A Licensing Model and Ecosystem for Data Sharing

- Alex Bertsch, grad. RA, MIT, Brown University
- Sam Madden, Lead PI, Massachusetts Institute of Technology
- 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

IIS/BD Spokes/Award #1636788
Overview

1. Motivation for “A Licensing Model and Ecosystem for Data Sharing” (SPOKE work)


3. Towards standardized licenses

4. Prototype software platform development

5. Conclusions and next steps
Data sharing advantages

Different Reasons

• More complete picture
• ROI
  • More data
  • More experts
  • Data reuse
• Better Insights into “Big Data”
Open data/open science
Data sharing barriers
Licensing, agreements, rights, privacy, policy, security, incentive...
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 lose control over data downstream
- Has to be updated
- What if data is redacted?
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
Sharing ‘restricted’ data today

• No sharing without a legal agreement

• Involve lawyers to create individual agreement!
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!
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 University as part of the Northeast Big Data Innovation Hub.

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

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often has restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry make 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 organizations. 2) Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and 3) Integrating 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 that work towards creating a 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!
Licenses: First Results
(Sam Grabus:
smg383@drexel.edu)
<table>
<thead>
<tr>
<th>Privacy &amp; Protection</th>
<th>Sensitive Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regulations</strong></td>
<td><strong>Preparing data</strong></td>
</tr>
<tr>
<td>• Regulation used to define sensitive data</td>
<td>• Identification of</td>
</tr>
<tr>
<td>(e.g., HIPAA, FERPA, etc.)</td>
<td>confidential/special</td>
</tr>
<tr>
<td>• Compliance with federal/state/international</td>
<td>categories of</td>
</tr>
<tr>
<td>data protection laws and regulations</td>
<td>information (e.g.,</td>
</tr>
<tr>
<td></td>
<td>pii, proprietary)</td>
</tr>
<tr>
<td></td>
<td>• Individual identifiers</td>
</tr>
<tr>
<td></td>
<td>removed/anonymized</td>
</tr>
<tr>
<td></td>
<td>prior to transfer</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td><strong>Avoiding re-identification</strong></td>
</tr>
<tr>
<td>• Anonymization of data</td>
<td>• No direct/indirect re-identification</td>
</tr>
<tr>
<td>• Confidentiality and safeguarding of PII/sensitive</td>
<td>• Statistical cell size (how many</td>
</tr>
<tr>
<td>data</td>
<td>people, in aggregated</td>
</tr>
<tr>
<td>• Removal/nondisclosure of company/personnel</td>
<td>form, can be released</td>
</tr>
<tr>
<td>identification in materials and publications</td>
<td>in groups</td>
</tr>
<tr>
<td>• No contact with data subjects</td>
<td>• Merging data with other sets</td>
</tr>
<tr>
<td></td>
<td>(e.g., allowed with aggregated</td>
</tr>
<tr>
<td></td>
<td>data—not in any way that will</td>
</tr>
<tr>
<td></td>
<td>re-identify</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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
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
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
<table>
<thead>
<tr>
<th>Technical</th>
<th>Agreement Clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access control &amp; rights management</td>
<td>Controlled access (who &amp; where)</td>
</tr>
<tr>
<td>Expiration</td>
<td></td>
</tr>
<tr>
<td><strong>Logging &amp; auditing</strong></td>
<td><strong>Tracking of access</strong></td>
</tr>
<tr>
<td>Provenance/Fingerprinting</td>
<td>Usage rights (e.g., publication, <strong>copying</strong>)</td>
</tr>
<tr>
<td>De-identification</td>
<td>Duration of use</td>
</tr>
<tr>
<td>“Noising”</td>
<td>Warrantees of correctness/completeness/availability</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Other requirements</td>
</tr>
</tbody>
</table>
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
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: Interactive DE-identification

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Street</th>
<th>City</th>
<th>State</th>
<th>P-Code</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J Smith</td>
<td>123 University Ave</td>
<td>Seattle</td>
<td>Washington</td>
<td>98106</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>Mary Jones</td>
<td>245 3rd St</td>
<td>Redmond</td>
<td>WA</td>
<td>98052-1234</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Bob Wilson</td>
<td>345 Broadway</td>
<td>Seattle</td>
<td>Washington</td>
<td>98101</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>M Jones</td>
<td>245 Third Street</td>
<td>Redmond</td>
<td>NULL</td>
<td>98052</td>
<td>299</td>
</tr>
<tr>
<td>5</td>
<td>Robert Wilson</td>
<td>345 Broadway St</td>
<td>Seattle</td>
<td>WA</td>
<td>98101</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>James Smith</td>
<td>123 Univ Ave</td>
<td>Seattle</td>
<td>WA</td>
<td>NULL</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>J Widom</td>
<td>123 University Ave</td>
<td>Palo Alto</td>
<td>CA</td>
<td>94305</td>
<td>NULL</td>
</tr>
</tbody>
</table>

...
### Manage Repository Licenses

<table>
<thead>
<tr>
<th>License Name</th>
<th>Applied To Tables</th>
<th>Manage Table Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERPA Data Anonymized</td>
<td>Applied to All Tables</td>
<td>Manage</td>
</tr>
<tr>
<td>HIPAA PII Removed</td>
<td>Not Applied to All Tables</td>
<td>Manage</td>
</tr>
<tr>
<td>again</td>
<td>License not applied</td>
<td>Apply To Table</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>changed</td>
<td>License not applied</td>
<td>Apply To Table</td>
</tr>
</tbody>
</table>

**Collaborators**

- user1
- user2

**Add Collaborators**

Username

Permissions for repo database tables:
- select
- update
- insert
- delete
- truncate
- references
- trigger

Permissions for repo files:
- read
- write

Add
Conclusions

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