SciLifeLab

Using clouds and VMs in bioinformatics training in Sweden

Ola Spjuth <ola.spjuth@farmbio.uu.se>
Science for Life Laboratory
Uppsala University

elixir SWEDEN

UPPSALA UNIVERSITET

Bioinformatics Compute and Storage

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

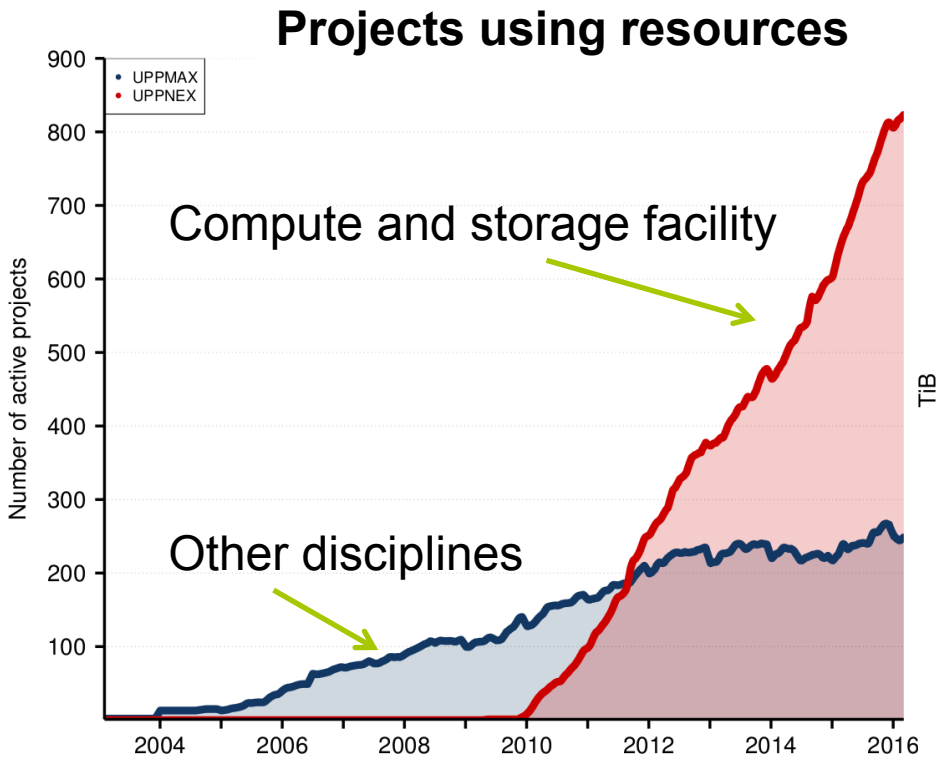
Ola Spjuth

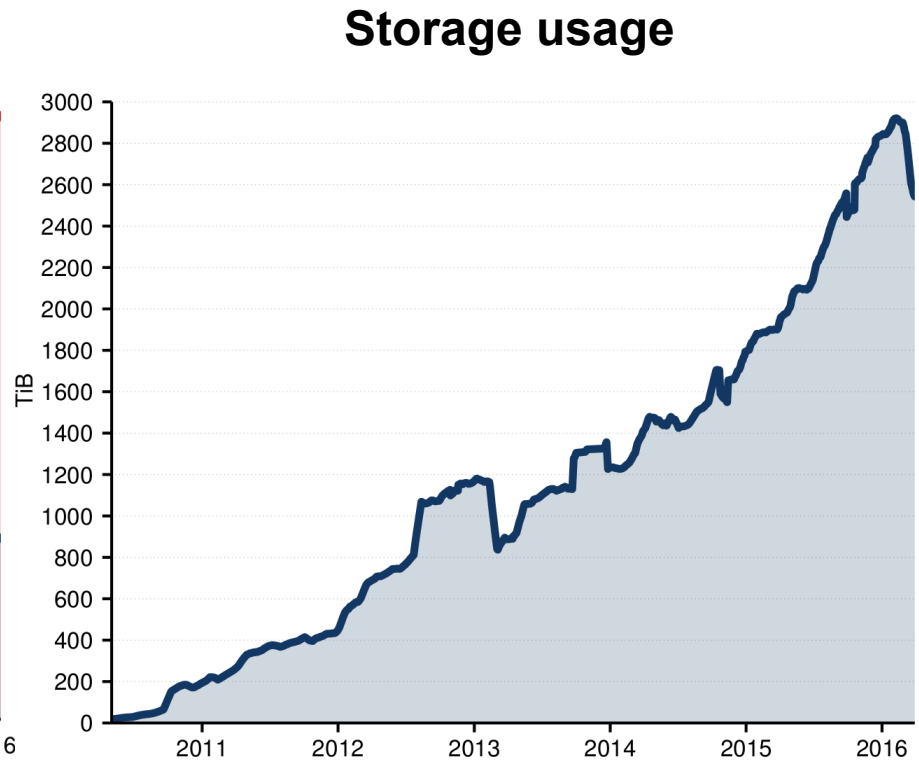
Specific needs for data-intensive bioinformatics:

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

KPIs

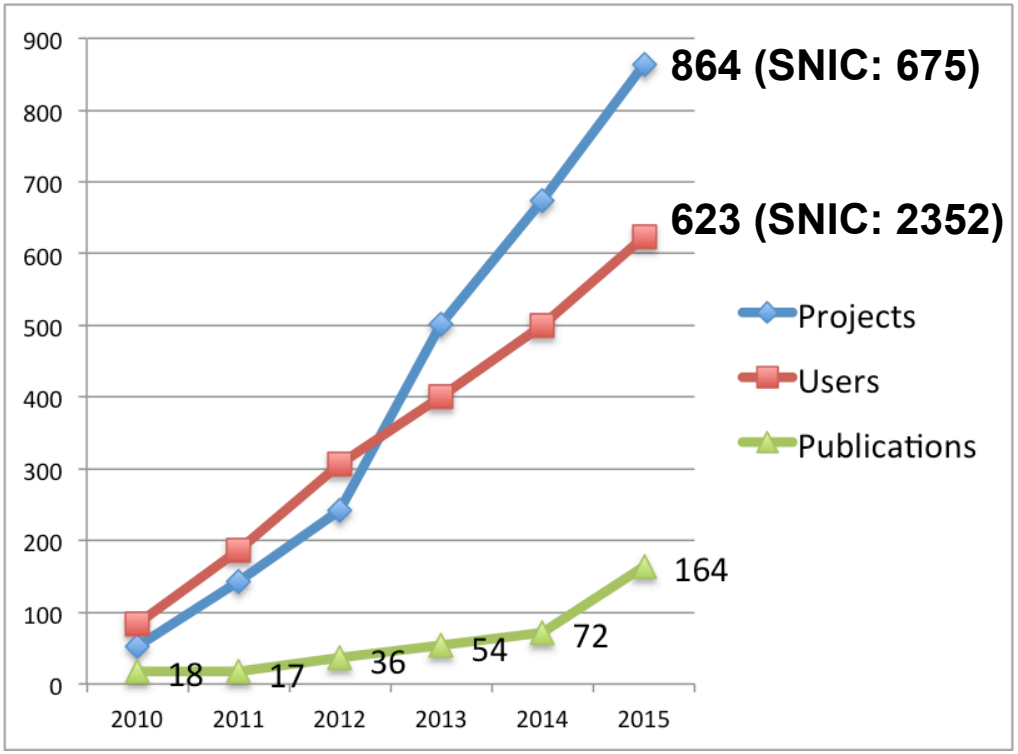
SciLifeLab



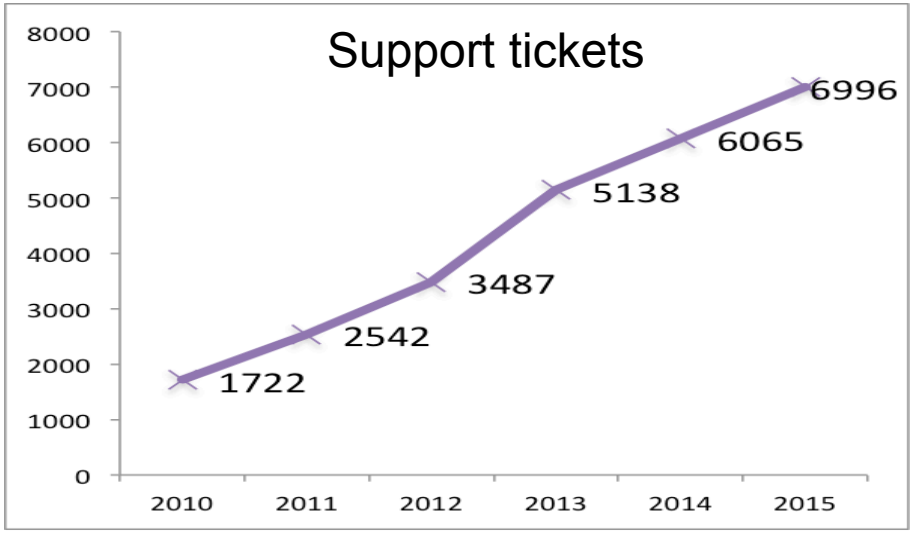


KPIs-2

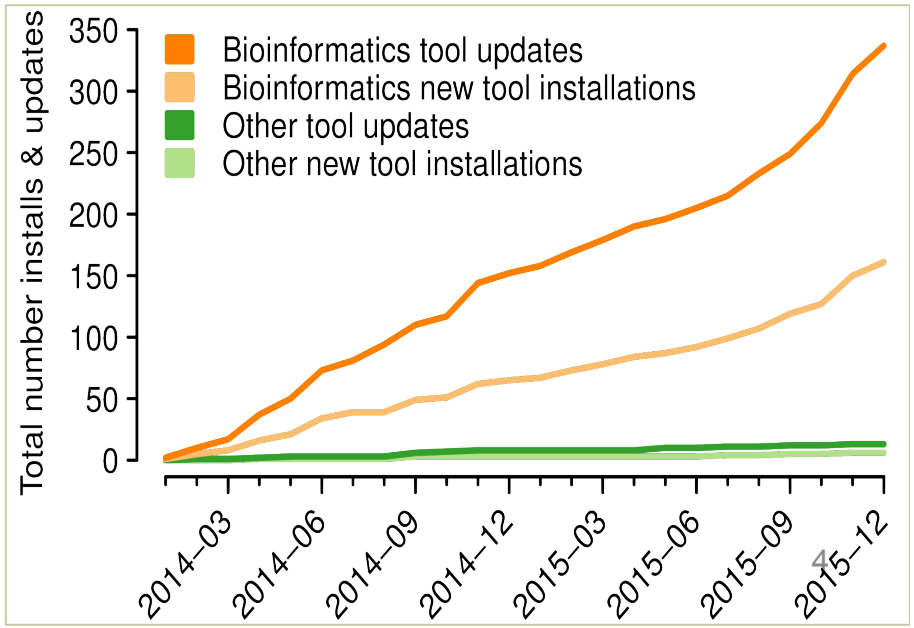
SciLifeLab



FTEs



Software suite



Sensitive data

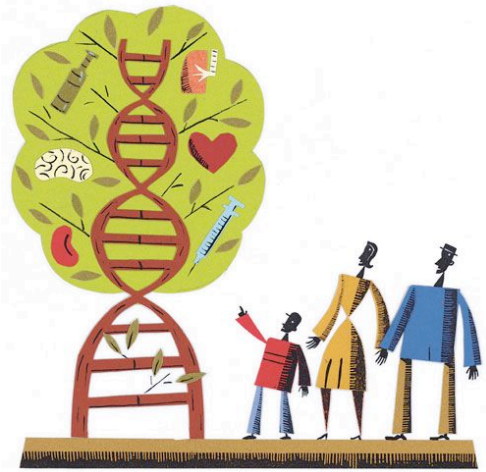
SciLifeLab

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

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

Initiatives:

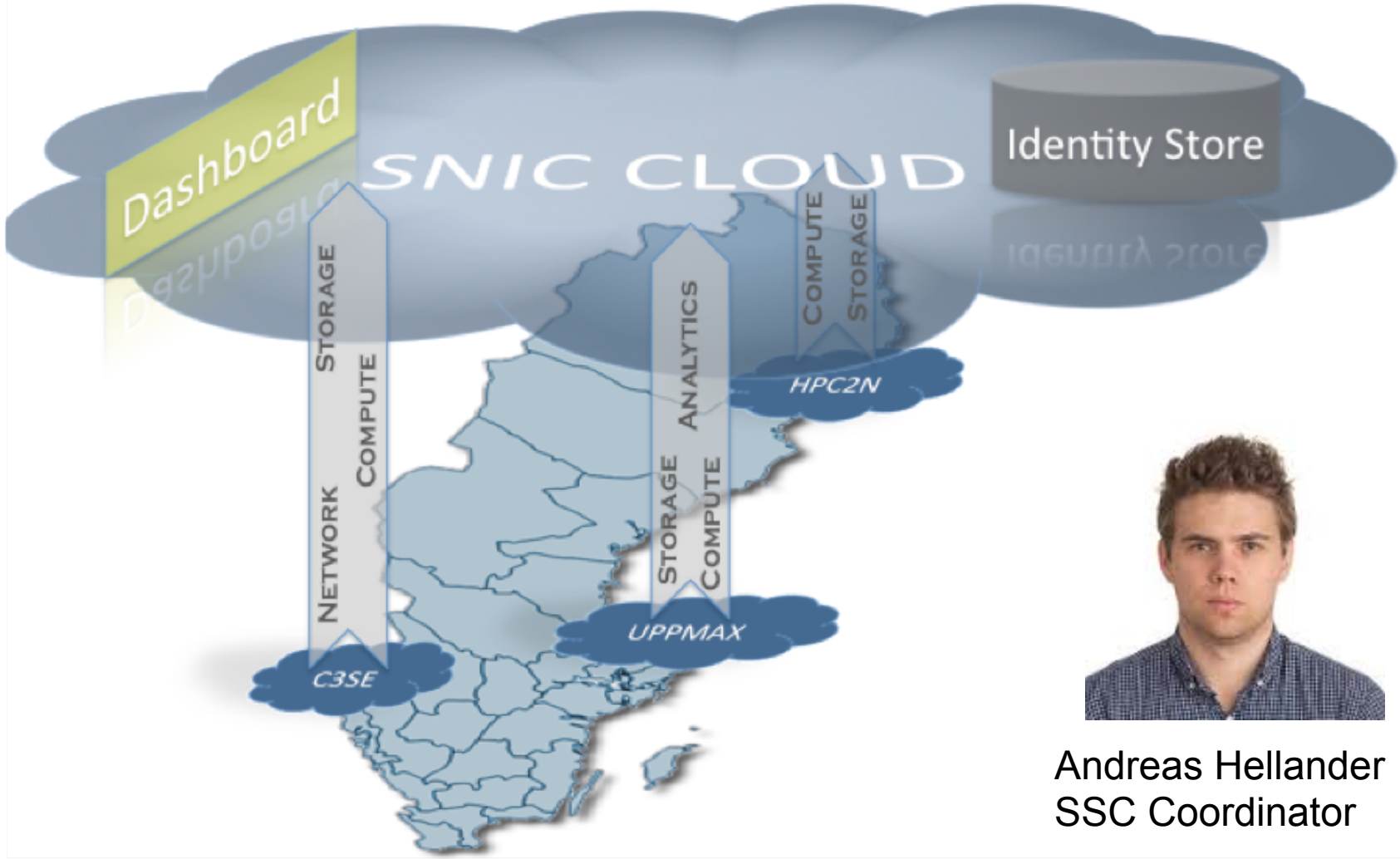
- Interact with Clinical Diagnostics facility
- Mosler pilot since 2013, project for scaling with CSC (FI) ongoing
- Facility part of *SNIC-Sens* national initiative, systems at UPPMAX
- Pilot with Google to sort out legal issues using public clouds



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

SciLifeLab

https://cloud.snic.se/



Andreas Hellander
SSC Coordinator

SSC Users

SciLifeLab

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

SciLifeLab

- 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

SciLifeLab

- **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.

Internet-based courses

SciLifeLab

- **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)

New courses

SciLifeLab

- 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

Online notebooks in training

SciLifeLab

- **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.)

Jupyter

SciLifeLab

spectrogram - Iceweasel

localhost:8889/notebooks/spectrogram.ipynb

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

IPython (Python 3)

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.

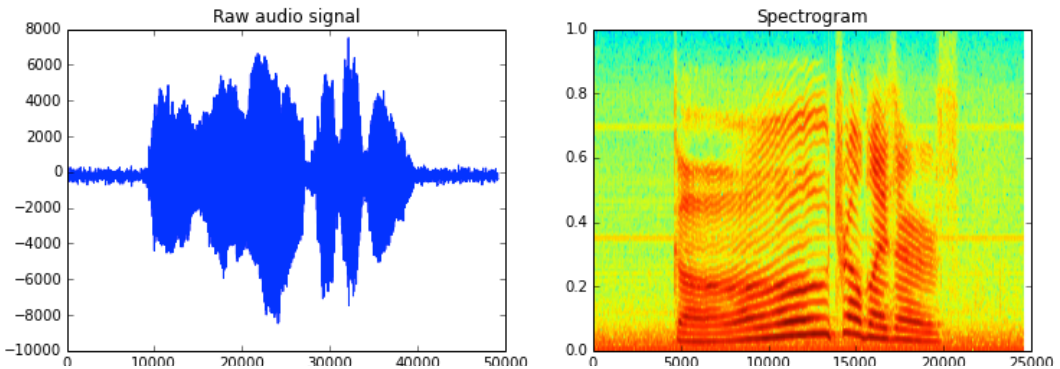
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

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:

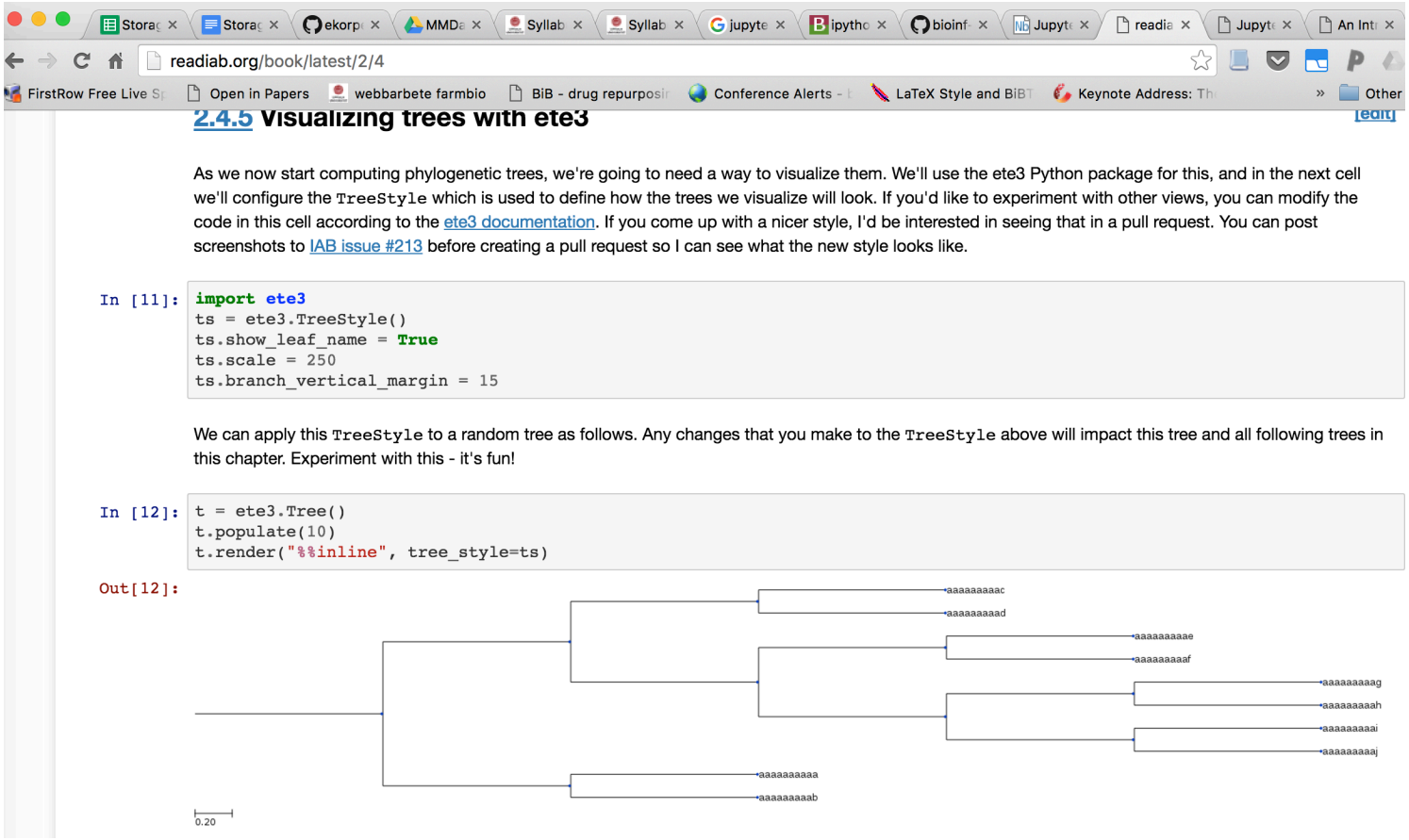
In [2]:
```
%matplotlib inline
from matplotlib import pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



Jupyter in Bioinformatics

SciLifeLab

- Several rich online resources exist



Our approach

SciLifeLab

- 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

SciLifeLab

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

BioImg.org

SciLifeLab

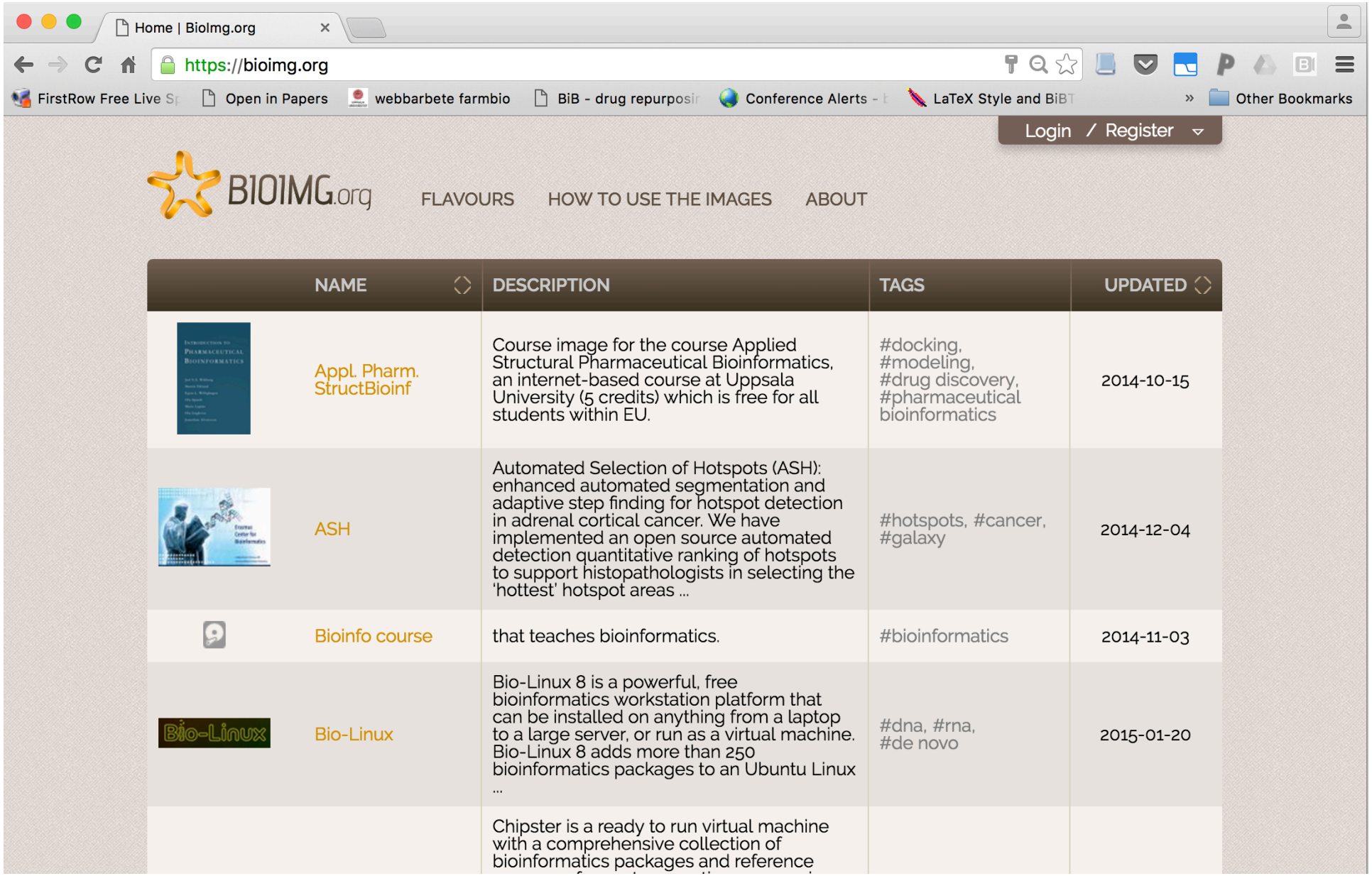
- Problems in FP7 AllBio:
  - Locating suitable VMIs
  - Publishing VMIs (bandwidth etc.)

- **BioImg.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.

BioImg.org

SciLifeLab



Discussion topics

SciLifeLab

- 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.

SciLifeLab

Thank You

ola.spjuth@farmbio.uu.se