eScience Activities Overview

David Rodriguez Gonzalez
(Edinburgh DIR/SBIRC)

National e-Science Centre

The following Universities are charitable bodies, registered in Scotland, with registration numbers as below.

Scottish Funding Council
Promoting further and higher education

UNIVERSITY OF ABERDEEN SC013683

UNIVERSITY OF DUNDEE SC015096

THE UNIVERSITY OF EDINBURGH SC005336

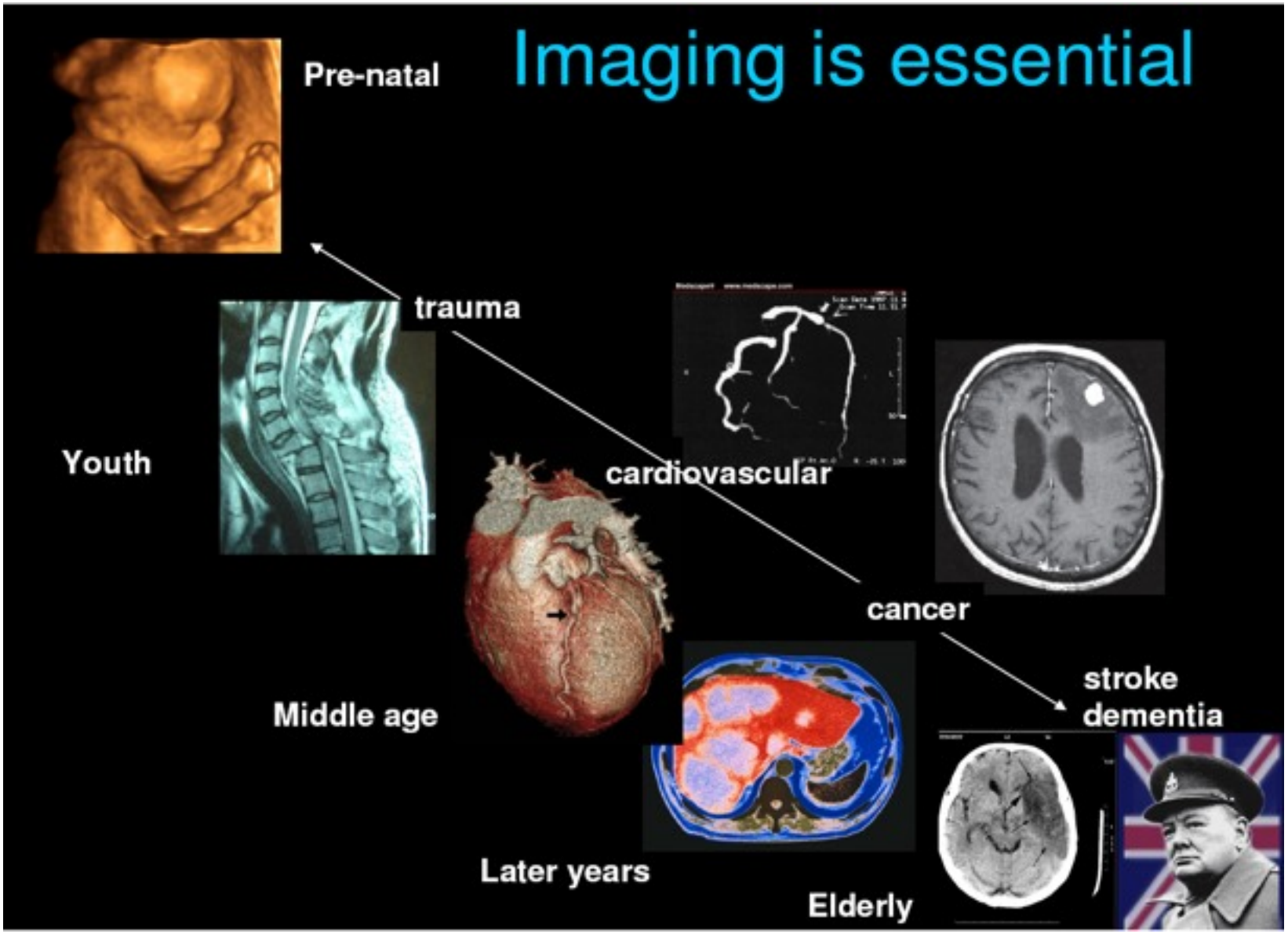
University of Glasgow SC021474

University of St Andrews SC013532

UNIVERSITY OF STIRLING SC011159

Chief Scientist Office

Slide by J. Wardlaw

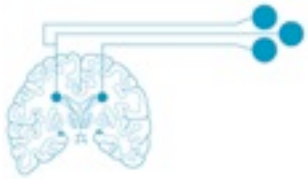


Slide by J. Wardlaw

Massive expansion in research imaging

- All branches of medicine – particularly brain
- Not just medicine – psychology, linguistics, engineering, parapsychology, etc.
- In Scotland too!!!
  - 8% UK population
  - 12.5% of all highest rated departments.
  - Highest concentration of biotech in Europe
- Neuroscience – much larger than NIH
  - But in 2006 there were machines, pockets of excellence, but little cohesion

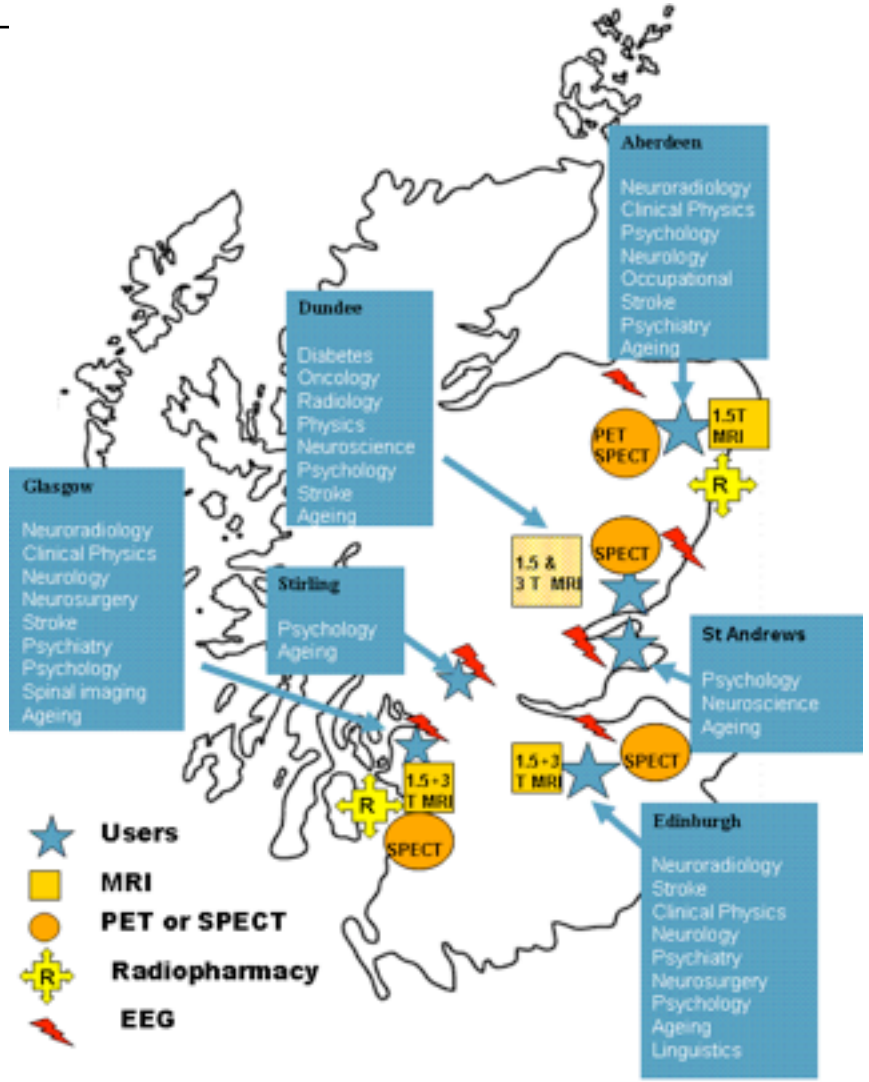
The SINAPSE Project



- Stands for ***Scottish Imaging Network: a Platform for Scientific Excellence***.
- Pooling initiative of six Scottish universities: Aberdeen, Dundee, Edinburgh, Glasgow, St. Andrews and Stirling.
- Main objectives:
  - develop imaging expertise,
  - support multi-centre clinical research in conjunction with the Clinical Research Networks,
  - improve the ability of neuroscientists to collaborate on clinical trials,
  - have a direct impact on patient health.

SINAPSE priority projects

- Stroke, the brain and the blood-brain interface
- Ageing brain to dementia
- Novel molecular imaging markers for major psychiatric disorders
- Innovative radiotracers for CNS inflammation



e-Science for SINAPSE

- Sharing of research data and applications between centres is an important part of the SINAPSE project's objectives
  - The increasing amount of data acquired in modern imaging facilities and the distributed nature of SINAPSE require a proper data management strategy
- National e-Science Centre actively involved in the SINAPSE collaboration
  - Mainly through the IT & Image Analysis Committee

eScience project activities

- Information governance & data de-identification
  - Networking
  - Development of de-identification tool
- Data sharing infrastructure
  - Facilitating multi-centre studies
- Portal for brain imaging
  - Improving usability
- Other
  - Analysis methods

Data Sharing e-Infrastructure

- Enabling multi-centre clinical research through data sharing
- Some features of the proposed of the SINAPSE e-infrastructure are:
  - **Privacy protection**, *automatic compliance with data protection policies*;
  - **Security**, *advanced authentication and authorisation within projects*;
  - **Usability**, *providing a user friendly environment to access data and applications*;
  - **Modularity**, *conforming to relevant standards and use of existing components*;
  - **Centralisation**, *leveraging existing compute clusters and storage*.

Centralised Architecture (pros & cons)

- Simpler Deployment
- Easier middleware release control
- Lesser impact in participant centres
- Easier to manage and use
- No default resilience
  - A second centre would be needed
  - But this is only necessary for critical services
  - With a good support a reasonable service can be provided using a single centre

Deployment

- ECDF (Edinburgh Compute and Data Facilities)
  - http://www.is.ed.ac.uk/ecdf/)
  - A singular facility along Scotland
- Disk space and CPU time can be rented
  - 1456 CPU cores
  - 275 TB of disk
- A dedicated server hosted by ECDF:
  - For SINAPSE specific services
  - ECDF provides basic hardware + software support

Data Formats

- Raw data
  - Imaging data usually in standard data format (DICOM)
  - EEG & spectroscopy format is manufacturer dependent
- Processed data
  - Varies from project to project

Digital **I**maging and **Co**mmunications in **M**edicine (**DICOM**)

- Standard for handling, storing, printing and transmitting medical imaging information
  - Supports CT, MRI, PET, Nuclear medicine, ultrasound,…
  - Several types of objects: Images, Presentation States, Structured Reports, Encapsulated Objects
- Data format:
  - "Header": includes metadata
  - Pixel data
- Also defines communications, confidentiality profiles, …

DICOM Files

- Enhanced Multi-frame DICOM CT & MR objects support storing a whole series in a single file
- Unfortunately this is still not widely adopted/supported
  - Thousands of very small files (even of 20KB)
  - Performance problems

Data Protection Act

- UK's Data Protection Act (1998). Implements the European Community Data Protection Directive 1995.
- Establish individuals' rights on data held about them and obligations for organisations or people processing personal data.
- Personal data must be processed in a fair and lawful manner.
  - 8 DPA principles.
- Other legislation pieces apply to medical data.
  - Common law: duty of confidentiality.
  - Human Rights Act 1998 (article 8).

DPA in research

- The DPA does not define the term "research purposes" apart from clarifying that it includes statistical or historical purposes.
- Data processing for research should be 'compatible' with the purpose for which the data were originally obtained.
- The data subjects should be aware that their personal information will be used for research purposes.

Anonymous Data

- Coded (pseudonymised or linked anonymised) data:
  - the identifiable information has been substituted by alphanumerical sequences with no plain meaning.
  - The data is anonymous to the research team.
  - The key to reverse the transformation shall be held securely by a third party to avoid falling into the DPA.
- (Fully) Anonymised data:
  - all personal identifiers or codes have been irreversibly removed.

Personal Data in DICOM

- As they are used in clinical workflows DICOM objects include many attributes with personal information
  - Some times personal data is found also “burned in” the pixel data
  - There is a potential risk for face recognition in 3D reconstructions (MRI)
- Considerable number of de-identification tools, but
  - Some do not do the job
  - Lack of flexibility
  - Bad performance
  - Linked to specific suites or frameworks

PrivacyGuard

- A DICOM de-identification toolkit
  - Implemented in Java
  - Highly configurable
    - Users can define their own de-identification methods
    - Mechanism for chaining different operations
- Privacy Policies expressed in XML documents
  - PolicyEditor: Privacy Policies authoring tool
- DICOM read/write through an interface that allows using different libraries
  - dcm4che2
  - pixelmed

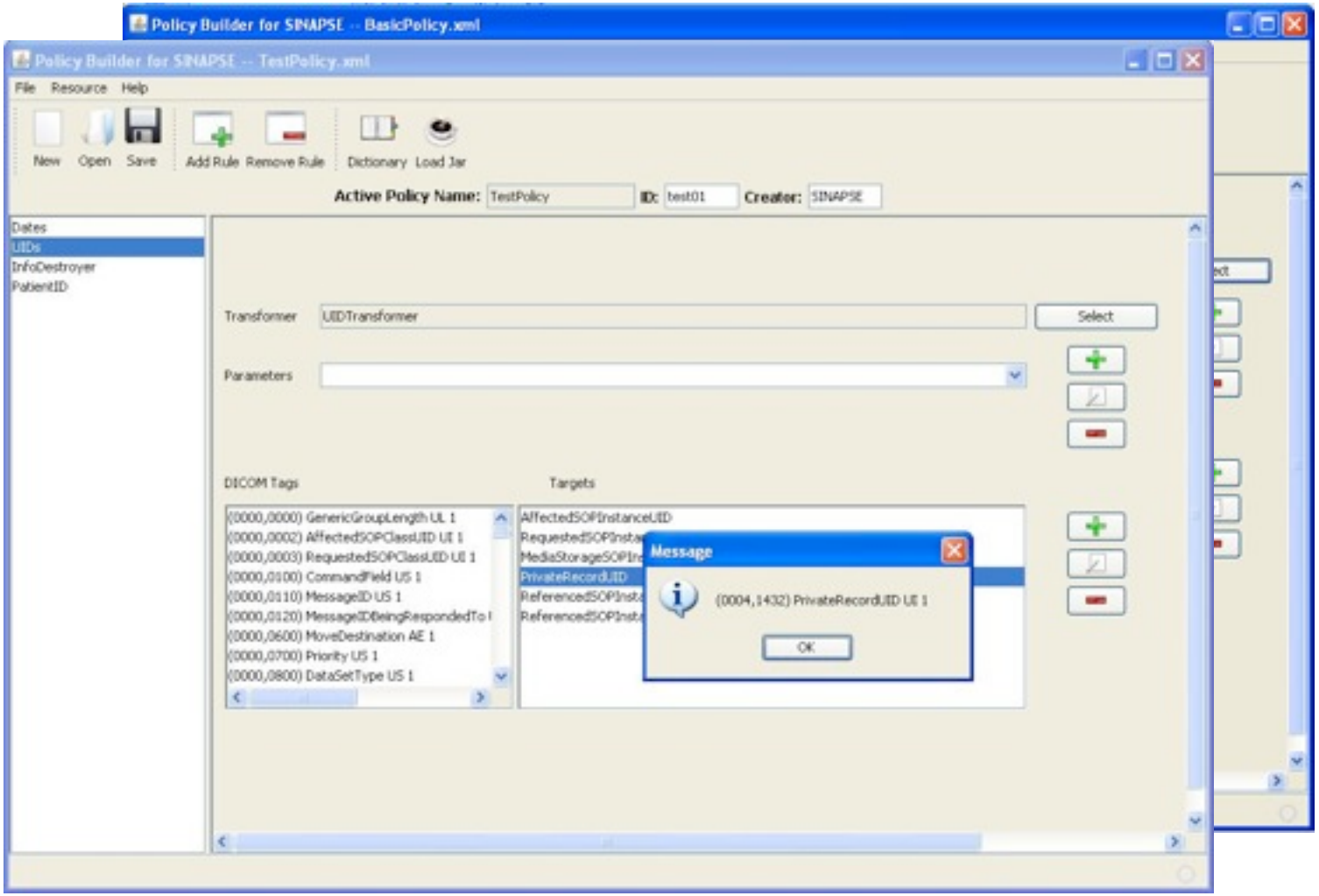
Privacy Policies

- XML documents containing the rules for anonymising the data
- A rule specifies:
  - The target fields
  - The class used for the transformation including:
    - Version
    - Digest
    - Resource (jar file)
  - Parameters

Policy Editor

- A Privacy Policies authoring tool
- Includes a DICOM dictionary
- Can look for de-identification classes in jar files
- Can sign and check the signatures of policies

Policy Editor



Pending: Face de-identification

- This is done by independent applications after volume reconstruction (Analyze or Nifti format)
  - So it is done once the data has already been passed to the researchers
- Would it be possible to do it before directly in DICOM?

Future problems: Data Publishing

- Data Publication increasing due to:
  - Regulators
  - Funding agencies
  - Collaboration spirit
- This also applies to the imaging data
- Even after the removal of personal data items there is risk of disclosure through linkage with other published data
  - The unknown background knowledge about the victim complicates the problem

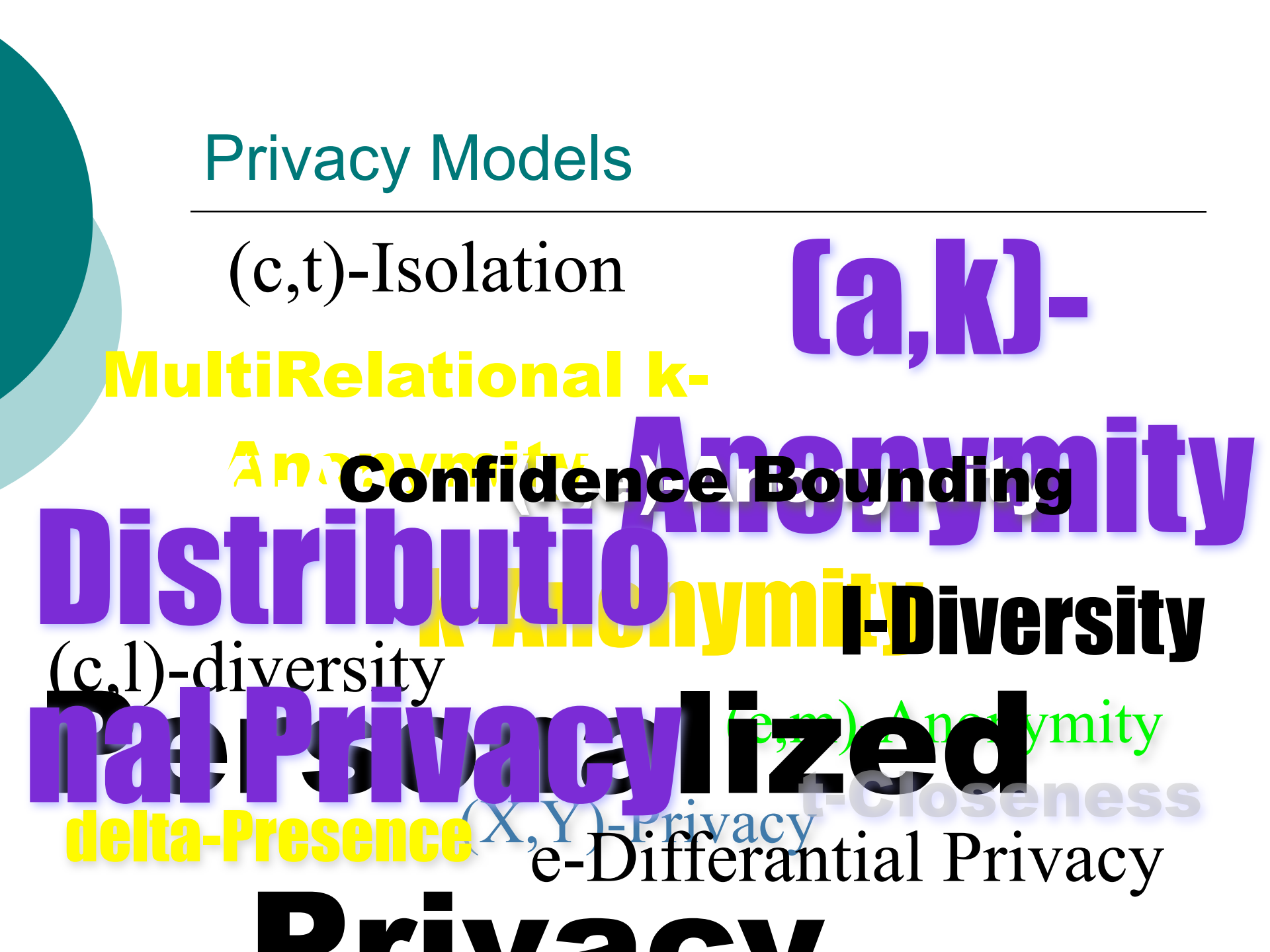
Privacy-Preserving Data Publishing

- See survey by Fung, Wang, Chen and Yu:
  - *Privacy-Preserving Data Publishing: A Survey of Recent Developments* (ACM Computing surveys, vol 42, 4, June 2010)
- Demand for publication of microdata
  - Publish data not the data-mining result
- Current practices lead to excessive distortion or insufficient protection
  - Truthfulness at the record level required in some scenarios
  - Unknown background knowledge

Privacy Protection

- When publishing data the attacker should not learn anything about the target (victim) compared with not publishing the data
  - Practical approach: assume limited background knowledge
- Two types of attacks:
  - Linkage (record, attribute or table)
  - Probabilistic
- Minimality attack on Anonymous data

Privacy Models



Anonymisation Operations

- **Generalisation**: replace a value with more general one
- **Suppression**: remove a value or record
- **Anatomization**: deassociates the relationship between the quasi-identifier and the sensitive attribute by grouping
- **Permutation**: the same by partitioning and shuffling the sensitive values
- **Perturbation**: preserve statistical information by replacing the original data with synthetic

Minimality Attack

- Most privacy models assume that the attacker knows the QID and/or the presence of the target in the table
- The attacker can possibly determine also
  - The privacy requirements
  - The anonymisation methods
  - The algorithm used
- This additional information can facilitate attacks
  - Many anonymisation algorithms follow an implicit minimality principle
  - This can be exploited to reverse the anonymisation

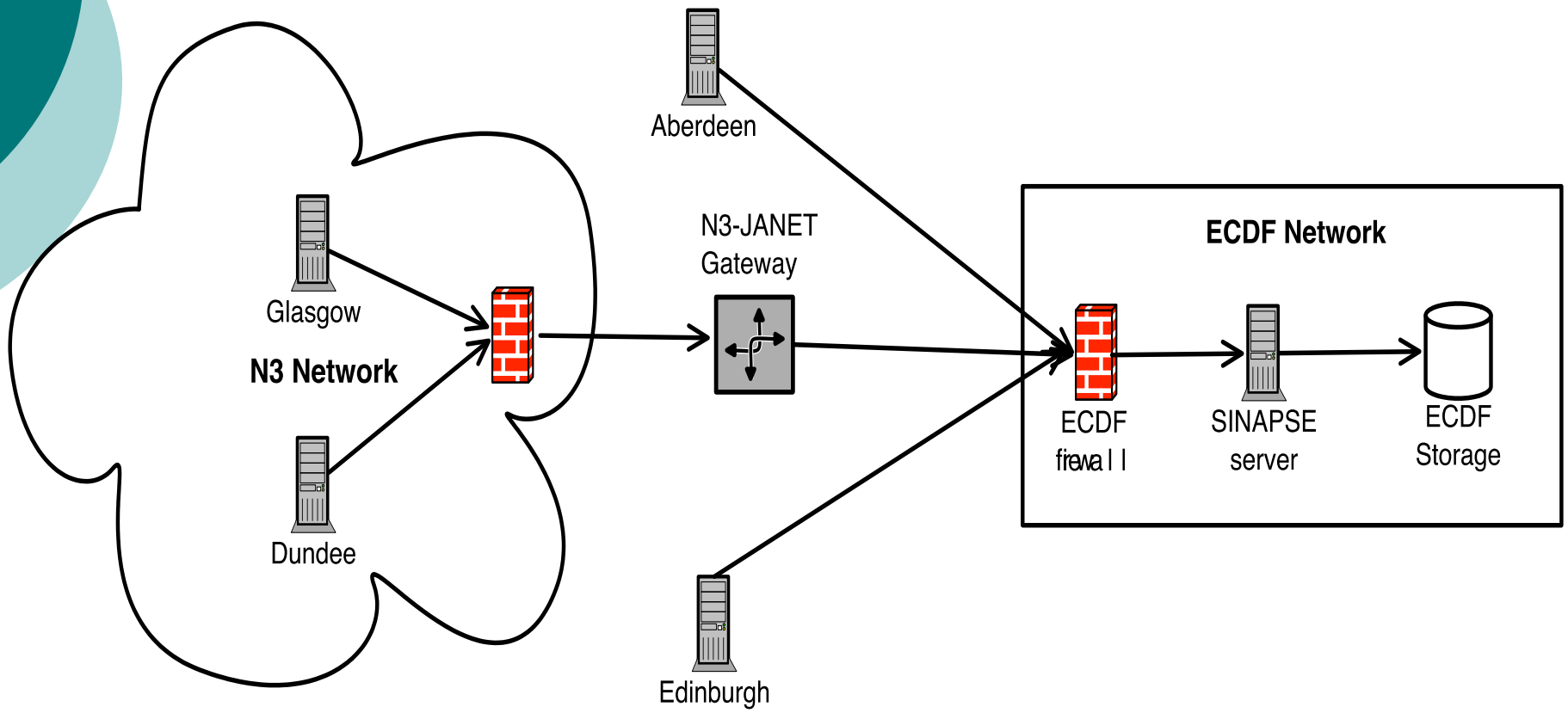
MRI QA in SINAPSE

- QA is used to monitor the performance of MRI scanners
  - particularly important in multicentre imaging studies
- Previous work in SINAPSE towards establishing a common QA protocol
  - 7 participant MR scanners in 4 centres
  - Framework for monitoring the quality of the data
  - It will facilitate the combination of data between centres

Motivation for an automatic system

- Remove the burden of some manual tasks currently being done in the centres
- Allow checking the correctness of the sequence parameters used
- Ensure the consistency of the software used for the analysis and
- Facilitate the reanalysis of the data
- Enforce (pseudo-)anonymisation policies across collaborations

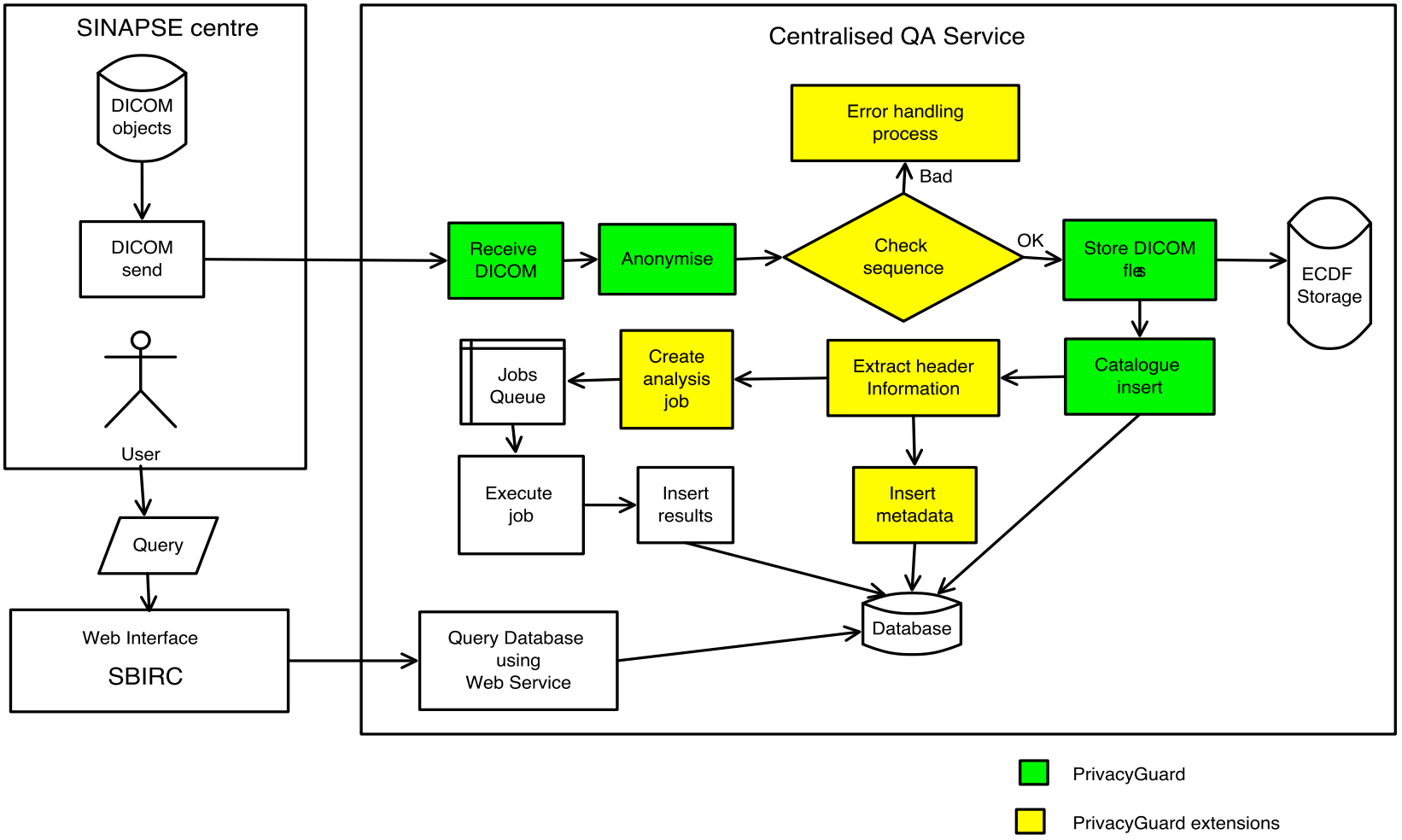
Network Configuration



Networking Issues

- MRI scanners are connected to N3 (NHS network) in some SINAPSE centres
  - Glasgow & Dundee
  - Application included in the N3-JANET gateway
  - Gateway already configured
- And port open in ECDF firewall
  - For the gateway?
  - Limited number of machines from the Universities
  - But still some connectivity issues

MRI QA flowchart & PrivacyGuard



Data storage and Analysis

- After checking the sequence
  - DICOM data is stored in ECDF
  - An entry is inserted in SINAPSE catalogue
- Predefined information is extracted from the header and inserted in the QA database
- An analysis job is created
  - Executed asynchronously
  - The results are also inserted in the QA database
- A web application is used to monitor the QA parameters evolution
  - Accessible from all SINAPSE centres
  - Uses the QA database as backend

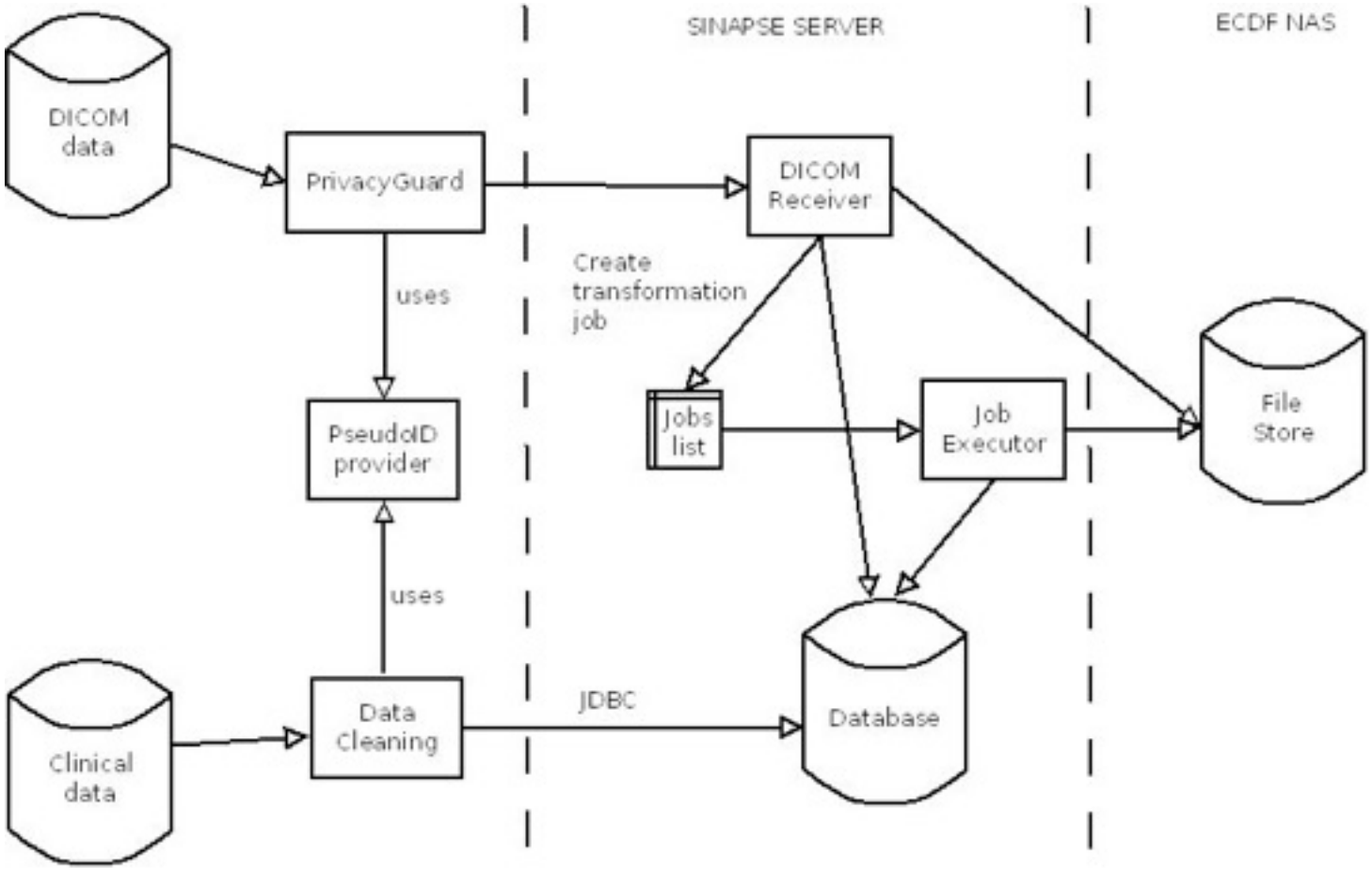
Extending PrivacyGuard

- PrivacyGuard provides a mechanism for adding new functionalities
- We are using it to:
  - Check the sequence correctness
  - Extract metadata from the DICOM header and insert it into the QA database
  - Create the analysis jobs

Image Bank

- Pilot project for a Normative Brain Imaging Bank using SINAPSE infrastructure
  - Server for databases
  - ECDF storage space
  - Portal
- Includes clinical and cognitive data along with imaging data
  - Cleaning process required before importing it to the centralised system

Data cleaning and import process



Other eScience activities in SINAPSE

- RapidBrain Portal:
  - Using Rapid a portlets building technology developed at NeSC
  - Proof-of-concept prototype last summer
  - A production portal to be deployed at ECDF is being built now
    - ECDF and EPCC collaborating
    - A general solution for portal single sign-on authentication to the cluster in place
- GPGPUs
  - Implementation of deconvolution algorithms for brain perfusion imaging (Fan Zhu, SINAPSE PhD student)