# NCBI in a Data Enabled World

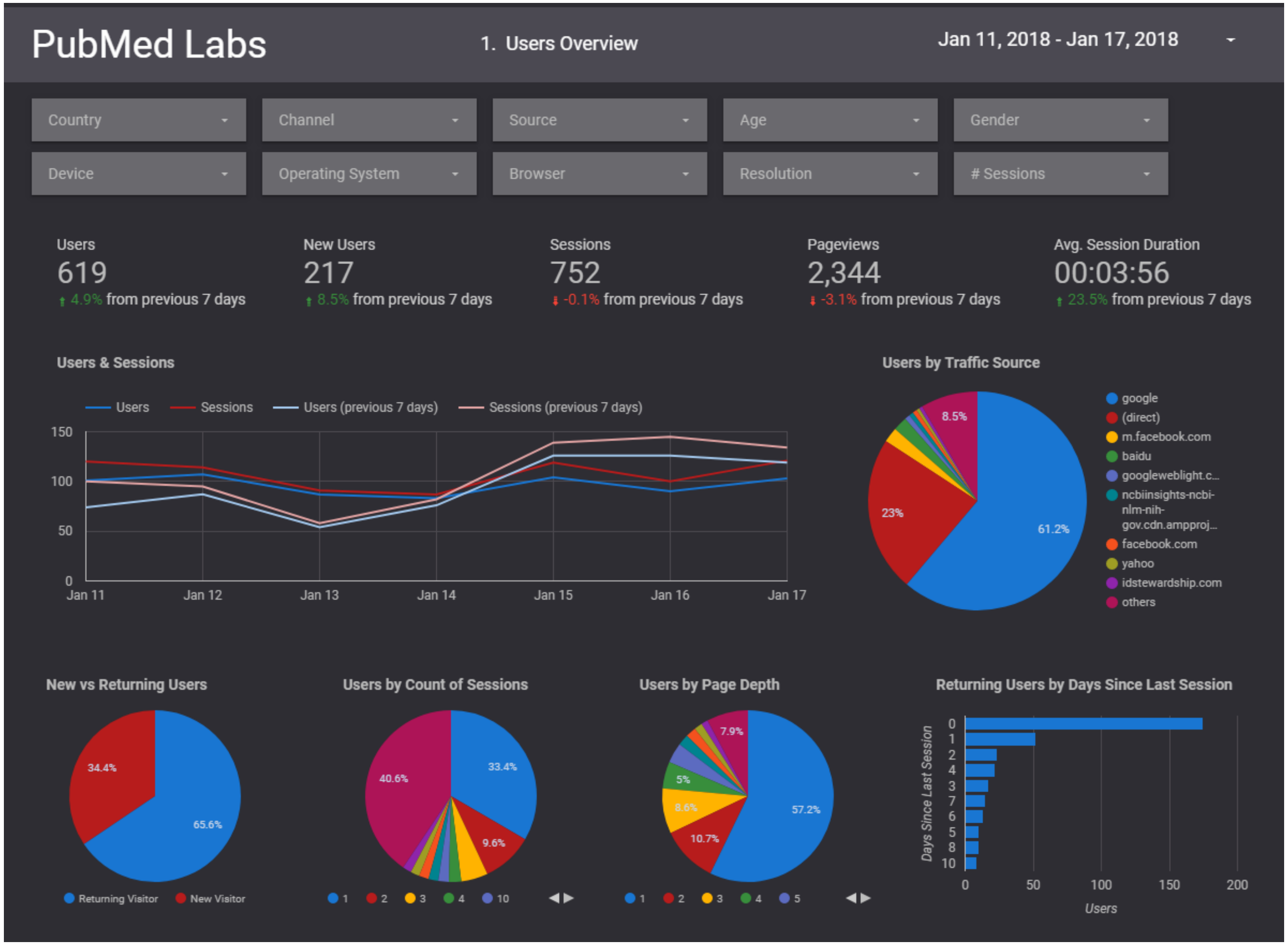
James Ostell
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
U.S. Department of Health and Human Services

# 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

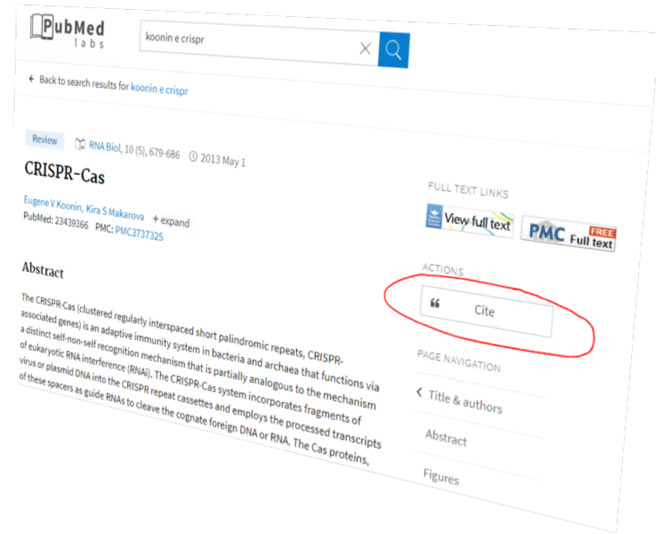
# All New Apps use CI/CD, web instrumented



# 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

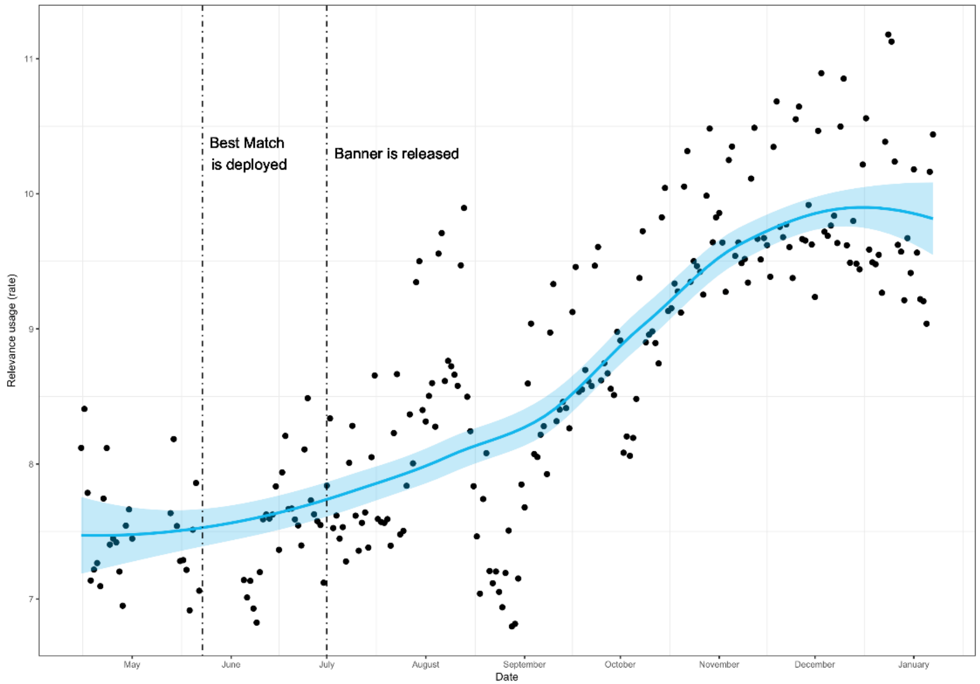



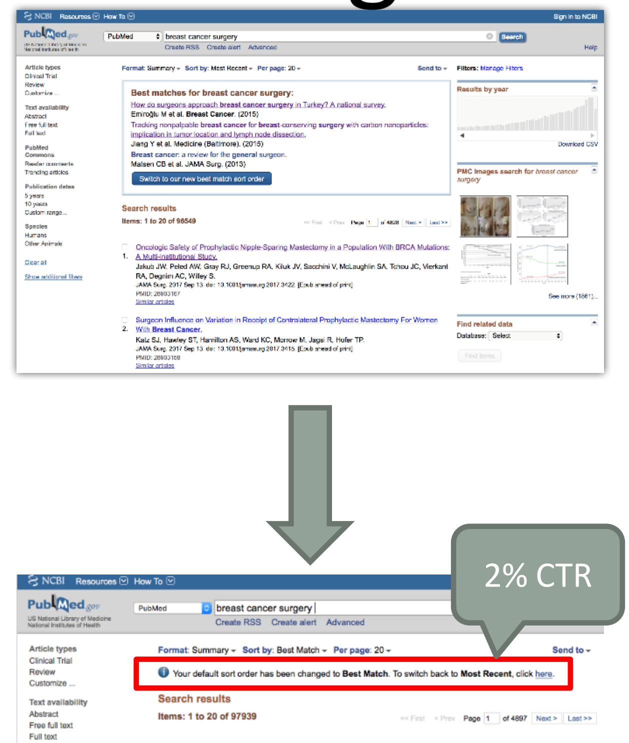

### Abstract Cite Button Label & Color

Nov 18, 2017 - Jan 18, 2018

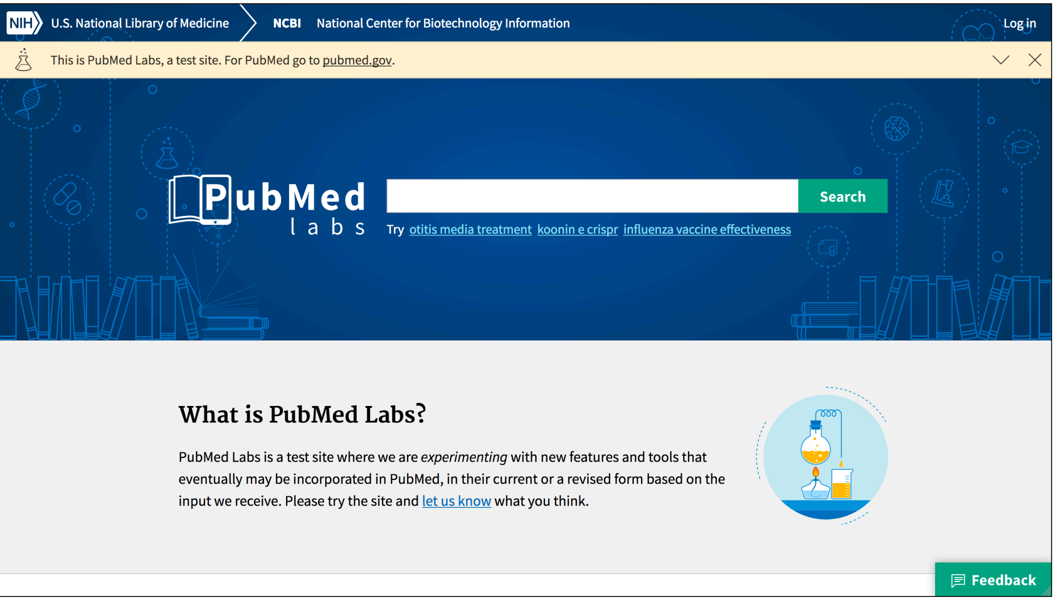
| Variant | Experiment Sessions | Conversions | Conversion Rate | Compare to Original |
|---|---|---|---|---|
| Original | 268 | 6 | **2.24%** | 0% |
| Blue button with white "Cite" label | 378 | 17 | **4.50%** | 100.88% |
| Gray (original) button with label "Cite article" instead of "Cite" | 381 | 16 | **4.20%** | 87.58% |
| Blue button with white "Cite article" label | 424 | 16 | **3.77%** | 68.55% |

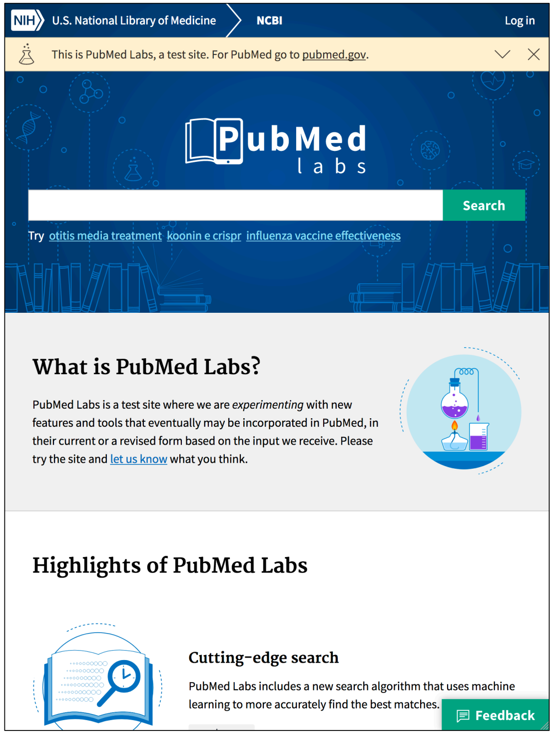
# All New Apps Monitor Feature Usage

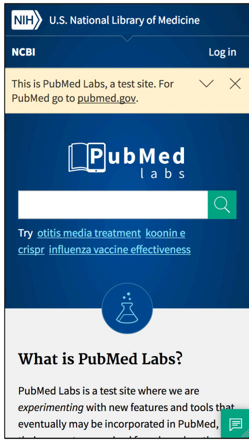


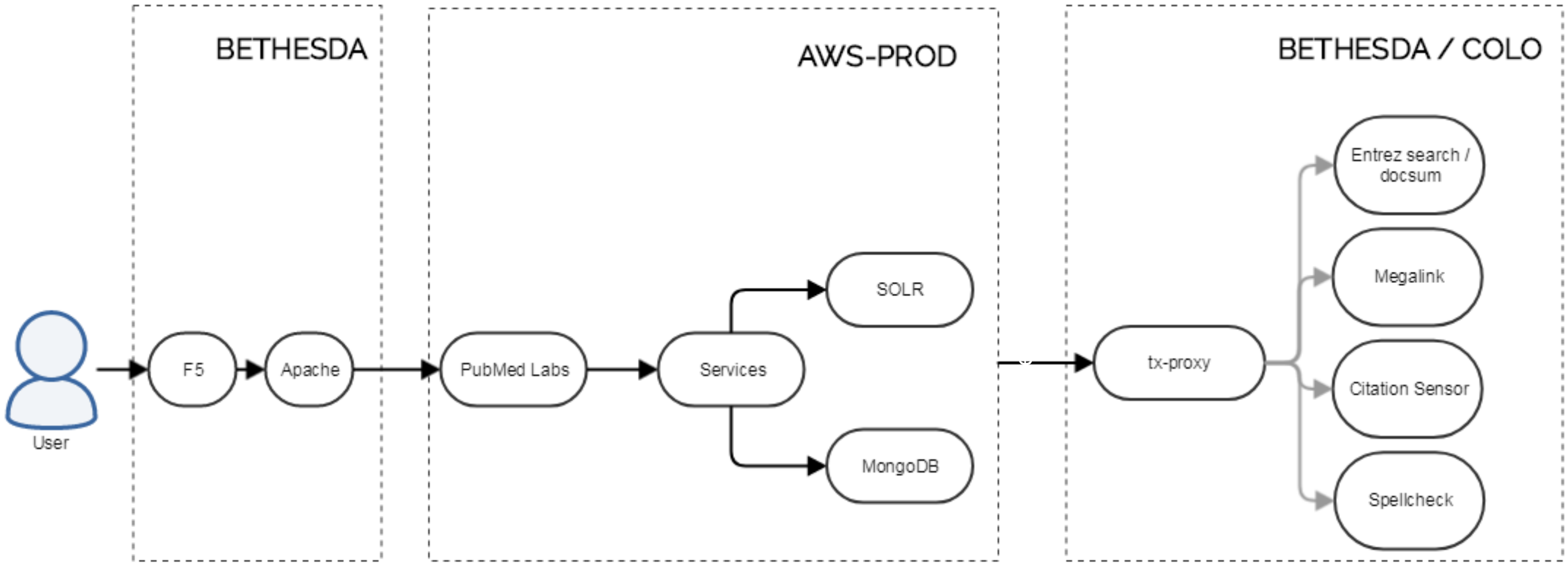
# Labs launched in Oct, 2017



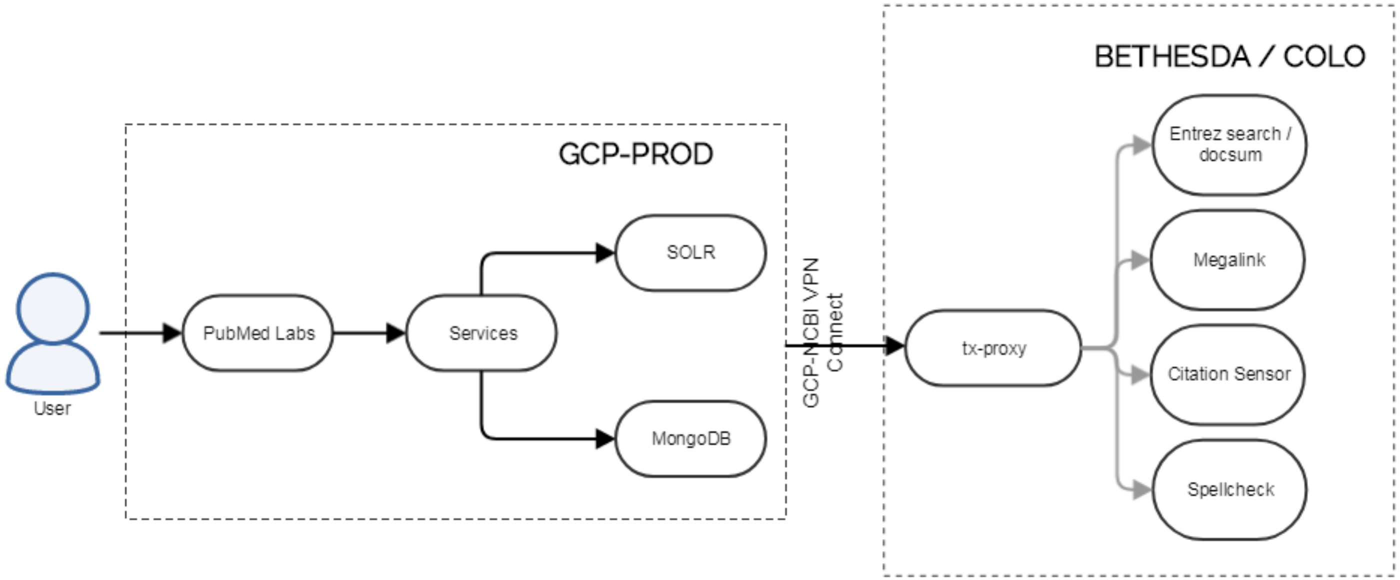



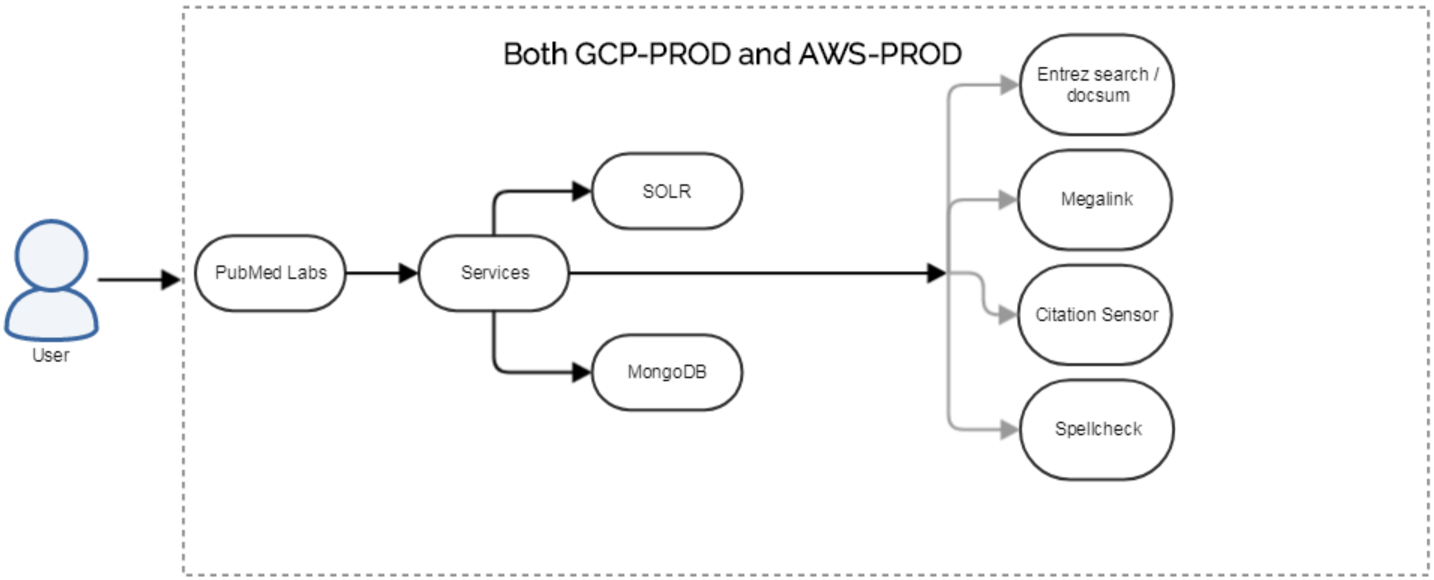
# Current architecture - AWS



# Proposed test architecture - GCP

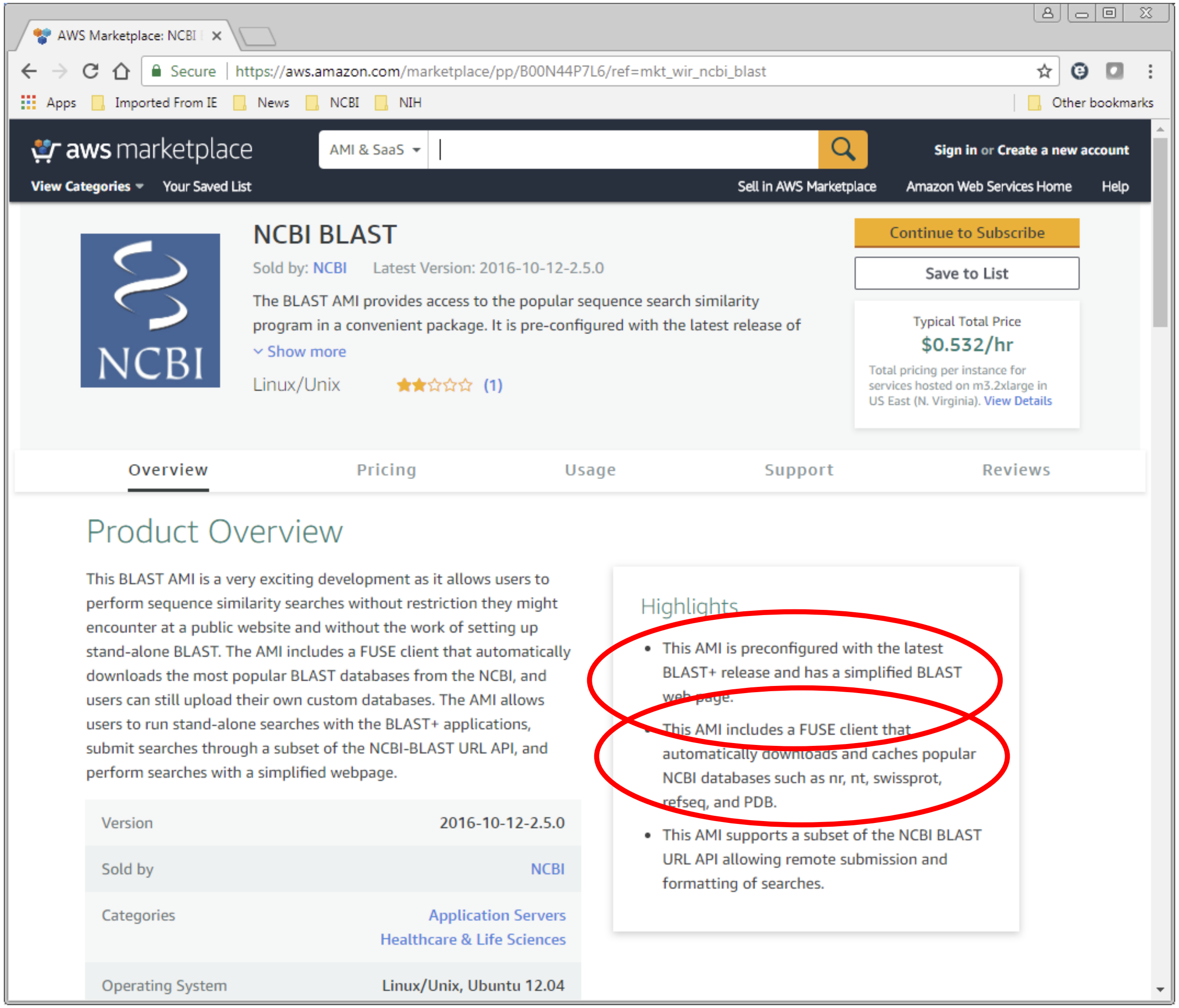


# Final architecture – both AWS and GCP

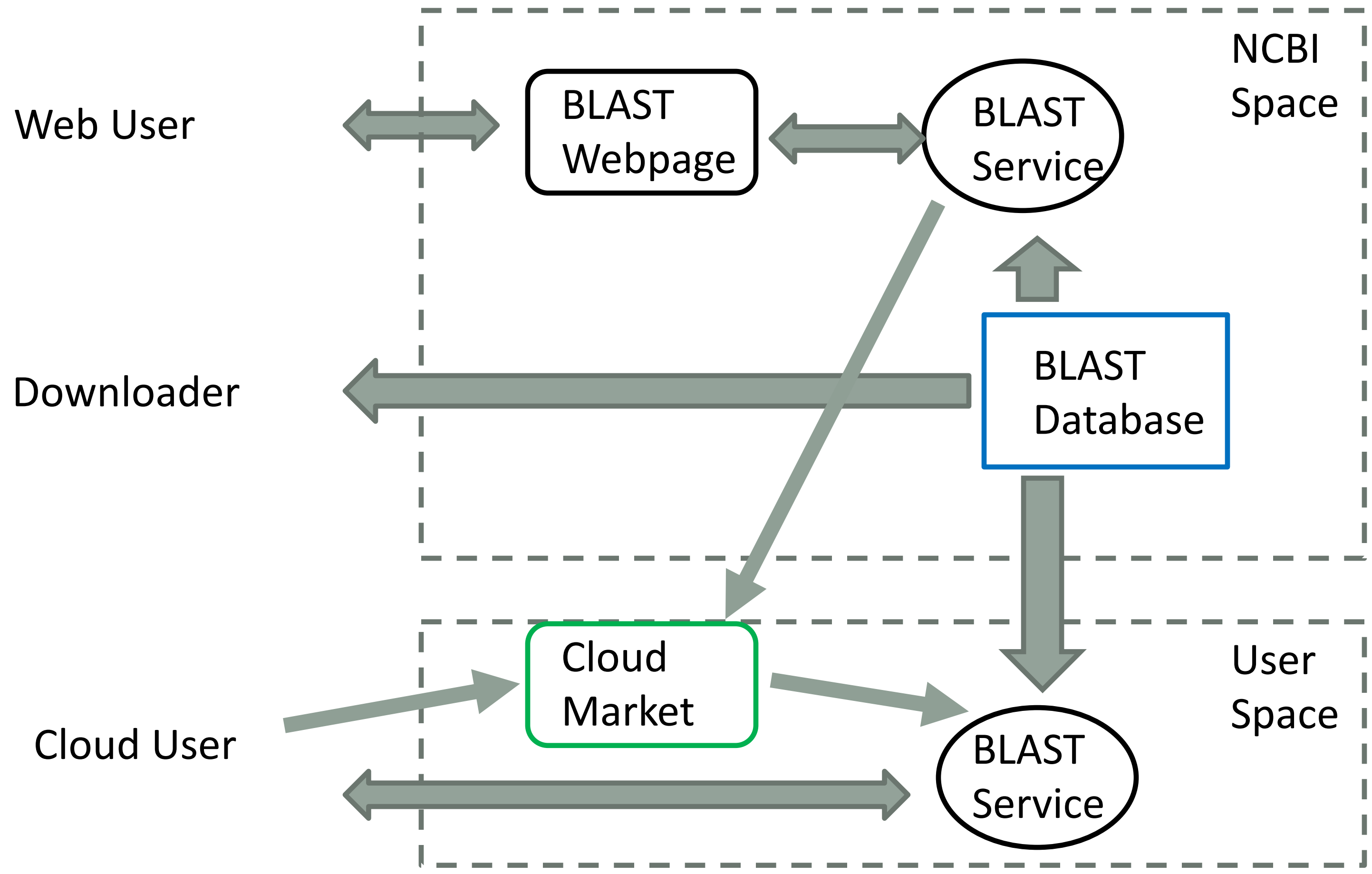


# Providing Access to Data for Others

# BLAST in the Cloud

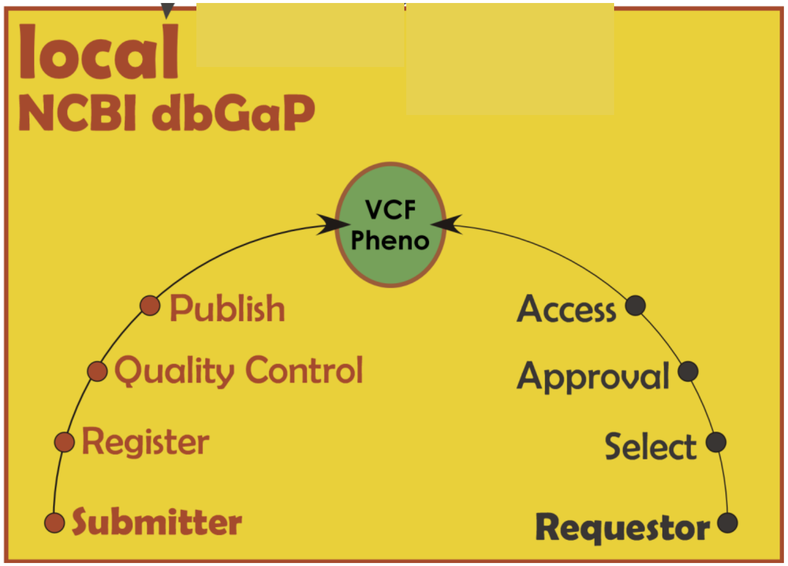


# BLAST in the Cloud

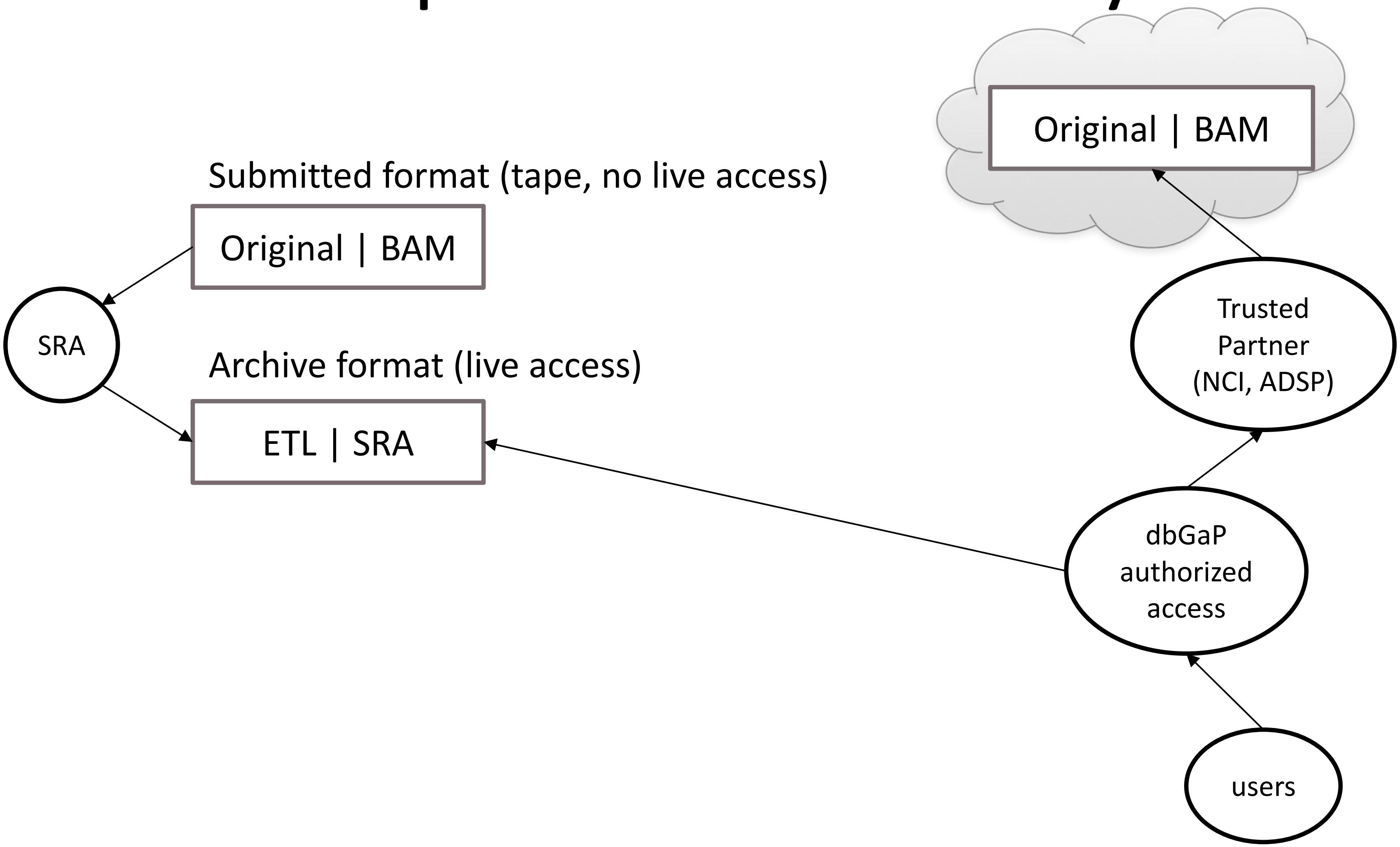


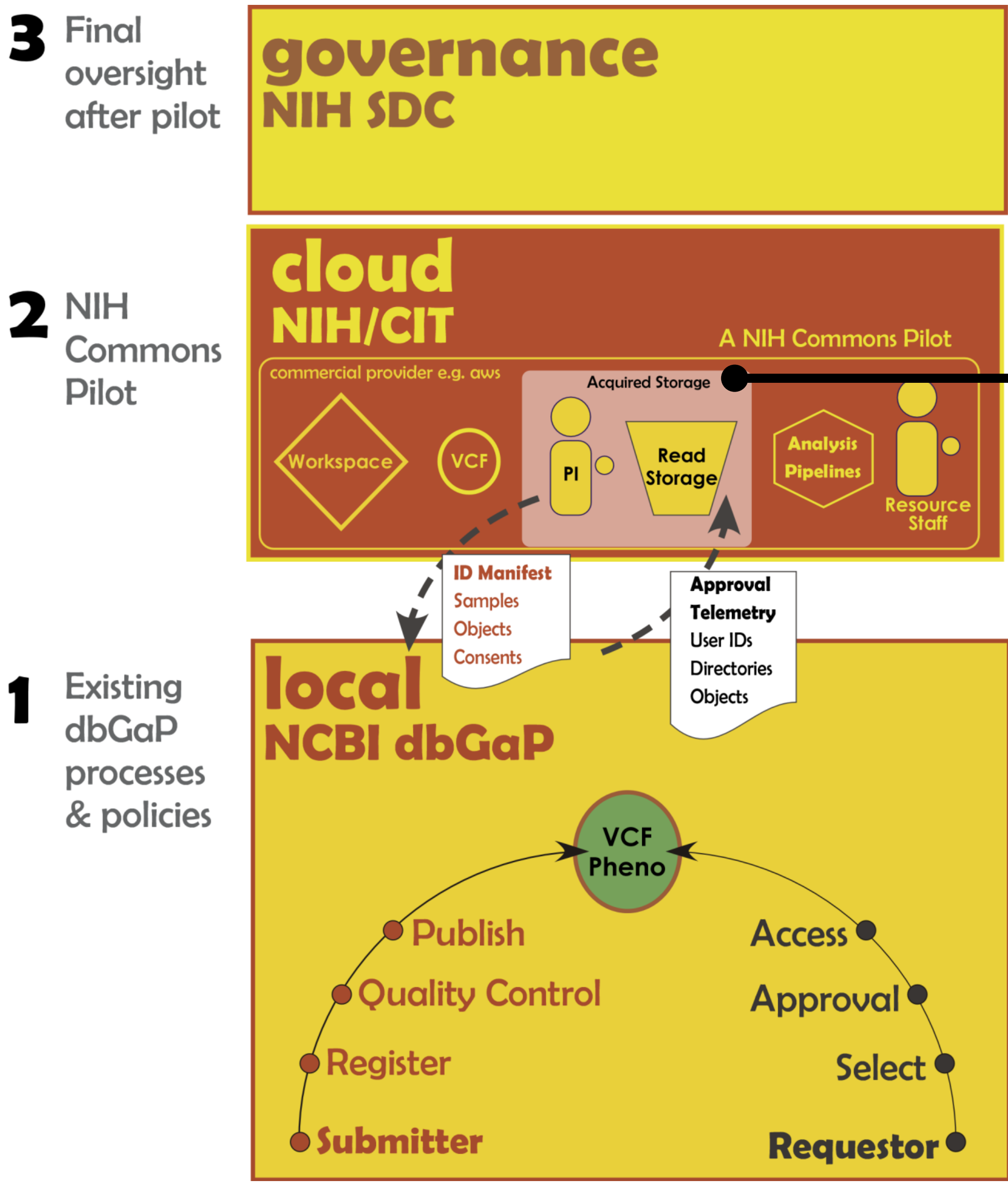
# dbGaP Process for Protected Human Data

1 Existing dbGaP processes & policies



# Human Sequence Reads - today

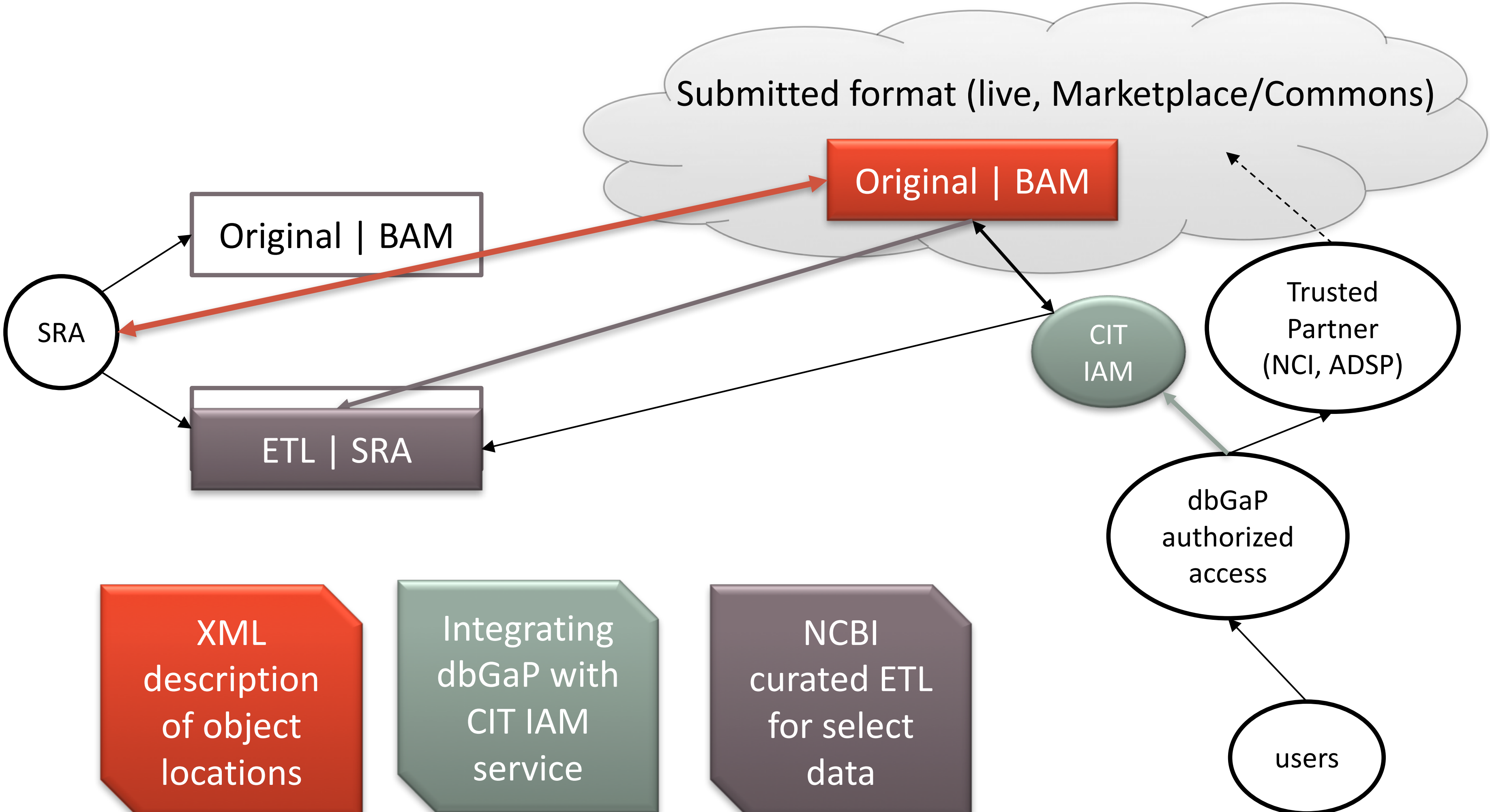





A variety of storage acquisition procedures are currently in use:

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

# Human Sequence Reads - future



# Federated Data Commons Model

**NIH Data Commons**

GTEx

ALLIANCE of GENOME RESOURCES

TOPMed

dbGaP GENOTYPES and PHENOTYPES

**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

All of Us THE FUTURE OF HEALTH BEGINS WITH YOU | The Precision Medicine Initiative

Diagram courtesy of Nick Weber, CIT