e-Rum2020
17/20 JUNE
MILANO

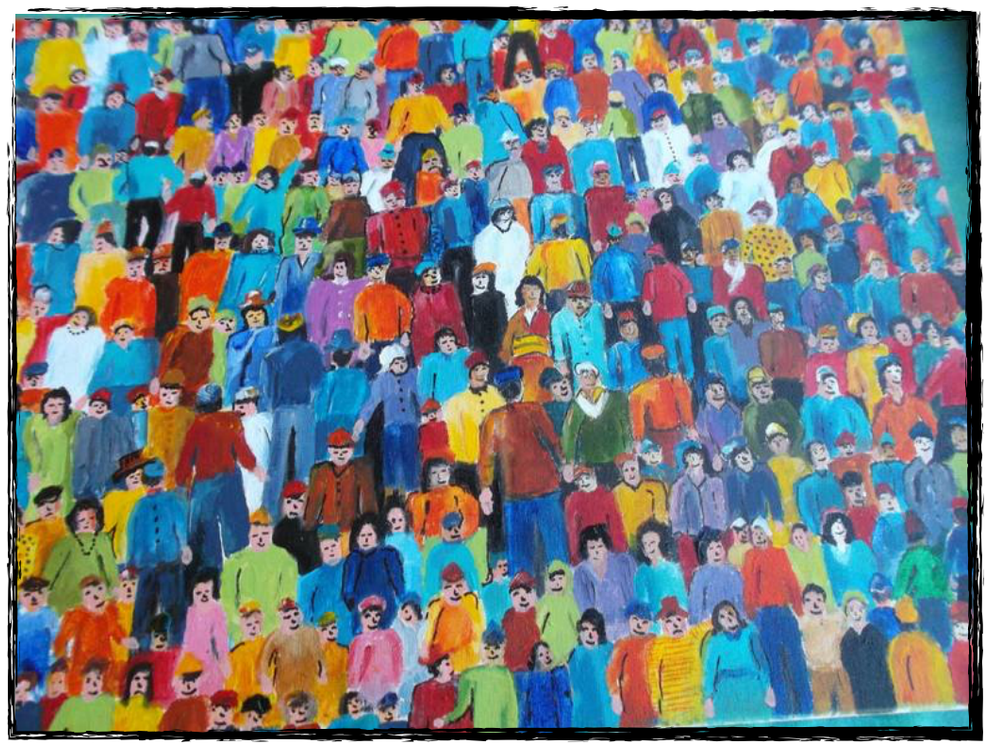
Newcastle University

Non-disclosive federated analysis in R

Patricia Ryser-Welch and the DataSHIELD team

Let’s bring some context

VS



British people will spend over four months of their lives talking about the weather, study says

https://www.independent.co.uk/extras/lifestyle/british-people-time-spent-talking-weather-conversation-topic-heatwave-a8496166.html

Let’s talk about a system that can feed their interest!

International collaboration of national weather services with some commercial agreement to access the data ….

National weather services

Data

Payment

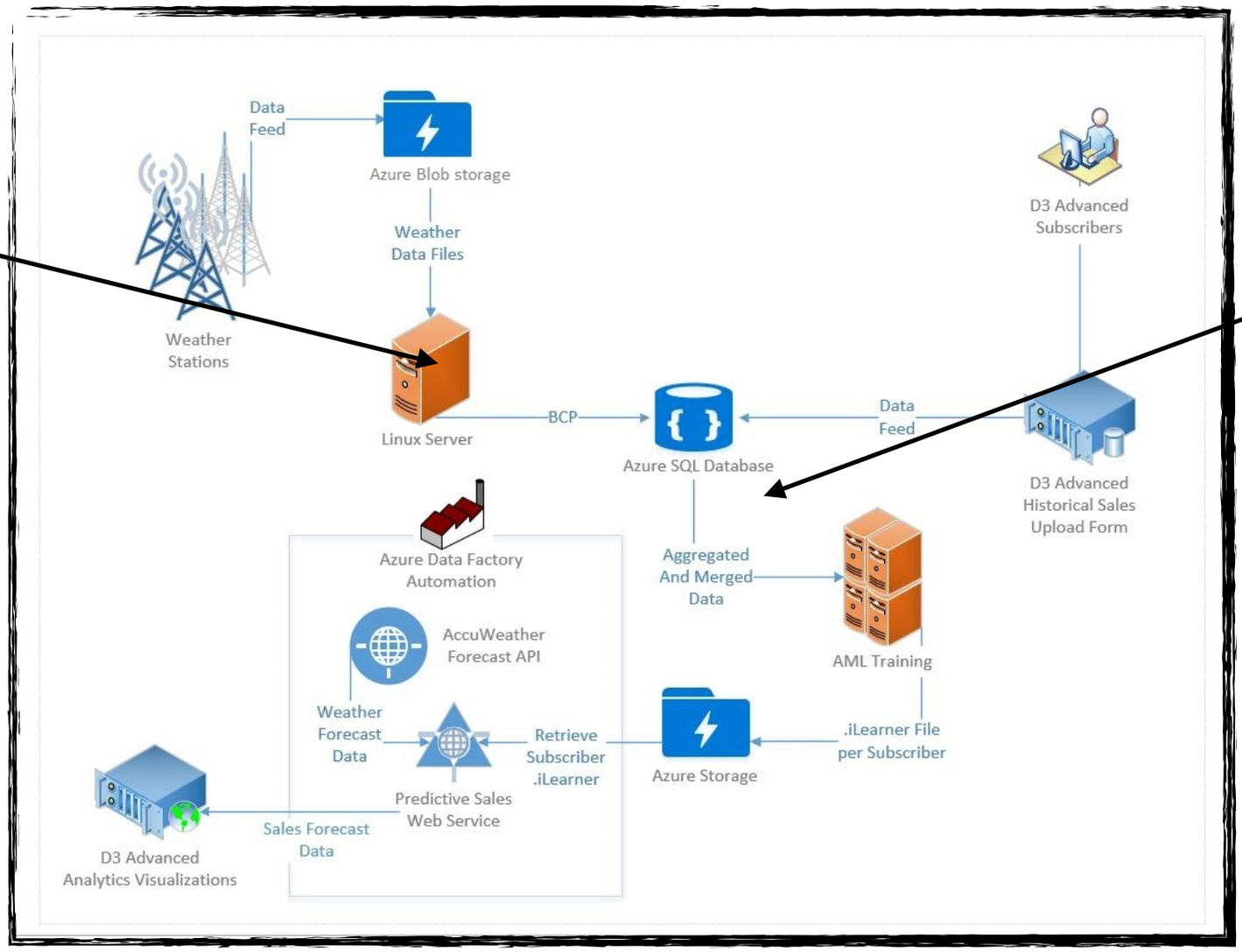
Legal agreement

Independence and autonomy of national weather services from the main meteorological system

A complex data science architecture based on Cloud technologies

Data collection

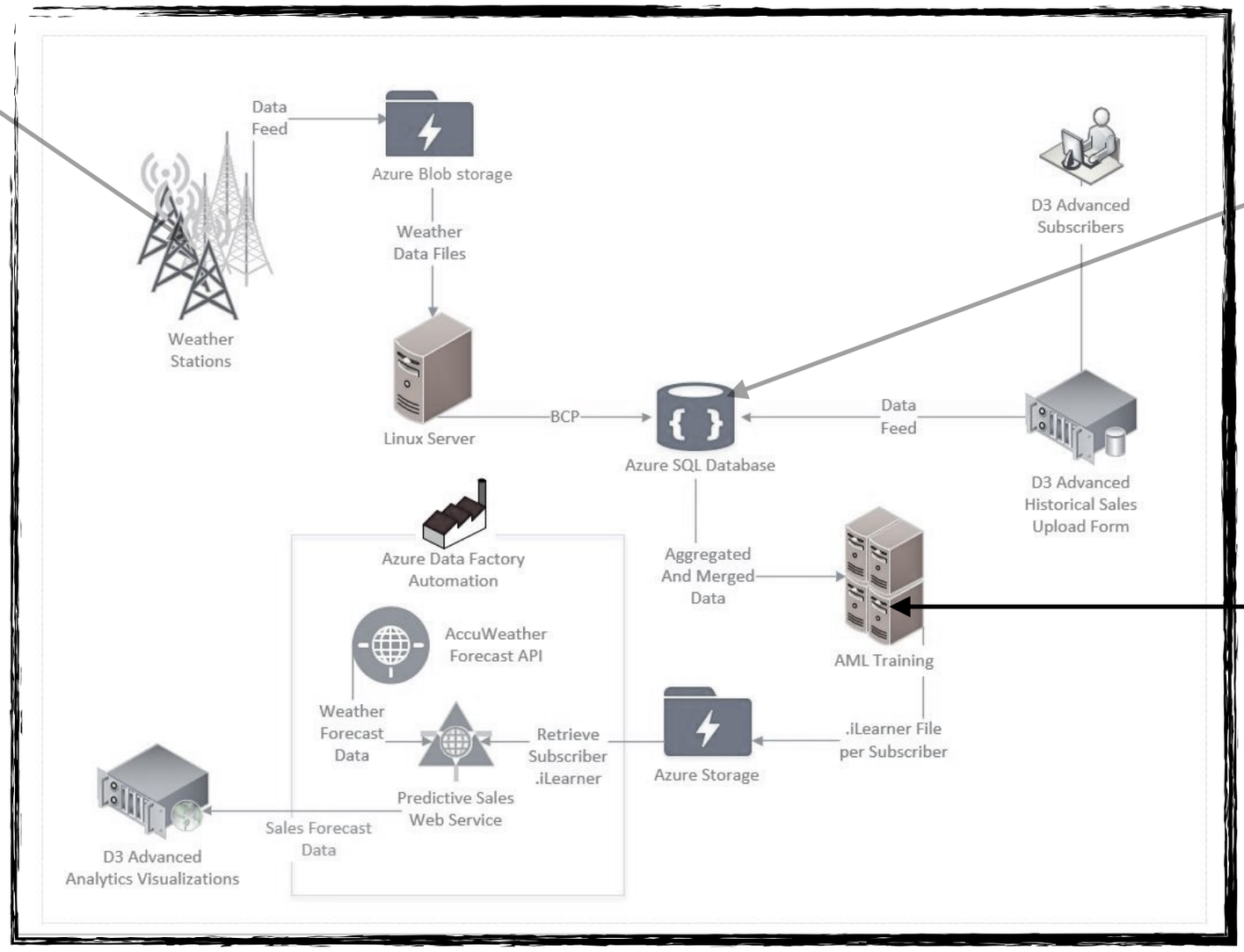
Storage



https://customers.microsoft.com/en-us/story/accuweather-partner-professional-services-azure

A complex data science architecture based on Cloud technologies

Data collection



Storage

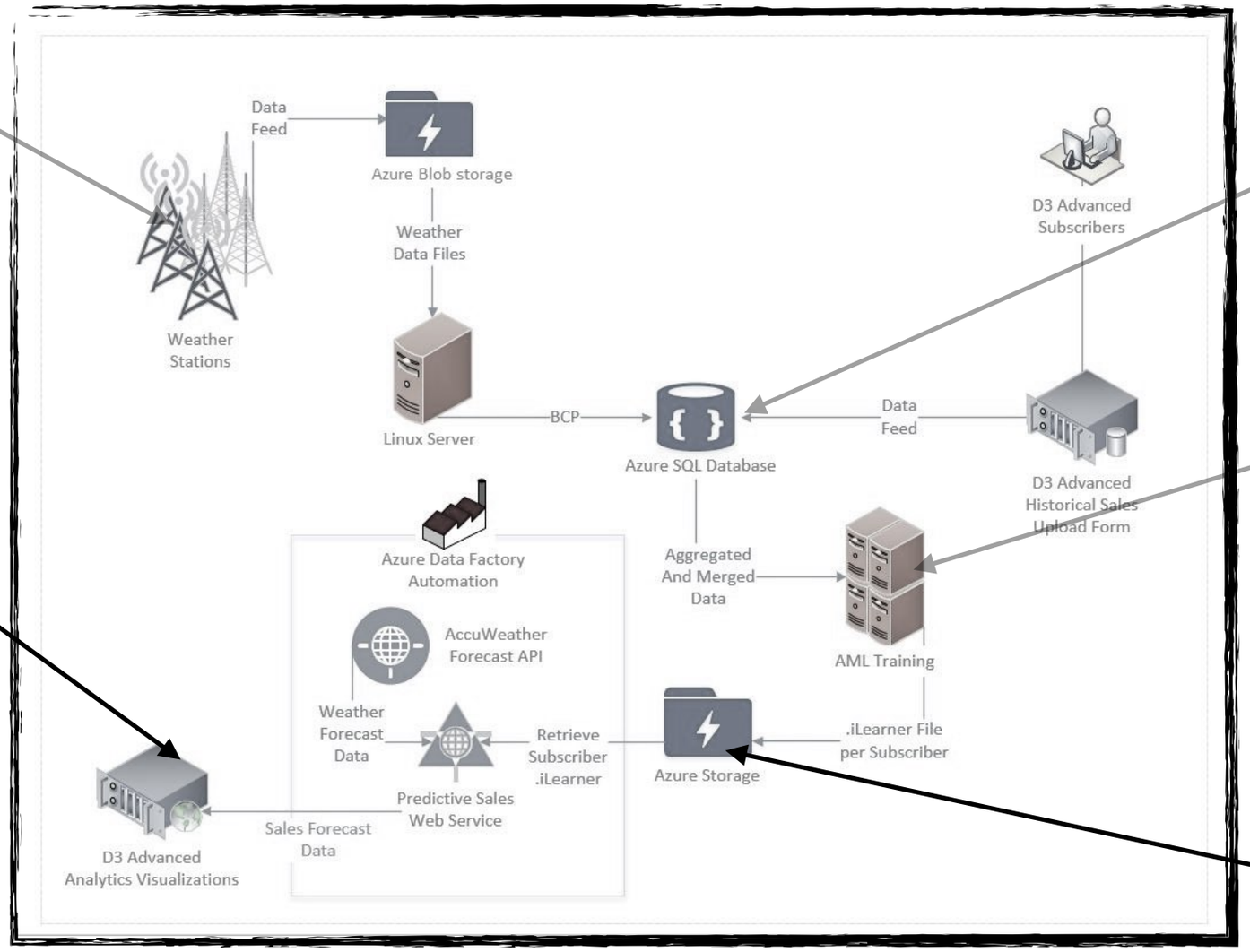
Combine dataset for data proces-sing

https://customers.microsoft.com/en-us/story/accuweather-partner-professional-services-azure

A complex data science architecture based on Cloud technologies

Data collection

Storage



Visualisation of results

Combine dataset For data proces-sing

Computations of Forecasts and statistical analysis

https://customers.microsoft.com/en-us/story/accuweather-partner-professional-services-azure

What about the data?

"Weather data" is not related to any identified or identifiable individual.

No individual can be directly or indirectly identified from the weather data processed.

Processing weather data do not need to consider (1) the content of information, (2) purpose of the processing, (3) impact or effect of that processing on some individual

No data needs to be truly anonymised before being shared or processed

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Researching Preterm babies

15 countries European Countries sharing information about preterm born babies

Large amount of data over 30-year time span

Inform about healthcare, social and education policy

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

EUCAN-Connect

13 highly experienced organisations and individuals across Europe and Canada

Make available existing populations data accessible to global scientific community.

FAIR federated data platform : findable, accessible, interoperable and reusable.
Enable large-scale pooled analyses with privacy-protecting features.

What about the data?

Population data is related to any identified or identifiable individual.

Individuals can be directly or indirectly identified from the population data processed.

Population data needs to consider (1) the content of information, (2) purpose of the processing, (3) impact or effect of that processing on Processingsome individuals.

Data needs to be truly anonymised before being shared or processed

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/

What is this workshop about?

Differences between individual person level data and other data

Explore issues with federated analysis and sensitive data

Experience Disclosive architectures

Simulate how the key ideas behind non-disclosive federated analysis

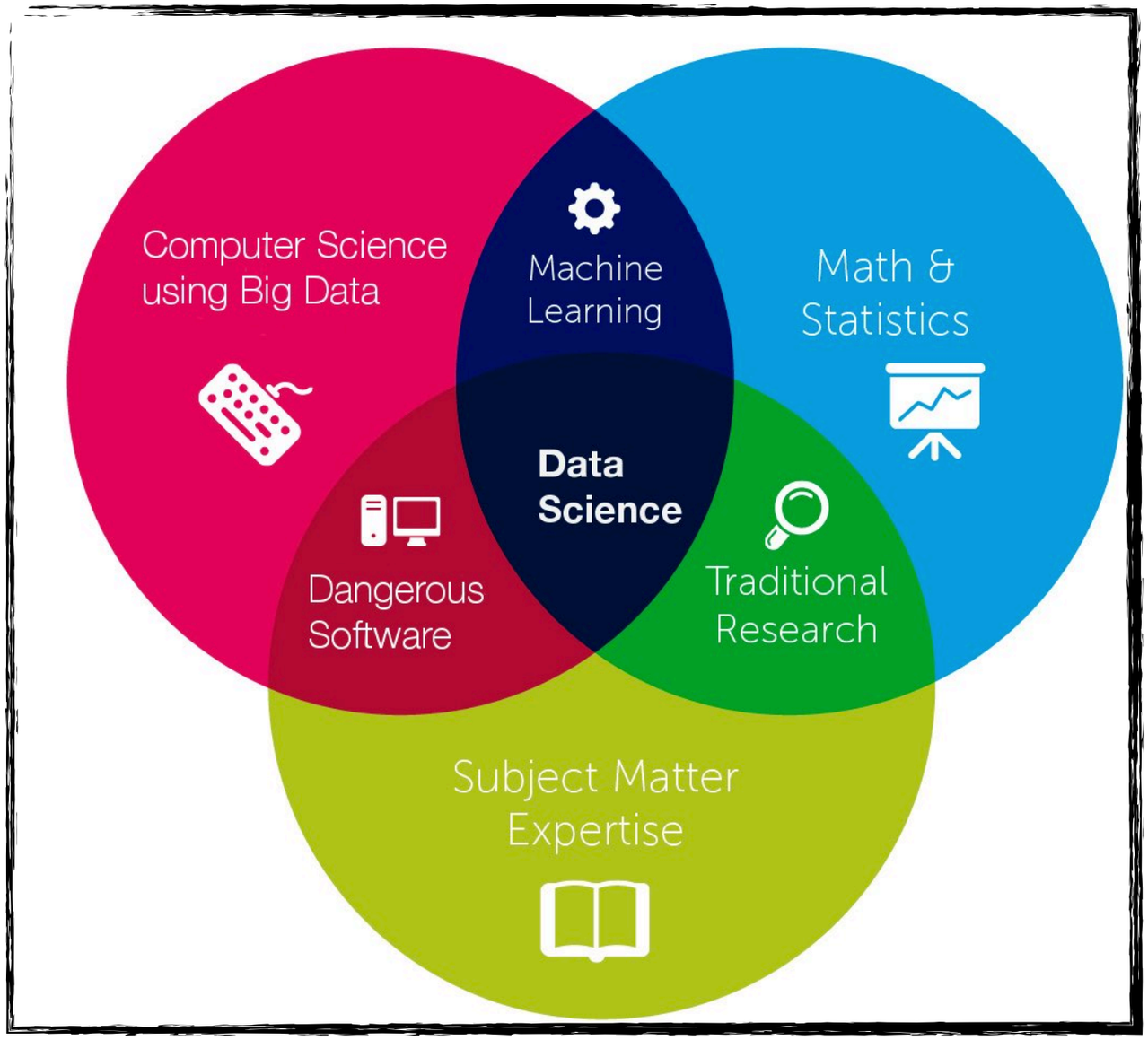
DataSHIELD architecture

DataSHIELD demonstrations

Let’s define ….

Data science, analytics, use of statistics ….

Ask the right questions, manipulate data sets, create some visualisation to communicate results.



e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Federated systems?

Collective agreement upon many standards

Heterogeneity in term of hardware, operating systems and network

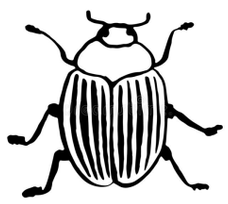
Independence and autonomy of each components of the federation.

Federated systems

Inter-operations of several distinct networked computer systems

Data distributed over multiple databases

Distribution of analysis



Examples of federated systems

A system, made of interconnected components, for the purpose of weather forecast.

Some data are being collected from people's mobile phones for commercial purposes.

Sharing population data or personal data. Personal information and patterns can help identify individuals.

Data brings many responsibilities …

https://www.ashurst.com/en/news-and-insights/legal-updates/data-bytes-12---m-and-a-and-data-protection-part-1/

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Disclosive analysis…

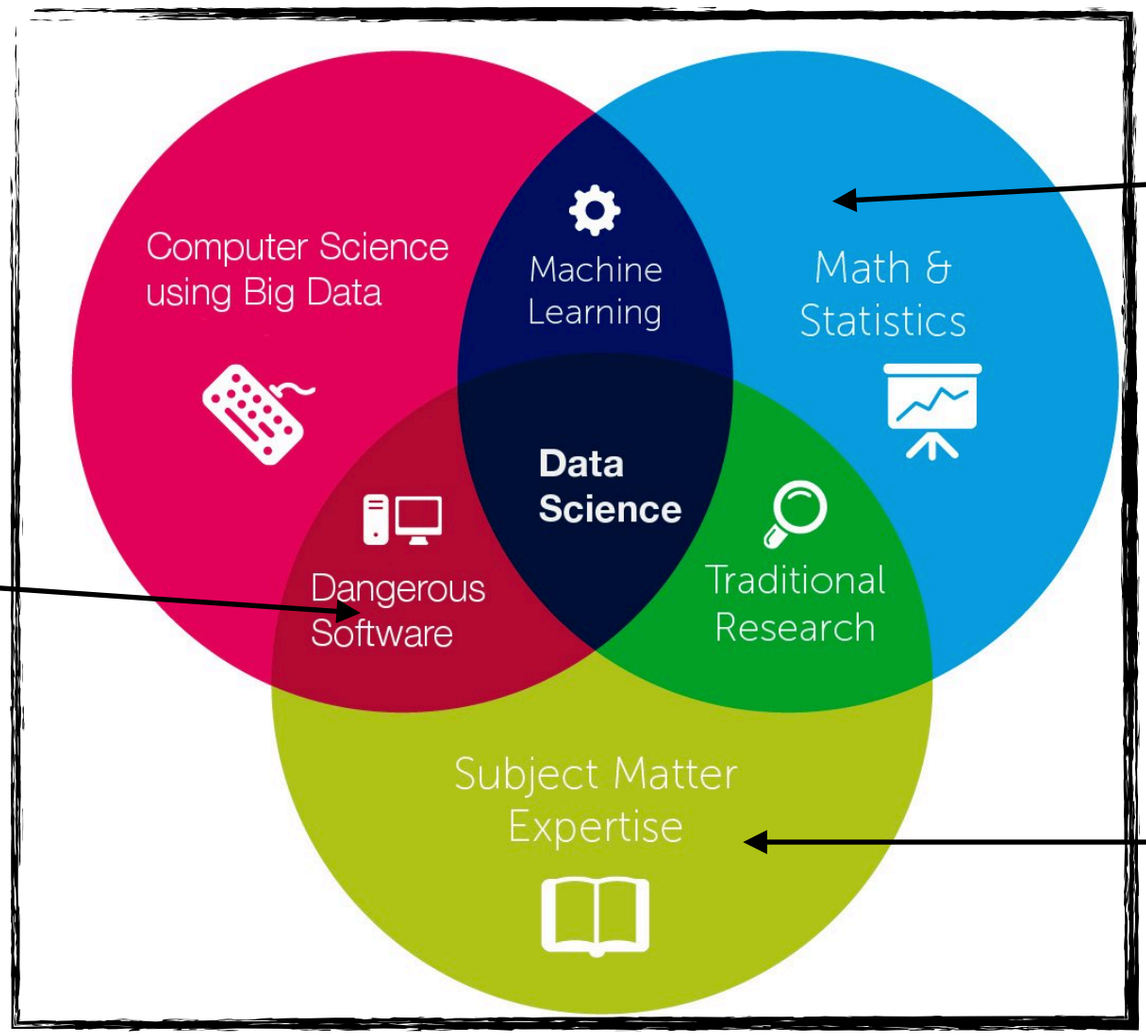
Provides information that violates any of the legal, contractual or ethical undertakings.

The data owner has entered into an agreement with third parties. Individuals can be identified.

Makes use of data without those agreements under which sensitive or personal information can be shared.

Data science, analytics, use of statistics ….

Ask the right questions, manipulate data sets, create some visualisation to communicate results.



disclosure of data

disclosure of data

disclosure of data

disclosure of data

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Non-disclosive analysis…

Provides information that respects all of the legal, contractual or ethical undertakings.

The data owner has entered into an agreement with third parties. No individual data should be inferred.

Makes use of data with those agreements under which sensitive or personal information can be shared.

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Different people, different skills

Data custodians and governance

Data analysts or data scientists

Developers

Data custodians and governance

- ☑ Controlled and authorised access to the data
- ☑ Data integrity is sustained
- ☑ Data content and changes are audited
- ☑ Technical controls safeguard data

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Data custodians and governance

- ☑ Controlled and authorised access to the data
- ☑ Data integrity is sustained
- ☑ Data content and changes are audited
- ☑ Technical controls safeguard data

Data analysts or data scientists

- ☑ Prevent revealing confidential and damaging information related to individuals
- ☑ Prevent revealing results that uses a small amount of data
- ☑ Prevent revealing information that can be combined with other datasets

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Data custodians and governance

- ☑ Controlled and authorised access to the data
- ☑ Data integrity is sustained
- ☑ Data content and changes are audited
- ☑ Technical controls safeguard data

Data analysts or data scientists

- ☑ Prevent revealing confidential and damaging information related to individuals
- ☑ Prevent revealing results that uses a small amount of data
- ☑ Prevent revealing information that can be combined with other datasets

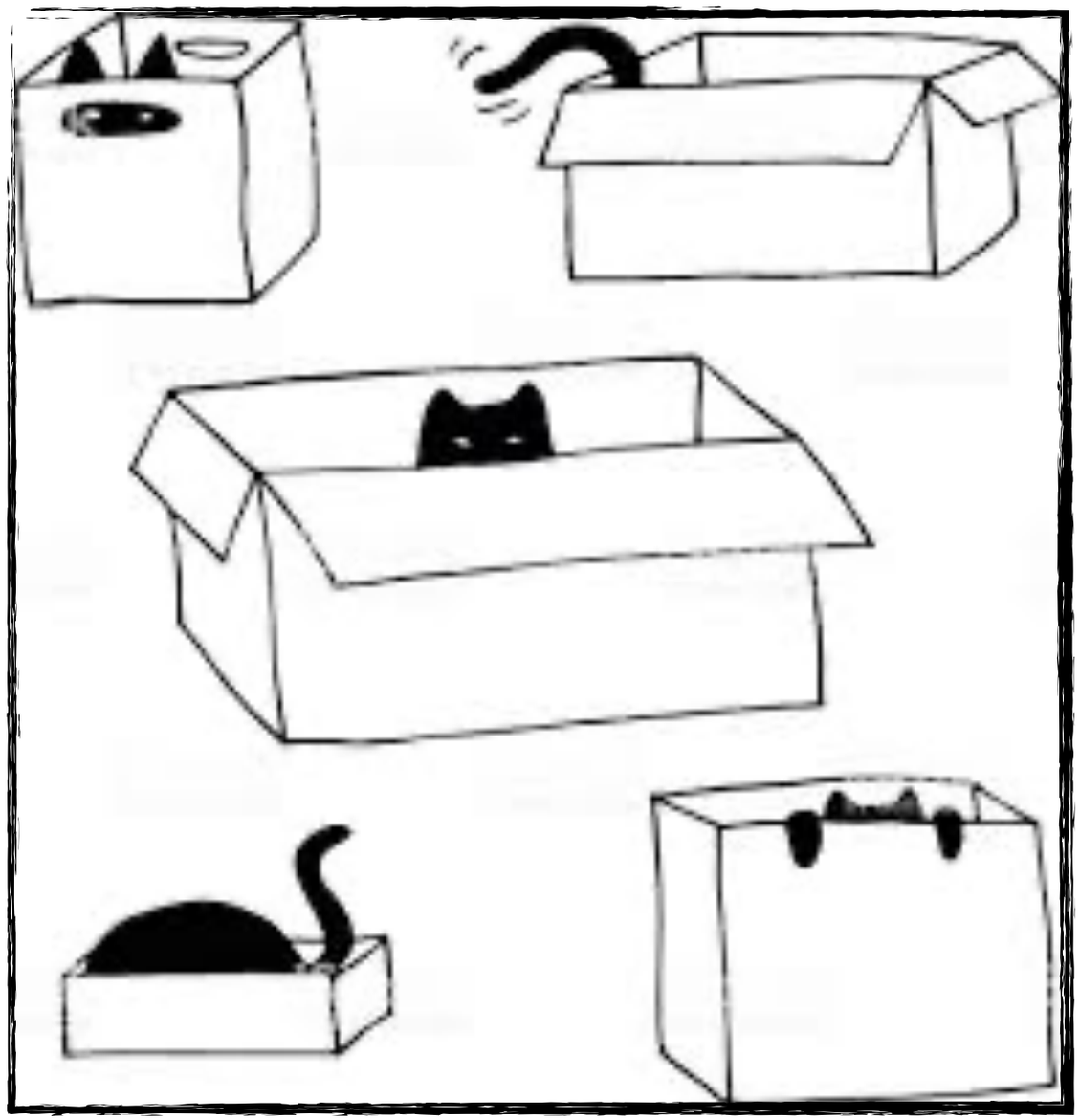
Developers

- ☑ Secure infrastructure that store data and monitor access
- ☑ Secure communication protocols
- ☑ Implement statistical control in statistical methodologies and visualisation tools
- ☑ Produce some tests that continuously assess these criteria

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Experiencing disclosure
And preventing it



e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Many forms of disclosure

Identity

Information about individual people we do not need to be aware of.

Providing access to the data

Client-Server Coding

Number of constraints on type of analysis to complete

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Some data

List of books with their authors, titles, approximated number of words, a Good Read score and their year of publication.

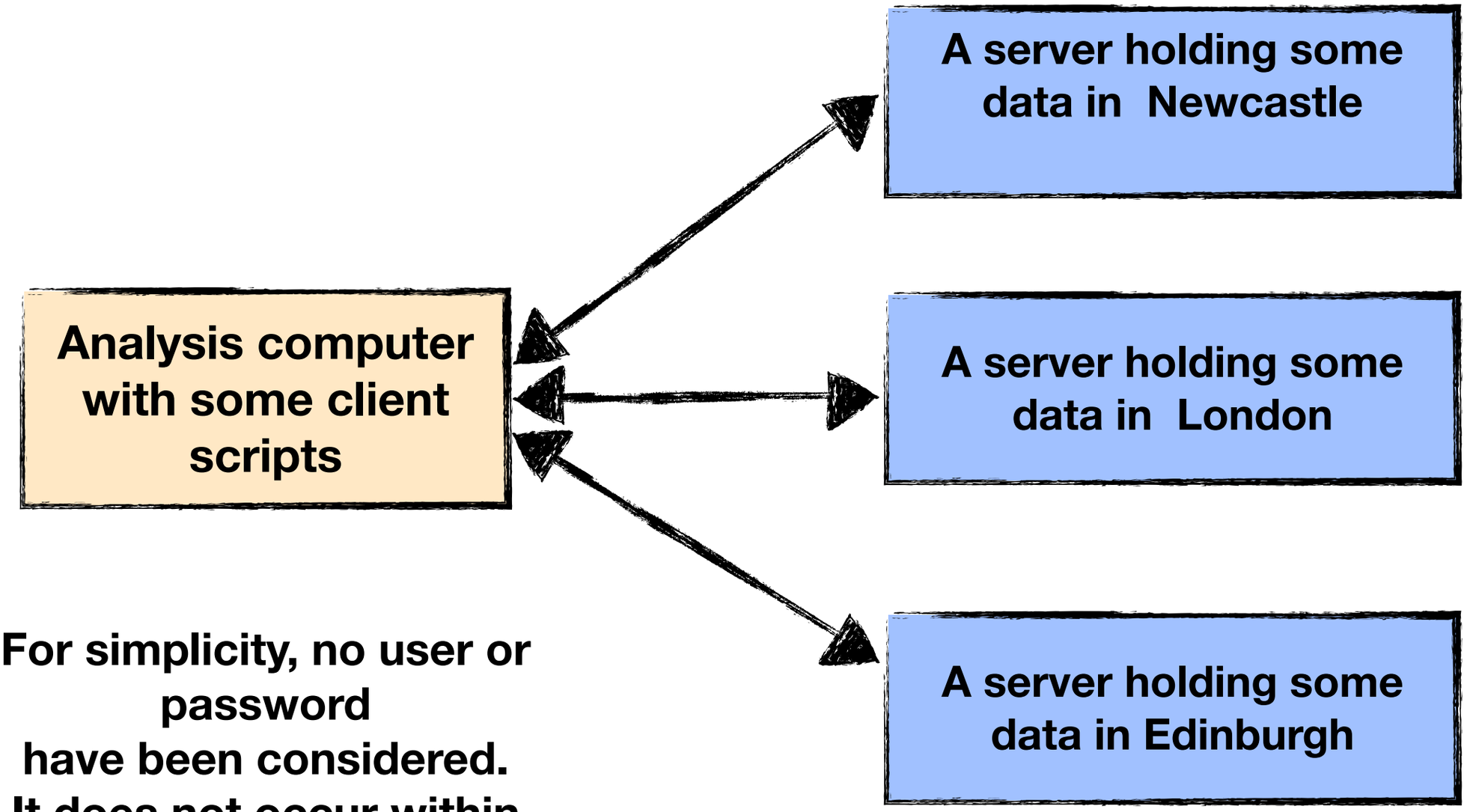
TITLE	AUTH	GR	WDS	YEAR
The Magnificent Ambersons	Booth Tarkington	3.73	98954	1918
Main Street	Sinclair Lewis	3.73	159901	1920
Under the Net	Iris Murdoch	3.75	75600	1954
To the Lighthouse	Virginia Woolf	3.75	69327	1927
The Prime of Miss Jean Brodie	Muriel Spark	3.76	45000	1961

https://www.goodreads.com/list

Simulation of ….

Client side

Server side



For simplicity, no user or password have been considered. It does not occur within DataSHIELD.

Purpose of the simulation

To simulate in a simple manner how disclosure of data can occurs
To simulate some techniques used to prevent the disclosure of data

Click on this link

https://github.com/patRyserWelch8/DataSHIELD.eRUM.2020

We will be back in this room in 40 minutes to review our finding and exploring how DataSHIELD implements these ideas

Break out room no 1 :

Analysing data

Break out room no 2 :

Omics

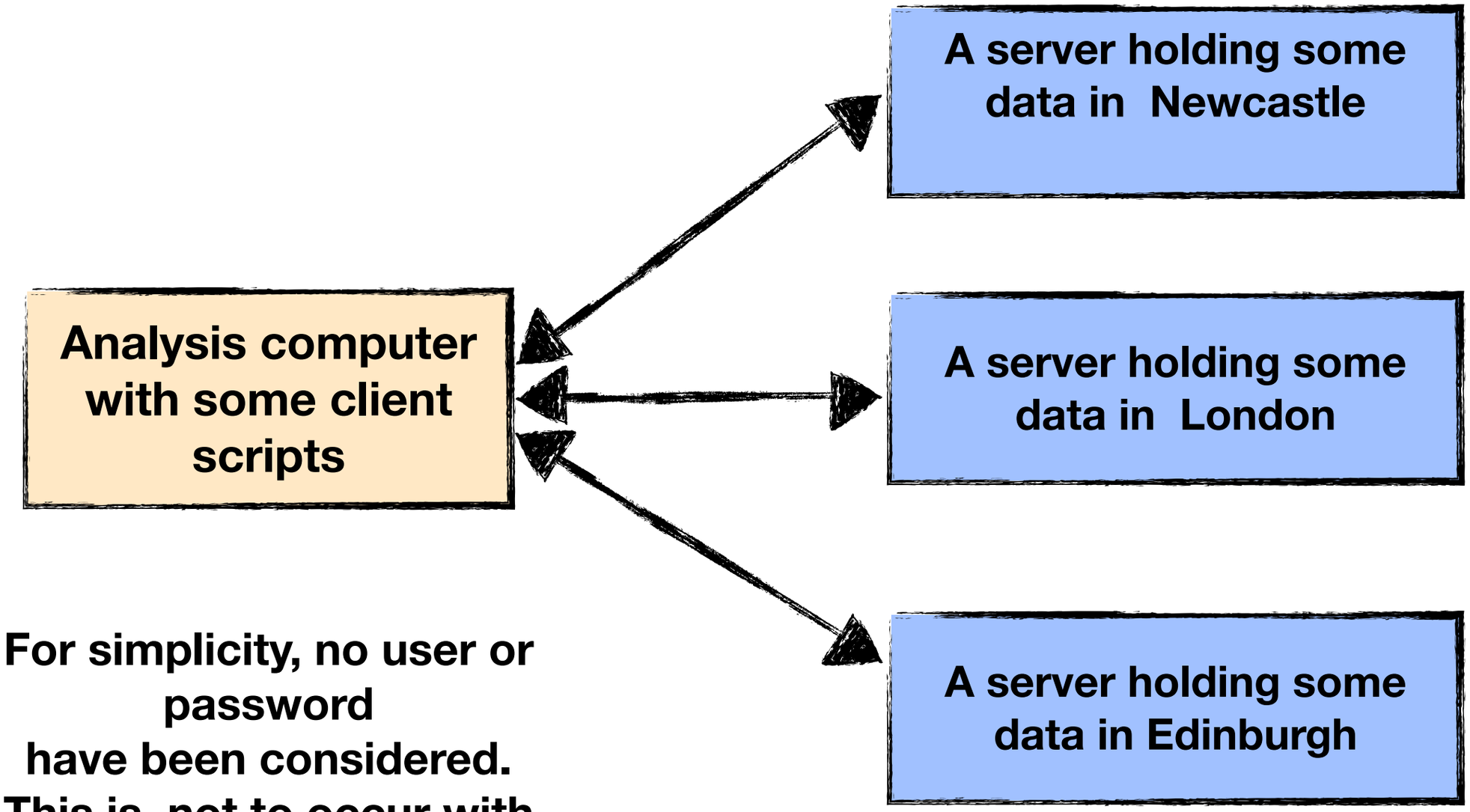
Break out room no 3 :

DataSHIELD developer

Simulation of ….

Client side

Server side

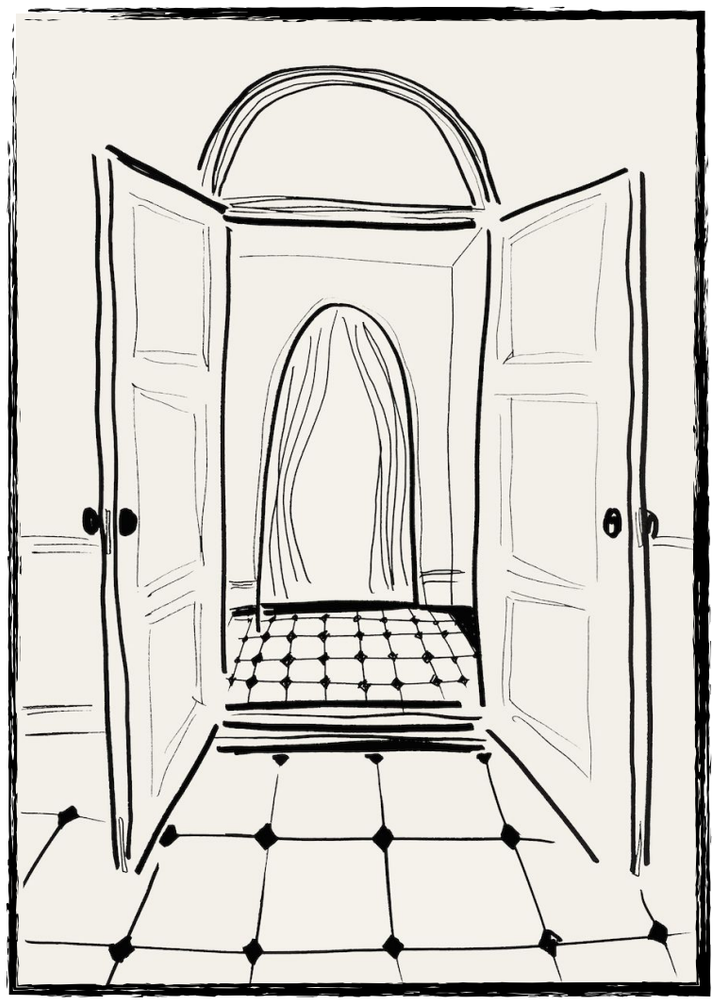


For simplicity, no user or password
have been considered.
This is not to occur with DataSHIELD.

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Providing access to the data



https://github.com/patRyserWelch8/DataSHIELD.eRUM.2020/wiki/Disclosive-simulation

Retrieve the data

Analysis computer with some client scripts

```
print("----  retrieve some of the data from the servers and display them -----")
print(connections$servers[["Newcastle"]]$datasets[["classic"]]$data)
```

```
[1] "----  retrieve some of the data from the servers and display them -----"
# A tibble: 30 x 5
   Title                                  Author          GreatReadScore  Words YearPub
   <chr>                                  <chr>                    <dbl>  <dbl>   <dbl>
 1 Heart of Darkness                      Joseph Conrad             3.39  53285    1899
 2 Wide Sargasso Sea                      Jean Rhys                 3.53  49665    1966
 3 Loving                                 Henry Green               3.55  67200    1945
 4 The Secret Agent                       Joseph Conrad             3.57  89230    1907
 5 Zuleika Dobson                         Max Beerbohm              3.58  75600    1911
 6 A Portait of the Artist as a Young Man James Joyce               3.58  84922    1916
 7 Sons and Lovers                        D.H. Lawrence             3.58 196200    1913
 8 Lord Jim                               Joseph Conrad             3.59 127949    1900
 9 The Way of All Flesh                   Samuel Butler             3.6   96000    1903
10 Lord of the Flies                      William Golding           3.61  59900    1954
```

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Identity

https://github.com/patRyserWelch8/DataSHIELD.eRUM.2020/wiki/Anonymisation-and-synthetic-data

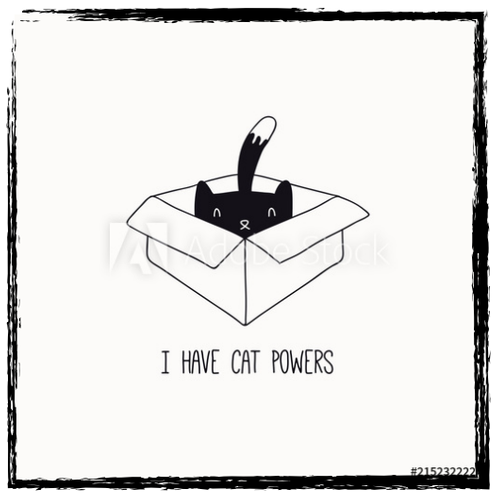
What did we do?

Remove the name of the author

A simple search using the title can identify the author from the year of publication and the title of the book.

Remove the title

It became much more challenging to infer the author and title from the data provided



Retrieve the data

Analysis computer with some client scripts

```
   GreatReadScore  Words YearPub
            <dbl>  <dbl>   <dbl>
 1           3.39  53285    1899
 2           3.53  49665    1966
 3           3.55  67200    1945
 4           3.57  89230    1907
 5           3.58  75600    1911
 6           3.58  84922    1916
 7           3.58 196200    1913
 8           3.59 127949    1900
 9           3.6   96000    1903
10           3.61  59900    1954
# … with 20 more rows
```

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Synthetic data

Synthetic data is data that contains all the characteristics of production minus the sensitive content. Synthetic data is generally made in order to validate mathematical models. This data is used to compare the behavior of the real data against the one generated by the model.

https://www.riaktr.com/synthetic-data-become-major-competitive-advantage/

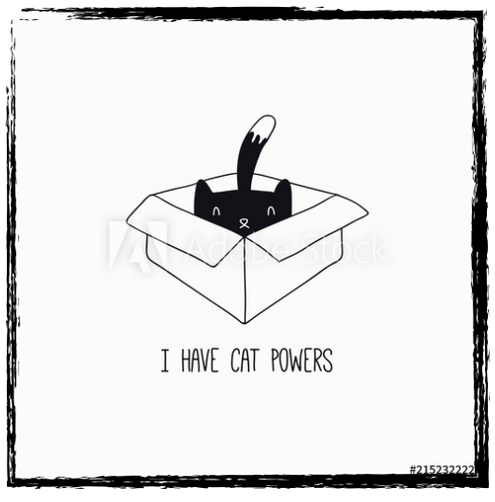
https://medium.com/@mubarakb/synthetic-data-a01c820fc6c6

Original

GreatReadScore	Words	YearPub
3.73	98954	1918
3.73	159901	1920
3.75	75600	1954
3.75	69327	1927
3.76	45000	1961

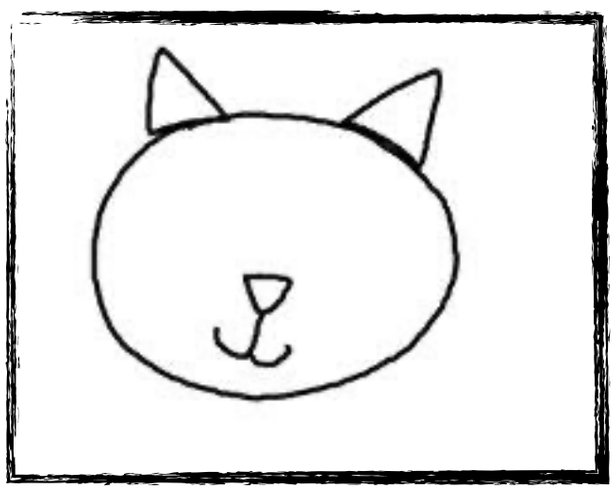
Normal distribution

GreatReadScore	Words	YearPub
3.749748	69015	1934
3.752870	134470	1942
3.742186	70379	1943
3.786993	102304	1937
3.738076	124322	1916



Multivariate

GreatReadScore	Words	YearPub
3.754071	81651.31	1934.574
3.749341	89774.82	1948.968
3.762247	28326.59	1968.638
3.759437	74133.90	1945.696
3.736903	133219.84	1910.132



O. Is original data N.D is normal distribution M. Multivariate

Dataset	Field	Min	Mean	S.d	Max	1st Q.	Median	IQR	3rd Q
O.	GR	3.73	3,744	0.01	3.760	3.73	3.75	0.02	3.75
	Words	45,000	89,756	43,662.8	1,599,901	639,327	75,600	29,627	98,954
	Year	1918	1936	20.07	1961	1920	1927	34	1954
N.D	GR	3.738	3.754	0.02	3.787	3.742	3.750	0.01	3.753
	Words	69,015	100,098	30,093.5	134,470	70,379	102,304	53,943	124,322
	Year	1916	1934	10.92	1943	1934	1937	8	1942
M.	GR	3.737	3.752	0.01	3.762	3.749	3.754	0.01	3.759
	Words	28,327	81,421	37,500.3	133,220	74,134	81,651	15,641	89,775
	Year	1910	1942	22	1949	1935	1946	14.39	1949

Are there any correlations between these statistical variables?

Datasets	Good read score and no of words	Good read score and year	Year and no of words
Original Dataset 1	0.2601	0.1943	-0.2953
Original Dataset 2	0.2133	0.2913	0.0556
Original Dataset 3	-0.8682	0.8637	-0.7071
ND Dataset 1	-0.1149	0.0630	0.1205
ND Dataset 2	-0.9347	0.0780	-0.0393
ND Dataset 3	0.0563	0.3193	-0.3208
M. Dataset 1	0.0558	0.4060	-0.1522
M Dataset 2	0.3051	0.4580	0.1164
M Dataset 3	-0.9360	0.8733	-0.9430

-1 : Perfect negative slope

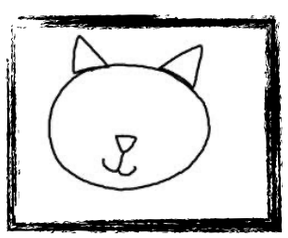
+1 : Perfect positive slope

Close to 0 : no linear correlation

Verdict ….

No sensitive content

The dataset has is made of some observations. Many of them have very little correlations….
Generating some synthetic data is more difficult than reproducing some data using some distributions….



All characteristics of production are not met:

- **Not enough data was used to compute some appropriate statistics**
- **The spread of the data does not matches the orginal data**
- **The range of values vary a lot between dataset and**
- **Some relationships between the fields have yet to be fully replicated**
- **Any mathematical models is unlikely to be validated correctly**

Some synthetic data suggested some male patients were incidentally expecting ….

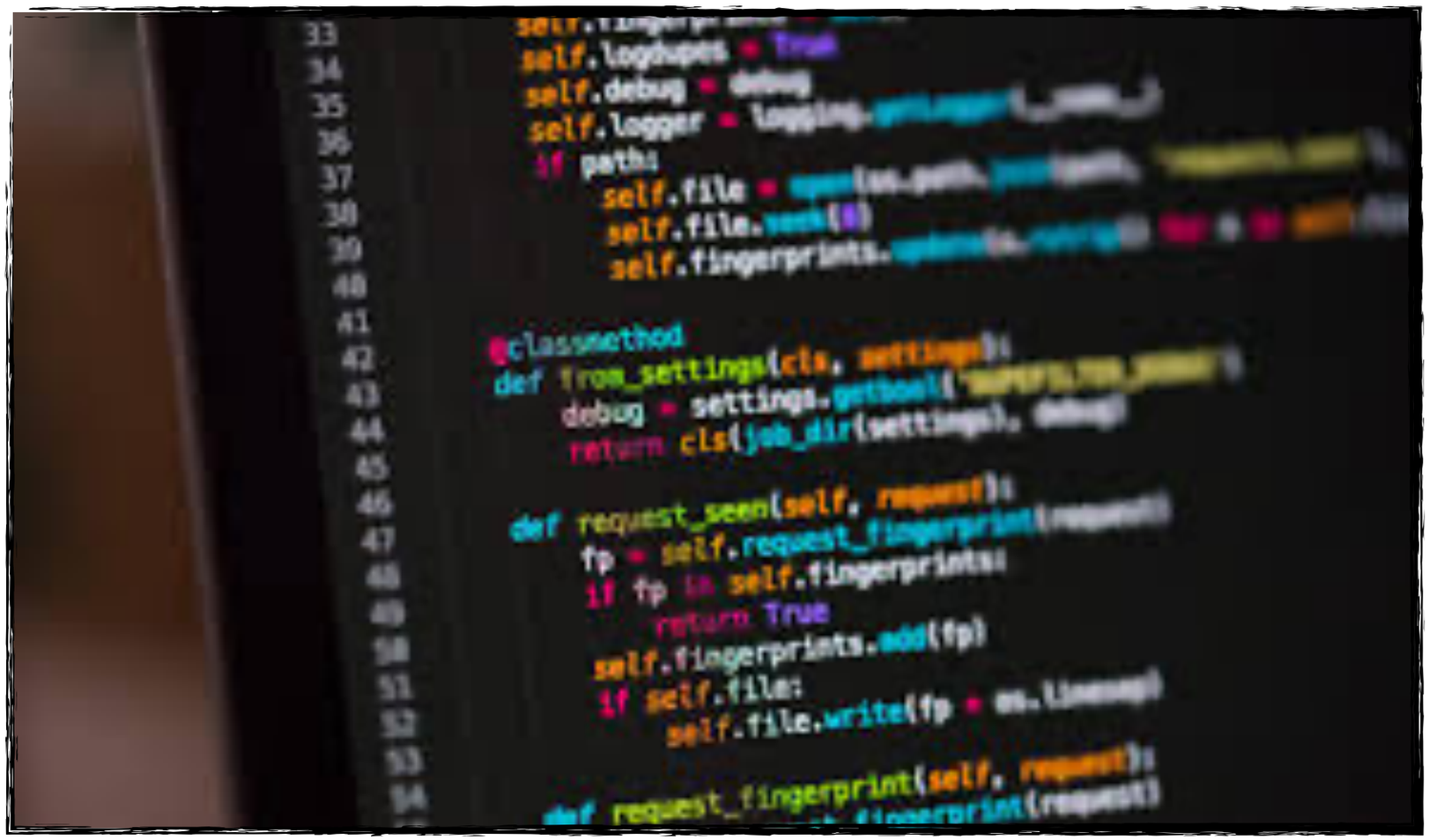
Some people died several times, in different places from different diseases …

Some pregnancies lasted 18 months…

e-Rum2020
17/20 JUNE
MILANO

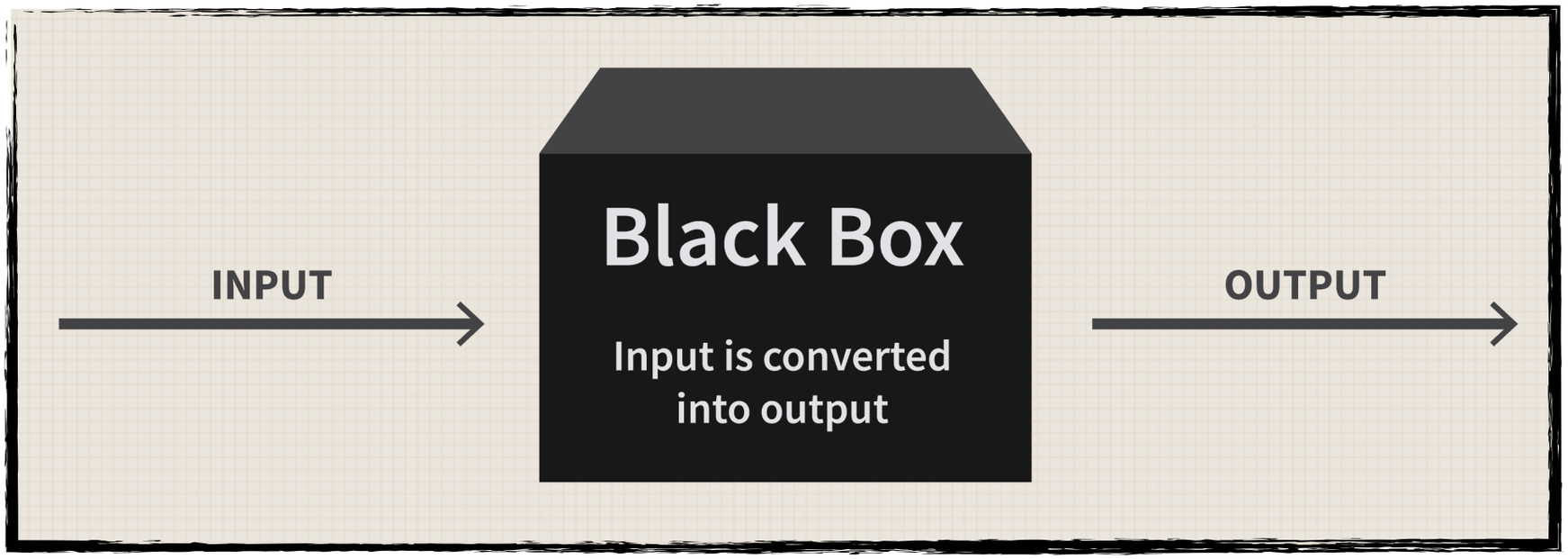
Newcastle University

Coding for non-disclosure of data



Black-box and parser

The access to the datasets is made available through the meta data of the data sets and some remote function calls. The client script or function must not have access to any data directly, but can only obtained some results of some computations.



Let's review the Server class

Data computer Newcastle

Previously

```
Server  <- R6Class("Server", list(
            datasets = NULL,
            initaliaze = function()
            {
              self$datasets <- list()
            },
            upload = function(meta.data, data, name)
            {
              new.dataset <- DataSet$new(meta.data, data)
              self$datasets[[name]] <- new.dataset
            }

  ))
```

Altered

```
private = list
(
    .datasets = NULL,
    .current = NULL
),
```

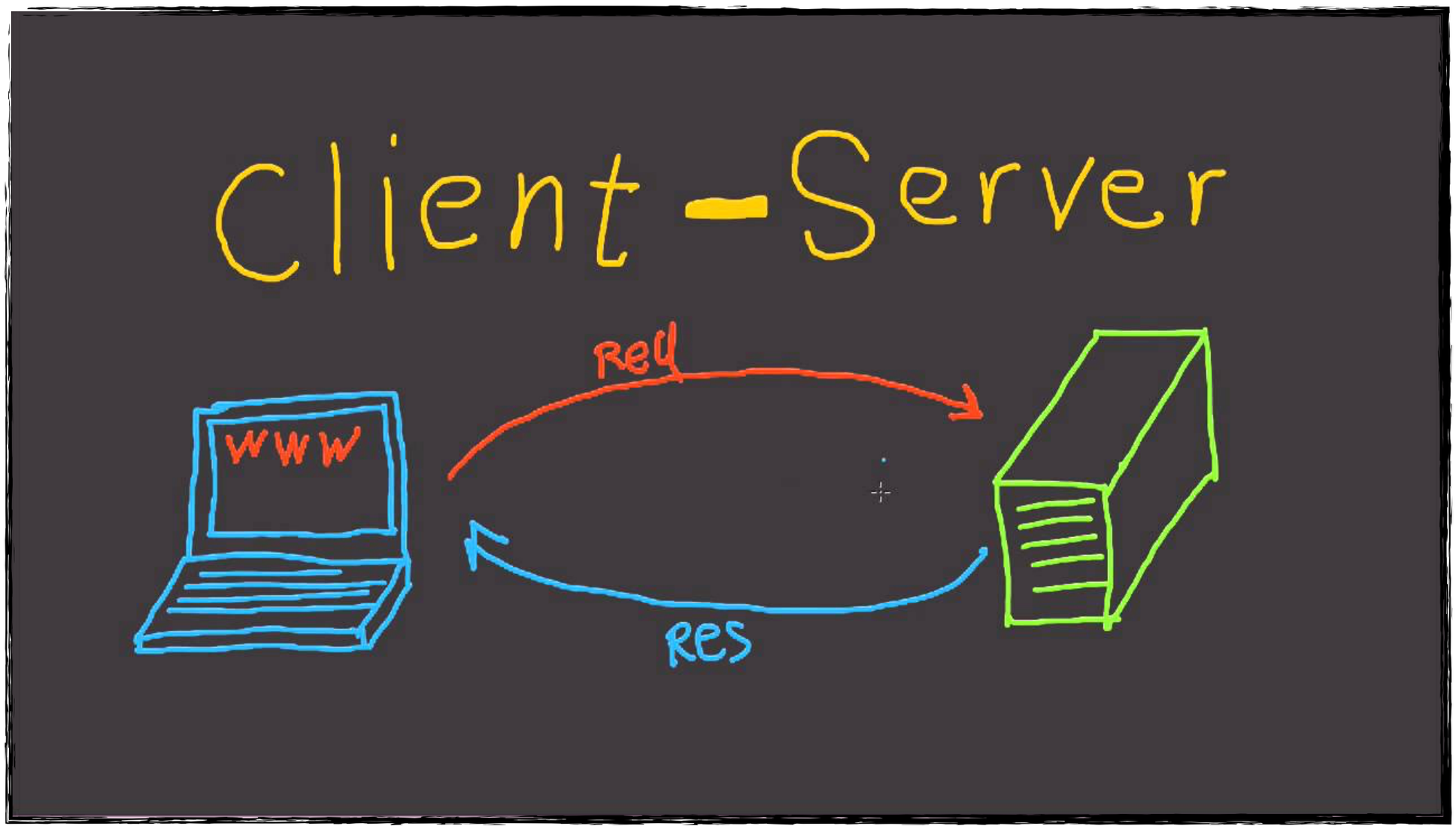
```
public = list
(
  initialize = function()
  {⟷},
  upload = function(meta.data, data, name)
  {⟷},
  set_dataset = function(dataset.name)
  {⟷},
  server.ls = function()
  {⟷},
  server.dim = function()
  {⟷},
  server.mean = function(variable)
  {⟷},
  server.sd = function(variable)
  {⟷},
  server.min = function(variable)
  {⟷},
  server.max = function(variable)
  {⟷},
  server.factor = function(variable)
  {⟷}
)
```

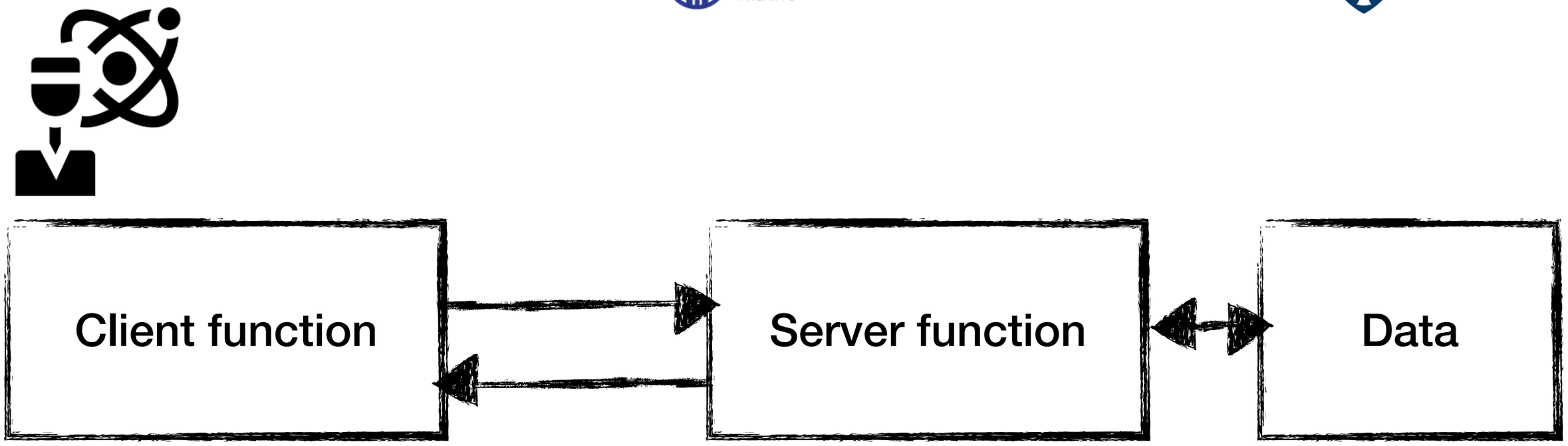
So...

```
> print(connections$servers[["Newcastle"]]$datasets[["classic"]]$data[,])
NULL
> print(connections$servers[["London"]]$datasets[["synth_classic"]]$data[,])
NULL
> print(connections$servers[["Edinburgh"]]$datasets[["synth_classic"]]$data[,])
NULL
>
```

But ...

```
> print(connections$get_server("Newcastle")$server.ls())
[1] "synth_classic" "classic"
> connections$get_server("Newcastle")$set_dataset("classic")
> print(connections$get_server("Newcastle")$server.dim())
[1] "GreatReadScore" "Words"          "YearPub"
> print(connections$get_server("Newcastle")$server.mean("Words"))
[1] 108710.5
> print(connections$get_server("Newcastle")$server.sd("Words"))
[1] 48438.31
> print(connections$get_server("Newcastle")$server.min("Words"))
[1] 49665
> print(connections$get_server("Newcastle")$server.max("Words"))
[1] 265391
>
```


Client-Server
WWW
Req
Res



ds.mean

server.sum

server.mean

server.length

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

Virtually-joined analysis

Generates the same results as if the data from all sources were physically transferred to a central warehouse and analysed jointly.

```
print(ds.mean(connections,"YearPub", combined =TRUE))
```

```
$combined
[1] 1936.594
```

Server-level analysis

Computes the analysis on each server.

```
print(ds.mean(connections,"YearPub", combined =FALSE))
```

```
$London
[1] 1940.682

$Newcastle
[1] 1927.7

$Edinburgh
[1] 1936
```

```
ds.mean <- function(connections = NULL, variable.name = NULL, combined = FALSE)
{
  stopifnot("Connection" %in% class(connections))
  stopifnot(is.character(variable.name))
  stopifnot(is.logical(combined))

  if(combined)
  {
    return(.combined.mean(connections,variable.name))
  }
  else
  {
    return(.split.mean(connections,variable.name))
  }

}
```

Statistis	Values
Factors	3.73, 3.75, 3.76
Minima	3.73
Maxima	3.76
Arithmetical mean	3.744
Standard deviation	0.01341641
Length	5

```
> print(mean(c(3.73,3.75,3.75,3.76,3.76)))
[1] 3.75
> print(mean(c(3.73,3.75,3,75,3.76,3.76)))
[1] 15.5
> print(mean(c(3.73,3.73,3.75,3.75,3.76)))
[1] 3.744
> print(sd(c(3.73,3.73,3.75,3.75,3.76)))
[1] 0.01341641
```

We were able to reconstruct a dataset using some inference

Client function

Server function
server.mean
server.sum

Data

ds.mean

ds.factor

Server Threshold
min.length (server.mean)
max.factor (server.factor)

Remove server.min and server.max

Analysis computer with some client scripts

```
> print("--- mean Great read score ---")
[1] "--- mean Great read score ---"
> print(ds.mean(connections,"GreatReadScore", combined =TRUE))
$London
[1] "total: 258.56"     "observations : 66"

$Newcastle
[1] "total: 109.3"      "observations : 30"

 $Edinburgh
 [1] "server warming: disclosive call"
```

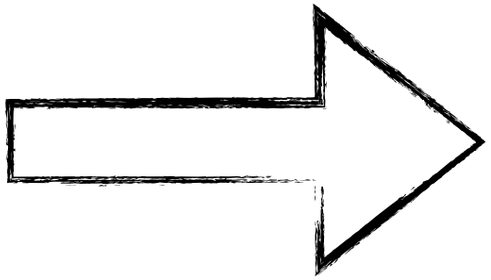
```
> print(ds.factor(connections,"GreatReadScore", combined =FALSE))
$London
[1] "server warming: disclosive call"

$Newcastle
[1] "server warming: disclosive call"

$Edinburgh
[1] "server warming: disclosive call"
```

Analysis computer with some client scripts

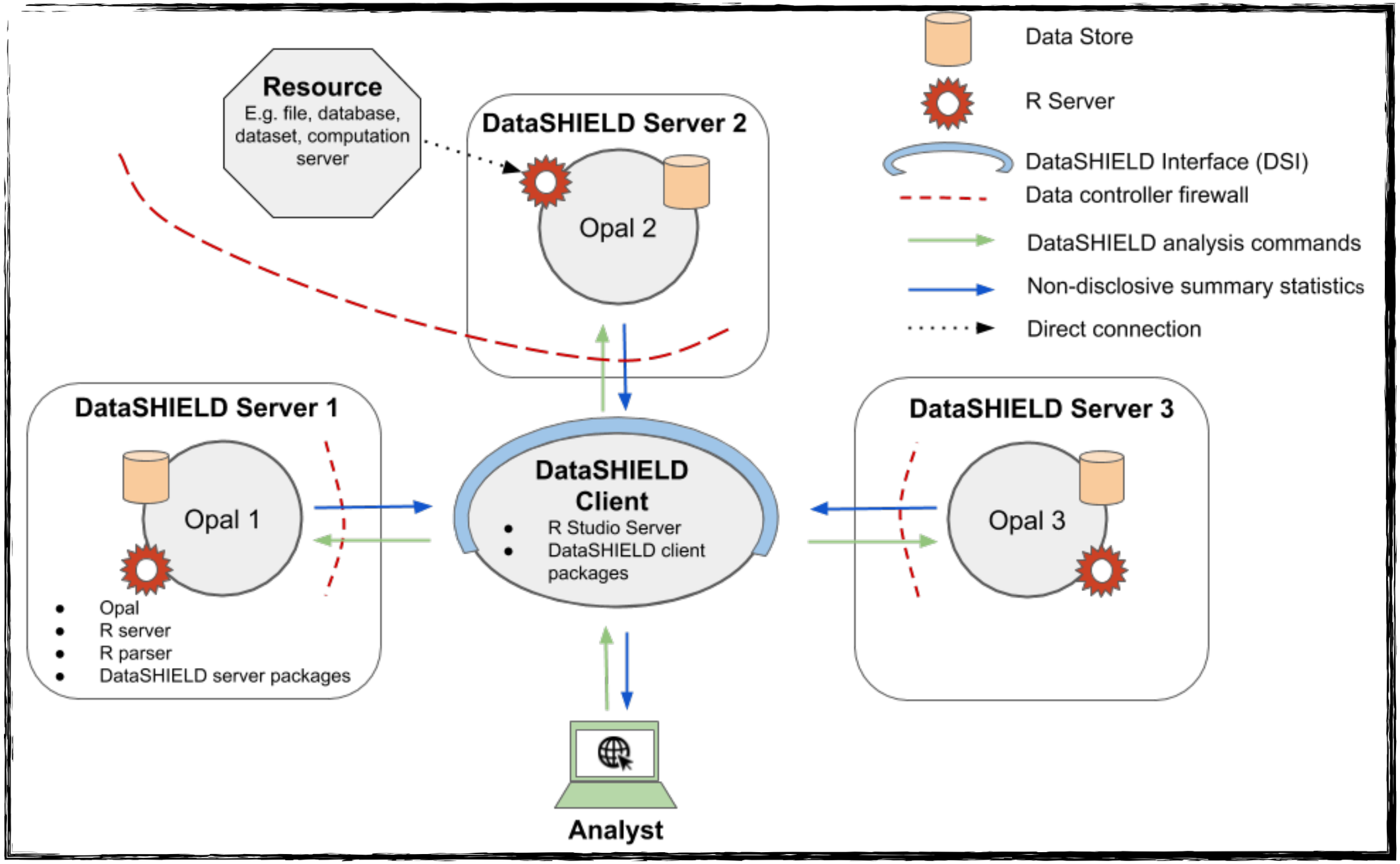
```
[1] "--- factor ---"
> print(ds.factor(connections,"Type", combined =TRUE)
$London
[1] "Electronic" "Hardback"   "PaperBack"

$Newcastle
[1] "Electronic" "Hardback"   "PaperBack"

$Edinburgh
[1] "Server warning: disclosive call"

$combined
[1] "Electronic" "Hardback"   "PaperBack"
```

What have we done?

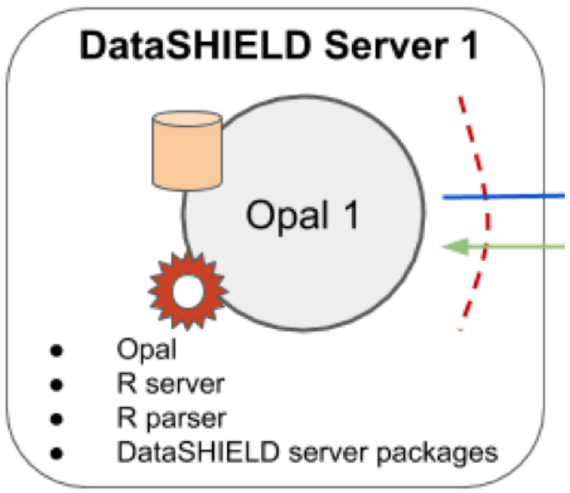
1) No direct access to data
2) Used a parser to constraint the number of analysis to complete
3) Limited data inference by using some thresholds controlled by the data governance.

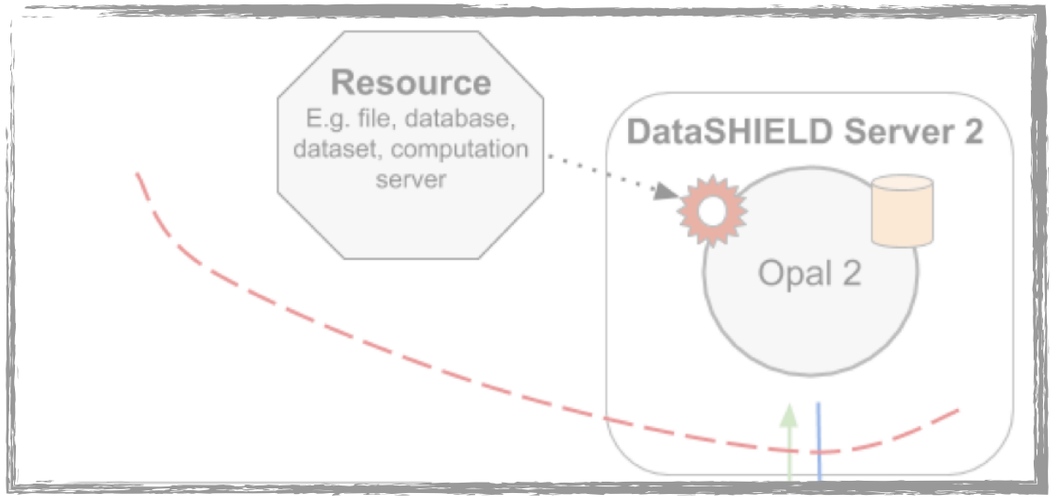


DataSHIELD Architecture



Inside the Opal Server: Data Warehousing





Traditional databases:

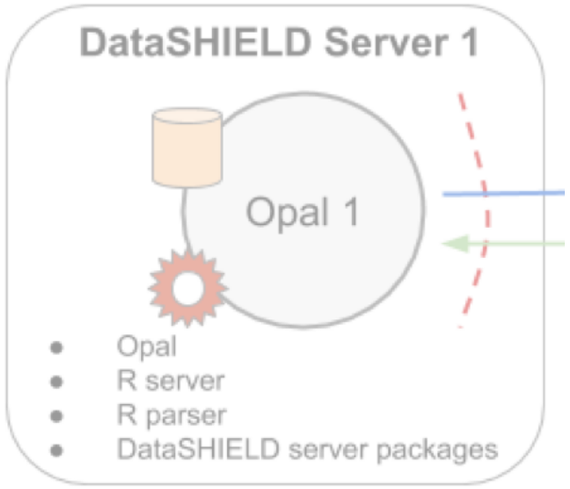
- Open-sourced document database
- Open-sourced RDBMS
- Table-format
- Opal database

More contemporary databases (Resources):

- Genomics data
- Genes databases
- Cloud-based storage
- Commands to some existing systems (SSH)
- Any organisation-specific data storage or system.

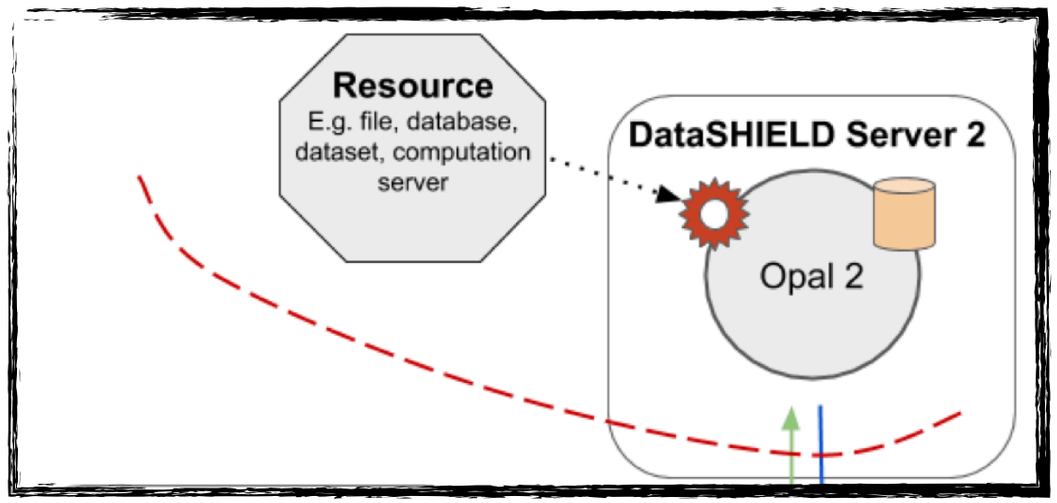
https://www.obiba.org/pages/products/opal/

Inside the Opal Server: Data Warehousing



Traditional databases:

- Open-sourced document database
- Open-sourced RDBMS
- Table-format
- Opal database

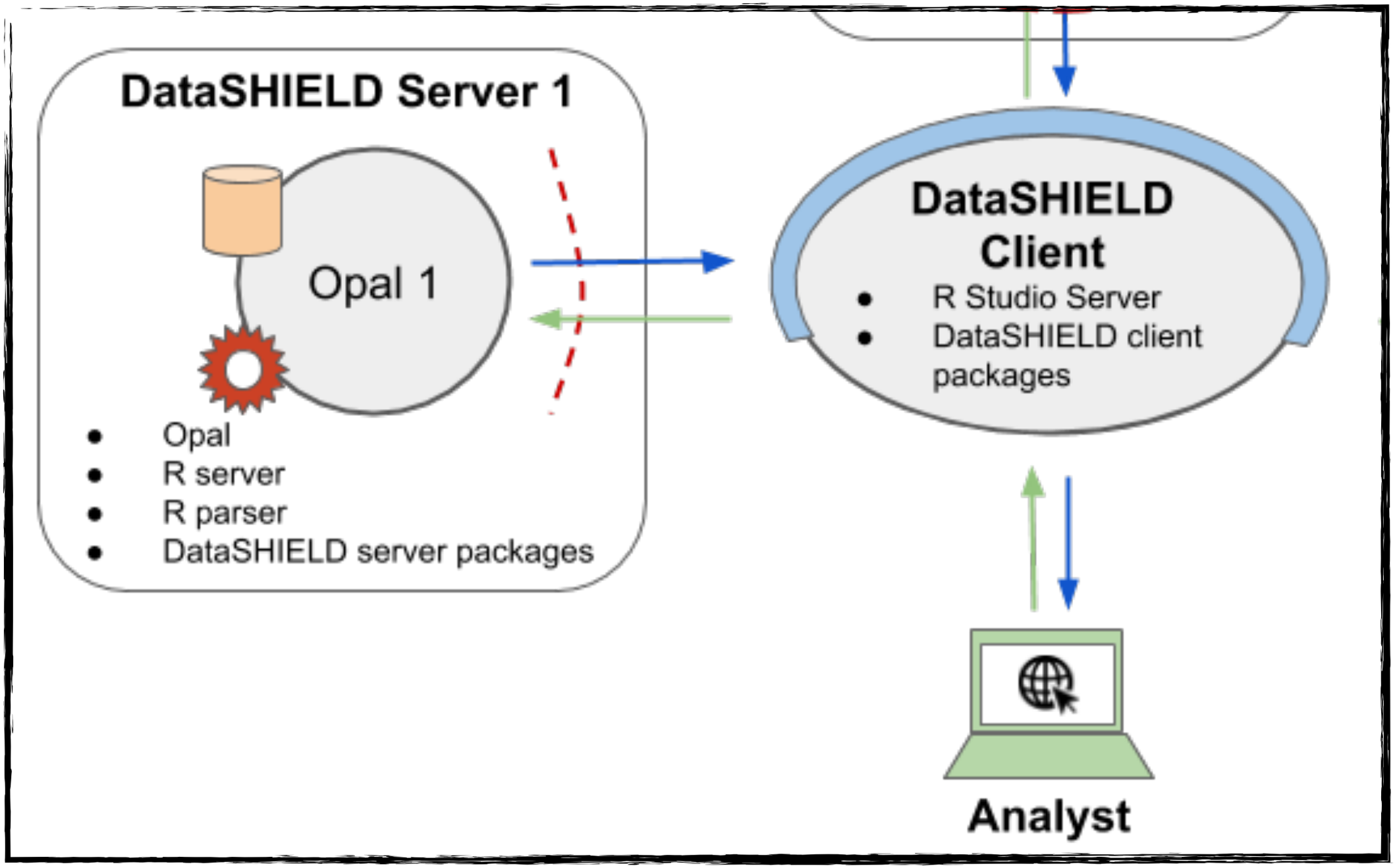


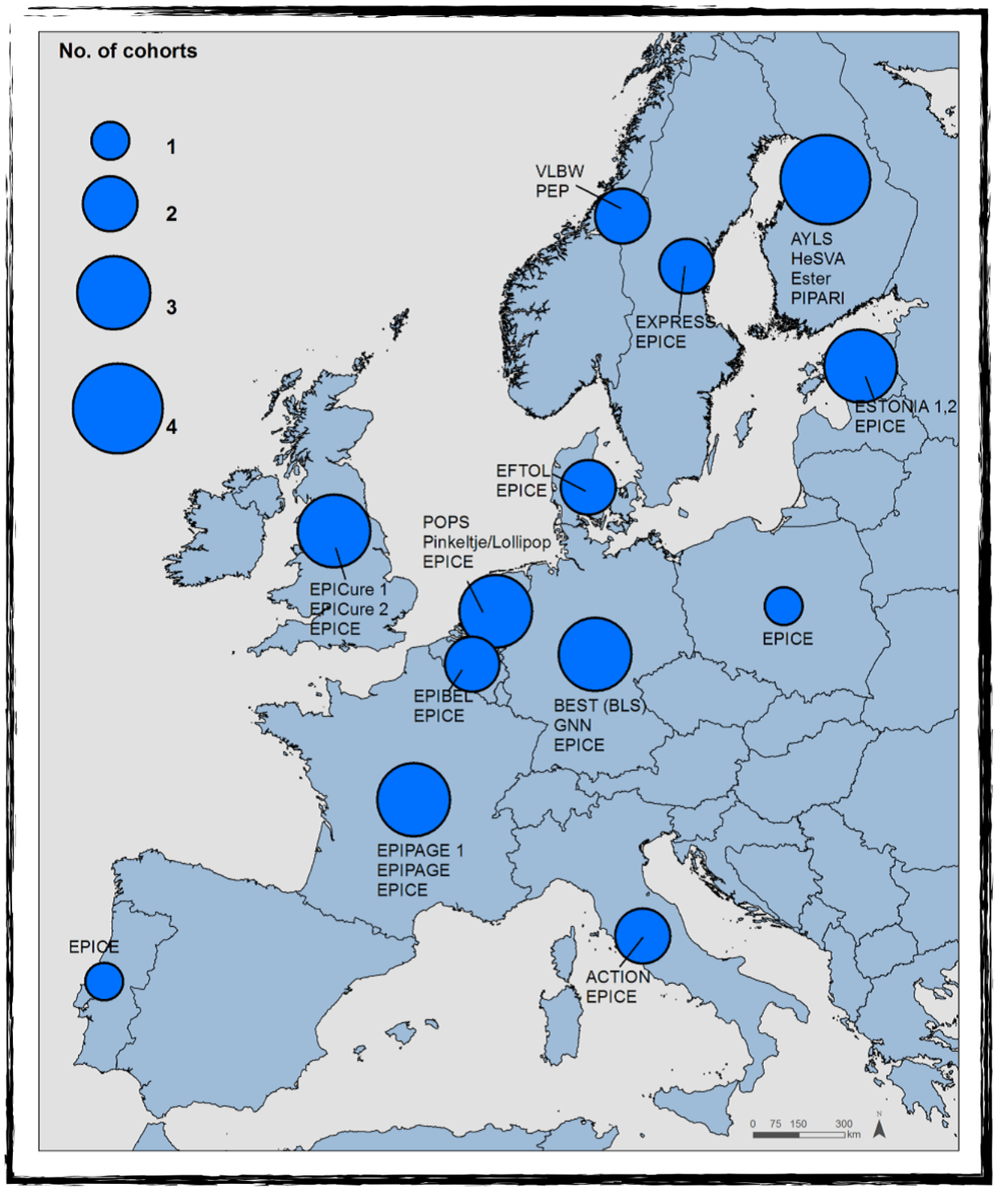
More contemporary databases (Resources):

- Genomics data
- Genes databases
- Cloud-based storage
- Commands to some existing systems (SSH)
- Any organisation-specific data storage or system.

https://www.obiba.org/pages/products/opal/

RESTful call of server functions





PRECAP preterm platform

- 20 different cohorts scattered around 13 countries

e-Rum2020
17/20 JUNE
MILANO

Newcastle University

To conclude….

DataSHIELD lowers the possibility of disclosure by bringing the computations to the data.

DataSHIELD provides some thresholds, so that data governance can control the parametrisation of the federated and server-level analysis.

DataSHIELD architecture prohibits the transfer of data, unless they are the results of some computations.

DataSHIELD enables analysis of data across countries and organisations, without transferring actual datasets across from one computer systems.

DataSHIELD has the new capability to connect and analyse a variety of datasets and databases systems using some resources.


KEEP
CURIOUS
AND
DEMO
TIME
KeepCalmAndPosters.com