

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

LearnSphere.org ~3.5 of 5 years of NSF



LearnSphere

About Explore

A community data infrastructure
to support learning improvement
online.



Existing Resources

DataShop@CMU
a data analysis service for the learning science community

MOOCDB



DATASTAGE

 **DiscourseDB**

DataShop@Memphis
a data analysis service for the learning science community

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 science 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 sets from science, social, and health disciplines with an online tutoring system, in many cases as part of online experiments of what learning works best. You can also submit studies at www.assistmentstestbed.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 digital content sets remotely. The project is based at New York University and at Penn State. The U.S. National Science Foundation (NSF BC 03-1238599) and the U.S. National Institutes of Health (NIH U01-HD-076595) 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-998009, 0324883) 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).

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 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.

Simon DataLab

The [Simon DataLab](#) is an emerging intellectual data commons to drive continuous improvement in student learning outcomes with a particular focus on supporting instructors and course developers in using data to improve their courses.

EDM Workbench

The Educational Data Mining (EDM) Workbench is a tool that helps researchers and practitioners define and analyze learning patterns in educational data. It provides a user interface for data exploration, visualization, and analysis. The workbench is designed to be used by researchers and practitioners who are interested in understanding and improving learning outcomes. It provides a user interface for data exploration, visualization, and analysis. The workbench is designed to be used by researchers and practitioners who are interested in understanding and improving learning outcomes.

Ways to Generate Data

Online Content Providers



Educational Technology Development Tools



Online Assessment and Tutoring Systems



DataShop: Not just “big”, but fine, wide, &



LearnSphere

About Explore

LearnSphere.org

A community data infrastructure to support learning improvement online.



Existing Resources

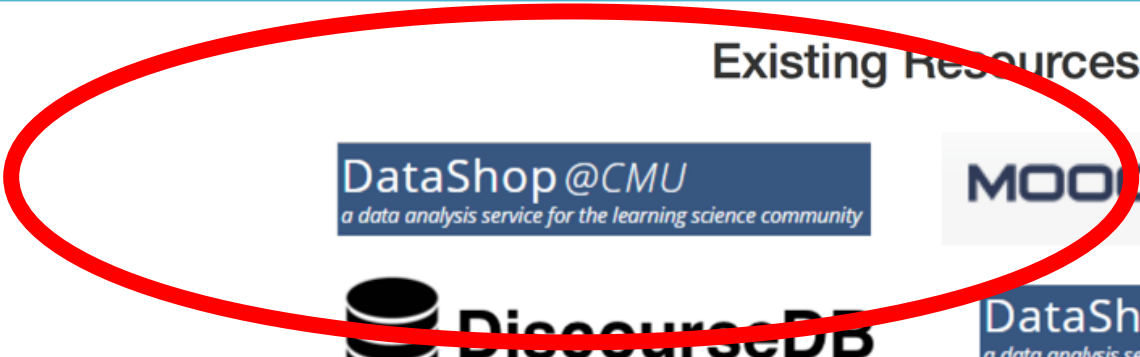
DataShop@CMU
a data analysis service for the learning science community

MOOCDB


DATASTAGE

 DiscourseDB

DataShop@Memphis
a data analysis service for the learning science community



Help

Explore

[Public Datasets](#)

[Private Datasets](#)

[External Tools](#)

[What can I do?](#)

Learn More

[Documentation](#)

[About DataShop](#)

[FAQ](#)

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 /student models](#)

[What is DataShop?](#)

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

K12 & college

[Help](#)**Explore**[Public Datasets](#)[Private Datasets](#)[External Tools](#)[What can I do?](#)**Learn More**[Documentation](#)[About DataShop](#)

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|

Recent dataset names

Description
(optional)

Recent descriptions

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

K12 & college

Help

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)

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 ...

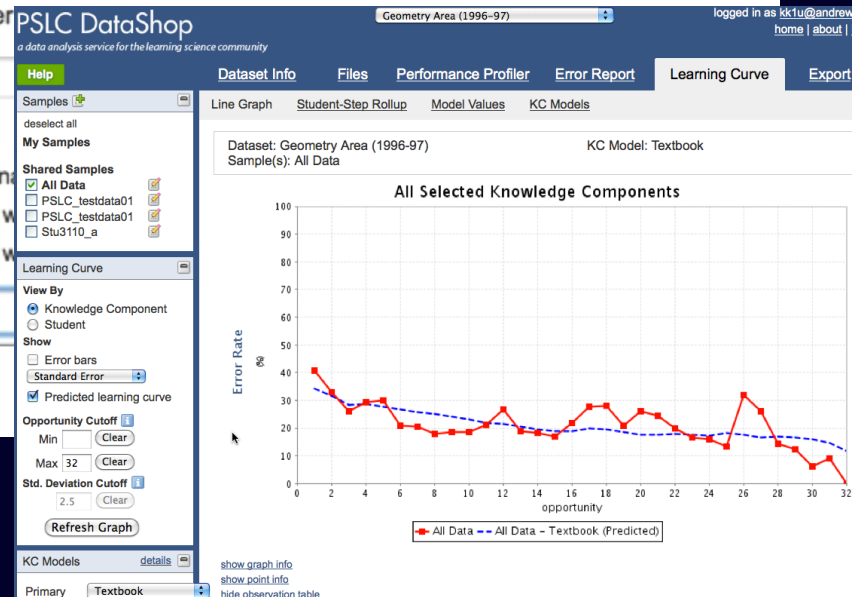
- new project existing project choose later

Project Name

- Data Collection Type
- Not specified
 - Not human subjects data (not original)
 - Study data collected under an IRB w/ consent
 - Study data collected under an IRB w/ no consent

Dataset Name

Description (optional)



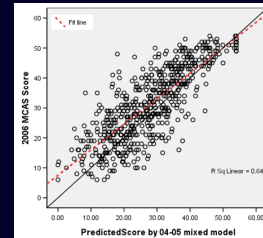
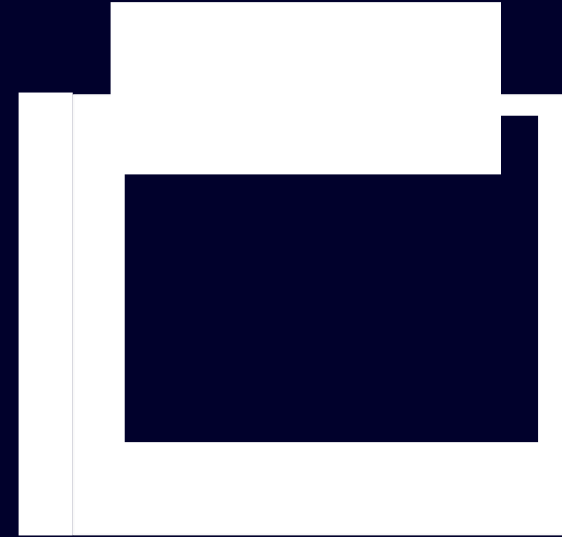
1600+ data sets
math, science, language ...

K12 & college

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



Doing

Low Extreme Trait	Factor	High Extreme Trait
Self-reliance	Autonomy	Interdependence
Individualism	Individualism	Collectivism
Privacy	Privacy	Openness
Control	Control	Flexibility
Order	Order	Spontaneity
Structure	Structure	Flexibility
Planning	Planning	Spontaneity
Logic	Logic	Emotion
Reasoning	Reasoning	Emotion
Analysis	Analysis	Intuition
Objectivity	Objectivity	Subjectivity
Neutrality	Neutrality	Emotionality
Impersonality	Impersonality	Warmth
Formality	Formality	Informality
Reserve	Reserve	Openness
Self-control	Self-control	Spontaneity
Self-discipline	Self-discipline	Spontaneity
Self-reliance	Self-reliance	Interdependence
Self-sufficiency	Self-sufficiency	Interdependence
Self-dependence	Self-dependence	Interdependence
Self-reliance	Self-reliance	Interdependence

6x better than

Watching

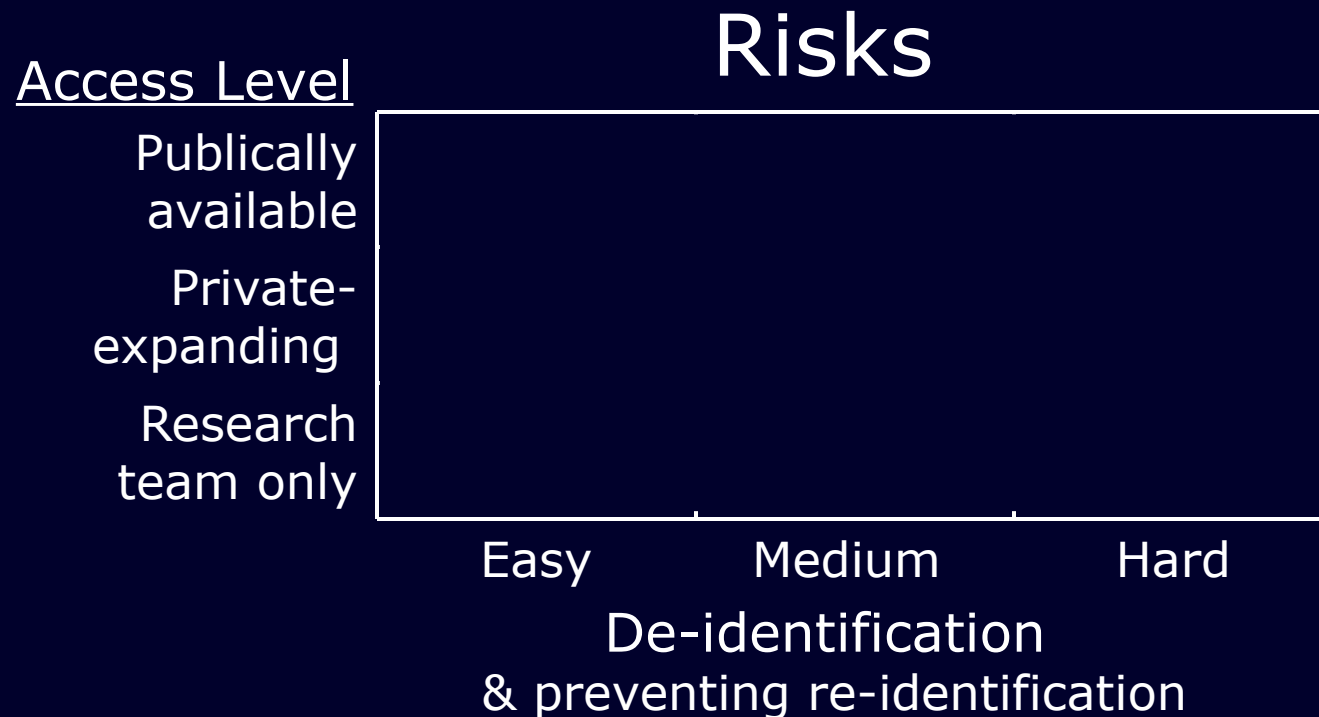
Psychology

- Brain & Behavior
- Sensation & Perception
- Learning & Memory
- Cognition & Language

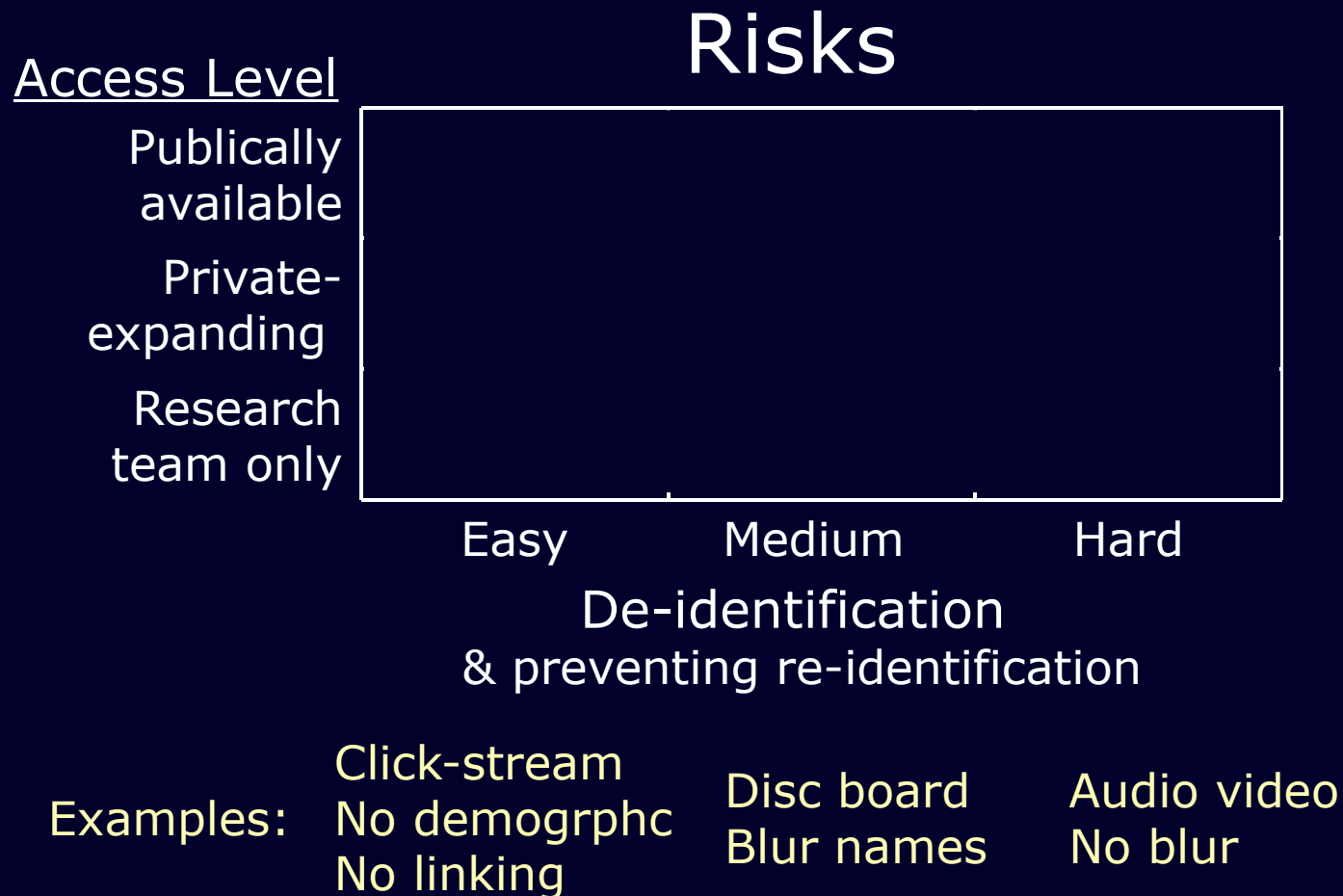
LearnSphere Key Points

- 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



Privacy risks of kinds of data & availability



Privacy risks of kinds of data & availability

<u>Access Level</u>	Risks		
	Easy	Medium	Hard
Publically available	Little	Lots	Too much
Private-expanding	Little	Some	Lots
Research team only	Little	Little	Little

De-identification
& preventing re-identification

Examples:	Click-stream	Disc board	Audio video
	No demogrphc No linking	Blur names	No blur

[Help](#)**Explore**[Public Datasets](#)[Private Datasets](#)[External Tools](#)[What can I do?](#)**Learn More**[Documentation](#)[About DataShop](#)

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|

Recent dataset names

Description
(optional)

Recent descriptions

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

K12 & college

Projects Have 3 Possible States

Research Manager determines Shareability

Data Owner determines Private/Public

Private

Public

Not-Shareable

Only PI/data owner may access

Default State: Private and Not-Shareable

Shareable

PI/data owner decides on case by case basis whether to share with non-team members

Any registered DataShop user may access project freely

Requirements for “Shareable” Designation

- Data collection
 1. 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*

Help

Please select a dataset...

My Data

[My Datasets](#)

[Upload a dataset](#)

[Create a new dataset](#)

[Access Requests](#)

[My Profile](#)

Explore

[Public Datasets](#)

[Private Datasets](#)

[External Tools](#)

[What can I do?](#)

Learn More

[Documentation](#)

[About DataShop](#)

[FAQ](#)

Advanced

[Metrics Report](#)

[Web Services](#)

[Logging Activity](#)

[Manage Terms](#)

[Edit Research Goals](#)

Admin

[Manage Users](#)

[Set Domain/LearnLab](#)

[IRB Review](#)

[All IRBs](#)

[Import Queue](#)

IRB Review

project filters

Filters

Public/Private

Project Name

PI/Data Provider

Shareability Review Status

Data Collection Type

Subject to DataShop

Project Created

Dataset Last Added

Needs Attention

useful default filter: Needs Attention

11 projects found.

<u>Project Name</u>	<u>Subject To DataShop 2012 IRB</u>	<u>Shareability Review Status</u>	<u>Data Collection Type</u>	<u>Unreviewed Datasets</u>	<u>Project Created</u>	<u>Dataset Last Added</u>	<u>Needs Attention</u>
<u>DALMOOC</u> Ryan Baker (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 1	2014-11-17	2014-11-17	Yes
<u>DyscalculiaData</u> Tanja Kaeser (pi) show datasets	Yes	Waiting for researcher	Study, consent req'd	0 of 1	2013-09-12	2013-09-20	Yes
<u>ENGAGE Beanstalk Game Study</u> Vincent Alevan (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 6	2013-07-23	2014-12-16	Yes
<u>Imbrogno - Cross Cultural Hint Usage</u>	Yes	Waiting for researcher	Not specified	0 of 1	2013-10-12	2013-10-12	Yes

DALMOOC project's IRB page

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

My Data

DALMOOC

[Datasets](#) [Permissions](#) [IRB](#) [Terms of Use](#)

Data Collection Type Study data collected under an IRB where consent was not required (IRB approval letter required) [edit](#)

Subject to 2012 DataShop IRB Yes, the data was added to DataShop after April 2012 [edit](#)

Shareability Review Status Waiting for researcher [edit](#)

Needs Attention Yes [edit](#)

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

IRB Documents [Add an IRB \(step 1\)](#) You can upload files (step 2) after adding an IRB

No IRBs uploaded.

Shareability Review History

2015-01-09 [Gail Kusbit](#) Waiting for researcher

Research Manager's Notes [edit](#)

1/9/15 asked Ryan for info. 1/12/15 Ryan said is study data, no consent. Suggest approval.

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

- [My Datasets](#)
- [Upload a dataset](#)
- [Create a project](#)
- [Access Requests](#)
- [My Profile](#)

- ### Explore
- [Public Datasets](#)
 - [Private Datasets](#)
 - [External Tools](#)
 - [What can I do?](#)

- ### Learn More
- [Documentation](#)
 - [About DataShop](#)
 - [FAQ](#)

- ### Advanced
- [Metrics Report](#)
 - [Web Services](#)
 - [Logging Activity](#)
 - [Manage Terms](#)
 - [Edit Research Goals](#)

- ### Admin
- [Manage Users](#)

My Data

- [My Datasets](#)
- [Upload a dataset](#)
- [Create a project](#)
- [Access Requests](#)
- [My Profile](#)

DALMOOC

- Datasets**
- [Permissions](#)
- [IRB](#)
- [Terms of Use](#)

DALMOOC project's Dataset page

[Request Access](#)

PI [Ryan Baker](#) [edit](#)

Data Provider [edit](#)

Description [edit](#)

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)
- [FAQ](#)

Advanced

- [Metrics Report](#)
- [Web Services](#)
- [Logging Activity](#)
- [Manage Terms](#)
- [Edit Research Goals](#)

Admin

- [Manage Users](#)
- [Set Domain/LearnLab](#)
- [IRB Review](#)
- [All IRBs](#)
- [Import Queue](#)
- [Manage Problem](#)
- [Content](#)

- Project Access
- [Rename](#)
- [Upload a](#)
- [Delete](#)

research manager accesses dataset settings

Datasets [edit](#)

Appears anonymous?	IRB Uploaded	Has Study Data	Dataset	Domain/LearnLab	Dates	Data Last Modified	Sta
N/A	TBD	Not Specified	DALMOOC		Nov 15, 2014 - Jan 9, 2015	Jan 10, 2015	

Appears Anonymous?

- N/A - Student user IDs were de-identified
- Yes - Data appears to be anonymous
- No - Data reveals student identities
- Not reviewed - Have not reviewed data for anonymity
- More info needed - Unclear whether data is anonymous

IRB Uploaded

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”

Adapterrex: Exploring the Learning Benefits of Erroneous Examples




[Datasets](#)
[Permissions](#)
[IRB](#)
[Terms of Use](#)
[Request Access](#)

Project Actions:

[Rename](#)
[Upload a dataset](#)
PI Bruce McLaren [edit](#)
Data Provider [edit](#)
Description [edit](#)
Tags [edit](#)
External Links [add](#)

Dataset page of a project that has multiple datasets. This project does not need attention

Datasets [edit](#)

Appears anonymous?	IRB Uploaded	Has Study Data	Dataset	Domain/ LearnLab	Dates	Data Last Modified	Status	Transactions	
N/A	Yes	Yes	AdaptErrEx	Math/ Other	Jul 21, 2010 - Mar 22, 2012	Mar 26, 2012	complete	537,302	
N/A	Yes	Yes	adapterrex2	Math/ Other	Oct 12, 2011 - Mar 30, 2012	Apr 3, 2012	complete	308,190	
N/A	Yes	Yes	adapterrex3	Math/ Other	Mar 26, 2012 - May 17, 2012	May 25, 2012	complete	369,106	

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

My Data

- [My Datasets](#)
- [Upload a dataset](#)
- [Create a project](#)
- [Access Requests](#)
- [My Profile](#)

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)
- [FAQ](#)

Advanced

- [Metrics Report](#)
- [Web Services](#)
- [Logging Activity](#)
- [Manage Terms](#)
- [Edit Research Goals](#)

Admin

- [Manage Users](#)
- [Set Domain/LearnLab](#)
- [IRB Review](#)
- [All IRBs](#)
- [Import Queue](#)
- [Manage Problem Content](#)

IRB Review

Filters

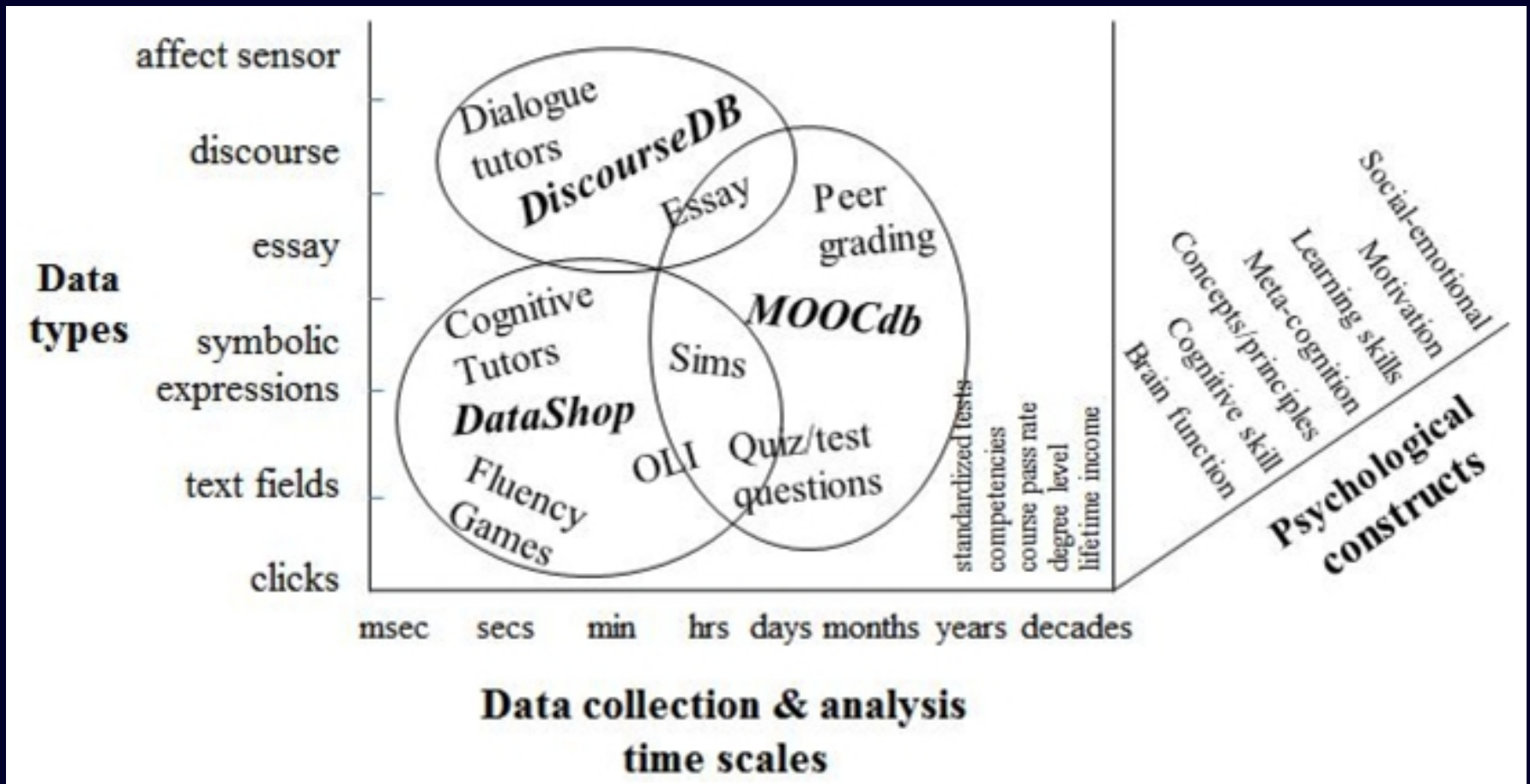
Public/Private	Project Name	PI/Data Provider
<input type="text"/>	<input type="text"/>	<input type="text"/>
	Search by project name	Search by PI or Data Provider
Shareability Review Status	Data Collection Type	Subject to DataShop
<input type="text"/>	<input type="text"/>	Yes <input type="text"/>
Project Created	Dataset Last Added	Needs Attention
Before <input type="text"/>	Before <input type="text"/>	Yes <input type="text"/>

System highlights any PUBLIC projects that need attention in yellow

12 projects found.

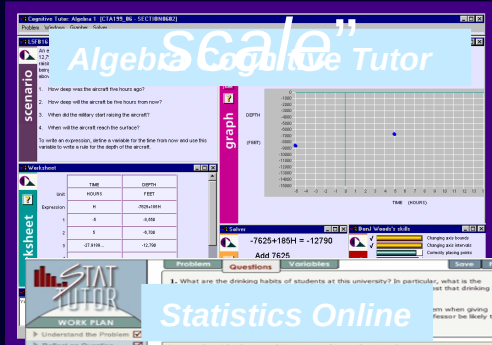
<u>Project Name</u>	<u>Subject To DataShop 2012 IRB</u>	<u>Shareability Review Status</u>	<u>Data Collection Type</u>	<u>Unreviewed Datasets</u>	<u>Project Created</u>	<u>Data Added</u>	<u>Needs Attention</u>
Digital Games for Improving Number Sense Derek Lomas (pi) show datasets	Yes	Shareable	Study, consent not req'd	0 of 1	2011-04-06	2011-04-06	Yes
DALMOOC Ryan Baker (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 1	2014-11-17	2014-11-17	Yes
DyscalculiaData Tanja Kaesser (pi) show datasets	Yes	Waiting for researcher	Study, consent req'd	0 of 1	2013-09-12	2013-09-20	Yes
ENGAGE Beanstalk Game Study Vincent Alevan (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 6	2013-07-23	2014-12-16	Yes
Imbrogno - Cross Cultural Hint Usage Jason Imbrogno (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 1	2013-10-12	2013-10-12	Yes
MathTutor Vincent Alevan (pi) show datasets	Yes	Waiting for researcher	Not specified	4 of 10	2010-01-19	2013-12-04	Yes

Vast space of data & questions =>
 need a *data infrastructure* to integrate =>
 produce *discoveries not possible within current data silos*

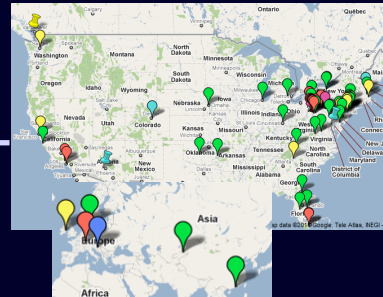


\$47M from NSF => 850+ datasets

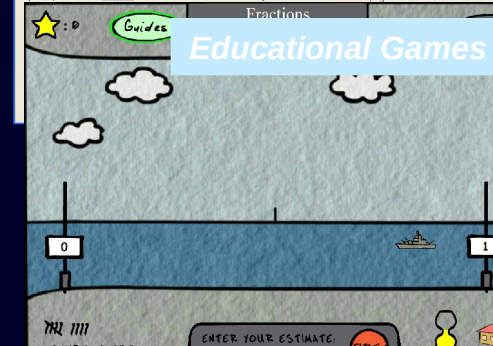
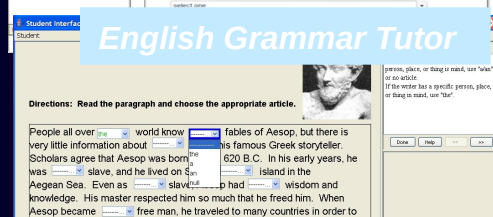
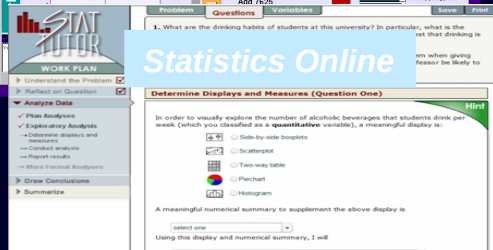
Ed tech + wide use = “Basic research at



+

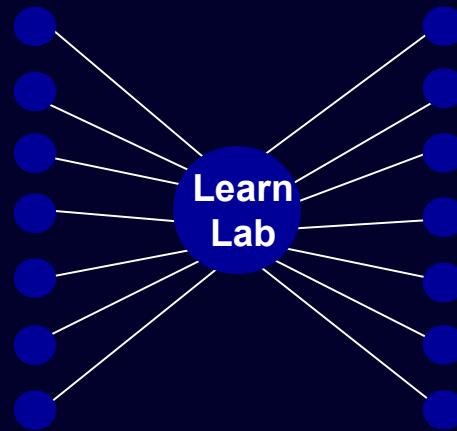


=



Researchers

Schools

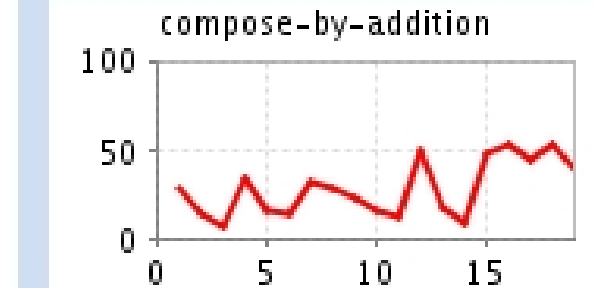
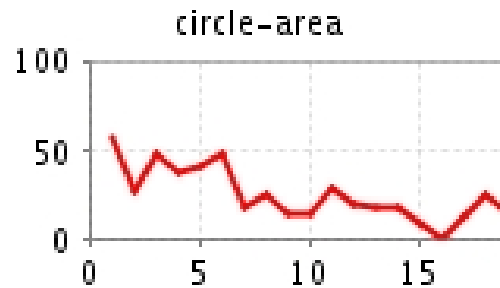
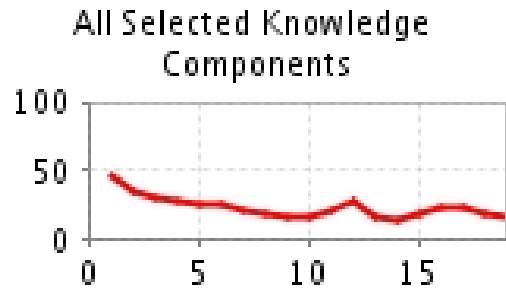
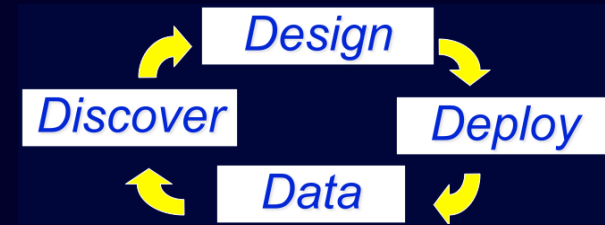


Since 2004

- ∅ 1600 ed tech data sets in DataShop
- ∅ 360 *in vivo* experiments

Koedinger et al. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.

Visualizing learning curves to find opportunities for improvement



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



Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education*

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

LearnSphere.org ~3.5 of 5 years of NSF



LearnSphere

About Explore

A community data infrastructure
to support learning improvement
online.



Existing Resources

DataShop@CMU
a data analysis service for the learning science community

MOOCDB



DATASTAGE



DiscourseDB

DataShop@Memphis

a data analysis service for the learning science community

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*

3

Others: DataBrary,
Upenn
Brian MacWhinney

Data repositories

DataShop

The LearnLab DataShop is a data repository and web application for learning science 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-0836312, 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 learning system, in many cases as part of online experiments of what learning works best. You can also submit studies. It www.assistments.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 animals. The project focuses on collecting data to store, manage, preserve, analyze, and share video and other media. The project is supported by grants from New York University and at Penn State. The U.S. National Science Foundation (N01-BC-9-1238569) and the U.S. National Institutes of Health (NIH U01-HD-070595) 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-998009, 0324883) 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).

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 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 to the educational data mining and intelligent tutoring systems communities.

ChatLab

The Simon DataLab is an emerging intellectual commons to drive continuous improvement in student learning outcomes with a particular focus on supporting instructors and course developers in using data to improve their courses.

EDM Workbench

The Educational Data Mining (EDM) Workbench provides a platform for data analysis, visualization, and reporting. It offers a variety of tools and services to support research and development in educational data mining.

Ways to Generate Data

Online Content Providers

Carnegie Learning



edX



Educational Technology Development Tools



Online Assessment and Tutoring Systems



Second Language Tutors

DataShop: Not just “big”, but fine, wide, &



LearnSphere

About Explore

LearnSphere.or

A community data infrastructure
to support learning improvement
online.



Existing Resources

DataShop@CMU
a data analysis service for the learning science community

MOOCDB



DATASTAGE



DiscourseDB

DataShop@Memphis
a data analysis service for the learning science community

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)
- [FAQ](#)

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 /student models

[What is DataShop?](#)

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

K12 & college

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)

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|

Recent dataset names

Description
(optional)

Recent descriptions

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

K12 & college

Help

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)

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

Data Collection Type

Not specified

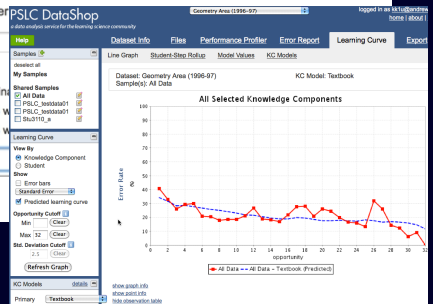
Not human subjects data (not original)

Study data collected under an IRB waiver

Study data collected under an IRB approval

Dataset Name

Description (optional)



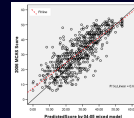
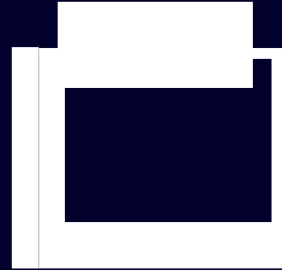
1600+ data sets
math, science, language ...

K12 & college

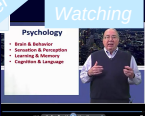
Data to improve learning

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

Sharing leverages interdisciplinary interaction



6x better than



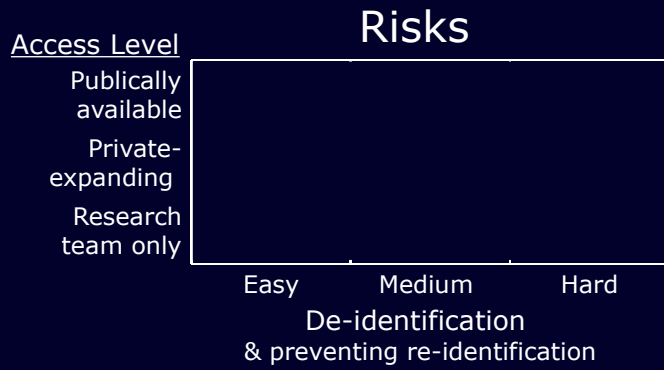
LearnSphere Key Points

- 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*

10

Others: DataBrary,
Upenn
Brian MacWhinney

Privacy risks of kinds of data & availability



Privacy risks of kinds of data & availability

		Risks		
Access Level		Easy	Medium	Hard
Publically available				
Private-expanding				
Research team only				
		De-identification & preventing re-identification		
Examples:	Click-stream No demogrphc No linking	Disc board Blur names	Audio video No blur	

Privacy risks of kinds of data & availability

Access Level	Risks		
	Easy	Medium	Hard
Publically available	Little	Lots	Too much
Private-expanding	Little	Some	Lots
Research team only	Little	Little	Little

De-identification
& preventing re-identification

Examples:	Click-stream No demogrphc No linking	Disc board Blur names	Audio video No blur
-----------	--	--------------------------	------------------------

Explore

- [Public Datasets](#)
- [Private Datasets](#)
- [External Tools](#)
- [What can I do?](#)

Learn More

- [Documentation](#)
- [About DataShop](#)

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

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

Recent dataset names

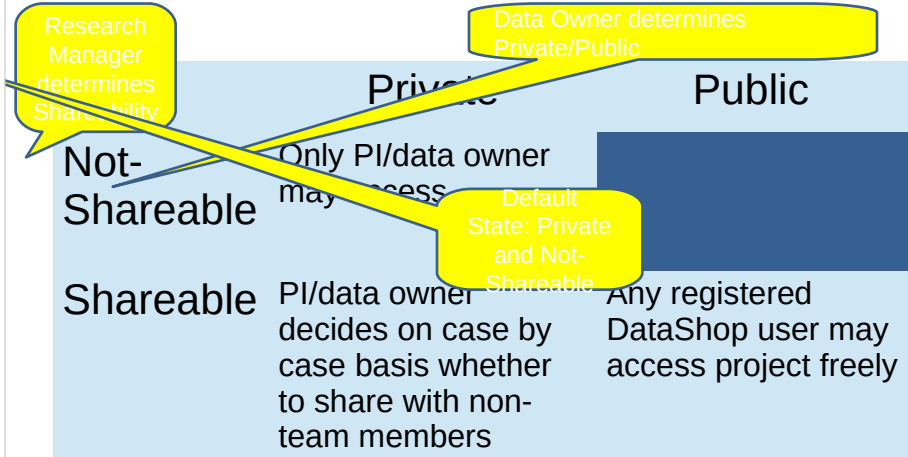
Description
(optional)

Recent descriptions

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

K12 & college

Projects Have 3 Possible States



Requirements for “Shareable” Designation

- Data collection

1. 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

PSLC DataShop
a data analysis service for the learning science community

logged in as [gshub@andrew.cmu.edu](#) | [home](#) | [about](#) | [help](#)

IRB Review

Filters

Public/Private: Public Private

Project Name: Search by project name

PIData Provider: Search by PI or Data Provider

Shareability Review Status: Shareable Not Shareable

Data Collection Type: Data Collection Not Data Collection

Subject to DataShop: Yes No

Project Created: Before

Dataset Last Added: Before

Needs Attention: Yes No

11 projects found.

Project Name	Subject To DataShop 2012 IRB	Shareability Review Status	Data Collection Type	Unreviewed Datasets	Project Created	Dataset Last Added	Needs Attention
DALMOOC Ryan Baker (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 1	2014-11-17	2014-11-17	Yes
DyscalculiaData Tania Kaiser (pi) show datasets	Yes	Waiting for researcher	Study, consent req'd	0 of 1	2013-09-12	2013-09-20	Yes
ENGAGE Beanstalk Game Study Vincent Alaven (pi) show datasets	Yes	Waiting for researcher	Not specified	0 of 6	2013-07-23	2014-12-16	Yes
Imbrogno - Cross Cultural Hint Usage	Yes	Waiting for researcher	Not specified	0 of 1	2013-10-12	2013-10-12	Yes

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”

PSLC DataShop
a data analysis service for the learning science community

logged in as gkusbit@andrew.cmu.edu

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

Datasets Permissions IRB Terms of Use

Data Collection Type Study data collected under an IRB where consent was not required (IRB approval letter required) [edit](#)

Subject to 2012 DataShop IRB Yes, the data was added to DataShop after April 2012 [edit](#)

Shareability Review Status Waiting for researcher [edit](#)

Needs Attention Yes [edit](#)

IRB Documents [Add an IRB \(step 1\)](#) You can upload files (step 2) after adding an IRB document.

No IRBs uploaded.

Shareability Review History

2015-01-09 [Gail Kusbit](#) Waiting for researcher

Research Manager's Notes [edit](#)

1/9/15 asked Ryan for info. 1/12/15 Ryan said is study data, no consent. Suggest approval.

My Data
My Datasets
Upload a dataset
Create a project
Access Requests
My Profile

Explore
Public Datasets
Private Datasets
External Tools
What can I do?

Learn More
Documentation
About DataShop
FAQ

Advanced
Metrics Report
Web Services
Logging Activity
Manage Terms
Edit Research Goals

Admin
Manage Users

https://pslcdatashop.web.cmu.edu/AccountProfile

Next slide—goes into Datasets tab

My Data

[My Datasets](#)
[Upload a dataset](#)
[Create a project](#)
[Access Requests](#)
[My Profile](#)

Explore

[Public Datasets](#)
[Private Datasets](#)
[External Tools](#)
[What can I do?](#)

Learn More

[Documentation](#)
[About DataShop](#)
[FAQ](#)

Advanced

[Metrics Report](#)
[Web Services](#)
[Logging Activity](#)
[Manage Terms](#)
[Edit Research Goals](#)

Admin

[Manage Users](#)
[Set Domain/LearnLab](#)
[IRB Review](#)
[All IRBs](#)
[Import Queue](#)
[Manage Problem](#)
[Content](#)

DALMOOC

Datasets Permissions IRB Terms of Use

Request Access

PI Ryan Baker [edit](#)

Data Provider [edit](#)

Description [edit](#)

research manager accesses dataset settings

Project Action [Rename](#) [Upload a](#) [Delete](#)

Datasets [edit](#)

Appears anonymous?	IRB Uploaded	Has Study Data	Dataset	Domain/LearnLab	Dates	Data Last Modified	Sta
N/A	TBD	Not Specified	DALMOOC		Nov 15, 2014 - Jan 9, 2015	Jan 10, 2015	

access dataset to confirm de-identification (if needed)

Appears Anonymous?

N/A - Student user IDs were de-identified
 Yes - Data appears to be anonymous
 No - Data reveals student identities
 Not reviewed - Have not reviewed data for anonymity
 More info needed - Unclear whether data is anonymous

IRB Uploaded

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

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”

DataShop > Project <https://pslcdatashop.web.cmu.edu/Project?id=67>

Adapterrex: Exploring the Learning Benefits of Erroneous Examples

[Datasets](#) [Permissions](#) [IRB](#) [Terms of Use](#)

Request Access

PI **Bruce McLaren** [edit](#)

Data Provider [edit](#)

Description [edit](#)

Tags [edit](#)

External Links [add](#)

Project Actions: [Rename](#) [Add a dataset](#)

Dataset page of a project that has multiple datasets. This project does not need attention.

Datasets [edit](#)

Appears anonymous?	IRB Uploaded	Has Study Data	Dataset	Domain/ LearnLab	Dates	Data Last Modified	Status	Transactions	
N/A	Yes	Yes	AdaptErrEx	Math/ Other	Jul 21, 2010 - Mar 22, 2012	Mar 26, 2012	complete	537,302	
N/A	Yes	Yes	adapterrex2	Math/ Other	Oct 12, 2011 - Mar 30, 2012	Apr 3, 2012	complete	308,190	
N/A	Yes	Yes	adapterrex3	Math/ Other	Mar 26, 2012 - May 17, 2012	May 25, 2012	complete	369,106	

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

PSLC DataShop > IRB Review

https://pslcdatashop.web.cmu.edu/IRBReview

PSLC DataShop
a data analysis service for the learning science community

logged in as gbrabbit@andrew.cmu.edu

IRB Review

Filters

Public/Private: [Dropdown] Project Name: [Search] PI/Data Provider: [Search]

Shareability Review Status: [Dropdown] Data Collection Type: [Dropdown] Subject to DataShop: [Dropdown]

Project Created: [Dropdown] Dataset Last Added: [Dropdown] Needs Attention: [Dropdown]

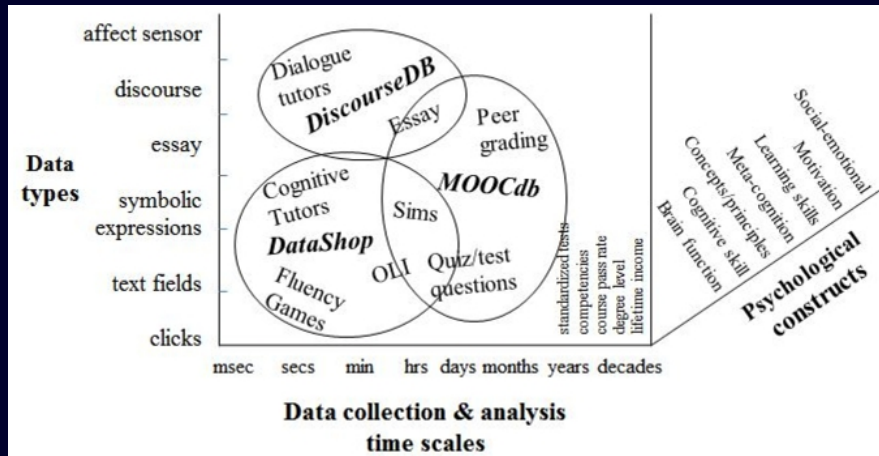
12 projects found.

Project Name	Subject To DataShop 2012 IRB	Shareability Review Status	Data Collection Type	Unreviewed Datasets	Project Created	Date Added	Needs Attention
Digital Games for Improving Number Sense <small>Derek Lomas (PI) show dataset</small>	Yes	Shareable	Study, consent not req'd	0 of 1	2011-04-06	2011-04-06	Yes
DALMOOC <small>Ryan Baer (PI) show dataset</small>	Yes	Waiting for researcher	Not specified	0 of 1	2014-11-17	2014-11-17	
Dyscalculia Data <small>Tania Kaiser (PI) show dataset</small>	Yes	Waiting for researcher	Study, consent req'd	0 of 1	2013-09-12	2013-09-20	Yes
ENGAGE Beanstalk Game Study <small>Vincent Aween (PI) show dataset</small>	Yes	Waiting for researcher	Not specified	0 of 6	2013-07-23	2014-12-16	Yes
Imbroqno - Cross Cultural Hint Usage <small>Jason Imbroqno (PI) show dataset</small>	Yes	Waiting for researcher	Not specified	0 of 1	2013-10-12	2013-10-12	Yes
Math Tutor <small>Vincent Aween (PI) show dataset</small>	Yes	Waiting for researcher	Not specified	4 of 10	2010-01-19	2013-12-04	Yes

System highlights any PUBLIC projects that need attention in yellow

System automatically highlights any PUBLIC project that needs attention.

Vast space of data & questions =>
 need a *data infrastructure* to integrate =>
 produce discoveries not possible within
 current data silos

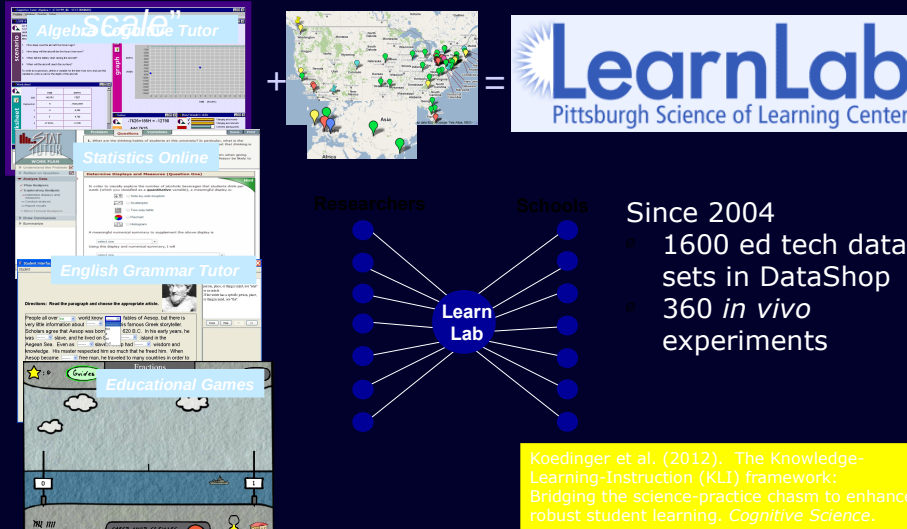


29

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.

\$47M from NSF => 850+ datasets

Ed tech + wide use = "Basic research at



Since 2004

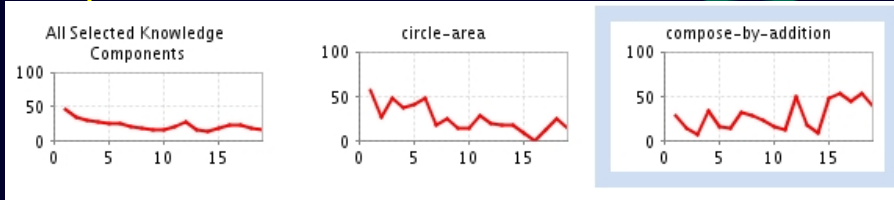
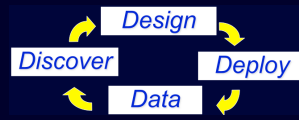
- 1600 ed tech data sets in DataShop
- 360 *in vivo* experiments

Koedinger et al. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.

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!



Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education*.