Computational Modeling and Data Mining Thrust

Kenneth R. Koedinger
Human-Computer Interaction
& Psychology
Carnegie Mellon University

Marsha C. Lovett
Eberly Center for Teaching Excellence
& Psychology
Carnegie Mellon University

From 3 Clusters to 4 Thrusts

Clusters
1. Interactive Communication
2. Coordinative Learning
3. Refinement & Fluency

Thrusts
1. Social Communicative Factors
2. Metacognitive & Motivation
3. Cognitive Factors
4. Computational Modeling & Data Mining

Motivation

• Transformative Opportunity of Technology
  – Key to 21st century education
  – Directly benefits education PLUS
  – Facilitates collection of vast data on learning that will dramatically accelerate the science of academic learning.

• PSLC Data Shop offers rich resource
  – Today
    • Vast amount of data already (see next)
    • Multiple measures of task performance, reasoning & problem solving & learning
  – Future
    • 100x more data in 5 years!
    • Multiple measures of motivation & metacognition

DataShop score card: Vast amount of free data!

<table>
<thead>
<tr>
<th>Domain</th>
<th>Data-sets</th>
<th>Papers linked to DS</th>
<th>Student Actions</th>
<th>Students</th>
<th>Student Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>50</td>
<td>8</td>
<td>2,300,000</td>
<td>2,684</td>
<td>5,000</td>
</tr>
<tr>
<td>Math</td>
<td>50</td>
<td>25</td>
<td>15,200,000</td>
<td>5,996</td>
<td>68,000</td>
</tr>
<tr>
<td>Science</td>
<td>21</td>
<td>11</td>
<td>2,900,000</td>
<td>3,267</td>
<td>16,000</td>
</tr>
<tr>
<td>other</td>
<td>17</td>
<td>13</td>
<td>1,500,000</td>
<td>2,669</td>
<td>8,000</td>
</tr>
<tr>
<td>Total</td>
<td>138</td>
<td>57</td>
<td>21,800,000</td>
<td>14,616</td>
<td>97,000</td>
</tr>
</tbody>
</table>
Plan

- Review relevant AB suggestions & status
- Describe CMDM high-level goals

- Breakout:
  - Probe goals
    - Illustrate with on-going work (as needed)
    - Discuss pros & cons of proposed work

Relevant Advice from the 2008 Advisory Board Meeting

- Extend PSLC work on the microgenetics of learning, such as data mining of event logs and development of DataShop tools, to apply to the field of assessing student learning.

- Expand current studies to include longitudinal research on students over time.

Assessment Project

- On-line assessment system that teaches as it tests
- Data from instructional interactions used to estimate end-of-year high stakes state test result

- Results
  - Reliably better prediction using interaction data
  - Model based only on interaction info makes better predictions than the traditional assessment model (only uses correctness)

Help-seeking tutor: Lasting effects of assessment & feedback!

- Roll, Aleven, McLaren, Koedinger

- Longitudinal:
  - Over 4 months
- Effects of help seeking tutor used in 2 units persists in future units
  - Students are better help-seekers even after immediate support has been removed
Other Metacognitive Assessment

- Sub-vocal self-explanation detector (Shih)
  - Individual differences in time after “bottom-out” hints predict learning!
- Gaming the system detectors (Baker)
  - General detector shown to work across different math courses & tutor units
  - Gaming is a state, not a trait, better predicted by features of curriculum than student
- Peer collaboration skill detector (Walker)
  - Language analysis of chat text can distinguish statements of tutor & tutee that are productive or not

Longitudinal Studies

Mostly within school year or semester so far
- Already mentioned
  - Assistments (Heffernan, Junker, Koedinger)
    - Months of data to predict spring standardized test
    - Embedded assessment in 8th grade predicts 10th grade test scores as well as the 8th grade test does
- On-going & planned
  - Mizera ESL study – across 3 semesters
    - Dev of L2 oral fluency can be tracked through increase in "formulaic sequences"
  - Tracking fluency prerequisites & effect on pre-algebra learning (Pavlik, Cen, Koedinger)
  - SC thrust – accountable talk analysis in class dialogs (Resnick, Rose)

Other Ed Data Mining News since last year

- Leadership in educational data mining
  - First Educational Data Mining Conference
    - Organized by Ryan Baker et al
  - New: Journal of Educational Data Mining
    - Baker is an Associate Editor
  - Coming: Handbook of Educational Data Mining
    - Several PSLC chapters
- Related on-going projects
  - Learning Factors Analysis (Cen, Koedinger & Junker, 06) in Geo
  - Improved Cognitive Task Analysis in Physics (van de Sande)
  - Transfer-enabling knowledge components (Hausmann, Nokes)
    - identify KCs common to both translational & rotational kinematics
    - Use to design self explanation & analogical comparison intervention

Focal Questions of this Thrust

1. How can we generate accurate cognitive models of students’ *domain-specific* knowledge?
2. What models of *domain-general* processes best capture student learning?
   - learning & metacognition
   - motivation & affect
   - social aspects and instructional talk
3. By integrating domain-specific & -general models into *predictive models*, how can we *engineer* instructional interventions with big impact?
Focal Research Questions:
Anticipated Outcomes

1. Cognitive models of domain-specific knowledge
   - Machine learning: New discovery algorithms, scale, efficiency
   - Learning science:
     - Produce better cognitive models for most of 90+ units/chapters across LearnLab courses
     - Use models to design provably better instruction
     - Conduct in vivo experiments to verify
2. Models of domain-general processes in learning
   - High fidelity SimStudent models that predict which of alternative instructional approaches yields better learning
   - Models (detectors) of motivation and affect that capture student's states accurately and create adaptive instruction
3. Engineering models
   - Specify Assistance Dilemma formula for ~ 5 dimensions
   - Show match to learning data
   - Generate and test novel predictions/instructional treatments

BREAK-OUT DISCUSSION --
Supporting slides as needed

Focal Questions of this Thrust

1. How can we generate accurate cognitive models of students’ domain-specific knowledge?
2. What models of domain-general processes best capture student learning?
   - learning & metacognition
   - motivation & affect
   - social aspects and instructional talk
3. By integrating domain-specific & -general models into predictive models, how can we engineer instructional interventions with big impact?

Domain-Specific Cognitive Models

- Question: How do students represent knowledge in a given domain?
- Answering this question involves deep domain analysis
- The product is a cognitive model of students’ knowledge
Discovering Knowledge Representations

- **Knowledge decomposability** hypothesis
  - Acquisition of academic competencies can be decomposed into units, called knowledge components, that yield accurate predictions about student task performance & transfer of learning
- Scientific importance: Not obviously true
  - “learning, cognition, knowing, and context are irreducibly co-constituted and cannot be treated as isolated entities or processes” (Barab & Squire, 2004)
- Practical importance: Optimal instructional design depends on deep understanding of domain knowledge

Future Goals in Discovering Domain Models

1. Improve model-discovery methods
   - Partial Order Knowledge Structures (POKS)
   - Exponential-family Principle Component Analysis
2. Improve human-machine interaction
   - Better process for task difficulty factor labeling
3. Show models yield improved student learning

Using learning curve data to evaluate knowledge component models

- Without decomposition, using just a single “Geometry” KC, no smooth learning curve.
- But with decomposition, 12 KCs for area concepts, a smooth learning curve.

Upshot: A decomposed KC model better fits learning data

Domain modeling projects

- Domain model discovery algorithm invention
  - LFA vs. ePCA (Cen, Singh, Gordon, Koedinger)
  - POKS, LFA, vs. PFA (Pavlik, Cen, Koedinger)
  - Clustering vs. IRT (Ayers, Nugent, Junker)
  - Time series, state-space models
- Computer science issues
  - Algorithm invention; software optimization
- Use of tools/algorithms by domain researchers
  - Van der Sante, Hausmann in Physics kinematics; Wylie in English article use; Matsuda in Algebra equation errors; Perfetti et al in Chinese; Lovett in Statistics
- Models yield improve learning
  - Pre-algebra conceptual prerequisites (Pavlik, Koedinger)
Focal Questions of this Thrust

1. How can we generate accurate cognitive models of students’ domain-specific knowledge?

2. What models of domain-general processes best capture student learning?
   - Learning & metacognition
   - Motivation & affect

3. By integrating domain-specific & -general models into predictive models, how can we engineer instructional interventions with big impact?

Models of domain-general processes

- Learning processes
  - SimStudent learns from algebra tutor (Matsuda et al.)

- Metacognition
  - Model of domain-general help-seeking (Aleven et al.)

- Motivation & affect
  - Using classroom observation & data mining to build detectors of motivation & affect (Baker)

Future domain-general model projects

- Models of learning, SimStudent
  - Is “weak” prior knowledge key to both domain-general learning & learner misconceptions? (Matsuda, Koedinger)

- Longitudinal models of affect & motivation
  - Detect affect & motivational behaviors (e.g., gaming the system, boredom, self-efficacy) over time (Baker)
  - Predict metacognition & learning

- Investigate relationships across data sets, domains, classrooms, teachers, & schools
  - Baker, Pavlik, Matsuda, Koedinger

Focal Questions of this Thrust

1. How can we generate accurate cognitive models of students’ domain-specific knowledge?

2. What models of domain-general processes best capture student learning?
   - Learning & metacognition
   - Motivation & affect

3. By integrating domain-specific & -general models into predictive models, how can we engineer instructional interventions with big impact?
**Assistance Dilemma: A Fundamental Unsolved Problem**

- "How should learning environments balance information or assistance giving and withholding to achieve optimal student learning?"
  
  - Koedinger & Aleven, 2007

<table>
<thead>
<tr>
<th>Instructional Support</th>
<th>Poor Learning Outcome</th>
<th>Good Learning Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>High assistance (less demanding)</td>
<td>crutch</td>
<td>scaffold</td>
</tr>
<tr>
<td>Low assistance (more demanding)</td>
<td>undesirable difficulty; extraneous load</td>
<td>desirable difficulty; germane load</td>
</tr>
</tbody>
</table>

**General plan of attack for the immense challenge**

1. **Decompose**: Identify & distinguish relevant dimensions of assistance
   - On-going: Practice spacing, practice timing, study-test, example-problem
   - Potential: Concrete-abstract, do-explain, immediate-delayed feedback, low-high variability, block-space, ...

2. **For each(1) dimension**
   1. **Integrate**: Collect & integrate relevant literature
   2. **Mathematize**: Characterize conditions, parameters, equations in precise predictive model
   3. **Test**: Make a priori predictions & test in experiments

**Inverted U for practice-interval dimension**

- Precise predictive formula

\[
\text{eff}_m = \frac{p_m b_{\text{sec}} g_m + (1-p_m) b_{\text{fail}} g_m}{p_m (t_m + \text{fixedsuccess}) + (1-p_m) \text{fixedfailcosts}}
\]

- \(\text{eff}_m\) = efficiency of robust learning
- \(p_m\) = probability of recall success
- \((1-p_m)\) = probability of recall failure
- \(b_{\text{sec}}\) = gain from success
- \(b_{\text{fail}}\) = gain from failure
- \(g_m\) = long-term increase in activation
- \(t_m\) = time of recall
- \(tsc\) = time for success
- \(tfc\) = time for failure

**General form of assistance formula**

For each learning event:

\[
\text{Robust learning efficiency gain} = \frac{p \cdot \text{benefit-of-success} + (1-p)\text{benefit-of-failure}}{p \cdot \text{cost-of-success} + (1-p)\text{cost-of-failure}}
\]

\(p\) = Probability of success during instruction
Future Engineering Modeling projects

- Instantiate equation & fit to data sets for 4 dimensions (Pavlik, Koedinger)
  - Practice spacing, practice timing, study-test, example-problem
- Collect missing data on example-problem dimension (Salden, Aleven, McLaren)
  - Parameterize adaptive example-fading
- Collect missing data on do-explain dimension (Wylie, Mitamura, Koedinger)

Summary of Anticipated Outcomes

1. Cognitive models of domain-specific knowledge
   - Machine learning: New discovery algorithms, efficiency
   - Learning science:
     - Produce better cognitive models for most of 90+ units/chapters across LearnLab courses
     - Use models to design provably better instruction
     - Conduct in vivo experiments to verify
2. Models of domain-general processes in learning
   - High fidelity SimStudent models that predict which of alternative instructional approaches yields better learning
   - Models (detectors) of motivation and affect that capture student's states accurately and create adaptive instruction
3. Engineering models
   - Specify Assistance Dilemma formula for ~ 5 dimensions
   - Show match to learning data
   - Generate and test novel predictions/instructional treatments

Possible Questions for the AB

- What aspects of the domain modeling are potentially interesting to the broader cognitive/learning science or psychometric audiences?
  - This is a quantitative approach to domain analysis -- can it be coupled with qualitative approaches like protocol or discourse analysis? Pros and cons?
- For some of us, the Barab quote is hard/impossible to make sense?
  - What does it mean? How to make progress in the field?
  - Better demonstrates of integrative knowledge components?
  - Better demonstrations of interactions with affect?
- Feedback on Assistance Dilemma agenda
  - Is this too big? Will this have traction?
  - Need to address cross dimension as well as within?