Community Labeling and Sharing of Security and Networking Test datasets (CLASSNET)

Jelena Mirkovic
mirkovic@isi.edu

John Heidemann
johnh@isi.edu

Wes Hardaker
hardaker@isi.edu

Robert Stovall
bes@merit.edu

Motivation

Network and security researchers lack *diverse, contemporary, labeled data*. Such data is needed to harness the power of machine learning and new algorithms to meet security and networking challenges like intrusion detection, denial-of-service-defense, routing security, and classification and detection.

Network and security datasets may also be *imperfectly labeled* – by running a custom detection algorithm or a commercial tool. Researchers should be able to *discuss* these *labels* and to *offer alternative labels*.

CLASSNET Contributions

❖ **Dataset Portal.** : https://comunda.isi.edu provides a collaborative environment to *find* and *share labels* about datasets.

❖ **Diverse Labels.** We support *multiple labels* on datasets, allowing researchers to present alternative hypothesis, and compare to, rate, and refine prior results. We plan to support incentives for sharing.

❖ **Diverse pathways to dataset.** We provide a *range of data access* pathways that enable research, while protecting data privacy and balancing cost. *Data privacy is ad*dressed via anonymization, researcher vetting, and data use agreements. Data can be shared by download, cloud use, researcher-login for data, and code-to-date.

❖ **Continuous and curated datasets.** Our datasets currently include: regular traffic and DDoS attack flows from FrontRange GigaPOP, Darknet data, DNS and DDoS events from B-Root, and Internet topology data.

Data Provider Experience

E-mail and SSH
**Our ticket system via browser**
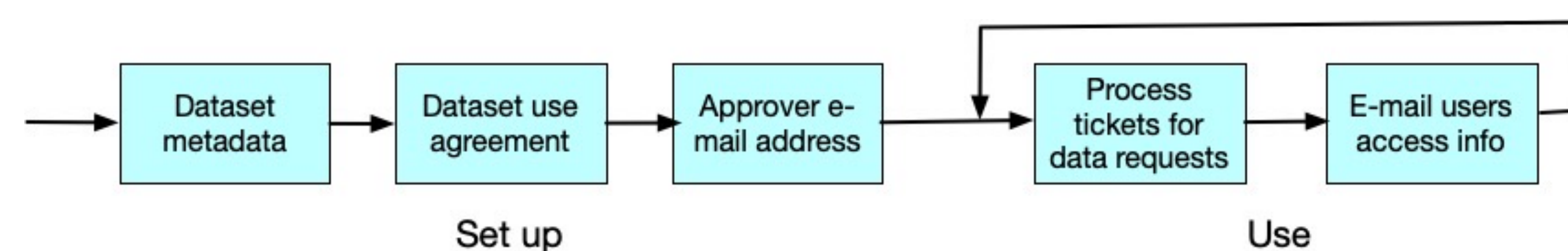
Benefits for Providers

❖ **Ready-to-use platform.** Providers can focus on data collection and offload tasks around data advertising and sharing to us.

❖ **Standardized metadata.** All our metadata is standardized, making it easy for providers to add new useful datasets.

❖ **Reach a wide audience.** Multiple and diverse datasets on our platform attract a wide research audience, benefiting all providers.

❖ **No loss of privacy.** Datasets remain at provider's facility – we only collect metadata, facilitate signing of data use agreements and connect researchers to providers.

❖ **Customized offloading of sharing activities.** We are prepared to handle user request approvals, data hosting and data distribution if the provider so desires.

Current Dataset Providers

USC Information Sciences Institute: B-root traffic (pcap and processed), regular Internet probing, Internet hitlists, outage data; **Merit Network Inc.:** Darknet data (pcaps and processed); **Front Range Giga POP:** Traffic flows (legitimate and attack). Coming soon: **Florida International University:** Traffic flows (legitimate and attack), **PIVOT project:** Vehicle security data.

Researcher Experience

Our portal at https://comunda.isi.edu
**Github for label corrections**
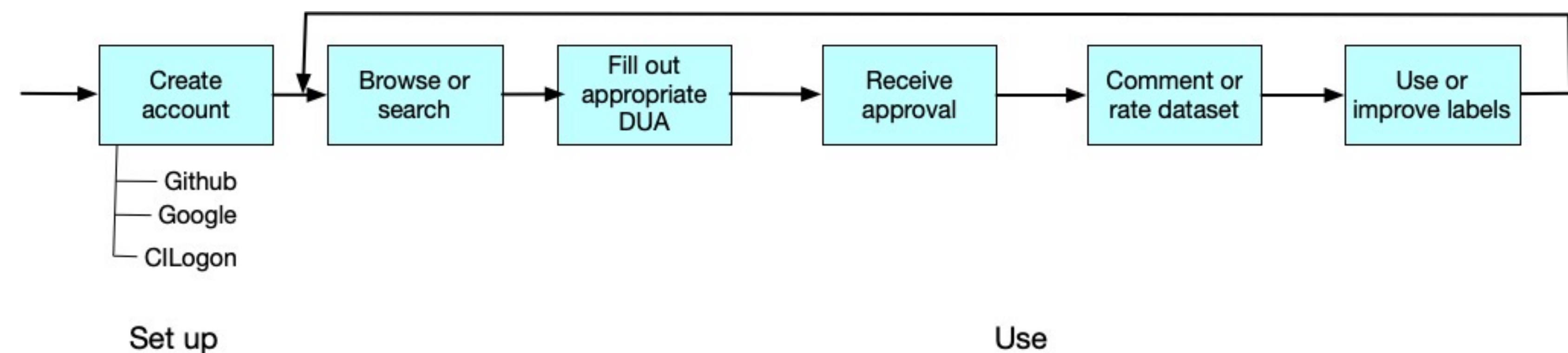
COMUNDA Portal
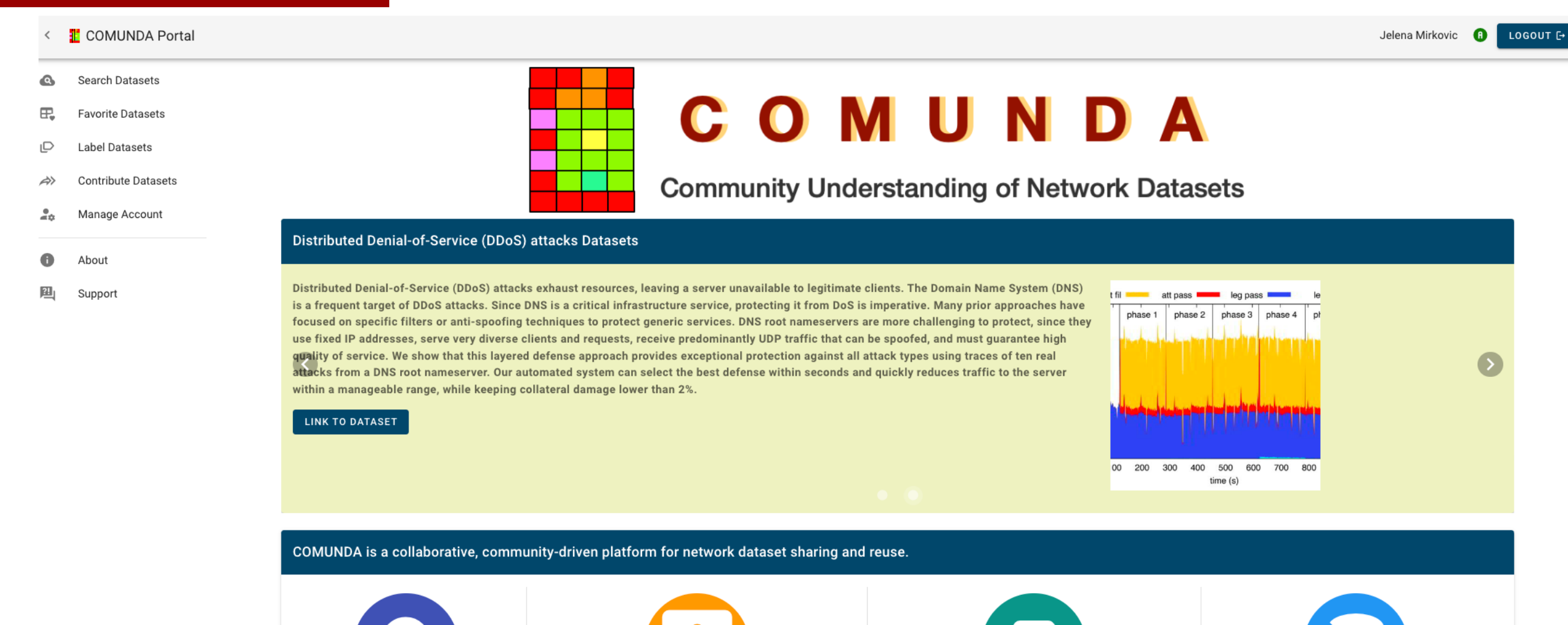
❖ **Dataset search.** Users can search for datasets using type and keywords, and see dataset metadata, such as size, start and end collection times, how data was collected, etc.

❖ **Access.** Users can log in using Github, Google or institutional login (inc. ORCID).

❖ **Dataset requests.** Users can request a dataset, which generates an email to data provider and a ticket in our system. Once the request is approved, provider contacts the user out of band to provide access instructions.

❖ **Rating and comments.** Users can rate and comment on datasets, which they were approved to access.

Dataset Labeling
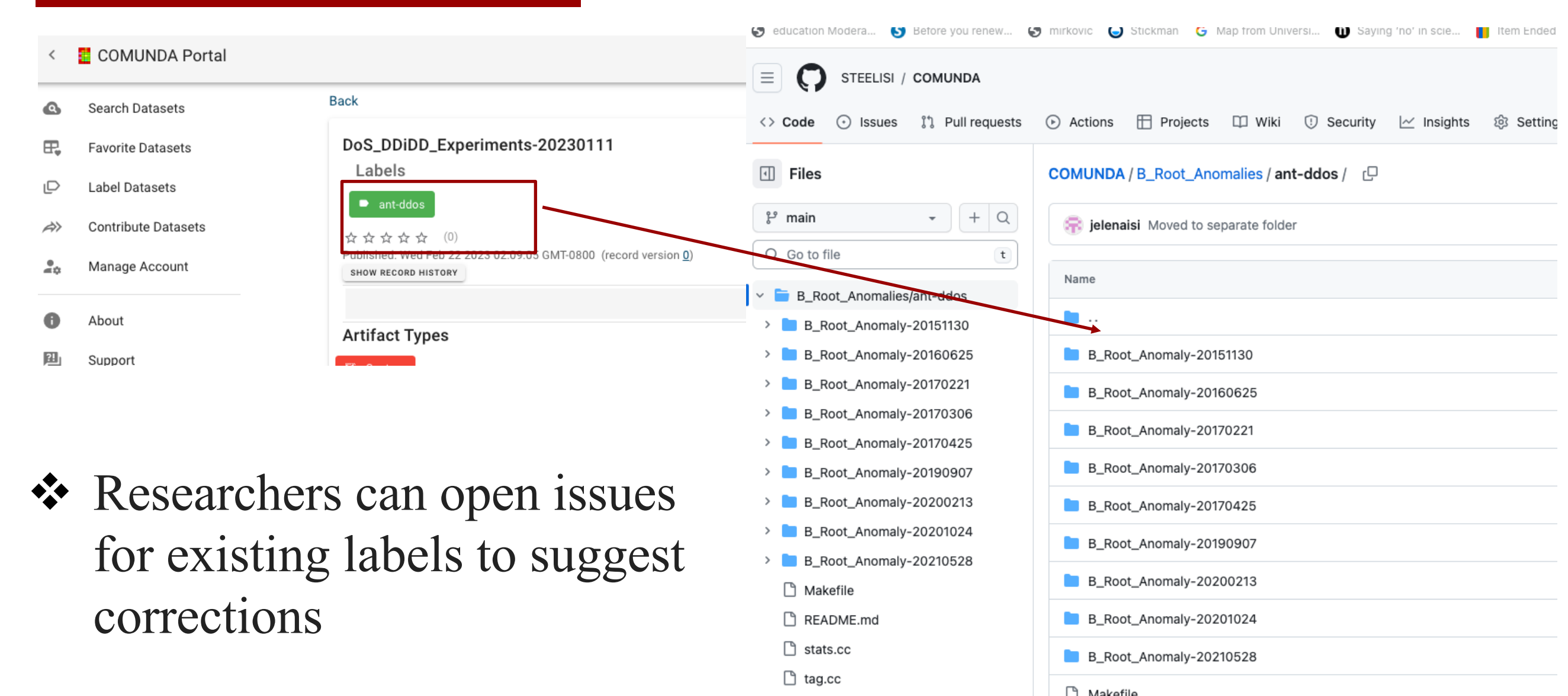
❖ Researchers can open issues for existing labels to suggest corrections

❖ Researchers can upload their own labels as new folder and new branch in COMUNDA and create a pull request

❖ Researchers can host new labels at their own location and provide a link to them

More Information

❖ Try our dataset portal at https://comunda.isi.edu

❖ You have data? We'd love to list it in our portal and send requests to you.

❖ Questions and comments? Reach us at classnet@isi.edu