VAMP IT UP!

A PRAGMATIC APPROACH TO REUSABLE RESEARCH OUTPUTS

Michael Hanke

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

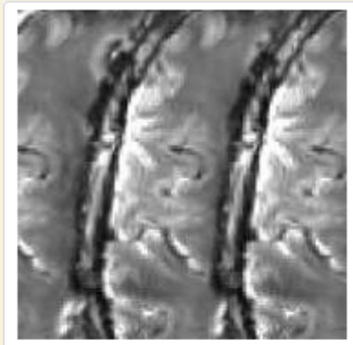
Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf

http://psychoinformatics.de

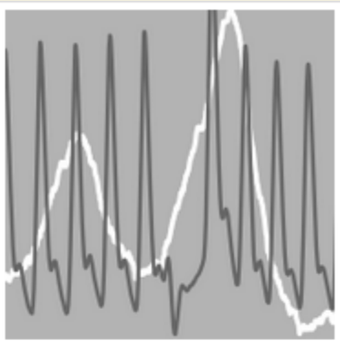
http://pymvpa.org

http://neuro.debian.org

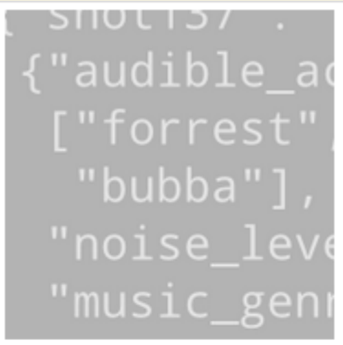
OPEN, "NATURALISTIC" DATA: STUDYFORREST.ORG



Functional MRI

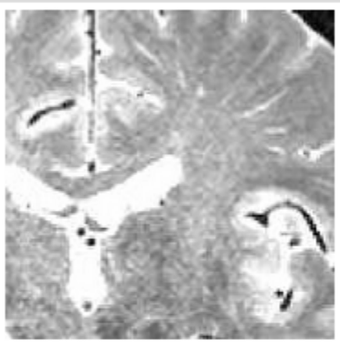


Physiological measurements



Annotations & surveys

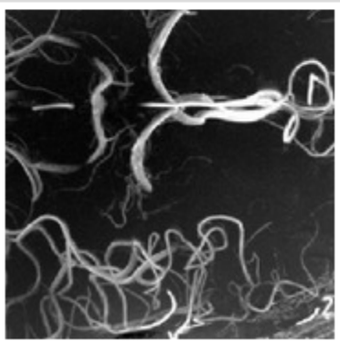
T1-weighted MRI



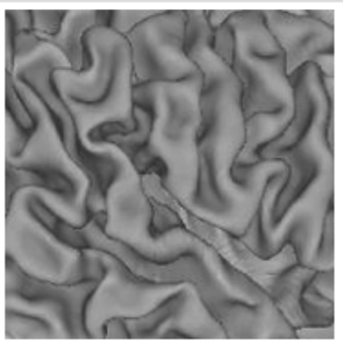
T2-weighted MRI

Susceptibility-weighted MRI

Diffusion tensor MRI



Angiography



Surface reconstruction

Hanke, Baumgartner, Ibe, Kaule, Pollmann, Speck, Zinke, & Stadler (2014) *A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie*. Scientific Data, 1:140003.
http://www.nature.com/articles/sdata20143

Findable

Accessible

Interoperable

Reusable

https://www.go-fair.org/fair-principles

F? I already have it, it's right here!

A? I am working with it already, I made it!

I? With what?

R? First let me finish this PhD and then we talk, OK?

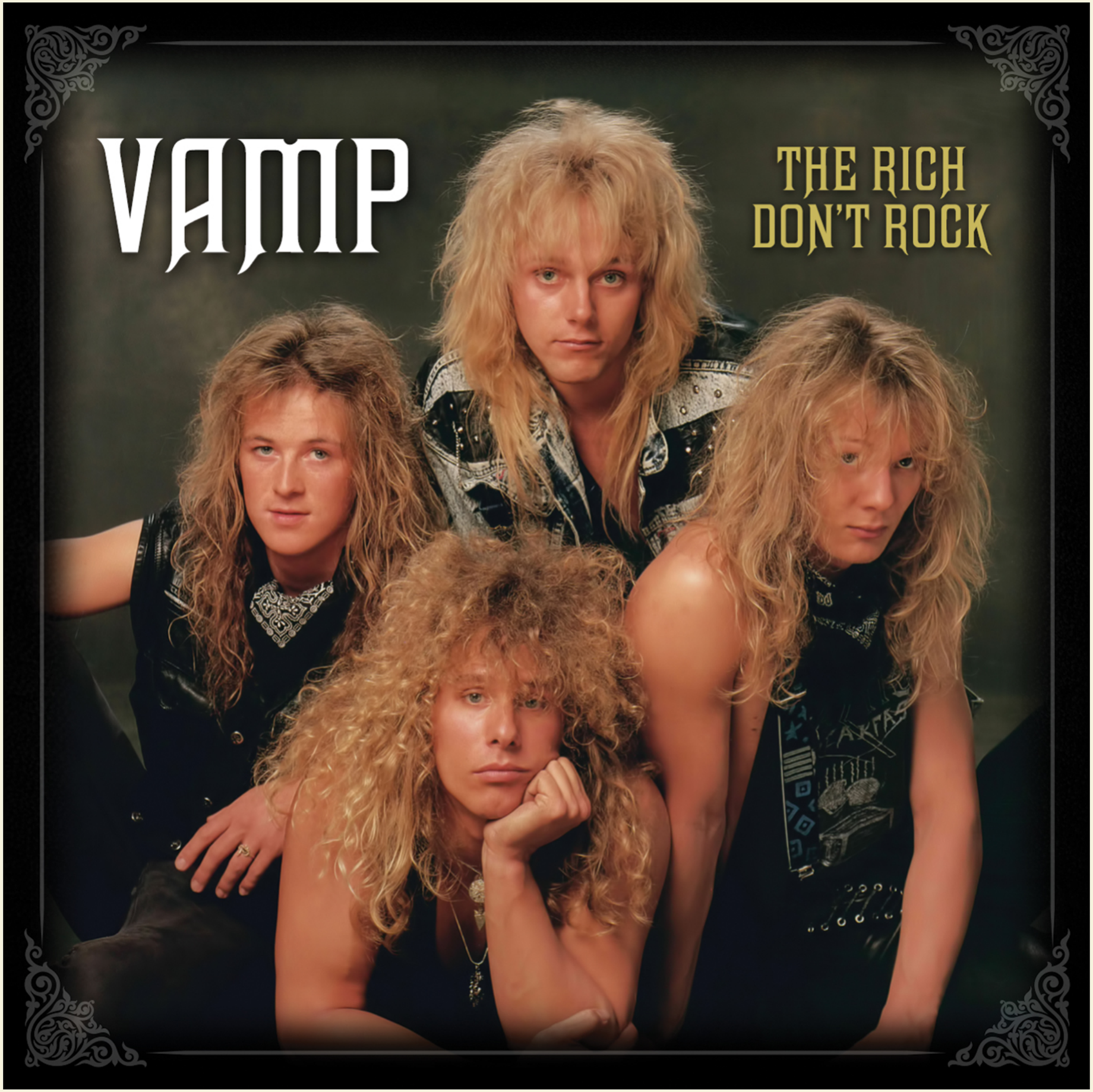


Image credit: Divebomb Records

Version-controlled

Actionable metadata

Modular

Pportable

EXHAUSTIVE TRACKING OF RESEARCH COMPONENTS

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

CAPTURE COMPUTATIONAL PROVENANCE

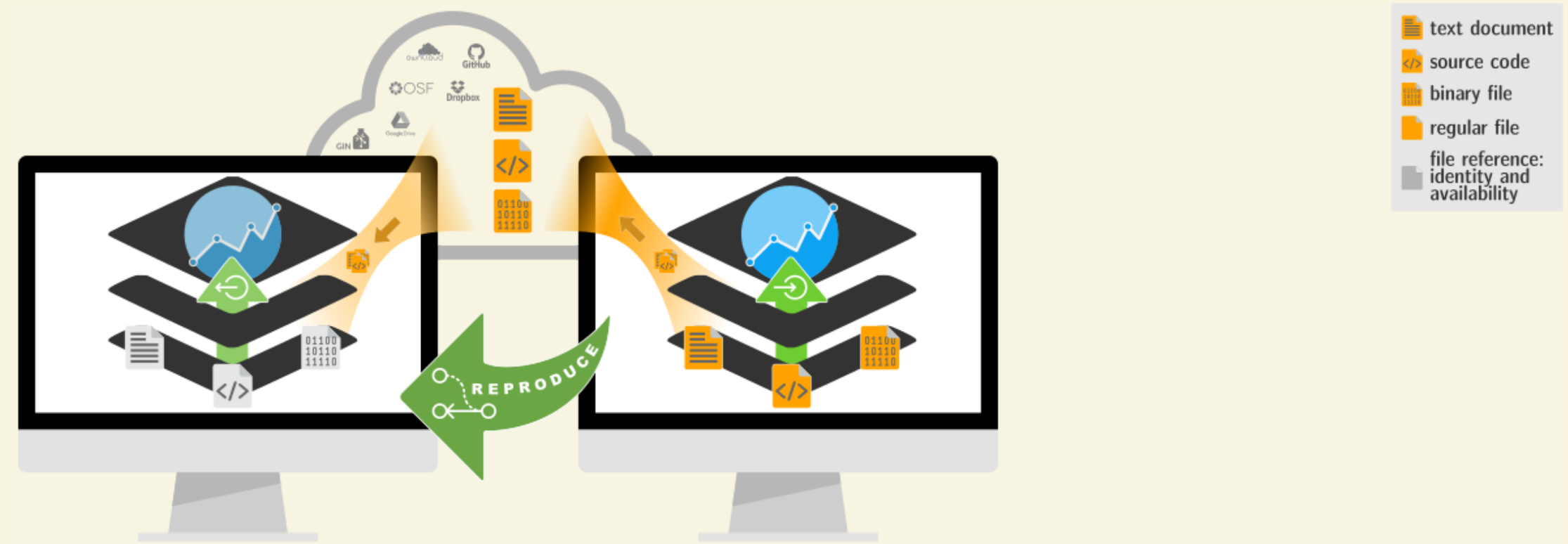
text document
source code
binary file
regular file
file reference: identity and availability

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

EXHAUSTIVE CAPTURE ENABLES PORTABILITY

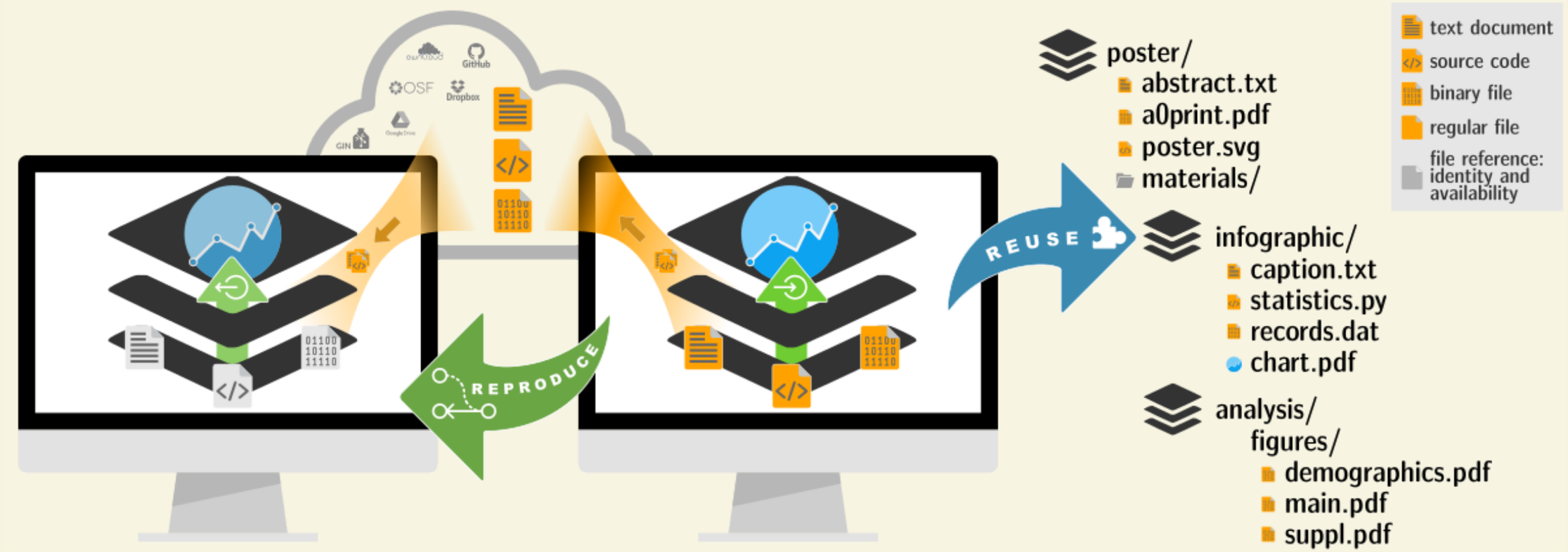
Precise identification of data and computational environments, combined for provenance records form a comprehensive and portable data structure, capturing all aspects of an investigation.

REPRODUCIBILITY STRENGTHENS TRUST



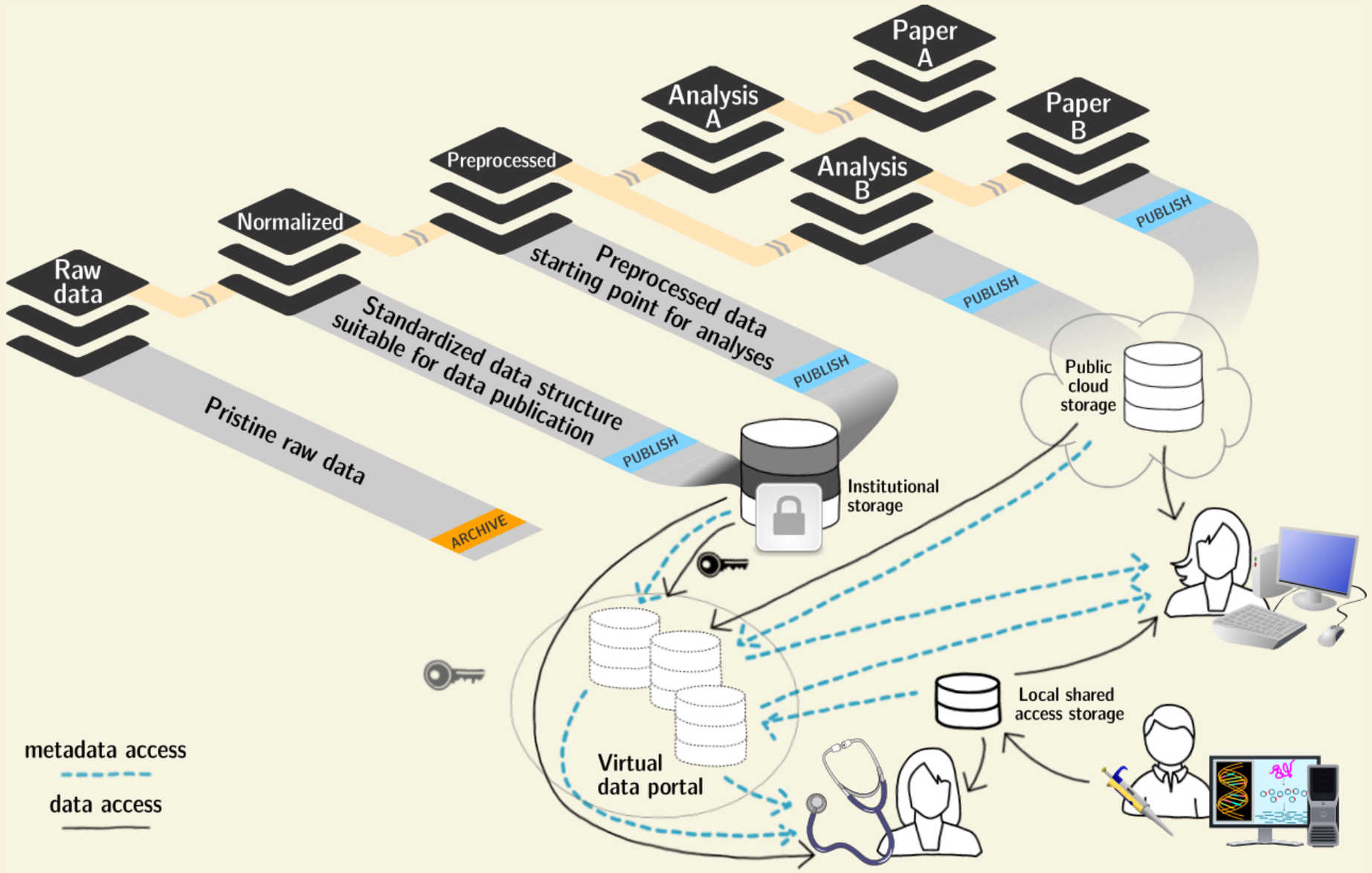
Outcomes of computational transformations can be validated by authorized 3rd-parties. This enables audits, promotes accountability, and streamlines automated "upgrades" of outputs

ULTIMATE GOAL: (RE-)USABILITY



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

DATALAD: MANAGE EVOLUTION OF DIGITAL OBJECTS



Consume, create, curate, analyze, publish, and query data with full provenance capture and "universal" metadata support.

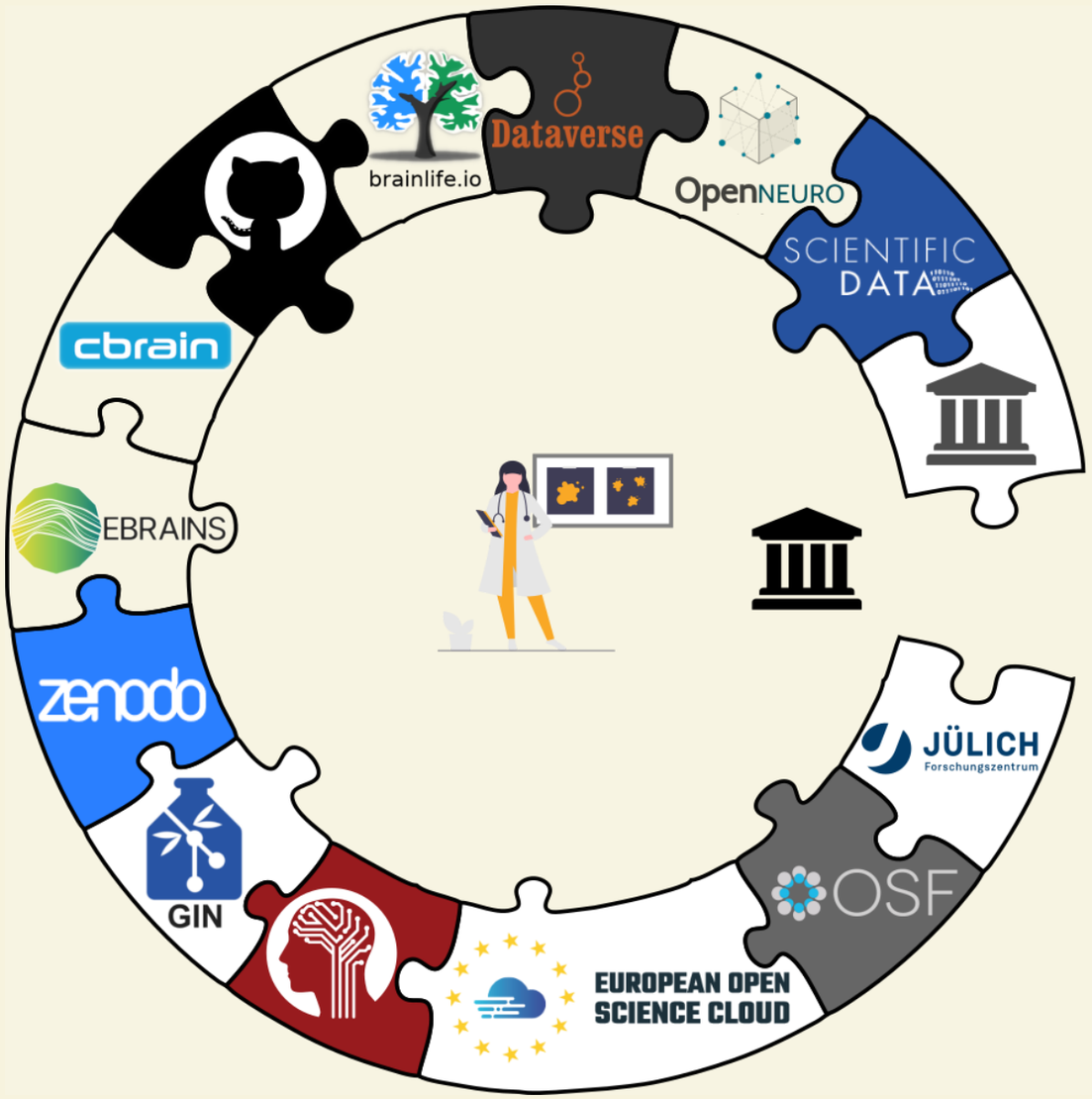
DataLad is free and open source (MIT-licensed).

http://datalad.org

ADVANTAGES OF THE VAMP ATTITUDE

- *Overlay data structure* hides away peculiarities of (current) environment choices for storage and computation
  — focus on content, not infrastructure
- Self-contained units that are *valid and complete without any external services*
  — federation-ready for improved resilience
- Metadata plurality puts *focus on metadata validity* (for your own work) without becoming a problem for global standardization efforts
  — ability to verify detailed metadata is more useful than today's choice of terminology and minimal description standard
- Promotes *long-term curation and stewardship* for flexibly reusable unit
  — yields proven and trusted resources for incremental science

"AUTOMATIC" INTEROPERABILITY WITH 3RD-PARTY SOLUTIONS



- Technology directly used by OpenNeuro, CBRAIN platform, BrainLife.io, and compatible with AWS/S3, GIN, Dropbox, etc. (optional strong encryption)
- 100+TB of research data, homogenously accessible regardless of hosting choices (datasets.datalad.org)

REPRODUCIBLE PAPER - A MAGIC TRICK?

Behavior Research Methods
https://doi.org/10.3758/s13428-020-01428-x

Published online: 24 July 2020

REMoDNaV: robust eye-movement classification for dynamic stimulation

Asim H. Dar[1] · Adina S. Wagner[2] · Michael Hanke[2,3]

Springer

- See for yourself: https://youtu.be/_I3JFhJJtW0?t=861
- Get step-by-step instructions: http://handbook.datalad.org/usecases/reproducible-paper.html

OPEN SCIENCE EDUCATION

http://handbook.datalad.org

- **educational materials** on technologies — **targeting scientists**, not developers (executable paper, student surpervisor workflow, ...)
- handbook with 400+ pages on concepts, workflows, and use cases (work in progress, led by Adina Wagner)

ACKNOWLEDGEMENTS

- Yaroslav Halchenko
- Benjamin Poldrack
- Kyle Meyer
- Adina Wagner
- 30+ Datalad contributors

EUROPEAN UNION
European Regional Development Fund

SACHSEN-ANHALT
Ministerium für Wissenschaft und Wirtschaft

Human Brain Project

CONP PCNO

VirtualBrainCloud

OpenNEURO

cbrain

brainlife.io

Website: http://datalad.org
Documentation: http://handbook.datalad.org
Open data: http://datasets.datalad.org