Modeling Complex Skill with Educational Data Mining

Ryan S. Baker
Teachers College, Columbia University
University of Pennsylvania
Modeling and Detecting Complex Student Skill

• In recent years, there has been an explosion of interest in inferring student inquiry and complex problem-solving skill automatically
A lot of this work

• Has used knowledge engineering and traditional psychometric approaches to isolate and measure desired behaviors

• One modern paradigm for this is Evidence-Centered Design (ECD) (Mislevy et al., 2013)
This approach

• Precisely define desired behaviors
• Create problem-solving situations where students can demonstrate these behaviors
Example

• Efforts to measure VOTAT – Vary One Thing at a Time (cf. Vollmeyer & Rheinberg, 1999; Greiff, Wustenberg, & Funke, 2012)
  – Also called CVS – Control for Variables Strategy (Chen & Klahr, 1999)
MicroDYN (Greiff & Funke, 2014)
The great thing about knowledge engineering

• The great thing about knowledge engineering is that you know exactly what you’ve got

• The models are
  – Clearly and cleanly defined
  – Can be communicated, debated, and ultimately agreed on by a community
The problem with knowledge engineering

• The problem with knowledge engineering is that it’s hard to use it to define the most complex skills
The problem with knowledge engineering

• The problem with knowledge engineering is that it’s hard to use it to define the most complex skills
  – Skills that can manifest appropriately in several ways
The problem with knowledge engineering

• The problem with knowledge engineering is that it’s hard to use it to define the most complex skills
  – Skills that can manifest appropriately in several ways
  – Skills that are compensatory, e.g. there are different paths to the same result
The problem with knowledge engineering

• The problem with knowledge engineering is that it’s hard to use it to define the most complex skills
  – Skills that can manifest appropriately in several ways
  – Skills that are compensatory, e.g. there are different paths to the same result
  – Skills where boundaries between correct and incorrect performance are fuzzy
Today’s Talk

• I will discuss how the methods of educational data mining

• Can support richer assessment of student inquiry skill and complex problem-solving skill
“the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs.”
Broad range of methods...

• Today’s talk will focus on prediction modeling and latent knowledge estimation methods

• For a general overview of EDM, see


Two parts

1. Microworlds: Beyond VOTAT
2. Science Inquiry in MUVEs
Two parts

1. Microworlds: Beyond VOTAT
2. Science Inquiry in MUVEs
One Challenge

• One challenge in applying rules like VOTAT/CVS has been that students often engage in
  – some behavior that looks like VOTAT/CVS
  – and some behavior that doesn’t
VOTAT/CVS?

• Changing two variables, changing one back
  – 50% VOTAT/CVS? 100% VOTAT/CVS?

• Changing three variables, changing two back
  – 50% VOTAT/CVS? 0% VOTAT/CVS?

• What if a student haphazardly changes the simulation parameters 18 times, always changing two or more variables, but systematically explores the variable space across trials?
  – 100% VOTAT/CVS? 0% VOTAT/CVS?

• What if a student explores the simulation but then gets down to business?
What rule should we use?

• Num. sequential controlled trials? (McElhaney & Linn, 2010)

• Avg. controlled trials over all trials? (Harrison & Schunn, 2004)

• How well do the rules assess?
Educational Data Mining Approach

• Machine learns rules based on student log data
  – Determine relationships and cut-offs from data, not ad-hoc
  – Verify model goodness and generalizability with
    • held-out test sets
    • student-level cross-validation
Student-Level Cross-Validation
Context: Science ASSISTments/Inq-ITS (Gobert et al., 2012)

• Web-based learning environment where Middle School students conduct inquiry within simulations for Physical, Life and Earth Science

• Assess, track, and scaffold inquiry skills

• Now a commercial product sold by Apprendis, Inc.
Science ASSISTments
Phase Change Activity
Goal: Determine how one variable you choose affects the boiling point of ice

HYPOTHESIZE: Build a testable hypothesis about the boiling point of ice. ...

My Hypothesis  If I change the Choose One... so that it Choose One...
the Choose

I need to explore more

I'm ready to run an experiment
Goal: Determine how one variable you choose affects the boiling point of ice

EXPERIMENT: Collect data to help you test your hypothesis. ... more

My Hypothesis

If I change the amount of ice so that it decreases, the boiling point decreases.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount of heat</td>
<td>High</td>
</tr>
<tr>
<td>amount of ice</td>
<td>100 grams</td>
</tr>
<tr>
<td>container cover</td>
<td>cover</td>
</tr>
<tr>
<td>size of the container</td>
<td>Large</td>
</tr>
</tbody>
</table>
Goal: Determine how one variable you choose affects the boiling point of ice

EXPERIMENT: Collect data to help you test your hypothesis. ... more

My Hypothesis

If I change the amount of ice so that it decreases, the boiling point decreases.

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Has Cover</th>
<th>Container Size</th>
<th>Heat Level</th>
<th>Liquid Amount</th>
<th>Melting Temp(°C)</th>
<th>Boiling Temp(°C)</th>
<th>Time(min) Melting</th>
<th>Time(min) Boiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>200 grams</td>
<td>0</td>
<td>100</td>
<td>10</td>
<td>68.75</td>
</tr>
<tr>
<td>3</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>100 grams</td>
<td>0</td>
<td>100</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>true</td>
<td>Large</td>
<td>High</td>
<td>100 grams</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Goal: Determine how one variable you choose affects the boiling point of ice

ANALYZE DATA: Determine if the data you collected support your hypothesis. ...

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Has Cover</th>
<th>Container Size</th>
<th>Heat Level</th>
<th>Liquid Amount</th>
<th>Melting Temp(°C)</th>
<th>Boiling Temp(°C)</th>
<th>Time(min) Melting</th>
<th>Time(min) Boiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
<td>16.25</td>
<td>102.5</td>
</tr>
<tr>
<td>2</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>200 grams</td>
<td>0</td>
<td>100</td>
<td>10</td>
<td>68.75</td>
</tr>
<tr>
<td>3</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>100 grams</td>
<td>0</td>
<td>100</td>
<td>5</td>
<td>35</td>
</tr>
</tbody>
</table>

Drag trials used in your analysis from here to the evidence table below.

Analysis

My Hypothesis
If I change the amount of ice so that it decreases, the boiling point decreases.

When I changed the amount of ice so that it decreased, the boiling point of the object did not change. This means that my data do not support my hypothesis.
**Goal:** Determine how one variable you choose affects the boiling point of ice

**ANALYZE DATA:** Determine if the data you collected support your hypothesis. ...

### Trial Data

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Has Cover</th>
<th>Container Size</th>
<th>Heat Level</th>
<th>Liquid Amount</th>
<th>Melting Temp(°C)</th>
<th>Boiling Temp(°C)</th>
<th>Time(min) Melting</th>
<th>Time(min) Boiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
<td>16.25</td>
<td>102.5</td>
</tr>
</tbody>
</table>

### Evidence Table

**Evidence Table:** Drag in the trials used to support your analysis from the trial table above. To remove a trial from the evidence table, click on it.

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Has Cover</th>
<th>Container Size</th>
<th>Heat Level</th>
<th>Liquid Amount</th>
<th>Melting Temp(°C)</th>
<th>Boiling Temp(°C)</th>
<th>Time(min) Melting</th>
<th>Time(min) Boiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
<td>16.25</td>
<td>102.5</td>
</tr>
<tr>
<td>2</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>200 grams</td>
<td>0</td>
<td>100</td>
<td>10</td>
<td>68.75</td>
</tr>
</tbody>
</table>

Go back. I need more data.

I'm done with analysis!
Data

• We have studied these issues with data from several studies conducted in urban and suburban schools in Central Massachusetts

• Primary data set I’ll discuss today is 148 students in suburban Massachusetts using Phase Change Activity

• Results also replicated on other activities and populations; I won’t talk as much about that today (Sao Pedro et al., 2012, 2013, 2014)
Our Models Can Detect

• Designing Controlled Experiments

• Testing Stated Hypothesis

• WTF Behavior (Wixon et al., 2012)

• Carelessness (Hershkovitz et al., 2012)
Today I’ll Focus On

• Designing Controlled Experiments
  – Includes VOTAT, but also other data collection behaviors that produce data that allows non-confounded inference
  – Assessed at grain-size of overall experimental sequence, rather than pairs of trials
Text Replay Tagging
(Baker et al., 2006; Sao Pedro et al., 2010)

• Leverages humans’ ability to recognize skills
• Human coders label segments of log files
  – Sampled from full data set, stratified so that
    • every student represented
    • full history of behavior represented
• Labels are treated as “ground truth” for whether
  student demonstrated skills
• 570 clips tagged for this research
  – Drawn from data from 148 eighth grade students,
    suburban public middle school in Massachusetts
  – Students completed 4 activities each
Hypothesis

- time since start: 0 Hypothesis iv: current=“Level of heat” previous=“unknown”
- time since start: 2 Hypothesis iv direction: current=“increases” previous=“unknown”
- time since start: 6 Hypothesis dv: current=“time the ice took to melt” previous=“unknown”
- time since start: 8 Hypothesis dv direction: current=“decreases” previous=“unknown”
- time since start: 12 Adding hypothesis: iv=“Level of heat” iv direction=“increases”
  dv=“time the ice took to melt” dv direction=“decreases”

time since start: 16 Experiment

- time since start: 22 Variable change: “Level of heat” current=“High” previous=“Low”
- time since start: 23 Run: Level of heat=“High”, Cover=“Off”, Container Size=“Small”, ...
- time since start: 27 Reset Experiment

- time since start: 30 Variable change: “Level of heat” current=“Med” previous=“High”
- time since start: 32 Run Experiment: Level of heat=“Med”
- time since start: 51 Reset Experiment
- time since start: 52 Run: Level of heat=“Med”, Cover=“Off”, Container Size=“Small”, ...

<<Clip Boundary>>


Done
(0s) Hypothesis

....

(61s) Adding hypothesis: iv=“Container Size”; iv direction=“increases”; dv=“melting point”; iv direction=“increases”

(63s) Experiment

(69s) Run Experiment: Amount of Substance=“300 g”; Container Size=“Large”
Cover=“on”; Level of heat=“Low”

...several variable changes...

(106s) Run Experiment: Amount of Substance=“200 g”; Container Size=“Medium”
Cover=“off”; Level of heat=“Medium”

...several variable changes and a couple of simulation runs...

(140s) Run Experiment: Amount of Substance=“300 g”; Container Size=“Large”
Cover=“on”; Level of heat=“High”

(146s) Simulation complete

(153s) Simulation reset

(155s) Variable Change: Container Size: current=“Medium” previous=“Large”

(156s) Run Experiment: Amount of Substance=“300 g”; Container Size=“Medium”
Cover=“on”; Level of heat=“High”

...ran another trial changing only container size...

(186s) Variable Change: Container Size: current=“Small” previous=“Large”

(187s) Run Experiment: Amount of Substance=“300 g”; Container Size=“Medium”
Cover=“on”; Level of heat=“High”

<< CLIP BOUNDARY >>
Inter-rater agreement

• Designing controlled experiments
  – Cohen’s Kappa = 0.69
Example Features (out of 73 distilled)

Counts
- Hypothesis field changes
- Hypotheses made
- IV changes
- Runs
- Pauses
- Complete runs
- Incomplete runs
- Pairwise controlled trials (no repeats)
- Repeat trials

Time
- Min, max
- Mean, mode, SD
  Computed for each action
Model Development Approach

- Features selected using combination of goodness of fit and construct validity as assessed by domain expert (Sao Pedro et al., 2012)

- Models tested using held-out test set

- J48 Decision Trees (Quinlan, 1993) built in RapidMiner 4.6
  - Search for qualitative cut-offs in data
Model Assessment

• Models assessed using

• $A'$
  – The model’s ability to distinguish when a behavior is present (e.g. is student designing controlled experiments or not)
  – Chance = 0.5, Perfect = 1.0, First-level medical diagnostics $> 0.8$

• Cohen’s Kappa
  – The degree to which the model is better than base rate
  – Base Rate = 0, Perfect = 1.0
Model Goodness (Most Recent Version)

- Designing Controlled Experiments
  - $A' = 0.94$
  - Kappa = 0.45
  - Can be used after three runs of simulation
Resultant Models are Difficult to Inspect...

**Figure 5.** Portion of the decision tree for the designing controlled experiments behavior.
But replicate human judgment even when applied to...

- New students (Sao Pedro et al., 2012)
- New science microworlds (Sao Pedro et al., 2013)
Using models in Latent Skill Estimation
(Figuring out what student knows)

• Detectors enable inference of whether student is showing appropriate inquiry behavior
  – during a single period of data collection

• Can we assess student’s overall proficiency?
  – Aggregating across demonstrations of student skill over time
  – Taking into account that student skill may improve over time
Approach: Bayesian Knowledge Tracing (Corbett & Anderson, 1995)

- Mathematical model that infers knowledge based on correctness over time
- Estimates probability that errors were due to “slips” and probability that correct answers were due to knowing the skill
- In order to infer learning as it occurs
Approach: Bayesian Knowledge Tracing

• Success at estimating changing skill in mathematics tutors (e.g. Corbett & Anderson, 1995; Koedinger & Corbett, 2006; Baker, Corbett, & Aleven, 2008; Ritter et al., 2009; Feng, Heffernan & Koedinger, 2009; Baker, Corbett, Gowda, et al., 2010; Pardos, Heffernan, Anderson, & Heffernan, 2010)

• Can predict post-test performance (Corbett & Anderson, 1995; Baker et al., 2010), and preparation for future learning (Baker, Gowda, & Corbett, 2011)
Model Development Approach

• Apply detector to full data set
  – Labels every student experimentation sequence with probability that it represents Designing Controlled Experiments

• Fit latent estimation model using BKT-BF software package (Baker et al., 2010)

• Can model predict performance on experimentation sequence N...
  – From performance on sequences 1 through N-1?
Model Goodness

• Designing Controlled Experiments
  – $A' = 0.74$

• Note that $A'$ values are lower than for detectors
  – Predicting the future is harder than making inferences about the present!
Knowledge Transfer
(Sao Pedro et al., 2014)

• Student knowledge of inquiry demonstrated in one microworld, according to this model

• Predicts student knowledge in a different microworld
Knowledge Transfer
(Sao Pedro et al., 2014)

• If a student *learns* inquiry skill in one microworld

• They can then demonstrate the skill in a different microworld
Two parts

1. Microworlds: Beyond VOTAT
2. Science Inquiry in MUVEs
Can we also

• Assess inquiry and problem-solving skills in more open-ended learning environments?
Virtual Performance Assessment
(Clarke-Midura et al., 2011)

- Explore
- Make observations
- Gather data
- Solve a scientific problem in context
You need to investigate what caused the six legged frog.

- Gather evidence from the farms
- Use the internet kiosk for research
- Do tests on your evidence at the lab
- Talk with the locals

Return to me once you think you know what caused the six legged frog and have enough evidence to support it!
Gather Data
Conduct tests in the lab
Frog Mutations: Pesticides

Commonly used pesticides disrupt the development of frogs, weaken their immune systems, delay or stunt development, cause mutations including cancerous growths and limbs and otherwise contribute to declining frog populations.

Evidence

Water: A number of pesticides show up in pond water due to runoff from nearby fields. Common pesticides found include:

Blood: Pesticides concentrate in the blood and tissue cultures in frogs and can be seen in tests. At sufficient concentrations it will lower white blood cell count.

Tadpoles: Pesticides such as Atrazine stop the metamorphosis from tadpole to frog. Stunted growth will result in shorter tails in tadpoles when compared to healthy tadpoles.

Frogs: Pesticides can stunt the growth of frogs making them smaller than healthy frogs. They can grow cancerous growths and other mutations including limb growth. This is thought to be caused by a weakened immune system which makes the frogs susceptible to viruses and bacteria.
Build a final claim
Not feasible to infer

• Science inquiry skill in terms of skills like VOTAT or our extension of VOTAT

• Students collect information in several ways, as in real-world problem-solving and inquiry
Text Replays

• Also aren’t a clear approach
Not easy to know

• What the correct procedures even are

• Are only data collection behaviors relevant?

• Or are other behaviors like talking to NPCs or reading information resources relevant too?
Plus

• Behavior unfolds over long periods of time

• A student might conduct comparable trials 20 minutes apart

• Making text replays hard to create and hard to use
Alternate Strategy

• Judge inquiry strategies by results

• Rather than a priori ideas of what good inquiry is

• Or expert human judgments of what good inquiry is
In this case

• Look at which students were able to
  – Correctly infer why frogs had 6 legs
  – Provide a complete and clear explanation for the evidence they used to infer the causal process

• Look at the inquiry behaviors they used
  – To attempt to predict from inquiry behaviors whether a student will
    • provide the correct final conclusion
    • and be able to design a causal explanation
Data Set

- 1,985 students used VPA as part of science classes
  - 7\textsuperscript{th}-8\textsuperscript{th} graders
  - 40 teachers and 138 classrooms
    - Northeastern USA
    - Midwestern USA
    - Western Canada

- Until they were ready to provide a final answer and causal explanation for why frogs had 6 legs
- Taking an average of 29 minutes, 29 seconds
  - SD = 14 minutes, 30 seconds
48 Features Distilled

• Location Visits
  – Did the student go everywhere? How much time did he/she spend?

• Objects
  – Did the student pick everything up?

• Lab Tests
  – Which tests did the student run? Overall? In the same lab session?
  – How much time did the student take to self-explain their results?

• Information Pages
  – How long did student spend reading information pages?
Dependent Measures

• Correctness of final conclusion for why frogs had six legs (0,1)

• Skill in designing casual explanations for why that claim was correct
  – Points were assigned based on whether the evidence provided supported the claim made (even if it was an incorrect claim)
Detectors

• Correct Final Claim
• Designing Causal Explanations
Detector of Correct Final Claim
Detector of Correct Final Claim

• JRip Decision Rules
  – Student-level cross-validation

• Develops set of rules, executed one after the other, for choosing whether correct final claim or not
  – Good at finding distinct paths to the same end point
Detector of Correct Final Claim

• Model achieved cross-validated Kappa of 0.548

• Model achieved cross-validated $A'$ of 0.79

• Comparable to detectors of science inquiry in Inq-ITS that I discussed a few minutes ago
Overall Model
(built on full data set)

1. **IF** the student spent at least 66 seconds reading the parasite information page, **THEN** the student will obtain the correct final conclusion (confidence = 81.5%)

2. **IF** the student spent at least 12 seconds reading the parasite information page **AND** the student read the parasite information page at least twice **AND** the student spent no more than 51 seconds reading the pesticides information page, **THEN** the student will obtain the correct final conclusion (confidence = 75.0%)

3. **IF** the student spent at least 44 seconds reading the parasite information page **AND** the student spent under 56 seconds reading the pollution information page, **THEN** the student will obtain the correct final conclusion (confidence = 68.8%)

4. **OTHERWISE** the student will not obtain the correct final conclusion (confidence = 89.0%)
Notes

- Detector only uses data on students’ time spent reading specific information pages!
  - Did students spend enough time reading about the correct final conclusion?
  - Did students spend “too long” reading pages linked to incorrect hypotheses?
Notes

• Does this imply that only information pages matter, and that we can dispense with the virtual environment?
Detector of
Designing Causal Explanations
Detector of
Designing Causal Explanations

• Linear regression model with variable selection using M5’ (Wang & Witten, 1997)
  – Student-level cross-validation
Detector of Designing Causal Explanations

- Model achieved cross-validated correlation of 0.531
Individual feature most correlated to DCE

• The number of times the student accessed the information page on parasites

• Again indicates that student ability to DCE is strongly connected with other information-seeking behaviors
But other features also correlated (and significant in model)

- maximum degree of coverage for a lab test
- percentage of time spent at farms
- whether student conducted blood tests on 6-legged frogs or other frogs
- whether student conducted water tests on lab water or farm water
- whether student conducted genetic test on non-sick frogs
- maximum number of distinct non-sick frogs brought into the lab
Overall

• Getting the right answer turns out to be much simpler to predict

• Than whether students can design causal explanation

• Many factors associated with student ability to design causal explanation
Conclusions
Conclusions

• Data mining enables us to assess student inquiry skill/complex problem solving skill in a more flexible fashion

• Makes it possible to credit positive data collection behavior that isn’t quite VOTAT

• Makes it possible to identify successful inquiry behaviors even in very open-ended MUVEs
Uses
Uses: Next-Generation Assessment

• Assess student inquiry skill in stealth assessments such as games (Shute, 2013)
Uses: Discovery with Models

• Study how general student inquiry skills are – e.g. do they transfer across domains? (Sao Pedro et al., 2014)
Uses: Intervention

• Automated detectors used as basis of intervention in Inq-ITS by cartoon character Rex
Goal: Determine how one variable you choose affects the boiling point of ice

EXPERIMENT: Collect data to help you test your hypothesis. ... more

My Hypothesis
If I change the amount of ice so that it decreases, the time the ice takes to melt decreases.

<table>
<thead>
<tr>
<th>amount of heat</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount of ice</td>
<td>300 grams</td>
</tr>
<tr>
<td>container cover</td>
<td>cover</td>
</tr>
<tr>
<td>size of the container</td>
<td>Small</td>
</tr>
</tbody>
</table>

![Graph showing temperature over time with a melting point at 0°C and a boiling point at 100°C, with a time of 59 minutes for a specific trial.]

I think the data you're collecting won't help you test your hypothesis because **you aren't designing a controlled experiment.**

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Has Cover</th>
<th>Container Size</th>
<th>Heat Level</th>
<th>Liquid Amount</th>
<th>Melting Temp(°C)</th>
<th>Boiling Temp(°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>true</td>
<td>Large</td>
<td>Low</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>false</td>
<td>Medium</td>
<td>Low</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>true</td>
<td>Small</td>
<td>High</td>
<td>300 grams</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
Hey, are you just playing with the buttons? Take your learning seriously or I will eat you!!!
See our free online MOOT “Big Data and Education”
Will Re-Launch as edX MOOC in 6-9 months
All lab publications available online – Google “Ryan Baker”
Extra Slides
Bayesian Knowledge Tracing (Corbett & Anderson, 1995)

Two Learning Parameters

\[ p(L_0) \] Probability the skill is already known before the first opportunity to use the skill in problem solving.

\[ p(T) \] Probability the skill will be learned at each opportunity to use the skill.

Two Performance Parameters

\[ p(G) \] Probability the student will guess correctly if the skill is not known.

\[ p(S) \] Probability the student will slip (make a mistake) if the skill is known.
Extension for use with Science ASSISTments (Sao Pedro et al., 2013)

• Classical BKT assumes that answer is correct or incorrect \{0,1\}

• Science inquiry skill detectors give us probabilistic estimate of whether student behavior is appropriate or inappropriate \{number between 0 and 1\}