Ktoré časti Dátovej analytiky prežijú (prvý) útok robotov?

@FilipVitek, Director Data Science

Kto do pekla je Filip Vítek ?

Filip Vítek

16 rokov sa venujem stratégii, analýze dát a rozvoju CRM systémov a BigData prístupov

Vybudoval som **analytické útvary** pre **6 rôznych odvetví,** teraz pracujem pre**Teamviewer (IT) :**

MORE INFO

V Data mining a BigData oblasti ma skôr budete poznať cez

300+ expert blogov

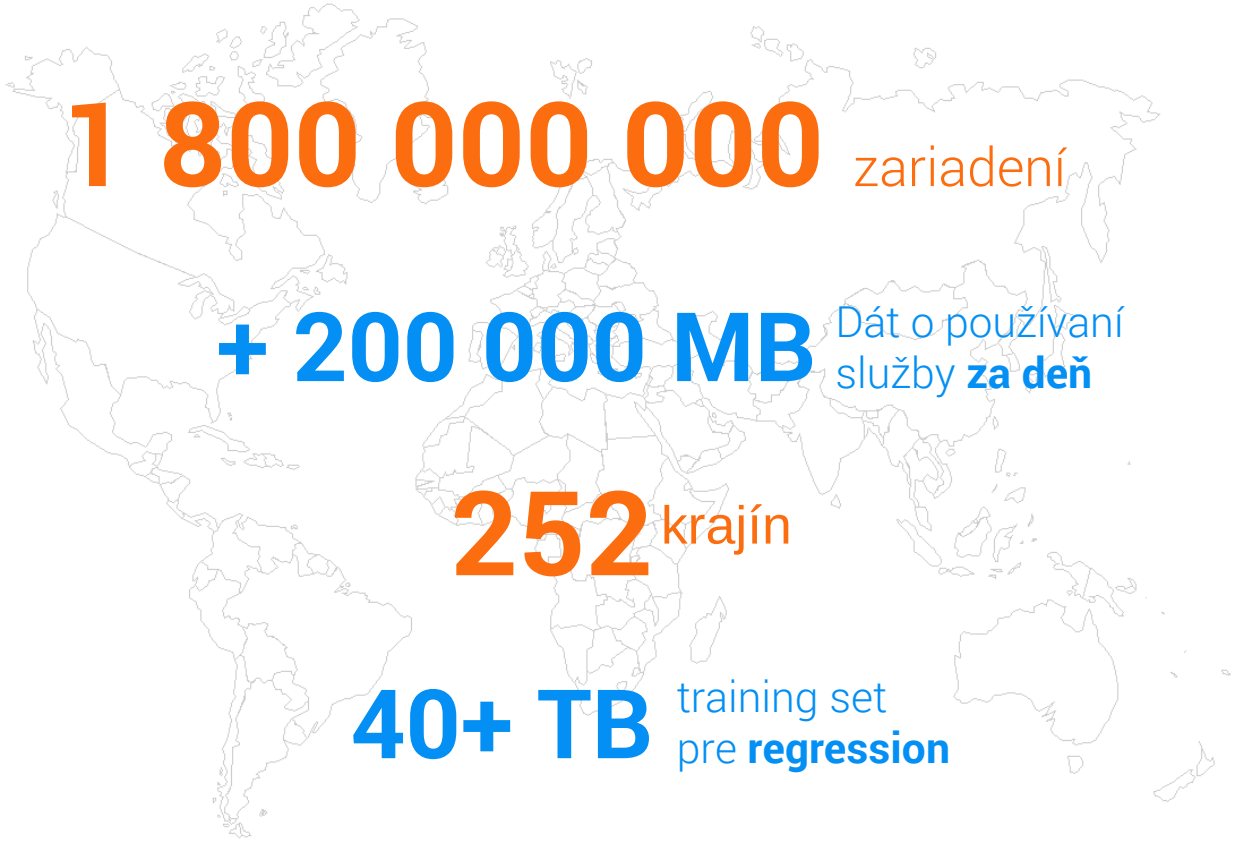
Ak nebudeme mať šancu ísť do hĺbky, budem sa dávať odkazy na blogy na dané témy.

TREND
Týždenník o ekonomike a podnikaní

Mocné dáta
www·MocneData·sk

TeamViewer

TeamViewer | Remote Management | TeamViewer IoT | TeamViewer Pilot | Blizz



1 800 000 000 zariadení

+ 200 000 MB Dát o používaní služby **za deň**

252 krajín

40+ TB training set pre **regression**

TECH STACK

Flume | kafka | amazon REDSHIFT | Microsoft SQL Server | Amazon S3 | AVRO | Parquet | Scala | python | tableau

Upozornenie:

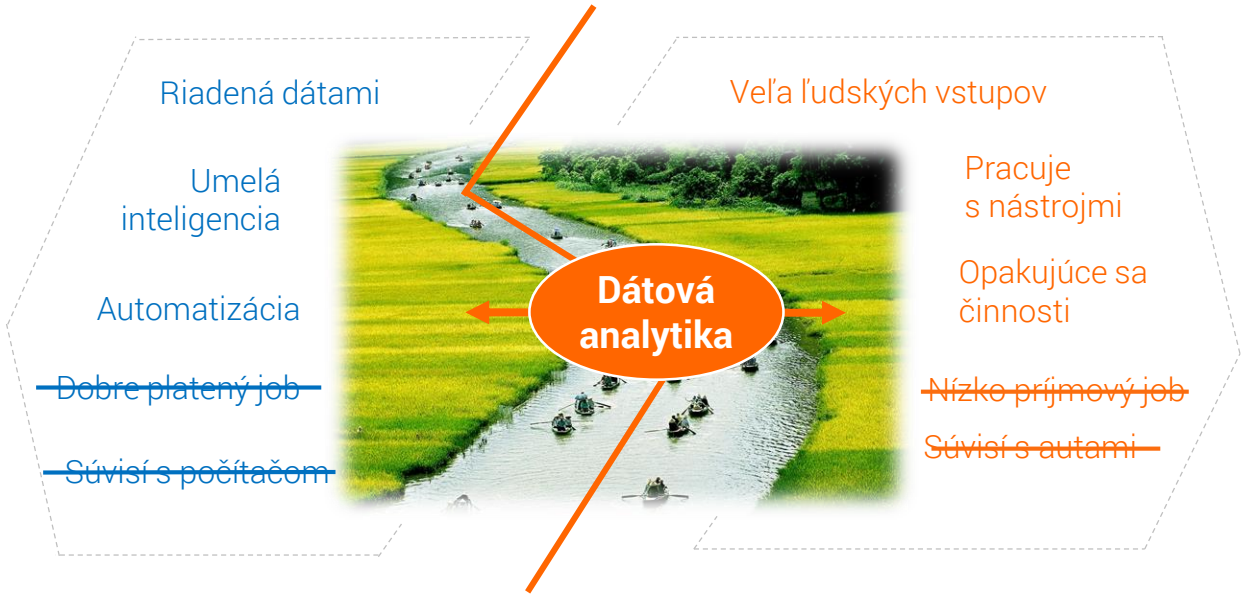
Cieľom tejto prezentácie **NIE JE VÁS VYSTRAŠIŤ.**

- - - Aj keď, niektoré veci, ktoré za chvíľu poviem,
sú **NAOZAJ STRAŠIDELNÉ**. - - -

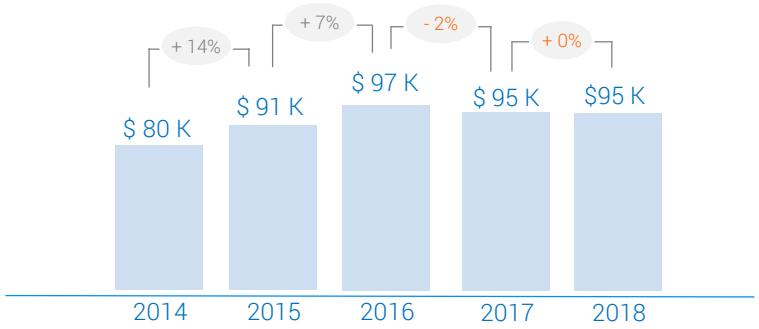
Ideálne by bolo, keby ste sa
ZARIADILI PODĽA NICH.

Môžete sa rovnako rozhodnúť ich aj ignorovať, ale iba na vlastné riziko.

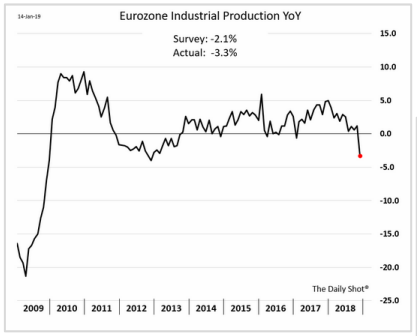
Prečo by sa vlastne dátoví analytici vôbec mali báť?



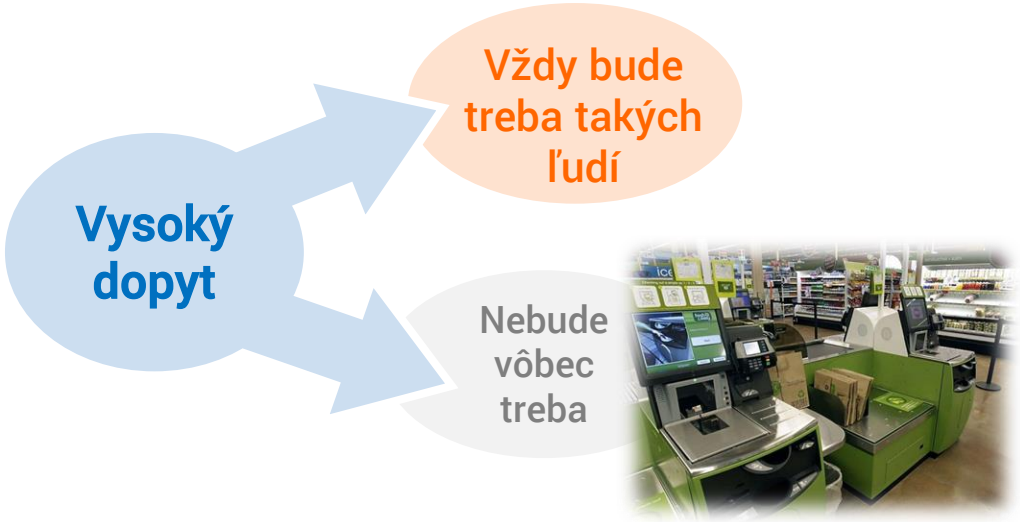
Vývoj ročného platu[1] Data Scientistu v US



Ekonomická kríza už prichádza



Source: Wall Street Journal



1 Ročný plat, dáta z USA trhu od

Vždy sme to vedeli “ nandať “ strojom. Doslova!

1800’s

1900’s

2000’s

21 útokov [1] *na Waymo (Google) v meste Chandler, Arizona*

1 https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html

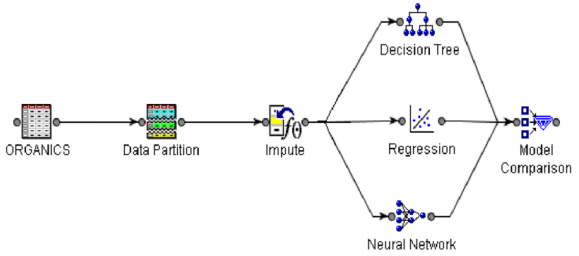
Ako sa teda tomu postaviť čelom?

Keď sa bojíme, že nás stroje porazia, prečo im to uľahčujeme tým, že znižujeme pre nich latku …

A Ignorancia externých dát

- Silná orientácia len na interné dáta
- Mylne sa domnievame, že je to veľa úsilia
  [V skutočnosti = 7 riadkov kódu na prehľadanie webstránok 900 tis klientov]
- Roboti nebudú leniví, im to príde úplne normálne

B Pandémia “Default option”



- Skontrolovali ste niekedy po MS Excel [TM] či počíta správne? =1*(0.5-0.4-0.1)
- Python je nový MS Excel [TM]

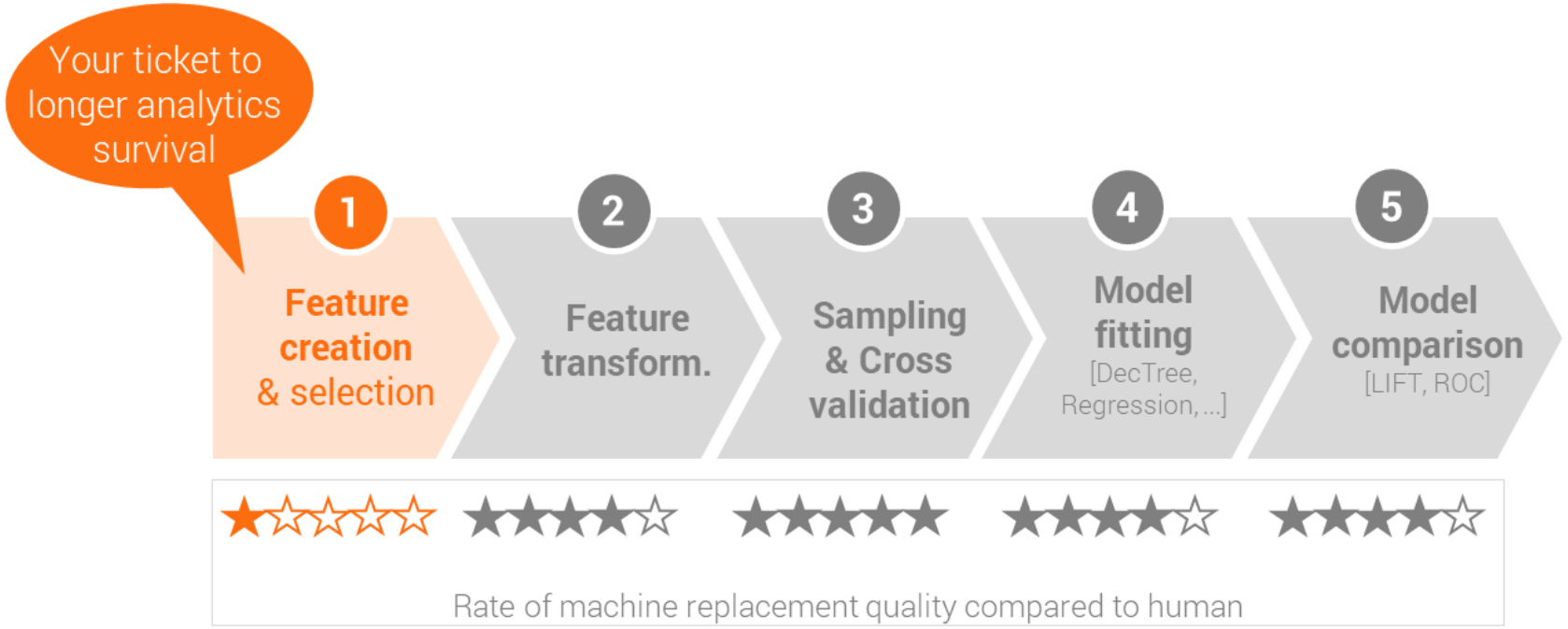
C Zlepšujeme sa v nesprávnych veciach

- Väčšina ML/AI kurzov je …
- Algoritmy sú komoditami
  [ako by človek mohol poraziť XG Boost ?]

SOFTECON

Čo teda v dátove analytike zostane pre ľudí?

Skúste si spomenúť na ostatný Machine learning model, ktorý ste robili. V ktorých z nasledovných krokov **ste sa spoľahli na nejaký predprogramovaný package/knižnicu?** (e.g. SciKit Learn, ...)



MORE INFO

"Akú alternatívu máme, ako ľudstvo, oproti AI?"
http://mocnedata.sk/en/what-alternative-do-we-have-to-ai/

SOFTECON

Byť všetkým znamená byť …

Muži **100m**

Muži **Desaťboj**

… je nutné vedieť, akým analytikom chcete NAOZAJ BYŤ

Analytika ako **Jeden kontinent**

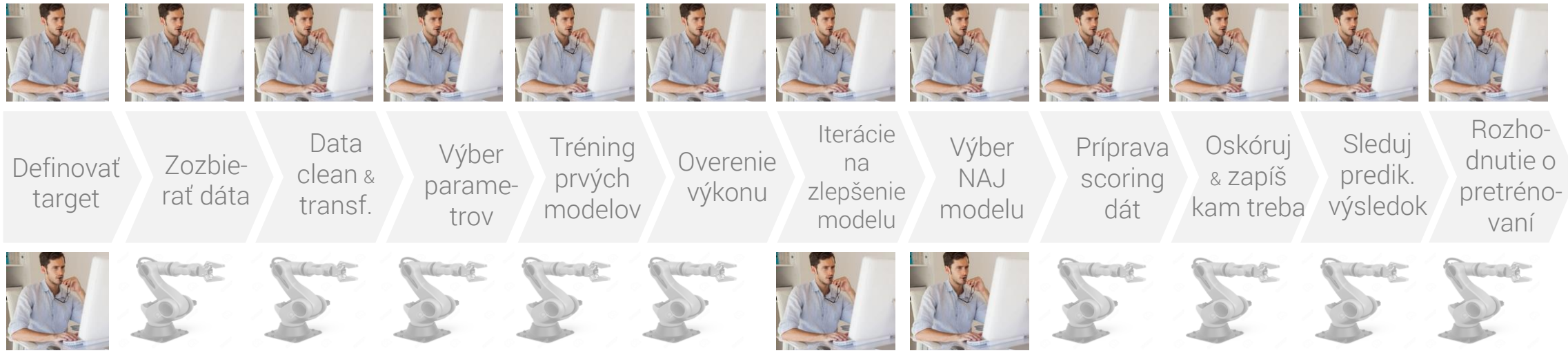
Analytika ako **Súostrovie**

V - I - B - A

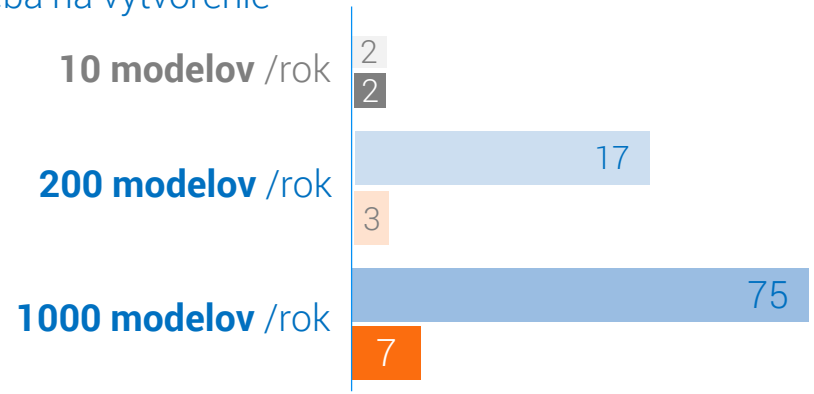
… the MBTI v oblasti analytiky. Over si aký typ Dátového analytika si TY sám:

http://mocnedata.sk/en/**VIBA-type-of-analyst/**

NEROBME, čo by sme NEMALI ROBIŤ



Koľko zamestnancov treba na vytvorenie

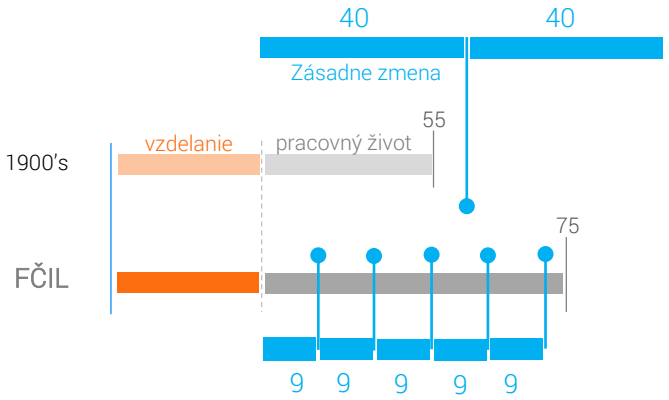


My, v TeamViewer,
sme prinútení rozsahom.
Ale väčšina teamov nie je ..

SOFTECON

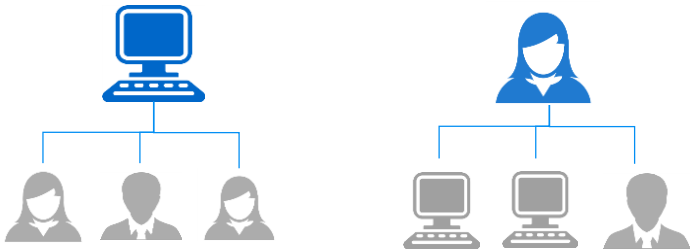
Zdokonalenie sa. Ako by sme sa my, ľudia, mali pripraviť?

Univerzity. Naozaj?



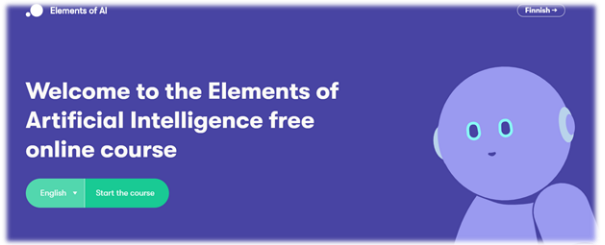
- Každý hovorí, že nam chýba "viac absolventov Data Science"
- Väčšina úspešných uchádzačov NEMÁ AI vzdelanie z Univerzity
  (keď už sme pri tom, ani ja nie)
- Ešte intenzívnejší boom Udemy, Coursera, Udacity, FutureLearn
- Vzniká konceptu "AI citizen" , ak môžete, využite toho

Manažérsky aspekt



- Každý chce byť data scientistom.
  … ale v konečnom dôsledku ich nebudeme potrebovať toľkých …
- Data Scientisti dnes reportujú NE-analytickým nadriadeným
  [syndróm Osamelého analytika + 4. druhy šéfa]
- Ak už dnes ovládate ML/DL, nesnažte sa zdokonaľovať v technických veciach. Trénujte soft skills, zoberte (aj keby horšie platený) job Teamlídra …

Fínske 1%



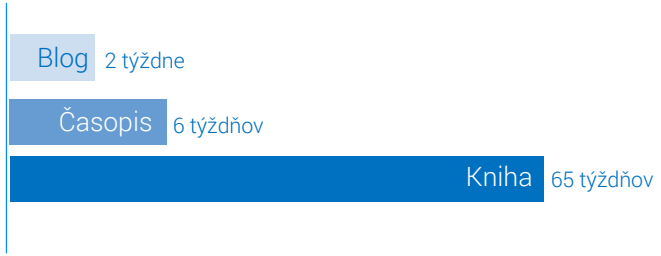
- Spomínate na kurzy IT gramotnosti?
  … budeme potrebovať niečo také…
- Fínsko vybralo náhodne 1% populácie, ktoré vyškolia na Umelú inteligenciu [zubár]
- Ak prepukne nezamestnanosť, dopadne to na štát. Takže ten má závažný záujem podporovať takúto AI do-kvalifikáciu
- Ak by to nakúpila vo veľkom objeme, je to super lacné (< 10 EUR per person)

MORE INFO *"Osamelosť Dátového analytika"*
http://mocnedata.sk/osamelost-datoveho-analytika/

MORE INFO *"4 druhy šéfov, čo nerozumejú analytike"*
http://mocnedata.sk/4-druhy-sefov-co-nerozumeju-analytike/

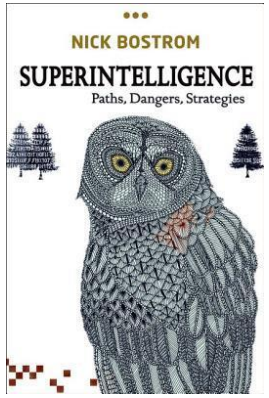
Zdokonalenie sa. Čo & Kde si prečítať?

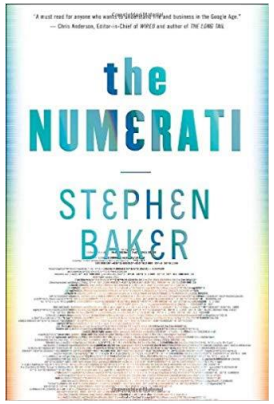
Koľko to trvá, kým NIEČO NOVÉ sa objaví v …



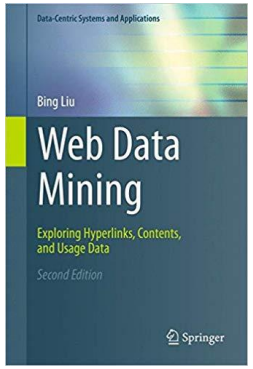
https://blog.feedspot.com/ai_blogs/

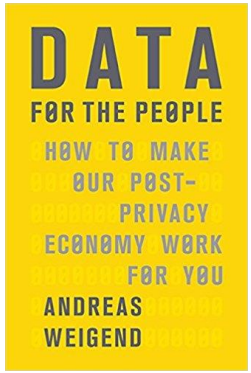
Mocné dáta
www.MocneData.sk

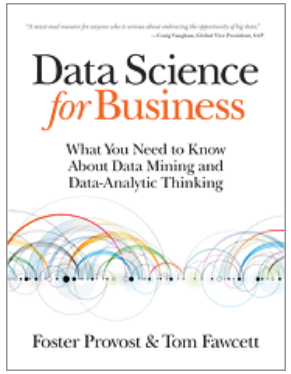




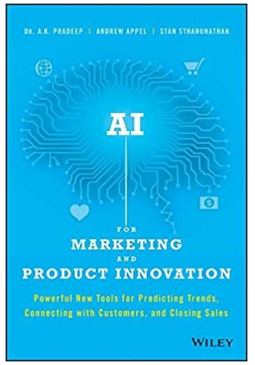
… pre expertov

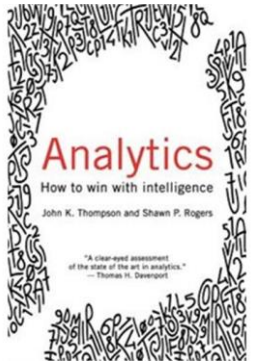






… pre manažérov



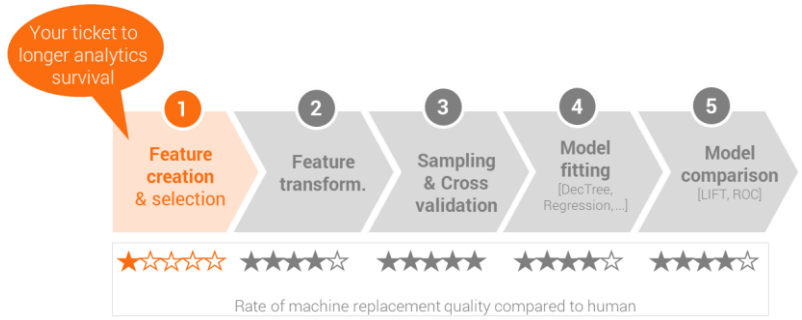


MORE INFO

4 kvalitné knihy, ktoré by ste si mali prečítať
http://mocnedata.sk/4-kvalitne-knihy-co-si-precitat-o-datovej-analytike/

SOFTECON

Experti sa zhodujú: "**Príprava premenných bude pravdepodobne jedna z posledných pevností človeka v analytike.**"

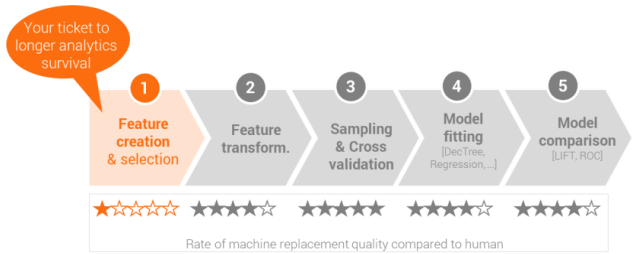


2 300 000 anotovaných vstupov

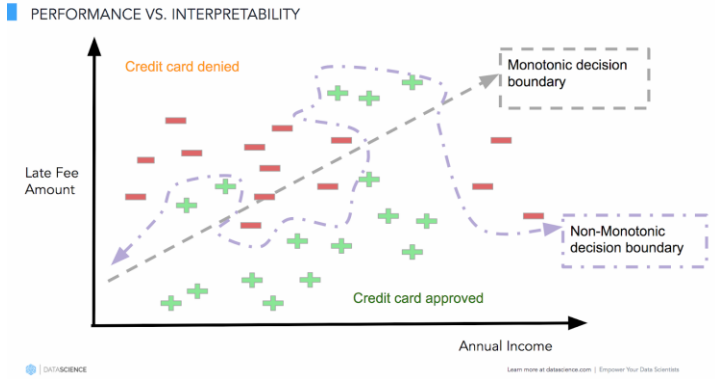
vs. **260** anotovaných vstupov

Aké joby **vyjdú víťazne z 1. útoku AI** na dátovú analytiku?

Feature inžinier
(alebo Data Scientist so silným Feature skillsami)



Model auditor/ Manažér vysvetliteľnosti



Front-End dizajnér pre AI produkty

Zamestnanec trhoviska algoritmov

SOFTECON

Zhrnutie toho najdôležitejšieho

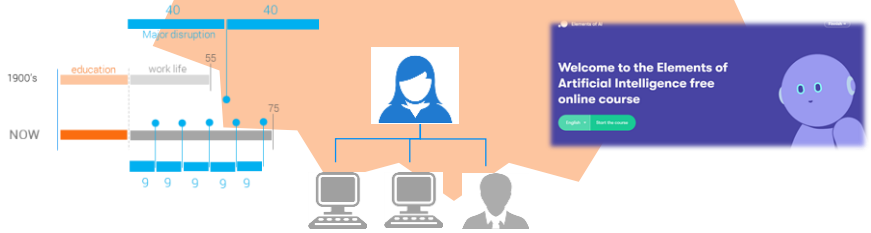
nerobiť
ČO BY SME NEMALI

neznižovať
LATKU

posledné
PEVNOSTI ĽUDÍ

zvážiť
DESAŤBOJ

prejsť
ZDOKONALENÍM



SOFTECON

SOFTECON

Ďakujem za Vašu pozornosť a som pripravený na

VAŠE OTÁZKY

Mocné dáta
www·MocneData·sk

Pridaj sa do komunity

http://mocnedata.sk/zostanme-v-kontakte/

Pokojne sa ozvi ...

Mgr. Filip Vítek
Data Science Director
TeamViewer, Berlin

+ 421 911 072 231
info@mocnedata.sk
https://sk.linkedin.com/in/vitekfilip
@FilipVitek

SOFTECON

BACK UP SLIDES

Čo očakávam od… COOL prípravy premenných

A **Rozšírila zoznam “surovín“**

(= viac informácií pre model)

B **Sorting out the hopeless cases**

C **Zúženie počtu hodnôt pre rýchlejšie**

trénovanie a správu modelu

D **Ohlási, že mi chýba niečo zásadné**

MORE INFO

“What (should) good Feature engineering look like?”

http://mocnedata.sk/en/good-feature-engineering/

SciKit Learn … Naše ICčko v Machine Learningu (?!)

✓ Supervised Learning

(GLM, LinDiscAnal, KernelRidge, SVM, StochGradDescent, NearestN, NaiveBayes, DecisionTrees, Feature selection, Ensemble methods, MulticlassAlgor, Isotonic Regress., Prob. Calibration, NeuralNetworks, ...)

✓ Unsupervised Learning

(GaussianMixture, ManifoldLearning, Clustering, Biclustering, MatrixFactorization, Covariance estimation, OutlierDetect, DensityEstimate, NeuralNet, ...)

✓ Model selection

(CrossValid, HyperParemeters, Model evaluation, Model persistence, Validation curves.)

✓ Dataset Transformations

(Pipelines, Feature extraction, Preprocessing, Impute, DimensionReduction, Projections, KernelApprox, PairWiseMetrics, TargetTransform)

✓ Datasets, Loading & Scaling

(Toy/Real datasets, Generated datasets, Loading, Incremental learning, PredicitonThrouput, Parallelism)

Feature selection tools

- Low Variance Removal
- Univariate feature selection
  (Select K-Best)
- Recursive Feature Elimination
  (only backwards)
- SelectFrom Model (Tree)
- Including into Pipeline

- Principal Component Analysis
- Independent Component Analysis

Ako to vykompenzovať v Python prostredí …

Predpripravené Lamba na opakujúce sa transformácie

Featuretools

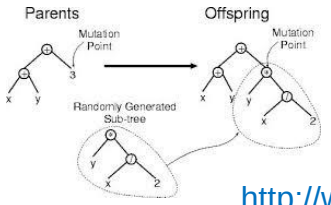
https://www.featuretools.com/

tsfresh

https://github.com/blue-yonder/tsfresh

TPOT

https://epistasislab.github.io/tpot/



Genetic Programming

http://www.philipkalinda.com/ds8.html

Použite **Deep Learning**

- Calculation costs
- Team skills
- Time-to-market
- Explain-ability

Postavte si **VLASTNÝ FEATURE** engine

1. **Calculate** variable statistics
   [see also transformations slide, ...]
2. **Generate obvious suspects**
   [aggregations, time windows, ...]
3. **Indicate** missing info categories
   [compare to dictionary, Expl. score …]
4. **Hard criteria** knock-out
   [Variance, NonNulls, distinct X, …]
5. **Binning & Categorical** decomposition
   [Forced binning if > N]
6. **Univariate** correlation & Log -P
   [Simple tree is enough]
7. **Bivariate** relations
   [Cut off for Categorical dummies by Support]
8. **Decision on ranking** of parameters
   [Simple, Stage based, …]

MORE INFO

"Unconventional methods of Feature engineering"
http://mocnedata.sk/en/feature-engineering-berlin/

Real examples of Unconventional Feature Generation

Unconventional approach to generating features

How old are you, **Bernard?**

- Nothing like "National ID" for German insurance companies = they have no clue about age of customer
- Important for setting proper communication (web vs. call vs. paper letter)

- First name + Region predicting 92% accurately the decade when the customer was born

[cut/off point for approx. 25% Individuals]

Fee increase tolerance

- Fee increase sensitivity for retail bank
- In search for metric that would tell: How "lazy" user is?

- Limited space, banking feels very un-emotional
- Lowest amount ever withdrawn from the ATM

[worked surprisingly well, due to large coverage]

Detecting **commercial** customer

- Quite a few small companies without license
- Too small to detect via IP address range
- Using standard desktop OS versions

- Pattern of use strong within working hours, weak outside

[nightmare of time zones from UTC]

Zodiac, are you kidding me?

- Probability to have car accident
- As Joker card for model
- Strong objection from Data Scientists: "*This is not serious work, we protest.*"
- Ended up as the Second strongest parameter in model.
- Later confirmed in 4 other countries in same issue

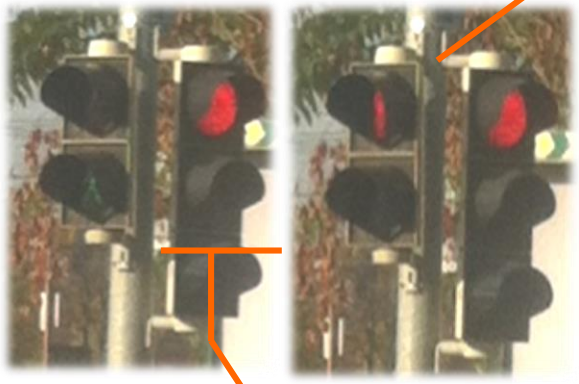
[I have a hypothesis why it works]

MORE INFO

Find more of interesting feature examples here:
http://mocnedata.sk/en/what-to-read-in-english-here/

Data underdogs … and their impact

Who will win the car race to nearest lights?



Has originally **other informational role**

Indicates **client behavior**
[or its change]

- ☐ Data fields that are "just identifiers"
- ☐ Contact & Transactional data
- ☐ No obvious relations as champion challengers (Joker cards)
- ☐ Unusual aspect of usage
- ☐ "Ryanair-like" data test

Social impact on other clients in portoflio

Jane.Angry@teamviewer.com
Martin.Neutral@teamviewer.com

1

2

Tone of voice

John.Warton @hotmail.com

Johny_geek @hotmail.com

3

Bank preference

(Online bank vs. **Postal bank)**

4

Relationship proxy

(133333333 /xxxx
133353333 /xxxx)

MORE INFO

Ospravedlňte sa svojej databáze
https://blog.etrend.sk/filip-vitek/data-underdogs.html