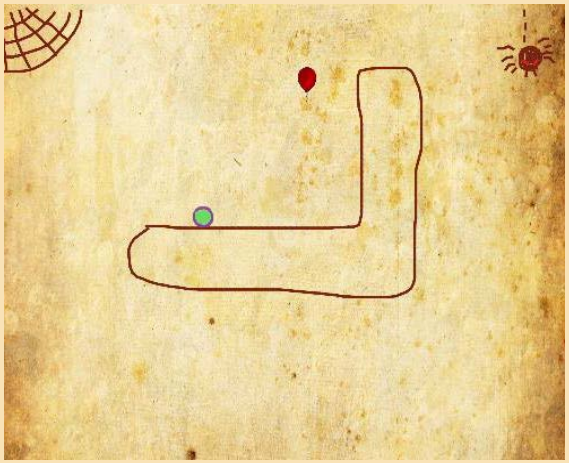
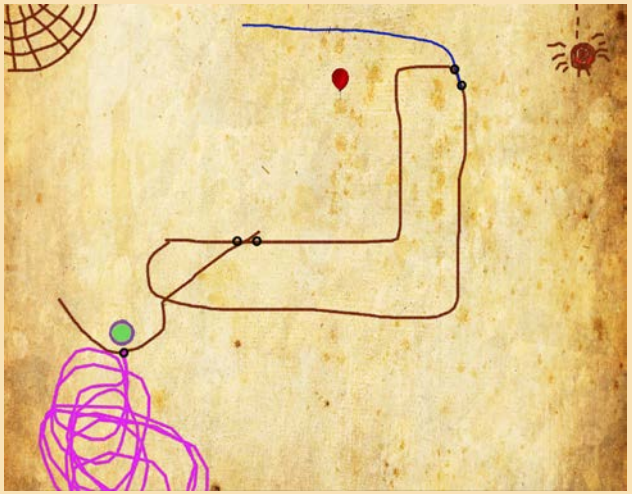
Debugging the Evidence Chain

Russell Almond, Yoon Jeon Kim, Valerie J. Shute, & Mathew Ventura
Florida State University
College of Education
Educational Psychology and Learning Systems
ralmond@fsu.edu
[1]With lots of help

Educational Assessment in a Game: *Newton's Playground*

- 2-D physics simulation game, similar to *Crayon Physics Deluxe*, based on Box2D physics engine for games
- Multiple levels. Goal of each level is to move ball to the balloon by drawing objects that obey the laws of physics





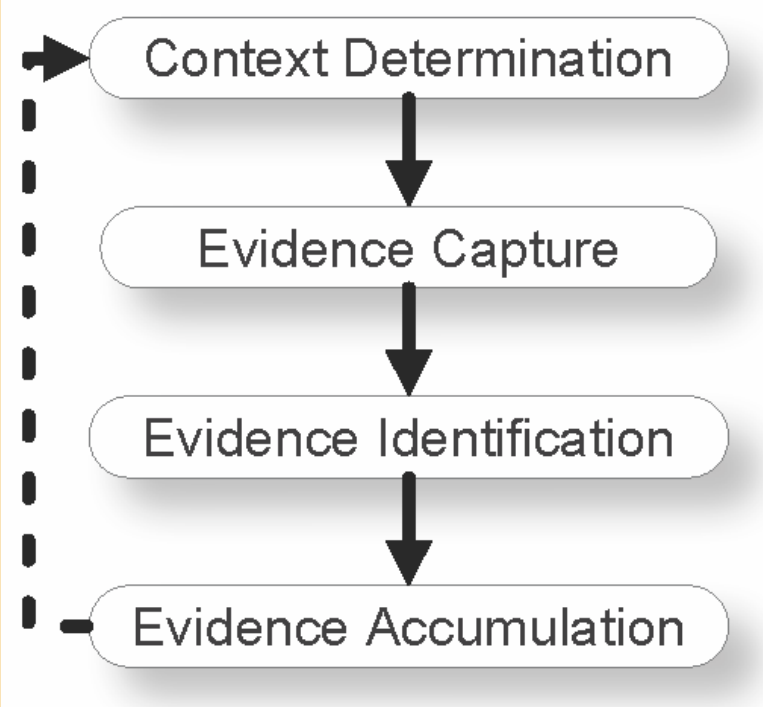
Newton's Playground Proficiencies Measured

- Qualitative Physics
  - Potential/Kinetic Energy
  - Angular momentum
- Creativity
  - Fluency
  - Flexibility
  - Originality
- Conscientiousness
  - Persistence
  - Perfectionism

Pilot Test Results

- Field trial with about 150 middle school students showed low (around 0.1) correlation between EAP(Physics) [from Bayes net] and short pre/post test on Intuitive Physics
- Possible Problems:
  - Low quality pre/post test (reliability α=0.5 (Form A), 0.4 (Form B)
  - Problems with design or software
  - Problems with game level design
  - Problems with students "gaming" the system

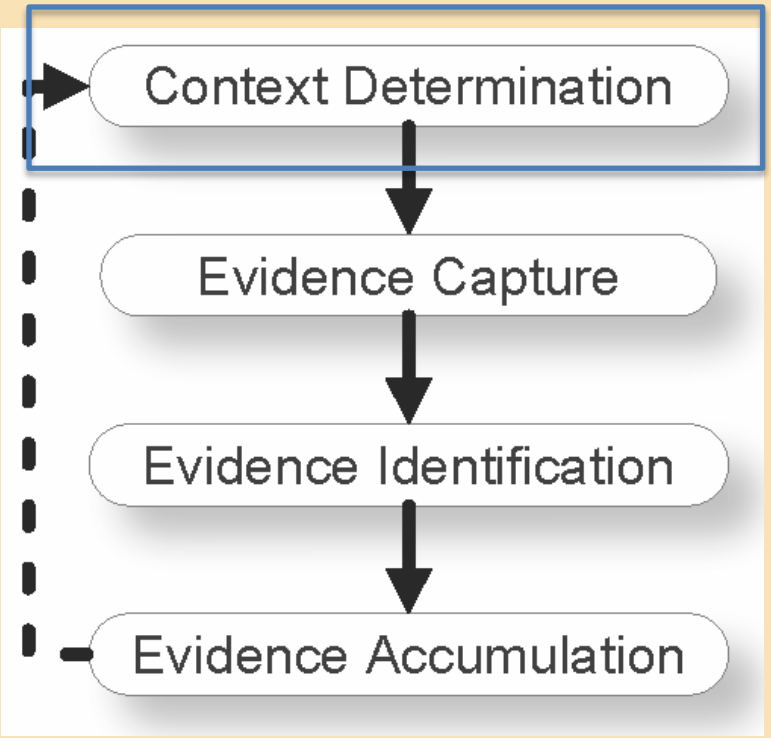
Four Process Model Revisited



- Context Determination (Activity Selection): Develop or recognize *task* contexts
- Evidence Capture (Presentation Process): Collect *work product*
- Evidence Identification (task level scoring): Determine values of *observable outcomes* from work products
- Evidence Accumulation: Aggregates observables across tasks to produce *scores*

In an adaptive system, the task selection/reward structure could be influenced by current "score"

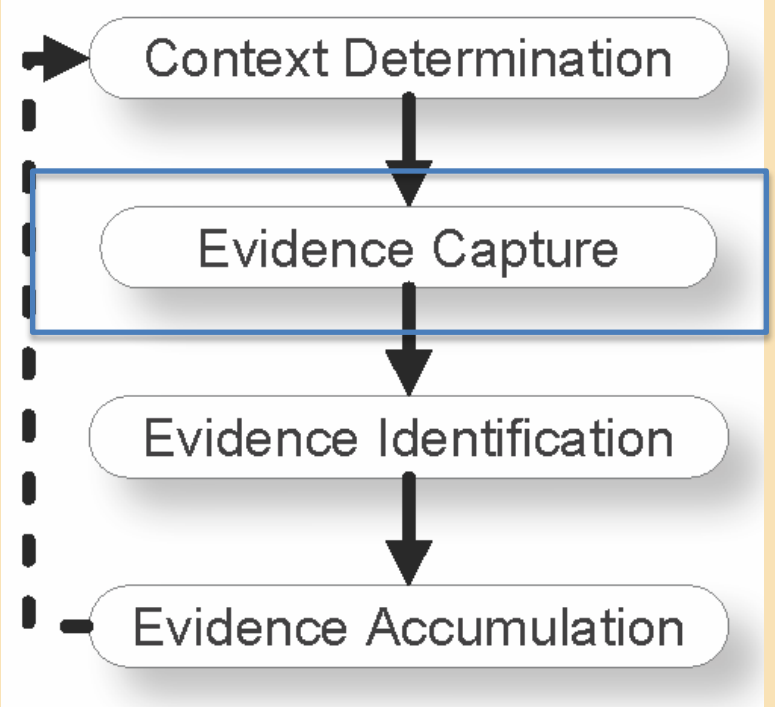
NP: Context Determination



- *NP* is a level-based game
- Authoring Game Levels
  - Coding them with key variables: targeted agents
  - Eliciting difficulty and discrimination expectations
  - Grouping them into "Playgrounds"
  - Testing to make sure levels and game engine work properly.

NP is not adaptive, but as we expect student will normally play through levels in order, need to ensure default sequence provides good evidence mix.

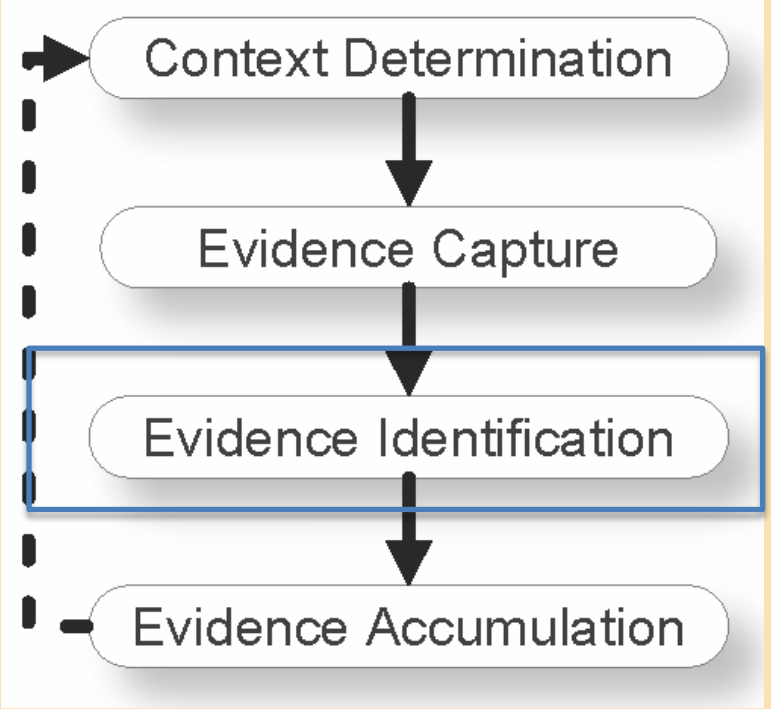
NP: Evidence Capture



- This is the game engine
- Work product includes:
  - Record of success/failures/attempts
  - All objects created and destroyed
  - Agent identification
  - Timing information

The log files are sufficient that we can use the game engine to replay any solution attempt

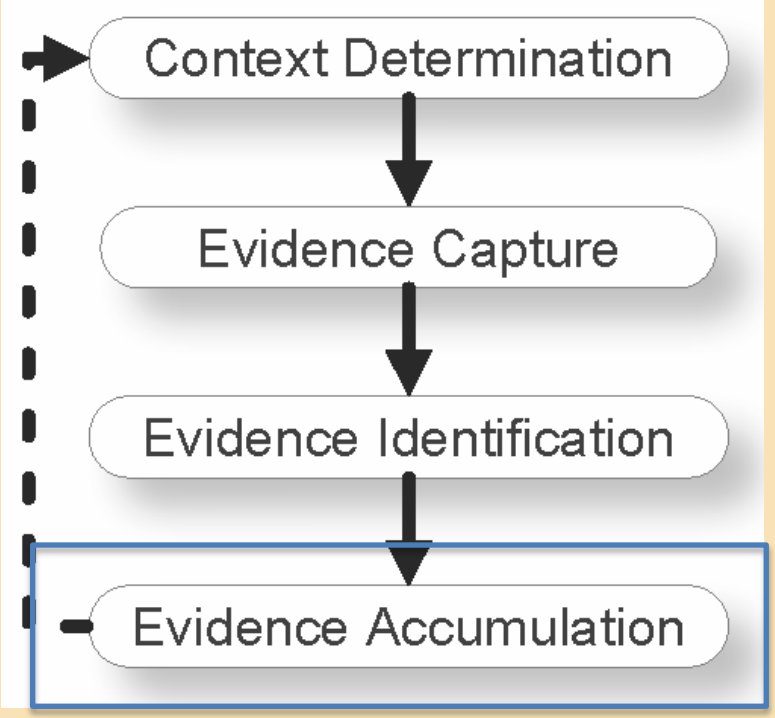
NP: Evidence Identification



EI and EA are currently done as post-processing steps on complete data. In future versions of NP they will provide live scoring.

- Series of Perl Scripts to extract observables from log file.
- Key challenges:
  - Defining observables
  - Aggregating over multiple attempts at the same level.
  - Filtering out off-track/irrelevant behavior ("gaming")
    - "Gaming" behavior produces noise in agent identification system.
  - Session-level observables

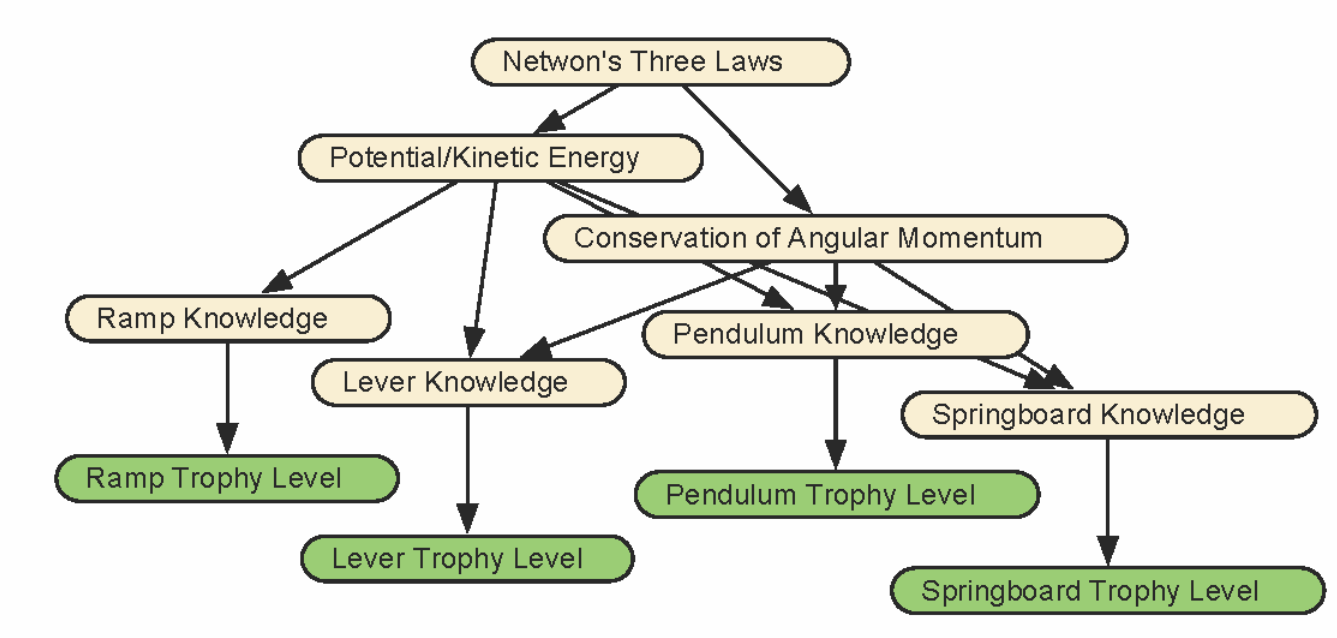
NP: Evidence Accumulation



Bayes net potentially produces score after observables from each level are absorbed.

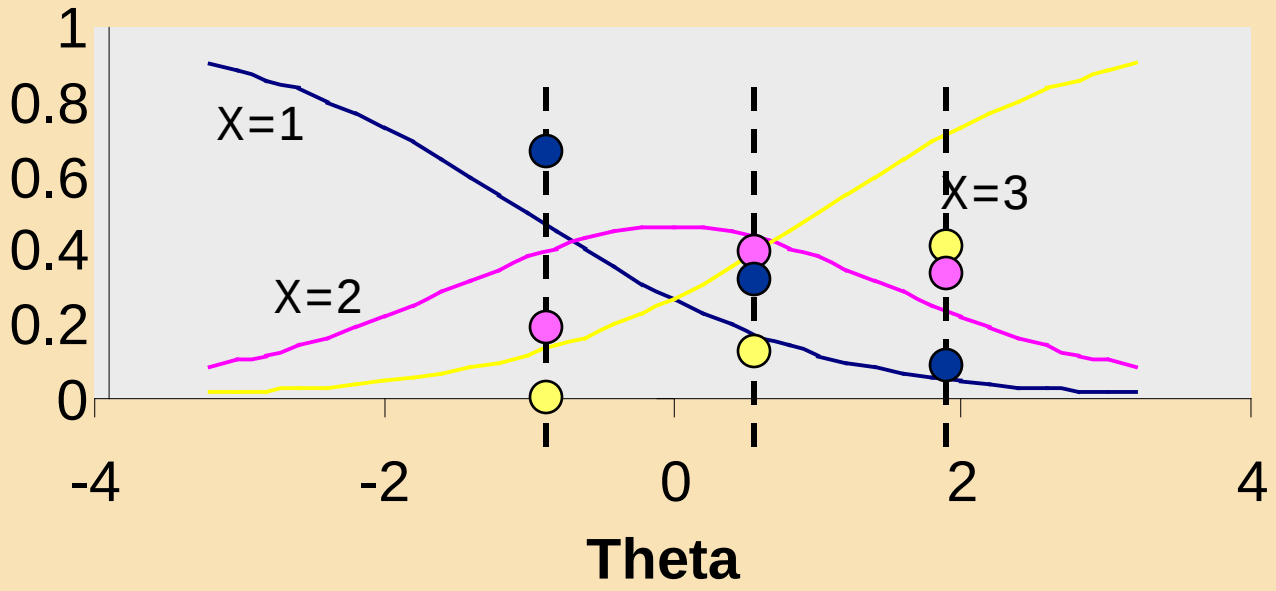
- Bayesian network
  - Student-specific student model
  - Level-specific evidence model
- Evidence models for all levels share same structure, but custom values for conditional probability tables (parameters)
- Can score based on parameters elicited from experts, or try to learn parameters from data:
  - Data are rather sparse, especially at later levels
  - "Gaming" behavior introduces noise into the estimation method
  - Students did not successfully complete many levels, so test-length is short.
  - Pre-post test is too short and too hard.

Physics Model



- Trophy nodes have three possible values: Gold, Silver or None
- Can use data from pre/post test as proxy for latent agent knowledge variables
- Depends heavily on agent identification system

The "Effective θ" Method for determining CPTs



Assign values of parent variable a value θ, on a unit normal scale

Pr(Any Trophy | Agent Ability)
$= \text{logit}^{-1}\ 1.7\ a_S\ (\theta - b_S)$

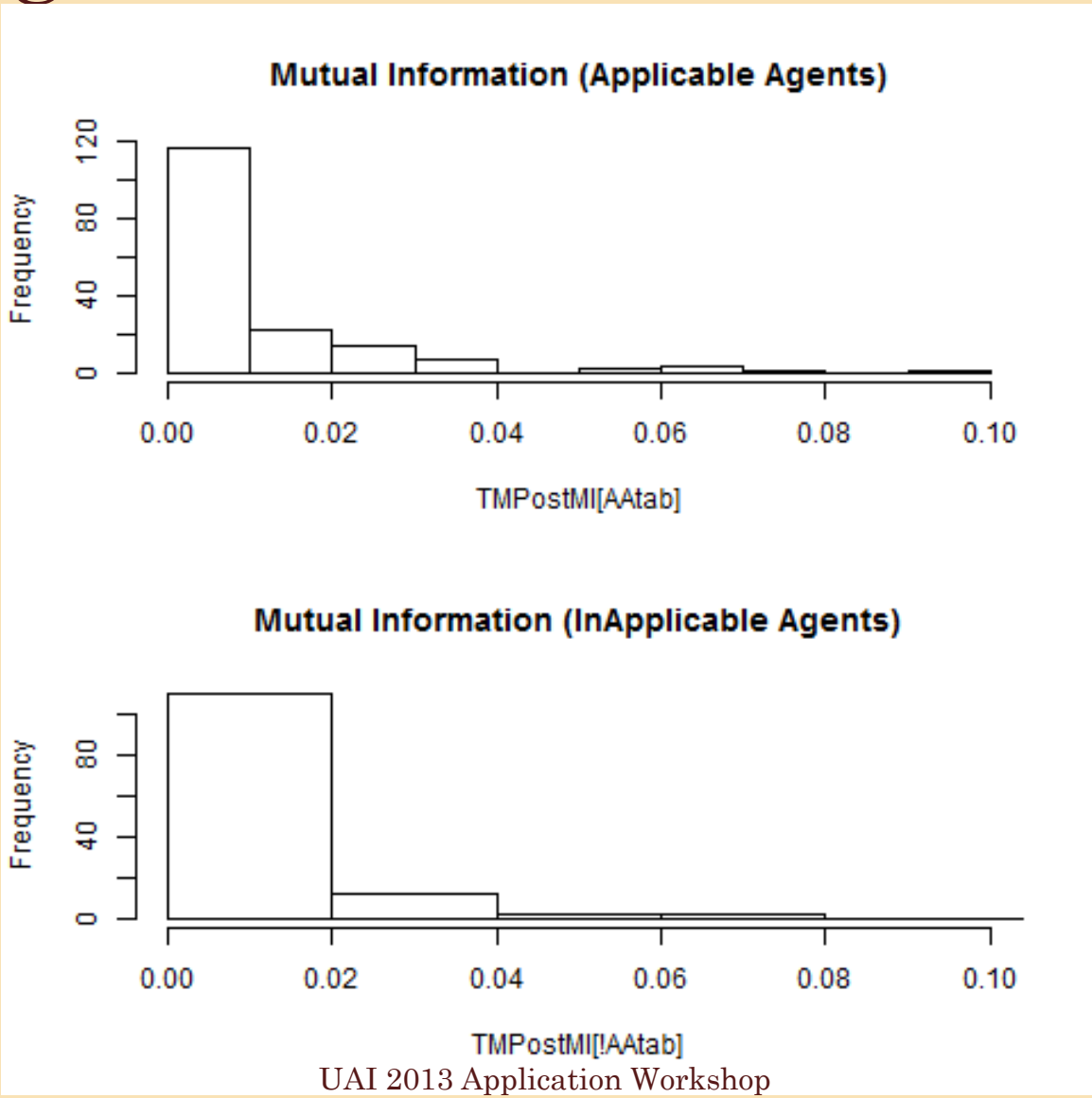
Pr(Gold Trophy | Any Trophy, Agent Ability)
$= \text{logit}^{-1}\ 1.7\ a_S\ (\theta - b_S)$

Some typical Numbers

Level	Lever Problem?	a_{Silver}	b_{Silver}	a_{Gold}	b_{Gold}	Mutual Information
Diving Board World	TRUE	0.90	5.04	0.024	1.97	0.000
Smiley	TRUE	3.37	7.26	0.002	1.48	0.000
St. Augustine	TRUE	0.90	5.04	0.024	1.97	0.000
Stairs	TRUE	11.08	10.76	0.000	0.77	0.064
Swamp People	FALSE	0.12	4.78	2.431	3.69	0.033
Ballistic Pendulum	FALSE	0.90	5.04	0.024	1.97	0.000

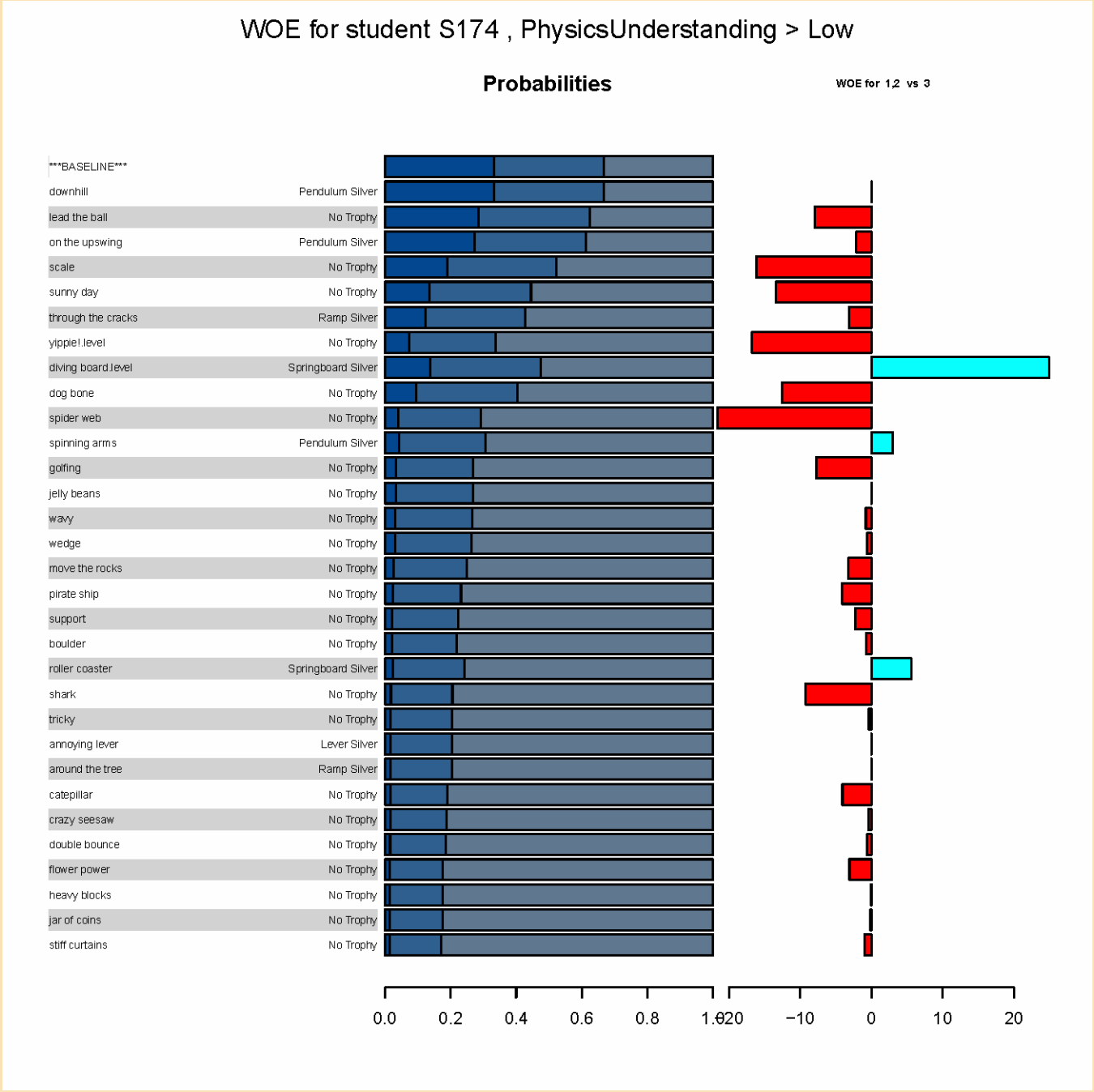
- High slope/intercepts could correspond to convergence problems in model
  - could just indicate we are the tails of the logistic curve
  - Use Mutual Information to screen for this case
- Low slope indicates that observable is not providing much information:
  - problematic if this observable is a target of inference.

Histogram of Mutual Information

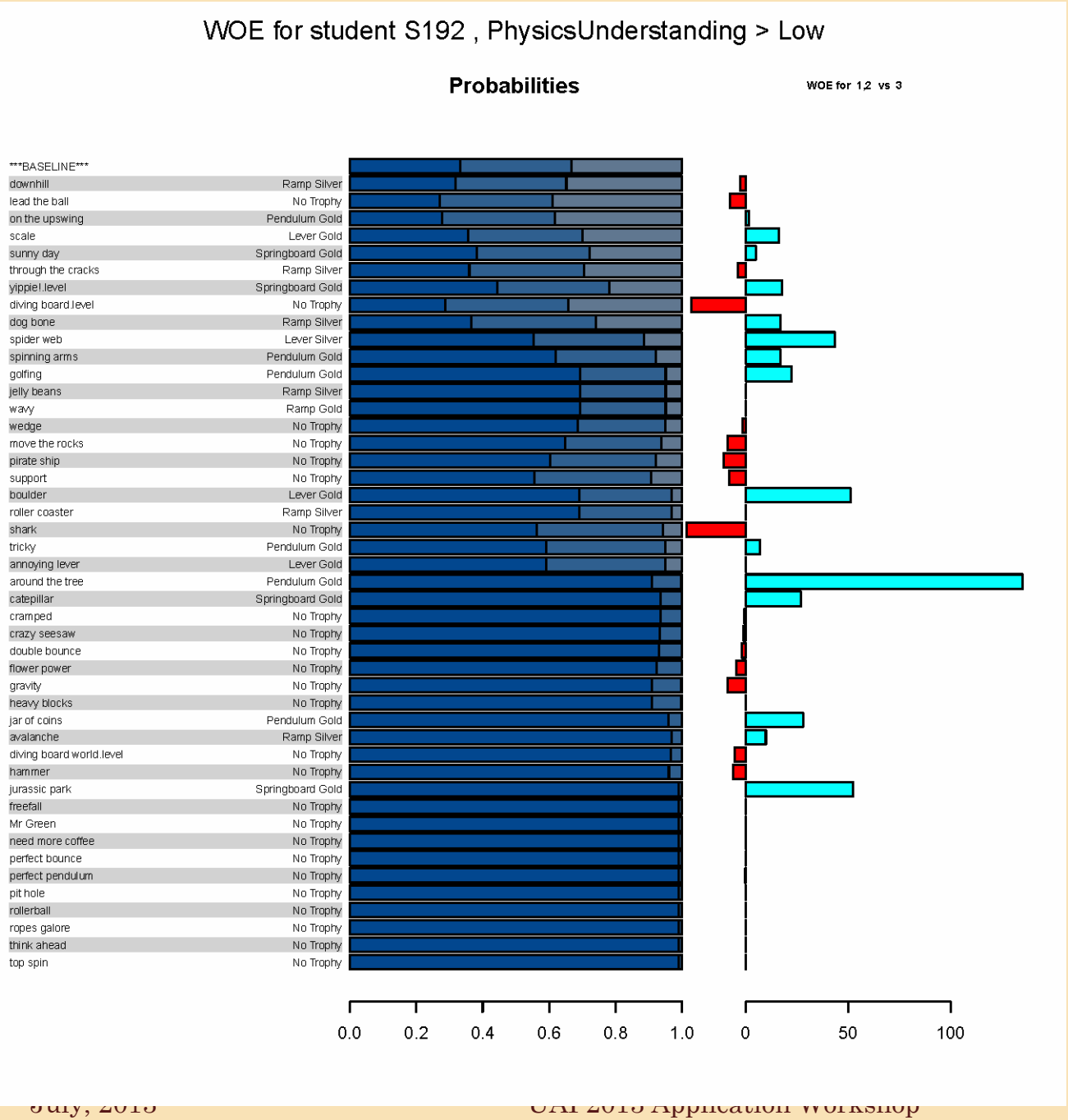


Evidence Balance Sheets

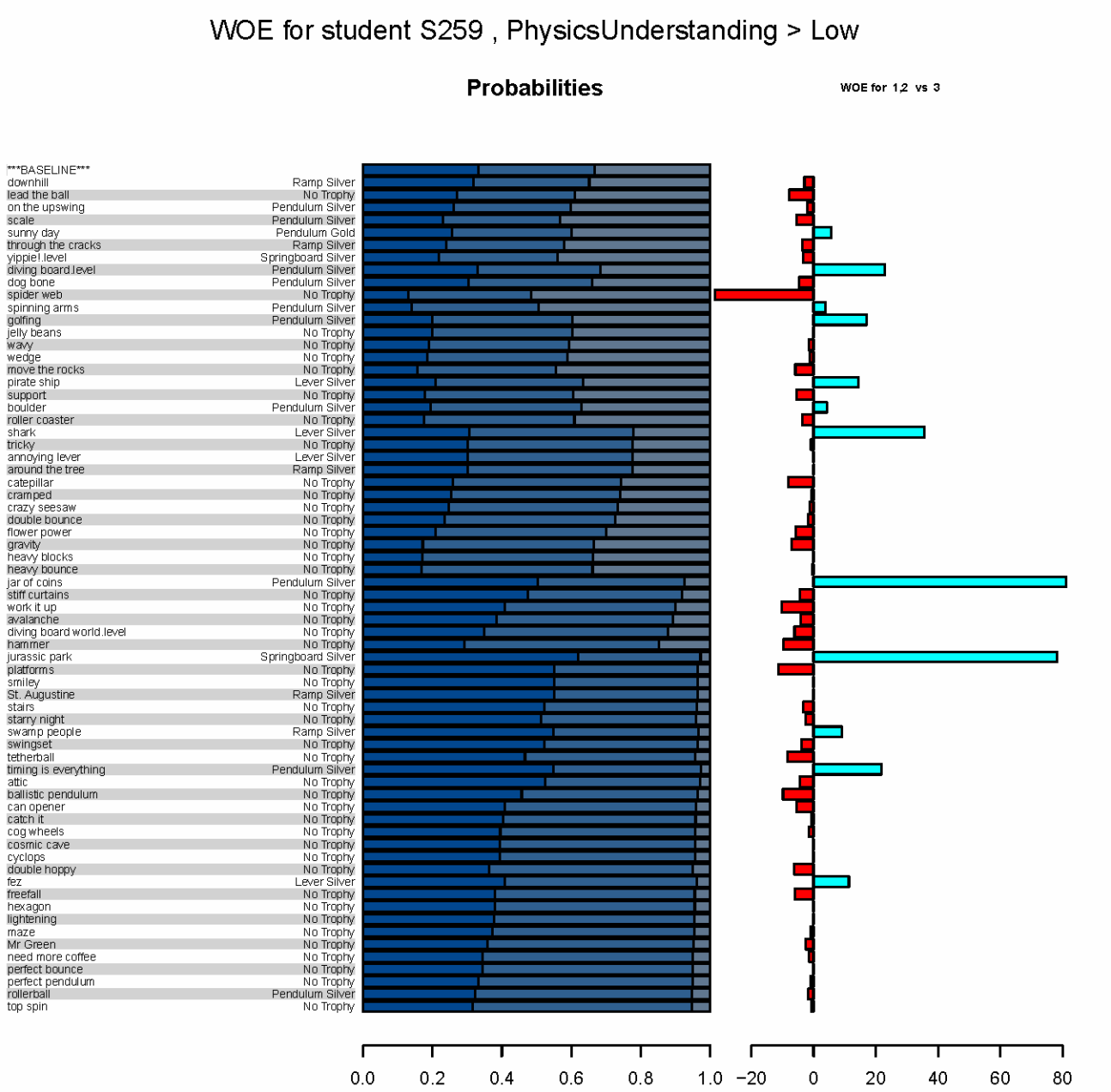
- Show changes in probability (Weight of Evidence) as scoring engine absorbs data from each level
  - (Earlier levels naturally have more evidence)
- Big Jumps indicate that something interesting is going on:
  - "Gaming" behavior
  - Nonstandard solutions
  - Problems with evidence accumulation or identification
  - Student Learning (possibly)



Low Performer



High Performer



Something strange is going on.
Lots of big jumps associated with silver trophies.

Problems found and fixed

- Insufficient weight given to expert priors in building CPTs
- Levels where lots of gaming going on:
  - Screened "gaming" cases out of pretest
  - Added gaming detection features to engine
  - Changed scoring rules to emphasize agent construction
- Levels with agent partially drawn on screen were problematic
  - And replaced
- Reworked Pre/post tests to make them less difficult
- Were able to bring correlations with pre/posttest up to around 0.4 (close to reliability of the pretest)

Thanks

- *Newton's Playground* team:
  - Val Shute (P.I.)
  - Matthew Small
  - Don Franceschetti
  - Lubin Wang
  - Pete Stafford
- Bill & Melinda Gates Foundation *U.S. Programs Grant Number 0PP1035331 Games as Learning/Assessment: Stealth Assessment*