

Continuous Improvement of Educational Technology through Discoveries with Big Data

John Stamper
Human-Computer Interaction Institute
Carnegie Mellon University

**Carnegie
Mellon
University**



AI, ML, and Big Data

- Educational systems are now generating data at scale (Big Data)
- We can harness Machine Learning and Data Mining to improve these systems.

DataShop

- Central Repository
 - Secure place to store & access research data
 - Supports various kinds of research
 - Primary analysis of study data
 - Exploratory analysis of course data
 - Secondary analysis of any data set
- Analysis & Reporting Tools
 - Focus on student-tutor interaction data
 - Data Export
 - Tab delimited tables you can open with your favorite spreadsheet program or statistical package
 - Web services for direct access



How big is DataShop?

Domain	Files	Papers	Datasets	Student Actions	Students	Student Hours
Language	72	14	170	18,222,928	19,117	33,728
Math	300	91	613	148,608,998	216,258	415,364
Science	192	19	297	30,887,757	69,704	93,458
Other Subjects	122	32	300	42,827,103	79,655	162,680
Unspecified	172	4	683	60,996,194	81,845	176,358
Total	858	160	2,063	301,542,980	466,579	881,590

As of July 2019



What kinds of data?

- By domain based on studies from the Learn Labs
- Data from intelligent tutors
- Data from online instruction
- Data from games

The data is fine grained at a transaction level!



DataShop Terminology

- **KC:** Knowledge component
 - also known as a skill/concept/fact
 - a piece of information that can be used to accomplish tasks
 - tagged at the step level
- **KC Model:**
 - a computational cognitive model or skill model
 - a mapping between correct steps and knowledge components

Getting the KC Model Right!

The KC model drives instruction in adaptive learning

- Problem and topic sequence
- Instructional messages
- Tracking student knowledge



What makes a good KC Model?

- A correct expert model is one that is consistent with student behavior
- Predicts task difficulty
- Predicts transfer between instruction and test

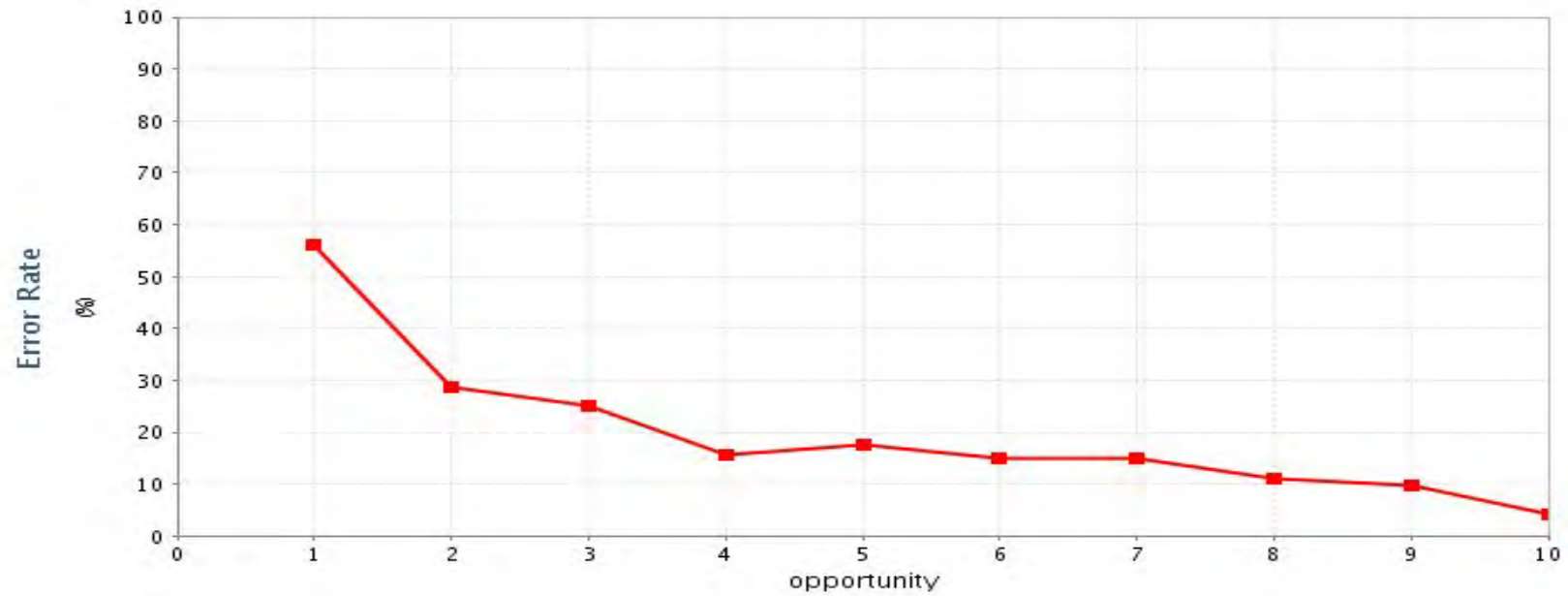
The model should fit the data!

Good KC Model = Good Learning Curve

- An empirical basis for determining when a KC model is good
- Accurate predictions of student task performance & learning transfer
 - Repeated practice on tasks involving the same skill should reduce the error rate on those tasks

A declining error rate learning curve should emerge

A Good Learning Curve



How do we make the Models?



Traditionally Cognitive Task Analysis

But CTA interview methods have some issues...

- Extremely human driven
- Highly subjective
- Leads to differing results from different analysts

And these human discovered models are often wrong!



If human centered intuitive design is not the answer...

How should student models be designed?

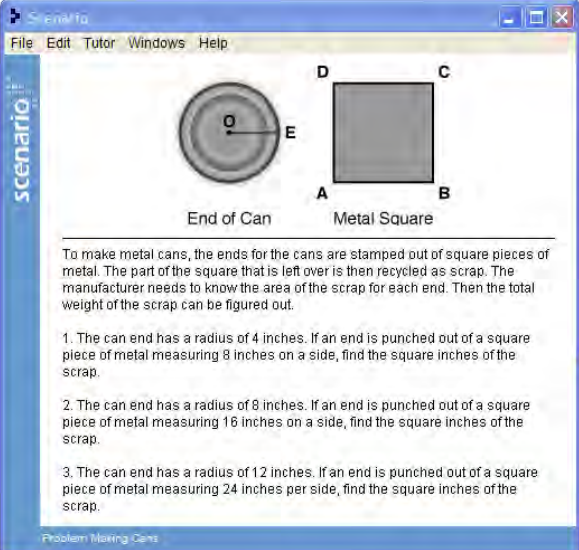
They shouldn't!

Student models should be discovered **not** designed!



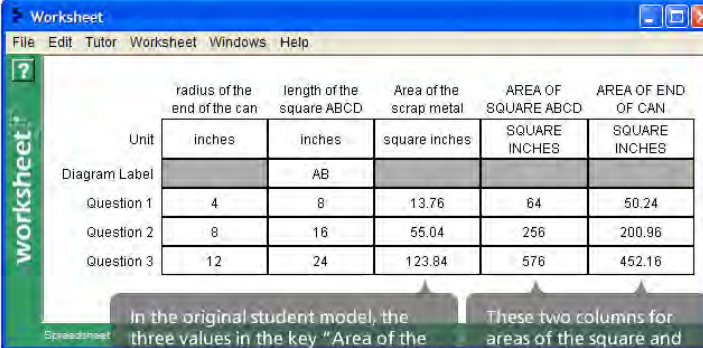
Solution – Use Data

Today we have lots of log data from edtech



The Scenario window displays a diagram with two parts: 'End of Can' and 'Metal Square'. The 'End of Can' is a circle with center 'O' and radius 'E'. The 'Metal Square' is a square with vertices 'A', 'B', 'C', and 'D'. Below the diagram, the text reads: 'To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The manufacturer needs to know the area of the scrap for each end. Then the total weight of the scrap can be figured out.'

1. The can end has a radius of 4 inches. If an end is punched out of a square piece of metal measuring 8 inches on a side, find the square inches of the scrap.
2. The can end has a radius of 8 inches. If an end is punched out of a square piece of metal measuring 16 inches on a side, find the square inches of the scrap.
3. The can end has a radius of 12 inches. If an end is punched out of a square piece of metal measuring 24 inches per side, find the square inches of the scrap.



	radius of the end of the can	length of the square ABCD	Area of the scrap metal	AREA OF SQUARE ABCD	AREA OF END OF CAN
Unit	inches	inches	square inches	SQUARE INCHES	SQUARE INCHES
Diagram Label		AB			
Question 1	4	8	13.76	64	50.24
Question 2	8	16	55.04	256	200.96
Question 3	12	24	123.84	576	452.16

In the original student model, the three values in the key "Area of the scrap metal" column were associated with the compose-by-addition KC that codes for situations where students must find an area by "composing" other areas

These two columns for areas of the square and circle ("end of can") were not initially present (making this an un scaffolded problem)

We can harness this data to validate and improve existing student models

Human-Machine Student Model Discovery

(Stamper & Koedinger, 2011)

DataShop provides easy interface to add and modify student models and ranks models

The screenshot displays the DataShop web interface. At the top, there are navigation tabs: 'Dataset Info', 'Performance Profiler', 'Error Report', 'Learning Curve', and 'Export'. Below these are sub-tabs: 'Overview', 'Papers and Files', 'KC Models', and 'Problem Breakdown'. The main content area shows the dataset 'Geometry Area (1996-97)' and a list of KC Models. Each model entry includes its name, creator, creation date, mapping type, LFA values, BIC, and status. An 'Add or modify a KC model' dialog box is open on the right, showing a file selection step and a progress window for importing a new model.

Dataset: Geometry Area (1996-97)

KC Models

Model Name	KCs	Status
Textbook New created by A. Skogsholm on 2008-06-10 15:00:32.0 mapping type: step-to-kc LFA values — AIC: 5161.16 BIC: 5677.64 show model details	10 KCs	ready to use
Area logged with dataset mapping type: correct-transaction-to-kc LFA values — AIC: 5642.15 BIC: 6054.03 show model details	2 KCs	ready to use
Decompose created by Y. Cho on 2008-07-10 10:37:57.0 mapping type: step-to-kc LFA values — AIC: 5131.29 BIC: 5687 show model details	13 KCs	ready to use
DecomposeArith created by K. Koedinger on 2008-12-06 14:23:14.0 mapping type: step-to-kc LFA values — AIC: 5085.89 BIC: 5628.53 show model details	12 KCs	ready to use

Add or modify a KC model

[Export \(Step 1\)](#) [Import \(Step 2\)](#)

1. Select a KC model file to import

DataShop will first try to verify the validity of your KC model file. Information about this process will be displayed below. After verifying, you will be given the opportunity to continue importing or cancel the process.

```
* Import New model "Improved Textbook"
If the information above is correct, press
the "Import" button to continue. Otherwise,
press "Cancel" to stop the import process for
this file.
Beginning insertion process.
Processing file... 100%
Initial import process complete!
141 rows in the file were processed.
Running learning factors analysis (LFA) and
building sample information for new models.
(Note: this may take a very long time
depending on the size of the dataset!)
LFA in progress.
```

Human-Machine Student Model Discovery

3 strategies for discovering improvements to the student model

- Lack of smooth learning curves
- No apparent learning
- Problems with unexpected error rates



A good KC model produces a learning curve

Without decomposition, using just a single "Geometry" skill, no smooth learning curve.

But with 12 skills for geometry area, a smooth learning curve.

(Rise in error rate because poorer students get assigned more problems)

Is this the correct or "best" model?

PSLC DATASHOP
a data analysis service for the learning science community

Dataset: Geometry Area (1996-97)
Sample(s): All Data

All Selected Knowledge Components

Opportunity Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of Observations	59	59	59	58	56	48	48	47	45	42	42	41	41	41	41	40	39	39		
Opportunity Number	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

All Selected Knowledge Components

Opportunity Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Number of Observations	59	59	59	58	56	48	48	47	45	42	42	41	41	41	40	39	39						

circle-area
circle-circumfe...
circle-diameter
compose-by-addi...
compose-by-mult...
equi-tri-height?
parallelogram-area
pentagon-area
rectangle-area
square-area

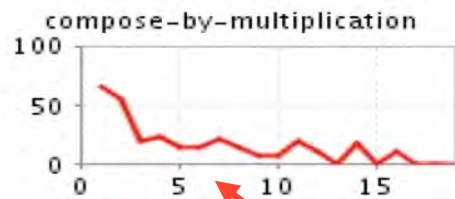
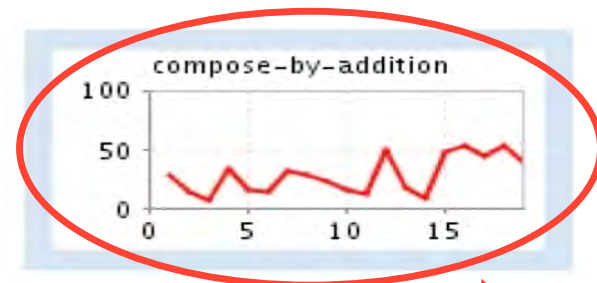
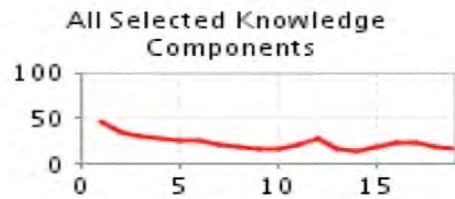
All Selected K.C.s

circle-area

rectangle-area

trapezoid-area

Inspect curves for individual knowledge components (KCs)



Many curves show a reasonable decline

Some do not => Opportunity to improve model!

No apparent Learning

KC Values For Textbook Model

KC Name	KC Category	Intercept (logit)	Intercept (probability)	Slope
Circle-area		0.47	0.61	0.02
Circle-circumference		0.1	0.53	0.11
Circle-diameter		1.07	0.74	0.02
Compose-by-addition		0.28	0.57	0
Compose-by-multiplication		0.77	0.68	0.01
Done		3.49	0.97	0.01
Geometric-Name		0.16	0.54	0.02
Given-unit-conversion		-1.78	0.14	0.01
Parallelogram		1.57	0.83	0
Pentagon		-0.28	0.43	0.03
Trapezoid		0.6	0.65	0.08
Triangle		0.08	0.52	0.03
Unit-name		0.88	0.71	0.04

These strategies suggest an improvement

- Hypothesized there were additional skills involved in some of the compose by addition problems
- A new student model (better AIC/BIC values) suggests the splitting the skill.

Decompose created by system on 2008-11-21 12:48:05.0 mapping type: correct transaction to kc LFA values – AIC: 14697.59 BIC: 15237.72 14875 observations labeled with KCs show model details	15 KCs status: ready to use	export
---	---------------------------------------	------------------------

Textbook created by system on 2008-11-21 12:48:05.0 mapping type: correct transaction to kc LFA values – AIC: 14865.38 BIC: 15375.07 14875 observations labeled with KCs show model details	13 KCs status: ready to use	export
--	---------------------------------------	------------------------

What does a better fit really mean?

The new model should be better at driving instruction to the students!

Redesign of the technology can be guided by the findings of the new model.

Redesign based on Discovered Model

Our discovery suggested changes needed to be made to the tutor

- Re-sequencing – put problems requiring fewer skills first
- Knowledge Tracing – adding new skills
- Creating new tasks – new problems
- Changing instructional messages, feedback or hints

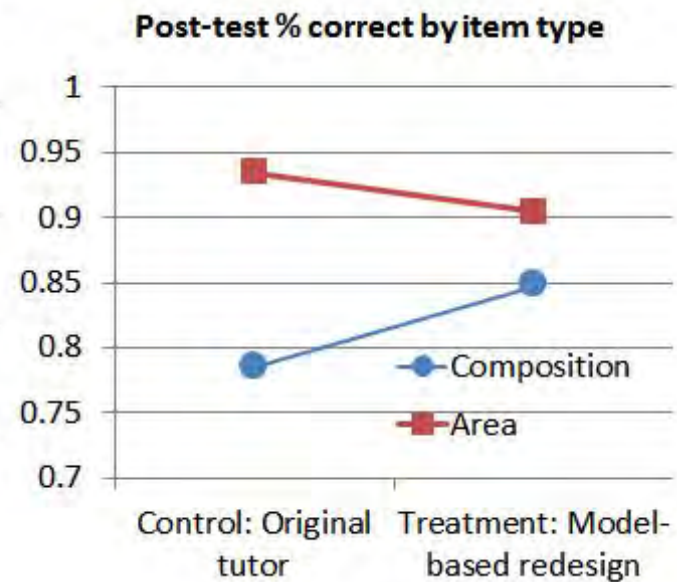
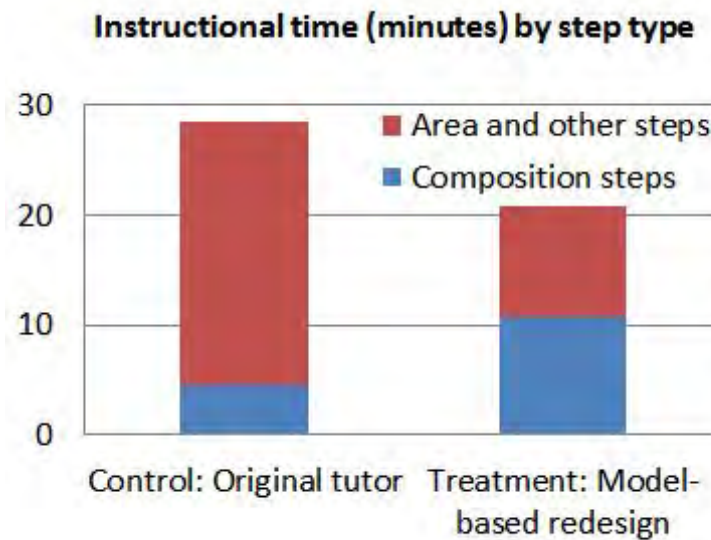
Closing the Loop, (Koedinger, Stamper, McLaughlin, 2013)

We implemented a new version of the Carnegie Learning Cognitive Tutor in Geometry

- Knowledge Tracing – added new skills for decomposing combined shapes
- Created new tasks – new problems isolating the new skills
- Changing instructional messages, feedback or hints



Results



- Significantly less time to mastery (25% less time) though more time on critical decomposition skills
- Better posttest performance on composition skills indicating better learning of decomposition skills

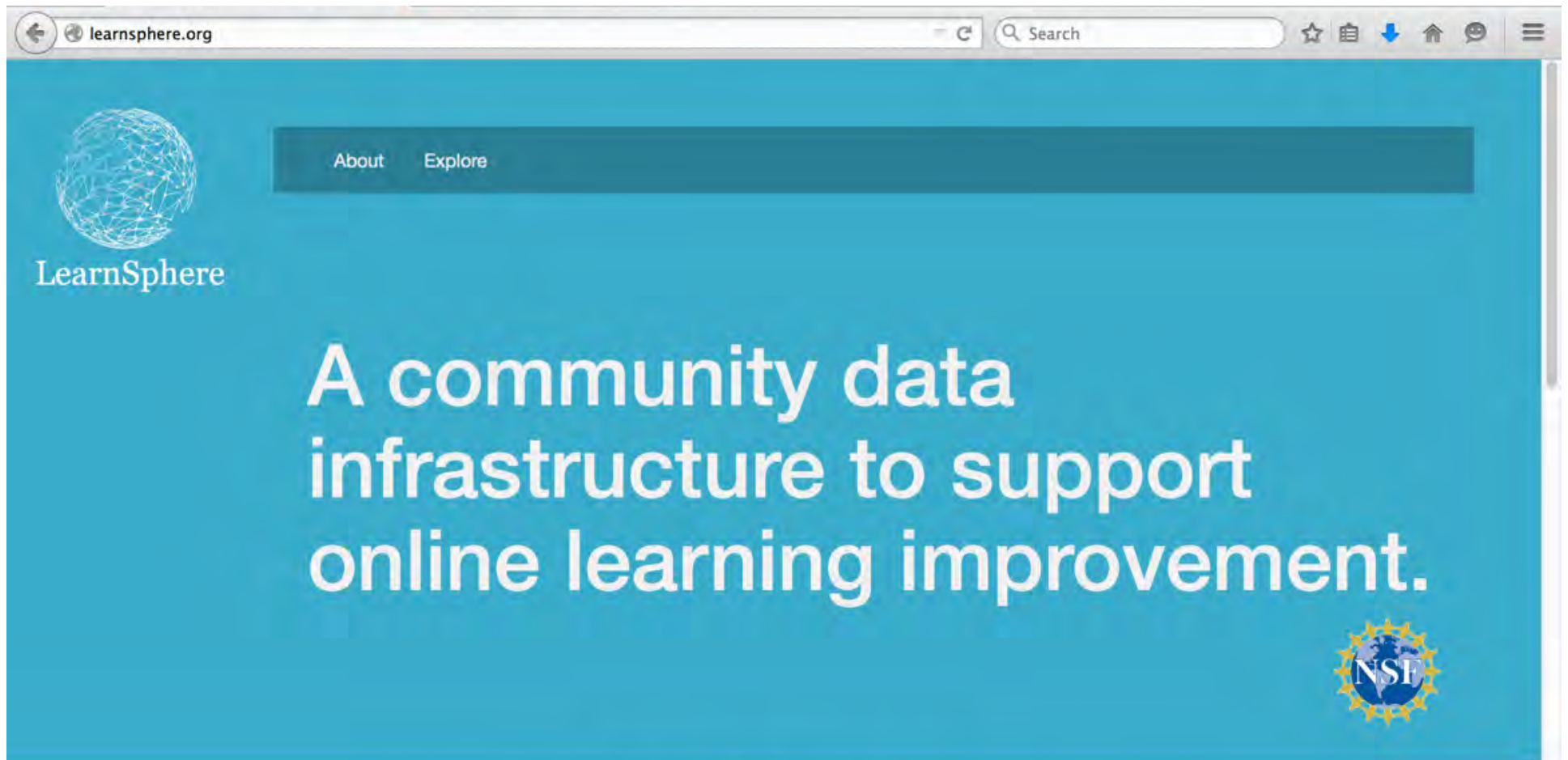
Other Data Driven Projects

- Learning Linkages: Integrating data streams of multiple modalities and timescales.
 - How to link multiple streams of data
 - What predicts what
- Data-Driven Methods to Improve Student Learning from Online Courses.
 - Applying previous methods to online courses
 - Tools for MOOCs

LearnSphere

“a community software infrastructure around the analysis of educational data that supports sharing and collaboration across the wide variety of educational data”

<http://learnsphere.org>



The image shows a browser window displaying the homepage of LearnSphere.org. The browser's address bar shows the URL "learnsphere.org" and a search bar with the text "Search". The website has a teal background. In the top left corner, there is a logo consisting of a sphere made of interconnected white lines, with the text "LearnSphere" below it. To the right of the logo, there is a dark teal navigation bar with the words "About" and "Explore" in white. The main content area features the text "A community data infrastructure to support online learning improvement." in large, white, sans-serif font. In the bottom right corner, there is a small logo for the National Science Foundation (NSF), which is a blue globe with yellow stars and the letters "NSF" in white.

learnsphere.org Search

About Explore

LearnSphere

A community data infrastructure to support online learning improvement.

NSF

Team

**Carnegie
Mellon
University**



**Stanford
University**

THE UNIVERSITY OF
MEMPHIS[®]

Data Silos → Data Integration

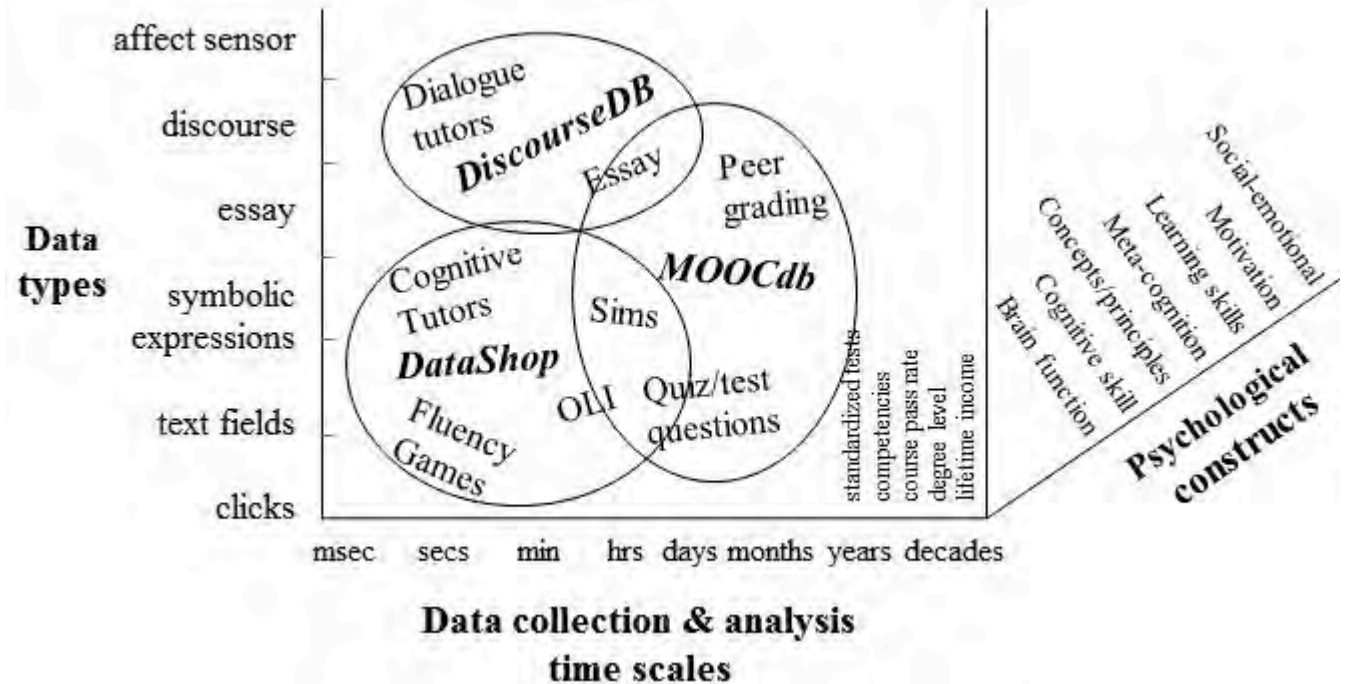
Many paradigms of data-driven education research differ in

- data types
- time scale
- research goals

Disciplinary silos are fostered by differences

Data infrastructure for analytics across these

Ultimate goal: *Produce discoveries not possible within current silos*

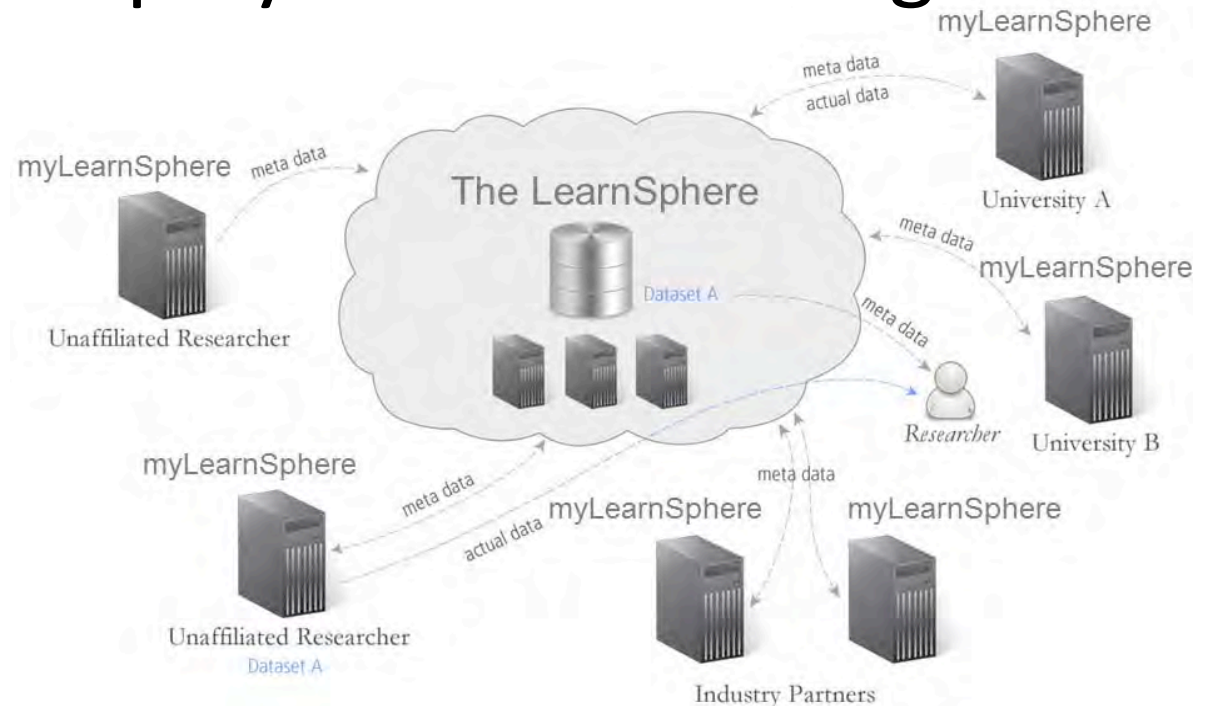


Distributed deployment and storage

Central LearnSphere portal

Multiple “myLearnSphere” installations

- Individual installations curate their own data or replicate to central repository
- Outside researchers can identify existing datasets through metadata provided by local versions



Learning Analytics Workflow Authoring Environment

The screenshot displays the LearnSphere@CMU Learning Analytics Workflow Authoring Environment. The interface includes a top navigation bar with the LearnSphere@CMU logo, a search bar, and navigation links for Tigris, Explore, Leadership, About, and Help. A user profile for "John Demo" is visible in the top right corner, along with a "logout" link. Below the navigation bar, there is a toolbar with icons for Save, Cancel, Save As, Run, and Results. The main workspace shows a workflow diagram with several nodes: "Student-Step" (Import #1), "AFM" (Analysis #2), "BKT" (Analysis #1), "Python AFM" (Analysis #3), "Learning Curves" (Visualization #1), and "Learning Curves" (Visualization #2). The workflow is connected by orange arrows, indicating the flow of data. The "Student-Step" node outputs to "AFM", "BKT", and "Python AFM". "AFM" outputs to "Learning Curves (Visualization #1)". "BKT" outputs to "Learning Curves (Visualization #2)". "Python AFM" outputs to "Learning Curves (Visualization #2)". The "Learning Curves" nodes are labeled "file". The workflow status is "Unsaved". A "Feedback" button is located in the bottom left corner. The interface also shows a "Filter by name or description" search bar and a "Search names only" option. The left sidebar contains a list of categories: Annotation, Import, Database, Transform, Analysis, Visualization, and Tetrad. The top right corner shows "Accessibility: Public Private", "Last updated: Jul 30, 2018 13:09:55", and "Workflow Status: Unsaved". A progress bar at the bottom right indicates 100% completion.

Take Aways

- The amount of data coming from educational technology is growing exponentially (Big Data is here in Education)
- Students are going to rely more and more on technology, so improving the learning in edtech systems is critical
- Human-Centered, Data-Driven approaches are most likely to be the ones that succeed in actionable improvements in edtech

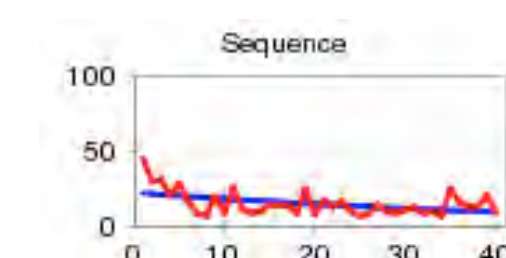
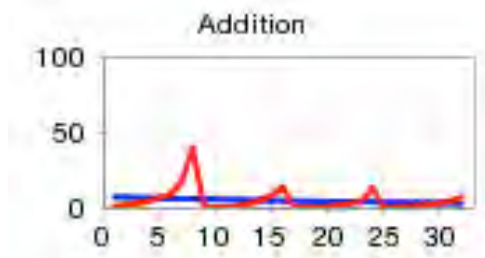
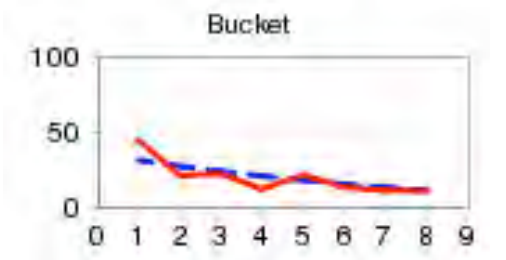
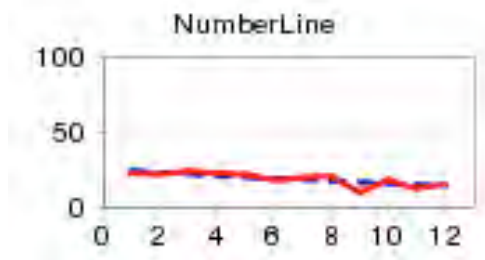


Questions?

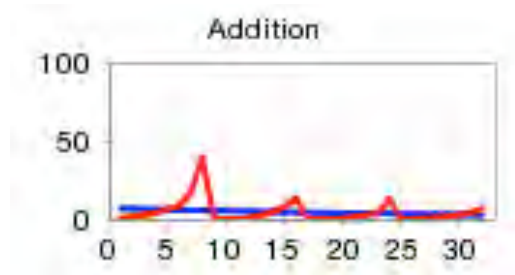
Contact me: jstamper@cmu.edu



Methods & Results - ProblemType Learning Curves



Methods & Results - Addition KC Decomposition



PUT CARRY VALUES HERE

1	1		
---	---	--	--

7.5 0 OZ
3.9 0 OZ

1 1.4 OZ

PUT CARRY VALUES HERE

	1		
--	---	--	--

1.9 0 OZ
1.9 0 OZ

3.8 OZ

PUT CARRY VALUES HERE

	1		
--	---	--	--

3.5 0 in
+ 5.6 0 in

9.1 in

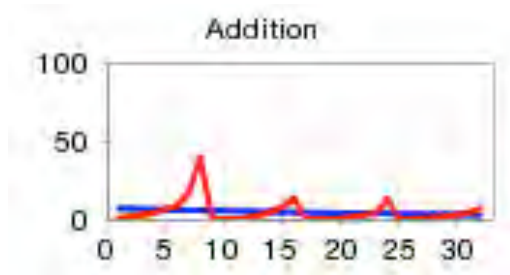
PUT CARRY VALUES HERE

	1	1	
--	---	---	--

1.3 2 in
+ 2.6 9 in

4.0 1 in

Methods & Results - Addition KC Decomposition



Addition_On
es

PUT CARRY VALUES HERE

1	1		
7	.	5	0 OZ
3	.	9	0 OZ
<hr/>			
1	1	.	4 OZ

PUT CARRY VALUES HERE

	1		
1	.	9	0 OZ
1	.	9	0 OZ
<hr/>			
	3	8	OZ

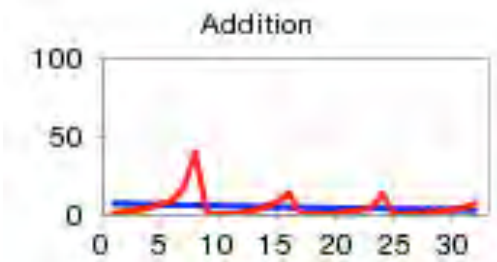
PUT CARRY VALUES HERE

	1		
3	.	5	0 in
+	5	.	6 0 in
<hr/>			
	9	.	1 in

PUT CARRY VALUES HERE

	1	1	
1	.	3	2 in
+	2	.	6 9 in
<hr/>			
	4	.	0 1 in

Methods & Results - Addition KC Decomposition



Addition_On
 es
 Addition_Tens_Non
 Zero

PUT CARRY VALUES HERE

1	1		
---	---	--	--

7.5 0 OZ
+ 3.9 0 OZ

1 1 4 OZ

PUT CARRY VALUES HERE

	1		
--	---	--	--

1.9 0 OZ
+ 1.9 0 OZ

3 8 OZ

PUT CARRY VALUES HERE

	1		
--	---	--	--

3.5 0 in
+ 5.6 0 in

9 1 in

PUT CARRY VALUES HERE

	1	1	
--	---	---	--

1.3 2 in
+ 2.6 9 in

4 0 1 in

