PARIS

MEETUP 04/08/2022

kaggle days

CHAMPIONSHIP

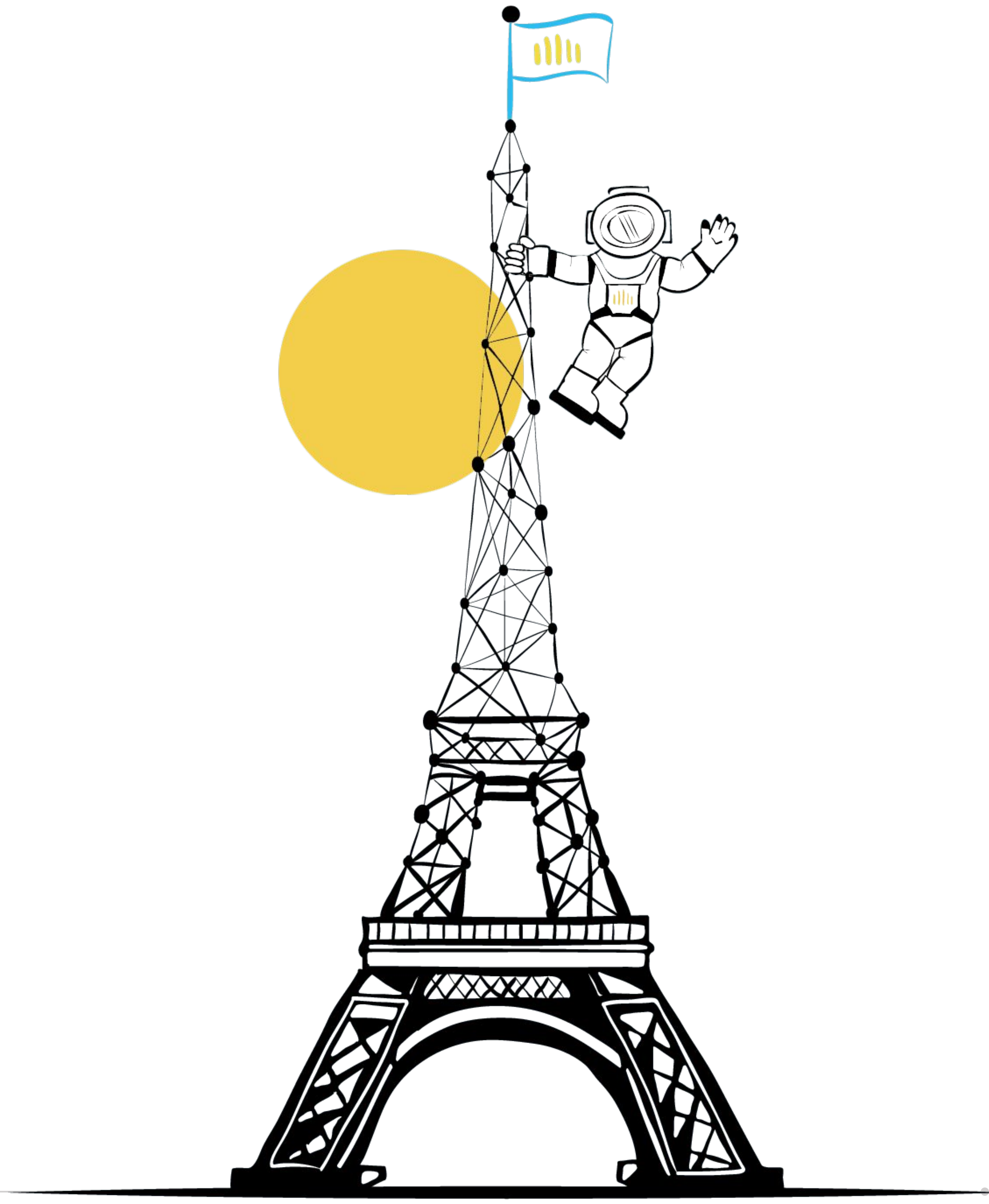
WITH Z hp

Data Centric AI Approach & Sustainability of AI

Amed Coulibaly
Ayon Roy

Agenda

- Basic Components of AI Systems
- Data Centric & Model Centric Approaches
- Need for Data Centric Approach
- Data Centric AI Approach in Kaggle Competitions
- Building Sustainable AI Solutions

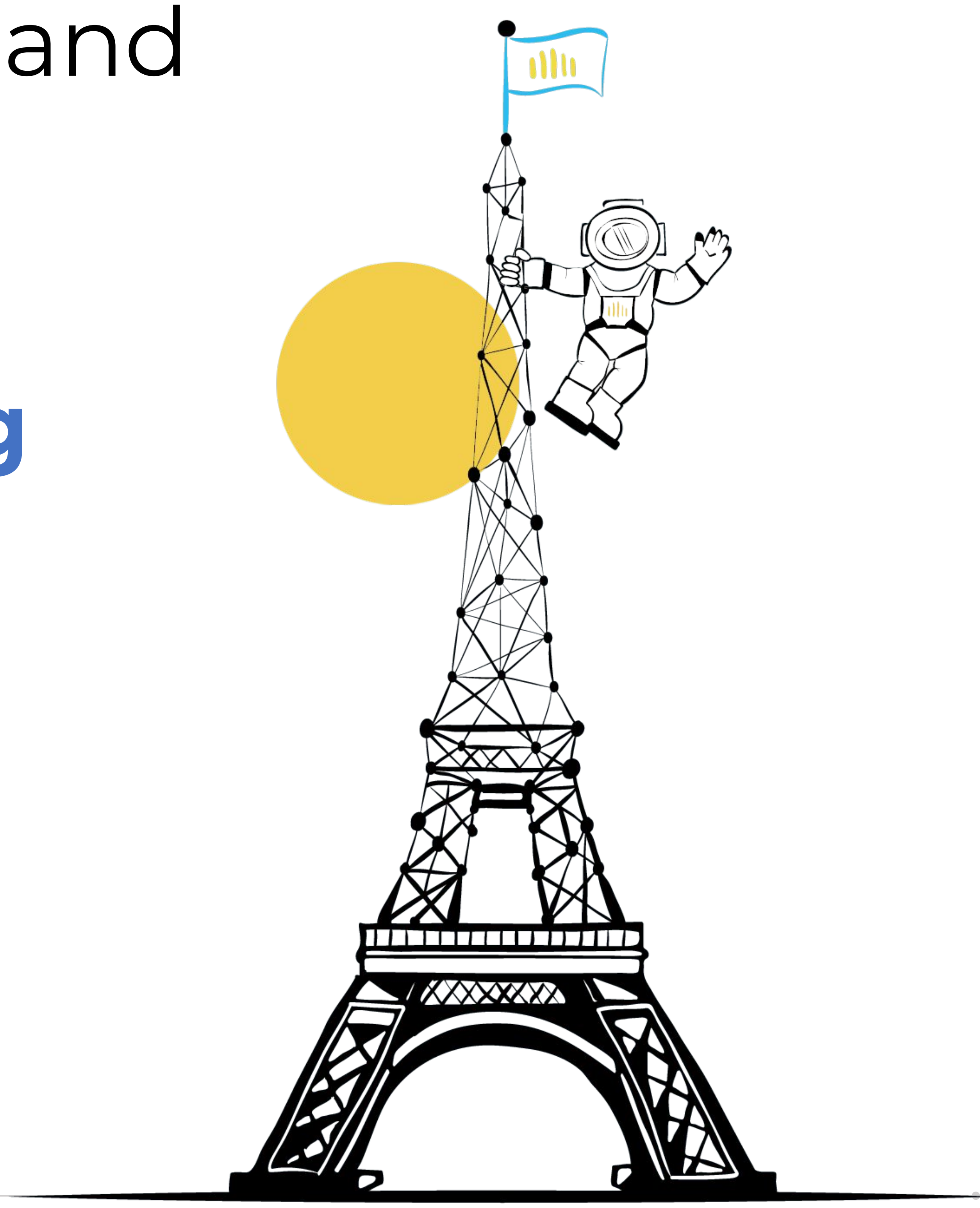


AI Systems

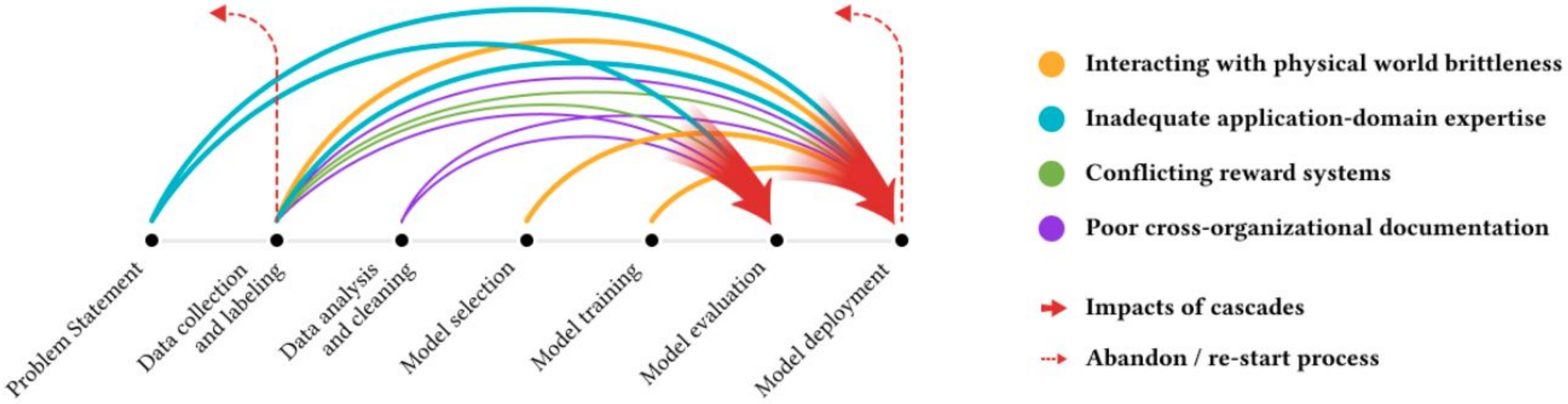
Artificial Intelligence Systems are projects which are undertaken with the long-term goal of simulating the human brain in real time, complete with artificial consciousness and artificial general intelligence.

How do we simulate the human brain in real time & bring artificial consciousness ?

Data + Model



Stages to build AI Systems



"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI

https://storage.googleapis.com/pub-tools-public-publication-data/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf

Model Centric & Data Centric Approach

kaggle days CHAMPIONSHIP WITH Z

Model-Centric Approach

This involves designing empirical tests around the model to improve the performance. This consists of finding the right model architecture and training procedure among a huge space of possibilities.

Data-centric approach

This consists of systematically changing/enhancing the datasets to improve the accuracy of your AI system. This is usually overlooked and data collection is treated as a one off task.

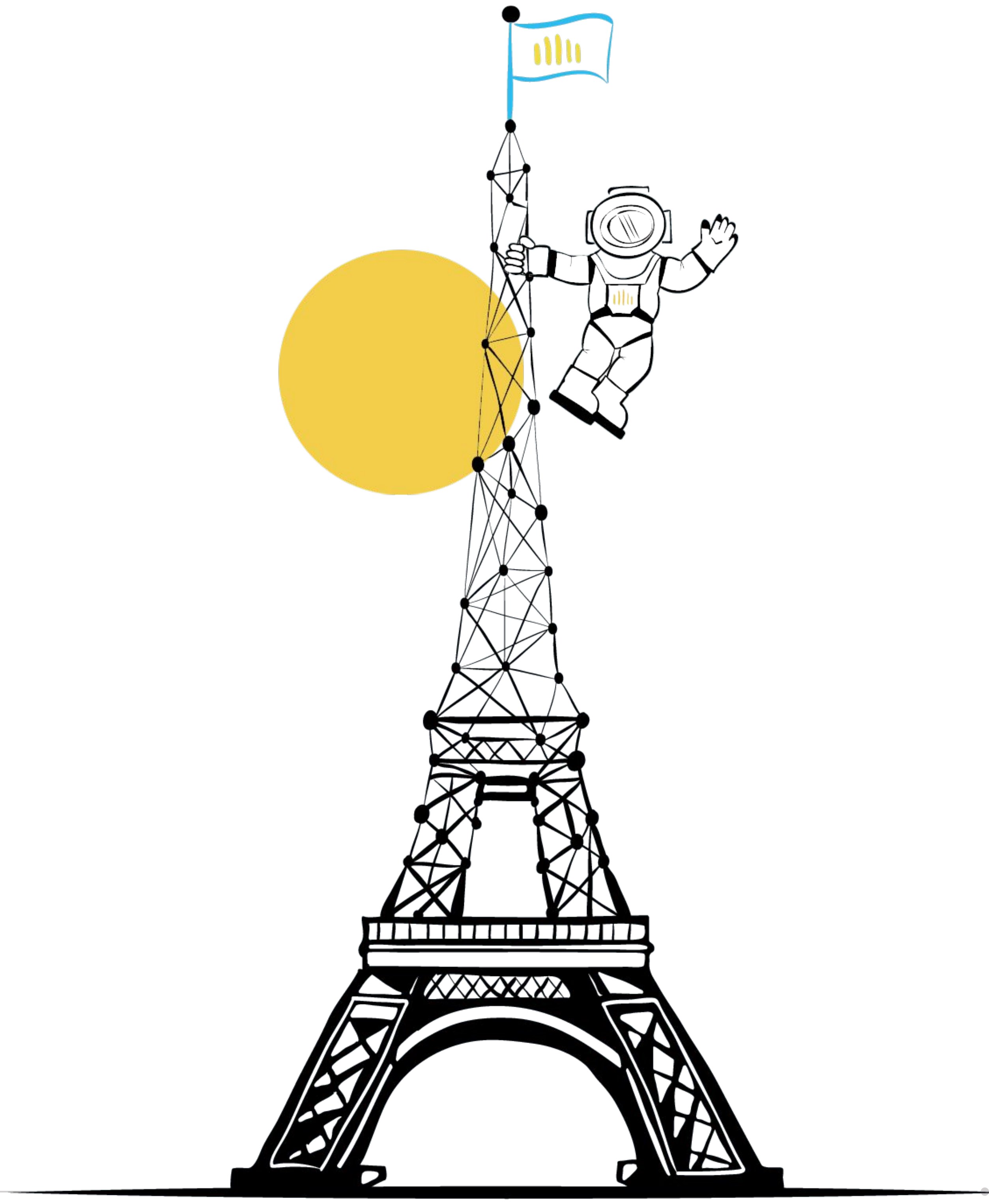
https://towardsdatascience.com/from-model-centric-to-data-centric-artificial-intelligence-77e423f3f593#:~:text=Data%2Dcentric%20approach,as%20a%20one%20off%20task

Community’s Bias towards Model Centric Approach

The steel sheets defect detection was one of the examples brought during the session — assuming a series of images from steel sheets we want to develop the best model to detect these defects that can happen during the process of steel sheets manufacturing. There are 39 different defects that we want to be able to identify. By developing a computer vision model with well-tuned hyperparameters, it was able to reach a **76.2% accuracy baseline system**, but the goal is to achieve **90% accuracy**. *How can this be done*?

Steel Sheets Detection Challenge

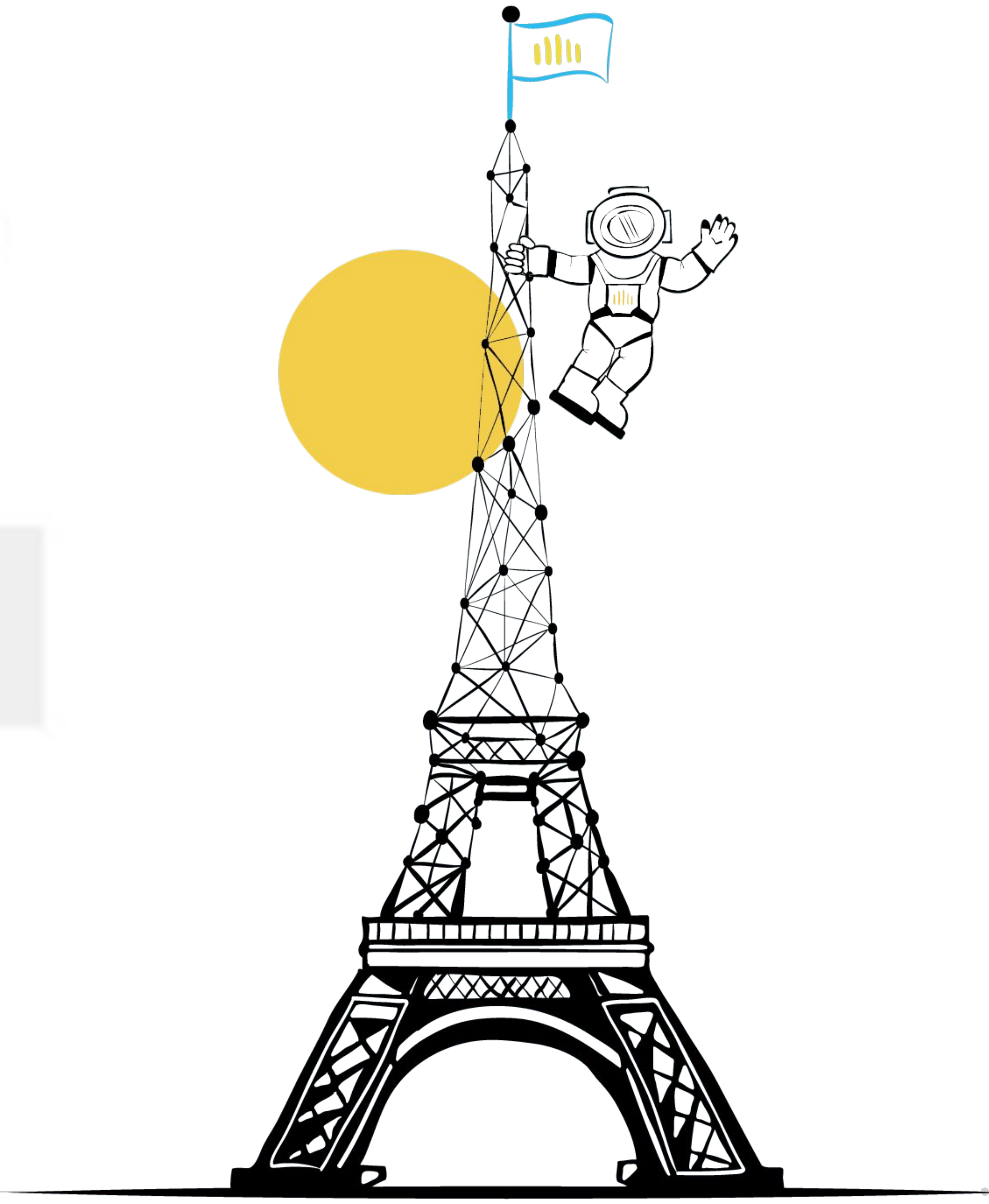
https://www.youtube.com/watch?v=06-AZXmwHjo&t=148s



Difference in Results

Knowing that the baseline model was already good, the task to have it improved to achieve 90% accuracy sound almost impossible — for the model-centric, the improvements based on Network Architecture search and using the state-of-the-art architectures, whereas, for the data-driven, the approach taken was to identify inconsistencies and clean noisy labels. The results were mind-blowing:

Steel sheets defects detection	Baseline	Model-centric	Data-centric
Accuracy	76.2%	+0% (76.2%)	+16.9% (93.1%)

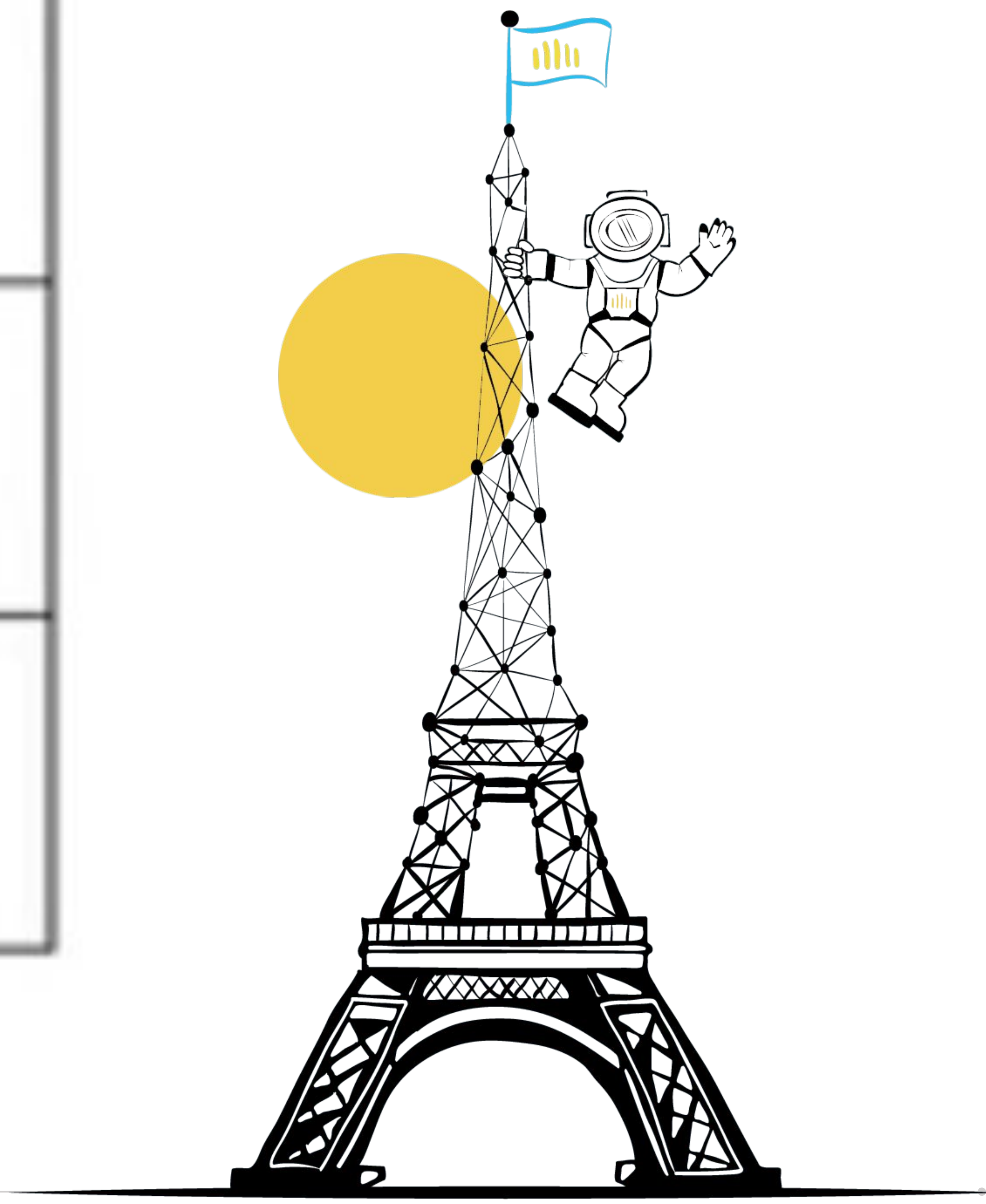


Importance of Data Centric Approaches

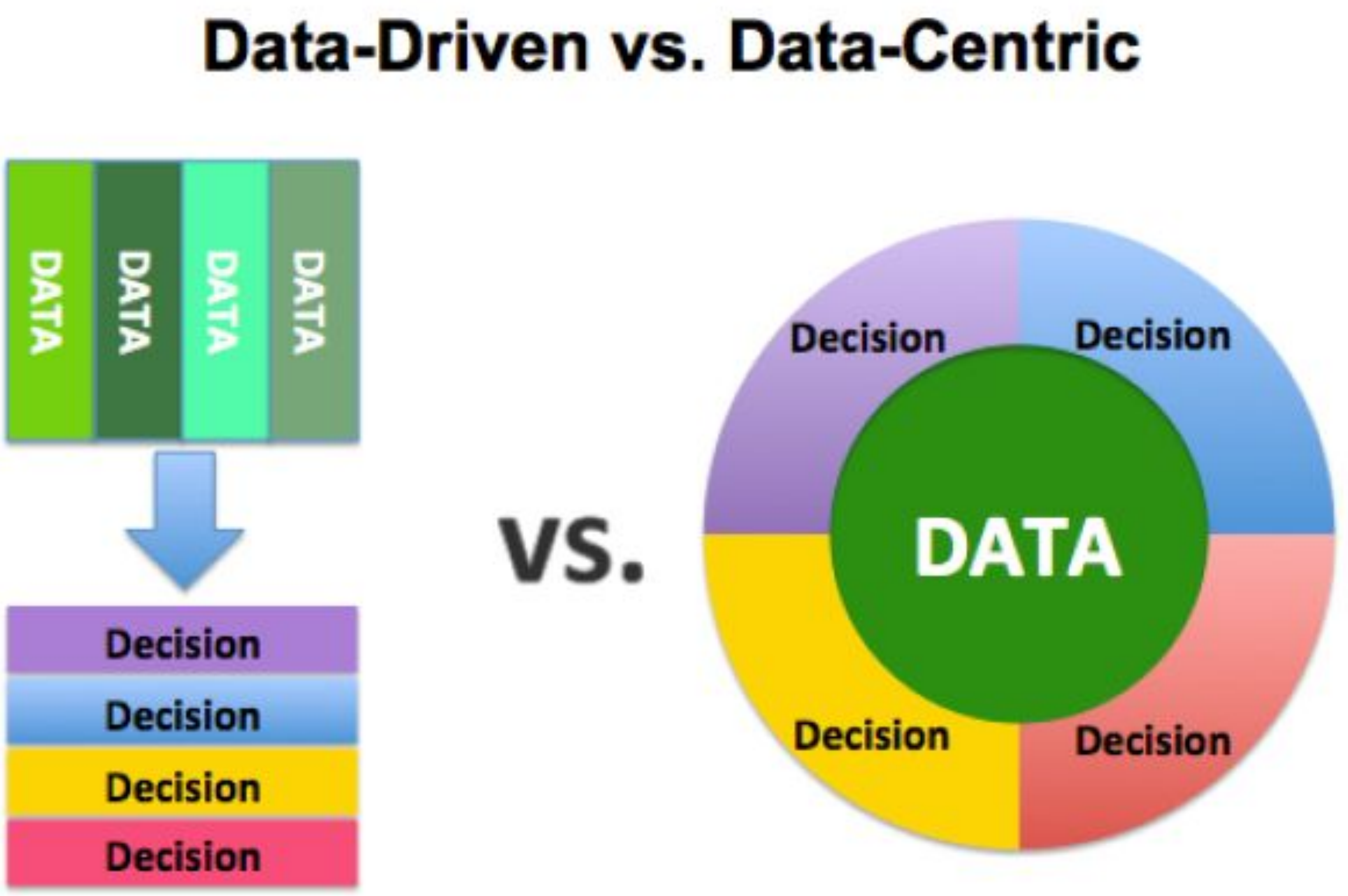
kaggle days
CHAMPIONSHIP WITH Z

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

https://www.youtube.com/watch?v=06-AZXmwHjo&t=324s



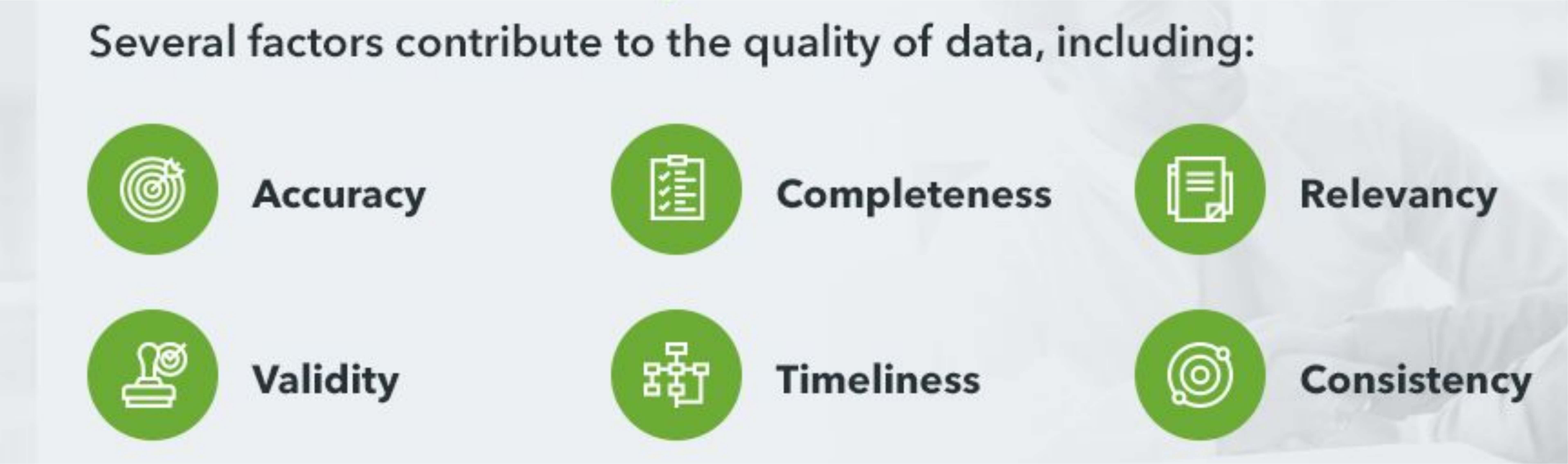
Beware of the Trade Off..



https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning

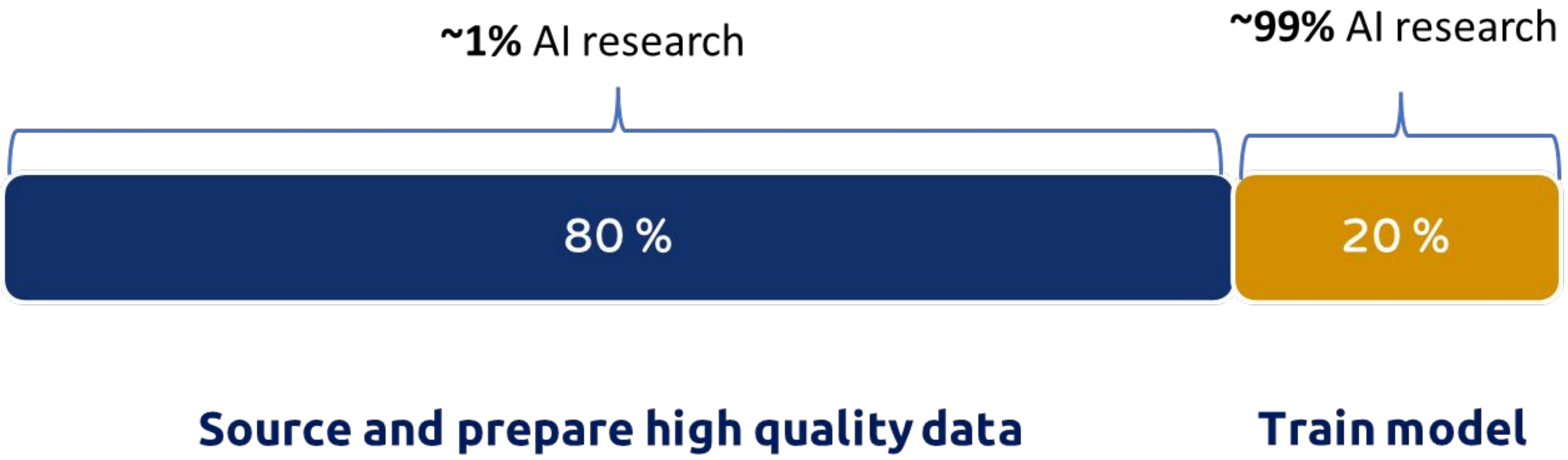
But..

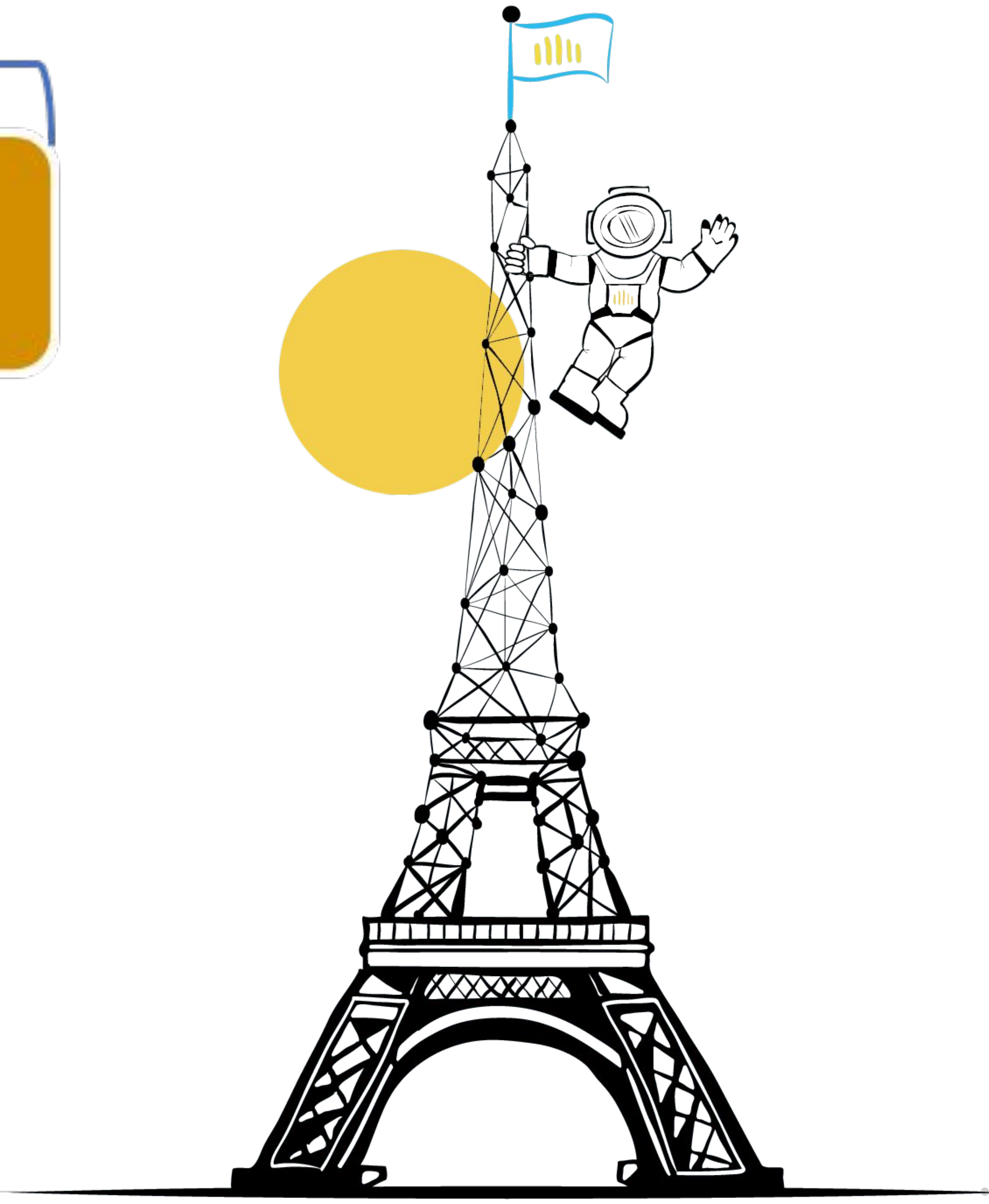
Check for Data Quality



https://www.lotame.com/why-is-data-quality-important/

Need of Data Centric Approaches





Model Centric AI Approaches in Kaggle Competitions

kaggle days
CHAMPIONSHIP
WITH Z

Model Centric AI

1

Data

Data Explorer

20.64 MB

- test
- train
- sample_submission.csv
- test.csv
- train.csv

2

Preprocessing

3

Modeling

Training → Evaluation → Hyperparameter Tuning → New Model

4

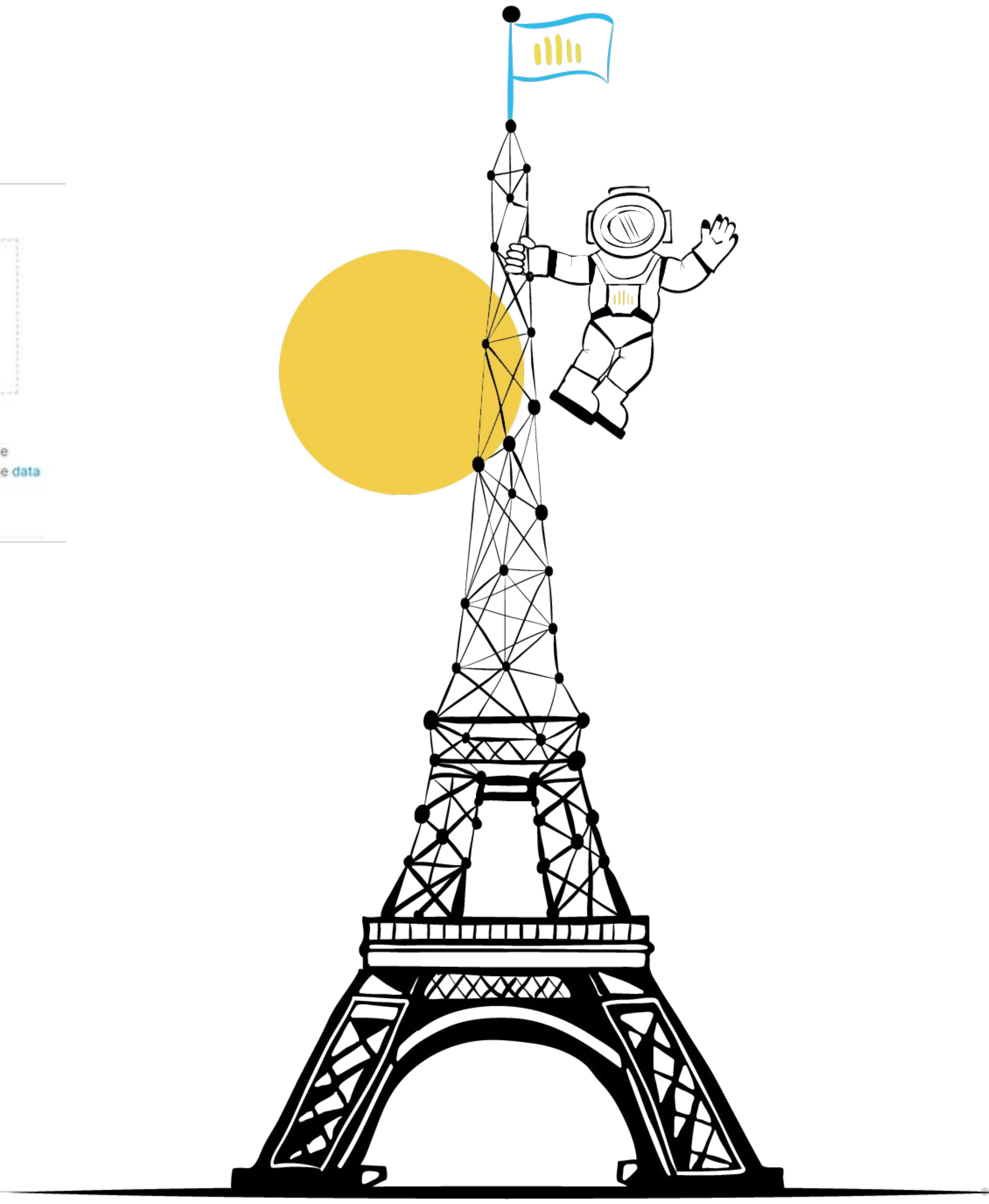
Submit Predictions

File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions

We expect the solution file to have 924621 prediction rows. This file should have a header row. Please see sample submission file on the data page.



Data Centric AI Approaches in Kaggle Competitions

kaggle days CHAMPIONSHIP WITH Z

Data Centric AI

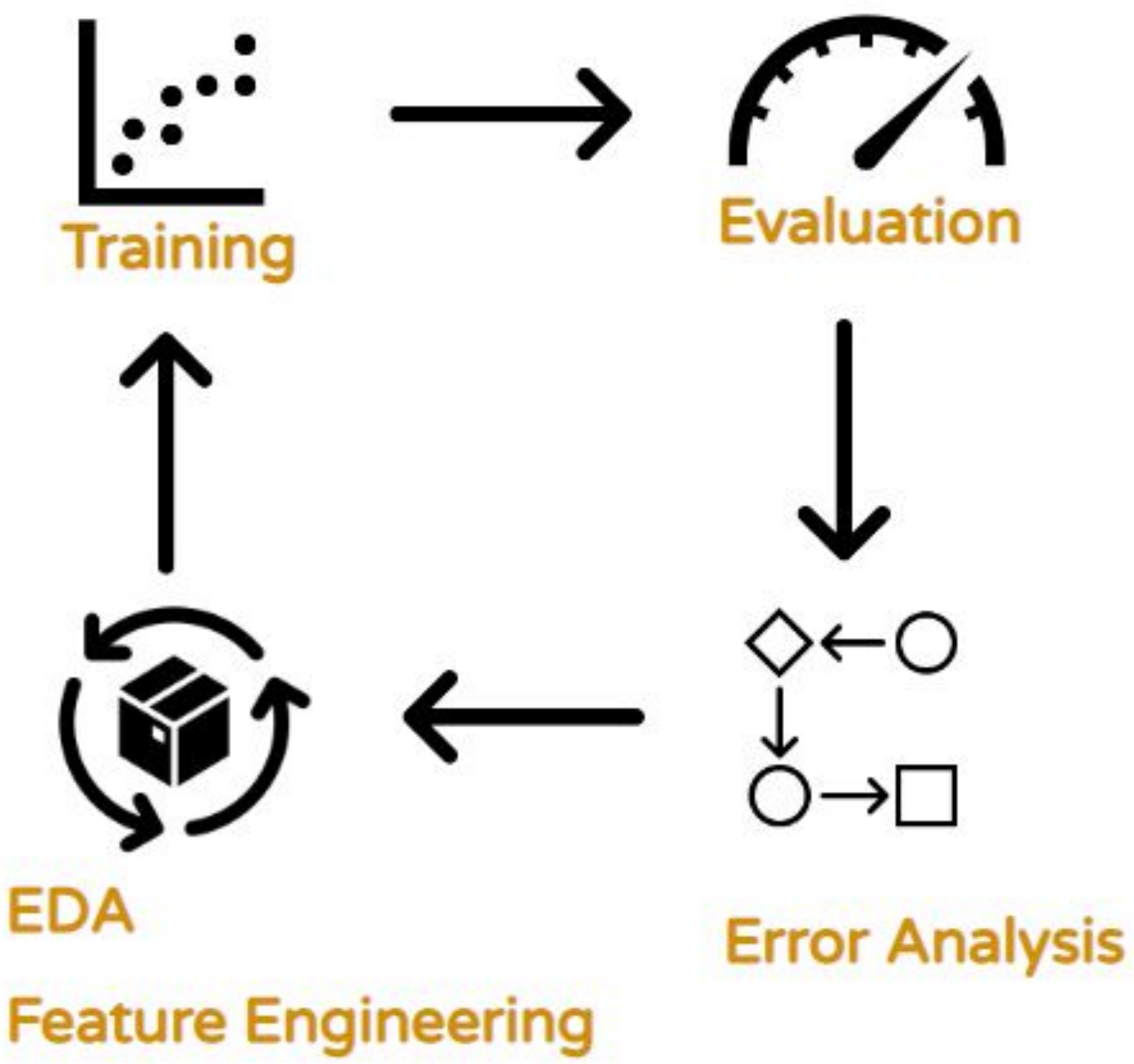
Data

Data Explorer

20.64 MB

- test
- train
- sample_submission.csv
- test.csv
- train.csv

Modeling



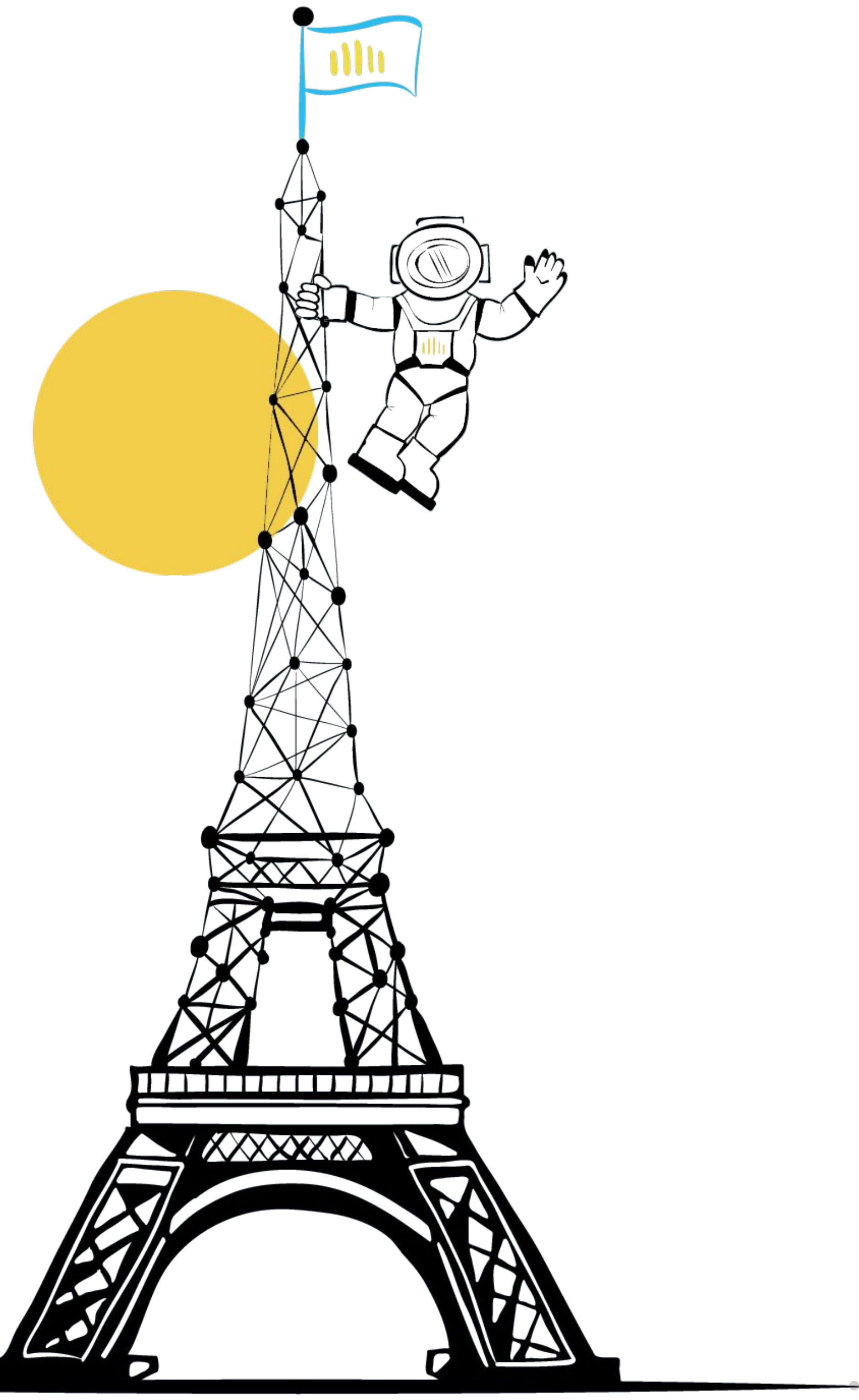
Submit Predictions

File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

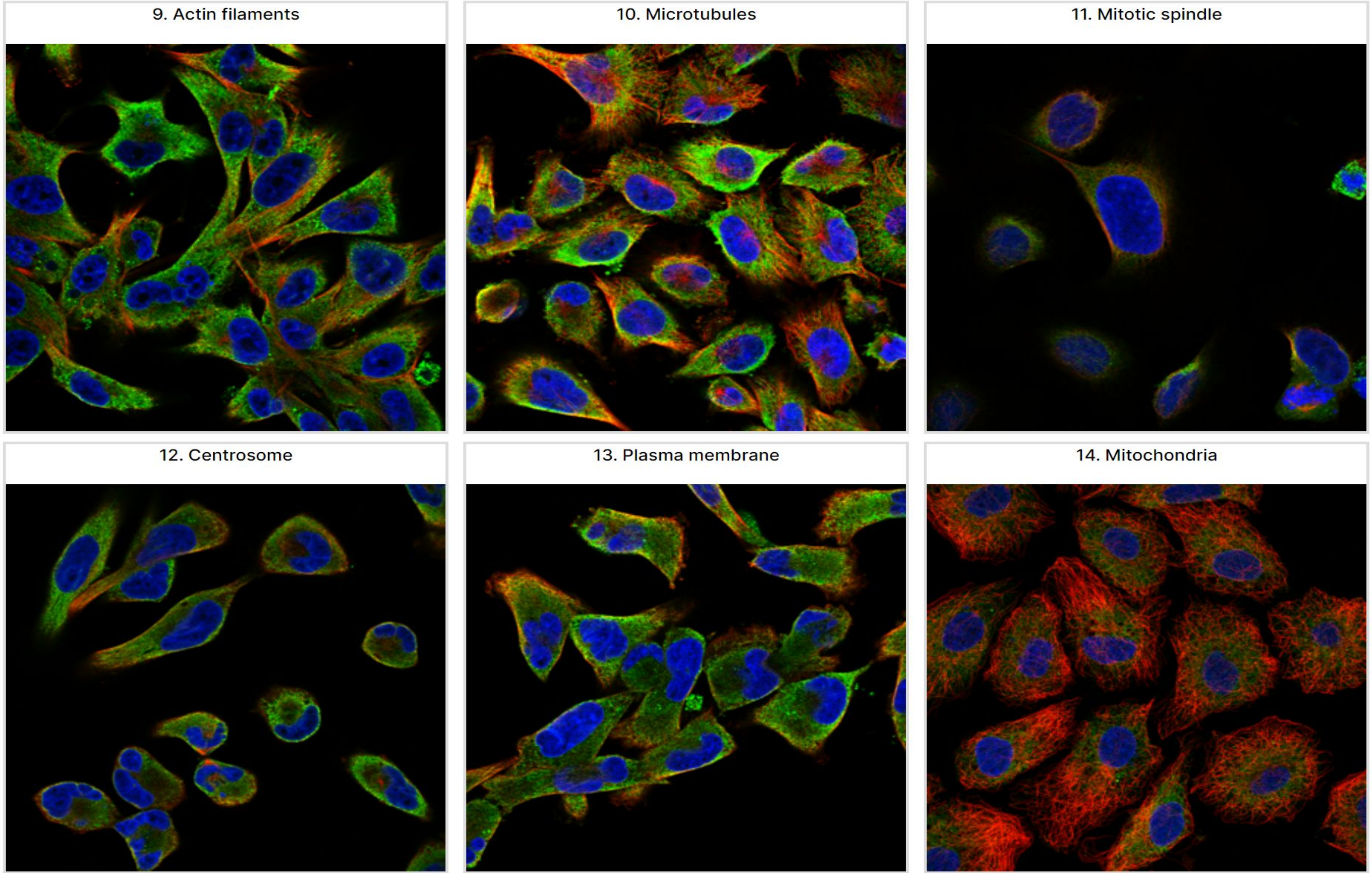
Number of Predictions

We expect the solution file to have 924621 prediction rows. This file should have a header row. Please see sample submission file on the data page.



Human Protein Atlas – Single Cell Classification [Kaggle Competition]

Data

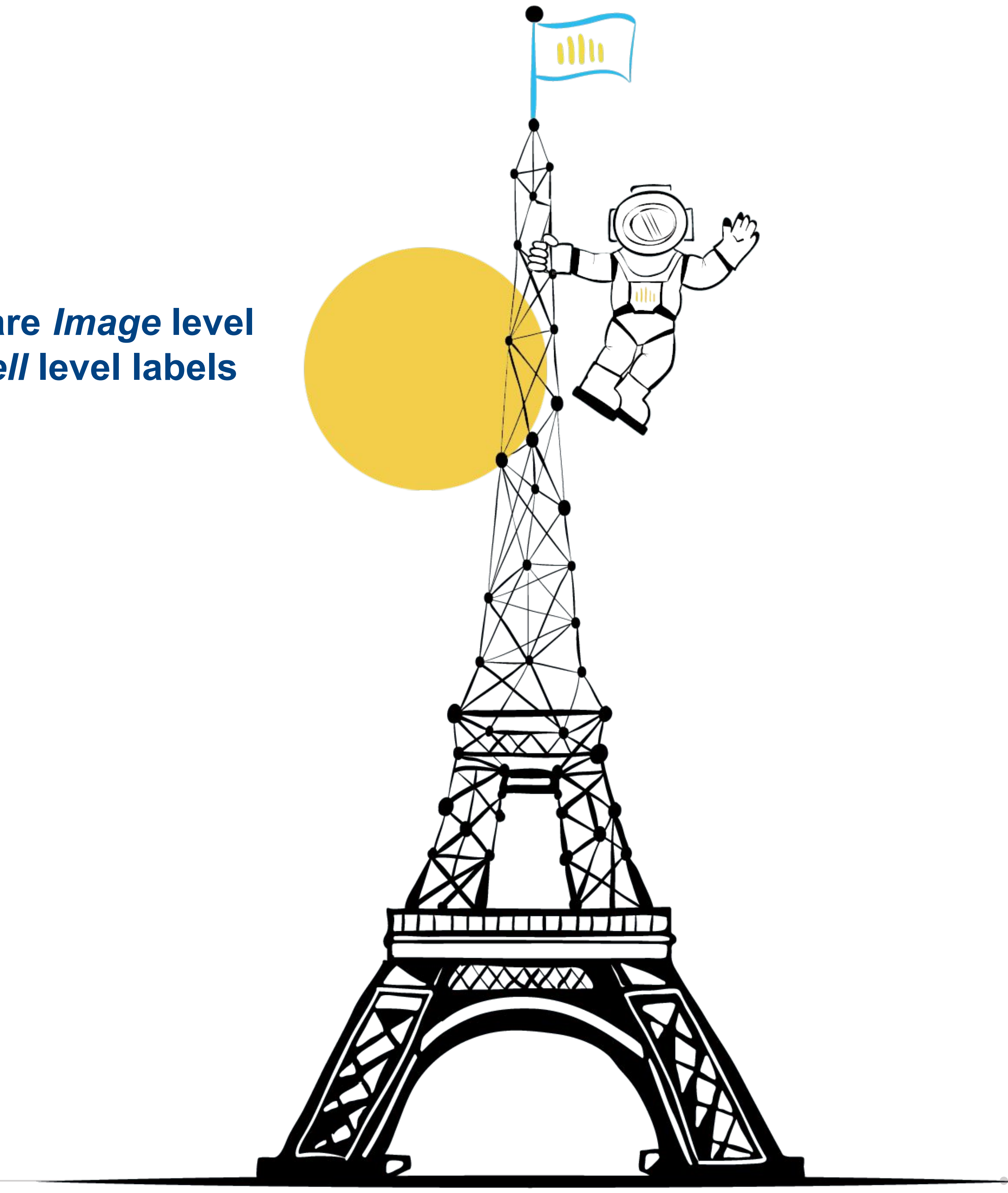


Task

Segment the cells in the images
and predict the labels of those segmented cells

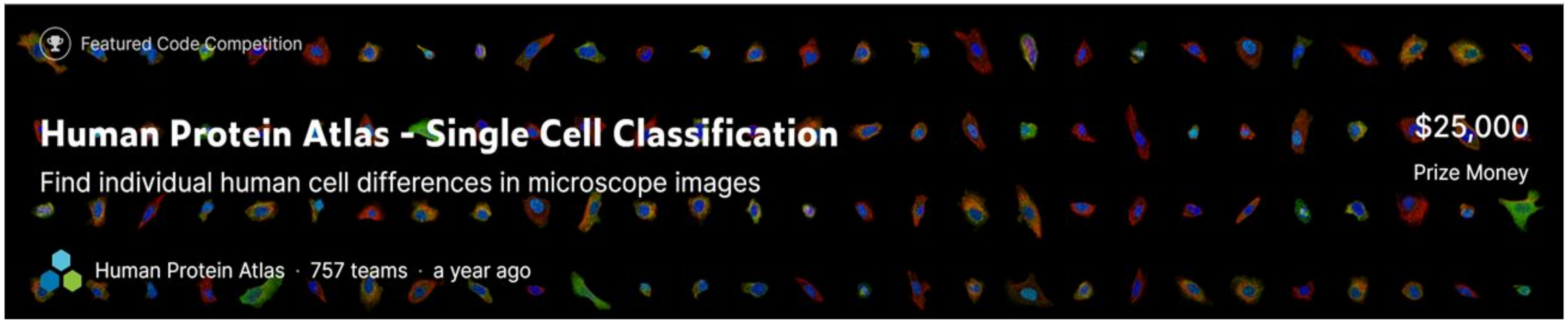
Challenge

The labels you will get for training are *Image* level labels while the task is to predict *cell* level labels

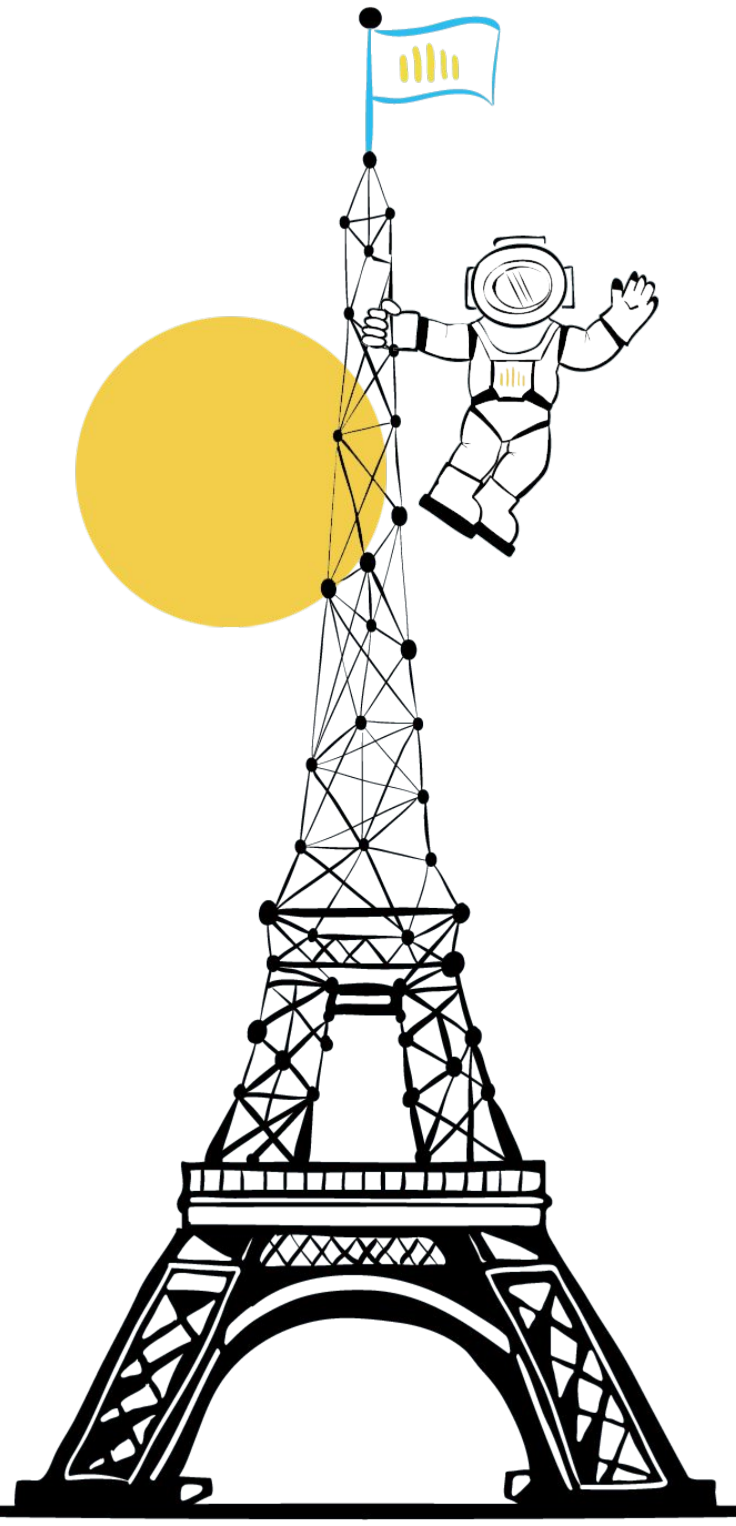


Kaggle Competition Leaderboard

kaggle days
CHAMPIONSHIP
WITH Z



#	△	Team	Members	Score	Entries	Last	Code
1	▲ 4	bestfitting		0.56670	480	1y	
2	—	[red.ai]		0.55328	459	1y	<>
3	—	MPWARE & ZFTurbo & Dieter		0.54995	500	1y	<>
4	▲ 2	MILIMED		0.54389	258	1y	



Kaggle Competition Solution Approach

kaggle days
CHAMPIONSHIP
WITH Z

CroDoc
Topic Author
4th place

4th Place Solution: MILIMED

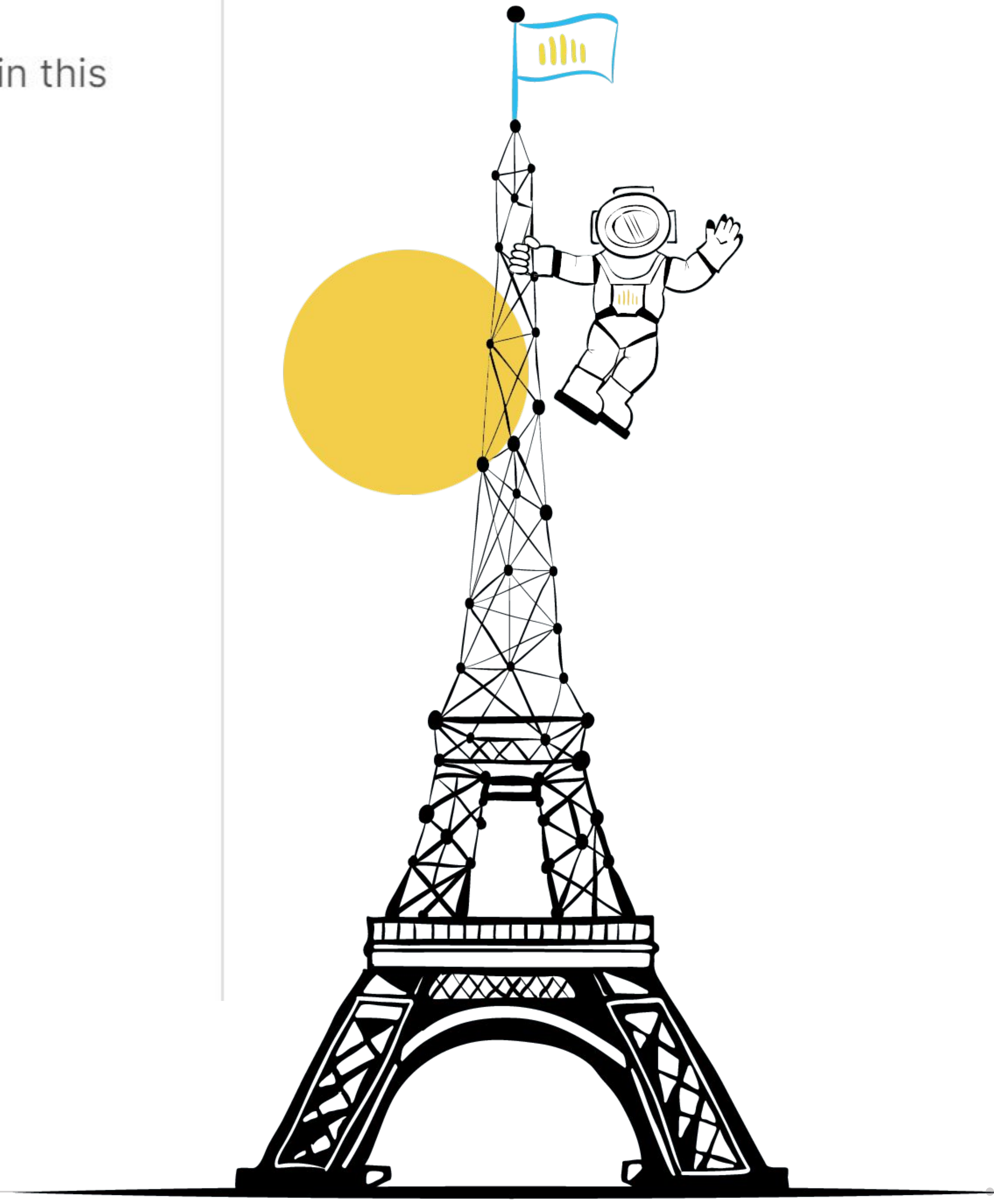
Posted in hpa-single-cell-image-classification a year ago

15

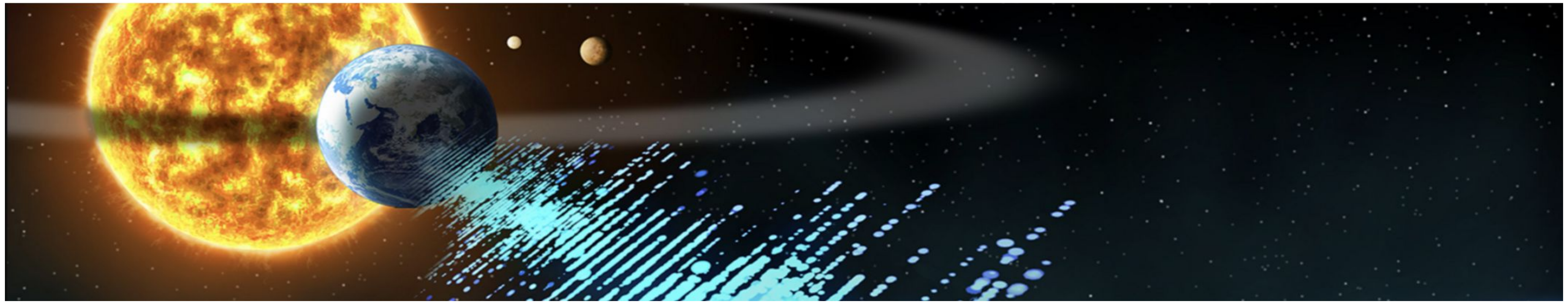
We are a very diverse team of computer scientists and medical doctor/students. It was our great pleasure to participate in this demanding challenge. Hope some of you find this solution useful and/or interesting.

Solution overview

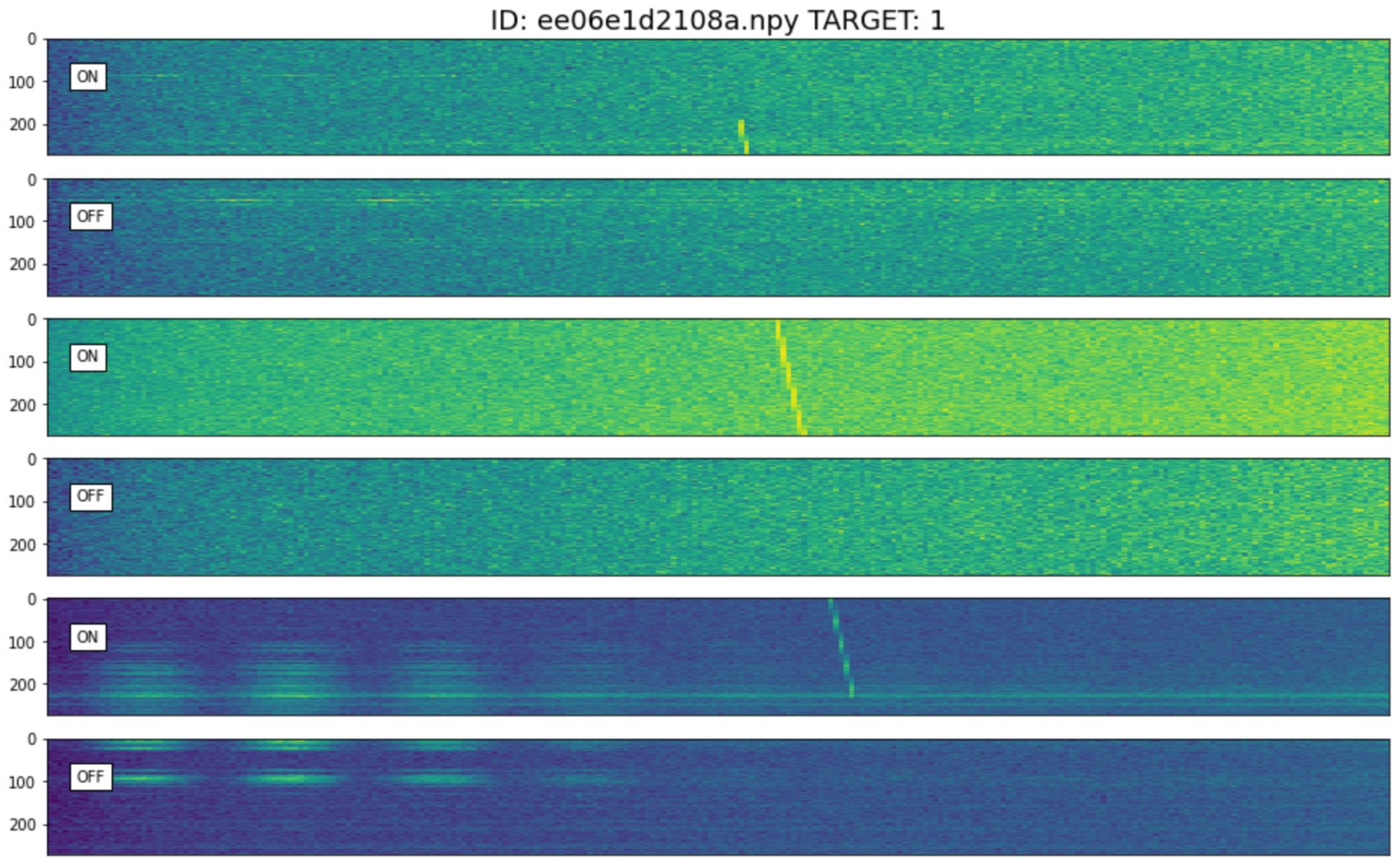
1. Segmentation → HPA-Cell-Segmentation
2. Dataset → 512×512 cell images (20% removed)
3. Parallelization → speed-up → 3h left for inference
4. Manual Labeling → smaller classes & validation (soft labels)
5. Pseudo-Labeling → negative labeling (& positive for mitotic spindle)
6. EfficientNetB0 Ensemble + semi-balanced data sampling
7. Fine-tuning → on manually labeled & non-labeled validation data
8. Cell/Image Weighting → final confidence = 0.7 * cell_confidence + 0.3 * image_confidence

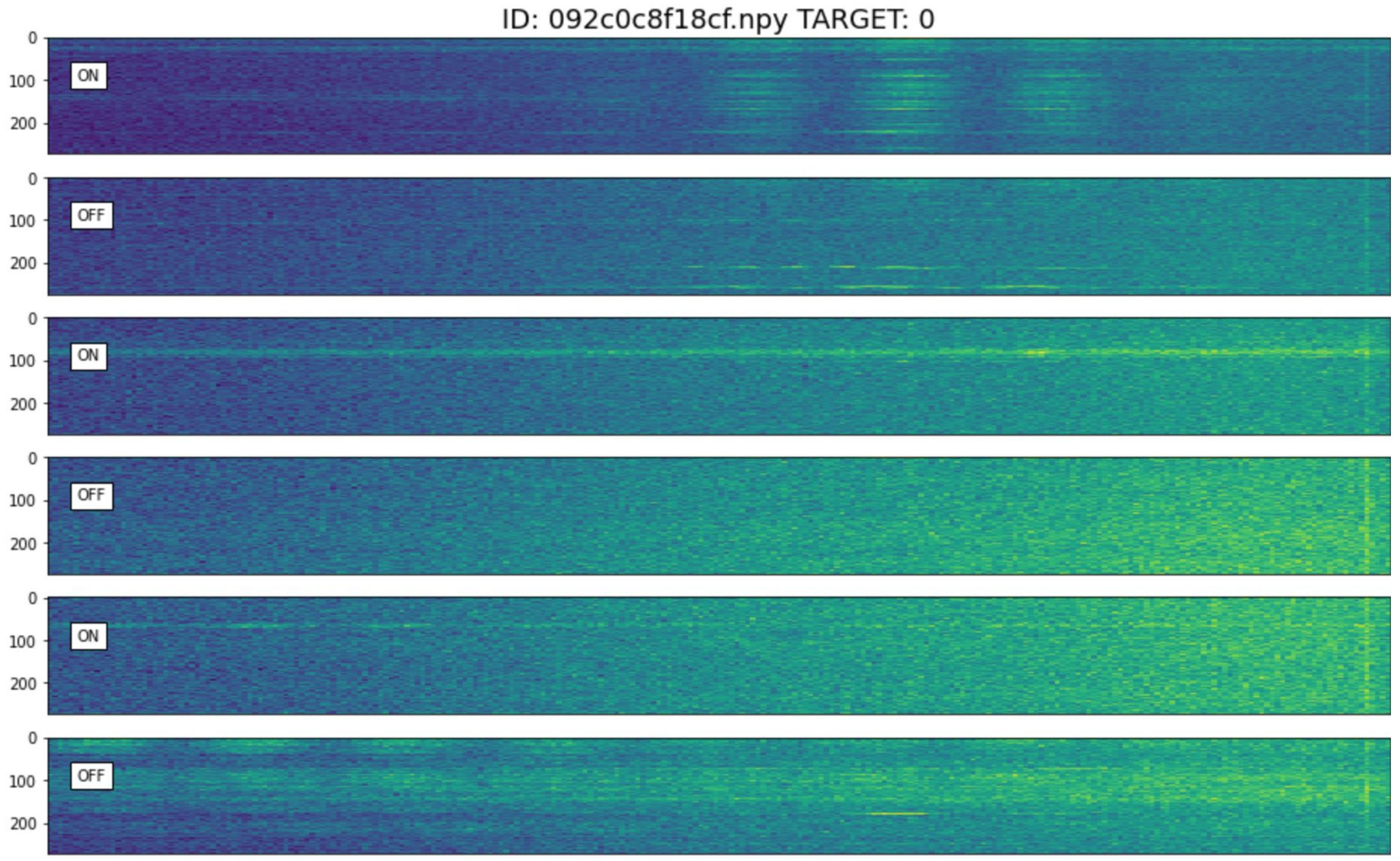


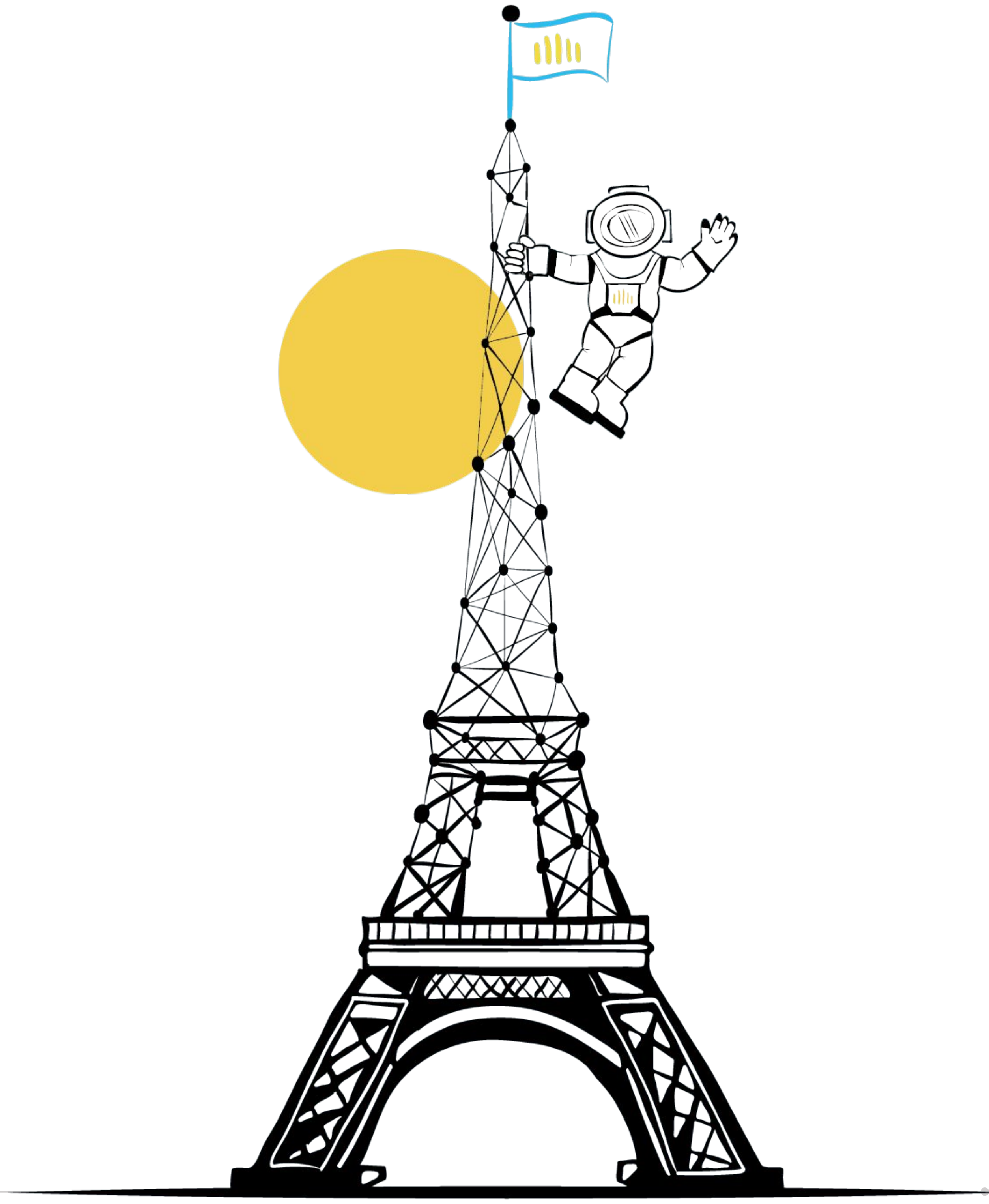
SETI Breakthrough Listen - E.T. Signal Search [Kaggle Competition]



Data

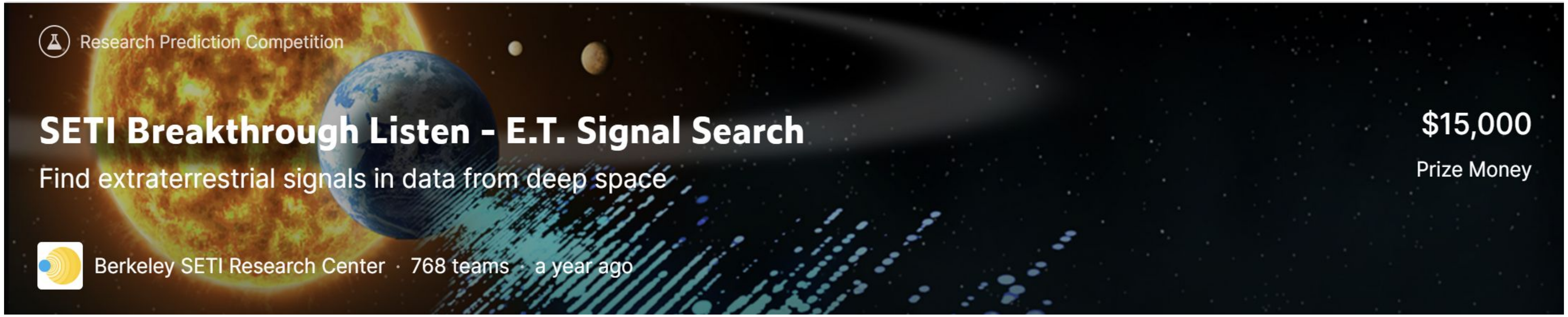




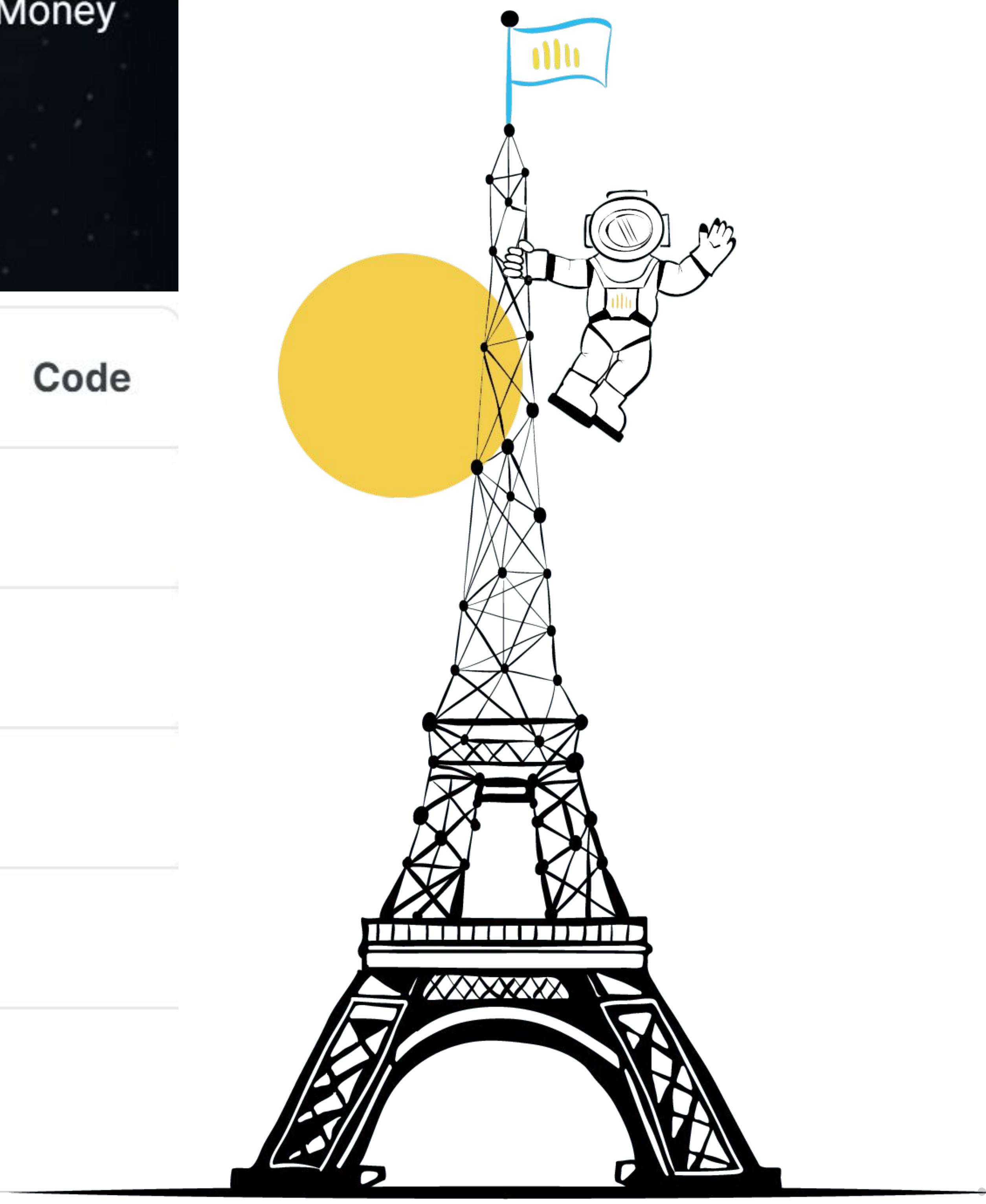


Kaggle Competition Leaderboard

kaggle days
CHAMPIONSHIP
WITH



#	△	Team	Members	Score	Entries	Last	Code
1	—	Watercooled		0.96782	93	1y	
2	—	未知との遭遇		0.81206	85	1y	
3	—	knj		0.80475	77	1y	
4	▲ 2	Steven Signal		0.80428	92	1y	



Kaggle Competition Solution Approach

1st Place Solution

Posted in seti-breakthrough-listen a year ago

268

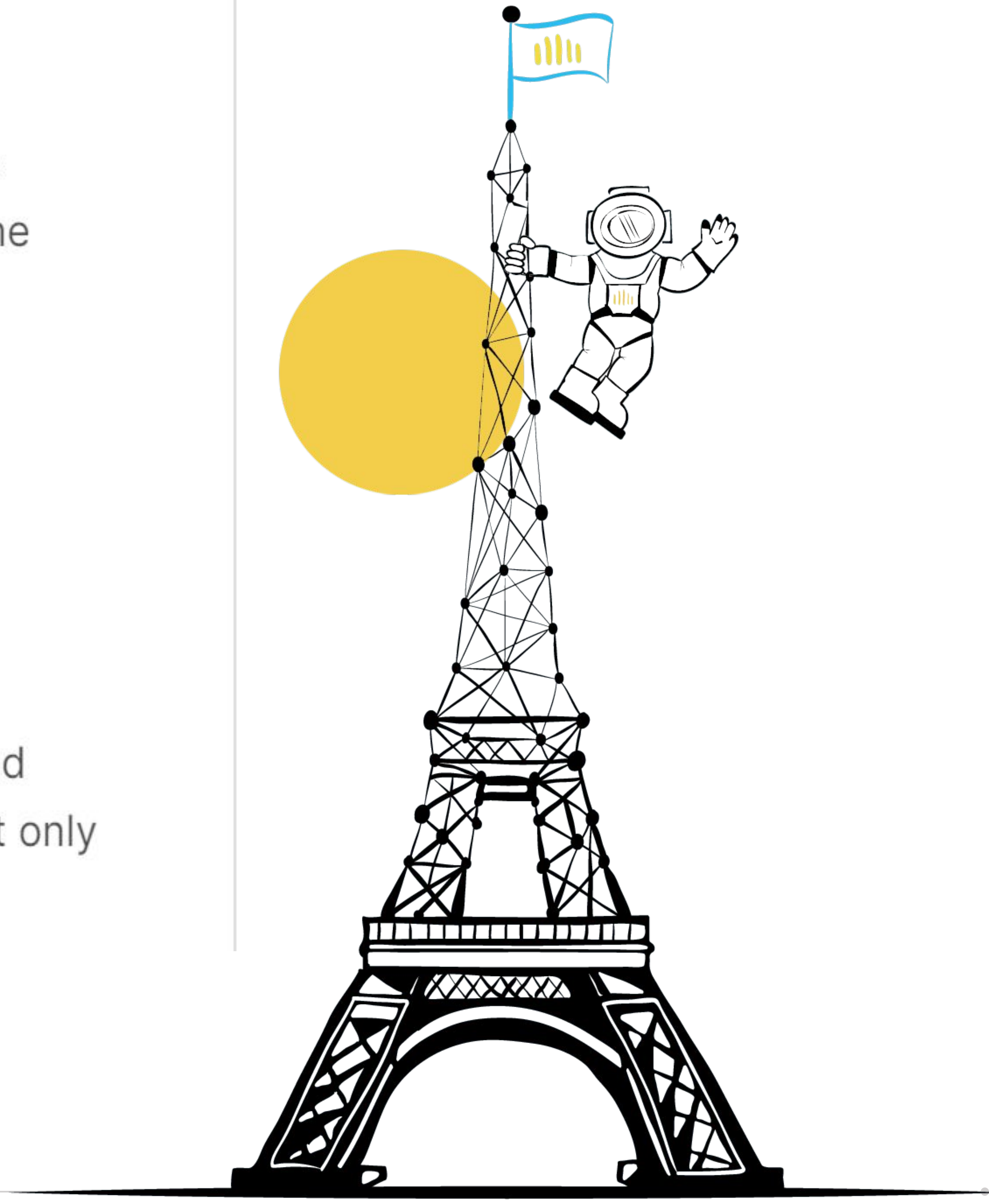
Ilu
Topic Author
1st place

Thanks to Kaggle and Berkeley SETI Research Center for this interesting competition. In the following, we want to give a summary of the winning solution of Team Watercooled. As always, thanks to all team members contributing equally to the solution.

@philippsinger @christofhenkel @ilu000

Summary

Our solution is based on large state-of-the-art classification models that were fine-tuned for this specific task. We pre-processed images by cleaning the backgrounds to boost signal to noise ratio. During training, we employed heavy augmentation in the form of Mixup. For faster training iterations, we only used the ON-channels. We only rely on provided competition data and do not utilize any external data. We additionally augment the training data with an extra signal that only appears in test files -- an "s-shape" signal -- by using a randomized signal generator.



Tabular Playground Series – January 2022 [Kaggle Competition]

kaggle days
CHAMPIONSHIP
WITH

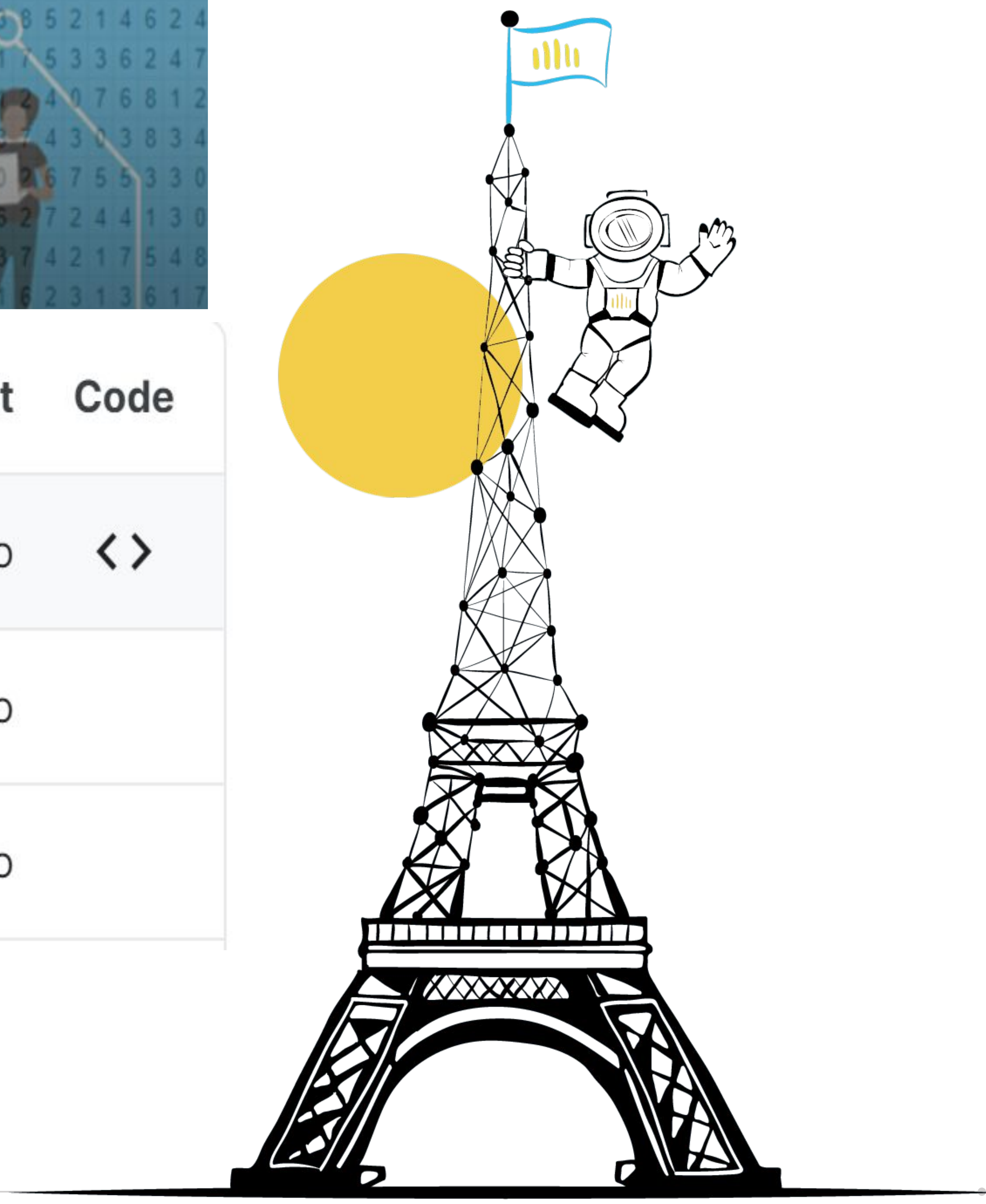
	row_id	date	country	store	product	num_sold
0	0	2015-01-01	Finland	KaggleMart	Kaggle Mug	329
1	1	2015-01-01	Finland	KaggleMart	Kaggle Hat	520
2	2	2015-01-01	Finland	KaggleMart	Kaggle Sticker	146
3	3	2015-01-01	Finland	KaggleRama	Kaggle Mug	572
4	4	2015-01-01	Finland	KaggleRama	Kaggle Hat	911

Data

Kaggle Competition Leaderboard

kaggle days
CHAMPIONSHIP
WITH Z

#	△	Team	Members	Score	Entries	Last	Code
1	▲ 129	AmbrosM		4.58941	67	6mo	<>
2	▲ 273	parth.tiwary		4.63253	13	7mo	
3	▲ 373	ZhangOS		4.63648	2	6mo	



PARIS
MEETUP 04/08/2022

Kaggle Competition Solution Approach

AmbrosM
Topic Author
1st place

#1 Solution Description: Advanced Linear Model

Posted in tabular-playground-series-jan-2022 6 months ago

103

The following lines describe the development of my final submission to this competition.

Importance of cross-validation

In this January TPS one could practice ignoring the public leaderboard. The public leaderboard is based on the first quarter of 2019, but all the interesting holidays occur in April of 2019 or later. This means that the public leaderboard gives no information at all about the quality of a model's holiday features. You can over- or underestimate the influence of Easter, Midsummer Day, National Day, Christmas and so on - for the public leaderboard it doesn't matter. The public leaderboard is only good to verify whether the model deals correctly with the yearly GDP.

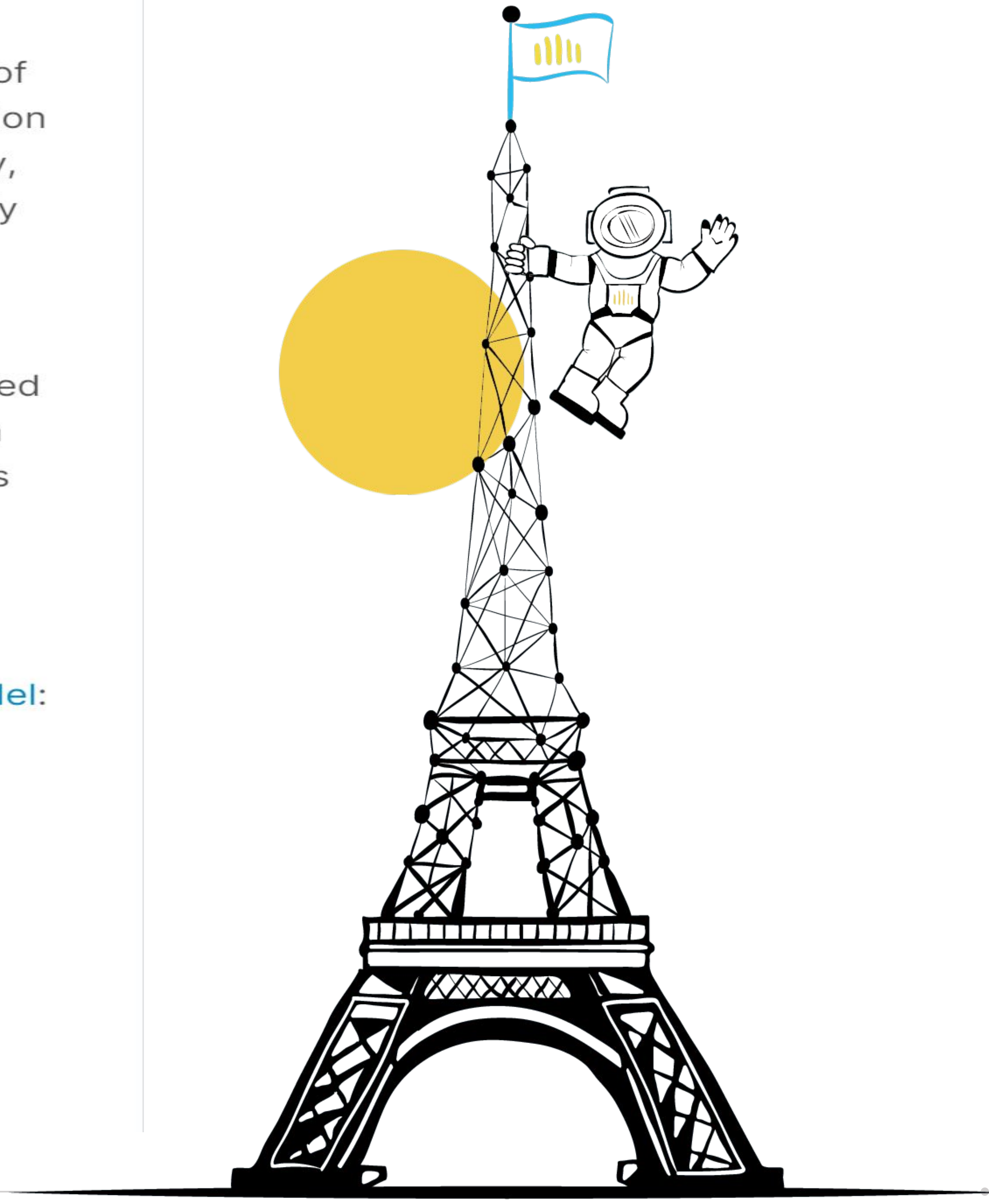
For this reason, I focused on cross-validation (`GroupKFold` with the years as groups), and in the cross-validation results, I evaluated the SMAPE for January through March separately from SMAPE for the rest of the year. Then I consistently optimized my model for the latter. For the final evaluation, I submitted the two notebooks with the best cv. The winning notebook has a public lb score of only 4.11991, which would rank it at position 306 of the public lb. It took quite some courage to mark this as the final submission...

Feature engineering

My final notebook still uses Ridge regression with a log-transformed target, but the features differ from my earlier linear model:

- The selection of Fourier coefficients has changed; the stickers get no Fourier coefficients at all (this means that the prediction for the stickers is constant over the whole year).
- There are small changes in the length of holidays.
- The Easter holiday in Norway differs from the Easter holiday in the other two countries.
- I added the OECD's consumer confidence index as external data, as suggested in this discussion.

All these features were found by a detailed analysis of the residuals.



Approaching Sustainability as you build AI Systems

kaggle days
CHAMPIONSHIP
WITH

One consequence of this increase in computing is the heavy environmental impact of training machine learning models. A recent research paper — Energy and Policy Considerations for Deep Learning in NLP — notes that an inefficiently trained NLP model using Neural Architecture Search can emit **more than 626,000 pounds of CO_2.** That's about **five times the lifetime emissions of an average American car!**

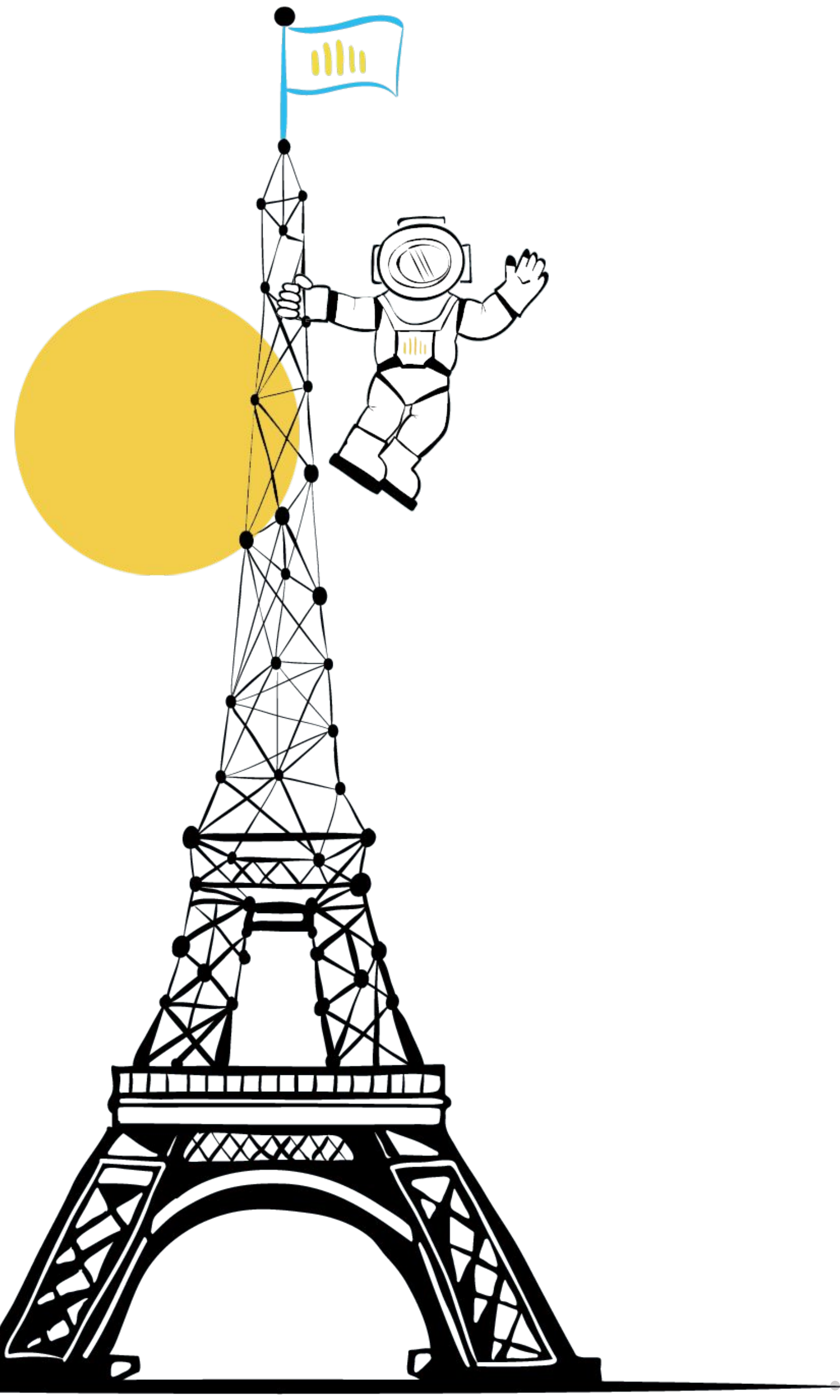
https://wandb.ai/amanarora/codecarbon/reports/Tracking-CO2-Emissions-of-Your-Deep-Learning-Models-with-CodeCarbon-and-Weights-Biases--VmlldzoxMzM1NDg3

Comparison of Certain NLP Models

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

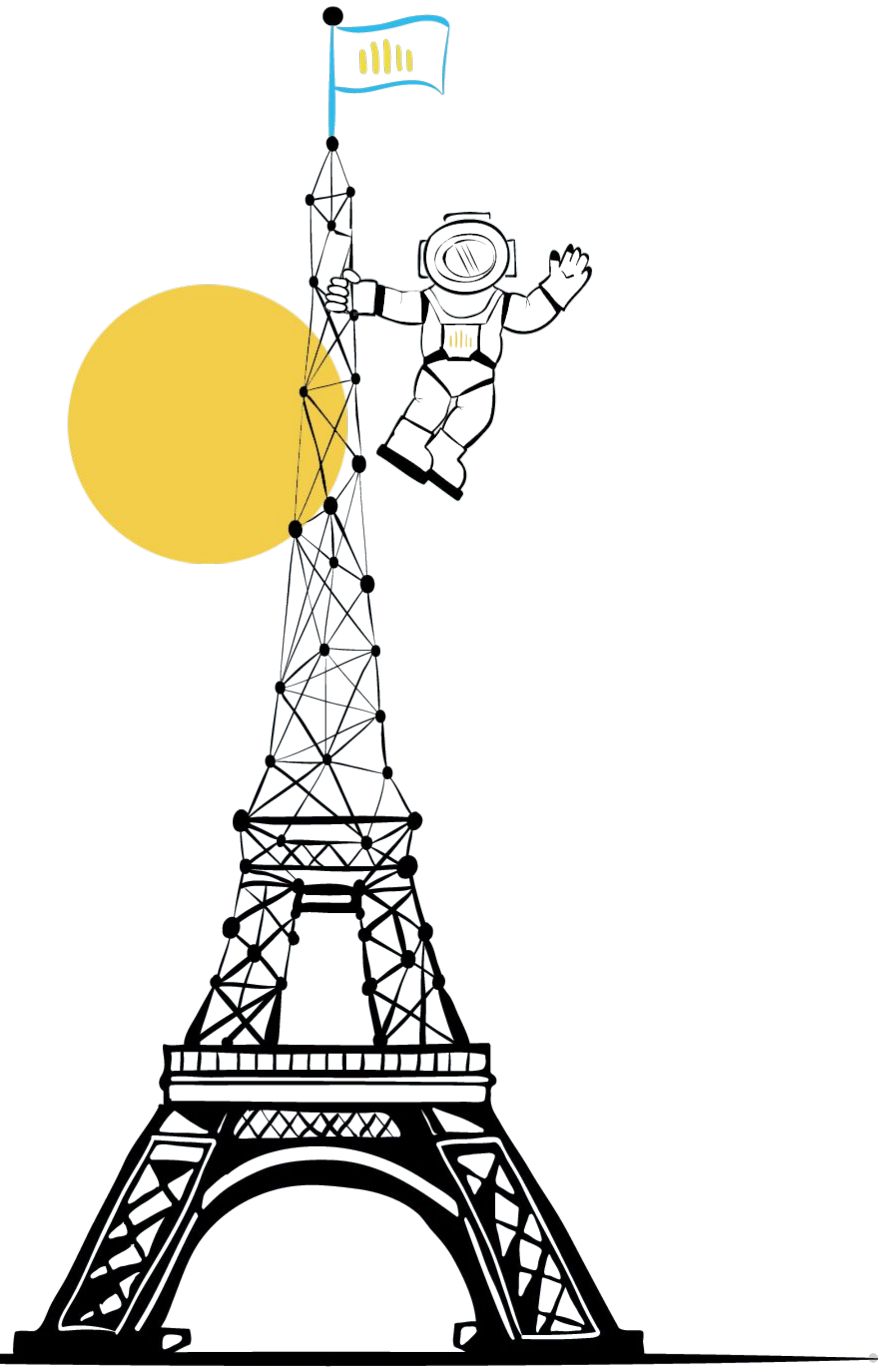
Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

https://arxiv.org/pdf/1906.02243.pdf



Tracking CO2 Emissions of Your Deep Learning Models with CodeCarbon and Weights & Biases

kaggle days
CHAMPIONSHIP
WITH

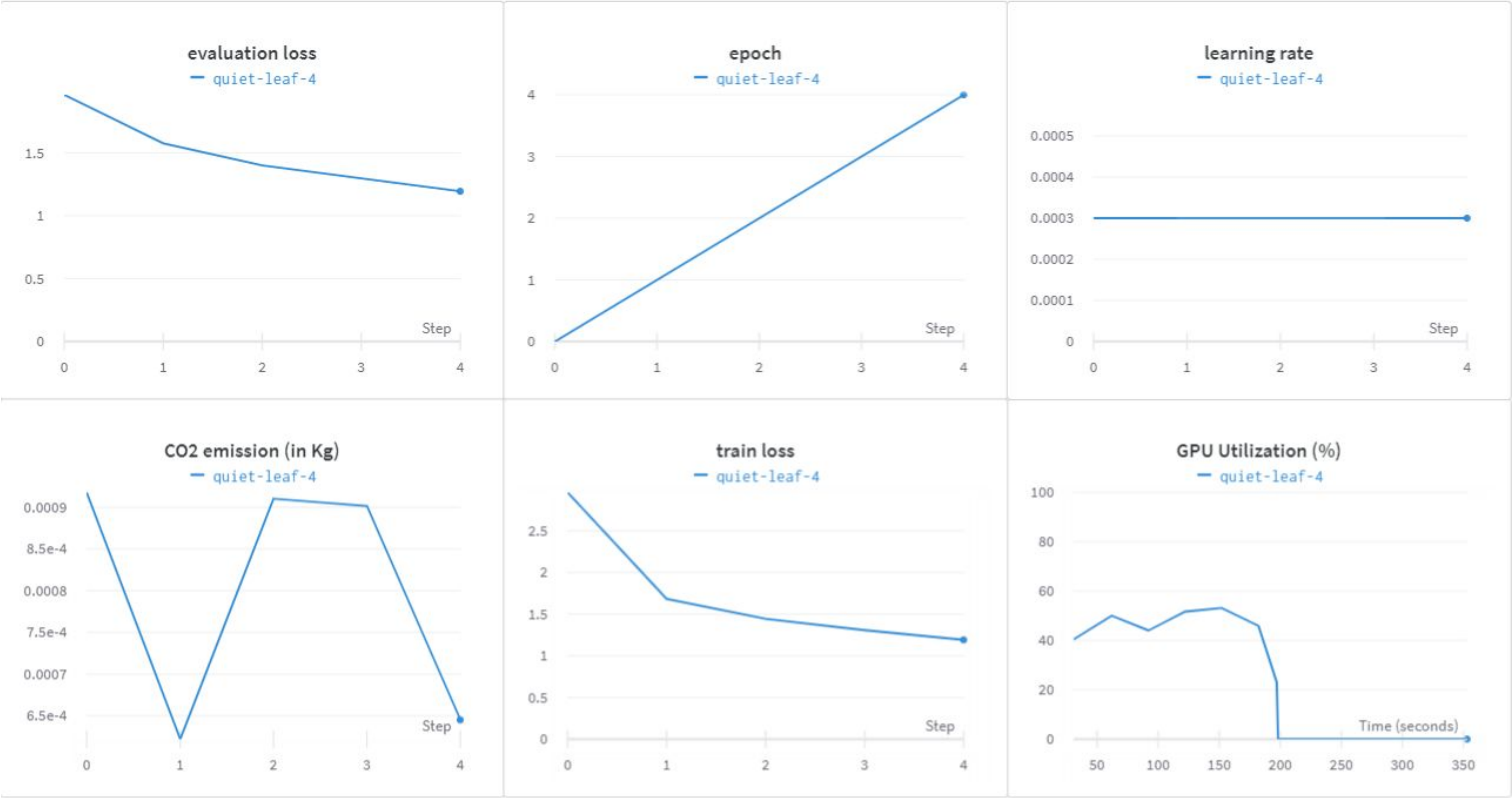


LOGICAI kaggle

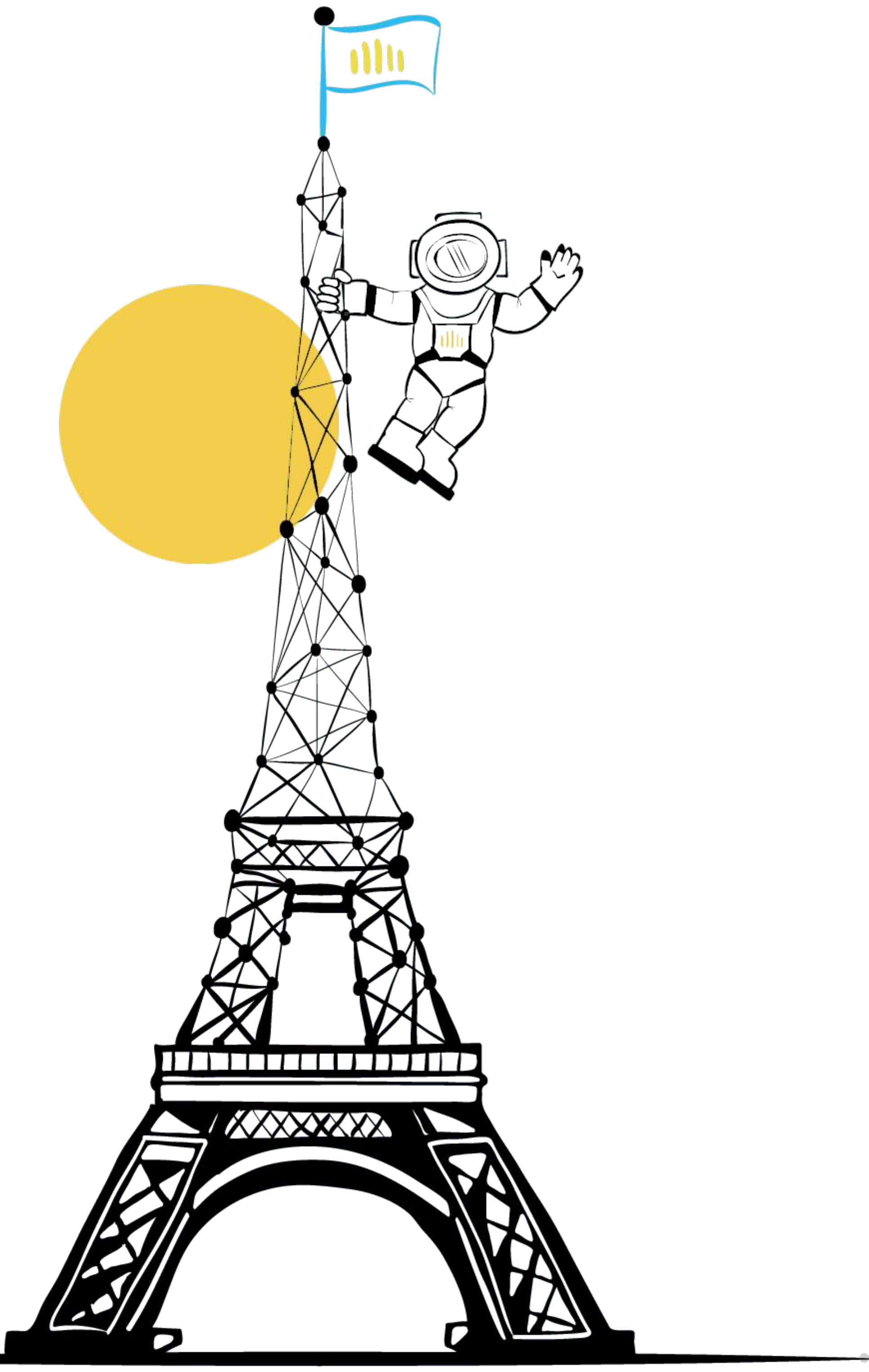
kaggle days

CHAMPIONSHIP

WITH Z



https://wandb.ai/amanarora/codecarbon/reports/Tracking-CO2-Emissions-of-Your-Deep-Learning-Models-with-CodeCarbon-and-Weights-Biases--VmlldzoxMzM1NDg3



LOGICAI Z kaggle

PARIS

MEETUP 04/08/2022

THANK YOU

PARIS

MEETUP 04/08/2022