# CrowdLearn: A Crowd-AI Hybrid System for Deep Learning-based Damage Assessment Applications

Daniel (Yue) Zhang, Yang Zhang, Qi Li, Thomas Plummer, Dong Wang

Department of Computer Science and Engineering

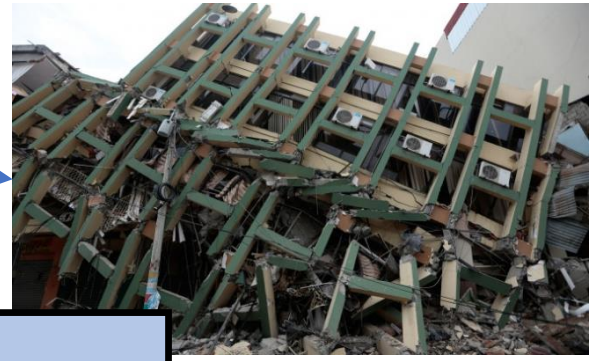
University of Notre Dame, IN, USA

ICDCS 2019, Dallas, TX, USA




# Background: Disaster Response

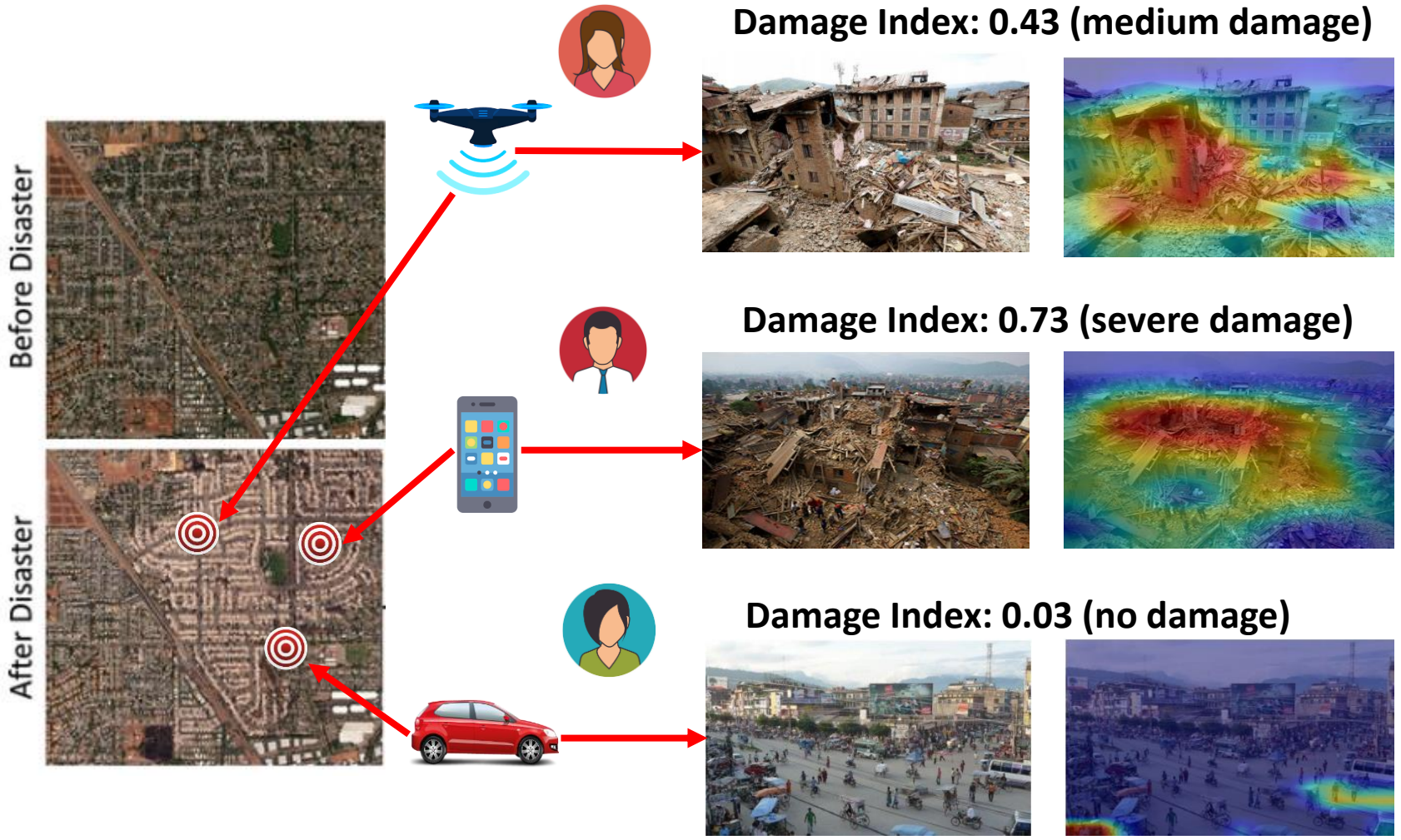
- **In disaster-related events (e.g., hurricanes, tornados), severe damages could simultaneous happen in many places.**
  - Broken bridge, damaged houses, fires, accidents
  - Require immediate attention from response teams (e.g., FEMA, Police, Ambulance)



How to prioritize the response team's operations?

# The Disaster Damage Assessment (DDA) Application

A new sensing application of automatically identifying the areas of maximum damage given images to prioritize rescue operations.



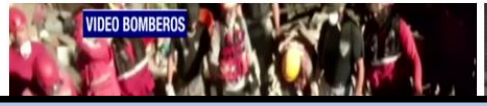
# Motivation – Limitation of AI

- We found existing DDA applications cannot deliver satisfactory accuracy and are prone to various failure scenarios.
  - Cannot distinguish fake images (Figure 2(a))
  - Easy to be mislead by camera angles (Figure 2(b))
  - Confused by low-resolution images (Figure 2(c))
  - Fail to understand implicit stories (Figure 2(d))

(a) Fake Image

(b) Close Up

| Figure | Actual | Reported |
|---|---|---|
| a | no | severe |




(c) Low Resolution

(d) Implicit

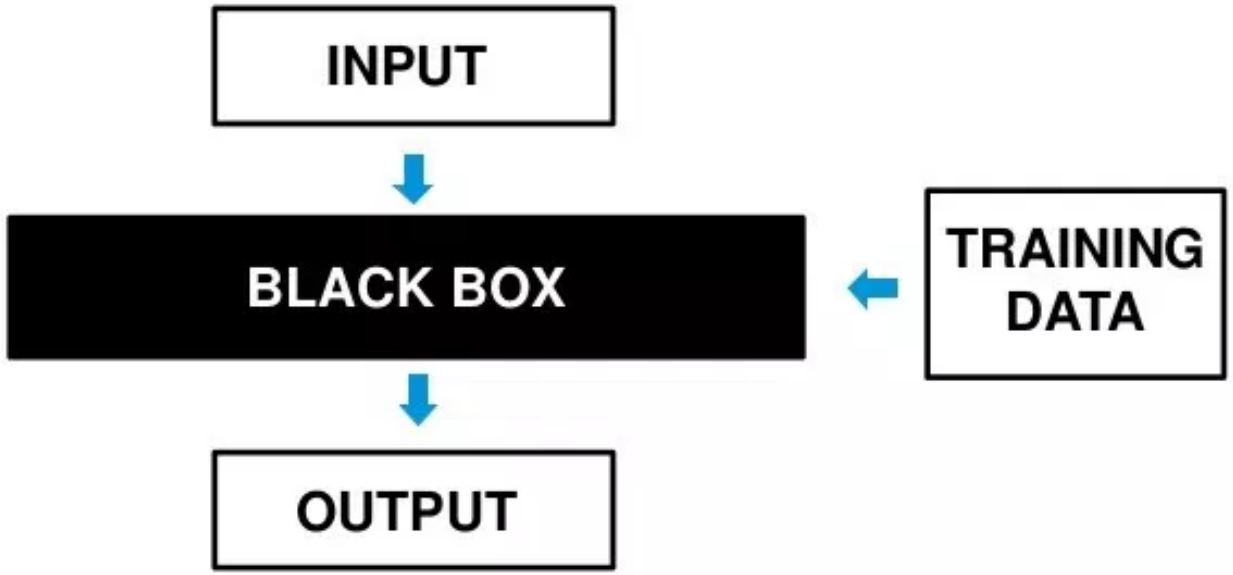
**Motivation: Can Human Intelligence (HI) help AI to deliver better DDA performance?**

**Wrong Outputs of AI** **Failure Scenarios of AI**

- Human on the other hand, can more reliably assess the actual scene described by the image and report the correct damage severity.

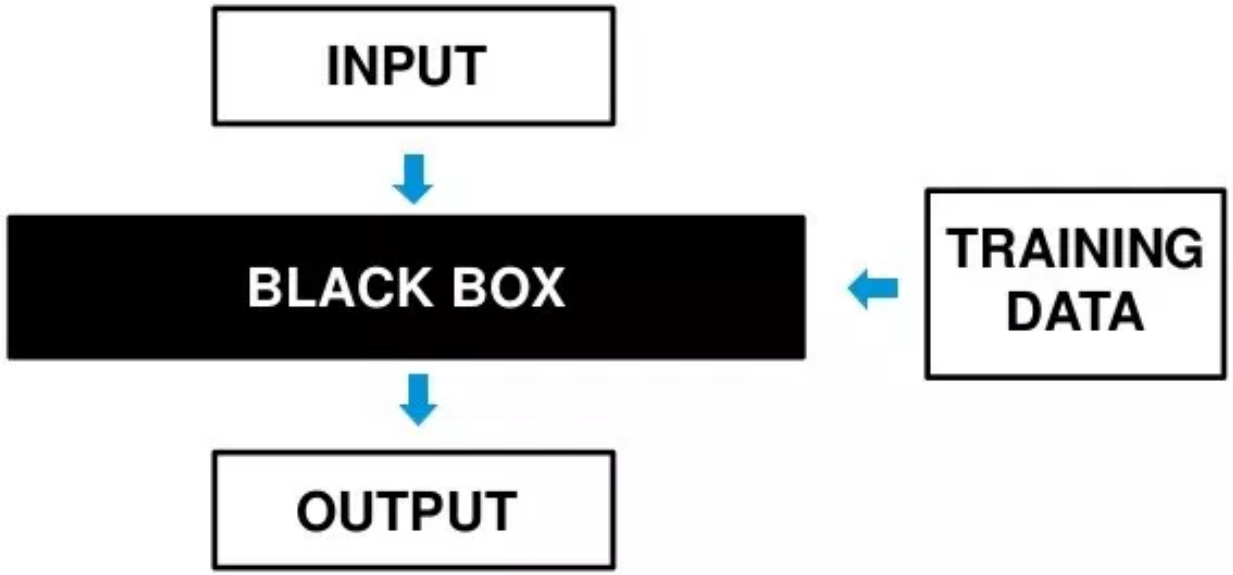
# Challenge 1: Black-Box AI

- The lack of interpretability of the results from AI algorithms makes it extremely hard to diagnose the failure scenarios such as performance deficiency.
  - Why the AI model fails? Is this due to lack of training data or the model itself?
  - How to identify and fix the failure scenarios?



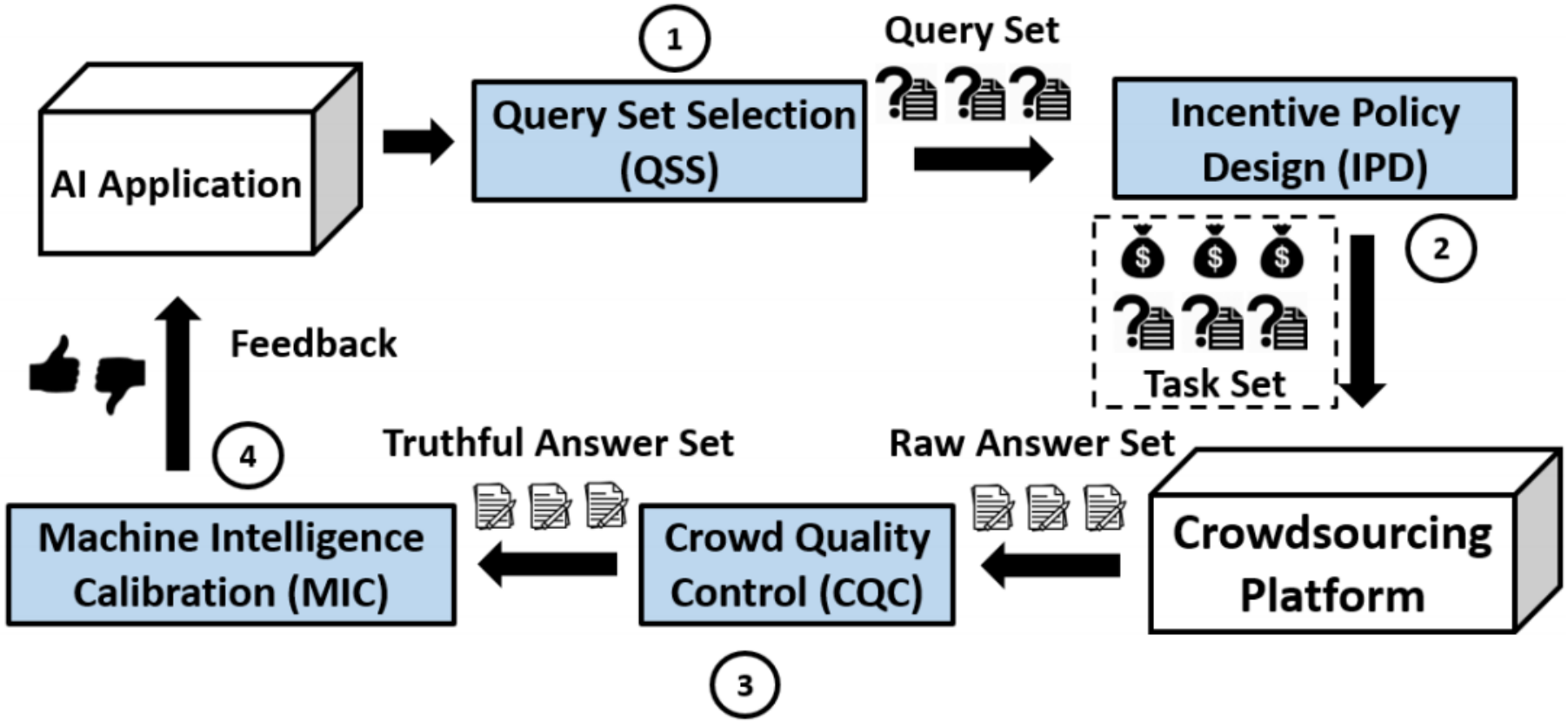
# Challenge 2: Black-Box HI

- the lack of control of crowdsourcing platform makes it difficult to acquire high quality and timely human intelligence.
  - Cannot directly select and manage the workers.
  - The time and quality of the responses from the crowd workers are highly dynamic and unpredictable.



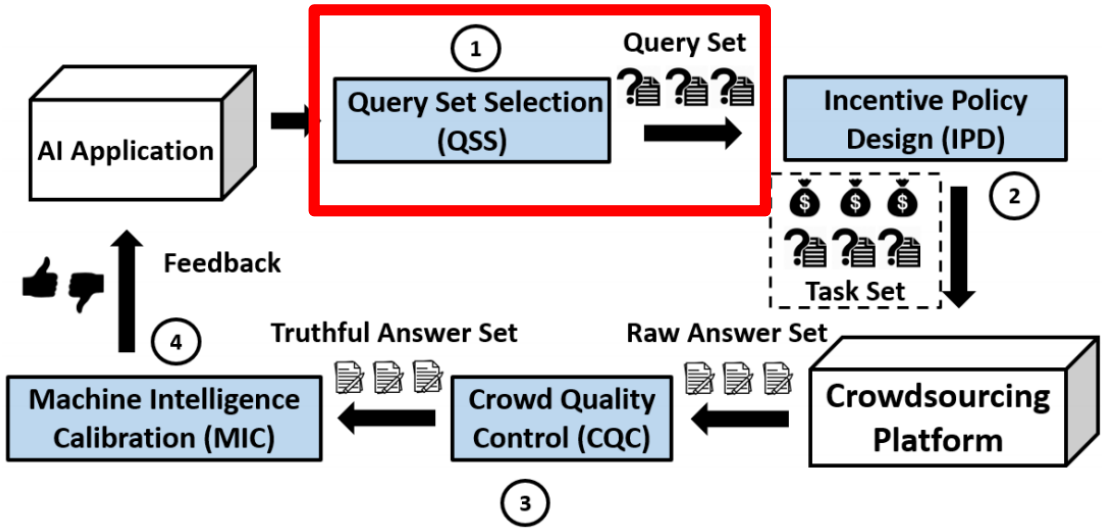
# The CrowdLearn Framework

- A Closed-loop crowd-AI hybrid system
- The AI algorithms selectively identify a subset of data to query the crowd
- The crowd's response is used to tune the AI model



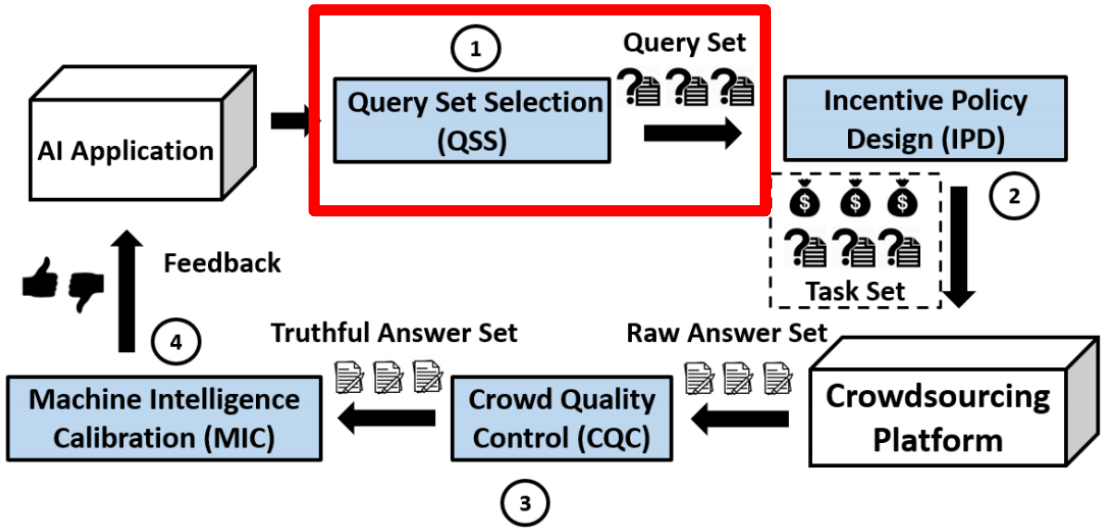
# Query Set Selection (QSS) Module

- QSS selects a set of images to query the crowd.
  - We select a diverse set of AI Algorithms for DDA as a committee
  - Each committee  member votes for the label of the image
  - Pick the images that the committee has the most disagreement



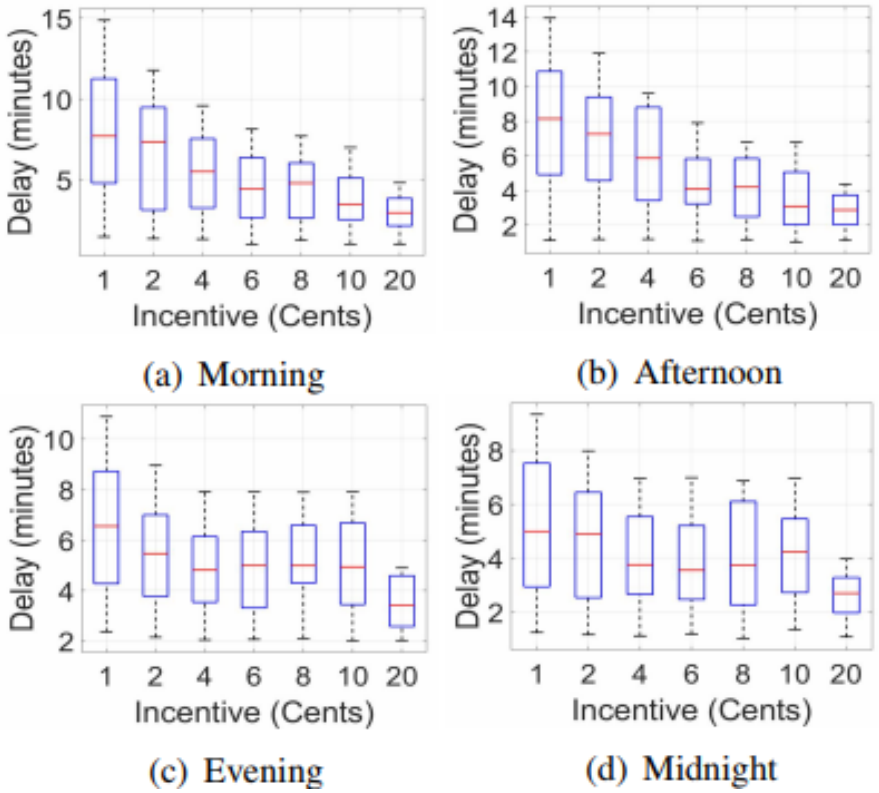
# Query Set Selection (QSS) Module

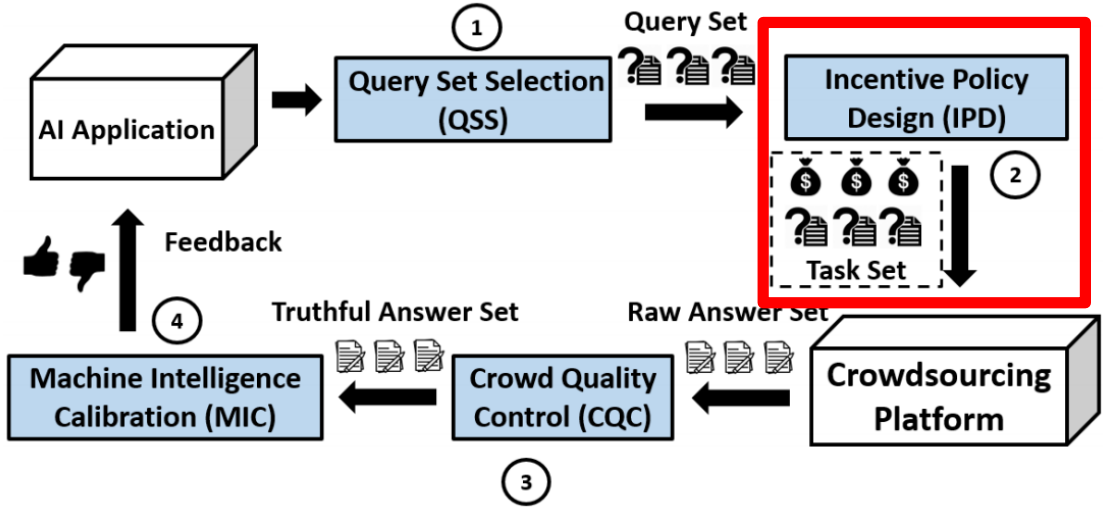
- QSS selects a set of images to query the crowd.
  - We select a diverse set of AI Algorithms for DDA as a committee
  - Each committee  member votes for the label of the image
  - Pick the images that the committee has the most disagreement



# Incentive Policy Design (IPD) Module

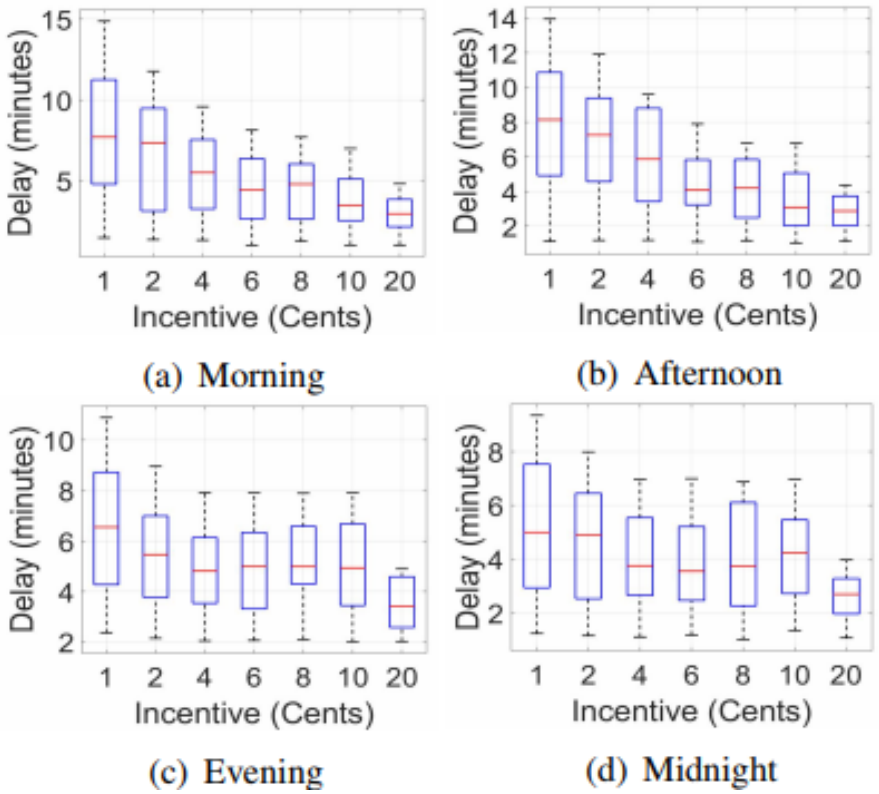
- IPD assigns the optimal incentive for each query to be sent to the crowdsourcing platform to get most timely response.
  - We observe that the incentive is context-aware and the expected level of incentives depends on time of the day.
  - We develop a Contextual Multi-armed Bandit to identify the optimal incentive for each query.

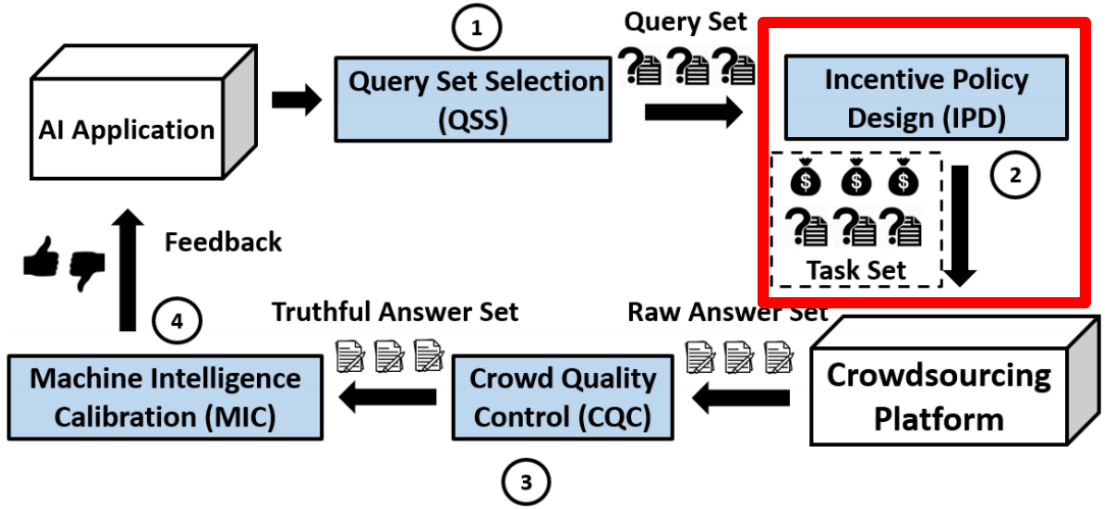


# Incentive Policy Design (IPD) Module

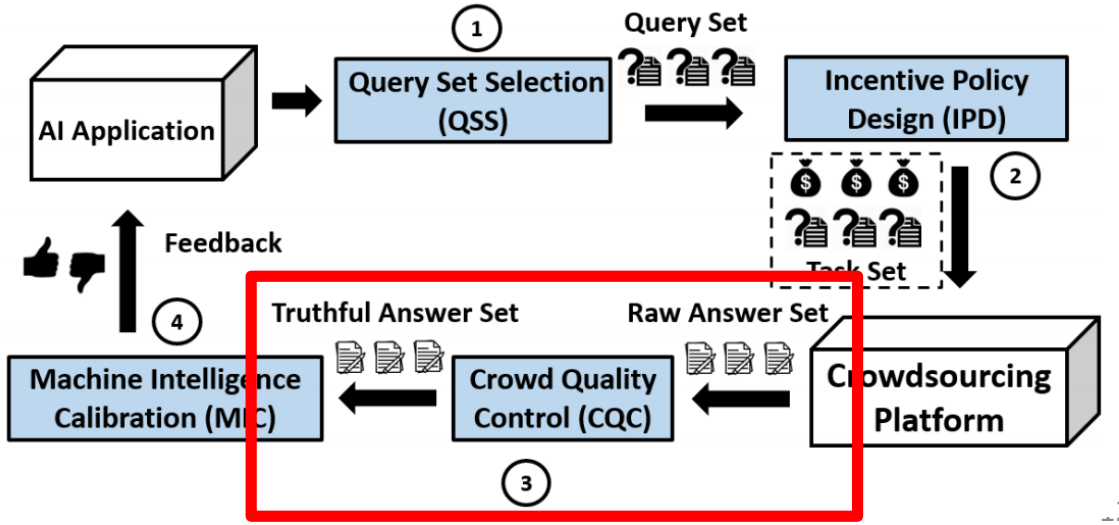
- IPD assigns the optimal incentive for each query to be sent to the crowdsourcing platform to get most timely response.
  - We observe that the incentive is context-aware and the expected level of incentives depends on time of the day.
  - We develop a Contextual Multi-armed Bandit to identify the optimal incentive for each query.



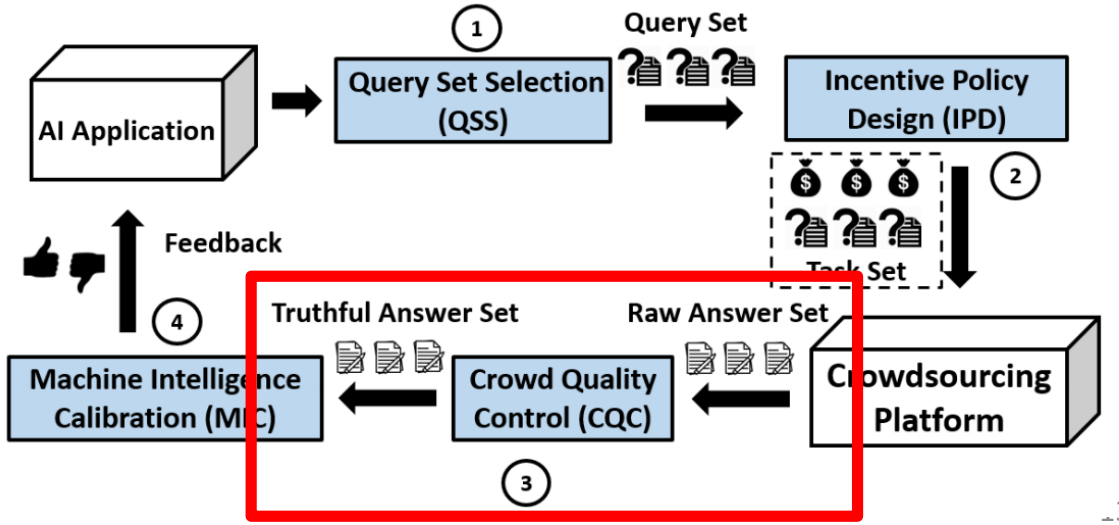
# Crowd Quality Control (CQC) Module

- CQC takes the answers from the crowd and provides quality control to generate truthful answers.
  - We collected both the label and contextual information about an image from human.
  - The label and contextual information are fed into a classifier (XGBoost) to decide the truthful label.

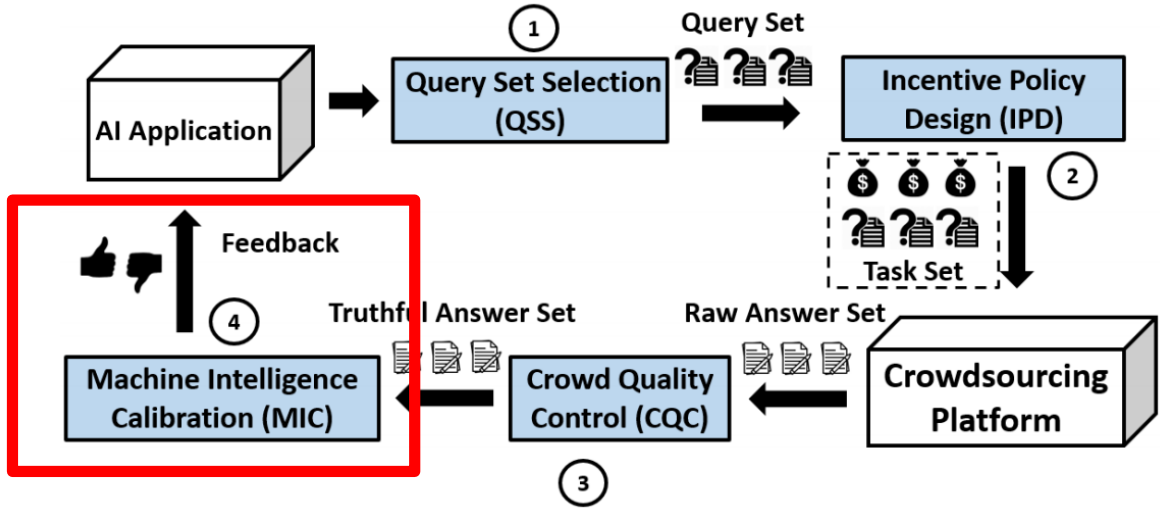
# Crowd Quality Control (CQC) Module

- CQC takes the answers from the crowd and provides quality control to generate truthful answers.
  - We collected both the label and contextual information about an image from human.
  - The label and contextual information are fed into a classifier (XGBoost) to decide the truthful label.

# Machine Intelligence Calibration (MIC) Module

- MIC compares the crowd answers with the results of the AI algorithms and improves AI.
  - Retrain the models with the new labels collected from CQC.
  - Update the expert weights of the AI algorithms in the committee.



# Evaluation

**Dataset**: 3,000 social media images collected from Equator earthquake

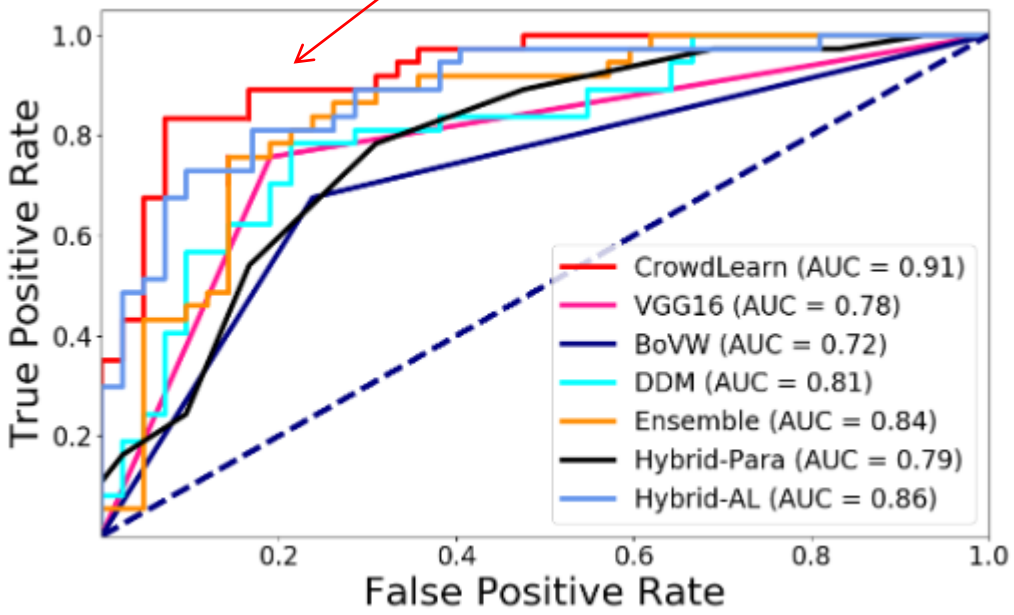
**Baselines**:

- **VGG16**: A DDA scheme that uses deep Convolutional Neural Networks (CNN)
- **BoVW**: A DDA scheme that uses handcrafted features (e.g., scale invariant feature transform, histogram of oriented gradients) to train a neural network classifier
- **DDM**: A DDA scheme that combines CNN and Gradient-weighted Class Activation Mapping (Grad-CAM) to produce a damage heatmap of a given image, which is used to derive the damage severity
- **Ensemble:** An aggregation of the above algorithms (VGG16, BoVW, DDM)
- **Hybrid-Para:** human-AI hybrid system where humans and AI independently label the images
- **Hybrid-AL**: a crowdsourcing-based active learning framework for AI algorithms where the annotated labels collected from MTurk are used to re-train the AI algorithm for the performance improvement

**Crowdsourcing Platform – Amazon Mechanical Turks**

# Results – Effectiveness

Our Solution




**ROC Curve**

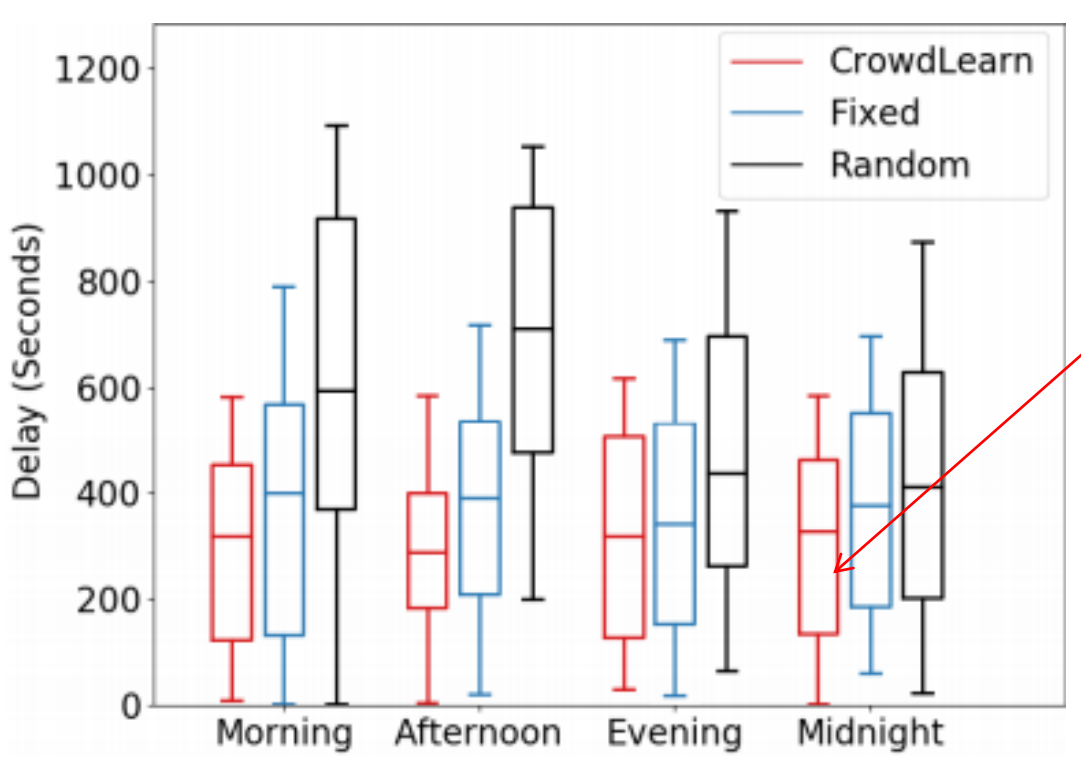
Our Solution

| Algorithms | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| **CrowdLearn** | **0.877** | **0.904** | **0.885** | **0.894** |
| VGG16 | 0.770 | 0.845 | 0.744 | 0.791 |
| BoVW | 0.670 | 0.707 | 0.744 | 0.725 |
| DDM | 0.807 | 0.891 | 0.765 | 0.823 |
| Ensemble | 0.815 | 0.892 | 0.778 | 0.831 |
| Hybrid-Para | 0.797 | 0.849 | 0.795 | 0.821 |
| Hybrid-AL | 0.823 | 0.883 | 0.803 | 0.841 |

**Accuracy, Precision, Recall, and F1**

**Our solution achieves better detection effectiveness.**

# Results – Delay vs. Context




**Our Solution**

**Crowd Delay w.r.t Temporal Contexts**

**Our solution achieves lower delay in various temporal contexts.**

# Thank You!

## Social Sensing Lab@ Notre Dame

https://www3.nd.edu/~sslab/