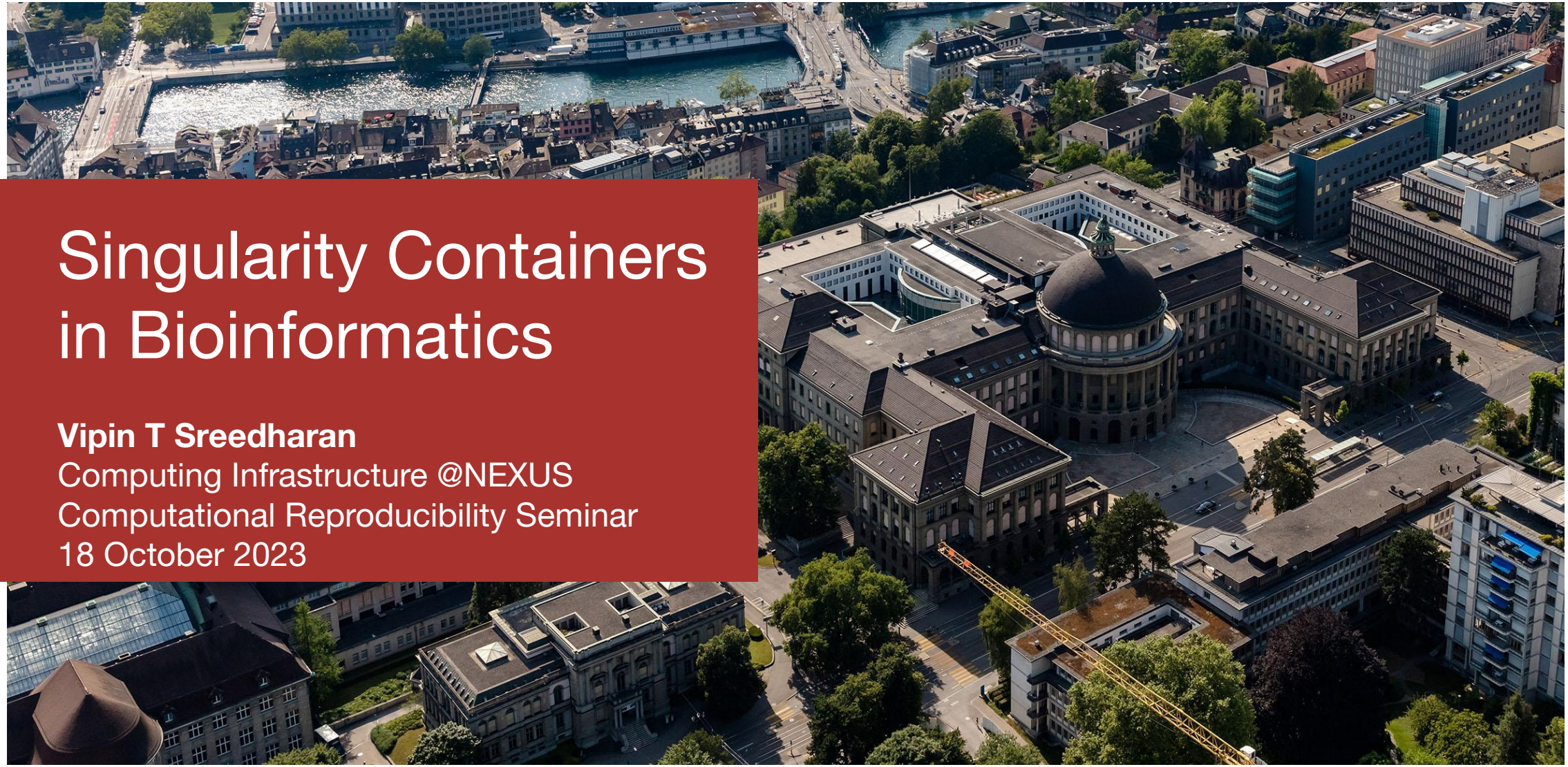
ETH zürich

nexus Personalized Health Technologies



Singularity Containers in Bioinformatics

Vipin T Sreedharan
Computing Infrastructure @NEXUS
Computational Reproducibility Seminar
18 October 2023

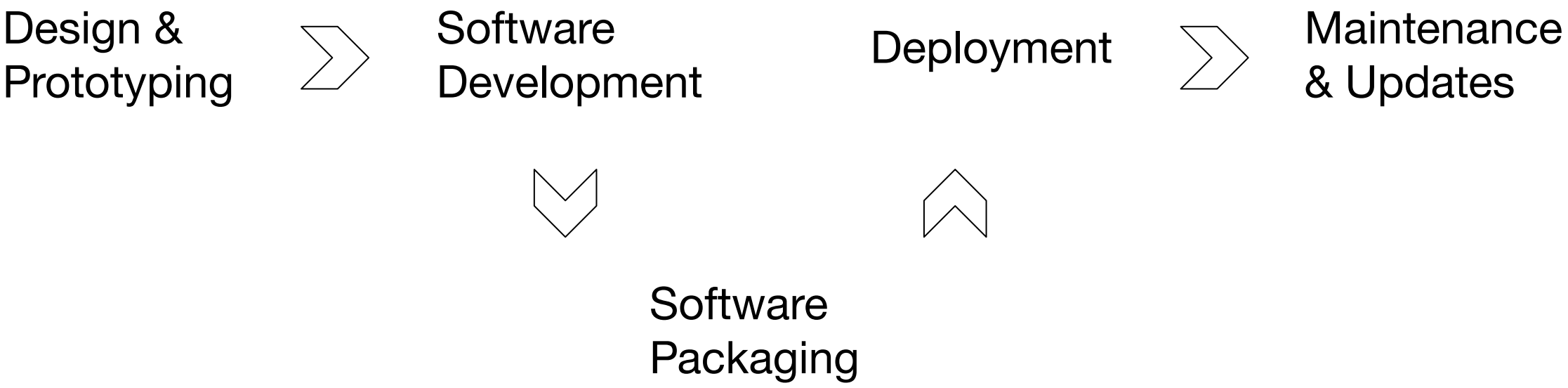
Objectives

- Software packaging
- Galaxy platform
- Container images for bioinformatics tools

Software Deployment Process

Design & Prototyping → Software Development → Deployment → Maintenance & Updates

Software Deployment Process



Software Deployment Process

Software Development → Software Packaging → Deployment

Package Manager

- Package manager performs dependency resolution checks and install
- In 1993, the earliest form of package manager began to appear from Linux
- Some of these early package managers (dpkg, rpm) live on today

Software Deployment before Package Managers

- Historically, software was provided either via FTP or basic websites
- The configure process starts using a C compiler and checks your system for application dependencies
- If the configure script completed successfully, a Makefile would be created

- Once a Makefile existed, you would then proceed to run the make command

- Finally, after the make process has been completed, you would need to run make install in order to actually install the software

Software Deployment Process

Software Development → Software Packaging → Deployment

32 Package Managers

npm
4.01M Packages

Go
472K Packages

CRAN
25.1K Packages

PyPI
481K Packages

CPAN
40.3K Packages

Maven
580K Packages

Language package managers

SUSE zypper

rpm

apt-get

yum yellowdog updater modified

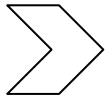
System package managers

https://libraries.io/

Software Deployment Process

Software Development

Software Packaging



Deployment

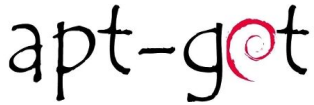
Go
472K Packages

CRAN
25.1K Packages

PyPI
481K Packages

CPAN
40.3K Packages

Maven
580K Packages



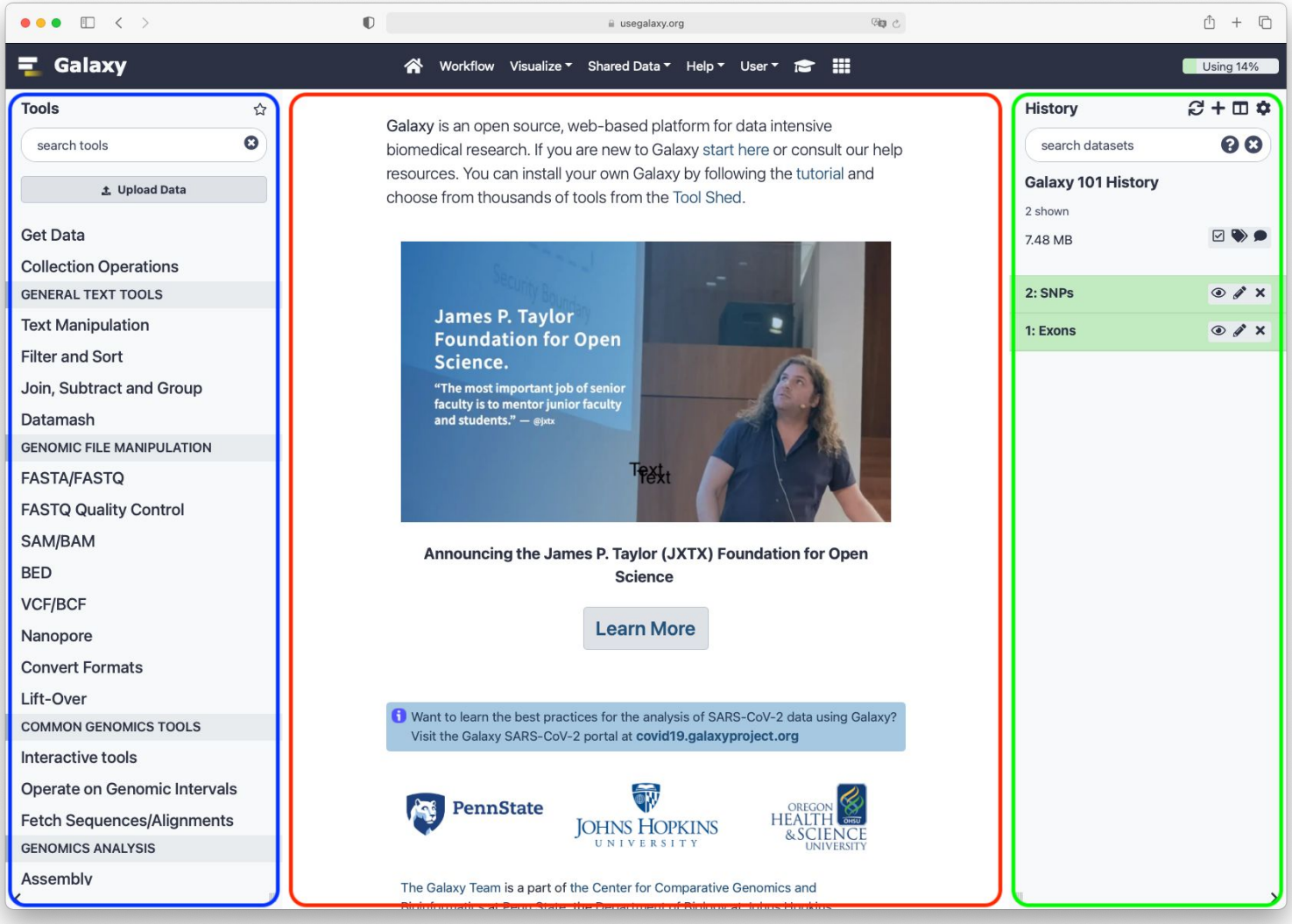
A short introduction to Galaxy platform

Galaxy is an open-source platform for FAIR data analysis that enables users to:

- Use tools from various domains (that can be plugged into workflows) through its graphical web interface
- Run code in interactive environments (RStudio, Jupyter...) along with other tools or workflows
- Manage data by sharing and publishing results, workflows, and visualizations
- Ensure reproducibility by capturing the necessary information to repeat and understand data analyses

https://galaxyproject.org/

Galaxy User Interface



Public servers

Galaxy EUROPE

Galaxy Australia

usegalaxy.org | usegalaxy.eu | usegalaxy.org.au | galaxy.inf.ethz.ch

Galaxy ToolShed

Galaxy

Galaxy Tool Shed Repositories Groups Help User

9588 valid tools on Sep 24, 2023

Search
Search for valid tools
Valid Galaxy Utilities
Tools
Custom datatypes
Repository dependency definitions
Tool dependency definitions
All Repositories
Browse by category
Available Actions
Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	198
Astronomy	Tools for astronomy	8
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	78
Climate Analysis	Tools for analyzing climate data	12
CLIP-seq	Tools for CLIP-seq	4
Combinatorial Selections	Tools for combinatorial selection	9
Computational chemistry	Tools for use in computational chemistry	180
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	11
Convert Formats	Tools for converting data formats	140
Data Export	Tools for exporting data to various destinations	17
Data Managers	Utilities for Managing Galaxy's built-in data cache	106
Data Source	Tools for retrieving data from external data sources	107
Ecology	Tools related to ecological studies	74
Entomology	Tools that involve insect studies	4
Epigenetics	Tools for analyzing Epigenetic/Epigenomic datasets	48
Fasta Manipulation	Tools for manipulating fasta data	121
Fastq Manipulation	Tools for manipulating fastq data	104
Flow Cytometry Analysis	Tools for manipulating and analyzing FCS files	45

https://galaxyproject.org/toolshed/

Software Sustainability in Science

Software Development

Software Packaging

Deployment

PyPI
481K Packages

CRAN
25.1K Packages

CPAN
40.3K Packages

docker

apt-get

yum
yellowdog updater modified

docker.com

Software Sustainability in Science

Standards needed for packaging bioinformatics software

- Programming language
- OS independent
- Multiple versions of the software
- HPC and Cloud compatible
- easy to maintain

Software Sustainability in Science

Software Development

Software Packaging

Deployment

PyPI
481K Packages

CRAN
25.1K Packages

CPAN
40.3K Packages

conda
18.8K Packages

BIOCONDA

docker

apt-get

yum
yellowdog updater modified

bioconda.github.io

Software Sustainability in Science

Code Blame 35 lines (29 loc) · 713 Bytes

```
{% set version = "0.2.2" %}

package:
  name: ribodiff
  version: {{ version }}

source:
  url: https://github.com/ratschlab/RiboDiff/archive/v0.2.2.tar.gz
  md5: b02833b4412959746032f0cf23a944d8

build:
  noarch: python
  number: 2

requirements:
  host:
    - python <3
    - pip
  run:
    - python <3
    - numpy >=1.8.0
    - scipy >=0.13.3
    - matplotlib >=1.3.0
    - statsmodels >=0.5.0

test:
  imports:
    - ribodiff
  commands:
    - TE.py --help

about:
  home: http://public.bmi.inf.ethz.ch/user/zhongy/RiboDiff/index.html
  license: GPL 3
  summary: 'RiboDiff is a statistical tool that detects the protein translational efficiency change from Ribo-Seq (ribosome footprinting) and RNA-Seq data.'
```

Software Sustainability in Science

bioconda.github.io

Edit me on GitHub

BIOCONDA

Navigation

FAQs
Contributing to Bioconda
Developer Docs
Tutorials

Browse packages
Bioconda @ Github
chat on gitter

Search
packages & docs

Go

BIOCONDA®

Bioconda lets you install thousands of software packages related to biomedical research using the conda package manager.

NOTE: *Bioconda supports only 64-bit Linux and macOS*

Usage

First, install conda.

Then perform a one-time set up of Bioconda with the following commands. This will modify your `~/.condarc` file:

```
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
conda config --set channel_priority strict
```

If you have used Bioconda in the past, note that the recommended configuration has changed over the years. You should run the above commands to ensure your settings follow the current recommendations.

- *How have the recommendations changed?*
- *What did these commands do?*
- *What if I don't want to modify my condarc?*

Now you can use `conda` to install and use any of the packages available in bioconda.

Software Sustainability in Science

Large scale bioinformatics data analysis

- Bioconda standard package management
- Different workflow managers
- Success story of the bioconda community
- easy to maintain

COMMON WORKFLOW LANGUAGE

nextflow

nextflow.io | snakemake.github.io | commonwl.org

Software Sustainability in Science

Software Sustainability in Science

Software Development

Software Packaging

Deployment

conda
18.8K Packages

BIOCONDA

docker

BioContainers

Galaxy Term - Mulled

- without Dockerfile
- layer donning approach for building containers

biocontainers.pro

Software Sustainability in Science

- 9874 bioinformatics packages
- 95404 singularity images

Software Sustainability in Science

nextflow

```
profiles {
  samtools {
    process.container = 'https://depot.nexus.ethz.ch/singularity/samtools:1.2'
    singularity.enabled = true
    singularity.cacheDir = "$PWD"
  }
}
```

snakemake

```
rule samtools:
    input:
        "inputs/{unmapped}.bam"
    output:
        "results/{unmapped}.bam"
    threads: 1
    singularity: "https://depot.nexus.ethz.ch/singularity/samtools:1.2"
    shell:
        """
        cat {input} > {output}
        """
        "samtools view {input} > {output}"
```

Software Sustainability in Science

Galaxy PROJECT

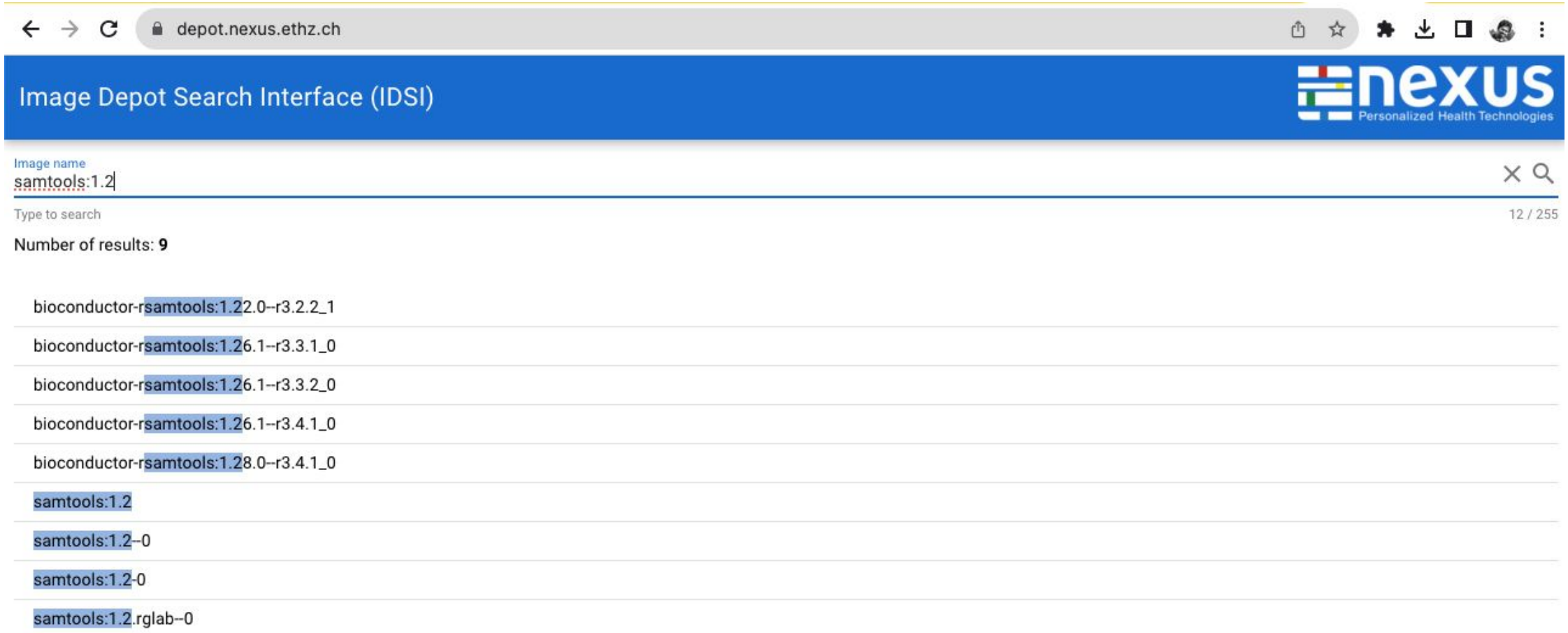
COMMON WORKFLOW LANGUAGE

Open Science Grid

nextflow

CYVERSE®

Software Sustainability in Science



depot.nexus.ethz.ch

Acknowledgments

Galaxy PROJECT

Galaxy EUROPE

BIOCONDA

ETH zürich

- NEXUS Software Engineering
- HPC Storage Service
- S4D Support Service
- ETH Gitlab Service