Science Services and Science Platforms

Using the cloud to accelerate and democratize discovery

Talk at NorduGrid annual conference, Košice, Slovakia, June 2nd, 2016

Ian Foster

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Thanks to co-authors and Globus team

Globus services (globus.org)

- Foster, I. **Globus Online: Accelerating and democratizing science through cloud-based services**. *IEEE Internet Computing*(May/June):70-73, 2011.
- Chard, K., Tuecke, S. and Foster, I. **Efficient and Secure Transfer, Synchronization, and Sharing of Big Data**. *Cloud Computing, IEEE*, 1(3):46-55, 2014.
- Chard, K., Foster, I. and Tuecke, S. **Globus Platform-as-a-Service for Collaborative Science Applications**. *Concurrency - Practice and Experience*, 27(2):290-305, 2014.

Publication (globus.org/data-publication)

- Chard, K., Pruyne, J., Blaiszik, B., Ananthakrishnan, R., Tuecke, S. and Foster, I., **Globus Data Publication as a Service: Lowering Barriers to Reproducible Science**. 11th IEEE International Conference on eScience Munich, Germany, 2015

Discovery engines

- Foster, I., Ananthakrishnan, R., Blaiszik, B., Chard, K., Osborn, R., Tuecke, S., Wilde, M. and Wozniak, J. **Networking materials data: Accelerating discovery at an experimental facility**. Big Data and High Performance Computing, 2015.

Thank you to our sponsors!

NIST

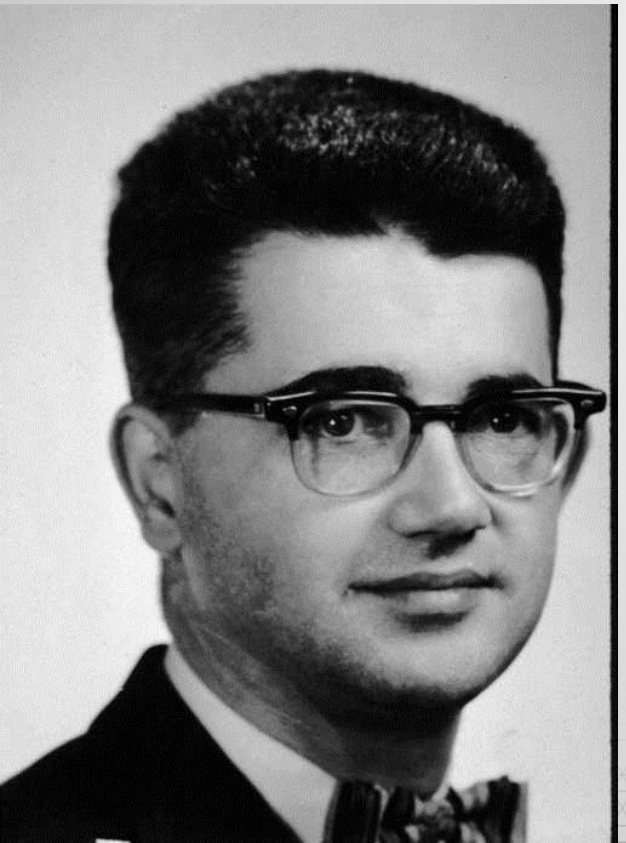
THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

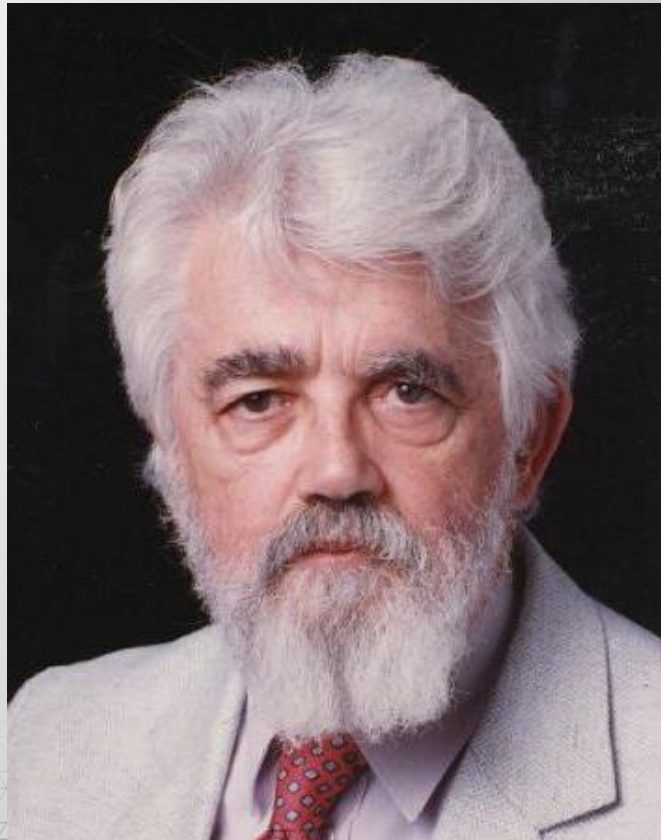
Civilization advances by extending the number of important operations which we can perform without thinking about them

Alfred North Whitehead (1911)

Computation may someday be organized as a public utility …
The computing utility could become the basis for a new and important industry.



John McCarthy (1961)



The grid vision

Accelerate discovery and innovation by providing on-demand access to computing

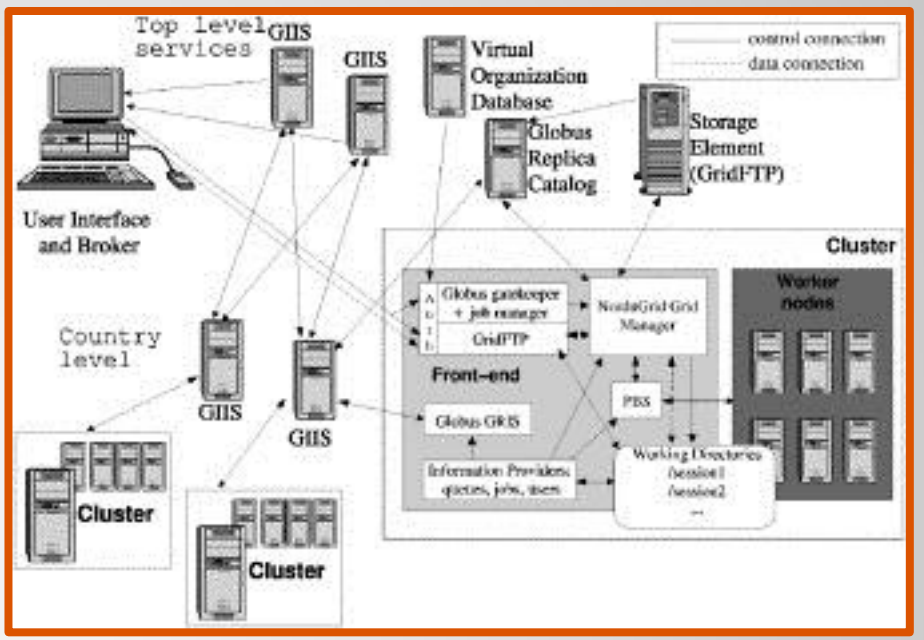
- “the average computing environment remains inadequate for [many] computationally sophisticated purposes”
- “if mechanisms are in place to allow reliable, transparent, and instantaneous access to high-end resources, then it is as if those resources are devoted to them”

[*The Grid*, Chapter 2, 1998]

Another pioneer: NorduGrid

NORDUGRID

Grid Solution for Wide Area Computing and Data Handling



The NorduGrid project: using Globus toolkit for building GRID infrastructure

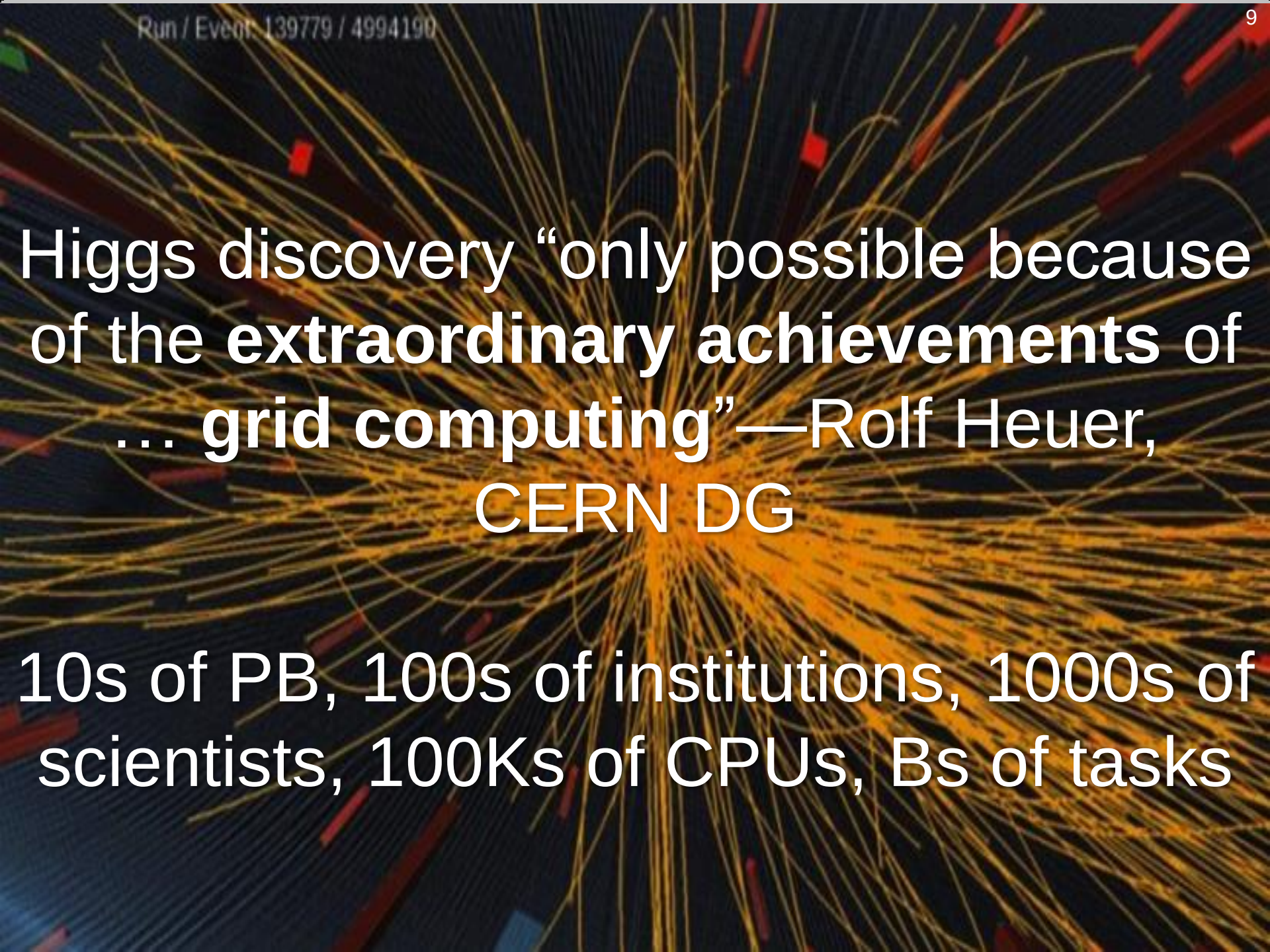
M. Ellert[a], A. Konstantinov[b,*], B. Kónya[c], O. Smirnova[c], A. Wäänänen[d]

[a] *Department of Radiation Sciences, Uppsala University, Box 535, 751 21 Uppsala, Sweden*
[b] *University of Oslo, Department of Physics, P.O. Box 1048, Blindern, 0316 Oslo, Norway*
[c] *Elementarpartikelfysik, Fysiska Institutionen, Lunds Universitet., Box 118, 22100 Lund, Sweden*
[d] *Niels Bohr Institutet for Astronomi, Fysik og Geofysik., Blegdamsvej 17, Dk-2100 Copenhagen Ø, Denmark*

Example grid scenarios

- “The application service providers, storage service providers, cycle providers, and consultants engaged by a car manufacturer to perform scenario evaluation during planning for a new factory”
- “Members of an industrial consortium bidding on a new aircraft”
- “A crisis management team and the databases and simulation systems that they use to plan a response to an emergency situation”
- “Members of a large, international, multiyear high-energy physics collaboration”

From: *The Anatomy of the Grid*, 2001



Higgs discovery “only possible because of the **extraordinary achievements** of … **grid computing**”—Rolf Heuer, CERN DG

10s of PB, 100s of institutions, 1000s of scientists, 100Ks of CPUs, Bs of tasks

What has changed?

- Thousands of people learned about the joys of large-scale distributed systems
- Virtual organization concepts and technologies
- Now routine to move 100s of terabytes (e.g., GridFTP moves >2 petabyte per day)
- High throughput computing is mainstream (e.g., Condor and Globus run millions of jobs per day)
- Large Hadron Collider found the Higgs
- Earth System Grid supports >25,000 users
- Commercial cloud computing has exploded

Looking forward

- Exploding data volumes and earlier successes mean that many more people face challenges of big data, big compute, big collaboration
- Networks are several orders of magnitude faster than when Grid started
- Commercial cloud providers provide a substrate on which powerful new capabilities can be built with new economies of scale

The BVP Cloudscape

Top 300 Privately Held Cloud Companies

SaaS

Business Users

Marketing

action, HubSpot, bizo, AGILONE, gigya, ensighten, SurveyMonkey, bloomreach, main street hub, Optimizely, Infusionsoft, BRIGHT EDGE, INVOCA, Owler, vidyard, CLICKTALE, KENSHOO, yodle, kapost, SAILTHRU, NewsCred, content RAVEN, MOZ, GinzaMetrics, CURALATE, Lattice, hootsuite, unbounce, doubledutch, DELIVERYAGENT, TEALIUM, DYNAMIC YIELD, Spredfast, janrain, sprinklr, infer, Zuberance, MailChimp, Simply Measured

Sales

clearslide, insidesales.com, xactly, toutapp, ConnectAndSell, clari, hearsaysocial, hoopla, RelateIQ, WorkinBox, nimble, Steelwedge, insightly, InsideView, Yesware, BOXER, SUGARCRM, pipedrive, selligy, Badgeville, CIRRUSPATH

Service & Support

Intercom, TOA, NEWVOICEMEDIA, gainsight, Lithium, TOTANGO, liveops, MEDALLIA, Preact, BLUENOSE, satisfaction, servicemax, uservoice

Finance

Adaptive Insights, zenpayroll, anaplan, FRESHBOOKS, aria, hostanalytics, kyriba, Avalara, Chargify, Intacct, Bill.com, BLACKLINE, coupa, tidemark, SAASU, zuora, Expensify, wave, FINANCIAL FORCE.COM, Recurly

HR

HireVue, entelo, SmartRecruiters, Parklet, LEVER, ZipRecruiter, greenhouse, evolv, bamboohr, workable, ZENEFITS, Jobvite, sumtotal, SilkRoad, PEOPLEMATTER, REPLICON, jobscience, Resumator, Avature, Maxwell | Health.

Vertical

Healthcare

DocuTAP, ClearCare, Welltok, hcl.com, omada, comprehend, WebPT, AMERICAN WELL, KINNSER, tigertext, CoreCloud, Medeanalytics, PointClickCare, evariant, MODERNIZING MEDICINE, GRAND ROUNDS, doximity, PROPELLER HEALTH, navicure, practicefusion

Education

KNEWTON, INSTRUCTURE, udemy, BETTERLESSON, Clever, Verificient, Desire2Learn, edmodo, topschool, BrightBytes, dreambox, UDACITY, schoology, coursera

Real Estate

HIGHTOWER, 42Floors, Rentlytics, YARDI, PROCORE, property, VIEW THE SPACE, RAINMAKER, appfolio, propertybase, Gehry Technologies

Other

shopify, Anaqua, oxtime, ServiceTitan, MINDBODY, clio, ncino, RSI, ONESTOP, socialtables, Kenandy, RPLY, Bigcommerce, Socrata, ROCKETLAWYER, JUDICATA, StyleSeat, PLEX, nulogy

Collaboration

box, EVERNOTE, 37signals, Alfresco, DESKAWAY, DocuSign, slack, moxie, Wrike, Speakeasy, asana, Dropbox, clarizen, Atlassian, ActiveTask, HIGHTAIL, huddle, Redbooth, sendthisfile, EGNYTE, LiquidPlanner, SugarSync, ZOHO, WatchDox

BI / Analytics

DOMO, Chartbeat, FLURRY, birst, INSIGHTSQUARED, UPSIGHT, SUMALL, pentaho, 1010data, GoodData, mixpanel, Datameer, Localytics, KISSmetrics, kognitio, looker, bime, PivotLink, SpatialKey, visier, Rosslyn Analytics

PaaS

IaaS

Developers

twilio, zapier, SendGrid, PAGERDUTY, Skytap, RIGHTSCALE, Cloudius, cloudera, npm, CrowdFlower, Iron.io, apprenda, NITROUS, PiCloud, Acquia, Hortonworks, thinq, Cloudy IDE, dotCloud, PANTHEON, stripe, Engine Yard, CloudBees, circleci, Sauce LABS, AUTOMATTIC, DigitalOcean

IT Ops

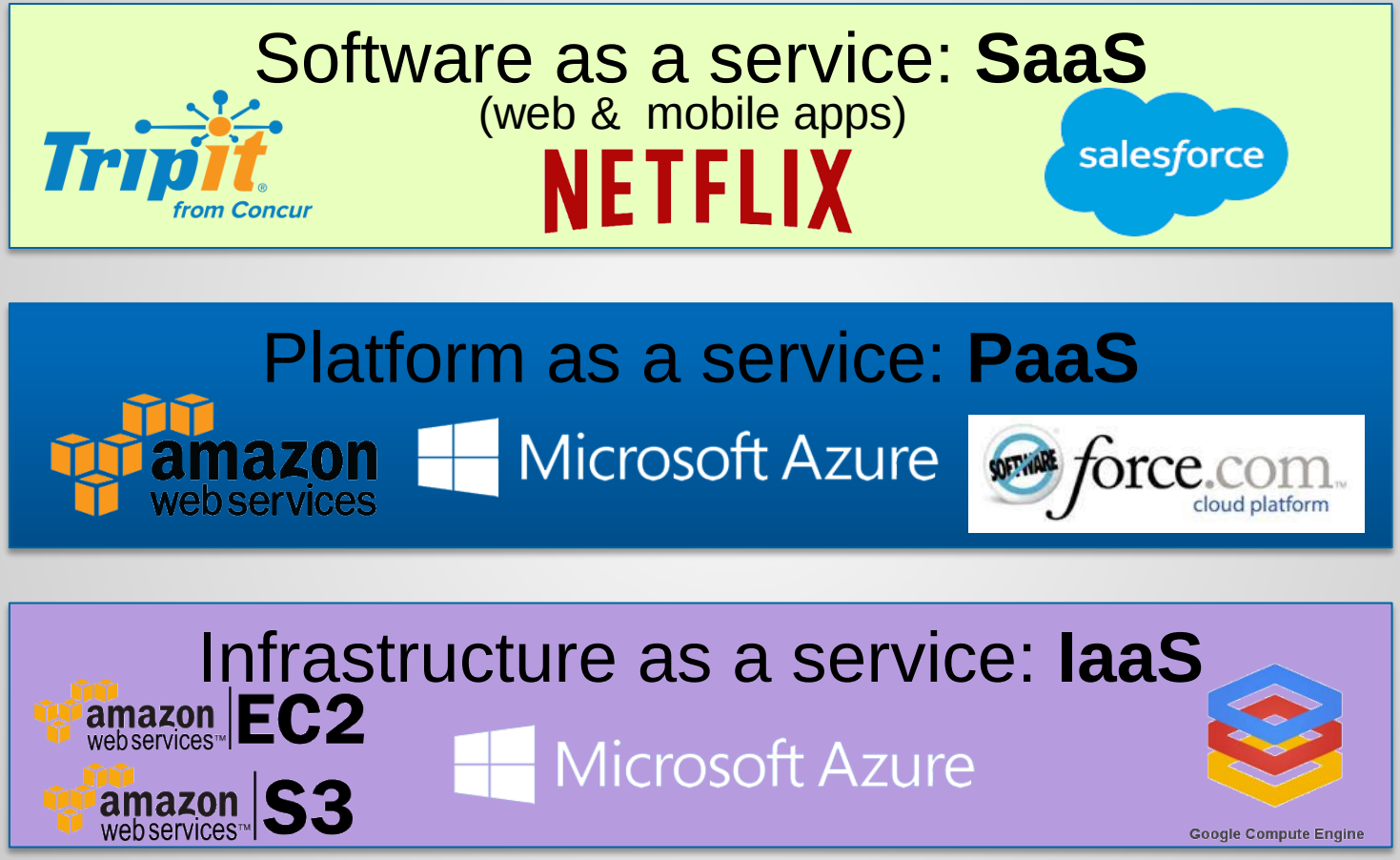
ravello, sumologic, VictorOps, APPERIAN, bounday, actifio, INFINIO, snapLogic, AXCIENT, BlueJeans, APPTIO, cloudshare, MuleSoft, symplified., APPDYNAMICS, Acronis, Zerto, GIGASPACES, EUCALYPTUS, thinking phone networks, SCALEXTREME, New Relic, SPICEWORKS, catchpoint, abiquo, Joyent, CRITTERCISM, SOASTA, veeam, linode.com, nebula, Pistoia, MIRANTIS, ONECLOUD LABS

Security

JID, CloudLock, defense.net, distil networks, CipherCloud, Auth0, MarkLogic, wandera, onelogin, CLOUDFLARE, WhiteHat, RAPID7, Vaultive, Lookout, CloudPassage, VERACODE, MobileIron, loggly, SilverSky, okta

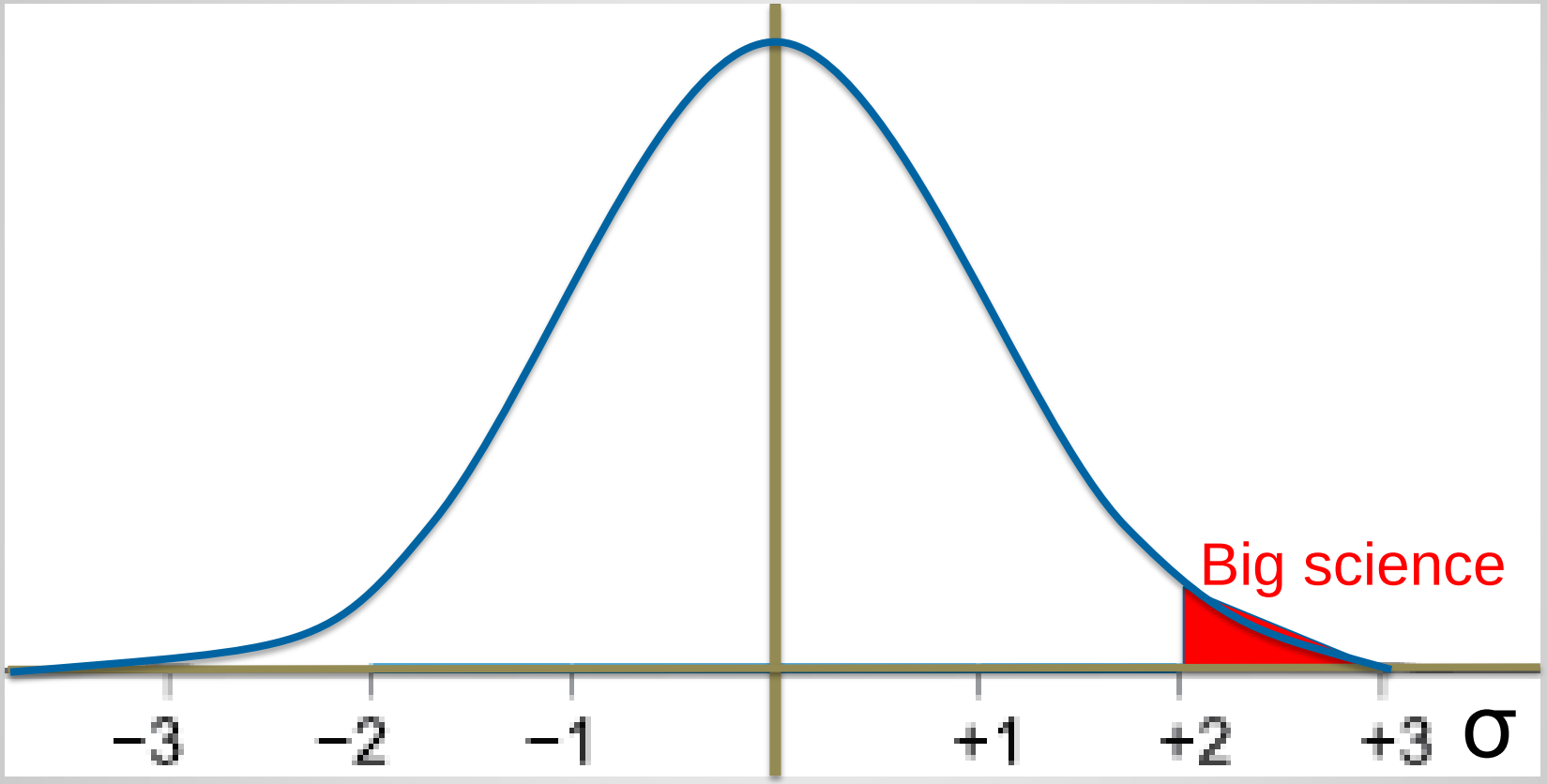
Download a digital copy or nominate your company: bvp.com/cloud

Cloud has transformed how software is developed and delivered



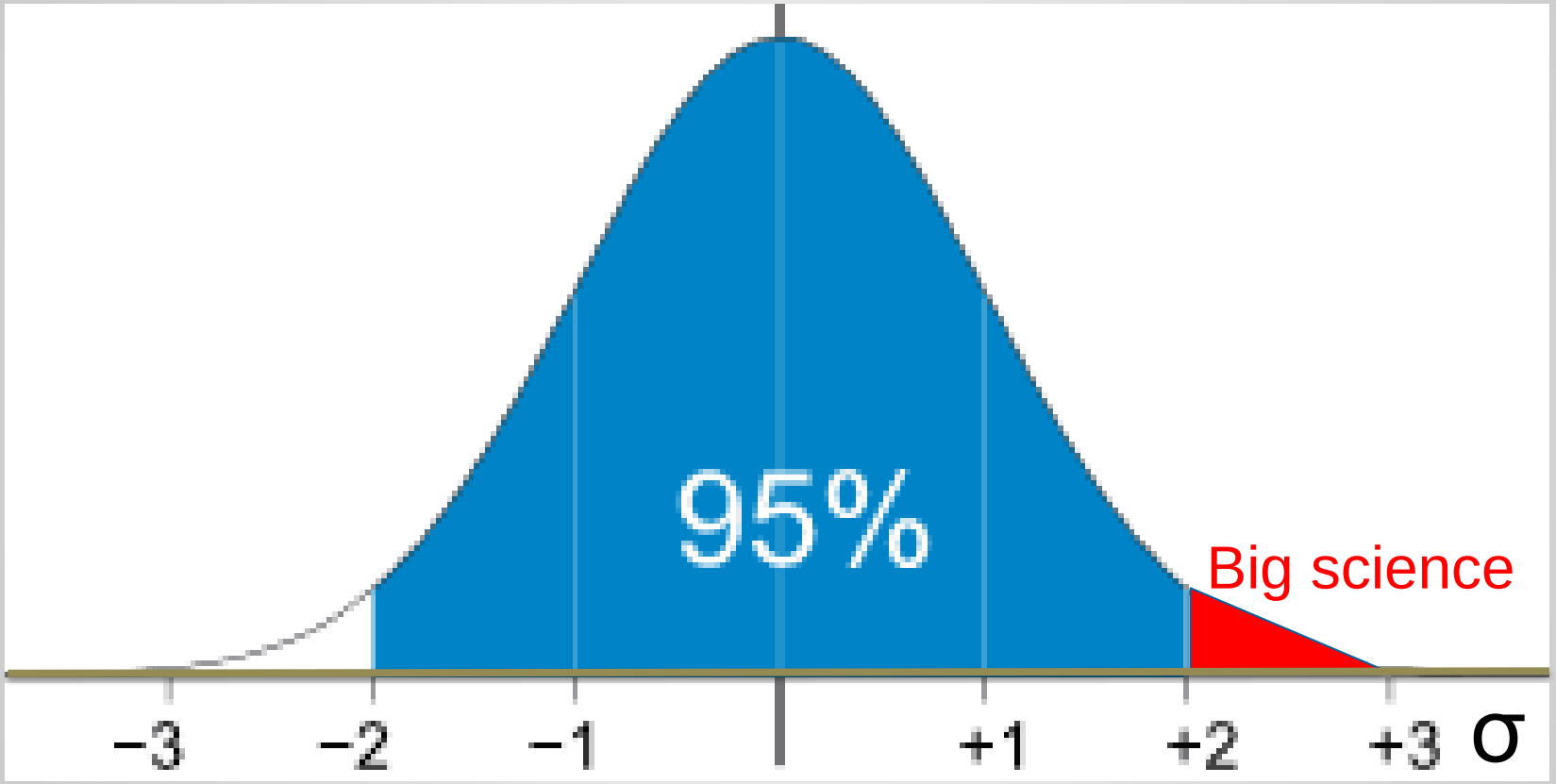
PaaS enables more rapid, cheap, and scalable delivery of powerful apps—as SaaS

The right platform can do the same for science



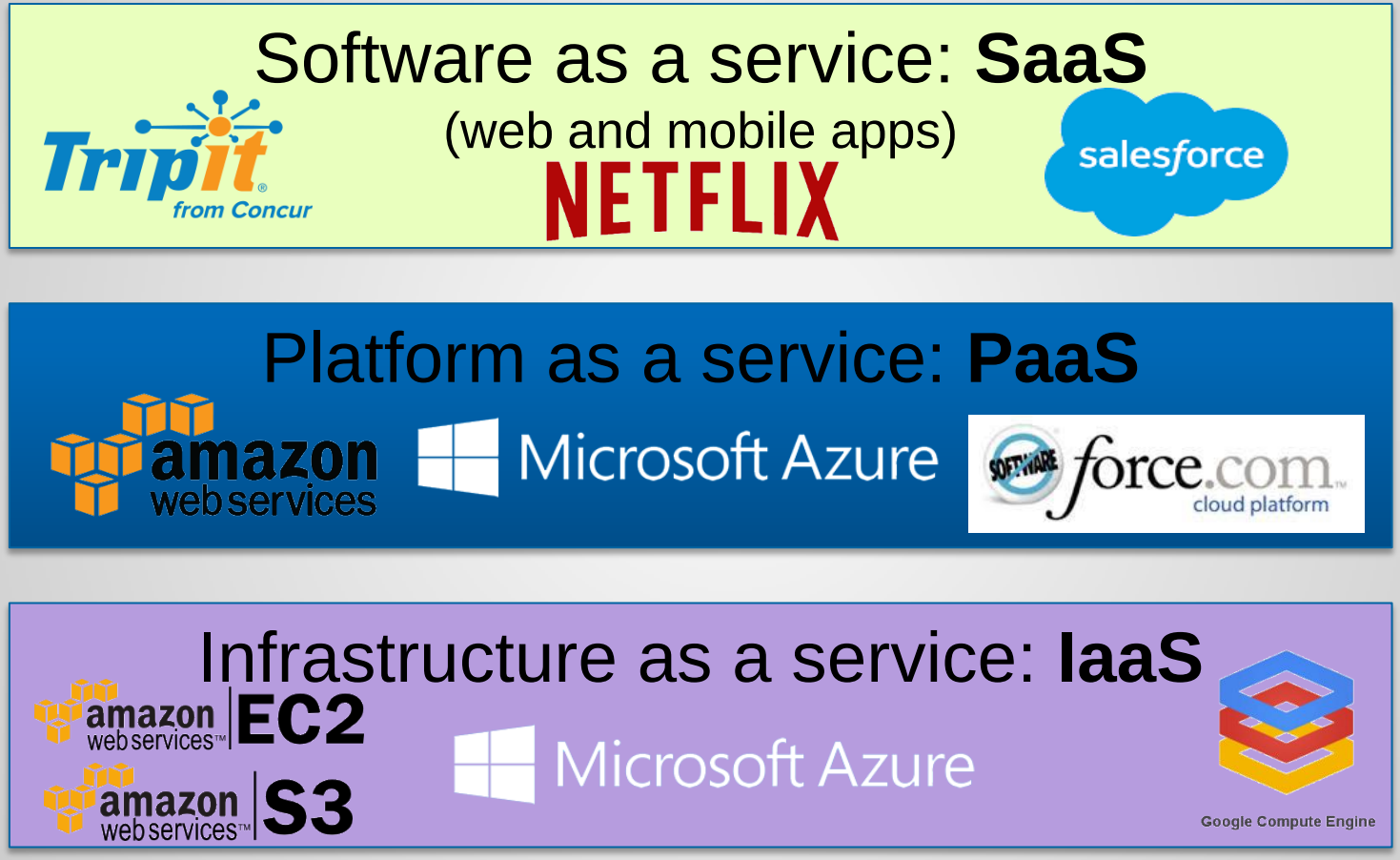
We can leverage cloud to provide solutions that span the vast majority of researcher needs

The right platform can do the same for science



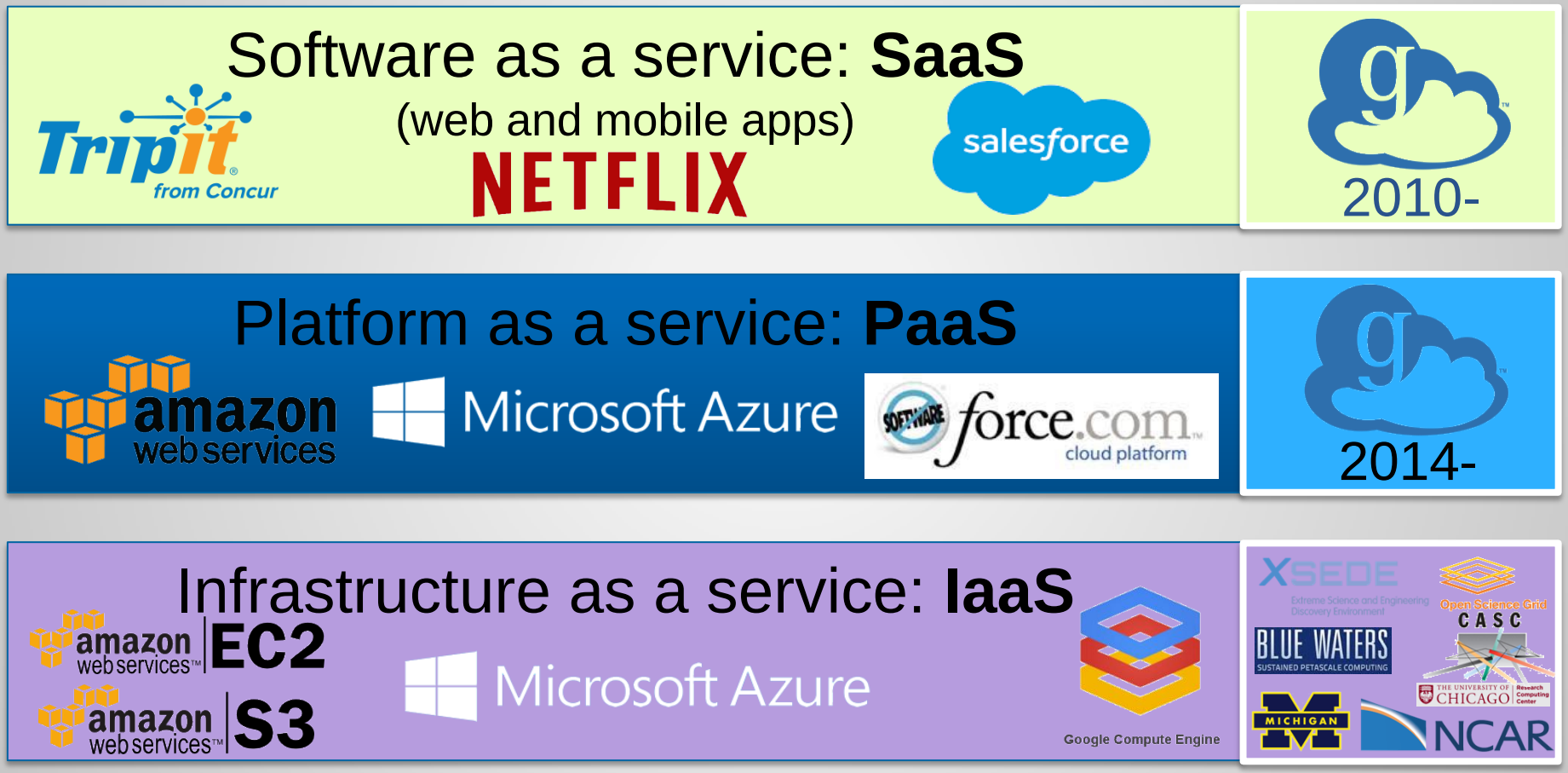
We can leverage cloud to provide solutions that span the vast majority of researcher needs

A science platform can spur a discovery cloud ecosystem



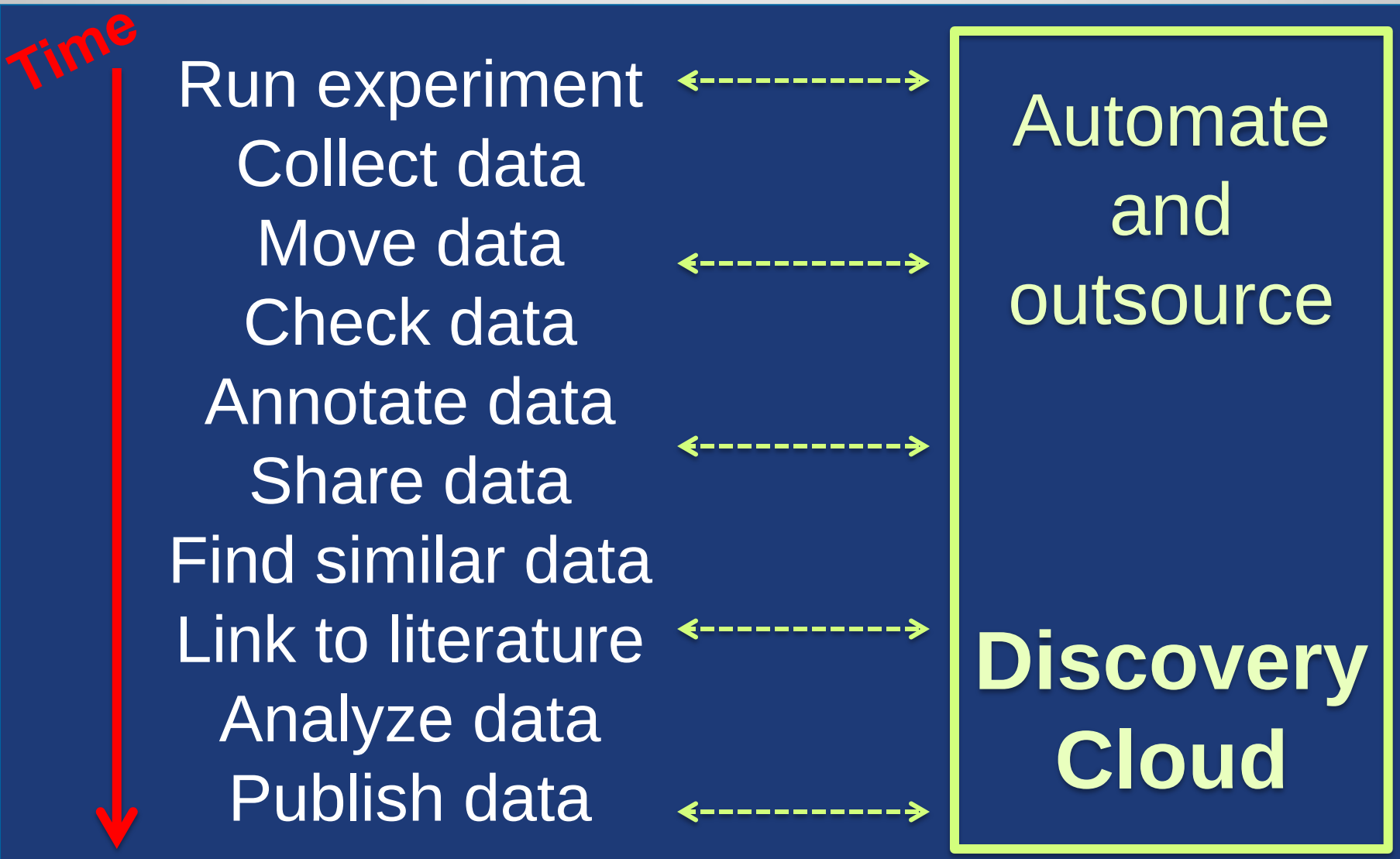
In so doing, we can slash costs, improve quality, and accelerate discovery across the sciences

A science platform can spur a discovery cloud ecosystem

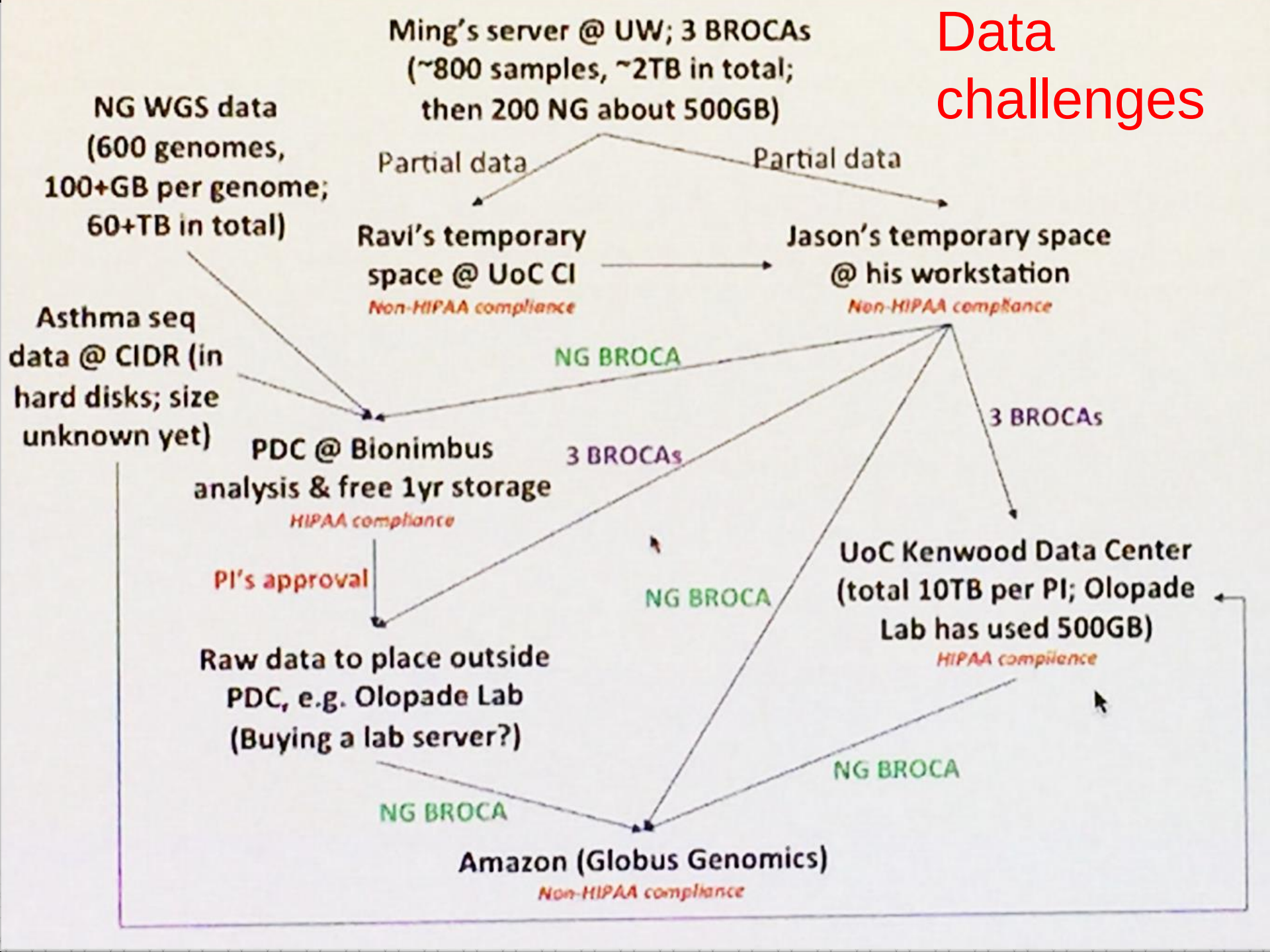


In so doing, we can slash costs, improve quality, and accelerate discovery across the sciences

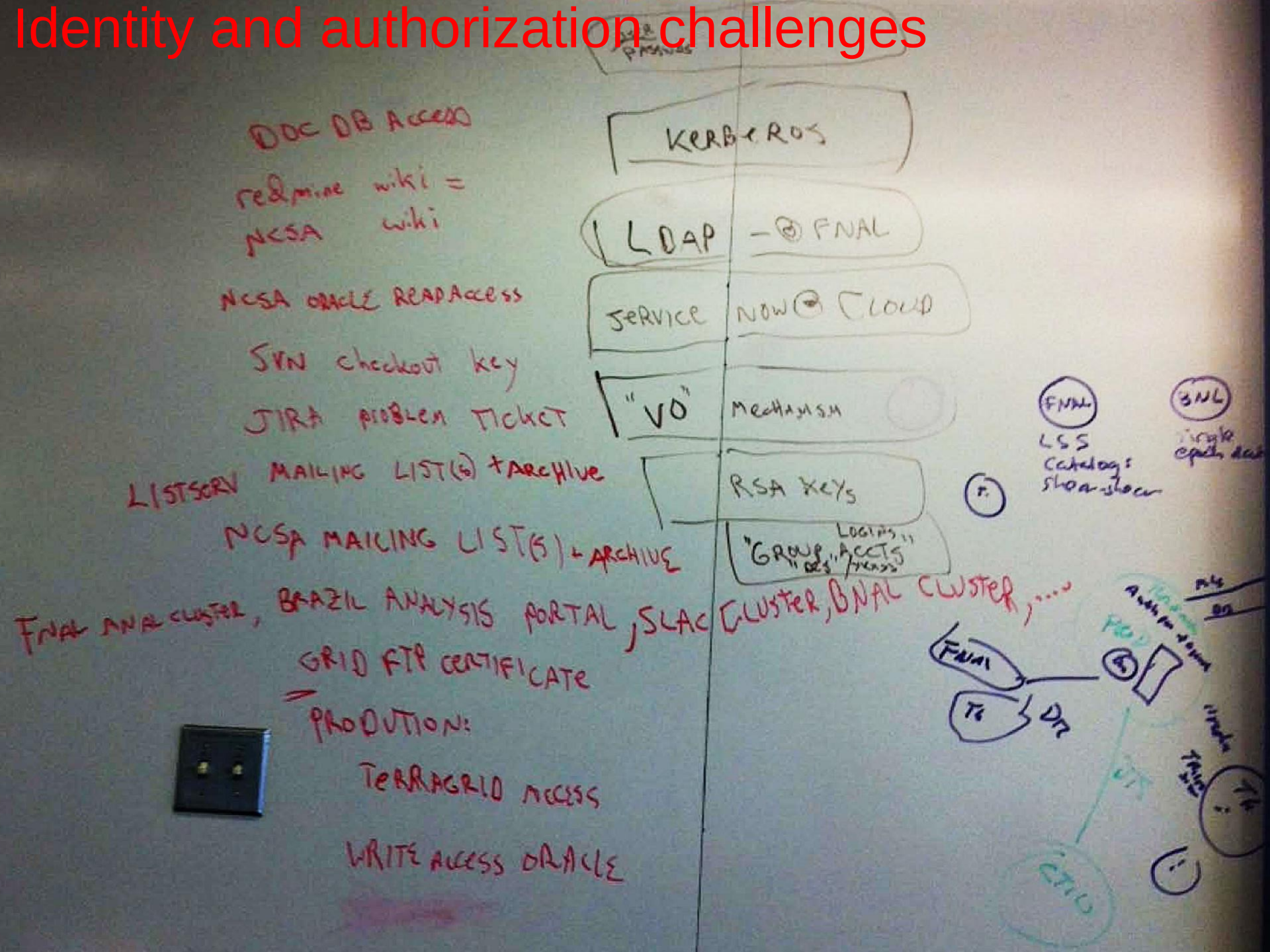
What can we automate and outsource in science?



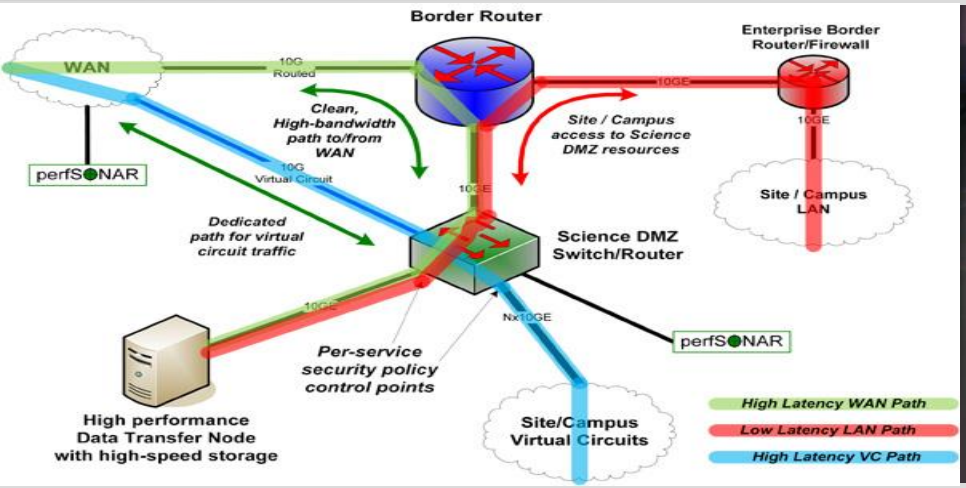
Data challenges



Identity and authorization challenges



Data access: we have the highways but not the delivery service



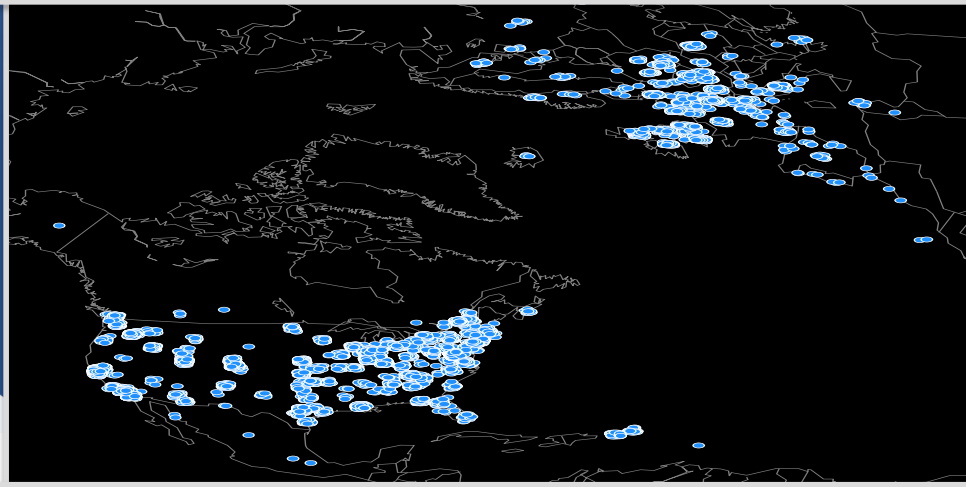
Our highways encompass the Internet, ultra-high-speed networks, science DMZs, data transfer nodes, high-speed transport protocols

A good **delivery service** automates, schedules, accelerates, adapts.

It provides APIs for experts and casual users.

Cuts costs and saves time.

Globus: Research data management as a service



Essential research data management services

- File transfer
- Data sharing
- Data publication
- Identity and groups

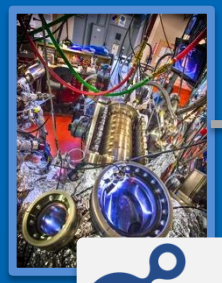
Builds on 15 years of research

Outsourced and automated

- High availability, reliability, performance, scalability
- Convenient for
  - Casual users: Web interfaces
  - Power users: APIs
  - Administrators: Install, manage

Globus and the research data lifecycle

Instrument



Globus transfers files reliably, securely

2

Compute Facility

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

7 Curator reviews and approves; data set published on campus or other system

Transfer

3

Share

1

Researcher initiates transfer request; or requested automatically by script, science gateway

Researcher selects files to share, selects user or group, and sets access permissions

6

Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

globus connect

Publication Repository

5

Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus

globus connect

Personal Computer

Publish

6

8

Peers, collaborators search and discover datasets; transfer and share using Globus

Discover

- **SaaS → Only a web browser required**
- **Use storage system of your choice**
- **Access using your campus credentials**

Globus by the numbers

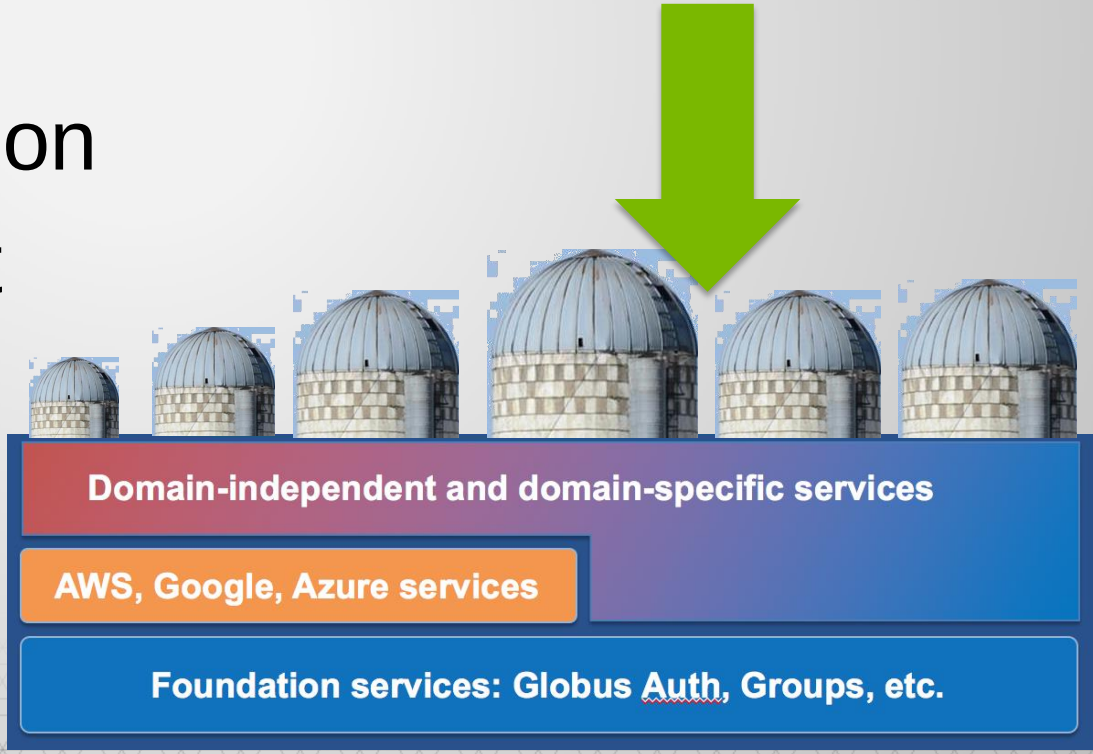
4 major services	**160 PB** transferred	**25 billion** files processed	**38,000** registered users
13 national labs use Globus	**10,000** active endpoints	**~400** active daily users	**99.9%** uptime
35+ institutional subscribers	**1 PB** largest single transfer to date	**3 months** longest continuously managed transfer	**130** federated campus identities

Platforms can slash costs, simplify access, increase interoperability

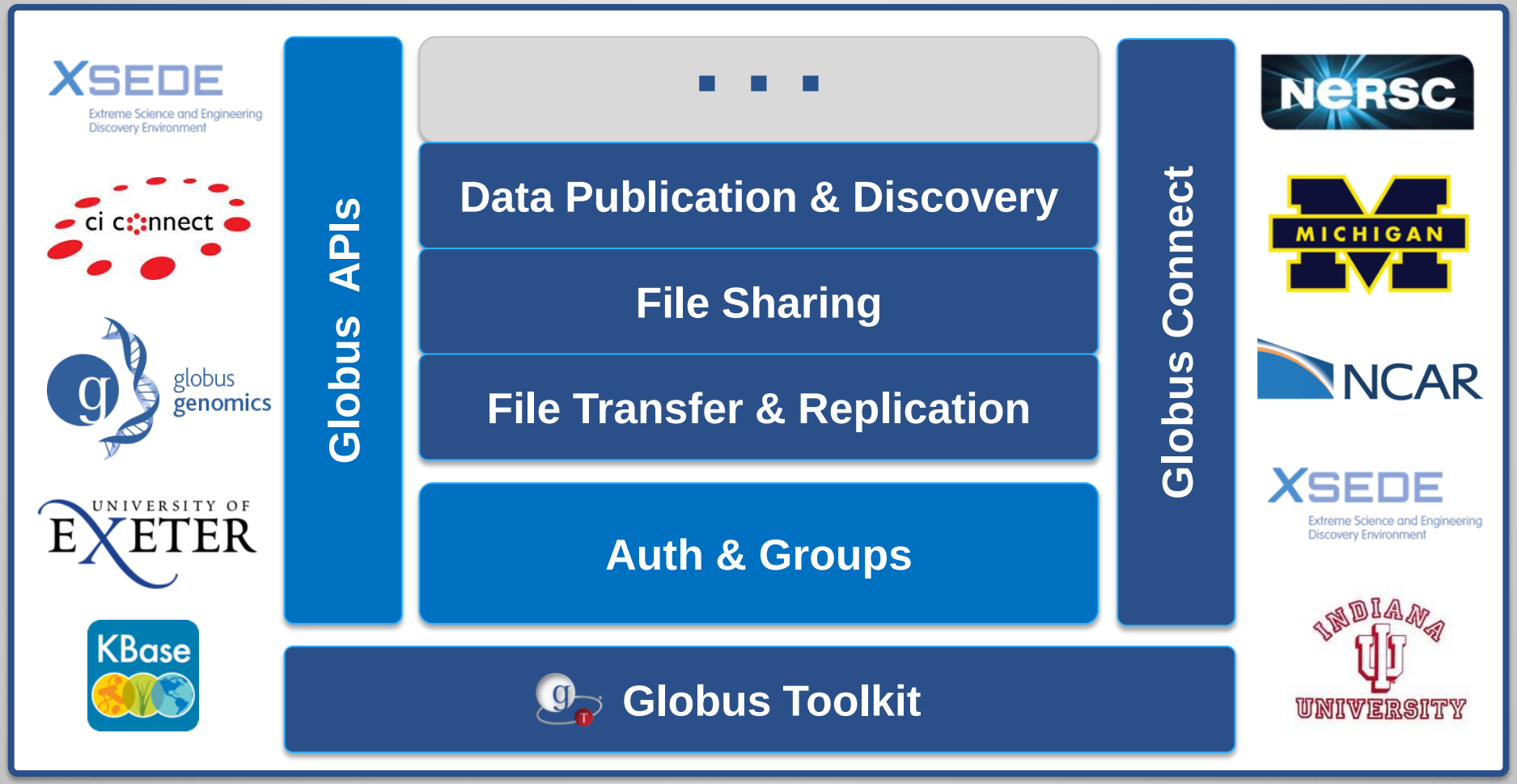
For example, by providing:

- Federated identity system with fine-grained authorization
- Data management easily integrated with application workflows

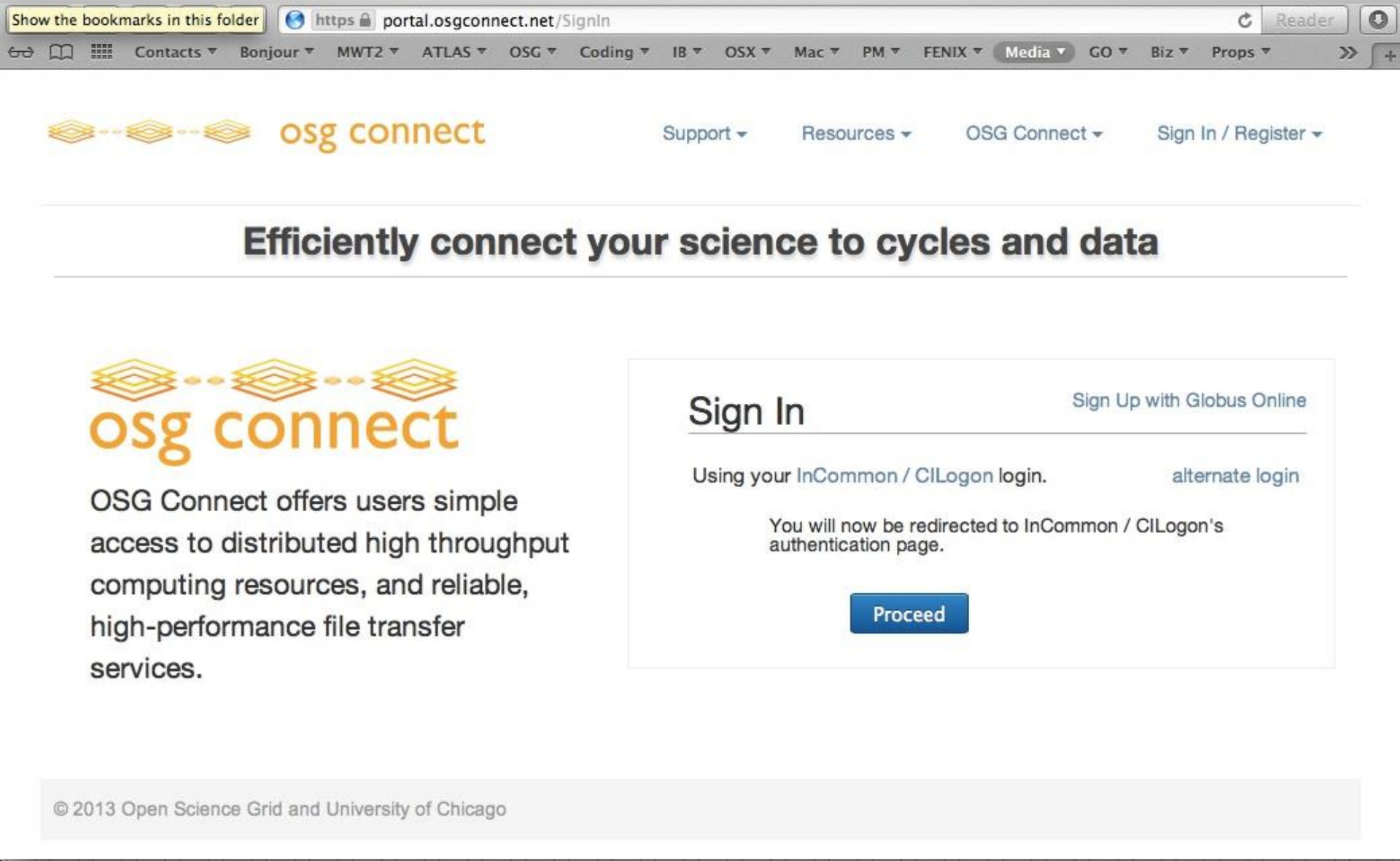
Via RESTful APIs



Globus PaaS: Ecosystem enabler



Globus PaaS and Open Science Grid



Simple web app server login

KBase
PREDICTIVE BIOLOGY

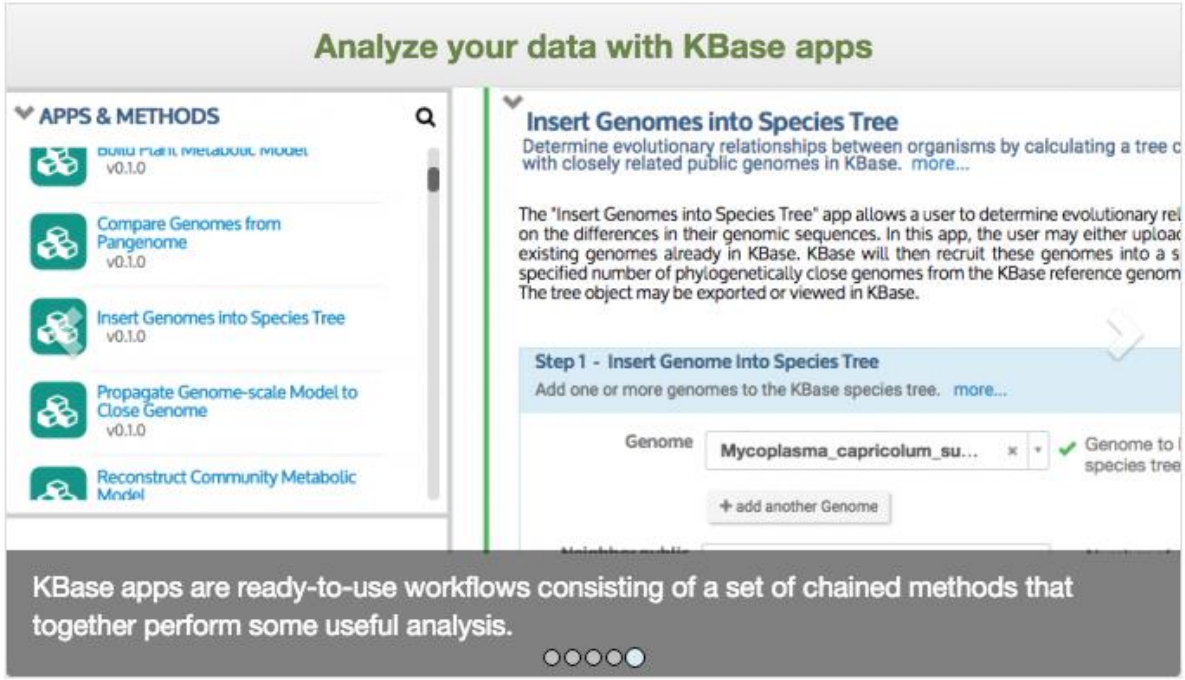
About ▾ Data & Tools ▾ Docs ▾ Help ▾

Sign Up Sign In

Search our site

Maintenance Window - February 13, 2016 in 2 days — Sat Feb 13 from 10:00am to 3:00pm

KBase: The Department of Energy Systems Biology Knowledgebase



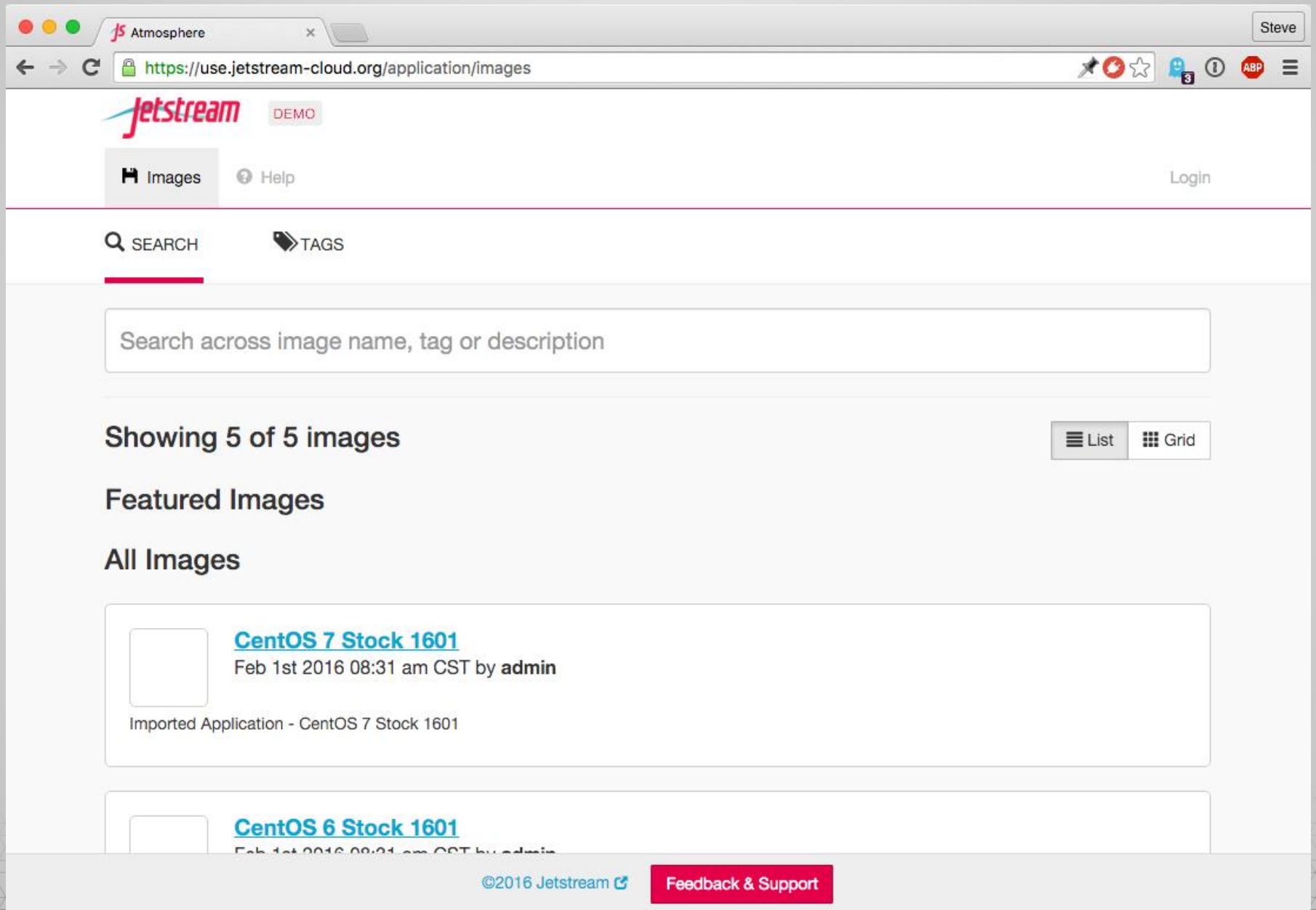
New to KBase?

Search Data

Sign In

KBase is an open platform for comparative functional genomics and systems biology for microbes, plants and their communities, and for sharing results and methods with other scientists.

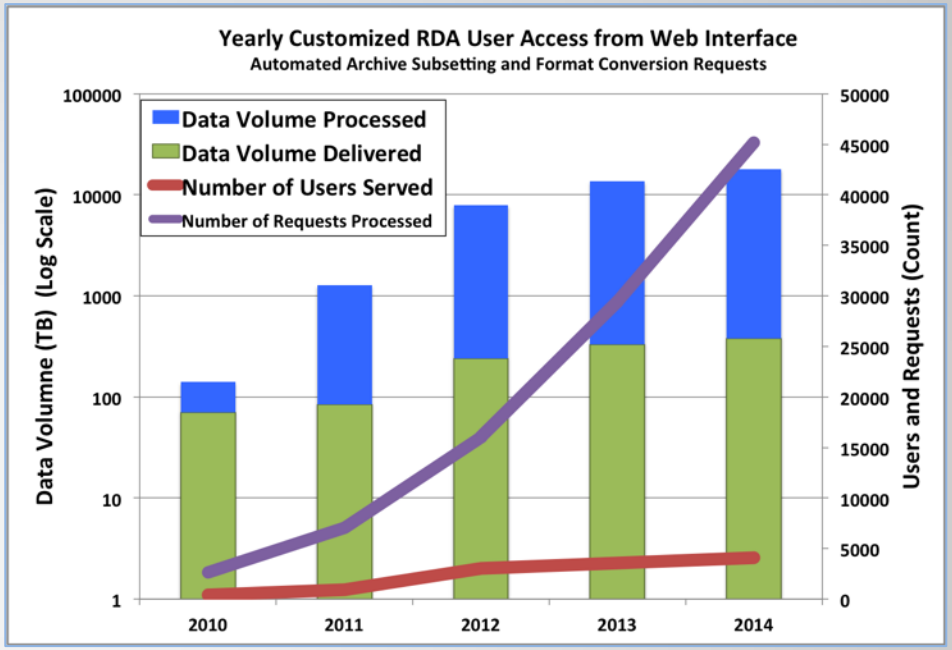
Jetstream cloud service



Serving a global community: NCAR's Research Data Archive

Serving a global community

- 17+ PB virtual processing
- 45,000+ custom orders, 4,000 users, 380 TB served in 2014



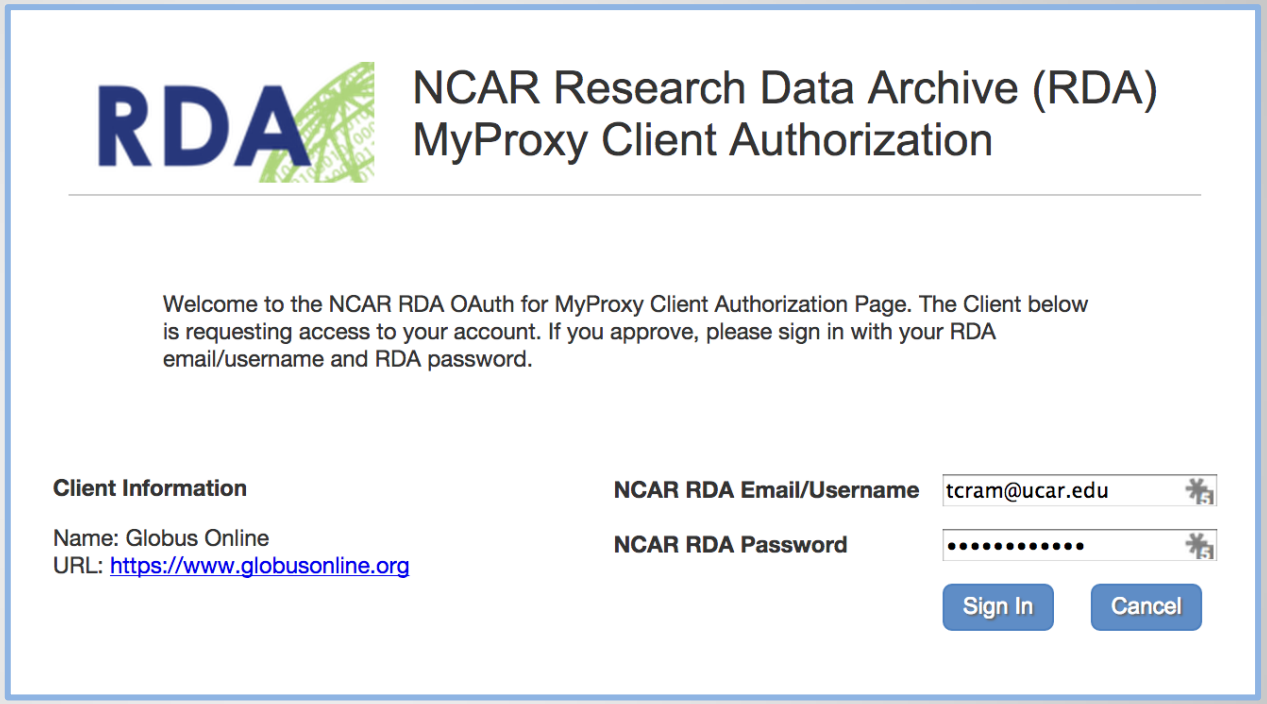
Courtesy of Thom[illegible]m, NCAR (2014)

Fully automated delivery using portal developed w/ PaaS

PaaS enabled automated workflow

User logs in with NCAR or other campus identity

Selected dataset copied to staging area (shared endpoint)



User granted read permission for shared endpoint

User receives email with link to access files

ACLs deleted after five days

Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the Wellcome Trust Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click here to learn more and follow us on Twitter.

Before you start

Be sure to read through the instructions.

You will need to set up a free account with Globus and have Globus Connect running at your institute or on your computer to transfer files to and from the service.

Ready to start?

If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the about page.

Full name

Organisation

Email address

What is this

Globus user identity

Next

News

@sangerimpute

11/05/2016
Thanks to EAGLE, we can now return **phased data**. The HRC panel has been updated to r1.1 to fix a known issue. See ChangeLog for more details.

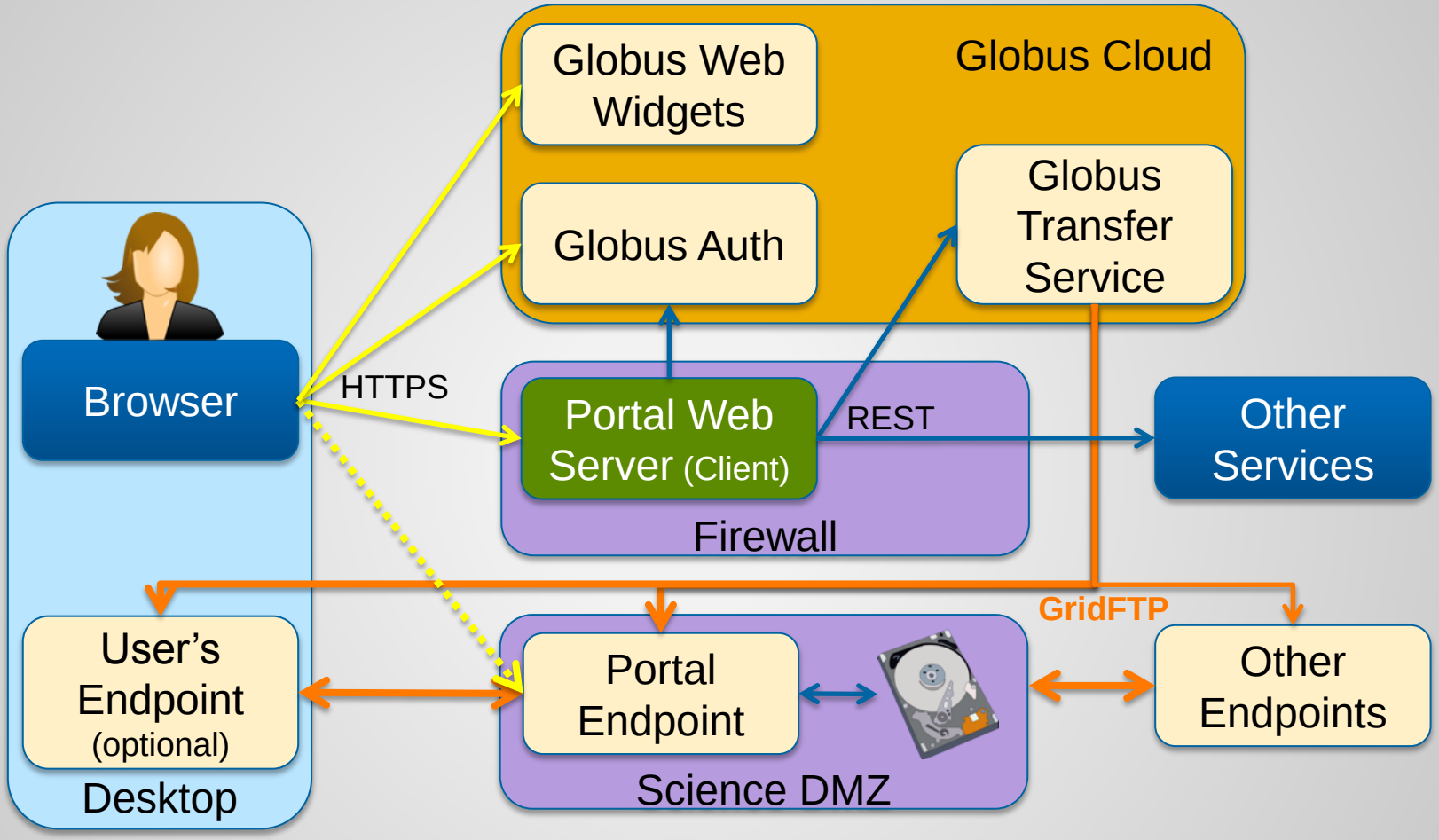
15/02/2016
Globus API changed, please see updated instructions.

17/12/2015
New status page and reworked internals. See ChangeLog.

09/11/2015
Pipeline updated to add some features requested by users. See ChangeLog.

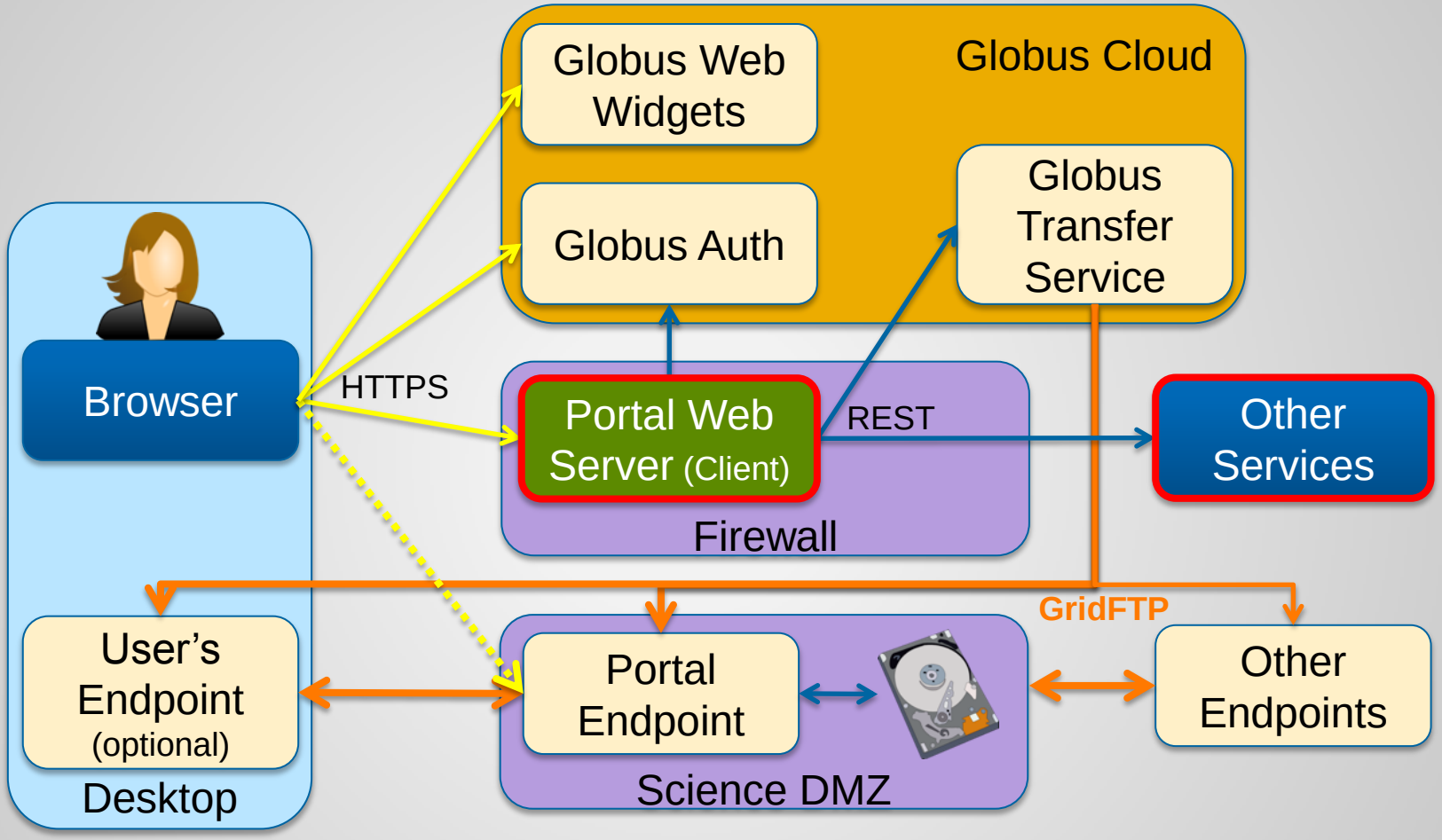
See older news…

Research data portal pattern



- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Research data portal pattern



- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

https://github.com/globus/globus-sample-data-portal

Globus Transfer API

API reference for transfer and sharing functions.

Globus Auth API

API reference for authentication and authorization.

Frequently Asked Questions

When all else fails...

globusworld 2016
APRIL 20-22
CHICAGO

Developer Workshop: Building the Modern Research Data Portal

New high-speed networks make it possible, in principle, to transfer and share research data at tremendous speeds and scales–but have also proved challenging to use in practice. Two new technologies now allow us to translate this potential into reality: Science DMZ architectures provide frictionless end-to-end network paths; and Globus APIs allow programmers to create powerful research data portals that leverage these paths for data distribu synchronization, and other useful purposes.

Introduction, Concepts, and Components

IMPERIAL 2

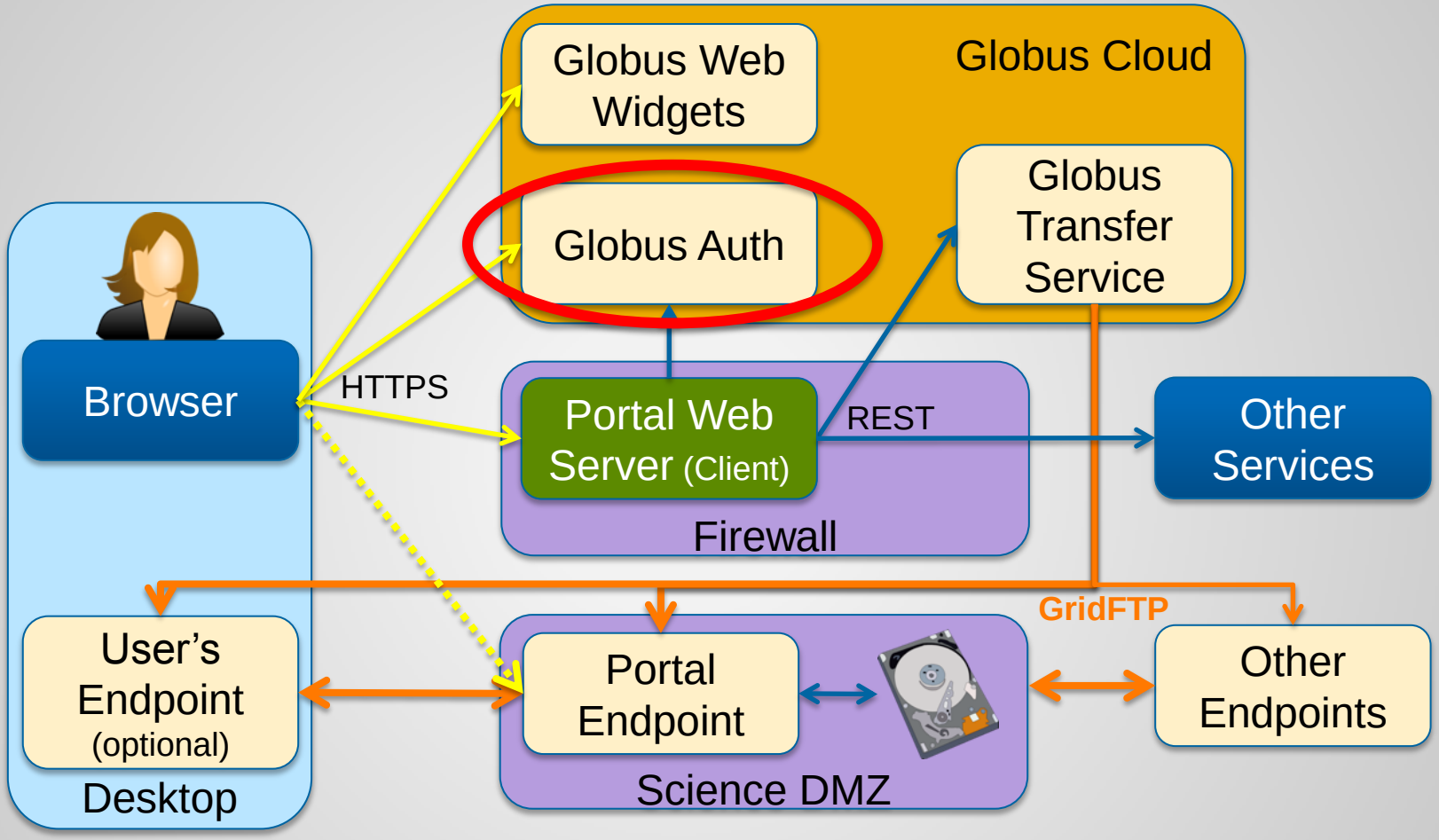
Led by: **TBD**

We will introduce the Modern Research Data Portal and set the context for how Globus and the ScienceDMZ combine to deliver unique data management capabilities. This will include:

- Overview of use cases: Common patterns like data publication/distribution, orchestration of data flows, etc.
- Overview of the Globus platform: Architecture and brief overview of available services
- Introduction to the Globus Auth API: Authenticating and authorizing a client
- Introduction to the Globus Transfer API: Make your first call and move data with Globus
- Introduction to the Python SDK for using Globus Auth and Transfer

Let us know if you'd like to participate in future workshops

Research data portal pattern



- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Globus Auth

- Foundational identity and access management (IAM) platform service
- Simplify creation and integration of advanced apps and services
- Brokers authentication and authorization interactions between:
  - End-users
  - Identity providers: XSEDE, InCommon, web apps
  - Resource servers: services with REST APIs
  - Clients: web, mobile, desktop, command line apps
  - Resource servers acting as clients to other resource servers

https://docs.globus.org/api/auth

Based on widely used web standards

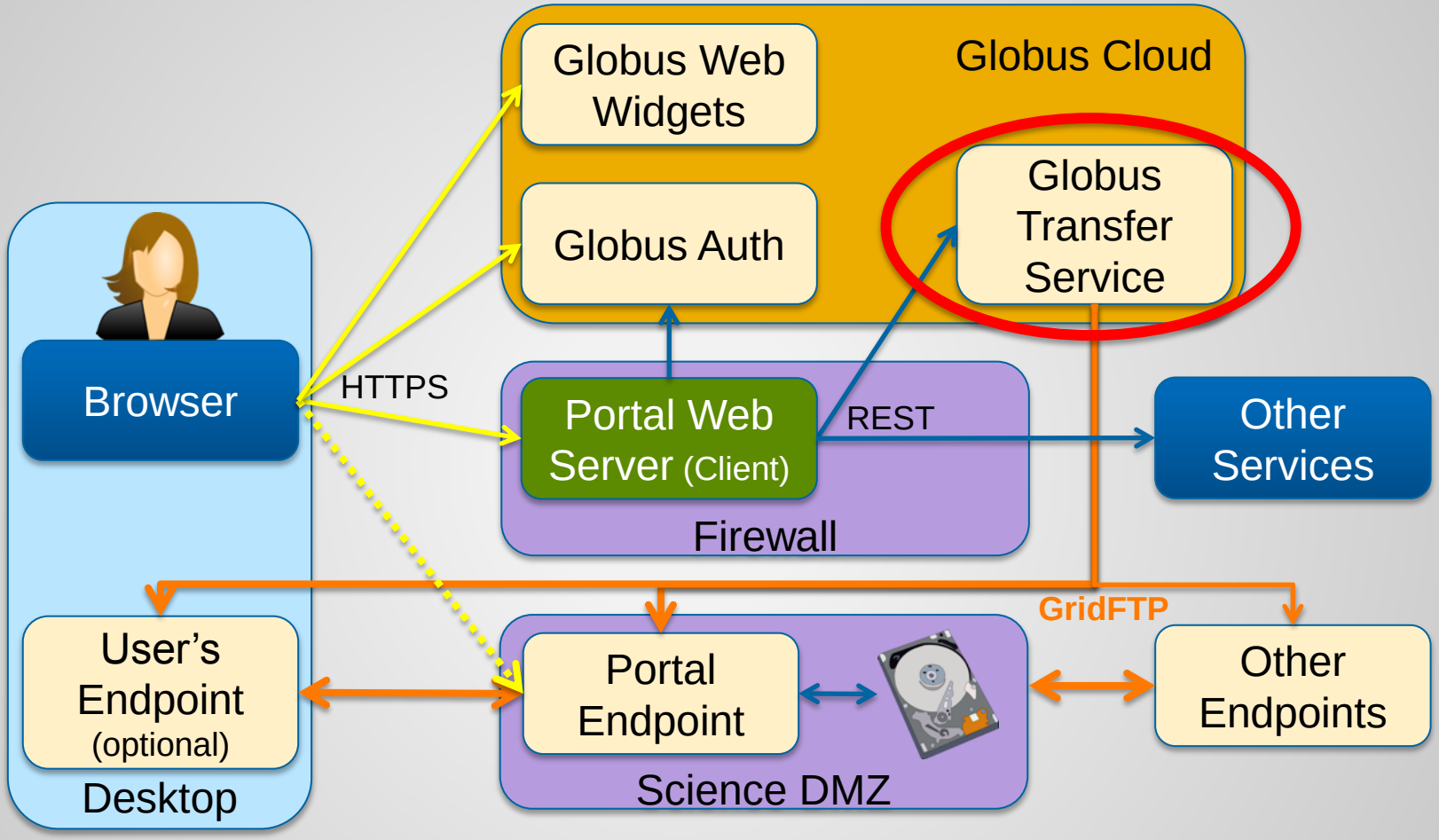
- OAuth 2.0 Authorization Framework
  - aka OAuth2
- OpenID Connect Core 1.0
  - aka OIDC

- Allows use of standard OAuth2 and OIDC libraries
  - E.g., Google OAuth Client Libraries (Java, Python, etc.), Apache mod_auth_openidc

Globus Auth uses

- Login to web app
  - “Log in with Globus”
  - Mobile, desktop, command line apps coming
- Protect all REST API communications
  - App → Globus service
  - App → non-Globus service
  - Service → service

Research data portal pattern

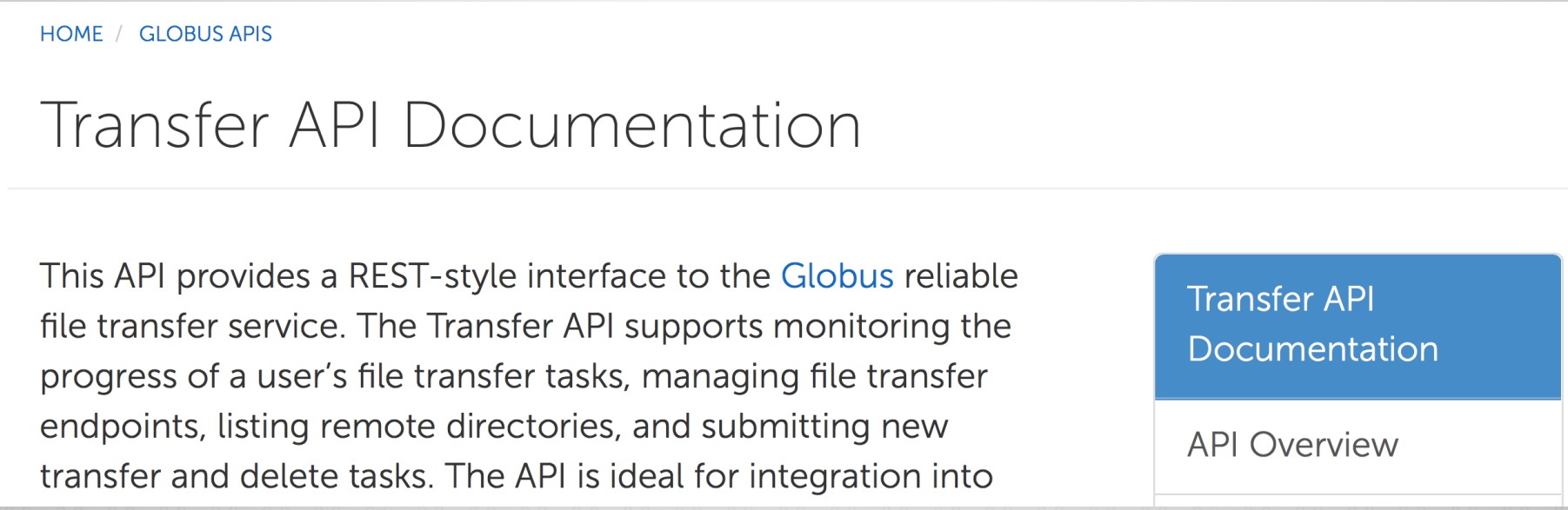


- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Globus transfer API

Nearly all Globus Web App functionality implemented via public Transfer API

https://docs.globus.org/api/transfer/



Globus Python SDK

Python client library for the Globus Auth and Transfer REST APIs

http://globus.github.io/globus-sdk-python/

Globus SDK for Python (Beta)

This SDK provides a convenient Pythonic interface to Globus REST APIs, including the Transfer API and the Globus Auth API. Documentation for the REST APIs is available at https://docs.globus.org.

Two interfaces are provided - a low level interface, supporting only `GET`, `PUT`, `POST`, and `DELETE` operations, and a high level interface providing helper methods for common API resources.

Source code is available at https://github.com/globus/globus-sdk-python.

Installation

The Globus SDK requires Python 2.6+ or 3.2+. If a supported version of Python is not already installed on your system, see this Python installation guide .

The simplest way to install the Globus SDK is using the `pip` package manager (https://pypi.python.org/pypi/pip), which is included in most Python installations:

```
pip install globus-sdk
```

Jupyter (iPython) notebooks

https://github.com/globus/globus-jupyter-notebooks

Globus SDK

https://github.com/globus/globus-sdk-python

Globus SDK Docs

http://globus.github.io/globus-sdk-python/

Requirements

- You need to be in the tutorial users group for sharing: https://www.globus.org/app/groups/50b6a29c-63ac-11e4-8062-22000ab68755
- Installed Globus Python SDK

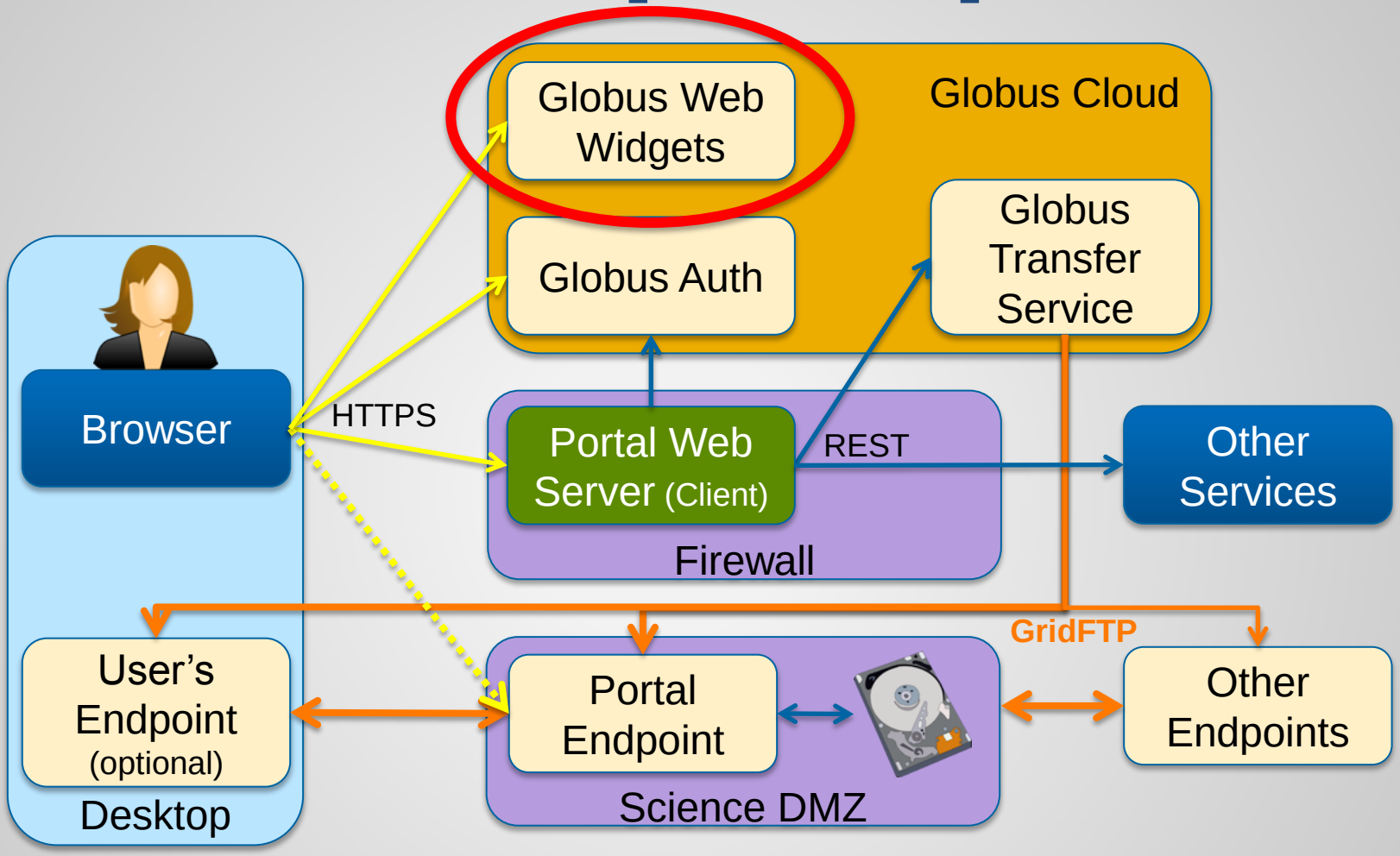
In [15]:
```
from __future__ import print_function # for python 2
tutorial_endpoint_1 = "ddb59aef-6d04-11e5-ba46-22000b92c6ec"      # endpoint "Globus Tutorial Endpoint 1"
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c6ec"      # endpoint "Globus Tutorial Endpoint 2"
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755"     # group "Tutorial Users"
```

Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can obtain access tokens via the tokens.globus.org website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and copy the resulting text into ~/.globus.cfg.

In [16]:
```
transfer_token = None # if None, tries to get token from ~/.globus.cfg file
```

Research data portal pattern

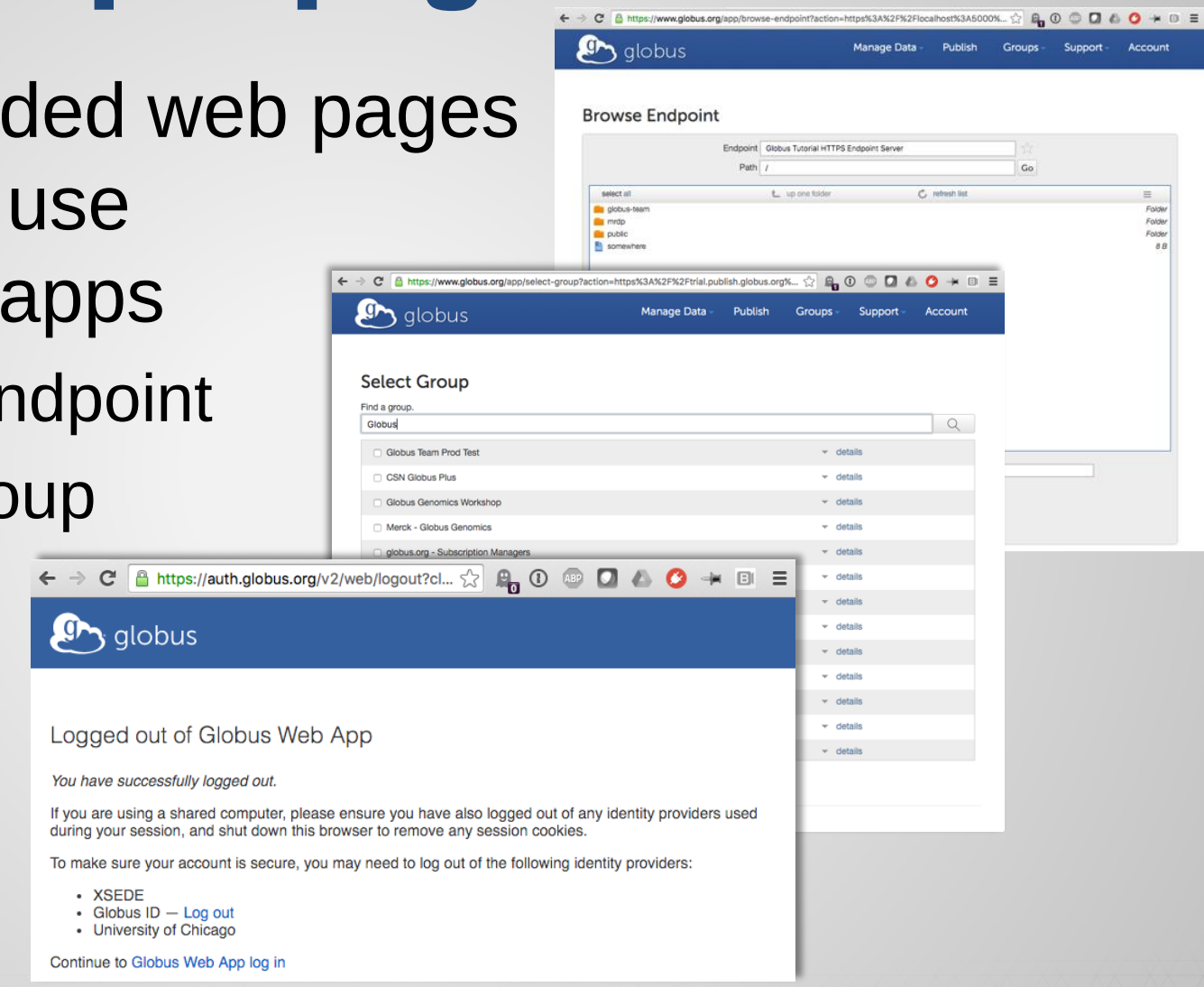


- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Globus helper pages

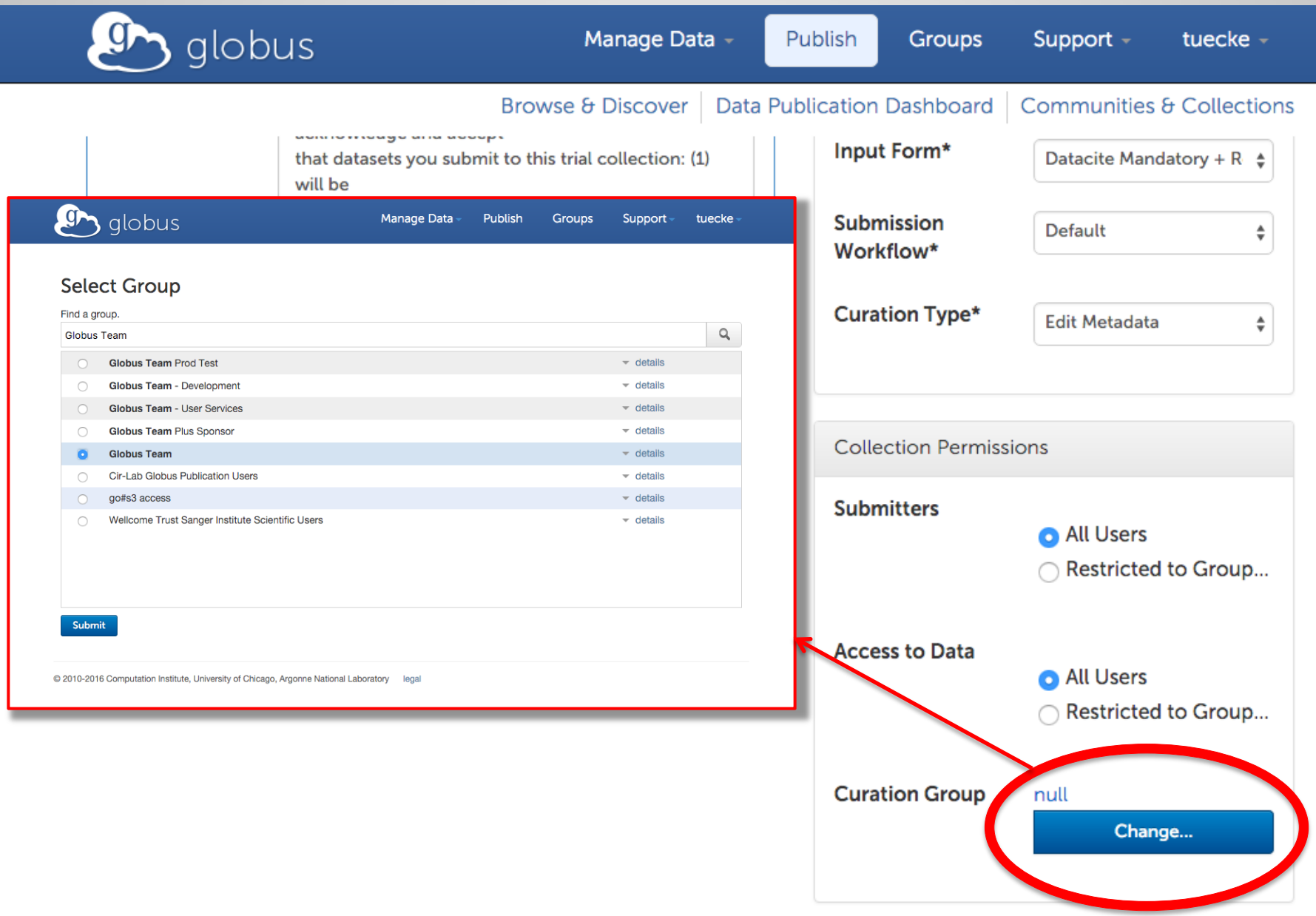
Globus-provided web pages designed for use by your web apps

- Browse Endpoint
- Select Group
- Logout



https://docs.globus.org/api/helper-pages/

Globus helper pages



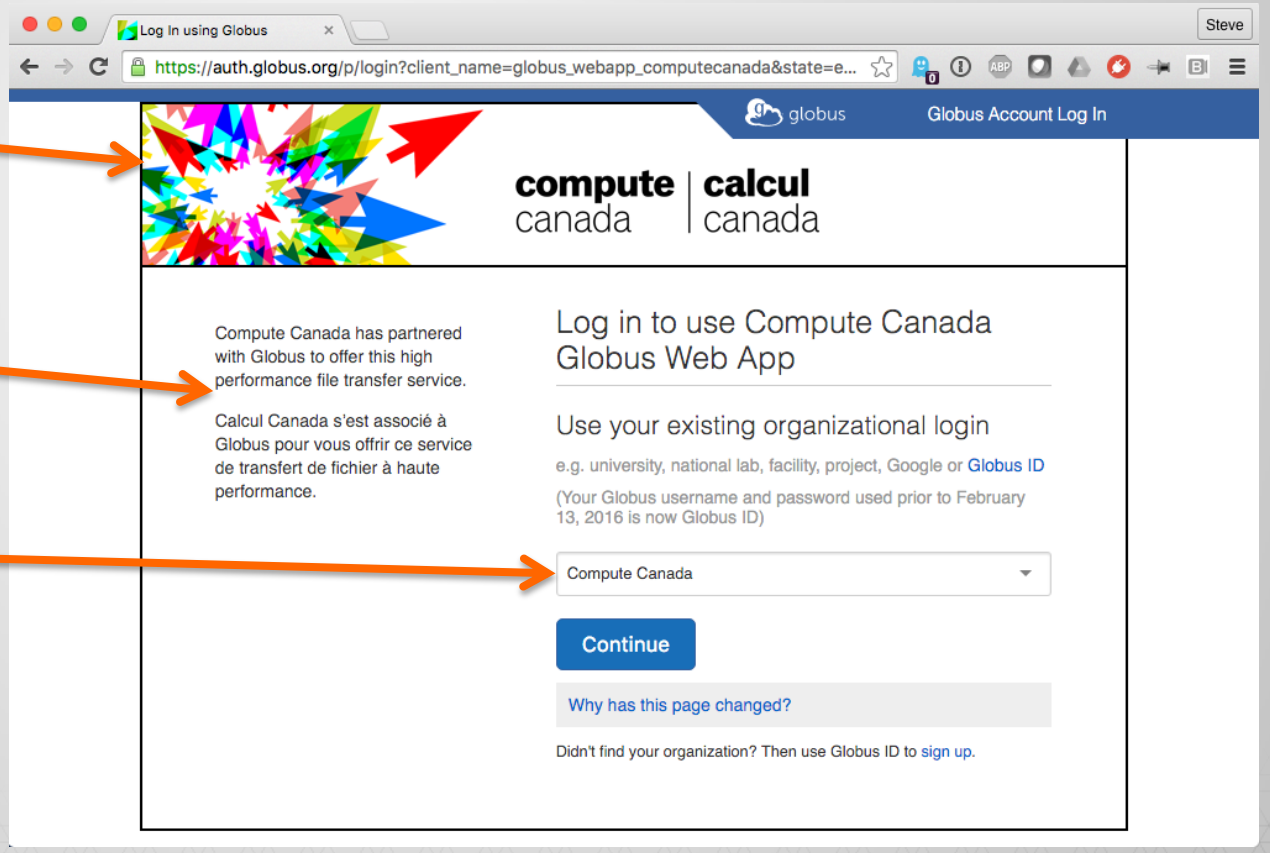
Branding

Can skin Globus Auth pages

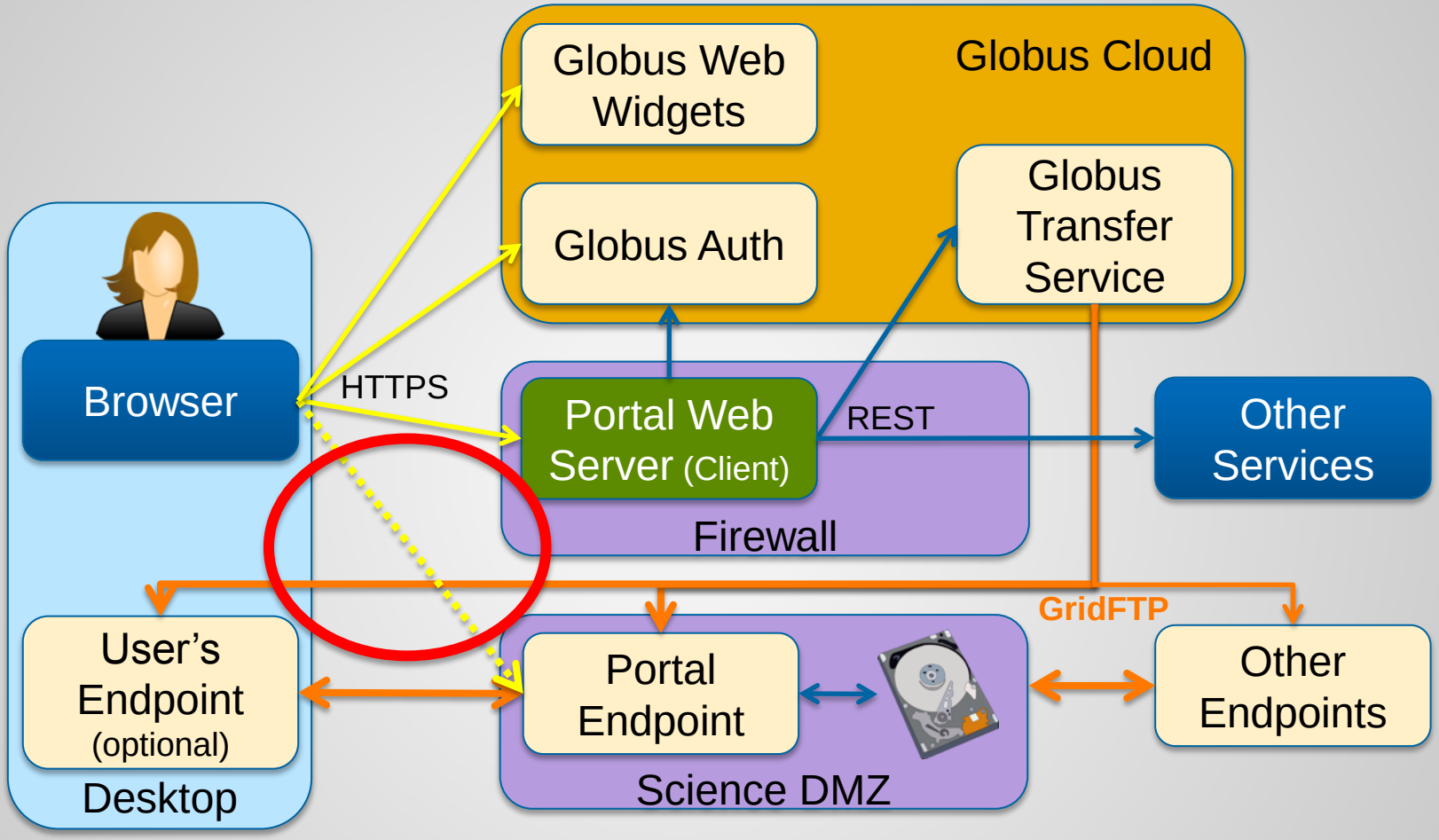
Header

Text

Default IDP



Research data portal pattern

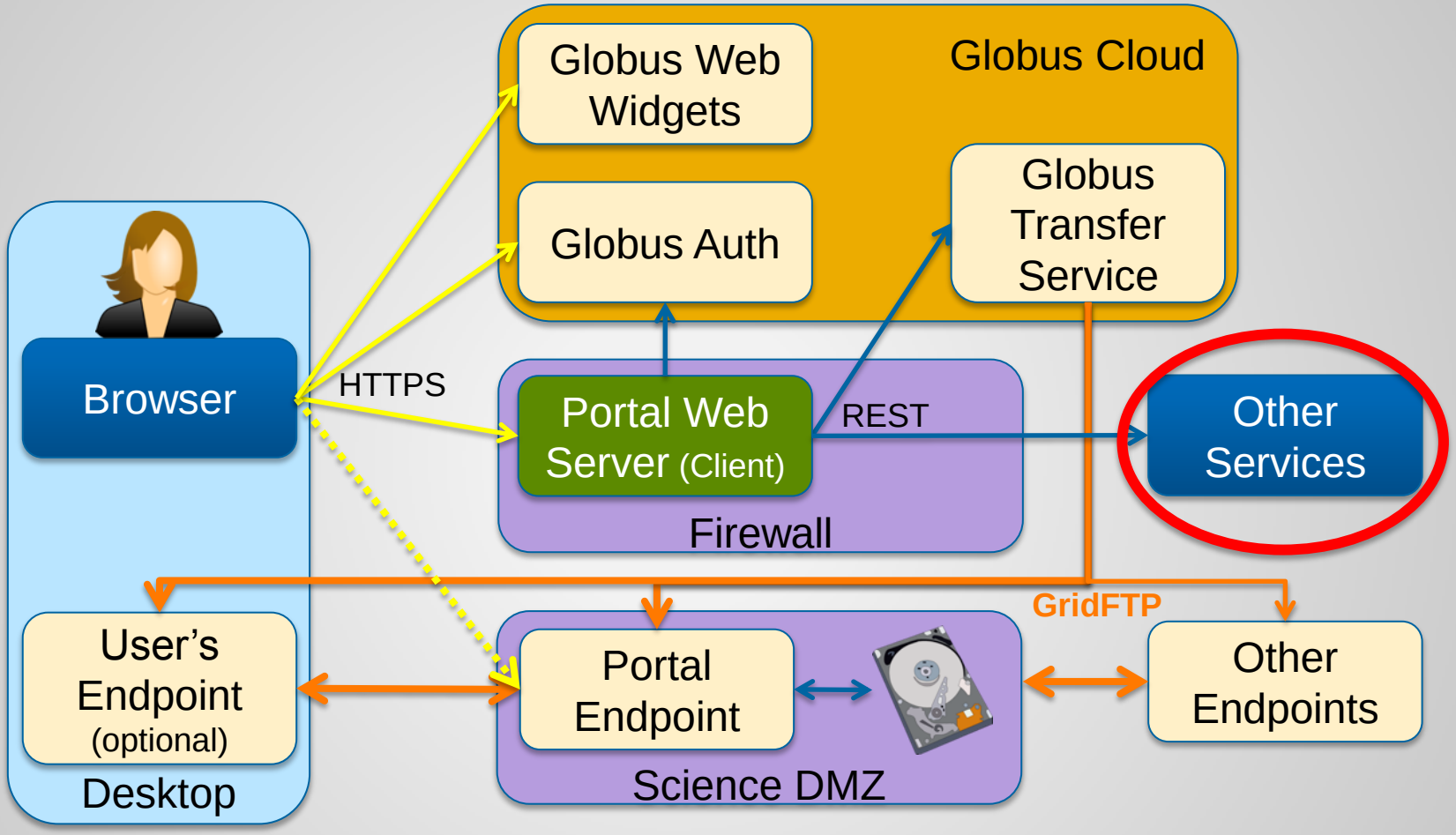


- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Globus Connect HTTPS

- The future of research CI is … the web
- Globus Connect HTTPS unlocks all research storage to the web
- Globus Auth provides security glue using standard web security
- GridFTP doesn’t go away – async, bulk data transfer is important, but its not the end-all, be-all

Research data portal pattern



- Move portal storage into Science DMZ, with Globus endpoint
- Leave Portal Web server behind firewall
- Globus handles the security and data heavy lifting

Why create your own services?

- Front-end / back-end within your portal
  - Remote backend for portal
  - Backend for pure Javascript browser apps

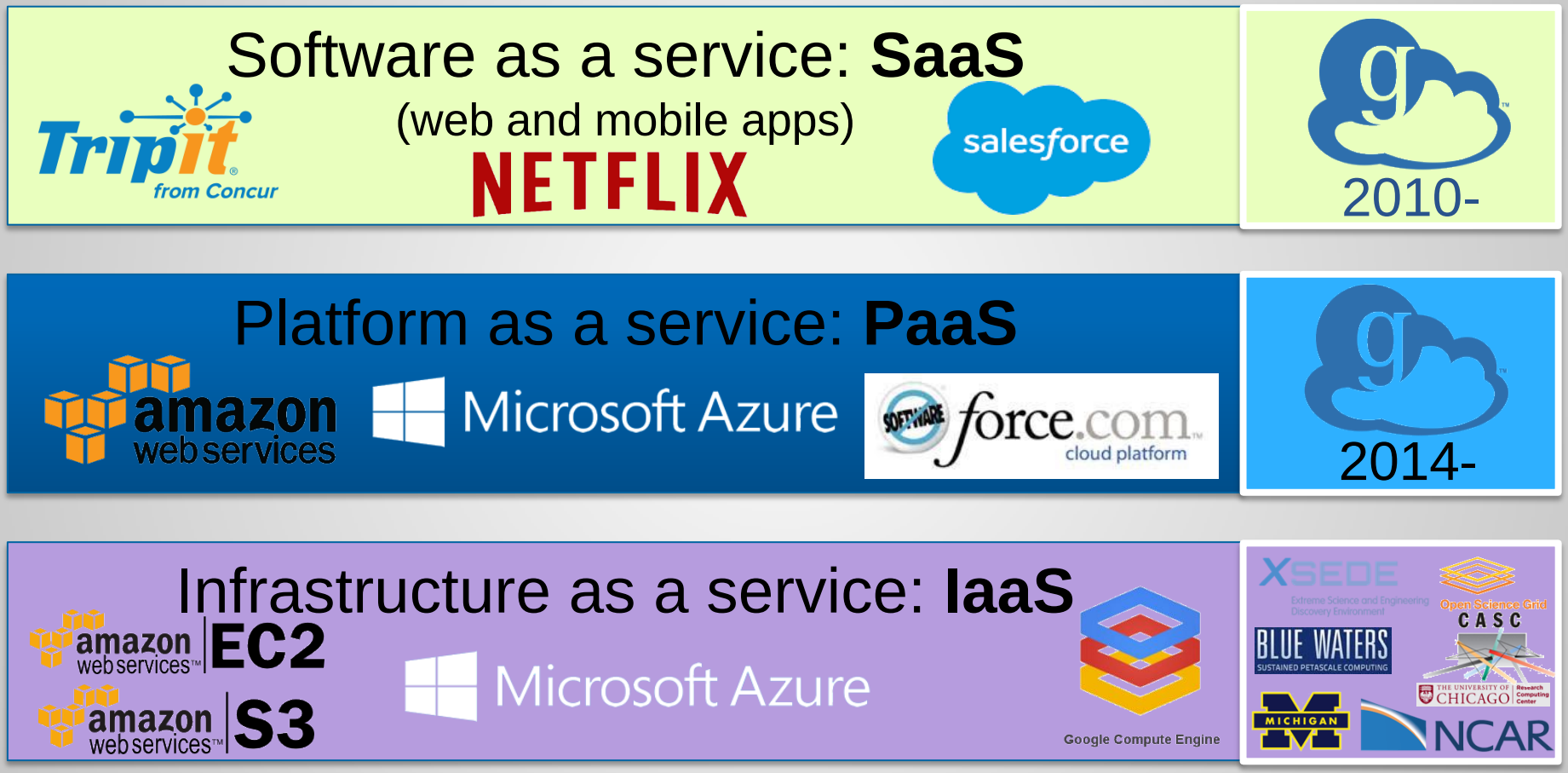
- Extend your portal with a public REST API, so that other app and service developers can integrate with and extend your portal

Why Globus Auth for your service?

- Outsource all identity management, authentication
  - Federated identity with InCommon, Google, etc.
- Outsource your REST API security
  - Consent, token issuance, validation, revocation
  - You provide service-specific authorization
- Apps use your service like all others
  - Its standard OAuth2 and OIDC
- Your service can seamlessly leverage other services
- Other services can leverage your service

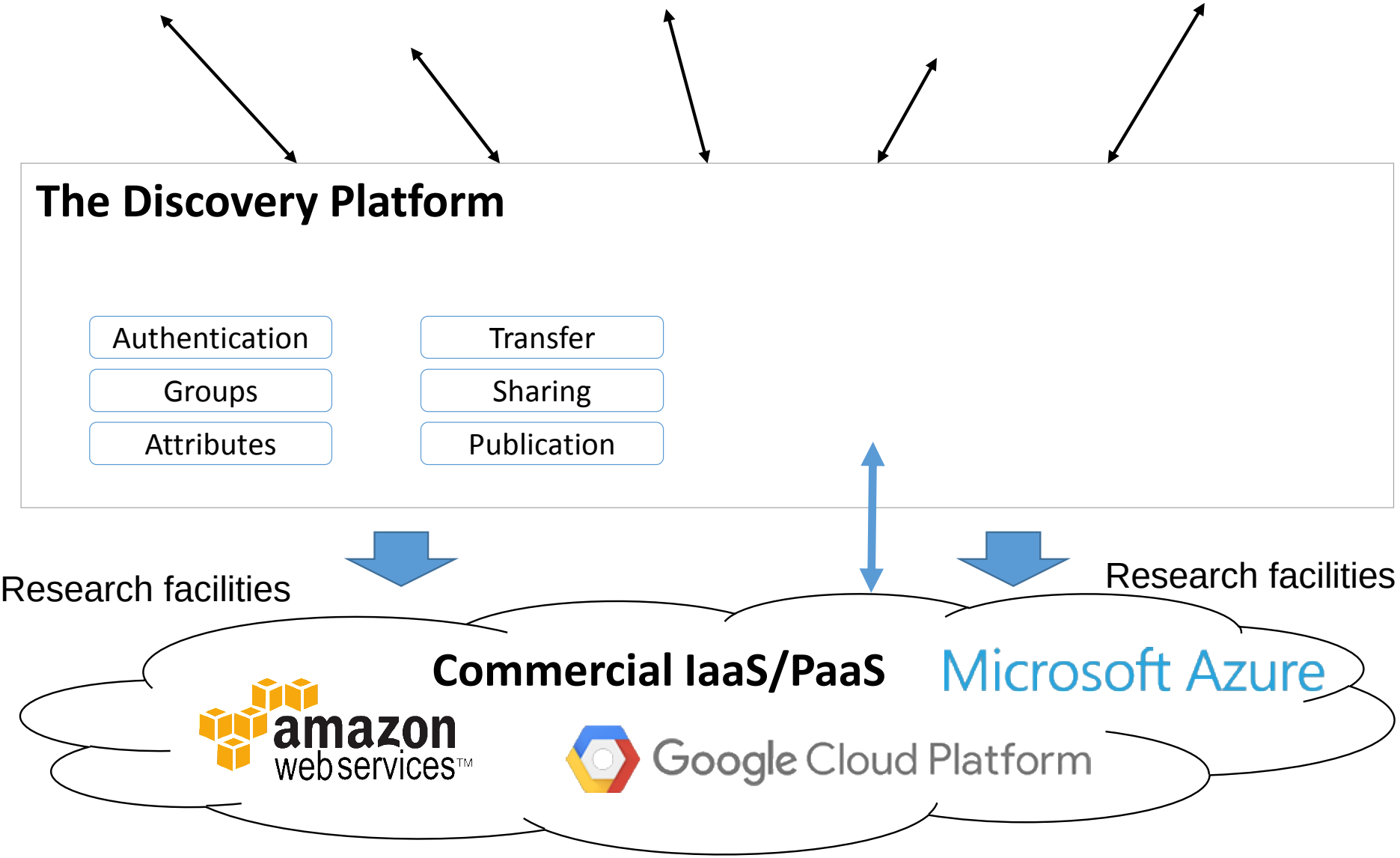
Add your service to the international science platform

A science platform can spur a discovery cloud ecosystem



In so doing, we can slash costs, improve quality, and accelerate discovery across the sciences

Enabling the Discovery Cloud



Together we can create
an integrated ecosystem
of services and applications
for the research and education
community

Thank you!

foster@uchicago.edu

@ianfoster