

Making an impact: Realising the potential of urban data science

Tom Smith, @_datasmith
Director, ONS Data Science Campus



**Data Science
Campus**

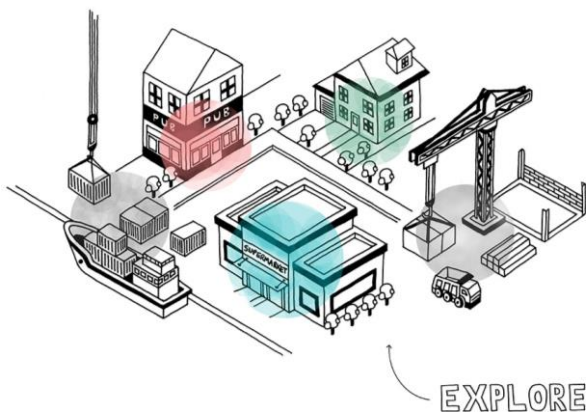
web: datasciencecampus.ons.gov.uk
email: datasciencecampus@ons.gov.uk
twitter: [@DataSciCampus](https://twitter.com/DataSciCampus)





How Data Science helped identify potential savings of over £581m for the NHS

Abi Giles-Haigh, 31 January 2018 - [Digital data and technology](#), [People and Skills](#)



Economy

GDP
Inflation
Labour market
+++



People

Population
Census
Incomes
+++



World

Trade
Sustainable
Development Goals
+++

Data Science Campus creation



“Although **better use of [data]** has the potential to transform the provision of economic statistics, ONS will need to **build up its capability** to handle such data.

This will take some time and will require not only **recruitment of a cadre of data scientists** but also **active learning and experimentation**.

That can be facilitated through **collaboration with relevant partners** – in academia, the private and public sectors, and internationally.”

*Independent Review Economic Statistics
Professor Sir Charles Bean, 2016, p.11*

The screenshot shows a Financial Times article. At the top is the 'FINANCIAL TIMES' logo and a navigation bar with links: HOME, WORLD, US, COMPANIES, MARKETS, OPINION, WORK & CAREERS, LIFE & ARTS. The article title is 'ONS 'unicorn' campus reimagines how to measure Britain'. Below the title is a sub-headline: 'Statisticians experiment with using Google Street View, shipping data and VAT returns'. The main image shows a man sitting in a red ergonomic chair at a desk with a laptop, looking out a large window at a green landscape. Below the image is a caption: 'The Data Science Campus in Newport © Gareth Iwan Jones/FT'. There are social media sharing icons for Twitter, Facebook, and LinkedIn, along with a 'Save to myFT' button. The article text begins with 'AUGUST 3, 2017 by Chris Giles in Newport, Wales' and continues with 'The inflatable rainbow unicorns near the entrance of its new £17m Data Science Campus are a jokey nod to the ambitions of Britain's statistics office.' and 'Here in Newport, South Wales, in a wing designed to look like the office of a Silicon Valley company, the Office for National Statistics is trying to imagine the future of measuring Britain.'



Purpose

We apply data science, and build skills, for public good across the UK and internationally

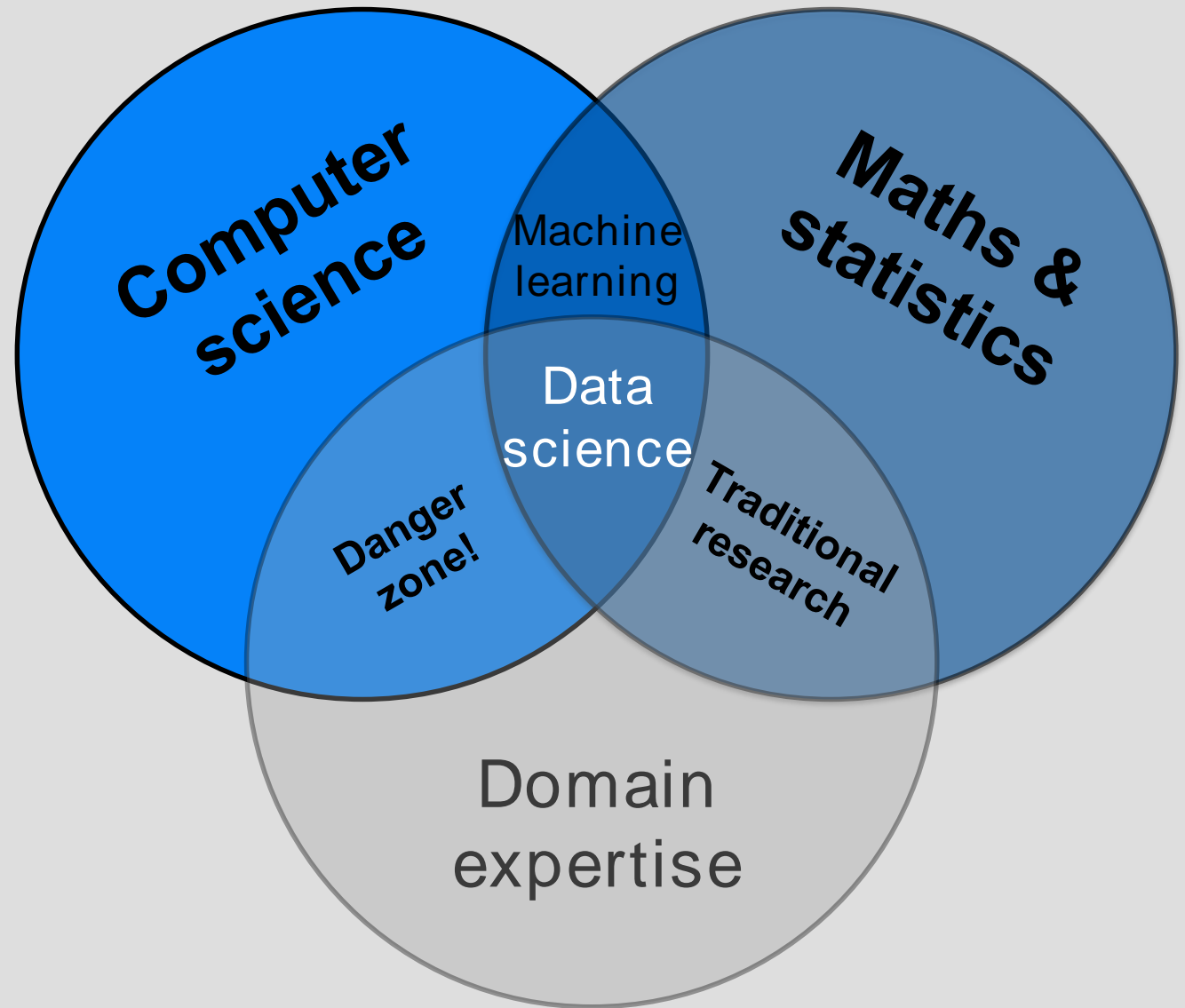
Mission

We work at the frontier of data science and AI - building skills and applying tools, methods and practices - to create new understanding which improves decision-making for public good

What is Data Science?

“Data scientists solve complex **business problems using a combination of domain expertise, coding knowledge, machine learning and statistics skills** on large and varied datasets.”

Government Data Science Partnership



(One of many) descriptions of data science, Drew Conway



1939 - London Transport workers manually examine over 4 million tickets to identify most and least popular routes
Gerry Cranham/Fox Photos/Hulton Archive/Getty Images

Transport for London

WiFi data collection

We are collecting WiFi data at this station to test how it can be used to improve our services, provide better travel information and help prioritise investment.

We will not identify individuals or monitor browsing activity.

We will collect data between Monday 21 November and Monday 19 December.

For more information visit: tfl.gov.uk/privacy

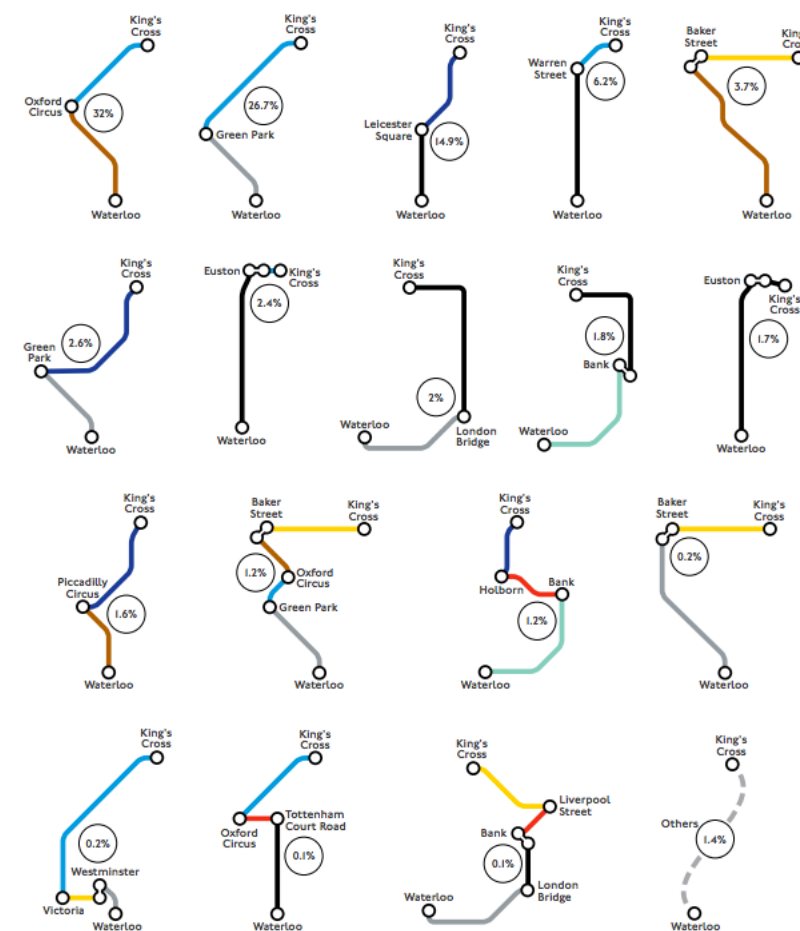


MAYOR OF LONDON



TRANSPORT
FOR LONDON
EVERY JOURNEY MATTERS

Figure 14: Route options between King's Cross St. Pancras and Waterloo, and the proportion of devices on each one



Transport for London 2016 pilot, assessing journeys by WiFi access



Why do we need Data Science?

- “Getting data right is the next phase of public service reform”
- Deliver more insight from the data we hold
- Drive more insight through use of new data sources

John Manzoni – Chief Executive
of the UK Civil Service



John Manzoni, Chief Executive of the UK Civil Service, Sprint 18, London, 2018

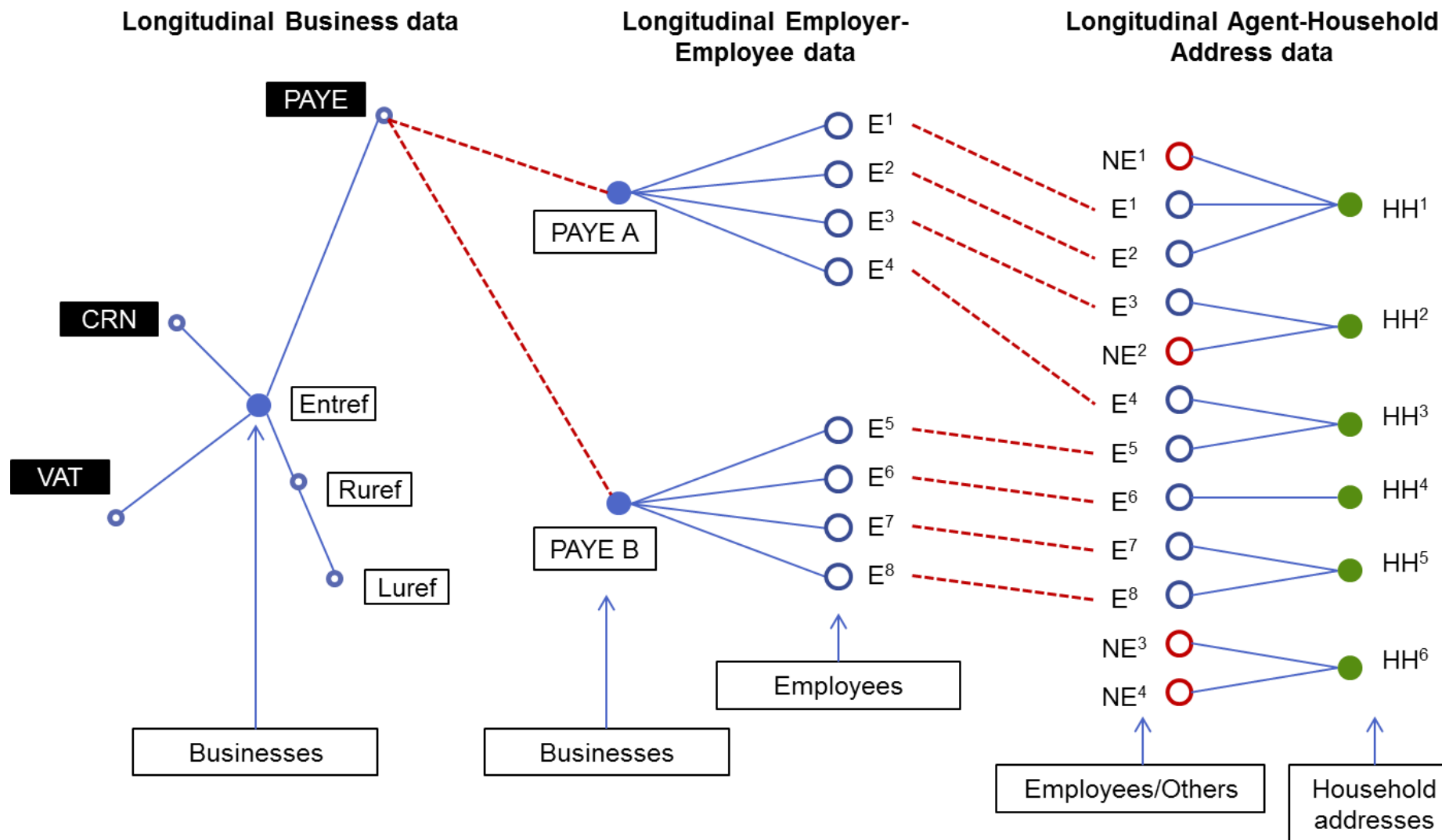


How Data Science helped identify potential savings of over £581m for the NHS

Abi Giles-Haigh, 31 January 2018 - [Digital data and technology](#), [People and Skills](#)



Linked administrative data sources (UK)





Linked administrative data is first prize

Using business tax data in GDP

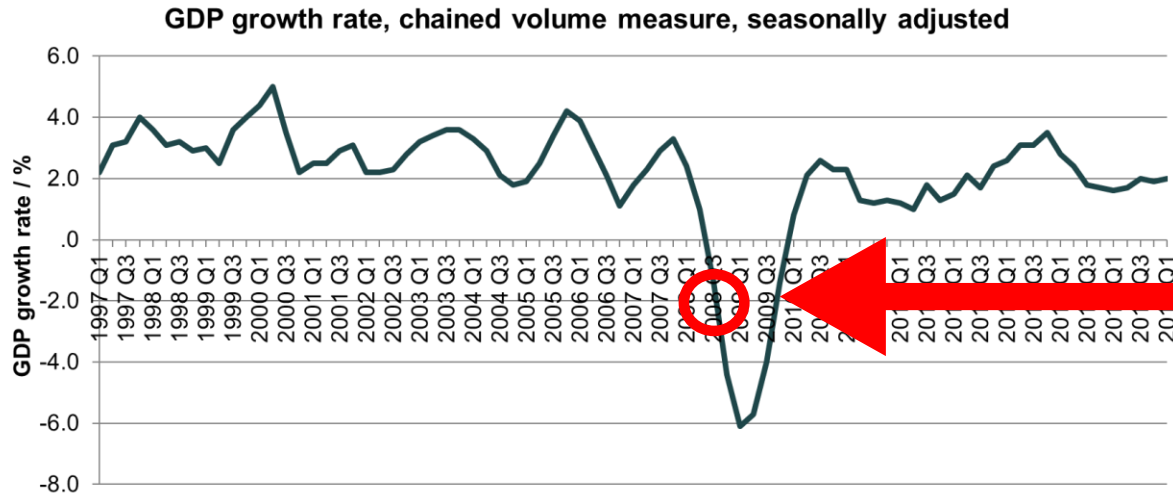


Fig 1. UK GDP Growth Rate

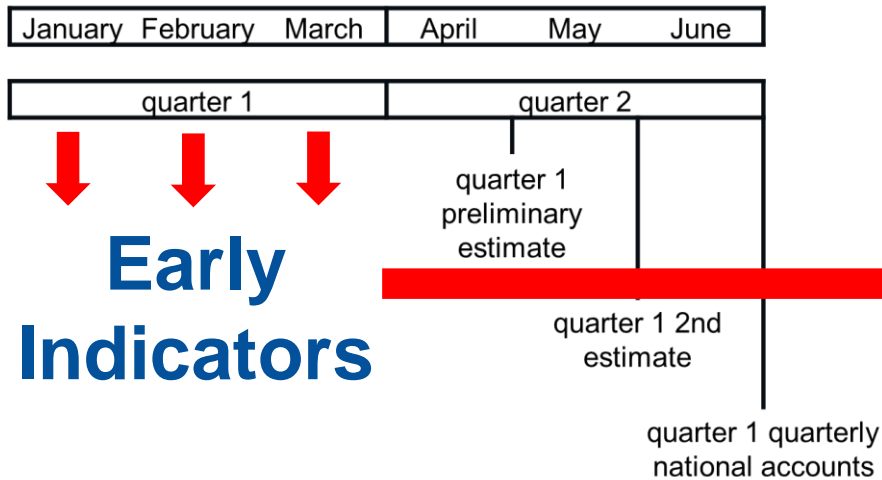


Fig 2. ONS National Accounts Publication Timetable

-6%

Change in UK GDP between first quarter of 2008 and second quarter of 2009

5 years

Length of time from 2008 for the UK economy to return to pre-recession size

£12b

Estimated value for earlier identification of 2008 downturn



There's a lot of new data sources ...

In a recent study produced for the Office for National Statistics (ONS) Natural Capital Accounts, the UK's trees were estimated to **remove 1.4 million tonnes** of air pollutants in a single year. This would result in an **annual saving of £1 billion** in avoided health damage costs. In another study, London's 8.42 million trees have been estimated to remove 2,241 tonnes of pollution per year, which in addition to other services, is estimated to provide £132.7 million in annual benefits.

For Cardiff, the annual benefit is close to **£8 million**.



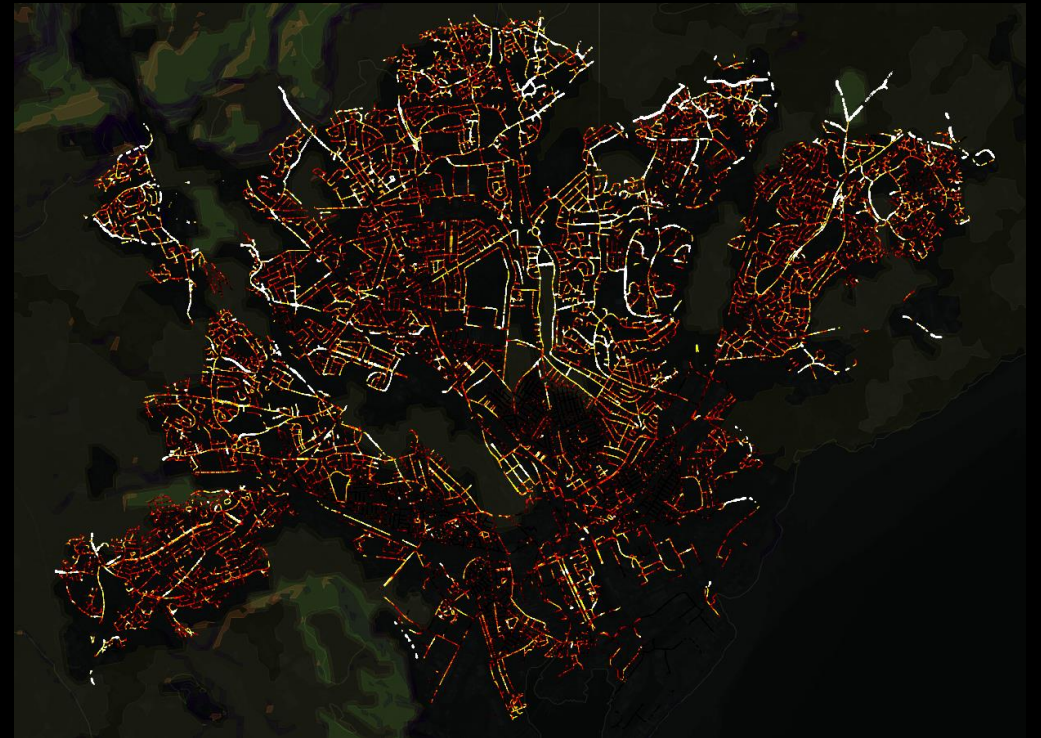


Aim: Generate a scalable,
consistent, automated,
urban vegetation index

Outcome: An end-to-end processing pipeline.

Making use of: **17 million images** from **Google Street View** for 112 cities in the UK.

... **OpenStreetMap** road network data
... Deep **image segmentation** methods



Current approach...

... Pyramid Scene Parsing Network

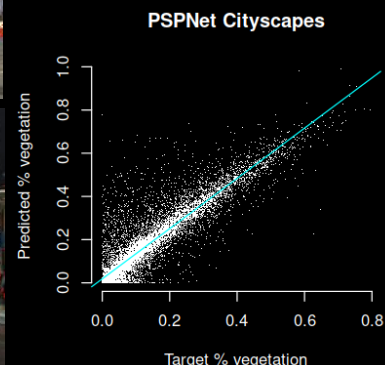
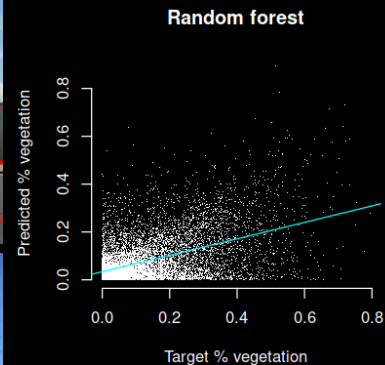
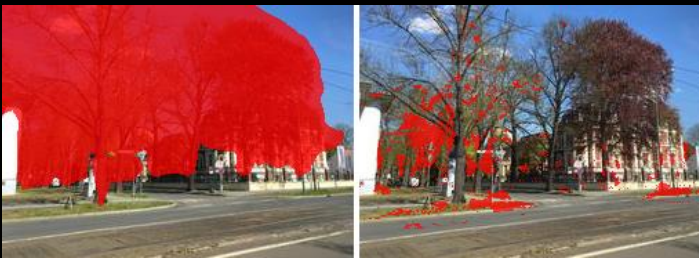
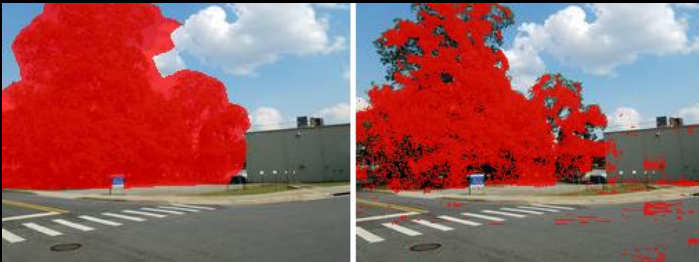
Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia.
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

Model	BACC	Pre	Rec	F1	MCC	R^2	τ
PSPNet (city)	0.90	0.66	0.87	0.75	0.72	0.83	0.77
PSPNet (ade20k)	0.85	0.82	0.73	0.77	0.74	0.83	0.76
Random forest	0.62	0.48	0.29	0.36	0.31	0.25	0.32
Lab (a* b*)	0.62	0.47	0.28	0.35	0.29	0.20	0.28
Lab (a*)	0.55	0.33	0.14	0.19	0.15	0.04	0.15

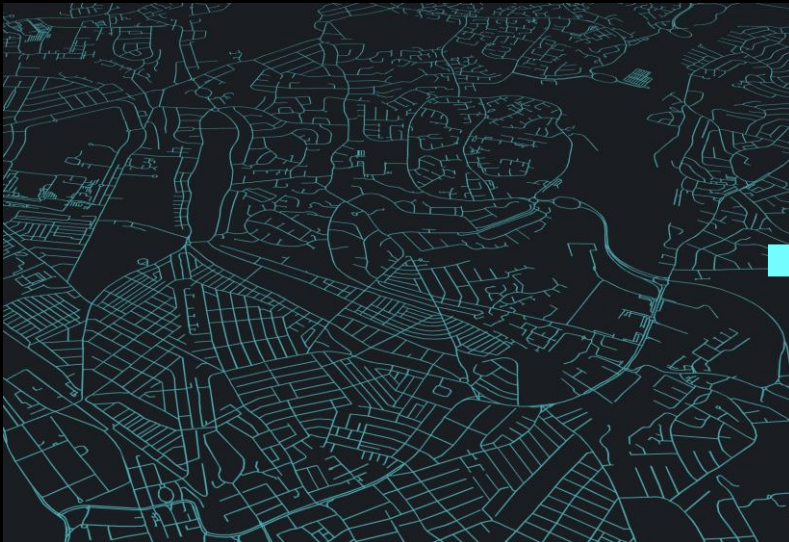
Images segmented by **cars**, buildings, **path**, **people**, **trees**.

90% vs 62% class balanced accuracy.

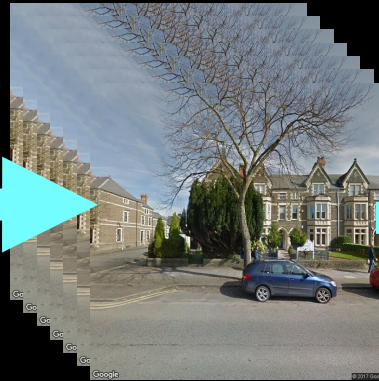
Validated using the Mapillary Vistas Dataset for semantic understanding of street scenes. <https://research.mapillary.com/>



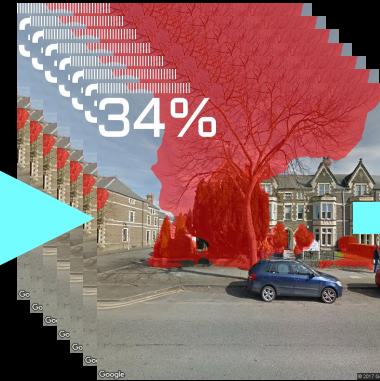
StreetView image processing pipeline



OpenStreetMap
road network data



17 million
StreetView
images



Percentage
trees for
each image



Urban vegetation
map

Enter your postcode or click on the map to explore

cf23 5ee

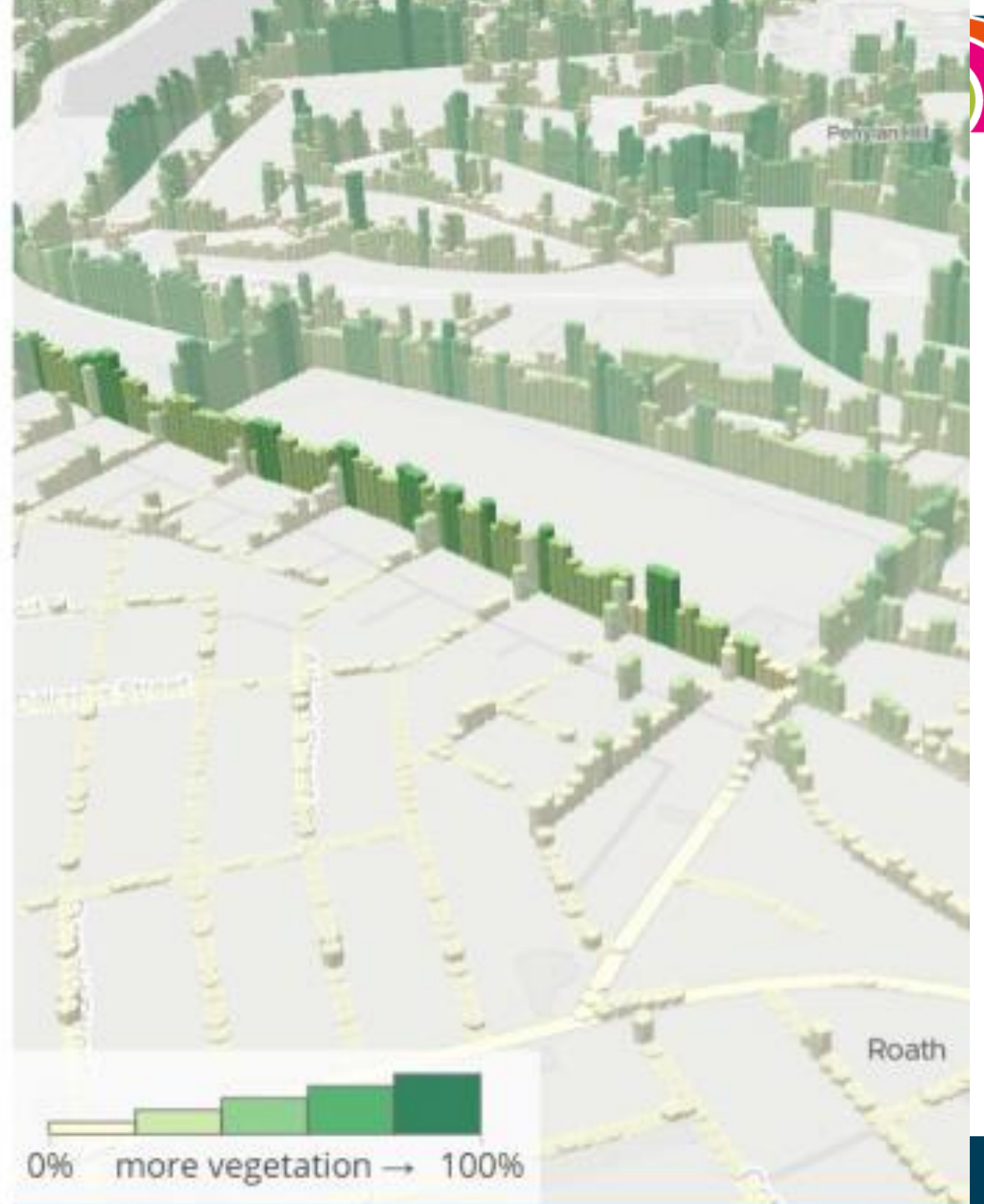


— Cardiff average **13%**

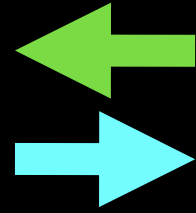


Ninian Road is the **182nd** greenest street out of **3,219** in **Cardiff**

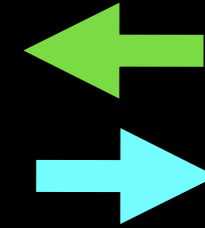
Share



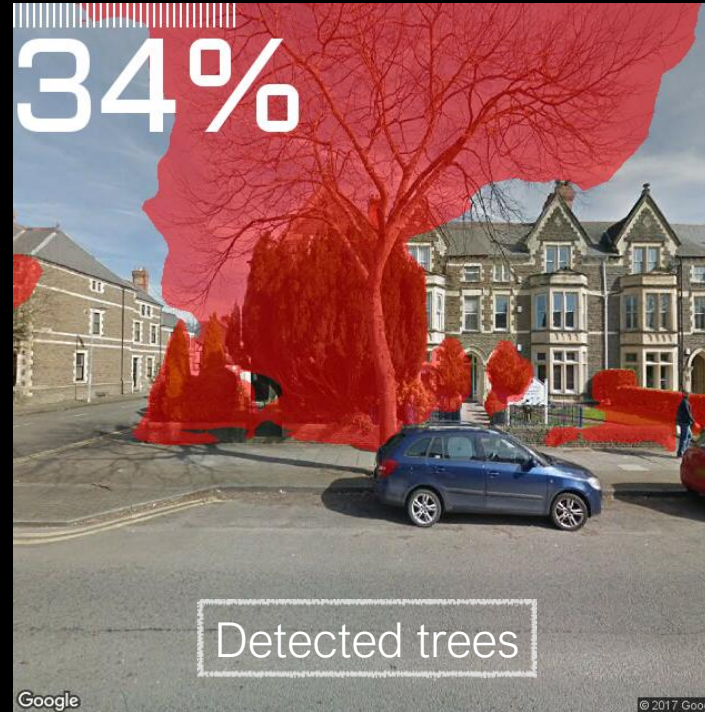
StreetView
processing
pipeline



UNGP
vegetation
service



UNGP
segmentation
service



1. Image processing pipeline pushes image to vegetation service
2. Vegetation service pushes image to Segmentation service
3. Vegetation service returns percentage trees in segmented image.

Early Indicators of GDP

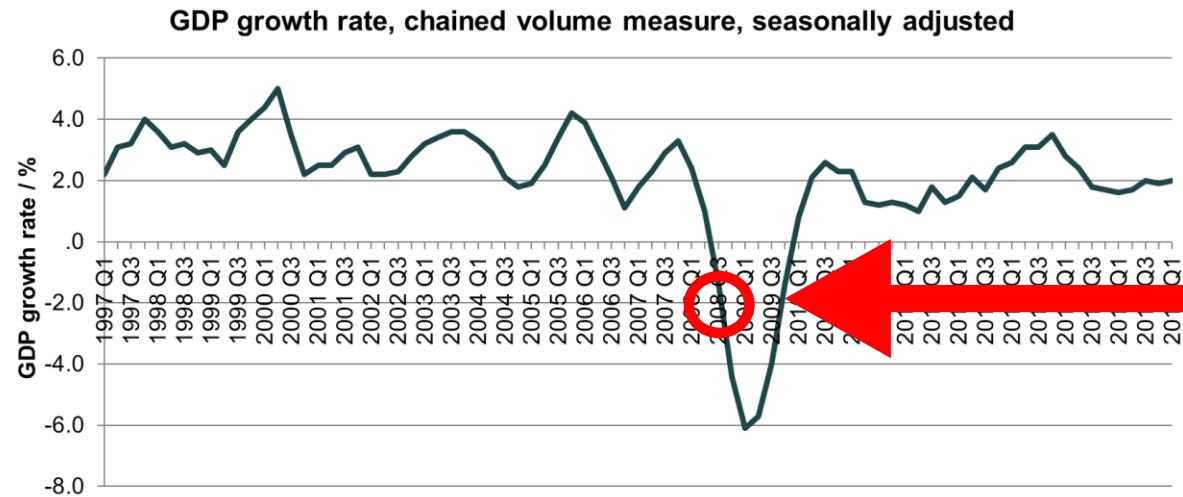


Fig 1. UK GDP Growth Rate

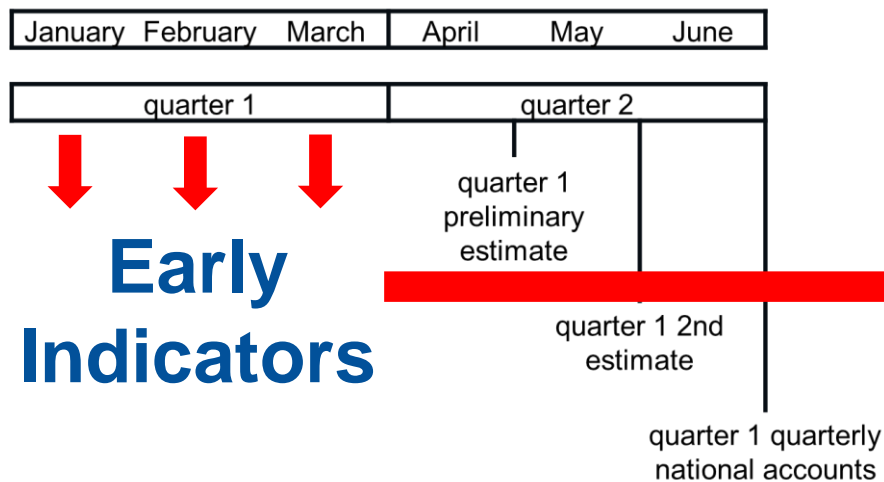


Fig 2. ONS National Accounts Publication Timetable

-6%

Change in UK GDP between first quarter of 2008 and second quarter of 2009

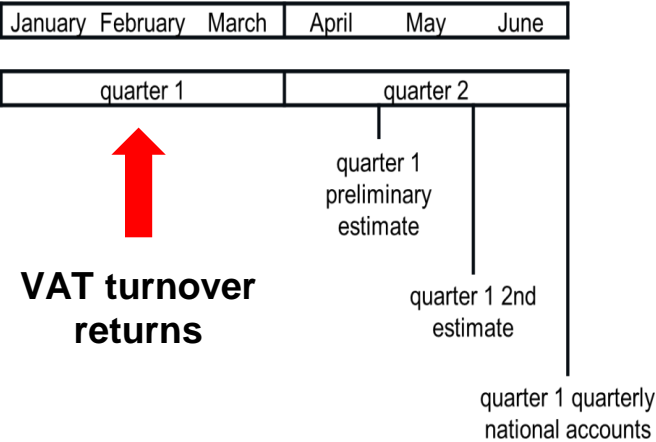
5 years

Length of time from 2008 for the UK economy to return to pre-recession size

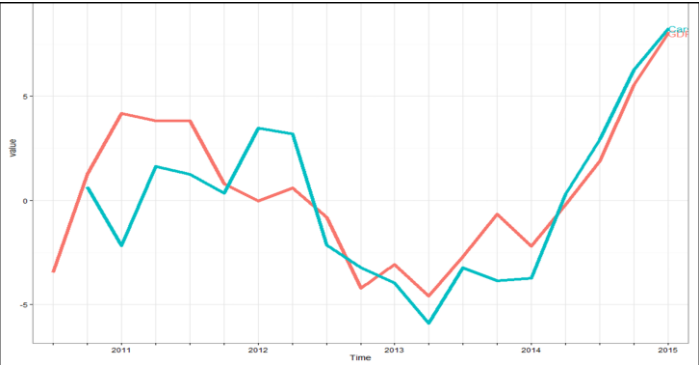
£12b

Estimated value for earlier identification of 2008 downturn

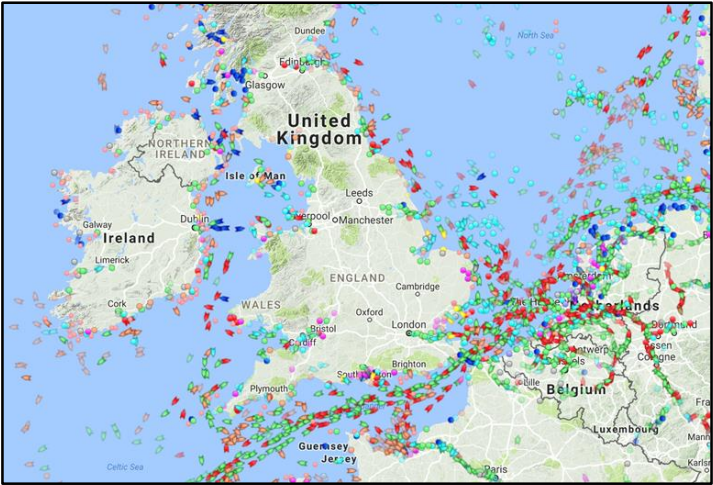
Early Indicators of GDP



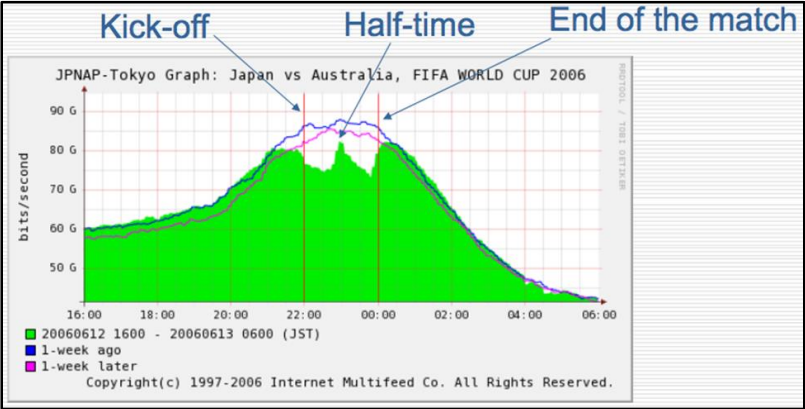
HMRC VAT Data



Road Traffic



AIS Ship Location



Broadband Traffic

-6%

Change in UK GDP between first quarter of 2008 and second quarter of 2009

5 years

Length of time from 2008 for the UK economy to return to pre-recession size

£12b

Estimated value for earlier identification of 2008 downturn

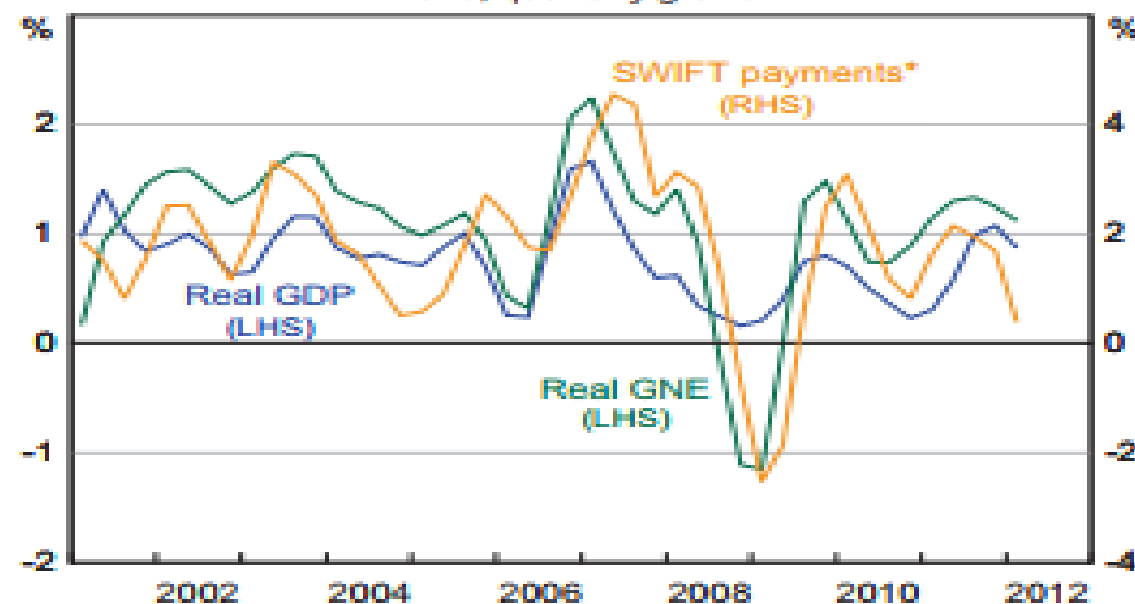
Payments data for regional indicators



- Identifying rapid, local economic indicators - breakdowns by geography, industry, product, credit / debit card, on-line payment, international
- Collaboration with Barclays, 2-way secondments
- What can we learn about payments data?



SWIFT Payments and Economic Activity*
Trend, quarterly growth



* Number of SWIFT interbank payments settled in RITS, 7-period
Henderson trend
Sources: ABS; RBA.

Payments data for regional indicators

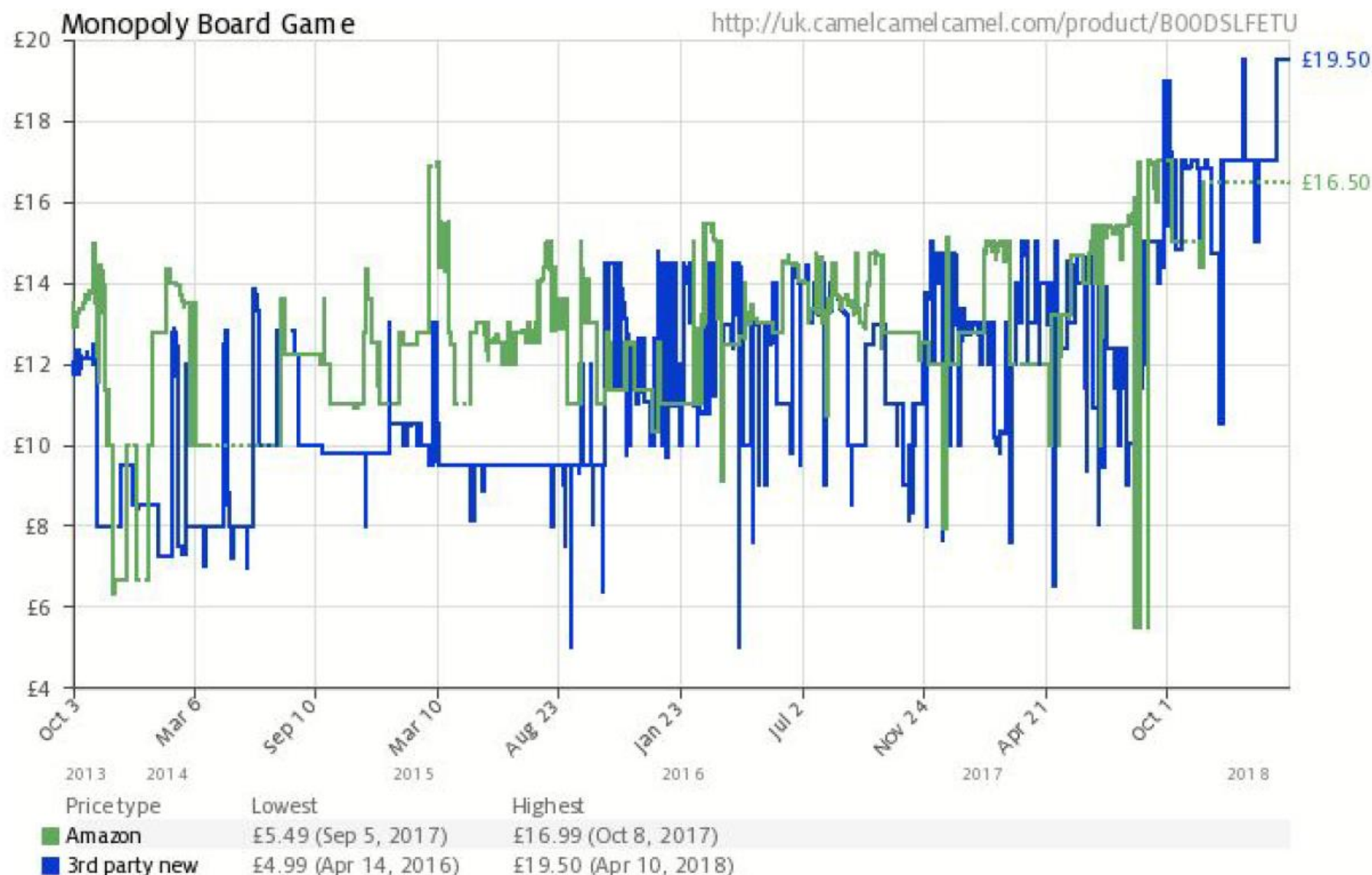


- Financial data held by banks
 - No sensitive or personally identifiable data shared
 - All outputs are aggregate and non-sensitive
- Hypotheses we are exploring include
 - Payments data as proxy for retail sales by sector & time (eg night time economy)
 - Payments data as proxy for private household consumption
 - Payments data can improve the accuracy of GDP nowcasting

- Data sources potentially available through secondments:

Consumer	Electronic payments	Business
Debit Card spend Credit card spend Personal Loans Mortgages Savings accounts Insurance	POS data ATM data Online gateway data (online purchases) Peer-to-peer	Merchant & Acquirer data Corporate Cards Business Bank products Corporate bank Products Investment bank products

Prices and volatility



Monopoly price fluctuation over 4 year period
High = £19.50
Low = £4.99
(Data from camelcamel)

Big Data is changing how consumer markets work

James Plunkett, 2017-18
Rybczynski Prize Essay

Predicting Viral Outbreaks



Fig 1. Wordcloud of Norovirus Keywords

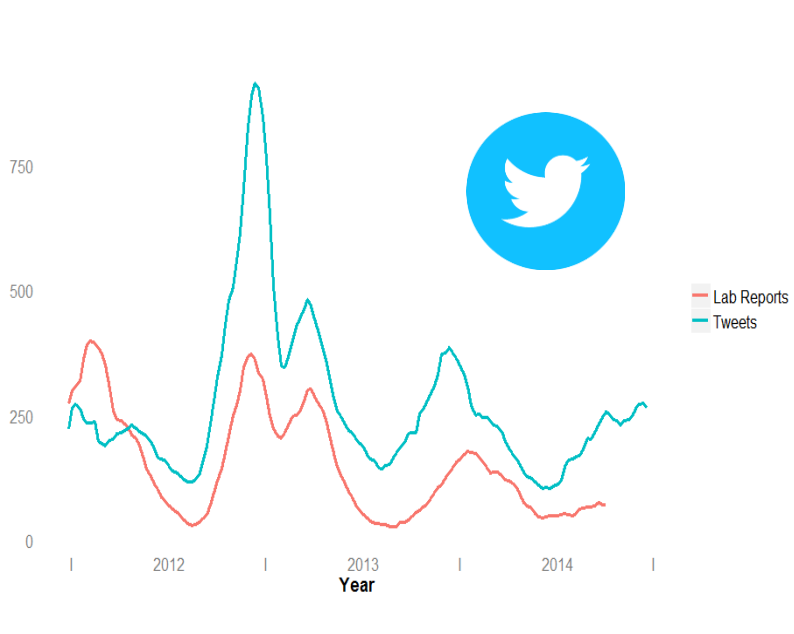


Fig 2. Norovirus Keywords in Tweets compared to reported incidents

2.8m

Cases of Norovirus per year in the UK

£120m


Estimated cost to the country in lost working hours due to Norovirus

£20k


Total cost of the project, including publicity campaign

Diarrhoea and vomiting?


There's no specific cure for stomach bugs such as Norovirus
Going to your GP puts others at risk of infection. Treat symptoms at home




stay hydrated



take paracetamol



prevent spread



stay at home for two days after symptoms clear

#EssentialKit

NLP Analysis of Ferry Cargo



The Challenge

Ferry operators collect information on the contents of lorries and trade vehicles boarding their Ferries

A single line description is recorded to detail the contents

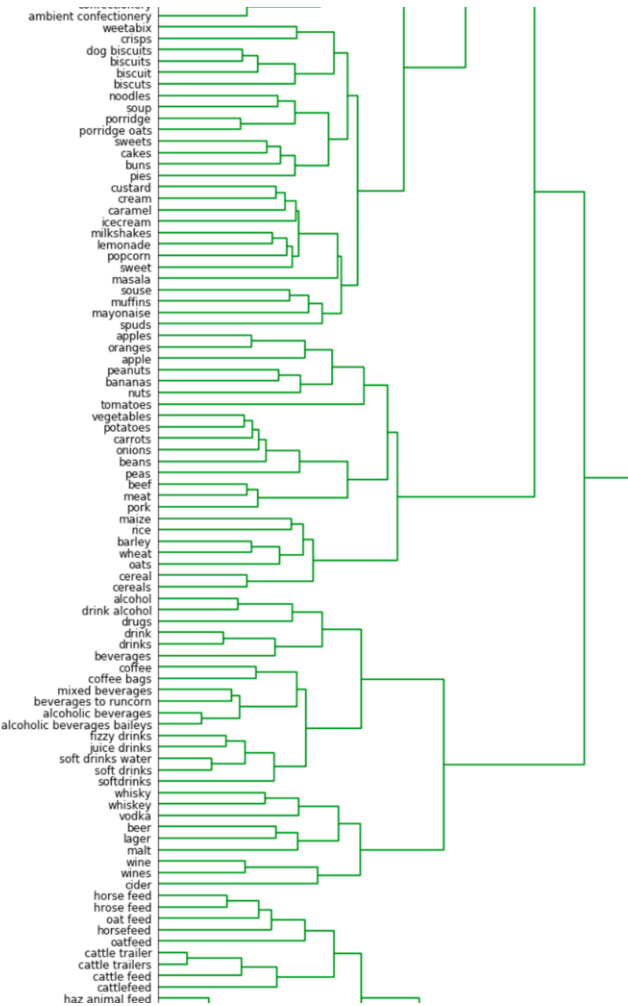
The data collection is not controlled enabling complete free text entries.

This significantly restricts the analysis that can be done.

The Solution

Optimus is a pipeline that can group items from free-text lists by context that do not have accompanying classifications or codes.

The tool can generate labels for groups of items based on common syntax or, in some cases, synonyms. It can also handle inconsistencies in text records such as spelling mistakes, plurality and other syntactic variation.



The Data

35k

Lorry journeys in single month analysed during Phase 1

450k

Lorry journeys in 2017 to be analysed during Phase 2

Identifying emerging technology trends through patent applications



90 million global patent applications assessed using text analytics, for Industrial Strategy Grand Challenges

Project aim

Can we use patents to explore popular and emerging terminology and technology?

Issue

Patent databases store information in various formats, we need to combine these.

Solution

Develop an app that can synthesise data types and produce an app that can extract key terms from patents abstracts.

	Term	TF-IDF Score
1.	electric power	0.275362
2.	power supply	0.274285
3.	fuel cell	0.269450
4.	storage device	0.255554
5.	electric vehicle	0.224883
6.	energy storage	0.223985
7.	exhaust gas	0.185377
8.	control unit	0.179616
9.	combustion engine	0.169071
10.	internal combustion	0.164826
11.	control device	0.158519

Testing

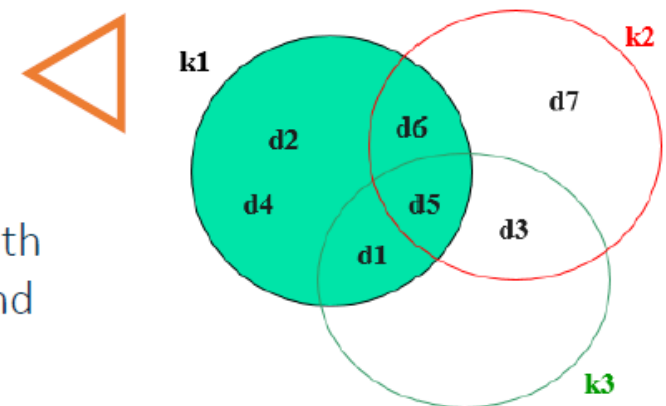
Tested on 100 random US patents, algorithm accuracy scores compared to human scores

The App

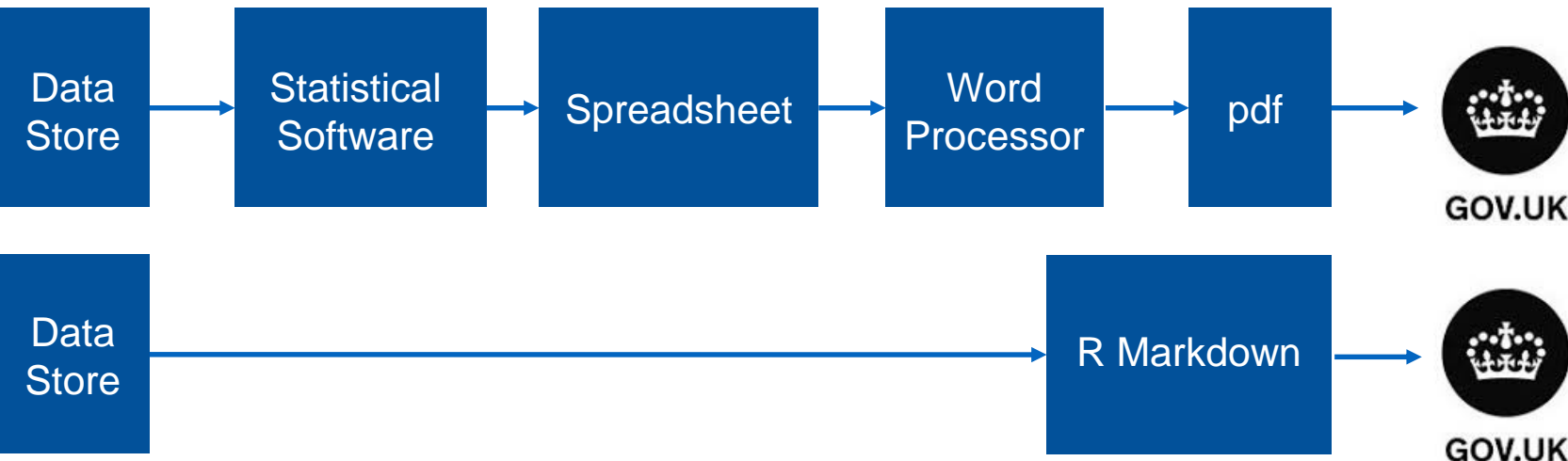
Python-based code with time-series analysis and citation weighting options to extract key terms

Method

Use NPL and TF-IDF to score terms



Reproducible Analytical Pipeline



Efficiency Savings

£118m

Estimated annual efficiency savings across government stats publications

The Challenge

- Producing official statistics for publications is a key problem: as it is a time consuming meticulous process
- It is time consuming as the analysis has to pass throw multiple systems and multiple individuals
- The systems are diverse and do not always conform to good software engineering practice

Solution

- Use of software engineering tools and techniques such as version control.
- Automated generation of tables/charts and statistical verification
- Process from data storage to report generation

£8.8k

Estimated average annual saving per publication

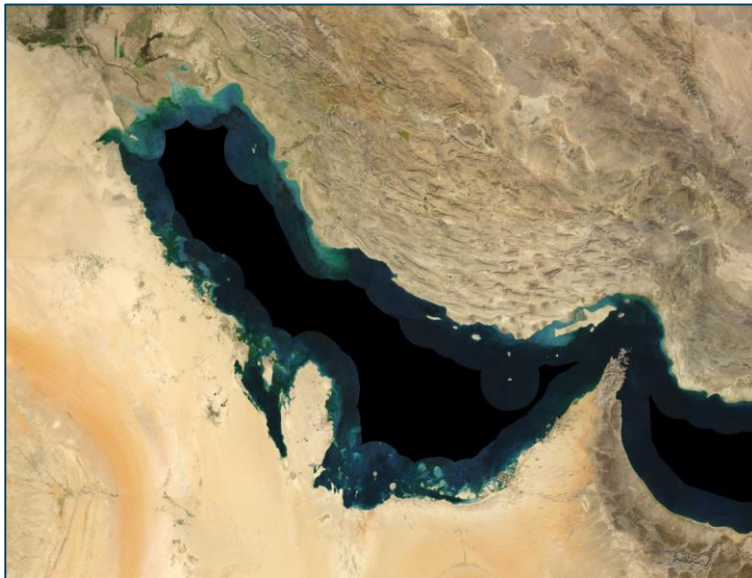
Challenge: automatically detect and digitise objects in the marine environment



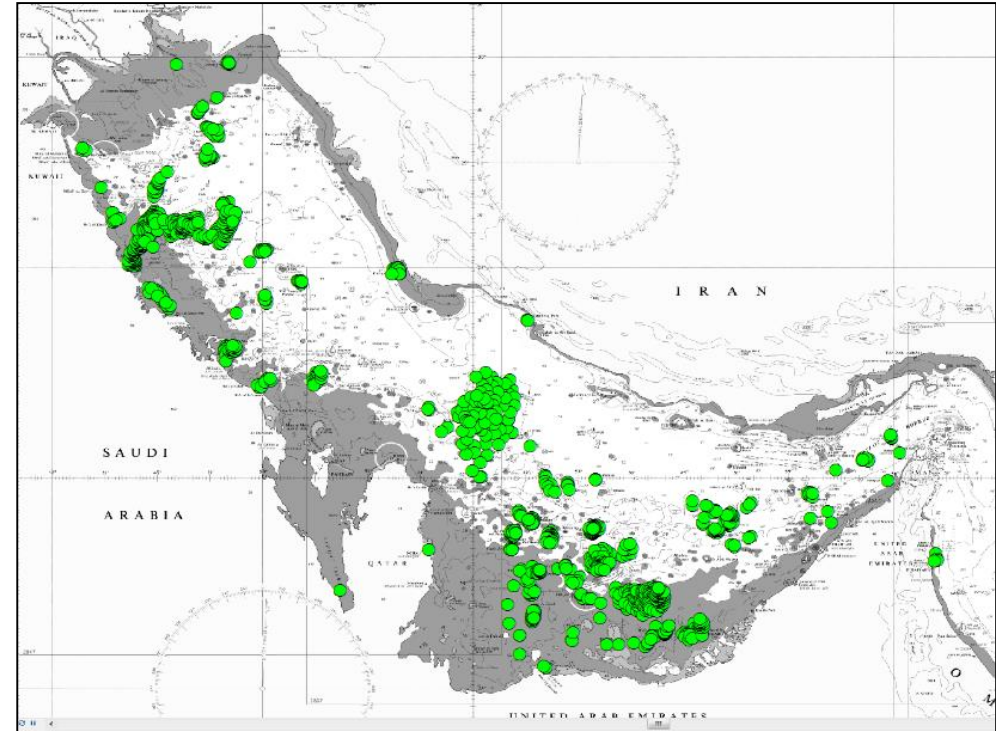
UK Hydrographic Office mentored by Data Science Campus

Process open source satellite data using image classification, object recognition and machine learning techniques

To validate and discover maritime hazards and create a dataset of global offshore infrastructure



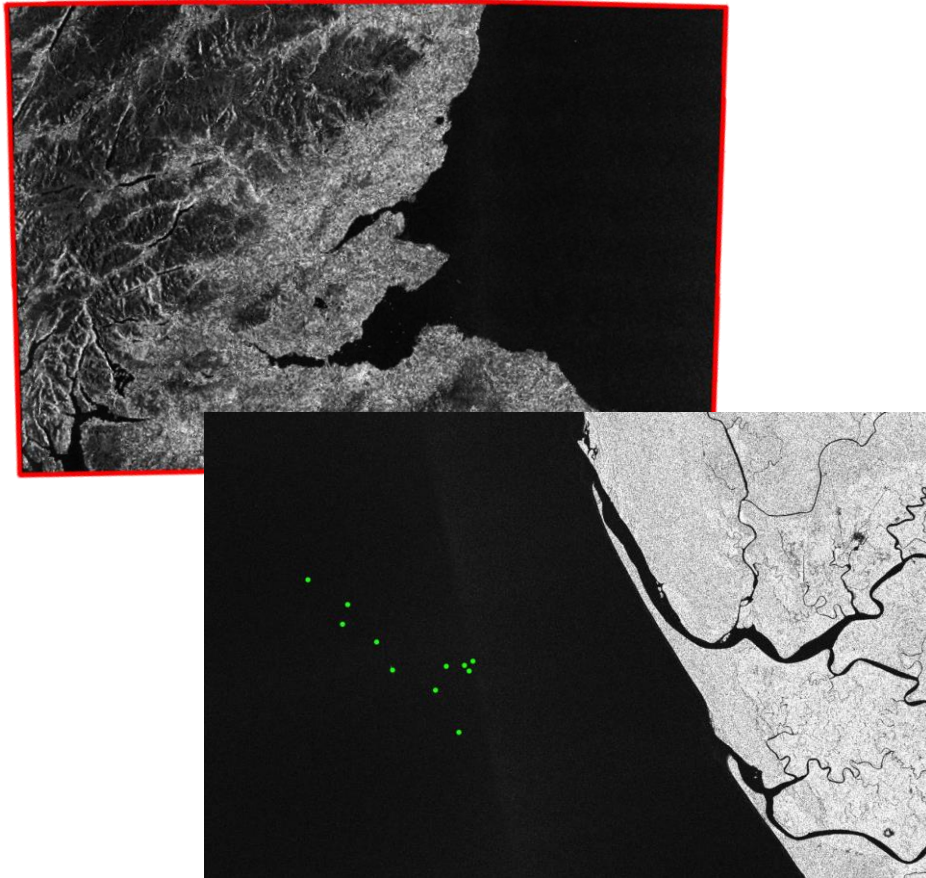
Automated sea object detection from satellite imagery



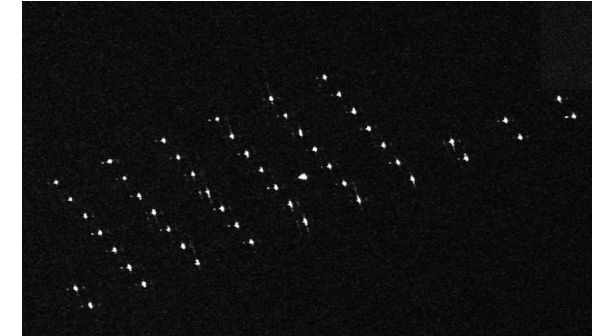
Challenge: automatically detect and digitise objects in the marine environment



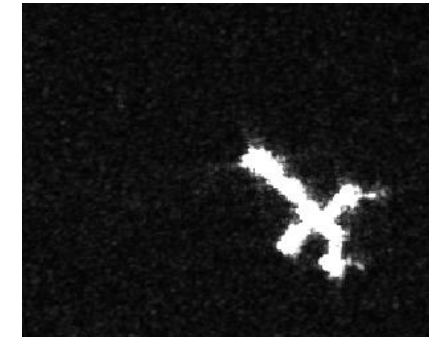
Blob detection, trained on UK data



Wind turbines



Oil platforms



Shipping



United Kingdom
Hydrographic Office

The Data Science Accelerator- Radar Imagery



Sprint
18

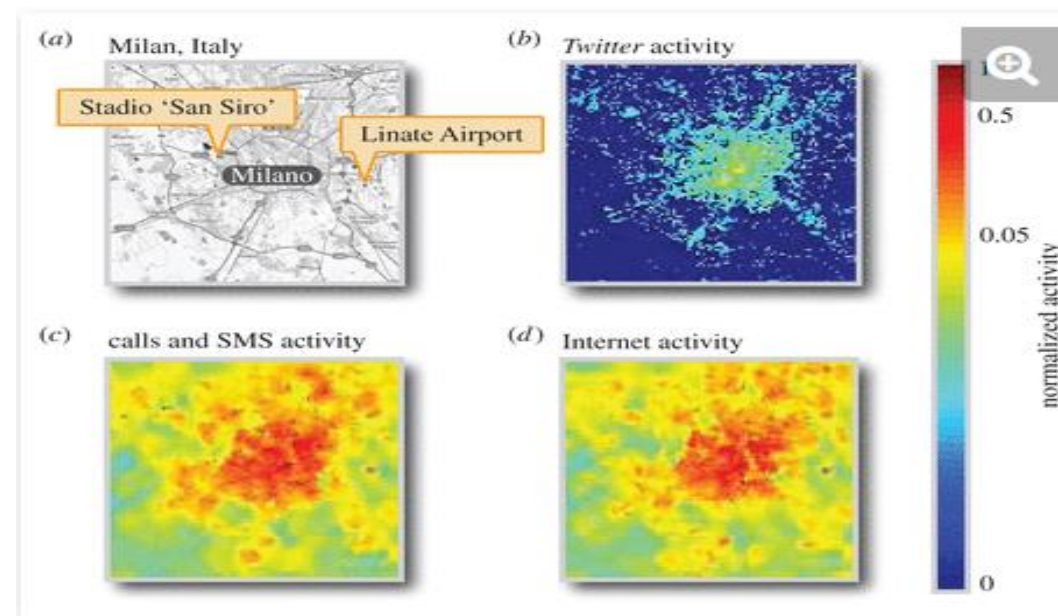
Catherine Seale, Senior Data Scientist at the UK Hydrographic Office, presenting at Sprint 18, London, May 2018

Estimating tourism levels through social media

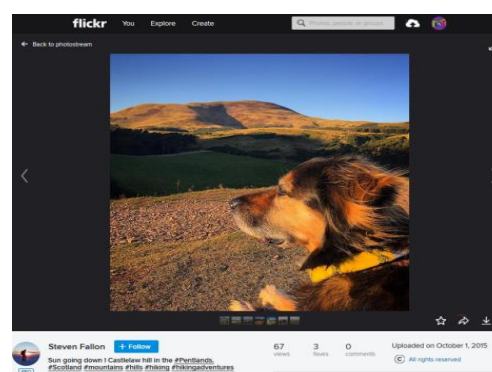


Research questions:

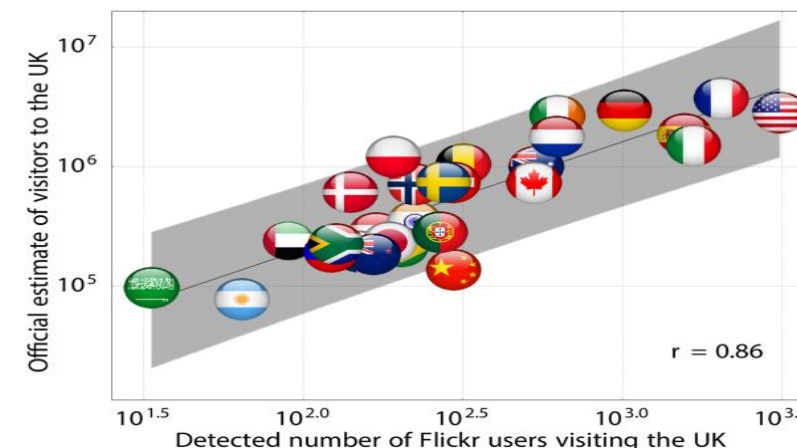
- Alternative data source for quality assurance of International Passenger Survey
- Nationality based under-representation
- Domestic travel trends
- Small area statistics, crowd size estimation
- Google analytics web journey



Visualisation of geo-located Flickr data



Machine Learning classification of photo tags

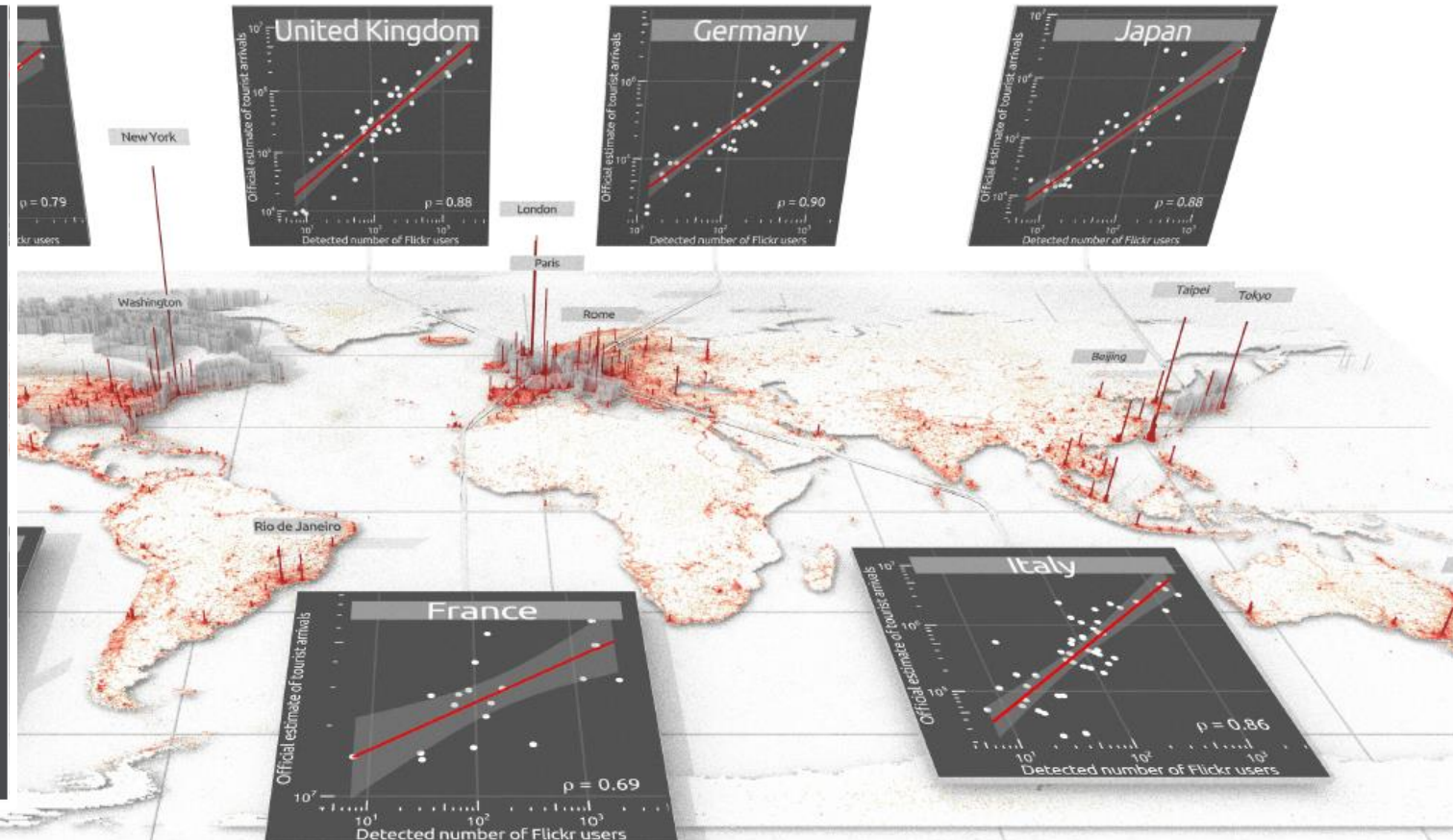
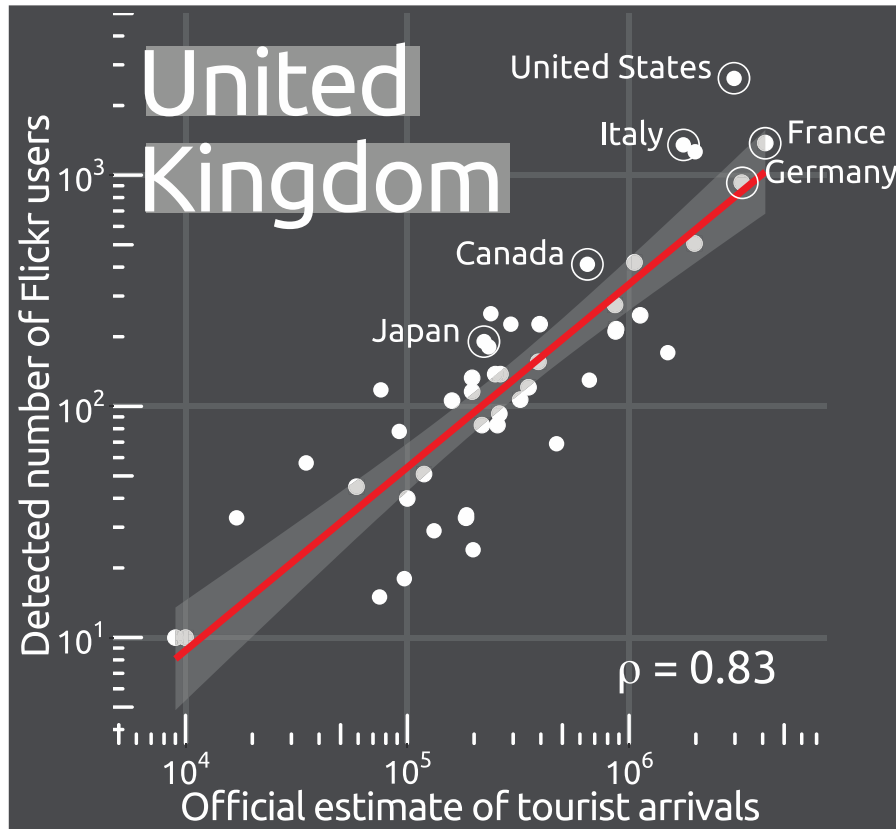


Sensing global tourism numbers with millions of publicly shared online photographs

Environment and Planning A
XX(X):1–4
© The Author(s) 2018
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE



Tobias Preis^{1,2}, Federico Botta¹, Lanthao Benedikt³ and Helen Susannah Moat^{1,2}



Preis et al (2018), Sensing global tourism numbers with millions of publicly shared online photographs



Collaborations with national, international and local government
Agreements with multiple external partners including universities, research institutes and international statistical institutions
NSIs working with / talking to: Netherlands, Rwanda, Canada, New Zealand & Norway
PhD programmes, Centres for Doctoral Training & PhD co-funding with partners
Projects for MSc students in Data Science
Commercial businesses – market engagement with Barclaycard, PwC ...



Growing Data Science Skills

Trainee programmes

- Level 4 Apprenticeship in Data Analytics
- Level 6 Apprenticeship in Data Science (Jan 19)

Data Science Training Unit

- In-house training programmes for ONS and government in coding (R and Python), basic data science skills (machine learning, NLP etc), and intro courses for policy makers

MSc in Data Analytics for Government

- Delivered by University of Southampton, Oxford Brookes and UCL
- 17 students funded, 8 more for 2018/19
- 82 students on CPD courses since Jan 2018

Mentoring Programmes

- Accelerator programme and Data Science Academy
- DECA Exemplar programme



Dr Suzy Moat delivering a guest lecture at the Data science Campus, January 2018



Data Science

Applying the tools, methods and practices of the digital and data age to create new understanding which improves decision-making

(h/t Tom Loosemoore, <https://twitter.com/tomskitomski/status/729974444794494976>)



Data has power & impact

All data is biased

Triangulation is key

Use your skills for good

Making an impact: Realising the potential of urban data science

Tom Smith, @_datasmith
Director, ONS Data Science Campus



**Data Science
Campus**

web: datasciencecampus.ons.gov.uk
email: datasciencecampus@ons.gov.uk
twitter: [@DataSciCampus](https://twitter.com/DataSciCampus)

