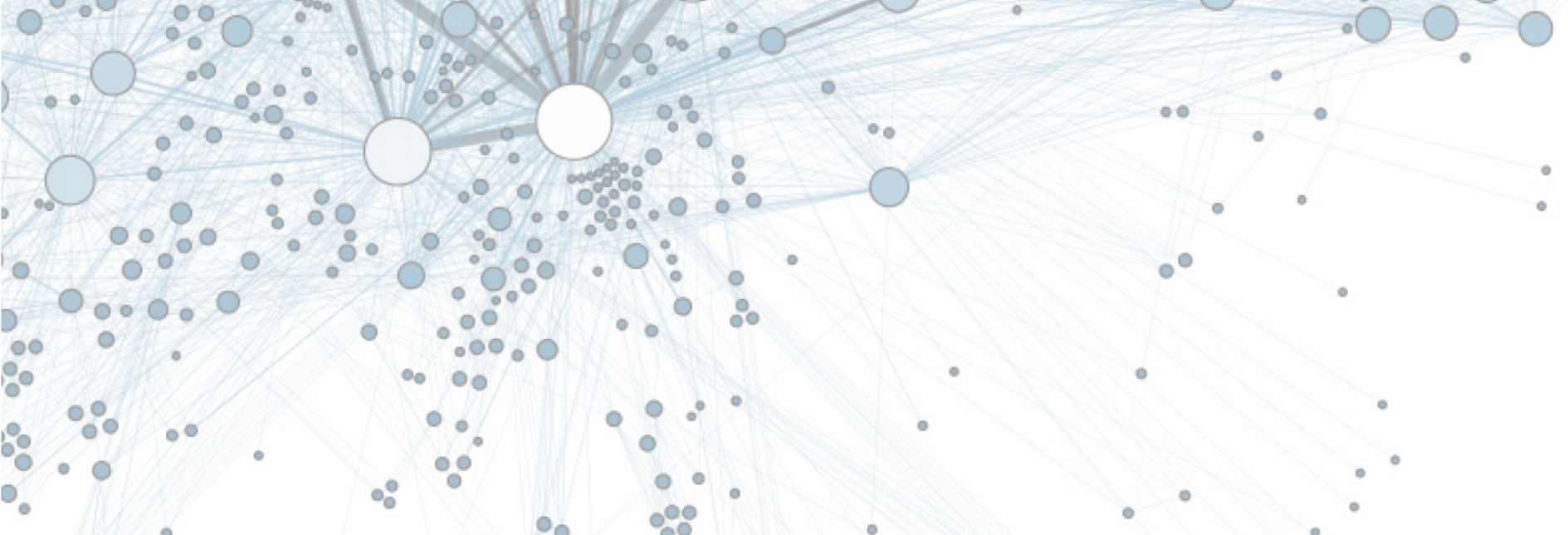


TOWARDS A LICENSING MODEL AND ECOSYSTEM FOR DATA SHARING

SAM MADDEN (MIT)

W/ CARSTEN BINNIG (BROWN), JANE GREENBERG (DREXEL), TIM KRASKA (BROWN)

NSF Big Data Hub Spoke Grant

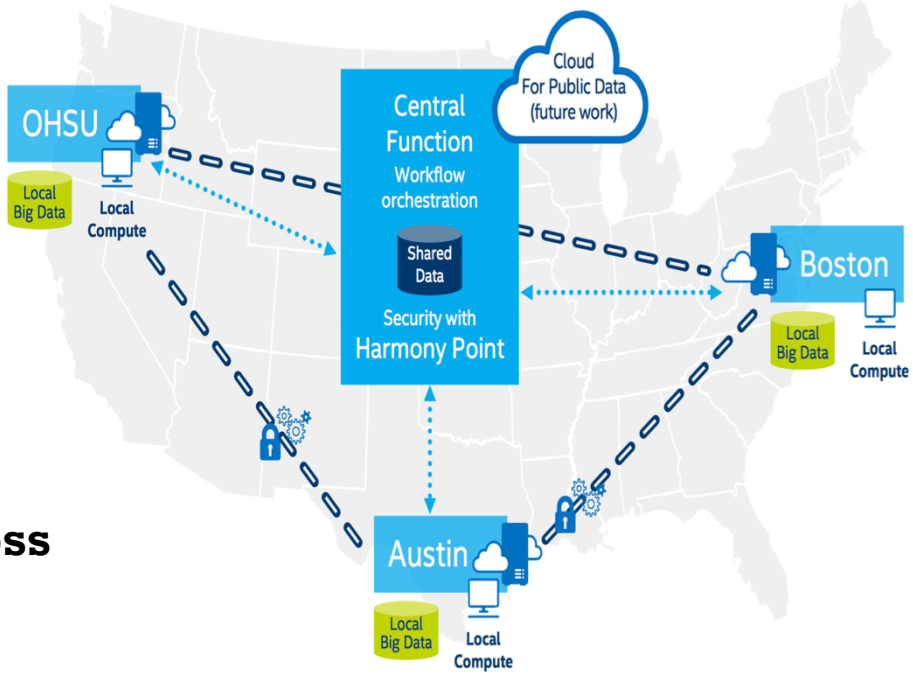
WHY IS DATA SHARING IMPORTANT!

Different Reasons

- Combining Data
- Sharing with Experts
- …

Promise: Better Insights into "Big Data"

COMBINING DATA: COLLABORATIVE CANCER CLOUD



Secure genomic data sharing across our three institutions

How open data on agriculture & nutrition can solve world hunger

07 SEPTEMBER 2016

Noi Jaitang, interviewed as part of the World Resources Institute report, waters his garden in Thailand // Laura Villadiego

How to Solve the Environmental Information Divide

TERESA MATHEW SEP 5, 2017

The New York Times

February 2016

SundayReview | OPINION

Give Up Your Data to Cure Disease

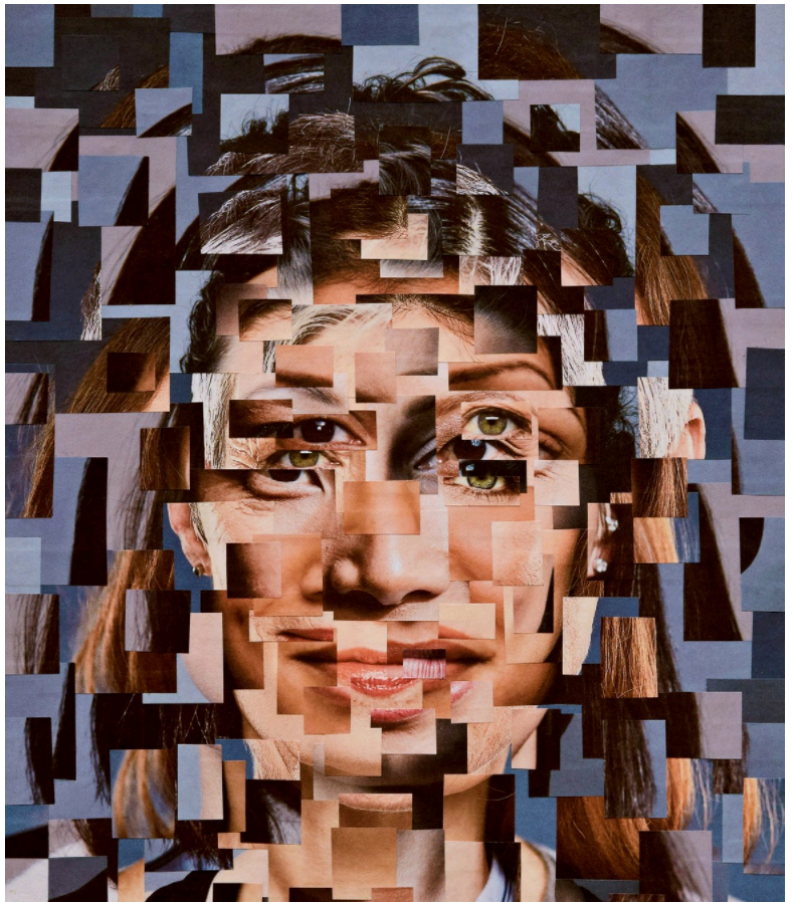
By DAVID B. AGUS FEB. 6, 2016

MARK WARREN **NATIONAL FRONTIERS** SCIENCE 10.19.16 6:55 AM

THE CURE FOR CANCER IS DATA—MOUNTAINS OF DATA

WIRED

October 2016



December 2013

Yes, Big Data Can Solve Real World Problems

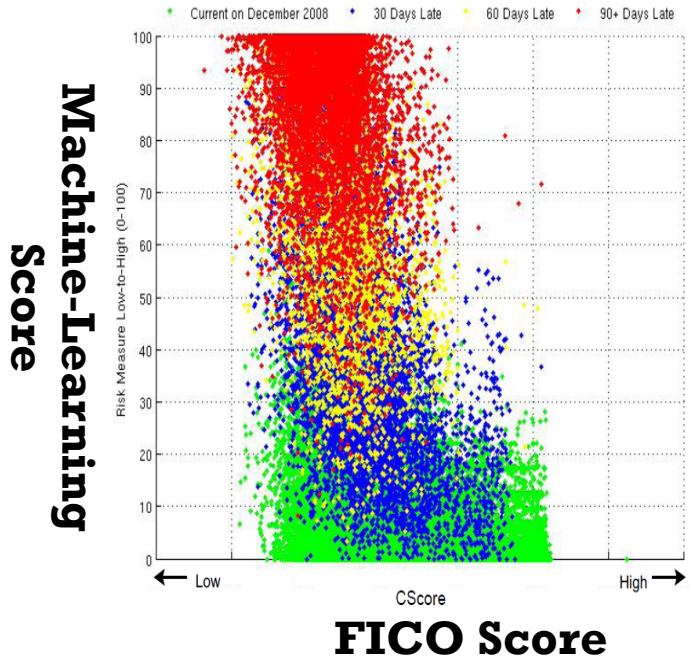
Greg Satell, CONTRIBUTOR
FULL BIO
Opinions expressed by Forbes Contributors are their own.

Forbes, *Working with IBM, the Memphis Police Dept. managed to reduce crime by 30% using big data analytics*

SHARING DATA WITH RESEARCHERS: FINANCIAL DATA

Challenge: Consumer credit risk analysis and forecasting

Approach: Machine learning



1% sample,10Tb
This graph:2008Q4

- Current
- 30-days late
- 60-days late
- 90+ days late

Andrew W. Lo et al.
(MIT Sloan School)

Machine-learning detects potential defaults more accurately than FICO scores!

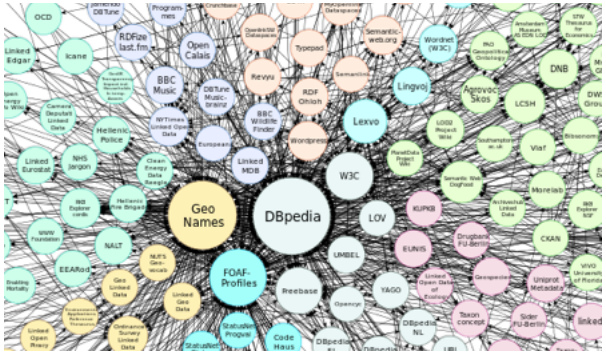
BUT DATA SHARING IS HARD!

OPEN DATA

DRYAD

DataONE

DFC DataNet FEDERATION CONSORTIUM



CLOSED DATA

Intel-Collaborative Cancer Cloud (CCC) (Dana-Farber, OHSU, Ontario Institute for Cancer Research (OICR))

Collaborative Genomics Cloud

(CGC)colocalizing massive genomics datasets)

FICO®

FICO score (Fair Isaac Corporation)

WHY NOT OPEN DATA?

creative commons

BARRIERS TO DATA SHARING

Must go beyond "creative commons"

Incentives – why would someone go to all the effort to share their valuable data?

Concerns over sensitive information (e.g., PII)

Regulations governing use of data in different domains

Not just "throwing it over the wall"!

- Do not want to loose control over data
- Can I get my data back?
- Has to be updated, requires training, redacted etc.

SHARING DATA TODAY

No data sharing without a legal agreement

Involve lawyers to create individual agreements → often prevents sharing!

DATA SHARING SPOKE: COMPONENTS

1. **Data-sharing Licensing Framework / Generator**
2. **Data-Sharing Platform (Enforce Licenses)**

Principle: Solve the 80% case!

GOAL: LICENSING FRAMEWORK

Standard terms that researchers, lawyers, and compliance teams conform with

- [x] Controlled access
- [ ] Tracking of access
- [x] Usage rights (e.g., publication, copying)
- [ ] Duration of use
- [x] Warrantees of correctness/completeness/availability
- [ ] Other requirements and regulations

LICENSES: **FIRST RESULTS**

Data-Sharing Workshop 2016 (Metadata Research Center @ Drexel):

- Approx. 60 participants form industry + academia
- Hear from the trenches
- What works? What doesn't? What are the biggest barriers? (What are the non-barriers?)
- Brainstorm solutions: would standardized licenses, use-cases/best practices help? Would better technologies help?
- Forge a path forward, together

Agenda and Report: http://cci.drexel.edu/mrc/news/2016-11-bigdatahubworkshop/

LICENSES: **FIRST RESULTS**

Collected sharing agreements from academic institutions

Compile list of standard terms for

- General (Time period, Use of data, …)
- Privacy & Protection (PII, Security, Training)
- Access (Who?, How?)
- Responsibility (Indemnity clause, Ownership, Rights)
- Compliance (Background checks, Right to audit, …)
- Data Handling (Allowed Methods of Data Transfer, …)

Initial analysis suggests there is much commonality

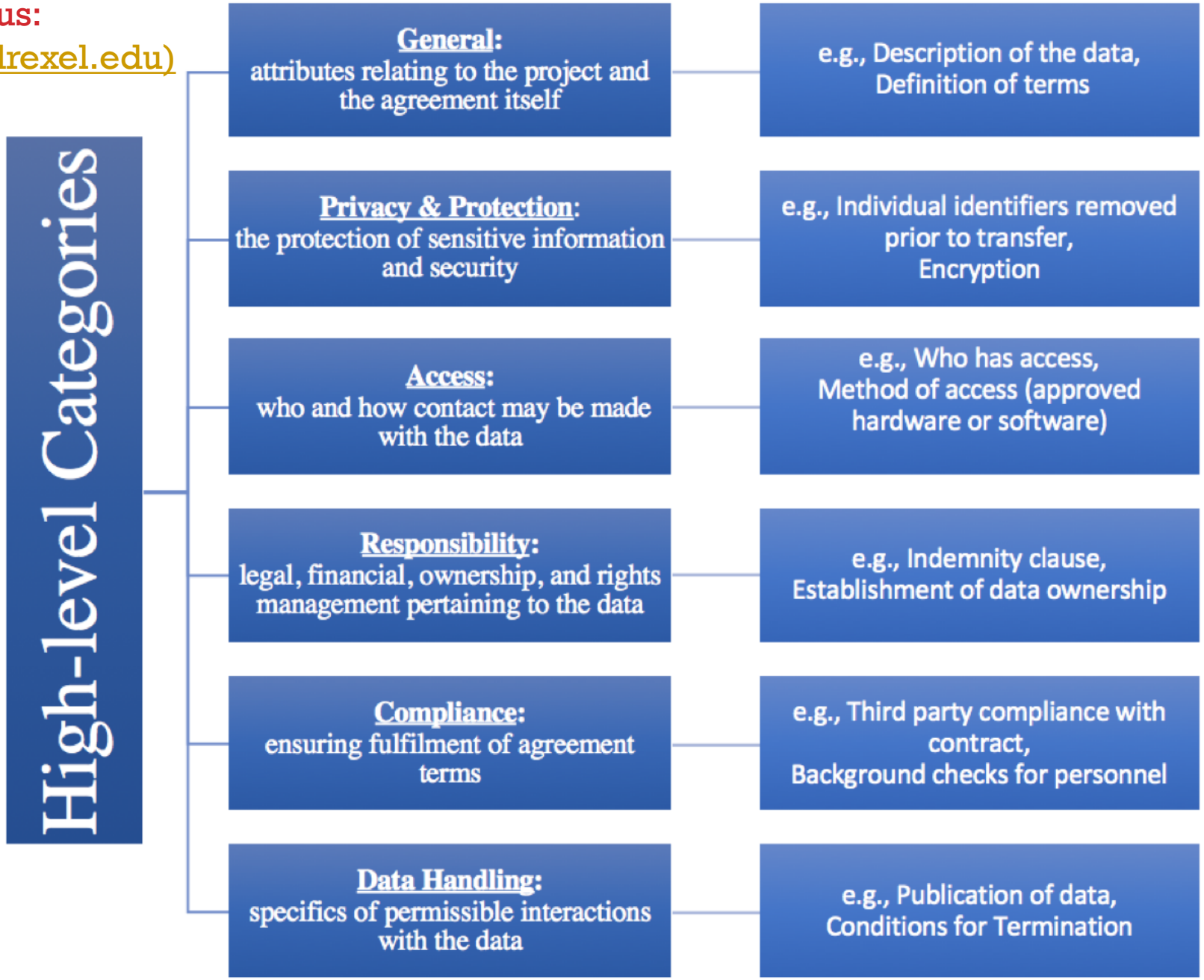
Send us your (anonymized) licenses: smg383@drexel.edu

CONTENT ANALYSIS

1. **Data collection**
   - 26 data sharing agreements, industry, academia, government
2. **Content analysis**
   - Confirm data sharing in closed environment
   - Focused, language parsed for higher-level general categories; mid, lower-level *to* → specifications to data handling
3. **Concept clustering**
   - Classes, sub-classes, attributes organized on a spreadsheet in a classified, hierarchical arrangement.
4. **Metadata labeling**
   - Language of the categories and attributes was refined

Licenses: First Results

(Sam Grabus: smg383@drexel.edu)



Privacy & Protection

Sensitive Information

Regulations

- Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)
- Compliance with federal/state/international data protection laws and regulations

Preparing data

- Identification of confidential/special categories of information (e.g., pii, proprietary)
- Individual identifiers removed/anonymized prior to transfer

Access

- Who has access to pii/confidential data
- Who has access to proprietary information

Privacy

- Anonymization of data
- Confidentiality and safeguarding of PII/sensitive data
- Removal/nondisclosure of company/personnel identification in materials and publications
- No contact with data subjects

Avoiding re-identification

- No direct/indirect re-identification
- Statistical cell size (how many people, in aggregated form, can be released in groups)
- Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify

Exceptions

- Exceptions to confidentiality
- Conditions of proprietary information disclosure
- Conditions of pii disclosure (who, what, and for what purpose?)
- Limitations on obligations if data becomes public
- Limitations on obligations if data is already known prior to agreement
- Limitations on obligations if data given by 3rd party without restriction

Security

- Sharing non-confidential data
- Password protection/authentication of files
- Encryption
- Security training for involved personnel
- Establishing infrastructure to safeguard confidential data

Data Handling

Use

- Each data field/elements to be accessed
- Use of data: only for project-specific/research, or analytical use
- Documenting all projects using the data
- Modification of data
- Compliance with data updates (changes, removal, corrections)
- Sharing data

Physical

- Copy/reproduction of data
- Storage of data
- Transfer of data (e.g., allowed methods)

Results

- Presentation of data
- Publication of data (e.g., prior approval needed or right to publically disclose publication)
- Results/reports and associated documents (e.g., must be provided copies)
- Right to remove/delete confidential data from proposed publications

Personal Gain

- Sale of/profit from data (e.g., noncommercial use only)
- Licensing of data
- No reverse engineering

Termination

- Conditions for termination
- Destruction or return of data after agreement
- 3rd party destruction or return of dataset
- Confirmation of data destruction
- Data retained or used for period of time after termination
- Which rights and obligations remain in effect after termination

6, ~ 40, 90+

Ontologizing

Privacy & Protection

- **Security**
  - Sharing non-confidential data →Sharing non-confidential data
  - Password protection/authentication of files → Password protection
  - Encryption → Encryption
  - Security training for involved personnel → Personnel Security Training
  - Establishing infrastructure to safeguard confidential data → Establishing Infrastructure

Data Handling

- **Use**
  - Each data field/elements to be accessed → Fields Accessed
  - Use of data: only for project-specific/research, or analytical use → Research Use Only
  - Documenting all projects using the data → Projects involved
  - Modification of data → Modification
  - Compliance with data updates (e.g., changes, removal, corrections) → Data Updates
- Sharing data → Data Sharing

NLTK – PARSING TERMS

Set maximum keywords length: 5
List top 1/5 of all the keywords

Result:

Keyword: research studies involving human subjects ,
score: 20.4583333333
Keyword: district assigned student identification numbers ,
score: 18.8387650086
Keyword: includes personally identifiable student information ,
score: 17.6168132942
Keyword: district initiated data research projects , score: 14.8577044025
Keyword: support effective instructional practices , score: 13.0
Keyword: personally identifiable information shared ,
score: 11.3440860215
Keyword: disclose personally identifiable information ,
score: 11.1440860215
Keyword: policy initiatives focused , score: 9.0
Keyword: informing education policies , score: 9.0

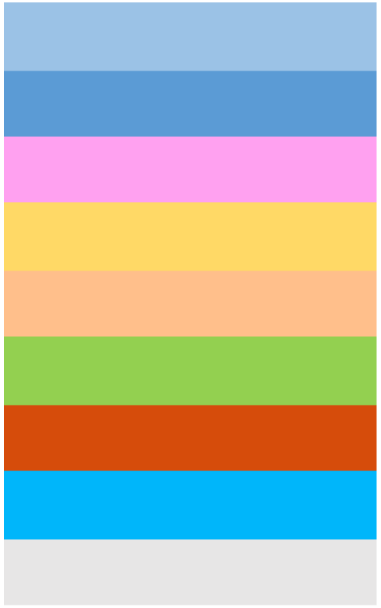
Sample 30 agreements

-5	-4	-3	-2	-1	0	1	2	3	4	5
			educational	right	privacy	act	health	insurance	portability	accountability
applicable	federal	law	regulation	protecting	privacy	citizen	including	family		
	license	agreement	authorized	protect	privacy	individual	subject	nd	study	
				applicable	privacy	law				
consistent	federal	family	educational	right	privacy	act	department	designates	education	alliance
subject	federal	family	educational	right	privacy	act	authorized			
education	record	covered	family	educational	privacy	act	amended			
recipient	agent	subcontractor	violation	agreement	privacy	rule	security	rule	implementing	regulation
comply	applicable	state	local	security	privacy	law	extent	protective	individual	privacy
		data	security	protection	privacy					
information	identified	family	educational	right	privacy	act				
		de	identified	applicable	privacy	law				
				applicable	privacy	law	permit	data	provider	provide
				federal	privacy	act	requirement	apply	agreement	entered
shared	state	subjected	applicable	requirement	privacy	confidentiality				
resolved	permit	covered	entity	comply	privacy	rule				
time	covered	entity	comply	requirement	privacy	rule	hipaa			
		reference	agreement	section	privacy	rule	mean	section	amended	renumbered
					privacy	rule	extent	information	created	received
					privacy	rule	standard	privacy	individually	identifiable
					privacy	rule	include	person	qualifies	personal
tern	defined	agreement	meaning	term	privacy	rule				
set	accordance	term	agreement	hipaa	privacy	security	rule			
hipaa	regulation	promulgated	thereunder	governing	privacy	security	health	information		

MOST FREQUENT TERMS

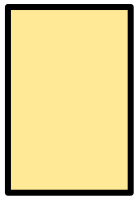
privacy	protection	set			
applicable	privacy	law			
privacy	rule	standard	privacy	individually	identifiable
definition	set	privacy	rule		
data	security	protection	privacy		

Frequency from the most to the least:



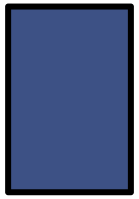
GOAL: HOSTED DATA-SHARING PLATFORM

data user

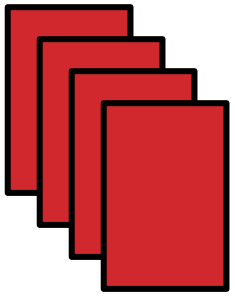


Suitably aggregated, de-identified, and fingerprinted data

ShareDB



Access log



data

data owner

IS THIS POSSIBLE: TECHNOLOGY ⋈ SHARING AGREEMENTS

Tech

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., publication, copying)

Duration of use

Warrantees of correctness/completeness/availability

Other requirements and regulations

IS THIS POSSIBLE: TECHNOLOGY ⋈ SHARING AGREEMENTS

Tech

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., publication, copying)

Duration of use

Warrantees of correctness/completeness/availability

Other requirements and regulations

IS THIS POSSIBLE: TECHNOLOGY ⋈ SHARING AGREEMENTS

Tech

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., publication, **copying**)

Duration of use

Warrantees of correctness/completeness/availability

Other requirements and regulations

IS THIS POSSIBLE: TECHNOLOGY ⋈ SHARING AGREEMENTS

Tech

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., **publication**, **copying**)

Duration of use

Warrantees of correctness/completeness/availability

Other requirements and regulations

IS THIS POSSIBLE: TECHNOLOGY ⋈ SHARING AGREEMENTS

Tech

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

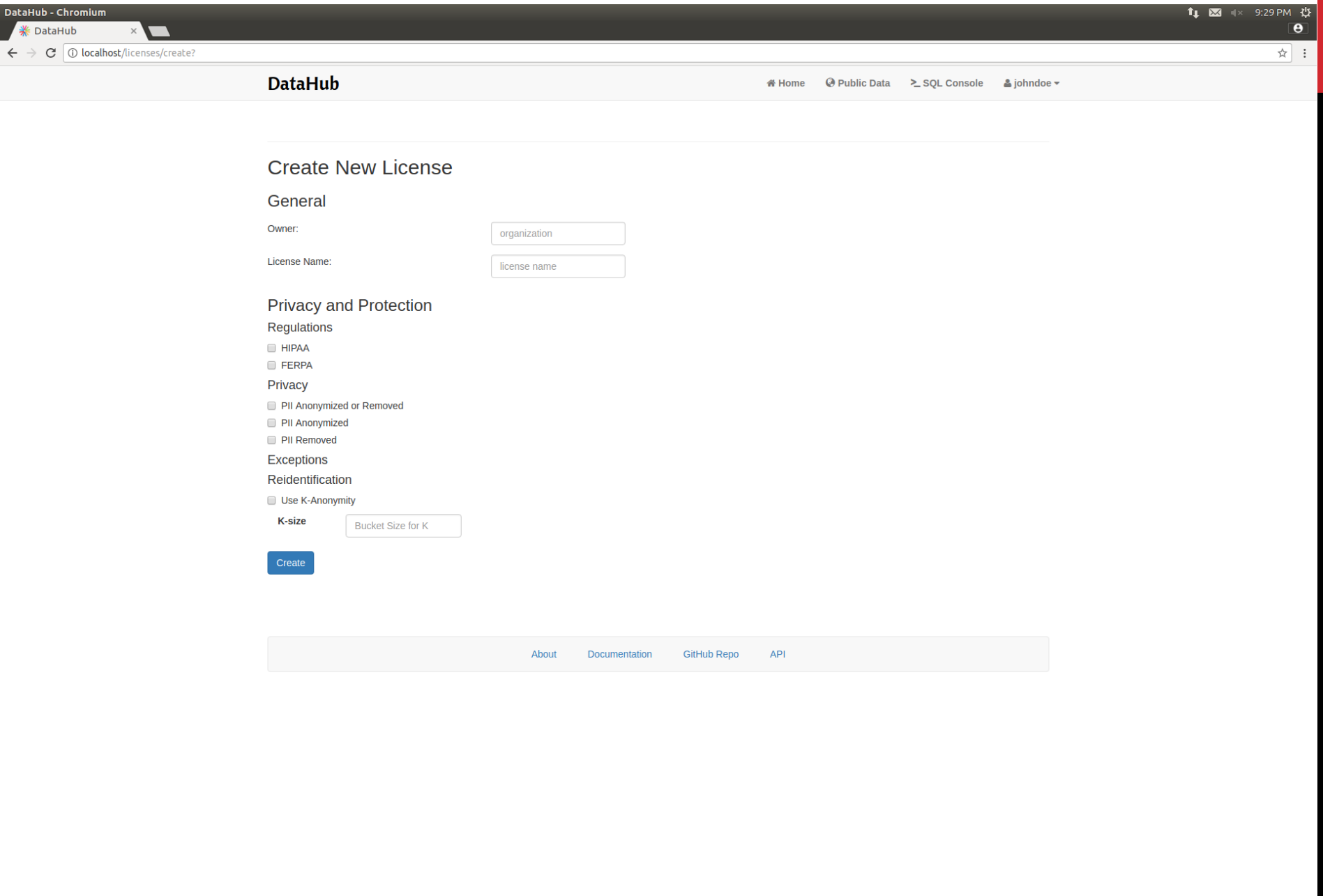
Usage rights (e.g., publication, copying)

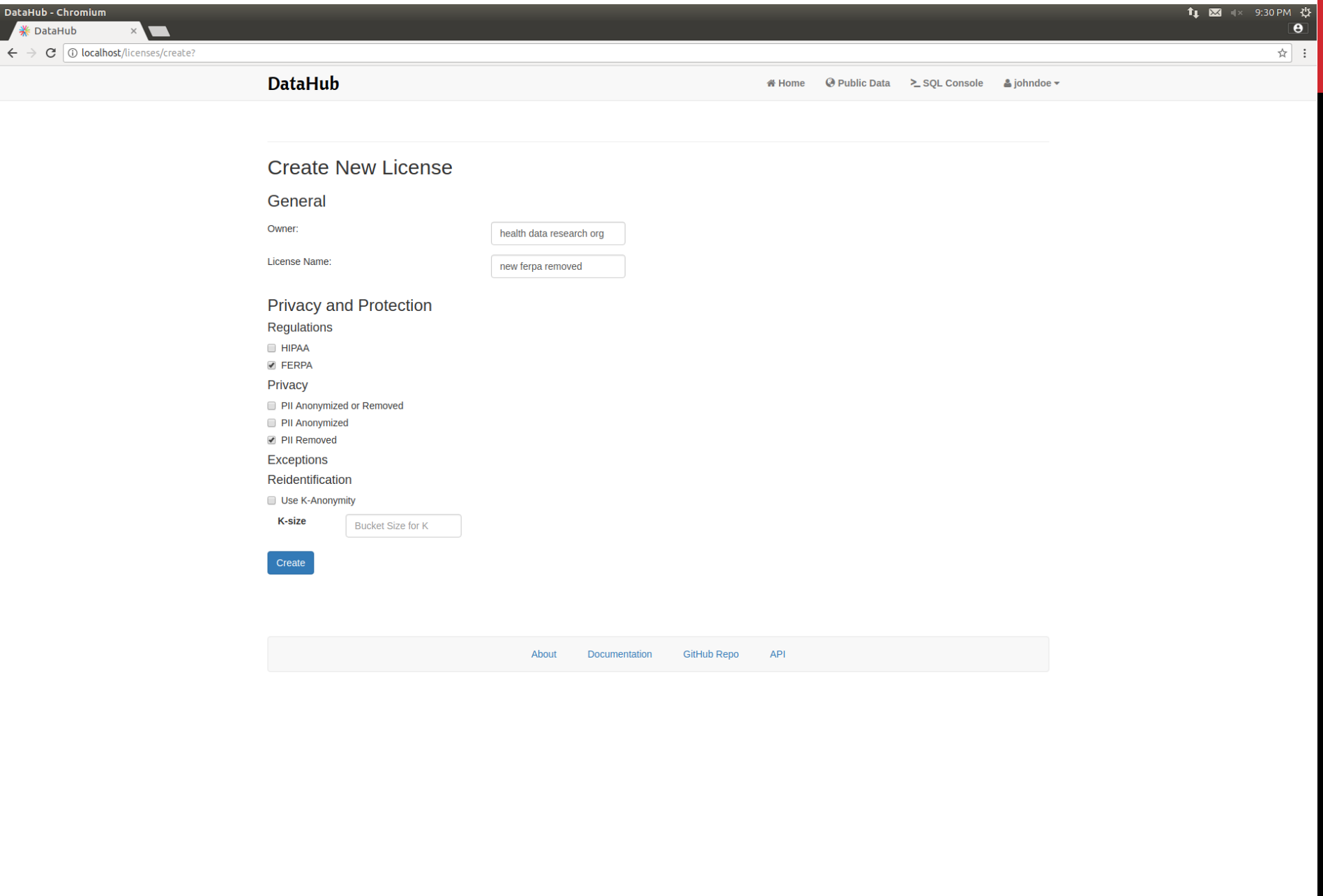
Duration of use

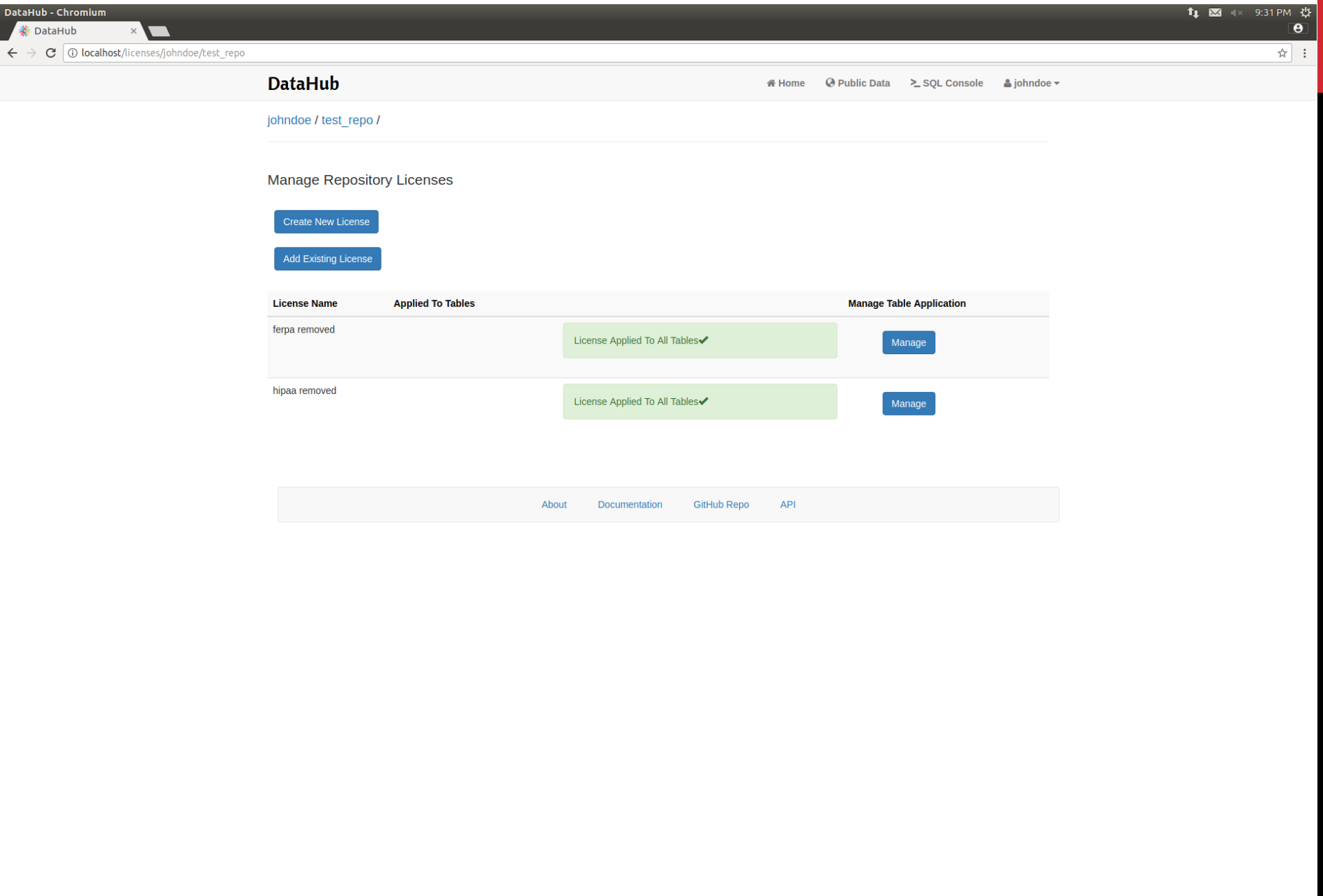
Warrantees of correctness/completeness/availability

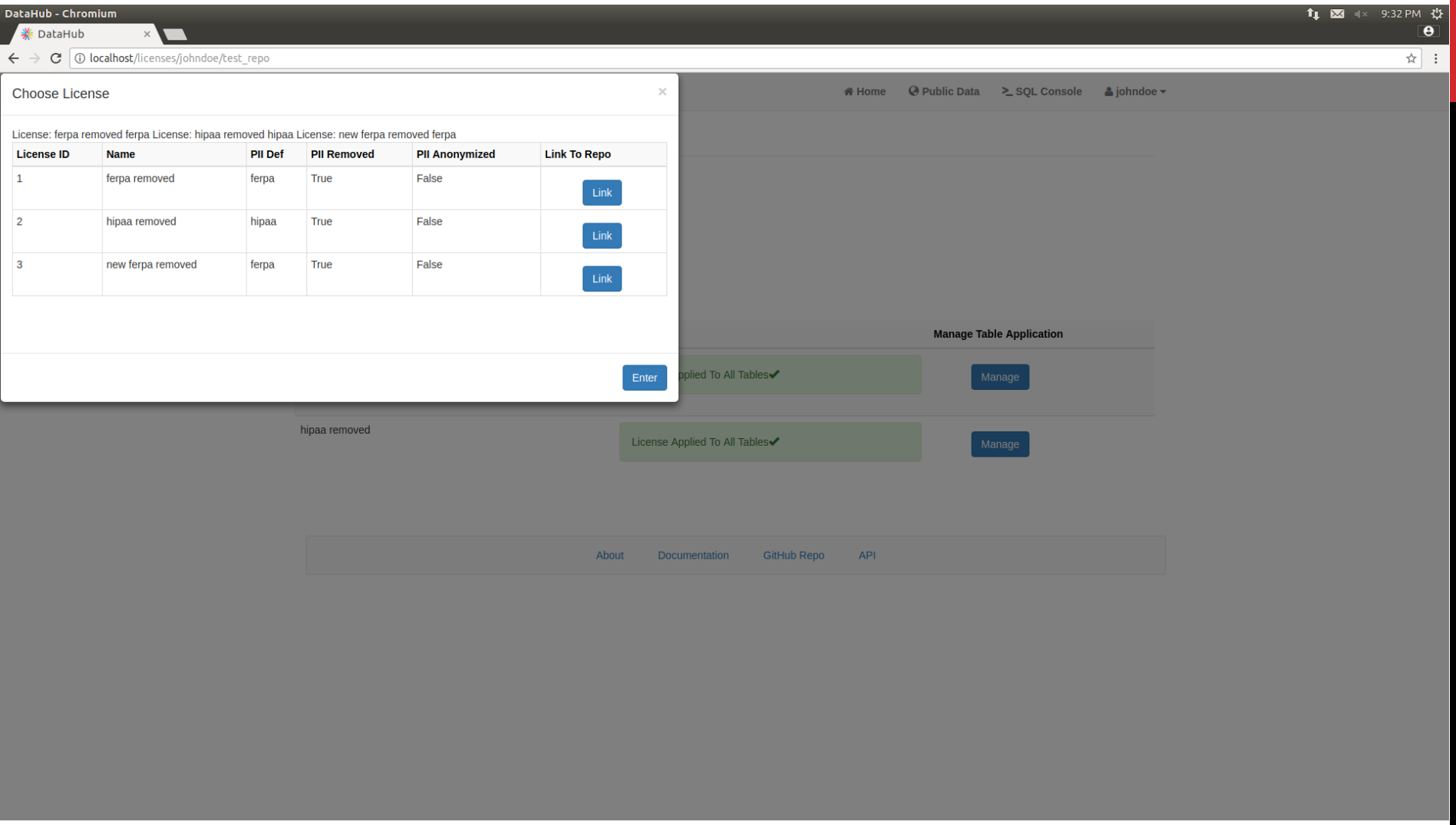
Other requirements and regulations

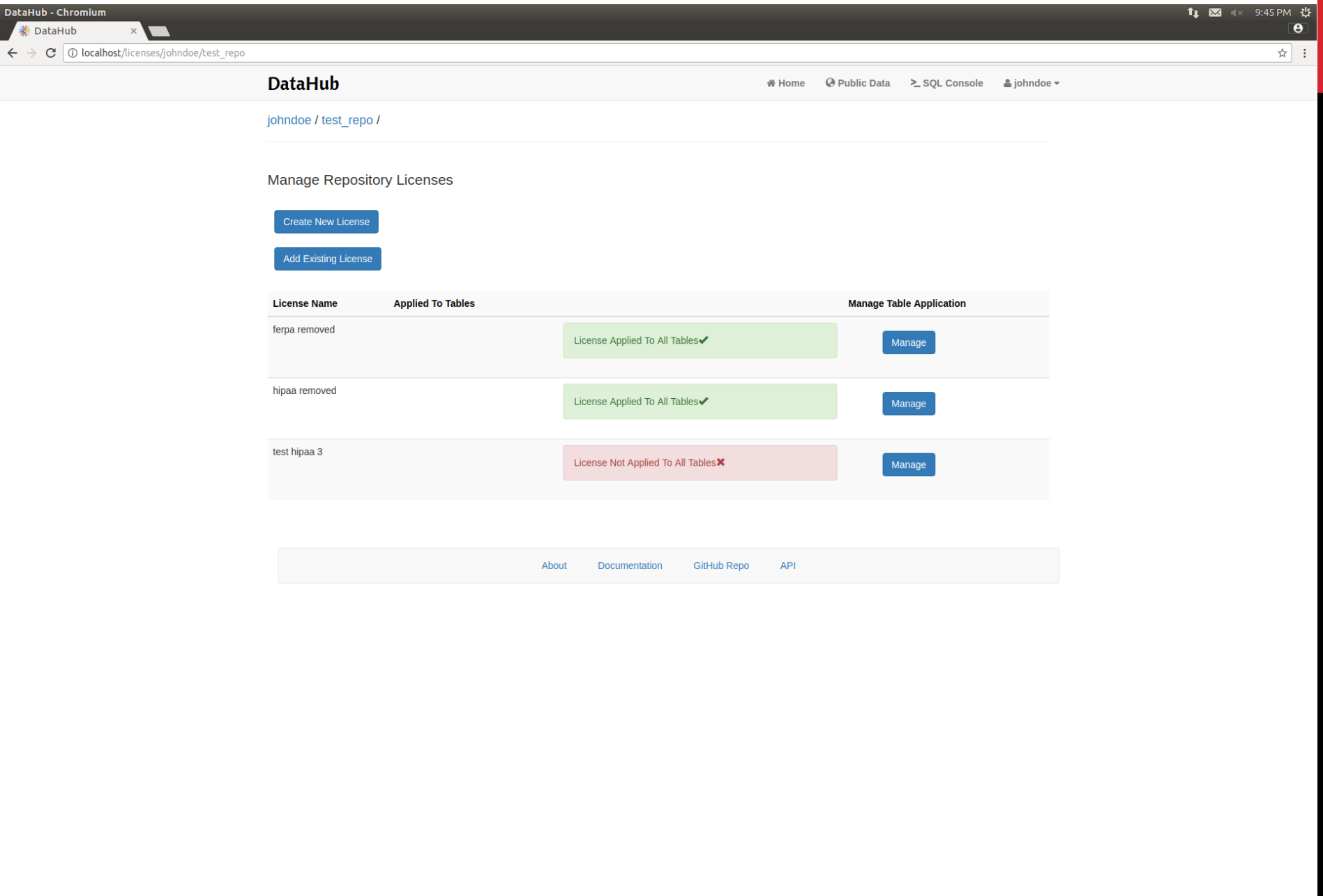
DEMO

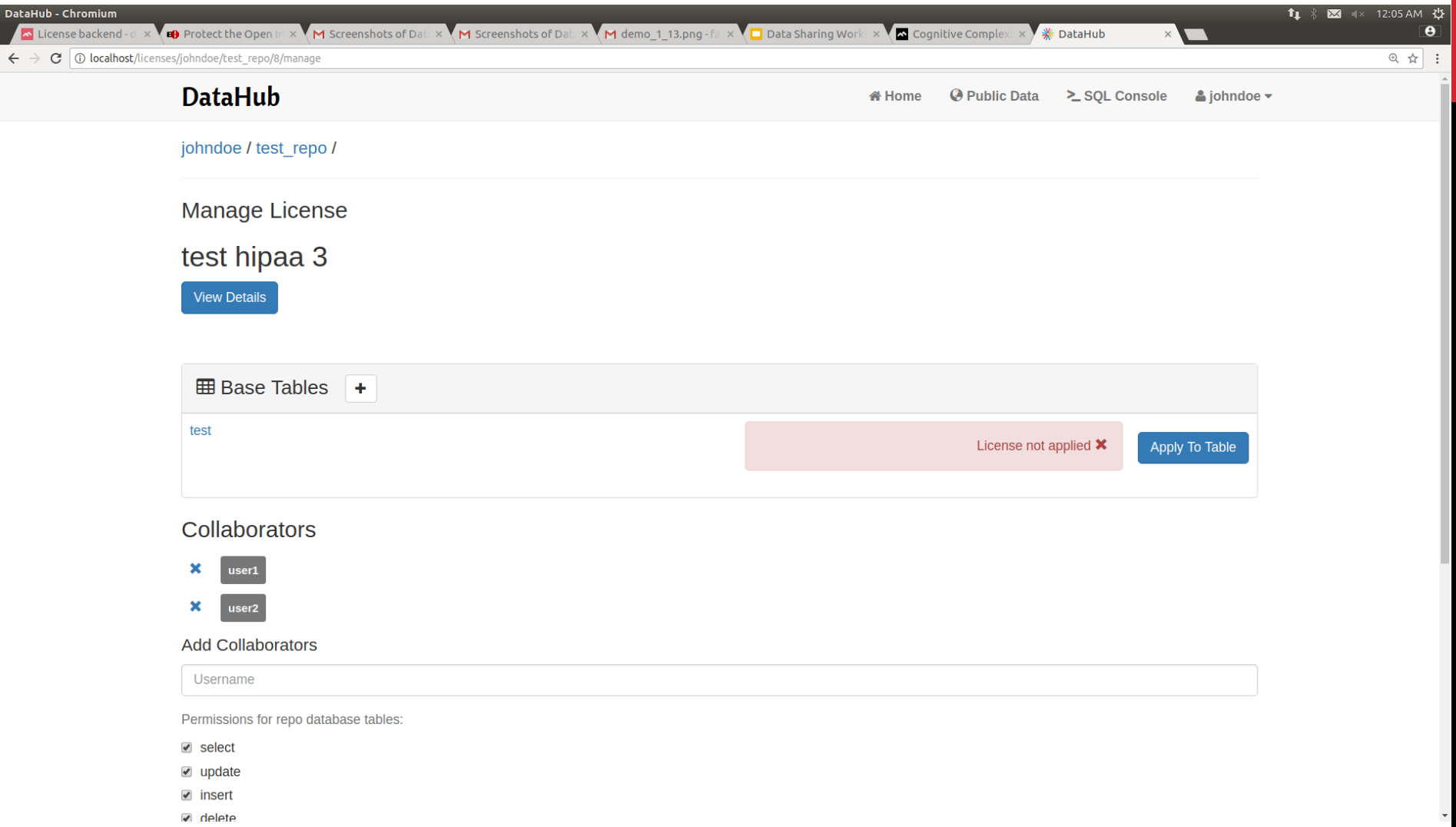

DataHub - Chromium
9:29 PM
DataHub
localhost/licenses/create?
DataHub
Home
Public Data
SQL Console
johndoe
Create New License
General
Owner:
organization
License Name:
license name
Privacy and Protection
Regulations
HIPAA
FERPA
Privacy
PII Anonymized or Removed
PII Anonymized
PII Removed
Exceptions
Reidentification
Use K-Anonymity
K-size
Bucket Size for K
Create
About
Documentation
GitHub Repo
API

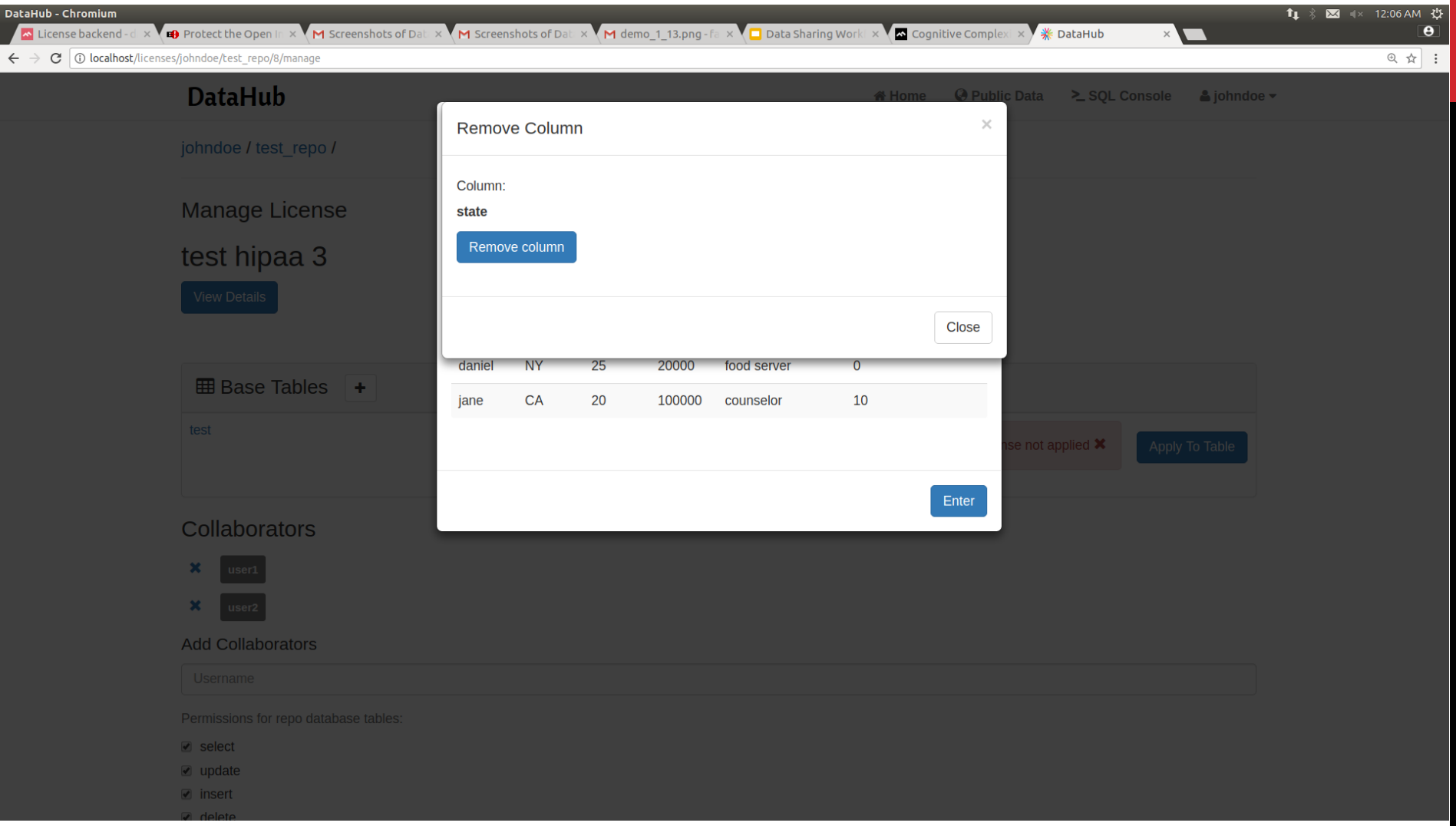

DataHub - Chromium
9:30 PM
DataHub
localhost/licenses/create?
DataHub
Home
Public Data
SQL Console
johndoe
Create New License
General
Owner:
health data research org
License Name:
new ferpa removed
Privacy and Protection
Regulations
HIPAA
FERPA
Privacy
PII Anonymized or Removed
PII Anonymized
PII Removed
Exceptions
Reidentification
Use K-Anonymity
K-size
Bucket Size for K
Create
About
Documentation
GitHub Repo
API

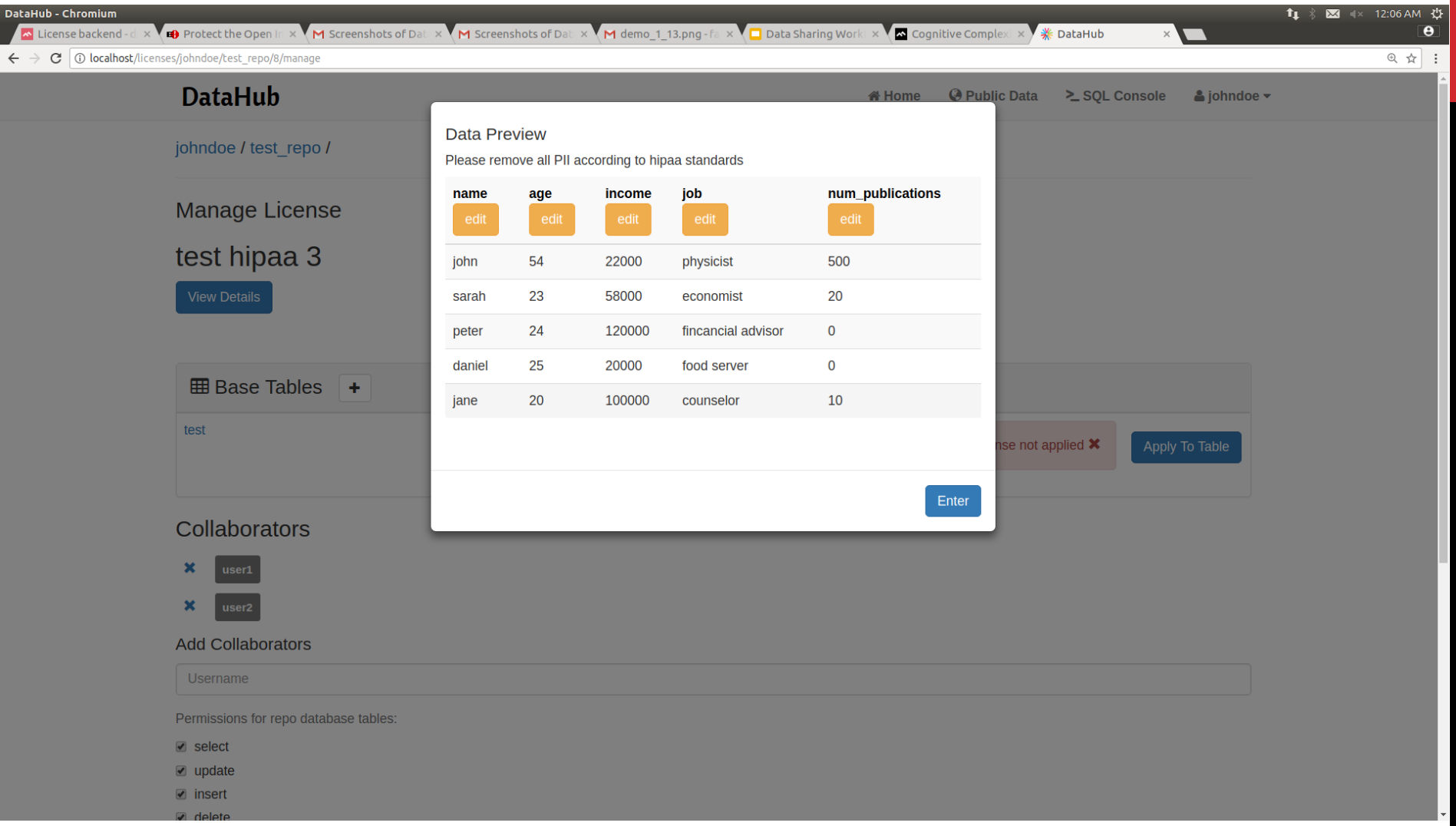

DataHub - Chromium
9:31 PM
DataHub
localhost/licenses/johndoe/test_repo
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage Repository Licenses
Create New License
Add Existing License
License Name
Applied To Tables
Manage Table Application
ferpa removed
License Applied To All Tables✔
Manage
hipaa removed
License Applied To All Tables✔
Manage
About
Documentation
GitHub Repo
API

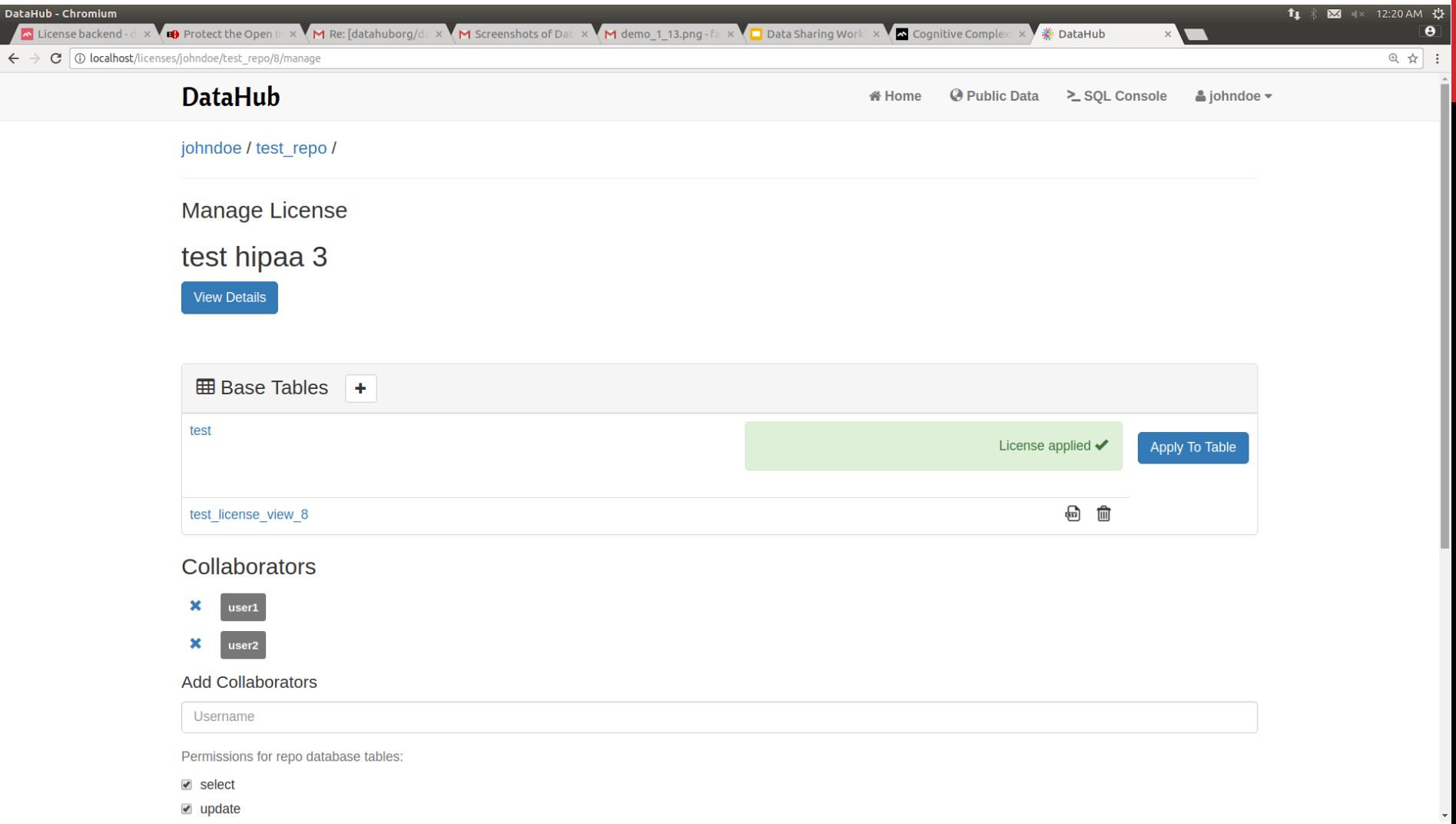

DataHub - Chromium
9:32 PM
DataHub
localhost/licenses/johndoe/test_repo
Choose License
License: ferpa removed ferpa License: hipaa removed hipaa License: new ferpa removed ferpa
License ID
Name
PII Def
PII Removed
PII Anonymized
Link To Repo
1
ferpa removed
ferpa
True
False
Link
2
hipaa removed
hipaa
True
False
Link
3
new ferpa removed
ferpa
True
False
Link
Enter
Home
Public Data
SQL Console
johndoe
Manage Table Application
Manage
hipaa removed
License Applied To All Tables
Manage
About
Documentation
GitHub Repo
API

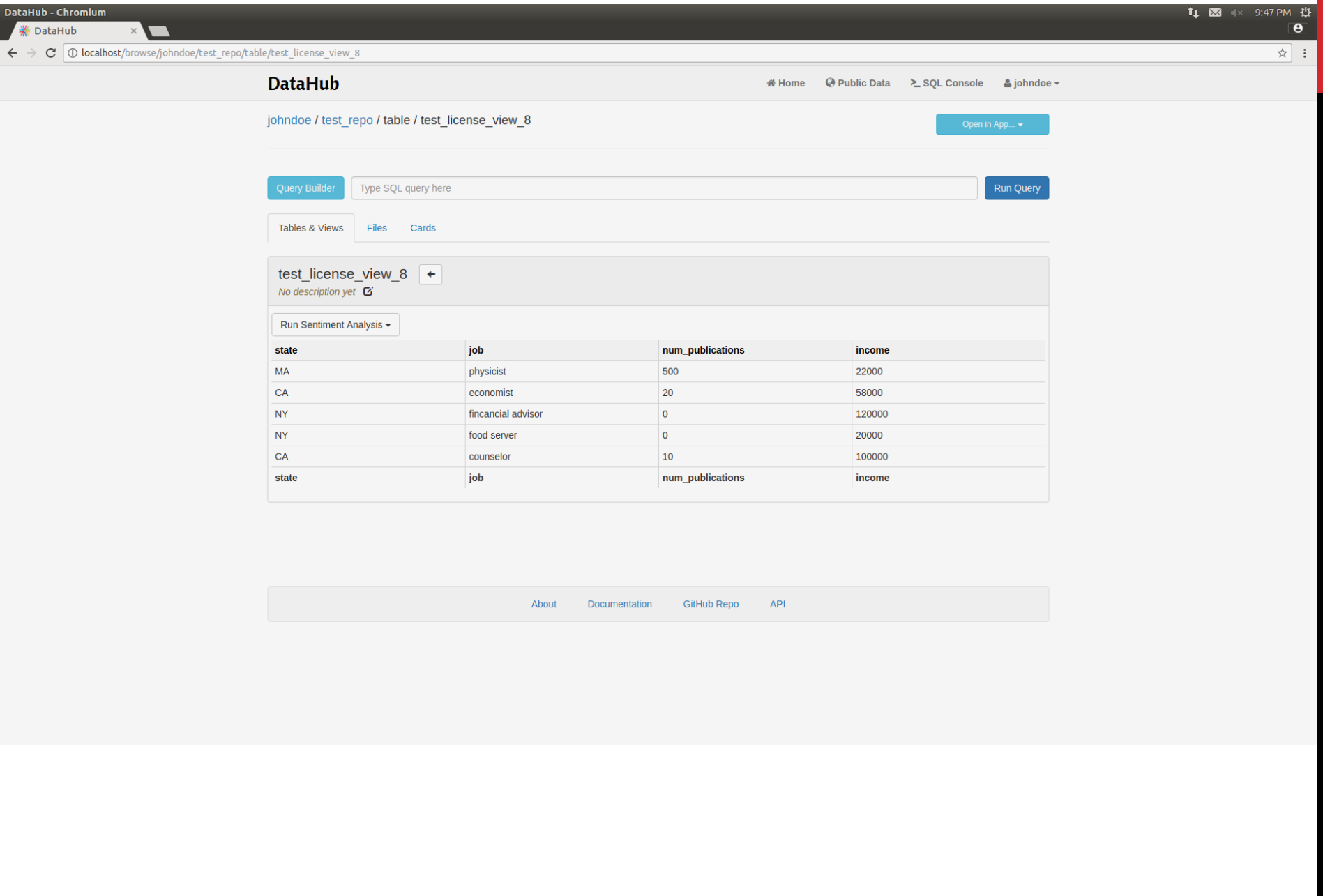

DataHub - Chromium
9:45 PM
DataHub
localhost/licenses/johndoe/test_repo
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage Repository Licenses
Create New License
Add Existing License
License Name
Applied To Tables
Manage Table Application
ferpa removed
License Applied To All Tables✔
Manage
hipaa removed
License Applied To All Tables✔
Manage
test hipaa 3
License Not Applied To All Tables✖
Manage
About
Documentation
GitHub Repo
API


DataHub - Chromium
12:05 AM
License backend - d
Protect the Open In
Screenshots of Dat
Screenshots of Dat
demo_1_13.png - fa
Data Sharing Work
Cognitive Complex
DataHub
localhost/licenses/johndoe/test_repo/8/manage
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage License
test hipaa 3
View Details
Base Tables
test
License not applied
Apply To Table
Collaborators
user1
user2
Add Collaborators
Username
Permissions for repo database tables:
select
update
insert
delete


DataHub - Chromium
12:06 AM
License backend - d
Protect the Open In
Screenshots of Dat
Screenshots of Dat
demo_1_13.png - fa
Data Sharing Work
Cognitive Complex
DataHub
localhost/licenses/johndoe/test_repo/8/manage
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage License
test hipaa 3
View Details
Remove Column
Column:
state
Remove column
Close
daniel NY 25 20000 food server 0
jane CA 20 100000 counselor 10
Base Tables
test
se not applied
Apply To Table
Enter
Collaborators
user1
user2
Add Collaborators
Username
Permissions for repo database tables:
select
update
insert
delete


DataHub - Chromium
12:06 AM
License backend - d
Protect the Open In
Screenshots of Dat
Screenshots of Dat
demo_1_13.png - fa
Data Sharing Work
Cognitive Complex
DataHub
localhost/licenses/johndoe/test_repo/8/manage
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage License
test hipaa 3
View Details
Base Tables
+
test
nse not applied
Apply To Table
Collaborators
user1
user2
Add Collaborators
Username
Permissions for repo database tables:
select
update
insert
delete
Data Preview
Please remove all PII according to hipaa standards
name
edit
age
edit
income
edit
job
edit
num_publications
edit
john 54 22000 physicist 500
sarah 23 58000 economist 20
peter 24 120000 fincancial advisor 0
daniel 25 20000 food server 0
jane 20 100000 counselor 10
Enter


DataHub - Chromium
12:20 AM
License backend - d
Protect the Open In
Re: [datahuborg/da
Screenshots of Dat
demo_1_13.png - fa
Data Sharing Work
Cognitive Complex
DataHub
localhost/licenses/johndoe/test_repo/8/manage
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo /
Manage License
test hipaa 3
View Details
Base Tables
test
License applied
Apply To Table
test_license_view_8
Collaborators
user1
user2
Add Collaborators
Username
Permissions for repo database tables:
select
update

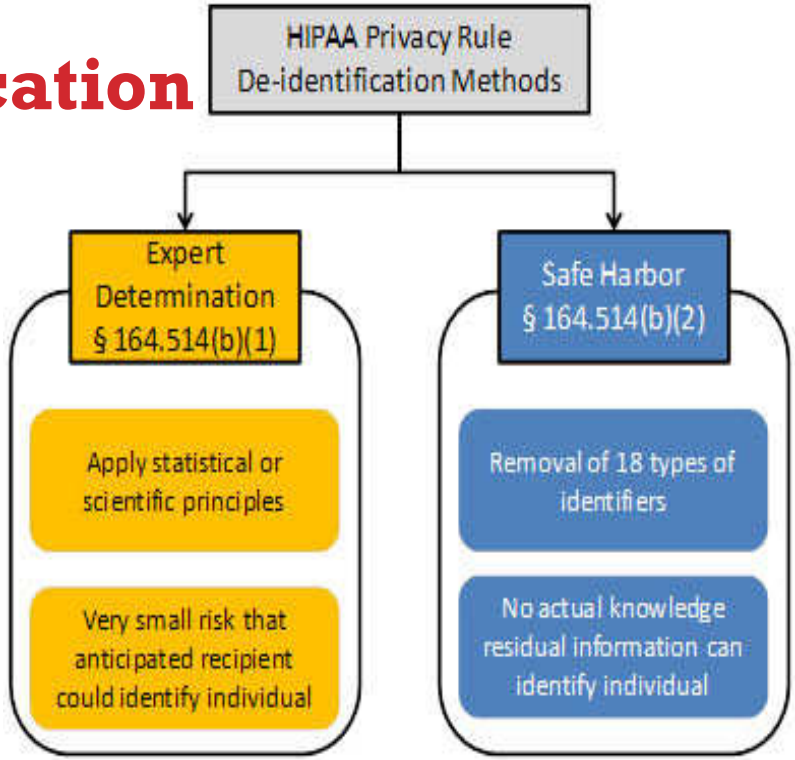

DataHub - Chromium
9:47 PM
DataHub
localhost/browse/johndoe/test_repo/table/test_license_view_8
DataHub
Home
Public Data
SQL Console
johndoe
johndoe / test_repo / table / test_license_view_8
Open in App...
Query Builder
Type SQL query here
Run Query
Tables & Views
Files
Cards
test_license_view_8
No description yet
Run Sentiment Analysis
state
job
num_publications
income
MA
physicist
500
22000
CA
economist
20
58000
NY
fincancial advisor
0
120000
NY
food server
0
20000
CA
counselor
10
100000
state
job
num_publications
income
About
Documentation
GitHub Repo
API

PLATFORM: **DEIDENTIFICATION**

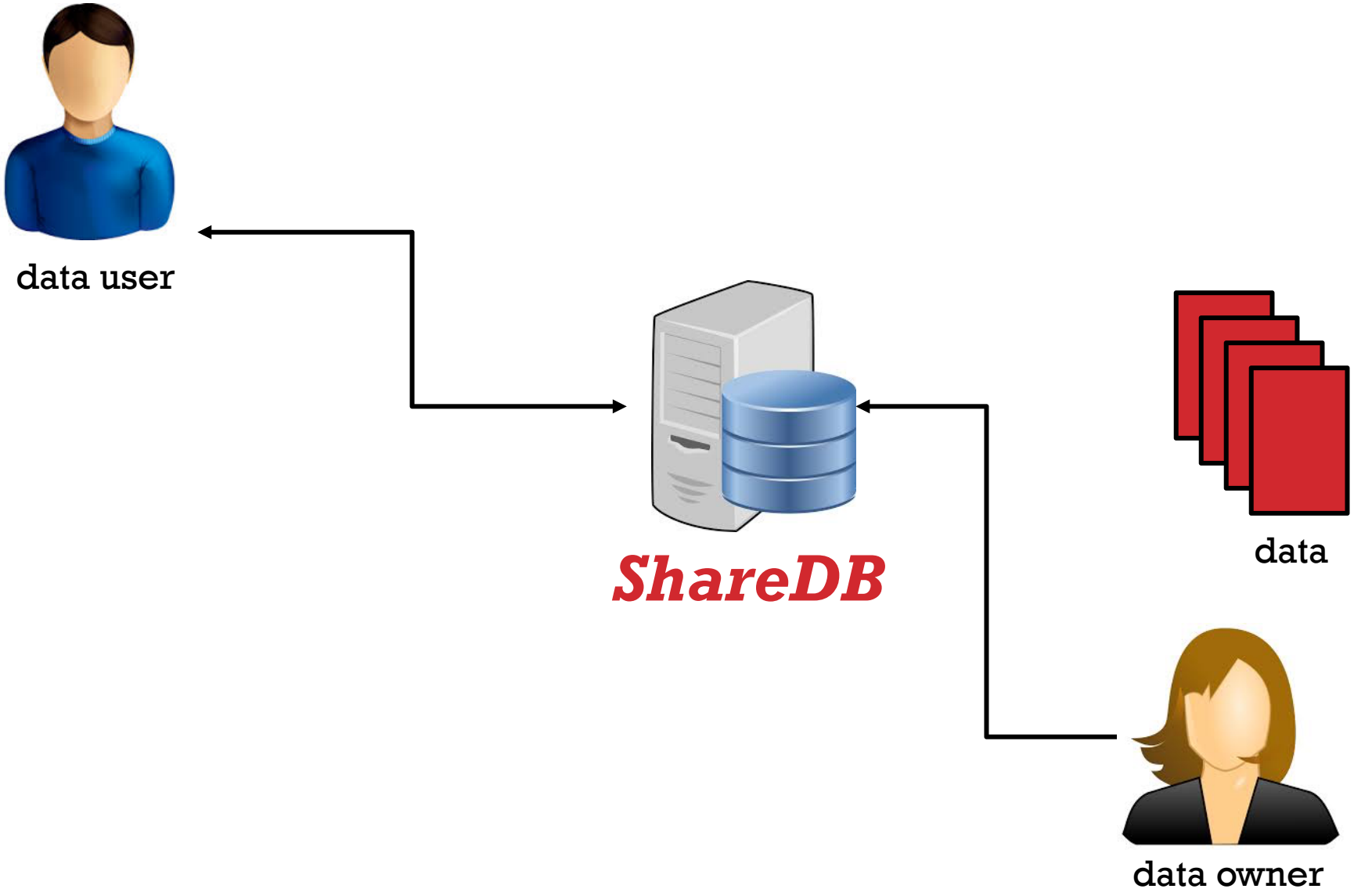
De-identification is a major obstacle for data sharing (e.g., HIPAA, FERPA, …)

Goal: Automatic De-identification

- Detect sensitive columns (rule catalog, user-defined, machine learning, …)
- Automatically de-identify



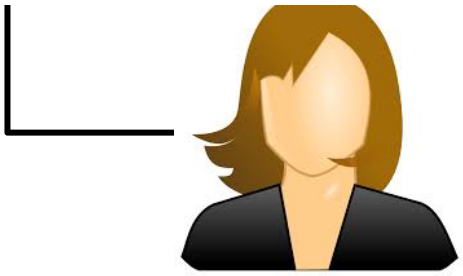
HIPAA: INTERACTIVE DE-IDENTIFICATION



HIPAA: INTERACTIVE DE-IDENTIFICATION

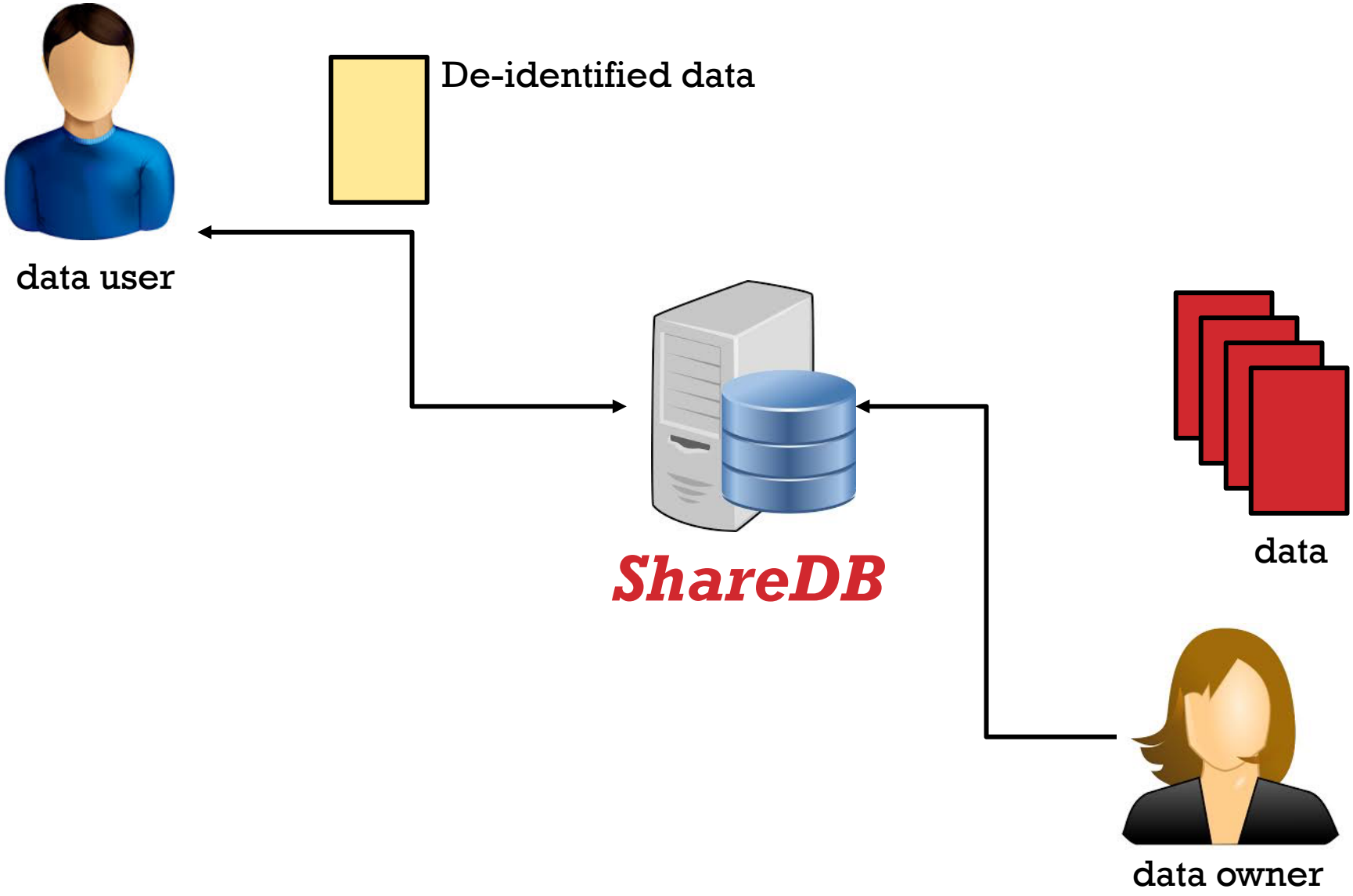
Id	ame	Street	City	State	P-Cod	Age
1	J Smith	123 University Ave	Seattle	Washington	98106	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle	WA	98101	19
6	James Smith	123 Univ Ave	Seatle	WA	NULL	41
7	J Widom	123 University Ave	Palo Alto	CA	94305	NULL
…	…	…	…	…	…	…

a



data owner

HIPAA: INTERACTIVE DE-IDENTIFICATION



ADDITIONAL TECHNICAL DIRECTIONS

- **Support support k-anonymous vies**
  - Upload or choose an existing generalization hierarchy

- **Support for differentially private views**
  - Currently support histogram-based views employing Laplacian noise
  - Adding support for automatically selecting the best differentially private mechanism given a target error rate
  - Based on analysis of data properties to determine best mechanism

NEXT STEPS

Next Data Sharing Spoke Workshop (Summer 2018)

Collect more agreements and create license framework 0.1

Extend tooling support, integrated into MIT datahub:

- **Watermarking**
- **Auditing**
- **Time-based fine-grained access control**
- **…**

Metadata support