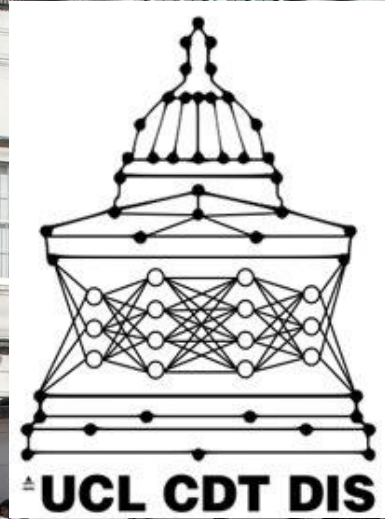


Industry and Wider Engagement

AI in HEP Exploratory Meeting

Gabriel Facini

October 1, 2024





Outline

- Facilitate fact-finding and discussion around challenges, barriers, opportunities in engaging with industry on AI matters
- HEP <-> Industry ML exchange
- Skills
- UCL Industry Projects
- Lessons
- Conclusions

What is the current status of industry/wider engagement in the HEP area?

- Advanced Tools/Algorithms Developed by Industry
 - pyTorch: Torch @ Idiap Research Institute (EPFL) -> pyTorch @ META -> Linue foundation
 - DeepMind [plasma control](#) w/ RL EPFL -> COSY [injection optimised](#)
 - **FastML open question – do we hit their level of complexity?**
- Computing Technology and Infrastructure
 - Cloud Computing Services - CERN utilizes Azure for scalable computing resources and Fermilab uses AWS (neutrino program?)
 - NVIDIA [blog](#) on LHC physics
- Quantum computing i.e. [IBM Quantum Computing and CERN](#)
- Data Challenges: [Higgs boson](#), [TrackML](#), [CMS collisions](#)
- Industry projects internship – not related to research, skills based

What is the current status of industry/wider engagement in the HEP area?

- **Summary: It is not a two way street. Industry leads, we follow**
 - Leading innovations in ML/AI via industry since the deep learning revolution
 - HEP is not regularly making CS-based publication that are foundational.
We apply the latest, greatest to our data
 - They make tools, hardware, services that we use
 - We make challenges to help them get interested in our problems

What are the challenges/barriers that we are facing in this area (e.g. generating interest in our research, engaging external/industry expertise, building lasting links/projects, generating income, industry churn, paperwork etc.)?

- Why should industry care about us?
 - Our research, more Higgs bosons, has zero street value
 - Honorary titles have some value for some people
 - Time is money and industry does not waste time (to first order)
- When are we useful?
 - **Our skills are in need – let me tell you about how I know.**
 - **Opportunity to upskill themselves**
 - Interests can align, but difficult to find. Hard requirement.

“Data is the driving force of the world’s modern economies.”

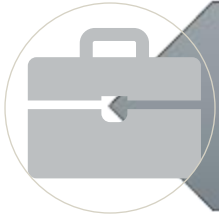
“It fuels innovation in businesses large and small and has been a lifeline during the global coronavirus pandemic. Effective use of data can boost productivity, create new businesses and jobs, improve public services and position the UK as the forerunner of the next wave of innovation.”

-- The Rt Hon John Whittingdale OBE MP
Minister of State for Media and Data



Department for
Digital, Culture,
Media & Sport₆

Skills Gap



Over [35,000](#) UK data scientist jobs on LinkedIn



only [10,000](#) new data scientists from university each year



Of the 54% of companies currently using the tech, 63% said they have a shortage in the skills required to properly implement it.



Some UK organisations i.e. Hartree Centre have a mandate to upscale industry. Training professionals to use AI.

Industry and Academia Hybrids (Hartree/SciML)

- Hartree/SciML are big organisations, tons of compute, top talent making collaborations/contract with big name organisations or for big problems (i.e. fusions)
 - Is this true?
- Is so, clearly something to offer industry – resources to solve big problems
- CERN OpenLab, Swiss Data Science Center – housed with experts

CASE and Knowledge Exchange

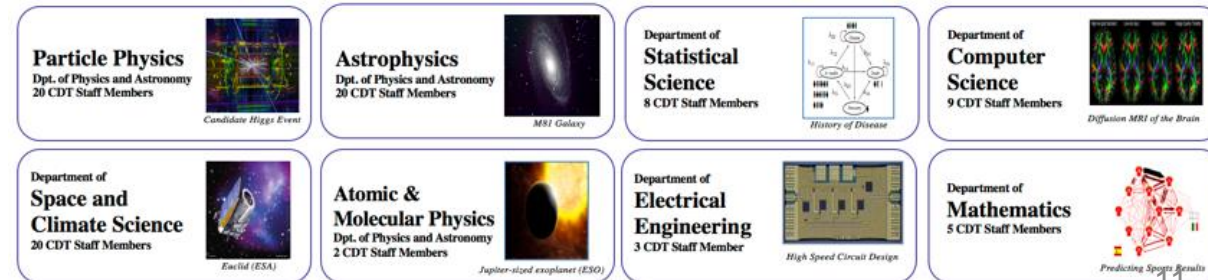
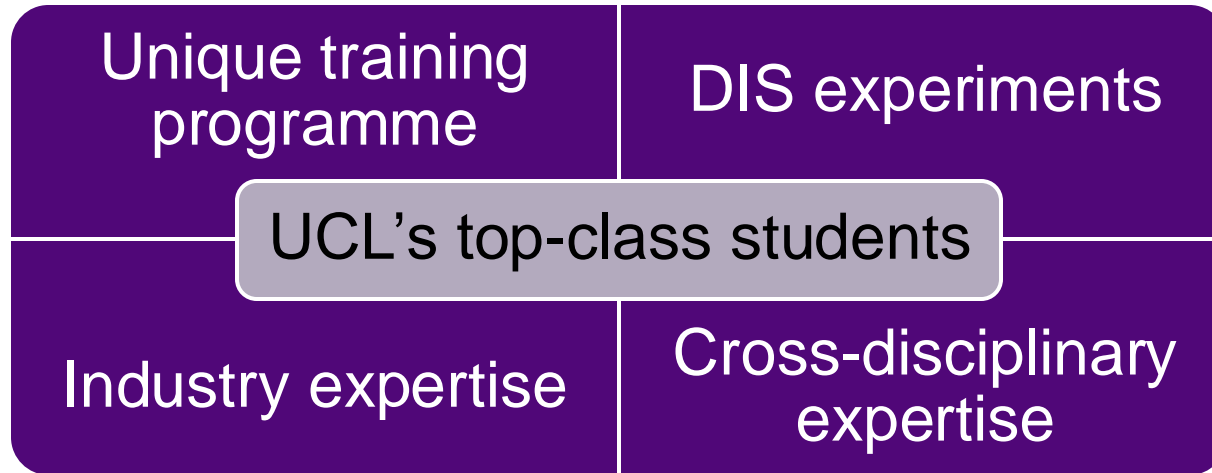
- CASE projects?
 - The STFC Industrial CASE (Cooperative Awards in Science and Technology) studentship competition provides support for PhD students to work in collaboration with a non-academic partner on projects that fall within the STFC core science programme in astronomy, particle physics, nuclear physics and accelerator science, or on projects that aim to apply technologies or techniques developed within the programme into other areas.
 - [list](#) "ML applied to galaxy formation and neutralising security threats."
- KIP – offer subsidies a PDRA-level hire in a company and 10% of academic time total for 2 academics. More money for smaller company
 - too slow for start-ups (9-11 months)
 - Haven't found a company who is interested – still looking.

Training Grounds: CDTs

- STFC Funded CDTs (2022, not 2016)
 - UCL CDT
 - University of Cambridge CDT
 - University of Liverpool CDT
 - Partnered with Liverpool John Moores University
 - University of Sussex CDT
 - Partnered with Queen Marys University of London, and the Open University
 - NUdata STFC Centre For Doctoral Training In Data Intensive Science
- UKRI artificial intelligence Centres for Doctoral Training
 - 12 funded in 2023 ([link](#))
- Will talk about UCL CDT as it is my experience

A Centre of Doctoral Training (CDT)

Year	Activities
Year 1	<ul style="list-style-type: none"> Taught courses Group project Exams PhD project assignment Software (SW) Carpentry CDT Summer School <p>Transferable Skills Communication skills, Scientific writing, Media training</p>
Year 2	<ul style="list-style-type: none"> MPhil to PhD transfer Placement assignment SW Carpentry (tutor) <p>Transferable Skills Entrepreneurship, Intellectual property, Science in the economy</p>
Year 3	<ul style="list-style-type: none"> Placement International training school CDT Summer School (tutor) <p>Transferable Skills Research planning, Proposal writing</p>
Year 4	<ul style="list-style-type: none"> International conference PhD Award <p>Transferable Skills Interview skills, Careers workshop</p>



Training Highlights

- PhD & MSc – Year 1, Term 1 Courses (frontload)
 - Research Computing with C++ and/or Python
 - Machine Learning and Big Data
 - High Performance Computing
 - Statistics and Data Analysis Techniques
 - Responsible Research & Innovation
 - Software Carpentry

• PhD – Year 1, Term 2: **Group Projects**

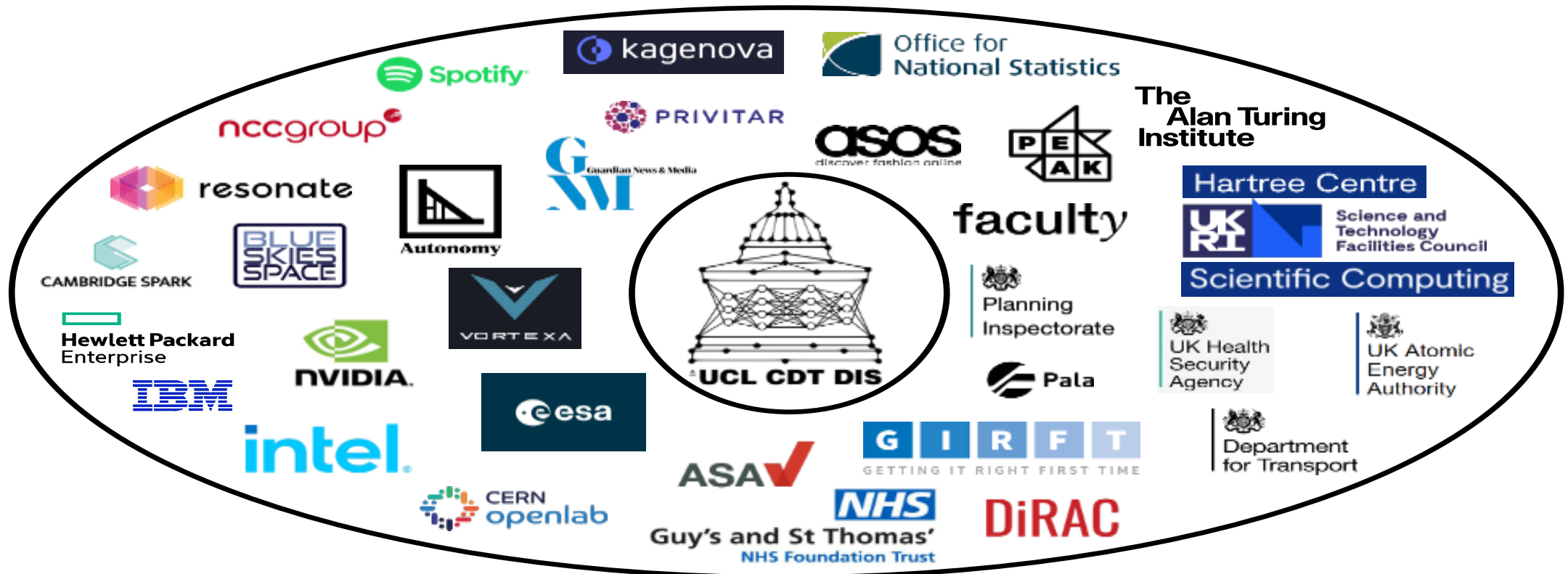
• MSc – Term 3: Thesis Project

• PhD – Year 2/3: **Industry Placement**

	Activities	
Year 1	<ul style="list-style-type: none"> • Software Carpentry • Taught courses • Industry Group Project (IGP) • Responsible Research & Innovation (Part 1) • DIS Summer School <p><u>Transferable Skills</u></p> <ul style="list-style-type: none"> • <i>Entrepreneurship</i> 	PhD research project - Discussion forums – Seminar series – Annual CDT events
Year 2	<ul style="list-style-type: none"> • MPhil to PhD transfer • Choice of Placement Partner • Responsible Research & Innovation (Part 2) • Software Carpentry as tutor • International Research Secondment <p><u>Transferable Skills</u></p> <ul style="list-style-type: none"> • <i>Communications and media training</i> 	
Year 3	<ul style="list-style-type: none"> • Placement • DIS Summer School as tutor <p><u>Transferable Skills</u></p> <ul style="list-style-type: none"> • <i>Management and Leadership</i> • <i>Research Planning, proposal writing</i> 	
Year 4	<ul style="list-style-type: none"> • IGP as mentor • PhD thesis writing, submission, defense • Award of PhD <p><u>Transferable Skills</u></p> <ul style="list-style-type: none"> • <i>Advance entrepreneurship</i> • <i>Interview skills, Careers workshop</i> 	

Partners

- Over 30 partners active in CDT
 - Projects, training, seminars
- Have invested over £1M in the Centre now
- **Continue to attract new partners**



Group Projects

What are they?

- Partner-defined & supervised project
- 3-4 first-year PhD students, 50% from January to mid-April
- 4-5 / year
- > 15 papers published from this work so far

MSc involvement

- extend projects from May – August

Benefits:

- Partner-student relationships start
- Partner project advances & can upscale skills
- Students get 1st taste of "real world"

Group Projects

16 different partner organizations

23 projects completed

- NCC: Cyber-security implications of deepfakes ([blog](#))
- ASOS: Detecting intent with natural language ([paper](#))
- TfL: Trains failure prediction
- Economist: Content Value/Influence Assessment

4 from 2023

- *Peak AI: Supply chain management via reinforcement learning*
- *SWA: Punch identification*
- *Guardian: Quote identification*
- *LDC: Identifying antibiotic-resistance bacteria in wastewater*

Group Projects: Take away

1st year PhD students can add value in an industry setting with 0.5 FTE over 12 weeks

Some will even pay them to continue at hourly rates

- *More useful to PhD student than waiting tables*
- *Would not agree on day 1, but do agree after working together*

Potential to upskill is useful for everyone

- upskill at the company (new techniques, supervision) is valuable
- Students skills enhanced
- Minimal impact on academic's skills (failure)

Industry Placements

What are they?

- Student work for 6-months in partner organization:
 - On specific project(s)
 - As member of team dealing with a class of projects
 - Anything that makes sense

Benefits:

- Cheap labour for partner
- Partner talent pipeline
- Knowledge transfer/stronger links with industry
- **Co-funding of student generates funds matched by UCL**

Placement Success Stories

- Including those agreed 42, with 23 partners. A few examples:
 - ASOS: Auto-detection and categorization of items in fashion images
 - Turing: Air-traffic control communication with reinforcement learning
 - Faculty: “AI Safety” representative learning to mitigate data compromise
 - SWA: punch identification, combat sport scoring and brain injury studies
 - Vortexa: Fill in gaps in ship data when tracking global trade
 - Babylon Health: ML to find causal relationships to improve diagnosis
 - HPE: HPC diagnostics

Placement Take Aways

- Great value, but only sometimes repeated often
 - In some cases, we have repeat customers. Others, limited by budget, effort (paperwork), and customisation (already have internships)
 - Cost are tough to discuss when decision maker is a few floors up
 - Start-ups are good for that reason
 - Head count limits are in affect
 - Gov't has neat work-arounds (it seems)

What are the challenges/barriers that we are facing in this area?

- **Research interest?** Curiosity driven, not business driven (not enough)
- **Spreading their influence?** Some platform companies will try hard to get you to use their platforms (esp new quantum companies).
 - Some will offer personal to help with problems if commercial gains are high enough to warrant their time (via payout from IP)
- **Lasting links?** Industry is dynamic. People move and situations change i.e. economic downturns
- **Generating income?**
- **Paperwork?** Yes! University paperwork and company paperwork.
- After have project, main challenge: demonstrating enough value that cost of ramp-up time is worth the effort.

What are the opportunities in this area for HEP

- **It is easier to make projects with no requirement on alignment**
 - More alignment, higher bar to satisfy due to needs in HEP
- **Enhanced outcomes:** yes, for student careers
- **Access to new tools/techniques/hardware:** case dependent (quantum?)
- **Co-creation of new tools:** maybe, see alignment above
- **Increased funding opportunities:**
 - I think so – 1st year PhDs have value
 - “Longer in academia, harder to break bad habits”
- **Knowledge exchange:** likely (but useful?)
- **Cultural exchange:** likely and useful!
- **Impact:** yes

More questions

- What is the most important gain from industry?
- How do we ensure we capture that?
- How do we build upon these examples and grow community-wide, to everyone can benefit?
- Can we share relationships?
- Can we get to the a level where industry comes looking for us?

Conclusions

- Industry does not need HEP to continue leading in AI/ML
 - They lead, we follow.
- The best that we can provide for them today is well-trained personal
 - Perhaps niche needs like FastML in the trigger?
 - HEP does not use cutting edge tools always
 - Definite shift from younger generations (great!)
 - Looking for baked in skills – hire who can do it today
 - We need to make sure the education and projects are on the mark – see UCL CDT
 - Soft skills are lacking
 - Business sense, communication, etc
 - Can grow with industry projects
- Bottom line: How does working with HEP at any level improve the bottom line. If not >> £0. No go...(to harsh?)

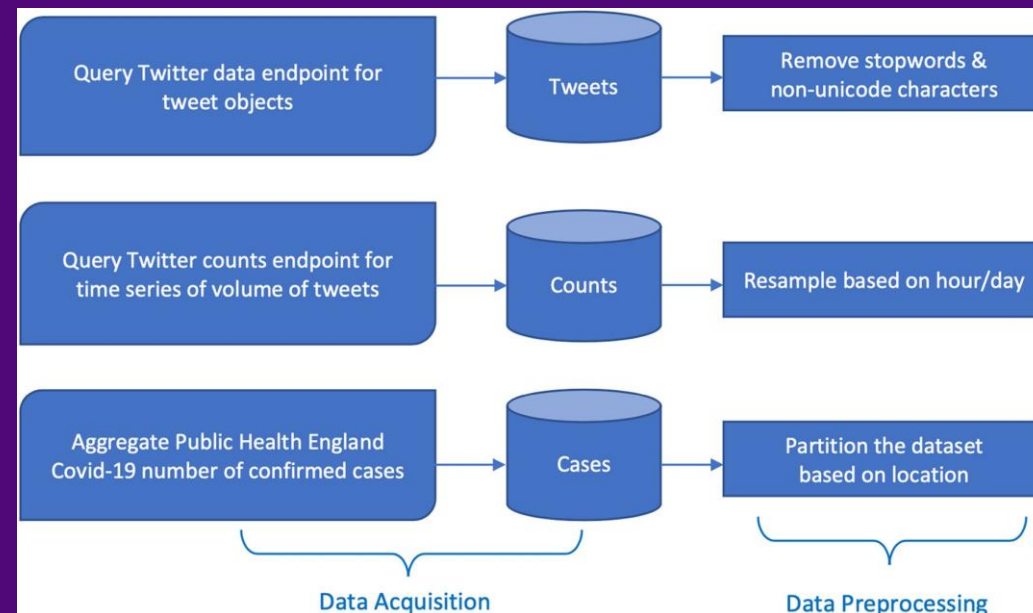
GP: UK Atomic Energy Agency

- UKAEA has been a partner since the get go. A few group projects:
 - Fast Regression of Tritium Breeding Ratio in Fusion Reactors ([paper](#))
 - Surrogate models of gyrokinetic turbulence with active learning ([poster](#))
 - Automated spatial calibration correction
 - Hyperspace mapping of MAST plasma shots

GP: Office of National Statistics

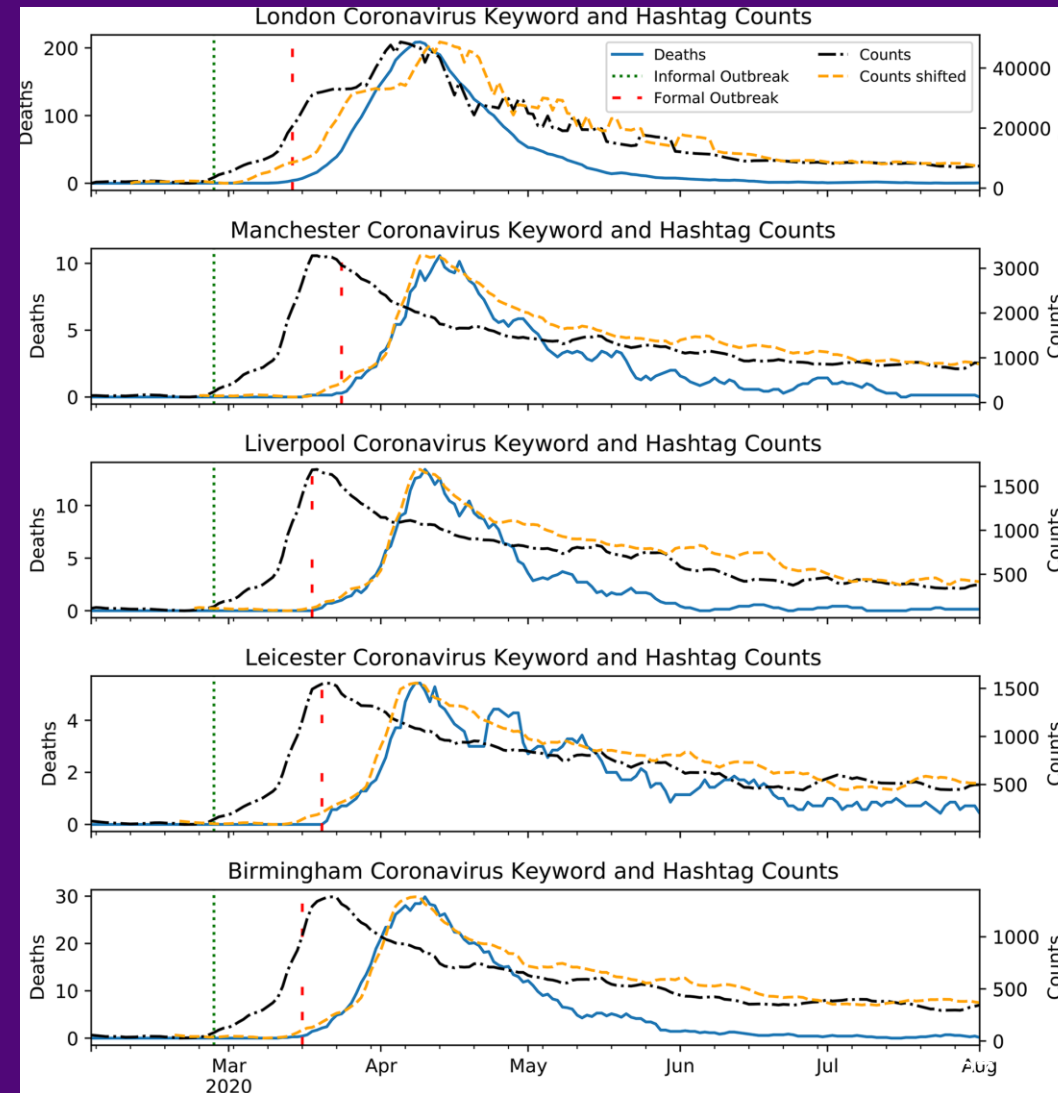
- Pitched in late 2019 – Can we use Twitter to make an "instant" poll?
- Designed Twitter pipeline to scrape for Brexit and COVID tweets
 - Published [paper](#) on COVID tweets from Jan – August 2020

- Designed custom pipeline avoiding limitations of twitter sampling and API



COVID: Wisdom of the Crowds

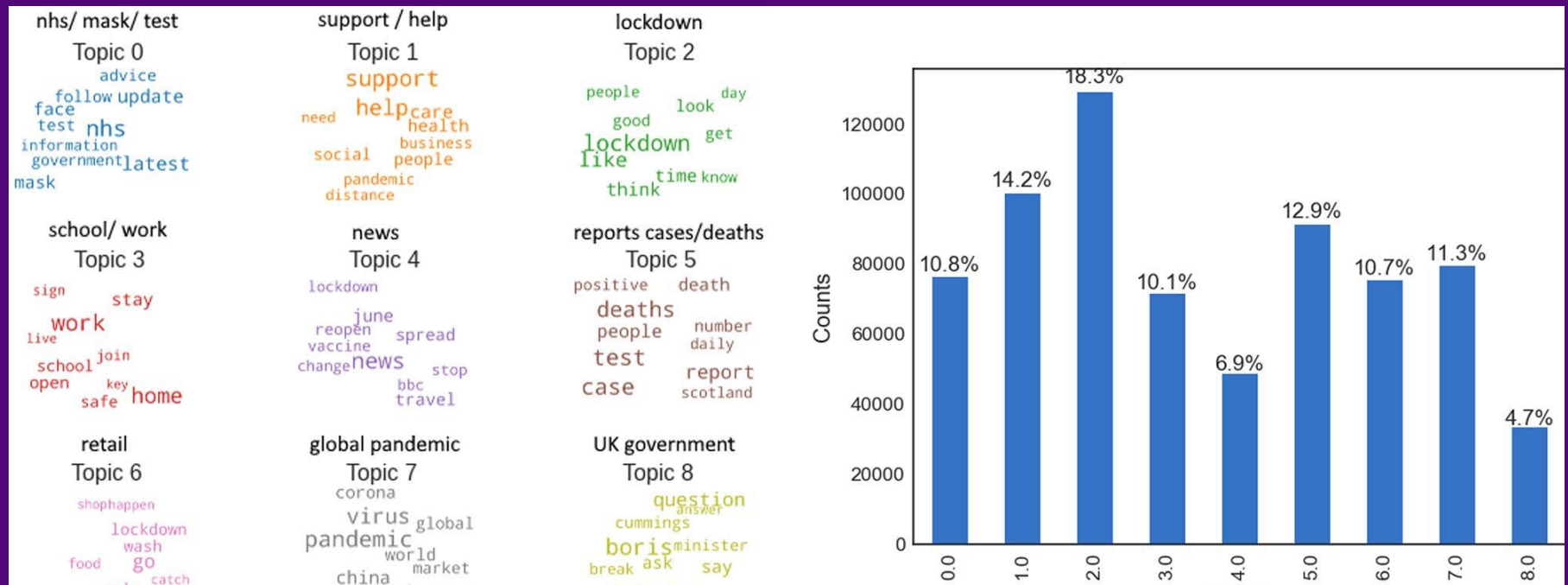
- (Recall early 2020...madness)
- Used occurrences of covid-related tweets to compare start of “informal” and “formal” outbreak
 - 6-27 day lag, matching other sources
- Lengthy discussion about the viability of the method
 - Hashtags evolve
 - No focus on symptoms



COVID: Topics and Sentiment

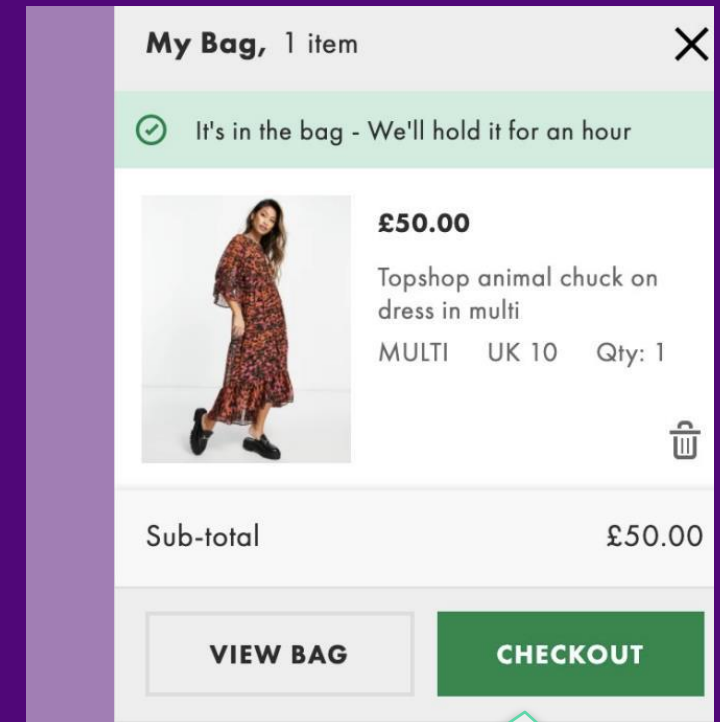
- Topic modelling via Latent Dirichlet Allocation (LDA) to isolate topics and Named Entity Recognition (NER) to learn the vector space of phrases to discriminate
- NER model returned F1 score of ~80% which matched human labelled results in literature

– Can isolate sub-sets i.e. “masks”



GP: ASOS Predicting Returns via GNN

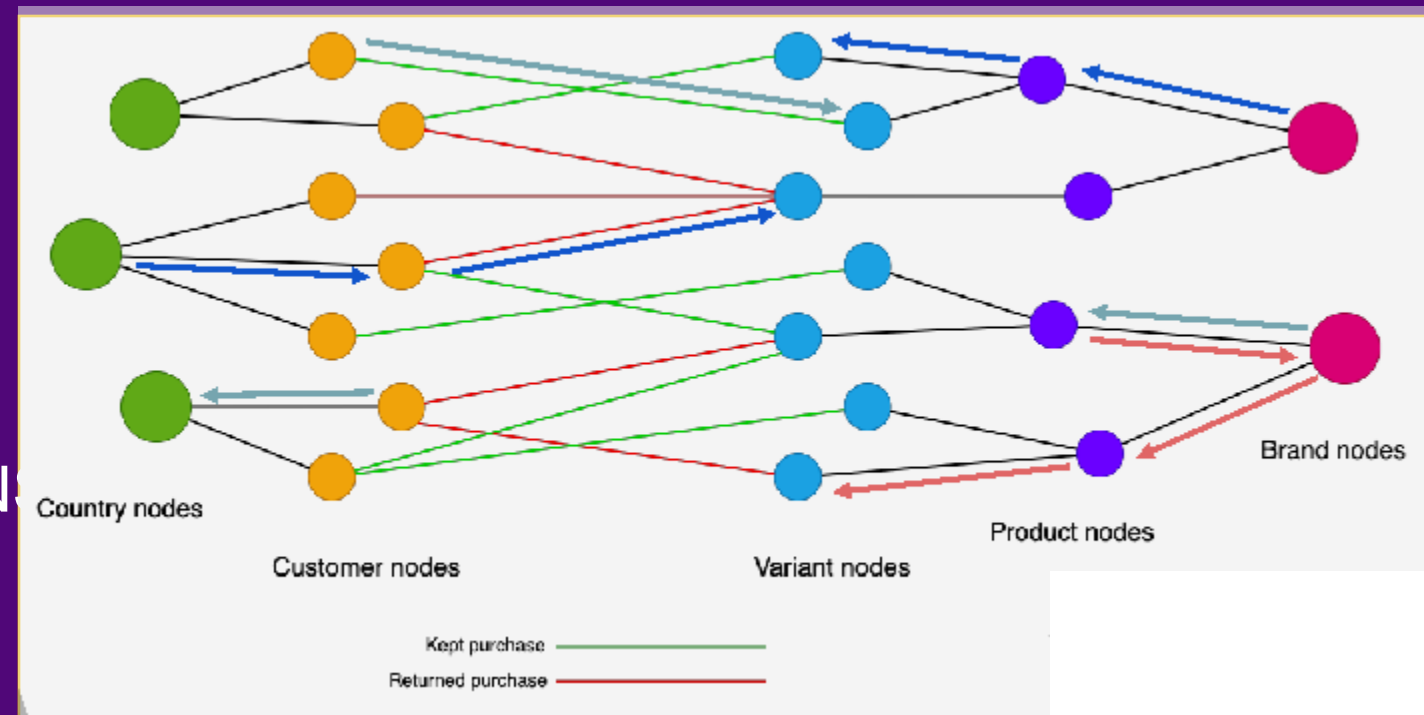
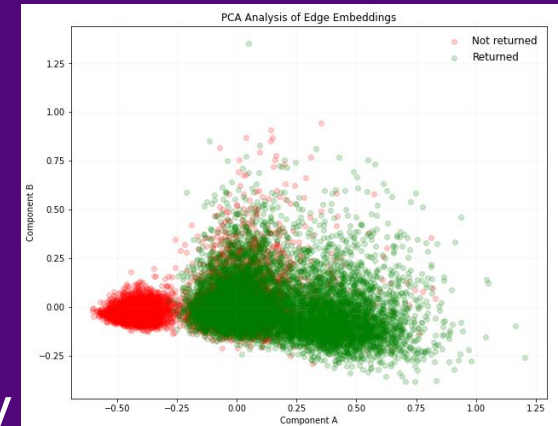
- British online fashion and cosmetic retailer, sells over 850 brands, ships to all 196 countries from fulfilment centres in the UK, USA and Europe.
 - In 2020, ASOS revenue was £3.26 billion
- Free returns
 - Reducing the number of returns saves on carbon and financial budget
- Goal: Identify possible return, suggest alternative
- Explore usage of graph neural networks
 - Address “cold-start problem”
 - Upscale team skills
- [Conference presentation](#)



We predict a **70% probability** you will return this item. Would you like to **see alternatives?**

Predicting Returns via GNN

- Designed graph structure balancing connectivity and generality
 - Homophily: recognise similar types of nodes in a network structure
 - structural equivalence: same/similar neighbours
- Designing a graph that showed good behaviour was a challenge
 - GNN performance improved on baseline models
 - More uniform vs country
- ASOS's first experience with GNN
 - **Upscale skills!**



Predicting Returns: MSc extension

- January – April
 - 3 1st year PhD students
 - 1 4th year physics theory student
- May – August
 - 3 MSc students worked on improvements
 - 4th year student continued as TA
- Results:
 - Useful results, better trained students
 - One paper/conference
 - Two job offers (one hire)



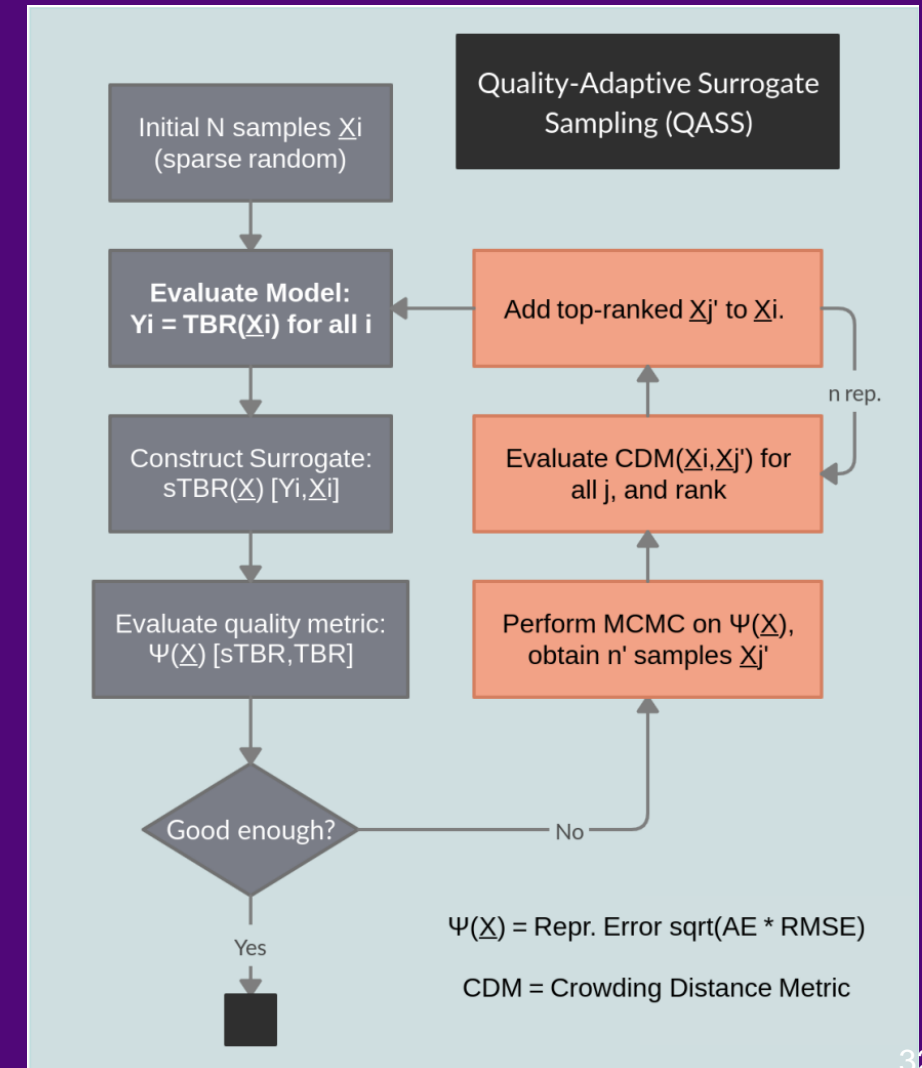
Established GNN method and baselines to compare to

Explore heterogeneous graphs, skip connections, mini-batch training, etc to obtain optimal model

Best GNN further developed within ASOS by new hire

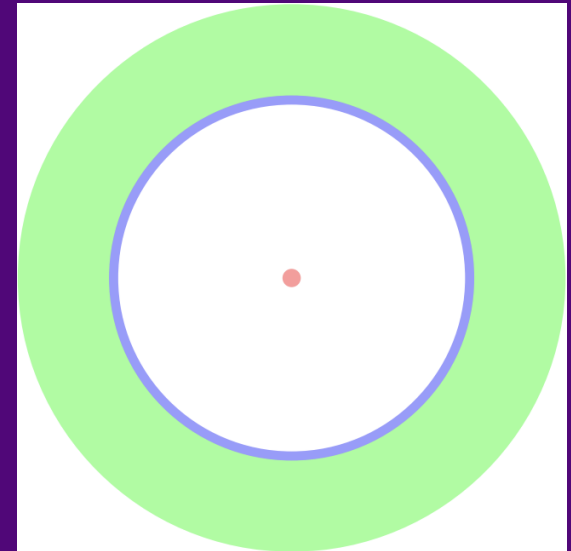
Fast Regression of Tritium Breeding Ratio

- Quality-Adaptive Surrogate Sampling algorithm
 - iteratively increment the training/test set with sample points which *maximize* surrogate error and minimize a crowding distance metric (CDM) in feature space
 - Markov Chain Monte Carlo performed to sample the error function generated by performing nearest-neighbor interpolation



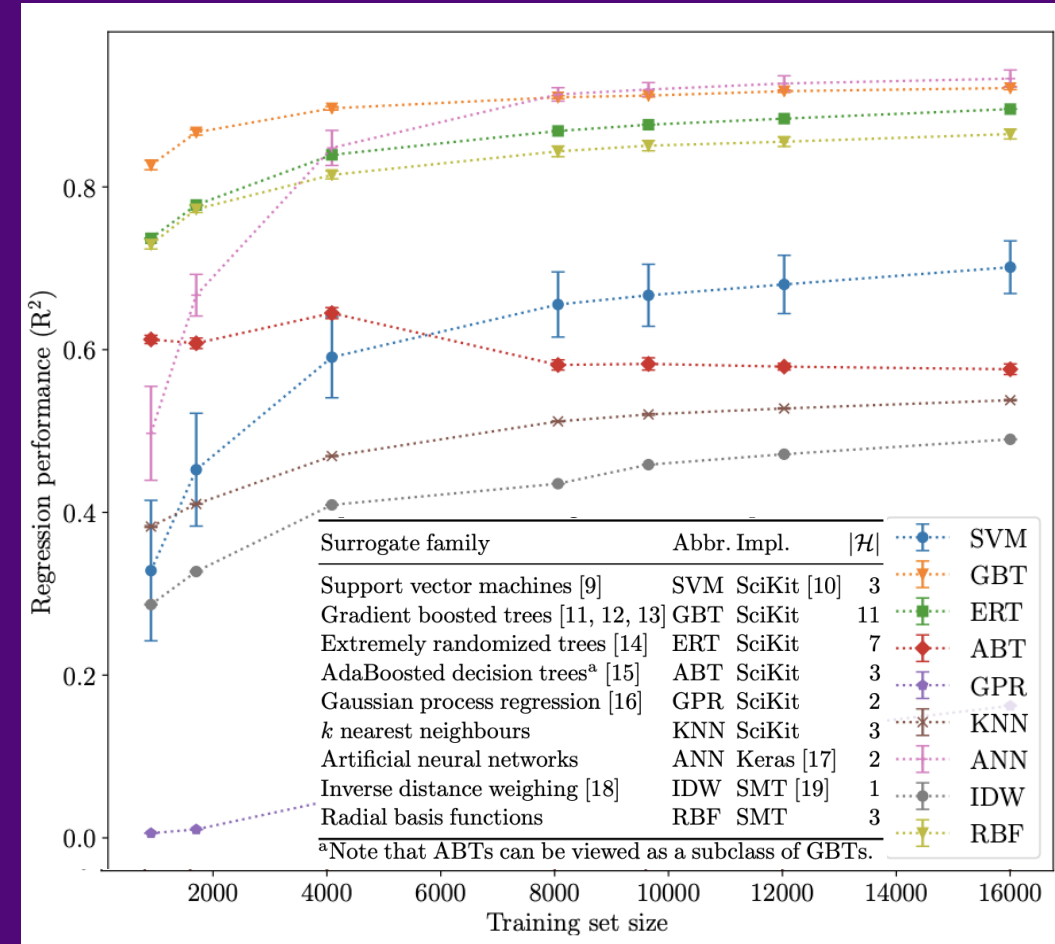
Fast Regression of Tritium Breeding Ratio

- Inertial confinement fusion (ICF) reactor with Deuterium-Tritium (D-T) plasma **neutron** source. (JET, ITER)
 - Tritium breeding **blankets**
 - **TBR = the ratio of tritium produced per source neutron**
- Monte Carlo neutronics simulation Paramak
 - predicts TBR for a spherical ICF reactor configuration
 - 12/6 parameters for **blanket/first wall**
 - Expensive to run simulation
- Surrogate models resolve computational limitations by replacing a resource-expensive procedure with a cheaper approximation



Fast Regression of Tritium Breeding Ratio

- Surveyed 9 surrogate family models
 - No variable reduction
 - 5-fold cross-validation
- Tree-based out perform NN in small dataset
 - NNs lead when dataset larger
- ANNs: 500k events
 - Best: $R^2 = 0.998$, $t=1.124 \mu\text{s}$ (7×10^6)
 - Fastest: $R^2 = 0.985$, $t=0.898 \mu\text{s}$ (9×10^6)
- GBT: 10k events
 - Fastest: $R^2 = 0.913$, $t=6.125 \mu\text{s}$ (1×10^6)
- Dev'ed adaptive algorithm: -40% surrogate error



Joint Biosecurity Centre



Placement in late 2020 and became a leader within the team.



“He is an extremely strong scientist, and we are lucky to have him with us working on the COVID response.”



prepared analysis for briefings for number 10

Generated analysis for the readiness of mass testing in Liverpool

Built dashboard for epidemiology data communication

Worked with #10 to incorporate dashboard

Prepared analysis for situational awareness reports across nation

Co-author Scientific Advisory Group for Emergencies (SAGE) paper

Joint Biosecurity Centre

- [Defra/JBC: Wastewater COVID-19 monitoring in the UK](#)
19 November 2020
- Wastewater based epidemiology in the UK has a long and notable history, dating back to the identification, by John Snow, of wastewater as the source of cholera outbreaks in the mid-19th Century.
- Focused on SARS-CoV-2 viral RNA in wastewater
- Established scientific and epidemiological insights depend on:
 - acuity and extent of the sampling,
 - laboratory analyses,
 - level of environmental determinism and stochasticity driving variability.
- Work anticipates the use of wastewater to track disease at scale as a proactive measure to ensure public health resilience.

Autonomy: Who are the renters?

Placement with think tank Autonomy in early 2021 ([blog](#), [news](#), [news](#))

- Findings showed insecure work is trapping low-paid workers in the London's "overpriced" rental sector, with many of them having to pay around two thirds of their wages to get a roof over their head.

Most manual workers ("working class") often own their home

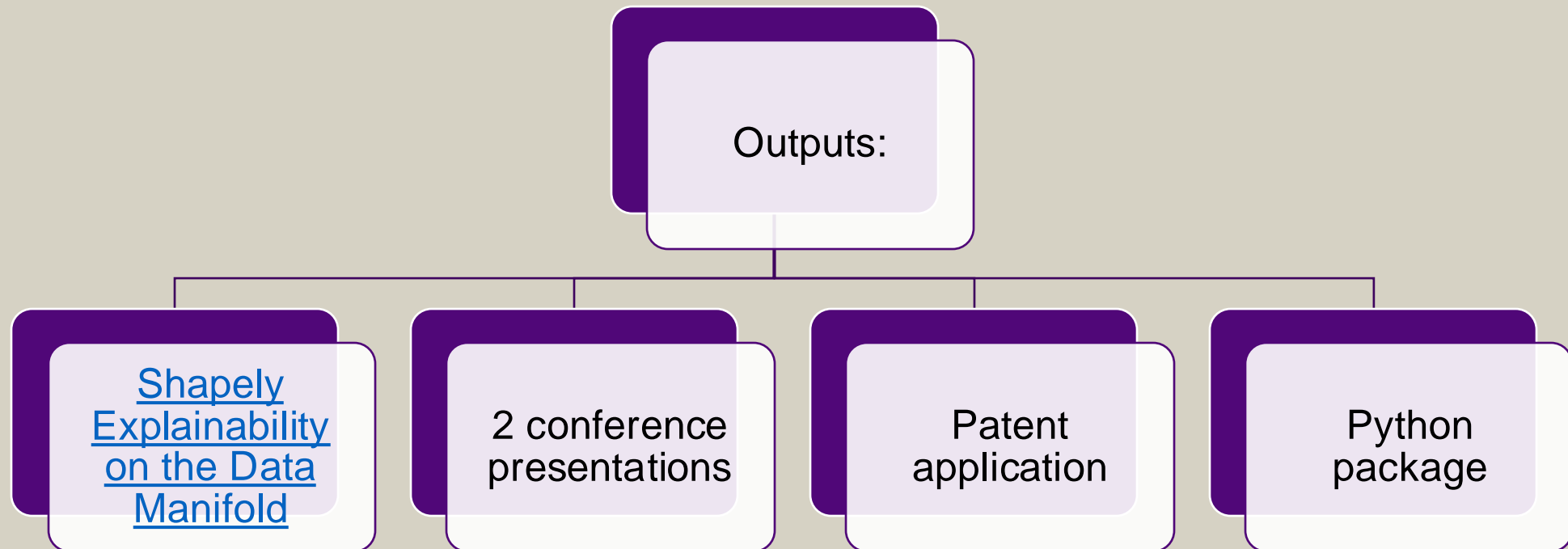
- i.e. 'Carpenters and joiners', 'Plasterers', 'Electricians and electrical fitters', 'Floorers and wall tilers' as well as 'Roofers, rooftilers and slaters'

The occupations in London with the highest prevalence of renters are:

- Cleaners and domestics (81% are women)
- Waiters and waitresses (73% are women)
- Kitchen and catering assistants (64% are women)
- Care workers and home carers (84% are women)
- Elementary construction occupations (98% are men)

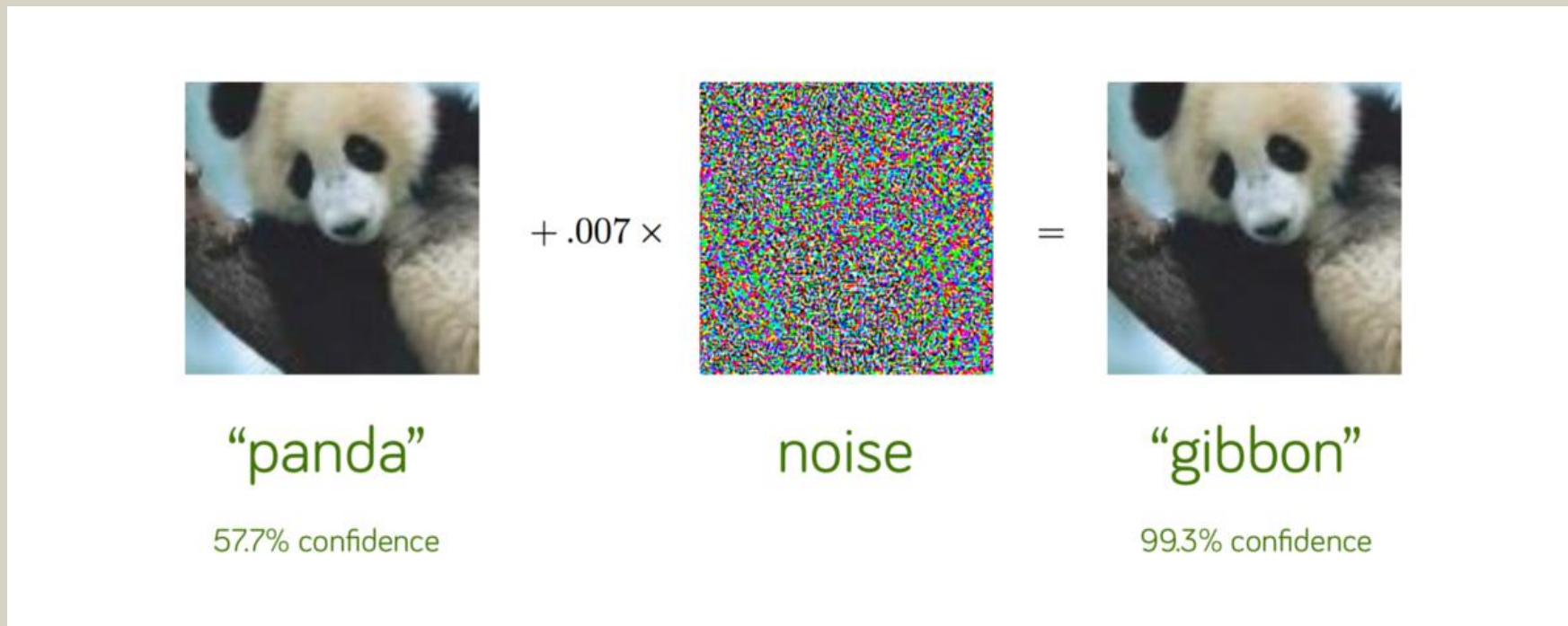
Faculty AI: ML Explainability

- Faculty: software, consulting, and services related to artificial intelligence.



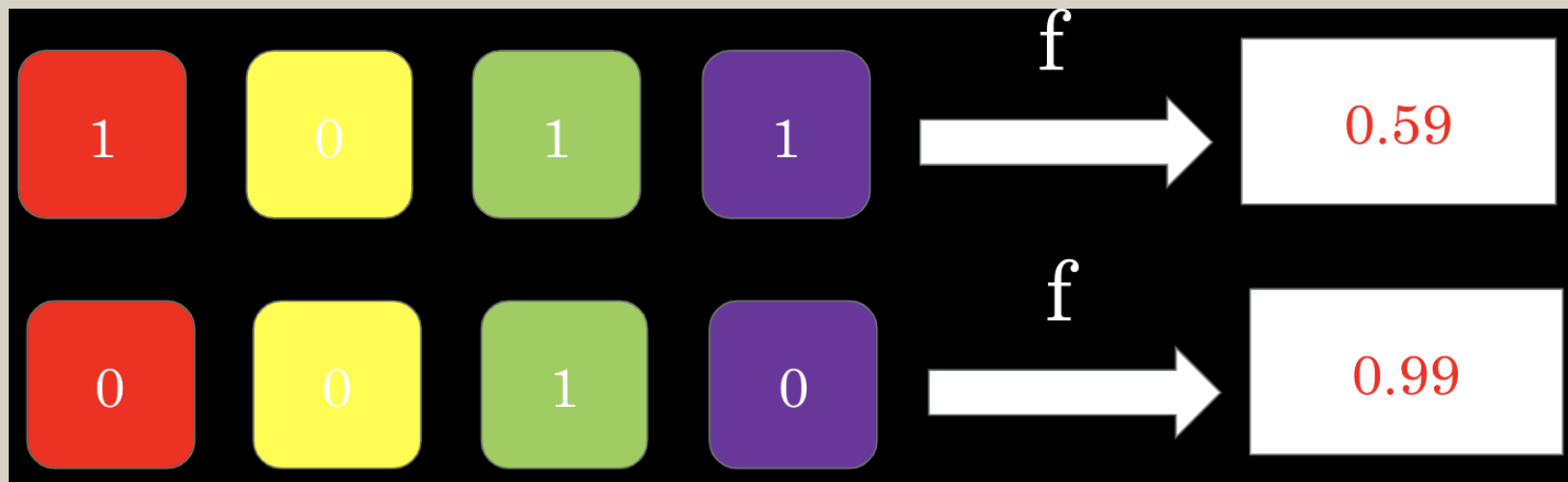
Faculty AI: ML Explainability

- ML models are performant but can be brittle, black-boxes
 - Especially dangerous when ML used in society




ML Explainability

- Shapley values - mathematically principled approach to model explainability
 - Based on game theory: distribute credit for total value earned by team
 - arXiv:2006.0127



ML Explainability

- Shapley values - mathematically principled approach to model explainability
 - Based on game theory: distribute credit for total value earned by team
 - arXiv:2006.0127

Importance of  ??

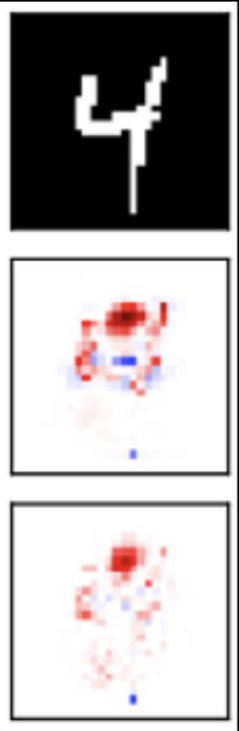
Discussion centers on how to calculate f with < 4 inputs.

$$f(\text{red } 1, \text{yellow } 0, \text{green } 1, \text{purple } 1) - f(\text{red } 1, \text{yellow } 0, \text{green } 1)$$

$$+ f(\text{red } 1, \text{yellow } 0, \text{purple } 1) - f(\text{red } 1, \text{yellow } 0)$$

...

Our method



Existing method

Precision, Recall & F1-score

Precision

- Out of all the **predicted positive labels**, how many are **correct**?

Recall

- Out of all of the **actual positive labels**, how many did we **correctly predict** as positive?

F1-score

- Weighted **average** of precision and recall.

		Predicted	
		Positive	Negative
Labels	Positive	<i>TP</i>	<i>FN</i>
	Negative	<i>FP</i>	<i>TN</i>

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$