



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

iRODS

- **Collaboration**

- ownership heterogeneity
- data sharing



Global Alliance
for Genomics & Health



docker



Bitbucket

- **Audience**

- accessibility
- impact



github
SOCIAL CODING

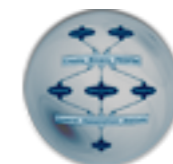


transSMART
v1.2

Galaxy
PROJECT



Pachyderm

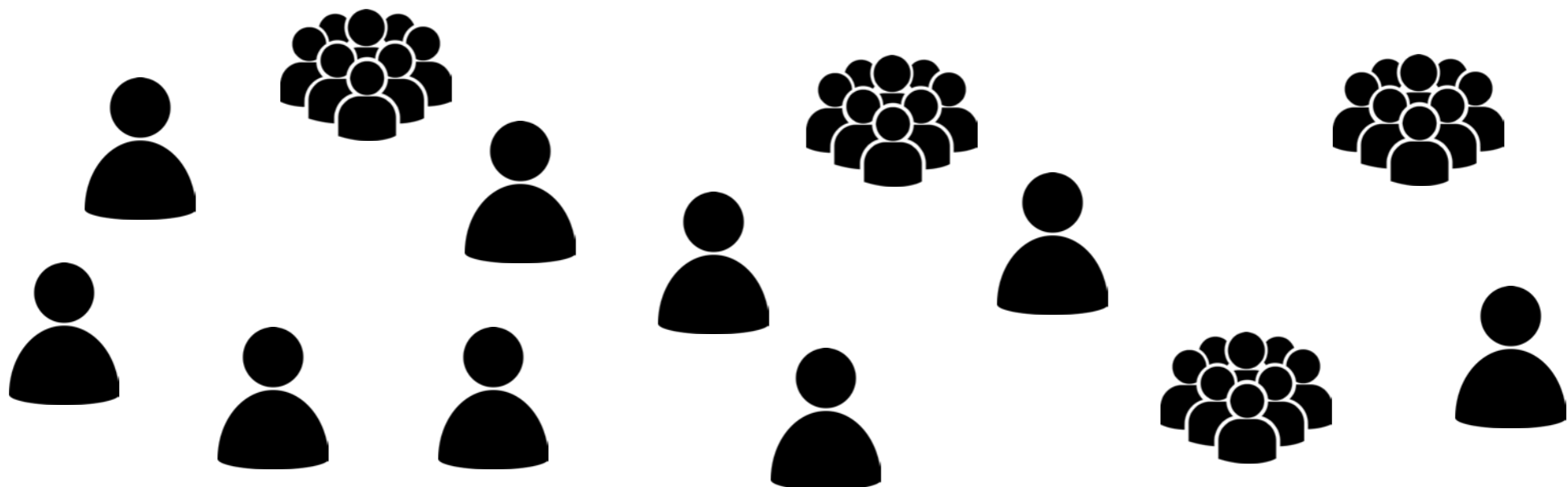


WINGS

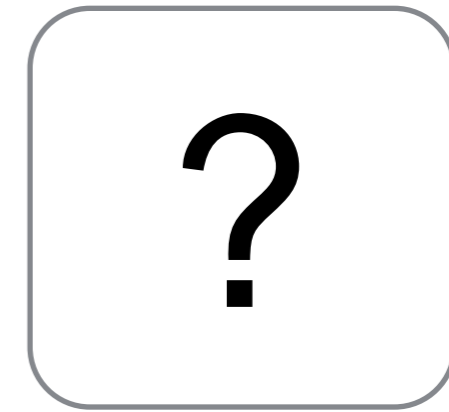
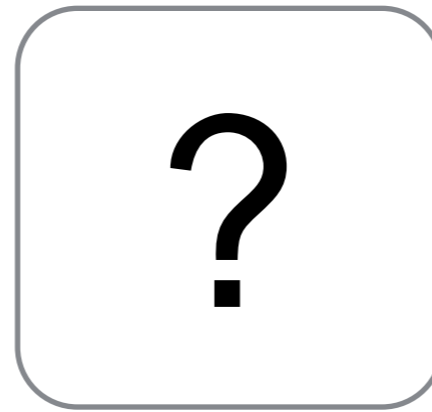
Data Management



Data Virtualisation Layer



Data Management



Data Virtualisation Layer



iRODS

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



renci



EMC²

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...

```
Terminal — bash — 103x36
-ChangeSettings ## Change system settings
-ObserveOnly ## Modify ControlObserve option to allow Observe mode only

-mask number ## Specify "naprivs" mask numerically instead (advanced)

-allowAccessFor ## Specify the Remote Management access mode
-allUsers ## Grant access to all local users
-specifiedUsers ## Only grant access to users with privileges

-computerinfo ## Specify all four computer info fields (default for each is empty)
-set1 -1 <text>
-set2 -2 <text>
-set3 -3 <text>
-set4 -4 <text>

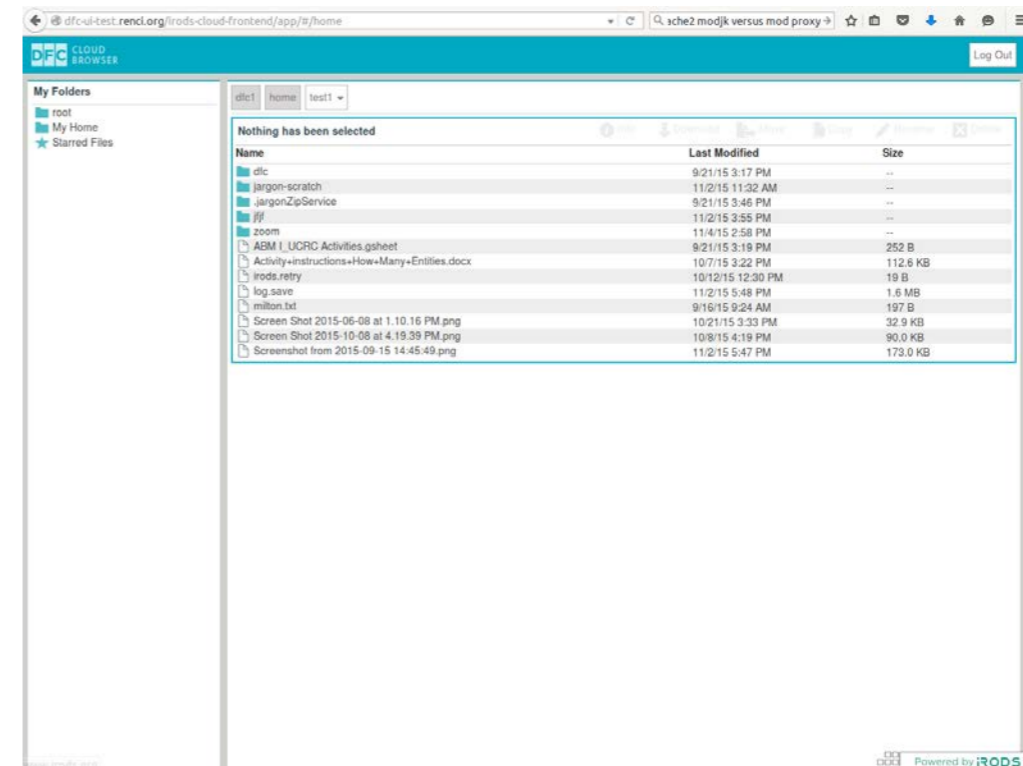
-clientopts ## Allow specification of several opts.
-setmenuextra -menuextra yes|no ## Set whether menu extra appears in menu bar
-setdirlogins -dirlogins yes|no ## Set whether directory logins are allowed
-setreqperm -reqperm yes|no ## Allow VNC guests to request permission
-setvnclegacy -vnclegacy yes|no ## Allow VNC Legacy password mode
-setvncpw -vncpw mynewpw ## Set VNC Legacy PW
-setwbem -wbem yes|no ## Allow incoming WBEM requests over IP

-stop ## Stop the agent and/or console program (N/A if targetdisk is not /)

-restart ## Enable the "restart" options: (N/A if targetdisk is not /)

-agent ## Restart the ARD Agent and helper
-console ## Restart the console application
-menu ## Restart the menu extra

-targetdisk ## Disk on which to operate, specified as a mountpoint in
## the current filesystem. Defaults to the current boot volume: "/".
## NOTE: Disables the -restart options (does not affect currently
## running processes).
```



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

The screenshot shows the iRODS Cloud Browser interface. The browser address bar displays `dfc-ui-test.renci.org/irods-cloud-frontend/app/#/home`. The interface includes a navigation bar with the DFC logo and a 'Log Out' button. On the left, a 'My Folders' sidebar lists 'root', 'My Home', and 'Starred Files'. The main content area shows a breadcrumb trail 'dfc1 > home > test1' and a table of files. The table has three columns: 'Name', 'Last Modified', and 'Size'. The files listed include folders like 'dfc', 'jargon-scratch', and 'zoom', and files such as 'ABM I_UCRC Activities.gsheat', 'Activity+instructions+How+Many+Entities.docx', and several screenshots.

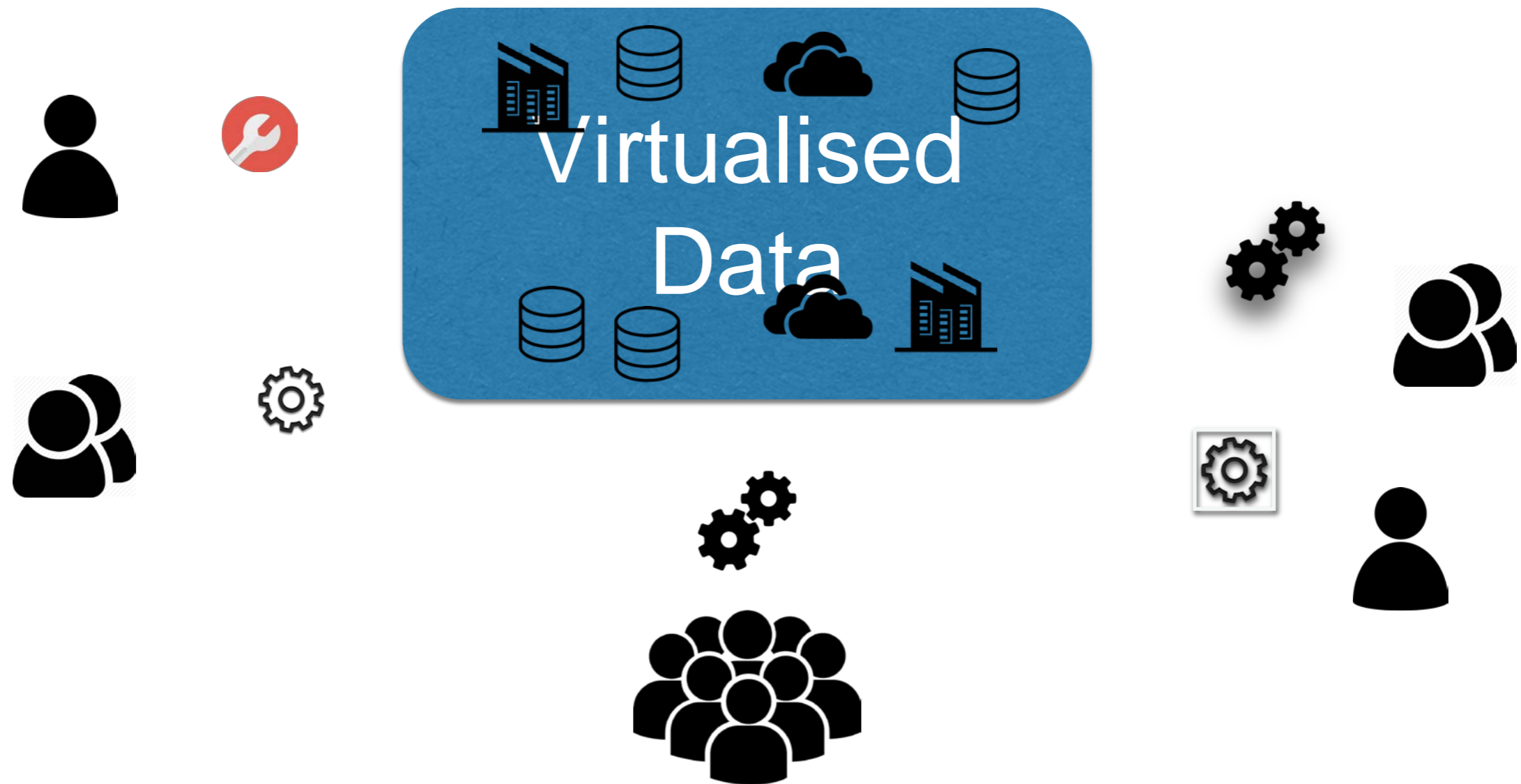
Name	Last Modified	Size
dfc	9/21/15 3:17 PM	--
jargon-scratch	11/2/15 11:32 AM	--
.jargonZipService	9/21/15 3:46 PM	--
ijjf	11/2/15 3:55 PM	--
zoom	11/4/15 2:58 PM	--
ABM I_UCRC Activities.gsheat	9/21/15 3:19 PM	252 B
Activity+instructions+How+Many+Entities.docx	10/7/15 3:22 PM	112.6 KB
irods.retry	10/12/15 12:30 PM	19 B
log.save	11/2/15 5:48 PM	1.6 MB
milton.txt	9/16/15 9:24 AM	197 B
Screen Shot 2015-06-08 at 1.10.16 PM.png	10/21/15 3:33 PM	32.9 KB
Screen Shot 2015-10-08 at 4.19.39 PM.png	10/8/15 4:19 PM	90.0 KB
Screenshot from 2015-09-15 14:45:49.png	11/2/15 5:47 PM	173.0 KB

www.irods.org

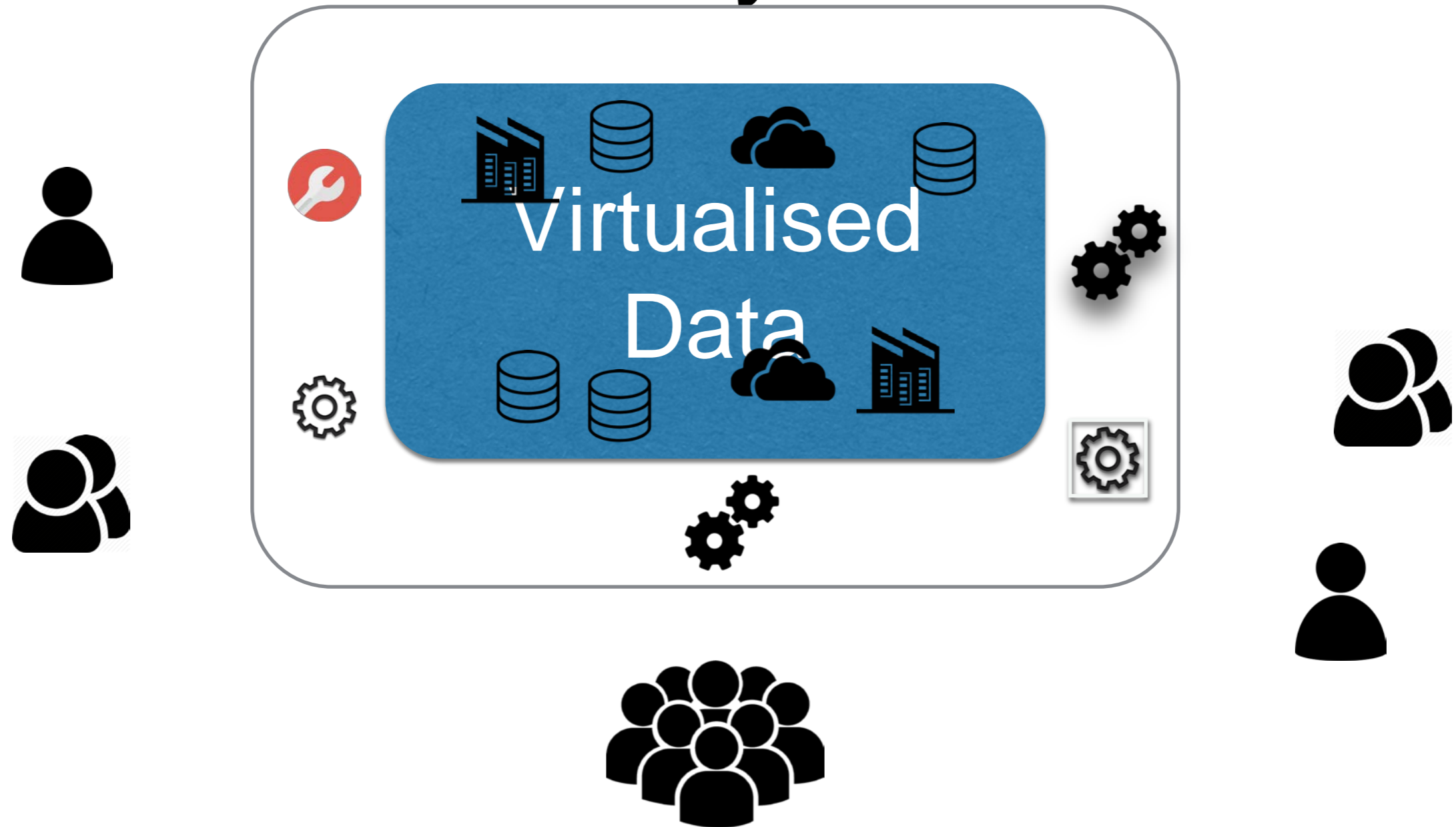
Powered by iRODS

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

Data Accessibility and Analysis



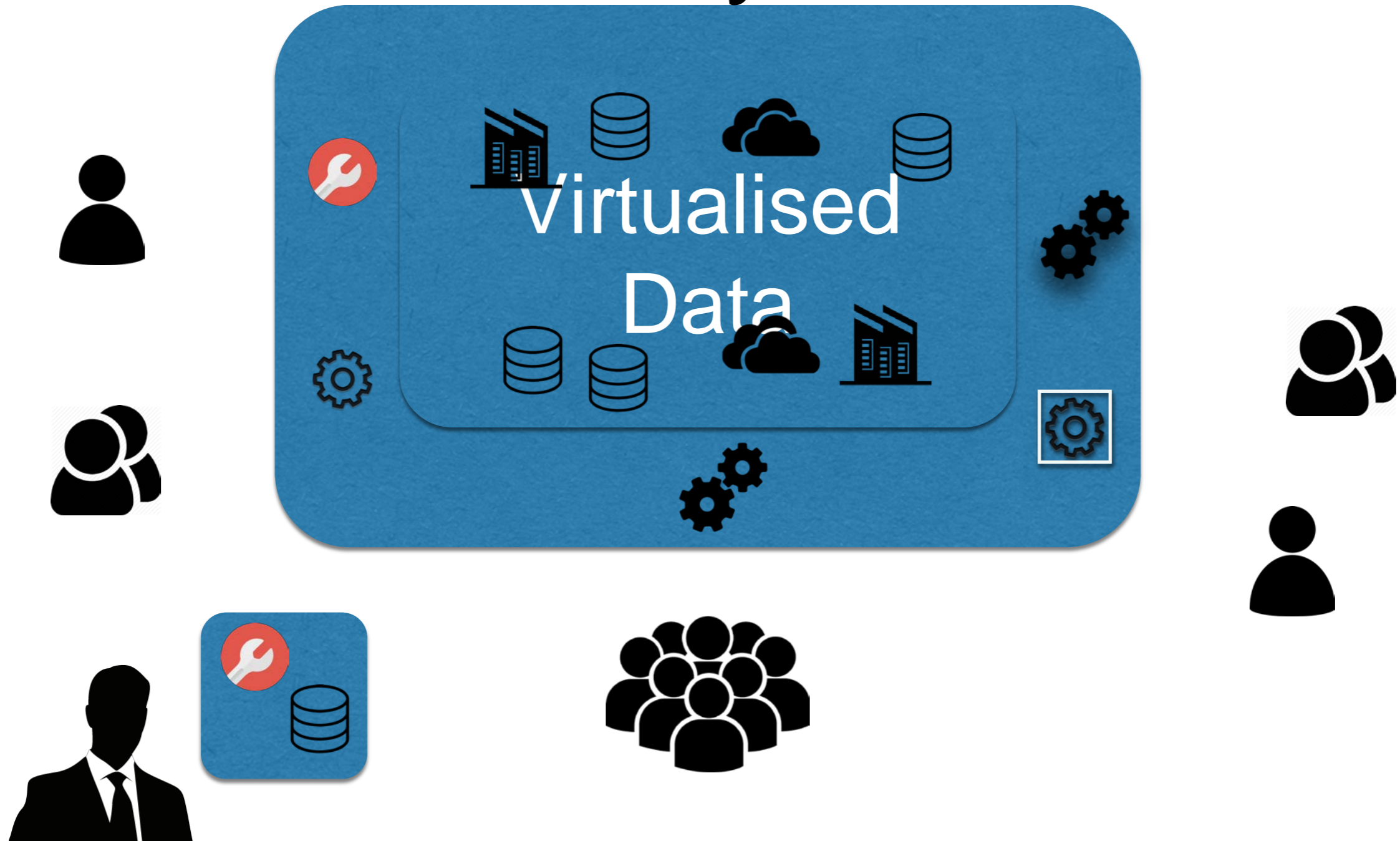
Data Accessibility and Analysis



Data Accessibility and Analysis



Data Accessibility and Analysis

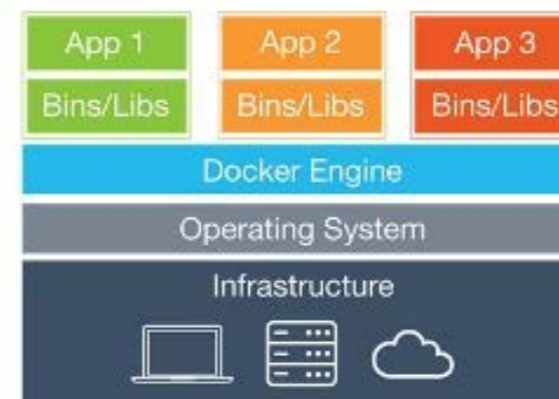




- Docker is an [open-source](#) project that automates the deployment of [Linux applications](#) inside [software containers](#)
- Docker provides an additional layer of abstraction and automation of operating-system-level virtualisation on [Linux](#)
- Docker can be integrated into various infrastructure 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.

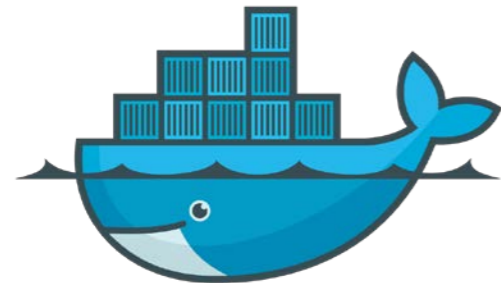


Virtual Machines



Containers

NGS easy



docker

+



=



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/config_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

SECURE
COLLABORATION



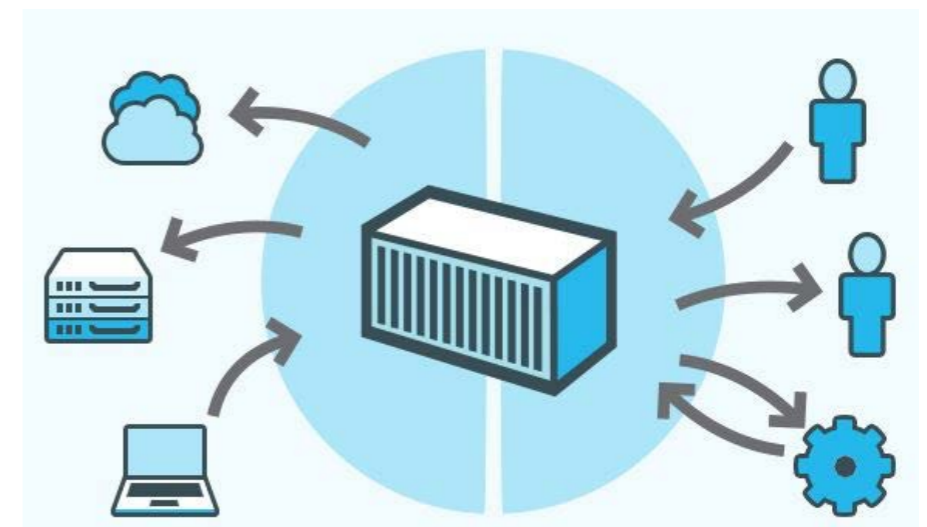
WORKFLOW
AUTOMATION



DATA
DISCOVERY



- Dockerized analysis pipelines

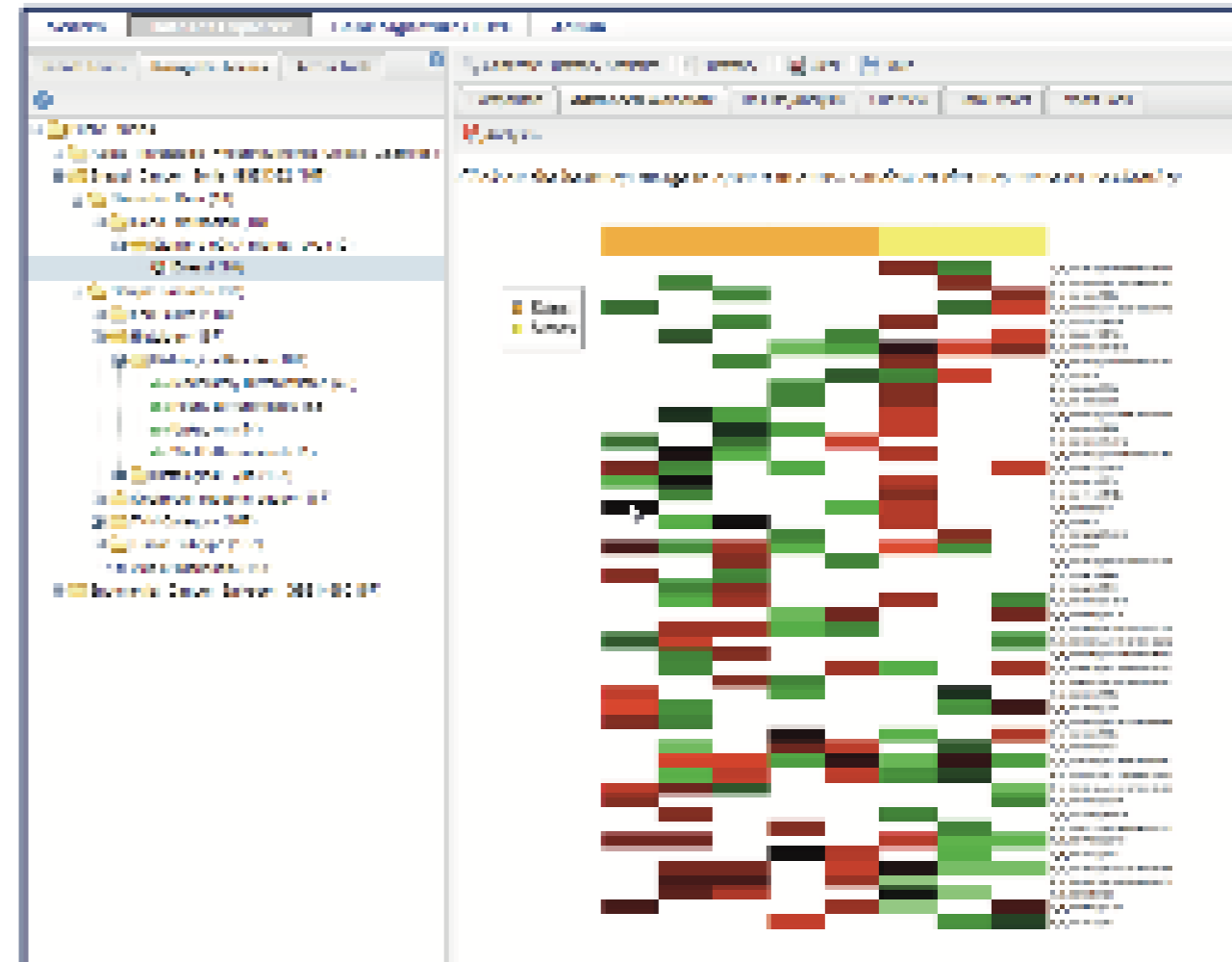




tranSMART

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

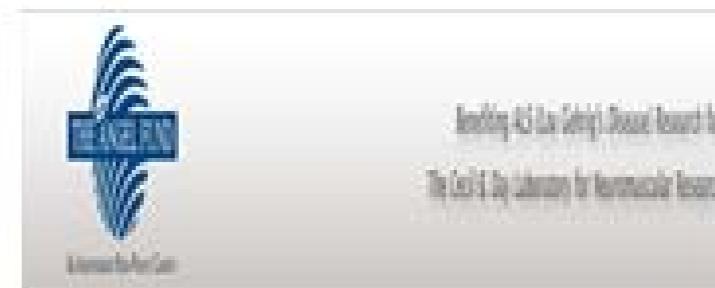
- tailor the system according to our needs

Future Plans

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



Leading science for better health



Economic and Social Research Council
Shaping Society





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

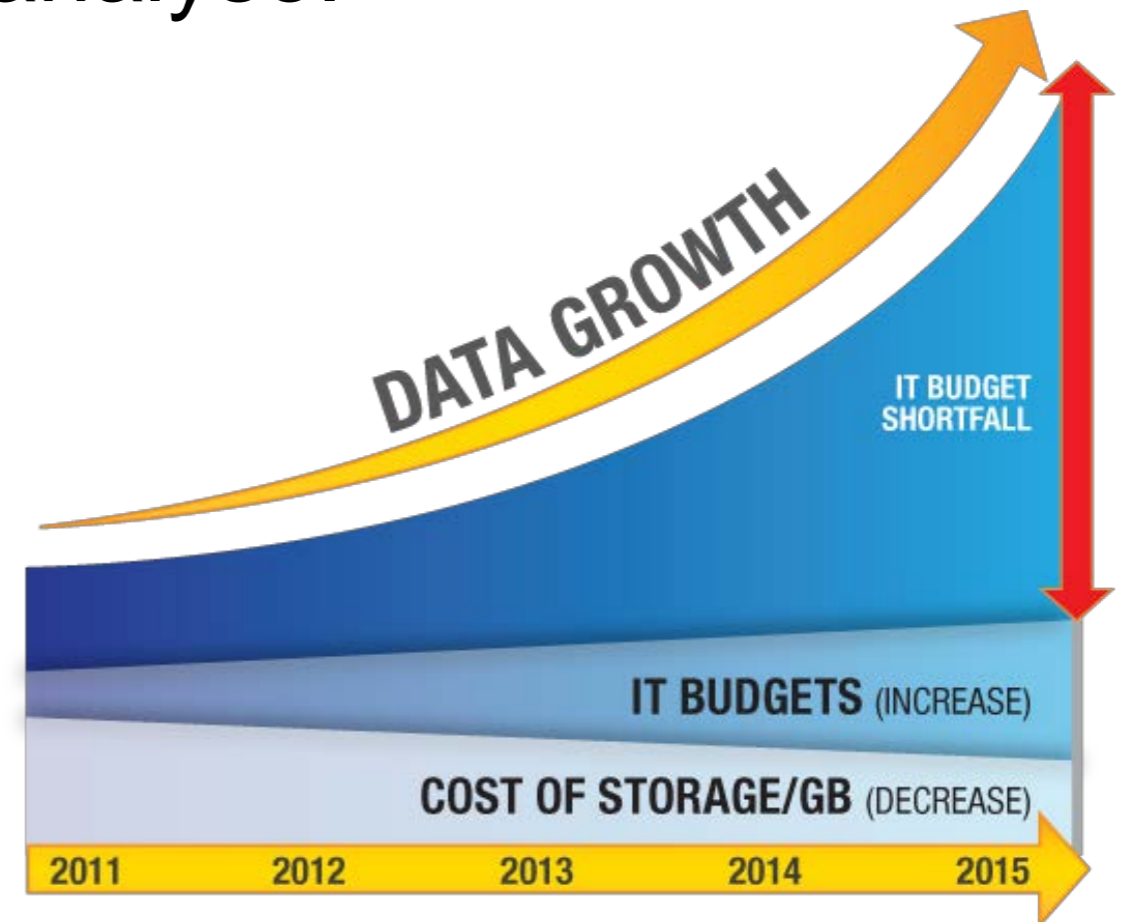
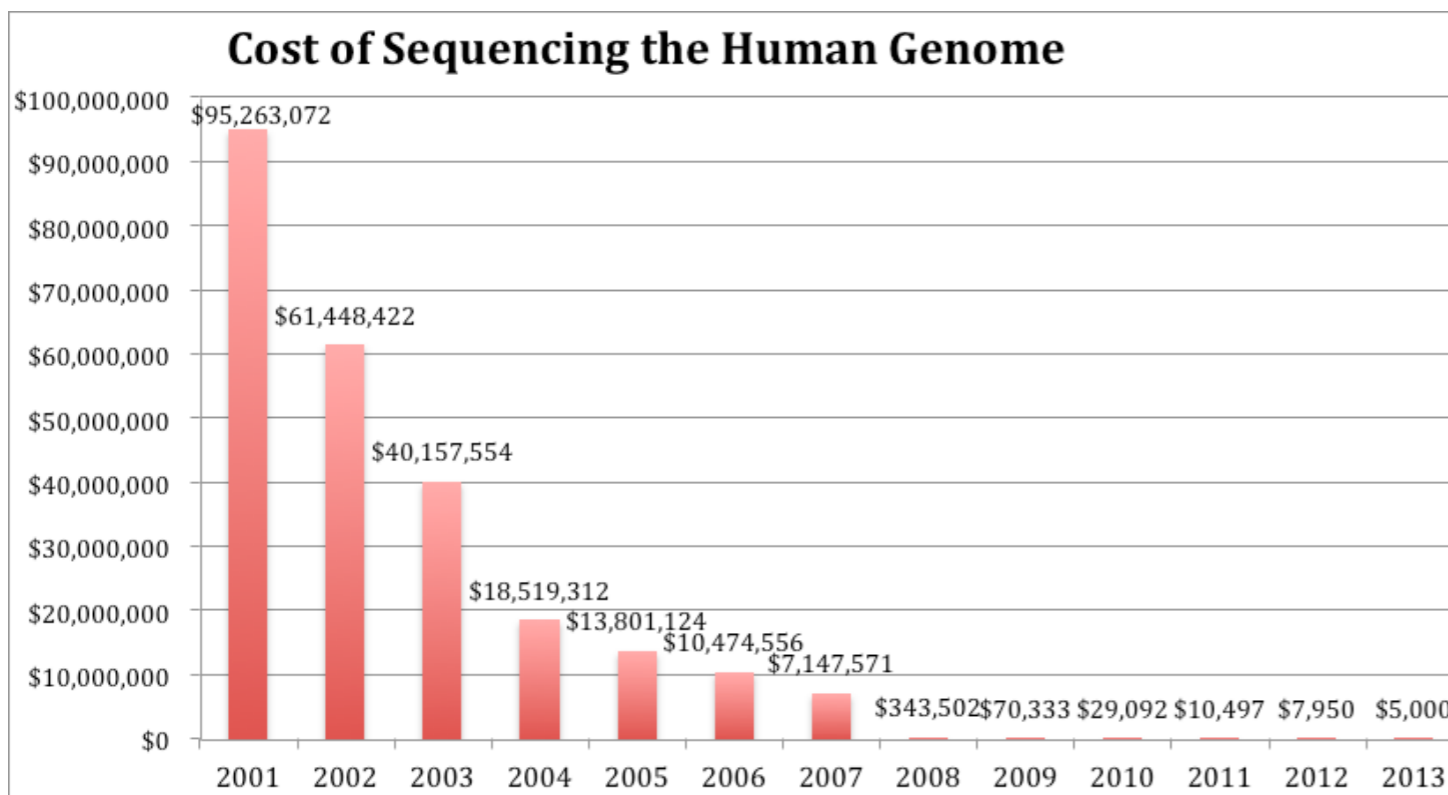
All our collaborators!!!!!!!!!!



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:**

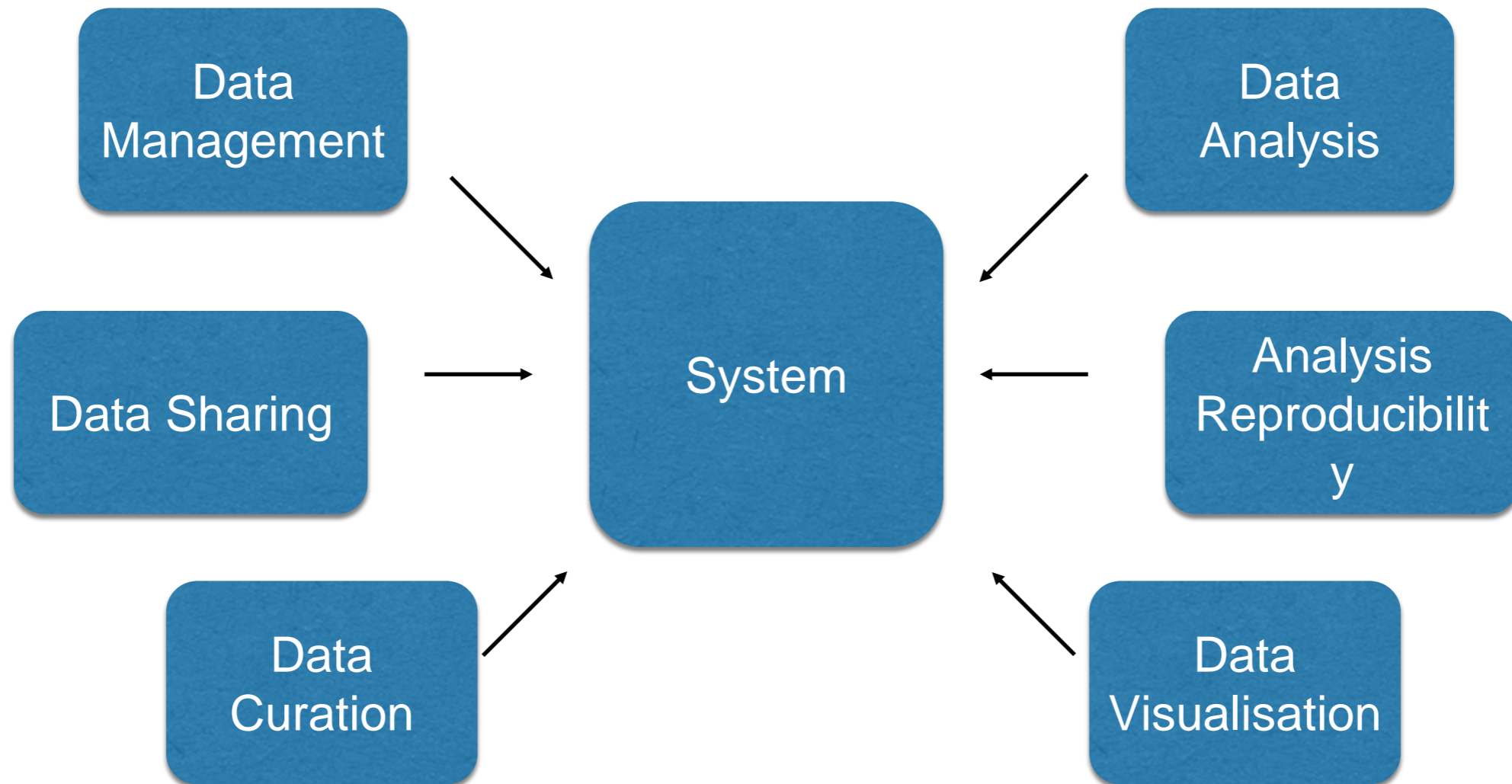
We are now collecting huge amounts of multilayered genetic, epigenetic, environmental and clinical data, much of which is difficult to share and therefore to analyse.



- **Project aim:**

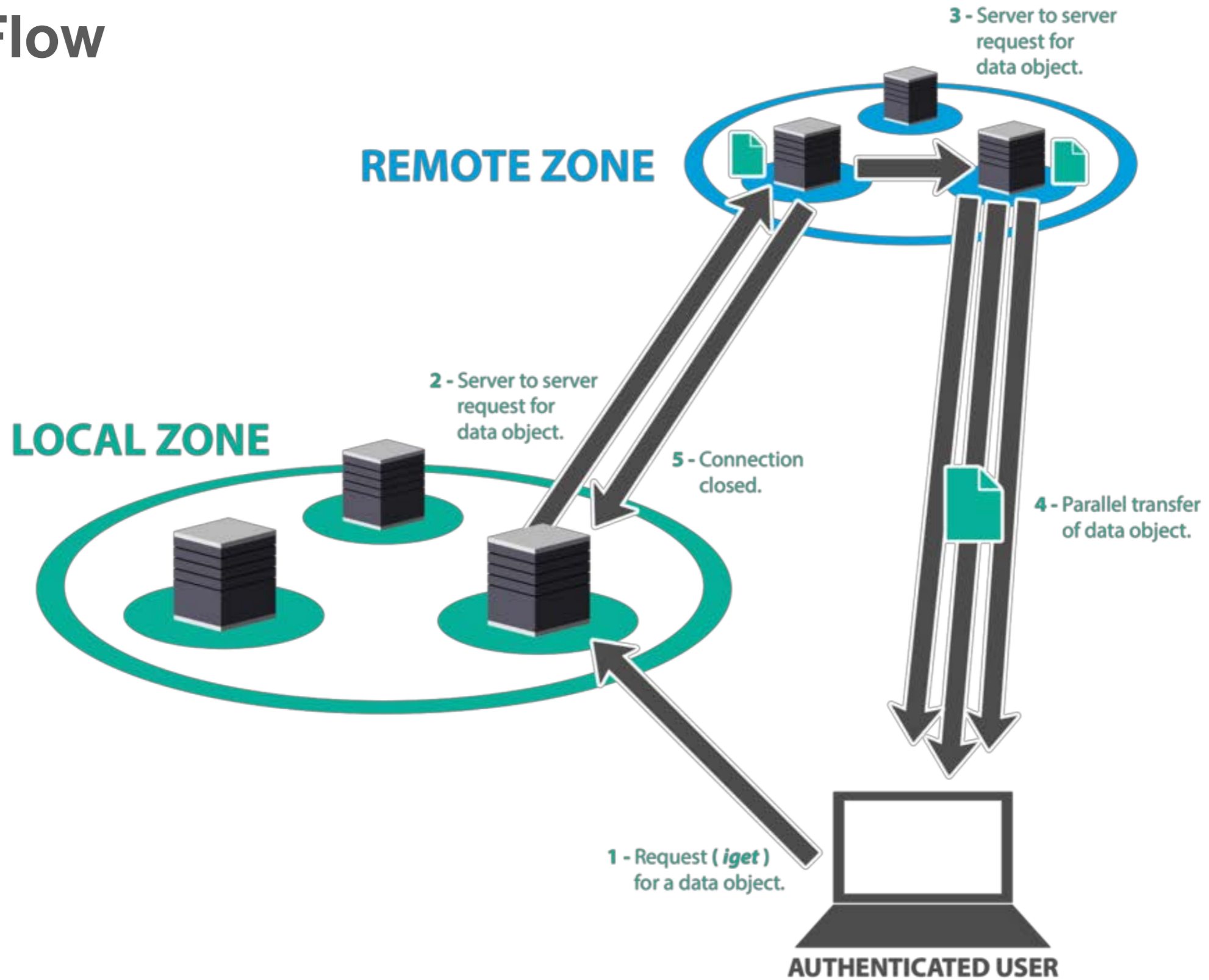
This project is a collaboration between ALS researchers and bi

Data Management and Analysis System



iRODS

Data Flow




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

