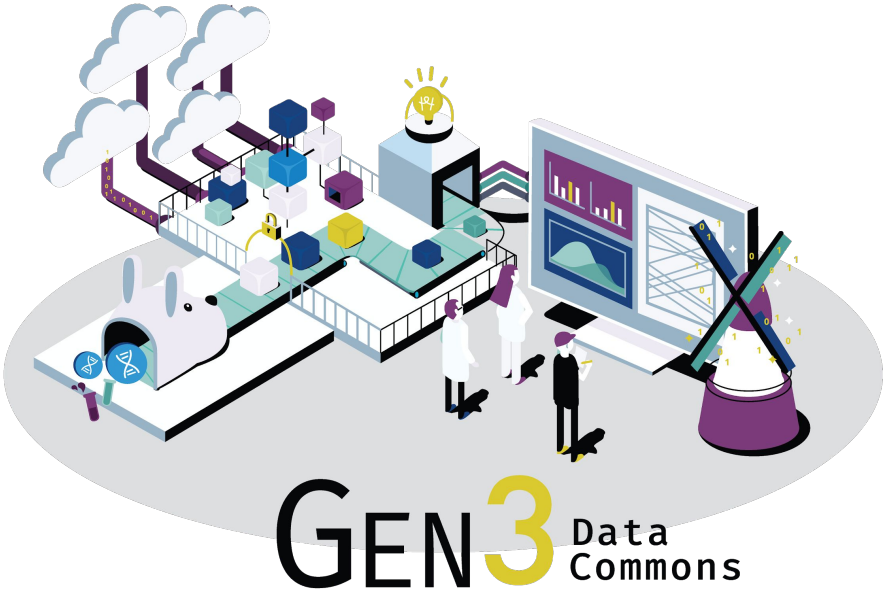
GEN3 Data Commons

Authentication, Authorization, and Data Access

Alexander VanTol
Rudyard Richter
Phillis Tang

Center for Translational Data Science,
University of Chicago



Data commons co-locate data, storage and computing infrastructure with commonly used software services, tools & apps for analyzing and ***sharing data*** *to create a resource for the research community.*

Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.

Sharing Data

GEN3 Data Commons

- Authorize users for access
- Make data files available for download
- Allow other platforms to access data in Gen3 on behalf of users

The Agenda

GEN3 Data Commons

- Authentication & Authorization
  - What are authentication (“authN”) and authorization (“authZ”)?
  - Gen3 implementation: fence
  - Interoperability
- Data Access
  - Motivation
  - Gen3 implementation: indexd
  - Interoperability

Fence

Gen3 authentication and authorization service

Fence

Gen3 **authentication** and authorization service

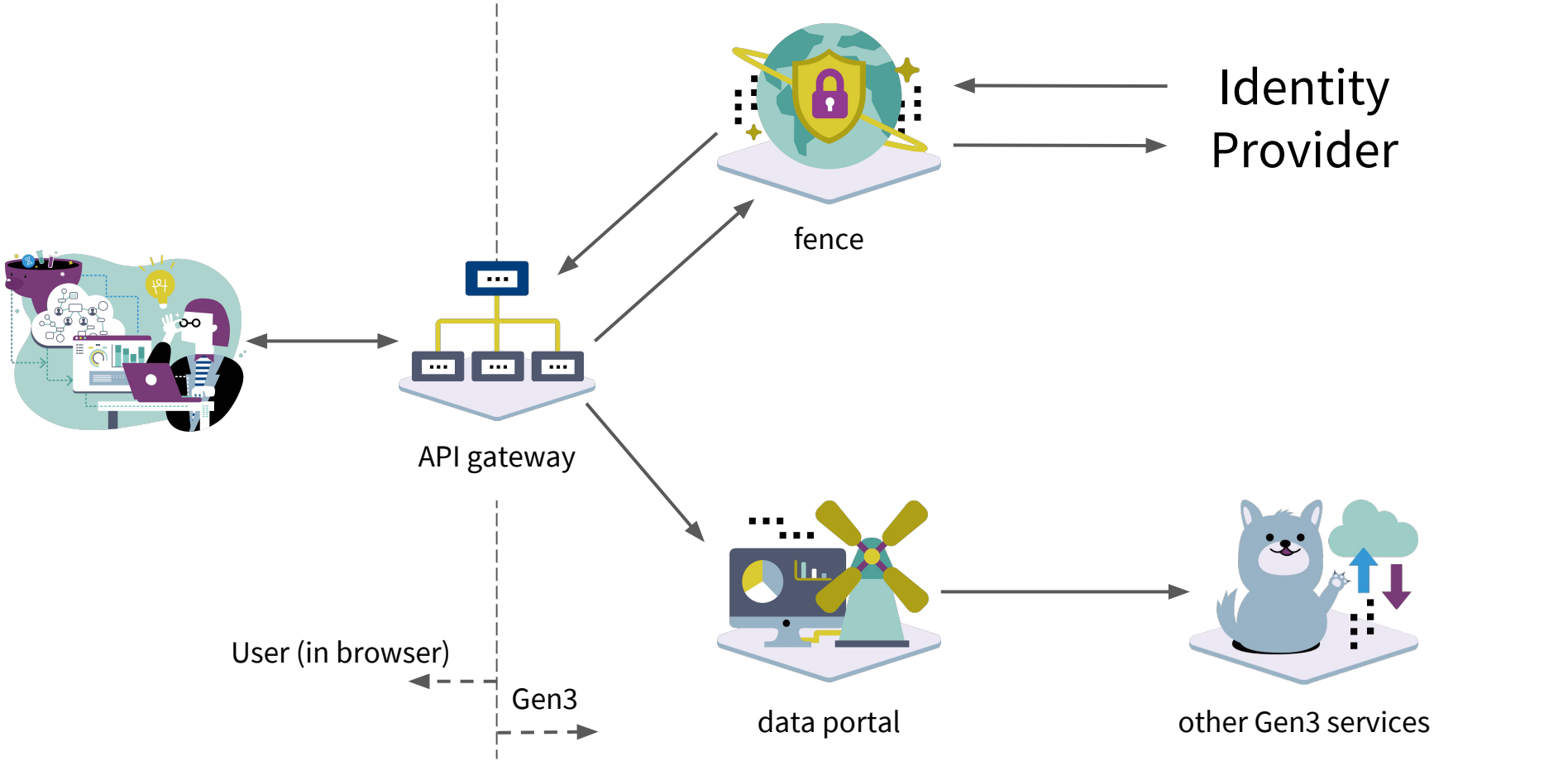
authentication: who you are

Fence

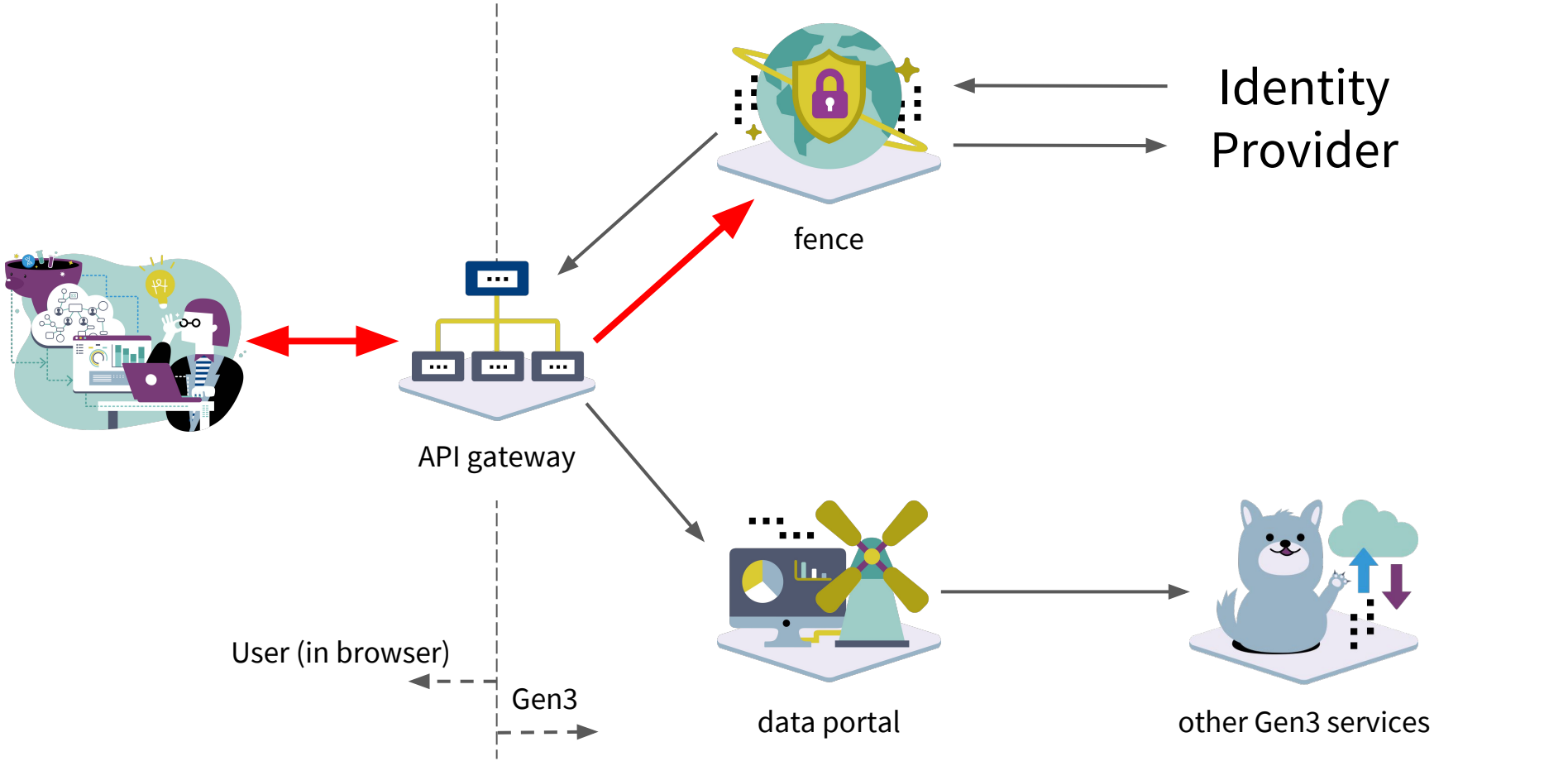
Gen3 authentication and **authorization** service

authorization: what you can do

Authentication: Gen3 Portal



Authentication: Gen3 Portal



Fence Login

GEN3 Data Commons

Submit Data | Documentation

GEN3 Data Commons | Generic Data Commons

Dictionary | Exploration | Files | Query | Workspace | Profile

Generic Data Commons

SEARCH, COMPARE, AND DOWNLOAD DATA

This website supports the management, analysis and sharing of human disease data for the research community and aims to advance basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cancer.

Login with Google

If you have any questions about access or the registration process, please contact support@datacommons.io.

Dictionary
v0.2.1-10-g76a9e30

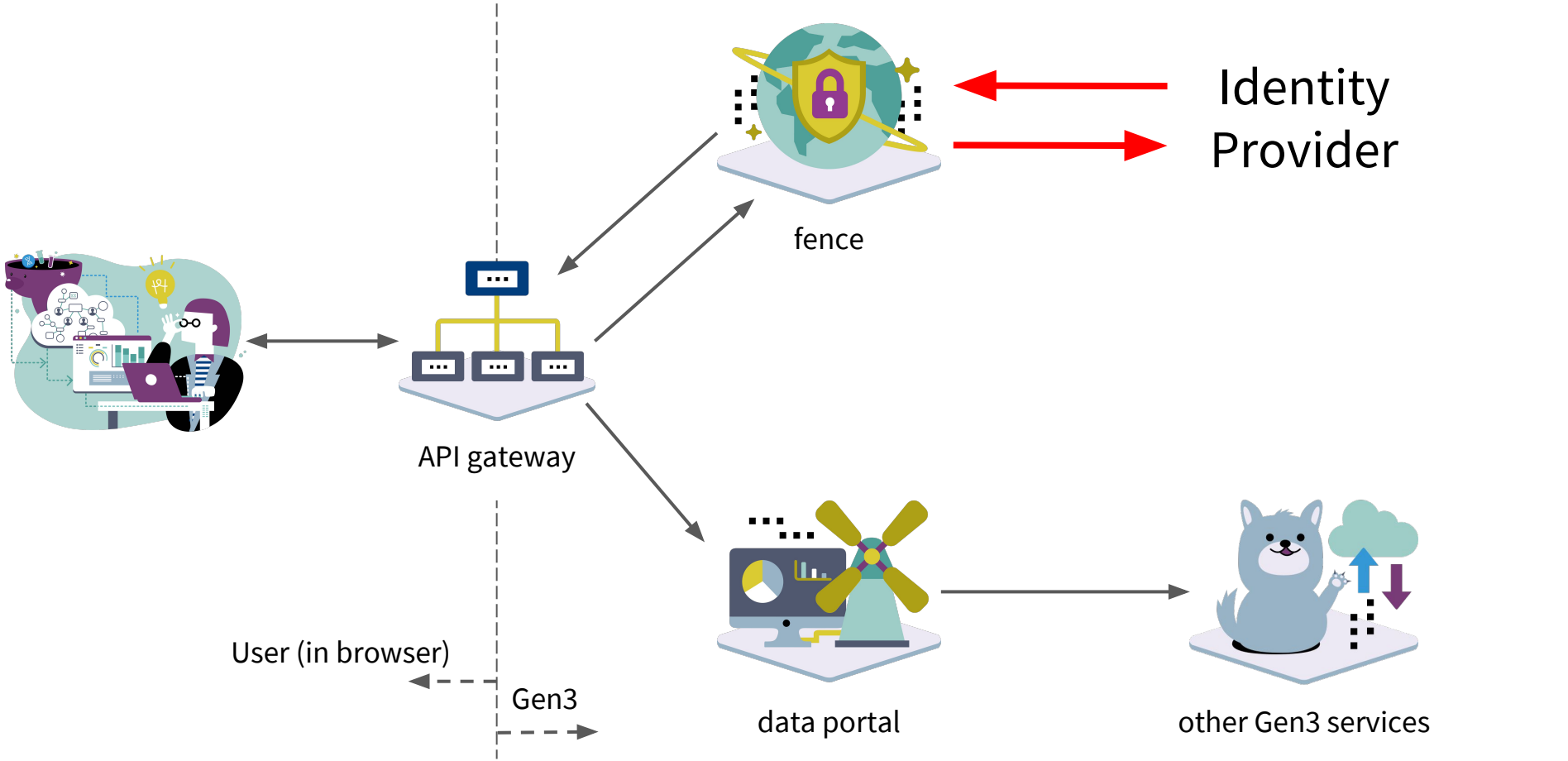
Submission
v1.1.8-1-ga1d9972

Portal
v2.5.6-1-g8e02bf0

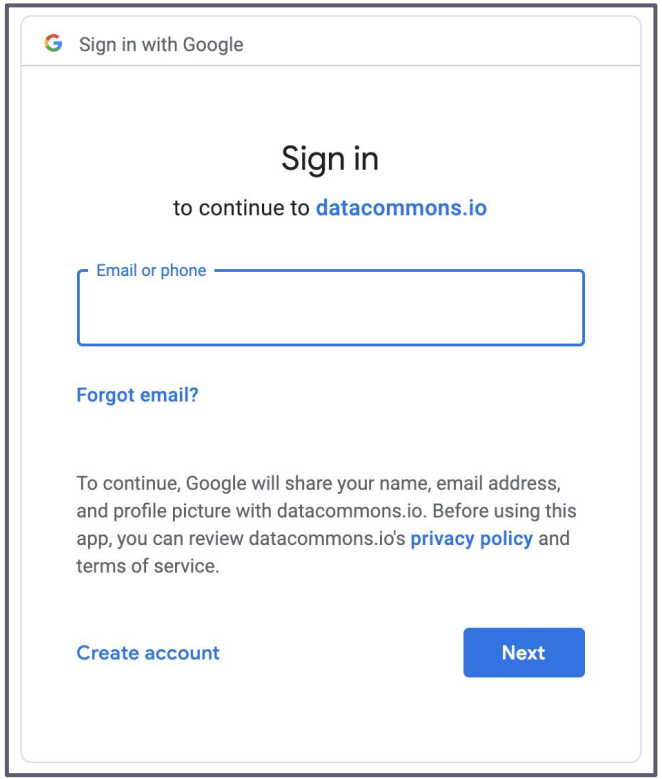
GEN3 Data Commons

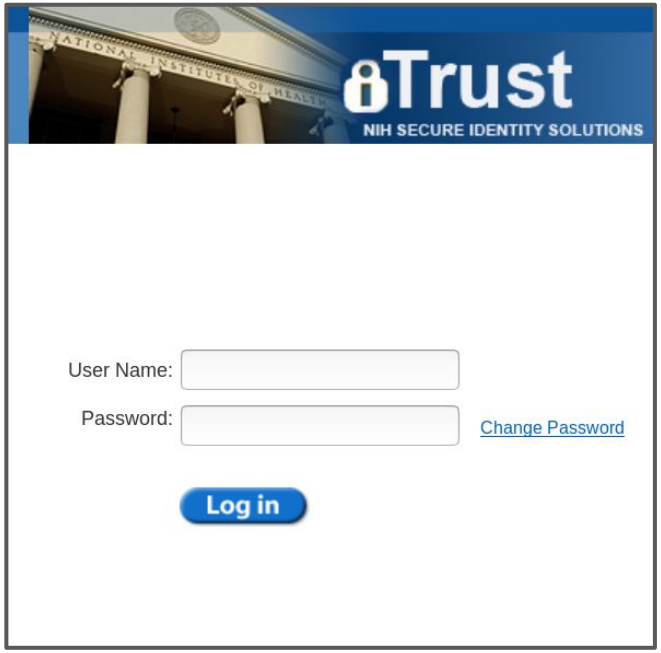
Center for Translational Data Science
AT THE UNIVERSITY OF CHICAGO

Authentication: Gen3 Portal

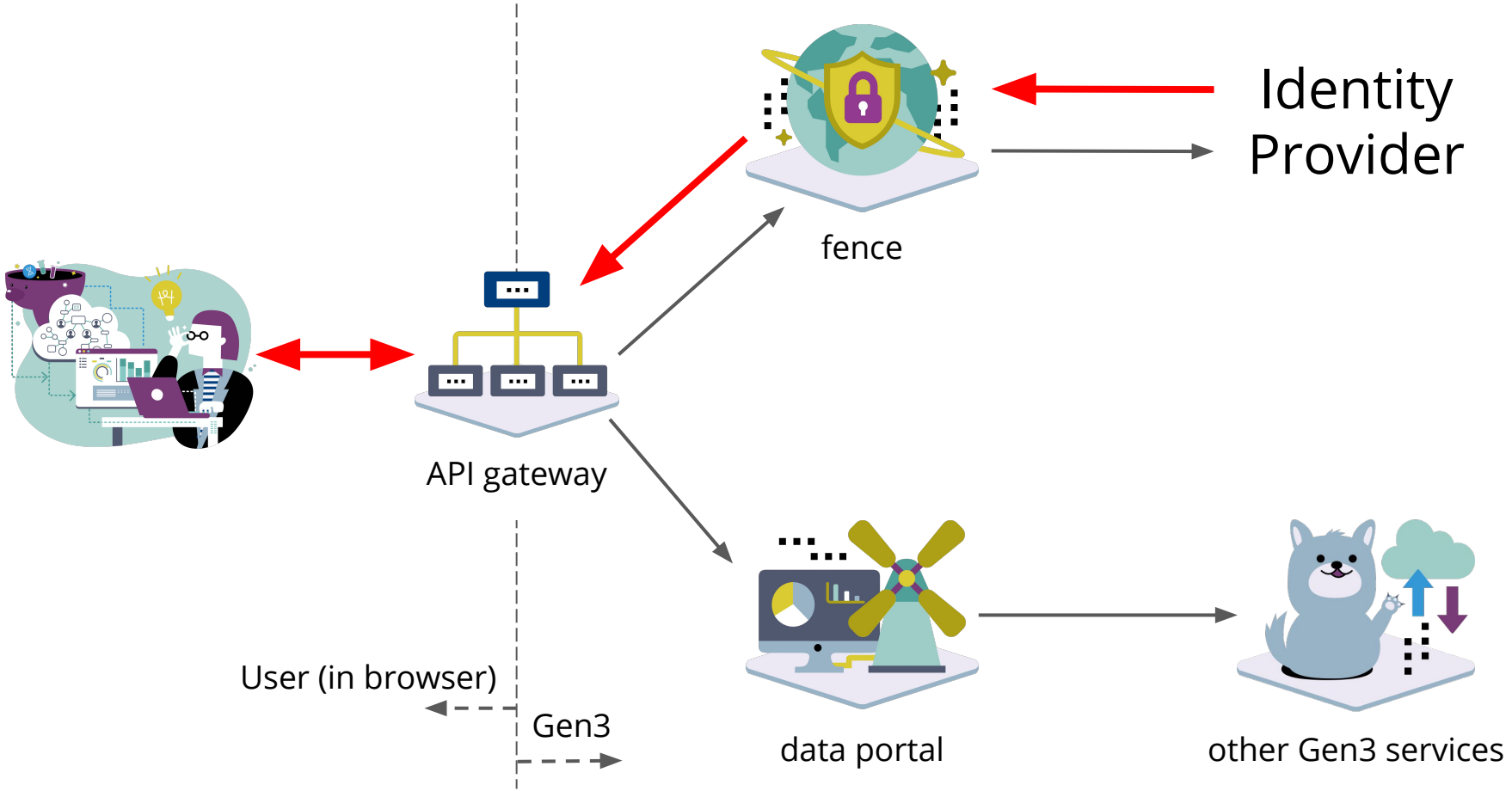


External IDP Login

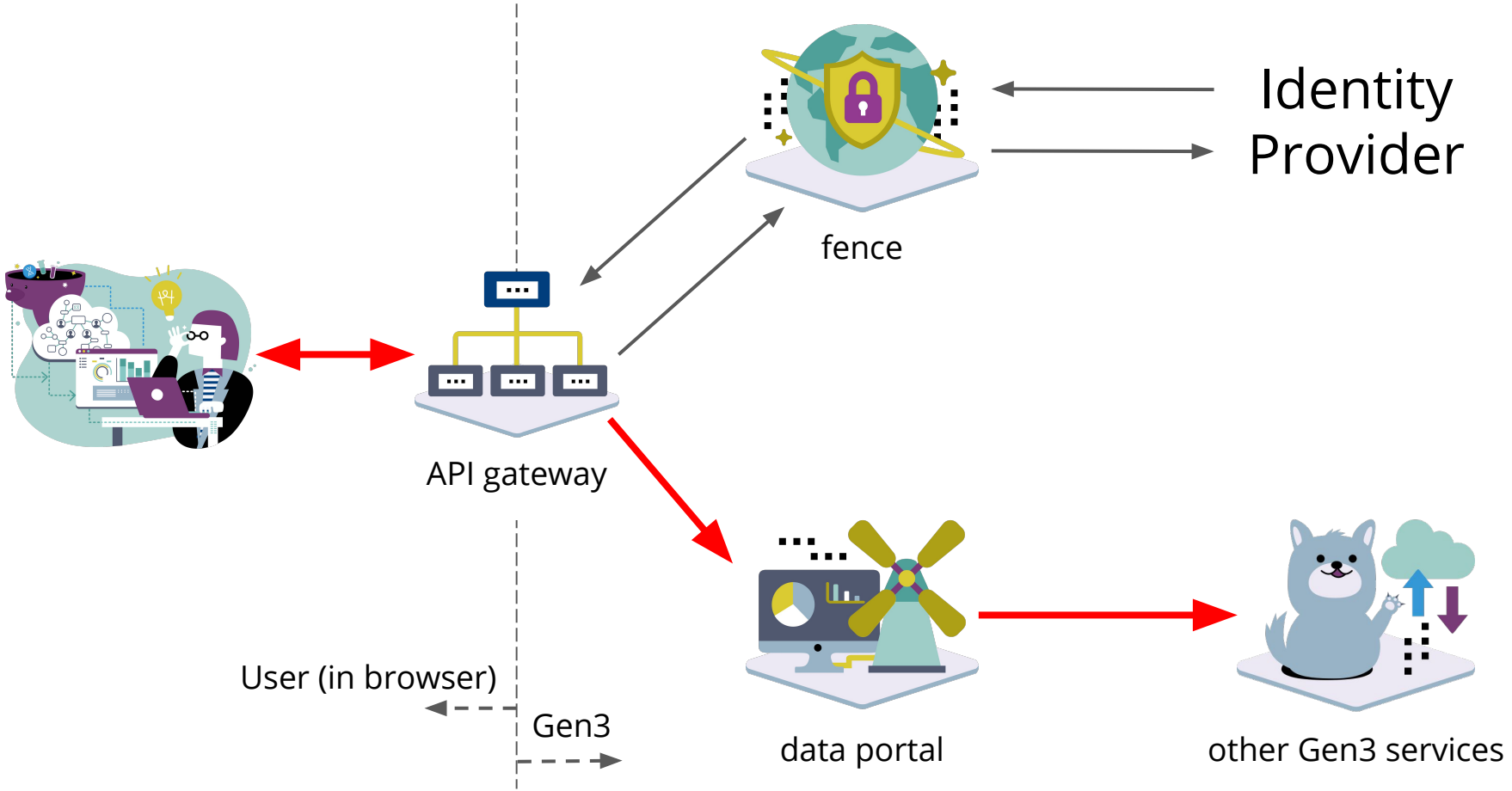




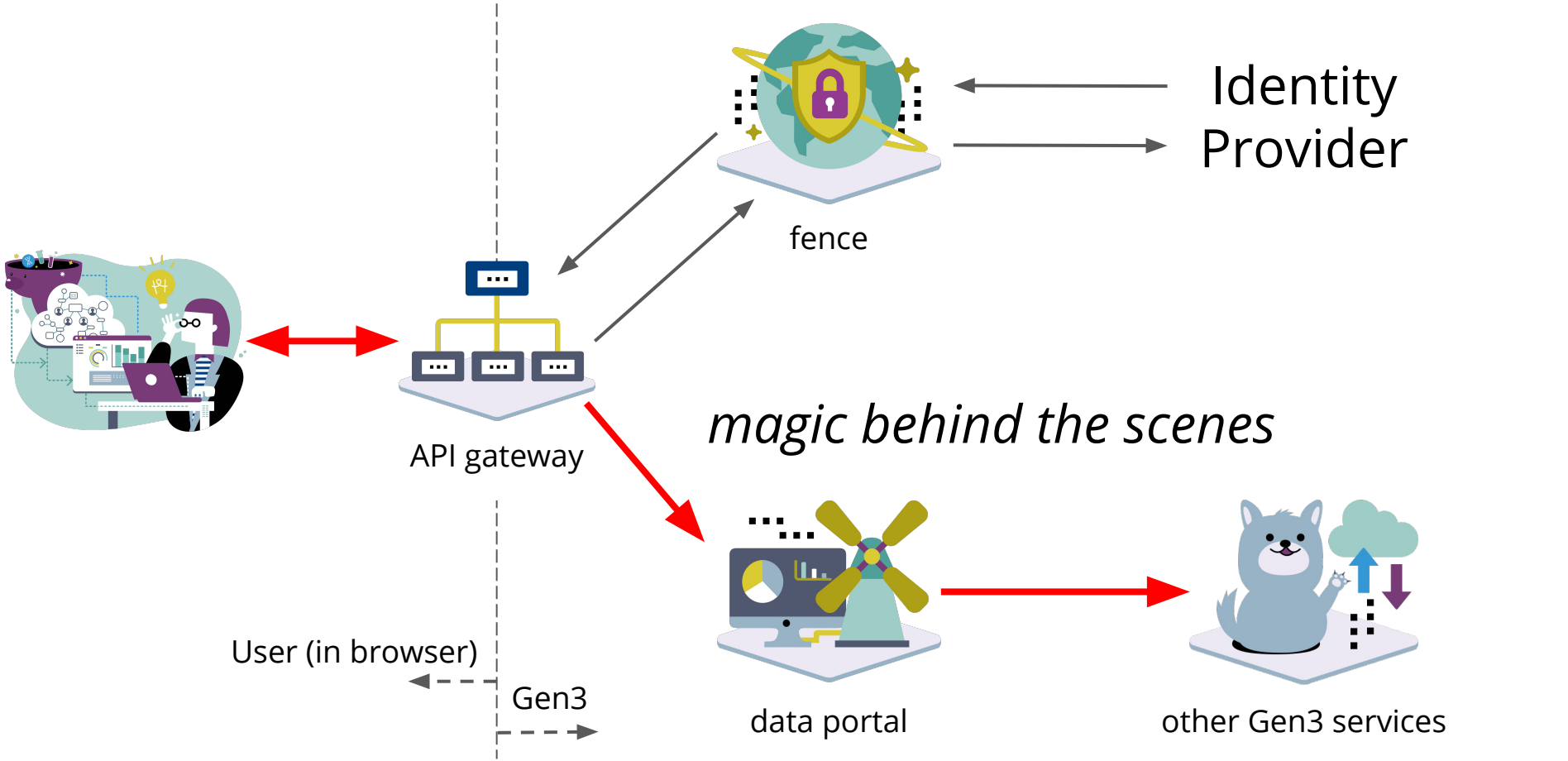
Authentication: Gen3 Portal



Authentication: Gen3 Portal



Authentication: Gen3 Portal



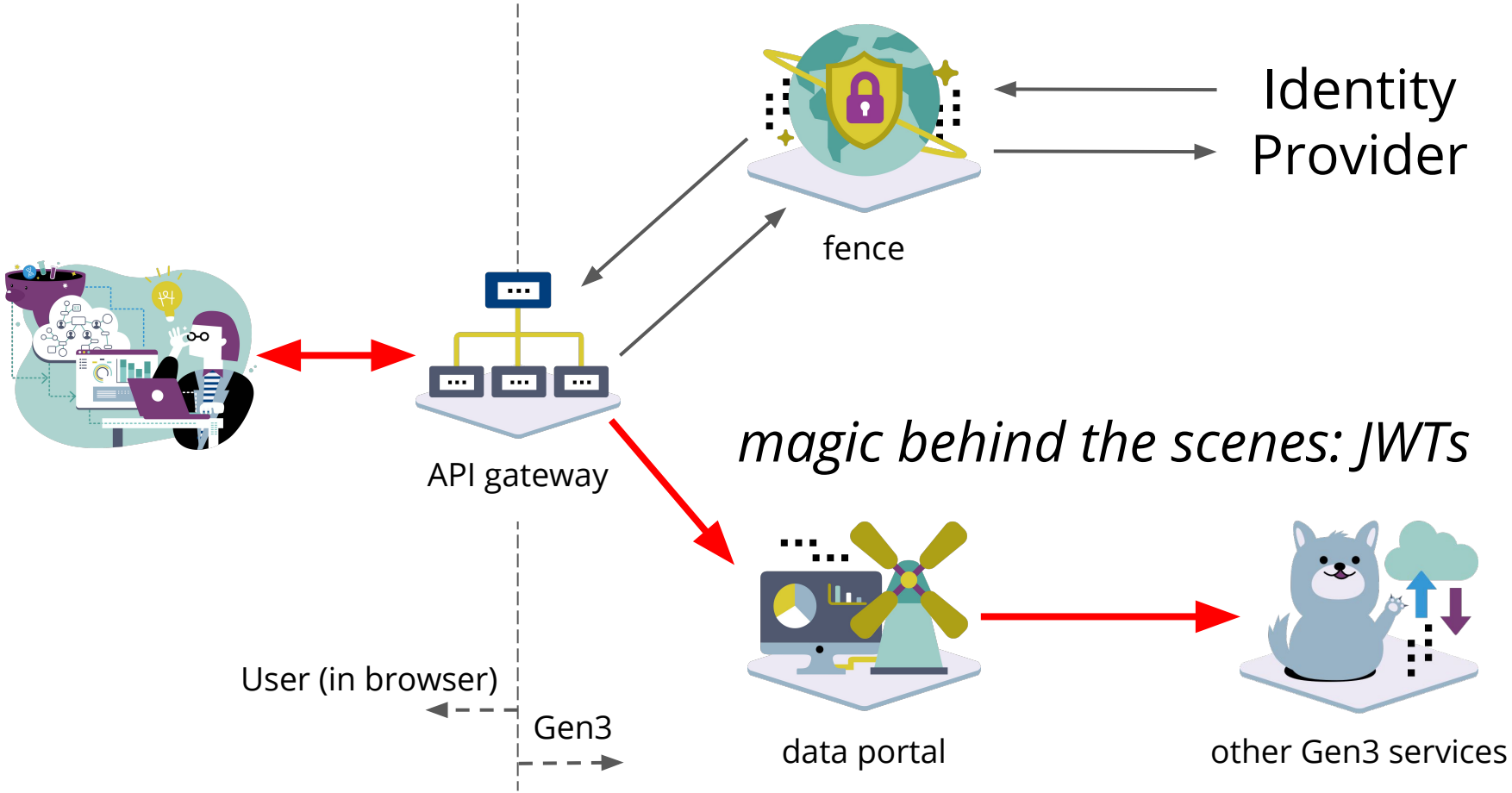
Behind the Curtain: JWTs

GEN3 Data Commons

- Cryptographically signed by fence
  - Use tokens for authentication
  - Any service can verify that a token was issued by the fence instance it expects
- Contains user information
  - User tokens for authorization
- Open source libraries for working with JWTs
  - jwt.io for list of all libraries
  - We use:
    - github.com/mpdavis/python-jose
    - github.com/jpadilla/pyjwt


```
{
  "sub": "7",
  "azp": "test-client",
  "pur": "access",
  "aud": ["openid", "user"],
  "context": {
    "user": {
      "is_admin": false,
      "name": "test",
      "projects": {
        "test": ["read", "create", "upload"]
      }
  },
  "iss": "https://portal.occ-data.org/",
  "jti": "2e6ade06-5afb-4ce7-9ab5-e206225ce291",
  "exp": 1516983302,
  "iat": 1516982102
}
```

...Auth Flow, Continued

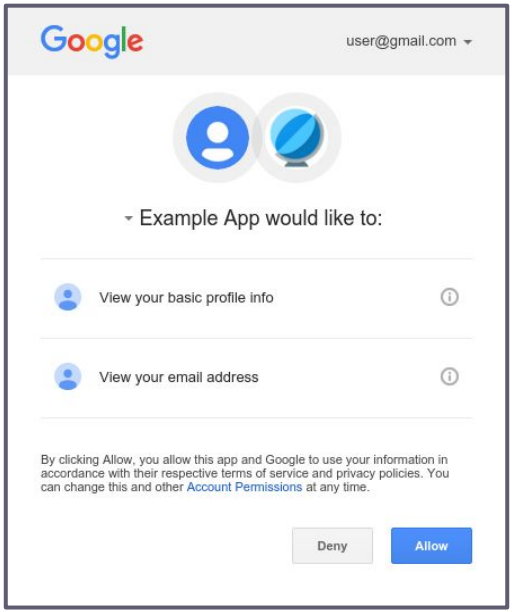


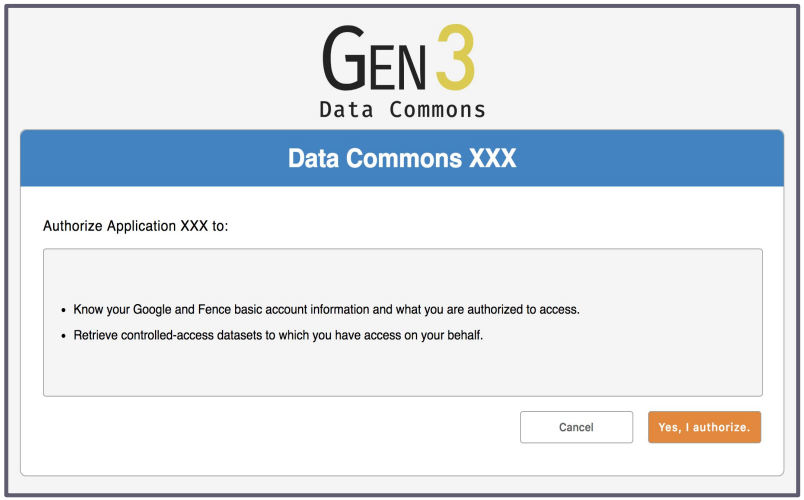
Interoperability Using OAuth2 & OpenID Connect

What is OAuth2?

GEN3 Data Commons

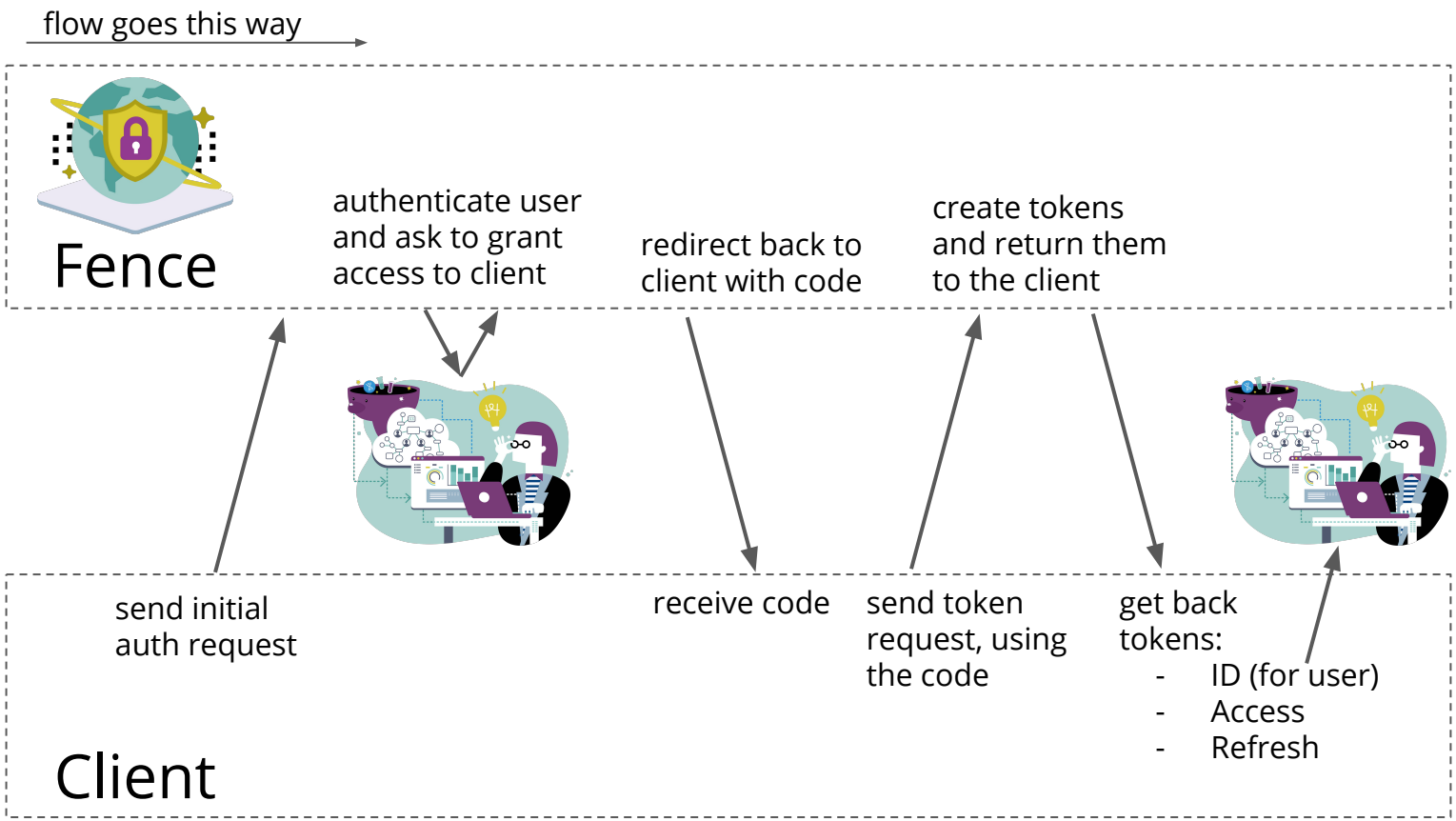
OAuth2 is a protocol allowing an application to securely access a resource on behalf of a user





Overview of OAuth2 & OpenID Connect

GEN3 Data Commons



OAuth/OIDC Clients

GEN3 Data Commons

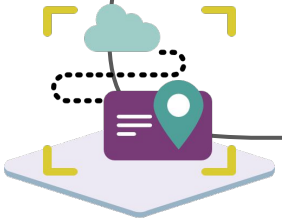
- Examples
  - Our service to handle auth in workspaces: github.com/uc-cdis/workspace-token-service
  - Data Commons Framework: dcf.gen3.org/data-access-with-dcf
- Creating an OAuth client
  - Python packages for OAuth clients
    - github.com/lepture/authlib
    - github.com/requests/requests-oauthlib
- OAuth client demo

3rd Party Data Access

Client Application

Gen3

fence

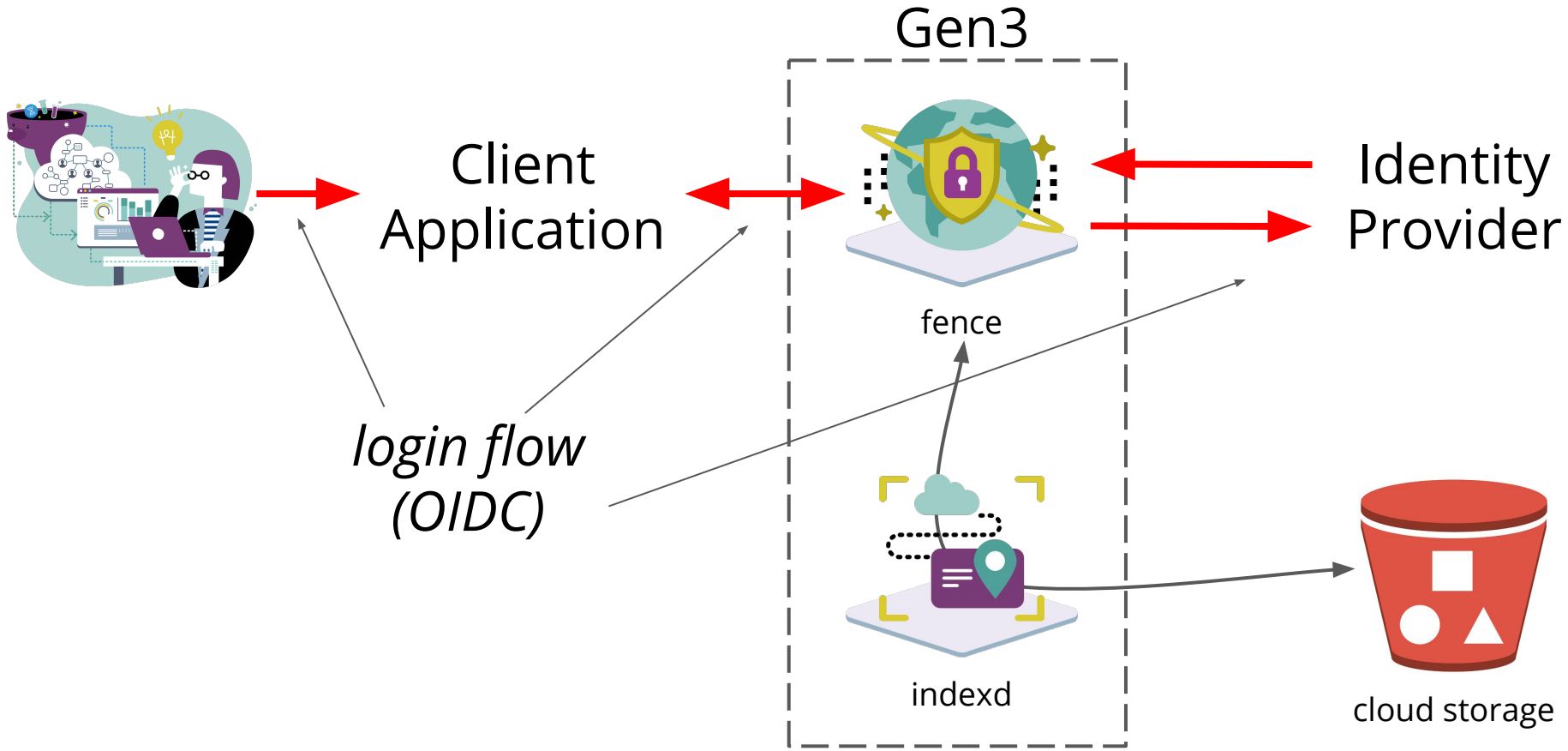


indexd

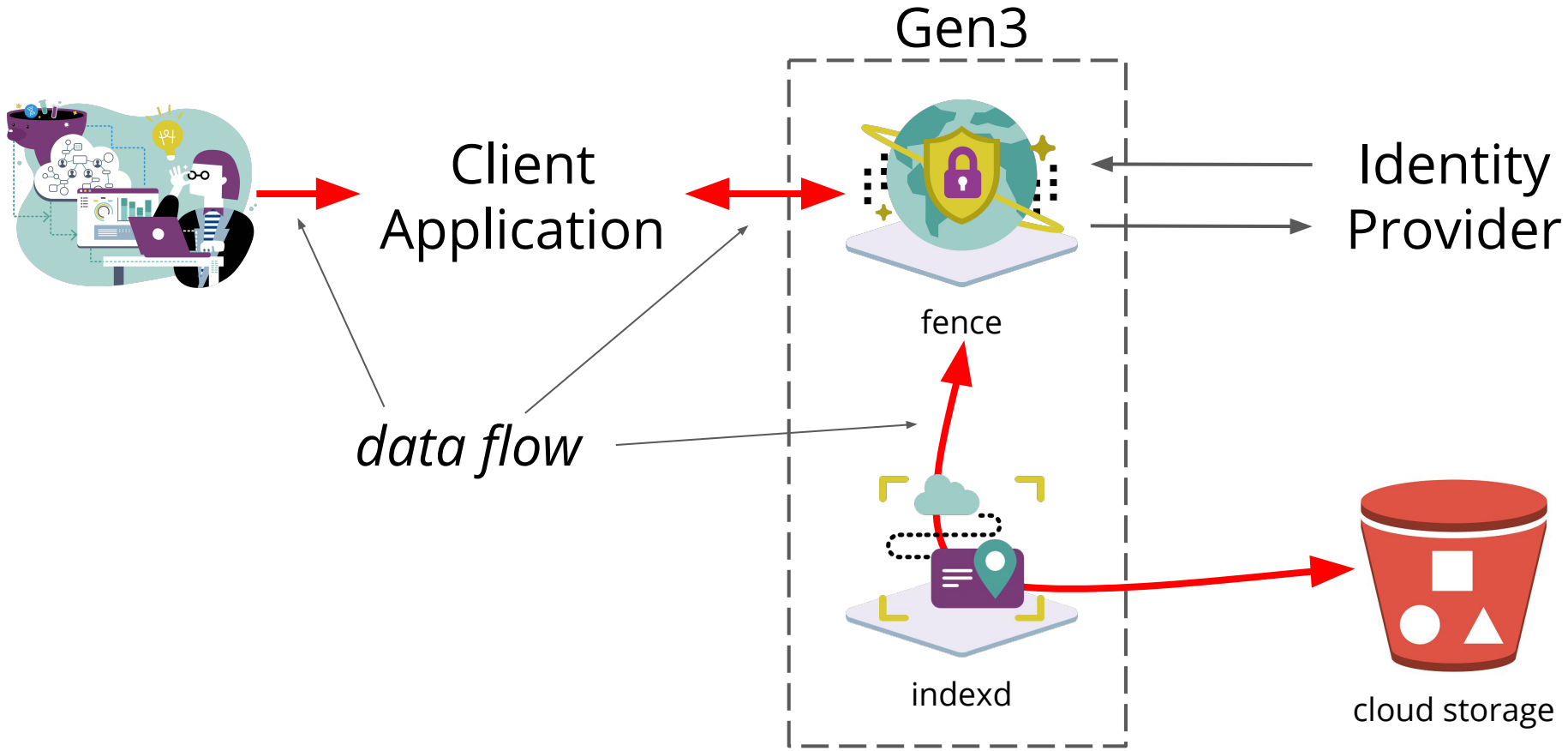
Identity Provider

cloud storage

3rd Party Data Access



3rd Party Data Access



Indexd

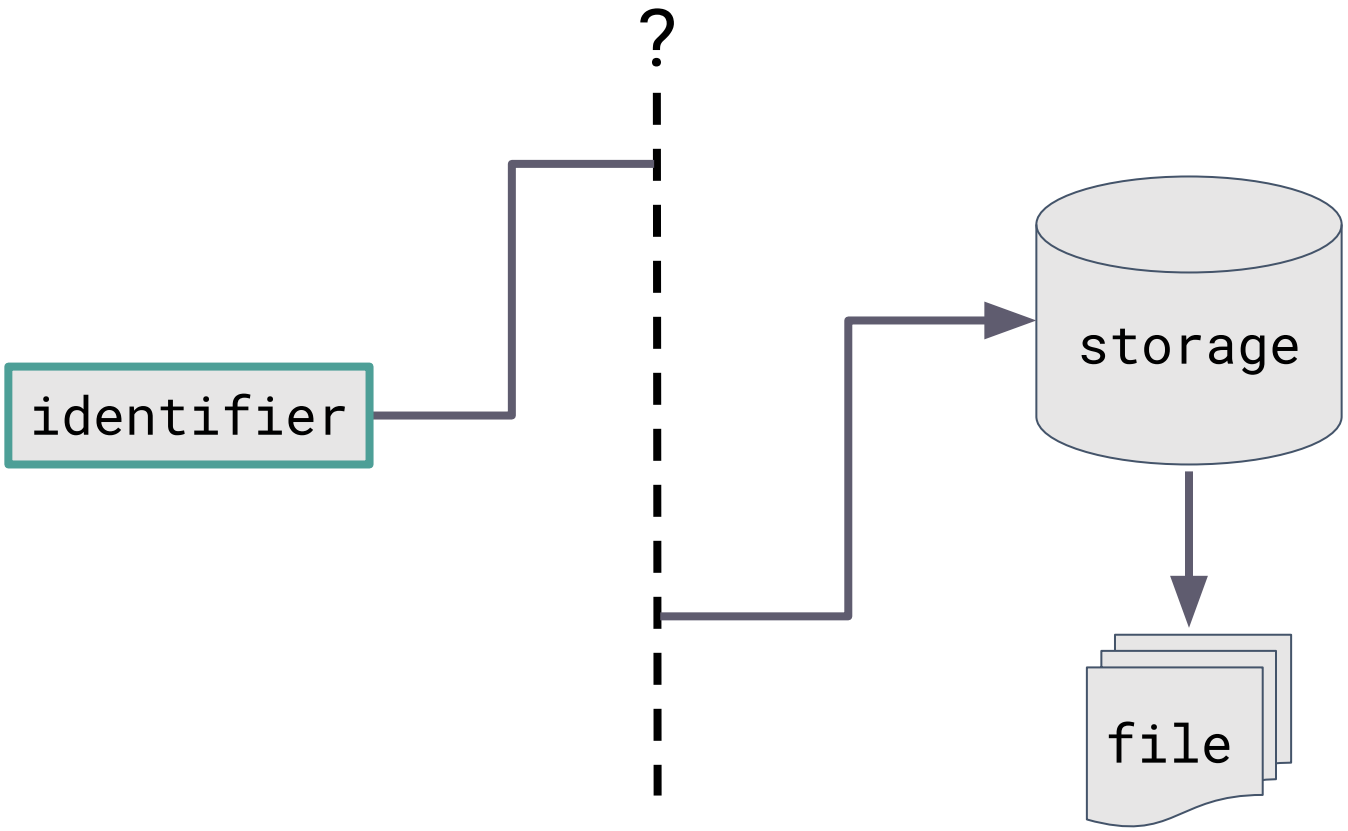
Gen3 data indexing service

Indexd

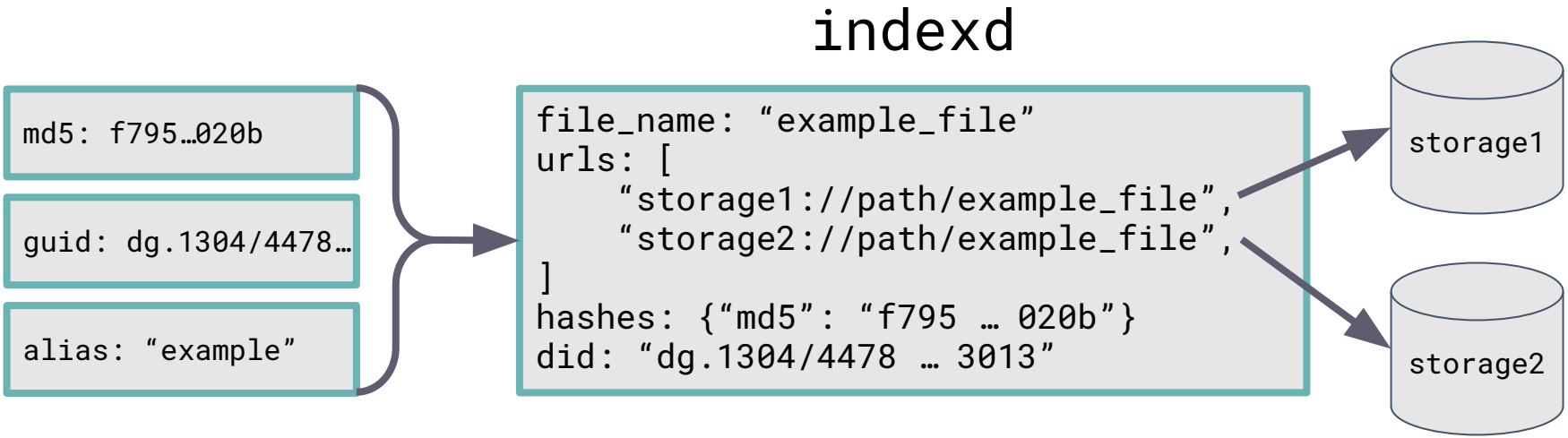
Gen3 data **indexing** service

indexing: locate data with easily used identifiers

How to Locate Data?



How to Locate Data? indexd!



- One level of abstraction over the data
- indexd maintains pointers to the data; if you can get the pointer (via hash/ID/alias), you know where to find the data
- Accessible by human-readable alias

Revisit: 3rd Party Data Access

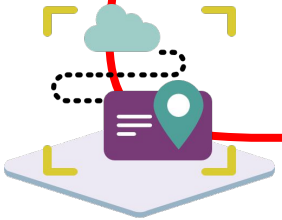
Client Application

Gen3

fence

Identity Provider

data flow

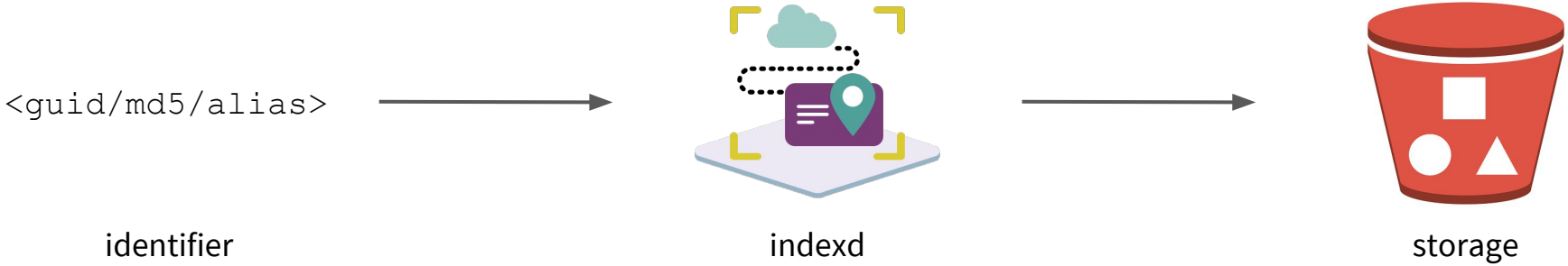


indexd

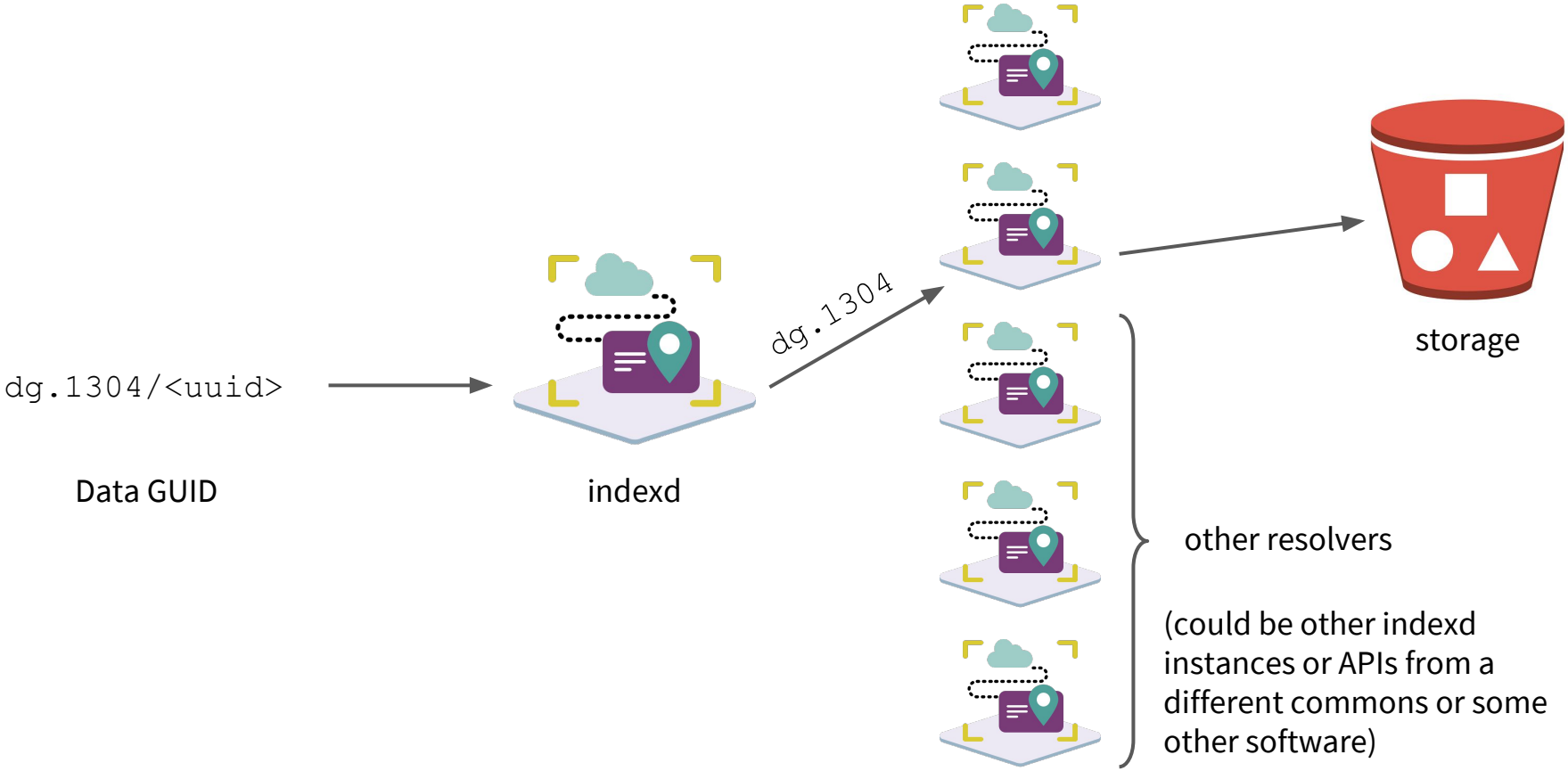
cloud storage

Distributed Resolution

Basic Case



Distributed Resolution



dataguids.org

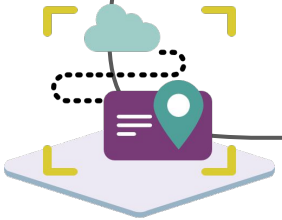
dg.4503/3625833a-404e-46c8-af16-3fb50a23f11c

3rd Party Data Access

Client Application

Gen3

fence



indexd

Identity Provider

cloud storage

Learn More

- github.com/uc-cdis
- gen3.org
- Gen3 Community on Slack
- dcf-support@datacommons.io
- ctds.uchicago.edu

Selected Data Commons Using Gen3

GEN3 Data Commons

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

Data STAGE

NIAID DATA HUB PILOT

NIH National Heart, Lung, and Blood Institute

AnVIL

BRAINCOMMONS

NIH National Human Genome Research Institute

Next Webinar

GEN3 Data Commons

Gen3 Data Modeling

Sheepdog& Peregrine

Herding Data Submissions

& Hunting Down Data

Thursday, May 9, 2019
1:00 PM-2:00 PM (CST)

Questions?

GEN3 Data Commons

Bonus Slides

Upcoming Features in Fence

GEN3 Data Commons

- Centralized authorization
  - Support role-based access control (RBAC) across an entire data ecosystem
  - Allow clients to configure their own resources and user access
  - Support more sophisticated access control in existing services, such as indexd
- Admin portal
  - Web interface for managing users, groups, resources, and access control

Fence: Extensible Configuration

GEN3 Data Commons

- Identity providers
  - Other OAuth2/OIDC providers, e.g. Google
  - Shibboleth, e.g. NIH iTrust
  - Multi-tenant: use other fence instances as IDP
- Specifying user access
  - Load a YAML file listing user privileges
  - Sync from dbGaP

“Narrow Middle Architecture”

GEN3 Data Commons