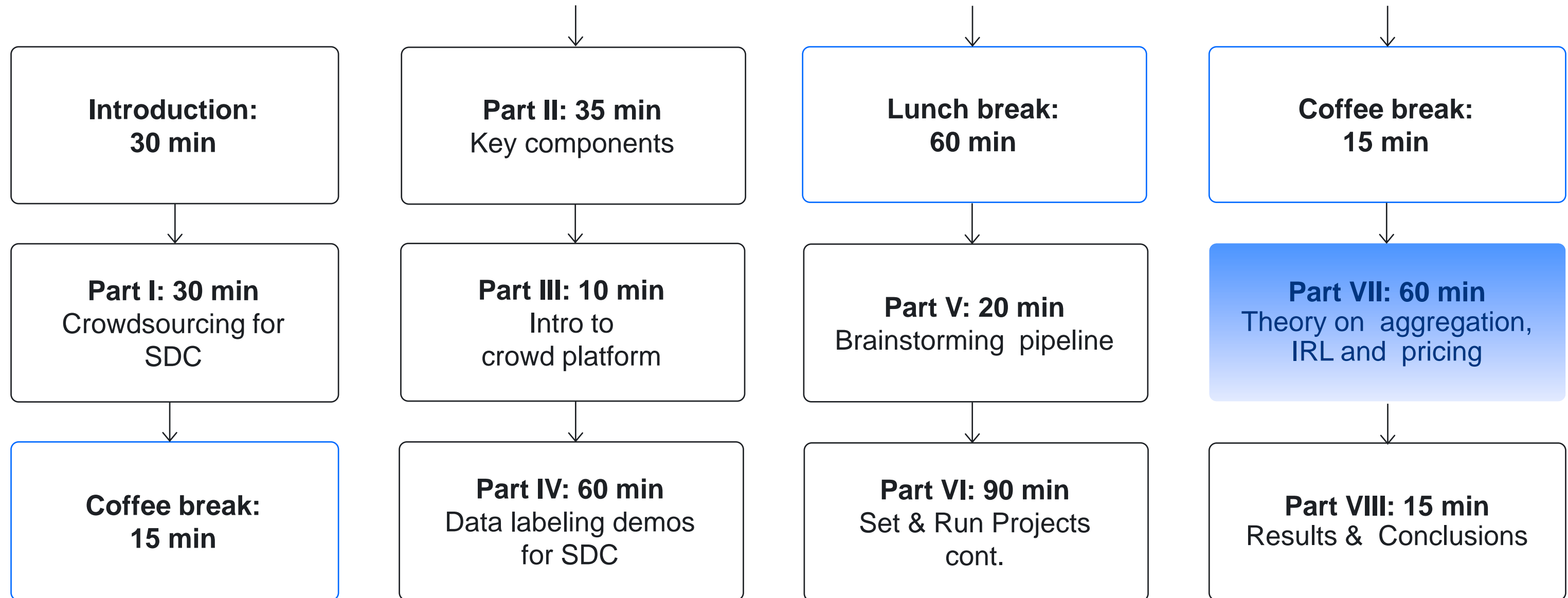


Part VII

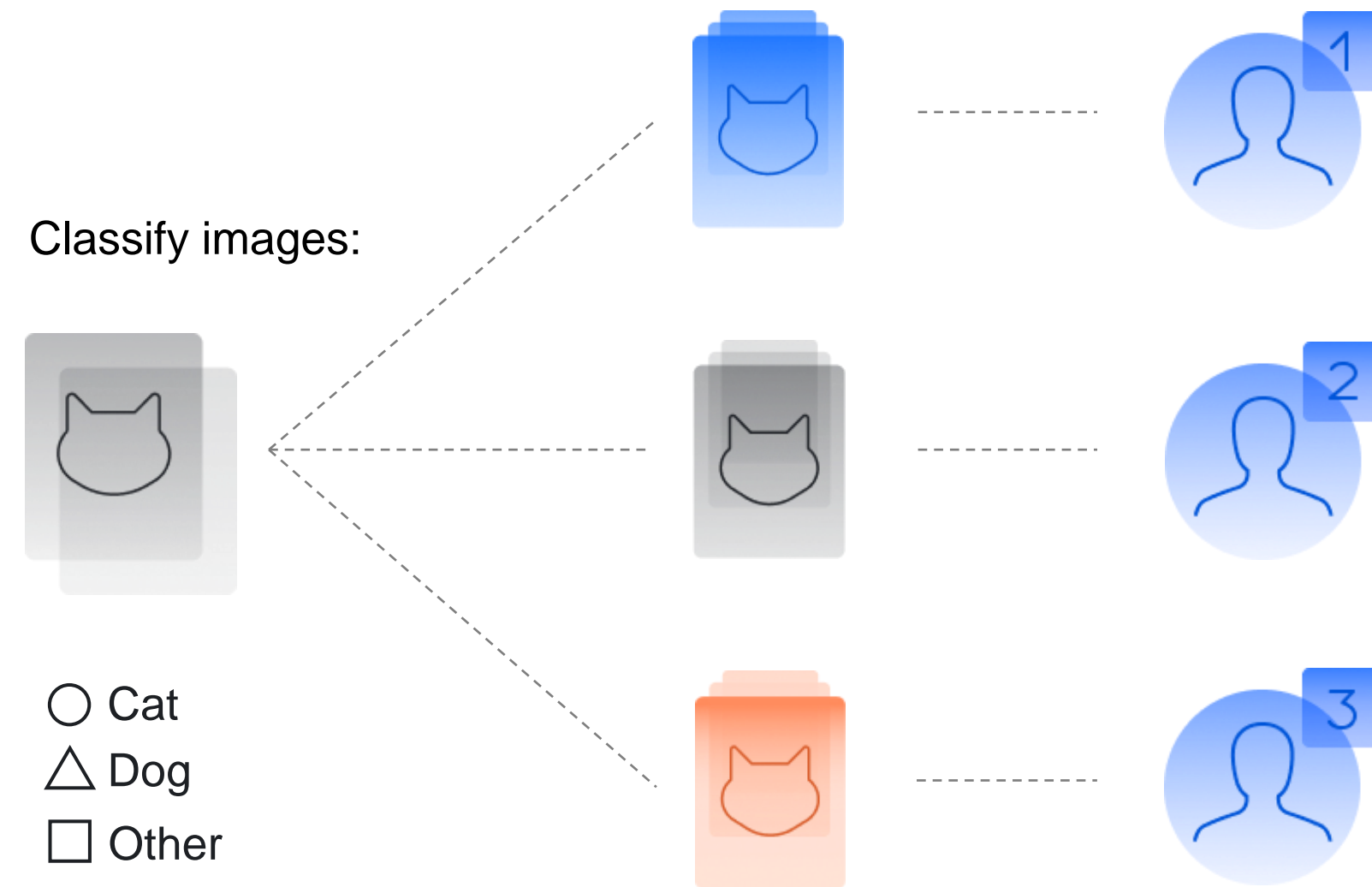
# Theory on Aggregation

Alexey Drutsa, Head of Efficiency and Growth Division, Toloka

# Tutorial schedule



# Labeling data with crowdsourcing



- ▶ How to choose a reliable label?
- ▶ How many labels per object?
- ▶ How much to pay for labels?
- ▶ ...

# Evaluation of labeling approaches



VS

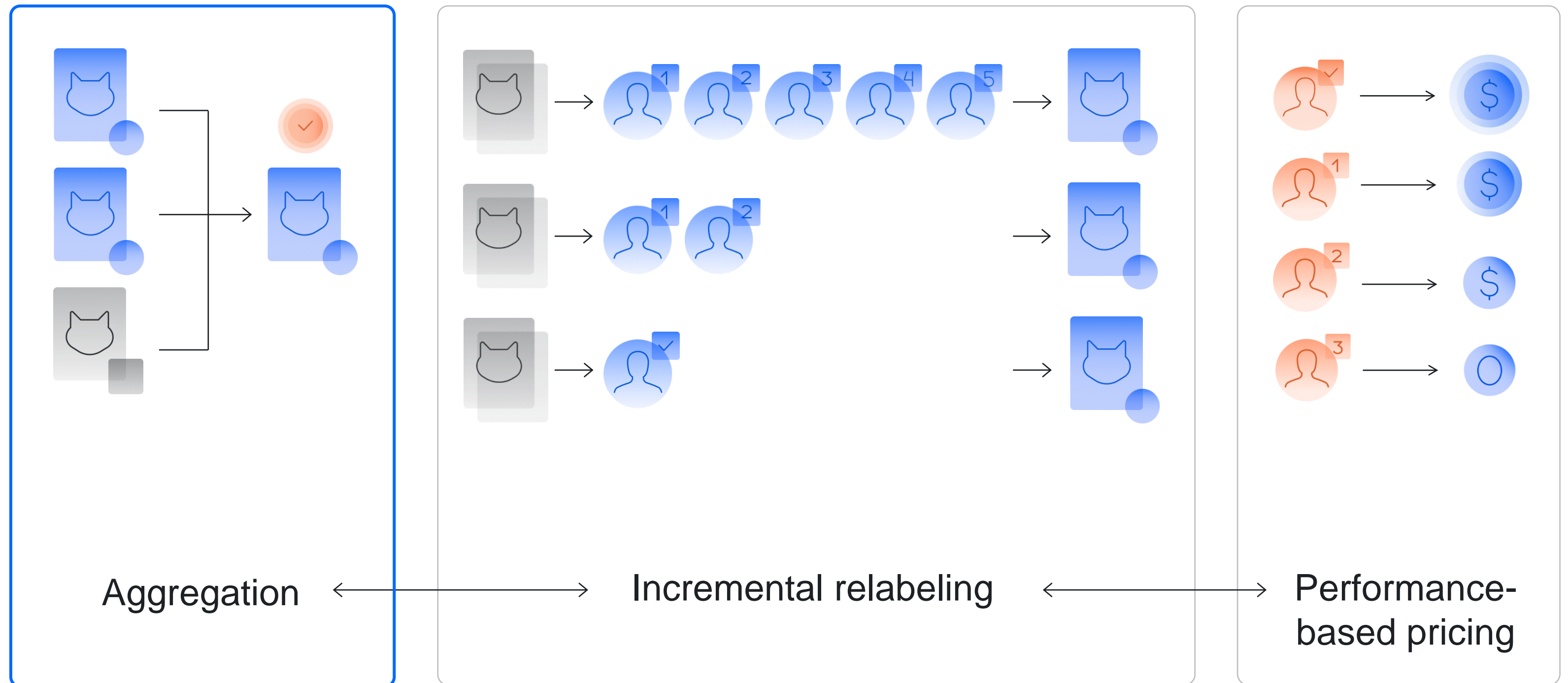


Accuracy

Cost

- ▶ Labels with a **maximal level of accuracy** for a **given budget**
- ▶ Labels of a **chosen accuracy level** for a **minimal budget**

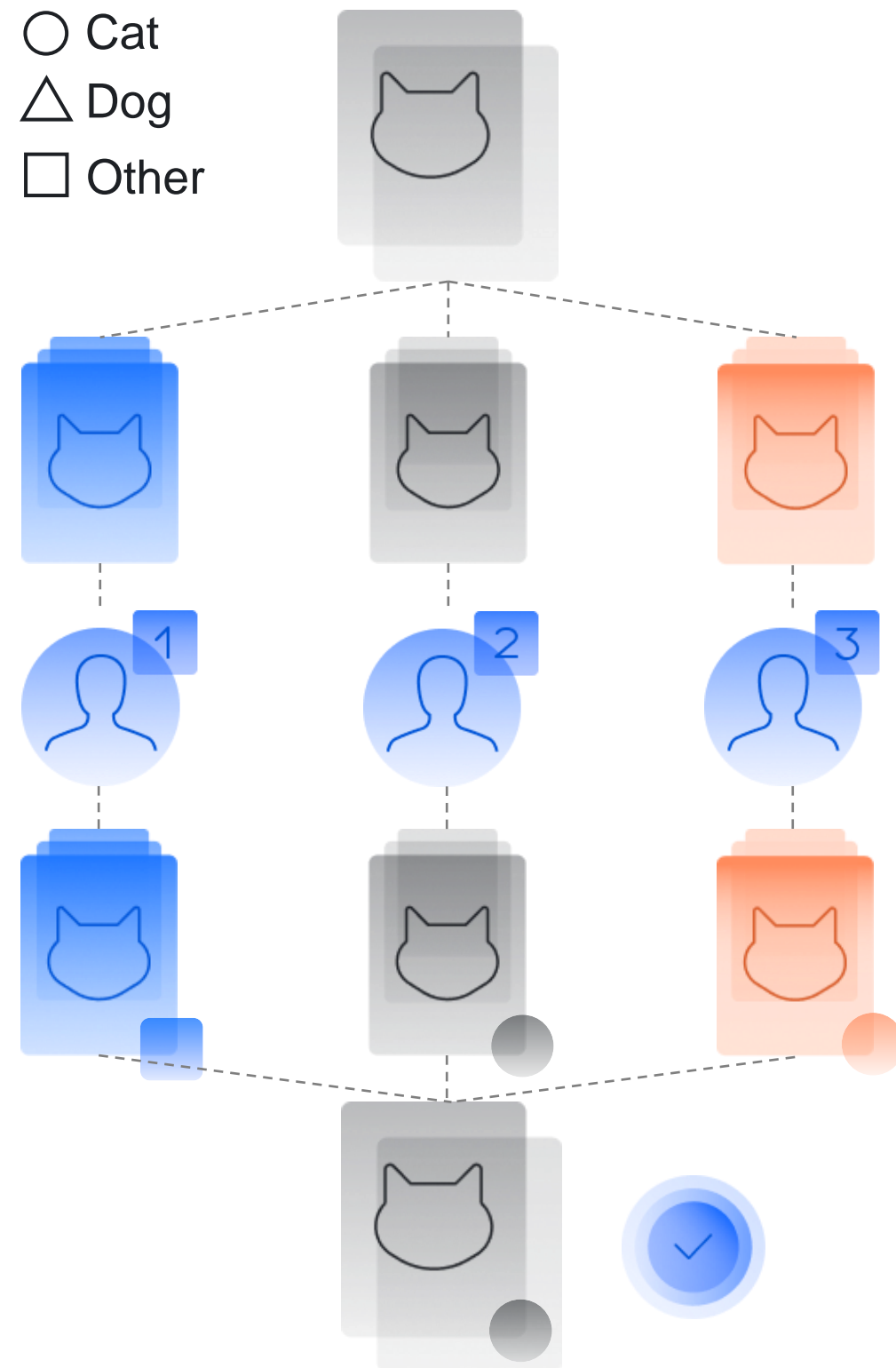
# Key components of labeling with crowds





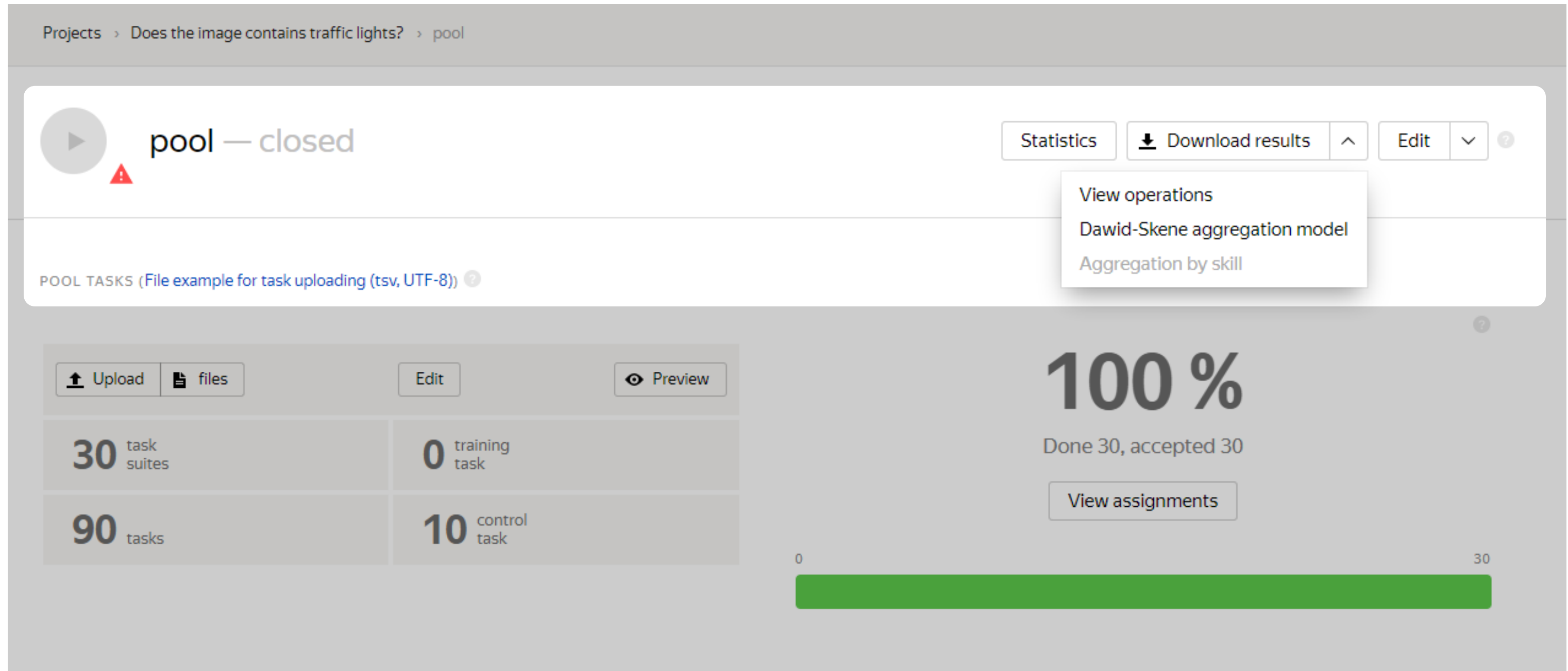
# Aggregation

# Labeling data with crowds



- Classify images
- Upload multiple copies of each object to label
- Performers assign noisy labels to objects
- Aggregate multiple labels for each object into a more reliable one

# Process results





# Multiclass labels

# Project 1: Filter images

Are there traffic lights in the picture?

Yes

No



# Notation

- ▶ Categories  $k \in \{1, \dots, K\}$ . E.g.:
- ▶ Objects  $j \in \{1, \dots, J\}$ . E.g.:
- ▶ Performers:  $w \in \{1, \dots, W\}$ . E.g.:
  - $W_j \subseteq \{1, \dots, W\}$  — performers labeled object  $j$

○ Cat



△ Dog



□ Other





# The simplest aggregation: Majority Vote (MV)

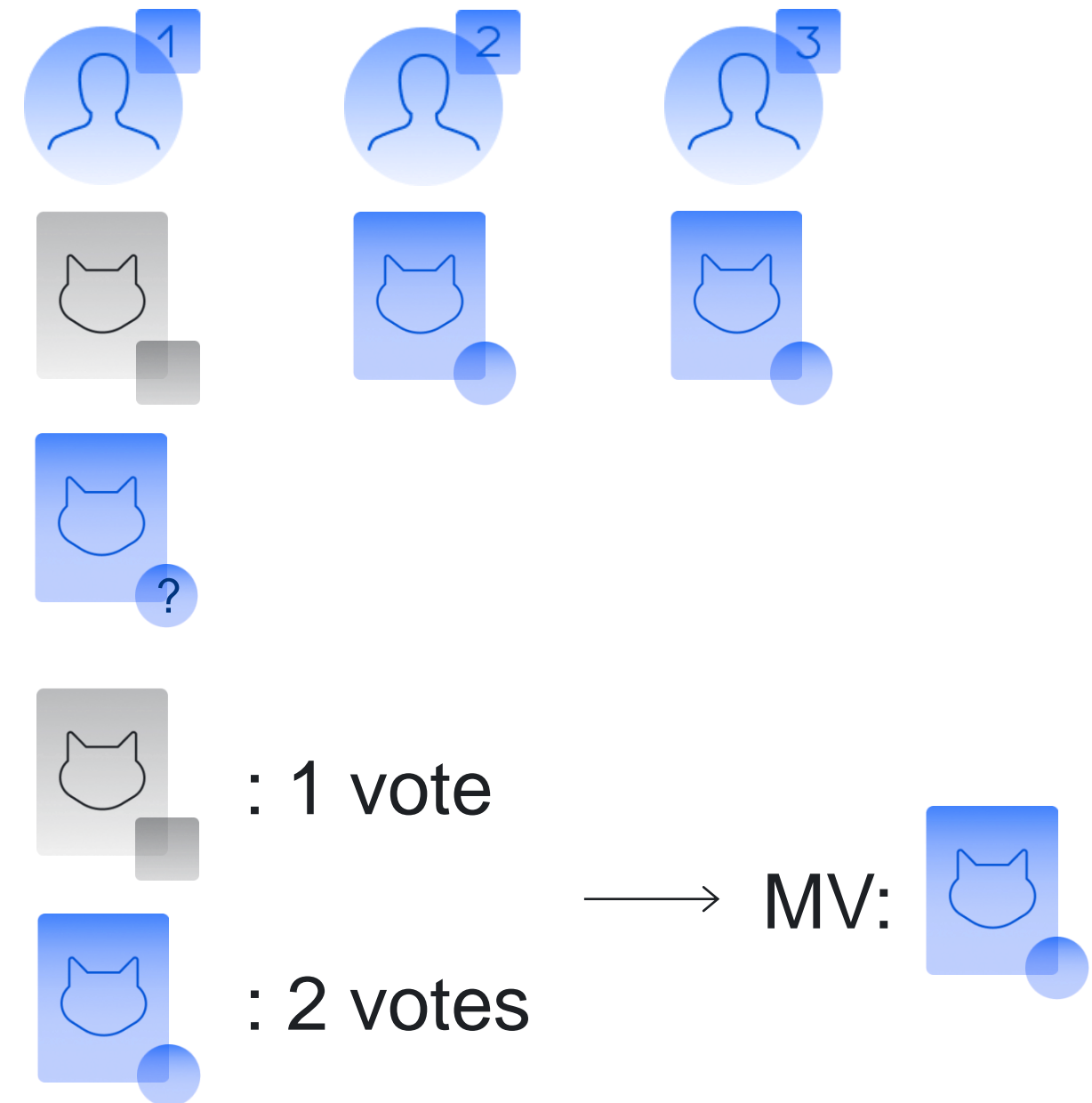
► The problem of aggregation:

- Observe noisy labels

$$y = \{y_j^w \mid j = 1, \dots, J \text{ and } w = 1, \dots, W\}$$

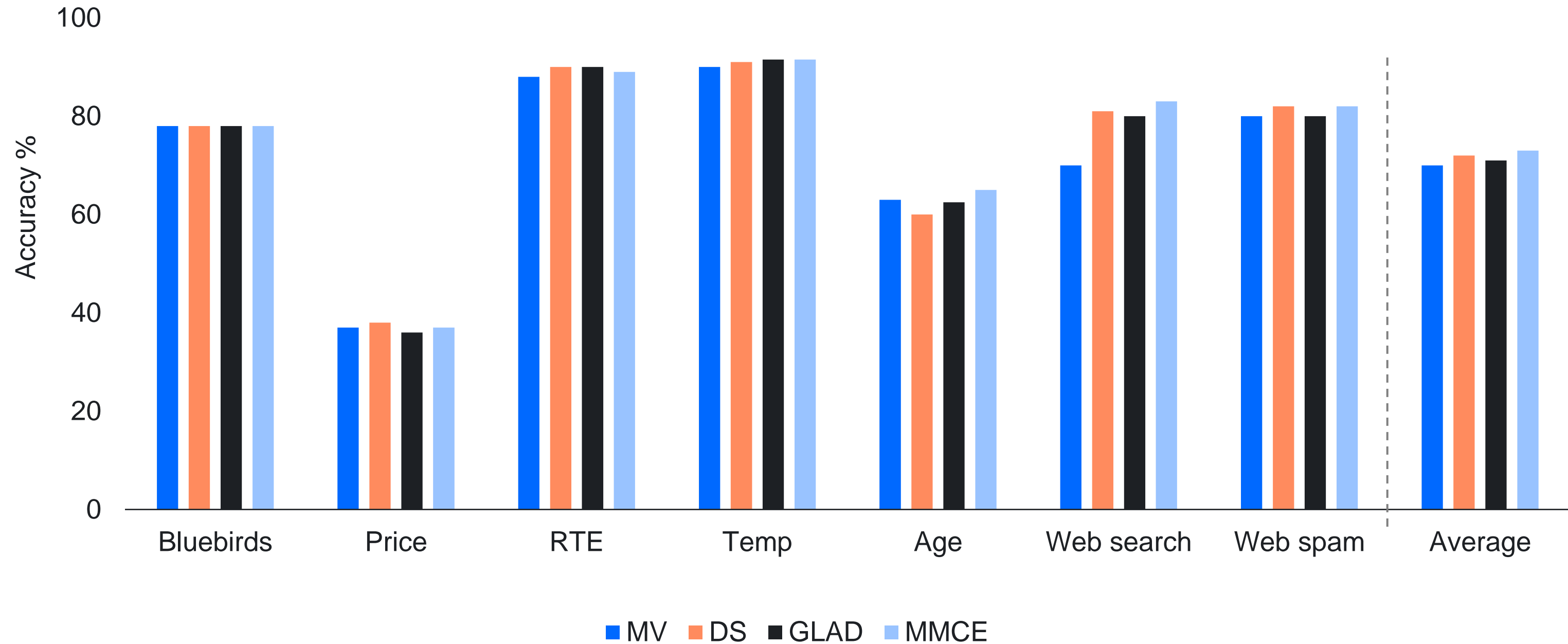
- Recover true labels  $z = \{z_j \mid j = 1, \dots, J\}$

► A straightforward solution:



$$\hat{z}_j^{MV} = \arg \max_{y=1, \dots, K} \sum_{w \in W_j} \delta(y = y_j^w), \text{ where } \delta(A) = 1 \text{ if } A \text{ is true and } 0 \text{ otherwise}$$

# Performance of MV vs other methods



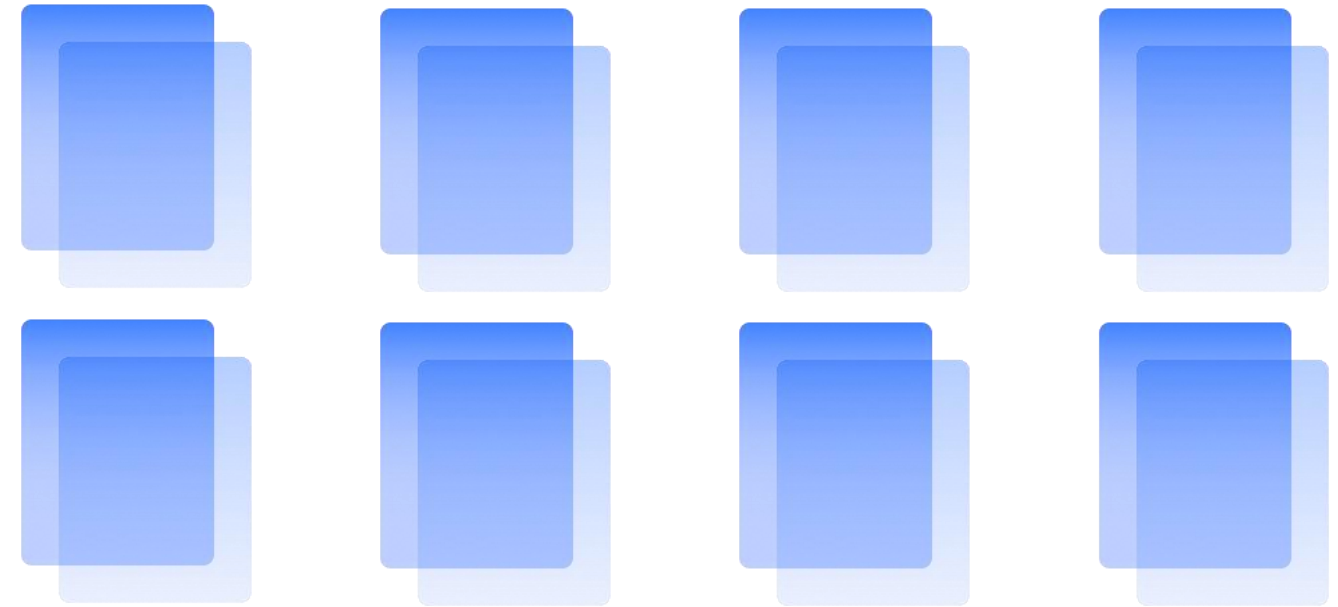


# Properties of MV

All performers are  
treated similarly

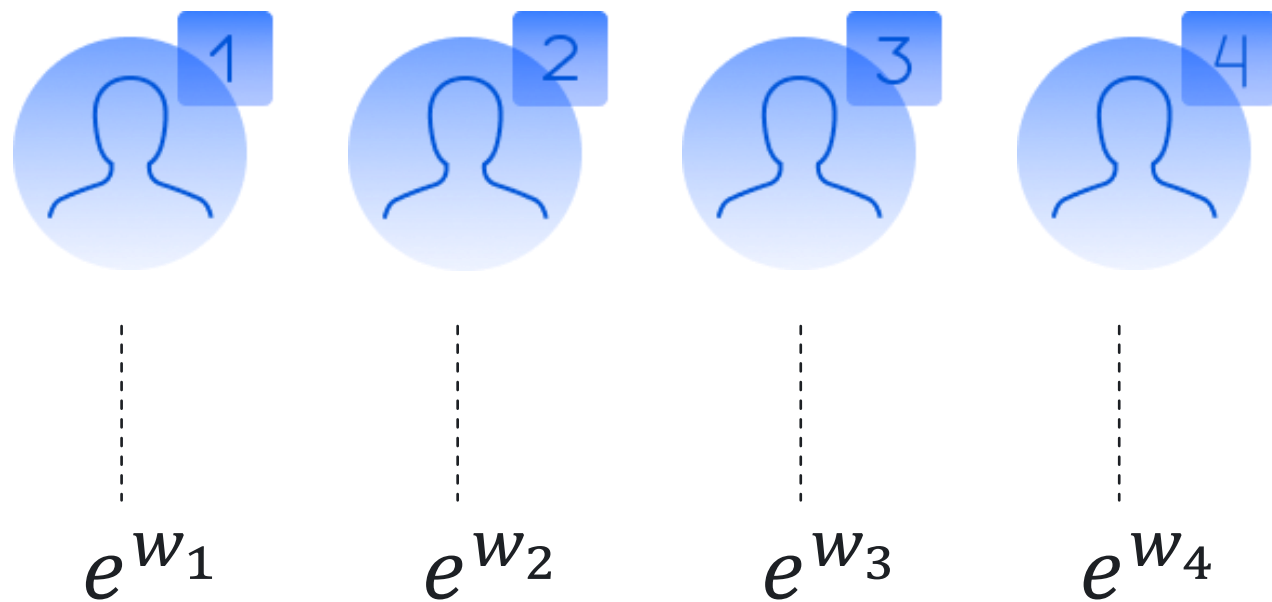


All objects are  
treated similarly

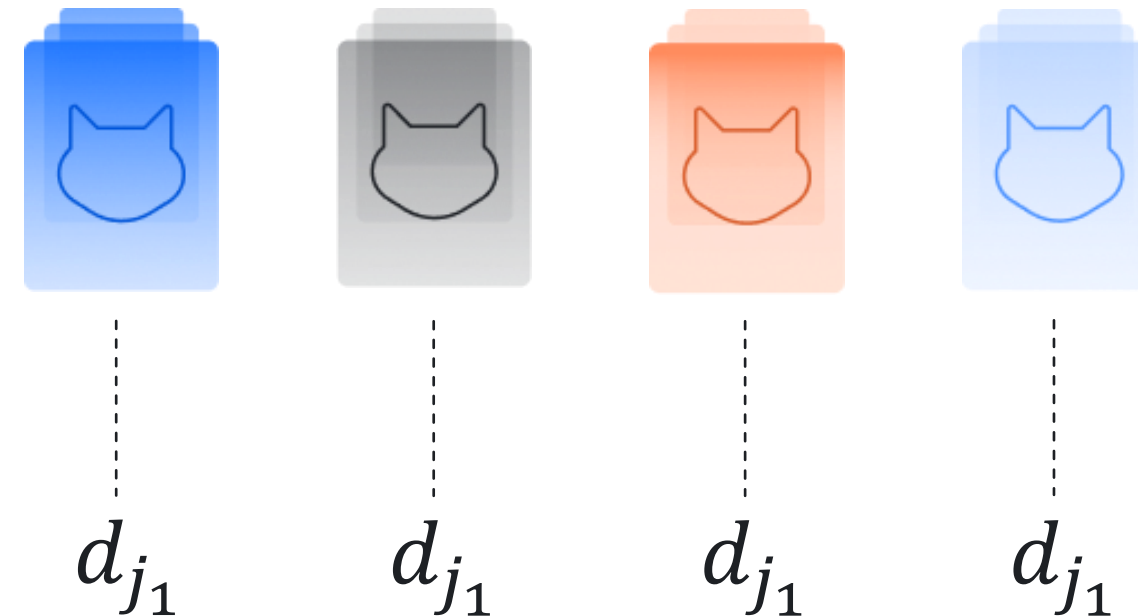


# Advanced aggregation: performers and objects

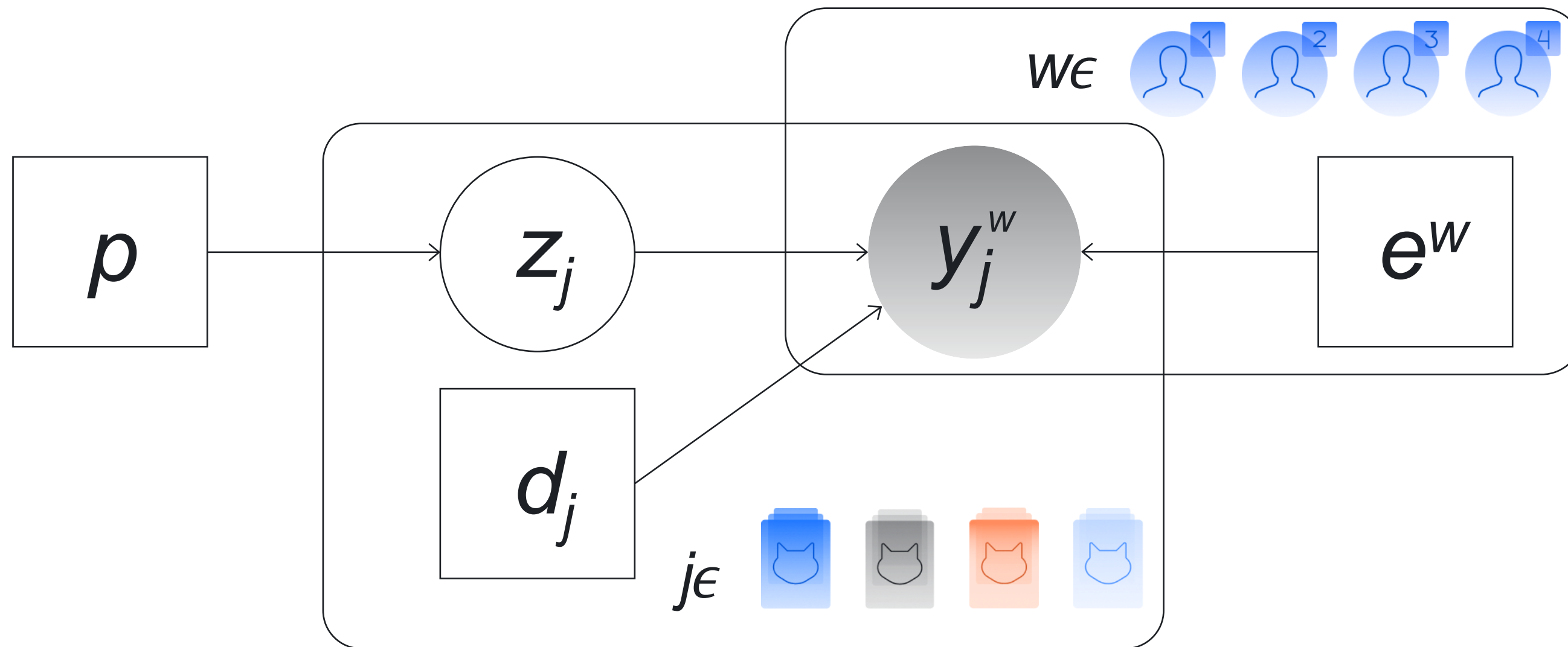
Parameterize expertise  
of performers by  $e^w$



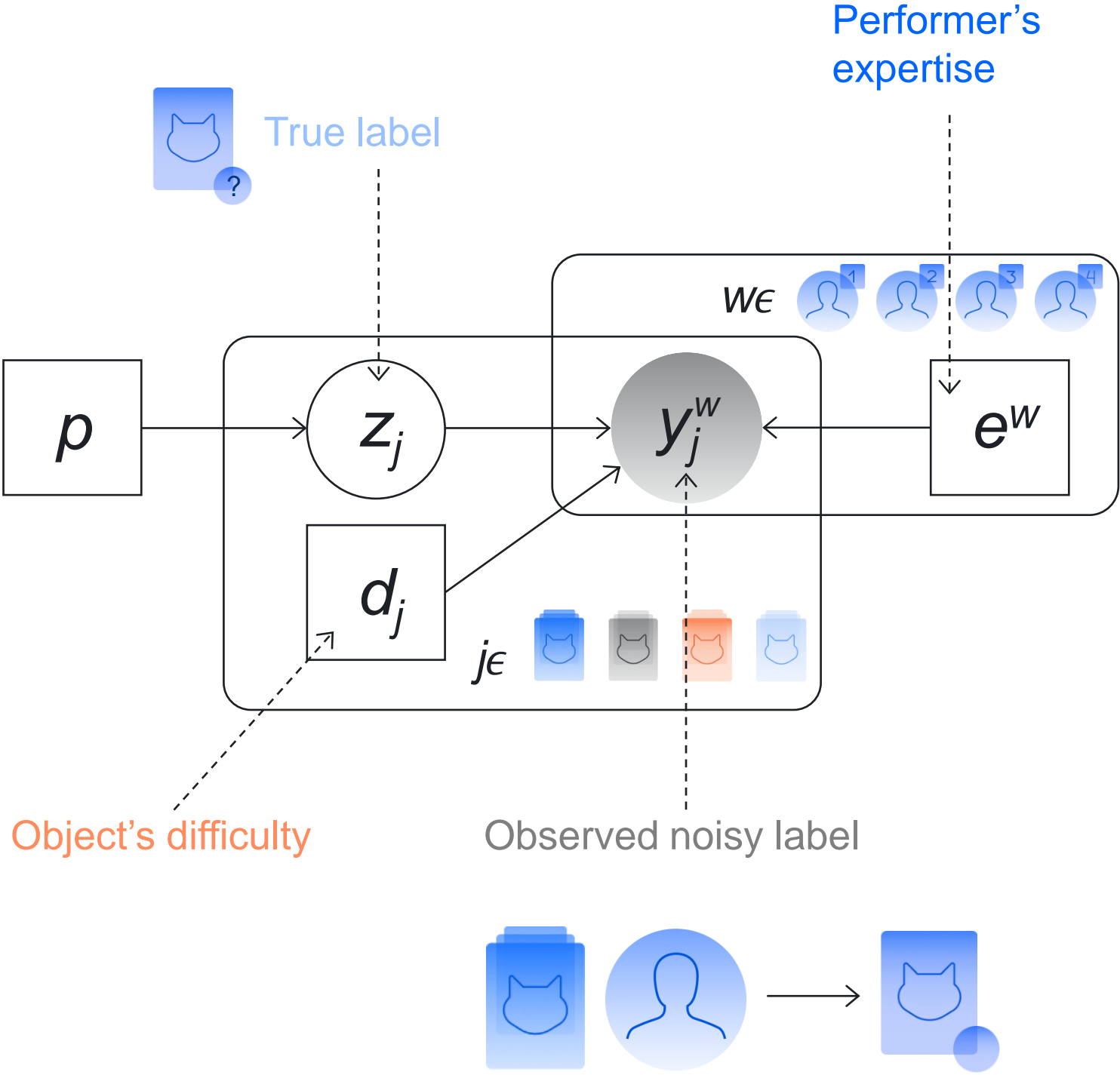
Parameterize difficulty  
of objects by  $d_j$



# Advanced aggregation: latent label models

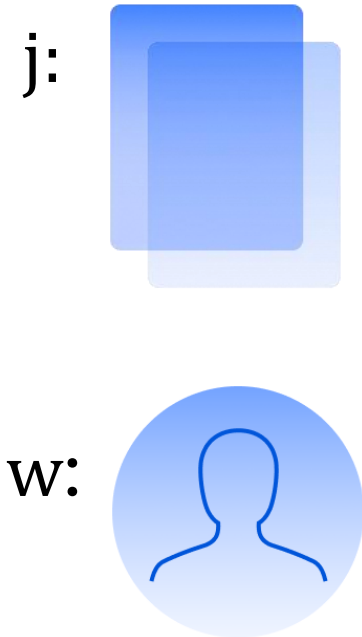


# Latent label models: noisy label model



A noisy label model  $M_j^w = M(e^w, d_j)$  is a matrix of size  $K \times K$  with elements

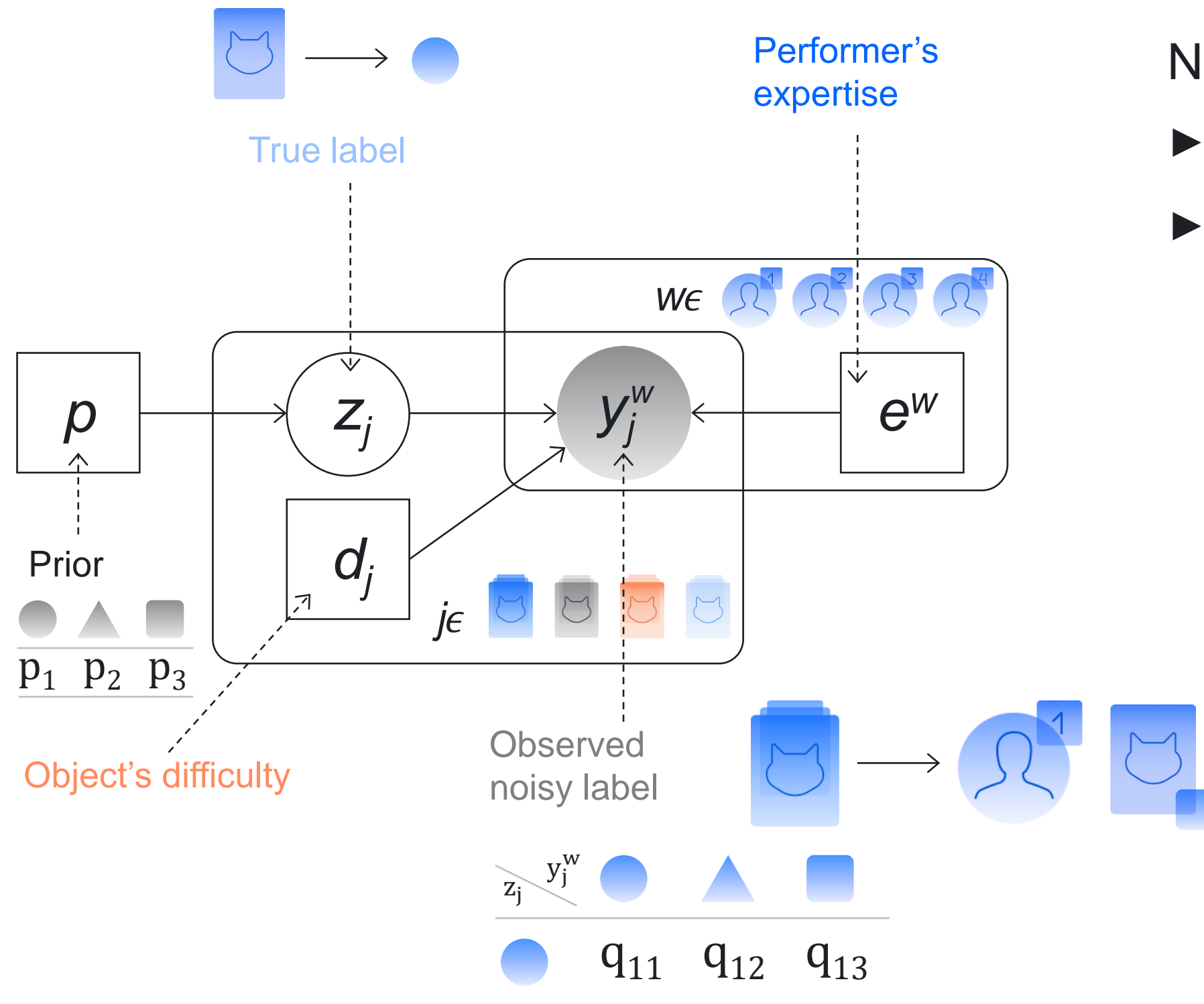
$$M_j^w[c, k] = \Pr(Y_j^w = k | Z_j = c)$$



Noisy True	●	▲	■
●	$q_{11}$	$q_{12}$	$q_{13}$
▲	$q_{21}$	$q_{22}$	$q_{23}$
■	$q_{31}$	$q_{32}$	$q_{33}$

$$q_{c1} + q_{c2} + q_{c3} = 1 \text{ for each } c$$

# Latent label models: generative process



Noisy labels generation:

- Sample  $z_j$  from a distribution  $P_Z(p)$
- Sample  $y_j^w$  from a distribution  $P_Y(M_j^w[z_j, \cdot])$

In multiclassification, a standard choice for  $P_Z(\cdot)$  and  $P_Y(\cdot)$  is a Multinomial distribution  $\text{Mult}(\cdot)$



# Latent label models: parameters optimization

- Assumption:  $y_j^w$  is cond. independent of everything else given  $z_j, d_j, e^w$
- The likelihood of  $y$  and  $z$  under the latent label model:

$$L\left(\{z_j\}_{j=1}^J, p, \{d_j\}_{j=1}^J, \{e^w\}_{w=1}^W\right) = \prod_{j \in J} \sum_{z_j \in \{1, \dots, K\}} \Pr(z_j | p) \prod_{w \in W_j} \Pr(y_j^w | z_j, d_j, e^w)$$

Observed noisy label

Latent true label

Latent parameters

Likelihood of noisy and true labels for object  $j$

- Estimate parameters and true labels by maximizing  $L(\dots)$

# Latent label models: EM algorithm

- Maximization of the expectation of log-likelihood (LL), a lower bound on LL of  $y$  and  $z$

$$\mathbb{E}_z \log \Pr(y, z) = \sum_{j \in J} \sum_{z_j \in \{1, \dots, K\}} \Pr(z_j | p) \log \prod_{w \in W_j} \Pr(z_j | p) \Pr(y_j^w | z_j, d_j, e^w)$$

- **E-step:** Use Bayes' theorem for posterior distribution of  $\hat{z}$  given  $p, d, e$ :

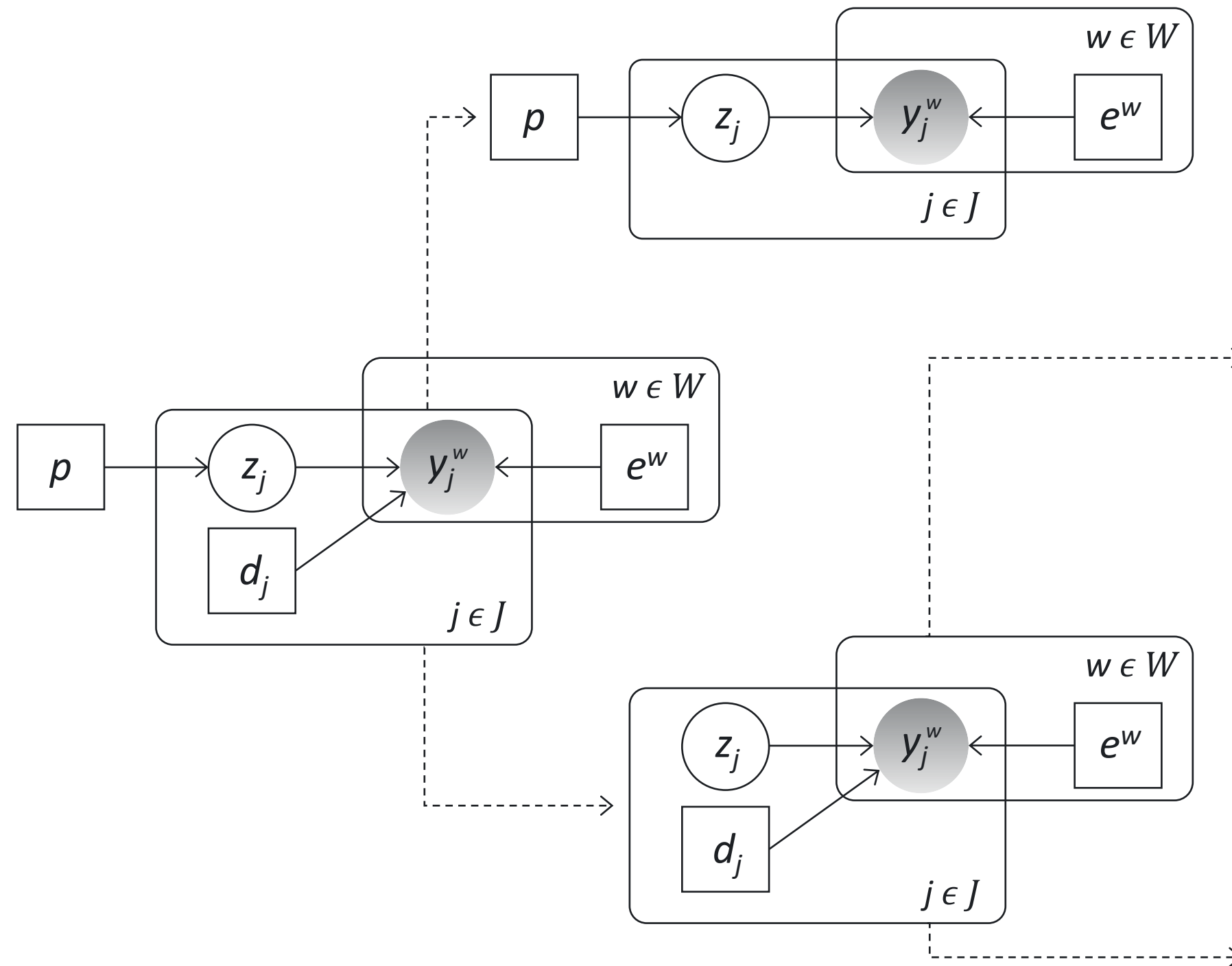
$$\hat{z}_j[c] = \Pr(Z_j = c | y, p, d, e) \propto \Pr(Z_j = c | p) \prod_{w \in W_j} \Pr(y_j^w | Z_j = c, d_j, e^w)$$

- **M-step:** Maximize the expectation of LL with respect to the posterior distribution of  $\hat{z}$ :

$$(p, d, e) = \operatorname{argmax} \mathbb{E}_{\hat{z}} \log \Pr(z_j | p) \prod_{w \in W_j} \Pr(y_j^w | z_j, d_j, e^w)$$

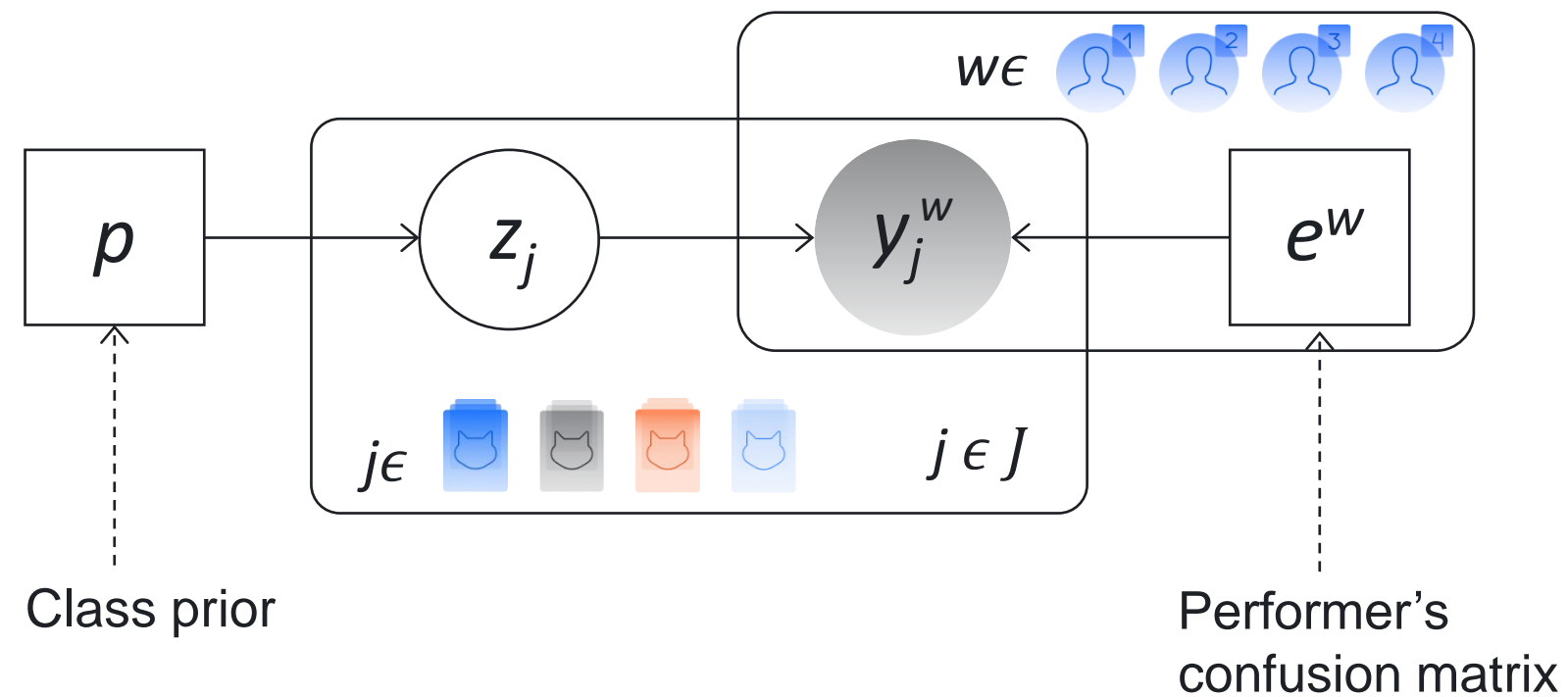
- Analytical solutions
- Gradient descent

# Latent label model (LLM): special cases



- Dawid and Skene model (DS):
  - Categories are **different**
  - Objects are **similar**
  - Performers are **different**
- Generative model of labels, abilities, and difficulties (GLAD):
  - Categories are **similar**
  - Objects are **different**
  - Performers are **different**
- Minimax conditional entropy model (MMCE):
  - Categories are **different**
  - Objects are **different**
  - Performers are **different**

# Dawid and Skene model (DS)



LLM with parameters:

- ▶  $p$  — vector of length  $K$ :  $p[i] = \Pr(Z = c)$
- ▶  $e^w$  — matrix of size  $K \times K$ :  $e^w[c, k] = \Pr(Y^w = k | Z = c)$

$z \backslash y_w$	●	▲	■
●	■	■	■
▲	■	■	■
■	■	■	■

# DS: parameters optimization

► **E-step:**

$$\hat{z}_j[c] = \frac{p[c] \prod_{w \in W_j} e^w[c, y_j^w]}{\sum_k p[k] \prod_{w \in W_j} e^w[k, y_j^w]}, \quad c = 1, \dots, K$$

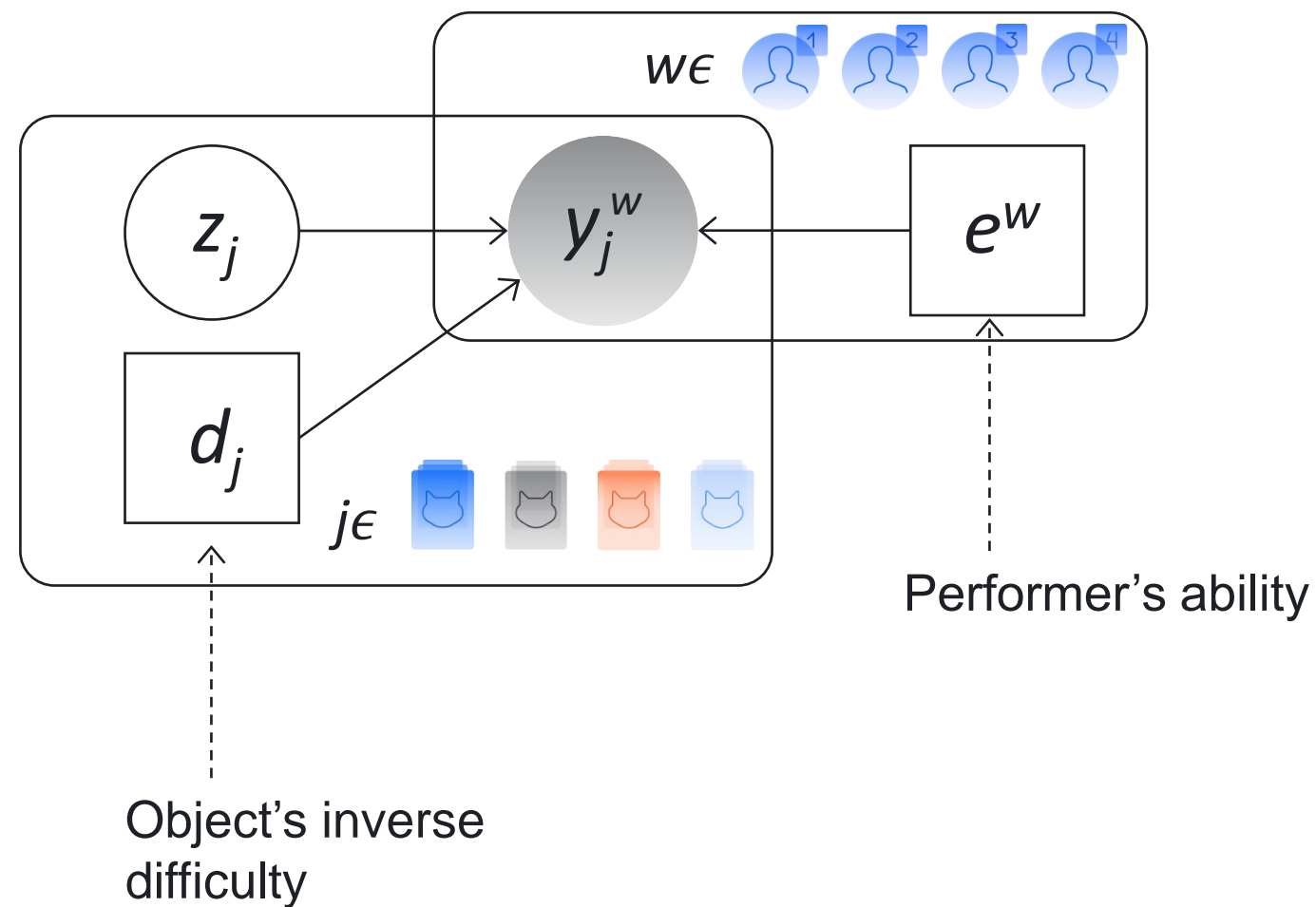
► **M-step:** Analytical solution

$$e^w[c, k] = \frac{\sum_{j \in J} \hat{z}_j[c] \delta(y_j^w = k)}{\sum_{q=1}^K \sum_{j \in J} \hat{z}_j[c] \delta(y_j^w = q)}, \quad k, c = 1, \dots, K$$

$$p[c] = \frac{\sum_{j \in J} \hat{z}_j[c]}{J}, \quad c = 1, \dots, K$$



# Generative model of Labels, Abilities, and Difficulties (GLAD)



LLM with parameters:

- Scalar  $d_j \in (0, \infty)$
- Scalar  $e^w \in (-\infty, \infty)$
- Model:

$$\Pr(Y_j^w = k | Z_j = c) = \begin{cases} a(w, j), & c = k \\ \frac{1 - a(w, j)}{K - 1}, & c \neq k \end{cases}$$

$$\text{where } a(w, j) = \frac{1}{1 + \exp(-e^w d_j)}$$

# GLAD: parameters optimization

► Let  $a(w, j) = \frac{1}{1 + \exp(-\mathbf{e}^w \mathbf{d}_j)}$  and  $P(z_j)$  be a predefined prior (e.g.,  $P(z_j) = 1/K$ )

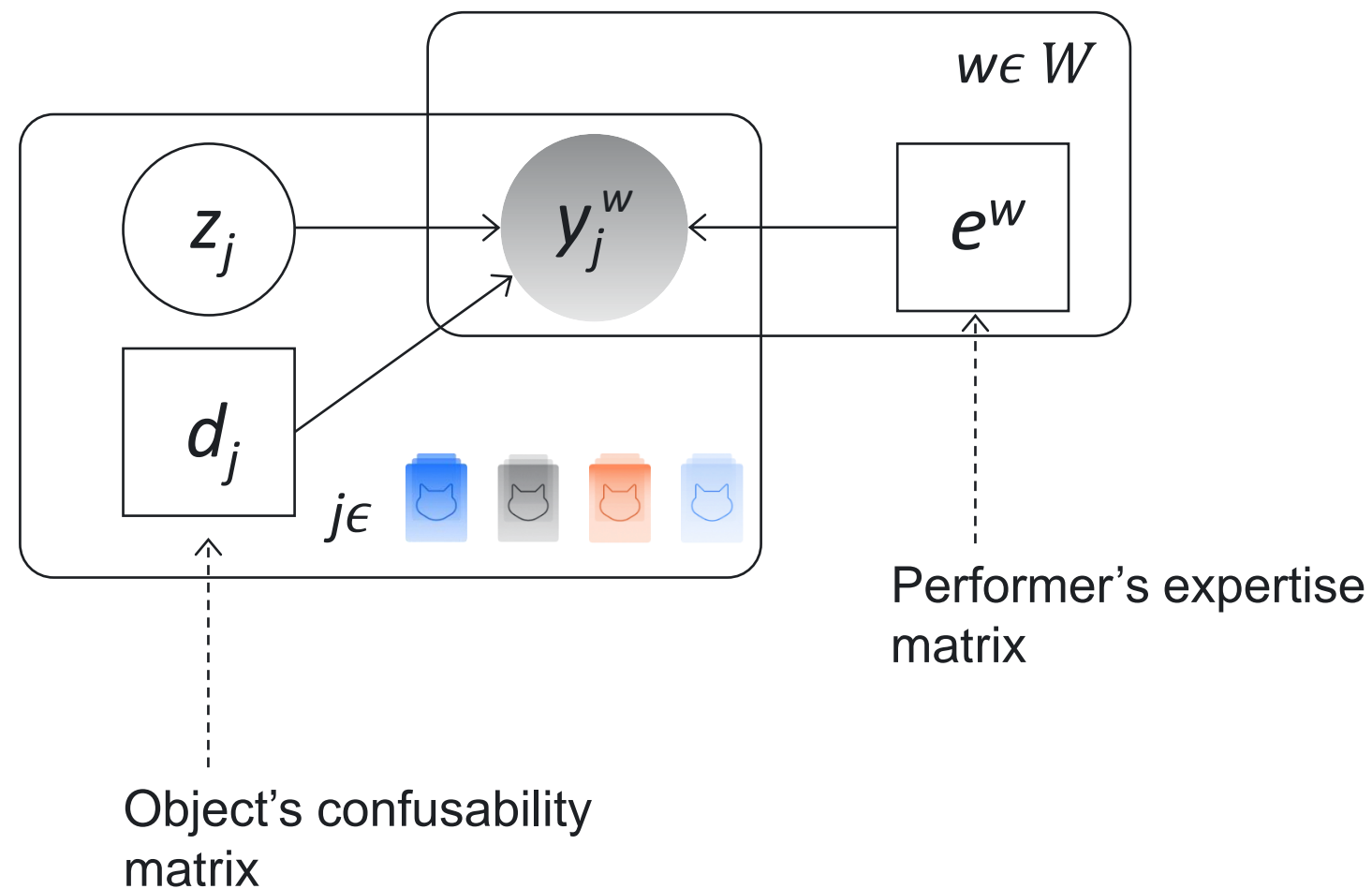
► **E-step:**

$$\hat{z}_j [c] \propto P(Z_j = c) \prod_{w \in W_j} a(w, j)^{\delta(y_j^w = c)} \left( \frac{1 - a(w, j)}{K - 1} \right)^{\delta(y_j^w \neq c)}, \quad c = 1, \dots, K$$

► **M-step:** estimate  $(\mathbf{d}, \mathbf{e})$  for given  $\hat{z}$  using gradient descent

$$(\mathbf{d}^t, \mathbf{e}^t) = \operatorname{argmax} \sum_{j \in J} \left[ \mathbb{E}_{\hat{z}_j} \log P(z_j) + \sum_{w \in W_j} \mathbb{E}_{\hat{z}_j} \log \Pr(y_j^w | z_j) \right]$$

# MiniMax Conditional Entropy model (MMCE)















- Find parameters that minimize the maximum conditional entropy of observed labels:

$$\min_Q \max_P - \sum_{\substack{j \in J \\ c \in \{1, \dots, K\}}} Q(Z_j = c) \sum_{\substack{w \in W \\ k \in \{1, \dots, K\}}} P(Y_j^w = k | Z_j = c) \log P(Y_j^w = k | Z_j = c)$$

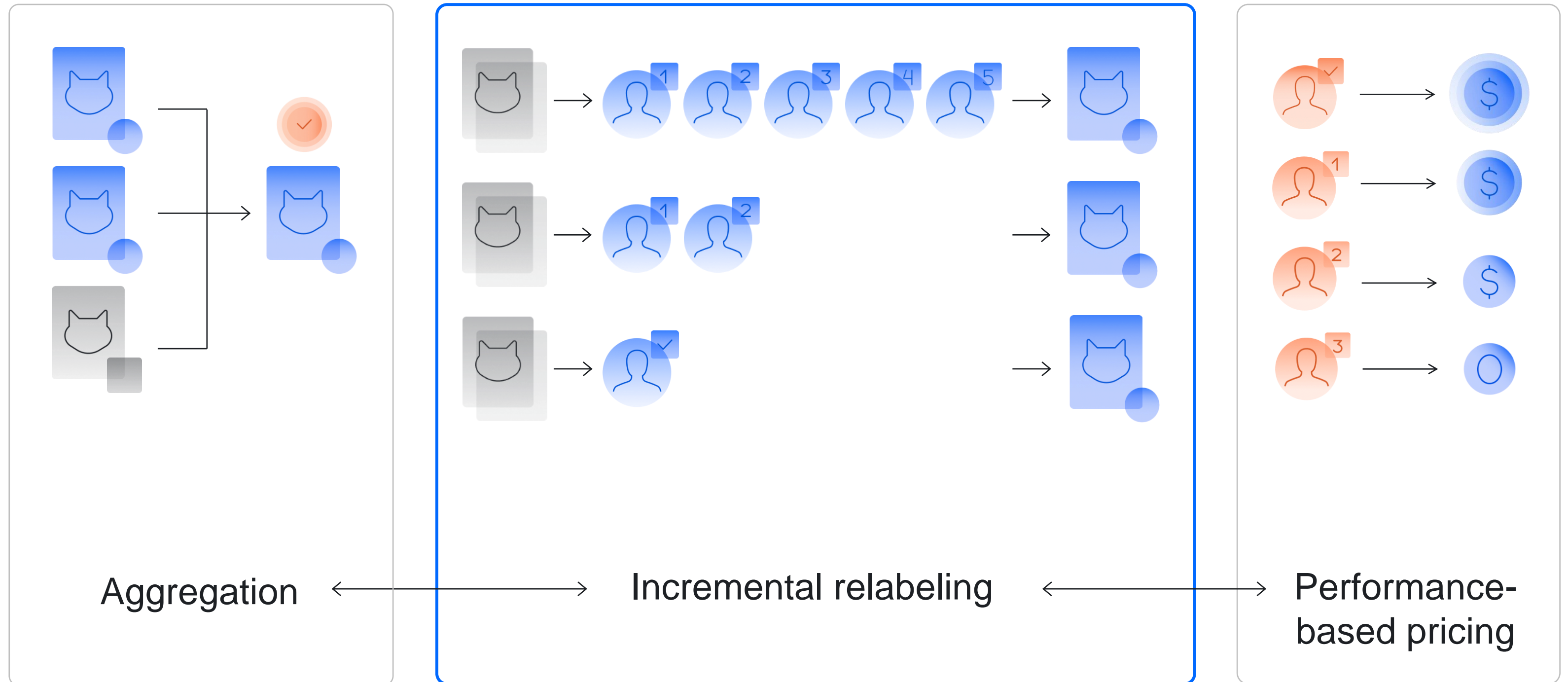
- LLM with parameters:
  - $d_j$  — matrix of size  $K \times K$
  - $e^w$  — matrix of size  $K \times K$
  - Noisy label model:

$$\Pr(Y_j^w = k | Z_j = c) = \exp(d_j[c, k] + e^w[c, k])$$

# Summary of aggregation methods

	MV	DS	GLAD	MMCE
Categories (K)				
Objects (J)				
Performers (W)				
Number of parameters	0	$WK^2 + K$	$W + J$	$(W + J)K^2$

# Key components of labeling with crowds





**Incremental relabeling**  
aka dynamic overlap

# Pool settings: dynamic overlap

### Quality control

Add rules to get more accurate responses.  
All rules work independently.

NON-AUTOMATIC ACCEPTANCE ?

☐ No

REVIEW PERIOD IN DAYS

CAPTCHA FREQUENCY ?

None

+

Add Quality Control Rule

### Overlap

Specify how many performers you want to complete each task in the pool.

OVERLAP ?

DYNAMIC OVERLAP ?

☐ Off

### Speed/quality ratio

Specify additional conditions for selecting performers by their rating in Toloka.  
This will improve quality, but may reduce the speed of task completion because there will be fewer performers available for completing tasks. [Learn more](#)

Top %

Online

Time

Specify the percentage of top-rated active users who can access tasks in the pool.

# Incremental relabeling problem

Obtain aggregated labels of a desired quality level using a fewer number of noisy labels



# Incremental relabeling scheme (IRL)

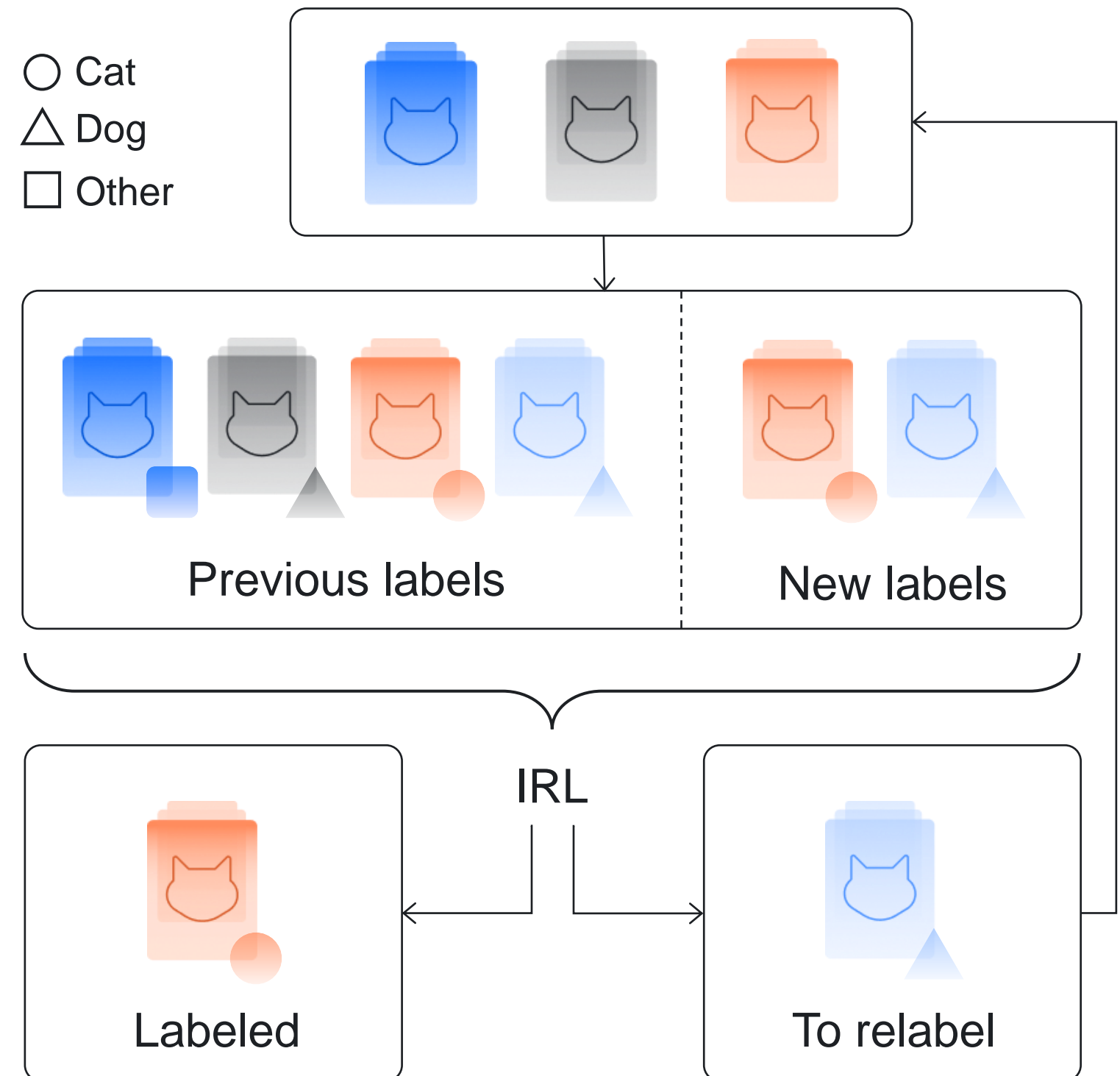
Request a label for each object

In real time IRL algorithm receives:  
(1) previously accumulated labels  
(2) new labels

Decides:

(1) which objects are labeled  
(2) which objects to relabel

Repeat until all tasks are labeled



# Notations

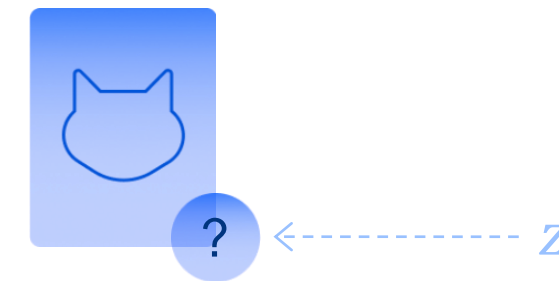
- Consider one object



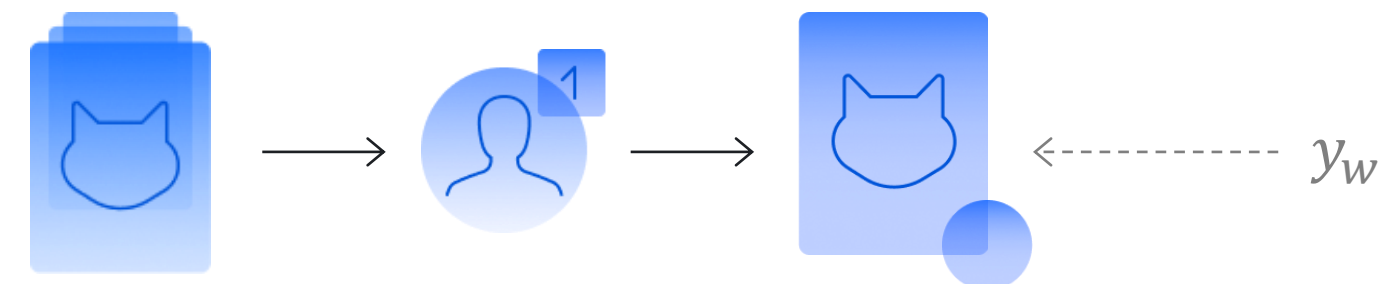
Classify images:

- Cat
- △ Dog
- Other

- $z \in \{1, \dots, K\}$  — latent true label



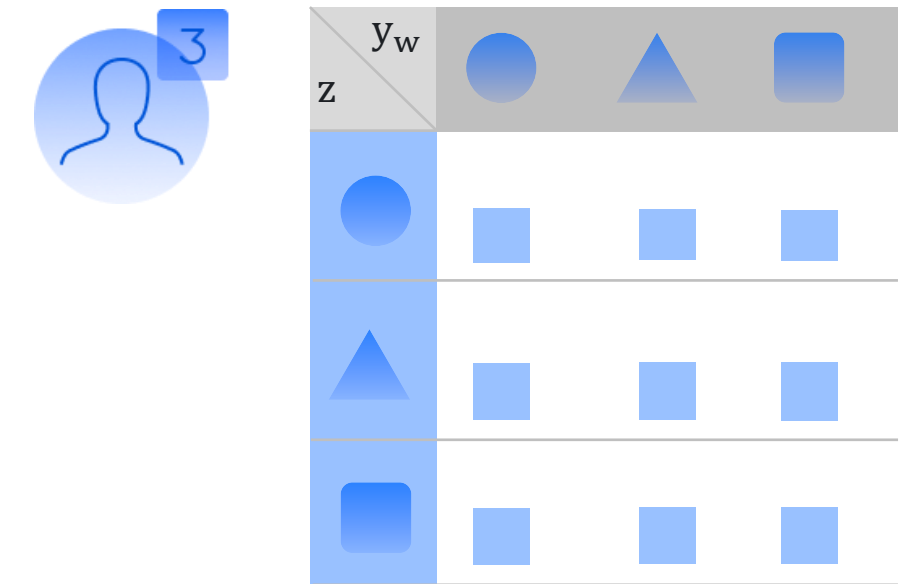
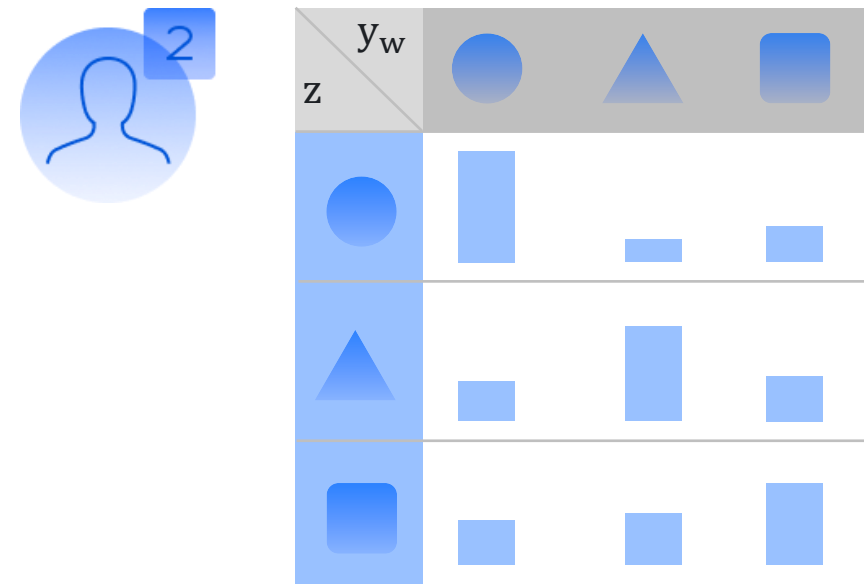
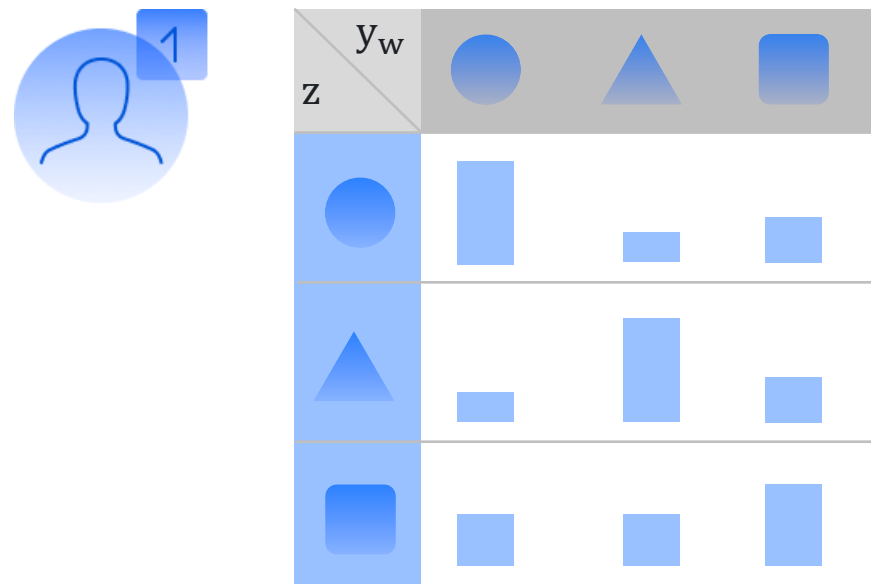
- $y_w \in \{1, \dots, K\}$  — observed noisy label from performer w:



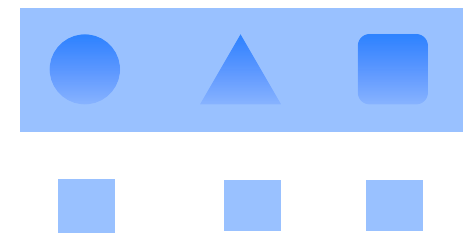
# Notations

- Noisy label model for performer  $w$ :

$$M_w \in [0,1]^{K \times K}: \Pr(Y_w = k | Z = c) = M_w[c, k]$$



- Prior distribution:  $\Pr(Z = k) = p_k$

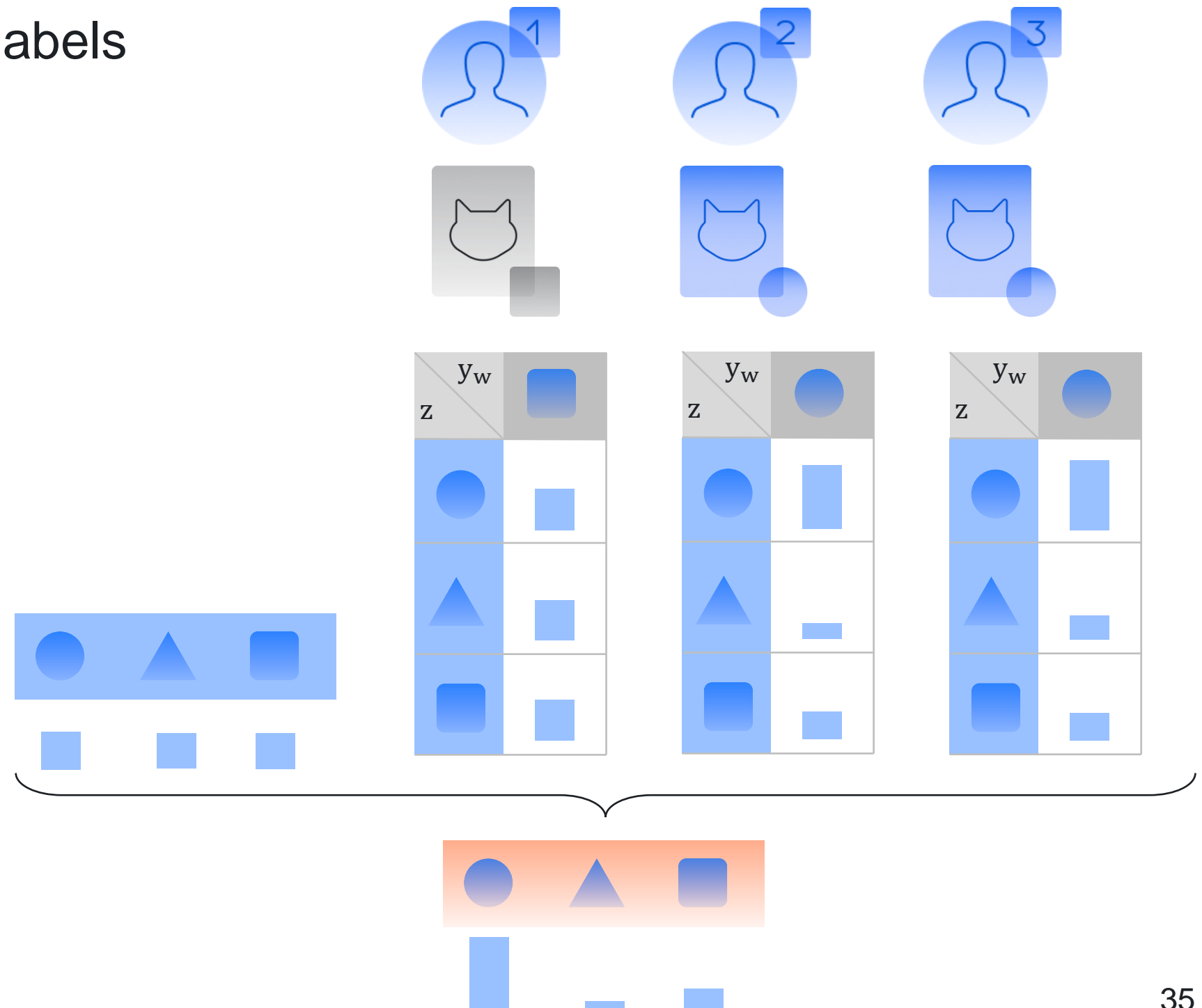




# Posterior distribution

- $\{y_{w_1}, \dots, y_{w_n}\}$  — accumulated noisy labels for the object
- Using Bayes rule:

$$\begin{aligned} & \Pr(Z = k | \{y_{w_1}, \dots, y_{w_n}\}) \\ &= \frac{\Pr(Z = k) \Pr(\{y_{w_1}, \dots, y_{w_n}\} | Z = k)}{\Pr(\{y_{w_1}, \dots, y_{w_n}\})} \\ &= \frac{p_k \prod_{i=1}^n M_{w_i}[k, y_{w_i}]}{\sum_{t=1}^K p_t \prod_{i=1}^n M_{w_i}[t, y_{w_i}]} \end{aligned}$$





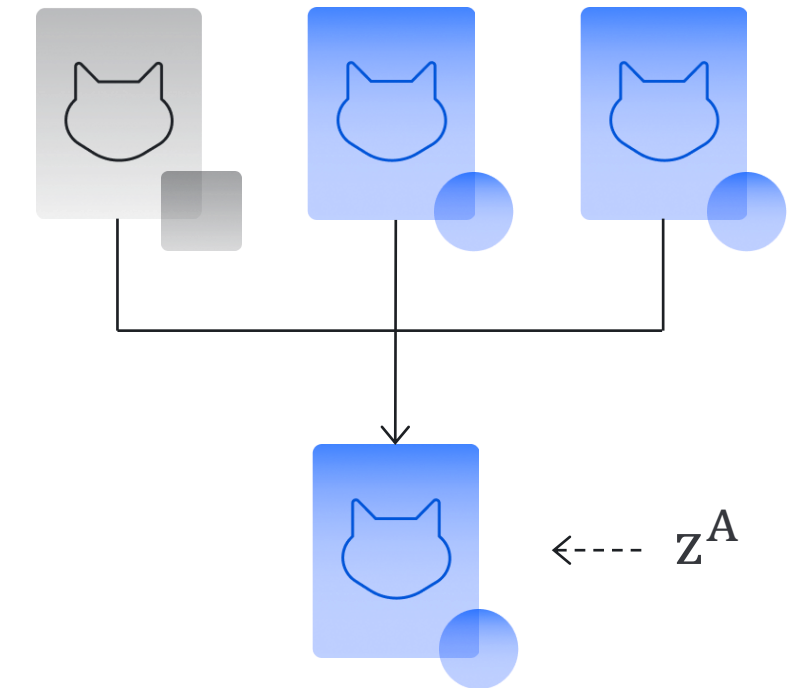
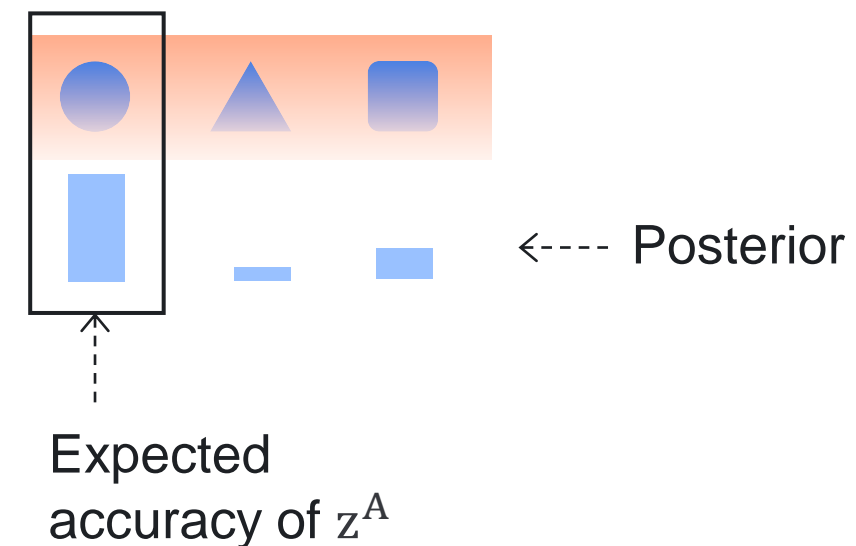
# Expected accuracy of aggregated labels

- ▶ Let  $A$  be an aggregation model, e.g. MV, DS, GLAD,...
- ▶ Denote aggregated label  $z^A = A(\{y_{w_1}, \dots, y_{w_n}\})$
- ▶ Expected accuracy of aggregated labels given noisy labels is

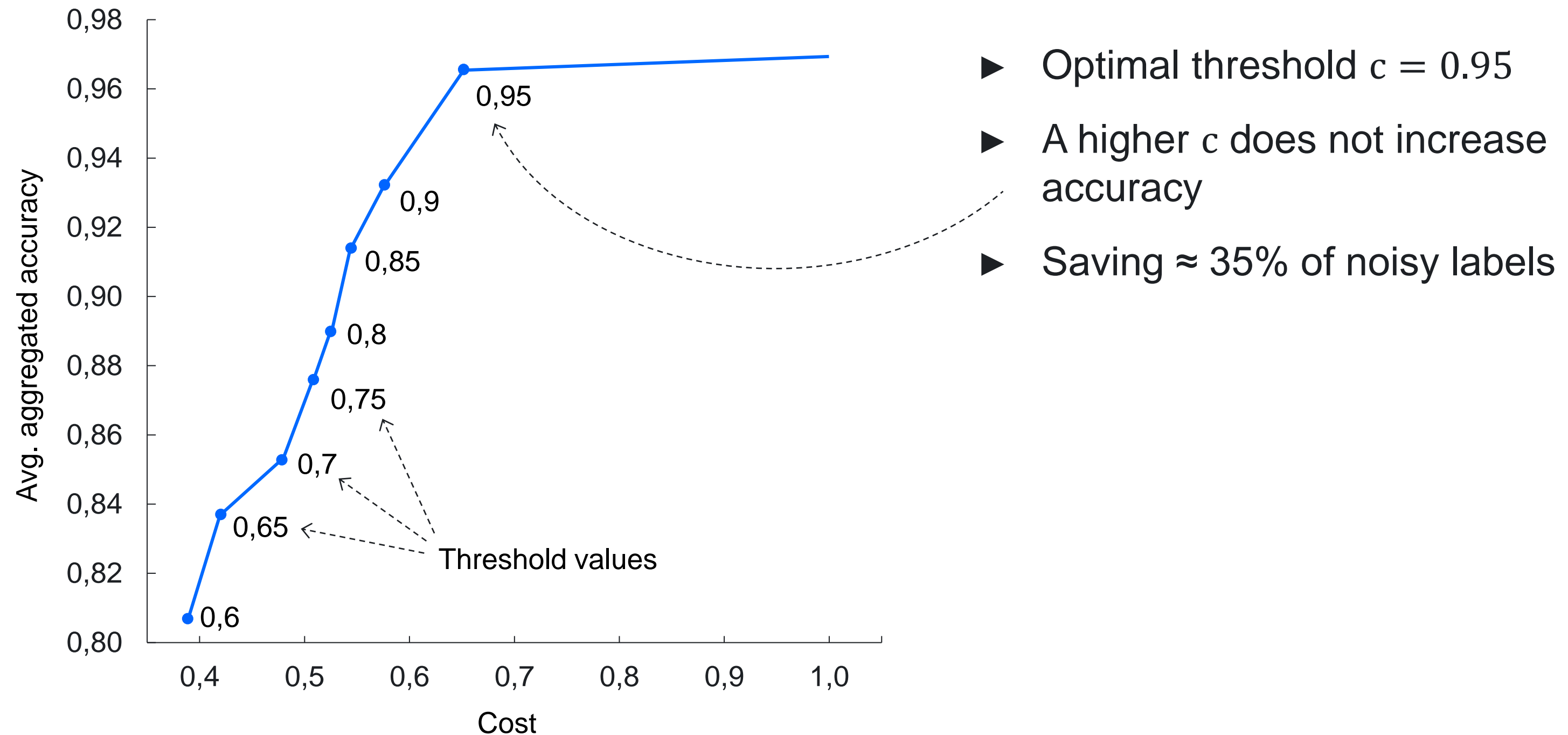
$$E(\delta(z = z^A) | \{y_{w_1}, \dots, y_{w_n}\}) = \Pr(z = z^A | \{y_{w_1}, \dots, y_{w_n}\})$$

- ▶ Stop labeling if  $E(\delta(z = z^A) | \{y_{w_1}, \dots, y_{w_n}\}) \geq C$

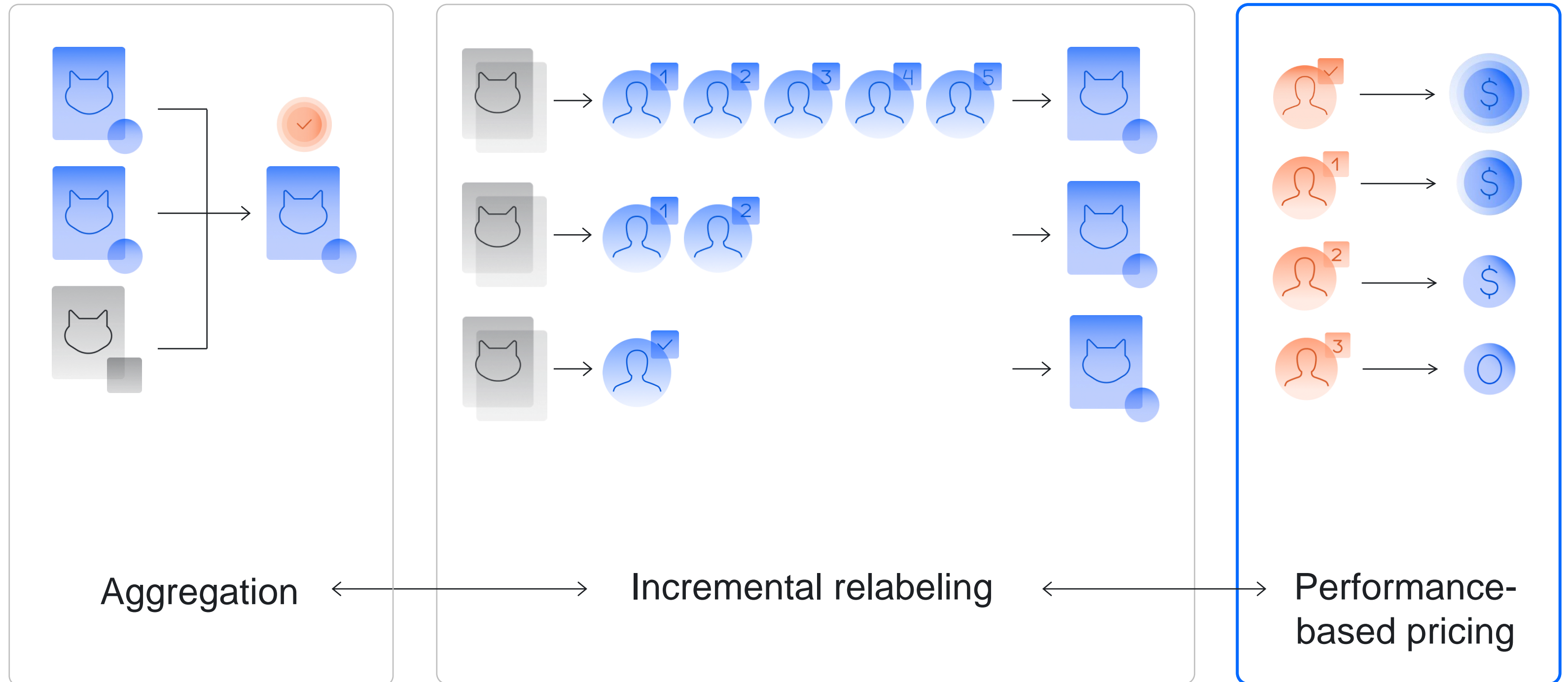
Parameter



# Threshold in IRL: cost – accuracy trade-off



# Key components of labeling with crowds



**Performance-based  
pricing**  
aka dynamic pricing

# Pool settings: dynamic pricing

POOL NAME (VISIBLE ONLY TO YOU) ?

Are there traffic lights in the picture? ✕

☒ Use project description

PUBLIC DESCRIPTION ?

Add a private description

**Price per task suite**

You can add one or more tasks to the page. Enter the total price for all tasks on the page.

PRICE IN US DOLLARS ?

0.07

FEE ?

+ Dynamic pricing

**Performers**

[Copy settings from...](#)

Filter performers who can access the task.  
Toloka has users from different countries, so don't forget to filter by language and region. [Learn more](#)

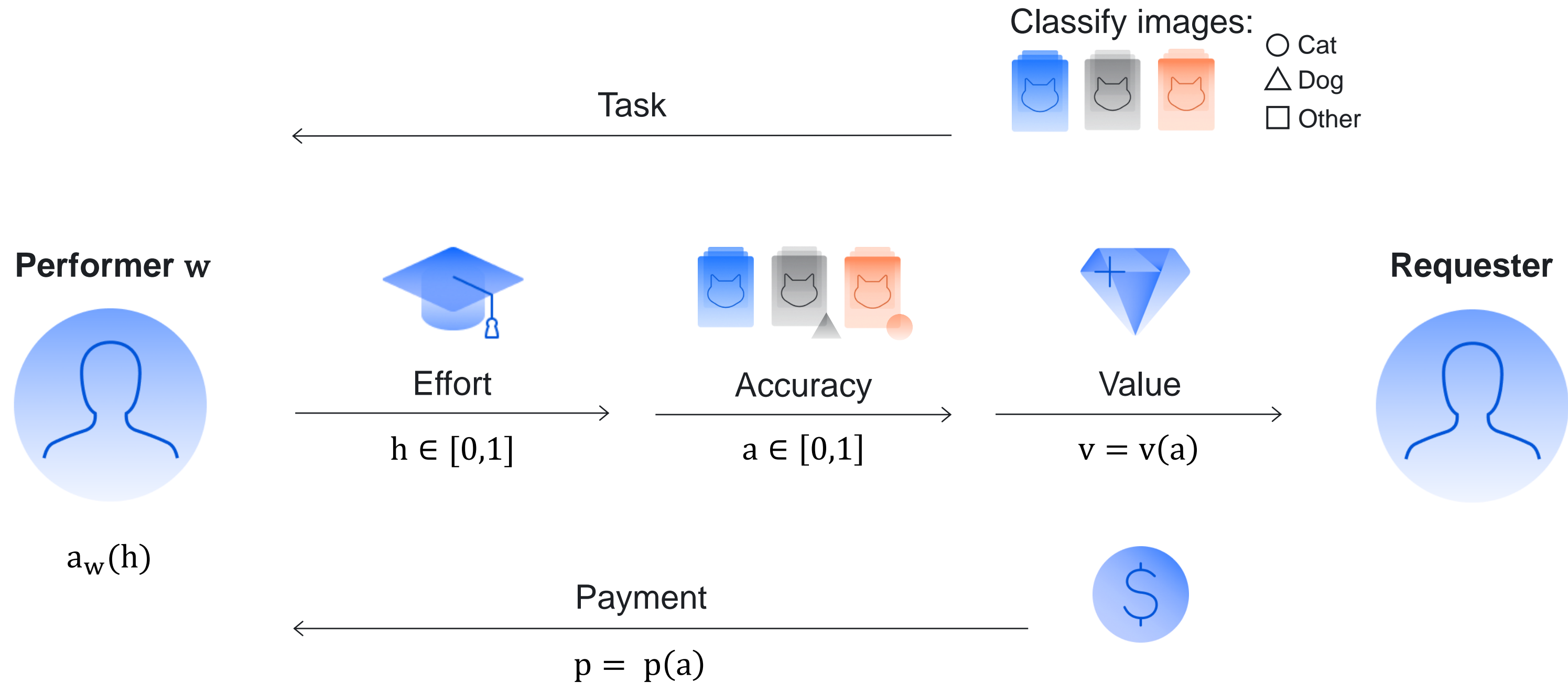
ADULT CONTENT ?

Yes ☐

Add filter ▼

Create skill

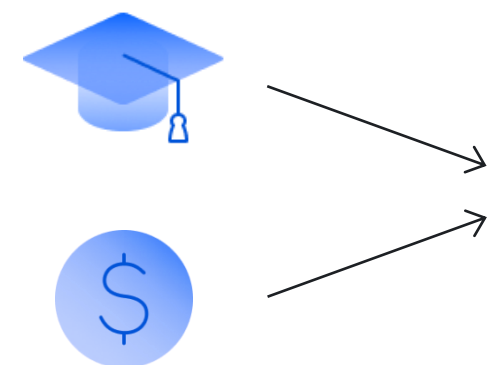
# Labeling as a game: notation



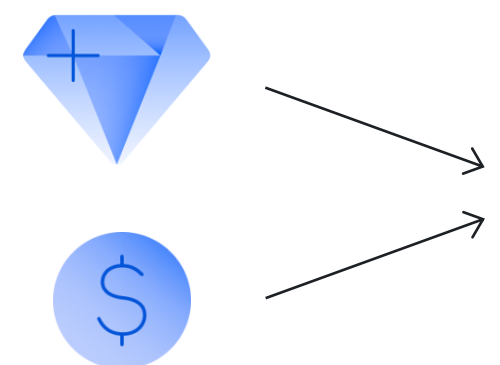


# Labeling as a game: formalization

- Each performer  $w$  chooses a level of effort  $h$  for labeling object to maximize earnings per unit of spent effort:


$$\frac{p(a_w(h))}{h} \rightarrow \max_{h \geq 0}$$

- The requester chooses a pricing  $p(a)$  to minimize payments per unit of obtained value



$$\frac{v(a)}{p(a)} \rightarrow \max_{a \in [0,1]}$$



# Labeling as a game: incentive compatible pricing

- Assume  $a_w(h)$  is a linear function of  $h$ :

$$a_w(h) = c_1 h + c_0$$

Accuracy 

The requester and the performers maximize their utility simultaneously if the pricing  $p(a)$  for each label is proportional to its accuracy  $a$

# Performance-based pricing in practice: settings

- Price  $p$  for the level of accuracy  $a_0$  :  $\Pr(\hat{z} = z) \geq a_0$  E.g.:



- $\hat{q}_w = \Pr(y^w = z)$  — estimated quality level of performer  $w$ , e.g. the fraction of correct labels for golden set (GS):



5 correct GS  
among 10  
 $\hat{q}_w = 0.5$



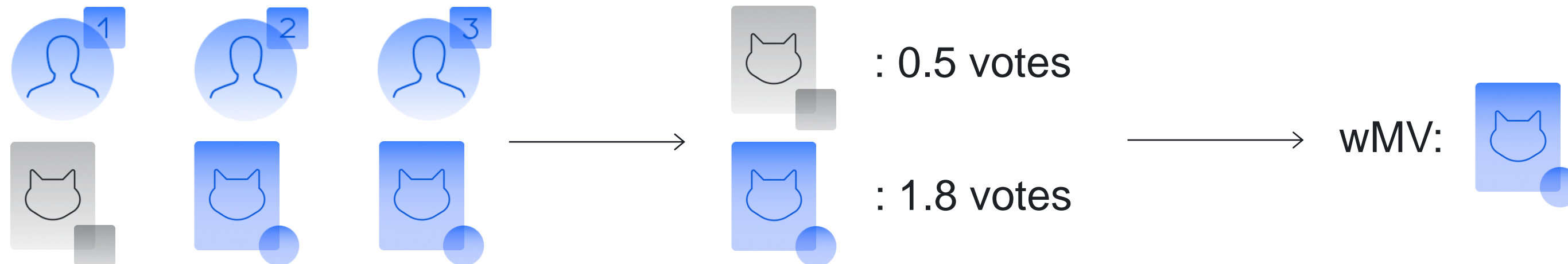
16 correct GS  
among 20  
 $\hat{q}_w = 0.8$



100 correct GS  
among 100  
 $\hat{q}_w = 1$

# Performance-based pricing in practice: settings

- Aggregation  $\hat{z}_j^{wMV} = \arg \max_{y=1,\dots,K} \sum_{w \in W_j} \hat{q}_w \delta(y = y_j^w)$



- IRL algorithm is based on the expected accuracy of  $\hat{z}_j^{wMV}$

# Performance-based pricing in practice

## ► Pricing rules

1. If  $\hat{q}_w \geq a_0$ , then the price is  $p$
2. Else find  $n$ :

$$\sum_{k=0}^{n/2} \binom{n}{k} \hat{q}_w^{n-k} (1 - \hat{q}_w)^k \geq a_0$$

Expected accuracy for MV

The price is  $p/n$

$$a_0 = 0.99$$



$$\hat{q}_w = 1$$



0.3\$



$$\hat{q}_w = 0.8$$

$$\Rightarrow n = 15$$



0.02\$



$$\hat{q}_w = 0.5$$

$$\Rightarrow n = \infty$$



0\$

# Key components of labeling with crowds

