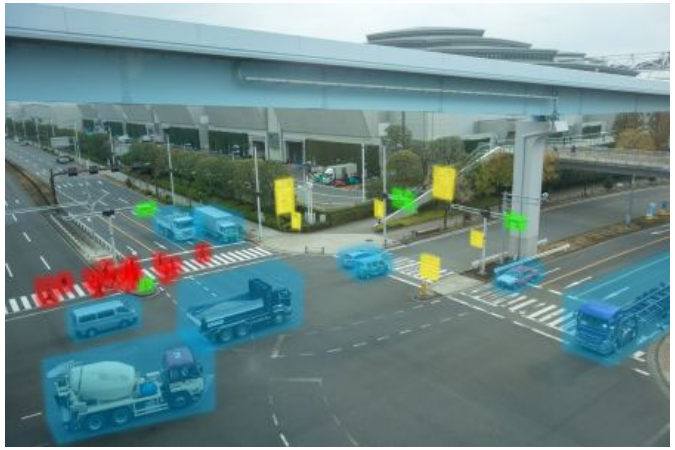
Concepts in Quality Assessment for Machine Learning - From Test Data to Arguments

Fuyuki Ishikawa
National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Background: (Engineering) Machine Learning

- Active effort to apply machine learning (ML)
  - Intensive support on libraries and platforms



- Now engineering support is essential
  - Defining and finalizing products with customers
  - Arguing quality of products
  - …

➧ Conceptual models !

Play the essential role of capturing the essence of the product and its quality

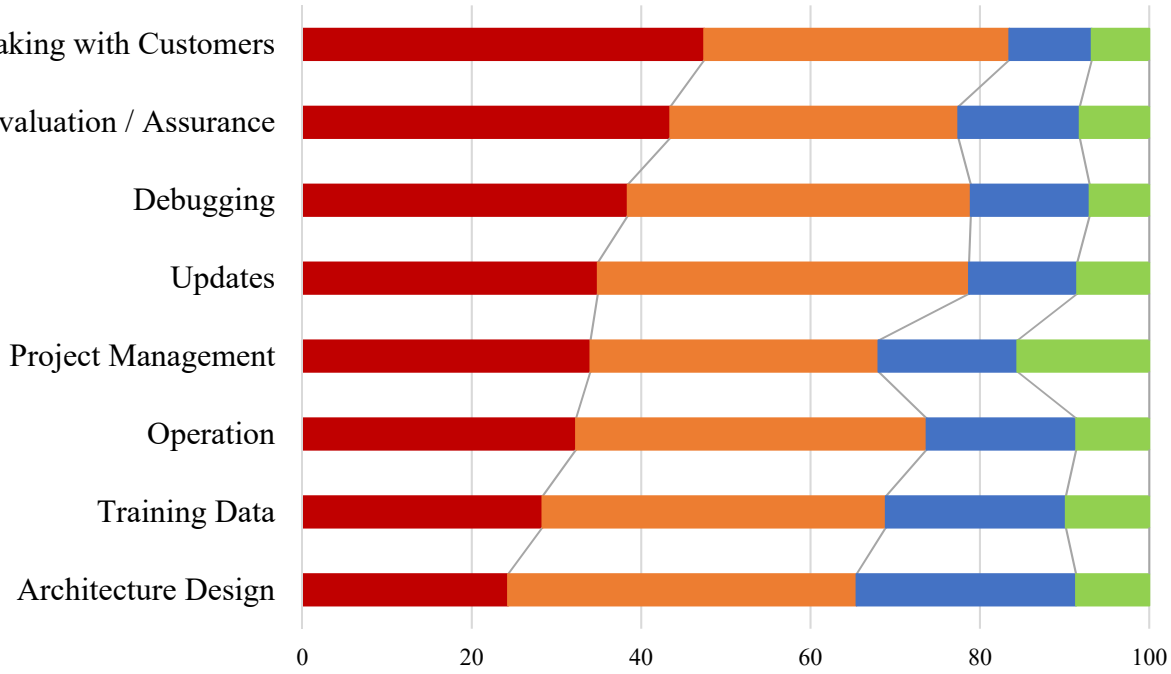
Motivation: Essential Difference in ML

- With ML, we obtain the behavior of a component (e.g., a neural net) **inductively** from training data

  Black-box, imperfect, non-testable (no oracle), unexplainable, has adversarial examples, …

Existing principles do not work

When arguing the product and its quality



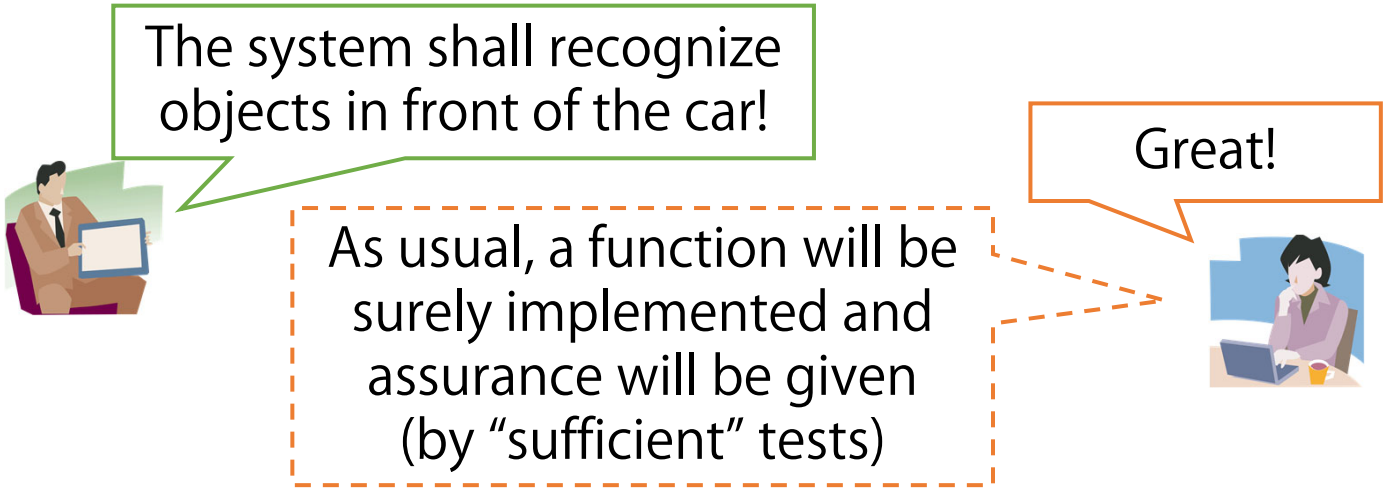
How difficult??

Questionnaire survey by SIG-MLSE, Japan, 2018

Motivation: Necessary Concepts?

- Argue the product and its quality!

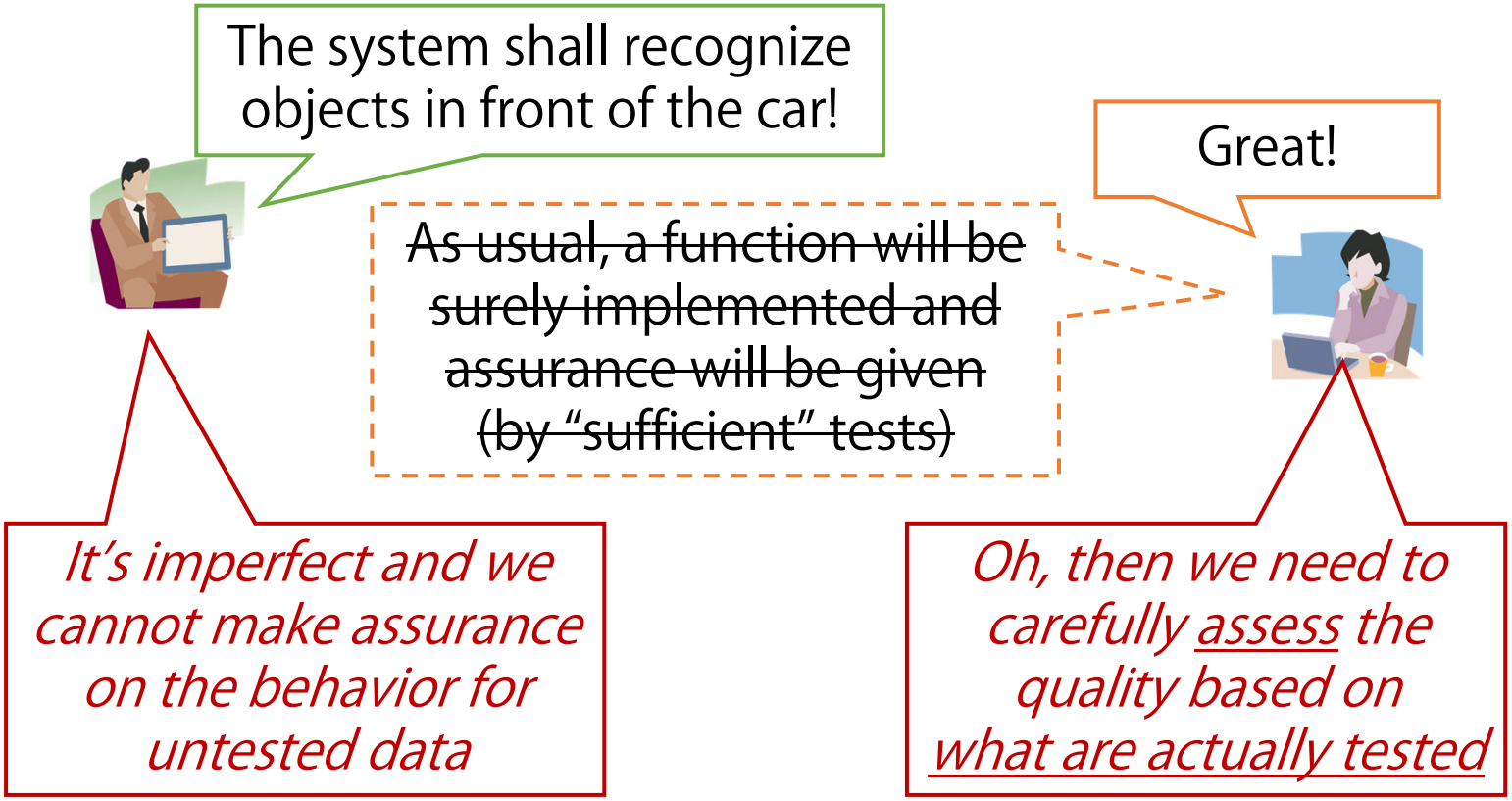
  (engineer-engineer or engineer-customer)



Motivation: Necessary Concepts?

- Argue the product and its quality!

  (engineer-engineer or engineer-customer)



Proposal: MLQ Framework

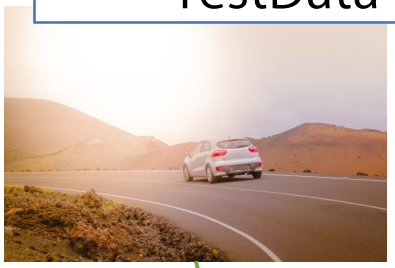
- Framework for assessing the quality of ML components and ML-based systems
  - Focuses on concepts to capture test data, or empirical evidence in more general
  - Reflects the state-of-the-art research on testing ML
  - Uses an argument model (e.g., assurance case in Goal-Structuring Notation) to describe the whole picture
  - To be linked to test-data management tools

Example: Test Data Attribution

- To specify & check constraints
- To describe current status
- To discuss validity
- To compare with operational data

We tested with 100,000 data!

What data … ? Did you test misty days?



carCameraImg1 : TestData

atr1 : Attribute

atr2 : Attribute

situation : AttributeType

sunrise : AttributeValue

weatherCondition : AttributeType

cloudy : AttributeValue

weatherReport : AttributionMethod

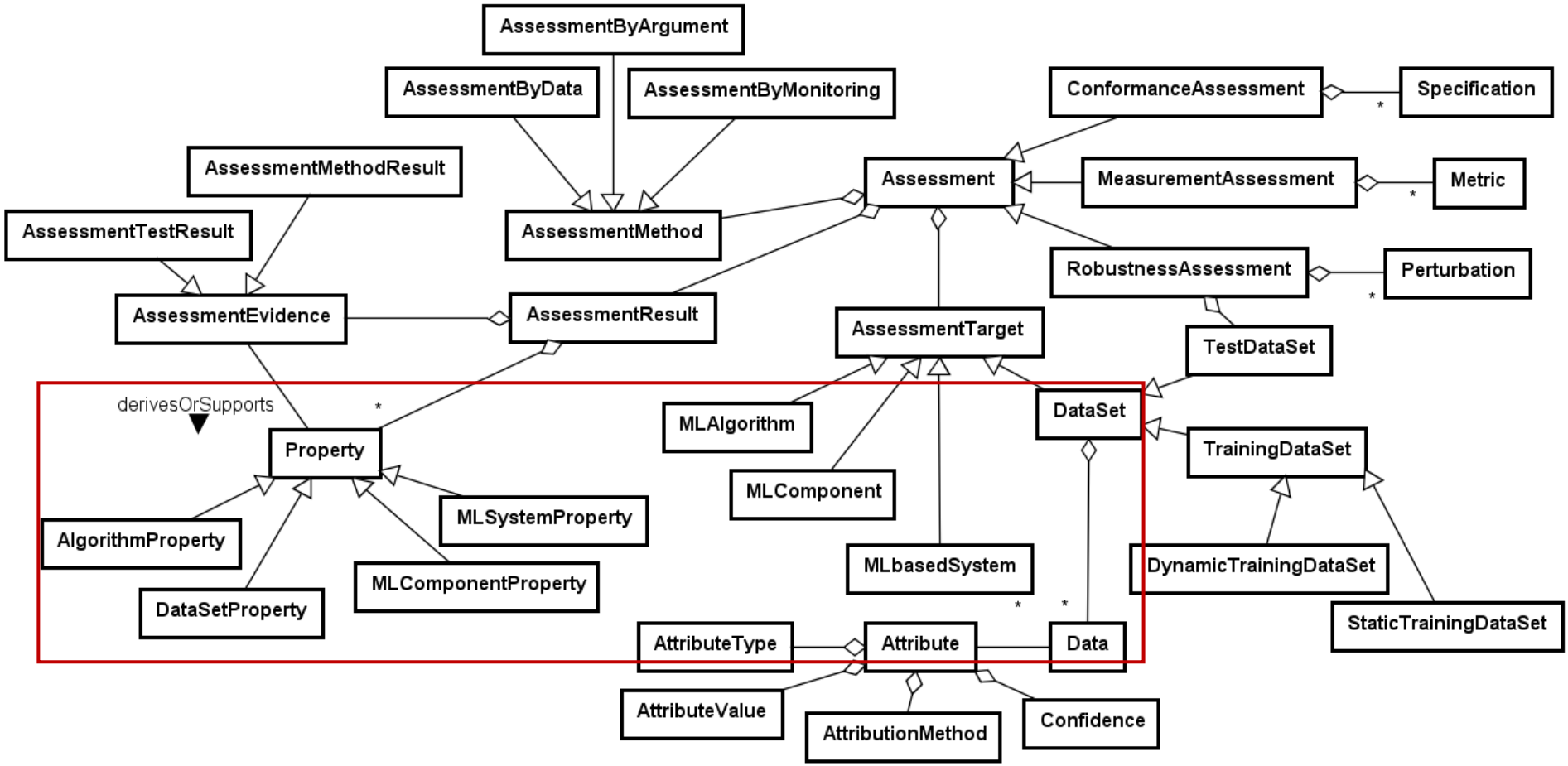
middle : Confidence

imgSet1 : TestDataSet

#misty = 2000 : DataSetProperty

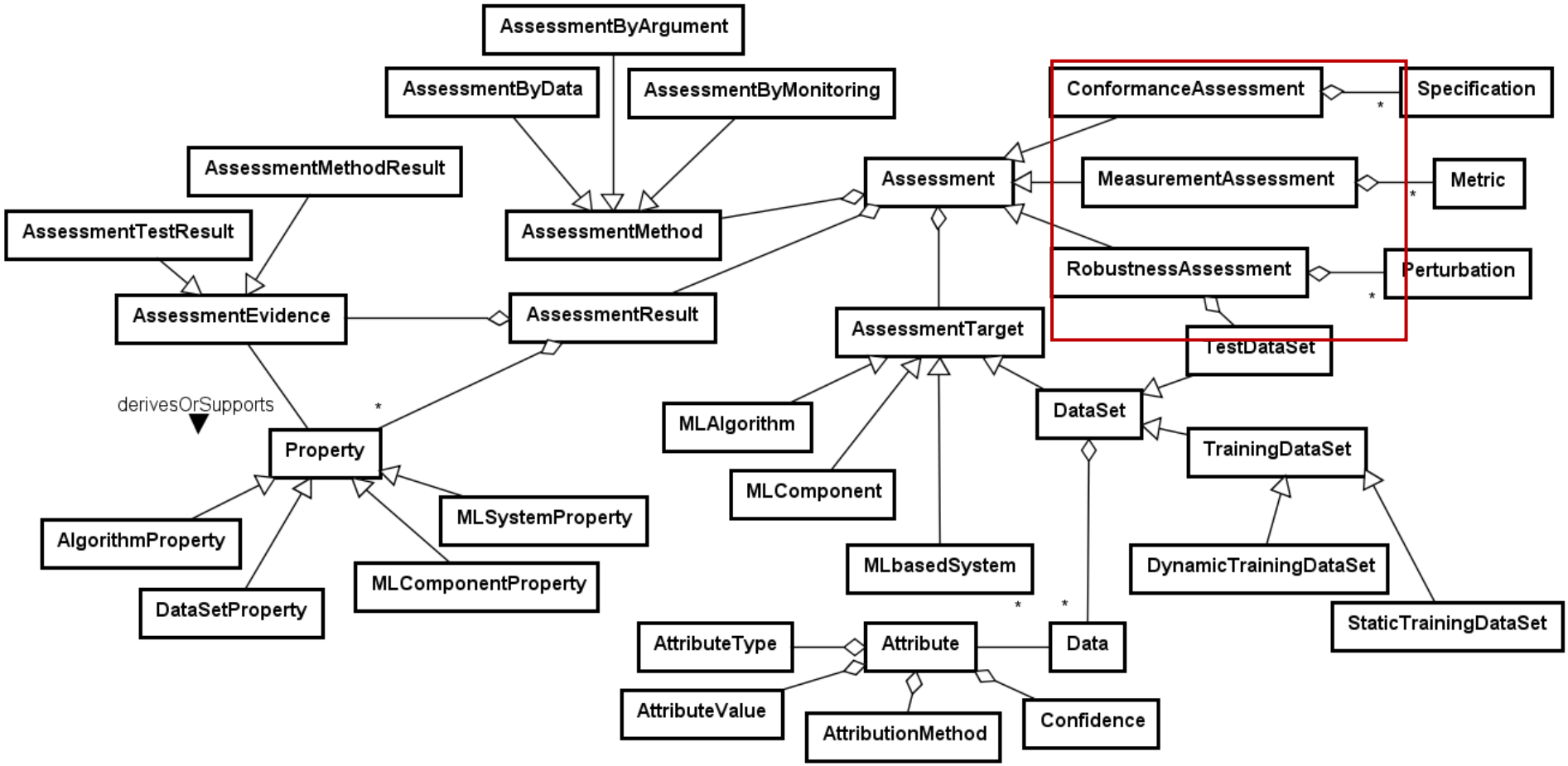
Actual representation may be spreadsheet, GUI, or whatever

Concepts: Whole Picture



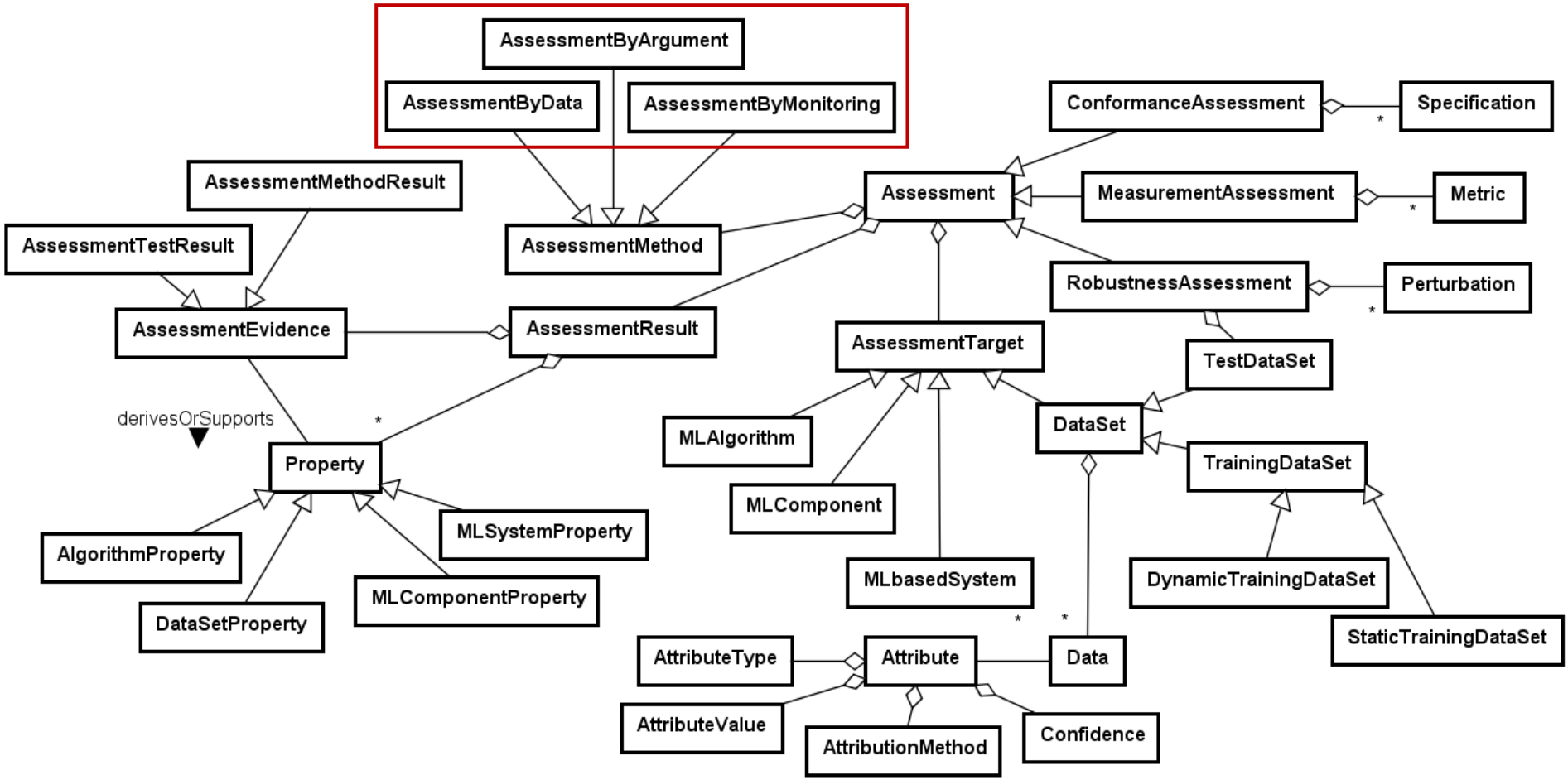
Different targets of assessment in the ML dev. process

Concepts: Whole Picture



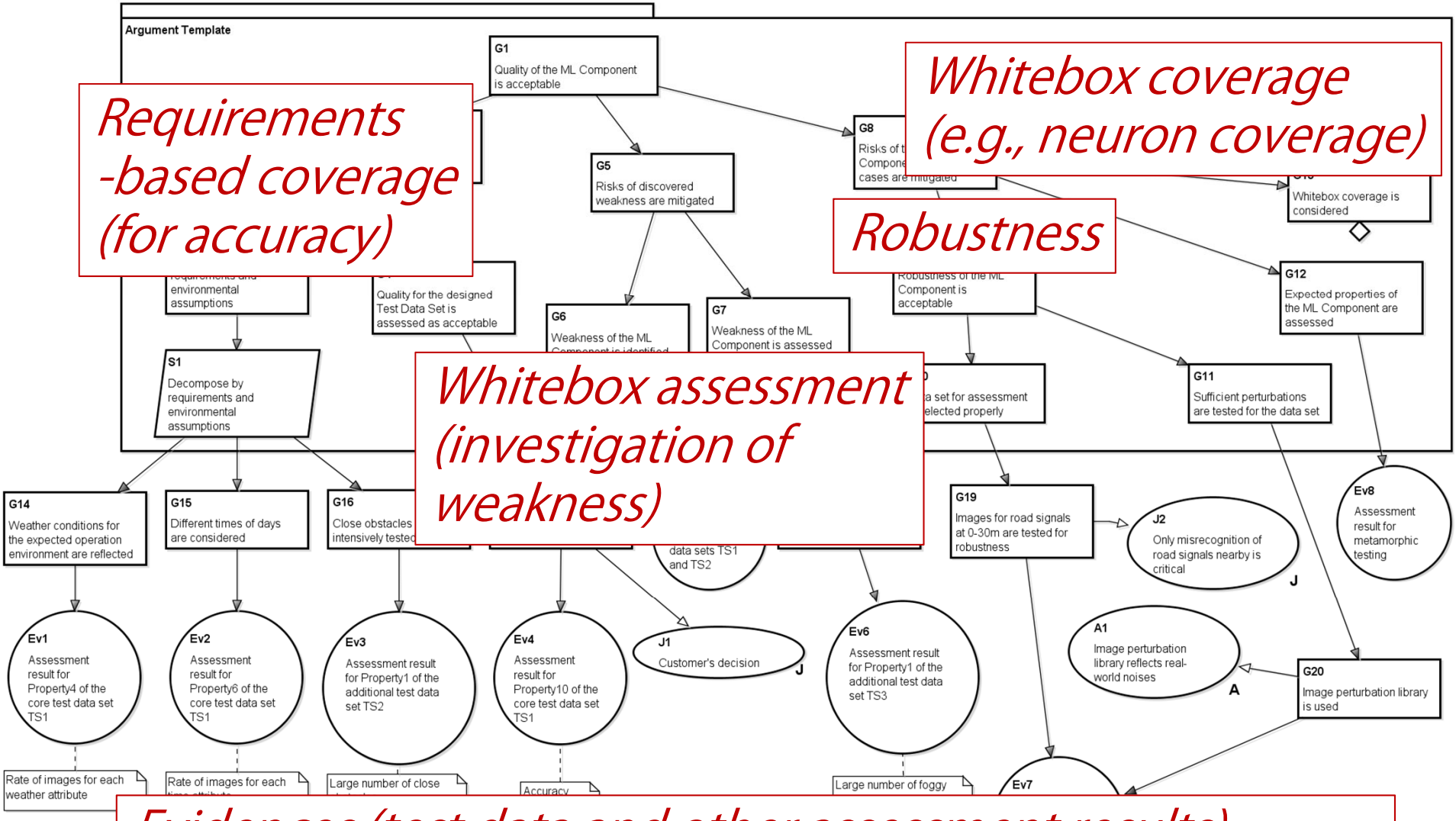
Different ways of assessment that covers the present techniques (accuracy, metamorphic relations, robustness, etc.)

Concepts: Whole Picture



Deductive/logical and inductive/empirical assessment (also covering runtime to tackle the uncertainty)

Argument Example: Whole Picture



Requirements -based coverage (for accuracy)

Whitebox coverage (e.g., neuron coverage)

Robustness

Whitebox assessment (investigation of weakness)

Evidences (test data and other assessment results) are managed with concepts annotated

Summary: MLQ Framework

- Framework for assessing the quality of ML components and ML-based systems

  From test data to arguments

- Ongoing/Future Work
  - Elaboration with Case Studies
  - Tools, especially connection with test data management
  - Uncertainty-aware argument modeling [ASSURE'18] (awareness of risks and continuous engineering)

Thank You!