DATA SHARING FOR BETTER SCIENCE AND BETTER HEALTH:

THE DATAVERSE PROJECT

Mercè Crosas, Institute for Quantitative Social Science, Harvard University

@mercecrosas

XXXVII JORNADAS DE ECONOMÍA DE LA SALUD, BARCELONA, 6-8 SEPTIEMBRE

THIS TALK

- **Importance of Data Sharing**
  - **Reproducibility** to verify science
  - **Reuse** to advance science and evidence-based policy
- **Enabling Data Sharing**
  - **Data Policies** from journals and funding agencies
  - **Data Citation** to find datasets, give credit to data authors
  - **Data Repositories** as publishers of data

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Print edition | Leaders ›

May 6th 2017

DATA SCIENCE, BIG DATA

"Every two years, the amount of digitized data is equal to all of the data ever collected before. The world's knowledge is at our fingertips, and **data science** allows us to effectively and efficiently make use of that knowledge. This is facilitating a societal shift as big as the Industrial Revolution. "

UVAToday Q&A, August 21, 2017

Phil Bourne

Data Science Director, UVA

Former Associate Director for Data Science, NIH

DATA SHARING, DATA PUBLISHING

Data sharing is "the release of research data, associated metadata, accompanying documentation, and software code for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way."

Data Publishing Group, 201 5

Since the Beginning of Modern Science ...

NULLIUS IN VERBA

"TAKE NOBODY'S WORD FOR IT"

(motto of the Royal Society, founded in 1660, launched first scientific journal in 1665)

AUG 22 2017

LEAVE A COMMENT

BY JOHN BORGHI

BEST PRACTICES

WHAT WE TALK ABOUT WHEN WE TALK ABOUT REPRODUCIBILITY

Campbell's Soup Cans (1962) by Andy Warhol. Created by replicating an existing object and then reproducing the process at least 32 times.

University of California Curation Center, DataPub blog, August 2017

REPRODUCIBILITY

REPRODUCIBILITY AND REPLICATION (BY THE NATIONAL SCIENCE FOUNDATION):

The ability of a researcher to duplicate the results of a prior study

... using the same materials and procedures used by the original investigator. **(reproducibility)**

... if the same procedures are followed but new data are collected. **(replication)**

EMPIRICAL, COMPUTATIONAL, AND STATISTICAL REPRODUCIBILITY (STODDEN, 2014):

Empirical: data and collection details are made freely available

Computational: code, software, hardware and implementations details are provided

Statistical: details on choice of statistics tests, model parameters are provided

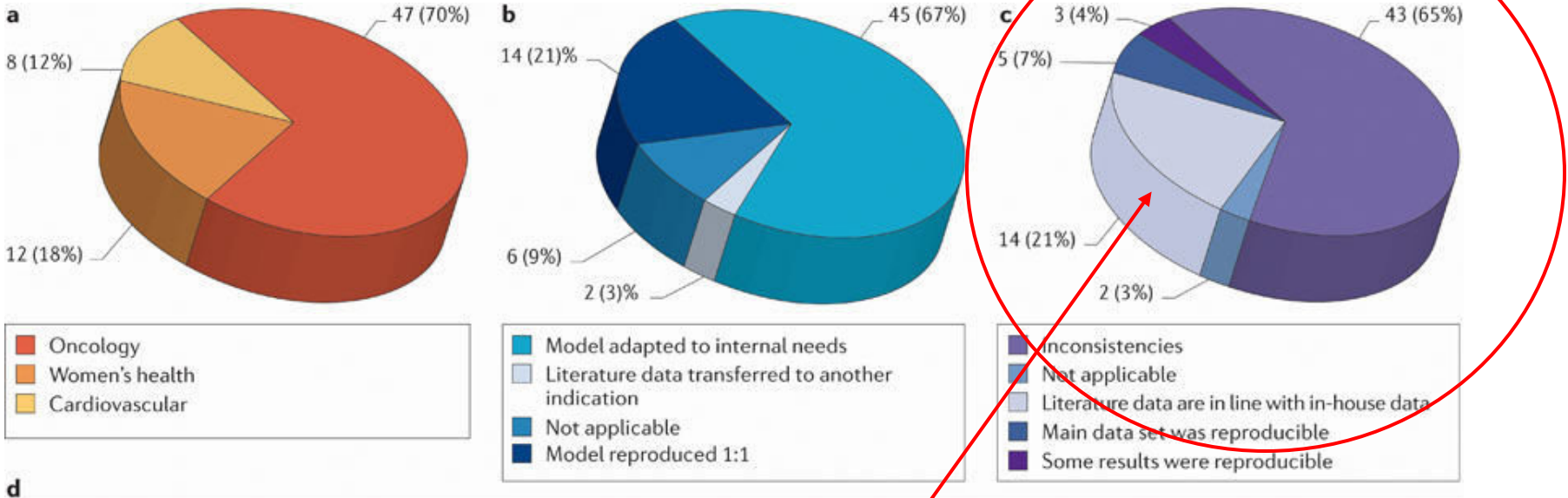

REPRODUCIBILITY IN
CANCER STUDIES

ONLY 6 (11%) OUT OF 53 LANDMARK STUDIES THAT CLAIM TO TREAT CANCER COULD BE REPRODUCED

Begley & Ellis, Nature, 2012

(from Scientists at Amgen biotechnology)

STUDY FROM BAYER: ONLY 14(21%) OF 67 PUBLICATIONS COULD BE REPRODUCED



d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

21% of Literature Data are in line with in-house data

Nature Reviews | Drug Discovery

Prinz, Schlange, Asadullah, 2011, Nature Reviews Drug Discovery

Reproducibility Project: Cancer Biology (RP:CB) Overview

The Reproducibility Project: Cancer Biology (RP:CB) is an initiative to conduct direct replications of 50 high-impact cancer biology studies. The project anticipates learning more about predictors of reproducibility, common obstacles to conducting replications, and how the current scientific incentive structure affects research practices by estimating the rate of reproducibility in a sample of published cancer biology literature. The RP:CB is a collaborative effort between the Center for Open Science and Science Exchange and will be published in eLife.

Through independent direct replication studies, the project aims to identify best practices that maximize reproducibility and facilitate an accurate accumulation of knowledge, enabling potentially impactful novel findings to be effectively built upon by the scientific community.

Additionally we expect to learn about:

- The overall rate of reproducibility in a sample of the published cancer biology literature.
- Obstacles that arise in conducting direct replications of original studies.
- The feasibility and practical challenges of getting proper materials, methods, and instrumentation for a replication.
- Predictors of replication success such as the journal in which the original finding was published, the citation impact of the original report, the number of direct replications that have been published elsewhere, the transparency of materials and methods included with the publication, and adherence to publishing checklists and guidelines.

Independently replicating a subset of experimental results from 50 high-profile papers in the field of cancer biology published between 2010-2012

Science EXCHANGE

eLIFE

COS CENTER FOR OPEN SCIENCE

Aims to release results by end of 2017

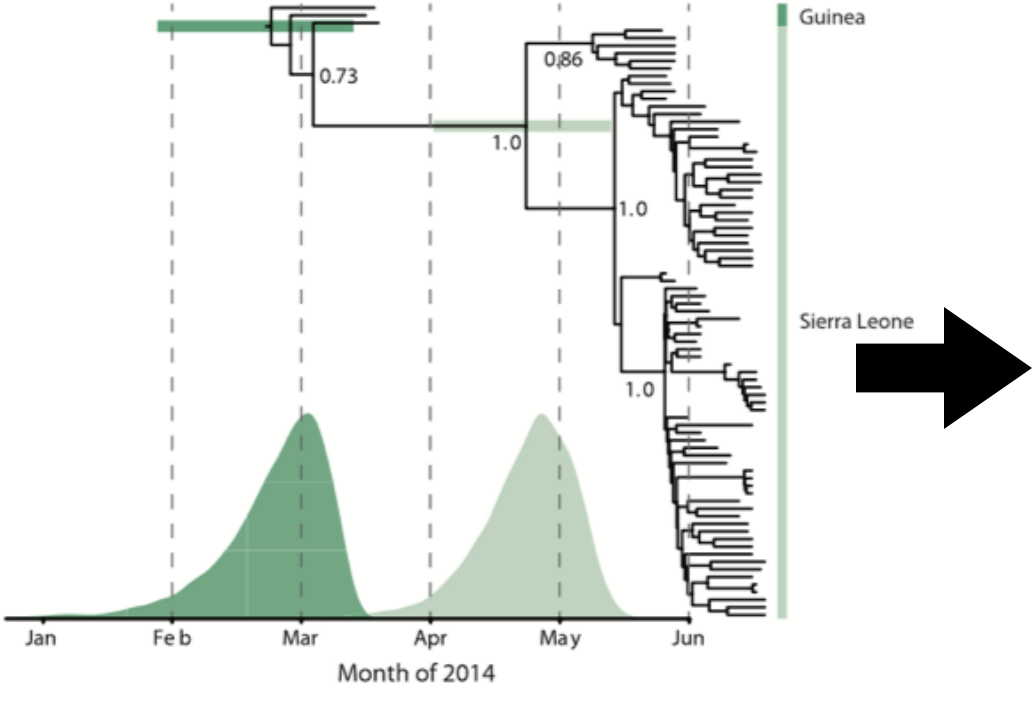
DATA SHARING TO ADVANCE SCIENCE AND POLICY MAKING

- **Outbreak Data**
- **City Data**


AB
SABETILAB
Home
Research
Outreach
BROAD INSTITUTE
ELCOME TO THE SABETI LA
INFECTIOUS DISEASE RESEARCH

SABETI LAB SHARED EBOLA DATA DURING OUTBREAK

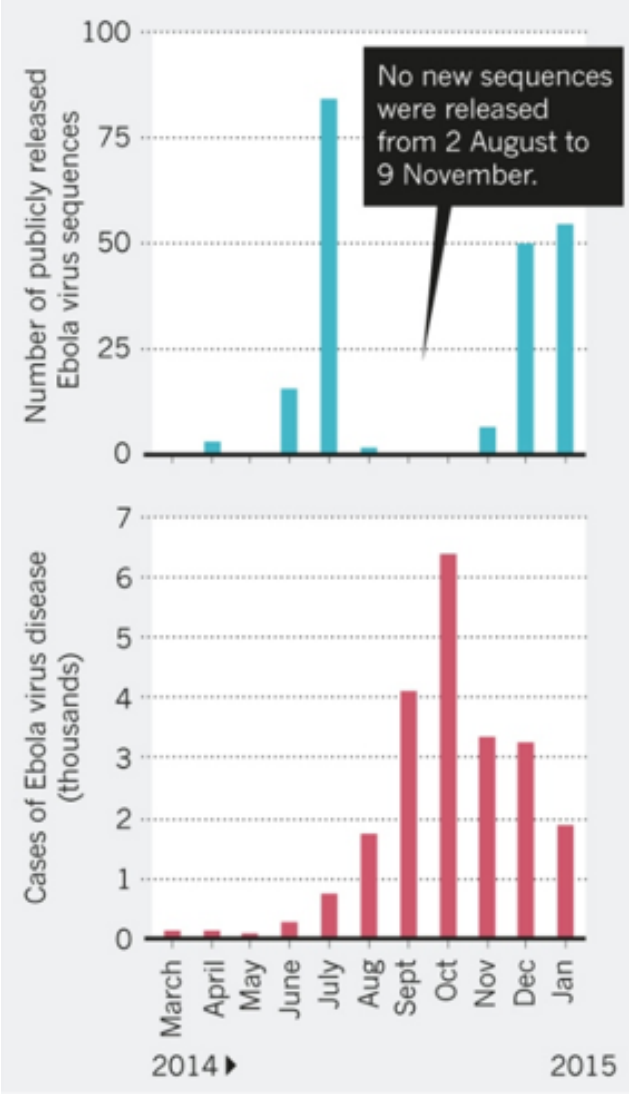
Lab released the first publicly available Ebola sequences (on GenBank), and clinical data (on Harvard Dataverse).



"We were amazed by the surge of collaboration that followed"

Yozwiak, Shaffner, Sabeti, 2015 "Make Outbreak Research Open Access" Nature

BUT, WE DON'T SHARE DATA OFTEN ENOUGH



Gaps in data sharing during the pike of the Ebola outbreak

WHY DATA SHARING DOESN'T HAPPEN MORE OFTEN DURING OUTBREAKS?

ONE REASON IS CONCERN ABOUT PATIENT PRIVACY

Image source: Andres Colubri, Sabeti Lab

Boston Area Research Initiative

Search

Area Research Initiative

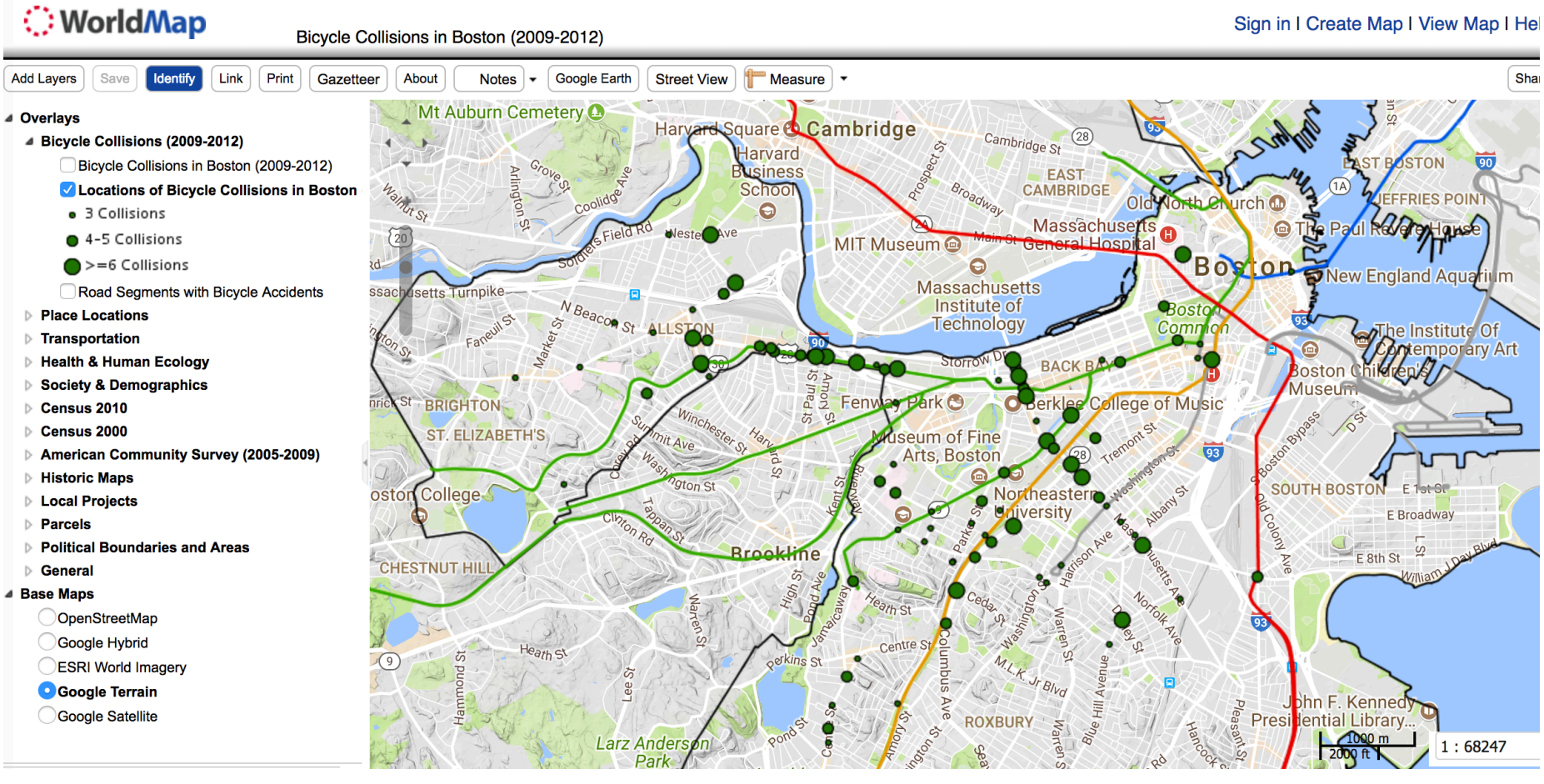
EM

About | News & Events | BARI's Spring Conference | Projects | Boston Data Portal | Grants & Fellowships | BARI Comm

Community

Read more »

BICYCLE ACCIDENT DATA PUBLICLY RELEASED: COLLISIONS IN BOSTON, 2009-2012



Data published at Harvard Dataverse

Boston

Cyclist Safety Report

2013

City Of Boston
Thomas M. Menino, Mayor

Bicycle data released by BARI was the centerpiece of Boston Mayor's Bike Safety Report

Ethical. Trustworthy. Real.

With crash data, Boston tries to make bicycling safer

Ridership up; 50% drop in injury rate is the goal

23

By Martine Powers | GLOBE STAFF MAY 15, 2013

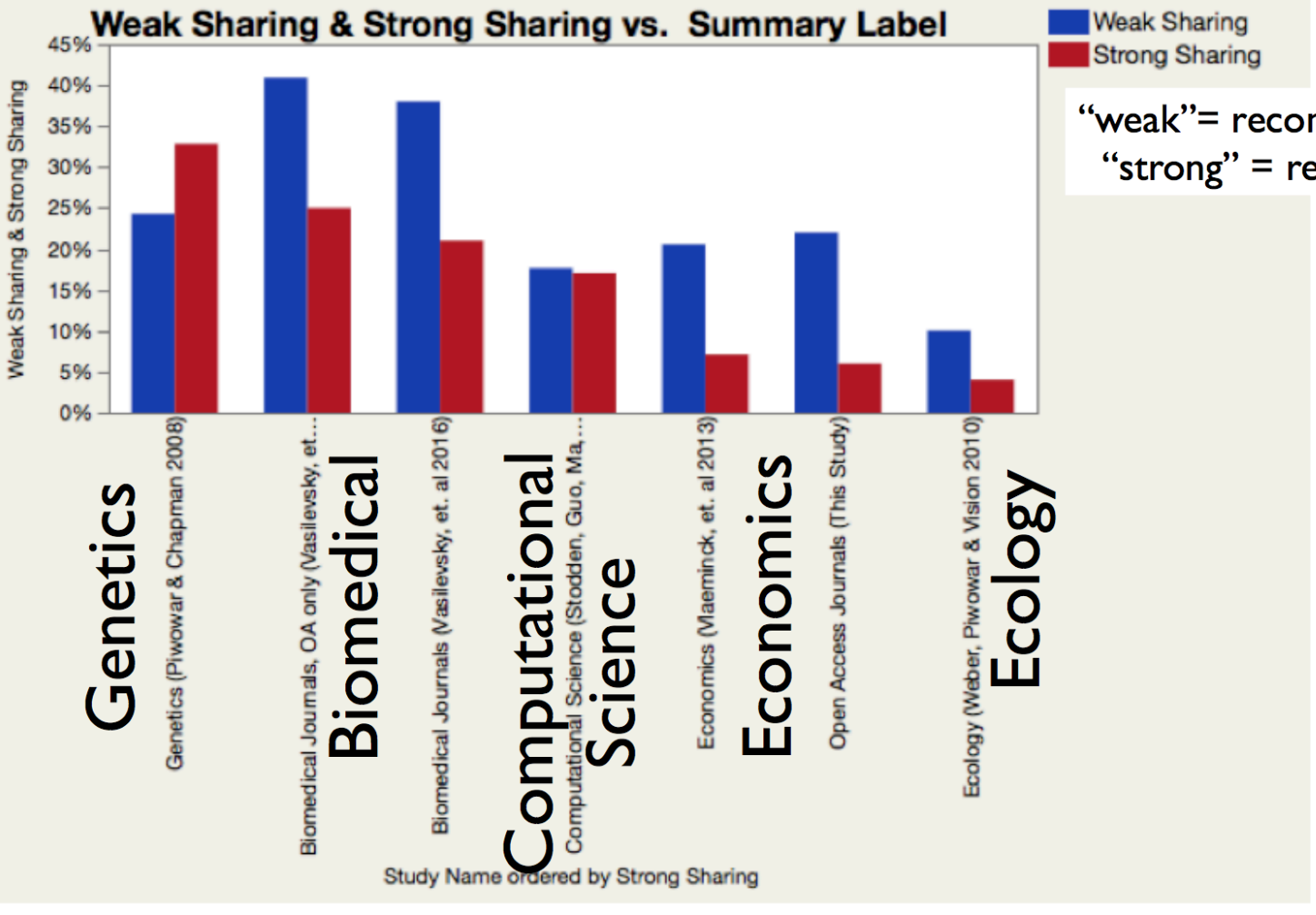
Bike crashes in Boston are most likely to occur on Massachusetts Avenue, between 4 and 7 p.m., to men between 19 and 31 years old. And they most often occur when a driver fails to see a cyclists or opens a door on a bike.

Those insights are contained in a report of unprecedented scale that compiles years of data on bike collisions in the city. With a slew of

HOW CAN WE INCREASE DATA SHARING?

- New Norms
- New Incentives
- New Technology

JOURNAL DATA POLICIES APPLIED ACROSS DISCIPLINES



"weak"= recommend
"strong" = require

Castro, Crosas, Garnett, Sheridan, Altman, 2017, Journal of Scholarly Publishing

MANY FUNDERS REQUIRE DATA SHARING & OPEN DATA

PRIVATE RESEARCH FUNDERS

- Bill and Melinda Gates Foundation Information Sharing Approach
- Sloan Foundation Data Sharing Policy
- Wellcome Trust Data Sharing Policy
- Arnold Foundation
- Moore Foundation
- Robert Wood Johnson Foundation
- HHMI Policy on the Sharing of Publication-Related Materials, Data and Software

PUBLIC RESEARCH FUNDERS

- Department of Agriculture
- Department of Commerce
- Department of Defense
- Department of Education
- Department of Energy
- Department of Health and Human Services
  - Agency for Healthcare Research and Quality (AHRQ)
  - Assistant Secretary for Preparedness and Response (ASPR)
  - Center for Disease Control and Prevention (CDC)
  - Food and Drug Administration (FDA)
  - National Institutes of Health (NIH)
- Department of Homeland Security
- Department of Housing and Urban Development
- Department of Interior
- Department of Labor
- Department of Transportation
- Department of Veterans Affairs
- Environmental Protection Agency (EPA)

Are reproducibility and open science starting to matter in tenure and promotion review?

July 14th, 2017, Brian Nosek

Tags: openscience, reproducibility

The NEW ENGLAND JOURNAL of MEDICINE

SOUNDING BOARD

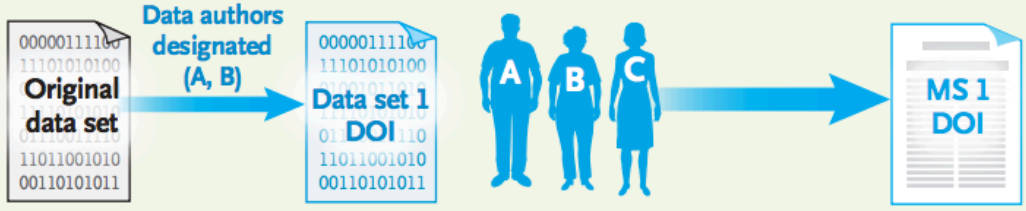
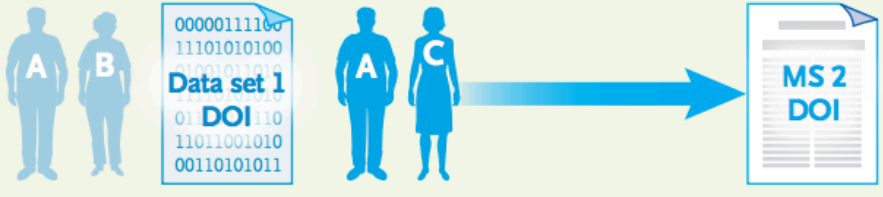
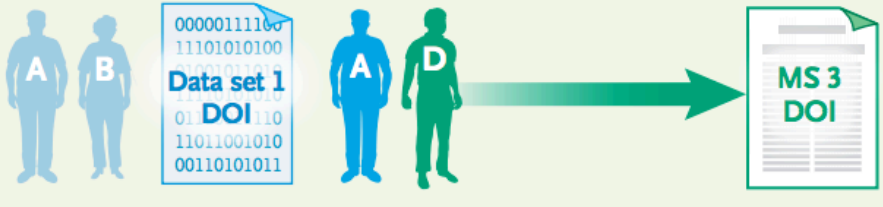
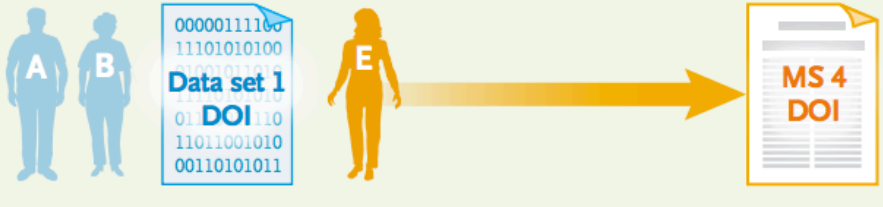
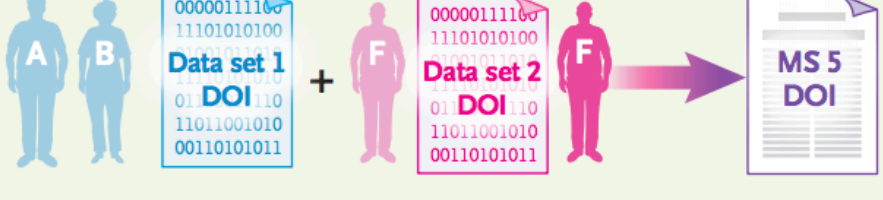
Data Authorship as an Incentive to Data Sharing

Barbara E. Bierer, M.D., Mercè Crosas, Ph.D., and Heather H. Pierce, J.D., M.P.H.

N Engl J Med 2017; 376:1684-1687 | April 27, 2017 | DOI: 10.1056/NEJMsb1616595

Data from well-designed and well-executed research not only are useful for the original purpose and applications, including independent replication, avoidance of duplicative studies, generation or [illegible] those who originally gathered the data, termed "data sharing," is currently encouraged or mandated [illegible] leadership through the draft data-sharing policy of the International Committee of Medical Journal Editors, charitable foundations such as the Robert Wood Johnson and the Bill and Melinda Gates Foundation, and the National Institutes of Health (NIH) in its recent request for information on data management and sharing strategies. Data sharing, whether elective or required, creates an obligation for the original investigators who obtain funding, design studies, collect and analyze data,

" We believe that both as a matter of fairness and as a matter of providing an incentive for data sharing, the persons who initially gathered the data should receive appropriate and standardized credit that can be used for academic advancement, for grant applications, and in broader situations."

1	**Primary authors generate data set and designate "data authors," DOI is assigned, and primary publication occurs**		3 Authors **(A, B, C)** 2 Data authors: **Data set 1 DOI (A, B)** 1 Manuscript DOI: **MS 1 DOI**
2	**Secondary analysis by members of primary group**		2 Authors **(A, C)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 2 DOI**
3	**Secondary analysis with collaborator**		2 Authors **(A, D)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 3 DOI**
4	**Secondary analysis by independent investigator**		1 Author **(E)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 4 DOI**
5	**Combination of data set with other data for new analysis by independent investigator**		1 Author **(F)** 3 Data authors: **Data set 1 DOI (A, B)** **Data set 2 DOI (F)** 1 New Manuscript DOI: **MS 5 DOI**

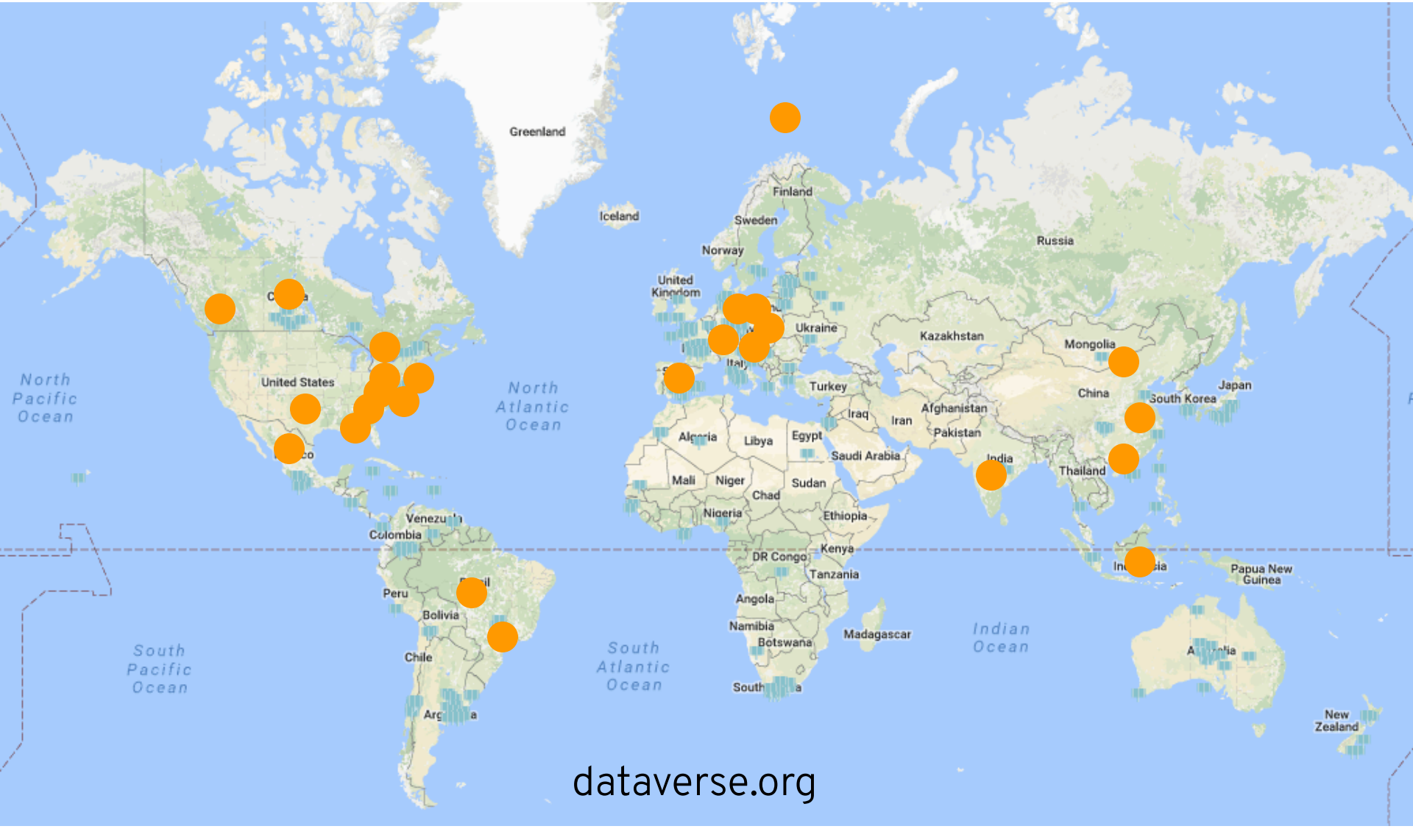

OUR INSTITUTE PROVIDES A TECHNOLOGY SOLUTION TO DATA SHARING

An open-source software to share, cite, and find data.

Developed at Harvard's Institute for Quantitative Social Science

2006 (we started)

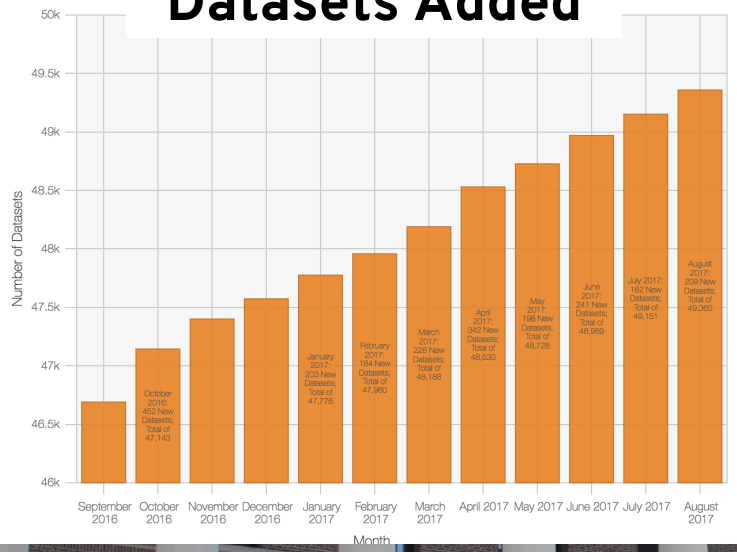
2017



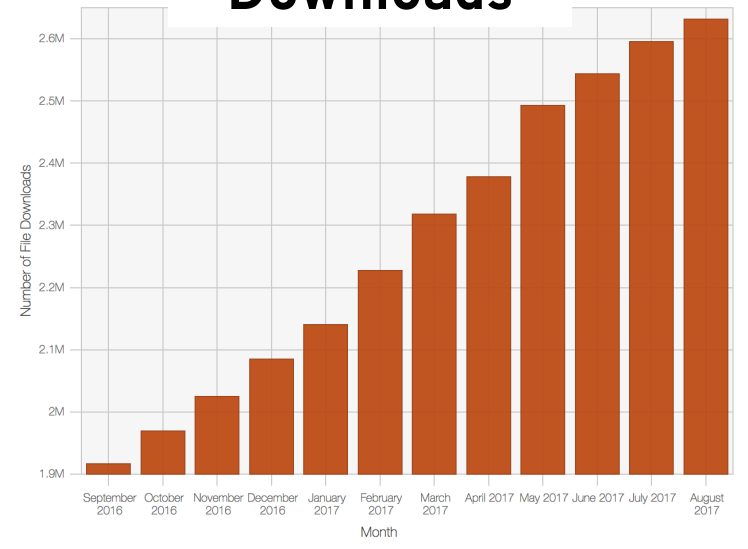
dataverse.org

HOW RESEARCHERS SHARE & USE DATA WITH DATAVERSE

Datasets Added



Downloads



Harvard Dataverse Repository

> 70,000 datasets total
> 49,000 datasets uploaded to Harvard Dataverse repository
200 datasets/month

> 340,000 files
4,000 files/month

> 2.5 M downloads
60,000 downloads/month

dataverse.harvard.edu

OUR CONTRIBUTIONS TO ENHANCE DATA SHARING

King, 1995, Replication, Replication

Altman et al, 2001, A Digital Library for the Dissemination and Replication of Quantitative Social Science

Altman and King, 2007, A Proposed Standard for the Scholarly Citation of Quantitative Data

King, 2007, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Crosas, 2012, The Dataverse Network: an open source application for sharing, discovering, and preserving research data

Crosas, 2013, A Data Sharing Story

Altman and Crosas, 2013, The Evolution to Data Citation: from principles to implementation

2014, Joint Declaration of Data Citation Principles

Pepe et al, 2014, How Do Astronomers Share Data?

Goodman et al, 2014, Ten Simple Rules for the Care and Feeding of Scientific Data

Crosas, Honaker, King, Sweeney, 2015, Automating Open Science for Big Data

Castro et al, 2015, Achieving Human and Machine Accessibility of Cited Data

Sweeney, Crosas, Bar-Sinai, 2015, Sharing Sensitive Data with Confidence: The DataTags System

Meyer et al. 2016, Data Publication with the Structural Biology Data Grid Supports Live Analysis

Wilkinson et al, 2016, The FAIR Guiding Principles for Scientific Data Management and Stewardship

Bierer, Crosas, Pierce, 2017, Data Authorship as an Incentive to Data Sharing

2017

Data should be ...

FINDABLE

ACCESSIBLE

INTERPOPERABLE

REUSABLE

Wilkinson et al. , 2016, "The FAIR Guiding Principles for Scientific Data Management and Stewardship"

Nature Scientific Data

FAIR DATA IN DATAVERSE

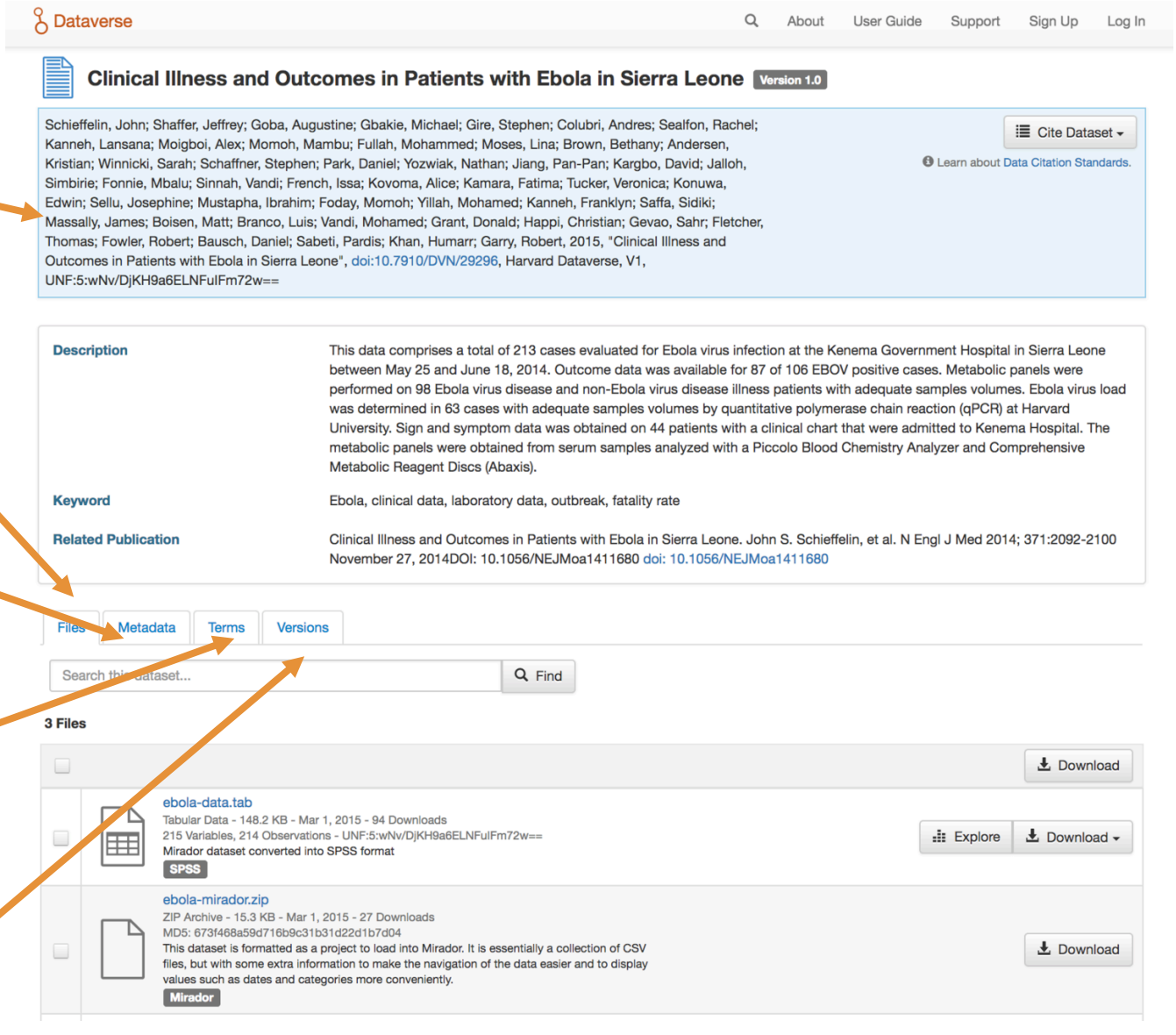
Data Citation with Persistent Identifier (DOI)

Data Files

Metadata

Data Licenses, User Agreements

Dataset Versions



WHAT ARE WE WORKING ON NOW?

DATA PRIVACY

CLASSIFY AND HANDLE DATASETS BASED ON THEIR PRIVACY LEVEL

Dataverse® as a DataTags repository

Data file deposit

Assistance to assign DataTag from:

- DataTags automated interview
- RobotLawyer auto-generated data user agreements (DUA)
- Review Board

orange **Direct Access**

Requires:

- User registration
- Approval needed for access
- Signed DUA

green **Privacy Preserving Access**

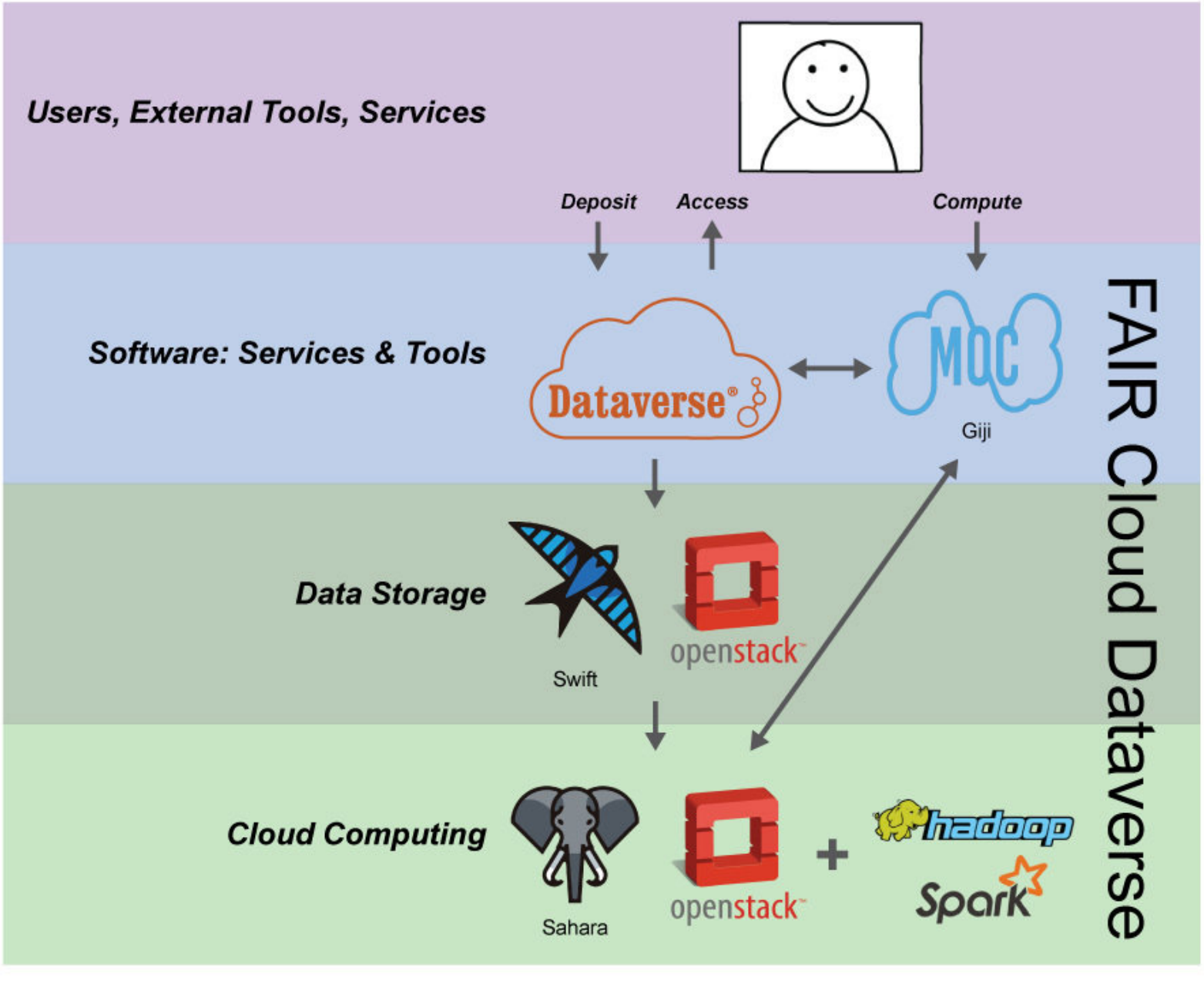
- Requires user registration
- Provides access to differentially private statistics using Private data Sharing Interface (PSI)

Harvard Data Privacy Tools Project: privacytools.seas.harvard.edu

DataTags Project: datatags.org

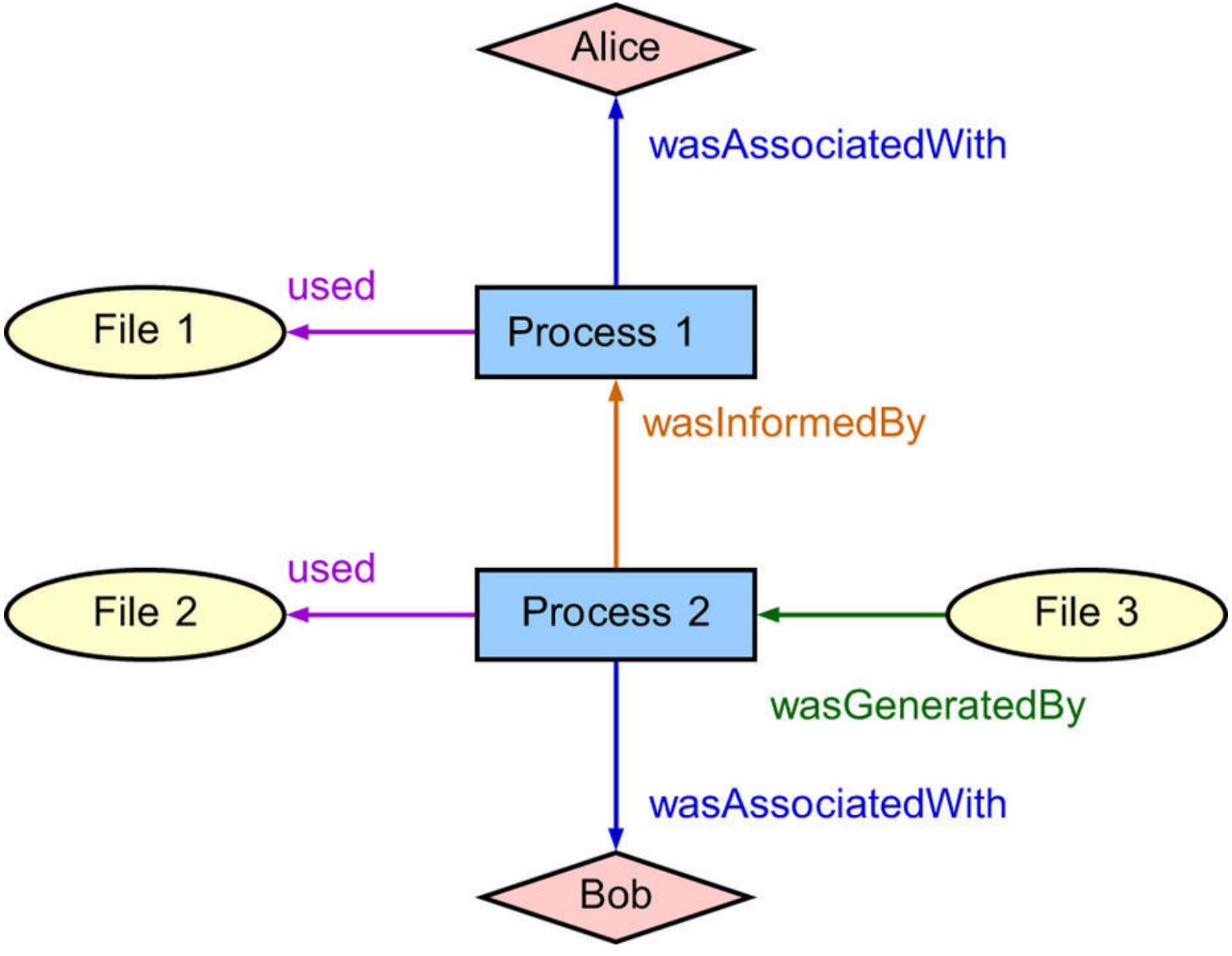
CLOUD DATAVERSE

COMBINE DATA REPOSITORIES WITH CLOUD COMPUTING


Users, External Tools, Services
Deposit
Access
Compute
Software: Services & Tools
Dataverse
MOC
Giji
Data Storage
Swift
openstack
Cloud Computing
Sahara
openstack
+
hadoop
Spark
FAIR Cloud Dataverse

DATA PROVENANCE

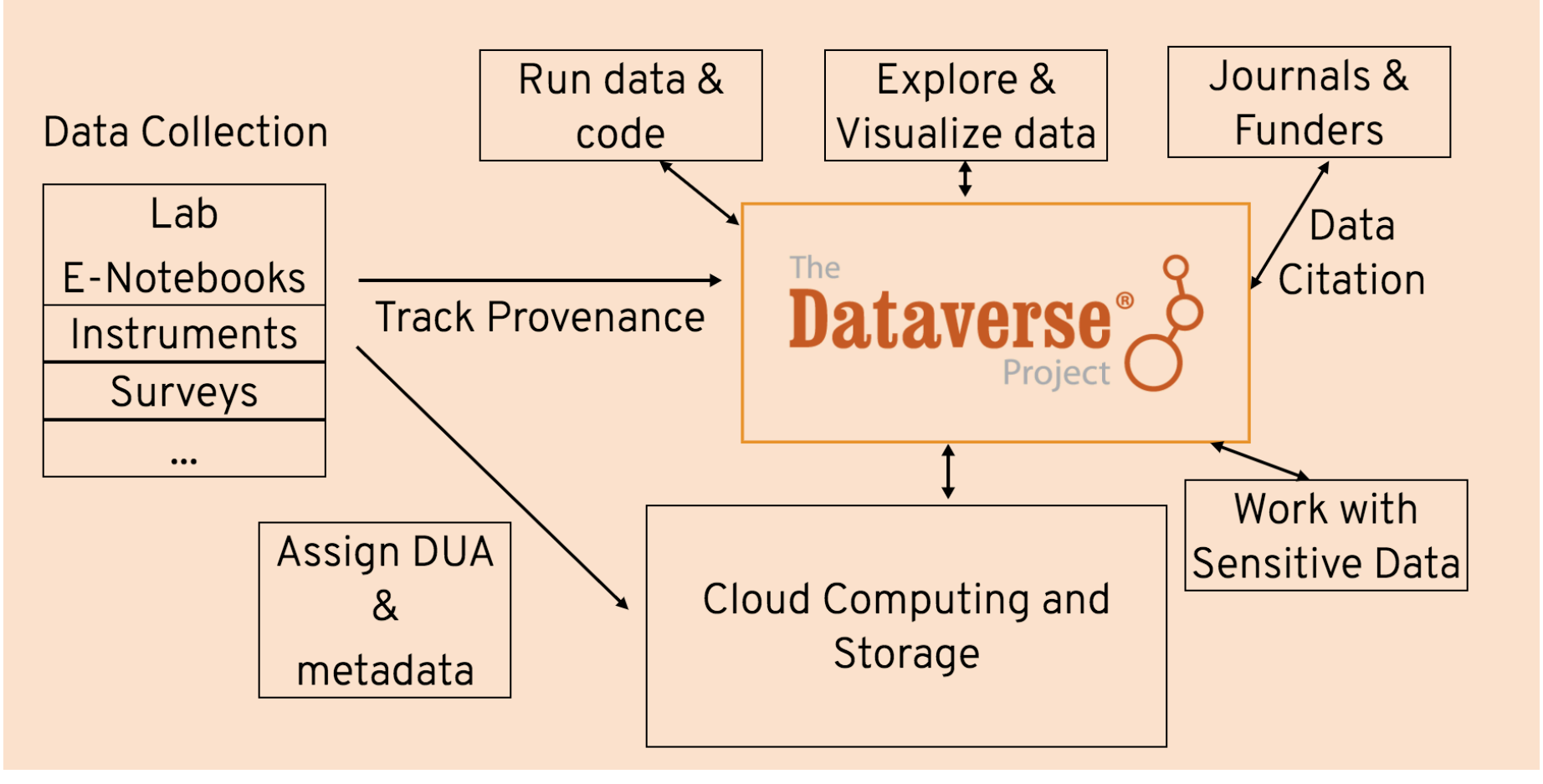
TRACK THE ORIGINAL SOURCE OF A DATASET



Pasquier, Lau, Trisovic, Boose, Coutierer, Crosas, Ellison, GIbson, Jones, Seltzer, 2017, *If These Data Could Talk*, Nature Scientific Data

INTEGRATION WITH TOOLS

DATAVERSE AS PART OF THE DATA LIFECYCLE


Data Collection
Lab
E-Notebooks
Instruments
Surveys
...
Run data & code
Explore & Visualize data
Journals & Funders
Data Citation
Track Provenance
The Dataverse® Project
Assign DUA & metadata
Cloud Computing and Storage
Work with Sensitive Data


DATAVERSE
COMMUNITY

49 SOFTWARE CONTRIBUTORS

BI-WEEKLY COMMUNITY CALLS

235 ATTENDEES
26 ORGANIZATIONS/UNIVERSITIES
11 COUNTRIES

ANNUAL COMMUNITY MEETING

NEXT: JUNE 13, 14, 15, 2018

THANKS

@mercecrosas

scholar.harvard.edu/mercecrosas

dataverse.org