Utrecht University

Data Science Day 2018

High Performance Computing for Data Science

Roel Brouwer, Kees van Eijden en Jonathan de Bruin

20 April 2018

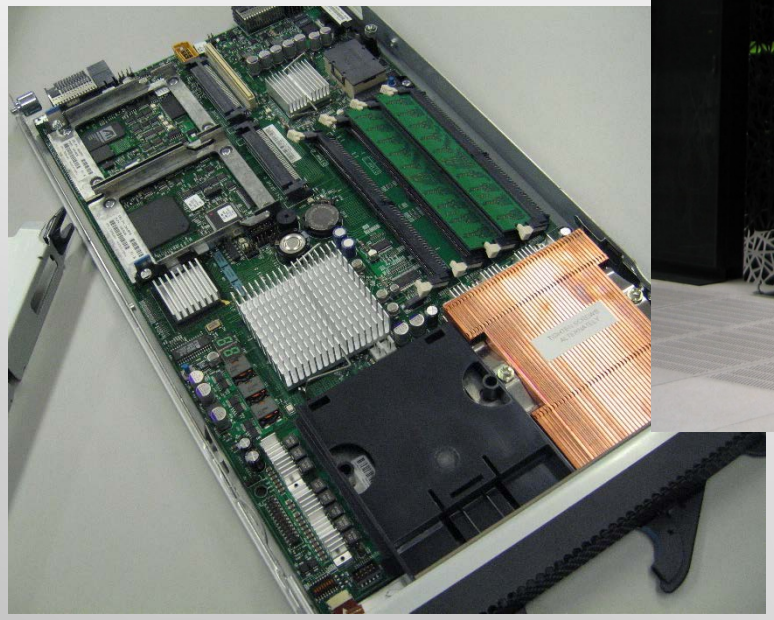
Outline

- **Welcome**
- **Introduction to HPC**
- **Data Science problem**
- **Hands-on session**

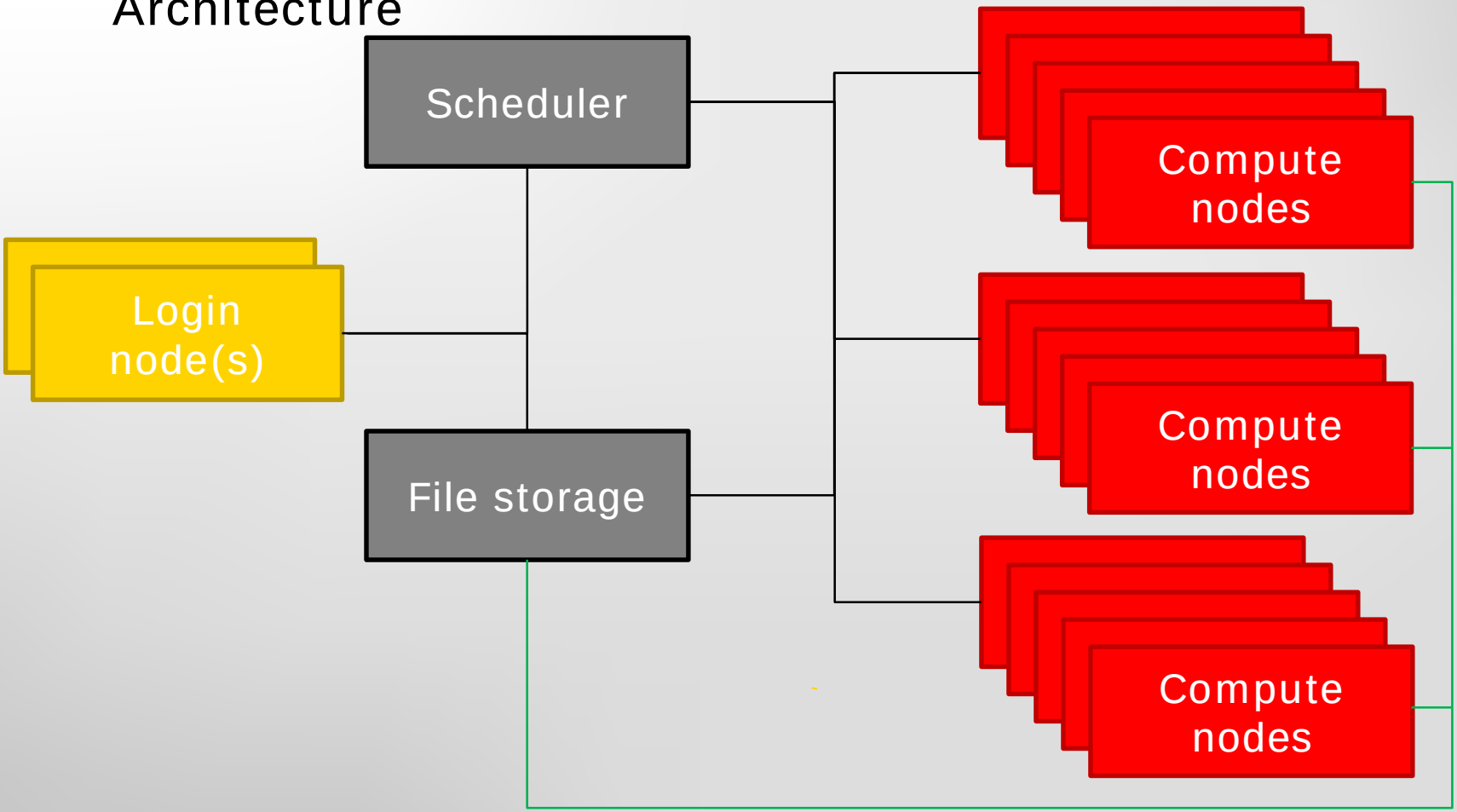
What is HPC?

"High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business."

Hardware



Architecture



Software

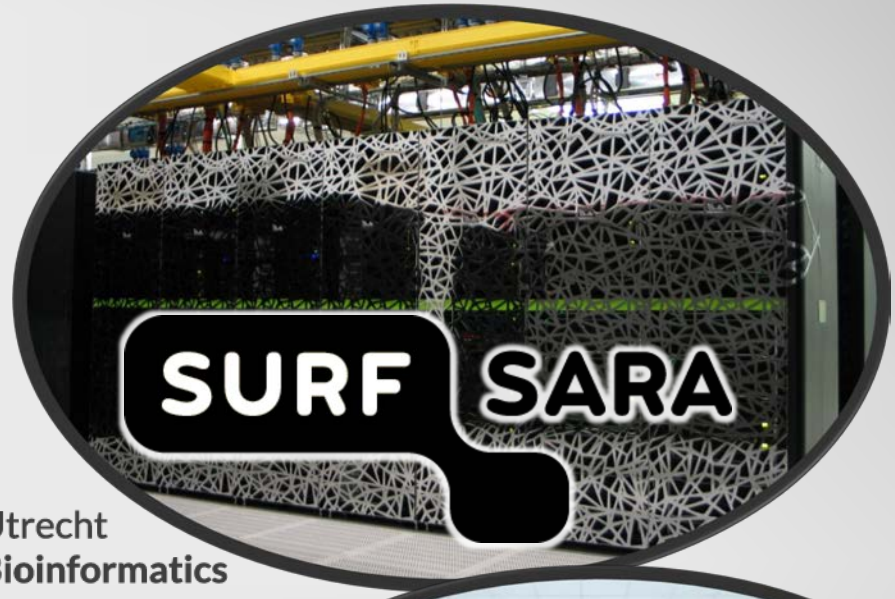
- **Operating system: Linux**
- **Scheduler / batch system**
  - SLURM, TORQUE, OGE/SGE, ...
  - Jobs are submitted to the scheduler
  - The scheduler determines when and where to execute jobs

Scheduler

```
     JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
   4128724 broadwell [illegible] [illegible] PD       0:00      1 (Resources)
   4128726    normal [illegible] [illegible] PD       0:00      3 (Priority)
   4128727    normal [illegible] [illegible] PD       0:00      1 (Priority)
   4128731    normal [illegible] [illegible] PD       0:00      4 (Priority)
   4128732    normal [illegible] [illegible] PD       0:00      4 (Priority)
   4128741    normal [illegible] [illegible] PD       0:00      5 (Priority)
   4127728    normal [illegible] [illegible] PD       0:00      1 (Dependency)
   4127730    normal [illegible] [illegible] PD       0:00      1 (Dependency)
   4127554    normal [illegible] [illegible]  R   11:47:45     32 tcn[1297-1300,1302,1305,1310,1359-1363,1513-1518,
   4127529    normal [illegible] [illegible]  R   14:49:28      8 tcn[743,748,766-768,1239,1242,1276]
   4127525    normal [illegible] [illegible]  R   15:19:20      8 tcn[740,796,830,1153,1173,1184-1185,1193]
   4128100    normal [illegible] [illegible]  R    3:40:20     32 tcn[1423-1424,1426-1428,1430,1432-1436,1438-1440,
   4127919    normal [illegible] [illegible]  R    5:13:01     16 tcn[1003-1004,1020,1022,1052-1053,1059,1062,1174,
   4109392       gpu [illegible] [illegible]  R 3-17:27:01     32 gcn[8,10-11,13-34,36-42]
4120830_14       gpu [illegible] [illegible]  R    7:05:09      1 gcn59
4120830_10       gpu [illegible] [illegible]  R   21:50:26      1 gcn48
4120830_11       gpu [illegible] [illegible]  R   21:50:26      1 gcn7
4120830_12       gpu [illegible] [illegible]  R   21:50:26      1 gcn35
4120830_13       gpu [illegible] [illegible]  R   21:50:26      1 gcn12
 4120830_7       gpu [illegible] [illegible]  R   21:54:29      1 gcn51
 4120830_8       gpu [illegible] [illegible]  R   21:54:29      1 gcn56
 4120830_9       gpu [illegible] [illegible]  R   21:54:29      1 gcn57
 4120830_6       gpu [illegible] [illegible]  R   21:54:30      1 gcn9
```

What is available?

- **Supercomputers / compute clusters**



- **Cloud computing**

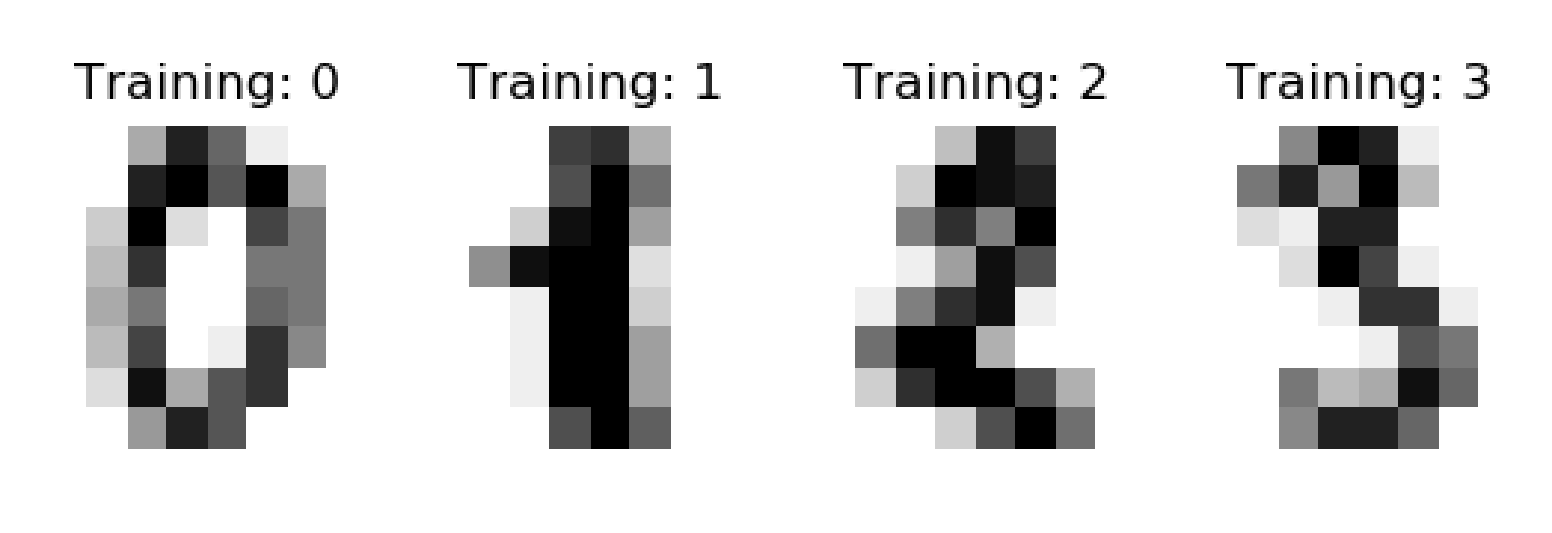
Utrecht Bioinformatics Center

amazon web services

Google Cloud

Azure

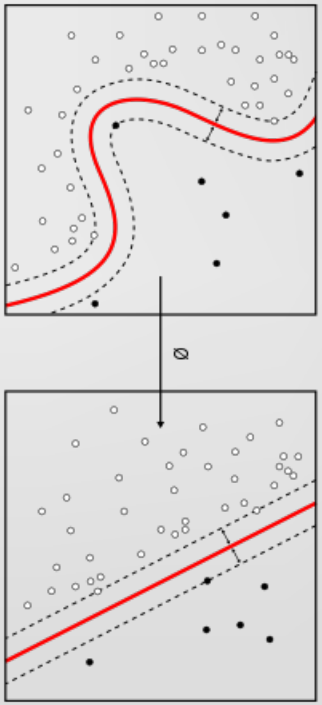
Data science problem: Recognizing hand-written digits.



Alpaydin, E., & Alimoglu, F. (1998). Pen-based recognition of handwritten digits. *Department of Computer Engineering, Bogazici University*.

Data science problem: Support Vector Machines

- Support vector machines for classification
- Find the optimal values of C and γ
- Hyperparameter optimization
  - One approach: **Grid Search**
  - $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$; $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$;



Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Data science problem: Scripts

- R and Python script available on github.com/UtrechtUniversity/datascienceday-hpc
- Pass the hyperparameters as command-line arguments

```
$ python digits_svm.py -C 1.0 -G 0.001
```

Hands-on

- **Instructions on your desk and available online**
- **Repository:**
  https://github.com/UtrechtUniversity/datascienceday-hpc
- **Credentials**
  Username: ws_hpcXX
  Password: DaSciDa#XX

MobaXterm

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

Record new macro

Saved macros

Sessions Tools Macros

Session settings

SSH Telnet Rsh Xdmcp RDP VNC FTP SFTP Serial File Shell Browser Mosh Aws S3

Basic SSH settings

Remote host * cartesius.surfsara.nl Specify username ws_hpc01 Port 22

Advanced SSH settings Terminal settings Network settings Bookmark settings

Secure Shell (SSH) session

OK Cancel

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

GitHub - UtrechtUniversi

GitHub, Inc. [US] | https://github.com/UtrechtUniversity/datascienceday-hpc

Features Business Explore Marketplace Pricing This repository Search Sign in or Sign up

UtrechtUniversity / datascienceday-hpc

Watch 2 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Insights

workshop material for Data Science Day 2018 HPC related

4 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Find file Clone or download

J535D165 Update README

R	Add workshop scripts
data	Add datasets
python	Add workshop scripts
README.md	Update README

Clone with HTTPS

Use Git or checkout with SVN using the web URL.

https://github.com/UtrechtUniversity/dat

Open in Desktop Download ZIP

README.md

Data Science Day 2018 Workshop - HPC: Speed up your data science problems!

Authors: Roel Brouwer and Kees van Eijden and Jonathan de Bruin

https://github.com/UtrechtUniversity/datascienceday-hpc/archive/master.zip

cartesius.surfsara.nl (ws_hpc29)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

4. cartesius.surfsara.nl (ws_hpc29)

/nfs/home1/ws_hpc29/

Name	Size
..	
.cache	
.local	
R	
.bash_history	1
.bash_logout	1
.bash_profile	1
.bashrc	1
.forward	1
.Xauthority	1

Parent directory
New directory
New empty file
Refresh current folder
Upload to current folder

Follow terminal folder

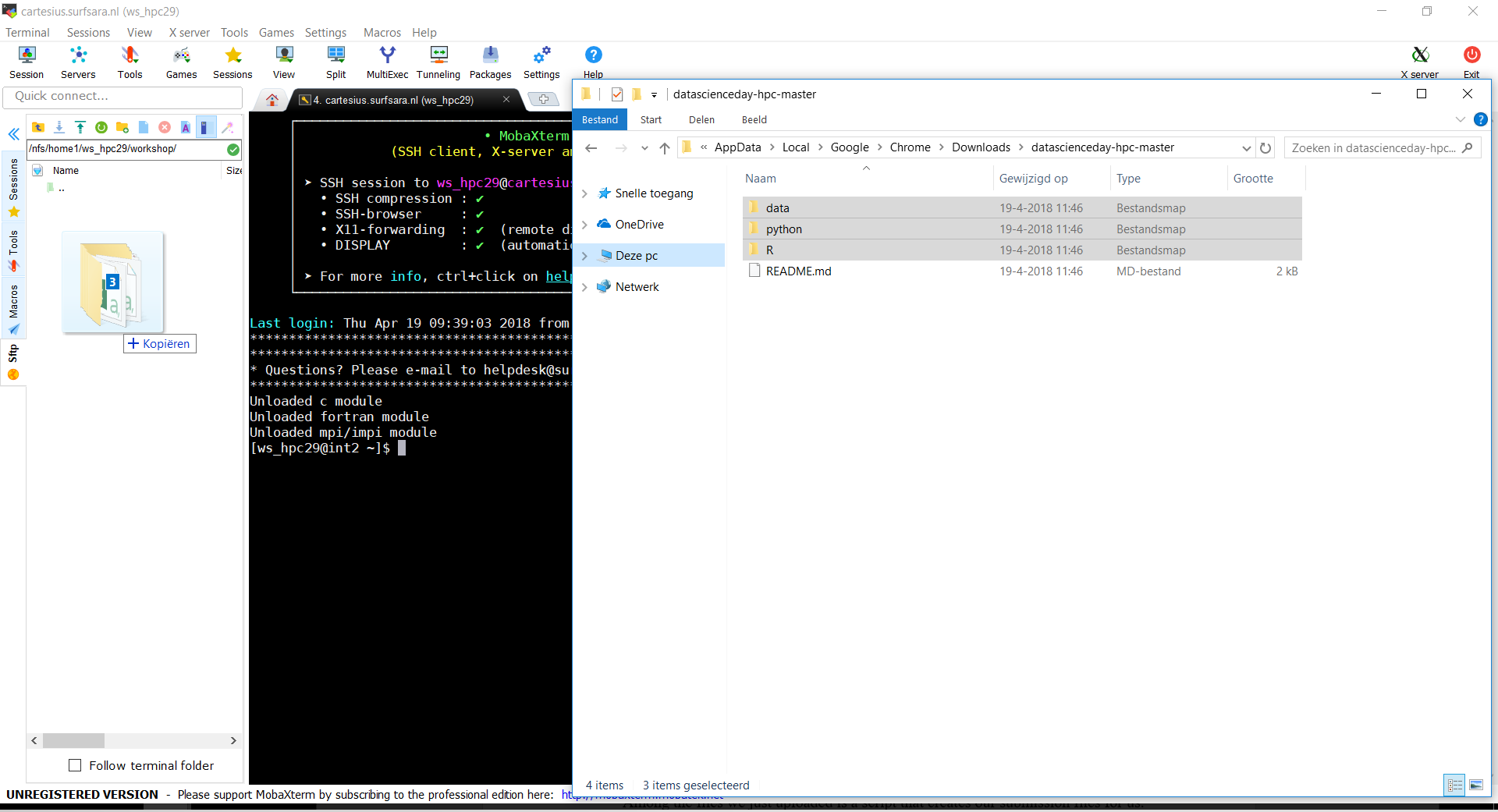
```
                    • MobaXterm 10.4 •
            (SSH client, X-server and networking tools)

 ➢ SSH session to ws_hpc29@cartesius.surfsara.nl
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote display is forwarded through SSH)
   • DISPLAY         : ✔  (automatically set on remote server)

 ➢ For more info, ctrl+click on help or visit our website

Last login: Thu Apr 19 09:39:03 2018 from 145.107.138.185
*******************************************************************************
*******************************************************************************
* Questions? Please e-mail to helpdesk@surfsara.nl, or call 020-8001400       *
******************************************** last modified   12-04-2018 21:48 ***
ded c module
ded fortran module
ded mpi/impi module
pc29@int2 ~]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net


cartesius.surfsara.nl (ws_hpc29)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
/nfs/home1/ws_hpc29/workshop/
Name
Kopiëren
Follow terminal folder
4. cartesius.surfsara.nl (ws_hpc29)
• MobaXterm
(SSH client, X-server a
➤ SSH session to ws_hpc29@cartesiu
• SSH compression : ✓
• SSH-browser : ✓
• X11-forwarding : ✓ (remote d
• DISPLAY : ✓ (automati
➤ For more info, ctrl+click on hel
Last login: Thu Apr 19 09:39:03 2018 from
* Questions? Please e-mail to helpdesk@su
Unloaded c module
Unloaded fortran module
Unloaded mpi/impi module
[ws_hpc29@int2 ~]$
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here:
datascienceday-hpc-master
Bestand Start Delen Beeld
AppData › Local › Google › Chrome › Downloads › datascienceday-hpc-master
Zoeken in datascienceday-hpc...
Snelle toegang
OneDrive
Deze pc
Netwerk
Naam Gewijzigd op Type Grootte
data 19-4-2018 11:46 Bestandsmap
python 19-4-2018 11:46 Bestandsmap
R 19-4-2018 11:46 Bestandsmap
README.md 19-4-2018 11:46 MD-bestand 2 kB
4 items 3 items geselecteerd

cartesius.surfsara.nl (ws_hpc29)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

4. cartesius.surfsara.nl (ws_hpc29)

/nfs/home1/ws_hpc29/workshop/

Name Size

..
batch_files
data
python
R

Follow terminal folder

```
                    • MobaXterm 10.4 •
            (SSH client, X-server and networking tools)

 ➢ SSH session to ws_hpc29@cartesius.surfsara.nl
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote display is forwarded through SSH)
   • DISPLAY         : ✔  (automatically set on remote server)

 ➢ For more info, ctrl+click on help or visit our website


Last login: Thu Apr 19 09:39:03 2018 from 145.107.138.185
*******************************************************************************
*******************************************************************************
* Questions? Please e-mail to helpdesk@surfsara.nl, or call 020-8001400       *
****************************************** last modified   12-04-2018 21:48 ***
Unloaded c module
Unloaded fortran module
Unloaded mpi/impi module
[ws_hpc29@int2 ~]$ cd workshop
[ws_hpc29@int2 workshop]$ python ./python/batch_script_generation.py -n 1
[ws_hpc29@int2 workshop]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

cartesius.surfsara.nl (ws_hpc29)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

5. cartesius.surfsara.nl (ws_hpc29)

/nfs/home1/ws_hpc29/workshop/

Name Size

..
batch_files
data
output
python
R
slurm-4173246.out 1

Follow terminal folder

```
                    • MobaXterm 10.4 •
            (SSH client, X-server and networking tools)

 ➢ SSH session to ws_hpc29@cartesius.surfsara.nl
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote display is forwarded through SSH)
   • DISPLAY         : ✔  (automatically set on remote server)

 ➢ For more info, ctrl+click on help or visit our website


Last login: Thu Apr 19 12:44:00 2018 from 145.107.138.185
*******************************************************************************
*******************************************************************************
* Questions? Please e-mail to helpdesk@surfsara.nl, or call 020-8001400       *
****************************************** last modified   12-04-2018 21:48 ***
Unloaded c module
Unloaded fortran module
Unloaded mpi/impi module
[ws_hpc29@int2 ~]$ cd workshop
[ws_hpc29@int2 workshop]$ python ./python/batch_script_generation.py -n 1
[ws_hpc29@int2 workshop]$ sbatch ./batch_files/batch-0.sh
Submitted batch job 4173246
[ws_hpc29@int2 workshop]$ squeue -u ws_hpc29
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           4173246    normal batch-0. ws_hpc29 PD       0:00      1 (Priority)
[ws_hpc29@int2 workshop]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

cartesius.surfsara.nl (ws_hpc29)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

5. cartesius.surfsara.nl (ws_hpc29)

/nfs/home1/ws_hpc29/workshop/output/

Name

..
digits_f1_plot.pdf
digits_svm0_C_0.03125_gamma_...
digits_svm100_C_32768.0_gamm...
digits_svm101_C_32768.0_gamm...
digits_svm102_C_32768.0_gamm...
digits_svm103_C_32768.0_gamm...
digits_svm104_C_32768.0_gamm...
digits_svm105_C_32768.0_gamm...
digits_svm106_C_32768.0_gamm...
digits_svm107_C_32768.0_gamm...
digits_svm108_C_32768.0_gamm...
digits_svm109_C_32768.0_gamm...
digits_svm10_C_0.125_gamma_3...
digits_svm11_C_0.125_gamma_0...
digits_svm12_C_0.125_gamma_0...
digits_svm13_C_0.125_gamma_0...
digits_svm14_C_0.125_gamma_0...
digits_svm15_C_0.125_gamma_0...
digits_svm16_C_0.125_gamma_0...
digits_svm17_C_0.125_gamma_0...
digits_svm18_C_0.125_gamma_2...
digits_svm19_C_0.125_gamma_8...
digits_svm1_C_0.03125_gamma_...
digits_svm20_C_0.5_gamma_3.0...
digits_svm21_C_0.5_gamma_0.0...
digits_svm22_C_0.5_gamma_0.0...
digits_svm23_C_0.5_gamma_0.0...
digits_svm24_C_0.5_gamma_0.0...
digits_svm25_C_0.5_gamma_0.0...
digits_svm26_C_0.5_gamma_0.1...
digits_svm27_C_0.5_gamma_0.5.txt
digits_svm28_C_0.5_gamma_2.0.txt
digits_svm29_C_0.5_gamma_8.0.txt
digits_svm2_C_0.03125_gamma_...

Follow terminal folder

```
                    • MobaXterm 10.4 •
            (SSH client, X-server and networking tools)

  ➢ SSH session to ws_hpc29@cartesius.surfsara.nl
    • SSH compression : ✔
    • SSH-browser     : ✔
    • X11-forwarding  : ✔  (remote display is forwarded through SSH)
    • DISPLAY         : ✔  (automatically set on remote server)

  ➢ For more info, ctrl+click on help or visit our website

Last login: Thu Apr 19 12:44:00 2018 from 145.107.138.185
*******************************************************************************
*******************************************************************************
* Questions? Please e-mail to helpdesk@surfsara.nl, or call 020-8001400       *
****************************************** last modified   12-04-2018 21:48 ***
Unloaded c module
Unloaded fortran module
Unloaded mpi/impi module
[ws_hpc29@int2 ~]$ cd workshop
[ws_hpc29@int2 workshop]$ python ./python/batch_script_generation.py -n 1
[ws_hpc29@int2 workshop]$ sbatch ./batch_files/batch-0.sh
Submitted batch job 4173246
[ws_hpc29@int2 workshop]$ squeue -u ws_hpc29
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           4173246    normal batch-0. ws_hpc29 PD       0:00      1 (Priority)
[ws_hpc29@int2 workshop]$ python ./python/aggregation.py
('Grid length:', 110)
('Optimal settings:', {'f1': 0.98553423654567174, 'cost': '2.0', 'gamma': '0.001953125'})
[ws_hpc29@int2 workshop]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

cartesius.surfsara.nl (ws_hpc29)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit

Quick connect...

/nfs/home1/ws_hpc29/workshop/output/

Name Siz

..
digits_f1_plot.pdf 12
digits_svm0_C_0.03125_gamma_... 3
digits_svm100_C_32768.0_gamm... 3
digits_svm101_C_32768.0_gamm... 3
digits_svm102_C_32768.0_gamm... 3
digits_svm103_C_32768.0_gamm... 3
digits_svm104_C_32768.0_gamm... 3
digits_svm105_C_32768.0_gamm... 3
digits_svm106_C_32768.0_gamm... 3
digits_svm107_C_32768.0_gamm... 3
digits_svm108_C_32768.0_gamm... 3
digits_svm109_C_32768.0_gamm... 3
digits_svm10_C_0.125_gamma_3... 3
digits_svm11_C_0.125_gamma_0... 3
digits_svm12_C_0.125_gamma_0... 3
digits_svm13_C_0.125_gamma_0... 3
digits_svm14_C_0.125_gamma_0... 3
digits_svm15_C_0.125_gamma_0... 3
digits_svm16_C_0.125_gamma_0... 3
digits_svm17_C_0.125_gamma_0... 3
digits_svm18_C_0.125_gamma_2... 3
digits_svm19_C_0.125_gamma_8... 3
digits_svm1_C_0.03125_gamma_... 3
digits_svm20_C_0.5_gamma_3.0... 3
digits_svm21_C_0.5_gamma_0.0... 3
digits_svm22_C_0.5_gamma_0.0... 3
digits_svm23_C_0.5_gamma_0.0... 3
digits_svm24_C_0.5_gamma_0.0... 3
digits_svm25_C_0.5_gamma_0.0... 3
digits_svm26_C_0.5_gamma_0.1... 3
digits_svm27_C_0.5_gamma_0.5.txt 3
digits_svm28_C_0.5_gamma_2.0.txt 3
digits_svm29_C_0.5_gamma_8.0.txt 3
digits_svm2_C_0.03125_gamma_... 3

Follow terminal folder

5. cartesius.surfsara.nl (ws_hpc29)

```
                    • MobaXterm
          (SSH client, X-server a

 ➢ SSH session to ws_hpc29@cartesiu
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote d
   • DISPLAY         : ✔  (automati

 ➢ For more info, ctrl+click on hel

Last login: Thu Apr 19 12:44:00 2018 from
*****************************************
*****************************************
* Questions? Please e-mail to helpdesk@su
*****************************************
Unloaded c module
Unloaded fortran module
Unloaded mpi/impi module
[ws_hpc29@int2 ~]$ cd workshop
[ws_hpc29@int2 workshop]$ python ./python
[ws_hpc29@int2 workshop]$ sbatch ./batch_
Submitted batch job 4173246
[ws_hpc29@int2 workshop]$ squeue -u ws_hp
             JOBID PARTITION     NAME
           4173246    normal batch-0. ws_
[ws_hpc29@int2 workshop]$ python ./python
('Grid length:', 110)
('Optimal settings:', {'f1': 0.9855342365
[ws_hpc29@int2 workshop]$
```

datascienceday-hpc-master

Bestand Start Delen Beeld

« AppData › Local › Google › Chrome › Downloads › datascienceday-hpc-master

Zoeken in datascienceday-hpc...

Snelle toegang
OneDrive
Deze pc
Netwerk

Naam	Gewijzigd op	Type	Grootte
data	19-4-2018 11:46	Bestandsmap	
python	19-4-2018 11:46	Bestandsmap	
R	19-4-2018 11:46	Bestandsmap	
README.md	19-4-2018 11:46	MD-bestand	2 kB

4 items 3 items geselecteerd

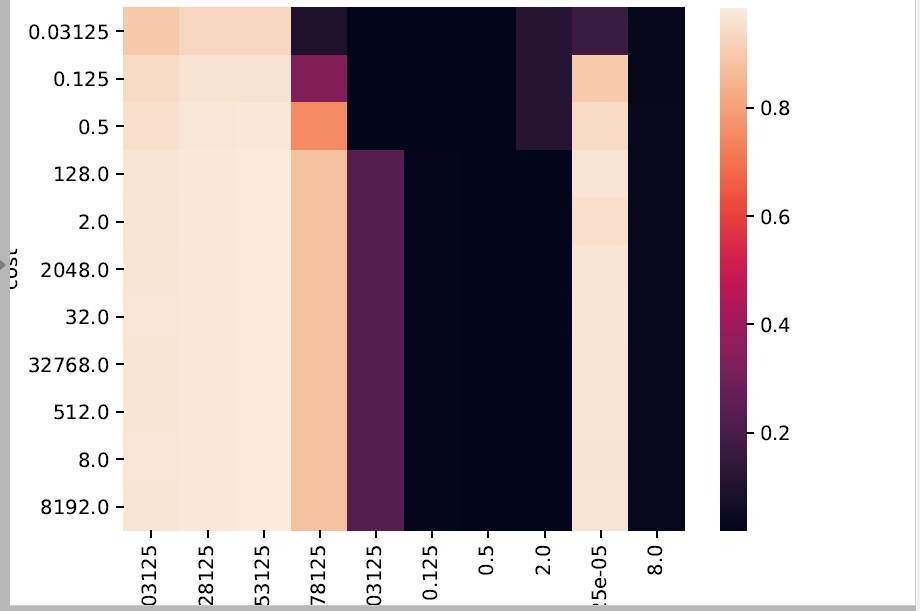
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

cartesius.surfsara.nl (ws_hpc29)

digits_f1_plot.pdf - Adobe Acrobat Pro DC

Bestand Bewerken Beeld Venster Help

Start Gereedschappen Document 1 / 1



X server Exit

y-hpc-master

eeld

ata > Local > Google > Chrome > Downloads > datascienceday-hpc-master

Zoeken in datascienceday-hpc...

Naam	Gewijzigd op	Type	Grootte
data	19-4-2018 11:46	Bestandsmap	
python	19-4-2018 11:46	Bestandsmap	
R	19-4-2018 11:46	Bestandsmap	
digits_f1_plot	19-4-2018 13:12	Adobe Acrobat D...	13 kB
README.md	19-4-2018 11:46	MD-bestand	2 kB

12.9 kB

Contact

RDM Support – High Performance Computing

Roel Brouwer and Kees van Eijden

info.rdm@uu.nl

20 April 2018