

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

Ken Koedinger

Professor of Human-Computer Interaction & Psychology
Carnegie Mellon University

Director of



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

Big Data for education

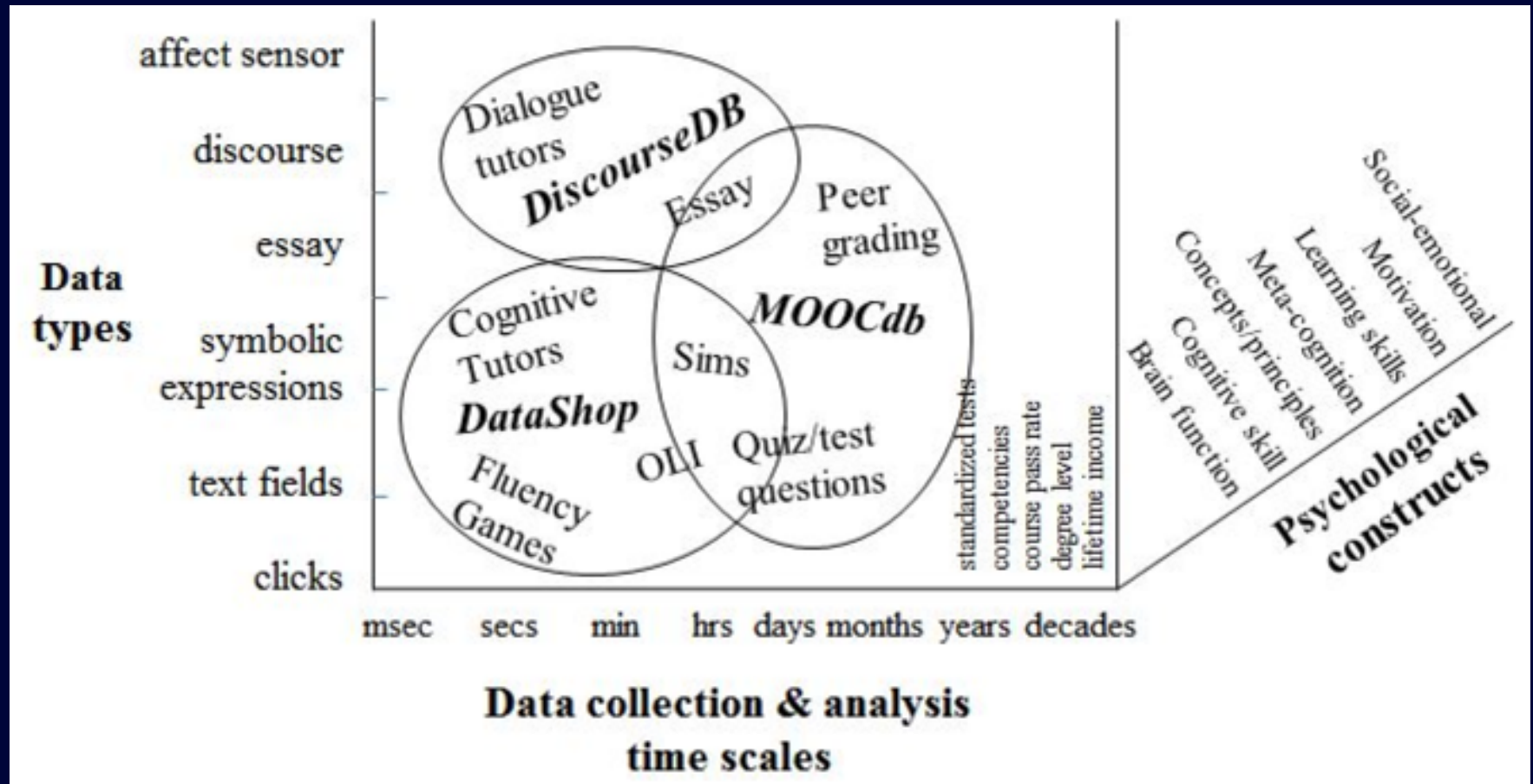
More important than “big”

- 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 a 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.

Quantity Name	Time	Current cost
Unit	minutes	\$
Expression	t	.13t
Question 1		
Question 2		
Question 3		
Question 4		

The cost from my current company increases by 0.13 each minute, but remember that it starts at 14.95 dollars.

Authentic problems

Feedback *within* complex solutions

Progress...

Personalized instruction

Challenging questions

4. After how many minutes of calls will the cost for both companies be the same?

Scenario
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 \$50? 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

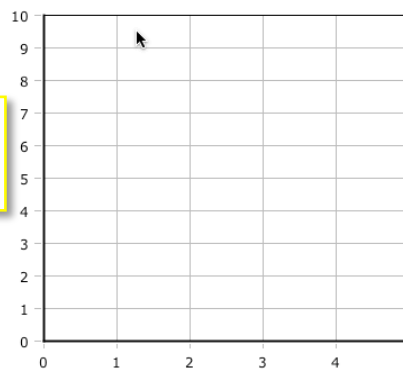
companies be the same?

Worksheet

Quantity Name
Unit
Expression
Question 1
Question 2
Question 3
Question 4

Grapher

10.0



0.0

0.0

Legend: $y =$ Enter Label
Equations: $y =$ Enter Equation

Hint
If the cost from my current company and the cost from PPS Cellular Phone Company are equal, then their expressions are equal. Write an equation and solve it to find the number of minutes.

Close << Previous Hint Next Hint >>

... individualization

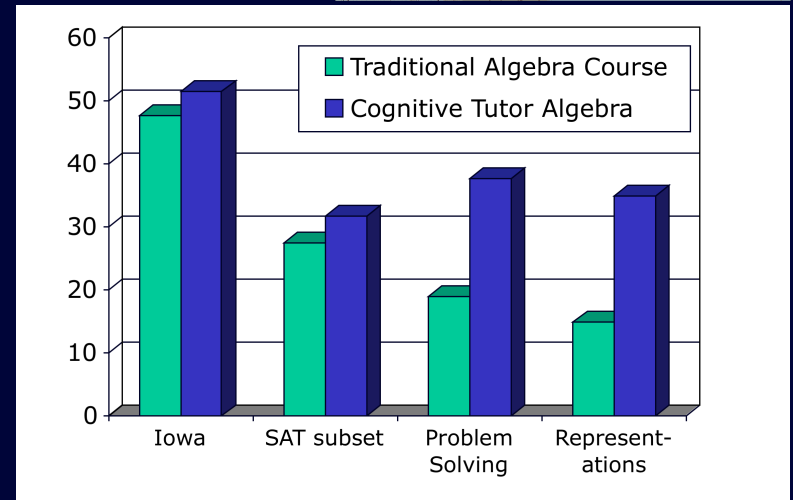
- Calculate input value.
- Writing expression, any form.
- Set axis bounds.
- Label point of intersection.
- Enter given.

Deep

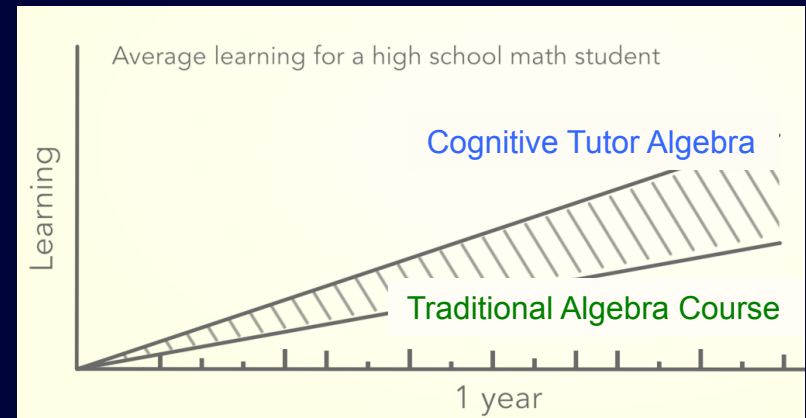
Real World Impact of Cognitive Science

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



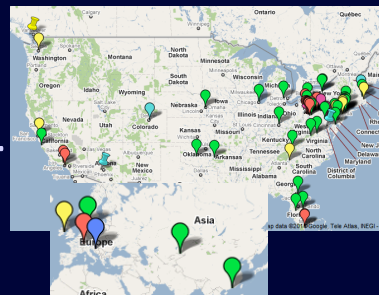
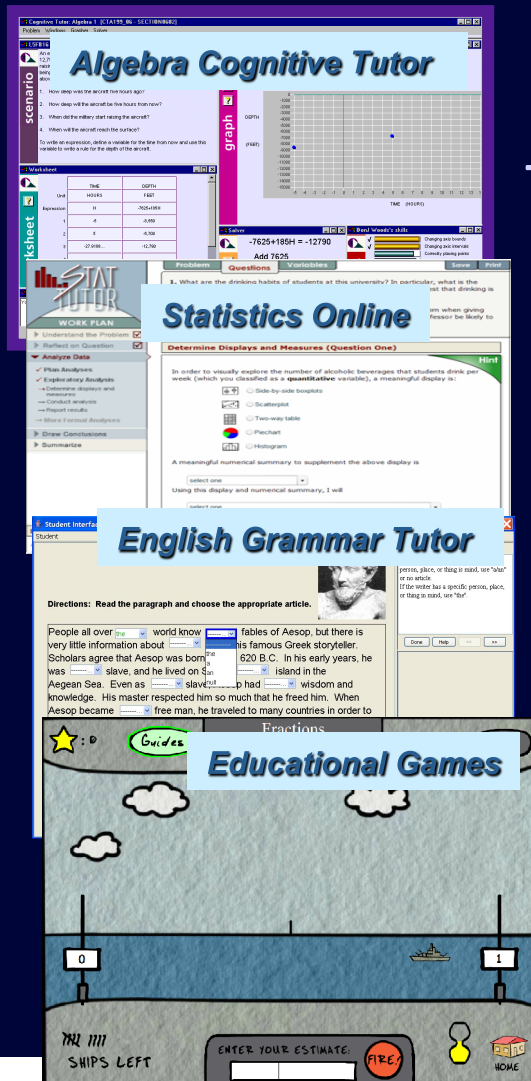
Koedinger, Anderson, Hadley, & Mark (1997).
Intelligent tutoring goes to school in the big city.



Pane et al. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. Santa Monica, CA: RAND Corp.

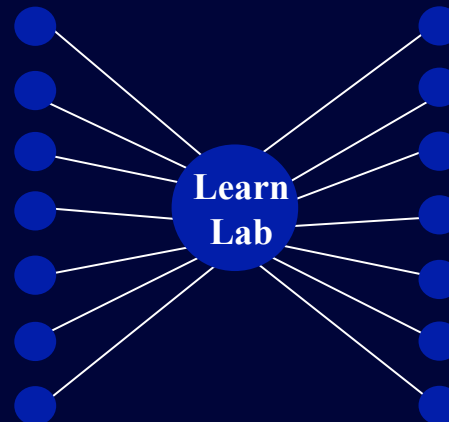
Social-technical infrastructure to discover conditions that cause *robust learning*

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



Researchers

Schools



Since 2004
> 680 ed tech data
sets in DataShop

> 320 *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*.

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 ...

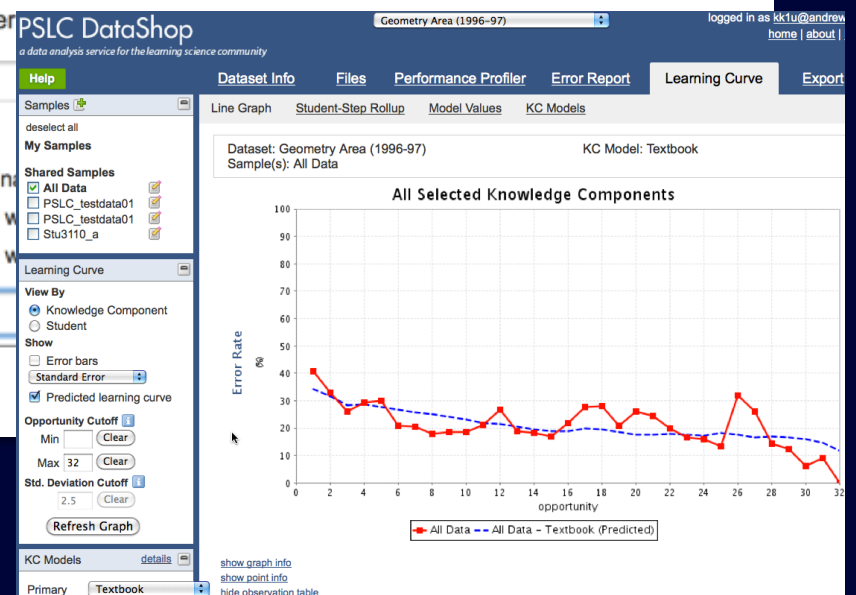
☒ new project ☐ existing project ☐ choose later

Project Name Psychology MOOC data

Data Collection Type
☐ Not specified
☐ Not human subjects data (not original)
☒ Study data collected under an IRB w/ approval
☐ Study data collected under an IRB w/ exemption

Dataset Name 2013 Psych|

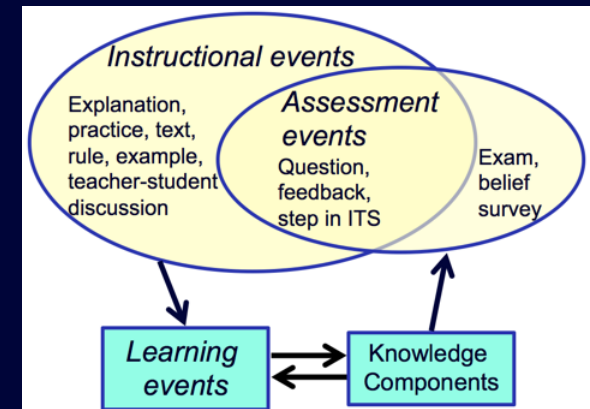
Description (optional)



680 data sets
 math, science, language ...
 K12 & college

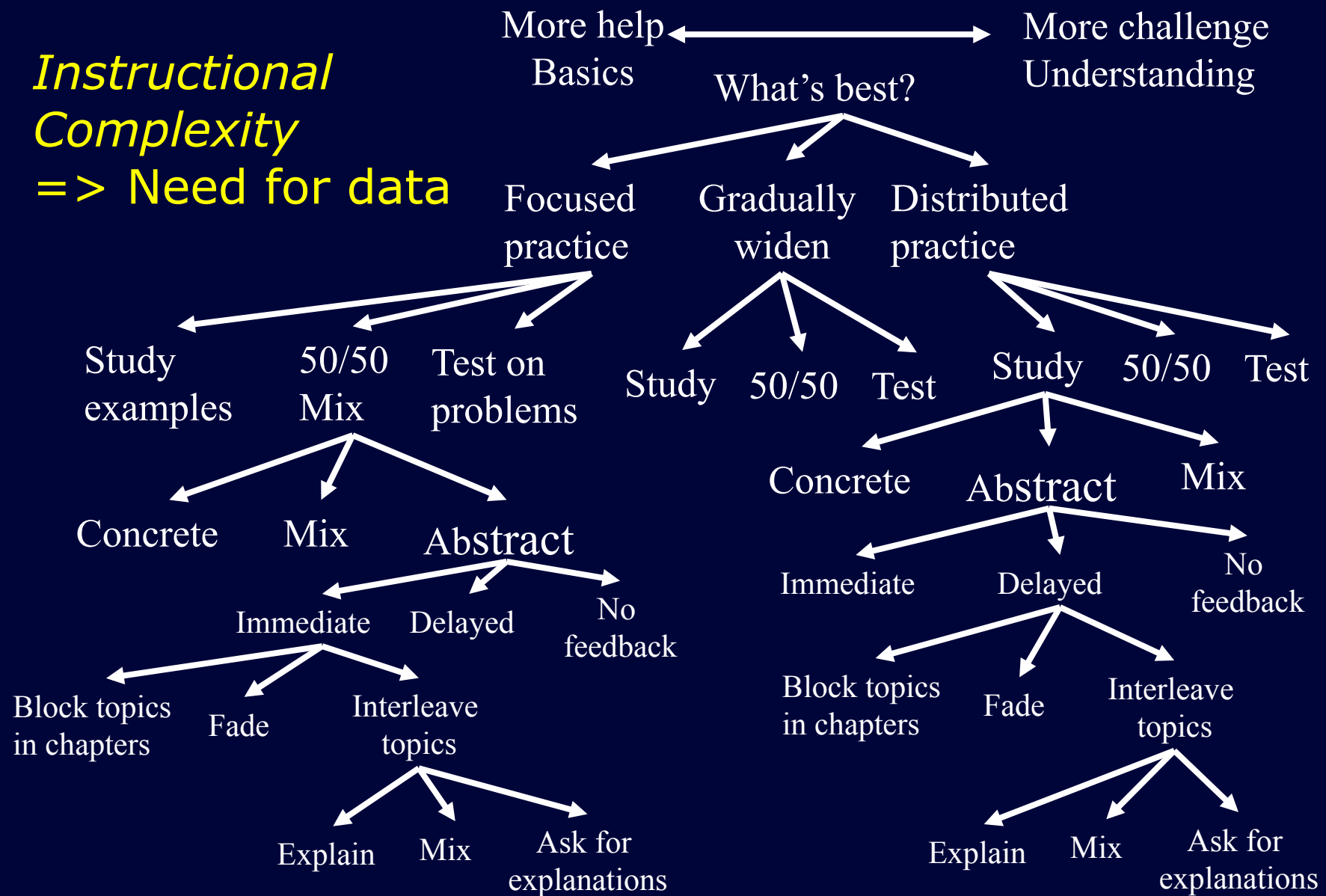
Integrate across data repositories to answer questions

- 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



Koedinger et al. (2012). The Knowledge-Learning-Instruction (KLI) framework. *Cognitive Science*.

Instructional Complexity => Need for data



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

Koedinger, Booth, Klahr (2013). Instructional Complexity and the Science to Constrain It. *Science*.

$> 3^{15 \times 2} = 205$ trillion options!

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

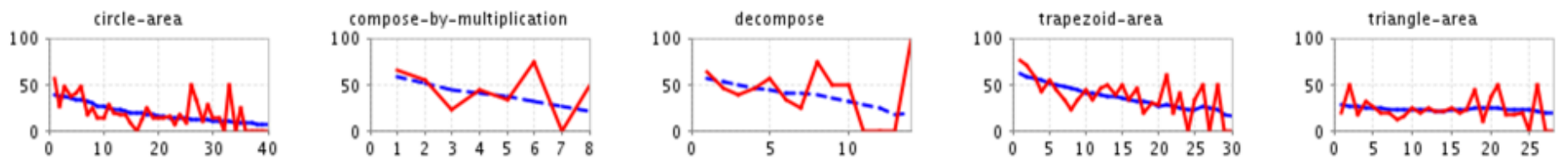
Cen, H., Koedinger, K., Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. *8th International Conference on Intelligent Tutoring Systems*.

Koedinger, McLaughlin, & Stamper (2012). Automated student model improvement. In *Proceedings of the Fifth International Conference on Educational Data Mining*. [Conference best paper.]

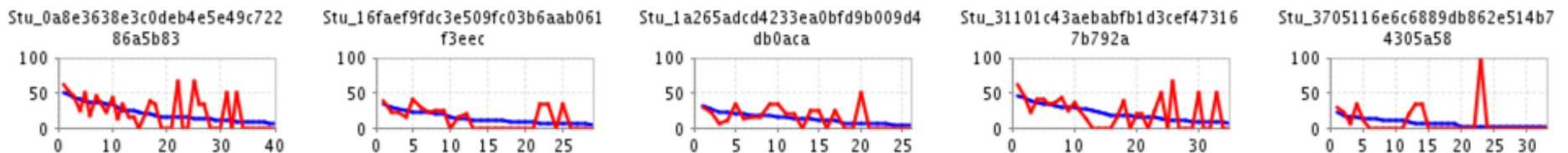
Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education*.

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

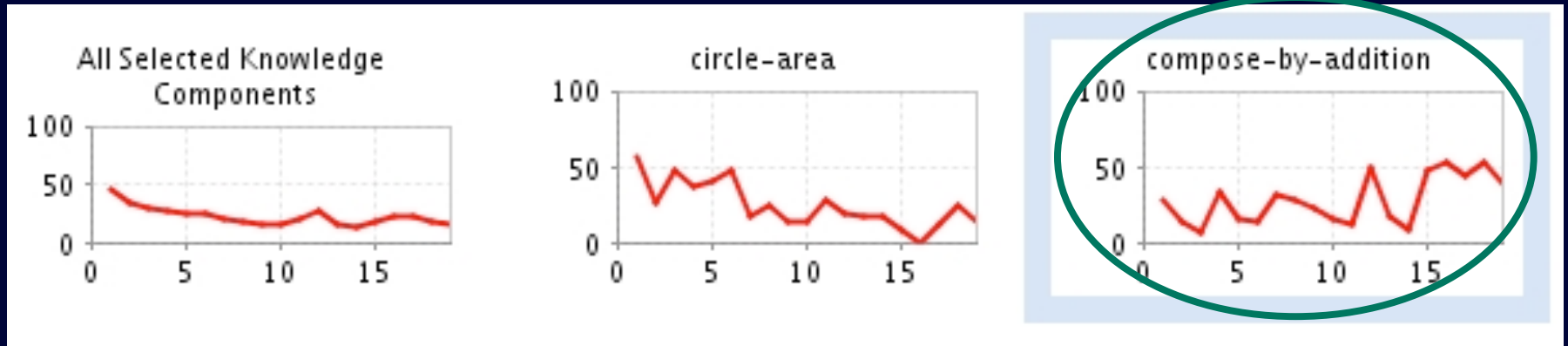
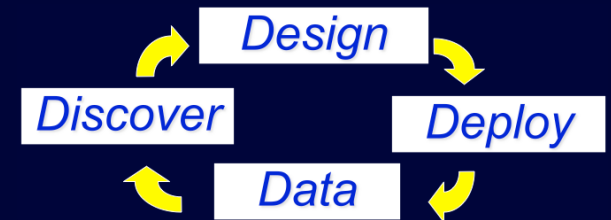


- Averaged across skills for different students – LESS variable



- What causes these variations?

Turning Discovery into Better Learning



High rough curve

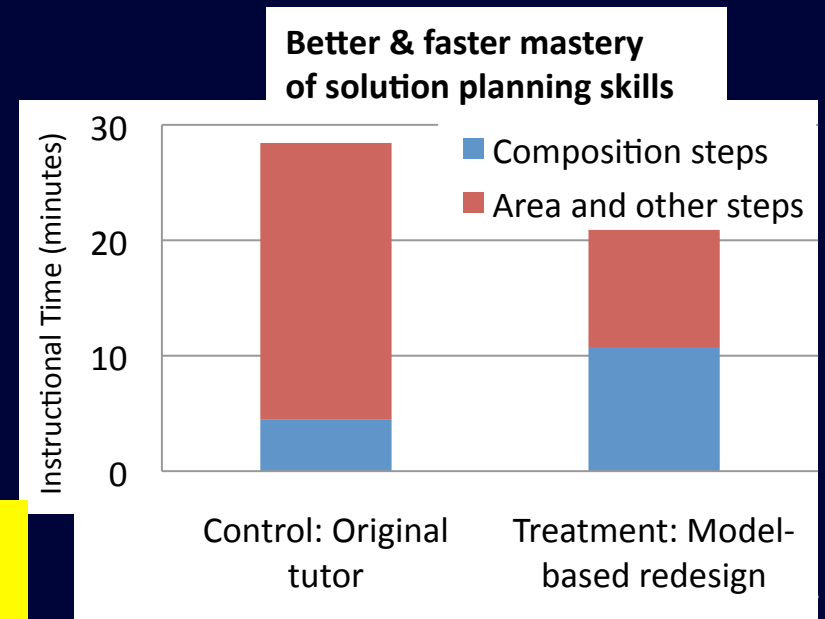
=> hidden skill

=> redesign instruction

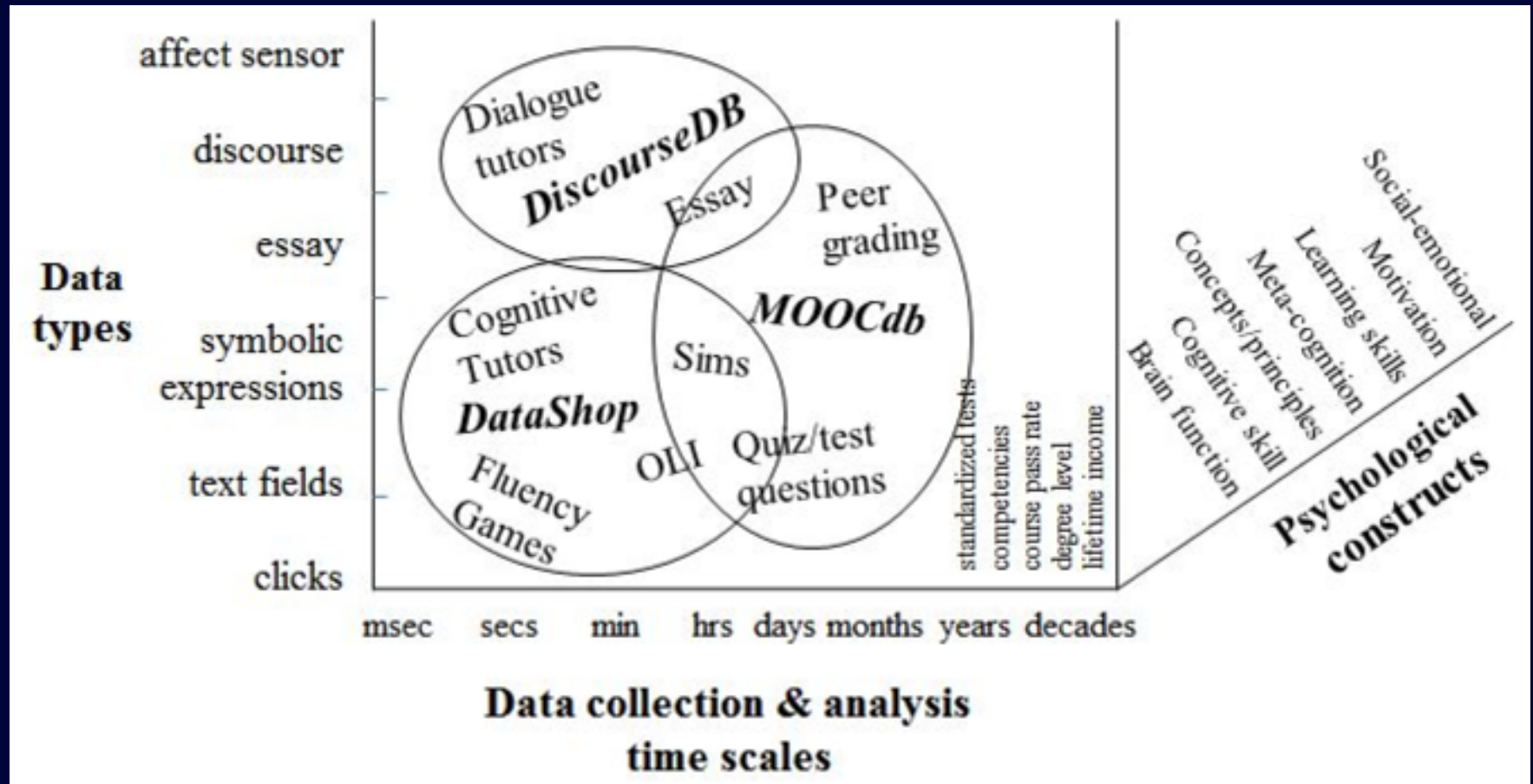
=> Experiment

Better student learning!

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



LearnSphere: Integrate across data repositories toward answering questions



We need a education data infrastructures to integrate analytic methods
=> *produce discoveries not possible within current data silos*



LearnSphere

About Explore

A community data infrastructure to support online learning improvement.



Existing Resources

PSLC DataShop
a data analysis service for the learning science community

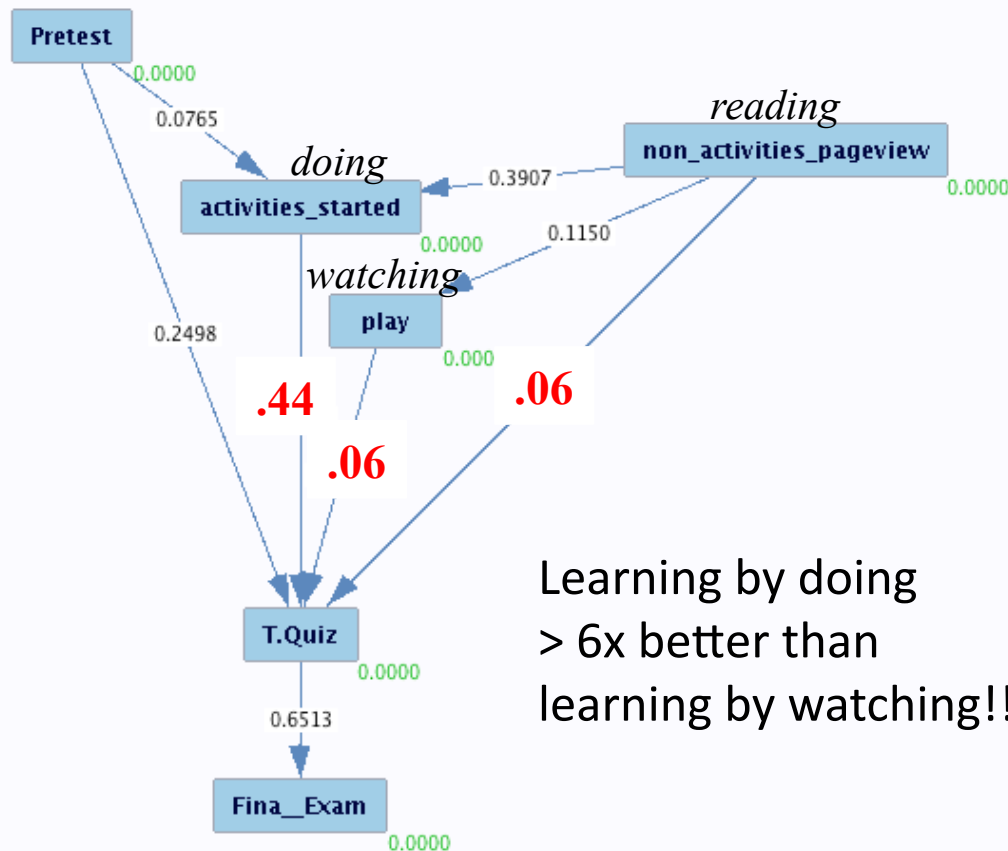
MOOCDB



Data Integration Example:

MOOC + OLI = Insight

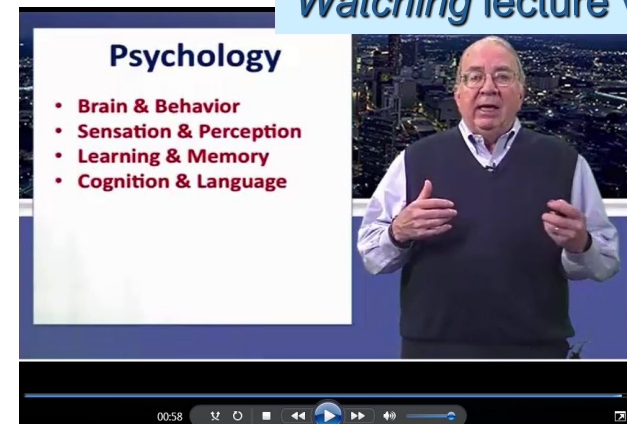
What student choices associate with most learning?



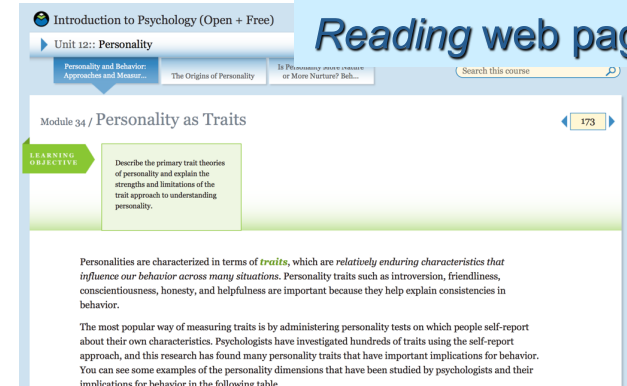
Learning by doing
> 6x better than
learning by watching!!

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*.

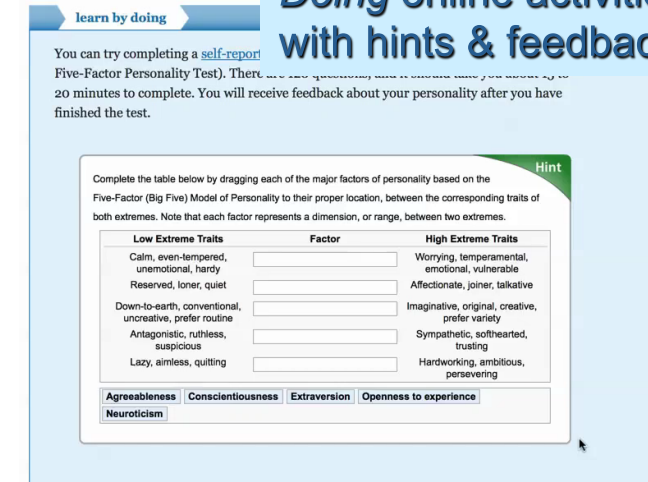
Watching lecture video



Reading web pages



Doing online activities with hints & feedback



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

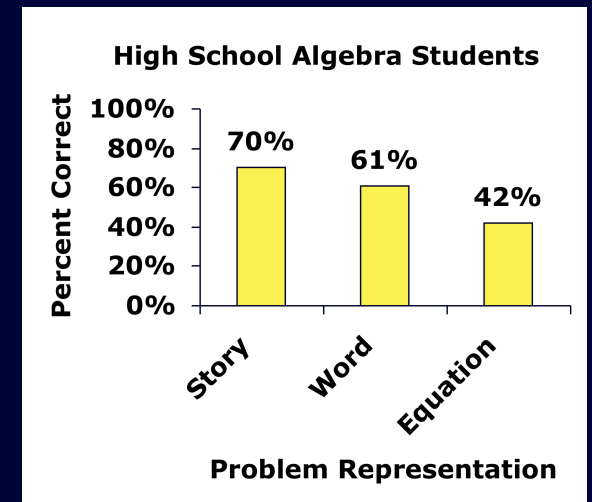
Starting with some number, if I multiply it by 6 and then add 66, I get 81.90. What number did I start with?

Equation

$$x * 6 + 66 = 81.90$$

Math educators say:
story or word is hardest

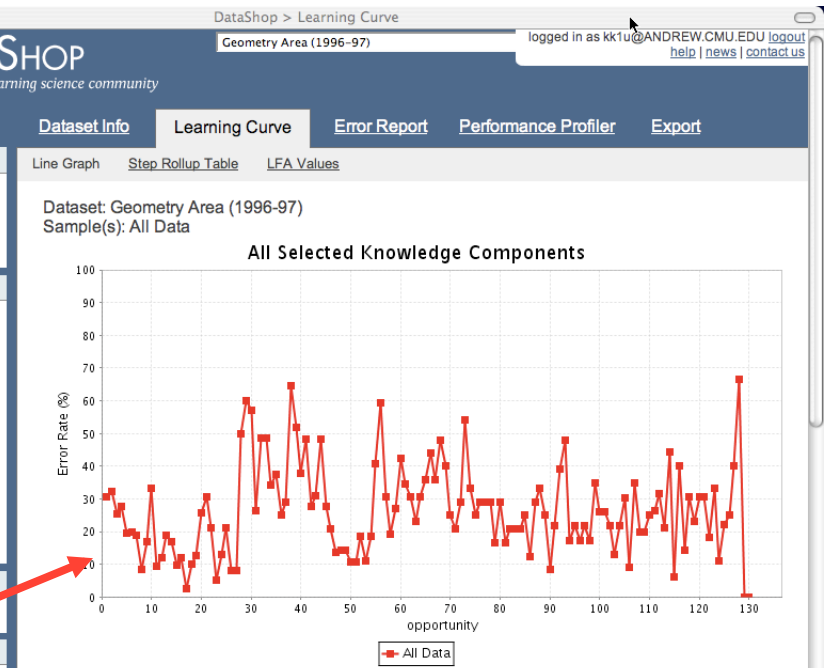
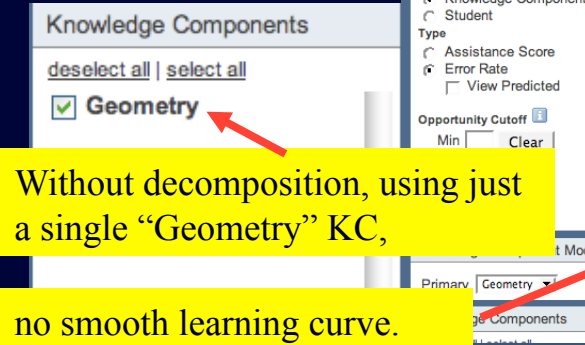
Students:
equations
are hardest



Expert blind spot!

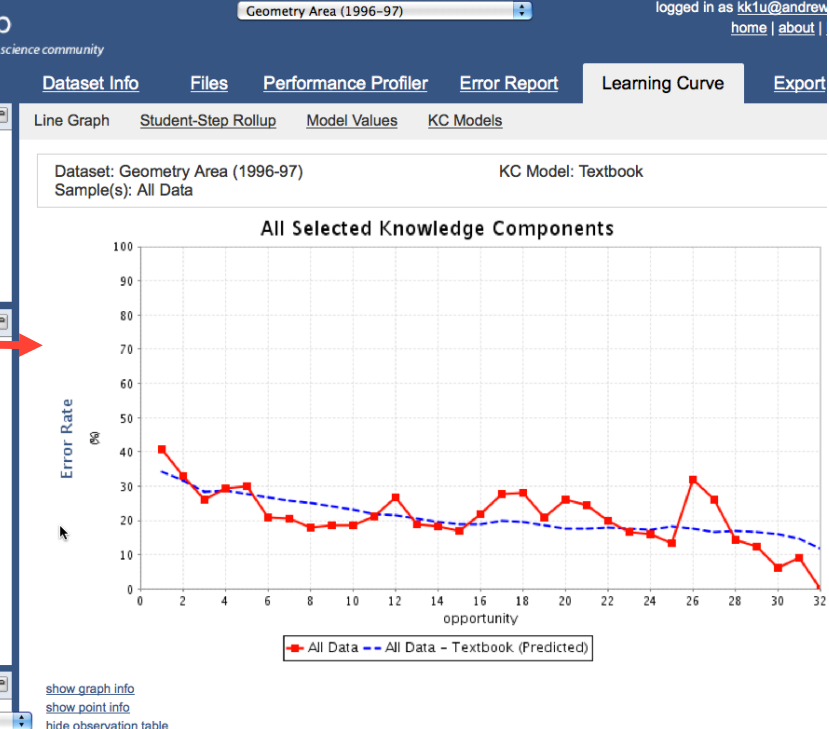
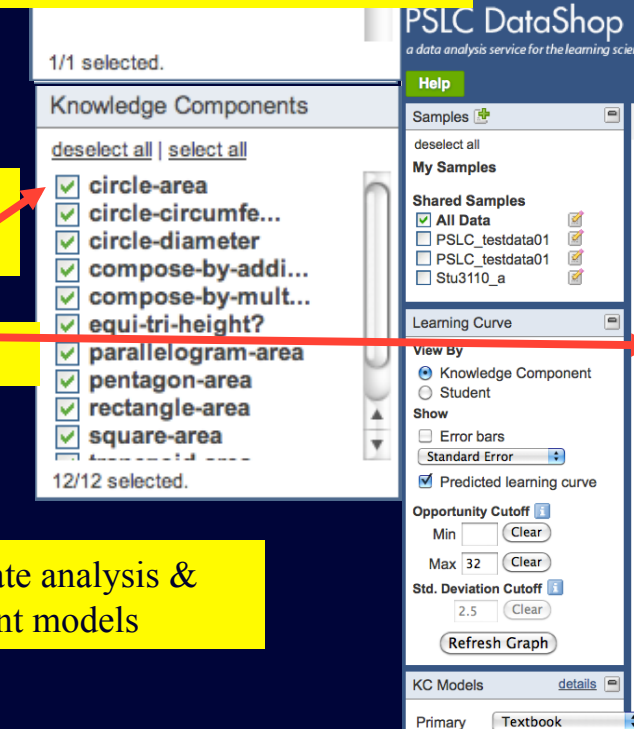
Algebra teachers, especially,
incorrectly think equations are easy

Cognitive Task Analysis using DataShop's learning curve tools



But with decomposition, 12
KCs for area concepts,

a smoother learning curve.

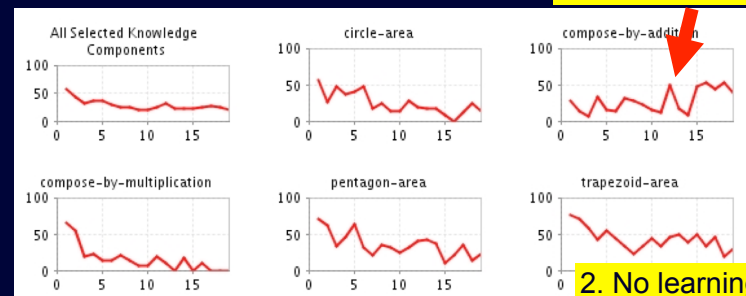


Upshot: Can automate analysis &
produce better student models

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

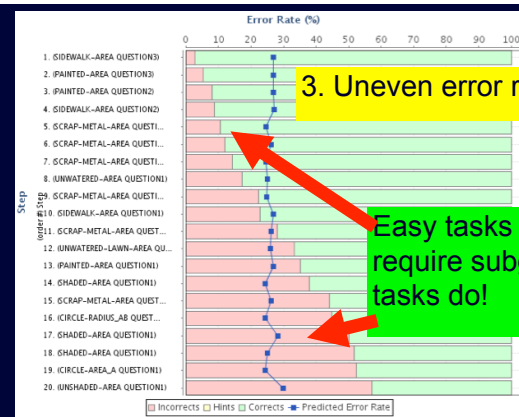
1. Not smooth



2. No learning

KC Name	Intercept	Slope
circle-area	0.58	0.068
compose-by-addition	0.74	0
compose-by-mult	0.6	0.114
pentagon-area	0.37	0.110
trapezoid-area	0.35	0.091

3. Uneven error rate



Easy tasks do not require subgoals, hard tasks do!

Geometry Tutor

Scaffolding problem decomposition

Problem decomposition support

The screenshot displays the Geometry Tutor interface, which is divided into three main windows: Scenario, Worksheet, and Skills.

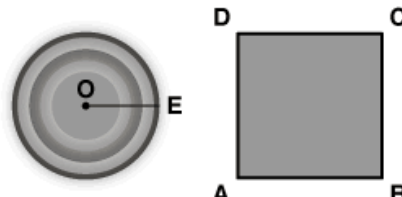
Scenario Window: This window contains the problem description and diagrams. It shows a diagram of a circular can end (labeled 'End of Can') and a square piece of metal (labeled 'Metal Square'). The square is labeled with vertices A, B, C, and D. The can end is labeled with center O and radius E. Below the diagrams, the problem 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." Three questions follow, each asking for the area of the scrap metal given a specific radius and square side length. A note at the bottom states: "NOTE: To find the area of the scrap metal remaining, you might have to first find the area of the can end, and the area of the metal square. For this problem use an approximate value for pi. $\pi \approx 3.14$ ".

Worksheet Window: This window is used for problem decomposition. It features a menu bar with 'File', 'Edit', 'Tutor', 'Worksheet', 'Windows', and 'Help'. The 'Worksheet' menu is open, showing options like 'Add Column' and 'Delete Column'. Below the menu is a table with four columns: 'Unit', 'length of the square ABCD', 'Area of the scrap metal', and 'AREA OF SQUARE ABCD'. The table has four rows: 'Diagram Label', 'Question 1', 'Question 2', and 'Question 3'. The 'Diagram Label' row has 'AB' in the second column. The 'Question 1' row has '4' in the first column, '8' in the second column, '4' in the third column, and '64' in the fourth column. The 'Question 2' and 'Question 3' rows are empty. A red circle highlights the 'Add Column' button and the 'Area of the scrap metal' column. Another red circle highlights the 'Question 1' row.

Skills Window: This window is titled 'Geo Unit01-6's skills' and contains a progress bar and the text 'Adding/subtracting areas'. It also shows the progress '1 Area / 6 Composite / Making Cans'.

New model discovery: Split “compose” into 3 skills

Scenario



End of Can Metal Square

To make metal cans, the ends for the cans are stamped out of square pieces of

Worksheet

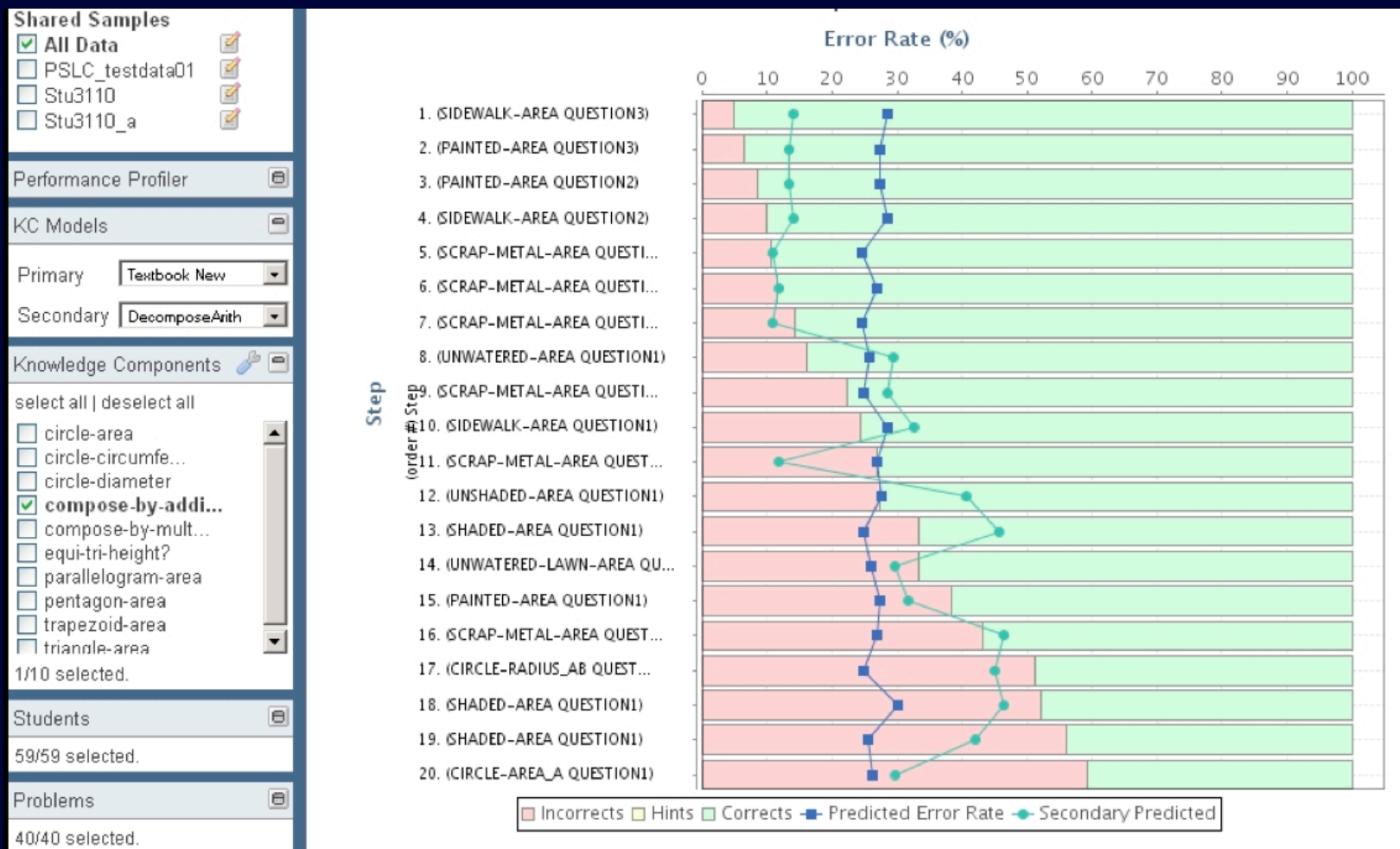
	radius of the end of the can	length of the square ABCD	Area of the scrap metal	AREA OF SQUARE ABCD	AREA OF END OF CAN
Unit	inches	inches	square inches	SQUARE INCHES	SQUARE INCHES
Diagram Label		AB			
Question 1	4	8	13.76	64	50.24
Question 2	8	16	55.04	256	200.96
Question 3	12	24	123.84	576	452.16

- 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

3

When prompts are initially present for component areas

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