DONDERS INSTITUTE

DELL EMC | intel

The HPC Summit – Netherlands

Thursday 27 October 2016
Amsterdam

# Life Cycle Research Data Management

*Hurng-Chun (Hong) Lee*

Radboud University Radboudumc

# Donders Institute

- a cognitive neuroscience research institute
- 600 researchers from more than 35 countries
- Radboud University, Radboud UMC and Max Planck Institute for Psycholinguistics
    - 3 centres in 4 administration domains

# “Data” of Donders Institute

- unstructured data
  - text, audio, video, imaging/signal data, etc.
  - ~100 TB per year

- files in various sizes (KB - GB)

- organisation depends on research project
  - a few hundred projects per year
  - raw data from ~50 labs
  - data stored on central filer, portal drives, desktop computers

# Research Data Management

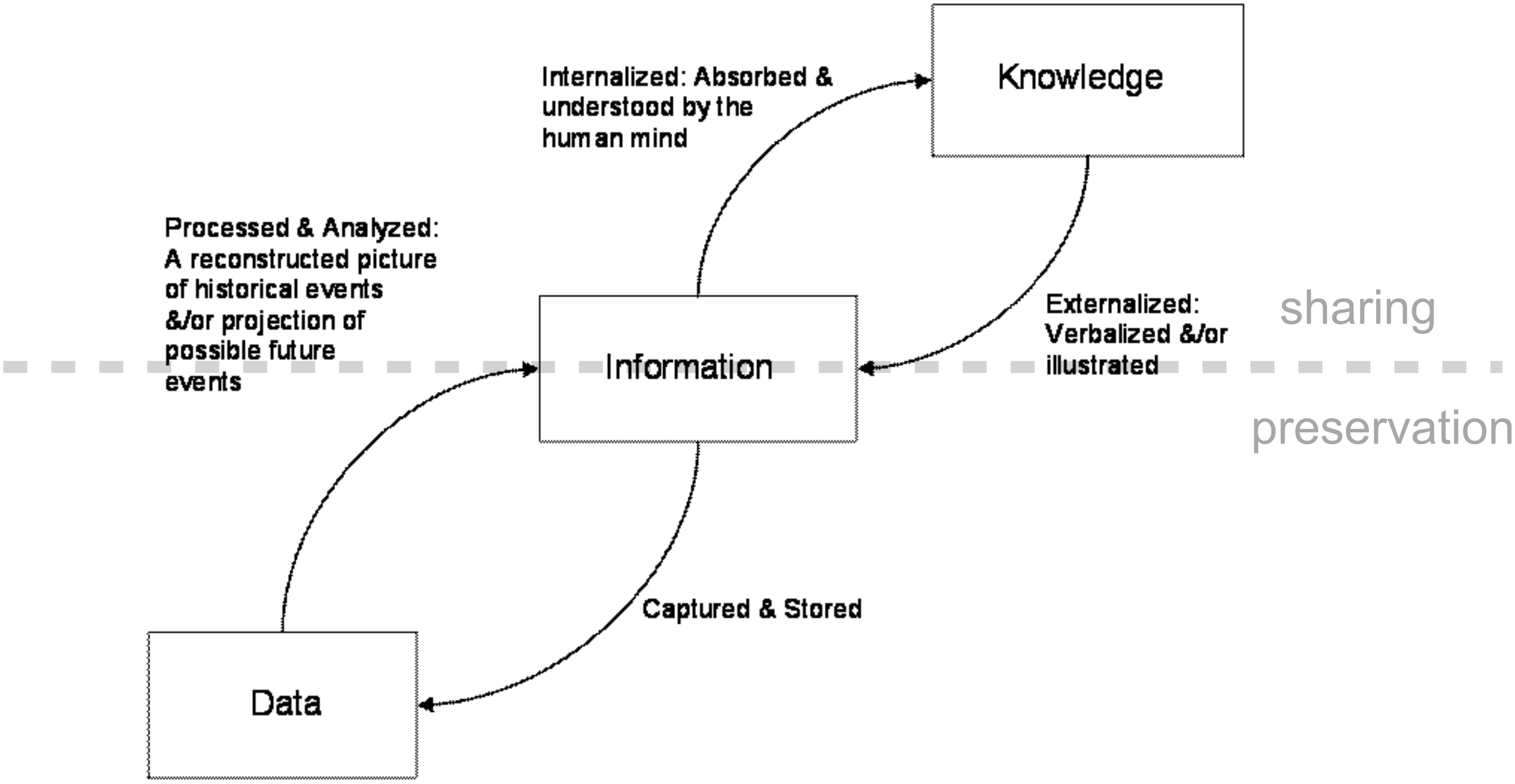
- Era of data-driven science
  - funding agency: data management plan
  - university: good research practice
  - researcher: data sharing

- In 2014, DI was selected to pilot a RDM dev. project funded by the University
  - administration heterogeneity
  - data complexity

# More than an ICT project

“Data management is the development, execution and supervision of **plans**, **policies**, **programs** and **practices** that control, protect, deliver and enhance the value of data and information assets.” - DAMA (www.dama.org)

# More than managing the “Data”



Knowledge

Internalized: Absorbed & understood by the human mind

Information

Processed & Analyzed: A reconstructed picture of historical events &/or projection of possible future events

Externalized: Verbalized &/or illustrated

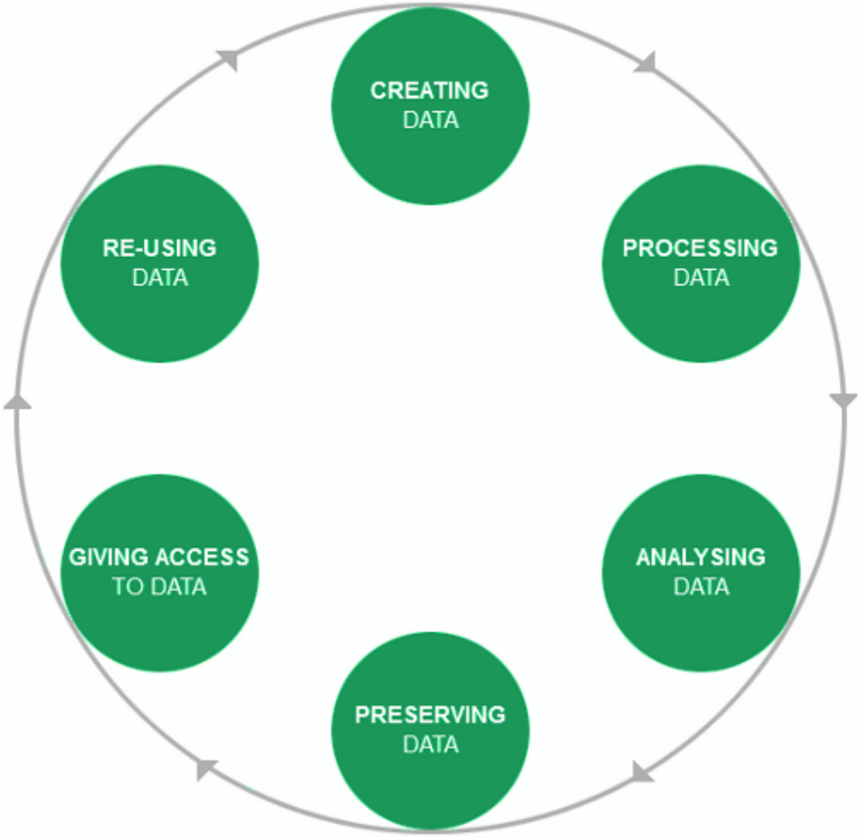
sharing

preservation

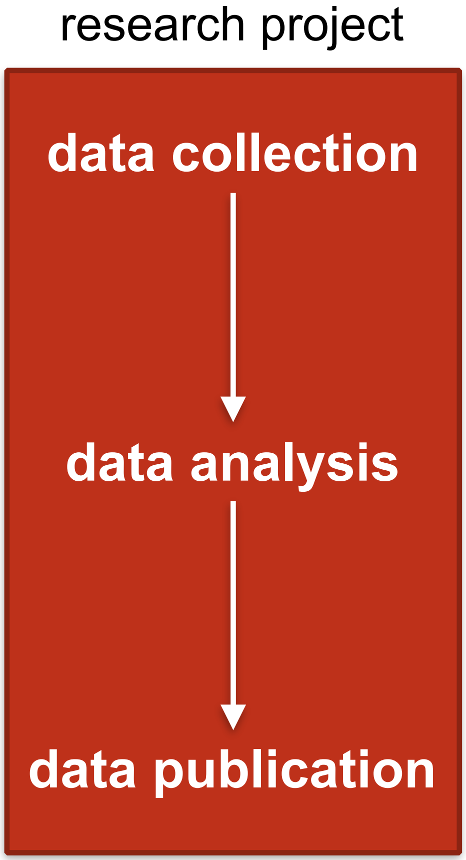
Data

Captured & Stored

# Lifecycle of research data



*source: data-archive.ac.uk*

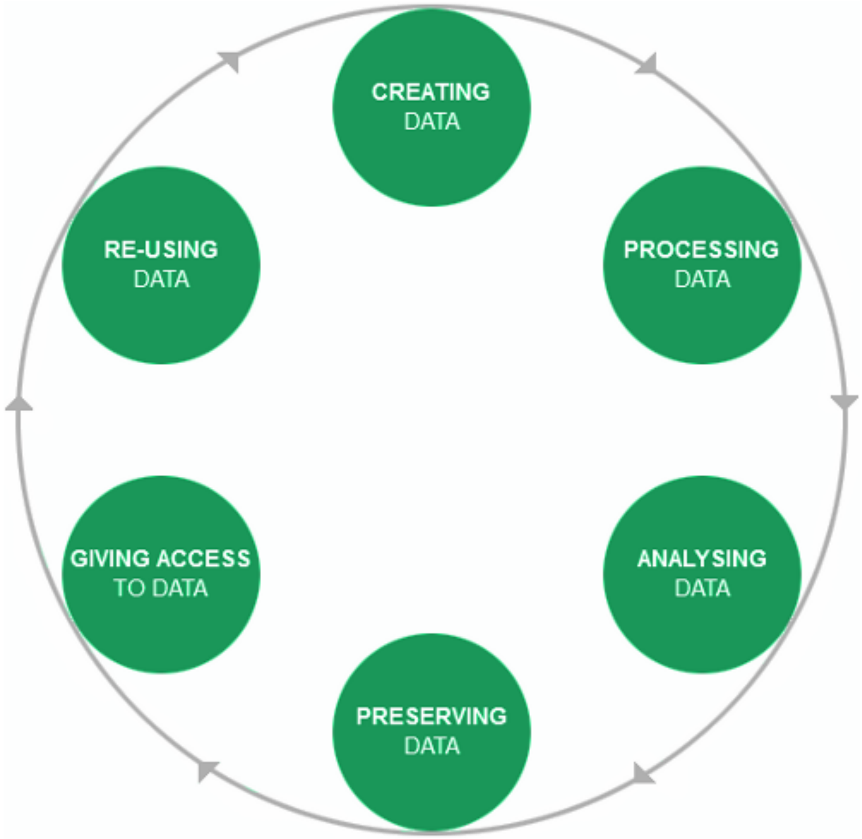

research project

data collection

data analysis

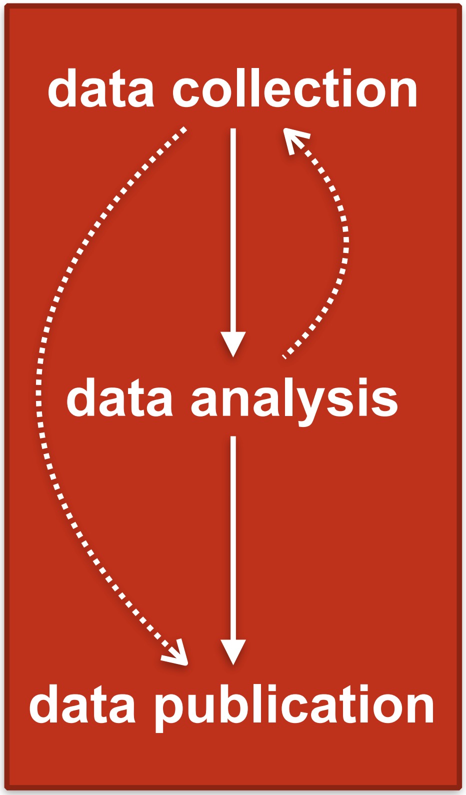
data publication

# Lifecycle of research data

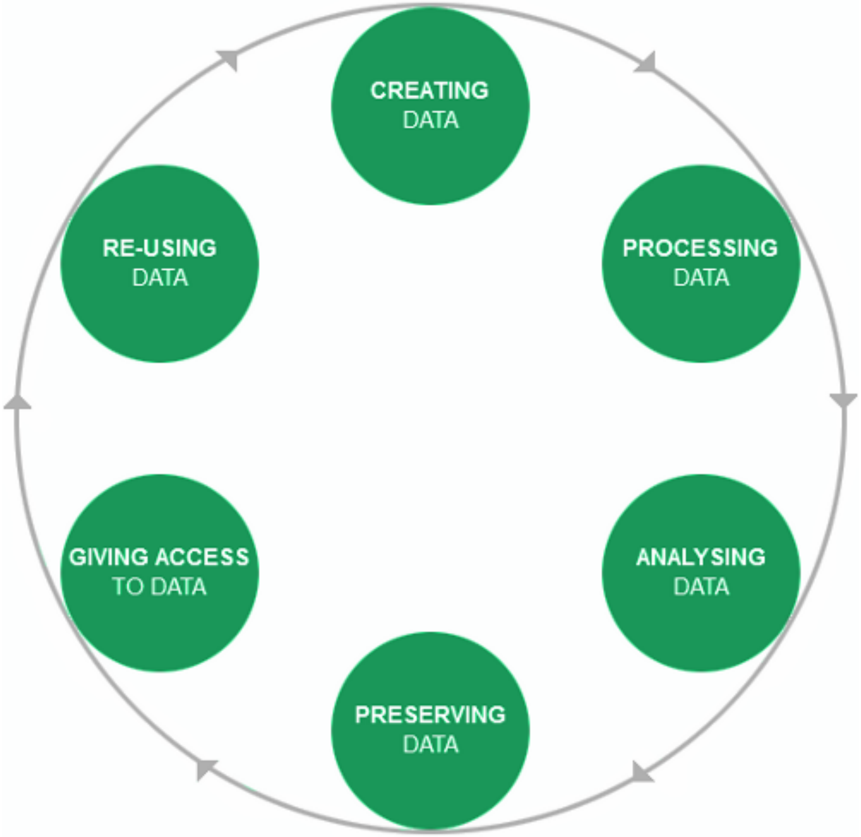


*source: data-archive.ac.uk*

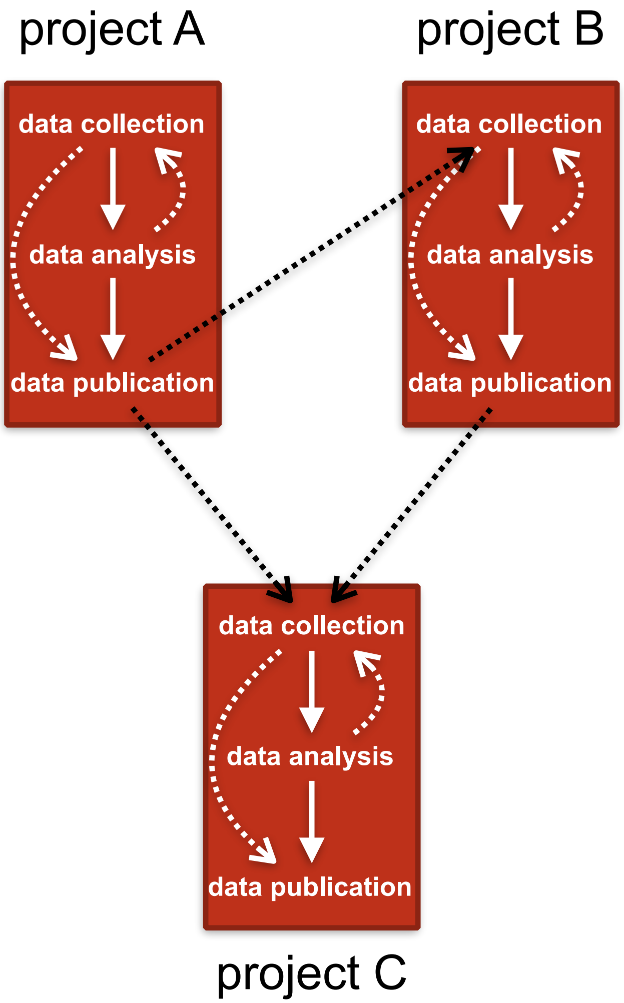
research project



# Lifecycle of research data



*source:* *data-archive.ac.uk*



# off-the-shelf (generic) solutions?

The Dataverse Project

EUDAT

zenodo

DANS

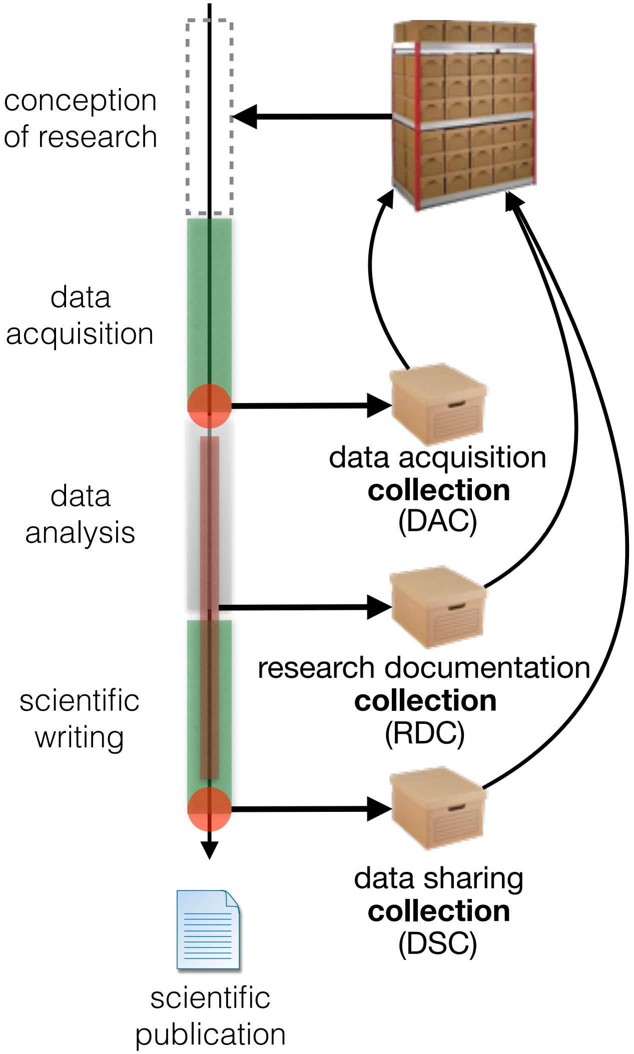
4TU. CENTRE FOR RESEARCH DATA

They are all very good for “publishing data” (and some other aspects); **but** …

# The DI-RDM approach

- protocol:
  - jointly developed by researchers, ICT support, University library, legal department

- ICT system:
  - decoupled from the HPC system, for the sake of simplicity


conception of research
data acquisition
data analysis
scientific writing
scientific publication
data acquisition **collection** (DAC)
research documentation **collection** (RDC)
data sharing **collection** (DSC)

# The protocol

- formal rules/procedures guiding researchers and administrators to work together for managing research data *(the “business model”)*
  - **terminologies** for communications/interactions between researcher and administrator
  - **workflows** of accessing, managing, documenting and sharing datasets
  - **authorisation** and **responsibility**

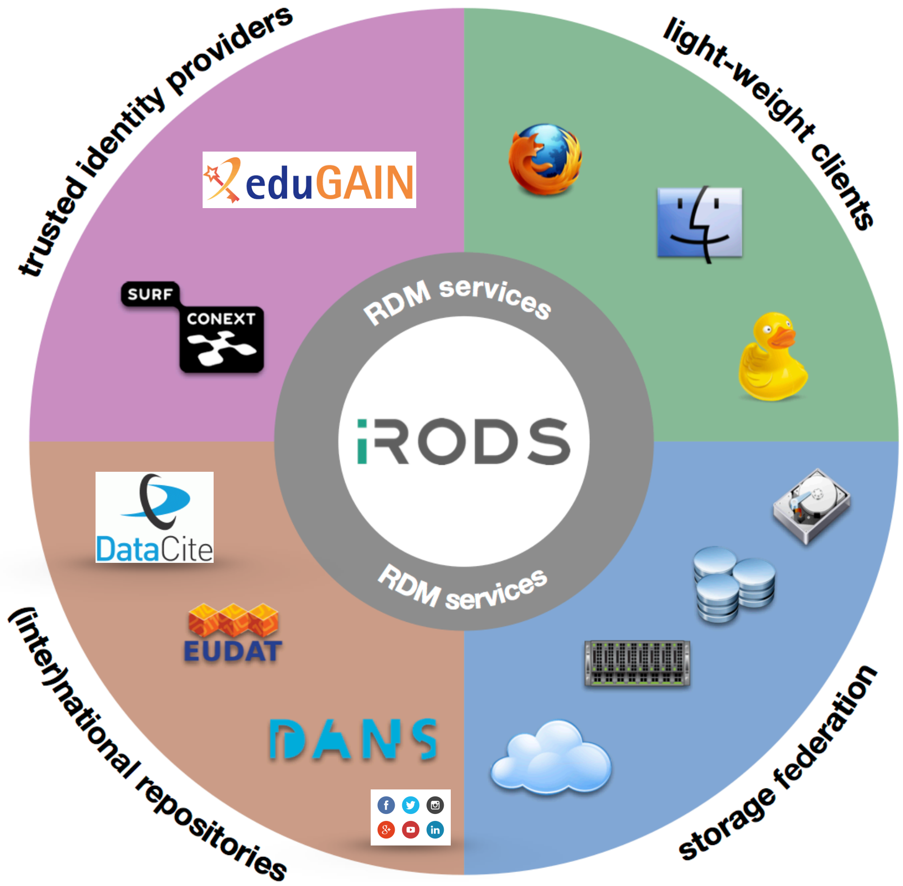
- from which the functional requirements of the ICT system are determined

# The ICT (eco-)system

separation of **management** and **data-access** interfaces

Why iRODS (integrated Rule-Oriented Data System)? http://irods.org

- storage abstraction, federation
- policy-driven data management
- multiple client interfaces



# Current status

- we are rolling out the core system to the Donders Institute
  - https://data.donders.ru.nl
  - enable your IdP for user sign-up via SURFConext!!

- RDM protocol (temporary location):
  - donders-institute.github.io/rdm-wiki

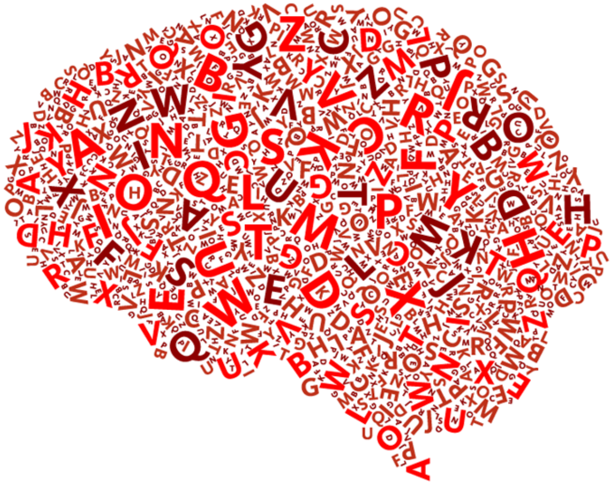
# Lessons learned so far …

- Well-defined terminology is important for communication

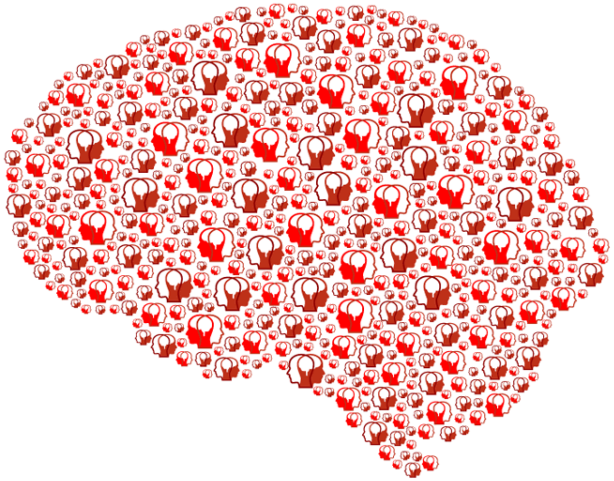
- It’s worth spending time on clarifying the “business model”

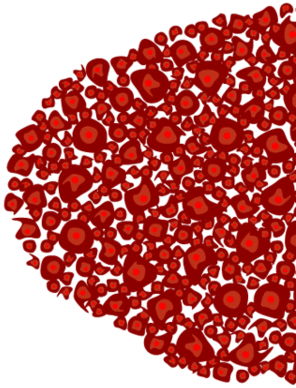
- ICT issues to overcome
  - Identifying user in a trustable way is a must
  - Data upload/download is not trivial
  - Seamless integration with HPC is a request

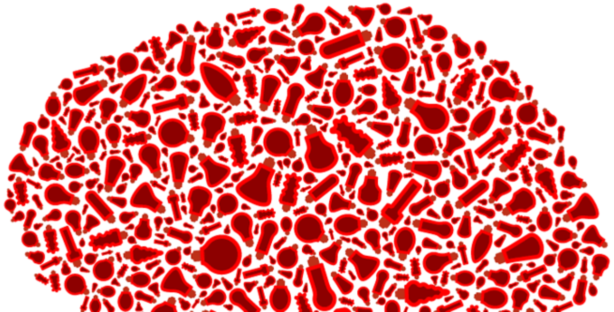
DONDERS INSTITUTE



data.donders.ru.nl







Radboud University Radboudumc