SUPPORTING DATA SCIENCE IN THE ENERGY SECTOR

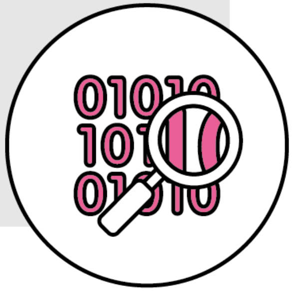
DR STEPHEN HABEN

DIGITAL AND DATA CONSULTANT

WEDNESDAY 27 SEPT 2022

AGENDA

- **The Academic-industry Gap**
- **Accessible and Reproducible Research**
- **The Value in Data Science Competitions**
- **A Brief Note on Data Science Skills**



CATAPULT NETWORK.

SUPPORTING BUSINESS IN TRANSFORMING GREAT IDEAS INTO VALUABLE PRODUCTS AND SERVICES.

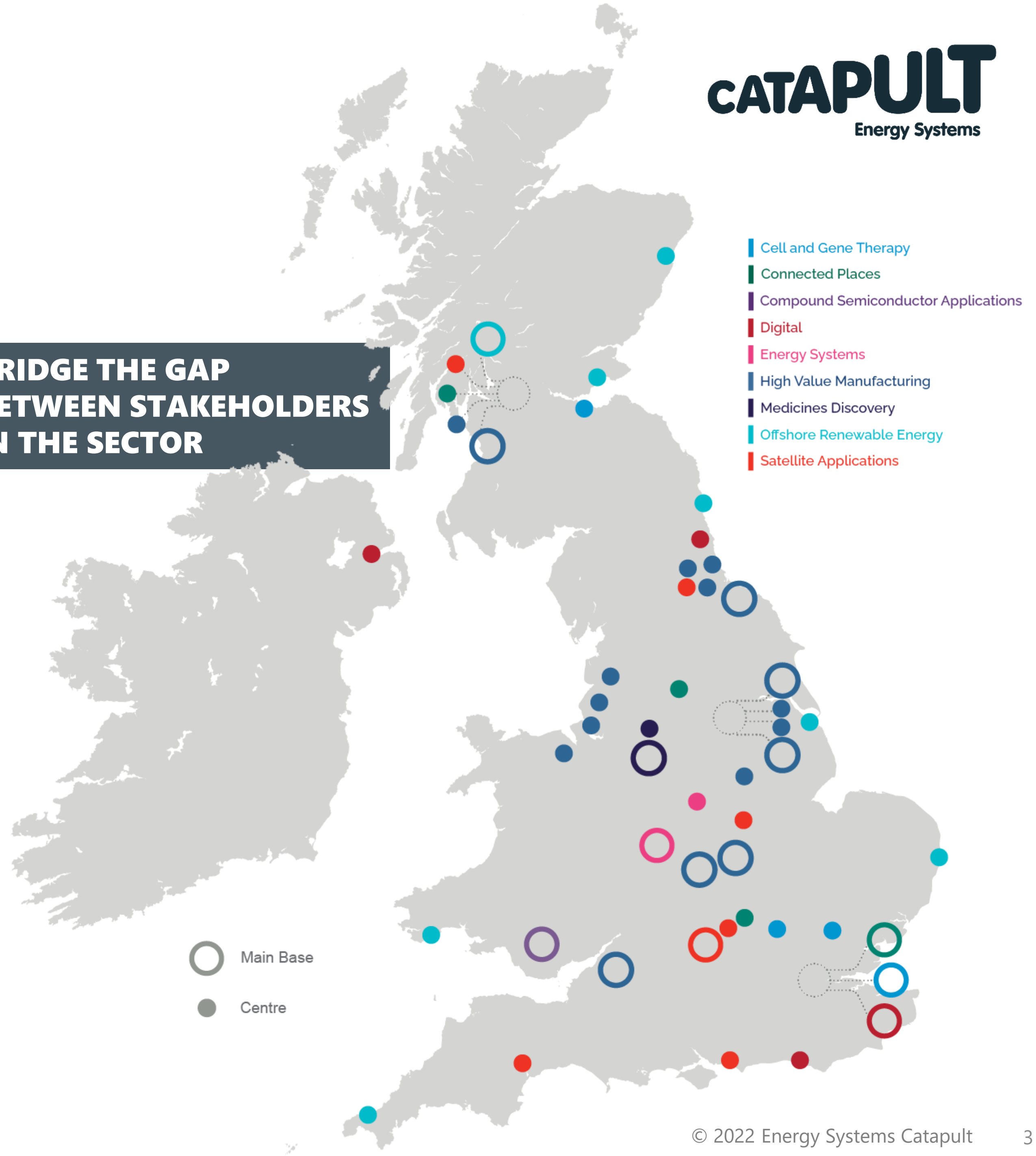
TECHNICAL CAPABILITIES, EQUIPMENT, AND OTHER RESOURCES

BRIDGE THE GAP BETWEEN STAKEHOLDERS IN THE SECTOR

OPEN UP OPPORTUNITIES FOR INNOVATORS, IN THE UK AND GLOBALLY

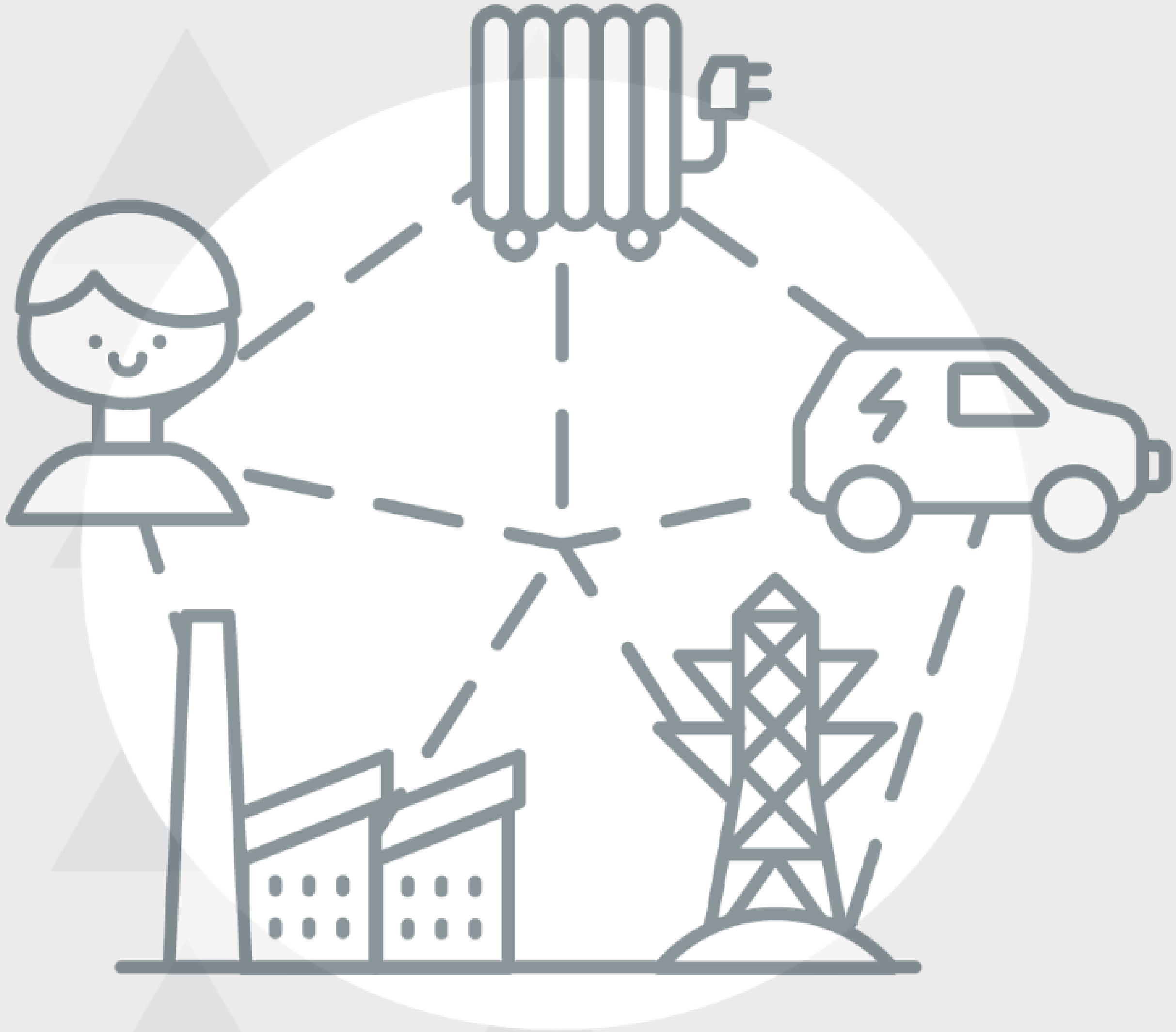
SOLVE KEY PROBLEMS AND DEVELOP NEW PRODUCTS AND SERVICES

CATAPULT Energy Systems



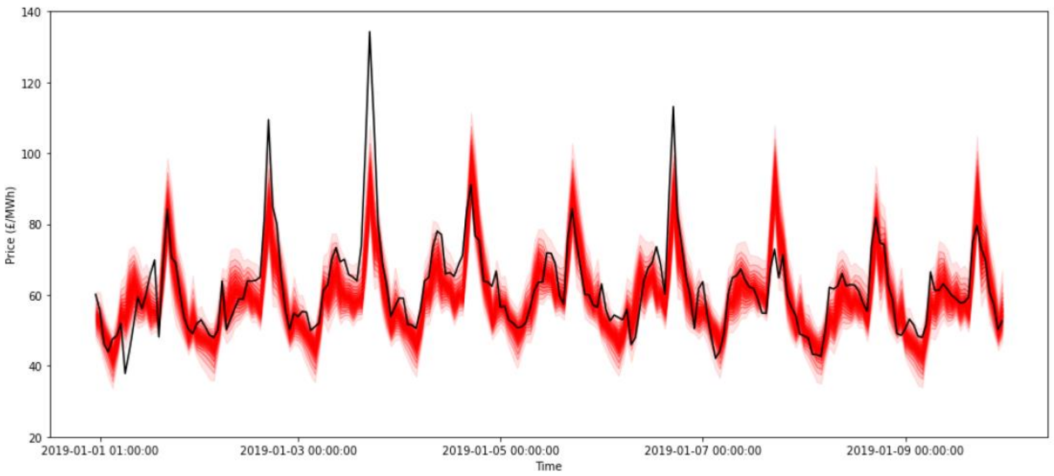
We work with Innovate UK

INTRODUCTION.



SUPPORTING THE ENERGY SECTOR

- Started a investigation on data driven probabilistic day ahead price forecasting to support projects on smart local energy systems
- Published in 2021
- Approached by Arenko an innovative flexibility service provider in UK about the work
- Started a conversation on the irreproducibility issues in machine learning in energy sector.



forecasting
MDPI

Article

Probabilistic Day-Ahead Wholesale Price Forecast: A Case Study in Great Britain

Stephen Haben [1,2,*], Julien Caudron [1] and Jake Verma [1]

1 Energy Systems Catapult, Cannon House, Birmingham B4 6BS, UK; julien.caudron@es.catapult.org.uk (J.C.); jake.verma@es.catapult.org.uk (J.V.)
2 Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK
* Correspondence: stephen.haben@es.catapult.org.uk

Abstract: The energy sector is moving towards a low-carbon, decentralised, and smarter network. The increased uptake of distributed renewable energy and cheaper storage devices provide opportunities for new local energy markets. These local energy markets will require probabilistic price forecasting models to better describe the future price uncertainty. This article considers the application of probabilistic electricity price forecasting models to the wholesale market of Great Britain (GB) and compares them to better understand their capabilities and limits. One of the models that this paper considers is a recent novel X-model that predicts the full supply and demand curves from the bid-stack. The advantage of this model is that it better captures price spikes in the data. In this paper, we provide an adjustment to the model to handle data from GB. In addition to this, we then consider and compare two time-series approaches and a simple benchmark. We compare both point forecasts and probabilistic forecasts on real wholesale price data from GB and consider both point and probabilistic measures.

Keywords: price forecasting; day-ahead forecasting; probabilistic price forecasting; electricity prices; supply and demand curves; price spikes; wholesale market

check for updates

Citation: Haben, S.; Caudron, J.; Verma, J. Probabilistic Day-Ahead Wholesale Price Forecast: A Case Study in Great Britain. *Forecasting* **2021**, *3*, 596–632. https://doi.org/10.3390/forecast3030038

Academic Editor: Sonia Leva

Received: 29 June 2021
Accepted: 25 August 2021
Published: 28 August 2021

1. Introduction

Since the liberalisation of the electricity market, forecasting electricity prices has been an important factor in decision making for energy suppliers and generators. Day-ahead wholesale electricity price forecasts are an essential component of the electricity market. In the wholesale market, electricity is traded between suppliers and generators by placing offers and bids, respectively, for different volumes of electricity. This is used to set the wholesale price at which day-ahead electricity is purchased. Energy suppliers generally hedge (purchase ahead) their best forecast of volumes and refine their positions closer to delivery, such as in the day-ahead market. The overall cost of wholesale energy is combined with other cost elements in the tariffs offered to consumers. The wholesale electricity price forecasts are a fundamental input for an energy company's decision making. Prices are relatively volatile, and hence, probabilistic forecasts are more useful, as they describe the uncertainty associated with different events. In this article, three day-ahead probabilistic electricity price forecasts are developed and tested for Great Britain's day-ahead wholesale electricity market. One of the focuses will be on a recent method developed in [1] called the X-model, whose focus is on the prediction of spikes in the electricity price.

Until recently, wholesale price forecasting was typically focused on point forecasts. However, in the last few years, probabilistic price forecasting has been gaining interest. The comprehensive 2014 review by Weron [2] showed that very few papers at the time considered probabilistic forecasts. The 2018 review update by Nowotarski and Weron [3], however, highlighted the importance of probabilistic forecasts due to the introduction of the so-called smart grid and the increased uncertainty in supply and demand. A more recent review [4] published after the development presented in this article underlines the

Forecasting **2021**, 3, 596–632. https://doi.org/10.3390/forecast3030038 https://www.mdpi.com/journal/forecasting

DATA SCIENCE: FROM ACADEMIA TO INDUSTRY

- Linking academic researchers and business is useful for the innovation process.
- "Knowledge transfer between academia and companies is not currently working", The Data Science & AI Section of the Royal Statistical Society, 2021
- Frustration with:
  - Academic outputs not being accessible or reproducible
  - Collaborations being suboptimal
  - Graduates not having sufficient coding skills
  - Lack of industry support

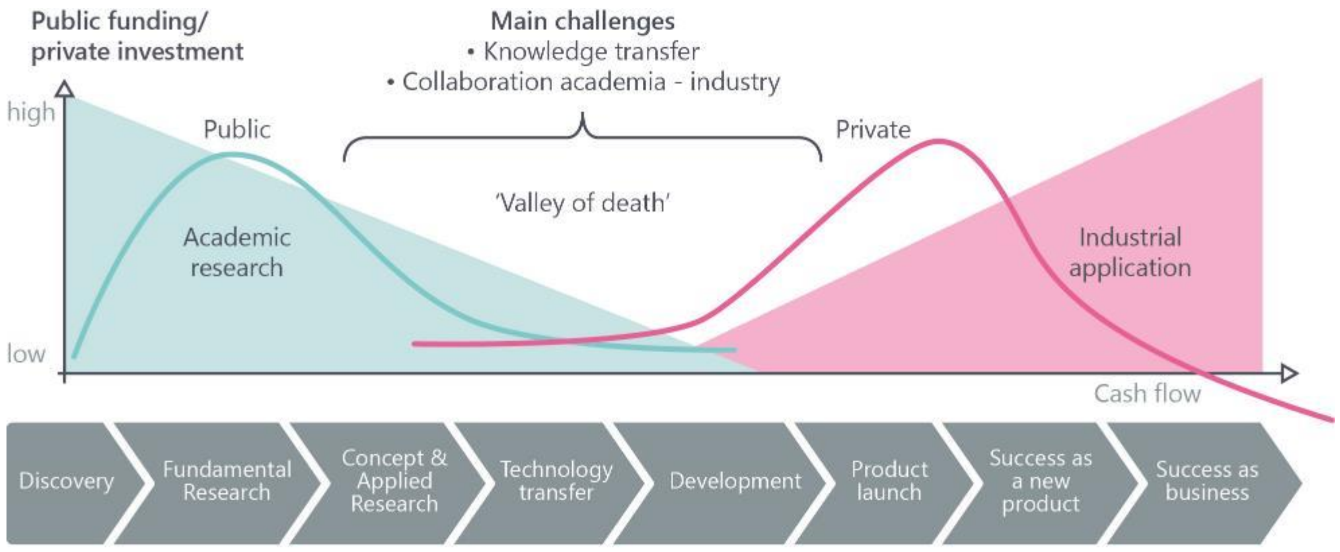


Figure: Valley of Death, reprinted and adapted from (Chirazi, Wanieck, Fayemi, Zollfrank, & Jacobs, 2019), under CC BY 4.0 license

FINAL REPORTS

https://es.catapult.org.uk/report/data-science-from-academia-to-industry/

Knowledge

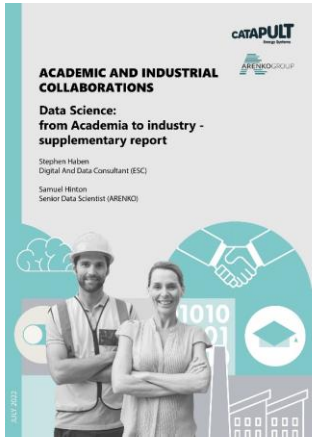
Skills

Accessibility & Reproducibility

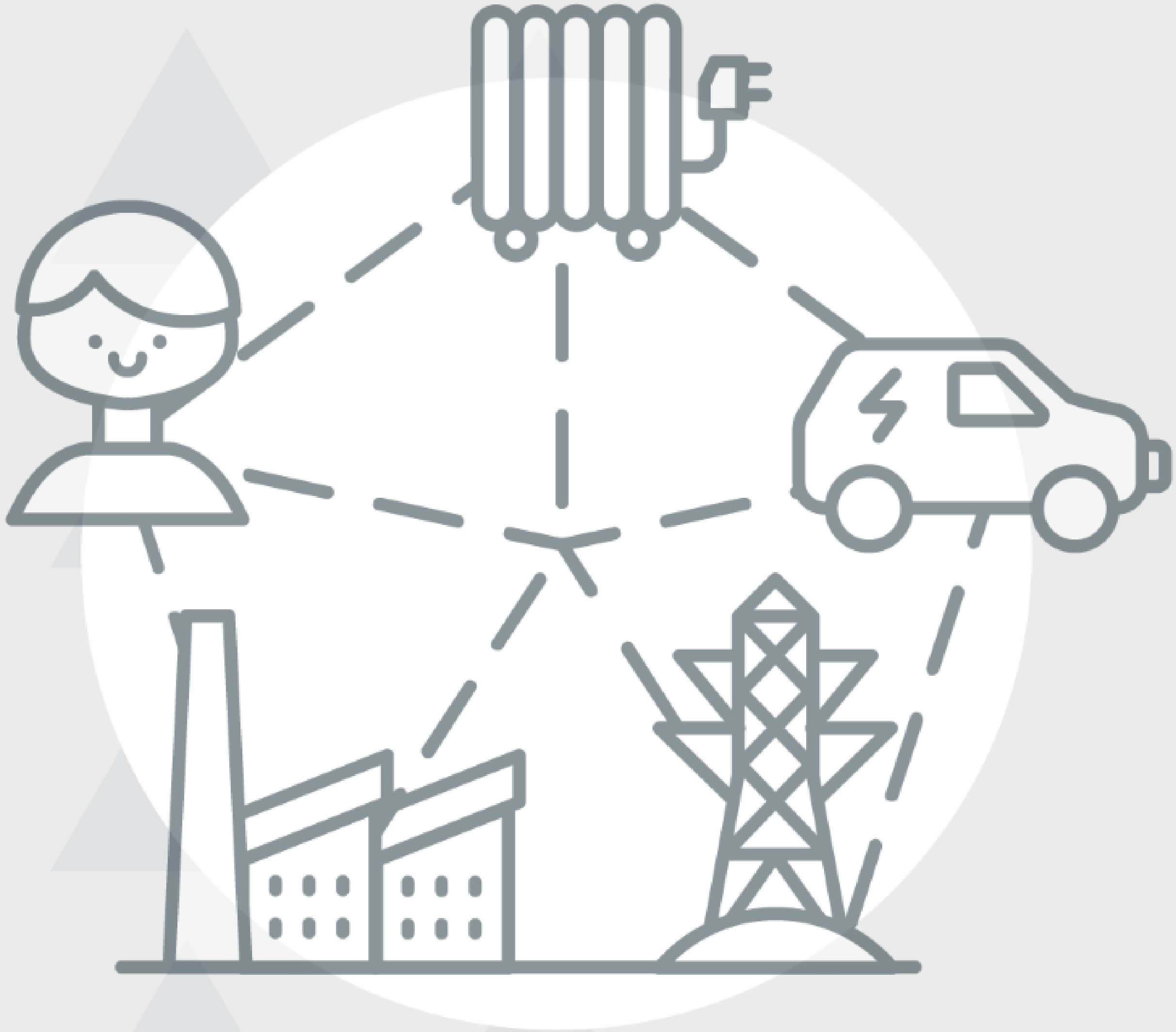
Collaborations

Industrial Support

Code Development

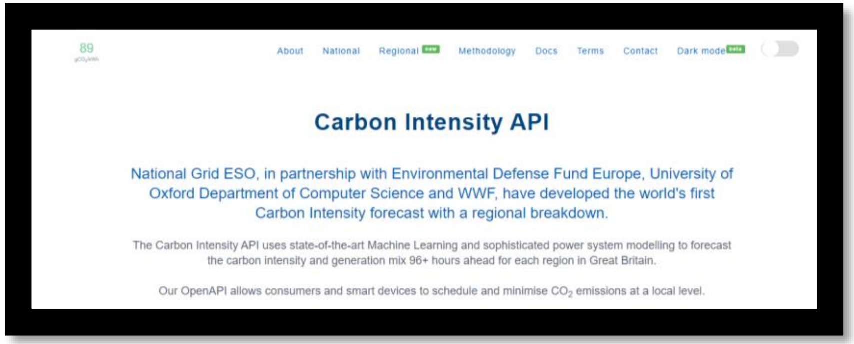


KNOWLEDGE SHARING



INDUSTRIAL NEEDS FROM ACADEMIA

CATAPULT Energy Systems



OPEN CLIMATE FIX

ARENKOGROUP

Innovation: can provide competitive edge. Most knowledge is known across same companies, but it is the unknown unknowns which can produce game changing innovation.

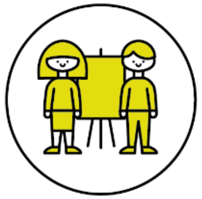
New Ideas: Companies are working on ideas that cannot be shared but working in silos. Academia can bring new ideas with very skilled individuals. Test new and potentially risky opportunities.

Benchmarking: Drive internal competition, and benchmark individuals who are rarely compared to others.

ACADEMIC NEEDS FROM INDUSTRY

Data Provider: Including clear documentation and ideally listed on data API marketplaces.

Domain Expertise: Practical constraints of markets, data availability details, subtleties about automatic/manual processes.



Problems: What are the most challenging problems facing industry?

Feedback: Very little information on what is used from academia. Feedback can support academic impact narratives, and help direct research.

Transparency: Industry process remain black boxes to all external parties. Restricts benchmarking, and development of state of the art. Could provide outputs rather than full models to preserve IP.

WHY SO LITTLE PROGRESS?

Misalignment between Academia & Industry

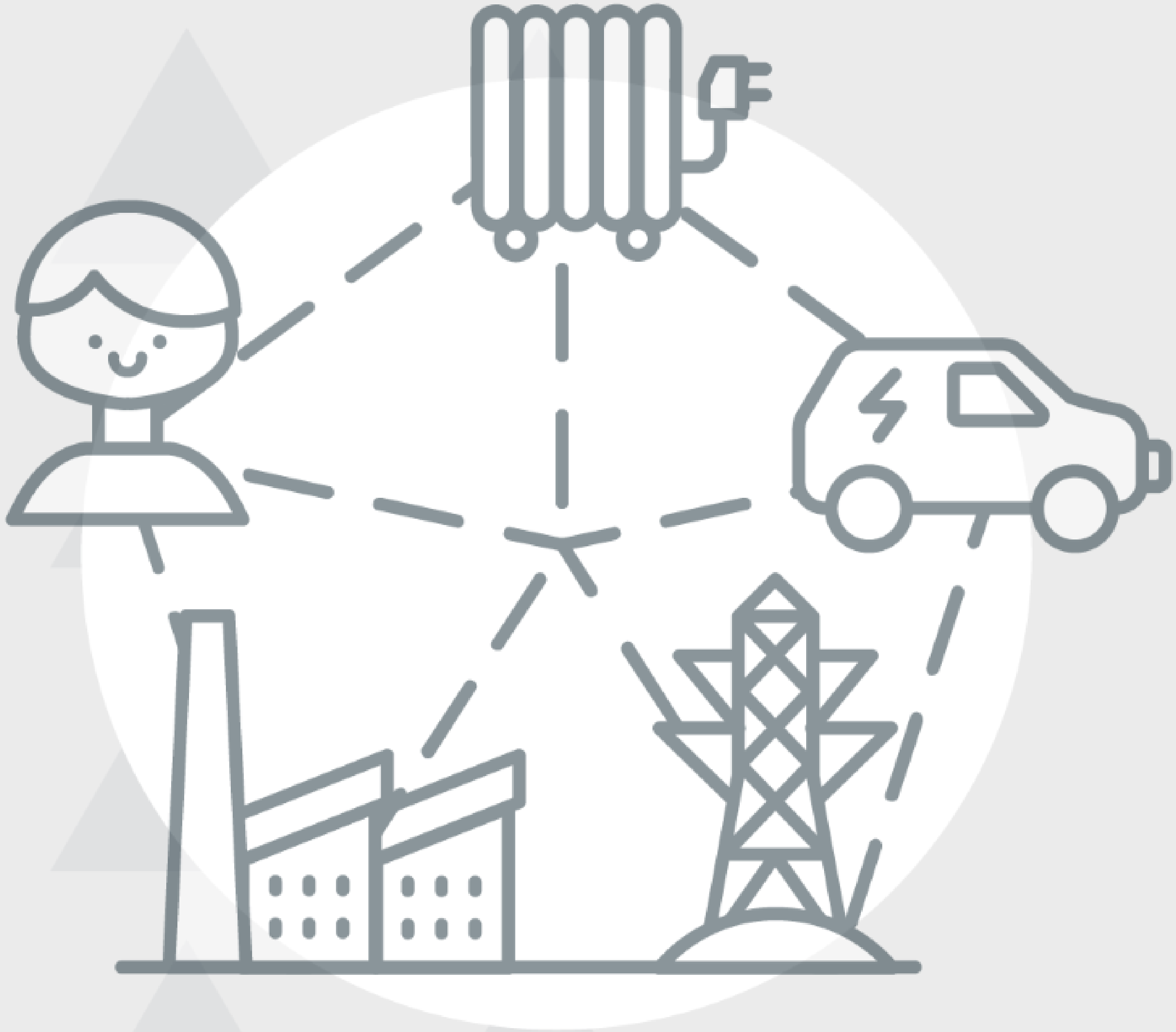
Openness Issues: Data, IP

Visibility & Discoverability

Accessibility & Reproducibility

Measuring Impact?

REPRODUCIBLITY AND ACCESSBILITY



"Reproducibility is the ability of ***independent investigators*** *to draw the* ***same conclusions*** *from an experiment by following the* ***documentation shared*** *by the original investigators"*

Odd Erik Gundersen*

- Princeton University "Leakage and the Reproducibility Crisis in ML-based Science" found 20 papers from 17 fields with reproducibility errors.
  - In turn effecting 329 papers.

- Example in Forecasting, M-competitions: "Although the test data and the submitted forecasts are all publicly available, the computed accuracy scores do not match those in the published paper" Hyndman, A brief history of forecasting competitions.

Limiting Issues (Accessibility and Reproducibility report):

- Data is **<u>not open</u>**
  - E.g. In review of low voltage load forecasting found only 24% use open dataset, and of those 42% come from a single source.

- Methodologies are not clear enough to be reproduced:
  - Methodology (Test-Training splits, benchmarks, hyperparameter optimisation, etc.)

- Code rarely shared (or if shared license may limit use, e.g. epftools)

Data leakage causes reproducibility failures in ML-based science

The running list below consists of papers that highlight reproducibility failures or pitfalls in ML-based science. We find 20 papers from 17 fields where errors have been found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. In each case, data leakage causes errors in the modeling process.

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Medicine	Bouwmeester et al.	2012	71	27	No train-test split
Neuroimaging	Whelan et al.	2014	—	14	No train-test split; Feature selection on train and test set
Autism Diagnostics	Bone et al.	2015	—	3	Duplicates across train-test split; Sampling bias
Bioinformatics	Blagus et al.	2015	—	6	Pre-processing on train and test sets together
Nutrition research	Ivanescu et al.	2016	—	4	No train-test split
Software engineering	Tu et al.	2018	58	11	Temporal leakage
Toxicology	Alves et al.	2019	—	1	Duplicates across train-test split
Satelitte imaging	Nalepa et al.	2019	17	17	Non-independence between train and test sets
Clinical epidemiology	Christodoulou et al.	2019	71	48	Feature selection on train and test set
Tractography	Poulin et al.	2019	4	2	No train-test split
Brain-computer	Nakanishi et al.	2020	—	1	No train-test split

From: Leakage and the Reproducibility Crisis in ML-based Science, Princeton University

* The fundamental principles of reproducibility, Odd Erik Gundersen, Royal Society Open Science, 2020.

LIMITATIONS IN LITERATURE: LV FORECASTING CASE STUDY: REVIEW 221 PAPERS

UNIVERSITY OF OXFORD

- Many papers use <u>no benchmark,</u> if they do it is non-competitive.
- Forecasts often ignored as part of an application
- Only 44 papers (<22%) utilised probabilistic forecasts.
- Missing Details:
  - Size of Validation/Testing period
  - Resolution of data
  - Forecast horizon
- Lack of Investigation into inputs and their effects, for example weather observations or forecasts.

Applied Energy 304 (2021) 117798

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

ELSEVIER

Review of low voltage load forecasting: Methods, applications, and recommendations

Stephen Haben [a], Siddharth Arora [a,*], Georgios Giasemidis [b], Marcus Voss [c], Danica Vukadinović Greetham [d]

[a] University of Oxford, UK
[b] Independent Researcher
[c] Technische Universität Berlin (DAI-Labor), Germany
[d] Tessella, Abingdon, UK

ARTICLE INFO

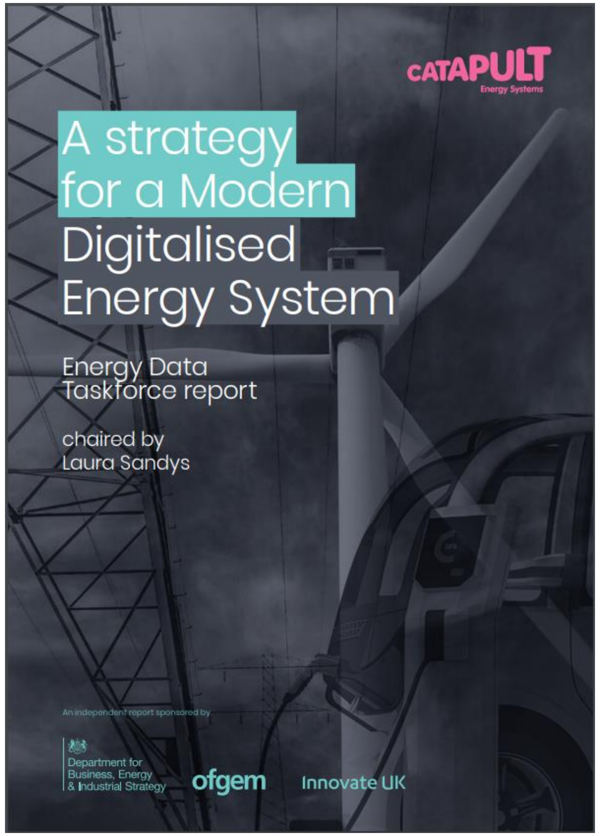
Keywords:
Low voltage
Smart meter
Load forecasting
Demand forecasting
Substations
Smart grid
Machine learning
Time series
Neural networks
Review
Survey

ABSTRACT

The increased digitalisation and monitoring of the energy system opens up numerous opportunities to decarbonise the energy system. Applications on low voltage, local networks, such as community energy markets and smart storage will facilitate decarbonisation, but they will require advanced control and management. Reliable forecasting will be a necessary component of many of these systems to anticipate key features and uncertainties. Despite this urgent need, there has not yet been an extensive investigation into the current state-of-the-art of low voltage level forecasts, other than at the smart meter level. This paper aims to provide a comprehensive overview of the landscape, current approaches, core applications, challenges and recommendations. Another aim of this paper is to facilitate the continued improvement and advancement in this area. To this end, the paper also surveys some of the most relevant and promising trends. It establishes an open, community-driven list of the known low voltage level open datasets to encourage further research and development.

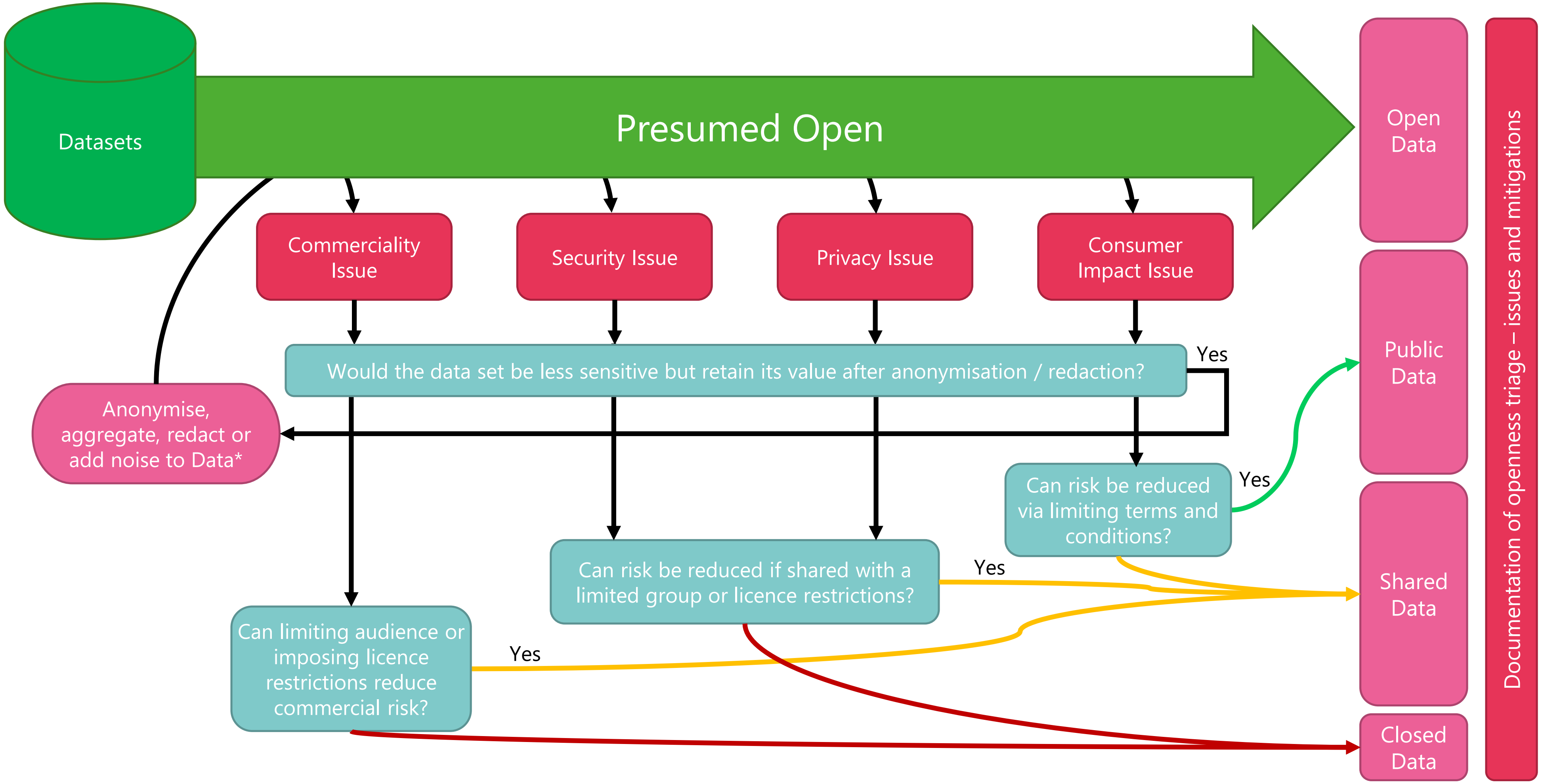
SOME SOLUTIONS

- **Opening Data:** Energy Data Taskforce (2019), "Presumed Open Data" principle. Now in Ofgem's "Energy Data Best Practice" guidance for energy networks.
- **Data Science Competitions:** Later Slides
- **Checklist for reproducible research (FATI Supplementary report 2)**
- **Tools for Releasing code (FATI Supplementary report 3)**: much more useful for understanding the nuances of the models
  - Version control systems like GitHub, or totally reproducible **binders**.
  - Utilise coding standards (Black)
  - Use appropriate license (try license selectors: http://ufal.github.io/public-license-selector/)
- **The Turing Way**: Collaborative open access book on reproducible research in data science https://the-turing-way.netlify.app/welcome



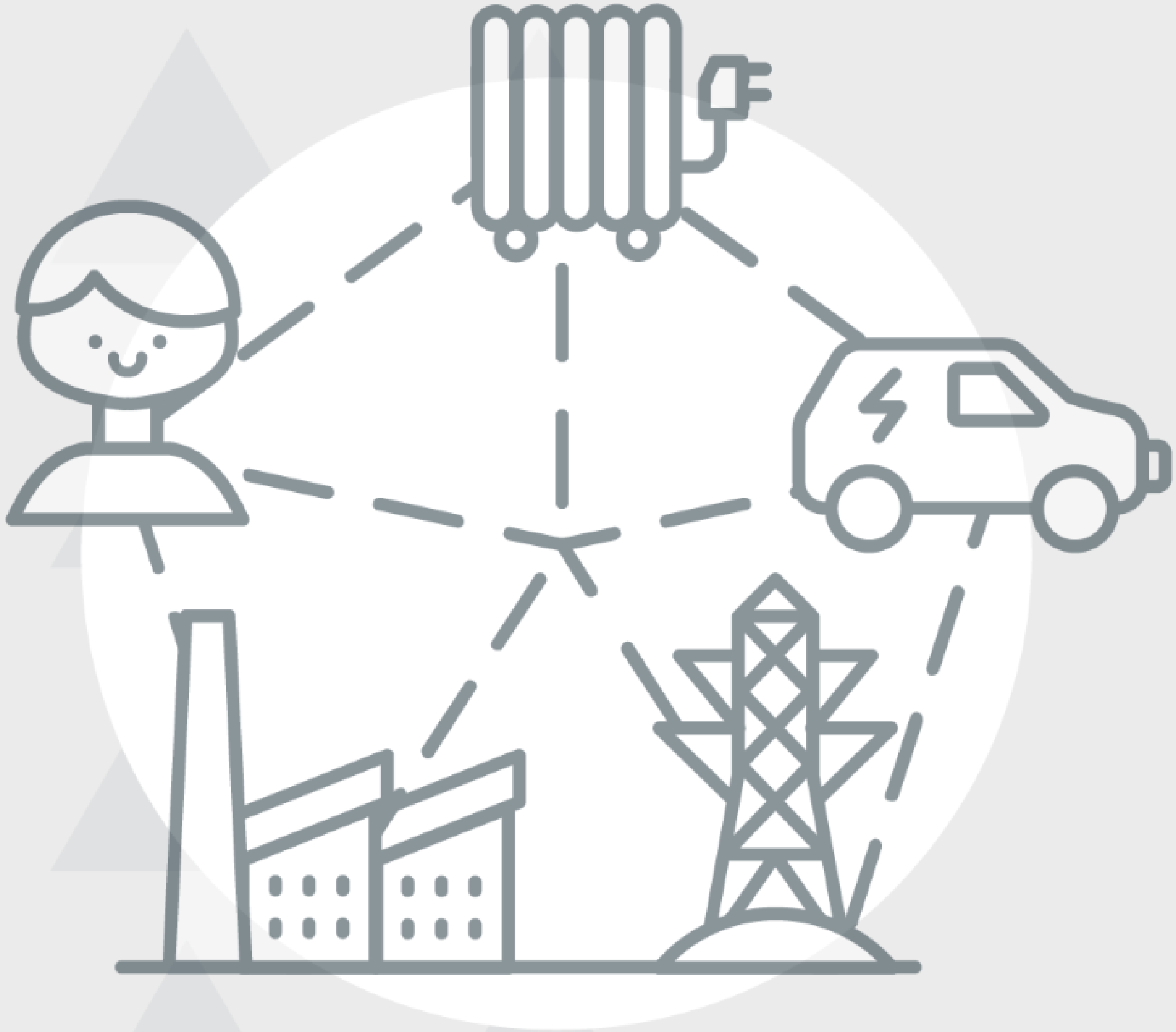
Category	Name	Description
Data	Modern	Ideally data is from the last few years. Data might be modern to when paper was published, but not modern anymore.
Data	No peaking	Data is available at time of predictions, respecting availability, trading times, data publication time etc.
Data	Open	Data is publicly available.
Data	Provided	Downloadable on Zenodo or other data platforms. I.e. you don't have to figure out how to get the data yourself, they make it explicit
Data	Size	Sufficient size for intended purposes, including for training and testing, and to take not account seasonalities and other features in the data.
Features	Split explicit	Train, test, validation split is explicitly stated in the paper.
Features	Realistic data split gaps	Realistic gaps and no overlap between training and testing data.
Features	Importance	Feature importance explicitly shown, including importance of lagged components if included.

PRESUMED OPEN DATA: TRIAGING OPENNESS



*Multiple stages of anonymisation / redaction may be required to address different issues (e.g. privacy and security) but repeated application should be limited

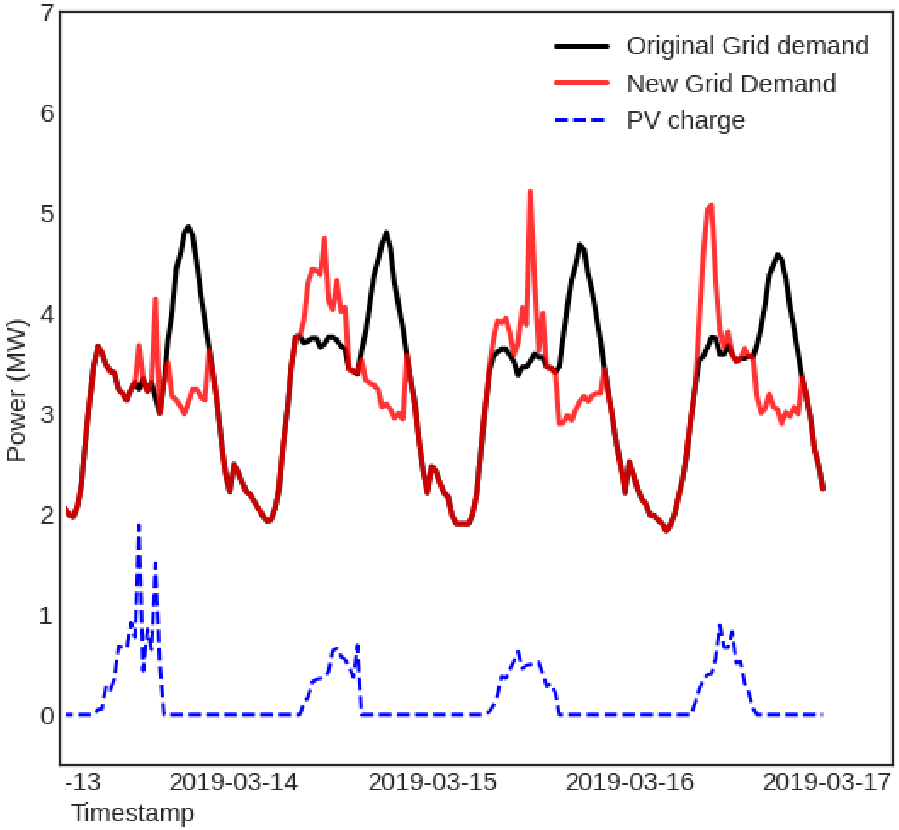
VALUE OF DATA SCIENCE COMPETITIONS.

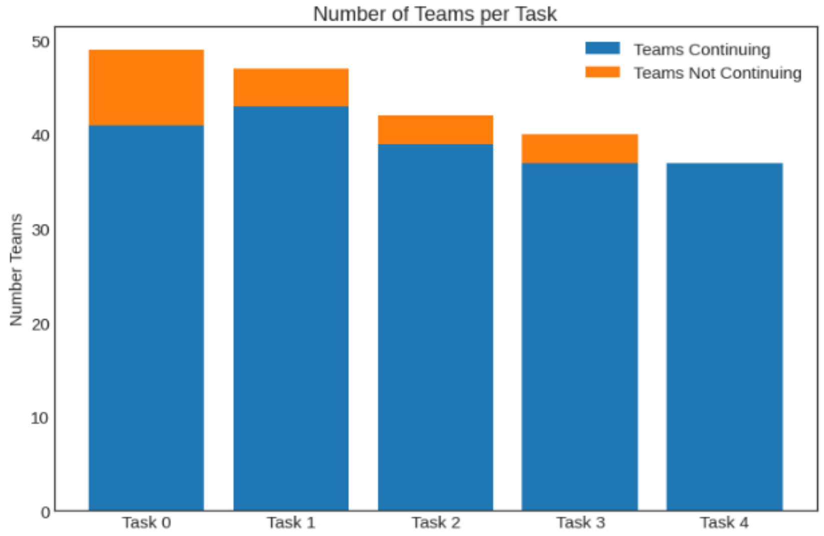


PRESUMED OPEN DATA

WPD INNOVATION
Transforming the electricity network

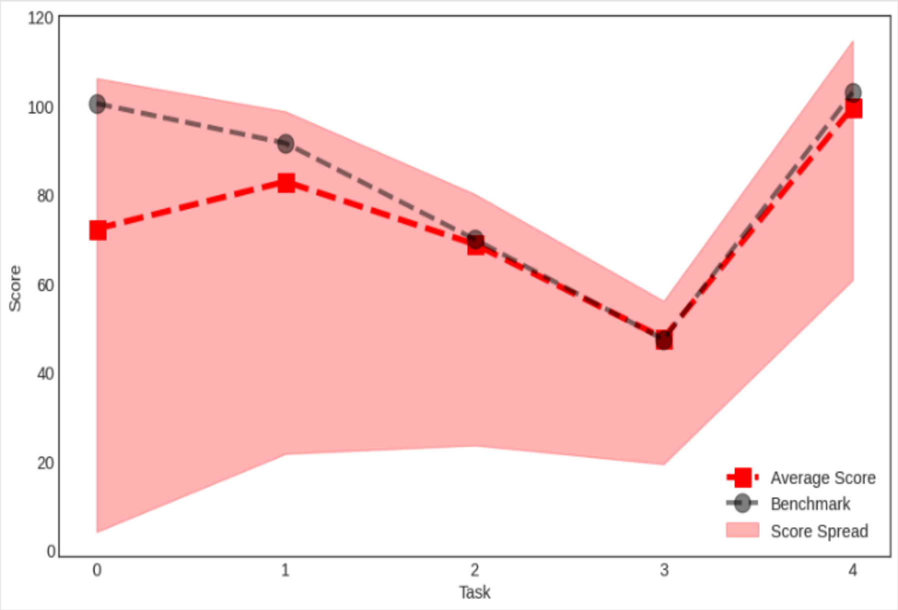
- Presumed Open Data Competition – an NIA project led by Western Power Distribution
- Two main aims:
  - Maximise the Visibility of Data.
  - **Maximise the Value of Data.**
- **Problem**: Design the control of a 6MWh/2.5MW battery storage device to support the distribution network to:
  - Maximise the daily evening peak reduction.
  - Using as much solar photovoltaic energy as possible.
- Five tasks over seven weeks.
- Prizes: Ideas Pitch and Publication in Energies Journal

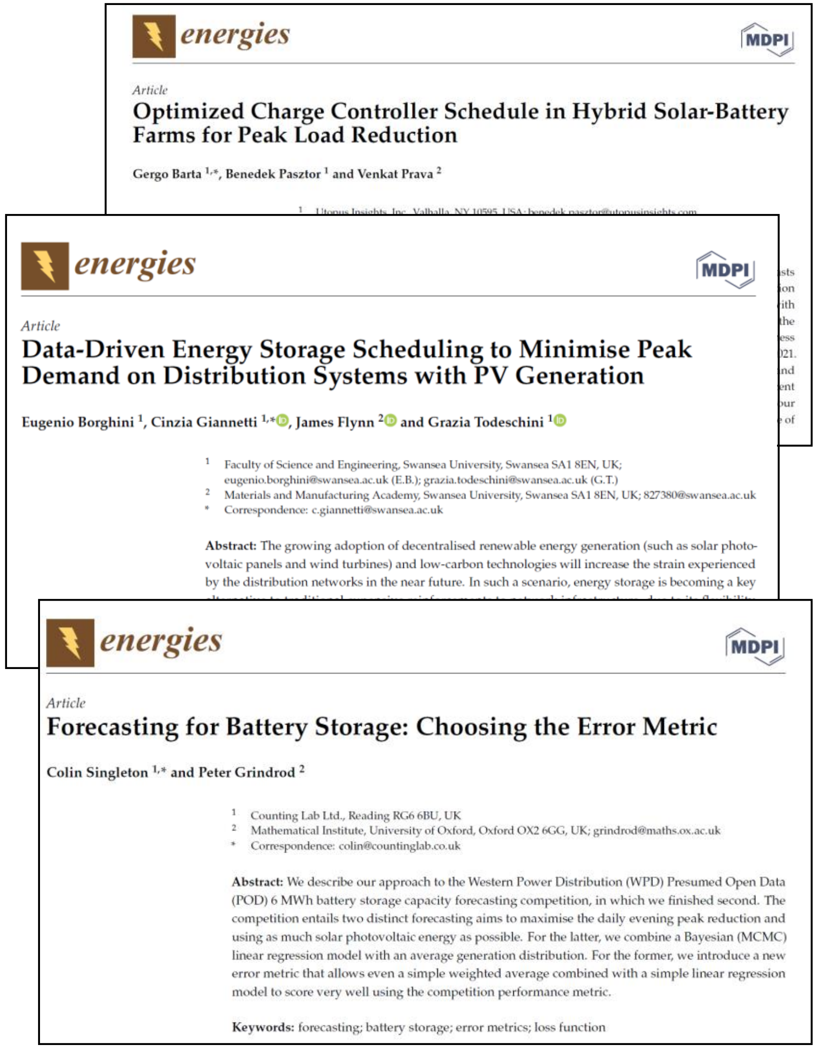




Outcomes

- Wide Participation: 55 teams - a total of 142 individuals – participated in a least one round. 72 different organisations/institutions
- Four teams openly released code (including winning team).
- Three peer-reviewed published papers.
- Illustrated diversity of solutions and approaches
- Illustrated improvement of solutions through experience.





energies MDPI

Article

Optimized Charge Controller Schedule in Hybrid Solar-Battery Farms for Peak Load Reduction

Gergo Barta [1,*], Benedek Pasztor [1] and Venkat Prava [2]

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

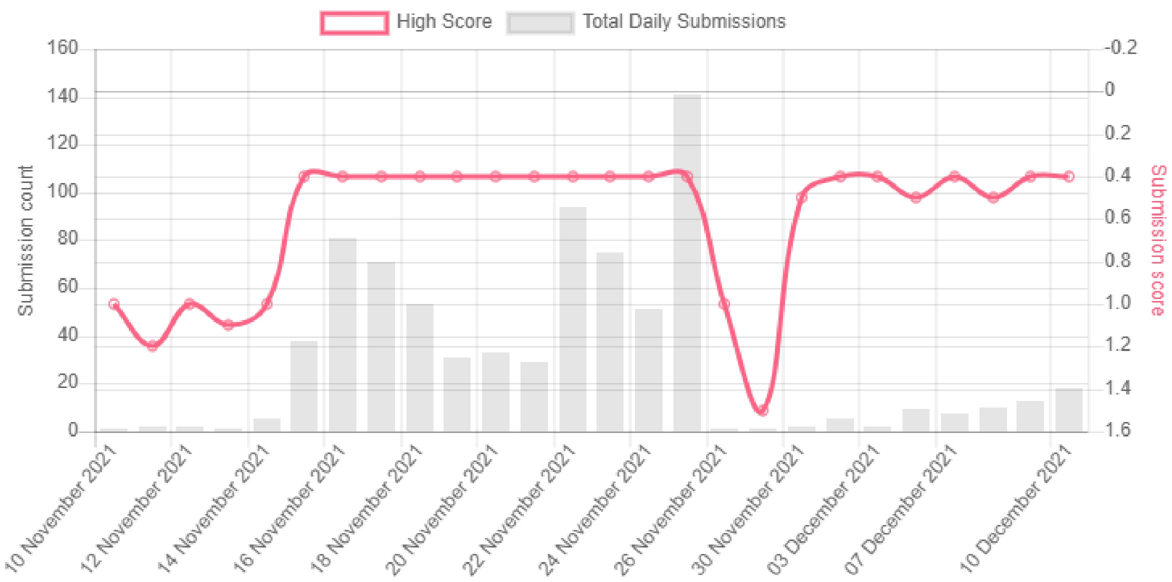
Abstract: We describe our approach to the Western Power Distribution (WPD) Presumed Open Data (POD) 6 MWh battery storage capacity forecasting competition, in which we finished second. The competition entails two distinct forecasting aims to maximise the daily evening peak reduction and using as much 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

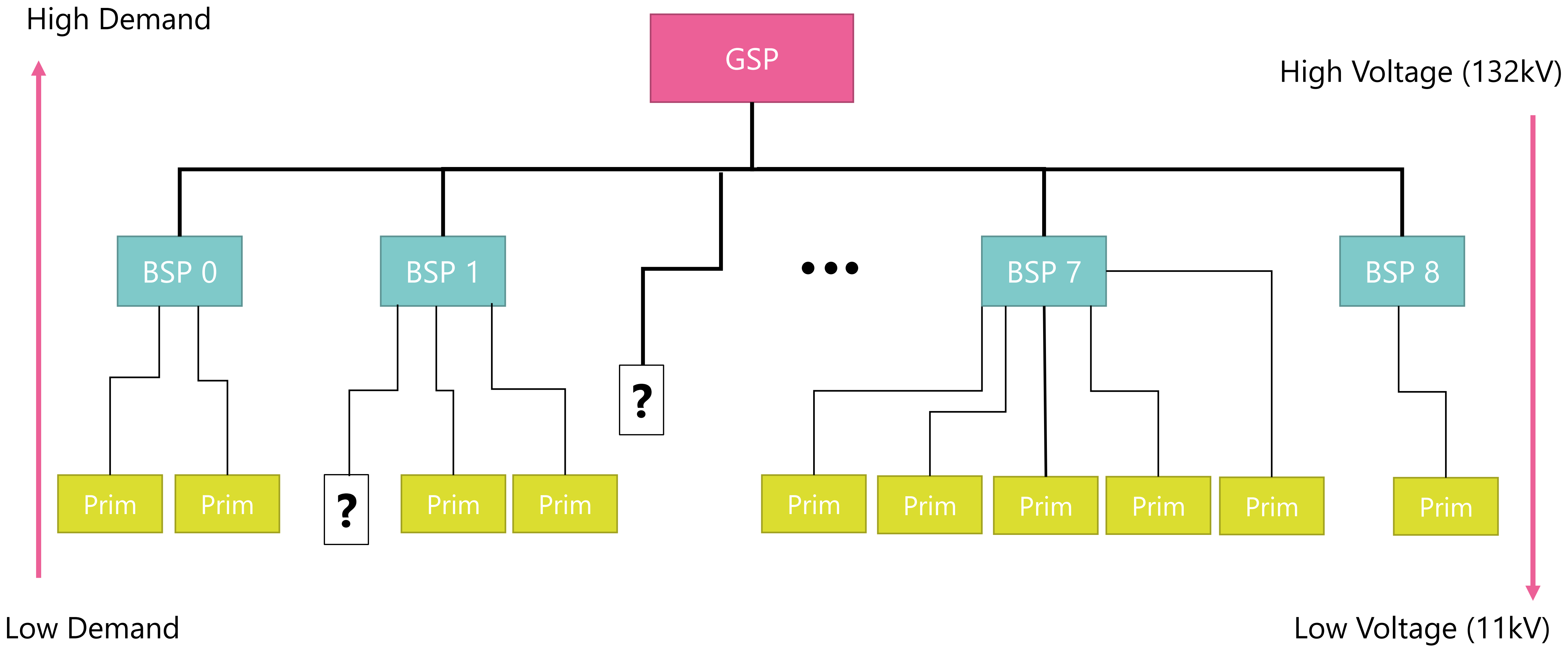
- **Three Network focused problems:**
  - High resolution feature estimation
  - EV detection
  - Missing data in network hierarchy

Data Science Learnings:

- Comparison of a variety of models:
  - Generalised additive models,
  - Artificial neural networks,
  - Ensemble methods,
  - k-nearest neighbours, etc.
- Demonstration of model combination
- Utilisation of weather and other data sources
- Novel and uncommon feature engineering methods

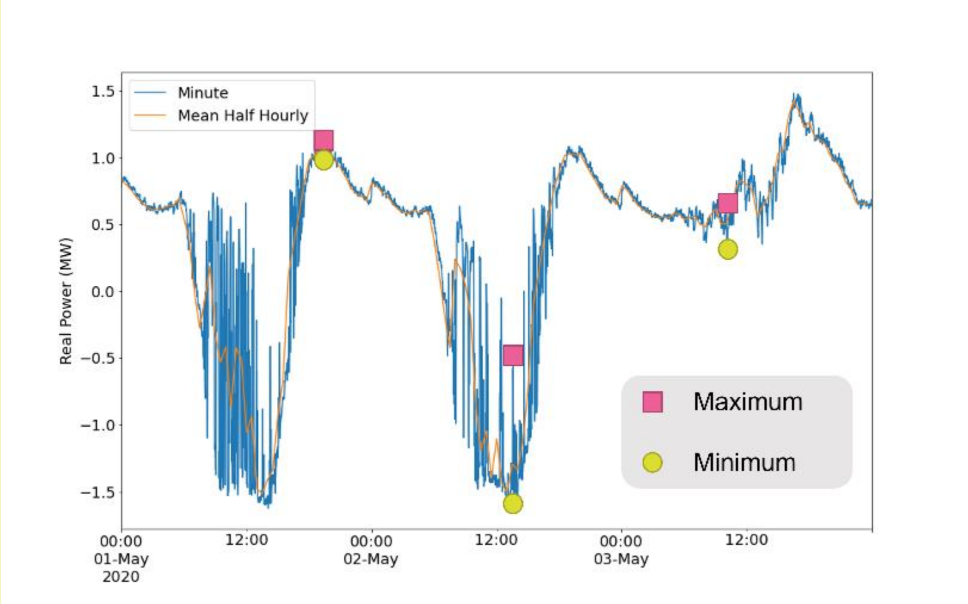


CHALLENGE 3: MISSING DATA IN NETWORK HIERARCHY



UNIQUE LINKED DATA SHARED

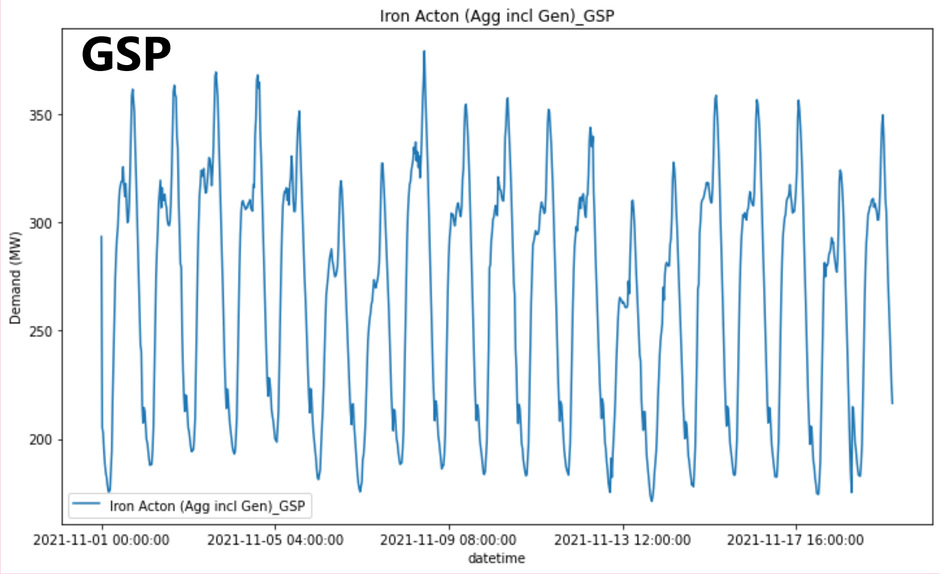
High-resolution Primary substation Data

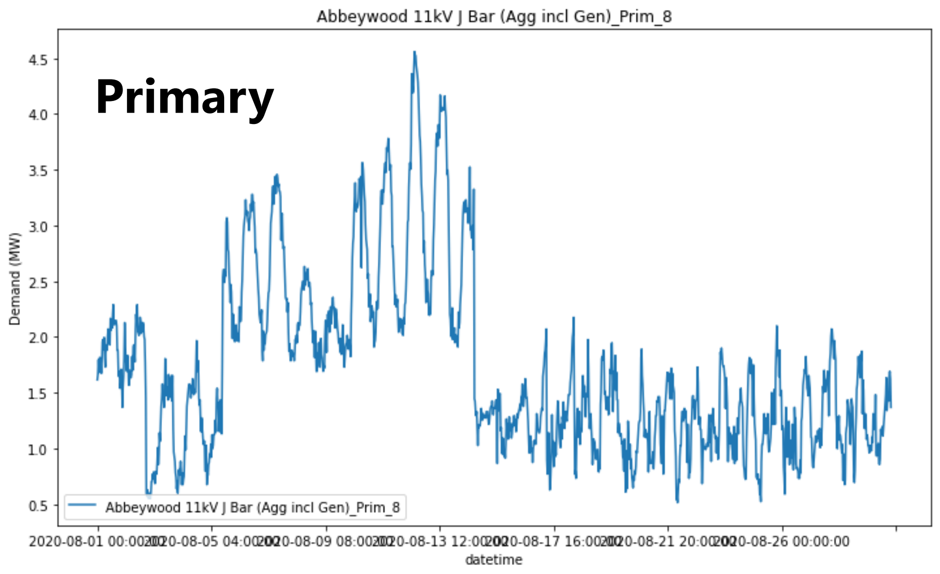


Linked Information:

- Location
- Distributed Generation
- Connectivity
- Capacities, asset ratings etc.
- Number and type of consumers
- Localised weather

Hierarchical Demand data: GSP-BSP-Primary

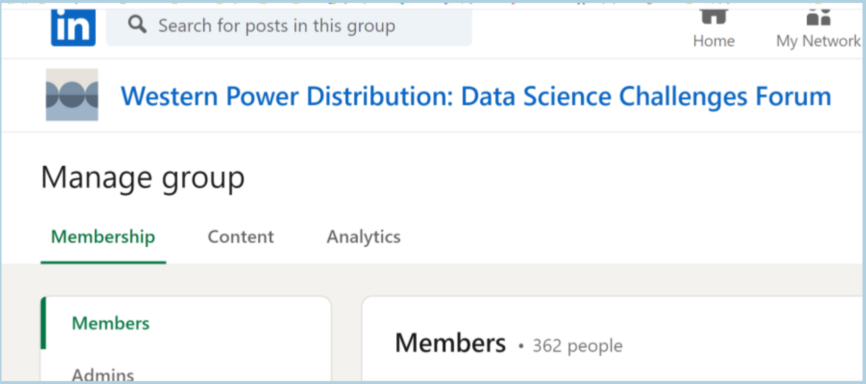




- Kick-off event participation:

Challenge	Date Kick-off	Registered	Attend	Video Views	Video Link
1	11th Nov 2021	321	100	324	https://www.youtube.com/watch?v=GkCC0odZOCo
2	20th Jan 2022	188	87	267	https://www.youtube.com/watch?v=KMCmlDhpN8o
3	15th March 2022	137	57	164	https://www.youtube.com/watch?v=2yc-K-x7Xaw

- From 20 to 100 hours spent on a single challenges by teams
- 373 LinkedIn members (Up from 120 prior to challenges)
- ~2000 Page Views and ~1000 Downloads of the Challenge data
- Over 2500 submissions and 37 Phase 2 Teams over all challenges



OUTPUTS AND PUBLICATIONS

High-Resolution Peak Demand Estimation Using Generalized Additive Models and Deep Neural Networks

Jonathan Berrisch*, Michał Narajewski*, Florian Ziel*
*University of Duisburg-Essen

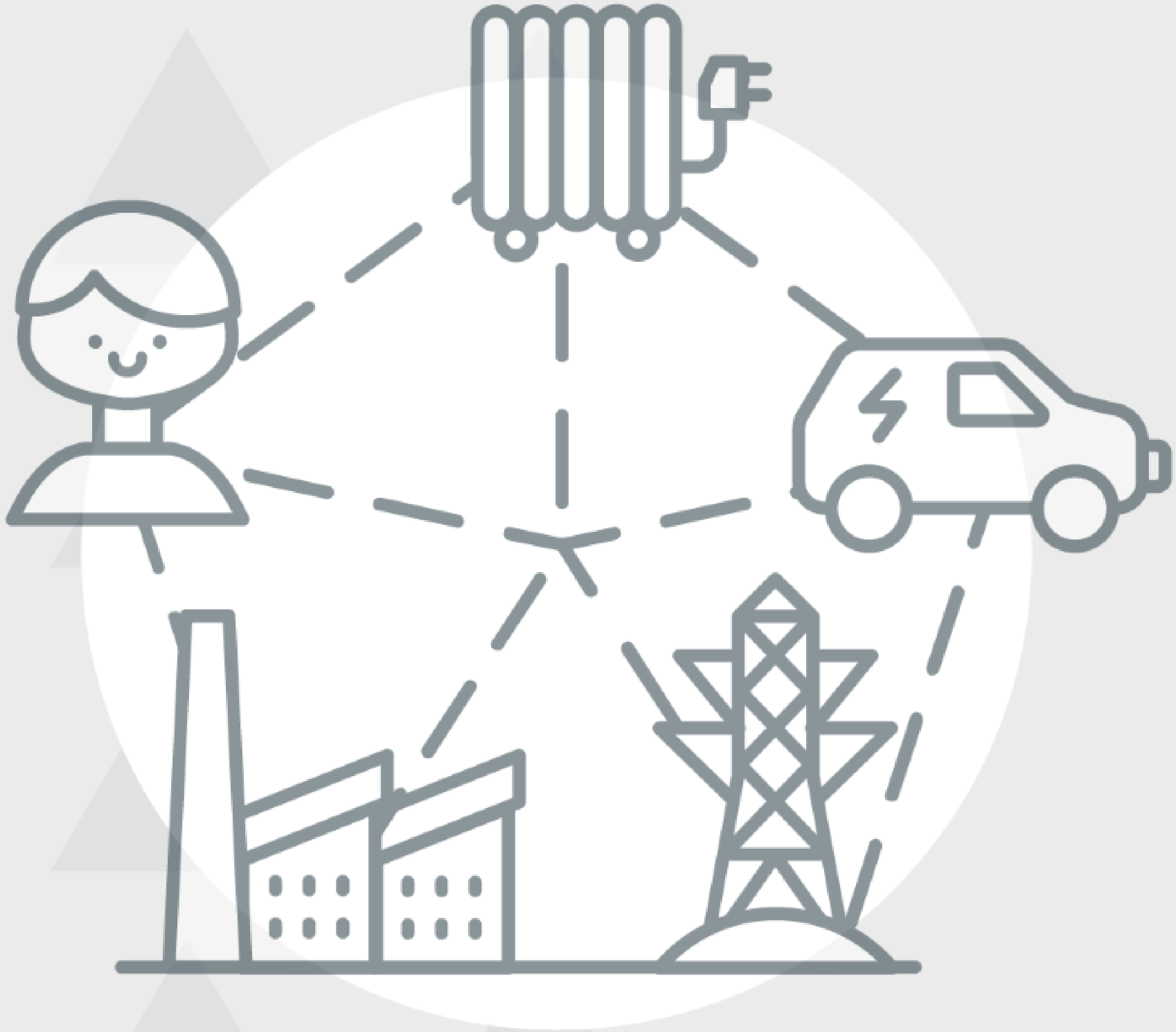
Knowledge, code and tools:

- High-Resolution Peak Demand Estimation Using Generalized Additive Models and Deep Neural Networks, Jonathan Berrisch, Michał Narajewski, Florian Ziel, Submitted paper from the winning team of challenge 1 with published preprint available here: https://arxiv.org/abs/2203.03342
- ESAIL Team Challenge 1 Code: https://github.com/AyrtonB/WPD-Hackathon Repository for the 4th placed team for the high-resolution feature estimation challenge.
- WOJJ Team Challenge 2 code: https://github.com/jsg16/WPD2-WOJJ Repository for the code for the second placed team from the EV estimation challenge.
- WOJJ Visualisation Challenge 2: https://jsg16.github.io/ Winning entry for challenge 2 visualisation prize.

Techniques and methods from top teams in each challenge:

- Challenge 1: https://www.youtube.com/watch?v=KMCmlDhpN8o
- Challenge 2: https://www.youtube.com/watch?v=2yc-K-x7Xaw
- Challenge 3: This Presentation (Will appear here: https://www.youtube.com/watch?v=gKVDLKnQJxo)

A BRIEF NOTE ON DATA SCIENCE SKILLS.

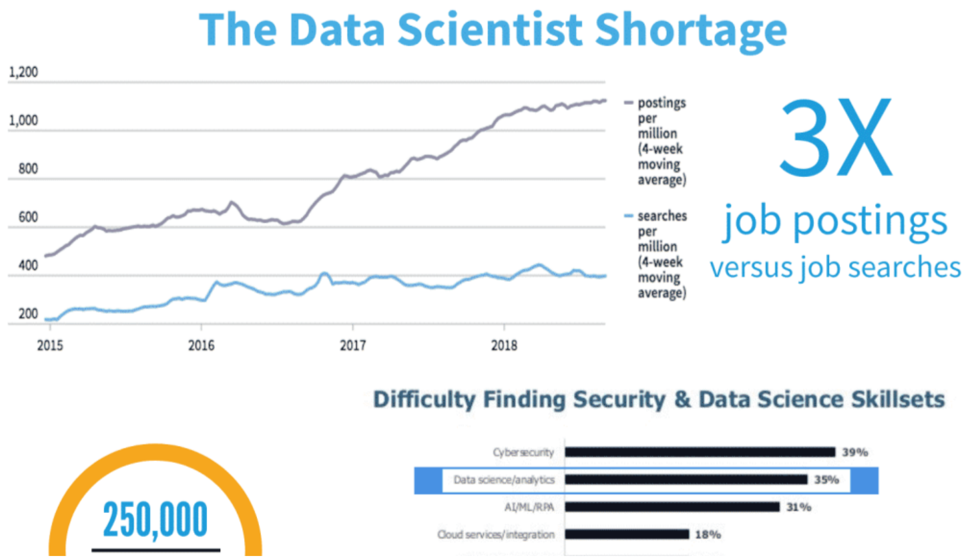


DATA SCIENCE SKILLS GAP IN ENERGY SECTOR

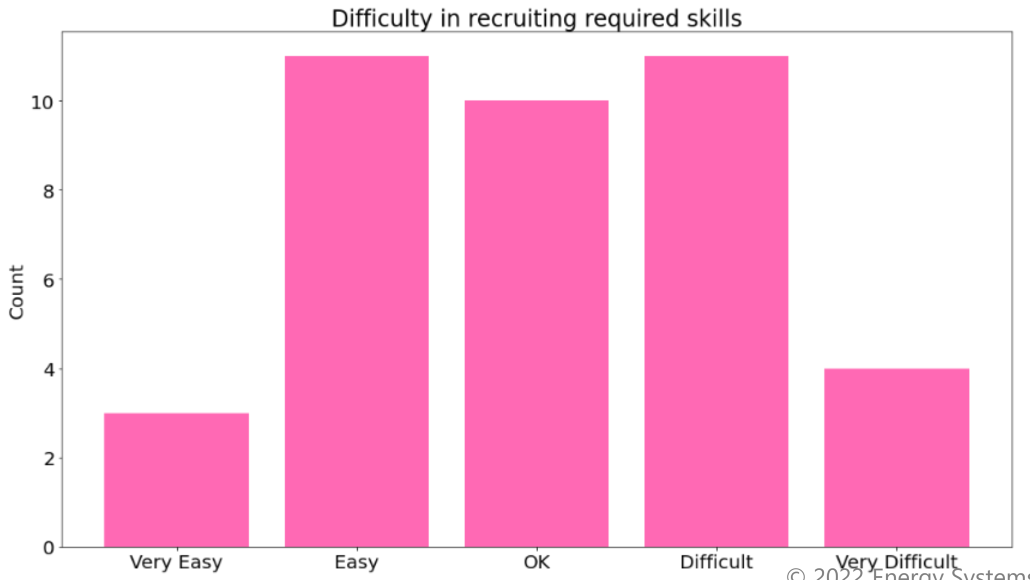
- Our own skills survey for data science in energy sector showed the three main traits found difficult to recruit:
  - Sufficient domain knowledge
  - Software/coding skills
  - Significant seniority and advanced data science skills

- Data Science report: Of ten data science focused master's programs from major universities in the UK we found only two that had dedicated introductory course to programming and none of them had any dedicated intermediate or advanced courses.

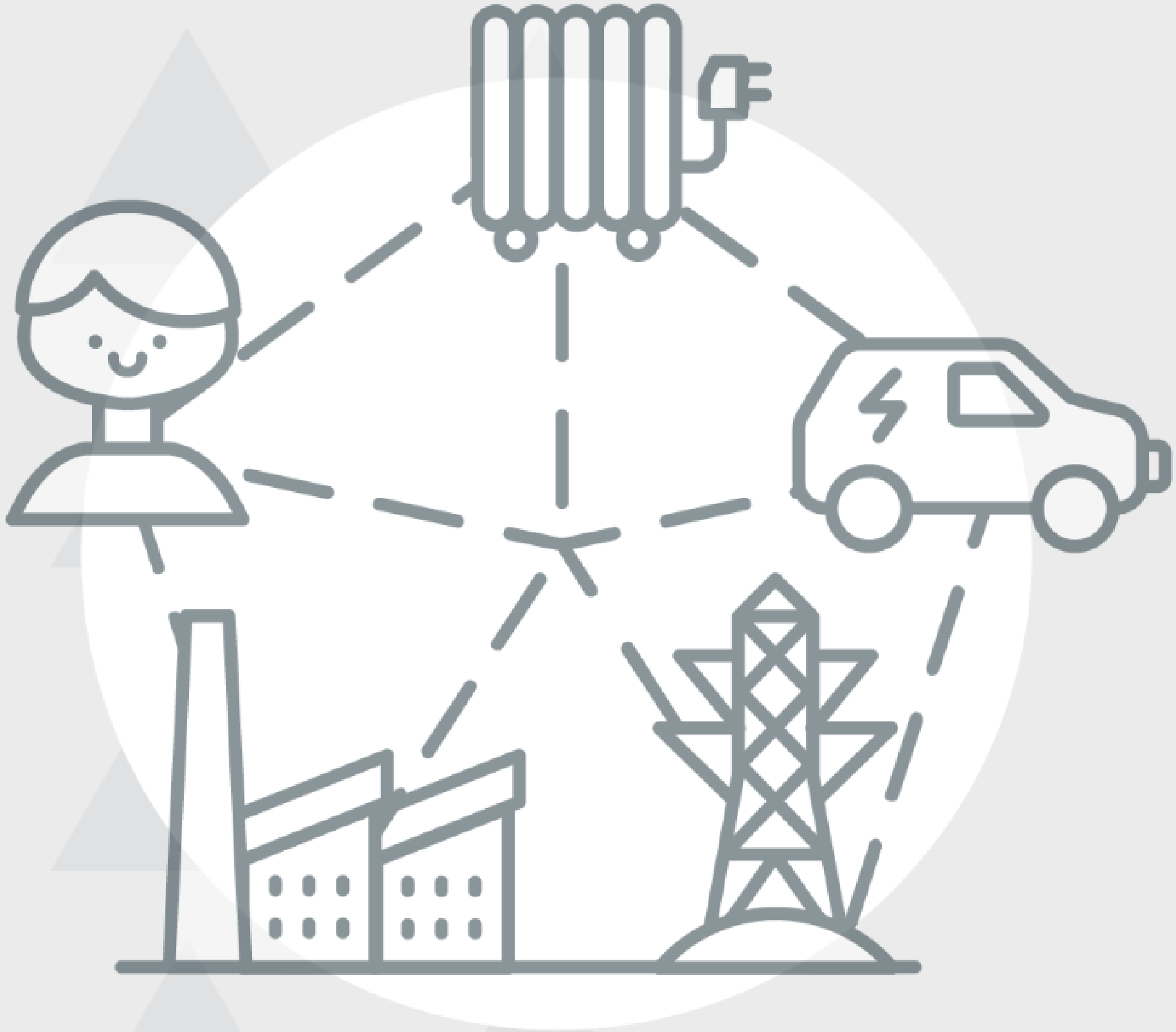
- Industry need to share **Latest technology and skills required:** "software development often 10 years or more ahead of the general academic research community"



https://quanthub.com/data-scientist-shortage-2020/



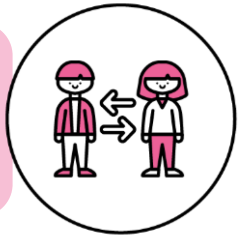
SOME CONCLUSIONS.



SOME SUMMARYS AND RECOMMENDATIONS

Accessibility and Reproducibility of Academic Outputs

- Focus on clear methodology and reproducibility published in open access journals or have an open preprints.
- Sharing (good quality) code with the paper using open data, and utilising common benchmarks.



Collaborations

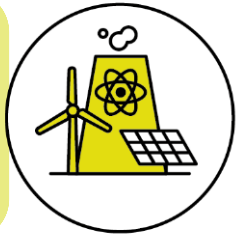
- Mutual understanding of culture and objectives, and utilise the best collaboration mechanism for the objective (Masters, PhD, Postdoc?)
- Consider strategic partnerships alongside the individual projects.

Code

- Focus on software skills and where possible, utilise industry knowledge and input of the latest technology.
- Academics should share code but there needs to be better incentivise code sharing

Industry Support

- Actively share (where possible) with the academic community: data, the current problems and code. Utilise data science challenges to facilitate this.
- Give feedback to academics when their code or research is used.



OTHER COMMENTS

Still Need More Data:
- Smart meter data still largely inaccessible
- Anonymisation could support further releases

Other Academic Support:
- Value in Energy Data Seminar series: https://www.youtube.com/playlist?list=PLkgx9FDaNeFmONK1-pXyn7p1eX_KnKaDZ
- Catalogue of Projects for Energy Data Platform: https://www.youtube.com/watch?v=iDn86IBzllg

Upcoming Papers:
- Data Science Skills
- Data ethics for Smart Local Energy Systems
- Algorithmic Governance

Going Forward
- Further Challenges.
- Supporting skills gap

OUR MISSION

TO UNLEASH INNOVATION AND OPEN NEW MARKETS TO CAPTURE THE CLEAN GROWTH OPPORTUNITY.

Energy Systems

THANKS TO: SAM YOUNG, SAM HINTON, CHRIS HARRISON, JON BERRY, LIAM MCSWEENEY, HOWARD THOMAS

STEPHEN HABEN

STEPHEN.HABEN@ES.CATAPULT.ORG.UK

ES.CATAPULT.ORG.UK

@ENERGYSYSCAT