HERIOT WATT UNIVERSITY

Data in research

George Streftaris
Professor of Statistics
Actuarial Mathematics and Statistics
Heriot-Watt University

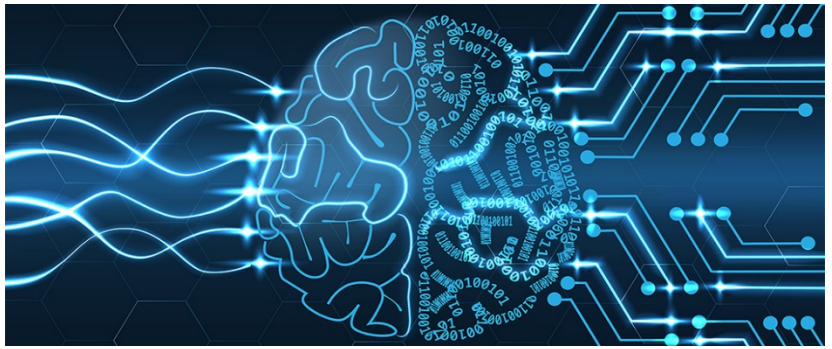
Research Skills Day, SMSTC, 3rd March 2023

LEADERS IN IDEAS AND SOLUTIONS

Data in research

Outline

- Acquiring data
  - availability and restrictions
- Managing data
  - storing, handling
- Analysing data
  - from descriptives to modelling
- Publishing data
  - safe outputs



Acquiring data

- Publicly available
  internet, literature, historical
- Upon request
- Public bodies
  ONS, UK Data Archive, UK Data Service, NHS digital,
  HMRC datalab, UK Biobank, MoH Malaysia …
- Industry
  Social media platforms, online data sites, insurance sector …
- Big data
  “four Vs” – volume, velocity, variety, veracity)
- Data collection

Acquiring data - process

- Public bodies
  - Certain processes in place
  - Address ethical, safety, confidentiality considerations
  - Requires time
  - And cost …

- Industry
  - As above
  - But also, often harder to obtain

Acquiring data – process
Case study: ONS

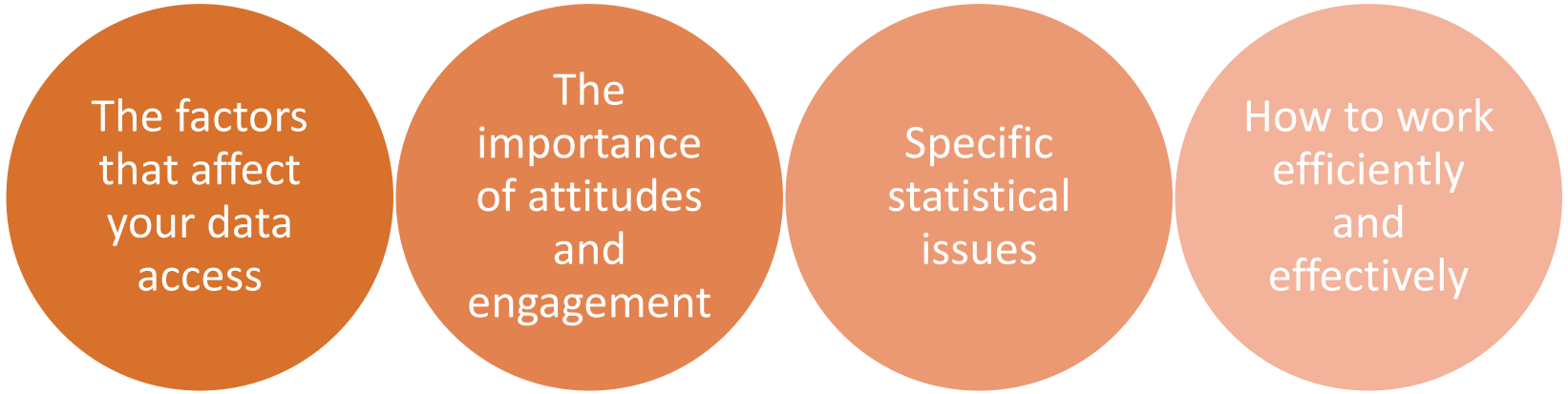
Process includes:

- Become an ‘accredited researcher’
  - Training, assessment
- Apply for a project
- Agree certain ‘Security Specifications of Controlled Access and Use’ terms
- Apply for ethical approval
  - often with both your institution and the data provider

Acquiring data – process
Case study: ONS

Accredited researcher training objectives

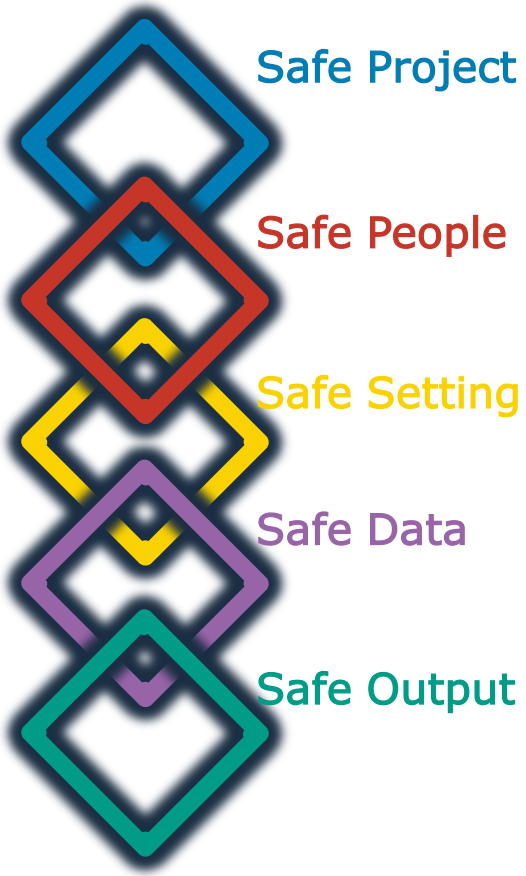
Understand:



Acquiring data – process
Case study: ONS

The 5-safes Framework:

working in a secure environment



'The framework is optimised for controlled data, de-identified data which are considered confidential or sensitive, but can be applied to any type of data.'

Acquiring data – process
Case study: ONS

Ethics

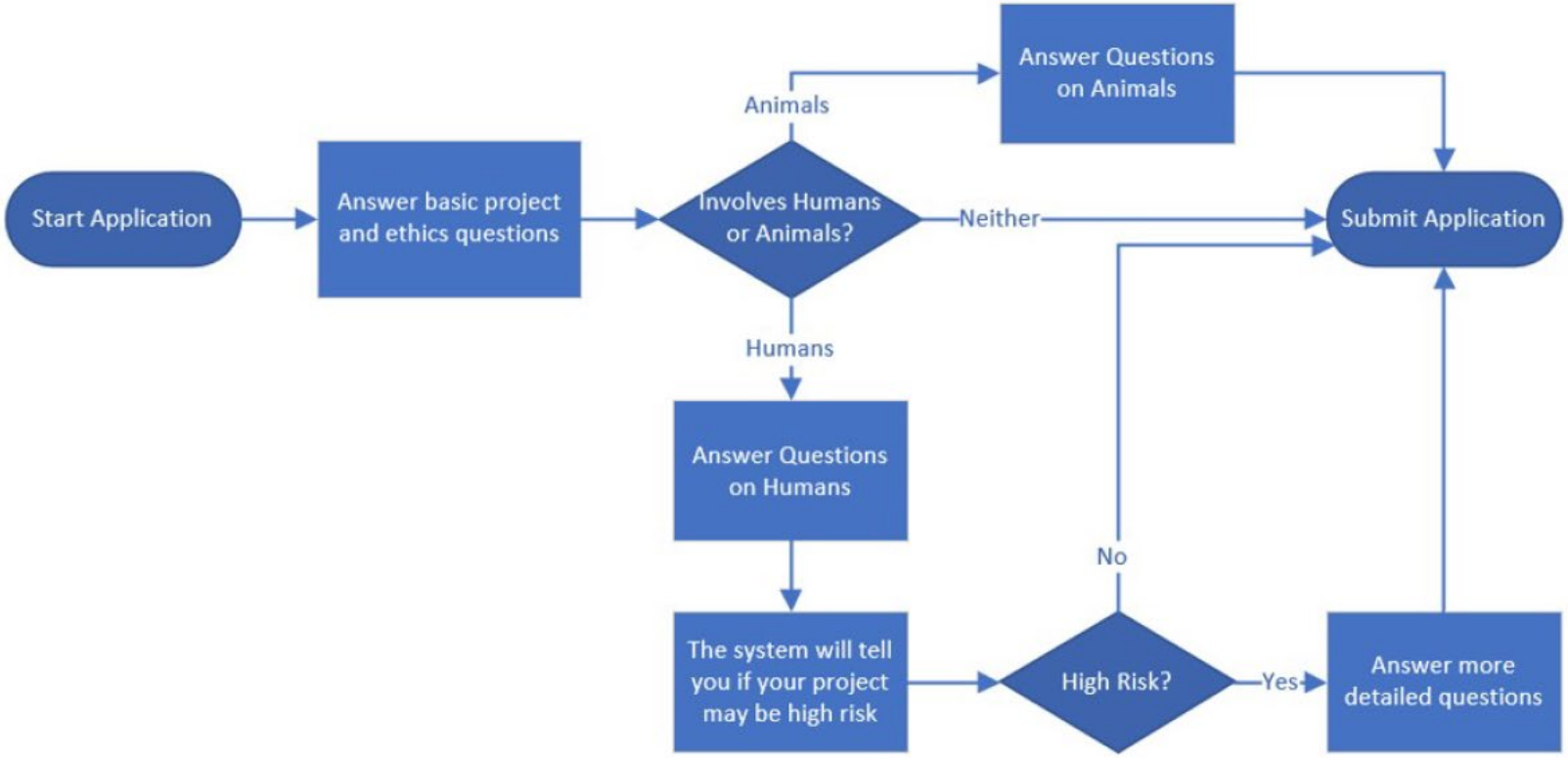
- Robust ethical approval
- Both with research institution (uni) and data provider (ONS)
- At minimum level, ensure data are
  - Anonymised
  - non-identifiable
  - conform with various GDPR policies

  (*General Data Protection Regulation*)

Acquiring data – process
Ethics

Ethics Application Form

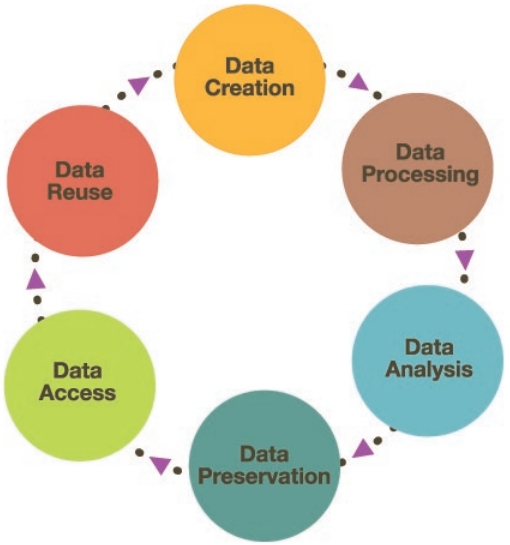
What is Going to Happen?



Managing data

Most research projects require a robust **Data Management Plan**

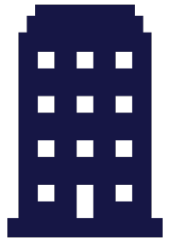
- What data will you collect, create, use?
- What documentation and metadata will accompany the data?
- How will you manage copyright and Intellectual Property Rights issues?
- How will the data be stored and backed up during the research?
- How will you share the data?
- What is the long-term preservation plan for the dataset?



Managing data - storing, handling

Several safeguards. Need to ensure:

- Appropriate encryption
- Safe storage
- Safe setting - e.g. for ONS access



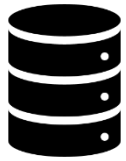
- **Safe Room / SafePod –** Protected rooms with terminal/s available to book in advance.

- **Assured Organisational Connectivity (AOC) –** Access options are agreed with your organisation, based onphysical and technical security standards.

- **AOC Remote Access** – An add-on agreement between ONS and your organisation, permitting access to SRS from home, if additional security standards have been met.

Managing data - storing, handling

Safe Data



Single dataset

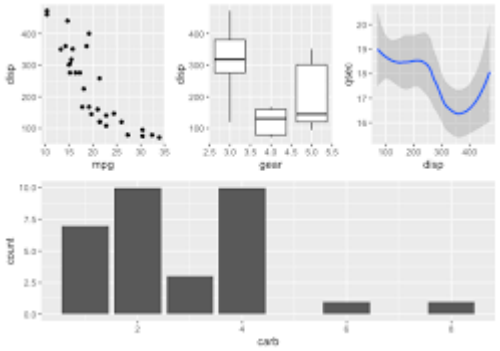
- Level of detail
- Which variables do you need?

Multiple datasets

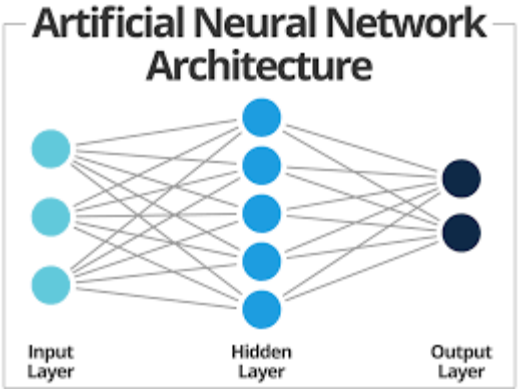
- Risk of using, linking or matching multiple data.
- Why do you need this combination of data?
- Risk of identification raises

Analysing data

- Organise & clean the data first!
- Need appropriate tools, techniques, methodology
- Statistics, Data science
- Modelling
  - complex statistical methods (more insights)
  - versus simpler methods (communicating to policymakers and other research users)



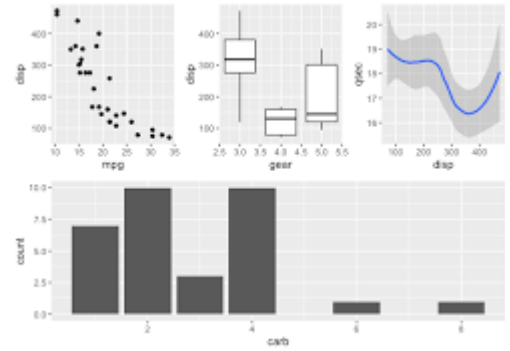

```
glm(formula = MPG.city ~ Weight, family = Gamma(link =
"log"))
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.29832  -0.06555  0.00177   0.04916   0.43407
Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 4.134e+00  5.997e-02   68.92   <2e-16 ***
Weight      -3.408e-04  1.917e-05  -17.78   <2e-16 ***
---
(Dispersion parameter for Gamma family taken to be
0.01176586)
    Null deviance: 4.9357  on 92  degrees of freedom
```



Organising the data

Example: Industry (health) data

AJ PhD, HWU



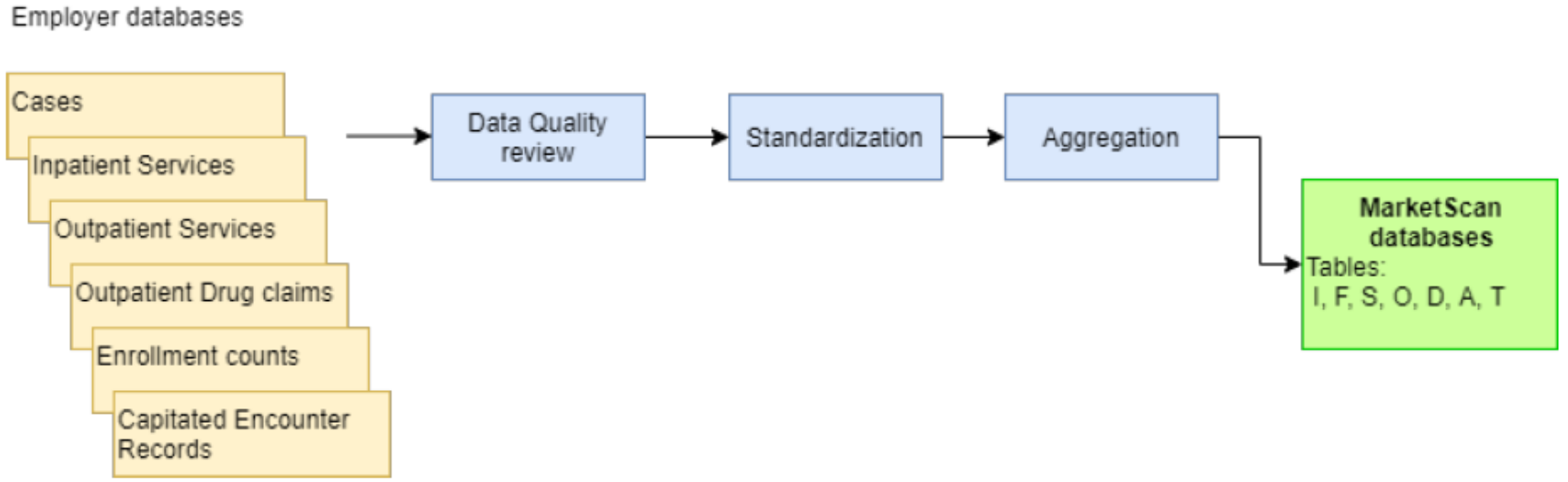
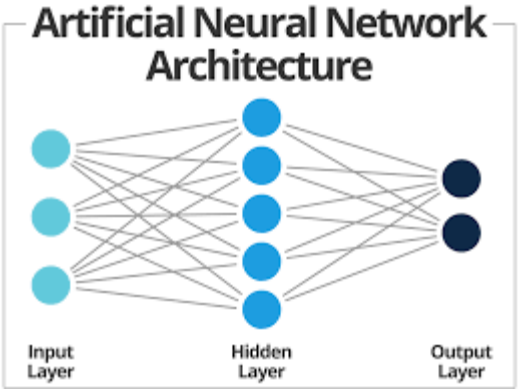


Figure 3.2: Data construction

```
glm(formula = MPG.city ~ Weight, family = Gamma(link =
"log"))
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.29832  -0.06555 0.00177   0.04916   0.43407
Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 4.134e+00  5.997e-02   68.92   <2e-16 ***
Weight      -3.488e-04  1.917e-05  -17.78   <2e-16 ***
---
(Dispersion parameter for Gamma family taken to be
0.01176586)
    Null deviance: 4.9357  on 92  degrees of freedom
```



Publishing data

- ‘Safe’ outputs
- What information is published?
- Can this be used to identify individuals?
- Is sensitive or confidential information released?

Ensure that outputs produced from confidential data pose a minimal risk of disclosure of identity and/or personal information.

Publishing data
Case study: ONS

Conform to **Statistical Disclosure Control (SDC)**

- A process applied to data outputs (statistical results)
- Mitigate risk of potentially disclosive information

HERIOT WATT UNIVERSITY

Publishing data
Case study: ONS

SDC Example: Class Disclosure

Table: Income distribution

Highest Qualification	Income quartile (lowest to highest)				Total
	1	2	3	4	
Postgrad	1	1	8	18	28
Degree	2	6	14	17	39
College	8	18	16	3	45
School	13	9	0	0	22
None	13	3	0	0	16
Total	37	37	38	38	150

HERIOT WATT UNIVERSITY

Publishing data
Case study: ONS

SDC Example: Indirect Disclosure

All persons

Age bands	Socio-economic group		Total
	Working class	Middle class	
50-54	21	11	32
55-59	25	11	36
60-64	28	12	40
65+	31	11	42
Total	105	45	150

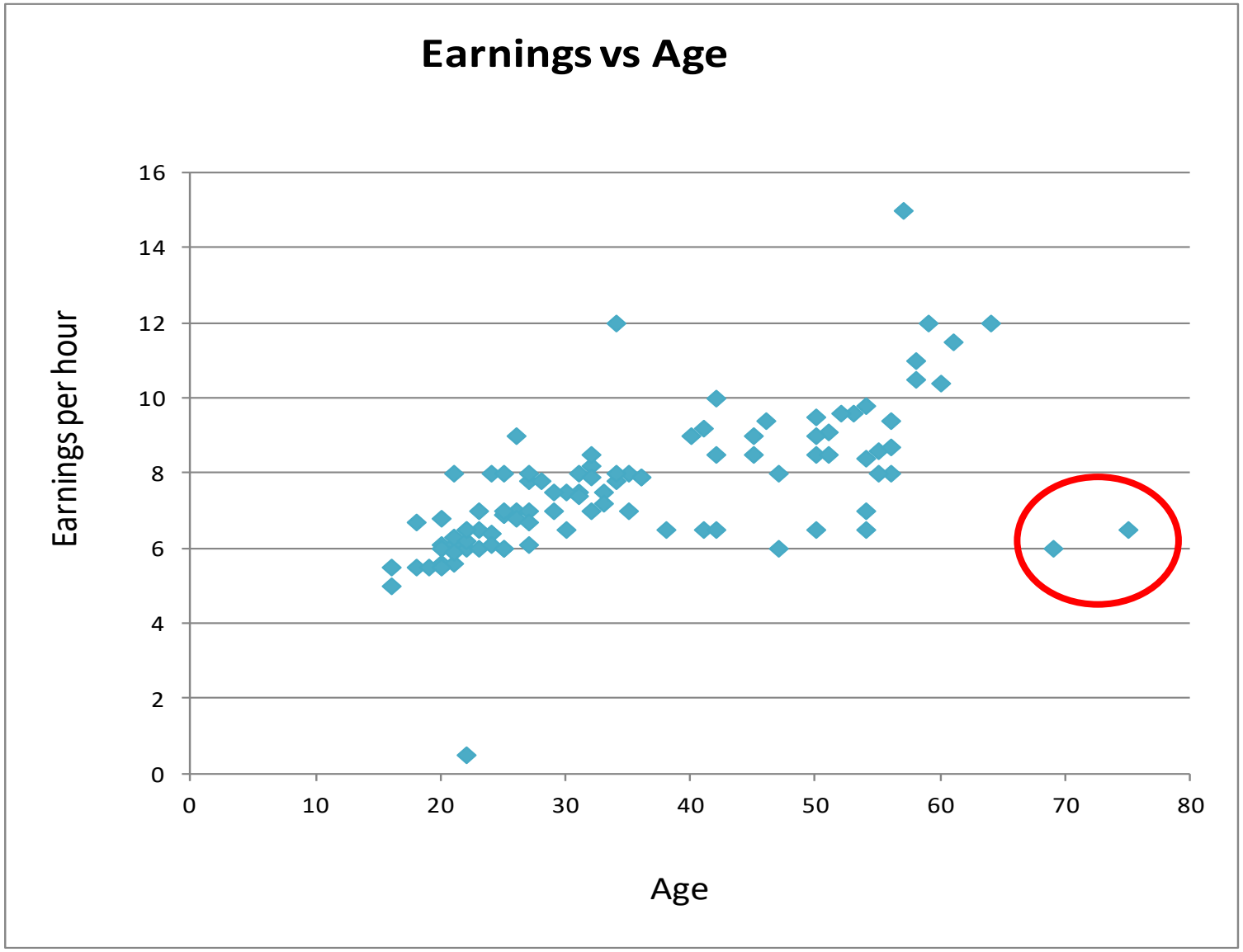
Non-diabetics

Age bands	Socio-economic group		Total
	Working class	Middle class	
50-54	17	7	24
55-59	19	9	28
60-64	23	8	31
65+	23	10	33
Total	82	34	116

Publishing data
Case study: ONS

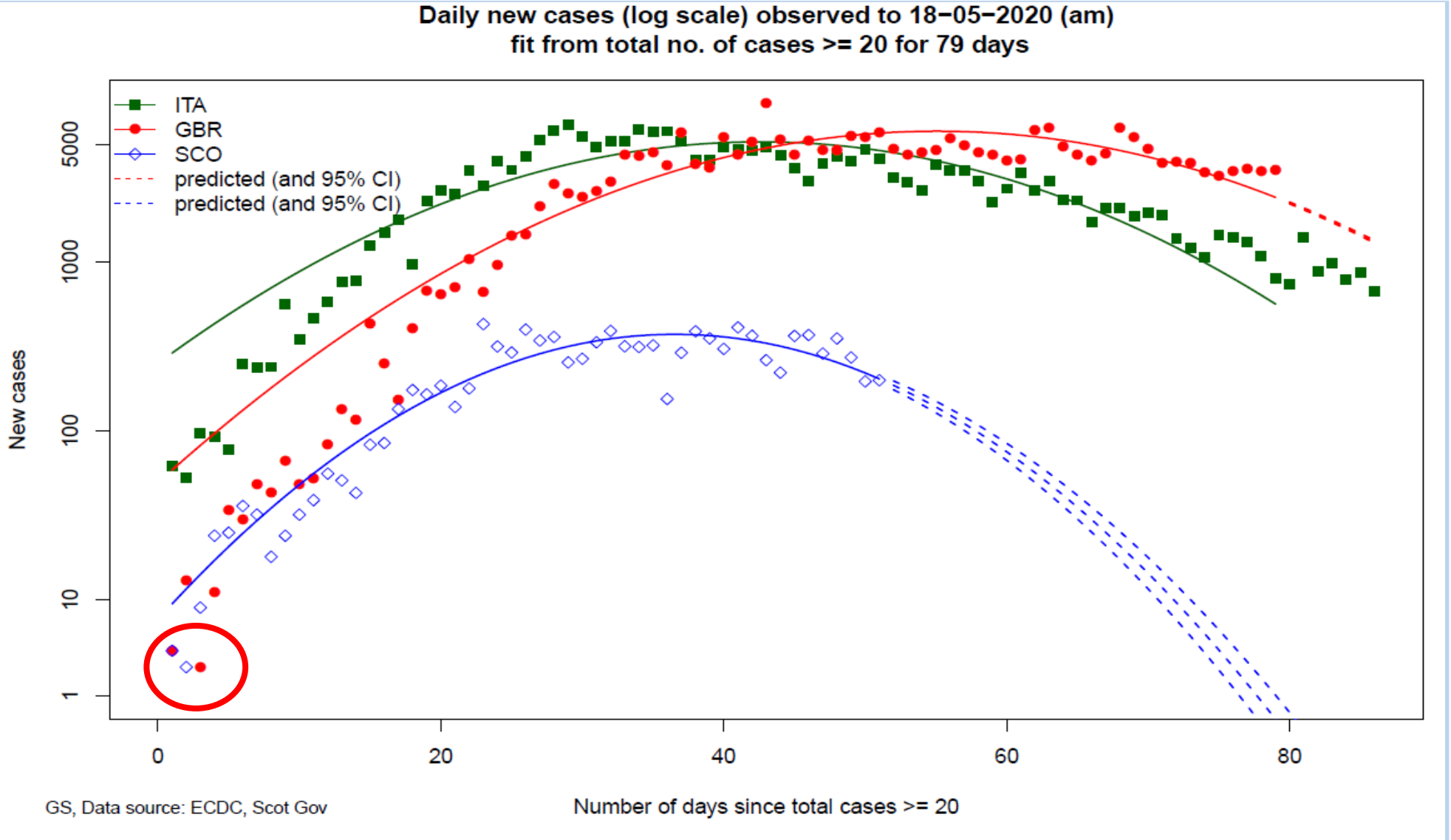
SDC Example: Scatterplots

Individual data



Publishing data

Example (HWU research) - Covid-19 – early days



Poisson generalised linear model

Publishing data

Example (HWU research)

Cancer trends by region and deprivation – how big is the gap?

- Are there regional or socio-economic differences?
- Is the gap getting wider?
- Data by age, year, deprivation, gender, region

Publishing data
Example (HWU research) - Modelling

Bayesian Generalised Linear Model:

$$C_{a,t,d,g,r} \sim \text{Poisson}(\theta_{a,t,d,g,r}\ E_{a,t,d,g,r})$$

$$\theta_{a,t,d,g,r} \sim \text{Lognormal}(\mu_{a,t,d,g,r}, \sigma^2)$$

$$\mu_{a,t,d,g,r} = \beta' X$$

$$\beta\text{'s} \sim \text{Normal}(0, 10^4)$$

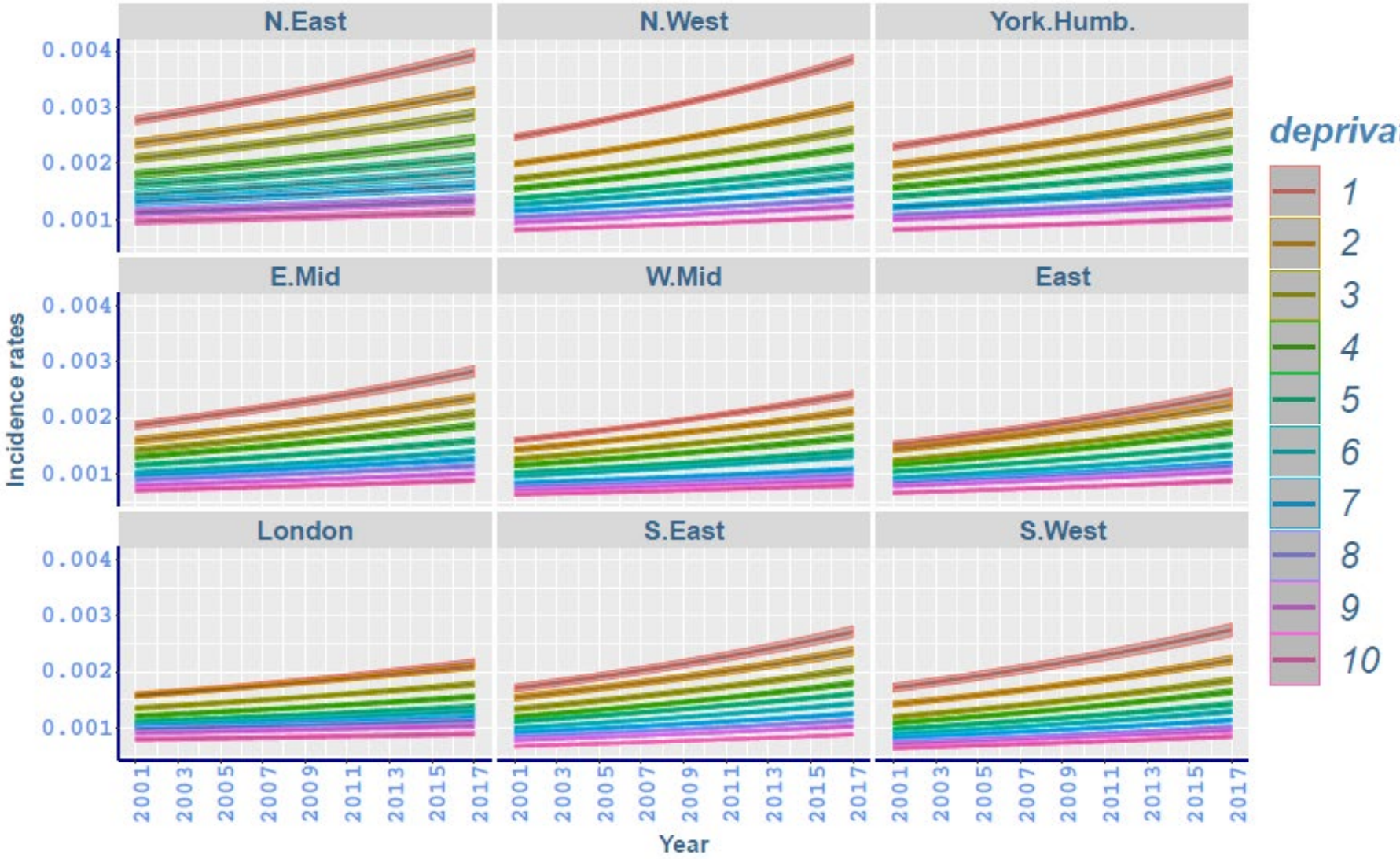
$$\sigma^2 \sim \text{Inv Gamma}(1, 0.001)$$

- Age: higher rates at older ages
- Time:
  - higher incidence in more recent years
  - lower mortality
- Gender: higher rates for men
- Region? Deprivation?

$$\mu_{a,t,d,g,r} = \beta_0 + \beta_1\ Age + \beta_2\ Year + \beta_3\ Deprivaton + \beta_4\ Gender + \beta_5\ Region$$

Publishing data
Example (HWU research) - Graphs

Deprivation inequality in cancer rates – lung cancer incidence, women, 2001-2017



Publishing data
Maps

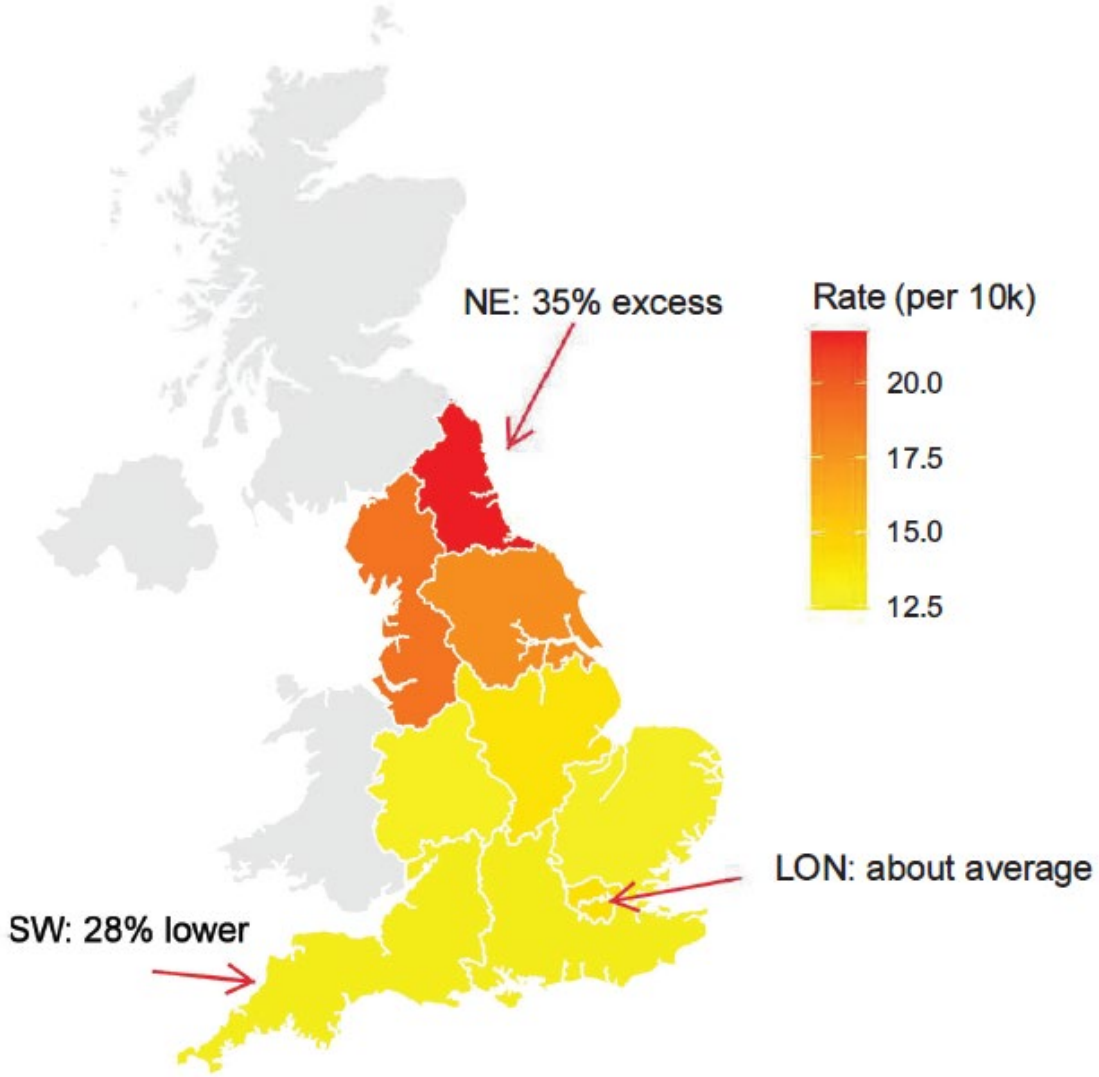
- Individual data

(here businesses registered for VAT)

- Are there risks with publishing this map?

Publishing data - Maps

Example (HWU research) – Cancer trends by region and deprivation



Regional variation in lung cancer incidence rates – Women, 2017:

Heatmap

Summary

Plethora of data sources

- data revolution age, big data
- need to be particularly careful how to use data in your research

Plan ahead

- acquiring, organising, analysing data
- time and effort

Safe data

- anonymise, de-identify, encrypt, store
- disclosure of sensitive data
- publication must conform with safeguards