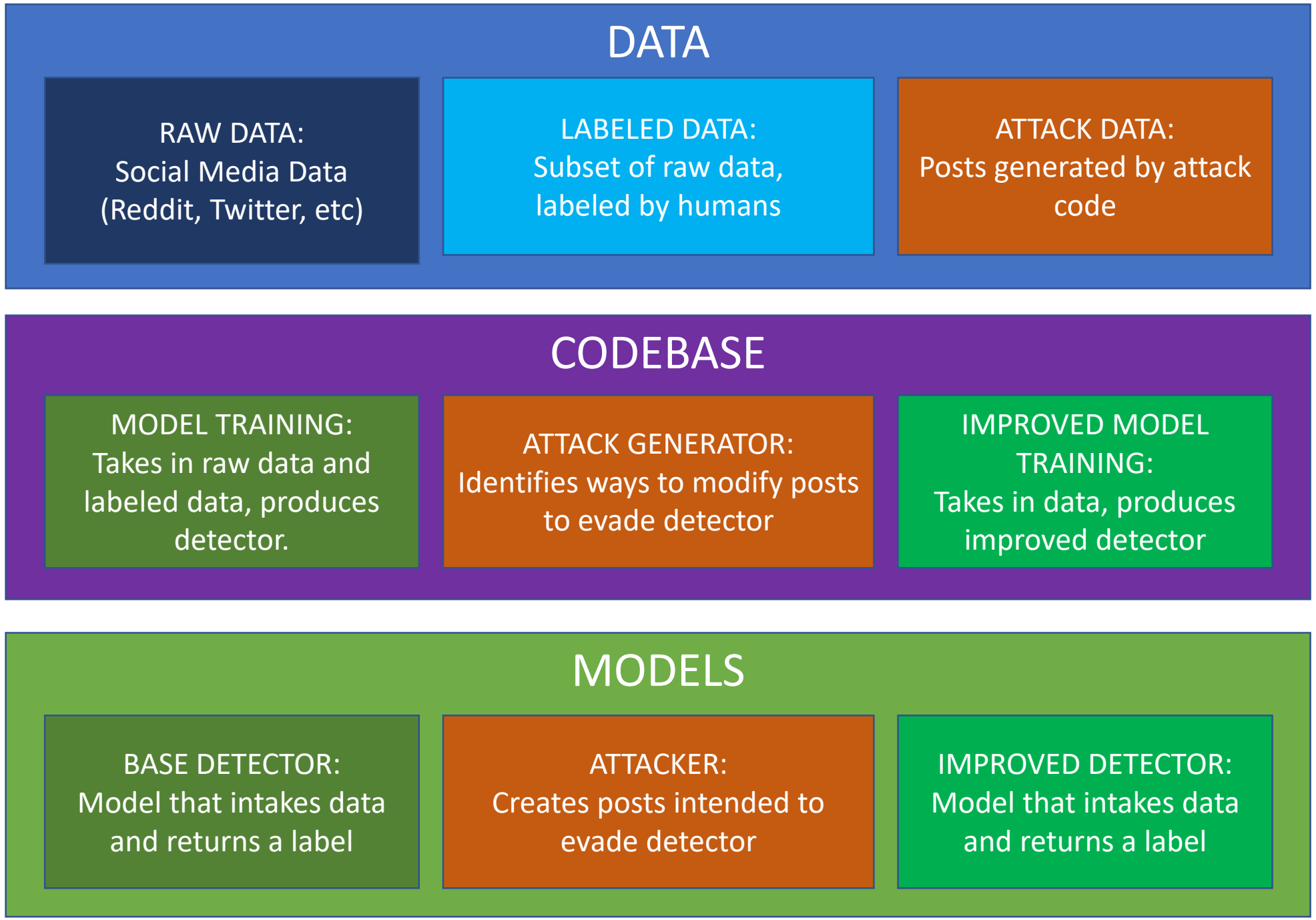
DATA
RAW DATA:
Social Media Data
(Reddit, Twitter, etc)
LABELED DATA:
Subset of raw data,
labeled by humans
ATTACK DATA:
Posts generated by attack
code
CODEBASE
MODEL TRAINING:
Takes in raw data and
labeled data, produces
detector.
ATTACK GENERATOR:
Identifies ways to modify posts
to evade detector
IMPROVED MODEL
TRAINING:
Takes in data, produces
improved detector
MODELS
BASE DETECTOR:
Model that intakes data
and returns a label
ATTACKER:
Creates posts intended to
evade detector
IMPROVED DETECTOR:
Model that intakes data
and returns a label