MRC | Cognition and Brain Sciences Unit

Introduction to CBU computing services

(pdf available at

http://http://imaging.mrc-cbu.cam.ac.uk/methods/MatlabLecturesSchedule)

Russell Thompson

Overview

- Computing resources
- Accessing resources
- Scientific Software
- Best practices

- Your responsibilities
  - Security and usage policy
  - Data protection

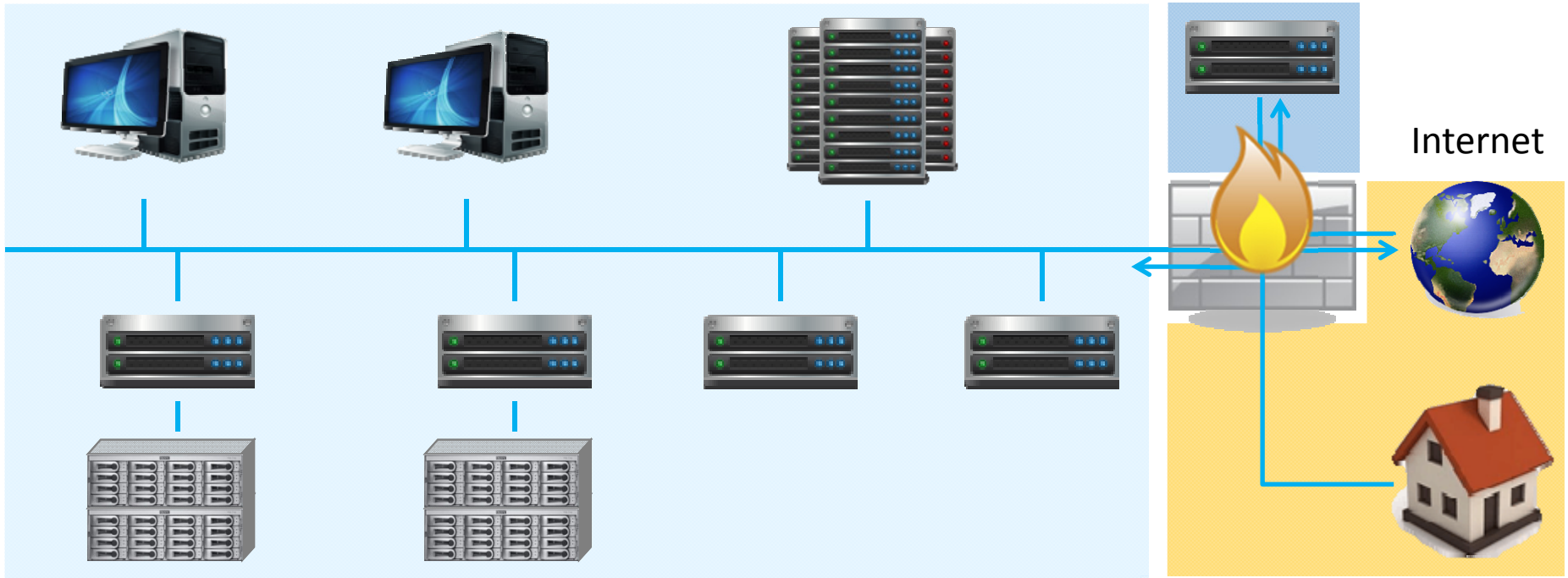
Computing Resources

Desktops
(mainly Windows 7, some Win XP, Linux, Mac OSX)

Testing and shared access machines
(Win 7, some legacy XP)

Central compute machines
(Linux)

DMZ Services
(www, FTP, DNS, SSH gateway, etc)



Imaging, group, & raw data storage

Home space storage

Email
(Microsoft Exchange)

Infrastructure Services
(DNS, DHCP, Authentication, etc)

Remote access

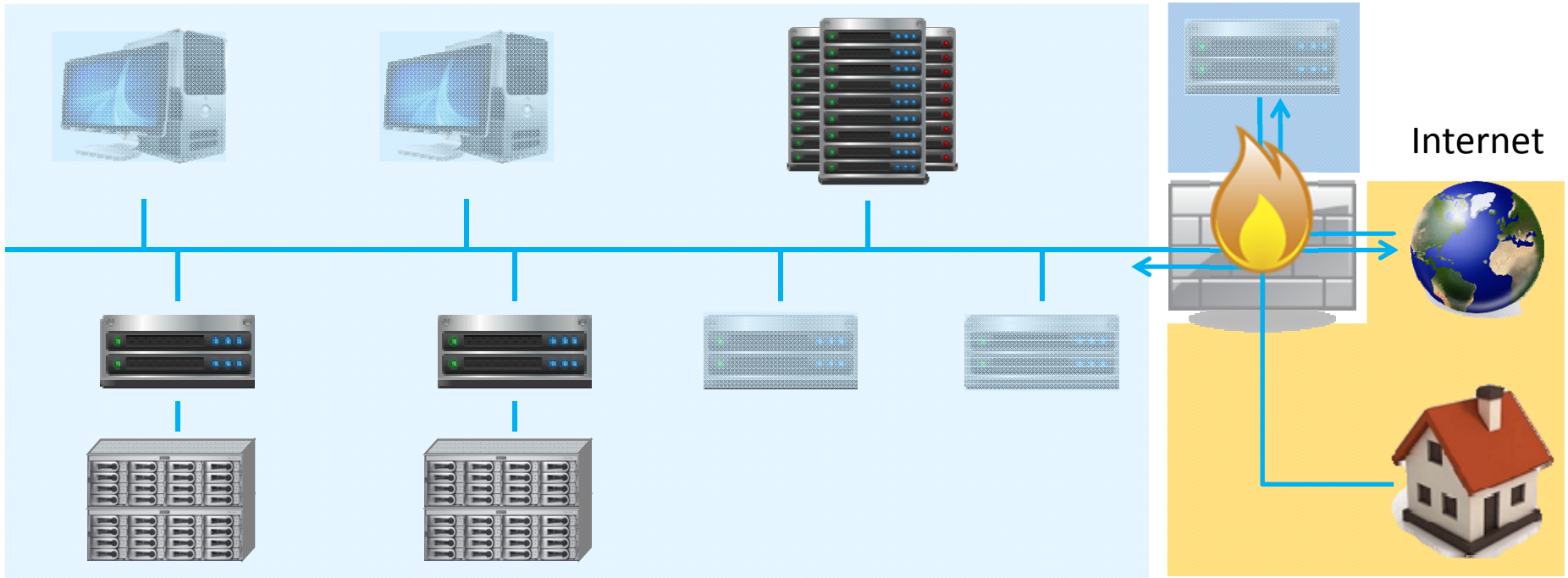
Computing Resources

Desktops
(mainly Windows 7, some Win XP, Linux, Mac OSX)

Testing and shared access machines
(Win 7, some legacy XP)

Central compute machines
(Linux)

DMZ Services
(www, FTP, DNS, SSH gateway, etc)



Imaging, group, & raw data storage

Home space storage

Email
(Microsoft Exchange)

Infrastructure Services
(DNS, DHCP, Authentication, etc)

Remote access

Network Storage

Home space:

- Permanent staff get 50GB quota
- Personal to you, by default not accessible by anybody else
- Snapshot backups – hourly / nightly / weekly
- Replicated hourly to offsite system
- Tape backups retained for 1 year
- Intended to store scripts, figures, documents etc – things that can't be recreated via script.
- Not really intended for large amounts of imaging data

Network Storage

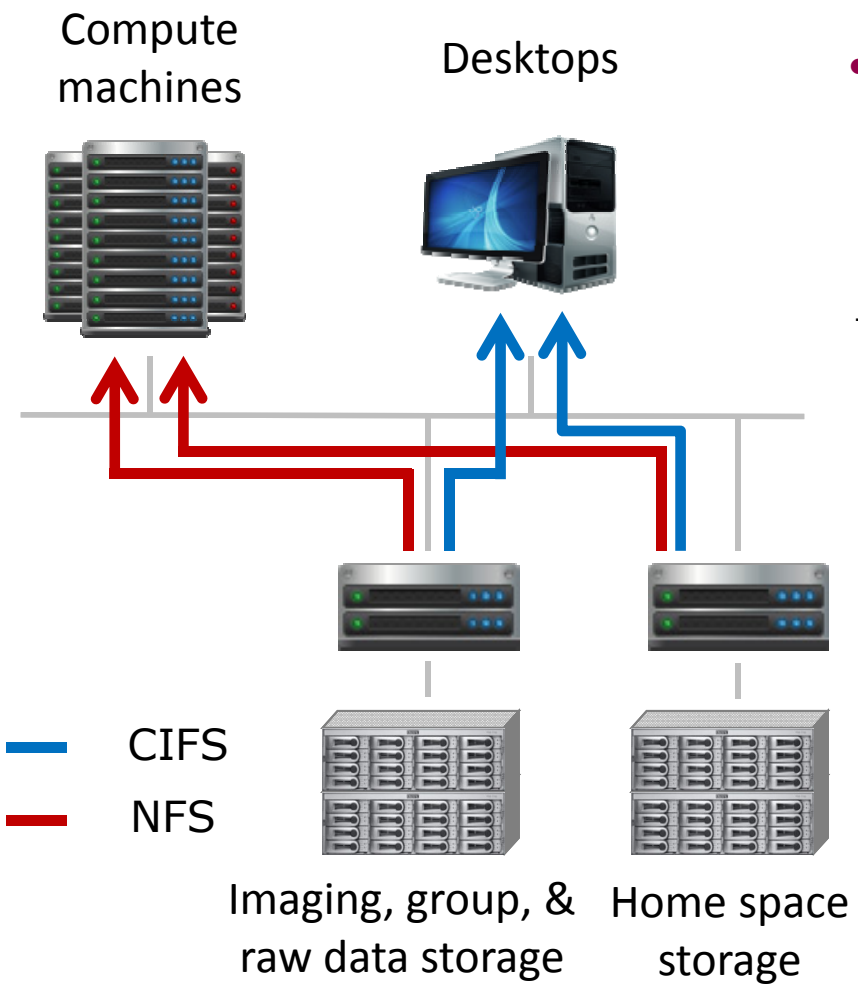
Imaging space:

- No quotas
- Not created by default – available on request for people doing imaging analysis
- Default permissions allow all members of the imagers group to read each others' directories
- 1.5PB Disk based storage (replicated off site)

Shared research group areas:

- No quotas
- Created to allow members of specific labs / research groups to share data
- Access limited to members of the relevant research group

Network storage



- Extra protection (snapshots, replication)
- Accessible from multiple locations (desktop PCs, compute machines, testing machines, etc)

To Access	From: Windows	From: Linux
Home Space	\\home*username* U:\	/home/*username*
Imaging Space	\\cbsu\data\imaging	/imaging/*username*
Group share	\\cbsu\data\group*groupname*	/groups/*groupname*
Raw Data	\\cbsu\data\scandata\mri	/mridata/cbu

Network Storage

Raw data:

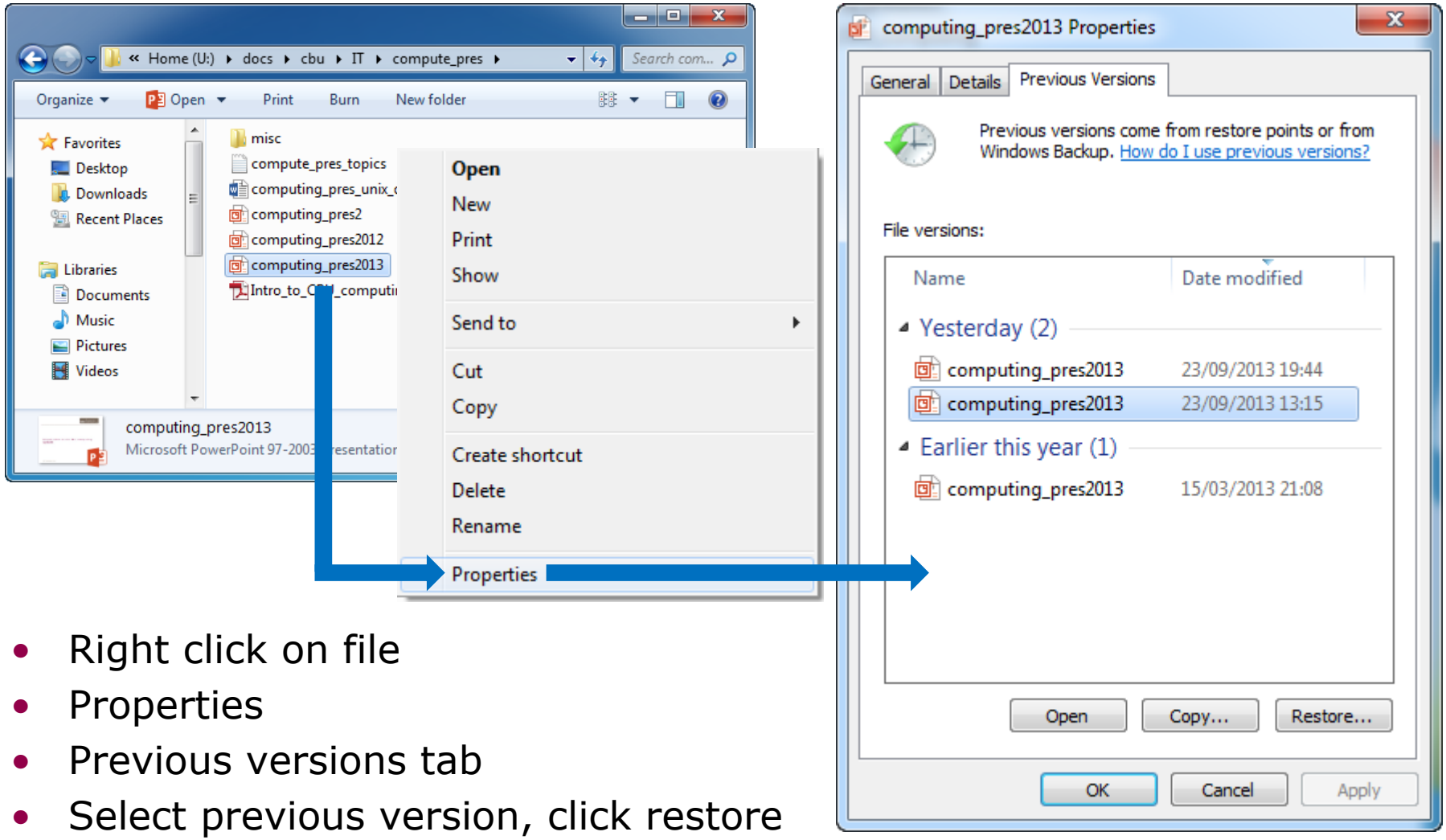
- Windows:
  \\cbsu\data\Scandata\<*MRI/MEG*>\<*institution*>
  e.g. \\cbsu\data\Scandata\CBU\

- Linux:
  /<*mridata/megdata*>/<*institution*>
  e.g. /mridata/cbu

- Typical path for MRI dataset:
  /mridata/cbu/CBU<*participant ID*>_<*project id*>/<*study date*>_<*study time*>/<*scan id*>

  e.g.
  /mridata/cbu/CBU160966_MR16010E/20161017_155824/Series_005_CBU_MPRAGE_32chn

Restoring from a snapshot - Windows



- Right click on file
- Properties
- Previous versions tab
- Select previous version, click restore

Network storage - Best Practice

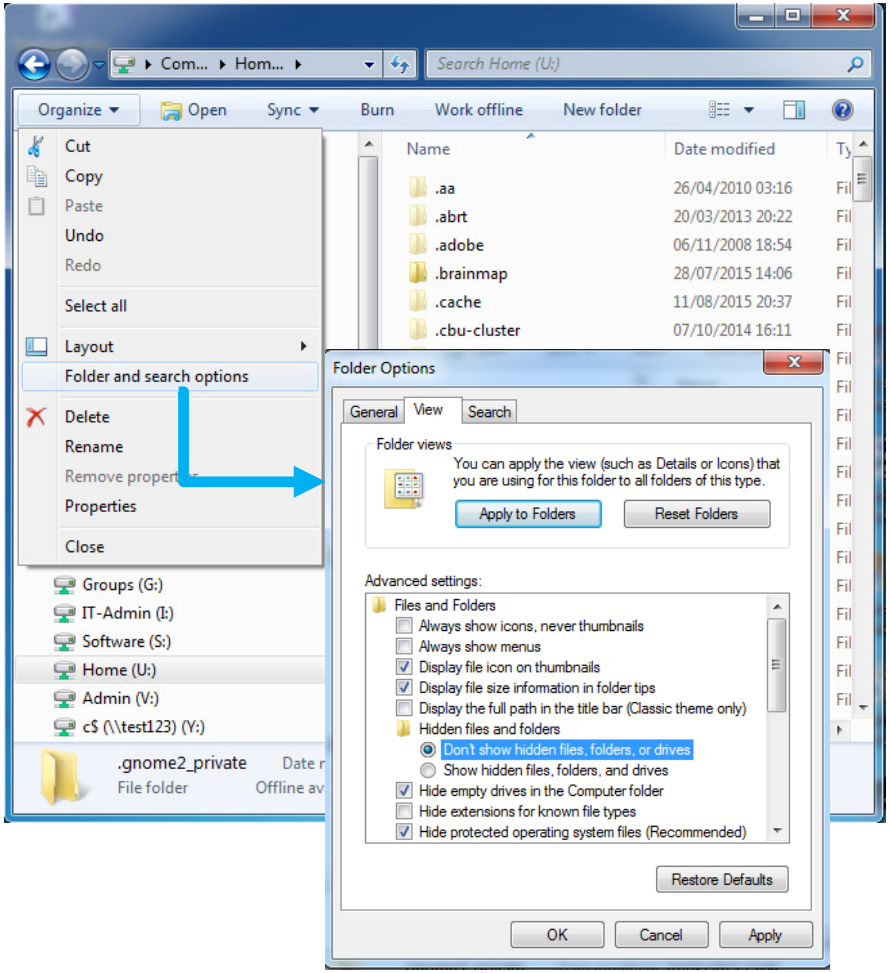
Home space:

- Try to get into the habit of storing documents in your home space.
  - Home space is backed up – desktop hard drives aren't.
  - You can access your home space from almost every machine – you don't have to create multiple copies of data, or move data around on removable media

- Use your home space to store anything you can't easily recreate (documents, figures, scripts). Don't use it for imaging data.

- Data is replicated off-site – in the worst case scenario, analyses could be re-created from raw data and a script stored in your home space

Network storage - Best Practice



- When you browse your home space in Windows, you may see a lot of files whose names start with a "."
- These are used by Linux for storing system settings, preferences, etc – don't be tempted to "tidy" or move them!
- Instead, mark anything you don't want to see as hidden
  - right click, properties, check "hidden"
- Configure windows explorer not to show hidden files
  - click the "Organise" menu in windows explorer, select "Folder and search options", click the "View" tab, select "Don't show hidden files, folders, or drives")

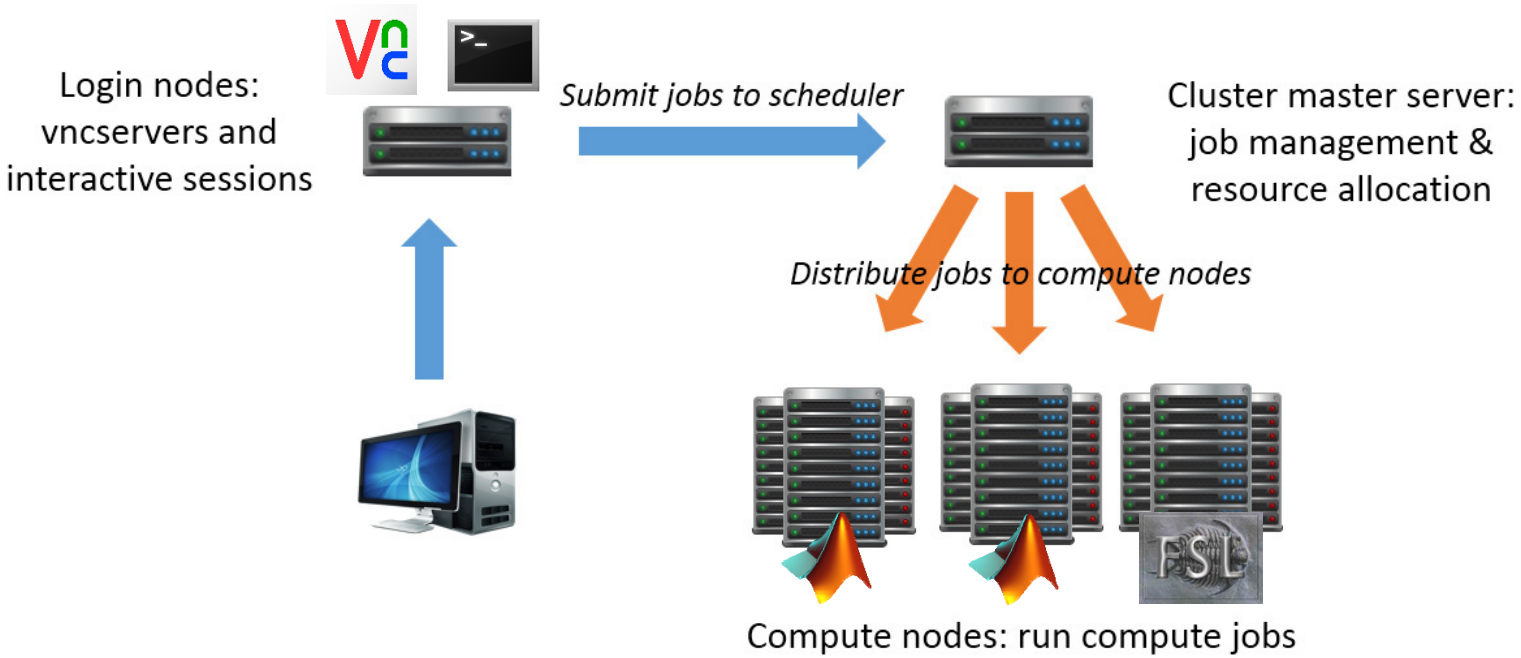
Network storage - Best Practice

Imaging / group storage:

- Unquota'd doesn't mean infinite…
- Clean up after your analyses – e.g. delete intermediate pre-processing images once you've finished with them
- If you are using AA version 4, make sure garbage collection is turned on
- Don't copy raw data from /mridata or /megata into your /imaging directory
- Don't create multiple copies of the same files
- You can read data from other peoples' imaging space – you don't need to copy data from their space to your own

Compute cluster

intranet.mrc-cbu.cam.ac.uk/compting/cluster



- Shared compute resource for intensive data analysis
- 88 machines, 1500 cores, c. 11TB RAM
- Login and run interactive sessions on a login node
- Run large compute jobs on compute nodes
- Submit compute jobs to a scheduling system (Torque / Maui) that manages allocation of compute resources

Login nodes

Name	CPU (MHz)	N Cores	RAM (GB)	Open GL graphics	CPU Architecture
Login11,12,14	2.67	12	48	No	Westmere
Login13	2.67	16	96	No	Sandy Bridge
Login15-login22	2.67	16	128	No	Ivy Bridge
Login23-24	2.40	20	256	No	Haswell
Login25-26	2.40	28	256	No	Broadwell
Login-gpu02-03	2.67	16	192	Yes	Ivy Bridge
Login-gpu04-05	2.00	12	256	Yes	Sandy bridge

- 332 cores @ ~9.5 GB/core
- All run Scientific Linux 6.4 (64 bit)

Compute nodes

Name	CPU (MHz)	N Cores	RAM (GB)	Open GL graphics	CPU Architecture
Node-cc01-04	2.67	16	96	No	Sandy bridge
Node-cc05-07	2.67	16	64	No	Sandy bridge
Node-d02-18	2.67	12	48	No	Westmere
Node-e01-20	2.67	16	96	No	Sandy bridge
Node-f01-08	2.67	16	192	No	Ivy Bridge
Node-g01-g06	2.60	20	256	No	Haswell
Node-h01-h08	2.40	28	192	No	Broadwell
Node-gpu01 – 02	2.67	16	64	Yes	Sandy bridge

- 1140 cores @ ~7 GB/core
- All run Scientific Linux 6.4 (64 bit)

Accessing compute machines

intranet.mrc-cbu.cam.ac.uk/compting/cluster-access

1. Access a **login** node

   Can pick a specific machine (login01, login17, etc), or use the alias "login"

2. Log in using ssh (=Secure SHell)
   - Windows – PuTTY
   - Linux – ssh command

   This provides a text only terminal

3. Create a graphical sessions using VNC (= Virtual Network Computing)

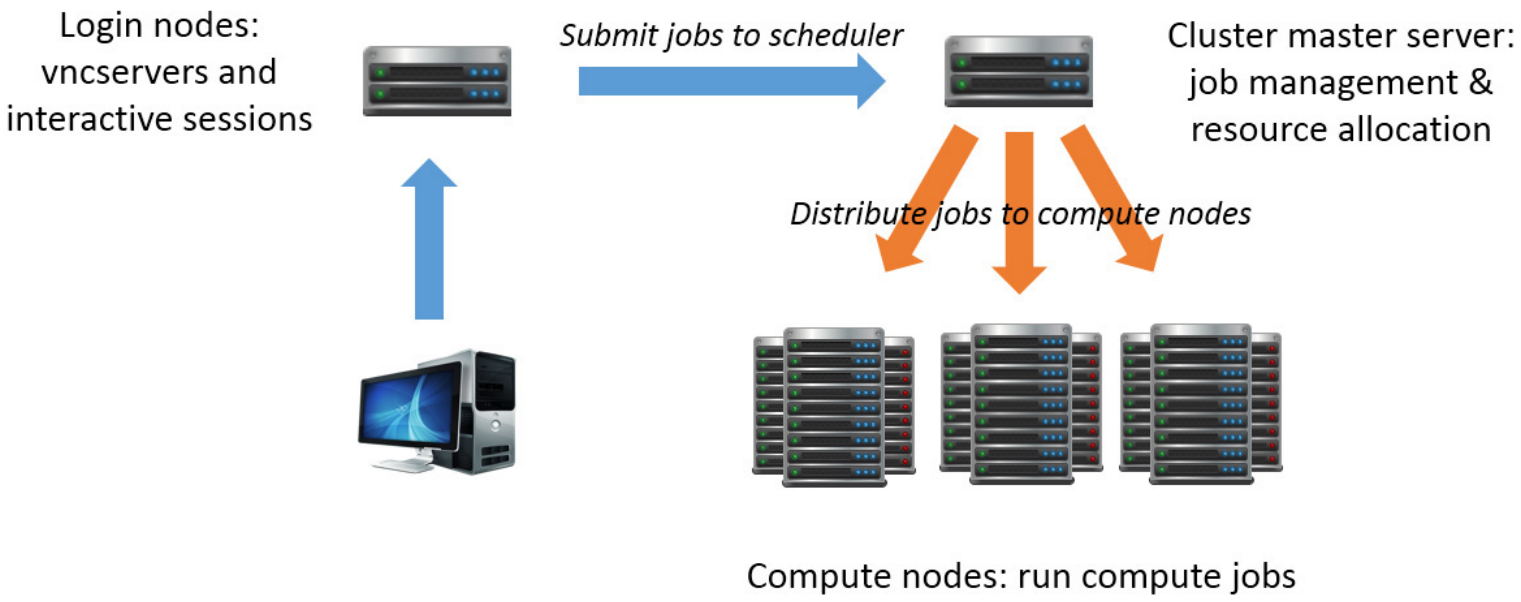
   The *vncserver* command will launch a graphical desktop on the login node

4. Connect to your vnc server using a vnc viewer running on your local machine.

Compute nodes are only accessible via ssh when you have a job running on them

Using the scheduling system

intranet.mrc-cbu.cam.ac.uk/compting/cluster-use



- Log in to a login node and start a vnc server
- Create a batch script to run your analyses
- Test the batch script and determine what resources it needs (esp. memory and CPU time)
- Submit the script to the scheduling system

Best Practice

Login nodes:

- These are a shared resource – think about other users when you're using them
- You should only need one or two vnc sessions at a time – re-use old vnc sessions (they will persist until the host machine is rebooted), or kill vnc sessions if you know you won't need to use them for a while
  - ssh *machine-name*
  - vncserver –kill :*desktop-number*
- Close any interactive SPM/Matlab sessions when you have finished using them, especially if your session has been using a lot of memory.
  - Open matlab sessions use 2 limited resources – memory and matlab licenses
  - If you don't want to close your session, run "clear all" to release memory
- Please don't run large compute jobs or matlabpools on the login nodes!

Best Practice

Scheduling system:

- Develop and debug your scripts on the login nodes before submitting to the scheduler
- Make a note of the resources your job requires – especially memory and cpu time
- Requesting the appropriate resources allows the scheduling system to operate most efficiently. The scheduler will try to launch as many jobs on each machine as possible, without overloading that machine
  - Under-requesting (e.g. requesting 4GB RAM when you need 16GB) can cause the machines to run out of memory and become unresponsive
  - Over-requesting (e.g. requesting 64GB RAM when you only need 16GB) means fewer jobs will run simultaneously
  - Over-requesting also means your job could wait for longer (there are more machines available to handle a 4GB job than a 60GB job, there are more machines with 12 cores than with 16 cores, etc)

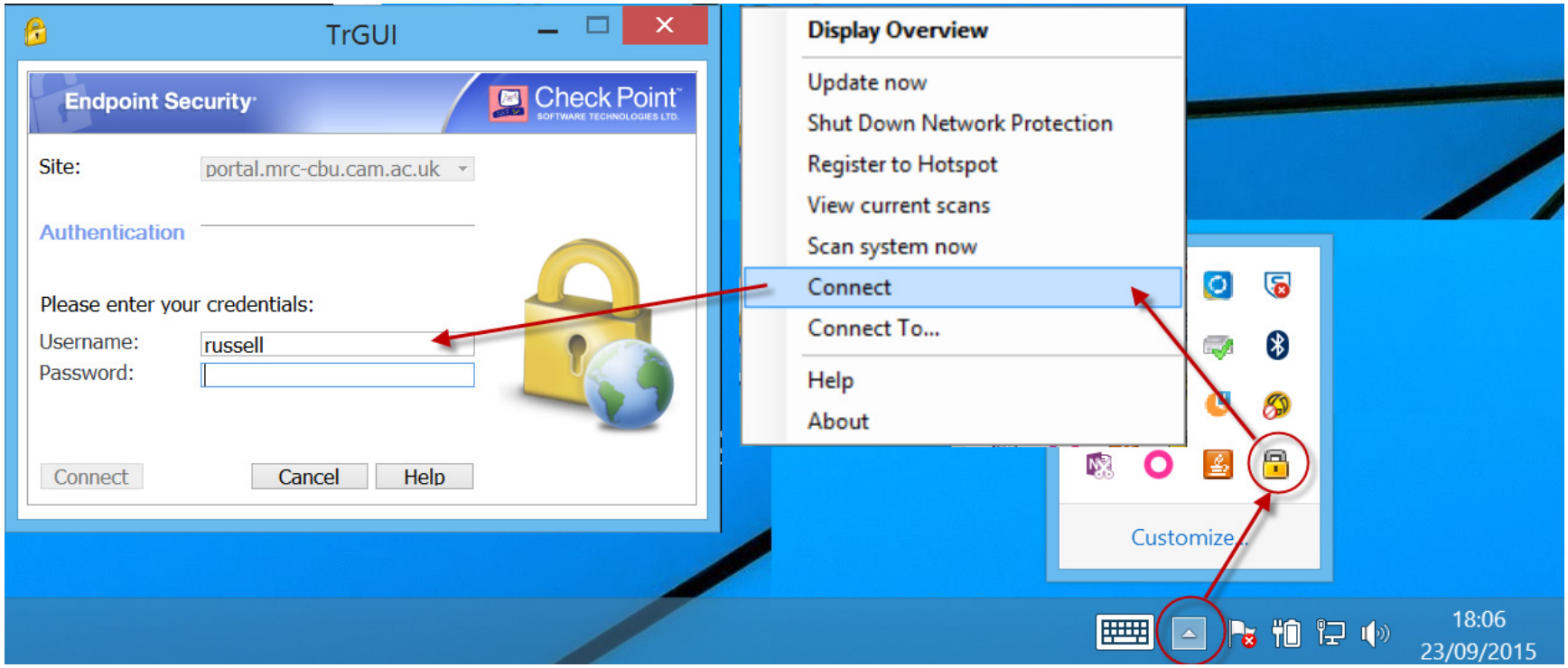
Remote working

- Connect to our network using a VPN (Virtual Private Networking) client:

- Traffic for bound for destinations on the CBU network is encrypted, re-encapsulated and sent over the internet

Remote working

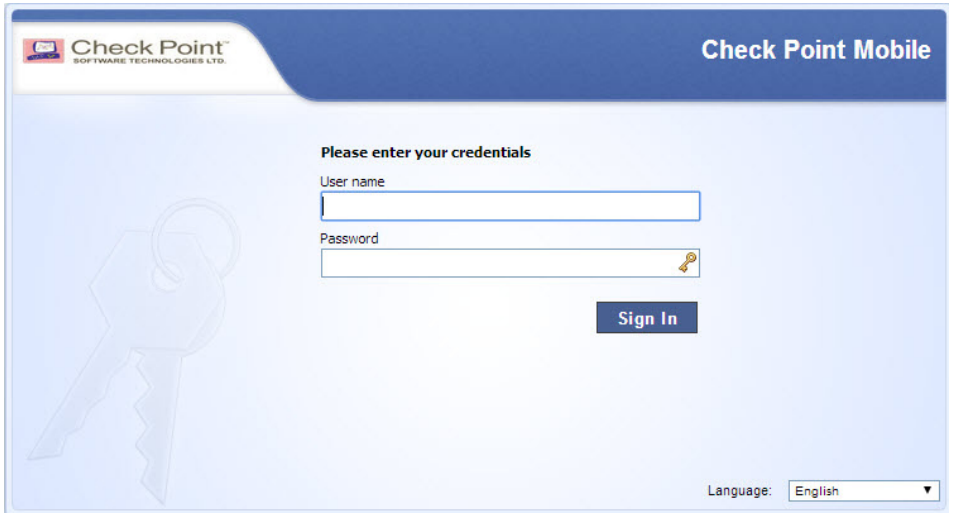
- From a CBU owned machine:

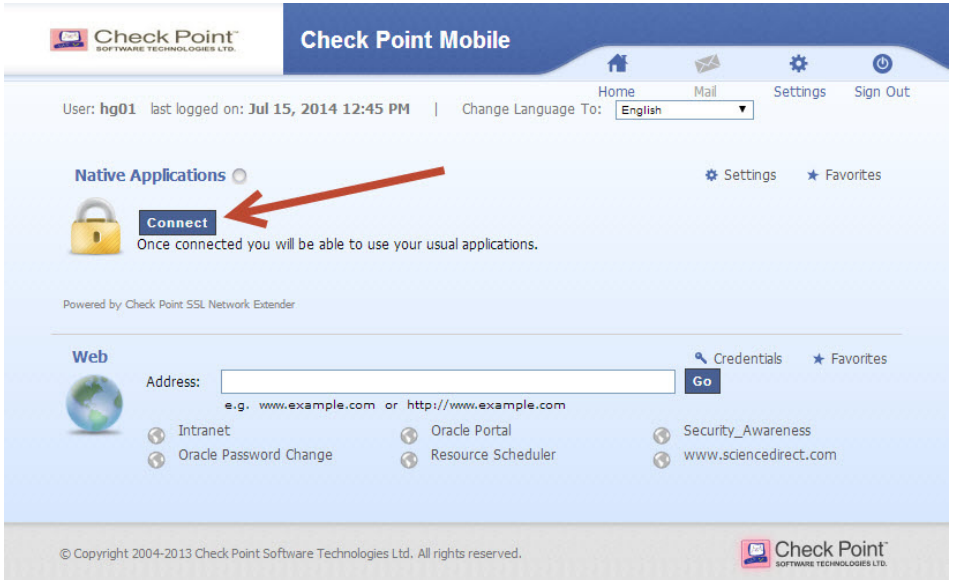


Remote working

intranet.mrc-cbu.cam.ac.uk/computing/Remote-Access/

- From a non-CBU machine:
  - browse to portal.mrc-cbu.cam.ac.uk
  - Sign in using your CBU credentials
  - Click “Connect”





Remote working

http://intranet.mrc-cbu.cam.ac.uk/computing/accessing-resources-remotely/

- Once you are connected and have an IP address on our network, you can access internal resources as if you were using a CBU desktop:
  - Compute cluster
  - Your CBU desktop
  - VNC servers on login nodes
  - Network storage
  - Resource scheduler and intranet
  - Journal articles
- No need to transfer data on removable media / cloud storage – just connect remotely to your CBU PC

Software

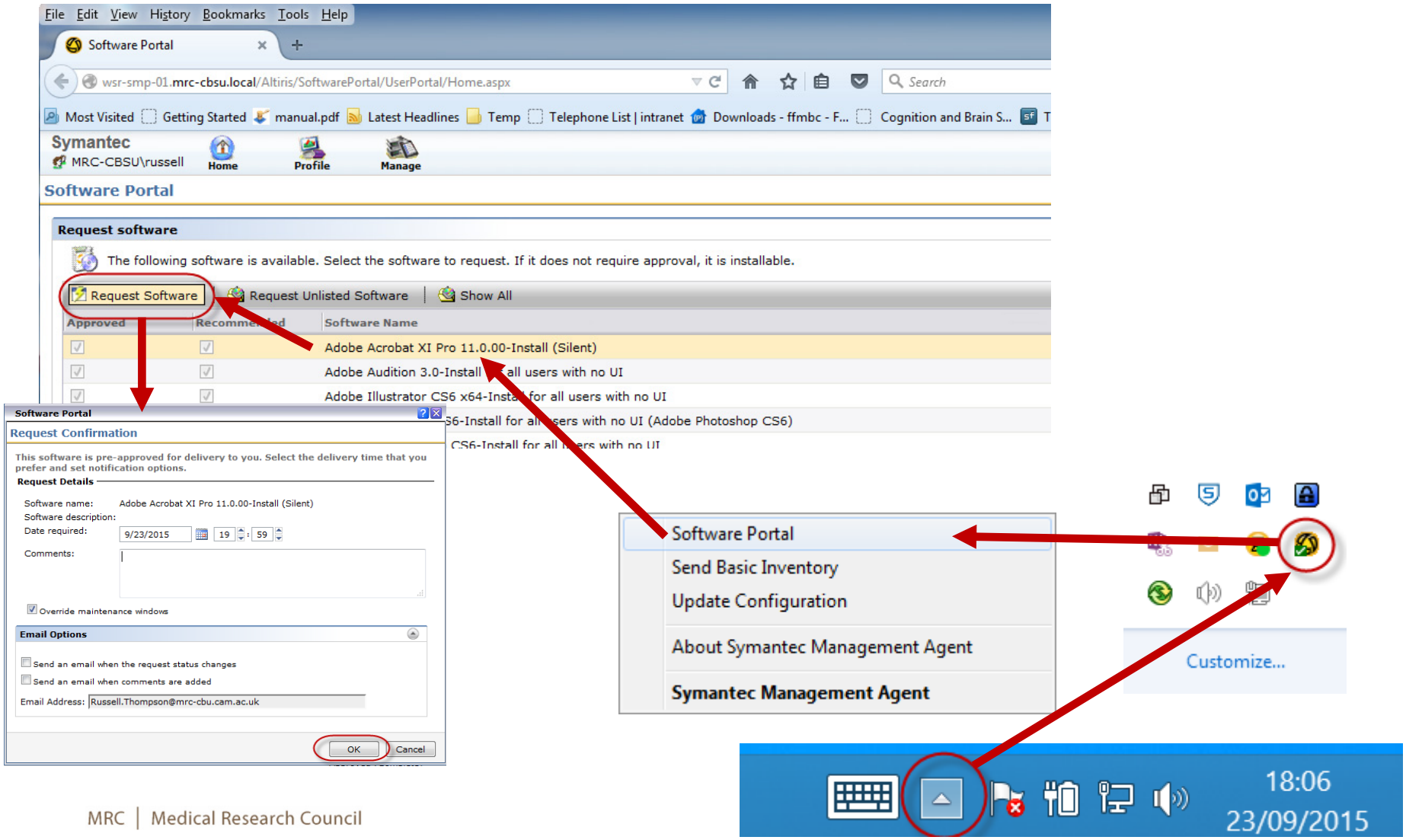
- Desktop PC – many common productivity /stats packages are available:
  - Office, Endnote
  - SPSS, Matlab
  - Adobe Photoshop/Illustrator/Acrobat

- Stimulus delivery software:
  - Eprime, Presentation, Matlab (Psychtoolbox, Cogent)
  - Write your own (Matlab, VB, python)

- Comptue cluster:
  - Matlab, SPM, FSL, Freesurfer, Python (Anaconda, inc Spyder), R/Rstudio
  - /imaging/local/software, or /hpc-software
  - http://imaging.mrc-cbu.cam.ac.uk/imaging/AvailableSoftware

Software Portal

http://wsr-smp-01.mrc-cbsu.local/Altiris/SoftwarePortal/UserPortal/Home.aspx



Scientific software on the compute cluster

- /imaging/local
- Readable by everyone, writeable by members of imagers_devel
- Some very old software in /imaging/local/linux
- Most current packages in /imaging/local/software
- /imaging/local/software/<*package name*>/<*package version*>/<*os arch*>
  e.g. /imaging/local/software/fsl/v5.0.8/x86_64
- FSL, Freesurfer, Python (Anaconda), R/Rstudio

- SPM:
  - Pre SPM 8: /imaging/local/spm
  - SPM 8 and above: /imaging/local/software/spm_cbu_svn

Scientific software on the compute cluster

- /hpc-software
- Readable by everyone, writeable by members of computing group

- Matlab, plus various utility scripts

- Matlab – /hpc-software/matlab/<version>
  e.g. /hpc-software/matlab/r2015a

- Launch matlab by typing matlab_<version>, e.g. matlab_2015a

Security and Usage Policies

- Full policies available on the intranet (http://intranet.mrc-cbu.cam.ac.uk/administration/induction/)
- By signing up for a CBU computing account, you are agreeing to abide by those policies

- **General principles:**
  - Protect other peoples' personal data
  - Protect our machines, data and users
  - Protect the integrity and reputation of the MRC and UoC
  - Avoid participating in, facilitating or encouraging illegal or inappropriate activities

Protect other peoples' personal data

- We have a **legal** obligation to protect the rights and privacy of staff, participants and members of the public, and there are serious consequences for non-compliance
- Data Protection Act (DPA; 1998)
- Designed to "protect the fundamental rights and freedoms of … persons, and in particular their right to privacy with respect to the processing of personal data."
- DPA Covers:

  "Obtaining, recording, holding, organisation, adaptation, alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment, combination, blocking, erasure or destruction"
- i.e. pretty much anything you might want to do with personal data…
- From May 2018 - General Data Protection Regulations (GDPR)

Protect other peoples' personal data

What counts as personal data?

- Data that is about or clearly relates to an identifiable, living individual
- Data that could be used to learn something about an individual
- Context is important – can an individual be identified by taking into account other data held by the same data controller, or other publically available information?
- **Anonymised** or aggregated data is not covered by the Data Protection Act

Protect other peoples' personal data

Defined in DPA:

- Name
- Address
- Email address
- Telephone number
- Official ID numbers (e.g. national insurance number, NHS number, passport number, driving license number, etc)
- Date of birth
- Birthplace
- Genetic information
- Face, fingerprints, or handwriting
- Credit card numbers

DPA also defines "Sensitive personal data"

- Medical information, information about political views, ethnicity, sexual orientation.

GDPR also covers:

- IP address, digital identity, login name, screen name, nickname, or handle

Protect other peoples' personal data

At the core of the act are 8 data protection principles.

Personal data shall be:

1. Processed fairly and lawfully
2. Processed only for specified, lawful and compatible purposes
3. Adequate, relevant and not excessive
4. Accurate and up to date
5. Kept for no longer than necessary
6. Processed in accordance with the rights of data subjects
7. Kept secure
8. Transferred outside the European Economic Area only if there is adequate protection.

Protect other peoples' personal data

Personal data shall be:

1. Processed fairly and lawfully
2. Processed only for specified, lawful and compatible purposes
3. Adequate, relevant and not excessive

- Get explicit consent to store and process participants' data.
- Explain what information you'll collect, what you'll use it for, and who will be able to access it
- Only use data for purposes to which participants have consented.
- Never share participants' personal information with anyone unless you have explicit authorisation to do so.
- Don't share participants contact details with anyone (e.g. sharing volunteer panel lists)

Protect other peoples' personal data

Personal data shall be:

4. Accurate and up to date
5. Kept for no longer than necessary

- Destroy personal data when you have finished with it.
- If you maintain a database of participants, contact them at regular intervals to ask them for updated details, or if they wish to leave.

Protect other peoples' personal data

Personal data shall be:

7. Kept secure
8. Transferred outside the European Economic Area only if there is adequate protection.

- Separate research and personal data.
- Wherever possible, **anonymise** research data completely, or use ID numbers linked to personal data stored elsewhere
- Store electronic personal data (and keys linking ID numbers to personal data) in our secure data area
- Store paper data in a locked drawer or filing cabinet
- Protect personal data with appropriate file permissions.
- Protect against unlawful disclosure of personal information, both accidental and deliberate:
  - Don't transfer personal data using laptops, removable media, email or cloud storage.
  - If you must transfer data this way, make sure it is encrypted.

Security enforcement

- Centrally enforced measures:
  - Traffic to / from external networks passes through our main firewall
    - Block certain ports / services
    - Block cloud storage sites including Dropbox and Google Docs.
    - URL filtering – block inappropriate content
  - Spam filtering of email
  - Logging of network traffic and email source / destination
  - Anti-virus software and local firewalls on all desktop machines
  - Policy to lock PC screen after 10 minutes inactivity
  - Full disk encryption
    - Desktops do single sign on, laptops enforce pre-boot authentication.
  - Removable media encryption
    - Not strictly enforced (do not have to encrypt all removable media)
    - Any media holding personal / personally identifiable data MUST be encrypted.
  - File permissions

Security best practice

- Other things you can do:
  - Lock your screen whenever you leave your desktop PC (windows + L)
  - Be very careful of installing software downloaded from the internet. If in doubt, ask.
  - Be very careful of "drive by installs" – unwanted software bundled with legitimate packages
  - Treat any emails asking you to download content / follow a link with extreme suspicion.
  - Do not transfer MRC data out of our system without the director's explicit permission.
  - Avoid transferring data on removable media. If there's no option, use the media encryption (NB - any media holding personal / personally identifiable data MUST be encrypted)
  - **Anonymise your data**
  - Secure any personal data – encryption, file permissions, locked cabinet, etc
  - Don't use any unblocked cloud storage

Further Information and Support

Computing group intranet page:

http://intranet.mrc-cbu.cam.ac.uk/computing/

Question and answer site:

http://forum.mrc-cbu.cam.ac.uk/qa

Imaging wiki:

http://imaging.mrc-cbu.cam.ac.uk/

Software gurus

http://imaging.mrc-cbu.cam.ac.uk/imaging/AvailableSoftware

IT helpdesk – it-help@mrc-cbu.cam.ac.uk

Computing group in room 58

Further Information and Support – Computing Group

- Try to provide as much diagnostic information as possible – exact circumstances under which an error occurs, what you have tried to fix the problem, error messages, etc.
- If you can't find the answer from one of the sources listed above, come and talk to us in Room 58
- Try to think about what IT resources (hardware, software, services, etc) you'll need as far in advance as possible.
- We'll do our best to help, but notice is appreciated!

Jeff

Howard

Russell

Demo

- http://intranet.mrc-cbu.cam.ac.uk/computing/intro-demo