The case for federated analysis: A DataSHIELD perspective

Dr Becca Wilson
PI DataSHIELD research project
University of Liverpool

Workshop: Safe access to sensitive research data
November 25, 2022

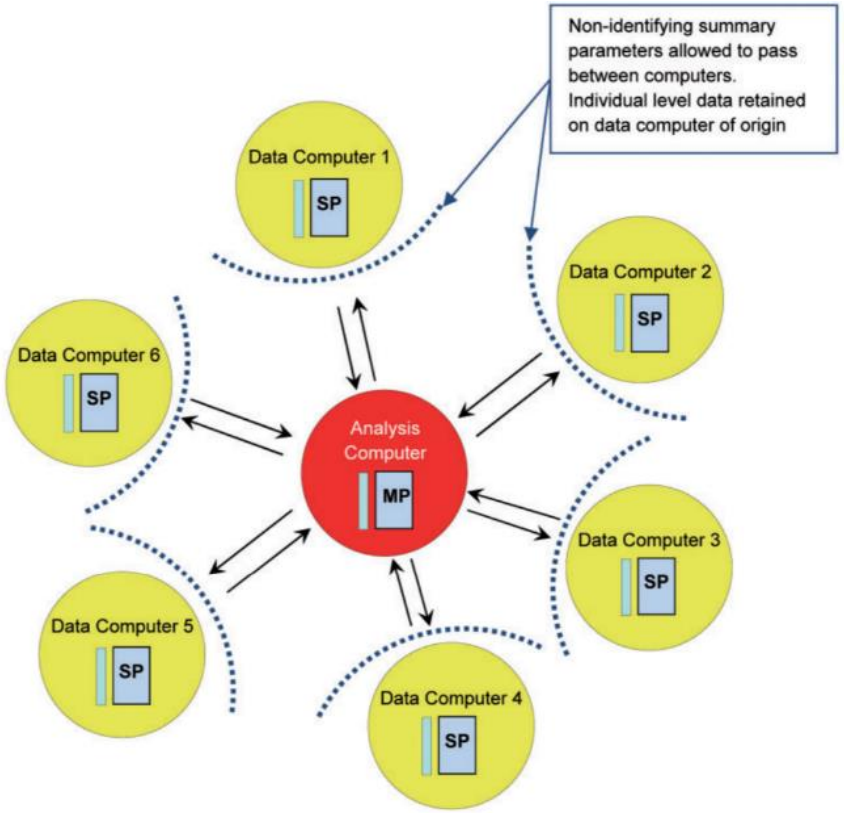
@DrBeccaWilson
datashield@liverpool.ac.uk
@DatashieldNews

Some fundamentals

- Data custodians must select solutions most appropriate for their data context
- No solution technical or non-technical can eliminate the risk of disclosure
  - Deductive disclosure, reidentification of anonymised data, data misuse, human error
- If data is so sensitive that elimination of these risks is required, these data should NOT be made available
- There is a three-point balance when mitigating disclosure
  - **the real risk** of disclosure
  - **the real costs** associated with implementation
  - **the real impact** on participants and utility/analysis

Taking the analysis to the data

- Co-built by the DataSHIELD Research Project (UK) and OBiBa (Canada)
- Federated analysis of data sets simultaneously, linked by non-disclosive summary statistics
- DataSHIELD is Open Source "takes the analysis to the data"
- We can get our statistical power
- Enables the use of data in a usual study level meta-analysis & individual patient data meta-analysis
- Includes a variety of disclosure mitigations including SDC
- Established user base across longitudinal studies in Europe & with secondary care data (research hospitals and covid19), SME data access help drive innovation



https://datashield.org/about/publications

Gaye et al., 2014; Wilson et al, 2017; Marcon et al, 2021

How does DataSHIELD help mitigate disclosure?

1. Complementary to formal data access or data governance agreements including 5 Safes Framework: User management, due diligence, data access requests, contracts.
2. Infrastructure best practice: implemented on robust hardware, information transmissions securely transferred (https), pseudonymisation, not the primary/live database for a study.
3. DataSHIELD features:
   - only valid characters/functions (via the R Parser) from client side to the server side
   - analysis environment server side (R) only called via Opal or Molgenis
   - DataSHIELD server side functions block directly disclosive
   - 11 disclosure settings to check outputs for direct disclosure https://bit.ly/DS-function-checks
   - Disclosure setting thresholds are customised by the data controller for their study
   - Unique to DataSHIELD – analysts can not directly view the individual level data
4. User commands on DataSHIELD server are logged, can be interrogated manually
5. Data controllers and analysts responsible for maintaining hardware/software. Releases - https://datashield.org/forum
6. Open source community - DataSHIELD package developers responsible for updating/maintaining their packages

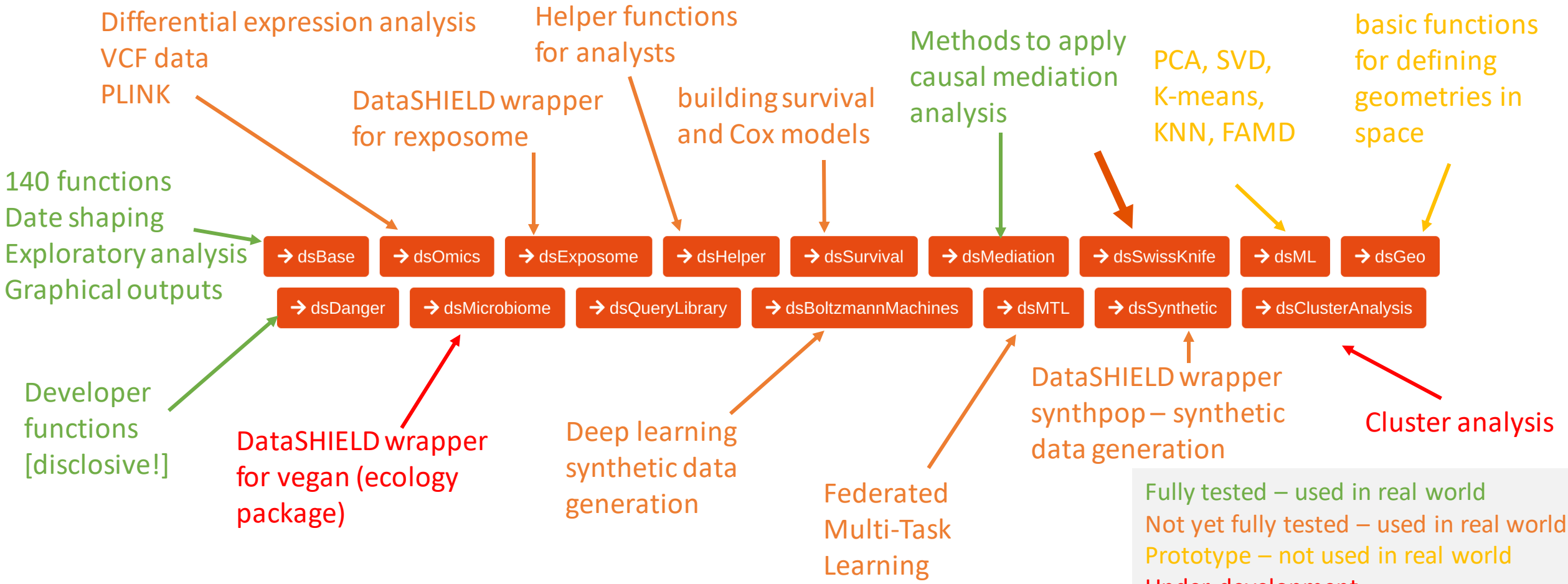
How does DataSHIELD help mitigate disclosure?

1. Complementary to formal data access or data governance agreements: User management, due diligence, data access requests, contracts. Researchers constrained by contractual and cultural expectations of employer
2. Infrastructure best practice: implemented on robust hardware, information transmissions securely transferred (https), not the primary/live database for a study.
3. DataSHIELD features:
   - only valid characters/functions (via the R Parser) from client side to the server side
   - analysis environment server side (R) only called via Opal or Molgenis
   - DataSHIELD server side functions block directly disclosive
   - 11 disclosure settings to check outputs for direct disclosure https://bit.ly/DS-function-checks
   - Disclosure setting thresholds are customised by the data controller for their study
   - Unique to DataSHIELD – analysts can not directly view the individual level data
4. User commands on DataSHIELD server are logged, only available to the data controller to interrogate manually
5. Data controllers and analysts responsible for maintaining hardware/software. Releases - https://datashield.org/forum
6. Open source community - DataSHIELD package developers responsible for updating/maintaining their packages

Comparison to common approaches

Feature	TRE/Safe Haven	Federated Data Network	DataSHIELD
Scalable, modular and interoperable infrastructure			✓
Data remains located with data controller		✓*	✓*
Manual disclosure checks on outputs (direct and inferential)	✓	✓	*
Real time direct disclosure checks			✓
Analyst can directly view individual patient data	✓	*	
Make use of individual patient data	✓	✓	✓
Analysis in real time	✓		✓
Utility for the researcher	High (1 study) Low (multiple)	Low	Medium

DataSHIELD R packages



@DrBeccaWilson
datashield@liverpool.ac.uk
@DatashieldNews

Who is using DataSHIELD?

EUCAN-Connect (€6M, 2019-2023)

173 cohort studies

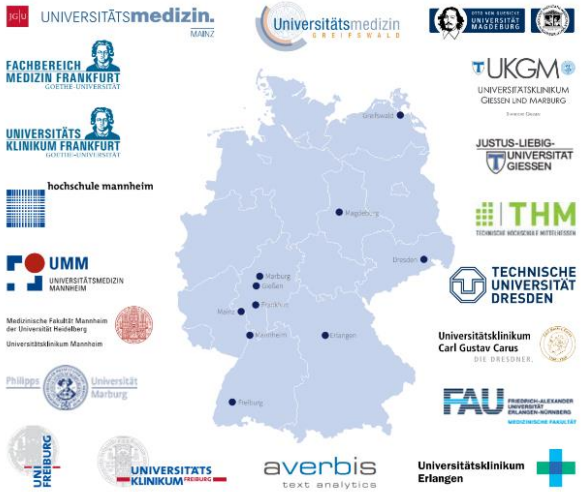
FAIR infrastructure: metadata, harmonisation, access, analysis

FAIR data

ATHLETE (€11M, 2020-2024)

develop toolbox of next-generation exposome tools

23 partners



MIRACUM (€37M, 2017-2022)

German network of University Hospitals: improve healthcare and strengthen informatics in research & education

UnCoVer (€3M, 2020-2022)

1.9M COVID-19 cases

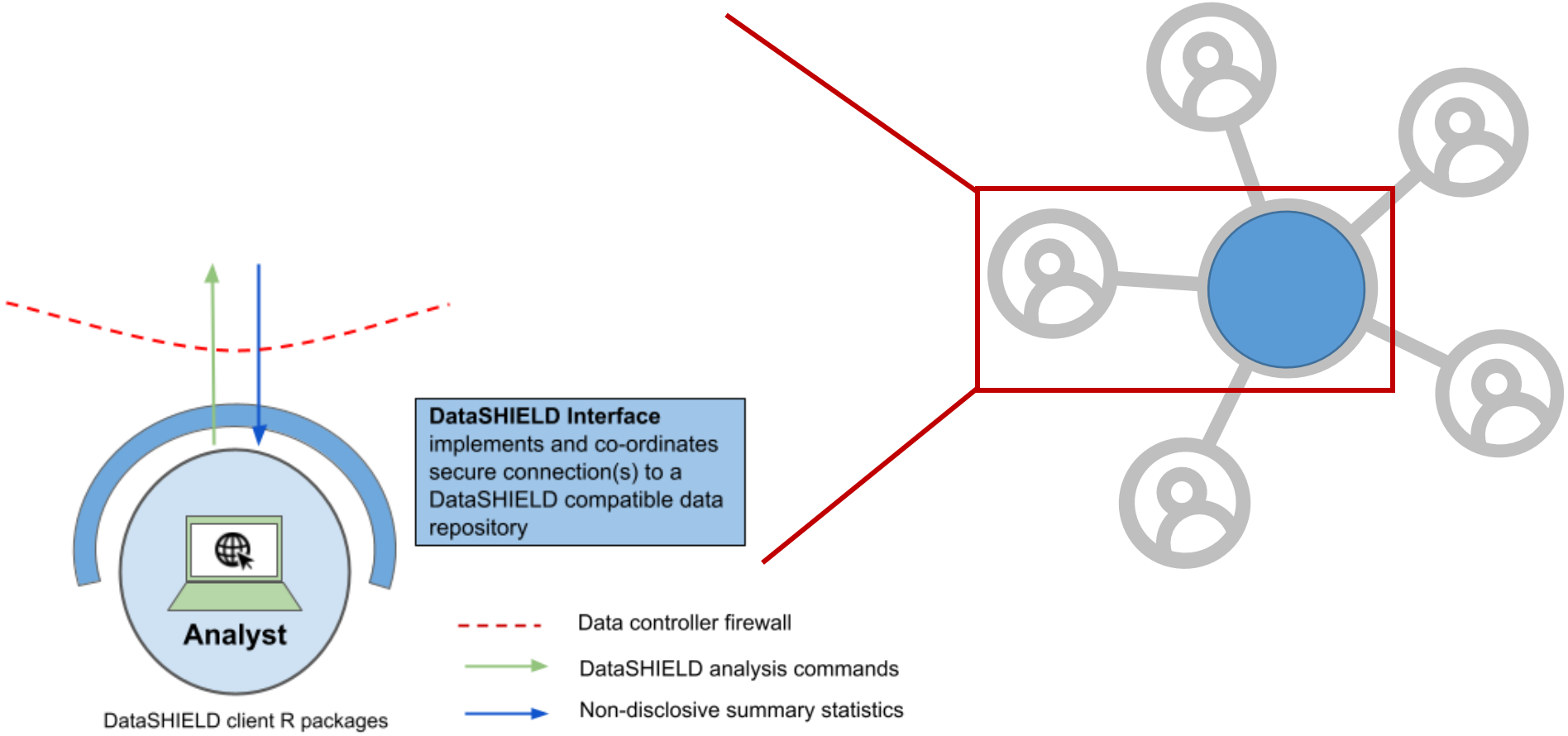
EHRs, image, omics

https://uncover-eu.net/

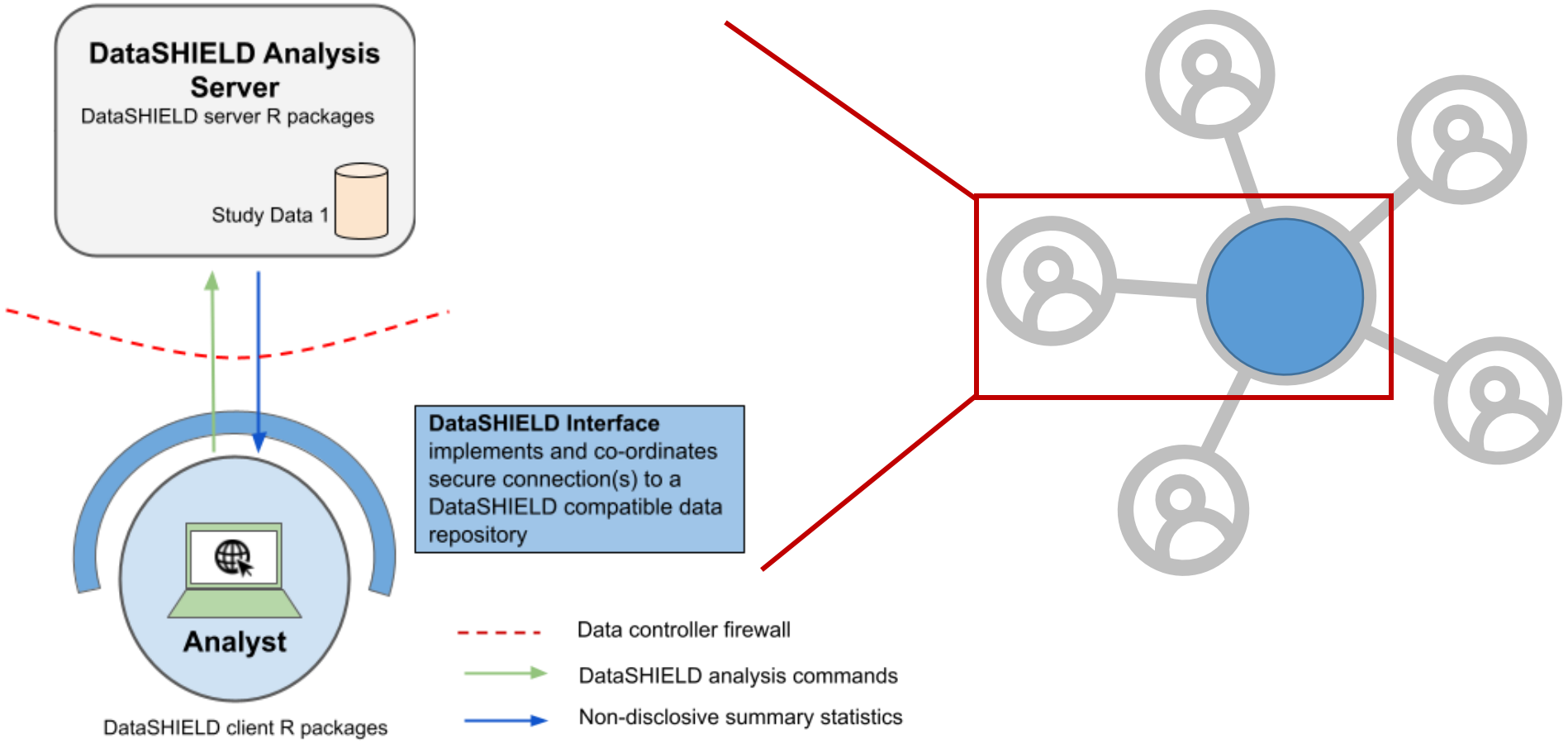
DataSHIELD Challenges

- High user demand: taking and not giving
  - Risks the future of the software
- Strategies:
  - Function and user demand – open source
  - Training needs – train the trainer https://datashield.org/forum
  - Paid support from various DS Community members
  - Paid support, paid datashield implementation/hosting
- Software sustainability
  - Formalised DS community
  - DS Community Steering Committee by end of year
  - Aim to formalise repository of community packages
  - Exploit our collaborations for funding sustainability
  - Encourage engagement in the community

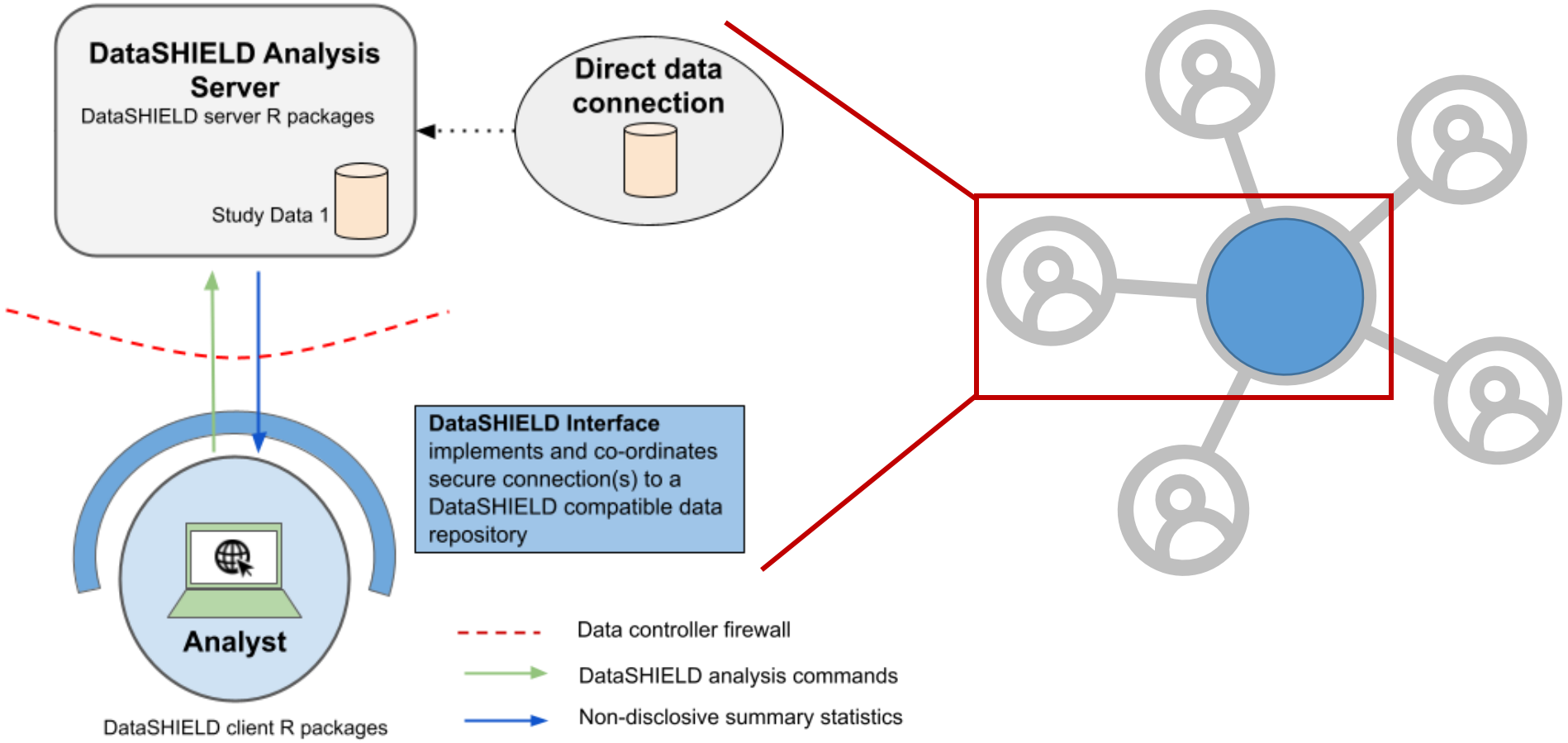
How does DataSHIELD work?



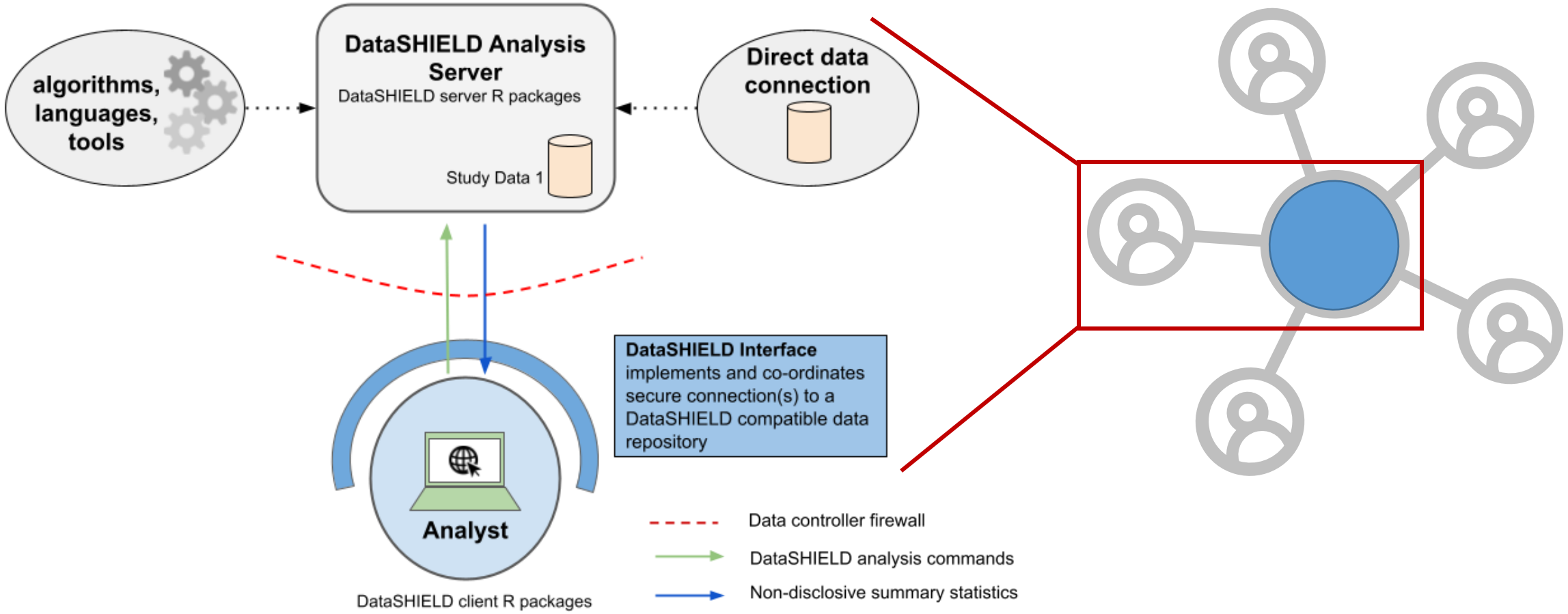
How does DataSHIELD work?



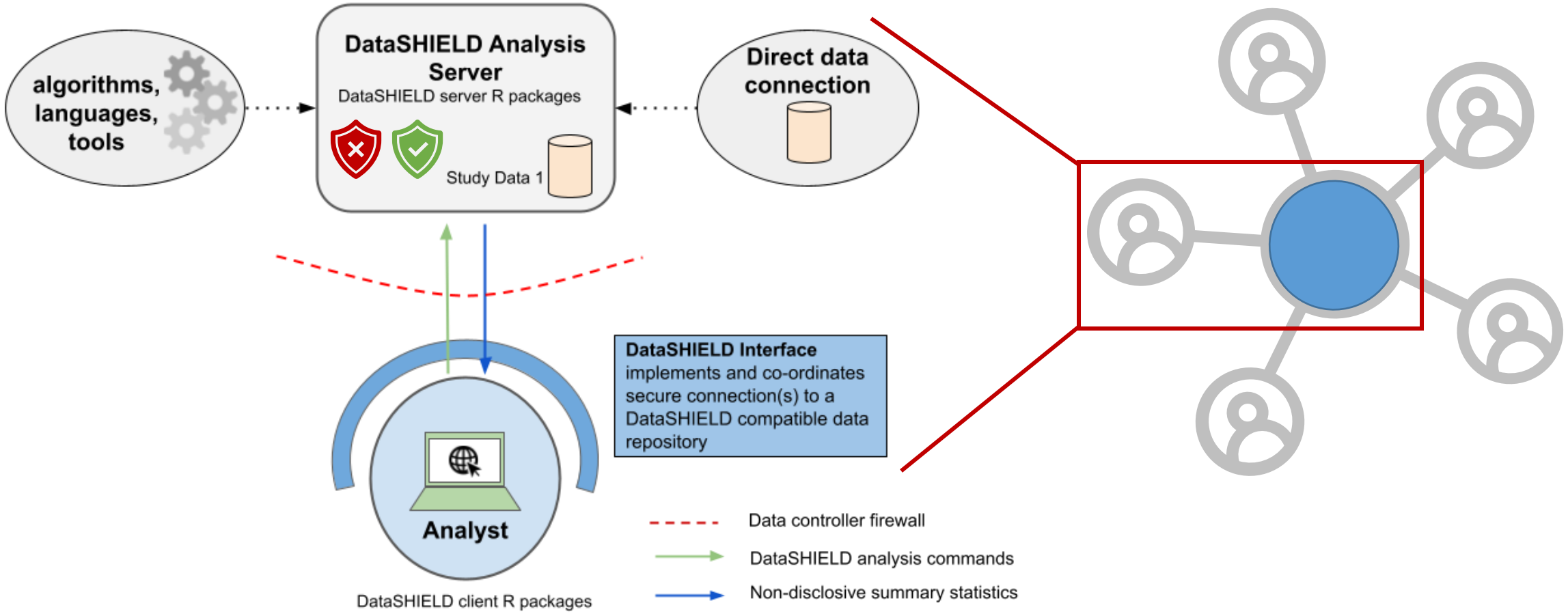
How does DataSHIELD work?



How does DataSHIELD work?

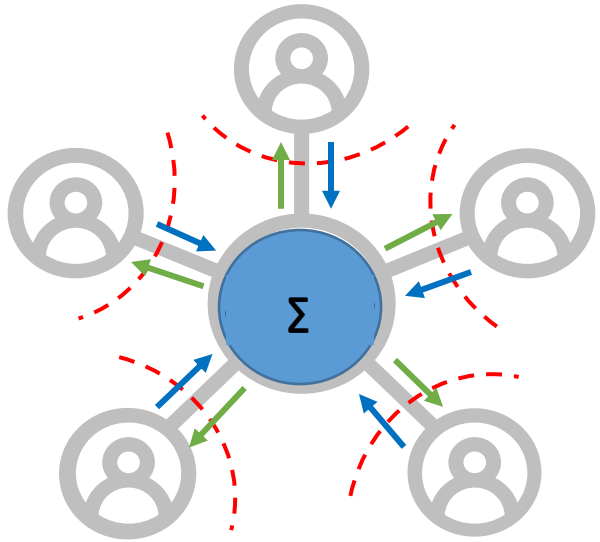


How does DataSHIELD work?



Example analysis

One step analysis

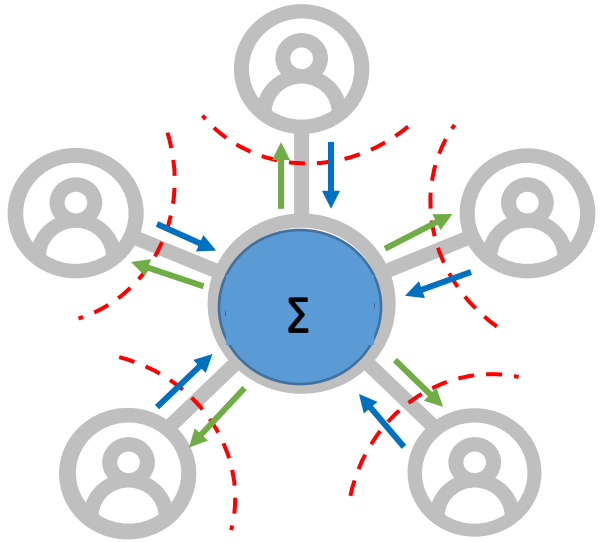


Gaye, et al (2014). *International Journal of Epidemiology*, https://doi.org/10.1093/ije/dyu188

Wilson, et al (2017). *Data Science Journal*, https://doi.org/10.5334/dsj-2017-021

Example analysis

One step analysis



Example score vector Study 1

[36, 487.2951, 487.2951, 149]

Example information matrix Study 1

500	70.56657	70.56657	297
70.56657	7646.29164	7646.29164	65.39412
70.56657	7646.29164	7646.29164	65.39412
297	65.39412	65.39412	382

Example final estimates

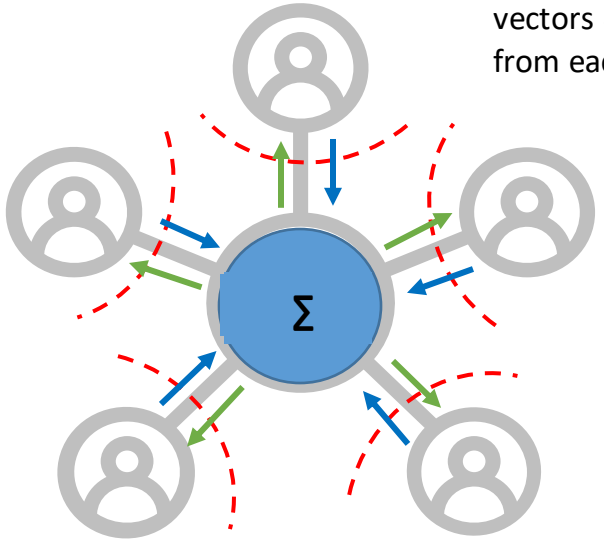
Coefficient	Estimate	Std Error
Intercept	-0.3296	0.02838
BMI	0.02300	0.00621
BMI.456	0.04126	0.01140
SNP	0.5517	0.03295

Gaye, et al (2014). *International Journal of Epidemiology*, https://doi.org/10.1093/ije/dyu188

Wilson, et al (2017). *Data Science Journal*, http://doi.org/10.5334/dsj-2017-021

Multi step analysis

Model 1 coefficients 0: score vectors & information matrices from each study returned to client



Model 2 coefficients from model 1 output: score vectors & information matrices from each study returned to client