Applying data science, machine learning and AI to cross-government challenges

Tom Smith, @_datasmith
Director, ONS Data Science Campus

Data Science Campus

web: datasciencecampus.ons.gov.uk
email: datasciencecampus@ons.gov.uk
twitter: @DataSciCampus
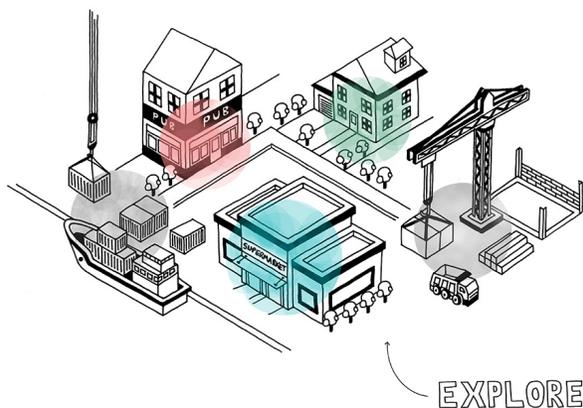
b

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

FINANCIAL TIMES

HOME   WORLD   US   COMPANIES   MARKETS   OPINION   WORK & CAREERS   LIFE & ARTS

ONS 'unicorn' campus reimagines how to measure Britain

Statisticians experiment with using Google Street View, shipping data and VAT returns

The Data Science Campus in Newport © Gareth Iwan Jones/FT

AUGUST 3, 2017 by Chris Giles in Newport, Wales

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.

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
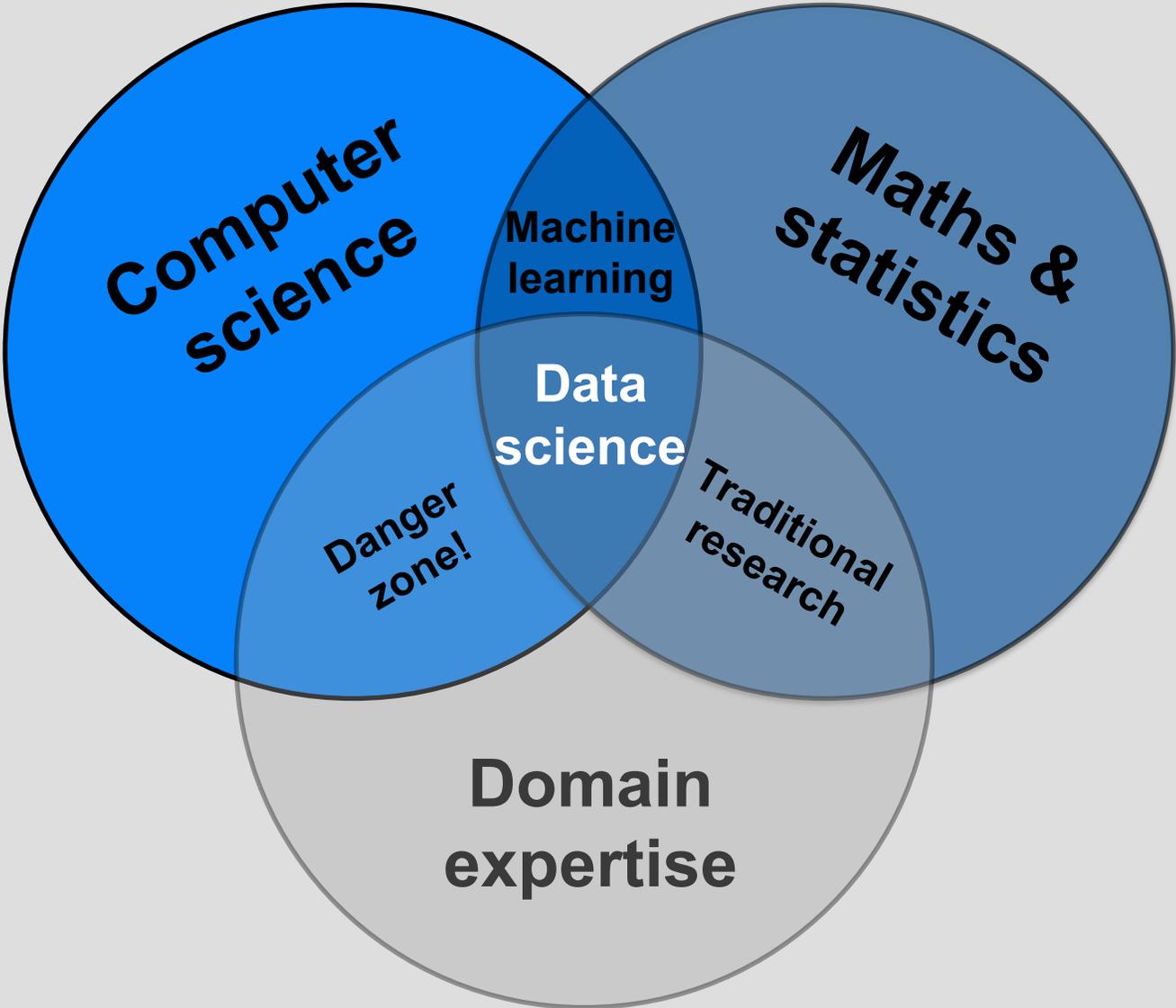
The Data Science Venn Diagram, designed by Drew Conway

"The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand"

- University of Warwick

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

"The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand"

- University of Warwick

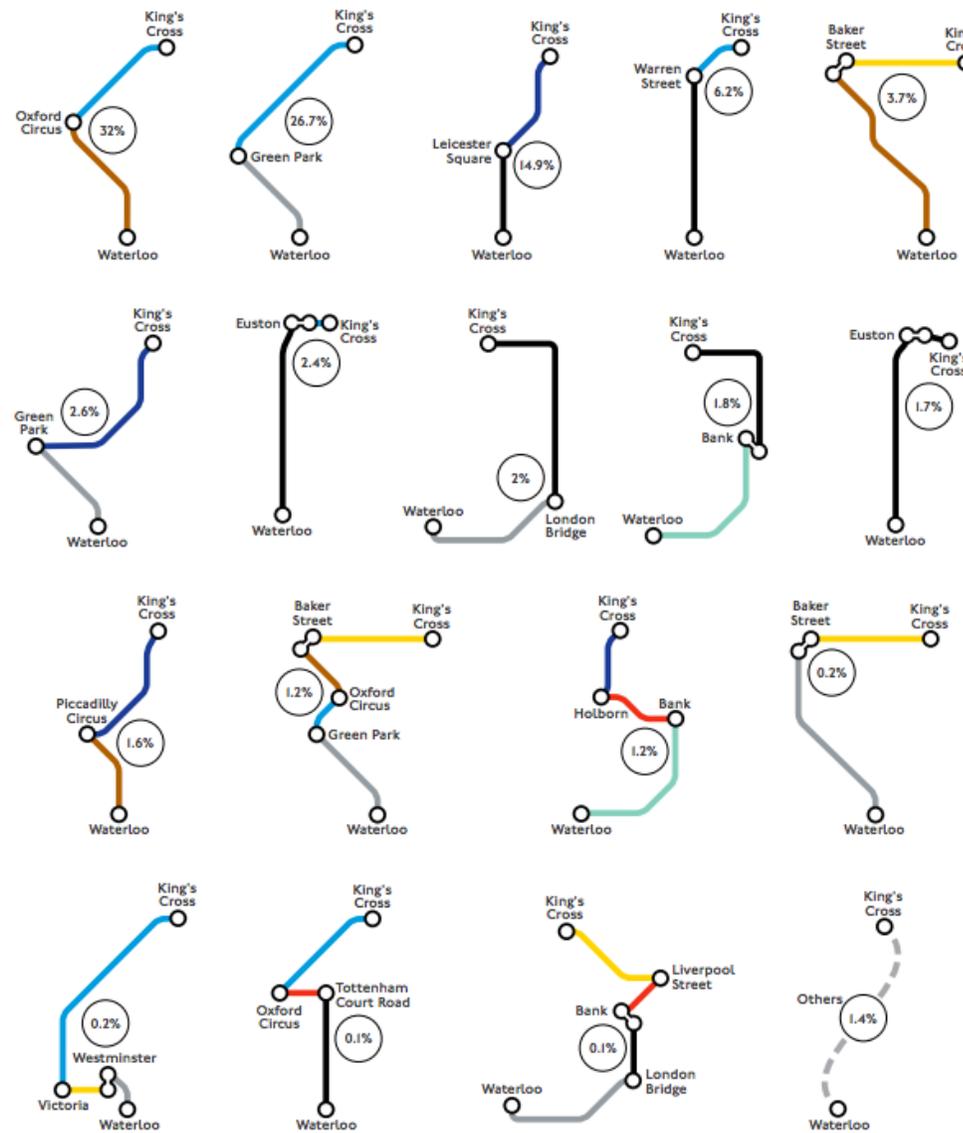Transport for London 2016 pilot, assessing journeys by WiFi access

"The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand"

- University of Warwick

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

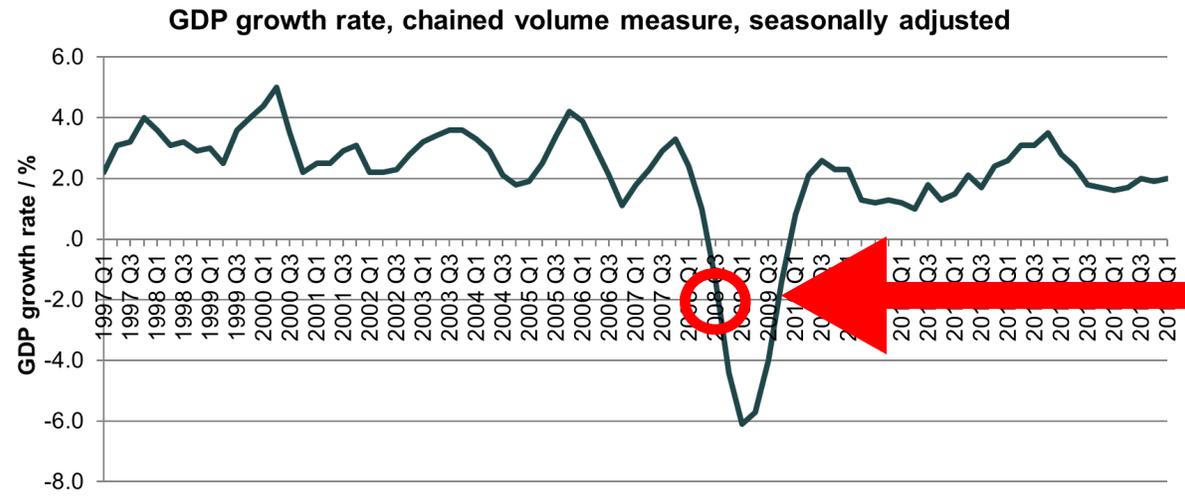Abi Giles-Haigh, 31 January 2018 · Digital data and technology, People and Skills

Predicting Viral Outbreaks

Fig 1. Wordcloud of Norovirus Keywords

Fig 2. Norovirus Keywords in Tweets compared to reported incidents

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

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

Early Indicators of GDP

GDP growth rate, chained volume measure, seasonally adjusted

Fig 1. UK GDP Growth Rate

Early Intervention

Early Indicators

Fig 2. ONS National Accounts Publication Timetable

January   February   March | April   May   June

quarter 1 | quarter 2

quarter 1 preliminary estimate

quarter 1 2nd estimate

quarter 1 quarterly national accounts

-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
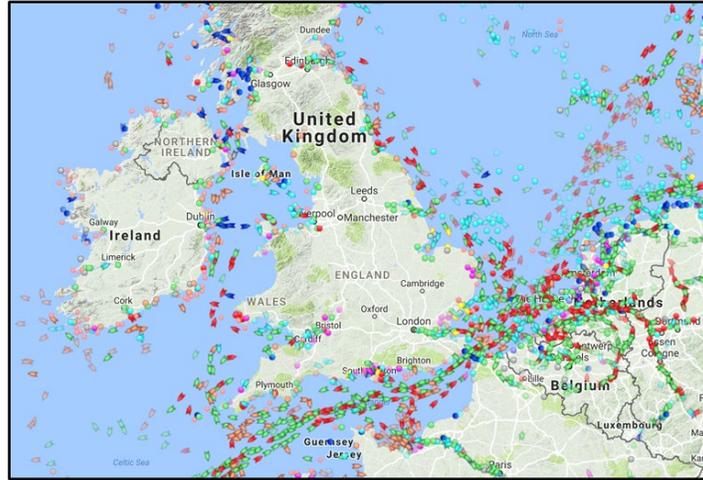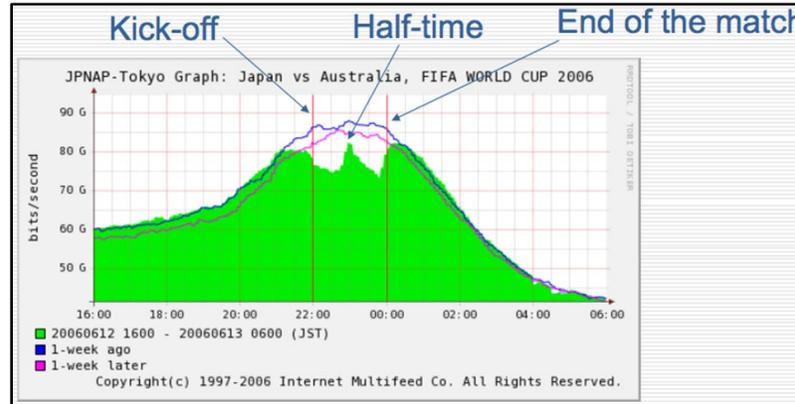Estimated value for earlier identification of 2008 downturn

Early Indicators of GDP

| January | February | March | April | May | June |

| quarter 1 | | | quarter 2 | | |

↑ **VAT turnover returns**

quarter 1 preliminary estimate

quarter 1 2nd estimate

quarter 1 quarterly national accounts
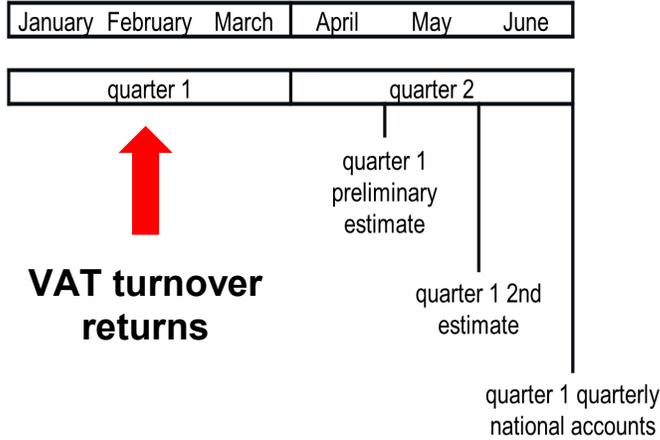
HMRC VAT Data
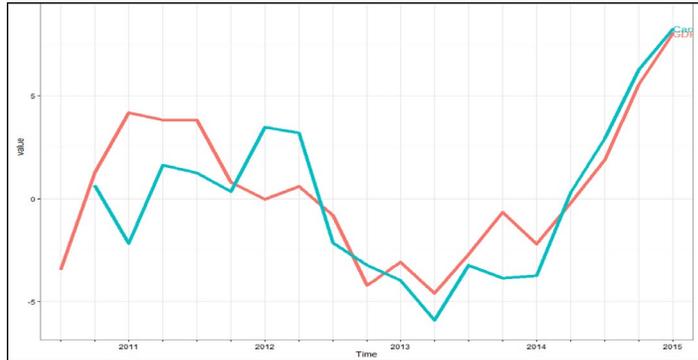
AIS Ship Location

Road Traffic

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

NLP Analysis of NI 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.
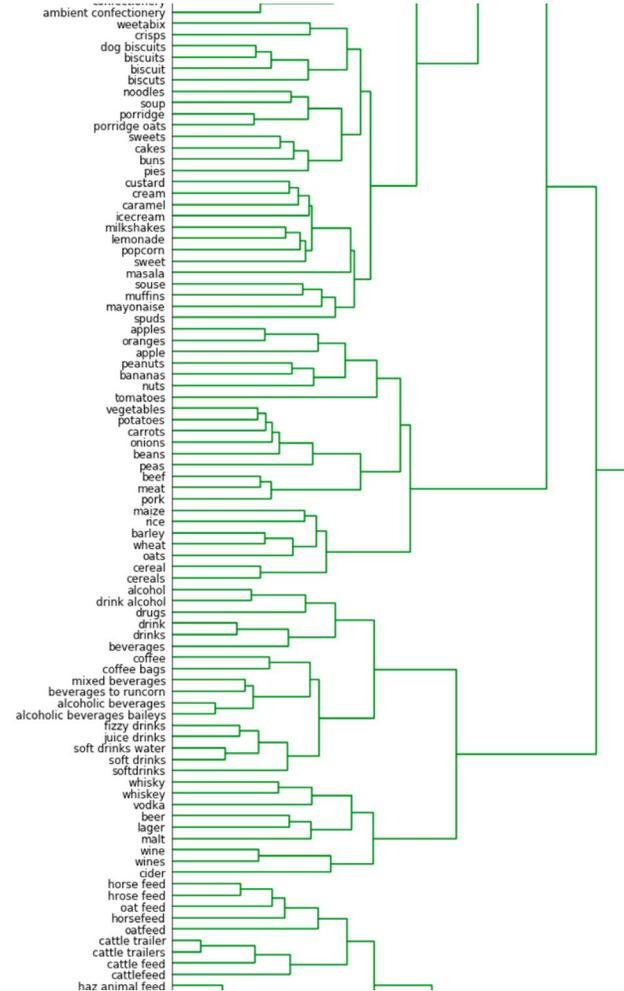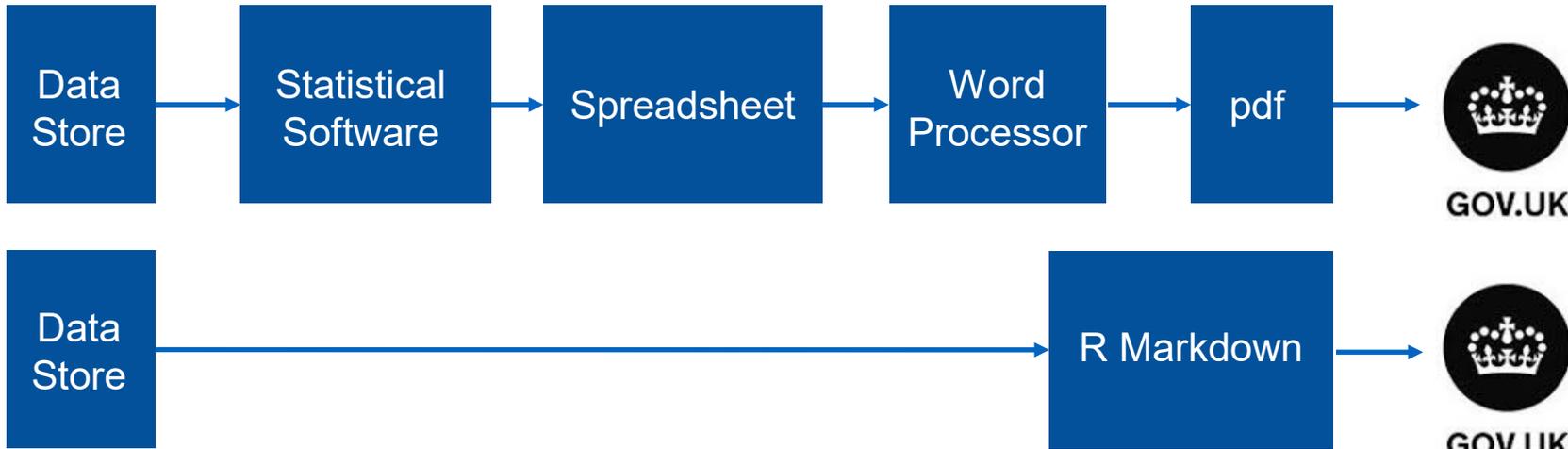

The Data

35k

Lorry journeys in single month analysed during Phase 1

450k

Lorry journeys in 2017 to be analysed during Phase 2

Reproducible Analytical Pipeline

Data Store → Statistical Software → Spreadsheet → Word Processor → pdf → GOV.UK

Data Store → R Markdown → GOV.UK

Efficiency Savings

£118m
Estimated annual efficiency savings across government stats publications

£8.8k
Estimated average annual saving per publication

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

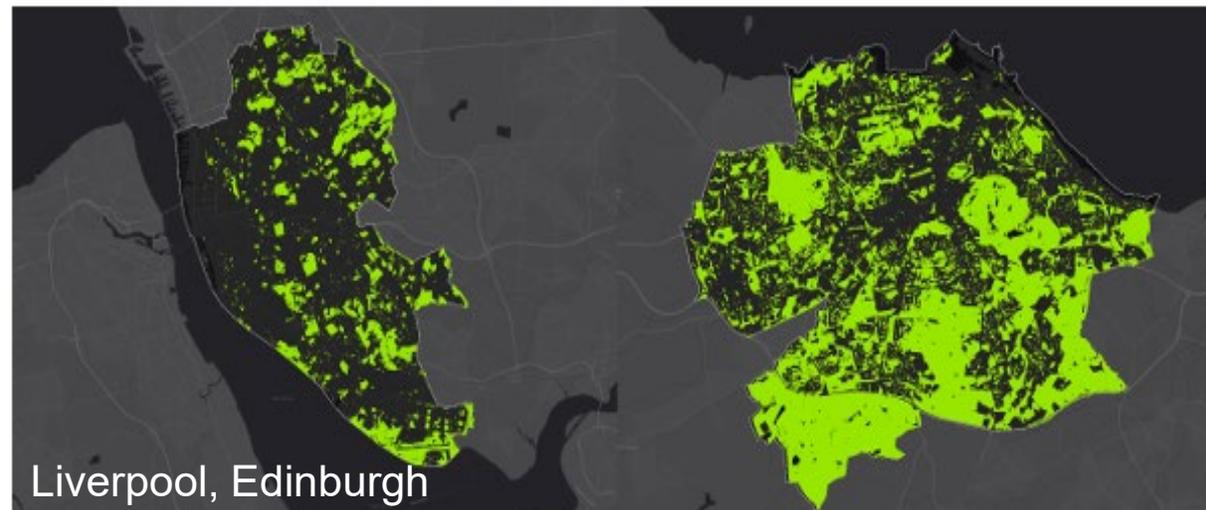# Projects & mentoring with 21 government departments
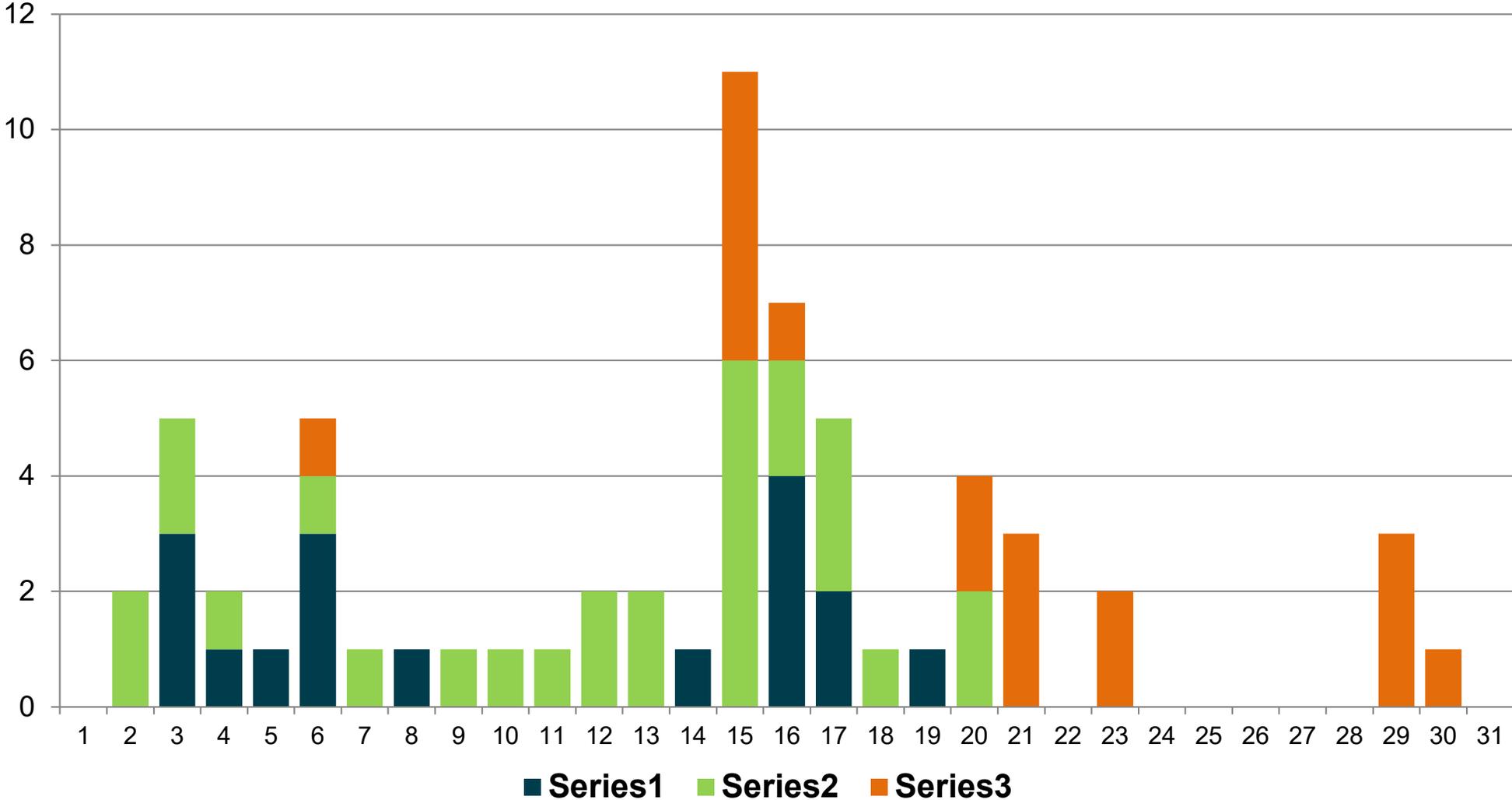
Collaborations with national, international and local government

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 …

Mapping the urban forest – indicators from images

- Analysing images to improve data on local environment
- £1Bn value trees in urban areas (air pollution, health, wellbeing)
- But poor data at local level on tree & urban greenery

Liverpool, Edinburgh

Cardiff

National Tree Map, Blue Sky
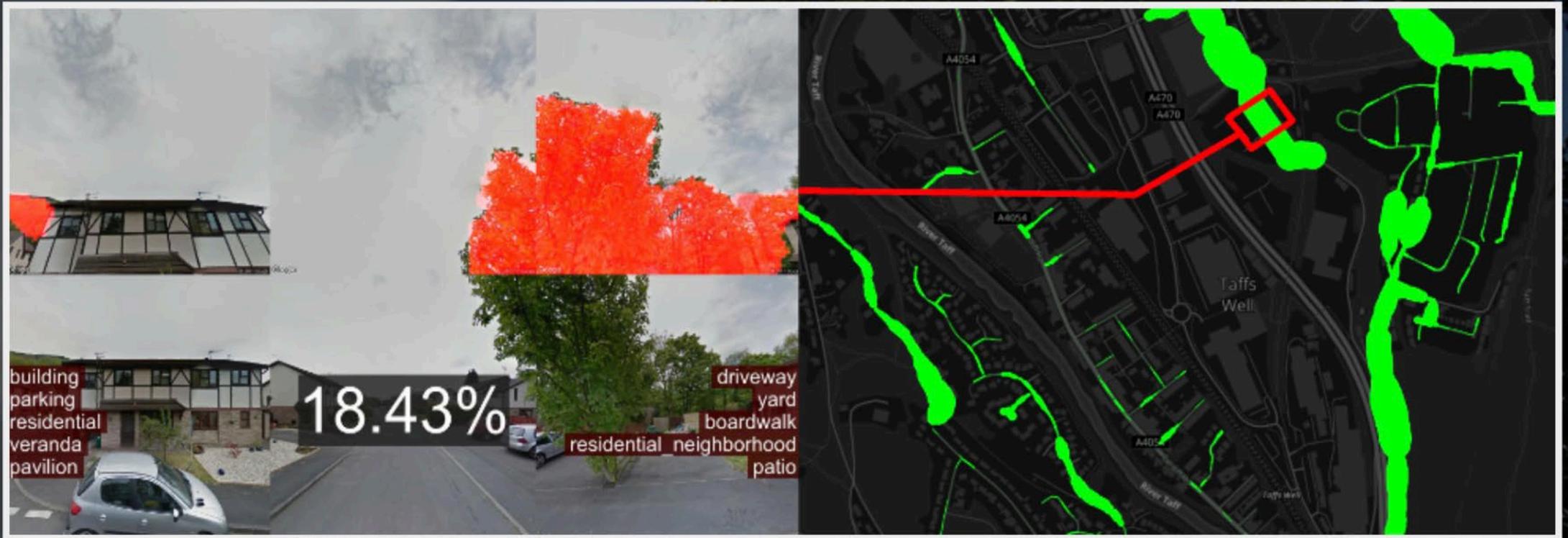
Makes use of:
1. Google streetview imagery
2. OpenStreetMap road network data

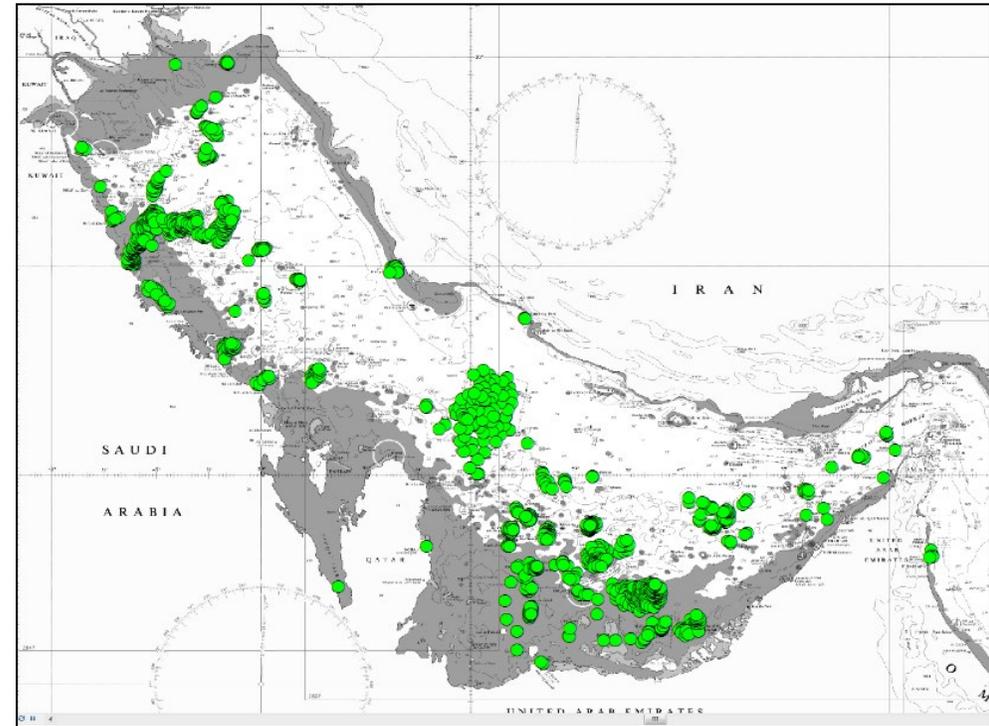# Challenge: automatically detect and digitise objects in the marine environment

UK Hydrographic Office mentored by Data Science Campus

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

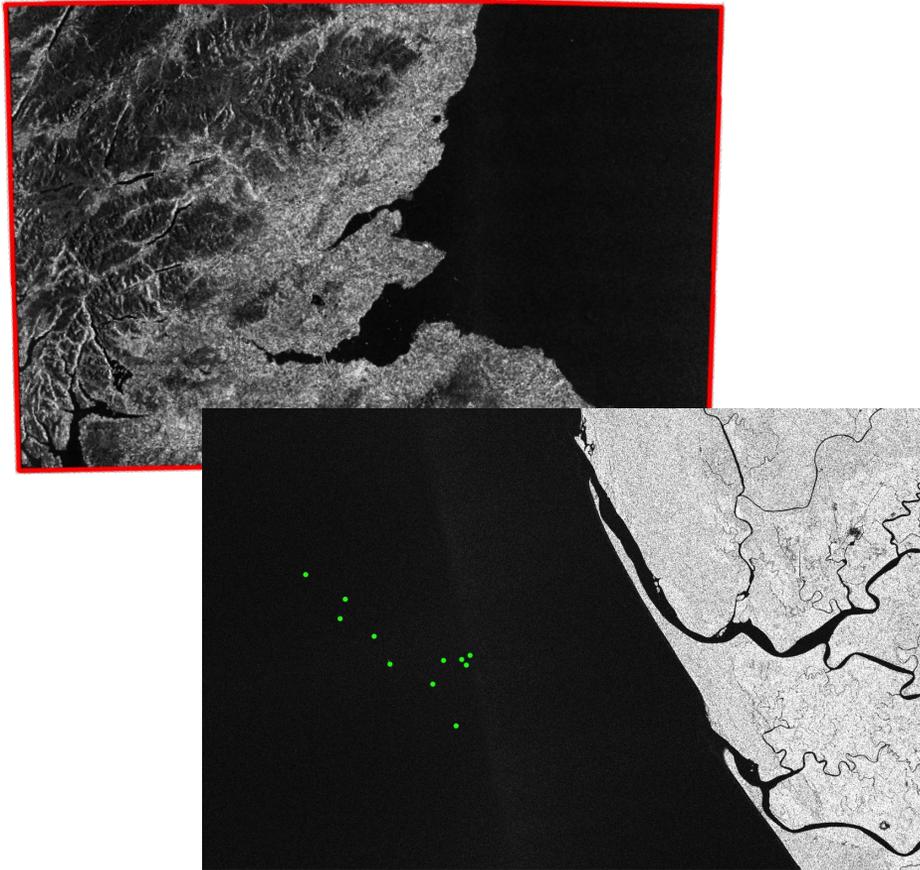To validate and discover maritime hazards and create a dataset of global offshore infrastructure
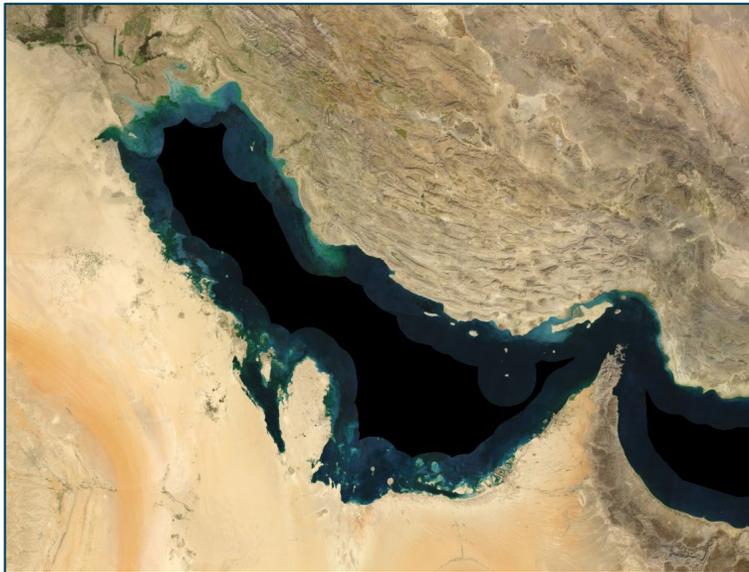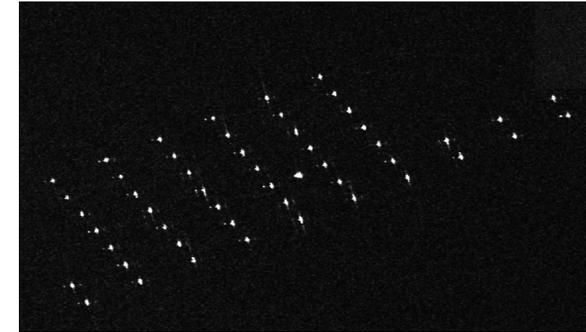
Automated sea object detection from satellite imagery

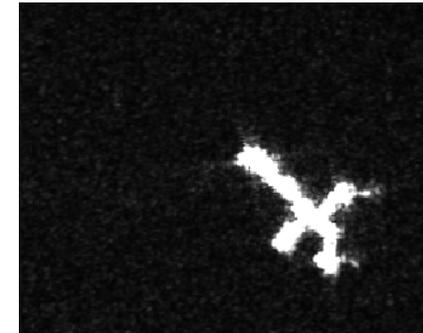# Challenge: automatically detect and digitise objects in the marine environment

Blob detection, trained on UK data

Wind turbines

Oil platforms

Shipping

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

Applying data science, machine learning and AI to cross-government challenges

Tom Smith, @_datasmith
Director, ONS Data Science Campus

web: datasciencecampus.ons.gov.uk
email: datasciencecampus@ons.gov.uk
twitter: @DataSciCampus