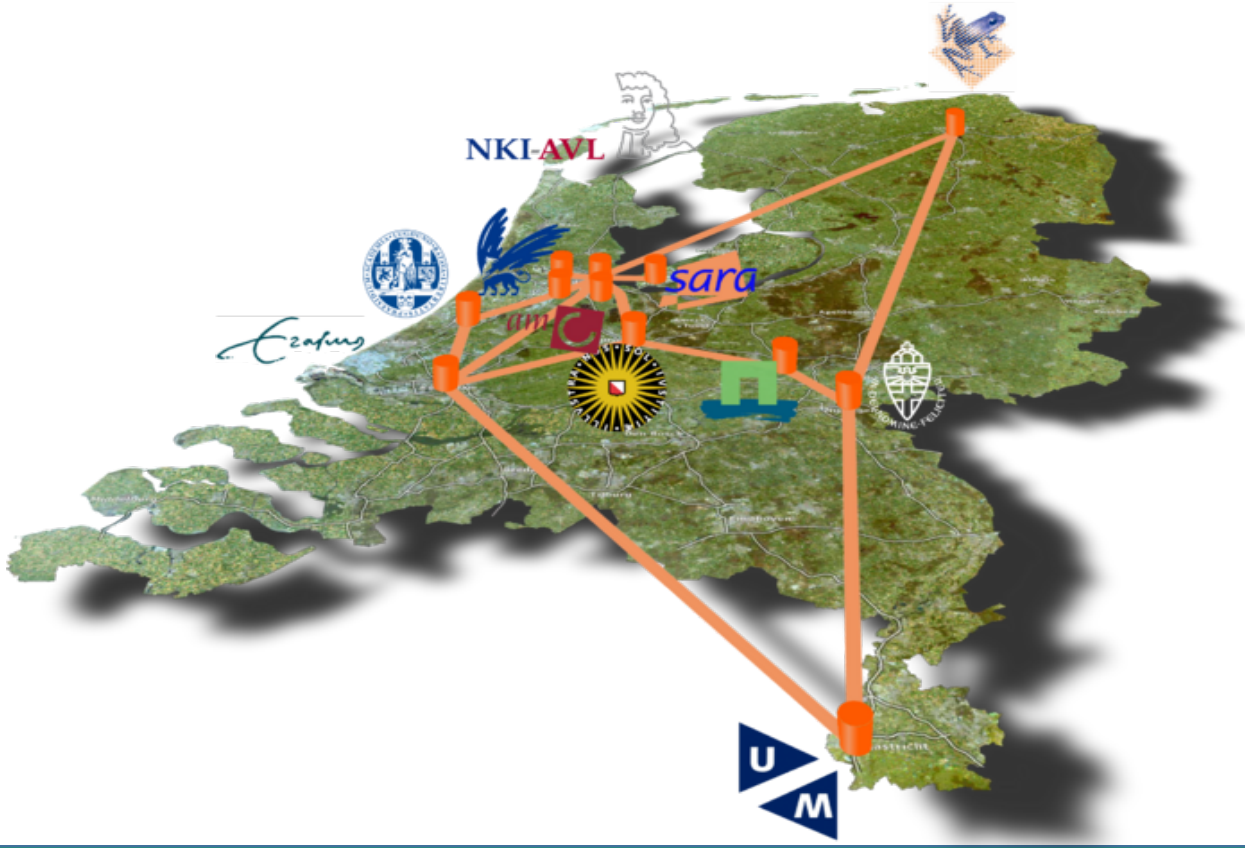
Grid and cluster Computing

Introduction to life science grid (LSG) and hands-on



Donders Institute for Brain, Cognition and Behaviour
21-22 November, 2016
Maithili Kalamkar, Natalie Danezi

SURF SARA

Content

- What is the Dutch e-Infrastructure?
- Introduction to Life Science Grid (LSG)
- Access the local cluster
- Submit your first job
- How can you sign up to use the Grid

Dutch e-Infrastructure

- **What is an e-Infrastructure?**
  Researchers in various locations across the world share distributed facilities in order to deal with increased complexity of scientific problems

- SURFsara offers the following e-Infrastructures to the researchers in the Netherlands:

  ❑ Compute services
    - Supercomputer (Cartesius)
    - Lisa cluster
    - Graphic Processing Unit cluster (part of Lisa)
    - **Grid**
    - High Performance Computing Cloud
    - Hadoop cluster

  ❑ Data services
  ❑ Visualization services
  ❑ Networking services
  ❑ Consultancy and Support

Introduction to Life science Grid

What is Grid?

- A collection of geographically distributed interconnected clusters

What is a cluster?

- A collection of computers coupled to work together
(Cluster-Animation http://web.grid.sara.nl/mooc/animations/)

So . . .

- Grid is a collection of geographically distributed clusters, data storage capabilities and services that can dynamically join in a transparent way

- A Grid user is able to use all this computing power simultaneously, without having to log in at all the different sites

- The clusters are interconnected by a high speed SURFnet network

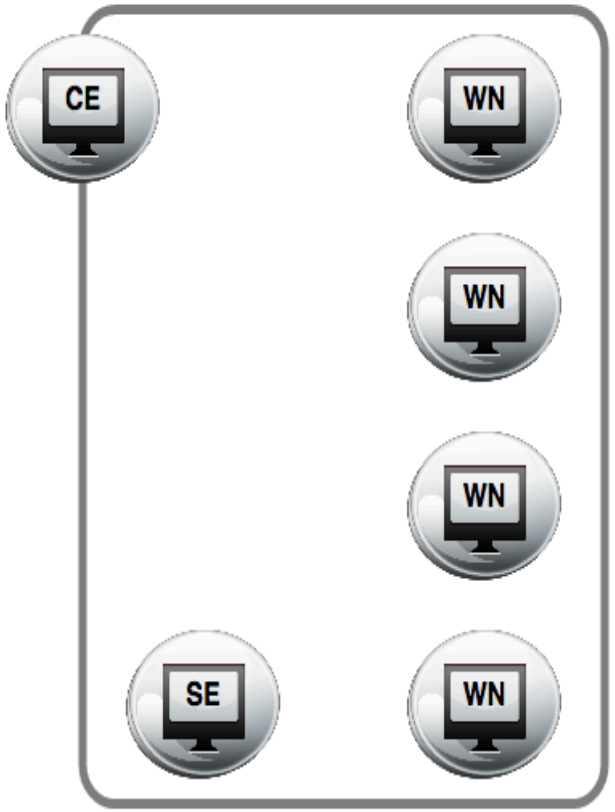


Fig: A cluster

Introduction to Life science Grid

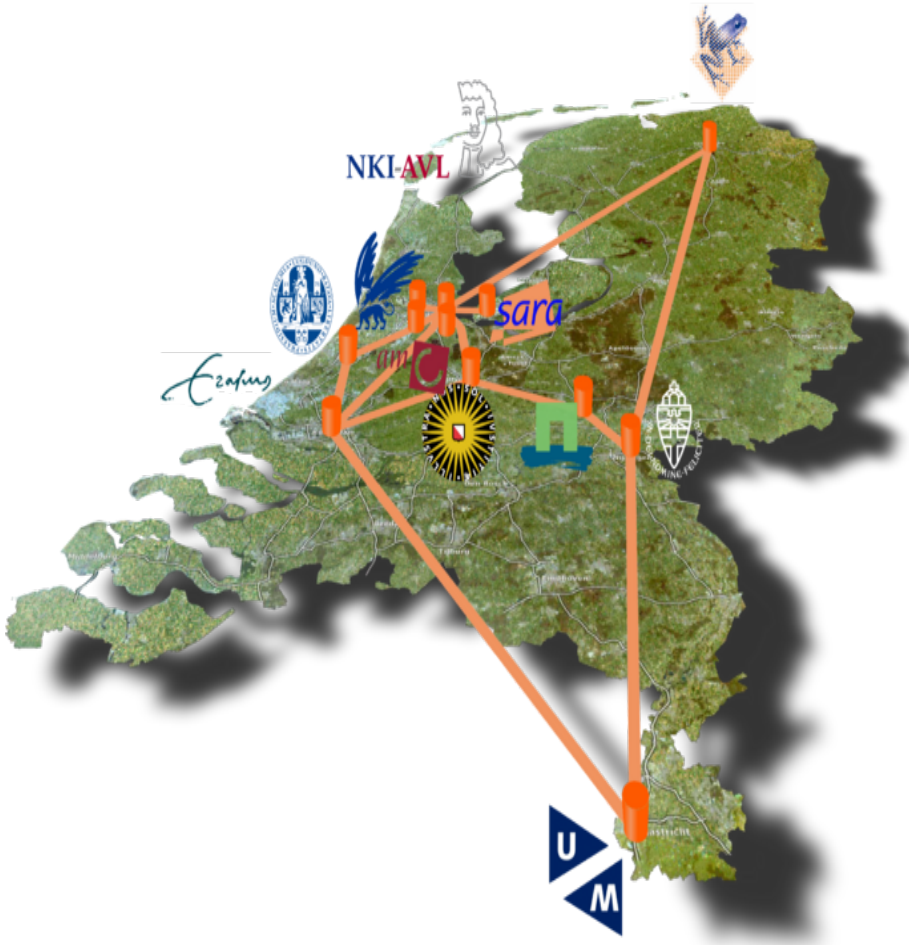
SURF SARA

Life Science Grid

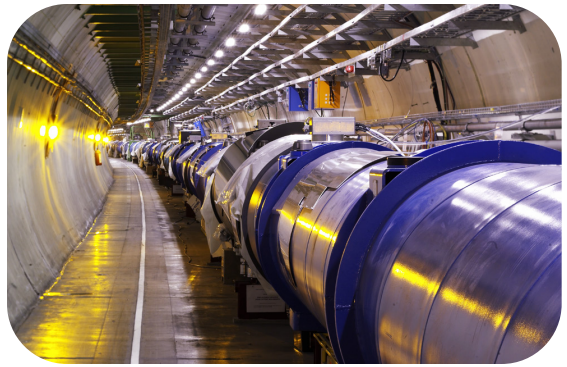
Dutch National Grid

- SURFsara
- Nihkef
- Groningen – Centre for Information Technology

LSG_BCBR	Utrecht
LSG_AMC	Amsterdam
LSG_AMS	Amsterdam
LSG_EMC	Rotterdam
LSG_KUN	Nijmegen
LSG_LUMC	Leiden
LSG_RUG	Groningen
LSG_TUD	Delft
LSG_UM	Maastricht
LSG_VU	Amsterdam
LSG_WUR	Wageningen

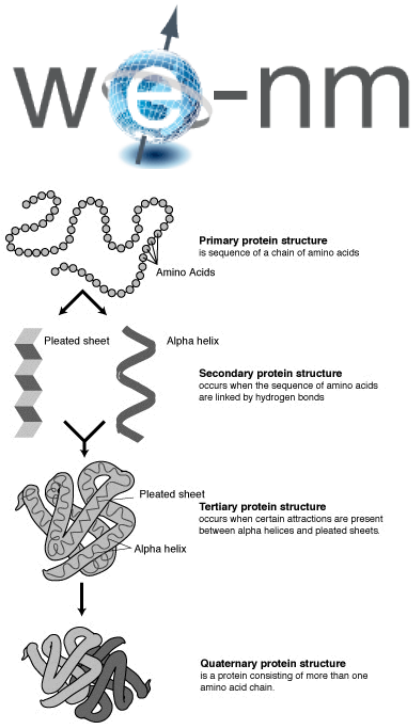


So who uses the Grid?



Large Hadron Collider

LOFAR



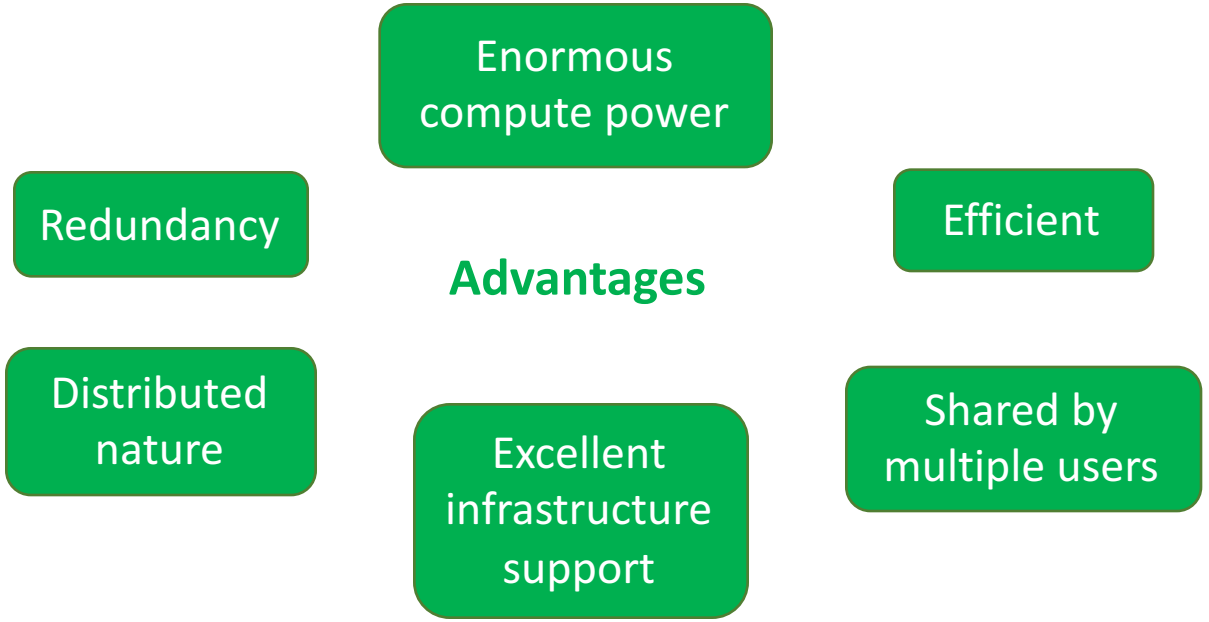
BBMRI

GoNL

project MinE
Make it yours

When to use the cluster/grid?

- **Embarrassingly parallel applications**: It is best suited for your problem when it is relatively easy to split it up in multiple, independent tasks (jobs).
- **Large scale computational problems**: Complex problems that require many services/resources in order to reduce computation time
- **Large scale data processing**: access large databases and collaborate with other users



Access the local cluster

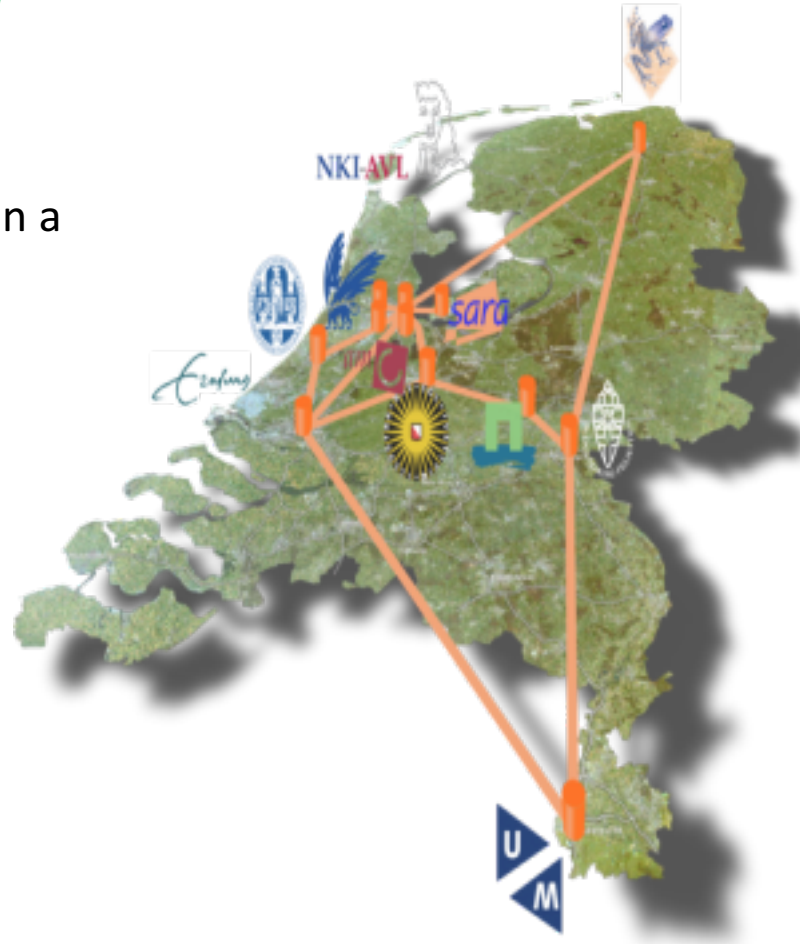
You are going to login to one of the clusters of LSG located in Nijmegen

How: Using an **ssh** client

- is a program for logging into a remote machine and for executing commands on a remote machine.
- is intended to provide secure communications

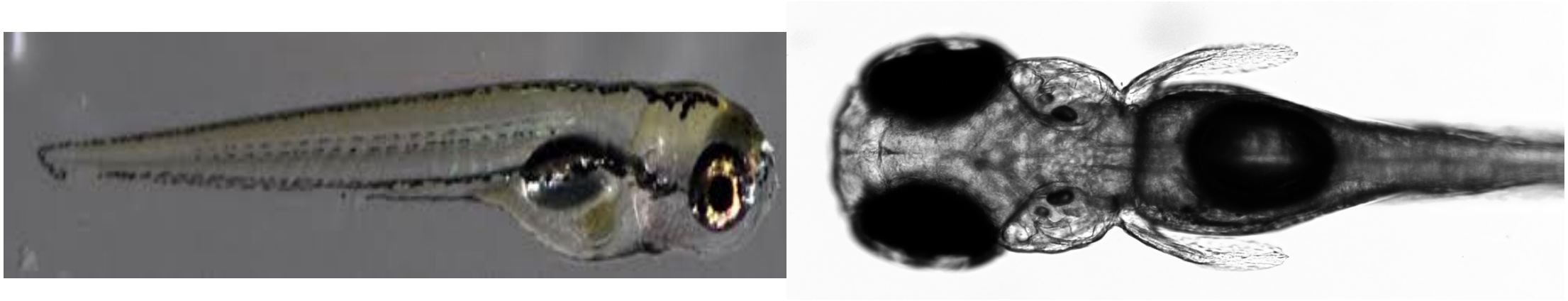
Where: From your web browser go to the link shown below for instructions:

https://github.com/sara-nl/lsg-course/blob/master/partA.md



Submit your first job

- This example characterize optomotor response of a larval zebrafish induced by the presentation of a moving visual stimulus
- The input data is a map of spatio-temporal responses (an array of numbers)
- The python script performs Principal Component Analysis to visualize the response as a map where color represents shape of the response and the brightness its strength
  (Ref: http://research.janelia.org/zebrafish/index.html)
- It will create an output image of the response



Larval zebrafish (Ref: pretty images off internet)

Submit your first job

Submitting a wrapper script to the cluster

```
#!/bin/sh
#Setting the environment variables
export JAVA_HOME=/usr
export SPARK_HOME=/cvmfs/softdrive.nl/lsg-crse/spark-2.0.1-bin-hadoop2.7
export PYTHONPATH=${SPARK_HOME}/python:$PYTHONPATH
export PATH=/cvmfs/softdrive.nl/lsg-crse/pbs-course-env/bin:$PATH
source activate course

# Printing node information
echo " The worker node running the script is:" ${HOSTNAME}
echo " The job is being run in the temporary directory:" ${TMPDIR}
echo " The job was submitted from the directory:" ${PBS_O_WORKDIR}

# Running zebrafish script
# For better performance, copy the script and execute it in the temporary directory
cd ${TMPDIR}
cp ${PBS_O_WORKDIR}/runfish/* .
# Create a random number in range {1 .. 10} as input argument
x=$(( ( RANDOM % 10 )  + 1 ))

# Run the actual script to calculate the principal components. The first argument is fixed and the second argument is random.
python runFish.py 1 ${x}
# Exit the program in case of errors
if [[ "$?" != "0" ]]; then
   echo "Problem on runtime. Exiting now..."
   exit 1
fi
#Copying the output back to the directory where the job was submitted from
cp -r *.png ${PBS_O_WORKDIR}/
echo "Done!"
exit 0
```

Submit your first job

Go to the website below for instructions to submit the job:

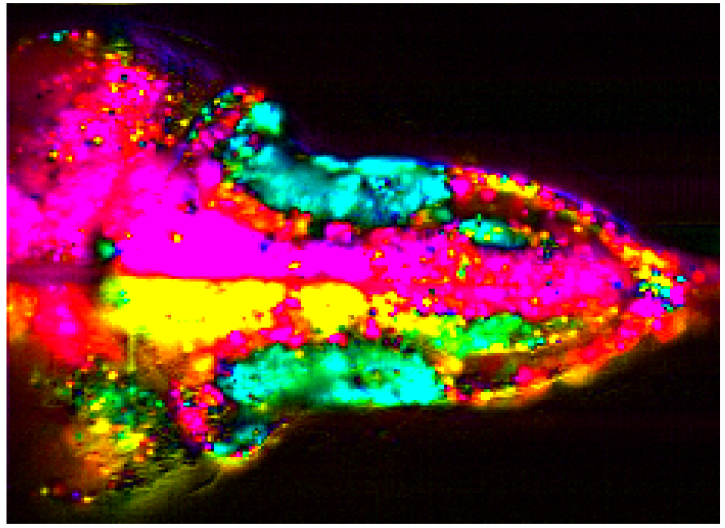
https://github.com/sara-nl/lsg-course/blob/master/partB.md

Recap:

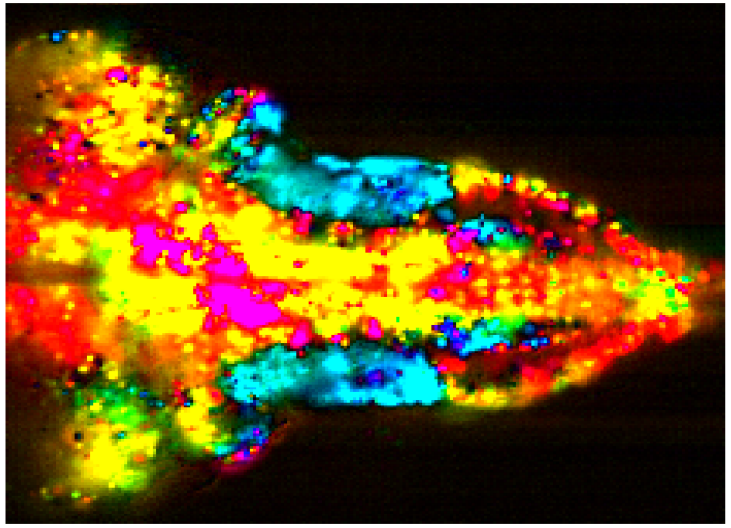
- ✓ Logged into the user interface of kun cluster
- ✓ Ran a simple “Hello World script” locally
- ✓ Ran the “Hello world” on the cluster
- ✓ Looked into the wrapper.sh script
- ✓ We submitted a job to the cluster via this script
- ✓ Learned about the various advanced capabilities
- ✓ Once the output was ready, we got to look at a beautiful image

Multiple components can be calculated & for all of you!!!

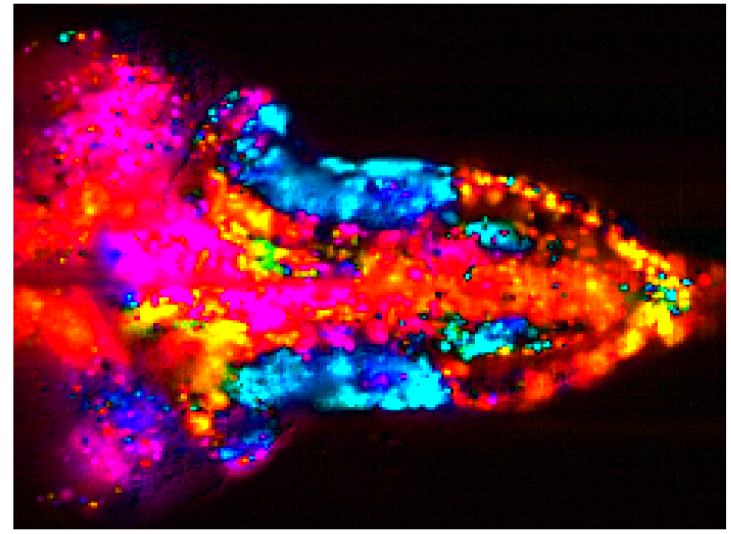
PC 1-2



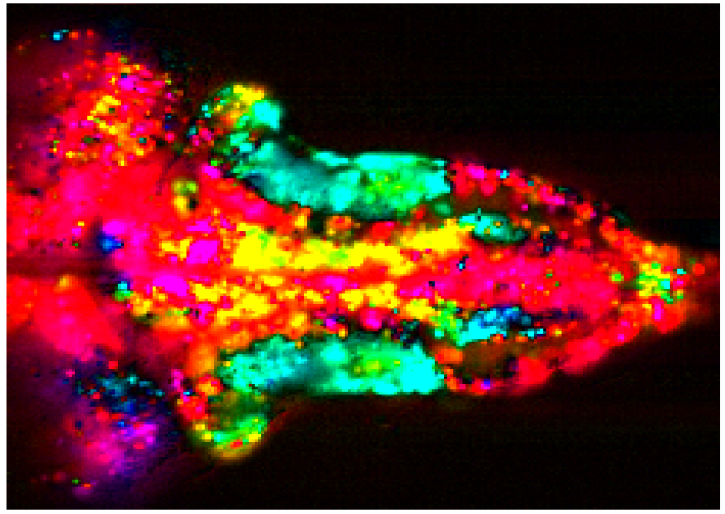
PC 1-3



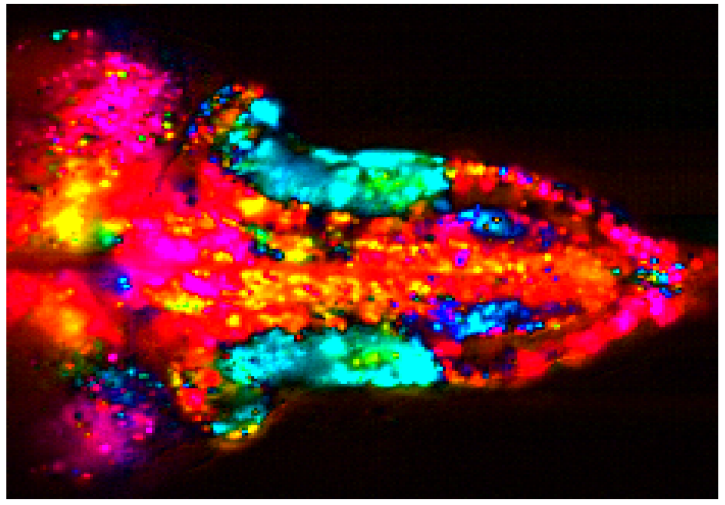
PC 1-4



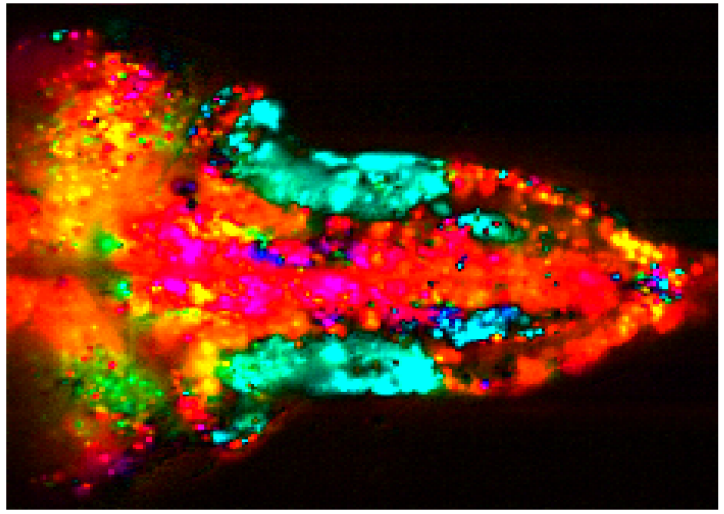
PC 1-5



PC 1-6



PC 1-7



Would you like to use the cluster and/or grid?

Local access to the LSG cluster

- Life science researchers with institutes hosting an LSG cluster are eligible
- To get an account on a local user interface, contact **helpdesk@surfsara.nl**

Access to the LSG

- Researchers at SURF-affiliated institutes can apply at no expense
- For scientists not affiliated to SURF, rates are based on tailor made packages.

Grid facilities

Need help? Contact us at helpdesk@surfsara.nl or +31-208001400