LearnSphere to Integrate DataShop, MOOCdb, DataStage, DiscourseDB ... Integrating Data Repositories Panel

Ken Koedinger
Professor of Human-Computer Interaction & Psychology
Carnegie Mellon University

Director of LearnLab
Pittsburgh Science of Learning Center

Workshop 2: Advancing Data-Intensive Research in Education
June 1, 2015
Big Data for education

More important than “big”

• Collected as part of natural activities
• Affords experimentation, “A/B testing”

Many dimensions of “big”

• Tall in number of participants (students)
• Wide in observations per participant (student)
• Fine in frequency of observation
• Long in spanning months or years
• Deep in theory-relevant variables
LearnSphere: Integrate across data repositories toward answering questions

We need education data infrastructures to integrate analytic methods => *produce discoveries not possible within current data silos*
Cognitive Tutors
Example source of educational data

My current cell phone company charges me $14.95 per month for service and $.13 per minute. PPS Cellular Phone Company has offered me $15.00 worth of free calls a month if I switch, but the charge is $.39 per minute.

1. How many minutes of calls can I get from PPS Cellular Phone Company for $5.00? What is the cost from my current company for that number of minutes?

2. How many minutes of calls can I get from my current company for fifty dollars? What is the cost from PPS Cellular Phone Company for that number of minutes?

3. What is the cost from both companies for sixty minutes? Which company is cheaper for sixty minutes?

4. After how many minutes of calls will the cost for both companies be the same?
Real World Impact of Cognitive Science

**Algebra Cognitive Tutor**

- Widespread intensive use
  - ~600K students per year
  - ~80 minutes per week
- Many field trials =>
  - Student learning is 2x better

- Still:
  - Could do better
  - Too many decisions driven by intuition


Pane et al. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. Santa Monica, CA: RAND Corp.
Social-technical infrastructure to discover conditions that cause robust learning

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

Researchers + Learn Lab = Schools

Since 2004

> 680 ed tech data sets in DataShop

> 320 in vivo experiments

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

What can I do with DataShop?

Upload a dataset

Dataset Name: 2013 Psych

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

Learn More
- Documentation
- About DataShop

680 data sets
math, science, language ...
K12 & college
Integrate across data repositories to answer questions

- Many complex open questions about the nature of:
  - Knowledge & cognition
  - Learning, metacognition
    - Motivation, & self-regulation
  - Instruction

- Need to work together to tackle these complex issues
  - Need to build on existing cognitive, social, education theory

Instructional Complexity
=> Need for data

More help
Basics

What’s best?
Focused practice

Gradually widen
Distributed practice

More challenge
Understanding

Study
examples

Test on
problems

Study
50/50

Test
50/50

Concrete
Mix

Abstract
Mix

Concrete
Mix

Abstract
Mix

Immediate
Delayed
No feedback

Immediate
Delayed
No feedback

Block topics
in chapters

Interleave
topics

Interleave
topics

Block topics
in chapters

Fade

Fade

Explain
Mix

Explain
Mix

Ask for
explanations

Ask for
explanations

Many other choices: animations vs. diagrams vs. not, audio vs. text vs. both, …

>3^{15*2} = 205 trillion options!

Automated support for cognitive task analysis: Discovering *hidden skills* using educational data


Learning is complex: Variations in task domains, knowledge demands, student characteristics

- Learning curves showing a decrease in error rate (y-axis) for each successive opportunity (x-axis) to learn
- Averaged across students for different skills – MORE variable
- Averaged across skills for different students – LESS variable

What causes these variations?
Turning Discovery into Better Learning

High rough curve
=> hidden skill
=> redesign instruction
=> Experiment

Better student learning!

LearnSphere: Integrate across data repositories toward answering questions

We need an education data infrastructure to integrate analytic methods => produce discoveries not possible within current data silos
A community data infrastructure to support online learning improvement.
Data Integration Example: 
**MOOC + OLI = Insight**
What student choices associate with most learning?

Learning by doing > 6x better than learning by watching!!

Koedinger et al. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. *Proceedings of Learning at Scale.*
Primary Suggestion for Action: Do data intensive research at our own universities

- Get college instructors involved!
  - Design course activities to collect data
  - Share data & seek analysis partners
  - Engage in discipline-based ed research

- Demonstrate success
  - Set a model for K12

- Incentives
  - NSF fund college-level data-driven innovation
  - Researchers enforce data reuse citation
Thank you!

Thanks to >200 researchers that have contributed!!

http://learnlab.org/DataShop

Ken Koedinger
koedinger@cmu.edu
Extras
Cognitive Model Discovery: From qualitative to quantitative

Traditional Cognitive Task Analysis
• Interviews or think alouds of experts & students
• Result: *Cognitive Model* of expert/student thinking
  – Experts aware of only ~30% of what they know
• Greatly improves instruction
  (~1.5 effect size, Clark et al)

Data-driven Cognitive Task Analysis
• Use student data from initial tutor
• Goal: more reliable & cost effective
• Employ machine learning & statistics to discover better cognitive models
Use data to develop models of learners – *because intuition is faulty!*

Which is harder for algebra students?

*Story Problem*

As a waiter, Ted gets $6 per hour. One night he made $66 in tips and earned a total of $81.90. How many hours did Ted work?

*Word Problem*

Starting with some number, if I multiply it by 6 and then add 66, I get 81.90. What number did I start with?

*Equation*

\[ x \times 6 + 66 = 81.90 \]

Math educators say: story or word is hardest

Students: equations are hardest

*Expert blind spot!*

Algebra teachers, especially, incorrectly think equations are easy

Cognitive Task Analysis using DataShop’s learning curve tools

Without decomposition, using just a single “Geometry” KC, no smooth learning curve.

But with decomposition, 12 KCs for area concepts, a smoother learning curve.

Upshot: Can automate analysis & produce better student models
Discovering a new knowledge component

- Each KC should have:
  - smooth learning curve
  - statistical evidence of learning
  - even error rates across tasks
- Find a feature common to hard tasks but missing in easy ones
Geometry Tutor
Scaffolding problem decomposition

Problem decomposition support
New model discovery: Split “compose” into 3 skills

- **Hidden planning knowledge:** *If* you need to find the area of an irregular shape, *then* try to find the areas of regular shapes that make it up.

- **Redesign instruction in tutor**
  - Design tasks that isolate the hidden planning skill
  - Given square & circle area, find leftover

---

**Worksheet:**

<table>
<thead>
<tr>
<th>Diagram Label</th>
<th>inches</th>
<th>length of square ABCD</th>
<th>Area of scrap metal</th>
<th>Area of SQUARE ABCD</th>
<th>Area of END OF CAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>4</td>
<td>0</td>
<td>13.76</td>
<td>64</td>
<td>50.24</td>
</tr>
<tr>
<td>Question 2</td>
<td>8</td>
<td>16</td>
<td>56.54</td>
<td>256</td>
<td>200.96</td>
</tr>
<tr>
<td>Question 3</td>
<td>12</td>
<td>24</td>
<td>123.84</td>
<td>576</td>
<td>452.16</td>
</tr>
</tbody>
</table>

When prompts are initially present for component areas.

To make metal cans, the ends for the cans are stamped out of square pieces of metal.
3-way split in new model (green) better fits variability in error rates than original (blue)
Where to go from here?

Possible partnerships/collaborations/relationships to pursue Cyberlearning advances through data sharing?
• Analyses that span levels of analysis

Key needs to be both effective & legal
• Data sharing cyberinfrastructure
  – Easy to use
  – Layered & managed access
  – Rigorous privacy review: IRB+
• Researcher incentives for sharing
  – Sticks: Funder requirements, journal requirements
  – Carrots: Data citation, badges, shared data/analytics counts toward tenure
What’s needed in Cyberlearning data partnerships?

As many as possible of:

• Shared datasets with
  – long-term robust learning & life outcomes
  – multiple assessments: performance, standardized, future learning
  – fine-grain, wide, & deep *click* data
  – fine-grain, wide, & deep *verbal* data
  – embedded experiments: 1 or more random variations

• Analytics sharing with *easy to*
  – access existing analytics
  – apply analytics to full space of Cyberlearning data sources
    • Online courses, simulations, games, tutors, inquiry, class video, ubiquitous computing...
  – recombine existing analytics without programming
  – contribute new analytics & new workflows

• Teams with compatible goals
  – interdisciplinary: education, computer science, psychology, economics ...
  – instructors drive research goals

• OTHERS???
Big Data for Learning Conclusions

• Big data can help unlock mysteries of human learning
  – Science & technology to support learning will transition from Model T to Jet Airplane

• Not the “big” that is important
  – Natural collection: tall, wide, fine, long, deep

• Future: Big data partnerships to tackle big interdisciplinary education questions
Five Recommendations

1. Search in the “function space”
2. Experimental tests of instructional function decomposability
3. Massive online multifactor studies
4. Learning data infrastructure
5. School-researcher partnerships