Privacy Tools for Sharing Research Data
A National Science Foundation Secure and Trustworthy Cyberspace Project, with additional support from the Sloan Foundation and Google, Inc.

The Dataverse Project

DataTags, the Tags Toolset, and Dataverse Integration Project

Michael Bar-Sinai
Computer Science Dept.
Ben-Gurion University of the Negev
Be'er-Sheva, Israel

Latanya Sweeney
Data Privacy Lab
Harvard University
Cambridge, MA

Mercè Crosas
Institute for Quantitative Social Science
Harvard University
Cambridge, MA

Managing Privacy in Research Data Repositories Workshop

July 13th, 2016

@michbarsinai

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences
Ben-Gurion University of the Negev
NSF
Data Privacy Lab
The Institute for Quantitative Social Science at Harvard University

Based in part on:

Sweeney L, Crosas M, Bar-Sinai M. **Sharing Sensitive Data with Confidence: The Datatags System**. *Technology Science [Internet]*. 2015. Technology Science

Bar-Sinai M, Sweeney L, Crosas M. **DataTags, Data Handling Policy Spaces, and the Tags Language**. *Proceedings of the International Workshop on Privacy Engineering*. 2016. IEEE Symposium on Security and Privacy ("Oakland")

We present a framework for
formally describing,
reasoning about, and
arriving at
data-handling policies

We present a framework for
formally describing,
reasoning about, and
arriving at
data-handling policies
Making it Easier to
store and share scientific datasets

Why Share Data?

- **Good Science**
  - Transparency
  - Collaboration
  - Research acceleration
  - Reproducibility
  - Data citation
- **Compliance** with requirements from sponsors and publishers

W ?

- **Good Science**
  - Transparenc
  - Collaboratio
  - Research ac
  - Reproducibi
  - Data citatior
- **Compliance** w s and publishers

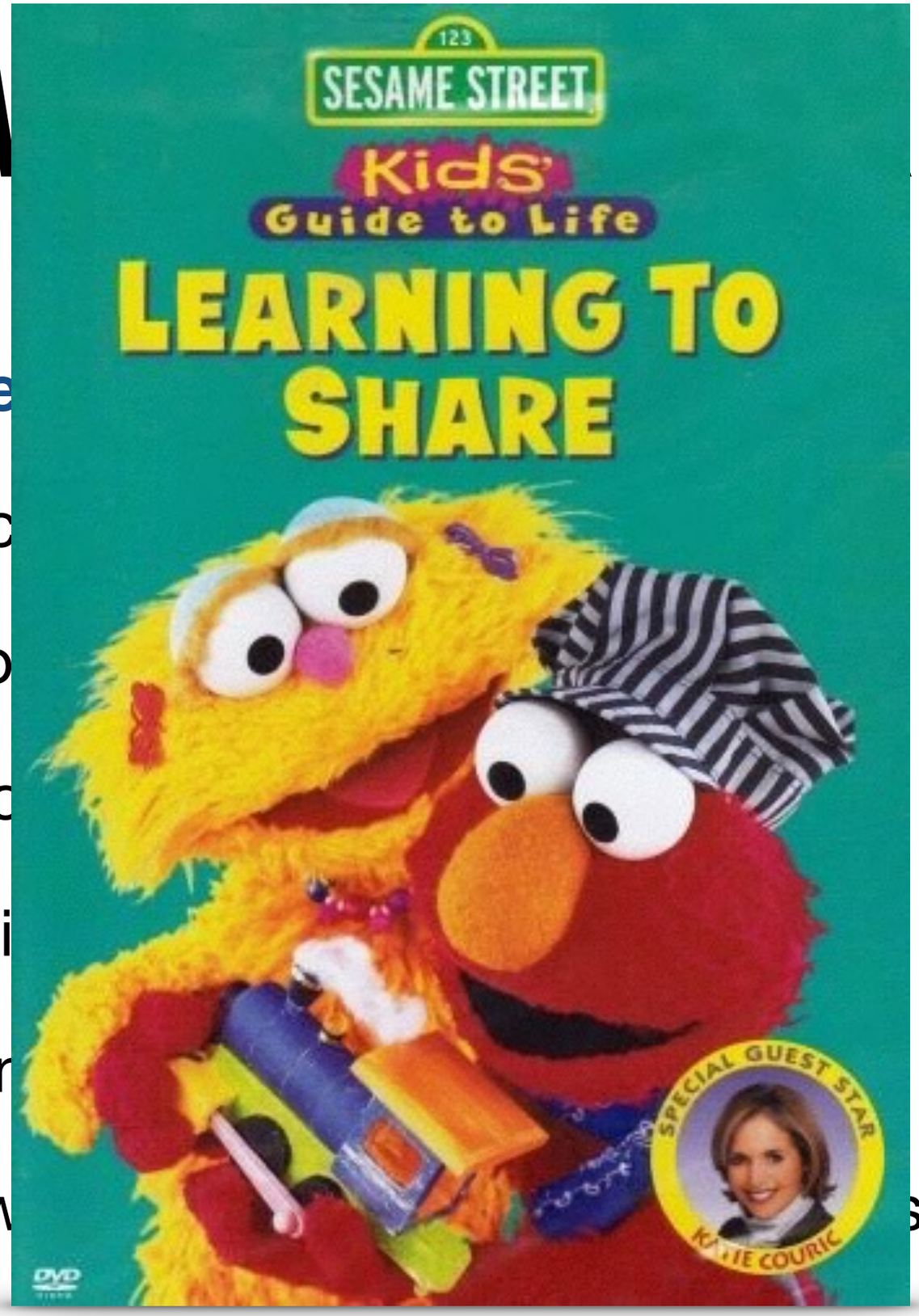


image: © Sesame Workshop

http://s4.evcdn.com/images/edpborder500/I0-001/003/710/831-6.jpeg_/avenue-q-31.jpeg

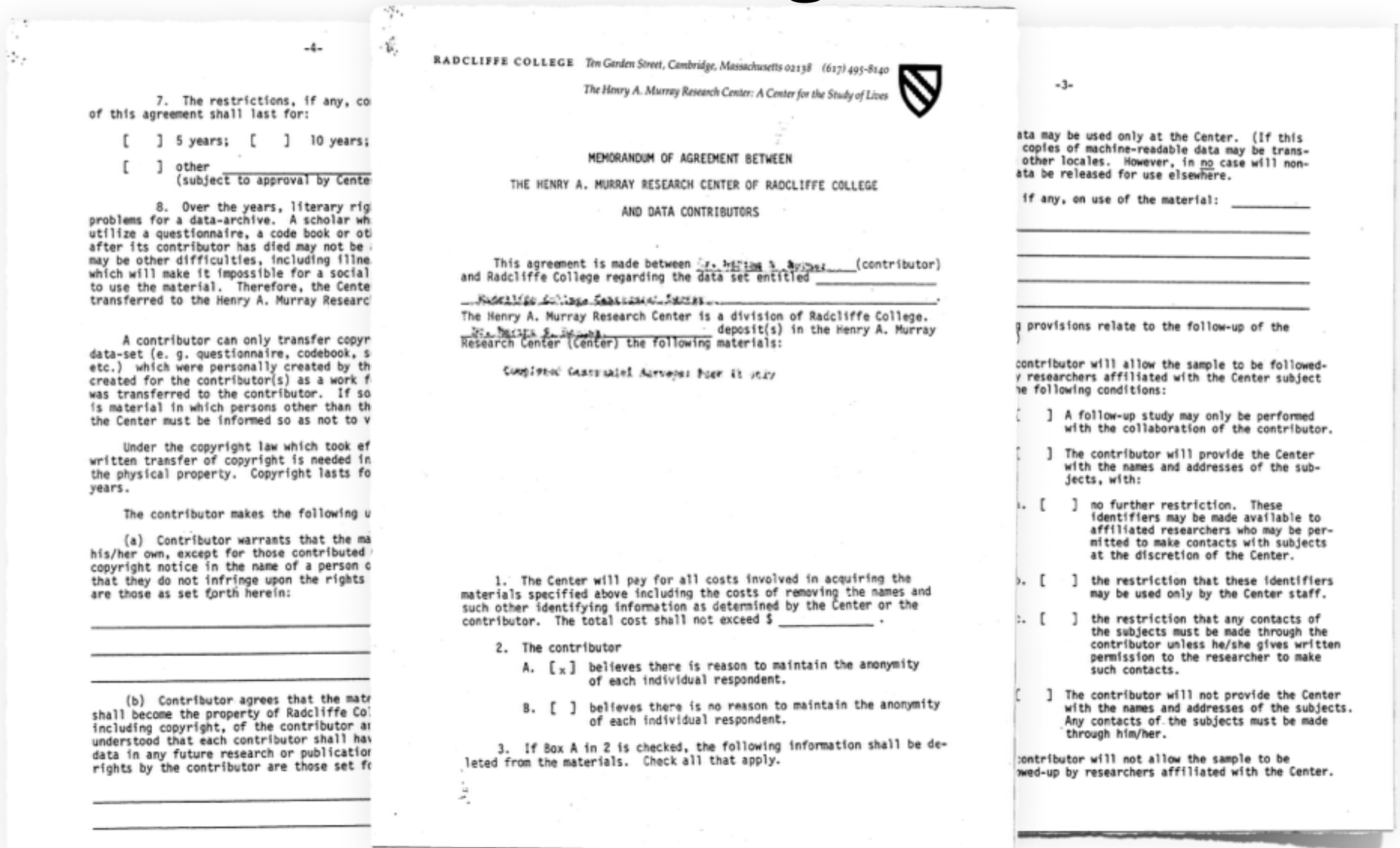
Sharing Data is Nontrivial

- ***Sharing may harm the data subjects***
- **Law** is complex
  - 2187 privacy laws in the US alone, at federal, state and local level, usually context-specific [Sweeney, 2013]
- **Technology** is complex
  - E.g. encryption standards change constantly, as new vulnerabilities are found
- **Specific dataset provenance** (may be) complex

Dataset handling policies
play the critical role of balancing
privacy risks and scientific value
of sharing datasets.

Here are some

Data Handling Policies



-4-

7. The restrictions, if any, co
of this agreement shall last for:

[ ] 5 years; [ ] 10 years;

[ ] other _____
(subject to approval by Cente

8. Over the years, literary rig
problems for a data-archive. A scholar wh
utilize a questionnaire, a code book or ot
after its contributor has died may not be
may be other difficulties, including illne
which will make it impossible for a social
to use the material. Therefore, the Cente
transferred to the Henry A. Murray Researc

A contributor can only transfer copyr
data-set (e. g. questionnaire, codebook, s
etc.) which were personally created by th
created for the contributor(s) as a work f
was transferred to the contributor. If so
is material in which persons other than th
the Center must be informed so as not to v

Under the copyright law which took ef
written transfer of copyright is needed in
the physical property. Copyright lasts fo
years.

The contributor makes the following u

(a) Contributor warrants that the ma
his/her own, except for those contributed
copyright notice in the name of a person c
that they do not infringe upon the rights
are those as set forth herein:

_____

_____

_____

(b) Contributor agrees that the matr
shall become the property of Radcliffe Co
including copyright, of the contributor a
understood that each contributor shall hav
data in any future research or publicatio
rights by the contributor are those set f

_____

_____

RADCLIFFE COLLEGE *Ten Garden Street, Cambridge, Massachusetts 02138 (617) 495-8140*

The Henry A. Murray Research Center: A Center for the Study of Lives

MEMORANDUM OF AGREEMENT BETWEEN

THE HENRY A. MURRAY RESEARCH CENTER OF RADCLIFFE COLLEGE

AND DATA CONTRIBUTORS

This agreement is made between [illegible] (contributor) and Radcliffe College regarding the data set entitled _____ [illegible].

The Henry A. Murray Research Center is a division of Radcliffe College. [illegible] deposit(s) in the Henry A. Murray Research Center (Center) the following materials:

[illegible]

1. The Center will pay for all costs involved in acquiring the materials specified above including the costs of removing the names and such other identifying information as determined by the Center or the contributor. The total cost shall not exceed $ _____ .

2. The contributor

A. [x] believes there is reason to maintain the anonymity of each individual respondent.

B. [ ] believes there is no reason to maintain the anonymity of each individual respondent.

3. If Box A in 2 is checked, the following information shall be deleted from the materials. Check all that apply.

-3-

ata may be used only at the Center. (If this
copies of machine-readable data may be trans-
other locales. However, in no case will non-
ata be released for use elsewhere.

if any, on use of the material: _____

_____

_____

_____

_____

g provisions relate to the follow-up of the

contributor will allow the sample to be followed-
y researchers affiliated with the Center subject
he following conditions:

[ ] A follow-up study may only be performed with the collaboration of the contributor.

[ ] The contributor will provide the Center with the names and addresses of the subjects, with:

a. [ ] no further restriction. These identifiers may be made available to affiliated researchers who may be permitted to make contacts with subjects at the discretion of the Center.

b. [ ] the restriction that these identifiers may be used only by the Center staff.

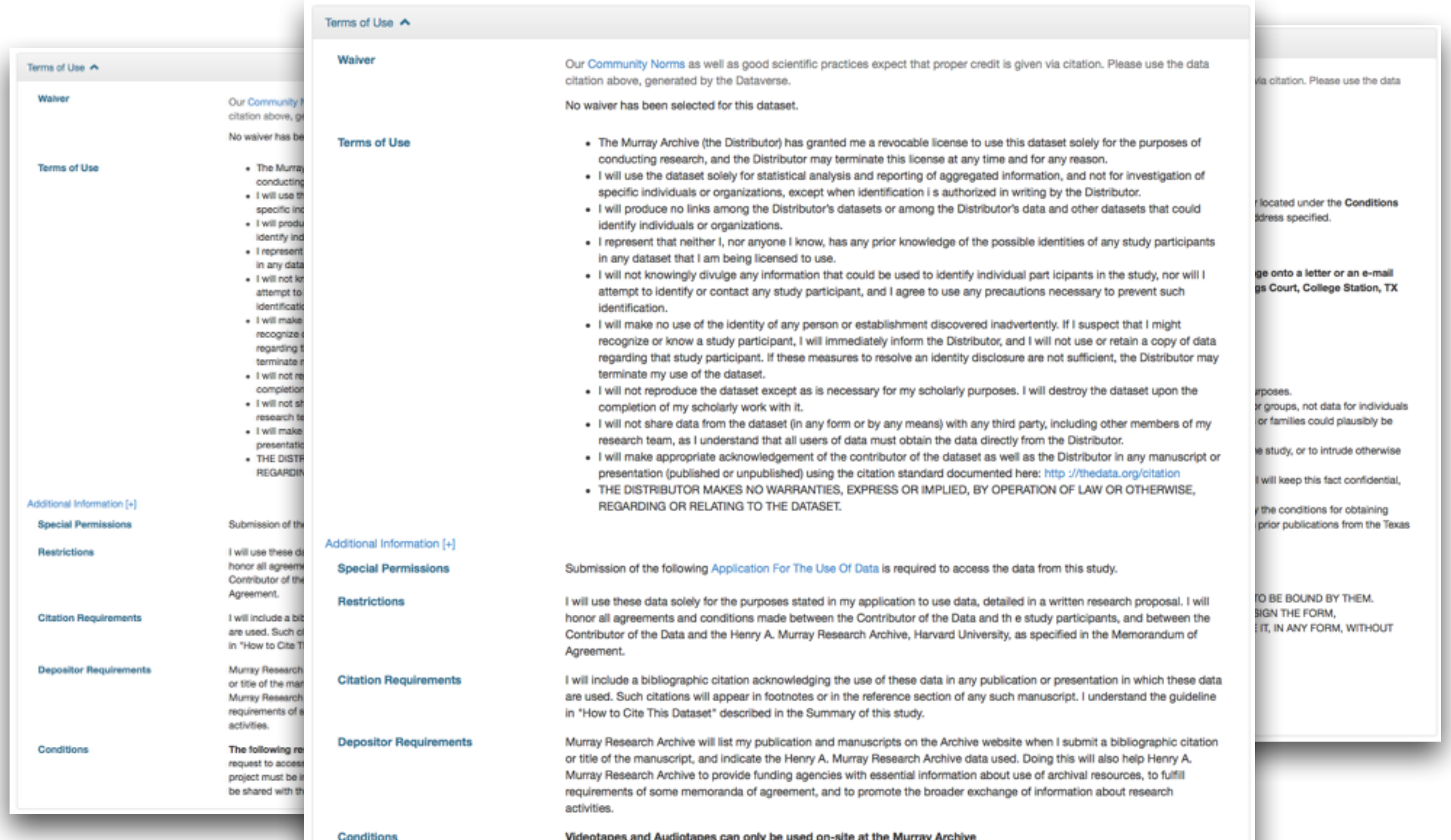
c. [ ] the restriction that any contacts of the subjects must be made through the contributor unless he/she gives written permission to the researcher to make such contacts.

[ ] The contributor will not provide the Center with the names and addresses of the subjects. Any contacts of the subjects must be made through him/her.

contributor will not allow the sample to be
wed-up by researchers affiliated with the Center.

Here are some new

Data Handling Policies



Terms of Use

Waiver

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

No waiver has been selected for this dataset.

Terms of Use

- The Murray Archive (the Distributor) has granted me a revocable license to use this dataset solely for the purposes of conducting research, and the Distributor may terminate this license at any time and for any reason.
- I will use the dataset solely for statistical analysis and reporting of aggregated information, and not for investigation of specific individuals or organizations, except when identification i s authorized in writing by the Distributor.
- I will produce no links among the Distributor's datasets or among the Distributor's data and other datasets that could identify individuals or organizations.
- I represent that neither I, nor anyone I know, has any prior knowledge of the possible identities of any study participants in any dataset that I am being licensed to use.
- I will not knowingly divulge any information that could be used to identify individual part icipants in the study, nor will I attempt to identify or contact any study participant, and I agree to use any precautions necessary to prevent such identification.
- I will make no use of the identity of any person or establishment discovered inadvertently. If I suspect that I might recognize or know a study participant, I will immediately inform the Distributor, and I will not use or retain a copy of data regarding that study participant. If these measures to resolve an identity disclosure are not sufficient, the Distributor may terminate my use of the dataset.
- I will not reproduce the dataset except as is necessary for my scholarly purposes. I will destroy the dataset upon the completion of my scholarly work with it.
- I will not share data from the dataset (in any form or by any means) with any third party, including other members of my research team, as I understand that all users of data must obtain the data directly from the Distributor.
- I will make appropriate acknowledgement of the contributor of the dataset as well as the Distributor in any manuscript or presentation (published or unpublished) using the citation standard documented here: http ://thedata.org/citation
- THE DISTRIBUTOR MAKES NO WARRANTIES, EXPRESS OR IMPLIED, BY OPERATION OF LAW OR OTHERWISE, REGARDING OR RELATING TO THE DATASET.

Additional Information [+]

Special Permissions

Submission of the following Application For The Use Of Data is required to access the data from this study.

Restrictions

I will use these data solely for the purposes stated in my application to use data, detailed in a written research proposal. I will honor all agreements and conditions made between the Contributor of the Data and th e study participants, and between the Contributor of the Data and the Henry A. Murray Research Archive, Harvard University, as specified in the Memorandum of Agreement.

Citation Requirements

I will include a bibliographic citation acknowledging the use of these data in any publication or presentation in which these data are used. Such citations will appear in footnotes or in the reference section of any such manuscript. I understand the guideline in "How to Cite This Dataset" described in the Summary of this study.

Depositor Requirements

Murray Research Archive will list my publication and manuscripts on the Archive website when I submit a bibliographic citation or title of the manuscript, and indicate the Henry A. Murray Research Archive data used. Doing this will also help Henry A. Murray Research Archive to provide funding agencies with essential information about use of archival resources, to fulfill requirements of some memoranda of agreement, and to promote the broader exchange of information about research activities.

Conditions

Videotapes and Audiotapes can only be used on-site at the Murray Archive

$\text{Formal}_{\text{cs}}$ DHPs

W3C's Privacy Preference Project (P3P)

Focuses on web data collection

Open Digital Rights Language (ODRL)

Models DRM, supports privacy and rule-based assertions

PrimeLife Policy Language (PPL)

Focuses on downstream usage, using rules

Data-Purpose Algebra

Models restriction transformation along data processing path

Robot Lawyers

See next session

DataTags

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

DataTags and their respective policies

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The Datatags System. Technology Science [Internet]. 2015.

Data-handling policies consist of independent *aspects*.

Encryption at rest, transfer type, access credentials, etc.

Data-handling policies consist of independent *aspects*.

Encryption at rest, transfer type, access credentials, etc.

Each aspect has multiple *possible requirements*, and can be defined such that these requirements are ordered.

DHPs: From Text to Space

Data-handling policies consist of independent *aspects*.

Encryption at rest, transfer type, access credentials, etc.

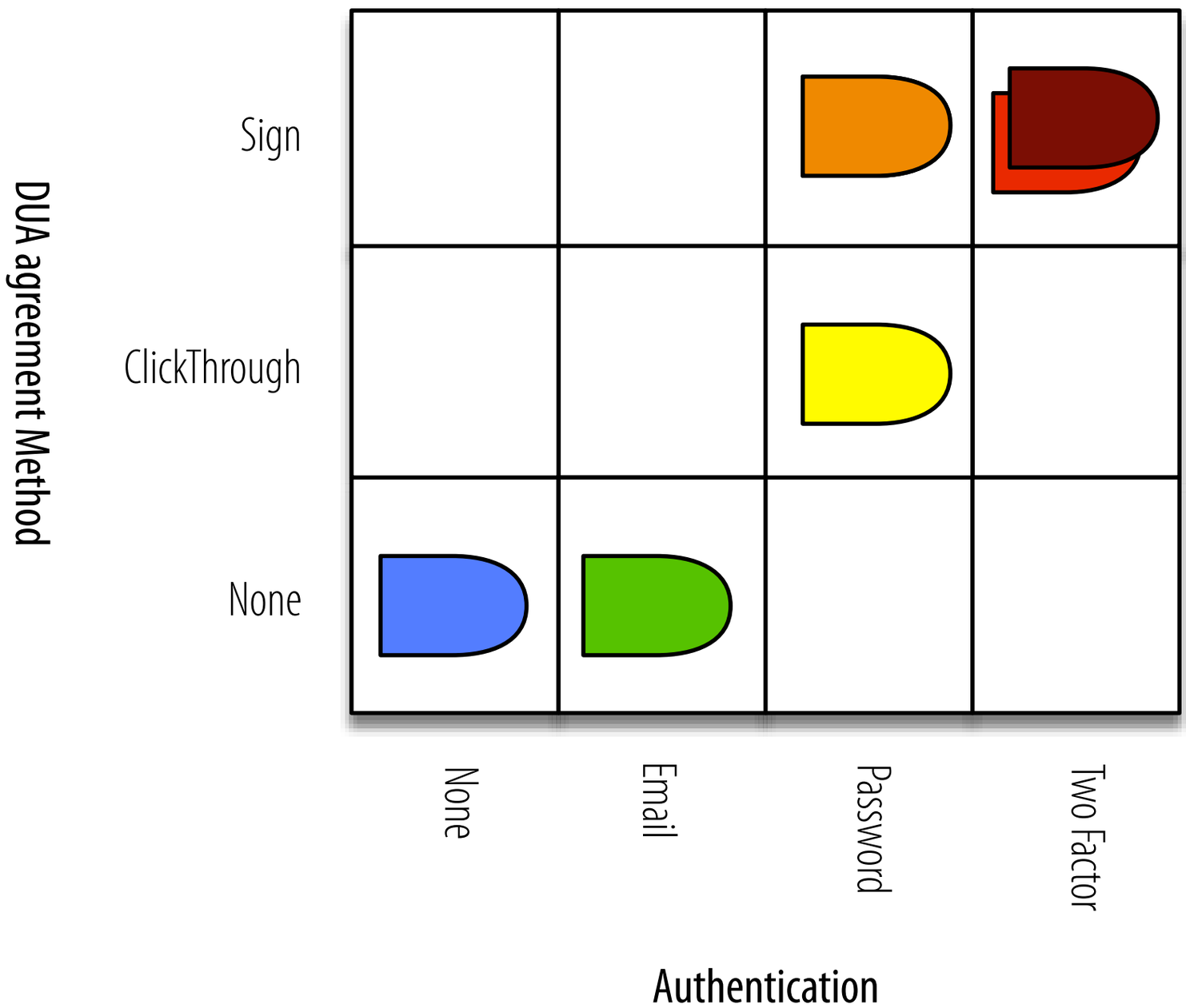
Each aspect has multiple *possible requirements*, and can be defined such that these requirements are ordered.

Construct a data-handling policy space by viewing aspects as axes, where each aspect's possible requirements serves as its coordinates.

Going from this…

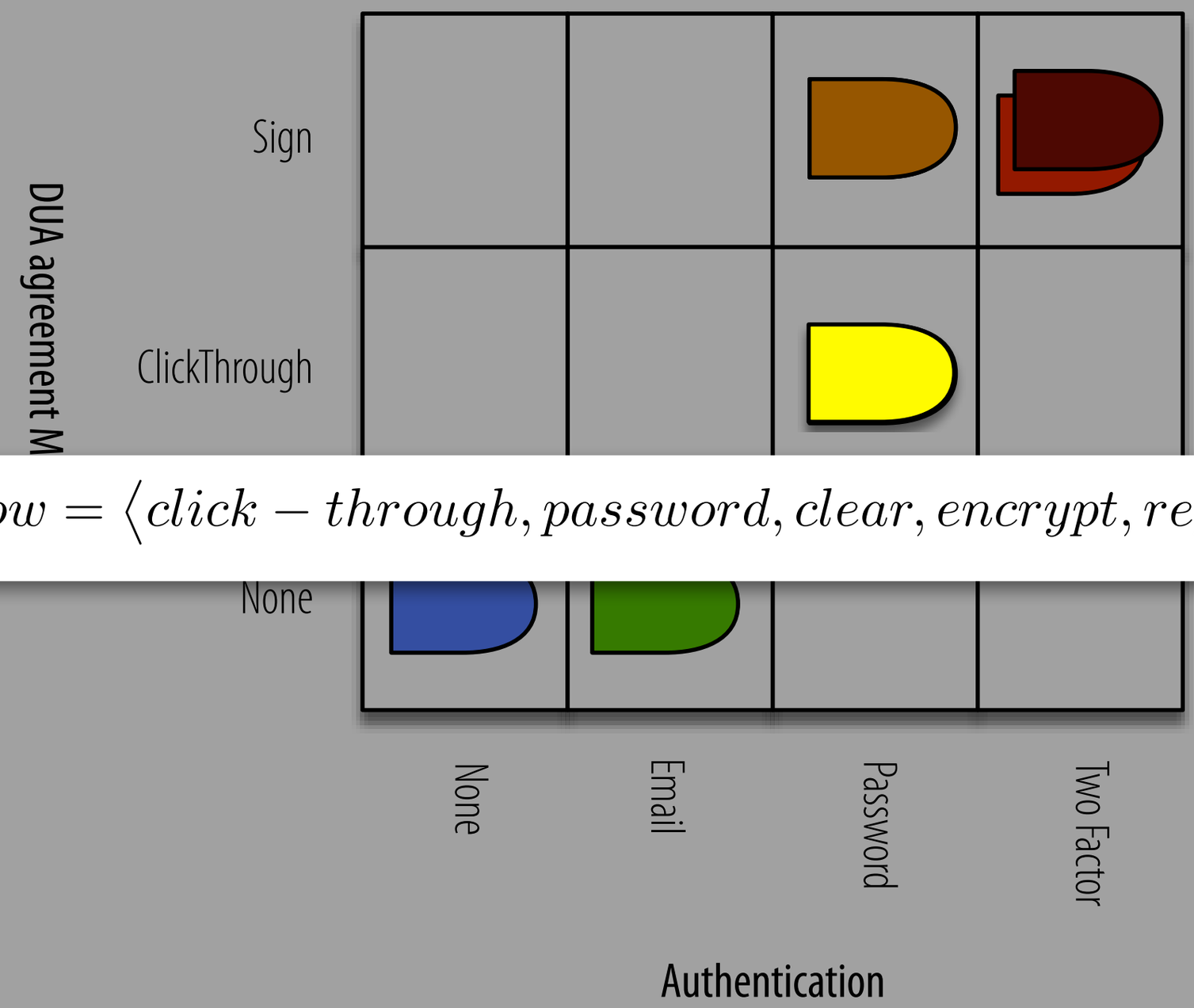
Tag Type	*Description*	*Security Features*	*Access Credentials*
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through 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

…to This*



** Shown here is a 2-D projection over the DUA Agreement Method and Authentication axes.*

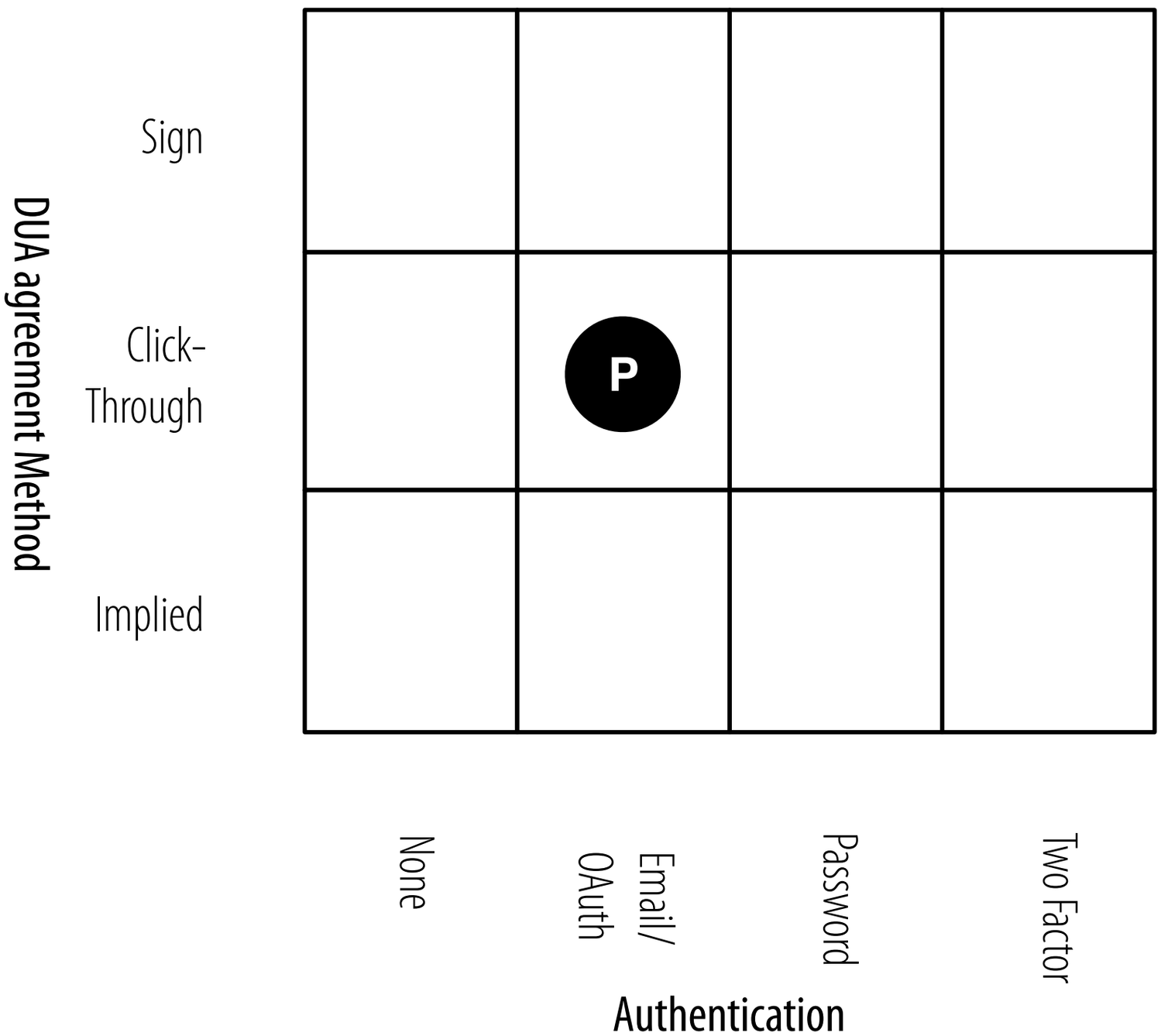
…to This*



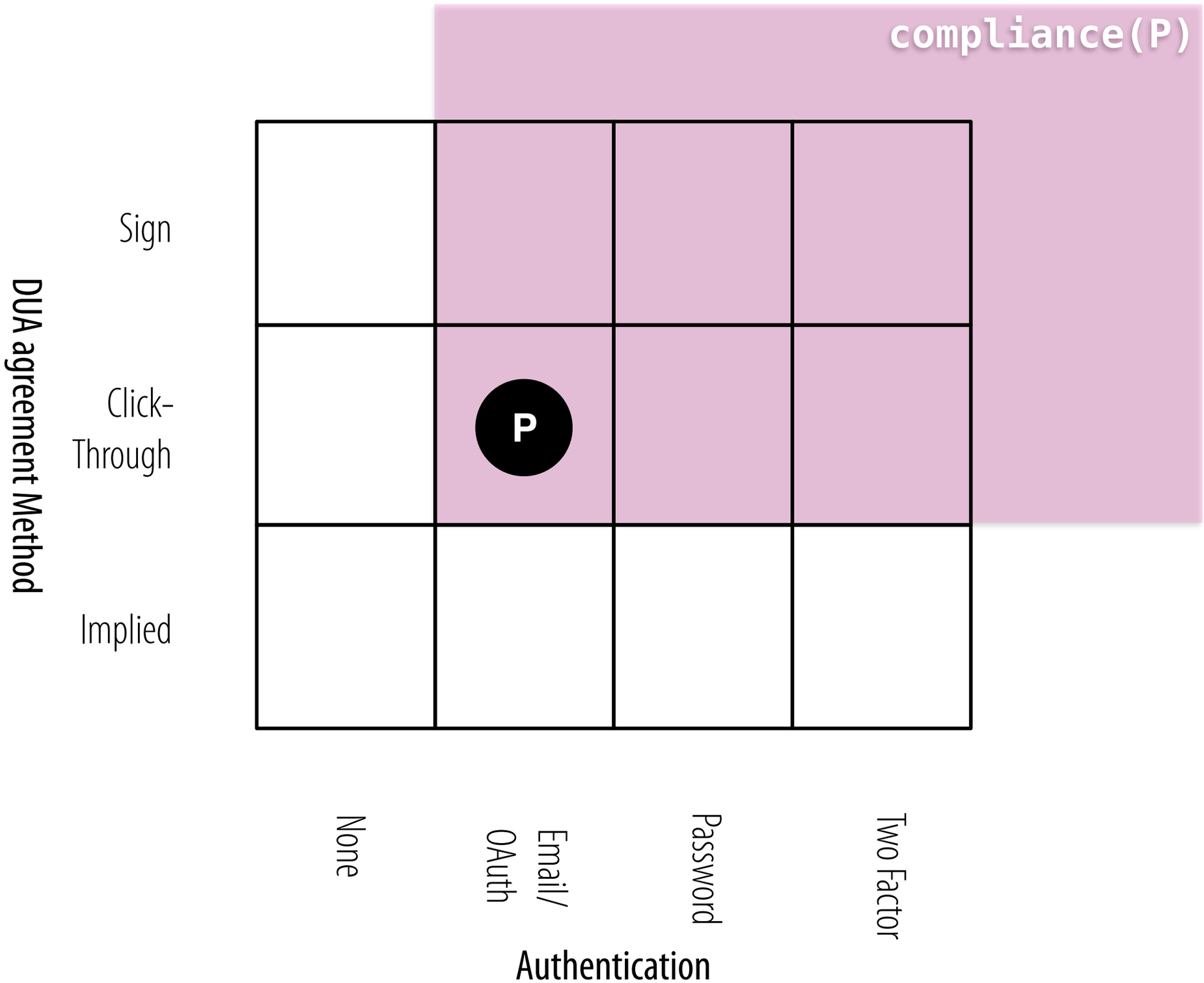
$$Yellow = \langle click - through, password, clear, encrypt, registered, approval \rangle$$

** Shown here is a 2-D projection over the DUA Agreement Method and Authentication axes.*

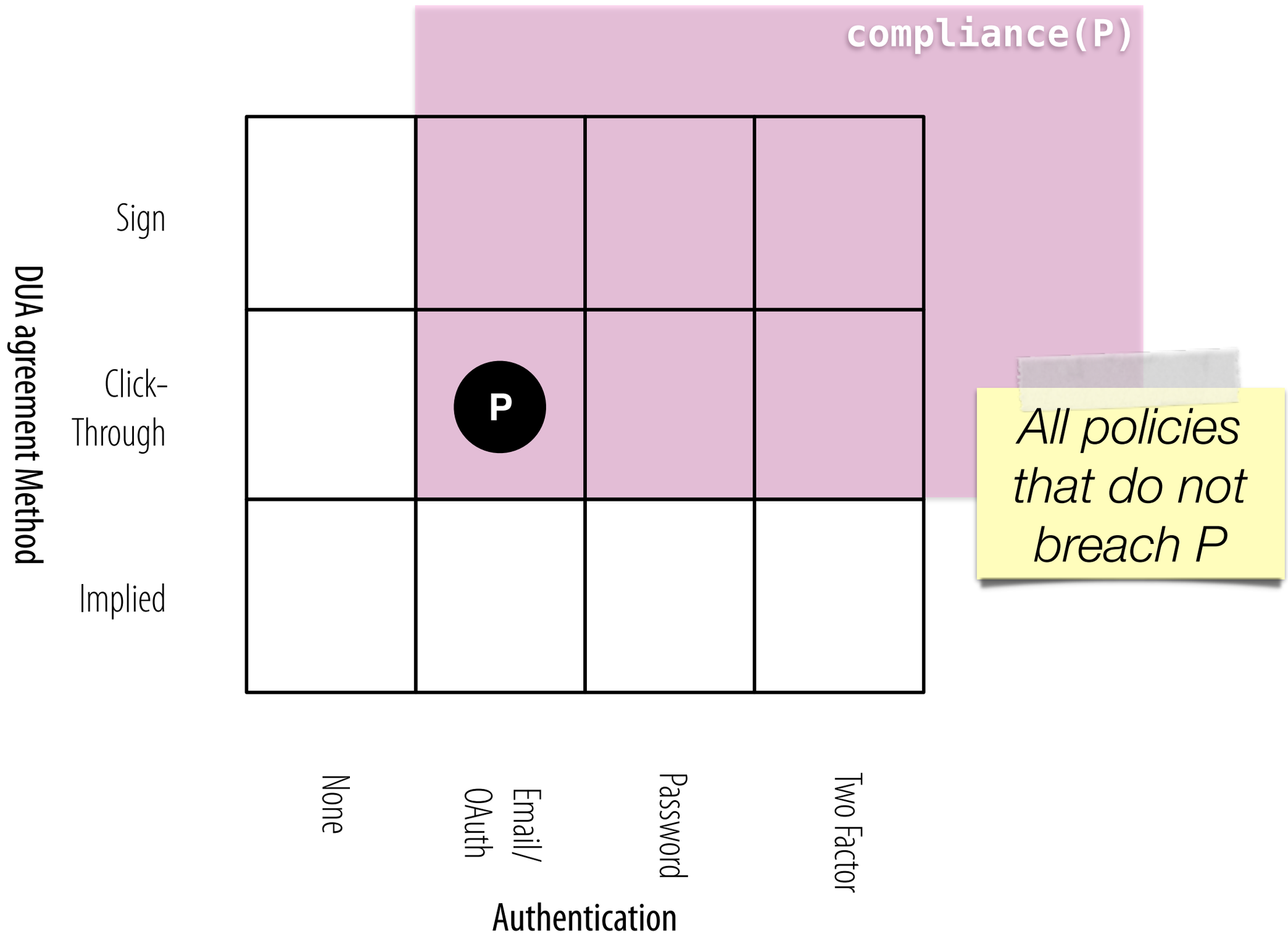
Strictness


DUA agreement Method
Sign
Click-Through
Implied
P
None
Email/ OAuth
Password
Two Factor
Authentication

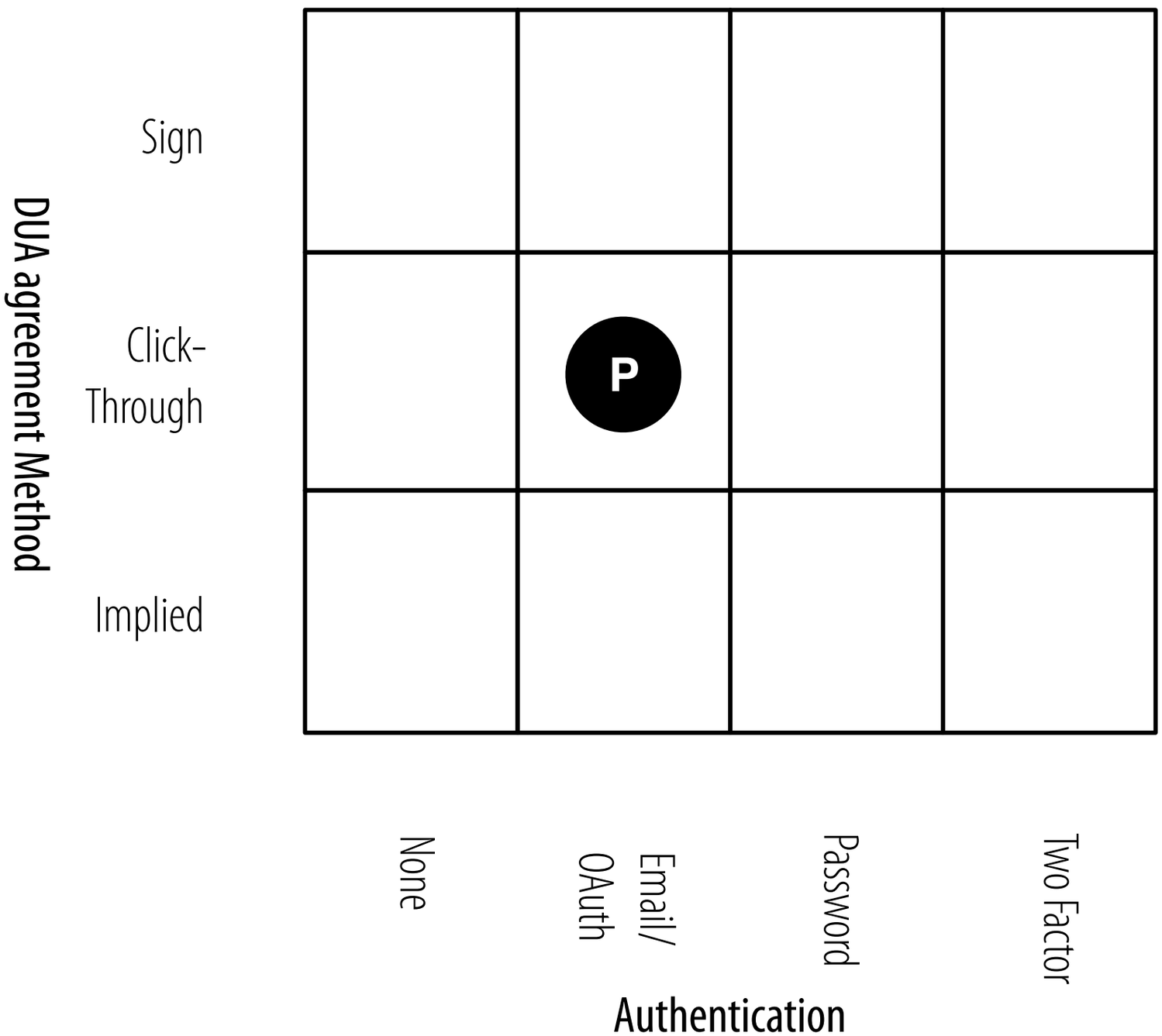
Strictness


compliance(P)
P
DUA agreement Method
Sign
Click-Through
Implied
None
Email/ OAuth
Password
Two Factor
Authentication

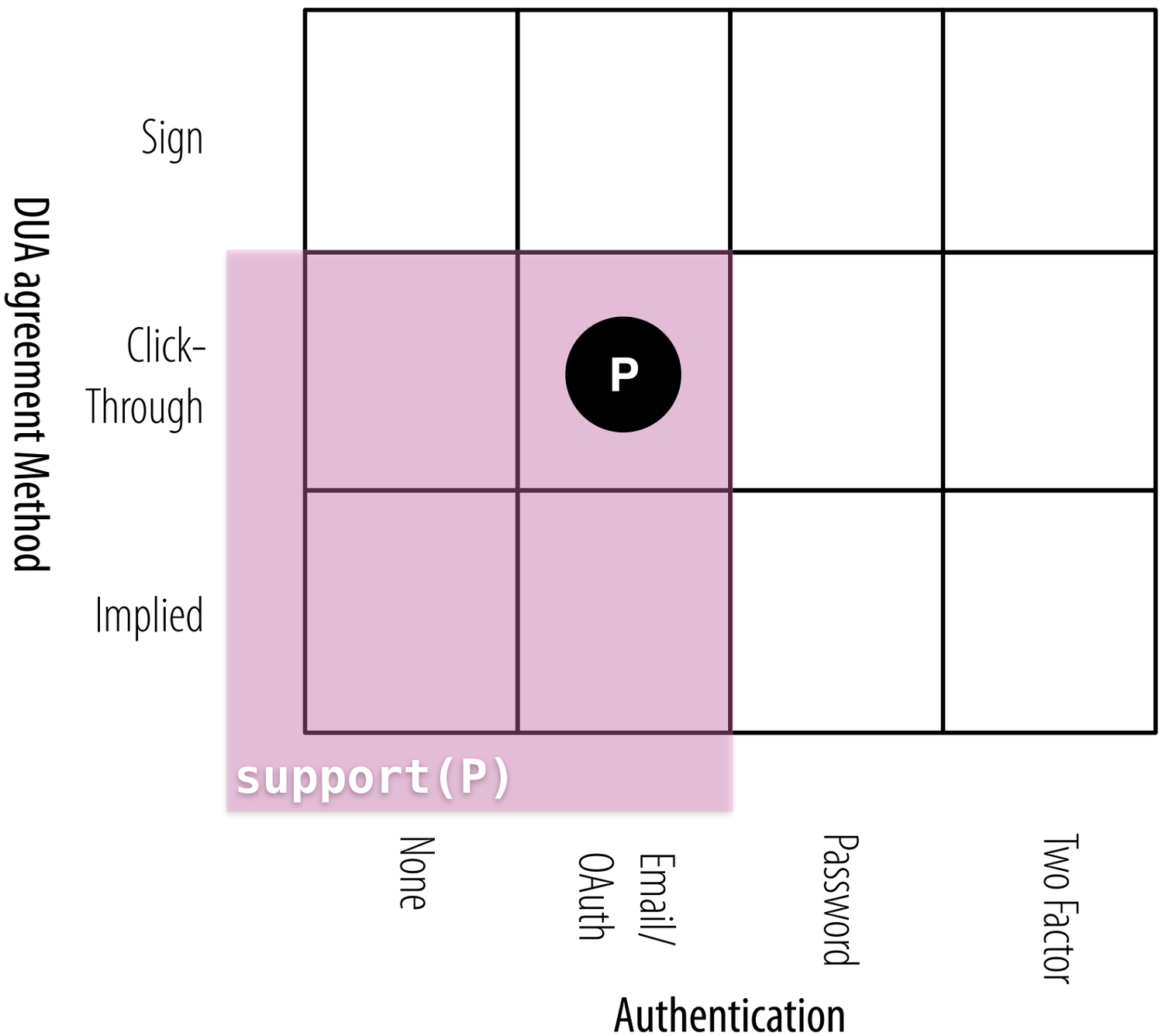
Strictness


compliance(P)
P
Sign
Click-Through
Implied
DUA agreement Method
None
Email/ OAuth
Password
Two Factor
Authentication
All policies that do not breach P

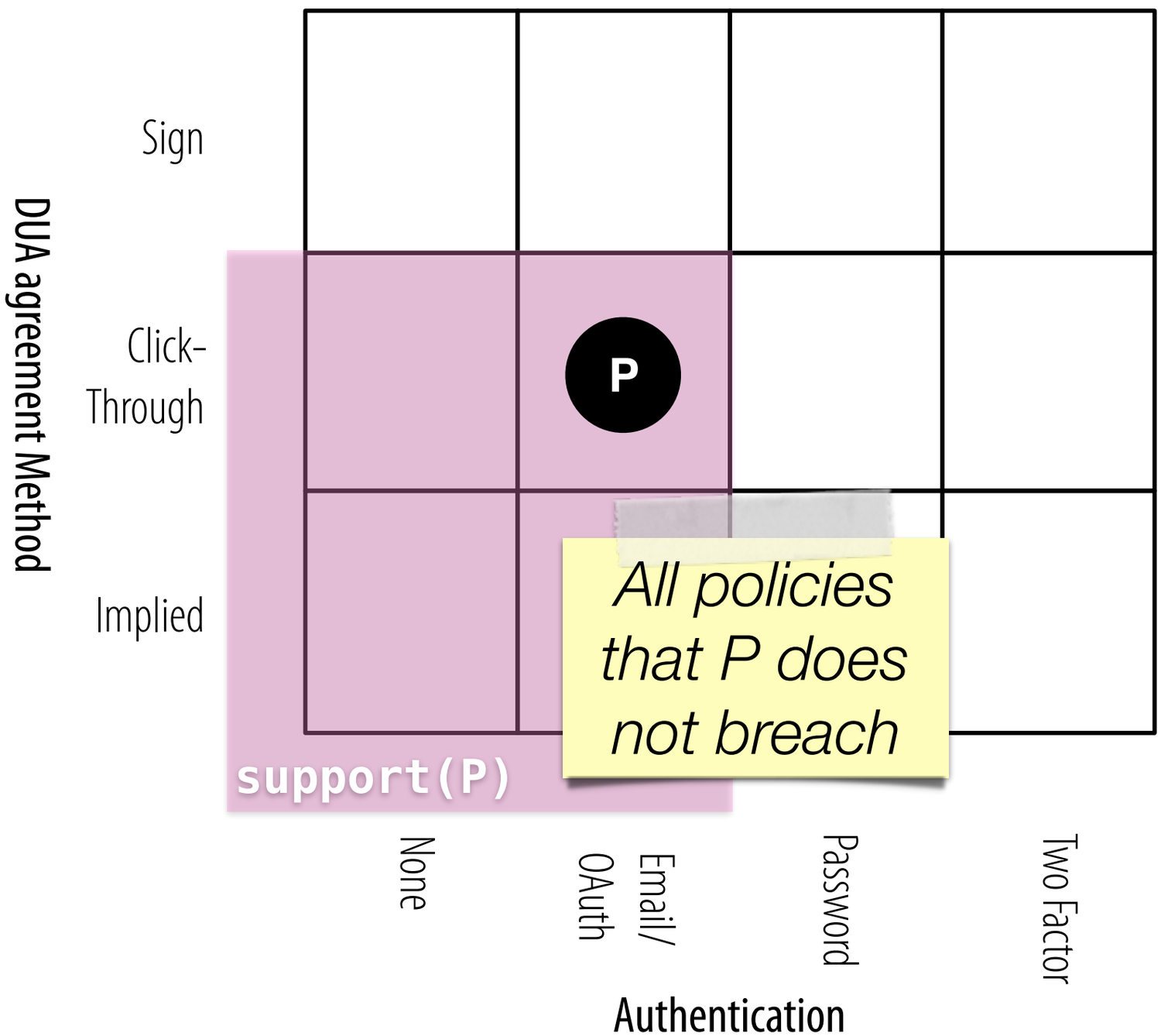
Lenience


DUA agreement Method
Sign
Click-Through
Implied
P
None
Email/ OAuth
Password
Two Factor
Authentication

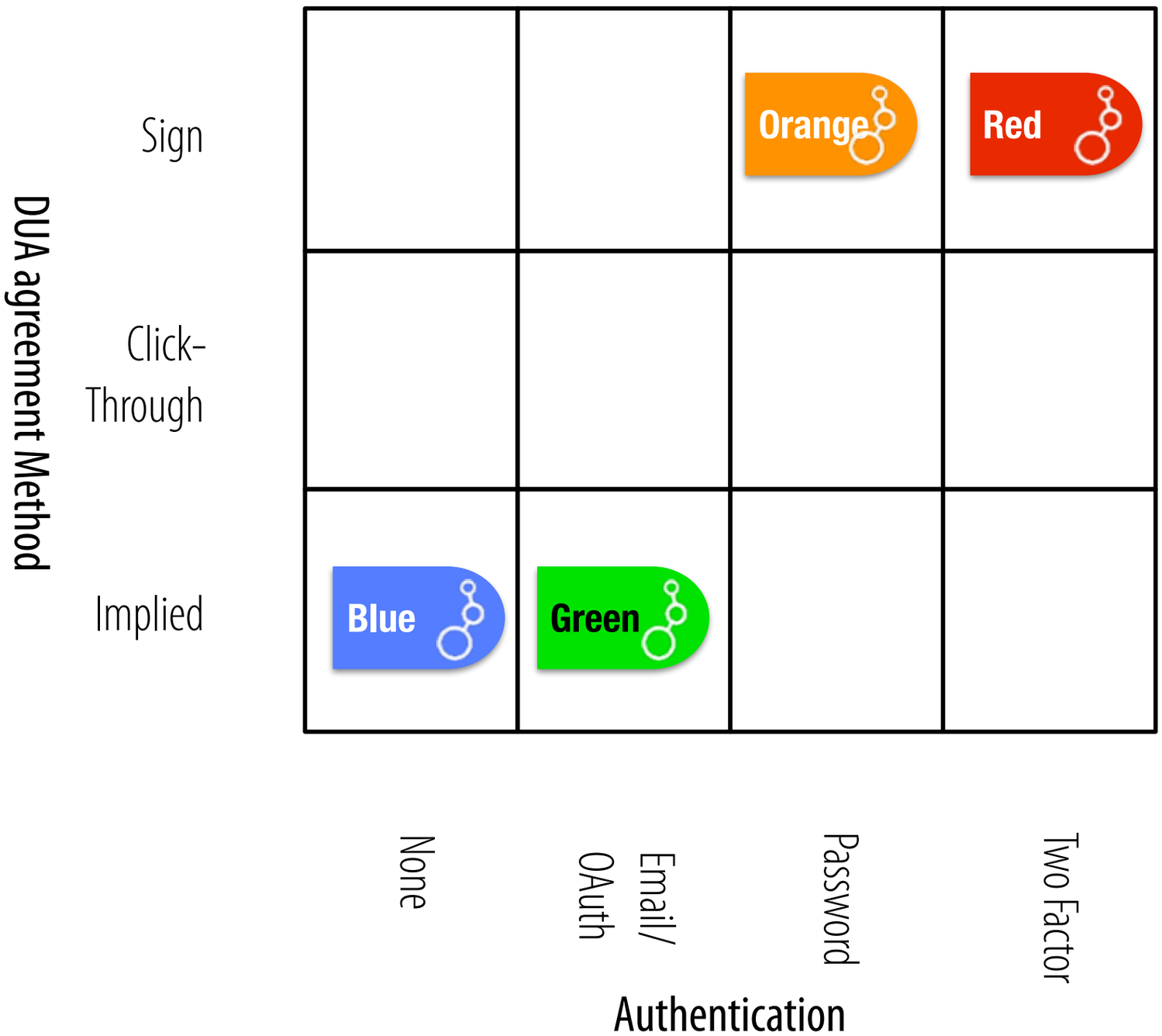
Lenience


Two Factor
Password
Email/
OAuth
None
Authentication
support(P)
P
Click-
Through
Implied
Sign
DUA agreement Method

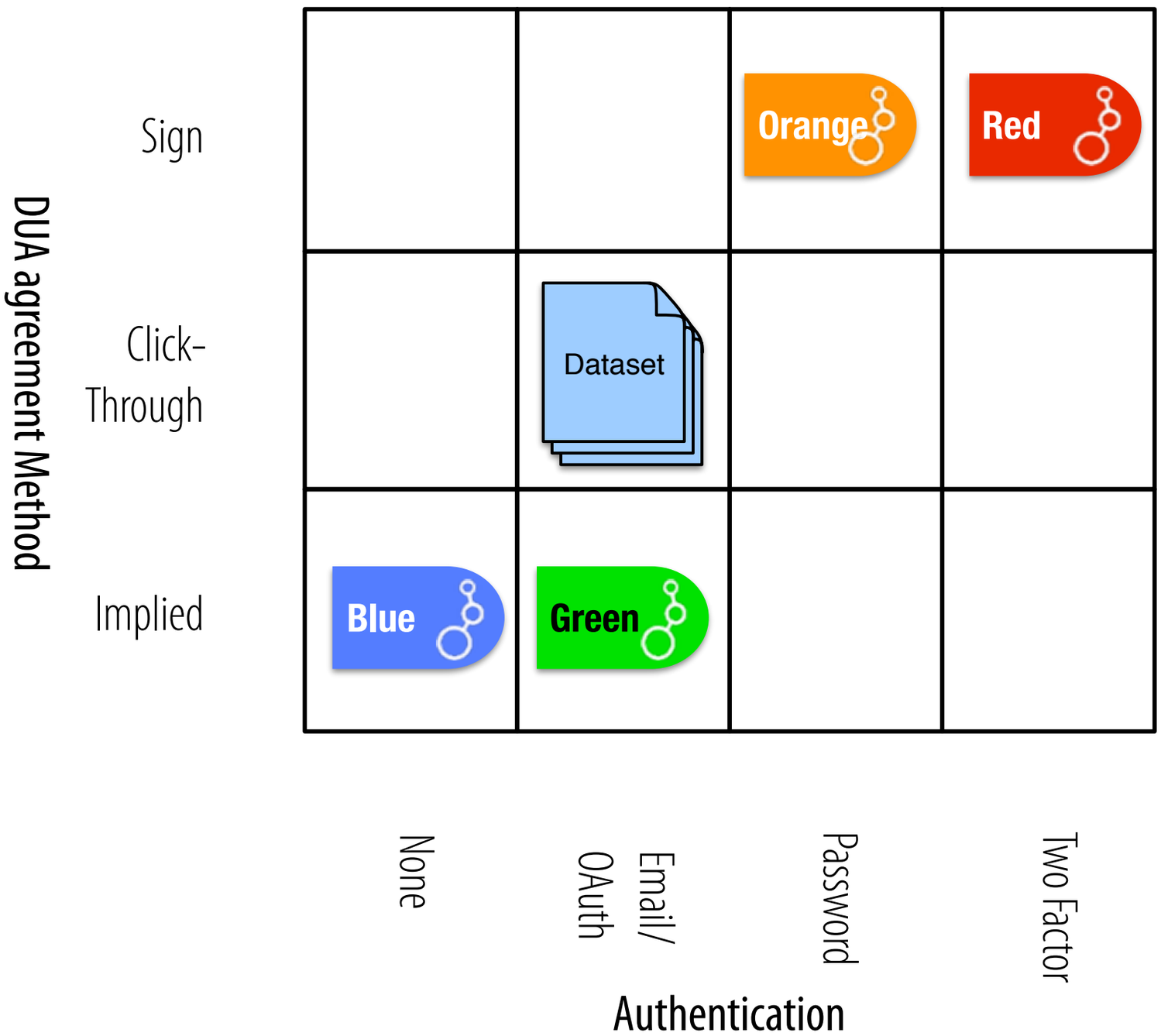
Lenience


DUA agreement Method
Sign
Click-Through
Implied
P
All policies that P does not breach
support(P)
None
Email/ OAuth
Password
Two Factor
Authentication

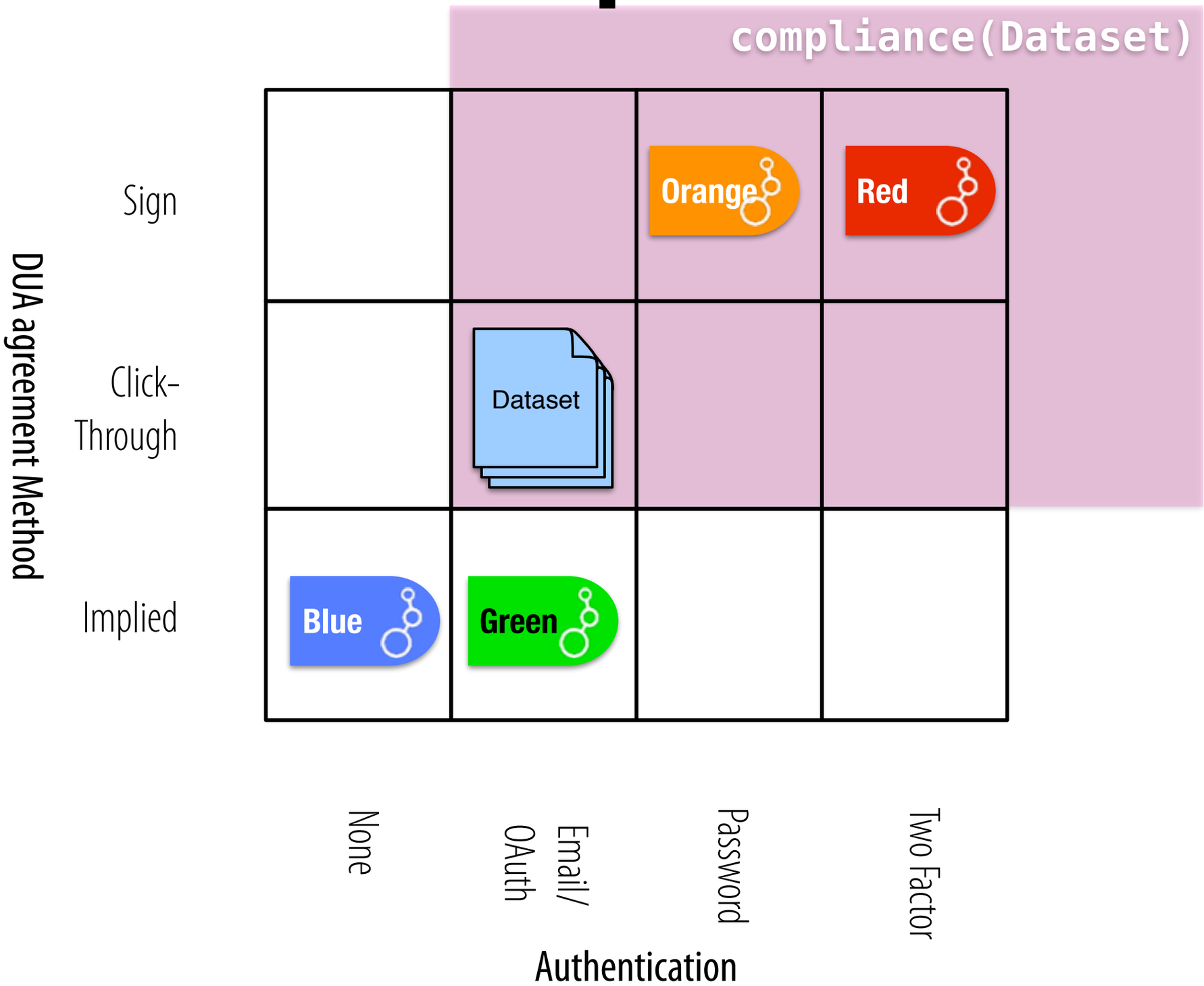
A Dataset and a Repository walk into a DHP space…



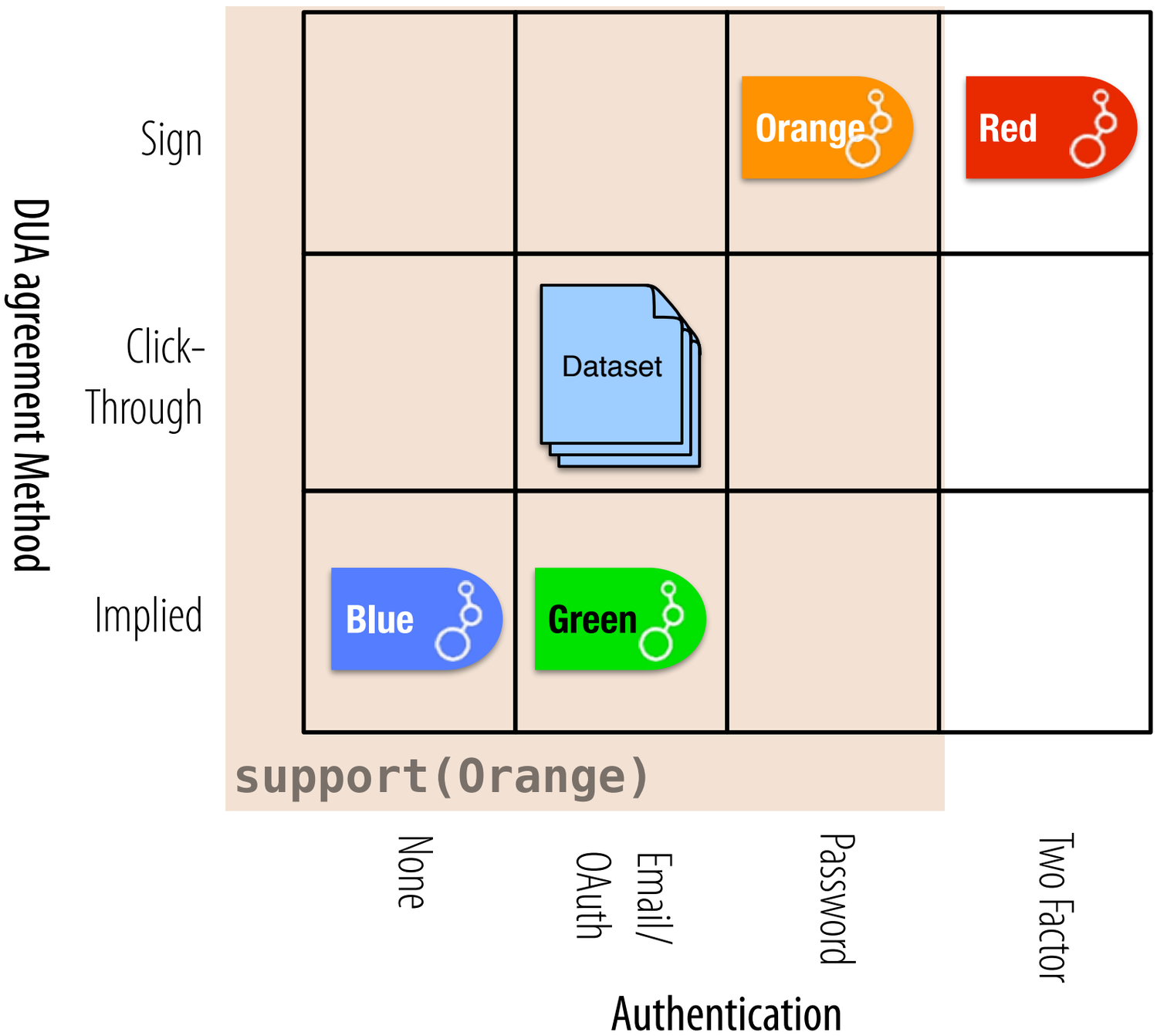
A Dataset and a Repository walk into a DHP space…



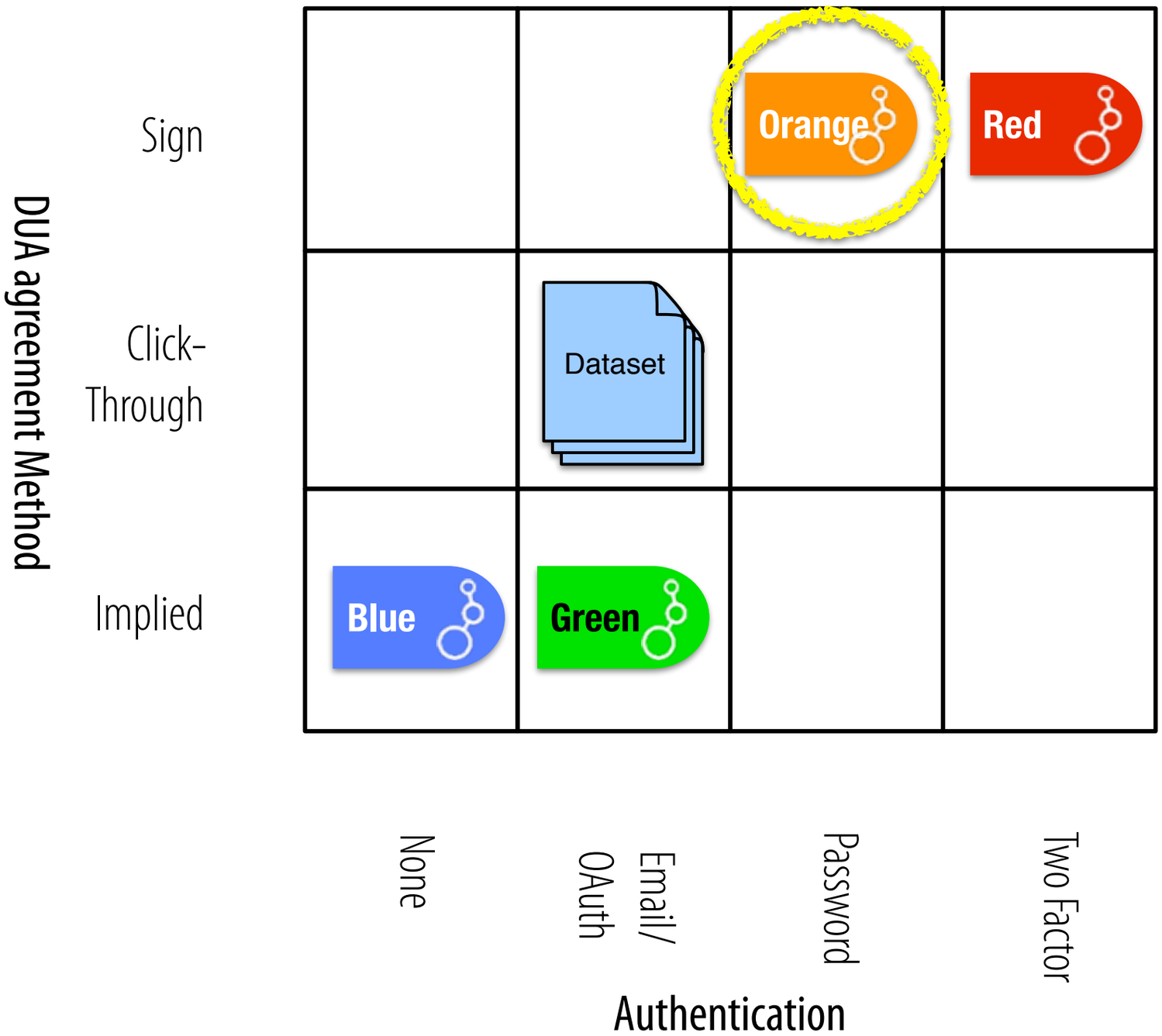
A Dataset and a Repository walk into a DHP space…



A Dataset and a Repository walk into a DHP space…



A Dataset and a Repository walk into a DHP space…



DataTagsTools

Open source *on* GitHub

Tag Space

A tag space is a hierarchical structure that defined a DHP space, with some assertion dimensions added.

BlueToCrimson.ts

```
<*
This is the tag space for the DataTags set proposed at:
 Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai. Sharing sensitive data with conf
 Science, 2015.
*>

DataTags: consists of Security, AccessCredentials. <-- This is the top-level slot

Security: consists of Storage, Transmit.

AccessCredentials: consists of Authentication, Registration, Approval, DUAAcceptance.

Storage[How are data stored on disk]: one of
 clear [No encryption used],
 encrypt [Data are stored encrypted on disk],
 multiEncrypt [Data are encrypted on disk, in a way that the server cannot unencrypt th

Transmit[How are data travelling through networks]:
  one of clear, encrypt.

Authentication: one of none, password, twoFactor.
```

Atom package screenshot: Gal Maman, Matan Toledano, BGU

Tag Space

BlueToCrimson.ts

```
<*
This is the tag space for the DataTags set proposed at:
 Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai. Sharing sensitive data with conf
 Science, 2015.
*>

DataTags: consists of Security, AccessCredentials

Security: consists of Storage, Transmit.

AccessCredentials: consists of Authentication, Registration, Approval, DUAAcceptance.

Storage[How are data stored on disk]: one of
 clear [No encryption used],
 encrypt [Data are stored encrypted on disk],
 multiEncrypt [Data are encrypted on disk, in a way that the server cannot unencrypt th

Transmit[How are data travelling through networks]:
  one of clear, encrypt.

Authentication: one of none, password, twoFactor.
```

Block Comment

```
<-- This is the top-level slot
```

Line Comment

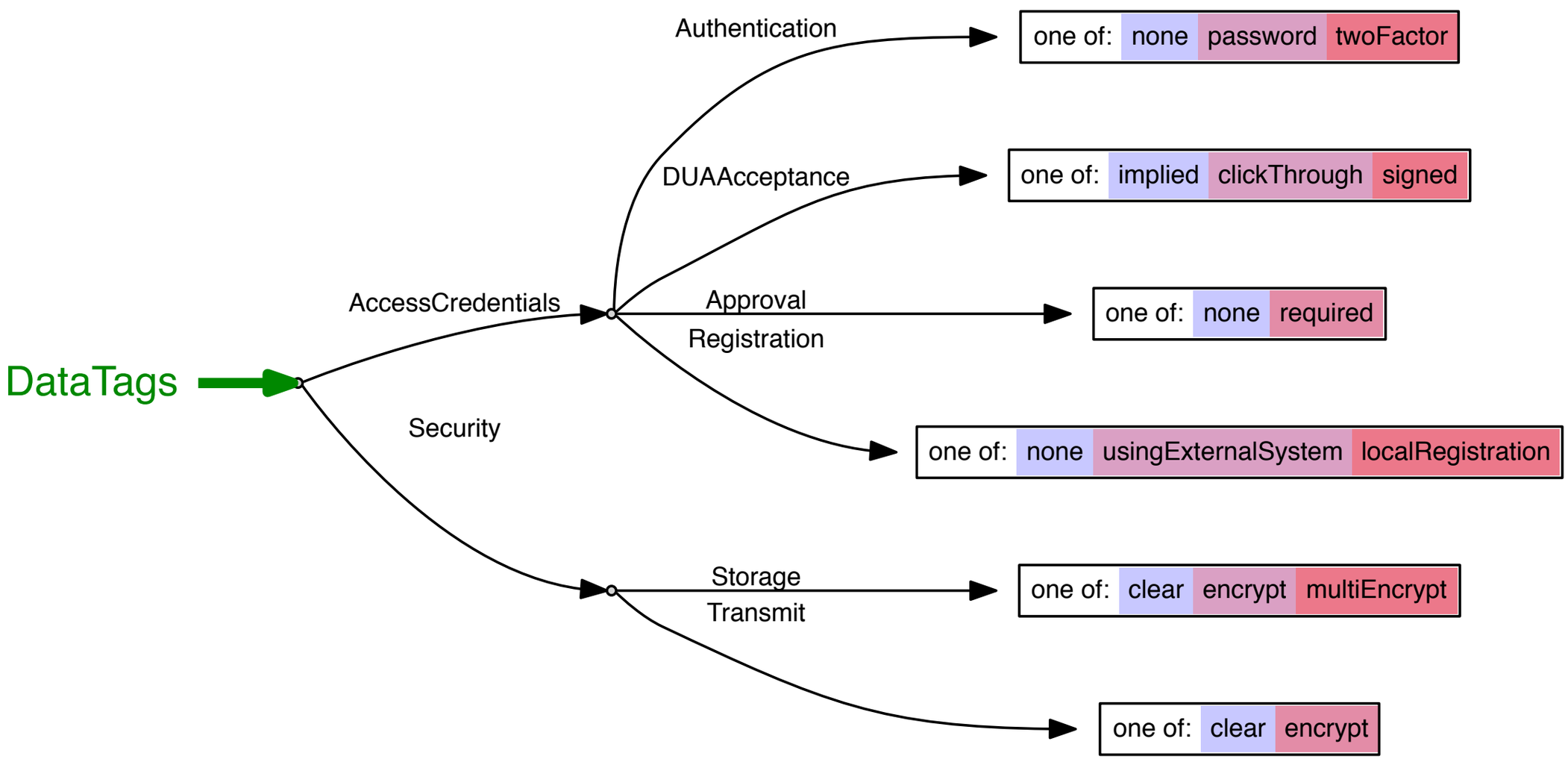
Compound Slot

Description

Atomic Slot

Atom package screenshot: Gal Maman, Matan Toledano, BGU

Tag-Space Visualized



Visualization using CliRunner (on a later slide) and Graphviz (www.graphviz.org).

Arriving at a DHP

www.datatags.org/interviews/questionnaireId/q/MR7

DataTags Feedback IQSS

Question: Please select one answer

Do the data contain information from a covered entity or business associate of a covered entity?

Terms

Business associate
A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.

Covered entity
A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

Yes | Not Sure | No

Answer Feed

Do the data contain information related to substance abuse diagnosis, referral, or treatment? No Revisit

Current Tags

DataTags

Arriving at a DHP

www.datatags.org/interviews/questionnaireId/accept

DataTags Feedback IQSS

Dataset Can be Accepted

Your dataset is tagged as Orange

May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement.

DataTags

Legal

MedicalRecords

HIPAA safeHarborDeidentified

EducationRecords

PPRA protectedDeidentified consent

ContractOrPolicy no

GovernmentRecords

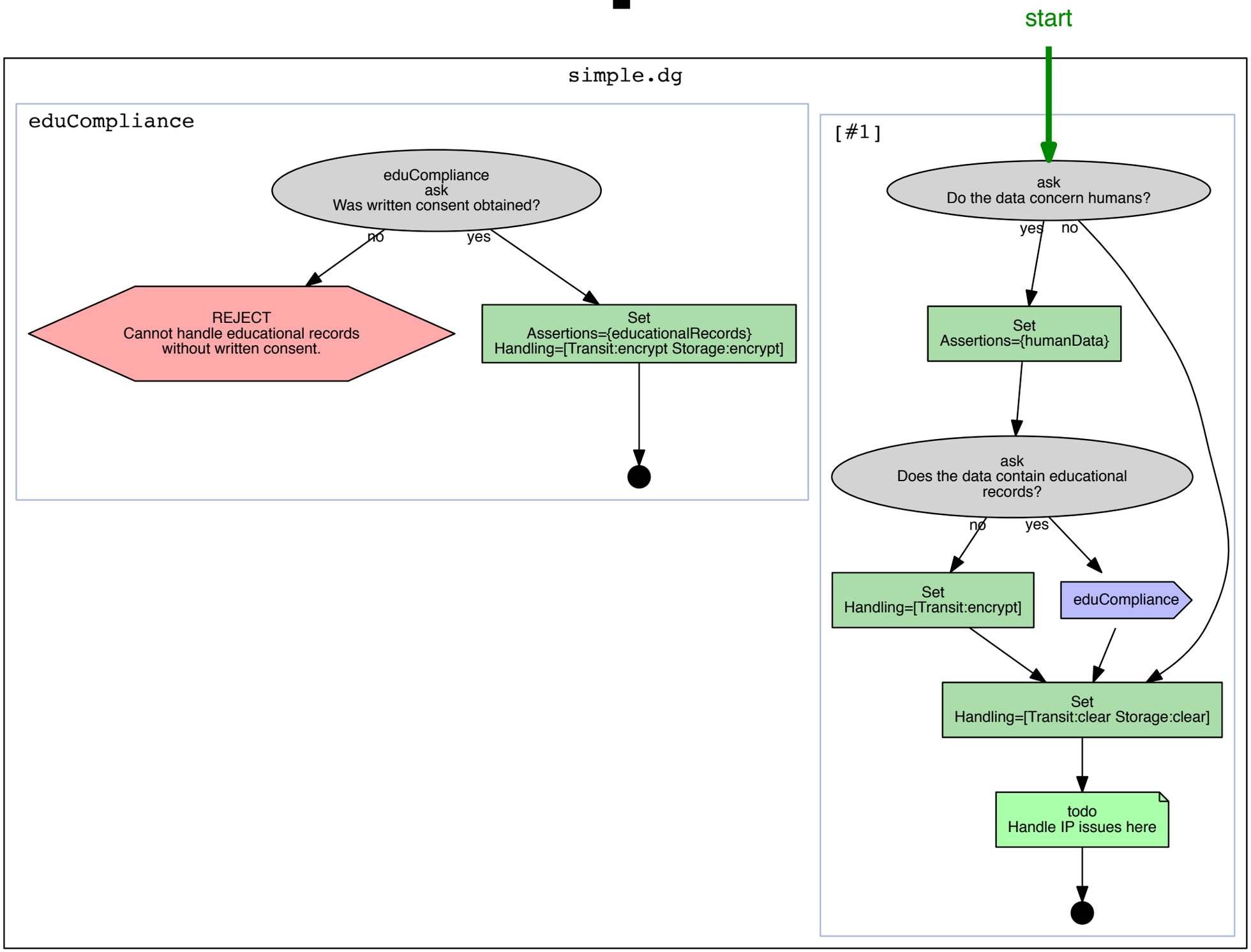
Tags Questionnaire

- “Interview with an expert” metaphor
- Consists of a tag space and a *decision graph*

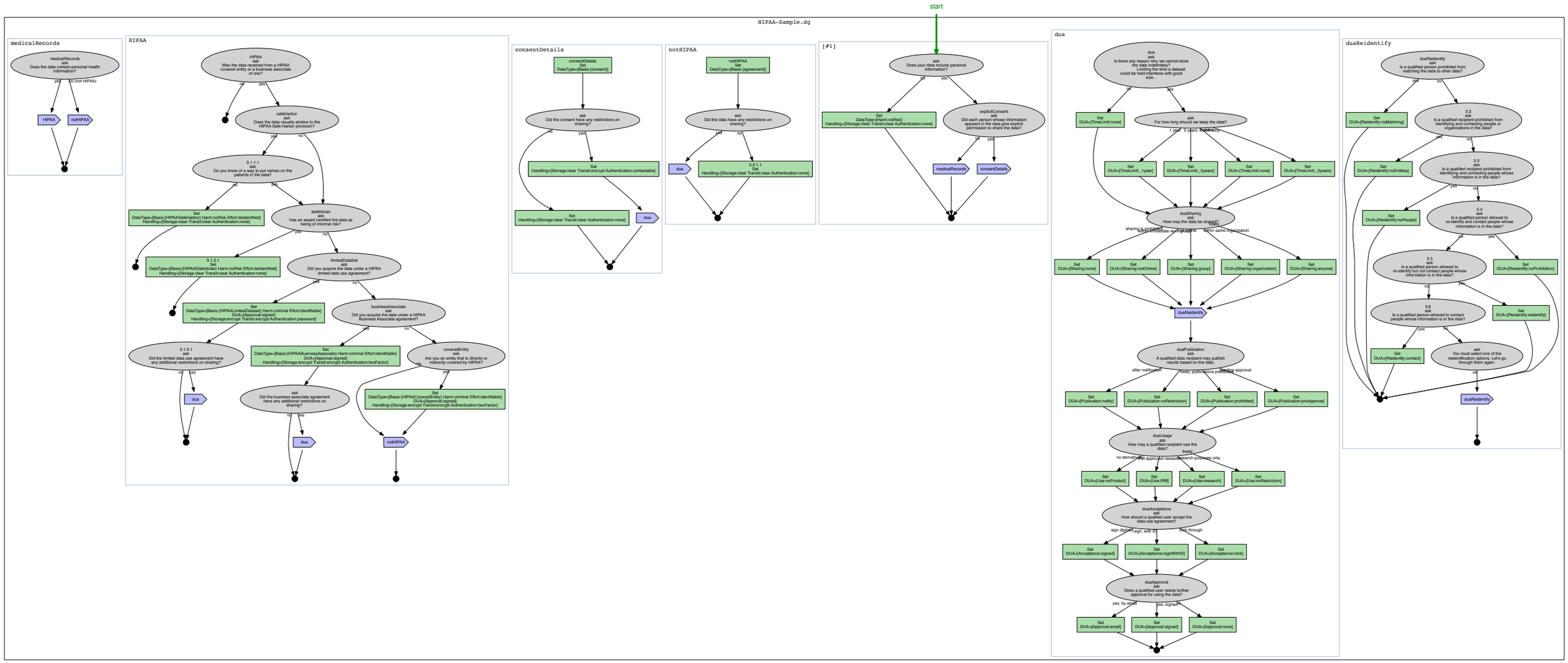
simple.dg

```
[ask:
  {text: Do the data concern humans?}
  {answers:
    {yes: [set: Assertions+=humanData]
          [ask:
            {text: Does the data contain educational records?}
            {answers:
              {yes: [call: eduCompliance]}
              {no:  [set: Transit=encrypt]}
            }]}
  }
]
[set: Storage=clear; Transit=clear] <-- defaults
[todo: Handle IP issues here]
[end]
<* Educational Compliance Section *>
[>eduCompliance< ask:
```

Decision Graph - Visualized

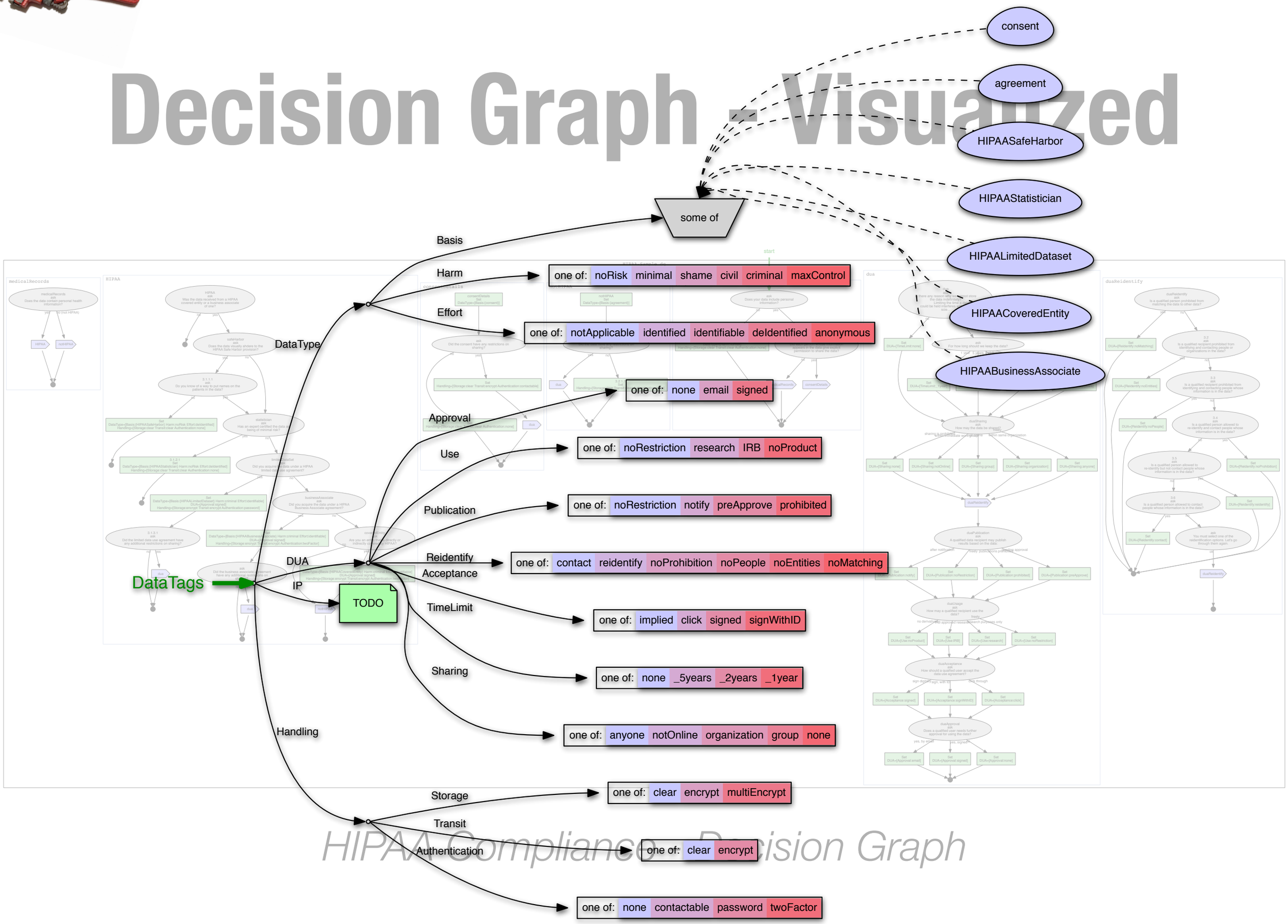


Decision Graph - Visualized



HIPAA Compliance - Decision Graph

Decision Graph - Visualized



HIPAA Compliance - Decision Graph

CliRunner

- Questionnaire Development Console
- Run, debug, visualize
- **Query:**
  What answer sequences result in encryption=clear, harm=severe?

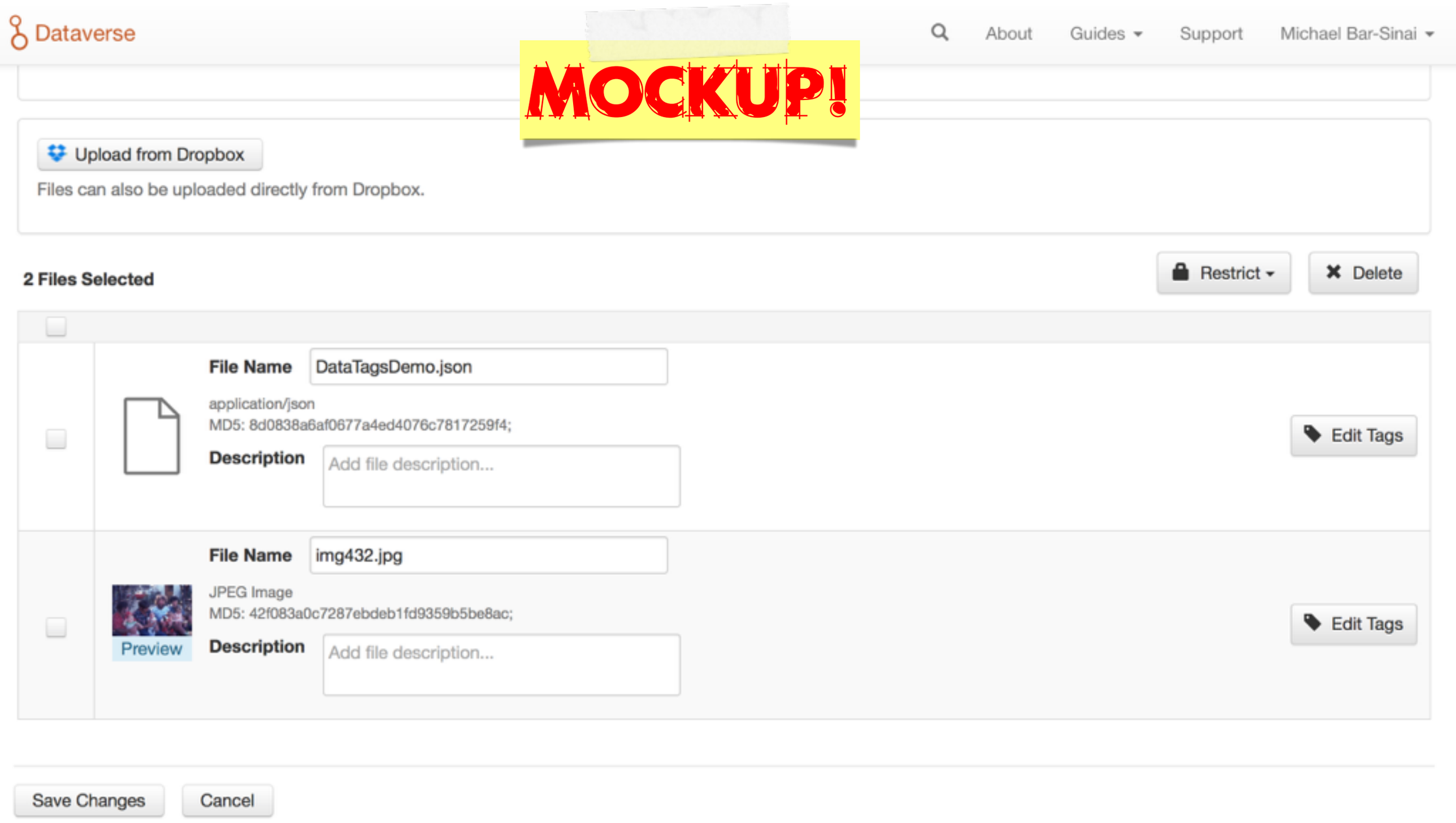
Collaborating

- Can use existing social coding tools

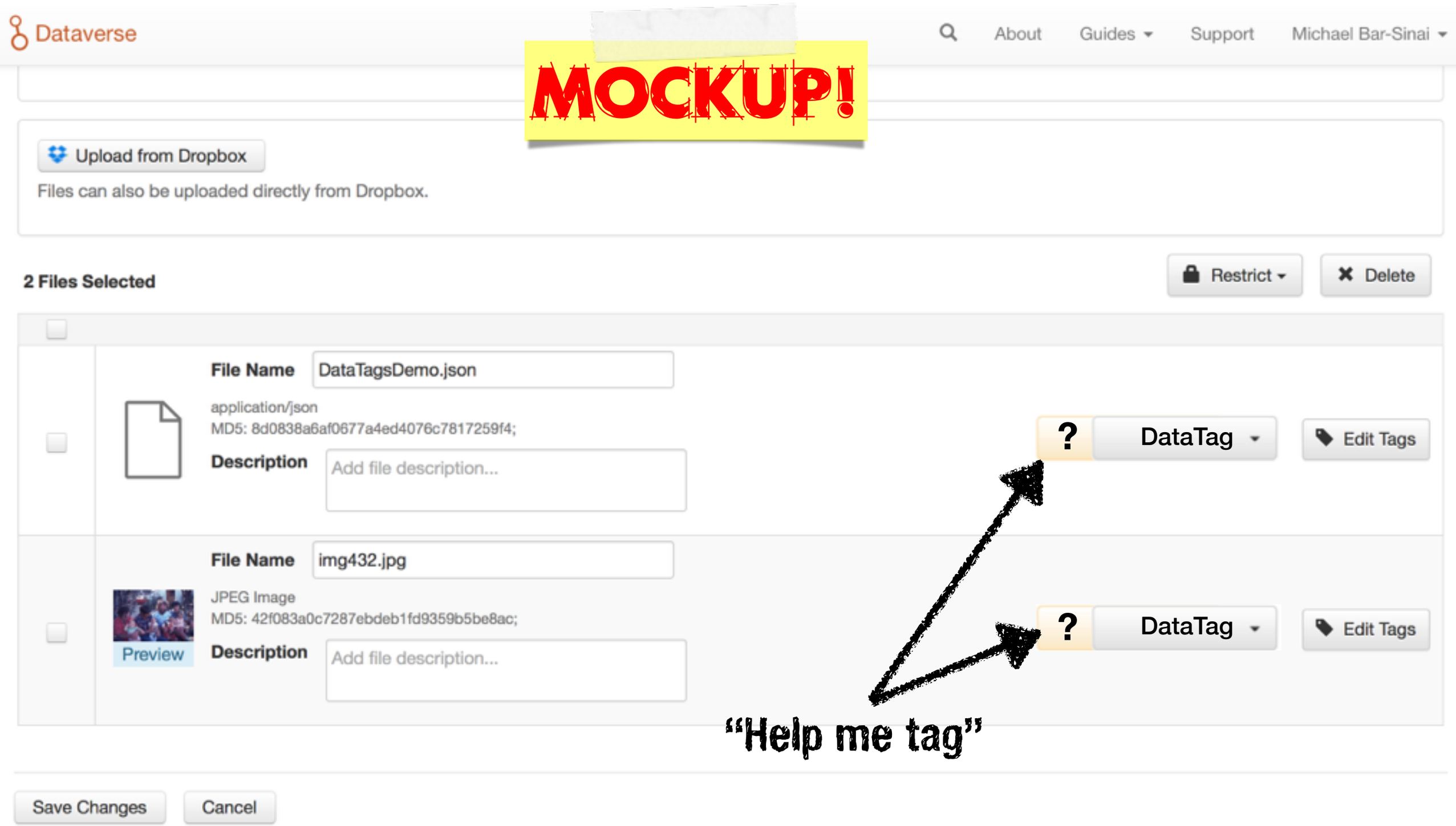
Screenshot: GitHub line comment

- Future Work: collaborative development environment

Dataverse Integration - Upload

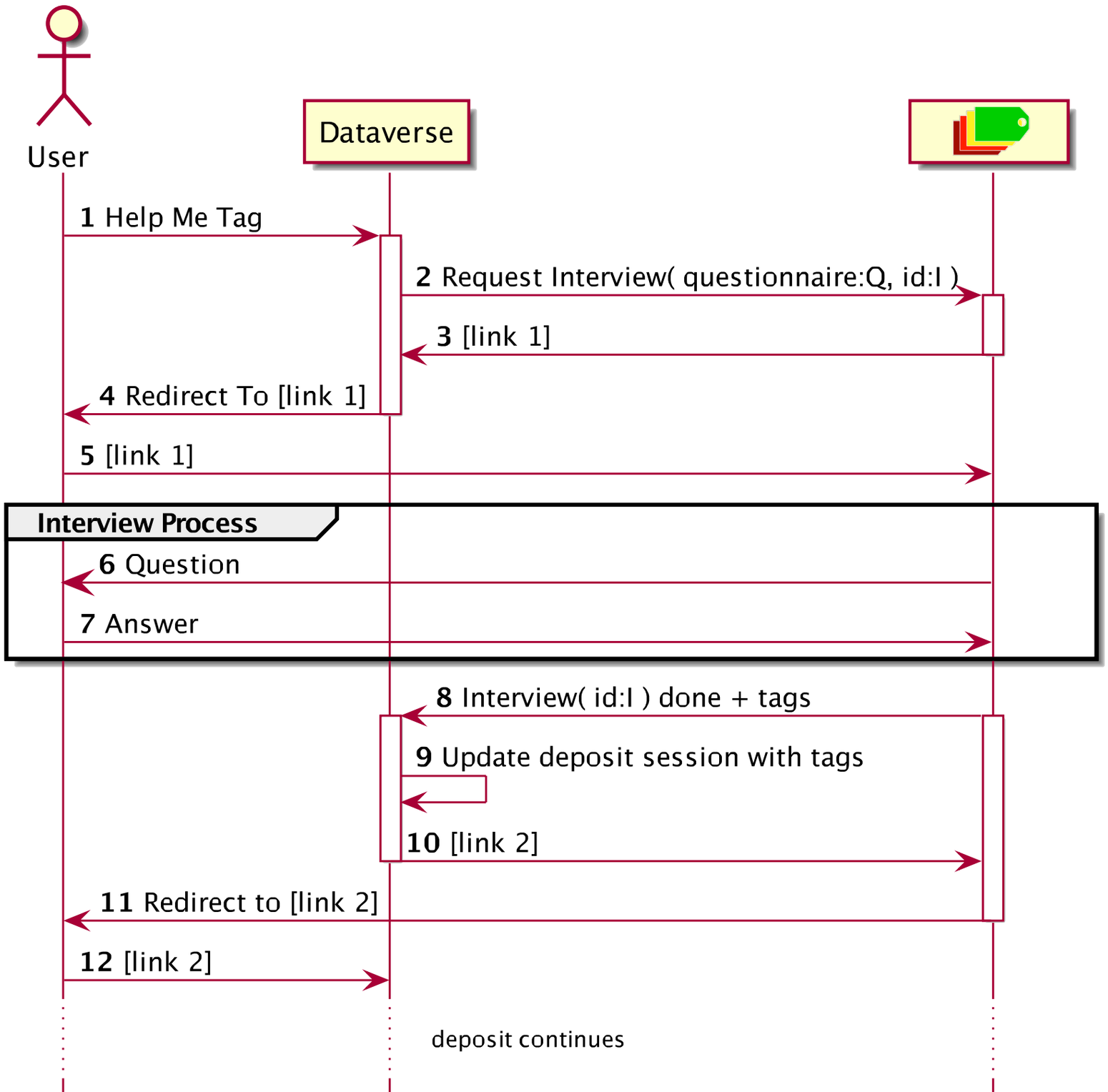


Dataverse Integration - Upload



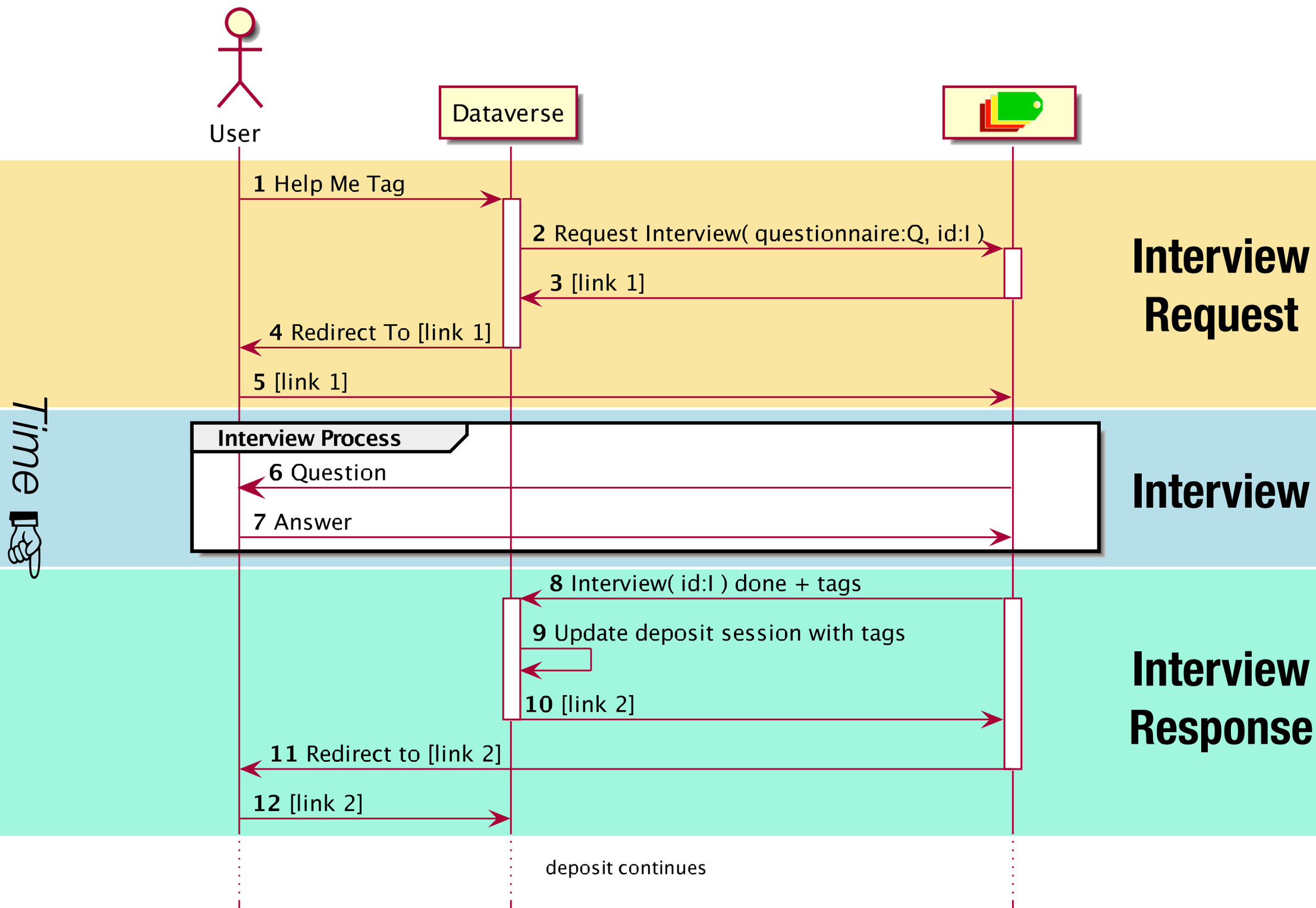
Dataverse Integration

Behind the scenes



Dataverse Integration

Behind the scenes



Dataverse Integration - View

Dataverse

About Guides Support Michael Bar-Sinai

MOCKUP!

Description This dataset contai referred to in the paper. They were not included in the paper, as it could also be understood without them. Thus, this dataset could be viewed as appendix for those who would like to deepen their understanding of the paper.

Subject Law; Computer and Information Science

Files | Metadata | Terms | Versions

Search this dataset... Find

14 Files

Upload Files | Edit Files | Download

algorithm-trace.html
text/html - 4.5 KB - Feb 7, 2016 - 0 Downloads
MD5: 61c867f8f541b4b72034033654f94e4a;
Viewable trace of the slot unfolding algorithm.
Download

algorithm-trace.md
text/markdown - 1.4 KB - Feb 7, 2016 - 0 Downloads
MD5: 23200ec0863f51c3ab8b6a497c4ca386;
Trace of the slot unfolding algorithm (editable markdown form).
Download

BlueToCrimson-ts.pdf
Adobe PDF - 16.6 KB - Feb 8, 2016 - 0 Downloads
MD5: d58367787995cd2f3268f6c537f3dd7e;
Visualization of the tagspace for the datatags set defined in
http://techscience.org/a/2015101601/
Download

BlueToCrimson.dg
Unknown - 1.7 KB - Feb 8, 2016 - 0 Downloads
MD5: d32d25bfcf80c4c7debc2023a5cc9345;
Download

Dataverse Integration - View

Dataverse

About Guides Support Michael Bar-Sinai

MOCKUP!

Description This dataset contai referred to in the paper. They were not included in the paper, as it could also be understood without them. Thus, this dataset could be viewed as appendix for those who would like to deepen their understanding of the paper.

Subject Law; Computer and Information Science

Orange

Files Metadata Terms Versions

Search this dataset... Find

14 Files

Upload Files Edit Files Download

algorithm-trace.html
text/html - 4.5 KB - Feb 7, 2016 - 0 Downloads
MD5: 61c867f8f541b4b72034033654f94e4a;
Viewable trace of the slot unfolding algorithm.

Green Download

algorithm-trace.md
text/markdown - 1.4 KB - Feb 7, 2016 - 0 Downloads
MD5: 23200ec0863f51c3ab8b6a497c4ca386;
Trace of the slot unfolding algorithm (editable markdown form).

Orange Download

BlueToCrimson-ts.pdf
Adobe PDF - 16.6 KB - Feb 8, 2016 - 0 Downloads
MD5: d58367787995cd2f3268f6c537f3dd7e;
Visualization of the tagspace for the datatags set defined in
http://techscience.org/a/2015101601/

Blue Download

BlueToCrimson.dg
Unknown - 1.7 KB - Feb 8, 2016 - 0 Downloads
MD5: d32d25bfcf80c4c7debc2023a5cc9345;

Blue

Privacy Tools
for Sharing Research Data
A National Science Foundation
Secure and Trustworthy Cyberspace Project

Ben-Gurion University
of the Negev

DATA PRIVACY LAB

The Institute
for Quantitative Social Science
at Harvard University

```
[set: Thank+=you]
[end]
```

I Data

http://datatags.org

http://datascience.iq.harvard.edu/about-datatags

This work was funded by grant CNS-1237235 from the National Science Foundation.