Statistics on jobs, businesses and people – where data science is adding value

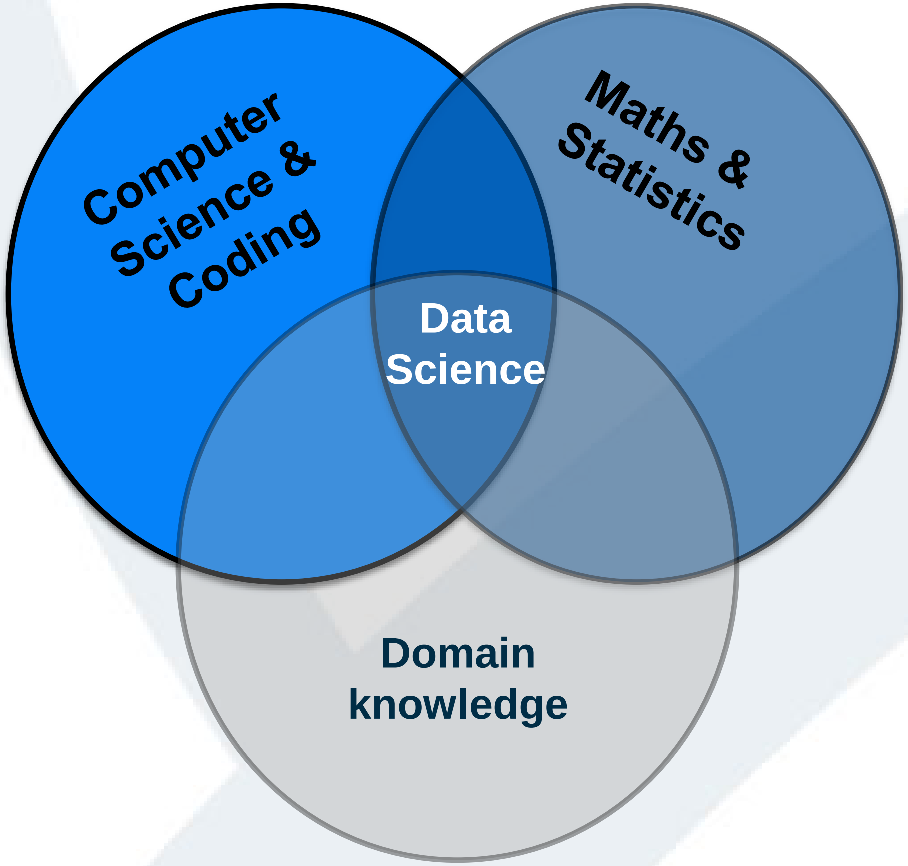
Karen Gask
Office for National Statistics
@GaskyK

Outline

- Introduction and scene setting
- Projects:
    - Online job vacancies
    - Address index for address matching
    - Plus a couple of others

Introduction

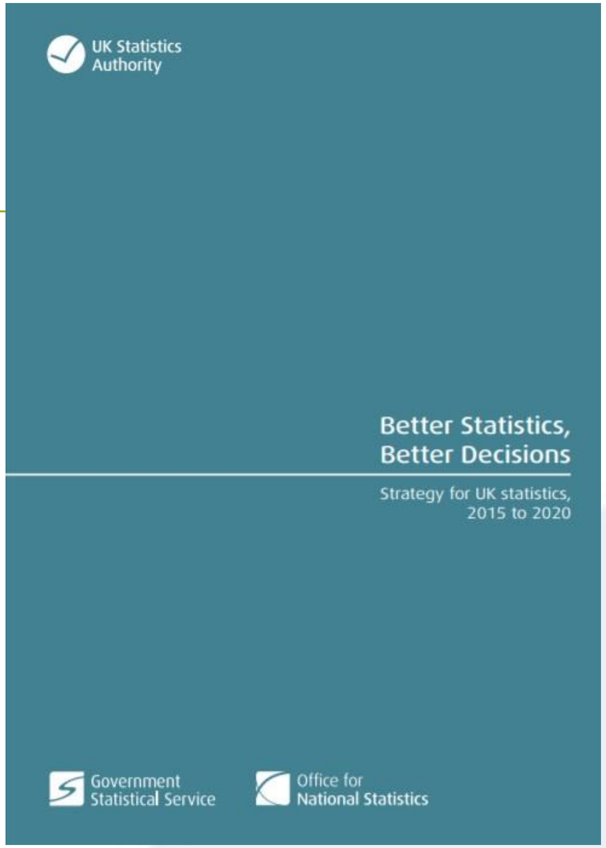
- There is a lot of data science hype!
- Data science is about applying best data, tools and techniques to a problem
- Projects undertaken at ONS illustrate where data science can add value



The Data Science Venn Diagram, modified from Drew Conway

Better Statistics, Better Decisions

Strategy for UK Statistics, 2015-2020


UK Statistics Authority

Better Statistics, Better Decisions

Strategy for UK statistics, 2015 to 2020

Government Statistical Service

Office for National Statistics

- Five perspectives:
  - Helpful
  - Professional
  - Innovative
  - Efficient
  - Capable
- Discusses building capability for exploiting richer and more complex data sources
- Keeping pace with advances in technology

ONLINE JOB VACANCIES

Potential of online job vacancy (OJV) data

- Currently official job vacancy data comes from a survey, but this has limitations

- Can online data replace / enrich the official survey-based data?

	Current Official Estimates (Survey)	Online data
Frequency	Monthly (Rolling Qtr)	Real-time?
Industry Sector	✓	✓
Enterprise Size	✓	✓
Job type / skills	×	✓
Geography	×	✓
National Totals	✓	×

More frequent
More timely
More granular
Less burden
Cheaper???

Status

- First collaborative project with other European National Statistics Institutes ran from February 2016 to May 2018
- Second collaborative project will start in Q4 2018 for 2 years
- ONS has acquired two datasets for free:
  - Adzuna data – through partnership, access for specific purposes approved by Adzuna
  - Burning glass data – through partnership with Nesta

Main challenges

- Not all jobs are advertised online. Coverage is incomplete and not representative
- There is no definitive source of OJV data
- Much OJV data is unstructured
- Live job adverts aren’t the same as the official definition of a job vacancy:
  - Job adverts might be for vacancies abroad
  - Job adverts might stay online after the vacancy is filled
  - Ghost vacancies – agencies sometimes post adverts solely to get CVs

Job vacancy jungle

- JVS
- Accessed data (Cedefop, CEB, Burning Glass, Adzuna)
- Web scraped data (7 portals)

What would “implementation” look like?

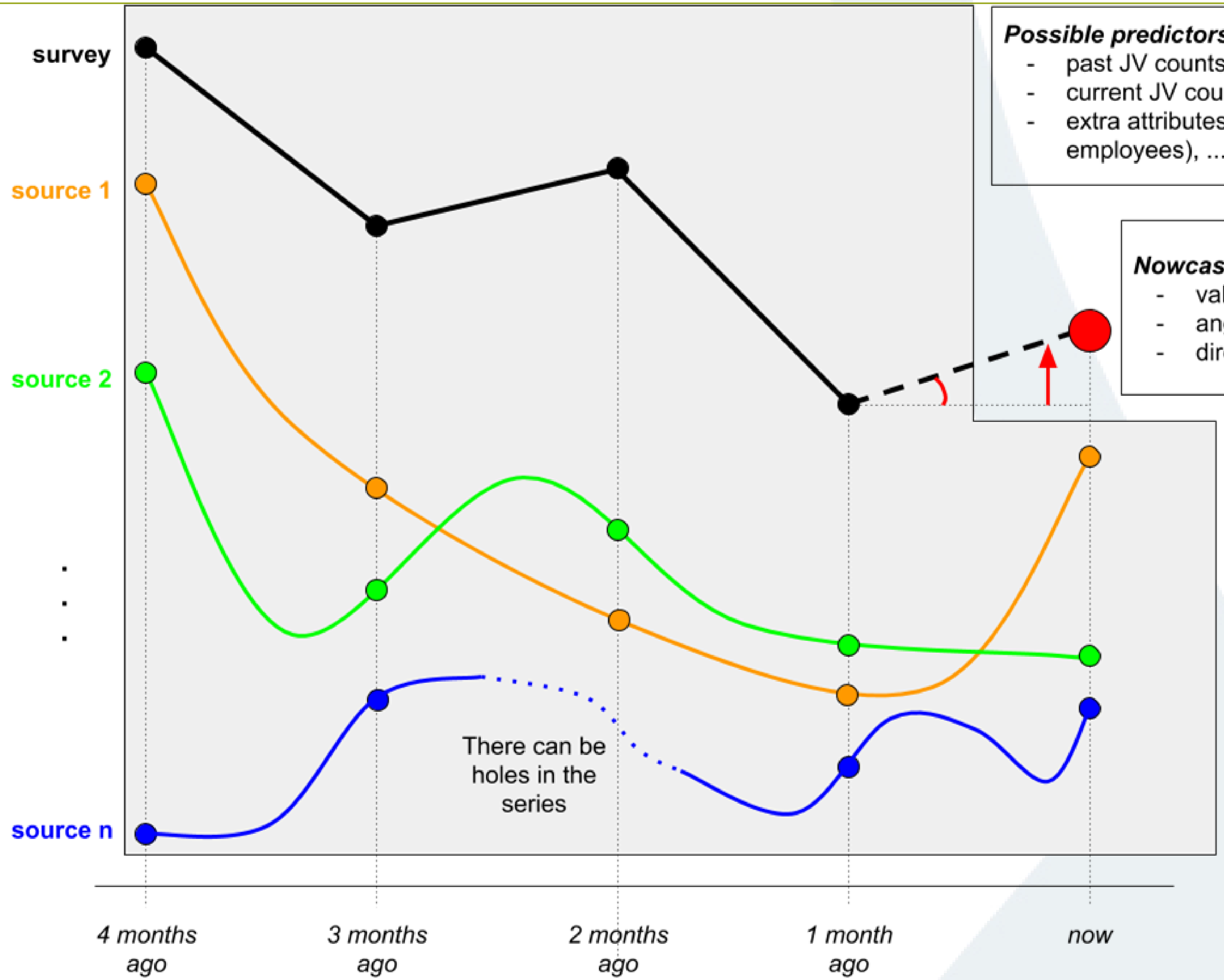
Options	Assessment
1. OJV data replaces the JVS	• Not feasible
2. Integration of OJV with JVS	• Not feasible (or at least extremely difficult)
3. Reduce frequency of JVS and use OJV data to produce modelled estimates	• Possibly feasible but needs investigation • Implies major change to business processes • Business benefits not clear
4. Produce new statistics of on-line vacancies/job ads to complement existing statistics	• Feasible • No change to JVS processes • Focus how statistics would be presented
5. OJV data to now-cast estimates	• Feasible

OJV = online job vacancies

JVS = job vacancy survey

Basic nowcasting idea

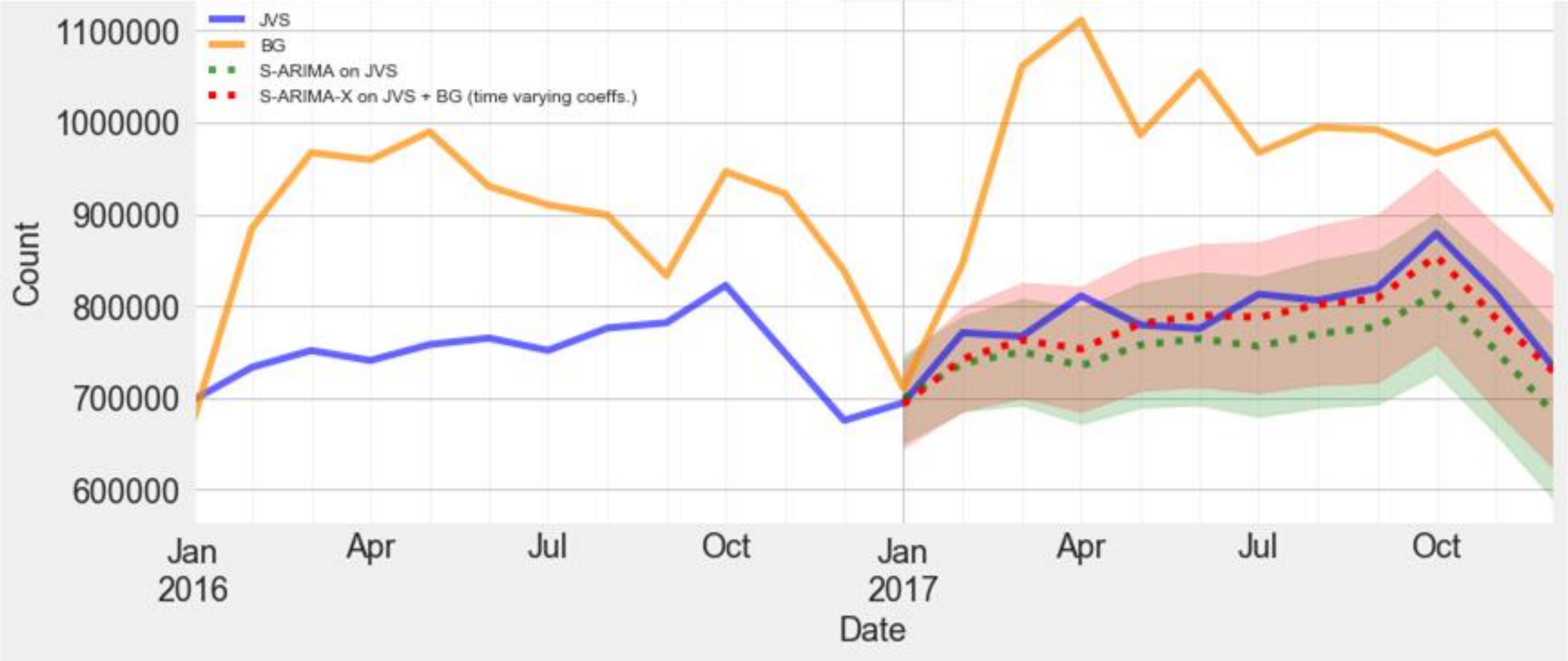
(given that survey data takes 11 weeks to be published)



Possible predictors:
- past JV counts
- current JV counts for online sources
- extra attributes (industry, no. of employees), ...

Nowcasting:
- value (regression)
- angle (bounded regression)
- direction (classification)

Nowcasting the official data using OJV data



Nowcasting job vacancies for individual companies

- Several methods used to nowcast vacancies for individual companies:
  - Linear regression
  - Classification (does the trend go up or down?)
  - Neural networks
- Data not suitable at individual company level
  - Too many gaps
  - High oscillation

Example of nowcasting for a particular company

Key Conclusions (and Questions)

- Agreed access arrangements are generally better than direct web scraping
- OJV data cannot replace the job vacancy survey
- OJV data does not correspond to target concepts and only measures part of the labour market. How useful are these measures?
- If useful, how should these measures be presented alongside the official estimates? As experimental statistics?
- How do we get the best possible quality data for official statistics purposes?

ADDRESS MATCHING

Why are addresses so important to ONS?

- A complete list of addresses is critical to many parts of ONS, from ensuring everyone receives a Census form to accurate geo-referencing
- Addresses are complicated, which makes matching really hard
- We have developed an address index service that matches an input address string to a validated address and Unique Property Reference Number (UPRN) from Address Base

Description of the problem

Input address string

- Unstructured text
- Messy, containing typos etc.
- Incomplete
- Range from historic to very recent addresses, including businesses

Reference data to match against (AddressBase)

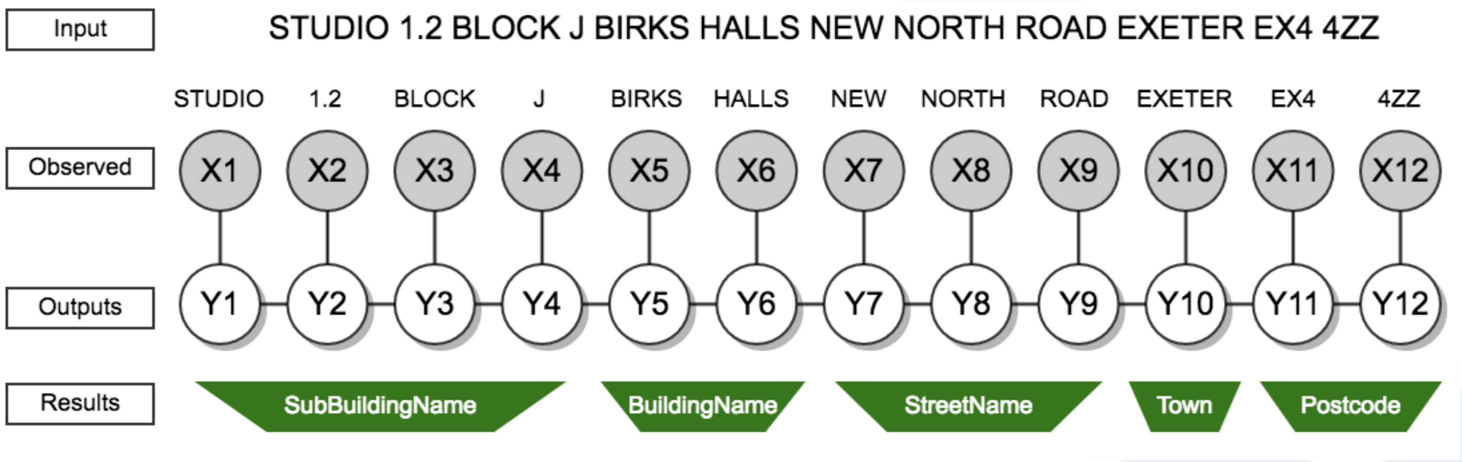
- Structured, tokenised
- Complete & correct (more or less)
- Snapshot of addresses at a given time
- Organisation / business names are not always part of the address

Methodology

- Control over the input is limited, hence one needs to parse the input string to tokens like building name and number, street name, town and postcode before linking can take place
- Rules based
  - Could use regular expressions or look-ups (for town names for example)
  - But addresses are very complex (eg. St Pauls St)
- Structured learning
  - Addresses are semi-structured (from small building to street then town)

Conditional Random Fields

- Uses features to calculate likely sequence of words (eg. number, street, town, postcode)



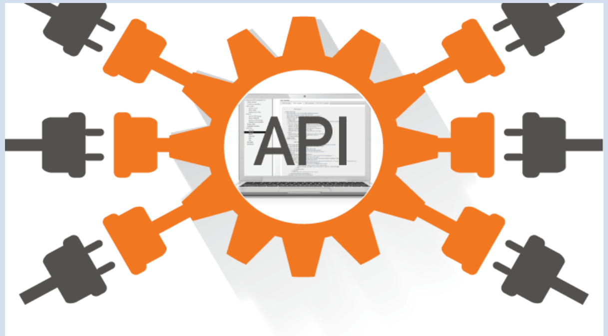
- F1-score: 0.992 (99.2% of tokens correct)
- Sequence accuracy: 0.975 (97.5% correct)

Matching

- We then compare the parsed input with the reference data
- Elastic search is used for this comparison:
  - Provides fast, scalable and reliable enterprise indexing and search technology
  - The reference data are indexed so that they could be retrieved quickly based on requested criteria
  - Similarity score is calculated between the query and all AddressBase records
- Finally a confidence score is calculated for the result UPRN to inform users about the estimated match quality

Address Index services (now in ‘public beta’)

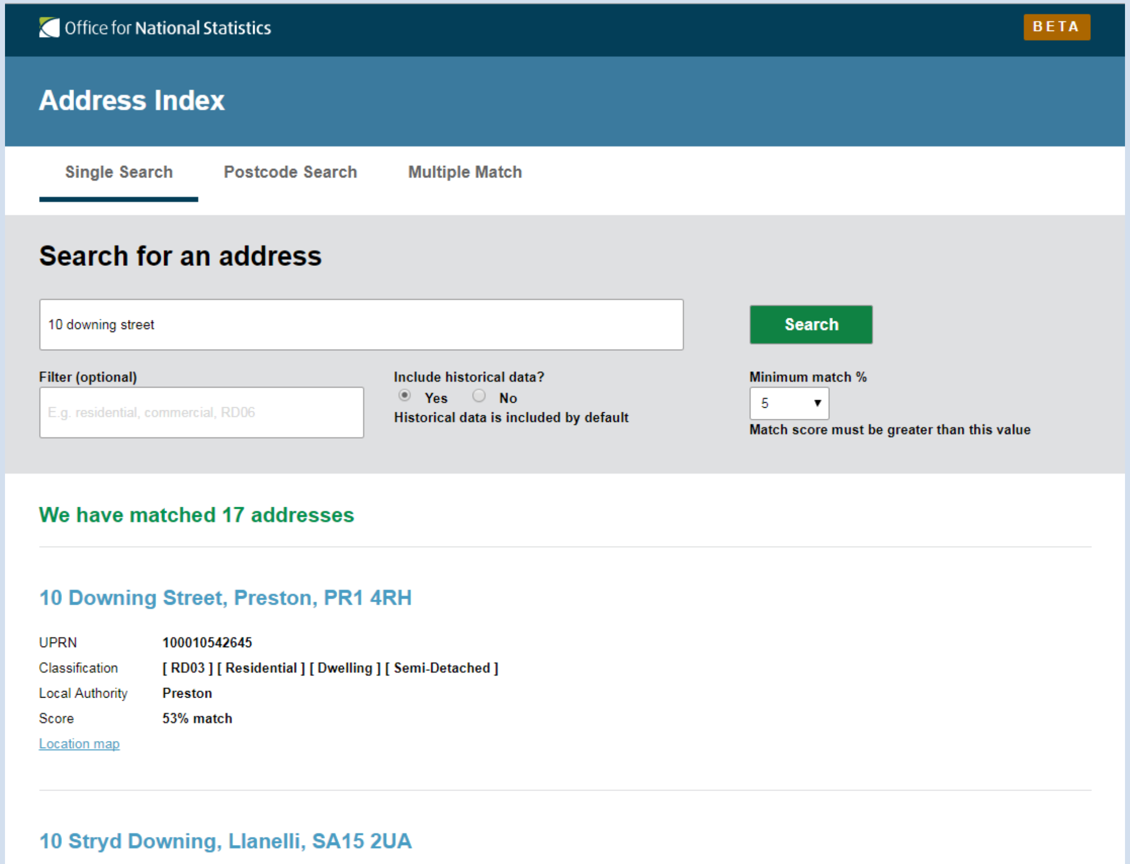
RESTful API



Bulk matching service

MOVEit FILE TRANSFER

User web interface



Next steps

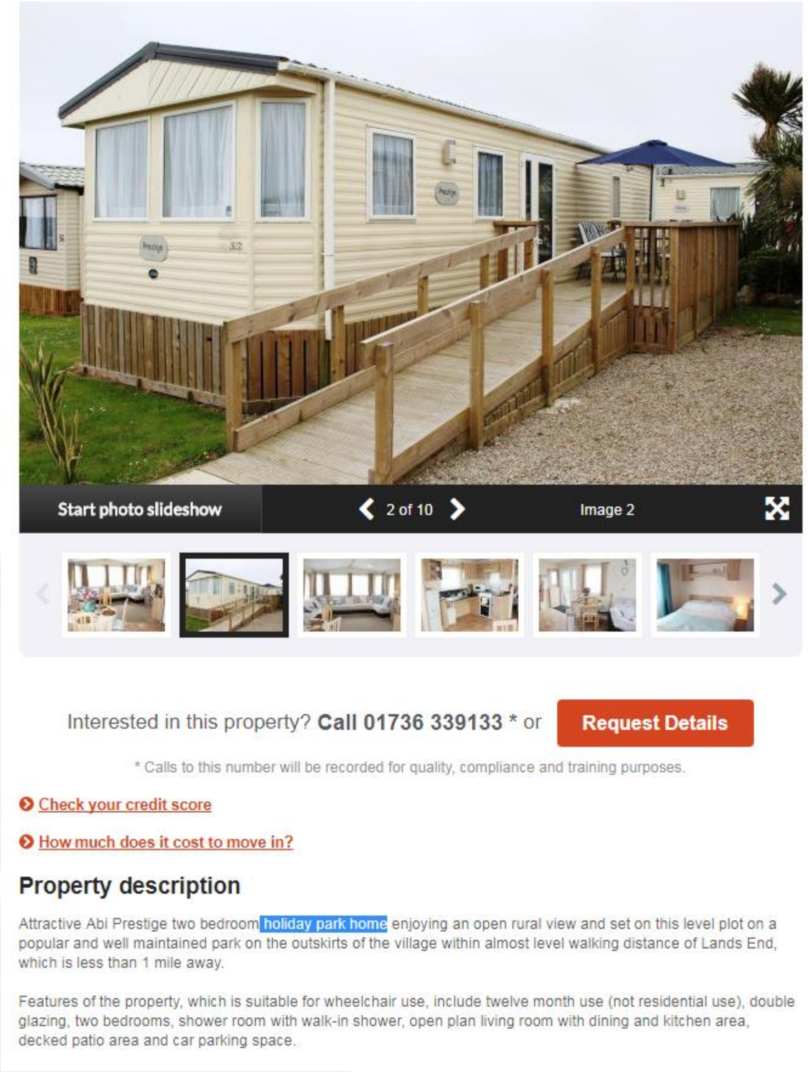
- ‘Time machine’ – to allow search in a specific (historic) time point
- Extending services offered based on arising customer needs (e.g. postcode search)
- Scaling up to make it available beyond beta users
- Continual improvements in matching quality and confidence score accuracy

Another example – automated survey coding

- Crime Survey for England and Wales asks respondents whether they have been a victim of crime
- Free text and closed question answers are all manually coded to a crime type
- Natural language processing and machine learning used
- If the model predicts one of 10 offence codes (around 40% of the total) where accuracy is >=97%, code automatically
- Saving an estimated £3,700 per year and freeing up statisticians for more complex coding tasks

Another example – using Zoopla to find caravan homes

- Accurate address list essential for 2021 Census but caravan homes recorded inconsistently in different datasets
- Obtained some data from Zoopla website
- Used natural language processing and machine learning on property description
- Accuracy of model above 90%
  - But difficulty separating residential / holiday caravans due to limited training examples
- Will feed into model of where to send thousands of census officers in 2021



Summary

- Shown examples of applying best data, tools and techniques to problems
  - Online job vacancy data
  - Address matching
- Data science beginning to be used in our processes and outputs
- Not a silver bullet – evolution not revolution!

QUESTIONS?