

LearnSphere: Infrastructure for Data-Driven Discovery & Innovation in Education

Carnegie Mellon: Ken Koedinger, John Stamper,
& Carolyn Rose

MIT: Una-May O'Reilly & Kalyan Veeramachaneni

Stanford: Candace Thille

U of Memphis: Phil Pavlik

Support from NSF Cyberinfrastructure, DIBBs, \$5M for 5 years

DIBBs PI Meeting: Jan 11-12, 2017

Panel 3: Remaining Challenges & Future Directions



Big data in Education *Opportunity*

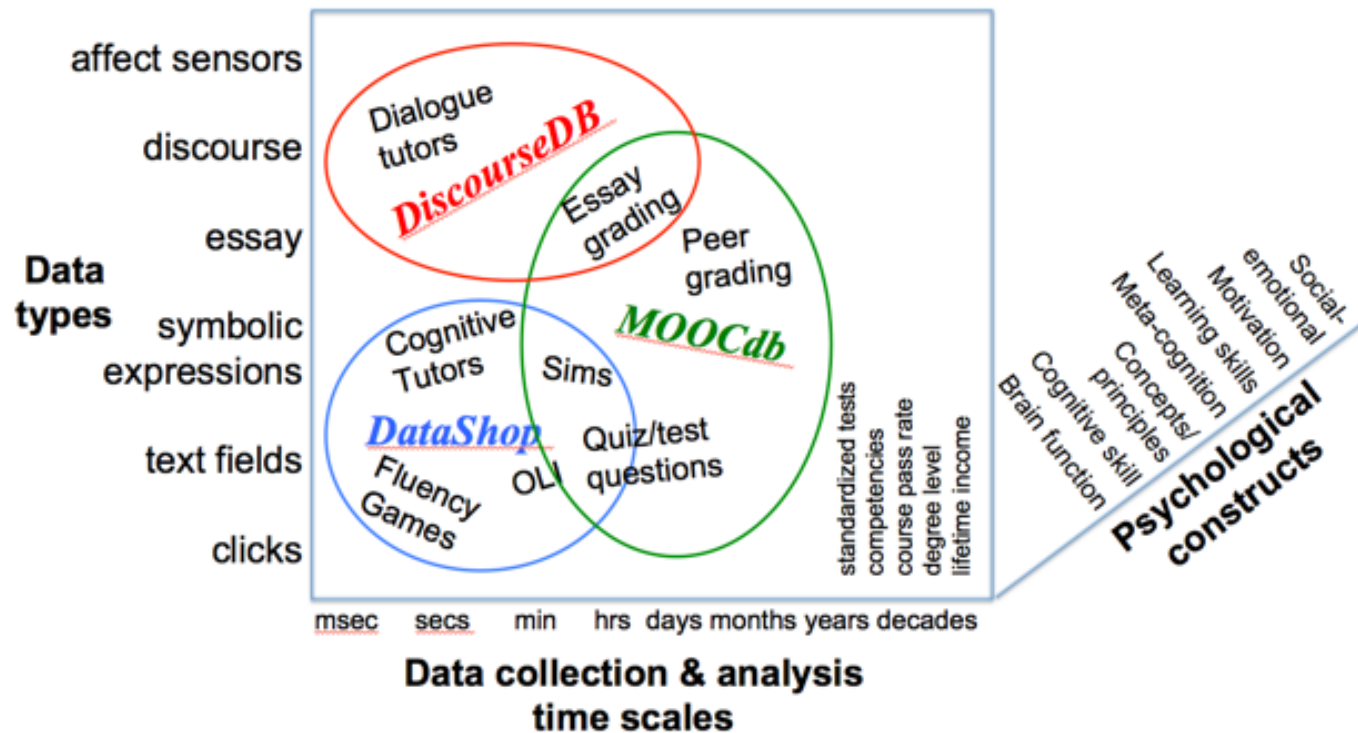
Problem: Data in **separate silos**

Click stream data in CMU's DataShop

MOOC analytics in MIT's MOOCdb

MOOC data in Stanford's DataStage

Language & discourse data in CMU's new DiscourseDB



LearnSphere *Solution*

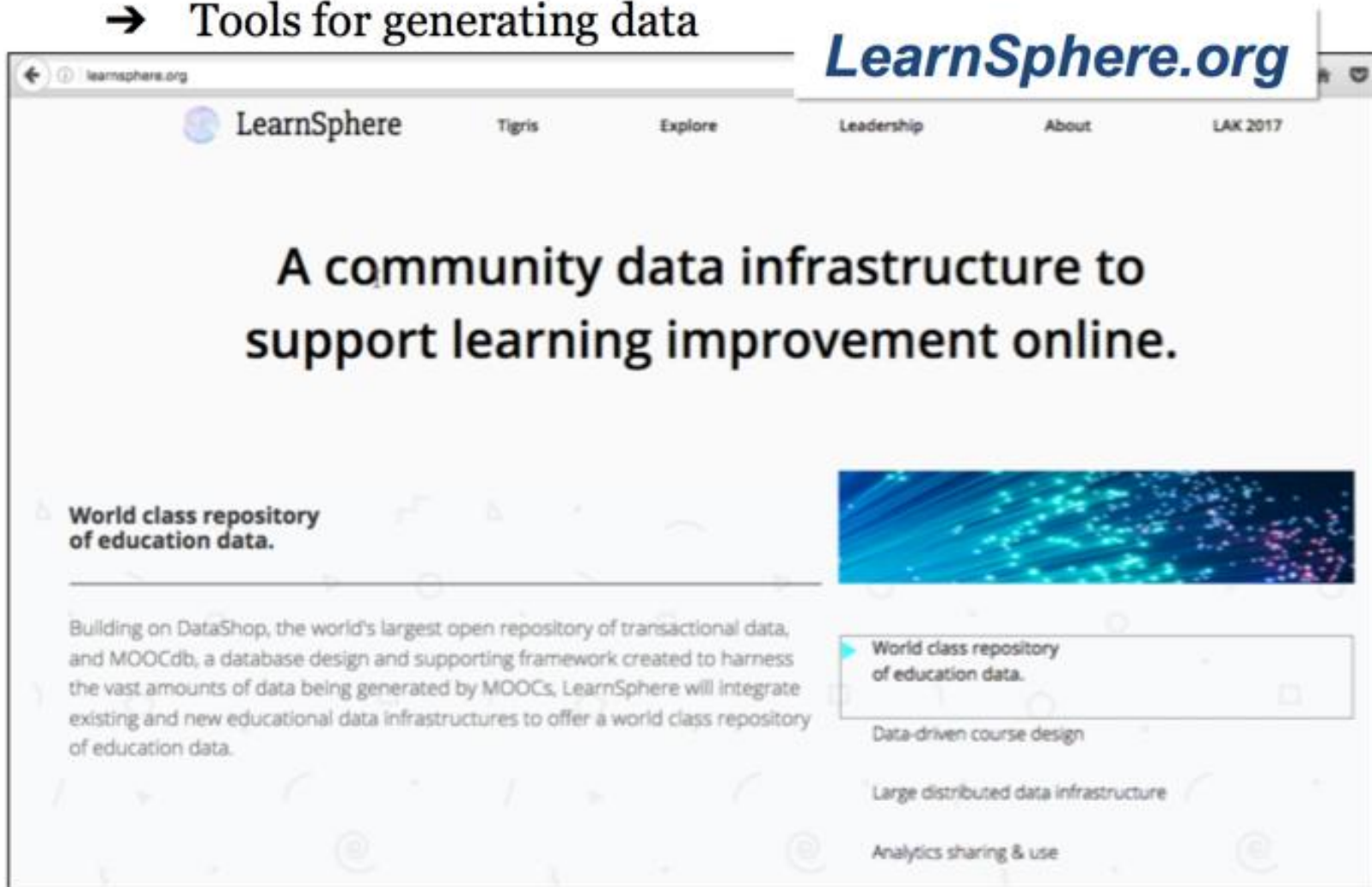
Web-based portal & workflow tool integrates DIBBs

Facilitate discoveries not possible within silos

LearnSphere.org: Web-based Portal

Hub for sharing:

- Learning data
- Learning analytic methods
- Tools for generating data



The screenshot shows the homepage of LearnSphere.org. The browser address bar displays 'learnsphere.org'. The website header includes the LearnSphere logo and navigation links for 'Tigris', 'Explore', 'Leadership', 'About', and 'LAK 2017'. The main heading reads: 'A community data infrastructure to support learning improvement online.' Below this, there is a section titled 'World class repository of education data.' followed by a paragraph: 'Building on DataShop, the world's largest open repository of transactional data, and MOOCdb, a database design and supporting framework created to harness the vast amounts of data being generated by MOOCs, LearnSphere will integrate existing and new educational data infrastructures to offer a world class repository of education data.' To the right of this text is a decorative image of blue and green light trails. Below the image is a list of features: 'World class repository of education data.', 'Data-driven course design', 'Large distributed data infrastructure', and 'Analytics sharing & use'.

LearnSphere.org

LearnSphere Tigris Explore Leadership About LAK 2017

A community data infrastructure to support learning improvement online.

World class repository of education data.

Building on DataShop, the world's largest open repository of transactional data, and MOOCdb, a database design and supporting framework created to harness the vast amounts of data being generated by MOOCs, LearnSphere will integrate existing and new educational data infrastructures to offer a world class repository of education data.

- World class repository of education data.
- Data-driven course design
- Large distributed data infrastructure
- Analytics sharing & use

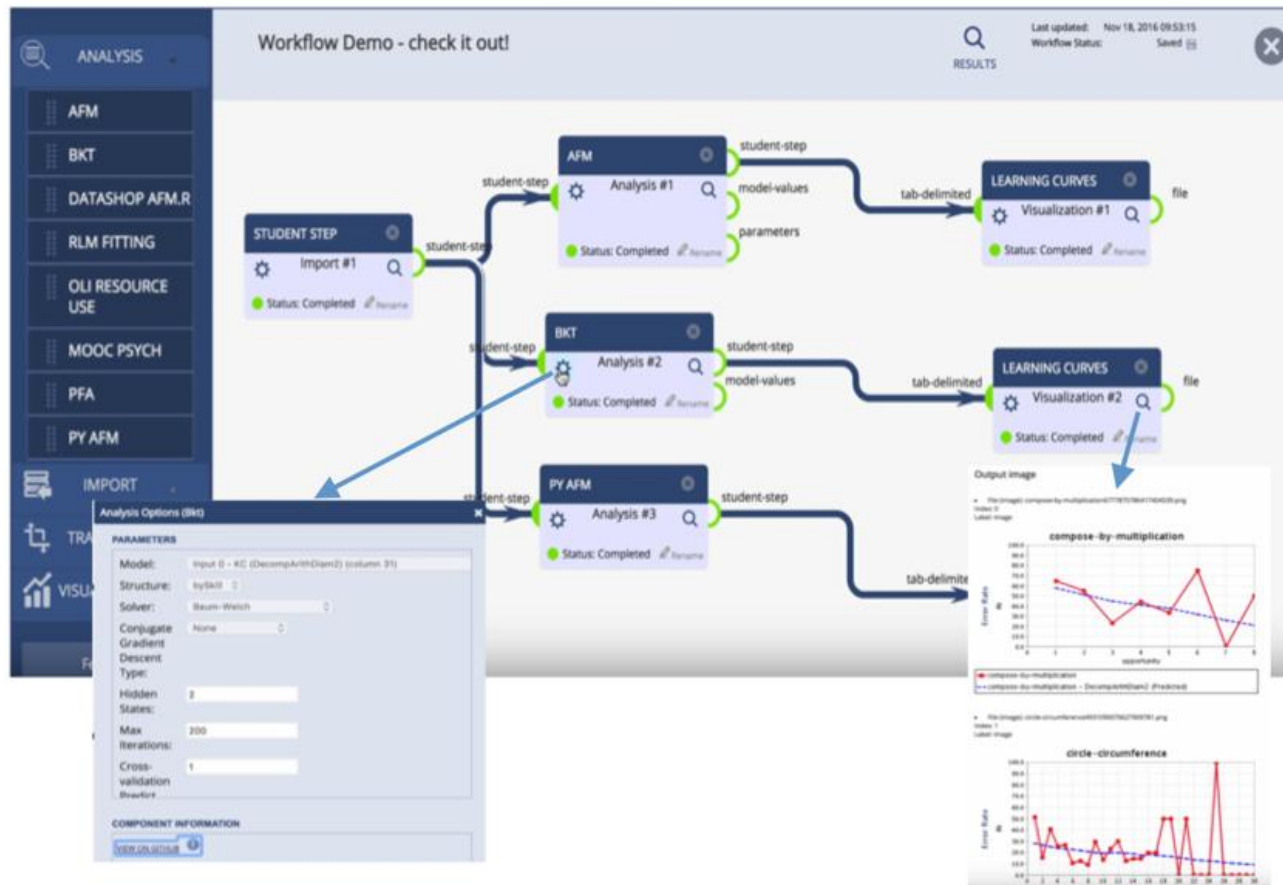
LearnSphere's Workflow Authoring

Web-based workflow authoring

- Method & data sharing & cross indexing
- **Emergent data standards** from convergent use

Easy for researchers, course developers, & instructors

- engage in learning analytics & ed data mining
- **without programming** skills
- component implementations in Java, C, R, Python, Matlab



Discovery Example

Workflow analytics that *bridge data silos* from 3 sources:

- tutor interaction, MOOC resource use and outcomes, & discussion boards

What student behaviors are associated with greatest learning?

- 4 courses with 5M interactions from 12K students
- **Striking discovery:** *6x better learning outcomes from active rather than passive learning*
 - ◆ Active = answering questions with feedback
 - ◆ Passive = lecture watching or text reading

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.*

Developments & Innovations

Doubled data sets available to over 1,300

Developed *DiscourseDB* from scratch

New MOOCdb capabilities

Distributed version of DataShop demonstrated at Memphis

Challenges & Future Directions

1. Uniformity-complexity challenge => *emergent standards*
 - Broad R&D community => no single data schema
 - *Uniformity* toward maximum reusability *with*
 - *flexibility* in representations to adapt to user needs
2. Sharing-privacy challenge
 - Maximize sharing of human data
 - *without* sacrificing student privacy
3. Sophistication-understanding trade-off
 - Advance sophistication & variety of analytic DIBBs
 - *yet* maintain understanding & trust
4. *Flexible “need for speed”*
 - Some jobs impractical within default processing
 - Transparent integration of cloud services needed

Future Directions

- Computer science is part & parcel of doing science
 - Computational biology, computational chemistry, computational X ...
- Publications are *not* just in English
 - Scientific insight is communicated through runnable models on available data