Interactive applications on HPC systems

Erich Birngruber
(erich.birngruber@gmi.oeaw.ac.at, @ebirn)
Vienna BioCenter

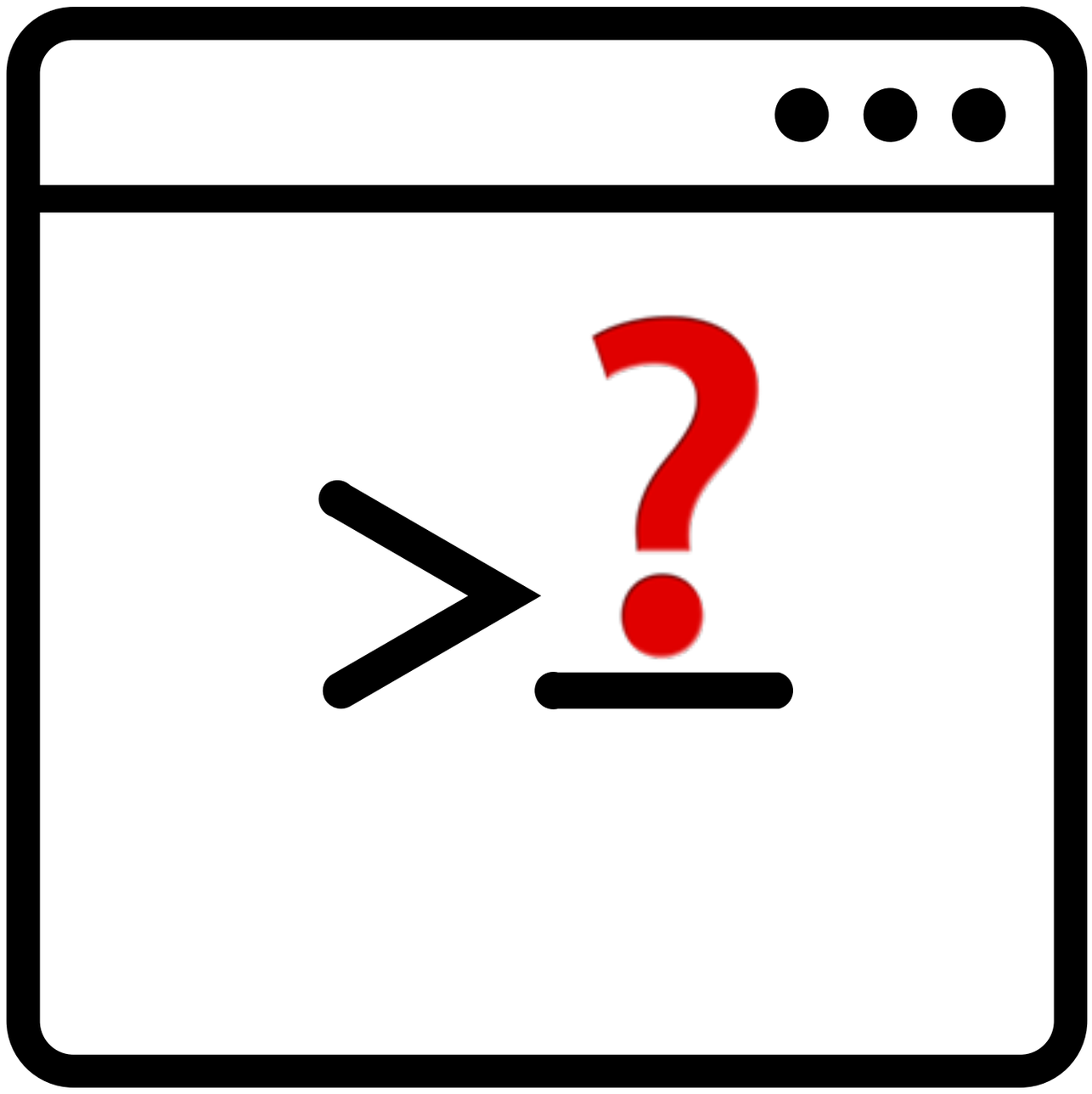
FOSDEM20

Interactive applications on HPC systems

Erich Birngruber
(erich.birngruber@gmi.oeaw.ac.at, @ebirn)
Vienna BioCenter

FOSDEM20

sh$ not good enough?



XPRA

XPRA

- https://xpra.org/
- “screen for X11”
- Allows disconnect / re-connect to existing X sessions
- Web interface for X11 rendering (HTML5 canvas)
- For arbitrary GUI applications
- Containerized in SLURM
- Custom middleware for job management

IT - Portal

it.vbc.ac.at/clip/cbe/xpra

Vienna BioCenter

erich.birngruber@gmi.oeaw.ac.at

- Dashboard
- Statistics
- Shop
- Documentation
- Announcements
- Cluster
- Infomail

Xpra - Run cluster jobs with a web UI

Application

Fiji

Cores

4

Memory (GB)

16

Walltime (h)

1

GPUs

0

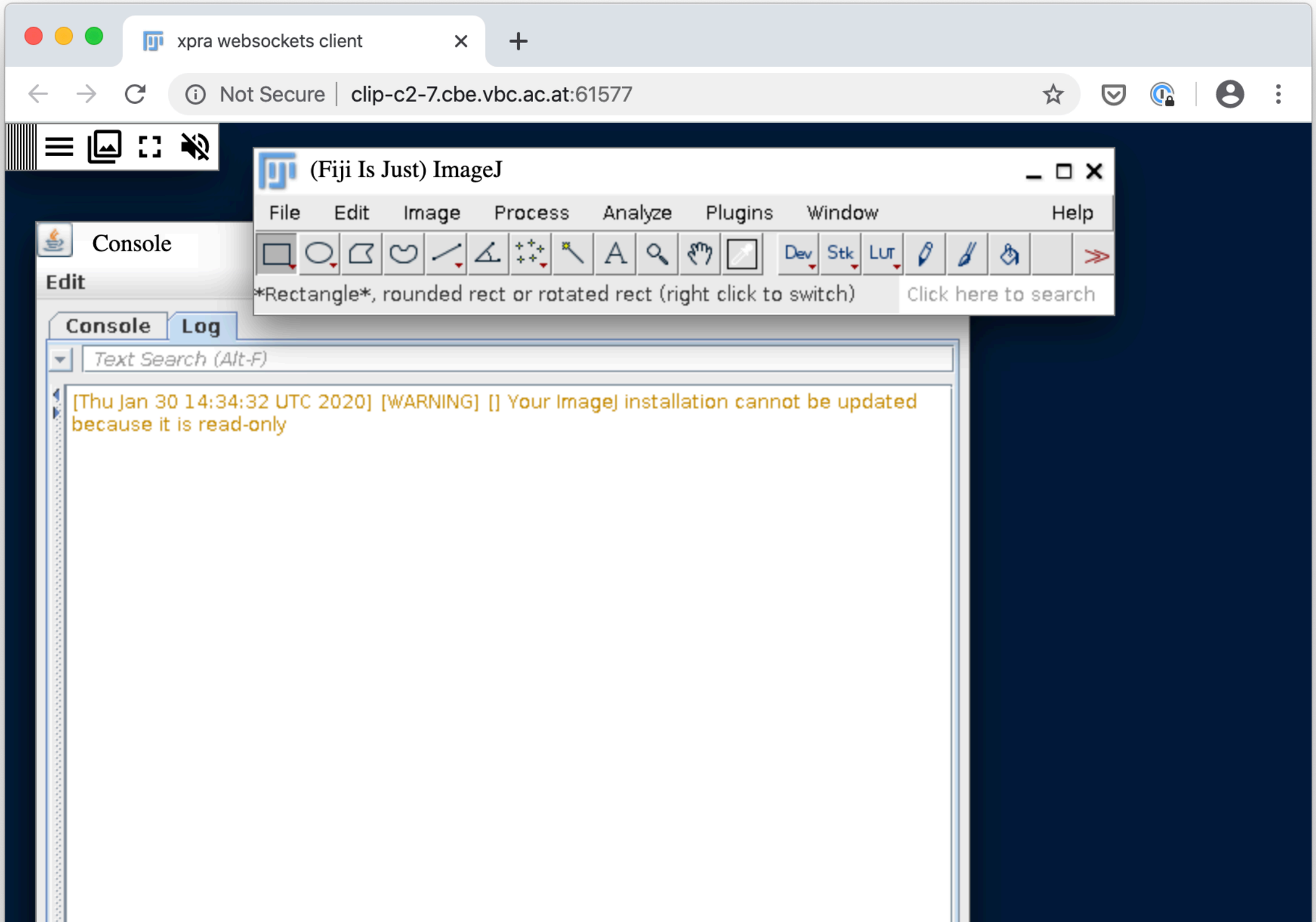
Start

XPRA job submitted

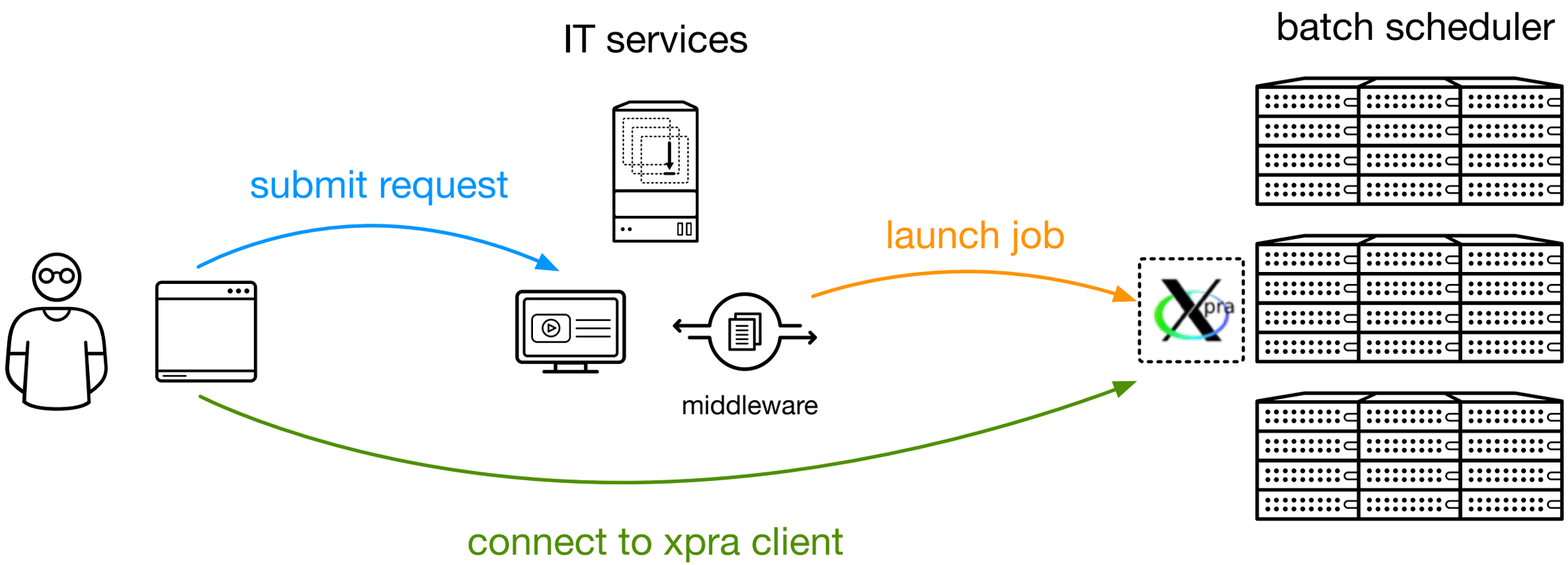
ID	Application	Hours	Cores	Memory (GB)	GPUs	State	Actions
313	Fiji	1	1	4	0	Job running	Join / Settings
56	X-Term	1	12	10	0	Job finished	

Items per page: 5 1 - 2 of 2

XPRA session



XPRA setup


IT services
batch scheduler
submit request
launch job
middleware
connect to xpra client

RStudio®

- https://rstudio.com/
- IDE for R language
- Desktop and Web version (RStudio server)
- Commercial version for advanced features
- RStudio company has become a public benefit company https://blog.rstudio.com/2020/01/29/rstudio-pbc

RStudio Server

rstudio.vbc.ac.at/s/57ea13c286bd33c286bd3/workspaces/

RStudio Server Pro

Logout

Sessions

New Session | Suspend all | Quit all

RStudio Session

Info | Suspend | Quit

IDLE R 3.5.1(R & Bioconducto)

(Home) CREATED: 3:48:11 PM LAST USED: 3:51:18 PM

(Home)

Info | Quit

SUSPENDED R 3.5.1(R & Bioconducto)

(Home) CREATED: 1/15/2020 LAST USED: 1/15/2020

Projects

Open a new project

RStudio Pro

rstudio.vbc.ac.at/s/57ee5cfc78a31a3dffeb1/?launcher=1

File Edit Code View Plots Session Build Debug Profile Tools Help

erich.birngruber Sessions (2)

Go to file/function Addins Project: (None) R 3.5.1

Console Terminal Launcher

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> print("I don't know R")
[1] "I don't know R"
> print()
```

princomp	{stats}
print	{base}
print.AsIs	{base}
print.by	{base}
print.condition	{base}
print.connection	{base}
print.data.frame	{base}

print(x, ...)

print prints its argument and returns it *invisibly* (via invisible(x)). It is a generic function which means that new printing methods can be easily added for new classes.

Press F1 for additional help

Environment History Connections

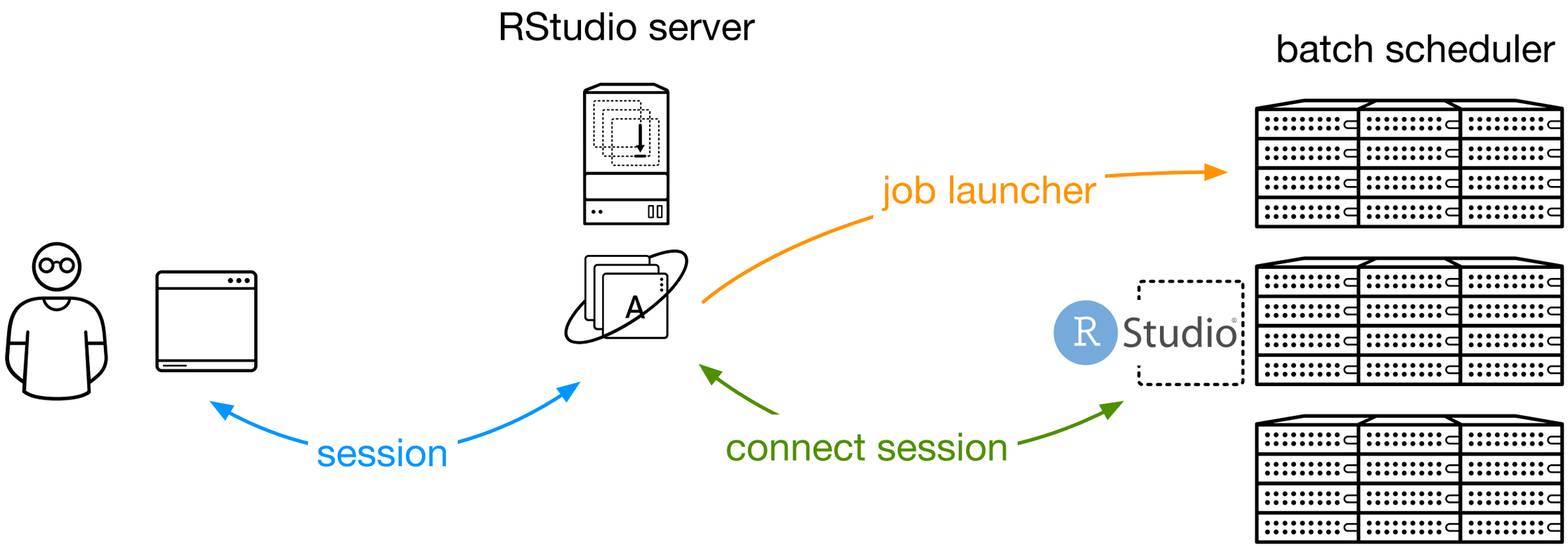
New Connection

Connection	Status

Files Plots Packages Help Viewer

Zoom Export

RStudio setup



Galaxy

PROJECT

Galaxy
PROJECT

- https://galaxyproject.org/
- Web based workflow tool
- Tools as building blocks (parameters, input, output)
- Tool definitions in XML
- Multiple instances: dev - testing - production

Galaxy

tds.galaxy.vbc.ac.at

Galaxy | Analyze Data | Workflow | Visualize | Shared Data | Help | User | Using 0%

Tools

search tools

Get Data

Public databases

Export Data

ALIGNMENT

Genome alignment

Blast Runs the selected BLAST search

Blat Aligns the reads to the selected reference

SPALN Maps the reads to the selected reference

Sequence alignment

NEXT-GENERATION SEQUENCING

NGS: Convert

BAM to FASTQ Extracts the reads (FASTQ) from a BAM file

BAM to BigWig Converts BAM/SAM files to BigWig

NGS: Hi-C

NGS: ChIP-seq

NGS: Expression

javascript:void(0)

Blast Runs the selected BLAST search (Galaxy Version 2.8)

Favorite | Options

Cluster Options

Memory (GB)

16

Walltime (h)

1

Source

File in your history

Query sequence(s) in FASTA format

No fasta dataset available.

Algorithm

BLASTn (DNA query against DNA database)

Select the BLAST algorithm

Database

Ambystoma mexicanum genome (AmexG_v3.0.0)

Job Resource Parameters

Use default job resource parameters

Execute

Galaxy | Workflow Editor

tds.galaxy.vbc.ac.at/workflow/editor?id=8396b9531999c3a9#

Galaxy

Analyze Data | Workflow | Visualize | Shared Data | Help | User | Using 0%

Tools

search tools

NGS: Expression

NGS: Bisulfite sequencing

NGS: QC and manipulation

GROUPS

Tanaka

Zuber

Busslinger

VBC

TOOLKITS

samtools

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to BAM

samtools BAM to CRAM convert BAM alignments to CRAM format

BedCov calculate read depth for a set of genomic intervals

CalMD recalculate MD/NM tags

CRAM to BAM convert CRAM alignments to BAM format

Extract FASTA or FASTQ from a SAM file

Demo workflow

Input dataset

output

Blast

Query sequence(s) in FASTA format

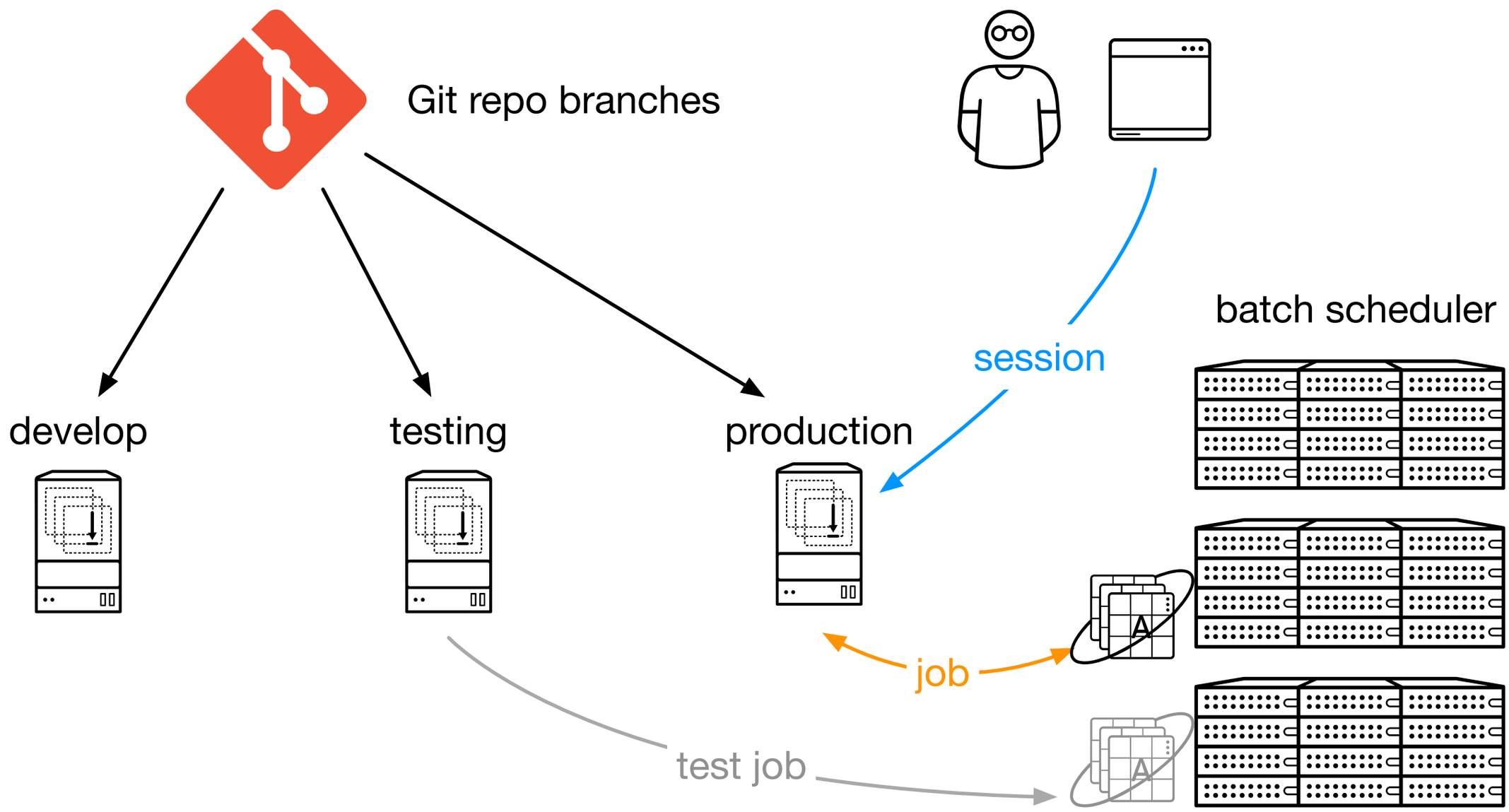
BLAST (${blast.algorithm}) results of ${on_string} against ${blast.db.name} (txt)

Export to Fileserver

File

logfile (txt)

Galaxy setup



jupyterhub

- https://jupyter.org/
- Web-Based IDE (standalone vs. hub)
- Notebooks = Code + Outputs
- Interpreters as “Kernels”

JupyterHub

jupyterhub.vbc.ac.at/hub/spawn/erich.birngruber

jupyter Home Token erich.birngruber Logout

Spawner Options

Job type

CPU short (4c, 16gb, 4h)

Jupyter environment

Environment based on CBE env modules (Python 3.6.6)

Logging

enable logging to $HOME/jupyterhub_{jobid}.log

Environment variables (one per line)

MY_VAR=myvalue123

Spawn

JupyterLab

jupyterhub.vbc.ac.at/user/erich.birngruber/lab?

File Edit View Run Kernel Tabs Settings Help

Launcher | matplotlib_demo.ipynb

Code | Python 3

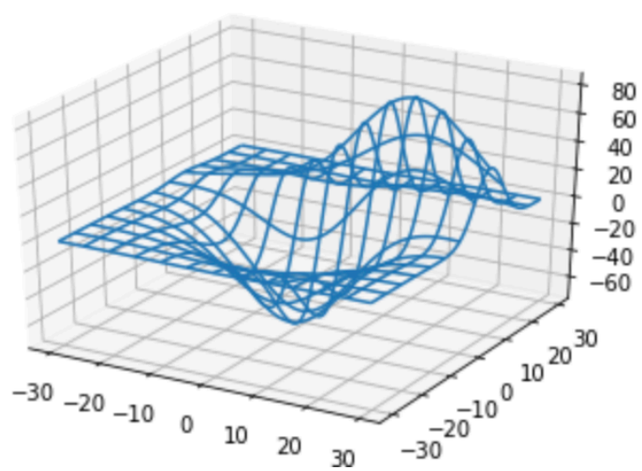
/

Name

- jpy_install
- my_jpy_kernel_env
- my_repo
- myRlibs
- pauli_env
- R
- razzle_dazzle
- rstudio-logs
- 7-ComparativeMeas...
- example.fasta
- h5ex_d_hyper.h5
- install.sh
- jpy_conda.yml
- jupyterhub_38180.log
- jupyterhub.sh
- matplotlib_demo.ipynb
- matplotlib_demo.py
- MEGA_note.ipynb
- my.py
- nested_data_ext.hdf5
- slurm-1244177.out
- spawn.sh
- Untitled.ipynb
- untitled.md

[4]:

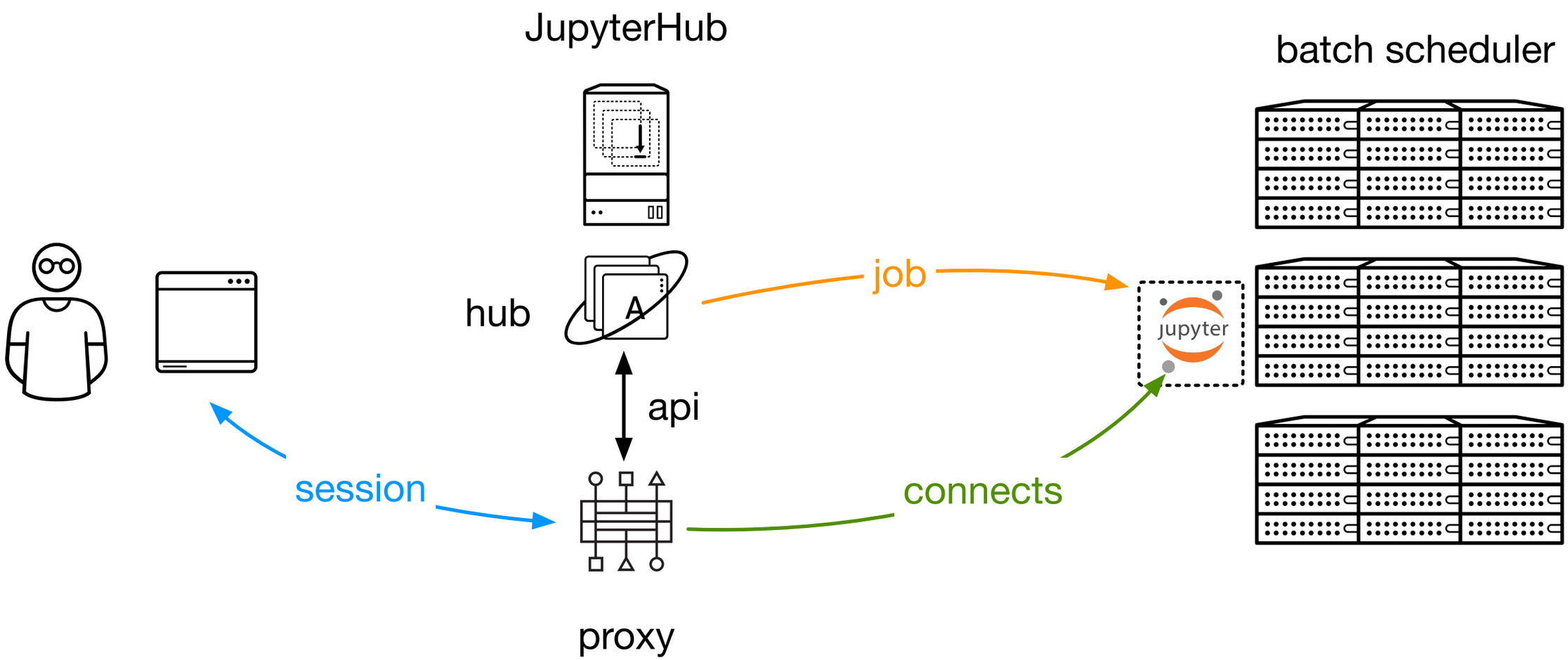
```
from mpl_toolkits.mplot3d import axes3d

fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')

# Grab some test data.
X, Y, Z = axes3d.get_test_data(0.05)

# Plot a basic wireframe.
ax.plot_wireframe(X, Y, Z, rstride=10, cstride=10)

# fig.canvas.layout.max_width = '1000px'

plt.show()
```



[ ]:

0 | 2 | Python 3 | Idle | Mode: Command | Ln 1, Col 1 | matplotlib_demo.ipynb

JupyterHub setup



Summary

- XPRA
  Special use cases: X11 applications (Fiji) in Containers
- RStudio
  R (from env modules), web-based IDE
- Galaxy
  pre-configured workflows
- JupyterHub
  Python (per-user kernels), plugins

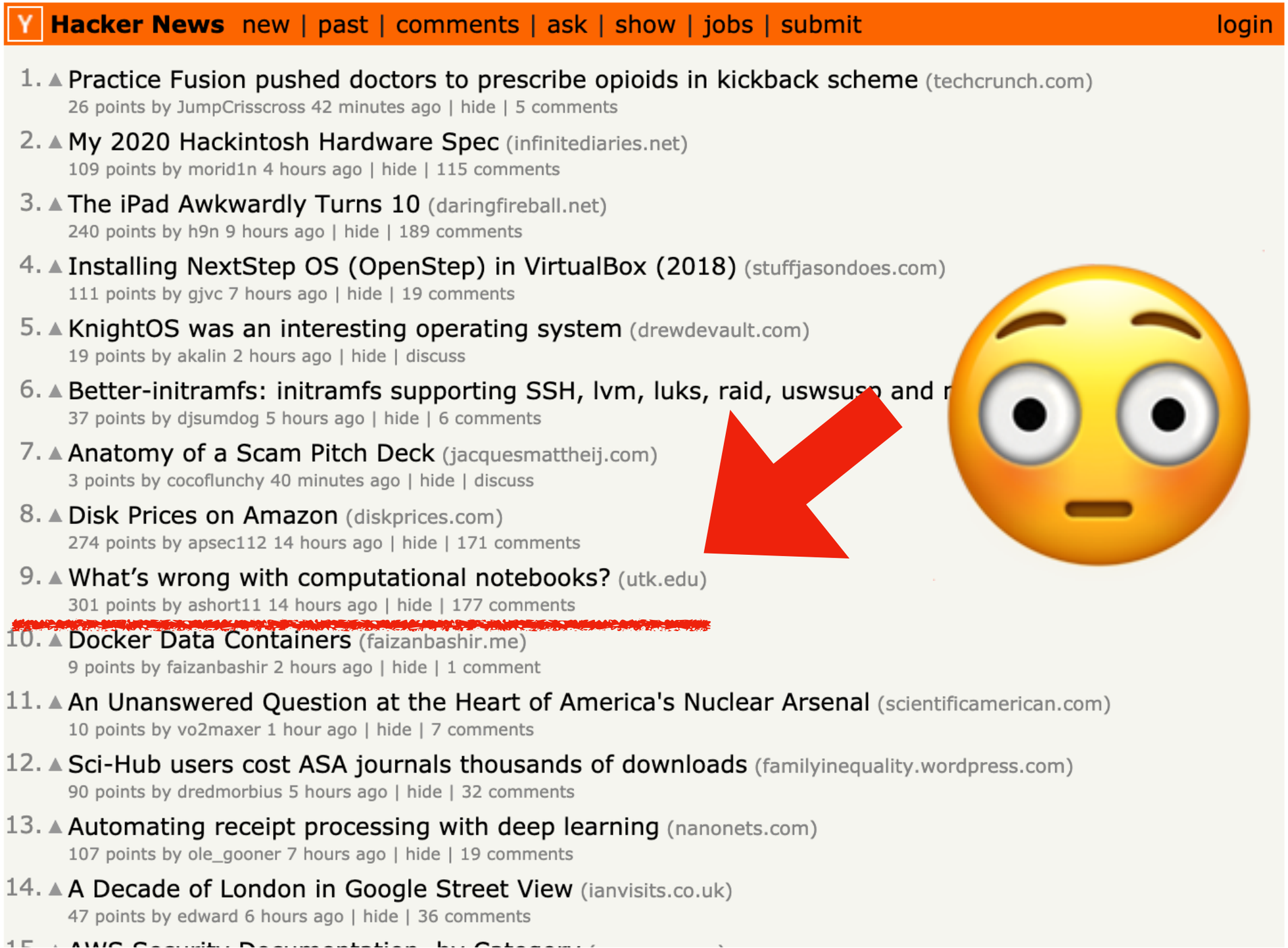
Others

- OpenOnDemand: interactive/remote desktop portal
  https://openondemand.org/

- Apache Zeppelin: data exploration “notebooks”
  https://zeppelin.apache.org/

- Eclipse Che: cloud-based editor
  https://www.eclipse.org/che/

Then this happened



What *is* wrong?

What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities

Souti Chattopadhyay[1], **Ishita Prasad**[2], **Austin Z. Henley**[3], **Anita Sarma**[1], **Titus Barik**[2]
Oregon State University[1], Microsoft[2], University of Tennessee-Knoxville[3]
{chattops, anita.sarma}@oregonstate.edu, {ishita.prasad, titus.barik}@microsoft.com, azh@utk.edu

ABSTRACT

Computational notebooks—such as Azure, Databricks, and Jupyter—are a popular, interactive paradigm for data scientists to author code, analyze data, and interleave visualizations, all within a single document. Nevertheless, as data scientists incorporate more of their activities into notebooks, they encounter unexpected difficulties, or pain points, that impact their productivity and disrupt their workflow. Through a systematic, mixed-methods study using semi-structured interviews ($n = 20$) and survey ($n = 156$) with data scientists, we catalog nine pain points when working with notebooks. Our findings suggest that data scientists face numerous pain points throughout the entire workflow—from setting up notebooks to deploying to production—across many notebook environments. Our data scientists report essential notebook requirements, such as supporting data exploration and visualization. The results of our study inform and inspire the design of computational notebooks.

Author Keywords

Computational notebooks; challenges; data science; interviews; pain points; survey

CCS Concepts

Azure,[1] Databricks,[2] Colab,[3] Jupyter,[4] and nteract.[5] While originally intended for exploring and constructing computational narratives [29, 31], data scientists are now increasingly orchestrating more of their activities within this paradigm [33]: through long-running statistical models, transforming data at scale, collaborating with others, and executing notebooks directly in production pipelines. But as data scientists try to do so, they encounter unexpected difficulties—pain points—from limitations in affordances and features in the notebooks, which impact their productivity and disrupt their workflow.

To investigate the pain points and needs of data scientists who work in computational notebooks, across multiple notebook environments, we conducted a systematic mixed-method study using field observations, semi-structured interviews, and a confirmation survey with data science practitioners. While prior work has studied specific facets of difficulties in notebooks [24, 17], such as versioning [18, 19] or cleaning unused code [13, 34], the central contribution of this paper is a taxonomy of validated pain points across data scientists' notebook activities.

Our findings identify that data scientists face considerable pain points through the entire analytics workflow—from setting up the notebook to deploying to production—across

References

- XPRA https://xpra.org/
- RStudio https://rstudio.com/
- Jupyterhub https://jupyter.org/hub
- Galaxy https://galaxyproject.org/
- What is wrong with computational notebooks? http://web.eecs.utk.edu/~azh/blog/notebookpainpoints.html