

Outbreak Data Analysis Platform

This document provides information for users of the ODAP hosted in EPCC at the University of Edinburgh.

| Revision | Author | Changes |
|------------|---------|---|
| 2023-11-03 | abrooks | Use of MFA |
| 2023-08-08 | abrooks | Note the project data should be in shared/workspace |
| 2023-01-30 | abrooks | New contacts, and troubleshooting steps. |
| 2022-09-14 | abrooks | SSH key size 2048 not 1024. |
| 2022-09-13 | abrooks | Installed bgen utilities and made the rbgen package available. |
| 2022-09-12 | abrooks | Changed logout image. |
| 2022-07-07 | abrooks | Installed CASSI from www.staff.ncl.ac.uk/richard.howey/cassi/ |
| 2022-04-19 | abrooks | Added email addresses |
| 2022-04-02 | abrooks | Renamed to ODAP |
| 2022-03-31 | abrooks | Added section on available software |
| 2022-03-24 | abrooks | Added link to ultra2 documentation. |
| 2022-03-16 | abrooks | Added troubleshooting for 403 Forbidden error. |
| 2022-03-09 | abrooks | Added an alternative way to start the remote R server. |
| 2022-02-02 | abrooks | Additional troubleshooting. |
| 2022-01-20 | abrooks | Added more details about using Ultra2 Terminal. |
| 2022-01-14 | abrooks | Added more screenshots. |
| 2021-12-08 | abrooks | Tidied up the sections on R. |

Contents

| | |
|--|----|
| Introduction | 3 |
| Terminology | 3 |
| TL;DR | 3 |
| Contacts | 3 |
| Procedure for Gaining Access | 4 |
| Logging Into the ODAP TRE at EPCC..... | 6 |
| Summary | 9 |
| Troubleshooting..... | 10 |
| Help using SAFE..... | 10 |
| Cannot connect to website | 10 |
| Cannot login | 10 |
| Virtual desktop problems..... | 10 |
| How to use the c19-desktop | 12 |
| Logging in | 12 |
| Logging out..... | 12 |
| Using desktop software | 12 |
| How to use Ultra | 14 |
| What you need to know..... | 14 |
| Directories..... | 14 |
| How to import and export data | 14 |
| Available Software | 16 |
| Using Ultra2 for complex analysis..... | 16 |
| Using Anaconda for R and Python | 16 |
| Using R Studio | 17 |

| | |
|--|----|
| Troubleshooting R..... | 18 |
| Using PyCharm on Ultra..... | 19 |
| Running time-consuming jobs on Ultra2 with RemoteR | 21 |
| Running large jobs on Ultra2 with Slurm | 21 |
| Accessing Ultra 2 from the desktop VM | 21 |
| Access to external databases from Ultra | 23 |

Introduction

The data analysis platform consists of several components:

- A database in the National Safe Haven where it is safe to store personally identifiable health data.
- A database and file storage outside the National Safe Haven for the storage of data which is not personally identifiable
- Processing systems which can operate safely on the personally identifiable data within the Safe Haven to link with other datasets, produce aggregated reports or to de-identify the data for further use.
- Access to desktops for approved researchers to work on the de-identified data.
- Access to High-Performance Computing (HPC) systems, Ultra2 and Eddie, for working with large datasets or the data which are not personally identifiable
- Access to desktops for deploying a web application for reporting

This guide describes how to use these systems, but it does not describe the data itself.

Terminology

- ODAP – Outbreak Data Analysis Platform, encompasses all of the above for the purposes of processing datasets such as ISARIC
- TRE – Trusted Research Environment, another name for a Safe Haven
- NSH – the National Safe Haven in Scotland, a TRE run by EPCC for eDRIS/Public Health Scotland
- FCS – Flexible Compute Space, the systems which lie outside the National Safe Haven making up the ODAP TRE
- PDA – Protected Data Access environment, the technical name for the how the FCS was built in EPCC
- Ultra2, SDF-CS1 – both names refer to the High Performance Computer accessed from the ODAP TRE
- EIDF – Edinburgh International Data Facility, the organization within EPCC which looks after the HPC and other systems
- MFA – Multi-factor authentication, meaning the use of text messages or apps to provide a numeric code which is used alongside a password when logging in.

TL;DR

Quick summary:

- Register for an account in SAFE, then apply within SAFE to join project u036 (Ultra PDA).
- Wait for your application to be approved and for your VDI account credentials to be sent to you.
- Login to the Ultra VDI service <https://eidf.epcc.ed.ac.uk/eidf01/> using the VDI credentials
- Select the c19-desktop (SSH) option, login using the u036 account, change your password, logout.
- Select the c19-desktop (RDP) option and login using the u036 account with new password.
- Inside this desktop you can SSH to ultra2, and you can use RStudio and PyCharm IDEs.
- Follow the guide to use Anaconda, and to use RStudio or PyCharm in “remote” mode.

Contacts

The helpdesk can be contacted at

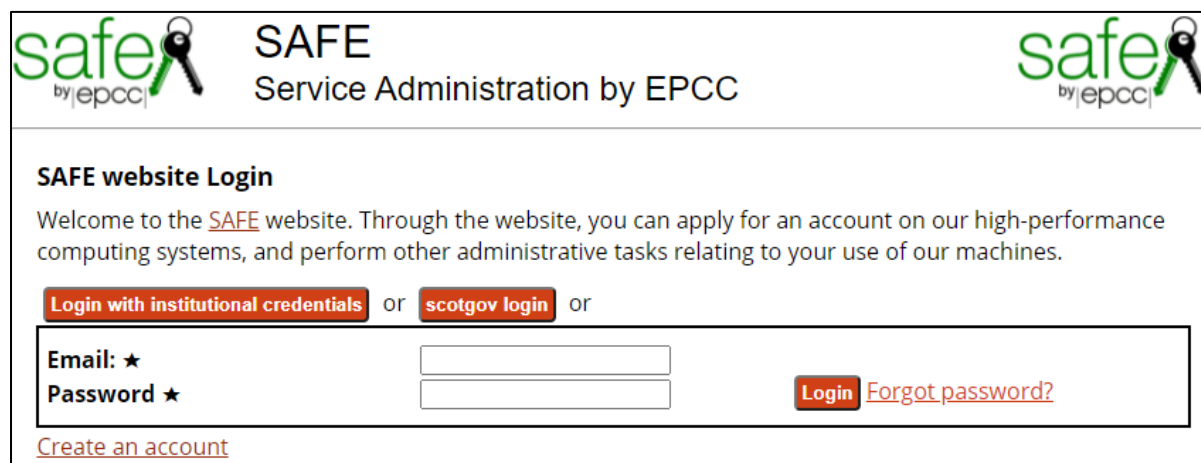
odap@ed.ac.uk

They can route your question to an appropriate Information Governance officer, a Data Manager, or a Technical Support agent.

Procedure for Gaining Access

Potential users first need to register in the EPCC “SAFE” which is a user registration and account management system.

<https://safe.epcc.ed.ac.uk/>

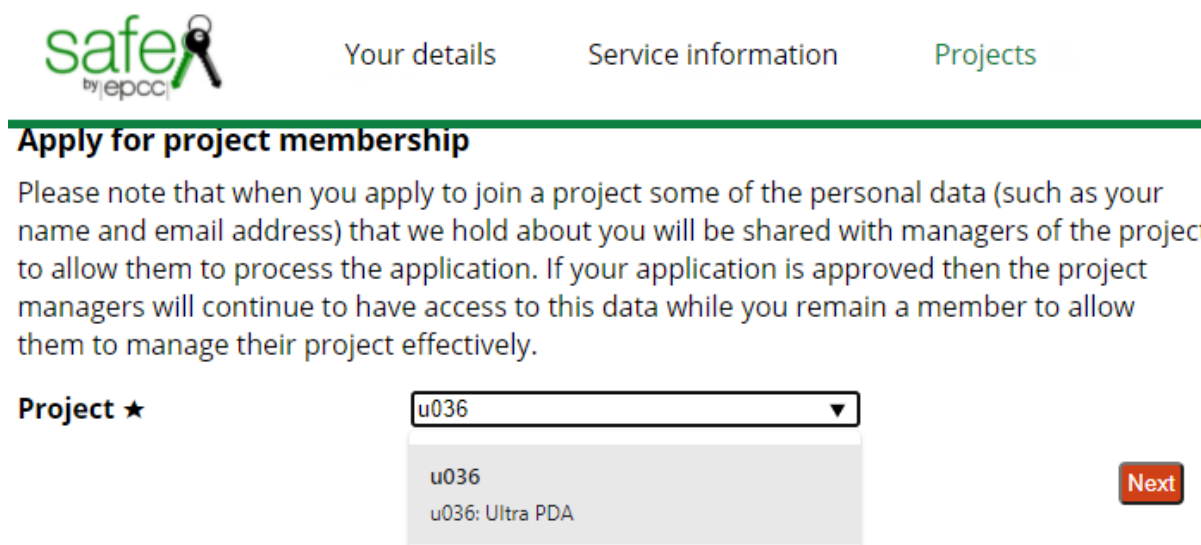


The image shows the SAFE website login page. At the top, there is a header with the 'safe by epcc' logo on the left and right, and the text 'SAFE Service Administration by EPCC' in the center. Below the header, the page is titled 'SAFE website Login'. A welcome message states: 'Welcome to the SAFE website. Through the website, you can apply for an account on our high-performance computing systems, and perform other administrative tasks relating to your use of our machines.' There are three login options: 'Login with institutional credentials', 'or', 'scotgov login', and 'or'. Below these options is a form with two input fields: 'Email: ★' and 'Password ★'. To the right of the password field is a 'Login' button and a link 'Forgot password?'. At the bottom left of the form is a link 'Create an account'.

Click on the link to Create an account. Once your account has been created you can Login.

(You can use your University of Edinburgh credentials (via EASE) to login, but only after you have created a SAFE account and registered your EASE credentials within SAFE).

Use the Projects menu to Request access:



The image shows the 'Projects' page of the SAFE website. At the top, there is a navigation bar with the 'safe by epcc' logo on the left and three menu items: 'Your details', 'Service information', and 'Projects'. Below the navigation bar, the page is titled 'Apply for project membership'. A paragraph of text explains that personal data will be shared with project managers if the application is approved. Below the text is a form with a label 'Project ★' and a dropdown menu. The dropdown menu is currently showing 'u036'. Below the dropdown menu, there is a list of options: 'u036' and 'u036: Ultra PDA'. To the right of the list is a 'Next' button.

Type the project code u036 which is a PDA (Protected Data Access) account on Ultra2.

Your project membership request will be sent to a Project Manager for review. The project manager may need to check with an approvals board so access may not be granted immediately.

The next step is to apply for a machine account. The SAFE system has only one option at this point, which is labelled “sdf-cs1”. If you see other options then please choose the “sdf-cs1” only.

New account policies

If a check-box does not appear beside a machine then the project you selected is allowed to use the machine but one of the policies that apply to the machine is preventing you from applying.

A cross will be marked against the policy that is preventing you from applying.

You would also be able to enable access to this machine by updating your account to meet any policy marked with an arrow.

Select a machine for the login account

| Select | Machine | Type | Description | Policies |
|----------------------------------|---------|------|-------------|--|
| <input checked="" type="radio"/> | sdf-cs1 | | | Users must have a public key registered to use the machine ✓ |

Next

When you click Next you can choose an account username. This is restricted to 8 letters. Please choose a username in the format: first initial plus surname, eg. "jsmith", if possible. The username must be unique across other machines in the SAFE so you may want to append some code or letter to indicate this is your ISARIC account.

SAFE Login account Request

This form is for requesting new login accounts. To request additional access for an existing account, select it from the navigation menu at the top of the page

Your username will be visible to other users on the system

This machine support ssh key authentication. You can upload a public key to use here.

A SSH public key is required to use this machine.

Requested username ★

SSH public key ★

Choose file No file chosen

Request

The system requires a SSH public key be supplied. This will not be used but unfortunately is a requirement that we cannot change, so at this stage it does not matter what you supply, as long as it looks like a valid key. A key can be generated on the website: <https://8gwifi.org/sshfunctions.jsp> Tick Algorithm:RSA and Size:2048, Click **Generate SSH Keys** and then copy and paste the **Public Key** text into the SAFE SSH public key field. You can save the Private and Public keys to files if you wish. If you get the error "Corrupt key" then check you are pasting a single line of text which begins with ssh-rsa.

Generate SSH Keys Online

SSH-Keygen Online Algorithm ☒RSA ☐DSA ☐ECDSA

RSA Key Size ☐1024 ☐2048 ☐4096

Passphrase

for generating encrypted keys

Generate-SSH-Keys

Email-SSH-Keys

SSH-keygen for [RSA-2048] (PrivateKey/PublicKey)

FingerPrint [6c:79:e3:a6:5a:3cdc:29:eb:22:8a:b9:14:be:59:d6]

Private Key

```
-----BEGIN RSA PRIVATE KEY-----
MIIEowIBAAKCAQEAqZYD+bItYqICE9LRSTV08JmEoz5LzRdvO9t6B3pXwB7o7L2
CVfAKqNbhYwzO6qRJWMLINfL79QQUIcePdnDG+TMew/S92jyDvNUEduOKxsvcta
wmyDvVsMirbB6bXFCn6S7e18Y/UyrdwJRnvS2UuU8My/GIPEMjImYjT6dSpblon
1eSsxI2iqMzZfdVPn7bJkUuGd3Ficvta:rcy/1pUCgYAepEc5MCpZ/LuUu7pupivK-r-0^
vd1TdDCEsc/97wMroSv0rLi3Jv7iCiAVLDuDRy3kCTWtsSBJ-448V5WJ3fLmO
```

Public Key

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAQ=CplgP5sm1gmogIT0tFJNXtWmYSjPkvNF28723pDelfAHujs
vVJV8Aqo1uHLDm7qpEIYwsg18vv1BBQhx492cMb5MwTD9L3aPIPBu1QTS44rGy9y1rCbIO9WwyKtsH
ptcUKfPlt7Xg9TKI3AIGe9LZS5TwzL8aI8QyNSZIKNPP1KluWfV5KzGlaKozNi91U+juFSRQY/YUjg4TG/
xqF2BMVf6RK3td/R1EuXZdNIZpYzRxDQ1PcwCD/DC7AcqV5QWNEWk4iY65+dmnx7QhZ+b/lfbw5v
FJR33HPRTBcJXm8jTc28Yv5fZ6KuM+Xarz13QCbsMsJuyZP1BZK79
```

Copy the
Public Key

If the system requires MFA (multi-factor authentication) please see the documents:

<https://epcced.github.io/safe-docs/safe-for-users/#how-to-turn-on-2-factor-authentication-on-your-machine-account>

Once your application has been approved you should login to SAFE and use the option to view your password from the Login Accounts menu. The machine name is “sdf-cs1” but the account may be listed as “username@eidf”. Note: Ignore the @eidf part when asked to enter your username. This is a one-time password; you will be required to change it when you first login.

Your machine account will give you a login to two computers, the “sdf-cs1” (which we will call “Ultra2” from now on) and a Linux desktop inside the ISARIC system. However, the only access to these systems, for security reasons, is via a *virtual* desktop. Access to the virtual desktop is through a VDI (Virtual Desktop Infrastructure)¹. Again, for security reasons, the VDI requires a separate username and password, and these will be sent to you by email.

From now on you only need to login to the VDI, not into SAFE, to access ISARIC.

Logging Into the ODAP TRE at EPCC

The Virtual Desktop Interface gives access to a virtual Linux desktop inside the secure archive area.

<https://eidf.epcc.ed.ac.uk/eidf01/>

If you get an error trying to access this page please see the Troubleshooting section.



The screenshot shows a web browser window with the address bar displaying "eidf.epcc.ed.ac.uk/eidf01/#/". The main content area features the EIDF logo, which consists of the letters "EIDF" in a bold, sans-serif font, with "Edinburgh International Data Facility" written in a smaller font below it. To the right of the text is a colorful, abstract graphic resembling a stylized 'C' or a series of overlapping arcs. Below the logo, the text "EIDF REMOTE SERVICE" is centered. Underneath this, there are two input fields: the first is for the username, with a cursor visible inside, and the second is for the password, with the placeholder text "Password" visible. At the bottom of the form is a dark grey button with the word "Login" in white text.

Use your VDI account username and password to login here. The VDI account is not the same as your SAFE account, and is not the same as the “sdf-cs1” machine account you requested within SAFE.

Please change your password by clicking your name in the top right, click Settings, and change your VDI password from the Preferences tab.

¹ Some people refer to this as *guacamole* because that is the name of the software which implements the VDI.

SETTINGS

Active Sessions

Preferences

Options below are related to the locale of the user and will impact how various parts of

Display language:

Timezone:

Europe

London

CHANGE PASSWORD

If you wish to change your password, enter your current password and the desired new password below, and click "Update Password". The change will take effect immediately.

Current Password:

New Password:

abrooks

Andrew Brooks

Home

Settings

Logout

The VDI home page will give a list of machines you can log into:

←

→

↺

🏠

secure.epcc.ed.ac.uk/eidf01/

RECENT CONNECTIONS

abr

ALL CONNECTIONS

Filter

C19

c19-desktop01 RDP

c19-desktop01 SSH

sdf-cs1 SSH

IMPORTANT NOTE: Please click on the “c19-desktop SSH” session first and login. This is the sdf-cs1 machine account you created within SAFE and the password which can be found in the accounts section of SAFE. The username might be listed as xxx@edif but you only type the xxx part (ignore the @eidf). Please enter the password carefully without using the Caps Lock key. You will be prompted to change your password. This procedure must be done in the SSH session as this will set your password and create your home directory. You can find your password in SAFE here:

safe

Your details

Service information

Projects

eCSE Funding

Login accounts

Request login account

abricadm@eidf

Use the button to view the initial password:

sshkey

Active

Add Credential

Delete Credential

Request Password Reset

View Login Account Password

Set Login Only

When you first login you will be prompted to choose a new password:

```
← → ↻ 🏠 secure.epcc.ed.ac.uk/eidf01/#/client/ODkAYwBteXNxbA==

Password: *****
Password expired. Change your password now.
Creating directory '/home/v004/v004/abricusr'.
Welcome to Ubuntu 20.04.2 LTS (GNU/Linux 5.4.0-80-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

System information as of Mon 23 Aug 21:13:04 BST 2021

System load:  0.08          Processes:           297
Usage of /:   58.9% of 31.51GB Users logged in:       0
Memory usage: 24%          IPv4 address for docker0: 172.17.0.1
Swap usage:   0%           IPv4 address for ens3:  192.168.133.84

WARNING: Your password has expired.
You must change your password now and login again!
Current Password: █
```

USEFUL TIP: Press the Shift + Ctrl + Alt keys together to get the settings menu. You can paste your password from SAFE into here to avoid mistakes typing it. You can also change the colour scheme if you find white-on-black difficult to read. Press the three keys together again to hide the menu.

c19-desktop01 SSH

abooks

Clipboard

Text copied/cut within Guacamole will appear here. Changes to the text below will affect the remote clipboard.

Press Ctrl-Alt-Shift to see this window.
Text from your clipboard appears here.
Then you can paste it into the virtual terminal with a mouse right-click.

Display

Color scheme:

Font name:

monospace

Font size:

12

8-generic x86_64)

om

MT 2021

308

in: 1

for ens3: 192.168.134.34

for ens9: 10.23.13.11

0.68

anaconda3/bin/activate

After logging in you will be prompted to choose a new password.

Type exit to close this session. Return to the VDI session page and select the RDP (Remote Desktop) option “c19-desktop RDP”. This will present a login screen to the Linux desktop. Again, use your “sdf-cs1” machine username and the password you have just chosen.

Summary

You will have three accounts:

1. Your SAFE website login (only needed during account creation)
2. Your VDI website login
3. Your machine account login (for the desktop and for the sdf-cs1/ultra2 computer)

These actions only need to be completed once:

1. Create an account in SAFE
2. Join the ODAP project u036 and create a machine account
3. Await approval and your VDI account credentials
4. Log into the VDI eidf01 using your VDI account
5. Change your VDI password
6. Choose the SSH session option and login with your machine account
7. Change the password for your machine account
8. Logout

These actions need to be done every time:

1. Log into the VDI eidf01 using your VDI account
2. Choose the RDP session option
3. Log into the desktop using your machine account

Troubleshooting

Help using SAFE

Please see the documents <https://epcced.github.io/safe-docs/> and contact the helpdesk if you have any questions.

Cannot connect to website

If you cannot contact the SAFE website or the VDI website then please try connecting to your institution's VPN.

Please try using the Chrome web browser.

If you get the error "403 Forbidden" then please try again using your institution's VPN. If you are using a VPN and still get this error then please contact the EPCC helpdesk and supply this information:

a) **before** connecting to the VPN

<https://ipv4.google.com/search?q=my+ip+address> and <https://www.whatismyip.com/>
(try both, in case the one listed by Google is different from whatismyip.com)

b) **after** connecting to the VPN

the IP address given by the same two websites

If you are interested, use this website to check where the IP address is located:

<https://www.maxmind.com/en/geoip-demo>

If you see something like Virgin Media or Vodafone (or your home ISP, if working from home) after connecting to the VPN then the VPN is not working. You would need to contact your local IT helpdesk about that.

Cannot login

If you are not sure about your username and/or password:

- SAFE website – use the Forgot Password? button on the SAFE website. If you have problems with this please contact the helpdesk, contact details on the SAFE website.
- VDI website – please contact the helpdesk and ask for your ticket to be assigned to Andrew Brooks.
- c19-isaric desktop – First try to login using the SSH option and see if your password has expired. If you cannot login there then use the password reset procedure provided on the SAFE website. You will need to request a password reset for the specific machine account, in this case on the "sdf-cs1" as part of the "u036" project. After a reset you will be able to log into SAFE and view the new password by selecting your username@eidf from the Login Accounts menu and clicking the View Login Account Password button. If you have problems with this please contact the helpdesk and ask for your ticket to be assigned to Andrew Brooks.

Virtual desktop problems

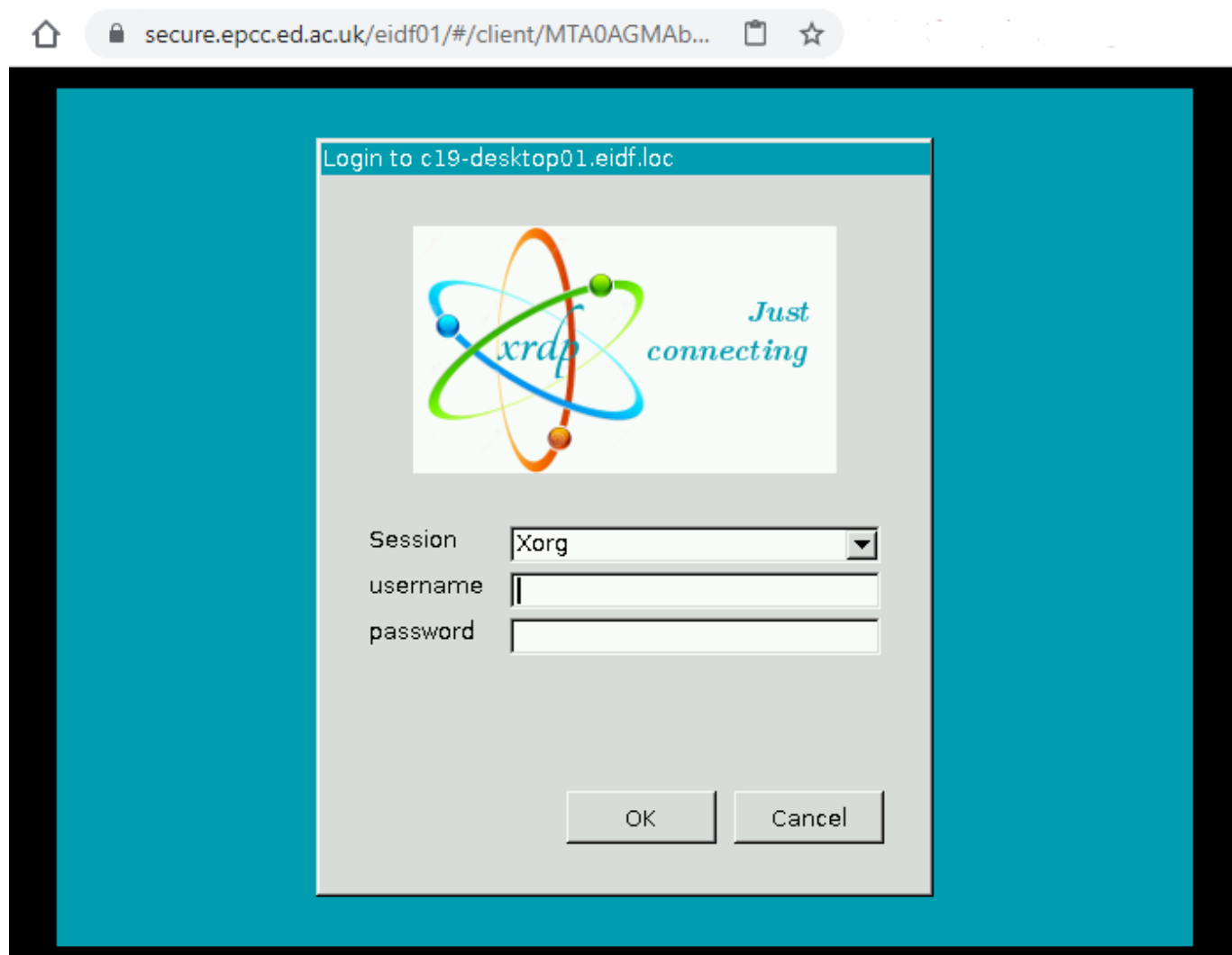
- **Color scheme**; cannot read text in the SSH window – Press the Shift-Ctrl-Alt keys together to get the Guacamole settings and scroll down to change the colour scheme. Press the same keys again to hide the settings.
- The **CAPS-LOCK** key seems to be stuck on. Even if you press it again, the CAPS state remains on. Press the Shift-Ctrl-Alt keys together to get the Guacamole settings and press CAPS LOCK. Press the same 3 keys again to hide the settings. Now CAPS LOCK is off in the virtual desktop and you can press CAPS LOCK again to turn it off on your local desktop.
- Unstable network connection: some users have reported better performance using the Chrome web browser instead of Firefox or Safari. You could also try different VPN settings, for example the Fortinet VPN in SSL mode. If your network connection drops then it is possible to log back into the desktop and continue where you left off. However, do not be tempted to rely on this and leave programs running overnight, as there are various reasons why you might come back and find the desktop has been restarted. (Technically, it might still be running but you can no longer access it). Please save your work and log off before disconnecting whenever possible.

How to use the c19-desktop

Logging in

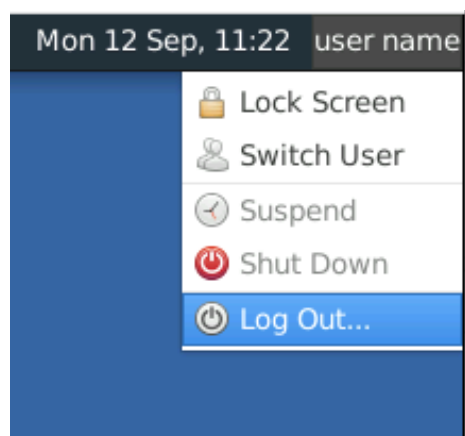
Login to the VDI at <https://secure.epcc.ed.ac.uk/eidf01/> using your VDI account credentials.

Select the c19-desktop (RDP) option and login using your sdf-cs1 (account@eidf) credentials.



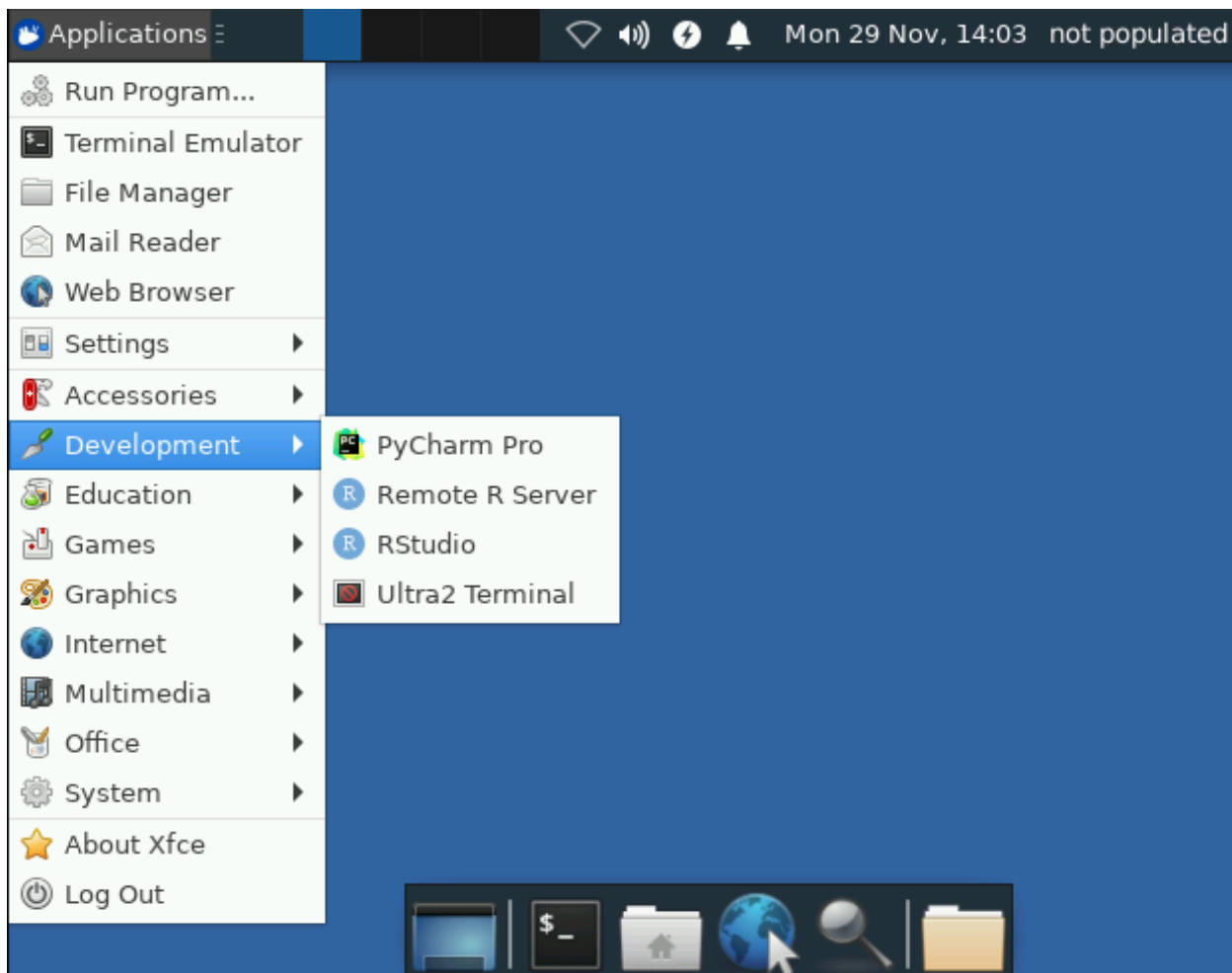
Logging out

To logout from the desktop use the menu at the top right of the screen. Click where it says “user name”.



Using desktop software

You can access RStudio and PyCharmPro from the Applications | Development menu:



The Development menu has options for:

- PyCharm Pro – a Python development environment. You need to bring your own license for this but it is free for academics.
- RStudio – an R-language development environment.
- Remote R Server – this runs R on ultra2 which you can access using the RStudio environment, giving you the ability to run compute-intensive jobs on ultra with the flexibility of a GUI on the desktop
- Ultra2 Terminal – this opens a command-line terminal window for using the ultra2 computer. It simply runs “ssh ultra2” in a window.

How to use Ultra

What you need to know

- There are two computer systems you will use. The “ultra2” computer is a HPC system (High Performance Computing) with a vast amount of memory and processing power. The virtual desktop “c19-desktop” is much more limited and shared with other users, but it does have RStudio and PyCharmPro. Please try to do your analysis on ultra2 if you can, **not** on the desktop, out of consideration for other users. See below for details.
- Your account will be a member of a sub-group of u036, with the name u036-*project*, for example u036-isaric, u036-phosp or u036-collab. By default, the files in one sub-group *cannot* be read by members of a different sub-group.
- Your home directory should not be used for storing *project* files, only personal files. Please use one of the shared directories for project files.
- Project files are visible to everyone else in the project but to nobody else.
- No personally identifiable data may be stored on the system; all data must be pseudonymised. Please check the governance agreements before storing unconsented data. Be aware that adding variables and datasets may increase the risk of re-identification.

Directories

Home directories and project files for the u036 project live under /home/u036. The same paths and files are available from both ultra2 and the c19-desktop so you can share files between the two systems.

There are several sub-groups, such as “isaric” and “phosp”, and there is an additional sub-group called “collab” which is for external collaborators.

Your project files will be in /home/u036/u036-*subgroup*/shared/... These files are *only* accessible to members of your sub-group (isaric/phosp/collaborator). Please use a workspace directory for project files. These are located in /home/u036/u036-[subproject]/shared/workspace/, and allow multiple users to collaborate in a shared location for a given project, and data managers to perform data transfers in and out of the environment. To have a workspace directory created, contact a data manager. Include with your request:

- Your username
- The usernames of any collaborators who also need access
- Any preferred name for the folder (it will be named the same as your username by default)

To share files across the whole project, i.e. members of u036-isaric and u036-phosp, etc., you can use /home/u036/shared, but this is discouraged unless approval has been granted.

Summary:

/home/u036

/home/u036/shared – files accessible to members of every sub-group

/home/u036/u036/*username* – your personal files

/home/u036/u036/shared – files accessible to members of every sub-group

/home/u036/u036/shared/bin – programs accessible to members of every sub-group

/home/u036/u036-*subproject*/shared – files accessible to members of *subproject* only

/home/u036/u036-collab/shared – files accessible to members of external collaborators only

How to import and export data

The environment is deliberately restricted to prevent the extraction of data. This is for security reasons and also to prevent publication of data which is not yet approved for publication. The restriction on extraction also implies that data cannot be imported, and thus there is no internet access. However data managers do have permission to import and export data on your behalf.

To import data please contact your data manager.

To export data please contact your data manager.

Available Software

A selection of software has been made available specifically for project members:

- RStudio has been installed onto the c19-desktop VMs
- PyCharm Pro has been installed onto the c19-desktop VMs (bring your own licence)
- Anaconda has been installed, providing R and Python languages
- More R packages can be installed from CRAN and BioConductor repositories
- More Python packages can be installed from the PyPi repository
- Analysis programs: regenie, plink, vcftools, bgen, etc. See below
- Please ask if there is anything else you require.

Software supplied by Anaconda (R in particular) is described below (see Using Anaconda for R and Python).

Some programs are available from `/home/u036/u036/shared/bin`. This directory has been added to your PATH so you can use them easily:

- `cassi` – the CASSI genome-wide interaction analysis software
- `plink` and `plink2` – genome association analysis toolset, for genotype/phenotype data
- `regenie.sh` – whole genome regression modelling of large genome-wide association studies
- `vcftools` – work with complex genetic variation data in the form of VCF files
- `bgen` (`bgenix`, `bgen_to_vcf`, `edit-bgen`, `cat-bgen`) – utilities for the binary GEN file format

Some software which is not available from CRAN or PyPi is available in `/home/u036/u036/shared/software`:

- `rbgen` – R package for working with the binary GEN file format

Using Ultra2 for complex analysis

As mentioned above, the Ultra2 computer has vast amounts of memory and CPU power so is a better place to do complex analysis, especially anything with large datasets or which takes a long time to run. Please try not to do such work on the desktop because it is shared and has limited resources.

You can login to ultra2 using the desktop menu: Development > Ultra2 Terminal. From there you have access to the same files as on the desktop.

You can use R and Python, amongst other things, on Ultra, see the next section. If you need to use a GUI (eg. RStudio or PyCharm) then please see the sections below.

To start running large jobs it is better to invoke the scheduling/batch system. See the section below.

The generic documentation for Ultra2 (not specific to ISARIC/ODAP) will be available here:

<https://epcced.github.io/eidf-docs/services/ultra2/run/>

Some useful utilities have been installed

Using Anaconda for R and Python

A shared copy of anaconda3 has been installed and can be used by issuing the command:

```
source /home/u036/u036/shared/anaconda3/bin/activate
```

That will activate the base conda environment giving you access to additional environments. Your command prompt will now show (base) to indicate this.

Then you can activate a specific environment to get additional software, for example to use R you can issue the command:

```
conda activate Rv4
```


You will see your prompt change from (base) to (...Rv4).

Use `conda deactivate` when finished with that environment (or simply logout).

Using R Studio

RStudio can be started from the Applications | Development menu. It is using the Anaconda version of R; see above.

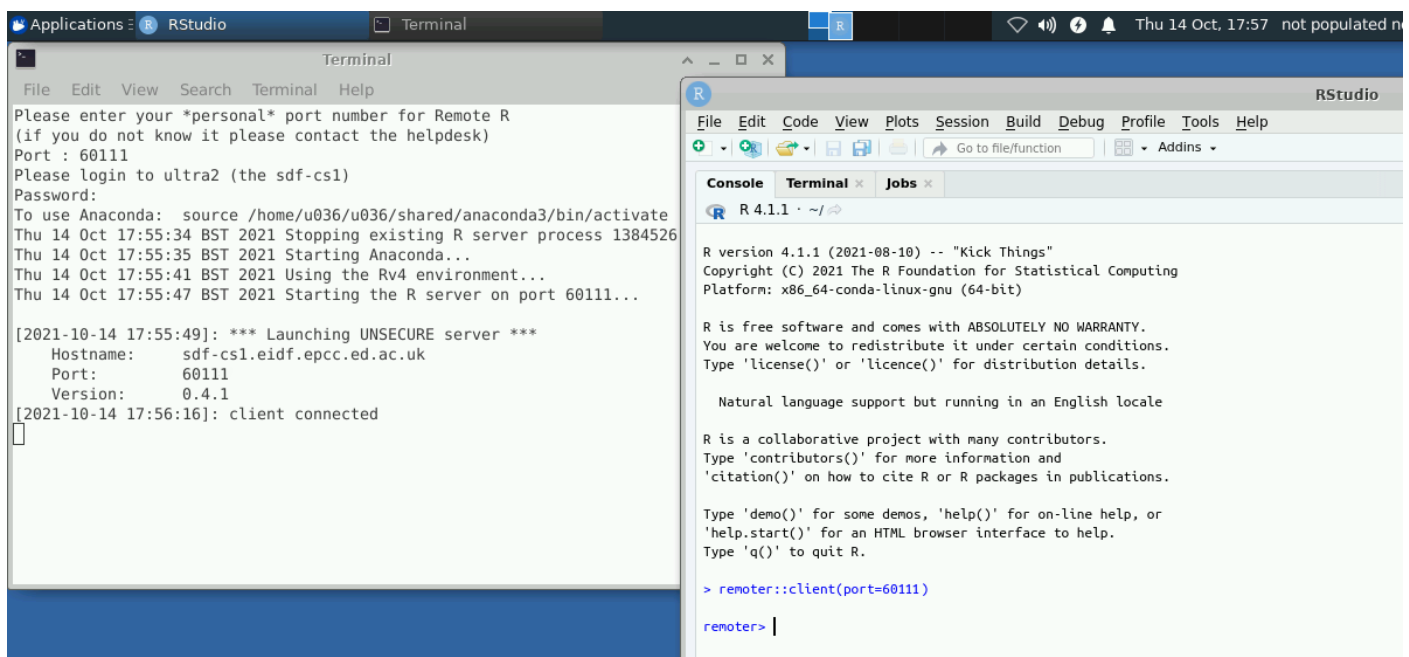
You can install additional R packages from CRAN into your personal directory using the normal command but it may be necessary to do this in a terminal window (not in RStudio):

```
source /home/u036/u036/shared/anaconda3/bin/activate
conda activate Rv4
R
> install.packages('DOSE')
```

As mentioned above, the resource constraints on the desktop mean that data-intensive and CPU-intensive work are better performed on the ultra2 computer. This can be done using RStudio as the GUI and connecting to an R Server running on ultra.

The first step is to ask EPCC's HPC Systems Team for a port number to be allocated to you (it will be something like 60123). (When logging the query ask them to check with abrooks).

Start the R Server using the Applications | Development | Remote R Server menu. This will prompt you for your personal port number. If you don't have one, please ask the helpdesk. Do not use somebody else's number!



When the server is running you can start RStudio and type:

```
remoter::client(port=N) # where N is your personal port number as above
```

Now all your variables are stored on ultra and all your R code will execute on ultra.

You can transfer a variable from the remote to the local using: `s2c(varname)` or from local to remote using `c2s()`.

You can see plots by using the `rpng()` command first, making the `plot()`, then retrieving it with `rpng.off()`. See the manual for more options.

When you have finished you can leave the remoter environment by typing:

```
exit()
```

and then close the Server window.

NOTE: If you have some analysis which will take a long time to run then please use the job scheduler; see the Slurm section below.

Troubleshooting R

Documentation:

<https://cran.r-project.org/web/packages/remoter/vignettes/remoter.pdf>

https://cran.r-project.org/web/packages/remoter/vignettes/remote_machines.pdf

<https://cran.r-project.org/web/packages/remoter/remoter.pdf>

install.packages() hangs or times out – you might need to specify the location of the CRAN mirror, for example:

```
install.packages('DOSE', repos='https://stats.bris.ac.uk/R/')
```

Bind failed: address already in use – this means that the R server is already running, please check you are using the correct port number, and if so then you don't need to start a new server. To see if the server is already running on ultra use this command and see if the output includes the command you used to start it: `pgrep -au$(id -u)`

channel 3: open failed: connect failed: Connection refused – this usually means that a client process is still running, i.e. inside your RStudio, but the remote server is not ready. If restarting RStudio does not help then the simplest solution is to reboot your computer.

Connection refused – this means that the R server is not running. If you previously started it then it may have crashed (this can happen due to uncaught R errors or if it would require interaction, such as trying to install a package without using the repos parameter). Try starting the server again, or exiting your RStudio.

Incompatible package versions – this happens when the versions of 'remoter' and 'pbdZMQ' on your RStudio do not match the versions on Ultra. In fact if your RStudio has newer versions than Ultra you will not see this message (however, see below). These packages are already installed so please contact the helpdesk.

Argument is of length zero (`get.status("method_plot_rpng") == "rasterImage"`) when plotting using `rpng.off()` – this happens when your RStudio version of 'remoter' is newer than the one on Ultra, typically if ultra is 0.4.0 and RStudio is 0.4.1. The solution, for now, is to downgrade your 'remoter' package in RStudio using the instructions above.

The R server keeps crashing – this happens when you try to execute an unknown command, particularly if a package has not been installed or loaded yet. In particular getting the parameters wrong for `ggplot()` will cause it to crash. This is a known bug, see <https://github.com/RBigData/remoter/issues/50> and a fix has been applied 2021-02-22.

If the server has crashed then you can restart it; there is no need to logout or login again.

If you wish to see error messages as they occur you can use the manual method for starting the R server as given above: login to ultra with ssh, source conda, activate Rv4, use the Rscript command to start the server. After a crash simply run the Rscript command again.

Using PyCharm on Ultra

It is not possible to use PyCharm on Ultra itself, because it is not a desktop environment, but it is possible to use PyCharm on the desktop and have it run the programs on Ultra.

The recommended way to use PyCharm on Ultra is to run it on the desktop and connect to a Python interpreter running on Ultra. This method has the benefit of a fast, responsive Python IDE running on the desktop, plus a Python interpreter running on the same machine as the data – the best of both worlds. You will need a full PyCharm license for this but it's free to students/teachers/etc. The full instructions are on the JetBrains website (links below) but the quick summary is:

File | New Project... | Name "*remote_ultra*"

File | New... | Python File | Name "*remote_ultra_test.py*" and add some code OR re-use existing project

File | Settings | Project: *name* | Project Interpreter

click the cog at the end of the Project Interpreter | Add...

In the Add Python Interpreter window choose SSH Interpreter in the left column

Enter Host: *ultra2* and Username: your existing username on *ultra*, click Next

Enter your *ultra* Password: and tick Save Password, click Next

Choose a Python interpreter, the default `/usr/bin/python` is v2.7.5 (old!),

or choose a Python interpreter from an installed Conda environment, such as

`/home/u036/shared/conda_environments/<environment name>/bin/python`, or

`/home/u036/shared/anaconda3/bin/python`, which is v3.7.6

Sync folders: click on the folder icon at the end of the Sync folders:

click in the Remote Path entry and change it from `/tmp/pycharm_project_N` to

`/home/u036/<your username>/PycharmProjects/<temporary project name>`, click Finish.

(Note! Change `/home/u036` to your own home directory)

(Note! Don't use the same project name as your local copy or they will clash)

File | Settings | Appearance and Behaviour | System Settings | HTTP Proxy

enter hostname *c19-desktop-proxy* and port 800

File | Settings | Build, Execution, Deployment | Deployment

click on the Mappings tab,

change the Deployment path: to the same path you entered in Sync folders.

Click OK (Wait until the Network Transfer tab has finished uploading all the deployment configuration to Ultra.)

Run | Run... | select the name of the configuration to run your code directly on *ultra*.

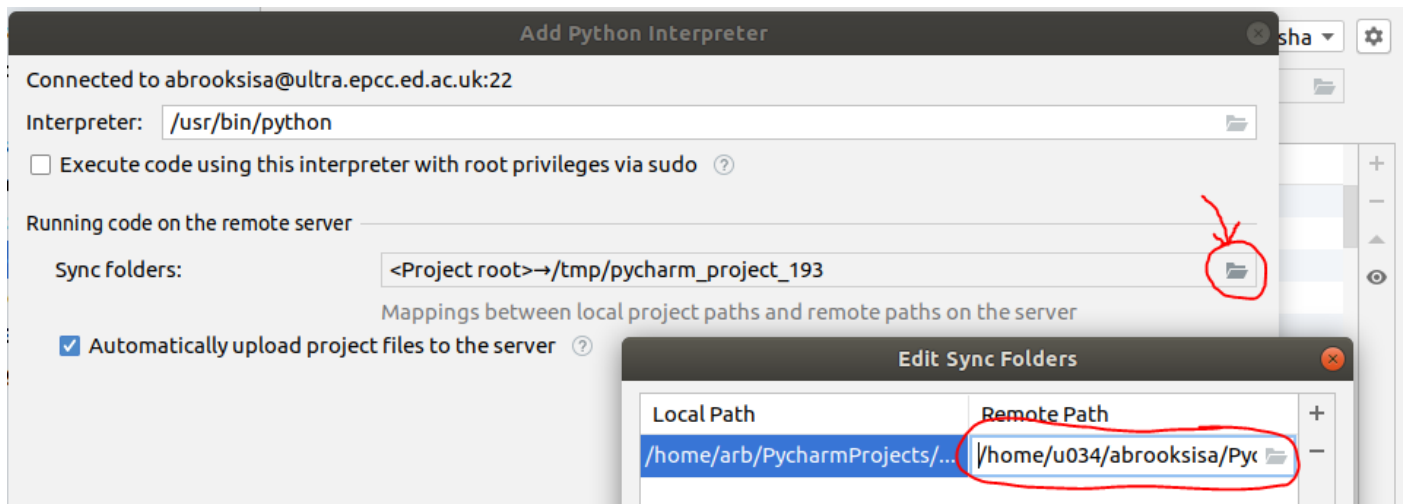
Control the upload of files to *ultra* from the Tools | Deployment menu.

References:

<https://www.jetbrains.com/help/pycharm/configuring-remote-interpreters-via-ssh.html>

and <https://www.jetbrains.com/help/pycharm/remote-debugging-with-product.html>

The Sync folders dialogue box:



NOTE: If you have some analysis which will take a long time to run then please use the job scheduler; see the Slurm section below.

Running time-consuming jobs on Ultra2 with RemoteR

As described above, you can run complex and time-consuming jobs on Ultra2 using the 'remoter' library in RStudio. However, the interactive nature of this workflow might make it unsuitable for some tasks. For example if it is going to take several hours you might be concerned that a loss of network connection will cause the job to be terminated. There are two solutions:

- run the job with a 'remoter' server that can be disconnected from the desktop
- run the job using a batch job scheduler. This is described in the next section.

To start a 'remoter' server which is not connected with the desktop environment:

- Use the "(Ultra2 tunnel)" entry in the Applications|Development menu
- Enter your personal port number
- If it reports an error then you already have a tunnel open
- It will prompt you to login to ultra2 so enter your password
- At the command prompt \$ enter the command:
`/home/u036/u036/shared/bin/run_R_server.sh N &`
where N is your personal port number. Make sure you add the ampersand character at the end. The R server is now running in the background. If your connection is interrupted, or you log off from the desktop, the server is still running and will preserve your variables. You can connect to it from RStudio in the same way as normal.

Running large jobs on Ultra2 with Slurm

Large, compute-intensive jobs should be run through Ultra's batch job scheduler, which on Ultra2 is Slurm. See the generic documentation for Ultra2 here: <https://epcced.github.io/eidf-docs/services/ultra2/run/>

To submit a large job, e.g. an R script on a large dataset, put all the commands used for your analysis into a batch file, e.g. `example_job.slurm`:

```
#!/bin/bash
#SBATCH --job-name=example_job
#SBATCH --cpus-per-task=2
#SBATCH --mem=8GB
#SBATCH --time=1:00:00
#SBATCH --output=/home/u036/u036/me/job_output.log
# go to the directory the script is being submitted from
cd $SLURM_SUBMIT_DIR
source /home/u036/u036/shared/anaconda3/bin/activate
conda activate Rv4
Rscript your_analysis.R
```

Scripts can be submitted by running, e.g. `sbatch example_job.slurm`. You can view the status of scheduled and currently running jobs with the command `squeue`, and finished jobs can be checked with the command `sacct`, or `sacct -j <job_id>`.

Documentation on using Slurm, including features like job arrays, can be found at <https://slurm.schedmd.com/documentation.html> and <https://slurm.schedmd.com/quickstart.html>

For information on migrating PBS scripts from Ultra 1 to Slurm, see SchedMD's Rosetta Stone of Workload Managers at <https://slurm.schedmd.com/rosetta.pdf>

Accessing Ultra 2 from the desktop VM

Log into Ultra 2 machine from the graphical desktop by opening the menu 'Application -> Development -> Ultra2 Terminal', or on the command line by running `ssh ultra2` (or `ssh sdf-cs1` but note that `ssh sdf-cs1.epcc.ed.ac.uk` will not work).

Access to external databases from Ultra

A database for ISARIC has been created on a separate host, called c19-isaric01. This host is accessible from the desktop VMs and from Ultra2 using the hostname c19-database-proxy. If it does not resolve then the IP address (as seen from Ultra2) is 172.16.28.56 and from the VMs it is 192.168.134.48

To request access to the ISARIC database, raise a request with the EPCC helpdesk with your Ultra username and whether or not you need to be able to insert data. You will receive back your login credentials once you have been given the relevant level of access.

The database name is "isaric" and the schema name is "isaric". Once connected to the "isaric" database you can refer to tables as "isaric.my_table".

Access from the command line

You can connect to the database from the Ultra command line by running:

```
psql -h c19-database-proxy -U <pg_username> -d isaric
```

where pg_username is the username you were given when requesting database access. As per the [Postgres docs](#), you can also store your database credentials in the file .pgpass in your home folder, in the format hostname:port:database:username:password, and you will not need to supply your password each time. The port in this case will be the default Postgres port 5432. Postgres will only accept this file if it's accessible only by you and nobody else. To assign the right file permissions, run:

```
chmod go-rwx ~/.pgpass
```

Once in you can run SQL queries, as well as psql commands like:

- \dt isaric.* to list tables
- \dv isaric.* to list views
- \d isaric.<table_or_view_name> to show column information for a table or view
- \password to change your Postgres password

The \d... commands can also be appended with a '+' to view extra information like any given table/column descriptions, or the explicit SQL query that a view is made up of:

- \dt+ isaric.*
- \d+ isaric.<table_or_view_name>

Access from R

The database server is PostgreSQL, so to connect to it from R on ultra:

```
library('RPostgreSQL')

pg_con <- DBI::dbConnect(RPostgreSQL::PostgreSQL(), dbname="isaric", host="c19-
database-proxy", user="myusername", password="mypassword")

library('tidyverse')

my_tbl <- tbl(pg_con, sql('select * from my_table'))

OR

my_tbl <- tbl(pg_con, 'my_table')

my_tbl %>% select(stuff) %>% filter(stuff)
```

```
dbDisconnect(pg_con)
```

to dynamically create SQL statements using R syntax.

The database name is "isaric" and the schema name is "isaric". Once connected to the "isaric" database you can refer to tables as "isaric.my_table".