

Expanding the role of HPC centres on training and collaboration for reproducibility

Computing Insight UK 2020



Science & Technology
Facilities Council

Lorena A Barba

SC20 invited talk

- ▶ Title: "Trustworthy computational evidence through transparency and reproducibility"
- ▶ Session: Responsible Application of HPC

Here, I expand on: how could teams at supercomputing facilities work with researchers to help them adopt better reproducibility practices?



SC20
Everywhere we are | more than hpc.

SC19 Reproducibility Chair

- ▶ Artifact Description appendix *required*
 - ▶ Standard form asks about software, data, or other digital artifacts
- ▶ Reviewed, innovative double-open model
- ▶ Reproducibility Challenge



SC19
Denver, CO | **hpc**
is now.

Open peer review of Appendices

Artifact Description

- ▶ Constructive: conversation with authors typical
- ▶ Artifact availability: many authors provide URLs to GitHub or lab website
- ▶ GitHub does not provide guarantees of persistence
- ▶ Researchers need advice and technical support!

Zenodo.org

Data repository

- ▶ funded by the European Union's Framework Programs for Research
- ▶ operated by CERN, largest high-energy physics laboratory of the world
- ▶ hosts any kind of data, under any license type (including closed-access)

zenodo

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science



- ▶ Study mandated by Congress in Jan. 2017, commissioned by the NSF to The National Academies of Sciences, Engineering and Medicine (NASEM)
- ▶ report released 7 May 2019

Reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis.

<http://doi.org/c5jp>



Reproducibility of scientific results in the EU

Scoping Report

- ▶ “reproducibility continuum” based on three research processes: reproduction, replication, re-use.
- ▶ “All three processes rest on the availability of data and methods from the original study.”
- ▶ “While the lack of reproducibility is a serious problem, it is not to the extent of a crisis.”

<https://op.europa.eu/s/ou0J>

Reproducibility hinges on transparency

Sharing of research objects

- ▶ Pre-emptive, via archive deposit in a persistent service that assigns a global identifier
- ▶ NOT “available upon request”
- ▶ Better: open-source license, public version control repository
- ▶ Even better: open development model

In HPC setting: why share code and data if other researchers cannot run?

Runs are costly and require allocations via competitive grant proposals

Projects are not “born HPC”

They mature over the years

- ▶ From code prototypes developed by grad students, to software collaborations, to large-scale projects
- ▶ Reproducible research practices should be applied from the start
- ▶ Graduate programs offer no adequate training
- ▶ Some groups have developed internal materials and onboarding

(See: Barba Group Reproducibility Syllabus, 2016;
Essential Skills for Reproducible Research, 2017.)

Proposal: HPC centres can serve a key role with educational programs in support of reproducible computational research

In addition to workshops and Summer schools in parallel computing, accessing systems, etc.

NASEM Recommendations

For various stakeholders

- ▶ ...researchers, journals and conferences, professional societies, academic institutions and national laboratories, and funding agencies
- ▶ **Recommendation 6-9:**
 - ▶ grant applications should discuss how they will assess and report uncertainties, and how the proposed work will address reproducibility and/or replicability issues
 - ▶ funders should use reproducibility and replicability in the **merit-review** criteria of grant proposals

Sandia National Laboratory

Laboratory Directed Research & Development grants

Thanks!

- ▶ proposers must include statements on how the project will address transparency and reproducibility
- ▶ discussed in the proposal review and a factor in the decision-making
- ▶ some teams are performing at a high level: artifact appendix for every publication



Mike Heroux, Senior Scientist at Sandia

Allocations of computing time

NSF science and allocation awards are separate

- ▶ Allocation Submission Guideline: intellectual merit, methodology, research plan and resource request, and efficient use of the computational resources
- ▶ not tied to performance, but to scale
- ▶ responsible stewardship via collaboration between PIs and the facility staff
- ▶ collection of system-wide performance data with TACC-Stats

Thanks!



John West, TACC
Director of
Strategic Initiatives

Opportunities

For infrastructure providers to lead

1. workflow-management/system monitoring to also supply automated provenance capture
2. expert staff at the facility broaden researcher support to include advice and training in transparency and reproducibility matters
3. cyber-infrastructure facilities could expand their training initiatives to include essential skills for reproducible research
4. small bump in allocation to engage in R&R activities?

In HPC settings

A blend of incentives and norms

- ▶ we can hardly ever reproduce results: machine access, cost, and effort
- ▶ consider especially the applications of high consequence to society
- ▶ support researchers with **infrastructure**
- ▶ arrive at a level of maturity to achieve the goal of trustworthy computational evidence, not by actually exercising the open research objects (artifacts) shared by authors (data and code), but by a research process that ensures *unimpeachable provenance*.

Infrastructure for interactive computing

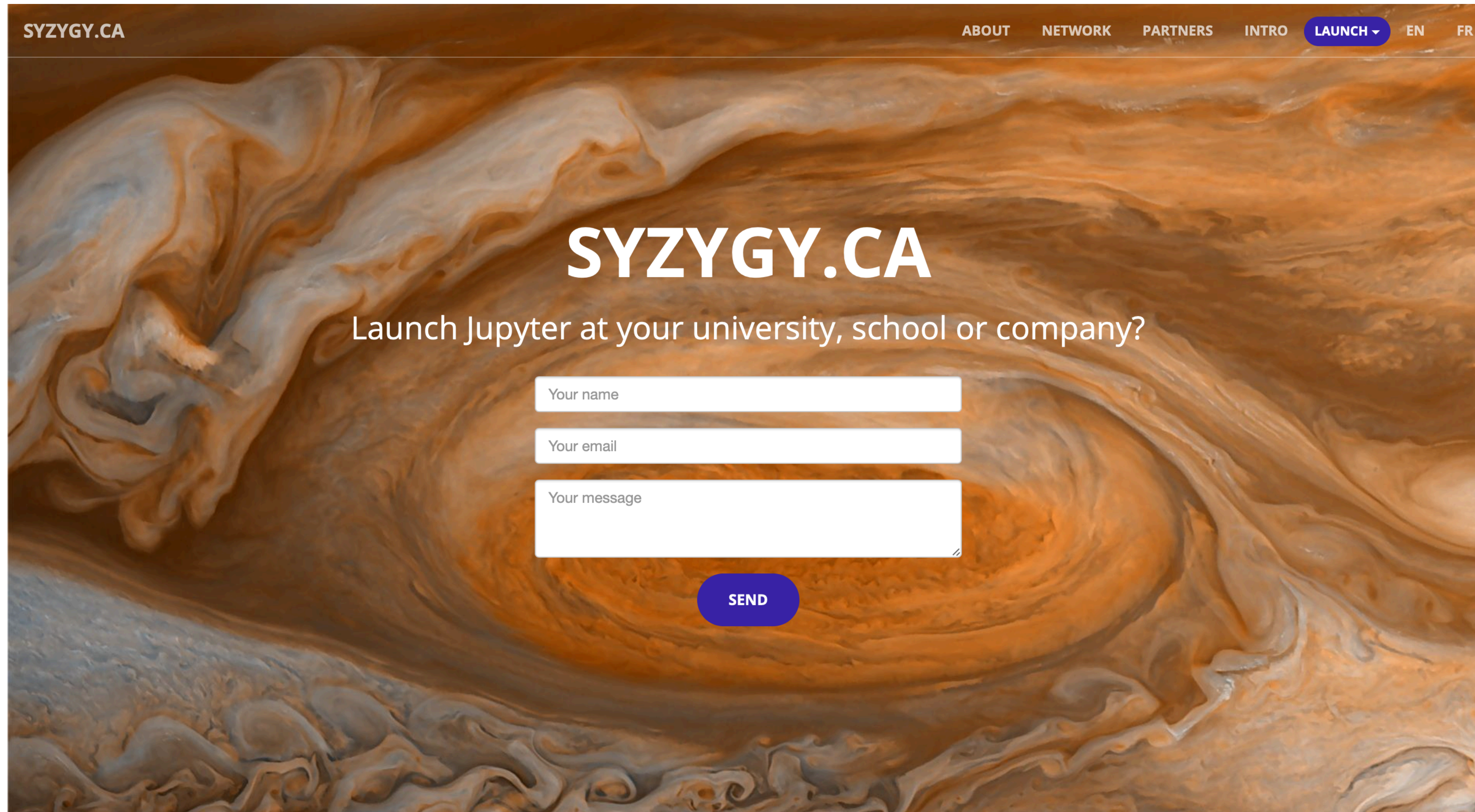
Scalable JupyterHub for education and research

- ▶ Jupyter: open-source tools for interactive and exploratory computing, at the heart of which is the Jupyter Notebook
- ▶ JupyterHub: to serve Jupyter Notebook for multiple remote users



Example: Canada's federated JupyterHub

Serving more than 20 universities, PIMS, Cybera

The image shows a web browser window displaying the SYZYGY.CA website. The background is a swirling, orange and white abstract pattern. The website has a dark brown header with the logo 'SYZYGY.CA' on the left and navigation links 'ABOUT', 'NETWORK', 'PARTNERS', 'INTRO', 'LAUNCH', 'EN', and 'FR' on the right. The 'LAUNCH' link is highlighted with a blue background and a white dropdown arrow. The main content area features the text 'Launch Jupyter at your university, school or company?' in white. Below this text are three white input fields: 'Your name', 'Your email', and 'Your message'. A blue 'SEND' button is positioned below the 'Your message' field. The overall design is clean and modern, with a focus on the swirling background image.

SYZYGY.CA

ABOUT NETWORK PARTNERS INTRO LAUNCH EN FR

SYZYGY.CA

Launch Jupyter at your university, school or company?

Your name

Your email

Your message

SEND

Example: Berkeley Data8 project

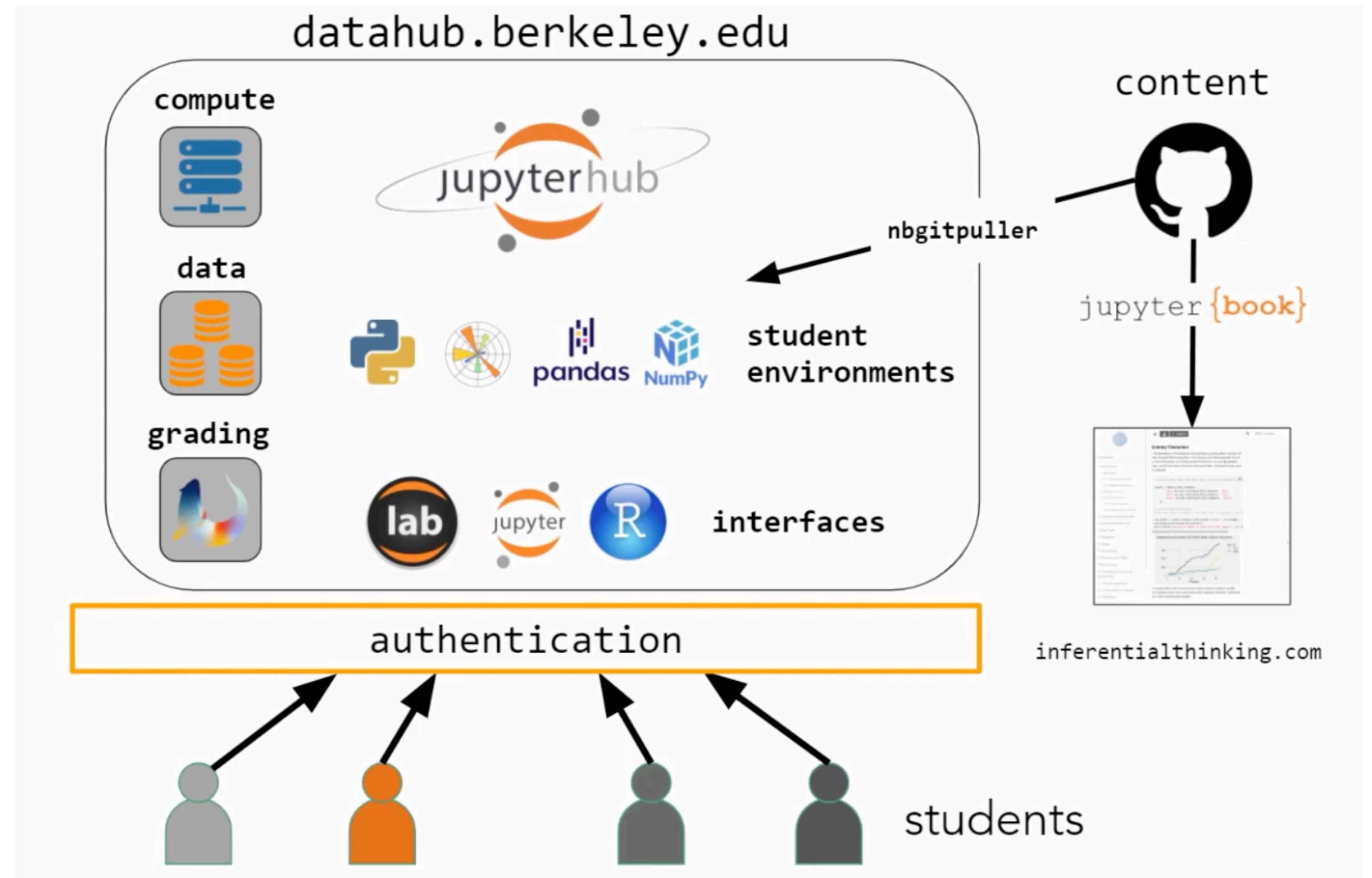
Teaching data science to thousands of students

- ▶ Provides *open infrastructure* for large-scale introductory course
- ▶ *Shared* computing environment for students, teaching assistants and faculty (no need for tech support or managing disparate environments)
- ▶ Learning content prepared in Jupyter, and auto-graded assignments using Jupyter

Incredible success!

Example: Berkeley Data8 project

Teaching data science to thousands of students



Credit: Chris Holdgraf

Who provides the research/teaching infra?

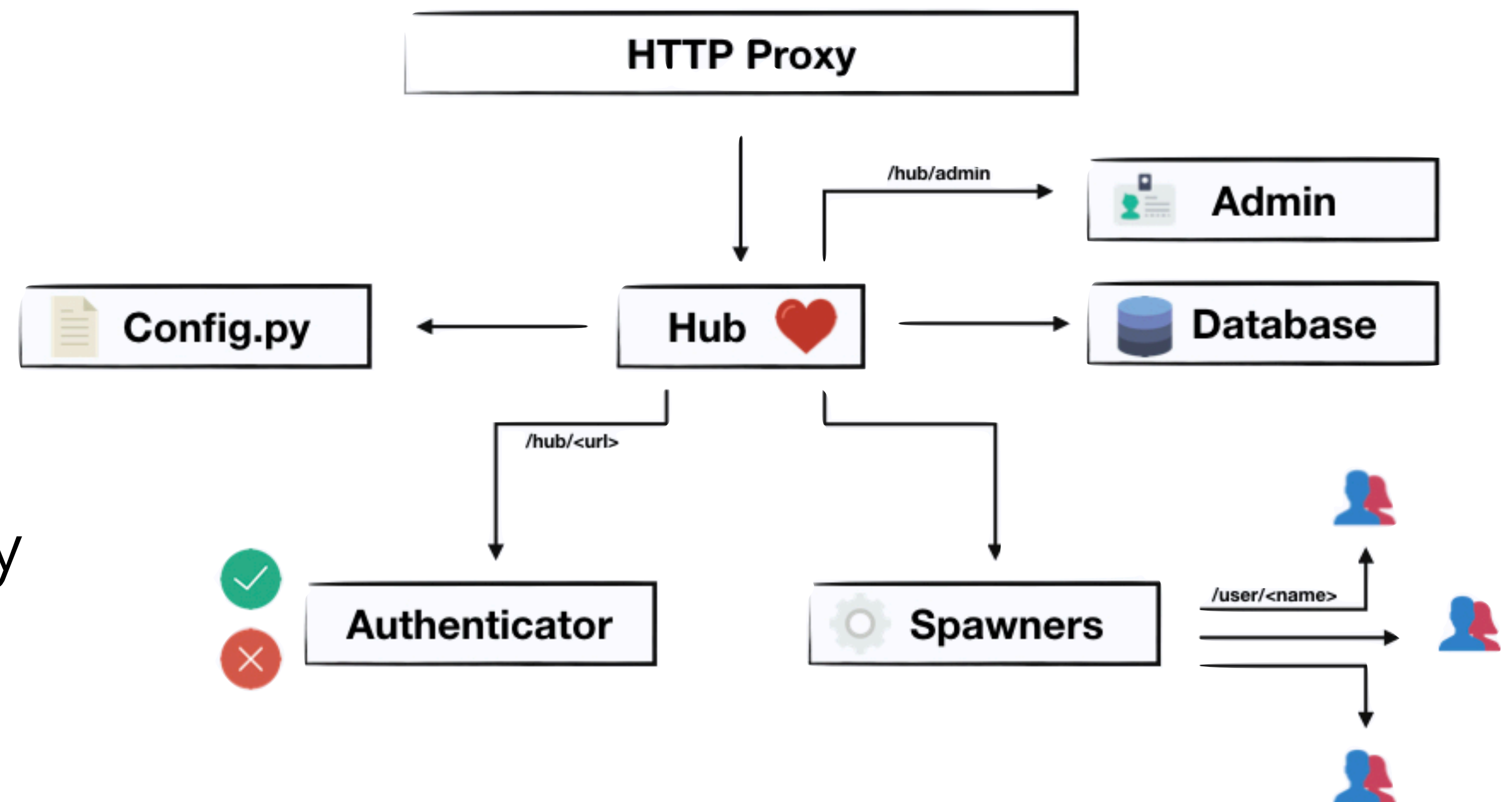
Expertise not often available at universities

<https://jupyterhub.readthedocs.io/>

Subsystems:

- ▶ Hub: tornado process
- ▶ configurable http proxy
- ▶ Authenticator
- ▶ Single-user servers monitored by spawners

JupyterHub



Fully open source stack preferable

Universities should avoid vendor lock-in

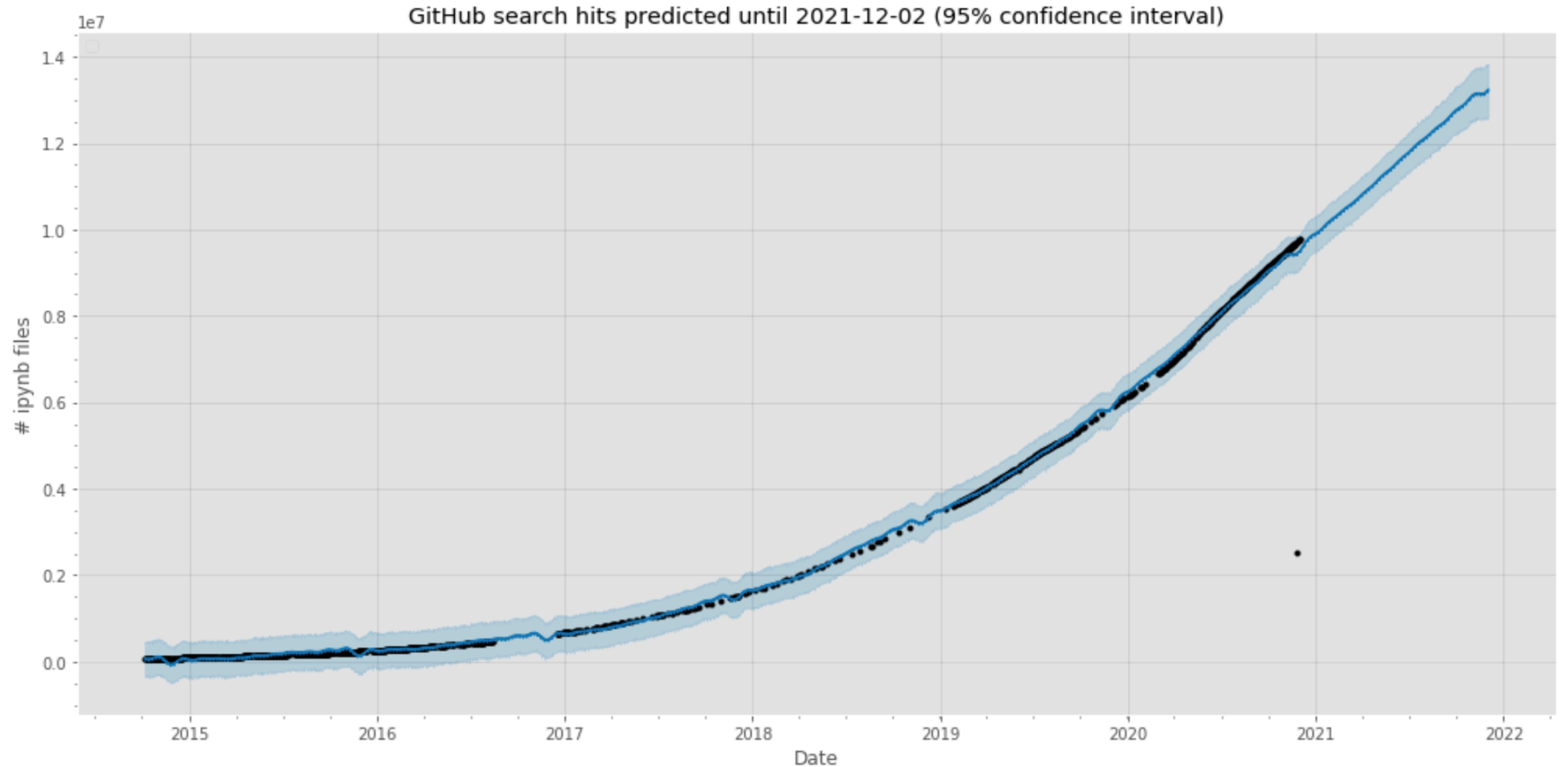
- ▶ Jupyter has become the *de facto* environment for data science
- ▶ Start-ups are popping up everywhere with SaaS solutions, and they market to universities
 - ▶ Tailored to industry workflows, not research/education
 - ▶ The interfaces/UX are custom, and not what students see in plain Jupyter
- ▶ Flexibility: can use any cloud provider

**Opportunity for HPC centres to partner
with universities and expand their mission
to a wider sector of users!**

The Jupyter ecosystem is now essential computing infrastructure for research and education

Number of Jupyter notebooks on GitHub

<https://github.com/parente/nbestimate>



**Opportunity for HPC centres to partner
with universities and expand their mission
to a wider sector of users!**

The Jupyter ecosystem is now essential computing infrastructure for research and education

Expanding the role of HPC centres on training and collaboration for reproducibility

Computing Insight UK 2020



Science & Technology
Facilities Council

Lorena A Barba