Presentation supported by US National Science Foundation (NSF) award #1759041
BDV PPP Summit, June 26 -28, Riga
Impact Empowered by data-driven artificial intelligence

Facilitating data access and reuse

with Dataverse and DataTags

Mercè Crosas, Ph.D., Harvard University

Chief Data Science and Technology Officer, IQSS

Harvard University Research Data Management Officer, OVPR

@mercecrosas. scholar.harvard.edu/mercecrosas

Access to data is critical for building AI solutions

Access to good quality reusable data is critical for building AI solutions

Responsible access to
good quality reusable data is critical
for building AI solutions

How can we get there?

- Facilitate data sharing, giving credit and control to data authors and providers
- Enable data reuse, following FAIR principles for machine-readable findable, accessible, interoperable, and reusable data
- Standardize data policies for sharing sensitive data
- Build open source platforms and tools for transparency and community engagement

Two projects that can help

Dataverse and DataTags

Dataverse: An open source platform for sharing, citing, and archiving research data

- Created and developed at Harvard's Institute for Quantitative Social Science since 2006
- An active, growing open source community, with 94 contributors
- 46 production sites in 6 continents, 14 of them federated
- Harvard Dataverse alone contains 80,000 datasets, with 500,000 files and 7 million downloads
- In the last year, newly formed Global Dataverse Community Consortium to support the Dataverse community

Dataverse.org

46 production sites (installations) of the Dataverse open source platform host data repositories around the world

Dataverse supports the FAIR principles

Findable:

- Described with rich metadata
- Assigned globally unique identifier (DataCite DOI)

Accessible:

- Retrievable through standard protocol
- Authentication and authorization procedure, where necessary
- Metadata are accessible, even when the data are no longer available

Interoperable:

- Commonly-used formats, schemas, and ontologies

Reusable:

- Richly described, with detailed provenance
- Released with a clear and accessible data usage license
- Domain-relevant community standards

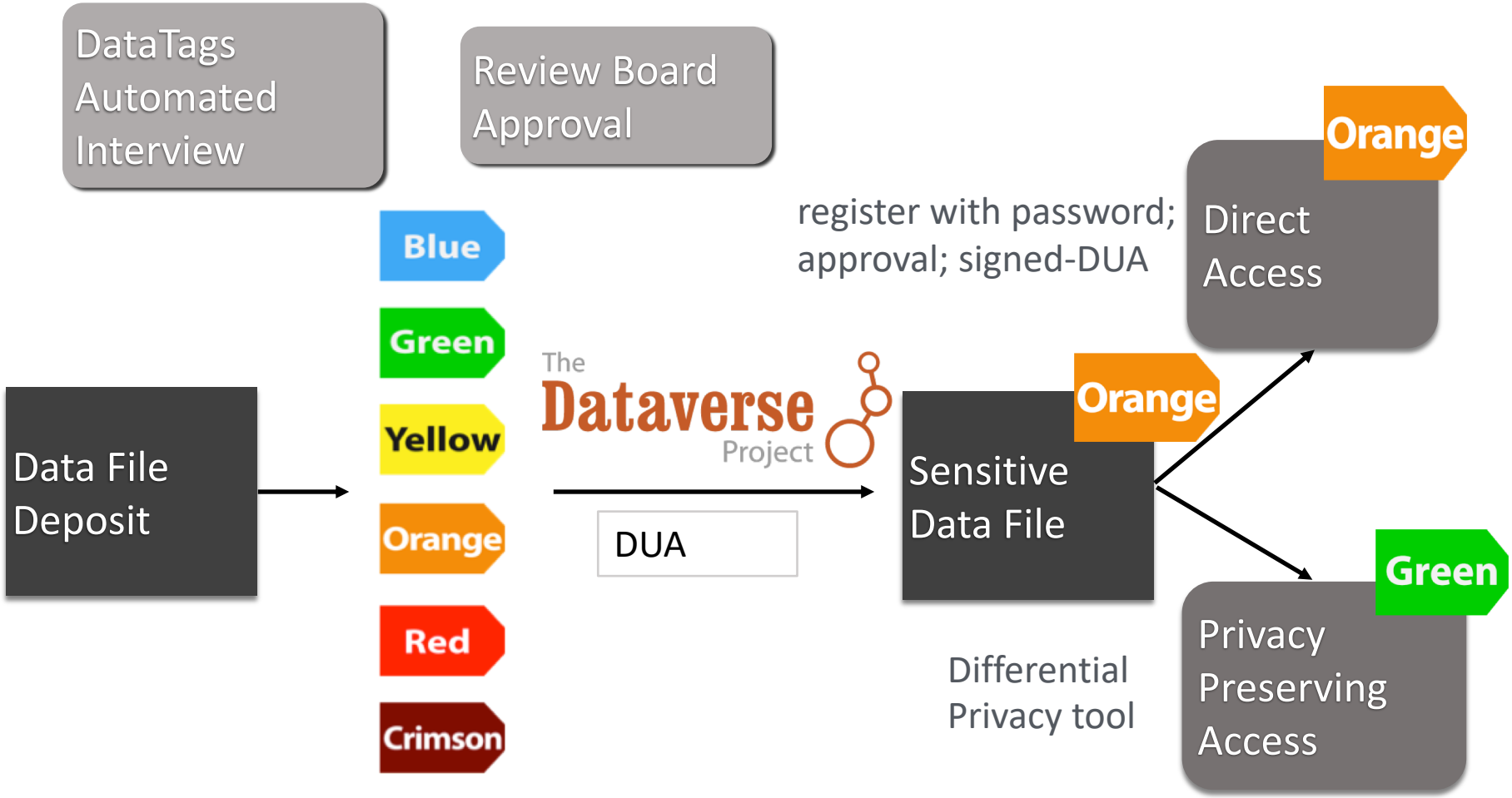
Wilkinson et al. 2016, The FAIR Guiding Principles of Scientific Data Management and Stewardship, Nature Scientific Data

DataTags: Standardized tiered access and security policies for sharing sensitive data

Blue	Public			
Green	Public	+ Register		
Yellow	Restricted	+ Approval Needed	+ Click-thru DUA	+ Encrypted transmit
Orange	Restricted	+ Approval Needed	+ Signed DUA	+ Encrypted transmit + Encrypted storage
Red	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transmit + Encrypted storage
Crimson	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transmit + Multi-encrypted storage

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

Datatags integrated with a Dataverse repository and privacy-preserving tools



Thanks

Presentation supported by US National Science Foundation (NSF) award #1759041

BDV PPP Summit, June 26 -28, Riga

Impact empowered by data-driven artificial intelligence