

MLJ Workshop

Resources: github.com/ablaom/MLJTutorial

Anthony Blaom with Quinn Asena

The Alan Turing Institute



THE UNIVERSITY OF
AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND



The Team

Core design: A. B., Franz Kiraly, Sebastian Vollmer

Lead development: A. B., Thibaut Lienart

Other contributors: Diego Arenas, Samuel Okon, Yiannis Simillides, Julian Samaroo, Ayush Shridar, Geoffroy Dolphin, Mosè Giordano...

Julia language consultants: Avik Sengupta

Some machine learning libraries in Julia

```
22-element Vector{String}:
"OutlierDetectionNeighbors"
"OutlierDetectionPython"
"OutlierDetectionNetworks"
"ScikitLearn"
"DecisionTree"
"MultivariateStats"
"MLJModels"
"BetaML"
"MLJLinearModels"
"LIBSVM"
"EvoTrees"
"NaiveBayes"
"MLJFlux"
"Clustering"
"ParallelKMeans"
"NearestNeighborModels"
"PartialLeastSquaresRegressor"
"LightGBM"
"GLM"
"TSVD"
"MLJText"
"XGBoost"
```

Some multi-paradigm machine learning toolboxes in Julia

- ScikitLearn.jl - thin wrapper for Python's scikit-learn
- AutoMLPipeline.jl
- MLJ.jl
- FastAI.jl - specific to neural networks

What?

What's a machine learning toolbox?

What?

What's a machine learning toolbox?

- A toolbox provides a **uniform interface** for **fitting**, **evaluating** and **tuning** machine learning models.
- Provides common **preprocessing tasks** (such as data cleaning and type coercion)
- Allows for model **composition** (e.g., pipelining)

Why?

Why learn the MLJ toolbox?

- Written in Julia (but does wrap non-native models):

Why?

Why learn the MLJ toolbox?

- Written in Julia (but does wrap non-native models):
 - Easy access to core algorithms
 - Easier to customize
 - Easier to add new performant models
 - Greater transparency and reproducibility
 - Better composability with other libraries

Why?

Why learn the MLJ toolbox?

- Written in Julia (but does wrap non-native models):
 - Easy access to core algorithms
 - Easier to customize
 - Easier to add new performant models
 - Greater transparency and reproducibility
 - Better composability with other libraries
- Start-of-the-art **model composition** that plays well with everything else.
- Meta-algorithms (e.g., tuning) are model wrappers

How to entertain your kids at home - the case for MLJ



3 models dinosaurs + 15 user contributed
models

We **support** you in building your snake powered train!



Workshop Overview

- Recap of supervised and unsupervised learning
- Preview of model composition (stacking)
- Part 1 - **Data Representation** + exercises
- Part 2 - **Selecting, Training and Evaluating Models** + exercises
- Break
- Part 3 - **Transformers and Pipelines** + exercises
- Part 4 - **Tuning hyper-parameters** + exercises
- Part 5 - **Advanced model composition** (time permitting)

Supervised Learning

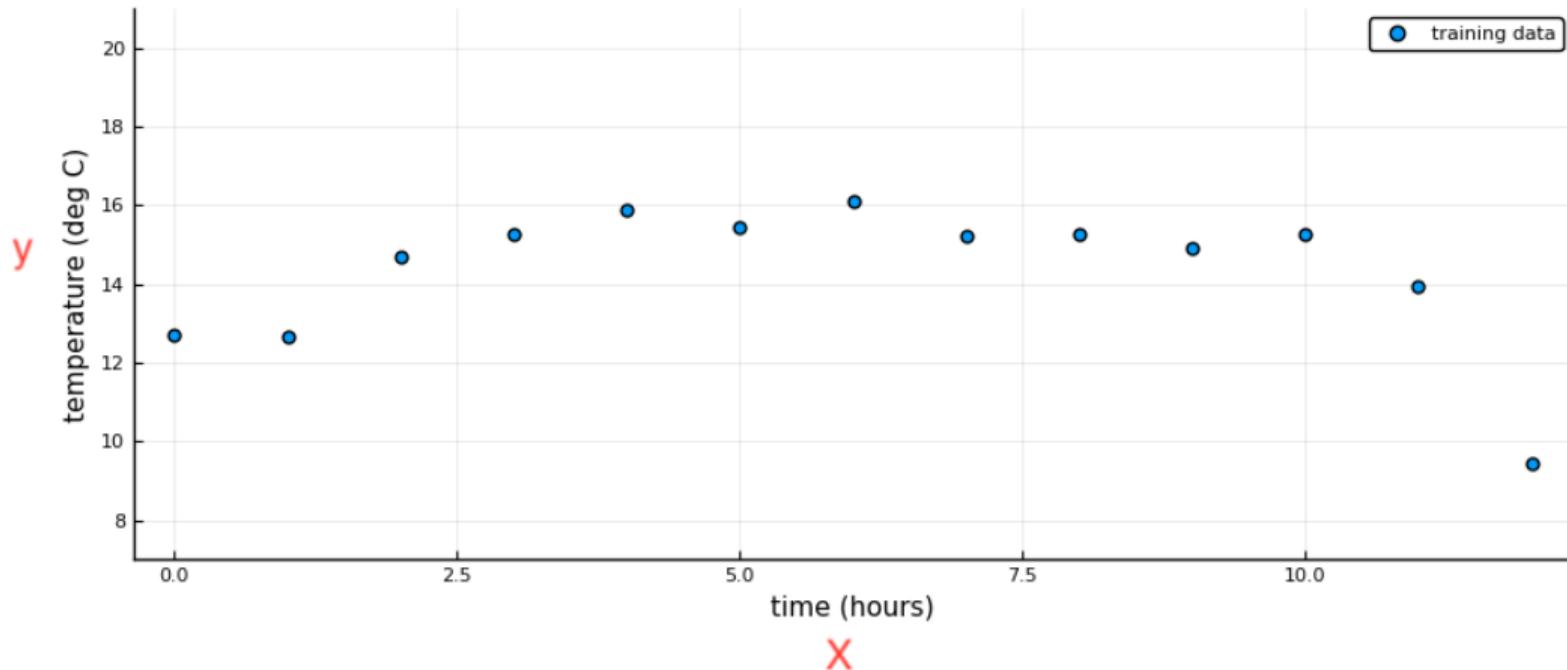
Learning to **predict** some target variable **y** from a knowledge of some other variables **X** (the *input features*).

Supervised Learning

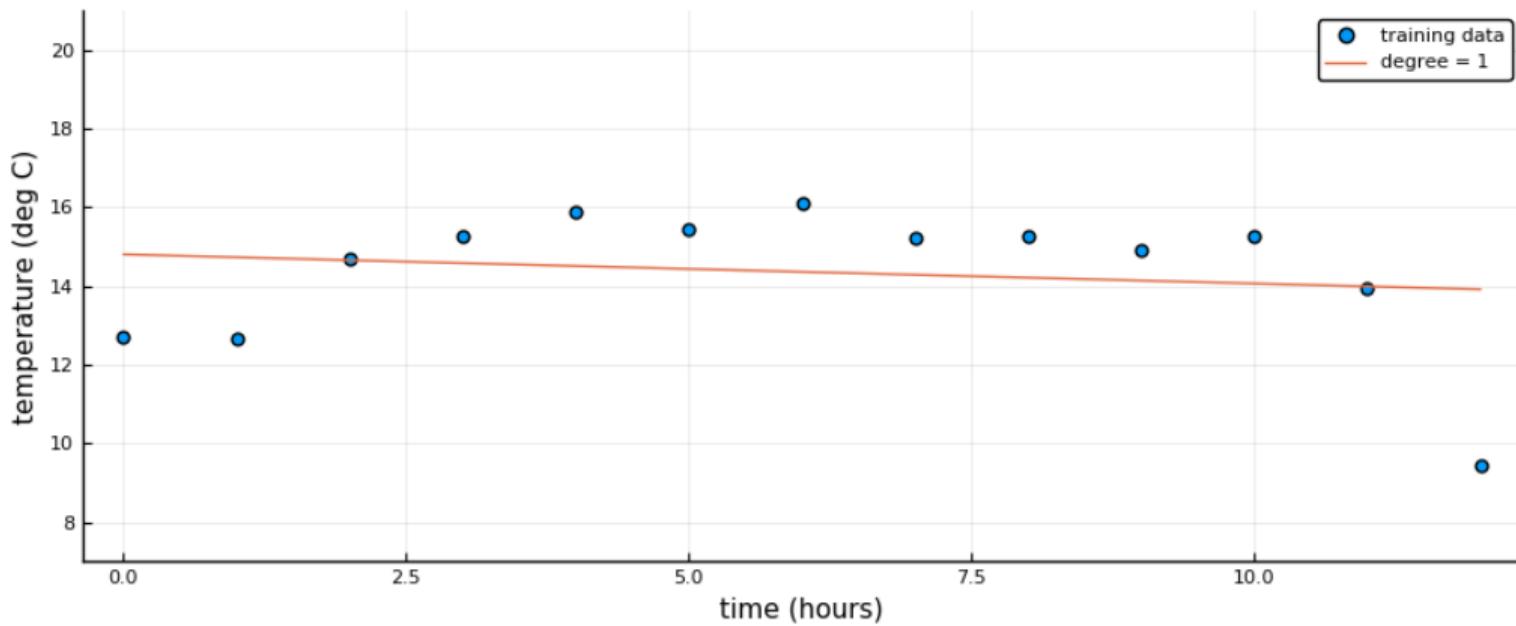
Learning to **predict** some target variable **y** from a knowledge of some other variables **X** (the *input features*).

X		y
time	room	temperature
5	kitchen	18.5
5	bathroom	18.3
5	bedroom_1	18.3
5	living_room	17.4
6	kitchen	16.6
6	bathroom	20.7
6	bedroom_1	18.9
6	living_room	20.2
7	kitchen	20.4
7	bathroom	19.9
7	bedroom_1	16.2
7	living_room	17.3

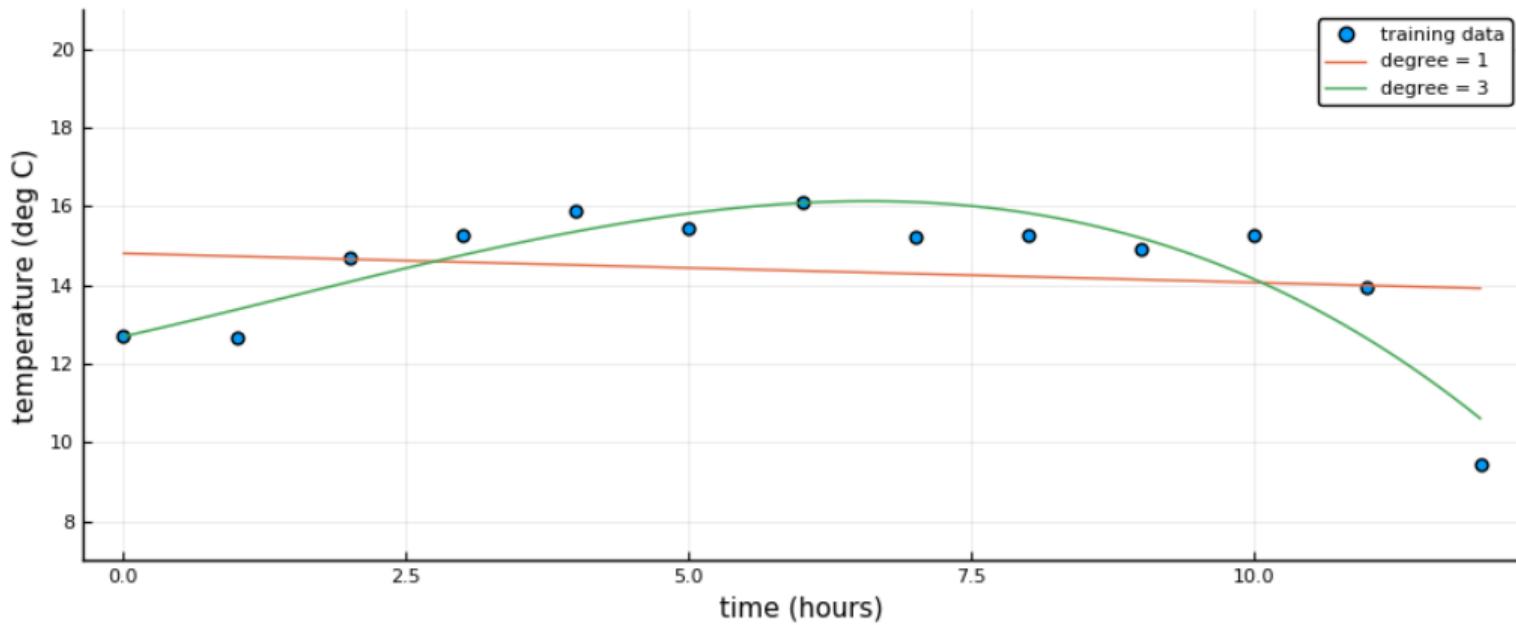
Supervised Learning



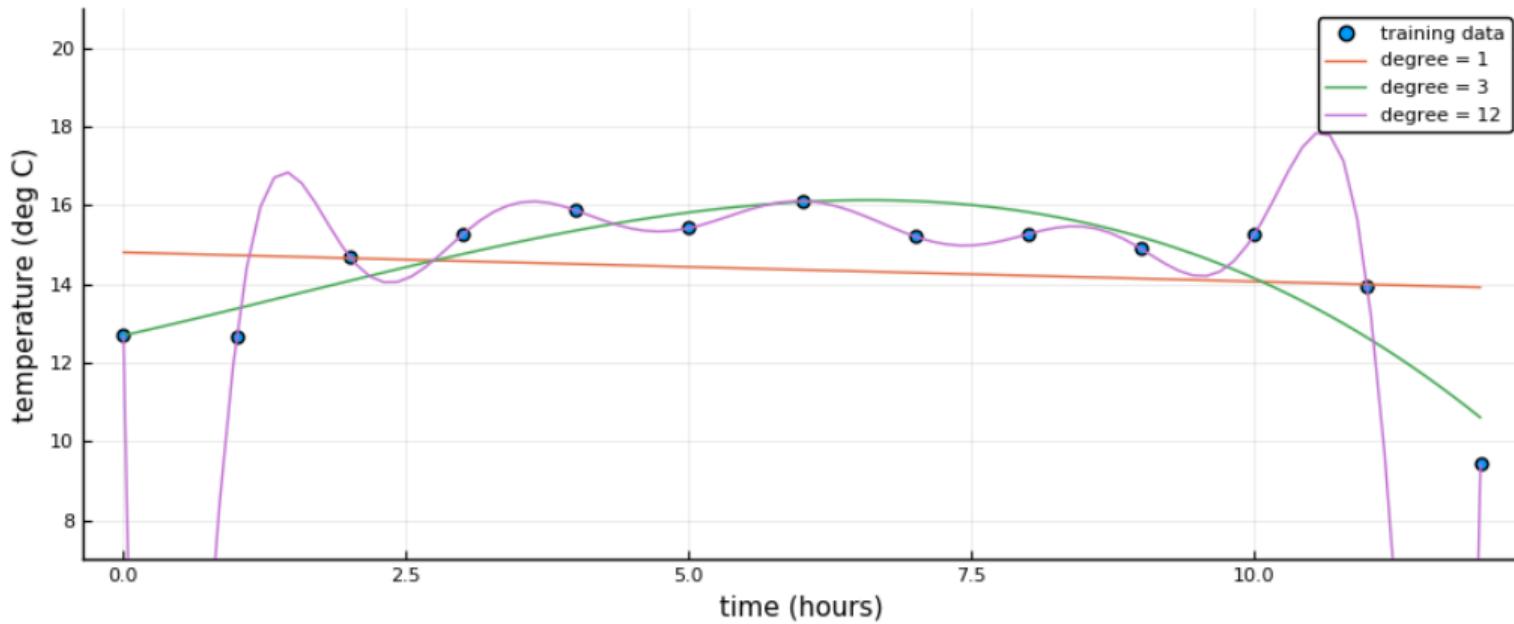
Supervised Learning



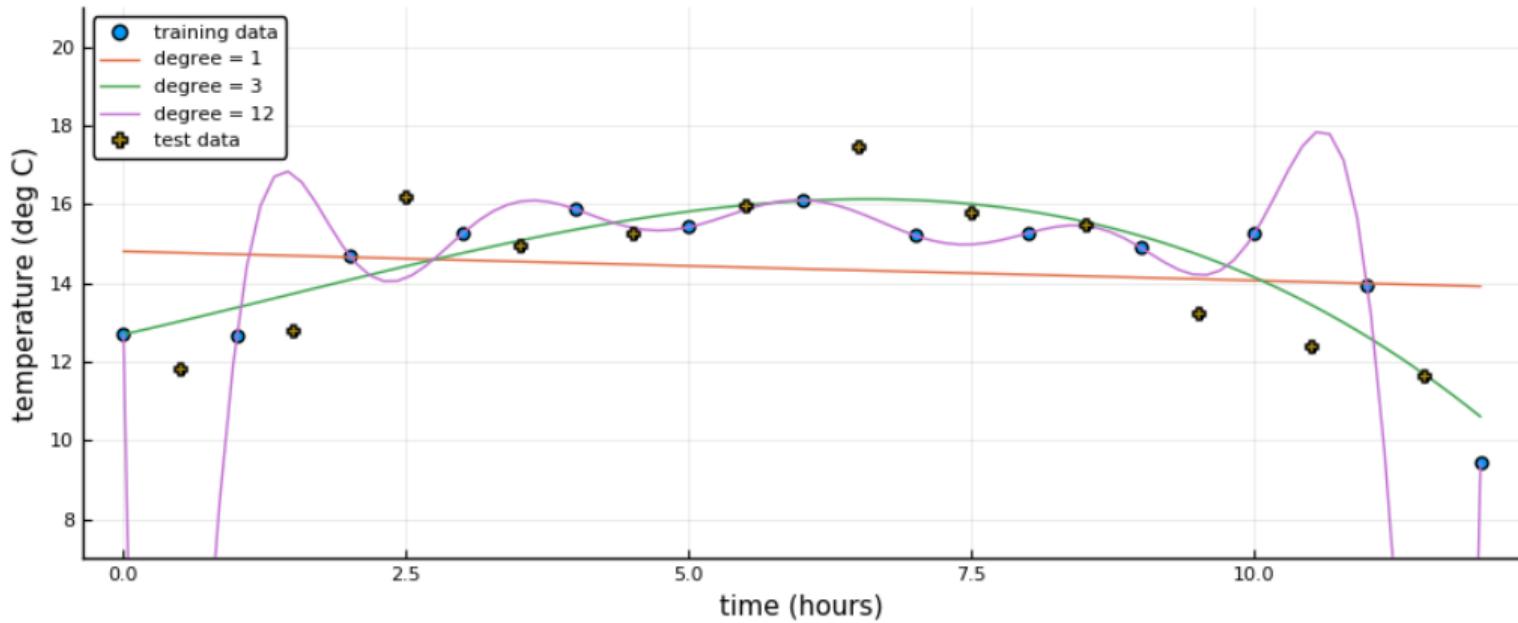
Supervised Learning



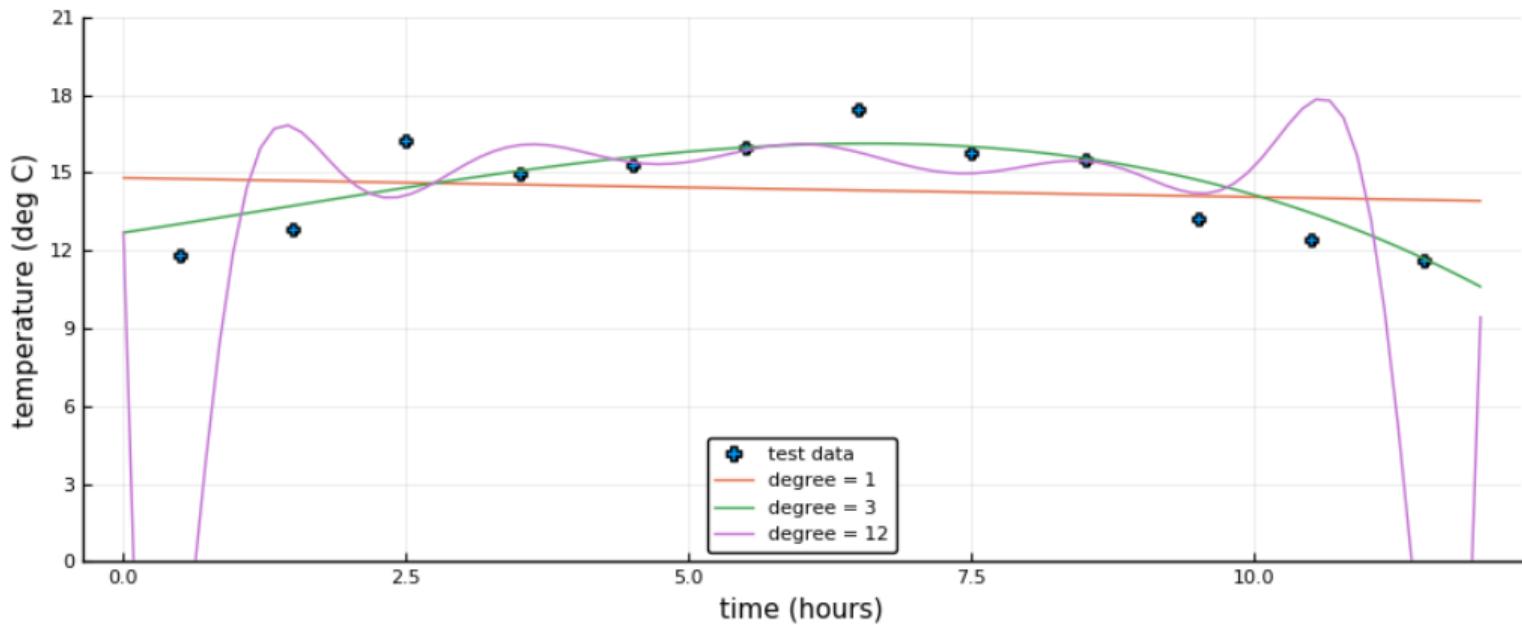
Supervised Learning



Supervised Learning

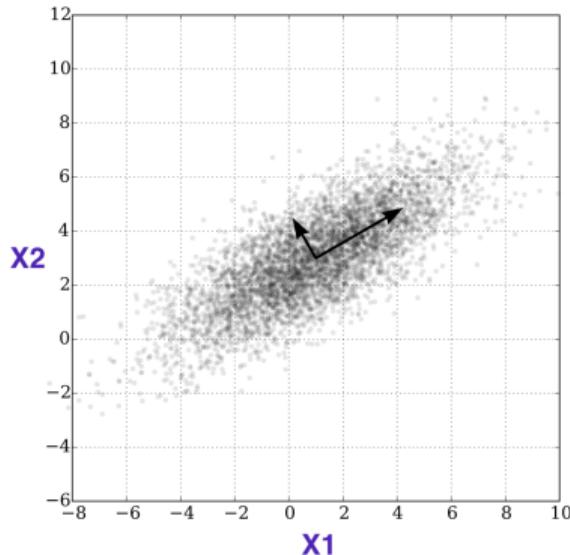


Supervised Learning



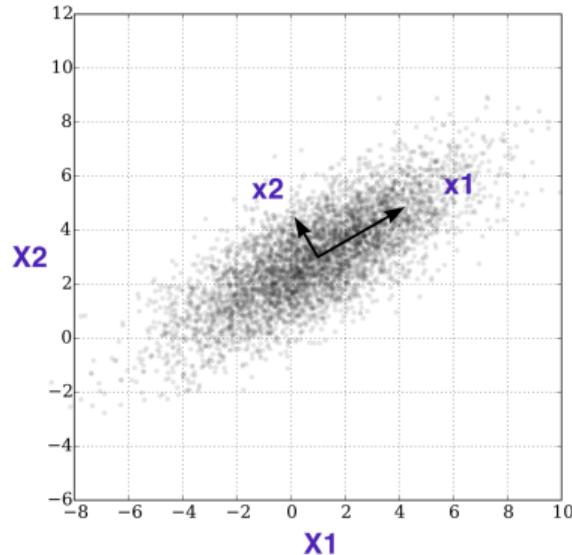
Unsupervised Learning

Learning data **transformations**, e.g., dimension reduction



Unsupervised Learning

Learning data **transformations**, e.g., dimension reduction



Data science competitions (kaggle)

IEEE-CIS Fraud Detection
Can you detect fraud from customer transactions?

IEEE Computational Intelligence Society - 70 Teams - 2 months to go 2 months to go until merge deadline

Overview Data Kernels Discussions Leaderboard Rules Join Competition

Public Leaderboard **Private Leaderboard**

This leaderboard is calculated with approximately 20% of the test data.
The final results will be based on the other 80%, so the final standings may be different.

Raw Data Refresh

In the money Gold Silver Bronze

#	Team Name	Kernel	Team Members	Score ⚡	Entries	Last
1	MingLiwei			0.9482	5	2h
2	Michael Jahn			0.9478	10	5h
3	TUW			0.9466	20	1h
4	[ods.ai] SinisterThree			0.9463	19	6h
5	José Pedro Peinado			0.9460	12	1d
6	3 LLamas			0.9455	10	5h
7	AL			0.9433	11	1d
8	THLUO			0.9432	7	7h
9	Aleksandr Koslapov			0.9431	20	4h
10	Li-Der			0.9431	19	1h
11	Raghavendra Singh			0.9430	17	2h
12	QayyamEl			0.9429	3	1d
13	Team Data			0.9428	15	2h
14	Patrick Chan			0.9427	11	2h
15	less			0.9426	11	4h
16	AndreaToacher			0.9424	10	10h

Cell Signal Prediction Genetics
CellSignal: Disentangling biological signal from experimental noise in cellular images

Recursion Pharmaceuticals - 469 teams - 2 months to go 13 days to go until merge deadline

Overview Data Kernels Discussion Leaderboard Rules Join Competition

Public Leaderboard **Private Leaderboard**

This leaderboard is calculated with approximately 22% of the test data.
The final results will be based on the other 78%, so the final standings may be different.

Raw Data Refresh

In the money Gold Silver Bronze

#	Team Name	Kernel	Team Members	Score ⚡	Entries	Last
1	[ods.ai] OndrejLerma			0.864	41	1d
2	gold diggers			0.860	81	12h
3	yu4e			0.690	13	16m
4	[attention heads] + shmyrko			0.684	61	1d
5	Dawid			0.672	34	5h
6	taski			0.668	9	6d
7	rapidaul			0.634	60	3h
8	Double strand			0.630	36	10h
9	mandarinente			0.620	34	7h
10	Road to NeuIPS			0.616	46	3d
11	Kirill Brodt (ahad nsk)			0.586	5	3d
12	-			0.586	25	10h
13	janning			0.573	41	10h
14	MikhailPapkov			0.564	5	18m
15	Konstantin Loshchkin			0.560	56	1d

Featured Code Competition
Jigsaw Unintended Bias in Toxicity Classification
Detect toxicity across a diverse range of conversations

Jigsaw/Conversation AI - 2,846 teams - 6 day ago

Overview Data Kernels Discussions Leaderboard Rules Late Submissions

Public Leaderboard **Private Leaderboard**

This is a Kernel Competition with two stages. The public leaderboard represents scores on the stage 1 test set. Your final private leaderboard score and ranking will be determined in stage 2, when selected kernels are re-run on a withheld private test set. For more information, review the details provided on the Description page.
This competition has completed. This leaderboard reflects the final standings.

Raw Data Refresh

In the money Gold Silver Bronze

#	pub	Team Name	Kernel	Team Members	Score ⚡	Entries	Last
1	→ 2622	[ods.ai] Toxicology			0.94734	2	1d
2	→ 2449	Linerider (XIEURB)			0.94720	2	1d
3	→ 2584	FJLS.DY			0.94707	2	23d
4	→ 2150	COMBAT WOMBAT			0.94706	2	1d
5	→ 2624	vecxox			0.94683	2	1d
6	→ 2409	yurikr			0.94678	2	1d
7	→ 2470	[DSU] (kaggle-ja) PTFAP			0.94660	2	23d
8	→ 2298	Gishen Ha			0.94660	2	1d
9	→ 2416	Kaz&Kan			0.94650	2	1d
10	→ 2393	Harness the beasts (HBB)			0.94649	2	23d
11	→ 2331	tosa_tobis			0.94635	2	1d
12	→ 2056	AAA Team			0.94634	2	1d
13	→ 2403	zhengtian			0.94633	2	1d

Installing the tutorials

github.com/ablaom/MLJTutorial

turing.ac.uk
@turinginst