

# NCBI in a Data Enabled World

---

James Ostell

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

U.S. Department of Health and Human Services



# Two Use Cases for Cloud Platform

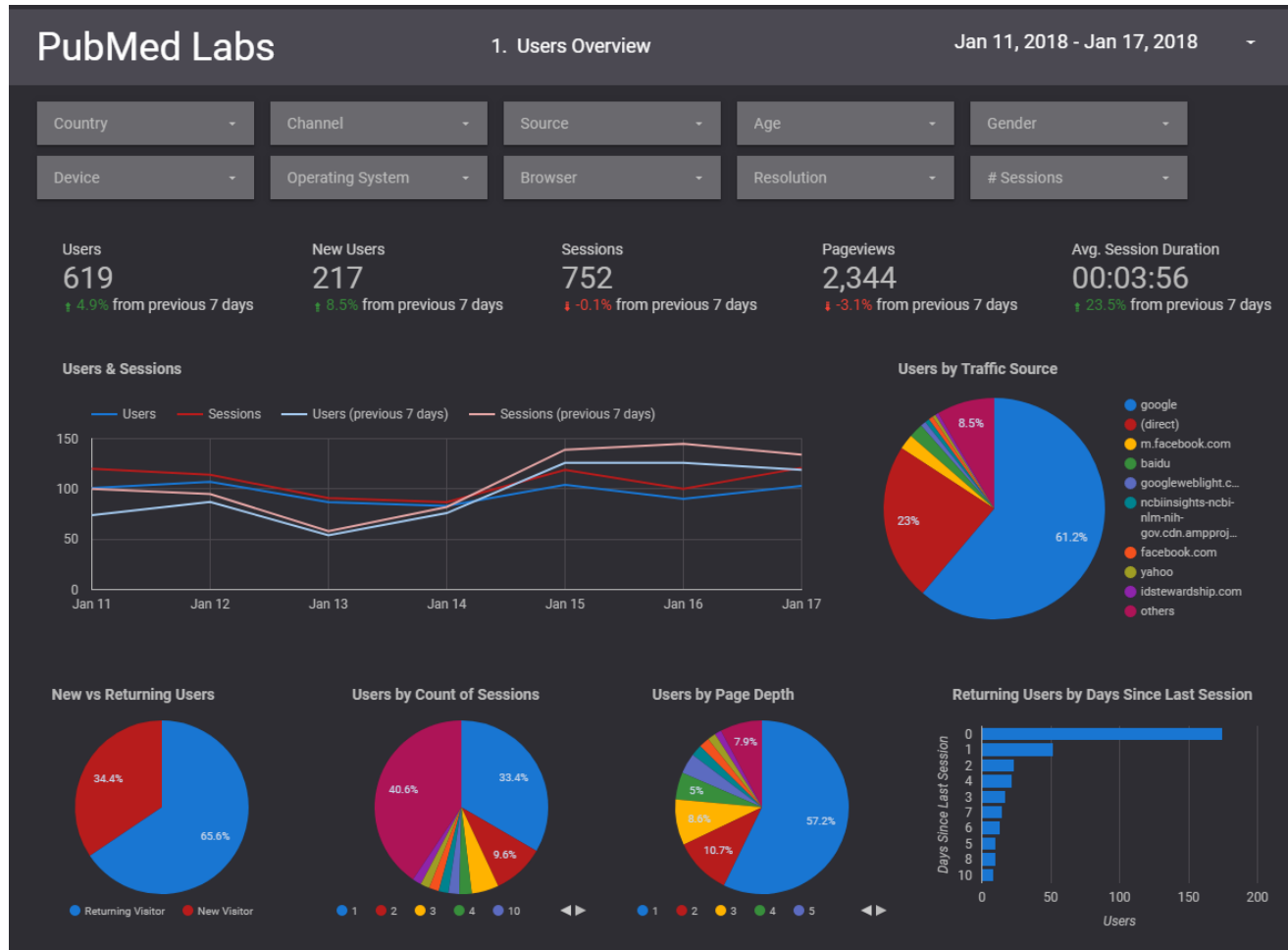
- To Deliver NCBI Services to the Public
  - PubMed 2.0
  - Modernize technology stack
  - High availability, scalability
  - Similar to other commercial sites (eg. Netflix, Amazon)
- To Provide Access to Data for Others
  - Access to NCBI data
    - No need to copy
    - No need to update
    - Convenient use without requiring NCBI servers/costs
  - Access to non-NCBI data
    - NCBI can provide indexing and access permissions without “owning” the data



# Delivering NCBI Services to the Public



# All New Apps use CI/CD, web instrumented



# All New Apps Enabled for Agile Development

## Google Optimize for UI A/B tests

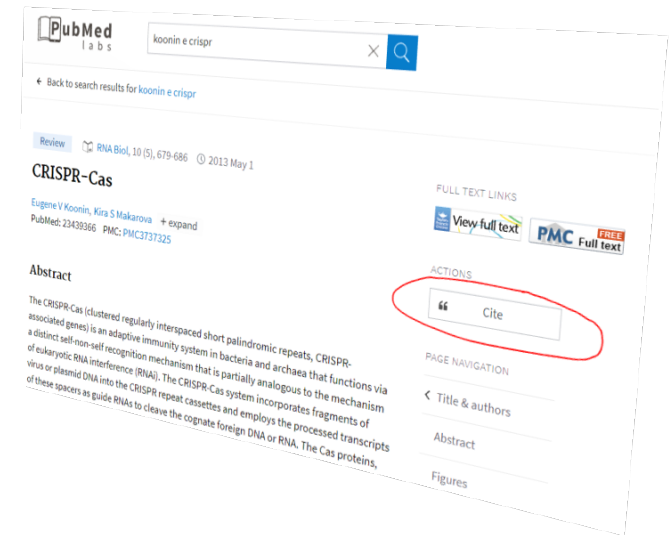
- UI changes can be done via Google Optimize interface
- Development and deployment cost is low
- Easy to create targeting audiences (particular percentage, etc.)
- Targeting is sticky – same users will see the same UI variant when they come back
- Plan running ads on PubMed and PubMed Mobile via Google Optimize in the future

“ Cite

“ Cite article

“ Cite

“ Cite article

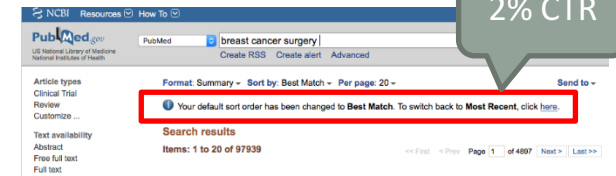
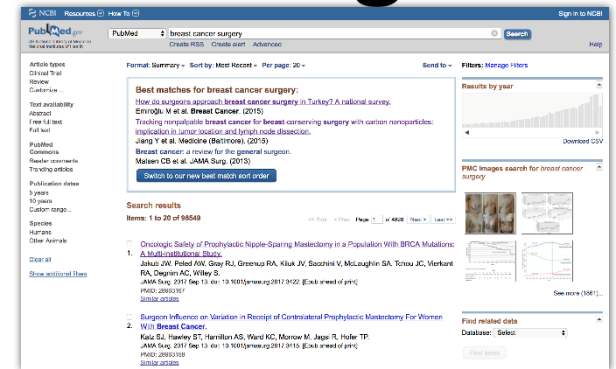
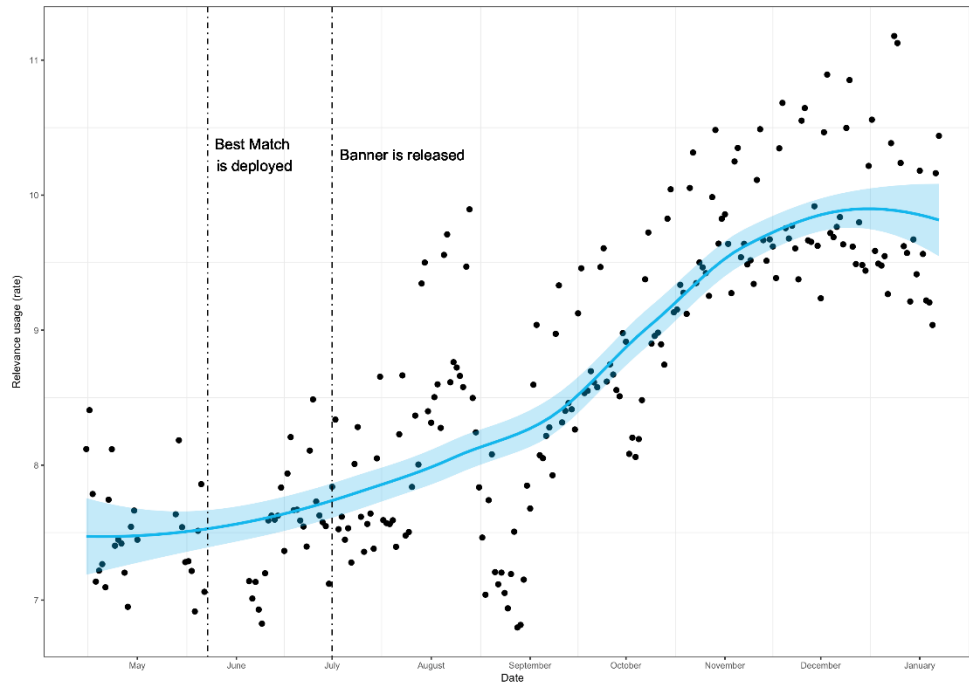


### Abstract Cite Button Label & Color

Nov 18, 2017 - Jan 18, 2018

Variant	Experiment Sessions	Conversions	Conversion Rate	↓	Compare to Original
<input checked="" type="checkbox"/> Original	268	6	2.24%		0%
<input checked="" type="checkbox"/> Blue button with white "Cite" label	378	17	4.50%		↑ 100.88%
<input checked="" type="checkbox"/> Gray (original) button with label "Cite article" instead of "Cite"	381	16	4.20%		↑ 87.58%
<input checked="" type="checkbox"/> Blue button with white "Cite article" label	424	16	3.77%		↑ 68.55%

# All New Apps Monitor Feature Usage



# Labs launched in Oct, 2017

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

This is PubMed Labs, a test site. For PubMed go to [pubmed.gov](http://pubmed.gov).

## PubMed labs

Try [otitis media treatment](#) [koonin e crispr](#) [influenza vaccine effectiveness](#)

**What is PubMed Labs?**

PubMed Labs is a test site where we are *experimenting* with new features and tools that eventually may be incorporated in PubMed, in their current or a revised form based on the input we receive. Please try the site and [let us know](#) what you think.

[Feedback](#)

NIH U.S. National Library of Medicine NCBI Log in

This is PubMed Labs, a test site. For PubMed go to [pubmed.gov](http://pubmed.gov).

## PubMed labs

Try [otitis media treatment](#) [koonin e crispr](#) [influenza vaccine effectiveness](#)

### What is PubMed Labs?

PubMed Labs is a test site where we are *experimenting* with new features and tools that eventually may be incorporated in PubMed, in their current or a revised form based on the input we receive. Please try the site and [let us know](#) what you think.

### Highlights of PubMed Labs

#### Cutting-edge search

PubMed Labs includes a new search algorithm that uses machine learning to more accurately find the best matches.

[Feedback](#)

NIH U.S. National Library of Medicine NCBI Log in

This is PubMed Labs, a test site. For PubMed go to [pubmed.gov](http://pubmed.gov).

## PubMed labs

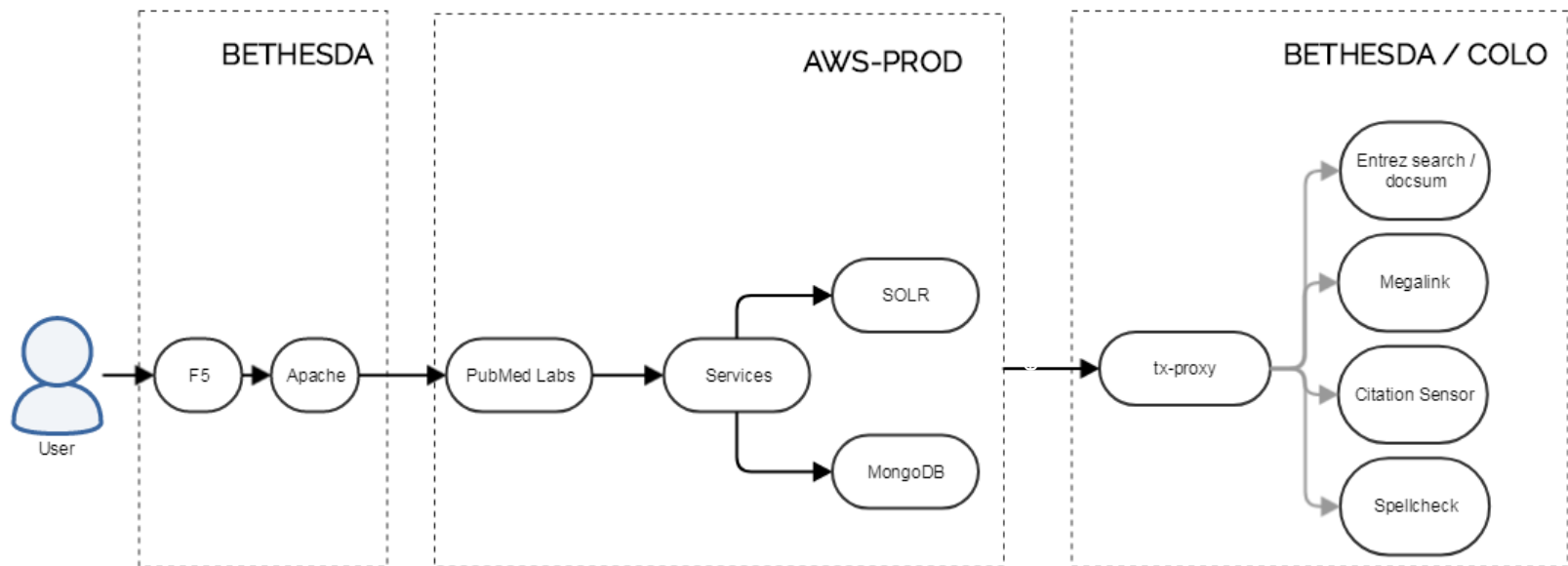
Try [otitis media treatment](#) [koonin e crispr](#) [influenza vaccine effectiveness](#)

### What is PubMed Labs?

PubMed Labs is a test site where we are *experimenting* with new features and tools that eventually may be incorporated in PubMed.

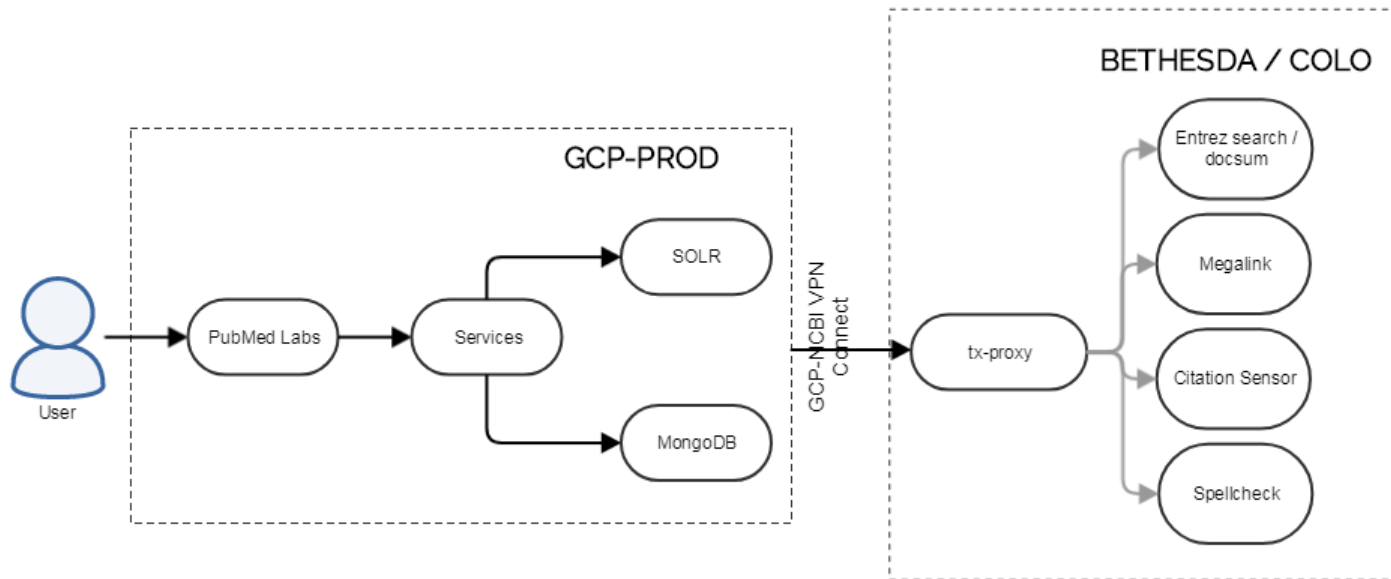
[Feedback](#)

# Current architecture - AWS

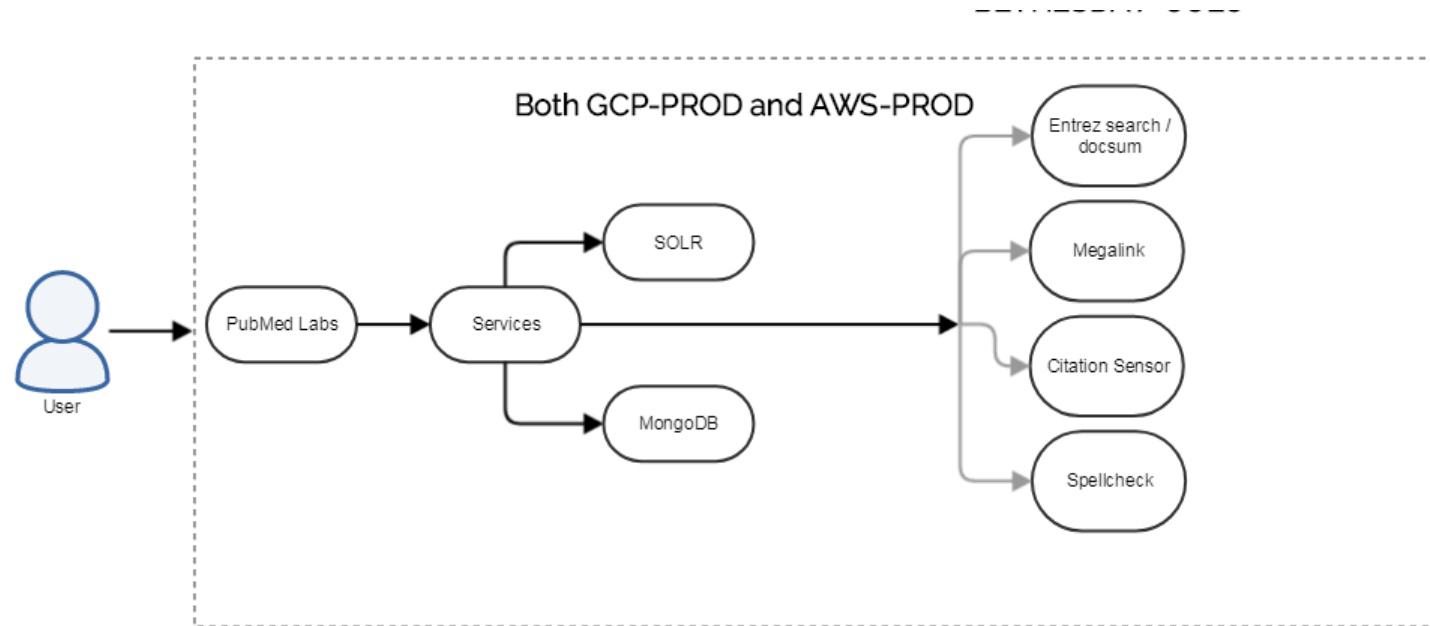




# Proposed test architecture - GCP



# Final architecture – both AWS and GCP



# Providing Access to Data for Others



# BLAST in the Cloud

The screenshot shows the AWS Marketplace page for the NCBI BLAST AMI. The page includes a search bar, navigation links, and a product overview section. The product overview section contains a description of the AMI, a highlights box, and a technical specifications table.

**NCBI BLAST**  
Sold by: NCBI Latest Version: 2016-10-12-2.5.0

The BLAST AMI provides access to the popular sequence search similarity program in a convenient package. It is pre-configured with the latest release of

Linux/Unix ★★★★☆ (1)

Typical Total Price  
**\$0.532/hr**  
Total pricing per instance for services hosted on m3.2xlarge in US East (N. Virginia). [View Details](#)

**Product Overview**

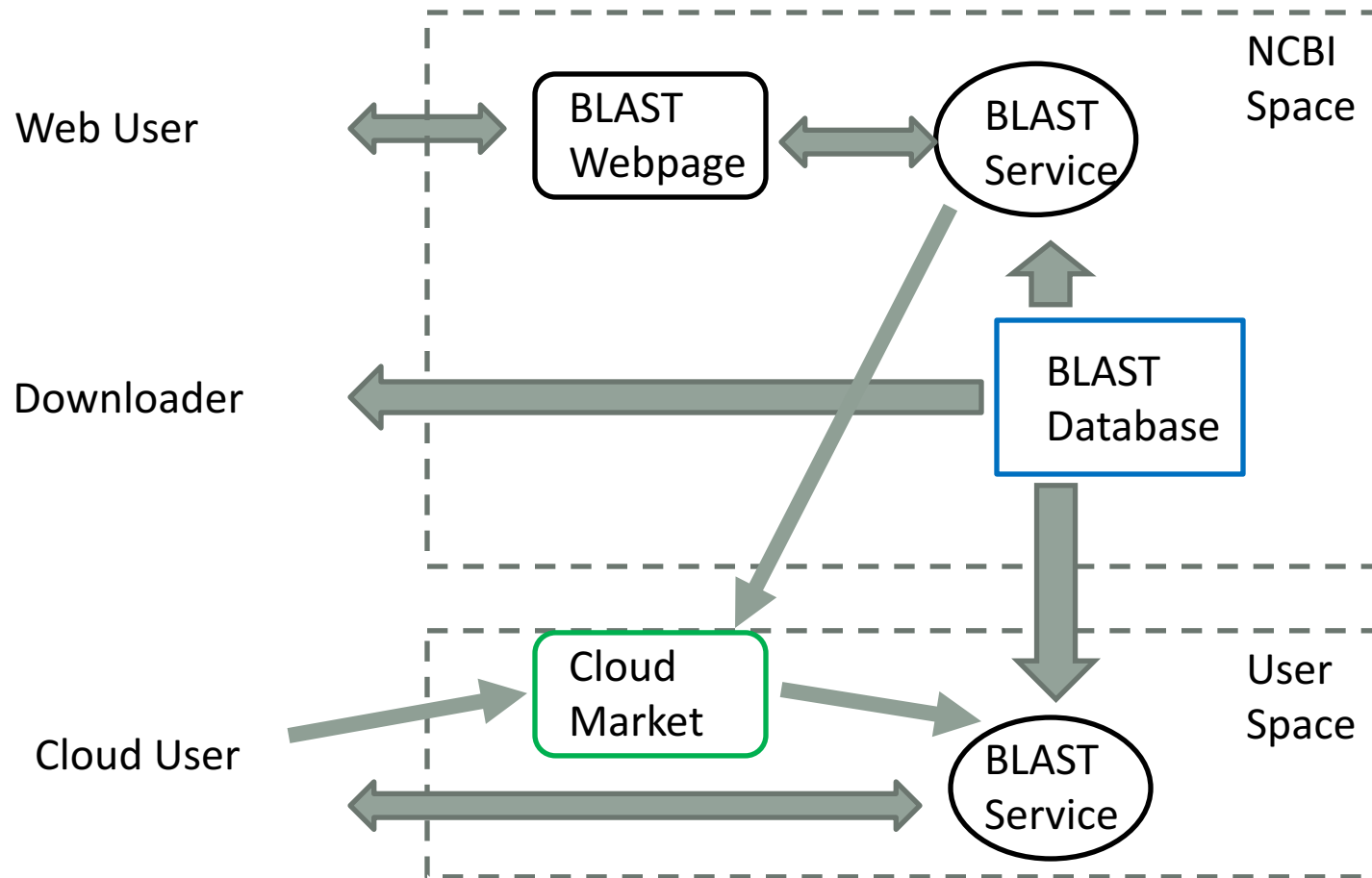
This BLAST AMI is a very exciting development as it allows users to perform sequence similarity searches without restriction they might encounter at a public website and without the work of setting up stand-alone BLAST. The AMI includes a FUSE client that automatically downloads the most popular BLAST databases from the NCBI, and users can still upload their own custom databases. The AMI allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified webpage.

**Highlights**

- This AMI is preconfigured with the latest BLAST+ release and has a simplified BLAST web page.
- This AMI includes a FUSE client that automatically downloads and caches popular NCBI databases such as nr, nt, swissprot, refseq, and PDB.
- This AMI supports a subset of the NCBI BLAST URL API allowing remote submission and formatting of searches.

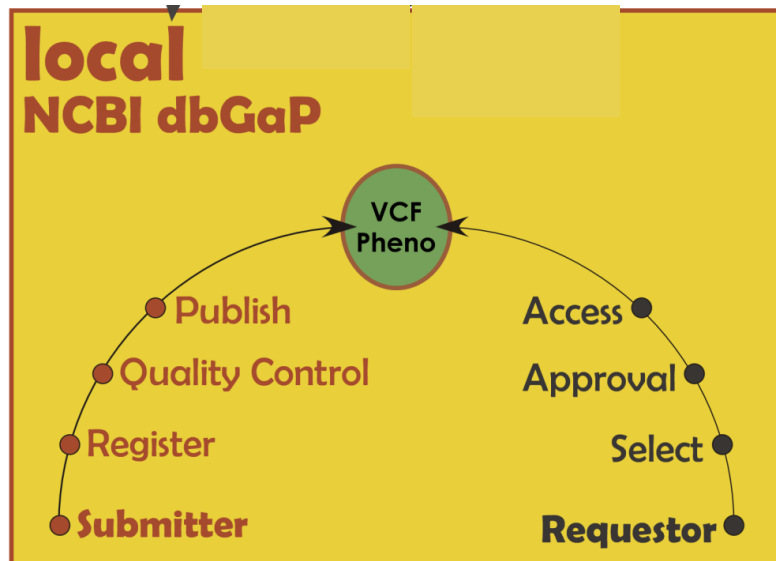
Version	2016-10-12-2.5.0
Sold by	NCBI
Categories	Application Servers Healthcare & Life Sciences
Operating System	Linux/Unix, Ubuntu 12.04

# BLAST in the Cloud

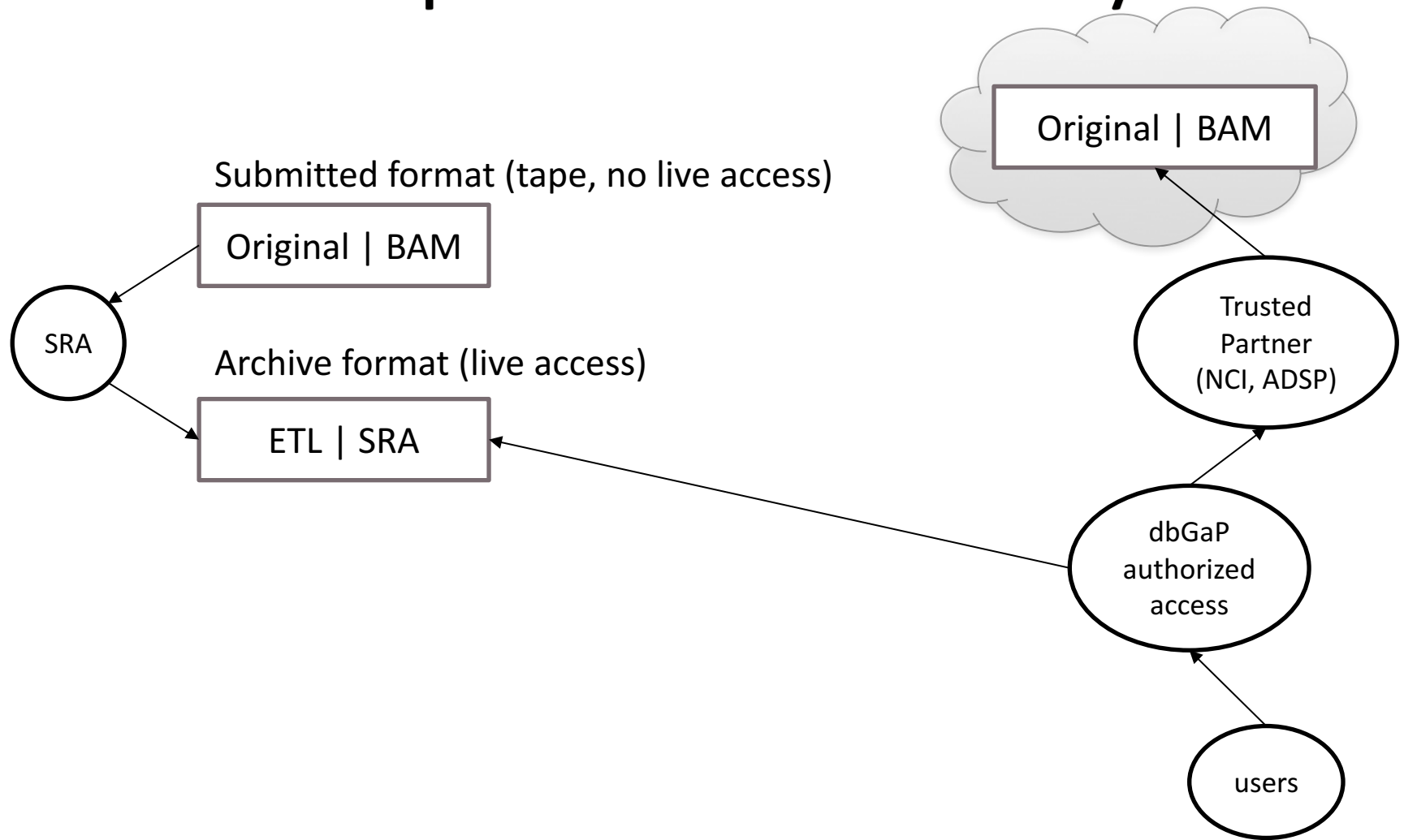


# dbGaP Process for Protected Human Data

**1** Existing dbGaP processes & policies



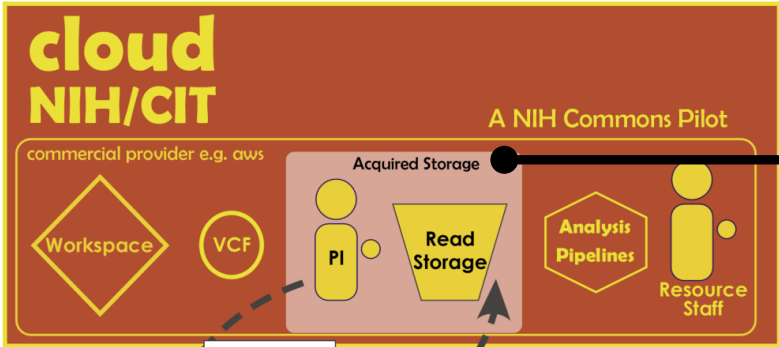
# Human Sequence Reads - today



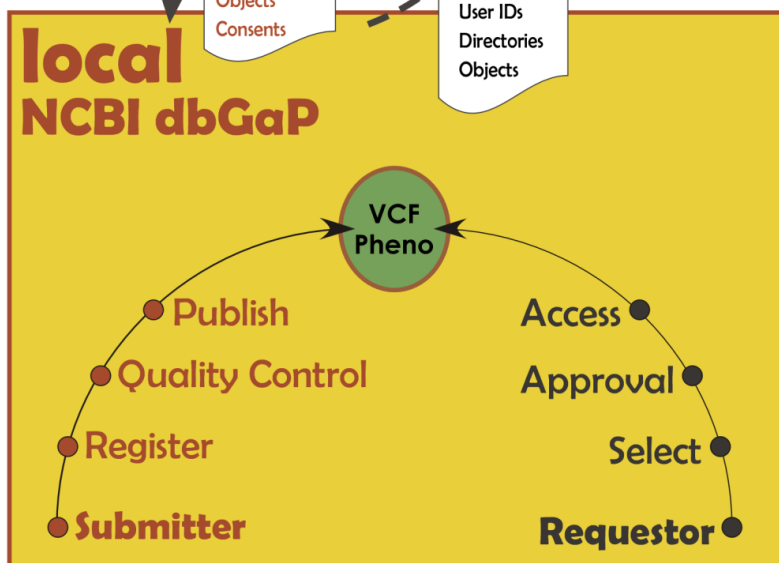
**3** Final oversight after pilot



**2** NIH Commons Pilot



**1** Existing dbGaP processes & policies

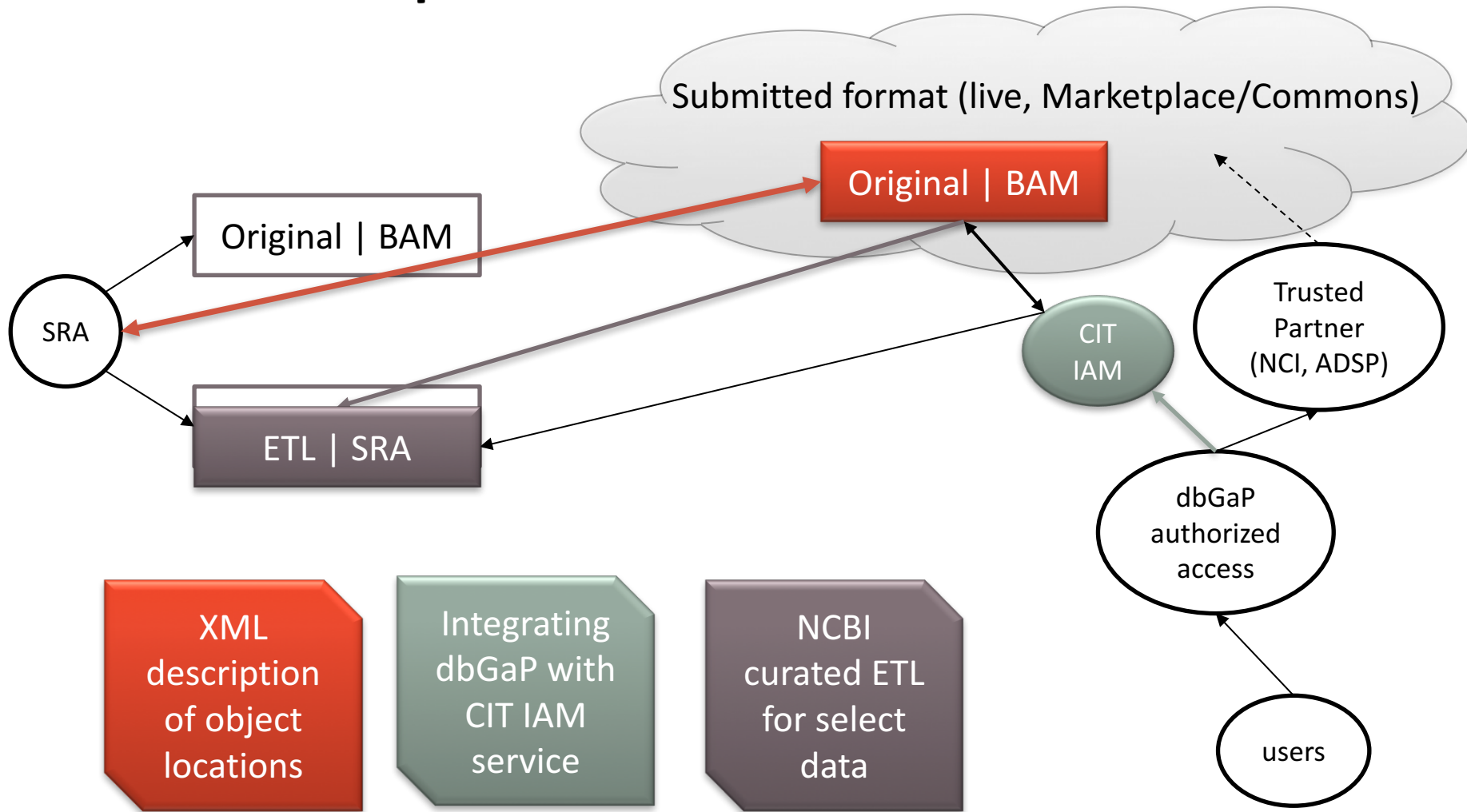


A variety of storage acquisition procedures are currently in use:

- Direct purchase by IC
- Indirect purchase by IC (Grantee)
- MITRE-acquired



# Human Sequence Reads - future



# Federated Data Commons Model

## NIH Data Commons



TOPMed



## Interoperability Tools

- Common user authentication system
- Shared APIs for data access & computing
- Adoption of FAIR Principles
- Docker containers
- Workflows management
- Digital objects catalogues and IDs
- Data standards and ontologies

## NHGRI Genomic AnVIL

