Data Sharing & Privacy Controls in LearnSphere

Carnegie Mellon: Ken Koedinger, John Stamper, & Carolyn Rose
MIT: Una-May O’Reilly & Kalyan Veeramachaneni
Stanford: Candace Thille
U of Memphis: Phil Pavlik

Support from NSF Cyberinfrastructure, DIBBs, $5M for 5 years

EDM Policy Workshop: July, 2018
A community data infrastructure to support learning improvement online.
LearnSphere Key Points

• Share data and analytic methods
• Data analytics are improving learning
• Data & analytic curation is complex
  – Inherent ambiguities in variable definition
  – Sharing forces better data & analytics
• Privacy management depends on reidentification risks & changes access
Data repositories

DataShop
The LearnLab DataShop is a data repository and web application for learning sciences researchers. It provides secure data storage as well as an array of analysis and visualization tools available through a web-based interface. DataShop was funded by a National Science Foundation grants (SBE-0836012, SBE-0354420) to LearnLab, the Pittsburgh Science of Learning Center.

DataStage
DataStage is provided by the Vice Provost Office for Online Learning (VPOL) at Stanford, which facilitates the teaching of online classes. The instruction delivery platforms are instrumented to collect a variety of data around participants’ interaction with the study material. Examples are participants manipulating video players as they view portions of a class, solution submissions to problem sets, uses of the online forum available for some classes, peer grading activities, and some demographic data. VPOL makes some of this data available for research on learning processes, and for explorations into improving instruction through Datastage.

ASSISTments Data
The ASSISTments data repository contains data from a tutoring system, in many cases as part of online experiments of what learning works best. You can z/tae submit studies at www.assistmentsstembed.org as well as get a lot of information on how to interpret your data.

Databrary
The Databrary project aims to promote data sharing, archiving, and reuse among researchers who study the development of humans and other animals. The project focuses on creating tools for scientists to store, manage, preserve, analyze, and share video and other forms of data, and is supported by funding from New York University and at Penn State. The U.S. National Science Foundation (NSF BC-1328599) and the U.S. National Institutes of Health (NIH U01-HD-076590) have provided the funding for this project.

TalkBank
TalkBank is an interdisciplinary research project to promote the study of human and animal communication. The subfields of study include first language acquisition, second language acquisition, conversation analysis, classroom discourse and aphasic language. TalkBank has been funded by grants from the National Science Foundation (including BCS-0998306, 0324888) as well as the National Institutes of Health.

CHILDES
The Child Language Data Exchange System (CHILDES) is the part of TalkBank focused on child language, or first language acquisition. CHILDES provides tools for studying conversational interactions, including a transcripts database, programs for analyzing transcripts, methods for linguistic coding and systems for linking audio and video. CHILDES is supported by grants from the National Institutes of Health (R01-HD23998, R01-HD051698).  

MOOcdb
The MOOcdb project aims to bring together education researchers, computer science researchers, machine learning researchers, technologists, database and big data experts to advance MOOC data science. The project founded at MIT includes a platform agnostic functional data model for data exhaust from MOOCs, a collaborative-open source-open access data visualization framework, a crowd sourced knowledge discovery framework and a privacy preserving software framework. The team is currently working to release a number of these tools and frameworks as open source.

DiscourseDB
DiscourseDB is a data infrastructure project, in the space of collaborative and Discussion-based learning, that aims to provide a common data model to accommodate diverse sources including but not limited to Chat, Threaded Discussions, Blogs, Twitter, Wikis and Text messaging. In the future, the project will make available analytics which will facilitate research questions related to the mediating and moderating effects of role taking, help exchange, collaborative knowledge construction and others.

DataShop External Tools
Free tools submitted by developers in the educational data mining and intelligent tutoring systems communities.

Ways to Generate Data
Online Content Providers

Educational Technology Development Tools

Online Assessment and Tutoring Systems
DataShop: Not just “big”, but fine, wide, & deep!

A community data infrastructure to support learning improvement online.

Existing Resources

DataShop@CMU
DataShop@Memphis
DiscourseDB
MOOCDB
DATASTAGE
LearnSphere.org
Welcome to DataShop, the world's largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

I'm a
- Data miner/computer scientist
- Cognitive scientist
- ITS/AIED researcher
- User modeling researcher
- Educational psychologist
- Course developer
- Psychometrician
- Learning analytics researcher

Here are topics of interest
- Test a theory of performance or learning
- Applications of Bayesian modeling
- Multiple skills
- Modeling the rate of learning
- Detecting motivation or engagement
- Discovering knowledge component/skill/cognitive/student models

What is DataShop?

1600+ data sets
math, science, language ...

K12 & college
Welcome to DataShop, the world's largest repository of learning interaction data.

What can I do with DataShop?

1600+ data sets
math, science, language ...

K12 & college
Welcome to DataShop, the world's largest repository of learning interaction data.

What can I do with DataShop?

Upload a dataset

Project
- Add this dataset to...
- new project
- existing project
- choose later

Project Name
- Psychology MOOC data

Data Collection Type
- Not specified
- Not human subjects data (not original)
- Study data collected under an IRB
- Study data collected under an IRB

Dataset Name
- 2013 Psych

Description (optional)

1600+ data sets
- math, science, language ...

K12 & college

http://learnlab.org/datashop
Data to improve learning

- Discover better models of learners
  - Data >> intuition alone
  - Design & deploy better learning activities
- Detect & remediate disengagement
- Improve assessment
- Improve MOOCs

Sharing leverages interdisciplinary interaction
LearnSphere Key Points

• Share data and analytic methods
• Data analytics are improving learning
• Data & analytic curation is complex
  – Inherent ambiguities in variable definition
  – Sharing forces better data & analytics
• Privacy management depends on reidentification risks & changes access
Privacy risks of kinds of data & availability

Access Level
- Publicly available
- Private-expanding
- Research team only

<table>
<thead>
<tr>
<th>Access Level</th>
<th>Risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>Medium</td>
</tr>
</tbody>
</table>

De-identification & preventing re-identification
Privacy risks of kinds of data & availability

Access Level
- Publicly available
- Private-expanding
- Research team only

Risks
- Easy
- Medium
- Hard

De-identification & preventing re-identification

Examples:
- Click-stream
  - No demographic
  - No linking
- Disc board
  - Blur names
- Audio video
  - No blur
Privacy risks of kinds of data & availability

<table>
<thead>
<tr>
<th>Access Level</th>
<th>Risks</th>
<th>De-identification &amp; preventing re-identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publically available</td>
<td>Little</td>
<td>Easy</td>
</tr>
<tr>
<td>Private-expanding</td>
<td>Lots</td>
<td>Medium</td>
</tr>
<tr>
<td>Research team only</td>
<td>Too much</td>
<td>Hard</td>
</tr>
</tbody>
</table>

- Examples:
  - Click-stream: No demographic, No linking
  - Disc board: Blur names
  - Audio video: No blur
Welcome to DataShop, the world's largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

Upload a dataset

Project: Add this dataset to...
- new project
- existing project
- choose later

Project Name: Psychology MOOC data

Data Collection Type:
- Not specified
- Not human subjects data (not originally collected for research purposes)
- Study data collected under an IRB where consent was not required (IRB approval letter required)
- Study data collected under an IRB where consent was required (IRB approval letter and consent form required)

Dataset Name: 2013 Psych

Description (optional)

1600+ data sets
math, science, language ...

K12 & college
Projects Have 3 Possible States

- **Private**: Only PI/data owner may access.
- **Public**: Any registered DataShop user may access project freely.
- **Not-Shareable**: Only PI/data owner decides on case by case basis whether to share with non-team members.

**Research Manager** determines shareability.

**Data Owner** determines private/public state.

Default state: Private and Not-Shareable.
Requirements for “Shareable” Designation

1. **Data collection types**
   
   a. **Not human subjects** (data not originally collected for research purposes, e.g. course data)
      
      - Data is de-identified

   b. **Study data** collected under an IRB where consent not required
      
      - Data is de-identified
      - IRB approval letter
      - Consent form
      - Non-prohibitive consent text
Project filters

Useful default filter: Needs Attention
DALMOOC project’s IRB page

1. researcher (or research manager) specifies “Data Collection Type”

2. researcher (or research manager) adds IRB information and uploads IRB documents

3. After reviewing IRB docs, research manager designates Shareability status
DALMOOC project’s Dataset page

- Research manager accesses dataset settings
- Access dataset to confirm de-identification (if needed)
LearnSphere/DataShop Progress

- Web portal for data & method sharing
- Analytic workflow beta
  - Cross language (R, Python, Java, C, Matlab, ...) integration
  - Non-programmer recombination
  - Integration projects
    - DataShop + MOOCdb + DataStream
      Apply doer effect workflow analytics to MOOCs
    - ... + DiscourseDB
      Analyze discussion board posts & video, activities
- Distributed services
  (LearnSphere@your_institution)
- Privacy control procedures & software
Thank you!

Carnegie Mellon: Ken Koedinger, John Stamper, & Carolyn Rose
MIT: Una-May O’Reilly & Kalyan Veeramachaneni
Stanford: Candace Thille
U of Memphis: Phil Pavlik

Support from NSF Cyberinfrastructure, DIBBs, $5M for 5 years

Asilomar 2: June 16, 2016
Extra slides
DataShop Terminology

• **Shareability**—determined by research manager (RM):
  
  § **Not-Shareable** = DataShop does not give the data owner the option of sharing their project outside their research team
  
  § **Shareable** = DataShop gives the data owner the option of sharing their project with people outside their team (regardless of whether data owner keeps their project private or makes it public).

• **Private vs Public**—determined by data owner (contingent on RM shareability designation):
IRB Review Interface built by DataShop team and used by DataShop Research Manager
“Needs Attention” triggered when...

- Researcher/data provider creates new project

- Researcher/data provider adds a new dataset to an “old” project after project had been designated “shareable”
Dataset page of a project that has multiple datasets. This project does not need attention.
Additional IRB management features

- one IRB can cover multiple projects: IRB entry page gives PI or RM option of applying existing IRB to new project or adding a new IRB.
  - “All IRB” page gives RM a listing of all IRBs and the projects associated with them.

- one project can have multiple IRBs related to it: A project’s IRB page shows all relevant IRB info w/links to documents
System highlights any **PUBLIC** projects that need attention in yellow.
Vast space of data & questions => need a data infrastructure to integrate => produce discoveries not possible within current data silos
$47M from NSF => 850+ datasets

Ed tech + wide use = “Basic research at scale”

Since 2004

Ø 1600 ed tech data sets in DataShop
Ø 360 in vivo experiments

Visualizing learning curves to find opportunities for improvement

High rough curve
=> revise skill model
=> redesign instruction
=> do A/B test

Better student learning!

Data Sharing & Privacy Controls in LearnSphere

Carnegie Mellon: Ken Koedinger, John Stamper, & Carolyn Rose
MIT: Una-May O’Reilly & Kalyan Veeramachaneni
Stanford: Candace Thille
U of Memphis: Phil Pavlik

Support from NSF Cyberinfrastructure, DIBBs, $5M for 5 years

EDM Policy Workshop: July, 2018
A community data infrastructure to support learning improvement online.

Existing Resources

DataShop @CMU
DataShop @Memphis
MOOCDB
DiscourseDB
DATASTAGE
LearnSphere Key Points

- Share data and analytic methods
- Data analytics are improving learning
- Data & analytic curation is complex
  - Inherent ambiguities in variable definition
  - Sharing forces better data & analytics
- Privacy management depends on reidentification risks & changes access

Others: DataBrary, Upenn
Brian MacWhinney
Data repositories

DataShop

DataShop is a data repository and web application for learning science researchers. It provides secure data storage as well as an array of analytics and visualization tools available through a web-based interface. DataShop was funded by a National Science Foundation grant (DUE 0941158, DUE-0943500) to LearnLab, the Pittsburgh Science of Learning Center.

DataStage

DataStage is provided by the Vice Provost Office for Online Learning (UPO) at Stanford, which facilitates the teaching of online classes. The instruction delivery platforms are instrumented to collect a variety of data around participants' interaction with the study material. Examples are participants' navigating video players as they view portions of a class, navigation attributes to problem sets, uses of the online forums available for some classes, peer posting activities, and some demographic data. UPO makes some of this data available for research on learning processes, and for explorations into improving instruction through DataStage.

ASSISTments Da:

The ASSISTments data report is a very useful tool for instructors. We can view student performance in real-time and make adjustments to our instruction. ASSISTments is a tool that allows for the assessment of student learning and provides feedback on performance, analysis, and strategies.

Databrary

The Databrary project aims to improve data sharing, archiving, and reuse among researchers who study the development of human skills. The project provides a platform for sharing data that is easy to use and understand. The project includes a variety of datasets related to human development, including language acquisition, and classroom interactions.

TalkBank

TalkBank is an interdisciplinary research project that promotes the study of human and animal communication. The subfields of study include the language acquisition, second language acquisition, conversation analysis, classroom discourse and pedagogical language. TalkBank has been funded by grants from the National Science Foundation (including BCS-0849533, 0954668) as well as the National Institutes of Health.

CHILDES

The Child Language Data Exchange System (CHILDES) is the part of TalkBank focused on oral language, or first language acquisition. CHILDES provides tools for studying conversational interactions, including a transcriber database, programs for analyzing transcriptions, methods for linguistic coding and systems for linking words and cases. CHILDES is supported by grants from the National Institute of Health (Y01 HD32886, R01 HD45698).

Data Processing and Analytic Methods

MOOCdb

The MOOCdb project aims to bring together education researchers, computer science researchers, machine learning researchers, technologists, database and big data experts to advance MOOC data science. The project is funded at WF by the National Science Foundation grant (DUE-1525393) to LearnLab, the Pittsburgh Science of Learning Center.

DiscourseDB

DiscourseDB is a data infrastructure project, in the space of collaborative and discussion-based learning, that seeks to provide a common data model to accommodate diverse sources including but not limited to Chat, Threaded Discussions, Blogs, Twitter, Wiki and Text messages. The project will make available analytics which will facilitate research questions related to the marketing and understanding of these tools, help exchange, collaboration, knowledge construction and other.

DataShop External Tools

The DataShop tool is designed for developers in the data analysis and educational technology fields. It offers a variety of features for analyzing data and improving learning.

EDM Workbench

The EDM Workbench is an emerging intellectual data command that helps researchers continue a continuous cycle in student learning and teaching by providing support for learning data and improving their use.

Educational Technology Development Tools

Online Assessment and Tutoring Systems

math tutor  Second Language Tutors
DataShop: Not just “big”, but fine, wide, & deep!

A community data infrastructure to support learning improvement online.
Welcome to DataShop, the world’s largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

I’m a
- Data miner/computer scientist
- Cognitive scientist
- ITS/AIED researcher
- User modeling researcher
- Educational psychologist
- Course developer
- Psychometrician
- Learning analytics researcher

Here are topics of interest (show all)
- Test a theory of performance or learning
- Applications of Bayesian modeling
- Multiple skills
- Modeling the rate of learning
- Detecting motivation or engagement
- Discovering knowledge component/skill/cognitive tutor models

What is DataShop?

1600+ data sets
math, science, language ...
K12 & college
Welcome to DataShop, the world’s largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

Upload a dataset

Add this dataset to ...  
- new project  
- existing project  
- choose later

Project Name: Psychology MOOC data

Data Collection Type:
- Not specified
- Not human subjects data (not originally collected for research purposes)
- Study data collected under an IRB where consent was not required (IRB approval letter required)
- Study data collected under an IRB where consent was required (IRB approval letter and consent form required)

Dataset Name: 2013 Psych

Description (optional):  
Recent dataset names  
Recent descriptions

1600+ data sets
math, science, language ...
K12 & college
Welcome to DataShop, the world's largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

Upload a dataset

Project

Add this dataset to:
- new project
- existing project
- choose later

Project Name: Psychology MOOC data

Data Collection Type:
- Not specified
- Not human subjects data (not original)
- Study data collected under an IRB
- Study data collected under an IRB

Dataset Name: 2013 Psych

Description (optional):

1600+ data sets
math, science, language ...
K12 & college

http://learnlab.org/datashop
Data to improve learning

- Discover better models of learners
  - Data >> intuition alone
  - Design & deploy better learning activities
- Detect & remediate disengagement
- Improve assessment
- Improve MOOCs

Sharing leverages interdisciplinary interaction
LearnSphere Key Points

- Share data and analytic methods
- Data analytics are improving learning
- Data & analytic curation is complex
  - Inherent ambiguities in variable definition
  - Sharing forces better data & analytics
- Privacy management depends on reidentification risks & changes access

Others: DataBrary, Upenn
Brian MacWhinney
Privacy risks of kinds of data & availability

Access Level
- Publically available
- Private-expanding
- Research team only

Risks

De-identification & preventing re-identification
### Privacy risks of kinds of data & availability

<table>
<thead>
<tr>
<th>Access Level</th>
<th>Risks</th>
<th>De-identification &amp; preventing re-identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publicly available</td>
<td>Easy</td>
<td>Medium</td>
</tr>
<tr>
<td>Private-expanding</td>
<td>Hard</td>
<td></td>
</tr>
<tr>
<td>Research team only</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Examples:
- Click-stream
  - No demographic
  - No linking
- Disc board
  - Blur names
- Audio video
  - No blur
### Privacy risks of kinds of data & availability

<table>
<thead>
<tr>
<th>Access Level</th>
<th>Risks</th>
<th>De-identification &amp; preventing re-identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publicly available</td>
<td>Little</td>
<td>Lots</td>
</tr>
<tr>
<td>Private-expanding</td>
<td>Little</td>
<td>Some</td>
</tr>
<tr>
<td>Research team only</td>
<td>Little</td>
<td>Little</td>
</tr>
</tbody>
</table>

**Examples:**
- Click-stream: No demographic, No linking
- Disc board: Blur names
- Audio video: No blur
Welcome to DataShop, the world's largest repository of learning interaction data.

Create an account or Log in to start analyzing data.

What can I do with DataShop?

Upload a dataset

- Project
  - Add this dataset to...
    - new project
    - existing project
    - choose later

- Project Name
  - Psychology MOOC data

- Data Collection Type
  - Not specified
  - Not human subjects data (not originally collected for research purposes)
  - Study data collected under an IRB where consent was not required (IRB approval letter required)
  - Study data collected under an IRB where consent was required (IRB approval letter and consent form required)

- Dataset Name
  - 2013 Psych

- Description (optional)

1600+ data sets
math, science, language ...
K12 & college
Projects Have 3 Possible States

- **Private**: Only PI/data owner may access.
  - Default State: Private and Not-Shareable.

- **Public**: Any registered DataShop user may access project freely.

- **Not-Shareable**: Only PI/data owner may access.
  - Data Owner determines Private/Public.

**Shareable**
- PI/data owner decides on case by case basis whether to share with non-team members.
- Research Manager determines Shareability.
Requirements for “Shareable” Designation

1. **Data collection types**
   - **Not human subjects** (data not originally collected for research purposes, e.g. course data)
     - Data is de-identified

2. **Study data** collected under an IRB where **consent not required**
   - Data is de-identified
   - IRB approval letter
Main IRB management page. Research manager can filter multiple ways. Most useful filter is Needs Attention.

Default filter settings: **Needs attention**—**YES**  
**Subject to DataShop 2012 IRB**—**YES**

“Needs Attention” triggered when:

- Researcher/data provider creates **new project**
- Researcher/data provider adds a **new dataset** to an “old” project **after** project had been designated “shareable”
- Researcher/data provider adds **new data** to an “old” dataset after its project had been designated “shareable”
1. The researcher (or research manager) specifies "Data Collection Type".

2. The researcher (or research manager) adds IRB information and uploads IRB documents.

3. After reviewing IRB docs, the research manager designates Shareability status.

Next slide—goes into Datasets tab.
In Dataset tab of this project—see items that might need RM attention

1. Appears anonymous: if not already de-identified by system, RM eyeballs datasets—if appears de-identified, indicate here “yes appears anonymous”. In this case, “N/A” indicates data already de-identified by system. In general, if RM finds any identifiable info, RM alerts DataShop staff for discussion with PI re de-identification process.

2. whether IRB has been uploaded for each dataset

3. whether each dataset has study data or not

4. If not already de-identified by system, RM accesses individual datasets to eyeball for de-identification
LearnSphere/DataShop Progress

- Web portal for data & method sharing
- Analytic workflow beta
  - Cross language (R, Python, Java, C, MatLab, ...) integration
  - Non-programmer recombination
  - Integration projects
    - DataShop + MOOCdb + DataStream
      - Apply doer effect workflow analytics to MOOCs
    - ... + DiscourseDB
      - Analyze discussion board posts & video, activities
- Distributed services
  (LearnSphere@your_institution)
- Privacy control procedures & software
Thank you!

Carnegie Mellon: Ken Koedinger, John Stamper, & Carolyn Rose
MIT: Una-May O’Reilly & Kalyan Veeramachaneni
Stanford: Candace Thille
U of Memphis: Phil Pavlik

Support from NSF Cyberinfrastructure, DIBBs, $5M for 5 years

Asilomar 2: June 16, 2016
Extra slides
DataShop Terminology

• **Shareability**--determined by research manager (RM):
  § **Not-Shareable** = DataShop does not give the data owner the option of sharing their project outside their research team
  § **Shareable** = DataShop gives the data owner the option of sharing their project with people outside their team (regardless of whether data owner keeps their project private or makes it public).

• **Private vs Public**--determined by data owner (contingent on RM shareability designation):
IRB Review Interface built by DataShop team and used by DataShop Research Manager
“Needs Attention” triggered when…

- Researcher/data provider creates new project

- Researcher/data provider adds a new dataset to an “old” project after project had been designated “shareable”
Dataset page of a project that has multiple datasets. This project does not need
Additional IRB management features

• one IRB can cover multiple projects: IRB entry page gives PI or RM option of \textit{applying existing IRB} to new project or \textit{adding a new IRB}.
  
  – “\texttt{All IRB}” page gives RM a listing of all IRBs and the projects associated with them.

• one project can have multiple IRBs related to it: A project’s IRB page shows all relevant IRB info w/links to documents
System automatically highlights any PUBLIC project that needs attention.
Figure 1. Variations in the types and time scales of educational data produce a challenge for data integration that, if met, will help advance understanding of the how the wide variety of psychological constructs underlying learning interact and can be best supported to produce effective learning. DataShop is an existing data infrastructure, which represents a portion of this space, MOOCdb is an emerging standard for MOOC data, and DiscourseDB represents a new area. Data infrastructure is needed to support analytic methods that integrate across this space so as to produce discoveries not possible within current data silos.
The PSLC builds on a rich history of cognitive science and education research in Pittsburgh, one manifestation of which has been the Cognitive Tutor technology that implements cognitive principles of learning in intelligent tutoring systems. The wide dissemination of our Algebra Cognitive Tutor inspired the idea that we could use such fielded technologies as a platform for doing basic research on learning with real students in real courses.

Two of our key goals are …
Visualizing learning curves to find opportunities for improvement

High rough curve
=> revise skill model
=> redesign instruction
=> do A/B test
Better student learning!