Good Practice in Health Data Privacy: A Guided Tour

Dr Becca Wilson, Newcastle University

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

EU General Data Protection Regulation

- Legal framework: guidelines for collection and processing of personal information of those in EU
  - who can do what with data
  - incl processing outside the EU

- **Data users** must be compliant – not only data controllers (custodians)
  - Privacy by design – accountability and responsibility to prevent disclosure
  - Breach reporting

https://bit.ly/2020cw-data-privacy

EU General Data Protection Regulation

Individuals given more protection

- Consent:
  - stricter rules (no auto tick consent)
  - **can withdraw consent anytime**
- Right to:
  - **access** data
  - **Be removed from dataset**
  - know **what happens** to the data

For us: stronger data governance frameworks, privacy by design infrastructure & processes, data processing logs, transparent algorithms

https://bit.ly/2020cw-data-privacy

EU General Data Protection Regulation

Non compliance = fines

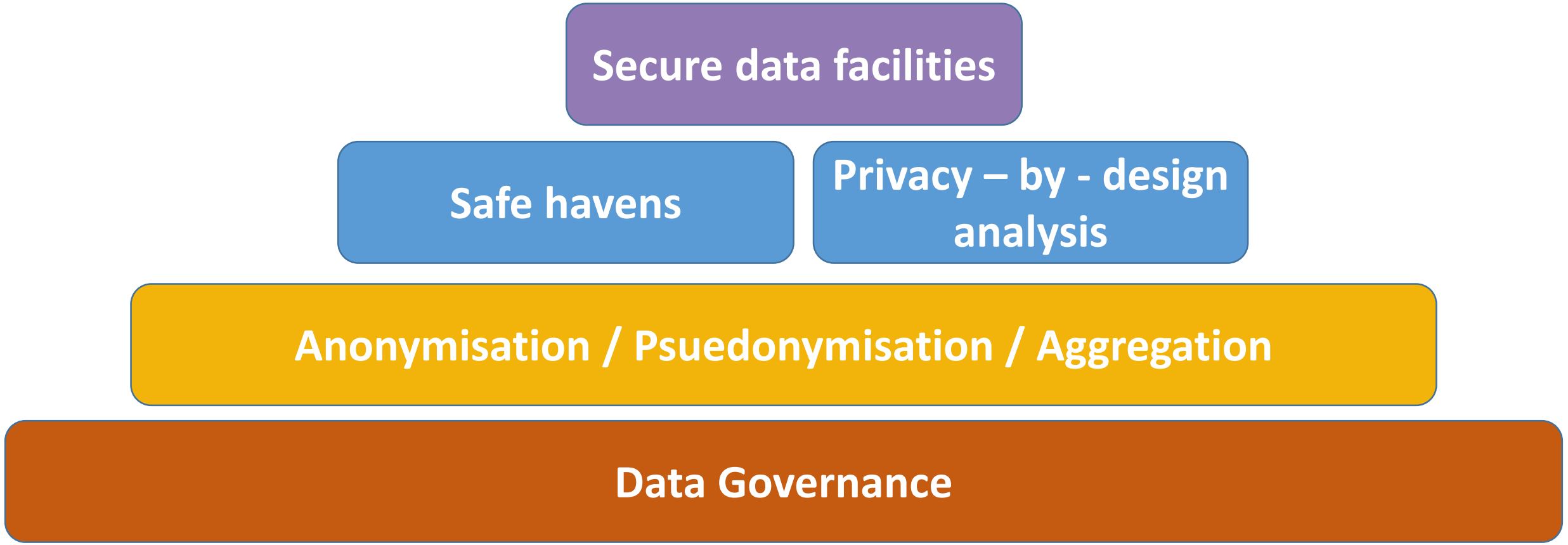
- UK applies GDPR via UK Data Protection Act, 2018. We follow both.
- DPA gives some exemptions:
  - For patient care, historical data

https://ico.org.uk/for-organisations/guide-to-data-protection/

- Post Brexit: Likely GDPR alignment
  - Google moved our data to US – as we are no longer in the EU.

https://bit.ly/2020cw-data-privacy

Data privacy measures in health research



@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Data privacy measures in health research

Policies and processes for good data management, useability, security

Secure data facilities

Safe havens

Privacy – by - design analysis

Ensure compliance with ethical-legal restrictions & data is not misused

Anonymisation / Psuedonymisation / Aggregation

Data Governance

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

Data privacy measures in health research

Secure data facilities

Safe havens

Privacy – by - design analysis

Anonymisation / Psuedonymisation / Aggregation

Data Governance

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

Anonymisation

Anonymised data: GDPR does not apply to the processing and storage

Anonymisation:
strips **all** identifiable information from a dataset

What is identifiable information?

- Name, date of birth, address
- IP address, email address

Name	DOB	Address	Smoker (Y/N)
James Smith	18-01-1978	123, Amazing Street, Brilliant Town	Y

Anonymisation

Anonymised data: GDPR does not apply to the processing and storage

Anonymisation:
strips **all** identifiable information from a dataset

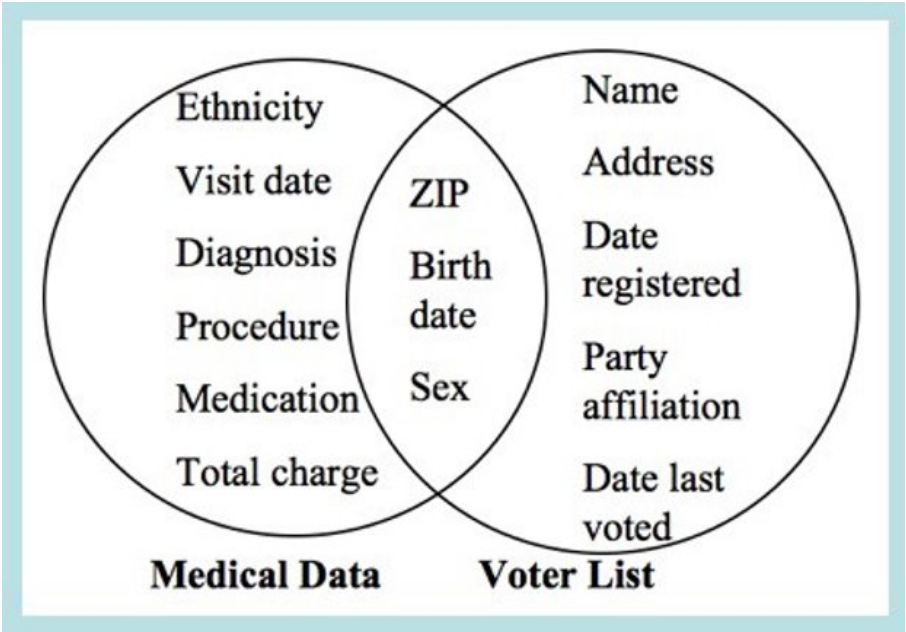
> ***Data can either be useful or perfectly anonymous, but never both.***
>
> Paul Ohm, 2009,
> https://ssrn.com/abstract=1450006

Name	DOB	Address	Smoker (Y/N)
XXXXX XXXXXXX	XX-XX-XXXX	XXXXXXX, XXXXXXXXX Brilliant Town	Y

Psuedonymisation

Psuedonymisation: replaces identifiable information with non identifiers e.g. references, aggregate or categorical information

Person Identifier	Age band	Gender	Location	Smoker (Y/N)
64982041	41-50	M	Brilliant Town	Y



Linking data increases risk of re-identification

1997 PG Latanya Sweeney re-identified Governor of Massachusetts combining publicly available data, **anonymised** health records & identifiable electoral register.

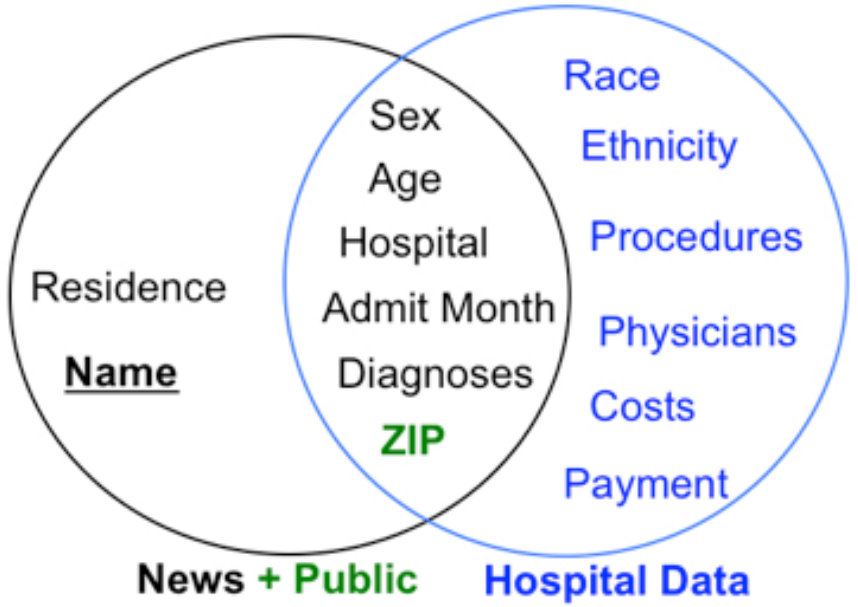
@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Linked data and re-identification

Record	
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due t loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]



Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. September 29, 2015. https://techscience.org/a/2015092903

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

Data privacy measures in health research

Secure data facilities

Safe havens

Privacy – by - design analysis

Anonymisation / Psuedonymisation / Aggregation

Data Governance

@DrBeccaWilson

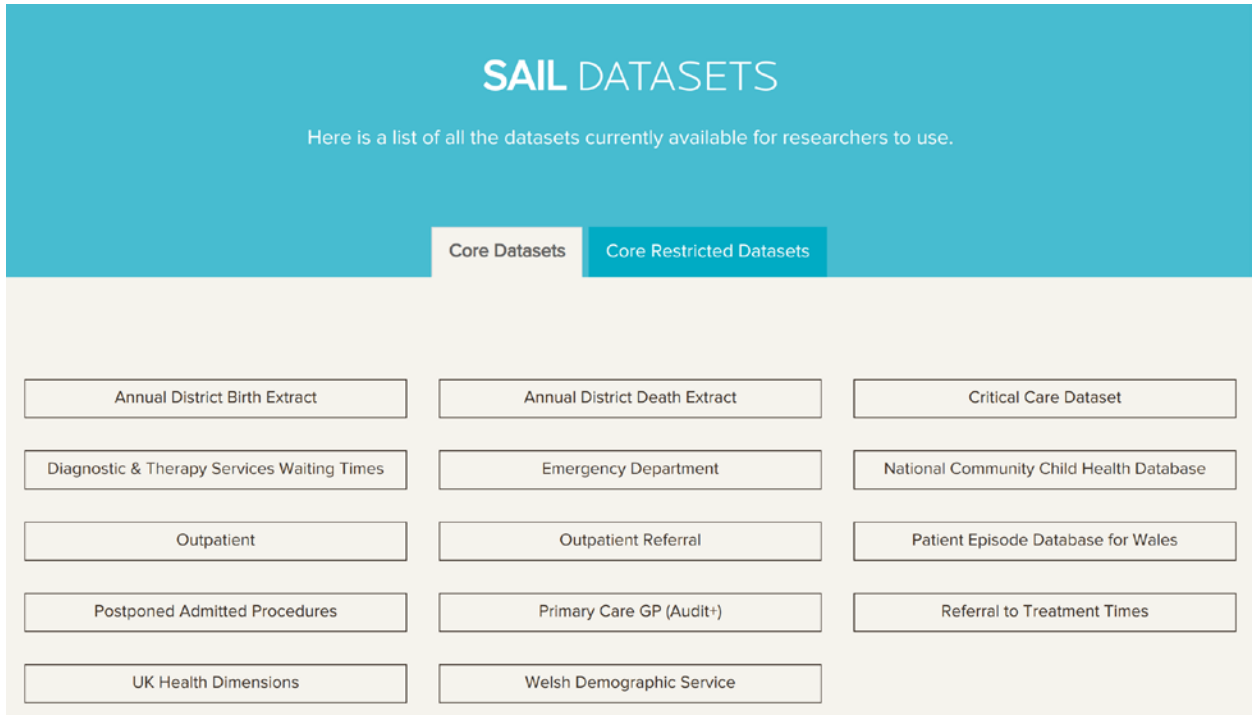
becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

Data safe havens

Secure infrastructure to store, remotely access and use research data

- Not always at scale but benefits of being national or regional
- Strong security and governance standards (e.g ISO 27001, restriction to approved users)
- Remote data analysis within the safe haven using standard tools
- Data usually anonymized, psudonymised and can be linked data
- Unable to remove data from environment

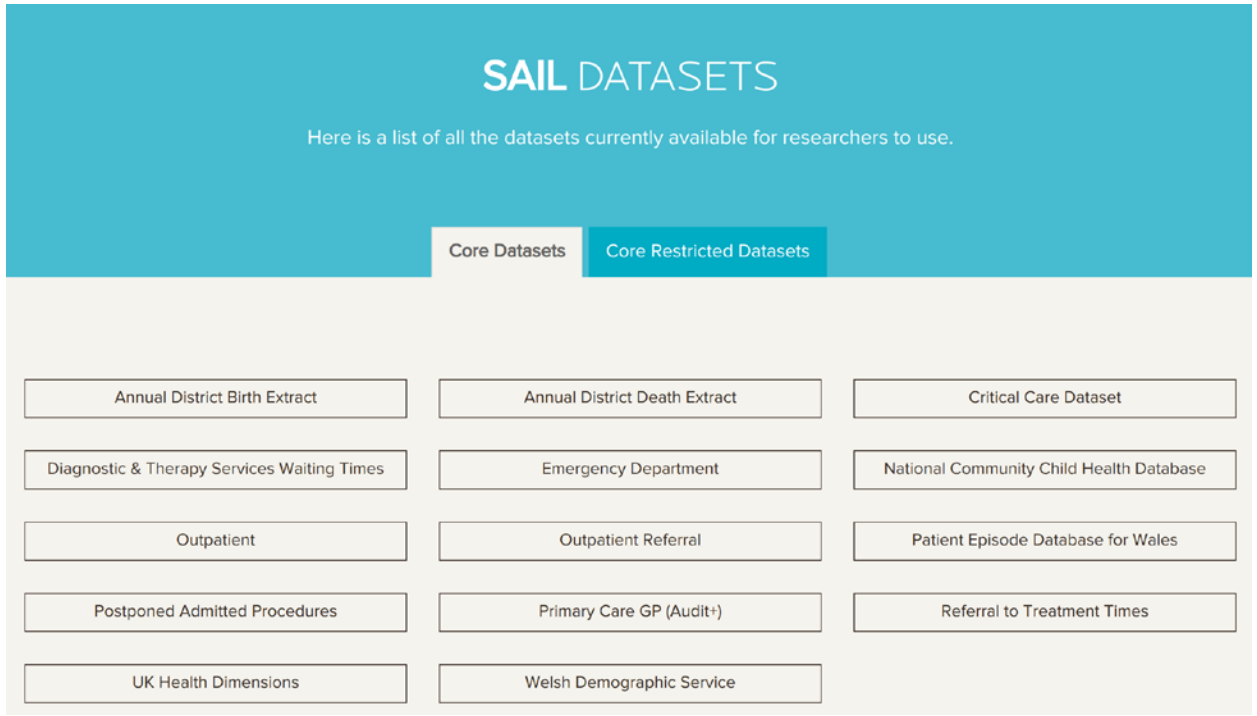


SAIL Databank: https://saildatabank.com/
Healthdata for Wales: Most requests ~12 weeks approval. Statistical analysis; medical texts for NLP.

@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Data safe haven limitations

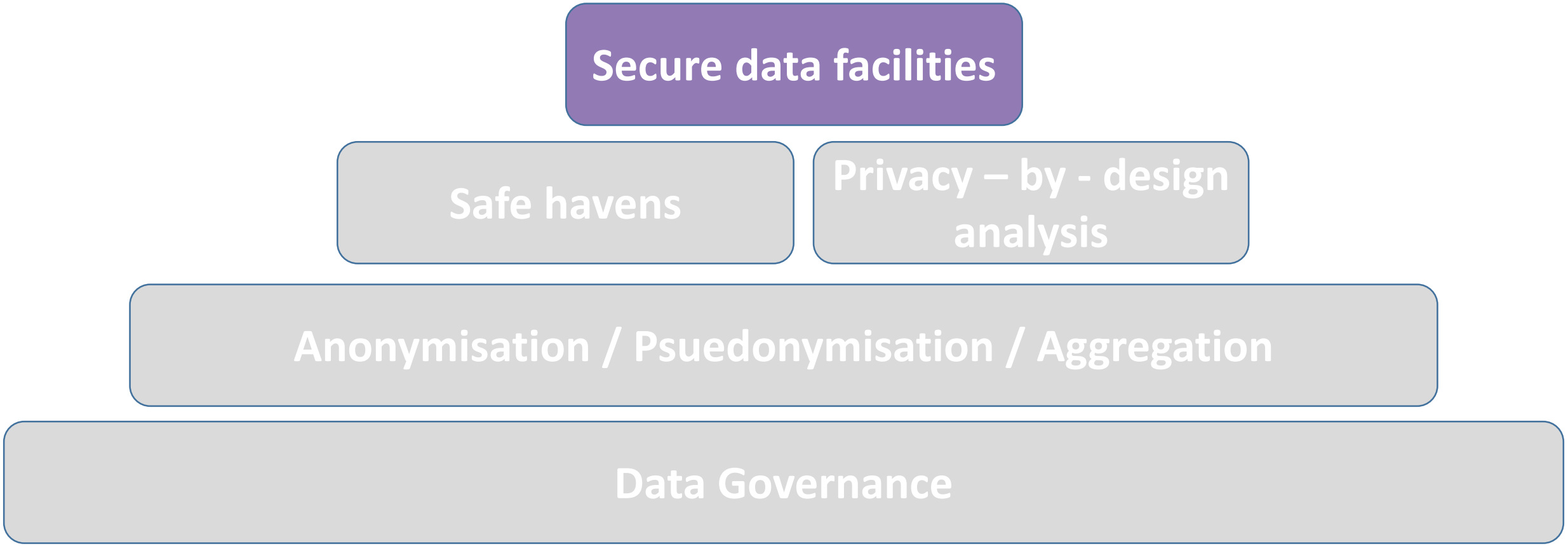
- Requires substantial investment to set up and maintain
- Safehavens - no access to outside world e.g. websites, github to bring scripts (often have an upload script mechanism with human scrutiny).
- Risk of data silos…
- Anonymisation only goes so far
- Can view the data, linked data increases reidentification, can screeenshot etc



SAIL Databank: https://saildatabank.com/
Healthdata for Wales: Most requests ~12 weeks approval. Statistical analysis; medical texts for NLP.

@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Data privacy measures in health research



@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Secure data facilities

Secure Pod: analysis of most sensitive data e.g. identifiable health care data; sensitive and confidential business, social and economic data

- Strong data governance
- Not externally networked (no internet)
- Secure connection to dataset - no data in pod
- Items prohibited e.g. keys, phone, electronic devices, usb stick, writing tools
- Often record you
- Analysis on the designated computer, results undergo human scrutiny

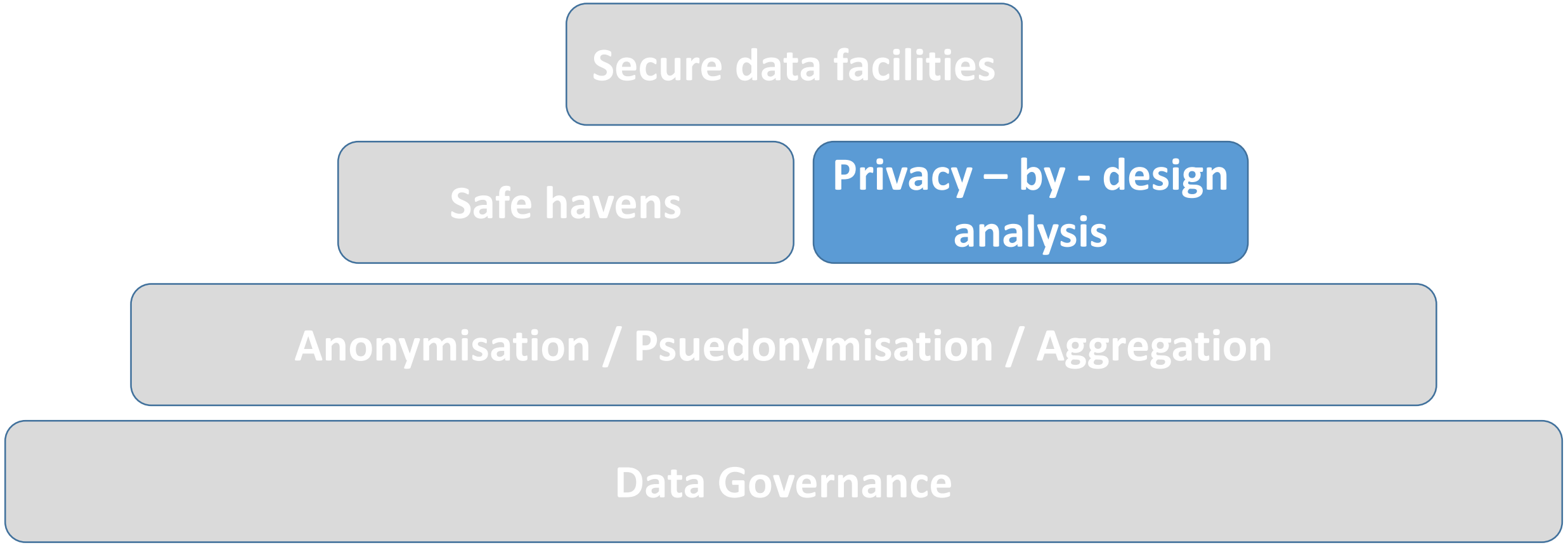
SafePod, St Andrews University

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

Data privacy measures in health research



@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Privacy by design analysis

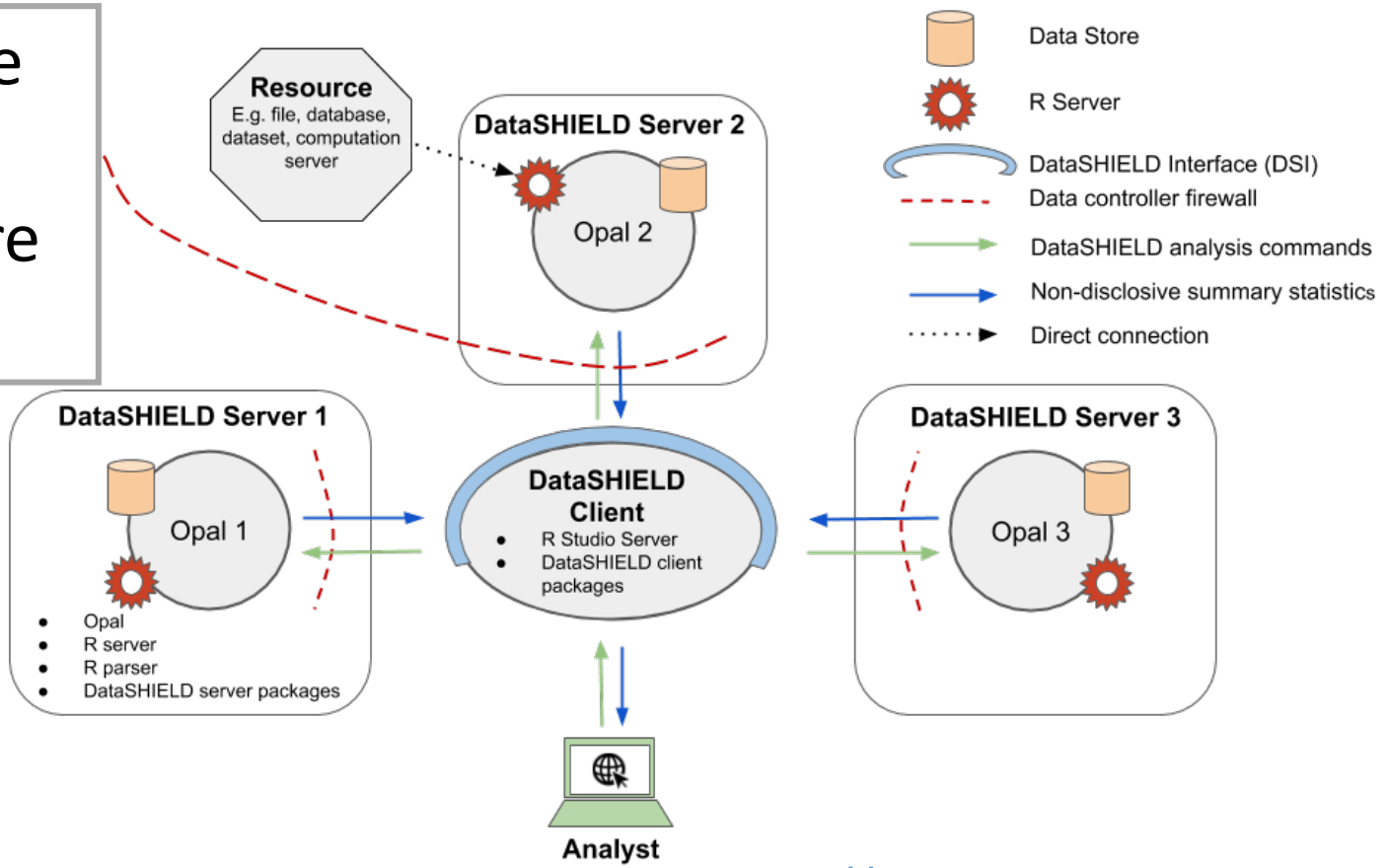
> Investment in novel technological approaches for the management of patient level data, which do not require the physical transfer of data ([DataSHIELD]), block chain and homomorphic encryption, and which meet national and international data protection legalisation are urgently required.

Cave, A et al., December 2019
for Heads of Medicines Agency - European Medical Agency joint big data taskforce:
DOI: 10.1002/cpt.1736

@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

Privacy by design analysis

DataSHIELD an open source software and infrastructure for distributed, remote analysis automated disclosure control



- 92 functions: exploratory, statistical modelling and data visualisation
- For horizontally partitioned data

Wilson et al 2017,
DOI: [10.5334/dsj-2017-021](https://doi.org/10.5334/dsj-2017-021)

http://www.datashield.ac.uk

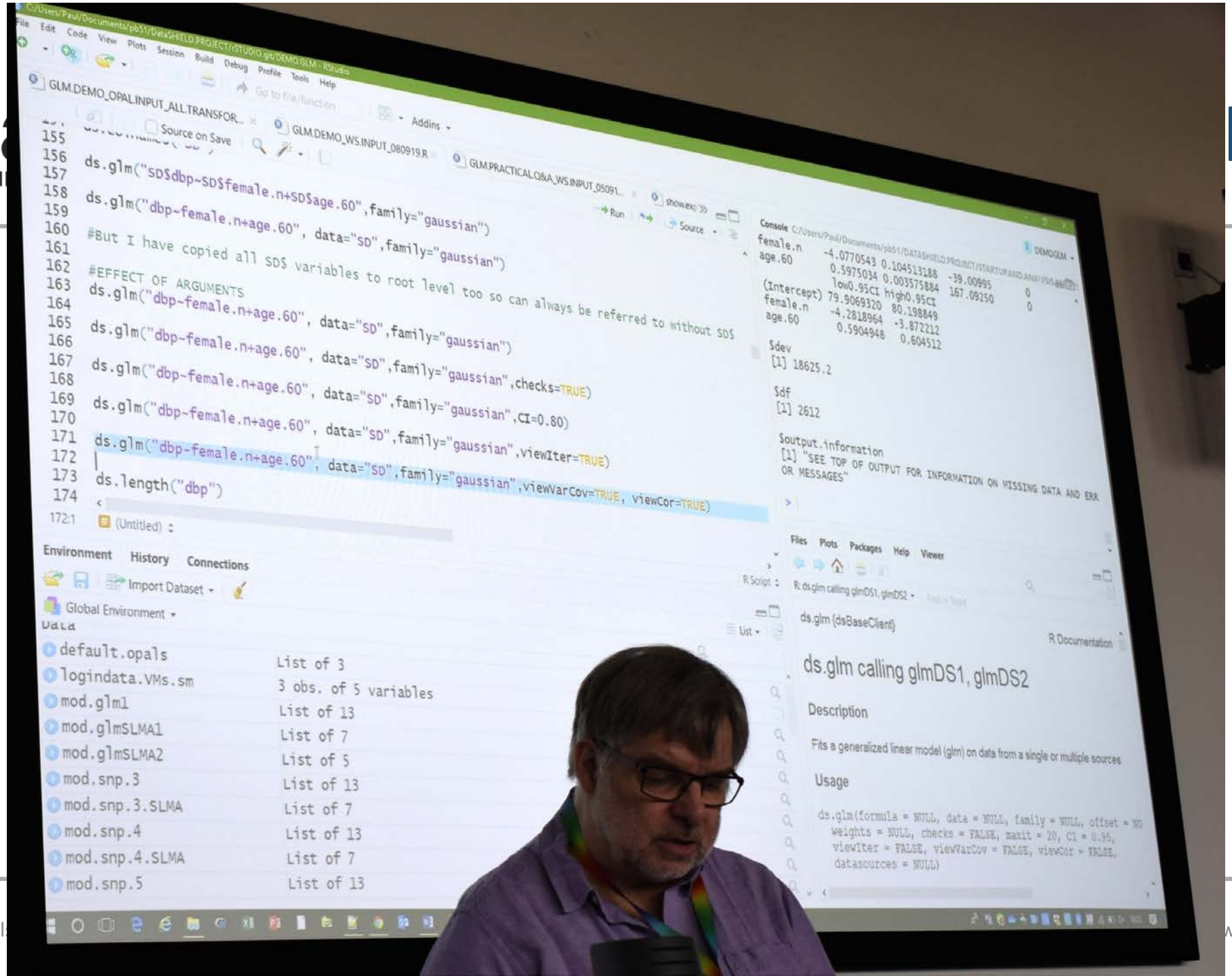
@DrBeccaWilson

becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

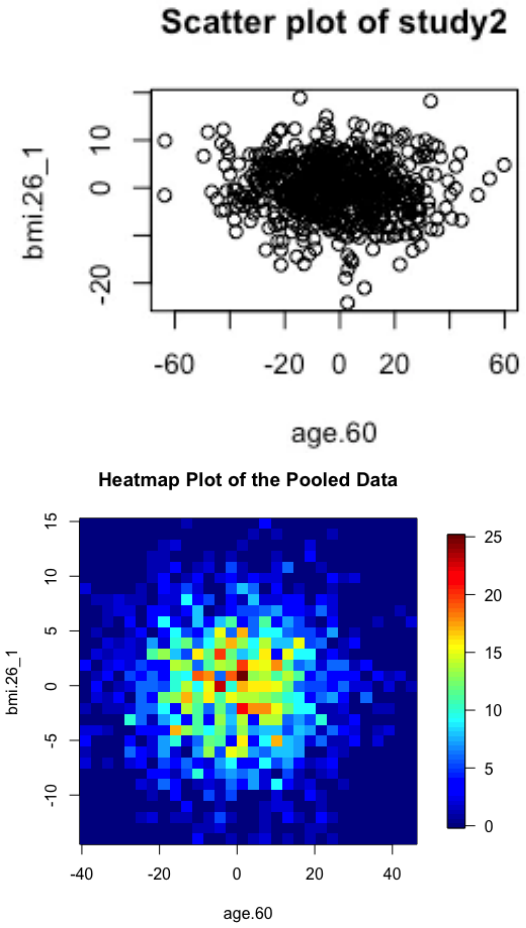
Data
Secu

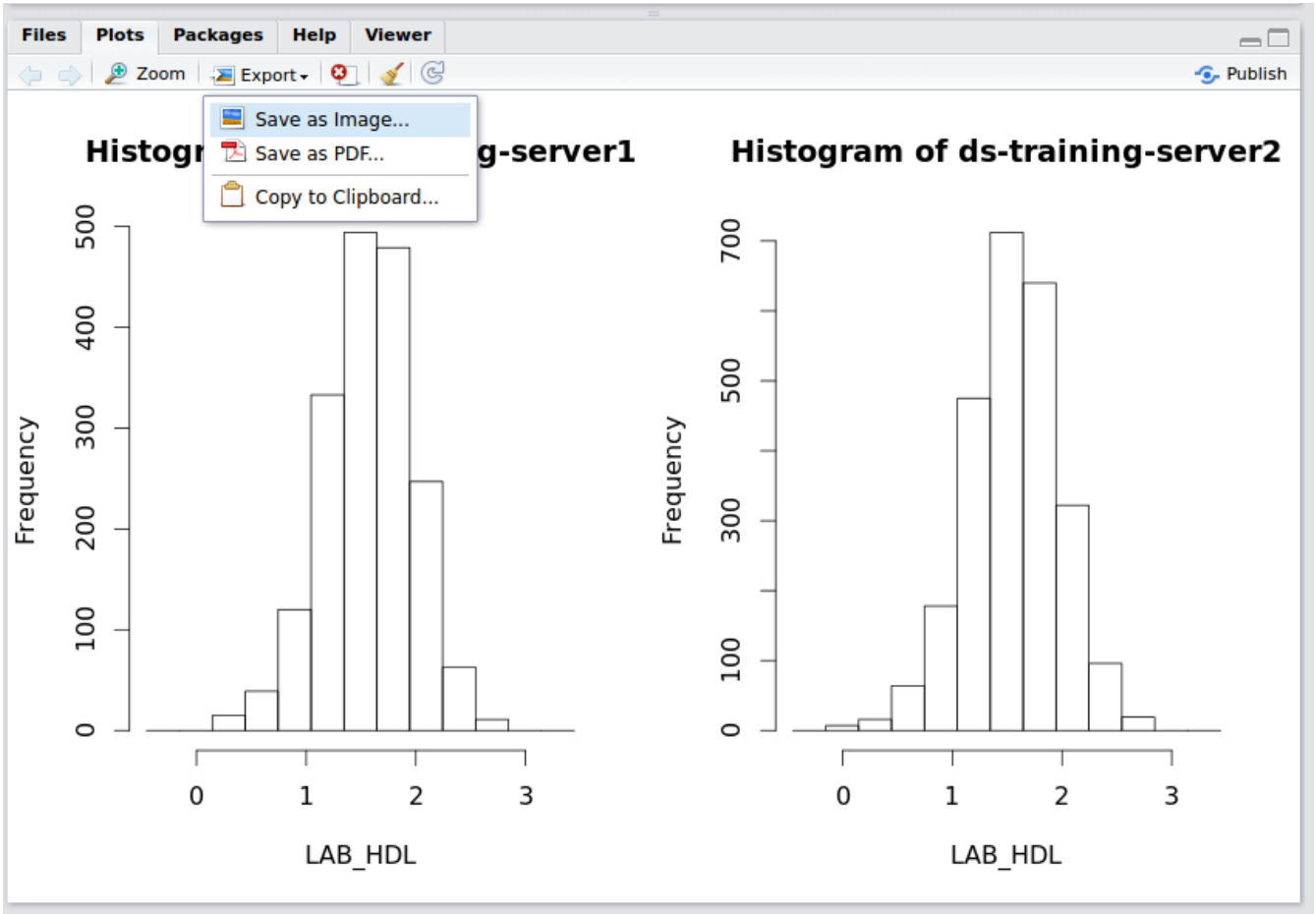
lysis



DataSHIELD
Secure Bioscience Collaboration

Working in DataSHIELD





@DrBeccaWilson

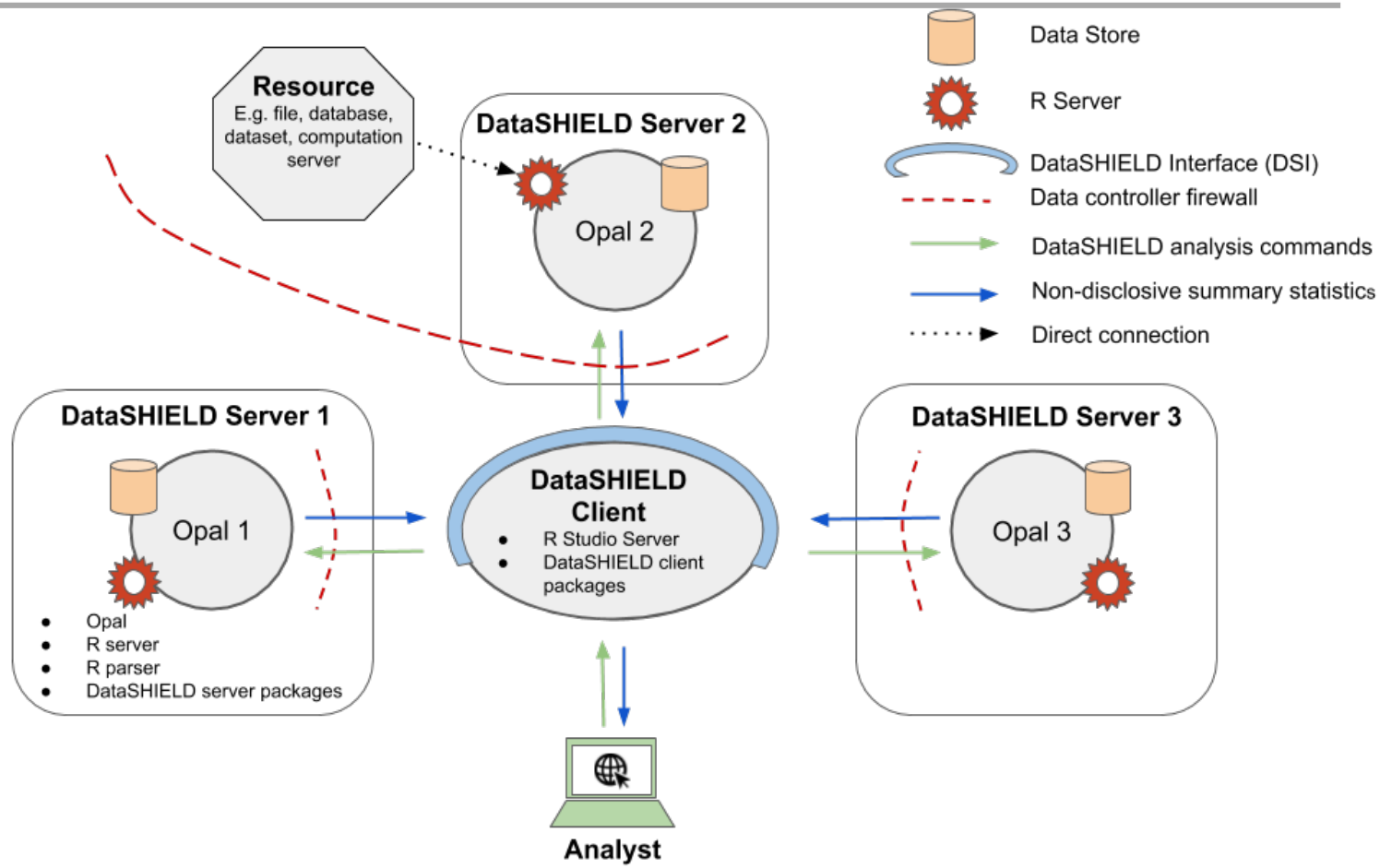
becca.wilson@newcastle.ac.uk

https://bit.ly/2020cw-data-privacy

DataSHIELD
Secure Bioscience Collaboration

Current Developments

- Resource: data connection locally hosted, via url etc
- high volume 'omics data held in standard formats such as vcf
- DsOmicsClient under dev led by DataSHIELD developers at ISGlobal



DataSHIELD Challenges

- **We're an open source software project!** SSI helped us grow
- From small University team to co-ordinating team in UK-France
  - Newcastle University, University of Cambridge
  - 2 x SMEs (operational and integration)
- Project contributers across Germany, Portugal, Spain, Netherlands, France, Denmark, Canada (https://bit.ly/datashield-team)
- Supporting research consortia Europe & Canada (https://bit.ly/DS-users)
- Limited by speed to write & test functions with disclosure control
- Funding sustainability

Join us https://datashield.discourse.group/

@DrBeccaWilson

becca.wilson@newcastle.ac.uk

Summary

- Need to have to have strong data governance and systems that are compliant with relevant ethical-legal restrictions for the data
- Tension between legislative requirements (= bare minimum) and addressing real re-identification and disclosure risks
- No one solution is perfect – combination of different methods that are suitable for your requirements

Summary

> **It is insufficient to protect ourselves with laws; we need to protect ourselves with mathematics ***

Bruce Schneier,
Applied Cryptography

* , strong data governance and secure infrastructure

@DrBeccaWilson
becca.wilson@newcastle.ac.uk
https://bit.ly/2020cw-data-privacy

DataSHIELD disclosure control

- No print to screen of individual level data
- Cell supression (consortia sets cell size themselves)
- Blocking overfitting of model: limits number of model parameters as proportion of sample size
- Maximum string length in arguments
- Minimum allowable *k* for functions that rely on k-nearest neighbour (plots)
- Minimum variance of added noise

bit.ly/DS-disclosure-control