FACULTY OF ENGINEERING
AND ARCHITECTURE

Third place solution: Liverpool Ion Switching

Gilles Vandewiele

1) My Journey on Kaggle

2) The Liverpool Competition

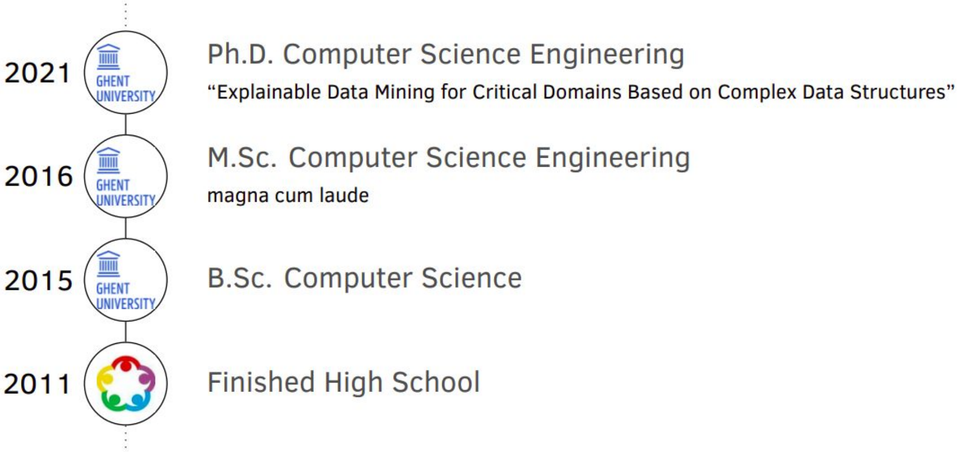
    a) Problem Statement
    b) Data Processing I
    c) Baseline
    d) HMMs
    e) Data Processing II
    f) “Advanced” HMMs
    g) The Leak
    h) Conclusion/Summary

3) Bis: Kaggle - General Tips & Tricks

My Journey on Kaggle

My background

- **Postdoctoral researcher @ IDLab**
- **Computer & Data Scientist, not an engineer**
- **Kaggle Master**

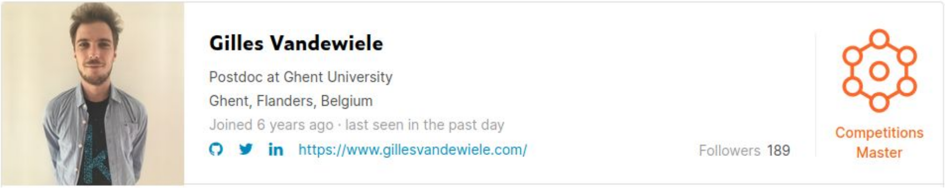


2021 Ph.D. Computer Science Engineering
“Explainable Data Mining for Critical Domains Based on Complex Data Structures”

2016 M.Sc. Computer Science Engineering
magna cum laude

2015 B.Sc. Computer Science

2011 Finished High School



Gilles Vandewiele

Postdoc at Ghent University

Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

Competitions Master

Kaggle - the home of data science

Over 5 million registered data scientists

Four different categories:

1) **Competitions**

   Prediction / Code / Analysis / Simulation

   Merit-based achievements

2) **Datasets**
3) **Notebooks**
4) **Discussion**

} popularity-based achievements

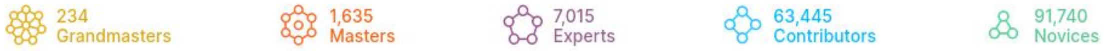
Novice

Contributor

Expert

Master

Grandmaster



234 Grandmasters

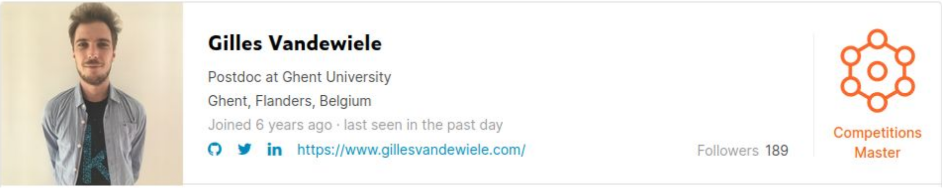
1,635 Masters

7,015 Experts

63,445 Contributors

91,740 Novices

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

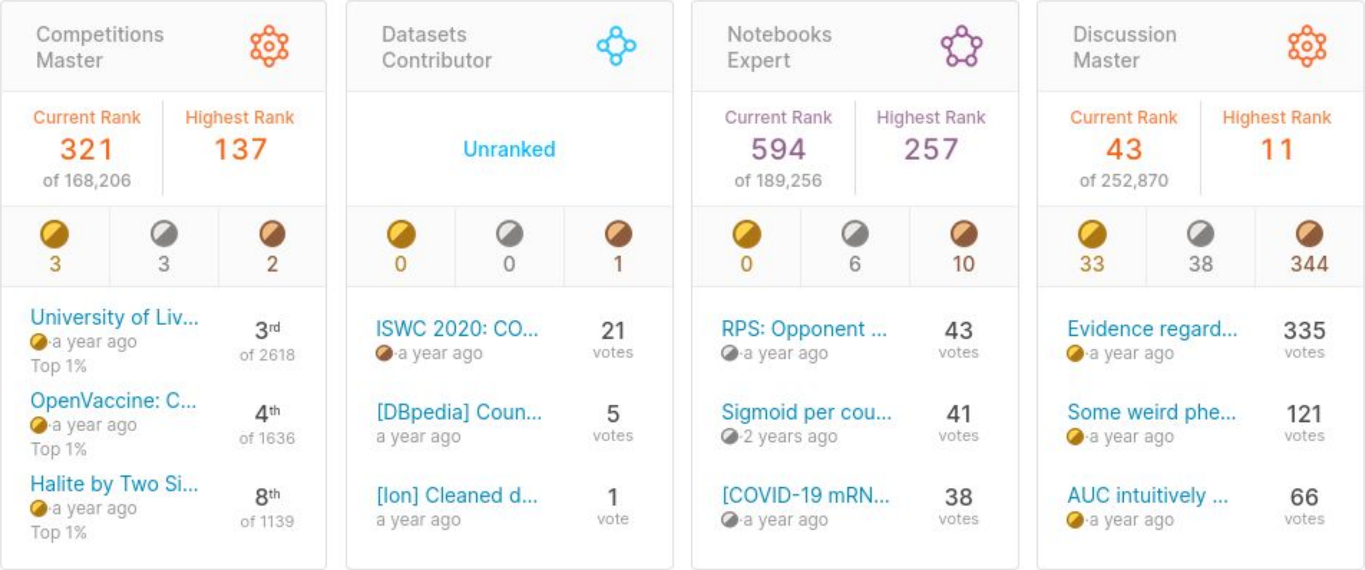
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

Competitions Master



Competitions Master

Current Rank	Highest Rank
321 of 168,206	137

●	●	●
3	3	2

University of Liv... a year ago Top 1%	3rd of 2618
OpenVaccine: C... a year ago Top 1%	4th of 1636
Halite by Two Si... a year ago Top 1%	8th of 1139

Datasets Contributor

Unranked

●	●	●
0	0	1

ISWC 2020: CO... a year ago	21 votes
[DBpedia] Coun... a year ago	5 votes
[Ion] Cleaned d... a year ago	1 vote

Notebooks Expert

Current Rank	Highest Rank
594 of 189,256	257

●	●	●
0	6	10

RPS: Opponent ... a year ago	43 votes
Sigmoid per cou... 2 years ago	41 votes
[COVID-19 mRN... a year ago	38 votes

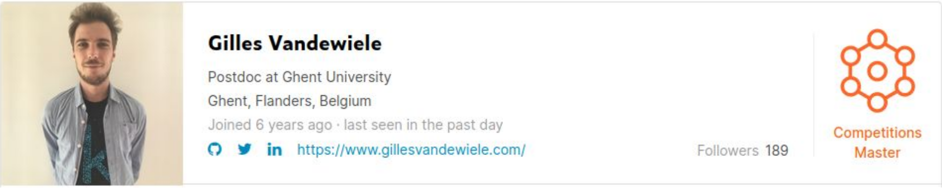
Discussion Master

Current Rank	Highest Rank
43 of 252,870	11

●	●	●
33	38	344

Evidence regard... a year ago	335 votes
Some weird phe... a year ago	121 votes
AUC intuitively ... a year ago	66 votes

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

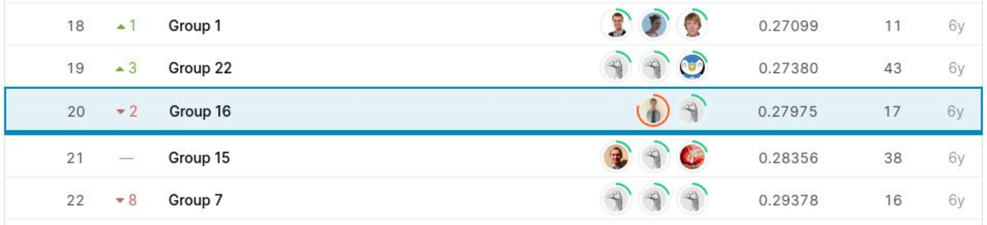
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

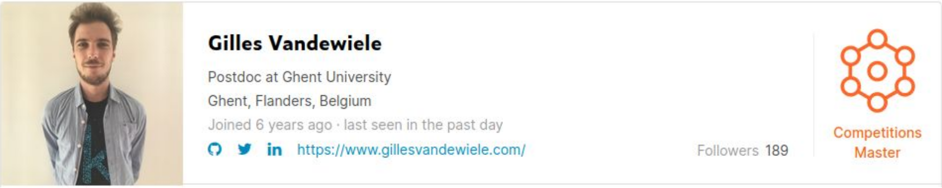
Competitions Master



18	▲ 1	Group 1		0.27099	11	6y
19	▲ 3	Group 22		0.27380	43	6y
20	▼ 2	Group 16		0.27975	17	6y
21	—	Group 15		0.28356	38	6y
22	▼ 8	Group 7		0.29378	16	6y

Oct. 2015 - created account for ML project at UGent (rank 20/31)

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

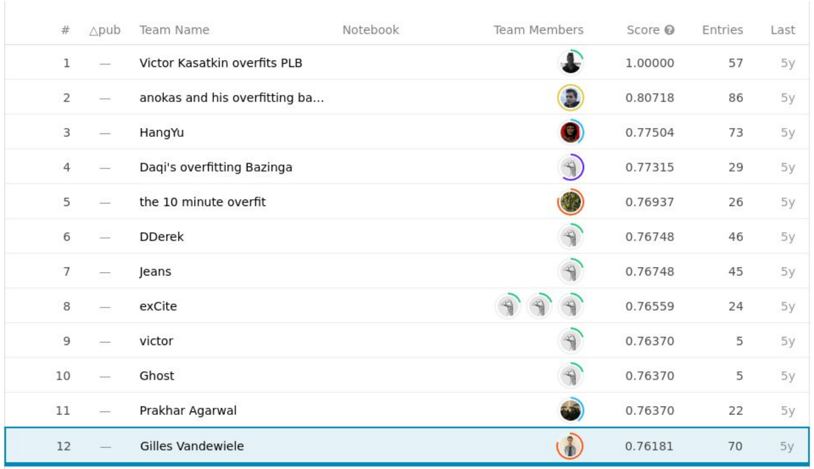
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

Competitions Master



#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Victor Kasatkin overfits PLB			1.00000	57	5y
2	—	anokas and his overfitting ba...			0.80718	86	5y
3	—	HangYu			0.77504	73	5y
4	—	Daqi's overfitting Bazinga			0.77315	29	5y
5	—	the 10 minute overfit			0.76937	26	5y
6	—	DDerek			0.76748	46	5y
7	—	Jeans			0.76748	45	5y
8	—	exCite			0.76559	24	5y
9	—	victor			0.76370	5	5y
10	—	Ghost			0.76370	5	5y
11	—	Prakhar Agarwal			0.76370	22	5y
12	—	Gilles Vandewiele			0.76181	70	5y

Oct. 2015

Oct. 2016 - halloween playground competition (rank 12/762)

The road to Kaggle Master...

Gilles Vandewiele

Postdoc at Ghent University

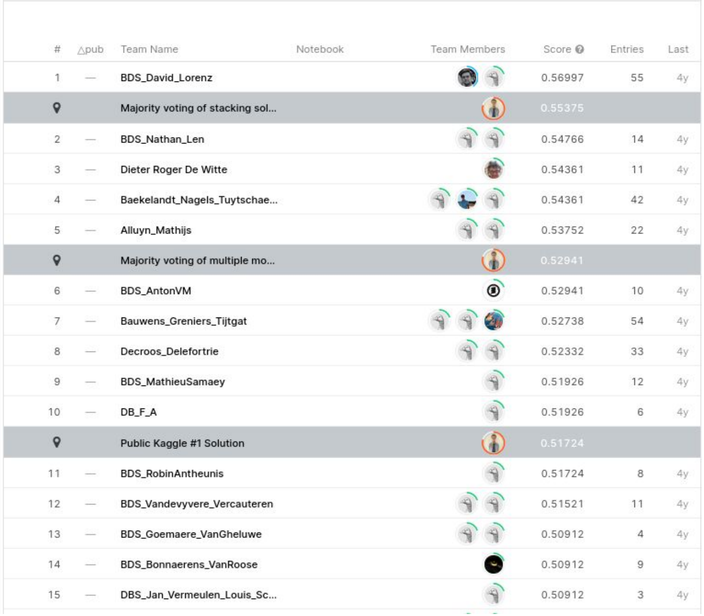
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

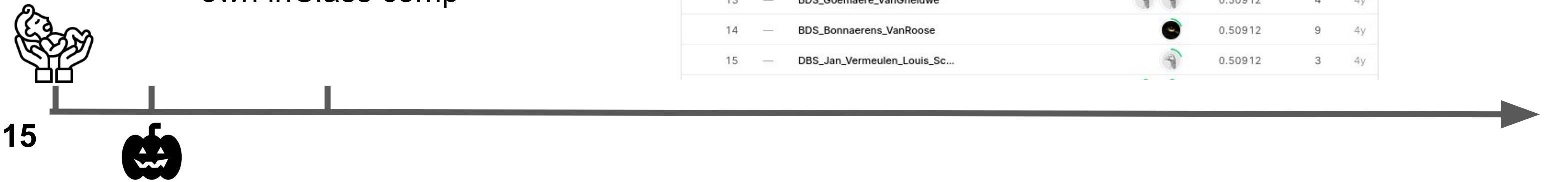
Followers 189

Competitions Master



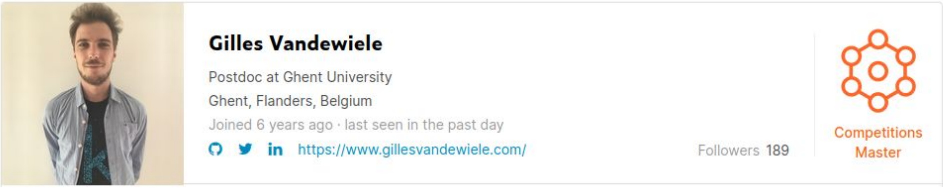
#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	BDS_David_Lorenz			0.56997	55	4y
		Majority voting of stacking sol...			0.55375		
2	—	BDS_Nathan_Len			0.54766	14	4y
3	—	Dieter Roger De Witte			0.54361	11	4y
4	—	Baekelandt_Nagels_Tuytschae...			0.54361	42	4y
5	—	Alluyn_Mathijs			0.53752	22	4y
		Majority voting of multiple mo...			0.52941		
6	—	BDS_AntonVM			0.52941	10	4y
7	—	Bauwens_Greniers_Tijtgat			0.52738	54	4y
8	—	Decroos_Delefortrie			0.52332	33	4y
9	—	BDS_MathieuSamaey			0.51926	12	4y
10	—	DB_F_A			0.51926	6	4y
		Public Kaggle #1 Solution			0.51724		
11	—	BDS_RobinAntheunis			0.51724	8	4y
12	—	BDS_Vandevyvere_Vercauteren			0.51521	11	4y
13	—	BDS_Goemaere_VanGheluwe			0.50912	4	4y
14	—	BDS_Bonnaerens_VanRoose			0.50912	9	4y
15	—	DBS_Jan_Vermeulen_Louis_Sc...			0.50912	3	4y

Jan. 2017 - hosted own inClass comp



Oct. 2015

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

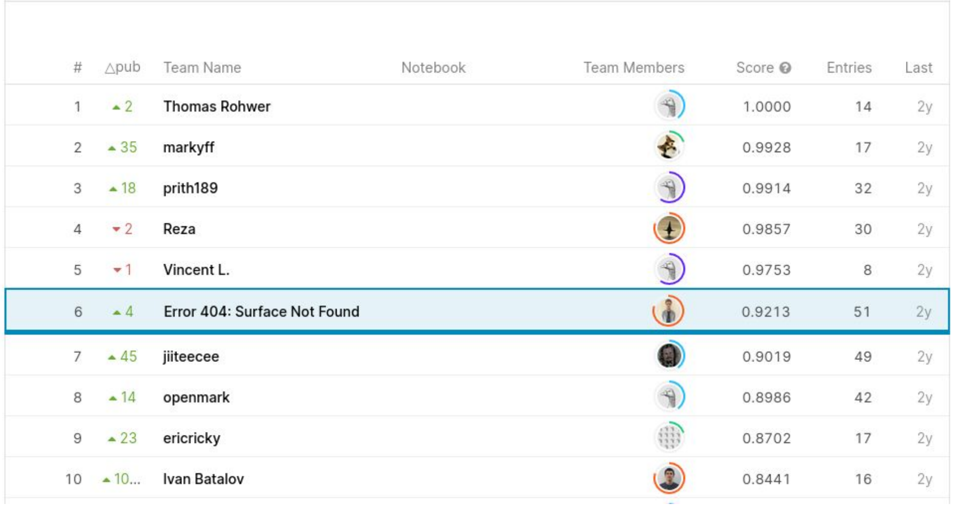
Competitions Master

Santander Customer Transaction Prediction

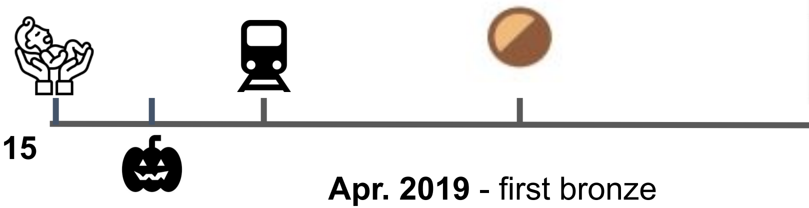
Can you identify who will make a transaction?

Featured · 2 years ago

796/8751



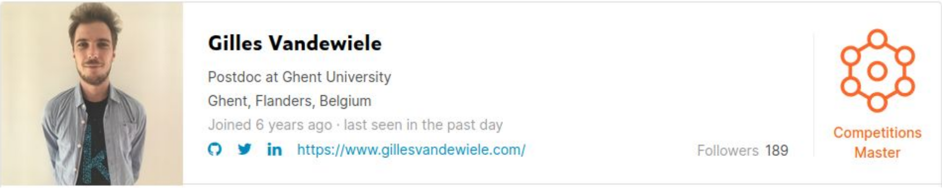
#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲2	Thomas Rohwer			1.0000	14	2y
2	▲35	markyff			0.9928	17	2y
3	▲18	prith189			0.9914	32	2y
4	▼2	Reza			0.9857	30	2y
5	▼1	Vincent L.			0.9753	8	2y
6	▲4	Error 404: Surface Not Found			0.9213	51	2y
7	▲45	jiiteecee			0.9019	49	2y
8	▲14	openmark			0.8986	42	2y
9	▲23	ericricky			0.8702	17	2y
10	▲10...	Ivan Batalov			0.8441	16	2y



Oct. 2015

Apr. 2019 - first bronze medal (Santander comp) and rank 6/1443 (+ swag) in CareerCon comp.

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

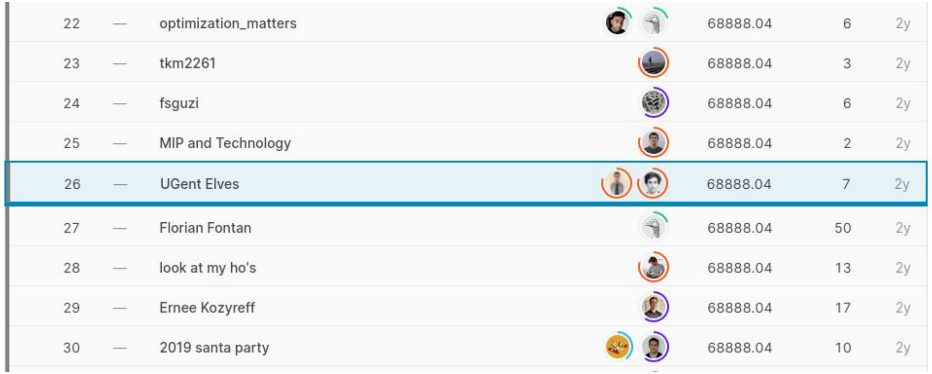
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

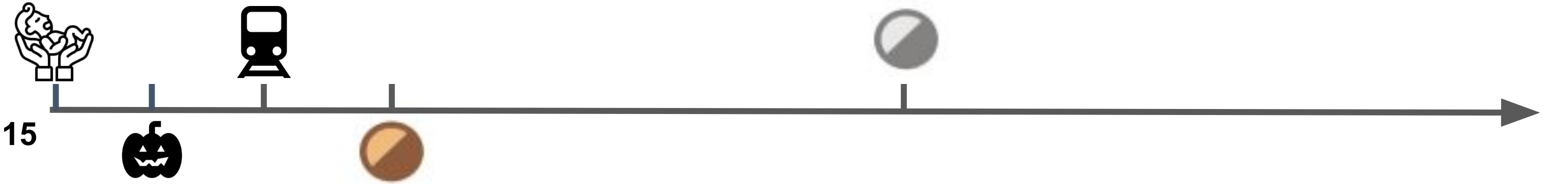
Followers 189

Competitions Master



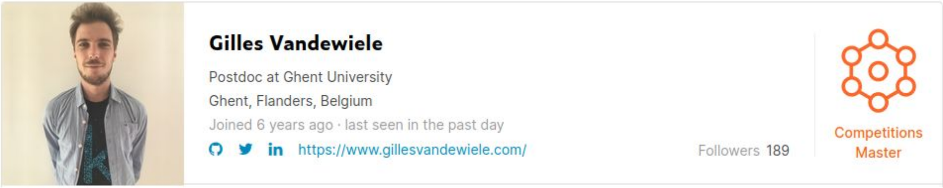
22	—	optimization_matters	68888.04	6	2y
23	—	tkm2261	68888.04	3	2y
24	—	fsguzi	68888.04	6	2y
25	—	MIP and Technology	68888.04	2	2y
26	—	UGent Elves	68888.04	7	2y
27	—	Florian Fontan	68888.04	50	2y
28	—	look at my ho's	68888.04	13	2y
29	—	Ernee Kozyreff	68888.04	17	2y
30	—	2019 santa party	68888.04	10	2y

Jan 2020 - first silver medal and competition **expert**!



Oct. 2015

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

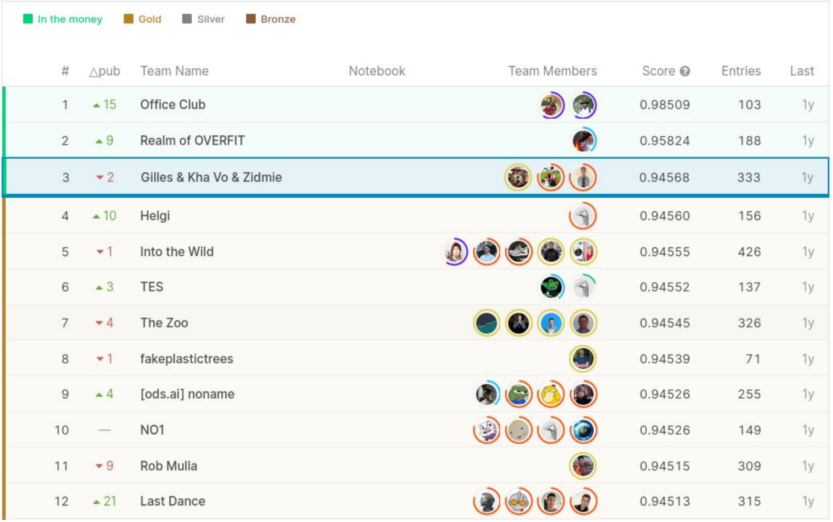
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

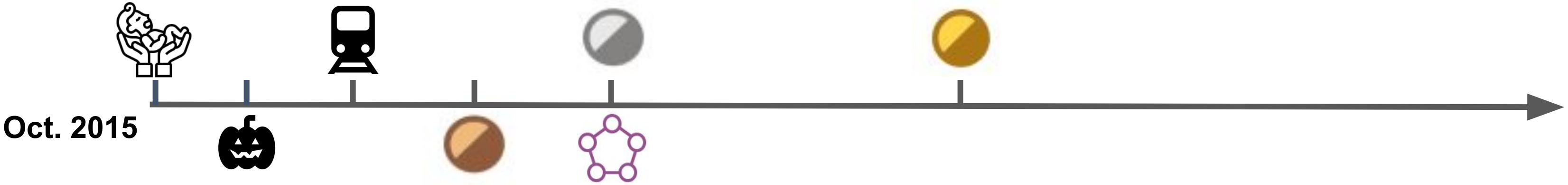
Competitions Master



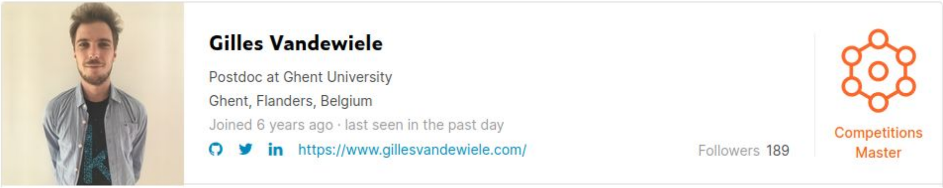
🟩 In the money 🟨 Gold ◼ Silver 🟫 Bronze

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 15	Office Club			0.98509	103	1y
2	▲ 9	Realm of OVERFIT			0.95824	188	1y
3	▼ 2	Gilles & Kha Vo & Zidmie			0.94568	333	1y
4	▲ 10	Helgi			0.94560	156	1y
5	▼ 1	Into the Wild			0.94555	426	1y
6	▲ 3	TES			0.94552	137	1y
7	▼ 4	The Zoo			0.94545	326	1y
8	▼ 1	fakeplastictrees			0.94539	71	1y
9	▲ 4	[ods.ai] noname			0.94526	255	1y
10	—	NO1			0.94526	149	1y
11	▼ 9	Rob Mulla			0.94515	309	1y
12	▲ 21	Last Dance			0.94513	315	1y

May 2020 - third place, first gold medal & first time in the money



The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

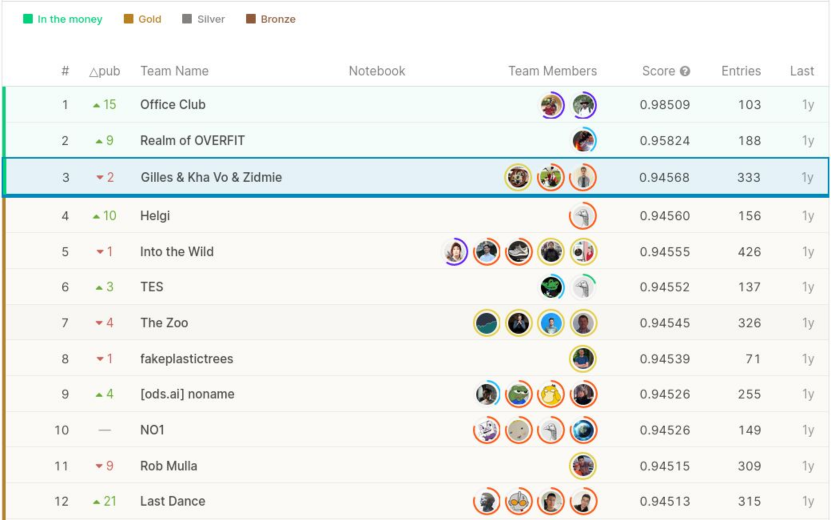
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

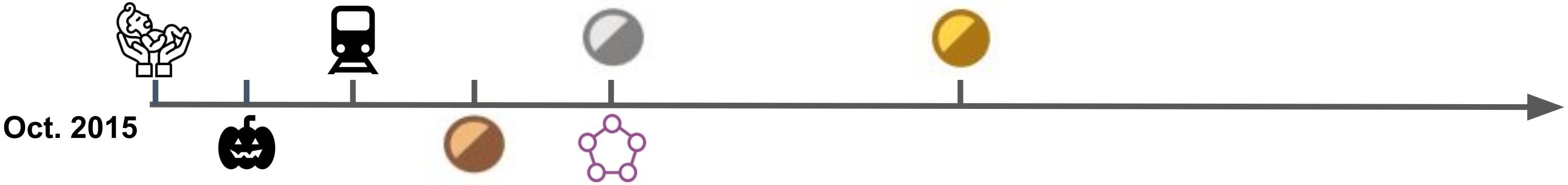
Competitions Master



🟩 In the money 🟧 Gold ⬜ Silver 🟫 Bronze

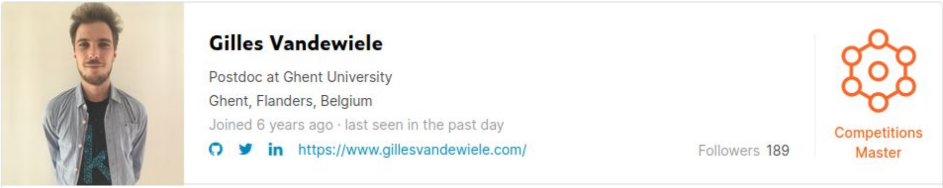
#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲15	Office Club			0.98509	103	1y
2	▲9	Realm of OVERFIT			0.95824	188	1y
3	▼2	Gilles & Kha Vo & Zidmie			0.94568	333	1y
4	▲10	Helgi			0.94560	156	1y
5	▼1	Into the Wild			0.94555	426	1y
6	▲3	TES			0.94552	137	1y
7	▼4	The Zoo			0.94545	326	1y
8	▼1	fakeplastictrees			0.94539	71	1y
9	▲4	[ods.ai] noname			0.94526	255	1y
10	—	NO1			0.94526	149	1y
11	▼9	Rob Mulla			0.94515	309	1y
12	▲21	Last Dance			0.94513	315	1y

May 2020 - third place, first gold medal & first time in the money



Focus of today's presentation!

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

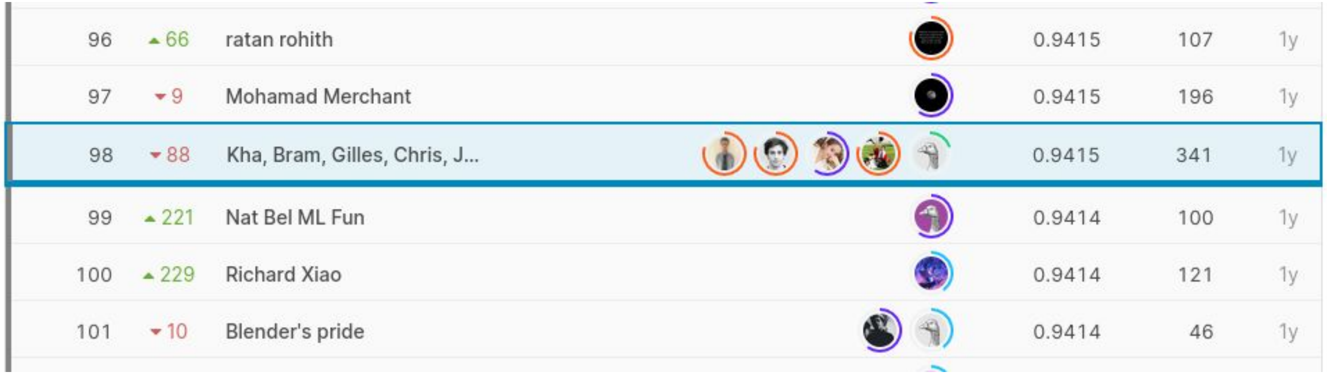
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

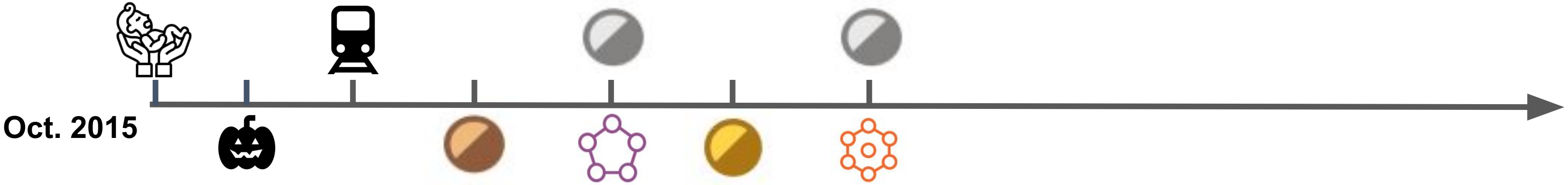
Followers 189

Competitions Master

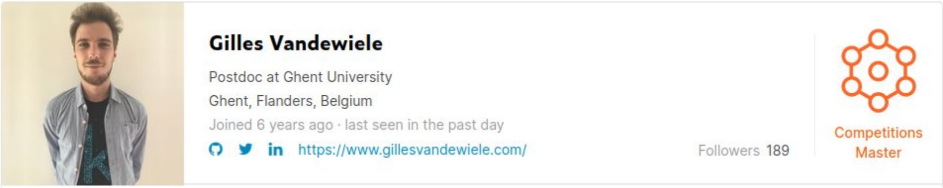


96	▲ 66	ratan rohith	0.9415	107	1y
97	▼ 9	Mohamad Merchant	0.9415	196	1y
98	▼ 88	Kha, Bram, Gilles, Chris, J...	0.9415	341	1y
99	▲ 221	Nat Bel ML Fun	0.9414	100	1y
100	▲ 229	Richard Xiao	0.9414	121	1y
101	▼ 10	Blender's pride	0.9414	46	1y

Aug. 2020 - silver medal and competition master!



The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

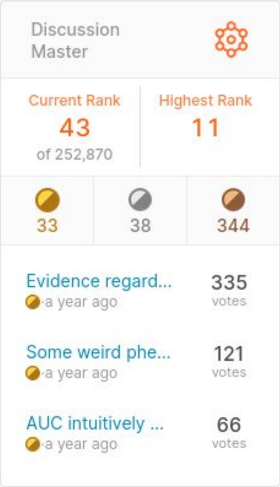
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

Competitions Master

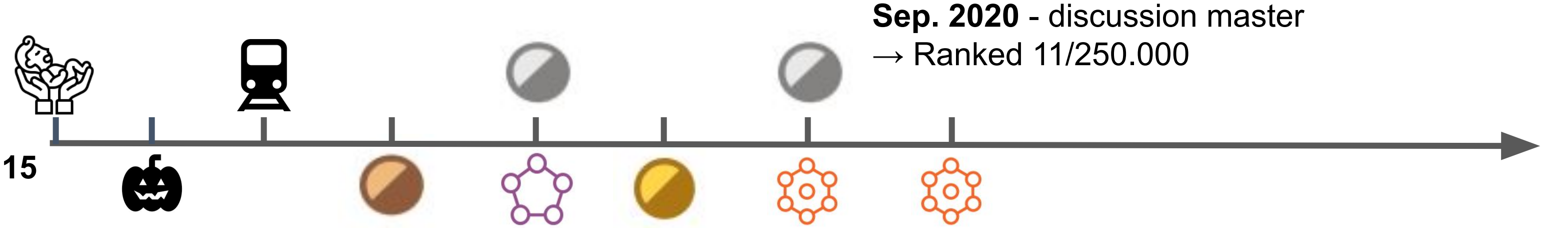


Discussion Master

Current Rank	Highest Rank
43 of 252,870	11

●	●	●
33	38	344

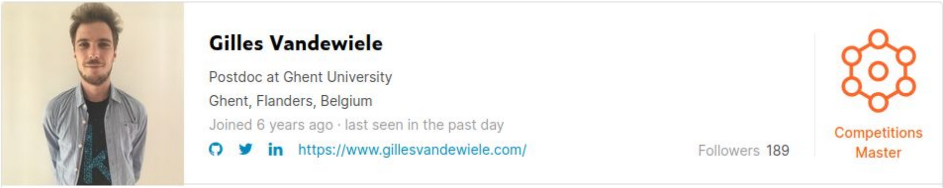
Evidence regard... a year ago	335 votes
Some weird phe... a year ago	121 votes
AUC intuitively ... a year ago	66 votes



Oct. 2015

Sep. 2020 - discussion master
→ Ranked 11/250.000

The road to Kaggle Master...



Gilles Vandewiele

Postdoc at Ghent University

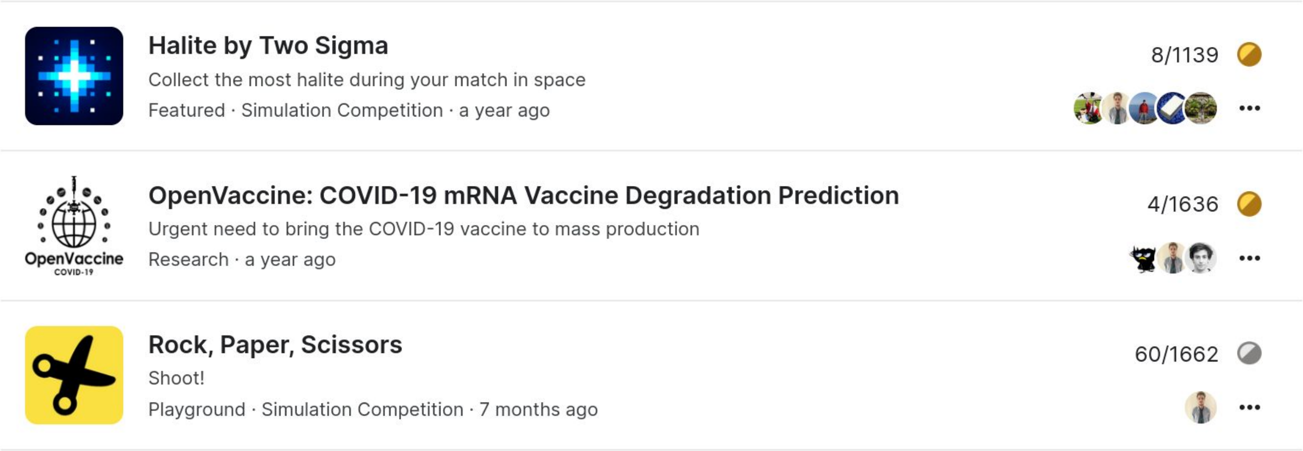
Ghent, Flanders, Belgium

Joined 6 years ago · last seen in the past day

https://www.gillesvandewiele.com/

Followers 189

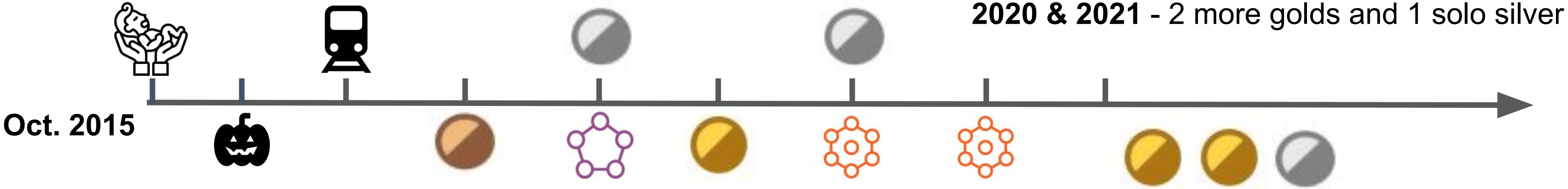
Competitions Master



Halite by Two Sigma
Collect the most halite during your match in space
Featured · Simulation Competition · a year ago
8/1139

OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction
Urgent need to bring the COVID-19 vaccine to mass production
Research · a year ago
4/1636

Rock, Paper, Scissors
Shoot!
Playground · Simulation Competition · 7 months ago
60/1662



Oct. 2015

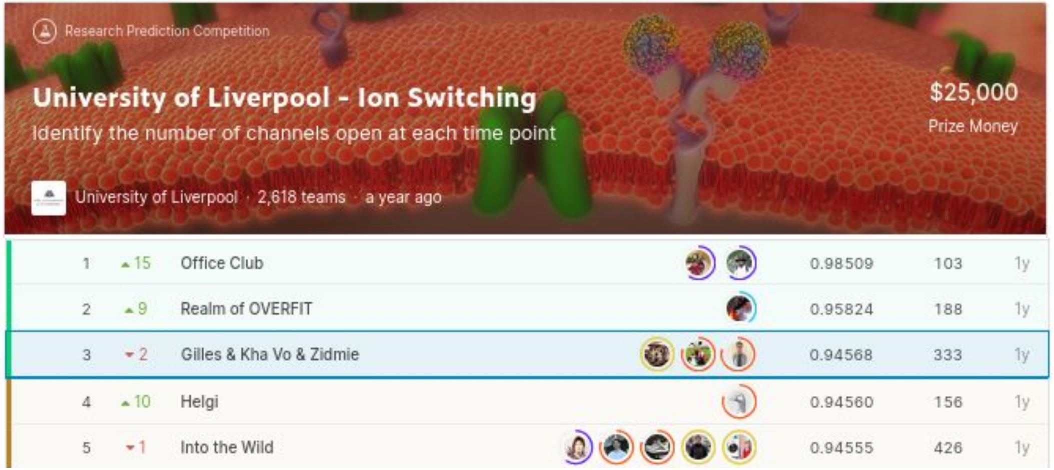
2020 & 2021 - 2 more golds and 1 solo silver

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

Problem statement

Teamwork makes the dream work



Research Prediction Competition

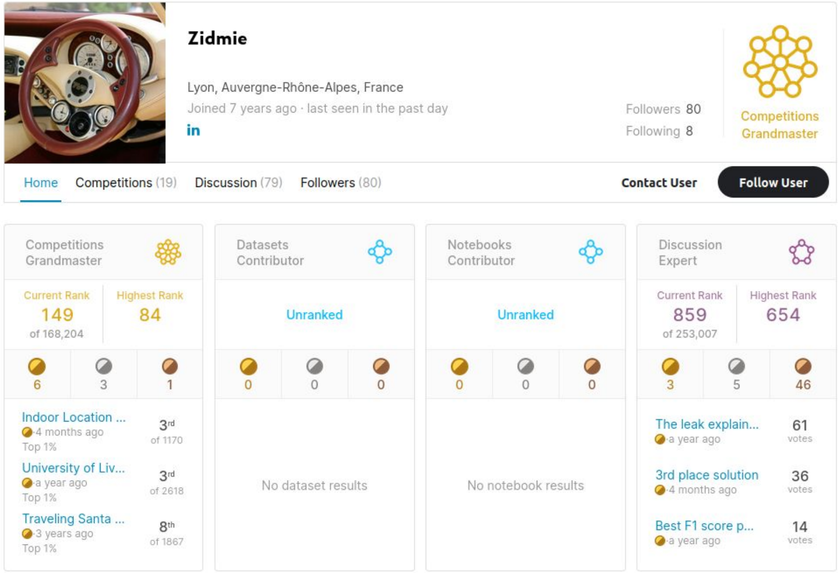
University of Liverpool - Ion Switching

Identify the number of channels open at each time point

$25,000 Prize Money

University of Liverpool · 2,618 teams · a year ago

1	▲15	Office Club	0.98509	103	1y
2	▲9	Realm of OVERFIT	0.95824	188	1y
3	▼2	Gilles & Kha Vo & Zidmie	0.94568	333	1y
4	▲10	Helgi	0.94560	156	1y
5	▼1	Into the Wild	0.94555	426	1y



Zidmie

Lyon, Auvergne-Rhône-Alpes, France

Joined 7 years ago · last seen in the past day

in

Followers 80

Following 8

Competitions Grandmaster

Home | Competitions (19) | Discussion (79) | Followers (80) | Contact User | Follow User

Competitions Grandmaster

Current Rank 149 of 168,204 · Highest Rank 84

● 6 · ● 3 · ● 1

- Indoor Location ... · 4 months ago · Top 1% · 3rd of 1170
- University of Liv... · a year ago · Top 1% · 3rd of 2618
- Traveling Santa ... · 3 years ago · Top 1% · 8th of 1867

Datasets Contributor

Unranked

● 0 · ● 0 · ● 0

No dataset results

Notebooks Contributor

Unranked

● 0 · ● 0 · ● 0

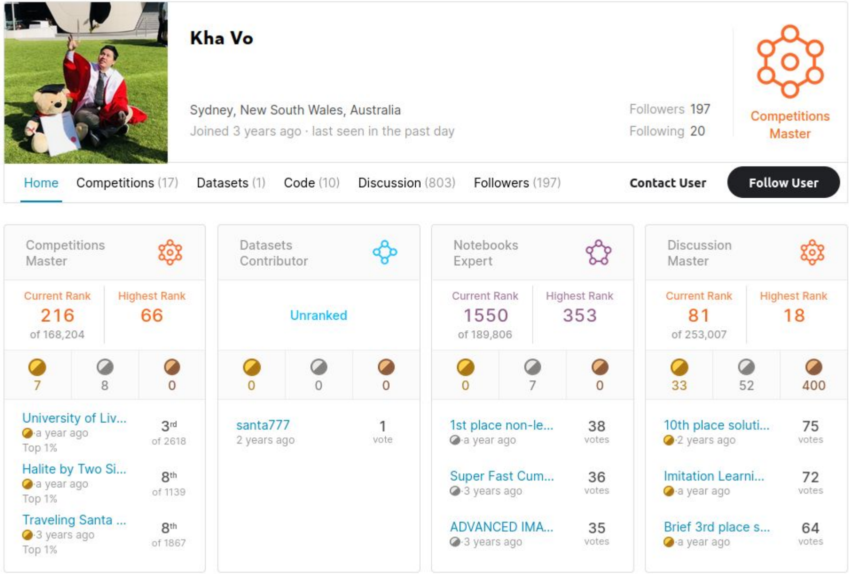
No notebook results

Discussion Expert

Current Rank 859 of 253,007 · Highest Rank 654

● 3 · ● 5 · ● 46

- The leak explain... · a year ago · 61 votes
- 3rd place solution · 4 months ago · 36 votes
- Best F1 score p... · a year ago · 14 votes



Kha Vo

Sydney, New South Wales, Australia

Joined 3 years ago · last seen in the past day

Followers 197

Following 20

Competitions Master

Home | Competitions (17) | Datasets (1) | Code (10) | Discussion (803) | Followers (197) | Contact User | Follow User

Competitions Master

Current Rank 216 of 168,204 · Highest Rank 66

● 7 · ● 8 · ● 0

- University of Liv... · a year ago · Top 1% · 3rd of 2618
- Halite by Two Si... · a year ago · Top 1% · 8th of 1139
- Traveling Santa ... · 3 years ago · Top 1% · 8th of 1867

Datasets Contributor

Unranked

● 0 · ● 0 · ● 0

- santa777 · 2 years ago · 1 vote

Notebooks Expert

Current Rank 1550 of 189,806 · Highest Rank 353

● 0 · ● 7 · ● 0

- 1st place non-le... · a year ago · 38 votes
- Super Fast Cum... · 3 years ago · 36 votes
- ADVANCED IMA... · 3 years ago · 35 votes

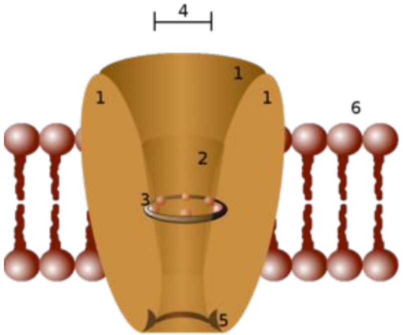
Discussion Master

Current Rank 81 of 253,007 · Highest Rank 18

● 33 · ● 52 · ● 400

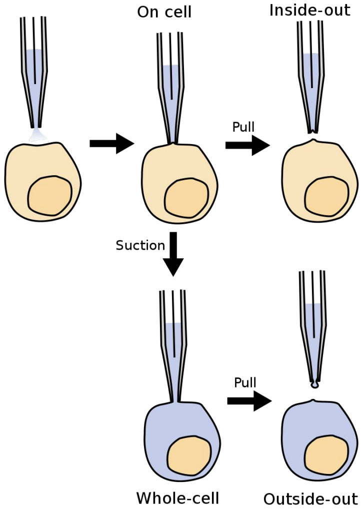
- 10th place soluti... · 2 years ago · 75 votes
- Imitation Learni... · a year ago · 72 votes
- Brief 3rd place s... · a year ago · 64 votes

Time for biology… Cells, Ion Channels & Patch Clamps

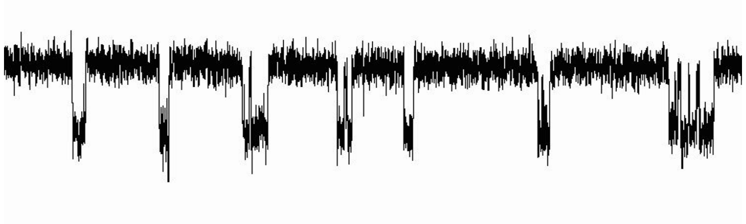


ion channel = gate that regulates flow of ions across cell membrane
→ encode learning and memory
→ help fight infections
→ enable pain signals
→ …

Studying how these ion channels behave within cells could have great impact on many areas of research.

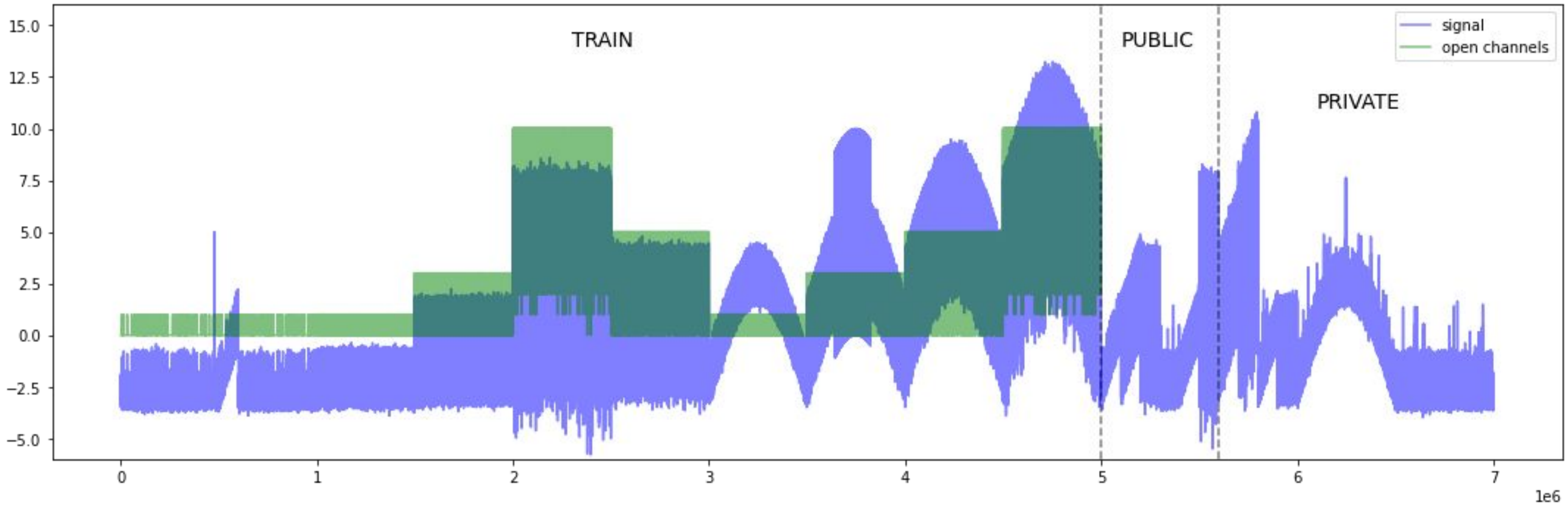


Patch clamp techniques allow us to study the behaviour of the ion channels by measuring electrical current.



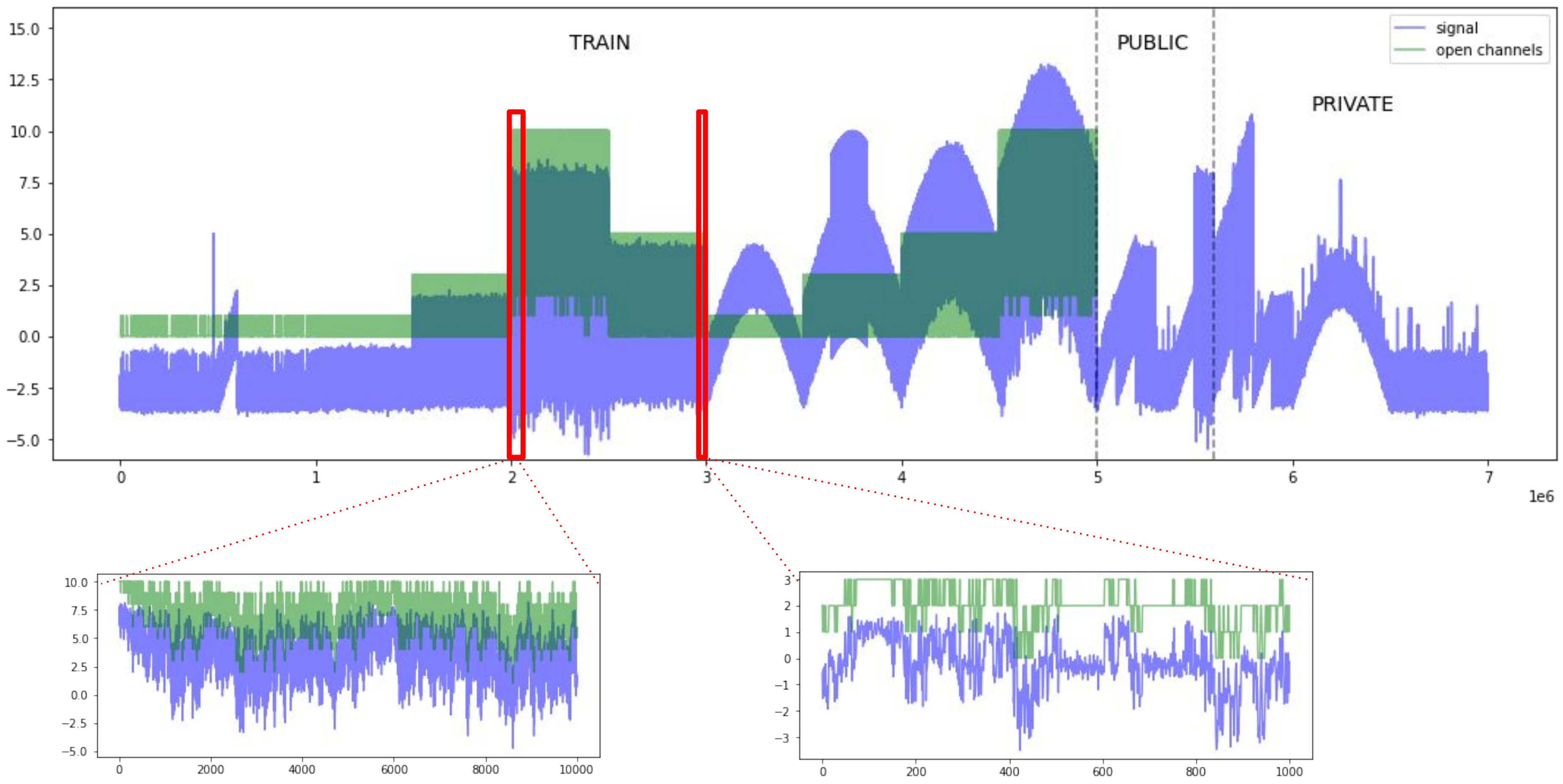
Analyzing this data manually is cumbersome and susceptible to human error & bias...

Three datasets: train, public test and private test

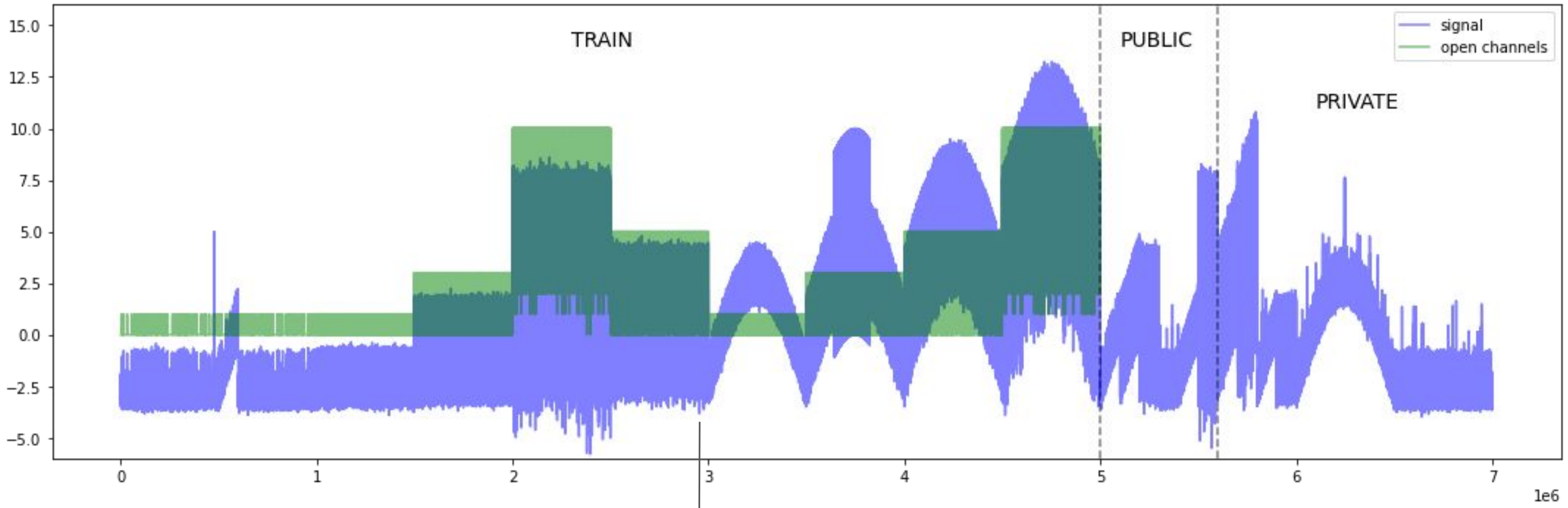


the data is from discrete batches of 50 seconds long sampled at 10 kHz

Three datasets: train, public test and private test



Goal: create model to predict open channels in public & private

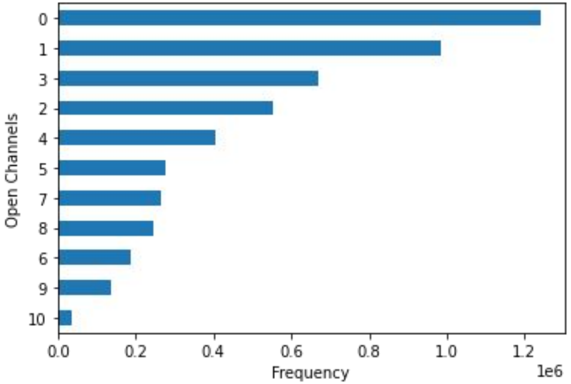


Number of open channels?

Competition metric: macro F_1 score

For each class, calculate its F_1 score and take the mean of class F_1 scores (equal weight for each class)

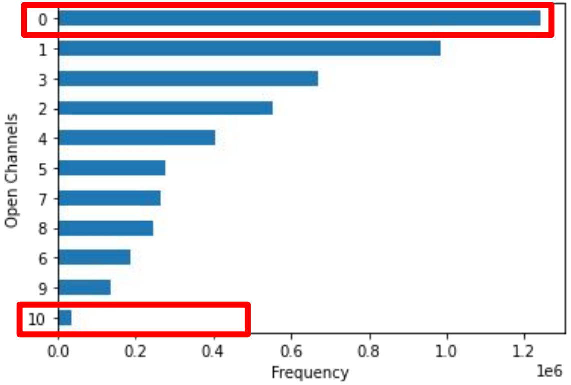
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



Competition metric: macro F_1 score

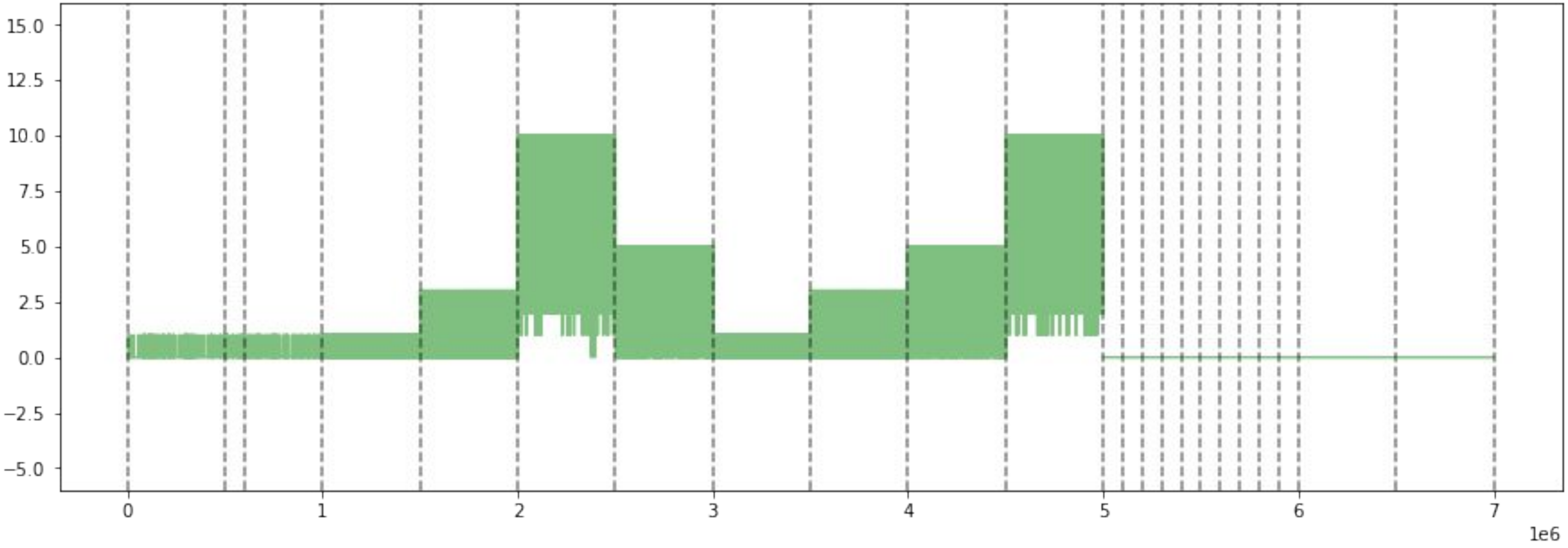
For each class, calculate its F_1 score and take the mean of class F_1 scores (equal weight for each class)

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

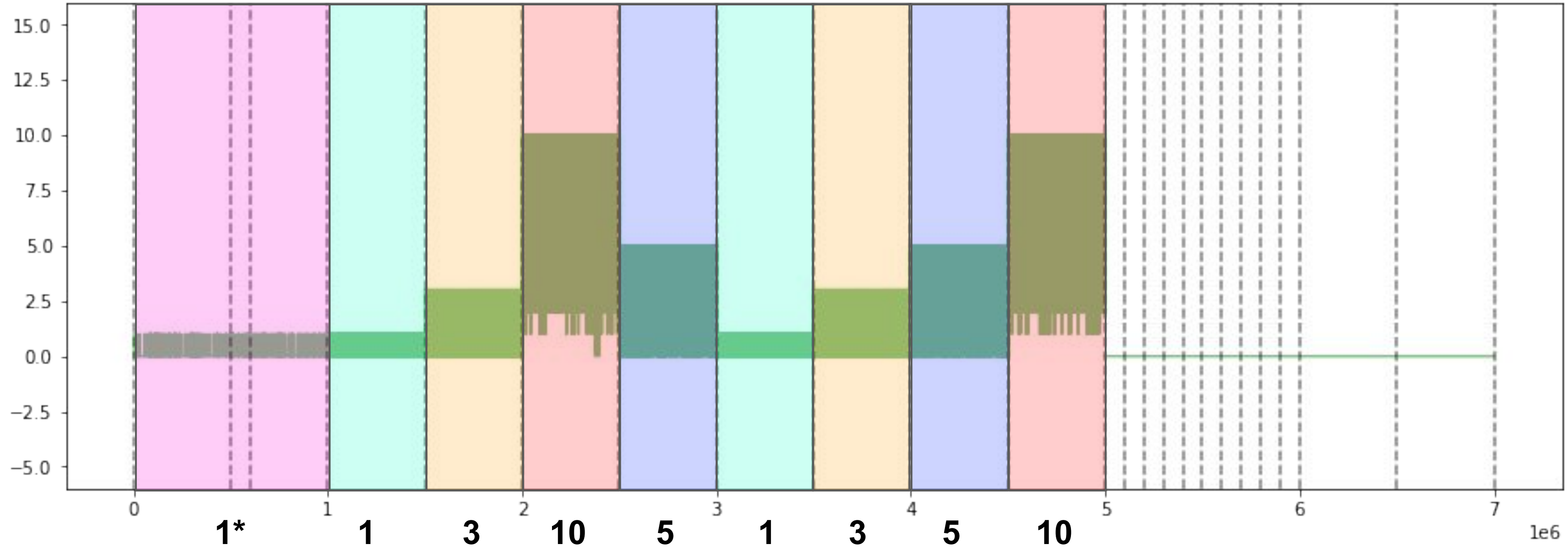


The small number of samples with open channels = 10 have the same impact on F1 score as the many samples with open channels = 0.

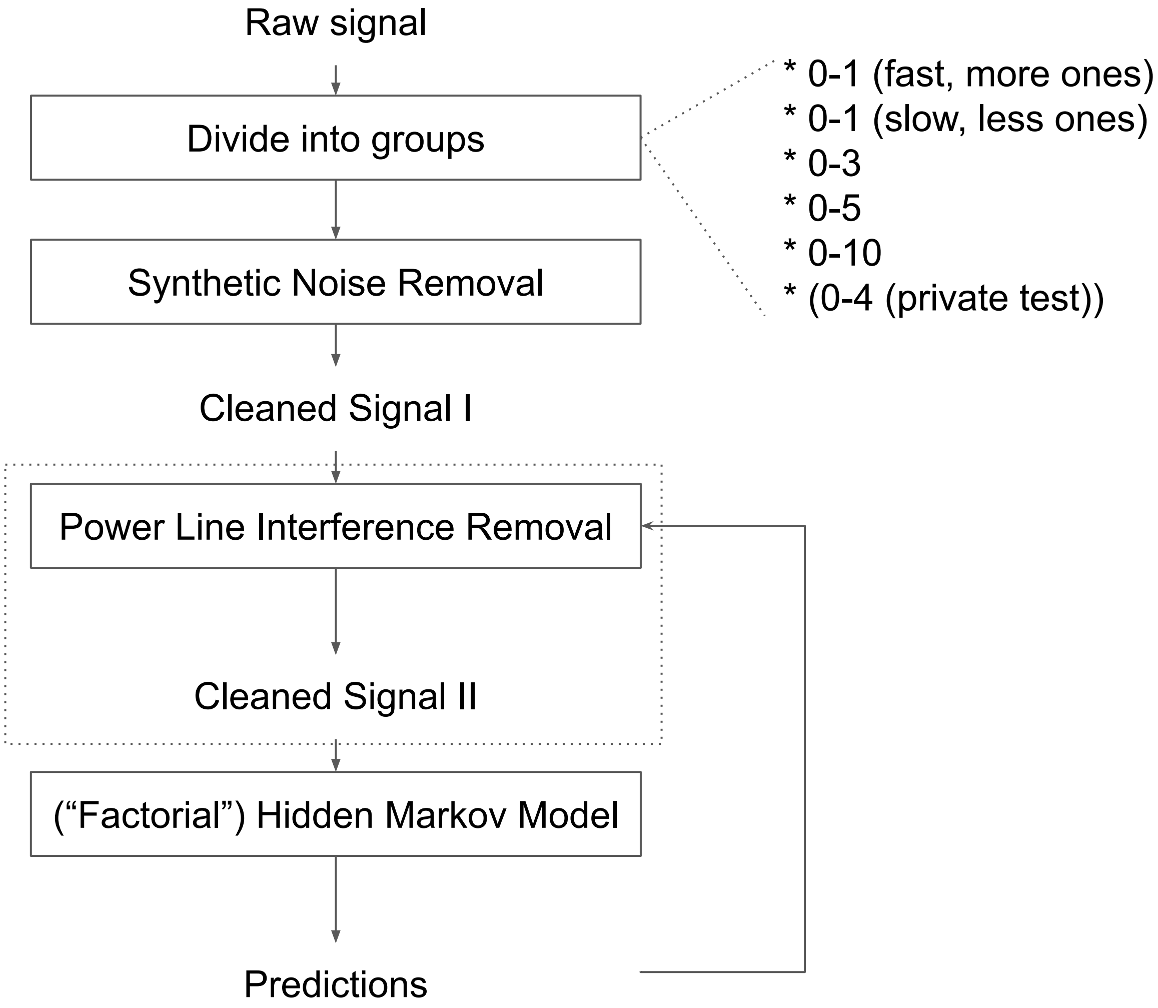
Identifying different groups in open channels…



Identifying different groups in open channels… **Max Value?**



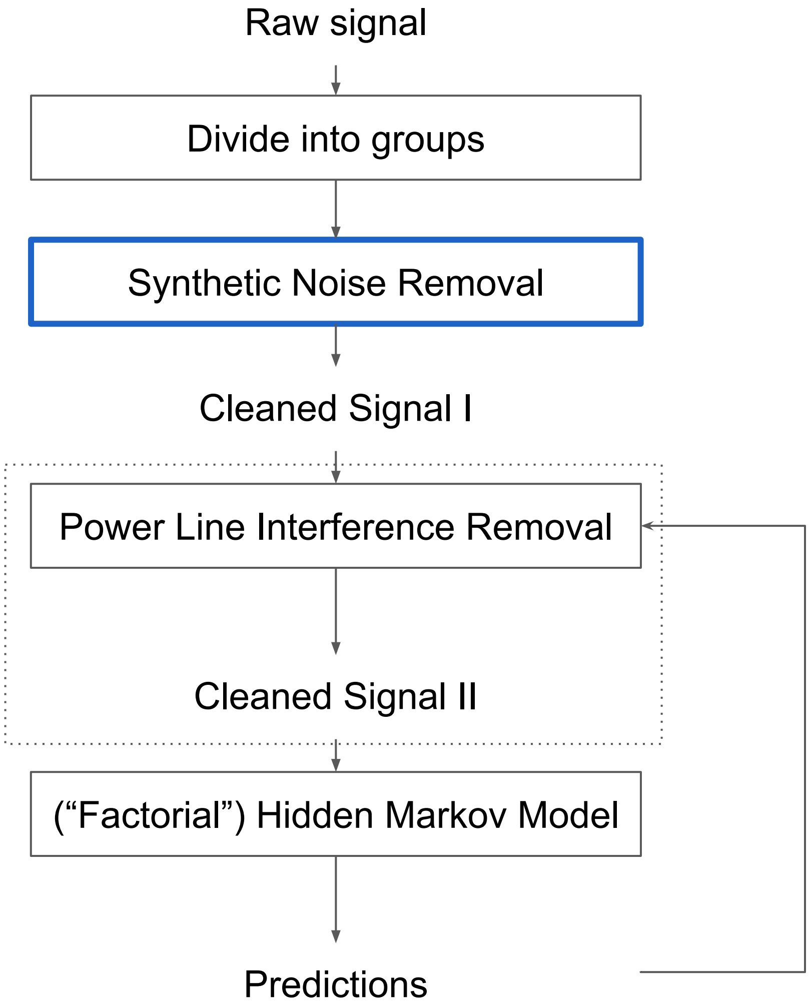
* Between the groups with a range of 0 to 1 open channels, there is one group (1*) with more zeroes.


Raw signal
Divide into groups
Synthetic Noise Removal
Cleaned Signal I
Power Line Interference Removal
Cleaned Signal II
("Factorial") Hidden Markov Model
Predictions
* 0-1 (fast, more ones)
* 0-1 (slow, less ones)
* 0-3
* 0-5
* 0-10
* (0-4 (private test))

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

Synthetic Noise Removal



Raw signal

Divide into groups

Synthetic Noise Removal

Cleaned Signal I

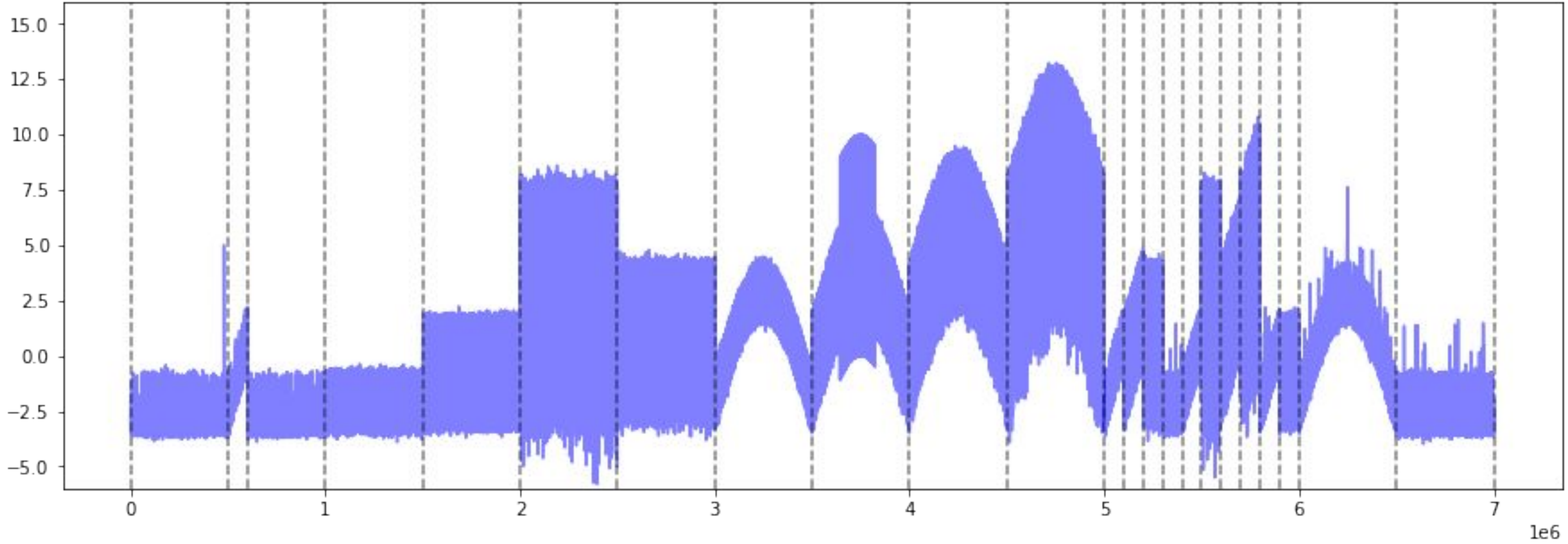
Power Line Interference Removal

Cleaned Signal II

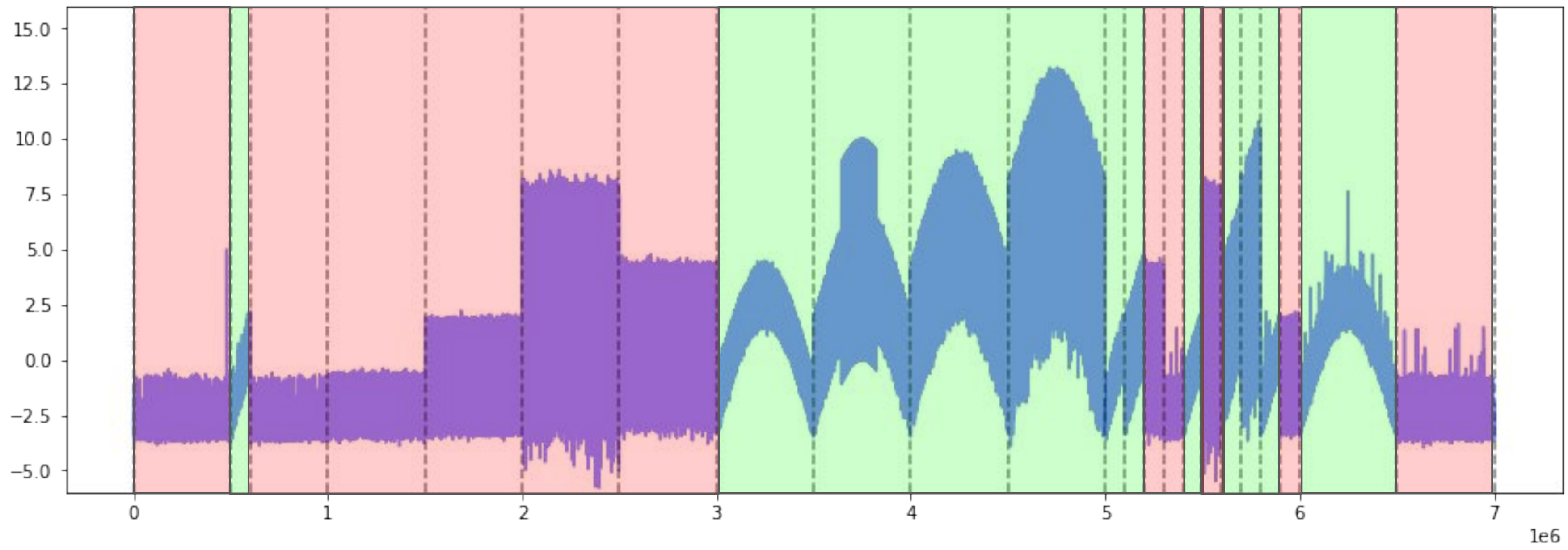
(“Factorial”) Hidden Markov Model

Predictions

Identifying different groups in signal...

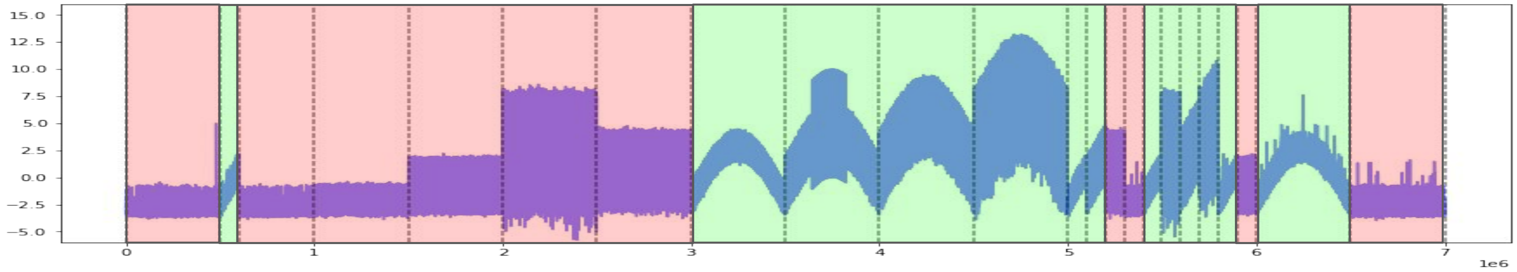


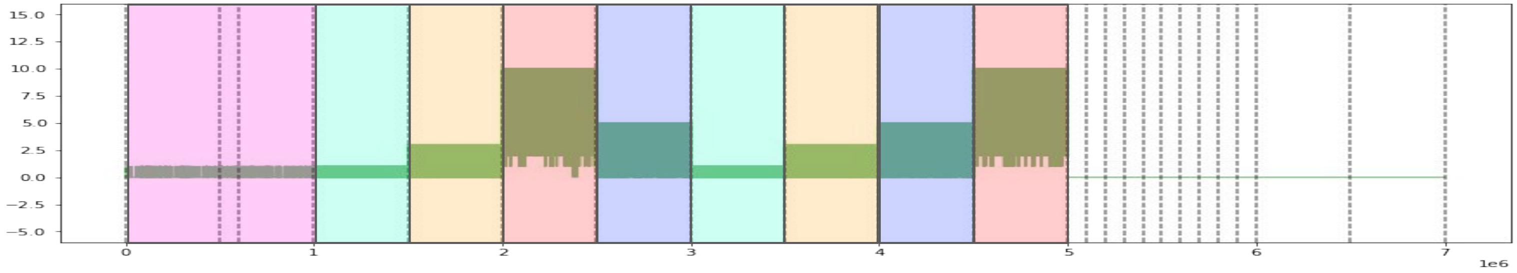
Identifying different groups in signal… **Weird Shape?***



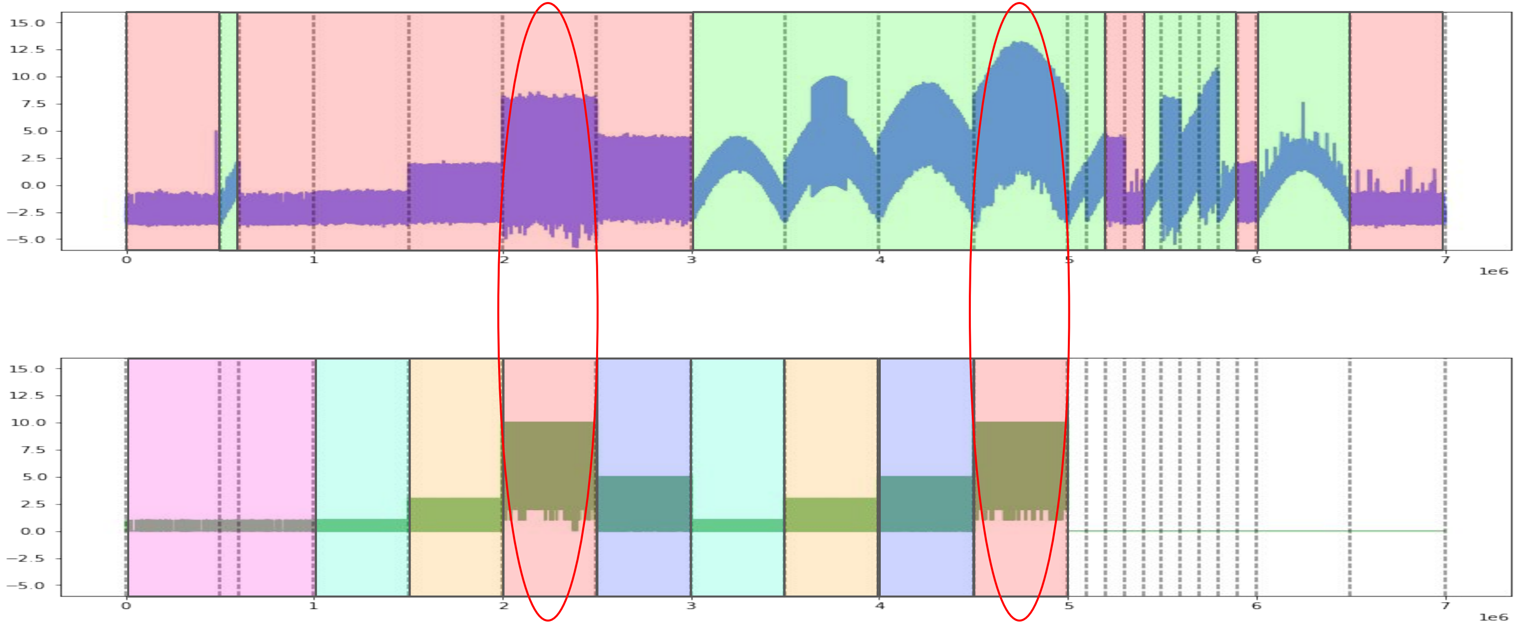
*** the signal values in some groups of data do not range from min-max across the entire range (more on this later)**
→ Continuously increasing trend?
→ Parabola?

Can we remove the “weird shape”?



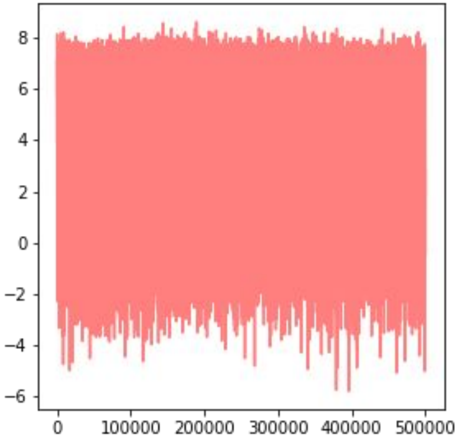


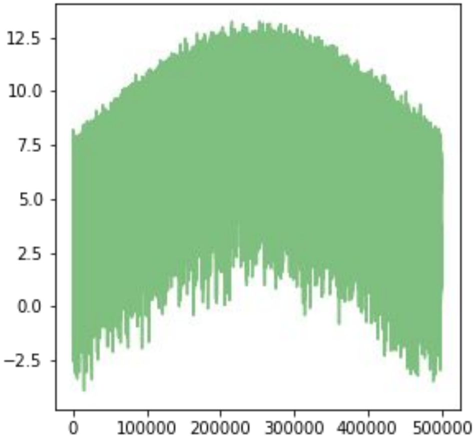
Can we remove the “weird shape”?



Each pair of two groups (based on open channels) has data <u>with</u> and <u>without</u> the weird shape!

Can we remove the “weird shape”?



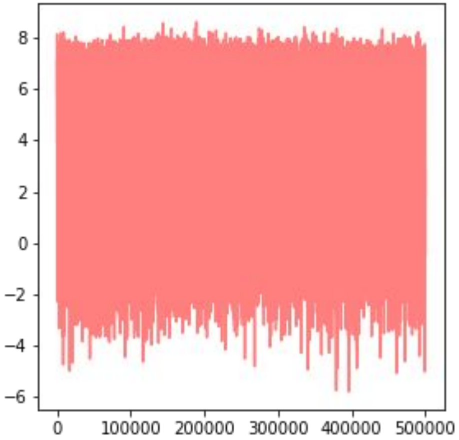


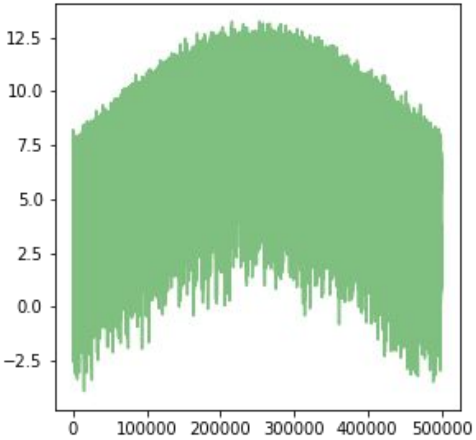
Can we remove the “weird shape”?

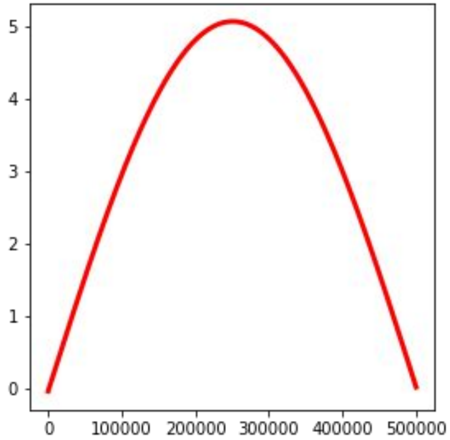
Fit sine function *A * sin(w * t + p)*

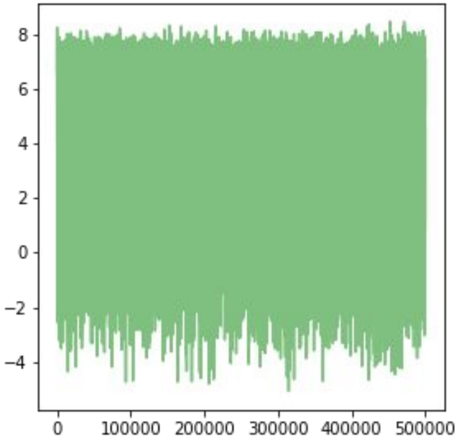
→ t = time

→ A = amplitude
→ w = frequency
→ p = phase

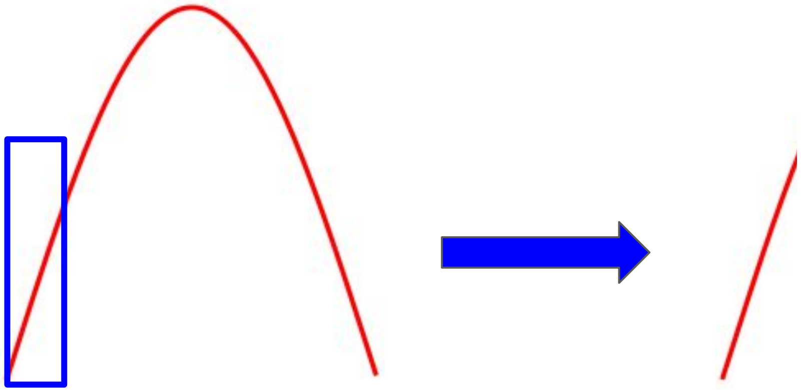






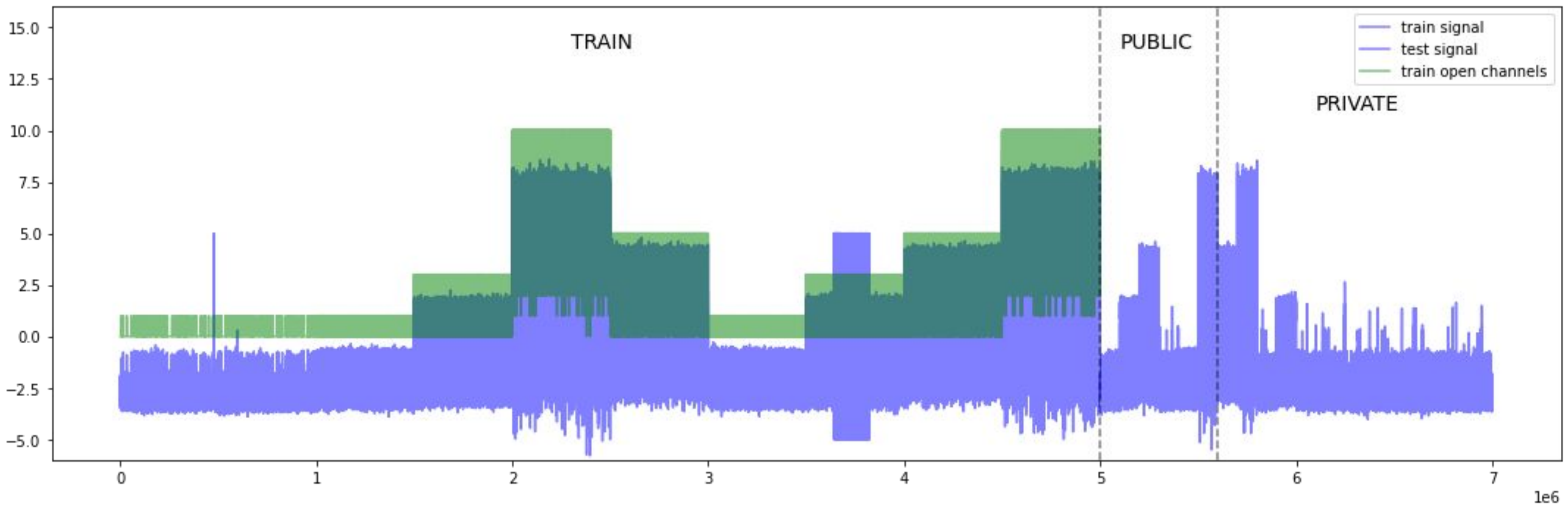


Can we remove the “weird shape”?

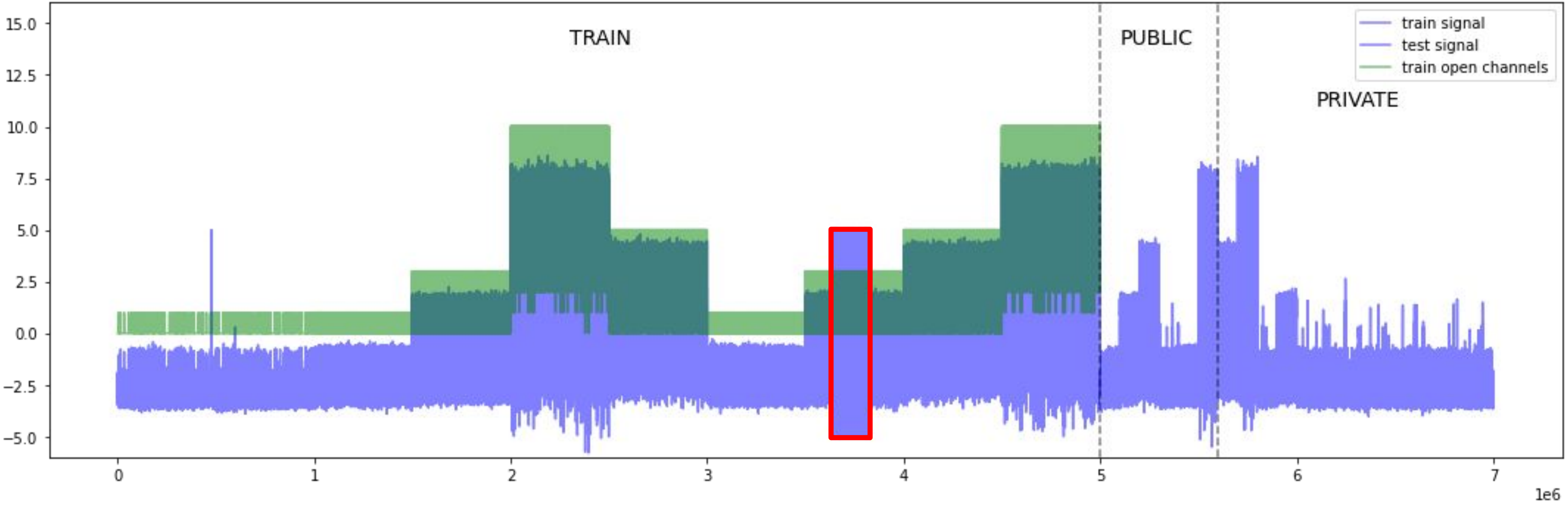


We use the sine function for the “linear” drift parts as well.

Result after cleaning data

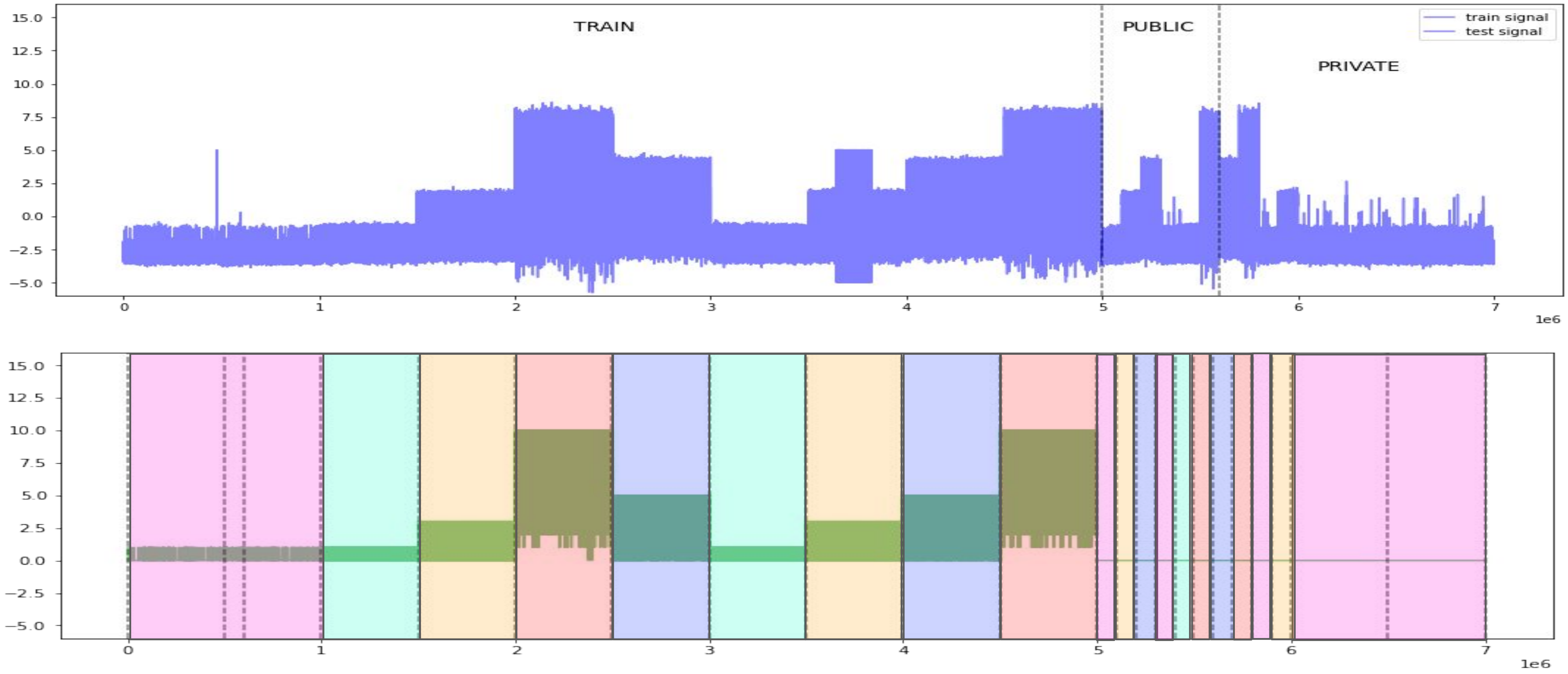


Result after cleaning data

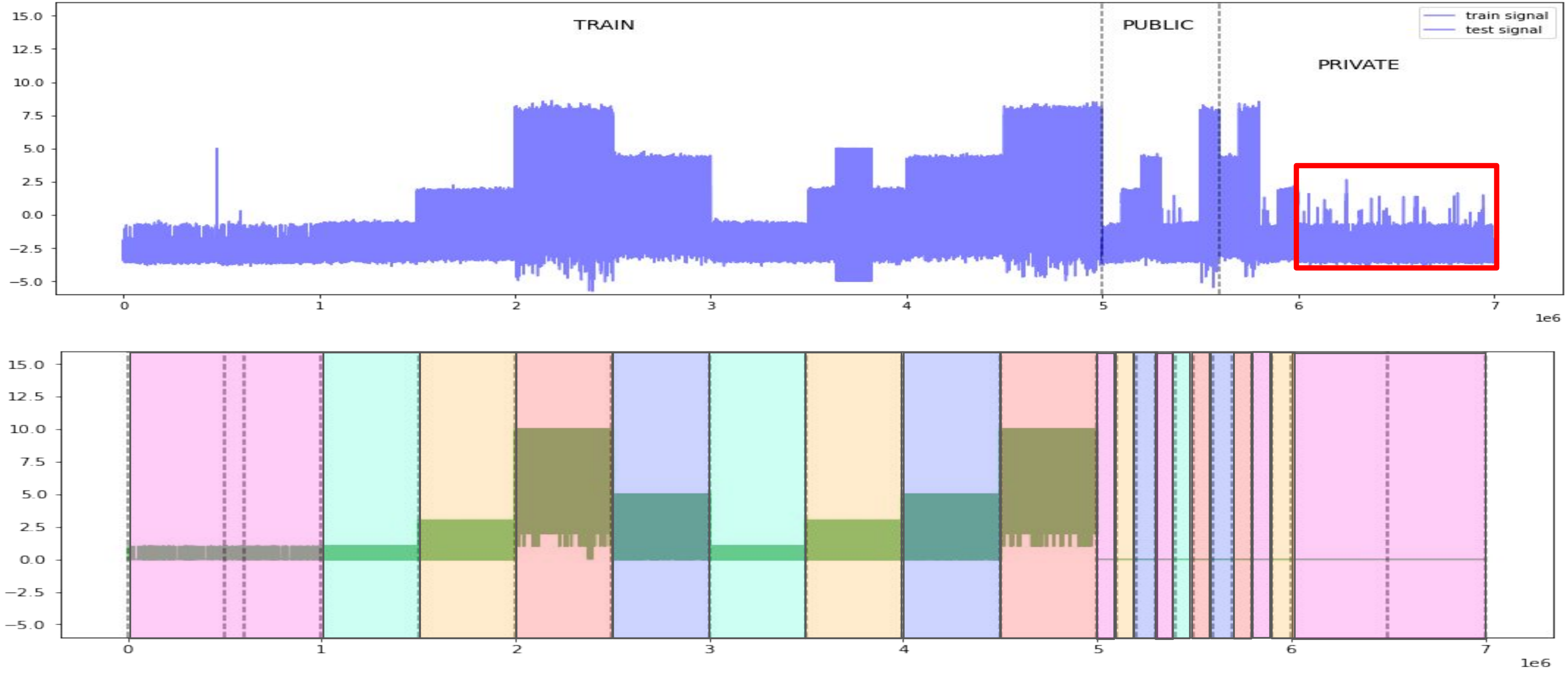


1 weird part remains → only in training data → ignore during modeling

Grouping train and test data



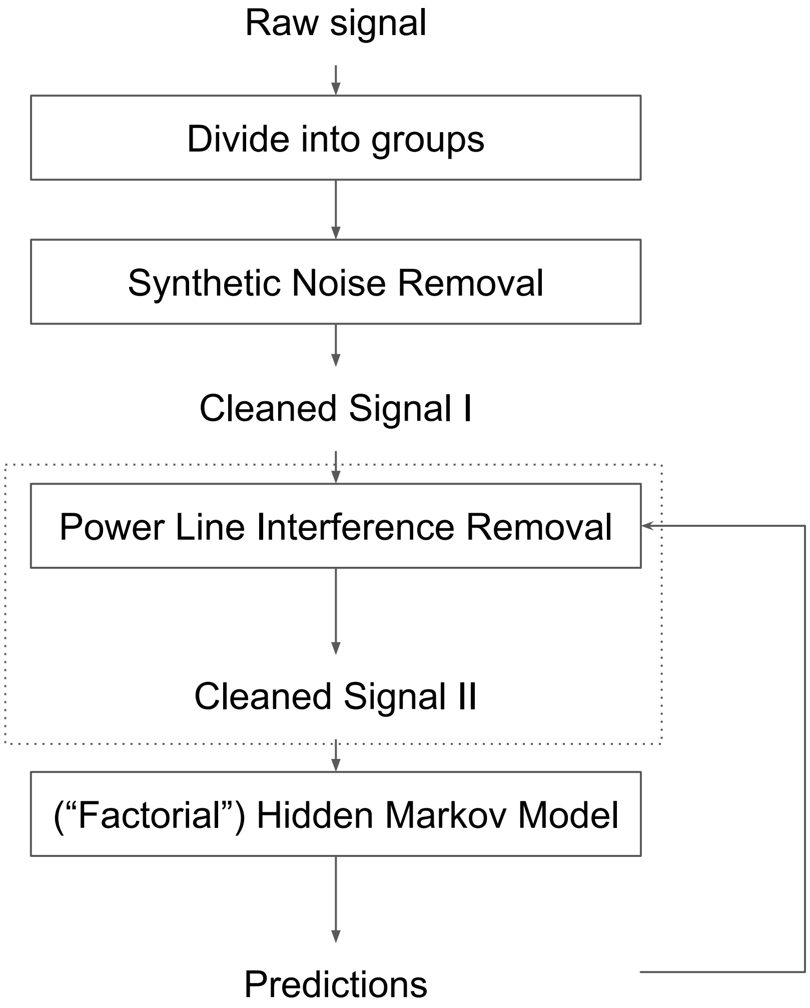
Grouping train and test data



Some of the 0/1 test data actually contains many spikes (2, 3 or 4 open channels), but impact on macro F1 will be minimal (so we will ignore for this presentation)

Liverpool - Ion Switching Competition

Simple Baseline: “Gaussian Mixture Model”



Raw signal

Divide into groups

Synthetic Noise Removal

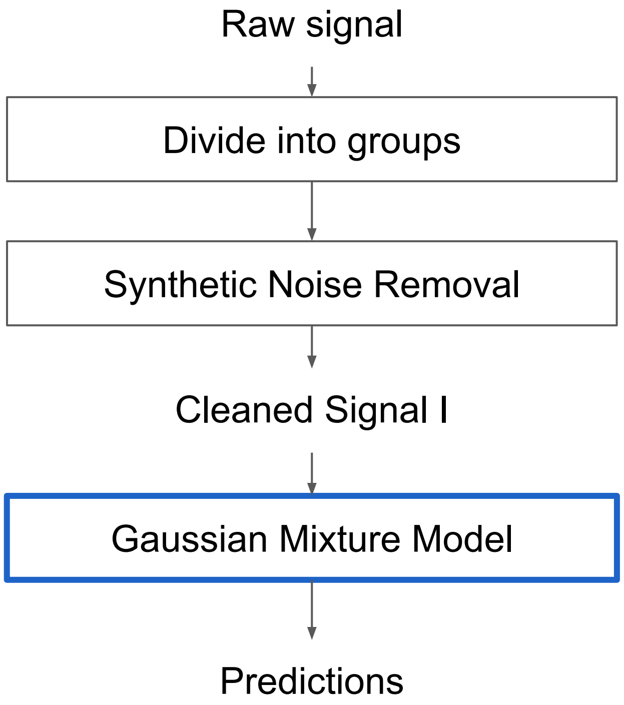
Cleaned Signal I

Power Line Interference Removal

Cleaned Signal II

("Factorial") Hidden Markov Model

Predictions



Raw signal

Divide into groups

Synthetic Noise Removal

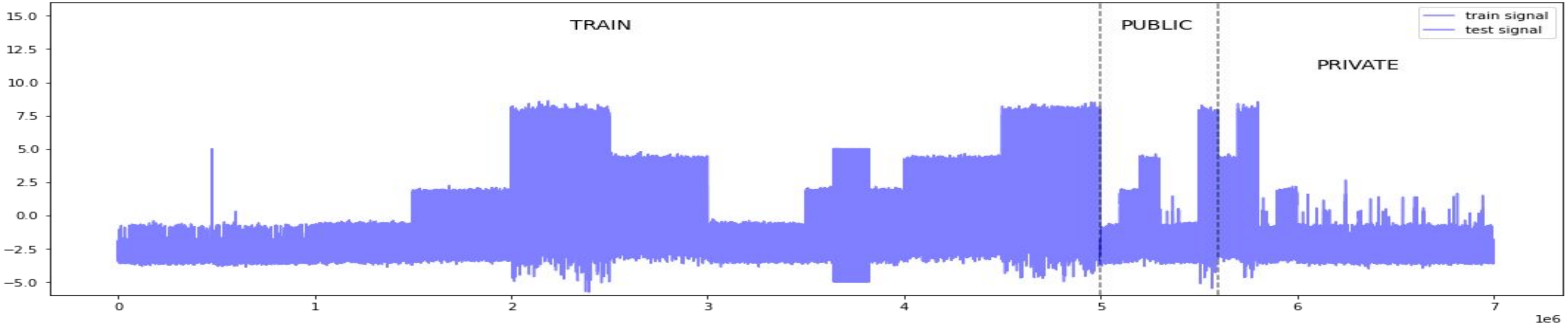
Cleaned Signal I

Gaussian Mixture Model

Predictions

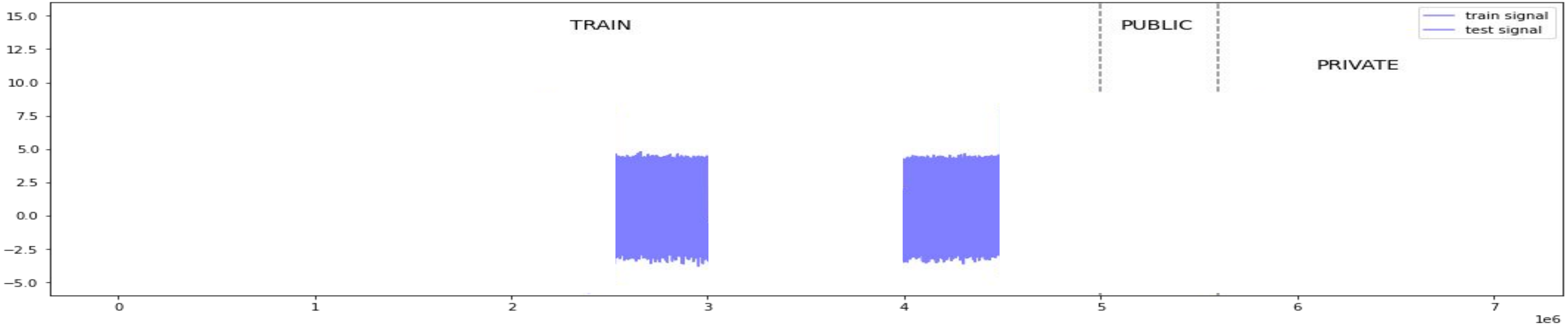
Baseline: “Gaussian mixture model” (per group of data)

1. Take a group of train data



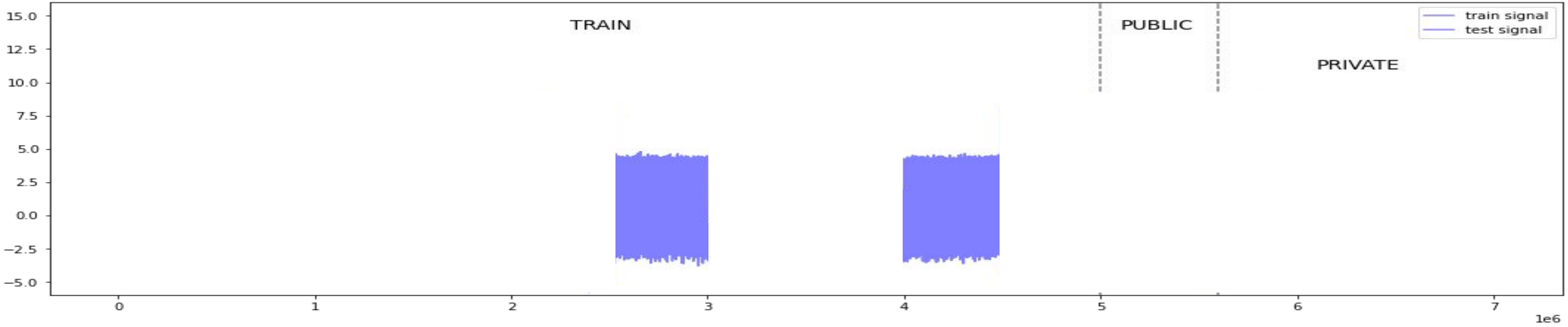
Baseline: “Gaussian mixture model” (per group of data)

1. Take a group of train data (here: 0-5 open channels)

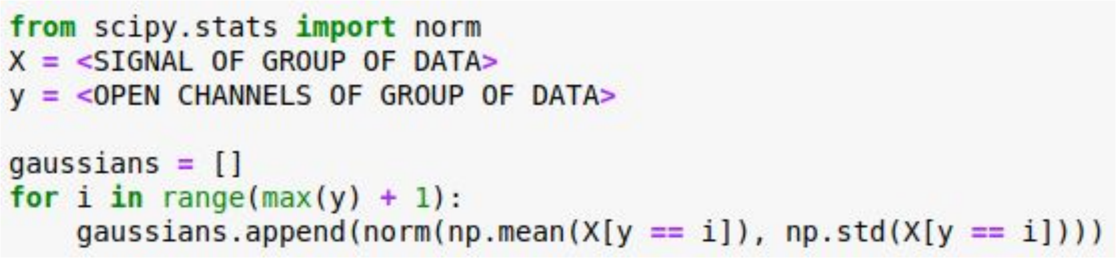


Baseline: “Gaussian mixture model” (per group of data)

1. Take a group of train data (here: 0-5 open channels)



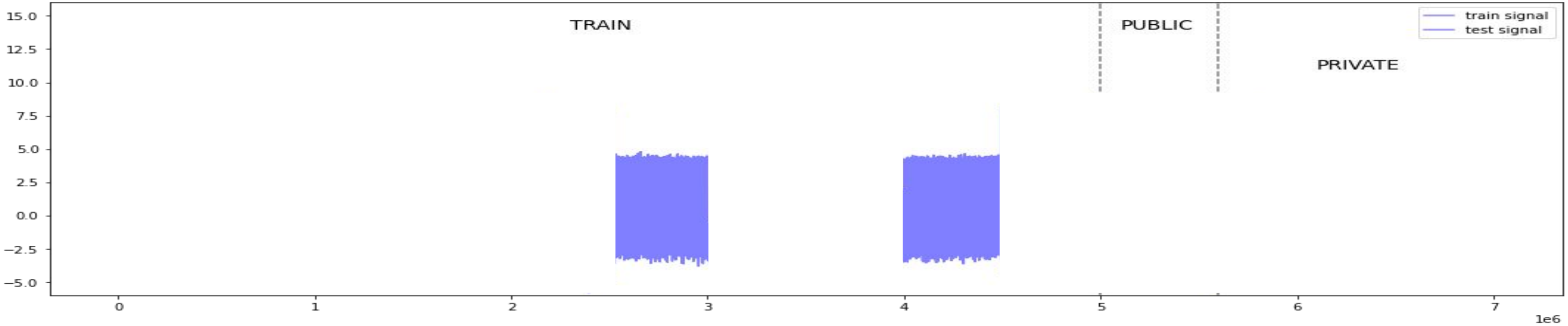
2. For each open channel value (0, … k), take all corresponding signal values & fit gaussian (calc mean and std)
→ N_0, …, N_k



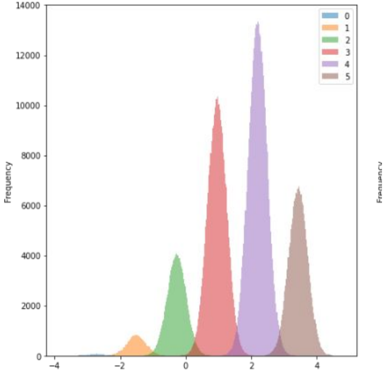
```
from scipy.stats import norm
X = <SIGNAL OF GROUP OF DATA>
y = <OPEN CHANNELS OF GROUP OF DATA>

gaussians = []
for i in range(max(y) + 1):
    gaussians.append(norm(np.mean(X[y == i]), np.std(X[y == i])))
```

Baseline: “Gaussian mixture model” (per group of data)

1. Take a group of train data (here: 0-5 open channels)



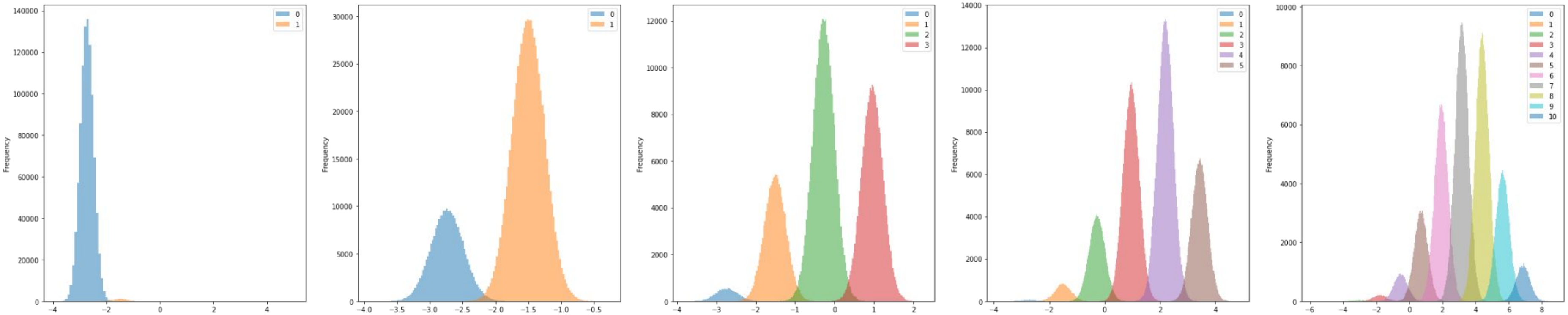
2. For each open channel value (0, … k), take all corresponding signal values & fit gaussian (calc mean and std)
→ N_0, …, N_k

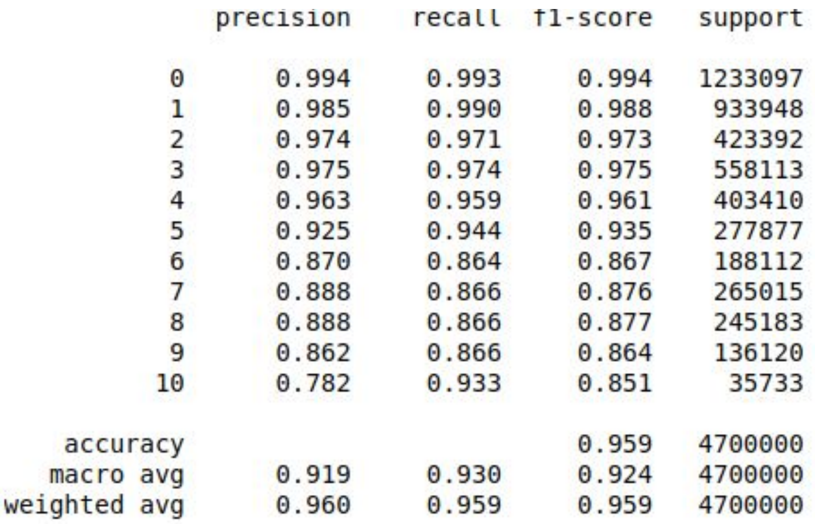


3. For a new signal value x:
Prediction corresponds to gaussian that gives us highest probability

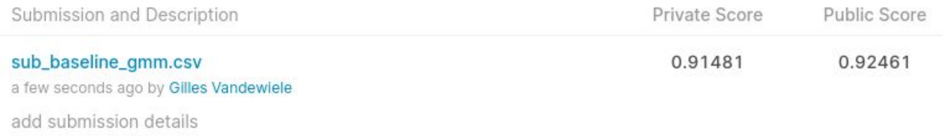
$$\underset{i=0,\ldots,k}{\operatorname{argmax}} f_{\mathcal{N}_i}(x)$$

Baseline: “Gaussian mixture model” (per group of data)





```
              precision    recall  f1-score   support

           0      0.994     0.993     0.994   1233097
           1      0.985     0.990     0.988    933948
           2      0.974     0.971     0.973    423392
           3      0.975     0.974     0.975    558113
           4      0.963     0.959     0.961    403410
           5      0.925     0.944     0.935    277877
           6      0.870     0.864     0.867    188112
           7      0.888     0.866     0.876    265015
           8      0.888     0.866     0.877    245183
           9      0.862     0.866     0.864    136120
          10      0.782     0.933     0.851     35733

    accuracy                          0.959   4700000
   macro avg      0.919     0.930     0.924   4700000
weighted avg      0.960     0.959     0.959   4700000
```

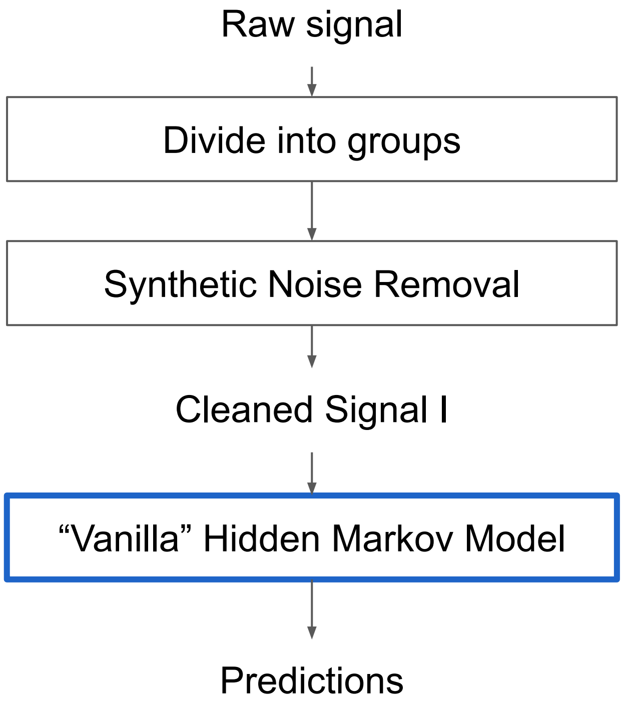


Submission and Description	Private Score	Public Score
sub_baseline_gmm.csv a few seconds ago by Gilles Vandewiele add submission details	0.91481	0.92461

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

Hidden Markov Models



Raw signal

Divide into groups

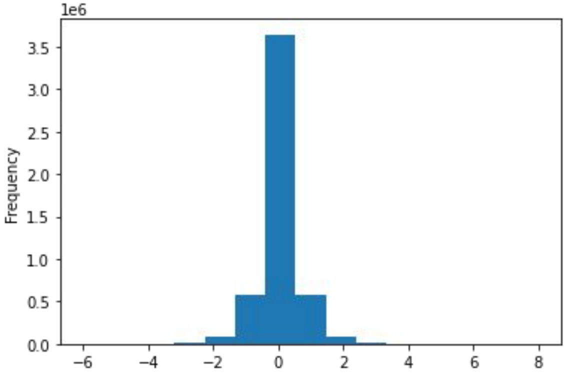
Synthetic Noise Removal

Cleaned Signal I

“Vanilla” Hidden Markov Model

Predictions

Hidden Markov Models: intuition

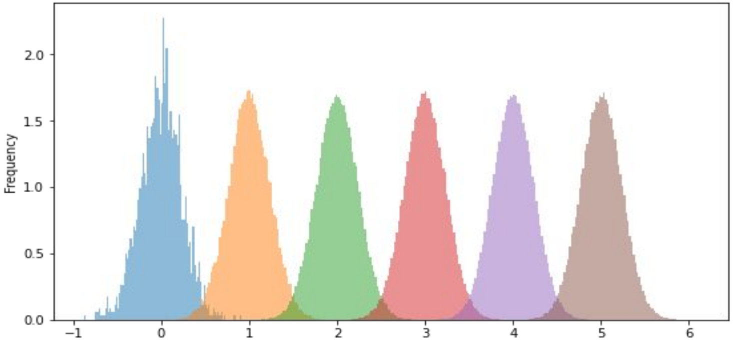


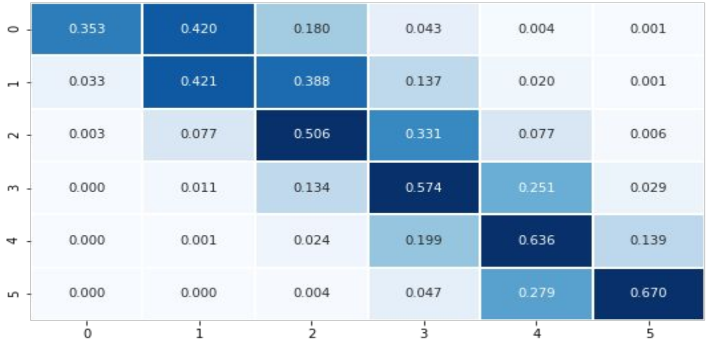
Histogram of y_{t+1} - y_t

- **Open channel values are temporally correlated.**
- **Going from 0 open channels at time *t* to 10 open channels at time *t + 1* is very unlikely.**
- **<u>Markov property</u>: we'll assume time *t + 1* only depends on 1 previous timestep *t***

Hidden Markov Models: theory

https://web.stanford.edu/~jurafsky/slp3/A.pdf





	0	1	2	3	4	5
0	0.353	0.420	0.180	0.043	0.004	0.001
1	0.033	0.421	0.388	0.137	0.020	0.001
2	0.003	0.077	0.506	0.331	0.077	0.006
3	0.000	0.011	0.134	0.574	0.251	0.029
4	0.000	0.001	0.024	0.199	0.636	0.139
5	0.000	0.000	0.004	0.047	0.279	0.670

emission probability: what is the probability that class is *i* if we observe *x* ($P(y_t = i \mid x_t)$)

→ our baseline!

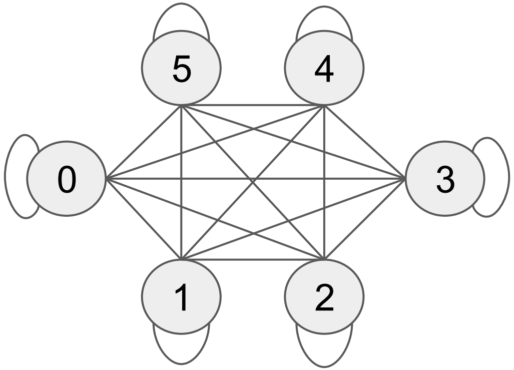
transition probabilities: what is the probability that class is *j* if prev. class was *i* ($P(y_{t+1} = j \mid y_t = i)$)

$$P(y_{t+1} = j) = \sum_{i=0}^{k} P(y_t = i) * P(y_{t+1} = j|y_t = i) * P(y_{t+1} = j|x)$$

prev. step **transition** **emission**

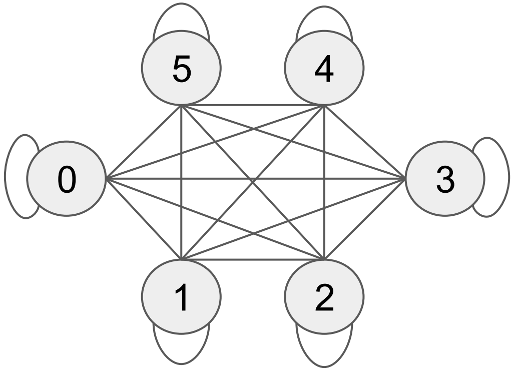
Vanilla HMM

*edges are directional!

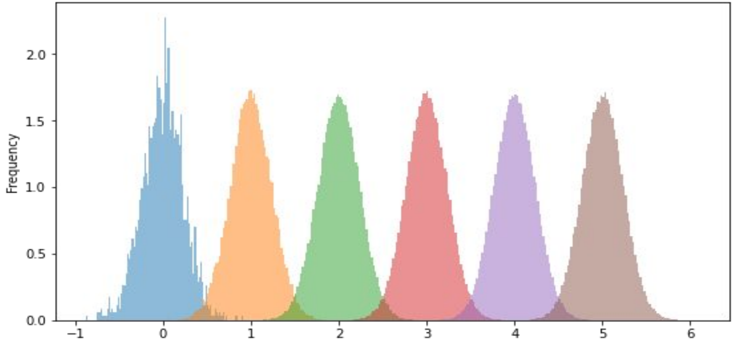


hidden states (Markov Chain)

Vanilla HMM



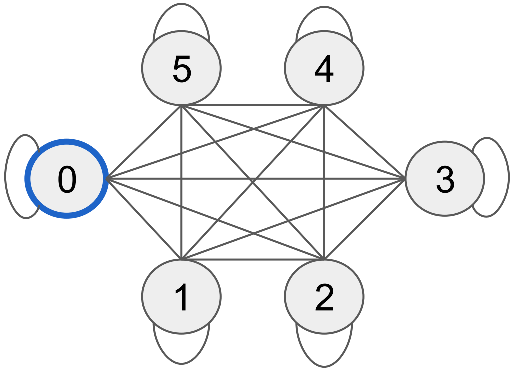
hidden states (Markov Chain)



observable data (Gaussians)

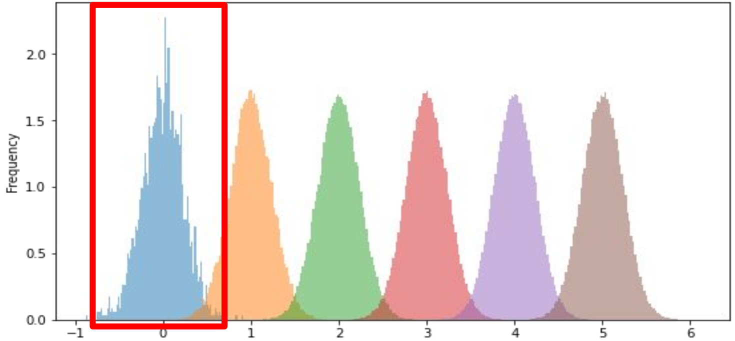
Vanilla HMM

t = 0



hidden states (Markov Chain)

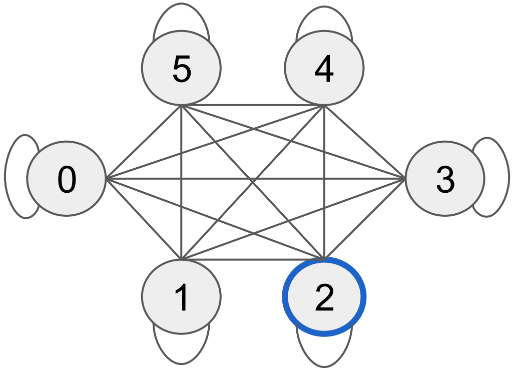
x ~ N(0, 0.15)



observable data (Gaussians)

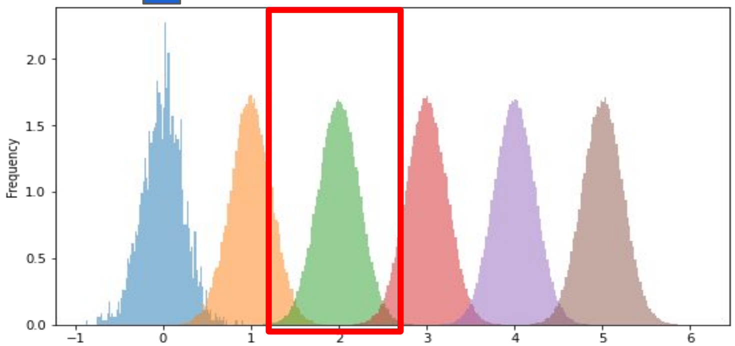
Vanilla HMM

t = 1



hidden states (Markov Chain)

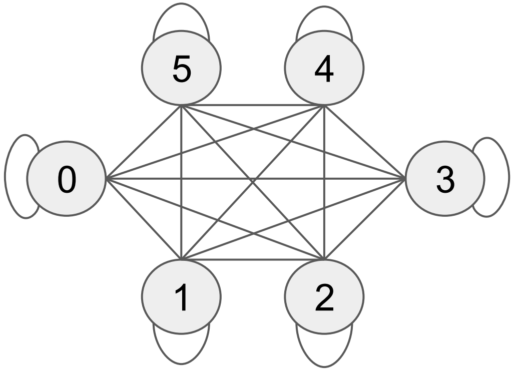
x ~ N(2, 0.15)



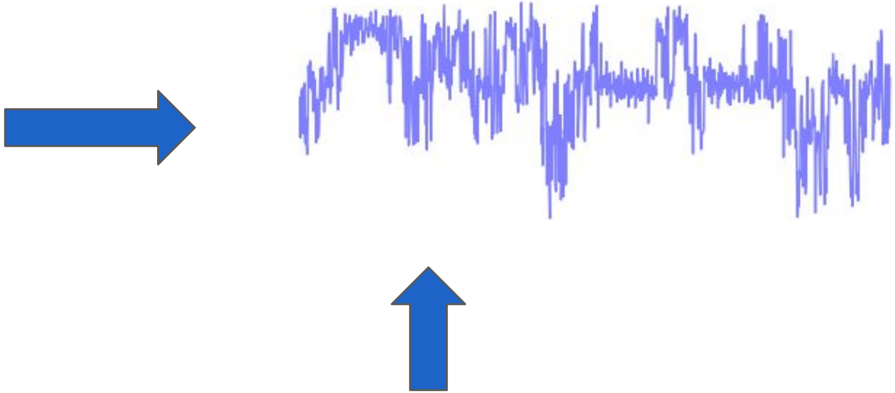
observable data (Gaussians)

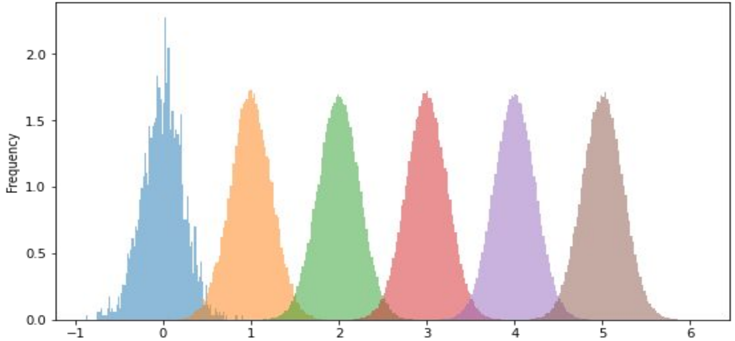
Vanilla HMM

t = N



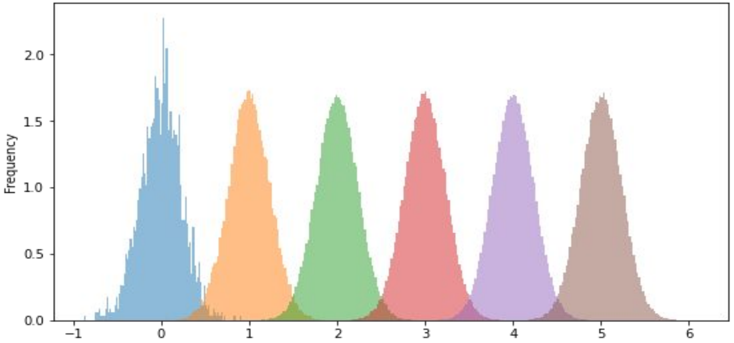
hidden states (Markov Chain)



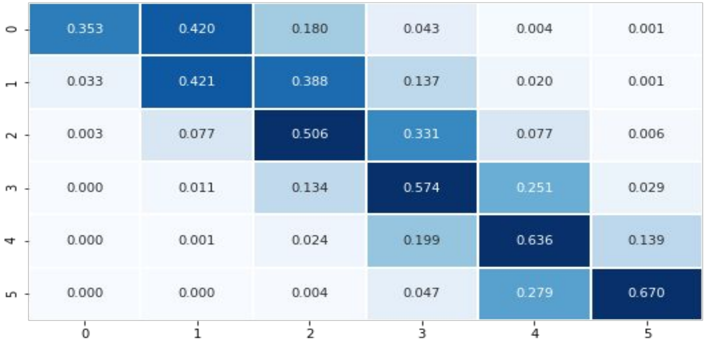


observable data (Gaussians)

Vanilla HMM

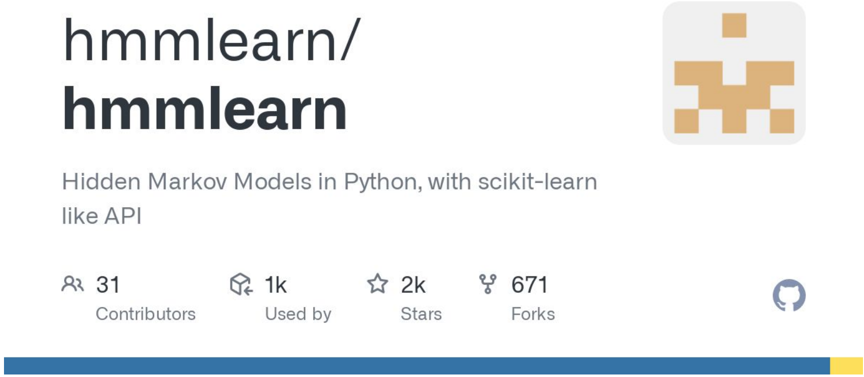


emission



	0	1	2	3	4	5
0	0.353	0.420	0.180	0.043	0.004	0.001
1	0.033	0.421	0.388	0.137	0.020	0.001
2	0.003	0.077	0.506	0.331	0.077	0.006
3	0.000	0.011	0.134	0.574	0.251	0.029
4	0.000	0.001	0.024	0.199	0.636	0.139
5	0.000	0.000	0.004	0.047	0.279	0.670

transition



hmmlearn/
hmmlearn

Hidden Markov Models in Python, with scikit-learn like API

31 Contributors

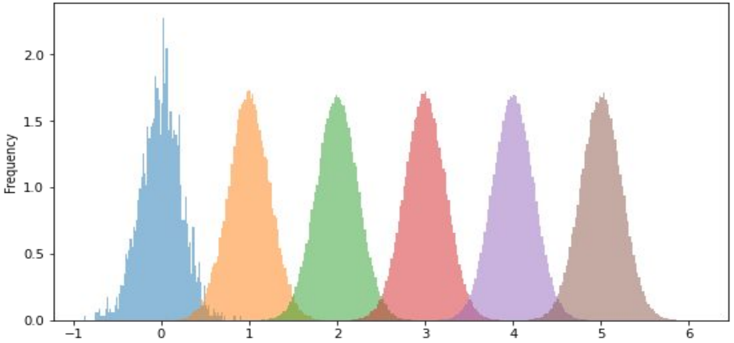
1k Used by

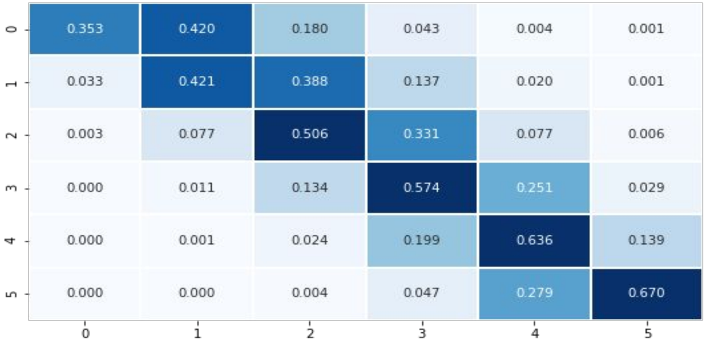
2k Stars

671 Forks

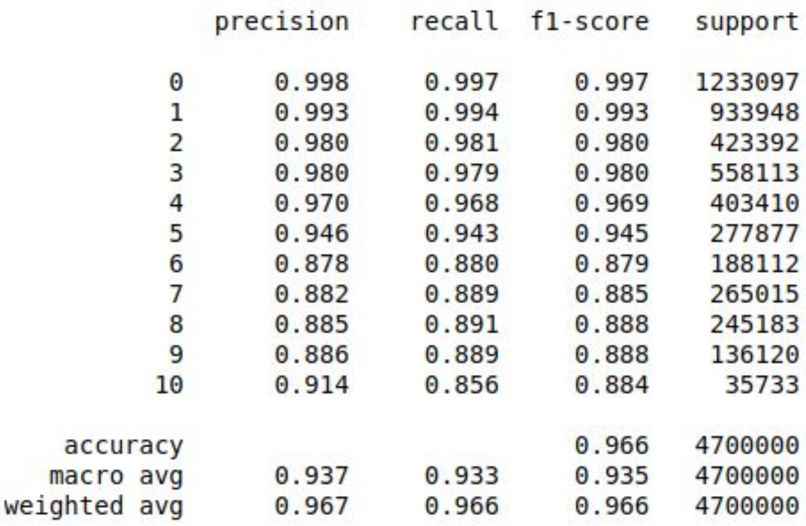
We use *k* + *1* hidden states for the HMM with *k* the maximum number of open channels.
→ 1 hidden state per open channel

Vanilla HMM

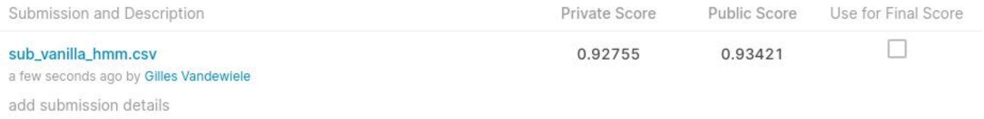




	0	1	2	3	4	5
0	0.353	0.420	0.180	0.043	0.004	0.001
1	0.033	0.421	0.388	0.137	0.020	0.001
2	0.003	0.077	0.506	0.331	0.077	0.006
3	0.000	0.011	0.134	0.574	0.251	0.029
4	0.000	0.001	0.024	0.199	0.636	0.139
5	0.000	0.000	0.004	0.047	0.279	0.670



```
              precision    recall  f1-score   support

           0      0.998     0.997     0.997   1233097
           1      0.993     0.994     0.993    933948
           2      0.980     0.981     0.980    423392
           3      0.980     0.979     0.980    558113
           4      0.970     0.968     0.969    403410
           5      0.946     0.943     0.945    277877
           6      0.878     0.880     0.879    188112
           7      0.882     0.889     0.885    265015
           8      0.885     0.891     0.888    245183
           9      0.886     0.889     0.888    136120
          10      0.914     0.856     0.884     35733

    accuracy                          0.966   4700000
   macro avg      0.937     0.933     0.935   4700000
weighted avg      0.967     0.966     0.966   4700000
```

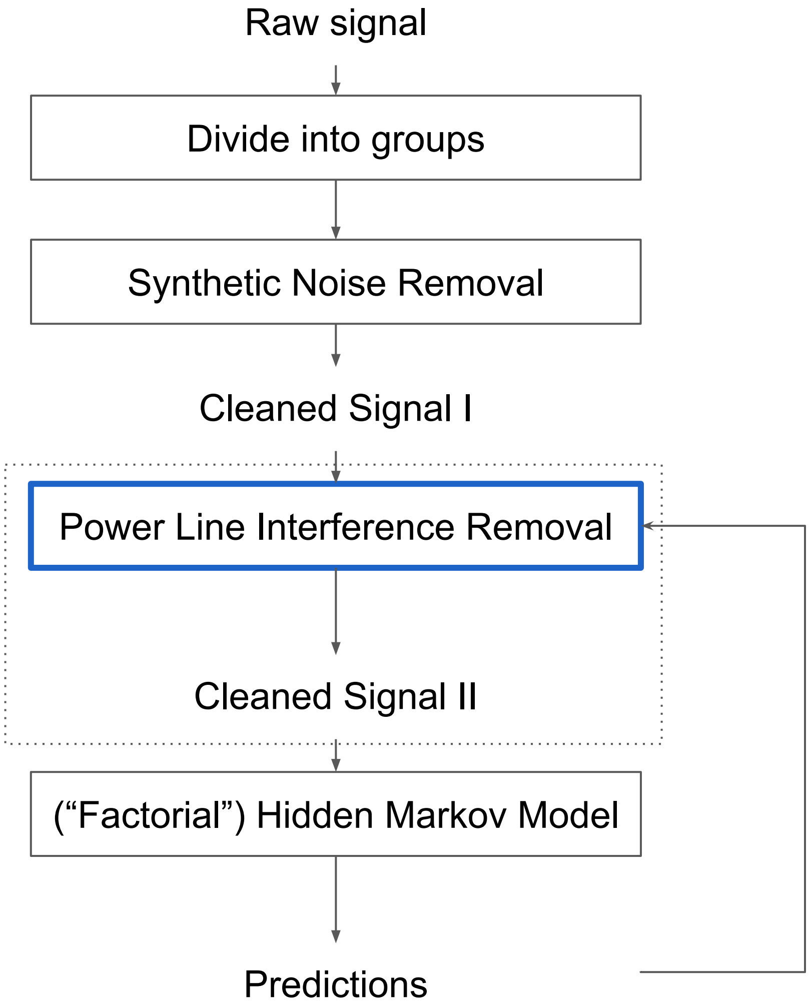


Submission and Description	Private Score	Public Score	Use for Final Score
sub_vanilla_hmm.csv a few seconds ago by Gilles Vandewiele add submission details	0.92755	0.93421	□

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

Power Line Interference Removal



Raw signal

Divide into groups

Synthetic Noise Removal

Cleaned Signal I

Power Line Interference Removal

Cleaned Signal II

("Factorial") Hidden Markov Model

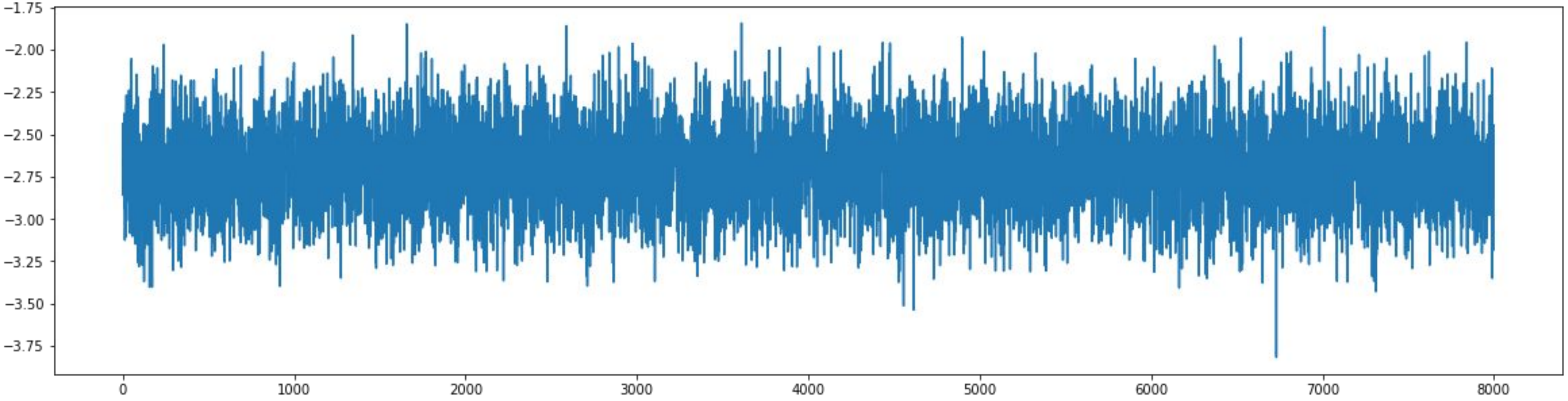
Predictions

Did we already remove all noise?

$$x_t = f(y_t) + e$$

→ **The observed signal values are a function of the ground truth with added noise (e).**

→ **Isolate *e* by calculating x_t - $f(y_t)$ with $f(y_t)$ the predictions of our strongest model.**

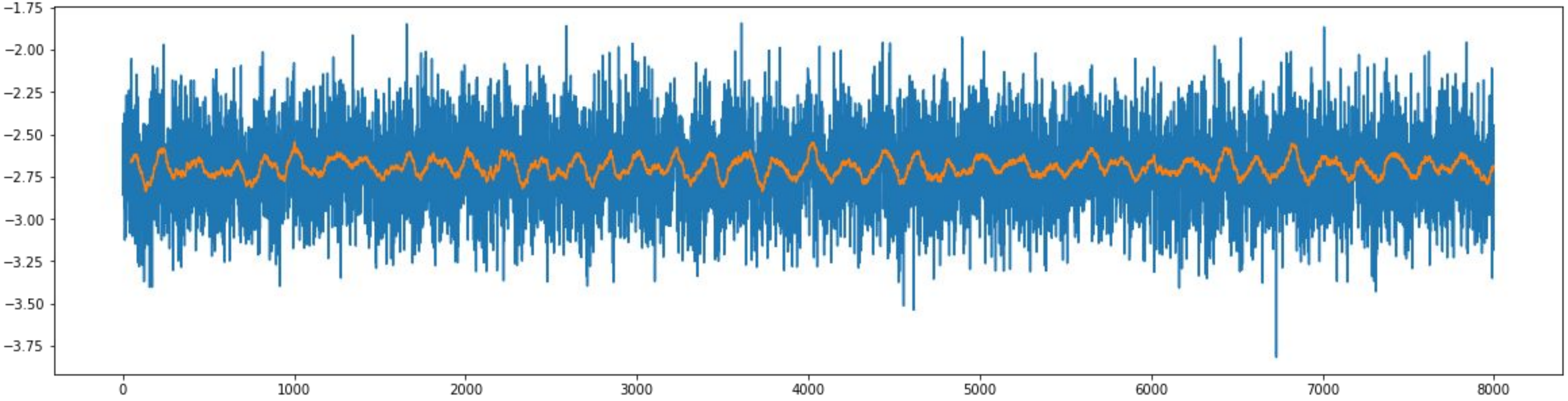


Did we already remove all noise?

$$x_t = f(y_t) + e$$

→ **The observed signal values are a function of the ground truth with added noise (e).**

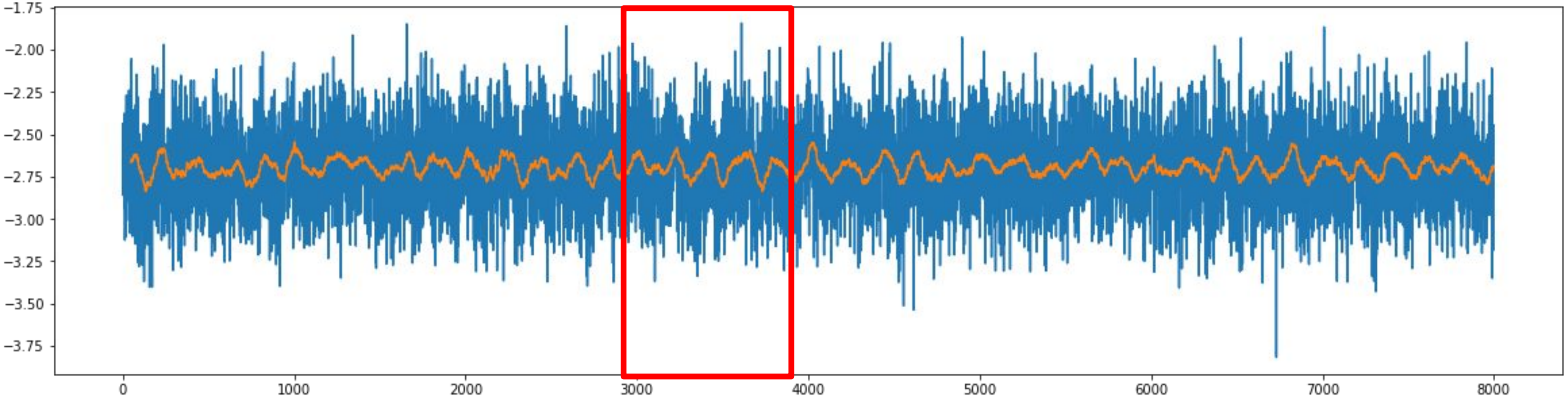
→ **Isolate *e* by calculating x_t - $f(y_t)$ with $f(y_t)$ the predictions of our strongest model.**
→ **Take rolling avg (window size = 50)**



Did we already remove all noise?

$$x_t = f(y_t) + e$$

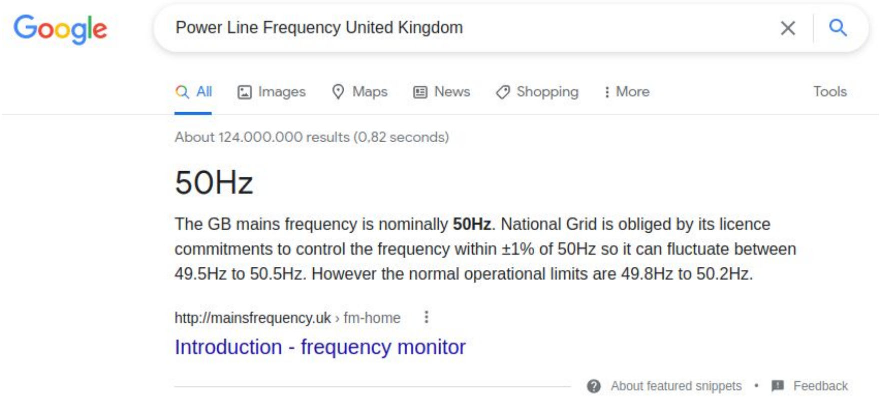
5 peaks per 1000 values
→ periodicity = 200 values
→ sampling rate = 10 kHz
→ frequency of this pattern = 50 Hz



Power line interference!

$$x_t = f(y_t) + e$$

5 peaks per 1000 values
→ periodicity = 200 values
→ sampling rate = 10 kHz
→ frequency of this pattern = 50 Hz



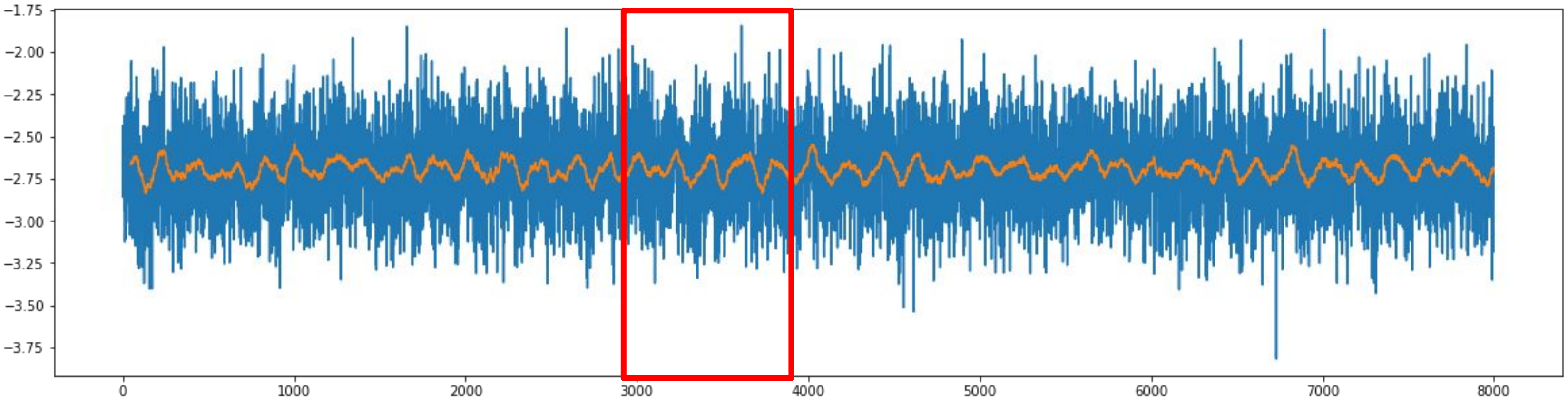
About 124.000.000 results (0,82 seconds)

50Hz

The GB mains frequency is nominally **50Hz**. National Grid is obliged by its licence commitments to control the frequency within ±1% of 50Hz so it can fluctuate between 49.5Hz to 50.5Hz. However the normal operational limits are 49.8Hz to 50.2Hz.

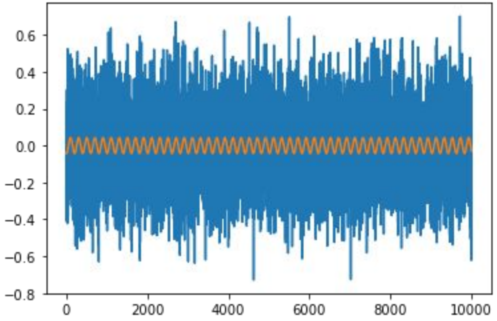
http://mainsfrequency.uk › fm-home

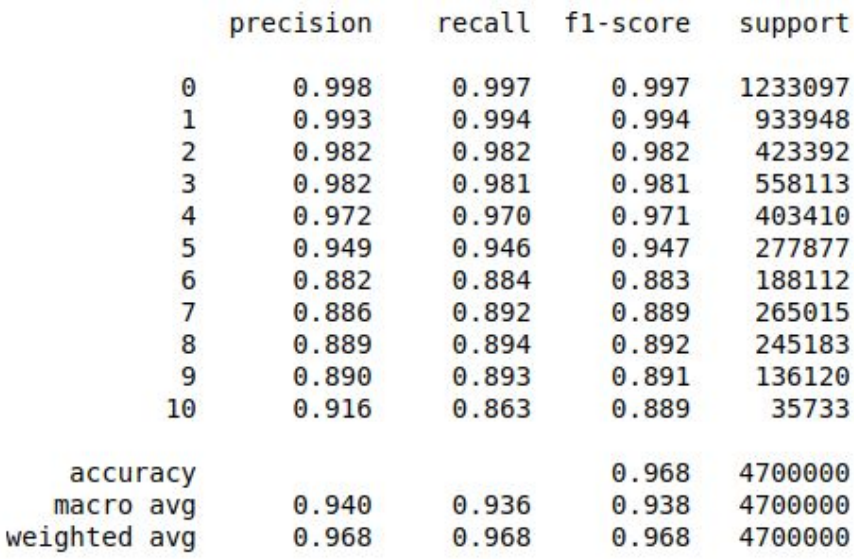
Introduction - frequency monitor



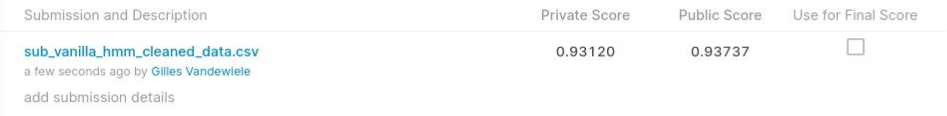
Impact of power line interference removal

→ **Reuse our sine fitter to remove power line interference**
→ **Re-fit our "vanilla" HMM**





```
              precision    recall  f1-score   support

           0      0.998     0.997     0.997   1233097
           1      0.993     0.994     0.994    933948
           2      0.982     0.982     0.982    423392
           3      0.982     0.981     0.981    558113
           4      0.972     0.970     0.971    403410
           5      0.949     0.946     0.947    277877
           6      0.882     0.884     0.883    188112
           7      0.886     0.892     0.889    265015
           8      0.889     0.894     0.892    245183
           9      0.890     0.893     0.891    136120
          10      0.916     0.863     0.889     35733

    accuracy                          0.968   4700000
   macro avg      0.940     0.936     0.938   4700000
weighted avg      0.968     0.968     0.968   4700000
```



Submission and Description	Private Score	Public Score	Use for Final Score
sub_vanilla_hmm_cleaned_data.csv a few seconds ago by Gilles Vandewiele add submission details	0.93120	0.93737	□

Quick recap

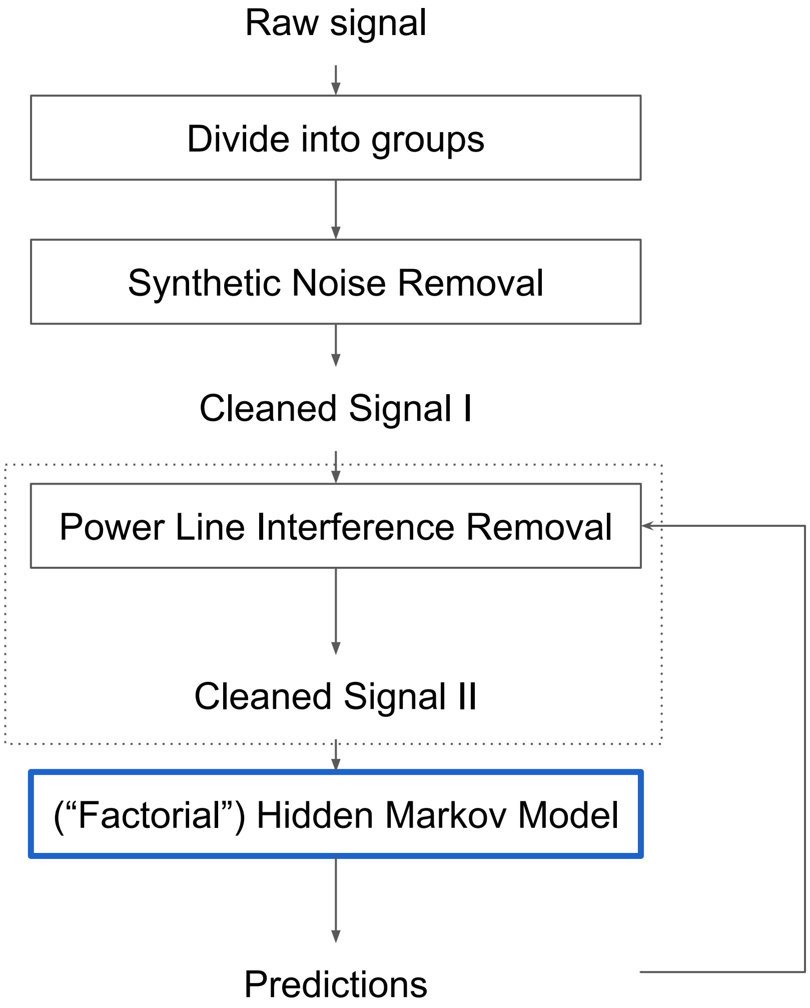
Method	Train	Public	Private
Baseline	0.924	0.925	0.915
Vanilla HMM	0.935	0.934	0.928
Power Line + Vanilla HMM	0.938	0.937	0.931

Good correlation between train, public & private scores!

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

Advanced HMMs (~ Factorial Hidden Markov Models)



Raw signal

Divide into groups

Synthetic Noise Removal

Cleaned Signal I

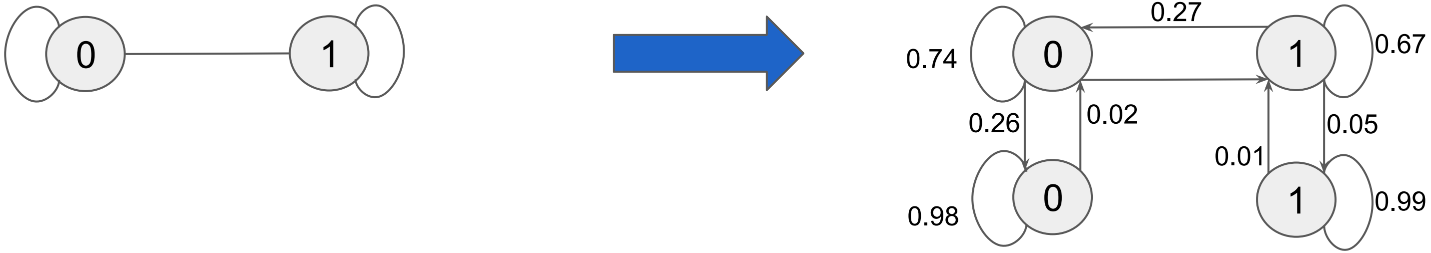
Power Line Interference Removal

Cleaned Signal II

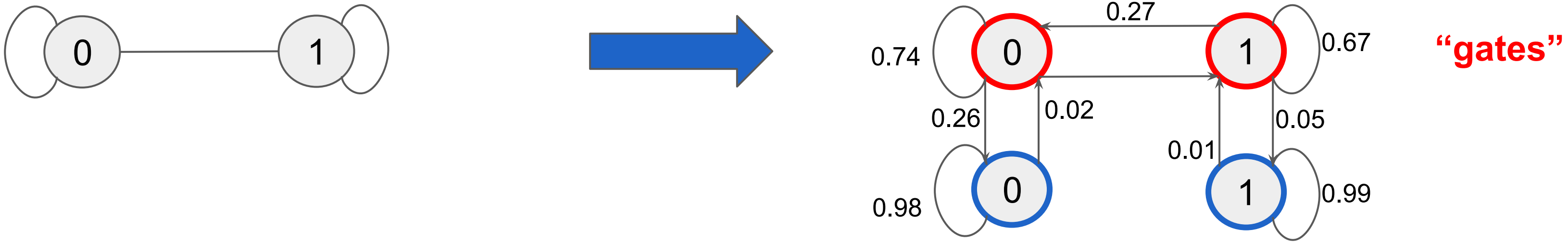
(“Factorial”) Hidden Markov Model

Predictions

Insight 1: data with 0/1 channels, has more than 2 hidden states

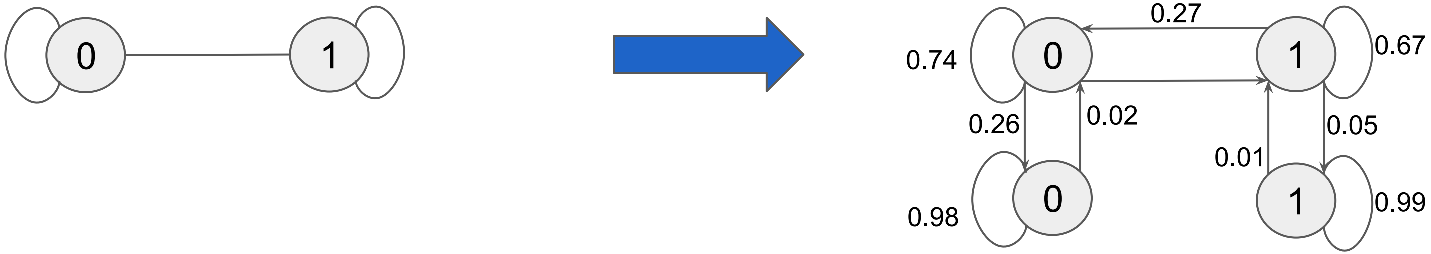


Insight 1: data with 0/1 channels, has more than 2 hidden states

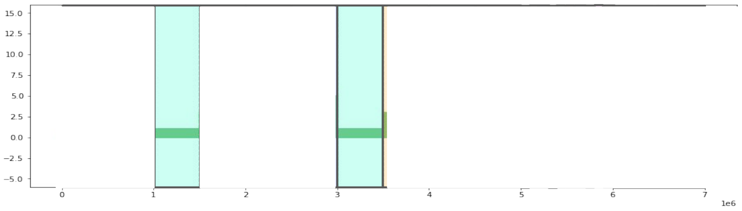


Produce longer sequence of 0's and 1's

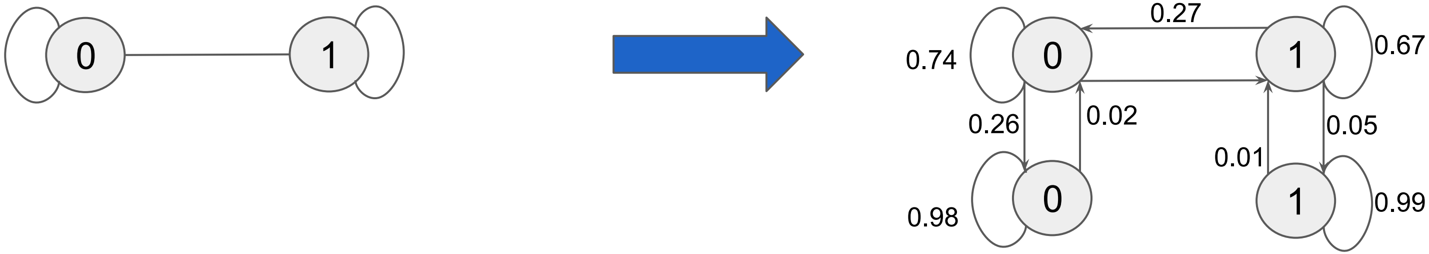
Insight 1: data with 0/1 channels, has more than 2 hidden states



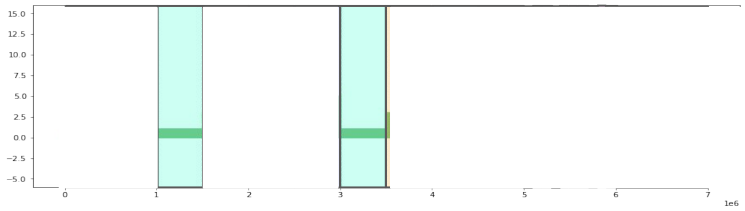
<u>Small experiment</u> on 0/1 data

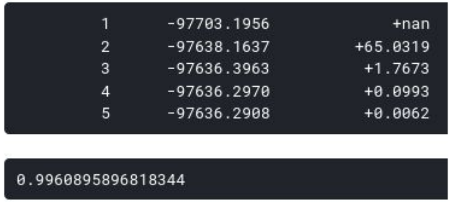


Insight 1: data with 0/1 channels, has more than 2 hidden states



Small experiment on 0/1 data

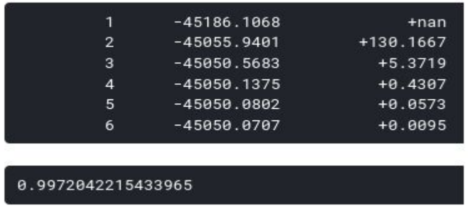




```
         1          -97703.1956             +nan
         2          -97638.1637          +65.0319
         3          -97636.3963           +1.7673
         4          -97636.2970           +0.0993
         5          -97636.2908           +0.0062
```

```
0.9960895896818344
```

hmmlearn with 2 states

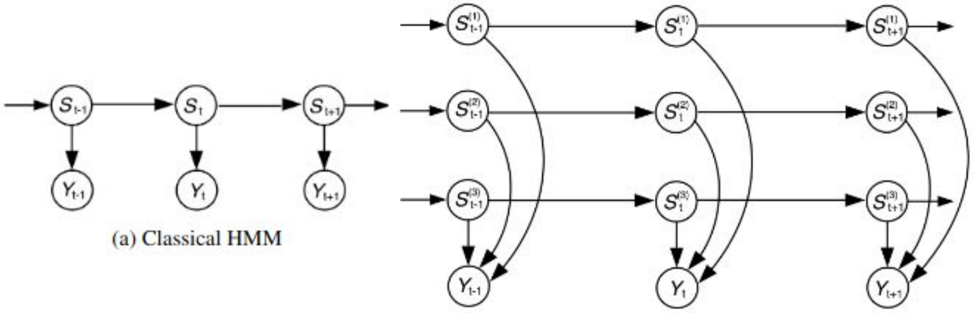


```
         1          -45186.1068             +nan
         2          -45055.9401          +130.1667
         3          -45050.5683            +5.3719
         4          -45050.1375            +0.4307
         5          -45050.0802            +0.0573
         6          -45050.0707            +0.0095
```

```
0.9972042215433965
```

hmmlearn with 4 states

Insight 2: data with k > 1 channels = sum of k 0/1 processes



(a) Classical HMM

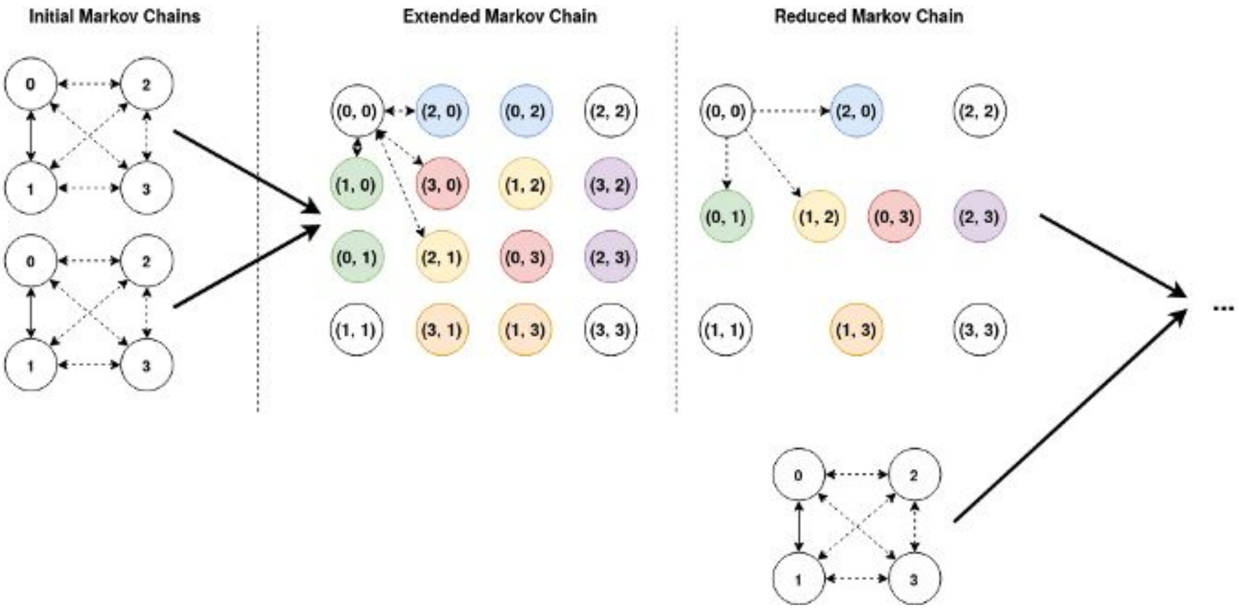
(b) Factorial HMM

https://emilemathieu.fr/files/fhmmreport.pdf

Insight 2: data with k > 1 channels = sum of k 0/1 processes

These Factorial HMMs were not trivial to implement…
→ We converted each of our “vanilla” k-state processes to n-state processes with

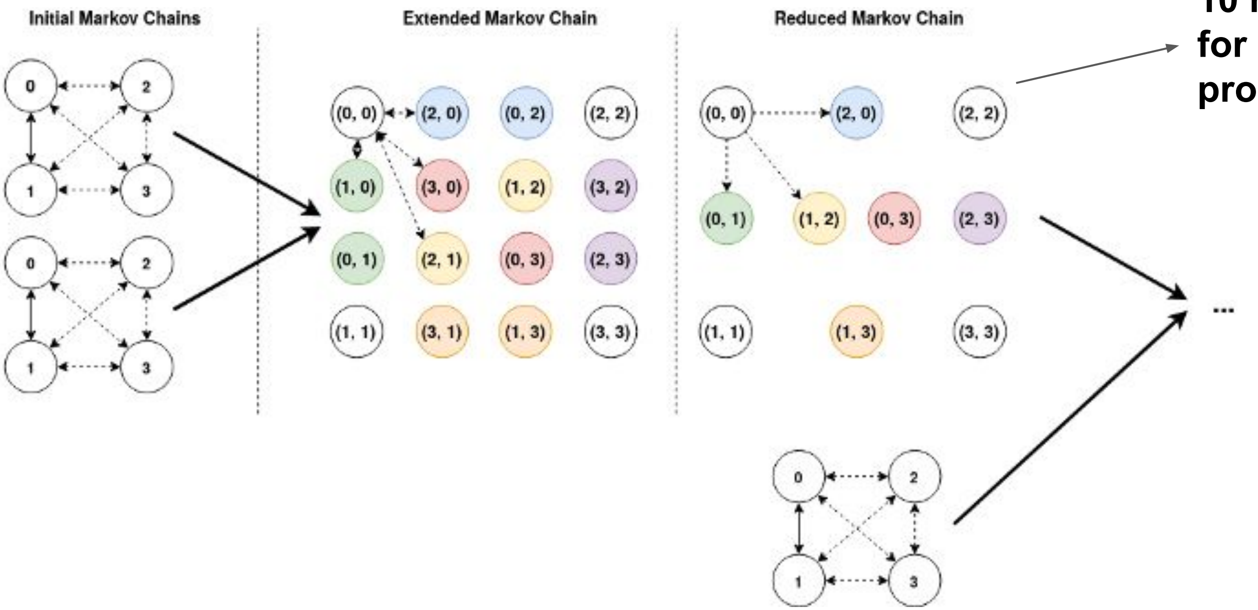
$$\binom{\bar{4}}{k} = \binom{4+k-1}{k}$$



Insight 2: data with k > 1 channels = sum of k 0/1 processes

These Factorial HMMs were not trivial to implement…
→ We converted each of our “vanilla” k-state processes to n-state processes with

$$\binom{\bar{4}}{k} = \binom{4+k-1}{k}$$



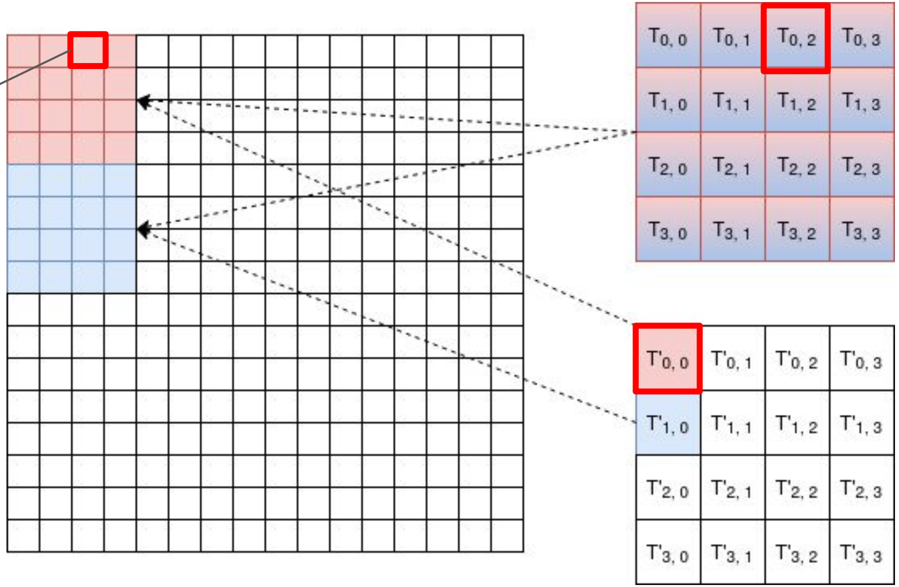
10 hidden states for sum of 2 processes

Insight 2: data with k > 1 channels = sum of k 0/1 processes

These Factorial HMMs were not trivial to implement…
→ We converted each of our “vanilla” k-state processes to n-state processes with

$$\binom{\bar{4}}{k} = \binom{4+k-1}{k}$$

Probability that our two chains were in state (0, 0) at time t and in state (0, 2) at time t + 1



$T_{0,0}$	$T_{0,1}$	$T_{0,2}$	$T_{0,3}$
$T_{1,0}$	$T_{1,1}$	$T_{1,2}$	$T_{1,3}$
$T_{2,0}$	$T_{2,1}$	$T_{2,2}$	$T_{2,3}$
$T_{3,0}$	$T_{3,1}$	$T_{3,2}$	$T_{3,3}$

$T'_{0,0}$	$T'_{0,1}$	$T'_{0,2}$	$T'_{0,3}$
$T'_{1,0}$	$T'_{1,1}$	$T'_{1,2}$	$T'_{1,3}$
$T'_{2,0}$	$T'_{2,1}$	$T'_{2,2}$	$T'_{2,3}$
$T'_{3,0}$	$T'_{3,1}$	$T'_{3,2}$	$T'_{3,3}$

Insight 2: data with k > 1 channels = sum of k 0/1 processes

These Factorial HMMs were not trivial to implement…
→ We converted each of our “vanilla” k-state processes to n-state processes with

$$\binom{\bar{4}}{k} = \binom{4 + k - 1}{k}$$

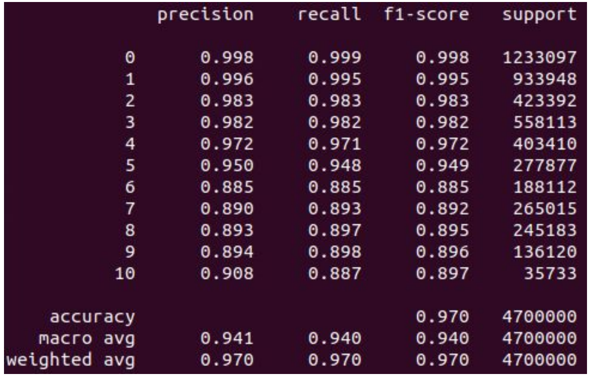
Data with 10 open channels has 286 hidden states!
→ hmmlearn becomes slow
→ implement our own custom algorithm

$$\alpha(t) = P_{\text{sig}}(t) * (\alpha(t-1) * P_{\text{tran}})^{c} * \beta(t)^{(1-c)}$$

$$\beta(t) = P_{\text{sig}}(t) * (\beta(t+1) * P_{\text{tran}}^{T})^{c} * \alpha(t)^{(1-c)}$$

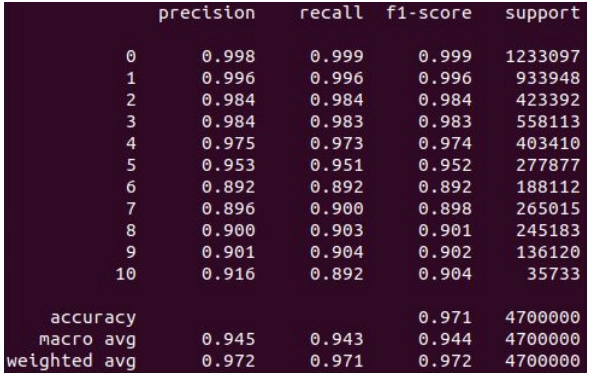
Final results...

iteration 1



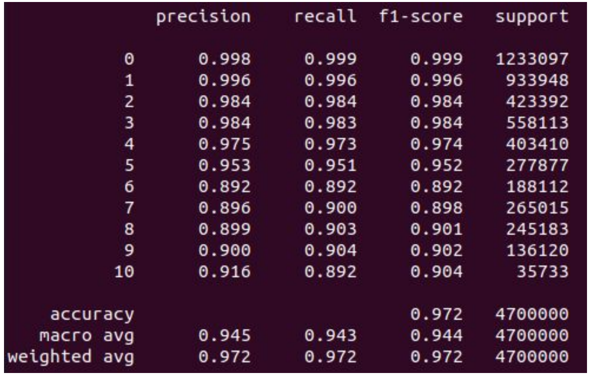
```
              precision    recall  f1-score   support

           0      0.998     0.999     0.998   1233097
           1      0.996     0.995     0.995    933948
           2      0.983     0.983     0.983    423392
           3      0.982     0.982     0.982    558113
           4      0.972     0.971     0.972    403410
           5      0.950     0.948     0.949    277877
           6      0.885     0.885     0.885    188112
           7      0.890     0.893     0.892    265015
           8      0.893     0.897     0.895    245183
           9      0.894     0.898     0.896    136120
          10      0.908     0.887     0.897     35733

    accuracy                          0.970   4700000
   macro avg      0.941     0.940     0.940   4700000
weighted avg      0.970     0.970     0.970   4700000
```

iteration 2

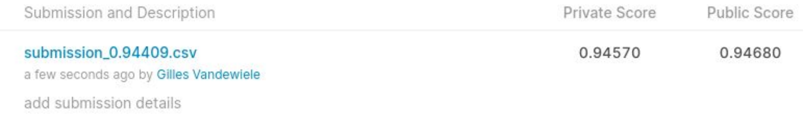


```
              precision    recall  f1-score   support

           0      0.998     0.999     0.999   1233097
           1      0.996     0.996     0.996    933948
           2      0.984     0.984     0.984    423392
           3      0.984     0.983     0.983    558113
           4      0.975     0.973     0.974    403410
           5      0.953     0.951     0.952    277877
           6      0.892     0.892     0.892    188112
           7      0.896     0.900     0.898    265015
           8      0.900     0.903     0.901    245183
           9      0.901     0.904     0.902    136120
          10      0.916     0.892     0.904     35733

    accuracy                          0.971   4700000
   macro avg      0.945     0.943     0.944   4700000
weighted avg      0.972     0.971     0.972   4700000
```

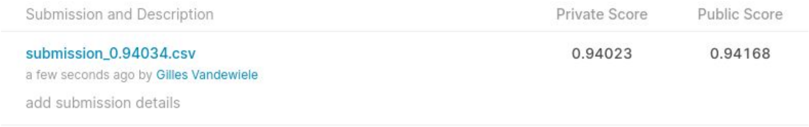
iteration 3



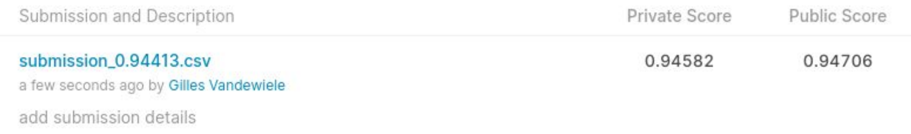
```
              precision    recall  f1-score   support

           0      0.998     0.999     0.999   1233097
           1      0.996     0.996     0.996    933948
           2      0.984     0.984     0.984    423392
           3      0.984     0.983     0.984    558113
           4      0.975     0.973     0.974    403410
           5      0.953     0.951     0.952    277877
           6      0.892     0.892     0.892    188112
           7      0.896     0.900     0.898    265015
           8      0.899     0.903     0.901    245183
           9      0.900     0.904     0.902    136120
          10      0.916     0.892     0.904     35733

    accuracy                          0.972   4700000
   macro avg      0.945     0.943     0.944   4700000
weighted avg      0.972     0.972     0.972   4700000
```



Submission and Description	Private Score	Public Score
submission_0.94409.csv a few seconds ago by Gilles Vandewiele add submission details	0.94570	0.94680



Submission and Description	Private Score	Public Score
submission_0.94034.csv a few seconds ago by Gilles Vandewiele add submission details	0.94023	0.94168



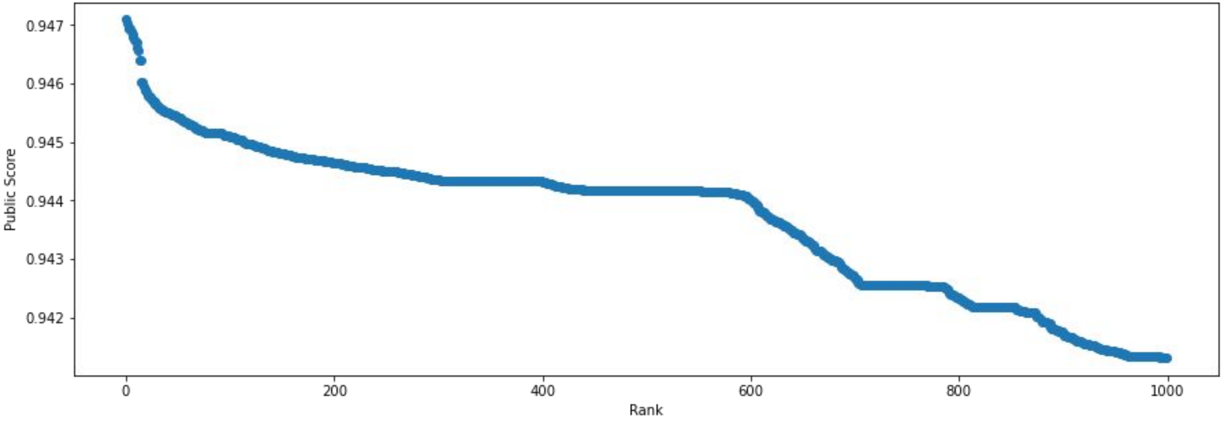
Submission and Description	Private Score	Public Score
submission_0.94413.csv a few seconds ago by Gilles Vandewiele add submission details	0.94582	0.94706

FACULTY OF ENGINEERING
AND ARCHITECTURE

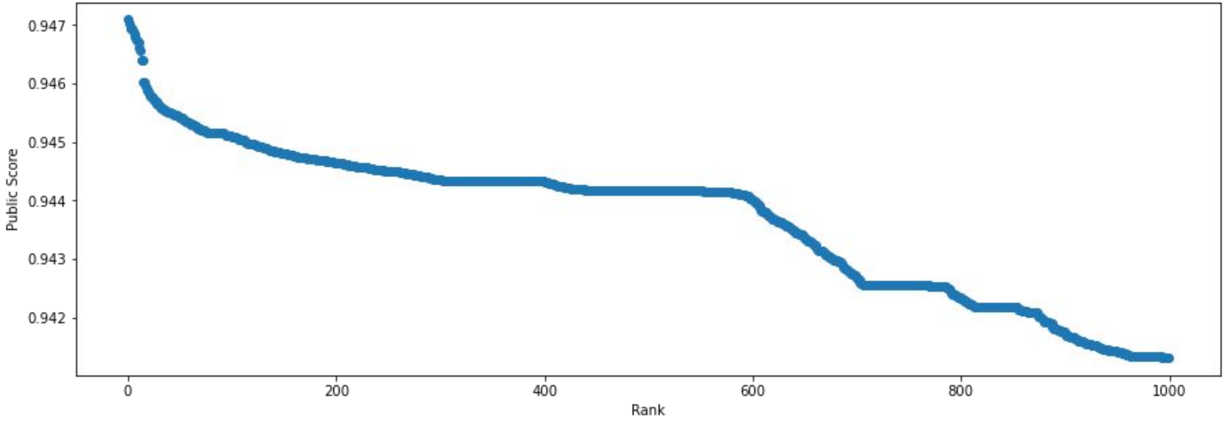
Liverpool - Ion Switching Competition

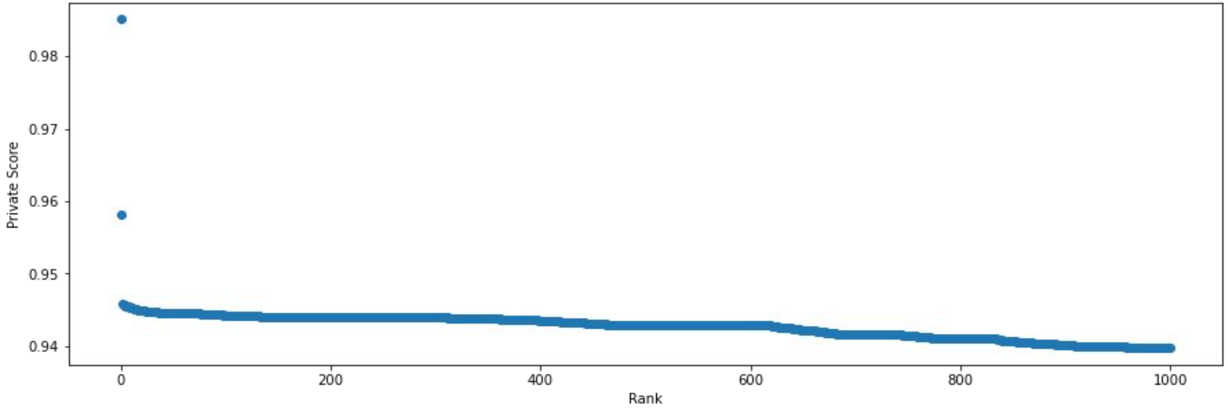
The leak

Two teams managed to obtain an extremely high private score. Kudos to them!

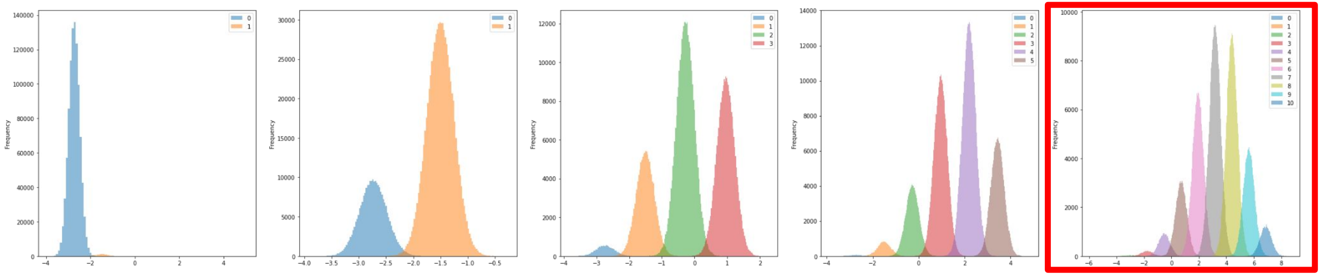


Two teams managed to obtain an extremely high private score. Kudos to them!





Something peculiar about the 0-10 data...



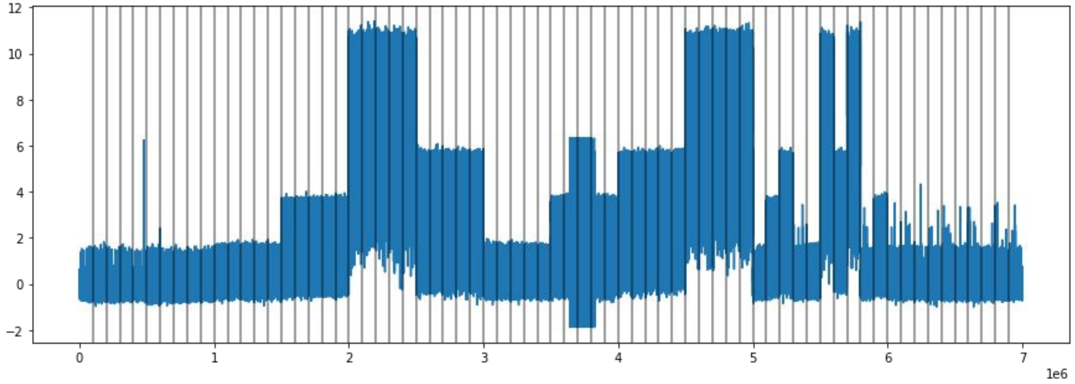
1) The stddevs of all gaussians are around 0.28, except for the 0-10 data, the stddevs are around 0.40

→ sqrt(0.28 ** 2 + 0.28 ** 2) ~ 0.40

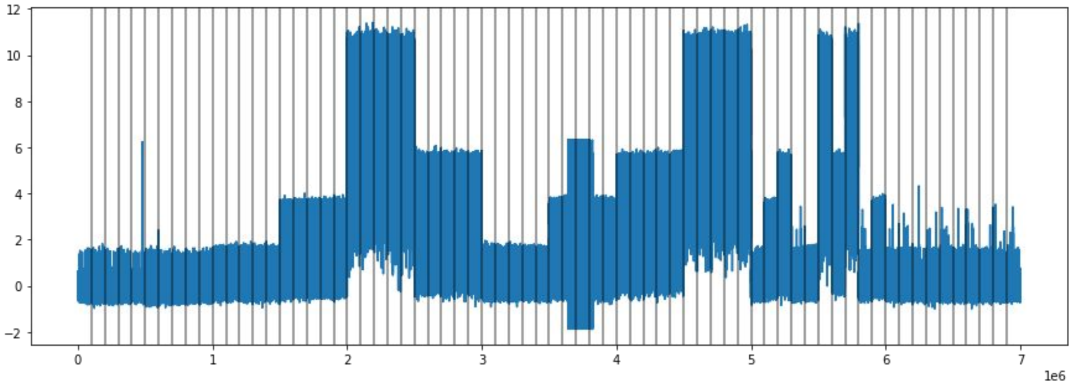
2) The mean of the all 0-10 data is roughly twice the mean of other data

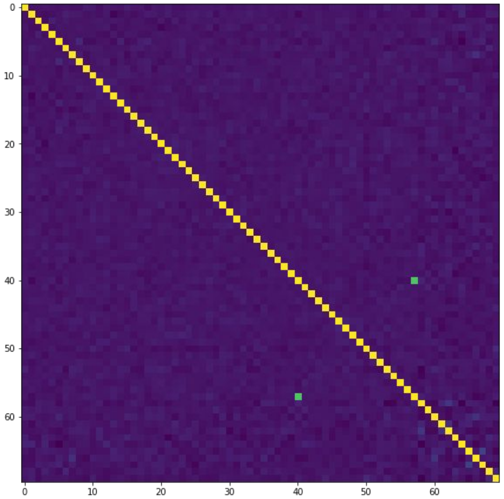
→ Turns out the 0-10 data is actually (0-5 data) + (0-5 data)
→ Organisers generated synthetic data using a matlab scripts, but they did it across multiple sessions and matlab is seeded BY DEFAULT (like C) so calls to random() will always give same results...

Let’s look at correlations between the data

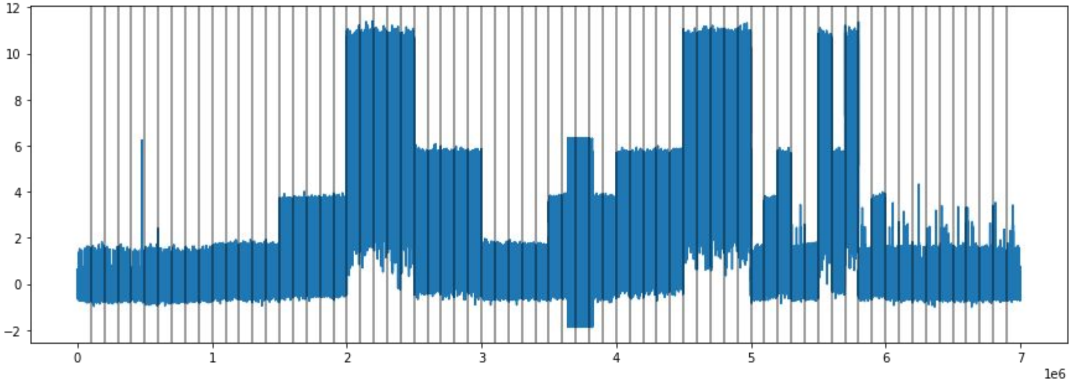


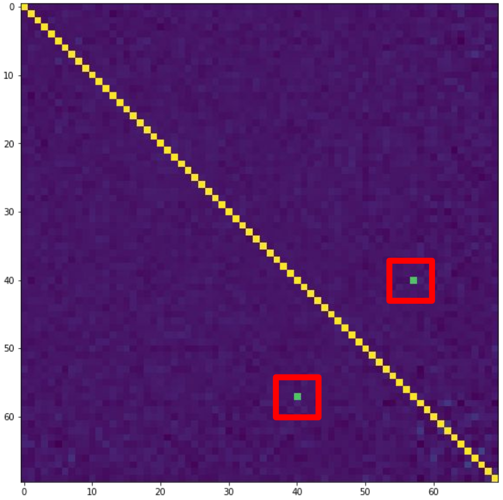
Let’s look at correlations between the data





Let’s look at correlations between the data

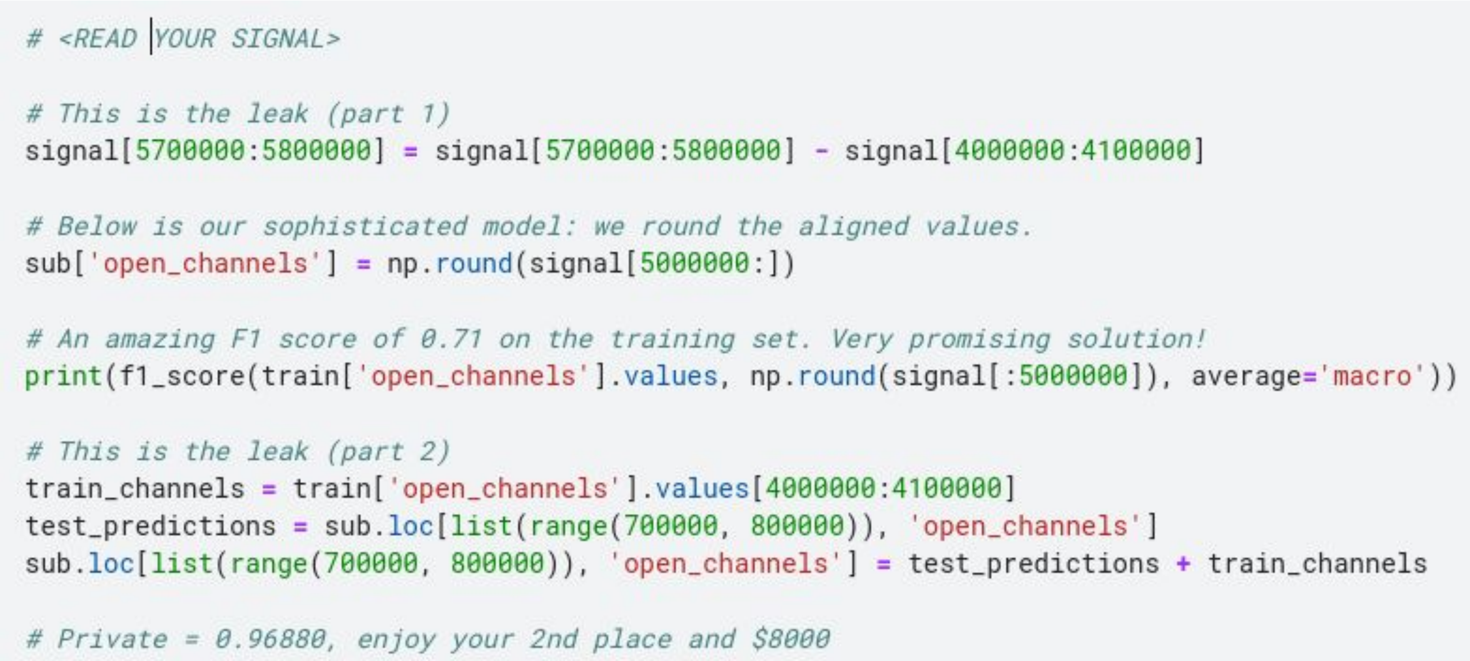




Turns out the 0-10 data is actually (0-5 data) + (0-5 data)

and one part is from the training data

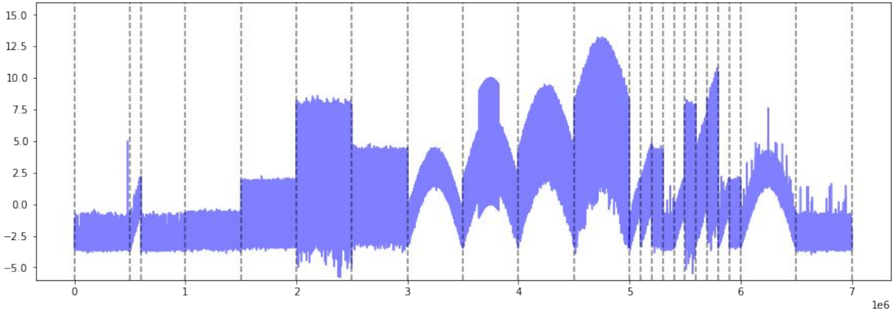
Rounding signal values gets 2nd spot!



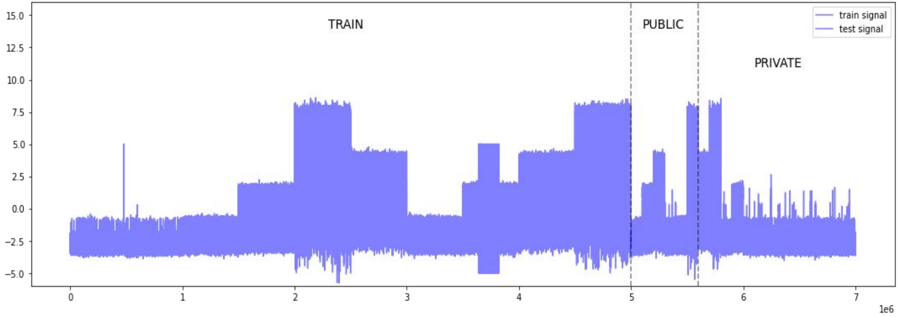
```
# <READ YOUR SIGNAL>

# This is the leak (part 1)
signal[5700000:5800000] = signal[5700000:5800000] - signal[4000000:4100000]

# Below is our sophisticated model: we round the aligned values.
sub['open_channels'] = np.round(signal[5000000:])

# An amazing F1 score of 0.71 on the training set. Very promising solution!
print(f1_score(train['open_channels'].values, np.round(signal[:5000000]), average='macro'))

# This is the leak (part 2)
train_channels = train['open_channels'].values[4000000:4100000]
test_predictions = sub.loc[list(range(700000, 800000)), 'open_channels']
sub.loc[list(range(700000, 800000)), 'open_channels'] = test_predictions + train_channels

# Private = 0.96880, enjoy your 2nd place and $8000
```

Conclusion / Summary

1. Start from the original data

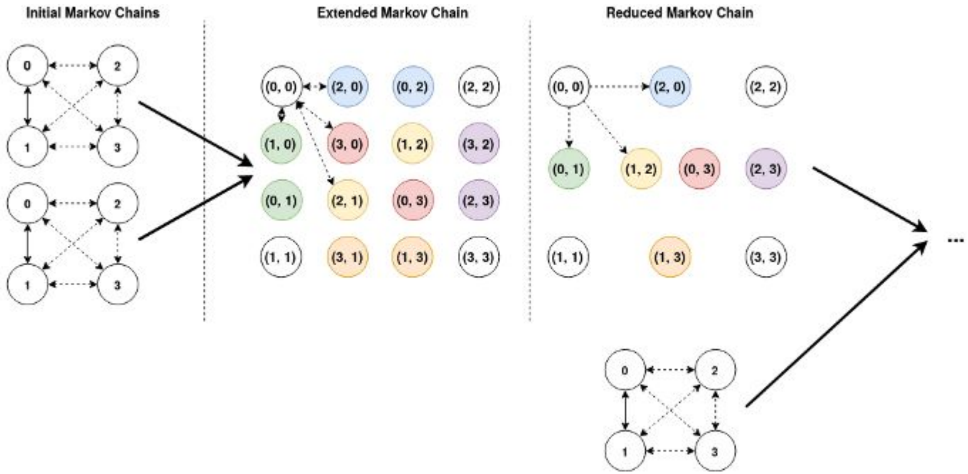


2. Remove the low-frequency sine noise



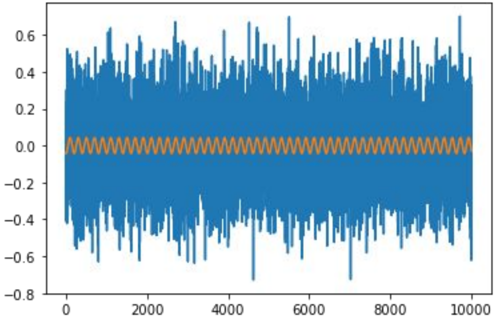
Conclusion / Summary

3. Create a HMM with $\binom{\bar{4}}{k}$ hidden states that represents k independent Markov Processes / Chains with 4 hidden states for each of our 5 (6) categories of data



Conclusion / Summary

4. Generate predictions & use these to isolate the error signal. From this signal, remove power line interference by fitting a sine function



5. Repeat steps 3 & 4 until convergence. Optionally, introduce the leak for a big boost.

FACULTY OF ENGINEERING
AND ARCHITECTURE

Liverpool - Ion Switching Competition

General Tips & Tricks

General Tips & Tricks

1. **Join for the learning experience, the community and the fun. Not the medals**

General Tips & Tricks

1. Join for the learning experience, the community and the fun. Not the medals
2. Priority: set up local evaluation that correlates with LB score

General Tips & Tricks

1. Join for the learning experience, the community and the fun. Not the medals
2. Priority: set up local evaluation that correlates with LB score
→ adversarial validation
→ identify “lottery” competitions

General Tips & Tricks

1. Join for the learning experience, the community and the fun. Not the medals
2. Priority: set up local evaluation that correlates with LB score
→ adversarial validation
→ identify “lottery” competitions
3. Strategies

General Tips & Tricks

1. **Join for the learning experience, the community and the fun. Not the medals**
2. **Priority: set up local evaluation that correlates with LB score**
→ adversarial validation
→ identify “lottery” competitions
3. **Strategies**
→ Join early vs late
→ Focus on one vs multiple competitions
→ Solo & Team

General Tips & Tricks

1. Join for the learning experience, the community and the fun. Not the medals
2. Priority: set up local evaluation that correlates with LB score
→ adversarial validation
→ identify “lottery” competitions
3. Strategies
→ Join early vs late
→ Focus on one vs multiple competitions
→ Solo & Team
4. Embrace the sharing mentality (discussions & notebooks)

Code, blog post & kaggle resources

https://github.com/GillesVandewiele/Liverpool-Ion-Switching

https://towardsdatascience.com/identifying-the-number-of-open-ion-channels-with-hidden-markov-models-334fab86fc85

https://www.kaggle.com/group16/lb-0-936-1-feature-forward-backward-vs-viterbi

https://www.kaggle.com/group16/private-0-9688-a-better-but-useless-solution

These slides will be published online shortly!

Thank You!

kaggle.com/group16

gilles.vandewiele@ugent.be

twitter.com/Gillesvdwiele

linkedin.com/in/gillesvandewiele

www.gillesvandewiele.com

kaggle.com/zidmie

yves-miehe-184808b4

kaggle.com/khahuras

kha-vo-phd-b91a38126