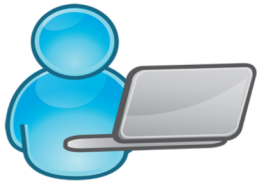
THE COMPUTE CLUSTER *LISA*

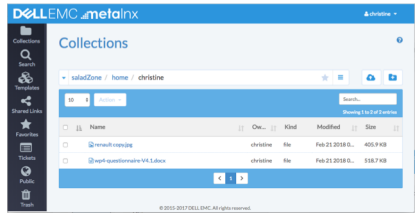
SURF SARA

Training setup – Remote workflow execution

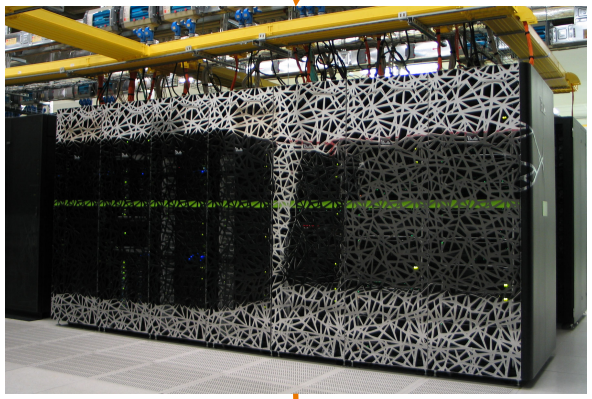


Log in

Return results to user

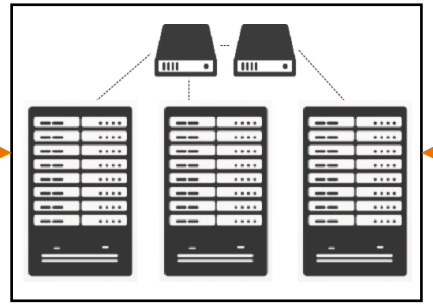


Lisa



iRODS

Compute nodes



Start jobs via PBS

Python script

1. Search for data in iRODS
2. Download data to lisa
3. Some calculations
4. Upload results to iRODS
5. Annotate results

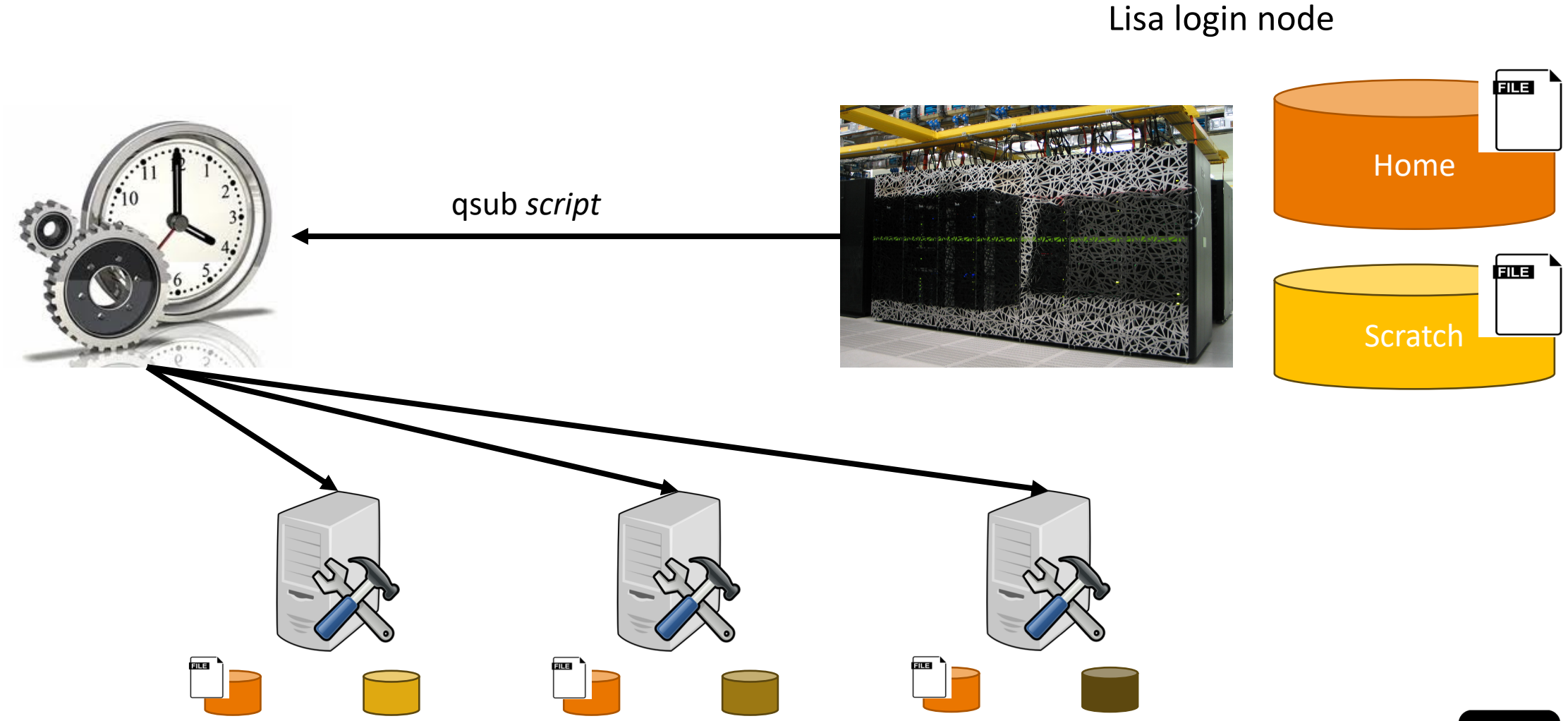
SURF SARA

Supercomputers

- What is a supercomputer
  - A fast computer (CPU)
  - A large computer (Memory)
  - An expensive computer (Money, man power, electricity)
- Why and when do you need a supercomputer
  - If your compute task takes weeks, months or years on a PC
  - If you need more memory than a PC offers

- Lisa versus Cartesius
  - Cartesius: Capability computing; few large-scale jobs
  - Lisa: Capacity computing, many smaller jobs

Lisa – File systems



Workflow 1 - Wordcount

Select cluster nodes

Set environment on compute node

Start workflow

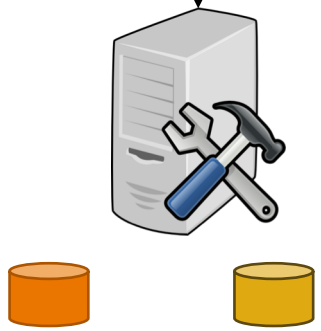
Jobscript

```
#PBS -S /bin/bash
#PBS -lwalltime=00:04:00 -lnodes=1

module load python/2.7.9
cd /home/sdemo110/RDM-Compute-training/iRODS-Compute-Tutorial-Words

python wordsWorkflow.py
```

1. Set parameters to connect to iRODS
2. Search for data in iRODS
3. Start computation
4. Write result data to iRODS



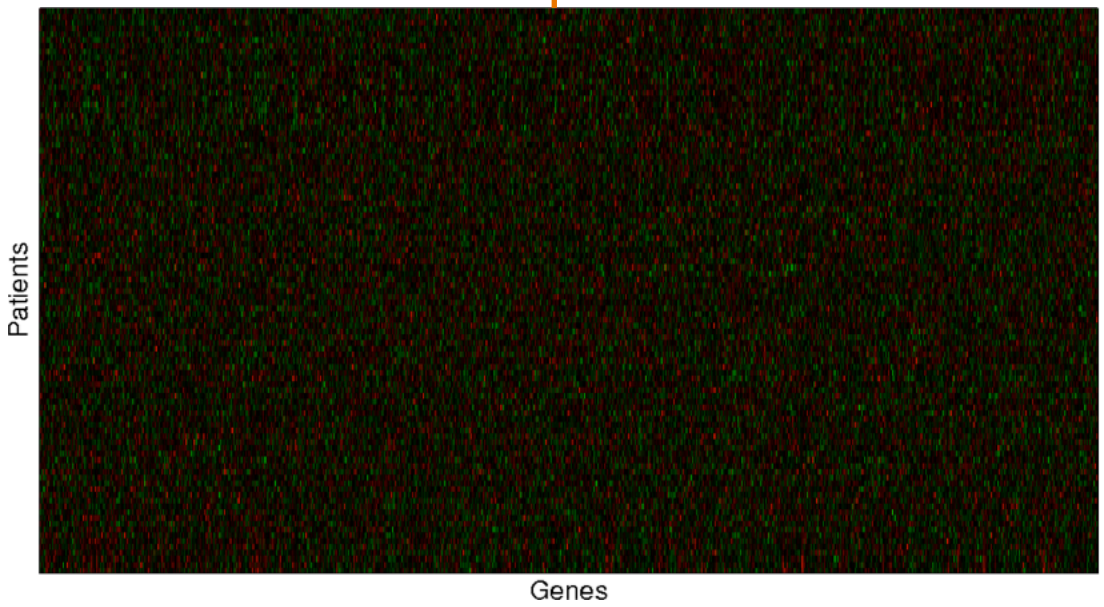
Home Scratch-1

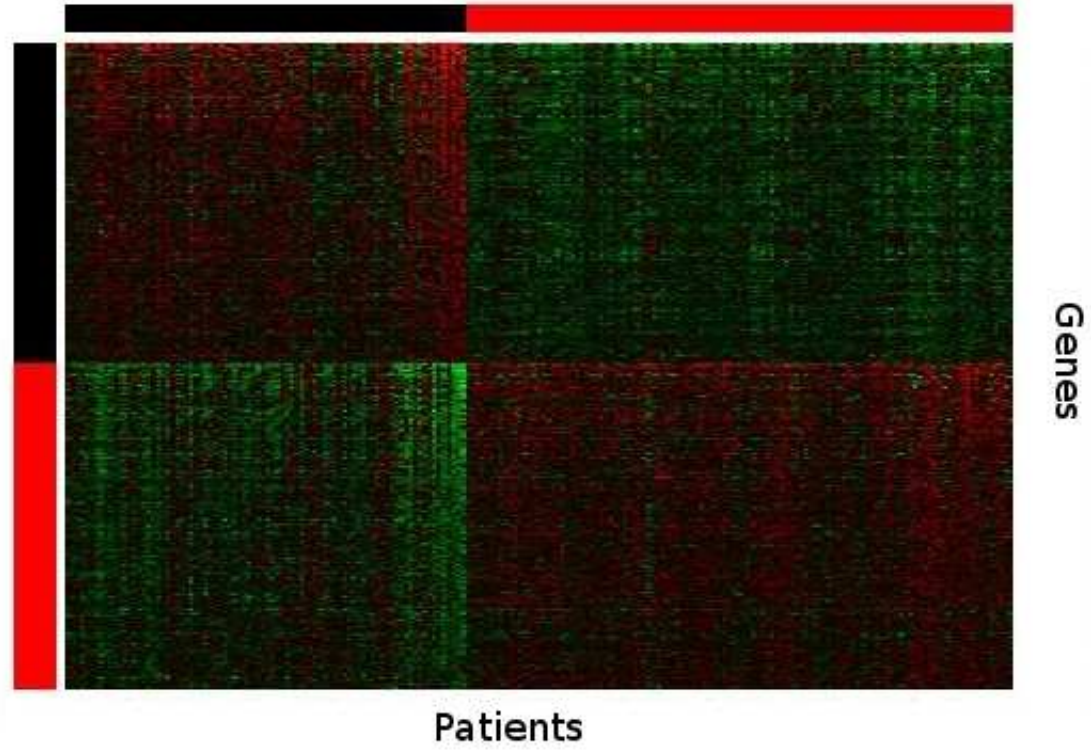
Workflow 2 - ACES

- Compute pipeline to extract features and classify cancer patients
- **Data source 1**: **Gene expression of 1000 breast cancer patients** on Figshare (external repository), classified as good outcome (no relapse) and poor outcome (relapse within 5 years)
- **Data source 2**: Generic **gene interaction networks** and **pathways,** grouping genes into functional clusters, archived in the iRODS instance
- **Task:** Find groups of genes that can discriminate well between the two patient classes and validate how well they perform in prediction.
- **Means:**
    - **Feature extraction** by combining the two data sources
    - **Cross validation** for performance evaluation

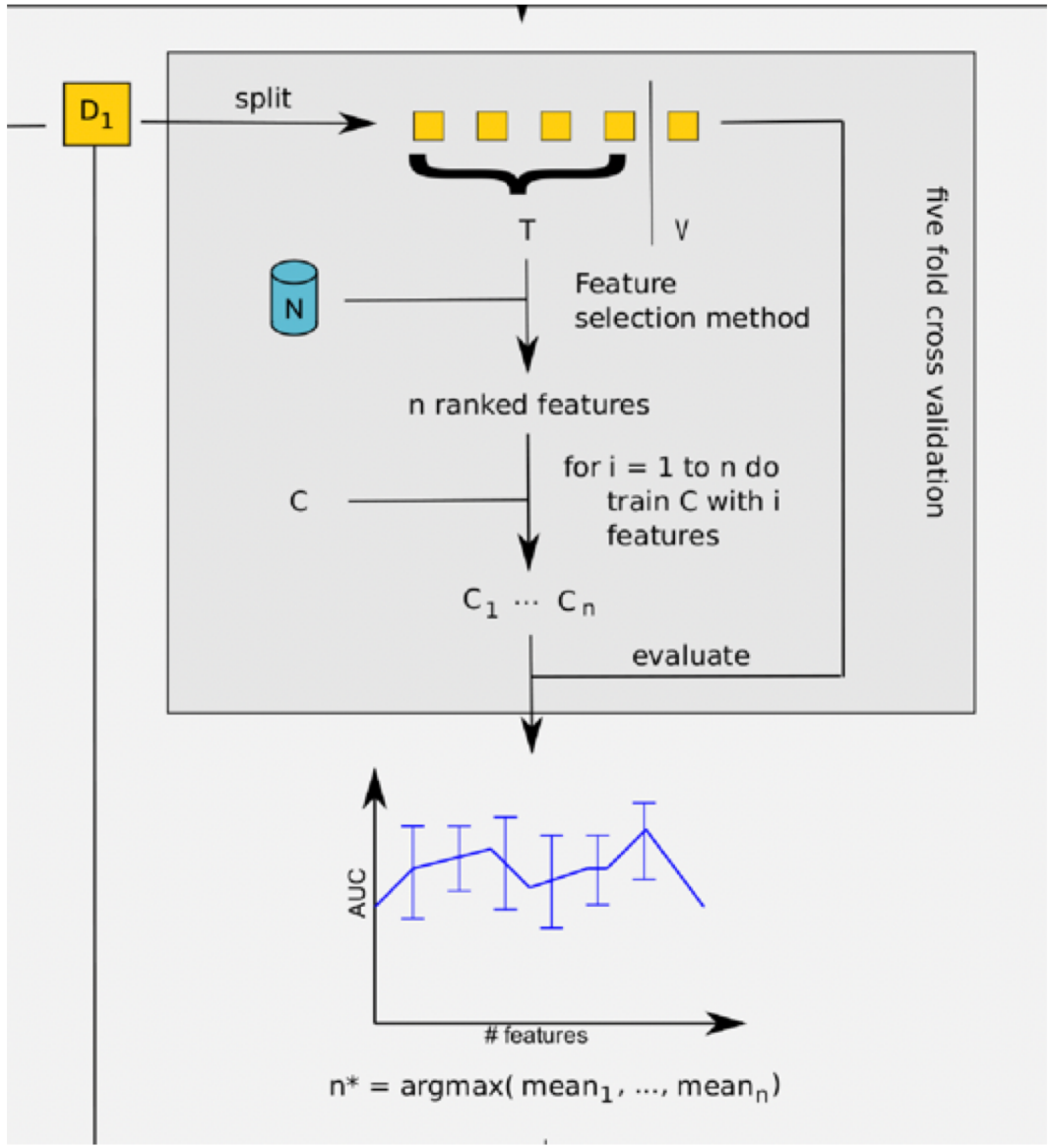
Workflow 2 - Data

Extract groups of genes which discriminate between patients





Workflow 2 – Measuring performance



- 5-fold cross validation
- Measure: area under the receiver operator curve (AUC); true positives vs false positives.

- References:

https://doi.org/10.1371/journal.pone.0034796

https://dx.doi.org/10.3389%2Ffgene.2013.00289

Workflow 2 - ACES

Home Scratch-1

Jobscript

Select cluster nodes

Set environment on compute node

Start workflow

```
#PBS -S /bin/bash
#PBS -lwalltime=00:04:00 -lnodes=1

module load python/2.7.9
cd /home/sdemo110/RDM-Compute-training/iRODS-Compute-Tutorial-Words

Set iRODS parameters
Download dataset from FigShare
                10.6084/m9.figshare.3119248

python acesWorkflow.py
```

1. Search for data in iRODS
2. Start computation
3. Write result data to iRODS

Material:

**git clone **
**https://github.com/sara-nl/iRODS-RDM-Compute-training **
-b SURF-research-bootcamp

Arthur Newton, Christine Staiger (SURFsara)
Arthur.newton(at)surfsara.nl, Christine.staiger(at)surfsara.nl

SURF SARA