Development of a high throughput gene, environment and epigenetics database and analysis system for international ALS research

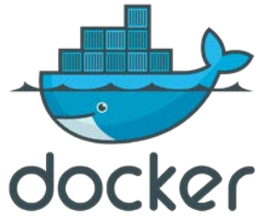
@ ENCALS 2017

Alfredo Iacoangeli
Al-Chalabi Group (Clinical Neuroscience)
Dobson Group (Health Informatics)

alfredo.iacoangeli@kcl.ac.uk

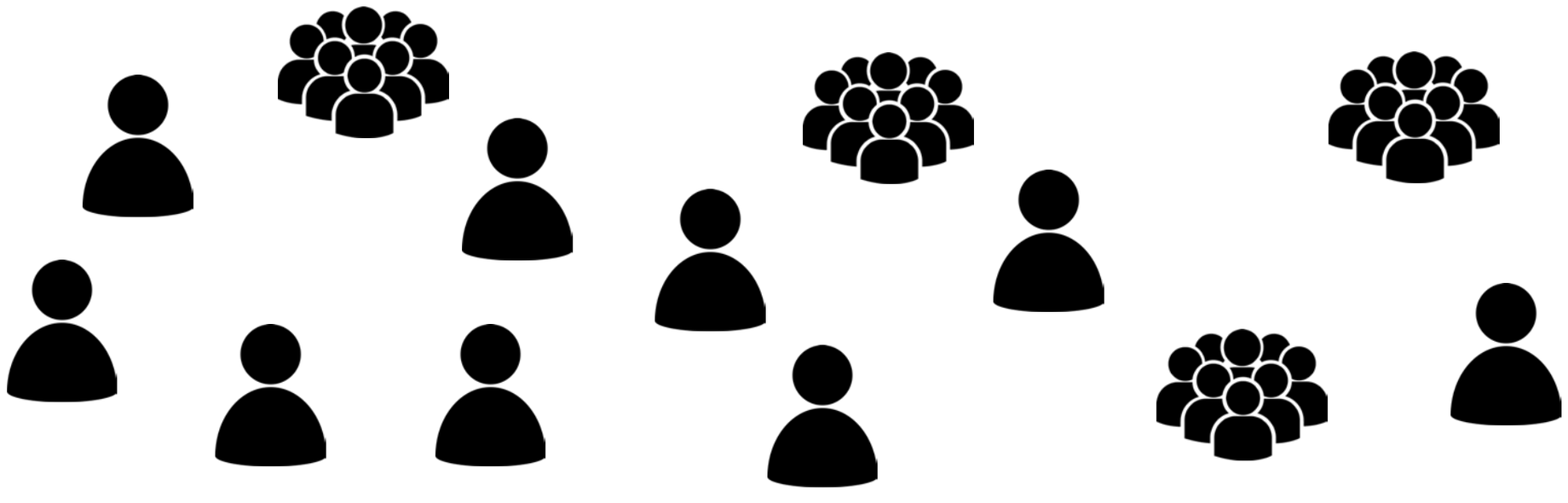
Issues

- Big Data
  - how to store it
  - how to manage it
  - how to analyse it
- Collaboration
  - ownership heterogeneity
  - data sharing
- Audience
  - accessibility
  - impact



Data Management

Data Virtualisation Layer



Data Management

?

?

project MinE Make it yours

BROAD INSTITUTE

wellcome trust sanger institute

Data Virtualisation Layer

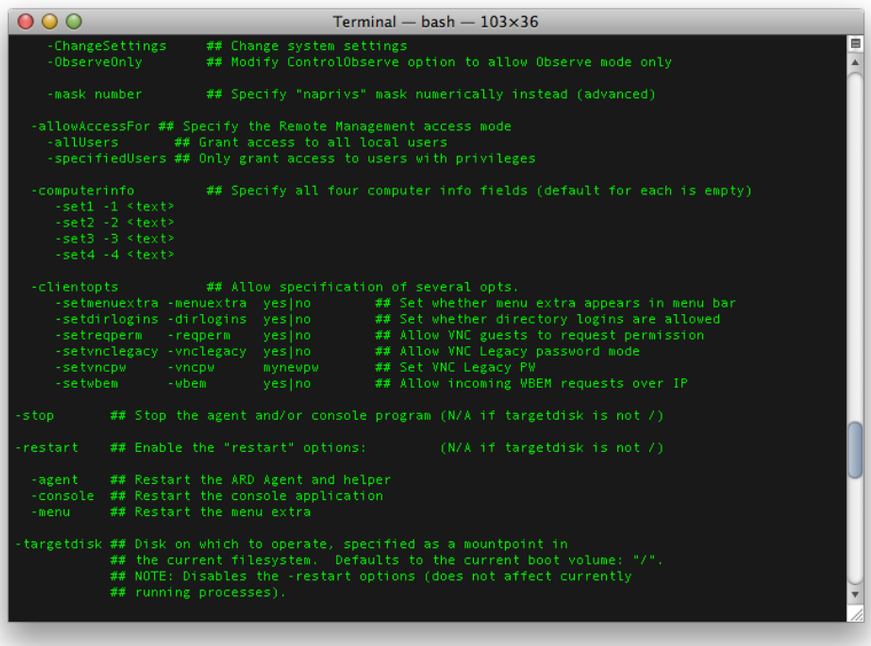
iRODS is an open source software for:

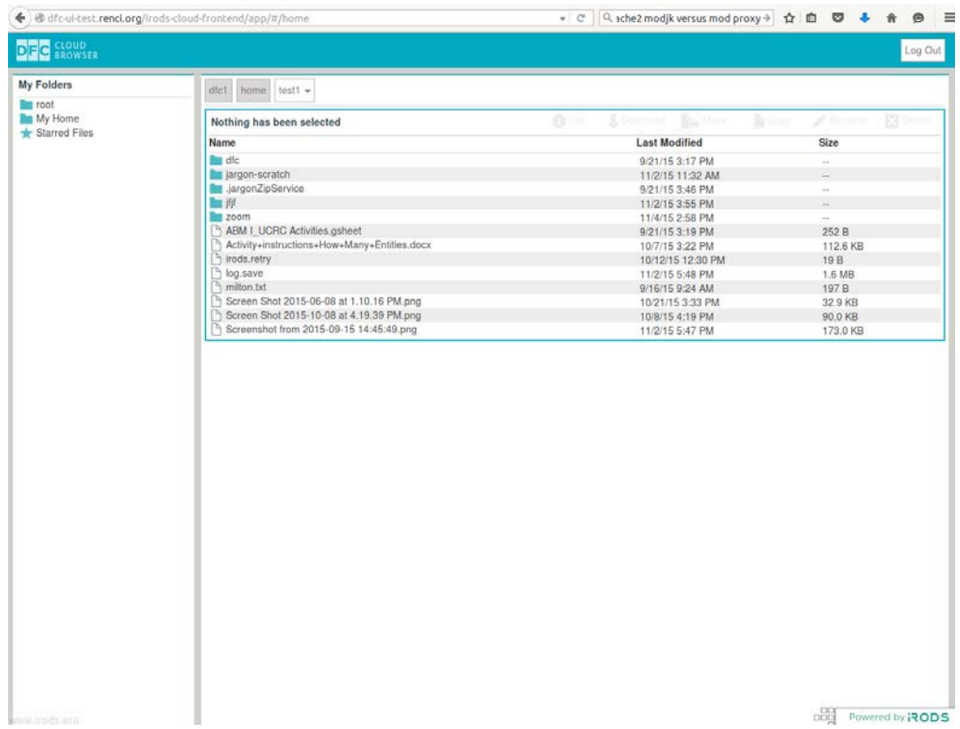
- Working with data distributed across storage technologies
- Annotating and searching data with rich metadata
- Implementing access control, auditing, preservation, organisation, and data movement policies
- Providing a single interface to share data between organisations

Universiteit Utrecht

How would we interact with all of this?

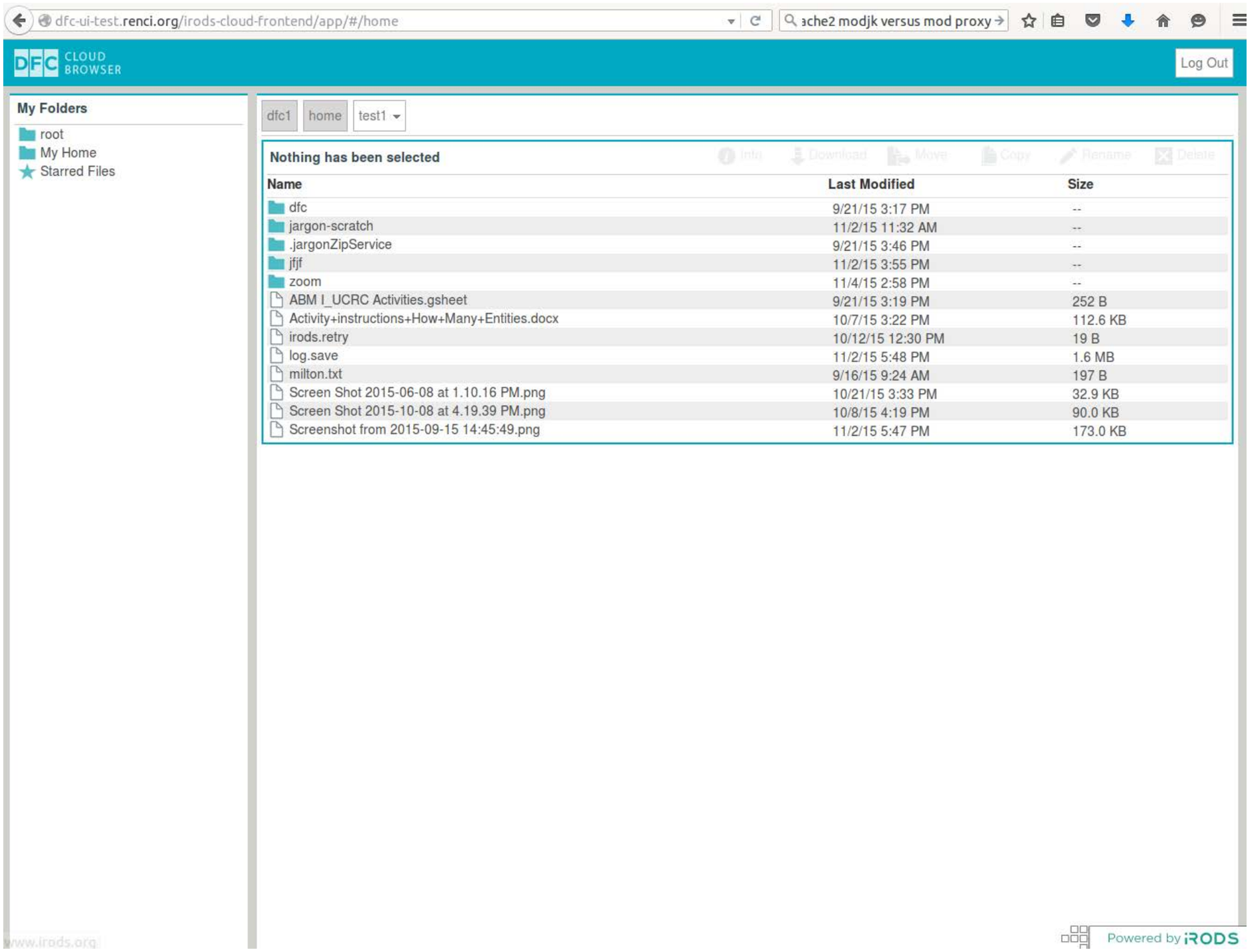
- through any iRODS zone of the network
- with the terminal command line: icommands
- web-browser: search, download/upload,write rules, add metadata, more to come…





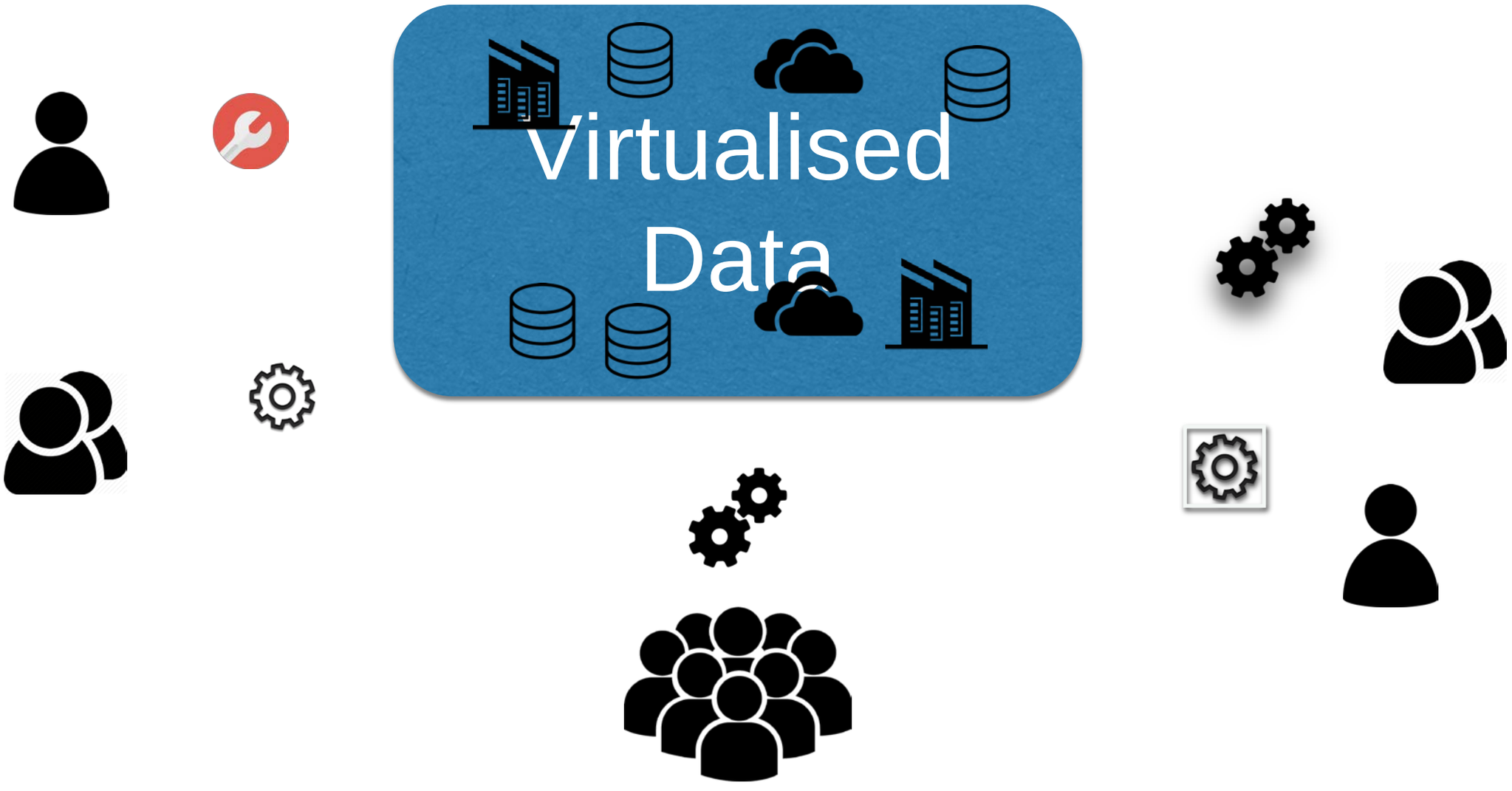
iRODS Web Browser

- Creation of new files
- File Editing
- Rule Execution
- Main Navigation Bar
- Metadata Search
- Recent Query Collections
- Docker deployment

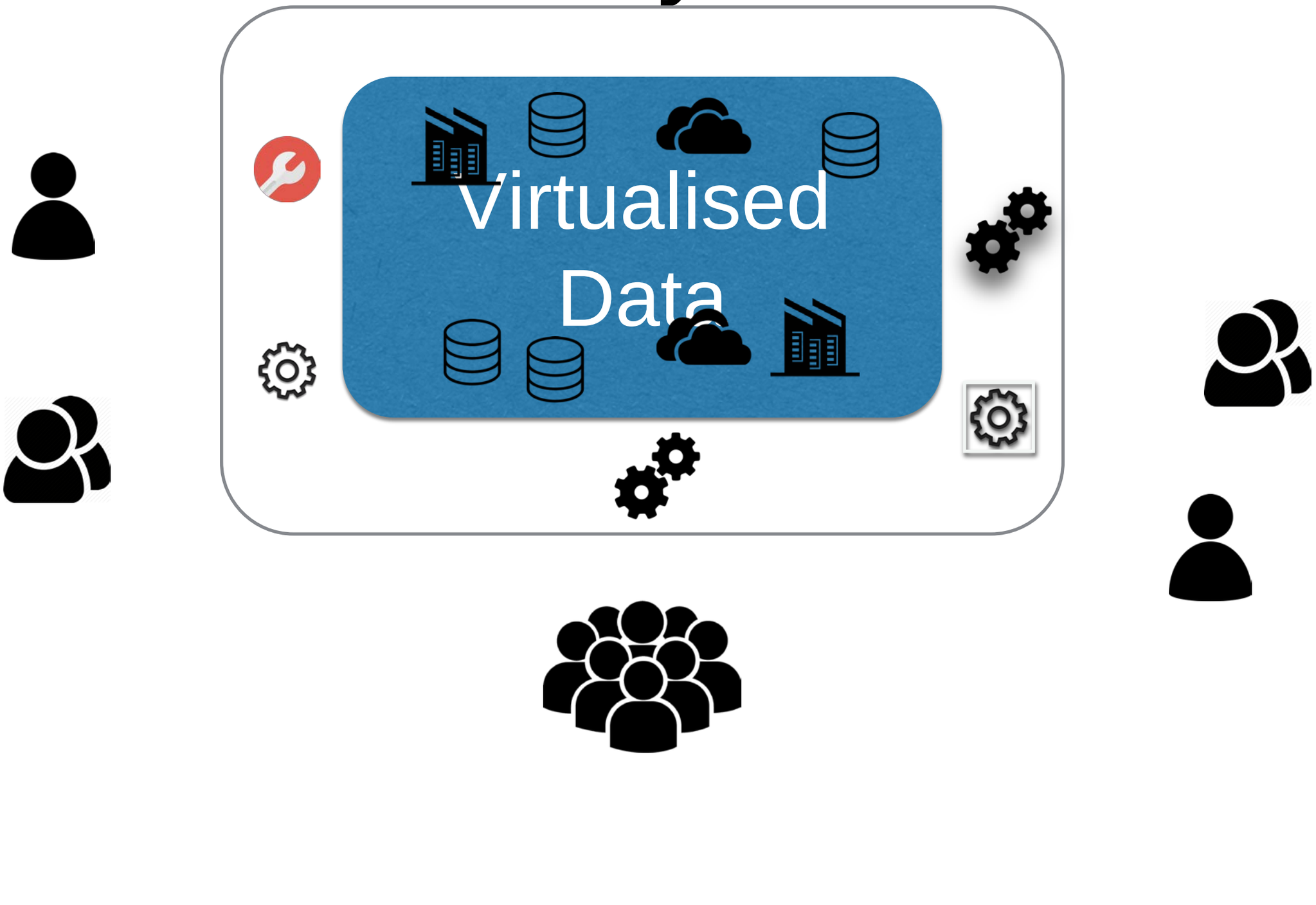


http://52.202.127.9:8552/irods-cloud-frontend/

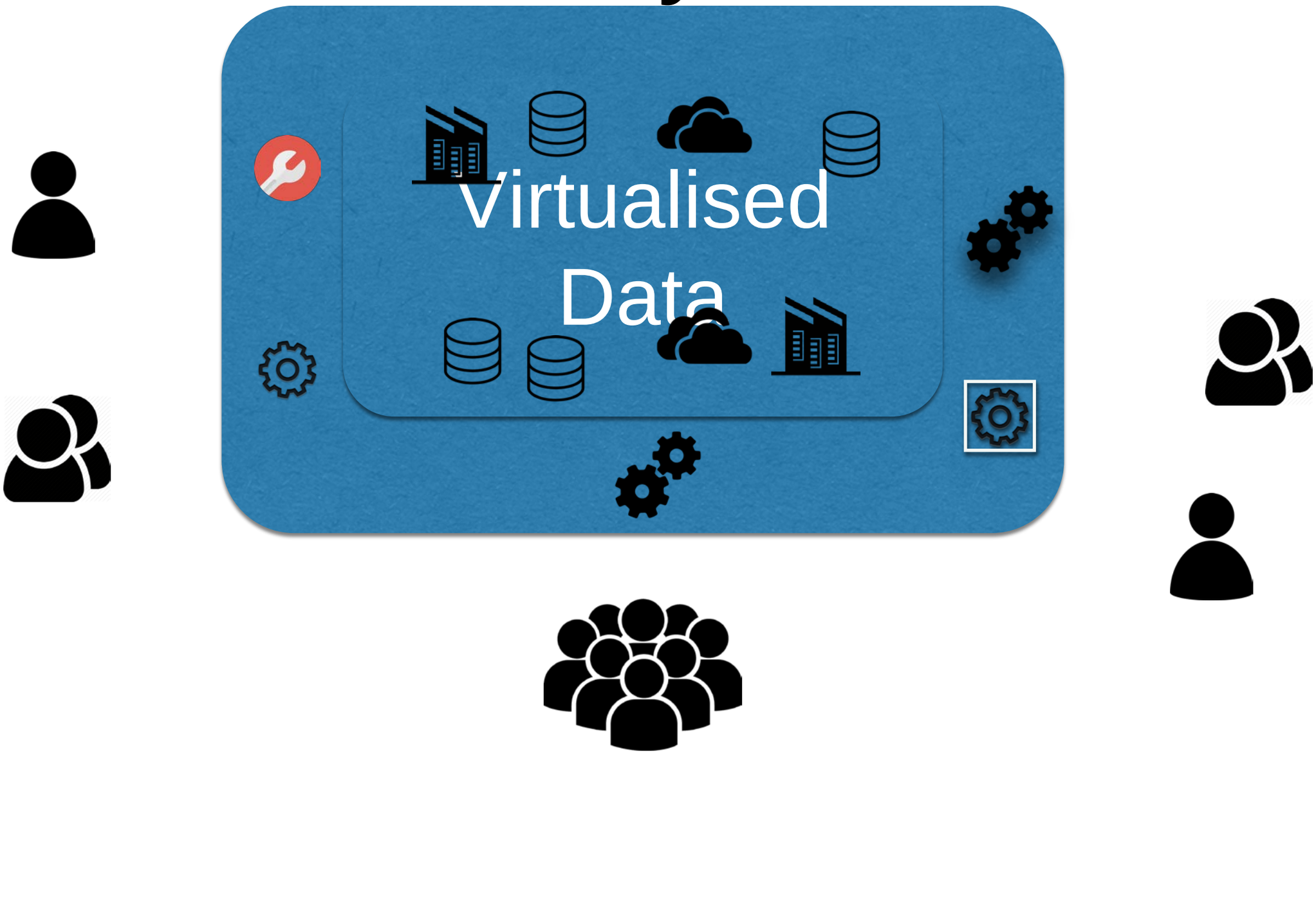
Data Accessibility and Analysis



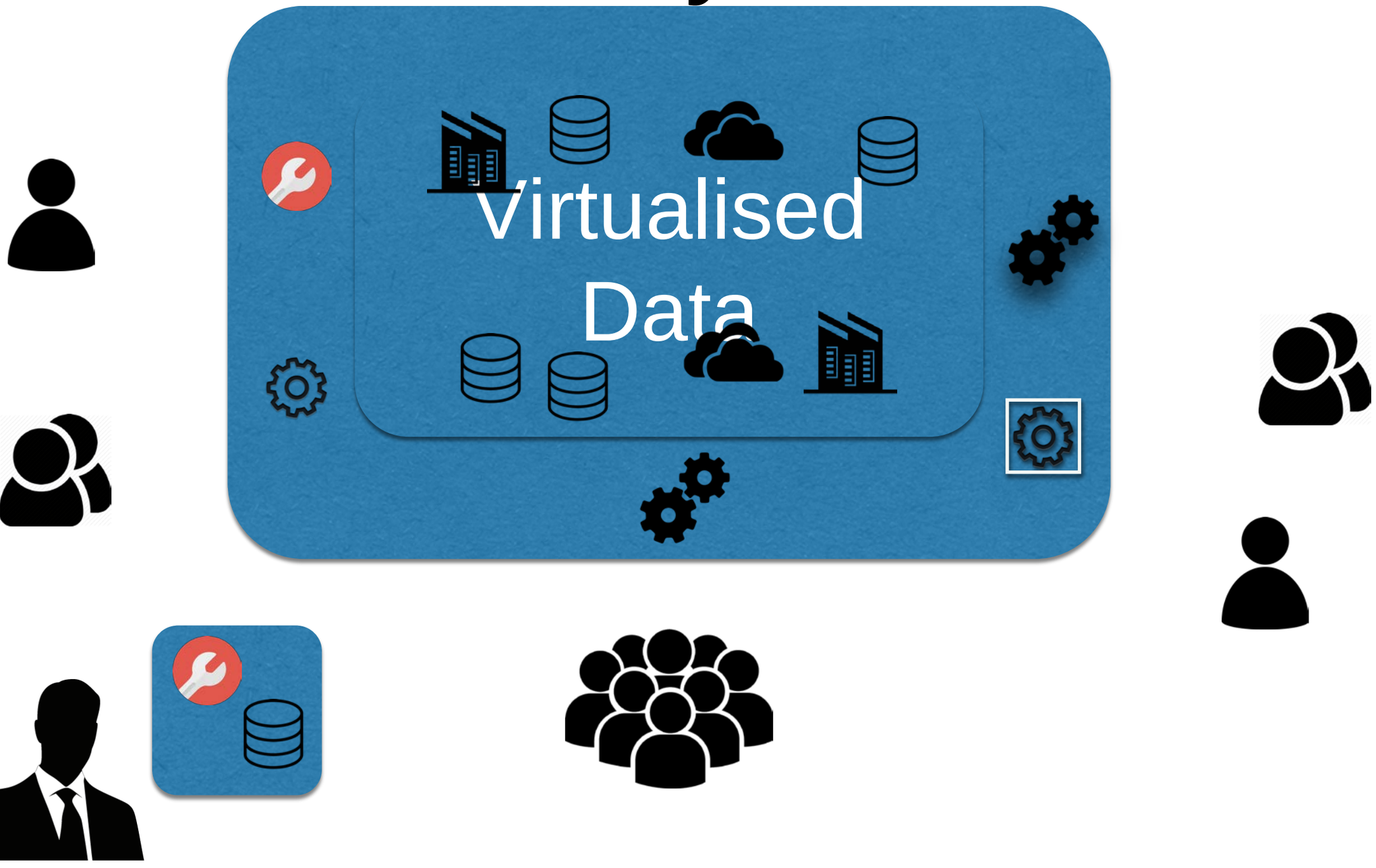
Data Accessibility and Analysis



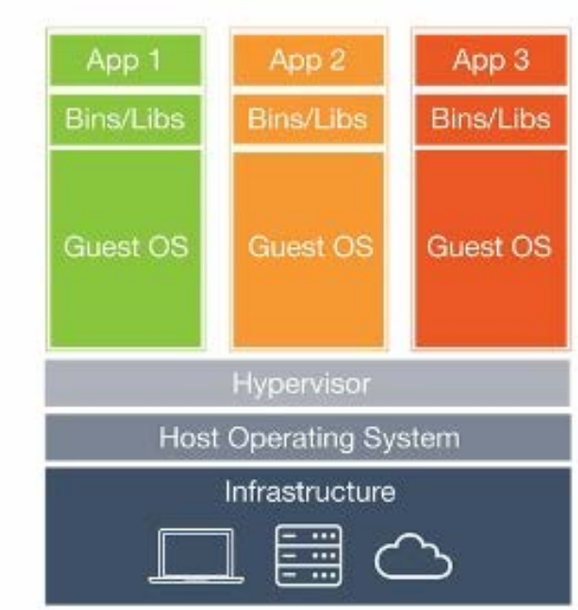
Data Accessibility and Analysis



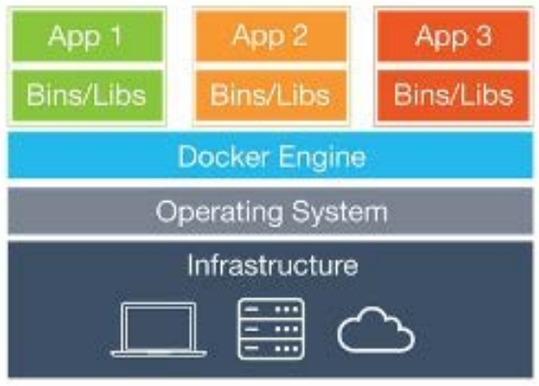
Data Accessibility and Analysis



- Docker is an open-source project that automates the deployment of Linux applications inside software containers
- Docker provides an additional layer of abstraction and automation of operating-system-level virtualisation on Linux
- Docker can be integrated into various infrastructure tools, including Amazon Web Services, Google Cloud Platform, OpenStack Nova, etc
- Containers running on a single machine share the same operating system kernel; they start instantly and use less RAM. Images are constructed from layered filesystems and share common files, making disk usage and image downloads much more efficient.



Virtual Machines



Containers

NGSeasy

Install it:

```
:~$ git clone https://github.com/KHP-Informatics/ngseasy.git
:~$ cd ngseasy
:~$ make INSTALLDIR=”/media/Data” all
:~$ sudo make install
```

Run it:

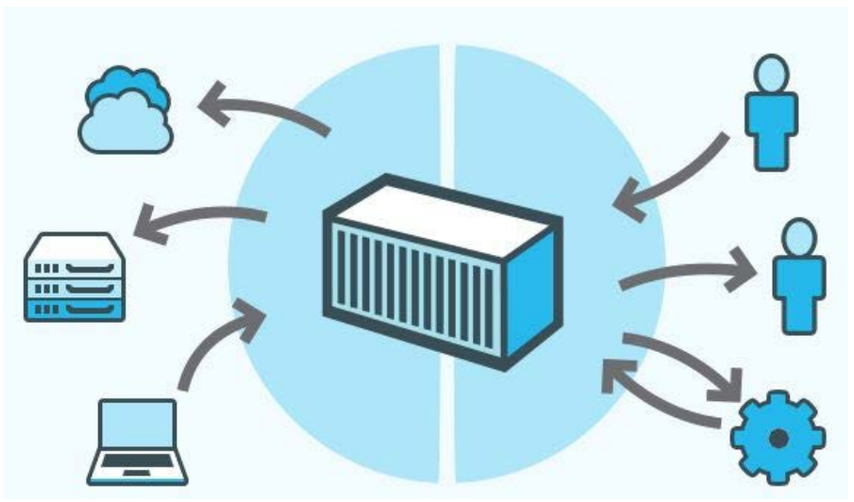
```
:~$ cd /media/Data/ngs_projects/congif_files
:~$ ngseasy -c my_config.tsv -d /media/Data/ngs_projects
```

https://github.com/KHP-Informatics/ngseasy

What is on the plate

- An iRODS system tailored for ALS research needs

- Dockerized analysis pipelines



https://github.com/KHP-Informatics/MND-DataManagementAnalysis-System

tranSMART v1.2

- **open-source, community-driven knowledge management platform for translational medicine**

- **organises clinical and research data on per patient base allowing:**

  - **Compare data from proteomics, metabolomics and other “omics” studies**

  - **Contrast patterns of gene expression in healthy and diseased individuals and in human tissue samples**

  - **Investigate correlations between genotype and phenotype in clinical trial data**

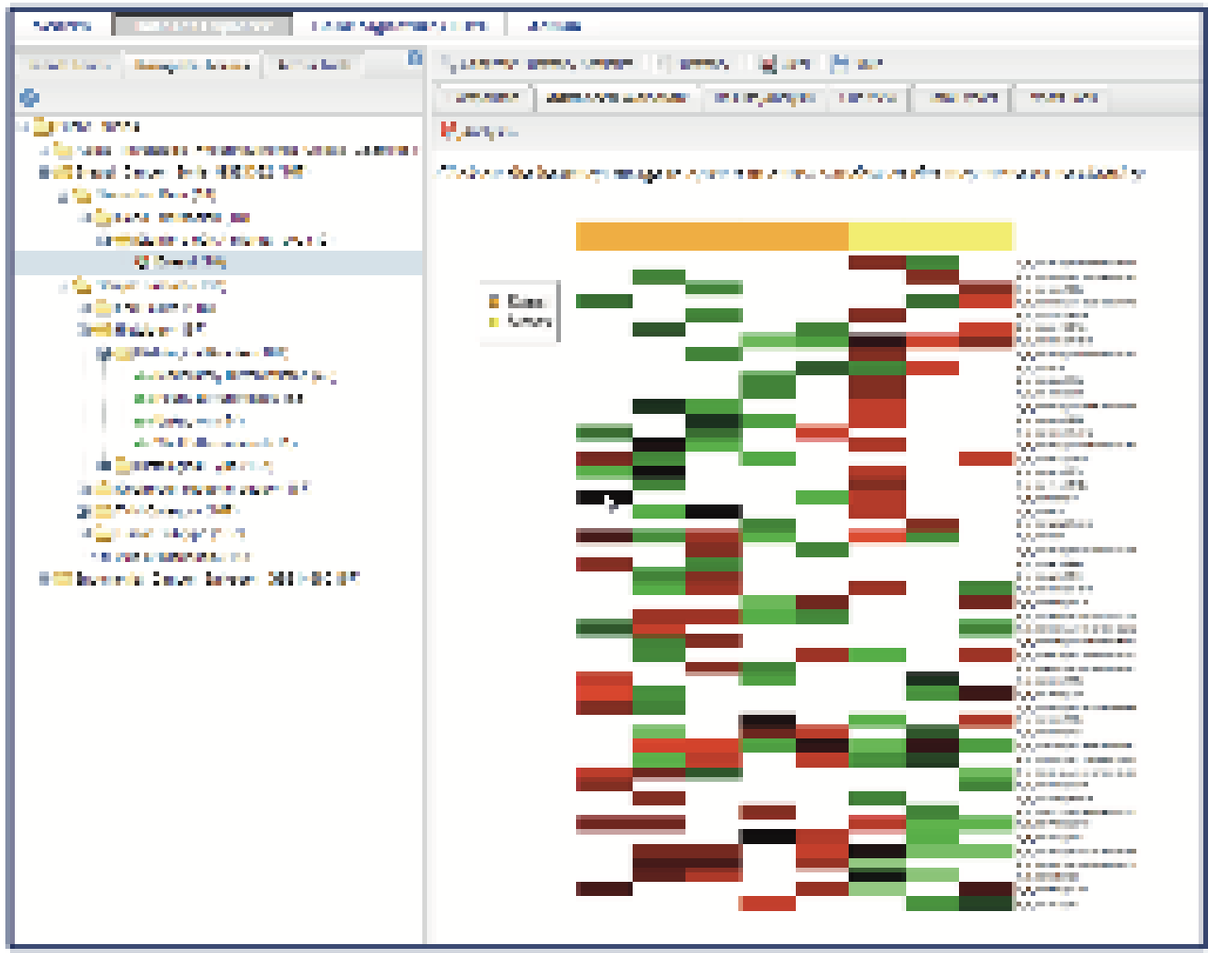
  - **Mine pre-clinical data for insights into the biology of human disease**

  - **Study genetic and environmental factors involved in human disease**

  - **Display data visually using a graphical interface**

  - **Stratify clinical data into molecular subtypes of a specific disease**

  - **Collaborate across academic, government and corporate research sectors**



Ok! All good but who is going to make this happen? and how?

SERVICE	Deployment time	Needed experience
iRODS local	hours	basic
Docker iRODS	minutes	basic
Docker	minutes	basic
Docker container (r)	minutes	basic
Docker image (r)	minutes	basic
iRODS rules (r)	hours	intermediate
iRODS rules (w)	hours	needs training
iRODS administration	minutes	needs training
Docker container (w)	hours	needs training
Docker image (w)	hours	needs training

Summary

A. Data management iRODS

- data sharing
- data accessibility
- data curation
- automatise workflow
- exploit metadata potentialities

B. docker, standardised pipelines, etc

- docker files
- docker images
- shared scrips
- Github

C. Community driven project

- tailor the system according to our needs

Future Plans

- TranSMART platform interactive analysis
- Develop portable analysis pipelines
- iRODS management system refinement
- User friendly interface

MRC Medical Research Council
Leading science for better health

mnda
motor neurone disease association

ALS ASSOCIATION

STRENGTH

MND REGISTER

project MinE
Make it yours

JPND research

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

50 E·S·R·C

Economic and Social Research Council
Shaping Society

ALS THERAPY ALLIANCE

NHS
National Institute for Health Research

Clinical Neuroscience

- Ammar Al-Chalabi
- Ahmad Al Khleifat
- Aleksey Shatunov
- Anna Kulka
- Anand Pandit
- Ashley Jones
- Sarah Martin
- William Sproviero

Health Informatics

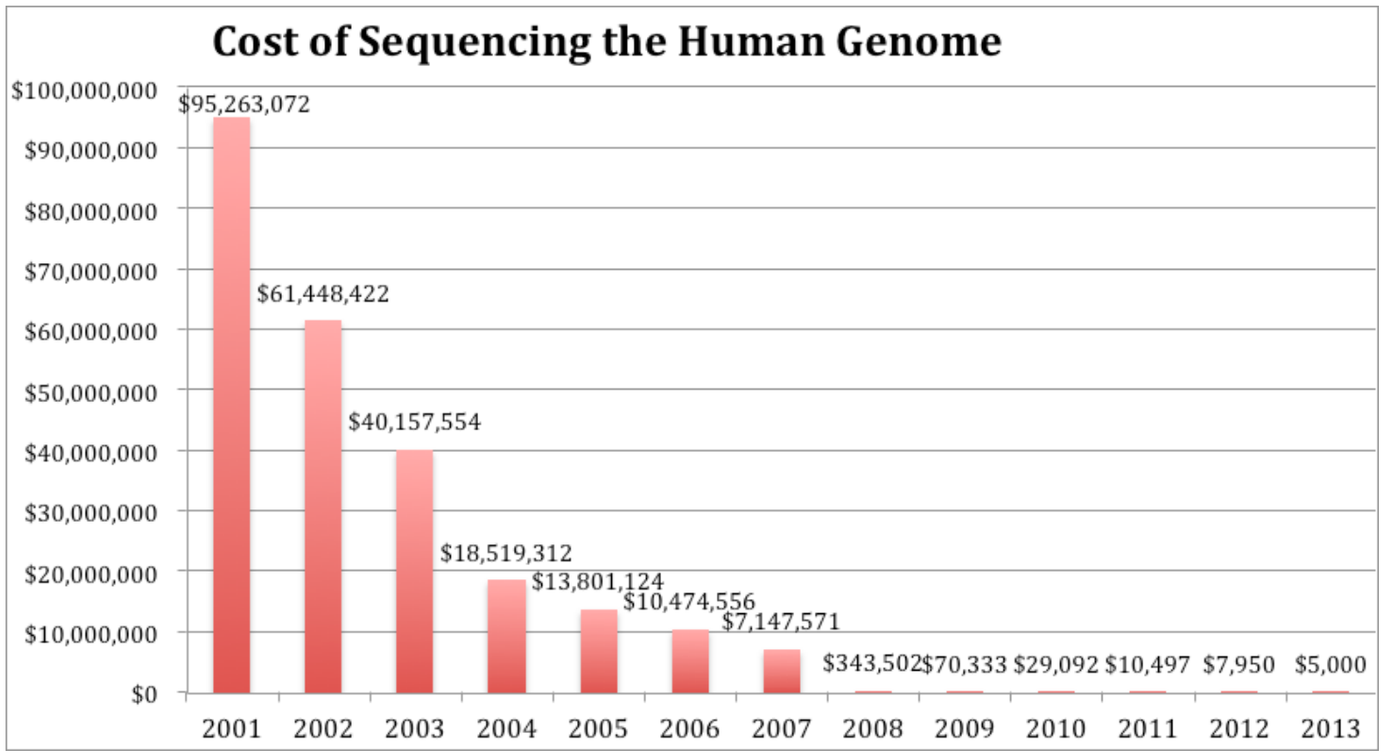
- Stephen J Newhouse
- Richard Dobson

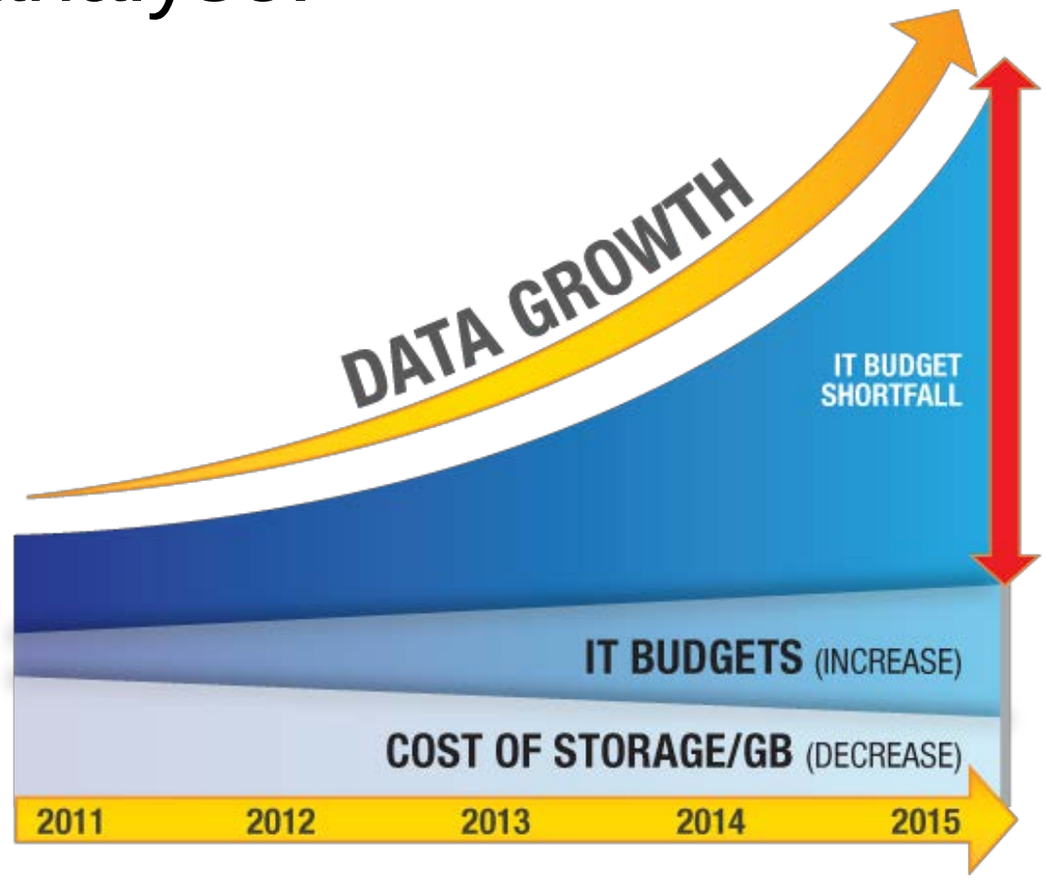
All our collaborators!!!!!!!!

Attribute	sample bam	sample vcf	sample XYZ
file type	bam	vcf	XYZ
lane	2	2	—
sample ID	LP00192	LP00192	LP00192
reference	hg19	hg19	—
study	MyFavStudy	YourFavStudy	TheirFavStudy
aligned	1	—	—
alignment	MyFavPipeline	MyFavPipeline	—
Genotyping	—	MyFavPipeline	—
user defined attribute 1	—	—	yellow
user defined attribute 2	—	—	bitter

- **Background**:

We are now collecting huge amounts of multilayered genetic, epigenetic, environmental and clinical data, much of which is difficult to share and therefore to analyse.

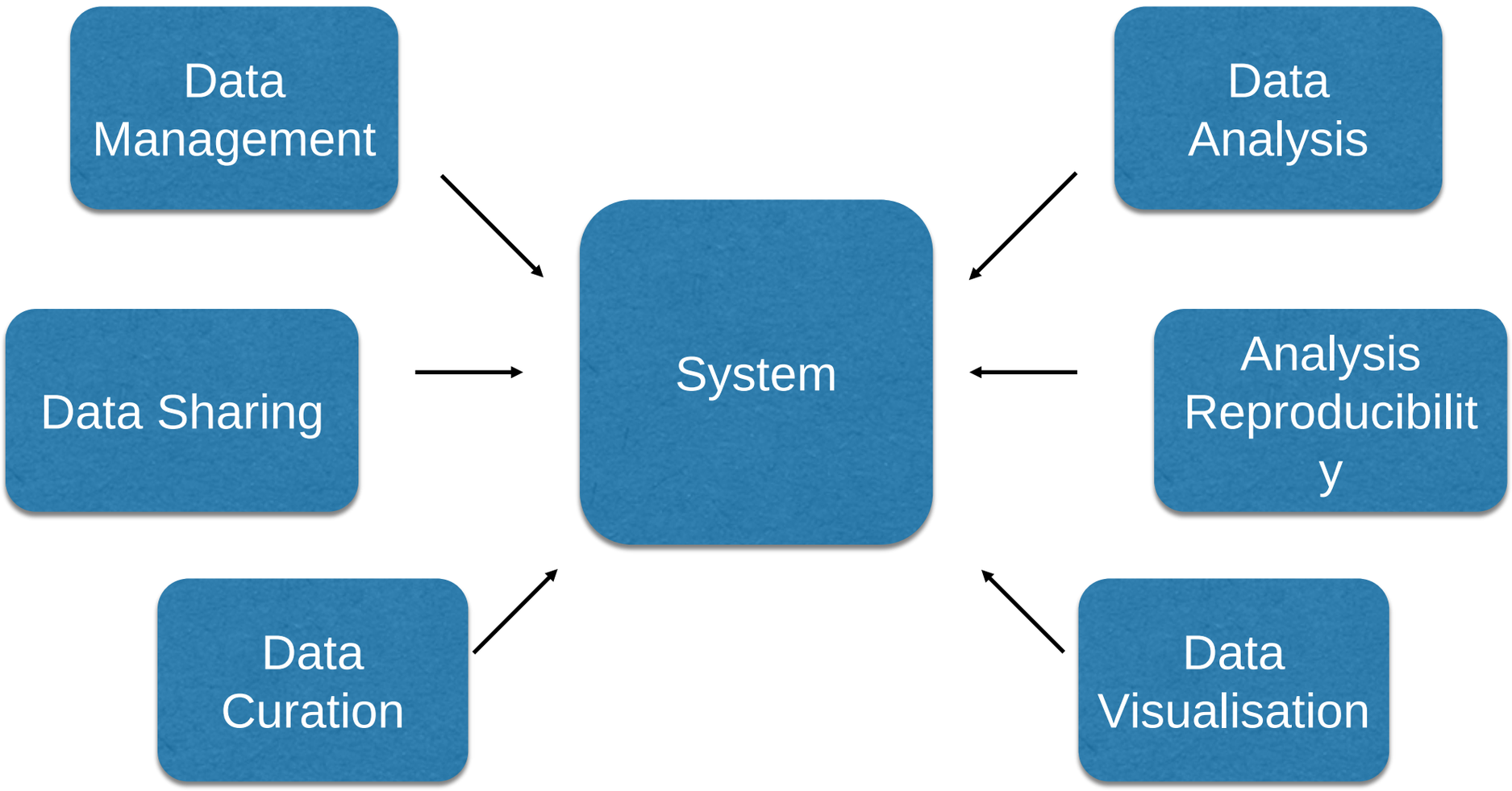




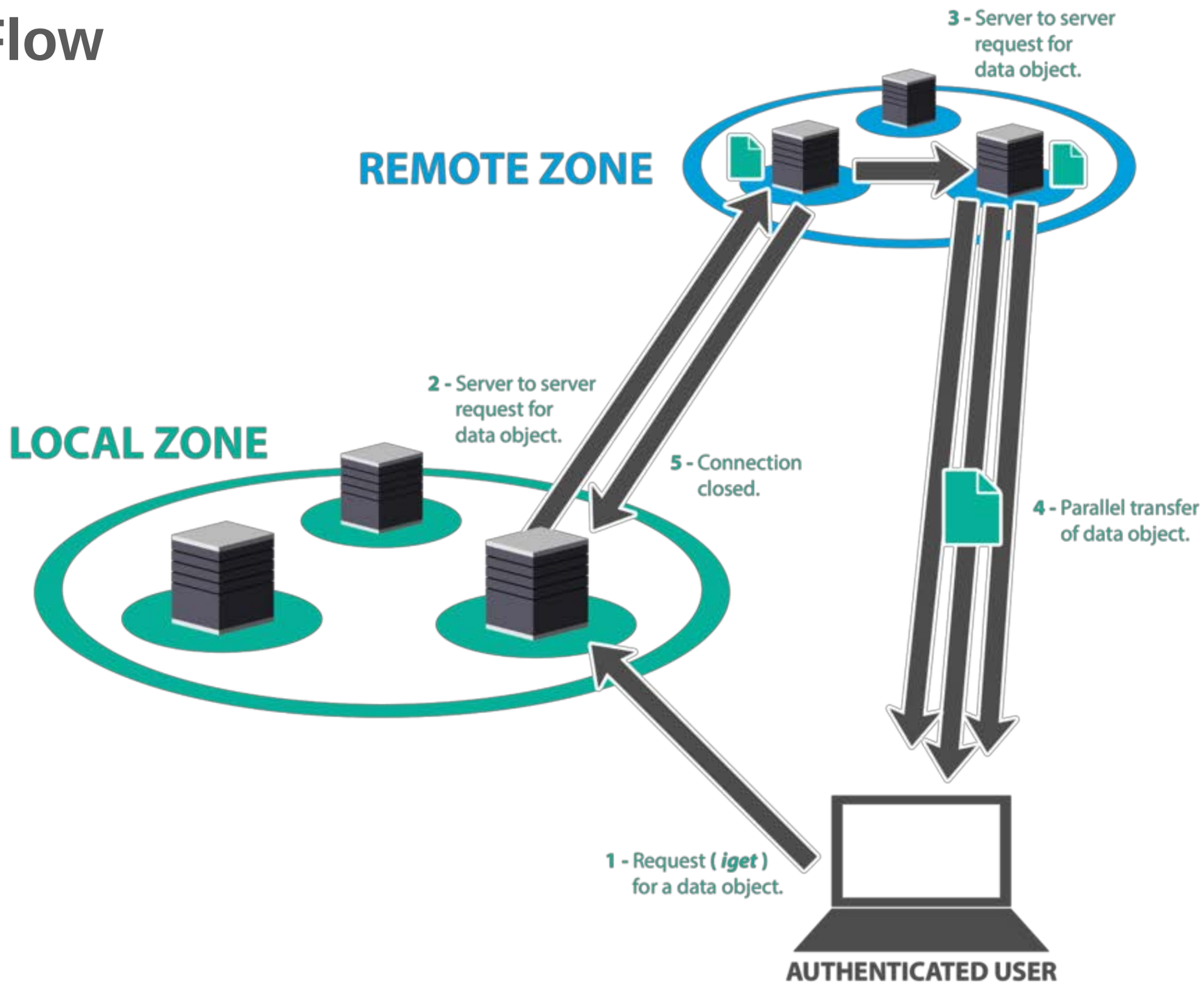
- **Project aim:**

This project is a collaboration between ALS researchers and bi

Data Management and Analysis System



Data Flow


3 - Server to server request for data object.
REMOTE ZONE
2 - Server to server request for data object.
LOCAL ZONE
5 - Connection closed.
4 - Parallel transfer of data object.
1 - Request (iget) for a data object.
AUTHENTICATED USER

Ok! All good but who is going to make this happen? and how?

who?

- on a central system -> King's IT + the community
- on your institution facilities -> ask your admins
- on local servers -> we can make it together

how?

- manual deployment -> www.iRODS.com
- Docker/guided deployment -> our github SOCIAL CODING

https://github.com/KHP-Informatics/MND-DataManagementAnalysis-System