Continuous Improvement of Educational Technology through Discoveries with Big Data

John Stamper
Human-Computer Interaction Institute
Carnegie Mellon University

Carnegie Mellon University

AI, ML, and Big Data

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

Carnegie Mellon University

DataShop

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

Carnegie Mellon University

How big is DataShop?

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

As of July 2019

Carnegie Mellon University

What kinds of data?

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

The data is fine grained at a transaction level!

DataShop Terminology

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

Carnegie Mellon University

Getting the KC Model Right!

The KC model drives instruction in adaptive learning

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

What makes a good KC Model?

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

The model should fit the data!

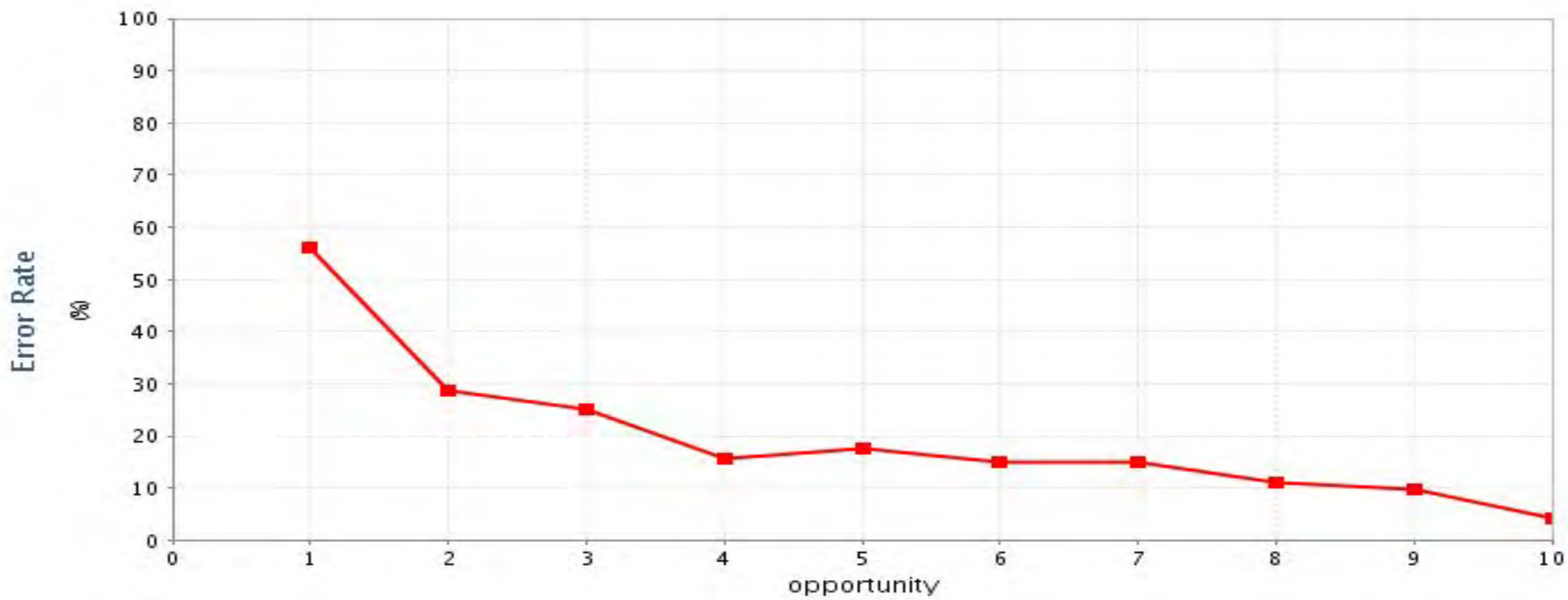
Carnegie Mellon University

Good KC Model = Good Learning Curve

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

A declining error rate learning curve should emerge

A Good Learning Curve



Carnegie Mellon University

How do we make the Models?

Traditionally Cognitive Task Analysis

But CTA interview methods have some issues…

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

And these human discovered models are often wrong!

Carnegie Mellon University

If human centered intuitive design is not the answer…

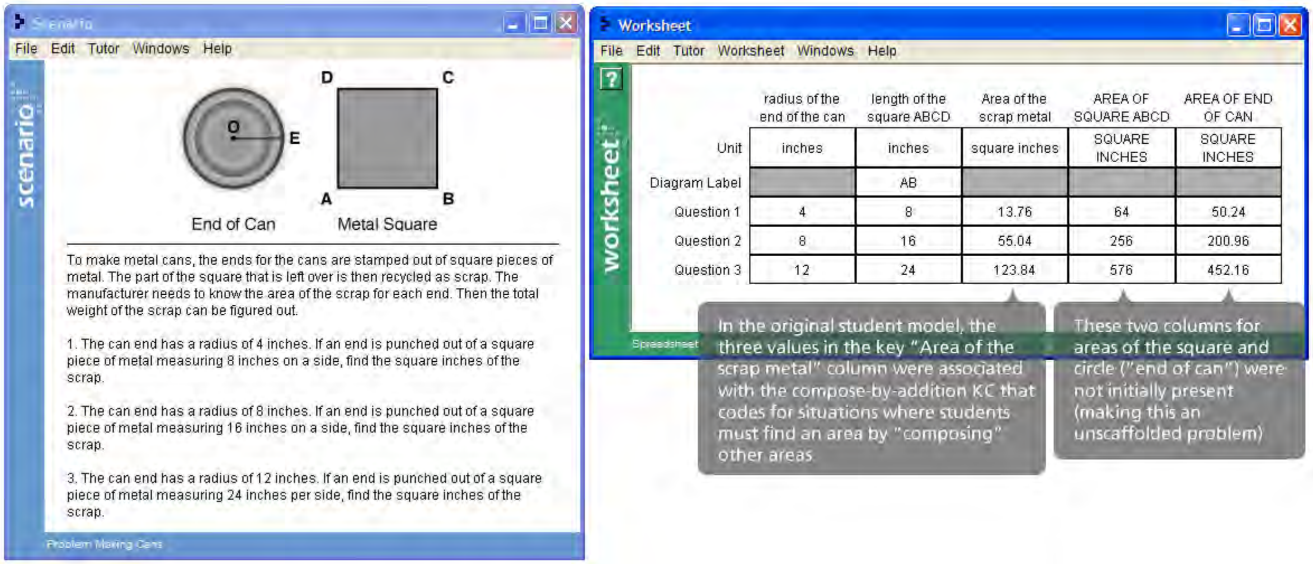
How should student models be designed?

They shouldn’t!

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

Solution – Use Data

Today we have lots of log data from edtech



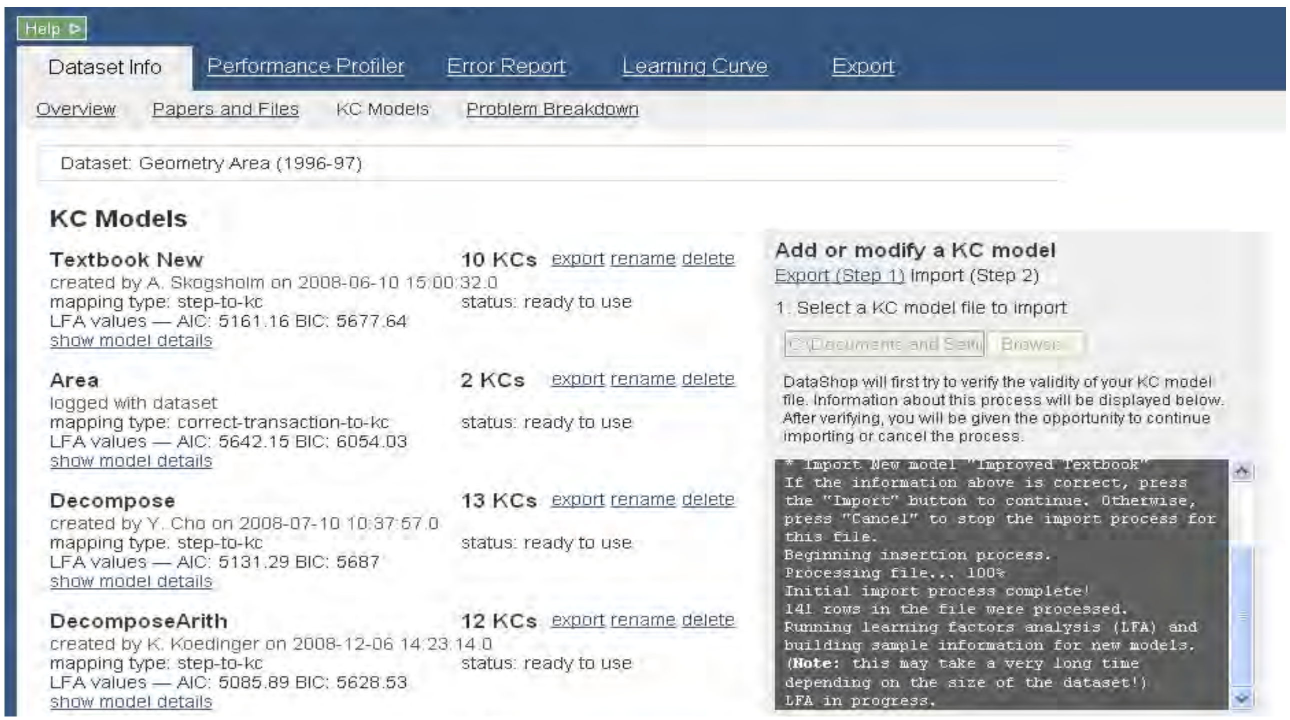
We can harness this data to validate and improve existing student models

Carnegie Mellon University

Human-Machine Student Model Discovery

(Stamper & Koedinger, 2011)

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



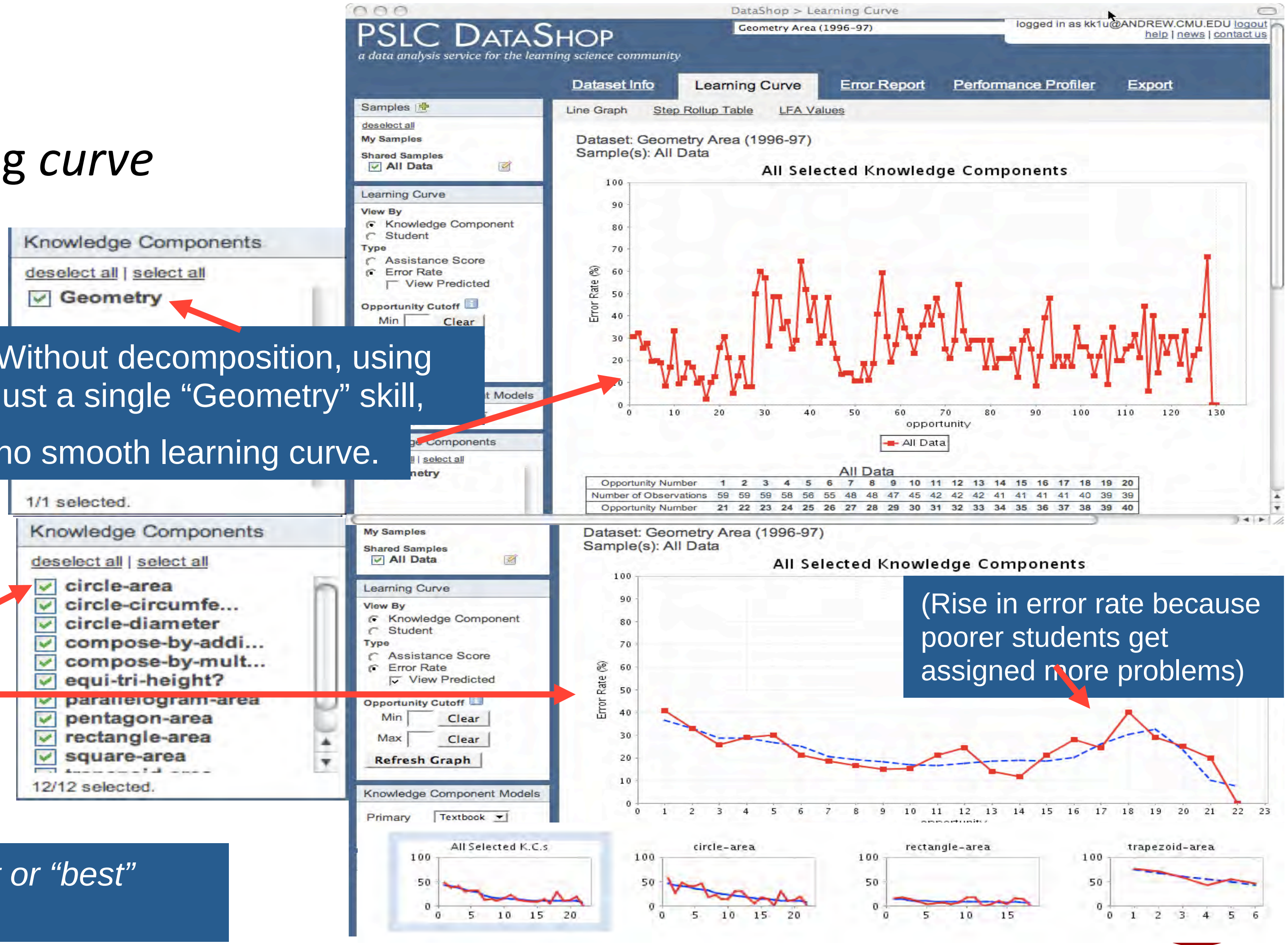
Human-Machine Student Model Discovery

3 strategies for discovering improvements to the student model

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

Carnegie Mellon University

A good KC model produces a learning *curve*



Without decomposition, using just a single “Geometry” skill,

no smooth learning curve.

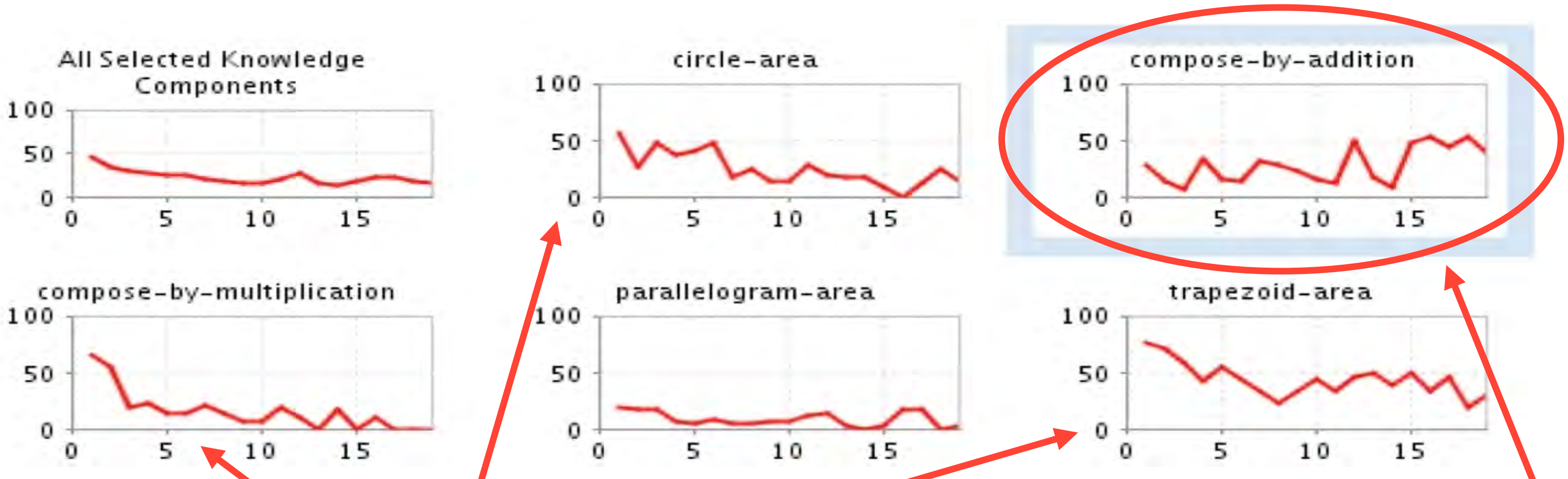
But with 12 skills for geometry area,

a smooth learning curve.

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

Is this the correct or “best” model?

Inspect curves for individual knowledge components (KCs)



Many curves show a reasonable decline

Some do not => Opportunity to improve model!

No apparent Learning

KC Values For Textbook Model

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

Carnegie Mellon University

These strategies suggest an improvement

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

What does a better fit really mean?

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

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

Redesign based on Discovered Model

Our discovery suggested changes needed to be made to the tutor

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

Carnegie Mellon University

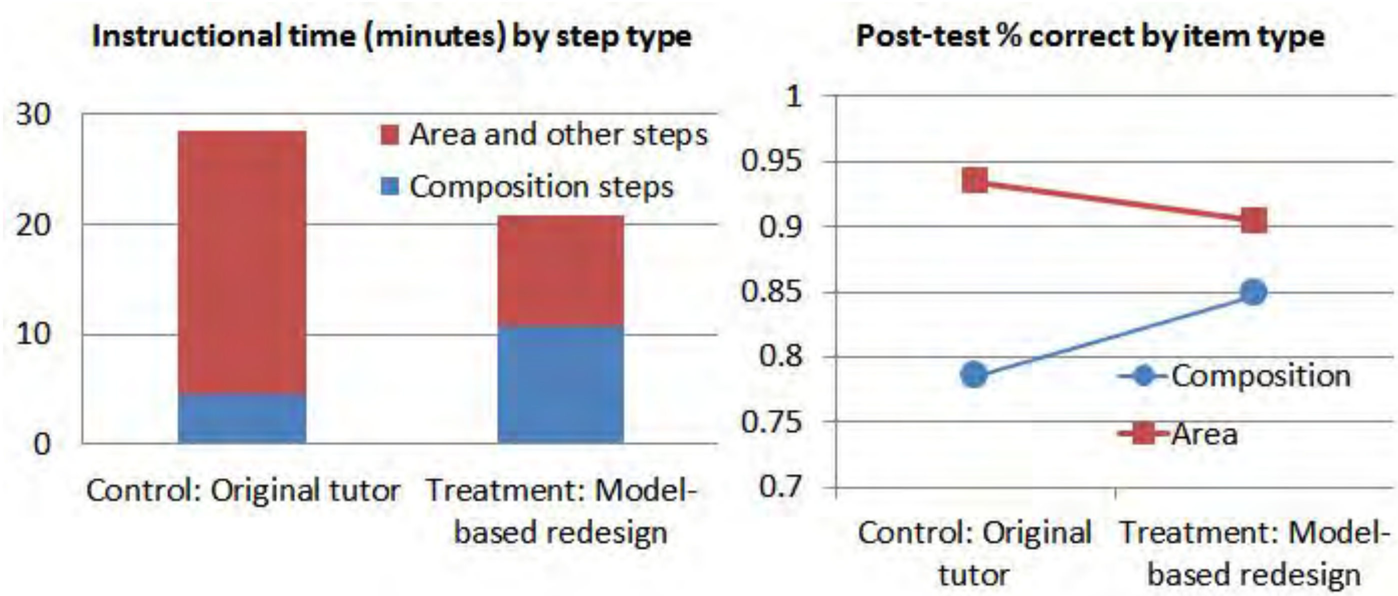
Closing the Loop, (Koedinger, Stamper, McLauglin, 2013)

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

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

Carnegie Mellon University

Results



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

Carnegie Mellon University

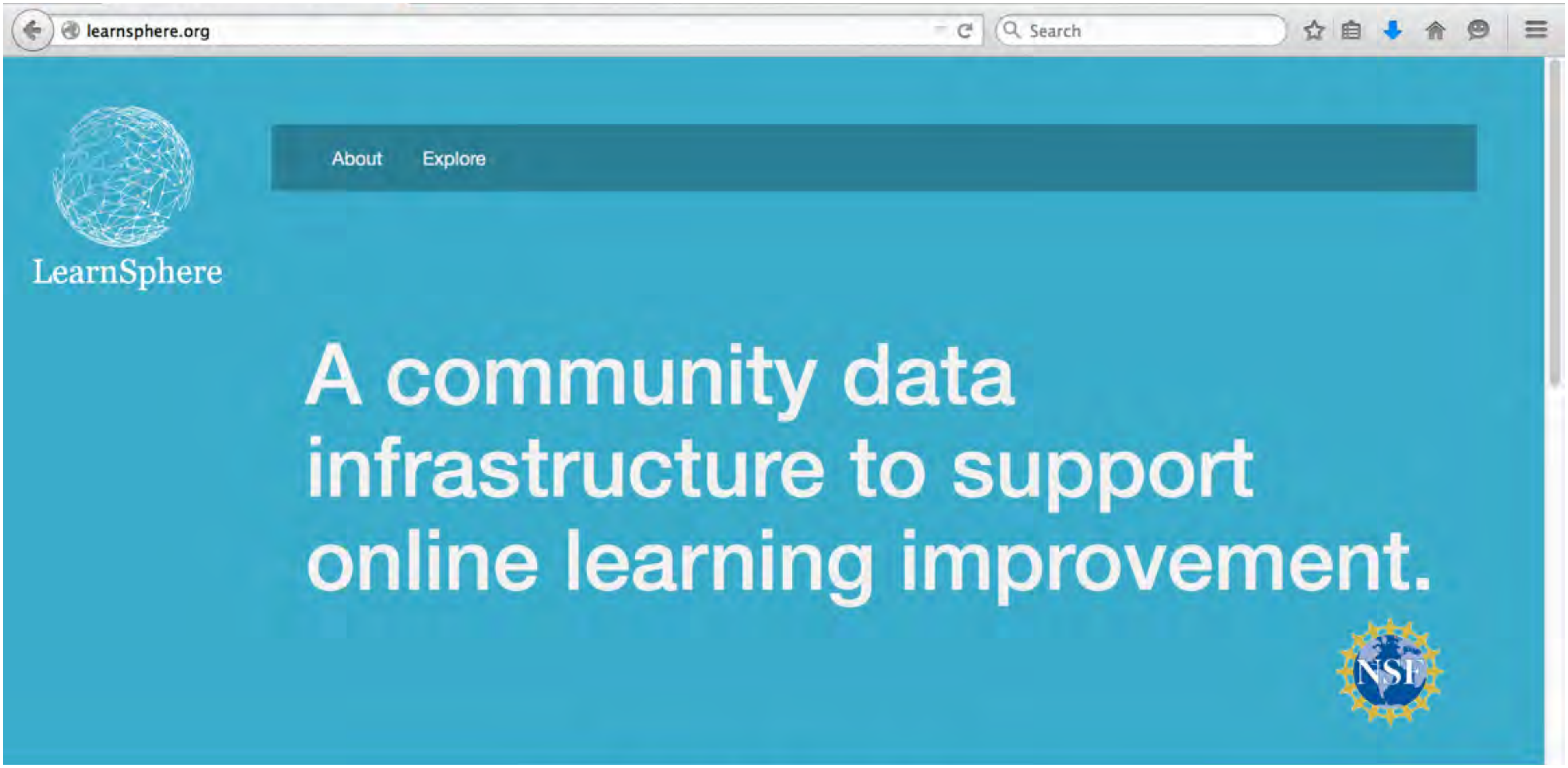
Other Data Driven Projects

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

Carnegie Mellon University

LearnSphere

"a community software infrastructure around the analysis of educational data that supports sharing and collaboration across the wide variety of educational data" http://learnsphere.org



Team

Data Silos → Data Integration

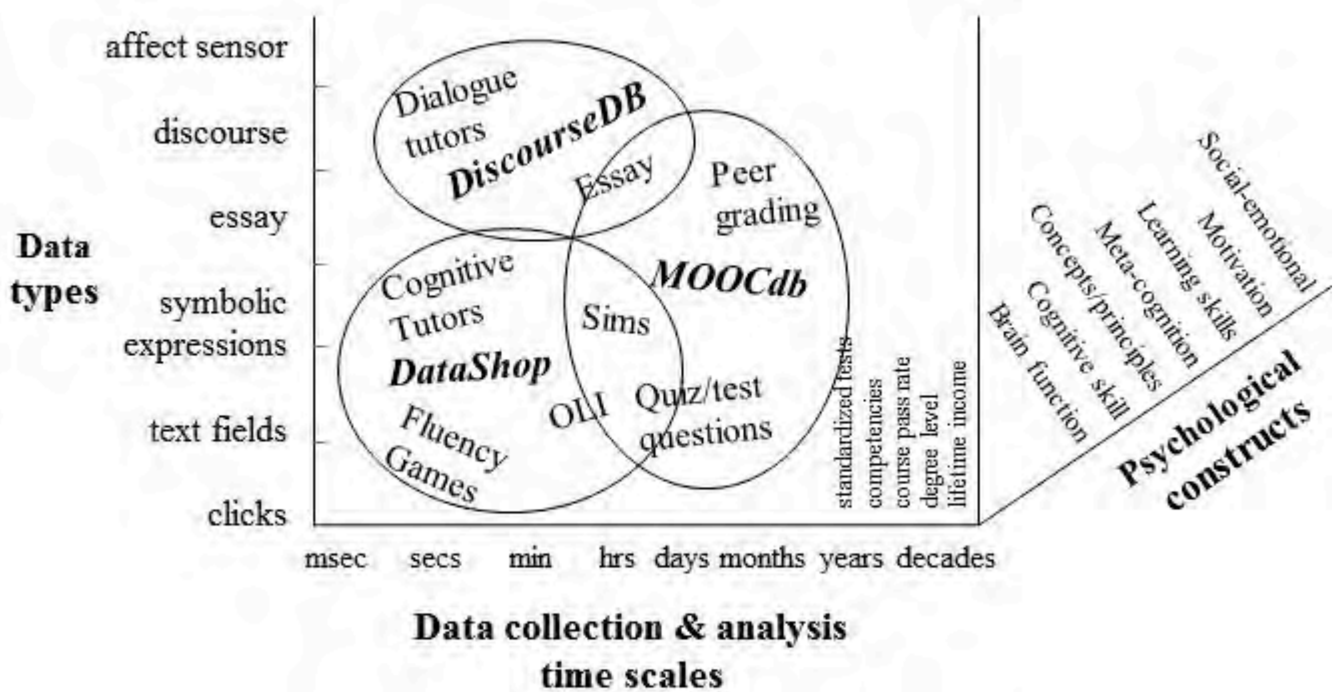
Many paradigms of data-driven education research differ in

- data types
- time scale
- research goals

Disciplinary silos are fostered by differences

Data infrastructure for analytics across these

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

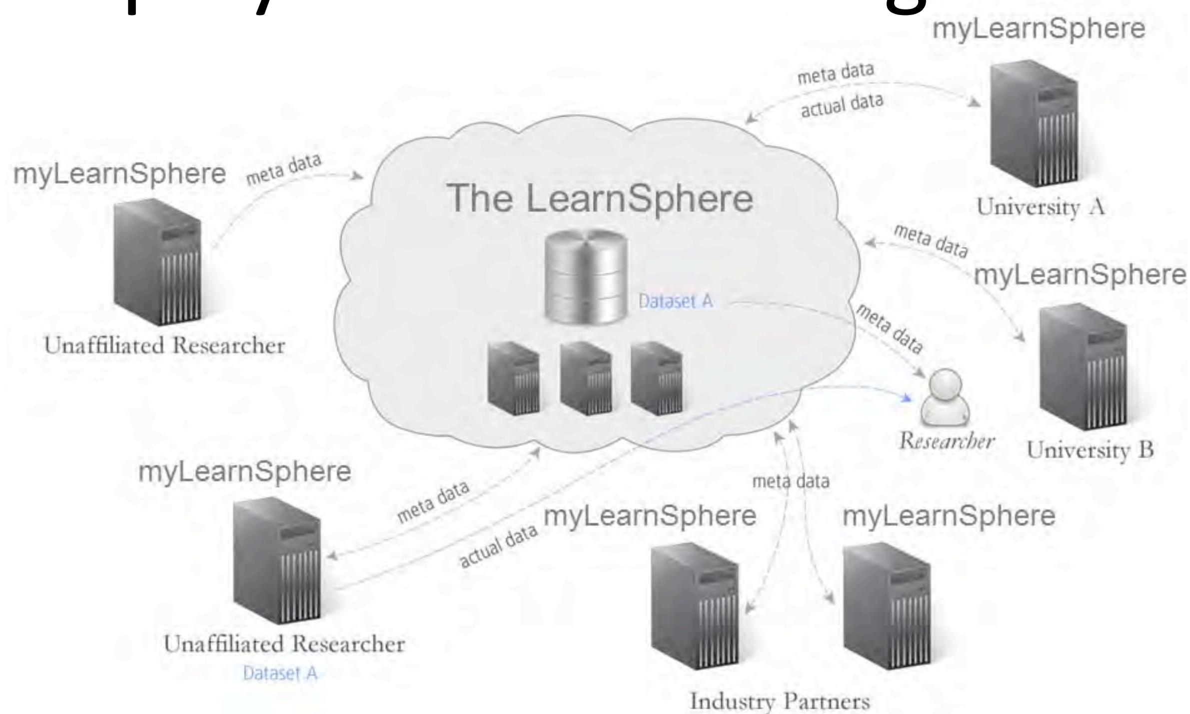


Carnegie Mellon University

Distributed deployment and storage

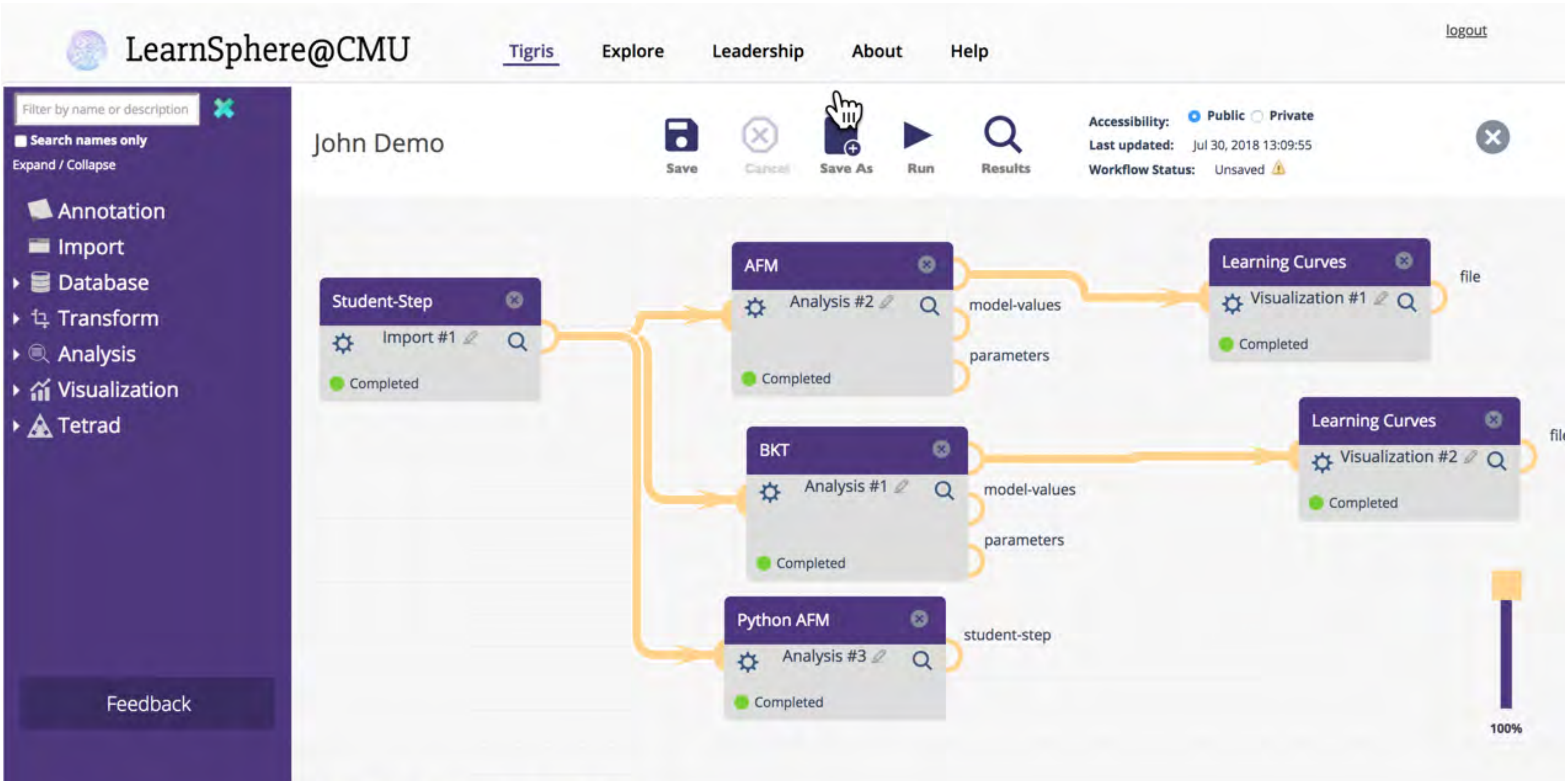
Central LearnSphere portal
Multiple "myLearnSphere" installations

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



Carnegie Mellon University

Learning Analytics Workflow Authoring Environment



Take Aways

- The amount of data coming from educational technology is growing exponentially (Big Data is here in Education)

- Students are going to rely more and more on technology, so improving the learning in edtech systems is critical

- Human-Centered, Data-Driven approaches are most likely to be the ones that succeed in actionable improvements in edtech

Carnegie Mellon University

HCII

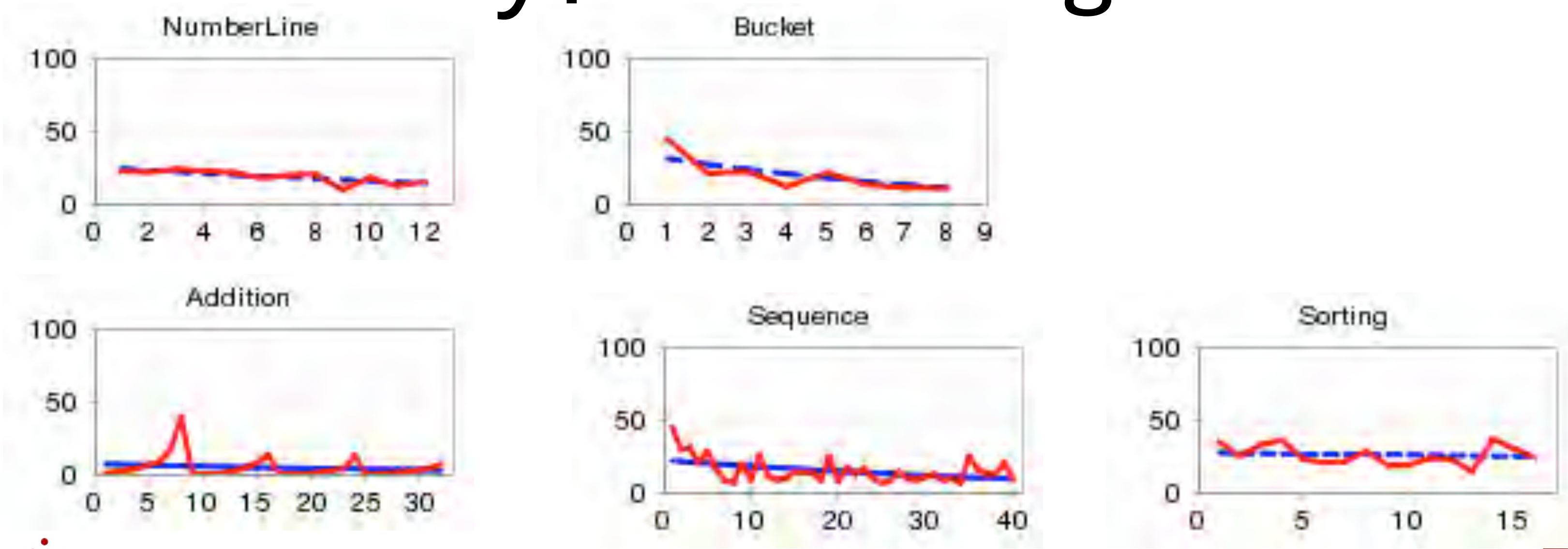
Questions?

Contact me: jstamper@cmu.edu

Carnegie Mellon University

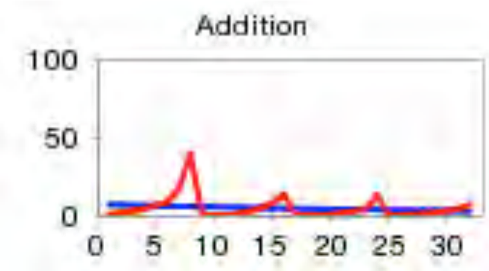
HCII

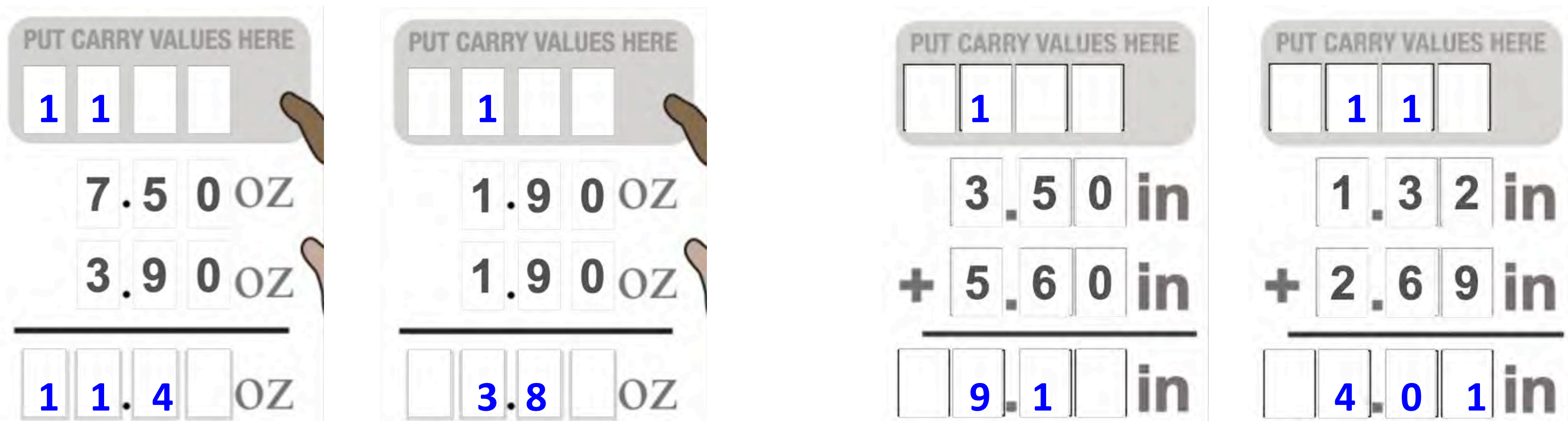
Methods & Results - `ProblemType` Learning Curves



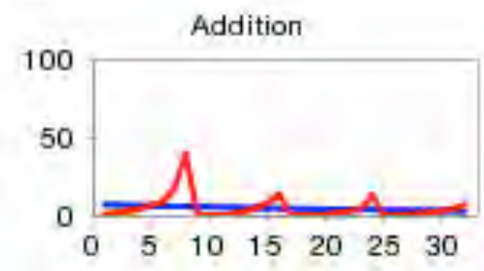
Carnegie Mellon University

Methods & Results - Addition KC Decomposition

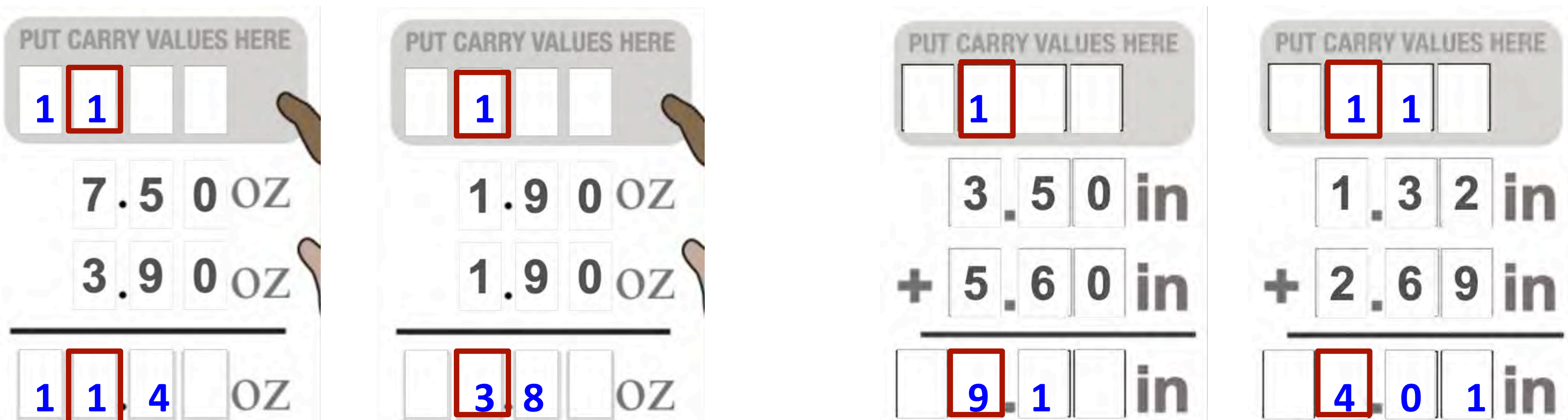




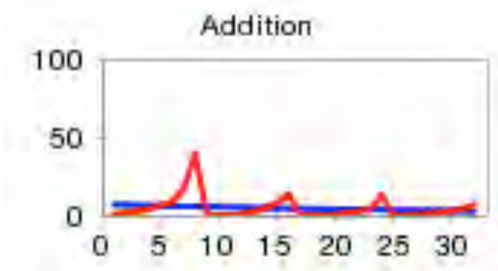
Methods & Results - Addition KC Decomposition



Addition_Ones



Methods & Results - Addition KC Decomposition



Addition_Ones

Addition_Tens_NonZero