OPEN RESEARCH SOFTWARE INFRASTRUCTURE IN NEURO-MEDICINE

Adina Wagner
mas.to/@adswa

Psychoinformatics lab,
Institute of Neuroscience and Medicine, Brain & Behavior (INM-7)
Research Center Jülich

JÜLICH
Forschungszentrum

Slides: DOI 10.5281/zenodo.10149349 (Scan the QR code)
files.inm7.de/adina/talks/html/zimannheim.html

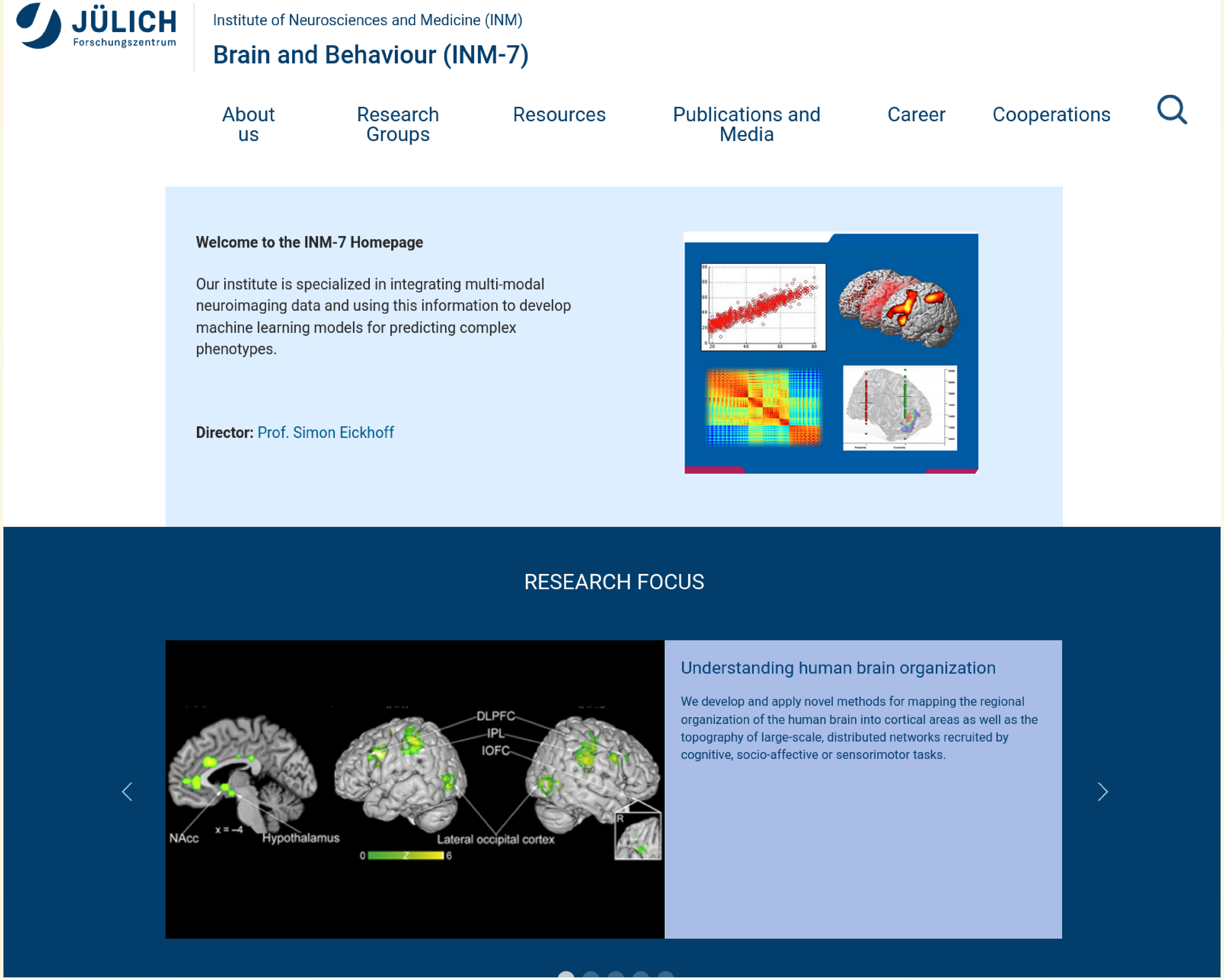
OPEN SCIENCE AND OPEN SOFTWARE GO HAND IN HAND

- Science has **specific requirements**; research software from within science ("from scientists, for scientists") can fulfill them. Open formats, protocols, and code allow re-use, interoperability, and customization.

- Open and reproducible science has specific needs for **transparency**: Open source software provides the necessary auditability.

- Creating software becomes **increasingly possible for scientists**: The San Francisco Declaration on Research Assessment (DORA; signed by FZJ), the Agreement on Reforming Research Assessment (CoARA), and the DFG recognize software as academic output.

THE INSTITUTE FOR NEUROSCIENCE & MEDICINE (INM-7)



Institute of Neurosciences and Medicine (INM)

Brain and Behaviour (INM-7)

Welcome to the INM-7 Homepage

Our institute is specialized in integrating multi-modal neuroimaging data and using this information to develop machine learning models for predicting complex phenotypes.

Director: Prof. Simon Eickhoff

RESEARCH FOCUS

Understanding human brain organization

We develop and apply novel methods for mapping the regional organization of the human brain into cortical areas as well as the topography of large-scale, distributed networks recruited by cognitive, socio-affective or sensorimotor tasks.

THE INSTITUTE FOR NEUROSCIENCE & MEDICINE (INM-7)

- Interdisciplinary institute with 11 research groups
- Research foci:
  - Infrastructure and method development: Digital biomarker, machine learning, meta analysis, research data management
  - Basic research in human brain mapping: Connectomics, genetic gradients, in-vivo brain mapping, multimodal integration
  - AI Applications in medical research: Cognition, Personality, Aging & neurodegenerative disease, Schizophrenia
  - Ethical implications of medical AI: Bias in AI applications, medical AI and society, individualized predictions

SOFTWARE @ INM-7

- The institute has a history of open source software, starting with the SPM Anatomy Toolbox (Eickhoff, 2005)
- Multiple groups develop and maintain open source research software for their respecitve subdomain
- Recent integration efforts connect our open software stack to open research software infrastructure for neuro-medicine

DataLad

- Domain-agnostic data management tool **(command-line** + **graphical user interface**), built on top of Git & Git-annex
- 10+ year open source project (100+ contributors), available for all major OS
- Born from rethinking data:
  - Just like code, **data is not static**.
  - Just like code, **data is subject to collaboration**. Stream-lined workflows for sharing and collaborating should be possible, mirroring those in software development.
  - **Provenance** of data is essential for reproducible, trustworthy, and FAIR science
  - Flexibility and **interoperability with existing tools** is the key to sustainability and ease of use

DataLad

- Domain-agnostic **command-line tool** (+ **graphical user interface**), built on top of Git & Git-annex
- 10+ year open source project (100+ contributors), available for all major OS
- Major features:

  Version-controlling arbitrarily large content

  Version control data & software alongside to code!

  Transport mechanisms for sharing, updating & obtaining data

  Consume & collaborate on data (analyses) like software

  (Computationally) reproducible data analysis

  Track and share provenance of all digital objects

  (... and much more)

EXHAUSTIVE TRACKING OF RESEARCH COMPONENTS

- text document
- source code
- binary file
- regular file
- file reference: identity and availability

Well-structured datasets (using community standards), and portable computational environments — and their evolution — are the precondition for reproducibility

```
# turn any directory into a dataset
# with version control


% datalad create <directory>
```

```
# save a new state of a dataset with
# file content of any size


% datalad save
```

CAPTURE COMPUTATIONAL PROVENANCE

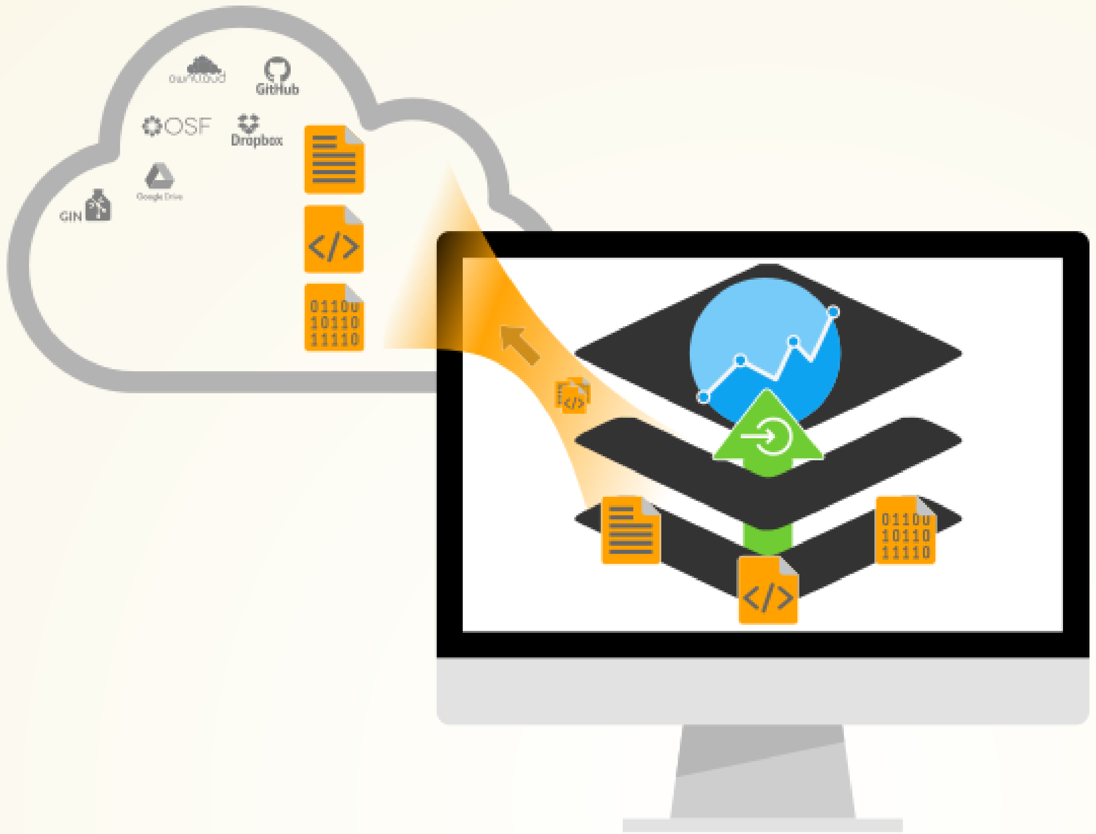
text document
source code
binary file
regular file
file reference: identity and availability

Which data was needed at which version, as input into which code, running with what parameterization in which computional environment, to generate an outcome?

```
# execute any command and capture its output
# while recording all input versions too


% datalad run --input ... --output ... <command>
```

EXHAUSTIVE CAPTURE ENABLES PORTABILITY



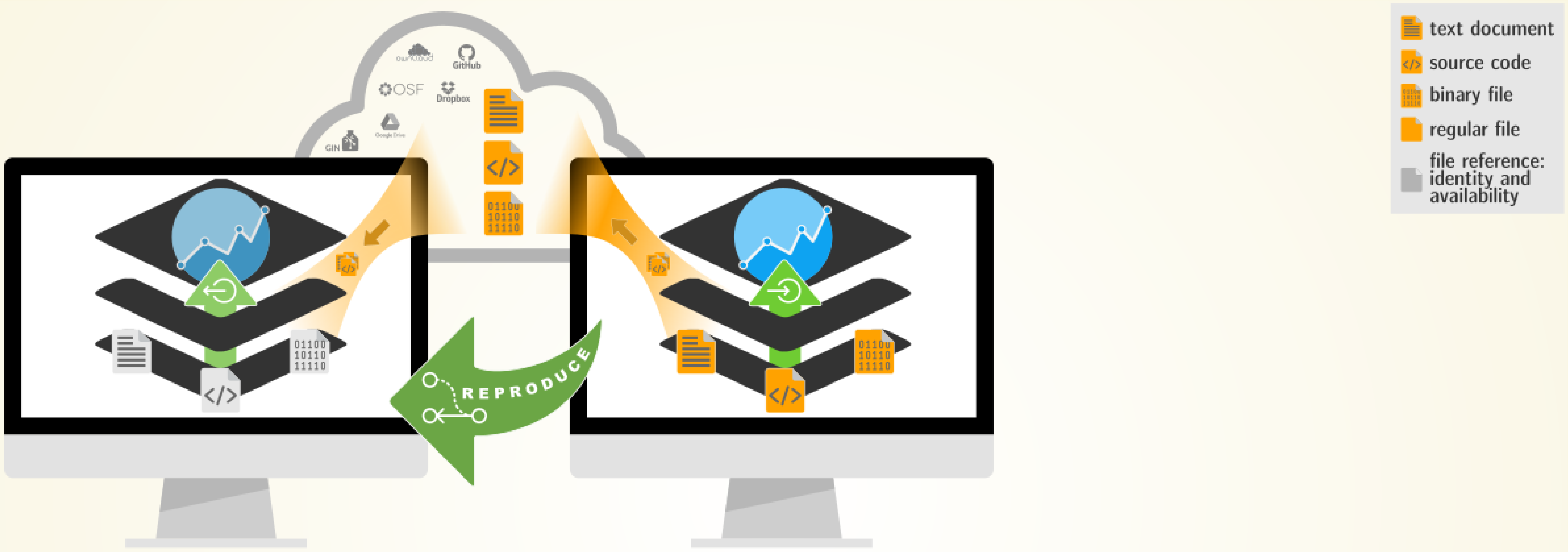
text document
source code
binary file
regular file
file reference: identity and availability

Precise identification of data and computational environments combined with provenance records form a comprehensive and portable data structure, capturing all aspects of an investigation.

```
# transfer data and metadata to other sites and services
# with fine-grained access control for dataset components


% datalad push --to <site-or-service>
```

REPRODUCIBILITY STRENGTHENS TRUST



Outcomes of computational transformations can be validated by authorized 3rd-parties. This enables audits, promotes accountability, and streamlines automated "upgrades" of outputs

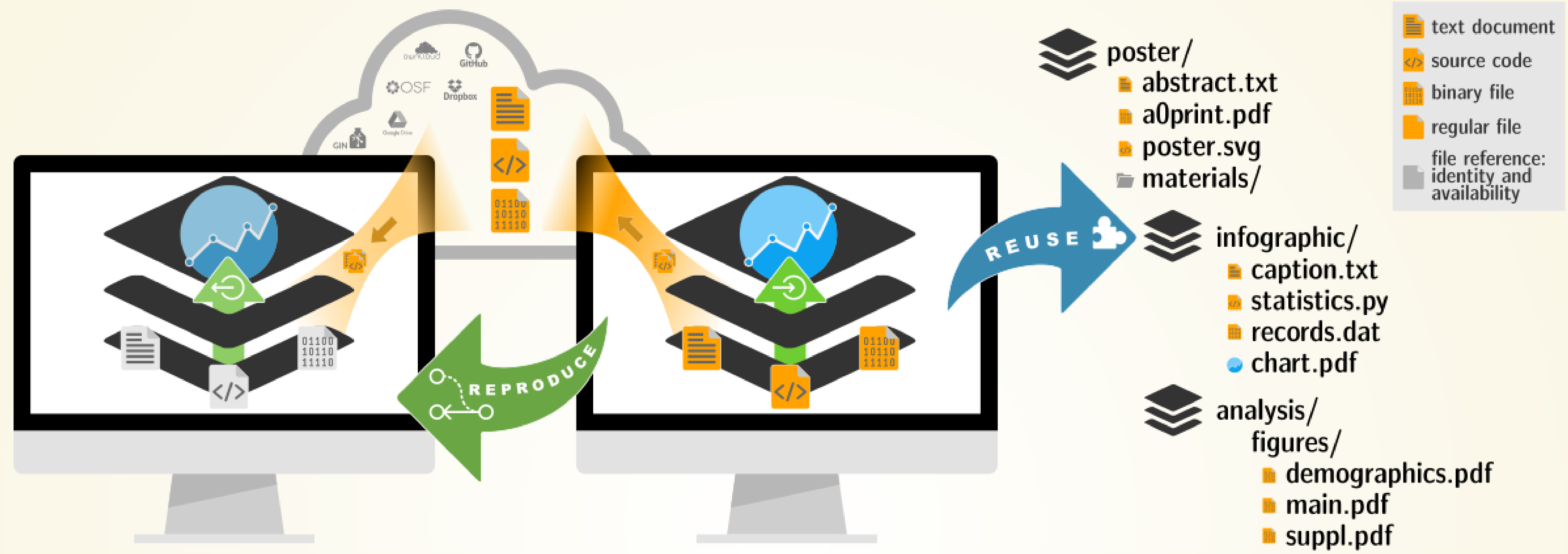
```
# obtain dataset (initially only identity,
# availability, and provenance metadata)


% datalad clone <url>
```

```
# immediately actionable provenance records
# full abstraction of input data retrieval


% datalad rerun <commit|tag|range>
```

ULTIMATE GOAL: (RE-)USABILITY



poster/
- abstract.txt
- a0print.pdf
- poster.svg
- materials/

infographic/
- caption.txt
- statistics.py
- records.dat
- chart.pdf

analysis/
- figures/
  - demographics.pdf
  - main.pdf
  - suppl.pdf

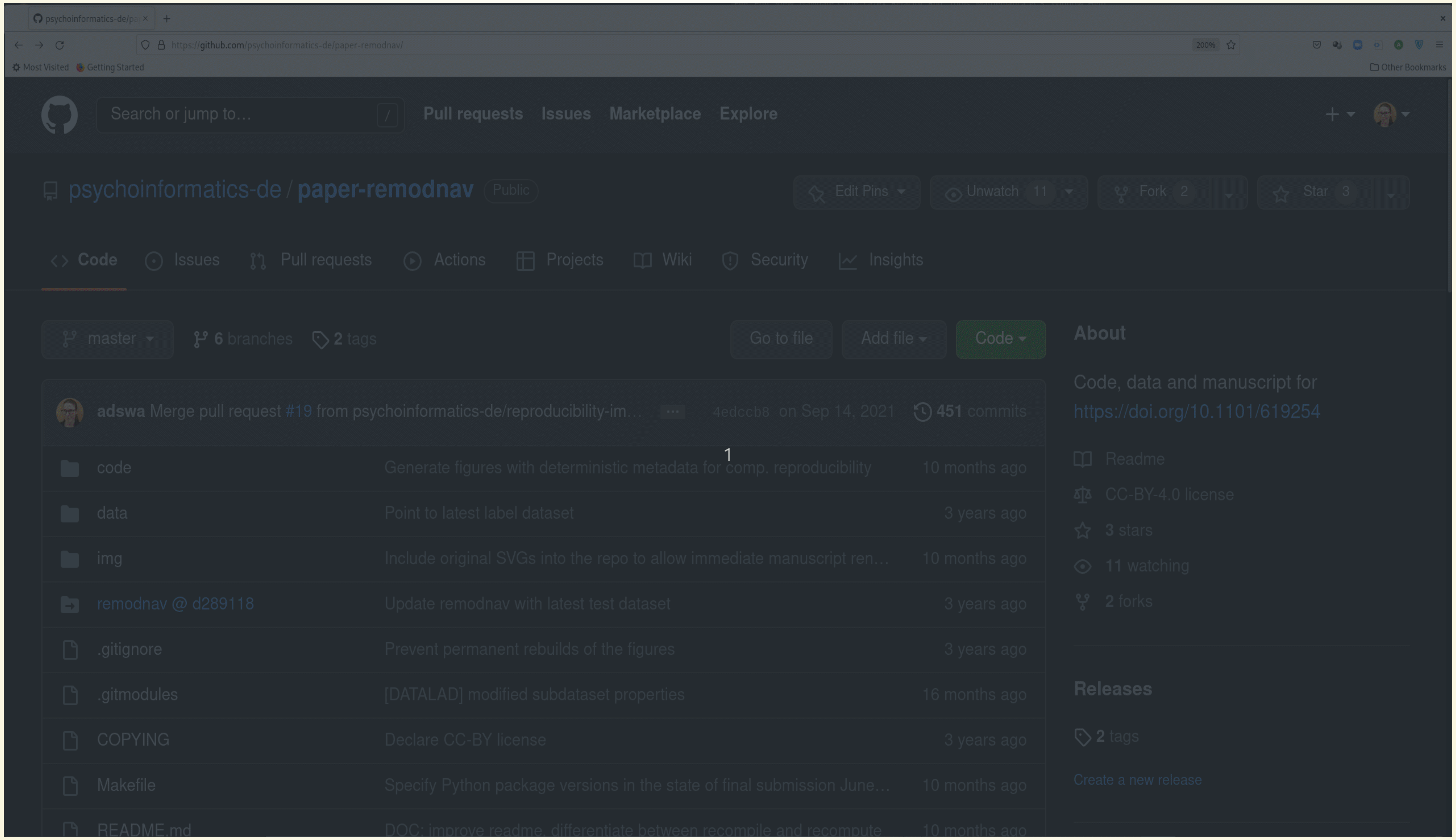
text document
source code
binary file
regular file
file reference: identity and availability

Verifiable, portable, self-contained data structures that track all aspects of an investigation exhaustively can be (re-)used as modular components in larger contexts — propagating their traits

```
# declare a dependency on another dataset and
# re-use it a particular state in a new context


% datalad clone -d <superdataset> <url> <path-in-dataset>
```

DATALAD USECASES



psychoinformatics-de / paper-remodnav Public

6 branches 2 tags

adswa Merge pull request #19 from psychoinformatics-de/reproducibility-im… 4edccb8 on Sep 14, 2021 451 commits

code	Generate figures with deterministic metadata for comp. reproducibility	10 months ago
data	Point to latest label dataset	3 years ago
img	Include original SVGs into the repo to allow immediate manuscript ren…	10 months ago
remodnav @ d289118	Update remodnav with latest test dataset	3 years ago
.gitignore	Prevent permanent rebuilds of the figures	3 years ago
.gitmodules	[DATALAD] modified subdataset properties	16 months ago
COPYING	Declare CC-BY license	3 years ago
Makefile	Specify Python package versions in the state of final submission June…	10 months ago
README.md	DOC: improve readme, differentiate between recompile and recompute…	10 months ago

About

Code, data and manuscript for https://doi.org/10.1101/619254

Readme
CC-BY-4.0 license
3 stars
11 watching
2 forks

Releases

2 tags

Create a new release

1

ACKNOWLEDGEMENTS

DataLad software & ecosystem

- Psychoinformatics Lab, Research center Jülich
- Center for Open Neuroscience, Dartmouth College
- Joey Hess (git-annex)
- *>100 additional contributors*

Funders

NSF

NSF 1429999

germany-usa

SPONSORED BY THE

Federal Ministry of Education and Research

BMBF 01GQ1411

JÜLICH Forschungszentrum

DFG

EUROPEAN UNION
European Regional Development Fund

cbbs center for behavioral brain sciences

SACHSEN-ANHALT
Ministerium für Wissenschaft und Wirtschaft

Collaborators

Human Brain Project

CONP PCNO

OpenNEURO

eBRAIN Health

MOTOR SFB 1451

brainlife.io

cbrain

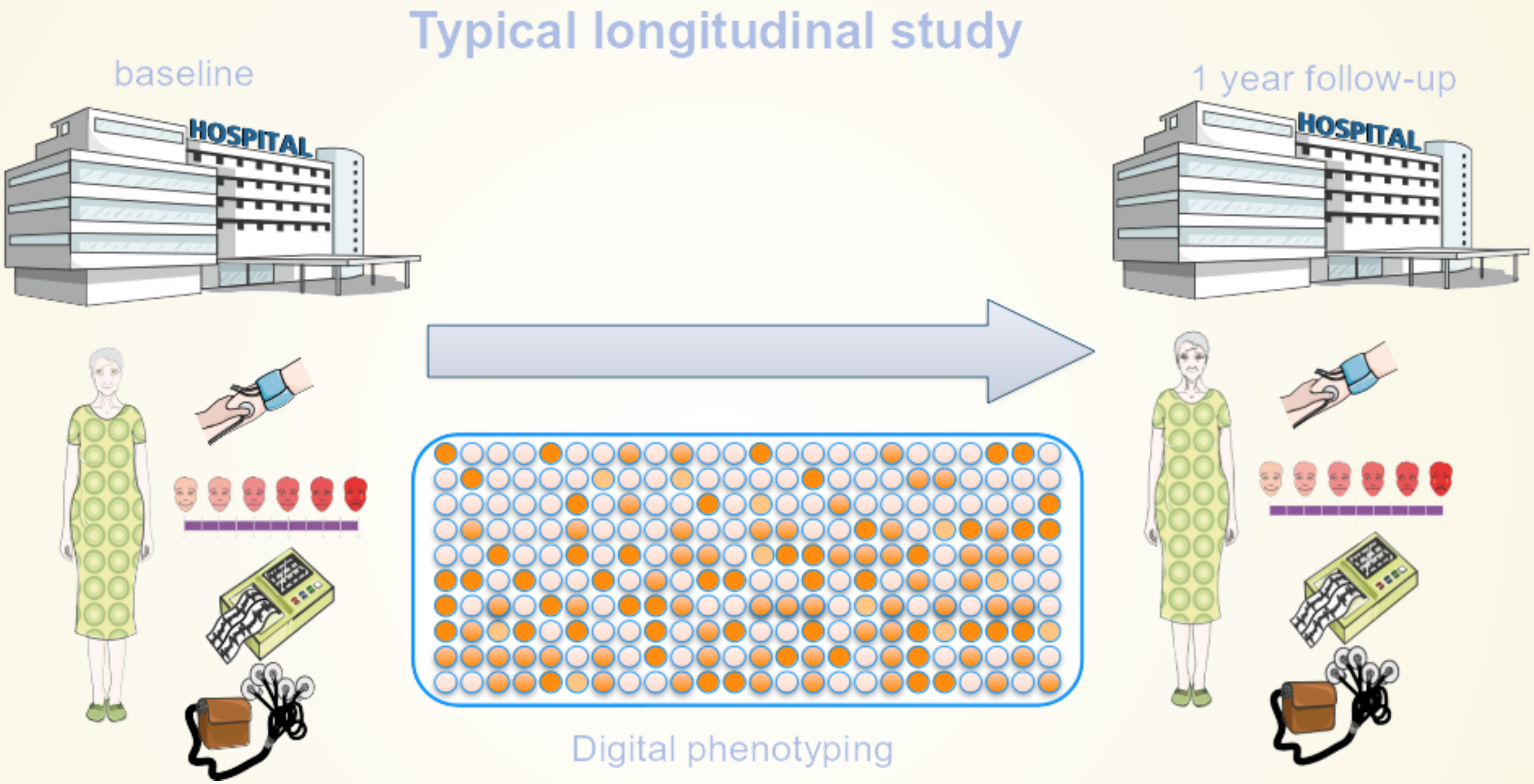
VirtualBrainCloud

JTRACK: DIGITAL BIOMARKERS FROM YOUR SMARTPHONE

- **Objective**: Close monitoring of patients/participants in non-clinical settings
- Modern smartphones contain a variety of sensors for passive monitoring and active acquisition:

  - gyroscope
  - accelerometer
  - location
  - human activity recognition
  - application usage
  - screen time
  - microphone

JTRACK



Flexible components for different users:

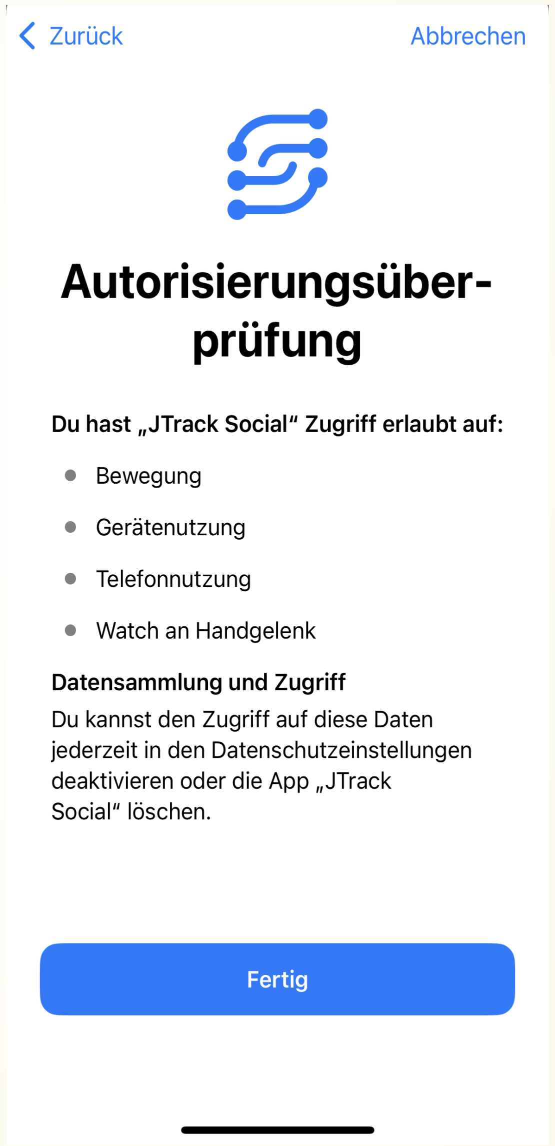
- **JTrack Social**: Smartphone app for participants
- **JTrack EMA**: Smartphone app for participants
- **JDash**: Monitoring and analytics tool for study owners

JTRACK COMPONENTS: JTRACK SOCIAL

Smartphone App for active labeling and passive monitoring

Available for Android + IPhone

- Sensor data (passive collection mode default: Accelerometer & Gyroscope)
- Application usage statistics
- Human activation recognition (e.g., walking, running, driving)
- Location information (anonymized)
- Active recording, e.g., free-speech generation tasks



Autorisierungsüber-prüfung

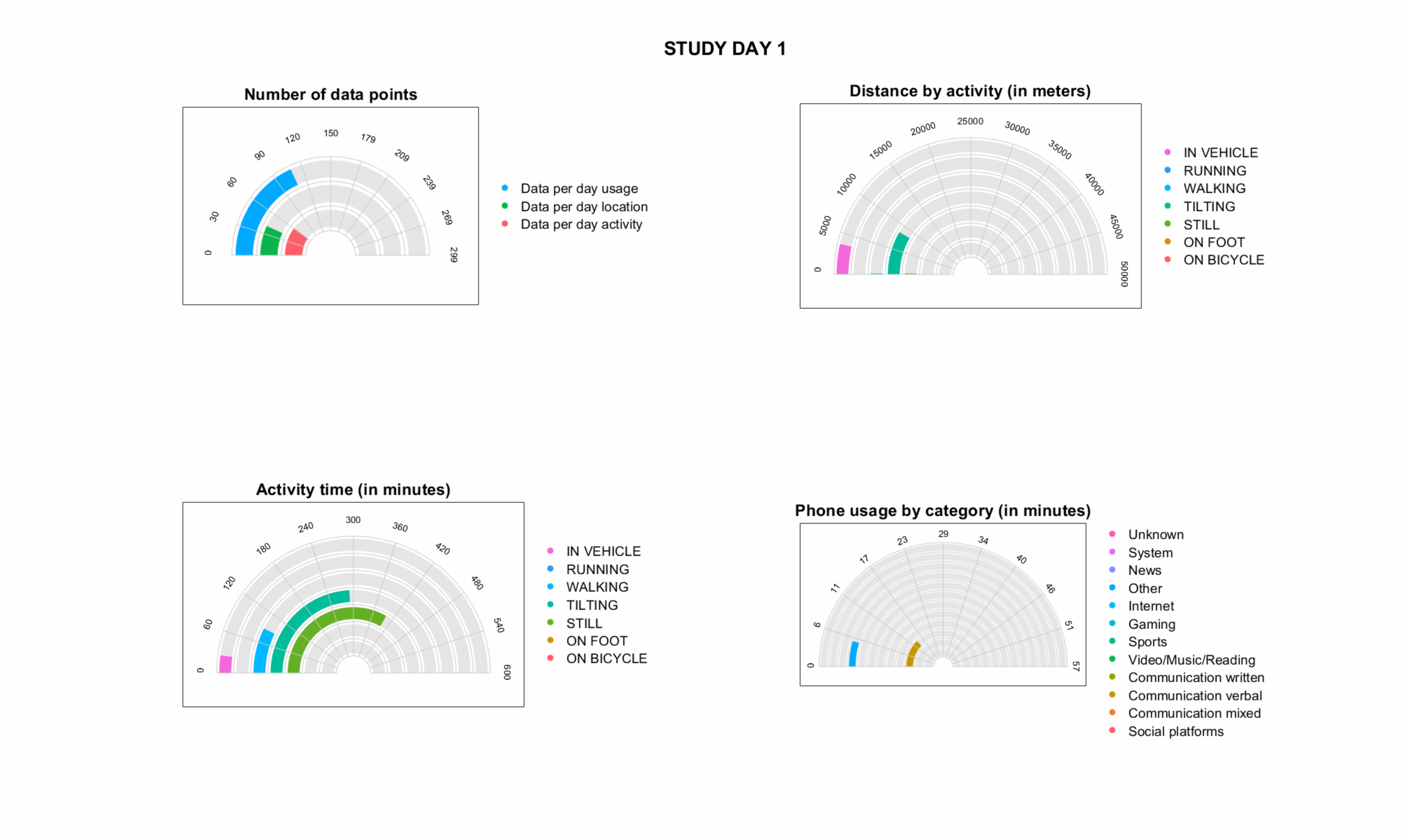
Du hast „JTrack Social" Zugriff erlaubt auf:

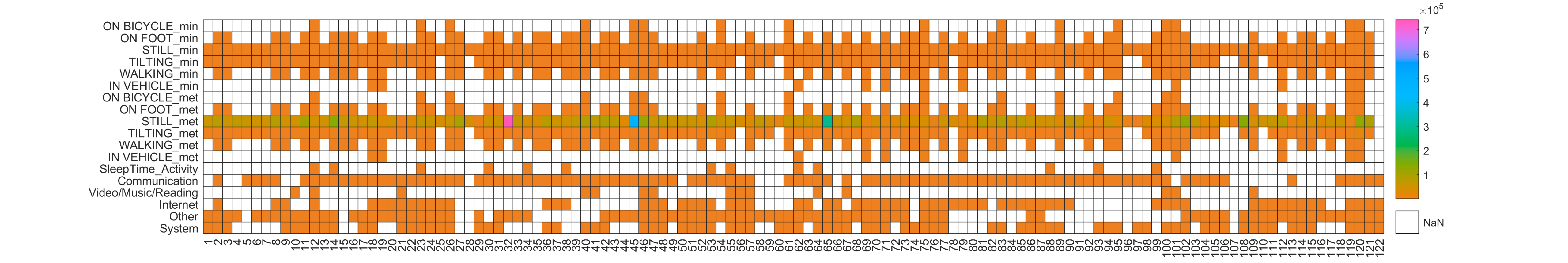
- Bewegung
- Gerätenutzung
- Telefonnutzung
- Watch an Handgelenk

Datensammlung und Zugriff

Du kannst den Zugriff auf diese Daten jederzeit in den Datenschutzeinstellungen deaktivieren oder die App „JTrack Social" löschen.

JTRACK COMPONENTS: JTRACK SOCIAL





JTRACK COMPONENTS: JTRACK EMA

Smartphone App for Ecological Momentary Assessment Available for Android + IPhone

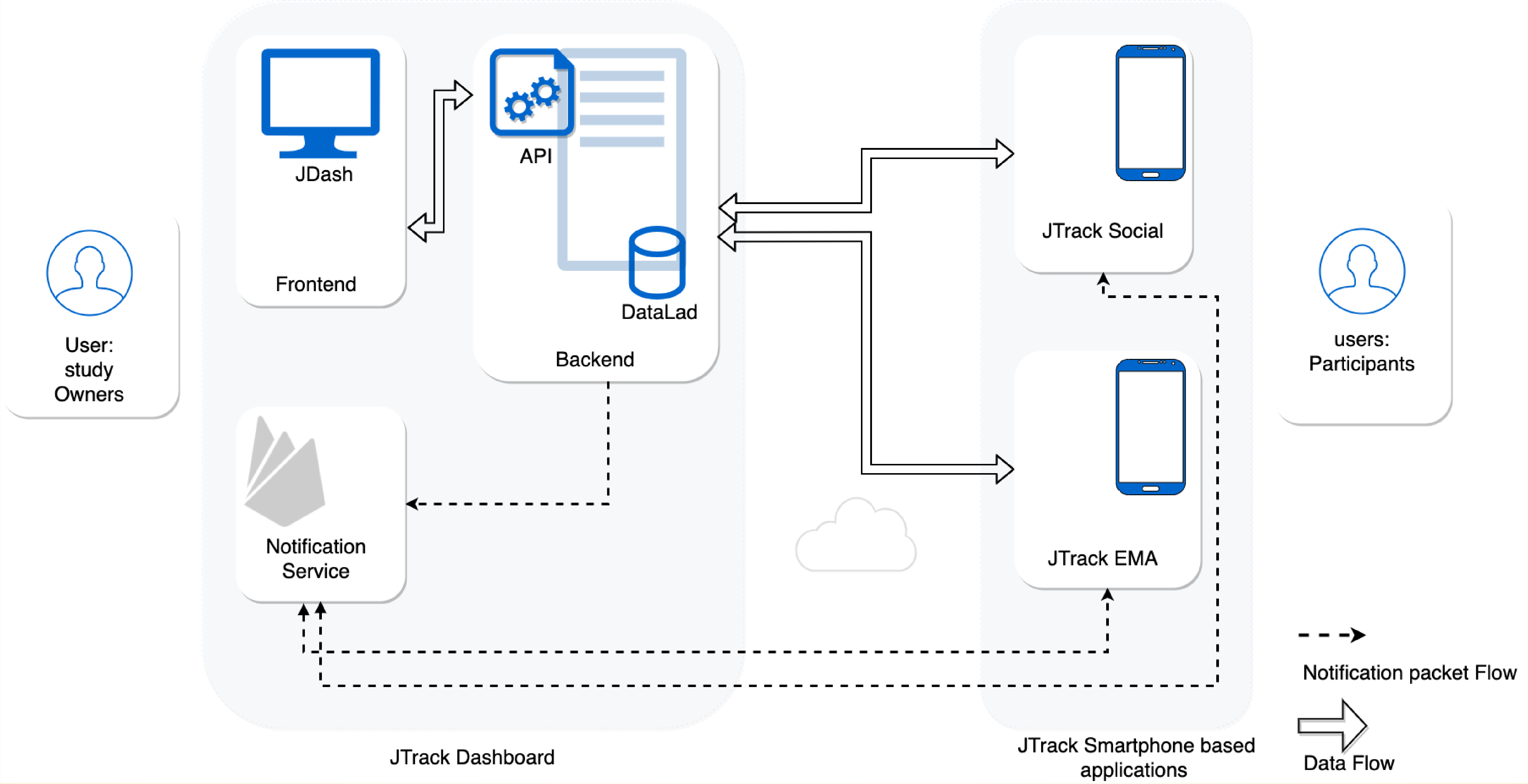
- Binary Questions
- Date and Time Questions
- Sliding Questions
- Multiple/Single coice questions

JTRACK COMPONENTS: JDASH

Dashboard for Study Administration

- Investigator's study & user management
- Data Quality Control
- Notification Center

BEHIND THE SCENES

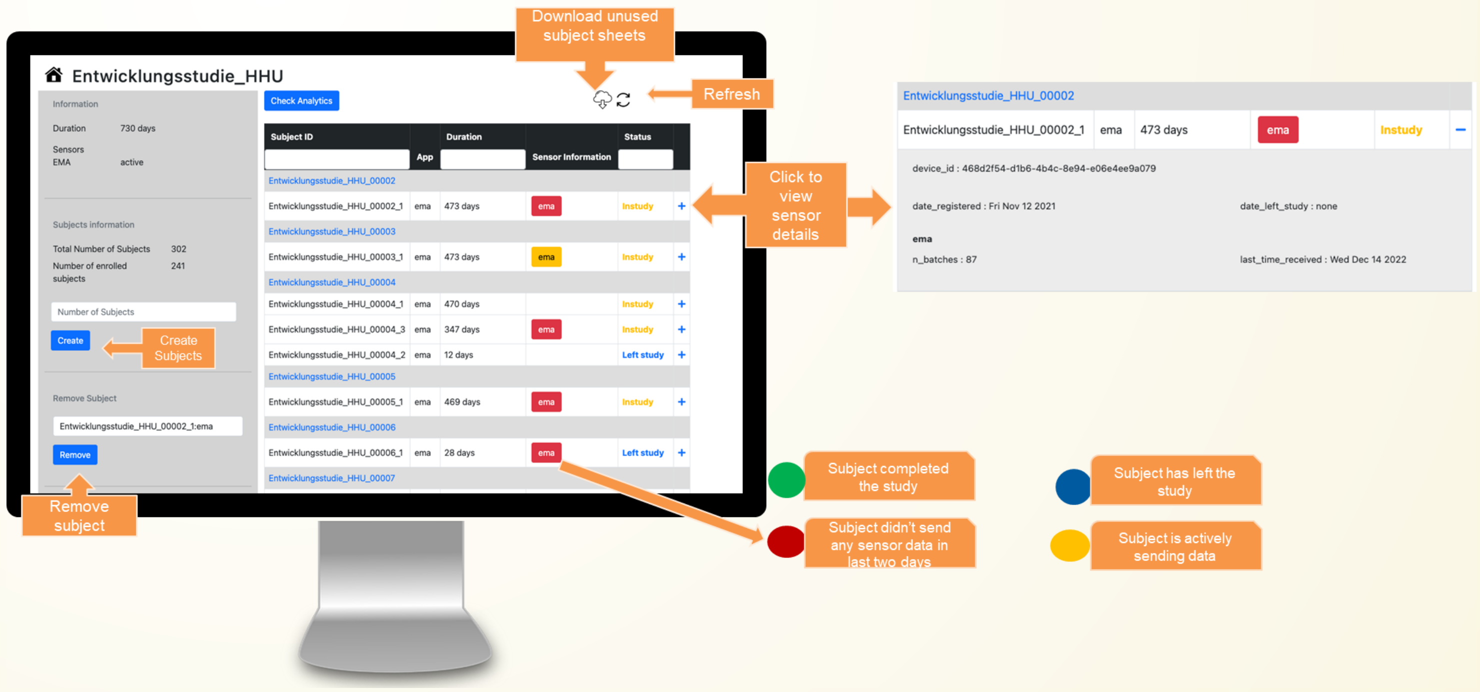


- Servers in Germany
- Data versioning via DataLad
- Authenticated data access via JDash
- Data transfer via HTTPS

INVESTIGATOR'S / PARTICIPANT'S POINT OF VIEW

- Install JTrack
- Scan QR code and give permissions to App
- Resume daily life

- Install JTrack with Participant
- Provide study- and subject-specific QR code from JDash
- Monitor study and communicate with participants via JDash



Entwicklungsstudie_HHU

Information

Duration	730 days
Sensors	
EMA	active

Subjects information

Total Number of Subjects	302
Number of enrolled subjects	241

Number of Subjects

Create

Remove Subject

Entwicklungsstudie_HHU_00002_1:ema

Remove

Check Analytics

Subject ID	App	Duration	Sensor Information	Status	
Entwicklungsstudie_HHU_00002					
Entwicklungsstudie_HHU_00002_1	ema	473 days	ema	Instudy	+
Entwicklungsstudie_HHU_00003					
Entwicklungsstudie_HHU_00003_1	ema	473 days	ema	Instudy	+
Entwicklungsstudie_HHU_00004					
Entwicklungsstudie_HHU_00004_1	ema	470 days		Instudy	+
Entwicklungsstudie_HHU_00004_3	ema	347 days	ema	Instudy	+
Entwicklungsstudie_HHU_00004_2	ema	12 days		Left study	+
Entwicklungsstudie_HHU_00005					
Entwicklungsstudie_HHU_00005_1	ema	469 days	ema	Instudy	+
Entwicklungsstudie_HHU_00006					
Entwicklungsstudie_HHU_00006_1	ema	28 days	ema	Left study	+
Entwicklungsstudie_HHU_00007					

Entwicklungsstudie_HHU_00002					
Entwicklungsstudie_HHU_00002_1	ema	473 days	ema	Instudy	−

device_id : 468d2f54-d1b6-4b4c-8e94-e06e4ee9a079

date_registered : Fri Nov 12 2021

date_left_study : none

ema

n_batches : 87

last_time_received : Wed Dec 14 2022

Download unused subject sheets

Refresh

Click to view sensor details

Create Subjects

Remove subject

Subject completed the study

Subject didn't send any sensor data in last two days

Subject has left the study

Subject is actively sending data

ADVANTAGES

- Easy to deploy and free environment for collection of real world data (RWD) basically at no cost
- Standardized data collection across centers
- High-density longitudinal data with fully customizable data collection
- Opportunity for citizen science

ACKNOWLEDGEMENTS

Publications:

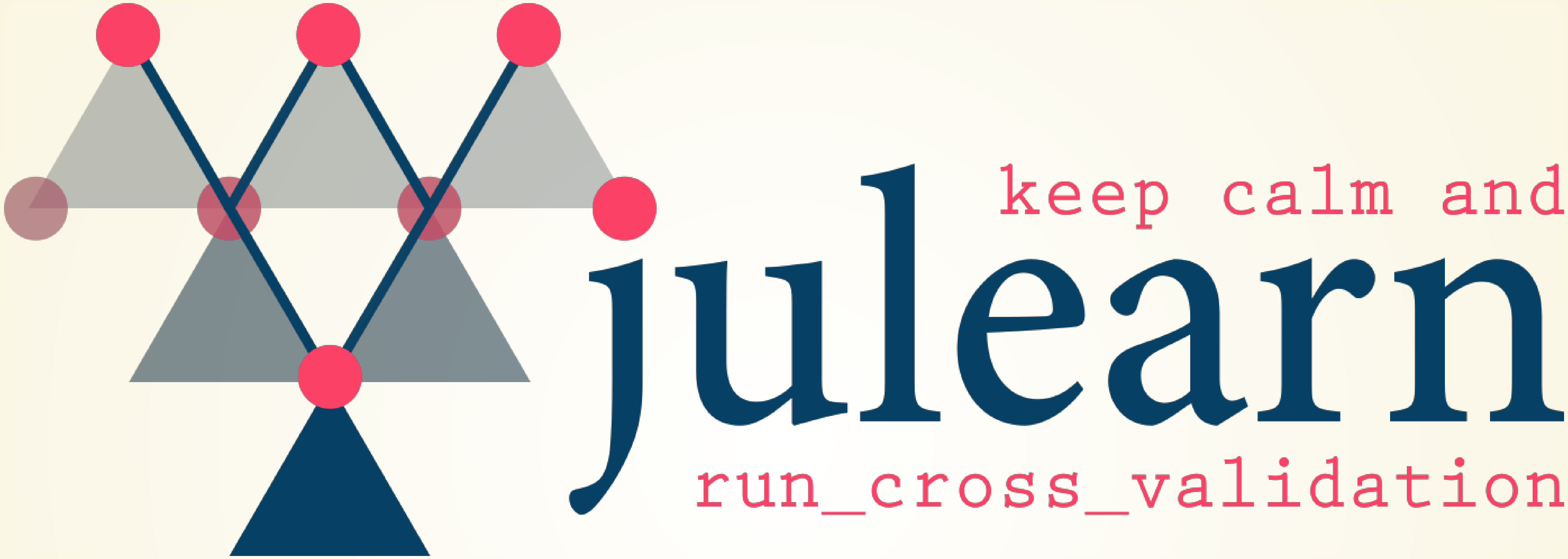
- JTrack Social: Sahandi Far et al. 2021
- JTrack EMA: Sahandi Far et al. 2023

Contact:

- via JDash: jdash.inm7.de

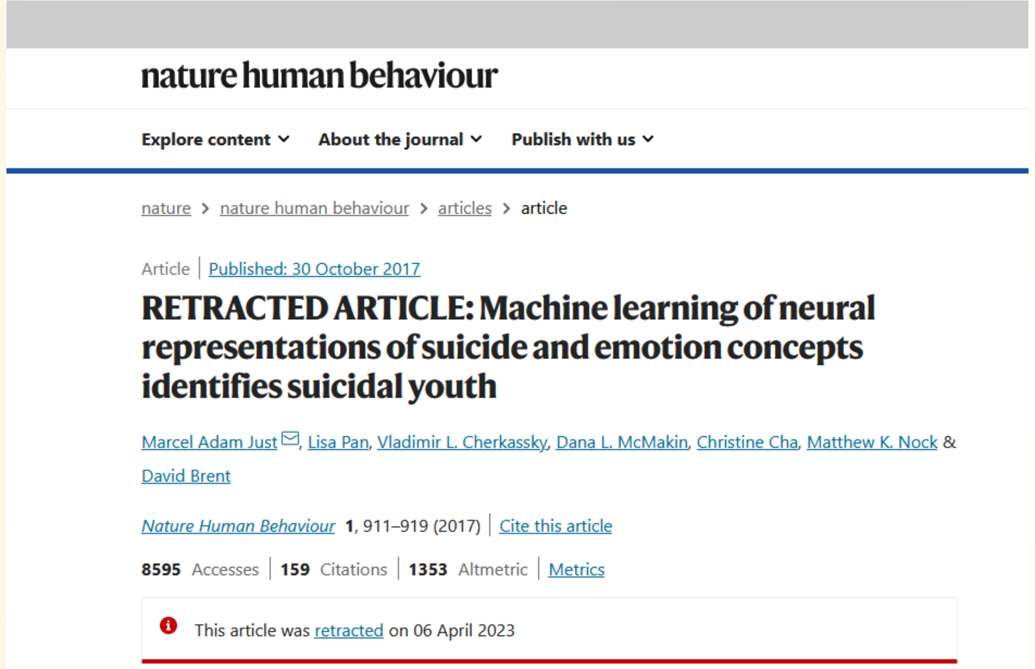
- **JTrack Hour** (open to everyone)
  Every second Tuesday at 1PM (even-numbered calendar weeks) -
  Join the Zoom meeting!

Team:



- Open source Python library for easy-to use ML-pipelines, built upon scikit-learn
- Domain-general, but aims to simplify entry into ML for domain scientists with built-in guarantees against most common pitfalls:
  - Data leakage
  - Overfitting of hyperparameters

The problem: Expensive AI mistakes



Article | Published: 30 October 2017

RETRACTED ARTICLE: Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock & David Brent

Nature Human Behaviour **1**, 911–919 (2017) | Cite this article

8595 Accesses | **159** Citations | **1353** Altmetric | Metrics

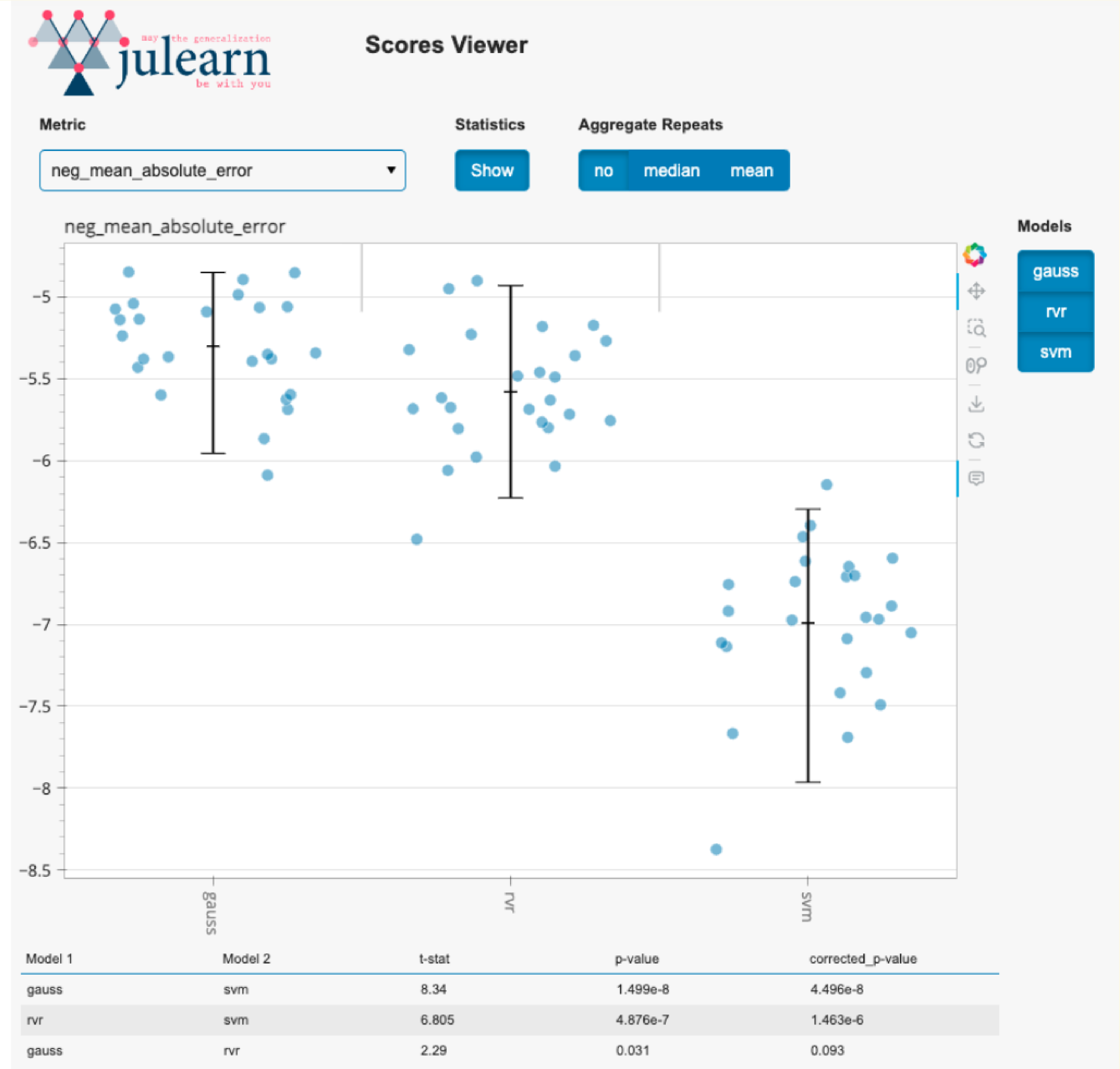
This article was retracted on 06 April 2023

A solution: User-friendly solutions to common complex use cases

- Simplifies common use cases for supervised ML pipelines, with feature such as:
  - Automatic usage of nested cross-validation for proper evaluation in hyperparameter tuning
  - Preprocessing based on feature types, incl. confound removal
  - Built-in visualization for model inspection and comparison
- Plug-and-play with scikit-learn transformers

VISUALIZATION

Interactive "Scores Viewer" for easier model comparison



Model 1	Model 2	t-stat	p-value	corrected_p-value
gauss	svm	8.34	1.499e-8	4.496e-8
rvr	svm	6.805	4.876e-7	1.463e-6
gauss	rvr	2.29	0.031	0.093

JULEARN VS SCIKIT-LEARN

Simple CV pipeline

```
from julearn import run_cross_validation
run_cross_validation(
    X=X, y=y, data=data,
    preprocess=["zscore"], model="svm",
    problem_type="classification",
    X_types={"continuous": X} # X_types optional here
```

```
from sklearn.model_selection import cross_validate
from sklearn.svm import SVC # SVR in case of regression
from sklearn.preprocessing import StandardScaler
from sklearn.pipelines import make_pipeline

pipeline = make_pipeline(StandardScaler(), SVC())
cross_validate(X=data.loc[:,X], y=data.loc[:,y], estimator=pipeline)
```

JULEARN VS SCIKIT-LEARN

Nested CV with hyperparameter tuning

```
from julearn import run_cross_validation, PipelineCreator
creator=PipelineCreator(problem_type="classification")
creator.add("zscore", with_mean=[True, False])
creator.add("pca", n_components=2)
creator.add("svm", C=[1,2], degree=[3,4])

# X_types optional
run_cross_validatoin(
    X=X, y=y, data=data, model=creator, X_types={"continuous": X})
```

```
from sklearn.model_selection import cross_validate, GridSearchCV
from sklearn.svm import SVC # SVR in case of regression
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.pipelines import make_pipeline

pipeline = make_pipeline(StandardScaler(), PCA(), SVC())
param_grid = {
    "standardscaler__with_mean": [True, False],
    "pca__n_components": [2],
    "scv__C": [1,2],
    "svc__degree": [3, 4]
}
grid_pipeline = GridSearchCV(estimator=pipeline, param_grid=param_grid)
cross_validate(X=data.loc[:,X], y=data.loc[:,y], estimator=pipeline)
```

ACKNOWLEDGEMENTS

- Preprint: Hamdan et al., 2023
- Documentation: juaml.github.io/julearn
- Source Code: github.com/juaml/julearn

Fede Raimondo

Sami Hamdan

Kaustubh Patil

Shammi More

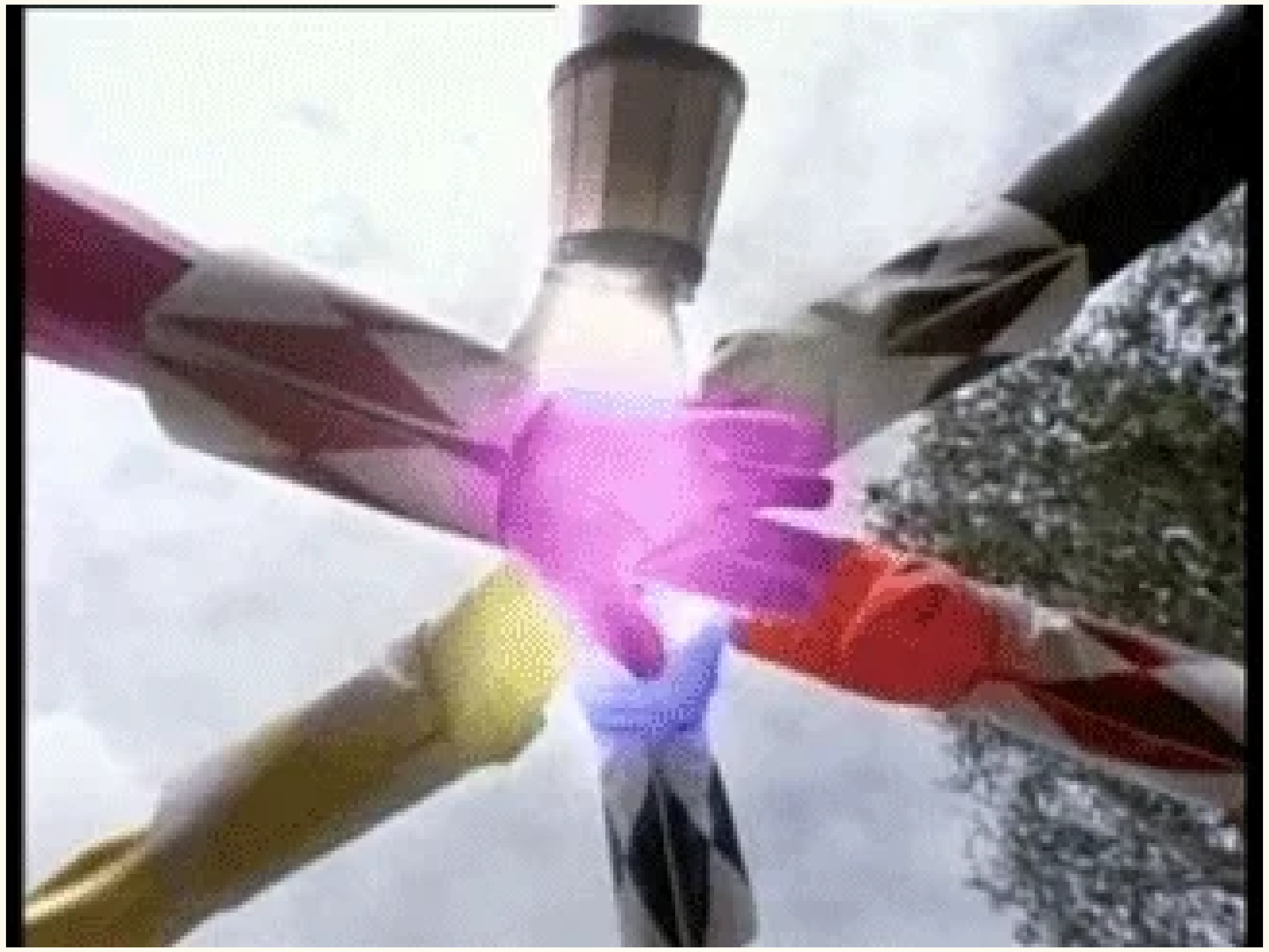
Vera Komeyer

Synchon Mandal

Leonard Sasse

ML Hours (open to everyone)
Consultancy on Machine-Learning, every second Thursday, 2-4pm
Chat: https://matrix.to/#/#ml:inm7.de

EVEN BETTER TOGETHER



THE ABCD-J PLATFORM

AN OPEN SOURCE PLATFORM FOR DIGITAL BIOMARKER FOR NEURO-MEDICINE IN NRW

A collaboration between clinical, academic, and industry partners:

Research Center Jülich

RWTH Aachen
University Bonn
University Cologne
HHU Düsseldorf

LVR Clinics
DZNE

PeakProfiling
CanControl
IXP

(open to future additions)

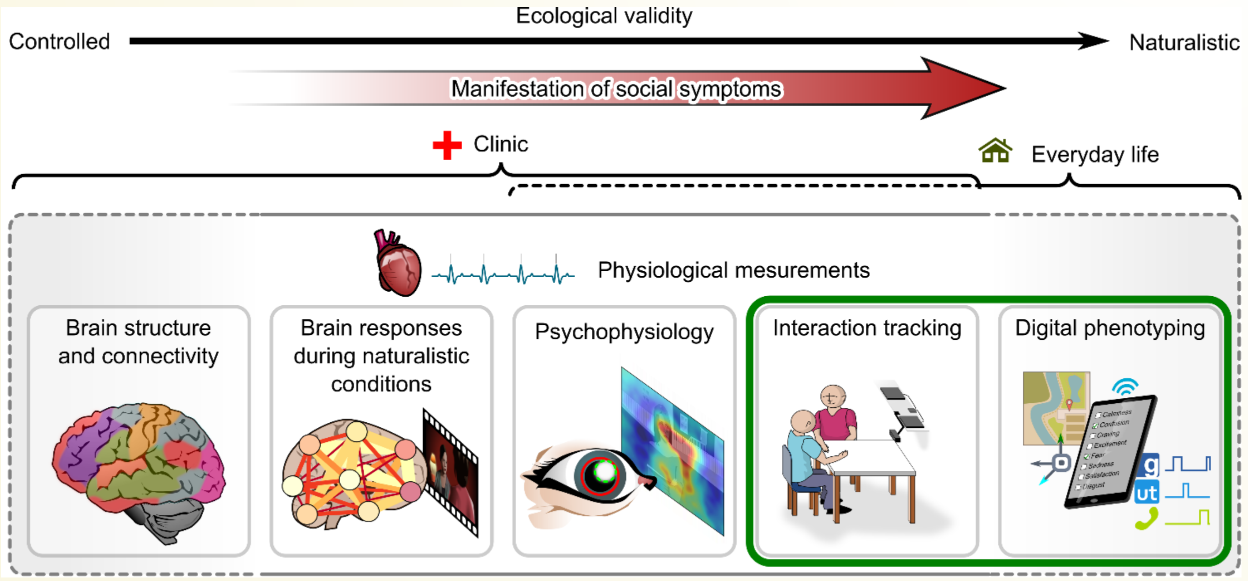
A
B
C
D
/ha/
/pa/
/k/
/tuin/
WEITERE MODULE

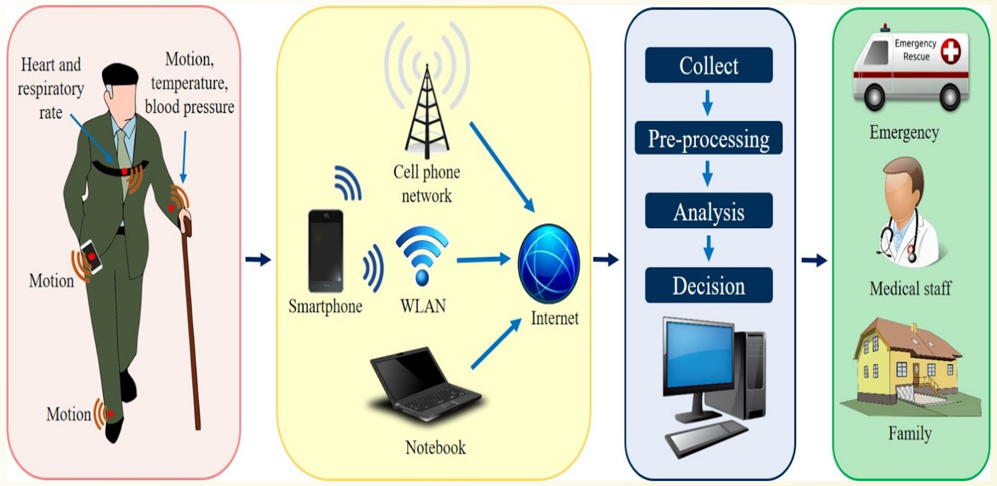
GOALS

- **Social:** Promote and facilitate collaboration between multiple centers
- **Technical:** Accelarate research through homogenization of workflows and processes, with emphasis on digital biomarker development; Elevate existing open technical solutions for research practice adoption

OPEN RESEARCH INFRASTRUCTURE

Clinicians' point of view	Patients' point of view
• "Deep phenotyping": ecologically valid, multimodal data • Decentral data acquisition, standardized and reproducible • Focus on patient well-being and optimal treatment	• Accurate diagnosis and optimal treatment • Strict data protection • Individual patient is central • Minimal disturbance in daily life


Ecological validity
Controlled
Naturalistic
Manifestation of social symptoms
Clinic
Everyday life
Physiological mesurements
Brain structure and connectivity
Brain responses during naturalistic conditions
Psychophysiology
Interaction tracking
Digital phenotyping

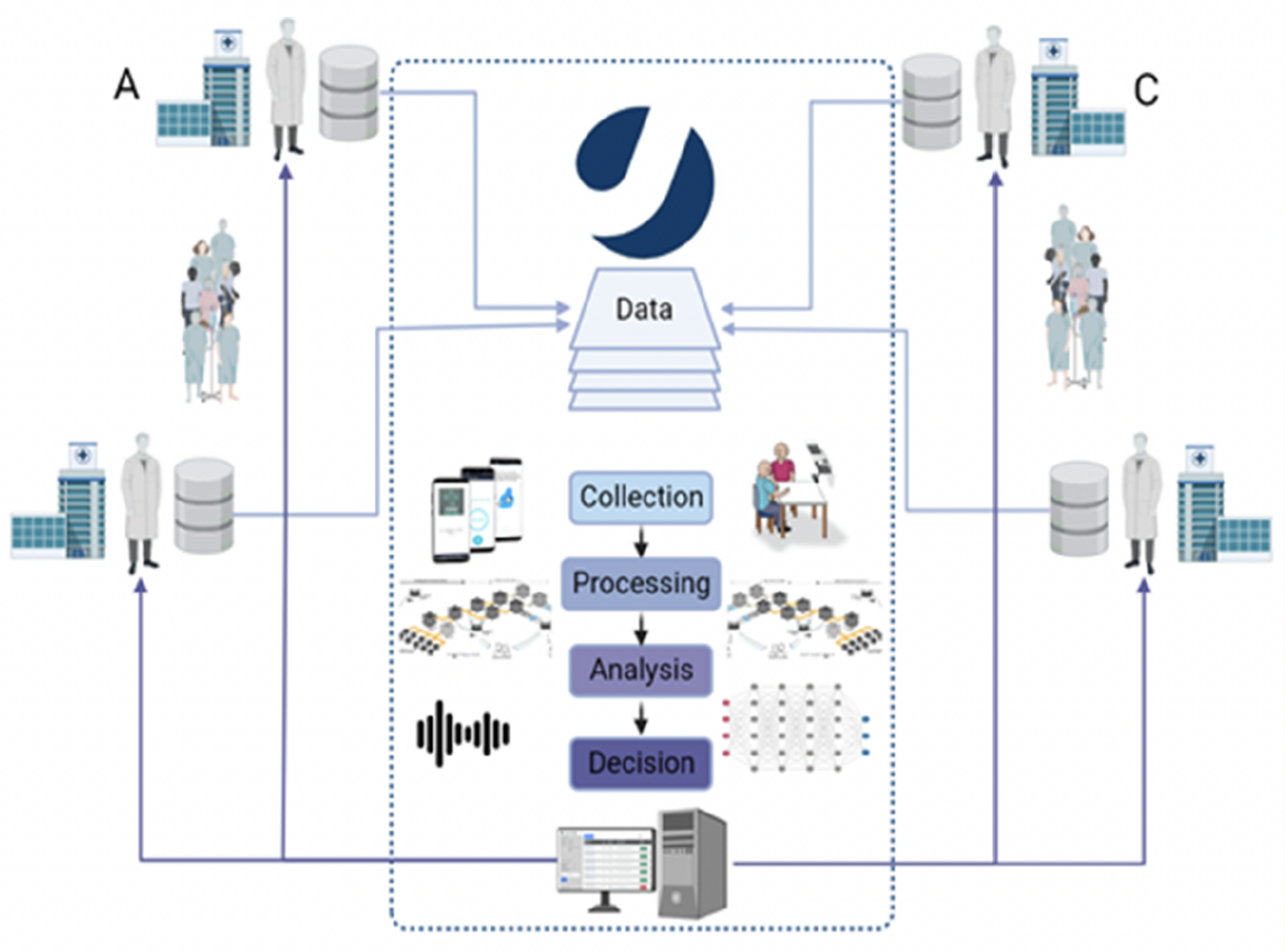


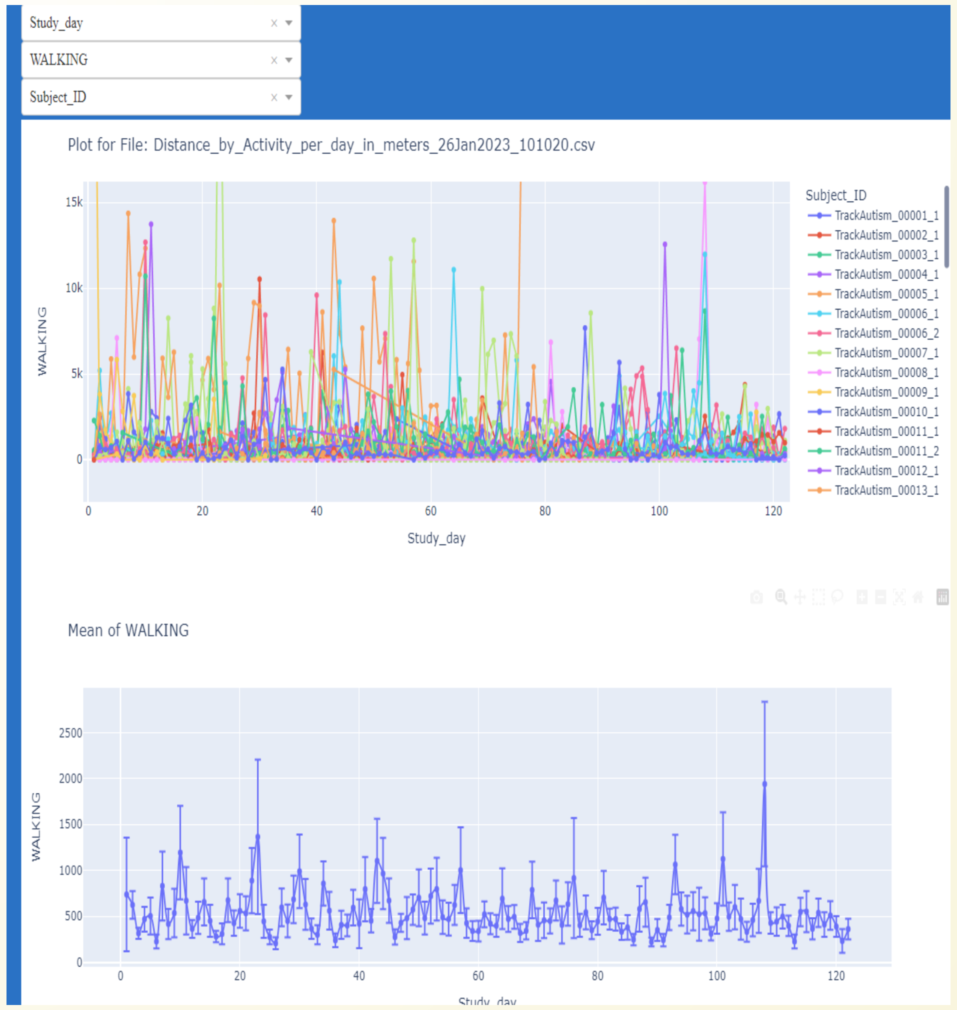
OPEN RESEARCH INFRASTRUCTURE

Research data management's point of view
• Technical solutions exists, need elevation for research practice adoption • Homogenization of workflows and processes fosters collaboration across sites • Decentralized approach with centralized services and web-based multi-center integration

FRONT-END AND BACK-END

- JTrack for decentral, ecologically valid acquisitions, complementing in-clinic assessments
- JDash for study management, participant management, and analytics overview (derived study data at subject & group level)
- Central data overview and analytics at FZJ
  - Provenance-tracked analysis and modeling
  - Automated meta-data extraction for data discoverability
  - Result overview for clinical decision making





Plot for File: Distance_by_Activity_per_day_in_meters_26Jan2023_101020.csv

Mean of WALKING

OPPORTUNITIES

Software improves with its use cases

- JTrack integration into different types wearables
- JTrack integration of cognitive tasks and feedback to participants
- Julearn integration into JDash
- More meta-data extractors for DataLad
- ...

CURRENT (FIRST) STEPS

Data cataloging

- Leveraging legacy data via data census and meta-data catalog
    - improved discovery without direct data transfer
    - homogenization of access request procedures
    - establishing a legal basis for (re-)use
    - Example: data.sfb1451.de
- Demonstrator for data infrastructure based on §21 data (standardized and anonymized performance data of hospitals, legally required, submitted yearly to InEK by all hospitals)

Feasibility/Proof-of-concept study

- Recommendations for common digital tools and workflows for common tasks and processes
- Selection of digital measures from clinical routine for first trials

SUMMARY

- Open Source (Research) Software aids in various domain-general or -specific applications.
- Open Science needs Open Source: For transparency and reproducibility, for science-specific requirements, for open formats, for re-use, and to enable interoperability across tools.
- Collaboration across clinical and research settings is a technical, social, and legal challenge. Technical solutions won't save us alone, but they are a good first step.

THANKS!

Questions?

Inputs, Resources, Synergies