ARIA - Participation

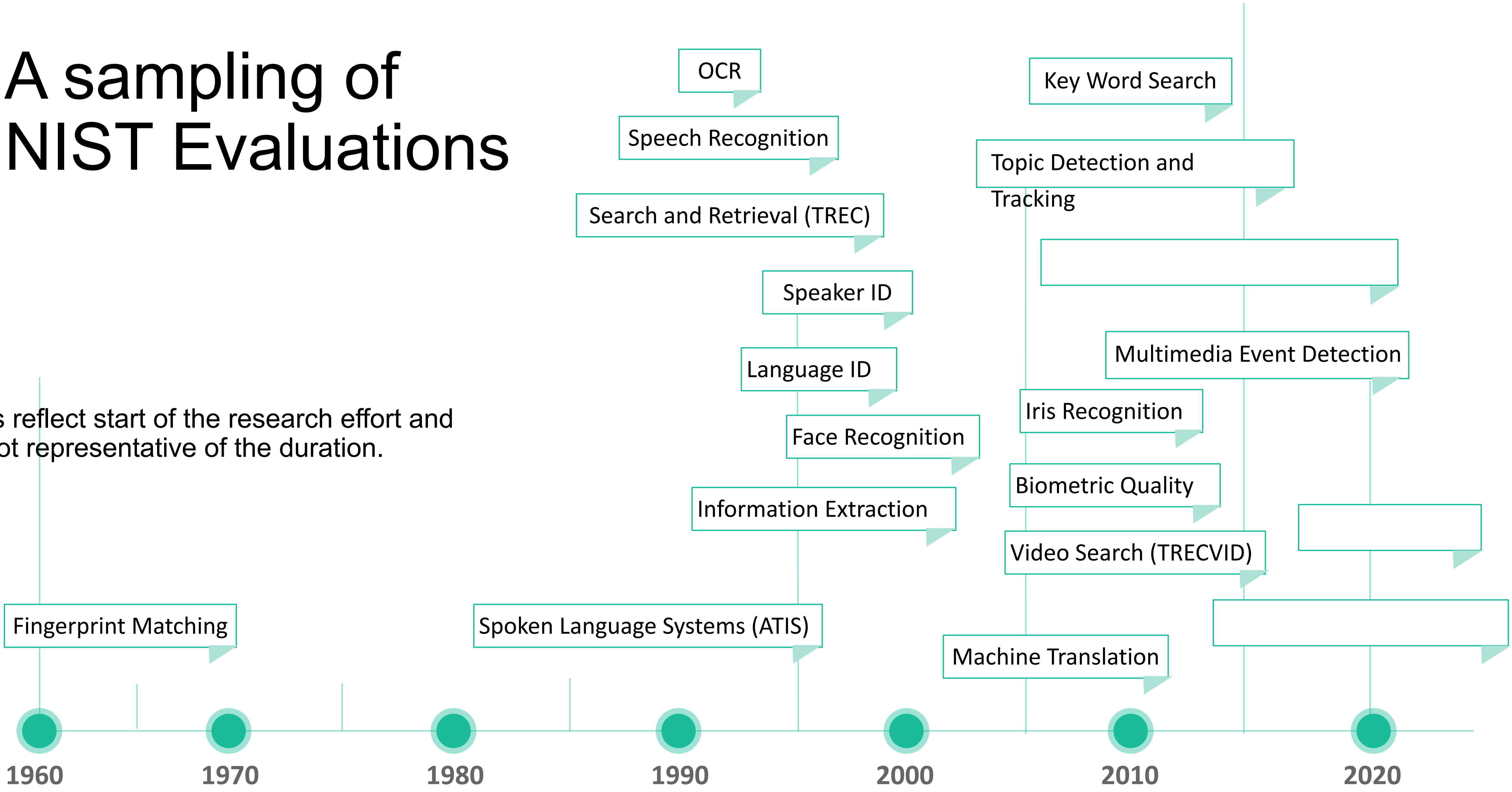
as a Model Team

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Knowledge Base Population (TAC)

Media Forensics

A sampling of NIST Evaluations

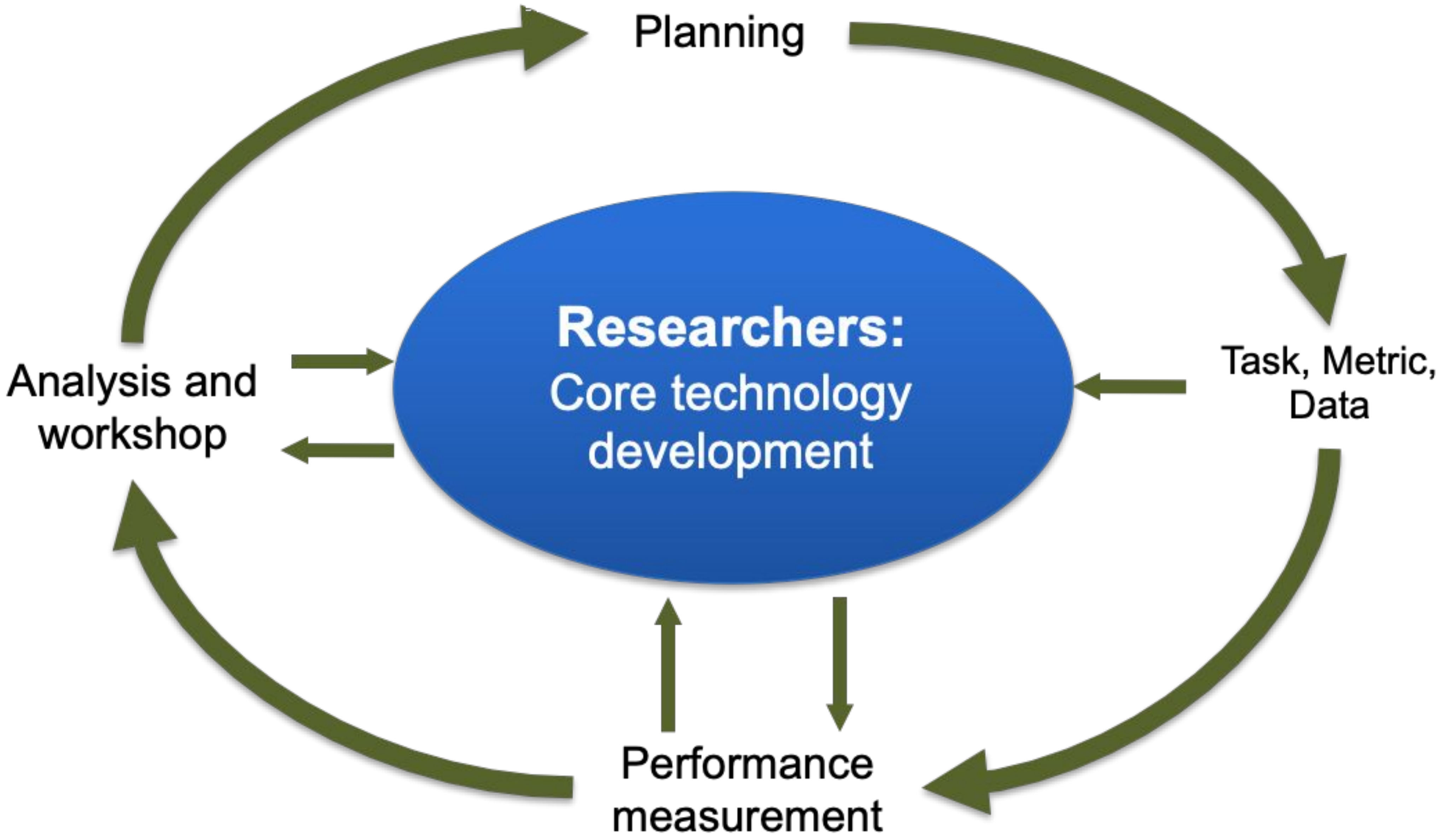


Dates reflect start of the research effort and are not representative of the duration.

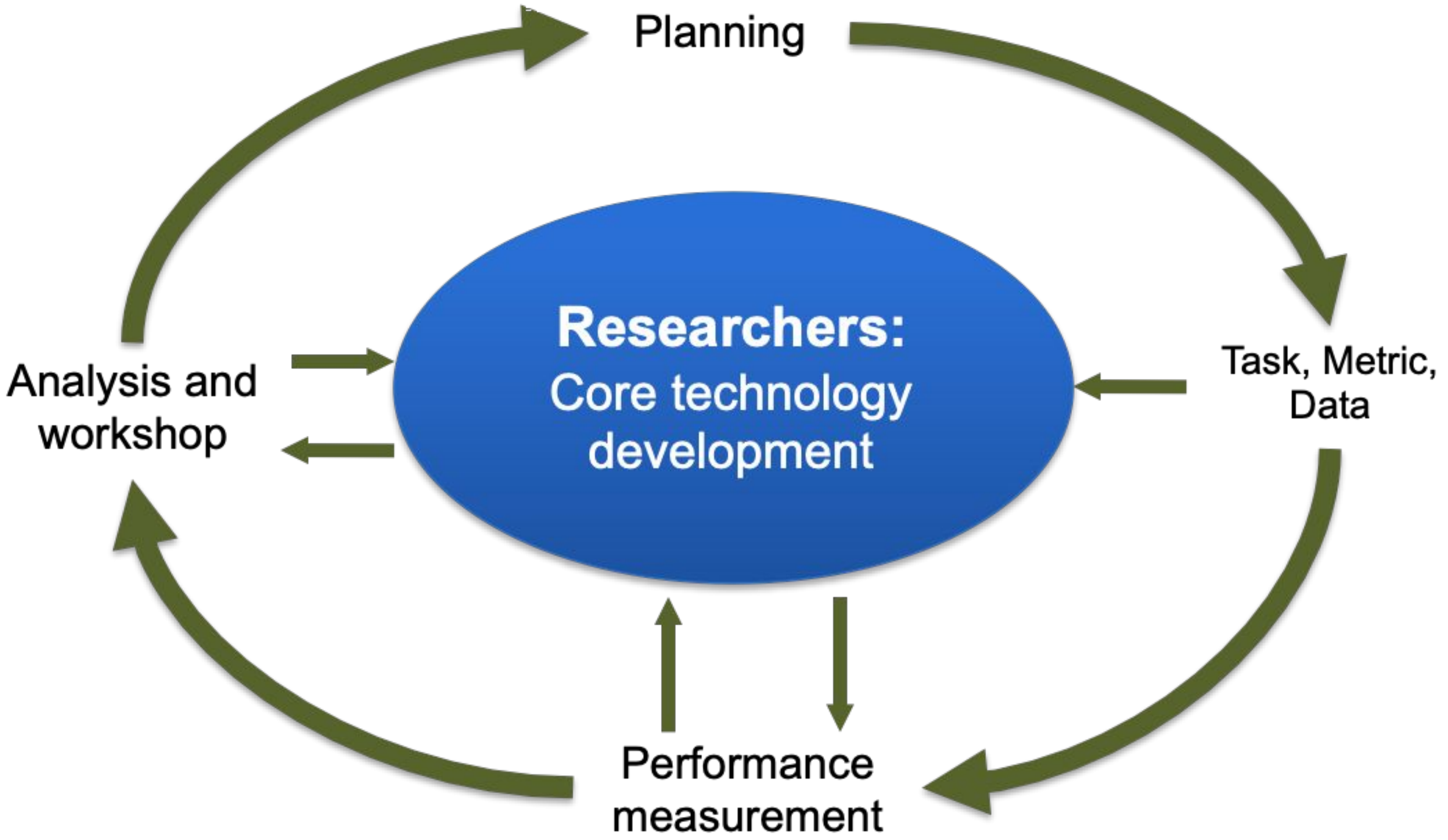
Researchers:

Core technology development

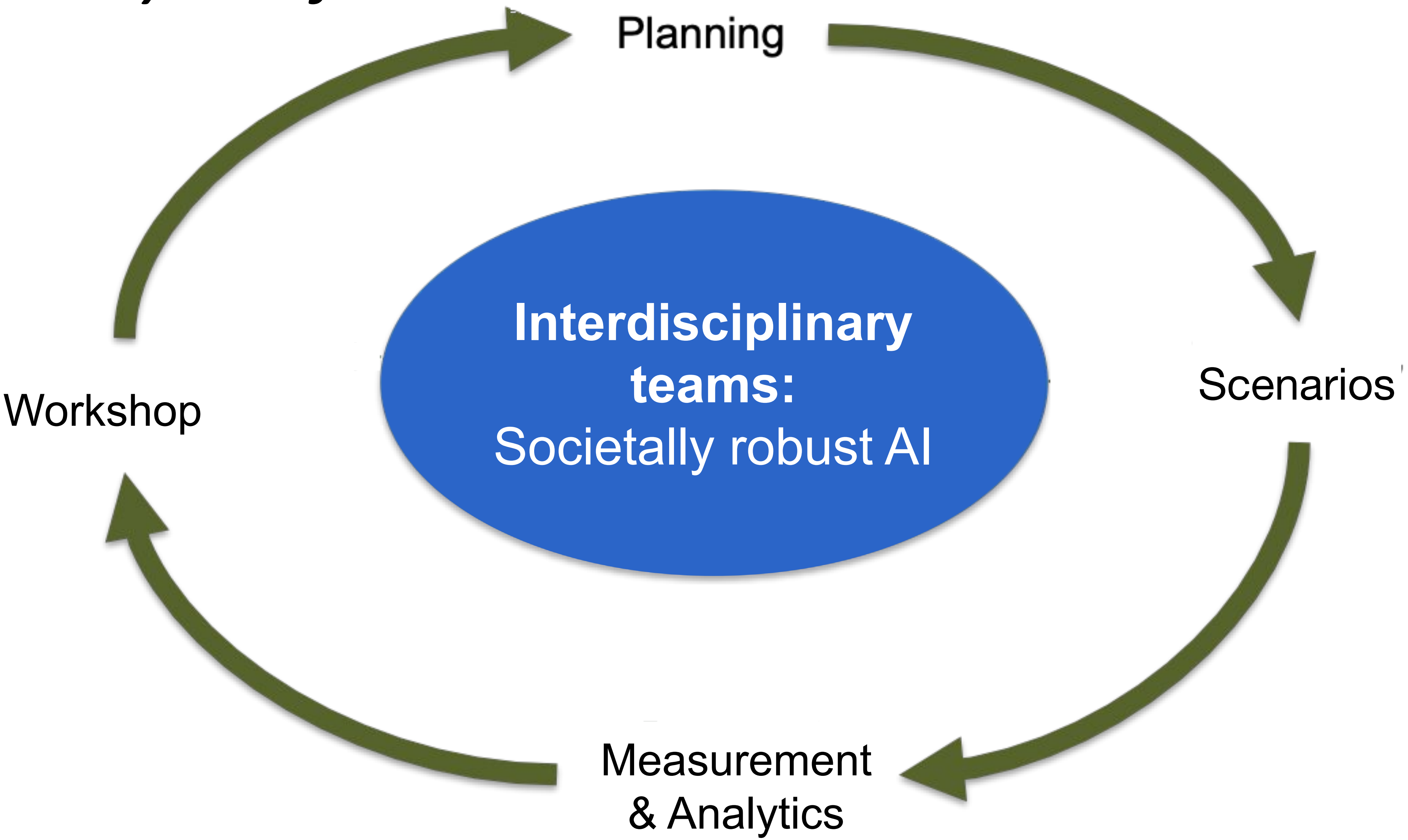
Virtuous Cycle of Evaluation



Virtuous Cycle of Evaluation



ARIA's Cycle of Evaluation



Planning

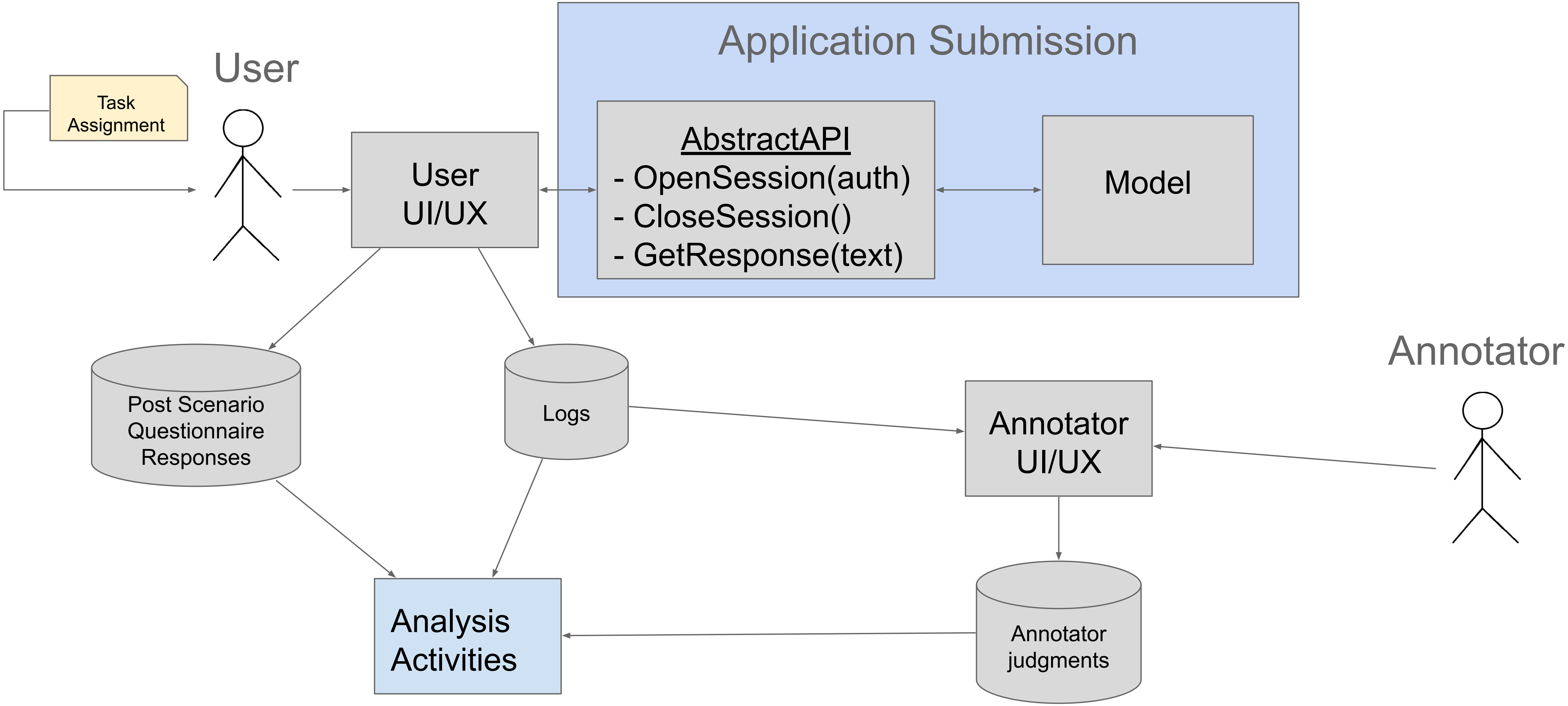
Scenarios

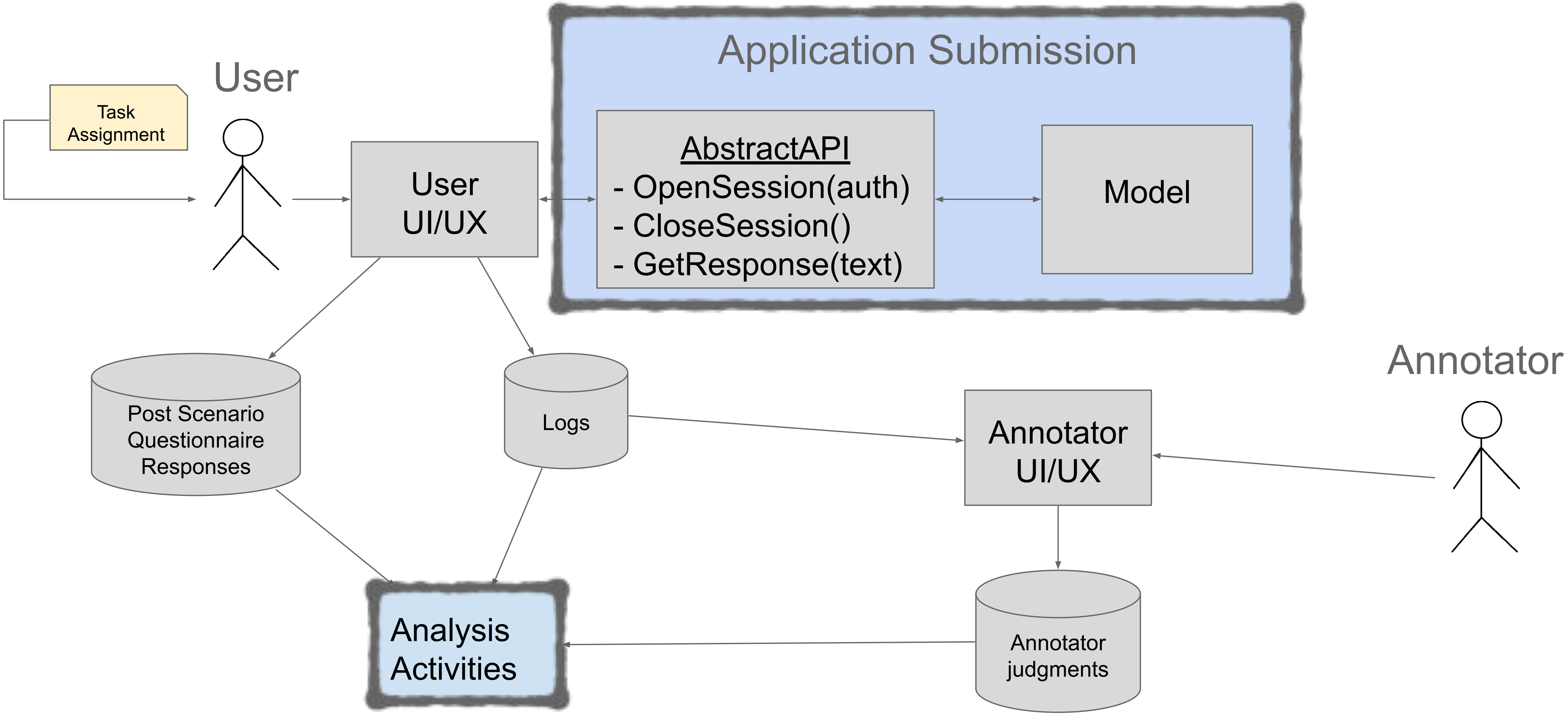
Measurement & Analytics

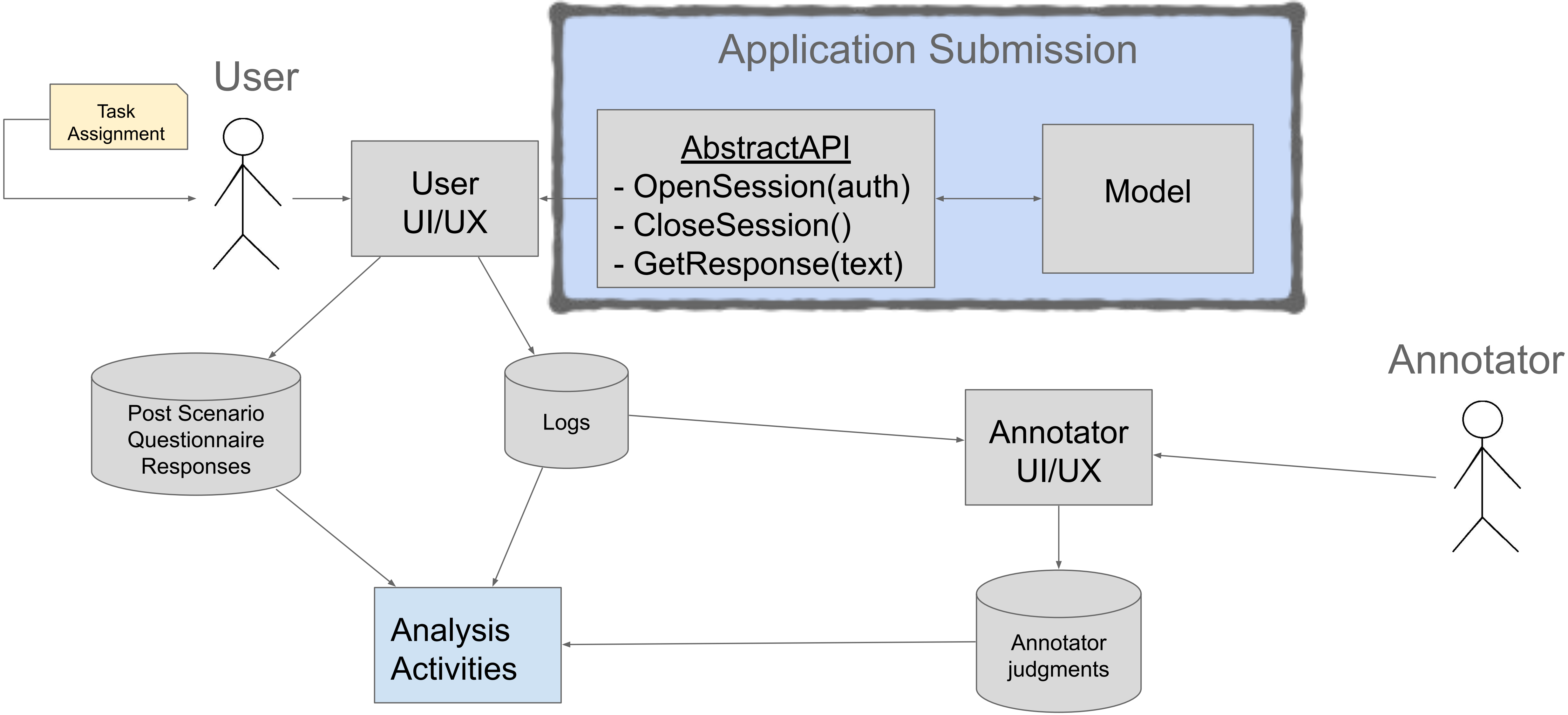
Workshop

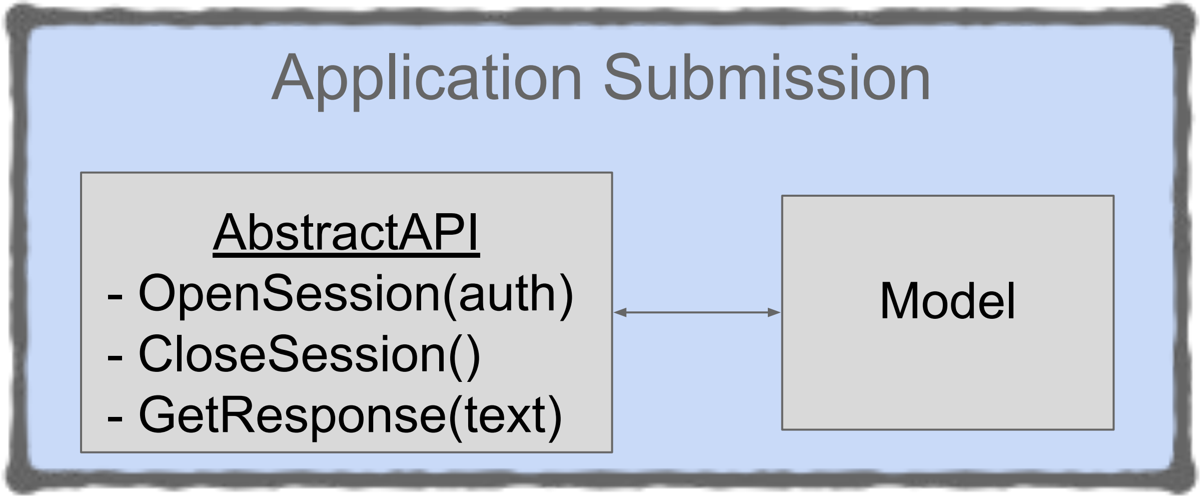
Interdisciplinary teams:
Societally robust AI

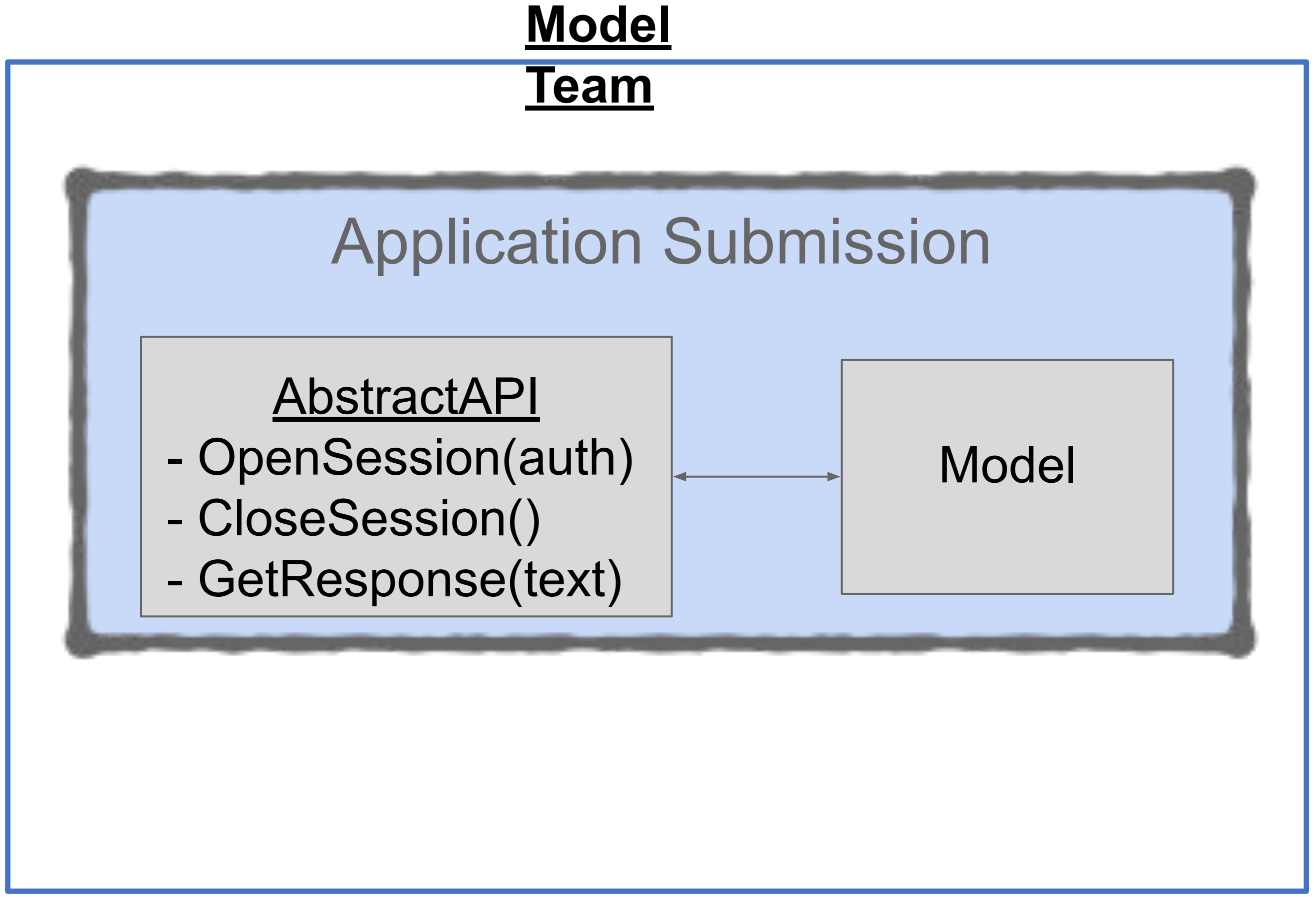
How to provide a model for ARIA …


Application Submission
User
Task Assignment
User UI/UX
AbstractAPI
- OpenSession(auth)
- CloseSession()
- GetResponse(text)
Model
Post Scenario Questionnaire Responses
Logs
Annotator
Annotator UI/UX
Analysis Activities
Annotator judgments


Application Submission
User
Task Assignment
User UI/UX
AbstractAPI
- OpenSession(auth)
- CloseSession()
- GetResponse(text)
Model
Post Scenario Questionnaire Responses
Logs
Annotator
Annotator UI/UX
Analysis Activities
Annotator judgments


Application Submission
User
Task Assignment
User UI/UX
AbstractAPI
- OpenSession(auth)
- CloseSession()
- GetResponse(text)
Model
Annotator
Post Scenario Questionnaire Responses
Logs
Annotator UI/UX
Analysis Activities
Annotator judgments


Application Submission
AbstractAPI
- OpenSession(auth)
- CloseSession()
- GetResponse(text)
Model



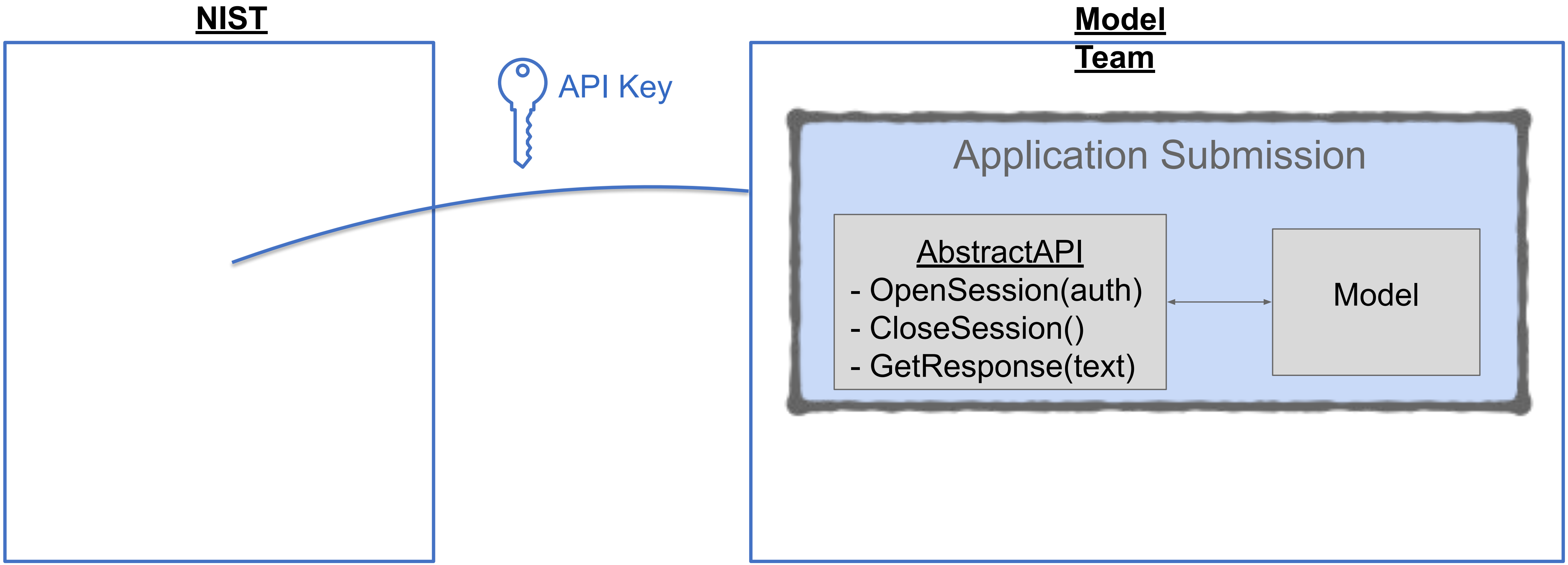
Model Team

Application Submission

AbstractAPI

- OpenSession(auth)
- CloseSession()
- GetResponse(text)

Model



NIST

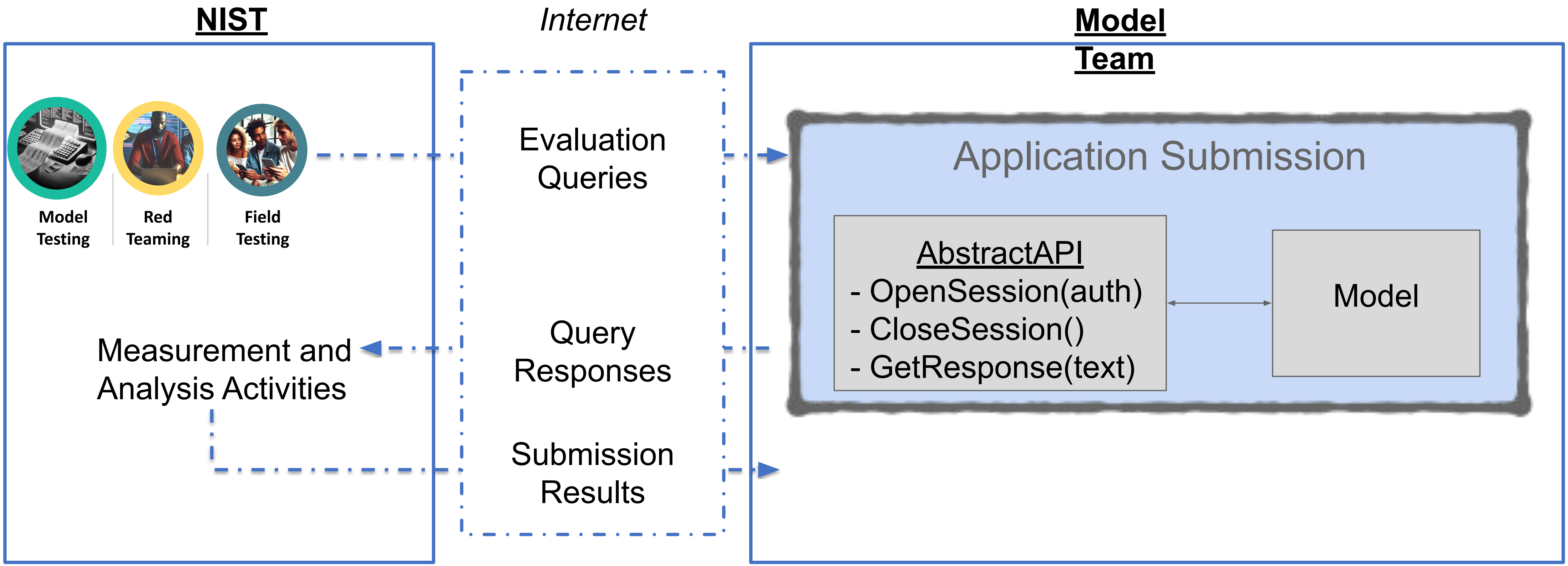
API Key

Model Team

Application Submission

AbstractAPI
- OpenSession(auth)
- CloseSession()
- GetResponse(text)

Model



NIST

Model Testing

Red Teaming

Field Testing

Measurement and Analysis Activities

Internet

Evaluation Queries

Query Responses

Submission Results

Model Team

Application Submission

AbstractAPI

- OpenSession(auth)
- CloseSession()
- GetResponse(text)

Model

Data Transfer Agreement (DTA)

Small lift, big benefit …

CoRIx is a new **multidimensional measurement instrument** measuring "contextual robustness" – the ability of an AI system to **maintain its level of functionality in a variety of real world contexts** and related user expectations.

Desiderata:

- Simple as possible, but no simpler
- Meaningful and informative
- Intuitive/provides the receiver with an accurate impression/interpretation
- Is valid
- Minimizes obfuscation of info
- Minimize opportunity for unproductive gaming of measurement
- Minimize likelihood of misrepresentation or misinterpretation
- Fit for purpose (e.g., rank ordering systems vs assessing properties of a given system)
- Appropriately captures context, including the social systems in which the AI operates
- Able to be aligned with application/task needs
- Repeatable and reproducible
- Is able to include estimates of uncertainty
- Provides a partial ordering
- Is well-conditioned

x

CoRIx is a new **multidimensional measurement instrument** measuring “contextual robustness” – the ability of an AI system to **maintain its level of functionality in a variety of real world contexts** and related user expectations.

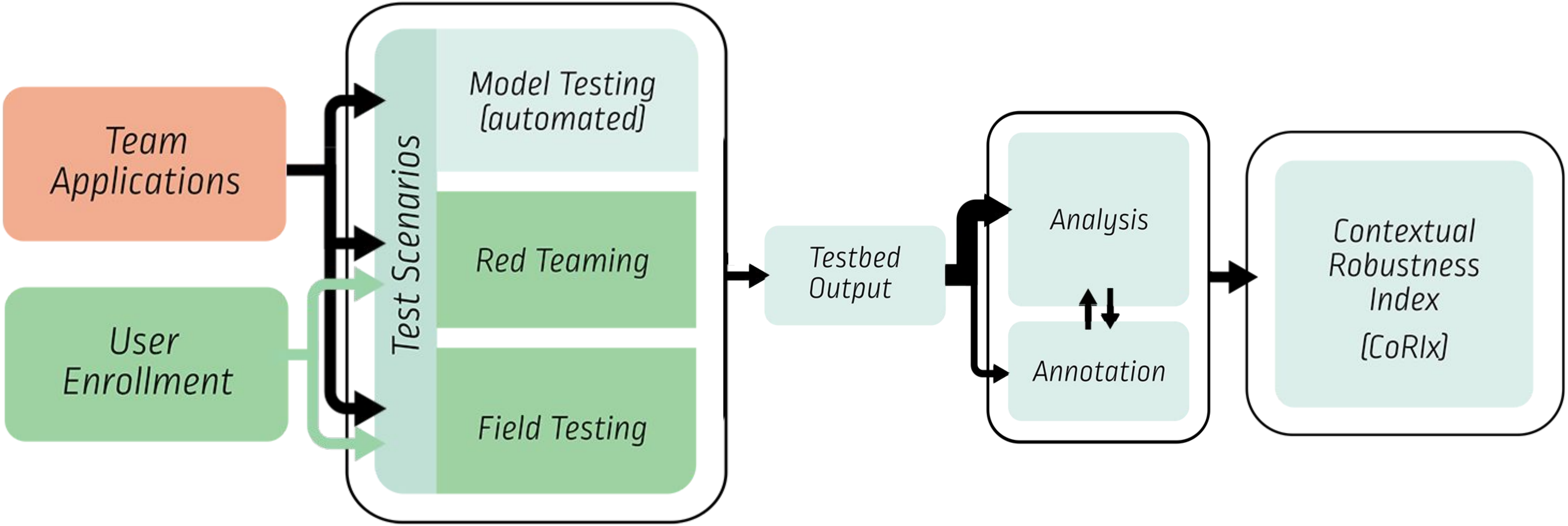
Planned Approach:

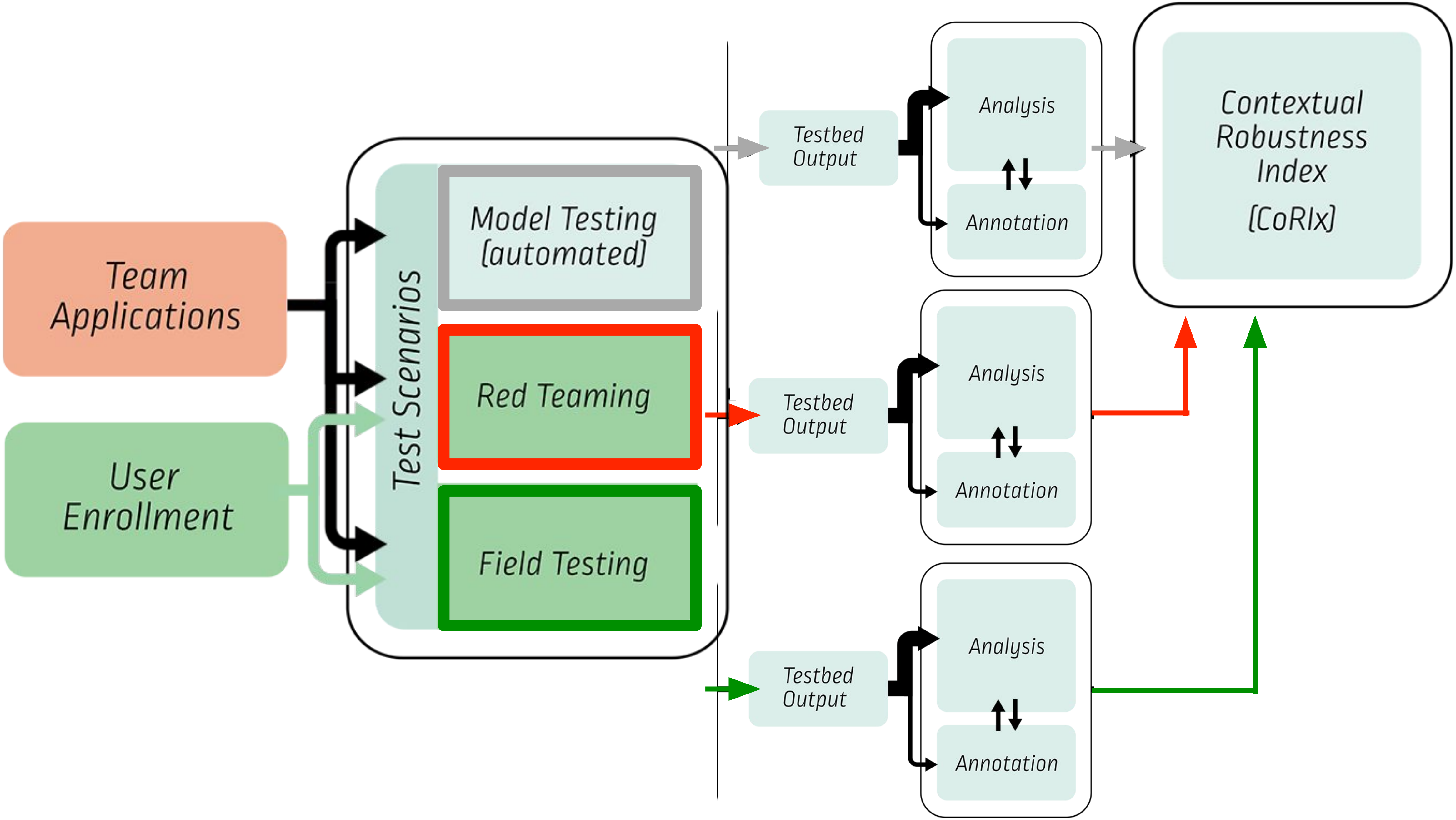
- **Mixed-methods**
- Measurement dimensions to include seven **AI RMF trustworthy characteristics**
- Not a single real-valued score CoRIx output is a **tree structure**, * where
  - each additional level in the tree provides more detailed information
  - each parent node provides a summary of its children,
  - associated with each node in the tree is a method for summarizing its children

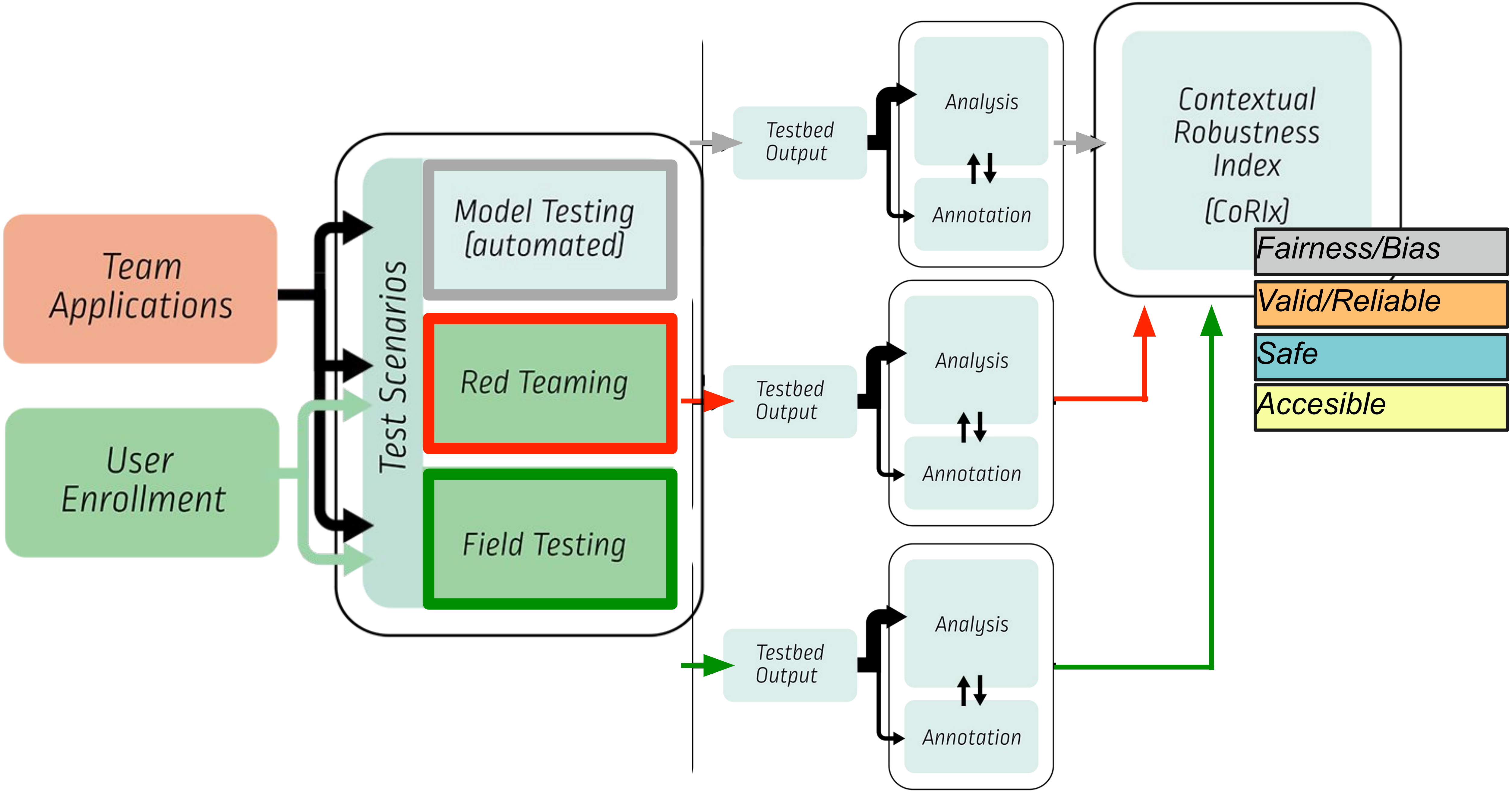
CoRIx scoring can be understood as **mapping between input data and tree-structures** with summary-annotated nodes.

* technically a directed acyclic graph

Now, an example …


Team Applications
User Enrollment
Test Scenarios
Model Testing (automated)
Red Teaming
Field Testing
Testbed Output
Analysis
Annotation
Contextual Robustness Index
(CoRIx)


Team Applications
User Enrollment
Test Scenarios
Model Testing (automated)
Red Teaming
Field Testing
Testbed Output
Analysis
Annotation
Testbed Output
Analysis
Annotation
Testbed Output
Analysis
Annotation
Contextual Robustness Index
(CoRIx)

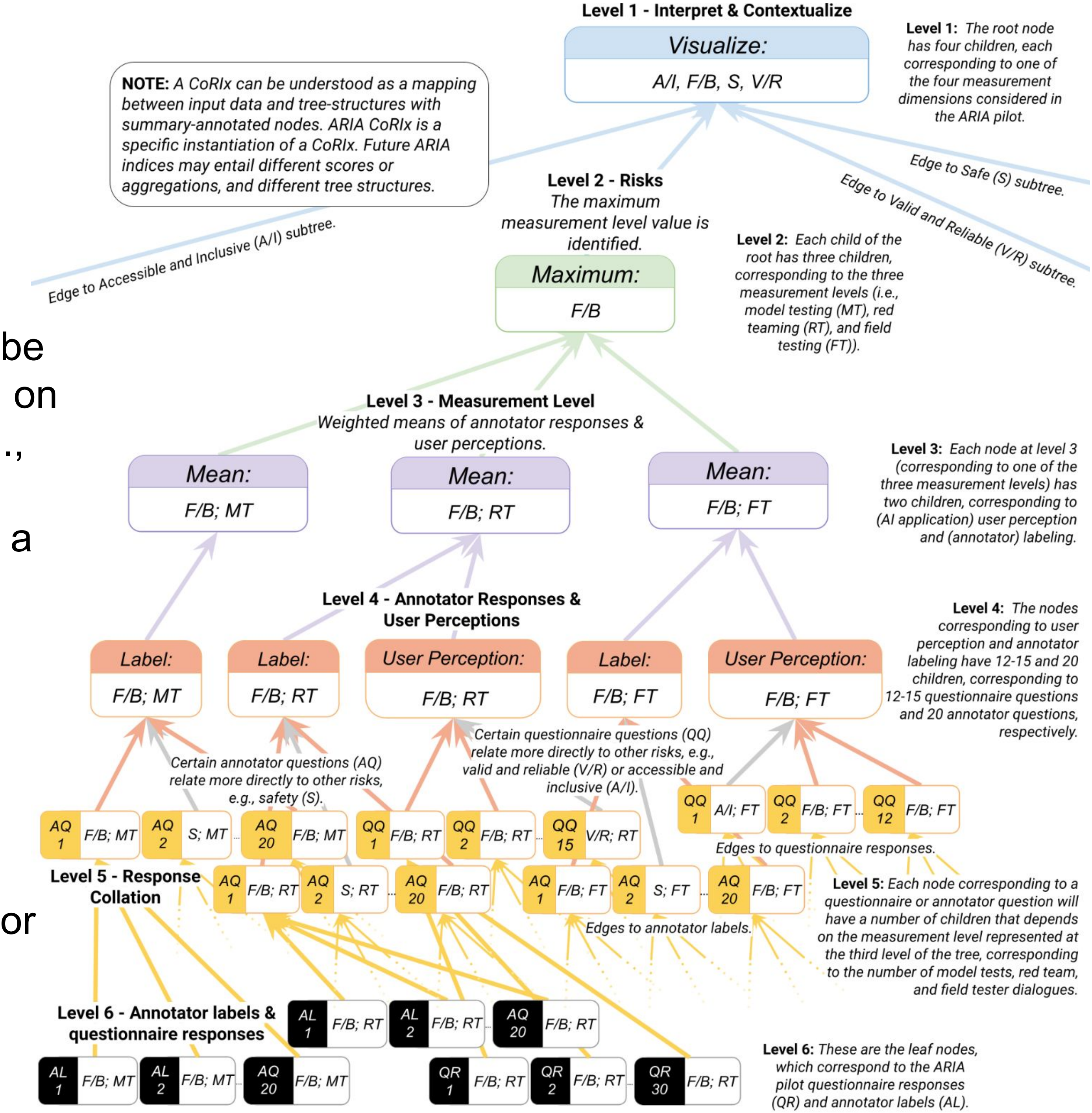

Team Applications
User Enrollment
Test Scenarios
Model Testing (automated)
Red Teaming
Field Testing
Testbed Output
Analysis
Annotation
Testbed Output
Analysis
Annotation
Testbed Output
Analysis
Annotation
Contextual Robustness Index (CoRIx)
Fairness/Bias
Valid/Reliable
Safe
Accesible

Summary methods can be many things, depending on their children; could, e.g., be weighted average, a textual summary, a plot, a combination thereof.

Similarly, many tree topologies are possible.

Score is whole tree

Can consider only root, or any tree depth, subtree, branch, …



NOTE: *A CoRIx can be understood as a mapping between input data and tree-structures with summary-annotated nodes. ARIA CoRIx is a specific instantiation of a CoRIx. Future ARIA indices may entail different scores or aggregations, and different tree structures.*

Level 1: *The root node has four children, each corresponding to one of the four measurement dimensions considered in the ARIA pilot.*

Level 2: *Each child of the root has three children, corresponding to the three measurement levels (i.e., model testing (MT), red teaming (RT), and field testing (FT)).*

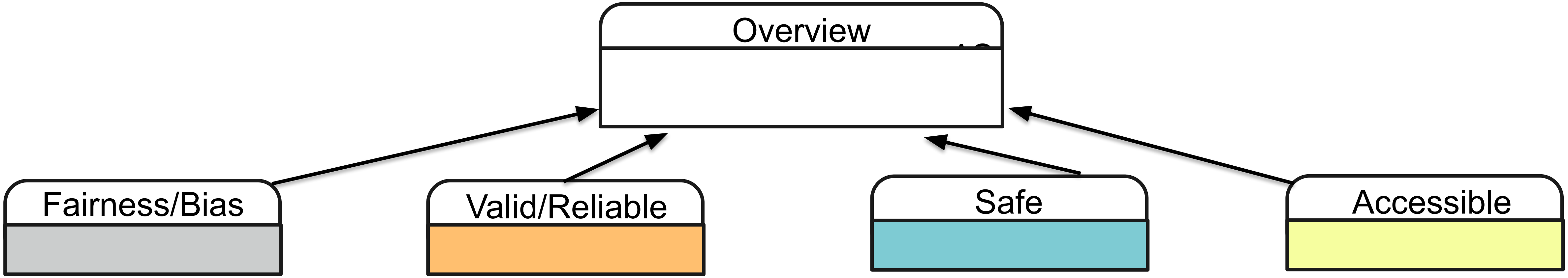
Level 3: *Each node at level 3 (corresponding to one of the three measurement levels) has two children, corresponding to (AI application) user perception and (annotator) labeling.*

Level 4: *The nodes corresponding to user perception and annotator labeling have 12-15 and 20 children, corresponding to 12-15 questionnaire questions and 20 annotator questions, respectively.*

Level 5: *Each node corresponding to a questionnaire or annotator question will have a number of children that depends on the measurement level represented at the third level of the tree, corresponding to the number of model tests, red team, and field tester dialogues.*

Level 6: *These are the leaf nodes, which correspond to the ARIA pilot questionnaire responses (QR) and annotator labels (AL).*

Level 1 - Interpret & Contextualize
Visualize:
A/I, F/B, S, V/R
Edge to Accessible and Inclusive (A/I) subtree.
Edge to Safe (S) subtree.
Edge to Valid and Reliable (V/R) subtree.
Level 2 - Risks
The maximum measurement level value is identified.
Maximum:
F/B
Level 3 - Measurement Level
Weighted means of annotator responses & user perceptions.
Mean:
F/B; MT
Mean:
F/B; RT
Mean:
F/B; FT
Level 4 - Annotator Responses & User Perceptions
Label:
F/B; MT
Label:
F/B; RT
User Perception:
F/B; RT
Label:
F/B; FT
User Perception:
F/B; FT
Certain annotator questions (AQ) relate more directly to other risks, e.g., safety (S).
Certain questionnaire questions (QQ) relate more directly to other risks, e.g., valid and reliable (V/R) or accessible and inclusive (A/I).
AQ 1 F/B; MT
AQ 2 S; MT
AQ 20 F/B; MT
QQ 1 F/B; RT
QQ 2 F/B; RT
QQ 15 V/R; RT
QQ 1 A/I; FT
QQ 2 F/B; FT
QQ 12 F/B; FT
Edges to questionnaire responses.
Level 5 - Response Collation
AQ 1 F/B; RT
AQ 2 S; RT
AQ 20 F/B; RT
AQ 1 F/B; FT
AQ 2 S; FT
AQ 20 F/B; FT
Edges to annotator labels.
Level 6 - Annotator labels & questionnaire responses
AL 1 F/B; RT
AL 2 F/B; RT
AQ 20 F/B; RT
AL 1 F/B; MT
AL 2 F/B; MT
AQ 20 F/B; MT
QR 1 F/B; RT
QR 2 F/B; RT
QR 30 F/B; RT



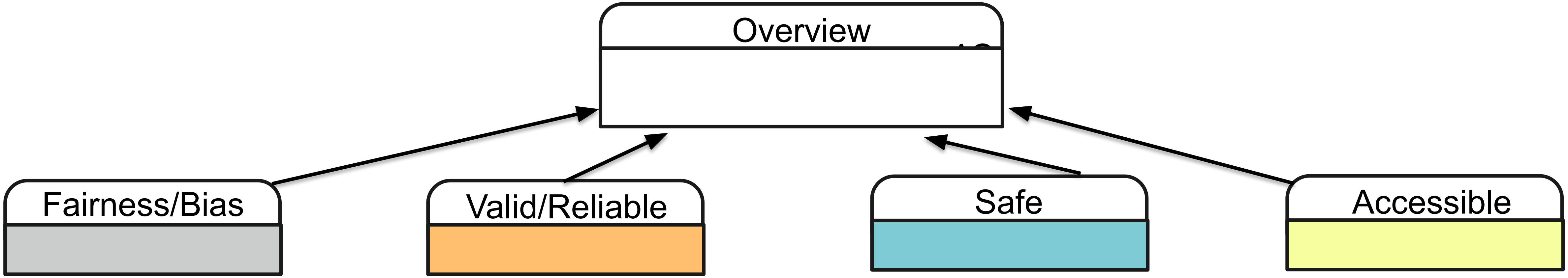
Overview

Fairness/Bias

Valid/Reliable

Safe

Accessible



Overview

Fairness/Bias

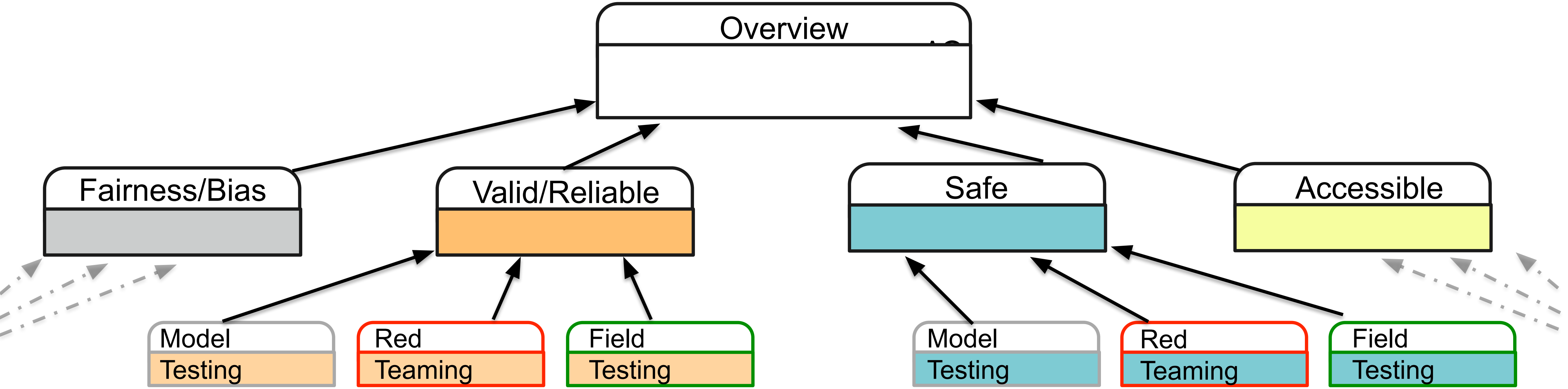
Valid/Reliable

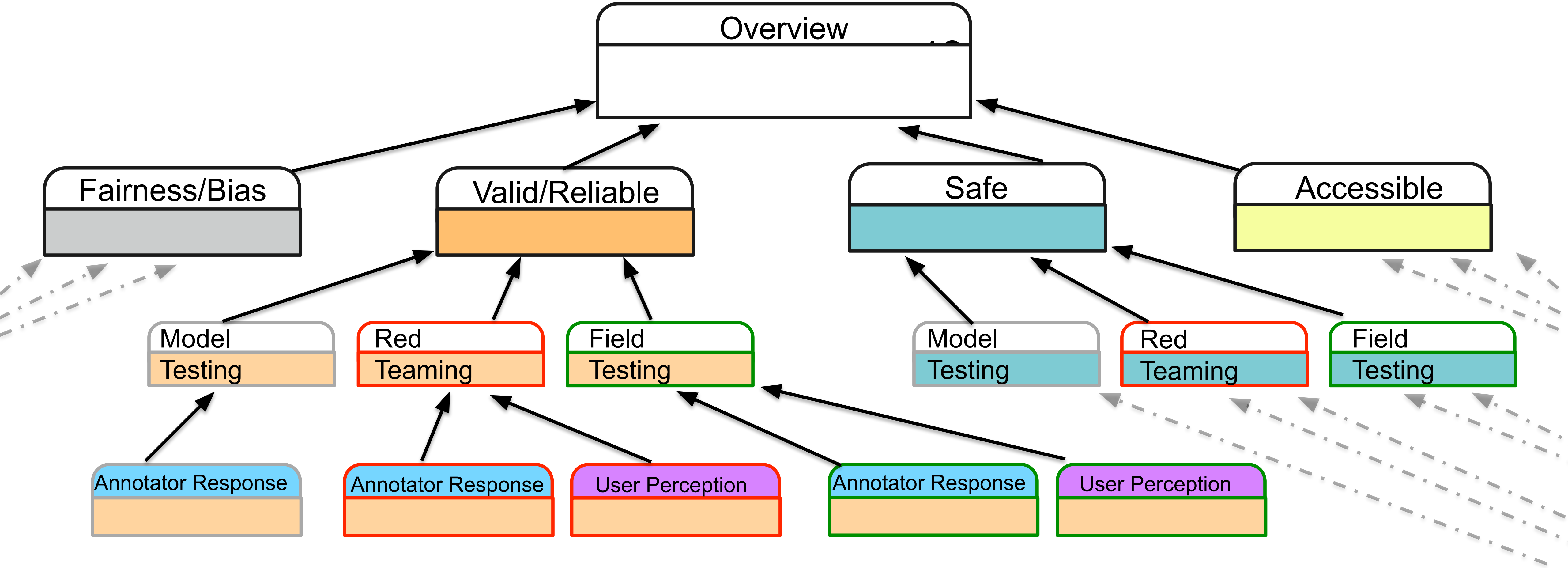
Safe

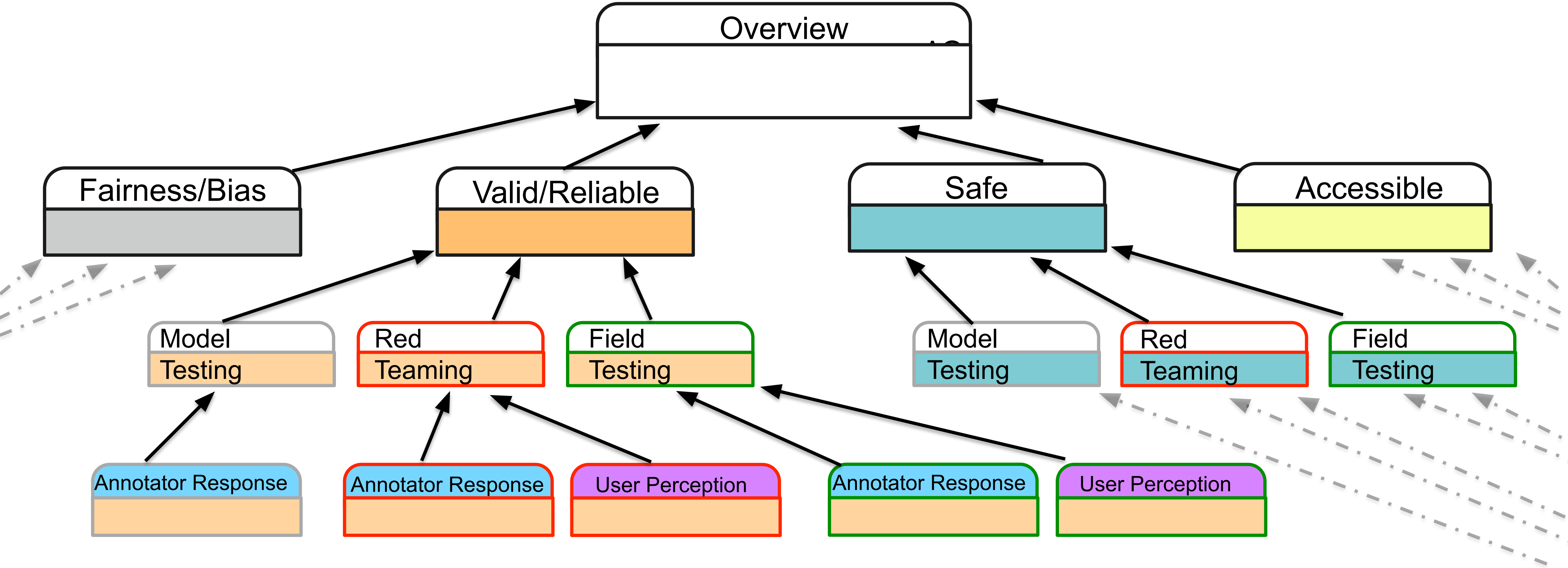
Accessible

Many summary methods are possible, e.g., a

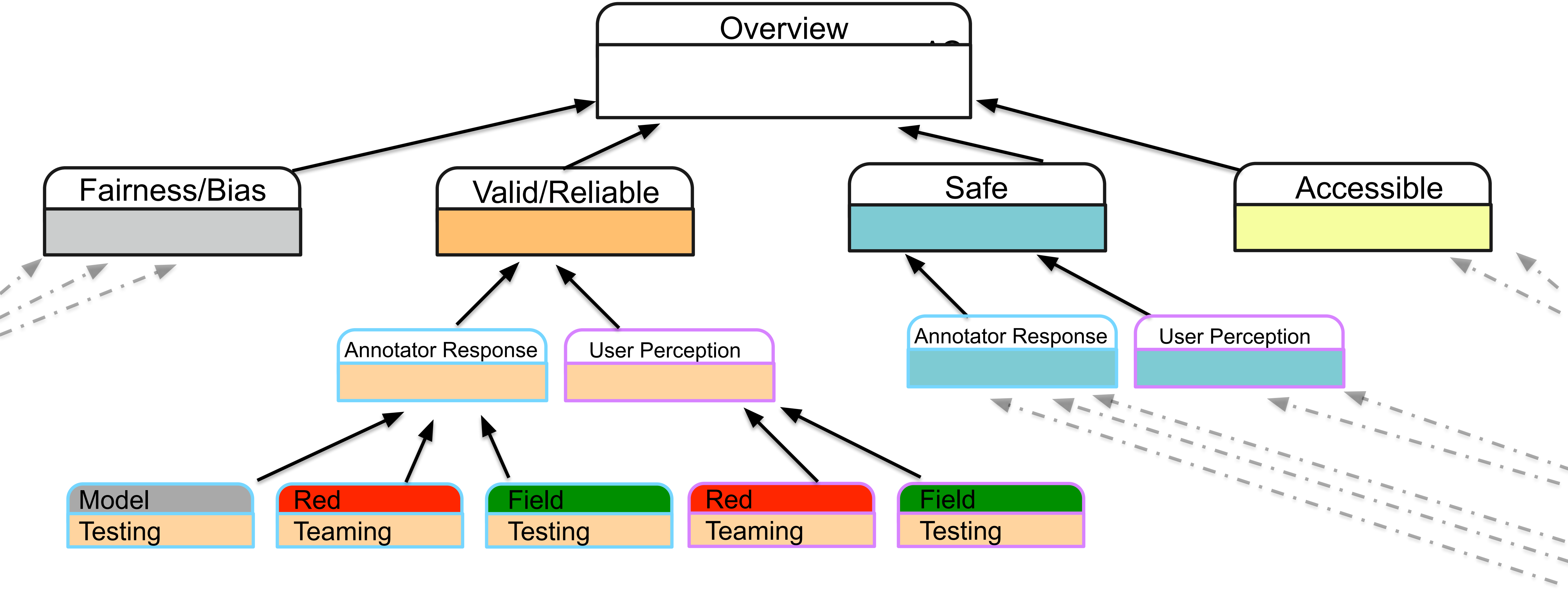
- *weighted average,*
- *textual summary*
- *plot*
- *combination thereof*


Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing

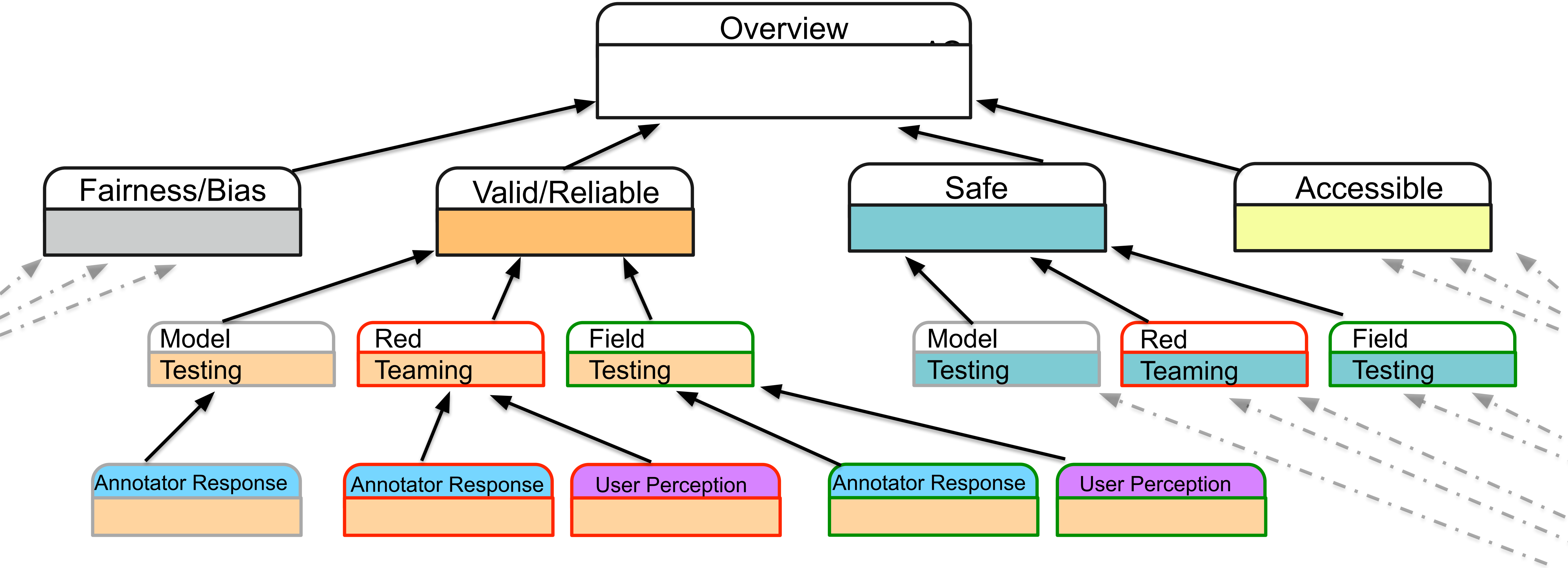

Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception



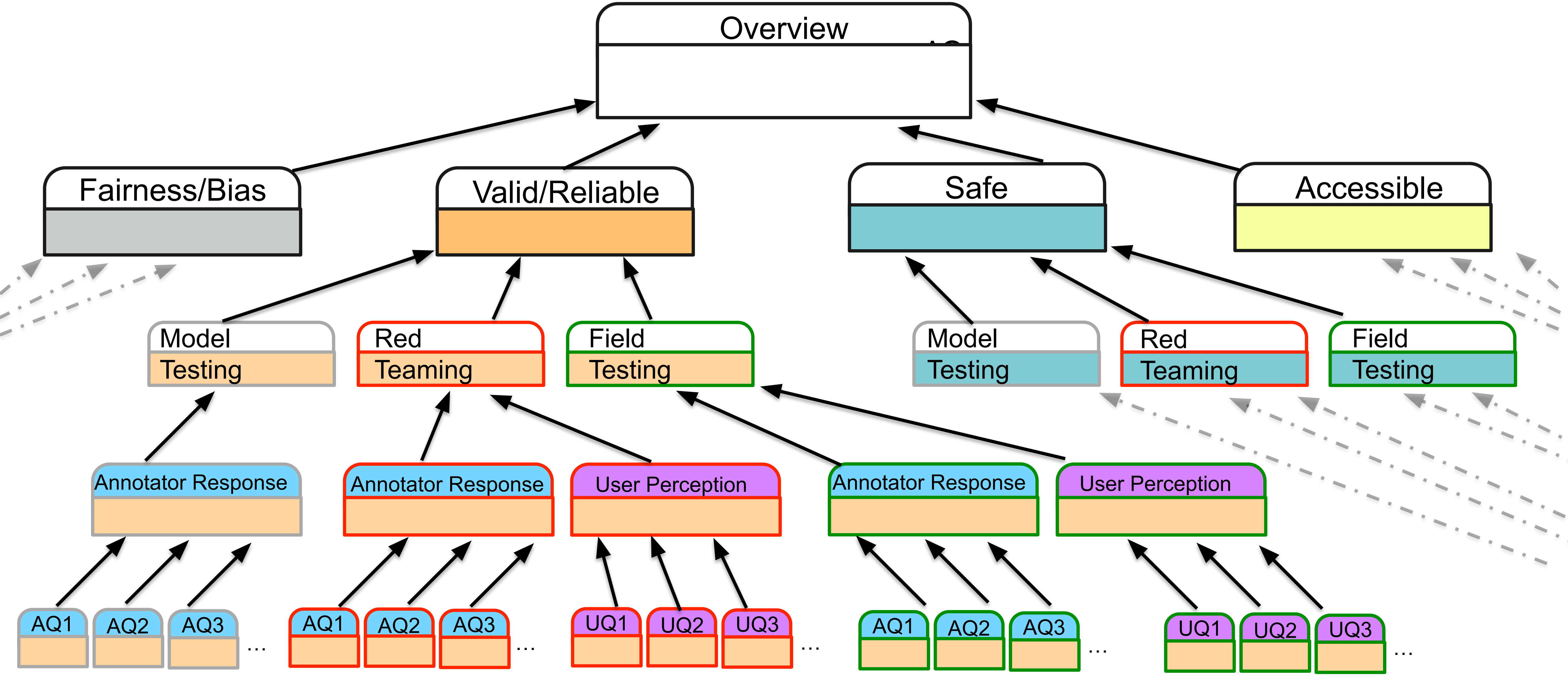
Many different tree topologies are possible



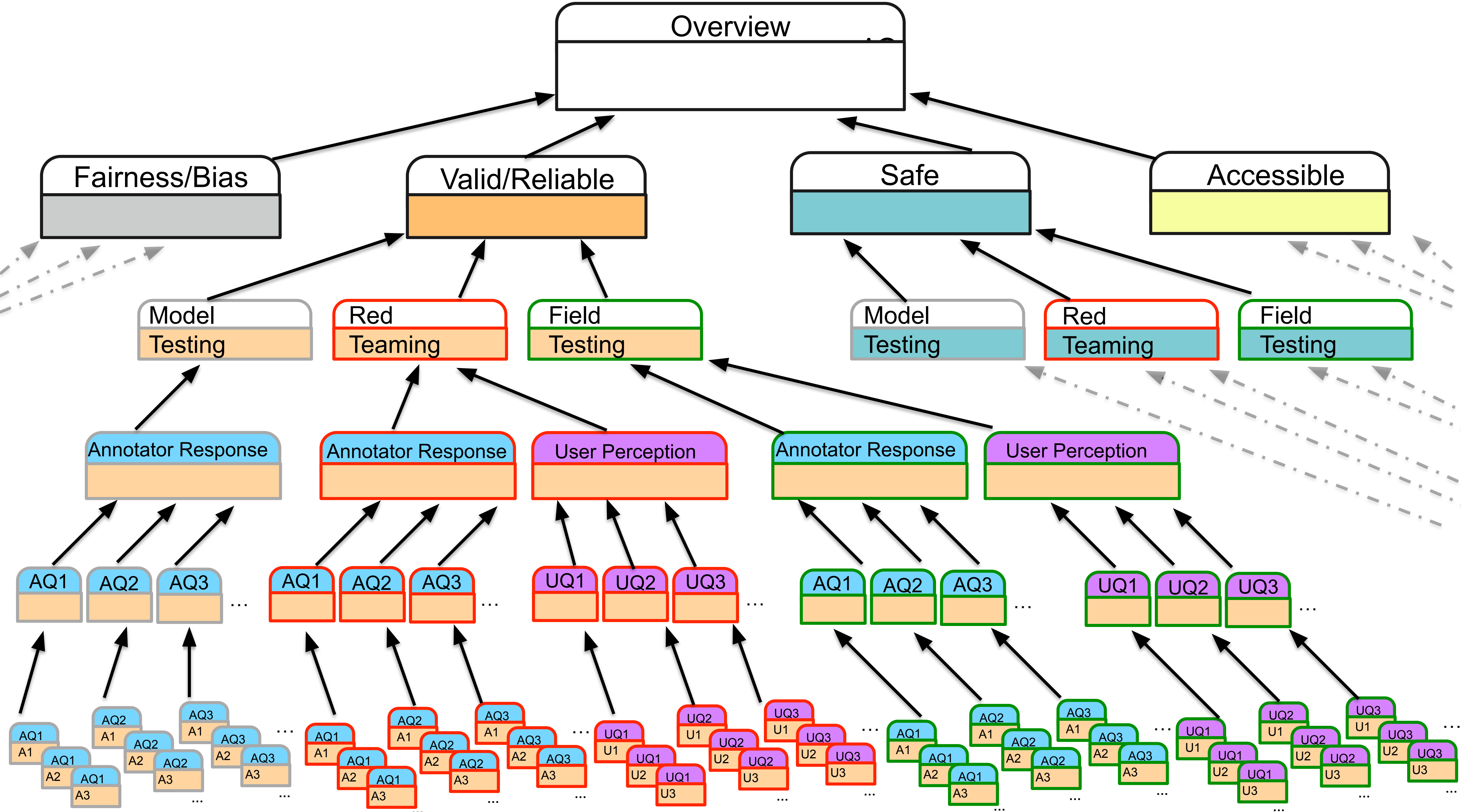
Many different tree topologies are possible



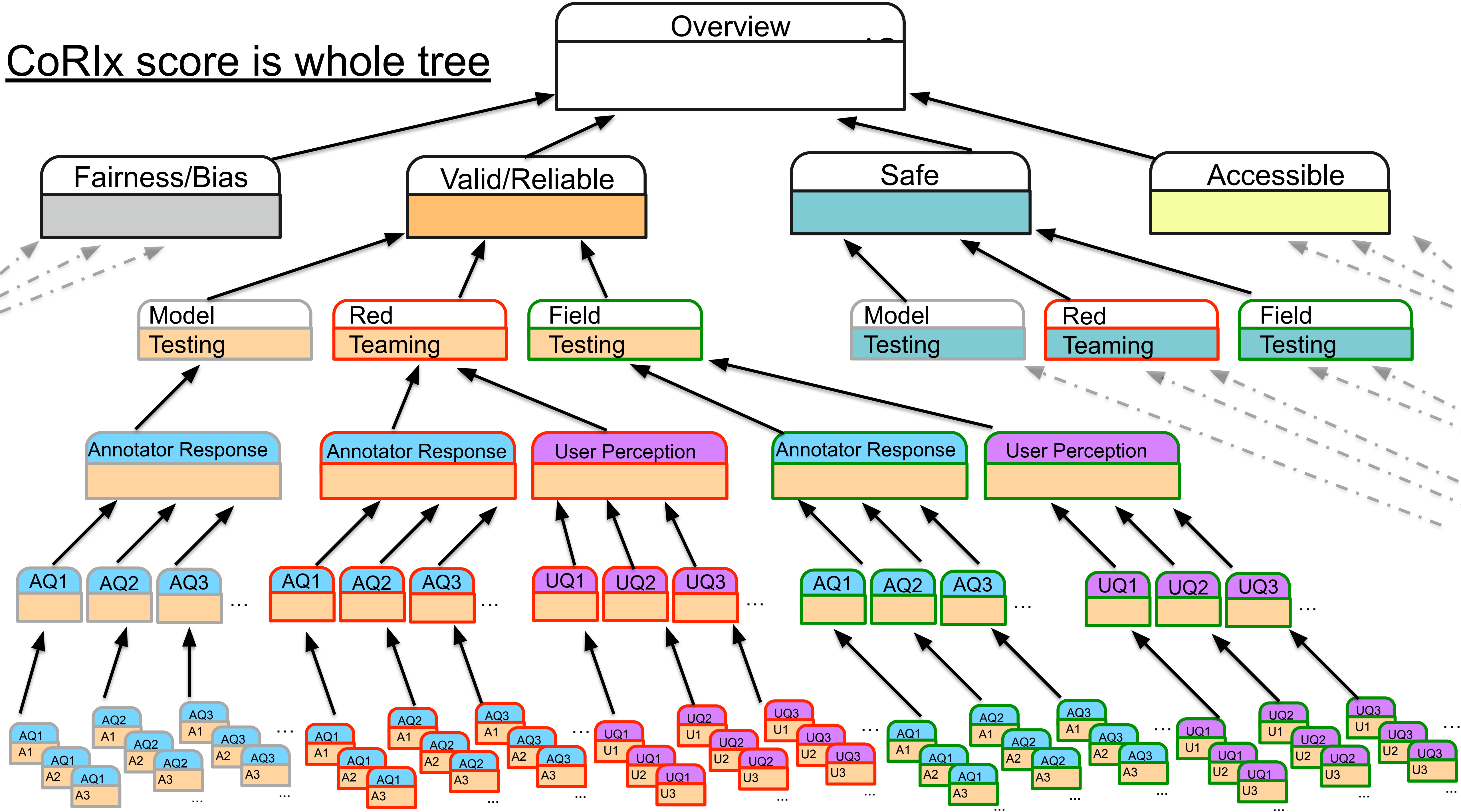
Many different tree topologies are possible

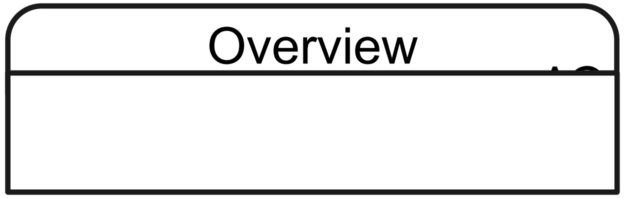

Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception
AQ1
AQ2
AQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...

...

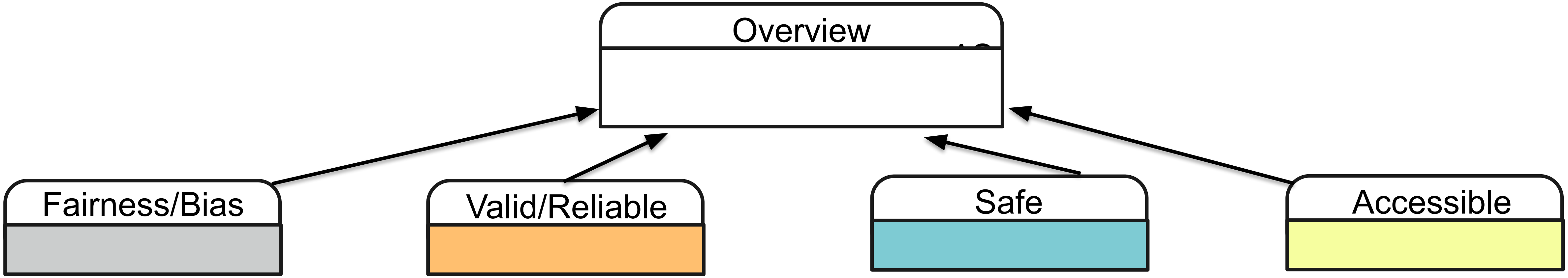

Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception
AQ1
AQ2
AQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
...
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
...
UQ1 U1
UQ1 U2
UQ1 U3
UQ2 U1
UQ2 U2
UQ2 U3
UQ3 U1
UQ3 U2
UQ3 U3
...
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
...
UQ1 U1
UQ1 U2
UQ1 U3
UQ2 U1
UQ2 U2
UQ2 U3
UQ3 U1
UQ3 U2
UQ3 U3
...

CoRIx score is whole tree





Overview



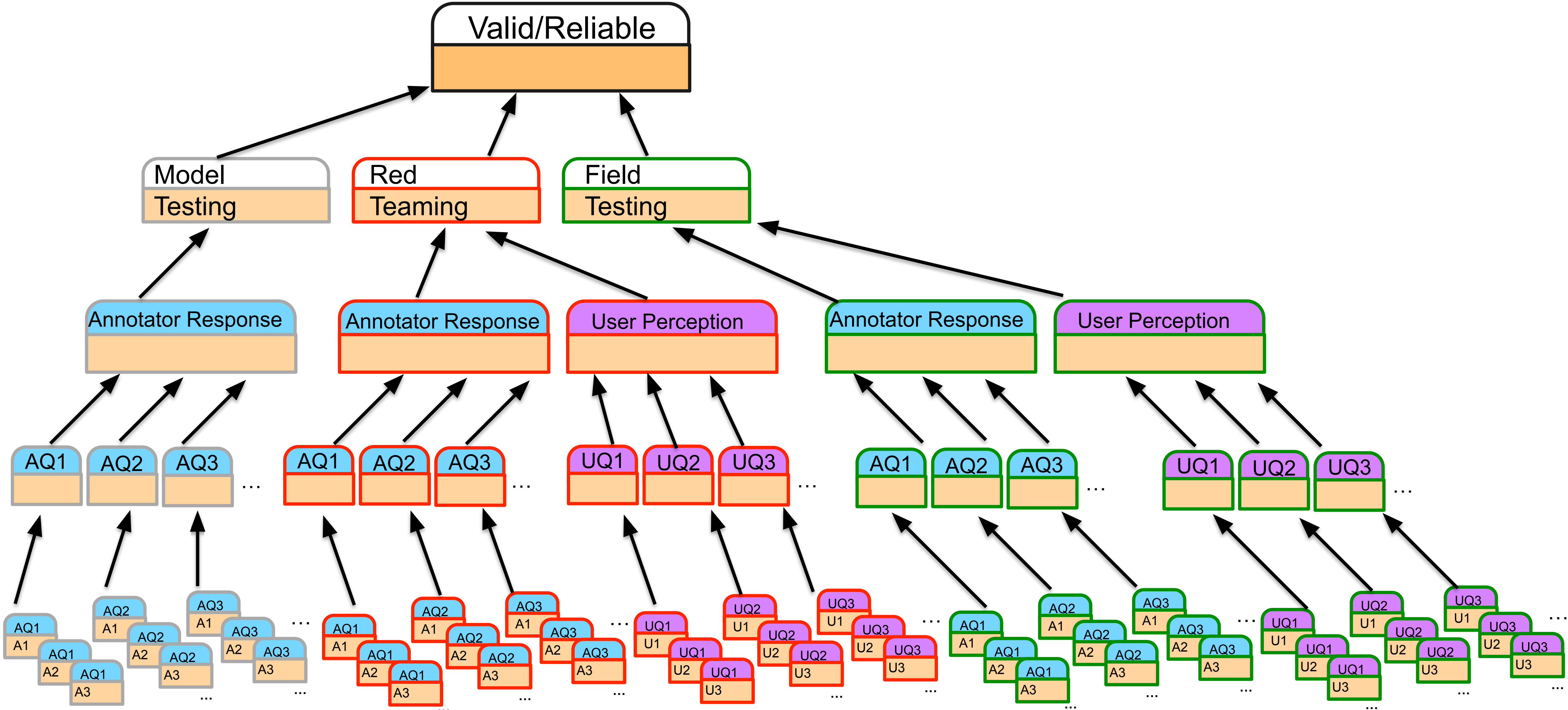
Overview

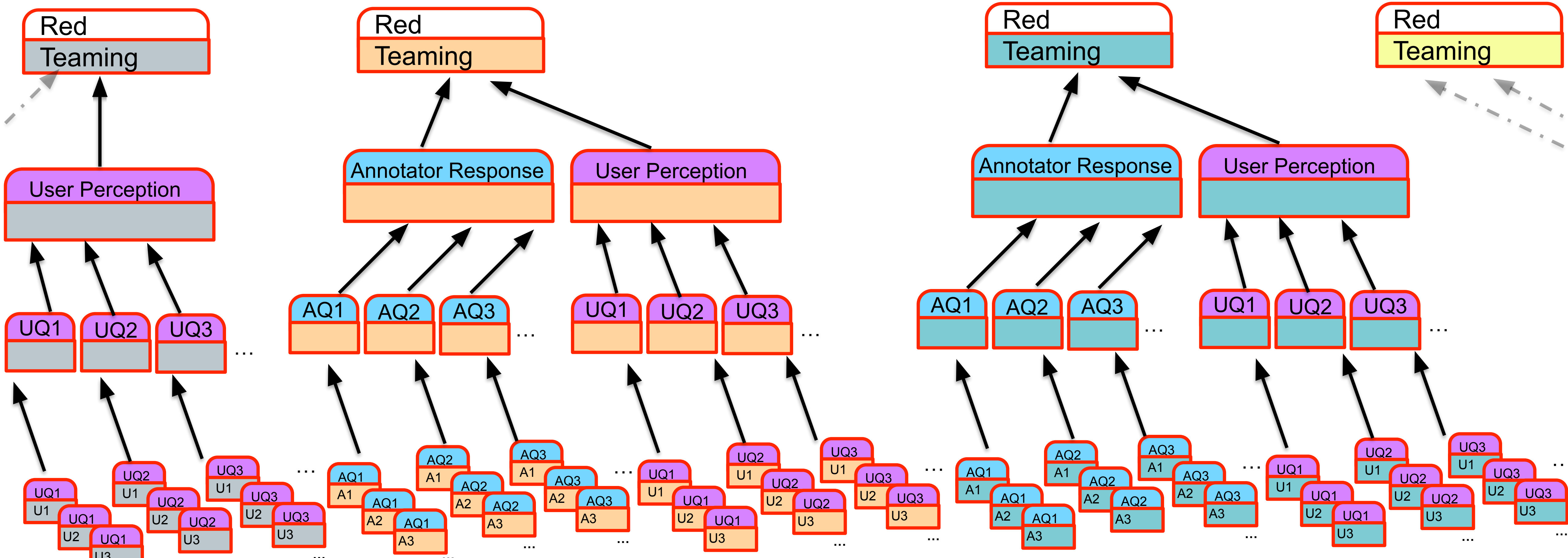
Fairness/Bias

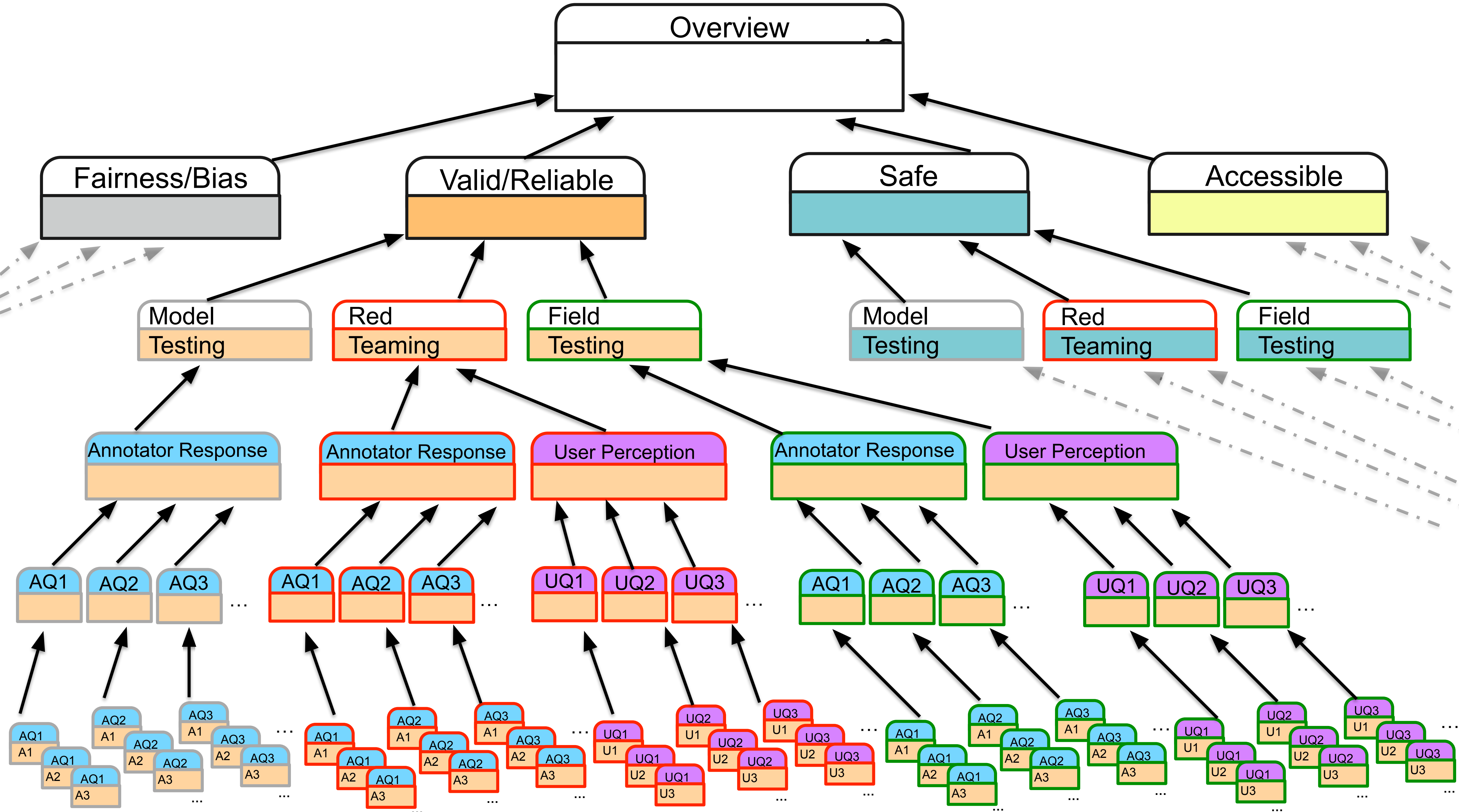
Valid/Reliable

Safe

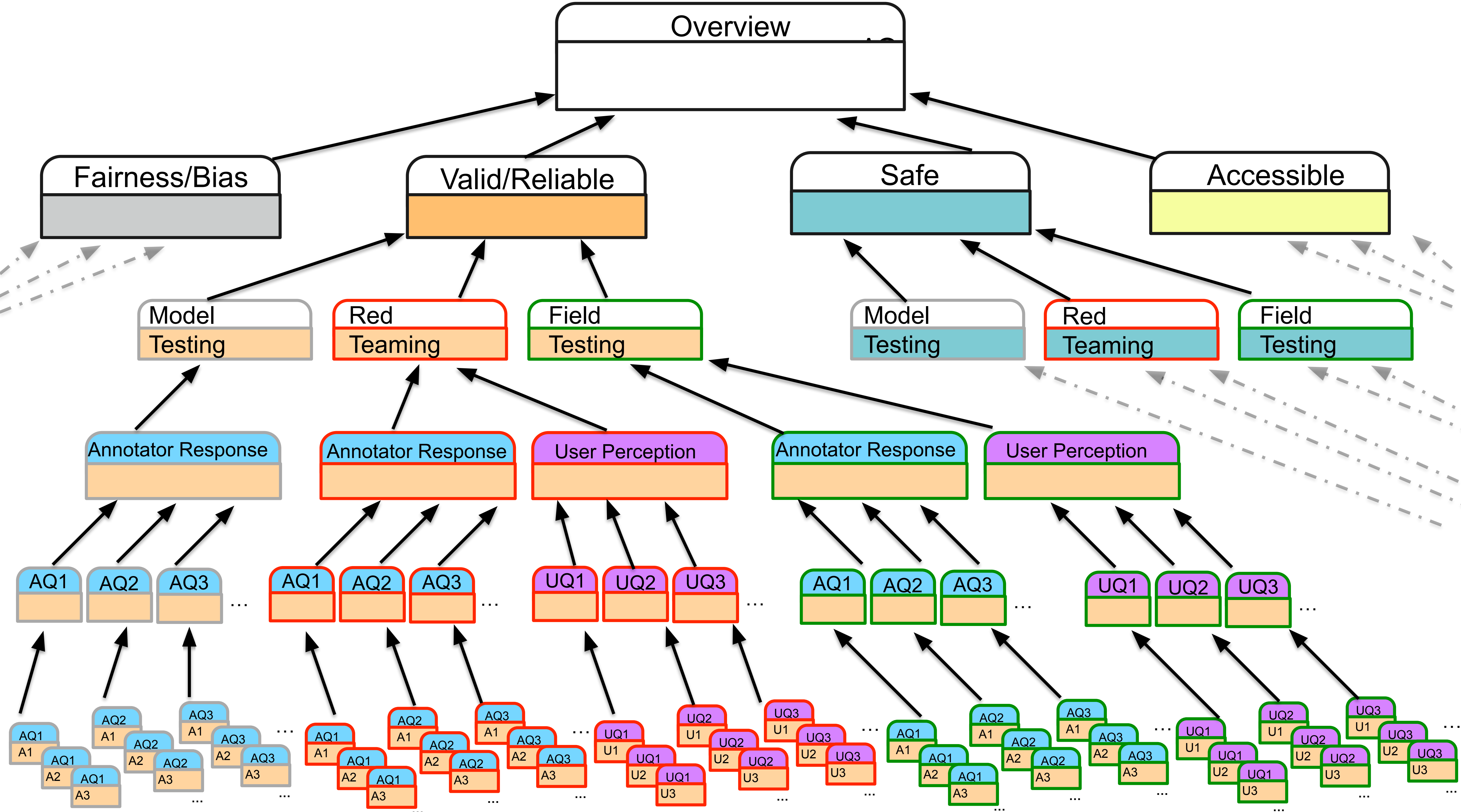
Accessible


Valid/Reliable
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
UQ1 U1
UQ1 U2
UQ1 U3
UQ2 U1
UQ2 U2
UQ2 U3
UQ3 U1
UQ3 U2
UQ3 U3


Red Teaming
User Perception
UQ1
UQ2
UQ3
…
UQ1 U1
UQ1 U2
UQ1 U3
UQ2 U1
UQ2 U2
UQ2 U3
UQ3 U1
UQ3 U2
UQ3 U3
Red Teaming
Annotator Response
User Perception
AQ1
AQ2
AQ3
UQ1
UQ2
UQ3
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
Red Teaming
Annotator Response
User Perception
Red Teaming



Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception
AQ1 AQ2 AQ3 ...
AQ1 AQ2 AQ3 ...
UQ1 UQ2 UQ3 ...
AQ1 AQ2 AQ3 ...
UQ1 UQ2 UQ3 ...
AQ1 A1 AQ1 A2 AQ1 A3
AQ2 A1 AQ2 A2 AQ2 A3 ...
AQ3 A1 AQ3 A2 AQ3 A3 ...
AQ1 A1 AQ1 A2 AQ1 A3 ...
AQ2 A1 AQ2 A2 AQ2 A3 ...
AQ3 A1 AQ3 A2 AQ3 A3 ...
UQ1 U1 UQ1 U2 UQ1 U3
UQ2 U1 UQ2 U2 UQ2 U3 ...
UQ3 U1 UQ3 U2 UQ3 U3 ...
AQ1 A1 AQ1 A2 AQ1 A3 ...
AQ2 A1 AQ2 A2 AQ2 A3 ...
AQ3 A1 AQ3 A2 AQ3 A3 ...
UQ1 U1 UQ1 U2 UQ1 U3 ...
UQ2 U1 UQ2 U2 UQ2 U3 ...
UQ3 U1 UQ3 U2 UQ3 U3 ...

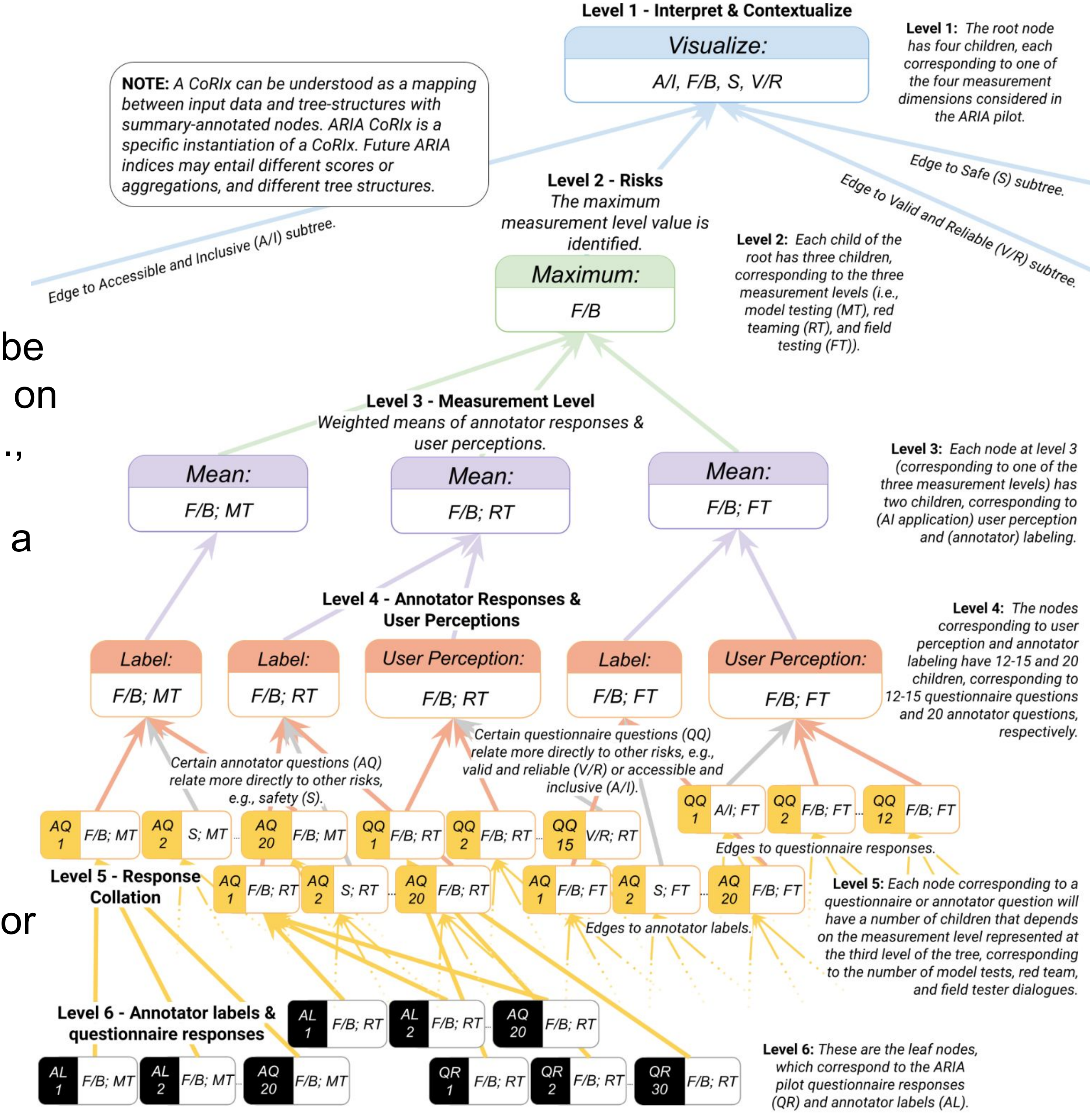

Overview
Fairness/Bias
Valid/Reliable
Safe
Accessible
Model Testing
Red Teaming
Field Testing
Model Testing
Red Teaming
Field Testing
Annotator Response
Annotator Response
User Perception
Annotator Response
User Perception
AQ1
AQ2
AQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...
AQ1
AQ2
AQ3
...
UQ1
UQ2
UQ3
...
AQ1 A1
AQ1 A2
AQ1 A3
AQ2 A1
AQ2 A2
AQ2 A3
AQ3 A1
AQ3 A2
AQ3 A3
UQ1 U1
UQ1 U2
UQ1 U3
UQ2 U1
UQ2 U2
UQ2 U3
UQ3 U1
UQ3 U2
UQ3 U3
...

Summary methods can be many things, depending on their children; could, e.g., be weighted average, a textual summary, a plot, a combination thereof.

Similarly, many tree topologies are possible.

Score is whole tree

Can consider only root, or any tree depth, subtree, branch, …



NOTE: *A CoRIx can be understood as a mapping between input data and tree-structures with summary-annotated nodes. ARIA CoRIx is a specific instantiation of a CoRIx. Future ARIA indices may entail different scores or aggregations, and different tree structures.*

Level 1 - Interpret & Contextualize

Level 1: *The root node has four children, each corresponding to one of the four measurement dimensions considered in the ARIA pilot.*

Level 2 - Risks
The maximum measurement level value is identified.

Level 2: *Each child of the root has three children, corresponding to the three measurement levels (i.e., model testing (MT), red teaming (RT), and field testing (FT)).*

Level 3 - Measurement Level
Weighted means of annotator responses & user perceptions.

Level 3: *Each node at level 3 (corresponding to one of the three measurement levels) has two children, corresponding to (AI application) user perception and (annotator) labeling.*

Level 4 - Annotator Responses & User Perceptions

Level 4: *The nodes corresponding to user perception and annotator labeling have 12-15 and 20 children, corresponding to 12-15 questionnaire questions and 20 annotator questions, respectively.*

Certain annotator questions (AQ) relate more directly to other risks, e.g., safety (S).

Certain questionnaire questions (QQ) relate more directly to other risks, e.g., valid and reliable (V/R) or accessible and inclusive (A/I).

Level 5 - Response Collation

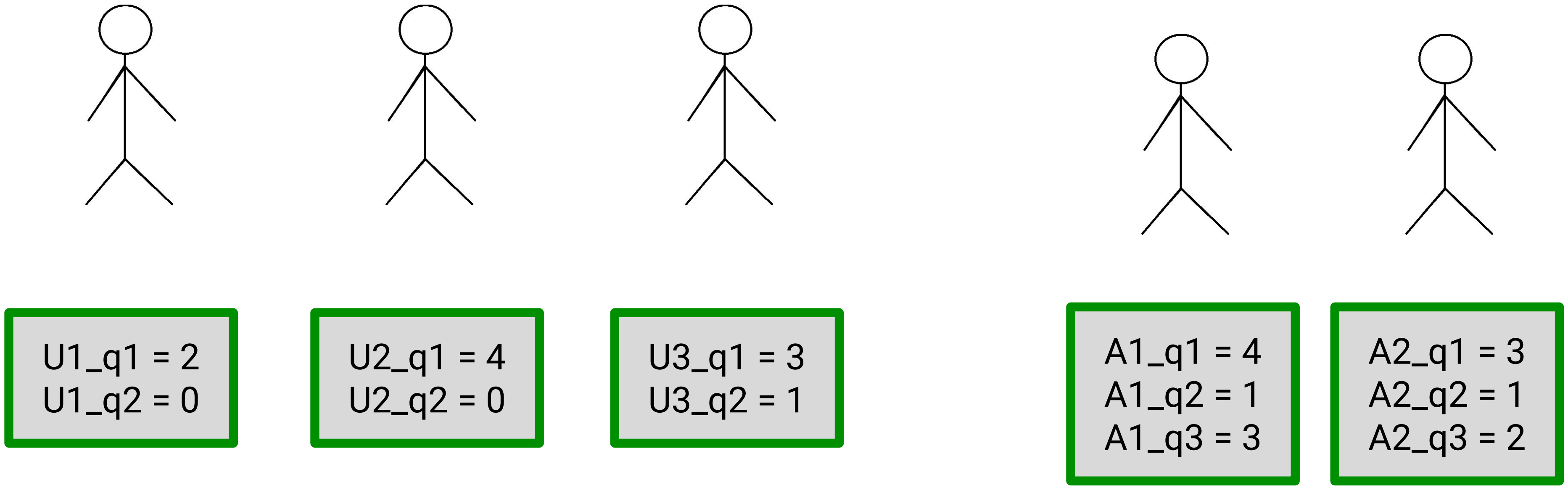
Level 5: *Each node corresponding to a questionnaire or annotator question will have a number of children that depends on the measurement level represented at the third level of the tree, corresponding to the number of model tests, red team, and field tester dialogues.*

Level 6 - Annotator labels & questionnaire responses

Level 6: *These are the leaf nodes, which correspond to the ARIA pilot questionnaire responses (QR) and annotator labels (AL).*

Users

Annotators



U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

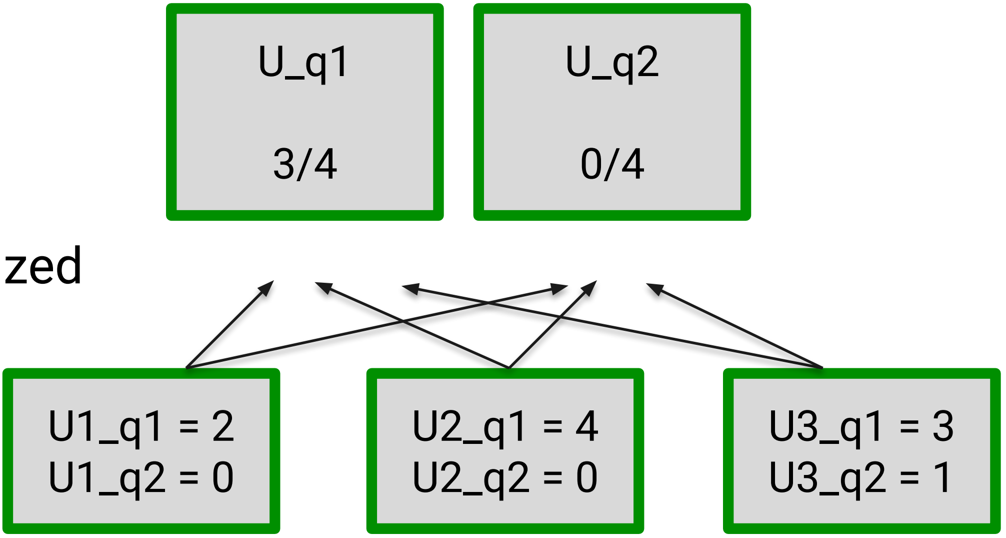
U3_q1 = 3
U3_q2 = 1

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Response Collation

Scale Normalized Median



U_q1

3/4

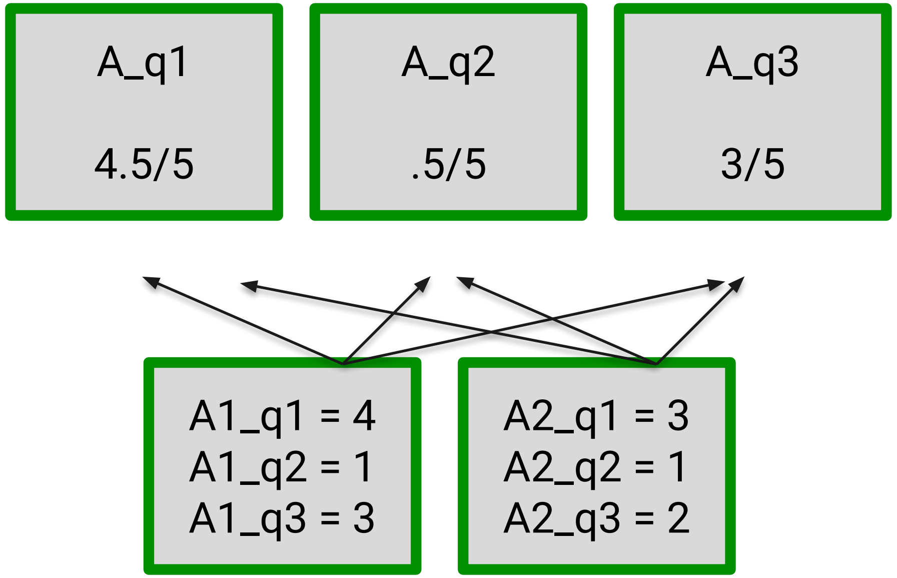
U_q2

0/4

U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1



A_q1

4.5/5

A_q2

.5/5

A_q3

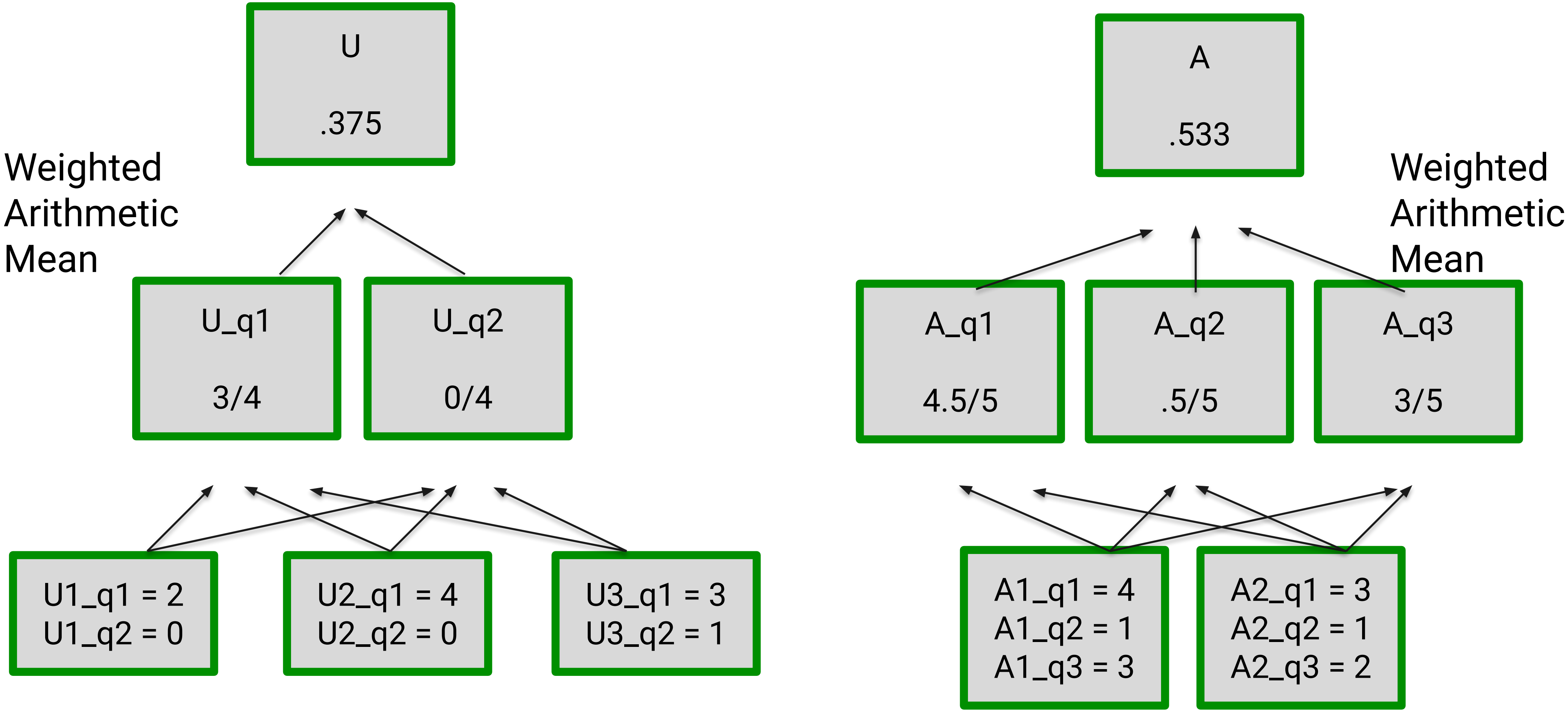
3/5

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

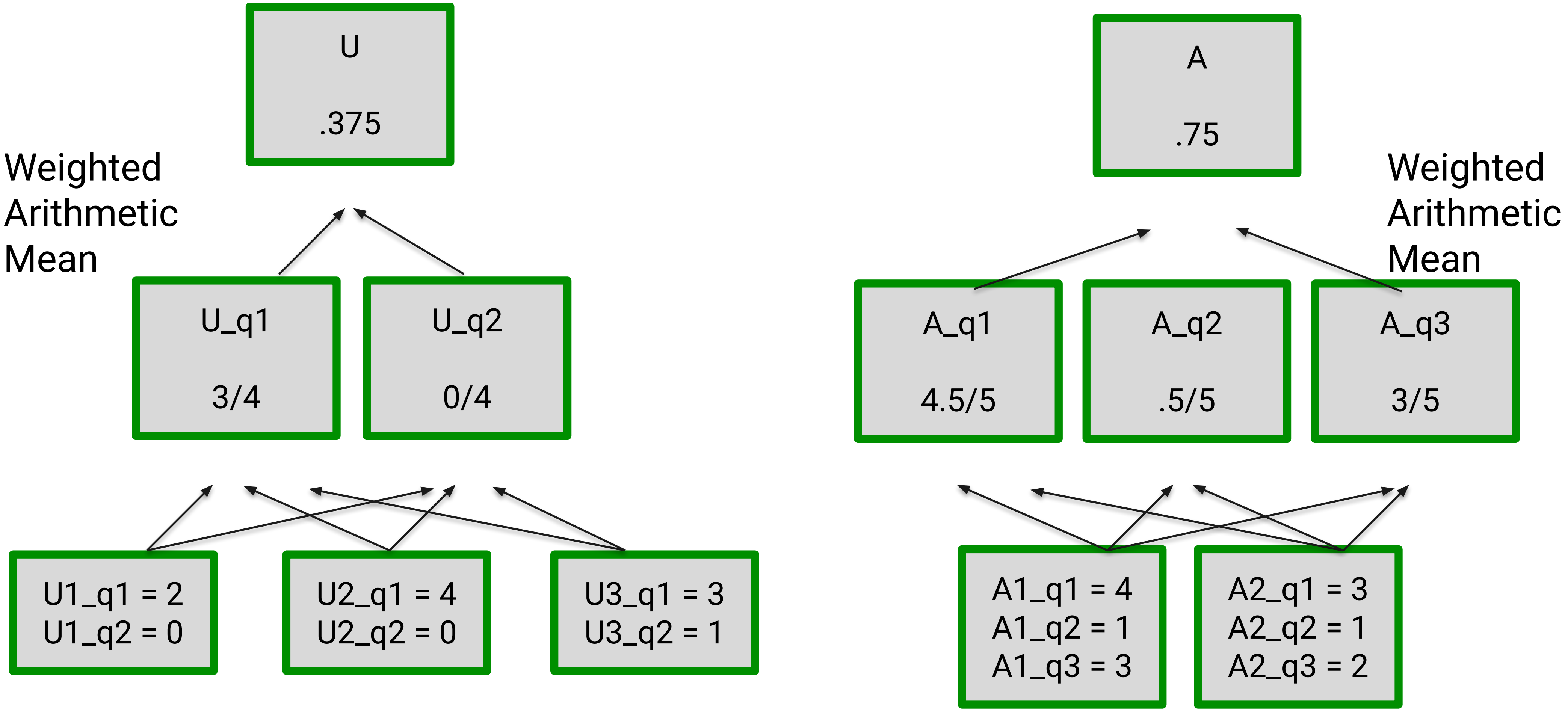
A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Scale Normalized Weighted Arithmetic Mean

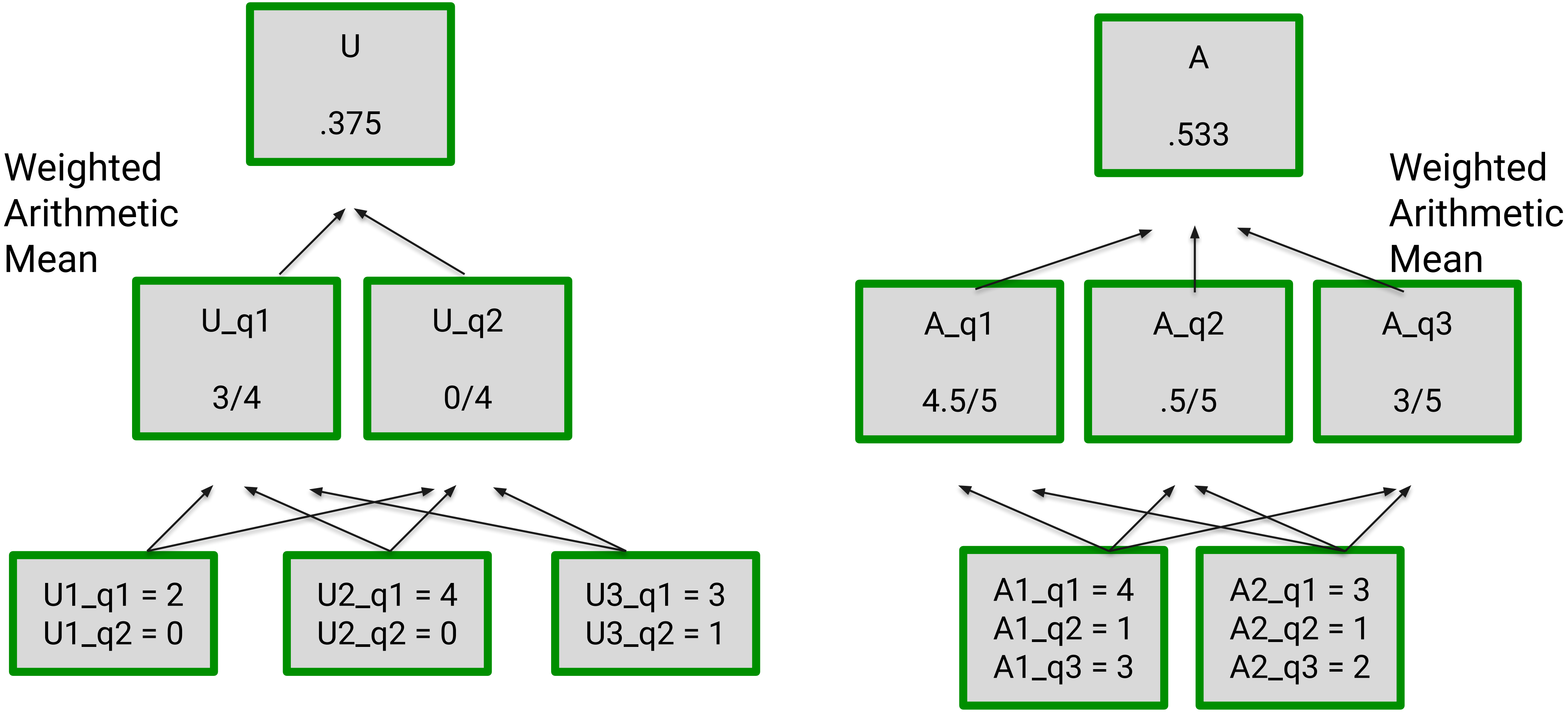
User Perception and Annotator Responses



User Perception and Annotator Responses



User Perception and Annotator Responses



U

.375

Weighted Arithmetic Mean

U_q1

3/4

U_q2

0/4

U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1

A

.533

Weighted Arithmetic Mean

A_q1

4.5/5

A_q2

.5/5

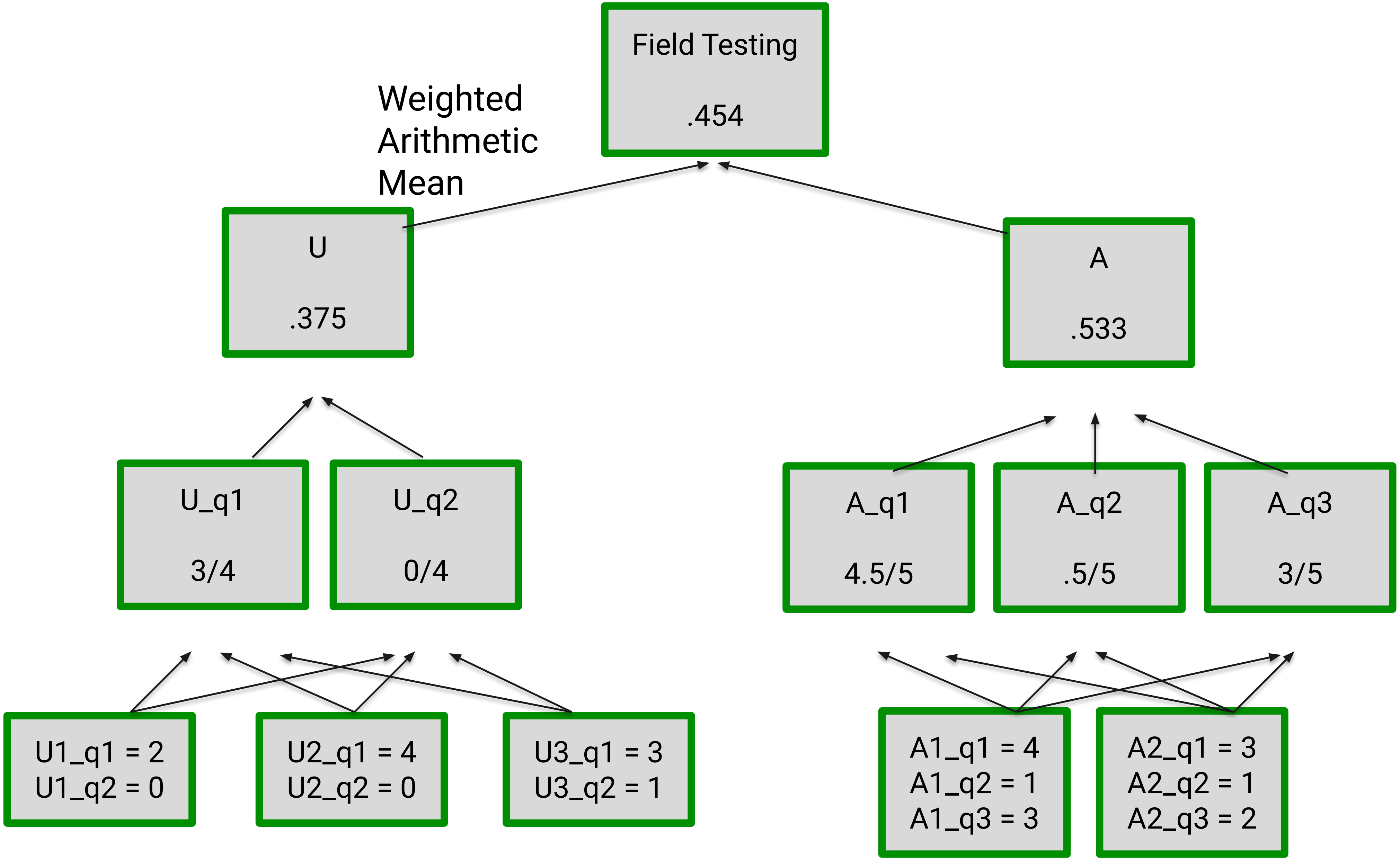
A_q3

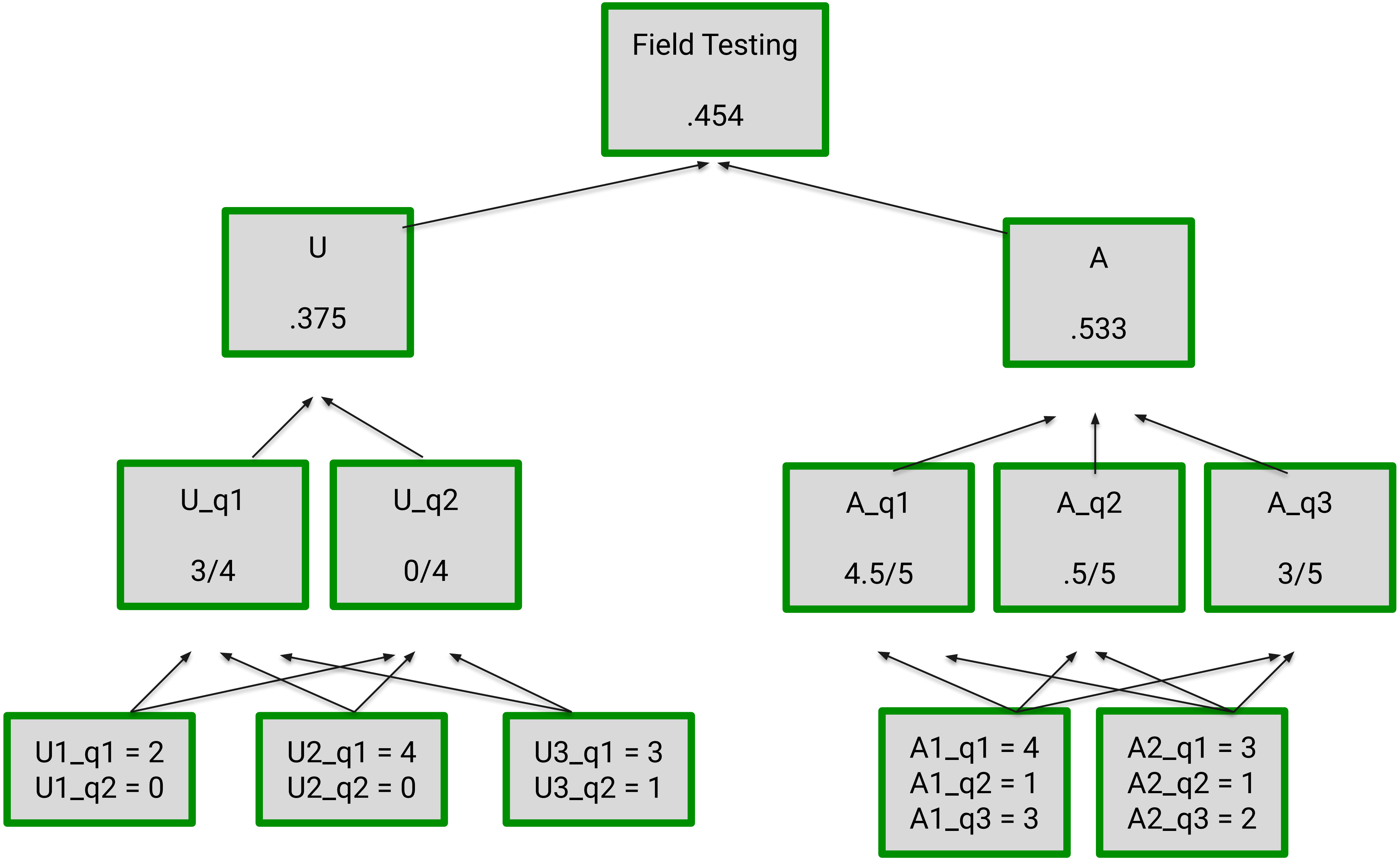
3/5

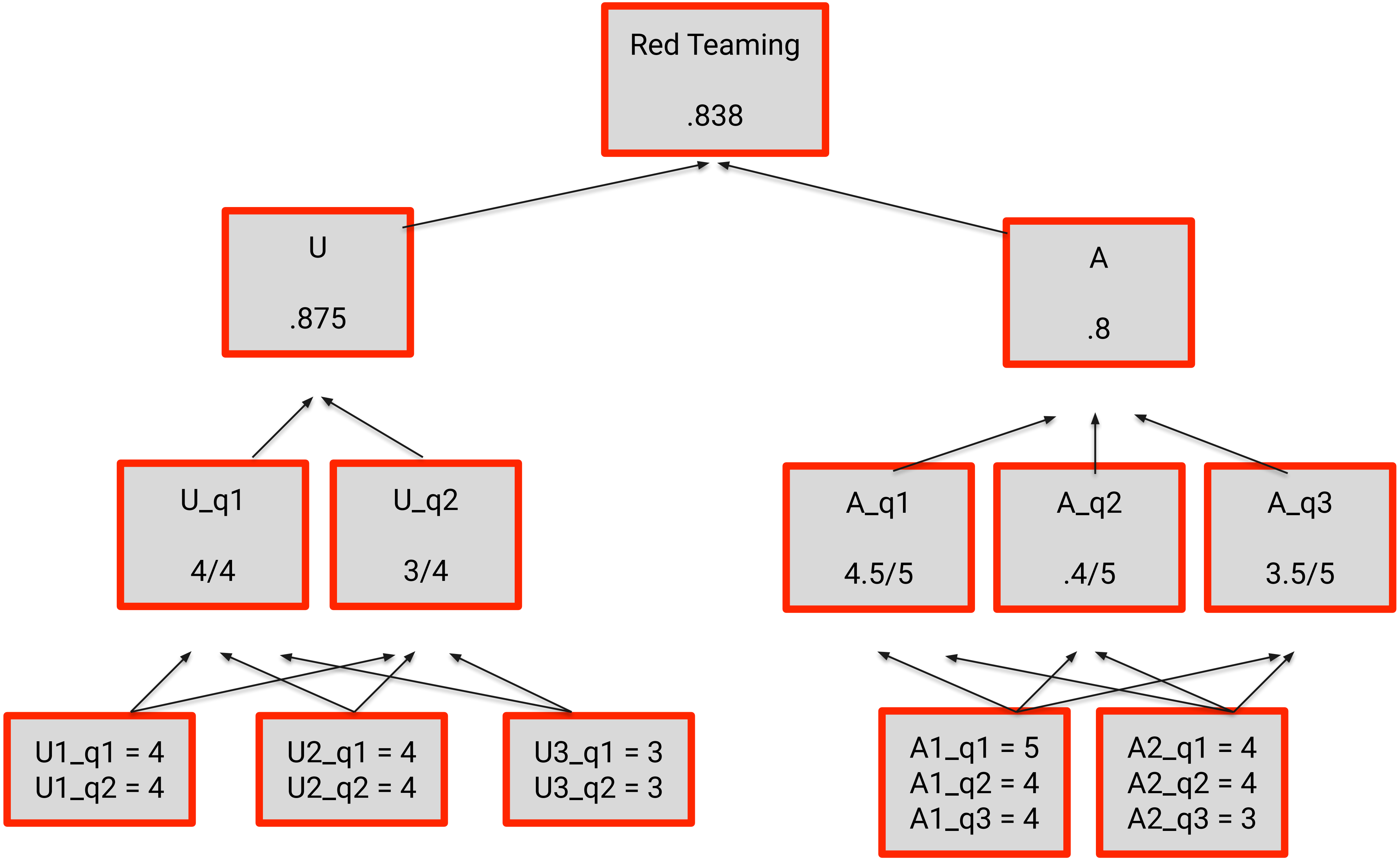
A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Measurement Level




Field Testing
.454
U
.375
A
.533
U_q1
3/4
U_q2
0/4
A_q1
4.5/5
A_q2
.5/5
A_q3
3/5
U1_q1 = 2
U1_q2 = 0
U2_q1 = 4
U2_q2 = 0
U3_q1 = 3
U3_q2 = 1
A1_q1 = 4
A1_q2 = 1
A1_q3 = 3
A2_q1 = 3
A2_q2 = 1
A2_q3 = 2



Red Teaming

.838

U

.875

A

.8

U_q1

4/4

U_q2

3/4

A_q1

4.5/5

A_q2

.4/5

A_q3

3.5/5

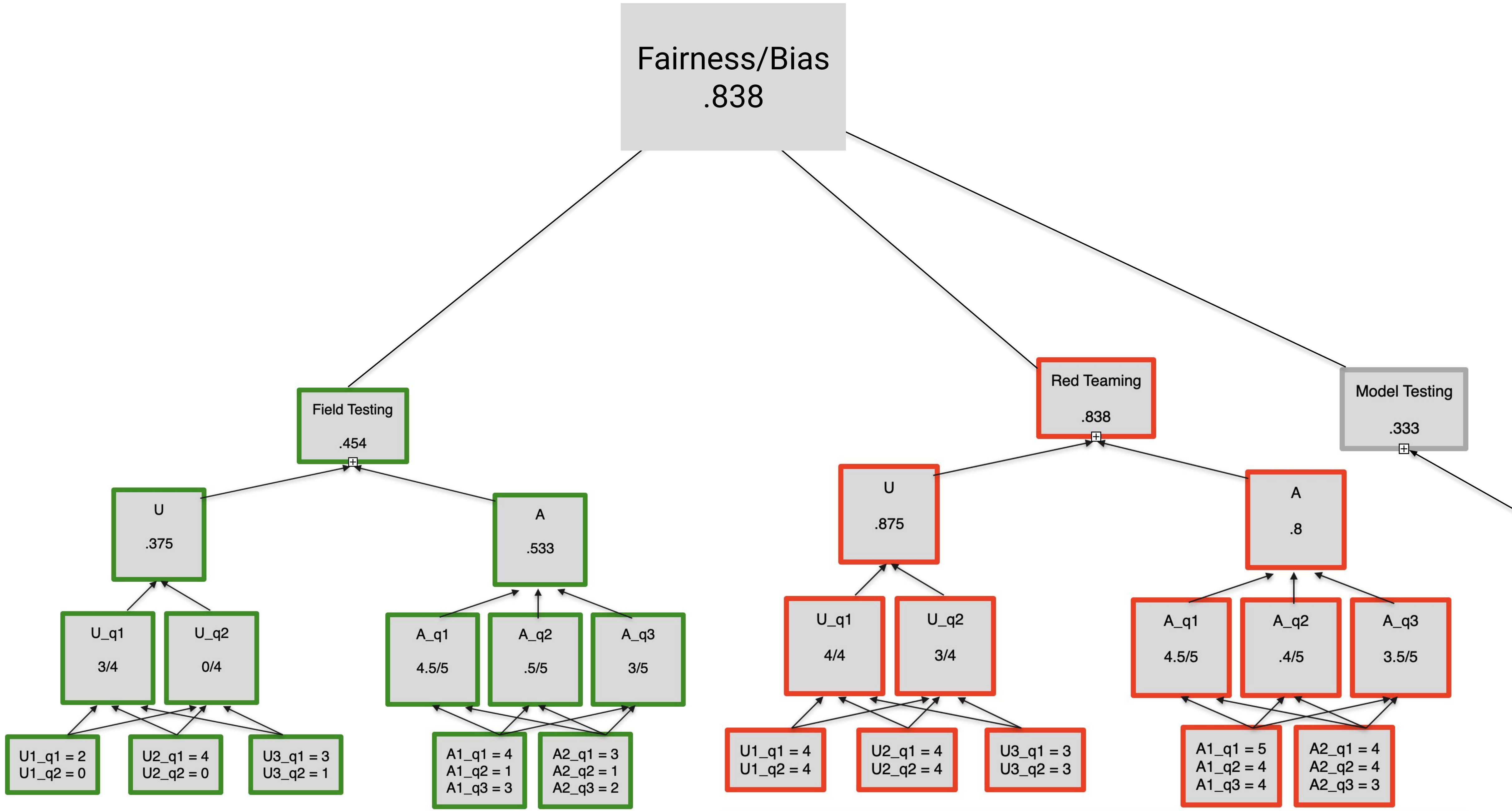
U1_q1 = 4
U1_q2 = 4

U2_q1 = 4
U2_q2 = 4

U3_q1 = 3
U3_q2 = 3

A1_q1 = 5
A1_q2 = 4
A1_q3 = 4

A2_q1 = 4
A2_q2 = 4
A2_q3 = 3



Fairness/Bias
.838

Field Testing
.454

U
.375

U_q1
3/4

U_q2
0/4

U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1

A
.533

A_q1
4.5/5

A_q2
.5/5

A_q3
3/5

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Red Teaming
.838

U
.875

U_q1
4/4

U_q2
3/4

U1_q1 = 4
U1_q2 = 4

U2_q1 = 4
U2_q2 = 4

U3_q1 = 3
U3_q2 = 3

A
.8

A_q1
4.5/5

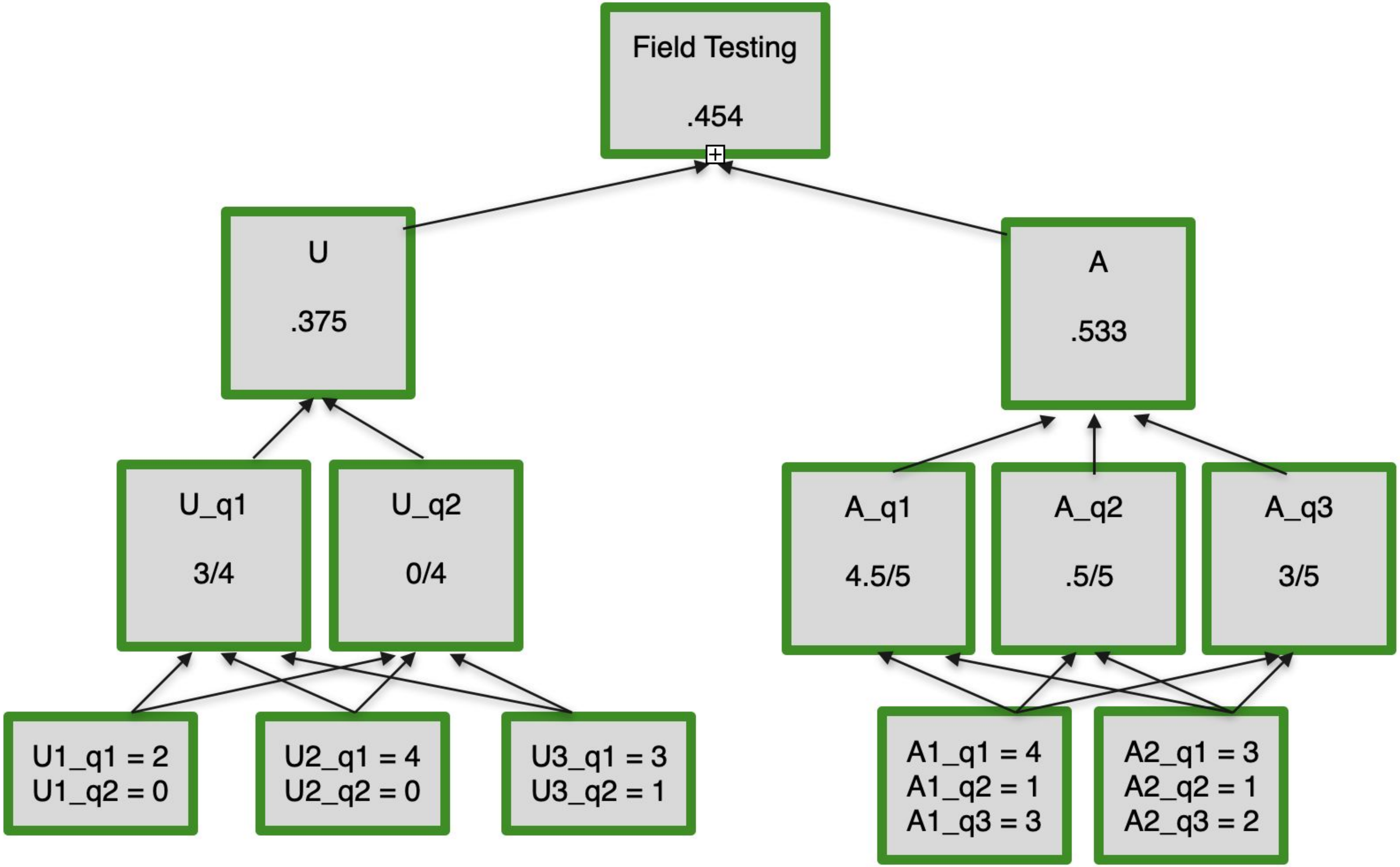
A_q2
.4/5

A_q3
3.5/5

A1_q1 = 5
A1_q2 = 4
A1_q3 = 4

A2_q1 = 4
A2_q2 = 4
A2_q3 = 3

Model Testing
.333



Field Testing

.454

U

.375

A

.533

U_q1

3/4

U_q2

0/4

A_q1

4.5/5

A_q2

.5/5

A_q3

3/5

U1_q1 = 2
U1_q2 = 0

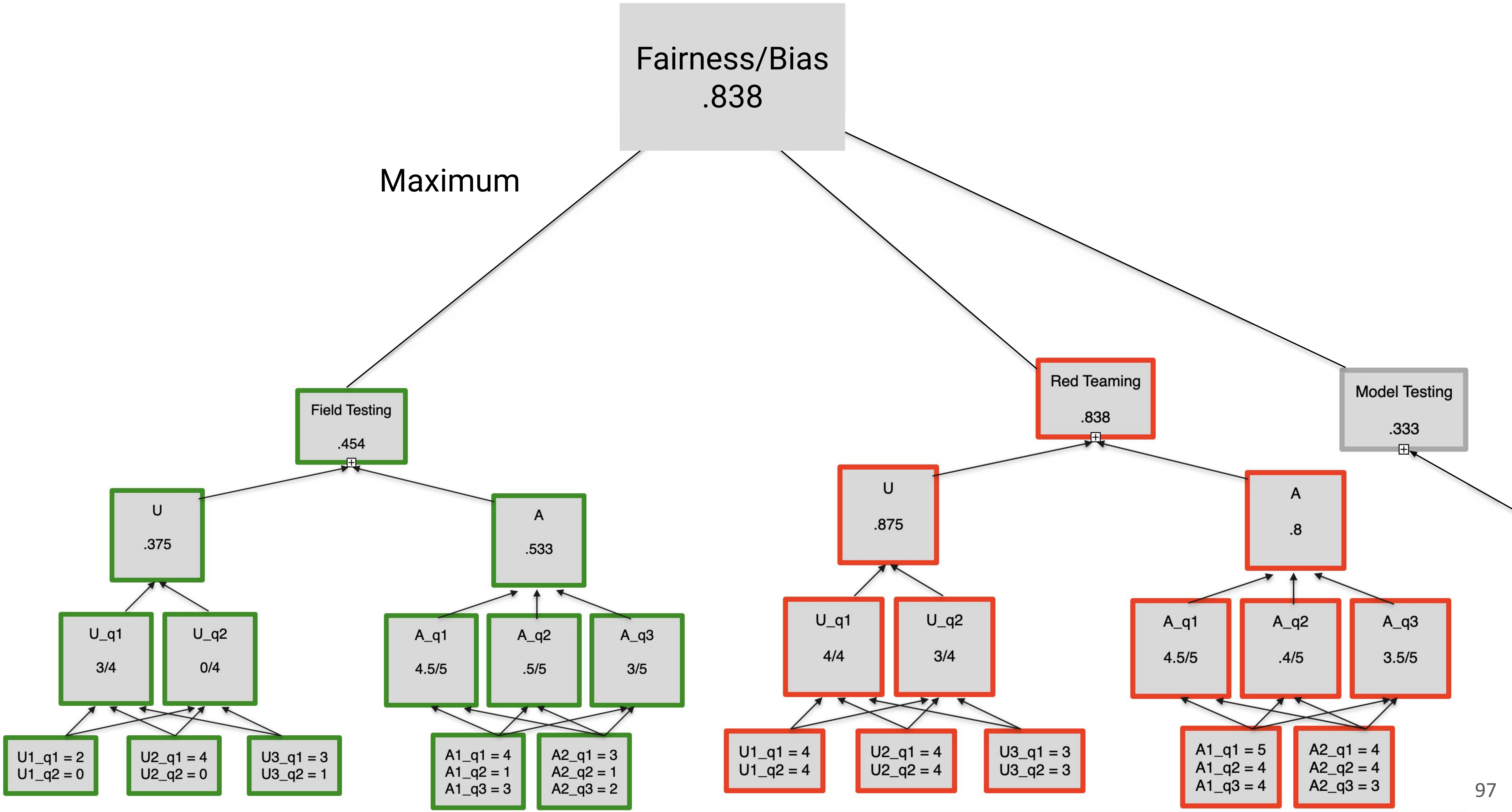
U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Risk Dimension



Fairness/Bias
.838

Maximum

Field Testing
.454

U
.375

U_q1
3/4

U_q2
0/4

U1_q1 = 2
U1_q2 = 0

U2_q1 = 4
U2_q2 = 0

U3_q1 = 3
U3_q2 = 1

A
.533

A_q1
4.5/5

A_q2
.5/5

A_q3
3/5

A1_q1 = 4
A1_q2 = 1
A1_q3 = 3

A2_q1 = 3
A2_q2 = 1
A2_q3 = 2

Red Teaming
.838

U
.875

U_q1
4/4

U_q2
3/4

U1_q1 = 4
U1_q2 = 4

U2_q1 = 4
U2_q2 = 4

U3_q1 = 3
U3_q2 = 3

A
.8

A_q1
4.5/5

A_q2
.4/5

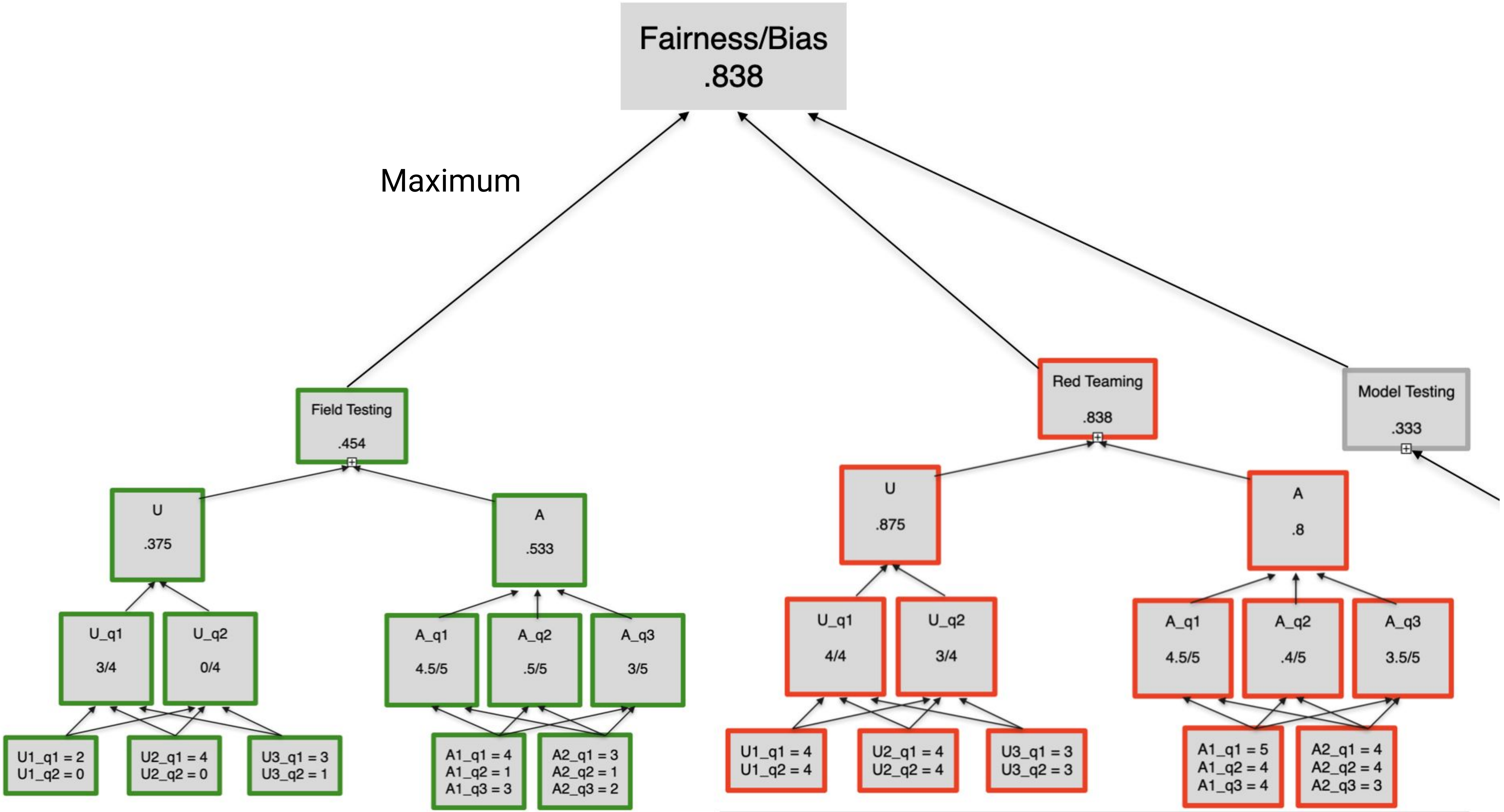
A_q3
3.5/5

A1_q1 = 5
A1_q2 = 4
A1_q3 = 4

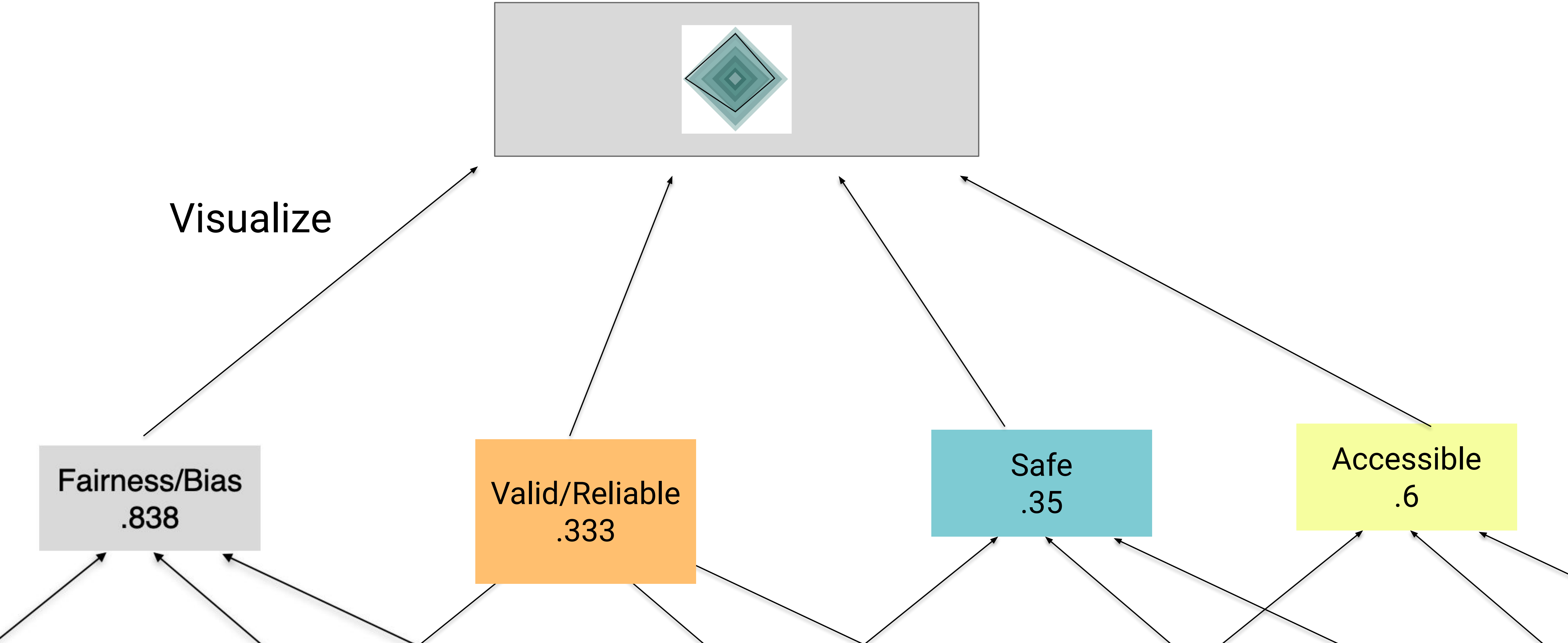
A2_q1 = 4
A2_q2 = 4
A2_q3 = 3

Model Testing
.333

Risk Dimension



Overview



Visualize

Fairness/Bias
.838

Valid/Reliable
.333

Safe
.35

Accessible
.6

Questions???