A Diagnostic Tool that Scales Student Voice through Semi-Automated Text Analysis and Qualitative Clustering

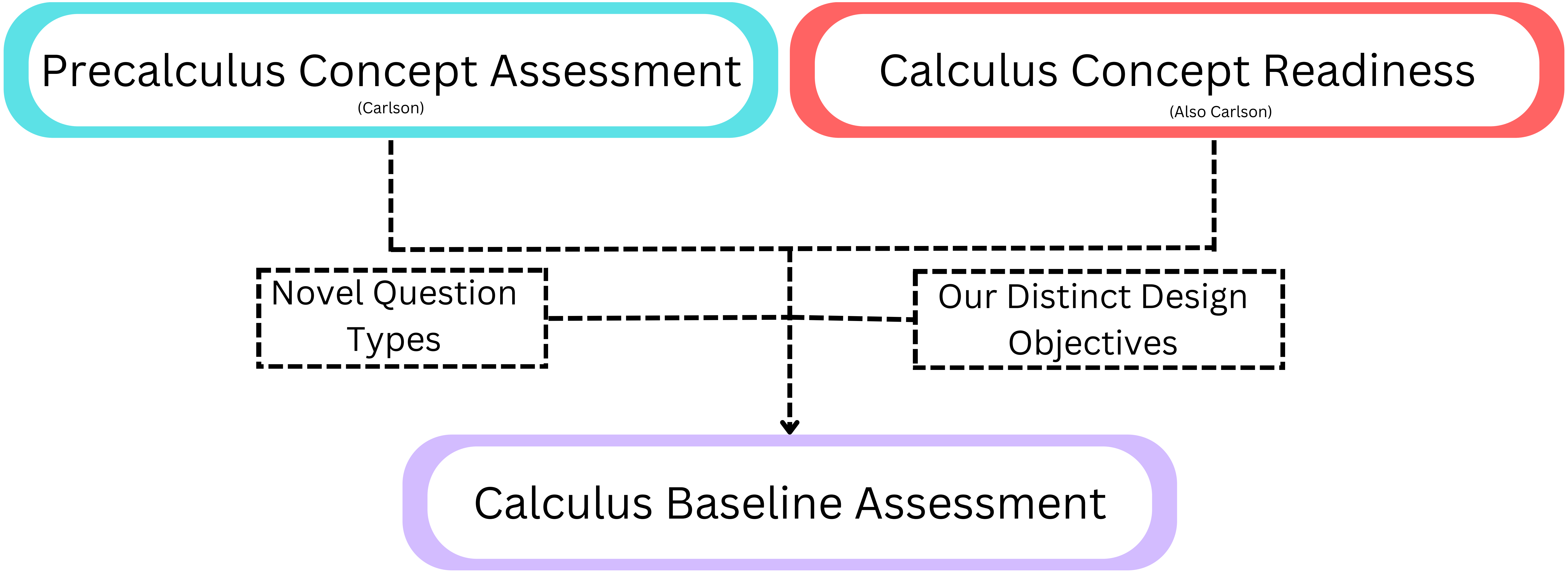
Connor Gregor (McMaster)
Lindsey Daniels (UBC)
Caroline Junkins (McMaster)
James Colliander (Crowdmark and UBC)

CBA Motivation

- Math is a cumulative subject:
    - Student readiness can be gauged based on their mastery of past subjects.
- Expediency is paramount for a diagnostic assessment that gauges said masteries:
    - This is achieved using short multiple choice questions.

CBA Motivation

- Math is a cumulative subject:
  - Student readiness can be gauged based on their mastery of past subjects.
- Expediency is paramount for a diagnostic assessment that gauges said masteries:
  - This is achieved using short multiple choice questions.



CBA Design Objectives

Our design objectives distinguish the CBA as its own distinct assessment.

CBA Design Objectives

Generate Feedback → Reduce Affective Drop → Accessible Deployment

- Multiple Choice Questions can be marked expediently, but do not reveal enough about student thought process

- CBA pairs questions with text prompt boxes.
- Prompts influence likelihood of student engagement with CBA.

CBA Design Objectives

- High-stakes settings degrade math performance.
- Stereotype threat similarly reduces student performance.

- CBA has a week long deadline and is completed open-book.
- CBA is incentivized by awarding course credit for completion.

CBA Design Objectives

Generate Feedback → Reduce Affective Drop → Accessible Deployment

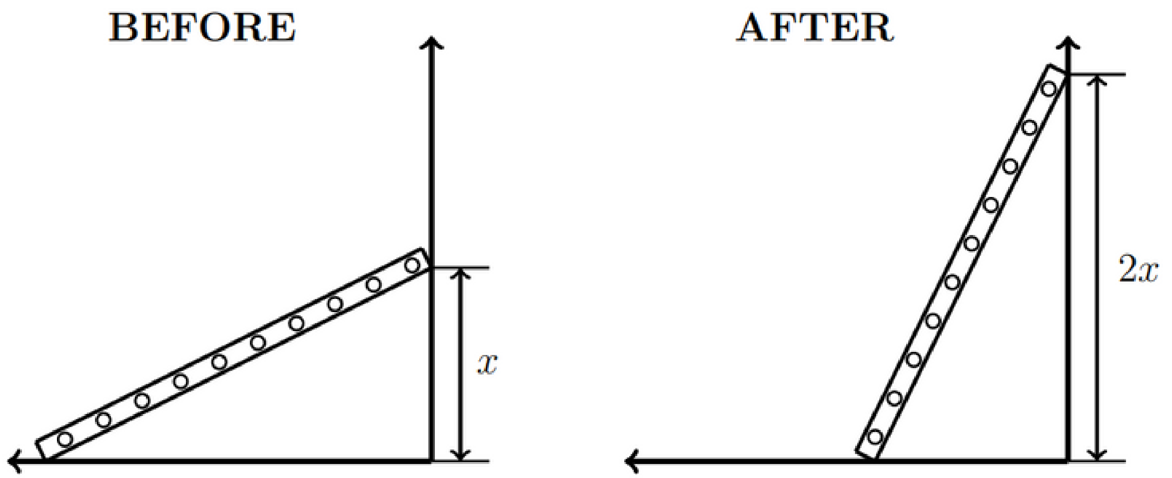
- CBA may be automatically graded and returned.
- Resource links may also be built-in to feedback.

- CBA is delivered through an online browser platform.
- Offers built in screen reading, text to speech, translations.

CBA Sample Question

Q13a (1 point)

A ladder - of fixed length - is leaning against a wall. The ladder is adjusted so that the distance of the top of the ladder from the floor is twice as high as it was before it was adjusted.



The slope of the ladder is:

- Less than twice what it was before
- Exactly twice what it was before
- More than twice what it was before
- The same as what it was before
- There is not enough information to determine if any of a through d is correct.

Q13b (0 points)

Explain the reasoning for your answer to Q13a in the box below.

Edit Preview

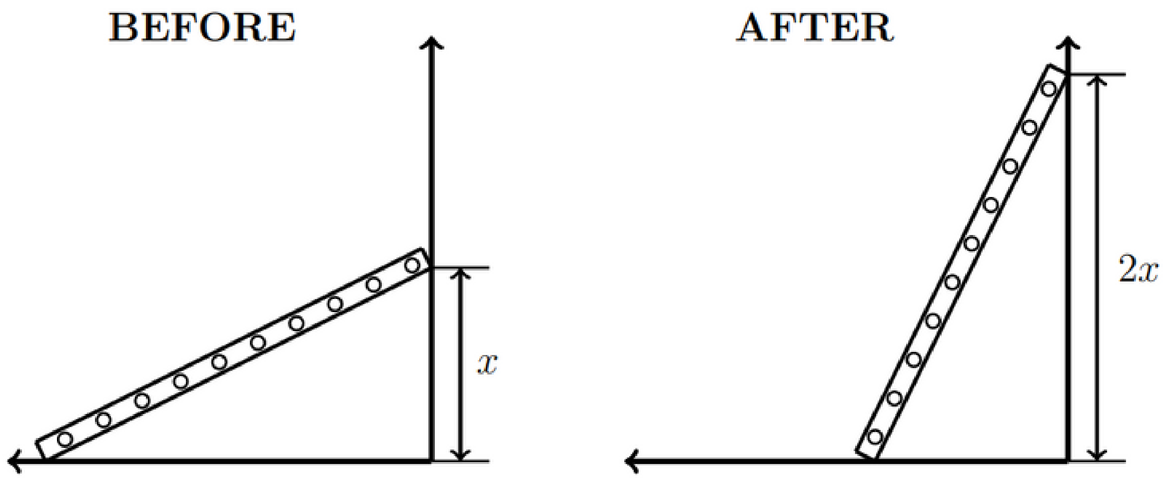
Please enter your response to Q13b

Attach files Formatting tips

CBA Sample Question

Q13a (1 point)

A ladder - of fixed length - is leaning against a wall. The ladder is adjusted so that the distance of the top of the ladder from the floor is twice as high as it was before it was adjusted.



The slope of the ladder is:

- ○ Less than twice what it was before
- ○ Exactly twice what it was before
- ● More than twice what it was before
- ○ The same as what it was before
- ○ There is not enough information to determine if any of a through d is correct.

Q13b (0 points)

Explain the reasoning for your answer to Q13a in the box below.

Edit Preview

Please enter your response to Q13b

Attach files Formatting tips

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
  - Implements Math Definitions

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
  - Implements Math Definitions
  - Contextual Covariational Reasoning

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
  - Implements Math Definitions
  - Contextual Covariational Reasoning
  - Understands Slope Behavior

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
  - Implements Math Definitions
  - Contextual Covariational Reasoning
  - Understands Slope Behavior
  - Provides Comprehensive Detail

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
  - Implements Math Definitions
  - Contextual Covariational Reasoning
  - Understands Slope Behavior
  - Provides Comprehensive Detail
- Similar codes are bundled into unifying themes.

The slope of a line equals rise over run. The height of the ladder (its rise) is being doubled, but since the ladder is fixed length this decreases the run. Because the directly proportional value is doubling and the inversely proportional value is decreasing, the slope will be more than twice its original value.

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
- Similar codes are bundled into unifying themes.

Algebra Skills

Math Relationship Skills

Solution Framework

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
- Similar codes are bundled into unifying themes.

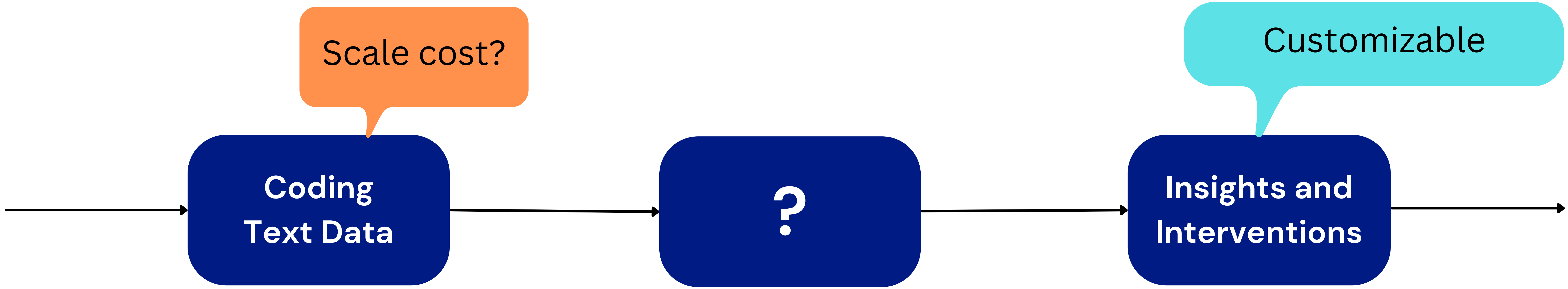
Algebra Skills	Algebra Traps
Math Relationship Skills	Solution Misinterpretation
Solution Framework	Knowledge Gaps

Qualitative Coding Procedure

- We curate a qualitative codebook from scratch.
- Similar codes are bundled into unifying themes.

Algebra Skills	Algebra Traps
Math Relationship Skills	Solution Misinterpretation
Solution Framework	Knowledge Gaps
Visualization	Mathematical Language
Contextual Reasoning	Heuristic View

Our Methodology



Our Methodology

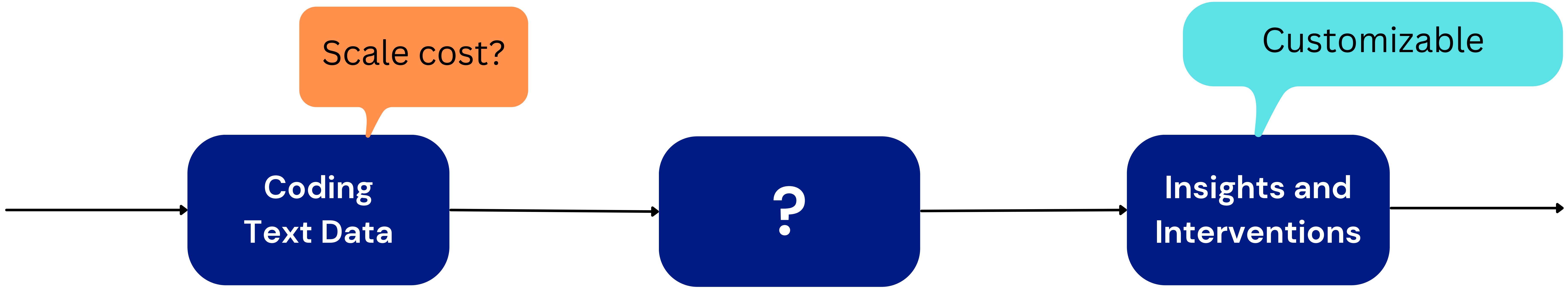
Scale cost?

Customizable

Coding Text Data → ? → Insights and Interventions

Qualitative coding provides quantifiable information about student aptitudes; however...

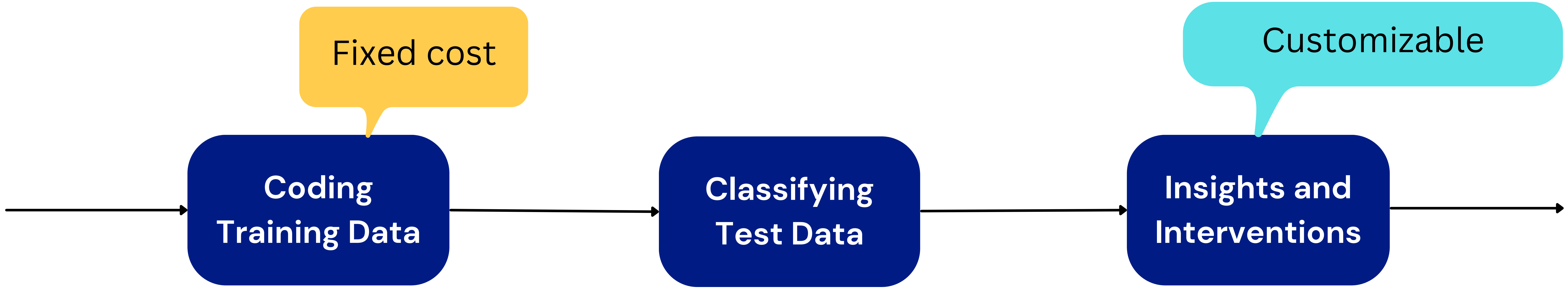
Our Methodology



Qualitative coding provides quantifiable information about student aptitudes; however...

This is a massive time investment!

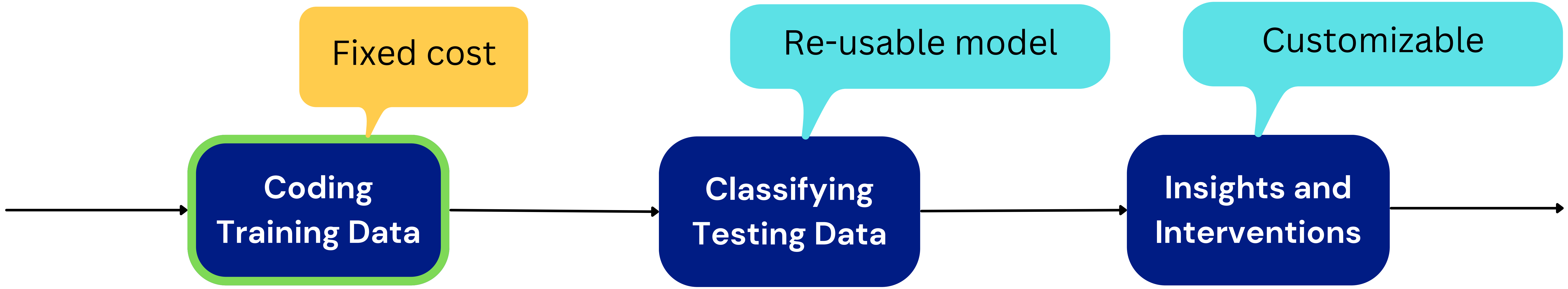
Our Methodology



Qualitative coding provides quantifiable information about student aptitudes; however...

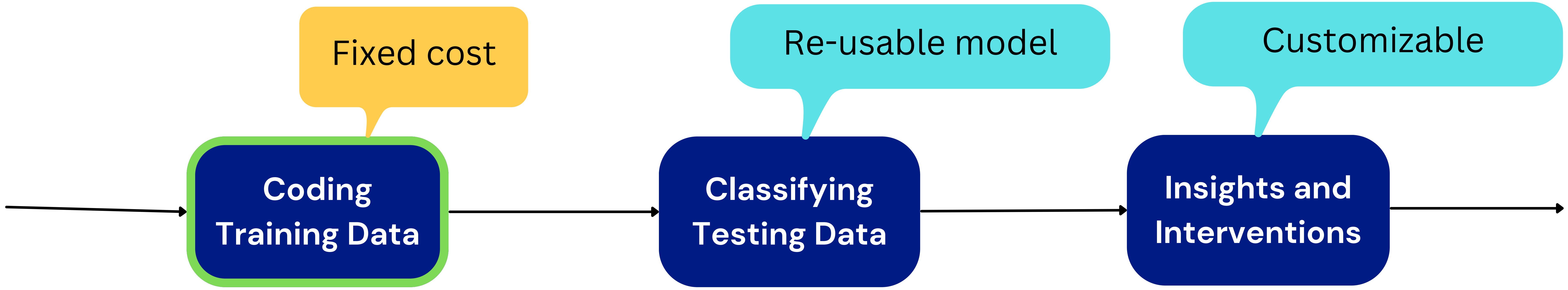
This is a massive time investment! Thankfully machine learning can automate this process!

Our Methodology



- For the sake of automation, each response is transformed using NLP.
    - The raw text is turned into a list of lemmatized tokens that have had their 'stop' words removed by a filter.

Our Methodology

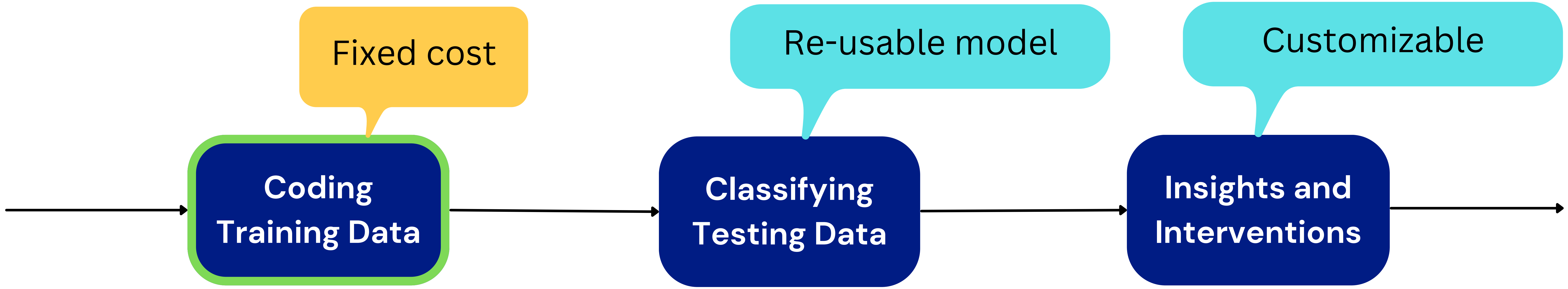


- For the sake of automation, each response is transformed using NLP.
  - The raw text is turned into a list of lemmatized tokens that have had their 'stop' words removed by a filter.

e.g. The previous statements are converted into this token list:

['sake', 'automation', 'response', 'transform', 'use', 'nlp', 'raw', 'text', 'turn', 'list', 'lemmatized', 'token', 'stop', 'word', 'remove', 'filter']

Our Methodology



- For the sake of automation, each response is transformed using NLP.
  - The raw text is turned into a list of lemmatized tokens that have had their 'stop' words removed by a filter.

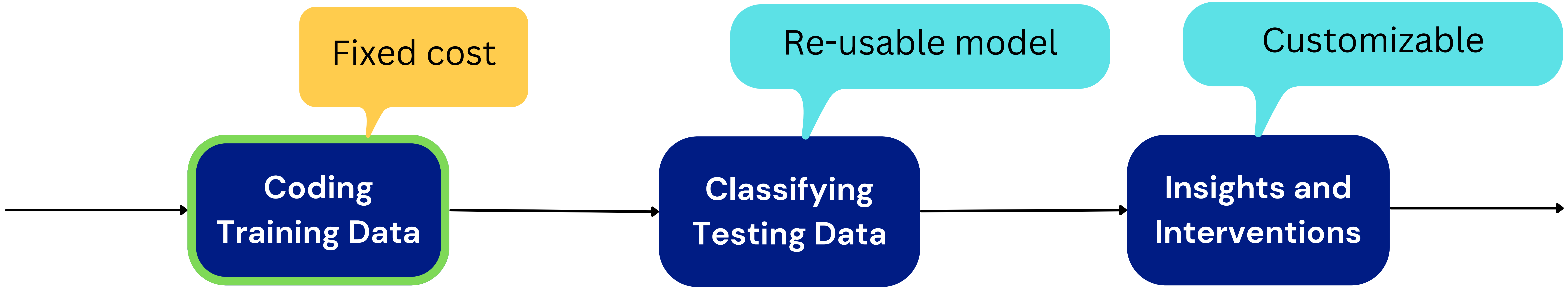
e.g.
stateme
into

['sake', 'automation', 'response', 'transform', 'use',
'token',

Token lists are transformed into vectors:

- Popular tokens are set as vector dimensions.
- Text entry has non-zero value in a dimension *if* it contains said token.
- Actual value varies based on choice of representation.

Our Methodology



- For the sake of automation, each response is transformed using NLP.
  - The raw text is turned into a list of lemmatized tokens that have had their 'stop' words removed by a filter.

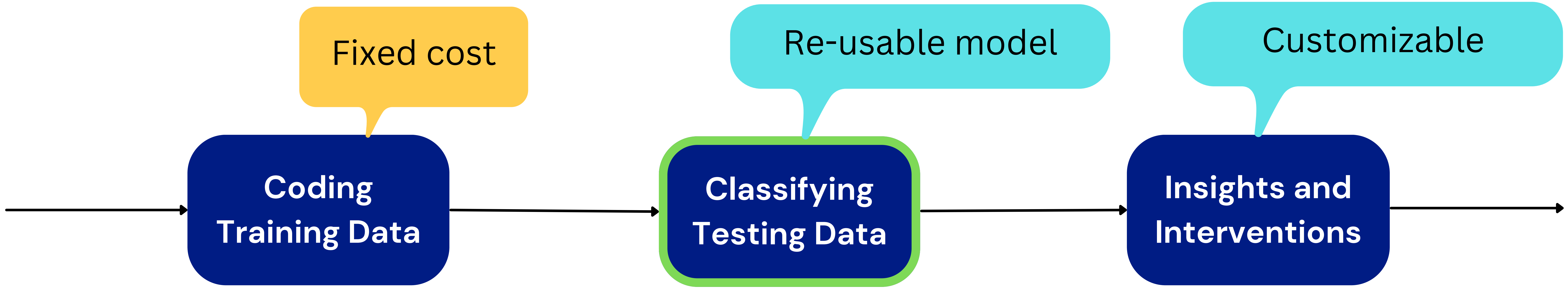
['sake', 'automation', 'response', 'transform', 'use',

Token lists are transformed into vectors:

e.g. The tokens can translate to a vector with these dimensions.

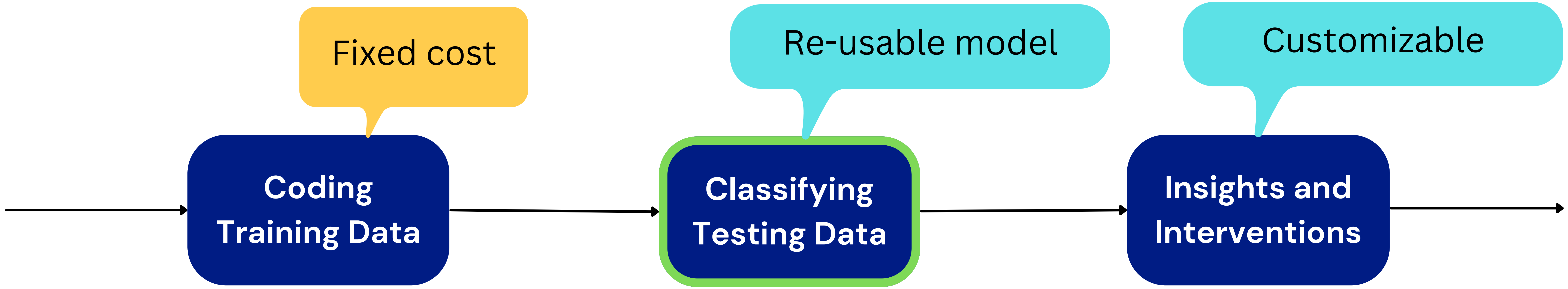
automation	computer	filter	math	nlp	plus	subtract	token	use
1	0	1	0	1	0	0	1	1

Our Methodology



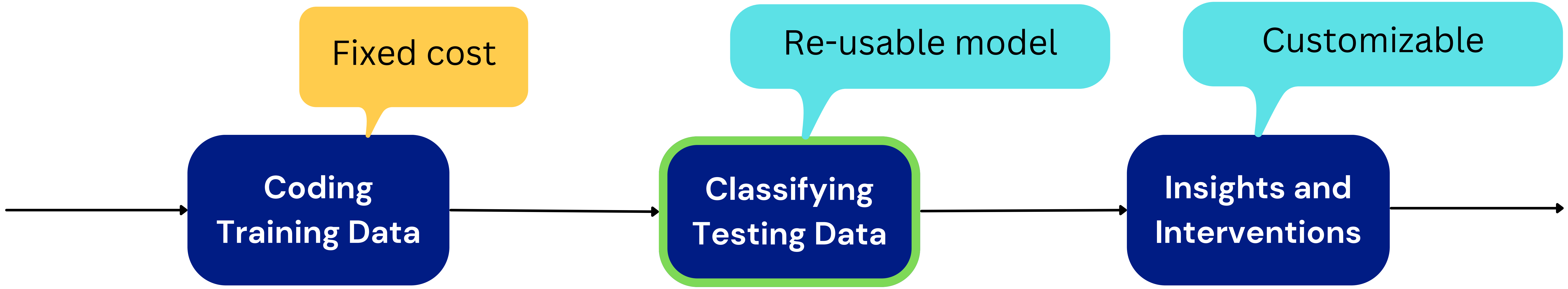
- Transformed vectors serve as input for our ML model (*x*)
- A separate vector serves as output (*y*)
  - Vector entries signal presence of a qualitative code.
  - Text responses possess own code signature.

Our Methodology

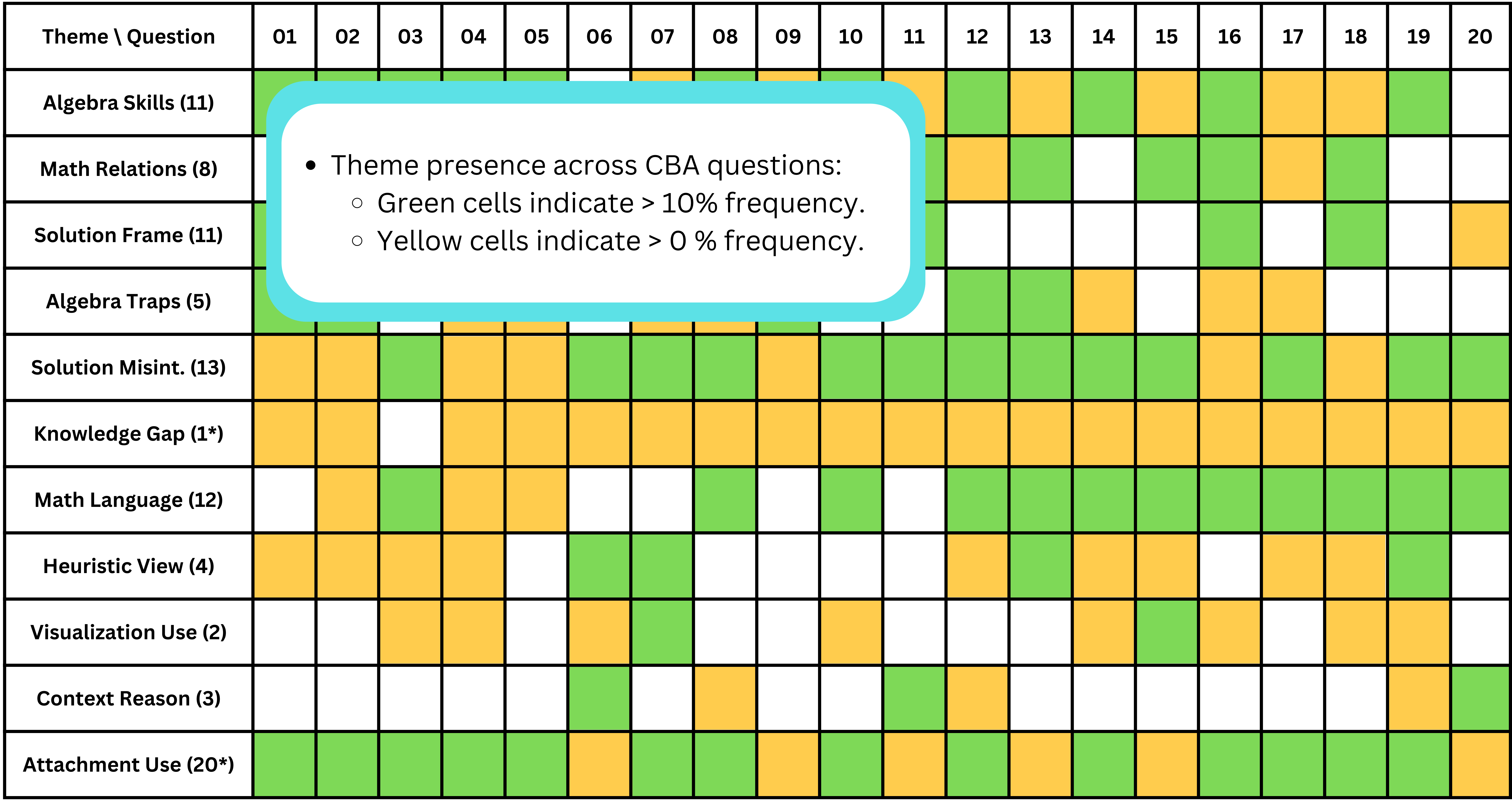


- Transformed vectors serve as input for our ML model (*x*)
- A separate vector serves as output (*y*)
  - Vector entries signal presence of a qualitative code.
  - Text responses possess own code signature.
- Using *(x,y)* pairs, a ML model can learn to qualitatively code!
  - We use gradient boosting machines to do this.
  - GBMs iteratively design sequence of decision trees which classify vectors.

Our Methodology

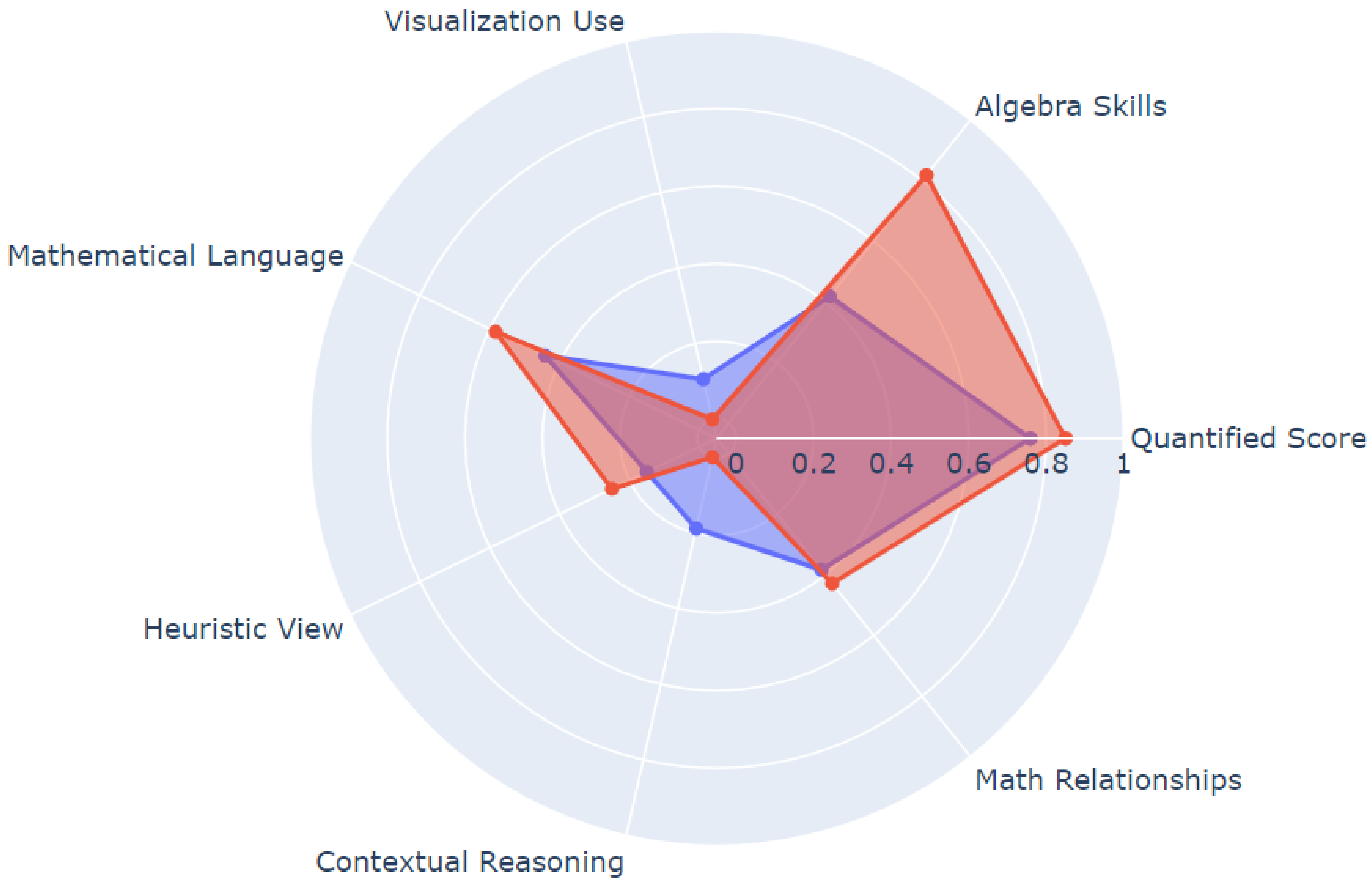


- Transformed vectors serve as input for our ML model (*x*)
- A separate vector serves as output (*y*)
  - Vector entries signal presence of a qualitative code.
  - Text responses possess own code signature.
- Using *(x,y)* pairs, a ML model can learn to qualitatively code!
  - We use gradient boosting machines to do this.
  - GBMs iteratively design sequence of decision trees which classify vectors.
    - We must discern where to apply models based on code frequency.


Theme \ Question
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Algebra Skills (11)
Math Relations (8)
Solution Frame (11)
Algebra Traps (5)
Solution Misint. (13)
Knowledge Gap (1*)
Math Language (12)
Heuristic View (4)
Visualization Use (2)
Context Reason (3)
Attachment Use (20*)
Theme presence across CBA questions:
Green cells indicate > 10% frequency.
Yellow cells indicate > 0 % frequency.

Theme \ Question	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Algebra Skills (11)																				
Math Relations (8)																				
Solution Frame (11)																				
Algebra Traps (5)																				
Solution Misint. (13)																				
Knowledge Gap (1*)																				
Math Language (12)																				
Heuristic View (4)																				
Visualization Use (2)																				
Context Reason (3)																				
Attachment Use (20*)																				

Data Visualization with Spider Plots

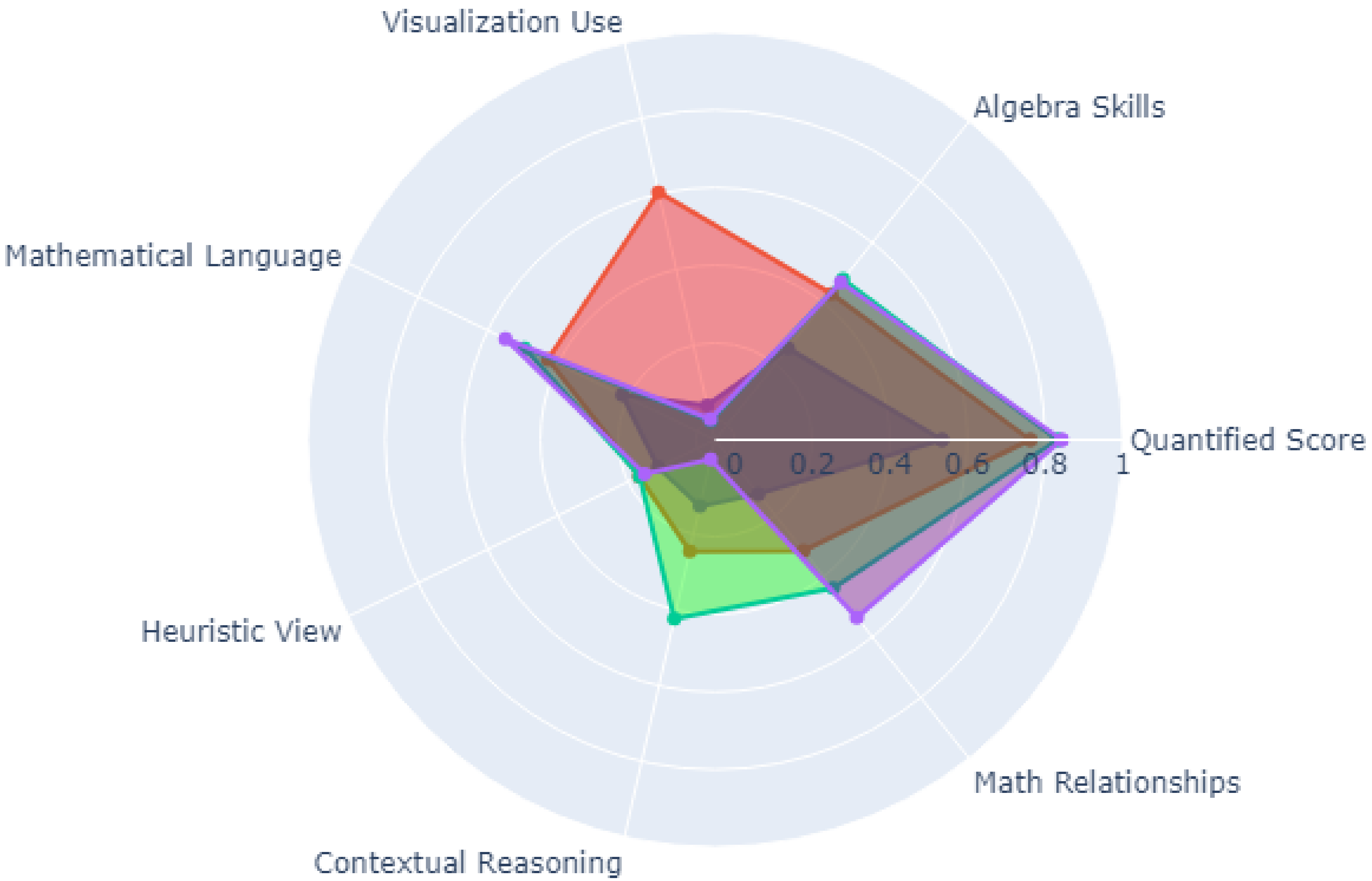


Six themes are joined with CBA score to create a profile of student readiness. The red plot shows the profile of a random student while the blue plot shows the averaged profile of every student in the cohort.

Collating Student Dimensions:

- We standardize the fraction of flagged theme instances as a single dimension.

Clustering Student Cohort



Four spider plots each depicting the same seven student attributes; each plot is the centroid of a student cluster where: cluster A (blue) contains 147 students, cluster B (red) contains 106 students, cluster C (green) contains 188 students, cluster D (purple) contains 222 students.

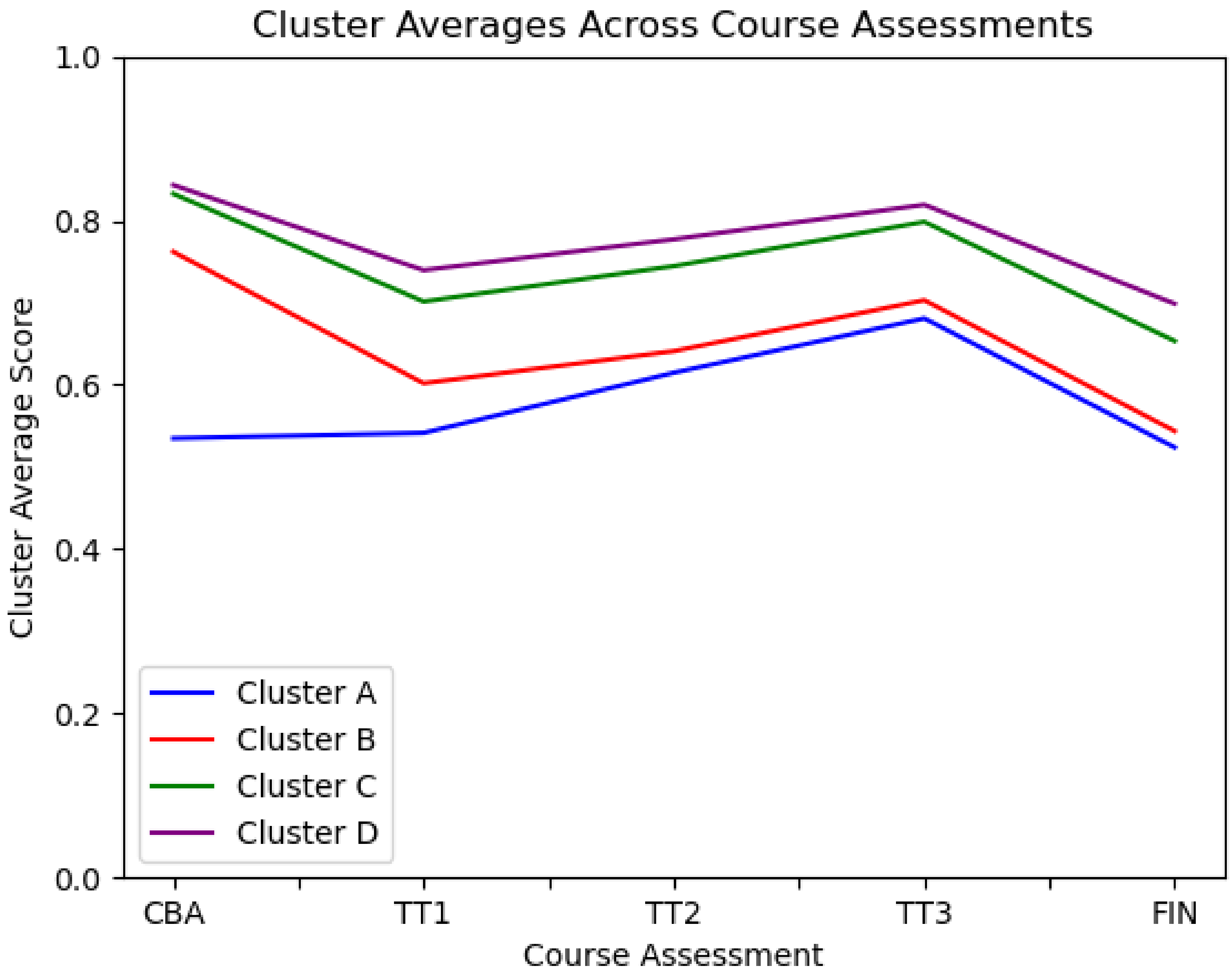
Collating Student Dimensions:

- We standardize the fraction of flagged theme instances as a single dimension.

Using k-clustering:

- We partition students based on dimensional distance from one another.

Clustering Student Cohort



Collating Student Dimensions:

- We standardize the fraction of flagged theme instances as a single dimension.

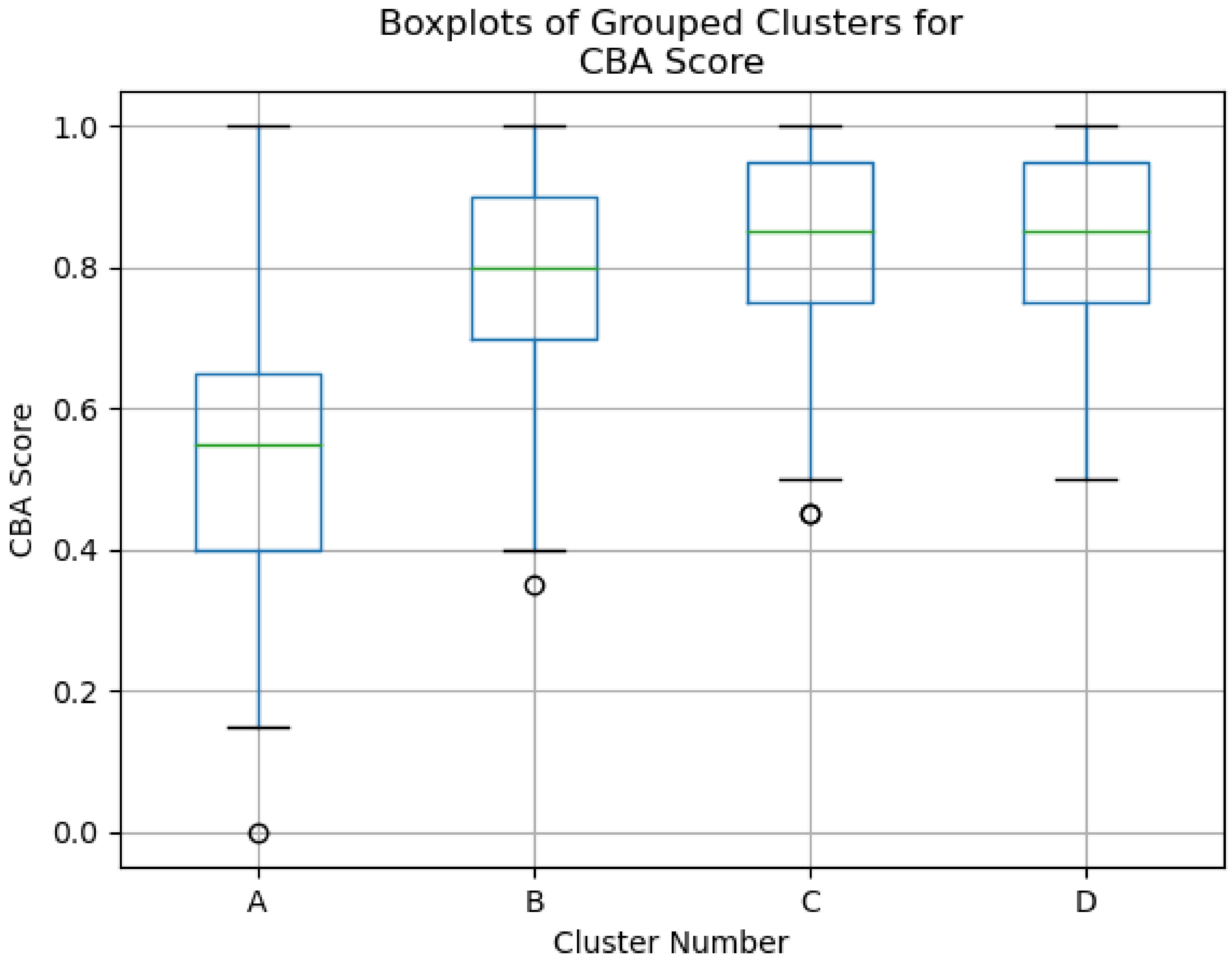
Using k-clustering:

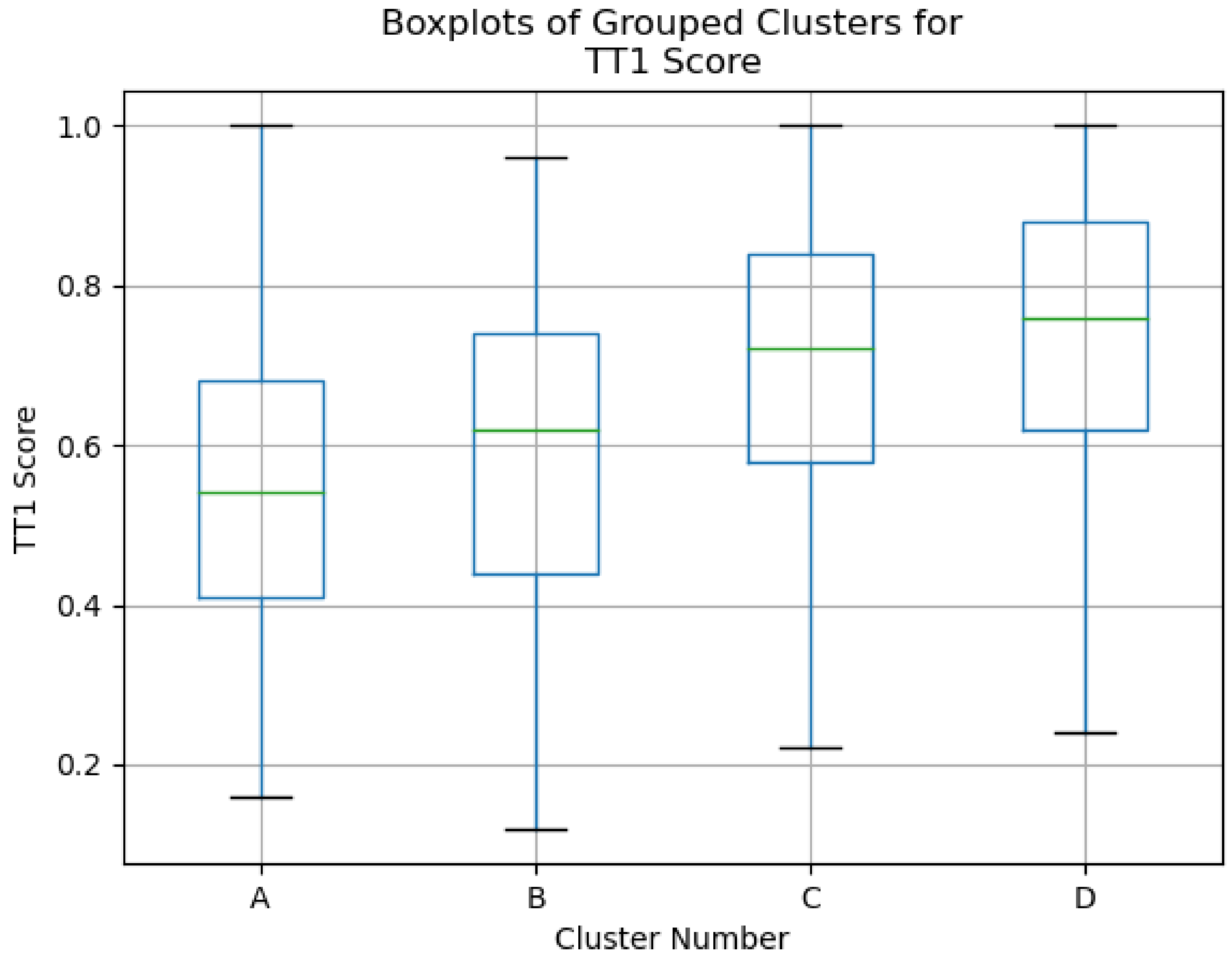
- We partition students based on dimensional distance from one another.

Tracking Cluster Performance:

- We track cluster performance across a semester for longitudinal observation.

Clustering Student Cohort





Boxplots depicting spread of student scores on math assessments when cohort is partitioned by student cluster.

Concluding Remarks

Using these clusters, targeted student interventions can be designed by instructors.

Future CBA implementations can allow for these interventions and supports to be automatically delivered upon completion.

Concluding Remarks

Using these clusters, targeted student interventions can be designed by instructors.

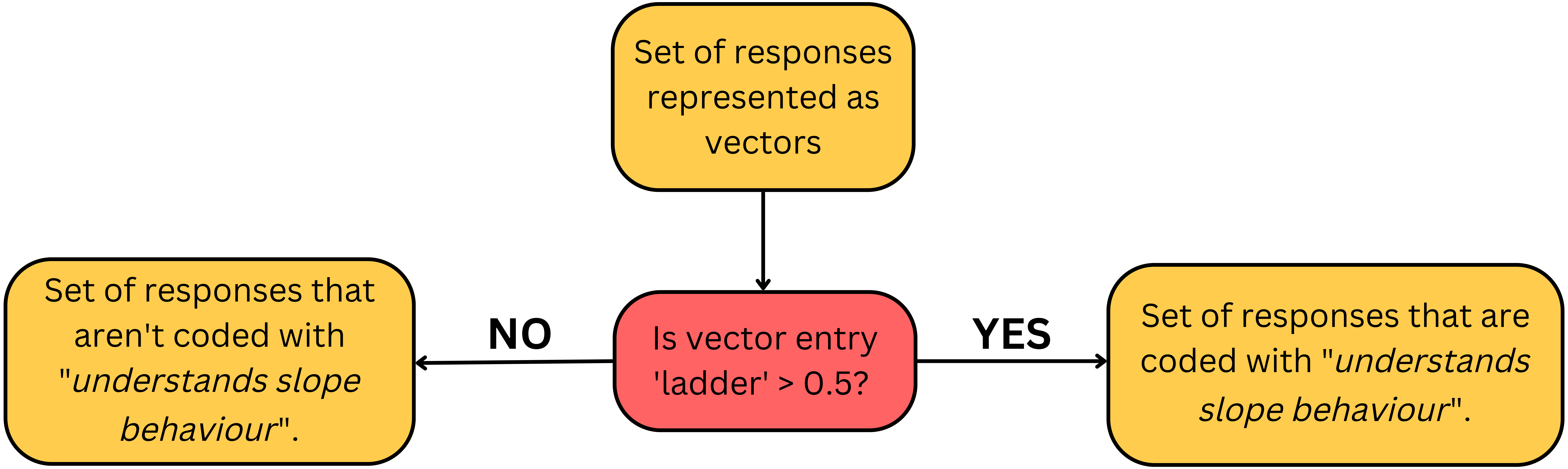
Future CBA implementations can allow for these interventions and supports to be automatically delivered upon completion.

THANK YOU FOR YOUR TIME!
Please ask away with any questions you may have!

Or visit our Github for a demo!

Gradient Boosting Machine Classification

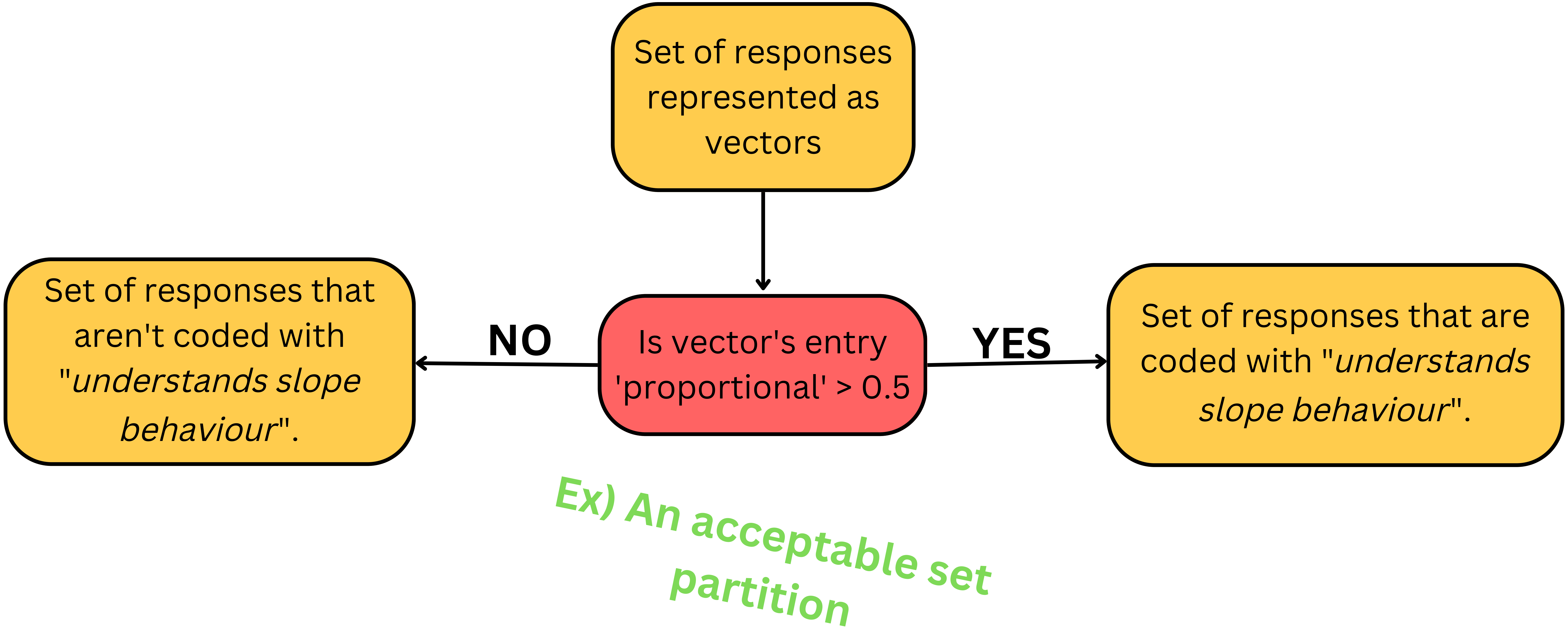
- The GBM is a decision tree that classifies the vector data.
  - The GBMs classification attempts to properly sort the labeled input.
  - It does this by partitioning the data with a one dimensional hyper-plane.



Ex) A terrible set partition

Gradient Boosting Machine Classification

- The GBM is a decision tree that classifies the vector data.
  - The GBMs classification attempts to properly sort the labeled input.
  - It does this by partitioning the data with a one dimensional hyper-plane.



Gradient Boosting Machine Classification

- The GBM is a decision tree that classifies the vector data.
    - The GBMs classification attempts to properly sort the labeled input.
    - It does this by partitioning the data with a one dimensional hyper-plane.
- Hundreds are built and discarded as a ***best*** classifier is iteratively designed.
    - Weak learners support the tree by reclassifying residuals.
- Separate, parallel GBMs must be trained for each code.
- Cross-validation grid search finds each model's best parameters:
    - parameters: sample size, tree depth, number of estimators, learning rate.
- Using the trained model:
    - Influential tokens for partitioning can be examined.
    - New data can be automatically coded by the model.