Od pomysłu przez hipotezę do modelu

co gdzie może pójść nie tak?

Kornel Kiełczewski
Grupa Allegro, 17.12.2015
kornel.kielczewski@allegrogroup.com

allegro.tech

wniosek

- 70% czasu poświęcisz na przygotowanie danych.

allegro

- dane clickstream - ~70 - 80GB dziennie
- oferty - 45M, a do tego jeszcze aukcje…
- użytkownicy - 30M

allegro

- królestwo trzech d:
    - data driven decisions
- fail fast (miesiąc to nie jest fast)

w2v

- word2vec - znając kontekst słów, buduje model mapujący słowa na wektory, których odległości zachowują ciekawe własności.

- przykład: v(król) - v(mężczyzna) + v(kobieta) = v(królowa)

pomysł

- a gdyby zastosować word2vec na opisach ofert?
- może pozwoli to znajdywać oferty podobne kontekstowo: czajnik + toster = …
- możemy dodawać cechy użytkownika do ofert: rower + mężczyzna = …
- eksperyment!

co jest potrzebne?

- wszystkie opisy ofert z Polski, aktywnych na dany dzień
- brzmi jak `select * from descriptions where country = 'PL' and date = 'xxx'`?

do roboty!

- ale zaraz… gdzie są opisy ofert?
- oracle? kafka? hdfs?

do roboty!

- zrzut z kafki na hdfs - brzmi jak BigData.
- opisy z procesu tworzenia ofert vs zmiany.
- co lepsze? Sprawdźmy!

snakebite ls /data/opisy/ofert

Source: giphy.com/gifs/big-rugby-collision-aoRWJpTQiMC8o

Uprawnienia…

Nie mam dostępu.

End of Story.

allegro.tech

Life Pro Tip #1

- uprawnienia są konieczne, ale mogą być niepotrzebnym hamulcem.
- dlatego bardzo istotny jest sprawny sposób nadawania (oraz odbierania) uprawnień.

opisy ofert

- `hdfs dfs -text … | head`
- {"offerId":"5856412830","eventTimestamp":"1449873757","eventMicroTimestamp":"1449873757817851","producer":null,"description":"<p><img src=\”http://images57.fotosik.pl/116….”}

Co tu jest nie tak?

- snappy - ok
- json - bardzo nie ok
- html - co zrobić, taka sytuacja
- brak schemy - w tym wypadku damy radę, raczej nie powinno być niespodzianek (mhmm)
- brak informacji z jakiego serwisu oferta pochodzi (polska, czechy, ukraina, …)?

Life Pro Tip #2

- json do przechowywania danych jest fatalny
- avro, parquet, orc, cokolwiek. będzie lepiej.

Szukamy kraju

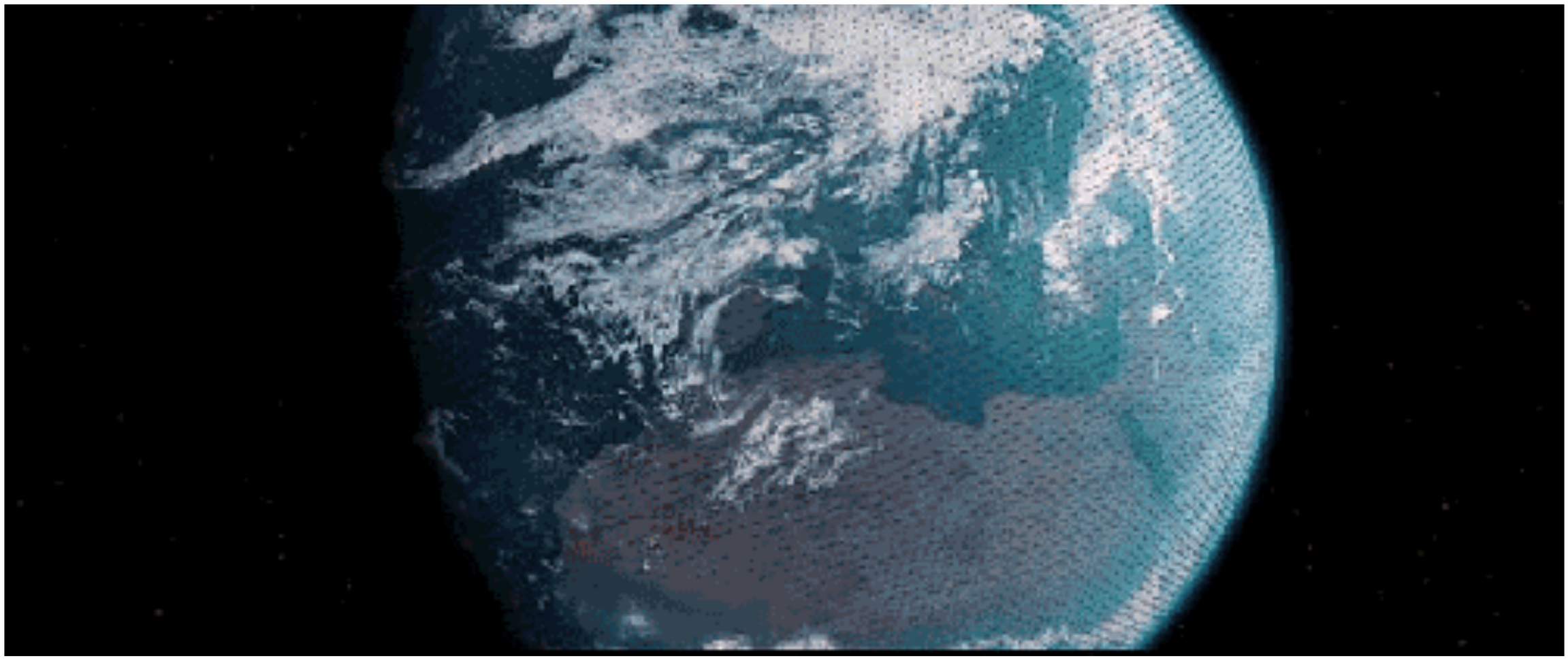
- wystarczy inner join z tabelą z ofertami z Polski
- Ocean możliwości:
    - hive, beeline w cli
    - hive w hue
    - spark job
    - spark shell
    - spark w jupyter notebook
    - spark w hue
    - …

co tu wybrać

- pro tip: spark-shell w hue na razie nie działa.

- użyję notebooka w jupyter.

jupyter



Source: giphy.com

allegro.tech

Life Pro Tip #3

- może to truizm, niemniej:
- jeśli cokolwiek gdziekolwiek nie działa tak jakbyśmy tego oczekiwali, zmieńmy to albo poprośmy kogoś aby poprawił.

w takim razie inaczej

- chcę zrobić join'a z Polskimi ofertami.
- spark-shell?
- tym razem lepiej hive.

Life Pro Tip #4

- jeśli Hive, to nigdy w życiu M/R. W allegro króluje obecnie Tez, może niebawem Impala?

- oczywiście, jak to w branży, to wszystko zależy, jednak gorzej od klasycznego M/R ciężko.

- trzeba próbować różnych rozwiązań.

join!

- {"offerId":"5856412830","eventTimestamp":"1449873757","eventMicroTimestamp":"1449873757817851","producer":null,"description":"<p><img src=\”http://images57.fotosik.pl/116….”}

- json - ok, offerId wyciągniemy za pomocą get_json_object.

- teraz czy mam na hdfs oferty z Polski?

join!

- dane są, ale partycja hive’a jest per data rozpoczęcia oferty, nie do użycia.

Source: http://giphy.com/gifs/agony-affliction-5AVgmlw7iAzdK

Life Pro Tip #5

- przygotowuj dane tak, aby od razu dało się z nich skorzystać.

- w szczególności: jeśli mamy dane z różnych rozłącznych źródeł, zawsze dostarczajmy dyskryminator skąd co pochodzi.

join był długi i bolesny

- hive ma ciężkie zadanie jeśli chodzi o optymalizację join'a
  - można mu pomóc, ale to wymaga dużo pracy, znajomości danych

- `insert overwrite table xxx select ...` - jest ryzyko dużej ilości małych plików. warto przerzucić

Life Pro Tip #6

- zwróc uwagę na to co hive wczytuje a co zapisuje - masz bardzo duży wpływ na wydajność zapytań.

ale co z tym html'em?

- to powinno być proste: UDF
- usunięcie <znaczników>JSoup</znaczników>
- build tool: Java? Gradle

zależności

- hive - tylko podczas kompilacji
- hadoop-core - tylko podczas kompilacji
- jsoup

Gradle what?

- gdzie jest scope provided?

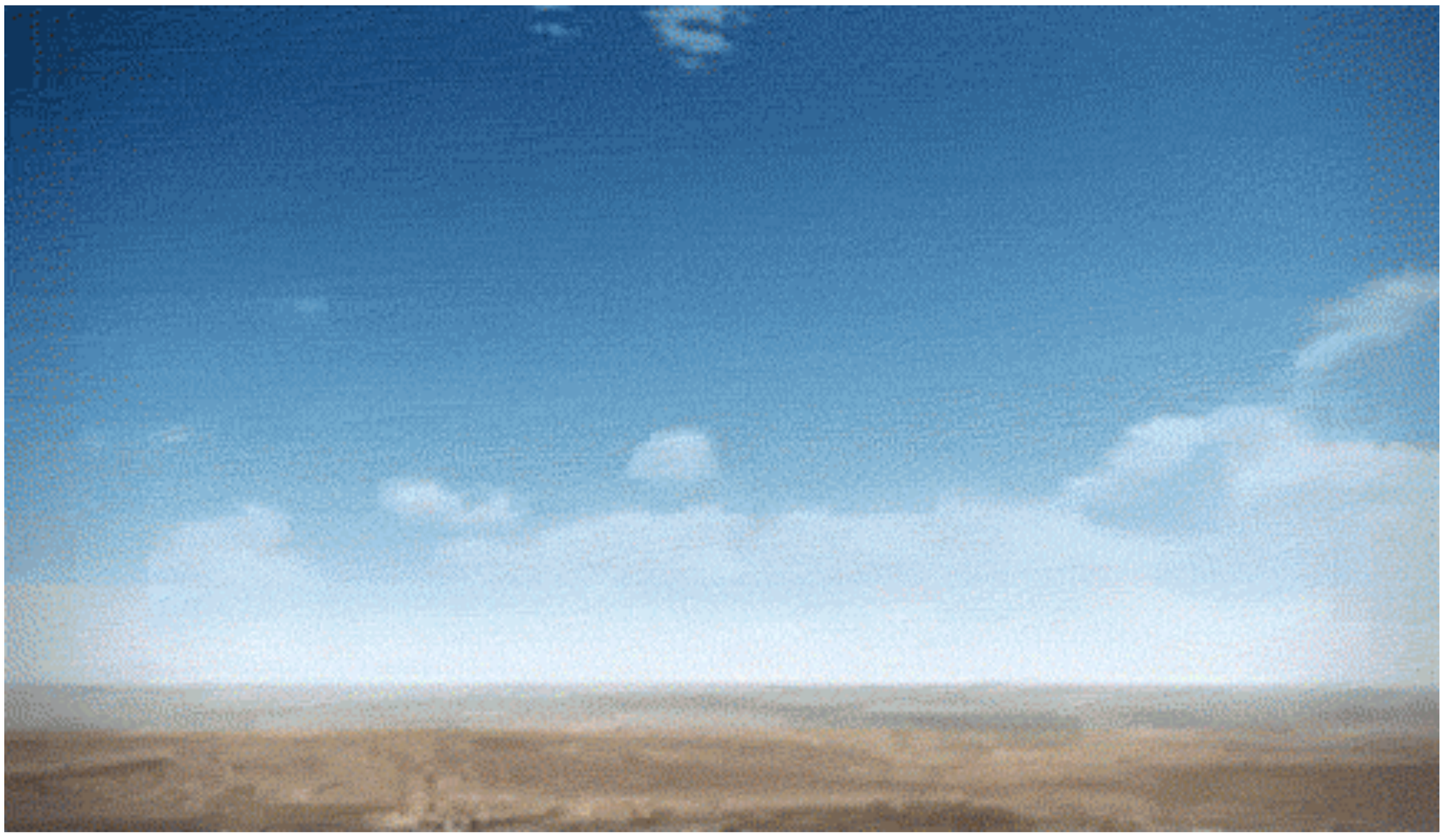
Life Pro Tip #7

- przetwarzając dane używaj narzędzi które znasz, ale nie bój się sprawdzić nowych.
- musisz być elastyczny i znać wiele rozwiązań aby szybko dotrzeć do celu.

JSoup UDF

- UDF gotowy, wrzucam.
- `> add jar hive-udf-1.0-SNAPSHOT-all.jar;`
- `> create temporary function striphtml as 'pl.allegro.hive.StripHtmlUDF';`
- `> select striphtml(description) from offer_descriptions_allegro limit 1;`

j7 v j8



Source: http://giphy.com/gifs/explosion-nuclear-12KiGLydHEdak8

Life Pro Tip #8

- pracując w ekosystemie Hadoop musimy znać (mniej więcej) wersję Hive'a, Hadoop'a, Javy, …

- warto zwrócić na to uwagę, aby uniknąć niepotrzebnego turn-around

co mamy

- wszystkie opisy ofert z Polski, aktywnych na dany dzień.

- bez HTML’a.

word2vec

- Jaka implementacja? Może ta ze spark'a?

```
@Since("1.1.0")
@Experimental
class Word2Vec extends Serializable with Logging {

  private var vectorSize = 100
  …

  /**
   * Sets vector size (default: 100).
   */
  @Since("1.1.0")
  def setVectorSize(vectorSize: Int): this.type = {
    this.vectorSize = vectorSize
    this
  }

  …
```

.collect()

- szybko zabrakło pamięci.

- może gensim? Próbka 2GB danych na laptopie nie powinna być problemem.

Life Pro Tip #9

- odpowiednie narzędzie do odpowiedniego zadania.

- cudownie jest przetwarzać 30GB danych, ale może 2GB wystarczy?

word2vec

- word2vec uruchomiony za pomocą gensim na 2GB opisów ofert

- mamy pierwsze wyniki!

Co z tego wyszło?

rower + mężczyzna	
starszy	0.76
prawdziwy	0.75
chłopak	0.75

szybko + mężczyzna	
dziewczynie	0.83
Kiedy	0.82
potrafi	0.82

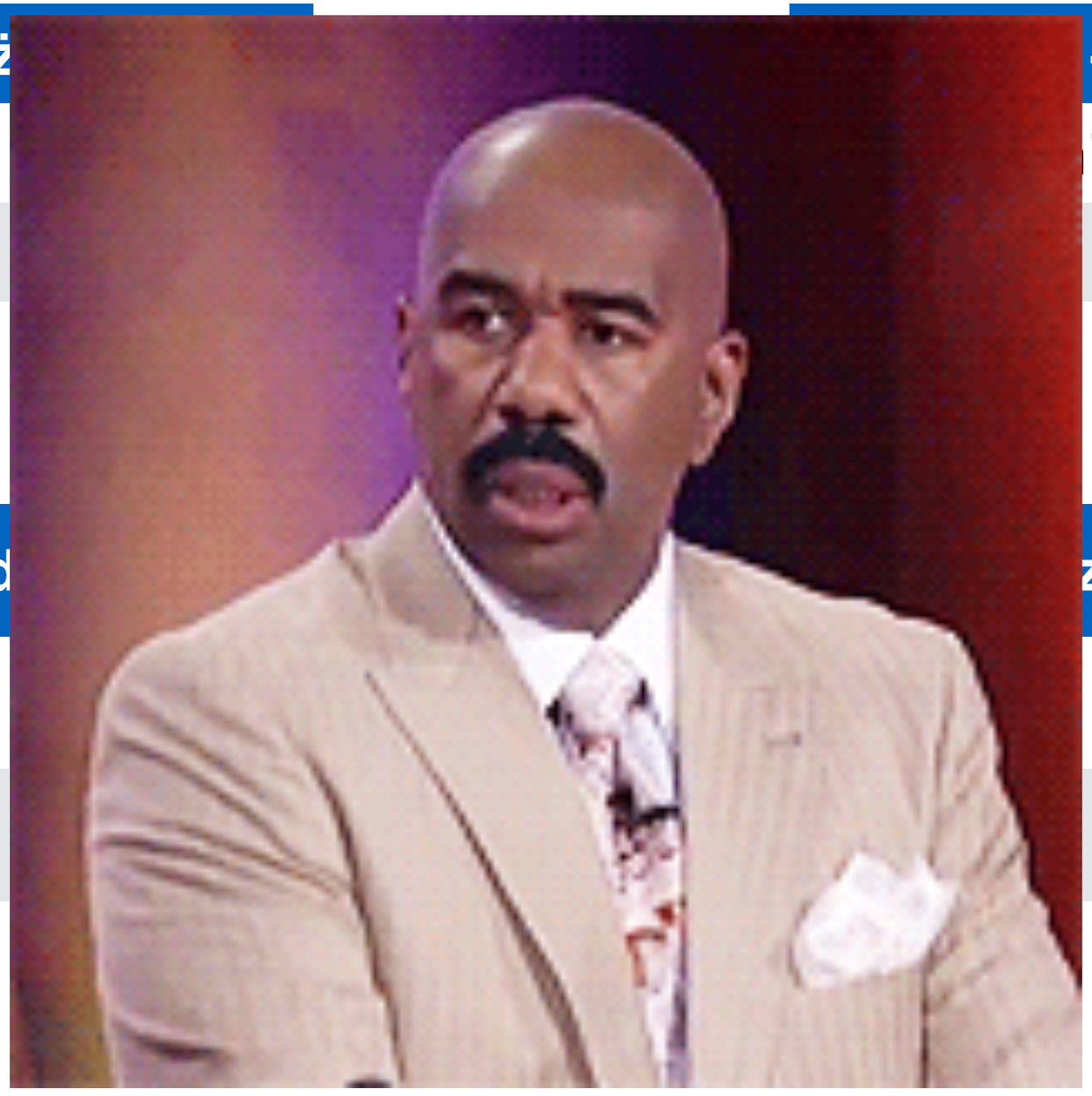
czajnik + toster	
składany	0.76
nóz	0.75
szybkowar	0.74

król - mężczyzna + kobieta	
ojca	0.77
powieści	0.76

allegro - drogo	
wpłaty	0.50
przelewem	0.49

Czyli Nic

rower + męż		+ mężczyzna	
starszy		ie	0.83
prawdziwy			0.82
chłopak			0.82



allegro - d		zna + kobieta	
wpłaty			0.77
przelewem			0.76

Source: http://giphy.com/gifs/everyone-brown-shaun-LyJ6KPlrFdKnK

wniosek

- minimum 70% czasu poświęcisz na przygotowanie danych

wniosek

- warto zadbać aby sprawdzenie dowolnej hipotezy było jak najszybsze.

- wtedy jest energia.

Pytania

Dziękuję

Kornel Kiełczewski
Grupa Allegro

kornel.kielczewski@allegrogroup.com

allegro.tech

A teraz…

- … z zupełnie innej beczki, czyli co zrobić gdy uda się zbudować sensowny model i go wdrożyć na produkcję?