Leveraging Data Science for Public Good

Deraj Wilson-Aggarwal
Senior Data Scientist
Reproducible Data Science and Analysis (RDSA)

01 November 2023

Office for National Statistics

Agenda

- Introduction to RDSA
- Considerations for Technical Solutions
- Introduction to our Development Framework

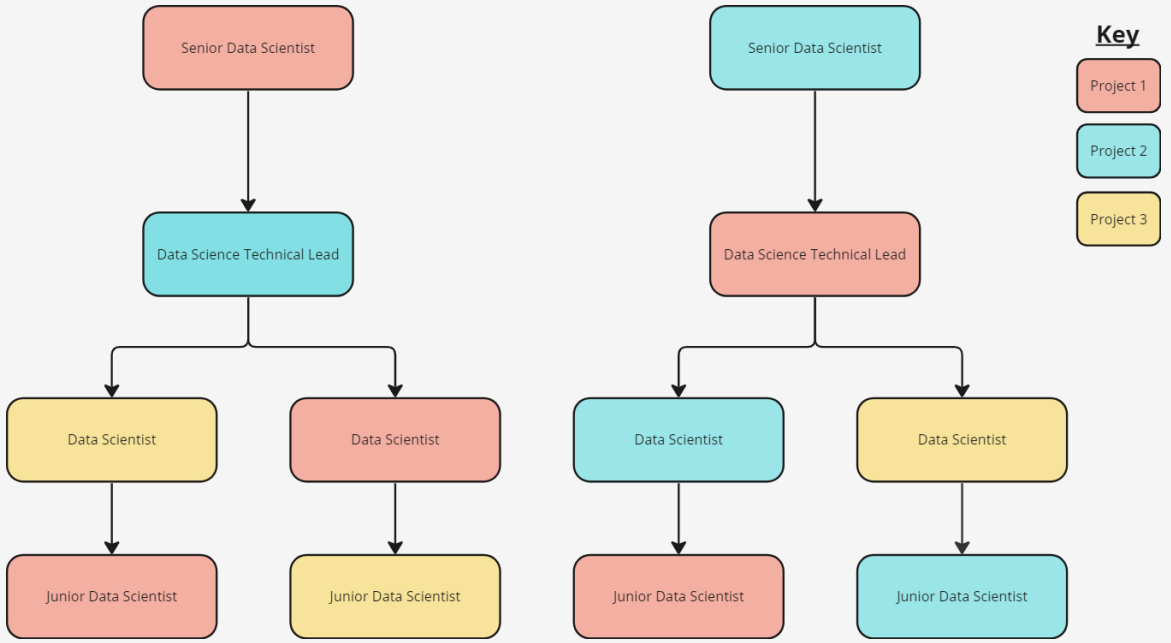
Introduction to RDSA

Who we are and what we do.

Office for National Statistics

Reproducible Data Science and Analysis (RDSA)

- RDSA is a sub-division of roughly 40 people split into 6 teams
- Sit within the Economic, Social and Environmental statistics Group (ESEG)
- Comprised of Data Scientists/Engineers from diverse academic and subject matter backgrounds
- Utilise a matrix-management system to reduce single points of failure



Our Role in Statistical Production

- Responsible for systems development and support across Economic Statistics.

- Build, maintain new systems, and transform older systems – developing **Reproducible Analytical Pipelines (RAPs).**

- Utilise data science and engineering skills to develop statistical pipelines to produce the next generation of economic statistics.
  - Implementation of cutting-edge methodologies
  - Pioneering the use of novel data sources

Reproducible Analytical Pipelines (RAPs)

- UK Government developed best practice principles for analytical systems

- Guidelines aim to:
  - Improve the quality of the analysis
  - Increase trust in the analysis
  - Create a more efficient process
  - Improve business continuity and knowledge management

https://analysisfunction.civilservice.gov.uk/policy-store/reproducible-analytical-pipelines-strategy/

Our Expertise

- Provide a centre of expertise in the use of our in-house data analysis platform
- Utilise Python, R, Spark and bash technologies to build solutions
  - Supporting ONS Capability and Training
  - Advising on best practices
- Design solutions on Cloud based platforms, and Inform Cloud First Strategy in ONS
- Implement Optimisation of models used in the production of economic statistics.

Considerations for Technical Solutions

The Right Tools: Open-Source

Harmonise tool usage across organisation

Facilitate Sharing Work – preventing duplicated efforts

Moving away from legacy-paid for systems - cost benefits

Utilise pre-existing communities for support

Facilitate community and external contributions

Choosing Appropriate Tools

- Size of data
  - Leveraging distributed computing
- Existing capability
  - Team capability
- Tool capability
  - Integrations with existing infrastructure
  - Breadth of available tool packages

Version Control

- Facilitate sharing of code
- Promote peer reviewing of code
- Adherence to organisational standardisation practices
- Guarantee an audit control and traceability
- Ability to reproduce results using a labelled iterations of codebase

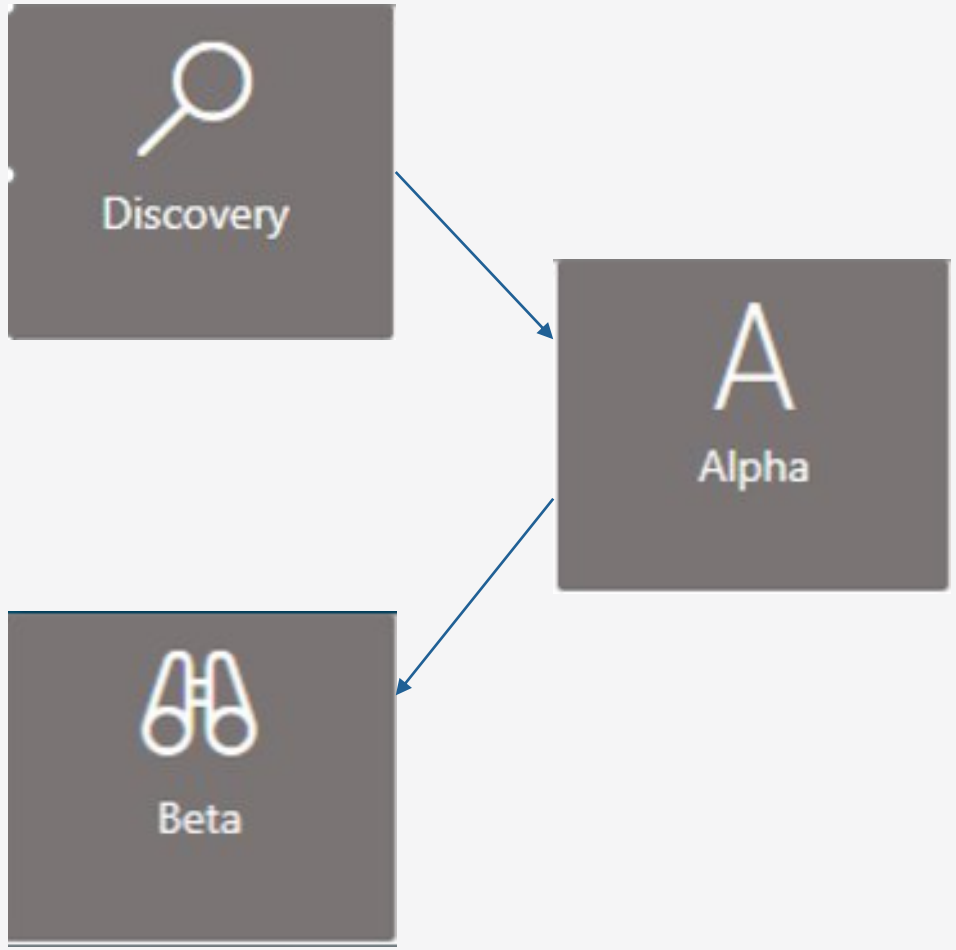
Further Considerations

- Unit Testing components/modules
- Logging systems for reproducibility and debugging
- Continuous Integration / Continuous Development (CI/CD)
- Cloud First / Cloud Agnostic Strategy
  - Scalability
  - Managed costs
  - Latest infrastructure and systems

Development Framework

Development Framework for transformation

- Discovery
  - Understand business requirements
  - Define Test Assurance Plan
- Alpha
  - Developers build system to requirements
  - Tests conducted (Unit Tests, Functional Tests)
  - Review build for compliance with Architecture
- Beta
  - Business run the system
  - Validate Acceptance criteria against real data



Discussion

Example Projects

Demonstrating the role Data Science can have in delivering statistics for public good

Office for National Statistics

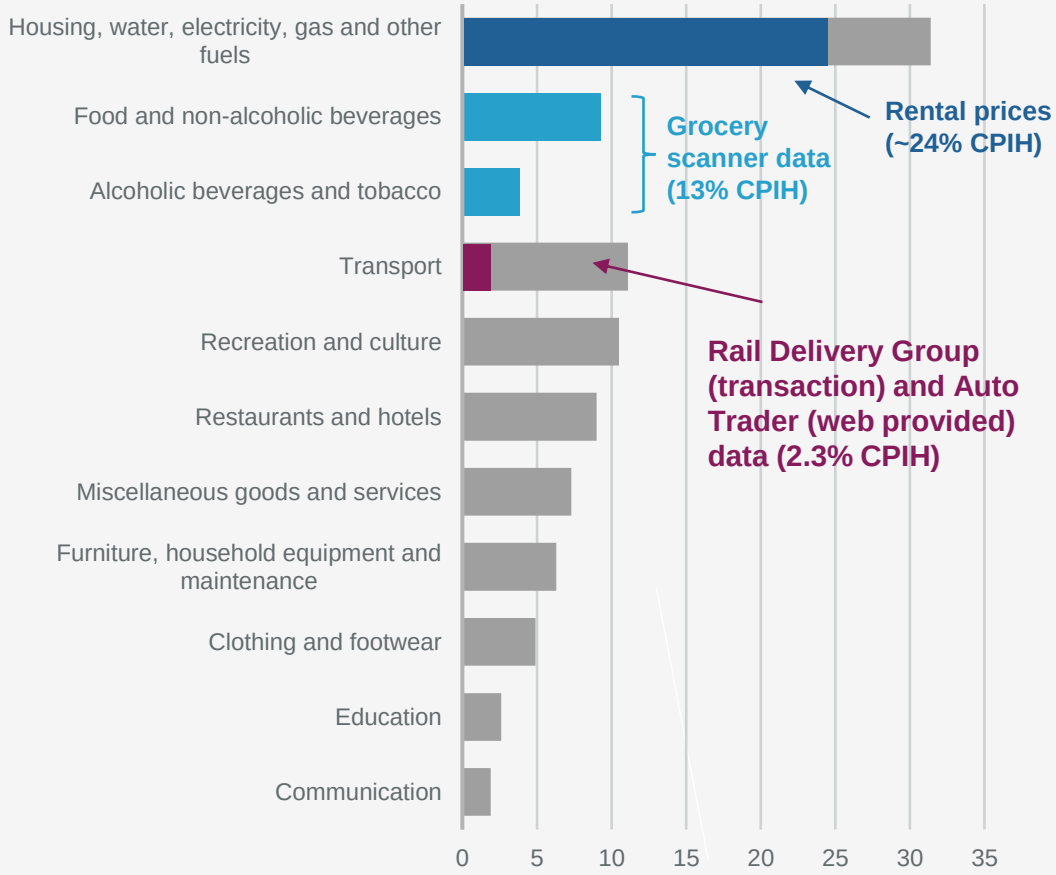
Transforming UK Consumer Price Statistics

Continuous programme of improvements for consumer price statistics over several years beginning with rail fares

Aims:

- Obtaining **robust sources of alternative data** (scanner/web-scraped data)
- **Researching methodologies** to most effectively incorporate the data
- Developing **statistical systems** for existing and new data and methods
- Embedding new **systems** and **processes**

Primarily, new data will help us to **inform the narrative** around what is driving inflation for our users



https://unece.org/sites/default/files/2023-05/7.4%20UK_un_systems_railfares_paper.pdf

Nowcasting: Signature Method

Collaboration with the Alan Turing Institute to produce an open-source tool for Signature Methods of Nowcasting

- Method handles samples of **irregular frequency** that cause missing values

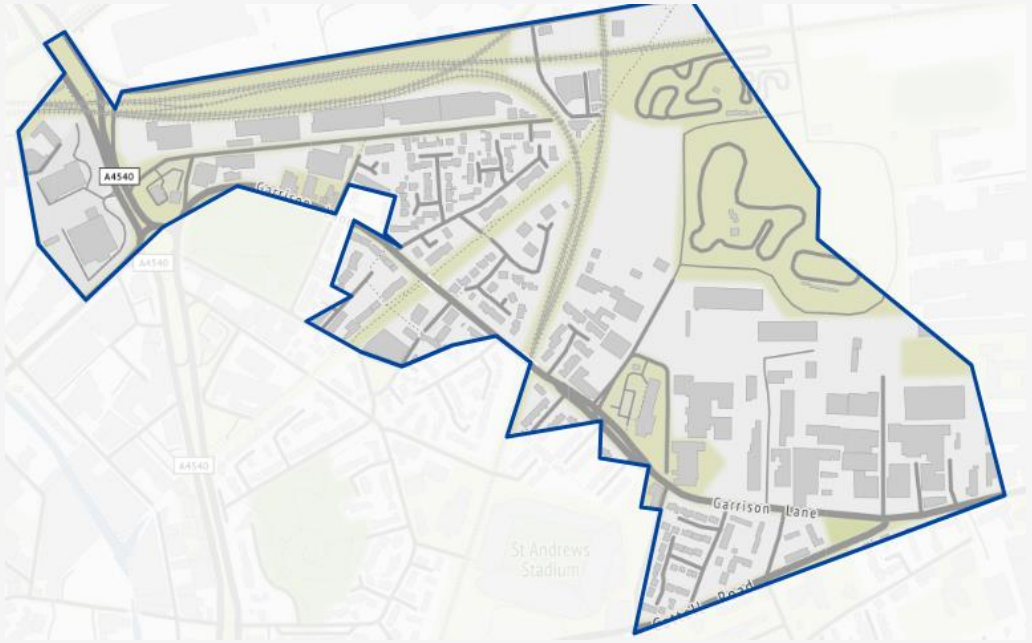
- Method can facilitate reliable estimates of delayed economic indicators

- https://github.com/datasciencecampus/SigNow_ONS_Turing
- https://arxiv.org/pdf/2305.10256.pdf

Regional Gross Value Added

- ONS now produce GVA data for **more detailed geographies**
  - Policymakers can **better track GVA** for areas affected by an intervention
  - More accurate monitoring and evaluation

- Targeted local investment and local infrastructure projects



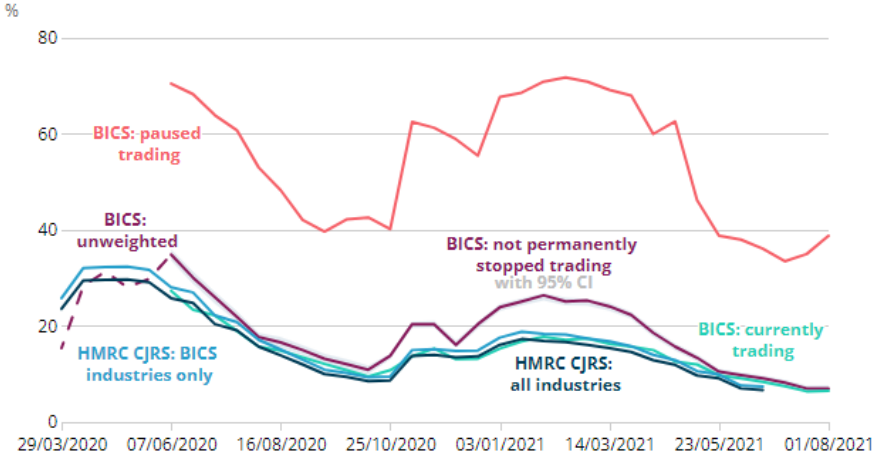
https://www.ons.gov.uk/releases/disaggregatingukannualsubnationalgrossvalueaddedgvatolowerlevelsofgeography1998to2020

Business Insights and Conditions Survey (BICS)

- Voluntary fortnightly survey to facilitate COVID-19 response
- Evolved to provide info on variety of economic factors
- Facilitate rapid data-driven decision
- Support evaluation and monitoring of policy

Figure 1: Comparison between CJRS statistics and BICS estimates on proportions of the workforce on furlough

Proportion of workforce on furlough, broken down and averaged by BICS wave, UK, 23 March 2020 to 8 August 2021



Source: Her Majesty's Revenue and Customs – Coronavirus Job Retention Scheme statistics and Office for National Statistics – Business Insights and Conditions Survey

https://www.ons.gov.uk/businessindustryandtrade/business/businessservices/articles/comparisonoffurloughedjobsdata/march2020tojune2021