



Office for  
National Statistics

# Better Scraping, Better Statistics?

## Using web-scraped data in statistical outputs

Matt Greenaway, ONS Big Data team  
GSS Conference, November 2017  
[matthew.greenaway@ons.gsi.gov.uk](mailto:matthew.greenaway@ons.gsi.gov.uk)





# Outline

---

- Web-scraping definition
- ONS web-scraping policy
  - Rationale
  - Advice
  - Policy
- Web-scraping applications
  - E-commerce
  - Job vacancies
- Lessons Learned





# Web-scraping definition

- Collection of data automatically from the internet
- For example – data on food prices from supermarket websites
- Tesco's website as it appears in your browser:

 <p>Hovis Soft White Medium Bread 800G <a href="#">Rest of shelf &gt;</a></p> <p><b>£ 1.05</b> £0.13/100g</p> <div><input type="text" value="1"/> <input type="button" value="Add"/></div>	 <p>Hovis Soft White Thick Bread 800G <a href="#">Rest of shelf &gt;</a></p> <p><b>£ 1.05</b> £0.13/100g</p> <div><input type="text" value="1"/> <input type="button" value="Add"/></div>	 <p>Tesco White Toastie Thick Bread 800G <a href="#">Rest of shelf &gt;</a></p> <p><b>£ 0.50</b> £0.06/100g</p> <div><input type="text" value="1"/> <input type="button" value="Add"/></div>	 <p>Warburtons Toastie Sliced White Bread 800G <a href="#">Rest of shelf &gt;</a></p> <p><b>£ 1.05</b> £0.13/100g</p> <div><input type="text" value="1"/> <input type="button" value="Add"/></div>
---	--	--	---

# Web-scraping definition

- Collection of data automatically from the internet
- For example – data on food prices from supermarket websites
- Tesco's website as it appears in your browser:

			
Hovis Soft White Medium Bread 800G	Hovis Soft White Thick Bread 800G	Tesco White Toastie Thick Bread 800G	Warburtons Toastie Sliced White Bread 800G
<a href="#">Rest of shelf &gt;</a>	<a href="#">Rest of shelf &gt;</a>	<a href="#">Rest of shelf &gt;</a>	<a href="#">Rest of shelf &gt;</a>
<b>£ 1.05</b> £0.13/100g	<b>£ 1.05</b> £0.13/100g	<b>£ 0.50</b> £0.06/100g	<b>£ 1.05</b> £0.13/100g
<input type="text" value="1"/> <input type="button" value="Add"/>	<input type="text" value="1"/> <input type="button" value="Add"/>	<input type="text" value="1"/> <input type="button" value="Add"/>	<input type="text" value="1"/> <input type="button" value="Add"/>

# Web-scraping definition

- Collection of data automatically from the internet
- For example – data on food prices from supermarket websites
- Tesco's website as HTML:

```
tile--wrapper data-reactid= 48/ ><div class= product-tile data-reactid= 488 ><div
class="flexi-tile" data-reactid="489"><div class="tile-content" id="256174499" data-
reactid="490"><a href="/groceries/en-GB/products/256174499" aria-hidden="true" class="product-
image-wrapper" tabindex="-1" target="_self" data-reactid="491"></a><div class="product-details--wrapper" data-reactid="493"><div
class="product-details--content" data-reactid="494"><a href="/groceries/en-
GB/products/256174499" class="product-tile--title product-tile--browsable" data-
reactid="495">Hovis Soft White Medium Bread 800G</a><!-- react-empty: 496 --><a
href="/groceries/en-GB/shop/bakery/bread-and-balls/white-bread" aria-label="Rest of
shelf White Bread" class="browse-category icon-chevron_right-link" data-reactid="497">Rest of
shelf</a><!-- react-empty: 498 --><!-- react-empty: 499 --></div></div><div class="product-
controls--wrapper" data-reactid="500"><form action="/groceries/en-GB/trolley/items/256174499?
_method=PUT" method="POST" data-reactid="501"><input type="hidden" name="_csrf"
value="URbZiVva-CVPWI7XIXj0APECxj7WNznCxyls" data-reactid="502"/><input type="submit"
class="hidden" data-reactid="503"/><input type="hidden" name="id" value="256174499" data-
reactid="504"/><input type="hidden" name="anchorId" data-reactid="505"/><input type="hidden"
name="returnUrl" value="/groceries/en-GB/search?query=white%20bread" data-reactid="506"/><input
type="hidden" name="backToUrl" value="#" data-reactid="507"/><input type="hidden"
name="oldValue" value="0" data-reactid="508"/><input type="hidden" name="oldUnitChoice"
value="pcs" data-reactid="509"/><input type="hidden" name="catchWeight" data-reactid="510"/>
<input type="hidden" name="adjustment" value="true" data-reactid="511"/><input type="hidden"
name="newUnitChoice" value="pcs" data-reactid="512"/><div class="controls" data-reactid="513">
<div class="price-details--wrapper" data-reactid="514"><div class="price-control-wrapper" data-
reactid="515"><div class="price-per-sellable-unit price-per-sellable-unit--price price-per-
sellable-unit--price-per-item" data-reactid="516"><div class="" data-reactid="517"><span data-
reactid="518"><span class="currency" data-reactid="519">£</span><span class="space" data-
reactid="520"> </span><span class="value" data-reactid="521">1.05</span></span></div></div>
</div><div class="price-per-quantity-weight" data-reactid="522"><span data-reactid="523"><span
class="currency" data-reactid="524">£</span><span class="value" data-reactid="525">0.13</span>
</span><span class="weight" data-reactid="526">/100g</span></div><!-- react-empty: 527 -->
```

# Web-scraping applications at the ONS

---

## Example 1

**Data:** Food price data scraped from supermarket websites

**Use:** To produce timely measures of food-price inflation

## Example 2

**Data:** Jobs vacancy data scraped from jobs portals

**Use:** to produce timely jobs vacancy statistics and provide a richer source of labour market information

## Example 3

**Data:** Data related to second/holiday homes scraped from holiday lettings and room-sharing websites

**Use:** To help inform census & social survey design & estimation

## Example 4

**Data:** Detailed information on contracts awarded to UK and non-UK companies scraped from procurement websites

**Use:** to evaluate impact of possible changes to procurement rules

## Example 5

**Data:** Data scraped from large numbers of business websites related to whether they conduct e-commerce

**Use:** To produce research and statistics on the digital economy

# Web-scraping policy



Office for  
National Statistics

# **Why do we need a web-scraping policy?**



# Challenge - Burden

- Code of Practice for Official Statistics emphasises the importance of minimising burden in data collection activities
- Web-scraping, if done 'badly', can overload websites - like a Distributed Denial of Service attack
- As well as being an ethical issue, this also implies reputational risk – for example, ONS accidentally taking down the website of a small business



# Challenge - Consent

---

- Relevant considerations -
  - robots.txt protocol
  - Terms and Conditions
- T&Cs and robots.txt not always consistent – which do we need to check?
- If we're scraping lots of websites ('crawling'), we can't check T&Cs for all of them – OK to just rely on robots.txt?
- In what circumstances (if any) should we try and gain explicit consent?
- What about consent from data subjects (e.g. – if we are scraping personal data?)

# Challenge - Legal

---

- Potentially relevant legislation:
  - **Contract law** – pertinent to terms and conditions of websites
  - **Copyright and Rights in Databases Regulation**  
criminalises extracting or utilising all or a substantial part of a 'protected database' without consent. Databases are protected under this law if 'a substantial investment in obtaining, verifying or presenting the contents of the database' has been made.
  - The **Computer Misuse Act**, which criminalises 'unauthorised access' to computer systems
  - **Data Protection Legislation**



# Challenge - Legal

- Web-scraping law 'inchoate', relatively little case-law but small number of high-profile cases

## LinkedIn and hiQ to take web scraping fight to court

By Gurkaran Singh · Aug 2, 2017

10

## Facebook Can Use Controversial Law to Punish Spammy Startup, Court Rules

Jeff John Roberts  
Jul 12, 2016



Should companies be able to use a federal hacking law to go after those who

HOME » BUSINESS

## Ryanair can sue over 'screen scraping' Supreme Court rules



Friday, February 20, 2015

## NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Education

### Technology

## LinkedIn told it cannot stop the bots



Dave Lee  
North America technology reporter

15 August 2017 | Technology | 70



100 F. Supp. 2d 1058 (2000)

EBAY, INC., Plaintiff,  
v.  
ORDER'S EDGE, INC., Defendant.

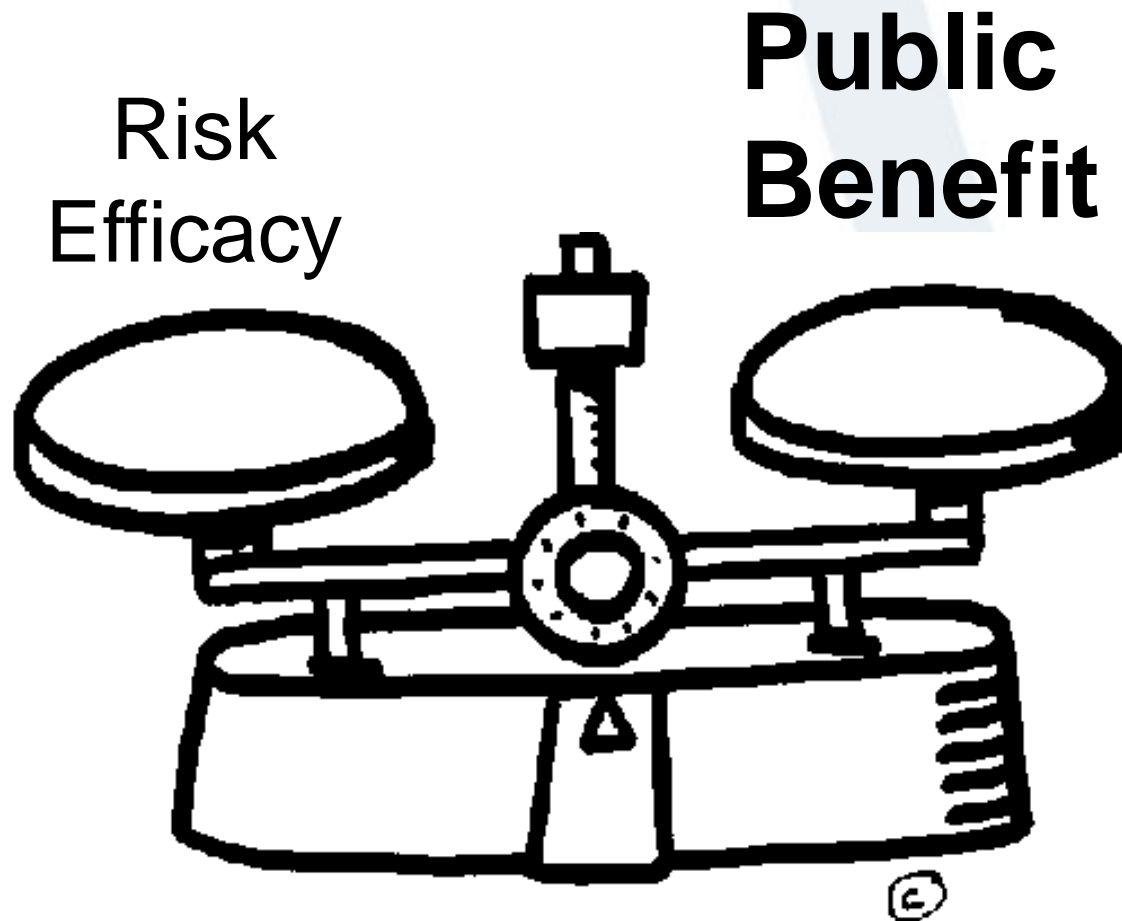
No. C-99-21200RMW.

States District Court, N.D. California.

# Advice

# Ethical Advice from NSDEC

---



# Ethical Advice from NSDEC

---

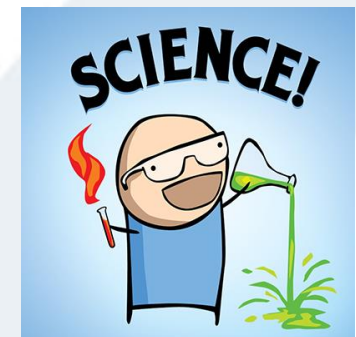
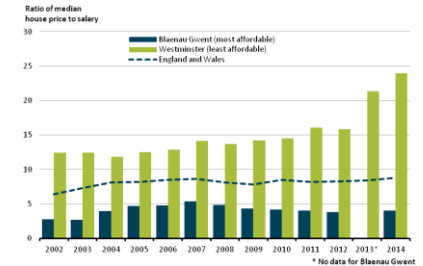
- Need a **transparent**, publically-available policy on how we will web-scrape, including how we will treat T&Cs and robots.txt
- Must always **identify ourselves**, provide a means for contact, and stop **scraping when asked to do so** by website owners
- **Limiting burden** important – should draw on best practice
- Where we are scraping **personal information** (data about people) – must always seek ethical review

# Policy



# Principals

- Use web-scraped data solely for the purpose of producing statistics, analysis and advice which has **clear benefit** for users.
- Seek to **minimise burden** on website owners
- **Honour requests** made by website owners to refrain from scraping their website
- **Protect all personal data** in all statistics and research outputs and seek ethical advice when scraping data which identifies individuals.
- Apply **scientific principles** in the production of statistics and research based on web-scraped data, and consider other sources of data.
- Abide by all applicable **legislation** and monitor the evolving legal situation



# Web-scraping applications

# Measuring ecommerce using data scraped from business websites

- ONS conduct an 'e-commerce' survey to capture whether businesses sell products or services through their own website:



*Notice is given under section 1 of the Statistics of Trade Act 1947*

## E-commerce Survey 2015

 Office for  
National Statistics

**Please do not discard this important document - your response is legally required**

16. Does this business' website have any of the following?

For each option, please ☒ either yes or no

Yes

No

On-line ordering or reservation/booking, for example using a shopping cart . . . . .



203

- Can we measure this by scraping business websites?
- Need to find business websites first (no administrative source)

# Measuring ecommerce using data scraped from business websites

---

- Steps:

1. Query search API with business name to find 'candidate' websites
2. Scrape 'candidate' websites and extract 'features' such as whether the businesses postcode and name is present on the website
3. Use supervised machine learning and extracted features to identify which of 'candidate' websites might be the business website
4. Use supervised machine learning with data scraped from business website to predict whether business engaged in e-commerce



Find business websites

The diagram consists of two light blue vertical bars with black outlines. The top bar is positioned to the right of steps 1 and 2, and the bottom bar is positioned to the right of steps 3 and 4. Both bars have a small notch on their right side, pointing towards the text labels. The background features a large, faint, light blue 'X' shape.

Detect whether engaged in ecommerce

# Measuring ecommerce data scraped from using business websites

- Identifying business websites – an example feature:

Postcode present on website?

False	806	49
True	17	143
	False	True

Correct website

Feature found

A confusion matrix for the feature 'Postcode present on website?'. The matrix is a 2x2 grid. The vertical axis is labeled 'Correct website' with 'False' at the top and 'True' at the bottom. The horizontal axis is labeled 'Feature found' with 'False' on the left and 'True' on the right. The cells contain the following counts: Top-left (False website, False feature found) is 806 and is black. Top-right (False website, True feature found) is 49 and is light gray. Bottom-left (True website, False feature found) is 17 and is white. Bottom-right (True website, True feature found) is 143 and is light gray.

# Measuring ecommerce data scraped from using business websites

---

- Can use search API and supervised machine learning to find business websites
- However – this method is typically better at finding e-commerce sites than non e-commerce sites

Proportion of websites  
which can be detected  
automatically

50%

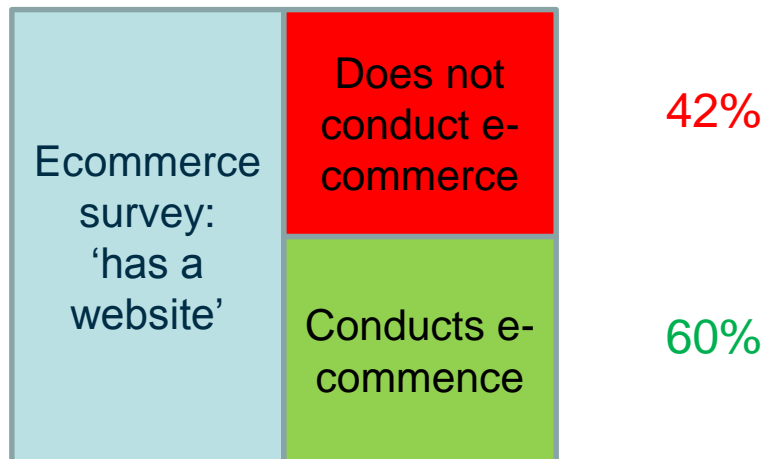
Ecommerce  
survey:  
'has a  
website'

# Measuring ecommerce data scraped from using business websites

---

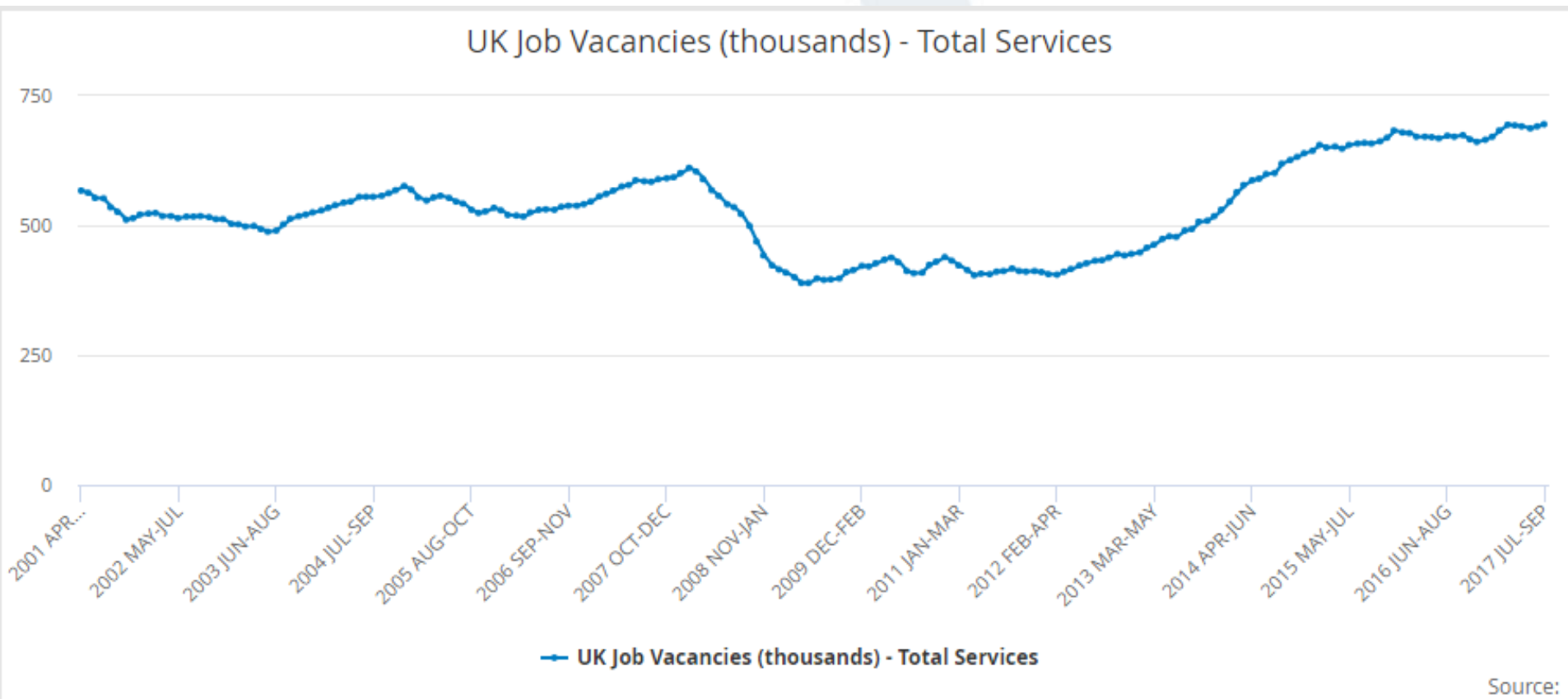
- Can use search API and supervised machine learning to find business websites
- However – this method is typically better at finding e-commerce sites than non e-commerce sites

Proportion of websites  
which can be detected  
automatically



# Measuring jobs vacancies using data scraped from business websites

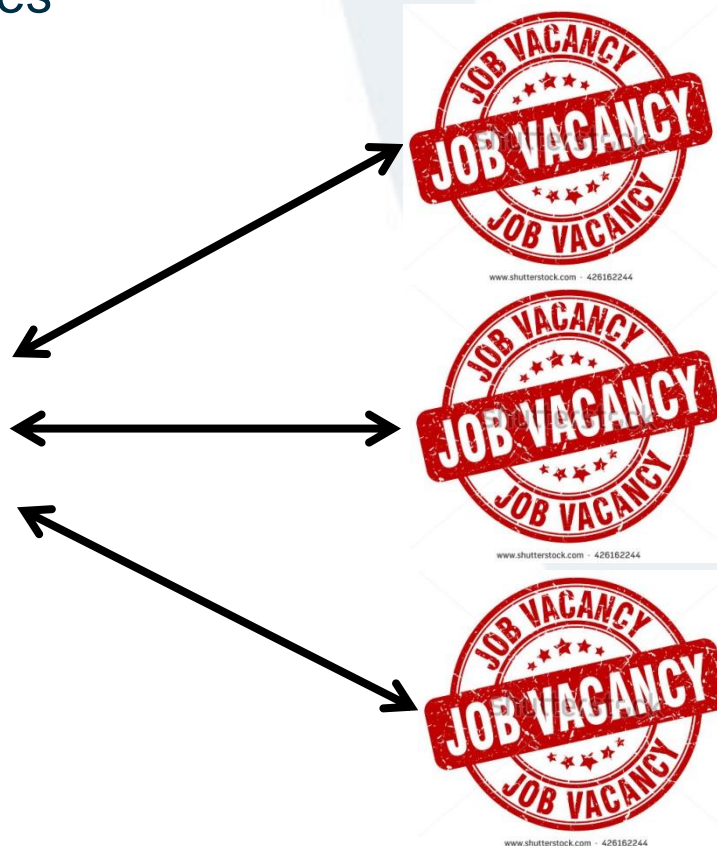
- ONS produce job vacancy statistics, partly based on a survey





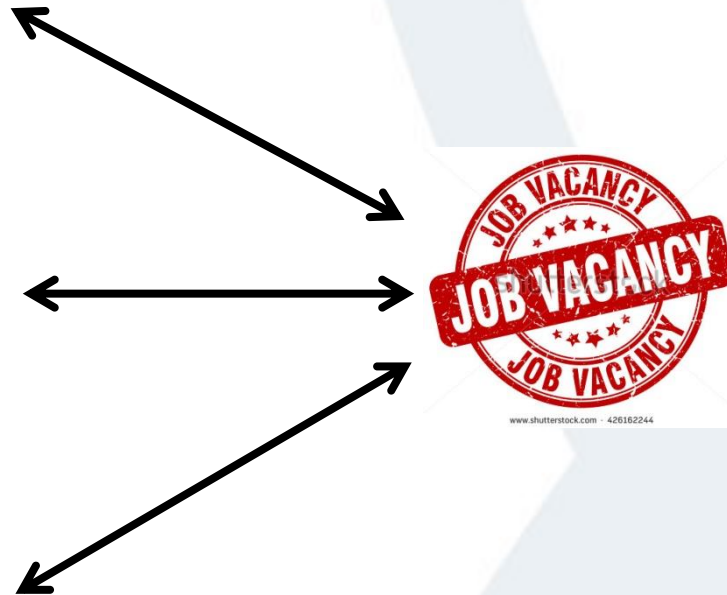
# Measuring jobs vacancies using data scraped from company websites

- Can we use data from jobs portals and business websites to improve our job vacancy statistics?
- Conceptual challenges -



# Measuring jobs vacancies using data scraped from company websites

- Can we use data from jobs portals and business websites to improve our job vacancy statistics?
- Conceptual challenges -

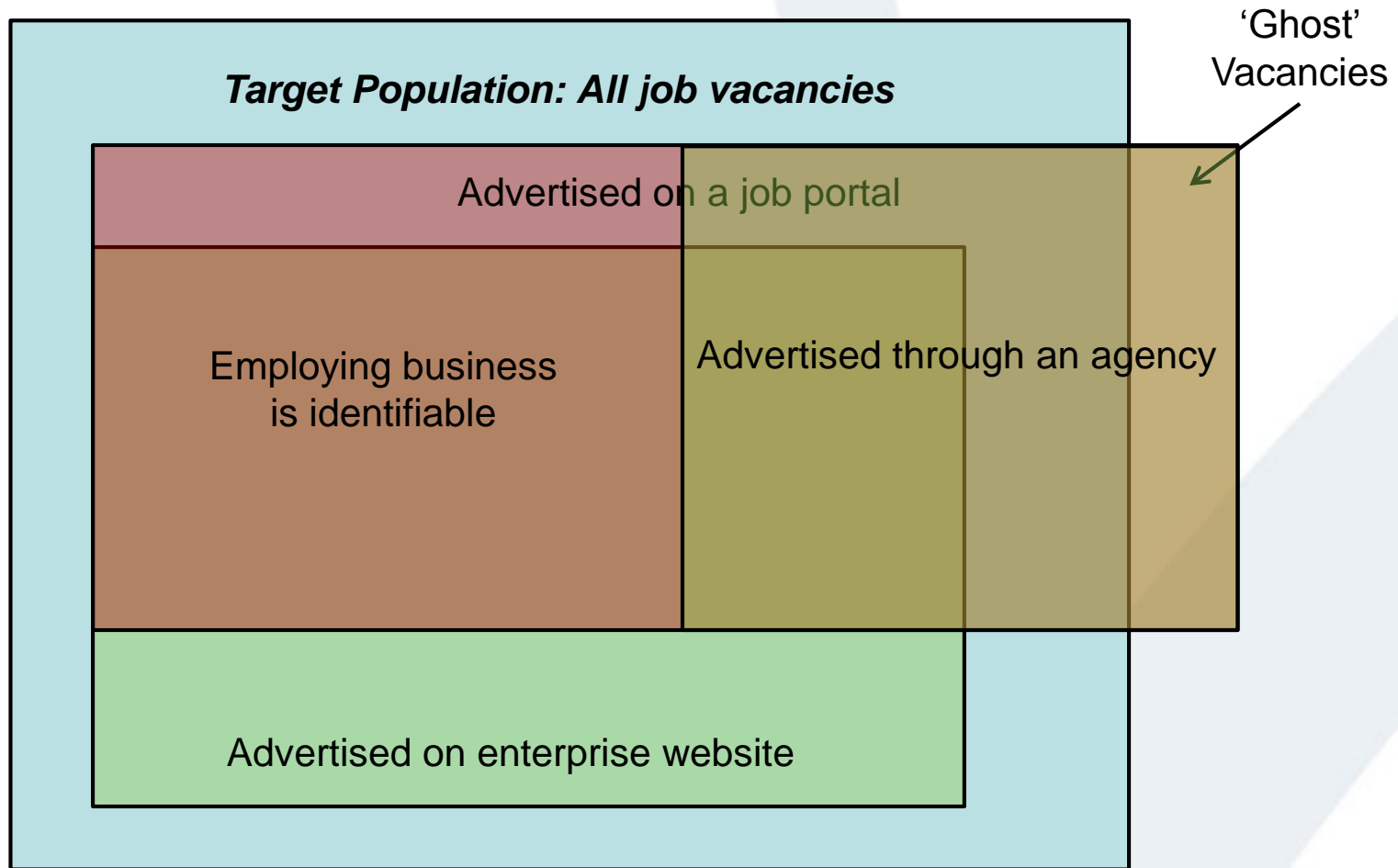


# Measuring jobs vacancies using data scraped from company websites

- Can we use data from jobs portals and business websites to improve our job vacancy statistics?
- Conceptual challenges -

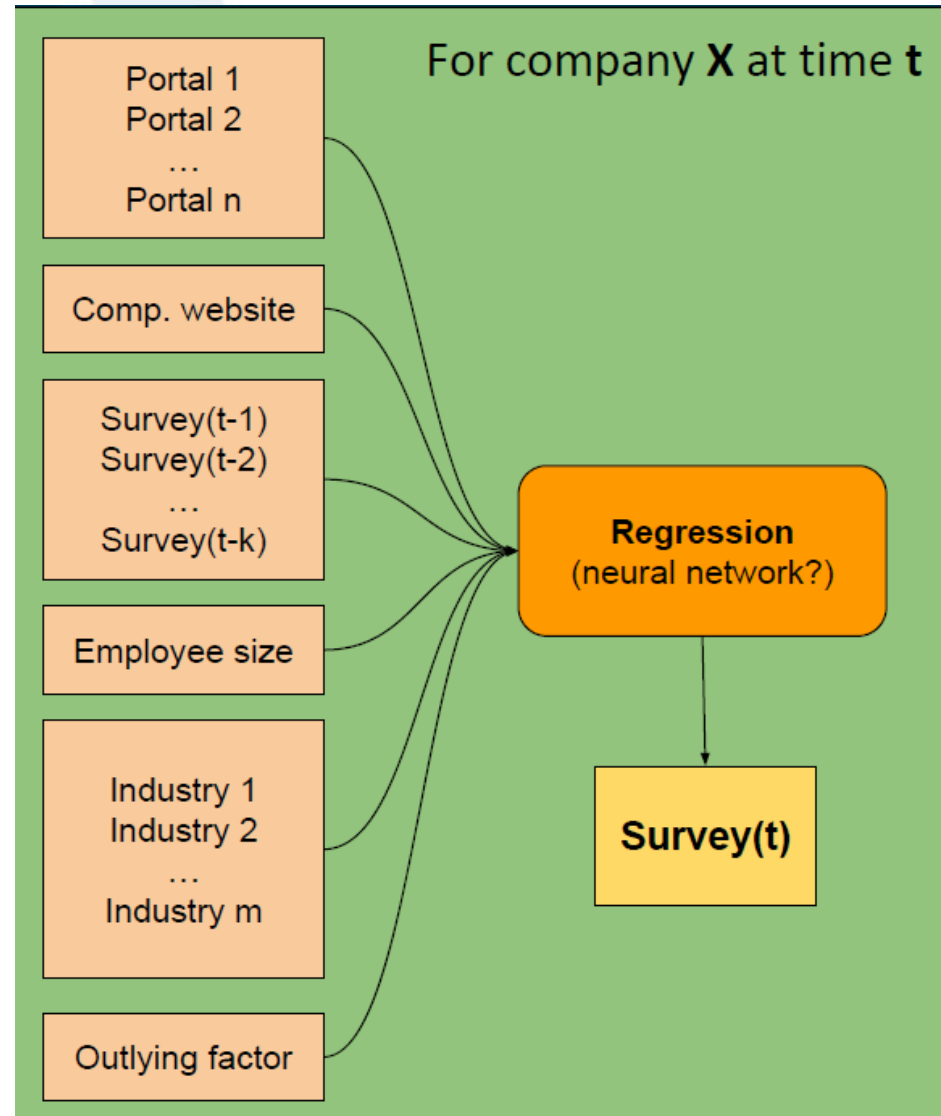


# Measuring jobs vacancies using data scraped from jobs portals



# Measuring jobs vacancies using data scraped from jobs portals

- 'Nowcasting' vacancies using survey returns and scraped data together
- Could do at company level (see diagram), industry level, overall level



# Lessons learned

---

- Drawing up web-scraping policy hard work, but worthwhile – helped bring certainty to web-scraping projects, brought much more certainty to research
- Ethical advice invaluable
- Understanding bias/coverage issues involved in using web-scraped data is not straightforward – requires significant effort separately for each use-case, often need to integrate with other data
- This also holds for data scraped by others!
- A lot of the value is in speed – potential for ‘real-time’ economic indicators
- Can also provide data that just isn’t available from elsewhere – for example, free-text descriptions

# Questions?

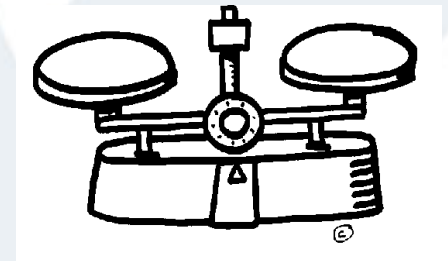
Feedback to:

[Matthew.Greenaway@ons.gsi.gov.uk](mailto:Matthew.Greenaway@ons.gsi.gov.uk)

# Legal Guidelines

---

- We will **cease scraping** whenever we are **asked to do so** by the website owner.
- We will **check the terms and conditions of websites** wherever it is practical for us to do **so**.
- Where it is not practical for us to check the terms and conditions of a website, we may scrape where we can justify that it is **ethical for us to do so** (with reference to ethical principles set out in the guidelines).





# Legal Guidelines – cont.

---

- We will carry out scraping in a manner which does **not cause financial** detriment to any website owner.
- We will abide by the Data Protection Act and other data sharing legislation.... this includes ensuring that **personal data is not revealed** in any published statistics or research.
- We will continue to monitor the legal situation as it evolves and amend our approach accordingly.



Data Protection Act 1998

