Open Science with MATLAB

From Open Data to Reproducible Workflows

Shubo Chakrabarti

13th April 2023

shuboc@mathworks.com

Agenda

- Challenges facing scientific research today
- Open Science : an attempt to address challenges
- Open Science requirements : a combined effort
- How scientists are using MATLAB for Open Science
  - Using Open Data with MATLAB
  - Accessing Open Infrastructure with MATLAB
  - Building Interoperable Workflows with MATLAB
  - Publishing Open Results with MATLAB
- MathWorks support for the research community

Challenges facing scientific research today

Data is bigger than ever

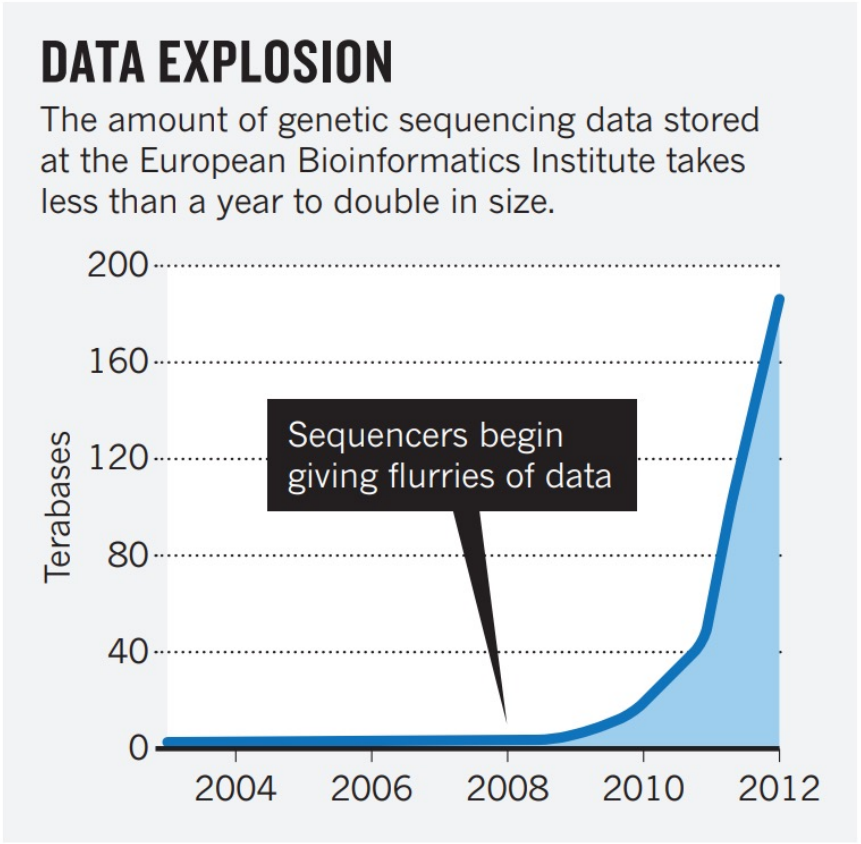
TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

EMBL–EBI

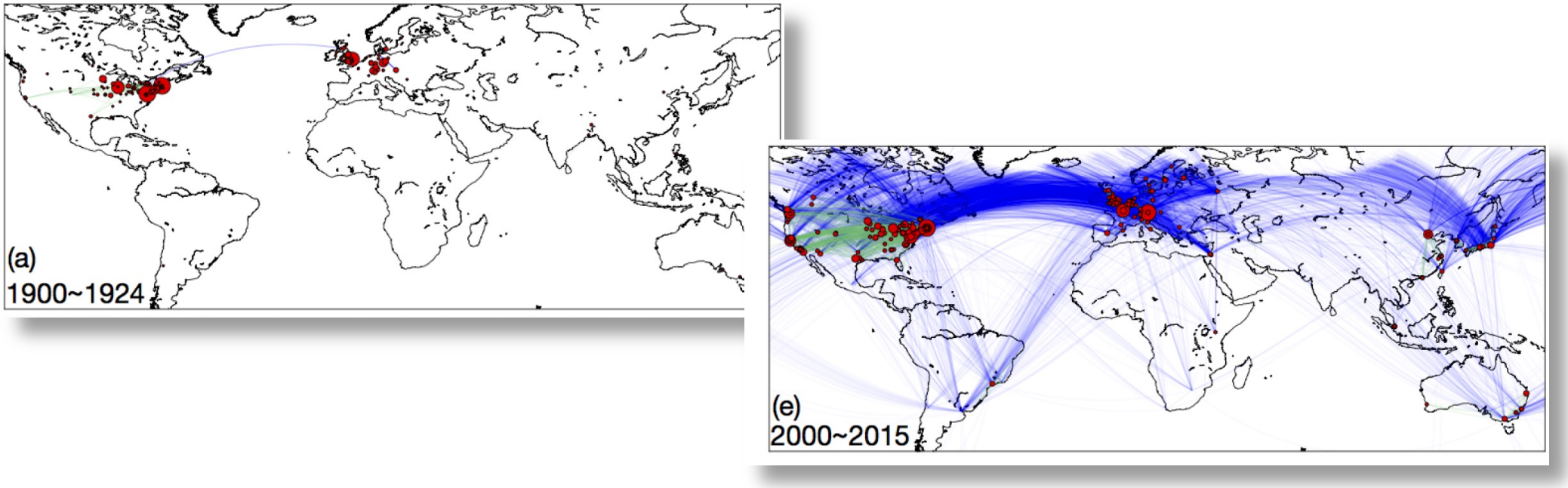
SOURCE: EMBL–EBI



Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013). https://doi.org/10.1038/498255a

Challenges facing scientific research today

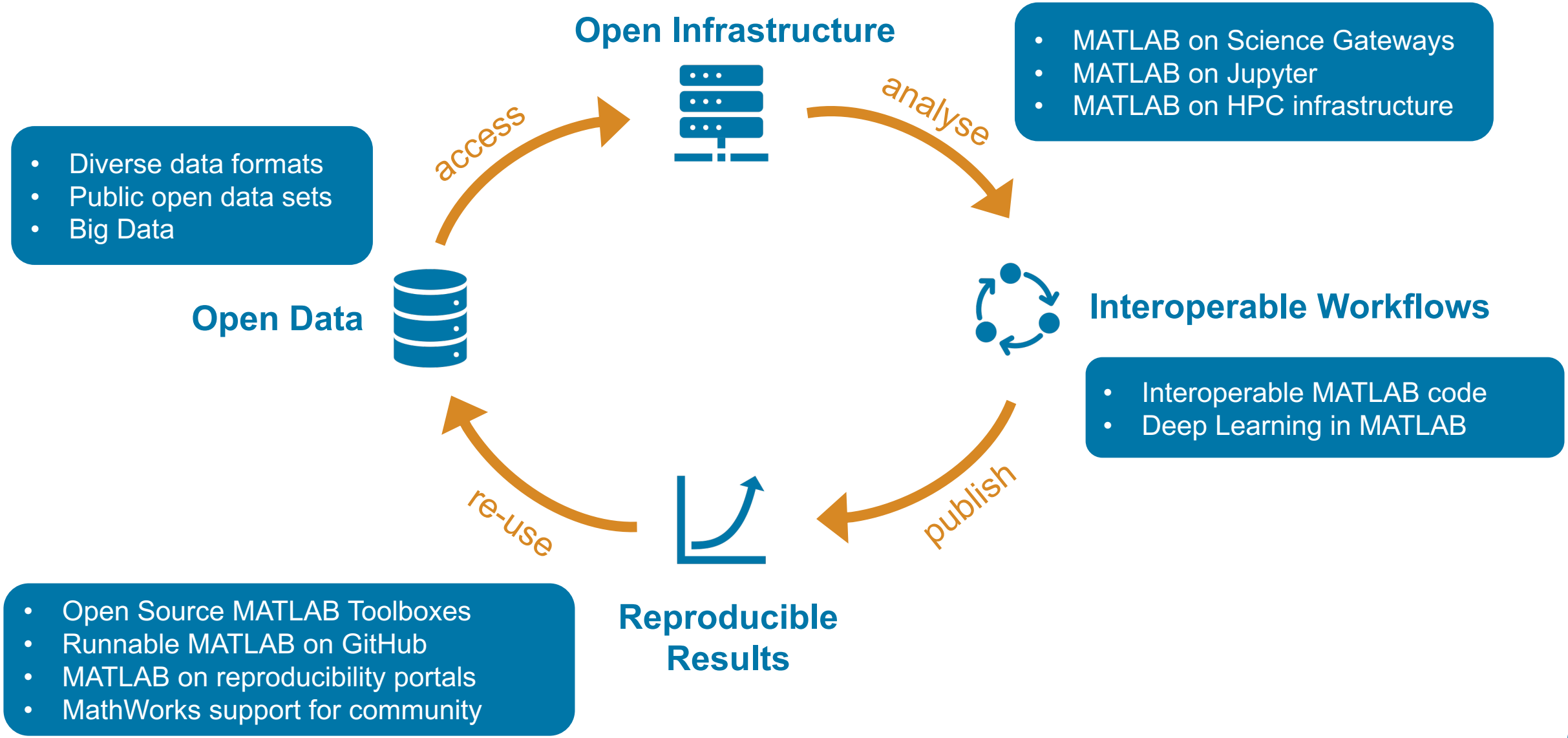
Research is a collaborative effort – more than ever today



A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations
Yuxiao Dong, Hao Ma, Zhihong Shen, Kuansan Wang, Microsoft Research. https://doi.org/10.1145/3097983.3098016

Open Science is an attempt to address these challenges

Open Science requires a combined effort of many stakeholders

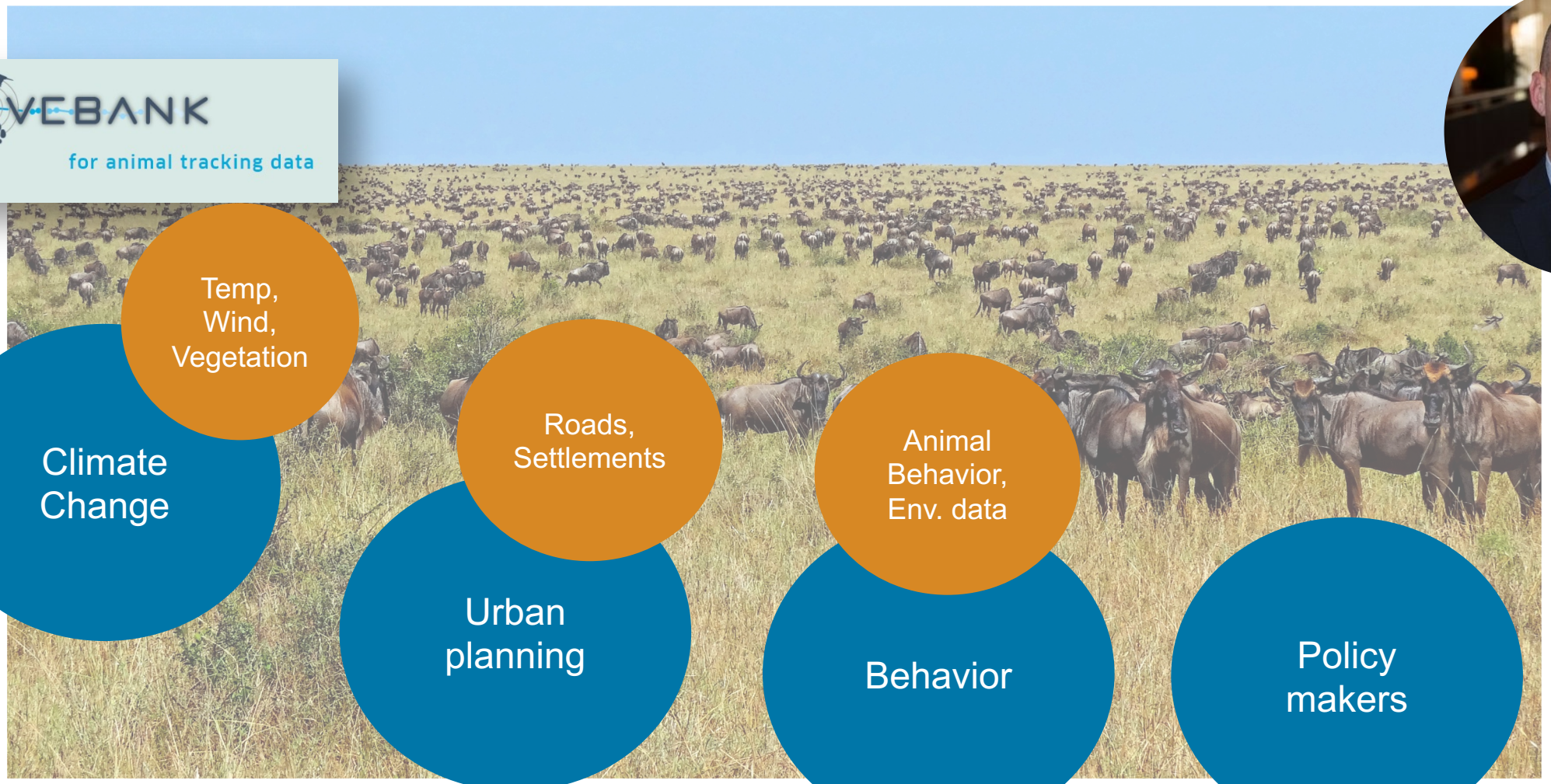


How scientists use MATLAB for Open Science

Using Open Data

Diverse uses of animal movement data for predictions and planning

MOVEBANK
for animal tracking data

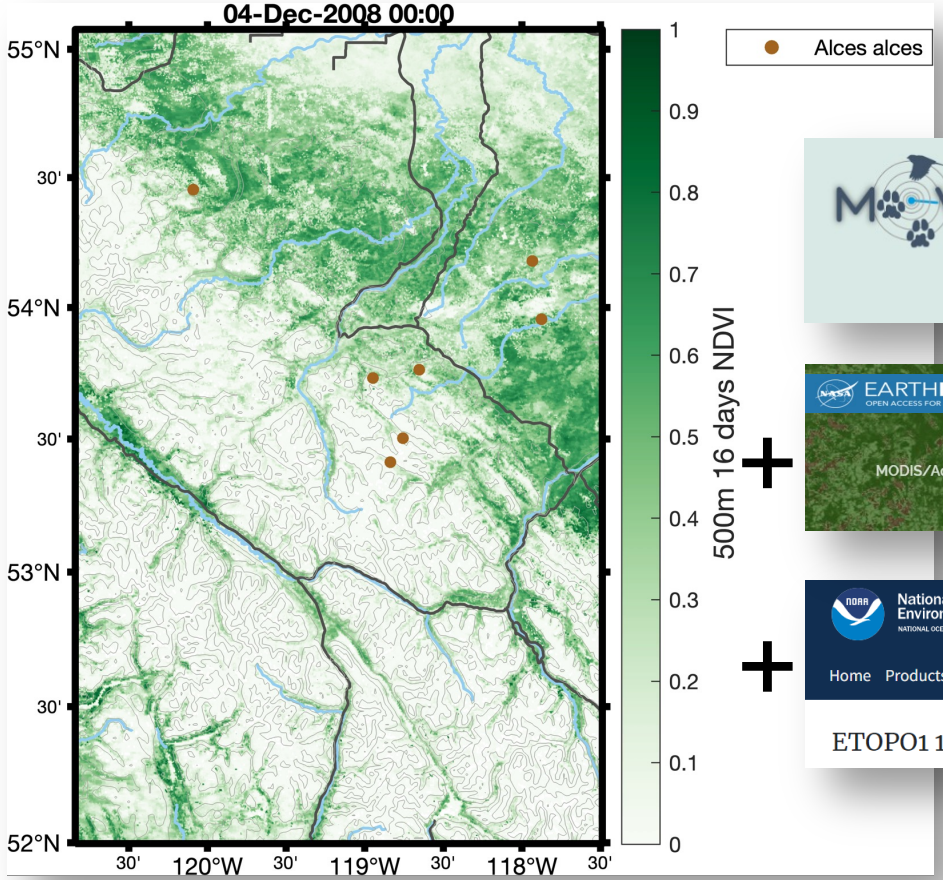


Prof.Gil Bohrer
Ohio State

Image: CC-BY-SA 3.0 License. Wikimedia

Movebank

ECODATA Animate: enabling users visualize animal movement data



App: download and install for all + open code on GitHub

MOVEBANK for animal tracking data

EARTHDATA Vegetation MYD13A1 v061
MODIS/Aqua Vegetation Indices 16-Day L3 Global 500 m SIN Grid
PI: Kamel Didan

National Centers for Environmental Information Topography
Home Products Services Resources News About Contact
ETOPO1 1 Arc-Minute Global Relief Model

Welcome to ECODATA-Animate's documentation!

ECODATA - Animate

ECODATA-Animate

ECODATA-Animate is a MATLAB® program for creating customized animated maps of animal movements. The program creates image frames that can be animated using the ECODATA-Prepare Movie Maker App. Define track visualization options and include additional layers from raster files, shapefiles, an elevation model and label lists. See ECODATA-Prepare for additional tools for preparing input data. Development is supported by MathWorks® and the NASA Earth Science Division, Ecological Forecasting Program, as part of the Room to Roam: Y2Y Wildlife Movements project.

Check out the documentation! App documentation Developer guide

Note: ECODATA-Animate is in the early stages of development, and any feedback is wel
have any suggestions or feature requests, enounter any bugs, or come across places
documentation is unclear, please submit a GitHub issue.

"I think MathWorks really got it right. Whatever we make can be compiled into an app with an executable. That app can be shared and does not require you to be a licensed user. For those who might want to work on the code further, we will publish the MATLAB code base." – **Gil Bohrer, PI, Ecodata/Prof. Ohio State University**

MathWorks

The Ohio State University

The Brain Observatory Toolbox: an interface to public cloud data

Allen Brain Observatory - Visual Coding AWS Public Data Set

electrophysiology | image processing | imaging | life sciences | Mus musculus | neurobiology | neuroimaging | signal processing

Description

The Allen Brain Observatory – Visual Coding is a large-scale, standardized survey of physiological activity across the mouse visual cortex, hippocampus, and thalamus. It includes datasets collected with both two-photon imaging and Neuropixels probes, two complementary techniques for measuring the activity of neurons in vivo. The two-photon imaging dataset features visually evoked calcium responses from GCaMP6-expressing neurons in a range of cortical layers, visual areas, and Cre lines. The Neuropixels dataset features spiking activity from distributed cortical and subcortical brain regions, collected under analogous conditions to the two-photon imaging experiments. We hope that experimentalists and modelers will use these comprehensive, open datasets as a testbed for theories of visual information processing.

Update Frequency

Annually

License

http://www.alleninstitute.org/

Documentation

Resources on AWS

Description
Project data files in a public bucket

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::allen-brain-observatory`

AWS Region
`us-west-2`

AWS CLI Access (No AWS account required)

Open in MATLAB Online | MATLAB® | File Exchange

Brain Observatory Toolbox

A MATLAB toolbox for accessing and using the neural recording public datasets from the **Allen Brain Observatory**[1].

Get oriented and get started with **3 lines of code**. You can:

- **Open in MATLAB Online**
- Enter `>>bot.README` on your own local/cloud installation

Either will orient you to several **live script examples** available to guide new users, including ***demos*** of neural data analysis

You can also individually view (👀) or run (▶) these examples on MATLAB Online:

Example Type	Data Type	View	Run	Data Type	View	Run
Quickstart	Calcium Imaging (Ophys)	👀	▶	Neuropixels Probe (Ephys)	👀	▶
Demo	Calcium Imaging (Ophys)	👀	▶	Neuropixels Probe (Ephys)	👀	(*)
Tutorial	Calcium Imaging (Ophys)	👀	▶	Neuropixels Probe (Ephys)	👀	(*)

- **Open Source**
- **Cloud workflow – no downloads**
- **Uses the AWS Open Data Registry**

Obtain Session of Interest

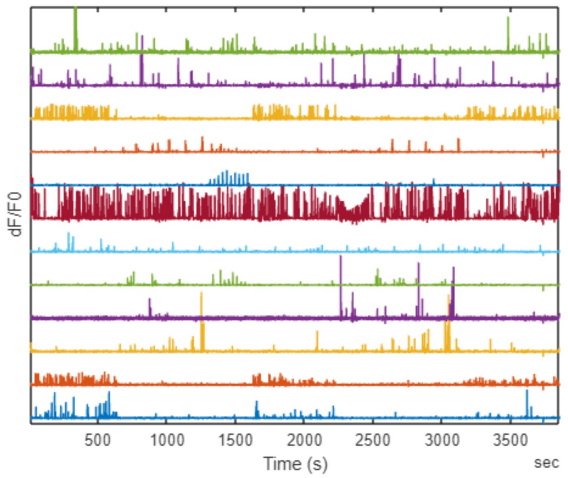
From the table of filtered items, one or more specific items can be obtained for further examination using the **`bot.get()`** functions:

```
sessionUnderStudy = bot.getSessions(filteredSessions(1,:));
```

```
Downloading file: [https://allen-brain-observatory.s3.us-west-2.amazonaws.com/visual-coding-2p/ophys_experiment_data/50341
to cache location: /external/neuralcoding/prod6/specimen_495727000/ophys_experiment_503412730/503412730.nwb...
```

```
sessionUnderStudy.info
```

```
ans = struct with fields:
                            id: 503412730
        experiment_container_id: 511510695
                  stimulus_name: three_session_A
    targeted_structure_acronym: VISal
             fail_eye_tracking: 0
                 imaging_depth: 175
                      cre_line: "Cux2-CreERT2"
           date_of_acquisition: 23-Feb-2016 19
                          name: "20160223_2224
                   specimen_id: 495727000
          experiment_container: [1×1 struct]
                      specimen: [1×1 struct]
           targeted_structure: [1×1 struct]
             well_known_files: [3×1 struct]
```

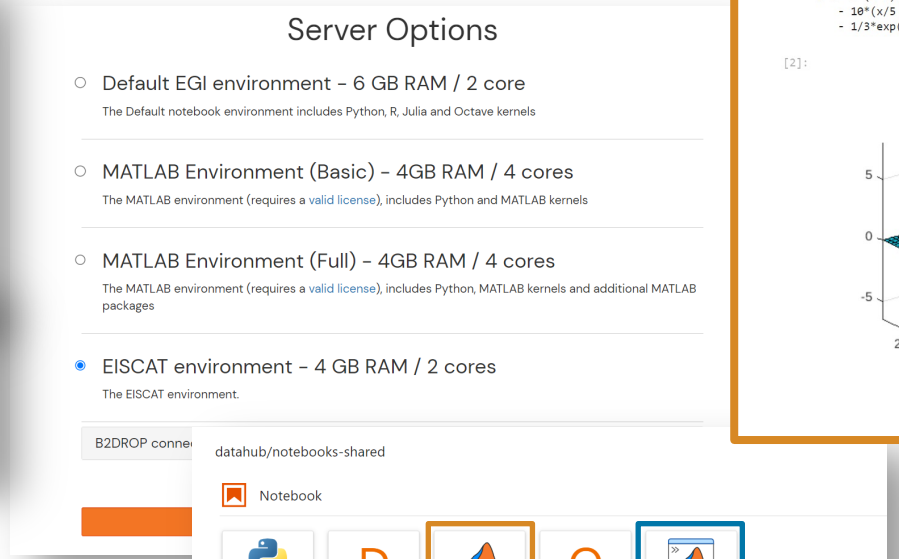


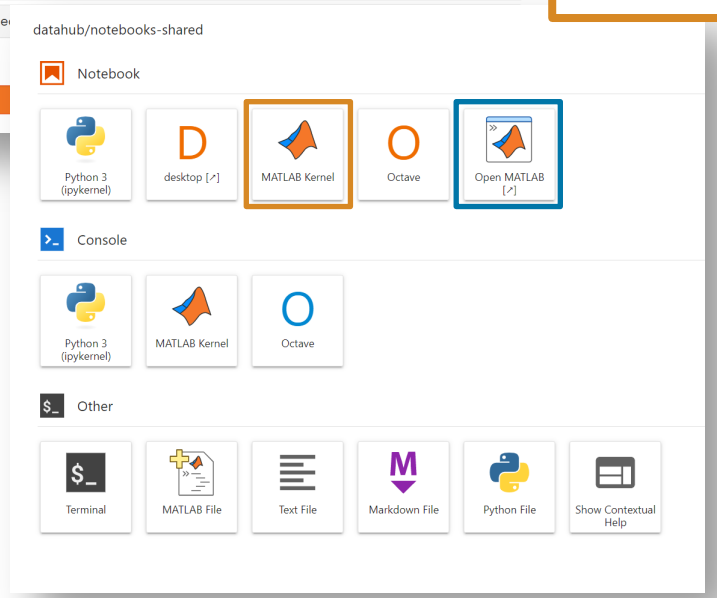
[AWS open data registry]
[Brain Observatory Toolbox]

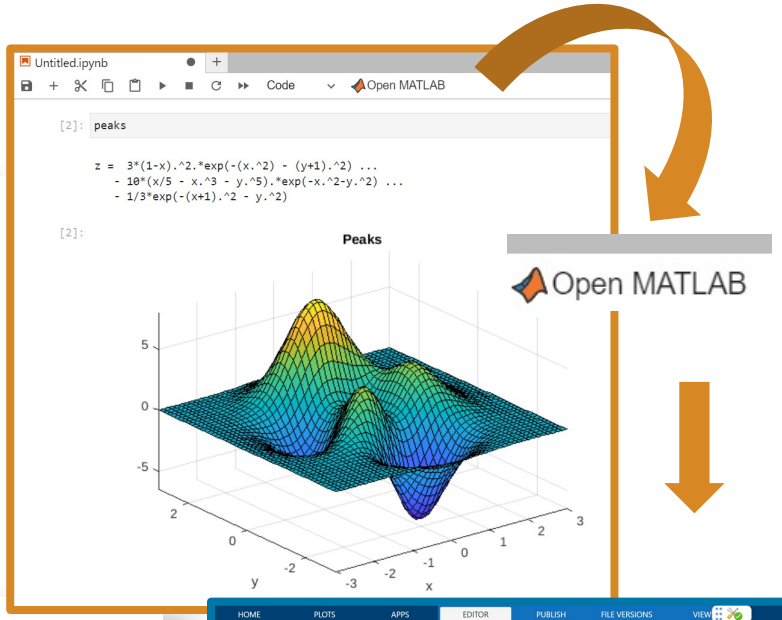
How scientists use MATLAB for Open Science

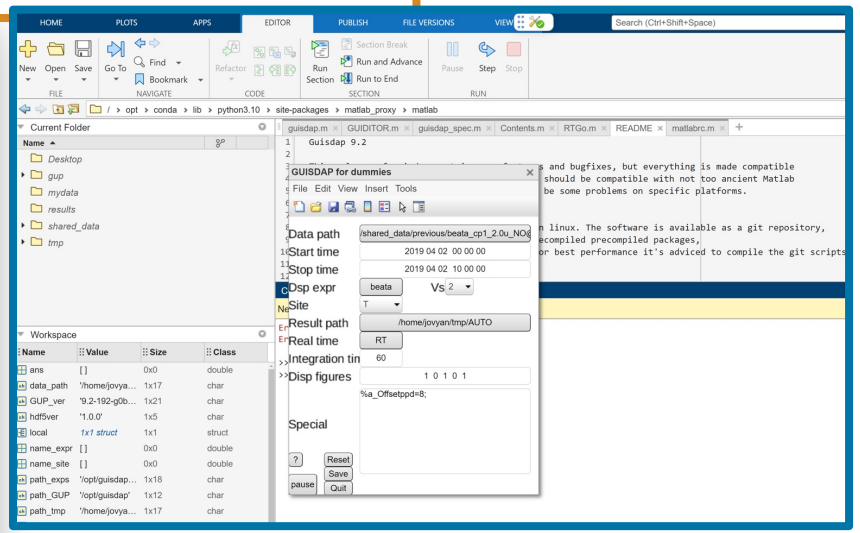
Accessing Open Infrastructure

Accessing Open Infrastructure –
MATLAB kernel on EISCAT3D JupyterHub







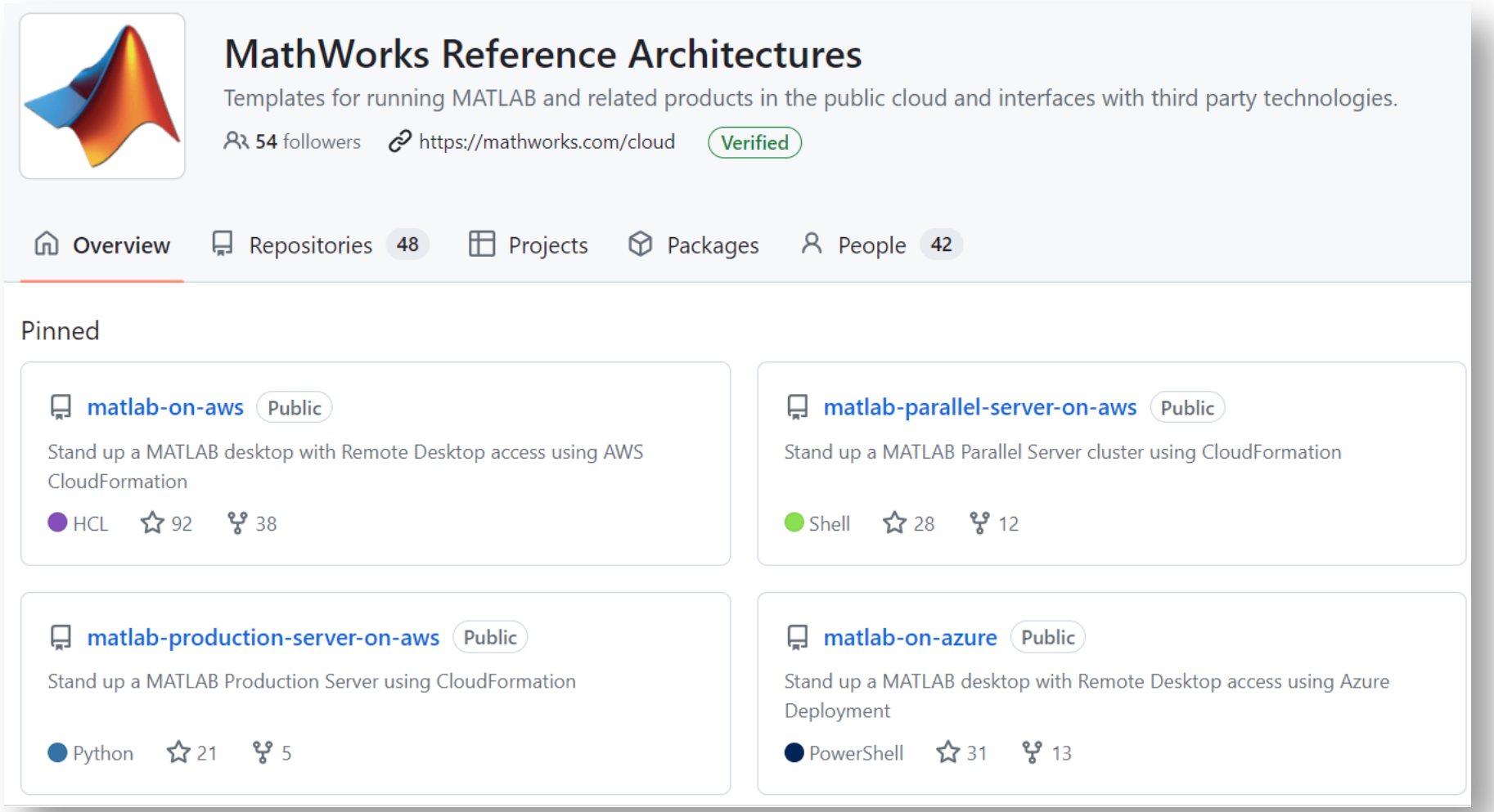


The EISCAT integration of Matlab into Jupyter is the perfect tool for new (and old) users and is a window to Open Science and FAIR principles.

- **Ingemar Häggström, Head of Operations**

Accessing Open Infrastructure: MathWorks reference architectures

Build your own integration



MathWorks reference architectures

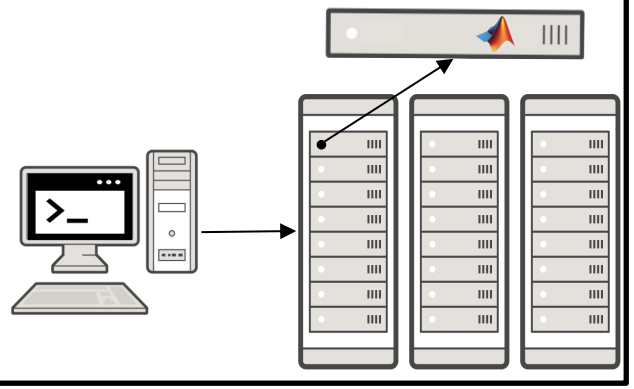
Accessing Open Infrastructure

Using HPC does not have to disrupt your research

More Similar ***Likeness to traditional HPC workflows****

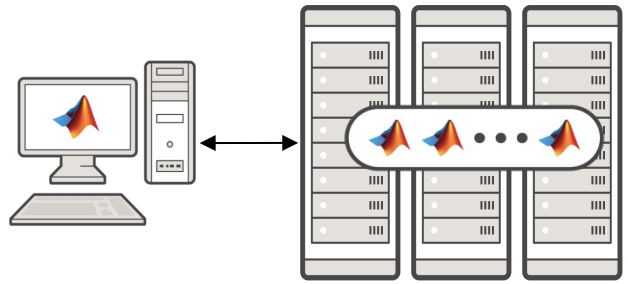
Run MATLAB on the cluster via batch or interactively using the scheduler

Required Knowledge:
- MATLAB
- Shell
- Scheduler
- File transfer



Submit work from MATLAB client to cluster running MATLAB Parallel Server

Required User Knowledge:
- MATLAB



Ease of use for MATLAB users *Less Difficult*

MATLAB Parallel Server

* Cluster scheduler manages workload for all options on this spectrum

University of Cambridge Accelerates Neuroimaging Data Analysis to Study Healthy Cognitive Aging

UNIVERSITY OF CAMBRIDGE

Challenge

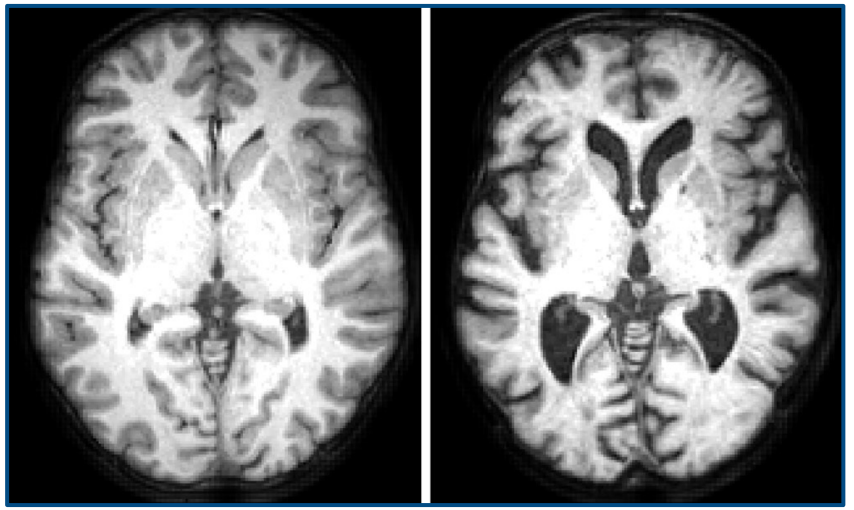
Analyze terabytes of MRI and MEG imaging data for a study on cognitive abilities in old age

Solution

Use MATLAB to **process the data on a high-performance computing cluster** and to apply advanced statistical, optimization, and machine learning techniques

Results

- Multistep image processing pipeline automated
- Data analysis accelerated on a 1200-core cluster
- Key influences on late-life cognitive health identified



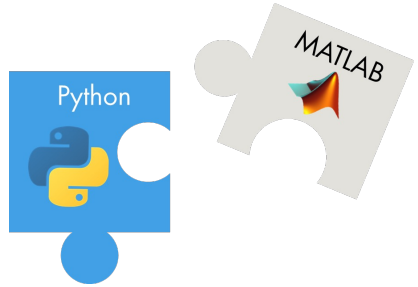
Structural MRI images of the brain of a typical healthy 19-year old (left) and typical healthy 86-year old (right).

"MATLAB Parallel Server not only reduces processing time; it also lowers the barrier to entry into parallel computing—an important consideration given the wide range of technical abilities among our scientists."

- Richard Henson, University of Cambridge

Link to technical article

Create Interoperable Workflows



Why researchers often need more than one tool

- Use functionality of one in another
- Run code available in another language (hybrid workflows)
- Collaborate

Call MATLAB from Python

```
$ pip install matlabengine
>>> import matlab.engine
>>> m = matlab.engine.start_matlab()
>>> x = m.sqrt(42)
```

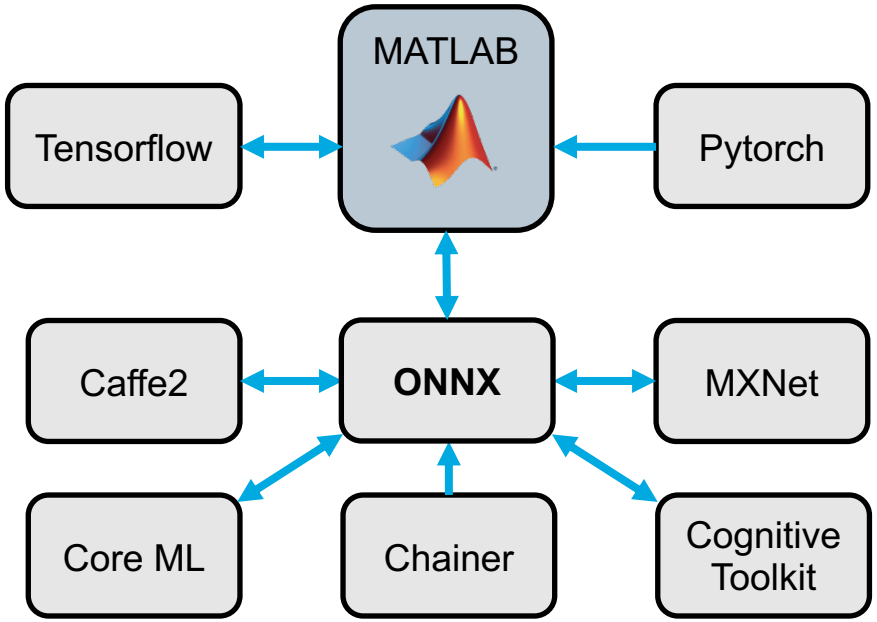
Call Python from MATLAB

```
>>> py.math.sqrt(42)
>>> py.importlib.import_module()
```

Using MATLAB with Python

Create Interoperable Workflows

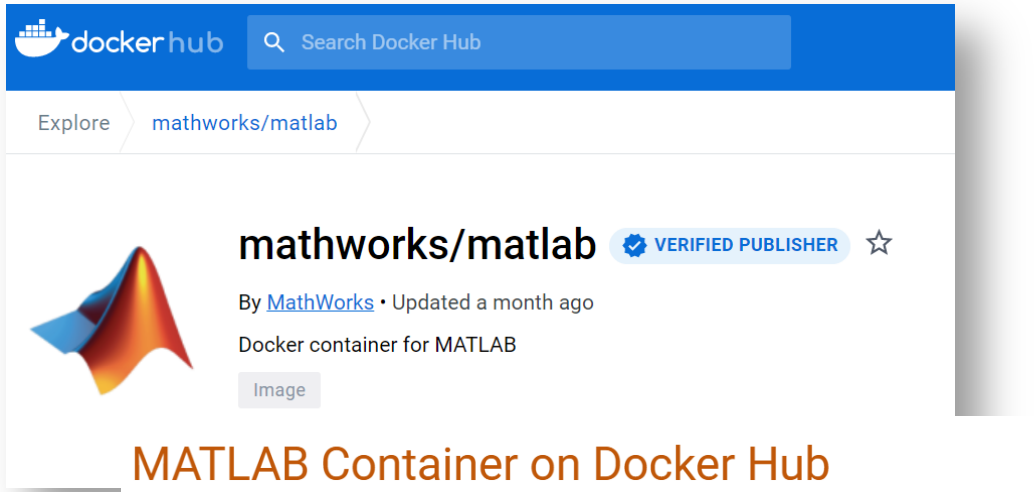
Import DL models from other frameworks




```
importTensorFlowNetwork
importNetworkFromPytorch
importONNXNetwork
```

[Interoperable with DL frameworks]

Pre-built containers on Docker-Hub



MATLAB Container on Docker Hub

Access MATLAB® on the cloud or in server environments by using the MATLAB container available on Docker Hub. The MATLAB container provides a simple and flexible solution to run MATLAB in cloud environments such as AWS® or Microsoft® Azure®. For more information on containers, see What is a Container?.

[MATLAB containers on DockerHub]

Publishing Reproducible Results: runnable code on GitHub

README.md

Open in MATLAB Online | MATLAB® File Exchange

Brain Observatory Toolbox

A MATLAB toolbox for accessing and using the neural

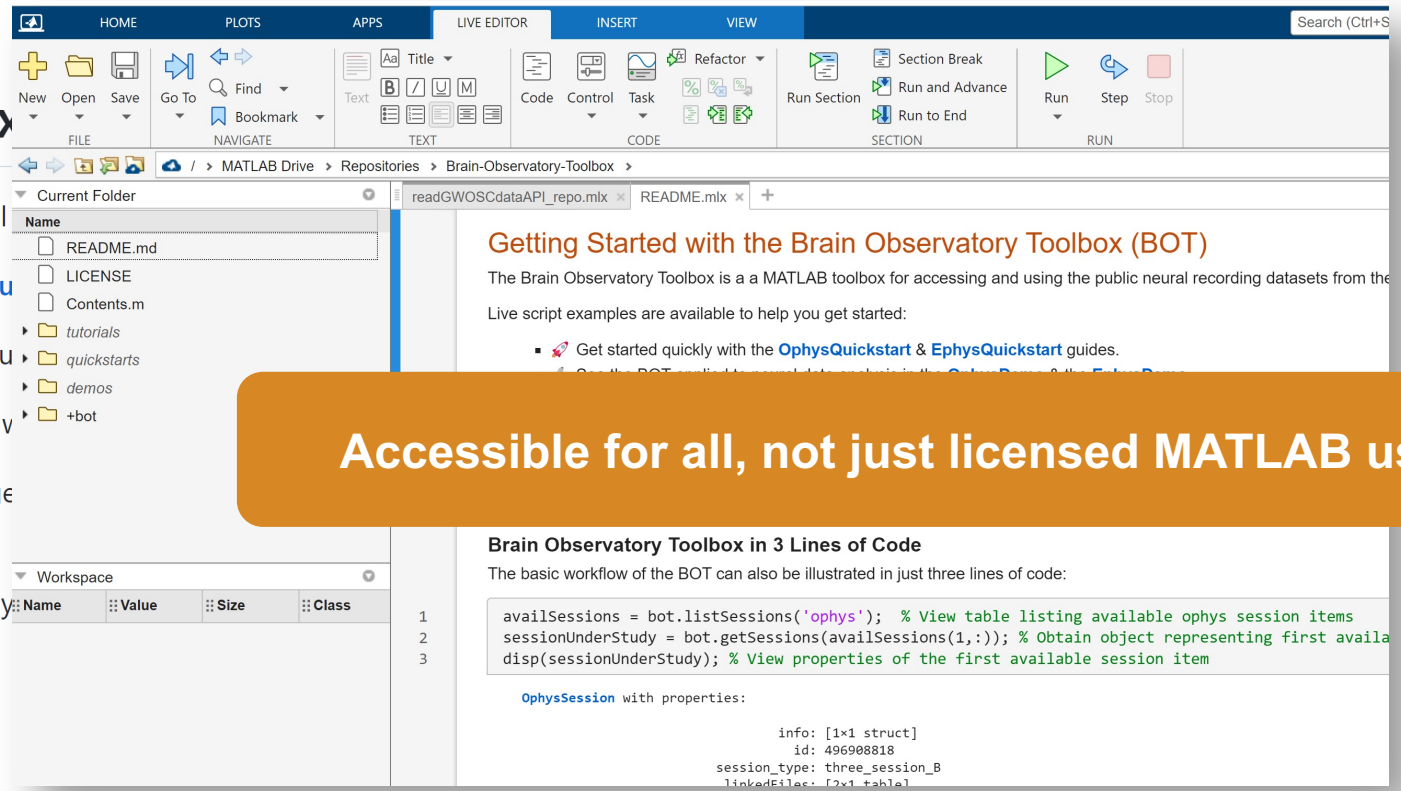
🚀 Get started with the EphysQuickstart & OphysQu

🔬 See the Brain Observatory Toolbox applied to neu

🏫 Learn how to use the Brain Observatory Toolbox v

🚧 The Brain Observatory Toolbox is at an early stage releases.

Questions, suggestions, and other feedback are highly



Accessible for all, not just licensed MATLAB users!

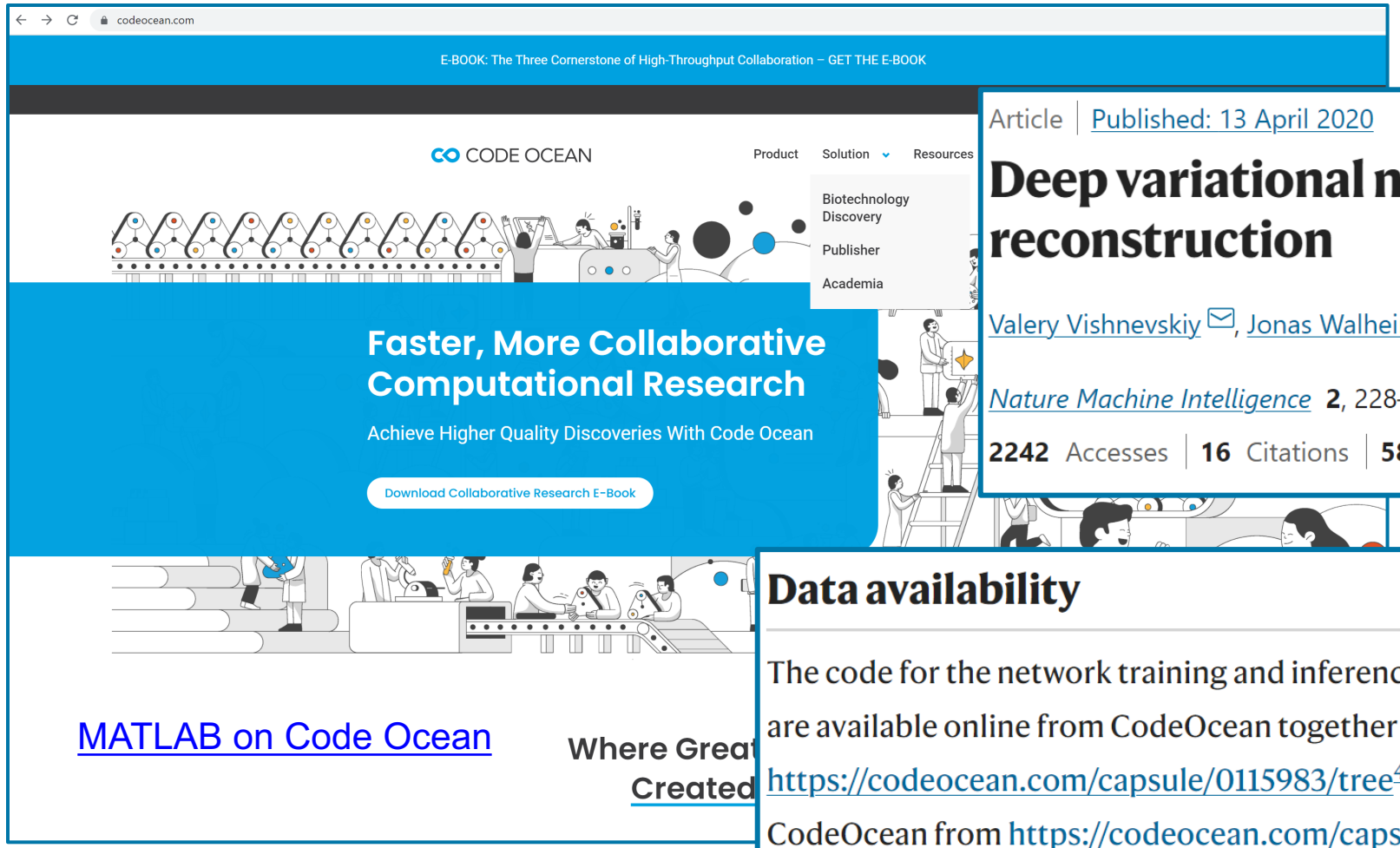
Open in MATLAB Online

```
[![Open in MATLAB Online](https://www.mathworks.com/images/responsive/global/open-in-matlab-online.svg)](https://matlab.mathworks.com/open/github/v1?repo=<authorname/reponame>&file=<path/to/filename.mlx>&line=88)
```

Make your MATLAB code reusable and reproducible

Use a reproducibility portal that supports MATLAB

CODE OCEAN



Article | Published: 13 April 2020

Deep variational network for rapid 4D flow MRI reconstruction

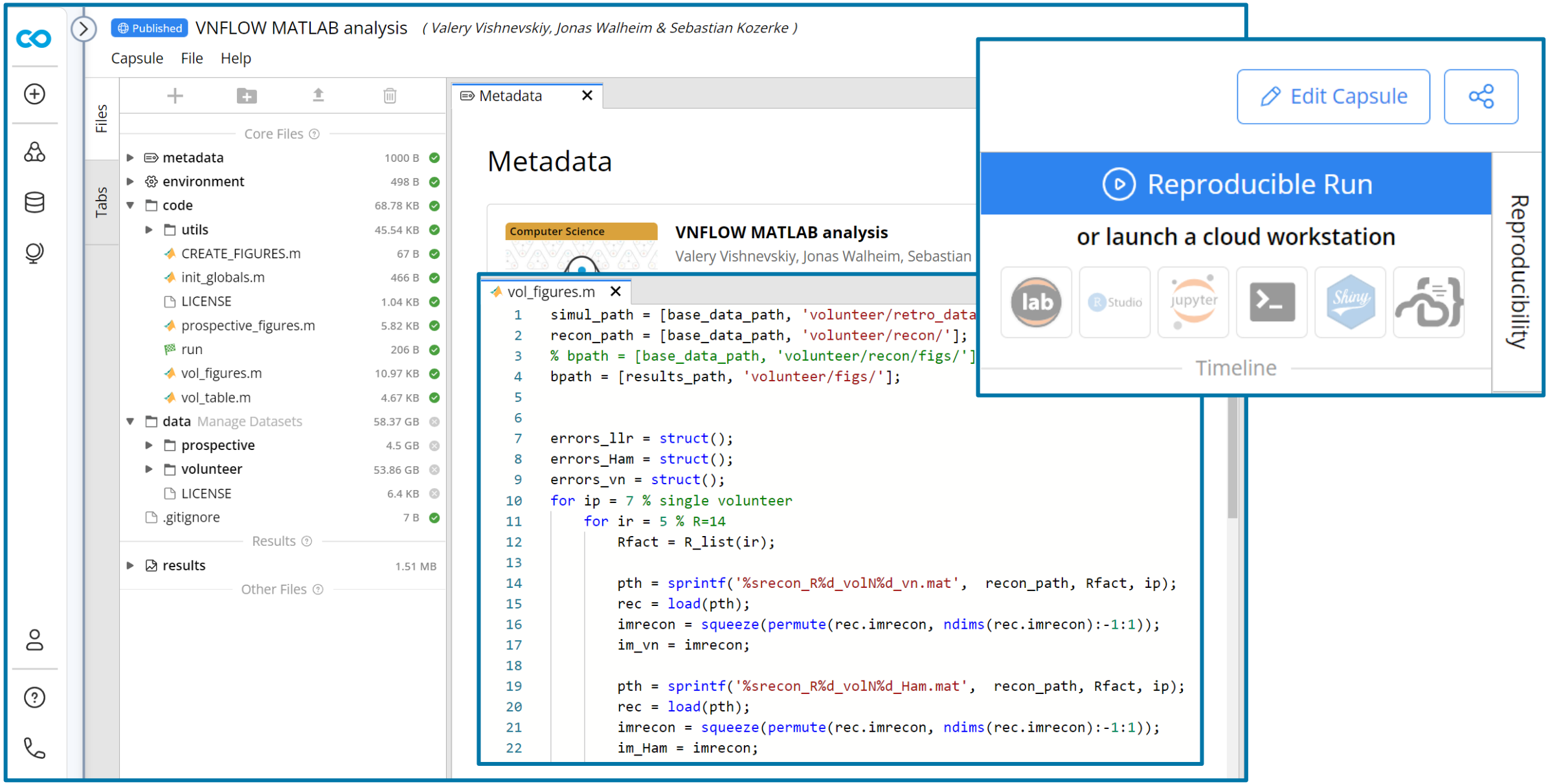
Valery Vishnevskiy, Jonas Walheim & Sebastian Kozerke

Nature Machine Intelligence **2**, 228–235 (2020) | Cite this article

2242 Accesses | **16** Citations | **58** Altmetric | Metrics

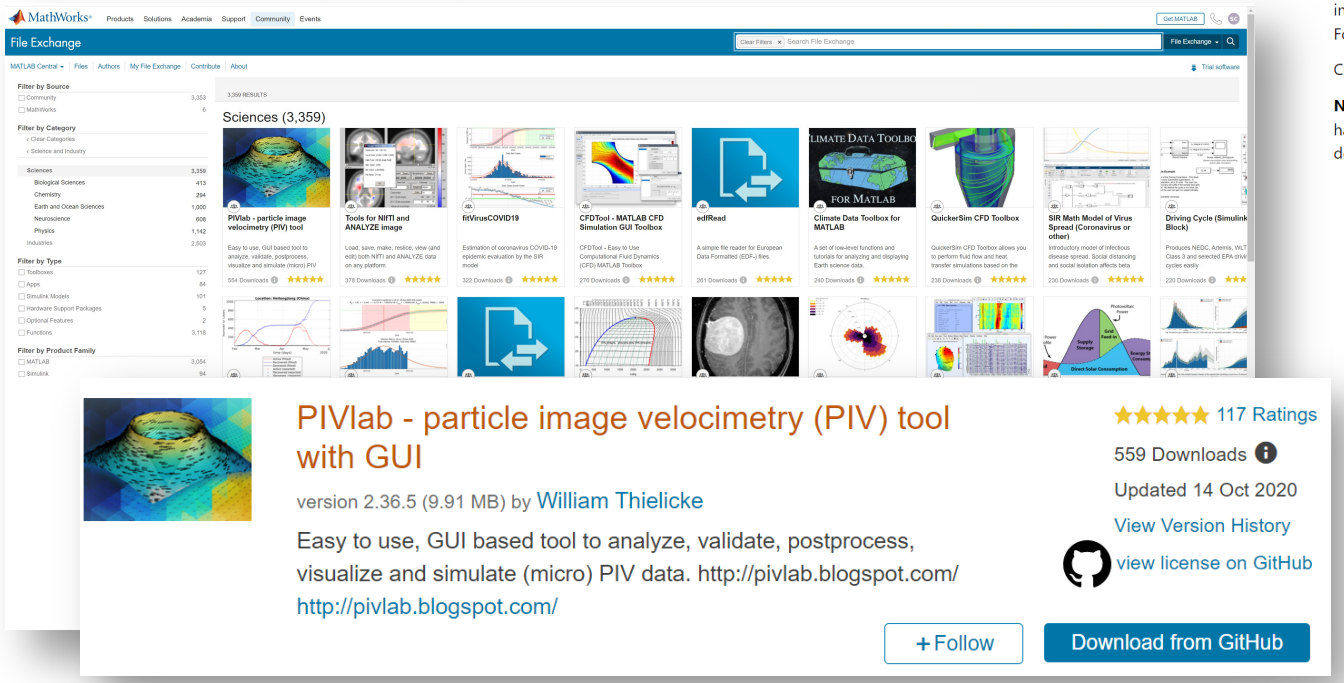
Data availability

The code for the network training and inference used in this study as well as network weights are available online from CodeOcean together with volunteer data: https://codeocean.com/capsule/0115983/tree[48]. The code for analysis is available on CodeOcean from https://codeocean.com/capsule/2587940/tree[49].

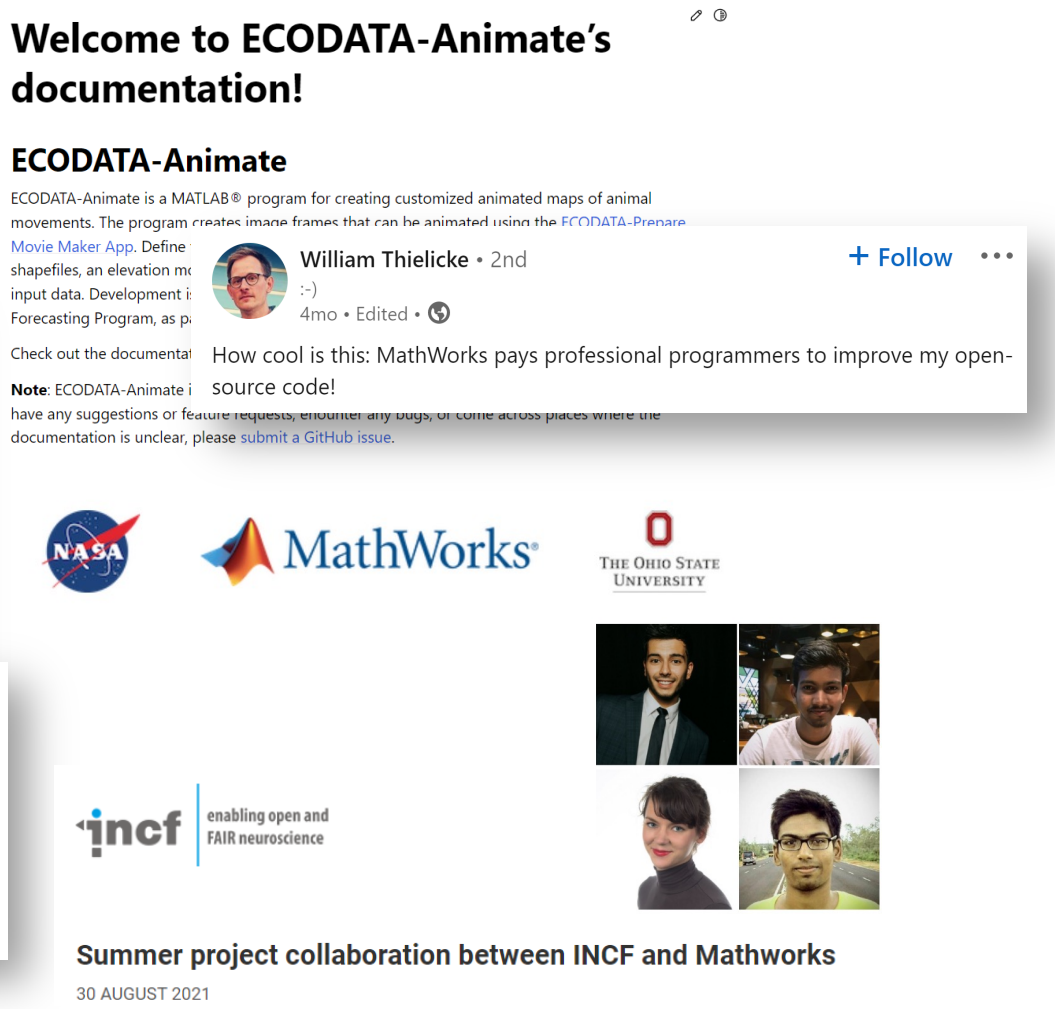

Published
VNFLOW MATLAB analysis
(Valery Vishnevskiy, Jonas Walheim & Sebastian Kozerke)
Capsule
File
Help
Files
Tabs
Core Files
metadata 1000 B
environment 498 B
code 68.78 KB
utils 45.54 KB
CREATE_FIGURES.m 67 B
init_globals.m 466 B
LICENSE 1.04 KB
prospective_figures.m 5.82 KB
run 206 B
vol_figures.m 10.97 KB
vol_table.m 4.67 KB
data Manage Datasets 58.37 GB
prospective 4.5 GB
volunteer 53.86 GB
LICENSE 6.4 KB
.gitignore 7 B
Results
results 1.51 MB
Other Files
Metadata
Metadata
Computer Science
VNFLOW MATLAB analysis
Valery Vishnevskiy, Jonas Walheim, Sebastian
vol_figures.m
1 simul_path = [base_data_path, 'volunteer/retro_data
2 recon_path = [base_data_path, 'volunteer/recon/'];
3 % bpath = [base_data_path, 'volunteer/recon/figs/']
4 bpath = [results_path, 'volunteer/figs/'];
5
6
7 errors_llr = struct();
8 errors_Ham = struct();
9 errors_vn = struct();
10 for ip = 7 % single volunteer
11 for ir = 5 % R=14
12 Rfact = R_list(ir);
13
14 pth = sprintf('%srecon_R%d_volN%d_vn.mat', recon_path, Rfact, ip);
15 rec = load(pth);
16 imrecon = squeeze(permute(rec.imrecon, ndims(rec.imrecon):-1:1));
17 im_vn = imrecon;
18
19 pth = sprintf('%srecon_R%d_volN%d_Ham.mat', recon_path, Rfact, ip);
20 rec = load(pth);
21 imrecon = squeeze(permute(rec.imrecon, ndims(rec.imrecon):-1:1));
22 im_Ham = imrecon;
Edit Capsule
Reproducible Run
or launch a cloud workstation
lab
Studio
jupyter
Shiny
Timeline
Reproducibility

MathWorks: hosting and funding open source community projects

File Exchange: Hosting > 40,000 open source community toolboxes



[File Exchange]



[MathWorks funds summer project to update community code]

Represented in CZ and SK by Humusoft

- MATLAB and Simulink are the tools of inspiration and innovation used by students, educators, and researchers around the world.

6500+
colleges and universities teach our software

2300+
MATLAB and Simulink-based books

Tens of Thousands
of skilled graduates enter the workforce each year

Thank you for your attention

Questions? Comments? Feedback? Slides?

shuboc@mathworks.com

danek@humusoft.cz