Statistical Quality Control for Human Computation and Crowdsourcing

Yukino Baba (University of Tsukuba)

Early career spotlight talk @ IJCAI-ECAI 2018

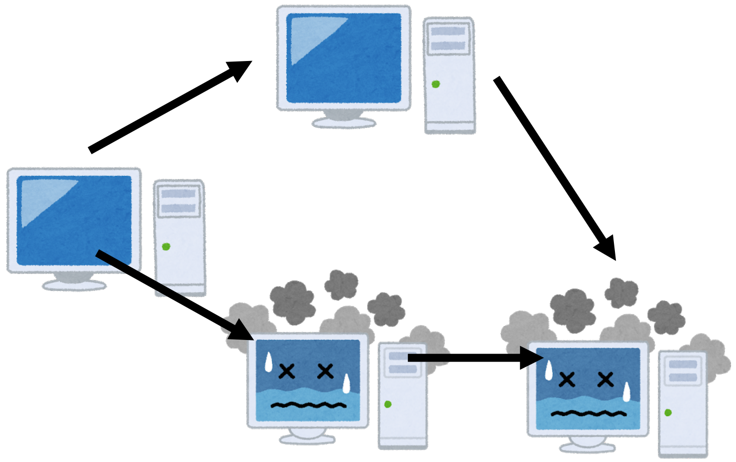
July 18, 2018

Humans and computers collaboratively solve problems

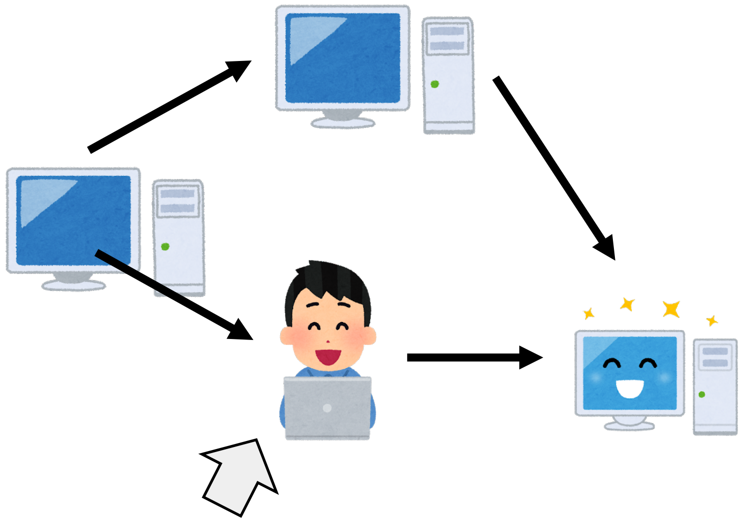
- Combining humans and computers for solving hard problems

☞ Querying human intelligence from computer systems

Computer system



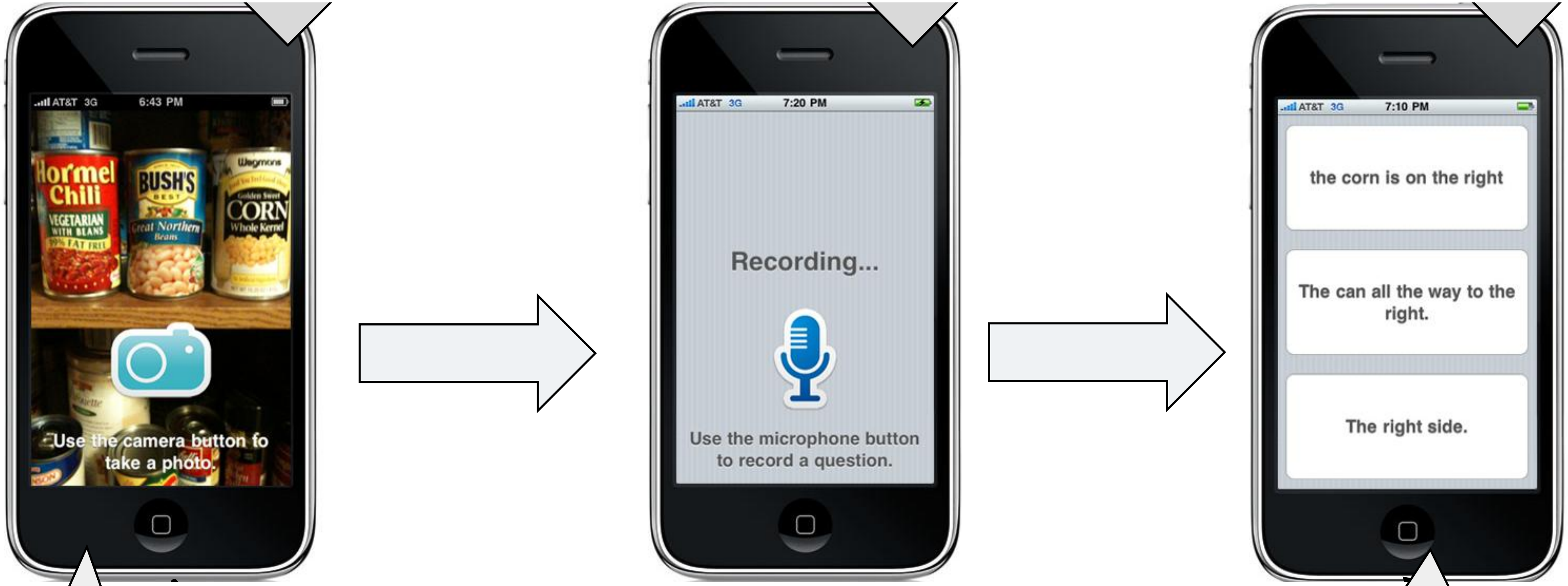
Human computation system



Crowdsourcing platform

EXAMPLE: VIZWIZ [Bigham+ 2010]

Human computation for supporting blind people



Use the camera button to take a photo.

Recording...

Use the microphone button to record a question.

the corn is on the right

The can all the way to the right.

The right side.

Step 1: user posts a visual question, e.g., *“which can is the corn?”*

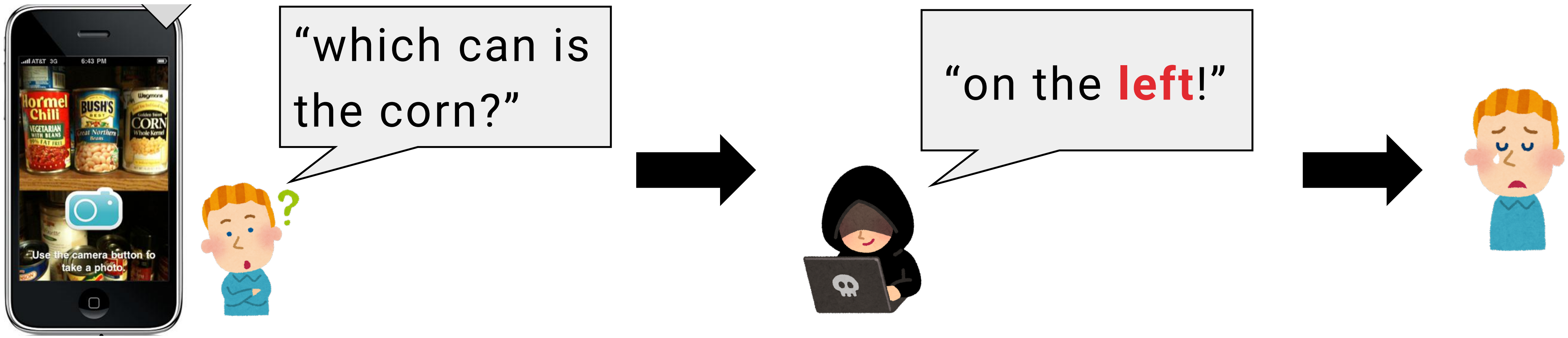
Step 2: humans inside the system answer the question, e.g., *“on the right”*

CHALLENGE

Quality control is a big challenge in human computation

- There is no guarantee all participants will answer correctly
  - Uncertainty: everyone can make mistakes
  - Diversity: people have different levels of reliability

Example: VizWiz with unreliable workers

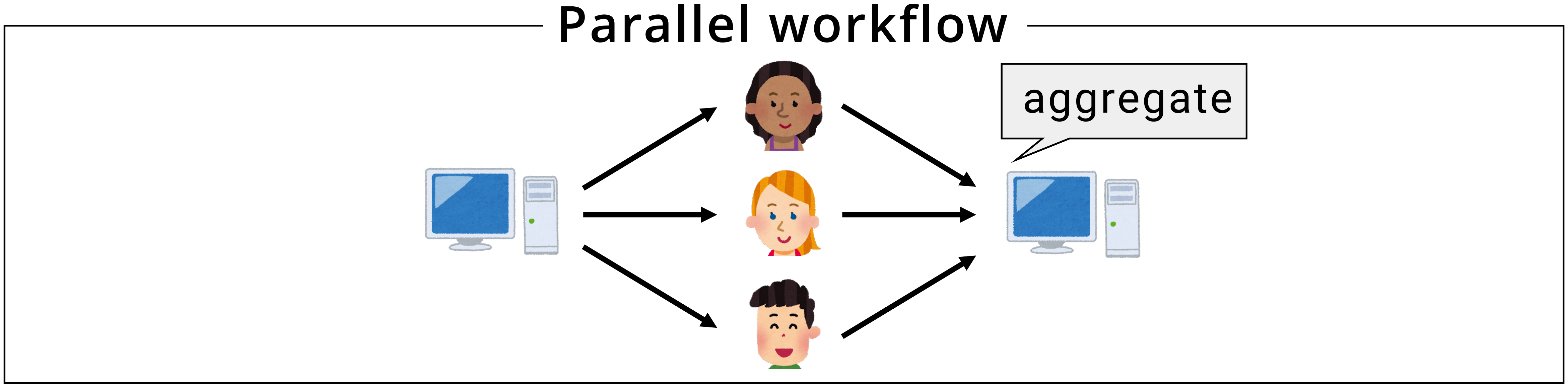


"which can is the corn?"

"on the **left**!"

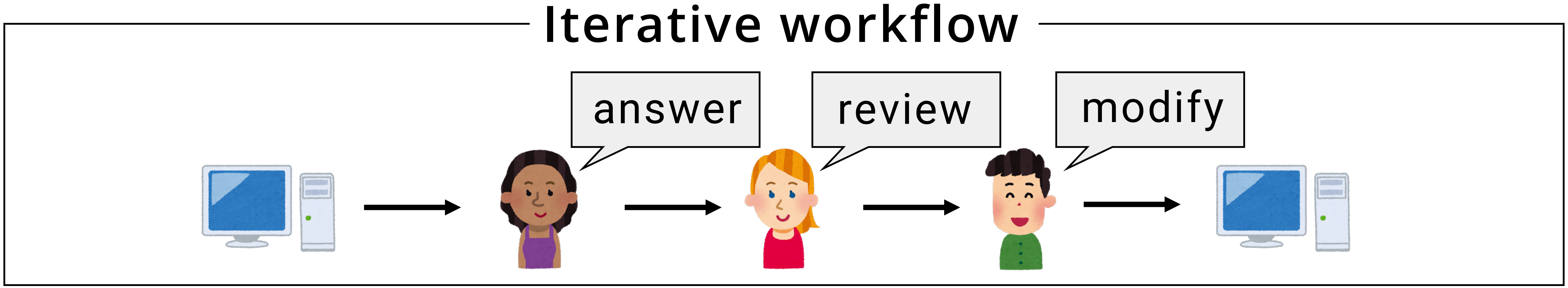
SOLUTION

Let multiple participants be involved in each task



Parallel workflow

aggregate



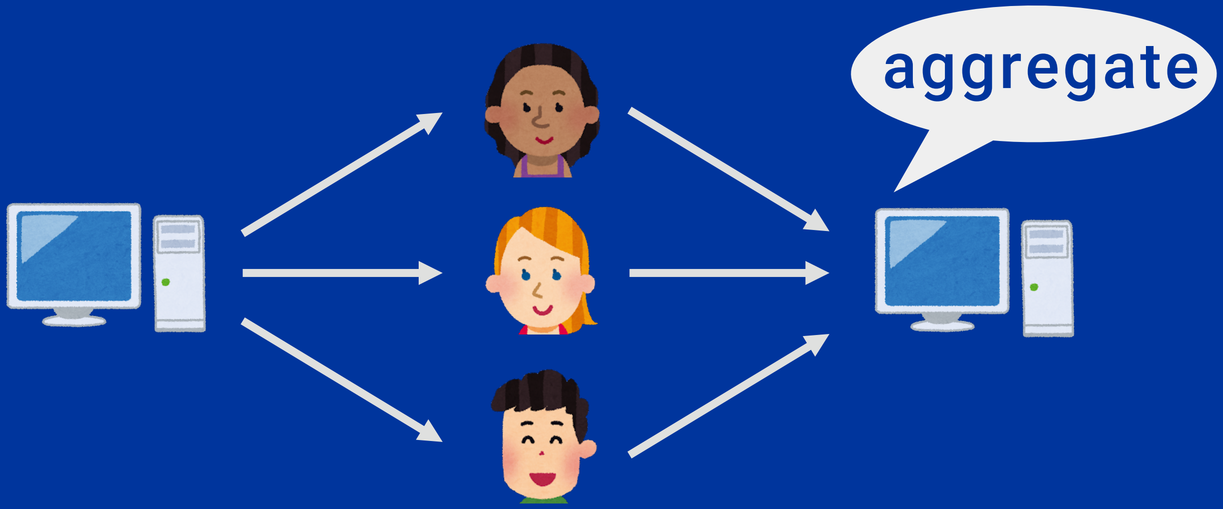
Iterative workflow

answer

review

modify

Statistical modeling for parallel workflow

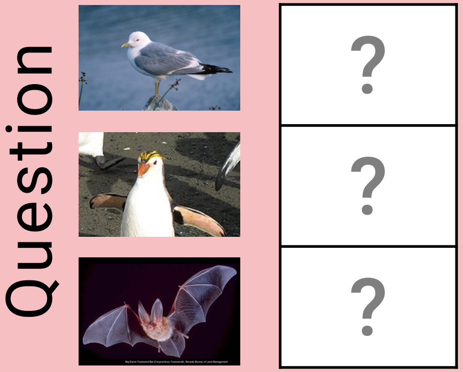


PROBLEM SETTING

We aim to estimate true answers from worker answers

Target: true answer

“Does this image have a bird?”

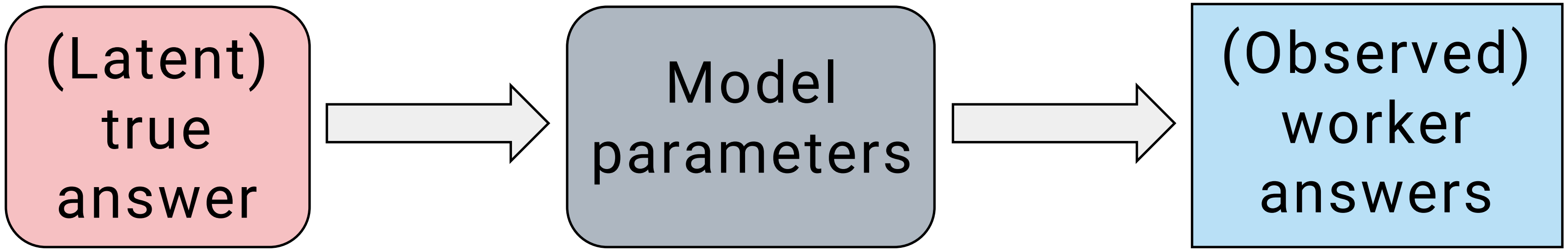


Estimate

Given: worker answers

Question				
	YES	YES	YES	NO
	NO	YES	YES	YES
	NO	YES	NO	YES

Generative model



(Latent) true answer → Model parameters → (Observed) worker answers

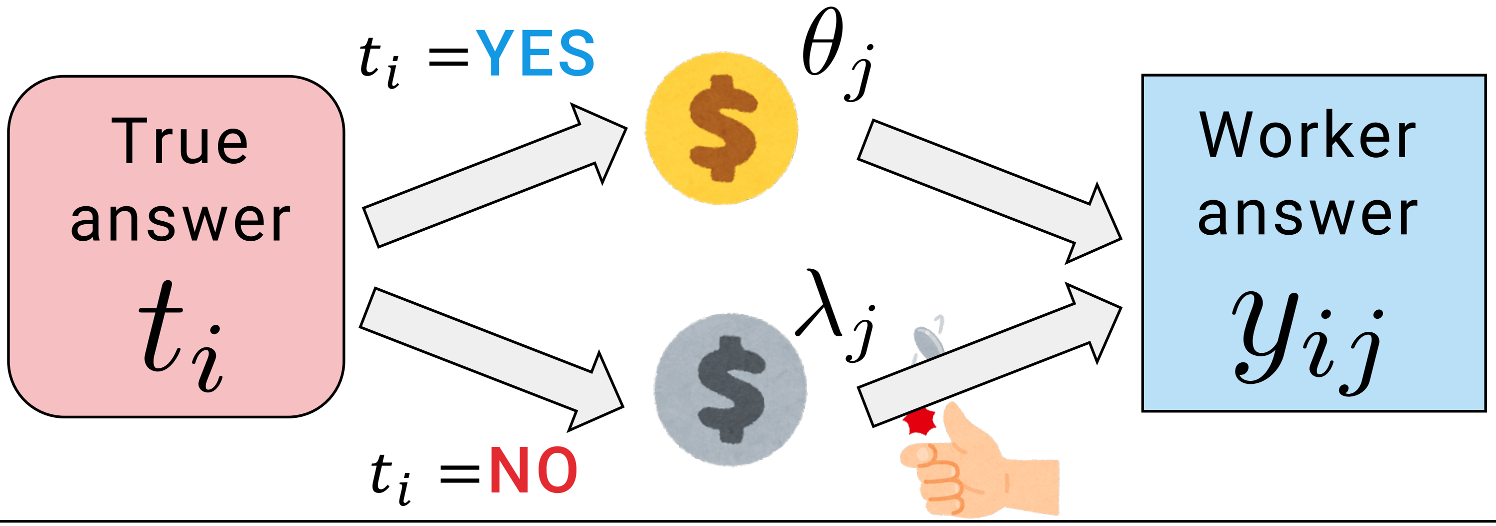
Worker reliability is incorporated into the model

Reliability parameters of each worker j

θ_j : Probability of answering YES when the true answer is YES

λ_j : Probability of answering NO when the true answer is NO

Generative model

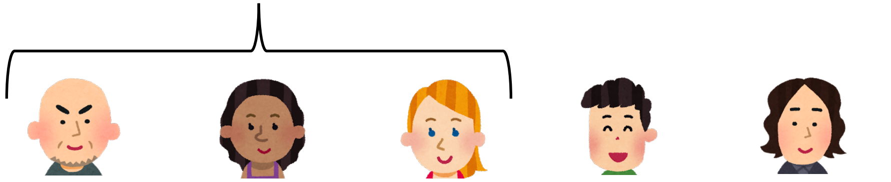


DRAWBACK OF EXISTING APPROACHES

They often fail when the majority is incorrect

- The DS method emphasizes the answers of the majority
  - Other sophisticated approaches work in a similar manner
- When the majority is incorrect, wrong workers can be considered reliable

Considered as reliable



Question	YES	YES	YES	NO	NO
	YES	YES	YES	NO	NO

Example of a difficult question

Q. Which of the following drugs is most likely to cause Cushing's syndrome with long-term use?
(a) Heparin, (b) Insulin, (c) Theophylline,
(d) Prednisolone

Oyama, Baba, Sakurai&Kashima IJCAI’13

Directly ask workers to report their confidence

- We ask workers to report the confidence with their answers

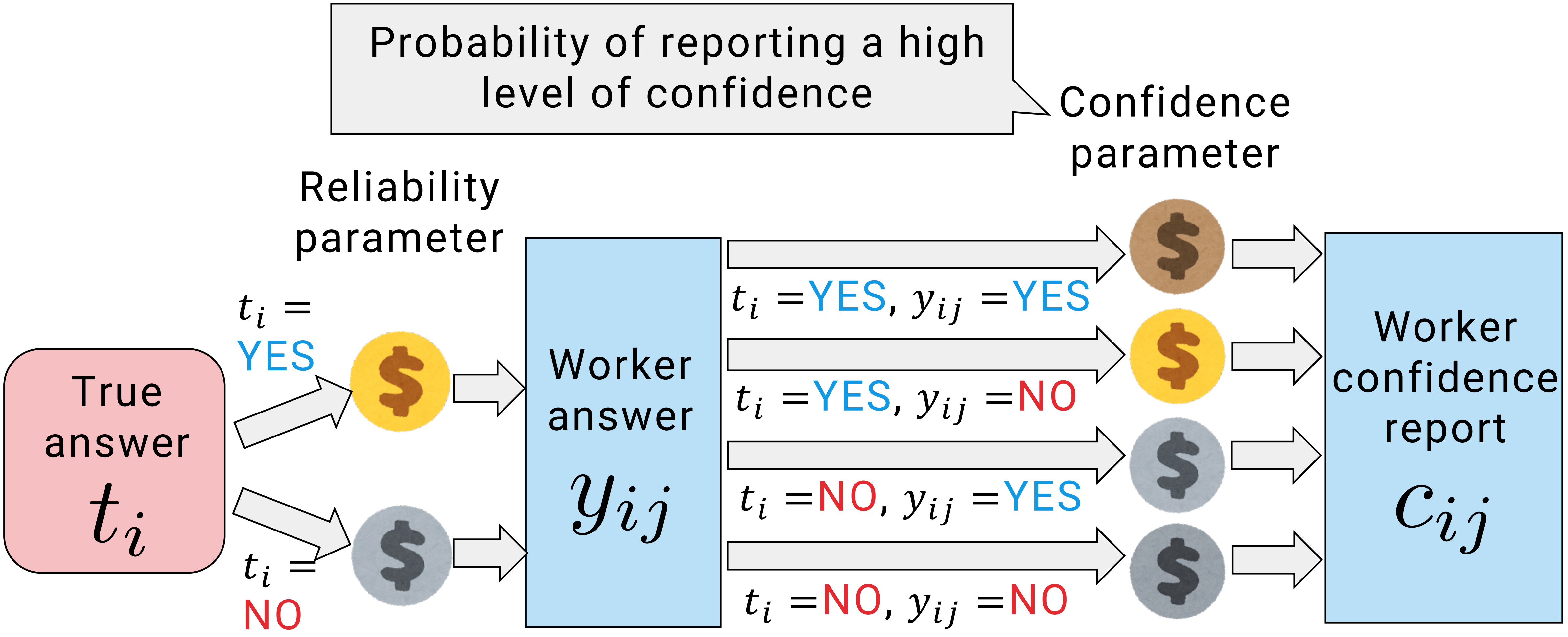
Q1. Is this “Blue-winged Warbler”? ○ YES ○ NO

Q2. Are you confident with you answer? ○ YES ○ NO

Confidence reports

- Confidence reports can be useful for targeting reliable workers (i.e., experts), but some workers report wrongly
  - Overconfident
  - Underconfident

Confidence parameters are incorporated into the model

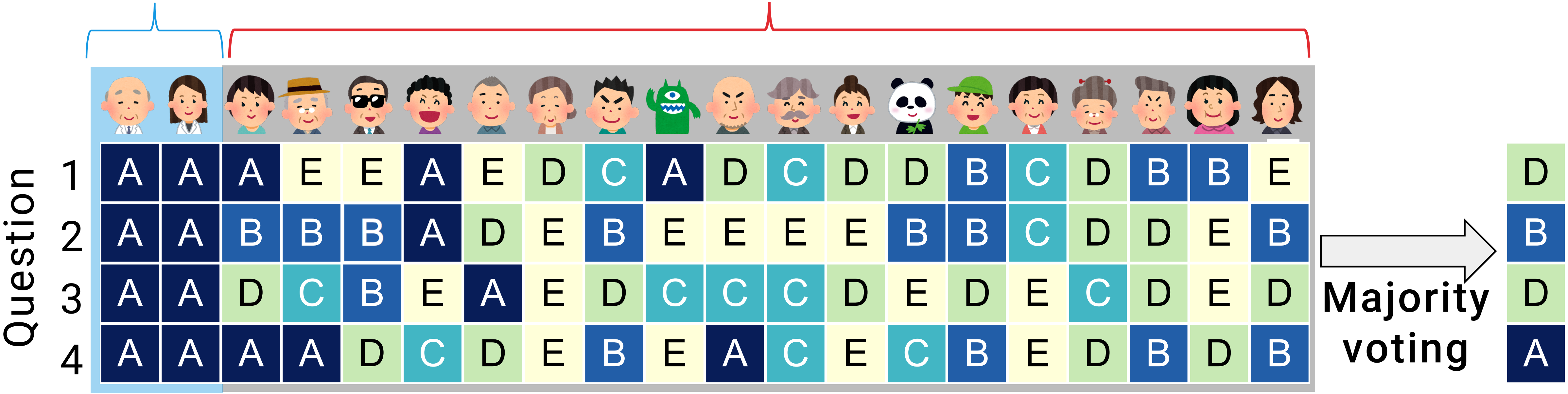


Experts are more likely to agree with each other

Example of an extreme case

Experts:
always answer correctly

Non-experts:
guess randomly

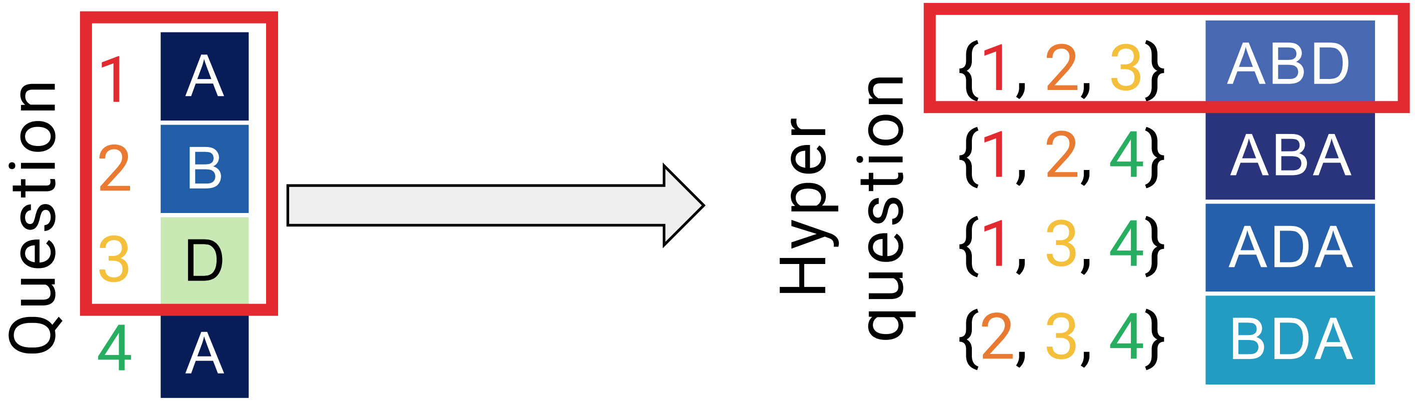


Question																					Majority voting
1	A	A	A	E	E	A	E	D	C	A	D	C	D	D	B	C	D	B	B	E	D
2	A	A	B	B	B	A	D	E	B	E	E	E	E	B	B	C	D	D	E	B	B
3	A	A	D	C	B	E	A	E	D	C	C	C	D	E	D	E	C	D	E	D	D
4	A	A	A	A	D	C	D	E	B	E	A	C	E	C	B	E	D	B	D	B	A

NOTE: “A” is the correct answer for all questions

We focus on sets of questions rather than single ones

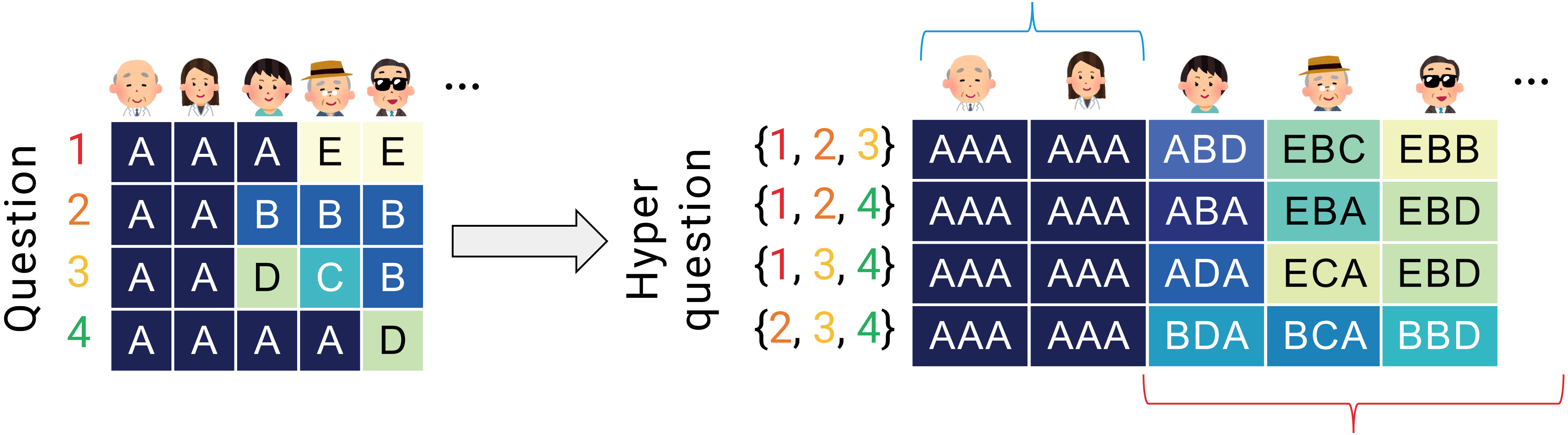
- Hyper question: random subset of single questions
  - E.g., 3-hyper questions of four questions {1, 2, 3, 4} are {1, 2, 3}, {1, 2, 4}, {1, 3, 4}, and {2, 3, 4}
- Answer to a hyper question: concatenation of the answers to the single questions



Li, Baba&Kashima CIKM’17

Hyper questions let experts win in majority voting

Experts can still reach a consensus on hyper questions and become majority

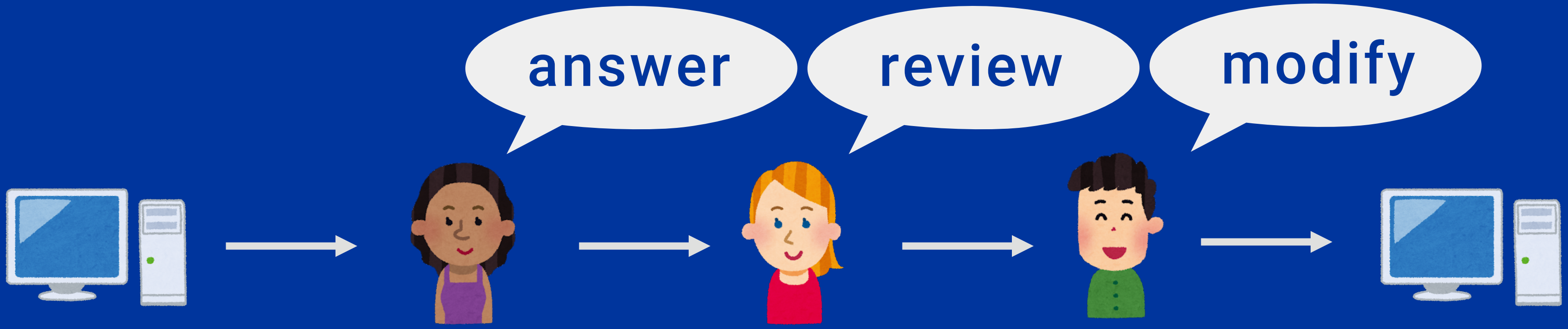


Question					
1	A	A	A	E	E
2	A	A	B	B	B
3	A	A	D	C	B
4	A	A	A	A	D

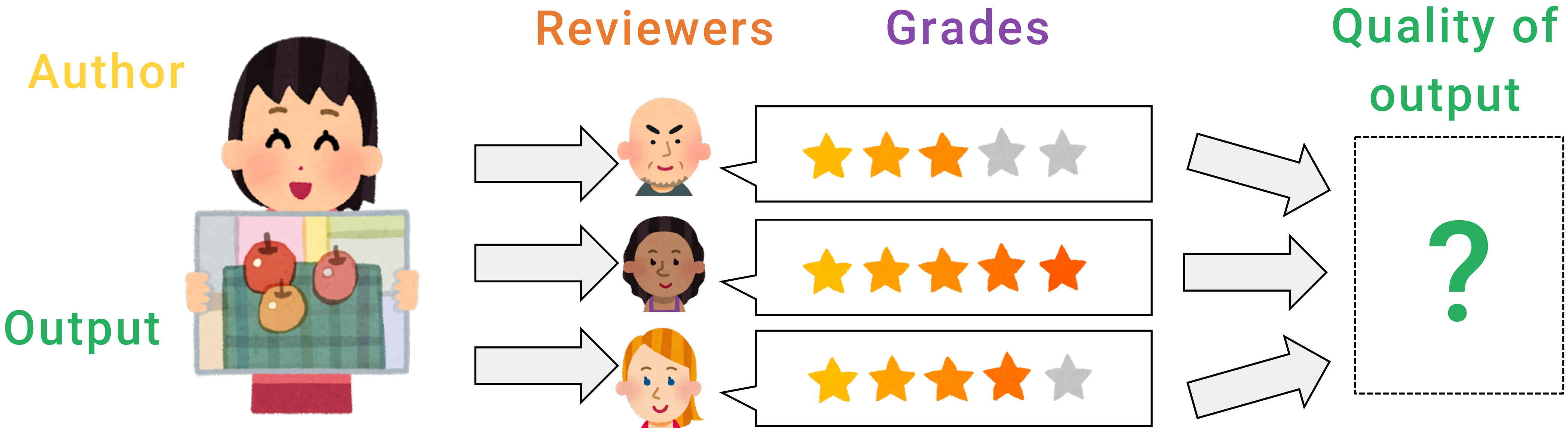
Hyper question					
{1, 2, 3}	AAA	AAA	ABD	EBC	EBB
{1, 2, 4}	AAA	AAA	ABA	EBA	EBD
{1, 3, 4}	AAA	AAA	ADA	ECA	EBD
{2, 3, 4}	AAA	AAA	BDA	BCA	BBD

Non-experts have less chance to reach a consensus on hyper questions

Statistical modeling for iterative workflow



Given grades, we aim to predict the quality of output



No guarantee that all reviewers are reliable

Baba&Kashima KDD’13

Each author has ability and variance parameters

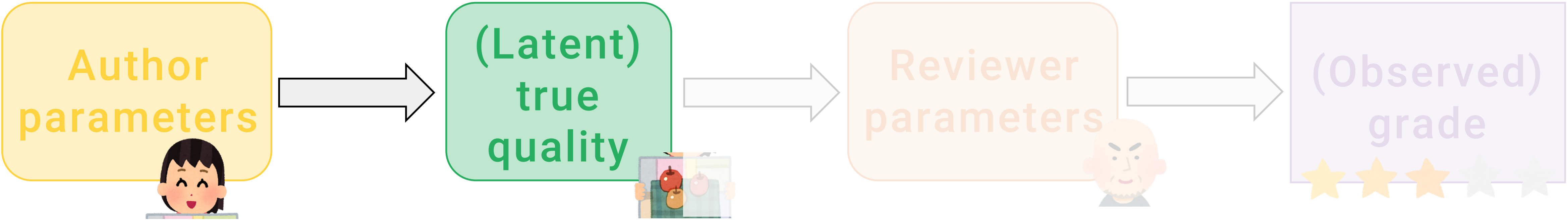
Step 1: Generative model of quality

Author’s variance

$$q_{ta} \sim \mathcal{N}\left(\mu_a, \sigma_a^2\right)$$

(Latent) true quality

Author’s ability



Author parameters

(Latent) true quality

Reviewer parameters

(Observed) grade

Baba&Kashima KDD’13

Each reviewer has bias and variance parameters

Step 2: Generative model of grade

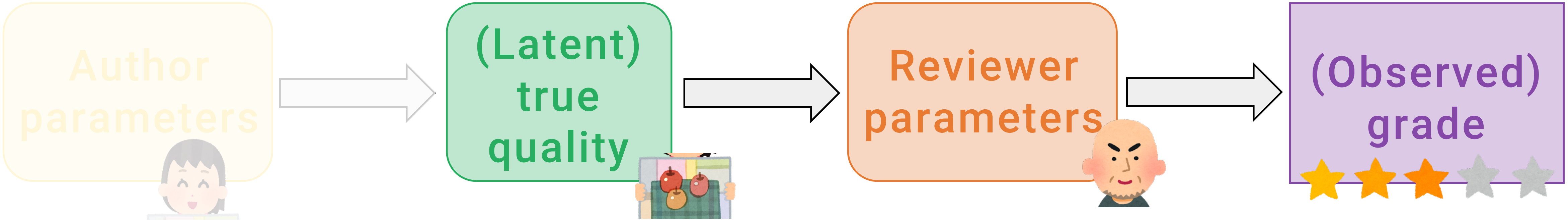
Reviewer’s variance

$$g_{tar} \sim \mathcal{N}\left(q_{ta} + \eta_r , \sigma_r^2\right)$$

(Observed) grade

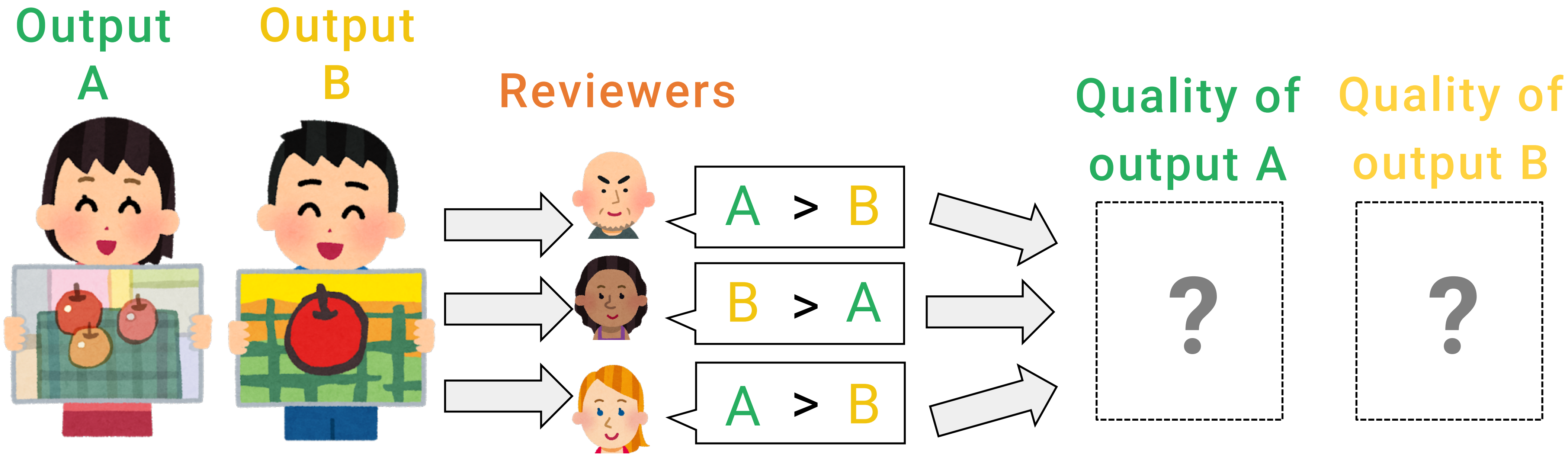
(Latent) true quality

Reviewer’s bias



Sunahase, Baba&Kashima AAAI’17

Comparison results are used for quality estimation



Idea

“Good reviewer votes for many good outputs”

“Good output is voted for by many good reviewers”

Quality is updated based on the weighted num. of votes

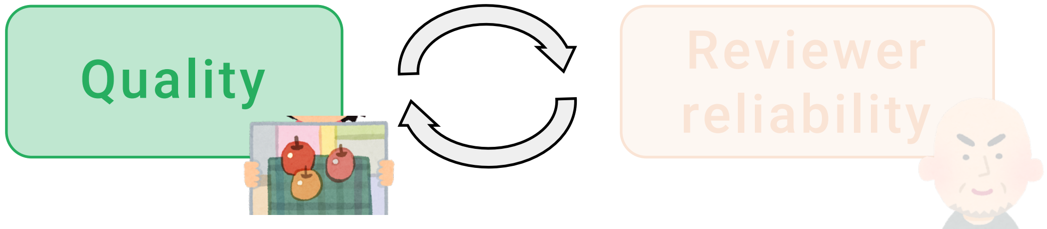
Step 1: update quality

$$q_j - q_k = \sum_{i \in V_{j \succ k}} r_i - \sum_{i \in V_{k \succ j}} r_i$$

Quality of output j

Reliability of reviewer voting for output j

Reliability of reviewer voting for output k



Reliability is updated by the proportion of correct votes

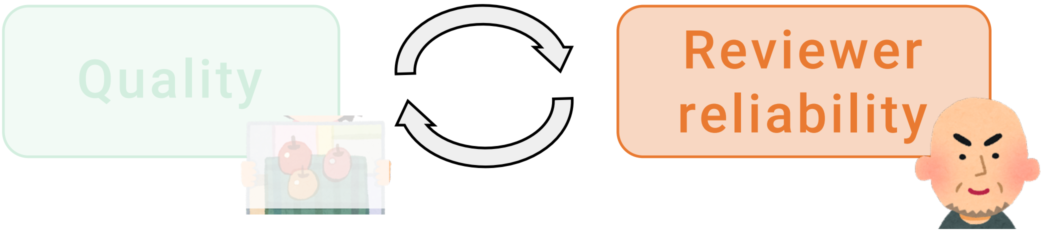
Step 2: update reviewer reliability

$$r_i = \frac{|\{(j \succ k) \in V_i \mid q_j > q_k\}|}{|V_i|}$$

Num. of correct votes given by the reviewer

Reviewer’s reliability

Num. of votes given by the reviewer



Statistical quality control in human computation

- Our approach
  - Statistical modeling for parallel and iterative workflow in human computation
- Open questions
  - How can we assign the reliability of each worker when there can be multiple correct answers?
  - How can we design a systematic way of letting people reach a consensus in complex questions?