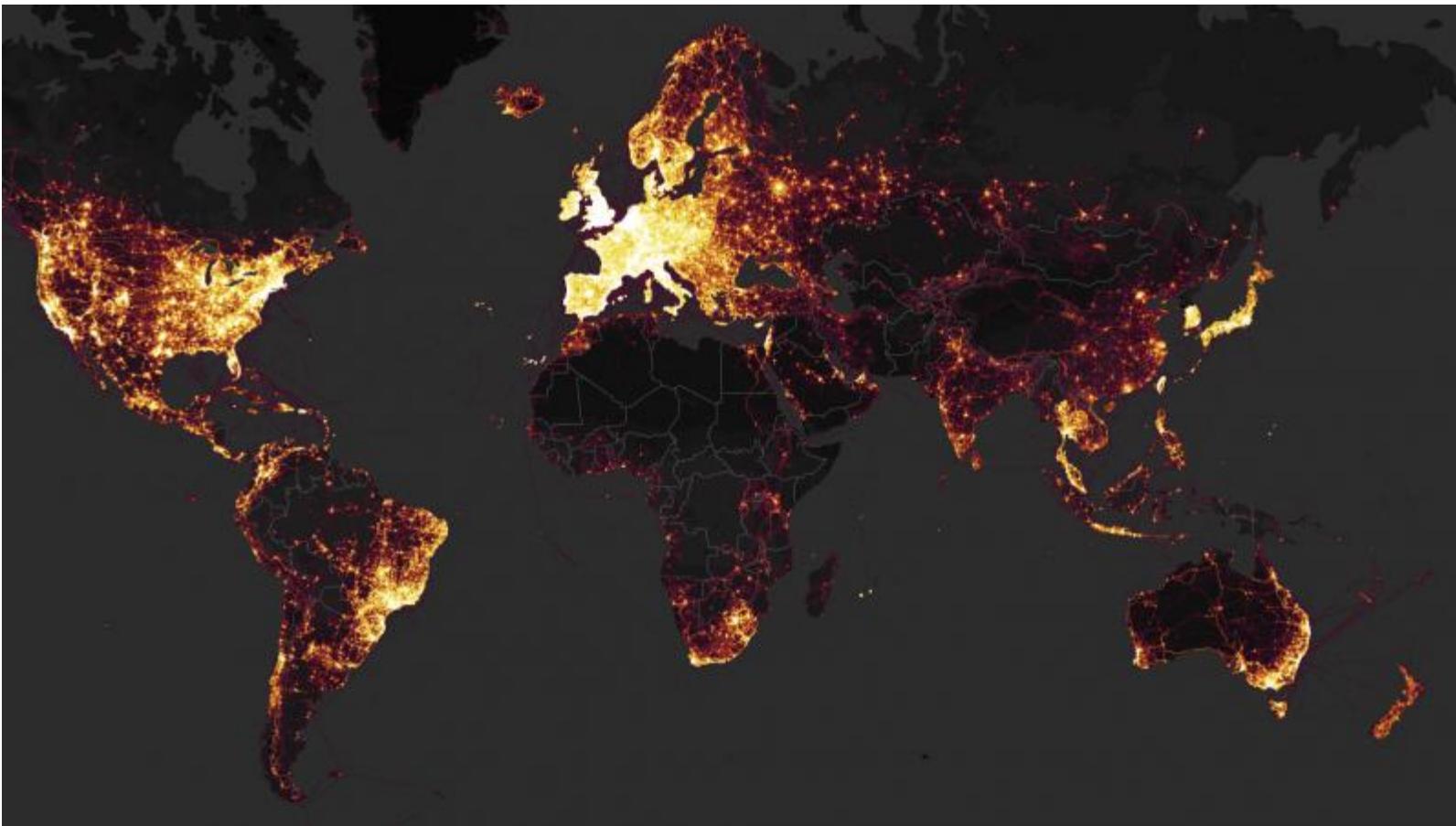
Problemy?

- Nie tylko Netflix umożliwia ocenianie filmów
- Okazuje się, że (prawie) każdy z nas oglądał Forresta Gumpa (i innych 99 popularnych filmów) ale każdy też ma unikalny zbiór mało popularnych produkcji.

W 2010 Netflix nie decyduje się na kolejną edycję konkursu.



O anonimizacji słów kilka...



<https://www.technologyreview.com/s/610090/stravas-privacy-pr-nightmare-shows-why-you-cant-trust-social-fitness-apps-to-protect-your/>



GDPR

W maju 2018 r. weszło w życie rozporządzenie w sprawie ochrony danych osobowych.

Zasdatnicza zmiana w stosunku do obecnych przepisów → m.in. dla banków, które posiadają szerokie możliwości przechowywania i analizy danych (np. specyficzna podgrupa “kredytobiorcy” na podstawie ustawy Prawo Bankowe)



GDPR

GDPR (ang. general data protection regulation) w pigułce:

1. Bezpośrednio stosowane.
2. UODO może nakładać kary do 4% globalnego obrotu firmy lub 20 mln Euro.
3. Obowiązek informacyjny UODO oraz indywidualnie osób, których dane zostały naruszone.
4. Osobna zgoda na każdy z celów przetwarzania danych (opt-in)
5. Prawo dostępu do danych, ich sprostowania oraz całkowitego i trwałego usunięcia (!) (a co z logami?)



A co na to prawo?

„**profilowanie**” oznacza dowolną formę zautomatyzowanego przetwarzania danych osobowych, które polega na wykorzystaniu danych osobowych do oceny niektórych czynników osobowych osoby fizycznej, w szczególności do analizy lub prognozy aspektów dotyczących efektów pracy tej osoby fizycznej, jej **sytuacji ekonomicznej, zdrowia, osobistych preferencji, zainteresowań, wiarygodności, zachowania, lokalizacji lub przemieszczania się**

Metadata

Data that provides information about other data.

Metadane są mniej istotne?



EXIF – metadane zaszyte niemal w każdym zdjęciu:

- data i czas
- lokalizacja
- typ aparatu
- itd



Hipokryzja?

“You want to look through target’s eyes?”

“You have to hack your target....

While your target is Browsing the web, Exchanging documents, Receiving documents”

Inni “klienci”: Sudan, Bahrain, Venezuela, Saudi Arabia, ... USA

<http://www.hackingteam.it/solutions.html>

Klaster Hadoop

To nie jest „magic bullet”

- algorytmy przeznaczone dla klastra muszą być napisane od nowa (podejście Map/Reduce; technologia RDD Spark) – choć istnieją gotowe biblioteki (Mlib – ale patrz poniżej)
- certyfikacja instalacji powoduje problemy (np. wersje oprogramowania)
- bank nie może korzystać ze skalowalnych serwisów zewnętrznych (np. EC2)
- to nie jest tania ani prosta w konfiguracji inwestycja

Niemniej jednak np. :

System analizy sekwencji zdarzeń w procesie kredytowym (zwykle około 30 tyś. klientów, m.in. FP-Growth, lasy losowe Breimana)

- serwer R @DHD - 48 godzin
- klaster Hadoop @DHD - 15 minut

Dane w mBanku...

Dane przechowywane w systemach IT Banku ~ 0.8 PB

Przetwarzanie surowych danych ~ 50 GB dziennie

Dziesiątki milionów operacji dziennie

Hurtownia danych mBanku w zasadzie nadąża za systemami transakcyjnymi (oczywiście nie „za darmo” - konieczna architektura Lambda - np. dla systemów Real Time Marketing, zrównoleglenie procesów ETL itd.)

Volume - ✓ Velocity - ✓

Big Data in mBank

If you look at the numbers alone...

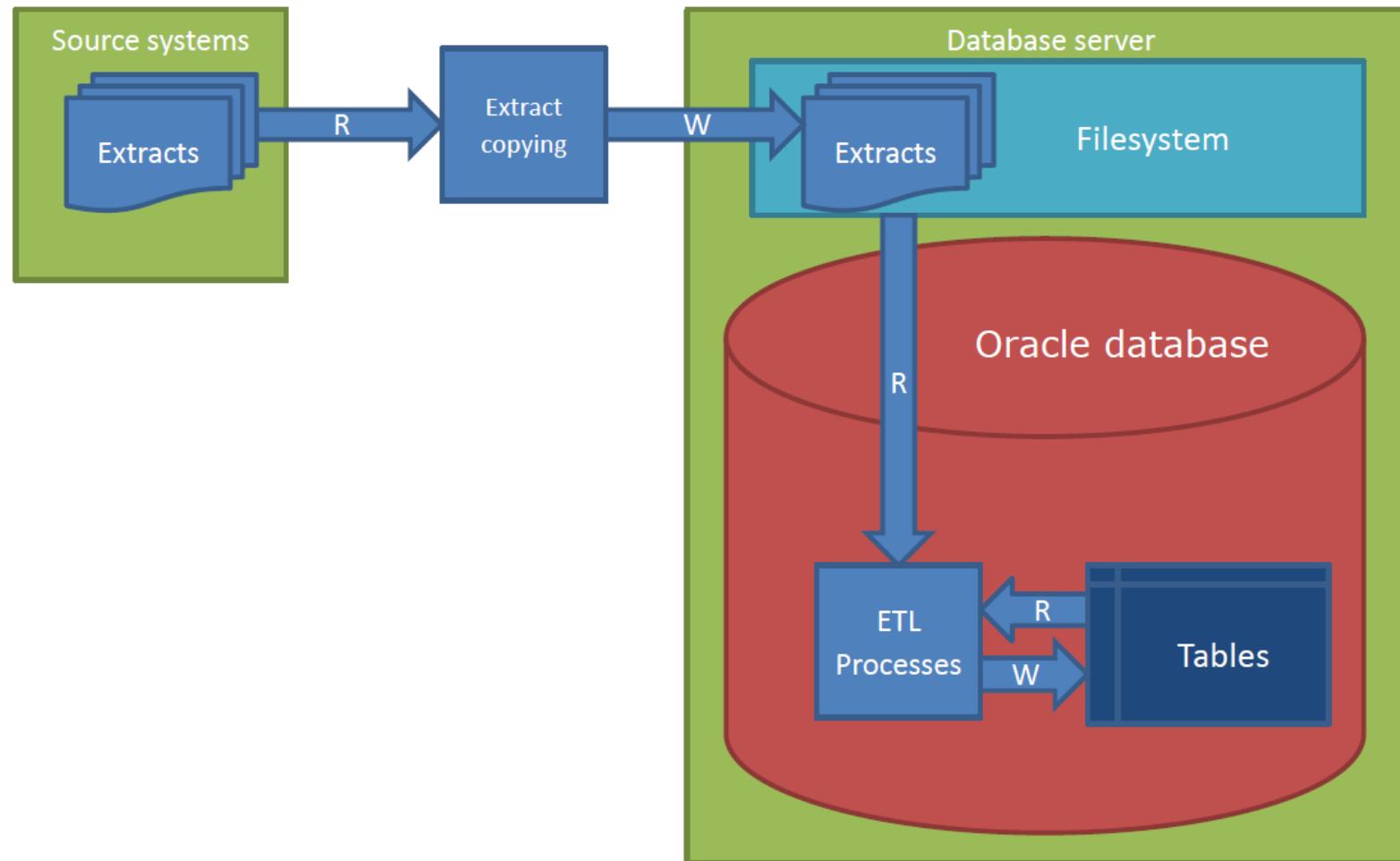
Typical Hadoop (Cloudera) cluster configuration @mBank

	Node type	Number	Cores	RAM (GB)	Storage (TB)
	Master node	9	180	1152	54
	Gateway	4	48	128	4
	Data node	27	648	6912	648
	Sum	40	876	8192	706

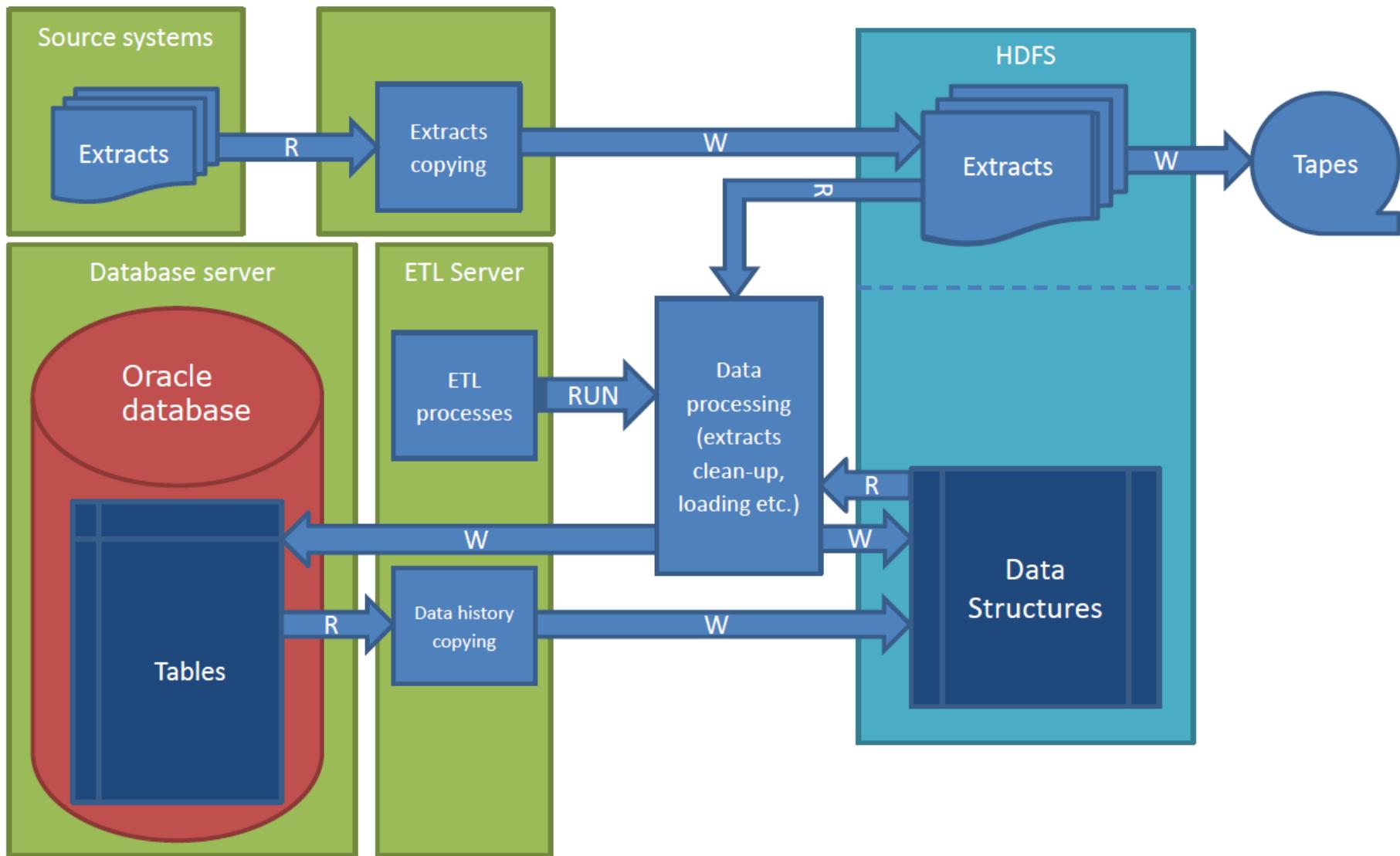
- Typical processing workloads per system: from tens of GB to couple of TB raw data daily
- Typical speedups resulting from parallel processing (e.g. ML on Spark vs. ML in SAS/R) – 1-2 orders of magnitude (e.g. from 24hrs down to 10 minutes)

... but where is this „Big Data” insight!?

Hurtownia danych – tak było



Klaster Hadoop – tak będzie/jest



Klaster Hadoop



Dane zewnętrzne

- media społecznościowe (kredyt „na hipstera” Facebook, Instagram lub „na profesjonalistę” LinkedIn) – ... ale compliance, trudności interpretacyjne, szerokie pole dla wszelkiego rodzaju fraudów
- platformy e-commerce – ... ale compliance (casus Aliora), ryzyko fraudów, niepewność co do wielkości obrotów oraz liczby kontrahentów etc. [a co z B2B?]
- usługi geokodowania (OpenLS w GUGiK)
- rejesty sądowe oraz ewidencje (KRS, CEIDG)

Dane zewnętrzne

Not so sexy...

Wywiadownie gospodarcze (BIG Info Monitor, BIK, KRD,
Deltavista i inne) – dla losowej grupy kredytobiorców

1.4% niespłaconych kredytów posiadało **pozytywny** wpis w
jednej z baz

oraz

2.5% spłaconych kredytów posiadało **negatywny** wpis w
jednej z baz

Cross-check konieczny!

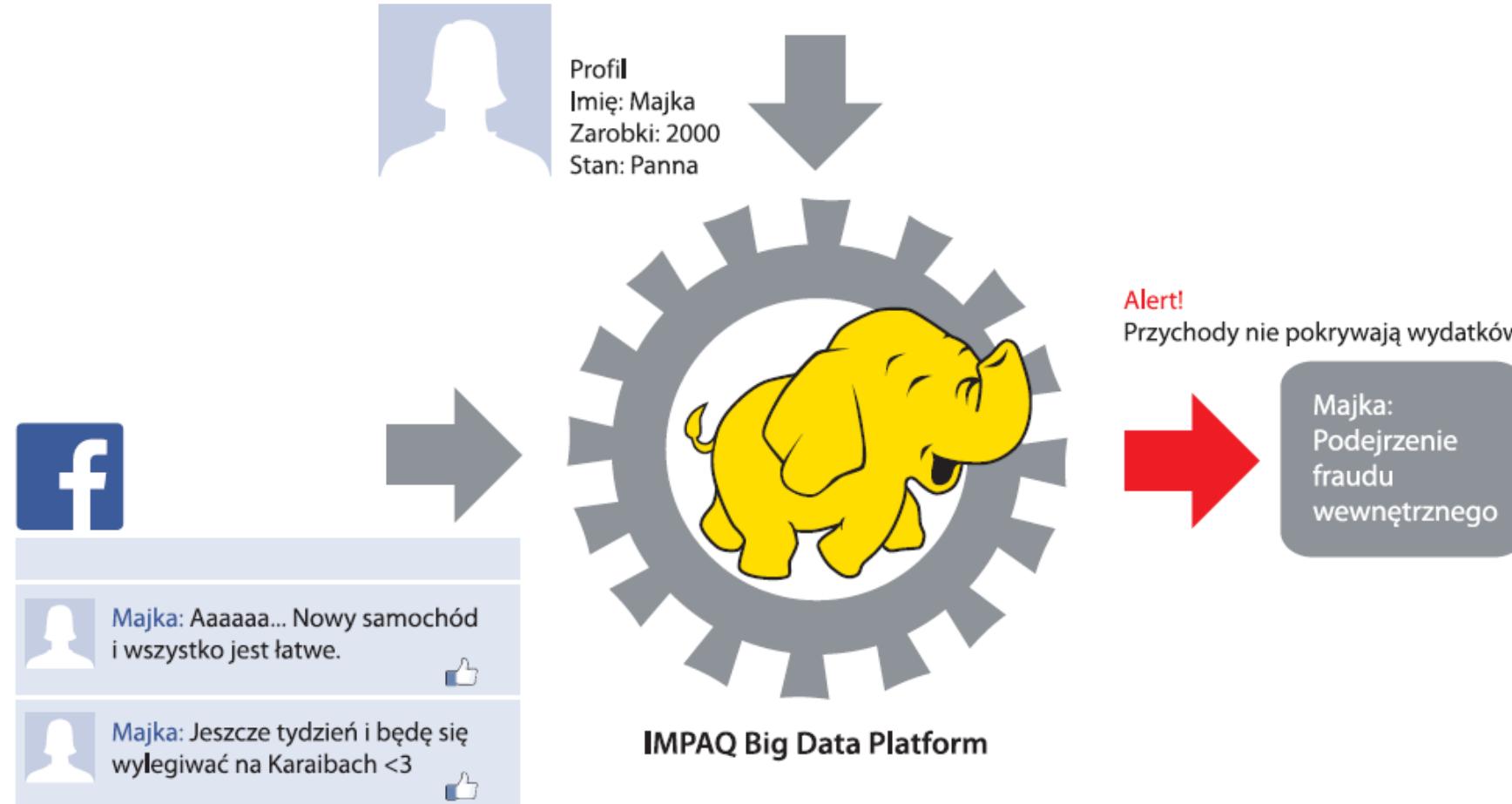
A propos bullshit...

źródło: http://www.impaqgroup.com/fileadmin/user_upload/Flyers/flyers_POL/Flyer_Big_Data_04_02_14_v_1.0.pdf

Np. IMPAQ

Big Data w praktyce

Majka pracuje w Banku. Zarabia 2000 PLN miesięcznie. Na swoim profilu na FB, w krótkich odstępach czasu, umieściła informację o kupnie nowego samochodu oraz wycieczce na Karaiby.



Przeanalizowanie wpisów na portalu społecznościowym i skonfrontowanie ich z informacjami pozyskanyimi z wewnętrznej bazy Banku pozwala zidentyfikować potencjalnego sprawcę nadużyć.



Majka: Aaaaaaa... Nowy samochód
i wszystko jest łatwe.



...semantyka

I podobnie... "W
Polsce kobieta rodzi
dziecko co 15
minut" <- znajdźmy
ją i zatrzymajmy!





Majka: Aaaaaaa... Nowy samochód
i wszystko jest łatwe.



...ironia



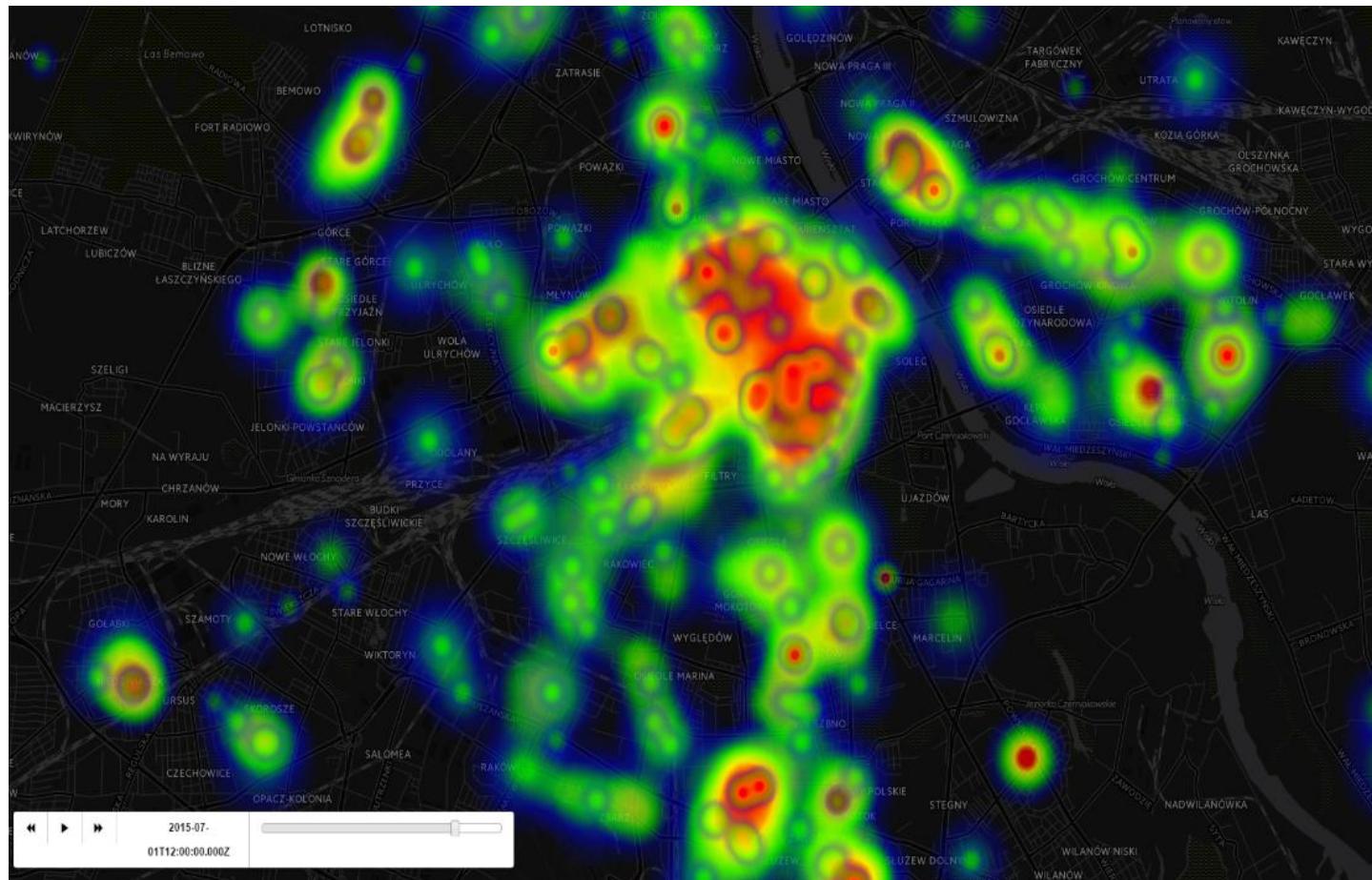
Wykorzystanie danych banku

Np. narzędzie analizy operacji ATM – not very sexy...



Wykorzystanie danych banku

Np. narzędzie analizy operacji ATM – sexy!!!

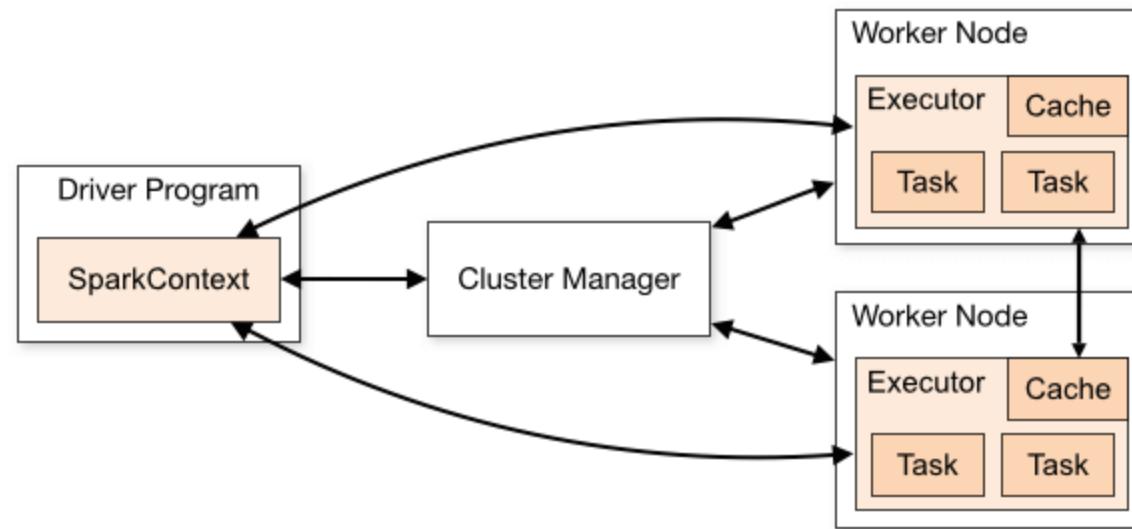


Problem



Features:

- Analytics
- Graphics & Visualization
- Libraries (cran.r-project.org)

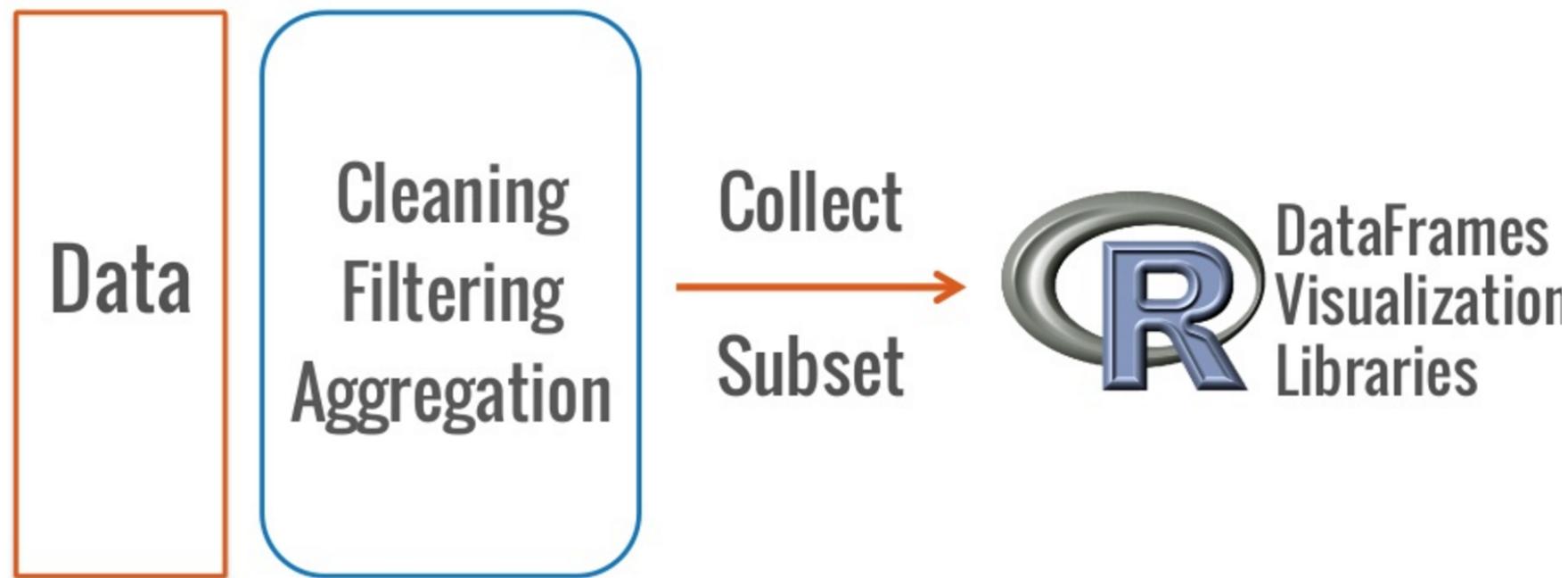


Solution - SparkR

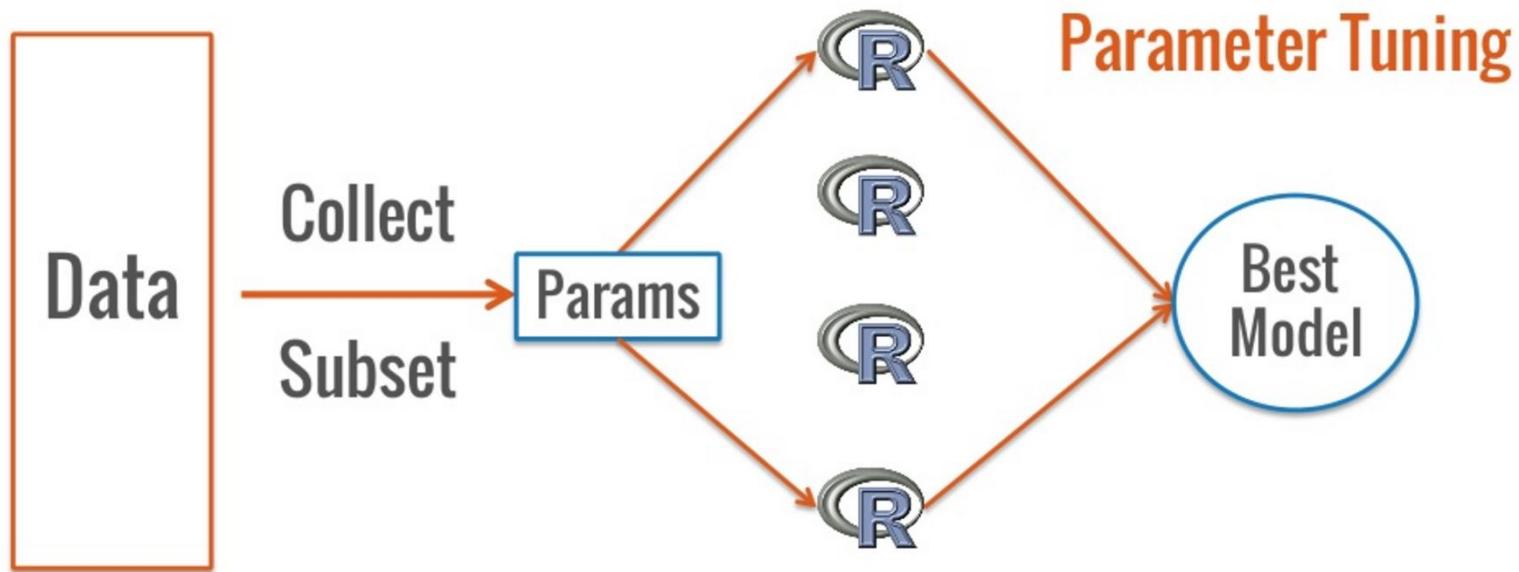
- Applications:
 - Big Data → small learning
 - partition aggregation
 - Large scale machine learning

source: SparkR: The Past, the Present and the Future, Spark Summit 2015

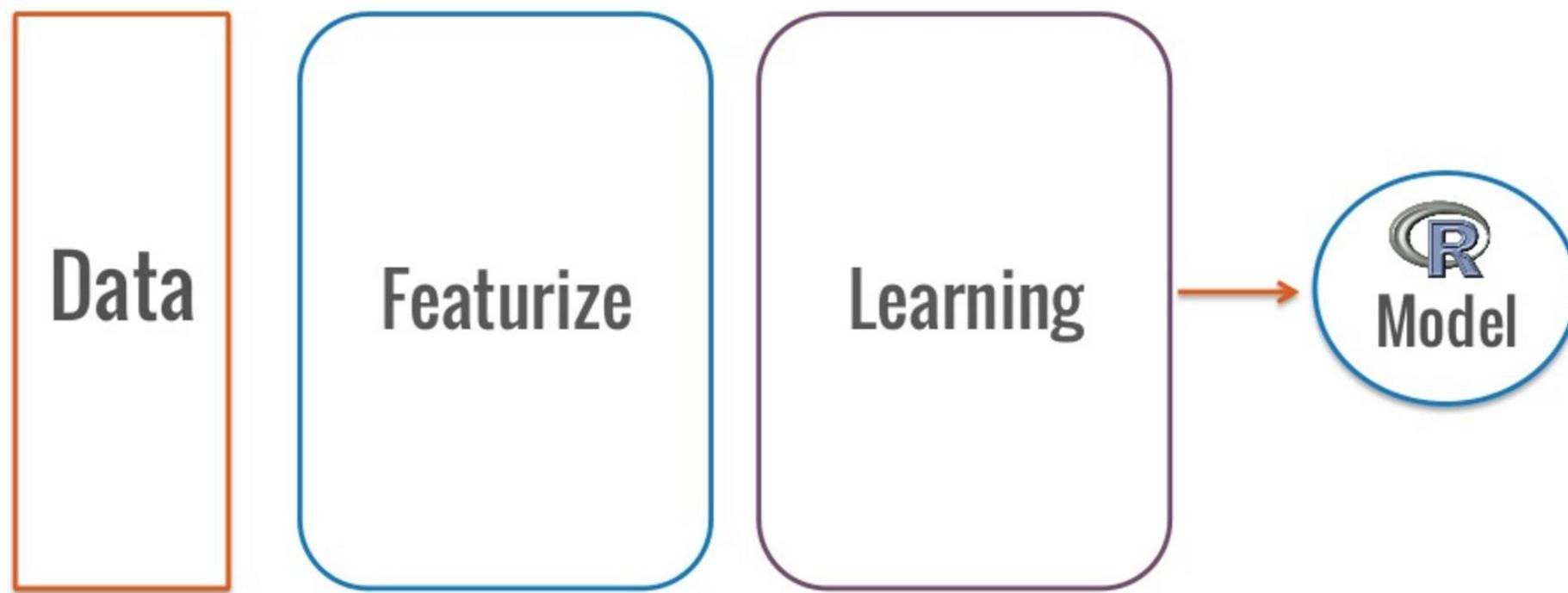
Big Data → small learning



Partition aggregation

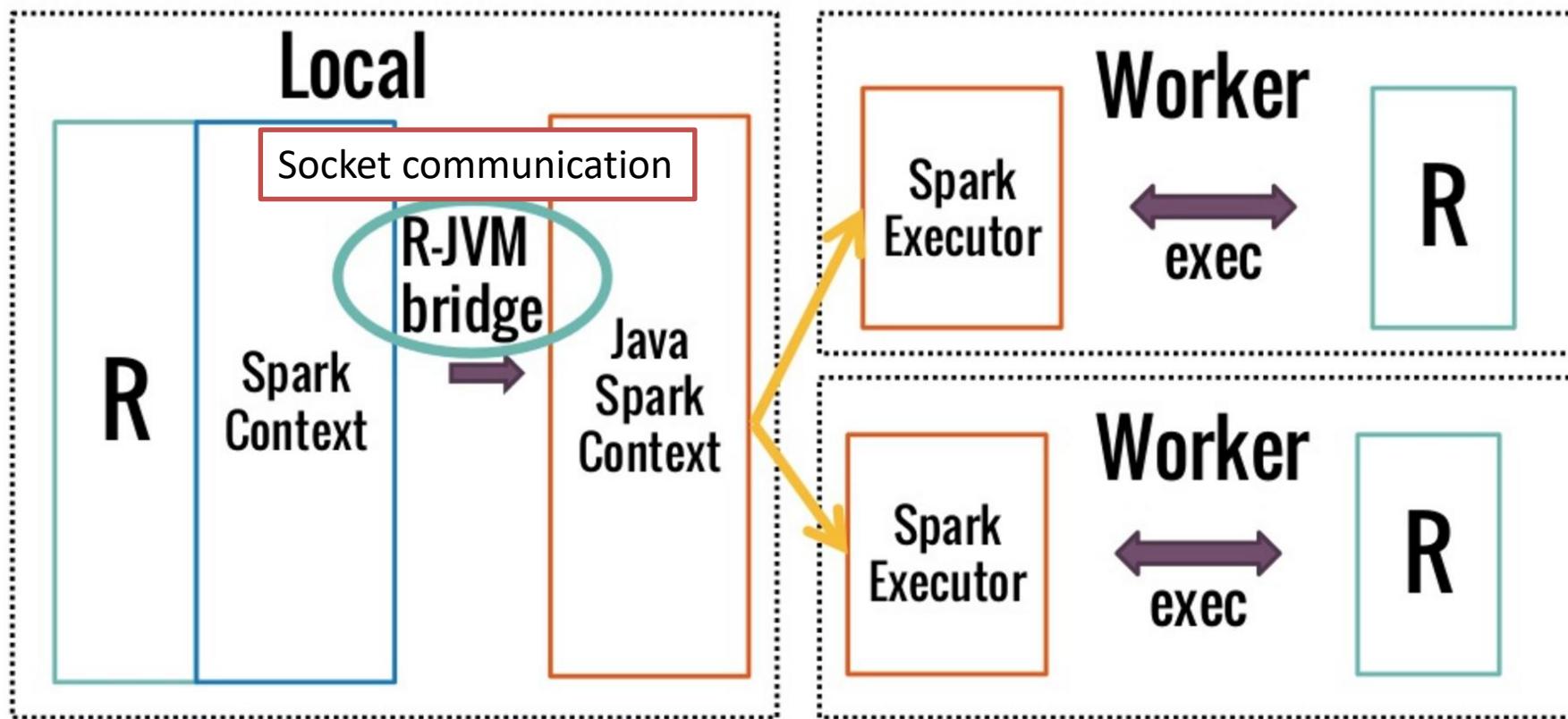


Large scale machine learning



```
> summary(model)
```

SparkR Architecture



SparkR – expressive power

```
> people <- read.df(sqlContext,  
"./examples/src/main/resources/people.json", "json")  
  
> head(people)  
##   age     name  
##1  NA Michael  
##2  30    Andy  
##3  19 Justin  
  
> hiveContext <- sparkRHive.init(sc)  
> results <- sql(hiveContext, "FROM src SELECT key, value")  
> head(results)  
##   key     value  
## 1 238 val_238  
## 2  86  val_86  
## 3 311 val_311
```

SparkR – expressive power

```
> df <- createDataFrame(sqlContext, iris)
> model <- glm(Sepal_Length ~ Sepal_Width + Species, data =
df, family = "gaussian")
> summary(model)
##$devianceResiduals
## Min      Max
## -1.307112 1.412532
##
##$coefficients
##                               Estimate Std. Error t value Pr(>|t|)
##(Intercept)              2.251393  0.3697543  6.08889 9.568102e-
09
##Sepal_Width                0.8035609  0.106339   7.556598 4.187317e-
12
##Species_versicolor  1.458743  0.1121079  13.01195 0
##Species_virginica   1.946817  0.100015   19.46525 0
```

Use case – credit scoring

Założmy, że merchant bez wcześniejszej historii kredytowej lub nawet bez żadnej historii (nowy klient) przychodzi do banku aby pożyczyc pieniądze.

... nie ma mowy ...

prawdą częstokroć jest, że posiada on bardzo realną i dobrze udokumentowaną historię sprzedaży... np. Ebay.com

Hipoteza:

Dla klientów, którzy sprzedają towary online w ramach platformy ebay.com jesteśmy w stanie oszacować czy dany klient spłaci kredyt wyłącznie na podstawie danych dotyczących historii sprzedaży na jego koncie.

Use case – credit scoring(Cont'd)

- OK. Ale w jaki sposób osiągniemy ten cel?



- Z pewnością dla części naszych klientów, którzy są merchantami prowadzącymi działalność w ramach Ebay jesteśmy w stanie dotrzeć do historii ich transakcji. Następnie możemy spróbować zinterpretować je jako historię sprzedaży w ramach platformy i zbudować model w oparciu o dane zewnętrzne, których de facto w banku nie posiadamy. **WOW!**

Use case – credit scoring (Cont'd)

Wektor uczący, który zawiera wyłącznie cechy

1. Od $b = 1$ do B : niezależne od danych bankowych.

- (a) Wylosuj próbę bootstrapową \mathbb{Z}^* o liczności N spośród wektorów uczących.
- (b) Wytrenuj drzewo lasu losowego T_b na próbie bootstrapowej powtarzając następujące kroki dla każdego węzła drzewa, aż do momentu gdy minimalny rozmiar węzła n_{min} zostanie osiągnięty (dla lasów losowych Breimana przyjmujemy $n_{min} = 1$):
 - i. Wylosuj $m = \lfloor \log_2 p + 1 \rfloor$ zmiennych spośród p wymiarów wektora cech.⁸
 - ii. Wybierz najlepszą zmienną/punkt podziału spośród m maksymalizując różnicę pomiędzy niebalansowaniem obserwacji w dzielonym węzle, a sumą niebalansowania obserwacji w węzłach potomnych:

$$\Delta Q_{w,w_L,w_R}(T_b) = Q_w(T_b) - Q_{w_L,w_R}(T_b), \quad (25)$$

gdzie $Q_{w_L,w_R}(T_b) = Q_{w_L}(T_b) + Q_{w_R}(T_b)$.

- iii. Podziel węzeł na dwa węzły potomne.
2. Zwróć las: $\{T_b\}_{b=1}^B$.

$$X = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

Use case – credit scoring (Cont'd)

- Results:

	Niespłacony	Spłacony
Prognoza - niespłacony	158	76
Prognoza - spłacony	290	6804

- **Quiz:** co jest nie tak z tokiem rozumowania przyjętym w tym modelu?

pytanie pomocnicze: jaką jest wspólna cecha dla wszystkich przypadków trenujących?

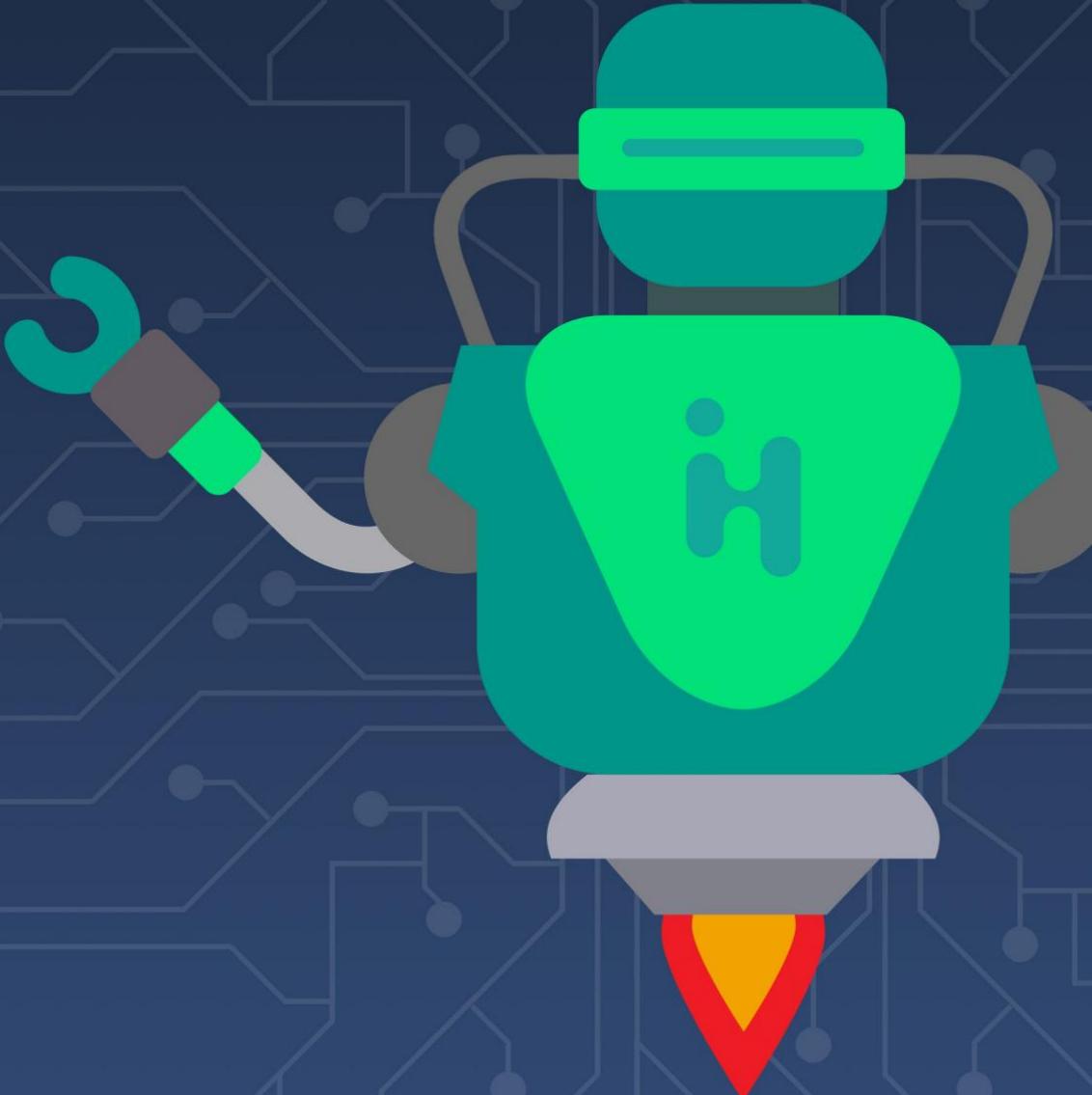


```
recruiterSpam.deleteAll();
```

a machine learning platform
that matches **IT Specialists**
with top tech companies

www.inhire.io

HAVE A LOOK INSIDE



Trusted by the most innovative tech companies

allegro



BCG GAMMA

BRAND24

codilime®
CREATING VALUE

SAMSUNG

netguru



binar::apps

COSMOSE

dynatrace

getindata

Apppsilon
DATA SCIENCE

esgroup

adform



DC POLCODE
DEV TEAM



HL TECH



BETTER
SOFTWARE
GROUP

PizzaPortal

SMEO.



MEDICALgorithms
INNOVATIVE SOLUTIONS IN MEDICINE

lingaro

addepto



SEQRED

IFM

sotrender

WLOG
SOLUTIONS

yameo

KERRIS



sunscrapers

ShelfWise

xchanger

nFront

ALGOLYTICS
Power your business with data!

FUTURE MIND™

idego

VMPL

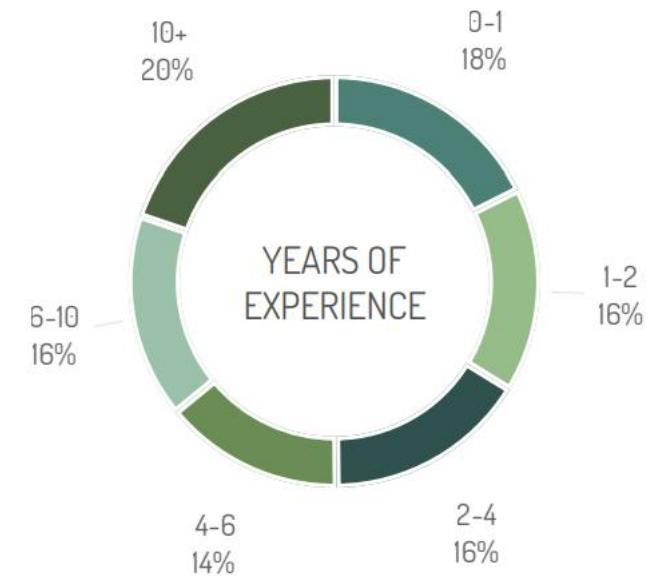
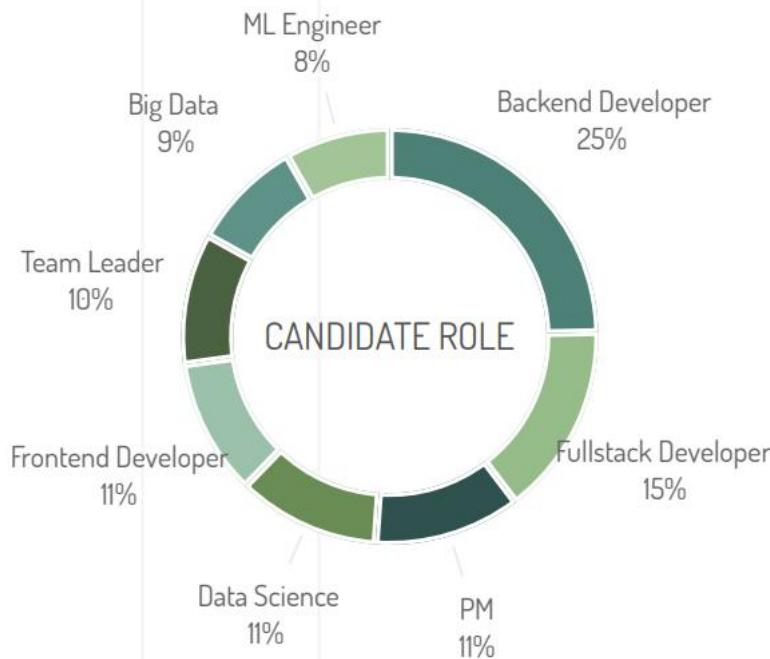
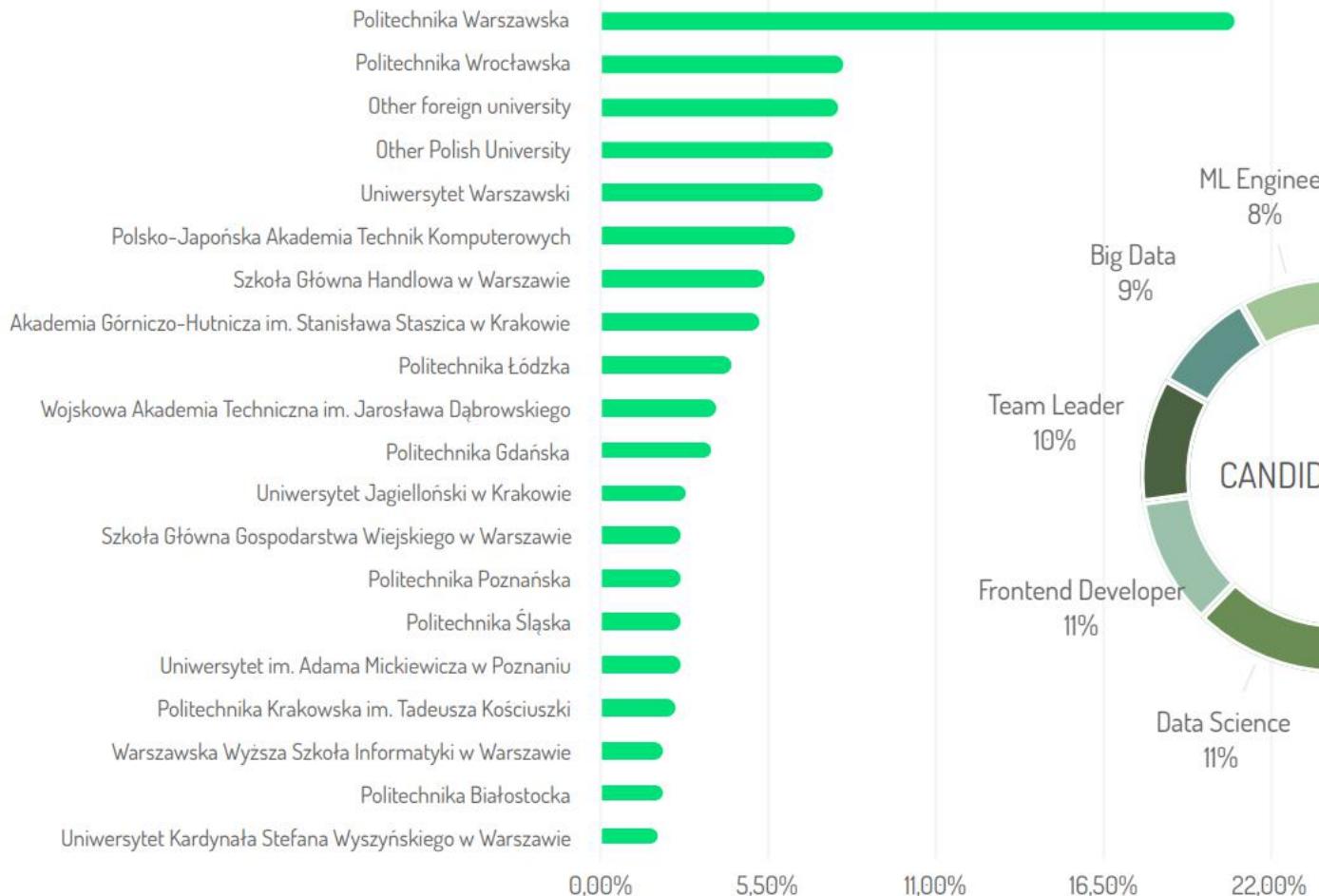
Draftable

How inhire works?



Candidates registered at inhire.io

TOP 20 UNIVERSITIES



Podsumowanie

1. Big Data nie jest taką nową koncepcją...
2. ... ale oczywiście użyteczną w wielu zastosowaniach.
3. Jednak znacząca liczba z technologii wykorzystywanych obecnie do procesowania w ramach Big Data nie została sprawdzona w boju.
4. Należy pamiętać, że GODO oraz organy nadzoru nie są wcale tak podekscytowane.