FJFI & PRAXE

Jakub Stech, Petra kostakova

7.3.2018

O CEM TO BUDE

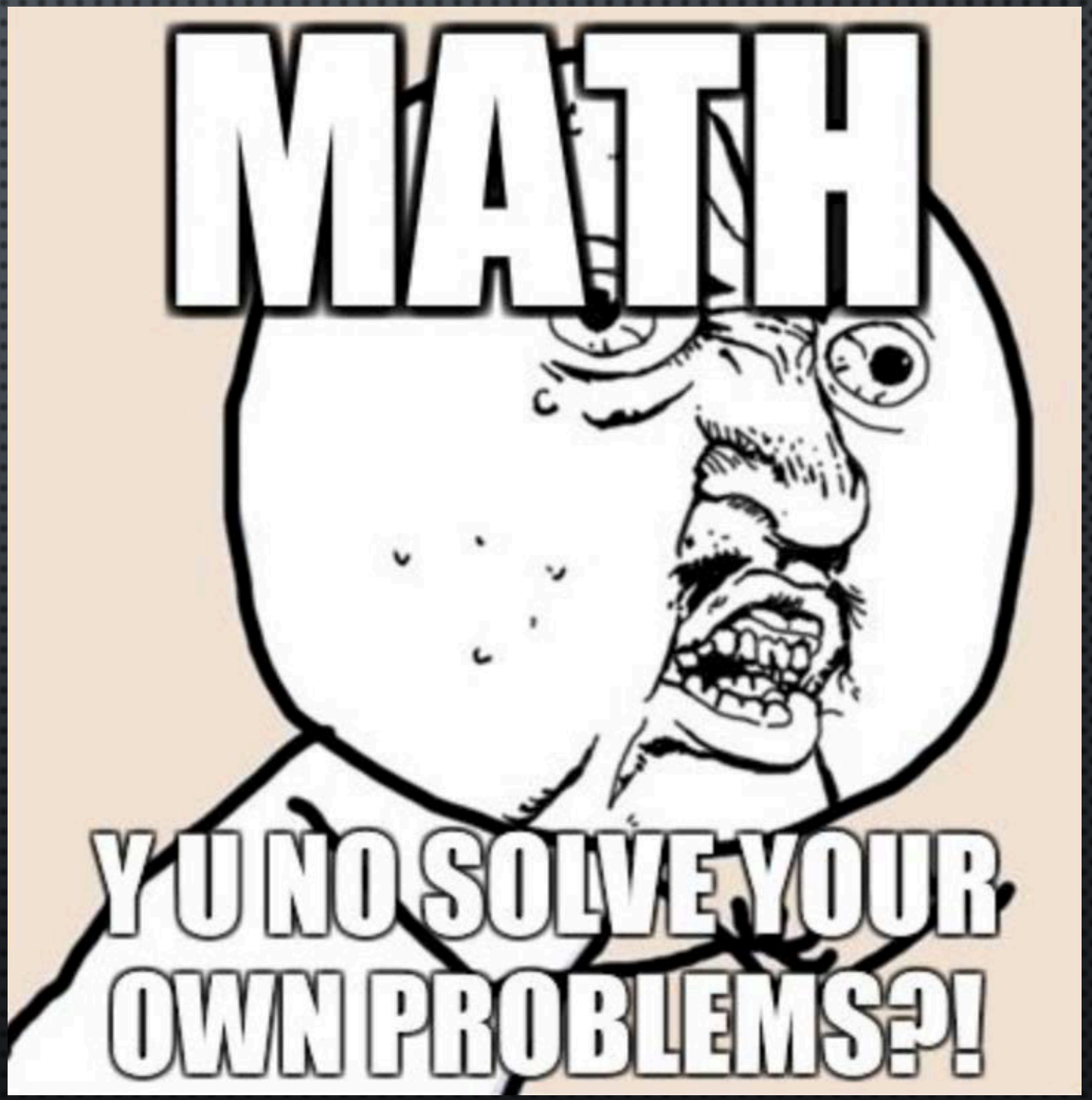
JAKUB

- PREDSTAVENI
- PRACE DATA ANALYTIKA+
- USE CASES

PETRA

- RANDOM FORESTS
- USE CASE

VAROVANI
(PRED MOJI CASTI)



KDO JSME

Jakub Stech

- ING NA FJFI AMSM 2016
- PHD 2016 - ???
- AS UTIA
- DATASENTICS 2016

KDO JSME

Petra Kostakova

- ING NA FJFI MM 2007
- RIP PHD
- DATASENTICS 2016

Petra Kocábová

rok: 2007

zaměření: Matematické modelování

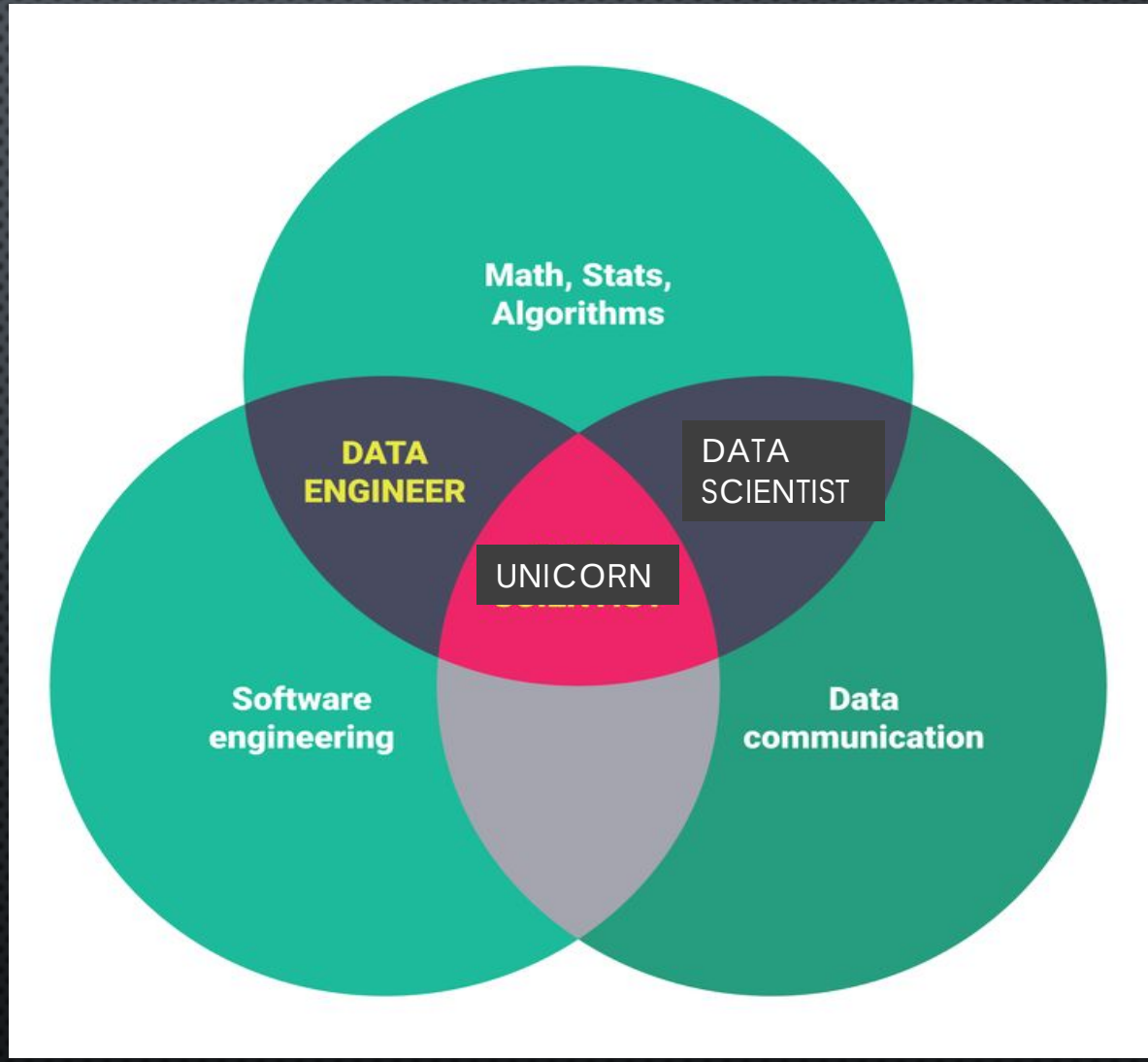
práce: Kvantová mechanika na násobně souvislých varietách Quantum mechanics on multiply connected manifolds

školitel: Prof. Ing P. Šťovíček, DrSc.

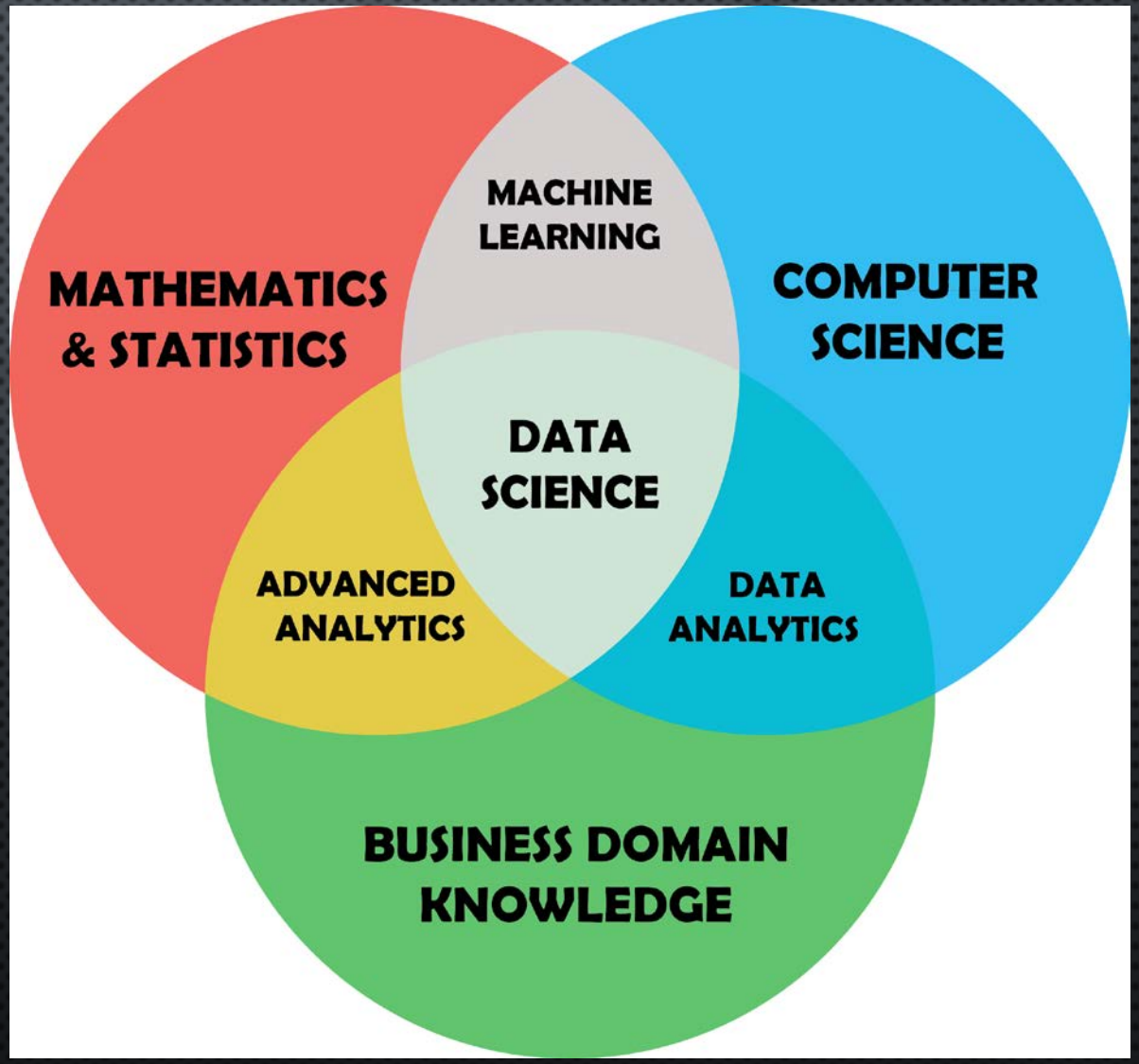
DATASENTICS

- MALY AGILNI BUTIK NA DATOVOU ANALYZU, ML, TECH (50% FJFI)
- DATA ENGINEERS vs DATA SCIENTISTS
- BIG DATA, LSTM 120L DEEP
- INSURANCE COMPANIES, ECOMMERCE, FMCG, HR, …

TEAM


Math, Stats,
Algorithms
DATA
ENGINEER
DATA
SCIENTIST
UNICORN
Software
engineering
Data
communication

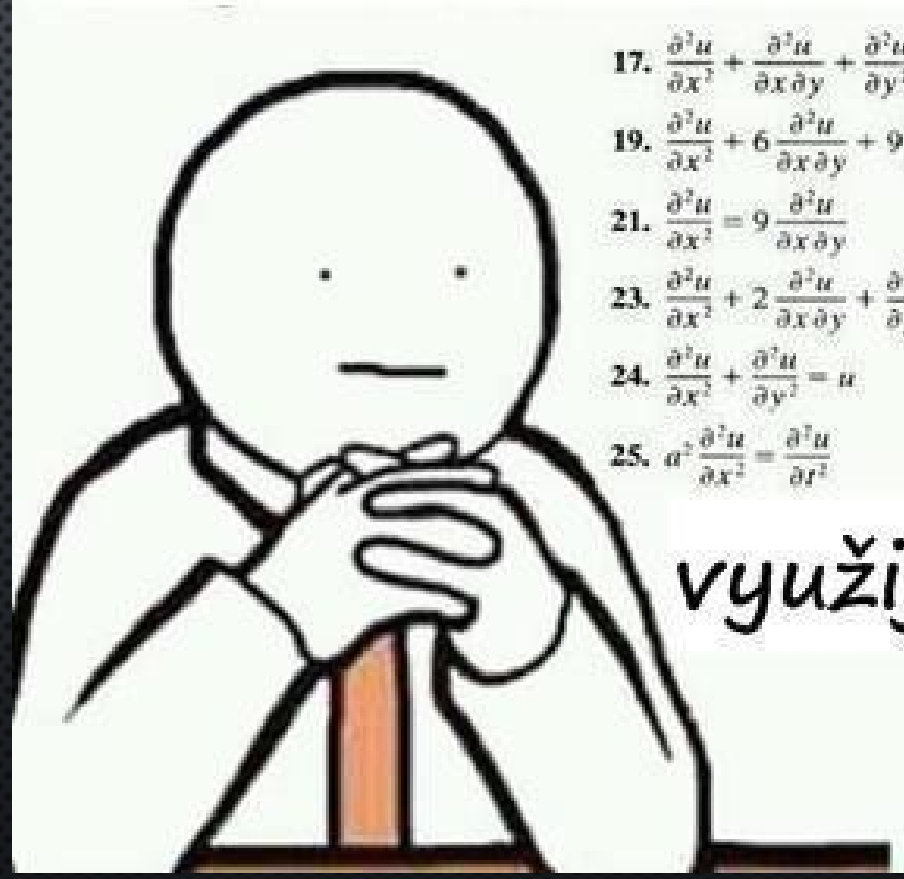
DATA SCIENTIST



…FJFI?

Stále čakám na ten deň,

kedy konečne toto



17. $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} = 0$

18. $3\frac{\partial^2 u}{\partial x^2} + 5\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} = 0$

19. $\frac{\partial^2 u}{\partial x^2} + 6\frac{\partial^2 u}{\partial x \partial y} + 9\frac{\partial^2 u}{\partial y^2} = 0$

20. $\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial x \partial y} - 3\frac{\partial^2 u}{\partial y^2} = 0$

21. $\frac{\partial^2 u}{\partial x^2} = 9\frac{\partial^2 u}{\partial x \partial y}$

22. $\frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^2 u}{\partial y^2} + 2\frac{\partial u}{\partial x} = 0$

23. $\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial x} - 6\frac{\partial u}{\partial y} = 0$

24. $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = u$

25. $a^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$

26. $k\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}, \quad k > 0$

využijem v reálnom

živote...

CLV
ARPU
AOV
CPC
CPA
YOY
RFM
RMSE
ROI…

AMSM

- ROZ (CLUSTERING, PCA)
- TIN, ZLIM (GLM)
- SKE, MEX
- DATS (SQL)
- DRO
- X NEUR, SSI, PRAKTICKE UKAZKY


PLEASE
TELL ME MORE
memegenerator.net

CESTA PROJEKTU

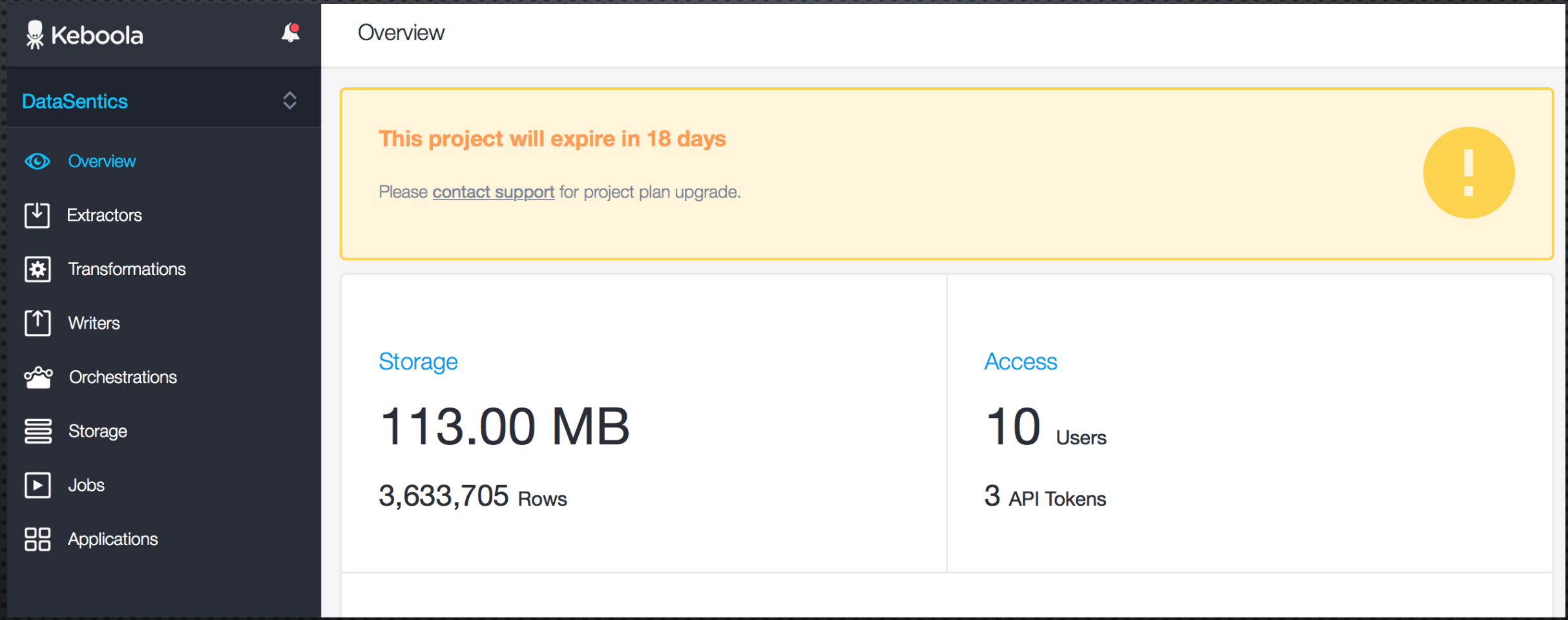
- 1 BUSINESS ANALYZA
- 2 SEHNANI A LOAD DAT
- 3 VIZUALIZACE
- 4 EXPLORACE
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE

CESTA PROJEKTU – BUSINESS ARCHITECT

- 1 BUSINESS ANALYZA – POCHOPENI POTREB ZAKAZNIKA, 80:20, POC
- 2 SEHNANI A LOAD DAT
- 3 VIZUALIZACE
- 4 EXPLORACE
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE

CESTA PROJEKTU – DATA ENGINEER

- 1 BUSINESS ANALYZA
- 2 SEHNANI A LOAD DAT (DB, KBC, CSV)
- 3 VIZUALIZACE
- 4 EXPLORACE
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE - PERSONALIZACE DLE VYSTUPU, POTREB, SYSTEM, PENEZ


Keboola
DataSentics
Overview
Extractors
Transformations
Writers
Orchestrations
Storage
Jobs
Applications
Overview
This project will expire in 18 days
Please contact support for project plan upgrade.
Storage
113.00 MB
3,633,705 Rows
Access
10 Users
3 API Tokens

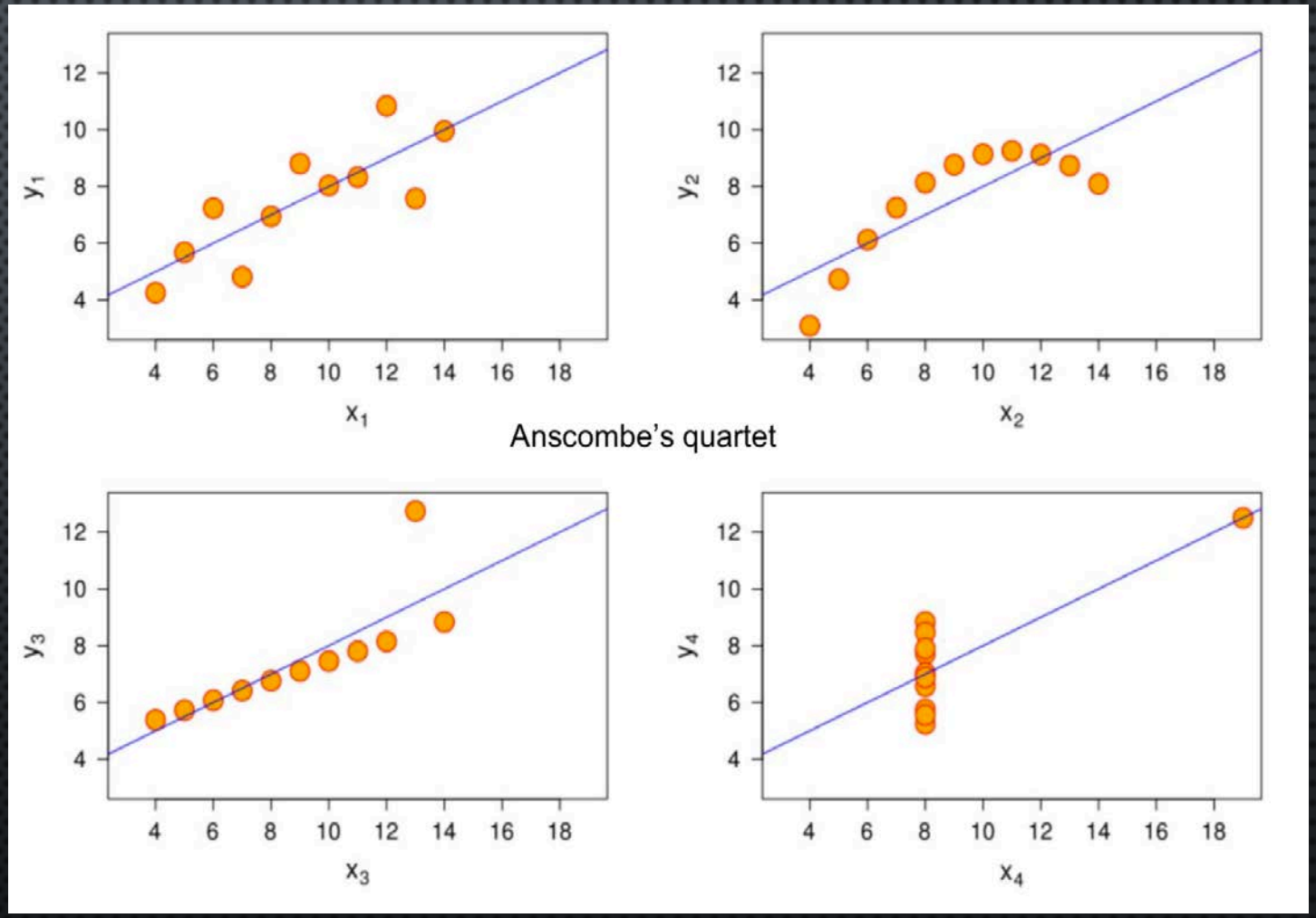
CESTA PROJEKTU – DATA SCIENTIST

- 1 BUSINESS ANALYZA
- 2 SEHNANI A LOAD DAT
- 3 VIZUALIZACE (QLIK, TABLEAU)
- 4 EXPLORACE (R, PYTHON)
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE

3 VIZUALIZACE

… EH, NESTACI CISLA?

3 VIZUALIZACE



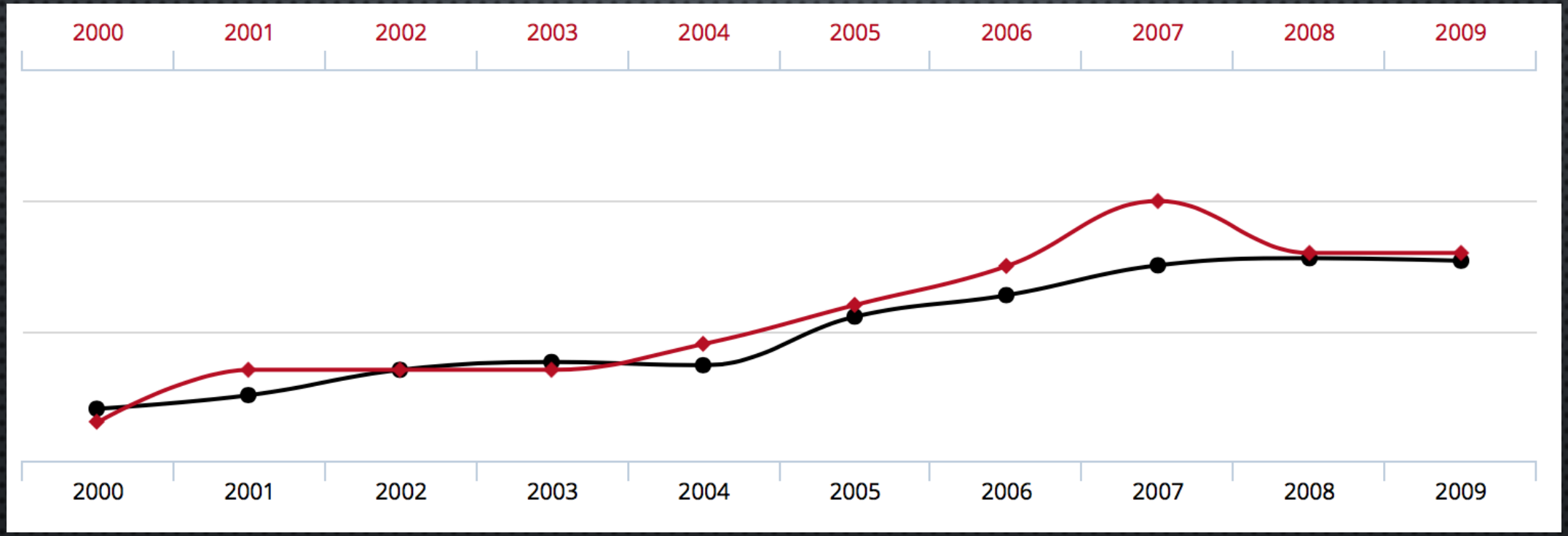
3 VIZUALIZACE

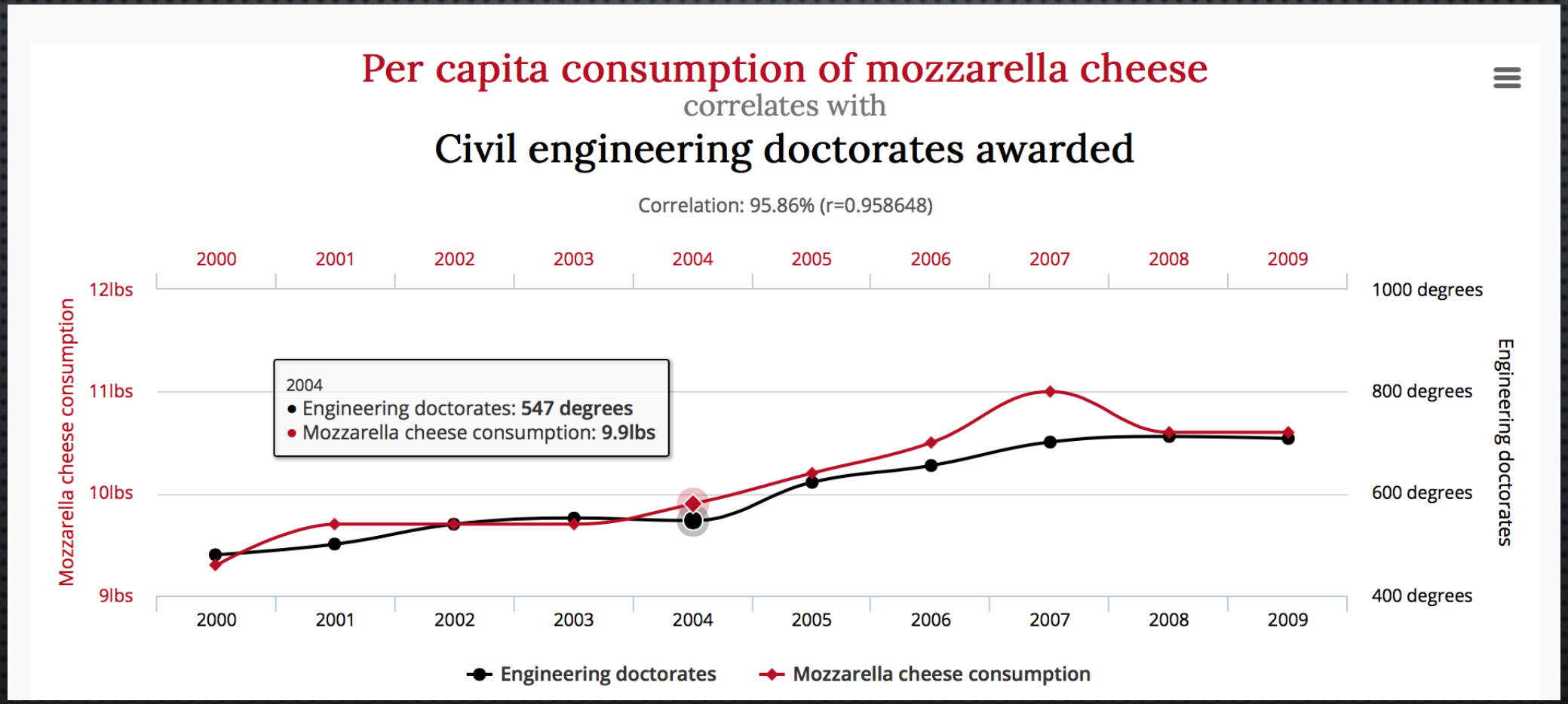
All the summary statistics you'd think to compute are close to identical:

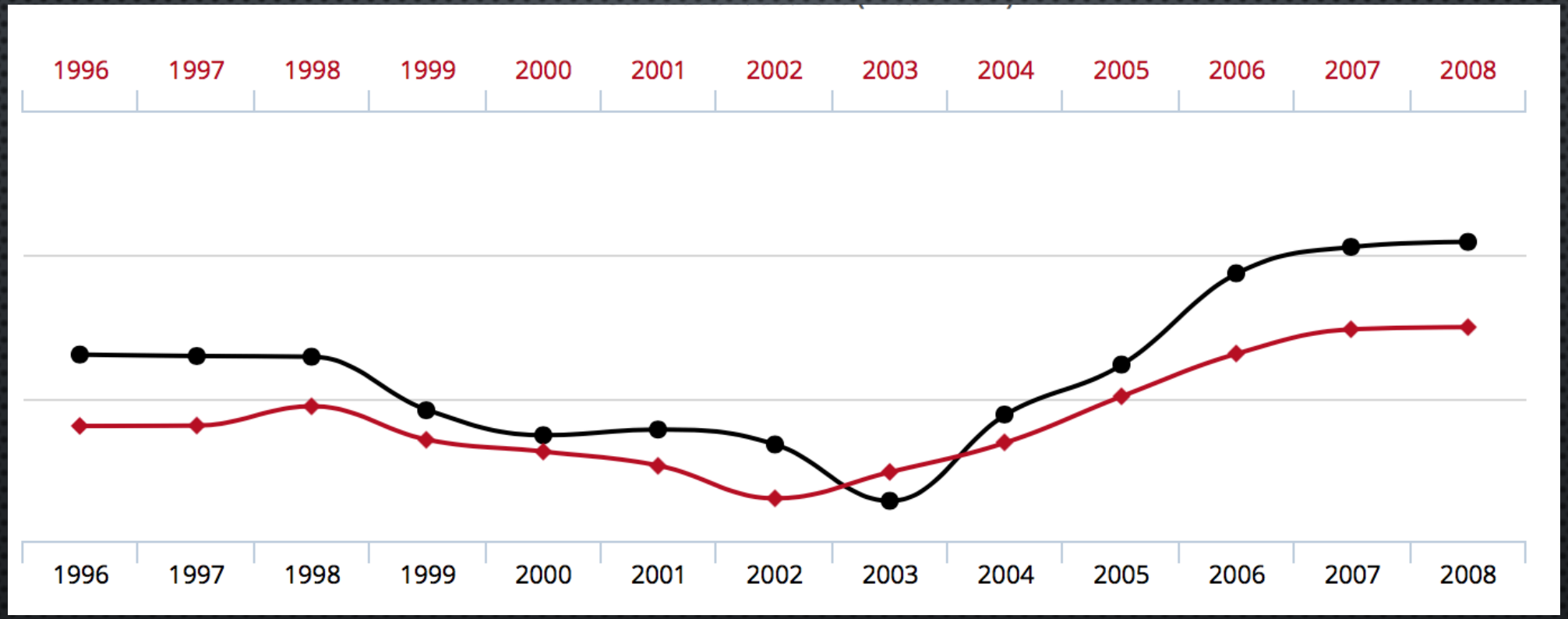
- The average *x* value is 9 for each dataset
- The average *y* value is 7.50 for each dataset
- The variance for *x* is 11 and the variance for *y* is 4.12
- The correlation between *x* and *y* is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation $y = 0.5x + 3$

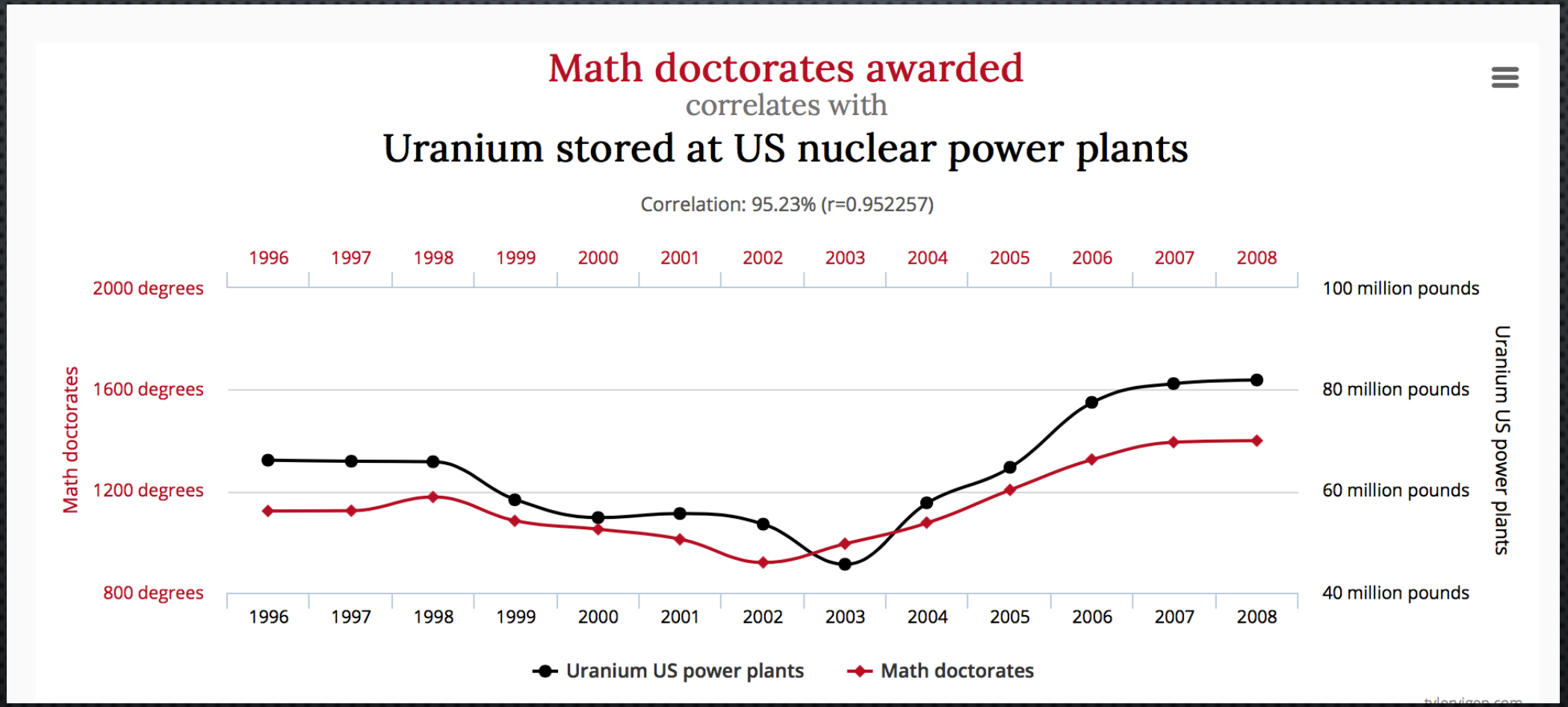
3 VIZUALIZACE

… POZOR NA BUSINESS KONTEXT


2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009


Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Correlation: 95.86% (r=0.958648)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
12lbs
11lbs
10lbs
9lbs
Mozzarella cheese consumption
1000 degrees
800 degrees
600 degrees
400 degrees
Engineering doctorates
2004
Engineering doctorates: 547 degrees
Mozzarella cheese consumption: 9.9lbs
Engineering doctorates
Mozzarella cheese consumption


1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008


Math doctorates awarded
correlates with
Uranium stored at US nuclear power plants
Correlation: 95.23% (r=0.952257)
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
2000 degrees
1600 degrees
1200 degrees
800 degrees
Math doctorates
100 million pounds
80 million pounds
60 million pounds
40 million pounds
Uranium US power plants
Uranium US power plants
Math doctorates

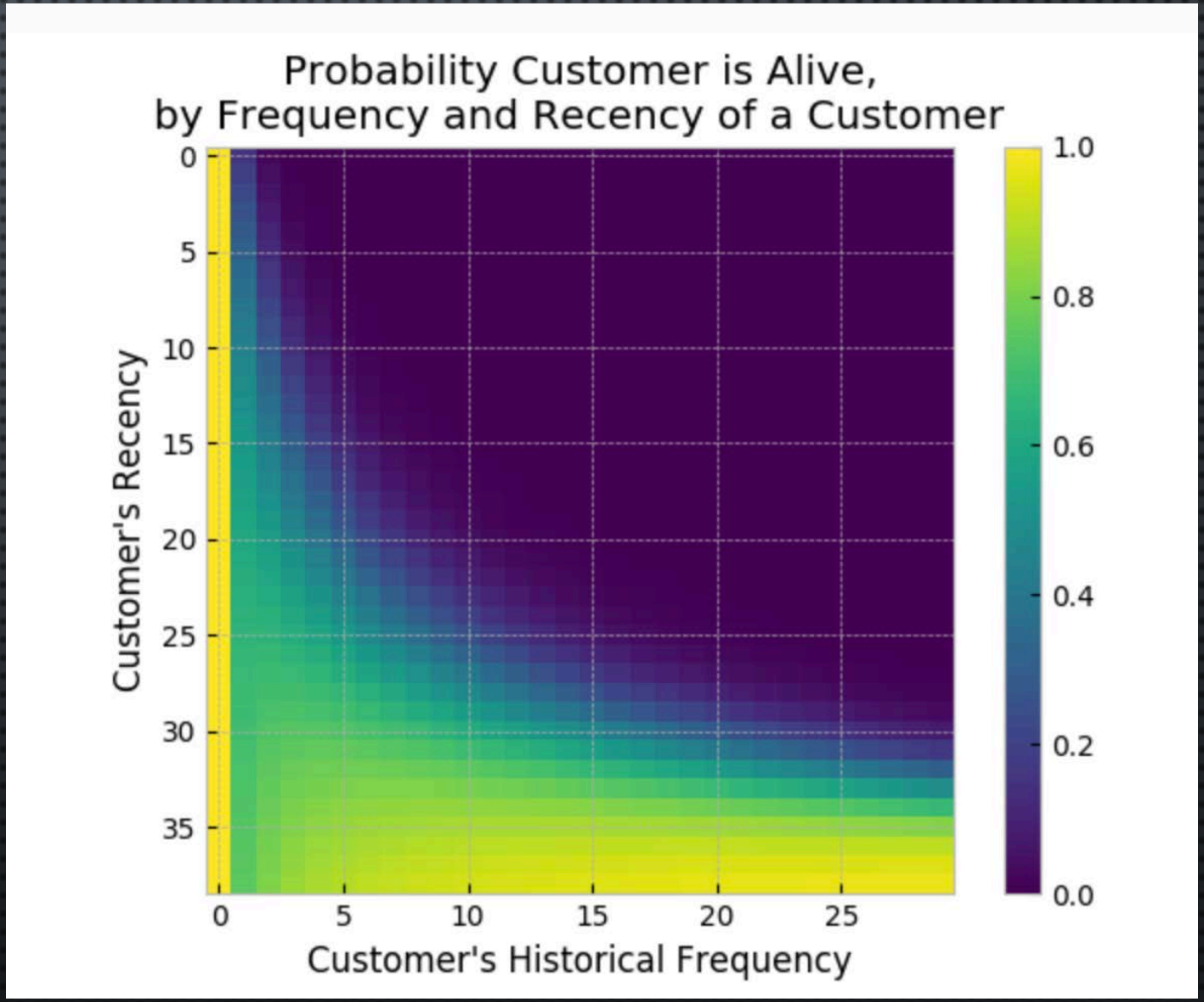
4 EXPLORACE

- CUSTOMER JOURNEY (WEB, NAKUPY)
- SKU, REVENUE MEZI N A N+1 NAKUPEM
- ACTIVE USER BASE

5 DATOVA ANALYZA, MACHINE LEARNING

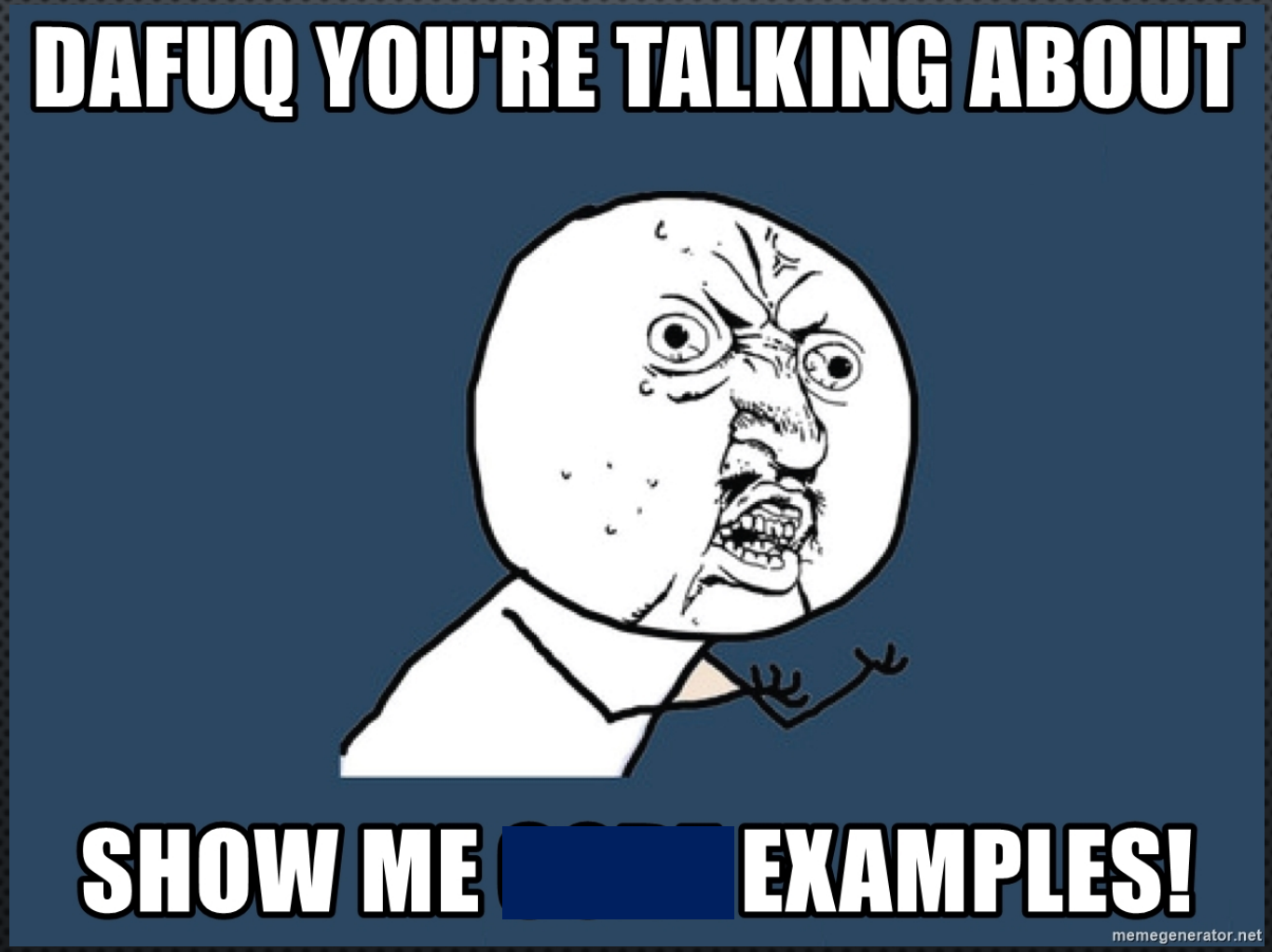
- PROBABILISTIC: CLV, CHURN RATES
- REGRESSION: (G)LM, (BAGGED) RANDOM FORESTS
- CLUSTERING: K-MEANS, SVM, RF
- NLP: TF-IDF, WORD2VEC

CHURN RATES

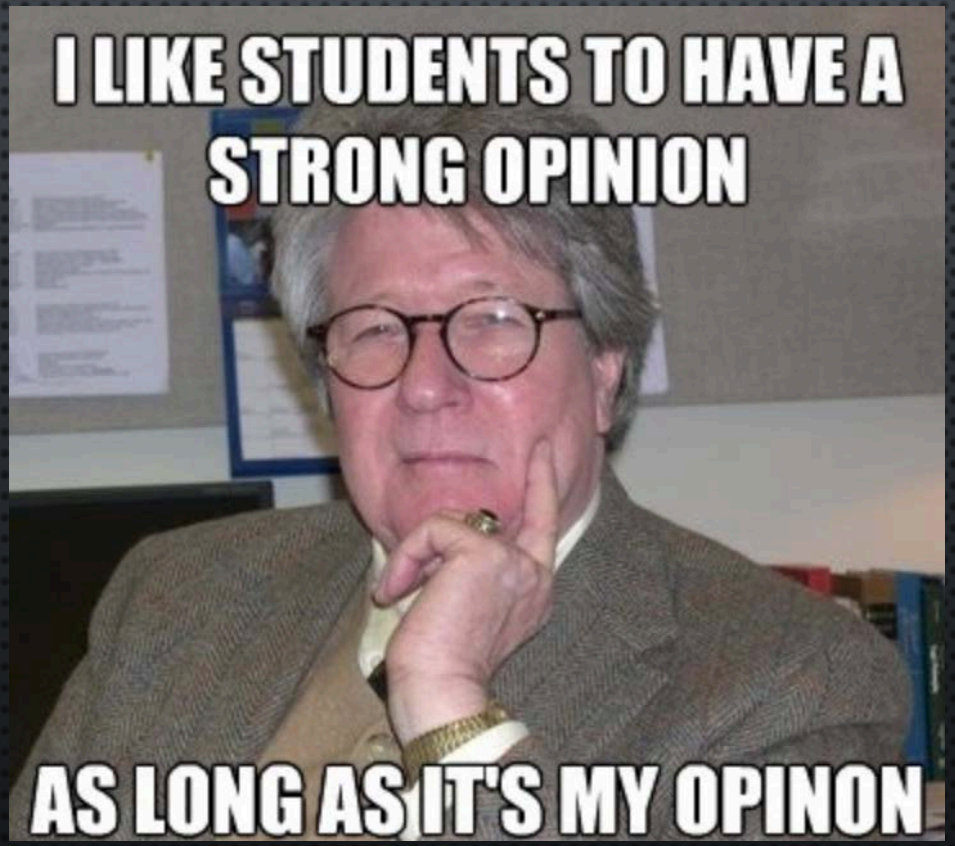

Probability Customer is Alive,
by Frequency and Recency of a Customer
Customer's Recency
0
5
10
15
20
25
30
35
Customer's Historical Frequency
0
5
10
15
20
25
1.0
0.8
0.6
0.4
0.2
0.0

CESTA PROJEKTU – DATA ENGINEER

- 1 BUSINESS ANALYZA
- 2 SEHNANI A LOAD DAT (DB, KBC, CSV)
- 3 VIZUALIZACE
- 4 EXPLORACE
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE - PERSONALIZACE DLE VYSTUPU, POTREB, SYSTEM, PENEZ


DAFUQ YOU'RE TALKING ABOUT
SHOW ME [redacted] EXAMPLES!
memegenerator.net

1 DYNAMICKY DIGITALNI MARKETING



1 DYNAMICKY DIGITALNI MARKETING

VYUZITI VICE DATOVYCH ZDROJU

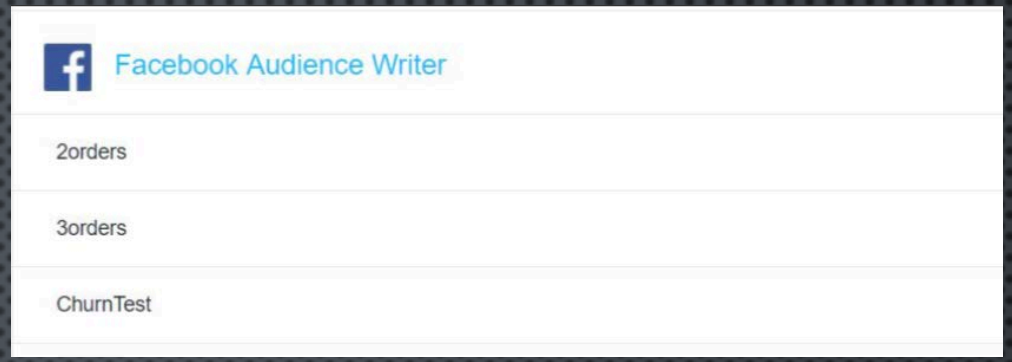
DORUCIT PRESONALIZOVANY OBSAH

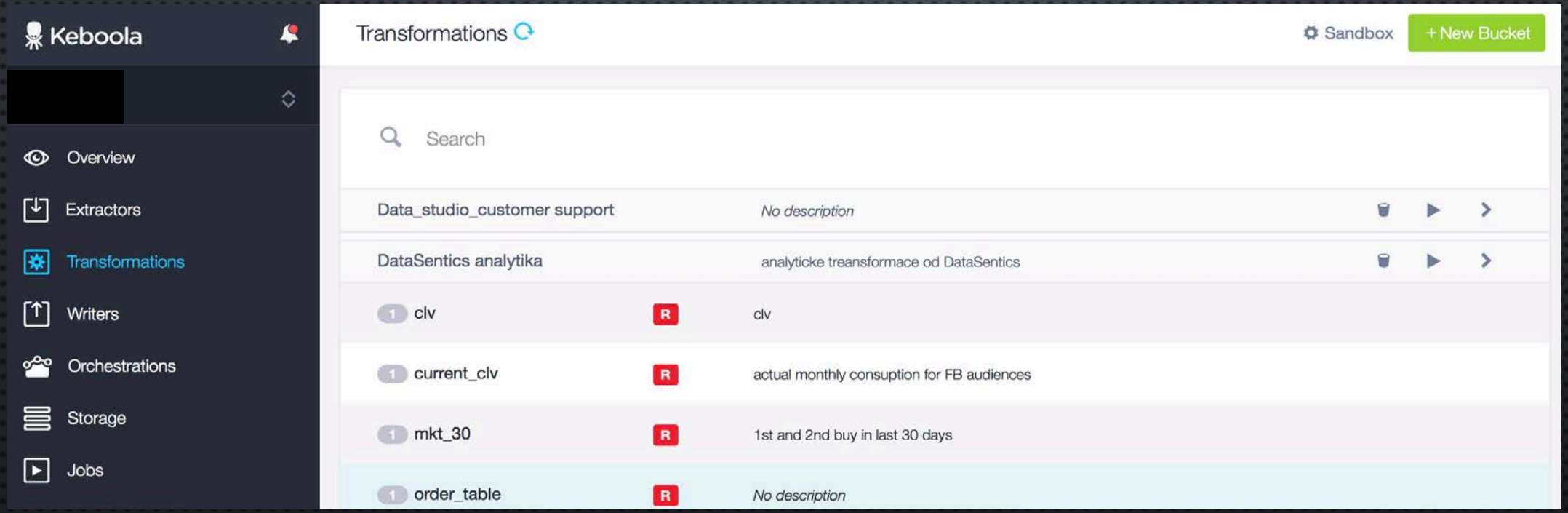
- NIZSI CHURN RATE
- VYSSI CTR, NIZSI CPA => VYSSI REVENUE

AUTOMATIZACE

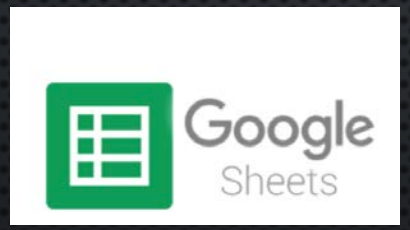
CO NEDELAME

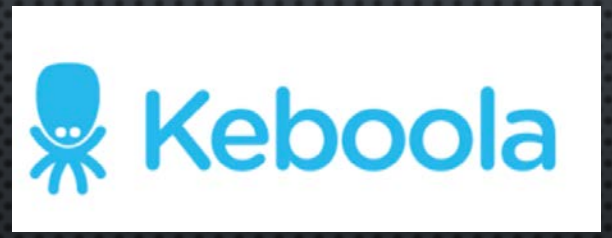
CO DELAME


Facebook Audience Writer
2orders
3orders
ChurnTest


Keboola
Overview
Extractors
Transformations
Writers
Orchestrations
Storage
Jobs
Transformations
Sandbox
+ New Bucket
Search
Data_studio_customer support
No description
DataSentics analytika
analyticke treansformace od DataSentics
clv
clv
current_clv
actual monthly consuption for FB audiences
mkt_30
1st and 2nd buy in last 30 days
order_table
No description

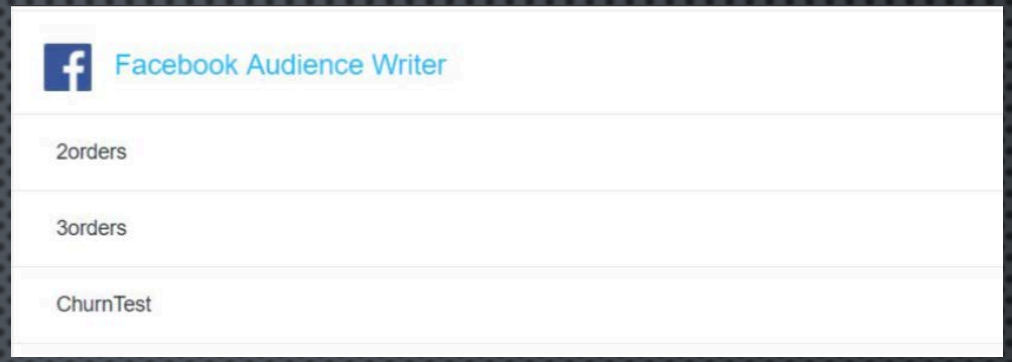
JAK TECOU DATA

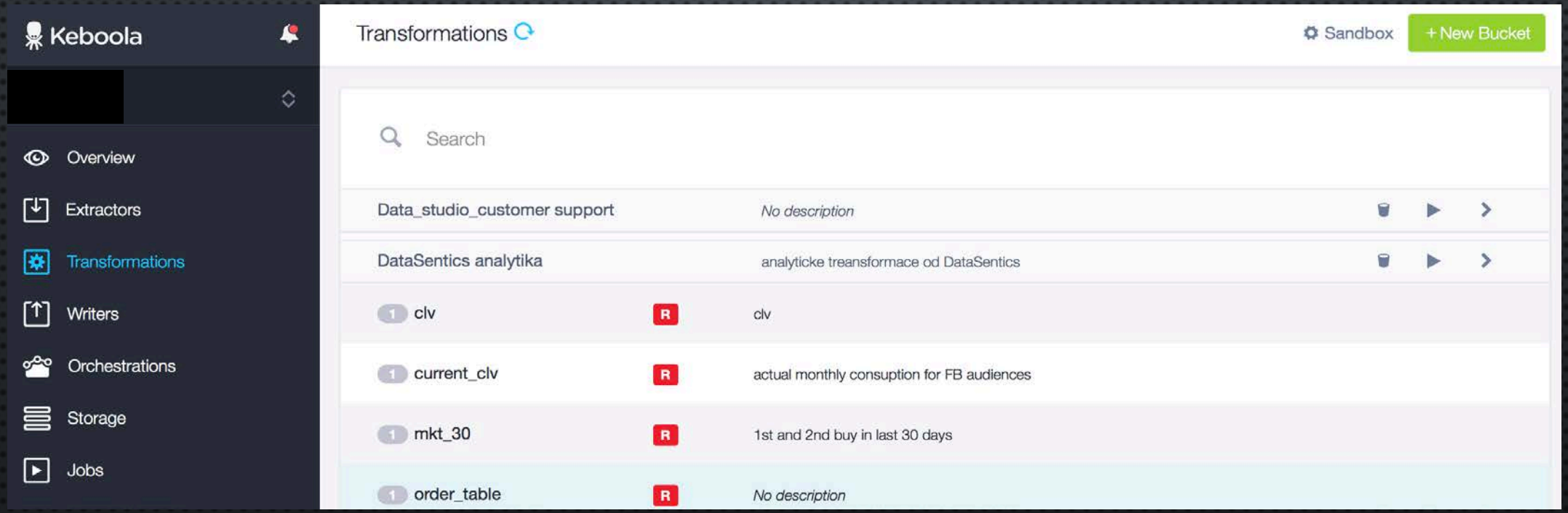




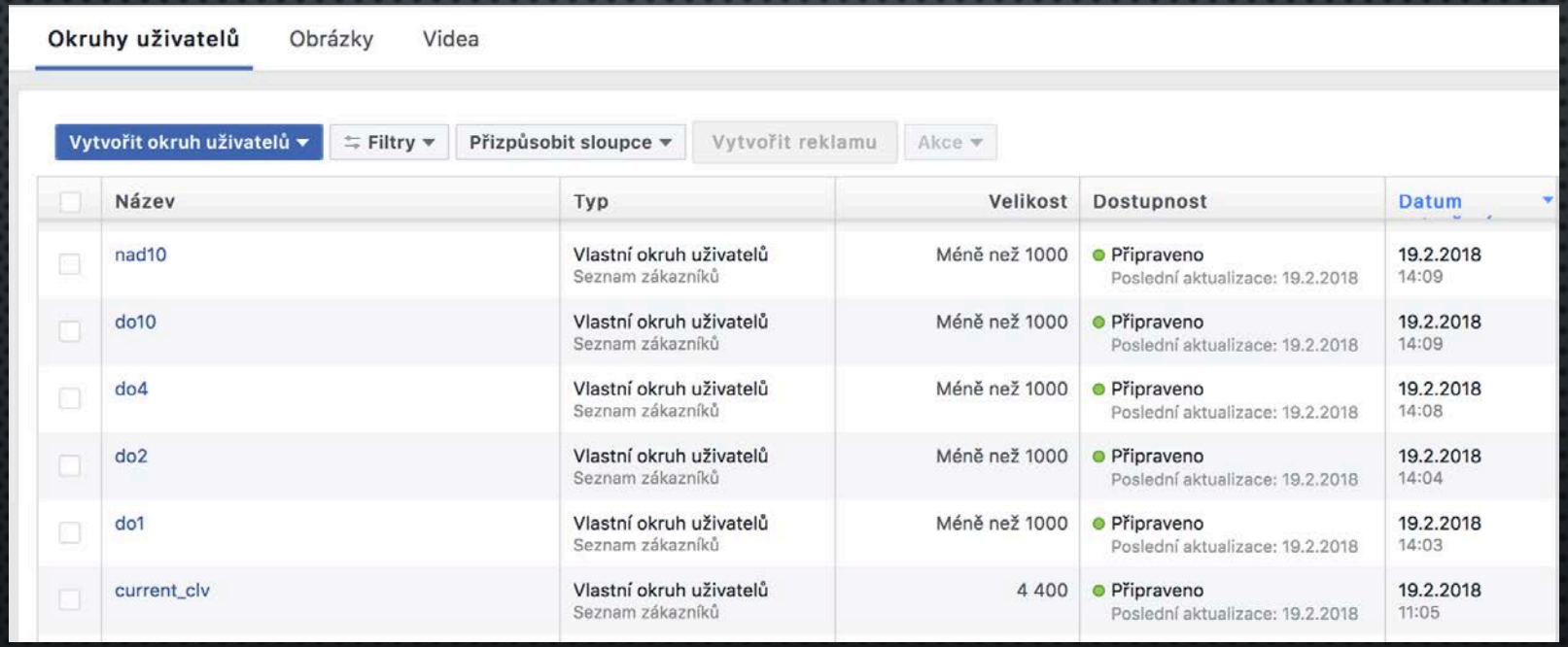
happy marketer

CO DELAME


Facebook Audience Writer
2orders
3orders
ChurnTest


Keboola
Overview
Extractors
Transformations
Writers
Orchestrations
Storage
Jobs
Transformations
Sandbox
+ New Bucket
Search
Data_studio_customer support
No description
DataSentics analytika
analyticke treansformace od DataSentics
clv
clv
current_clv
actual monthly consuption for FB audiences
mkt_30
1st and 2nd buy in last 30 days
order_table
No description

CO DELAME


Okruhy uživatelů
Obrázky
Videa
Vytvořit okruh uživatelů
Filtry
Přizpůsobit sloupce
Vytvořit reklamu
Akce
Název
Typ
Velikost
Dostupnost
Datum
nad10
Vlastní okruh uživatelů
Seznam zákazníků
Méně než 1000
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
14:09
do10
Vlastní okruh uživatelů
Seznam zákazníků
Méně než 1000
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
14:09
do4
Vlastní okruh uživatelů
Seznam zákazníků
Méně než 1000
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
14:08
do2
Vlastní okruh uživatelů
Seznam zákazníků
Méně než 1000
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
14:04
do1
Vlastní okruh uživatelů
Seznam zákazníků
Méně než 1000
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
14:03
current_clv
Vlastní okruh uživatelů
Seznam zákazníků
4 400
Připraveno
Poslední aktualizace: 19.2.2018
19.2.2018
11:05

rohlik.cz
31 říjen v 9:38 ·

Oblíbené pesto Barilla v akci za 69,90 Kč! Skvělé jídlo s ním vykouzlíte za pár minut👌

Objevte naší cenovku dne!

Nakupujte zde >>>

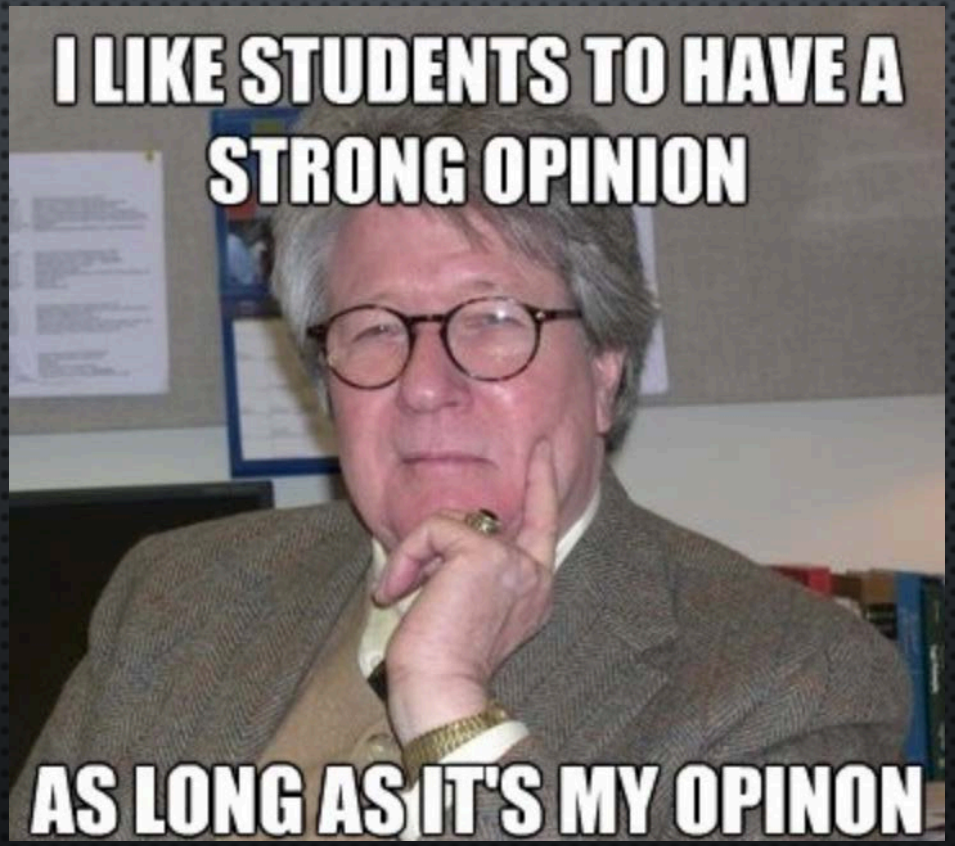
ROHLIK.CZ

Super Komentář Sdílet

Vy a 8 dalších

Hlavní komentáře

2 TAXI


I LIKE STUDENTS TO HAVE A STRONG OPINION
AS LONG AS IT'S MY OPINON

2 SENSITIVITA ZAKAZNIKU

EXTRAKCE PREFERENCI ZAKAZNIKU

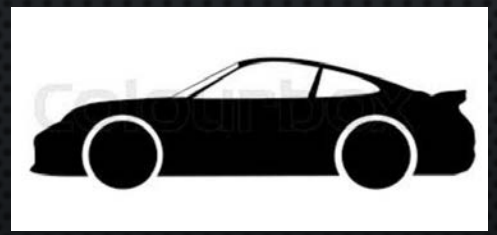
- NIZSI CHURN RATE
- ZVYSENI OBSAZENOSTI
- ZVYSENI REVENUE

CC POJISTOVNA, TAXISLUZBA

GLM, (RF)

TAXI PROBLEM

A → B



Cena
Prijezd
Delka jizdy
Komfort auta
Hodnoceni ridice

Taxi model

- Logisticka regrese
- P(jizda = accepted) = a * Cena + b * Prijezd + c *Delka jizdy + D * Komfort auta
- Bussiness jizdy = kladne “a”
- Nebussiness = zaporne “b” kam az zdrazit viz thresh

OTAZKY…?

Trees, Random Forest, Gradient Boosted Trees

Aplikace v pojišťovnách

Business motivace

- Jaké otázky řeší pojišťovny a jaké data mají?

Business motivace

- Jaké otázky řeší pojišťovny a jaká data mají?

Data

- Data klientská – nastavení smlouvy, počet smluv, klientská historie, demografické údaje, interakce se zákazníkem – CRM databáze o cca 400 sloupcích
- Obohacení o statistiky, průzkumy, banko produkty

Business motivace

- Jaké otázky řeší pojišťovny a jaká data mají?

Otázky

- Reporting ☺
- Cross sell, Up sell, správný čas na oslovení klienta
- Optimální nastavení produktu
- Predikce rizikových skupin atd.

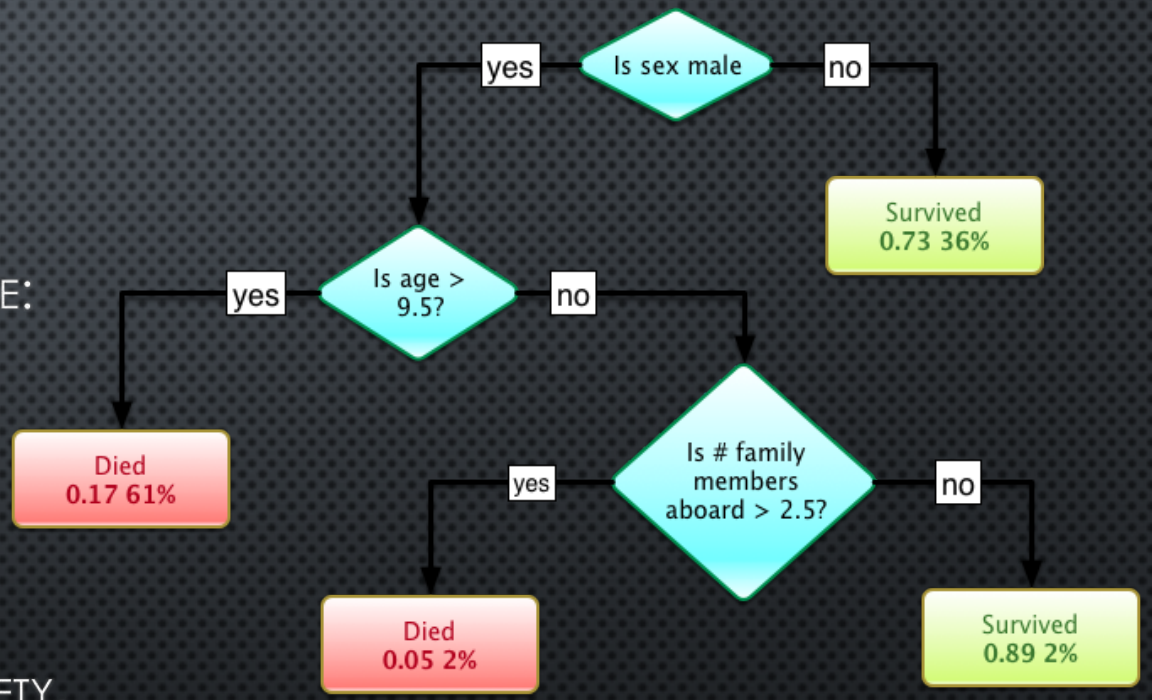
Machine learning

- Obecná koncepce:
- F(x)=y
- Bias – variance tradeoff (underfitting, overfitting)

STROMY

RŮZNÉ ALGORITMY PRO ROZDĚLENÍ DO LISTŮ – NEJBĚŽNĚJŠÍ MAXIMÁLNÍ NÁRŮST INFORMAČNÍ ENTROPIE:

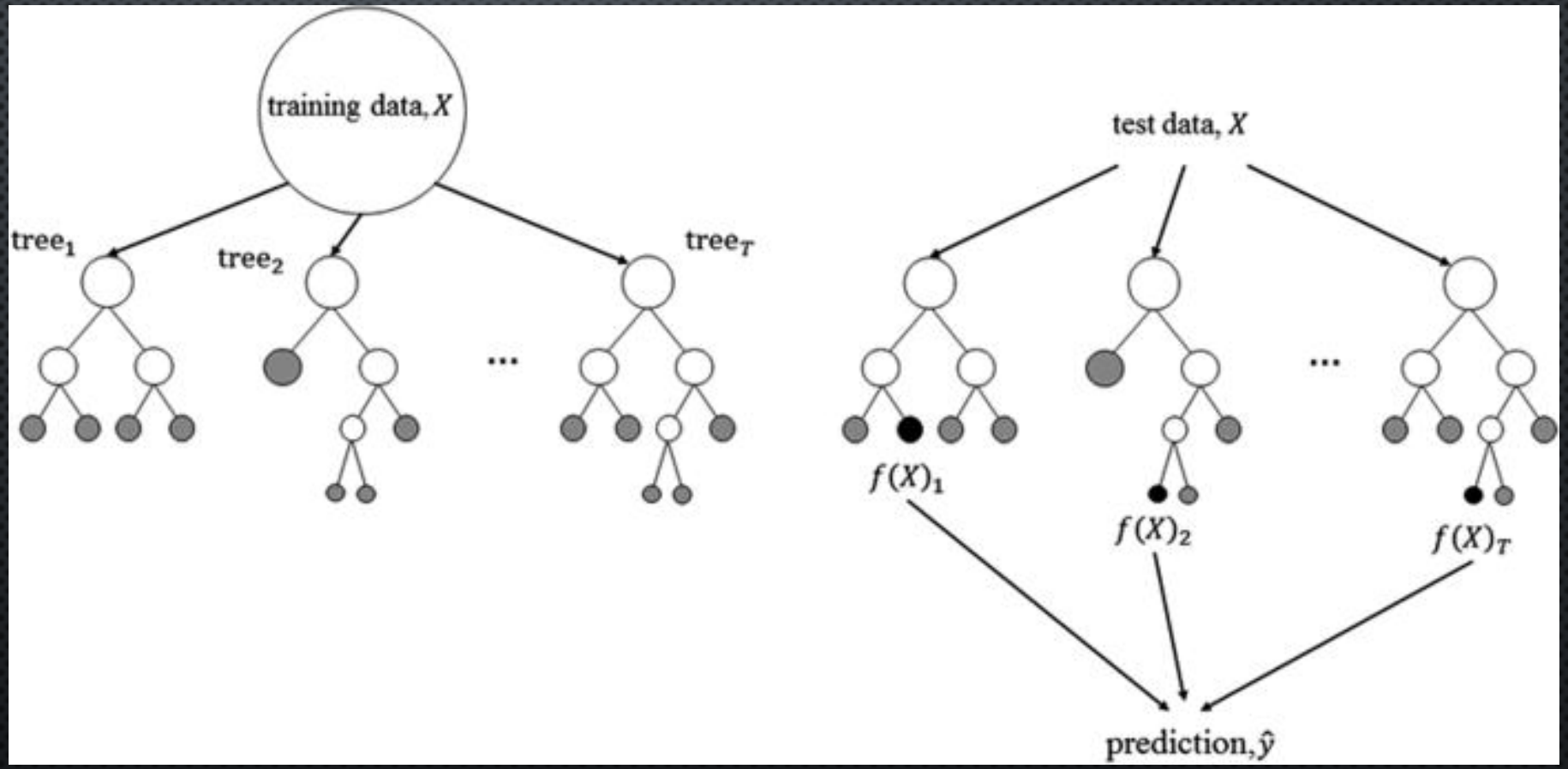
$$H(T) = I_E(P_1, P_2, .., P_J) = \sum_{i=1}^{j} P_I log_2 p_i$$



- VÝHODY: LZE DO NĚJ VLOŽIT KATEGORICKÉ I NUMERICKÉ HODNOTY, SNADNÁ INTERPRETOVATELNOST, FUNGUJE PRO VELKÉ DATASETY, DATA NENÍ TŘEBA MOC PŘEDUPRAVOVAT
- NEVÝHODY: NON-ROBUST, OVERFITTING, NEMUSÍ

RANDOM FOREST

NOVÉ STROMY VZNIKAJÍ SAMPLOVÁNÍM DATASETU/PROMĚNNÝCH



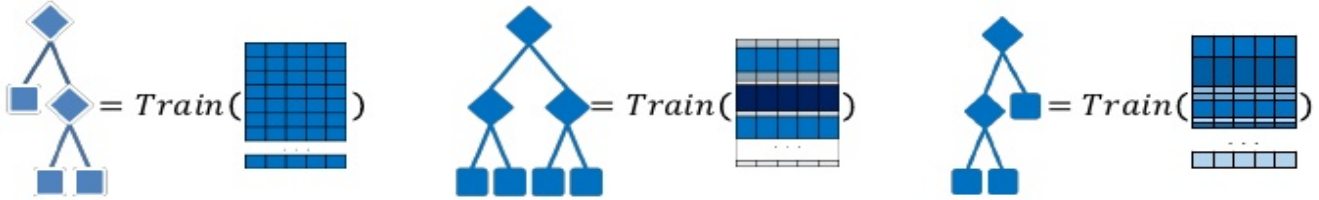
GRADIENT BOOSTED TREES

- JEDEN Z NEJLEPŠÍCH ALGORITMŮ PRO KLASIFIKACI – KAGGLE
- OPEN SOURCE IMPLEMENTACE PRO R A PYTHON (A DALŠÍ)
- PARALELNÍ -> RYCHLÉ
- UMÍ POSTIHNOUT I KOMPLEXNÍ PROBLÉMY
- DOCELA SLUŠNĚ INTERPRETOVATELNÉ

Boosting: Iterative Tree Construction

"Best off-the-shelf classifier in the world" – Breiman

- Reweight examples for each subsequent tree to focus on **errors**



- Numerically: gradient descent in function space
  - Each subsequent tree approximates a step in $-\frac{\partial L}{\partial f}$ direction
  - Recompute *target labels* $y^{(m)} = -\left[\frac{\partial L(y,f(x))}{\partial f(x)}\right]_{f(x)=f^{(m-1)}(x)}$
  - Logistic loss: $L(y,f(x)) = \log(1+\exp(-yf(x))$ $y^{(m)} = \frac{y}{1+\exp(yf(x))}$
  - Squared loss: $L(y,f(x)) = \frac{1}{2}(y-f(x))^2$ $y^{(m)} = y - f(x)$

Vyhodnocení modelů

- Roc Křivka

Co to znamená v praxi

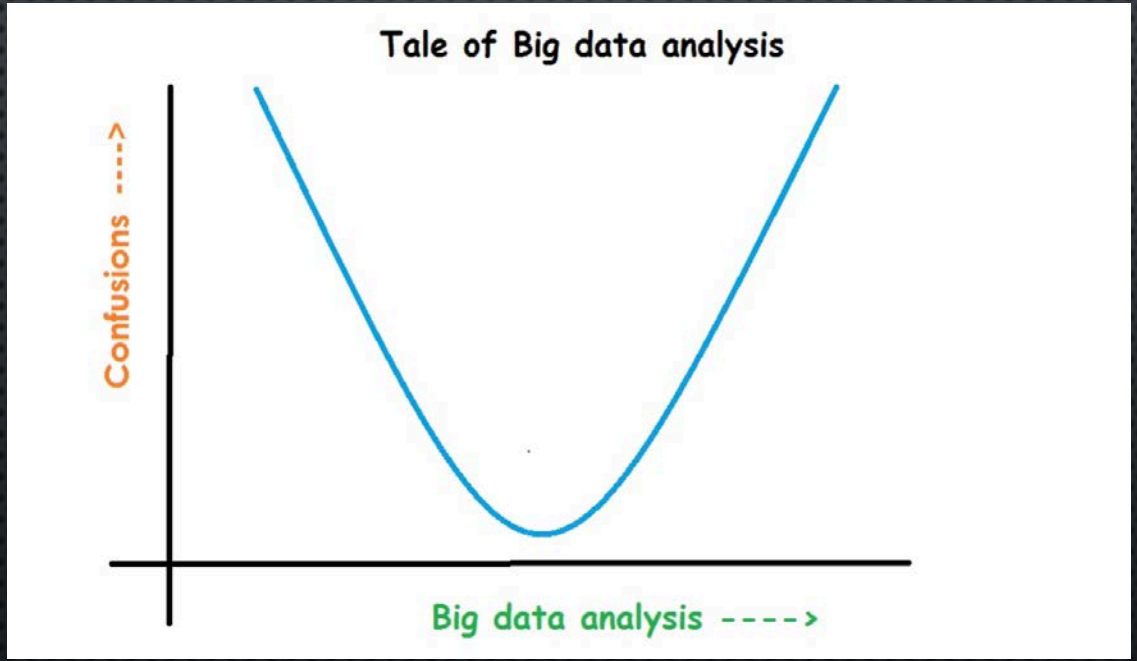
- Mít k dispozici data v analytické databázi
- Mít napočítanou L2 vrstvu crm
- 1000 – 2000 řádků v Rku (takový drobný základ)
- Důraz na interpretaci
- Až sedminásobně větší úspěšnost při oslovení, nespamujeme všechny

OTAZKY…?

Zajimave zdroje

- Datacamp
- https://www.r-bloggers.com/
- Meetup Praha
- Data Analysts, Data Engineers & Data Scientists - Czech&Slovak Group
- Jakub.stech@datasentics.com

DEKUJEME ZA POZORNOST



KDE JSME SKONCILI

- FJFI DOBRY PREDPOKLAD
- DRO
- DATA SCIENCE WORKFLOW

CESTA PROJEKTU – DATA SCIENTIST

- 1 BUSINESS ANALYZA
- 2 SEHNANI A LOAD DAT
- 3 VIZUALIZACE (QLIK, TABLEAU)
- 4 EXPLORACE (R, PYTHON)
- 5 DATOVA ANALYZA, MACHINE LEARNING
- 6 PRODUKCIONALIZACE

rohlik.cz
31 říjen v 9:38 ·

Oblíbené pesto Barilla v akci za 69,90 Kč! Skvělé jídlo s ním vykouzlíte za pár minut👌

Objevte naší cenovku dne!

Nakupujte zde >>>

ROHLIK.CZ

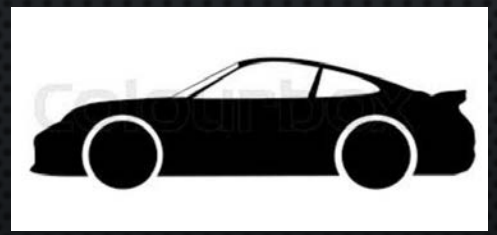
Super Komentář Sdílet

Vy a 8 dalších

Hlavní komentáře

TAXI PROBLEM

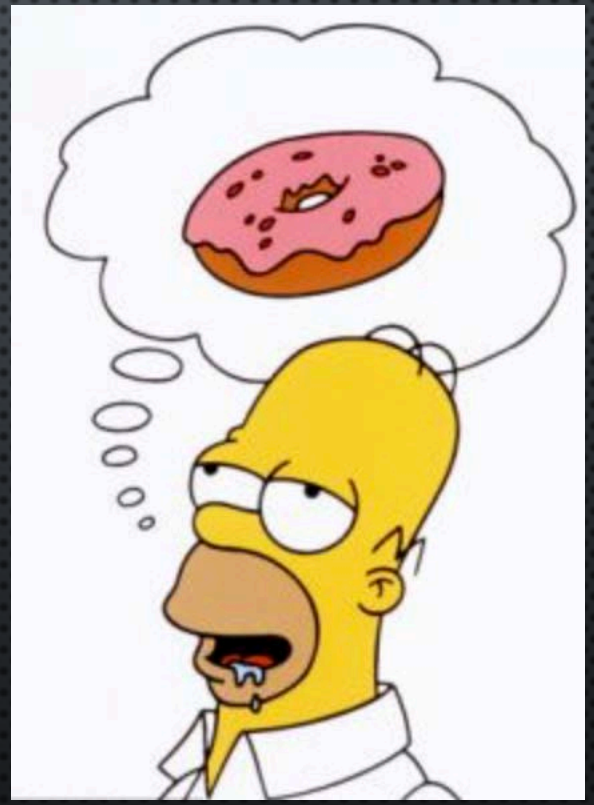
A → B



Cena
Prijezd
Delka jizdy
Komfort auta
Hodnoceni ridice

KDE JSME SKONCILI

…KOHO TO ZAUJALO?



CO DNES?

- GLM v POJISTOVNE
- PROBABILISTIC CHURN RATE + CLV, DATASENTICS CLV v ECOMMERCE

1 CALLCENTRUM POJISTOVNA

SITUACE

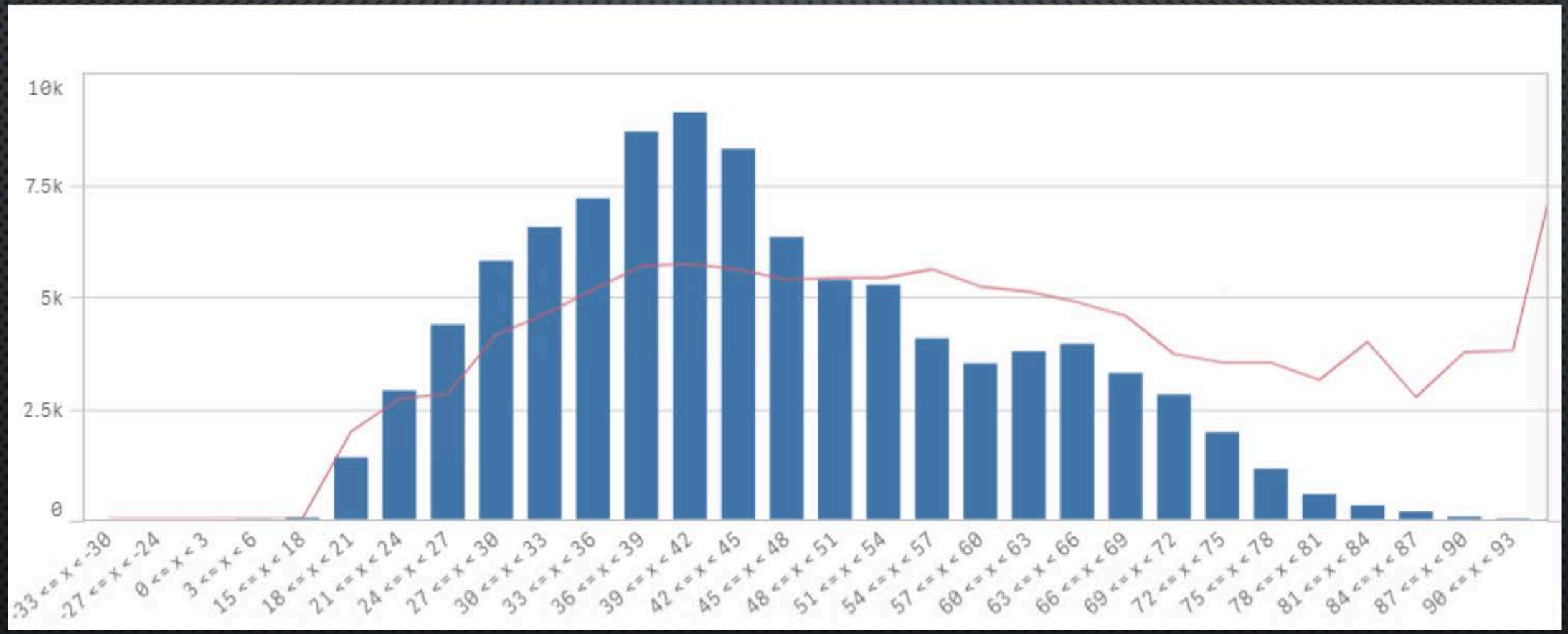
- KALKULACE NA WEBU –> PRVNI NABIDKA
- KDYZ NESJEDNA -> CALLCENTRUM -> DRUHA NABIDKA

1 CALLCENTRUM POJISTOVNA

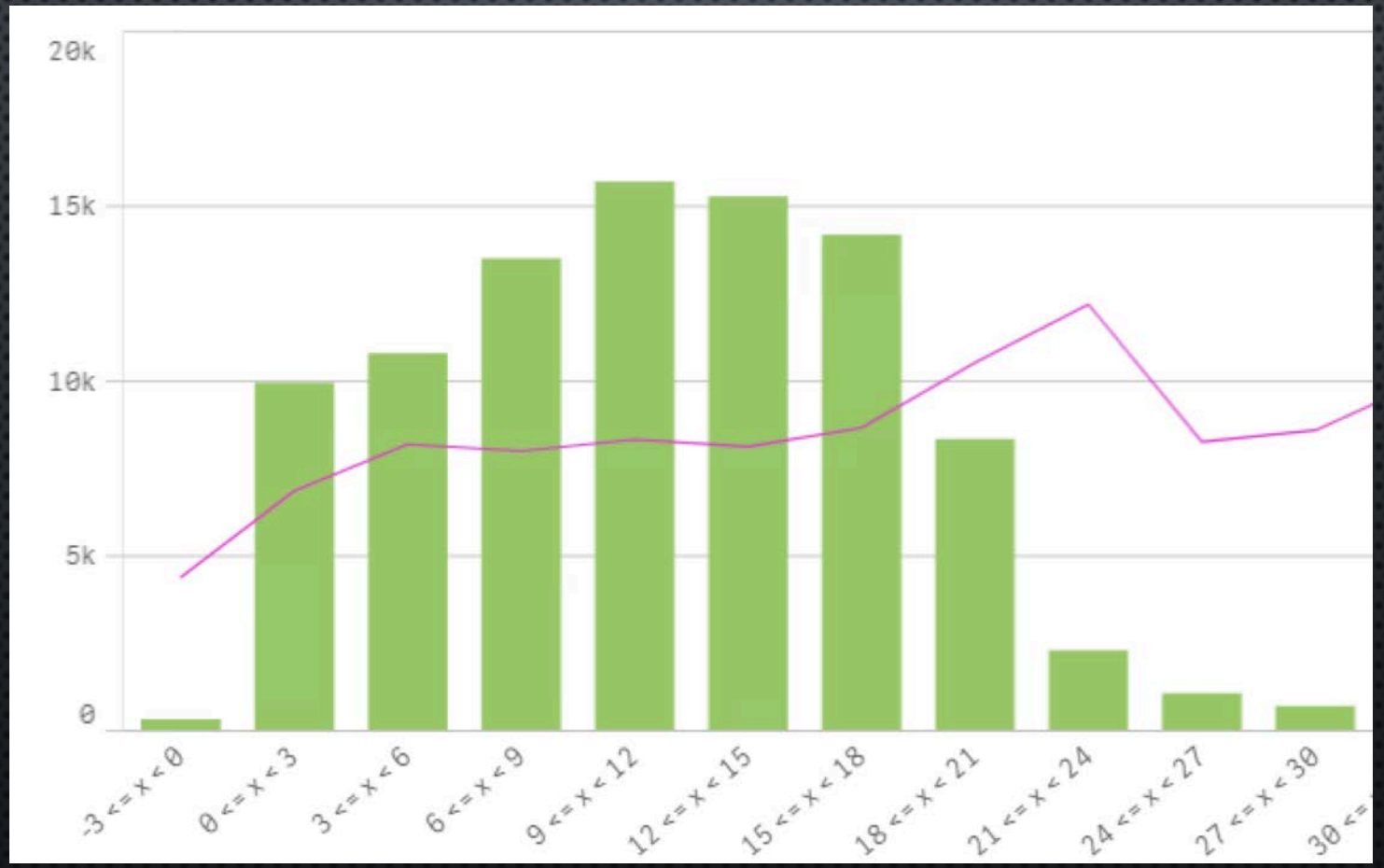
BUSINESS MOTIVACE

- UPRAVA/PERSONIFIKACE NABIDKY NA WEBU/CALLCENTRU
- PRIORITIZACE NAVOLAVANI
- UPSELL HAV, PPC, …

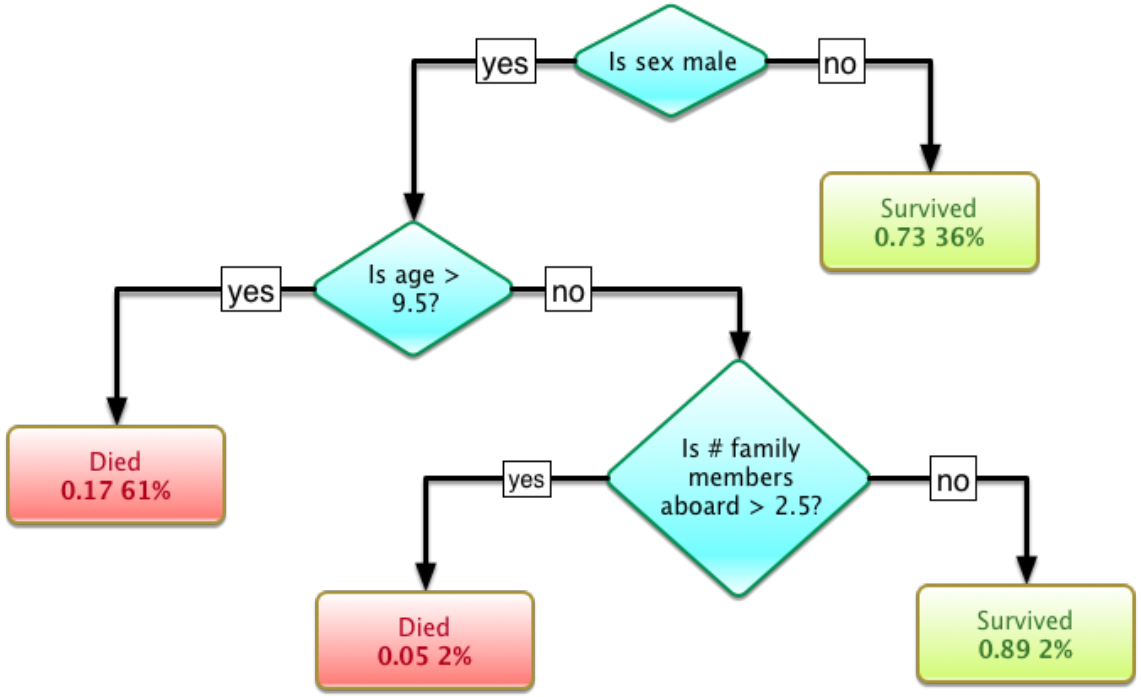
VEK


10k
7.5k
5k
2.5k
0
-33 <= x < -30
-27 <= x < -24
0 <= x < 3
3 <= x < 6
15 <= x < 18
18 <= x < 21
21 <= x < 24
24 <= x < 27
27 <= x < 30
30 <= x < 33
33 <= x < 36
36 <= x < 39
39 <= x < 42
42 <= x < 45
45 <= x < 48
48 <= x < 51
51 <= x < 54
54 <= x < 57
57 <= x < 60
60 <= x < 63
63 <= x < 66
66 <= x < 69
69 <= x < 72
72 <= x < 75
75 <= x < 78
78 <= x < 81
81 <= x < 84
84 <= x < 87
87 <= x < 90
90 <= x < 93

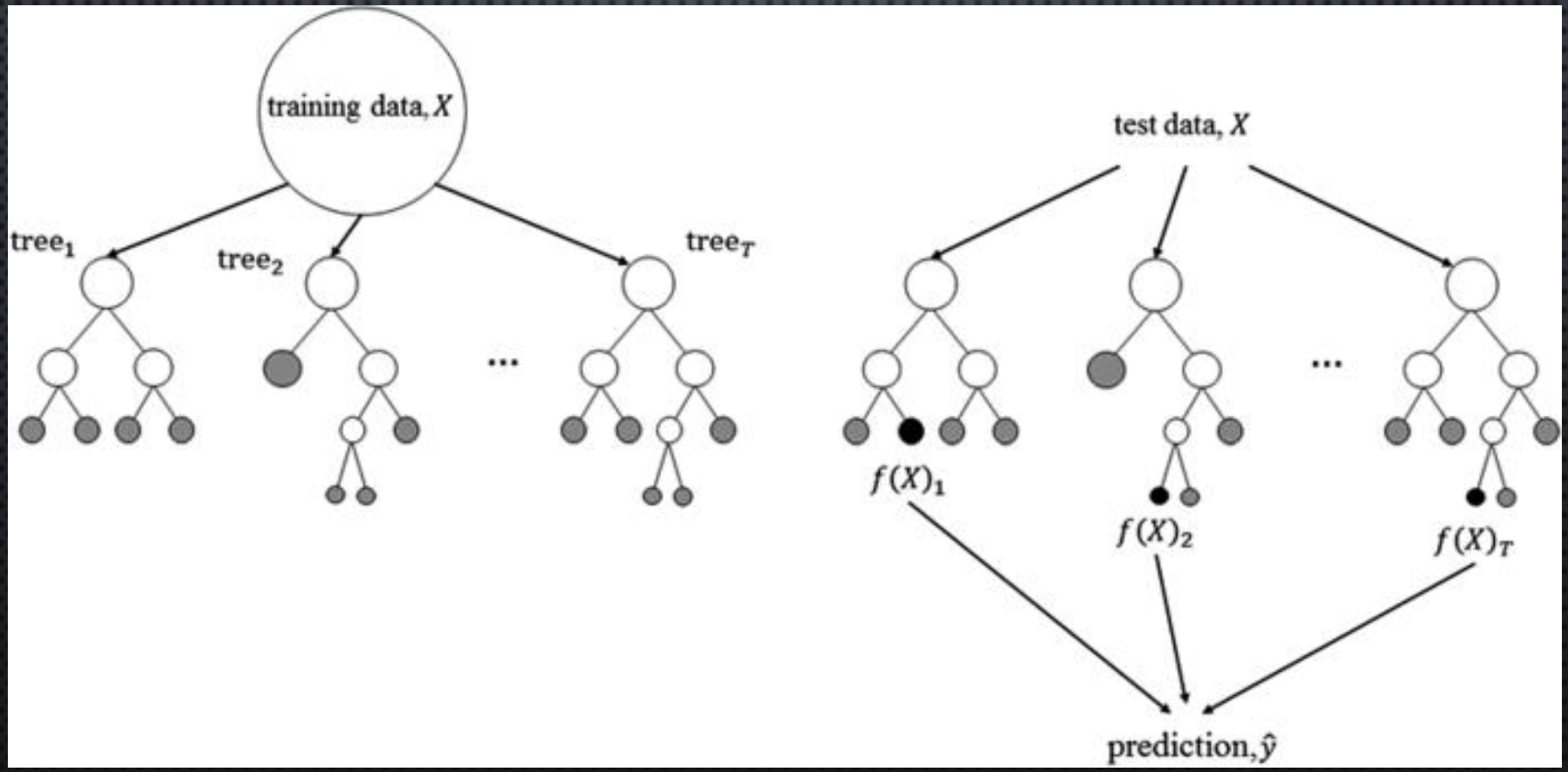
STARI AUTA


20k
15k
10k
5k
0
-3 <= x < 0
0 <= x < 3
3 <= x < 6
6 <= x < 9
9 <= x < 12
12 <= x < 15
15 <= x < 18
18 <= x < 21
21 <= x < 24
24 <= x < 27
27 <= x < 30

STROMY


Is sex male
yes
no
Survived
0.73 36%
Is age >
9.5?
yes
no
Died
0.17 61%
Is # family
members
aboard > 2.5?
yes
no
Died
0.05 2%
Survived
0.89 2%

RANDOM FOREST

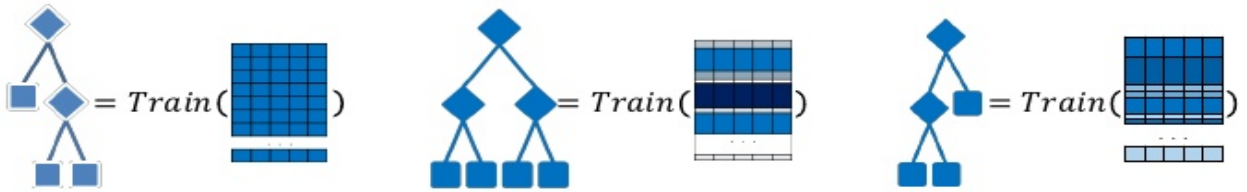


GRADIENT BOOSTED TREES

Boosting: Iterative Tree Construction

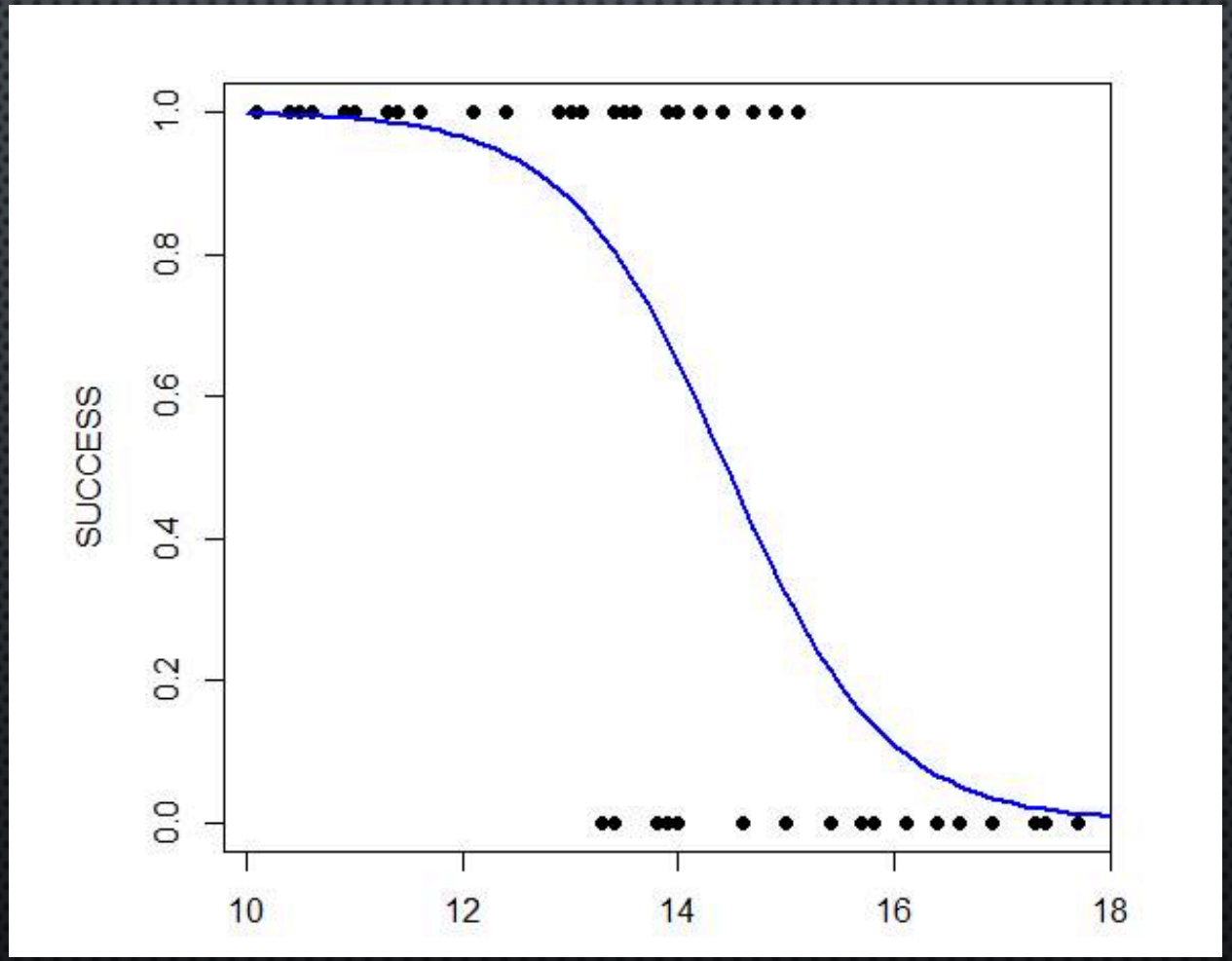
"Best off-the-shelf classifier in the world" – Breiman

- Reweight examples for each subsequent tree to focus on **errors**



- Numerically: gradient descent in function space
  - Each subsequent tree approximates a step in $-\frac{\partial L}{\partial f}$ direction
  - Recompute *target labels* $y^{(m)} = -\left[\frac{\partial L(y,f(x))}{\partial f(x)}\right]_{f(x)=f^{(m-1)}(x)}$
  - Logistic loss: $L(y,f(x)) = \log(1+\exp(-yf(x))$ $y^{(m)} = \frac{y}{1+\exp(yf(x))}$
  - Squared loss: $L(y,f(x)) = \frac{1}{2}(y-f(x))^2$ $y^{(m)} = y - f(x)$

GLM



GLM

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

GLM

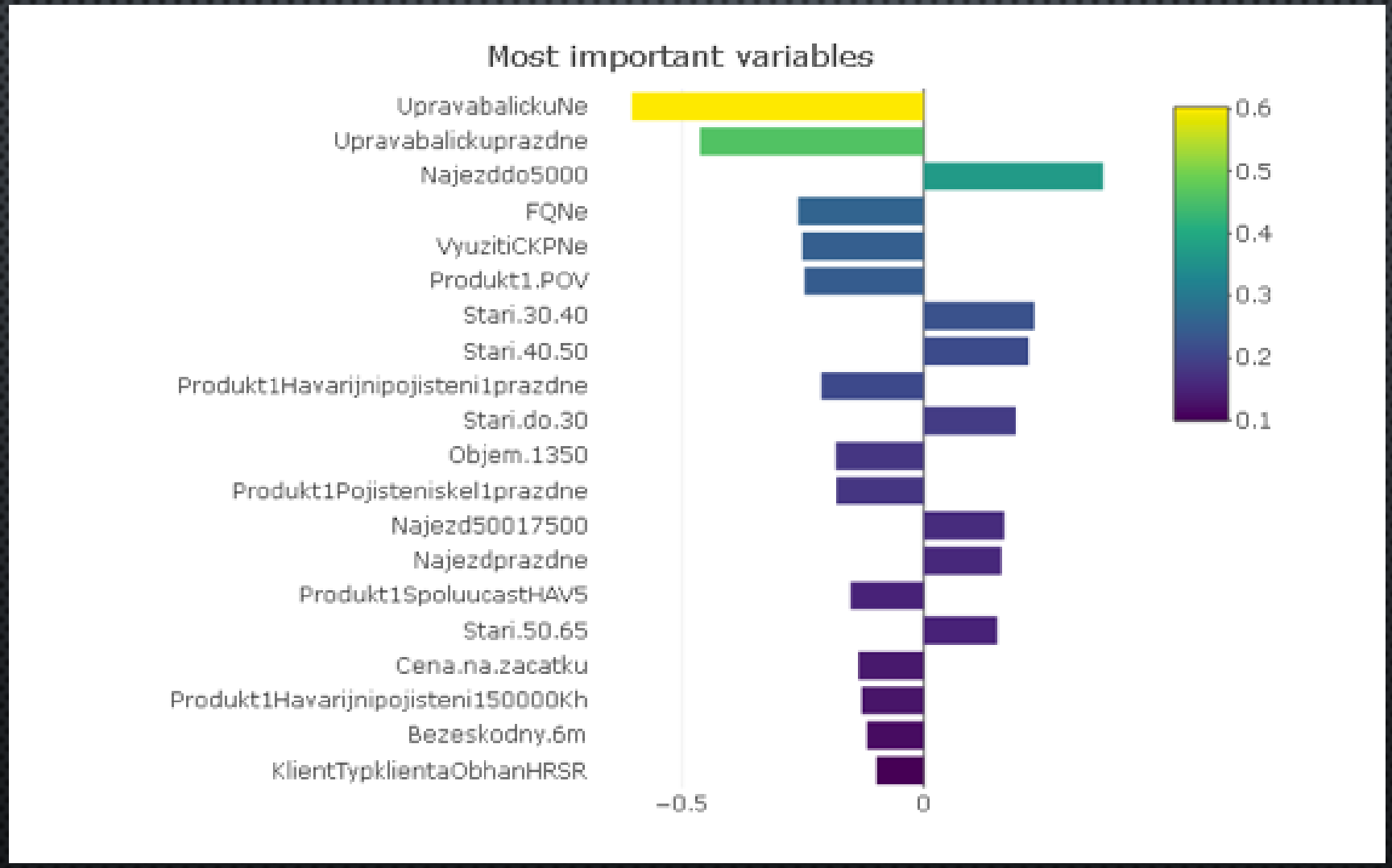
- P/(1-P) = ODDS

$$odds = e^{(b0 + b1 * X)}$$

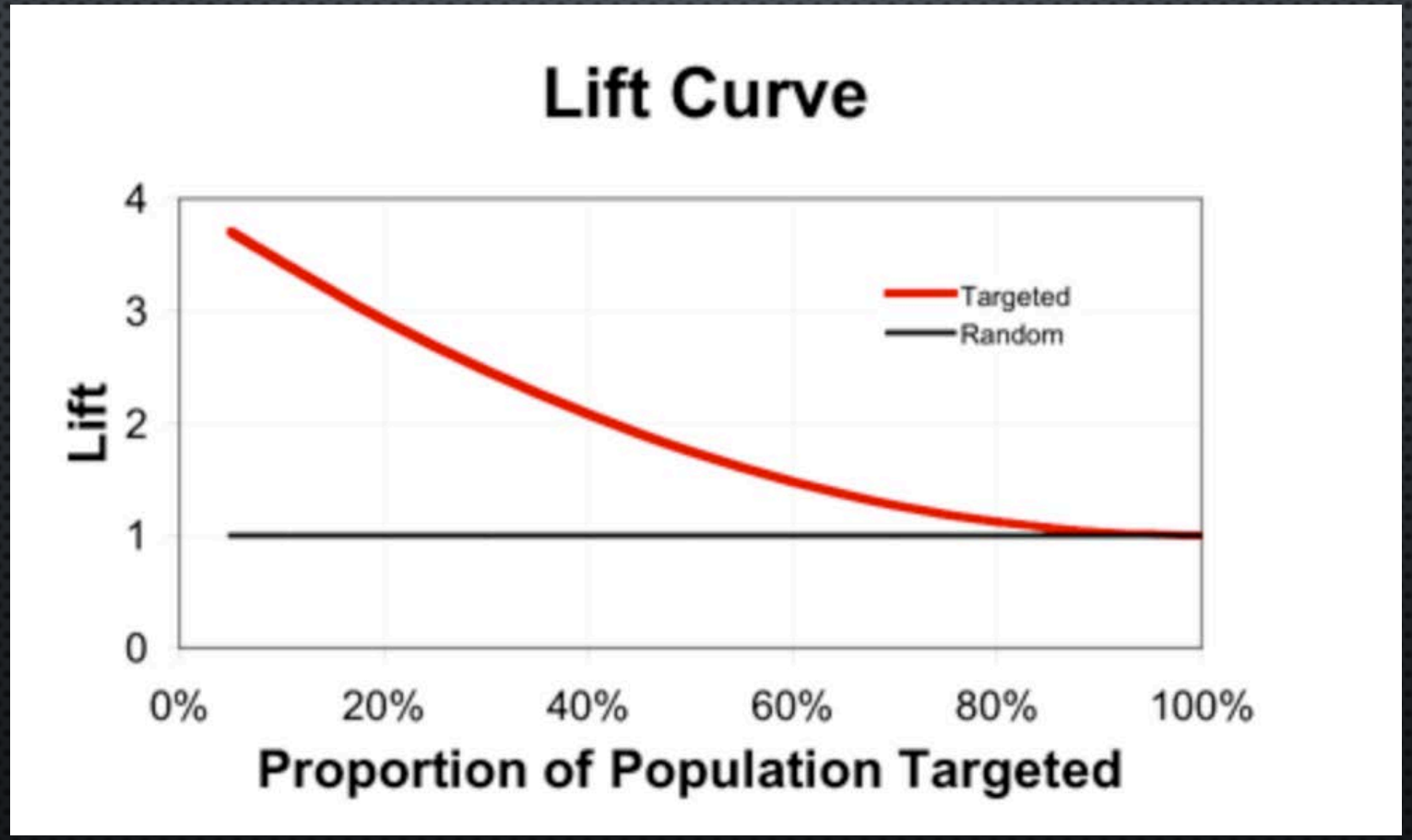
1 CALLCENTRUM POJISTOVNA

GLM v CC

- LOGISTICKA REGRESE
- TARGET = SJEDNANO
- CENA, VEK,

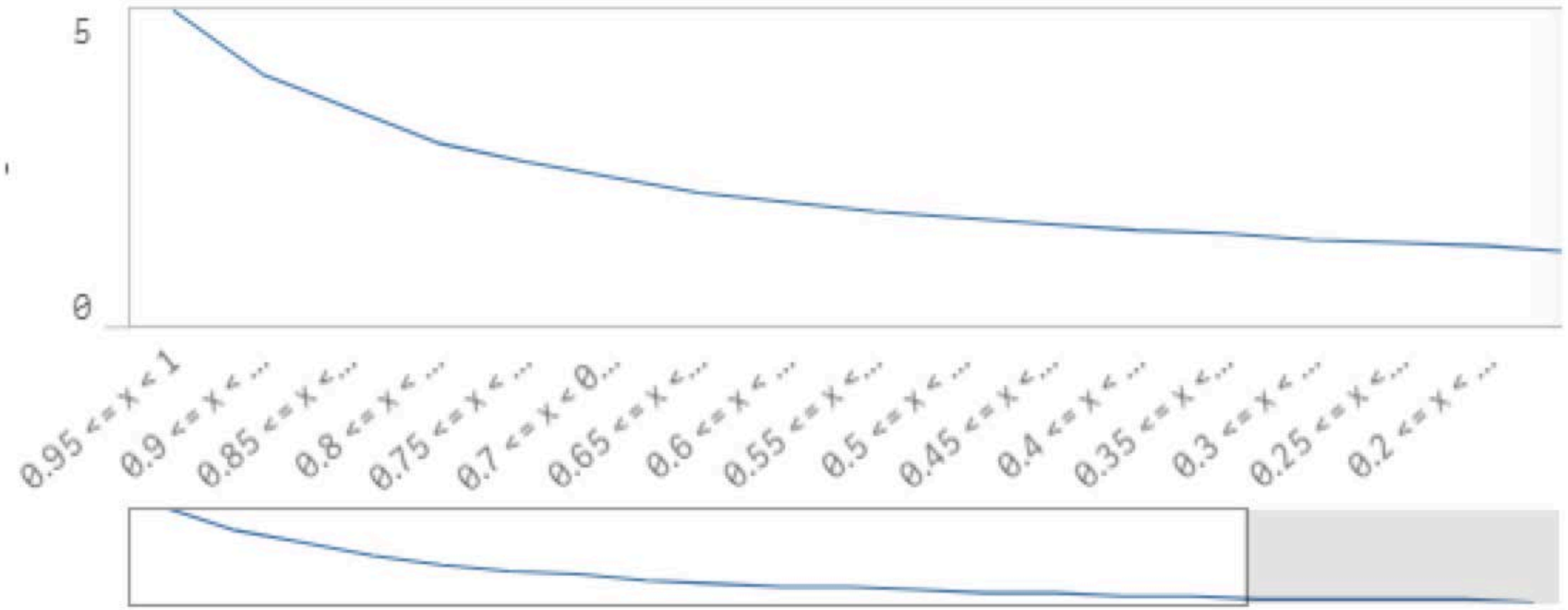

Most important variables
UpravabalickuNe
Upravabalickuprazdne
Najezddo5000
FQNe
VyuzitiCKPNe
Produkt1.POV
Stari.30.40
Stari.40.50
Produkt1Havarijnipojisteni1prazdne
Stari.do.30
Objem.1350
Produkt1Pojisteniskel1prazdne
Najezd50017500
Najezdprazdne
Produkt1SpoluucastHAV5
Stari.50.65
Cena.na.zacatku
Produkt1Havarijnipojisteni150000Kh
Bezeskodny.6m
KlientTypklientaObhanHRSR
−0.5
0
0.6
0.5
0.4
0.3
0.2
0.1

LIFT KRIVKA


Lift Curve
Lift
4
3
2
1
0
Targeted
Random
0%
20%
40%
60%
80%
100%
Proportion of Population Targeted


Lift curve - probability of new contract
3
2
1
0.95 <= x < 1
0.9 <= x < ...
0.85 <= x <...
0.8 <= x < ...
0.75 <= x < ...
0.7 <= x < 0...
0.65 <= x <...
0.6 <= x < ...
0.55 <= x <...
0.5 <= x < ...
0.45 <= x <...
0.4 <= x < ...
0.35 <= x <...
0.3 <= x < ...
0.25 <= x <...
0.2 <= x < ...
0.15 <= x < ...
0.1 <= x < 0...
0.05 <= x <...
0 <= x < 0.05
NULL

Lift curve - probability of new contract * yearly payment



2 MANA

SITUACE

- EXPLORACE
- CASHFLOW
- OPTIMALIZACE MKT SPENDU

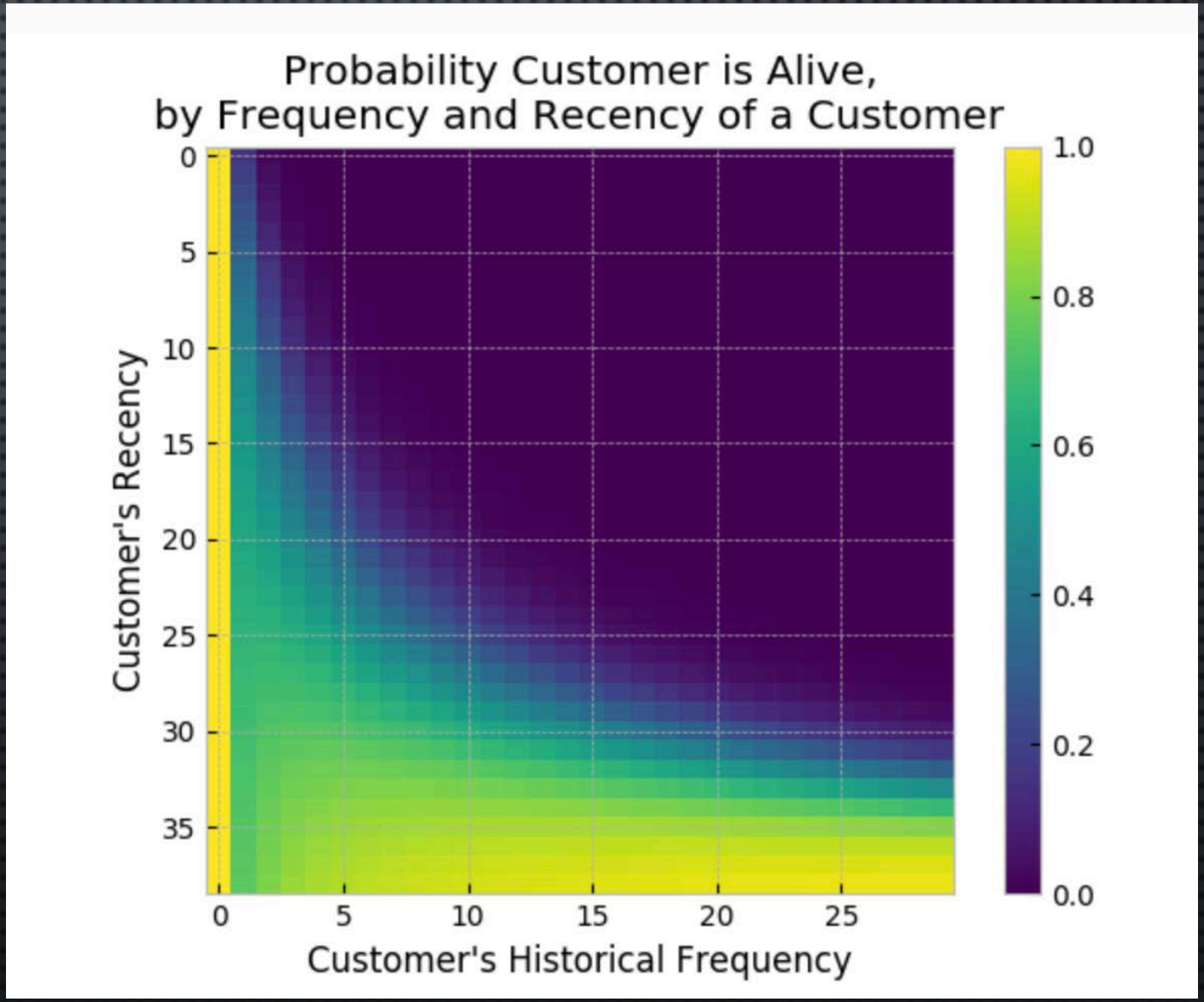
CLV/CHURN RATES – PROBABILISTIC APP

- number of transactions - Poisson process λ.
- time between transaction - exponential λ,

The basic Pareto/NBD model

- “coin” customer churns - pareto
- “dice” orders - negative binomial

CHURN RATES


Probability Customer is Alive,
by Frequency and Recency of a Customer
Customer's Recency
0
5
10
15
20
25
30
35
Customer's Historical Frequency
0
5
10
15
20
25
1.0
0.8
0.6
0.4
0.2
0.0

Datasentics CLV

- Spotreba
- Prechodove matice spotreby
- Mrtvy neni mrtvy jen malo spotrebuje

DATASENTICS CLV

	do 1	do 2	do 4	do <10	do >10
z 1	94%	1%	1.1%	1.4%	2.5%
z 2	1.5%	95%	1.5%	0.9%	1.1%
z 4	1.9%	6.5%	79.3%	8.6%	3.7%
z <10	1.1%	3.1%	11.8%	70.4%	13.7%
z >10	3.6%	2.5%	4.3%	19.4%	70.3%

DATASENTICS CLV

- NAMODELOVAT X NOVYCH ZAKAZNIKU
- STREDNI HODNOTA = CLV

SROVNANI OBOU PRISTUPU

- POTREBA HODNE ZAZNAMU PRO PROBABILISTIC APP
- DEFINICE AKTIVNIHO/MRTVEHO ZAKAZNIKA – 3 NAKUPY
- PROBLEM PREDIKCE AKTUALNI SPOTREBY

OTAZKY…?

DEKUJI ZA POZORNOST