Share BIG DATA

Metadata Solutions for Sharing Restricted Data

BIGG DATA: Balancing Impacts, Investments and Education

2017 ESS/SAES/ARD Fall Meeting: A Question of Balance

September 26, 2017

Jane Greenberg
Alice B Kroger Professor
CCI/Drexel

IIS/BD Spokes/Award #1636788

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

NSF

NORTHEAST
BIG DATA
INNOVATION HUB

Team members

- Alex Bertsch, grad. RA, MIT, Brown University
- Sam Madden, Lead PI, Massachusetts Institute of Technolog
- Carsten Binnig, PI, Brown University
- Sam Grabus, grad. RA, Drexel University
- Jane Greenberg, PI, Drexel University
- Hongwei Lu, grad. RA, Drexel University
- Famien Koko, grad. RA, MIT
- Tim Kraska, PI, Brown University
- Danny Weitzner, PI, MIT

Overview

1. Motivation for "A Licensing Model and Ecosystem for Data Sharing" (SPOKE work)
2. Northeast hub/Drexel workshop, "Enabling Seamless Data Sharing in Industry and Academia" (https://doi.org/10.17918/D8159V)
3. Towards standardized licenses
4. Prototype software platform development
5. Conclusions and next steps

NORTHEAST BIG DATA INNOVATION HUB

Data sharing advantages

Different Reasons

- More complete picture
- ROI
  - More data
  - More experts
  - Data reuse
- Better Insights into "Big Data"

NORTHEAST BIG DATA INNOVATION HUB

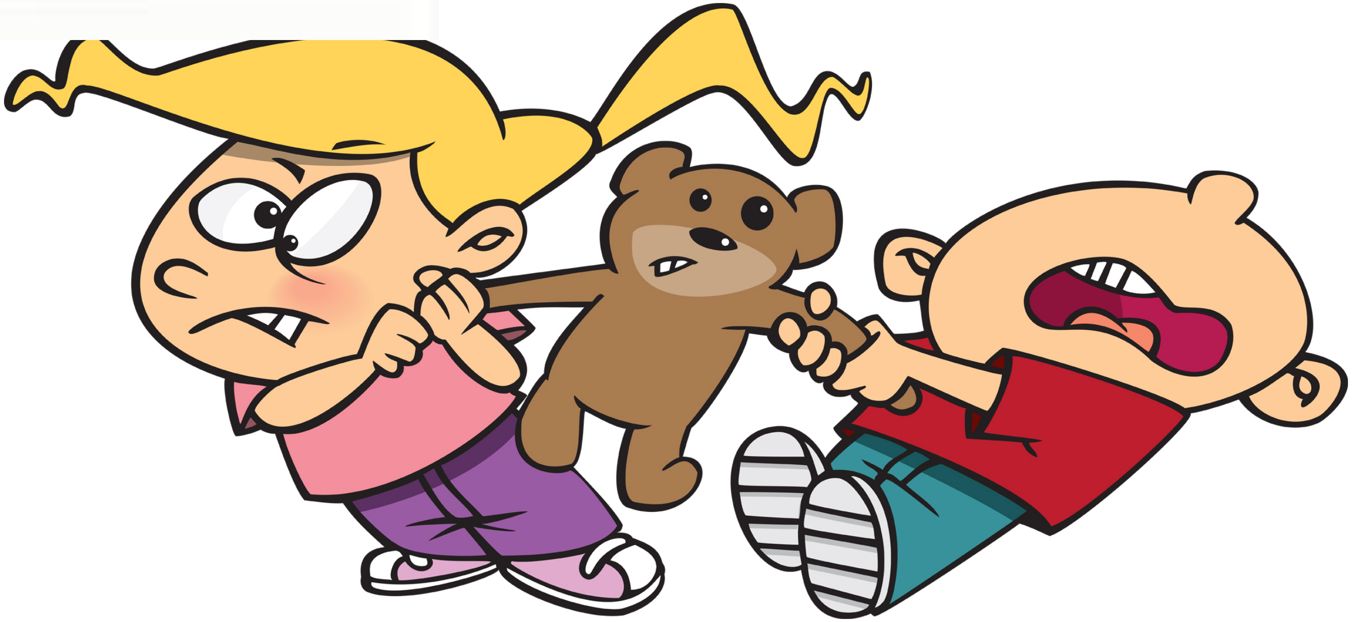
Open data/open science

DataONE

DFC DataNet FEDERATION CONSORTIUM

DRYAD

Data sharing barriers



Licensing, agreements, rights, privacy, policy, security, incentive…

BIG DATA INNOVATION HUB

Significant barriers to data sharing, particularly with **industry**…and other partners

1. Licensing, agreements

- "Creative commons" does not address need

2. Rights, privacy

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

Significant barriers to data sharing

3. Policy

- Complex regulations governing use of data in different domains

Data lifecycle - living thing

- Do not want to loose control over data downstream
- Has to be updated
- What if data is redacted?

NORTHEAST BIG DATA INNOVATION HUB

Significant barriers to data sharing

4. Security

- Technical and systematic aspects (~ policy, regulations, confidentiality/rights)

5. Incentives

- Why would someone go to all the effort to share their valuable data?

Still, merit in sharing

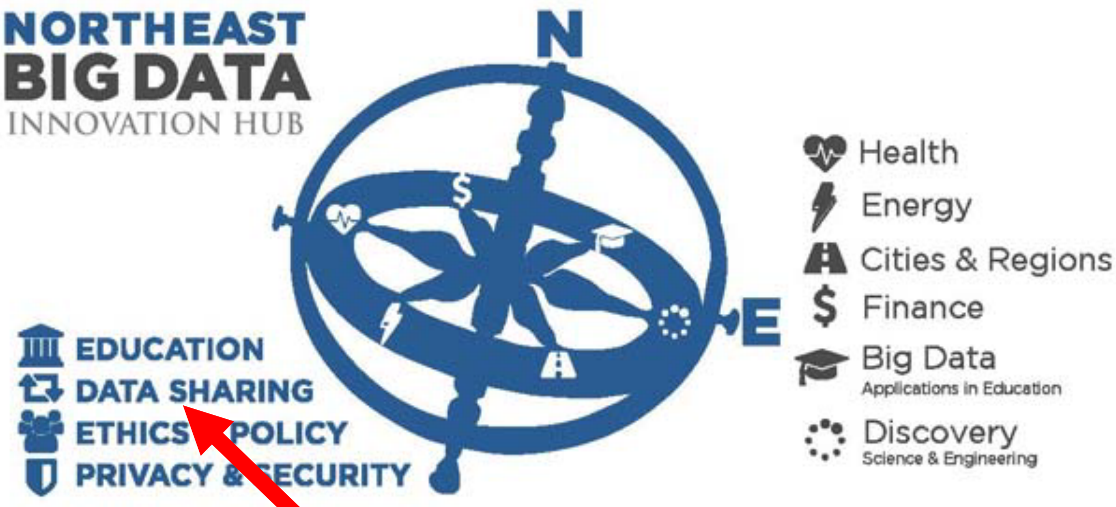
NORTHEAST BIG DATA INNOVATION HUB

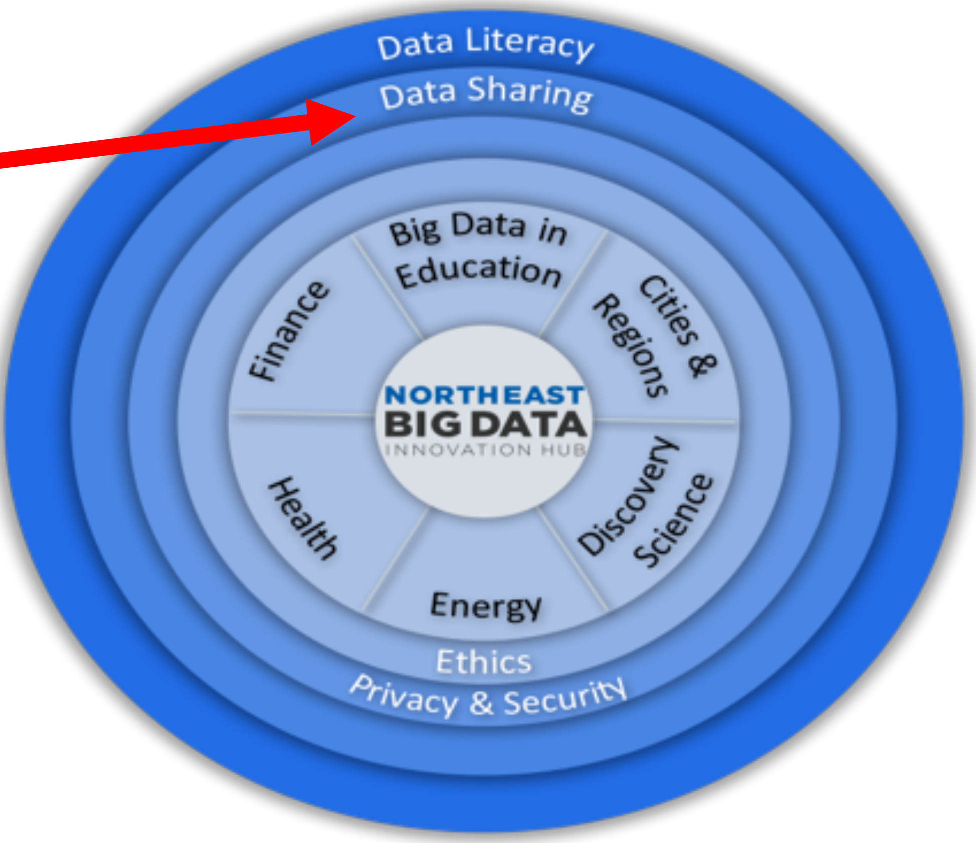
Sharing ‘restricted’ data today

- No sharing without a legal agreement

- Involve lawyers to create individual agreement!

NORTHEAST BIG DATA INNOVATION HUB





Spokes and rings

Co-Chairs
Jane Greenberg, Drexel
Sam Madden, MIT

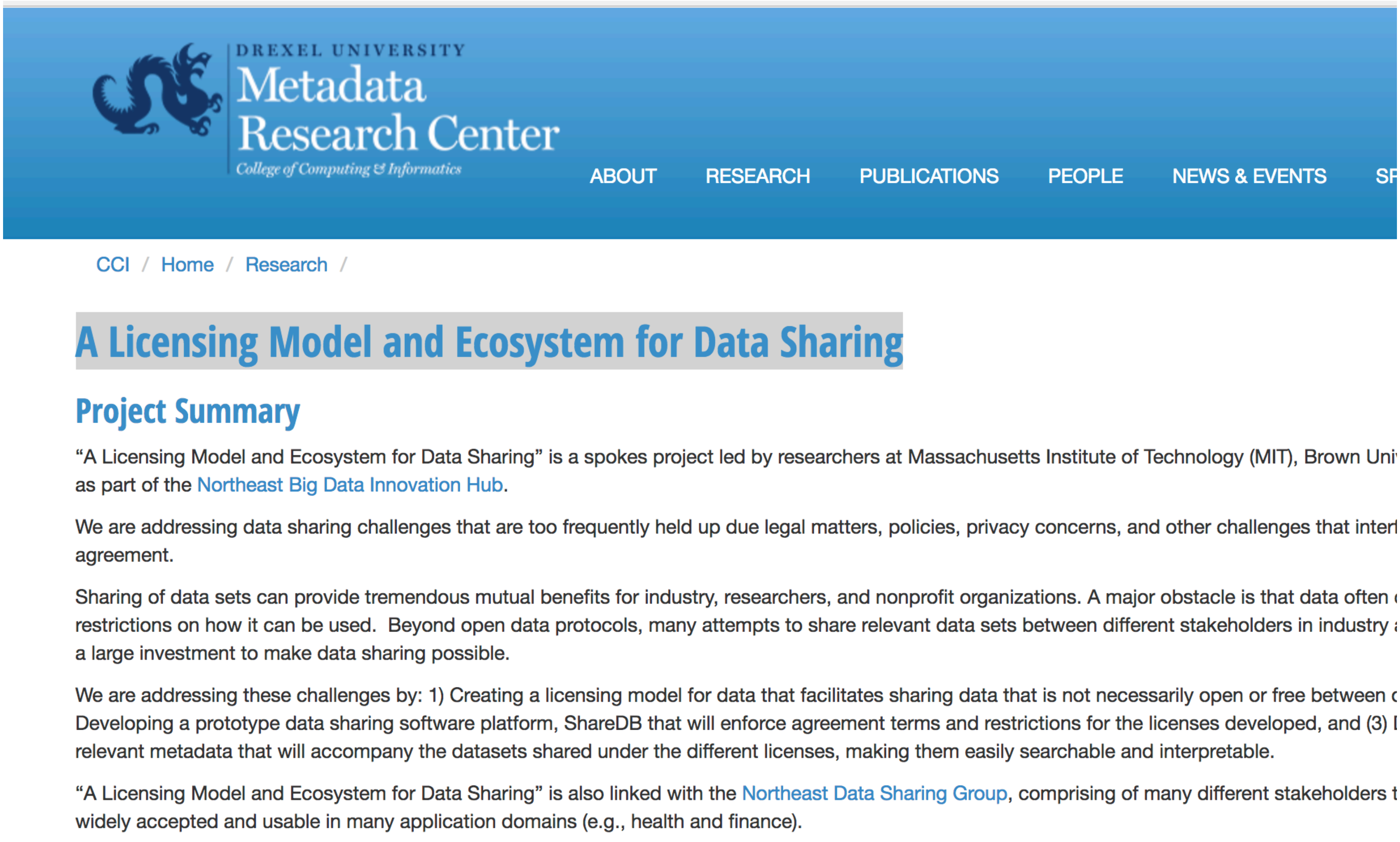
A Licensing Model and Ecosystem for Data Sharing

1. Licensing Framework / Generator
2. Data-Sharing Platform (Enforce Licenses)
3. Metadata (Search Licenses & Data)

- Principle: Solve the 80% case!

NORTHEAST BIG DATA INNOVATION HUB

http://cci.drexel.edu/mrc/projects/a-licensing-model-and-ecosystem-for-data-sharing/



A Licensing Model and Ecosystem for Data Sharing

Project Summary

"A Licensing Model and Ecosystem for Data Sharing" is a spokes project led by researchers at Massachusetts Institute of Technology (MIT), Brown Uni
as part of the Northeast Big Data Innovation Hub.

We are addressing data sharing challenges that are too frequently held up due legal matters, policies, privacy concerns, and other challenges that interf
agreement.

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often
restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry
a large investment to make data sharing possible.

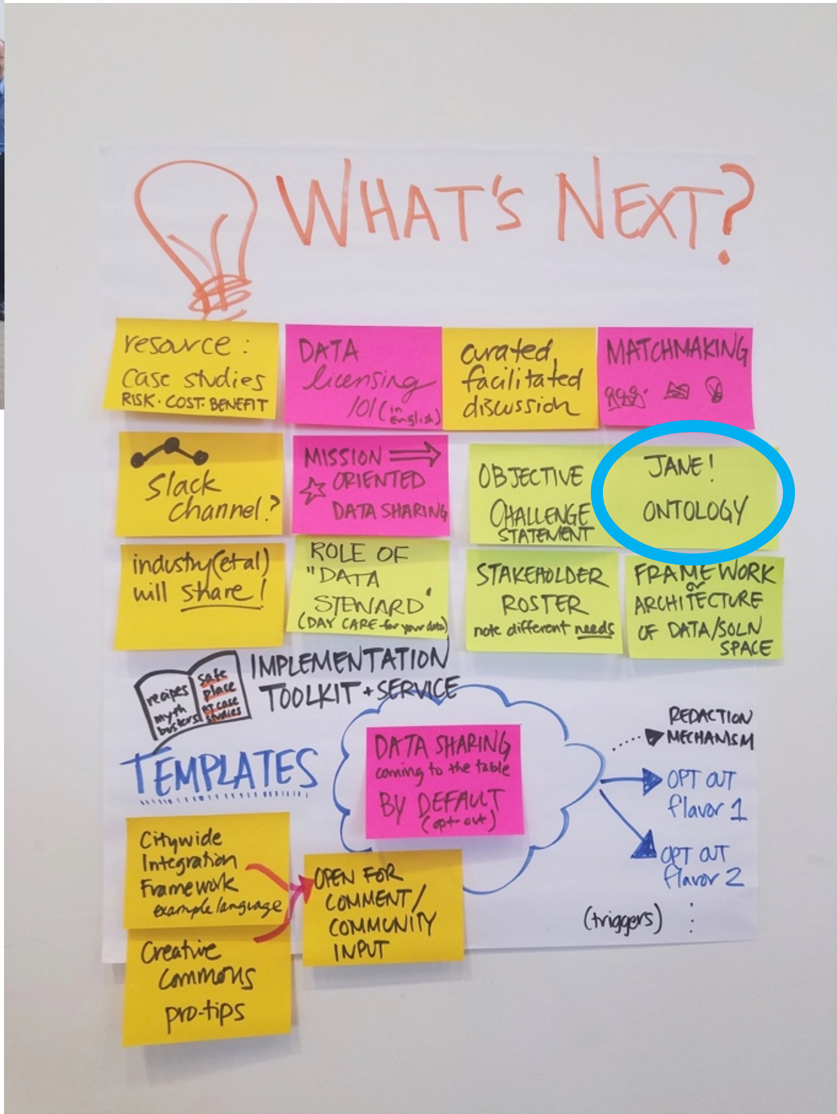
We are addressing these challenges by: 1) Creating a licensing model for data that facilitates sharing data that is not necessarily open or free between
Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and (3)
relevant metadata that will accompany the datasets shared under the different licenses, making them easily searchable and interpretable.

"A Licensing Model and Ecosystem for Data Sharing" is also linked with the Northeast Data Sharing Group, comprising of many different stakeholders t
widely accepted and usable in many application domains (e.g., health and finance).

Enabling Seamless Data Sharing in Industry and Academia (Fall 2017)

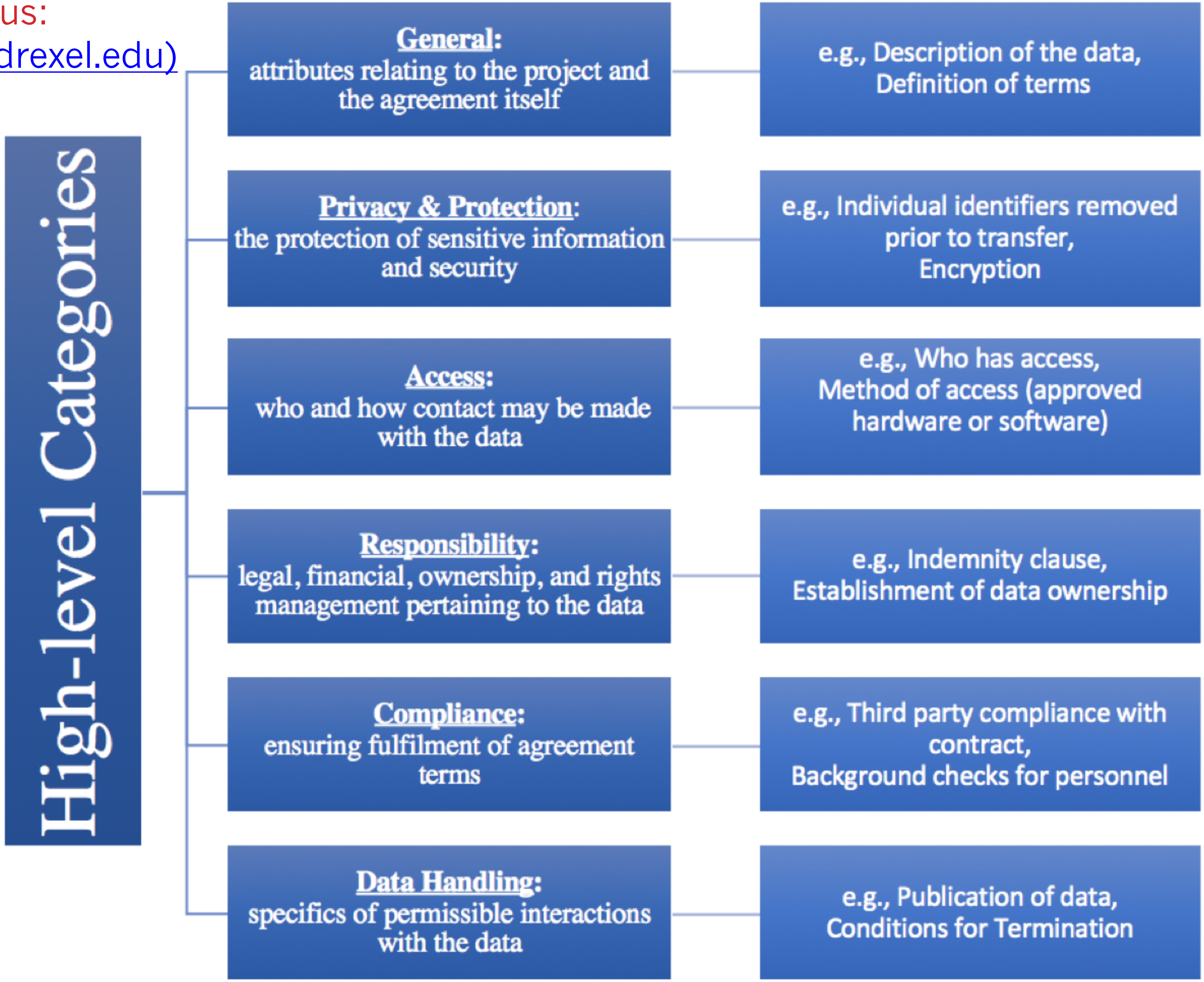
Heard from the trenches…

- Collect agreements
- Build a trusted platform
- Good metadata!


WHAT'S NEXT?
resource : case studies RISK · COST · BENEFIT
DATA licensing 101 (in english)
curated, facilitated discussion
MATCHMAKING
Slack channel?
MISSION ⟹ ORIENTED DATA SHARING
OBJECTIVE CHALLENGE STATEMENT
JANE! ONTOLOGY
industry (et al) will share!
ROLE OF "DATA STEWARD" (DAY CARE for your data)
STAKEHOLDER ROSTER note different needs
FRAMEWORK or ARCHITECTURE OF DATA/SOLN SPACE
IMPLEMENTATION TOOLKIT + SERVICE
recipes
myth busters
safe place
of case studies
TEMPLATES
DATA SHARING coming to the table BY DEFAULT (opt-out)
REDACTION MECHANISM
OPT OUT flavor 1
OPT OUT flavor 2
(triggers)
Citywide Integration Framework example language
OPEN FOR COMMENT/ COMMUNITY INPUT
Creative Commons pro-tips

Licenses: First Results

(Sam Grabus: smg383@drexel.edu)



High-level Categories

General: attributes relating to the project and the agreement itself

e.g., Description of the data, Definition of terms

Privacy & Protection: the protection of sensitive information and security

e.g., Individual identifiers removed prior to transfer, Encryption

Access: who and how contact may be made with the data

e.g., Who has access, Method of access (approved hardware or software)

Responsibility: legal, financial, ownership, and rights management pertaining to the data

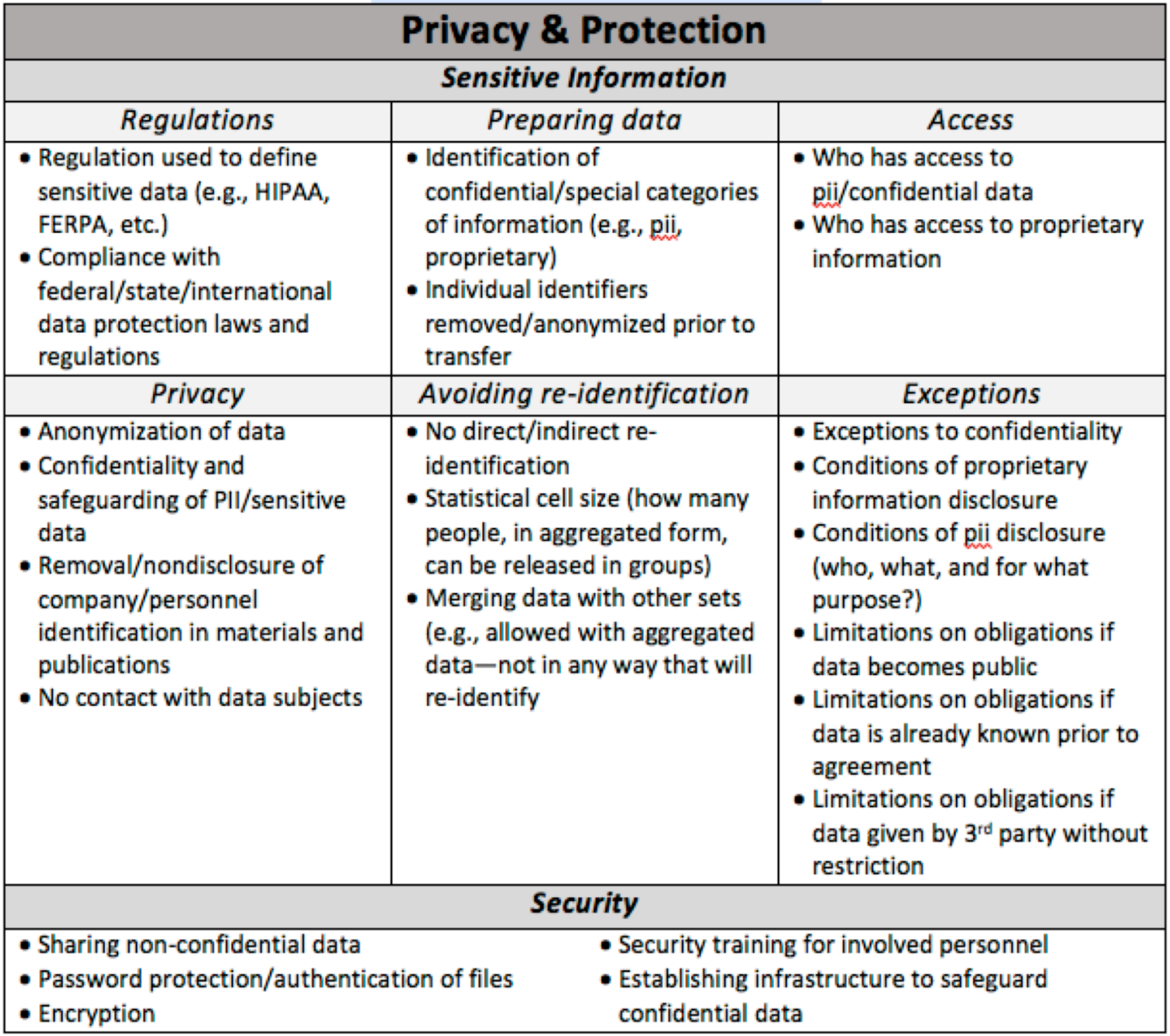
e.g., Indemnity clause, Establishment of data ownership

Compliance: ensuring fulfilment of agreement terms

e.g., Third party compliance with contract, Background checks for personnel

Data Handling: specifics of permissible interactions with the data

e.g., Publication of data, Conditions for Termination



Privacy & Protection

Sensitive Information

Regulations	*Preparing data*	*Access*
• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.) • Compliance with federal/state/international data protection laws and regulations	• Identification of confidential/special categories of information (e.g., pii, proprietary) • Individual identifiers removed/anonymized prior to transfer	• Who has access to pii/confidential data • Who has access to proprietary information
Privacy	***Avoiding re-identification***	***Exceptions***
• 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	• 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 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

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
```

System brainstorming, building on DBHub

Goal: Licensing Framework

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

- ☑ Controlled access
- □ Tracking of access
- ☑ Usage rights (e.g., publication, copying)
- □ Duration of use
- ☑ Warrantees of correctness/completeness/availability
- □ Other requirements

NORTHEAST BIG DATA INNOVATION HUB

Is this possible: Technology $\bowtie$ Sharing Agreements

Technical

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

Is this possible: Technology ⋈ Sharing Agreements

Technical

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

Is this possible: Technology ⋈ Sharing Agreements

Technical

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

Is this possible: Technology ⋈ Sharing Agreements

Technical

Access control & rights management

Expiration

Logging & auditing

Provenance/Fingerprinting

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

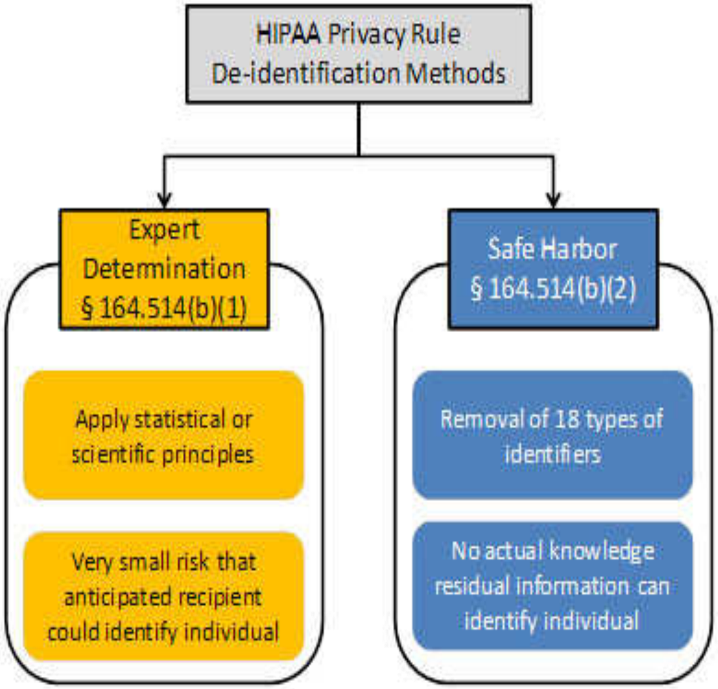
NORTHEAST BIG DATA INNOVATION HUB

Platform: **First Results**

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

De-identification tool

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

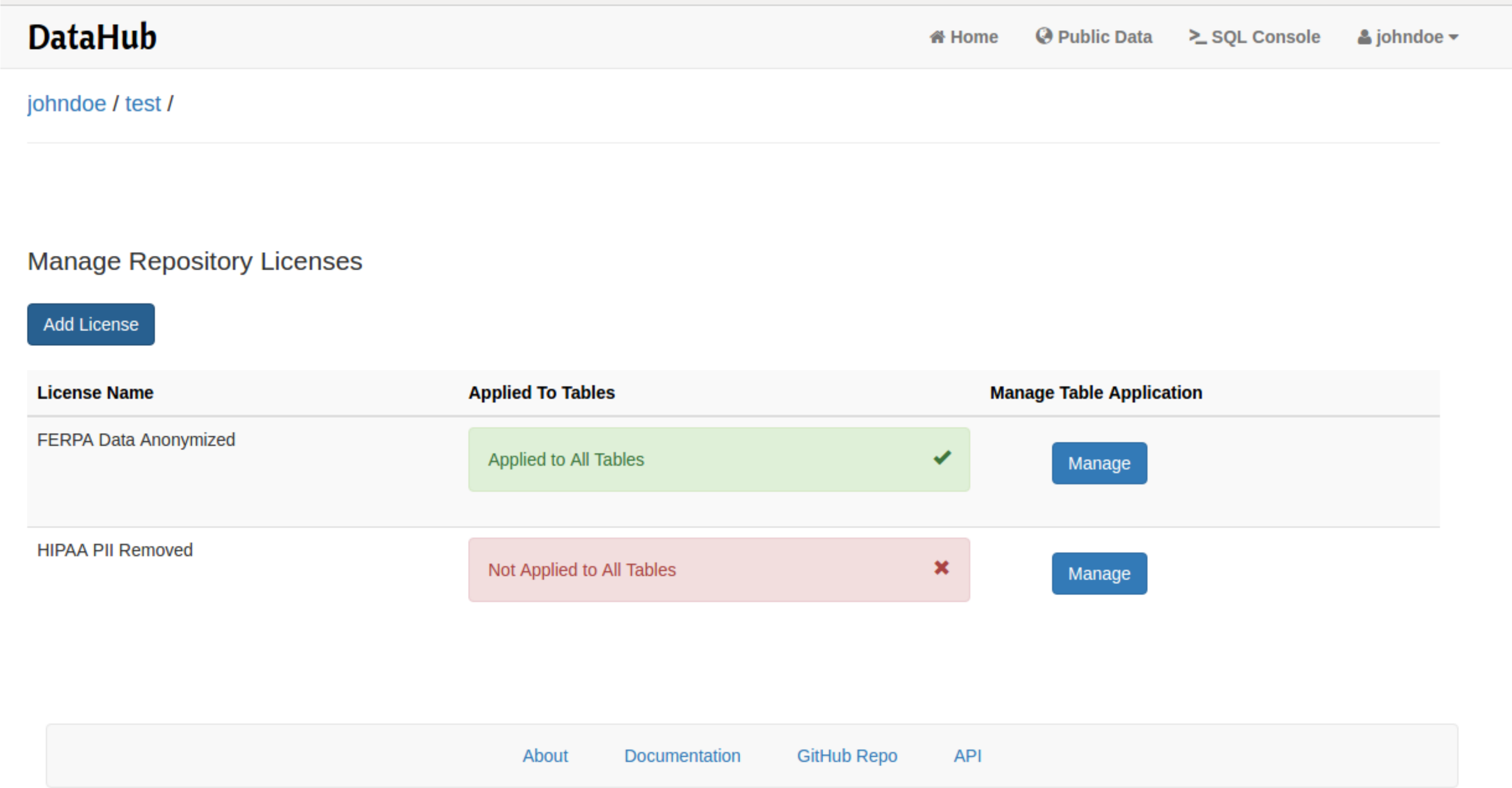

HIPAA Privacy Rule
De-identification Methods
Expert Determination § 164.514(b)(1)
Safe Harbor § 164.514(b)(2)
Apply statistical or scientific principles
Removal of 18 types of identifiers
Very small risk that anticipated recipient could identify individual
No actual knowledge residual information can identify individual

HIPAA: Interactive DE-identification

Id ✓	Name ⚠	Street ⚠	City ⚠	State ✓	P-Code ?	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

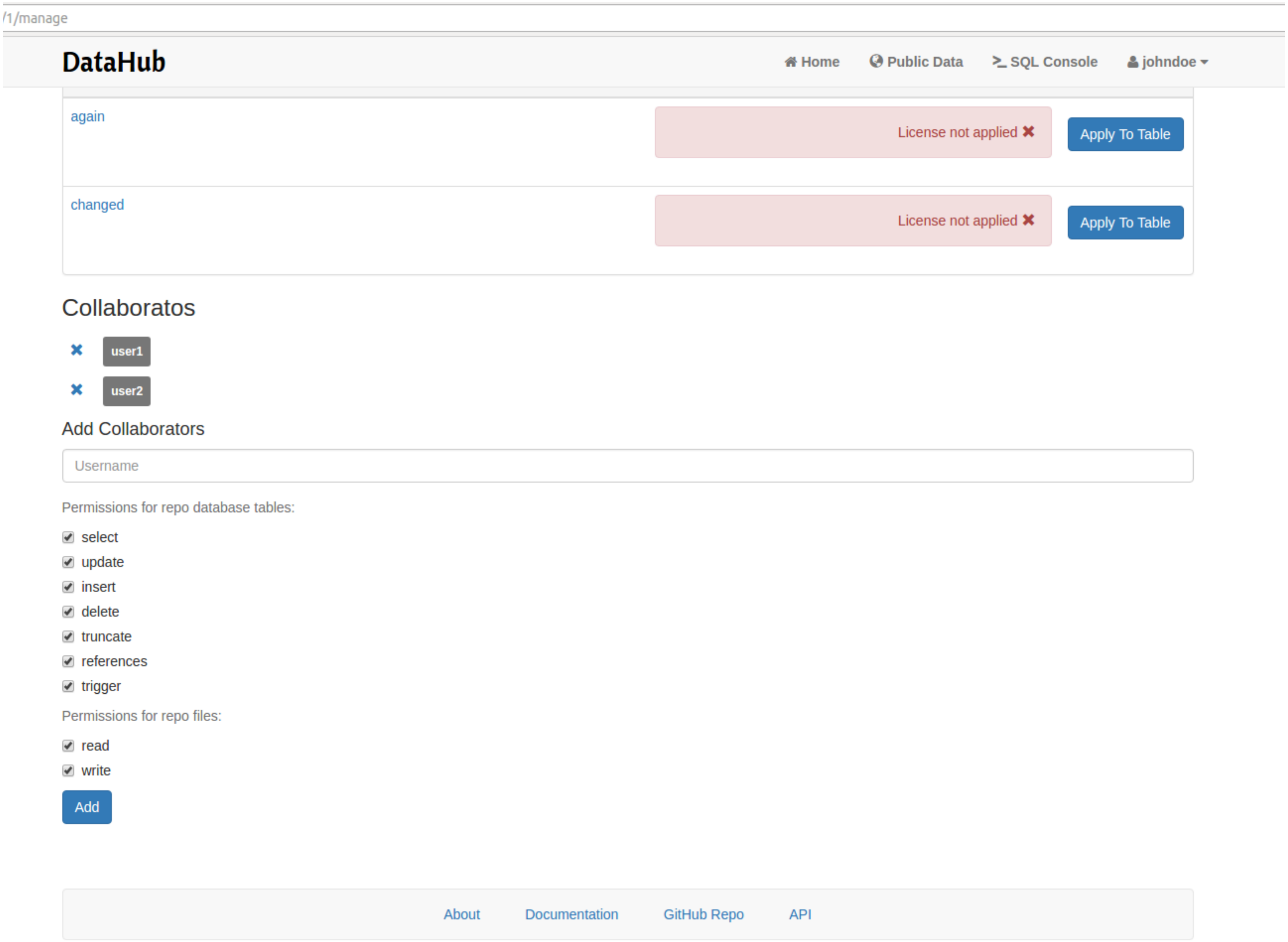
data owner



johndoe / test /

Manage Repository Licenses

License Name	Applied To Tables	Manage Table Application
FERPA Data Anonymized	Applied to All Tables ✓	Manage
HIPAA PII Removed	Not Applied to All Tables ✗	Manage



again

License not applied ✖

changed

License not applied ✖

Collaboratos

✖ user1

✖ user2

Add Collaborators

Permissions for repo database tables:

☑ select
☑ update
☑ insert
☑ delete
☑ truncate
☑ references
☑ trigger

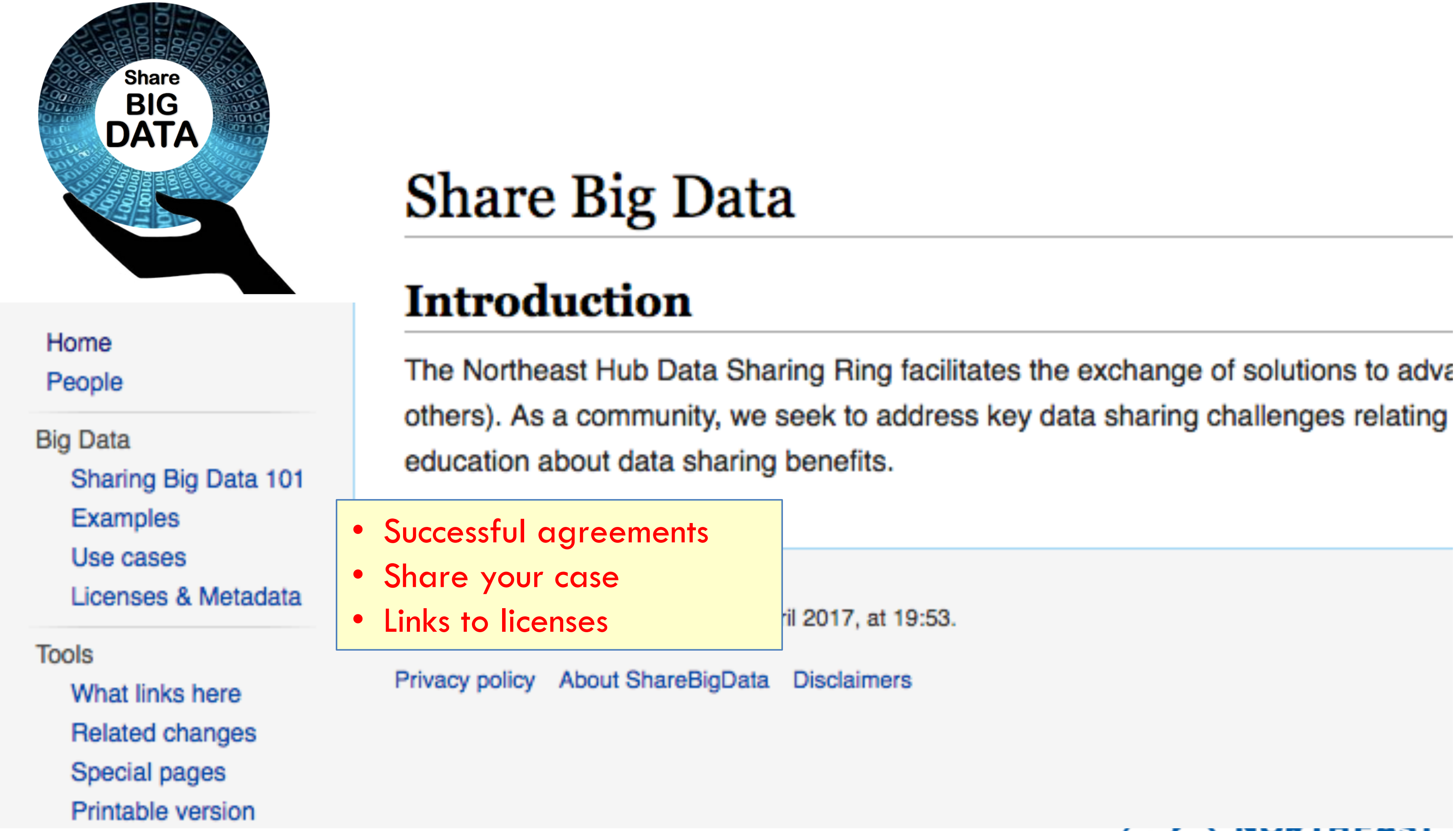
Permissions for repo files:

☑ read
☑ write

Conclusions

- Work underway, a lot of heavy lifting…
- Infrastructure to build on
- Metadata expertise
- Mining licenses shows great diversity
- Community building and connecting

https://cci.drexel.edu/ShareBigData



Share Big Data

Introduction

The Northeast Hub Data Sharing Ring facilitates the exchange of solutions to adva
others). As a community, we seek to address key data sharing challenges relating
education about data sharing benefits.

il 2017, at 19:53.

- Successful agreements
- Share your case
- Links to licenses

Final comment - Next Steps

- Data Sharing Spoke Workshop (Spring 2018)
  - Workshop agenda, slides: http://cci.drexel.edu/mrc/news/2016-11-bigdatahubworkshop/
  - Final report: *Enabling Seamless Data Sharing in Industry and Academia* is at: https://doi.org/10.17918/D8159V.
- Collect more agreements and create license framework 0.1 (Grabus, Sam smg383@drexel.edu)
- Extend tool support, continue prototyping
- IRB/RDA connection, metadata check-list

NORTHEAST
BIG DATA
INNOVATION HUB