The Center for Innovative Research in Cyberlearning

#cyberlearning

CIRCL• **Our purpose**

The **Center for Innovative Research in Cyberlearning** seeks to amplify research-based voices by:

- Nurturing community among projects, investigators and those new to the field
- Addressing common needs
- Planning for the future
- Creating broader impact together

SRI Leads, EDC brings best practices, NORC evaluates

SRI Education

EDC Learning transforms lives.

NORC at the UNIVERSITY of CHICAGO

CIRCL • **Priority Activities**

- **Events**: *annual major meetings, working groups, webinars*
- **Brokering:** *helping connect investigators, projects and newcomers to knowledge and resources*
- **Synthesis and Web Site**: *creating a public space to highlight contributions, share findings, build community and capacity*
- **Portfolio Analysis:** *understanding the funded projects*
- **Sharing Data:** *as needed by NSF and others*
- **Broadening Participation:** *in the cyberlearning CoP to include institutions and individuals currently underrepresented*

CIRCL • **What can CIRCL do for you?** http://circlcenter.org

Perspectives

Learn about researchers, teachers, industry, informal learning and other stakeholders in the cyberlearning community, what drives their work, and what they think the community should be doing.

What's your view on Cyberlearning? Use this quick form to let us know.

administrator | graduate student | industry | informal | postdoc | researcher | teacher | All

Projects

CIRCL Spotlights illuminate some of the different cyberlearning projects across NSF, including projects funded by the NSF Cyberlearning Program and projects funded by other NSF programs whose work has a cyberlearning theme. A tag map of funded projects is also available.

Want us to spotlight your project?
Contact us to contribute your story.

spotlight | All

LEARN ABOUT OUR COMPLEX WORLD THROUGH MAP-BASED GAMES!
CIRCL Spotlights illuminate some of the different projects...

HEAD-MOUNTED DISPLAYS IN DEAF EDUCATION
CIRCL Spotlights illuminate some of the different projects...

REVOLUTIONIZING EDUCATION IN HAITI
CIRCL Spotlights illuminate some of the different projects...

UNDERSTANDING SUSTAINABILITY THROUGH DISCOVERY AND PLAY
CIRCL Spotlights illuminate some of the different projects...

LINKING SUPERHEROES AND TECHNOLOGY TO STEM ASPIRATIONS
CIRCL Spotlights illuminate some of the different projects...

SYNERGISTIC TEACHING OF COMPUTATIONAL THINKING AND SCIENTIFIC MODELING
CIRCL Spotlights illuminate some of the different projects...

MIXED REALITY BRINGS SCIENCE CONCEPTS TO LIFE
CIRCL Spotlights illuminate some of the different projects...

ACTIVITY MONITOR GAME INCREASES YOUTH FITNESS
CIRCL Spotlights illuminate some of the different projects...

Newsletter

Subscribe to the CIRCL newsletter to get updates 6 times a year on cyberlearning-related news.

Subscribe

Have some news to share with the community?
Contact CIRCL

newsletter | All

CIRCL NEWSLETTER – ISSUE 7, SEPTEMBER 2014
CIRCL News Diversity makes our community strong. In this month's newsletter, we...

CIRCL NEWSLETTER – ISSUE 6, JULY 2014
CIRCL News: Summit Recap On June 9 and 10, 2014, more than...

CIRCL NEWSLETTER – ISSUE 5, MAY 2014
CIRCL News On March 26 and 27, 2014, CIRCL hosted an intensive...

CIRCL NEWSLETTER – ISSUE 4, MARCH 2014
CIRCL News "It's just four questions." We're grateful for the many ways...

CIRCL NEWSLETTER – ISSUE 3, JANUARY 2014
CIRCL News Please consider joining us for two important gatherings. First, the...

CIRCL NEWSLETTER – ISSUE 2, NOVEMBER 2013
CIRCL News CIRCL met with leaders across NSF on September 25 and...

CIRCL NEWSLETTER – ISSUE 1, SEPTEMBER 2013
Ninety people attended the first CIRCL Synthesis and Envisioning Meeting on June...

Like | Share | 327

Cyberlearning Center @CIRCLCenter 18h
Apply to attend our the next #cyberlearning meeting. See ow.ly/CILnq

Shuchi Grover @shuchig 19h
2nd Call for Proposals CSCL 2015 @ Gothenburg, Sweden, Jun 7-11 2015 (Paper deadline extended to Nov 17) isls.org/cscl2015/index... @CIRCLCenter
Retweeted by Cyberlearning Center
Expand

Jianwei Zhang @JianweiZhang 4 Oct
@CIRCLCenter

Tweet to @CIRCLCenter

Follow Us!

Resources

Browse CIRCL Synthesis statements, watch NAPLeS webinars, search the digital collection of education resources from Informal Commons, and see other resources below.

Have resources to suggest?
Contact CIRCL

Synthesis Statements

CIRCL synthesis statements summarize effective use of advanced learning technologies that are integrative, innovative, empirically grounded, and widely useful. Want to contribute? Let us know.

synthesis | All

AI APPLICATIONS IN EDUCATION
Contributors: Chad Lane, Shuchi Grover, and Jeremy Roschelle

EDUCATIONAL DATA MINING AND LEARNING ANALYTICS
Contributors: Mimi Recker, Andrew Krumm, Mingyu Feng, Shuchi Grover

LEARNING SCIENCES
Contributors: Jeremy Roschelle, Shuchi Grover

DESIGN-BASED IMPLEMENTATION RESEARCH
Contributors: Barry Fishman, Britte Cheng, William Penuel

Events

Learn about upcoming CIRCL events like the 2015 Synthesis & Envisioning meeting, and access archives from past events, including the 2014 Cyberlearning Summit and the 2012 Cyberlearning Summit.

Browse our calendar of other cyberlearning-related conferences and events.
Please let us know about other cyberlearning events in the community.

event | All

CYBERLEARNING 2015: CONNECT, COLLABORATE, AND CREATE THE FUTURE
January 27-28, 2015 in Arlington, VA A gathering of participants with a...

CYBERLEARNING SUMMIT 2014
On June 9-10, 2014, CIRCL hosted the 2014 Cyberlearning Summit at the...

NSF CYBERLEARNING INTEGRATION (INT) PROPOSAL WEBINAR
Monday, June 2nd from 3pm – 4pm ET An informational webinar on...

PARTNERING FOR IMPACT 2014
On March 26 and 27, 2014, SRI hosted an intensive two-day workshop...

NSF CYBERLEARNING SOLICITATION INFORMATION WEBINAR
Tuesday, February 18th from 1pm – 2:30pm ET An informational webinar on...

SYNTHESIS AND ENVISIONING 2013
A gathering of NSF-funded cyberlearning projects to synthesize what is known and...

CYBERLEARNING SUMMIT 2012
The 2012 Cyberlearning Research Summit was a high-profile gathering in Washington DC...

CIRCL • **Connect, collaborate, create**

Perspectives

Learn about researchers, teachers, industry, informal learning and other stakeholders in the cyberlearning community, what drives their work, and what they think the community should be doing.

What's your view on Cyberlearning? Use this quick form to let us know.

administrator | graduate student | industry | informal | postdoc | researcher | teacher | All

CIRCL • **Identify synergistic projects**

Projects

CIRCL Spotlights illuminate some of the different cyberlearning projects across NSF, including projects funded by the NSF Cyberlearning Program and projects funded by other NSF programs whose work has a cyberlearning theme. A tag map of funded projects is also available.

Want us to spotlight your project?
Contact us to contribute your story.

spotlight | All

LEARN ABOUT OUR COMPLEX WORLD THROUGH MAP-BASED GAMES!
CIRCL Spotlights illuminate some of the different projects...

HEAD-MOUNTED DISPLAYS IN DEAF EDUCATION
CIRCL Spotlights illuminate some of the different projects...

REVOLUTIONIZING EDUCATION IN HAITI
CIRCL Spotlights illuminate some of the different projects...

UNDERSTANDING SUSTAINABILITY THROUGH DISCOVERY AND PLAY
CIRCL Spotlights illuminate some of the different projects...

LINKING SUPERHEROES AND TECHNOLOGY TO STEM ASPIRATIONS
CIRCL Spotlights illuminate some of the different projects...

SYNERGISTIC TEACHING OF COMPUTATIONAL THINKING AND SCIENTIFIC MODELING
CIRCL Spotlights illuminate some of the different projects...

MIXED REALITY BRINGS SCIENCE CONCEPTS TO LIFE
CIRCL Spotlights illuminate some of the different projects...

ACTIVITY MONITOR GAME INCREASES YOUTH FITNESS
CIRCL Spotlights illuminate some of the different projects...

CIRCL • **Access integrative, empirically grounded resources**

Big Ideas

Read CIRCL synthesis statements, review resources for writing strong proposals to the NSF Cyberlearning Program, watch NAPLeS webinars, subscribe to edSurge, read the Cyberlearning Educators blog, search the digital collection of resources from Informal Commons, browse NSF project and program data in DIA2, and more.

Have ideas or resources to suggest?
Contact CIRCL

Synthesis Statements

CIRCL synthesis statements summarize effective use of advanced learning technologies that are integrative, innovative, empirically grounded, and widely useful. Want to contribute? Let us know.

THE CUTTING-EDGE OF INFORMAL LEARNING: MAKERS, MOBILE, AND MORE!
Cyberlearning spans in-school and out-of-school learning -- and these days, a lot of meaningful learning is...

GAMES AND VIRTUAL WORLDS
Computer-based games and virtual worlds provide opportunities for players to think about choices, take action, and...

PARTNERING FOR IMPACT: INCREASING CYBERLEARNING'S INFLUENCE IN EDUCATION MARKETS
Many Cyberlearning researchers know that their work could make

TECHNOLOGY ENABLED FORMATIVE ASSESSMENT
Formative assessment occurs when teachers check student understanding and guide decision making to improve learning.

COLLABORATIVE LEARNING
Learning to explain, justify, critique, etc. are essential skills for today's citizens, for scientists, and in...

EDUCATIONAL DATA MINING AND LEARNING ANALYTICS
EDM is the use of multiple analytical techniques to better understand relationships, structure,

AI APPLICATIONS IN EDUCATION
AI techniques can enable educational technologies to better track, adapt to, and support individual learners.

LEARNING SCIENCES
The Learning Sciences is a field of scientific research that developed in the 1980s, from influences...

CIRCL • **Join a vibrant community of practice**

Events

Learn about upcoming CIRCL events like Cyberlearning 2015, and access archives from past events.

Browse a calendar of other cyberlearning-related conferences and events.
Please let us know about other cyberlearning events in the community.

event | All

CYBERLEARNING 2015: CONNECT, COLLABORATE, AND CREATE THE FUTURE

January 27-28, 2015 in Arlington, VA A gathering of participants with a...

CYBERLEARNING SUMMIT 2014

On June 9-10, 2014, CIRCL hosted the 2014 Cyberlearning Summit at the...

NSF CYBERLEARNING INTEGRATION (INT) PROPOSAL WEBINAR

Monday, June 2nd from 3pm – 4pm ET An informational webinar on...

PARTNERING FOR IMPACT 2014

On March 26 and 27, 2014, SRI hosted an intensive two-day workshop...

NSF CYBERLEARNING SOLICITATION INFORMATION WEBINAR

Tuesday, February 18th from 1pm – 2:30pm ET An informational webinar on...

SYNTHESIS AND ENVISIONING 2013

A gathering of NSF-funded cyberlearning projects to synthesize what is known and...

CYBERLEARNING SUMMIT 2012

The 2012 Cyberlearning Research Summit was a high-profile gathering in Washington DC...

CIRCL • **Follow us, contribute, stay connected!**

Like Share 327

Cyberlearning Center @CIRCLCenter 18h
Apply to attend our the next #cyberlearning meeting. See ow.ly/CILnq

Shuchi Grover @shuchig 19h
2nd Call for Proposals CSCL 2015 @ Gothenburg, Sweden, Jun 7-11 2015 (Paper deadline extended to Nov 17) isls.org/cscl2015/index... @CIRCLCenter
Retweeted by Cyberlearning Center
Expand

Jianwei Zhang @JianweiZhang 4 Oct
@CIRCLCenter

Tweet to @CIRCLCenter

Follow Us!

Newsletter

Subscribe to the CIRCL newsletter to get updates 6 times a year on cyberlearning-related news.

Subscribe

Have some news to share with the community?
Contact CIRCL

newsletter All

CIRCL NEWSLETTER – ISSUE 7, SEPTEMBER 2014
CIRCL News Diversity makes our community strong. In this month's newsletter, we...

CIRCL NEWSLETTER – ISSUE 6, JULY 2014
CIRCL News: Summit Recap On June 9 and 10, 2014, more than...

CIRCL NEWSLETTER – ISSUE 5, MAY 2014
CIRCL News On March 26 and 27, 2014, CIRCL hosted an intensive...

CIRCL NEWSLETTER – ISSUE 4, MARCH 2014
CIRCL News "It's just four questions." We're grateful for the many ways...

CIRCL NEWSLETTER – ISSUE 3, JANUARY 2014
CIRCL News Please consider joining us for two important gatherings. First, the...

CIRCL NEWSLETTER – ISSUE 2, NOVEMBER 2013
CIRCL News CIRCL met with leaders across NSF on September 25 and...

CIRCL NEWSLETTER – ISSUE 1, SEPTEMBER 2013
Ninety people attended the first CIRCL Synthesis and Envisioning Meeting on June...

http://circlcenter.org

circl-info@sri.com

Cyberlearning Data Sharing and Privacy

Ken Koedinger

Professor of Human-Computer Interaction & Psychology

Carnegie Mellon University

Director of LearnLab

Pittsburgh Science of Learning Center

LearnLab Researchers: Charles Perfetti (Psych), Vincent Aleven (HCI), Geoff Gordon (Machine Learning), David Klahr (Psych), Tim Nokes-Malach (Psy), Lauren Resnick (Psy/Ed), Carolyn Rose (Language Tech) + some 200 others!

CIRCL Webinar: May 20, 2015

Outline

- Why data sharing?
- Data curation & privacy management
  - LearnSphere: DataShop, MOOCdb, DataStage, DiscourseDB
- Future of Cyberlearning data partnerships

Big Data for Cyberlearning

More important than "big"

- Collected as part of *natural* activities
- Affords experimentation, "A/B testing"

Big Data for Cyberlearning

More important than "big"

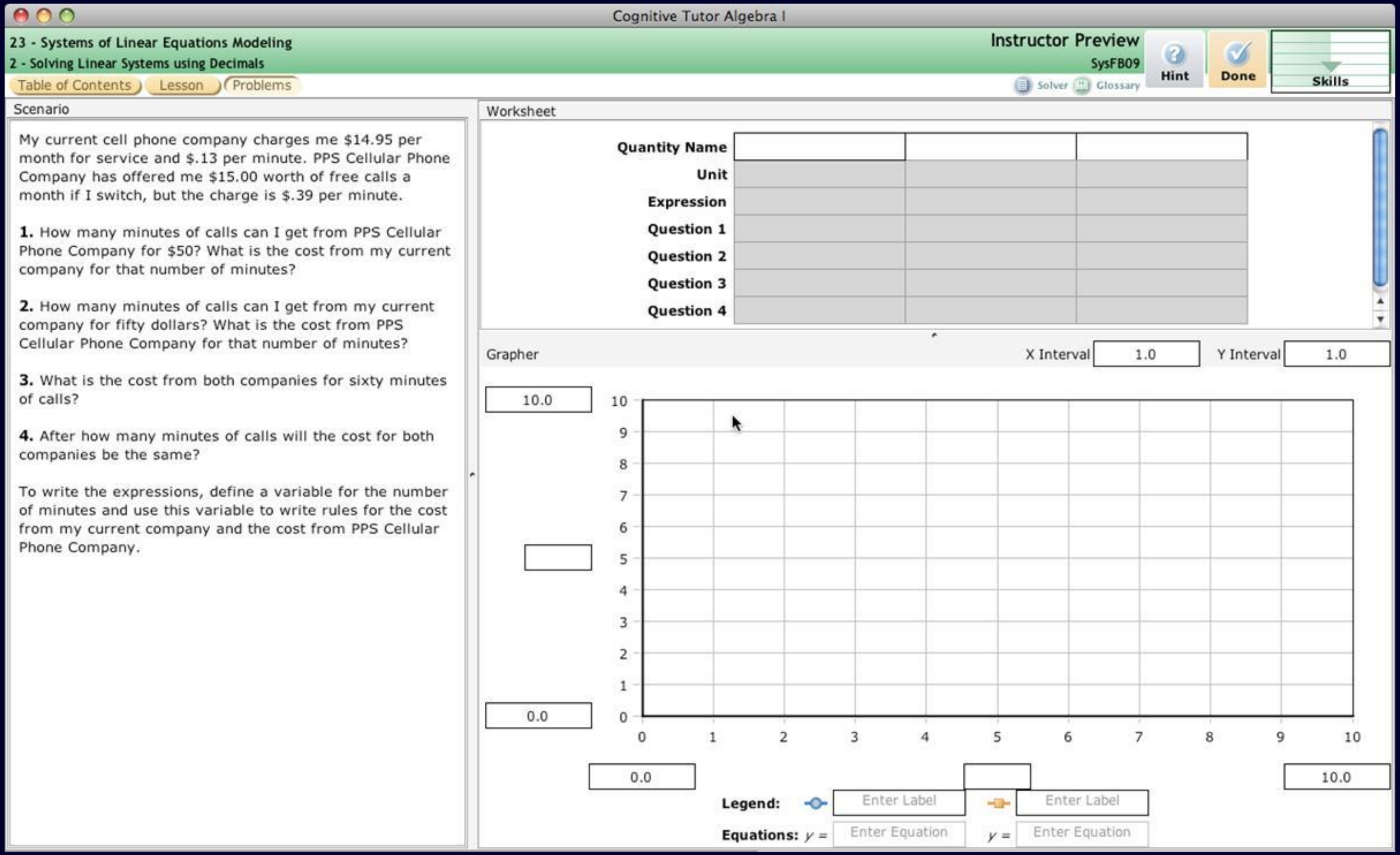
- Collected as part of *natural* activities
- Affords experimentation, "A/B testing"

Many dimensions of "big"

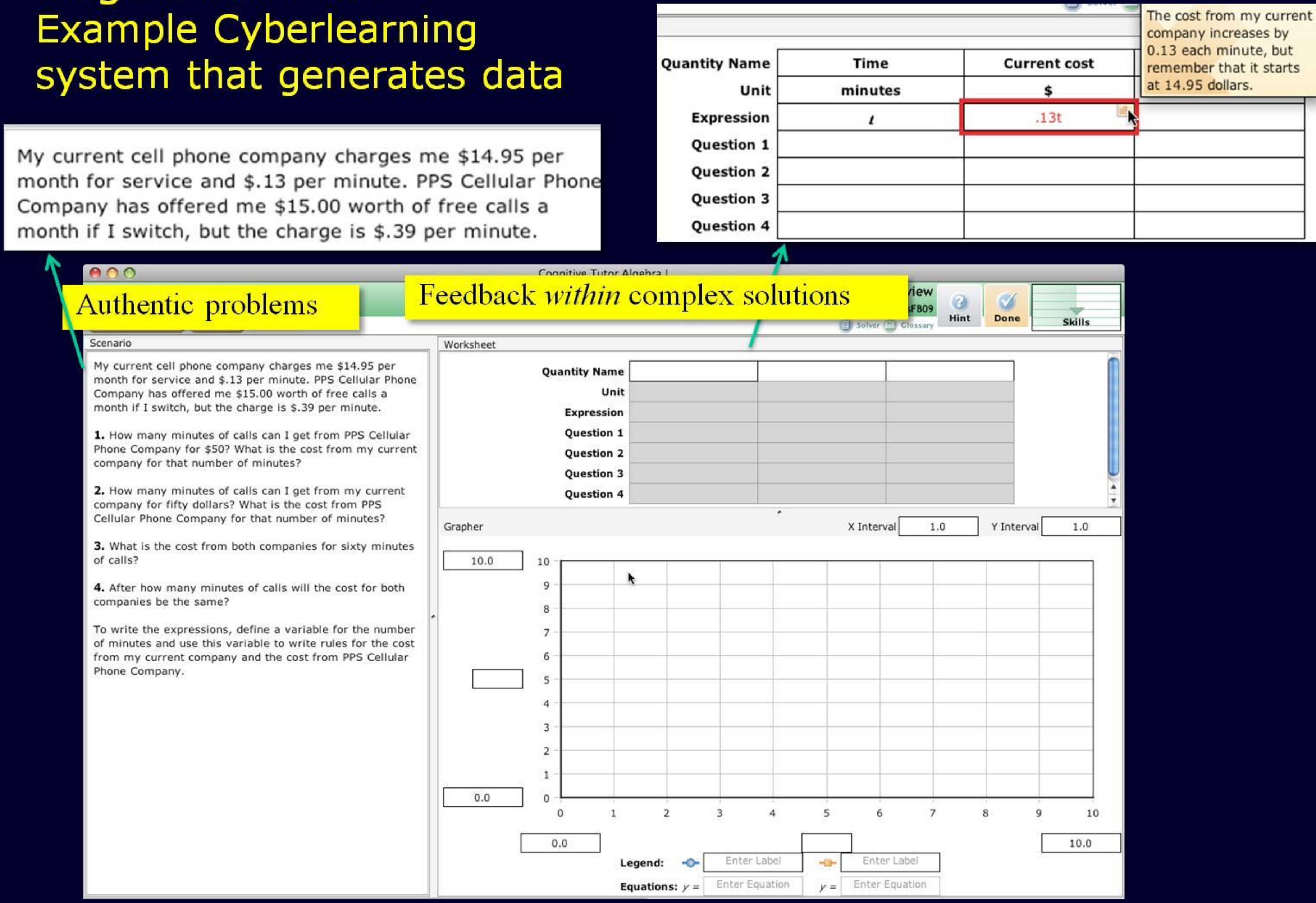
- *Tall* in number of participants (students)
- *Wide* in observations per participant (student)
- *Fine* in frequency of observation
- *Long* in spanning months or years
- *Deep* in theory-relevant variables

Privacy protection is a distinct challenge

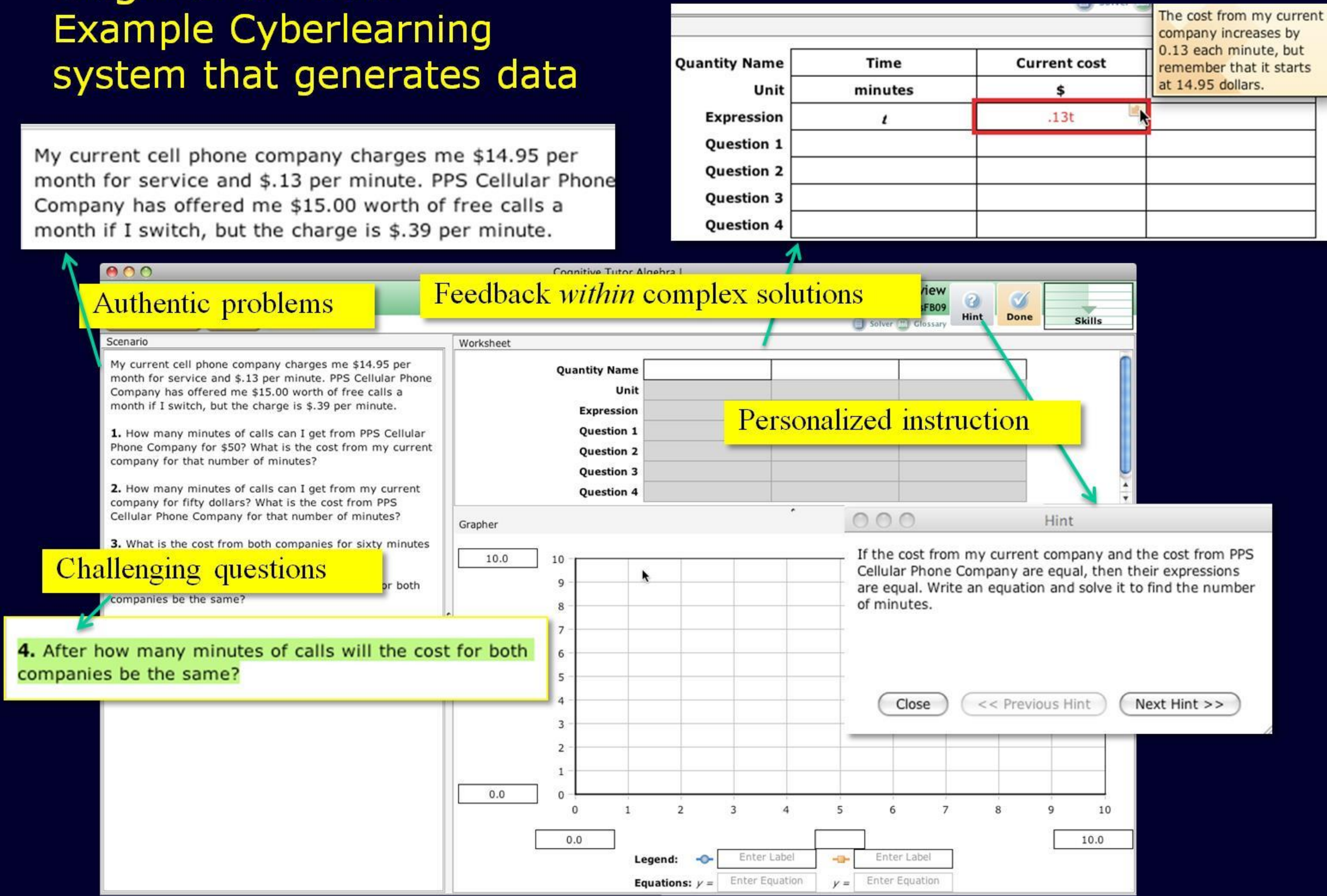
Cognitive Tutor
Example Cyberlearning system that generates data



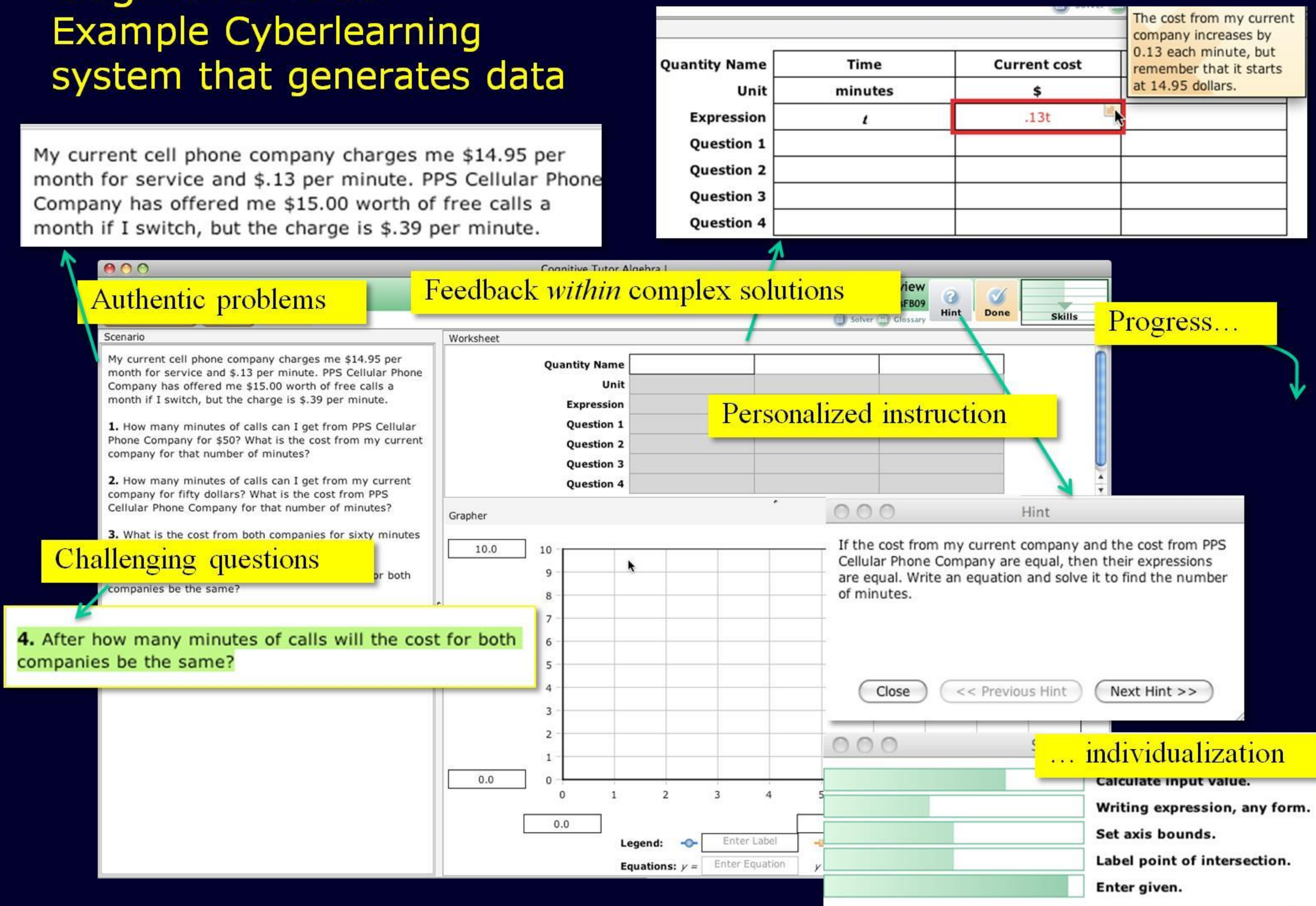
Cognitive Tutor
Example Cyberlearning system that generates data



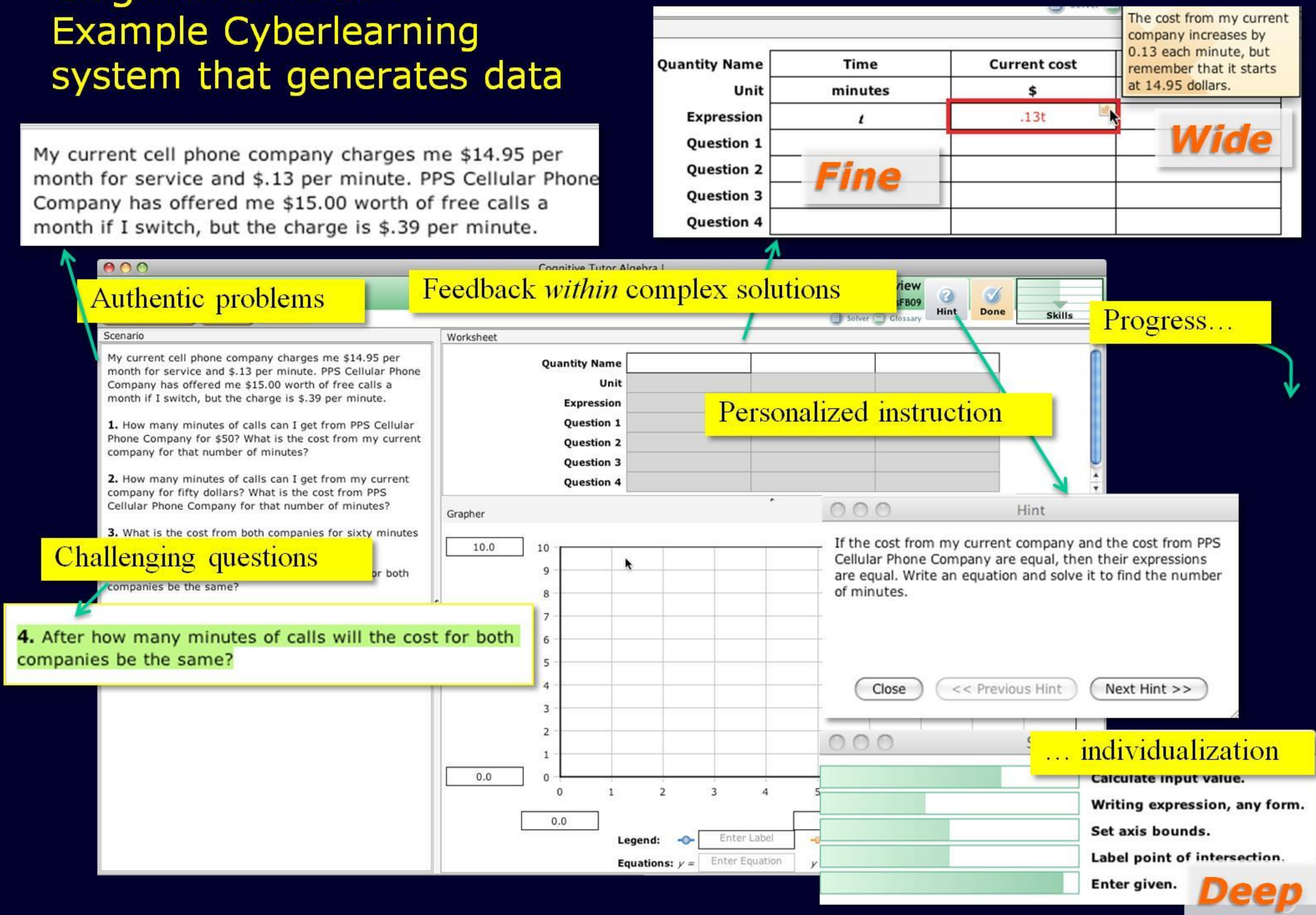
Cognitive Tutor
Example Cyberlearning system that generates data



Cognitive Tutor
Example Cyberlearning system that generates data



Cognitive Tutor
Example Cyberlearning system that generates data



Cognitive model
=> adaptive support of learning by doing

- Cognitive Model: A system that can solve problems in the various ways students can

$$3(2x - 5) = 9$$

If goal is solve a(bx+c) = d
Then rewrite as abx + ac = d

If goal is solve a(bx+c) = d
Then rewrite as bx+c = d/a

If goal is solve a(bx+c) = d
Then rewrite as abx + c = d

$$6x - 15 = 9 \qquad 2x - 5 = 3 \qquad 6x - 5 = 9$$

Cognitive model
=> adaptive support of learning by doing

- Cognitive Model: A system that can solve problems in the various ways students can

$$3(2x - 5) = 9$$

If goal is solve a(bx+c) = d
Then rewrite as abx + ac = d

If goal is solve a(bx+c) = d
Then rewrite as bx+c = d/a

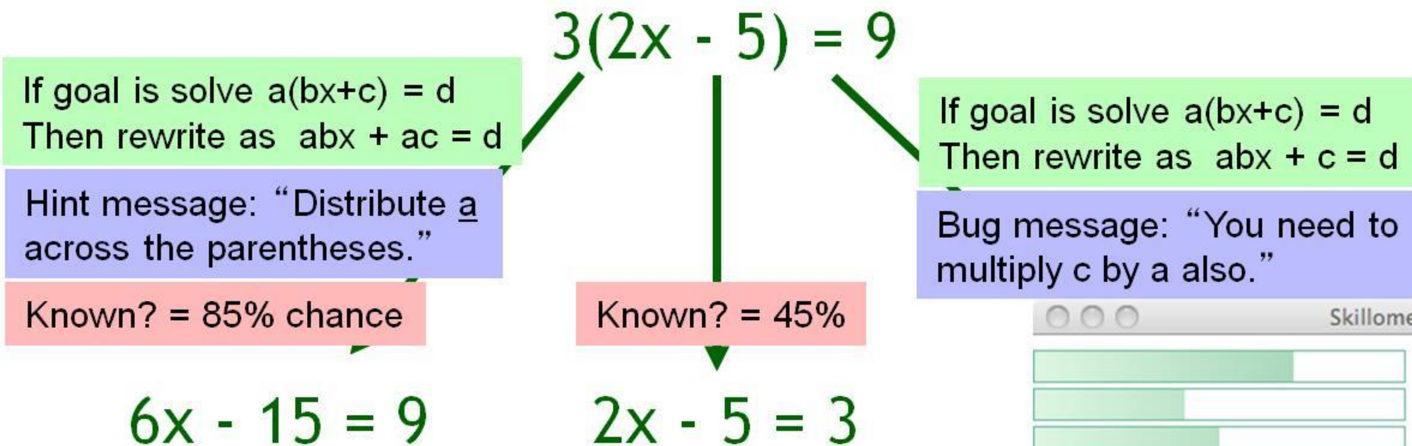
If goal is solve a(bx+c) = d
Then rewrite as abx + c = d

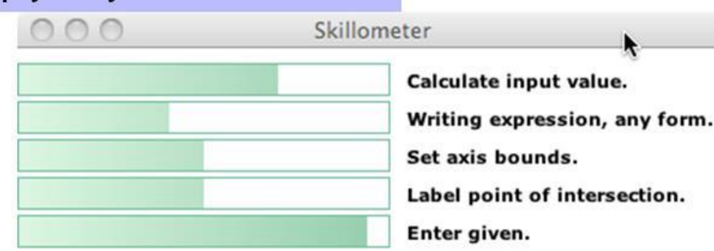
$$6x - 15 = 9 \qquad 2x - 5 = 3 \qquad 6x - 5 = 9$$

- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction

Cognitive model
=> adaptive support of learning by doing

- Cognitive Model: A system that can solve problems in the various ways students can

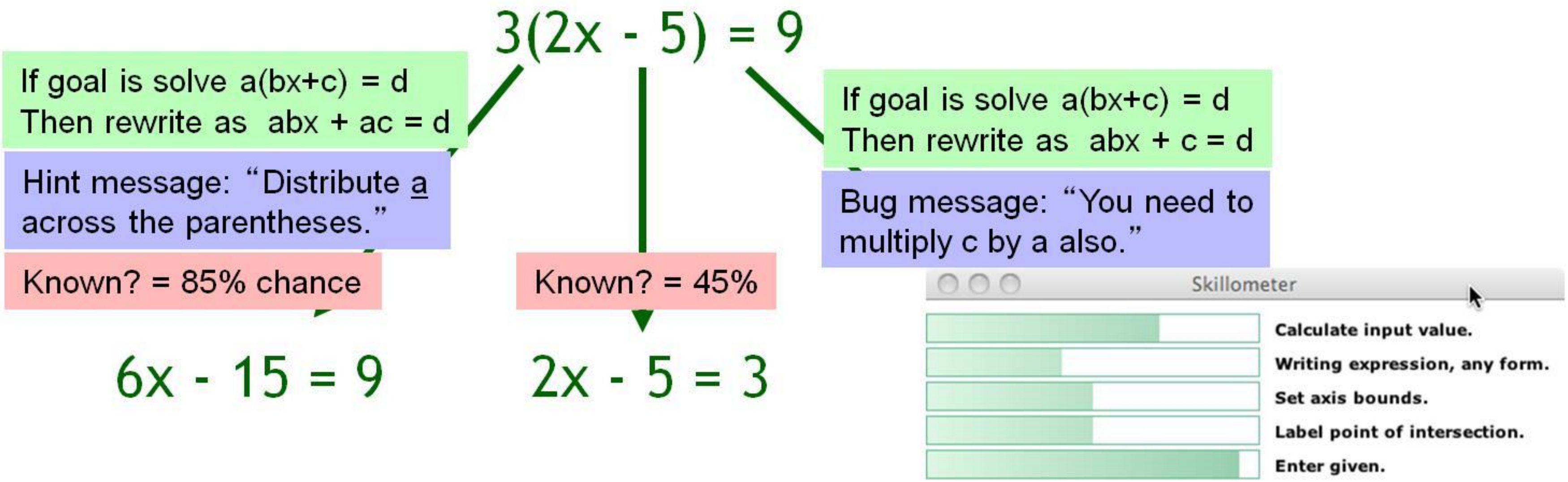




- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction
- Knowledge Tracing: Assesses student's knowledge growth -> individualized activity selection and pacing

Cognitive model
=> adaptive support of learning by doing
=> *deep* theory-coded data stream

- Cognitive Model: A system that can solve problems in the various ways students can



- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction
- Knowledge Tracing: Assesses student's knowledge growth -> individualized activity selection and pacing

Use data to develop models of learners

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?

Answer question online!

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

$x * 6 + 66 = 81.90$

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

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

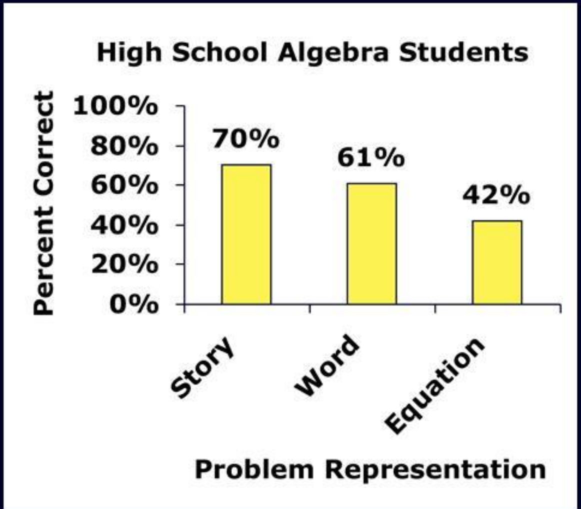
Starting with some number, if I multiply it by 6 and then add 66, I get 81.90. What number did I start with?

Equation

x * 6 + 66 = 81.90

Math educators say:
story or word is hardest

Students:
equations
are hardest



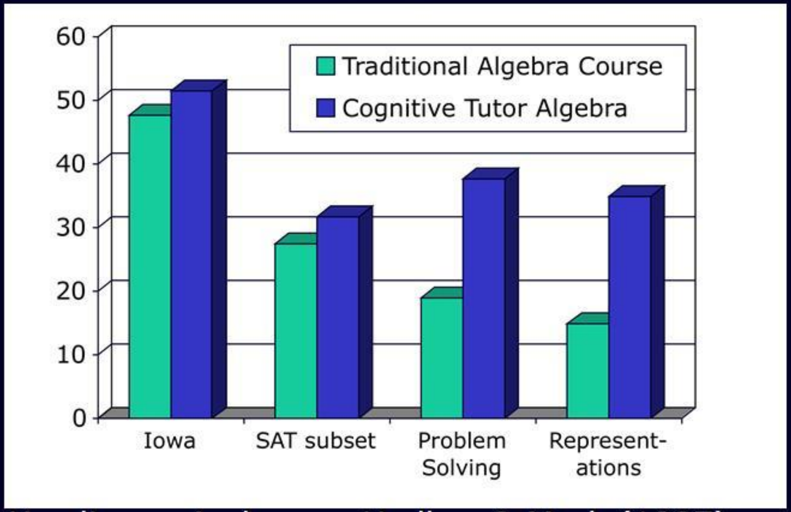
Expert blind spot!
Algebra teachers, especially, incorrectly think equations are easy

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

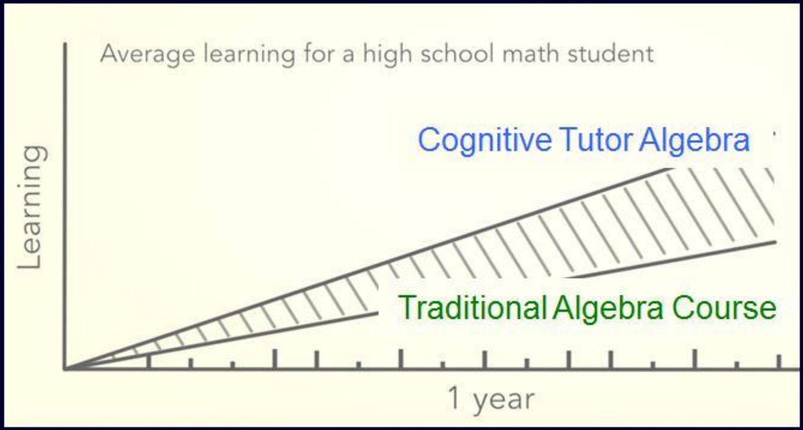
Cyberlearning in use

Algebra Cognitive Tutor

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



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

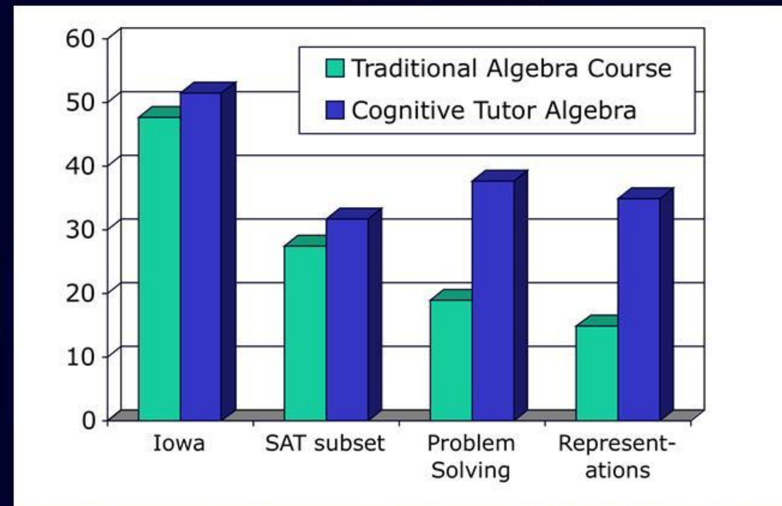


Pane et al. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. Santa Monica, CA: RAND Corp.

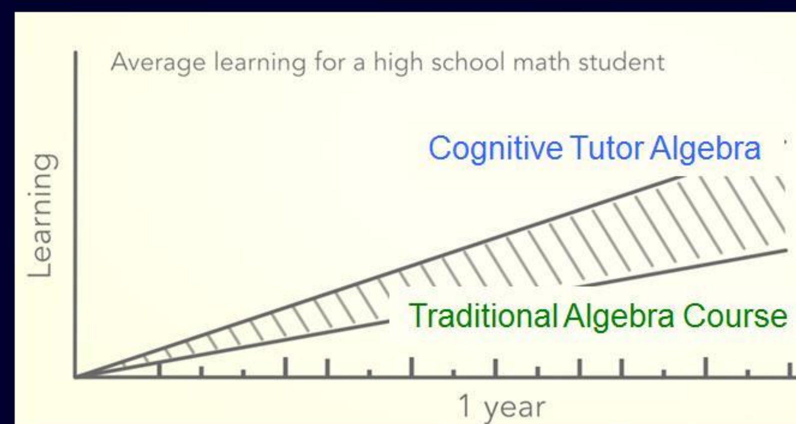
Cyberlearning in use: Good, but can do much better

Algebra Cognitive Tutor

- Widespread intensive use
  ~600K students per year **Tall**
  ~80 minutes per week **Long**
- *Many* field trials => Student learning is 2x better
- Still:
  Could do better
  Too many decisions driven by intuition



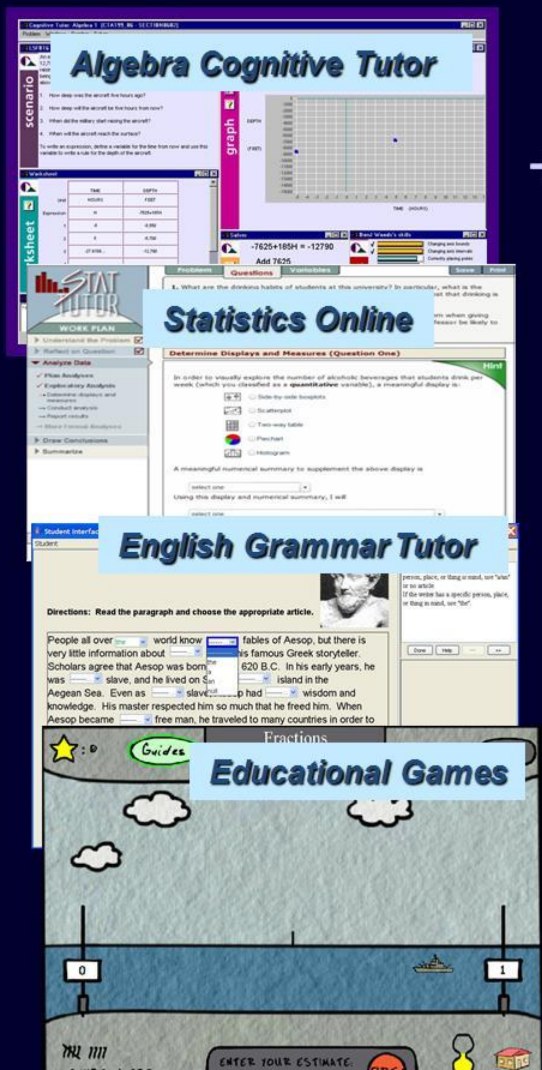
Koedinger, Anderson, Hadley, & Mark (1997). Intelligent tutoring goes to school in the big city.



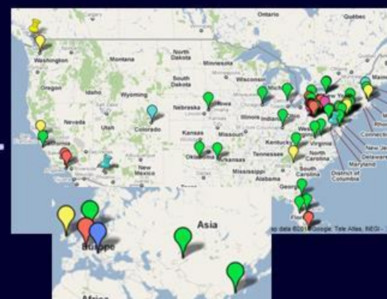
Pane et al. (2013). Effectiveness of Cognitive Tutor Algebra I at Scale. Santa Monica, CA: RAND Corp.

LearnLab: Use fielded Cyberlearning systems to advance learning theory

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



+



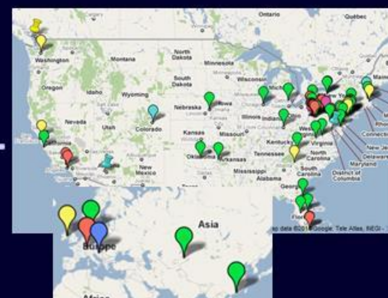
=

LearnLab
Pittsburgh Science of Learning Center

LearnLab: Use fielded Cyberlearning systems to advance learning theory

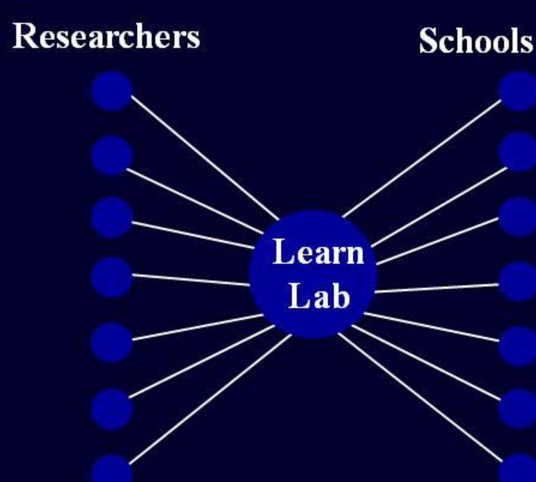
Ed tech + wide use = "Basic research *at scale*"

+



=

LearnLab
Pittsburgh Science of Learning Center

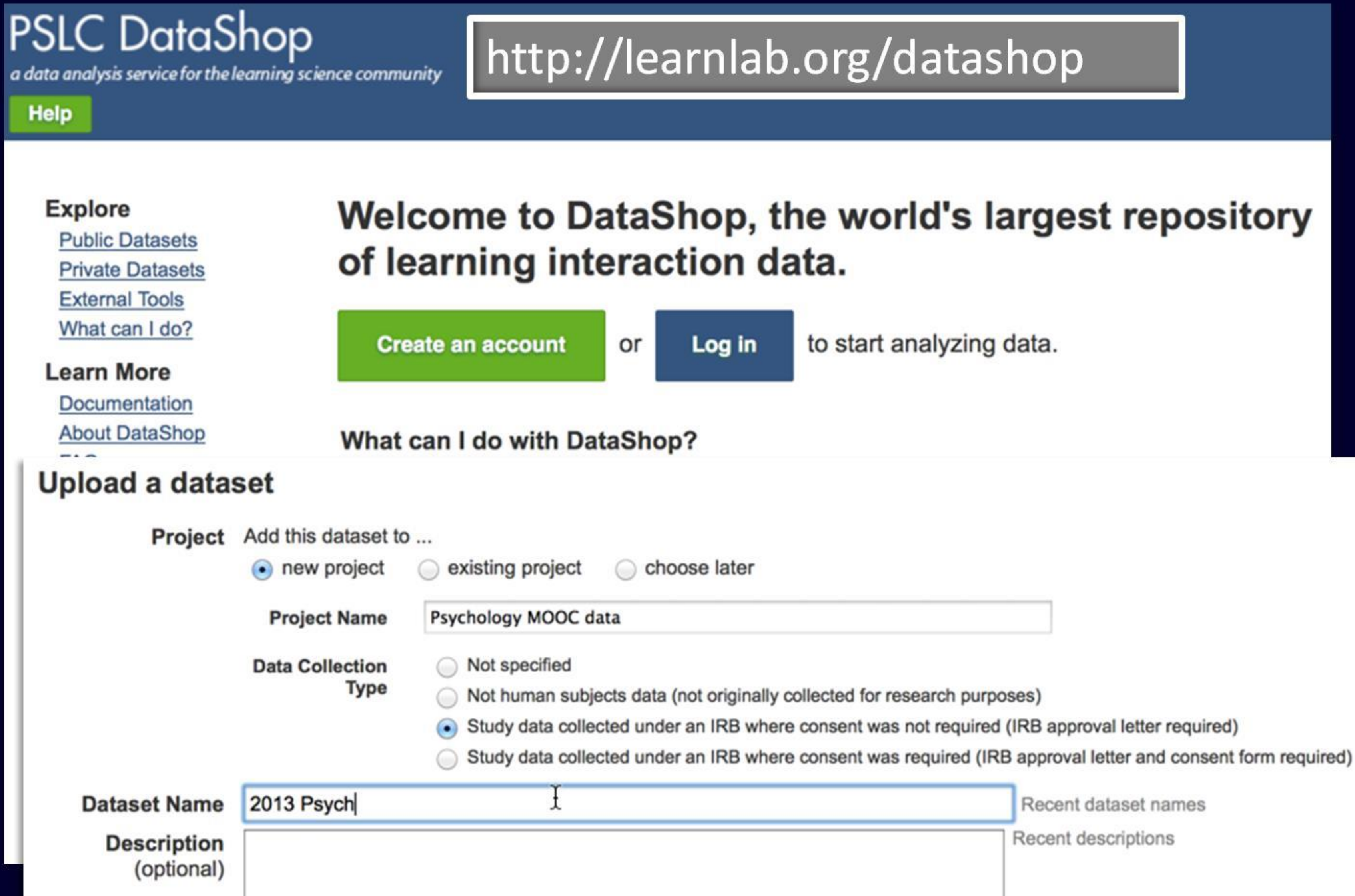


Since 2004
- 680 ed tech data sets in DataShop
- 360 *in vivo* experiments

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

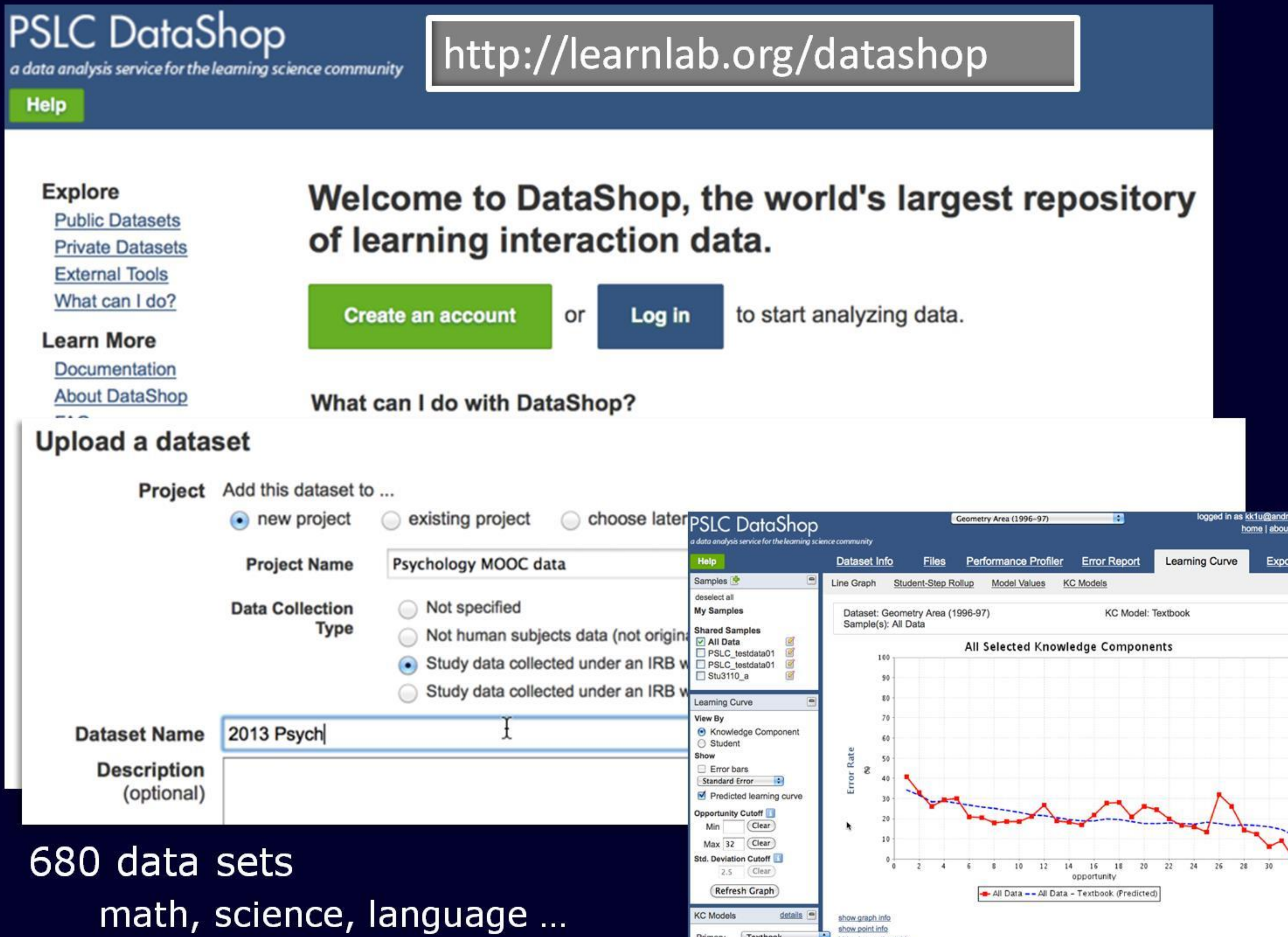
680 data sets

- math, science, language …
- K12 & college



680 data sets

- math, science, language …
- K12 & college



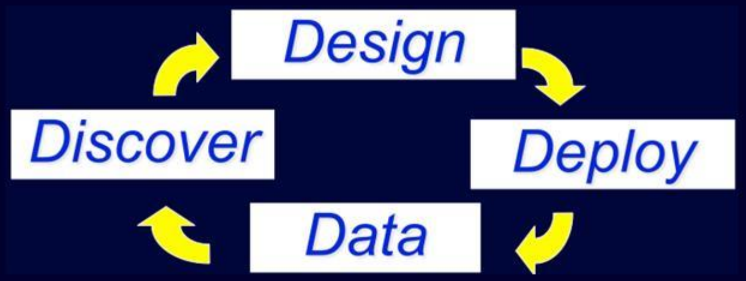
680 data sets

math, science, language …

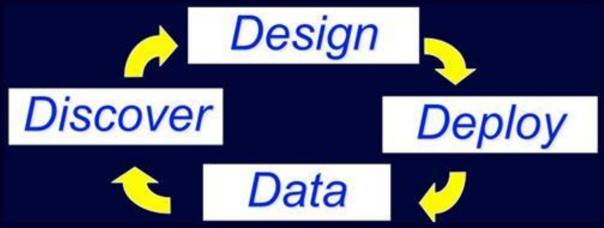
K12 & college

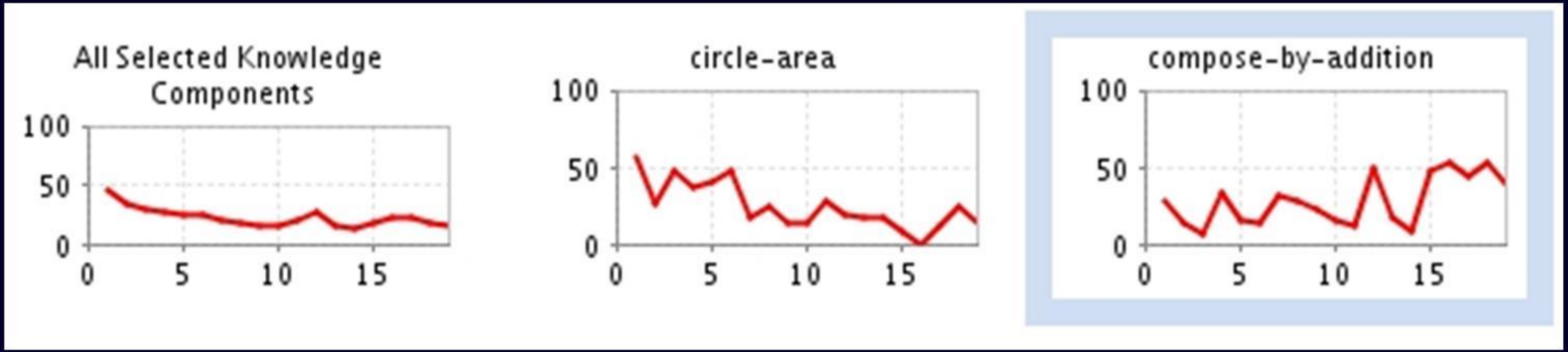
Closing the loop
Data-driven continuous improvement

- *Deploy* v1 of online course
  - Use *data* to make *discoveries*
  - Make *design* changes
- *Deploy* v2 vs. v1 in randomized controlled experiment
  - Use *data* to evaluate improvement



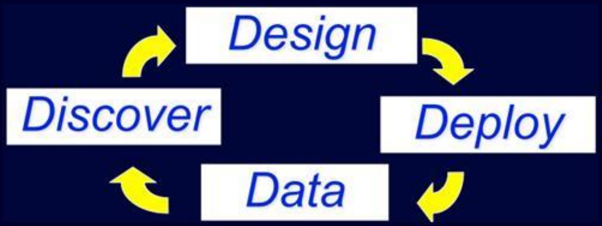
Visualizing learning curves to find opportunities for improvement

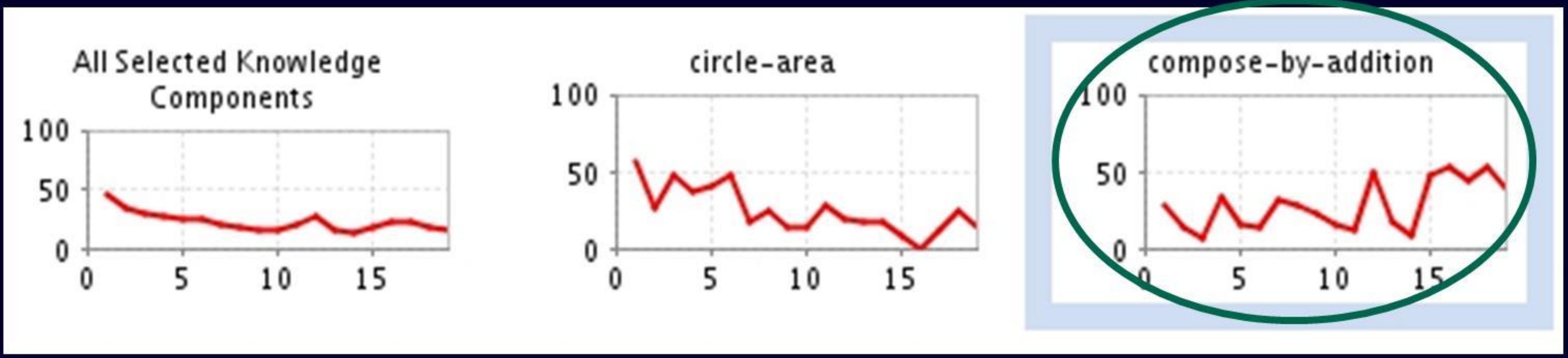




Koedinger, Stamper, McLaughlin, & Nixon. (2013). Using data-driven discovery of better student models to improve student learning. *Proceedings of Artificial Intelligence in Education.*

Visualizing learning curves to find opportunities for improvement





High rough curve

=> revise skill model

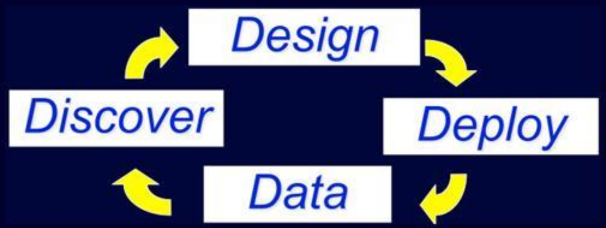
=> redesign instruction

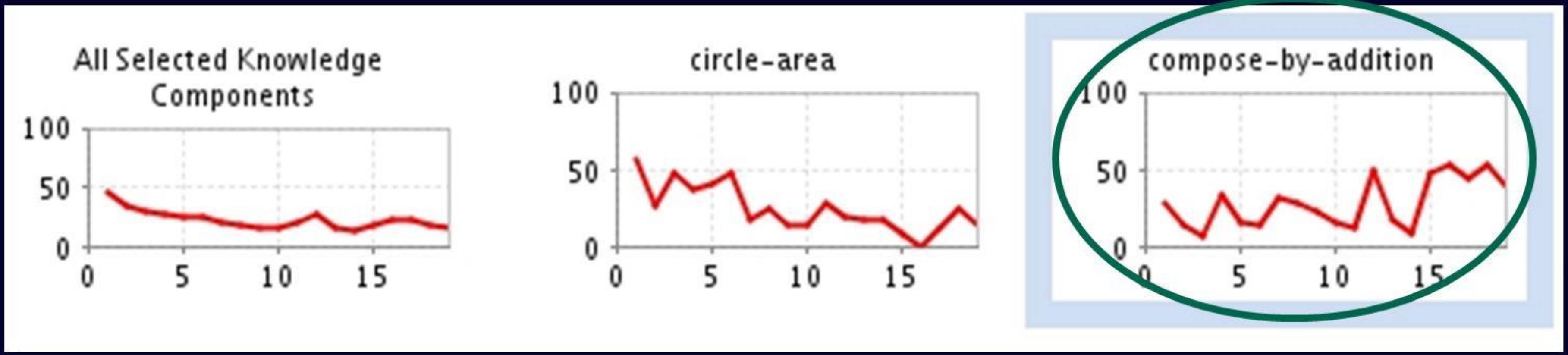
=> do A/B test

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.*

Visualizing learning curves to find opportunities for improvement





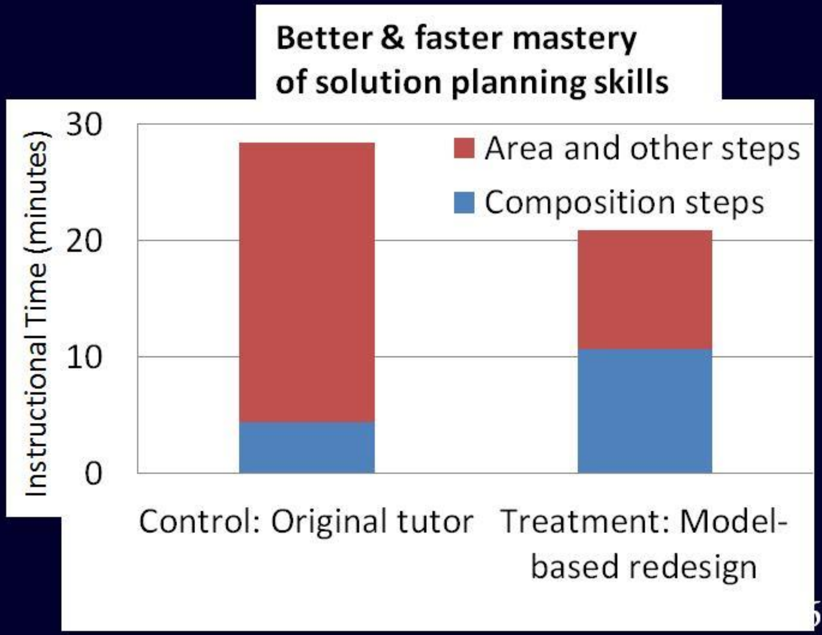
High rough curve

=> 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.*

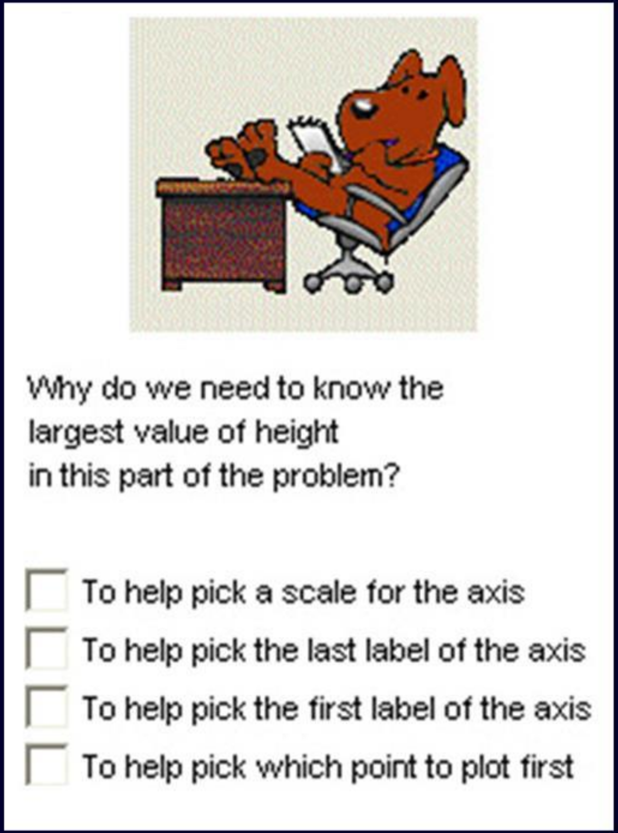
Machine learning on clickstream => diagnose engagement, “gaming the system”, & effectively respond

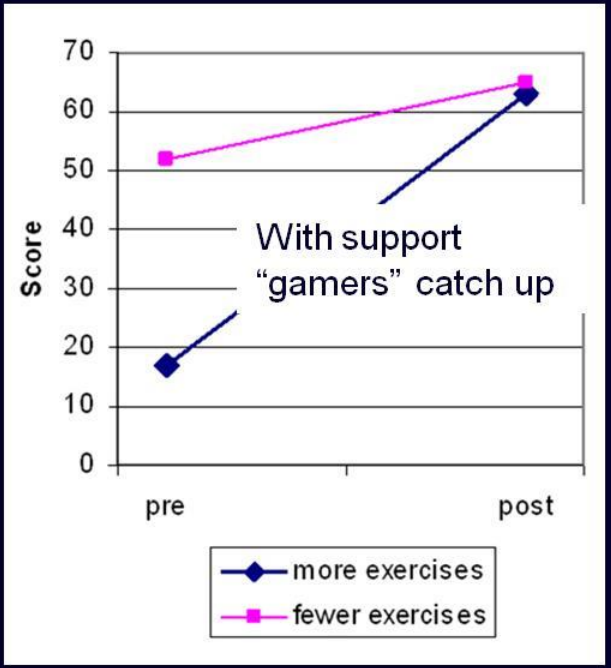
Why do we need to know the largest value of height in this part of the problem?

- [ ] To help pick a scale for the axis
- [ ] To help pick the last label of the axis
- [ ] To help pick the first label of the axis
- [ ] To help pick which point to plot first

(Baker et al., 2006; Rodrigo et al., 2011)

Machine learning on clickstream => diagnose engagement, “gaming the system”, & effectively respond

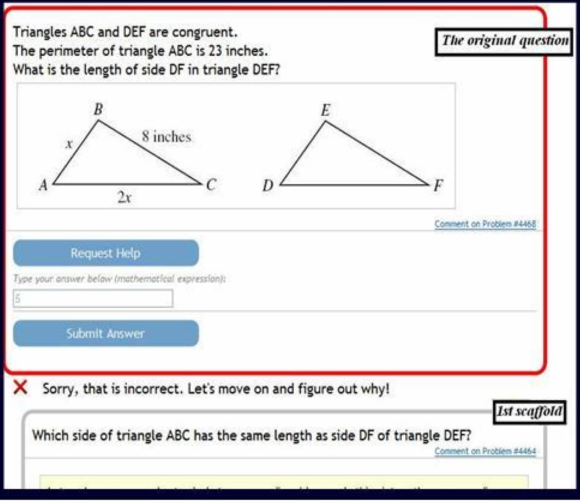


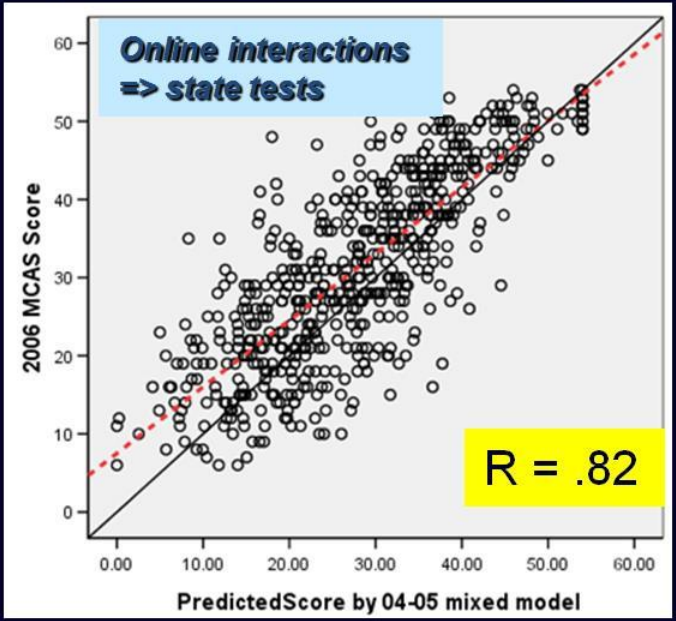


(Baker et al., 2006; Rodrigo et al., 2011)

Tech-based formative assessment: Aids learning & informs teachers

- Assess *during* learning
  - Data from Oct-Nov predicts end-of-year tests
  - Student effort matters
- Student feedback
  - Learning is better than traditional homework
- Teacher use
  - Low use students benefit from high use teachers

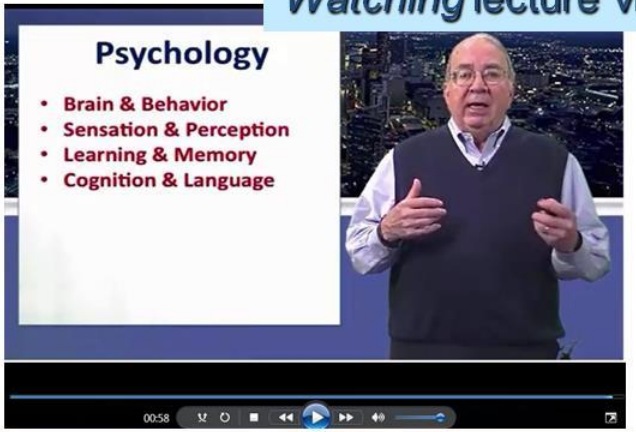


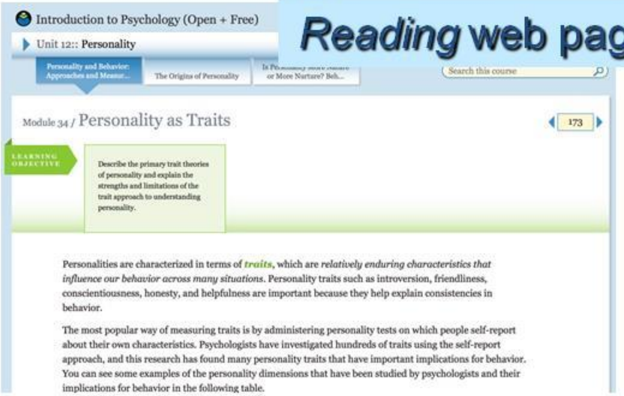


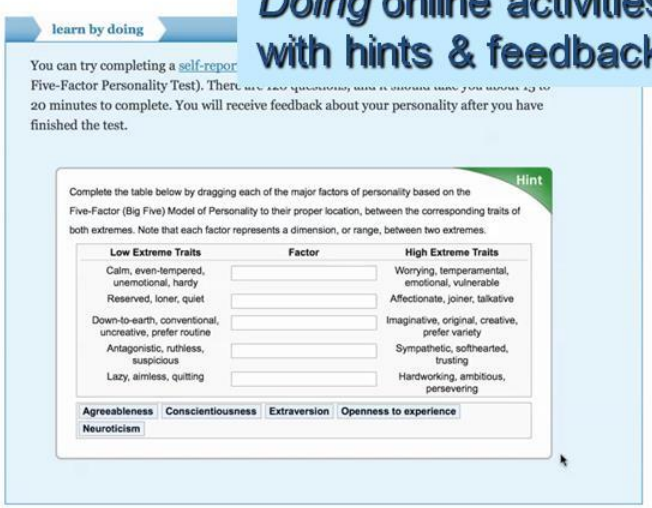
Koedinger, McLaughlin, & Heffernan (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research.*

Feng, Heffernan, & Koedinger (2009). Addressing the assessment challenge in an online system that tutors as it assesses. In *User Modeling and User-Adapted Interaction: The Journal of Personalization Research.*

MOOC Analysis:
What student choices associate with most learning?



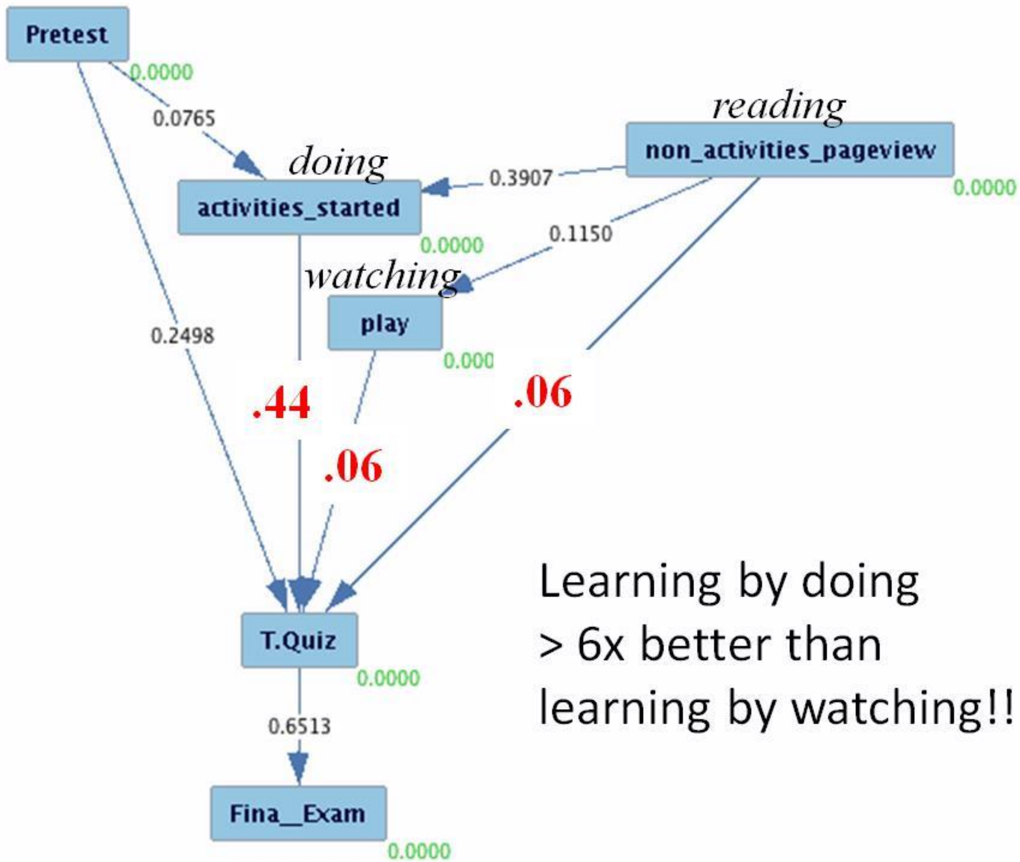




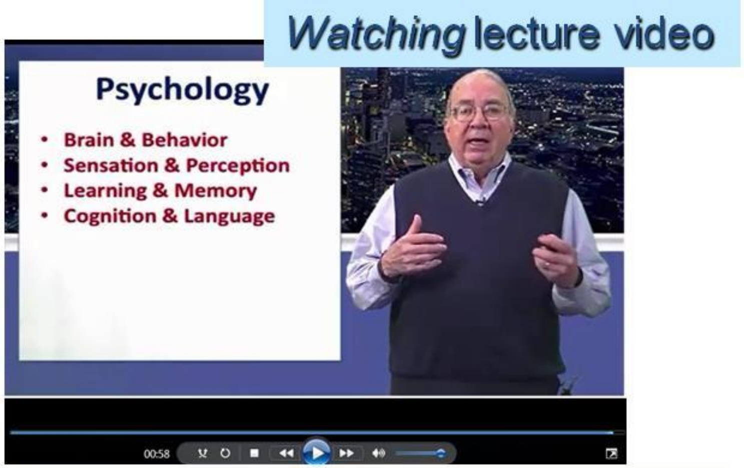
Koedinger et al. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. *Proceedings of Learning at Scale.*

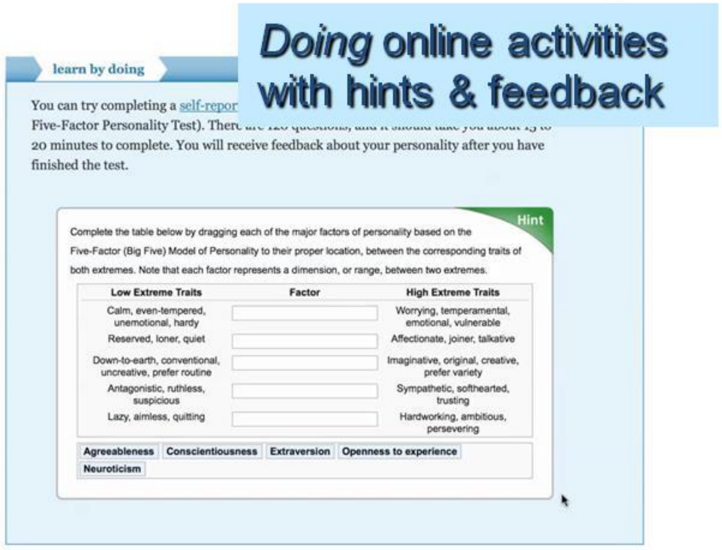
MOOC Analysis:

What student choices associate with most learning?



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



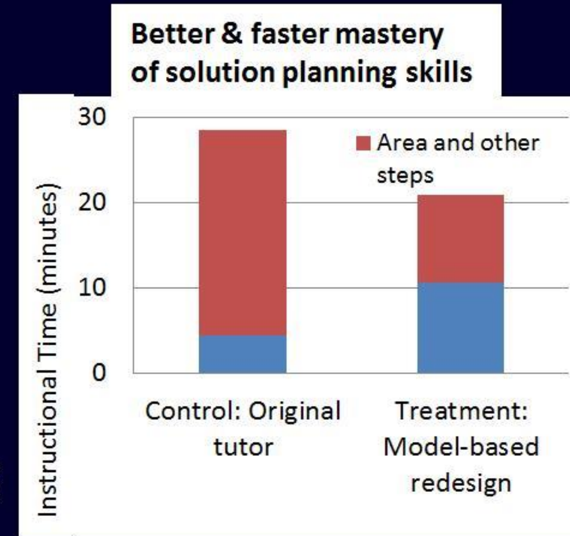


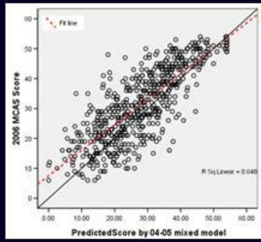
Koedinger et al. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. *Proceedings of Learning at Scale.*

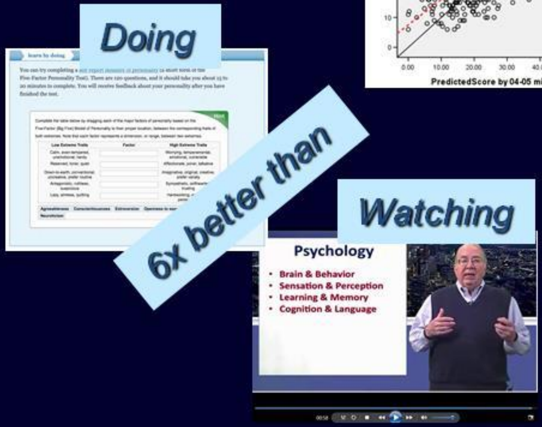
Summary So Far & Questions?

Why data sharing?

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





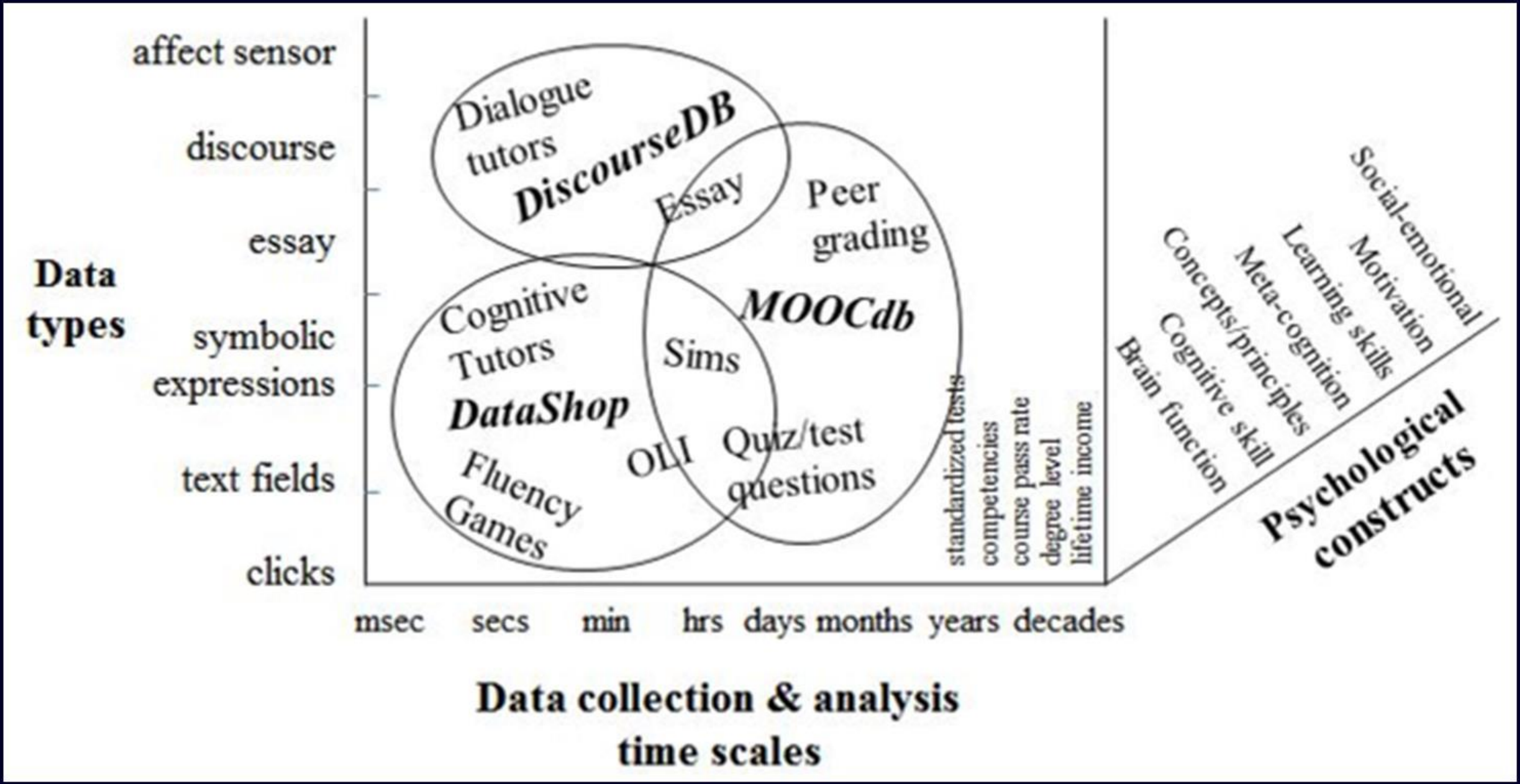


Sharing leverages interdisciplinary interaction

Outline

- Why data sharing?
- Data curation & privacy management
  - LearnSphere: DataShop, MOOCdb, DataStage, DiscourseDB
- Future of Cyberlearning data partnerships

Many kinds of data, time scales, and goals of analysis



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

Check out LearnSphere.org!

learnsphere.org

LearnSphere

About Explore

A community data infrastructure to support online learning improvement.

NSF

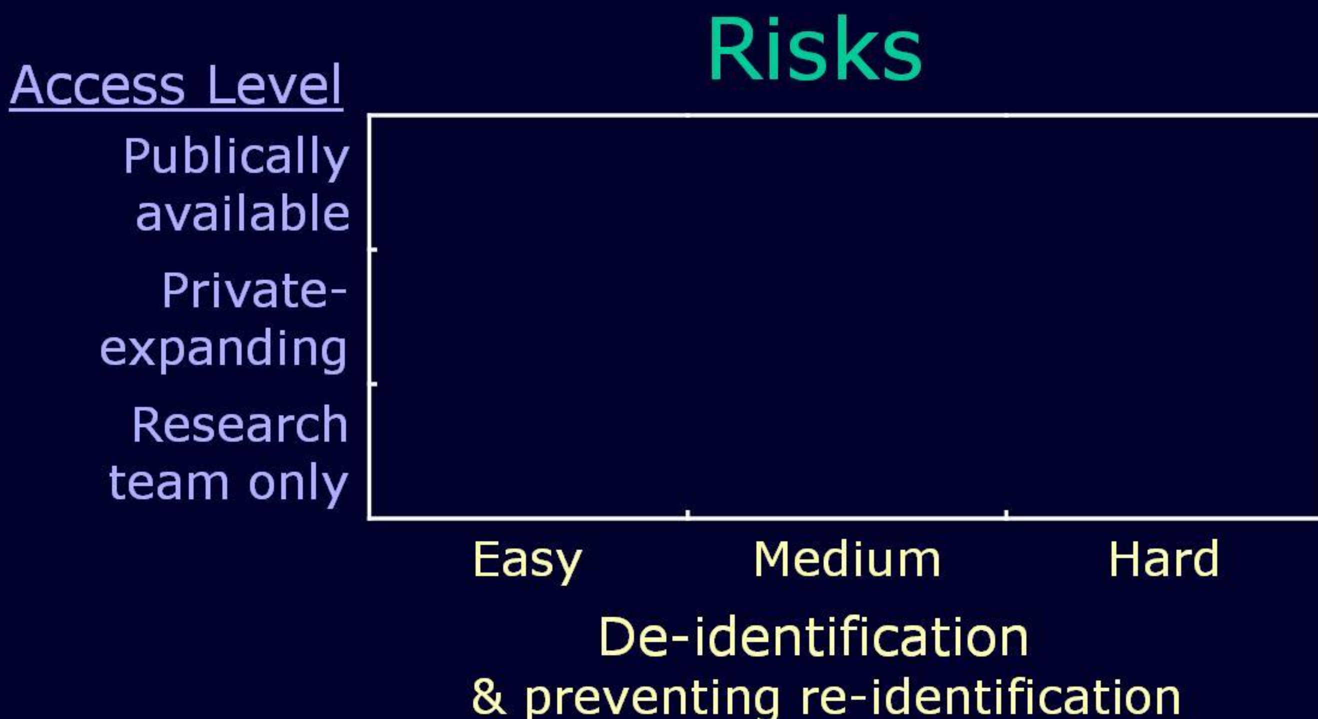
Existing Resources

PSLC DataShop
a data analysis service for the learning science community

MOOCDB

DATASTAGE

Privacy risks of kinds of data & availability



Privacy risks of kinds of data & availability

Risks

Access Level

Publically available

Private-expanding

Research team only

Easy | Medium | Hard

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

Risks

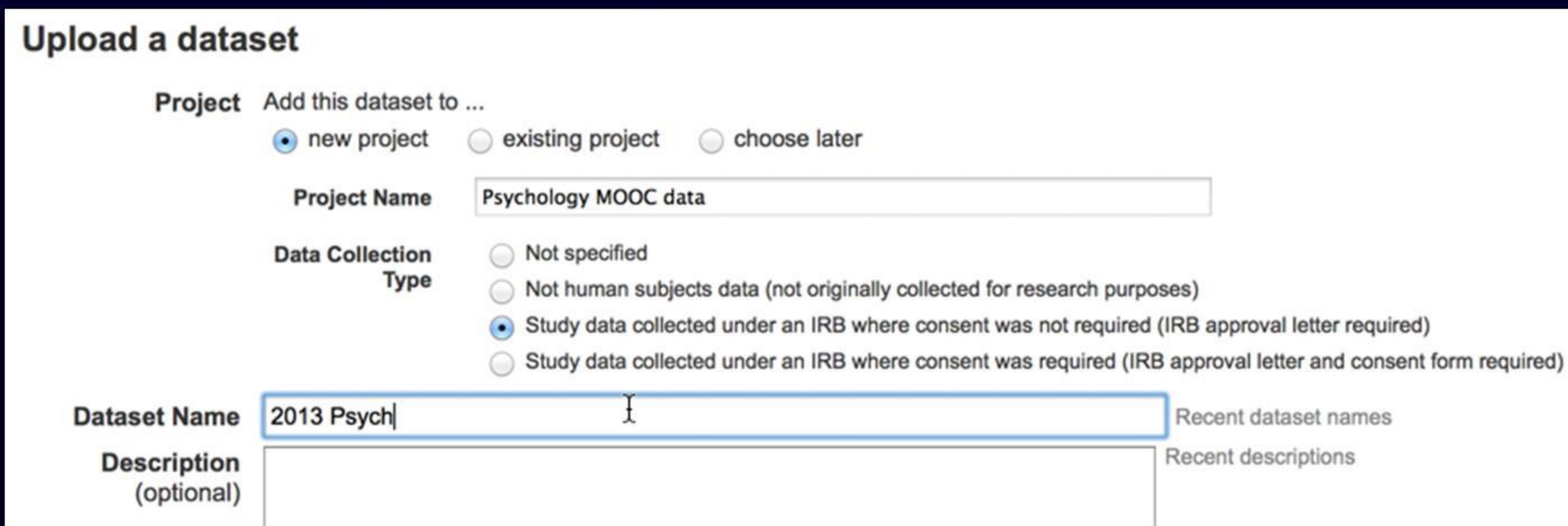
Access Level			
Publically available	Little	Lots	Too much
Private-expanding	Little	Some	Lots
Research team only	Little	Little	Little
	Easy	Medium	Hard

De-identification
& preventing re-identification

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

Review process starts when data is uploaded

- Automatic upload for some courses
- By-hand upload is easy using tab-delimited format



Data Management Distinctions

- **Shareability**--determined by *shareability review**
  - **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

- **Private vs Public**--determined by data owner, if shareable
  - **Private** = registered DataShop users cannot access project without data owner approval
  - **Public** = registered DataShop users may freely access project without owner approval (only “shareable” projects may be made public).

*Research manager conducts *shareability review* according to shareability criteria (see below)

Data Sharing Procedures in DataShop

	Private	Public
Not-Shareable	Only PI/data owner may access	
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

Data Sharing Procedures in DataShop

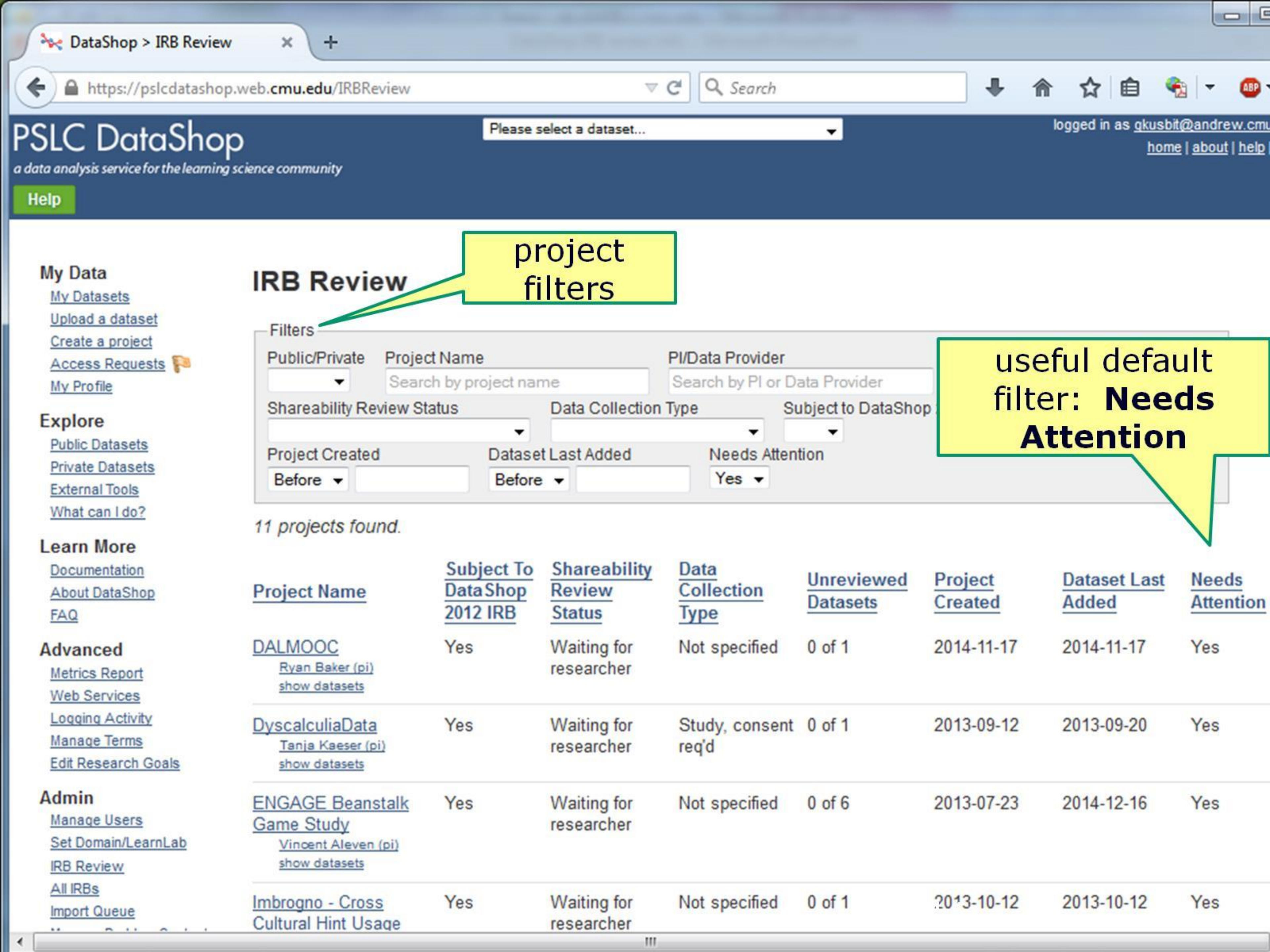
Research Manager determines Shareability

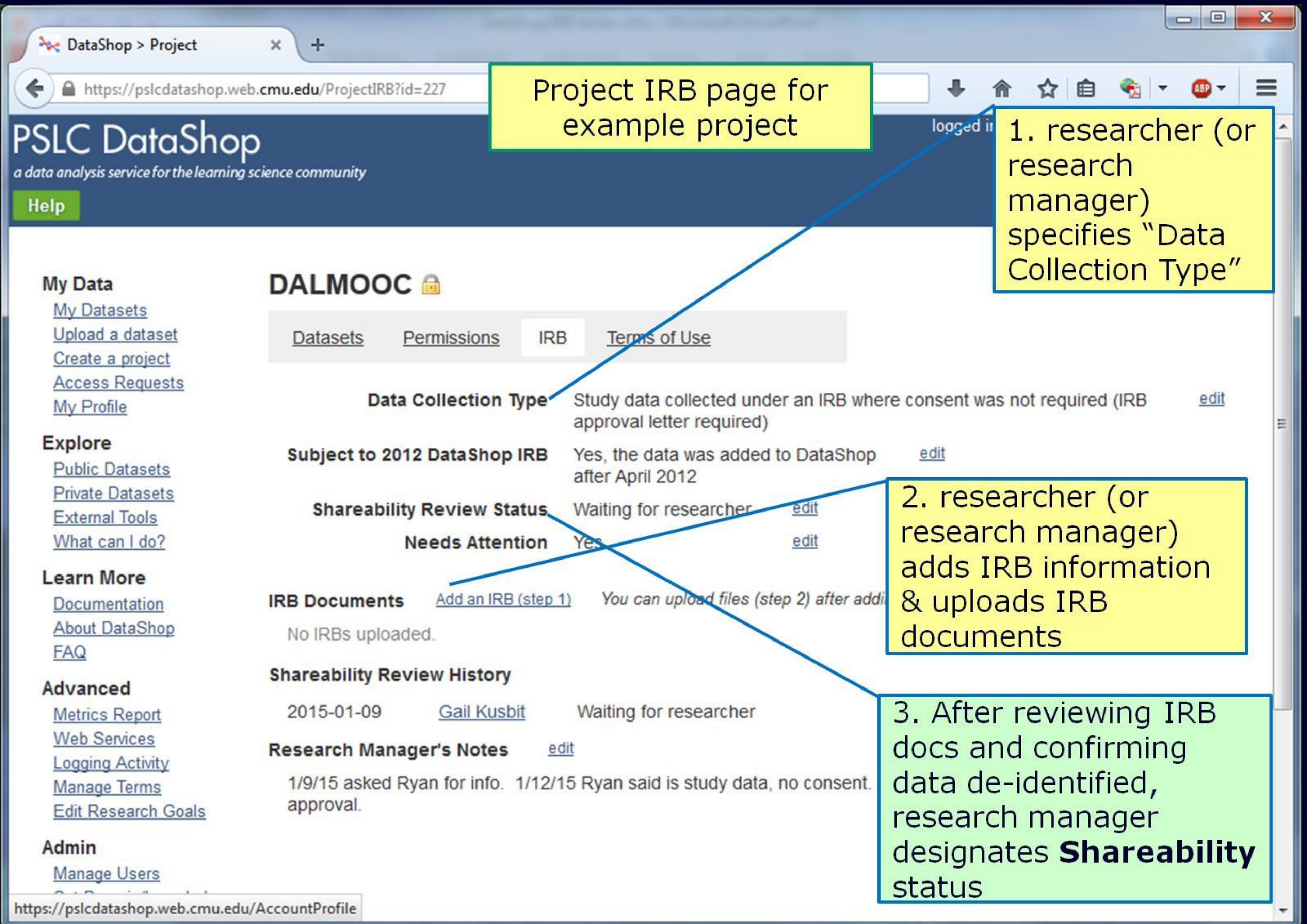
Data Owner determines Private/Public

	Private	Public
Not-Shareable	Only PI/data owner may access	
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

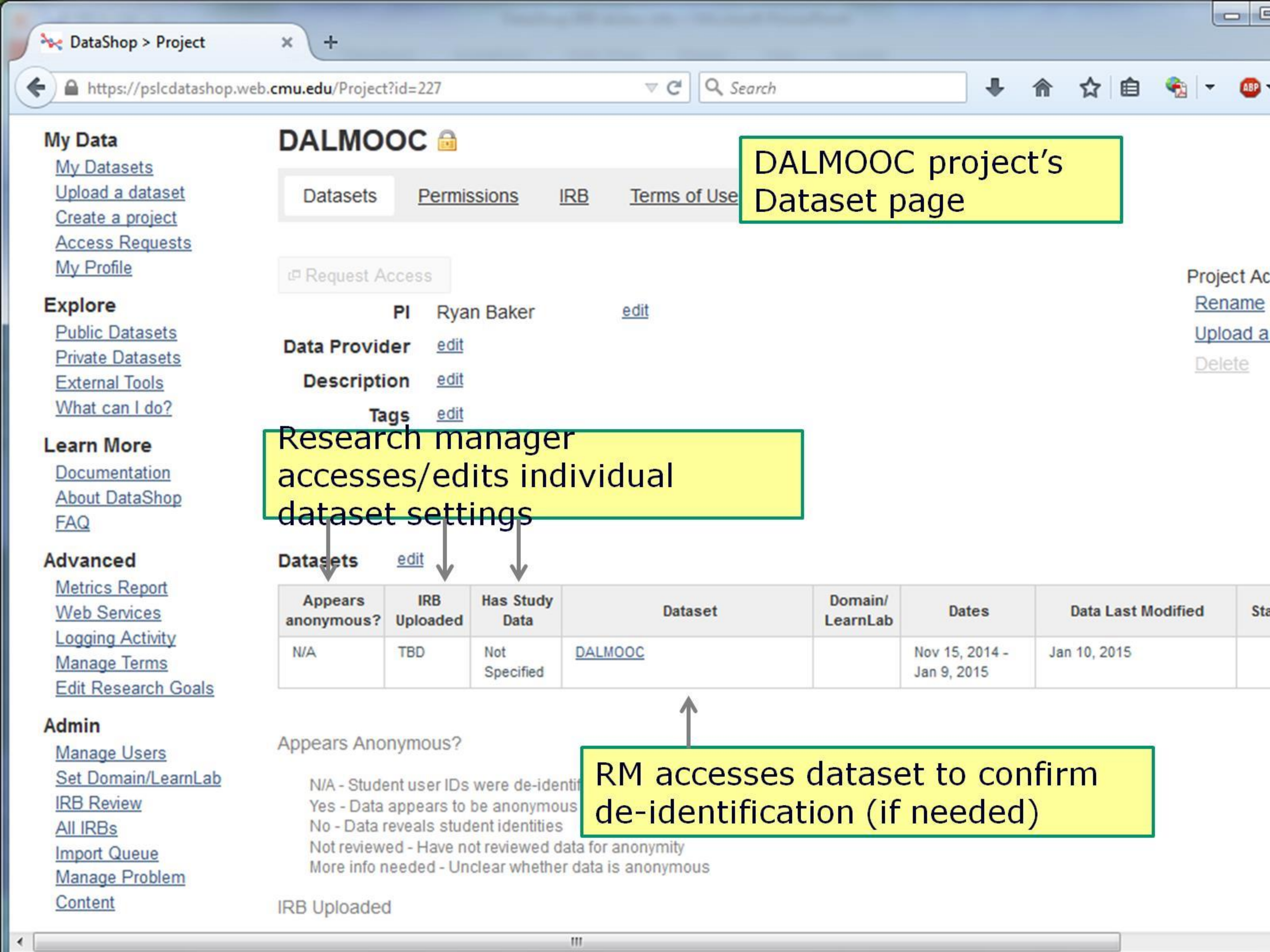
Default State: Private and Not-Shareable

Cyberinfrastructure supports for data privacy review & access management …


DataShop > IRB Review
https://pslcdatashop.web.cmu.edu/IRBReview
PSLC DataShop
a data analysis service for the learning science community
Please select a dataset...
logged in as gkusbit@andrew.cmu
home | about | help
Help
project filters
useful default filter: **Needs Attention**
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
IRB Review
Filters
Public/Private
Project Name
Search by project name
PI/Data Provider
Search by PI or Data Provider
Shareability Review Status
Data Collection Type
Subject to DataShop
Project Created
Before
Dataset Last Added
Before
Needs Attention
Yes
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 | Tanja Kaeser (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 Aleven (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



Project IRB page for example project
1. researcher (or research manager) specifies "Data Collection Type"
2. researcher (or research manager) adds IRB information & uploads IRB documents
3. After reviewing IRB docs and confirming data de-identified, research manager designates **Shareability** status
DataShop > Project
https://pslcdatashop.web.cmu.edu/ProjectIRB?id=227
PSLC DataShop
a data analysis service for the learning science community
Help
logged i
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
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 addi
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. approval.
https://pslcdatashop.web.cmu.edu/AccountProfile


DataShop > Project
https://pslcdatashop.web.cmu.edu/Project?id=227
Search
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
DALMOOC project's Dataset page
Request Access
PI Ryan Baker edit
Data Provider edit
Description edit
Tags edit
Research manager accesses/edits individual dataset settings
Project Ac
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
RM accesses dataset to confirm de-identification (if needed)
Appears Anonymous?
N/A - Student user IDs were de-identif
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

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

Project Actions:
Rename
Upload a dataset
Delete

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.

Adding a new dataset or new data triggers "needs 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	

DataShop > IRB Review

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

PSLC DataShop

a data analysis service for the learning science community

Please select a dataset...

logged in as gkusbit@andrew.cmu.e
home | about | help | c

Help

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: Search by project name | PI/Data Provider: Search by PI or Data Provider

Shareability Review Status | Data Collection Type | Subject to DataSho: Yes

Project Created: Before | Dataset Last Added: Before | Needs Attention: Yes

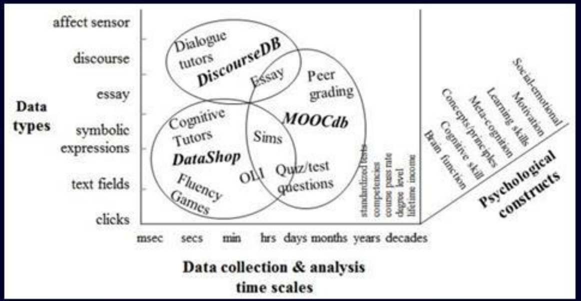
System highlights any **PUBLIC** projects that need attention in yellow

12 projects found.

Project Name	Subject To DataShop 2012 IRB	Shareability Review Status	Data Collection Type	Unreviewed Datasets	Project Created	Dataset L Added	Needs Attention
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 Kaeser (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 Aleven (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 Aleven (pi) show datasets	Yes	Waiting for researcher	Not specified	4 of 10	2010-0[illegible] 19	2013-12-04	Yes

Summary So Far & Questions?

- Avoid data silos to facilitate multi-disciplinary discovery
- Privacy protection depends on data & access
- Cyberinfrastructure helps manage privacy protection

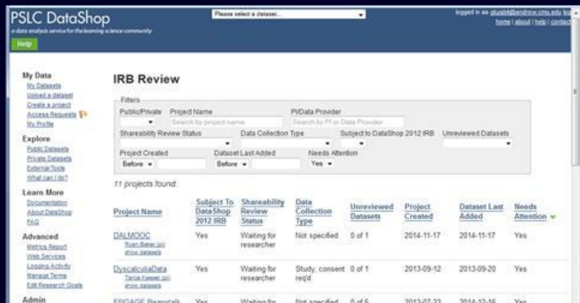


Risks

Access Level	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



Outline

- Why data sharing?
- Data curation & privacy management
  - LearnSphere: DataShop, MOOCdb, DataStage, DiscourseDB
- Future of Cyberlearning data partnerships

Where to go from here?

Possible partnerships/collaborations/relationships to pursue Cyberlearning advances through data sharing?

- Analyses that span levels of analysis

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

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

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
- Privacy:
  Limit access as de-identification increases
- Future: Big data partnerships to tackle big interdisciplinary education questions

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

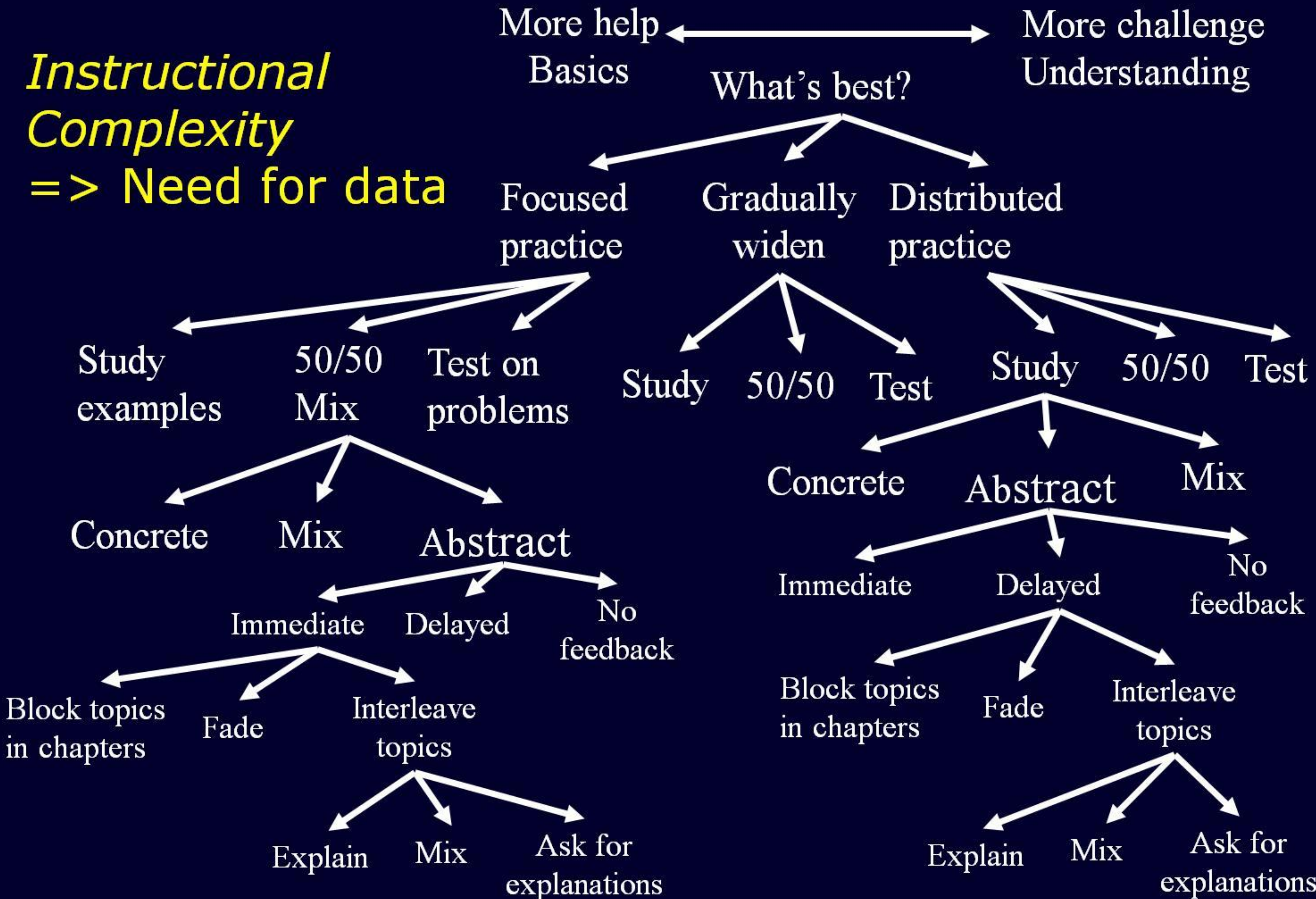
Ken Koedinger
koedinger@cmu.edu

Extras …

New professional masters!

www.metals.cs.cmu.edu

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*.

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

DataShop IRB Procedures

Gail Kusbit, LearnLab Research Manager

As of 1/16/15, DataShop has

- 187 projects containing a total of 641 datasets
- 23 projects are public, 154 are private

Shareability Review Criteria

<u>Data collection types</u>

1. ***Not human subjects*** (data not originally collected for research purposes, e.g. course data)
   - *the data is de-identified*

2. ***Study data*** collected under an IRB where ***consent not required***
   - *the data is de-identified*
   - *IRB approval letter*

3. ***Study data*** collected under an IRB where ***consent required***
   - *the data is de-identified*
   - *IRB approval letter*
   - *Consent form*
   - *Non-prohibitive consent text*

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”

- Researcher/data provider adds **new data** to an “old” dataset after its project had been designated “shareable”

Confirming dataset de-identification

For datasets not automatically de-identified by DataShop system, research manager…

- exports a sample of the dataset
- looks through 200 rows of data
- makes *"good faith effort"* to search for any identifiable information
  - if no identifying info found, RM manually changes "appears anonymous" to "yes"
  - if identifying info found, RM consults with DataShop personnel and data owner—dataset removed from DataShop

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 as well as listing of the projects associated with each IRB.

- one project can have multiple IRBs related to it: A project's IRB page shows all relevant IRB info w/links to documents

Non-exempt IRB Protocol Example

Protocol description :

- "Pilot studies, up to 30 paid students per year. Conducted at CMU to test our study materials before conducting the large-scale classroom studies."

Standard consent wording (permits sharing):

- By your child's participating, you understand and agree that the data and information gathered during this study may be used by Carnegie Mellon and published and/or disclosed by Carnegie Mellon to others outside of Carnegie Mellon.

Newer IRB-suggested consent wording (specific to DataShop, permits sharing):

- Your de-identified data may be stored indefinitely in the DataShop repository at Carnegie Mellon University. Only registered users of DataShop will have access to the data for analysis purposes.

Exempt IRB Protocol Example

Protocol description:

- "Students will either be completing the standard curriculum unit (already shown to improve learning) or the modified curriculum unit, which has been designed to be at least as good as the standard curriculum."

Protocol confidentiality/privacy wording:

- Carnegie Learning personnel will de-identify all student data before giving it to the research team. De-identified data will be stored on secure DataShop server, and may be shared with registered DataShop users for analysis purposes.

Count of project types in DataShop

- 187 projects total
- Projects subject to DataShop's 2012 IRB = 112
  - Not human subjects projects = 24
  - Exempt projects (consent not required) = 43
  - Non-exempt projects (consent required) = 21
  - Not specified (waiting for researcher) = 8
  - Misc (DataShop personnel testing, projects created but no datasets inside) = 16