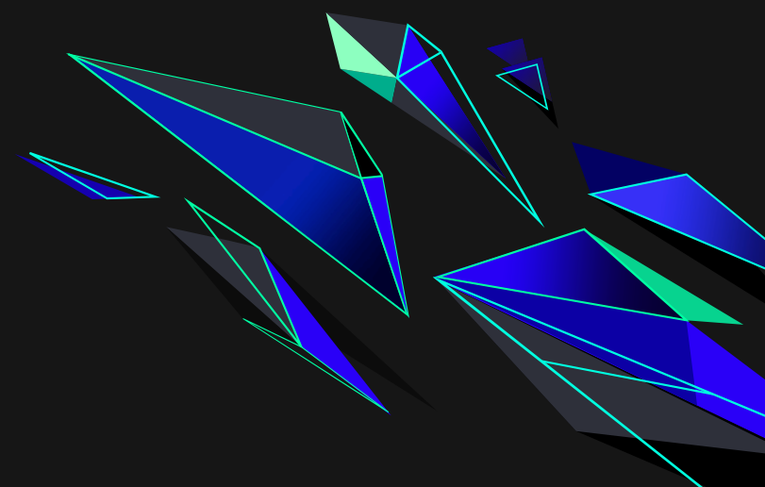
Audience Forecasting dla reklamy RTB

Przemysław Piotrowski

Adform Research



WARSZAWSKIE DNI INFORMATYKI.PL

sponsor diamentowy

Jobs for demanding talents

sponsor złoty

Wymarzona praca na pograniczu biznesu i informatyki

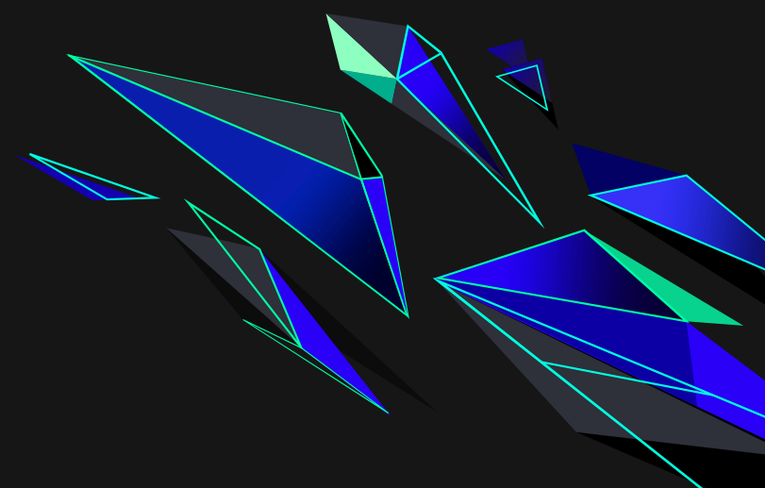
sponsor srebrny

sponsor srebrny

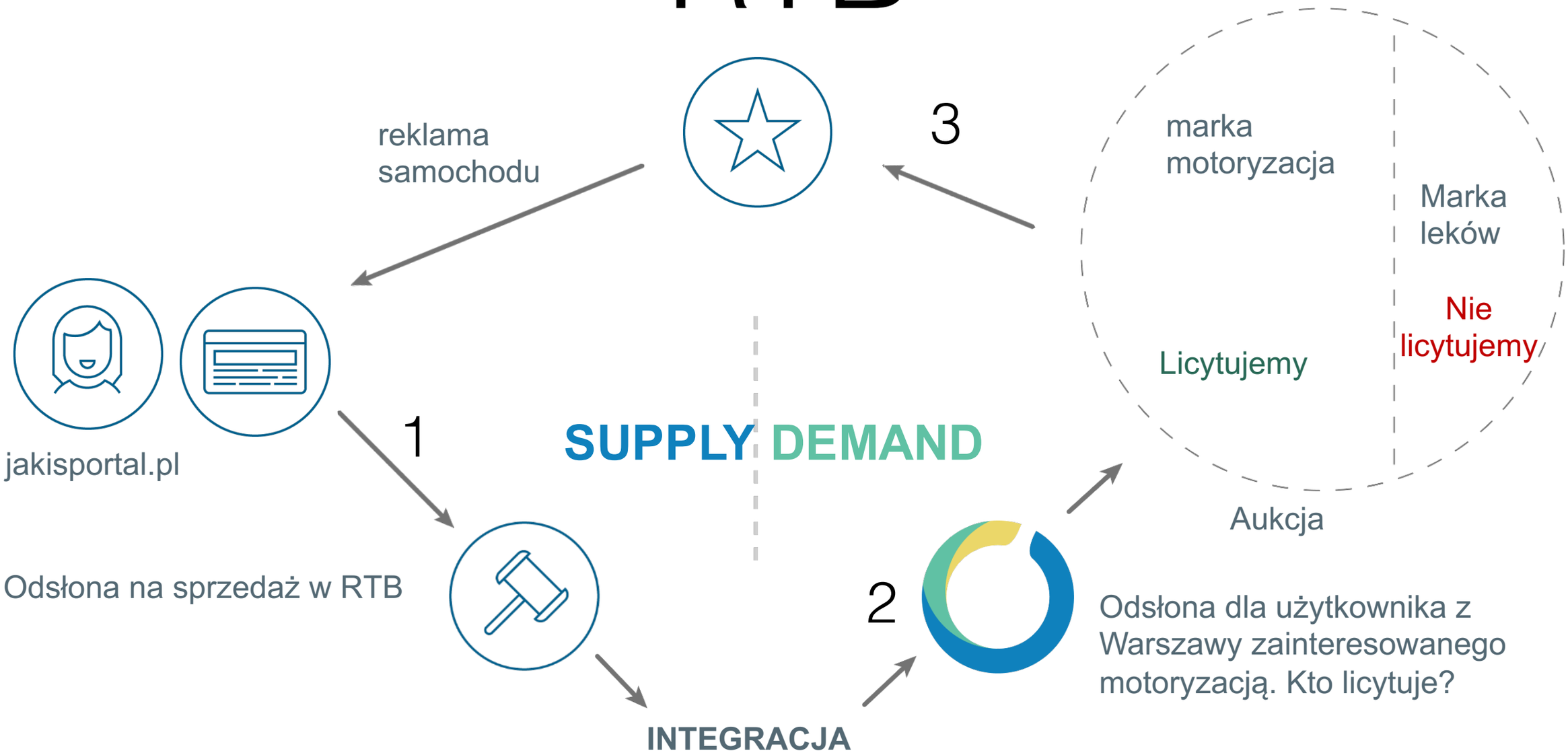
Come for career, stay for passior

Dzisiaj

- kampanie RTB
- przewidywanie zasięgu
- rozwiązanie
- jakość



RTB


reklama
samochodu
3
marka
motoryzacja
Marka
leków
Nie
licytujemy
Licytujemy
jakisportal.pl
1
SUPPLY
DEMAND
Aukcja
Odsłona na sprzedaż w RTB
2
Odsłona dla użytkownika z
Warszawy zainteresowanego
motoryzacją. Kto licytuje?
INTEGRACJA

3 mln RPS

Kampania

Źródło ruchu

- aplikacje mobilne
- wydawcy
- konkretne domeny
- URL
- data, czas
- kategorie

Ciasteczko

- ustawione?
- remarketing
- segment ciasteczka

Powierzchnia

- przygotowane banery
- odtwarzacze wideo
- above/below the fold

Limity

- frequency capping
- brand safety

Grupa docelowa

- kraj, województwo, miasto
- GPS
- język
- system operacyjny
- przeglądarka
- ISP
- łącze internetowe

Hyperlocal

GPS

RTB Audience Forecasting

Przewidywanie liczby

- ciasteczek
- dostępnych odsłon
- wygranych odsłon [1]

dla zdefiniowanej kampanii

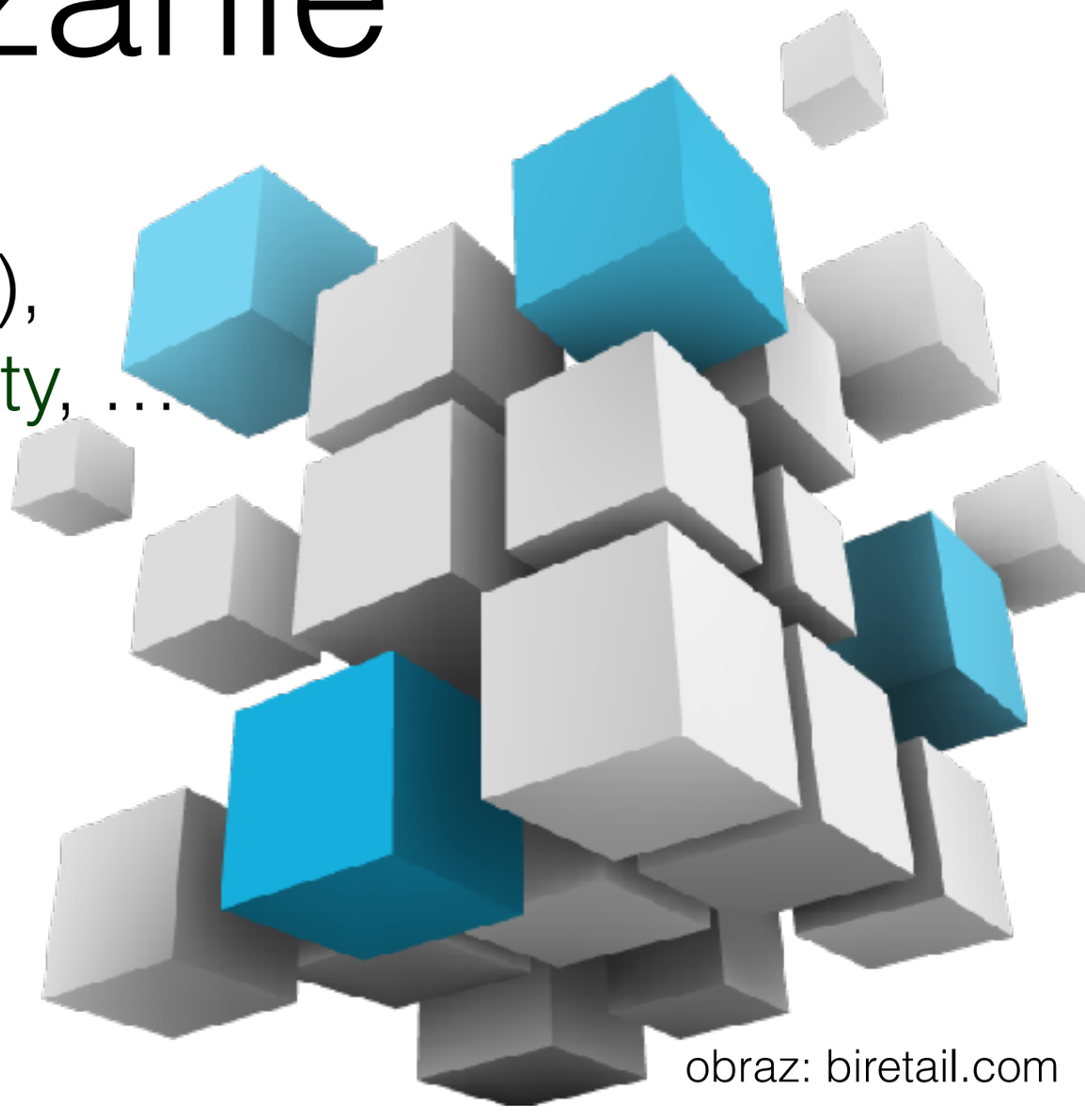
[1] Bid Landscape jest poza zakresem tej prezentacji

Wymagania

- odpowiedź do 1s podczas definiowania kampanii
- reprezentowanie
  - 3 mln RPS z EU, US, APAC
  - miliardów zdarzeń remargetingowych
  - ponad 100 000 segmentów
- dane wejściowe dla innych algorytmów

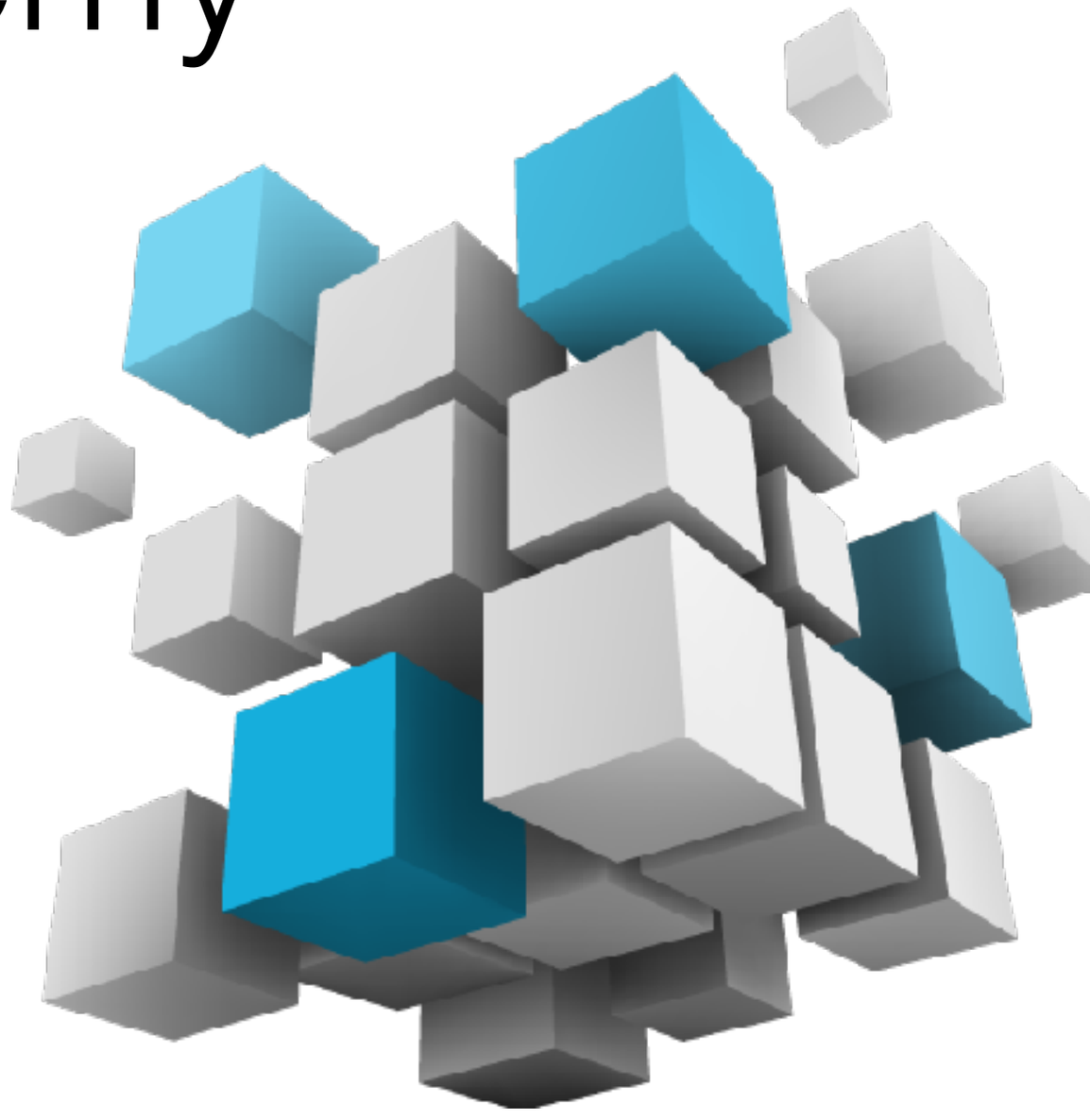
Rozwiązanie

```
SELECT
  count(1), count(distinct cookie_id),
  inventory, hour, country, region, city, …
FROM bid_requests
```



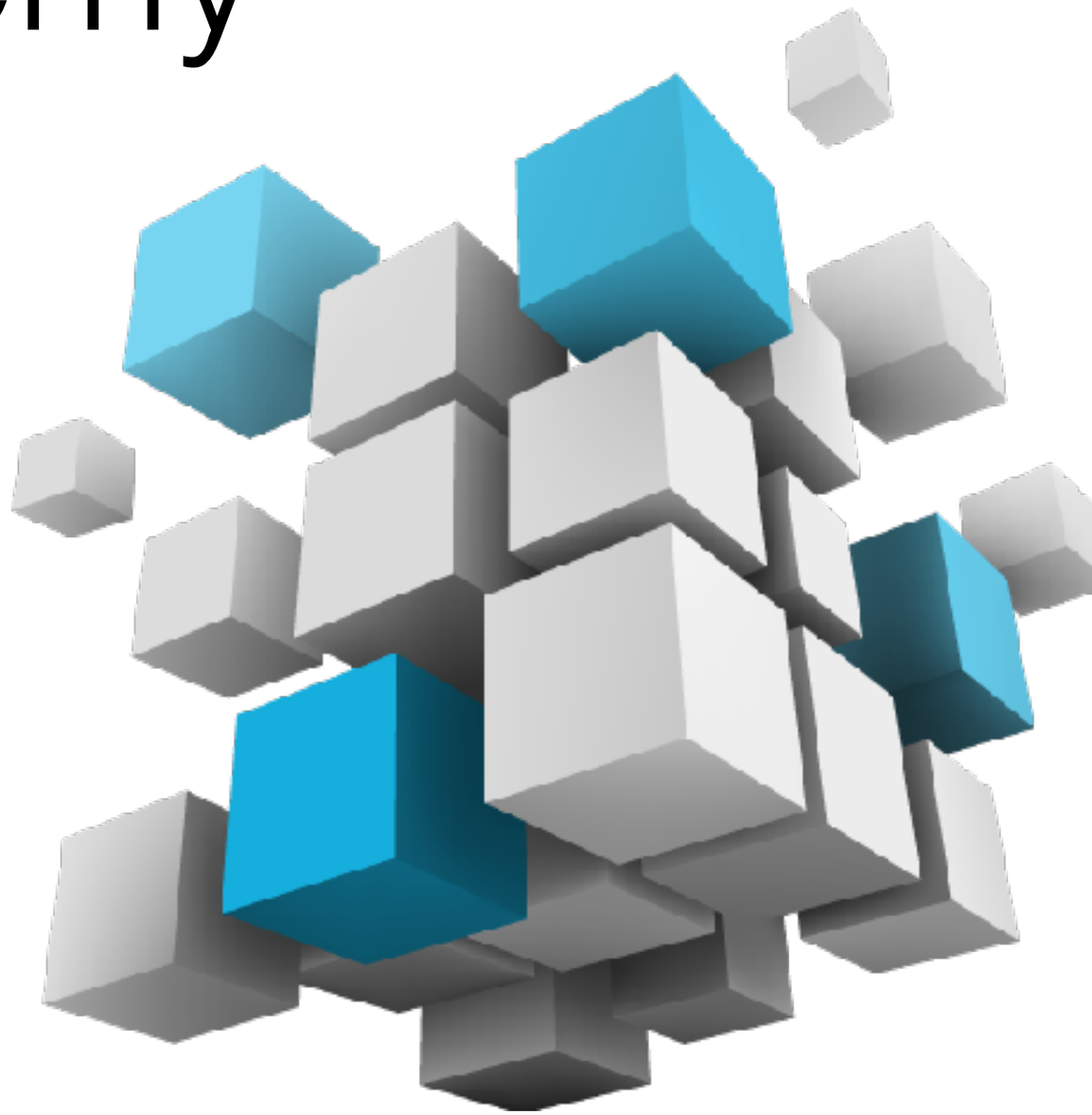
obraz: biretail.com

Problemy



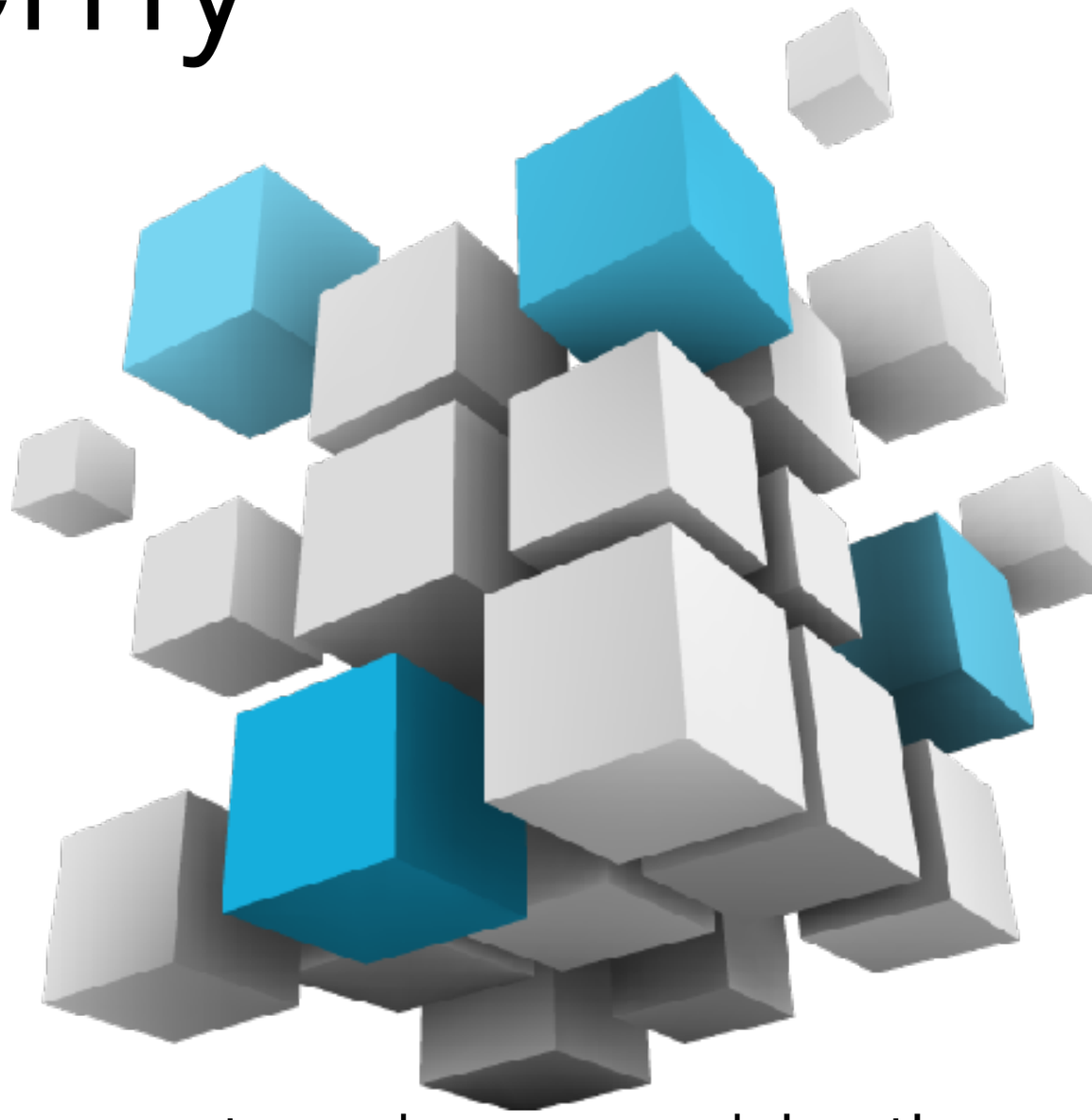
Zbyt wiele wymiarów

Problemy



Wspólne ciasteczka pomiędzy źródłami ruchu

Problemy



Złączenie ruchu z remarketingiem

„It will be challenging to have your forecast work on combinations of inventory and tracking points. Consider letting your users forecast either on inventory or on tracking.”[1]

[1] https://www.quora.com/How-can-we-do-reach-forecasting-in-a-DSP-demand-side-platform-Is-it-even-possible

Rozwiązanie

odsłona	iOS	Polska	hobby|ML	wynik
0	0	1	1	0
1	0	0	1	0
2	1	0	0	0
3	0	0	0	0
4	1	1	1	**1**
5	1	0	0	0
6	0	0	0	0
7	1	1	1	**1**
…	…	…	…	…
500mln	0	0	0	0

dostępne odsłony = `cardinality(wynik)`

Remarketing

odsłona	hobby|ML
0	1
1	1
2	0
3	0
4	1
5	0
6	0
7	1
…	…
500mln	0

hobby|ML

=

zapytania o odsłony ciasteczek, które zostały sprofilowane jako zainteresowane Machine Learning

Indeksy bitmapowe

ang. bitmap index = bitmap = bitset = bitarray

Cecha (kraj, producent)

Wartość cechy (`kraj|Polska`, `producent|samsung`)

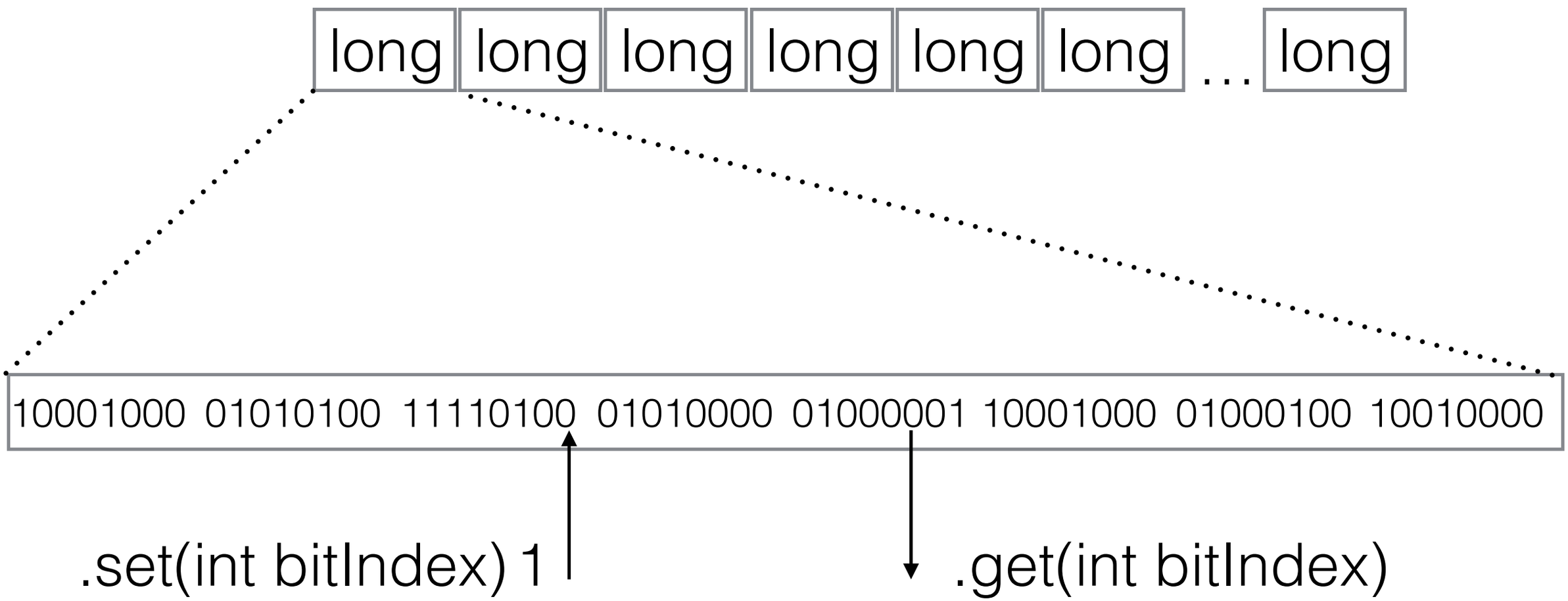
- jedna bitmapa dla każdej wartości cechy
- suma, iloczyn i różnica zbiorów
- wsparcie predykatu równości
- brak wsparcia dla operatorów porównania > <

Zastosowania

- pobranie konkretnych elementów z innego źródła danych
  - ***row_id*** w bazach danych
- zliczenie elementów spełniających kryteria wyszukiwania
- budowanie reguł asocjacyjnych na dużych zbiorach

~~java~~.util.BitSet
anylang

- private long[] words (64-bit)



java.util.BitSet

.set(int bitIndex)

.get(int bitIndex)

.cardinality()

.isEmpty()

.and(BitSet other)

.andNot(BitSet other)

.or(BitSet other)

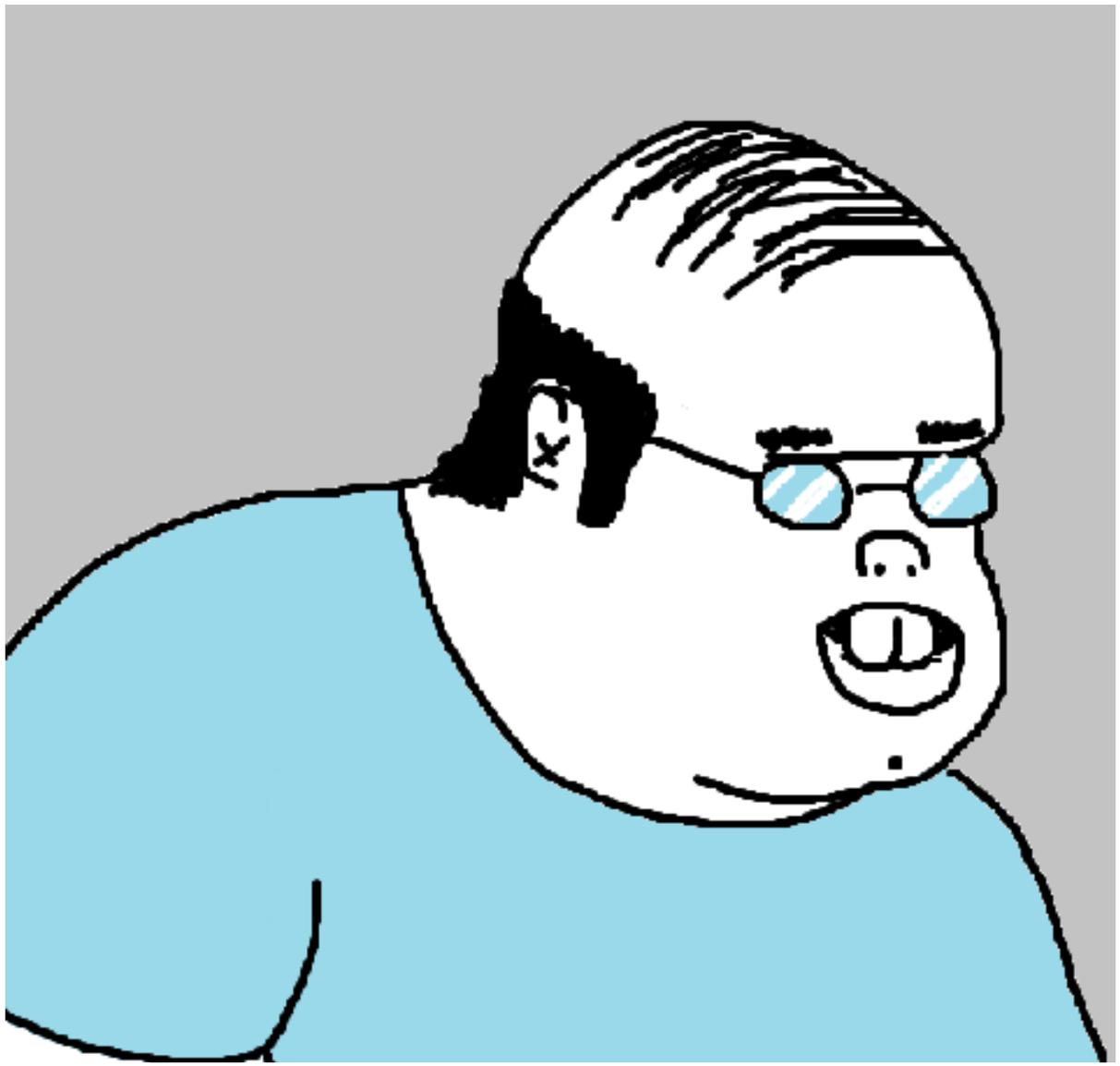
.xor(Bitset other)

Tylko jeden element

```
BitSet tiny = new BitSet();
tiny.set(Interger.MAX_INT);
```

2 147 483 648 bits = 256 MiB

ACKCHYUALLY



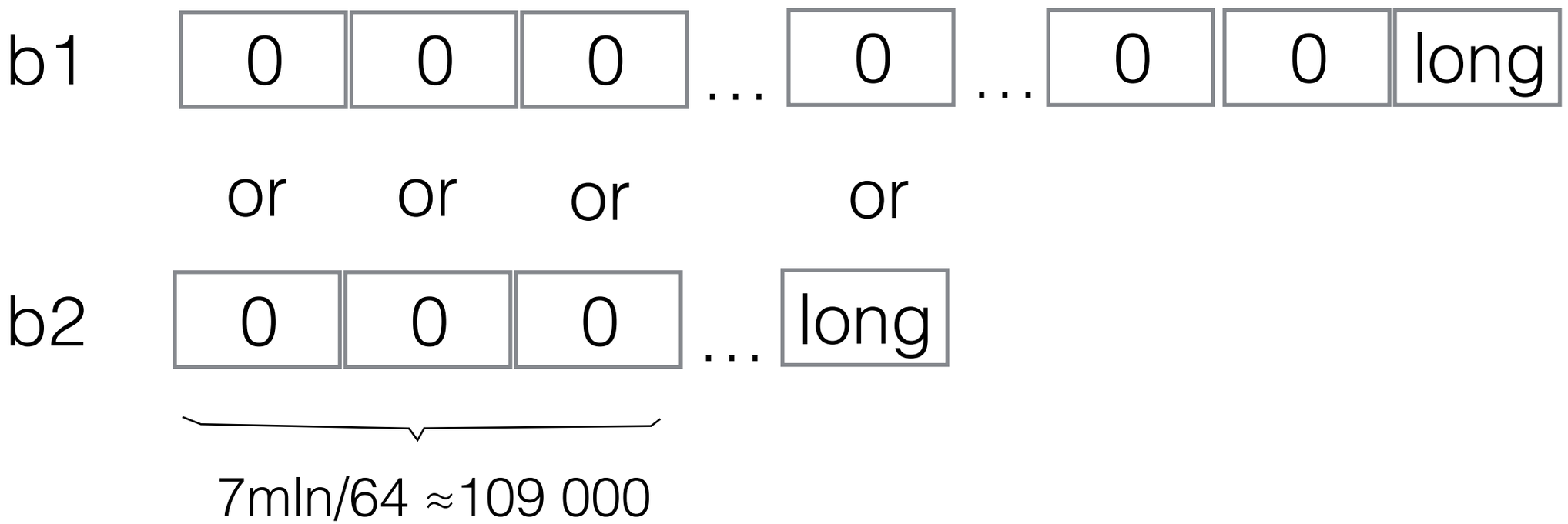
???

… ale mnie nie obchodzi zużycie pamięci

```
BitSet b1 = new BitSet();
BitSet b2 = new BitSet();
b1.set(9_000_000);
b2.set(7_000_000);
b1.or(b2);
```

CPU?

109k zbędnych alternatyw



Compressed Bitmaps

ROARING BITMAPS

Roaring Bitmap

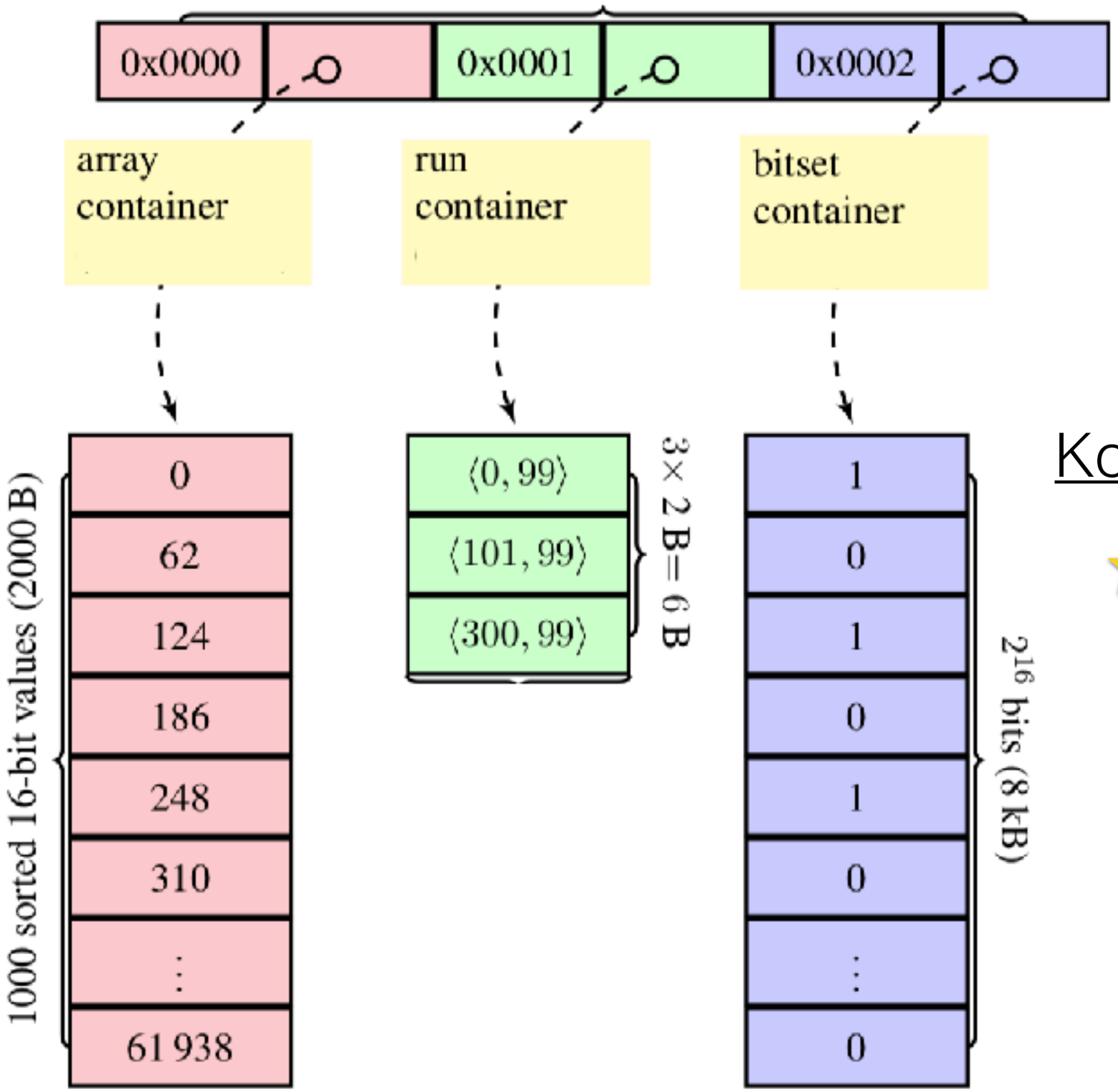
Otwarta, efektywna implementacja indeksów bitmapowych

roaringbitmap.org

„Use Roaring for bitmap compression whenever possible.
Do not use other bitmap compression methods” [1]

[1] Wang, Jianguo, et al.
"An experimental study of bitmap compression vs. inverted list compression." 2017.

Implementacja

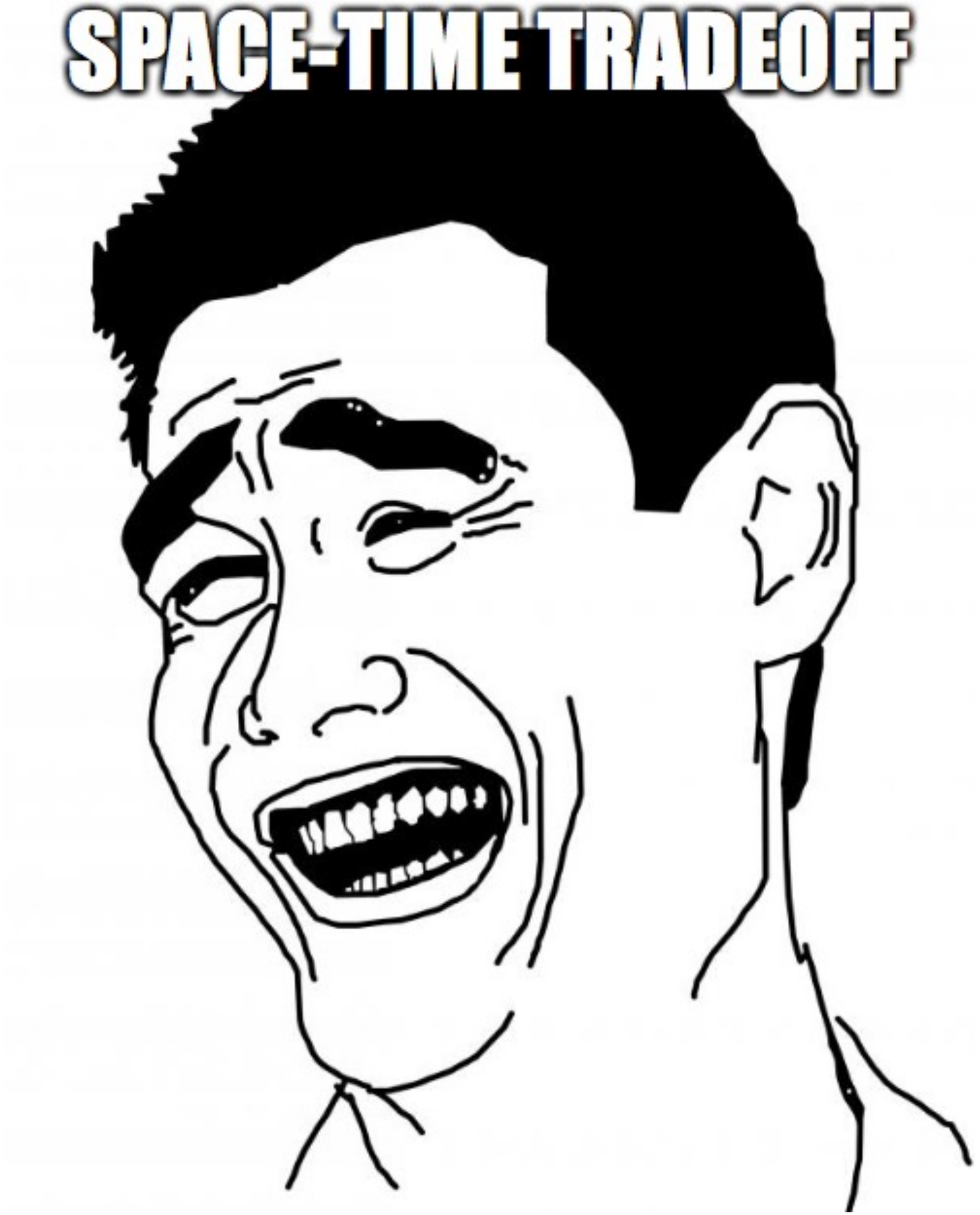


Każdy kontener reprezentuje obszar $2^{16} = 65536$ pozycji z przestrzeni 2^{32} liczb

Kontenery

- Brak kontenera
- **Array** - rzadkie (ang. sparse) regiony
- **Bitset** - gęste regiony
- **Run** - regiony z ciągłymi obszarami

https://arxiv.org/pdf/1709.07821.pdf


SPACE-TIME TRADEOFF

Optymalizacje

	array	bitset	run
array	?	?	?
bitset	?	?	?
run	?	?	?

dla każdej operacji
- sumy
- iloczynu
- różnicy

specjalizowana implementacja

- bit level parallelism
- dedykowane instrukcje CPU

5:33 https://www.youtube.com/watch?v=1QMgGxiCFWE

Kod

Implementacje

- C (SIMD[1])
- Java
- Go
- Python (C wrapper)
- …

DB

- Pilosa
- Roaring Redis

Użytkownicy

- Adform
- Apache
- Netflix
- Druid
- Linkedin
- Microsoft

[1] Lemire, Daniel, et al. "Roaring bitmaps: Implementation of an optimized software library." Software: Practice and Experience (2017).

Usprawnienia

Sorting improves word-aligned bitmap indexes

Daniel Lemire[a,*], Owen Kaser[b], Kamel Aouiche[a]

[a] *LICEF, Université du Québec à Montréal (UQAM), 100 Sherbrooke West, Montreal, QC, H2X 3P2 Canada*
[b] *Dept. of CSAS, University of New Brunswick, 100 Tucker Park Road, Saint John, NB, Canada*

Przykład:

40 000 kontenerów typu **array** może nieść tę samą informację co 1 kontener typu **run**.

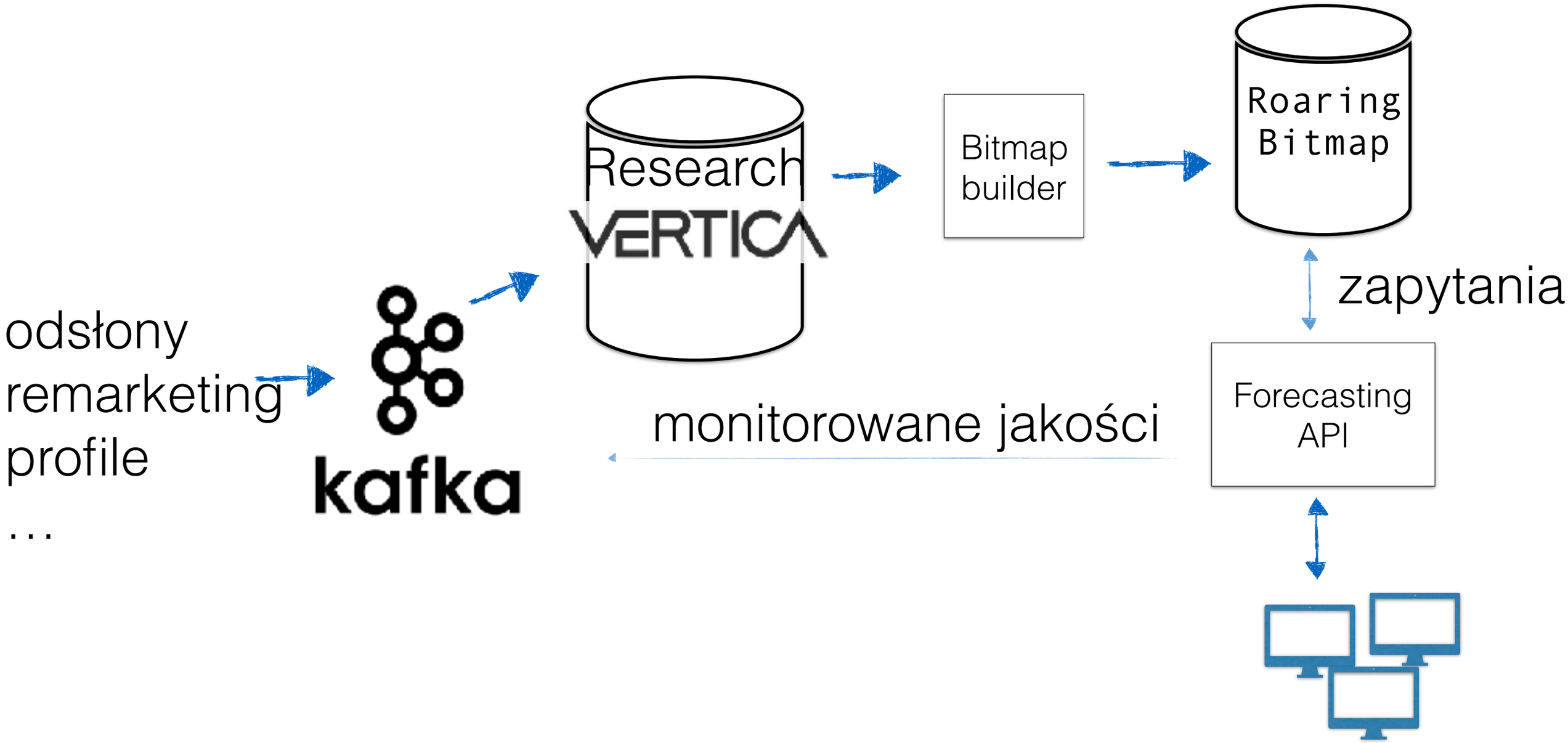
https://arxiv.org/pdf/0901.3751.pdf

Sortowanie

ORDER BY country_id, device_id

odsłona	Polska	Niemcy	samsung	apple
0	1	0	1	0
1	1	0	1	0
2	1	0	0	1
3	1	0	0	1
4	0	1	1	0
5	0	1	0	1
6	0	1	0	1
7	0	1	0	1
…	…	…	…	…
500mln	0	0	0	0

Architektura



Zapytania

```
And(
  Or("inventory|1", "inventory|2", "inventory|3"),
  AndNot(
    Universe,
    Or("kategoria|motoryzacja", "kategoria|wedkarstwo")
  ),
  Or(
    "godzina_tygodnia|41",
    "godzina_tygodnia|42",
    "godzina_tygodnia|43",
    "godzina_tygodnia|44"
  ),
  Or("domena|com", "domena|testowo.org"),
  And(
    "kraj|polska",
    AndNot(
        Universe,
        "miasto|warszawa"
    )
  ),
  ...
)
```

Liczby

- **10 GB** rozmiar bitmap dla 7 dni
- **200 TB** podejście naiwne
- **6h** czas budowy bitmap dla 7 dni
- **3 mln** cech
- **500 mln** * *próbkowanie* zapytań
- **5 mln** * *próbkowanie* ciasteczek

Zasięg kampanii

- czy niepusty segment?
  (US - .com)
- liczba ciasteczek

odsłona	iOS	Polska	hobby|ML	wynik
0	0	1	1	0
1	0	0	1	0
2	1	0	0	0
3	0	0	0	0
4	1	1	1	**1**
5	1	0	0	0
6	0	0	0	0
7	1	1	1	**1**
…	…	…	…	…
500mln	0	0	0	0

dostępne odsłony = `cardinality(wynik)`

Zliczanie ciasteczek

odsłona	wynik
0	0
1	0
2	0
3	0
4	**1**
5	0
6	0
7	**1**
…	…
500mln	0

odsłony

{4, 7, …}

f: R_{pos} ->C_{pos}

ciasteczka

{1, 17,…}

300k ciastek

$R_{pos} \in$ <0;500mln)

$C_{pos} \in$ <0;5mln)

Uczenie maszynowe

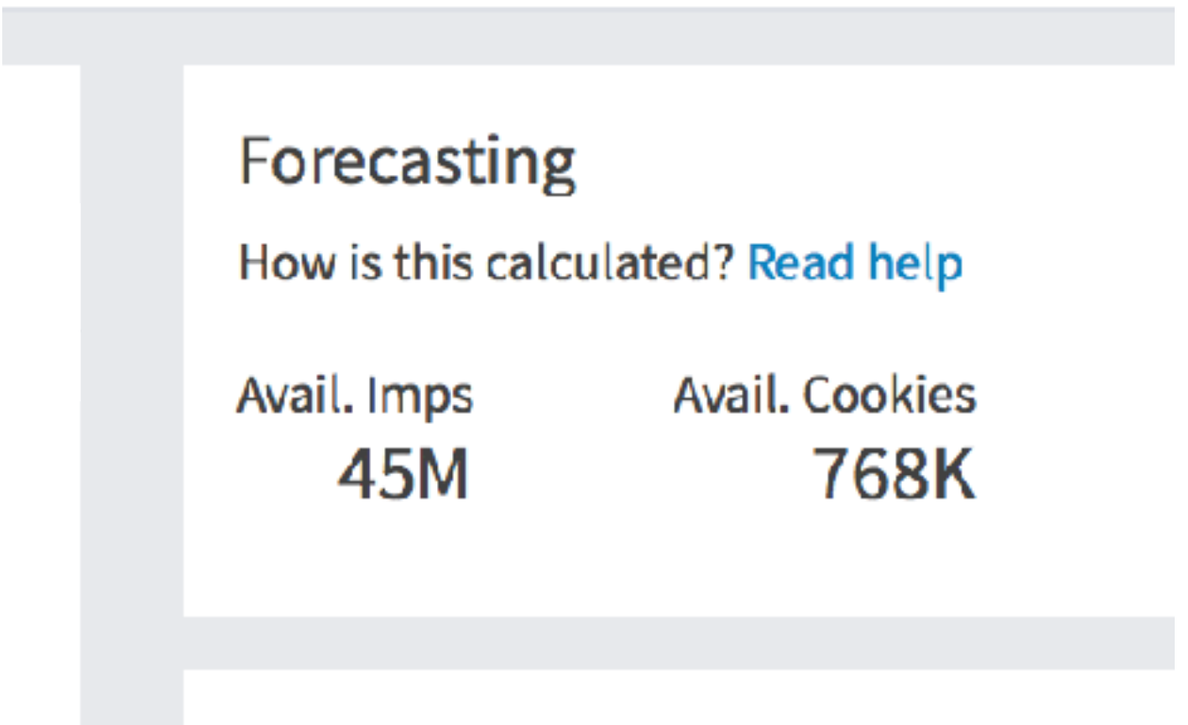
- czas trwania kampanii
- limit odsłon dla ciasteczka

Jakość

- eksperymenty vs. produkcyjnie działająca kampania
- kampanie są ciągle modyfikowane
- logowanie na Kafce pełnego zapytania wraz z wynikami pośrednimi i parametrami użytymi do estymacji
- monitorowanie impresji następnego dnia
- powtórzenie na bitmapach z kolejnych dni
- trend

Podsumowanie

- Roaring to zalecana implementacja indeksów bitmapowych
- obecna w wielu znanych aplikacjach
- zastosowana z sukcesem do zliczania ruchu w RTB
- 5 różnych wdrożeń w Adform



>>>Pytania?

Zadawajje, oceniaj prelekcję, komentuj lub polub poprzez sli.do:

WarszawskieDniInformatyki.pl/slido

WARSZAWSKIE DNI INFORMATYKI.PL