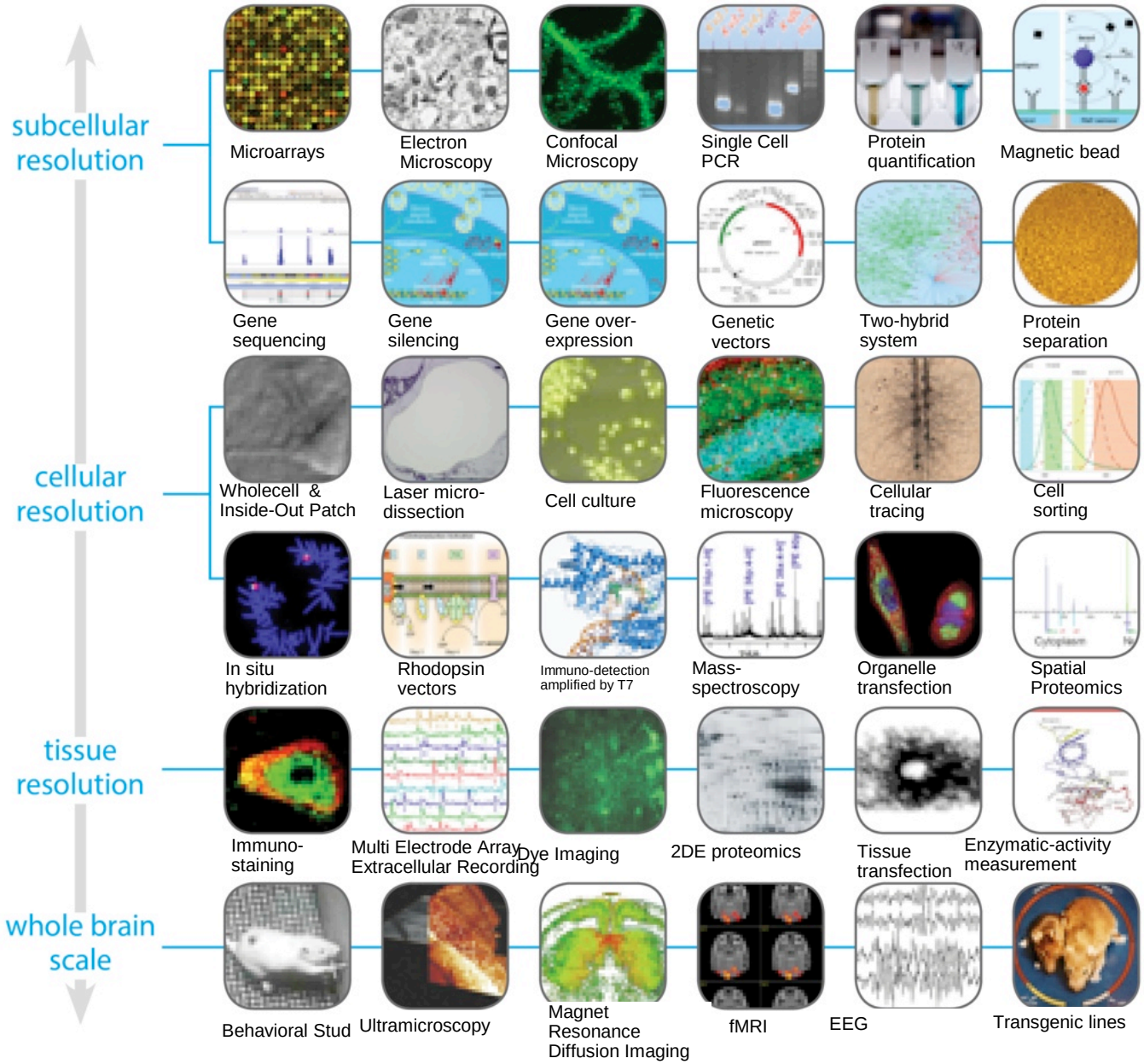
Safe Replication, PIDs and INCF

Raphael Ritz, Scientific Officer
International Neuroinformatics Coordinating Facility
Stockholm, Sweden

raphael.ritz@incf.org

2nd EUDAT User Forum, March 11, 2013, London, UK

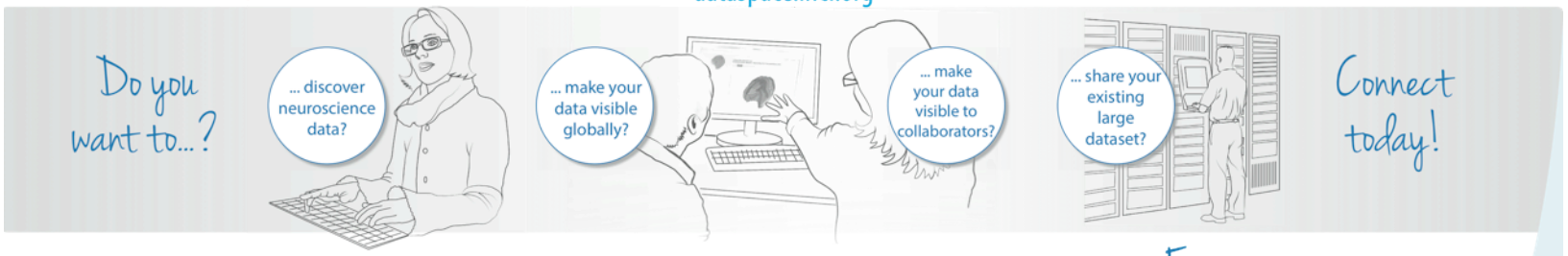
Multiomic Neuroscience Data

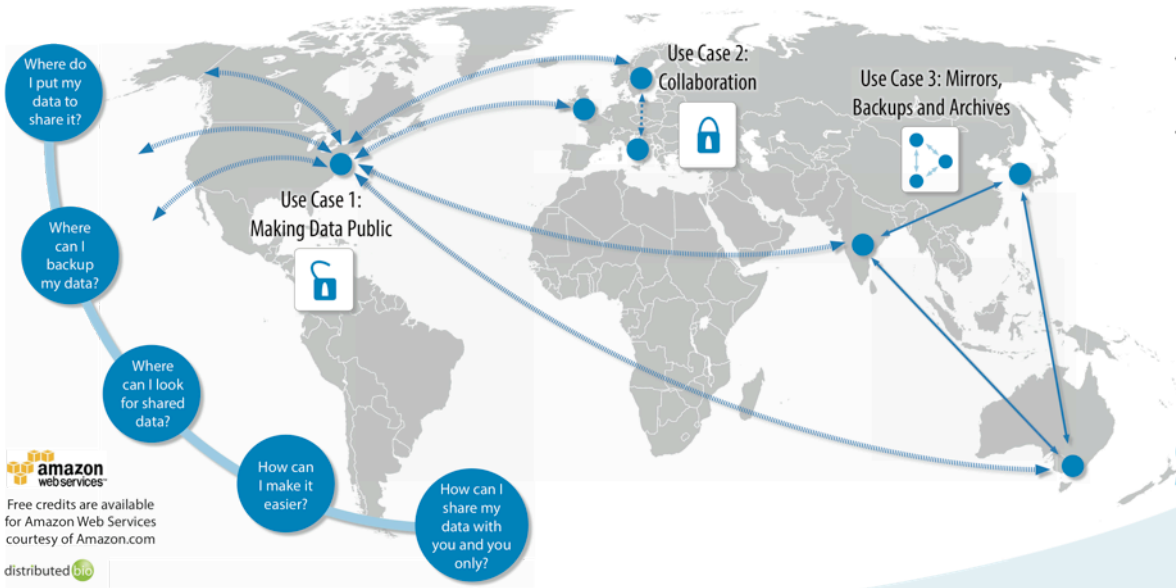


International Neuroinformatics Coordinating Facility

DATASPACE

dataspace.incf.org





Features...

- Access diverse data repositories from around the world through a single resource
- Browse and access data using different user interfaces
  - Web
  - File navigator
  - Command line
- Upload and download data worldwide
- Set/get arbitrary metadata for files and folders
- Search metadata
- Manage large data
- Keep directories synchronized
- Create temporary public or private links to share data

connect share win

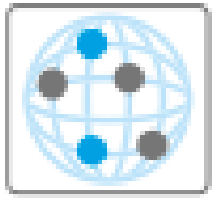
Are you producing valuable data?

Are you producing reusable data?

dataShareAward

incf.org/datashareaward

incf



DATASPACE

- INCF central authentication
- User defined access control (Private, Public, Group)
- Policy based group data access (e.g. data use agreement)
- Standardized navigation structure and policies
- Globally distributed zones - distributed data storage costs

Growing the Federation

- Challenges
  - People already have “some systems” – need to fit existing environments
  - EC2 is hard to pay for - and not necessarily cheaper than a university environment
  - Integrate at application rather than file level
- EUDAT
  - Simple Storage
  - Safe Replication
  - Persistent Identifiers

INCF – EUDAT Federation

- Server at PDC running the eudat.pdc.kth.se zone federated with the INCF data space
- PDC assigns PIDs to data stored on its own resource
- PDC replicates to CSC
- Replication policies are currently ad-hoc

Data considered

- Reference data from the Waxholm Mouse Brain Atlas (Nissl stains)
  - About 600 files/250 GB
- Waxholm Rat next – ca. 1 TB
- Mindboggle project (fMRI) – ca. 1.5 TB
- Allen Brain Atlas Systems – ca. 70 – 150 TB

PIDs - current status

Handled by PDC

- Path in namespace is registered
- Stored as attribute on the data object
- “Known” but not resolved at hdl.handle.net
- “Known” but not resolved at dx.doi.org

PIDs – possible improvements

- Allow logical path to change
- What about collections?
- Register further metadata (e.g., license, related)
- Provide a descriptive, web-based landing page that resolvers can redirect to
- iget 11140/9f7387e8-81a3-11e2-a643-842b2b12ea0c
  - including specific attribute access?
- Can they be made to look nicer?

Show me the code

- It seems like code is handled internally only
- Why not develop in the open?
- Some components are certainly of general interest
- Could be self-hosted or at Github, Launchpad, Bitbucket, Google Code, SourceForge and the likes.

Documentation

- For end users: video tutorials
  - http://www.youtube.com/user/INCForg
- Design documents
  - http://dev.incf.org/trac/infrastructure/wiki
- For administrators: data&zone servers
  - http://github.com/INCF/ids-tools/wiki
- Background reading: a workshop report
  - http://www.incf.org/programs/workshops/scientific-workshops/ci-1

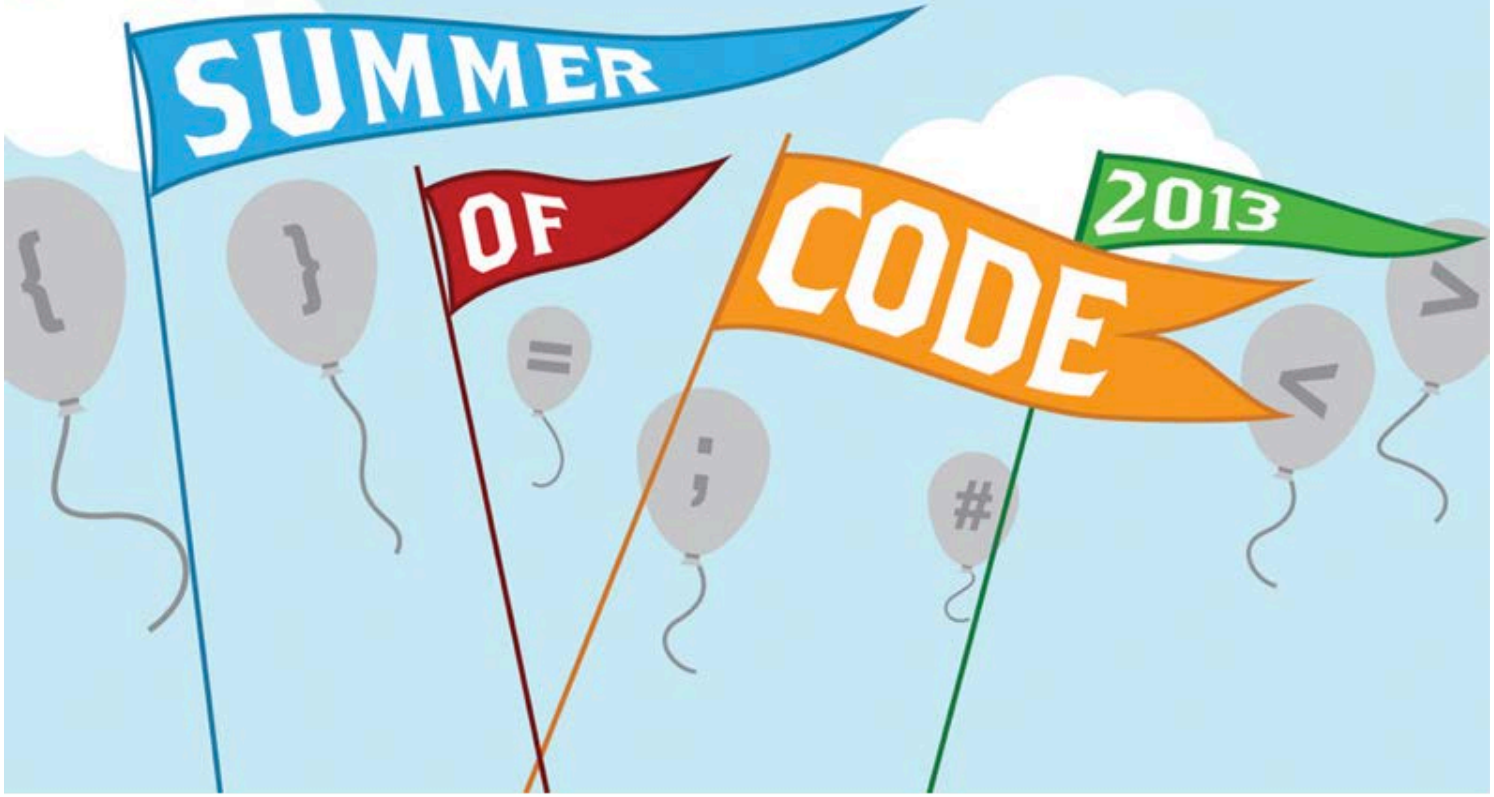
Contributors

- Sean Hill
- Chris Smith
- Sina Khaknezhad
- Ylva Lillberg
- Beatriz Martin
- Mathew Abrams
- EUDAT
  - Jani Heikkinen
  - Johannes Reetz
  - Dejan Vitlacil

incf @ Google

Contact info: gsoc@incf.org

Web:
www.incf.org/gsoc



incf