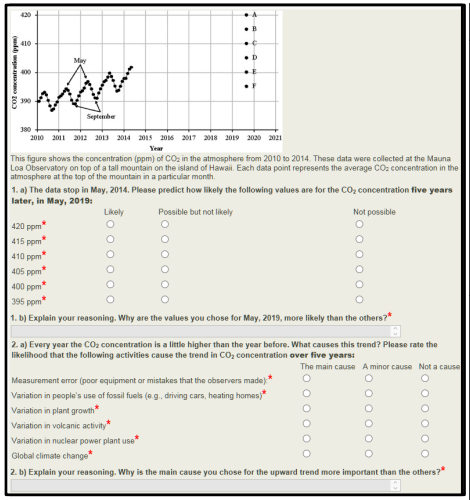
Using machine learning to interpret NGSS tasks at scale

Jay Thomas
Jay.Thomas@act.org
ACT Next, ACT Inc.

Challenges with assessing NGSS with ML

- Unlike essay scoring which can usual general algorithms and engines across multiple prompts (Attali et al, 2010; Shermis, 2015; Shaw, et al, 2019), NGSS based items need a separate scoring rubric for each task because the integration of science content with argumentation is critical.
- Scoring multidimensional constructs that involve SEP, DCI, and CCC
- Assessment tasks should include multiple components to fully assess a given concept (NRC, 2014) using authentic data
- Student errors in spelling, typing, etc. should not negatively impact scoring if not relevant to the construct
- Must be able to maintain acceptable reliability (QWK>.7) across multiple testing cycles
- Need to be able to use composite items with forced choice and constructed response to assess multiple facets of constructs and concepts in time efficient way
- Provide feedback for future instruction (formative)

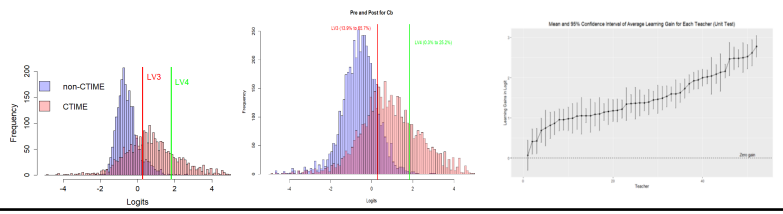
Items, rubrics, and scoring



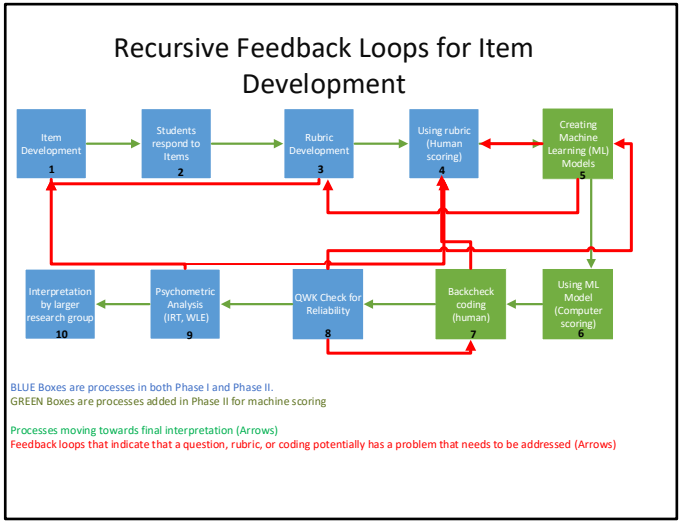
Levels	Indicators	Sample Student Responses for Indicator
Level 4: Students recognize the periodicity of the figure and identify plant processes as the primary cause.	1. Explains that an increase in photosynthesis/plant growth/CO2 uptake during the summer is the main reason for the variation in CO2 concentration in the atmosphere. 2. Explains that plant growth is the only process that can account for the periodic nature of the graph.	4.1) Not a cause/ A minor cause/ The main cause/ A minor cause/ Not a cause/ Not a cause / Variation in plant growth is more important because there aren't many people living on a volcano, so it's mostly natural causes, and in the winter, the CO2 level in the atmosphere goes up because there is less photosynthesis. 4.2) 340) Not a cause/ A minor cause /The main cause/ Not a cause/ Not a cause/ Not a cause /The reason i chose the answer i did is because most of the other answers would not account for the repeating pattern over multiple years. Plant growth is something we as scientists can predict while global climate change would not explain how consistently may is the peak and September is when it falls.
Level 3: Students recognize the periodicity of the figure but make mistakes explaining the mechanism for its cause. Or they recognize plant processes as the primary cause, but don't explicitly relate those processes to the seasonal pattern.	1. Accounts for the seasonal pattern in the figure with an incorrect mechanism (e.g., people's fossil fuel use). 2. Describes an inaccuate mechanism for how plants impact the seasonal CO2 patterns, OR don't recognize plant processes as the primary driver of the annual pattern. 3. Explains that plants take in CO2 with no mention of the seasonality of this process.	3.1) Not a cause/ The main cause/ The main cause /A minor cause/ A minor cause/ A minor cause /, etc. instead of diving due to the nice worm weather. Also people will use less energy warming homes while it is summer causing less fossil fuels to be burnt. seasonal pattern 3.2) A minor cause/ A minor cause/ The main cause/ A minor cause /Not a cause/ A minor cause/ For me the major source of CO2 would be plants decaying during the time between May and September and the CO2 levels in the atmosphere rising as a result. 3.3) Not a cause /A minor cause /The main cause/ Not a cause /A minor cause/ A minor cause /Because, plants are the ones that use CO2 for photosynthesis so they absorb it.
Level 2: Students identify fossil fuels as a carbon source.	1. Explains that fossil fuel use produces CO2/carbon (may also identify other sources, too)	2.1) Not a cause /The main cause/ The main cause/ A minor cause/ A minor cause / The main cause /The main causes /are people use of fossil fuel, plant growth, and global climate change because they all affect the amount of carbon dioxide that enters the atmosphere.

- Must incorporate features from both FC and CR
- Identify student responses at both the Level and Indicator to identify specific error, misconception, etc. that can be addressed with future instruction
- ML scoring should mirror expert human coding with acceptable reliability
- ML engine uses FC even when human rubric did not include that as a criteria because the ML engine can find patterns in large data that humans could not
  - FC responses usually lower in decision tree (tie breaker)
  - Lower weight in logistic regression
  - Proxies for words (hoto,*rgy) that are misspelled or have multiple forms
- ML codes at the indicator level tied to specific errors or misconceptions that can be used to inform instructional decisions (formative assessment)
- ML scoring can serve as an indicator of the quality of items, rubrics, and scoring procedures

Examples of larger group analysis



Feedback loops in assessment system using ML



ML engines CANNOT score items that humans score poorly . This does not mask problems in assessment but it will help to identify problematic issues: poor item design, incomplete rubrics, inconsistent human scoring.

This allows for iterative development of items that are able to assess the desired constructs consistently. Many items in assessment fail either during review or pre-test stages. These feedback loops allow for some of these items to be used through improvements in the rubric or human scoring while driving some items to be replaced so that they better measure the desired constructs.

Using ML scoring to assess at scale

- Increase in the size of the usable data set to increase power of statistics
- Increased confidence in reliability of scoring through back-checking samples and revising models
- Reduced costs by needing fewer human coders
- Model to show that the kinds of assessments envisioned by Pellegrino et al (2014) for NGSS can be reached at scale with low cost
- Allows for comparison of learning gains because of scope of data
- Models that fail to meet reliability guidelines can be replaced and all responses rescored quickly
- Every student response from the entire year can be used for statistical analyses
  - Unit test (pre and post)
  - Full year (pre and post)

School year	Responses scored	Unique items scored	Assessments scored
15-16	175,265	33	27,981
16-17	532,825	39	61,475
17-18	693,086	41	66,335
18-19	409,266	39	42,117
Total	1,810,442	57	197,908

Citations available. Please email the author.

ACKNOWLEDGEMENT This poster is adapted from an earlier version of a paper co-authored with Andy Anderson, Qinyun Lin, and Kenneth Frank (MSU) as well as Karen Draney and Shruti Bathia (BEAR). When submitted for publication rather than a presentation all co-authors will receive the credit that they richly deserve. Karen and Shruti's work is primarily contained in another poster in this symposium. Qinyun, Andy, and Kenneth are presenting in another session at NARST.