FORNAX

How to win Kaggle competitions?

A handful of Machine Learning tips

by Rafał Cycoń

WWW.FORNAX.AI

kaggle | Search kaggle | Competitions | Datasets | Kernels | Discussion | Jobs | •••

Competitions

12 active competitions

Sort by Prize

Active | All | Entered | Hosted | All Categories | Search

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 8 months to go
$1,200,000
271 teams

Intel & MobileODT Cervical Cancer Screening
Which cancer treatment will be most effective?
Featured · 25 days to go
$100,000
691 teams

Google Cloud & YouTube-8M Video Understanding Challenge
Can you produce the best video tag predictions?
Featured · 6 days to go
$100,000
633 teams

Planet: Understanding the Amazon from Space
Use satellite data to track the human footprint in the Amazon rainforest
Featured · 2 months to go
$60,000
294 teams

Instacart Market Basket Analysis
Which products will an Instacart consumer purchase again?
Featured · 3 months to go
$25,000
328 teams

Rafal Cycon (blaine)

CTO at fornax.ai
Poland
Joined 3 years ago · last seen in the past day
http://fornax.ai

Home | Competitions (11) | Kernels (0) | Discussion (23) | Datasets (2) | More

Competitions Master

Current Rank	Highest Rank
204 of 57,910	30

3	3	2

Grasp-and-Lift EEG Detec... · 2 years ago · Top 1%	1st of 379
BCI Challenge @ NER 2015 · 2 years ago · Top 1%	1st of 260
Allstate Purchase Predicti... · 3 years ago · Top 1%	10th of 1568

Kernels Contributor

Unranked

0	0	0

No kernel results

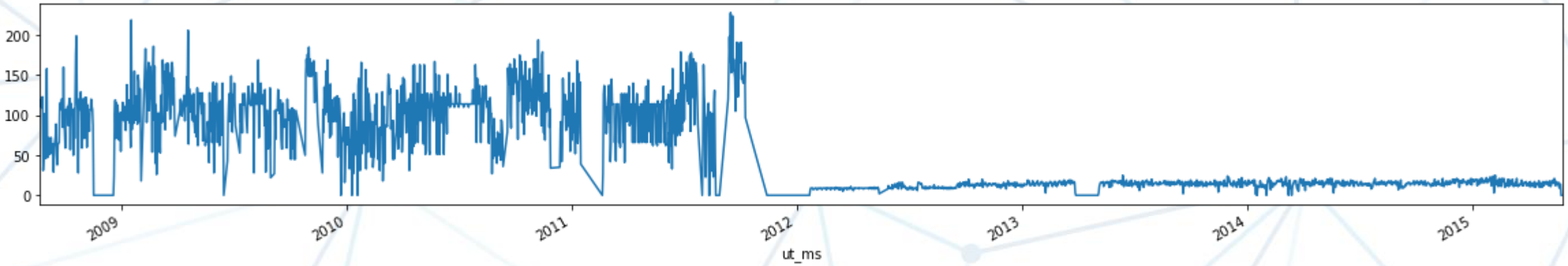
Know your data

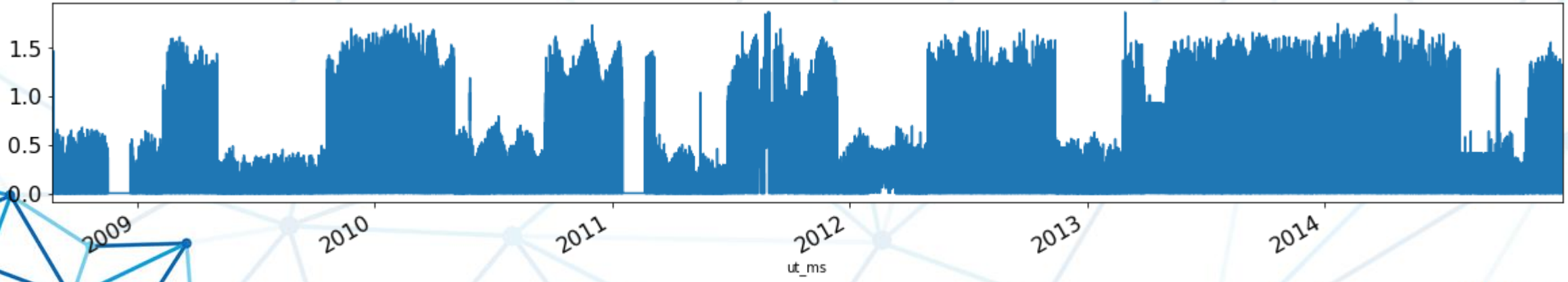
- Have a look on features
- Look on the target variable(s)
- Check for outliers
- Normalize where necessary

- Plot things and statistics

Know your data

MEX

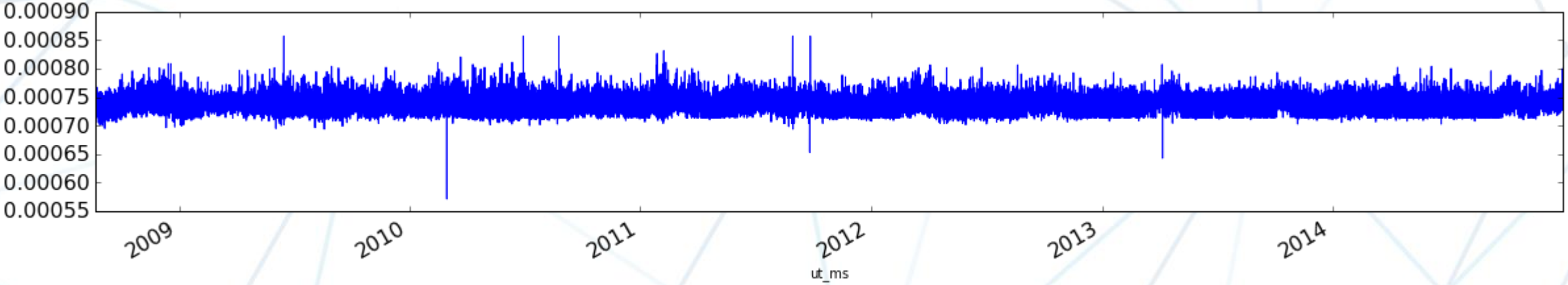


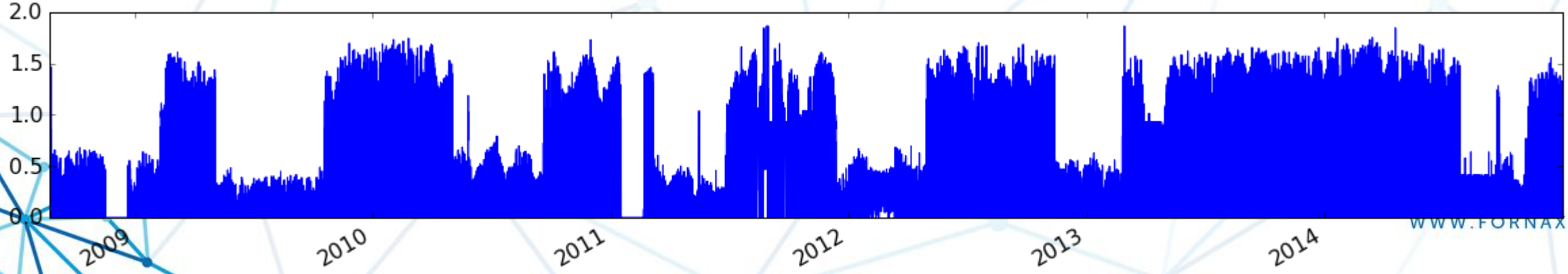


Know your metric...

...and understand it

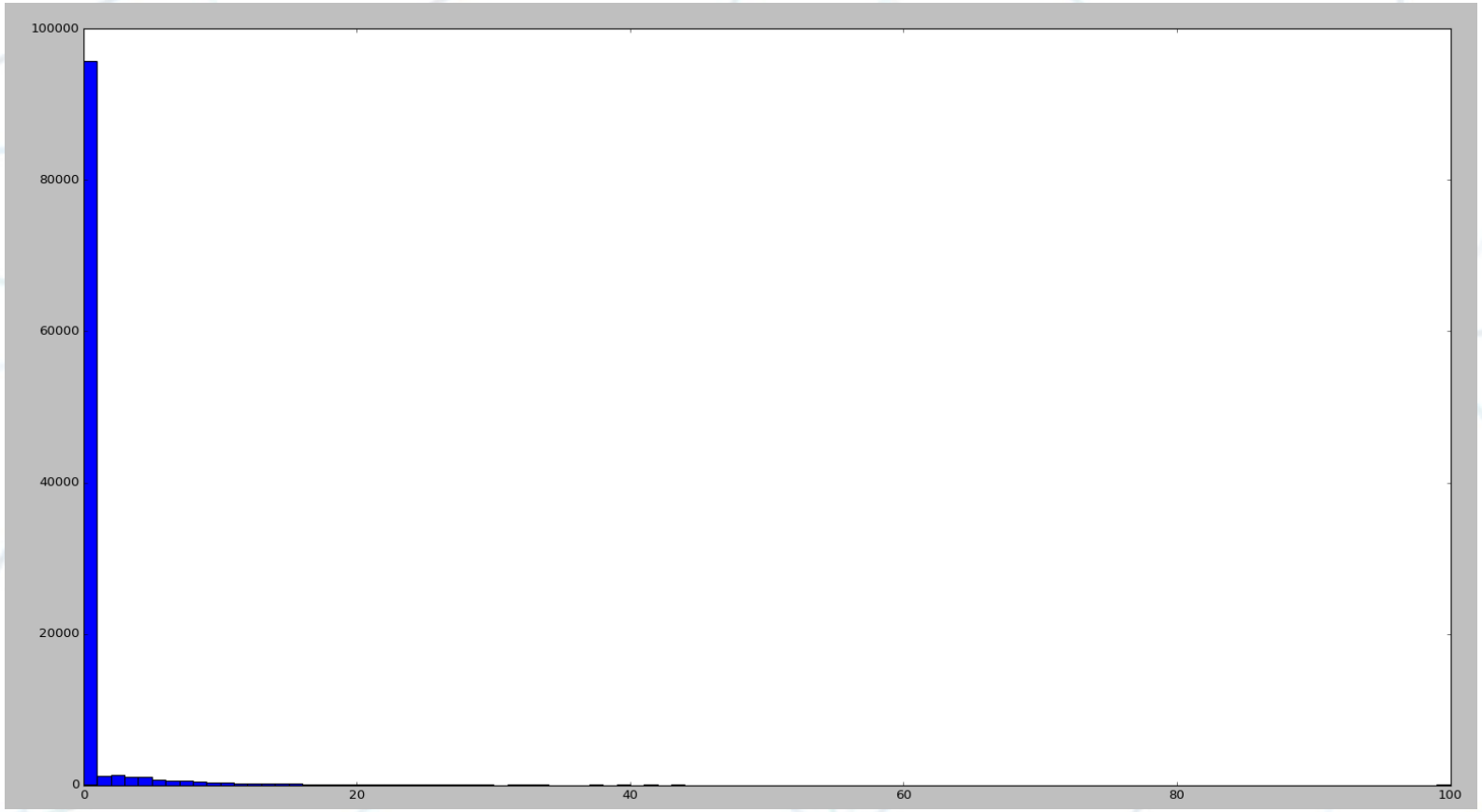
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$





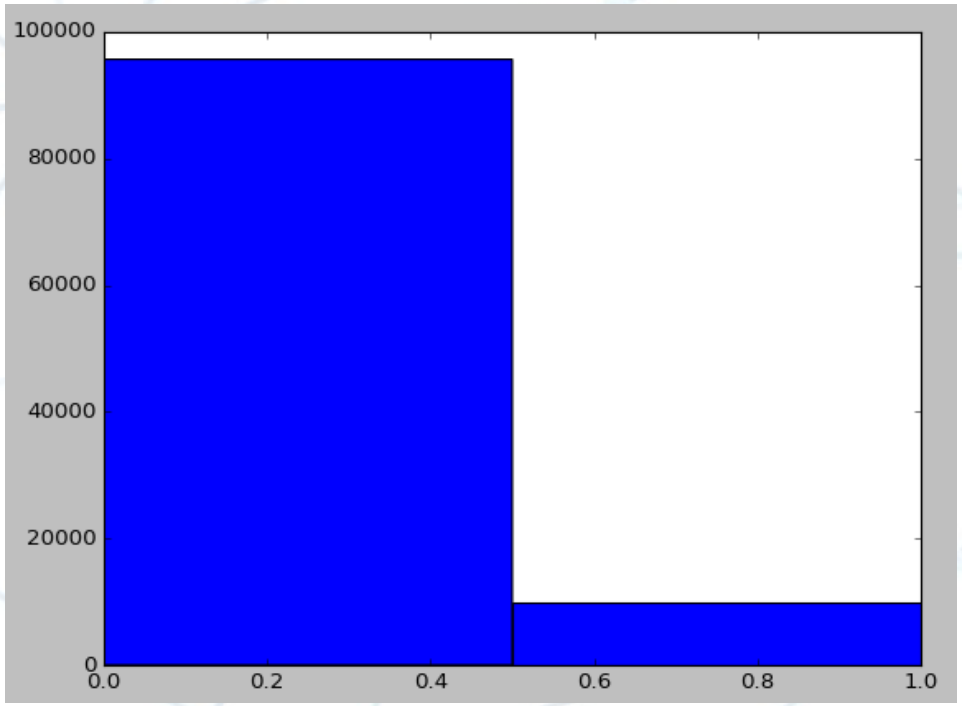
Think out of the box

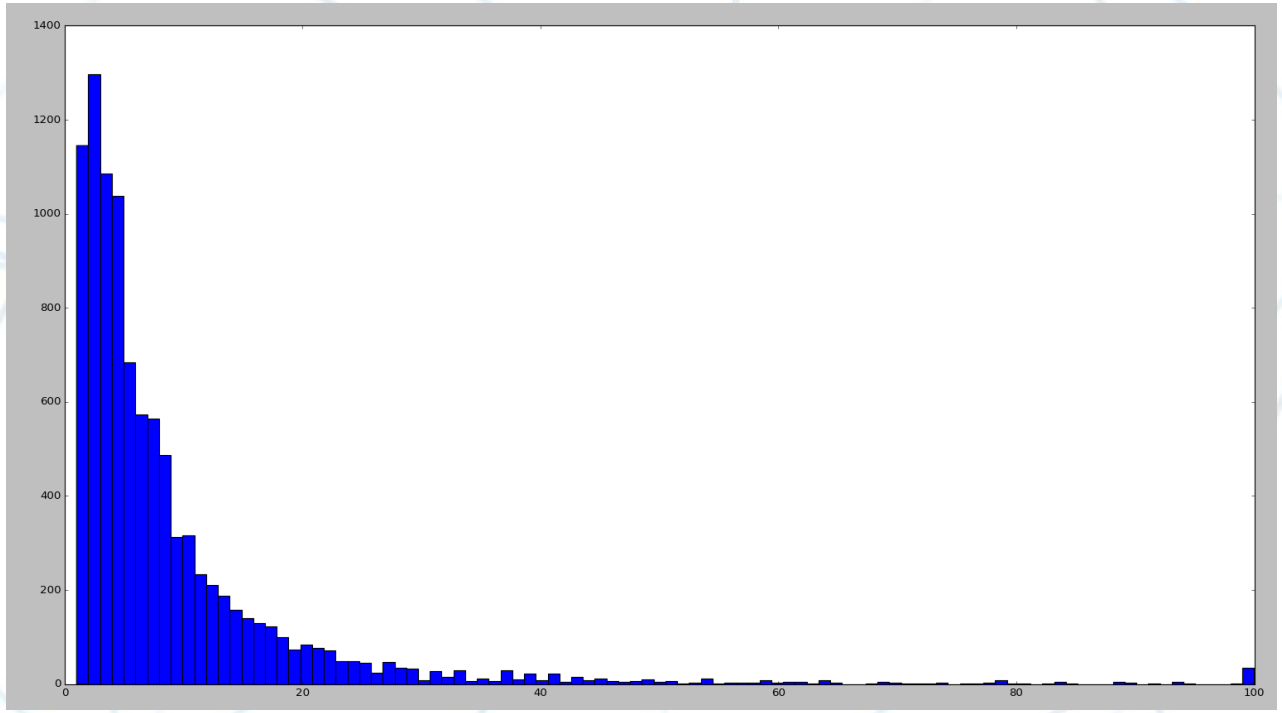
Loan default prediction



Think out of the box

Loan default prediction





Think out of the box

Allstate purchase prediction

Visit 1: A1, B2, C1

Visit 2: A2, B2, C1

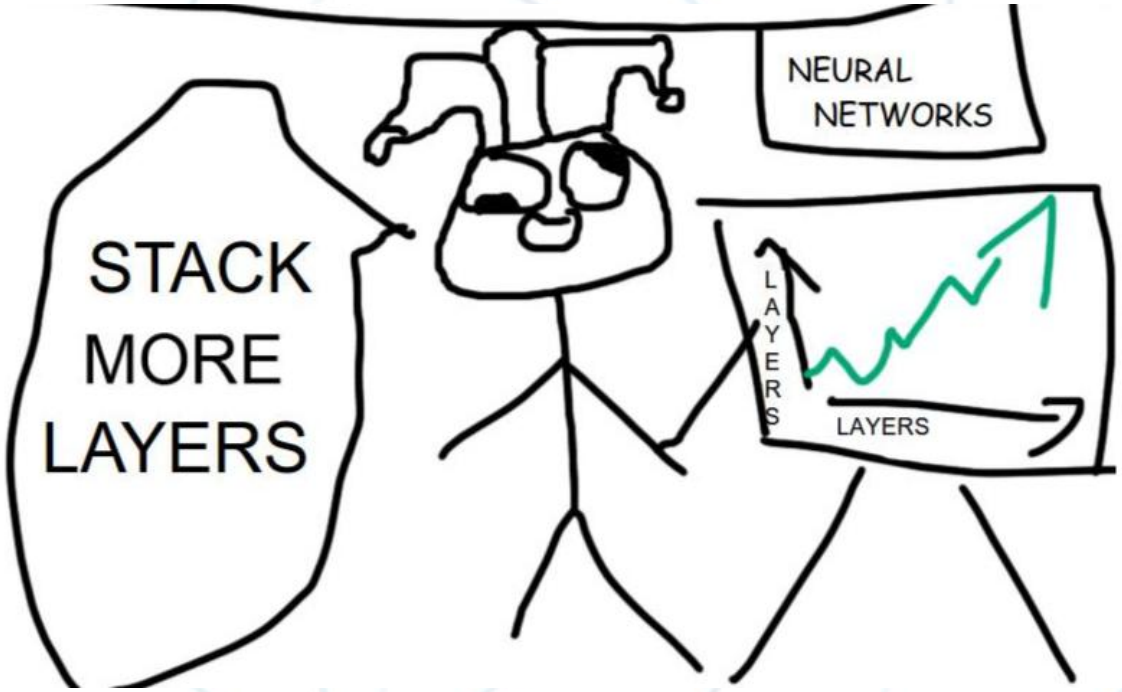
....

Visit 4: A2, B2, C1 (purchase)

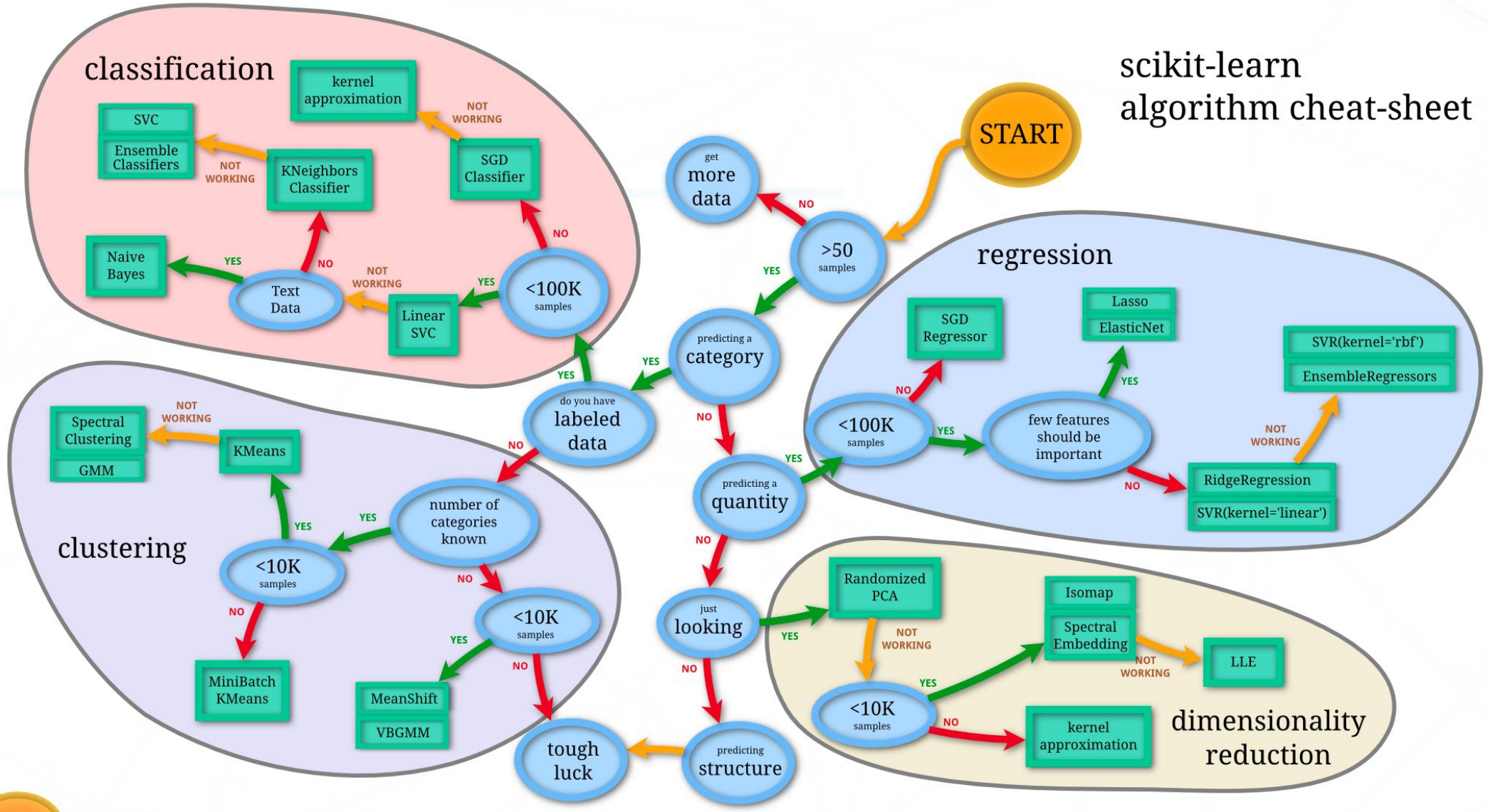
Which algorithm to use?

Deep learning!

Or else...



scikit-learn algorithm cheat-sheet



Back

Which algorithm to use?

Images, signals, text: deep learning

Otherwise:

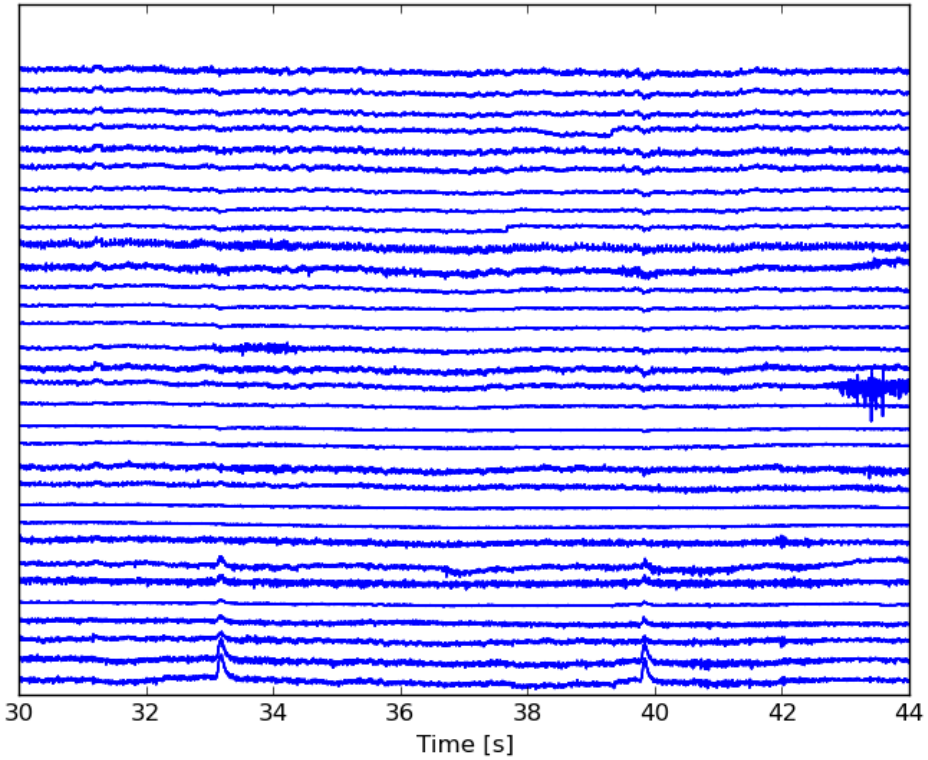
- linear/logistic regression
- tree-based methods (gradient boosting)
- (surprise surprise) deep learning
- and whatever else works

Start with the simplest things, then move to more complex algorithms.

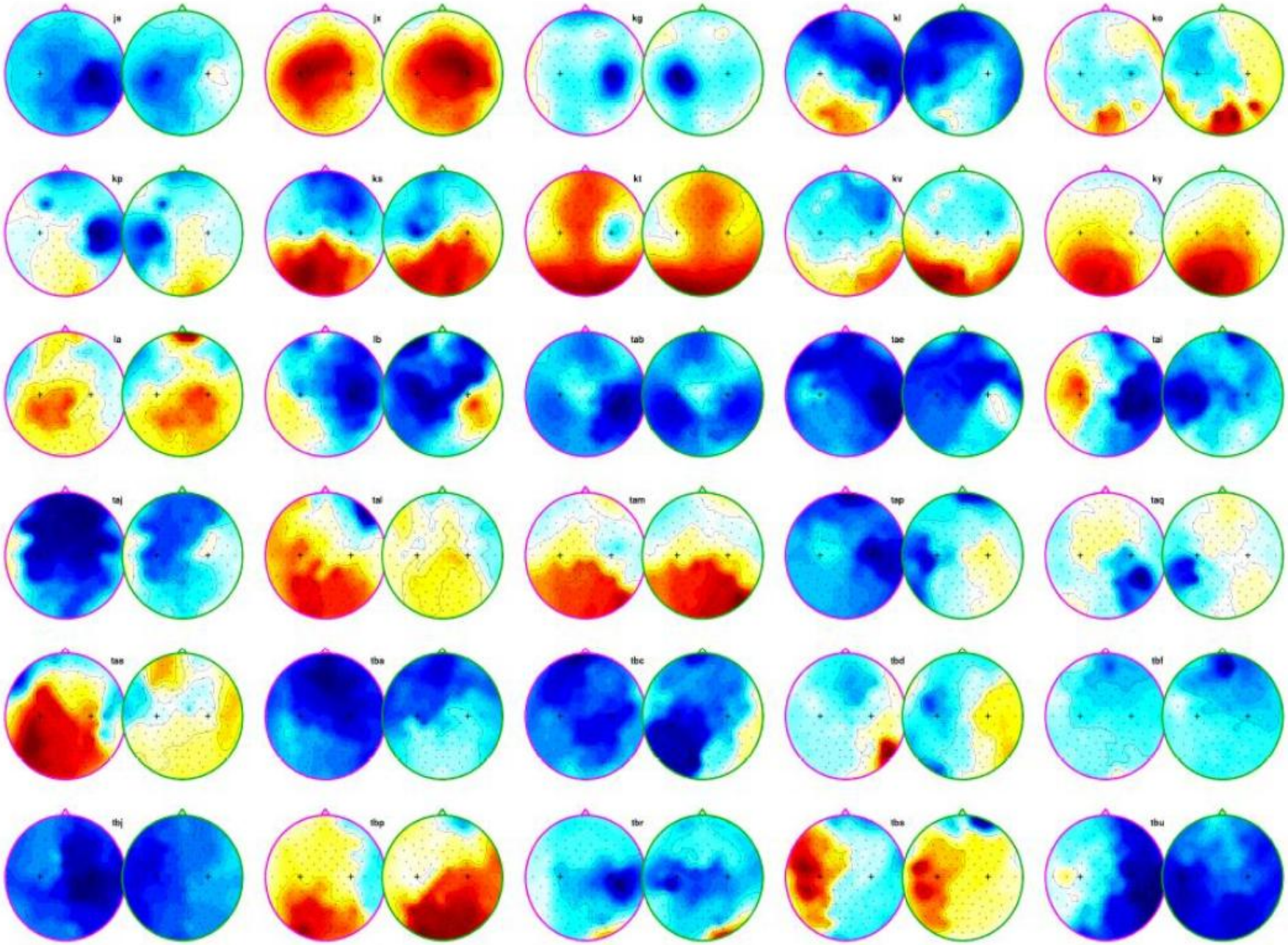
Data > algorithm

- Create new features
- Deep learning: data augmentation

EEG



EEG


jx
jx
kg
kl
ko
kp
ks
kt
kv
ky
la
lb
tab
tae
tai
taj
tal
tam
tap
taq
tas
tba
tbc
tbd
tbf
tbj
tbp
tbr
tbs
tbu

EEG

+ Clever features
+ Linear model (regularized)

#	△pub	Team Name ✱ in the money	Kernel	Team Members	Score
1	—	✱ the overfitting avengers			0.87224
2	—	✱ Devin			0.85669
3	—	✱ H2O.ai			0.81850
4	—	barrack_d(NER)			0.76921
5	—	Jose M.			0.74790

Ensembling

- Averaging
- Weighted averaging
- Stacking

Netflix prize

$1mln prize for improving recommendations by 10%

Solution: ensemble all the things!

- >500 models
- Unusable in production
- Ended up implementing a worse, but an elegant and practical solution

Seizure prediction

#	—	Team Name ✱ in the money	Kernel	Team Members	Score	Entries	Last
1	—	✱ Medrr			0.90316	264	3y
2	—	✱ cgp & Alexandre & blaine			0.86350	393	3y

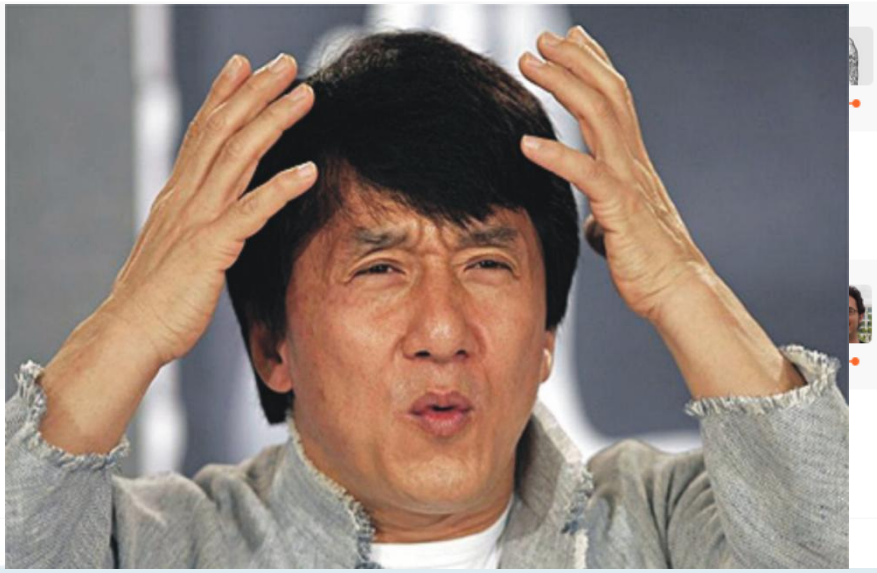
Your Best Entry ↑

Your submission scored 0.80582, which is not an improvement of your best score. Keep trying!

3	—	✱ Michael Hills			0.86248	427	3y
4	—	QMSDP			0.85951	501	3y
5	—	Carlos Fernandez			0.84225	299	3y

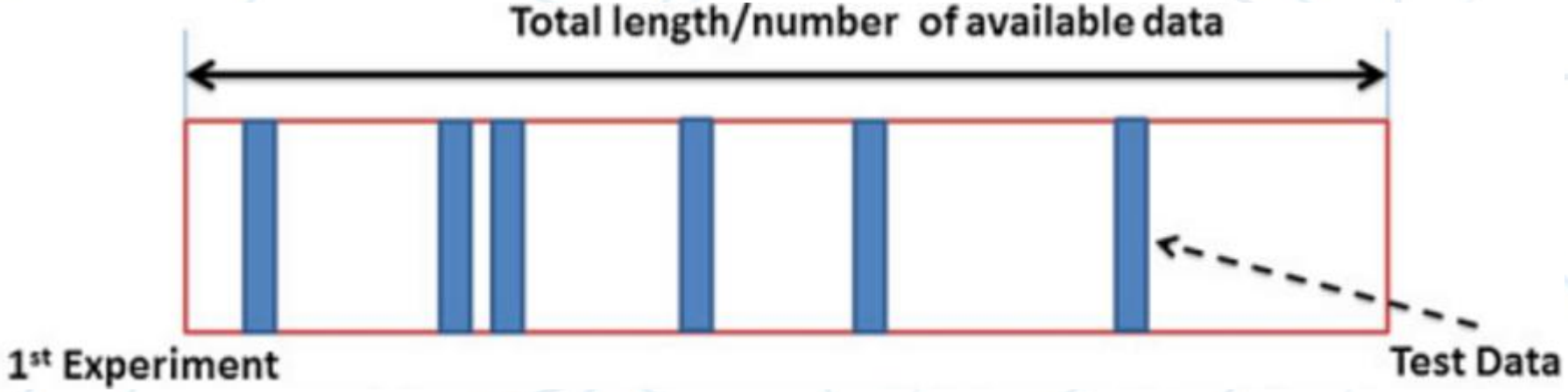
Seizure prediction

#	△pub	Team Name ✱ in the money	Kernel	Team Members	Score	Entries	Last
1	—	✱ Medrr			0.83993	264	3y
2	—	✱ QMSDP			0.81962	501	3y
3	—	✱ Birchwood			0.80079	160	3y
4	—	ESAI CEU-UCH			0.79347	182	3y
5	—	Michael Hills			0.79251	427	3y
26	—	cgp & Alexandre & blaine			0.75120	393	3y



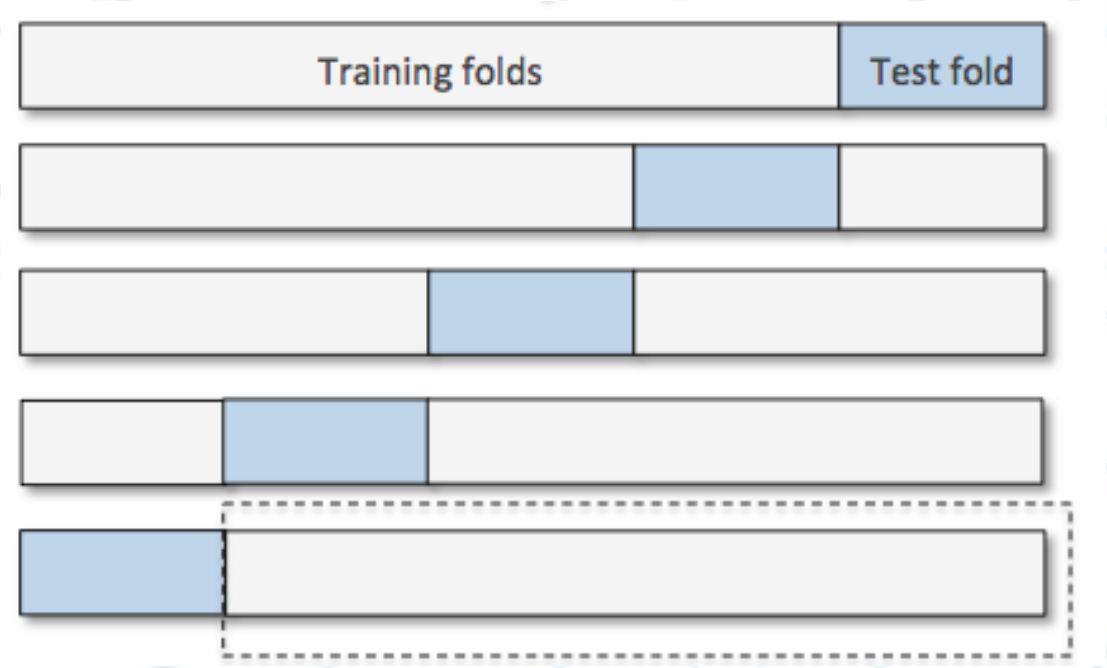
How do you know if it works?

Validate on a random sample of data

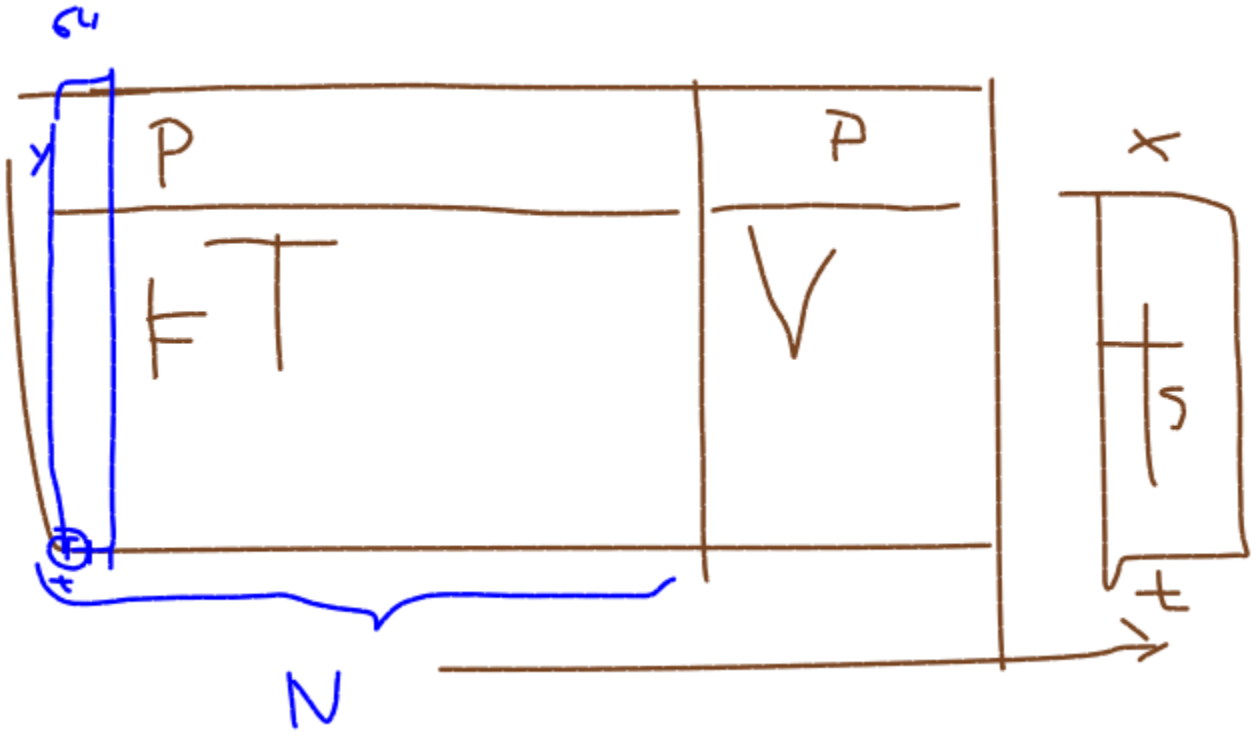


How do you know if it works?

Cross validation



Train/validation split in MEX



Final golden rules

Validate your model on a “real” scenario

this is the only valid approach of testing your model.

If it unexpectedly works REALLY great

then probably something is wrong.

FORNAX

Rafał Cycoń
CTO / co-founder

+48 607 697 054

WWW.FORNAX.AI