



---

# Using clouds and VMs in bioinformatics training in Sweden



The background features a large, abstract graphic composed of numerous thin, curved lines in shades of green, blue, and orange, forming organic, flowing shapes across the slide.

Ola Spjuth <[ola.spjuth@farmbio.uu.se](mailto:ola.spjuth@farmbio.uu.se)>  
Science for Life Laboratory  
Uppsala University



# Bioinformatics Compute and Storage

[www.scilifelab.se/facilities/uppnex/](http://www.scilifelab.se/facilities/uppnex/)

SciLifeLab

*High-performance resources for  
data-intensive biological research*

**Managing access to high-performance  
computing and large scale storage  
dedicated for bioinformatics**

- Computer clusters (8928 cores in total)
- >7 Petabytes of storage
- Backup
- >150 bioinformatics software packages
- Reference genomes, data collections
- Technical user support and training

Free of charge for all users

*Director*



Jukka Komminaho  
(acting)

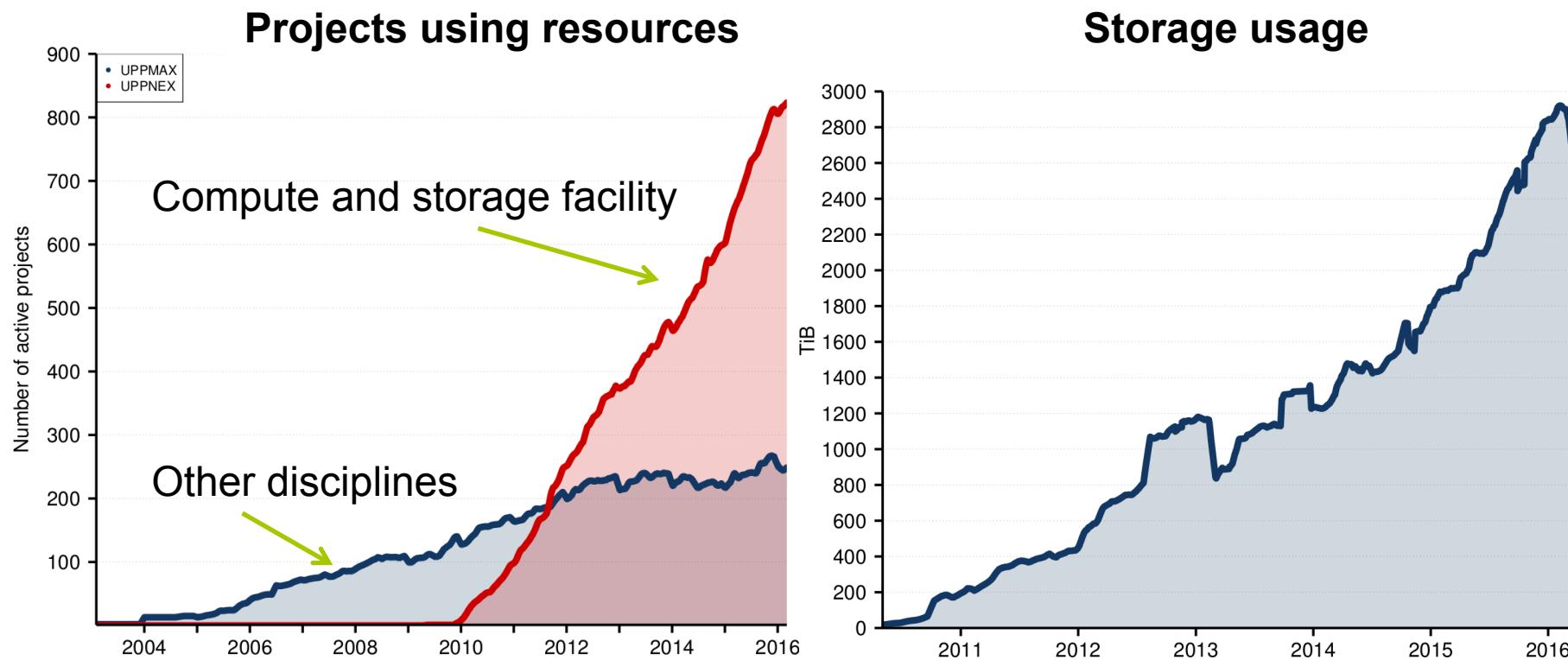
*Head of Facility*



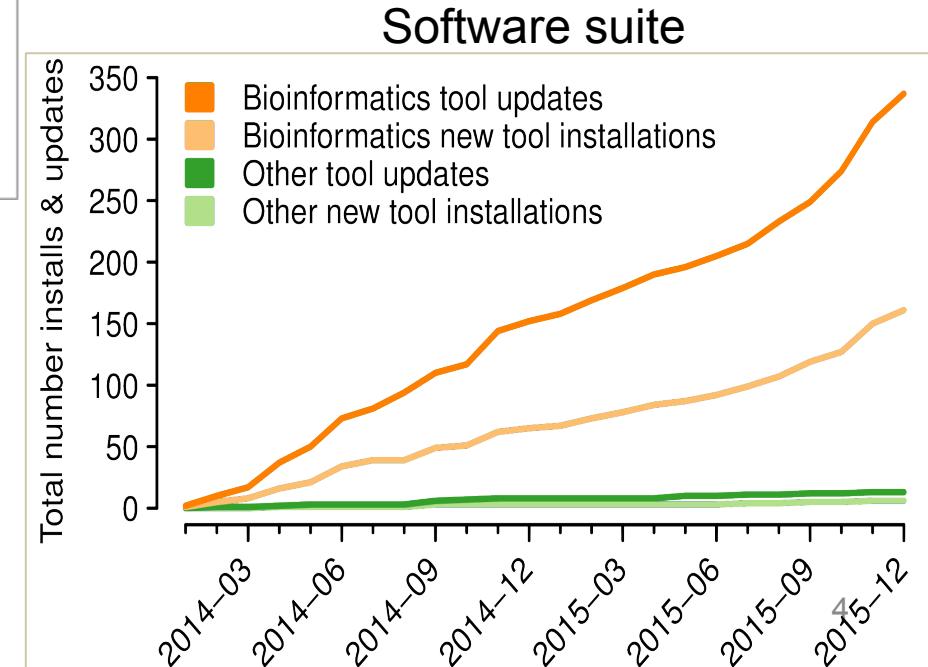
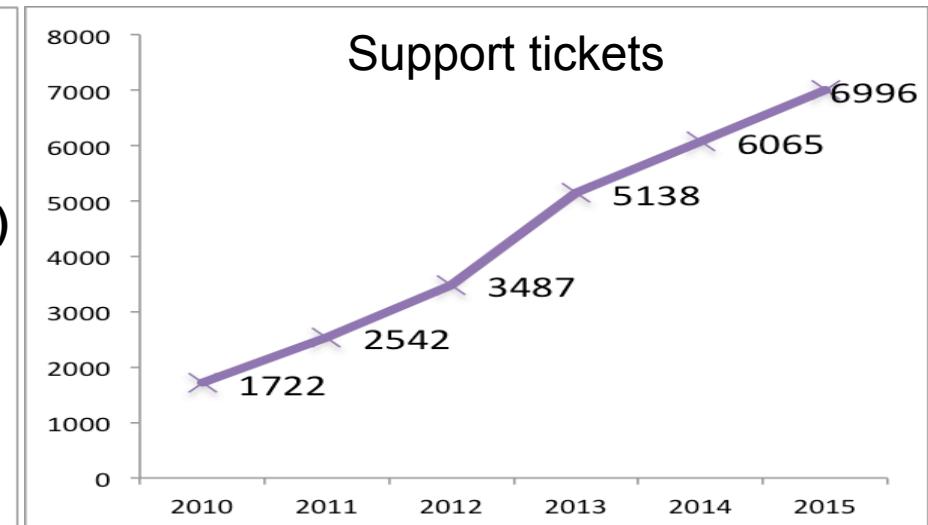
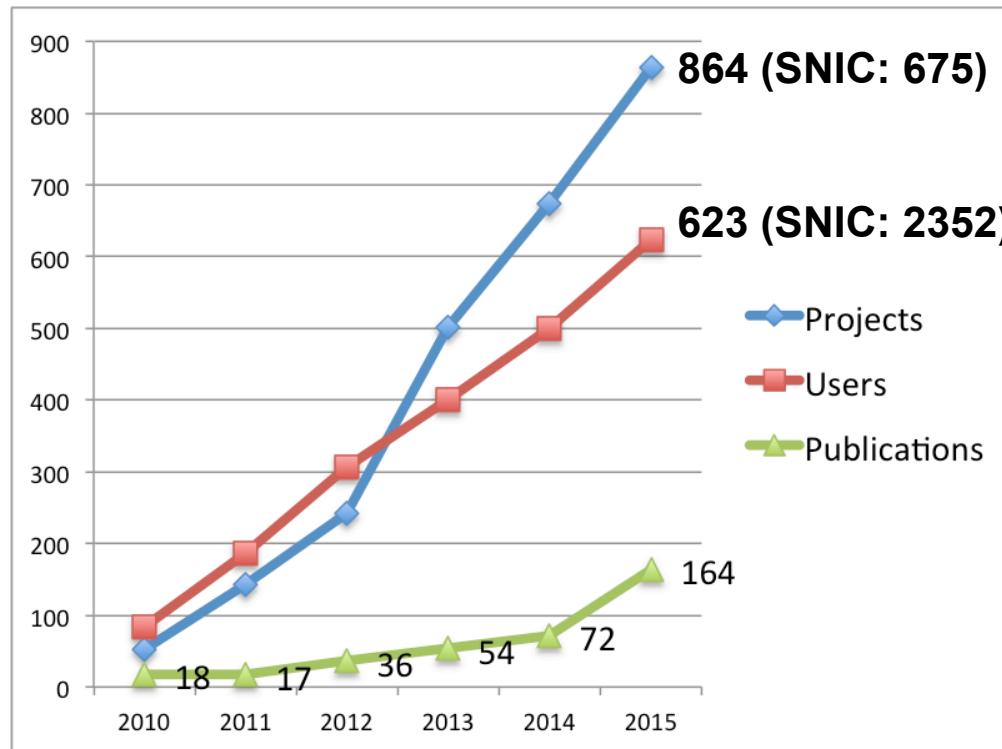
Ola Spjuth

**Specific needs for data-  
intensive bioinformatics:**

- Rapid access to high-memory compute cluster
- Large scale storage
- The latest bioinformatics software available



# KPIs-2

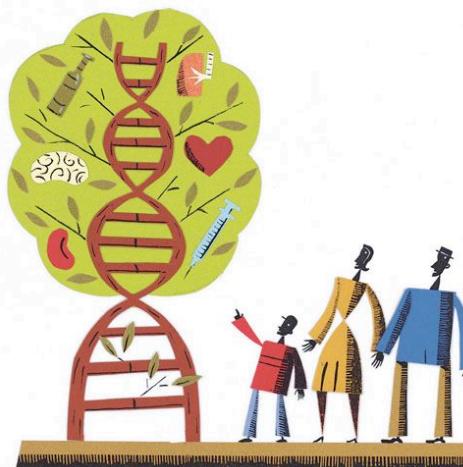


**Systems for analysis of sensitive data is highly prioritized at the facility**

- a large part of data-intensive bioinformatics involves sensitive data

## Initiatives:

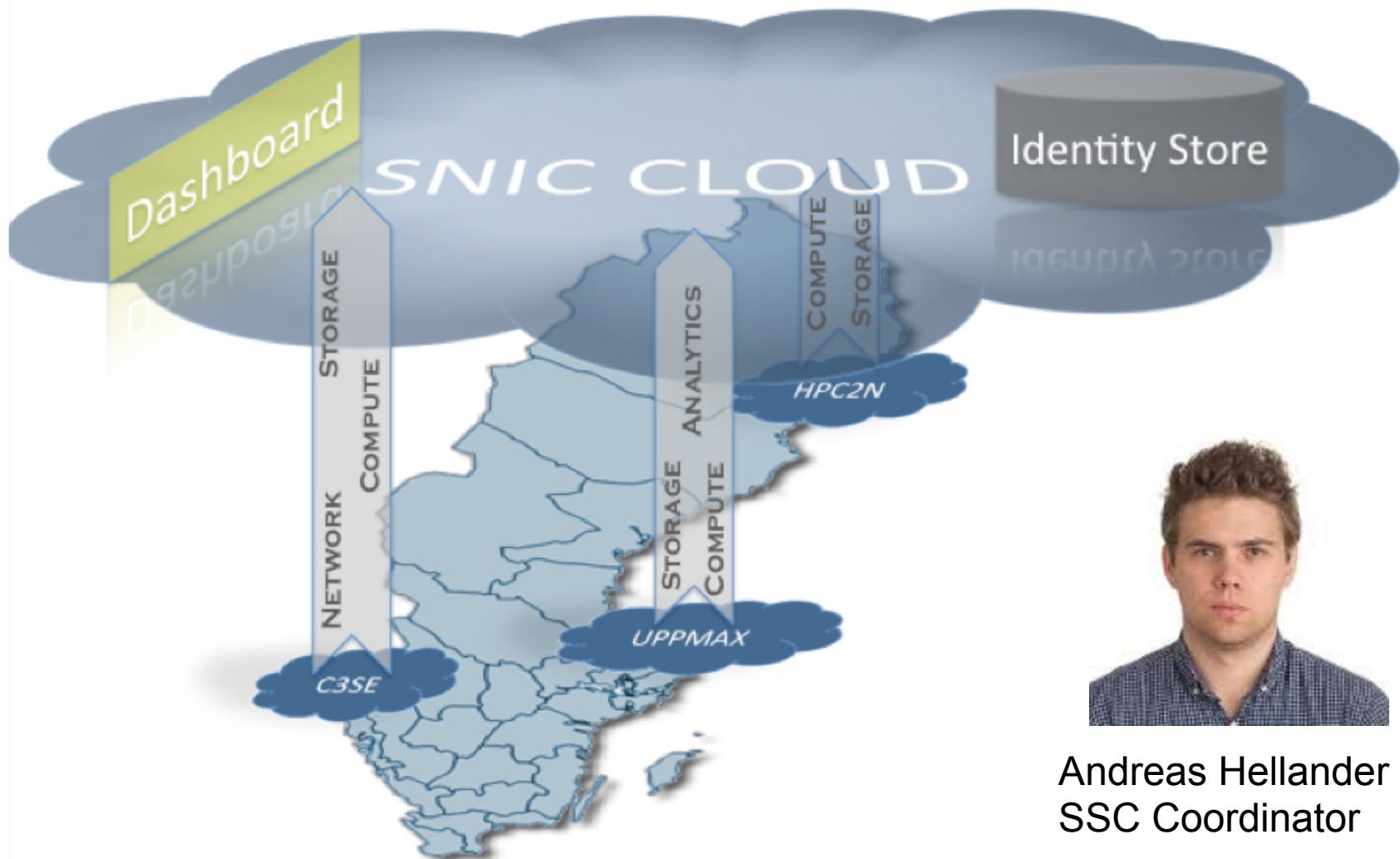
- Interact with Clinical Diagnostics facility
- Mosler pilot since 2013, project for scaling with CSC (FI) ongoing
- Facility part of SNIC-Sens national initiative, systems at UPPMAX
- Pilot with Google to sort out legal issues using public clouds



# Swedish Science Cloud (SSC): A National-Scale IaaS

SciLifeLab

<https://cloud.snic.se/>



Currently **35 projects** and **>100 users** (UU, UmU, GU, Chalmers, KTH, SLU, SciLife, BILS), with different levels of activity

- Life Science dominates, but we also have Comp. chemistry,social science,IoT, Math and more.

Today SSC is used e.g. to:

- Develop SaaS (both PIs and RIs)
- Explore e.g. Apache Spark, Mesos, Docker, Kubernetes
- Chipster (<https://cloud.snic.se/index.php/chipster/>)
- Galaxy (<https://cloud.snic.se/index.php/galaxy/>)

# Selected bioinformatics-related education on SSC

---

- Uppsala University
  - Large Datasets for Scientific Applications (M.Sc)
  - Applied Cloud Computing (M.Sc)
- Karolinska Institutet
  - Gene Expression course using Chipster (Penny Nymark, 2015, 14 students)
- Uppmax workshops
  - Chipster tutorial (Eija Korpelainen, 2015, 20 participants)
  - Microservices (PhenoMeNal-H2020, 2016, 18 participants)

# Undergraduate/graduate courses

---

- **Large Datasets for Scientific Computing (UU)**
  - The data management and data processing part of Data Science
  - *Three labs:*
    - Introduction to cloud computing/Hadoop basics
    - Hadoop/MapReduce (batch analysis of Twitter data)
    - Spark, basic MLlib examples (accessed through Notebooks, Python, Scala kernels)
- *Cloud motivation:*
  - students to be able to work on their own, small, virtual private clusters to allow experimentation.
  - Teachers need to provide multiple technologies, it is not feasible to have to go to IT-support for all of this.

- **Uppsala University**
  - Pharmaceutical Bioinformatics
  - Applied Pharmaceutical Bioinformatics
  - Applied Structural Pharmaceutical Bioinformatics
- Cloud motivation
  - Providing tools and data as VMI easier than support users to install tools (however, need to support users on e.g. virtualbox and linux)

- Courses that run on HPC:
  - Cloud computing: Backup for scheduled or unscheduled maintenance of HPC
  - Would allow students to practice before/after course ends
  - SciLifeLab in Sweden interested for national bioinformatics courses

- **Jupyter (previously iPython) notebooks**
  - Human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis
  - Many different kernels (Python, Perl, Java, C, R, Matlab, Hadoop, Spark etc.)

spectrogram - Iceweasel

localhost:8889/notebooks/spectrogram.ipynb

IP[y]: Notebook spectrogram Last Checkpoint: a few seconds ago (autosaved)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

## Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#) using windowing, to reveal the frequency content of a sound signal.

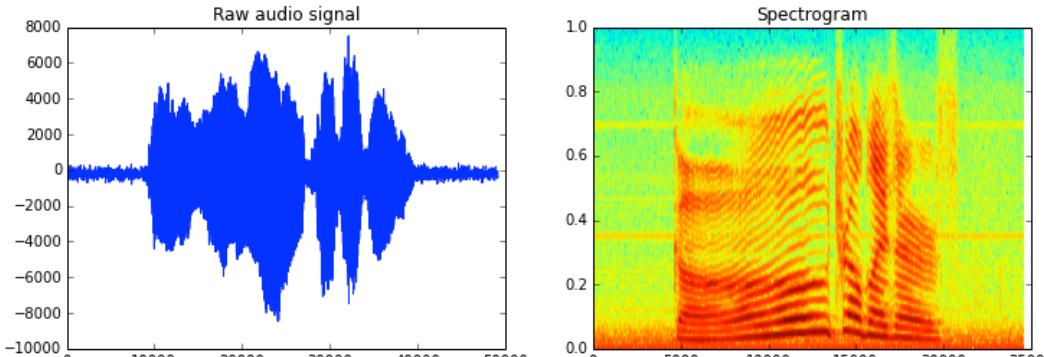
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N-1$$

We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile  
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [2]: %matplotlib inline  
from matplotlib import pyplot as plt  
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))  
ax1.plot(x); ax1.set_title('Raw audio signal')  
ax2.specgram(x); ax2.set_title('Spectrogram')
```



- Several rich online resources exist

The screenshot shows a Jupyter Notebook interface with a browser tab for 'readiab.org/book/latest/2/4' open. The notebook cell In [11] contains Python code to import ete3 and define a TreeStyle:

```
In [11]: import ete3
ts = ete3.TreeStyle()
ts.show_leaf_name = True
ts.scale = 250
ts.branch_vertical_margin = 15
```

The notebook cell Out[12] shows the resulting phylogenetic tree. The tree has four main branches. The leftmost branch leads to two leaves labeled 'aaaaaaaaaa' and 'aaaaaaaaab'. The middle-left branch leads to two leaves labeled 'aaaaaaaaac' and 'aaaaaaaaad'. The middle-right branch leads to two leaves labeled 'aaaaaaaaae' and 'aaaaaaaaaf'. The rightmost branch leads to three leaves labeled 'aaaaaaaaag', 'aaaaaaaaah', and 'aaaaaaaaai'. A scale bar at the bottom indicates a distance of 0.20.

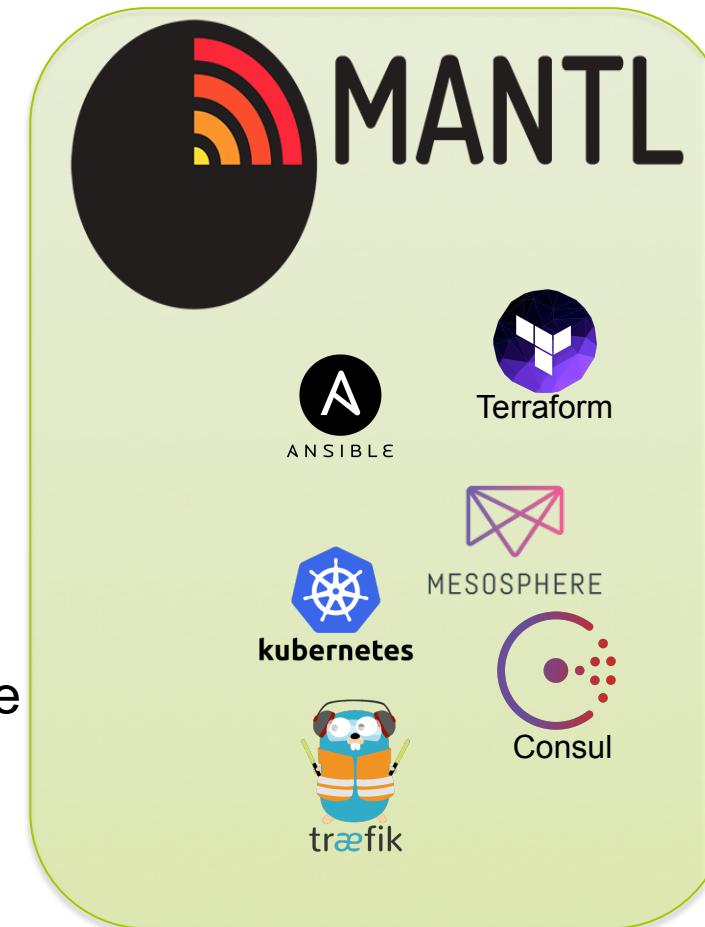
- Use Jupyter to deliver a scripting/programming environment for students at Uppsala University (UU)
- **Large Datasets for Scientific Applications (UU)**
  - Access to SPARK environment for Big Data analytics
    - Students handed in assignments as notebooks
- **Introduction to R and Python (UU)**
  - We plan to use Jupyter to offer computing environments to students
  - Investigating how assignments best should be sent in

# Provisioning virtual infrastructures

## MANTL (<https://mantl.io/>) by Cisco Cloud

A modern platform for rapidly deploying globally distributed services

- Defines the infrastructure (compute nodes, storage, networks, DNS, firewall etc)
- Docker orchestration through Mesosphere and Kubernetes
- High-availability, scalability and fault tolerance
- Provides a way to deploy on your favourite cloud provider



- Problems in FP7 AllBio:
  - Locating suitable VMIs
  - Publishing VMIs (bandwidth etc.)
- **Biolmg.org** is a catalogue where users can search for or add new bioinformatics VMIs.
  - Also offers a free service to host images on servers with large storage and bandwidth



M. Dahlö, F. Haziza, A. Kallio, E. Korpelainen, E. Bongcam-Rudloff, and O. Spjuth,  
**“Bioimg.org: A catalog of virtual machine images for the life sciences”**  
*Bioinform Biol Insights*, vol. 9, pp. 125–8, 2015.

Screenshot of the Biolmg.org website interface:

The browser address bar shows <https://biolmg.org>. The page title is "Home | Biolmg.org". The top navigation bar includes links for "FirstRow Free Live Sp", "Open in Papers", "webbarbete farmbio", "BiB - drug repurposir", "Conference Alerts", "LaTeX Style and BiBT", "Login / Register", and "Other Bookmarks".

The main content area features the Biolmg.org logo (a yellow star icon) and navigation links for "FLAVOURS", "HOW TO USE THE IMAGES", and "ABOUT". Below this is a table listing various bioinformatics resources:

NAME	DESCRIPTION	TAGS	UPDATED
Appl. Pharm. StructBioinf	Course image for the course Applied Structural Pharmaceutical Bioinformatics, an internet-based course at Uppsala University (5 credits) which is free for all students within EU.	#docking, #modeling, #drug discovery, #pharmaceutical bioinformatics	2014-10-15
ASH	Automated Selection of Hotspots (ASH): enhanced automated segmentation and adaptive step finding for hotspot detection in adrenal cortical cancer. We have implemented an open source automated detection quantitative ranking of hotspots to support histopathologists in selecting the 'hottest' hotspot areas ...	#hotspots, #cancer, #galaxy	2014-12-04
Bioinfo course	that teaches bioinformatics.	#bioinformatics	2014-11-03
Bio-Linux	Bio-Linux 8 is a powerful, free bioinformatics workstation platform that can be installed on anything from a laptop to a large server, or run as a virtual machine. Bio-Linux 8 adds more than 250 bioinformatics packages to an Ubuntu Linux ...	#dna, #rna, #de novo	2015-01-20
	Chipster is a ready to run virtual machine with a comprehensive collection of bioinformatics packages and reference		

- How do others use virtualized environments in teaching?
- Experiences with Jupyter and iPython notebooks in education?
- Course environment in cloud, students can submit assignments etc.

---

# Thank You

**ola.spjuth@farmbio.uu.se**

