



Whole Tale: Merging Science and Cyberinfrastructure Pathways

Matthew Turk on behalf of Bertram
Ludaescher, Kyle Chard, Niall Gaffney,
Matthew B. Jones, Jaroslaw
Nabrzyski, Victoria Stodden



wholetale.org





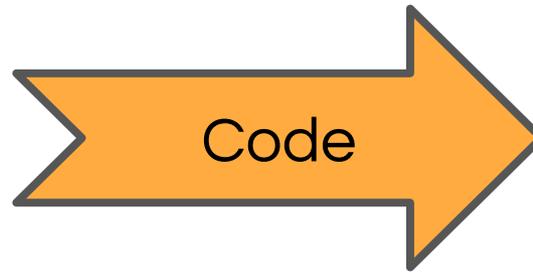
Vision



“It used to be, you’d publish a paper...”

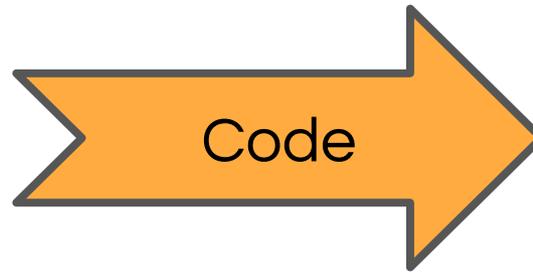
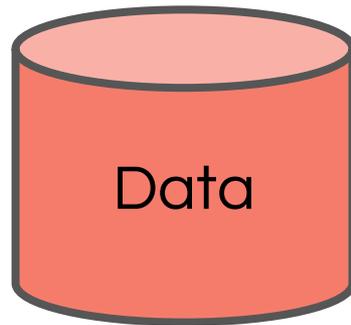


Whole Tale Vision



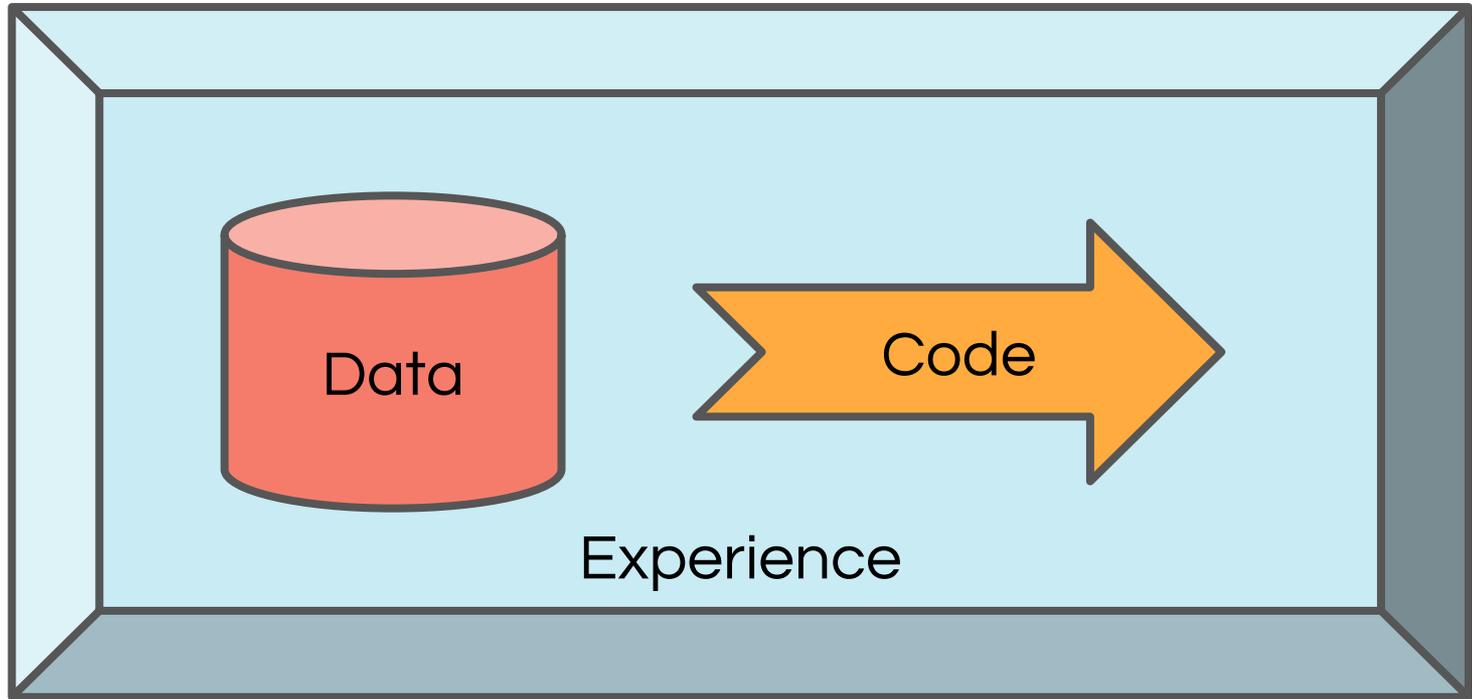


Whole Tale Vision





Whole Tale Vision





Core to our mission is *active, meaningful* engagement with open source and research communities.



Girder

Data management platform



- Ingestion of Data
- Frontends to Data
- Abstracted Storage
- Data Namespacing
- Identity Management
- Event System

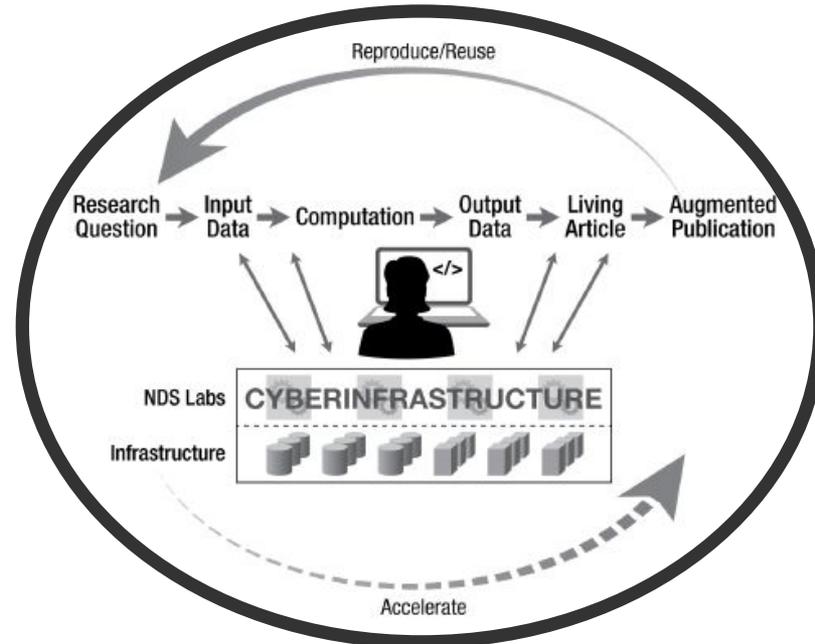


Overview



Whole Tale: What's in a name?

- (1) Whole Tale \Leftrightarrow Whole **Story**:
 - **Support** (computational & data) **scientists**
 - ... along the **complete research lifecycle**
 - ... from **experiment** to (new kind of) **publication**
 - ... and back!





Whole Tale: What's in a name?

- (2) Whole Tale ⇔ For the Long Tail of Science
 - *“Big data & compute for mere mortals”*

nature neuroscience

Home | Current issue | Comment | Research | Archive | Authors & referees | About the journal

home » archive » issue » commentary » full text

NATURE NEUROSCIENCE | COMMENTARY

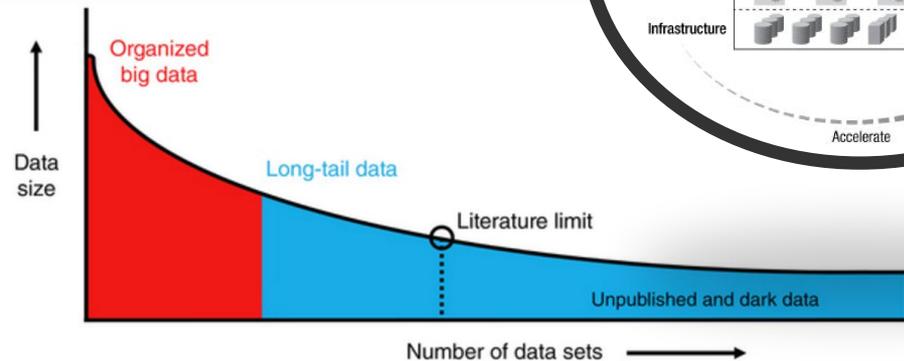
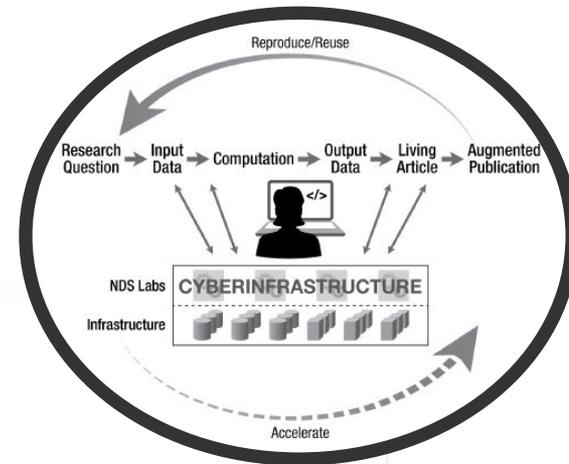
Focus on big data

Focus issue: November 2014 Volume 17, No 11

- Contents
- Editorial
- Commentaries
- Perspective
- Reviews

Big data from small data: data-sha... tail' of neuroscience

Adam R Ferguson, Jessica L Nielson, Melissa H Cragin, An Martone



Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world.



Whole Tale Vision

- The Old Way:
 - Scholarly **Publication** .. || .. Data .. || .. Code
 - The Emerging Way:
 - Scholarly **Publication** ↔ **Data** .. | .. Code
 - The New Way:
 - “Living” **Publication** ↔ **Data** ↔ **Code**
= *Computational Narrative*
 - (more easily) *Reproducible Science*
- .. participate in and share the *experience of inquiry*



Problems Facing Data Researchers

Workflow for data research is **fragmented**:

- Data comes from many sources and is **“integrated the old fashioned way”** (*email, Excel, ...*)
- Use cloud services **copying data** from (*Drop)Box, Google-Drive, ...* to local storage with a distributed directory structures to organize (and provide discovery) to data
- Data provenance is **not captured** (custom scripts, some version of a community developed and supported codebase)
- Publication of data with link to publication (never mind DOIs, DMP) is **not sufficient for reproducibility**



So what do we do about this?

- WT will leverage & contribute to **existing CI and tools** to support the **whole science story** (= run-to-pub-cycle), and providing access to big data via CI and compute for **long tail** researchers.

➔ ***Integrate tools to simplify usage and promote best practices***

- NSF CC*DNI DIBBS:
 - 5 Institutions, 5 Years (\$5M total)
 - Cooperative Agreement

The Whole Tale
Merging Science and Cyberinfrastructure Pathways

Whole Tale will enable researchers to examine, transform, and then seamlessly republish research data that was used in an article. As a result, these "living articles" enable new discovery by allowing researchers to construct representations and syntheses of data.



Specific Goals of Whole Tale

- **Expose existing CI**
 - ... through popular frontends (Jupyter, RStudio, ..)
- **Develop necessary “software glue”**
 - ... for seamless access to different CI-backend capabilities
- **Enhance data-to publication lifecycle**
 - ... by empowering scientists to create computational narratives in their usual programming environments



Approach



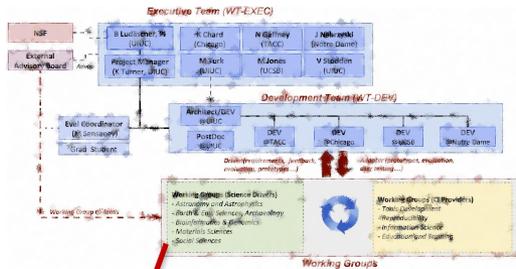
WT will integrate established CI components, creating a simple, unified environment to use, share, and publish data and workflows:

1. **Unified Authentication** via Globus Auth
2. **Abstracted Storage** Layer with a unified namespace
3. **Integrated** Python and R APIs with **Jupyter Notebook Environments**
4. **Ingest and publication** service linking data, computations, and scholarly articles
5. **NextCloud integration** for “Dropbox like interface”
6. **Event System** to react to changes (e.g. new data published)
7. **Data Dashboard** to ease data management and service interactions

→ Capture full workflow via notebooks, scripts, and applications to be published along with data and research publications



Iterative Design through Working Groups



*Merging Science & CI Pathways
... through Working Groups!*

Working Groups (Science Drivers)

- Astronomy and Astrophysics
- Earth & Env. Sciences, Archaeology
- Bioinformatics & Genomics
- Materials Sciences
- Social Sciences



Working Groups (CI Providers)

- Tools Development
- Reproducibility
- Information Science
- Education and Training

*Working Groups Driving Use Cases
and Adoption*

***Iterative
Design!***

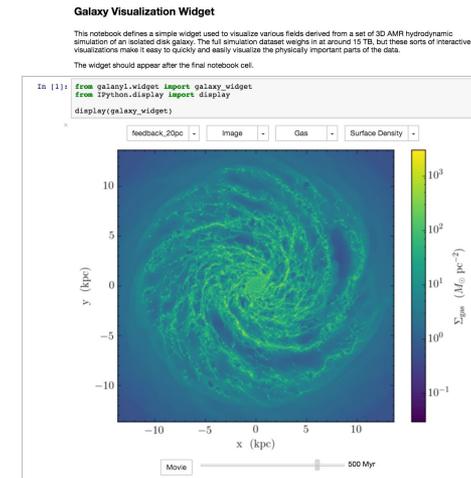
*Working Groups to Provide Key
Components*



Science Pathway (WG Example): Astronomy



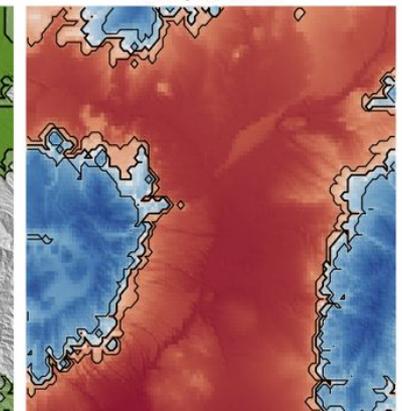
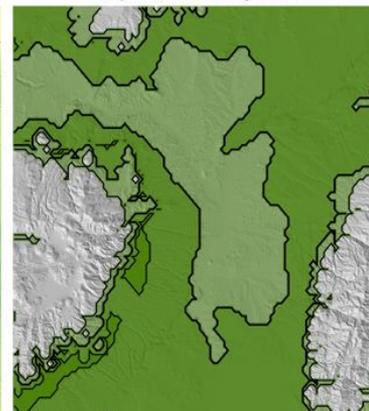
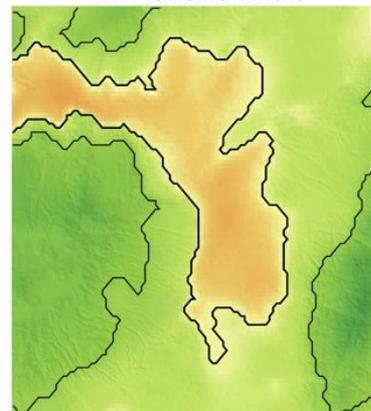
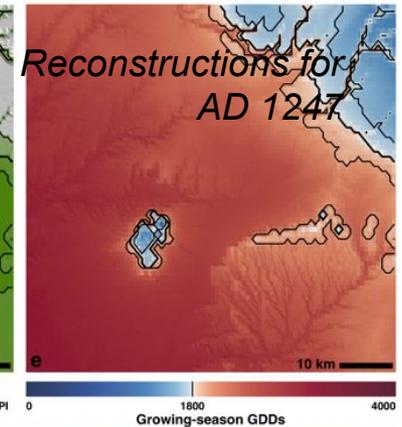
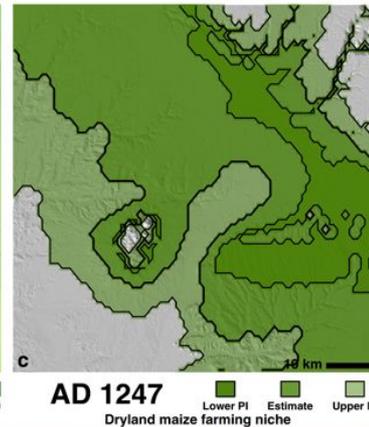
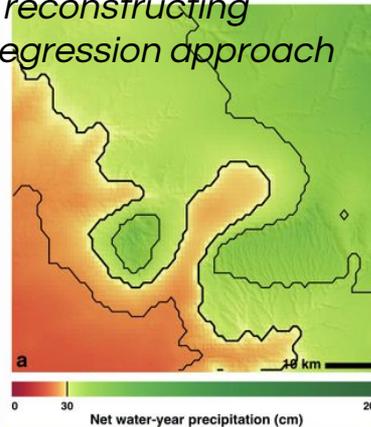
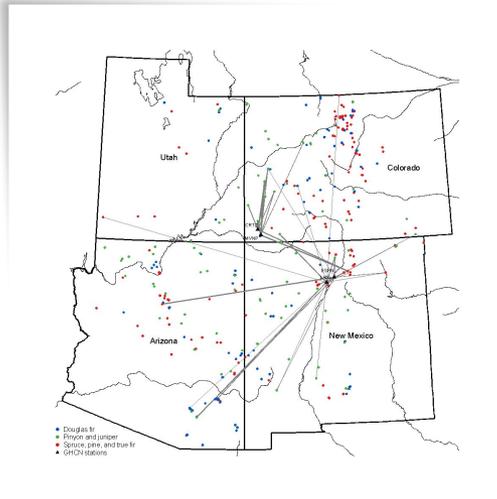
- Enabling direct analysis and collaborative research on simulation outputs stored in Whole Tale enabled repositories via user-supplied Python scripts.
- Tools, e.g., yt, astropy, will provide advanced, customizable analysis and visualization, leveraging Jupyter for provide the scripting support.
- Federation will allow jobs to move to data or visa versa where appropriate





Science Pathway (WG Example): Computational Archaeology

Map showing the "selected" trees for reconstructing precipitation at four sites in the CAR regression approach (Correlation-Adjusted corRelation).



K. Bocinsky, T. Kohler, A 2000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. *Nature Communications*. doi:10.1038/ncomms6618



Science Working Groups Logistics

- Virtual meetings
- Optionally:
 - Hackathon and opportunistic face-to-face
 - Summer Internships
- **Co-Leads:**
 - 1 community member
 - 1 WT-exec member
- **EAB-Link:**
 - 1 EAB member (oversight)
- Several WGs starting in November



WT External Advisory Board (EAB)

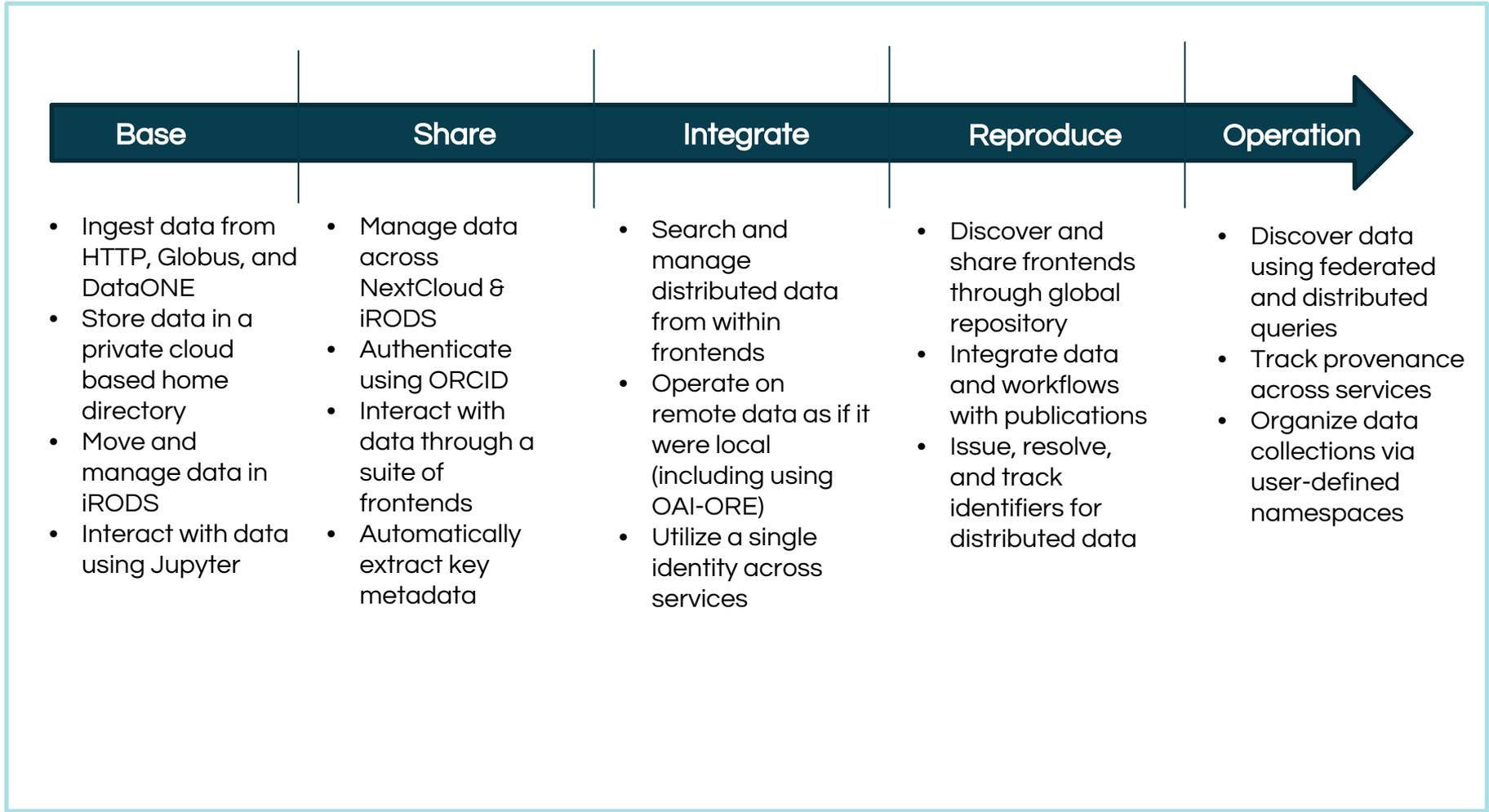
- The **WT-EAB**
 1. advises the WT-exec on project direction and outcomes
 2. suggests and monitors Working Group activities and outcomes

- 1st (virtual) meeting planned **November 2016**
 - Introduce EAB to WT vision, progress, WG plans
 - **Solicit feedback & suggestions on Working Groups**

- 2nd (virtual) meeting planned during **2017**
 - **Present progress, solicit feedback**



High-level Yearly Milestones, Phases





Identity



Whole Tale Identity, Authentication and Authorization Landscape



Identities

Cyberinfrastructure



Identity, Authentication and Authorization Requirements

- Single identity that can be used across to access many services
- Support for diverse researchers' identities
- Enable delegated actions on behalf of users
- Common representations for identities
- Common interfaces for integrating in external services
- Support headless authentication





Globus Auth: Brokering Authentication and Authorization

- Brokers authentication and authorization interactions between:
 - end-users
 - identity providers: InCommon, XSEDE, Google, portals
 - services: resource servers with REST APIs
 - apps: web, mobile, desktop, command line clients
 - services acting as clients to other services
- Provides web interfaces for managing identities and approving access to “resources”
- Provides APIs for obtaining delegated access tokens that can be used within a requested “scope”
- Supports standard auth models (OAuth, OIDC)



Authn and Authz workflow

username: **chard@uchicago.edu**
id: **de305d54-75b4-431b-adb2-eb6b9e546014**
identity_provider: **uchicago.edu**
display_name: **Kyle Chard**
email: **chard@uchicago.edu**



1. Login

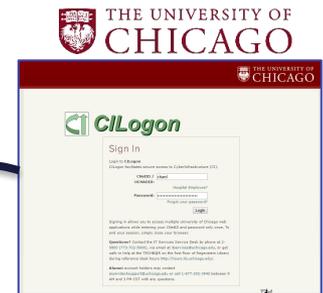
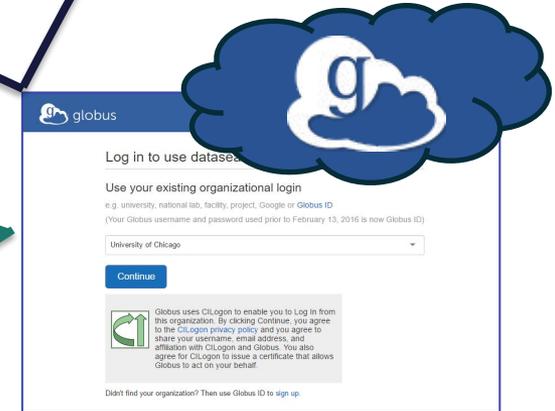


2. Request authn and authz for scopes (identity, pub repo)

4. Return auth and issue access tokens

3. Authenticate with IDP

5. Access other authorized services





Identity Linking

- Link identities from several federated IDP
 - E.g., InCommon (SAML), Google (OpenID), XSEDE (OAuth MyProxy)
- Use linked identity to authenticate to services using any one of the identities
- Perform actions permitted to any of the included identities
- Users can select a “primary identity”
- Applications can choose to require identities from a particular IDP (“effective identity”)



Current Status

- Globus Auth integrated in: Girder, Dashboard, Data Search
- ORCID integrated in DataONE
- Working to add ORCID support to Globus Auth
- Investigating methods to harmonize representations between Globus Auth and DataONE

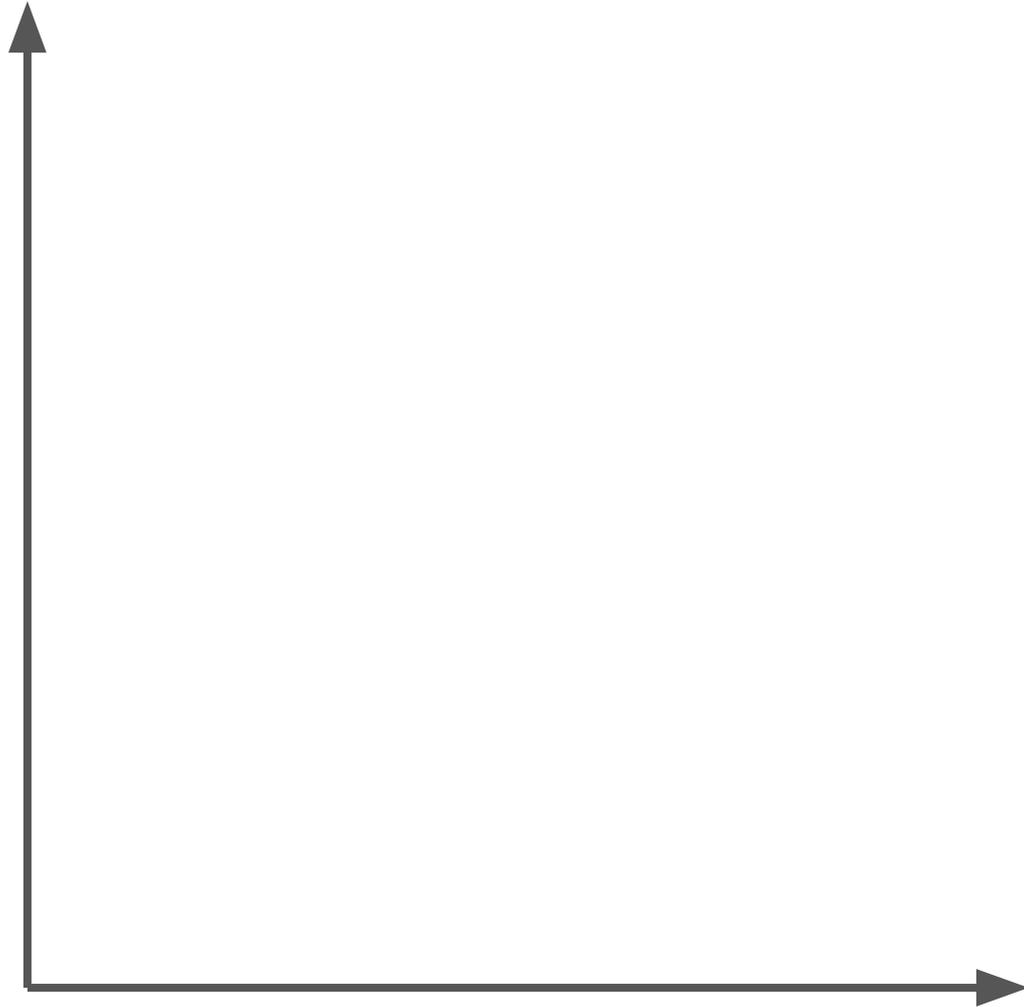


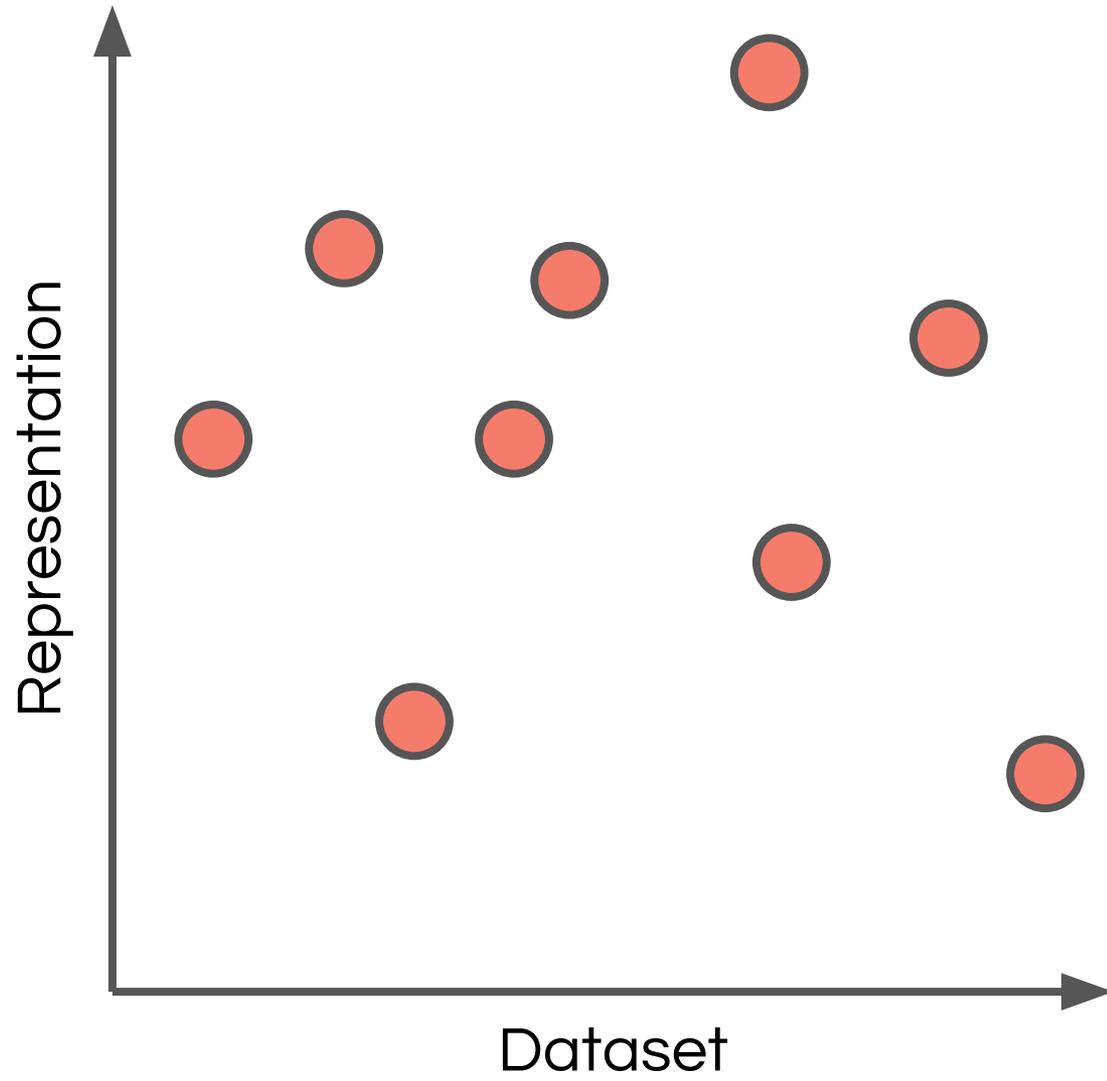
Frontends

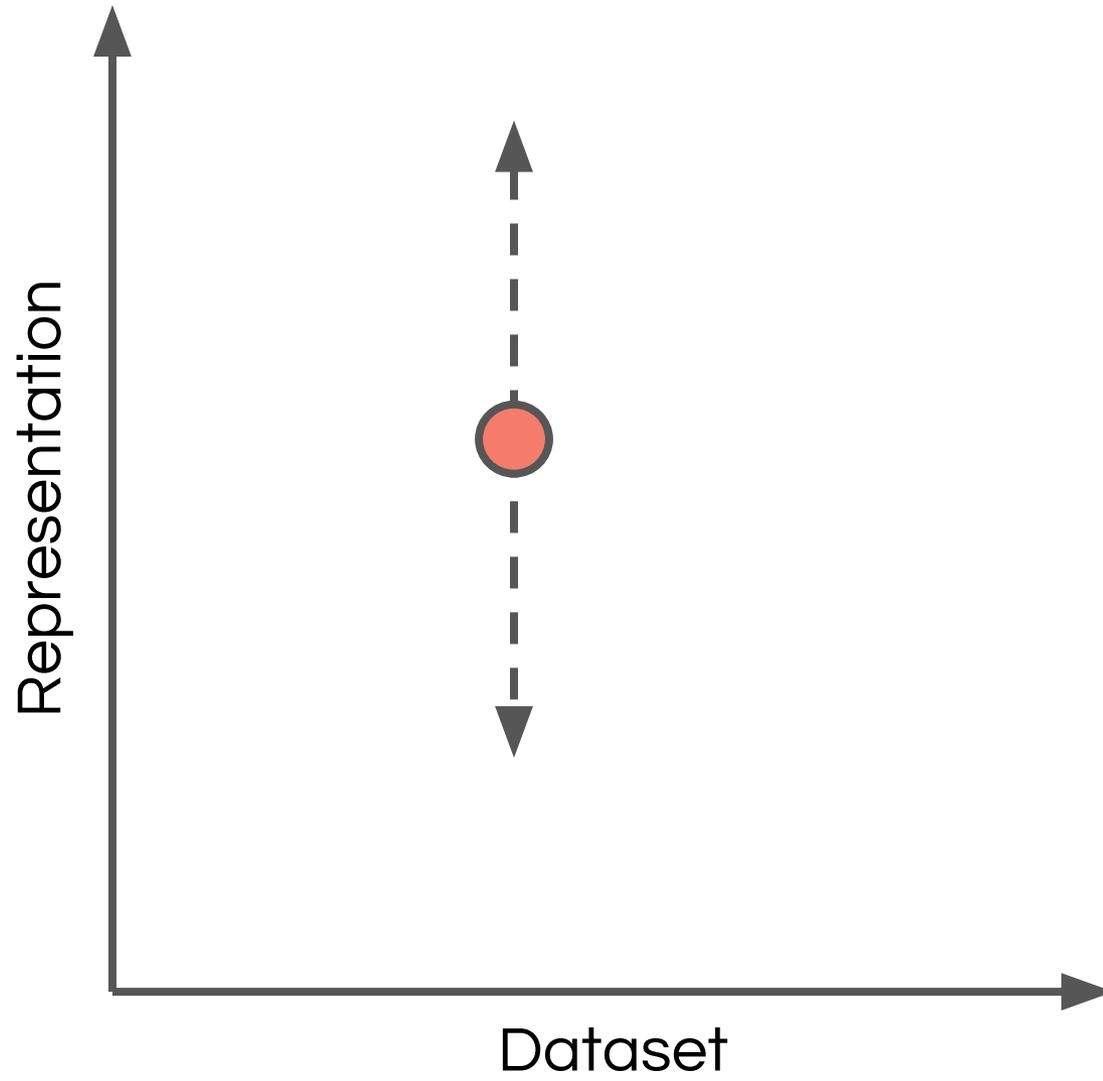


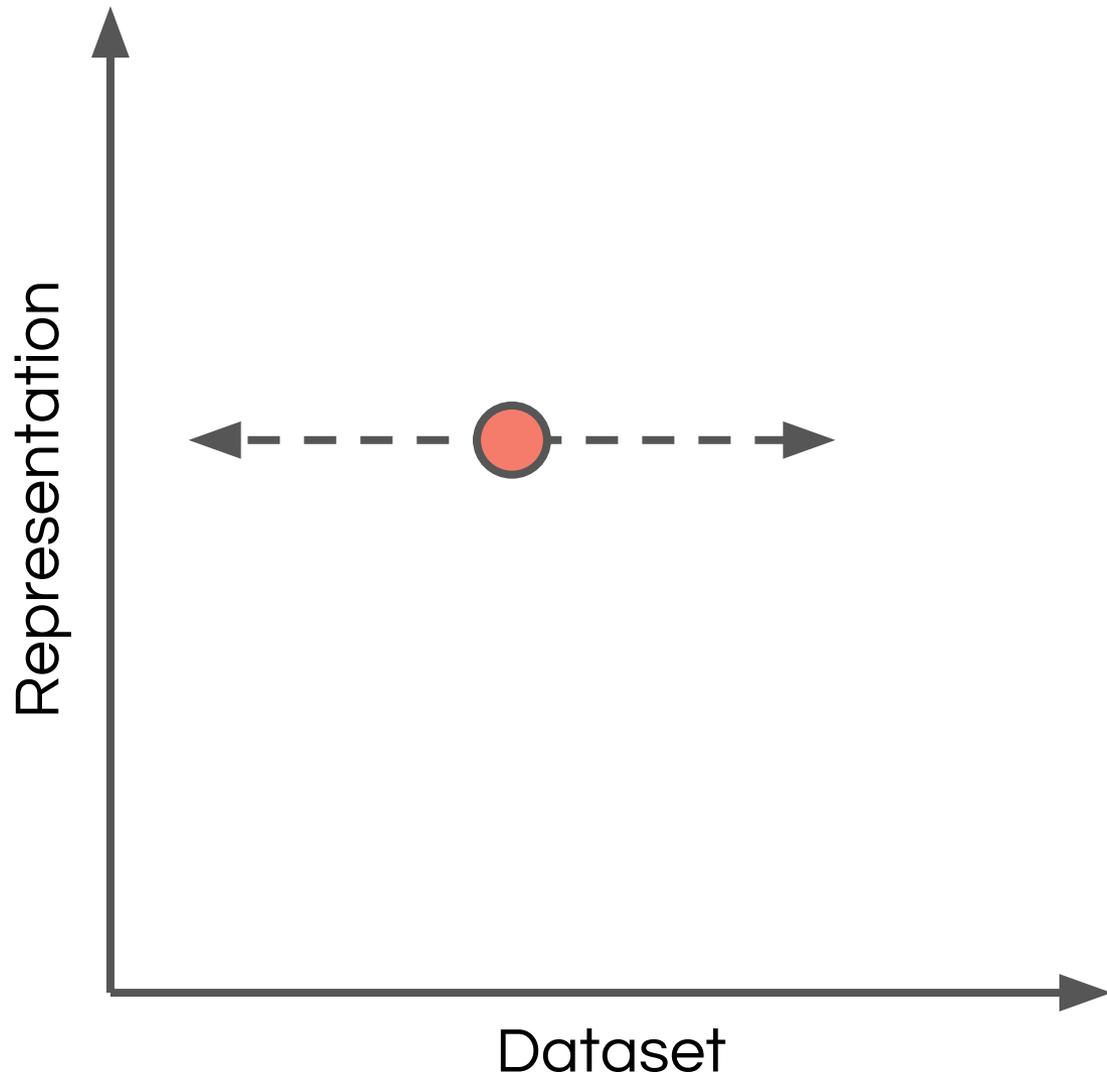
Representation

Dataset







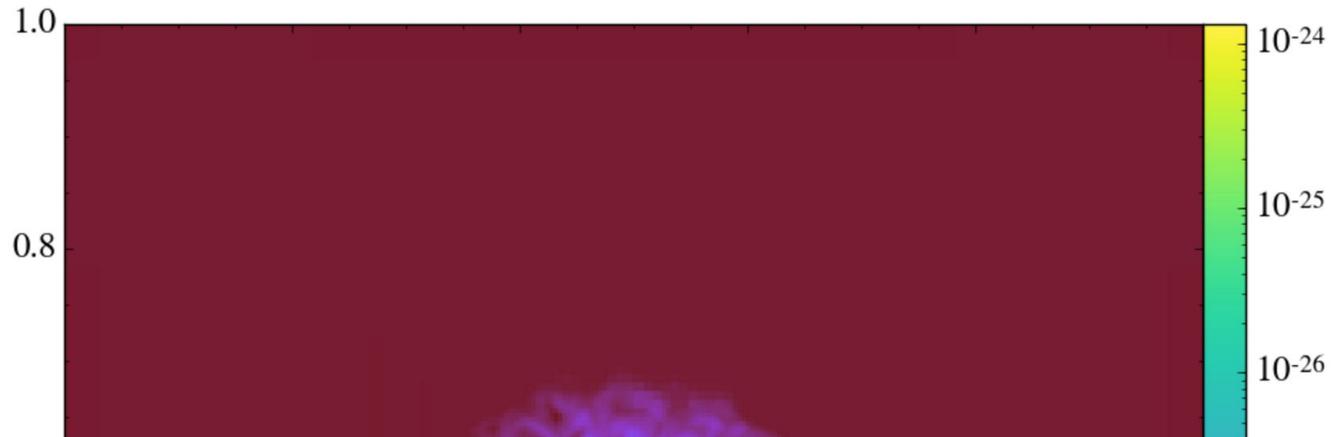


Concepts

1. Logging in
2. Talking to data
3. Storing artifacts
4. Provenance

```
In [1]: import yt
ds = yt.load("data/IsolatedGalaxy/galaxy0030/galaxy0030")
ds.r[:,0.5,:].plot("density")
```

```
yt : [INFO ] 2016-09-27 14:12:38,067 Parameters: current_time = 0.00600002000283
yt : [INFO ] 2016-09-27 14:12:38,069 Parameters: domain_dimensions = [32 32 32]
yt : [INFO ] 2016-09-27 14:12:38,071 Parameters: domain_left_edge = [ 0.  0.  0.]
yt : [INFO ] 2016-09-27 14:12:38,072 Parameters: domain_right_edge = [ 1.  1.  1.]
yt : [INFO ] 2016-09-27 14:12:38,073 Parameters: cosmological_simulation = 0.0
/opt/conda/envs/py2-dev/lib/python2.7/site-packages/matplotlib/font_manager.py:273: UserWarning: Matplotlib is building the font cache using fc-list. This may take a moment.
  warnings.warn('Matplotlib is building the font cache using fc-list. This may take a moment.')
yt : [INFO ] 2016-09-27 14:12:39,189 xlim = 0.000000 1.000000
yt : [INFO ] 2016-09-27 14:12:39,192 ylim = 0.000000 1.000000
Parsing Hierarchy : 100%|██████████| 173/173 [00:00<00:00, 4740.75it/s]
yt : [INFO ] 2016-09-27 14:12:39,301 Gathering a field list (this may take a moment.)
yt : [INFO ] 2016-09-27 14:12:43,657 Making a fixed resolution buffer of (density) 800 by 800
yt : [INFO ] 2016-09-27 14:12:50,631 Making a fixed resolution buffer of (('gas', 'density')) 800 by 800
```



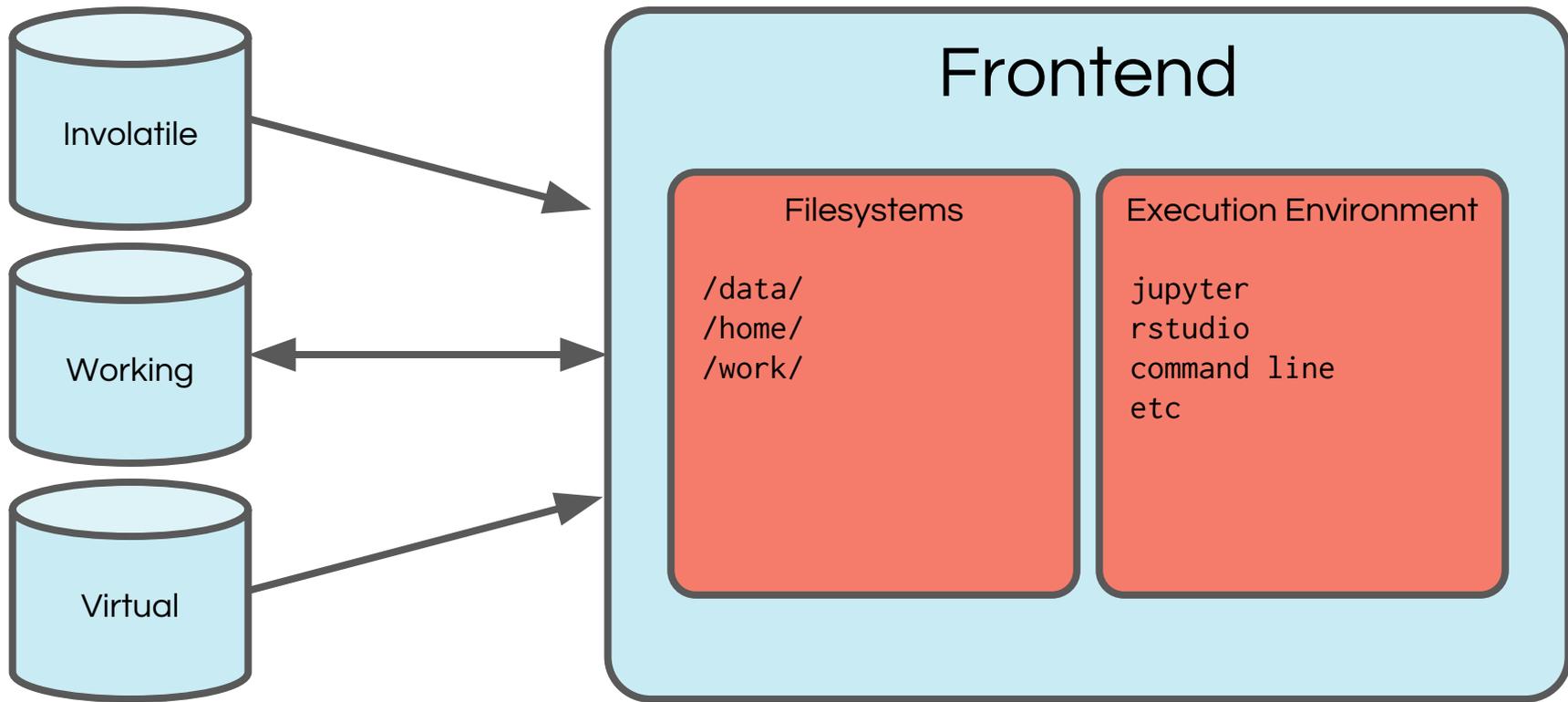
Frontend

Filesystems

/data/
/home/
/work/

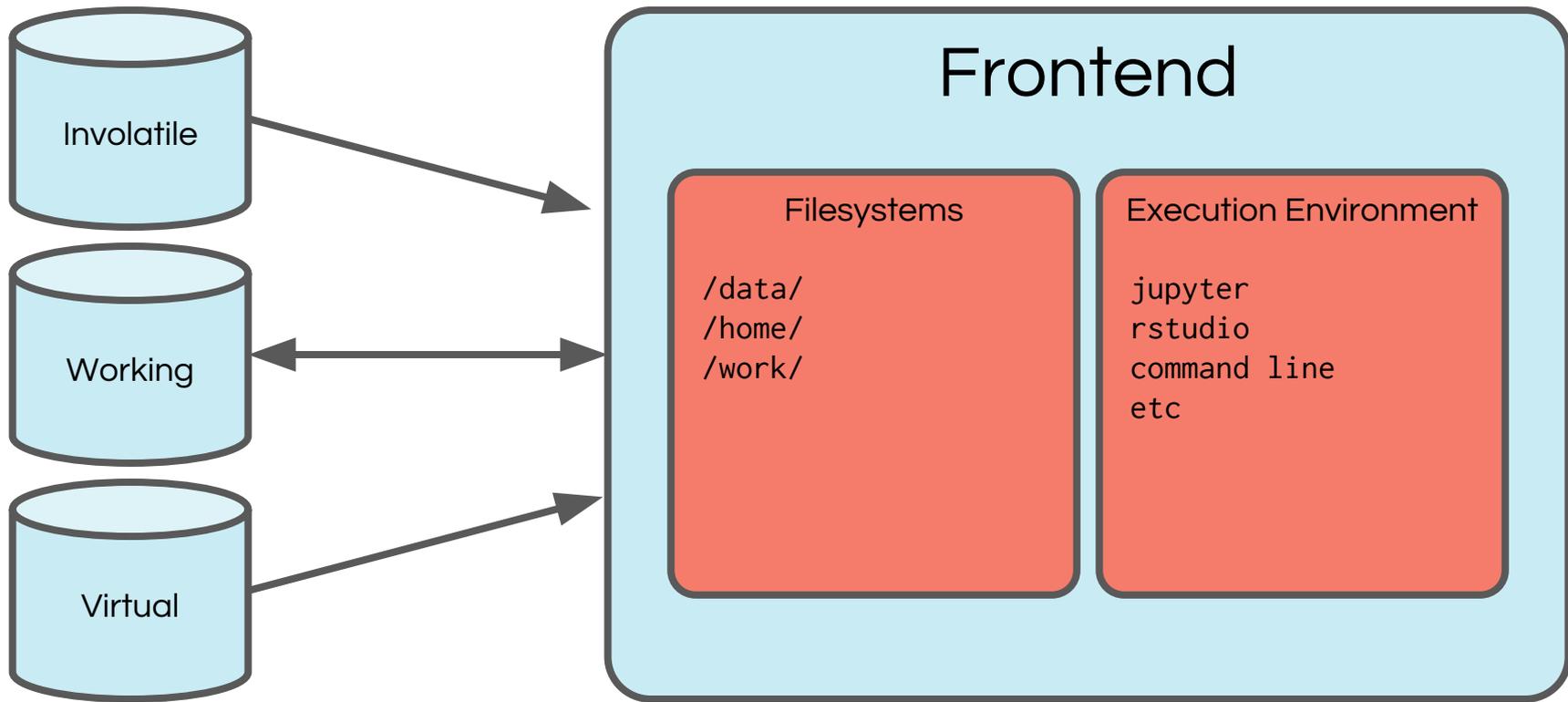
Execution Environment

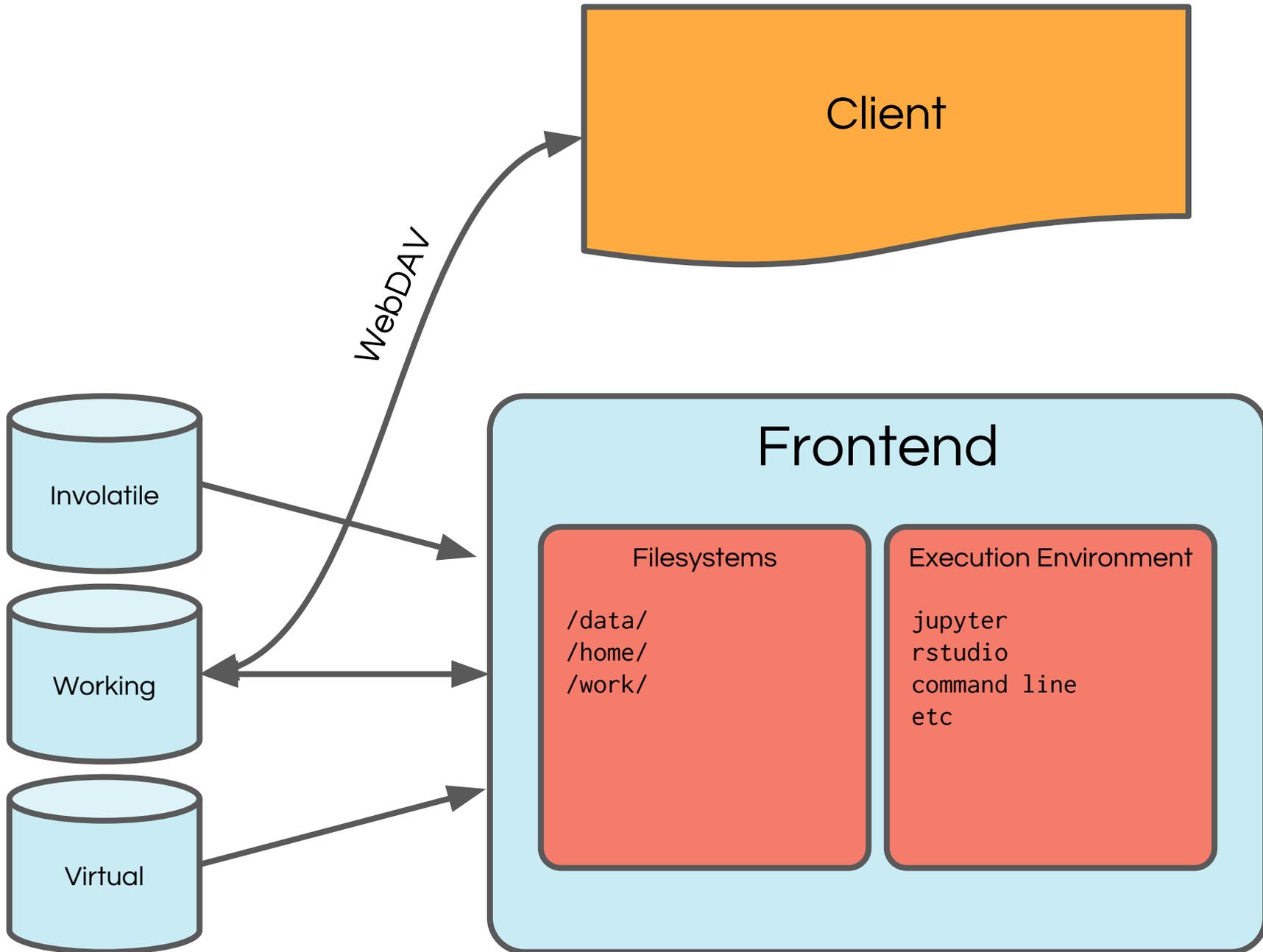
jupyter
rstudio
command line
etc

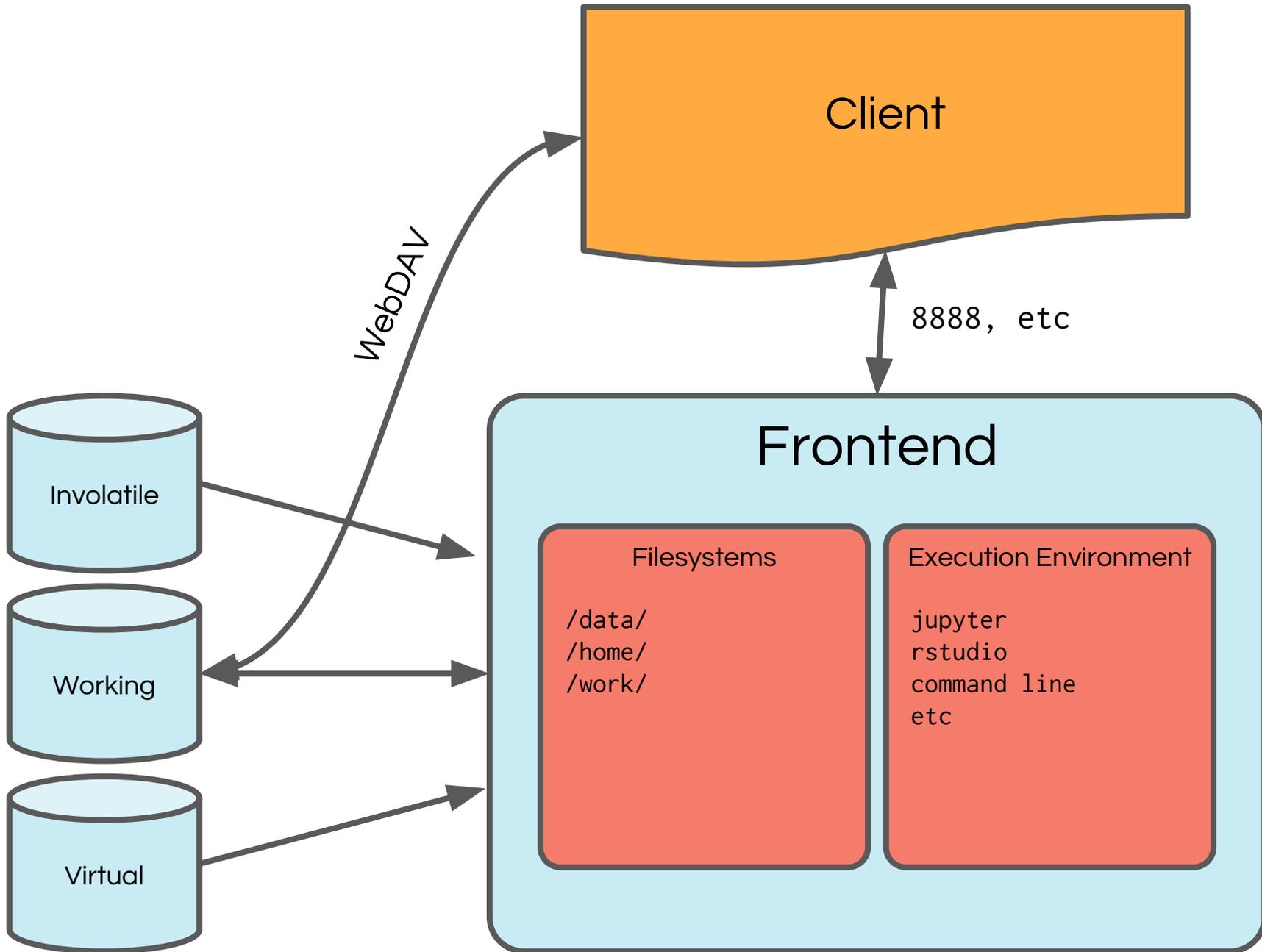


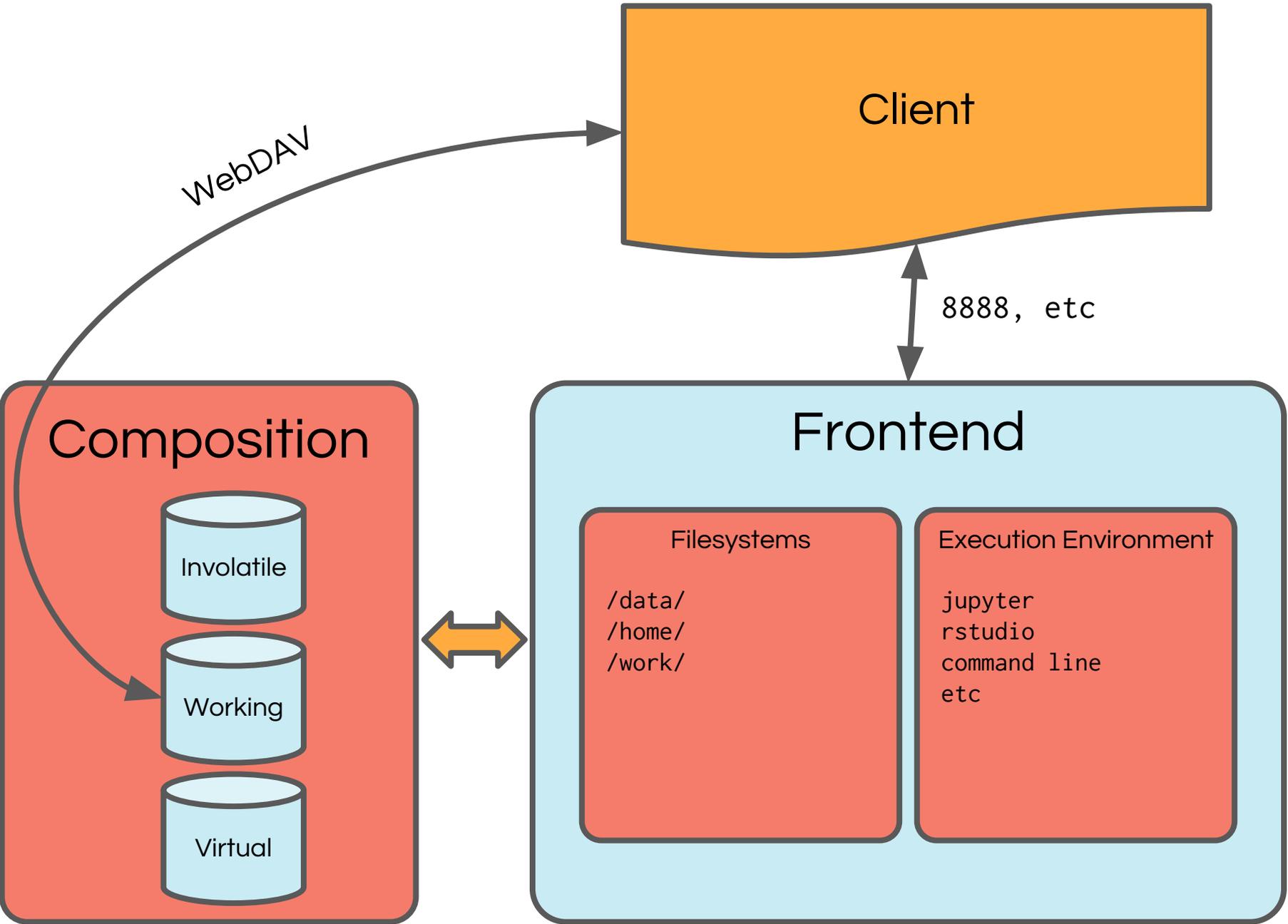
Classes of Data

1. Virtual data cache
2. Exported involatile storage
3. Bring my own data
4. Virtual data sources
5. User settings
6. Frontends/containers

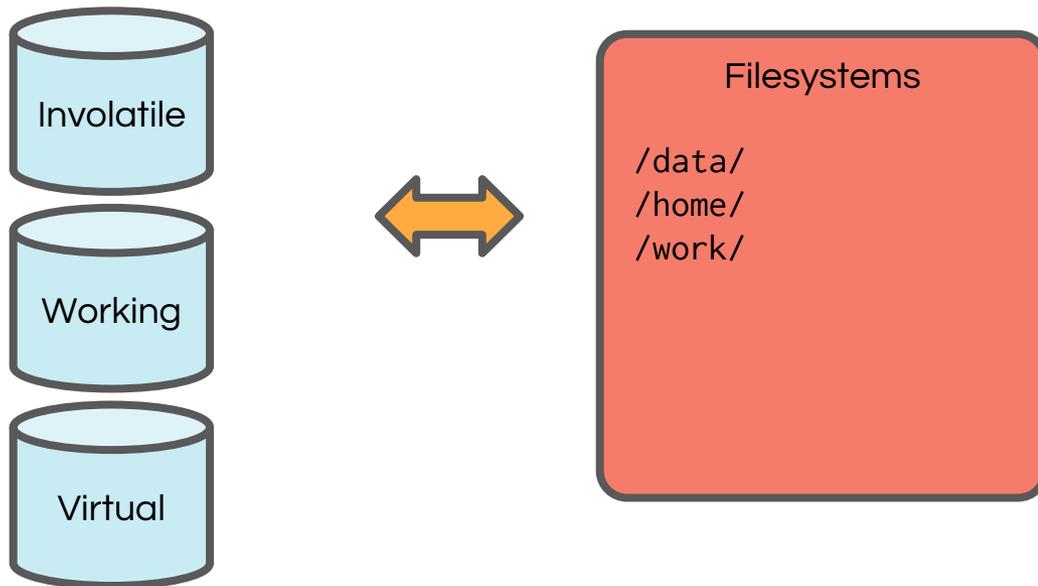


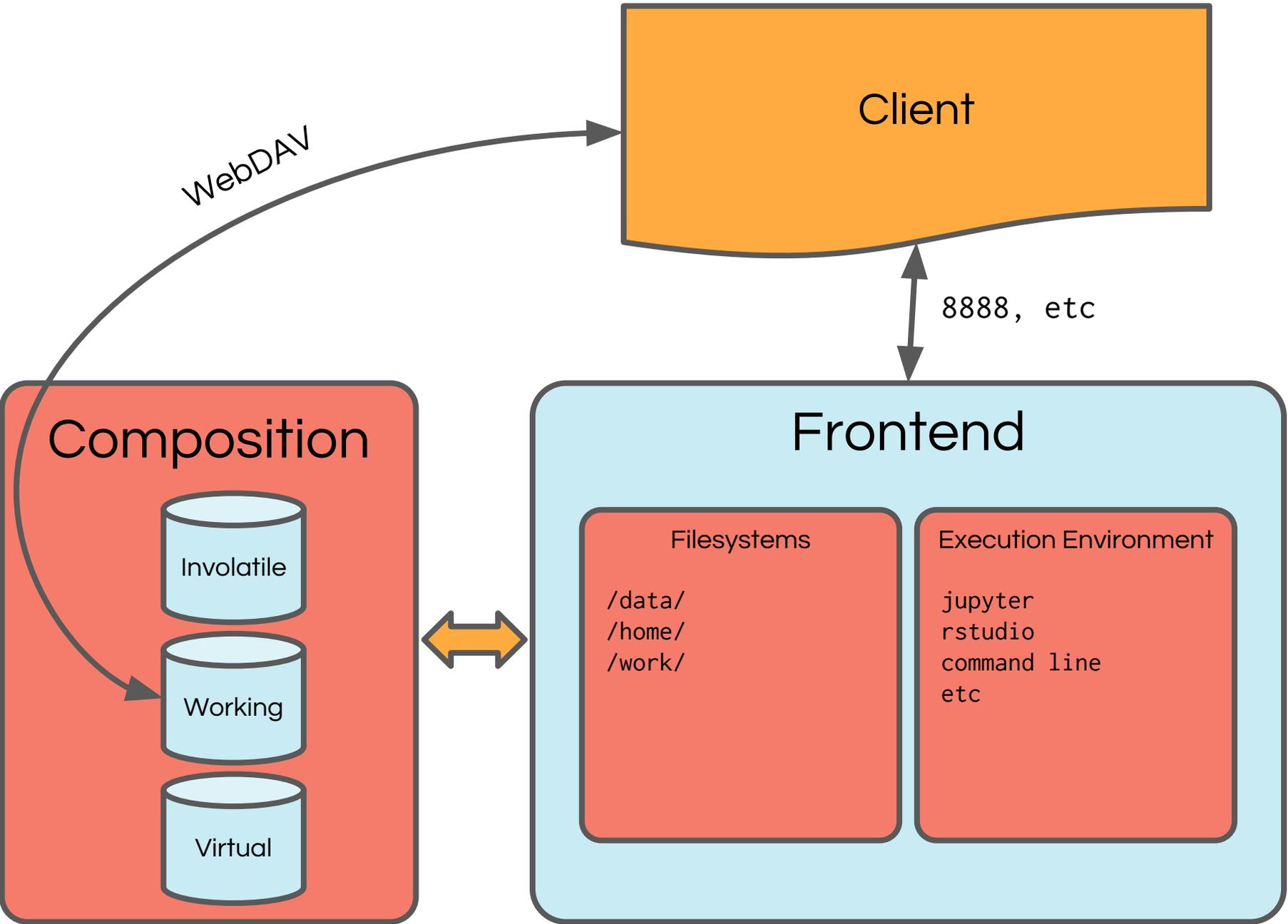


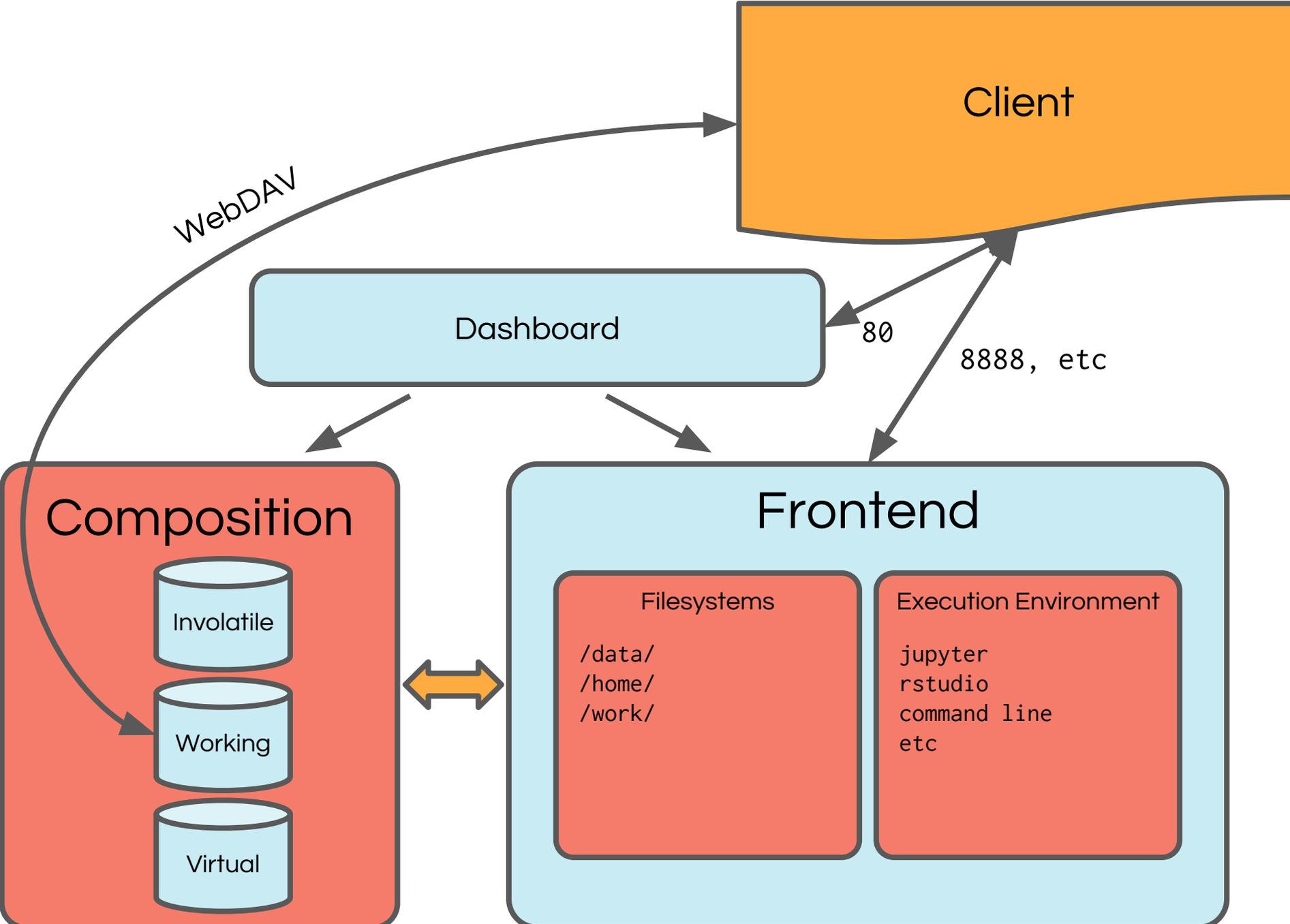




Filesystem-level Integration







Collections and Filesystems

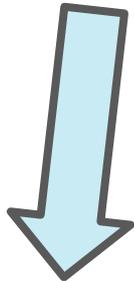
	Dataset 1
	Dataset 2
	Dataset 3

Collections and Filesystems

X	Dataset 1
	Dataset 2
X	Dataset 3

Collections and Filesystems

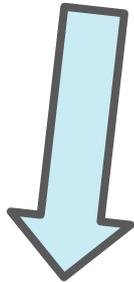
X	Dataset 1
	Dataset 2
X	Dataset 3



```
{ [ "dataset1", [ ... ], ...,  
  [ "dataset3", [...], ... ] }
```

Collections and Filesystems

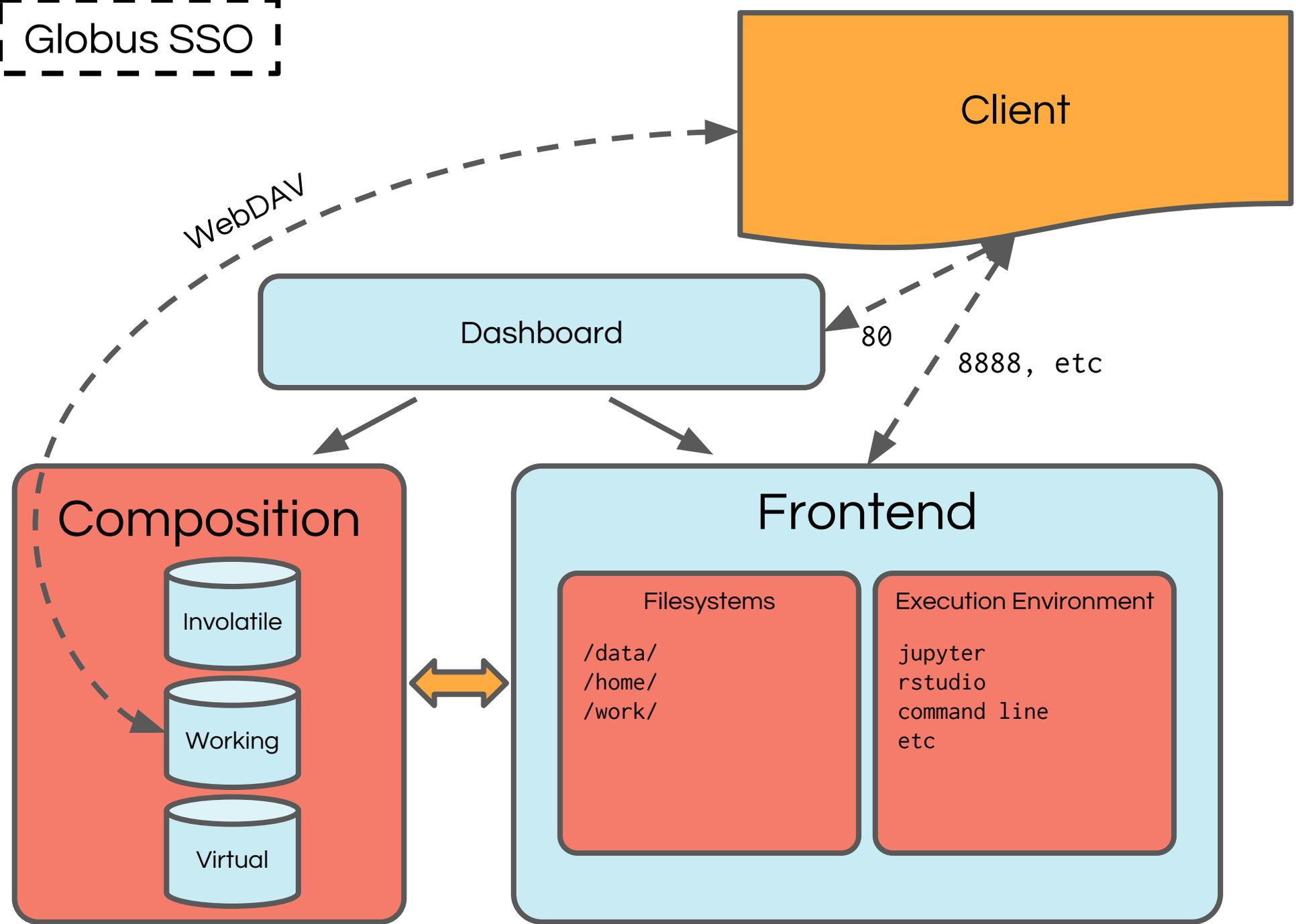
X	Dataset 1
	Dataset 2
X	Dataset 3



```
{ [ "dataset1", [ ... ], ...,  
  [ "dataset3", [...], ... ] }
```



```
/data/  
/data/collection.json  
/data/dataset1/...  
/data/dataset1.json  
/data/dataset3/...  
/data/dataset3.json
```



Globus SSO

Client

WebDAV

Dashboard

80

8888, etc

Composition

Involatile

Working

Virtual

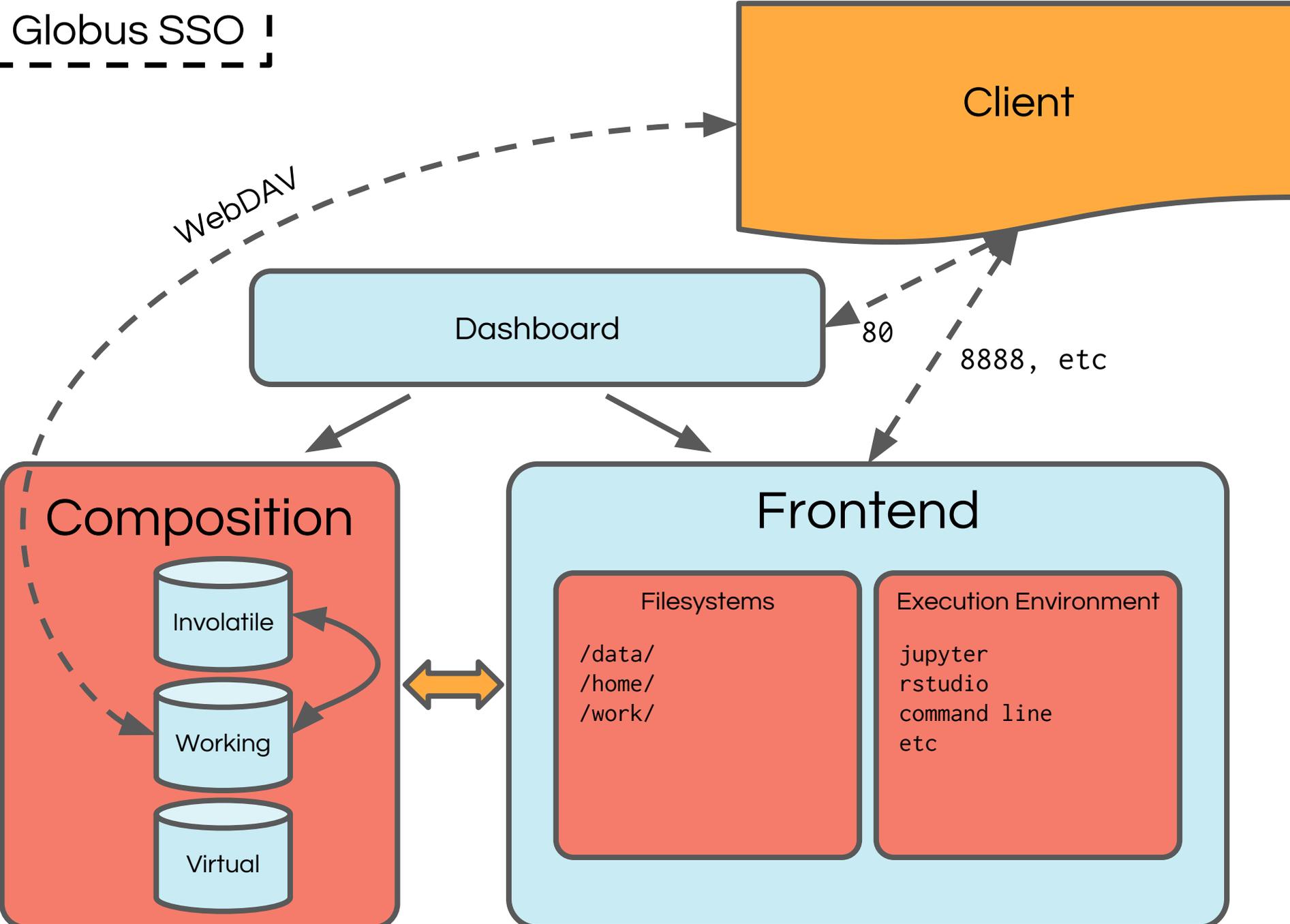
Frontend

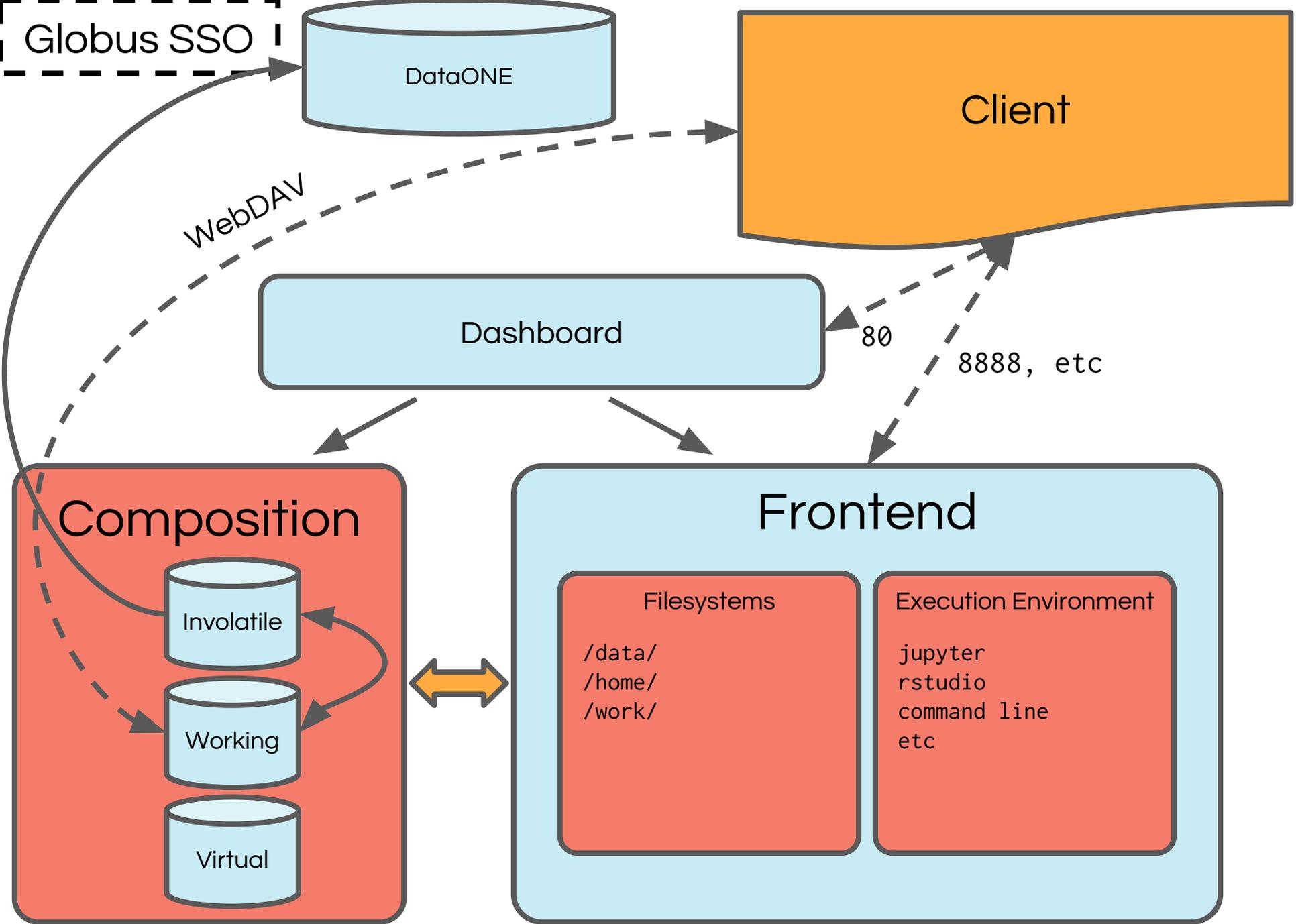
Filesystems

/data/
/home/
/work/

Execution Environment

jupyter
rstudio
command line
etc







Dashboard Front End

Semantic UI

JavaScript

HTML5

ember

BACKBONE.JS

User Data Workspace

owncloud

nextcloud

WHOLETALE API

Data Management

User Management

Container Management

Authorization Management

Metadata & Search

Girder

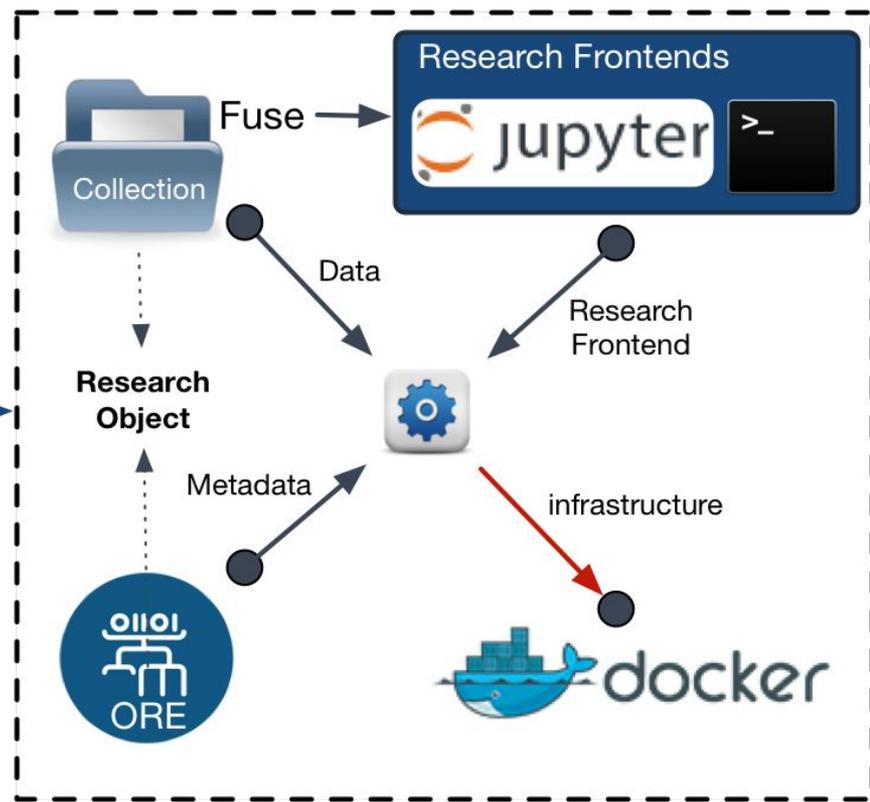
CherryPy

Data

DataOne	Globus
Fuse	HDFS
Amazon S3	iRODS

User & Authorization

Globus





Things I Didn't Mention



- DataONE integration
- iRODS federation
- DOI resolution
- Namespacing
- Globus integration
- Provenance tracking
- Engagement with publishers

Thank you.

mjturk@illinois.edu

wholetale.org

github.com/whole-tale/

Interested in joining a
working group?

kyturner@illinois.edu