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

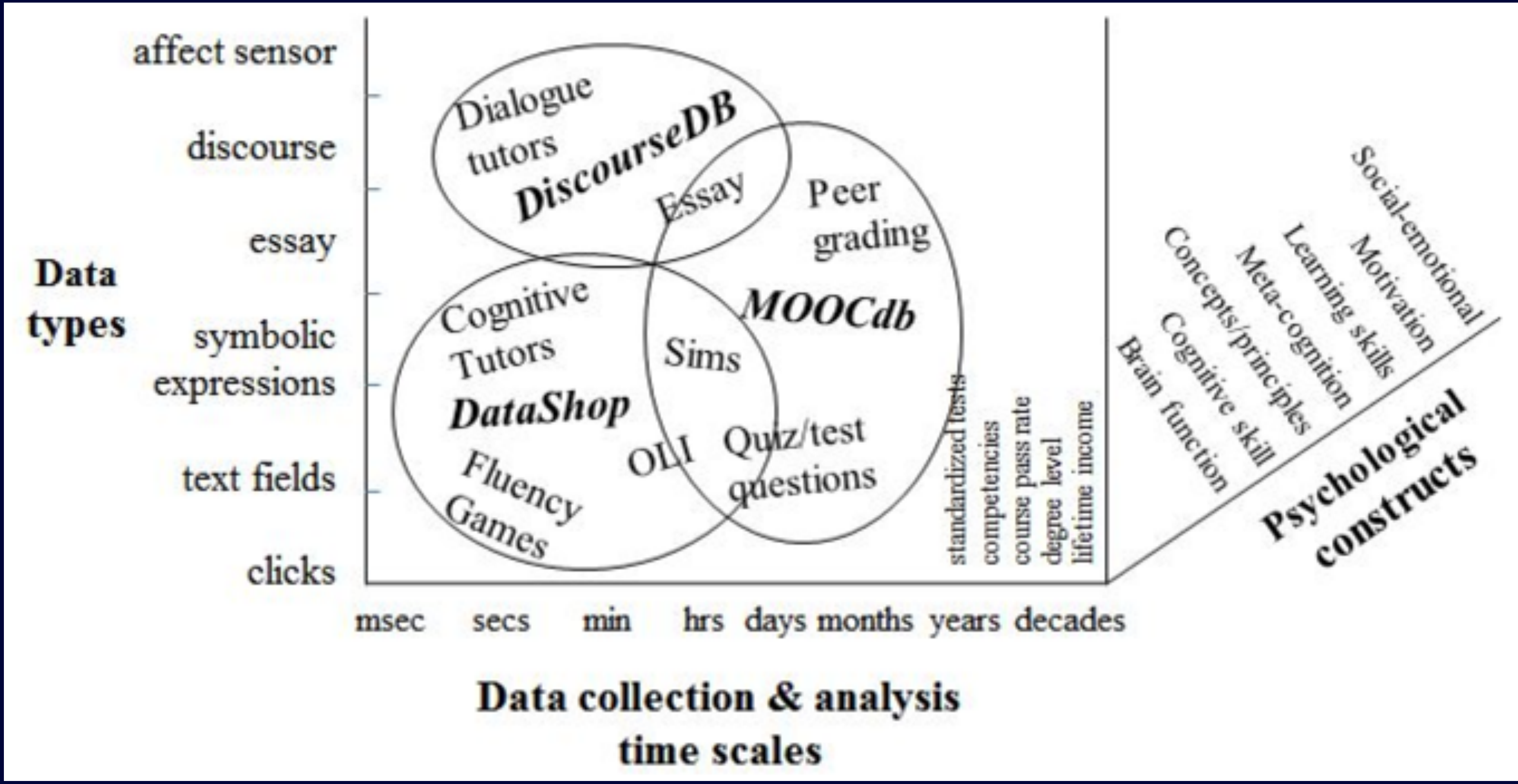
More important than "big"

- 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

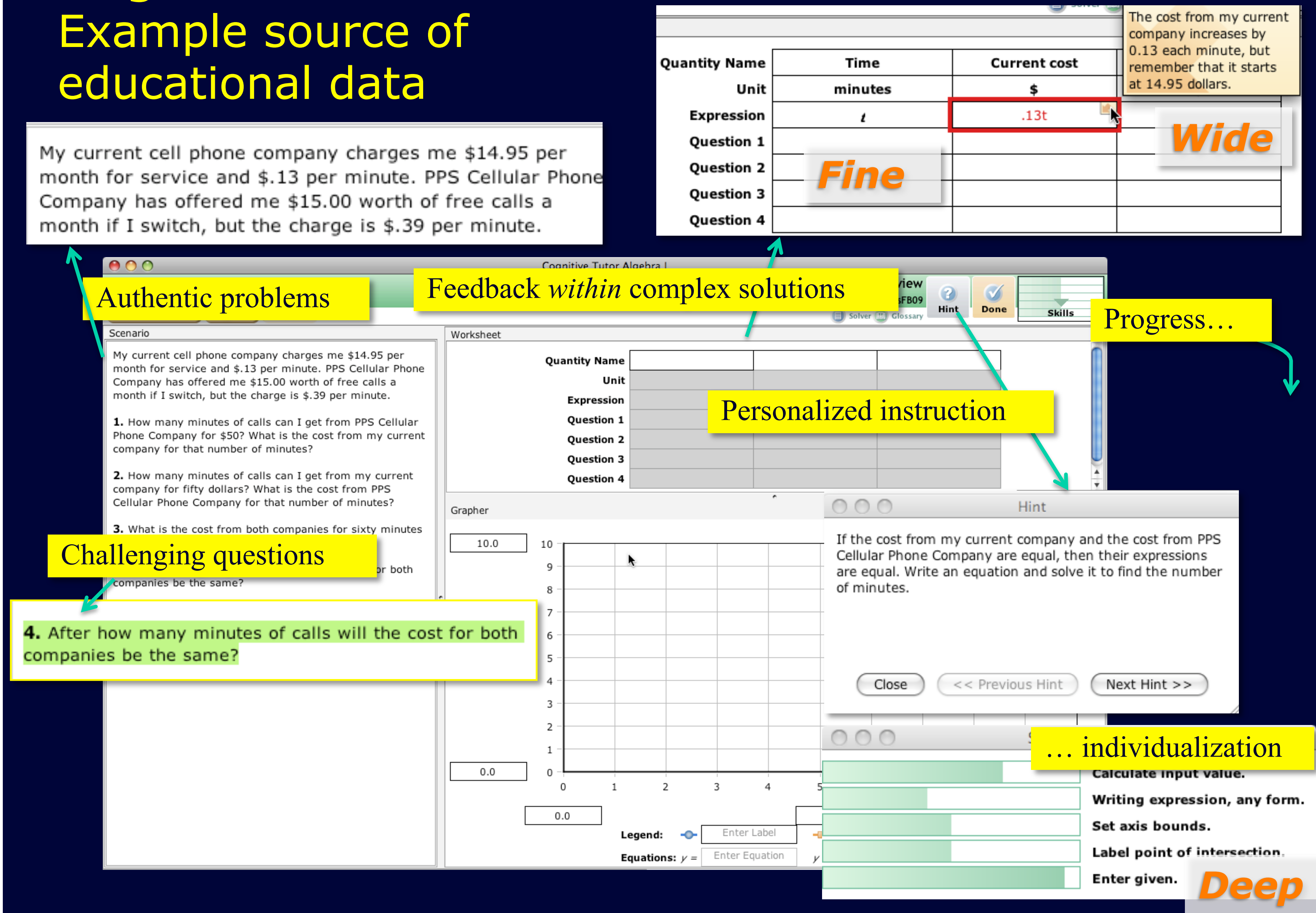
LearnSphere: Integrate across data repositories toward answering questions



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

Cognitive Tutors

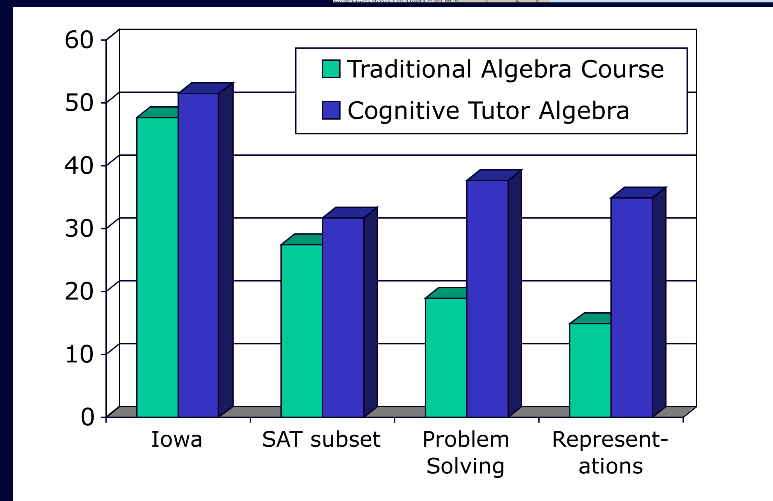
Example source of educational data



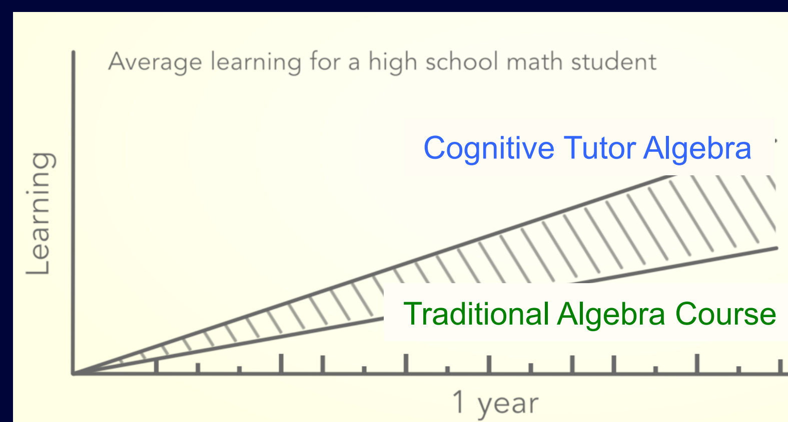
Real World Impact of Cognitive Science

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



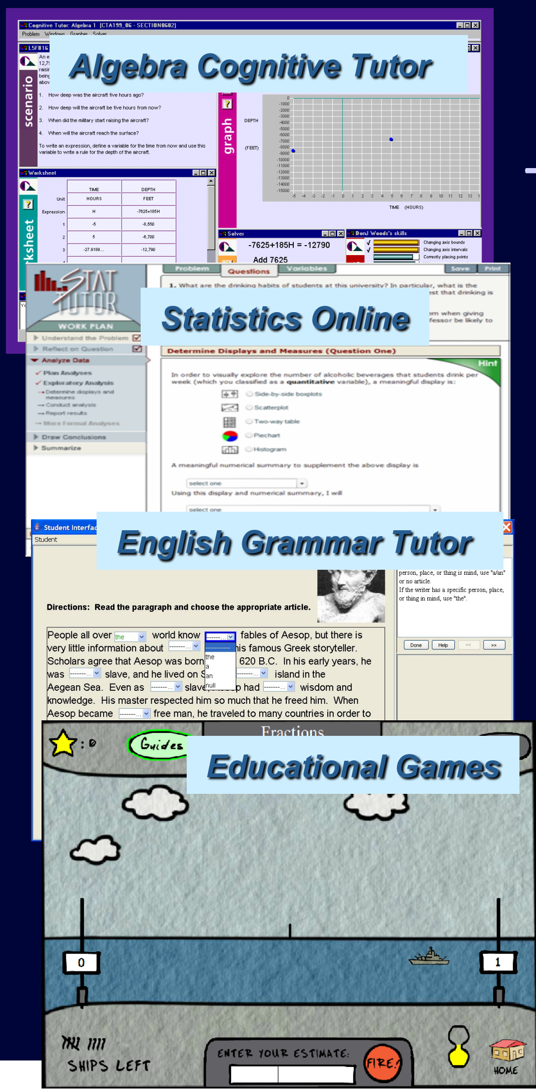
Koedinger, Anderson, Hadley, & Mark (1997). Intelligent tutoring goes to school in the big city.



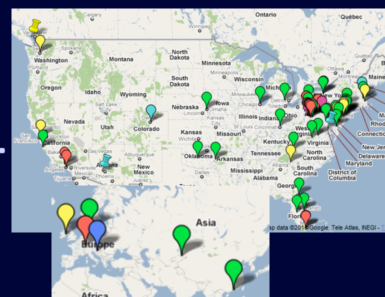
Pane et al. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. Santa Monica, CA: RAND Corp.

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

Ed tech + wide use = "Basic research *at scale*"

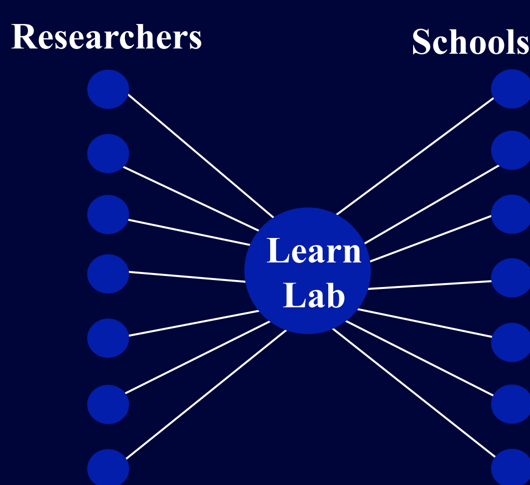


+



=

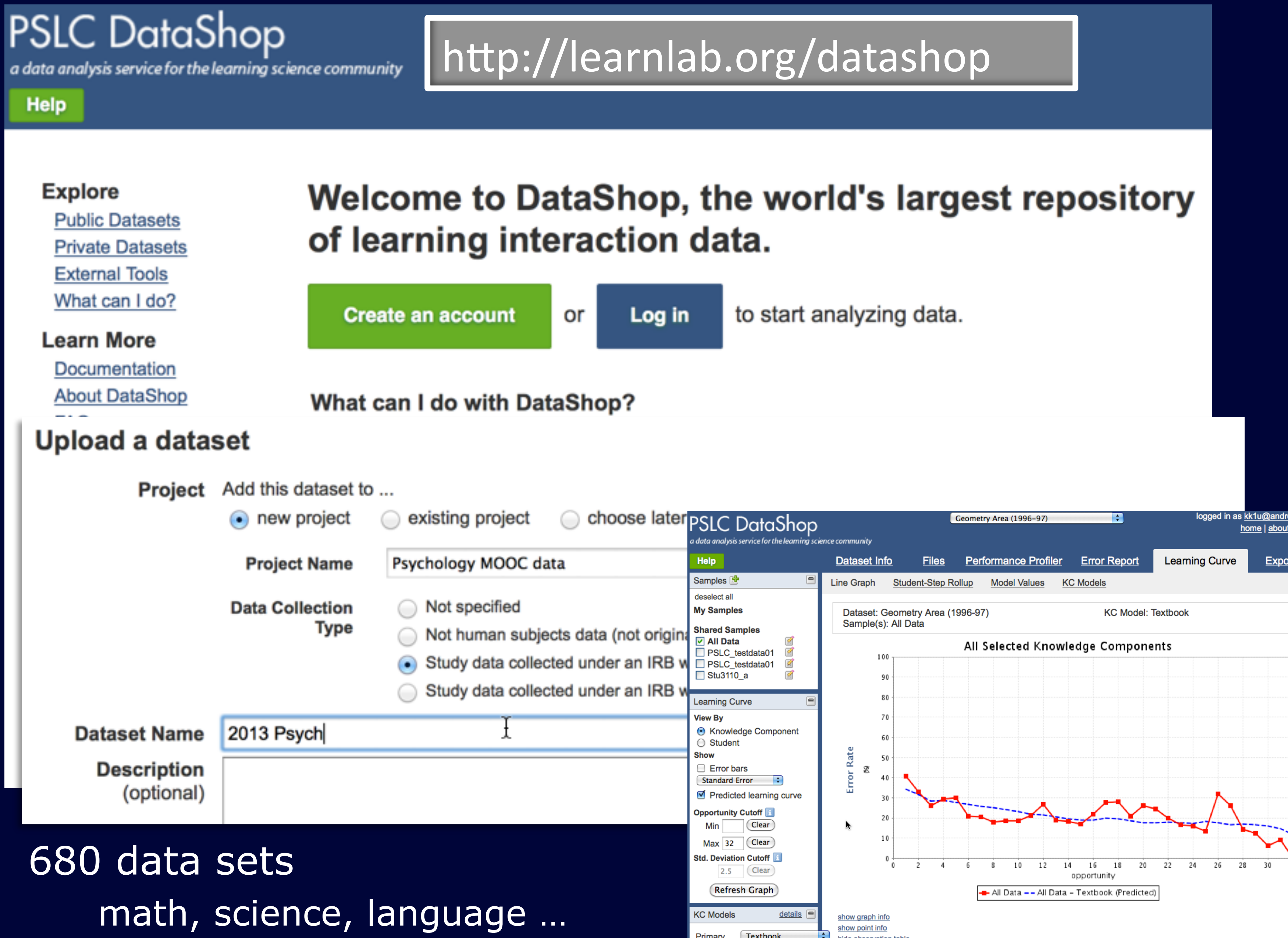
LearnLab
Pittsburgh Science of Learning Center



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*.



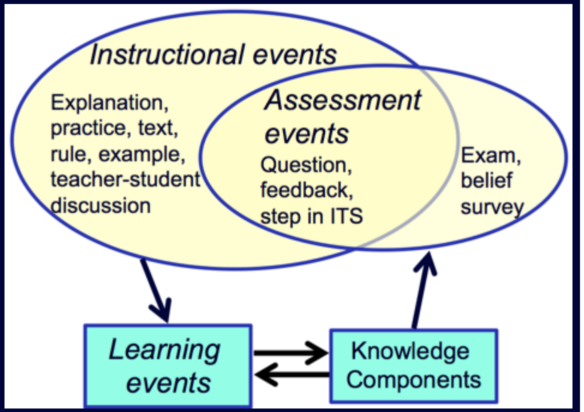
680 data sets

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

More help ↔ More challenge
Basics — Understanding

- What's best?
  - Focused practice
    - Study examples
    - 50/50 Mix
      - Concrete
      - Mix
      - Abstract
        - Immediate
          - Block topics in chapters
          - Fade
          - Interleave topics
            - Explain
            - Mix
            - Ask for explanations
        - Delayed
        - No feedback
    - Test on problems
  - Gradually widen
    - Study
    - 50/50
    - Test
  - Distributed practice
    - Study
      - Concrete
      - Abstract
        - Immediate
        - Delayed
          - Block topics in chapters
          - Fade
          - Interleave topics
            - Explain
            - Mix
            - Ask for explanations
        - No feedback
      - Mix
    - 50/50
    - Test

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.*

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

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

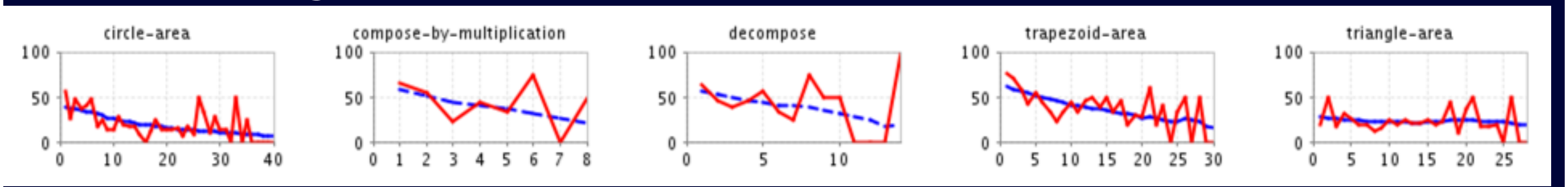
Cen, H., Koedinger, K., Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. *8th International Conference on Intelligent Tutoring Systems*.

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

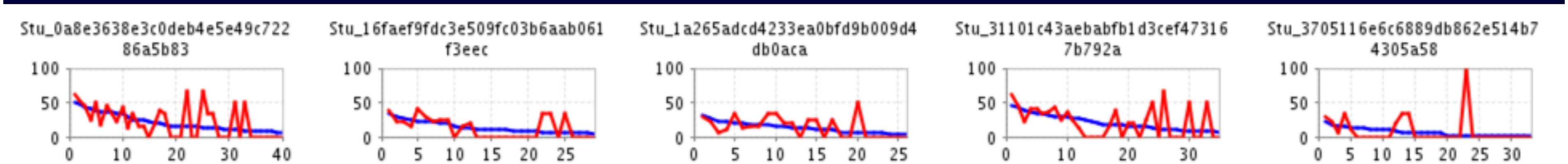
Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education.*

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

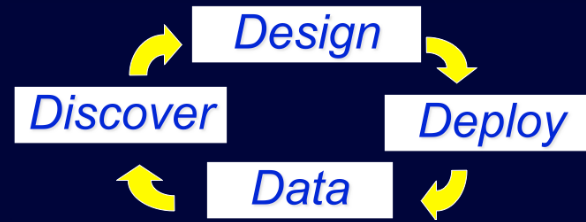


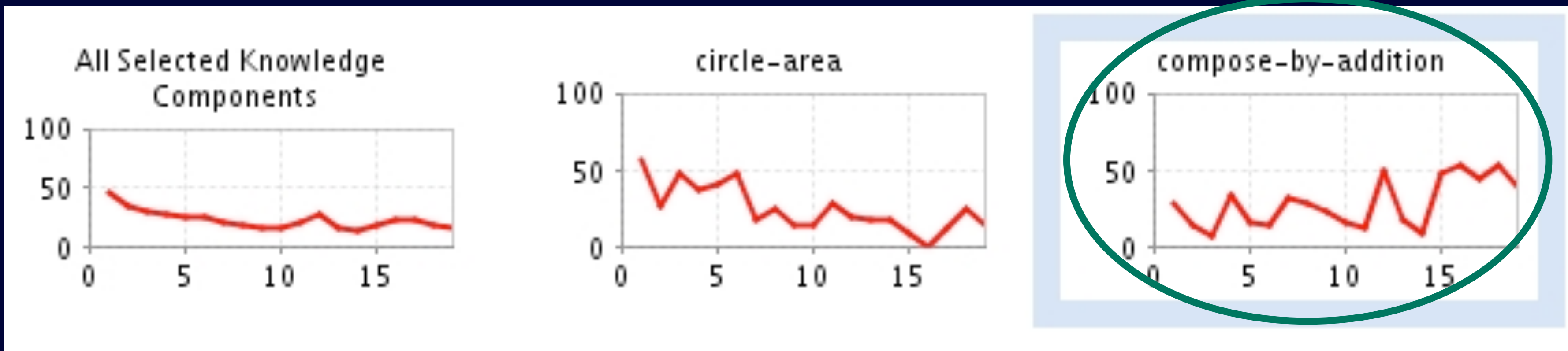
- Averaged across skills for different students – LESS variable



- What causes these variations?

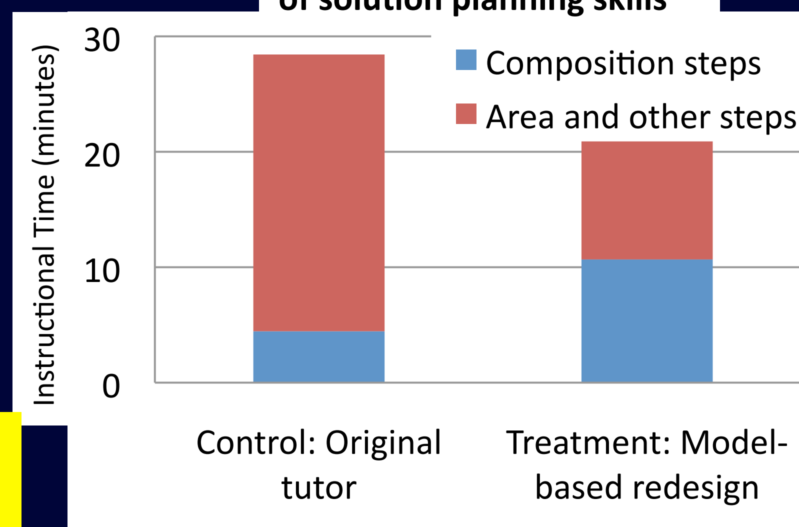
Turning Discovery into Better Learning





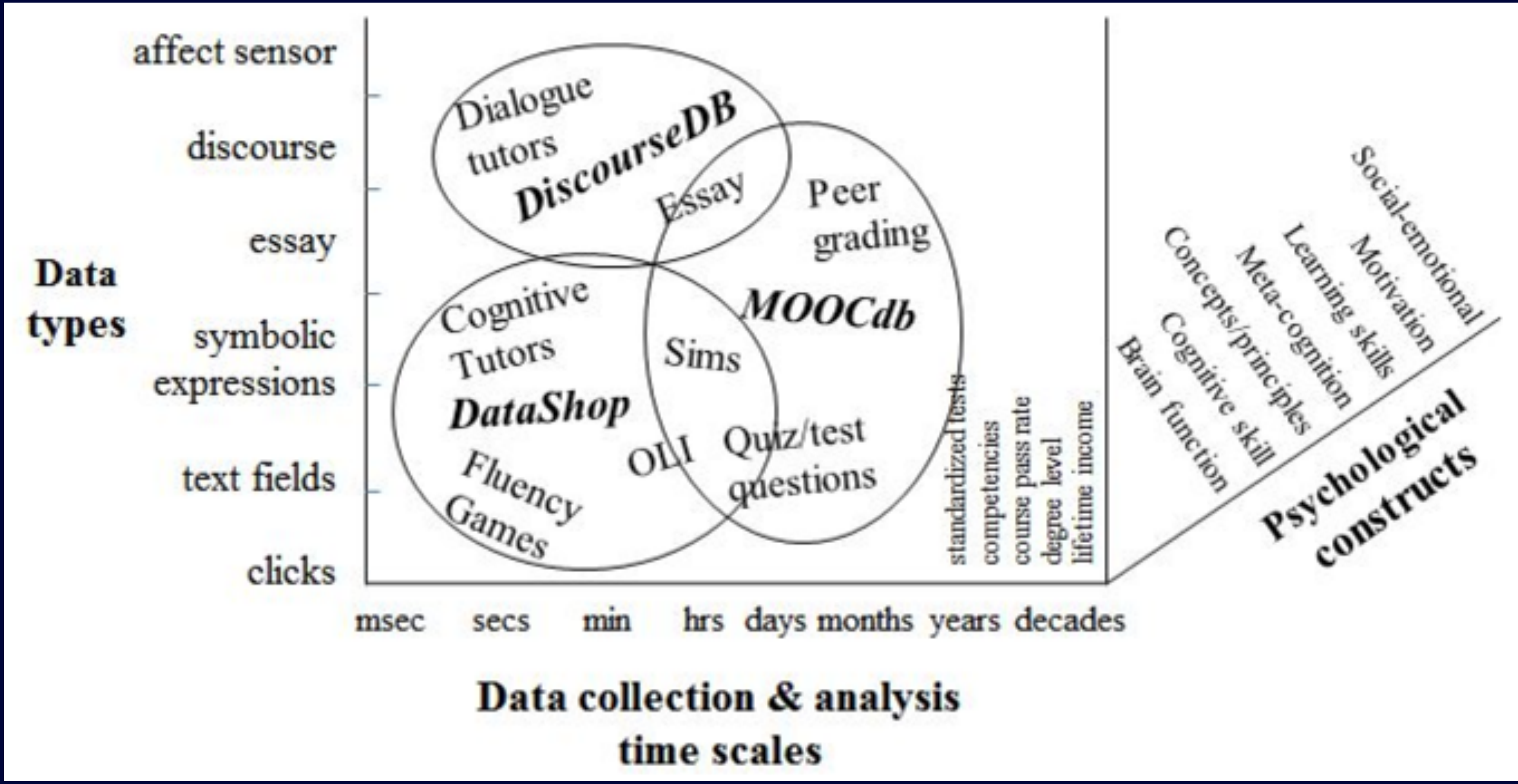
High rough curve
=> hidden skill
=> redesign instruction
=> Experiment
Better student learning!

Better & faster mastery of solution planning skills



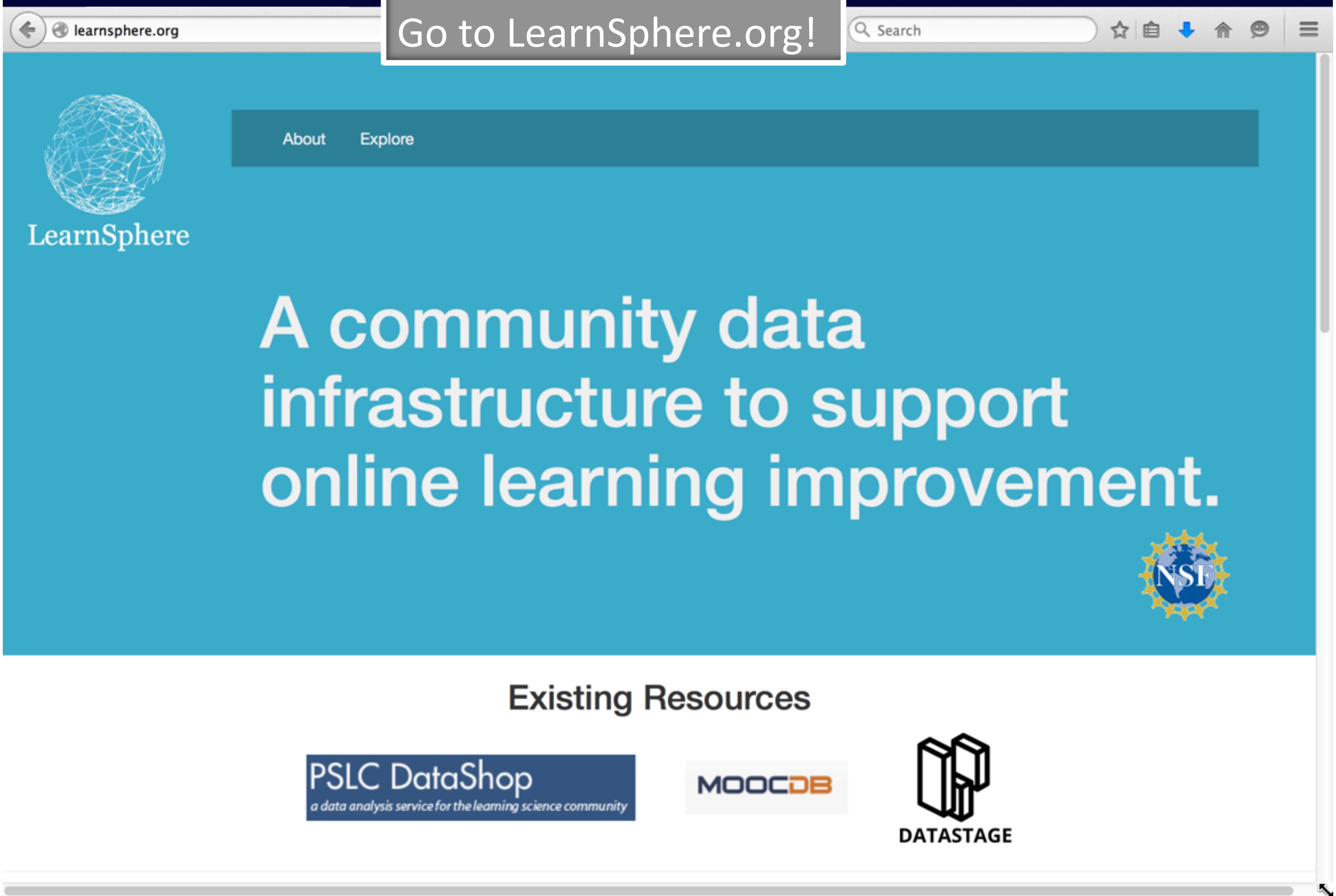
Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education.*

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

Go to LearnSphere.org!

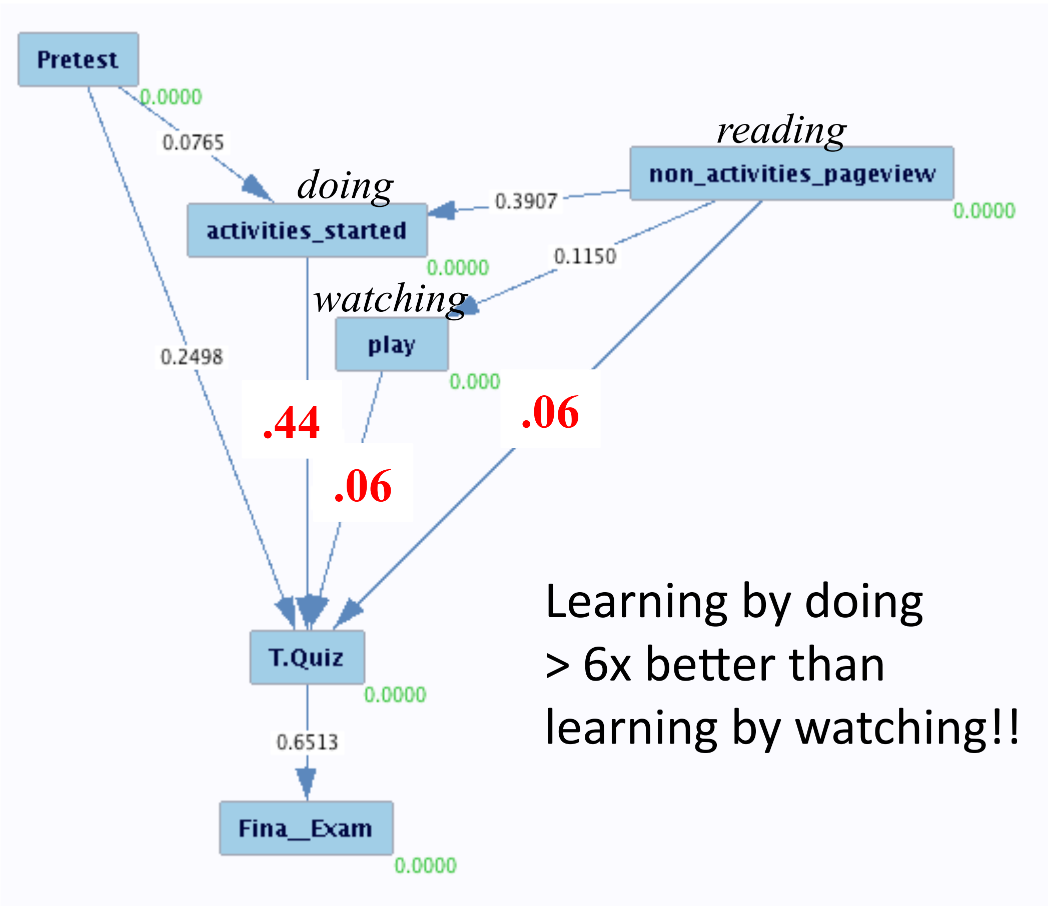


14

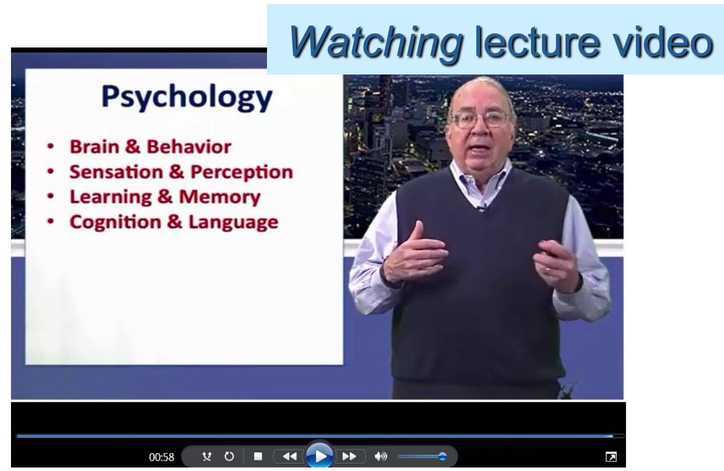
Data Integration Example:

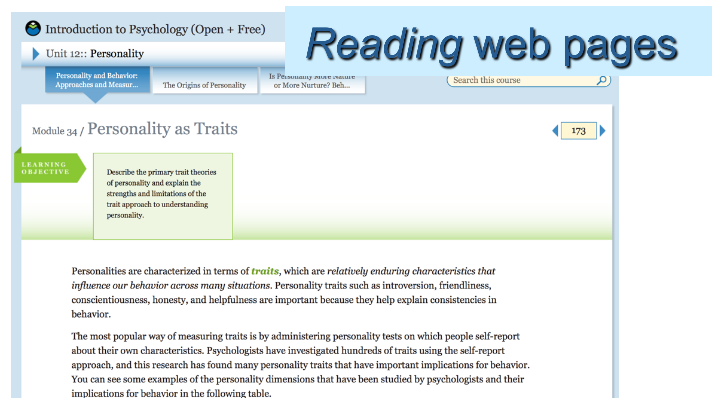
MOOC + OLI = Insight

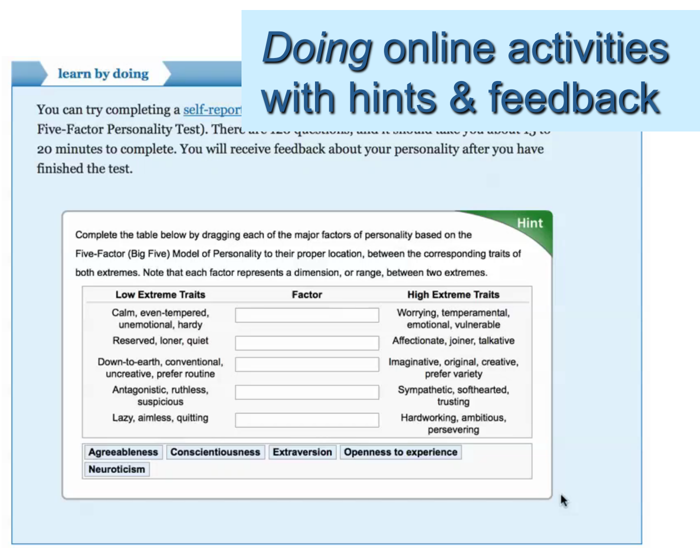
What student choices associate with most learning?



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!

LearnLab
Pittsburgh Science of Learning Center

National Science Foundation
WHERE DISCOVERIES BEGIN

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

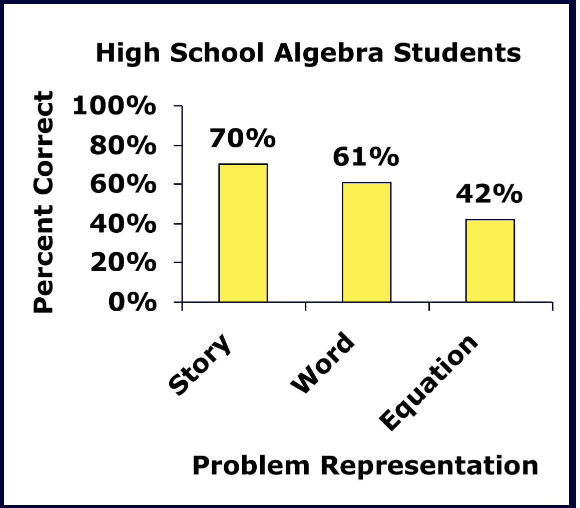
Starting with some number, if I multiply it by 6 and then add 66, I get 81.90. What number did I start with?

Equation

x * 6 + 66 = 81.90

Math educators say:
story or word is hardest

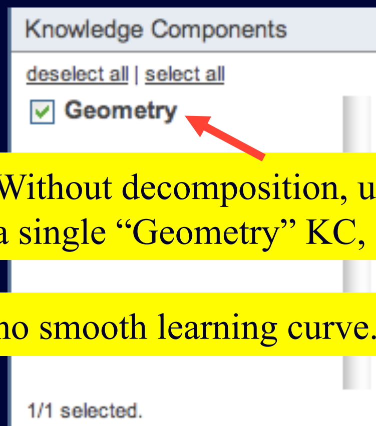
Students:
equations
are hardest

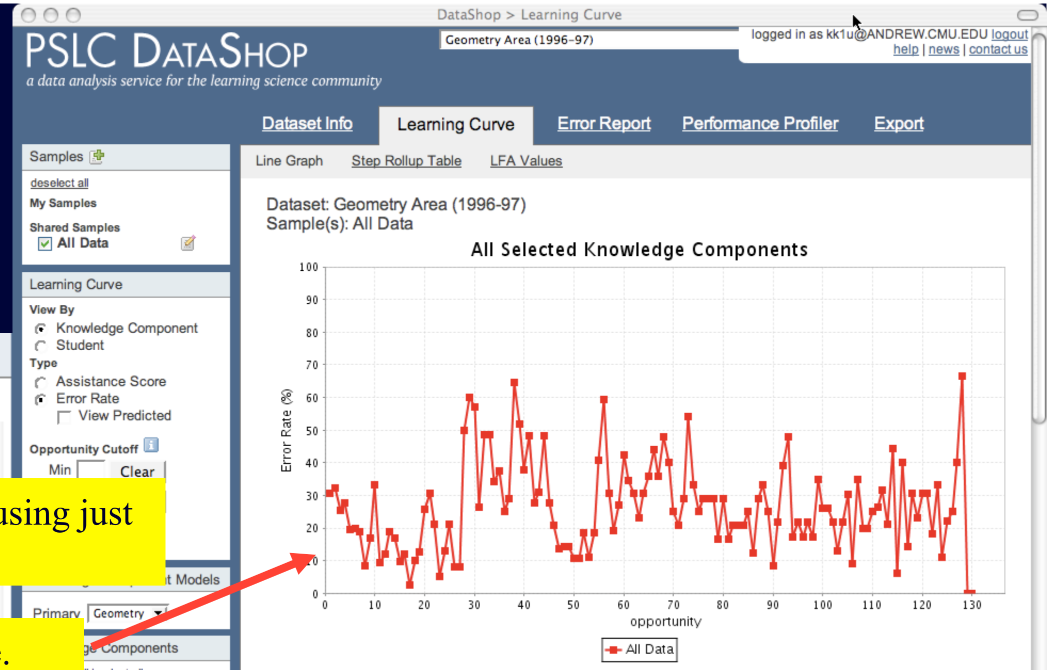


Expert blind spot!
Algebra teachers, especially, incorrectly think equations are easy

Koedinger & Nathan (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Learning Science.*

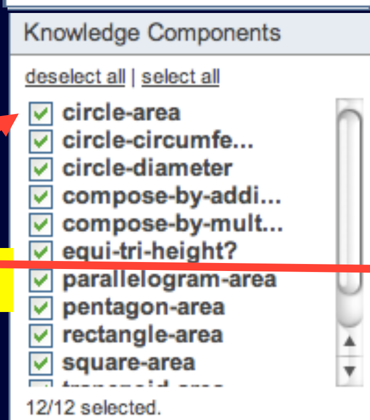
Cognitive Task Analysis using DataShop's learning curve tools

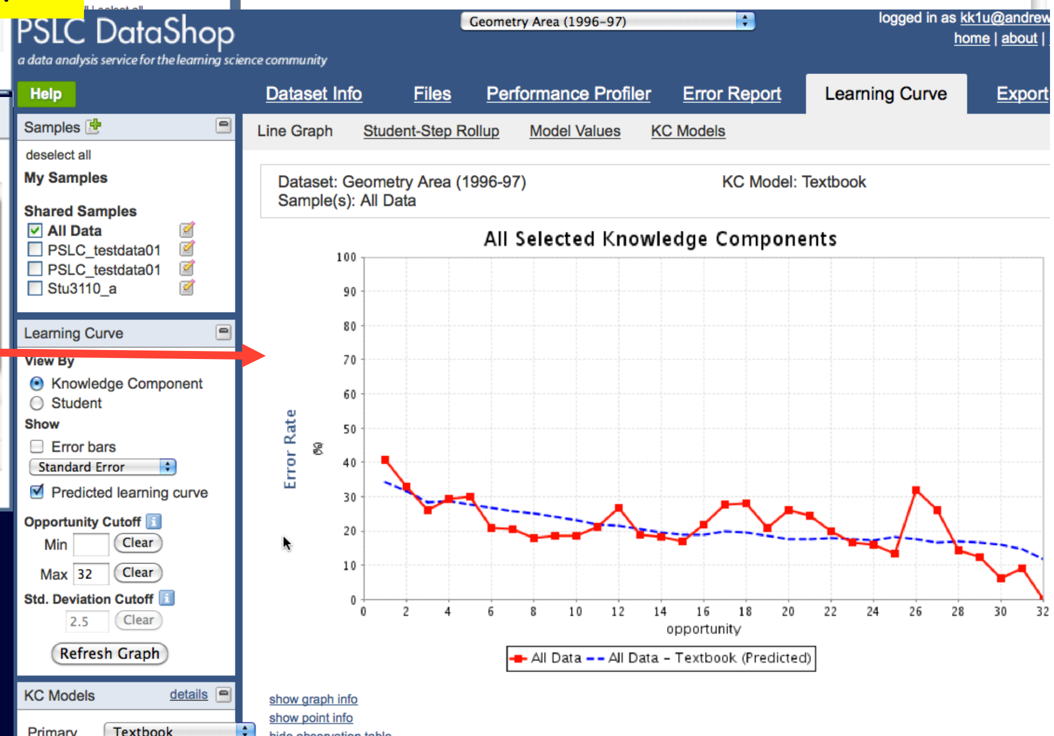




Without decomposition, using just a single "Geometry" KC,

no smooth learning curve.





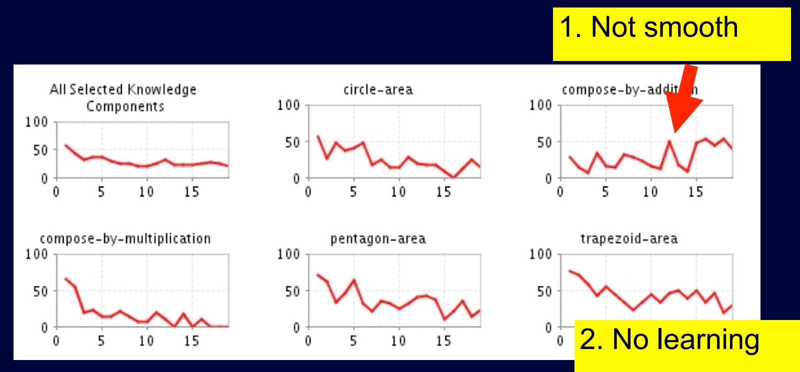
But with decomposition, 12 KCs for area concepts,

a smoother learning curve.

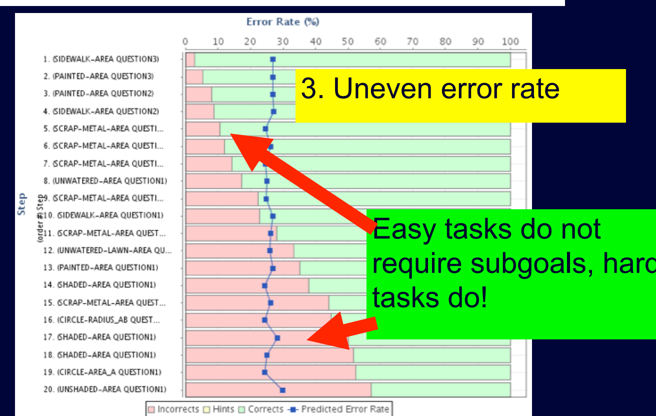
Upshot: Can automate analysis & produce better student models

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

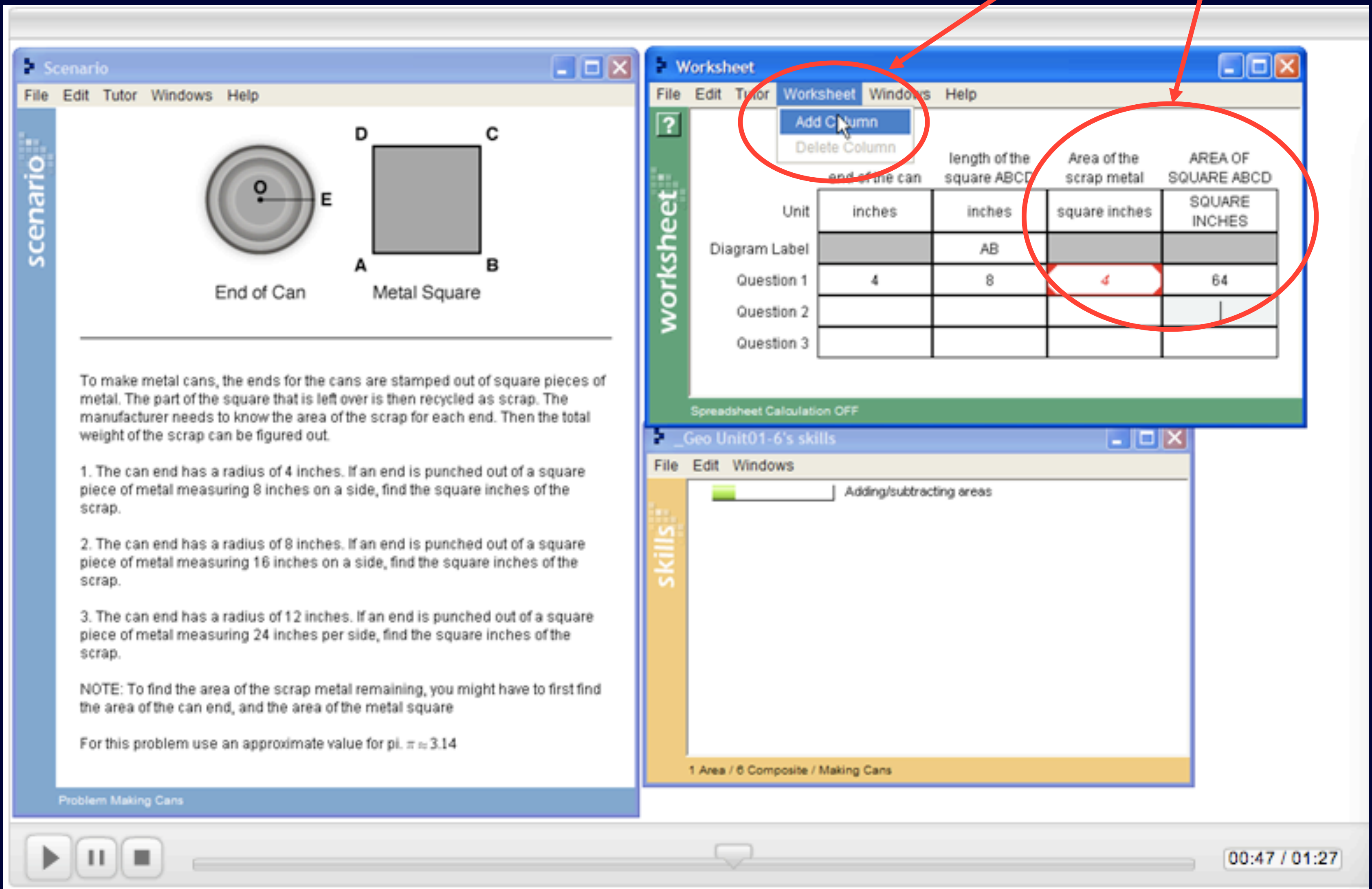


KC Name	Intercept	Slope
circle-area	0.58	0.068
compose-by-addition	0.74	0
compose-by-mult	0.6	0.114
pentagon-area	0.37	0.110
trapezoid-area	0.35	0.091

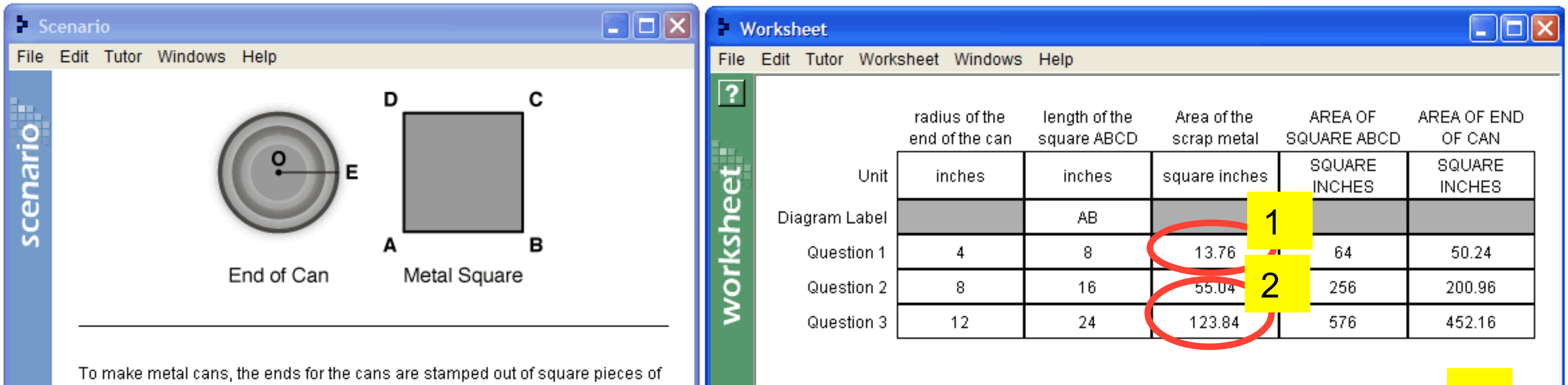


Geometry Tutor
Scaffolding problem decomposition

Problem decomposition support



New model discovery: Split "compose" into 3 skills

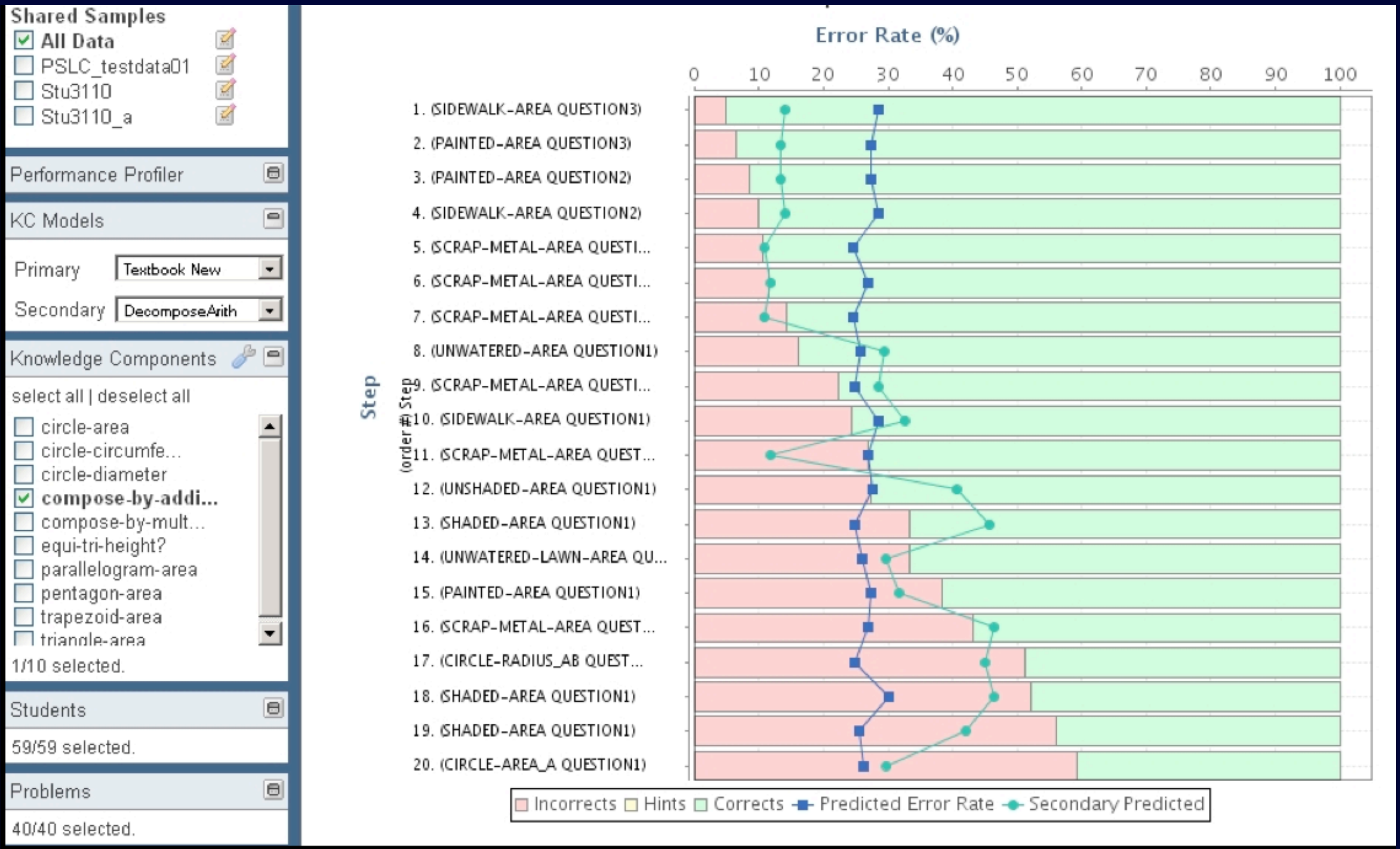


3

When prompts are initially present for component areas

- 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-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

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