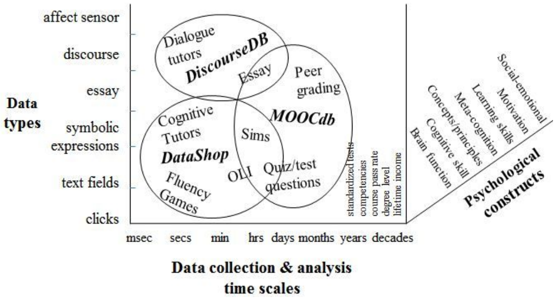
LearnSphere: Data-Driven Discovery and Innovation in Education

CMU: Ken Koedinger, John C. Stamper, Carolyn Rose
MIT: Kalyan Veeramachaneni, Una-May O'Reilly
Stanford: Candace Thille; U Memphis: Phil Pavlik

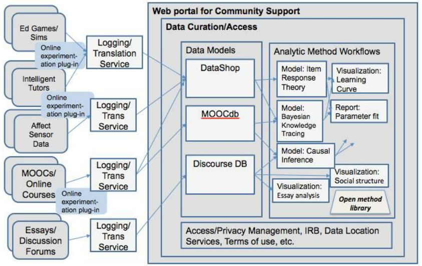
Big data in education *opportunity*
Problem: Data in **separate silos**

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 is funded by NSF award CISE-ACI-1443068.

LearnSphere.org: Web-based Portal

Hub for sharing

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

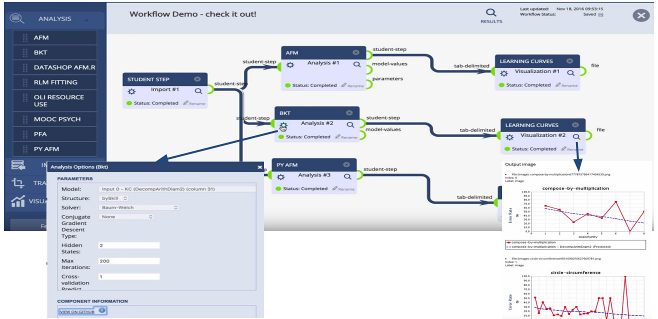
LearnSphere's Workflow Authoring

Web-based workflow authoring

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

- ➔ 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

Developments & Innovations

Doubled data sets available to over 1,300
Developed *DiscourseDB from scratch*
New MOOCdb capabilities
Distributed version of DataShop demonstrated at Memphis
Released workflow authoring tool

- ➔ Workflow components *contributed by many*
- ➔ Published results associated with analytics

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