When Will AI Kill Data Scientists?

Your Ticket to survival in Data science

@FilipVitek, Director Data Science

Who the hell is Filip Vitek ?

Mr. Filip Vítek

16 years building business strategies, Data Science, CRM systems development and BigData projects

Built **analytical units** in **6 different industries,** now working for **Teamviewer (IT) :**

MORE INFO

Data mining is my hobby and passion, I wrote more than

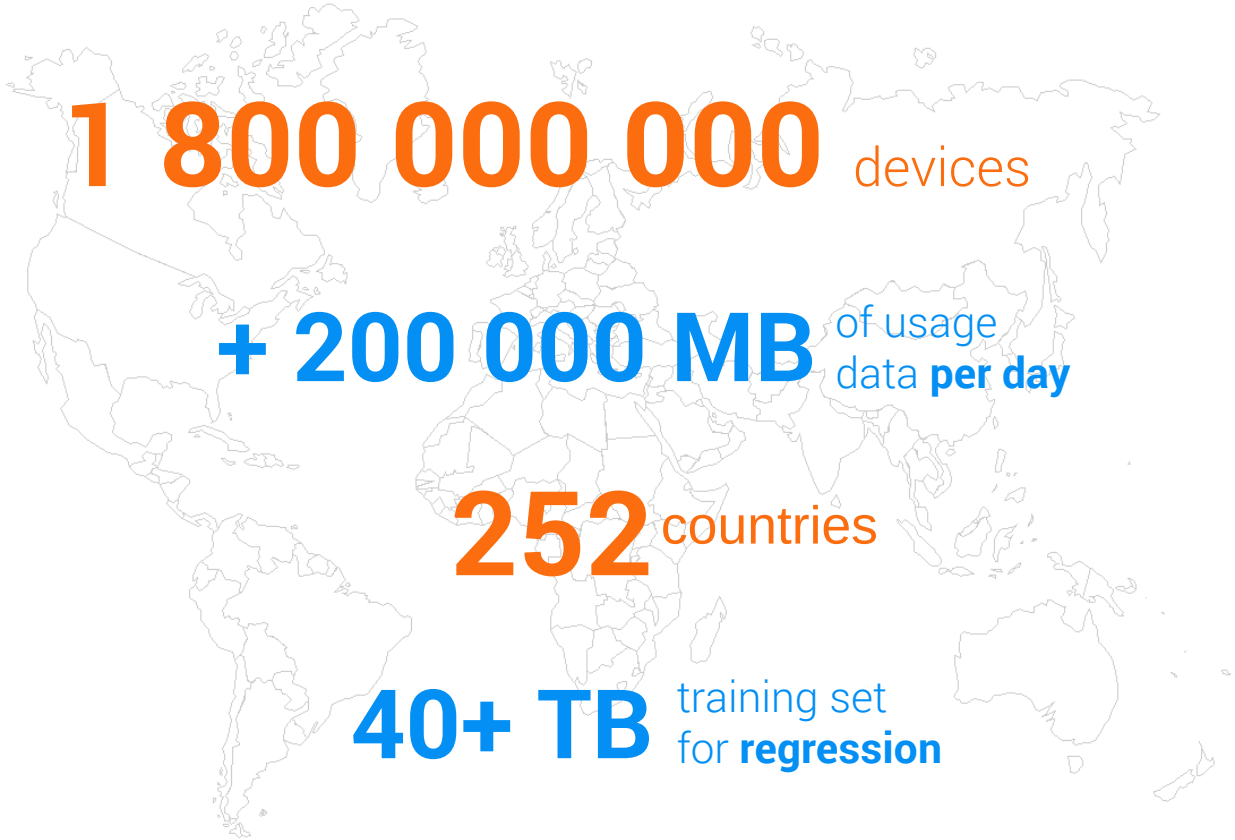
300+ expert blogs

If no time to go into details, I will leave you an link to read further on given topic.

Mighty data
www.TheMightyData.com

TeamViewer

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



1 800 000 000 devices

+ 200 000 MB of usage data **per day**

252 countries

40+ TB training set for **regression**

TECH STACK

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

Disclaimer:

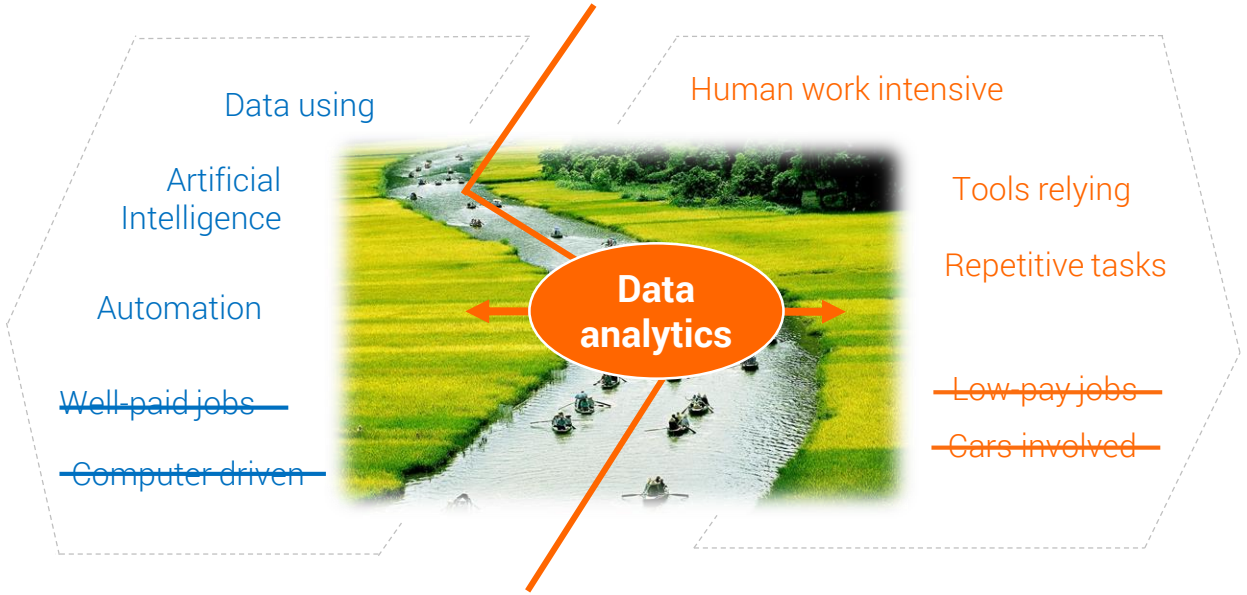
The goal of this presentation is **NOT TO SCARE YOU.**

- - - Though, some things I am just about to say
REALLY ARE scary. - - -

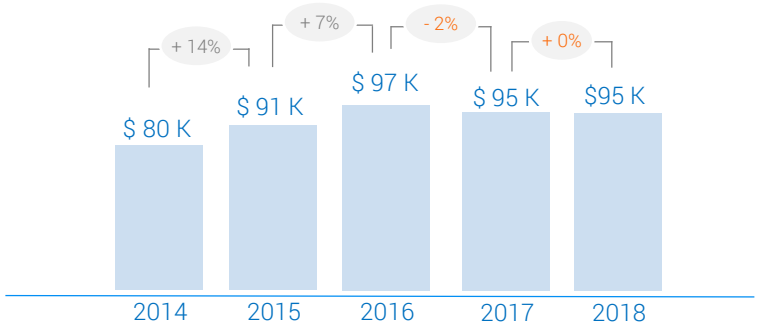
Ideally, I would like **YOU TO ACT** ON THEM.

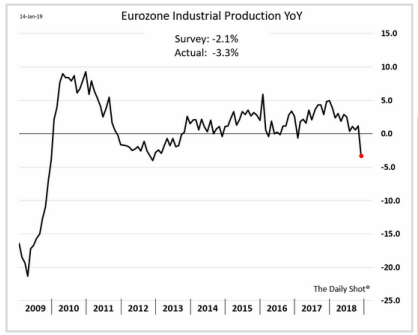
But feel free to ignore them, on your own risk.

Why should Data Scientists be in Danger at all?



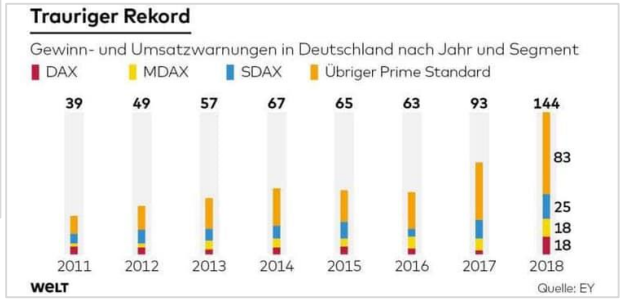
DS Salary development[1]





Source: Wall Street Journal

Economy crises coming



1 Data from US market by © 2018 Burtch Works LLC

We always wanted to Beat The Machines. Literally!

1800's

1900's

2000's

21 attacks [1] *at Waymo (Google) in Chandler, Arizona*

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

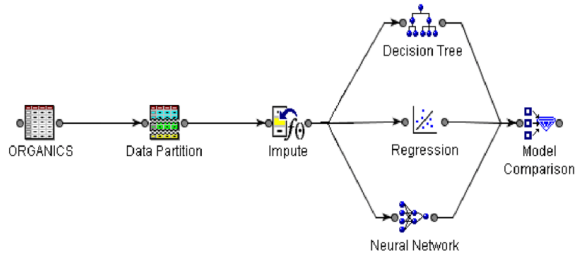
How should we Face it properly?

We fear machines get better. Let's not help them to do so with keeping the bar so low …

A Ignoring external data

- Strong internal locus of data
- We feel it is too much of effort
  [In reality = 7 lines of code to get all your clients webpages]
- Robots will not be lazy to do it, it is natural for them

B "Default option" pandemics



- Did you ever check if MS Excel [TM] calculates properly? =1*(0.5-0.4-0.1)
- Python is the new MS Excel [TM]

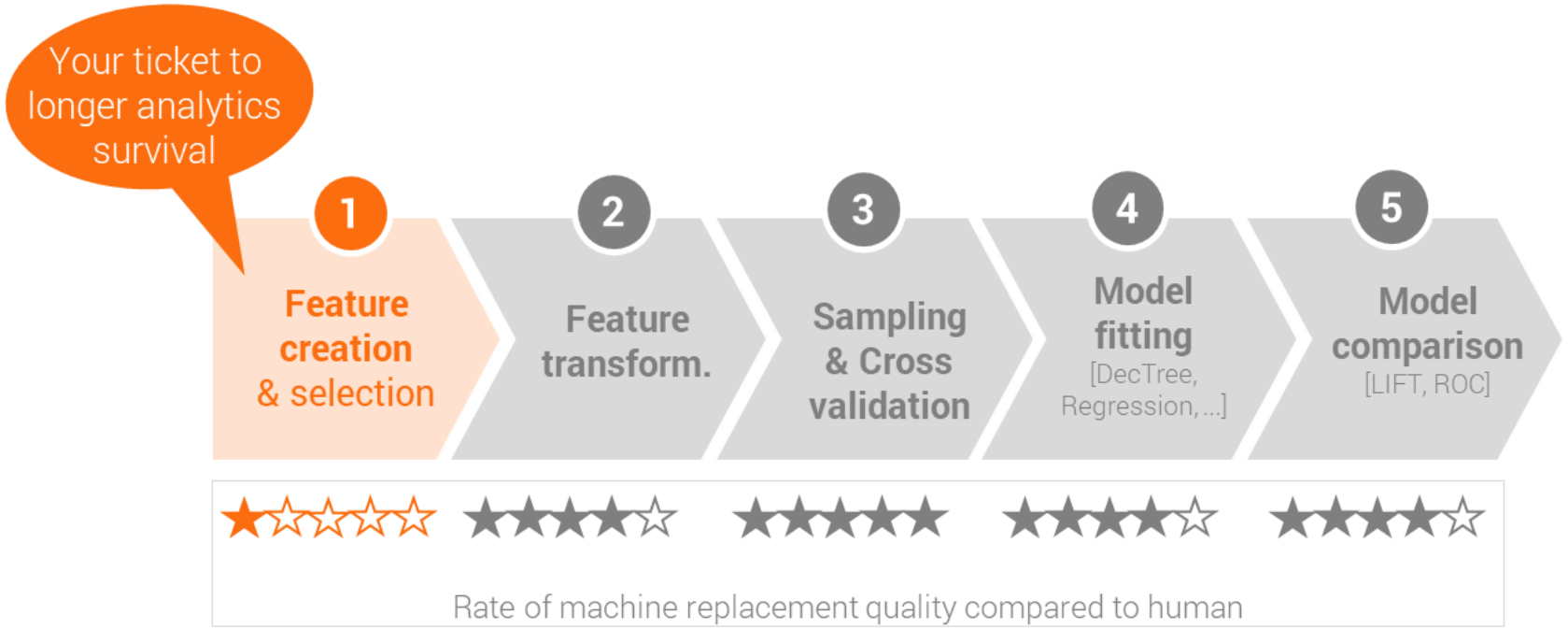
C Getting better in wrong things

- Self-study courses of ML/AI
- Algorithms are commodities
  [think XG Boost and how can human do better]

DATA festival 2019

What is left for Humans to take part ?

Think of last time you were designing the Machine learning model. In which of the following steps **did you rely on some of the ready-made packages?** (e.g. SciKit Learn, ...)



MORE INFO

"What alternative to AI do we, humans, have?"
http://mocnedata.sk/en/what-alternative-do-we-have-to-ai/

Being everything means being …

MEN's 100m

MEN's Decathlon

… need to be true to your self on WHO YOU REALLY ARE

Analytics as **Single continent**

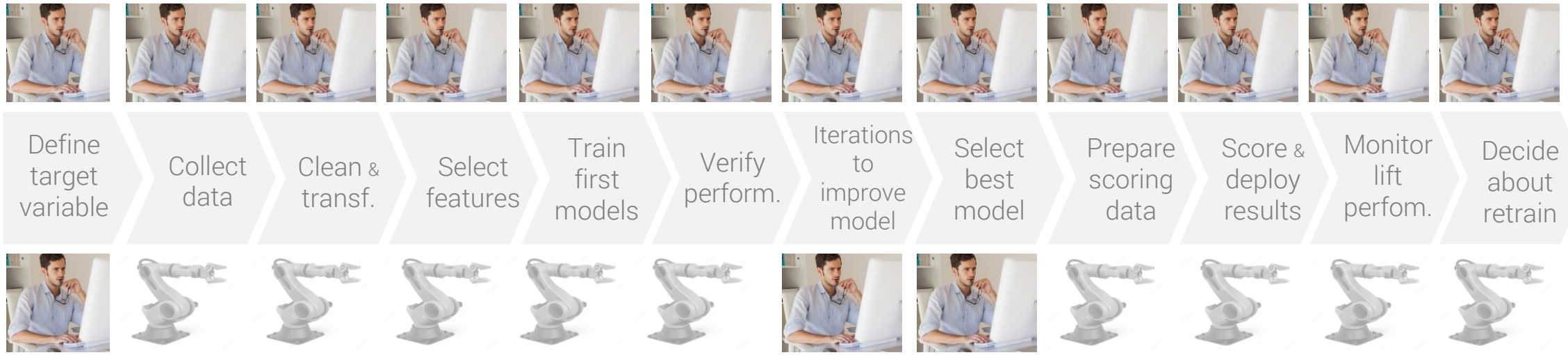
Analytics **Archipelago**

V - I - B - A

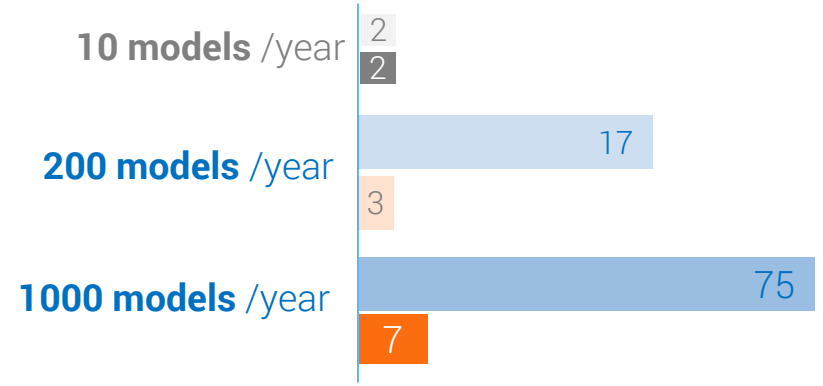
… the MBTI of the analytics. Find out which type of Data Scientist YOU are:

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

DON'T DO what you SHOULD NOT DO



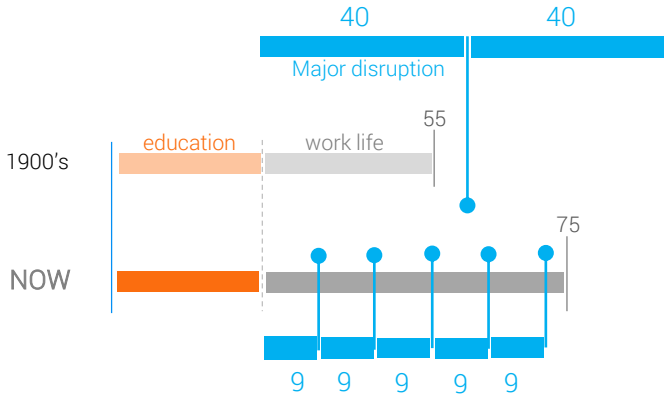
FTEs needed to create



We, in TeamViewer,
are forced by sheer volume.
But most of the teams are not ..

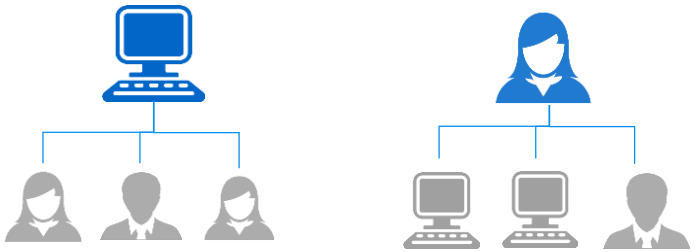
Up-skilling. How shall we, humans, get better prepared?

University. Really?



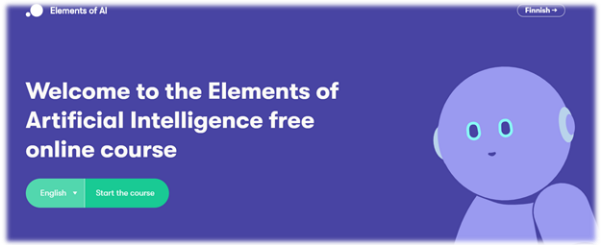
- Everybody relies on "**more graduates**"
- Most people applying have **NO AI degree** from University (btw, nor do I)
- Further boom of **Udemy**, **Coursera**, **Udacity**, **FutureLearn**
- **"AI citizen"** concept arising, take benefit of it

Managerial issue



- Everybody wants to be data scientist.
  … but we will not need that many of them later in time …
- Data Science people are reporting to Non-analytics managers [syndrome of Data analyst loneliness]
- If you already understand ML/DL, don't get more expert-ish. Train soft skills, get (even if worse paid) Team Lead job, …

Finland 1%

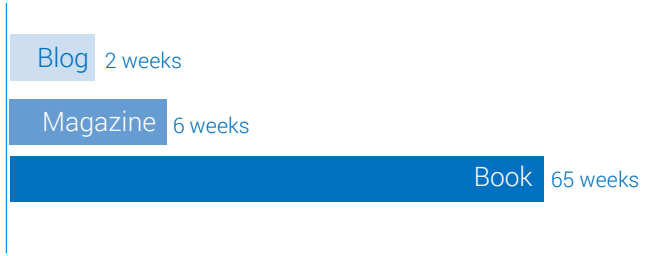


- Do you remember IT literacy courses?
  … we would need something like that …
- Finland picked 1% of population at random and train them AI fundamentals
- Government would have to face the unemployment burden, so they have vested interest to more here
- It is super cheap, if purchased in bulk
  (< 10 EUR per person)

MORE INFO *"Loneliness of Data Analyst Syndrome"*
http://mocnedata.sk/en/lonely-analyst/

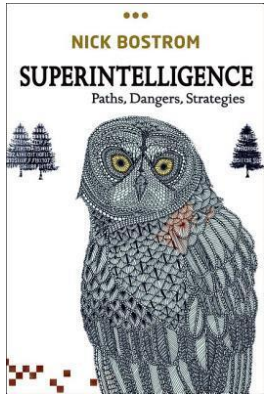
Up-skilling. How & What to read?

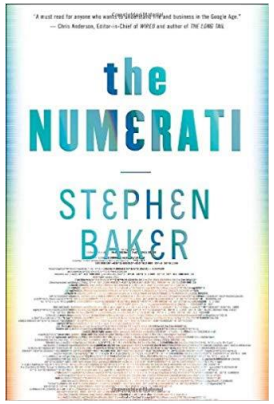
Time for SOMETHING NEW to make it into ...



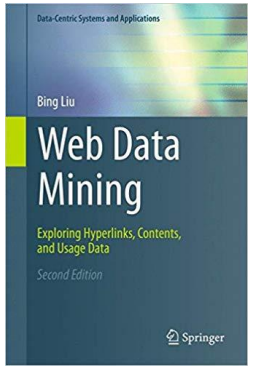
https://blog.feedspot.com/ai_blogs/

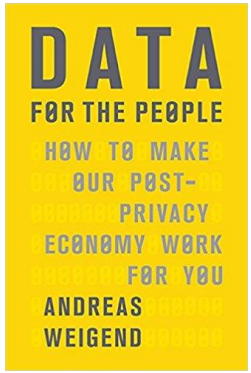
Mighty data
www.TheMightyData.com

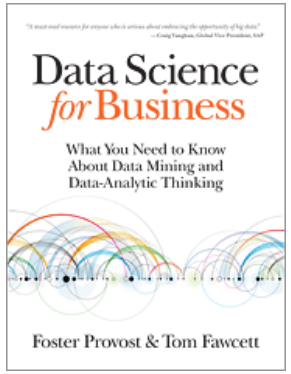




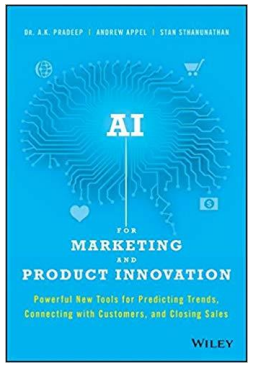
... for experts

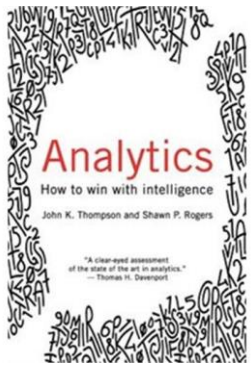






... for managers



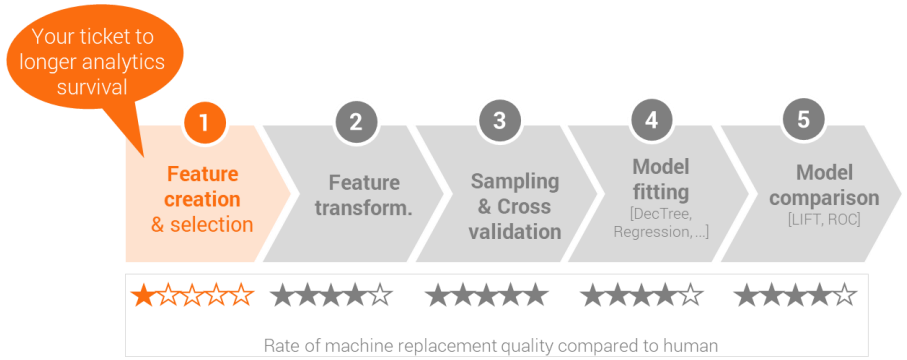


MORE INFO

4 analytics books that you should read [google translate]
http://mocnedata.sk/4-kvalitne-knihy-co-si-precitat-o-datovej-analytike/

DATA festival 2020

Experts agree: "Feature engineering will be one of The Last Human Fortress standing."



CAPTCHA

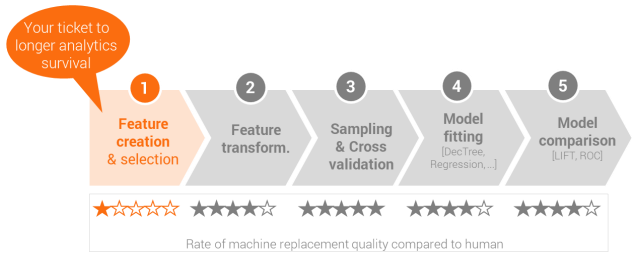
2 300 000 annotated cases

vs. **260** annotated cases

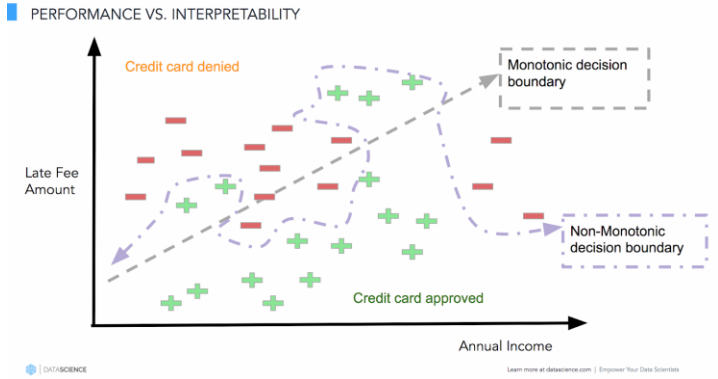
DATA festival 2019

Where to survive the 1. wave of AI assault on analytics?

Feature engineer
(or feature strong Data Scientist)



Model auditor/ Explainability Curator



Front-End Designer for AI products

Algorithm exchange platforms

How should we Face it properly?

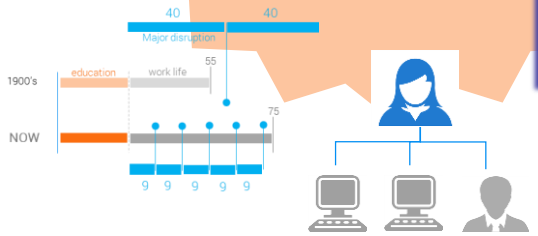
lowering
THE BAR

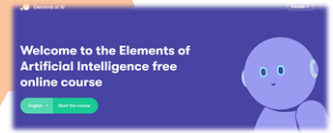
doing
WHAT WE SHOULDN'T

the last
FORTRESS OF HUMAN

rethink
DECATHLON

going through
UP SKILLING





DATA festival 2019

Thanks for Your attention and I am ready to answer

YOUR QUESTIONS

Mighty data

www.TheMightyData.com

Join the community

http://mocnedata.sk/en/lets-keep-in-touch/

Feel free to contact me ...

Mgr. Filip Vítek

Data Science Director

TeamViewer, Berlin

+ 49 1525 309 8505

filip.vitek@teamviewer.com

https://sk.linkedin.com/in/vitekfilip

@FilipVitek

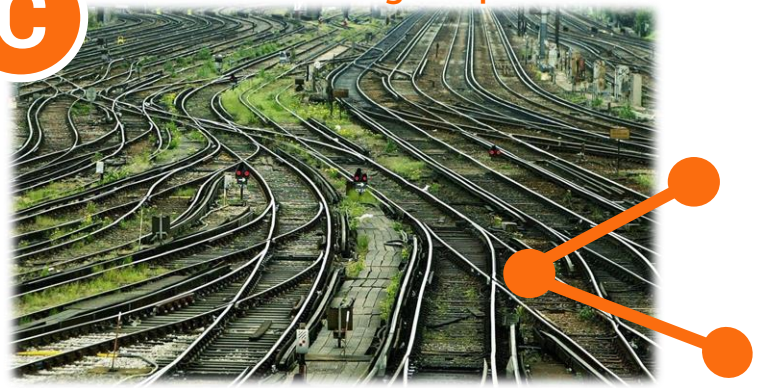
BACK UP SLIDES

What do I expect … of COOL Feature Engineering

A Extending set of ingredients
(= insights available)

B Sorting out the hopeless cases

C Prune to have more efficient
model training & operation



D Signal, if I missed anything relevant

MORE INFO

"What (should) good Feature engineering look like?"
http://mocnedata.sk/en/good-feature-engineering/

Python is … the Microsoft Excel™ of our era

It Became standard. There are options, but why to bother to even try. (Think Lotus 1-2-3)

Everybody claims knowledge of it, but knowledge of most people is very shallow.
(Frustrating to test Data Science hires for basic Python … and see them struggle with RegEx)

Tool is really powerful, but you need to possess certain skills beyond elementary use.

To rely hone the power of it, you need to know more than the default options/libraries.
(… which most people don't)

=1*(0.5-0.4-0.1)

VLOOKUP blinds
(back search, MIX, Case Sensitive)

Z-score glitch

"Don't CHALLENGE or REVIEW, just CONSUME."
(Have you ever checked what Excel calculates? / Have you challenged any Sci-kit learn routines?)

SciKit Learn … Our U-bahn of the Machine Learning (?!)

✓ 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

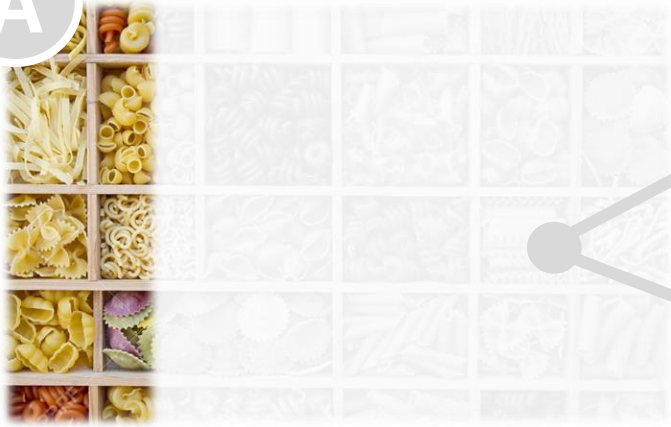
- Principal Component Analysis
- Independent Component Analysis

How does SciKit Learn … meet our expectations?

A

Extending set of ingredients

(= insights available)



B

Sorting out the hopeless cases

C

Prune to have more efficient

model training & operation

D

Signal, if I missed anything relevant

How to compensate for that in Python space …

Pre-cooked Lambda on
always done transformations

Featuretools

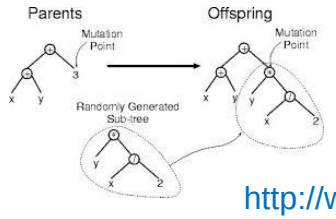
https://www.featuretools.com/

tsfresh

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

TPOT

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



Genetic Programming

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

Go via **Deep Learning**

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

Build your **OWN 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/

Is Principal Component Analysis your Friend or Foe?

PCA

- Pure **Machine to Machine** interface
- Data-space **visualization** required
- Overcomes **mutual correlations** of features without even explicitly checking for them

- **Feature selection** procedure [even in SciKit Learn]. Reduction ≠ selection
- **Humans using** the result of predictions
- **Had to do** oversampling in process of the model preparation
- **Neural network** one of the rival models
- **Non linear** effects of the variables

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

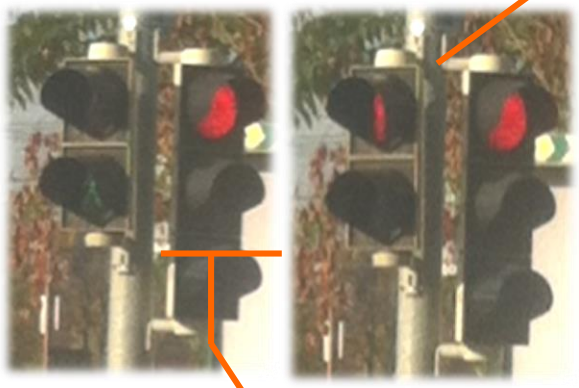
[I have a hypothesis why it works]

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

Bank preference

(Online bank vs. Postal bank)

3

Relationship proxy

4

(133333333 /xxxx
133353333 /xxxx)

MORE INFO

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

Variable Transformations … simply & shortly

Variance X_i, Y
Kurtosis, Skewness
UNI Pearson correlation
5^{th} /95^{th} percentile

Square Root

- Uneven “units“
- To compress variance

SQRT

1 / x transf.

- Opposite causality
- Indirect growth

PWR ?

Logarithm

- Rovnaké/podobné “units“, iná dynamika

LOG

POWER of

- Uneven “units“
- Zväčšiť /zosilniť prem.

SQR / PWR

Normalizing

- Large discrepancy in 5th
- Different decim prec.

NORM