SHARING SENSITIVE DATA WITH CONFIDENCE:
THE DATATAGS SYSTEM

Mercè Crosas
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University

Michael Bar-Sinai
PhD candidate in Computer Science at the Ben-Gurion University of the Negev, Israel
Fellow at the Institute for Quantitative Social Science at Harvard University.

Latanya Sweeney
Professor of Government and Technology in Residence
Director of Data Privacy Lab
Harvard University

Data sharing:
good for you and good for the world

Researchers	→	Get credit for their data
Publishers and Journals	→	Verify published work
Federal funding agencies	→	Make public assets accessible
Science	→	Validate, reuse and extend previous work

The Dataverse Project

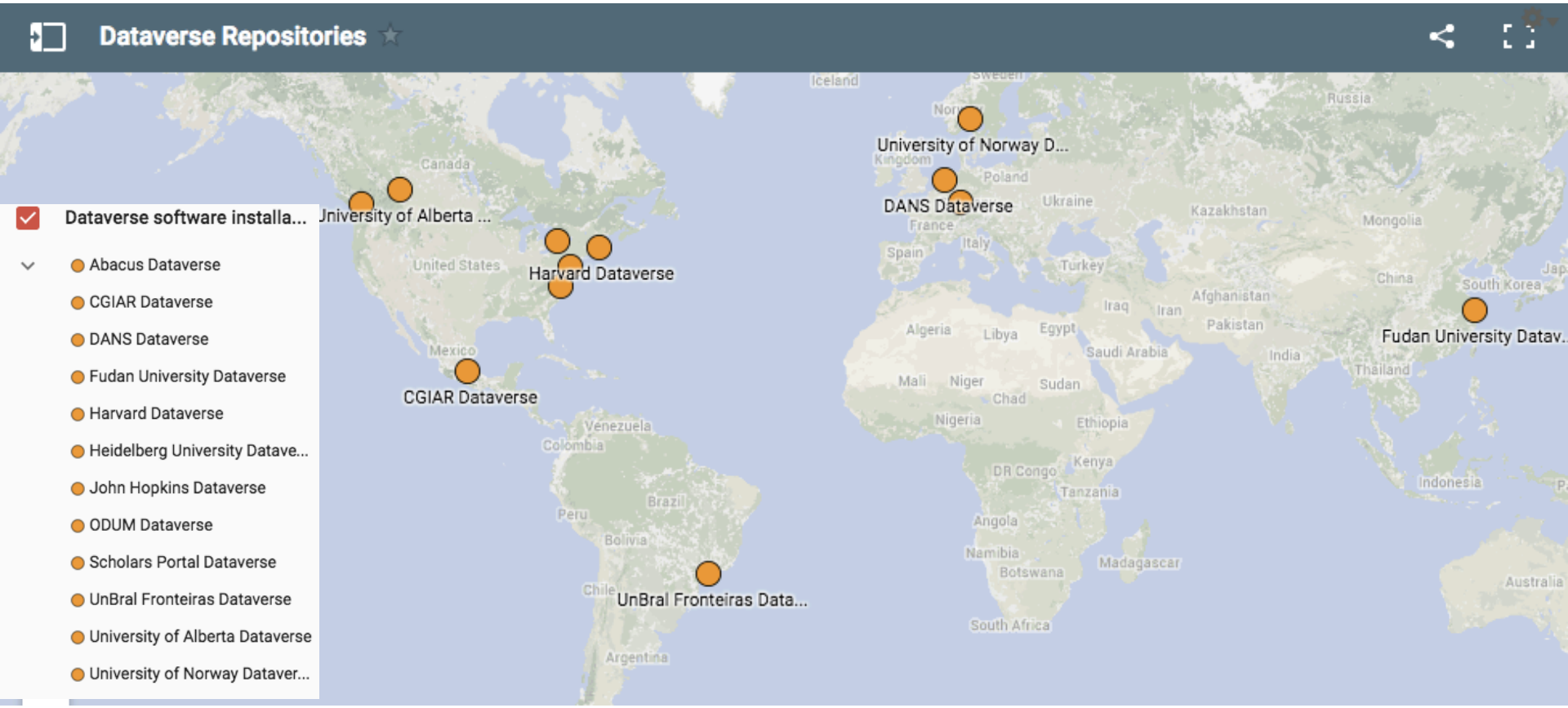
dataverse.org

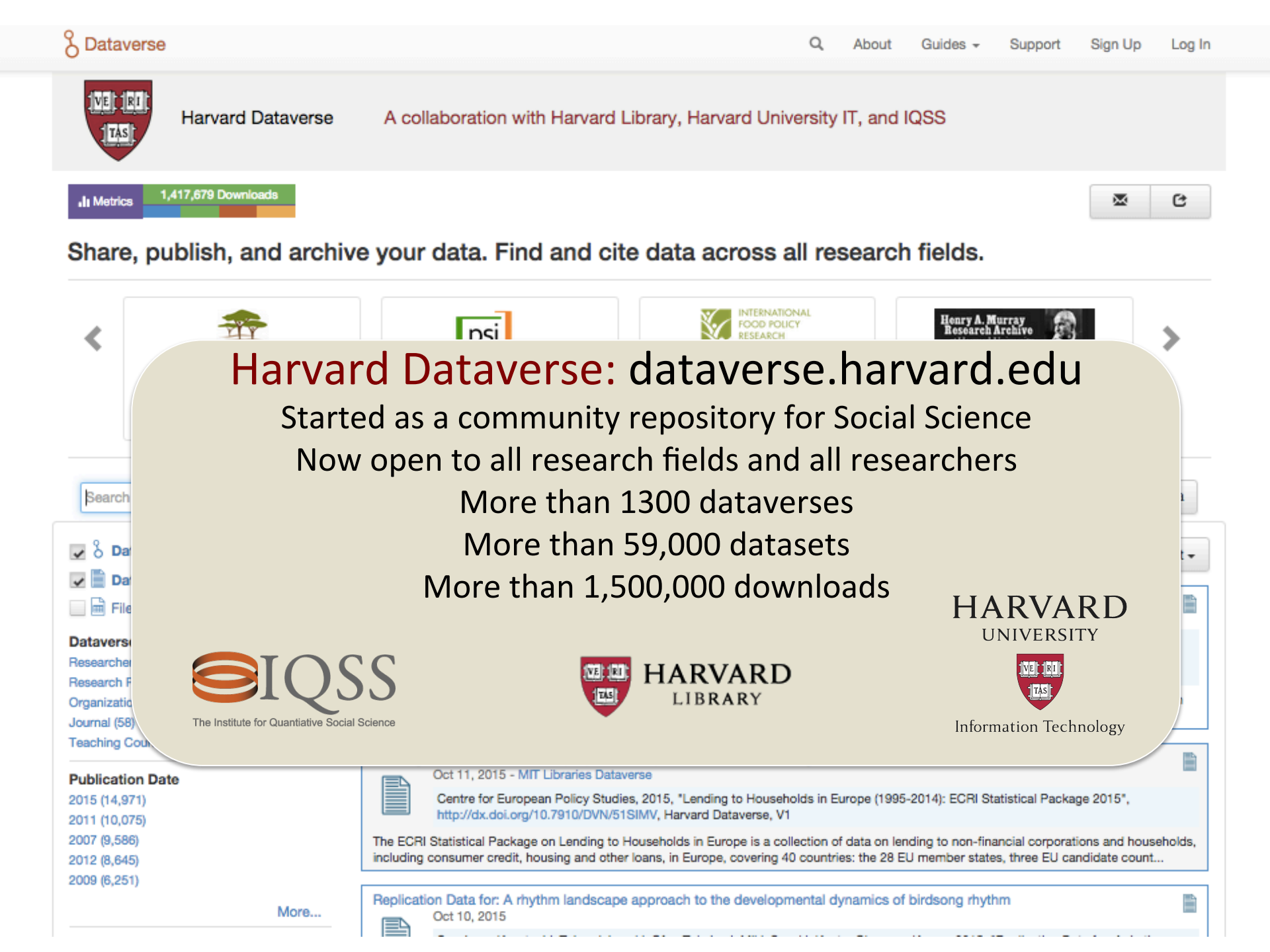
Open-source software developed at Harvard's IQSS since 2006

Used to share, publish, cite and archive research data

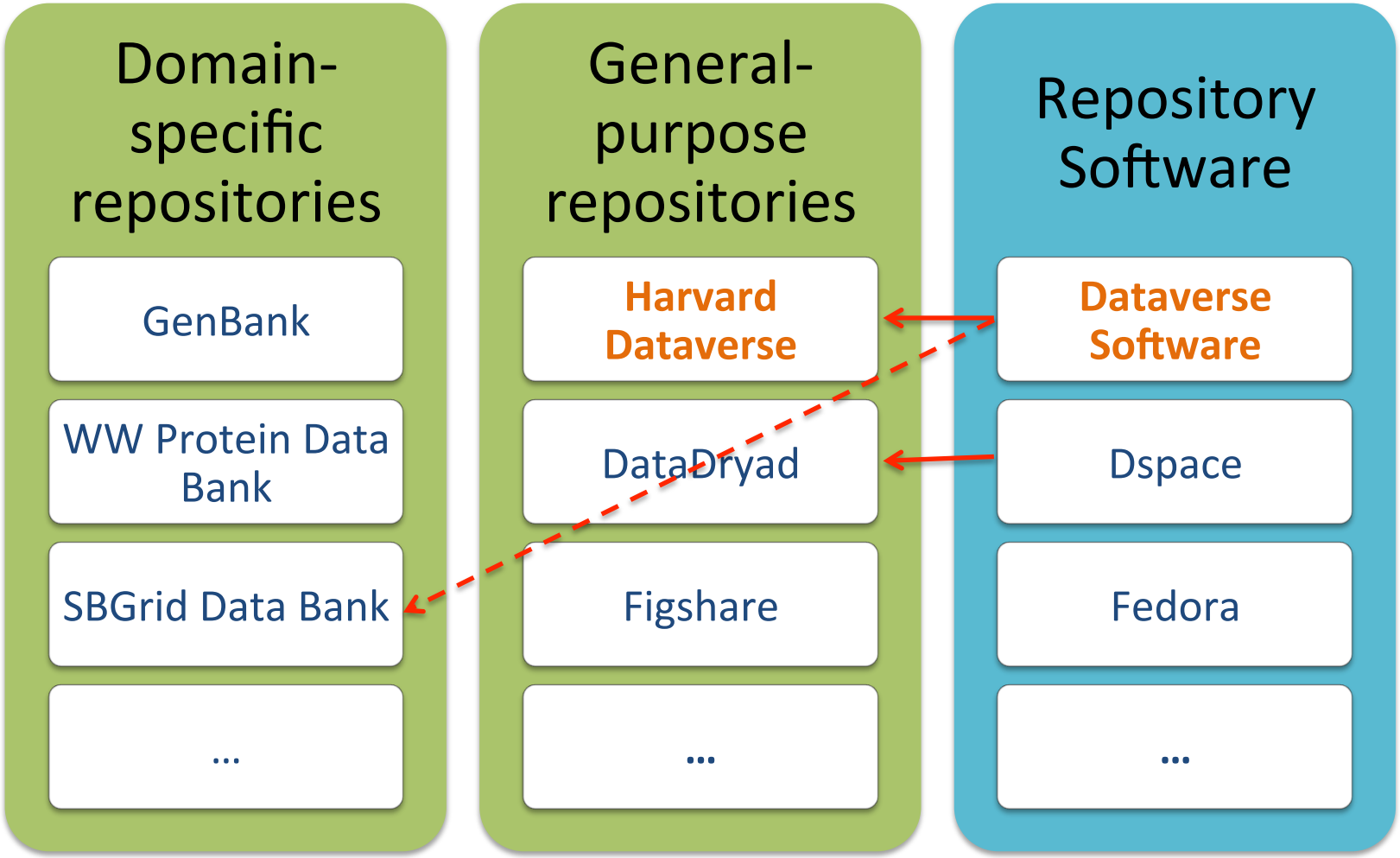
Installed in 12 production sites world wide

Serving 100s of universities and organizations

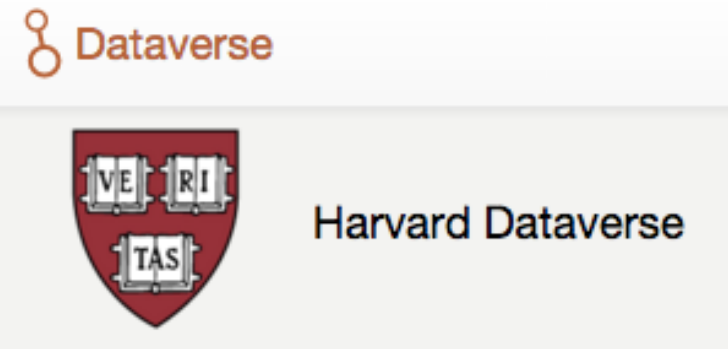



Dataverse
About
Guides
Support
Sign Up
Log In
Harvard Dataverse
A collaboration with Harvard Library, Harvard University IT, and IQSS
Metrics
1,417,679 Downloads
Share, publish, and archive your data. Find and cite data across all research fields.
Harvard Dataverse: dataverse.harvard.edu
Started as a community repository for Social Science
Now open to all research fields and all researchers
More than 1300 dataverses
More than 59,000 datasets
More than 1,500,000 downloads
IQSS
The Institute for Quantiative Social Science
HARVARD LIBRARY
HARVARD UNIVERSITY
Information Technology
Publication Date
2015 (14,971)
2011 (10,075)
2007 (9,586)
2012 (8,645)
2009 (6,251)
More...
Oct 11, 2015 - MIT Libraries Dataverse
Centre for European Policy Studies, 2015, "Lending to Households in Europe (1995-2014): ECRI Statistical Package 2015", http://dx.doi.org/10.7910/DVN/51SIMV, Harvard Dataverse, V1
The ECRI Statistical Package on Lending to Households in Europe is a collection of data on lending to non-financial corporations and households, including consumer credit, housing and other loans, in Europe, covering 40 countries: the 28 EU member states, three EU candidate count...
Replication Data for: A rhythm landscape approach to the developmental dynamics of birdsong rhythm
Oct 10, 2015

Data Repositories vs Repository Software



But, existing community repositories do **not** support sensitive data



“User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or user should not be possible.”

“Submitter represents and warrants that the Content does not contain any information (i) which identifies, or which can be used in conjunction with other publicly available information to personally identify, any individual;”

GenBank

"If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. It is our assumption that you have received any necessary informed consent authorizations that your organizations require prior to submitting your sequences."

HOW CAN WE MAXIMIZE SHARING SENSITIVE DATA WHILE BEING MINDFUL OF PRIVACY?

Technology Science
...how technology impacts humans.
Home

Tweet Published 2015-10-16. Views 3,490. Downloads 408. Suggestions 0.

Sharing Sensitive Data with Confidence: The Datatags System

Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Definitions for each of six ordered Blue to Crimson sample datatags.

- We introduce datatags as a means of specifying security and access requirements for sensitive data
- The datatags approach reduces the complexity of thousands of data-sharing regulations to a small number of tags
- We show implementation details for medical and educational data and for research and corporate repositories

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. Technology Science. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

A datatag is a set of security features and access requirements for file handling

A datatags repository is one that stores and shares data files in accordance with a standardized and ordered levels of security and access requirements

A DataTags Repository must satisfy the following conditions:

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
    a. additional requirements cannot weaken the file security
    b. and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must:
    a. satisfy file's access requirements,
    b. produce sufficient credentials as requested,
    c. and agree to any terms of use required to acquire the file.
4. Provides technological guarantees for requirements 1, 2 and 3.

Datatags Levels

Tag Type	Description	Security Features	Access Requirements
Blue	**Public**	Clear storage Clear transmission	Open
Green	**Controlled public**	Clear storage Clear transmission	Email, OAuth verified registration
Yellow	**Accountable**	Clear storage Encrypted transmit	Password, Registered , Approval, Click DUA
Orange	**More accountable**	Encrypted storage Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	**Fully accountable**	Encrypted storage Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	**Maximally restricted**	MultiEncrypt store Encrypted transmit	Two-factor authentication, Approval, Signed DUA

DATATAGS WITH HARVARD DATAVERSE

DataTags vs Harvard Security Levels

Level 1:
No sensitive data; open data

Level 1:
De-identified data

Level 2:
Confidential information by University standards; no material harm

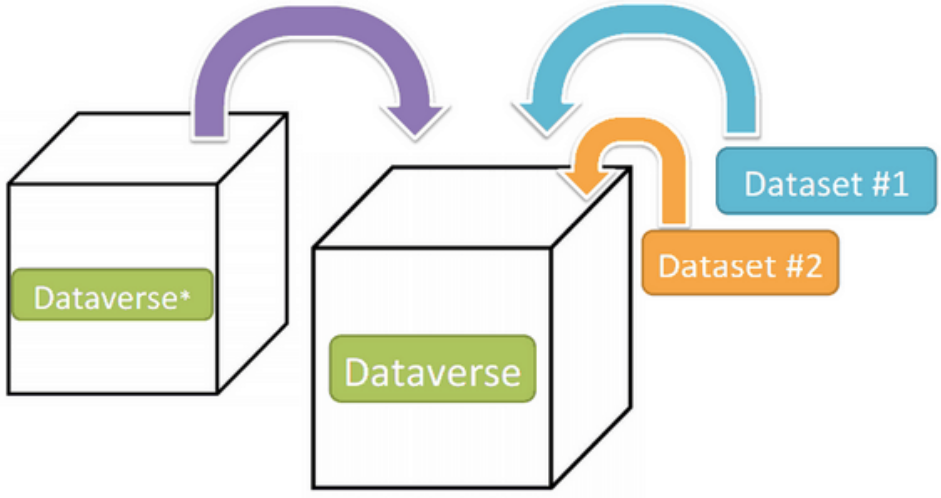
Level 3:
Confidential information that could cause material harm (non-level 4 FERPA)

Level 4:
High-risk confidential information (SSN)

Level 5* (Level 4.5, on the network)
Information that would cause severe harm

Dataverses, Datasets, Data Files and DataTags

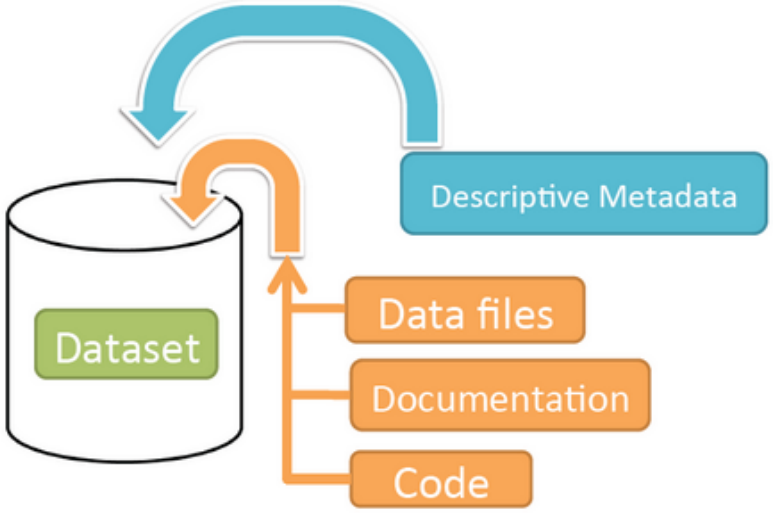
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

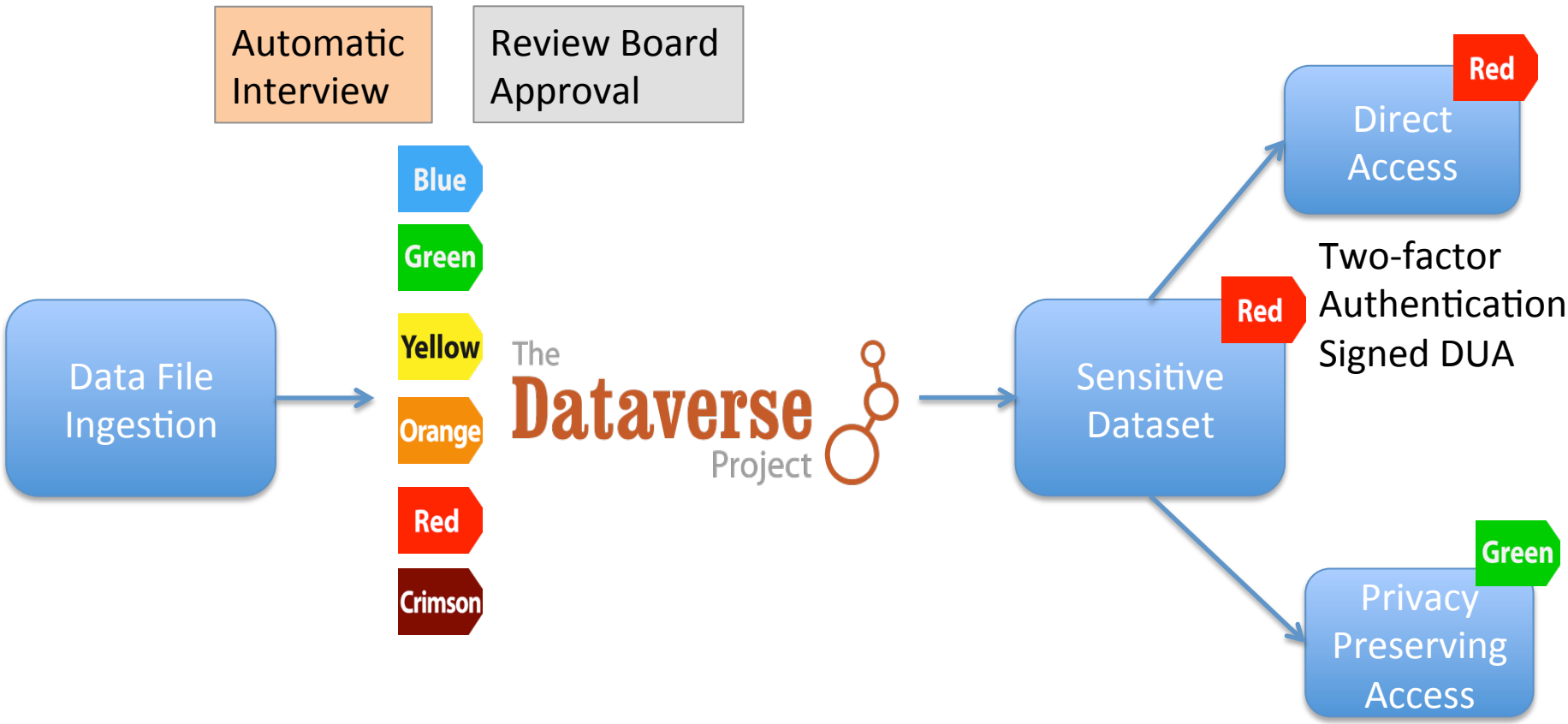
Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

A Datatag is assigned to each Data File (not to the Dataset)

DataTags Workflow with Dataverse



DATA PRIVACY LAB

IQSS
The Institute for Quantitative Social Science

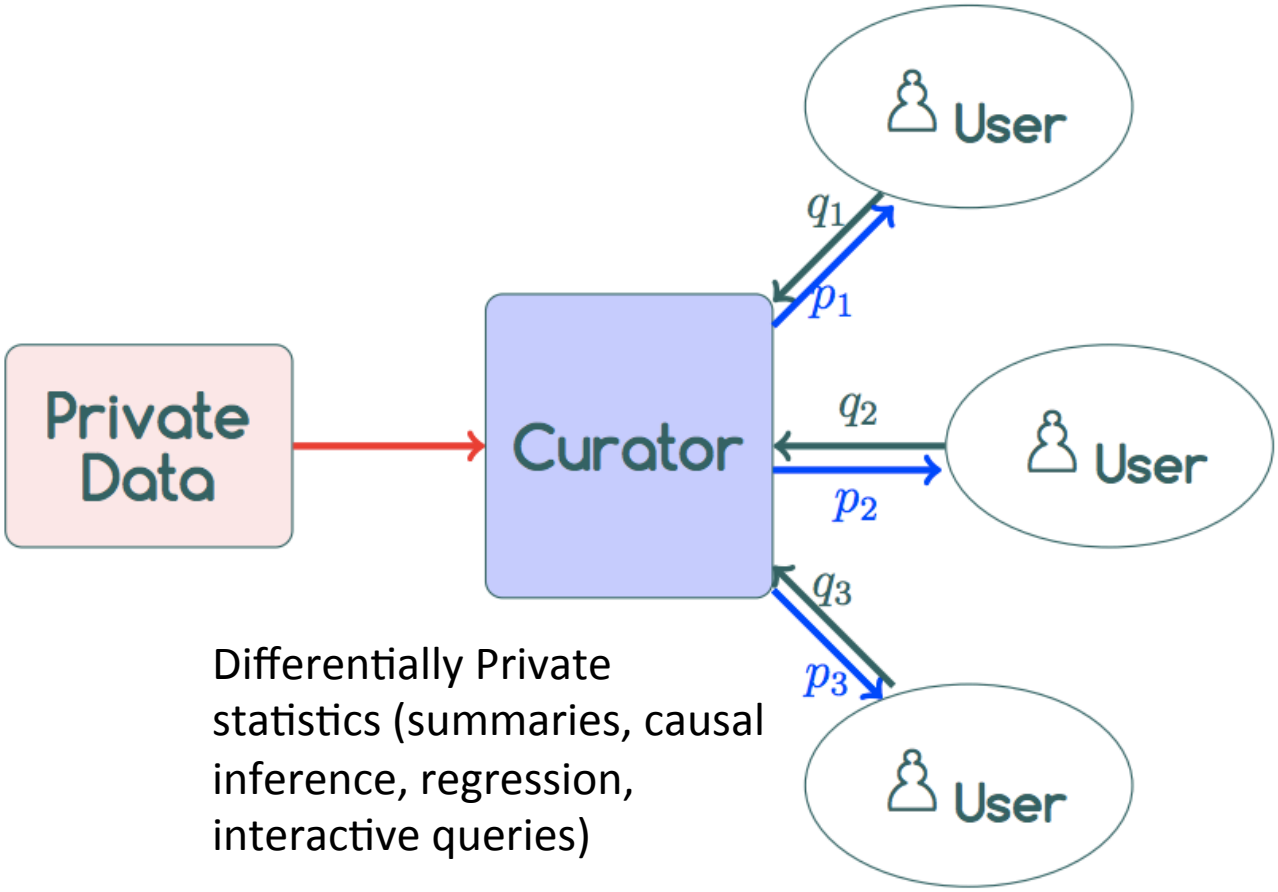
BERKMAN CENTER FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

CRCS Center for Research on Computation and Society
at Harvard School of Engineering and Applied Sciences

Program on Information Science
MIT Libraries

http://datatags.org
http://privacytools.seas.harvard.edu

A Curator Model for Privacy-Preserving Analysis



Acknowledgement: Honaker, J. and Nissim, K., Data Privacy Tools Project

Credentials and Retrieval in Dataverse

Blue

Green

Data File not restricted;
Guestbook – Email to access

Yellow

Data File restricted; Dataverse/InCommon account; Request access; Click DUA

Orange

Data File restricted; Dataverse/InCommon account; Request access; Sign DUA

Red

Data File restricted; InCommon account; Request access; Two-Factor authentication Sign DUA

Crimson

OTHER TYPE OF DATATAGS REPOSITORIES

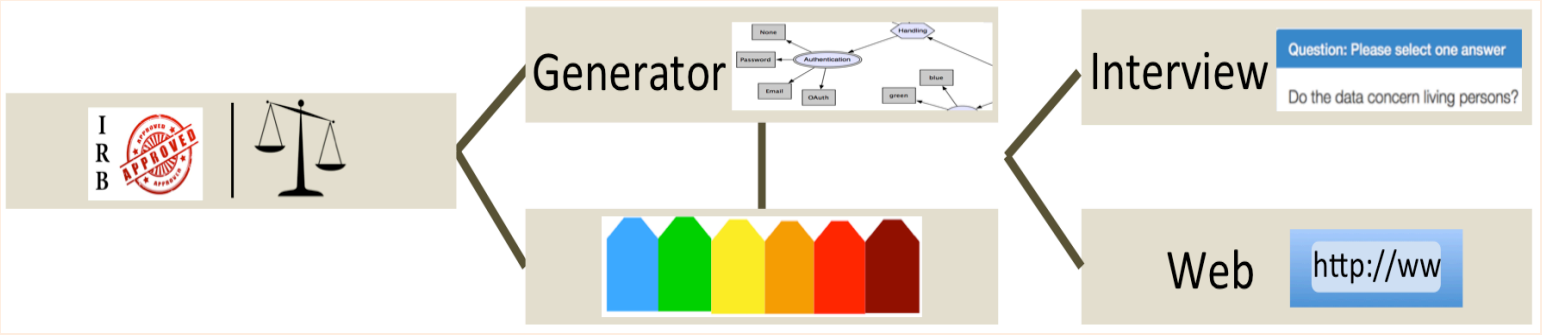
Betty: Sole Researcher

- Received consent from participants
- Repository for sharing highly sensitive data (not necessarily Harvard Dataverse)

Betty: Global Research Repository

Ingestion and Decision-making Knowledge	IRB determination or an interview system.
Codification and Infrastructure	Blue, Green, Yellow, Orange, Red, Crimson.
Credentials and Retrieval	Different files may additionally require specific terms of use based on legal or regulatory requirements or adopted best practices.

(Same use case as Dataverse)

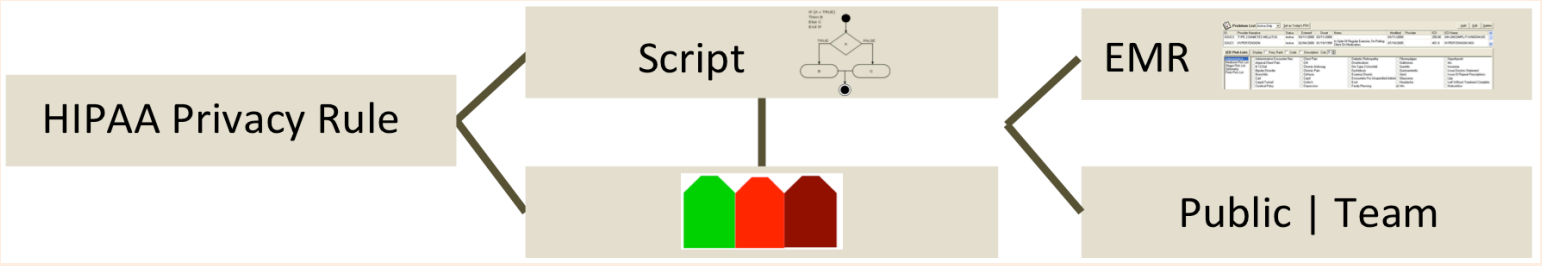


Adam: Large Medical Research Group

- Repository for sharing local data
- Repository for published data
- Repository for sharing with collaborators

Adam: Large Medical Research Group

Ingestion and Decision-making Knowledge	A HIPAA-consistent Safe Harbor script redacts data files to produce a version for sharing under the Green tag. It assigns a Crimson tag to any file if finds that contains clinical notes, psychiatric notes, or HIV-AIDS information. It assigns a Red tag to all other data files and to the original non-redacted files that are not Crimson.
Codification and Infrastructure	Green, Red and Crimson tags.
Credentials and Retrieval	Data-use agreements. Red and Crimson are limited to those who qualify based on IRB review and their data-use agreements describe handling requirements beyond the repository for downloaded files.

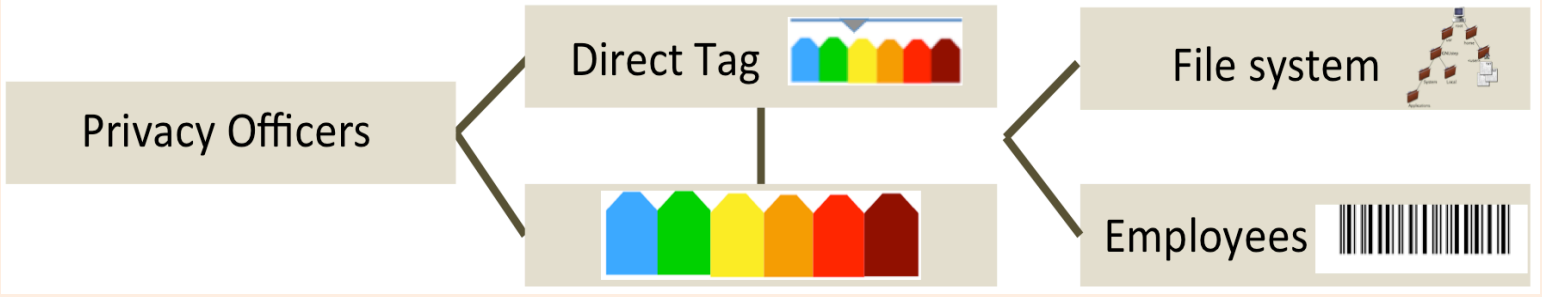


Diane: Multinational Corporation

- Cloud contains data from all over the world, collected under a variety of terms, subject to different laws
- Repository that enforces requirements on employee access

Diane: Multinational Corporation

Ingestion and Decision-making Knowledge	Local privacy officers and project leaders determine which datatags apply to which data sets and specify any additional restrictions or notices that apply.
Codification and Infrastructure	Blue, Green, Yellow, Orange, Red, Crimson with access based in part on the company's role-based access system.
Credentials and Retrieval	Employees having appropriate credentials in the company's role-based access system may access a file in the datags repository after acknowledging receipt of any notices about special handling required for the file. Employees may not share the files, even with other employees.

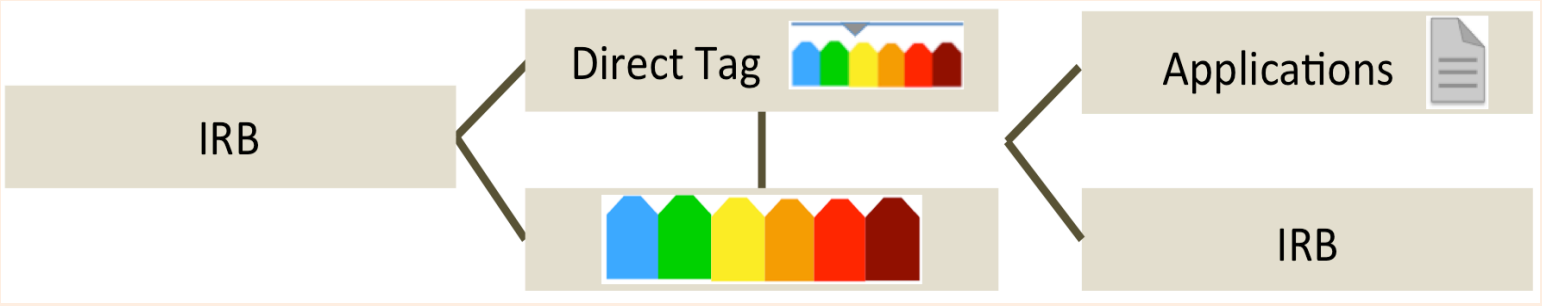


Charles: Institutional Review Board

- Document committee decisions
- Recommend handling based on prior decisions

Charles: Institutional Review Board

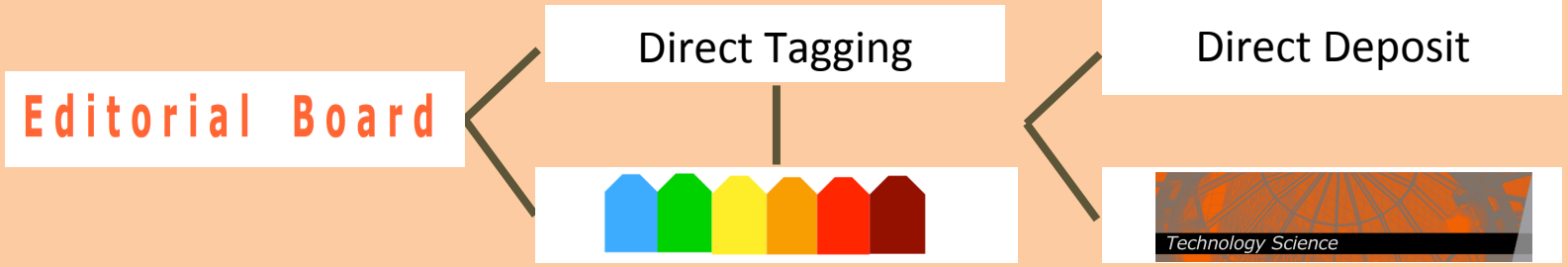
Ingestion and Decision-making Knowledge	The IRB determines which datatags apply to which data sets and specify any additional restrictions that apply. A copy of IRB documents appears as files in the repository, and not the data themselves.
Codification and Infrastructure	Blue, Green, Yellow, Orange, Red, Crimson. However, the access requirements associated with the tags are not used to access the IRB files. IRB committee members have password access to any file in the repository.
Credentials and Retrieval	IRB members can retrieve documents describing the data, as well as summary reports about the nature of data archived at each level.



DATA

Technology Science

How technology impacts humans.



Open.
Agree to cite.

Register email.
Agree to handling.

Confirm email.
Be approved.
Agree to handling.

Confirm email.
Be approved.
Sign agreement.

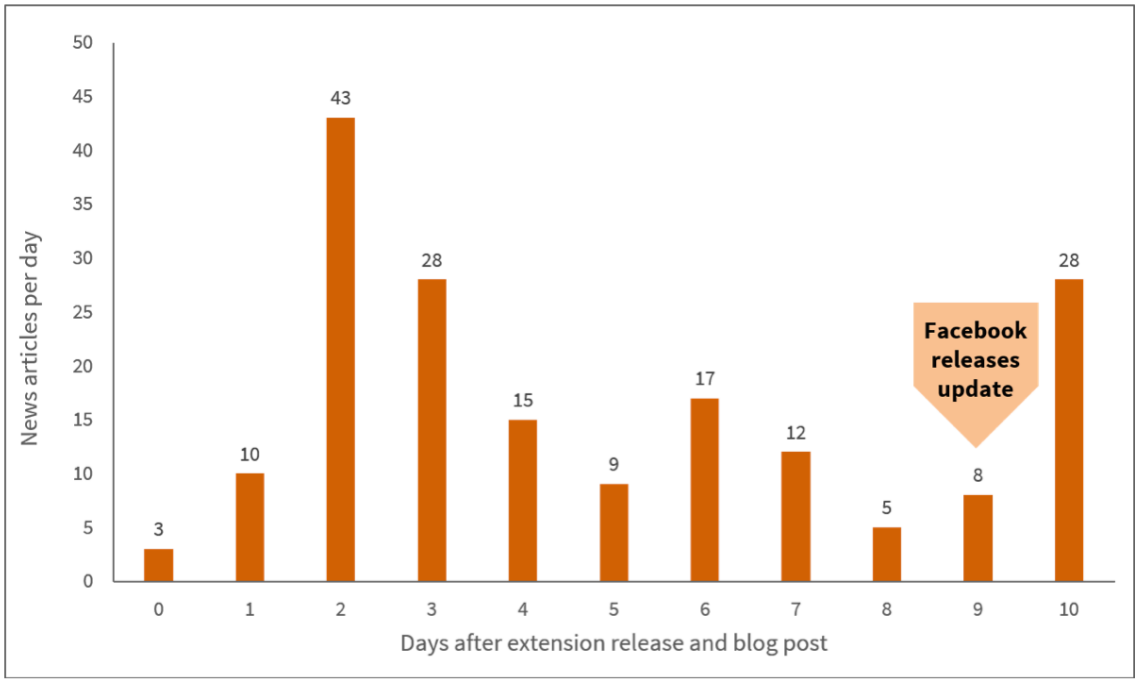
Confirm email, phone.
Be approved.
Sign agreement.

Confirm email, phone.
Be approved.
Sign agreement.

Facebook's Privacy Incident Response: a study of geolocation sharing on Facebook Messenger

Aran Khanna

Blue



News coverage by day

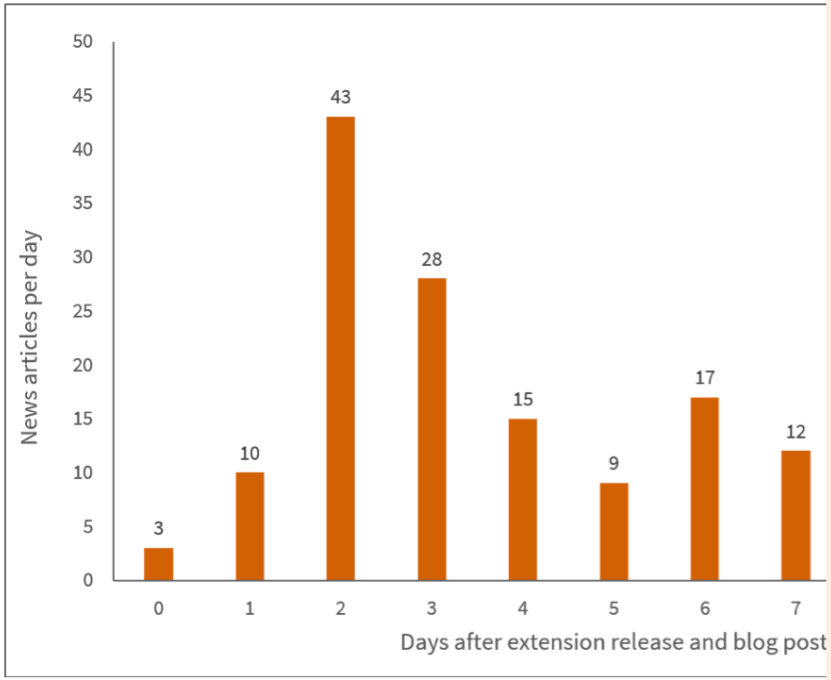
- In 2012, a media outlet reported that Facebook Messenger shared personal geolocations by default
- In 2015, my demonstration displayed Facebook's shared data on a map; it was downloaded over 85,000 times
- After 9 days of news coverage, Facebook released an update that requires a user's permission to share geolocations

techscience.org

Khanna A. Facebook's Privacy Incident Response: a study of geolocation sharing on Facebook Messenger. Technology Science. 2015081101. August 11, 2015. http://techscience.org/a/2015081101

Facebook's Privacy Incide
geolocation sharing on Fac

Aran Khanna



News coverage by day

techscience.org

Khanna A. Facebook's Privacy Incident Response: a
Technology Science. 2015081101. August 11, 2015

Dataverse About Guides Support Sign Up Log In

Technology Science

Technology Science Dataverse (Harvard University) Home Page

Harvard Dataverse > Technology Science Dataverse >

Replication Data for: Facebook's Privacy Incident Response, a study of geolocation sharing on Facebook Messenger

Metrics 88 Downloads

Replication Data for: Facebook's Privacy Incident Response, a study of geolocation sharing on Facebook Messenger

Khanna, Aran, 2015, "Replication Data for: Facebook's Privacy Incident Response, a study of geolocation sharing on Facebook Messenger", http://dx.doi.org/10.7910/DVN/D2SNRI, Harvard Dataverse, V1 [UNF:6:hiXa2O0z0wPt9CL8yBGHDA==]

Download Citation

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description	This dataset was used for this paper published on 8/11/2015 on Technology Science http://techscience.org/a/2015081104/
Subject	Computer and Information Science

Files Metadata Terms Versions

Download

20 Files

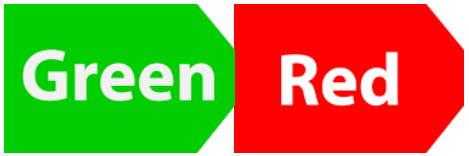
bing_news_search_results_1.docx
MS Word (docx) - 153.0 KB - Aug 10, 2015 - 13 Downloads
MD5: 8ef886cc26c1b493bca0b0366d080785;
Download

bing_news_search_results_2.docx
MS Word (docx) - 138.3 KB - Aug 10, 2015 - 3 Downloads
MD5: e0412b149afc92247f831aad184f5112;
Download

De-anonymizing South Korean Resident Registration Numbers Shared in Prescription Data

처방전 데이터의 주민등록번호 익명성 해제 연구

Latanya Sweeney and Ji Su Yoo



- South Korea's national identifier, the Resident Registration Number (RRN) includes encoded demographic information and a checksum with a publicly known pattern
- We conducted two de-anonymization experiments on 23,163 encrypted RRNs from prescription data of South Koreans
- We demonstrate the data's vulnerability to de-anonymization by revealing all 23,163 unencrypted RRNs in both experiments

Letter	Number
a	1
b	2
c	3
d	4
e	5
f	6
g	7
h	8
i	9
j	0

Odd-digit

Letter	Number
f	0
g	9
h	8
i	7
j	6
k	5
l	4
m	3
n	2
o	1

Even-digit

Coding table that replaced digits of South Korean national identifiers with letters in shared prescription data.

Published 2015-09-29

techscience.org

Sweeney L, Yoo J. De-anonymizing South Korean Resident Registration Numbers Shared in Prescription Data. Technology Science. 2015092901. September 29, 2015. http://techscience.org/a/2015092901

DATATAGGING TOOLS

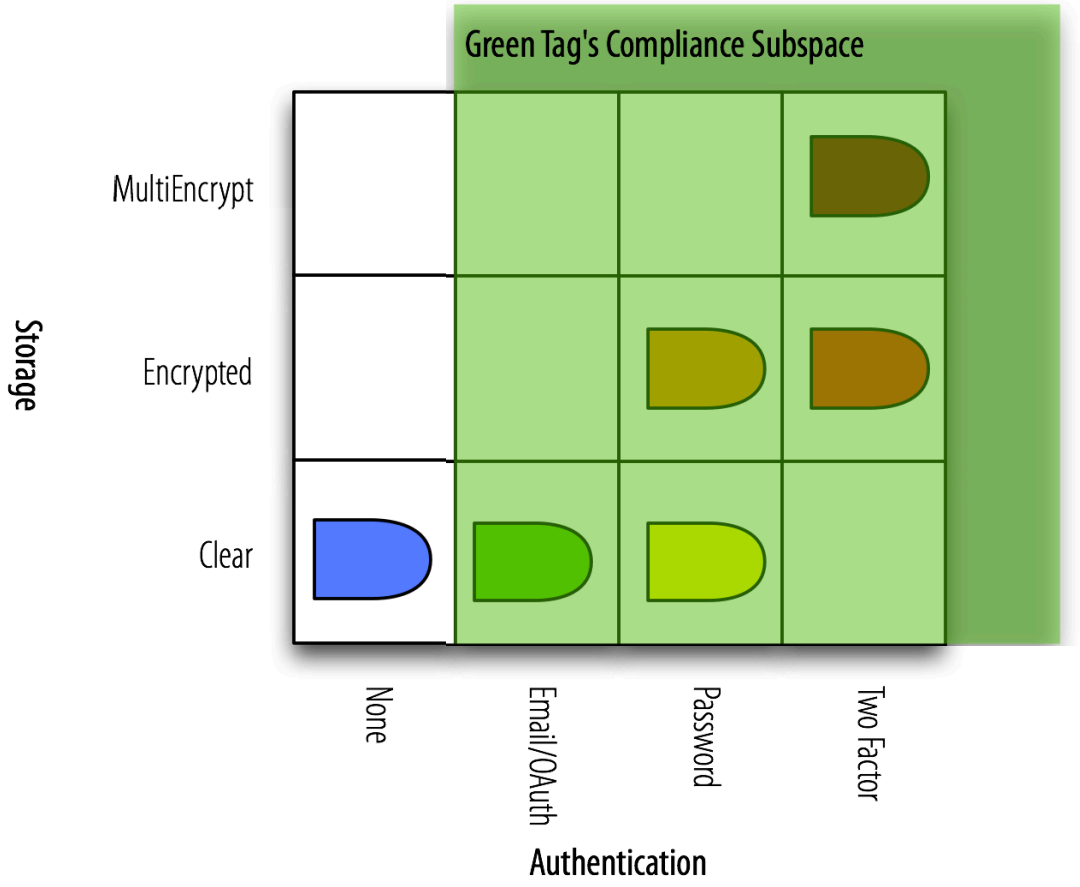
A Datatagging tool needs:

- Formal description of a Datatag
  - Capture the data handling policy of the tag
  - Capture the "stricter-than" ordering
- Interview creation tool
  - Support user-friendly interviews
  - Decide on the datatag based on the answers only

Formal Description of a Datatag

- Model data handling policies as a set of orthogonal aspects
  - Storage encryption, access requirements…
- Describe implementation options for each aspect; order implementations from lenient to strict
  - Clear < Encrypted < Multi Encrypt

Data Handling Policy Space



Tags: TagsSpace file (.ts)

- Describe a tag space
- Convenience features: hierarchy, “slots” of different types, top-down design support, comments…

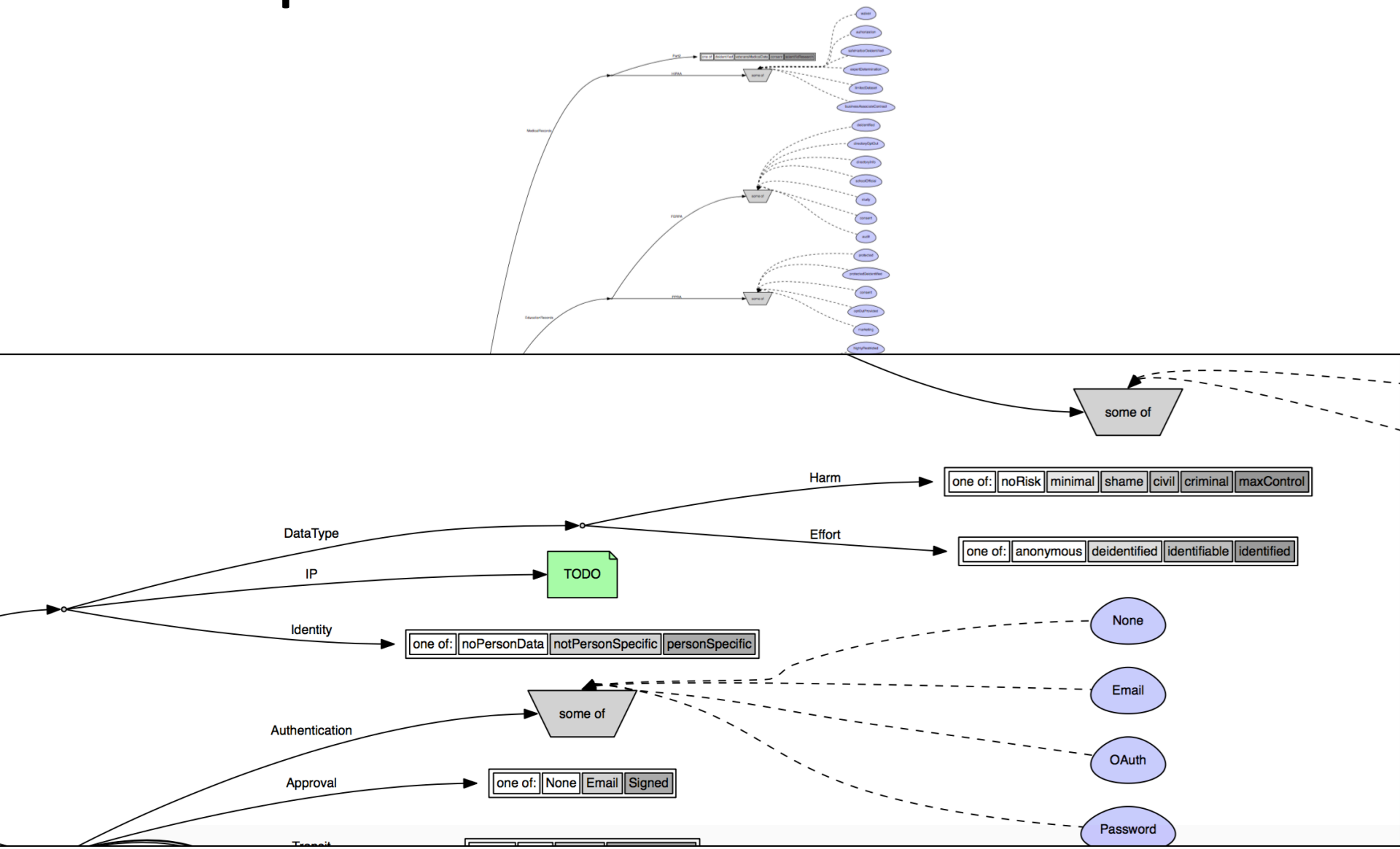
```
Authentication: some of
  None     [Available to anonymous individuals.],
  Email    [Available to individuals with verified email address.],
  OAuth    [Available to individuals with verified online identity or a mobile phone.],
  Password [Available to individuals having a password accounts on system.]
.

DataType: consists of
  Effort, Harm.

Effort:    one of  anonymous, deidentified, identifiable, identified.
Harm:      one of  noRisk, minimal, shame, civil, criminal, maxControl.
```

Screenshot from actual Atom package: Gal Maman, Matan Toledano, BGU

Comprehension Aid: Visualization



Finding the Right Tag – Decision Graph

- Directed, Acyclic Graph
- Node Types:
  - Ask
  - Set
  - Convenience: Call, End, Reject, Todo

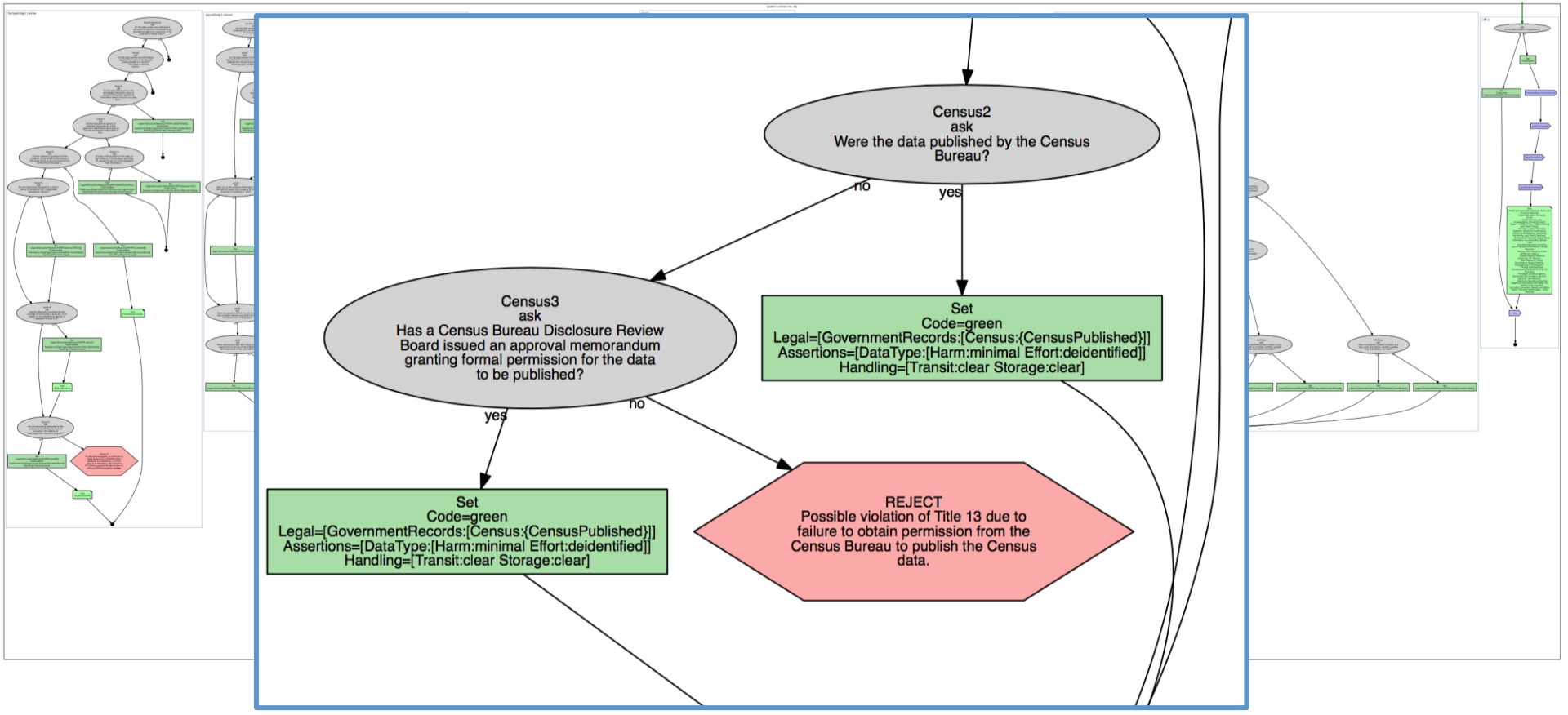
Finding the Right Tag – Decision Graph

```
<*
Sample toy interview
*>
[call: ensureLegality]
[ask:
  {text: Do the data contains personally identifiable information?}
  {terms:
    {Personally identifiable information: Any information about an individual... }
  }
  {answers:
    {yes: [set: Storage=encrypt; Transfer=encrypt]
    {no: [set: Storage=clear; Transfer=clear]}}]
[todo: Test for additional ...] <-- Issue #42 follows
[end]

[>ensureLegality< ask:
  {text: Did you get parental consent?}
  {answers:
    {no: [reject: Must get parental consent before collecting data from subjects under 18.]}}]
[end]
```

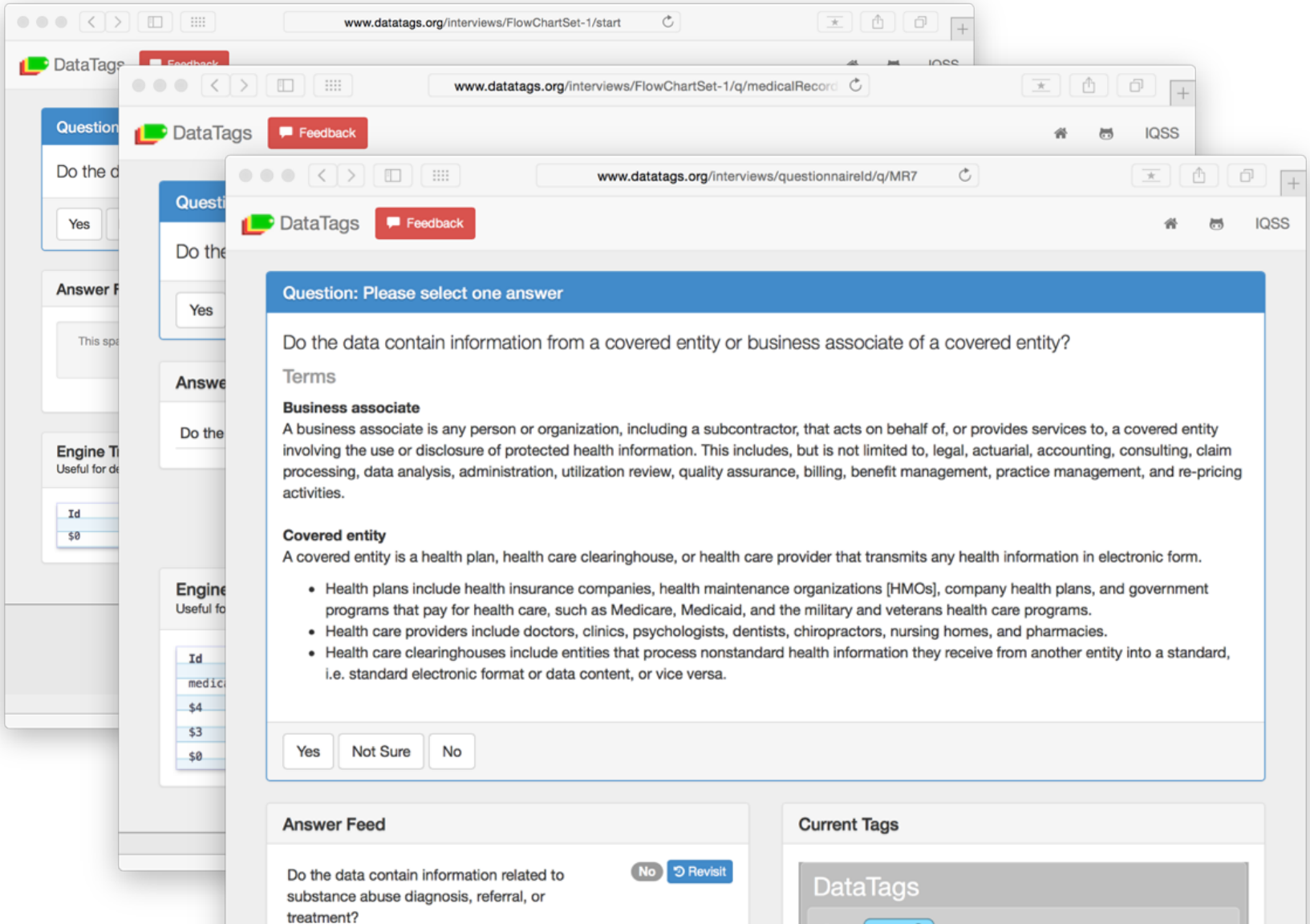
Screenshot from actual Atom package: Gal Maman, Matan Toledano, BGU

Interview Visualization

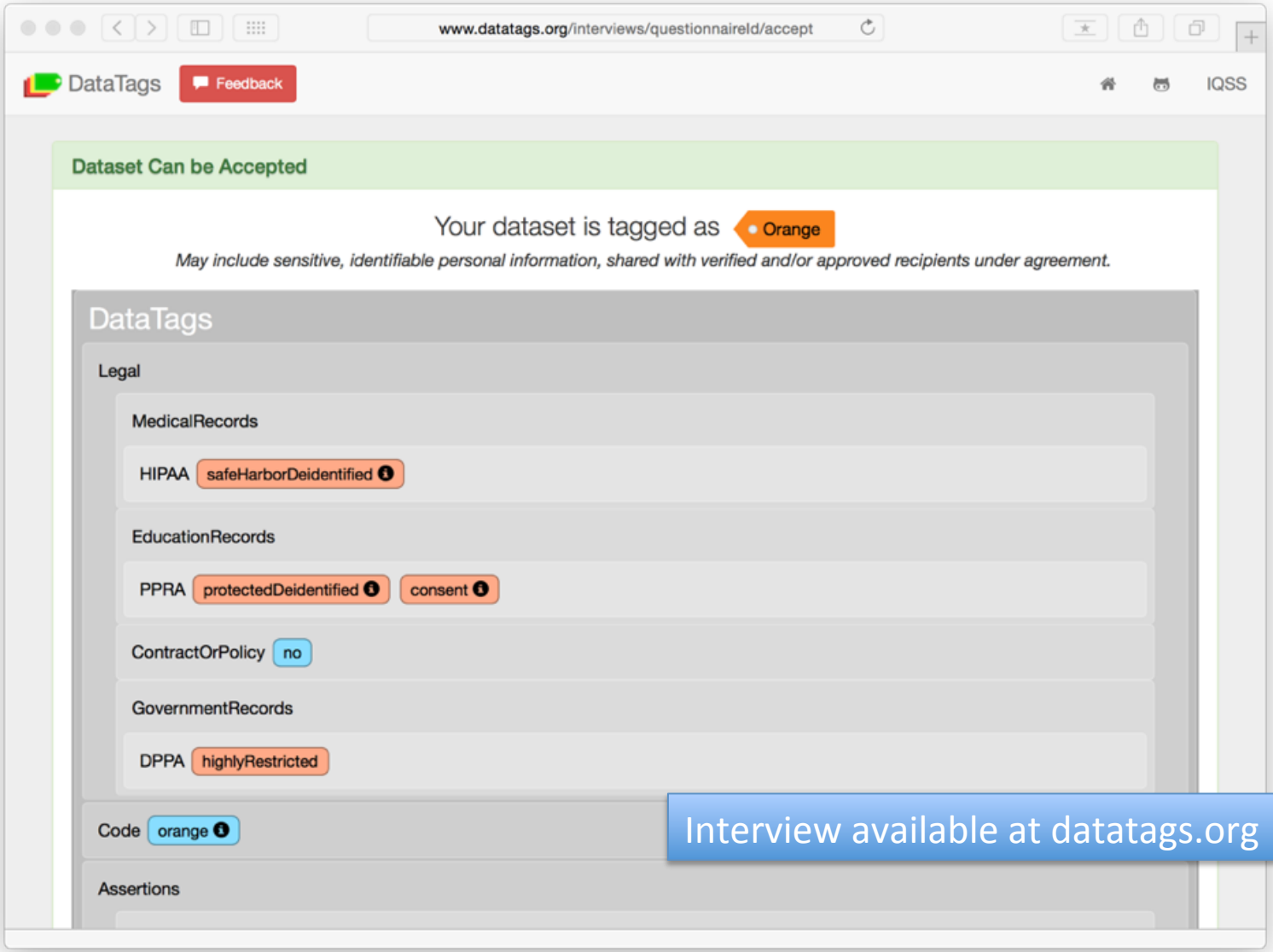


Interview credit: The Data Privacy Lab @ Harvard (Latanya Sweeney, Sean Hooley), Berkman Ctr. for Internet and Society (Alexandra Wood, David O'Brien, clinical students), IQSS (Mercè Crosas, Michael Bar-Sinai). Part of the Privacy tools for sharing research data project

Interview on the Web

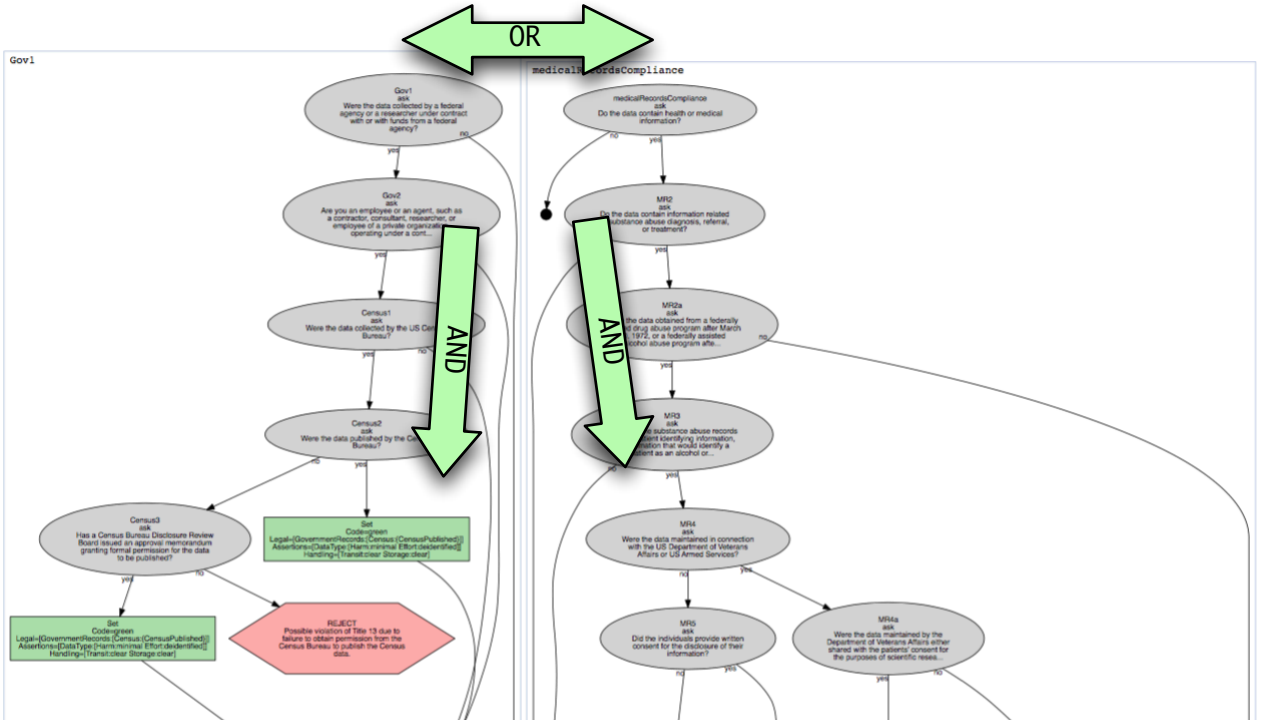


Interview on the Web



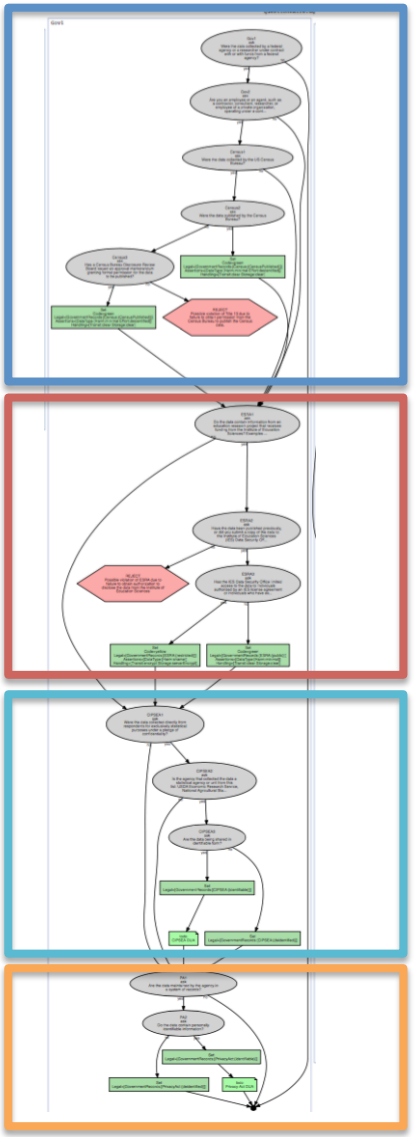
Decision Graph Points

- Familiar “interview with a specialist” metaphor
- Implicitly describe logic inference



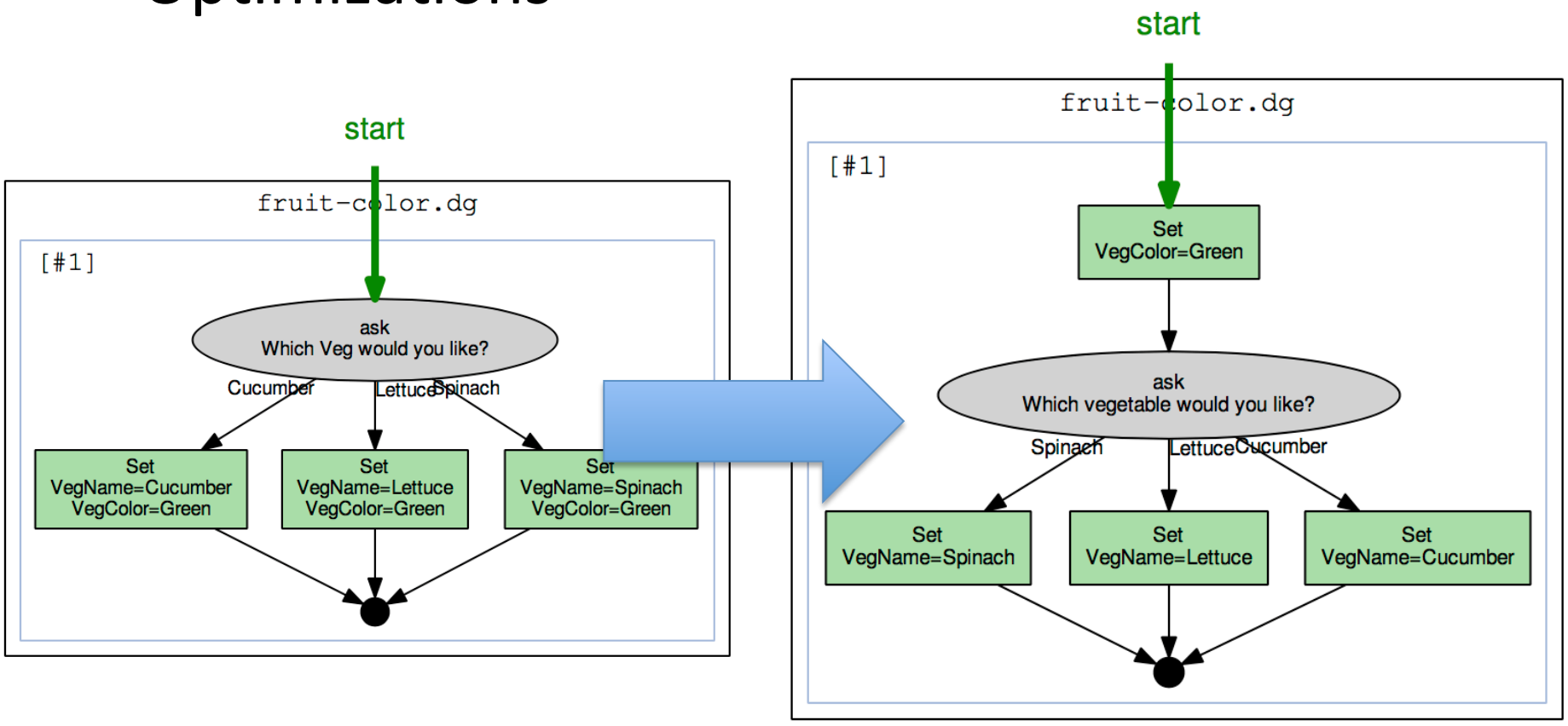
Decision Graph Points

- Analysis: Detection of Independent parts
- Queries, such as “what series of answers will create a datatags that allows clear storage?”



Decision Graph Points

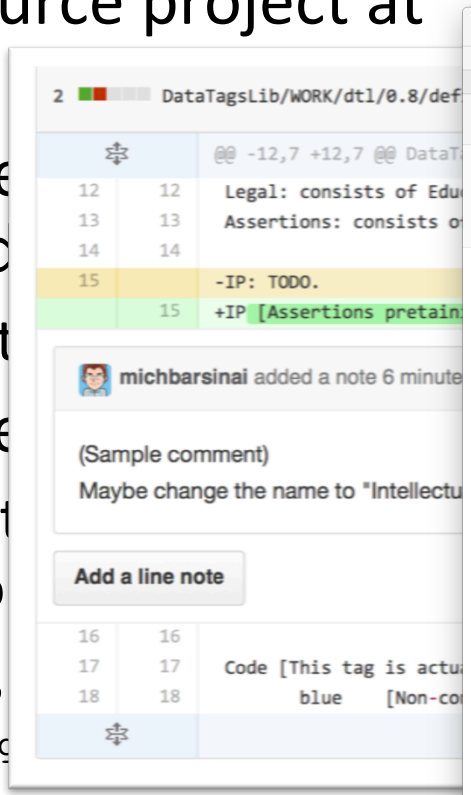
- Optimizations

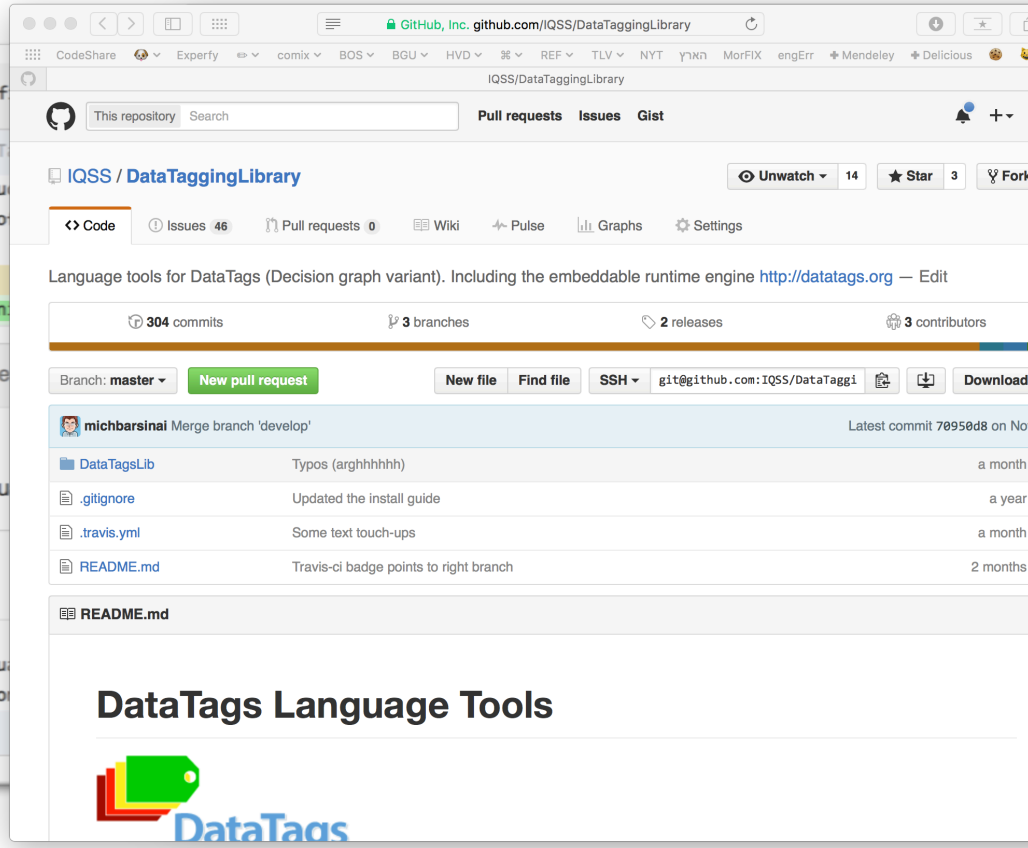


Example created by Eyal Ben-Simon, BGU

State of the Tags Tool

- Open-source project at GitHub
- Language tools and
  - Project
- Language
  - Inspect develo
- Tutorials
  `datatagging`
- Collaboration via, e.g. GitHub





Future of the Tags Tool

- Update web interview application
  - Include upload and inspection features
- On-line collaboration environment
  - A-la Google docs?

Mercè Crosas, Michael Bar-Sinai, Latanya Sweeney

THANKS