Western Power Distribution: Data Science Challenges Kick-off

Dr Stephen Haben

Digital and Data Consultant

Dr Chris Harrison & Sam Young

Data Scientist & Practice Manager for Data Science and AI

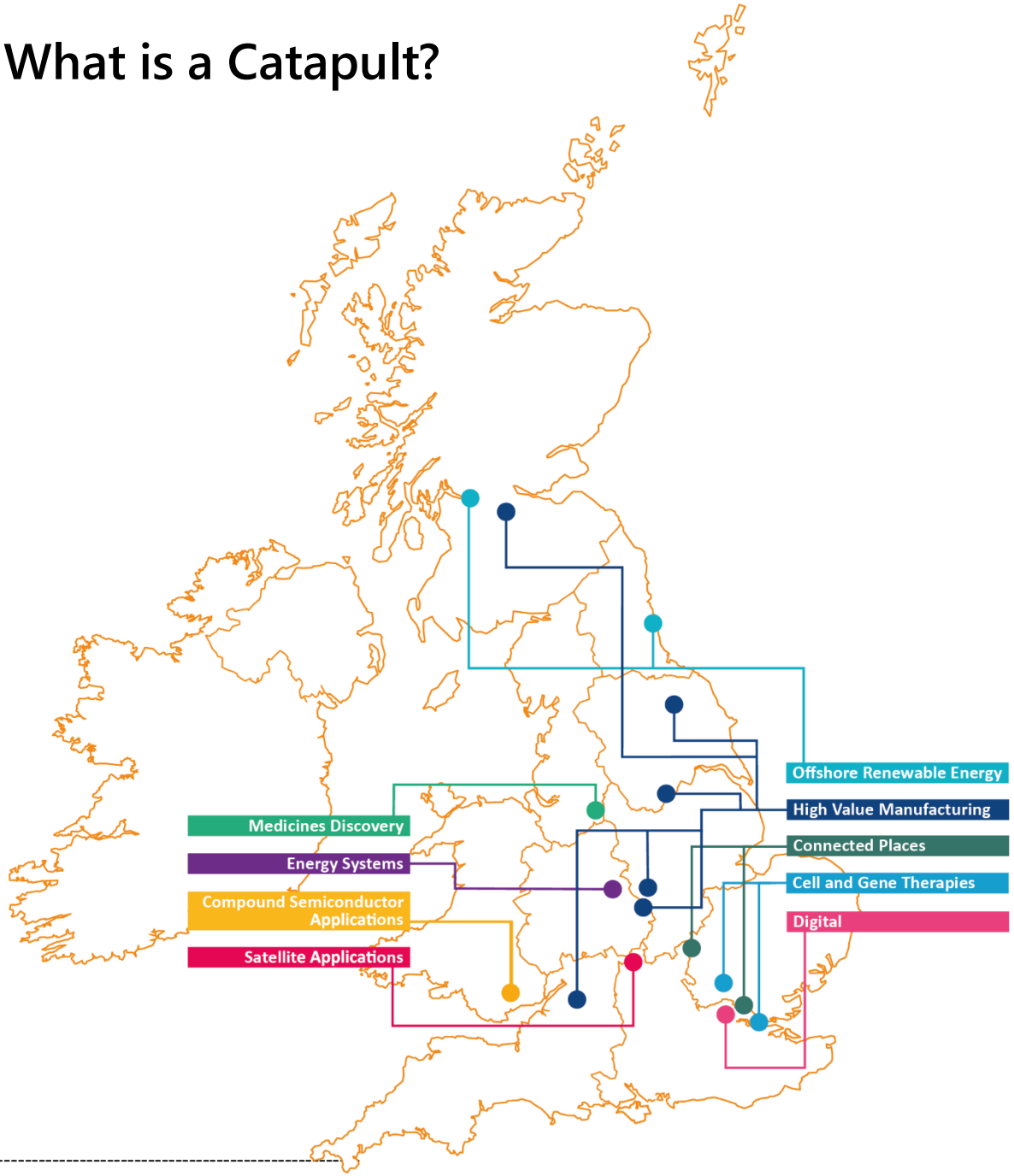
Thurs 11th November 2021

@EnergySysCat

Contents

- Aims of the Challenges.
- Previous Challenge overview – POD.
- Challenge Details.
- Platform: Codalab and submission process.
- Other details: dates, rules, etc.
- Looking forward: Challenges 2 & 3

What is a Catapult?

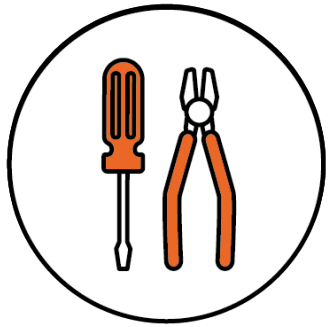


CATAPULT
Energy Systems

We work with Innovate UK

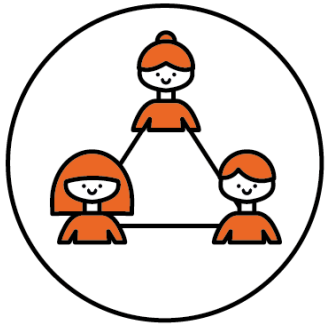
Established and overseen by Innovate UK

9 catapults



Technical capabilities, equipment, and other resources

Solve key problems and develop new products and services



Bridge the gap between stakeholders in the sector

Open up opportunities for innovators, in the UK and globally

Our Capabilities and Assets

Modelling

National Energy System Modelling
Local Area Energy Planning and Modelling
Building Energy System Modelling

Energy System Modelling Environment™
EnergyPath Networks™
Home Energy Dynamics
Storage and Flexibility Model

Markets, Policy and Regulation

Policy and Regulatory Knowledge
Economic Appraisal

Digital and Data

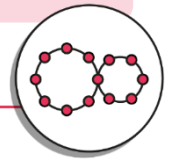
Data Science
Data Systems

Living Lab
Energy Knowledge eXchange™

Systems Integration

Systems Engineering and Integration
Dynamic Energy System Simulation
Dynamic Energy System Architecting
Business Model Innovation
Energy System Integration Guides

EnergyPath Operations™



Consumer Insight

Research
Design
Trials

People Lab
Home Truths®

Infrastructure and Engineering

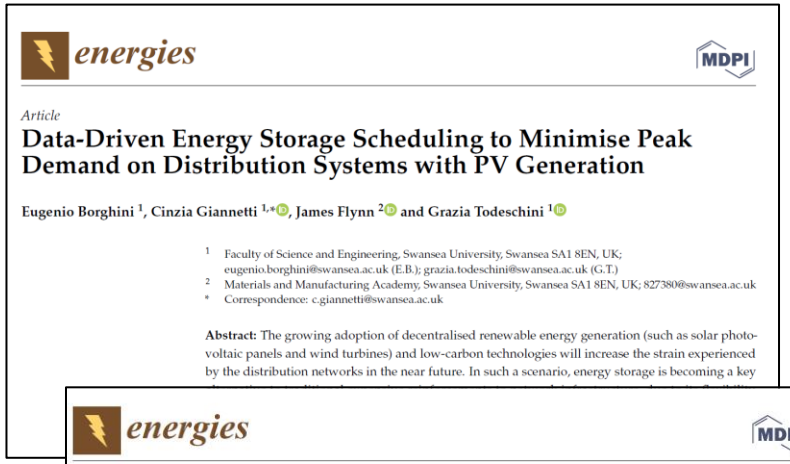
Networks and Energy Storage
Renewables
Transport
Nuclear
Carbon Capture and Storage, Industry and Hydrogen
Bioenergy

Welcome and Aims

The data science challenge from Presumed Open Data Project (next slide) – demonstrated that there is value in making data open and tackling data driven problems.

Aim of these challenges:

- Demonstrate the value of making unique data sets available.
- Highlight some of the major data driven challenges facing distribution network operators.
- Build a community of energy data science enthusiasts.
- Showcase state-of-the-art data science and machine learning methods for energy system problems.
- Identify some competitive and useful benchmarks.



energies

MDPI

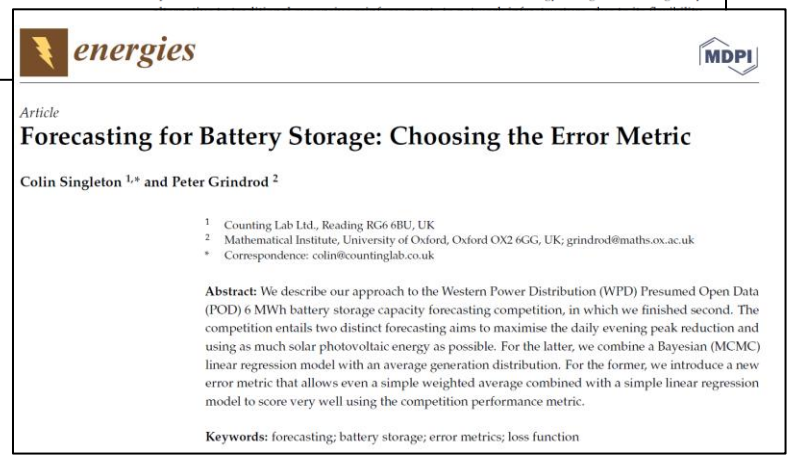
Article

Data-Driven Energy Storage Scheduling to Minimise Peak Demand on Distribution Systems with PV Generation

Eugenio Borghini [1], Cinzia Giannetti [1,*], James Flynn [2] and Grazia Todeschini [1]

[1] Faculty of Science and Engineering, Swansea University, Swansea SA1 8EN, UK; eugenio.borghini@swansea.ac.uk (E.B.); grazia.todeschini@swansea.ac.uk (G.T.)
[2] Materials and Manufacturing Academy, Swansea University, Swansea SA1 8EN, UK; 827380@swansea.ac.uk
* Correspondence: c.giannetti@swansea.ac.uk

Abstract: The growing adoption of decentralised renewable energy generation (such as solar photovoltaic panels and wind turbines) and low-carbon technologies will increase the strain experienced by the distribution networks in the near future. In such a scenario, energy storage is becoming a key



energies

MDPI

Article

Forecasting for Battery Storage: Choosing the Error Metric

Colin Singleton [1,*] and Peter Grindrod [2]

[1] Counting Lab Ltd., Reading RG6 6BU, UK
[2] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK; grindrod@maths.ox.ac.uk
* Correspondence: colin@countinglab.co.uk

Abstract: We describe our approach to the Western Power Distribution (WPD) Presumed Open Data (POD) 6 MWh battery storage capacity forecasting competition, in which we finished second. The competition entails two distinct forecasting aims to maximise the daily evening peak reduction and using as much solar photovoltaic energy as possible. For the latter, we combine a Bayesian (MCMC) linear regression model with an average generation distribution. For the former, we introduce a new error metric that allows even a simple weighted average combined with a simple linear regression model to score very well using the competition performance metric.

Keywords: forecasting; battery storage; error metrics; loss function

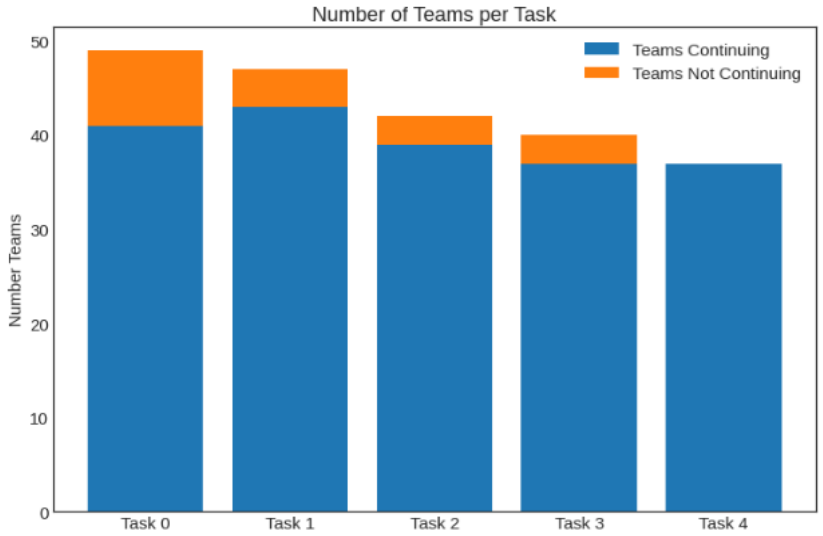
Overview: Presumed Open Data

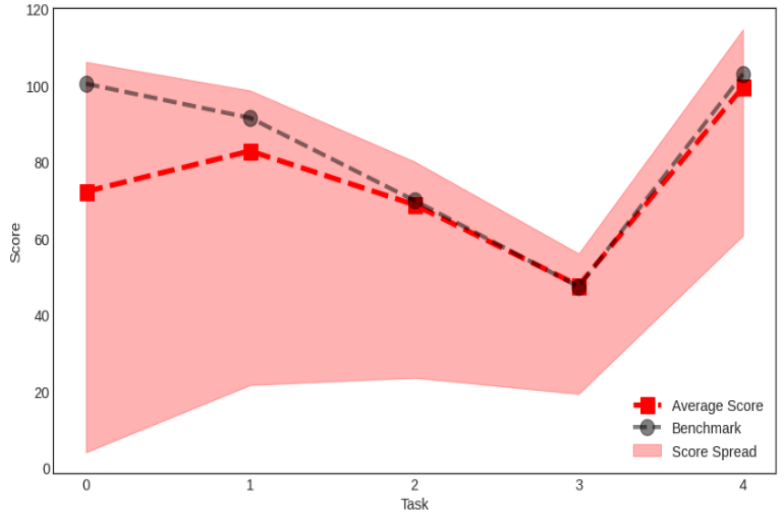
Outcomes

- Wide Participation: 55 teams - a total of 142 individuals – participated in a least one round. 72 different organisations/institutions
- Released Code (including winning team).
- Illustrated diversity of solutions and approaches
- Illustrated improvement of solutions through experience.
- Community (LinkedIn page ~120 people, now >180 after advertising this round).

Limitations/Lesson's Learned:

- A lot of work, requiring variety of techniques and skills (7 weeks, 5 submissions).
- Not accessible to many.
- Submission process not automated
- Scoring process – not scaled according to magnitude/ volatility of current task.





Challenges Series Overview

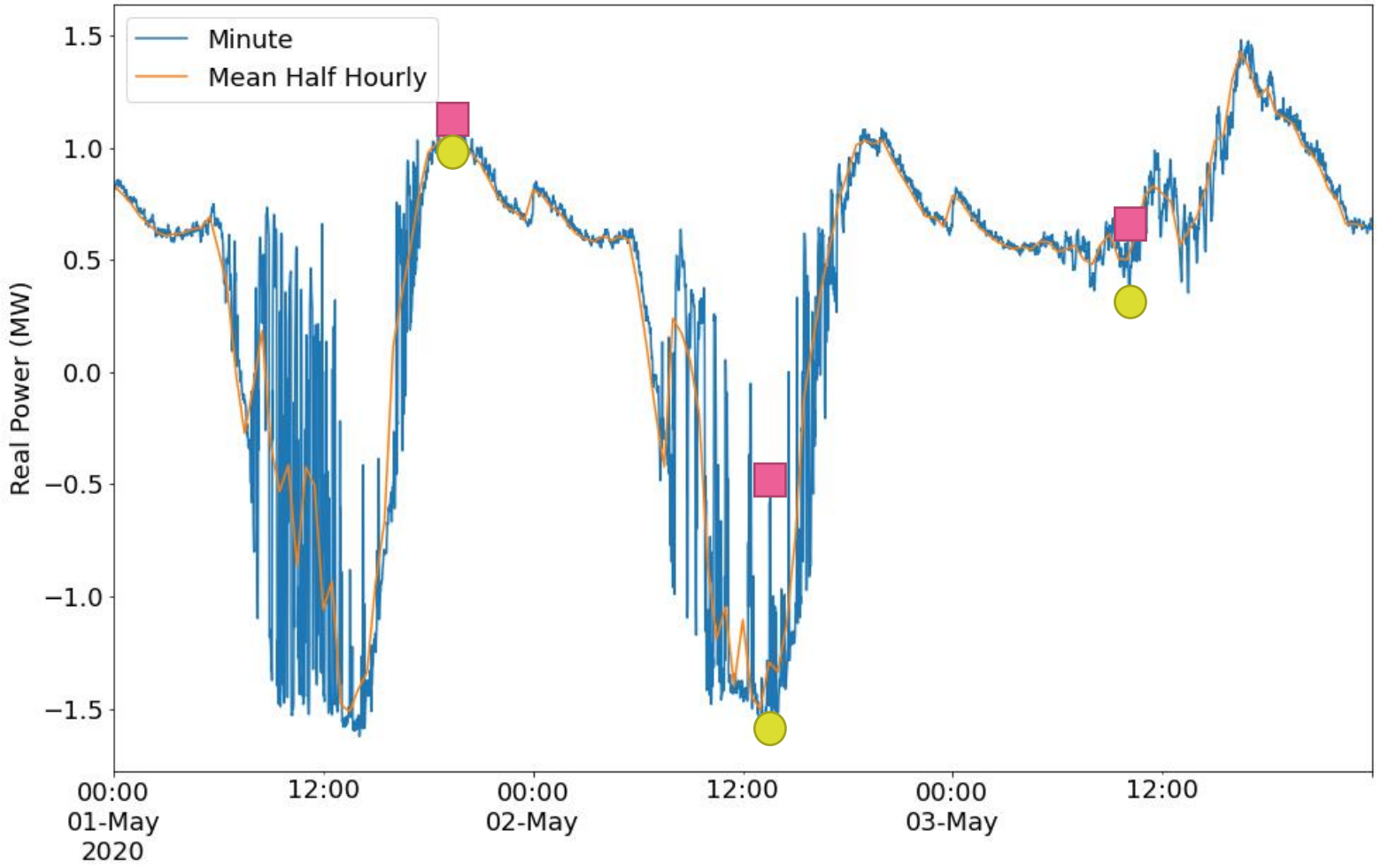
- Plan for **three challenges** over next few months.
- Three weeks each challenge with two weeks of validation and one/two week for testing.
- Simplified compared to POD but all on problems related to electricity networks.
- Automation of submission process.
- Aim for new data to be released for each challenge.
- No obligation to attempt all challenges.
- Encourage to release code at end to support building of benchmarks and community but no obligations!

Introduction to Challenge 1!

Outline of Challenge 1

- High resolution monitoring can be expensive and requires increased requirements.
- Important features to power system modelling are the peaks and troughs in electricity demand.
- Can some of the features (in particular maximum and minimum) at high resolution be accurately estimated given only 30 minute averages, and weather data?
- **Aim is to estimate the maximum and minimum demand at minute resolution for each 30 minute period for a whole month**.

Illustration of three days



Maximum

Minimum

This document is marked as confidential

Site Details

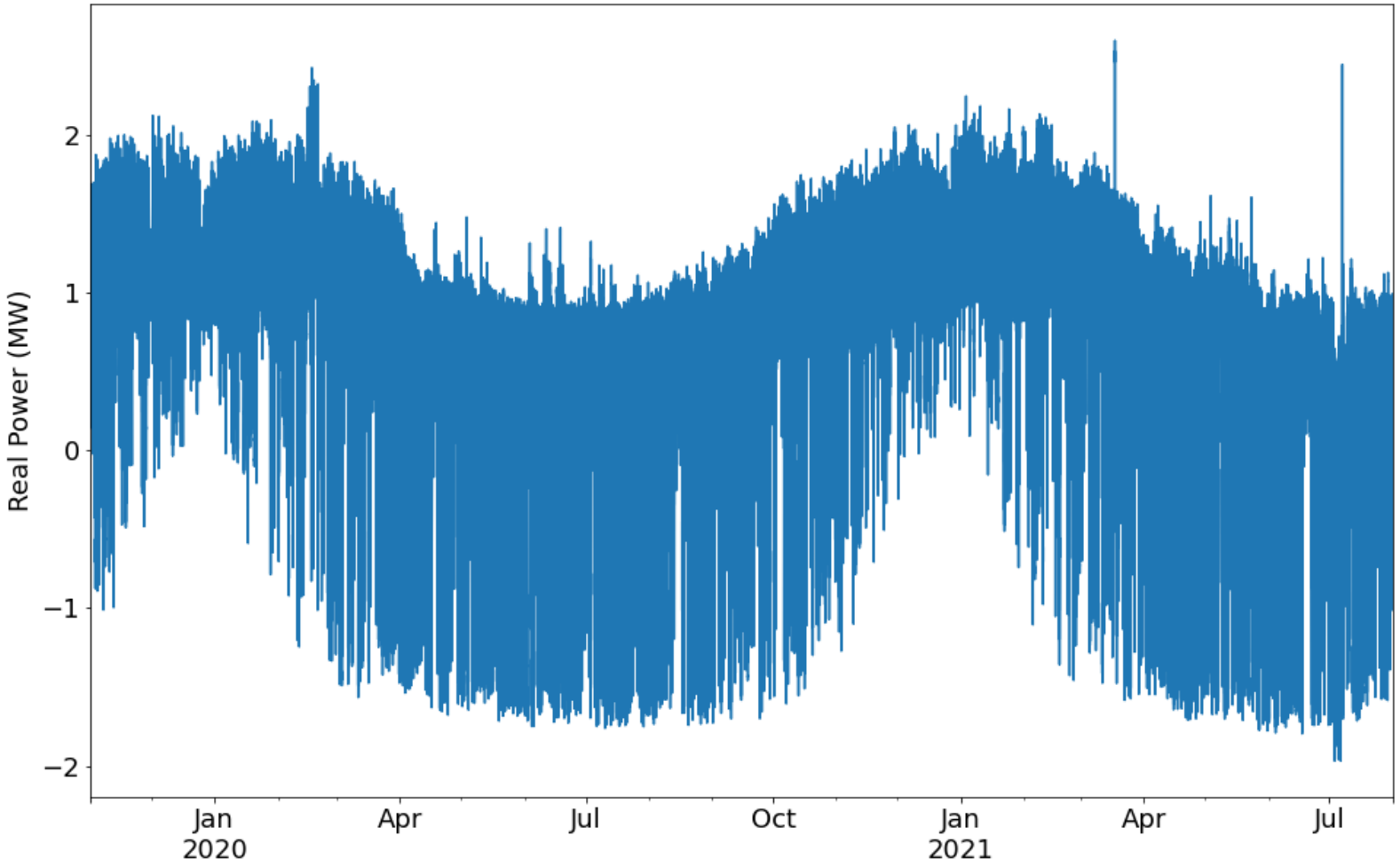
- Primary Substation 33/11kV feeder.
- (Latitude, Longitude) = (51.0254, -3.1204).
- Large amount of distributed generation on 11kV side.
- Mixed generation – more than one type, but don't have complete visibility.
- Mix of different consumer types (left).

Staplegrove	
Generation Type	kVA
Hydro	13
Mixed	8.68
Photovoltaic	6433.9
Storage (Battery)	3.68

Profile Class	Number of Customers
1 (domestic unrestricted)	8066
2 (domestic economy 7)	952
3 (non-domestic unrestricted)	499
4 (non-domestic economy 7)	105
5 (Non-dom, max demand customers with Peak Load factor <20%)	0
6 (Non-dom, max demand with PLF between 20 and 30%)	1
7 (Non-dom, max demand with PLF between 30 and 40%)	2
8 (Non-dom, max demand with PLF over 40%	6

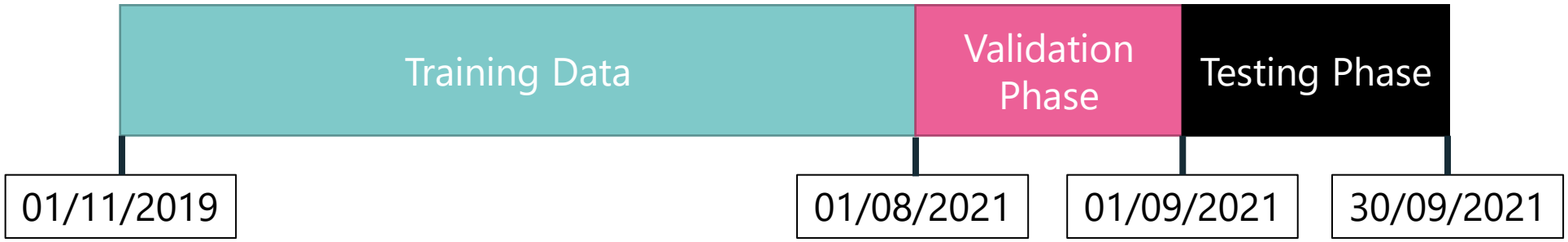
More details on Profile classes:
https://www.elexon.co.uk/knowledgebase/profile-classes/

Half-hourly demand data



Demand Data Details

- Provided on UTC time stamps. (No need for daylight savings corrections).
- Demand provided:
  - Half hourly demand data for full period (1st Nov 2019 to 30th Sept 2021 inclusive).
  - Minute Resolution training data (1st Nov 2019 to 31st July 2021). August data released after validation phase.
  - Solution data (max and min minute resolution values for each half hour period). Validation solutions released end of validation phase. Testing data solutions released at end challenge.
- Templates (CSV files for submitting solution)
- Also release data for two other sites for practice. These are not included in challenge.
- More details on the codalab page (see later).



Weather Data

Name	Latitude	Longitude
staplegrove_1	51	-3.125
staplegrove_2	51	-2.5
staplegrove_3	51.5	-3.125
staplegrove_4	51.5	-2.5
staplegrove_5	51	-3.75

- Weather data important for demand/generation.
- Used extensively in POD data science challenge
- Hourly Merra-2 reanalysis data, from 5 sites around substation.
- For full period: 1st Nov 2019 to 30th Sept 2021 inclusive.
- Timestamps in UTC.
- Further information on platform website.

Substation Site

Weather Grid point

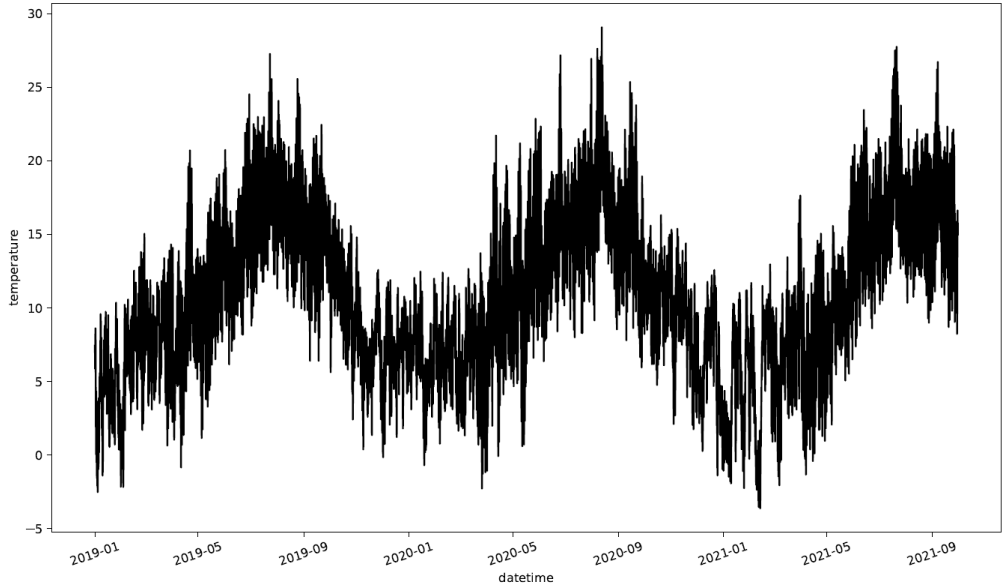
Weather Data (further details)

Variable	Units	Description	MERRA Database	MERRA field ID
Temperature	Celcuis	Instantaneous temperature reading	inst1_2d_asm_Nx	T2M
Solar Irradiance	W m-2 (watts per square meter)	Surface incoming shortwave flux	tavg1_2d_rad_Nx	SWGDN
Eastward Wind Speed	m s-1	Average (2-meter) Wind Speed in East Direction	tavg1_2d_slv_Nx	U2M
Northward Wind Speed	m s-1	Average (2-meter) Wind Speed North Direction	tavg1_2d_slv_Nx	V2M
Surface Pressure	Pa	Average Surface Pressure	tavg1_2d_slv_Nx	PS
Specific Humidity	kg kg-1	Average 2-meter specific humidity	tavg1_2d_slv_Nx	QV2M

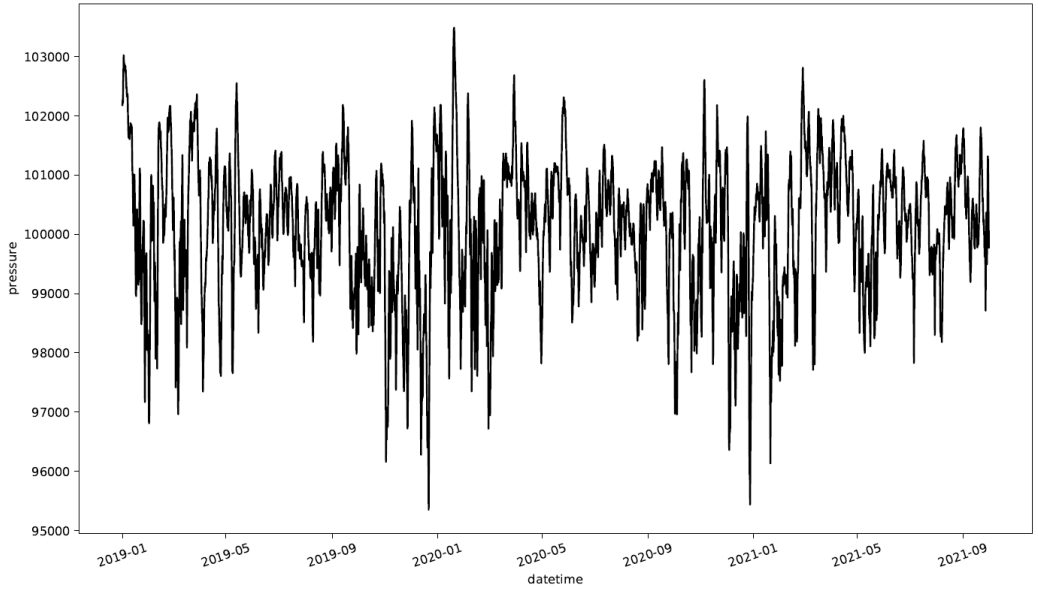
MERRA-2 data policy found here: https://disc.gsfc.nasa.gov/information/documents?title=data-policy

Examples (Staplegrove 1)

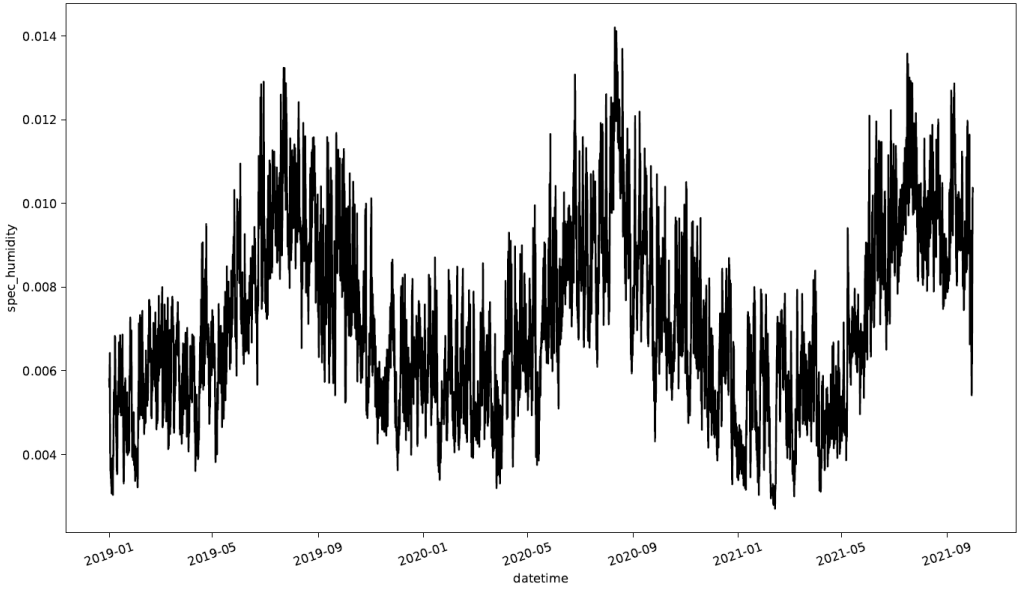
Temperature



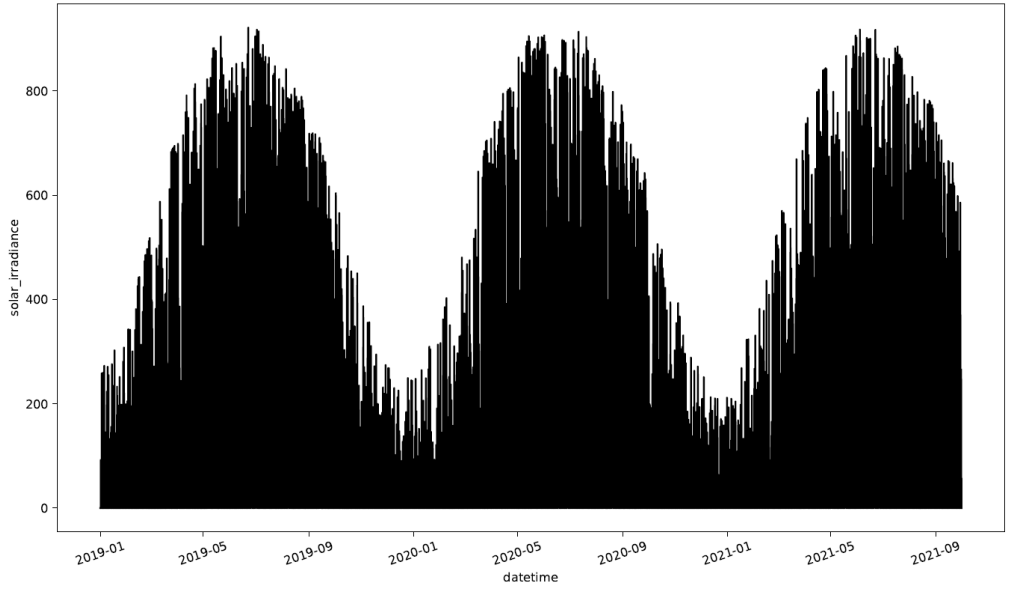
Pressure



Specific Humidity



Solar Irradiance



Challenge Formulation Specifics

- Wordy version: Given half hourly demand data and hourly weather data over the month, and historical minute resolution demand data, accurately estimate the largest and smallest demand value (at minute resolution) which occurs over each half hour period in the month.

- More precision: Let $P_{d,h,m}$ define the average power value (in MW) for minute m ($m \in \{1,..,30\}$), within half hour h ($h \in \{1,..,48\}$) of day d ($d \in \{1,..,D\}$) (D is number of days in our training data.

- Next define the half hourly average power series $Q_{d,h}$ (in MW) for half hour h ($h \in \{1,..,48\}$) of day d ($d \in \{1,..,D\}$) where

$$Q_{d,h} = \frac{1}{30}\sum_{m=1}^{30} P_{d,h,m},$$

Objective

Then given the test data is days $d = D_1, \ldots, D_2$ the aim of the challenge is to generate an solution $\hat{P}$ which accurately estimates:

- Maximum values

$$P_{d,h}^{max} = \max_{m \in \{1,..,30\}} P_{d,h,m}$$

For each ($h \in \{1,..,48\}$) and day d ($d \in \{D_1, \ldots, D_2\}$), and

- Minimum values

$$P_{d,h}^{min} = \min_{m \in \{1,..,30\}} P_{d,h,m}$$

For each ($h \in \{1,..,48\}$) and day d ($d \in \{D_1, \ldots, D_2\}$).

Benchmark & Scoring

Define the RMSE error of an estimate :

$$RMSE(\hat{P}, P) = \sqrt{\sum_{d=D_1}^{D_2} \sum_{h=1}^{48} (\hat{P}_{d,h}^{max} - P_{d,h}^{max})^2 + \sum_{d=D_1}^{D_2} \sum_{h=1}^{48} (\hat{P}_{d,h}^{min} - P_{d,h}^{min})^2}$$

The scoring will be a *skill score* relative to a simple benchmark $\hat{B}$ given by:

$$\hat{B}_{d,h}^{max} = Q_{d,h}$$

For the maximum value. And

$$\hat{B}_{d,h}^{min} = Q_{d,h}$$

For the minimum value, for $h(h \in \{1,..,48\})$ and day d $(d \in \{D_1, \dots, D_2\})$. In other words simply the average half hourly demand.

Then the skill score $S(\hat{P}, \hat{B})$ is define as

$$S(\hat{P}, \hat{B}) = \frac{RMSE(\hat{P}, P)}{RMSE(\hat{B}, P)}$$

CodaLab Example

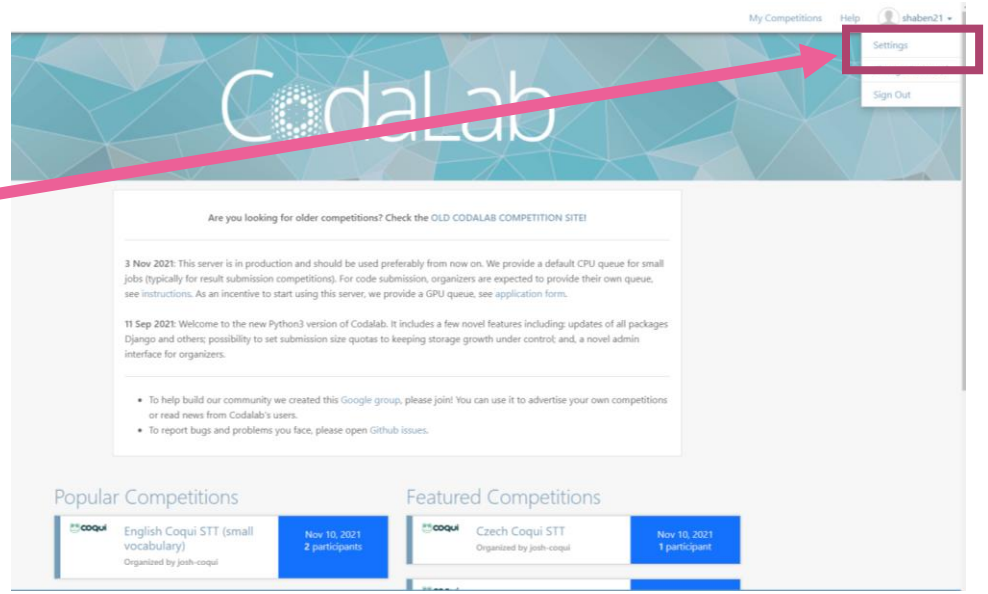
Sign Up and Team Creation

- First Sign up: https://codalab.lisn.upsaclay.fr/accounts/signup/
- Go to Challenge link: https://codalab.lisn.upsaclay.fr/competitions/213
  Click "Participate" tab -> tick to "accept terms and conditions" -> Click "Register"
- Wait for approval.
- Once approved (email notification) challenge should show up in "Competitions I'm In"

To Create Team:

- Maximum team size of five.
- Click your username in top right then "Settings"
- Add a team name in "Team Name" box. All individuals in same team need to give same name.
- **Teams must be finalised by end Validation phase.**
- **Only one member of each team allowed to make submissions in final testing phase.**

More Info from Git pages for codalab:
https://github.com/codalab/codalab-competitions/wiki

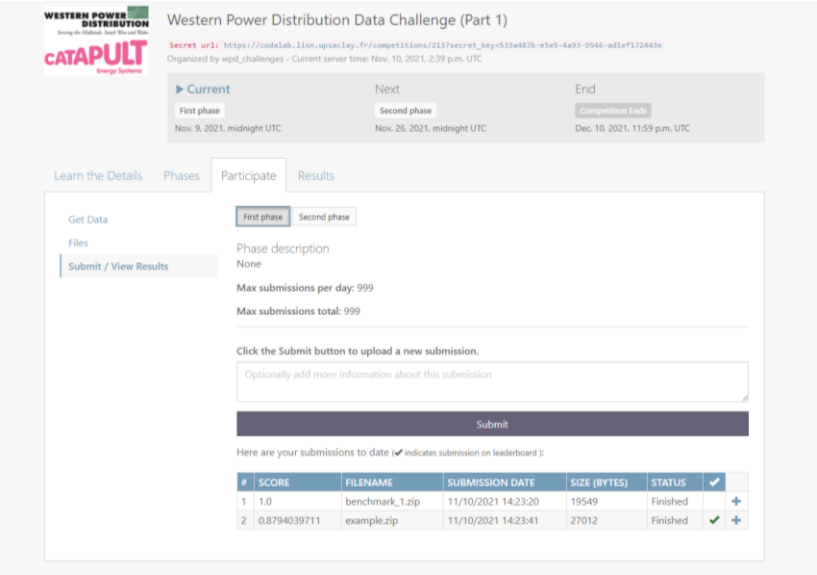


Codalab Demonstration

Submission Process

- Individual submissions:
  - Take solution template
  - Add your solutions
  - **Rename file to “Predictions.csv”**
  - Zip it (zip filename can be anything you want).
  - Go to challenge website
  - Click “Participate” Tab
  - Click “Submit/View Results”
  - Click Submit and add zip file. (May require refreshing or take time)

- Two Phases
  - Phase 1: Validation phase – Unlimited submissions (don’t count towards final score)
  - Phase 2: Test Phase – maximum of 6 submissions (please nominate one submitter per team)



Rules and Other Information

- Teams of up to 5 only.
- Additional data not allowed, future challenges will allow this with certain restrictions.
- Email Short 1 page outline of method/approach.
- Code: Not obligatory to share but great way to promote methods and work, and utilise as benchmarks/comparison.
- Top teams invited to present at end of challenge workshop.
- Prizes: Choices
  - Up to £500 cloud computing resource.
  - Colab Pro membership for team members.
  - Suggestions welcome for future challenges.

Links and Content

- Data Hosted at the WPD data hub https://connecteddata.westernpower.co.uk/dataset/western-power-distribution-data-challenge-1-high-resolution-peak-estimation includes:
  - Demand data (30minute data) for entire period.
  - Demand data (1minute) for the period up to pre August.
  - Template for submissions
  - Additional data of two other sites (Greevor and Mousehole) for practice.
- Weather data (In Data tab on Codalab)
- Challenge Link: https://codalab.lisn.upsaclay.fr/competitions/213
- Email: wpd-challenges@es.catapult.org.uk
- LinkedIn Challenge Forum group: https://www.linkedin.com/groups/9025332/
- Slides and video link will be shared on LinkedIn page.

Final Reminders/Thoughts

Important Dates:

- **Validation Period:** 2 weeks started from 11th November 2021 (Now!) to 25th November 2021 (Midnight UK time).
- **Test Period:** 2 week from 26th November to 10th December 2021 (23:59:59 UK time)
- Submit using template and don't mess with order!

- Utilise the practice session with unlimited submissions.
- Discuss on LinkedIn Forum and/or use to find team mates.

Sneak Peak future challenges:

- Suggestion period for new input data.
- Code submission – final challenge.
- New Prizes.

