New light on old problems
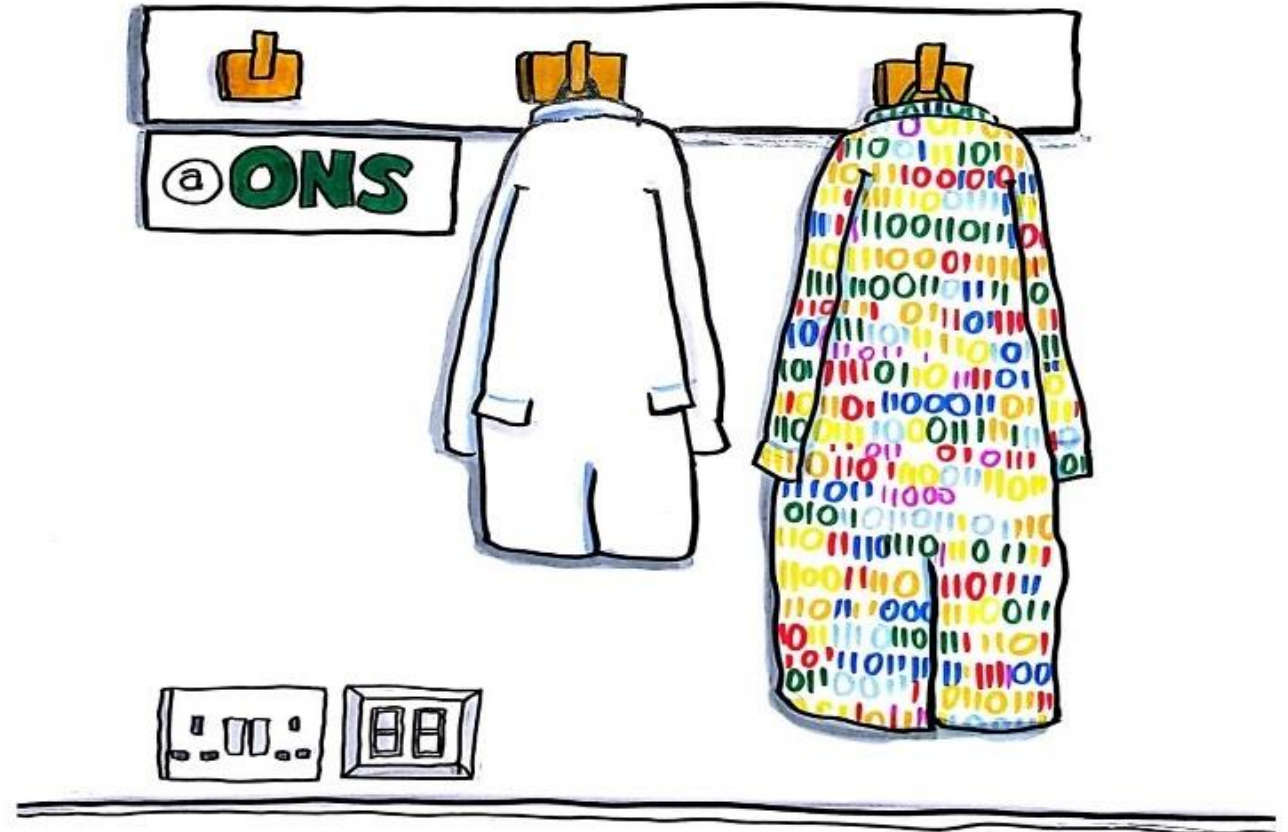# Novel data sources and official statistics

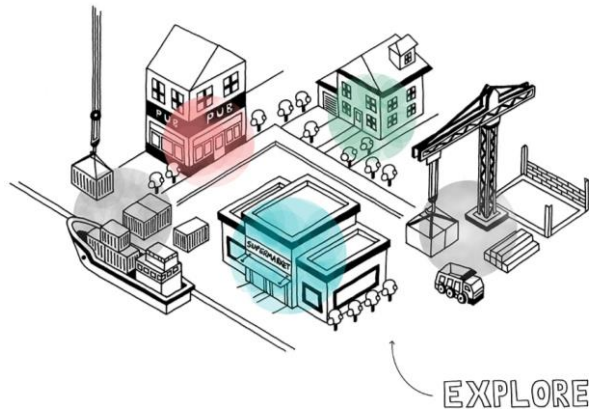Tom Smith, @_datasmith
Director, ONS Data Science Campus

Data Science Campus

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

Economy

GDP
Inflation
Labour market
+++

People

Population
Census
Incomes
+++

World

Trade
Sustainable
   Development Goals
+++
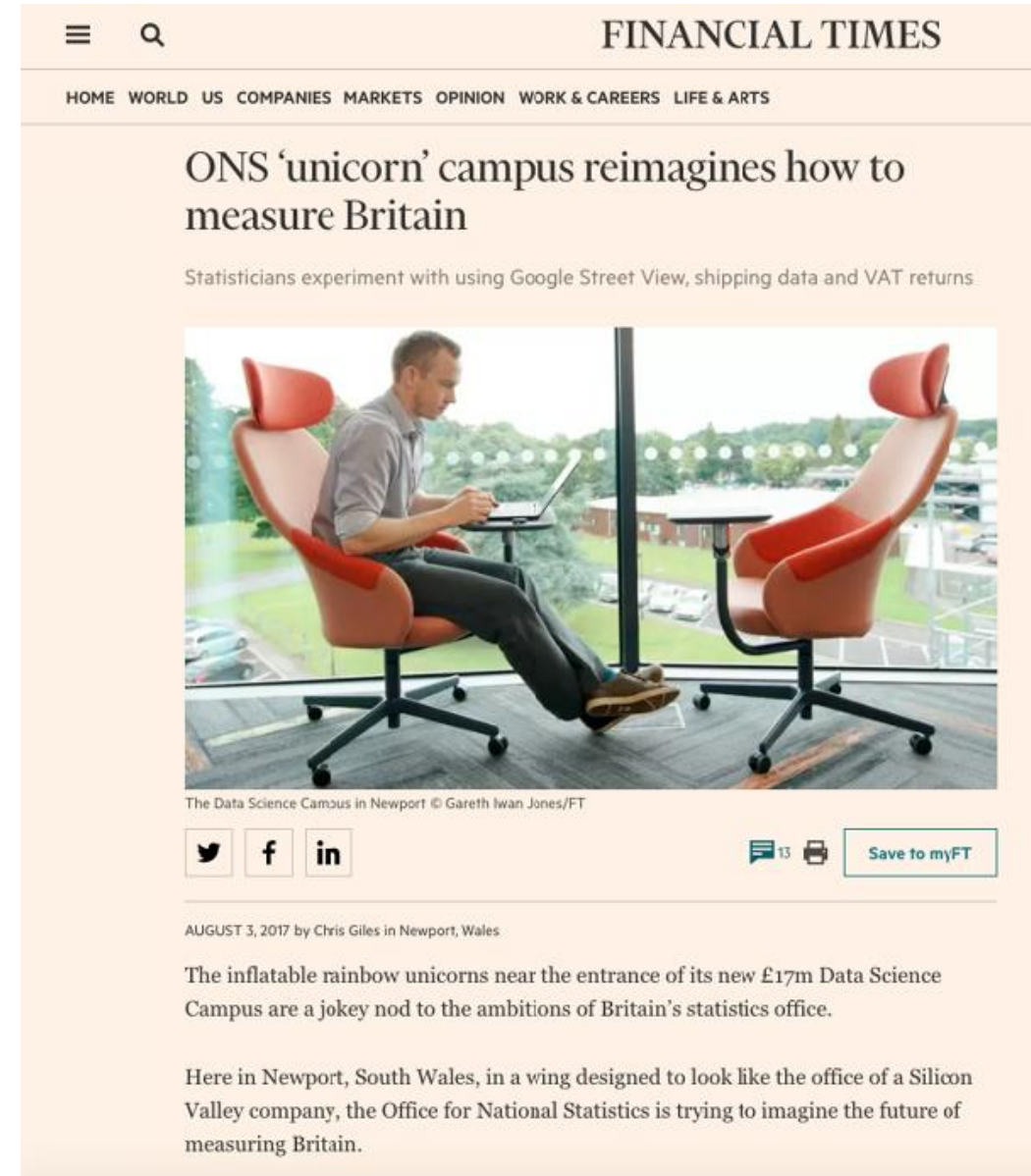
Data Science Campus creation

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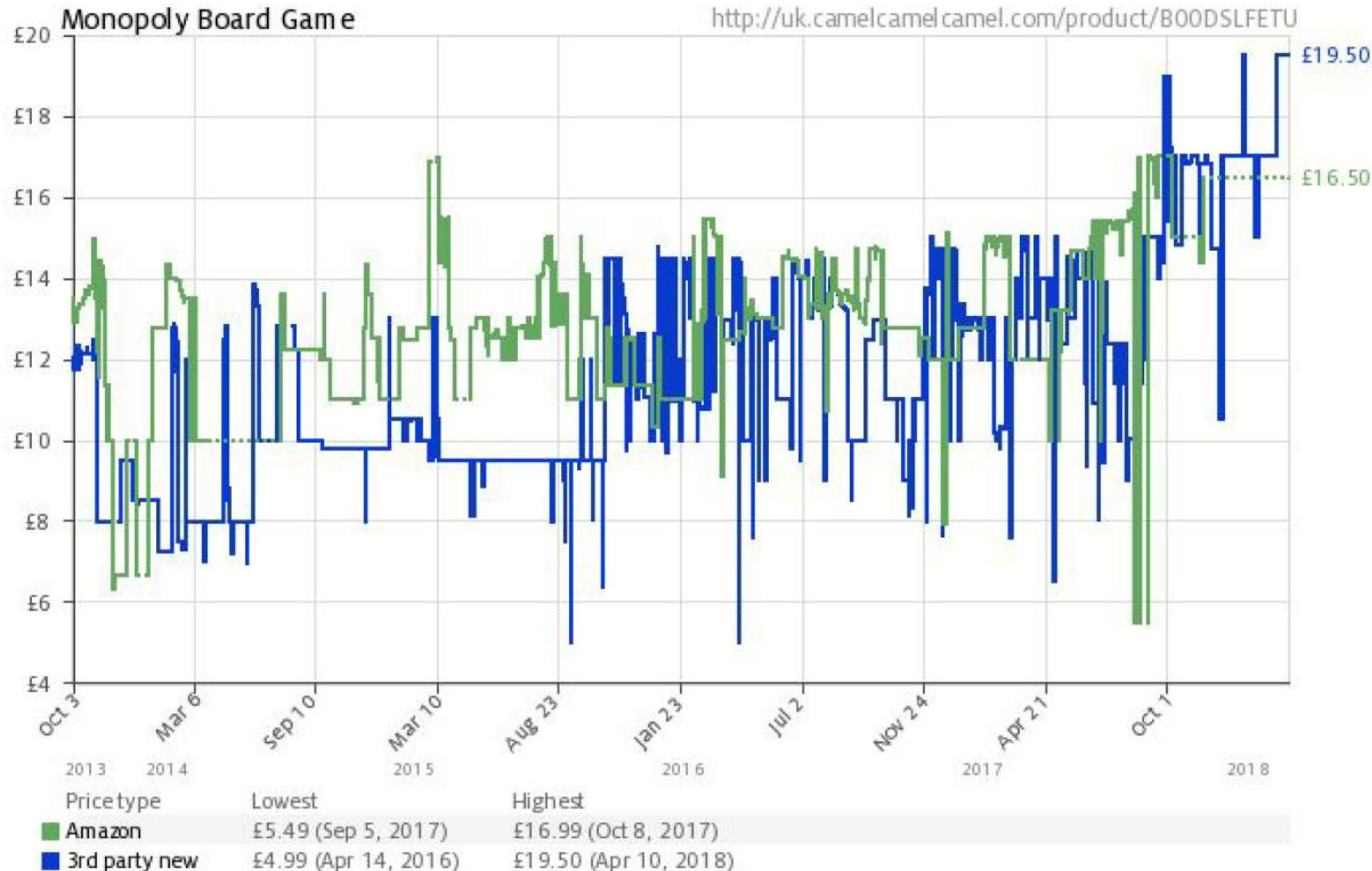
That can be facilitated through **collaboration with relevant partners** – in academia, the private and public sectors, and internationally."

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.

We need big data to understand the world

Monopoly Board Game
http://uk.camelcamel.com/product/B00DSLFETU

Price type	Lowest	Highest
Amazon	£5.49 (Sep 5, 2017)	£16.99 (Oct 8, 2017)
3rd party new	£4.99 (Apr 14, 2016)	£19.50 (Apr 10, 2018)

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
Rybczynksi Prize Essay

"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 longer a decision-maker has to wait for the statistics, the less useful they are likely to be

Professor Sir Charles Bean (2016)

Review of UK economic statistics

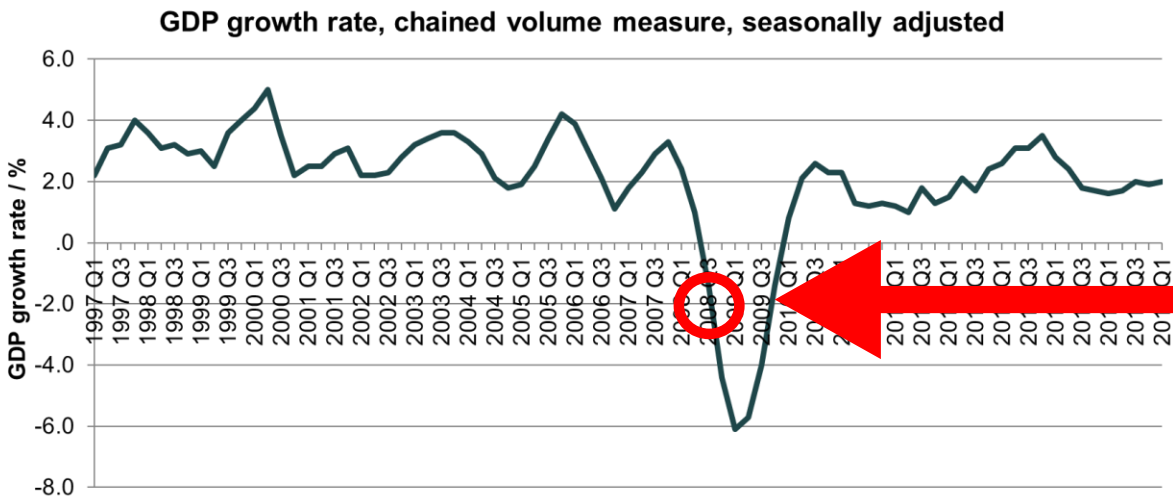# Faster indicators of the economy

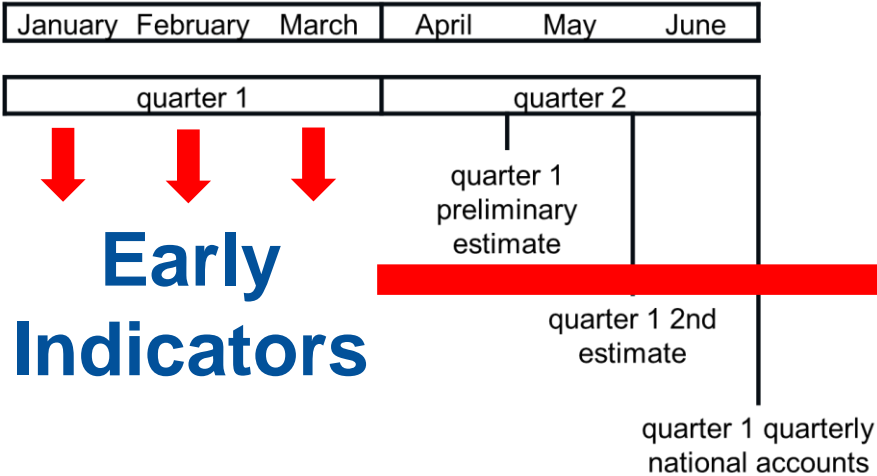GDP growth rate, chained volume measure, seasonally adjusted

Fig 1. UK GDP Growth Rate

Fig 2. ONS National Accounts Publication Timetable

Early Intervention

Early Indicators

-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

Business value added tax (VAT) returns

- Expenditure and turnover diffusion indices
- Reporting behaviour
- Available up to 1 month before GDP

VAT due in this period on **sales** and other outputs — 1

VAT due in this period on **acquisitions** from other **EC Member States** — 2

Total VAT due (**the sum of boxes 1 and 2**) — 3

VAT reclaimed in this period on **purchases** and other inputs (including acquisitions from the EC) — 4

Net VAT to be paid to HMRC or reclaimed by you (**Difference between boxes 3 and 4**) — 5

Total value of **sales** and all other outputs excluding any VAT. **Include your box 8 figure** — 6

Total value of **purchases** and all other inputs excluding any VAT. **Include your box 9 figure** — 7

Total value of all **supplies** of goods and related costs, excluding any VAT, to other **EC Member States** — 8

Total value of all **acquisitions** of goods and related costs, excluding any VAT, from other **EC Member States** — 9
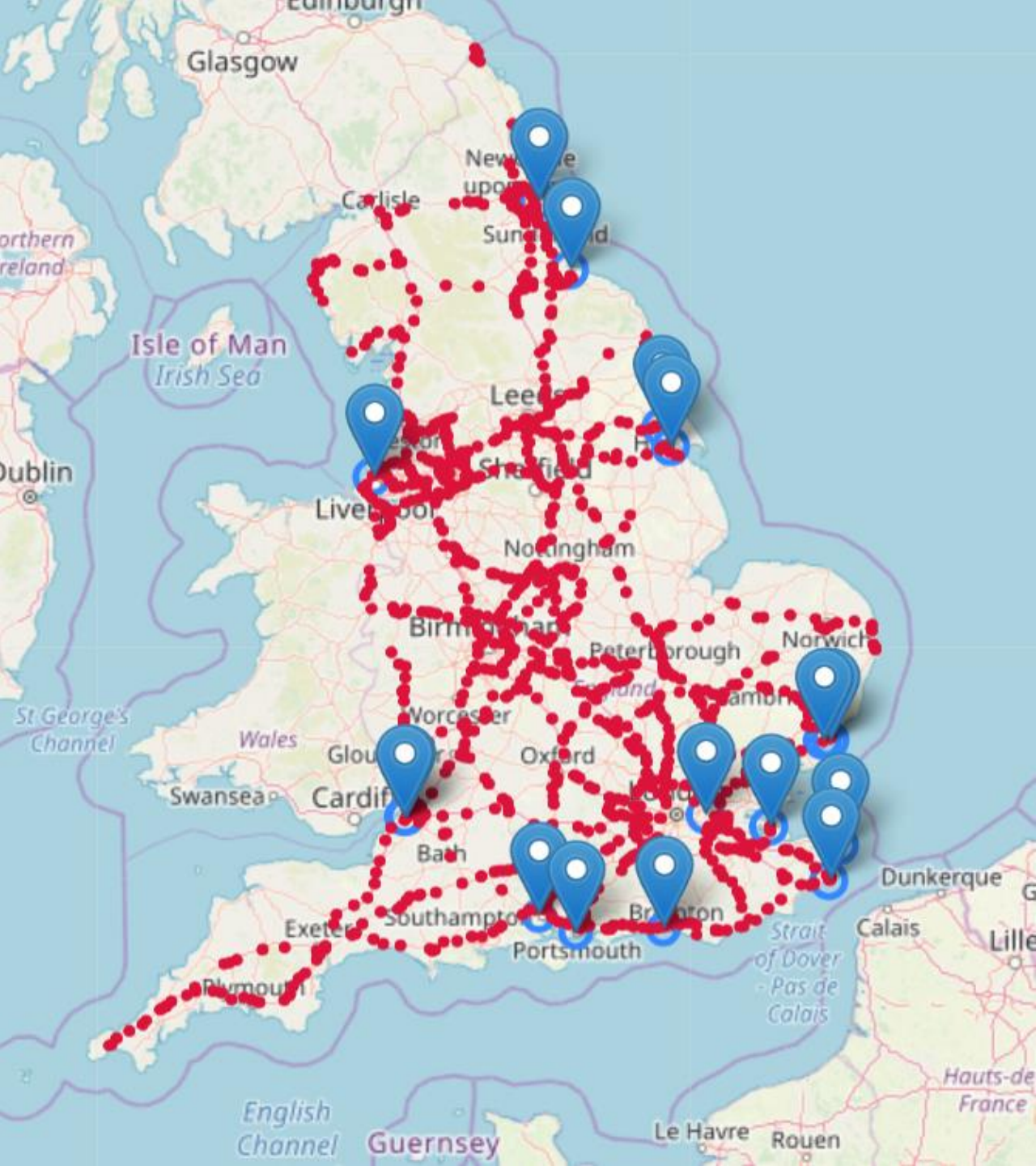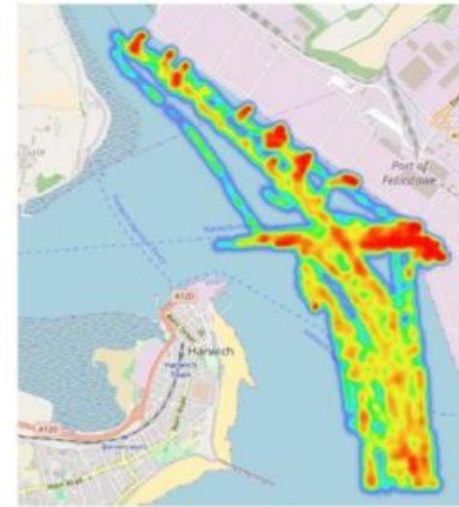
Road traffic data

- Highways England sensor data

- Average speeds

- All-England and English ports

- By vehicle length

- Available 2 months before GDP

Shipping Automated Information System (AIS) data

- Marine and Coastguard Agency, ORBCOMM, Global Platform
- Ship tracking data
- Port traffic frequency
- Time in port
- Real time

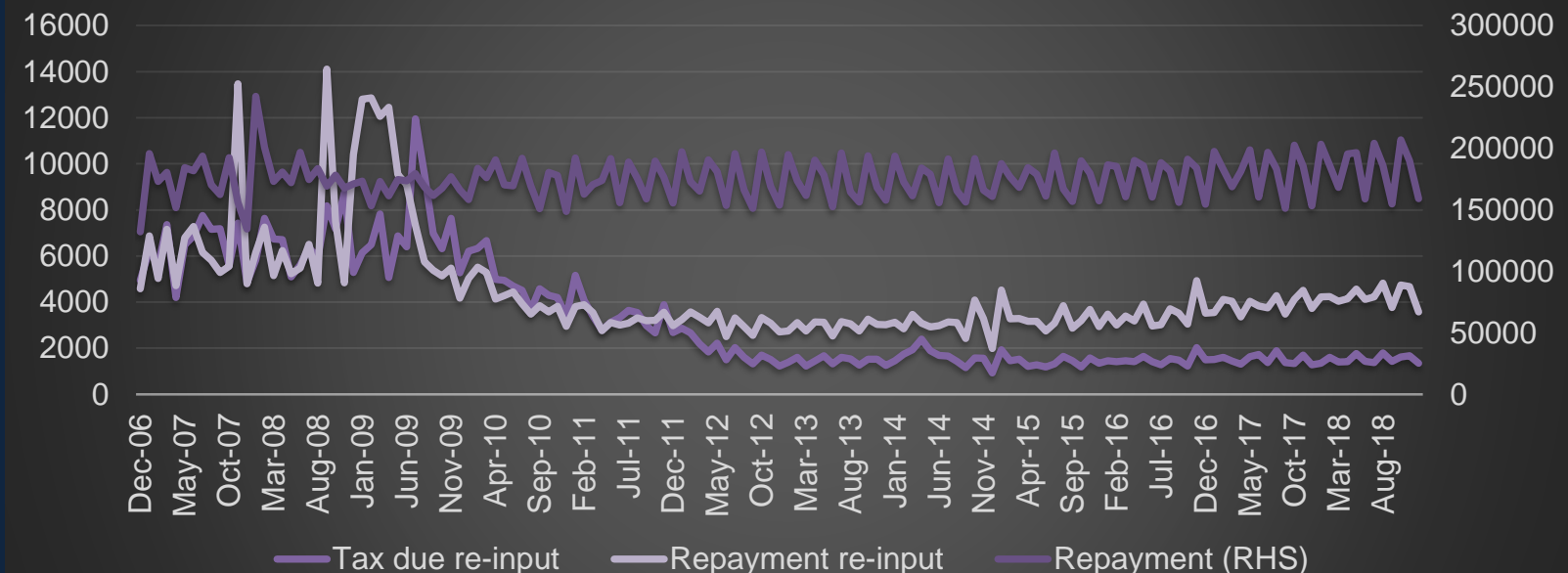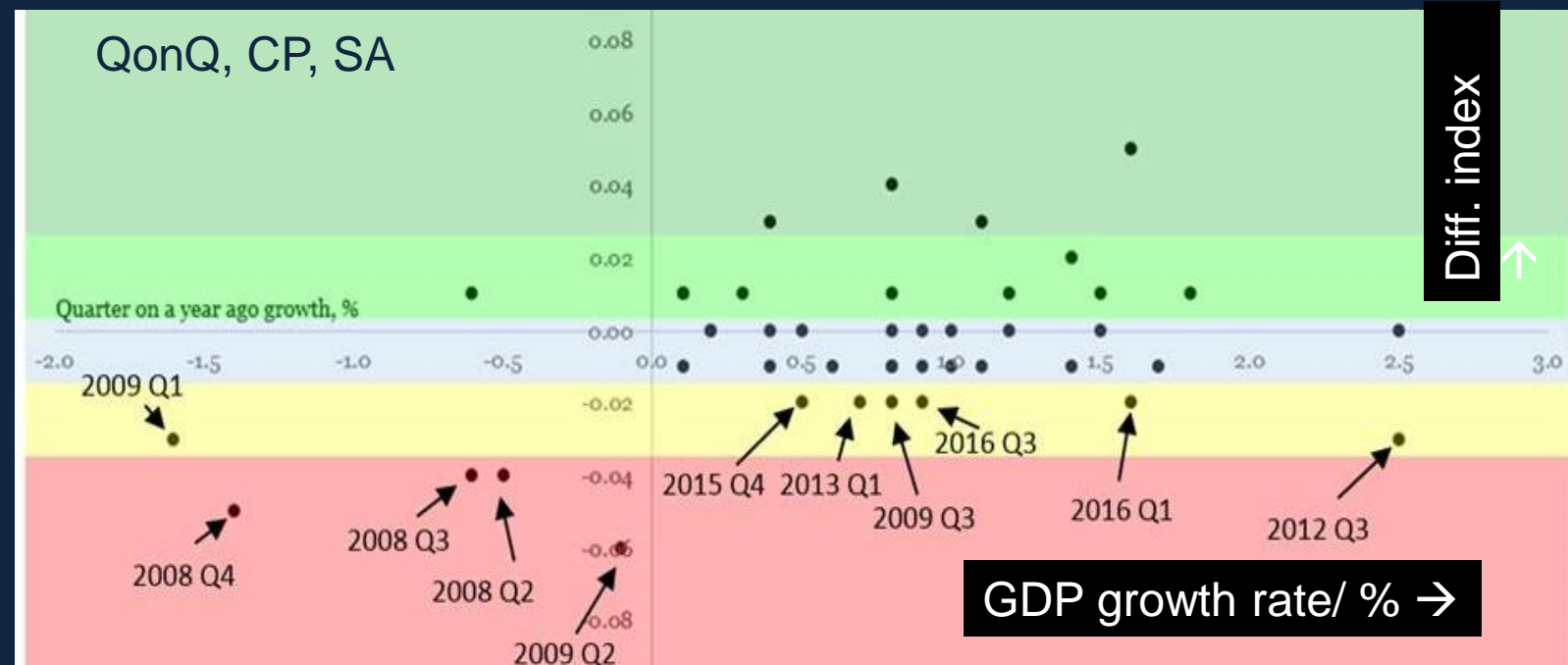Monday 12 December 2016

Monday 19 December 2016

Sunday 25 December 2016

Sunday 1 January 2016

- VAT is a good indicator of large change
- Last recession identified 5 months before official statistics
- Novel repayments indicators show financial stress
- Care with over-interpretation and beware bias!

FINANCIAL TIMES

UK COMPANIES TECH MARKETS GRAPHICS OPINION WORK & CAREERS LIFE & ARTS HOW TO SPEND IT

nomic growth

ppers unfazed by Brexit as UK
il sales jump

UK house price growth weakest in
seven years

Slowing UK productivity rings
alarm bells

fastFT UK economic growth   + Add to myFT

Economy lost momentum in first quarter, ONS 'big data' show

New system based on VAT returns aims to be early warning for looming problems

Bloomberg

Economics

Faster Indicators for U.K. Point to Modest Decline in Turnover

By Andrew Atkinson and Fergal O'Brien
April 15, 2019, 10:28 AM GMT+1

The Telegraph   HOME   NEWS   S

Business

Economy | Companies | Opinion | Open economy | Markets | Alex | Telegraph

FTSE 100 ▲   FTSE 250 ▲   GBP/USD ▼   GBP/EUR ▲   BRENT OIL ▼
7523.07 +0.85%   19908.02 +0.36% $1.2943 -0.31%   €1.1543 +0.09% $74.20 -0.07%

◆ PREMIUM

🏠 › Business

Early warning indicators not flashing red: few signs of economic slowdown in March

THE TIMES   SUBSCRIBE   LOG IN

MARCH 20 2019, 12:01AM, THE TIMES

Even a skip-counter like me agrees economic indicators need updating

DAVID SMITH

Today I want to take you away from the Alice in Wonderland world of Brexit, populated by strange creatures often called Steve or John, and deal with something that will still be important long after the dust has settled. My interest was piqued

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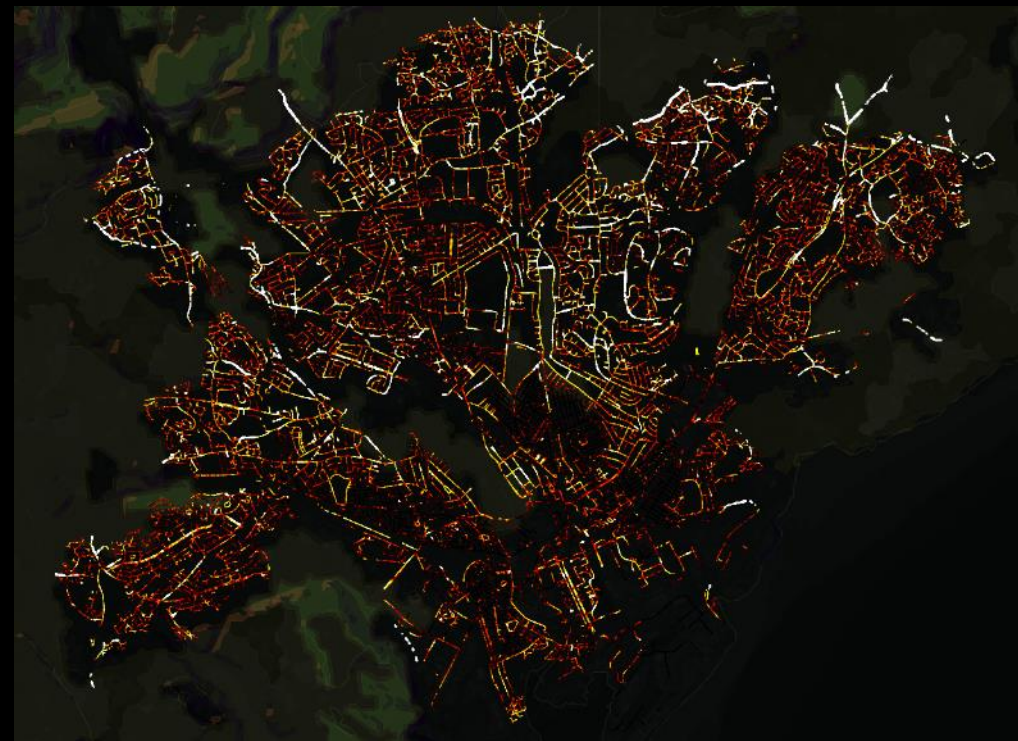
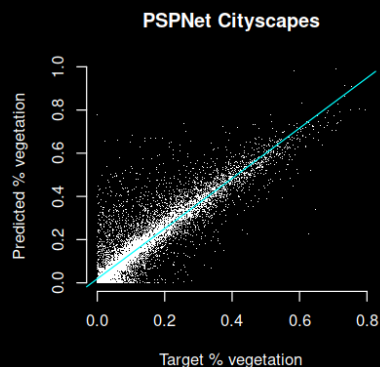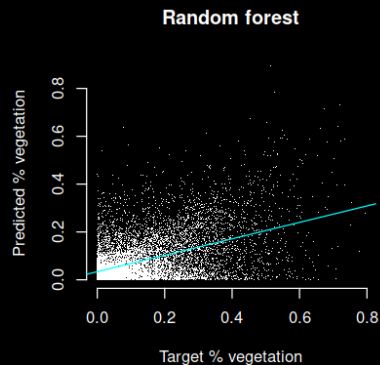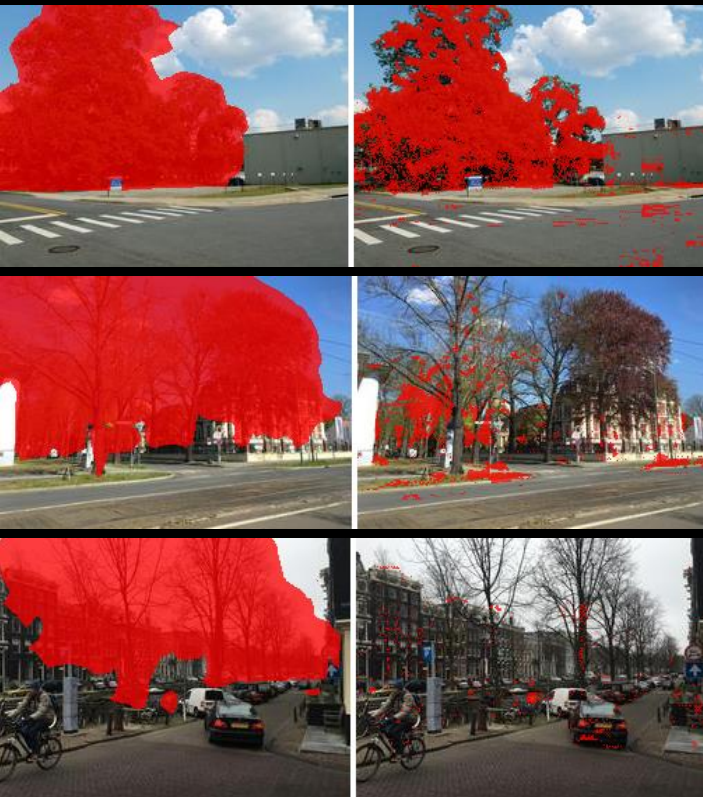For Cardiff, the annual benefit is close to £8 million.

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

Outcome: An end-to-end processing pipeline.

Making use of: **17 million images** 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/

Article
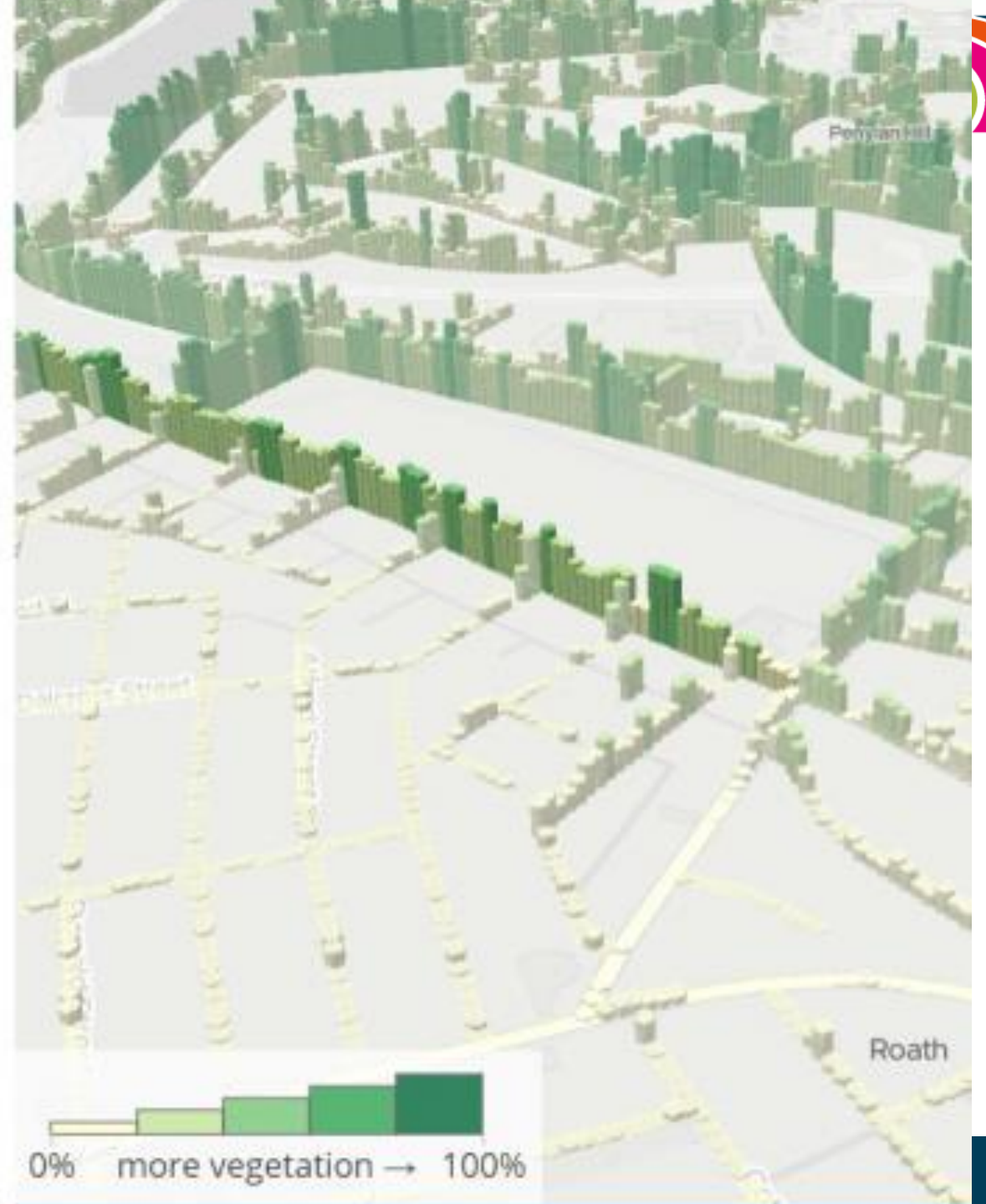
A Hierarchical Street Tree Index Using Street-Level Imagery and Deep Learning

Philip Stubbings [a,1]*(ID), **Joe Peskett** [a,2](ID), **Francisco Rowe** [b,3] (ID) and **Dani Arribas-Bel** [b,4](ID)

[a]   Data Science Campus. Office for National Statistics
[b]   Geographic Data Science Lab, Department of Geography & Planning. University of Liverpool
[1]   philip.stubbings@ons.gov.uk
[2]   joseph.peskett@ons.gov.uk
[3]   F.Rowe-Gonzalez@liverpool.ac.uk
[4]   D.Arribas-Bel@liverpool.ac.uk
*   Correspondence: philip.stubbings@ons.gov.uk; Tel.: +44-1633-65-1977

1   **Abstract:** In this study, we describe a computer vision based method to derive an Urban Street Tree
2   Vegetation Index which aims to quantify the amount of vegetation visible from the point of view of a
3   pedestrian. We first generalise the problem of detecting trees and vegetation at street level as a form
4   of semantic image segmentation. Using labelled data, we then compare two optimised colour-space
5   thresholding methods to identify green pixels with a deep learning based approach. While image
6   thresholding is computationally efficient and effective in controlled environments, we find a deep
7   learning based approach to image segmentation superior in this context, particularly with respect to
8   the problem of detecting diverse, possibly non-green tree species in different seasons and varying
9   lighting conditions. Having identified a robust method to determine the quantity of vegetation
10 present in arbitrary images, we then sample 220,068 street-level images for a large city in the UK and
11 produce a vegetation index for 10 metre intervals along the entirety of the city road network. Using
12 this high resolution dataset, we then generate indices for Lower Layer Super Output Areas (LSOAs)
13 using a hierarchical linear model which we use to correct for visual obstructions typical in the urban
14 environment.

Text 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 an NLP 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.
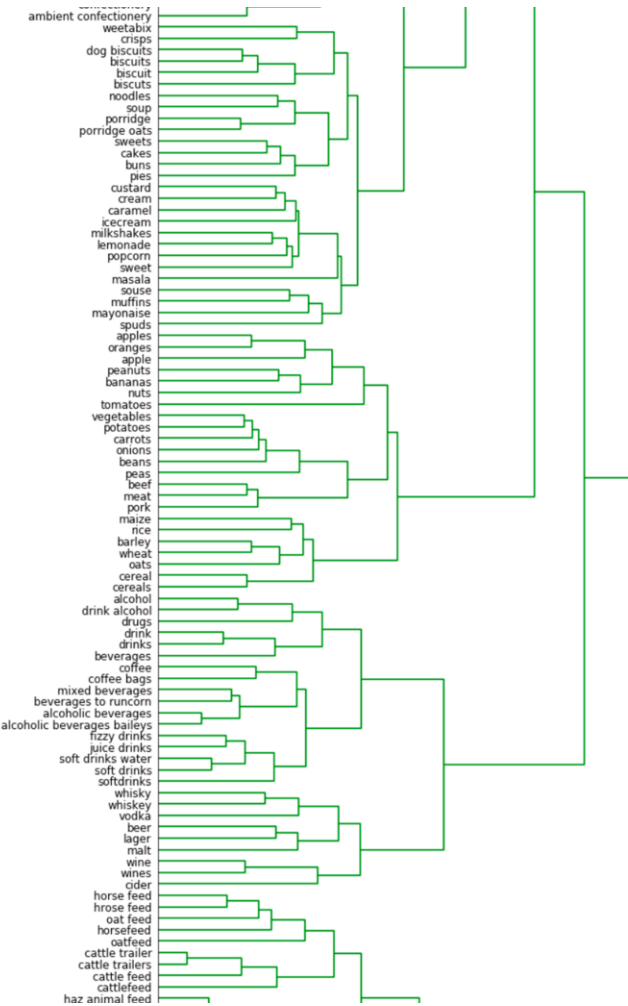
The Data

35k

Lorry journeys in single month analysed during Phase 1

450k

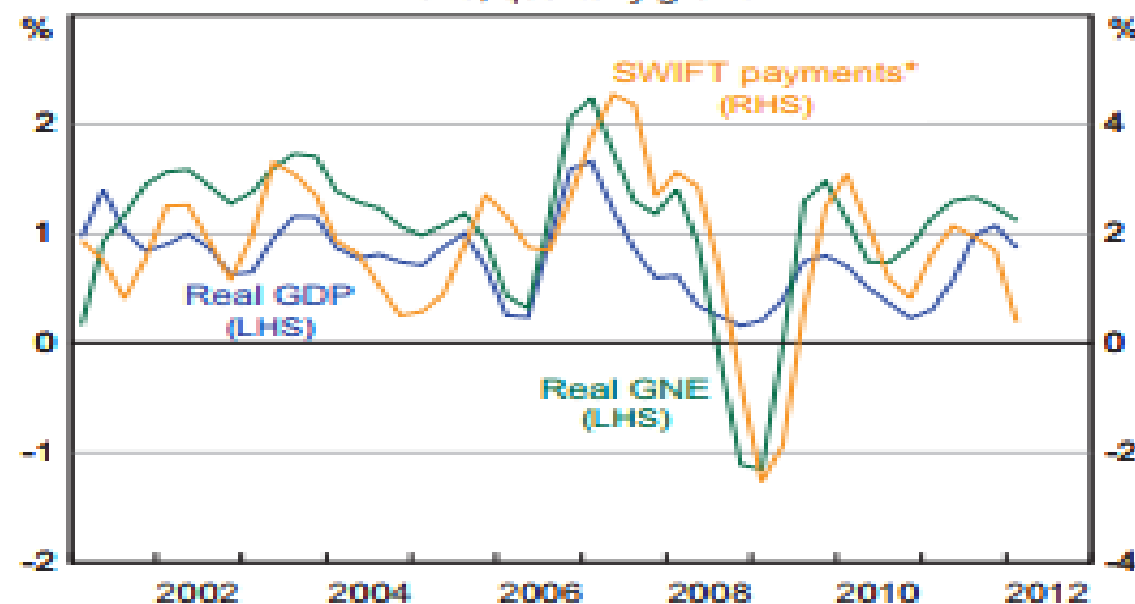Lorry journeys in 2017 to be analysed during Phase 2

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?

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

Growing Data Science Skills

Level 4 Apprenticeship in Data Analytics:

Two-year programme, 12 months at the Campus followed by 2 x 6 month rotations across ONS

Level 6 Apprenticeship in Data Science:

Three-year programme, approved for England in 2018, launching in England and Wales in 2019

Data Science Faculty:

In-house training unit delivering operational courses in programming (R, Python) and fundamentals of Machine Learning, NLP etc

Accelerator:

12-week mentoring programme for Government analysts delivered with GDS, in parallel with in-house ONS Data Science Academy

Masters in Data Analytics for Government:

2 year part-time MSc in Data Science and Statistics for government analysts, delivered by UCL, Oxford Brookes and University of Southampton. Expanding in 2020/21

Elund Morgan, Minister for Welsh Language and Lifelong Learning and Alexis Fernquest, ONS Data Analytics Apprentice, Nov 2018

Attracting talent tomorrow, Accessing skills today

University research programmes and PhD funding:

Research partnerships with over a dozen universities and the Alan Turing Institute, which additionally hosts the *Data Science for Public Good* joint PhD programme

Centres for Doctoral Training:

7 UKRI, EPSRC and STFC centres for Doctoral Training in AI, Data Science, Statistics and Physics, with PhD co-supervision and Campus placements

Graduate Student Placement Programme:

Three-month MSc thesis and MSc/PhD industry placements at the Campus

Welsh Data Science Graduate Programme:

2 year MSc programme with Welsh Contact Centre Forum that includes three eight month industry placements.

STEM Outreach:

Girls into STEM (Science, Technology, Engineering & Maths) programme for GCSE students; Nuffield Research Placements for A-level students
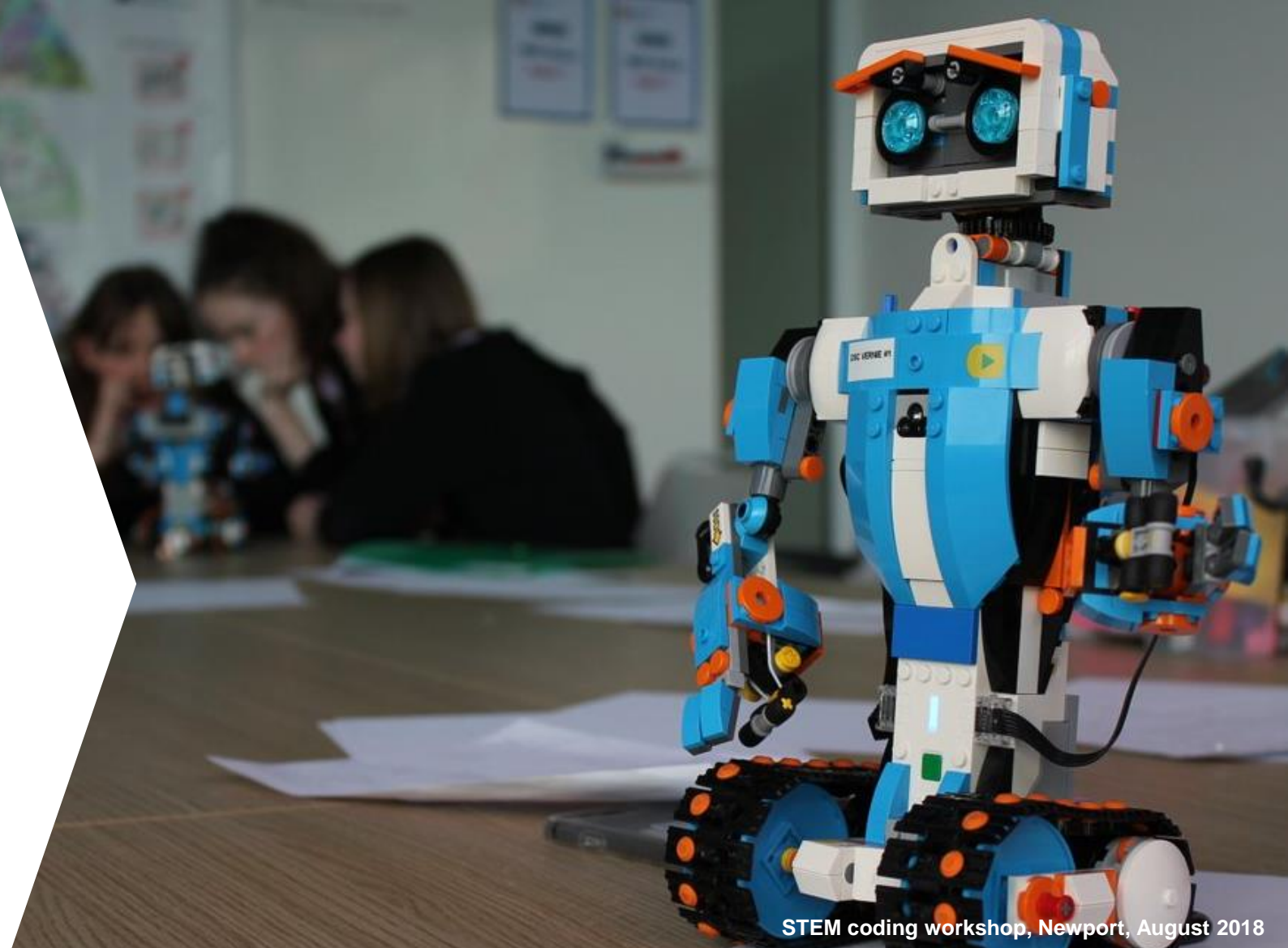
STEM coding workshop, Newport, August 2018

Developing the next generation of Data Scientists

New light on old problems
Novel data sources and official statistics

Tom Smith, @_datasmith
Director, ONS Data Science Campus

Data Science Campus

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