IQSS
The Institute for Quantitative Social Science

The Dataverse Project

Sensibly Storing Sensitive Stuff

Gustavo Durand, Technical Lead / Architect, IQSS. gdurand@iq.harvard.edu
Michael Bar-Sinai, PhD Candidate, Ben Gurion University
Fellow, IQSS.
@michbarsinai m@mbarsinai.com
Tania Schlatter, UX & UI lead, IQSS

Dataverse

Overview

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard’s Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- 15 on the core team - developers, UX & UI designers and researchers, metadata specialists, curation manager

Dataverse Features - Data

- Persistent IDs / URLs
  - DataCite
  - Handle
- Automatically Generated Citations with attribution
- Compliant with FAIR and data citation principles
- Domain-specific Metadata
- Versioning
- File Storage
  - Local
  - Swift (OpenStack)
  - S3 (Amazon)

Dataverse Features - Users

- Multiple Sign In options
    - Native
    - Shibboleth
    - OAuth (ORCID)
- Dataverses within Dataverses
- Branding
- Widgets

Dataverse Features - Workflows

- Permissions
- Access Controls and Terms of Use
- Publishing Workflows
- Private URLs
- Upload / Download Workflows
  - Browser
  - Dropbox
  - Rsync (for big data “packages”)

Dataverse Features - Interoperability

- APIs
    - SWORD
    - Native
- Harvesting (OAI-PMH)
    - Client
    - Server
- Modular External Tools
    - Explore
    - Configure

Dataverse Technology

Glassfish Server 4.1

Java SE8

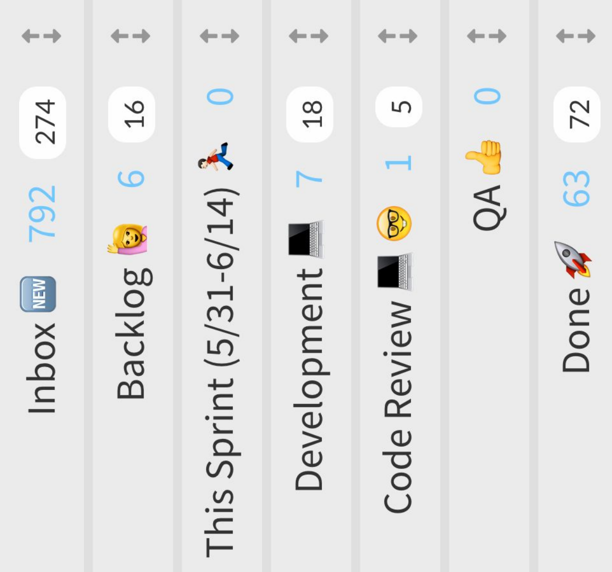
Java EE7

- Presentation: JSF (PrimeFaces), RESTful API
- Business: EJB, Transactions, Asynchronous, Timers
- Storage: JPA (Entities), Bean Validation

Storage: Postgres, Solr, File System / Swift / S3

Dataverse Development Process

- Inbox
- Backlog
- This Sprint
- Development
- Code Review
- QA
- Done



Inbox 🆕 792 274

Backlog 🙋 6 16

This Sprint (5/31-6/14) 🏃 0

Development 💻 7 18

Code Review 💻 🤓 1 5

QA 👍 0

Done 🚀 63 72

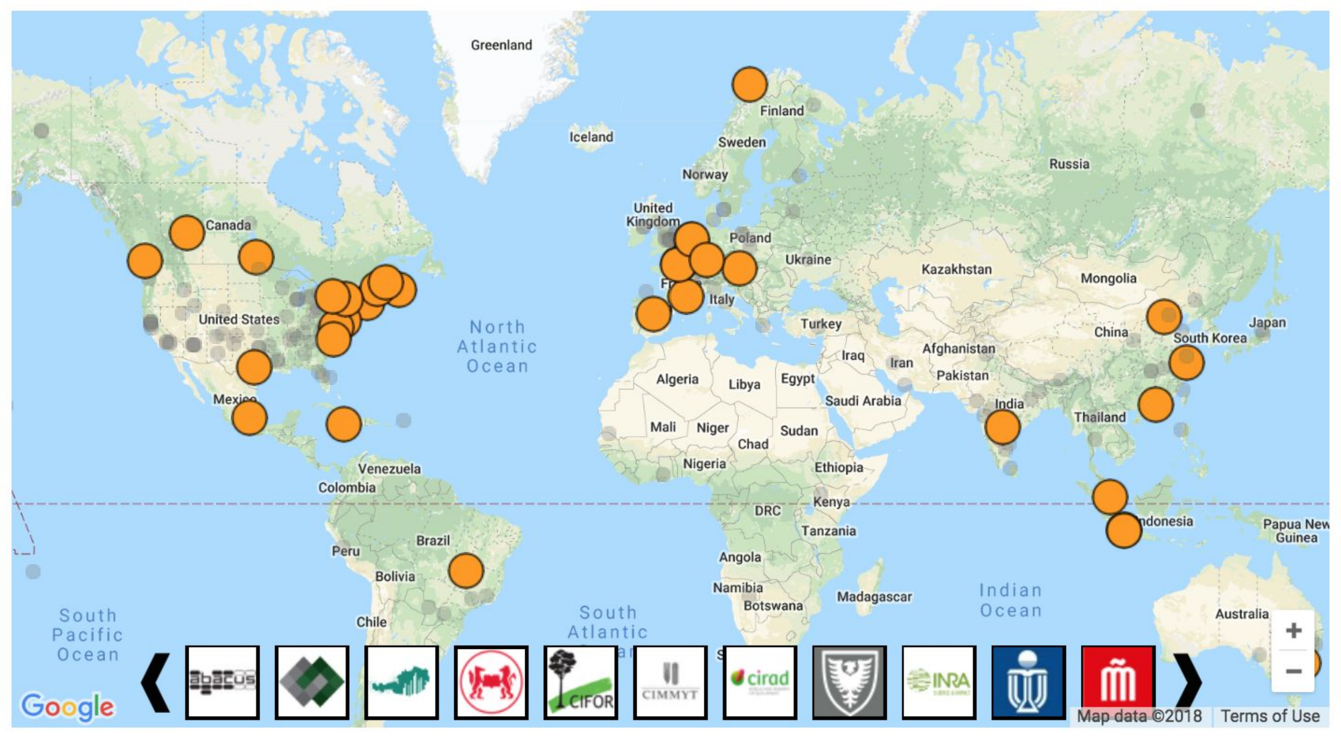
https://waffle.io/IQSS/dataverse

(some) Collaborations

- SBGrid Data
  - Large Data and Support
- Massachusetts Open Cloud
  - Big Data Storage and Compute Access (OpenStack)
- DANS/CIMMYT
  - Handles Support
- ResearchSpace
  - API Java Client Library
- Provenance
  - W3C PROV

Dataverse Community

- 34 installations around the world



Dataverse Community

- 75+ code contributors outside of the Core Team
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
    - Dataverse Google Group
    - Dataverse Community Calls
    - Dataverse Community Meeting
- Global Dataverse Community Consortium

Sensitive Data

What are We Talking About When we Talk About **Sensitive Data**

Broadly speaking:

Data is sensitive if it can cause harm

Wait, there’s more

What is “harm”?

Harm to whom?

Not necessarily the data subject

Data is sensitive if it can **cause harm**

Legal and contractual restrictions may apply

Legal issues vary by country (and state)

The **metadata** may be more sensitive than the data

Can't we Simply Anonymize Data?

87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex [Sweeney, 2000]

De-anonymization of anonymized Netflix data [Narayanan, Shmatikov, 2008]

We used 15 months of data from 1.5 million people to show that 4 points--approximate places and times--are enough to identify 95% of individuals in a mobility database.

[de Montjoye, Hidalgo, Blondel, Verleysen, 2013]

No.

Can’t we Simply Anonymize Data?

No.

- Linkage attacks (Sex, DoB, Zip-code) – enough for identification [Sweeny 00]
- AOL Debacle [2006]
- Netflix award [Narayanan, Shmatikov 08] Anonymized bi-partite graph of users and movies
- Social networks [Backstrom, Dwork, Kleinberg 07]
- Genetic data (GWAS) [Homer, Szelinger, Redman, Duggan, Tembe, Muehling, Pearson,Stephan, Nelson,Craig 08]
- Microtargeted advertising Facebook advertising mechanism, aimed at providing privacy, actually violates privacy [Korolova 11]
- Recommendation Systems [Calandrino, Kiltzer, Naryanan, Felten, Shmatikov 11]
- Israeli CBS [Mukatren, Nissim, Salman, Tromer]
- New York Taxicabs [NeustarResearch 14]
- Multiple innocuous queries (queries are random) [Dinur, Nissim 03], [Dwork, McSherry, Talwar 07]
- Unique in the Shopping Mall: Reidentifying credit card data [de Montejoye, Pentland, Radaelli, Singh, 2015]

In summing up,

It’s complicated

Application Developers **Have to Face** These Challenges

Growing Regulations:

HIPAA, CalOPPA, GDPR…

Growing User Concern:

E.g. Facebook & Cambridge Analytica

Classifying Data By Harm

Harvard's Five Security Levels

Assigned by IRB/Researcher

Level 1
Public Information

Level 2
Information the University has chosen to keep confidential but the disclosure of which would not cause material harm

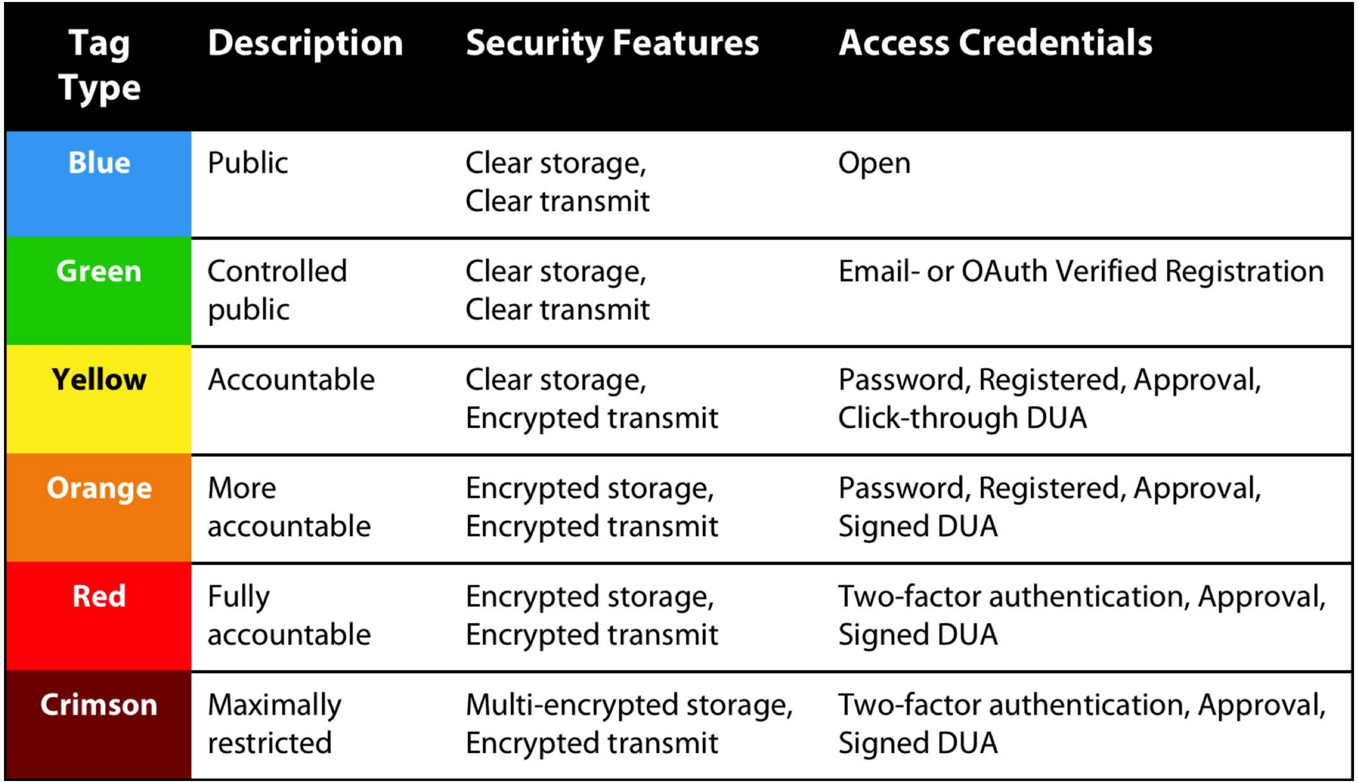
Level 3
Information that could cause risk of material harm to individuals or the University if disclosed

Level 4
information would likely cause serious harm to individuals or the University if disclosed

Level 5
Information that would cause severe harm to individuals or the University if disclosed

Source: https://policy.security.harvard.edu/view-data-security-level

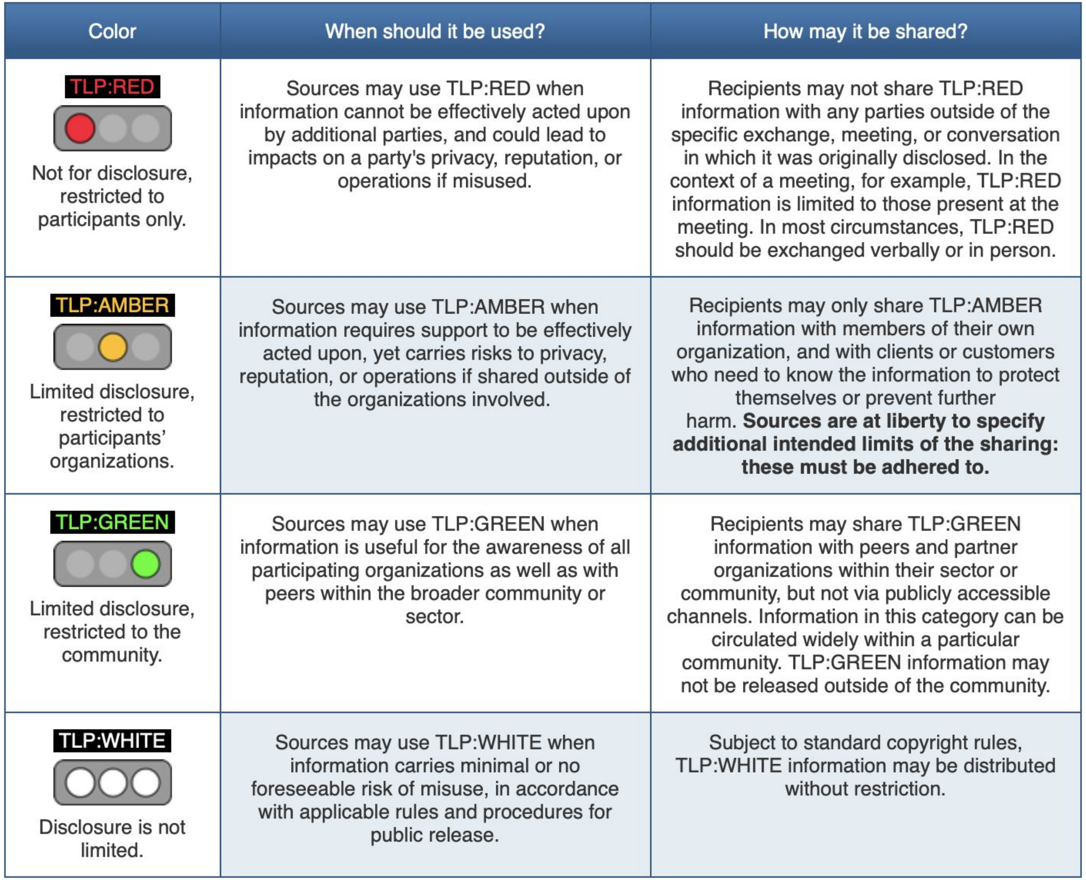
DataTags Levels



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

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

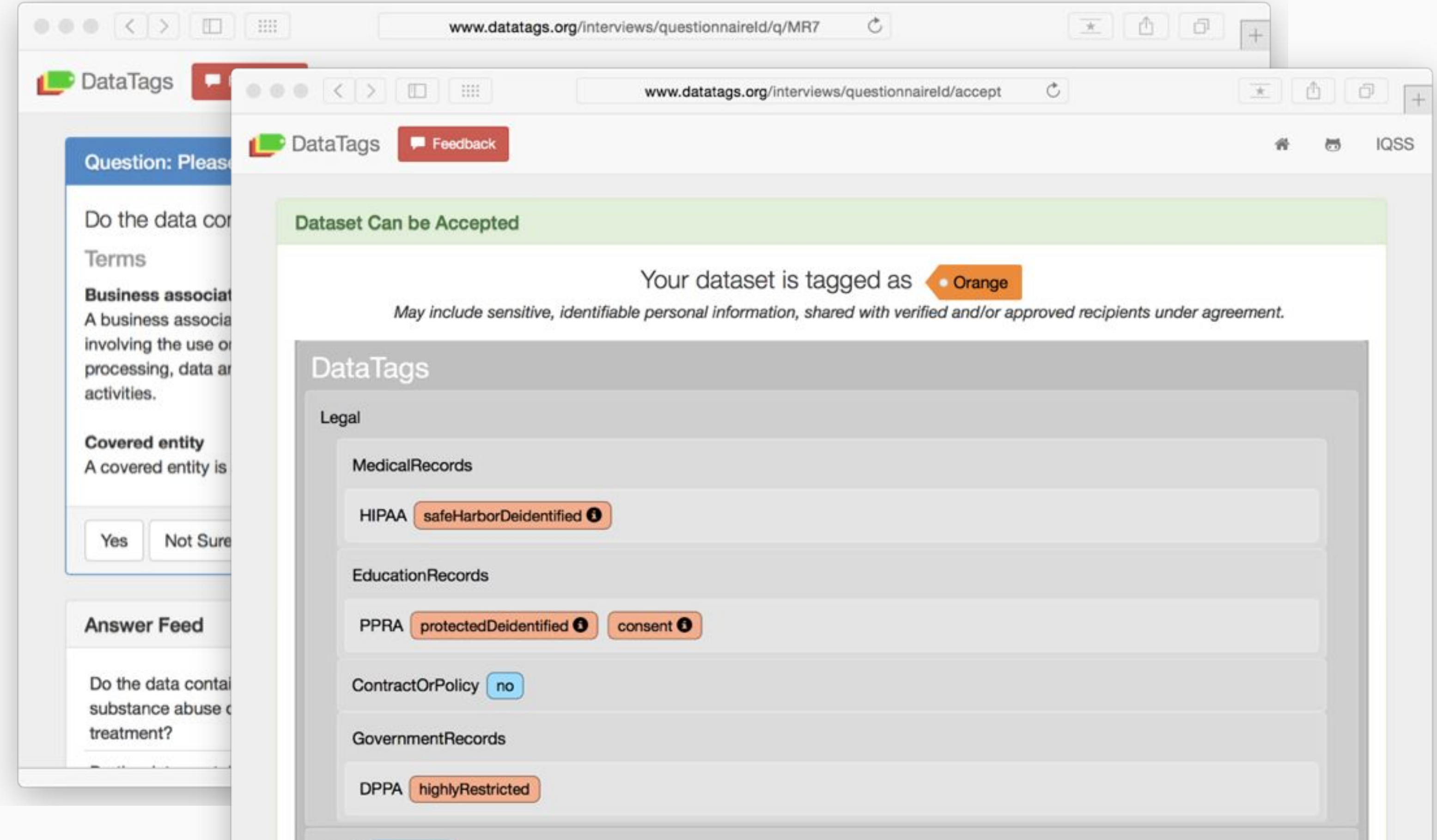
US-CERT Traffic Light Protocol



Color	When should it be used?	How may it be shared?
TLP:RED Not for disclosure, restricted to participants only.	Sources may use TLP:RED when information cannot be effectively acted upon by additional parties, and could lead to impacts on a party's privacy, reputation, or operations if misused.	Recipients may not share TLP:RED information with any parties outside of the specific exchange, meeting, or conversation in which it was originally disclosed. In the context of a meeting, for example, TLP:RED information is limited to those present at the meeting. In most circumstances, TLP:RED should be exchanged verbally or in person.
TLP:AMBER Limited disclosure, restricted to participants' organizations.	Sources may use TLP:AMBER when information requires support to be effectively acted upon, yet carries risks to privacy, reputation, or operations if shared outside of the organizations involved.	Recipients may only share TLP:AMBER information with members of their own organization, and with clients or customers who need to know the information to protect themselves or prevent further harm. **Sources are at liberty to specify additional intended limits of the sharing: these must be adhered to.**
TLP:GREEN Limited disclosure, restricted to the community.	Sources may use TLP:GREEN when information is useful for the awareness of all participating organizations as well as with peers within the broader community or sector.	Recipients may share TLP:GREEN information with peers and partner organizations within their sector or community, but not via publicly accessible channels. Information in this category can be circulated widely within a particular community. TLP:GREEN information may not be released outside of the community.
TLP:WHITE Disclosure is not limited.	Sources may use TLP:WHITE when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release.	Subject to standard copyright rules, TLP:WHITE information may be distributed without restriction.

Source: https://www.us-cert.gov/tlp

Assigning a Tag Semi-automatically



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

DataTags

Question: Pleas

Do the data co

Terms

Business associat
A business associa
involving the use o
processing, data a
activities.

Covered entity
A covered entity is

Yes | Not Sure

Answer Feed

Do the data contai
substance abuse
treatment?

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

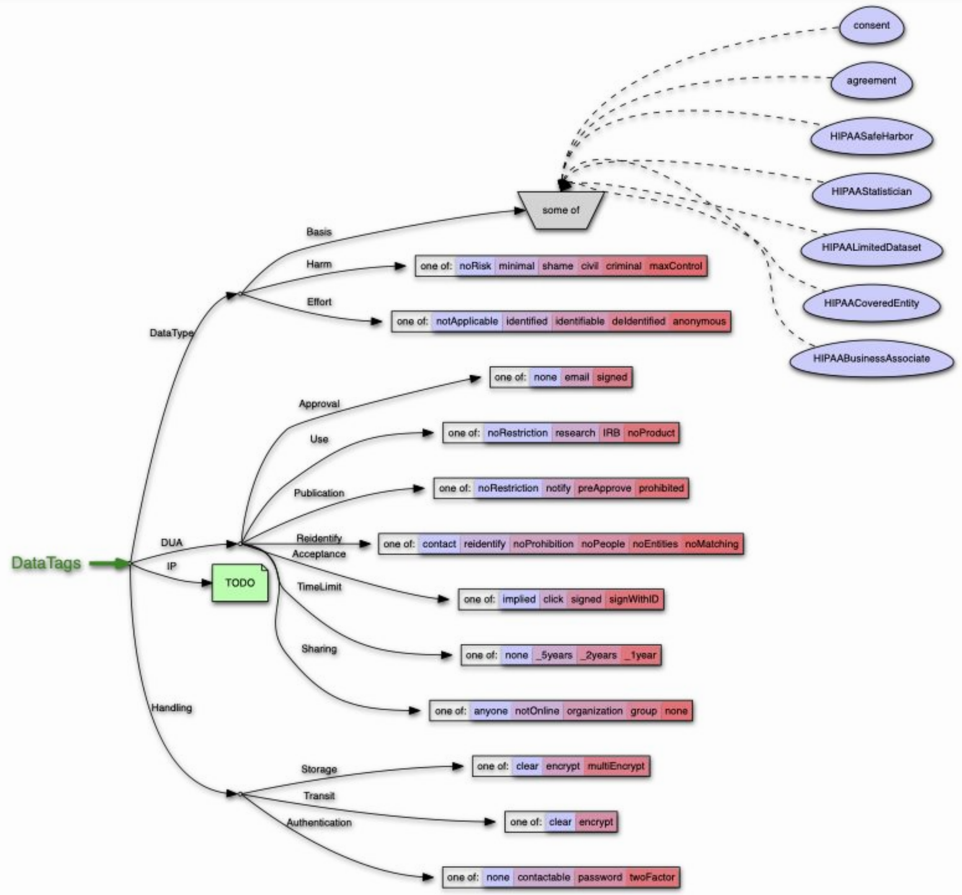
GovernmentRecords

DPPA highlyRestricted

Assigning a Tag Semi-automatically

PolicyModels

HIPAA Compliance - Tag Space



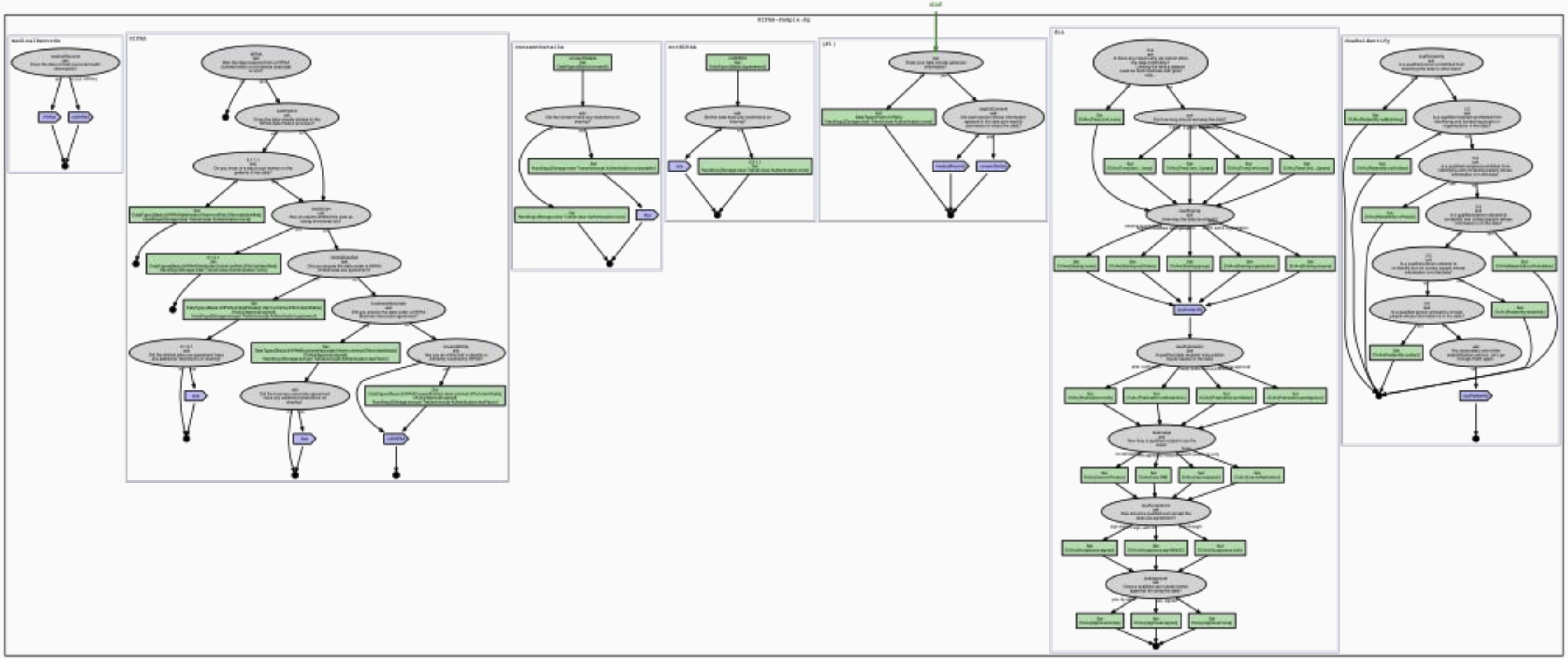
consent
agreement
HIPAASafeHarbor
HIPAAStatistician
HIPAALimitedDataset
HIPAACoveredEntity
HIPAABusinessAssociate
some of
Basis
Harm
one of: noRisk minimal shame civil criminal maxControl
Effort
one of: notApplicable identified identifiable deIdentified anonymous
DataType
Approval
one of: none email signed
Use
one of: noRestriction research IRB noProduct
Publication
one of: noRestriction notify preApprove prohibited
Reidentify
one of: contact reidentify noProhibition noPeople noEntities noMatching
Acceptance
one of: implied click signed signWithID
TimeLimit
one of: none _5years _2years _1year
Sharing
one of: anyone notOnline organization group none
DataTags
DUA
IP
TODO
Handling
Storage
one of: clear encrypt multiEncrypt
Transit
one of: clear encrypt
Authentication
one of: none contactable password twoFactor

Does not replace professional legal advice!

Assigning a Tag Semi-automatically

PolicyModels

HIPAA Compliance - Decision Graph

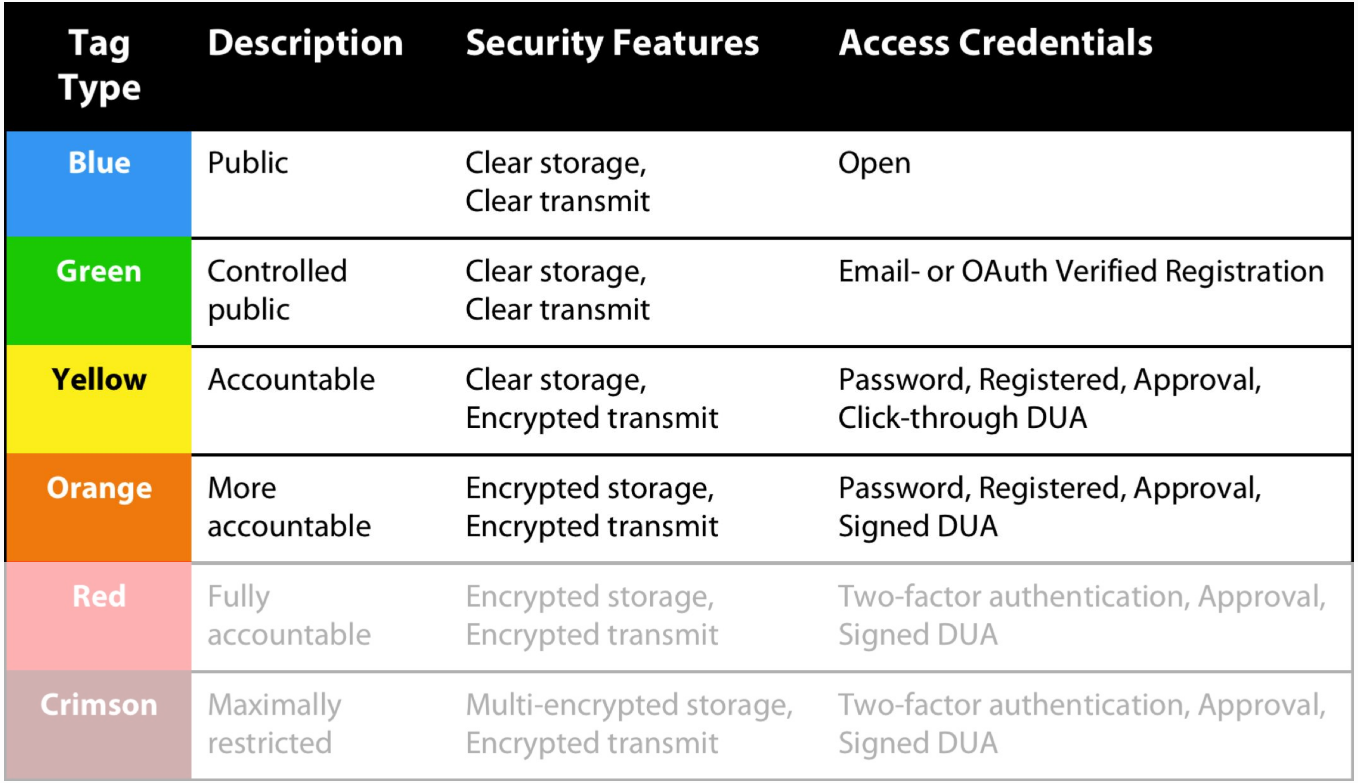


Does not replace professional legal advice!

https://github.com/IQSS/DataTaggingLibrary

What We Are Doing

Goal for Next Year



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

Infrastructure

- Encrypted Transit
  - Inter-server communication must be “protected”
- Encrypted Storage (#4113, #4379)
- Require verification of e-mail address (#3300)
- Complex Passwords
  - :PVMinLength, :PVMaxLength
  - :PVCharacterRules, :PVNumberOfCharacteristics
  - :PVDictionaries
  - :PVGoodStrength
- Mitigate against password guessing
- Bulk Removal of Roles / Permissions

**non-exhaustive list*

Dev<->Ops<->Management

- Server operators need to know who is in charge of the application
- Must prove “business need” to access server/application
- Software must be patched and updated regularly
- Server must run updated maleware-detection software
- Periodical scans
  - “high risk confidential information”
  - Activity logs
- Training and screening of system administrators
- Protection for server *and backups*. Delete data prior to disposing devices.
- Logging of user and administrative access
- Any actual or suspected loss, theft, or improper use of or access to confidential information must be reported promptly.

Confidential information must only be accessed for authorized purposes

**non-exhaustive list*

Engineering

- Administrators should not be able to retrieve user passwords
- Force re-identification after idle period
- Is it OK to have a super-user in your application?
- Allow API keys in HTTP headers, not just as query parameters
- Multi Party encryption
- Multiple Storage Back-Ends

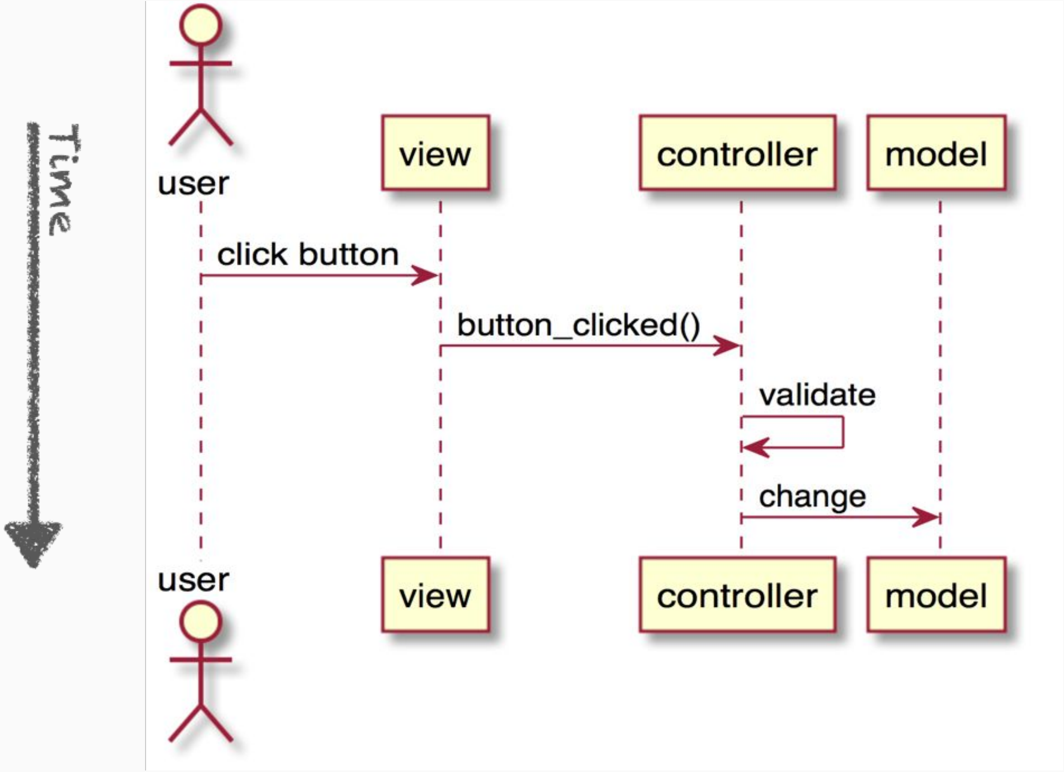
Bake security into application architecture from the start
Use common-sense and healthy paranoia
(so your users won’t have to)

Engineering - Command Design Pattern with Permission Metadata

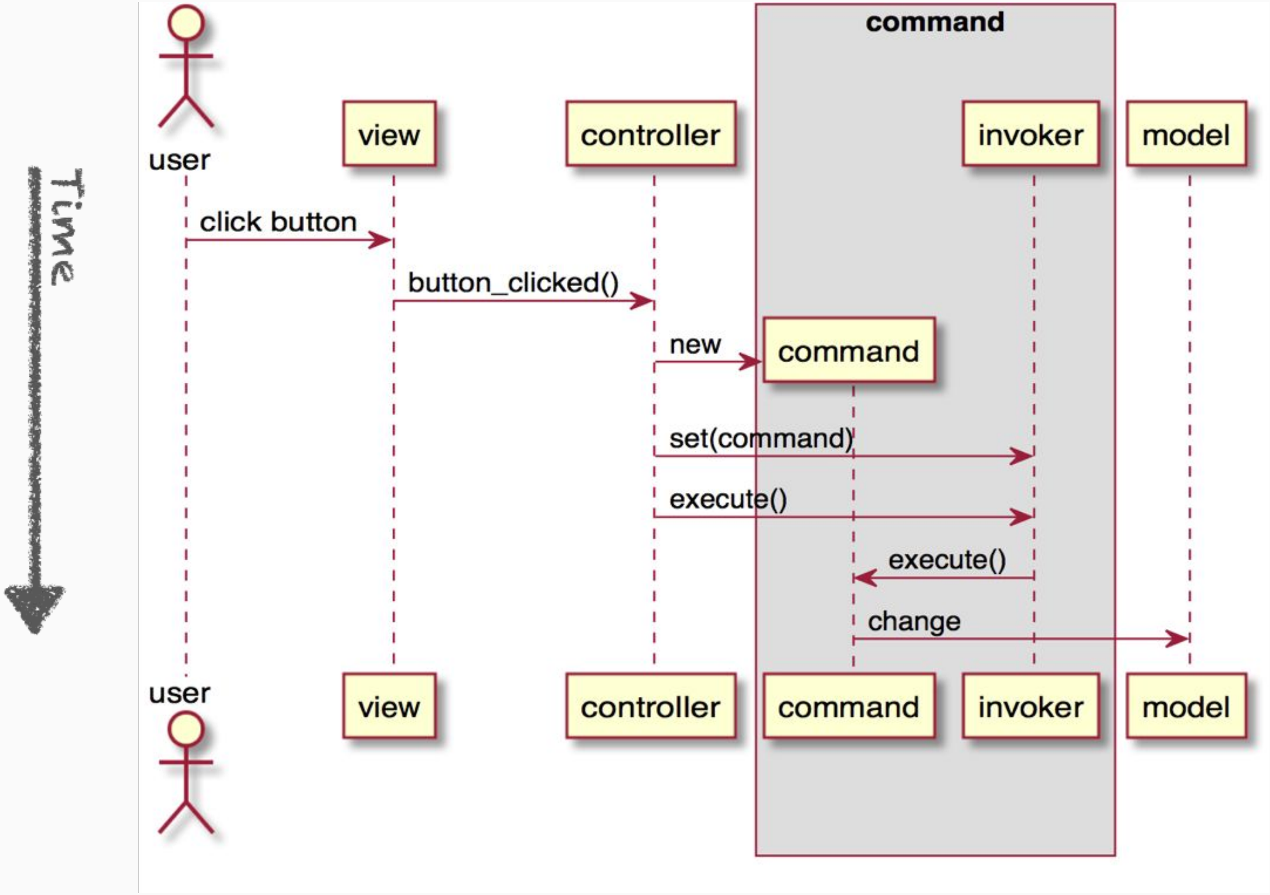
Problem Permission checks are spread throughout the applications; Miss a single permission check - and you may have a severe security vulnerability

Design Pattern Use the command pattern instead of direct manipulation; Add permission metadata to command objects, and have a central point that validates permissions prior to executing commands

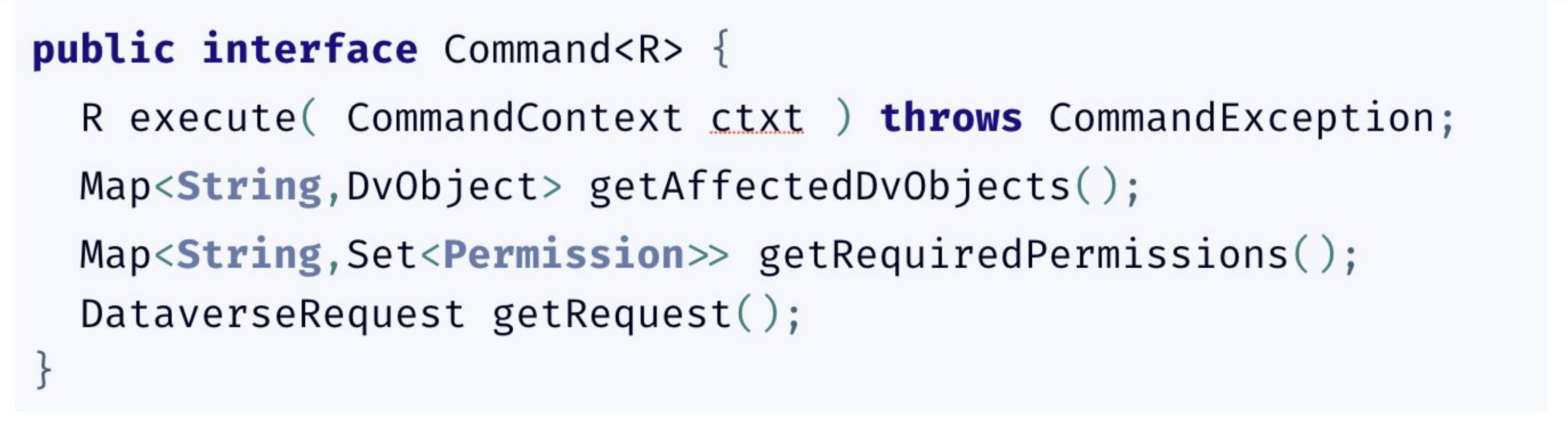
Command Pattern - Recap



Command Pattern - Recap



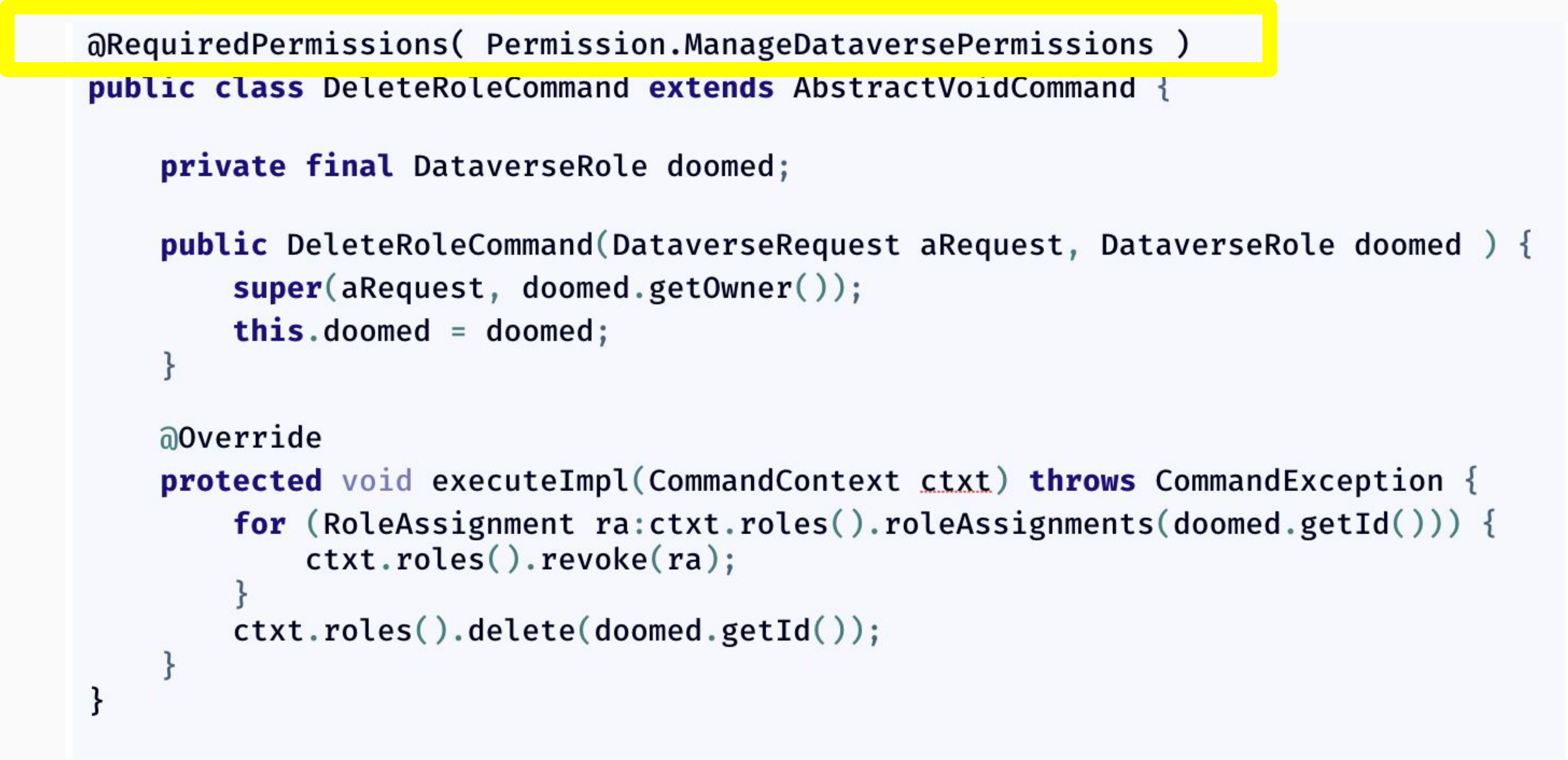
Adapted Command Pattern



```
public interface Command<R> {
  R execute( CommandContext ctxt ) throws CommandException;
  Map<String,DvObject> getAffectedDvObjects();
  Map<String,Set<Permission>> getRequiredPermissions();
  DataverseRequest getRequest();
}
```

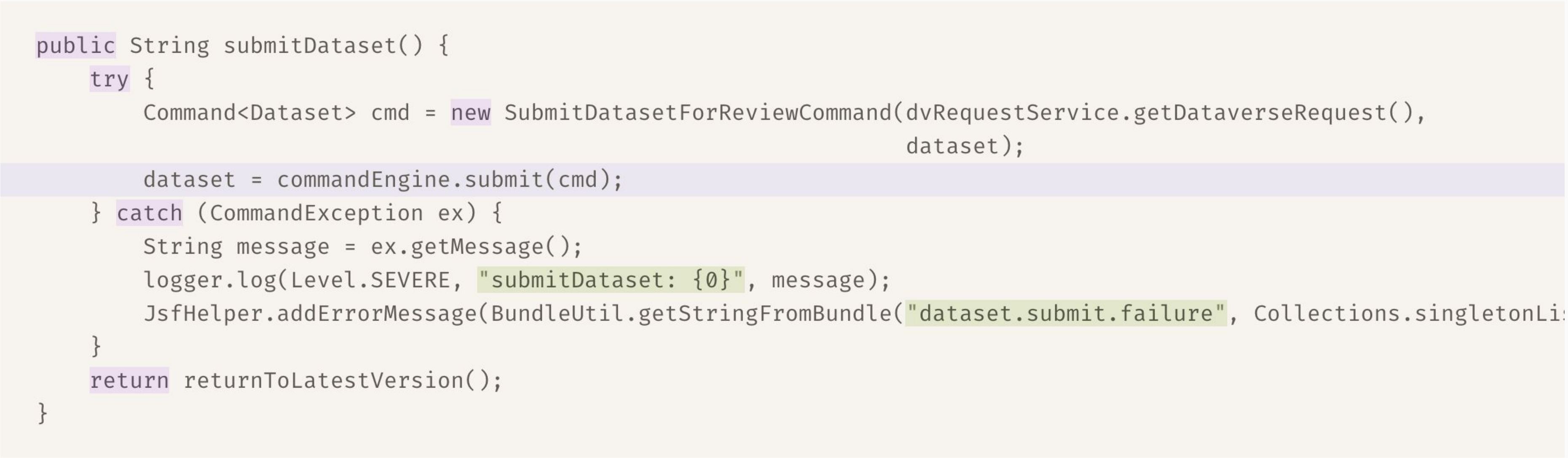
- `execute()` is where the work is done.
  - Server resources and injected objects are made available via `CommandContext`
- `getAffectedDvObjects()` and `getRequiredPermissions()` detail which objects are affected and what permissions are needed to affect them
- `getRequest()` allows the permission system to detect which permissions the user has.

A Sample Command



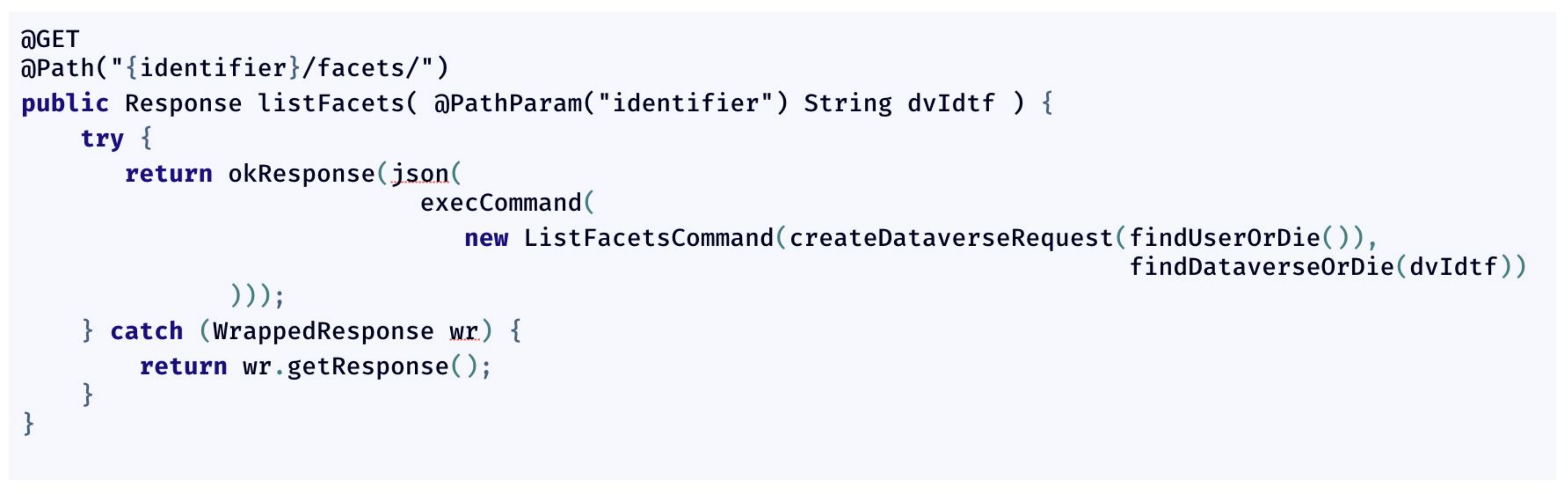
```java
@RequiredPermissions( Permission.ManageDataversePermissions )
public class DeleteRoleCommand extends AbstractVoidCommand {

    private final DataverseRole doomed;

    public DeleteRoleCommand(DataverseRequest aRequest, DataverseRole doomed ) {
        super(aRequest, doomed.getOwner());
        this.doomed = doomed;
    }

    @Override
    protected void executeImpl(CommandContext ctxt) throws CommandException {
        for (RoleAssignment ra:ctxt.roles().roleAssignments(doomed.getId())) {
            ctxt.roles().revoke(ra);
        }
        ctxt.roles().delete(doomed.getId());
    }
}
```

Command Usage



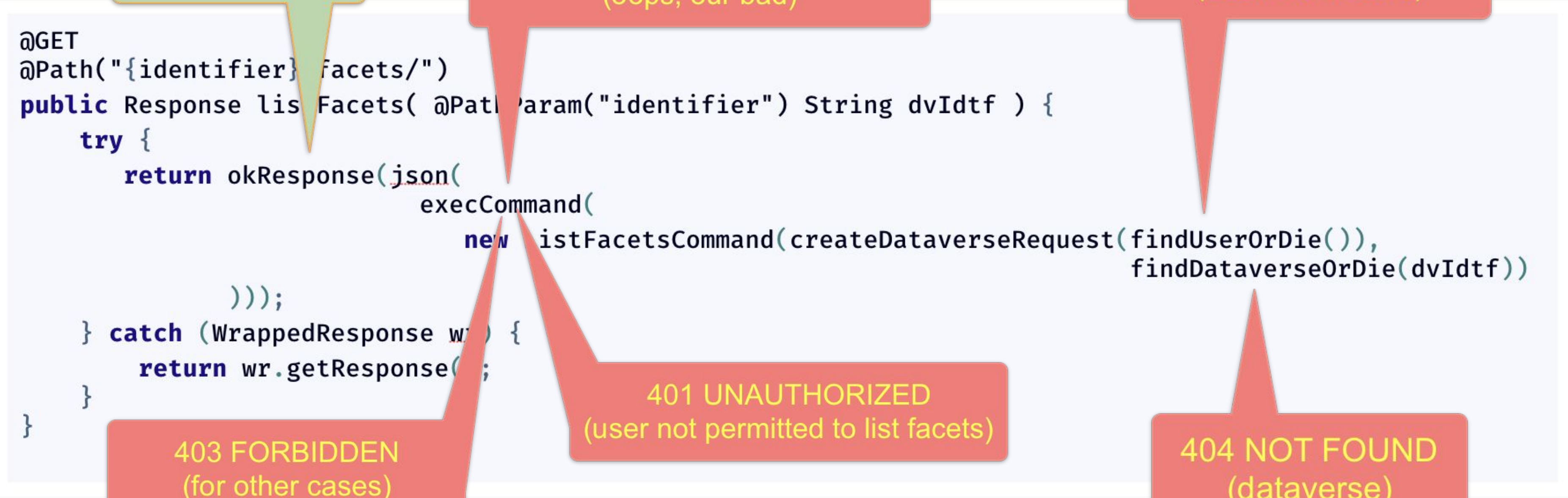
```java
public String submitDataset() {
    try {
        Command<Dataset> cmd = new SubmitDatasetForReviewCommand(dvRequestService.getDataverseRequest(),
                                                                 dataset);
        dataset = commandEngine.submit(cmd);
    } catch (CommandException ex) {
        String message = ex.getMessage();
        logger.log(Level.SEVERE, "submitDataset: {0}", message);
        JsfHelper.addErrorMessage(BundleUtil.getStringFromBundle("dataset.submit.failure", Collections.singletonLi
    }
    return returnToLatestVersion();
}
```

Command Pattern Extra Points: Abstraction of Model Actions



```
@GET
@Path("{identifier}/facets/")
public Response listFacets( @PathParam("identifier") String dvIdtf ) {
    try {
       return okResponse(json(
                           execCommand(
                              new ListFacetsCommand(createDataverseRequest(findUserOrDie()),
                                                                             findDataverseOrDie(dvIdtf))
              )));
    } catch (WrappedResponse wr) {
        return wr.getResponse();
    }
}
```

Command Pattern Extra Points: Abstraction of Model Actions



```
@GET
@Path("{identifier}/facets/")
public Response listFacets( @PathParam("identifier") String dvIdtf ) {
    try {
        return okResponse(json(
                                    execCommand(
                                        new ListFacetsCommand(createDataverseRequest(findUserOrDie()),
                                                                                          findDataverseOrDie(dvIdtf))
                     )));
    } catch (WrappedResponse wr) {
        return wr.getResponse();
    }
}
```

200 OK
(JSON content)

500 INTERNAL SERVER ERROR
(oops, our bad)

401 UNAUTHORIZED
(user not found)

403 FORBIDDEN
(for other cases)

401 UNAUTHORIZED
(user not permitted to list facets)

404 NOT FOUND
(dataverse)

More on this Design Pattern

- JavaOne 2014: BOF5619 - Lean Beans (are made of this): Command pattern vs. MVC

  [Durbin, Bar-Sinai]
- JavaOne 2016: BOF4161 - REST in Peace with Java EE

  [Durand, Bar-Sinai]
- **Securing Dataverse with an Adapted Command Design Pattern**

  [Durand, Bar-Sinai, Crosas, Proceedings of 2017 IEEE Cybersecurity Development]

Differential Privacy

$$Pr[T(M(X)) = 1] \leq e^{\epsilon} Pr[T(M(X')) = 1] + \delta, \qquad \forall\, T.$$

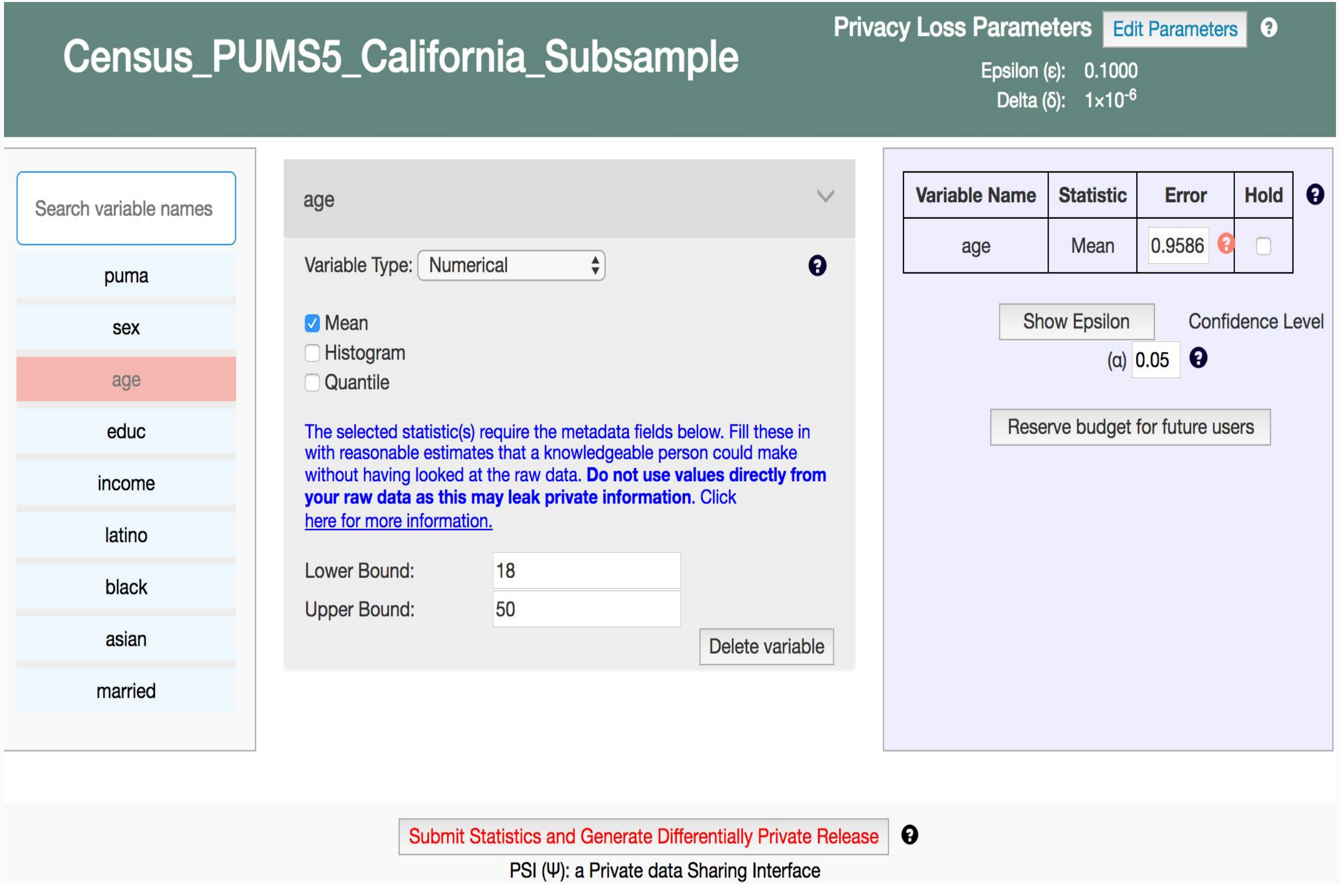
Differential Privacy is a formal, mathematical conception of privacy preservation. It **guarantees** that any reported result does not reveal information about any one single individual, regardless of auxiliary information.

https://privacytools.seas.harvard.edu/differential-privacy

https://privacytools.seas.harvard.edu/psi

External Tools: PSI Budgeteer

The budgeteer allows users to select which statistics they would like to calculate and are given estimates of how accurately each statistic can be computed. They can also redistribute their privacy budget according to which statistics they think are most valuable in their dataset.



Census_PUMS5_California_Subsample

Privacy Loss Parameters

Epsilon (ε): 0.1000

Delta (δ): 1×10^{-6}

Variable Type: Numerical

☑ Mean

□ Histogram

□ Quantile

The selected statistic(s) require the metadata fields below. Fill these in with reasonable estimates that a knowledgeable person could make without having looked at the raw data. **Do not use values directly from your raw data as this may leak private information.** Click here for more information.

Lower Bound: 18

Upper Bound: 50

Variable Name	Statistic	Error	Hold
age	Mean	0.9586	□

Confidence Level (α) 0.05

PSI (Ψ): a Private data Sharing Interface

UX & UI Implications

Steve Krug
UPDATED with three new chapters!
~~DON'T~~
MAKE
ME
THINK
A Common Sense Approach to Web Usability
SECOND EDITION

UX & UI Implications

- Requiring users to “stop and think” without scaring them
  - Complicates application flow
- How to raise these important issues so that users are prepared and not surprised by additional requirements?
- New functionality/UI pattern(s)
  - For Dataverse, we created an explicit step by step guide in the UI when depositing data
- **Might not play nicely with existing UI**
  - Color-based systems mess up your existing UI
  - Term collision: e.g. “Tagging” is already over used

Legal Implications

Does not replace professional legal advice!

- Have a professional legal team look at your system and the data it gathers, stores, shares, and processes
- Implications depend on country, state
- User agreements
- Terms of use
- Explicit acceptance of “warrant”

When Appropriate, Shift Liability to Users

Restrictions section of the Harvard Dataverse current EULA

Restrictions

In contributing Content to the site, you must ensure that the Content complies with the Terms of Use. ... ... you make the following representations and warranties to Harvard Dataverse:

- ...
- User Uploads must be void of all identifiable information, such that re-identification of any subjects ... should not be possible. Specifically, User Uploads cannot contain social security numbers; credit card numbers; medical record numbers; health plan numbers; other account numbers of individuals; or biometric identifiers.... The only exceptions for when identifiable information is allowed are when:
    - the information has been previously released to the public;
    - the information describes public figures, where the data relates to their public roles or other non-sensitive subjects ;
    - ...

Information in this slide does not replace professional legal advice!

Lessons Learned

- Understand risks, legal background, and technologies that apply to your application's data. These also affect what you can do with the data once gathered
- Expect a considerable effect on application UI/UX
- No more "roll your own" permission enforcement, this has to be done in a single place using a well-tested algorithm
- Design UX, UI, and application so that not all installations/accounts have to support the most sensitive data (and, thus, do not need to suffer the UX/performance impact)
- Use what you need, not what's available
  - Do you absolutely *have* to comply with FIPS 140-2 Level 3?

IQSS

The Institute for Quantitative Social Science

@dataverseorg

Thanks!

Questions?

The Dataverse Project

Michael's travel funded in part by The Frankel Fund for BGU students