Part V

Theory on efficient aggregation, incremental relabeling, and pricing

Valentina Fedorova,
Researcher

Project 1: Filter images

Does the image contain traffic signs?

Yes

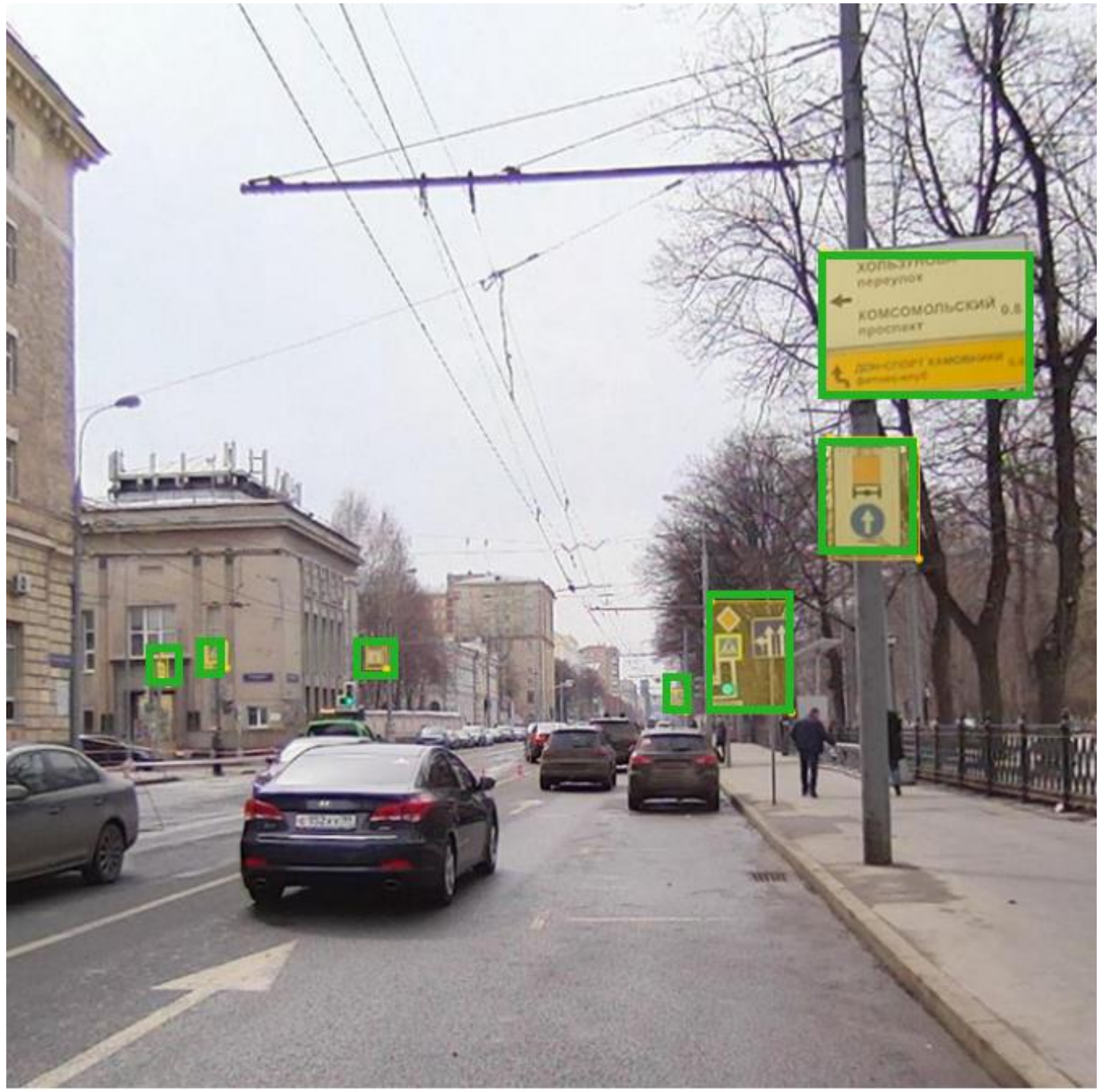
No

Project 3: Verification

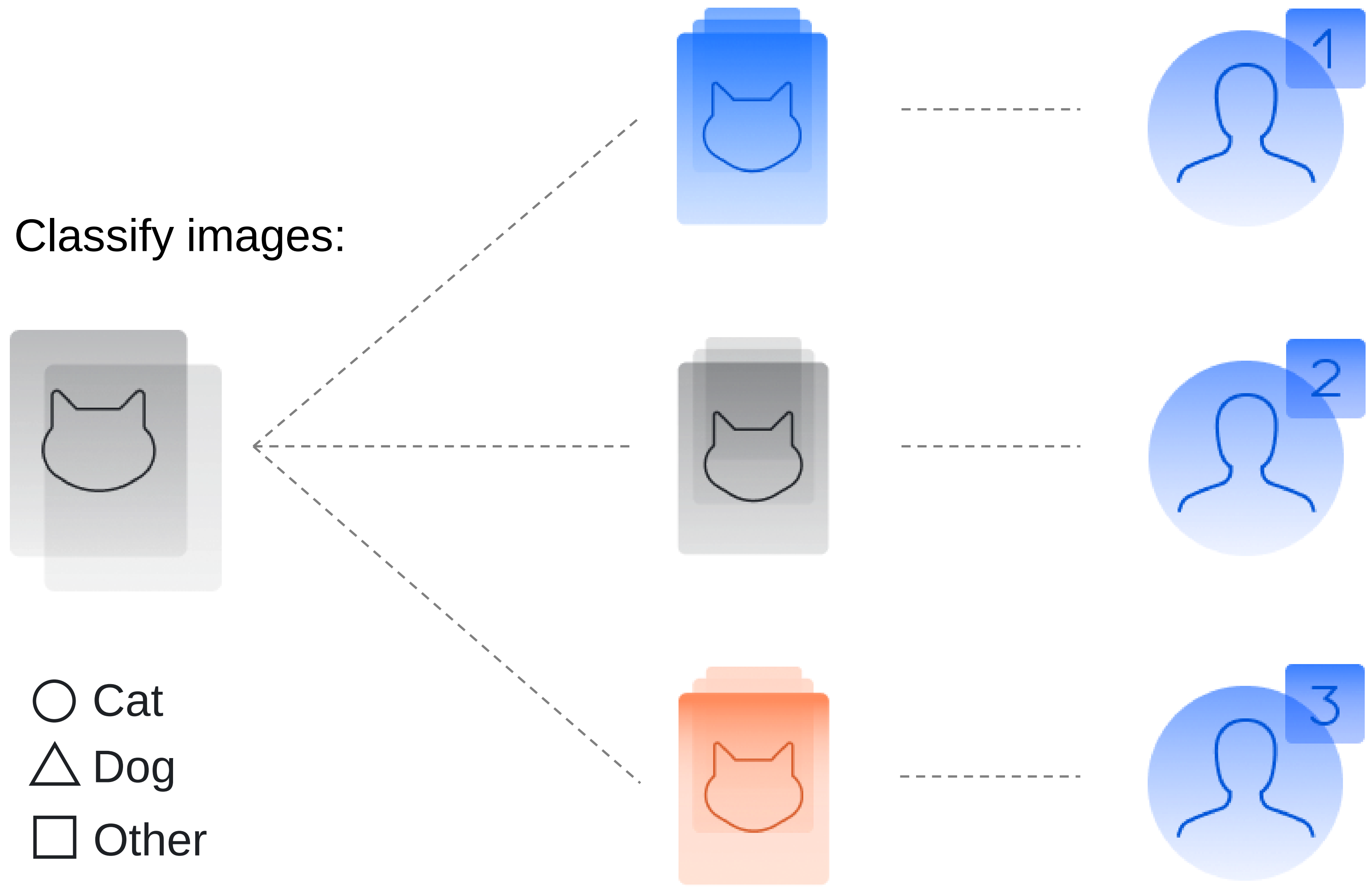
Are the bounding boxes correct?

Yes

No



Labeling data with crowdsourcing



- How to choose a reliable label?
- How many workers per object?
- How much to pay to workers?
- …

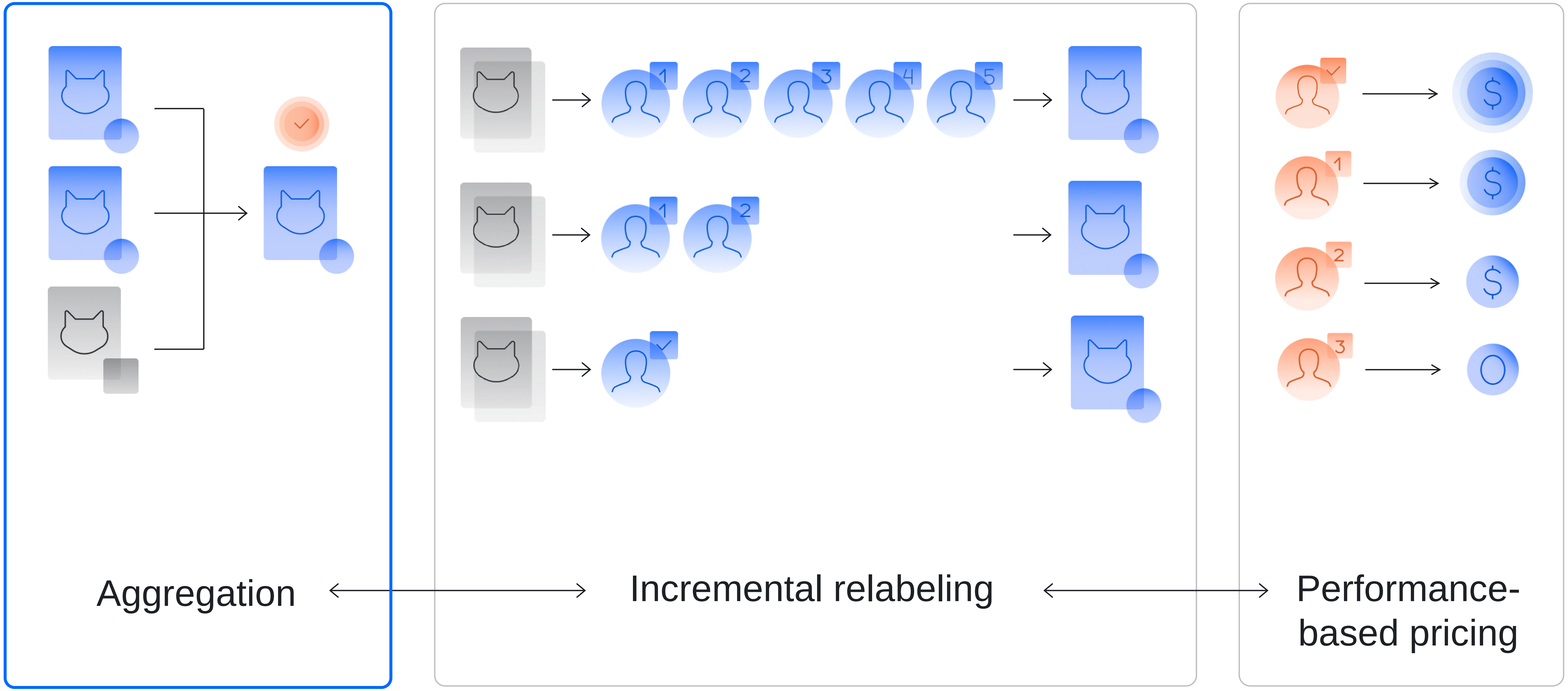
Evaluation of labeling approaches

Accuracy vs Cost

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

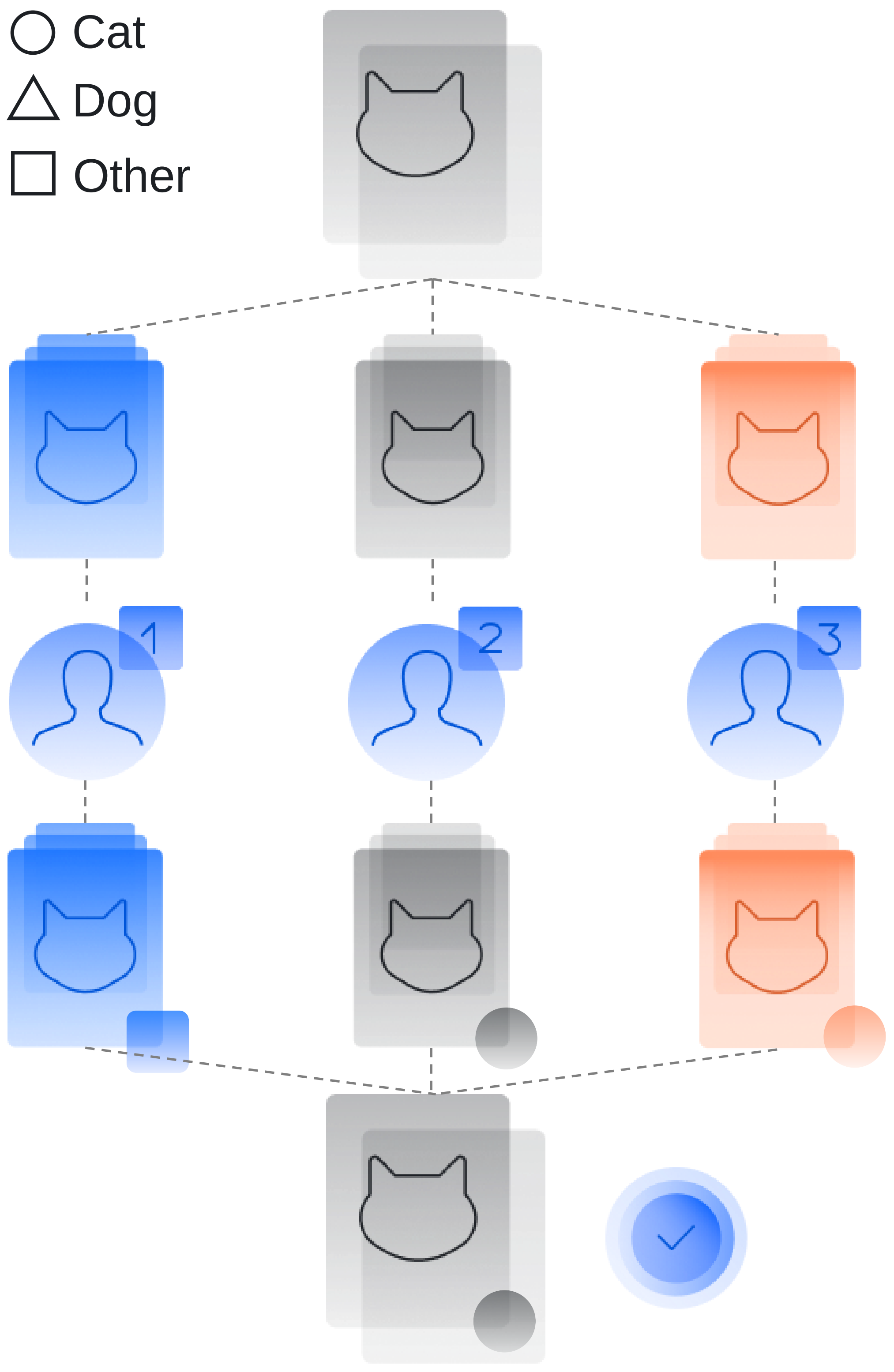
  or
- Labels of a chosen accuracy level for a **minimal budget**

Key components of labeling with crowds



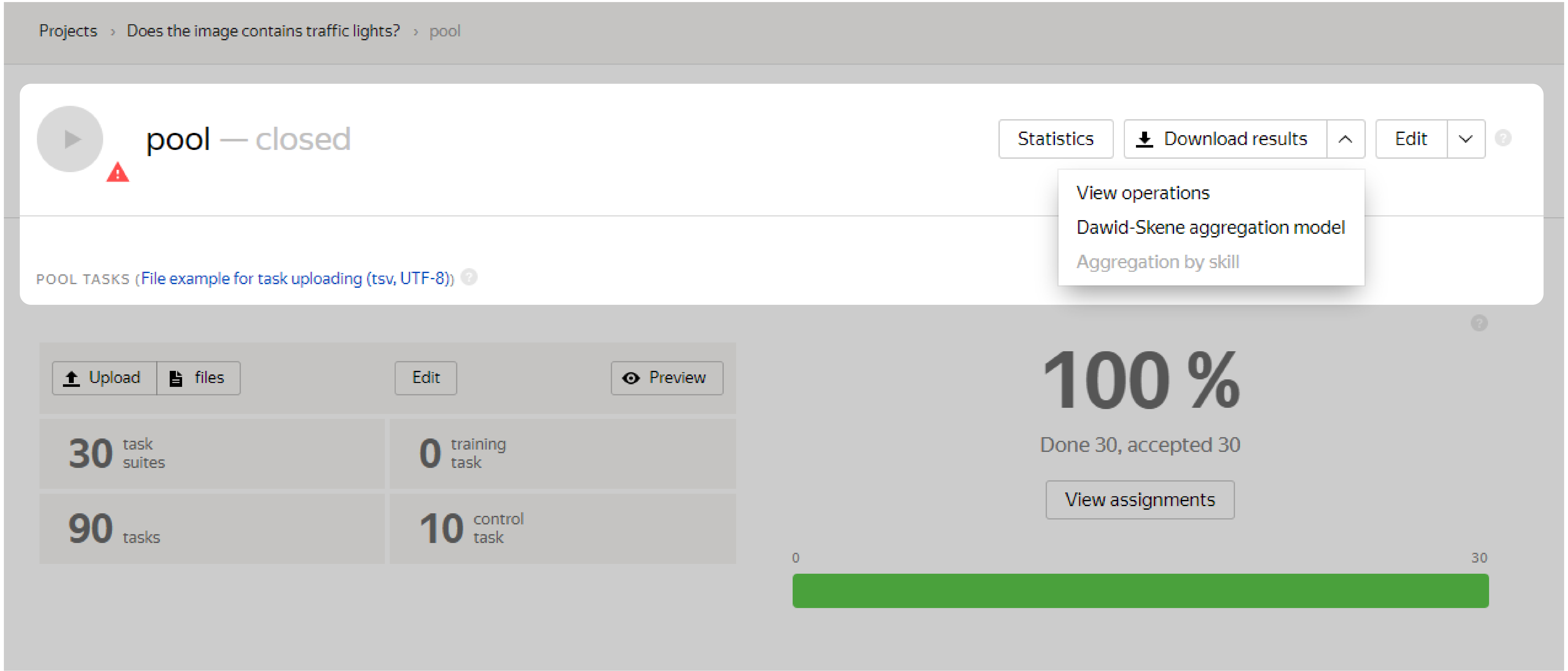
Aggregation

Labeling data with crowds



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

Process results

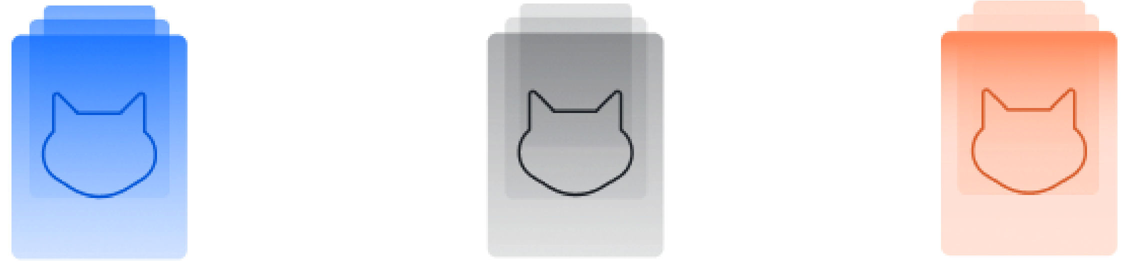


Notation

- Categories $k \in \{1,\dots,K\}$. E.g.:

○ Cat △ Dog ☐ Other

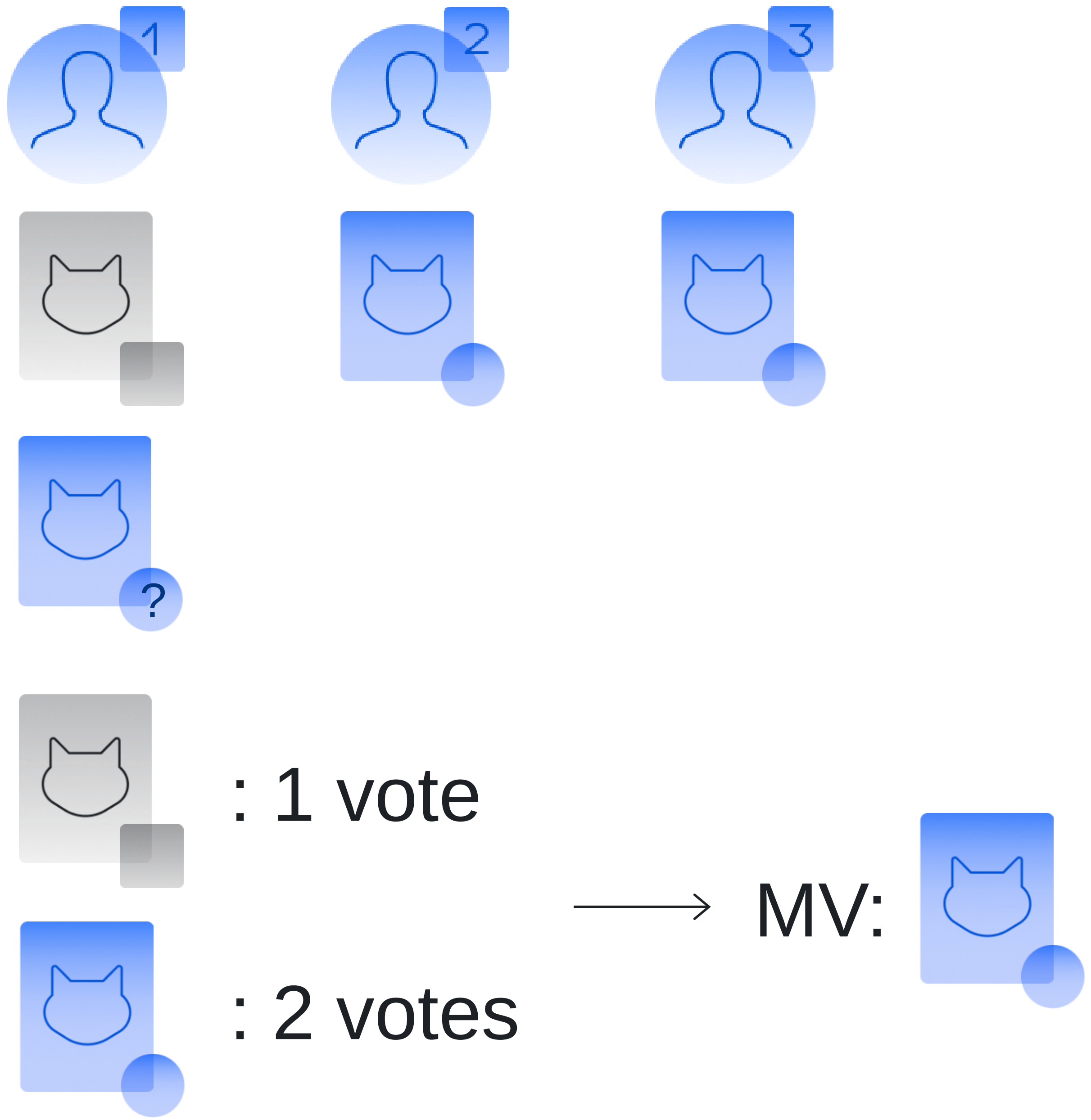
- Objects $j \in \{1,\dots,J\}$. E.g.:



- Workers: $w \in \{1,\dots,W\}$. E.g.:
  - $W_j \subseteq \{1,\dots,W\}$ — workers labeled object j

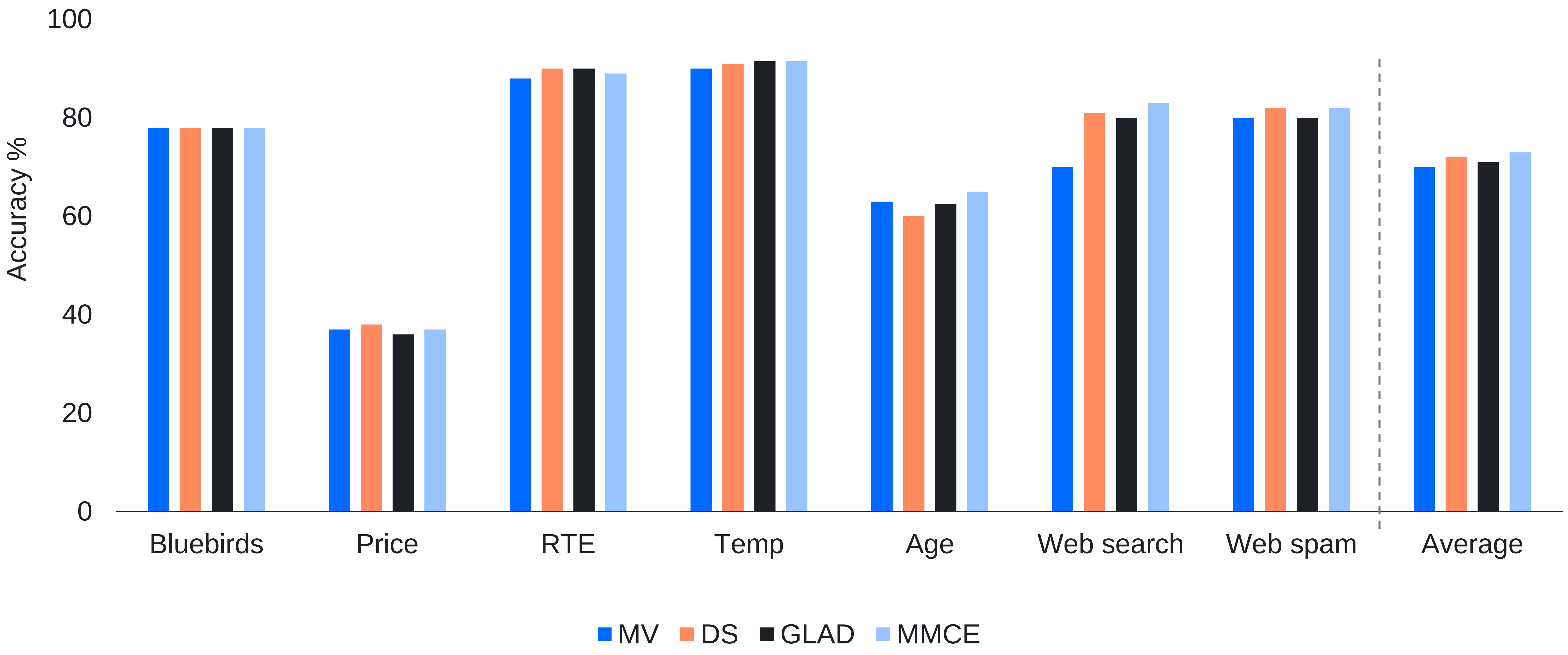
The simplest aggregation: Majority Vote (MV)

- 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:



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

Performance of MV vs other methods

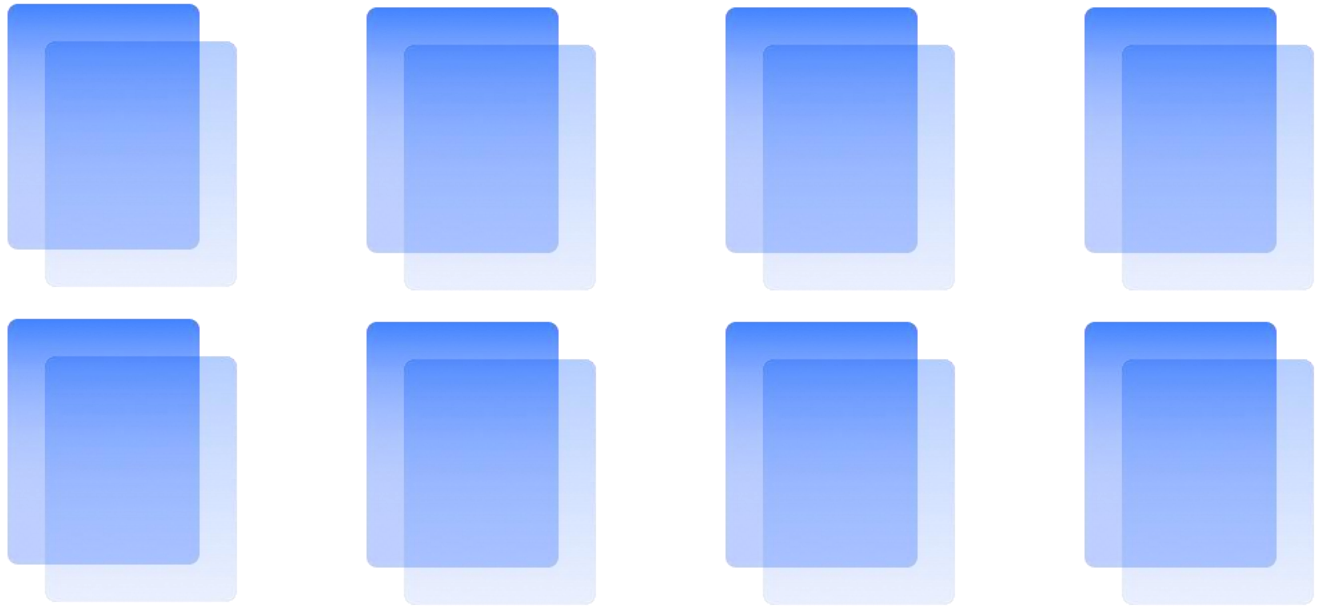


Zhou D. et al. Regularized minimax conditional entropy for crowdsourcing. 2015

Properties of MV

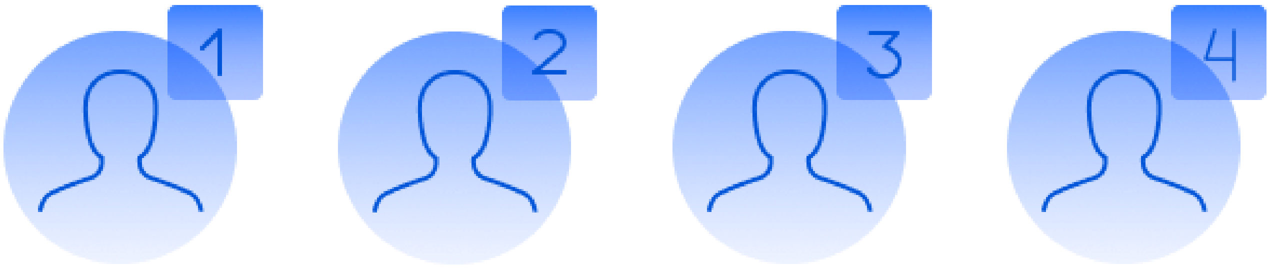
All workers are treated similarly

All objects are treated similarly



Advanced aggregation: workers and objects

Parameterize expertise of workers by e^{w}

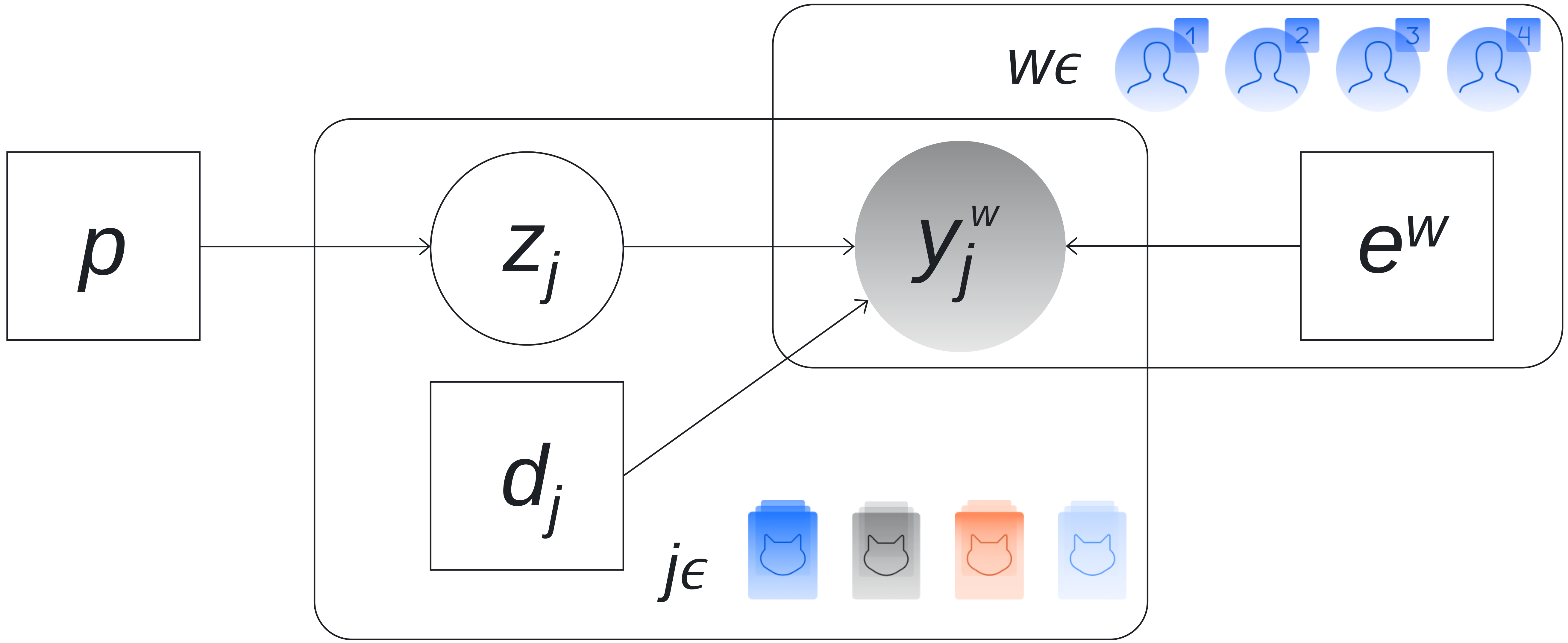


e^{w_1} e^{w_2} e^{w_3} e^{w_4}

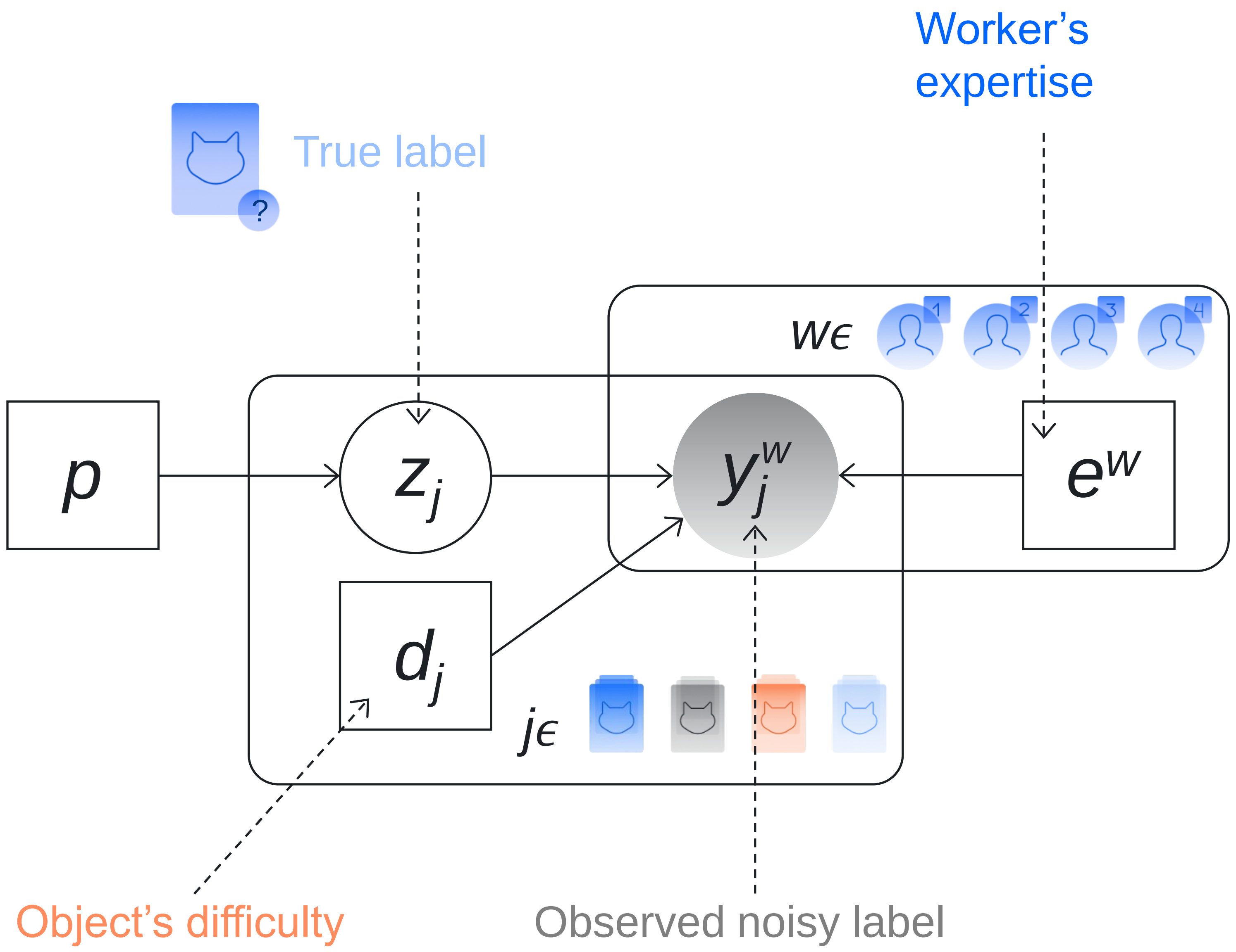
Parameterize difficulty of objects by d_j

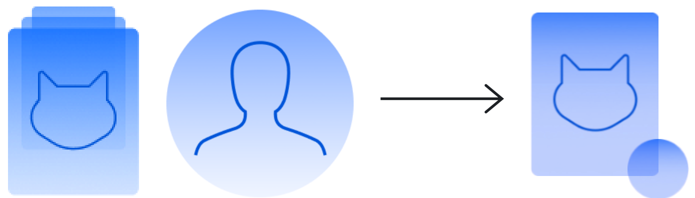
d_{j_1} d_{j_1} d_{j_1} d_{j_1}

Advanced aggregation: latent label models



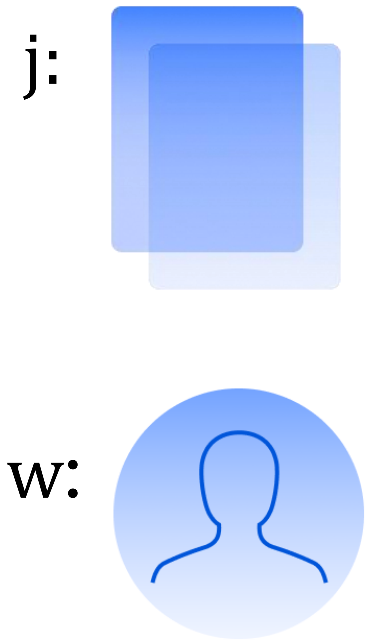
Latent label models: noisy label model





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

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

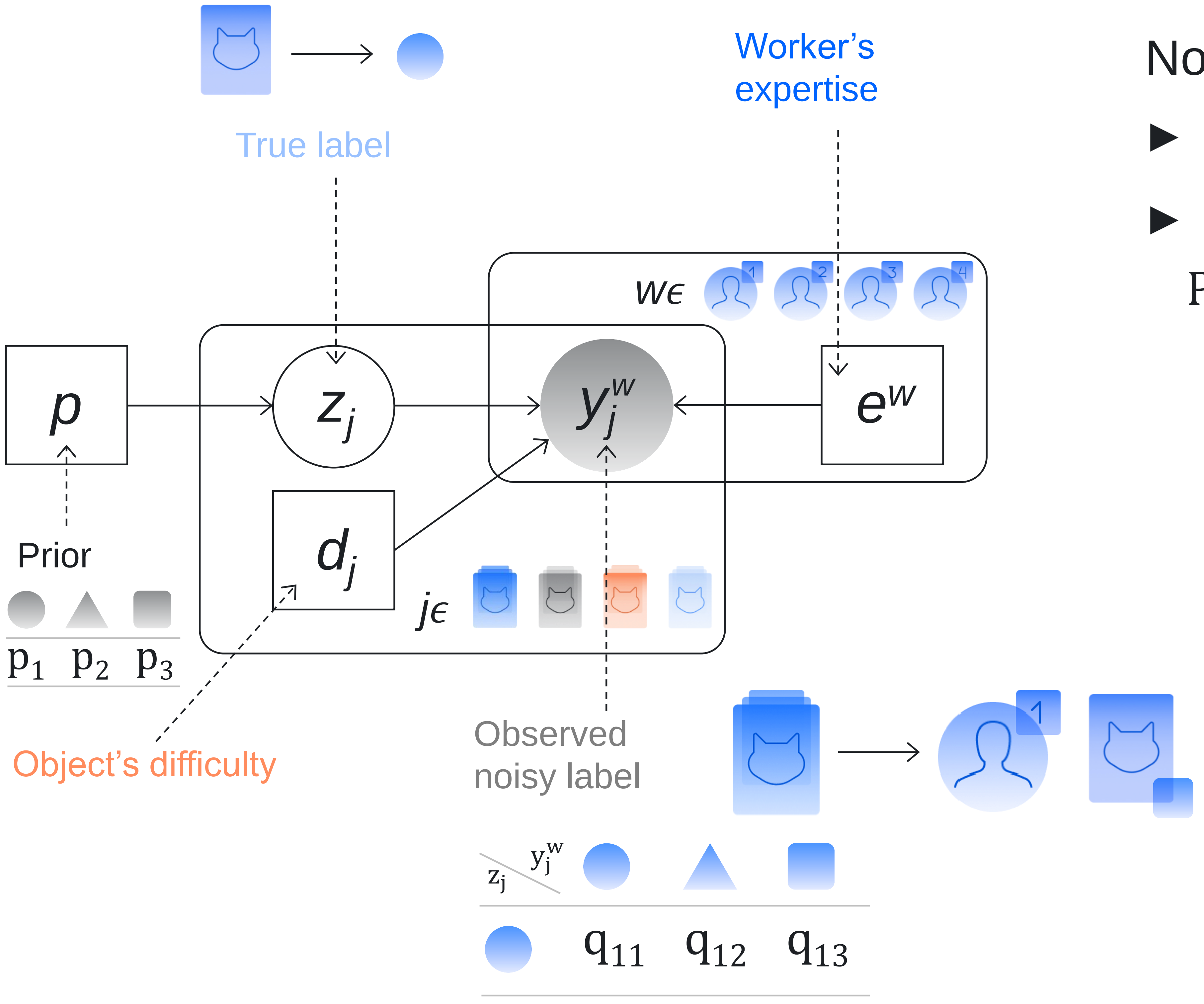


M_j^w:

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

$q_{c1} + q_{c2} + q_{c3} = 1$ 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 $P_Y\left(M_j^w\left[z_j,\cdot\right]\right)$

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

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:

Observed noisy label

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

Latent true label

Latent parameters

Likelihood of noisy and true labels for object j

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

Latent label models: EM algorithm

- Maximization of the expectation of log-likelihood (LL)*

$$\mathbb{E}_{\mathbf{z}} \log \Pr(\mathbf{y}, \mathbf{z}) = \sum_{j \in J} \sum_{z_j \in \{1,\ldots,K\}} \Pr\left(z_j | \mathbf{p}\right) \log \prod_{w \in W_j} \Pr\left(z_j | \mathbf{p}\right) \Pr\left(y_j^w | z_j, \mathbf{d}_j, \mathbf{e}^w\right)$$

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

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

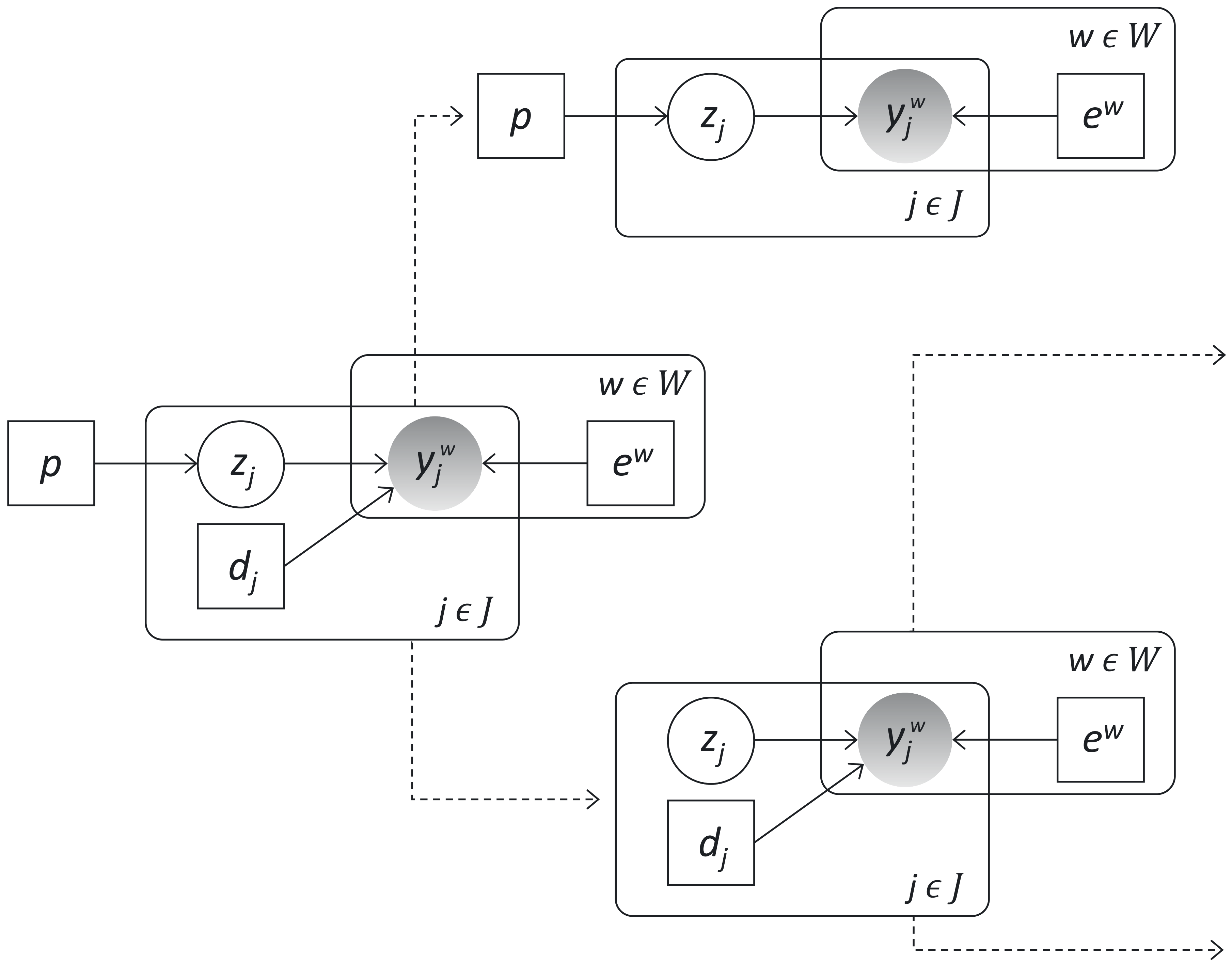
- **M-step:** Maximize the expectation of LL with respect to the posterior distribution of $\hat{\mathbf{z}}$:

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

  - Analytical solutions
  - Gradient descent

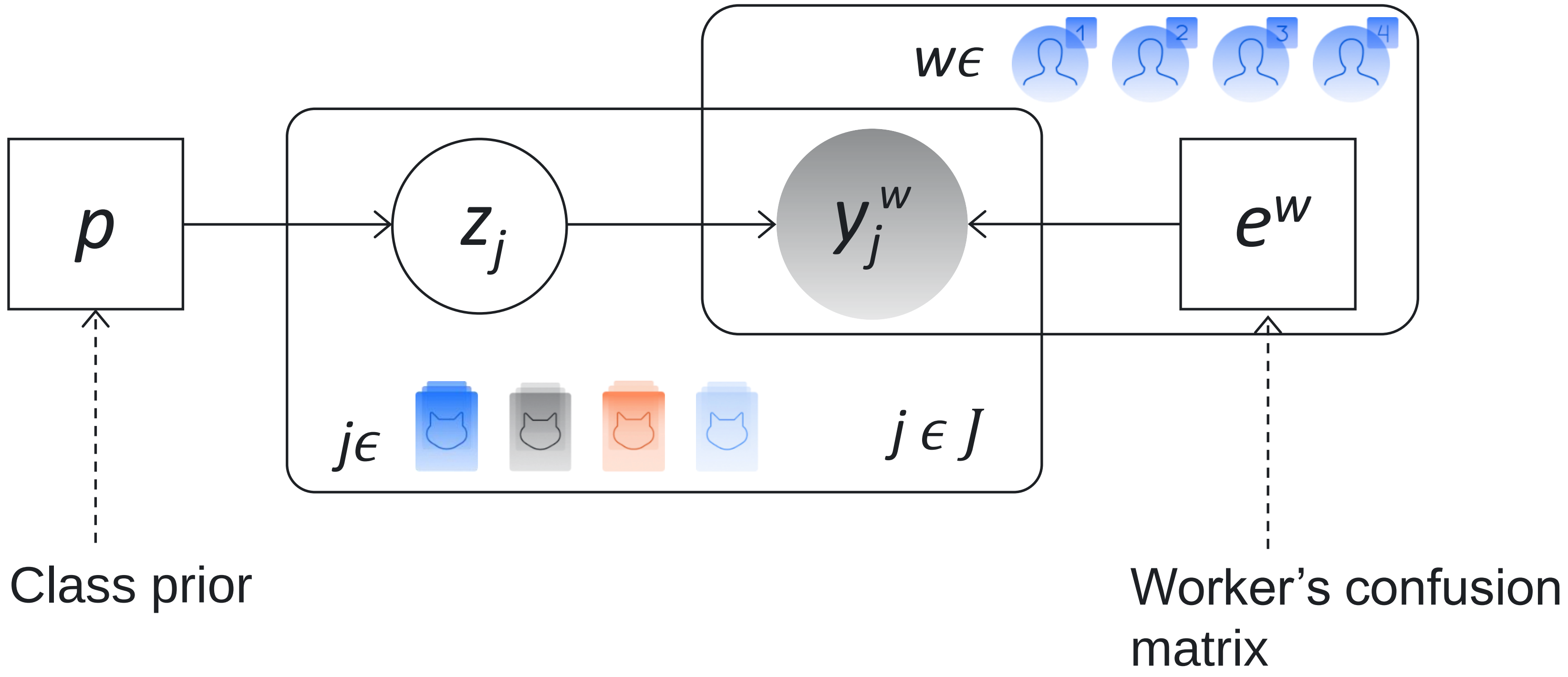
* it is a lower bound on LL of $\mathbf{y}$ and $\mathbf{z}$

Latent label model (LLM): special cases



- Dawid and Skene model (DS):
  - Categories are different
  - Objects are **similar**
  - Workers are different
- Generative model of labels, abilities, and difficulties (GLAD):
  - Categories are **similar**
  - Objects are different
  - Workers are different
- Minimax conditional entropy model (MMCE):
  - Categories are different
  - Objects are different
  - Workers 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 $K \times K$: $e^w[c, k] = \Pr(Y^w = k | Z = c)$
- ▶ Model:
  - Z_j ~Mult(p)
  - y_j^w ~Mult$(e^w[z_j, \cdot])$

Dawid and Skene, Maximum Likelihood Estimation of Observer Error-Rates Using the *EM* Algorithm,1979

DS: parameters optimization

- **E-step:**

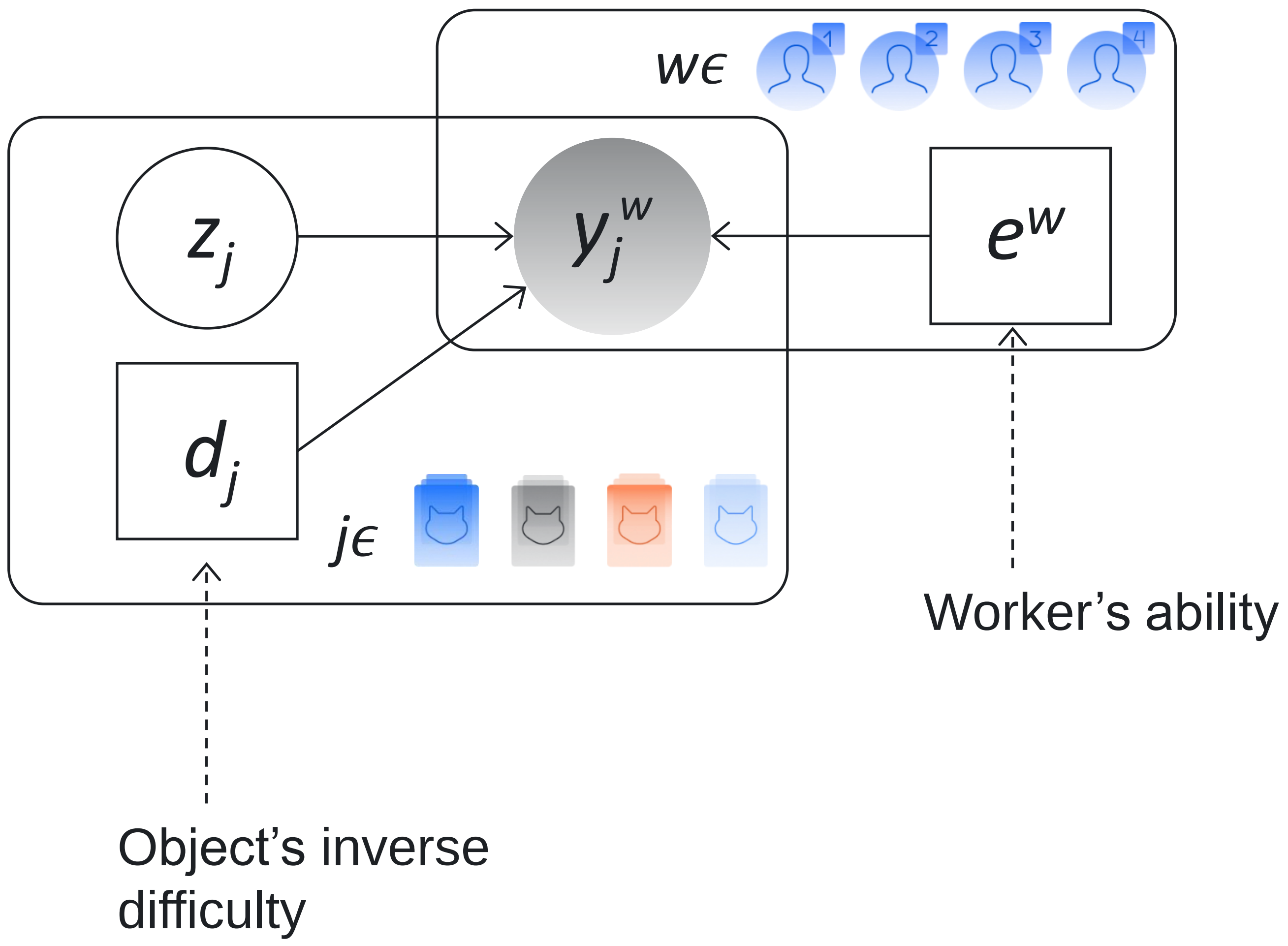
$$\hat{z}_j[c] = \frac{p[c]\ \prod_{w \in W_j} e^w[c, y_j^w]}{\sum_k p[k]\ \prod_{w \in W_j} e^w\left[k, y_j^w\right]}, \qquad c = 1, \dots, K$$

- **M-step:** Analytical solution

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

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

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



LLM with parameters:

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

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

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

Whitehill et al., Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, 2009

GLAD: parameters optimization

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

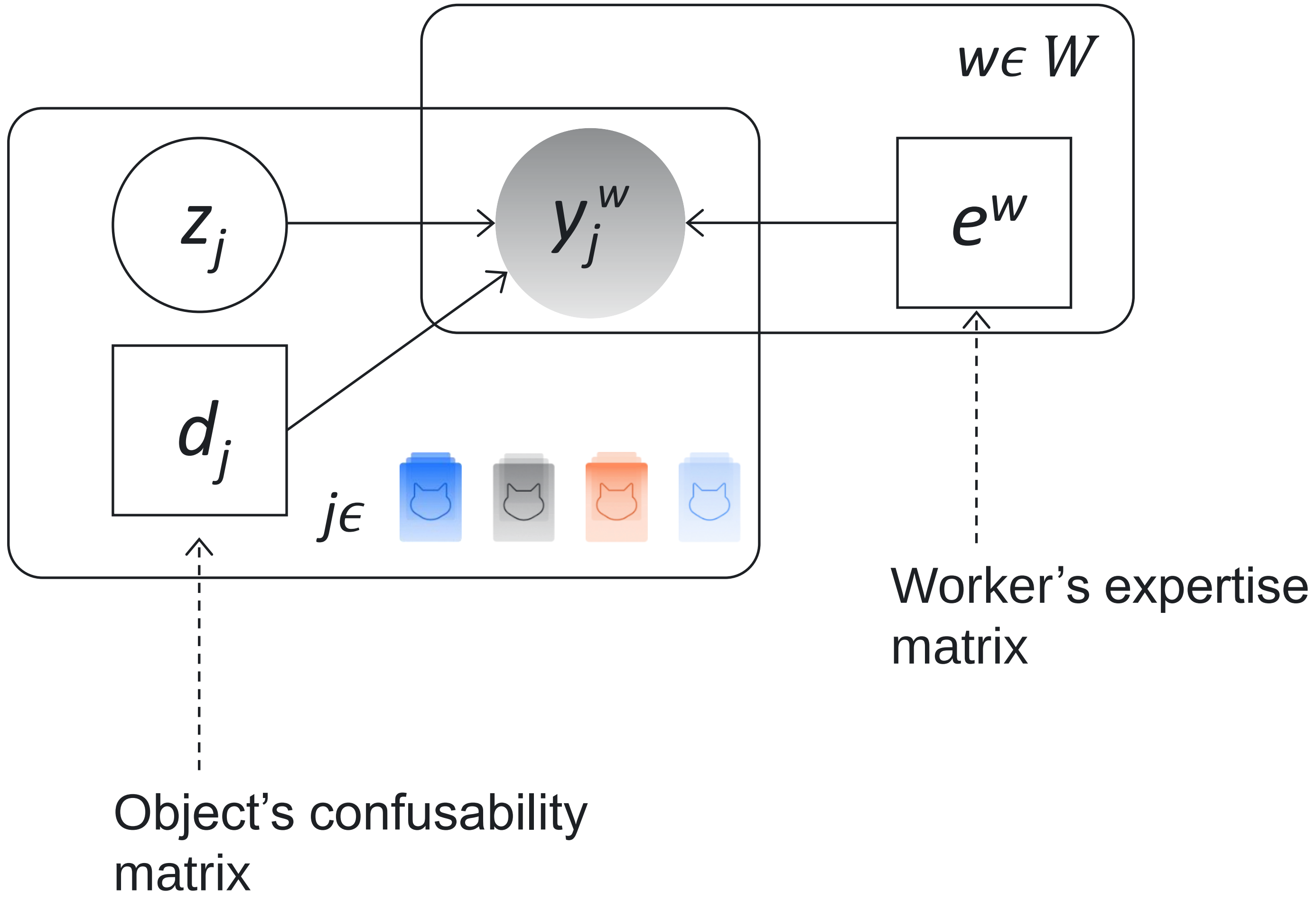
- **E-step:**

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

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

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

MiniMax Conditional Entropy model (MMCE)



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

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

*The model was derived by minimizing the maximum conditional entropy of observed labels

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

Zhou et al., Learning from the Wisdom of Crowds by Minimax Entropy, 2012

Summary of aggregation methods

	MV	DS	GLAD	MME
Categories (K)				
Objects (J)				
Workers (W)	1 1 1	1 2 3	1 2 3	1 2 3
Number of parameters	0	$WK^2 + K$	$W + J$	$(W + J)K^2$

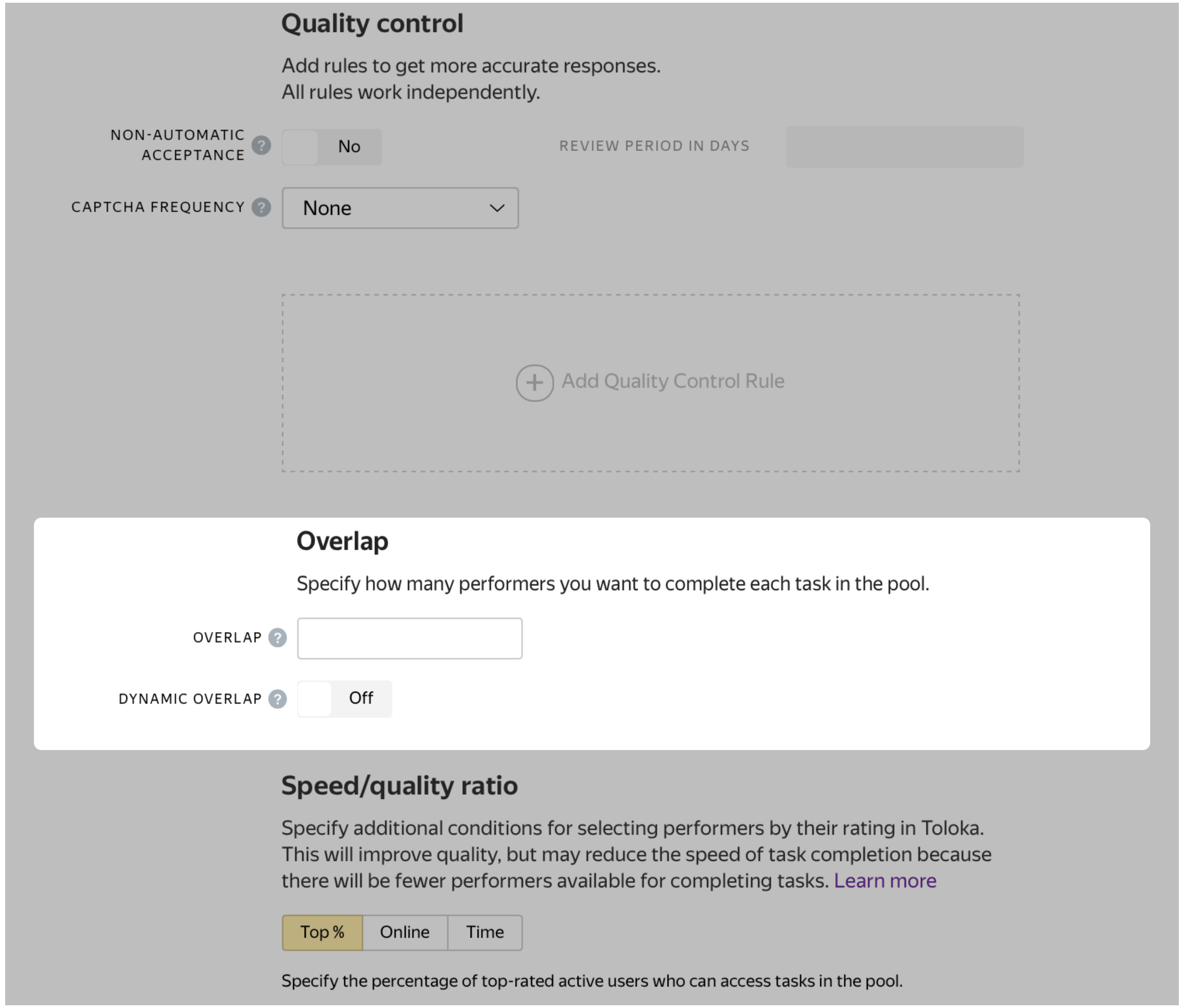
Key components of labeling with crowds

Aggregation ↔ Incremental relabeling ↔ Performance-based pricing

Incremental relabeling

aka dynamic overlap

Pool settings: dynamic overlap



Incremental relabeling problem

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

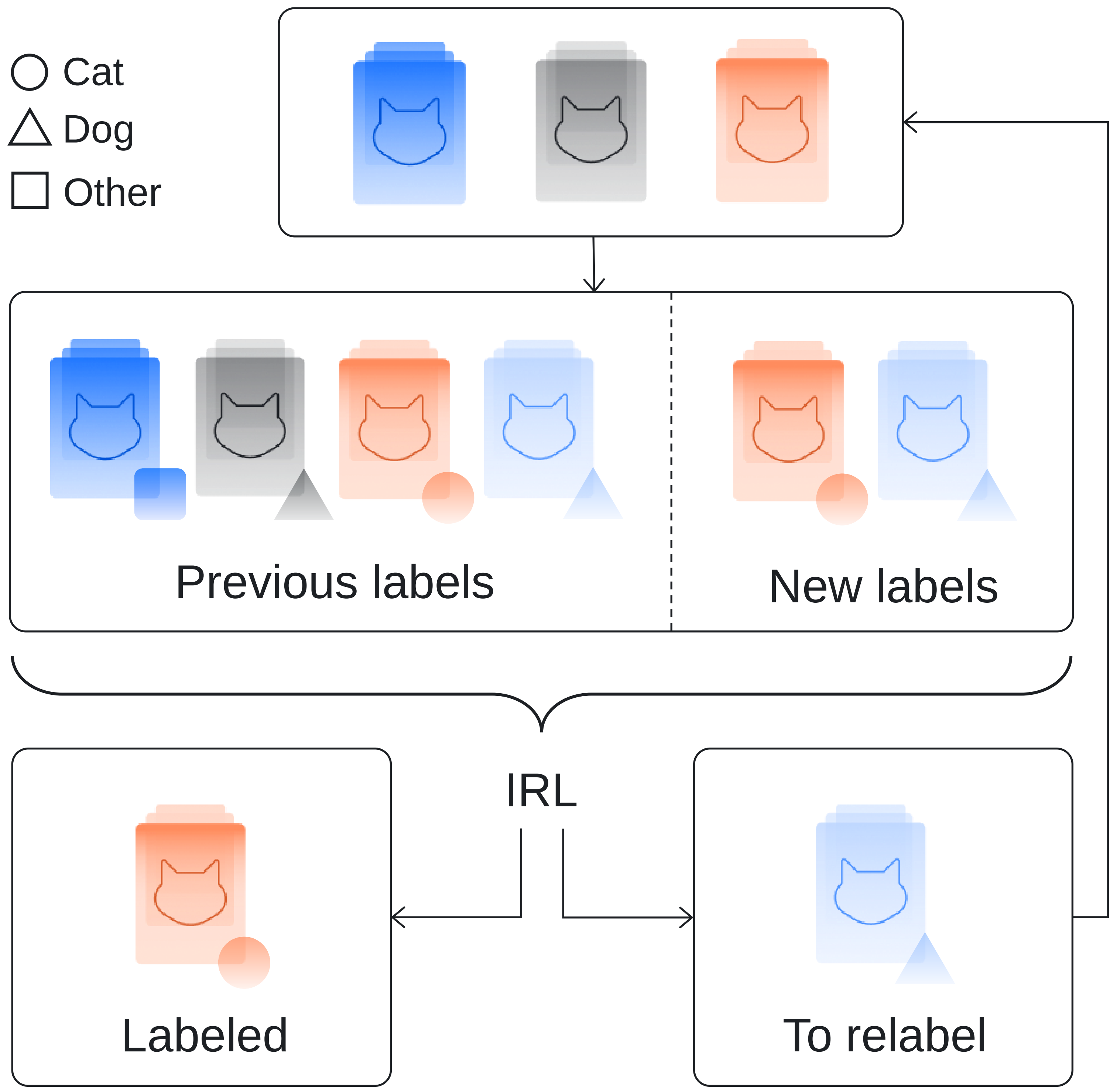
Incremental relabeling scheme (IRL)

Request 1 label for each object

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

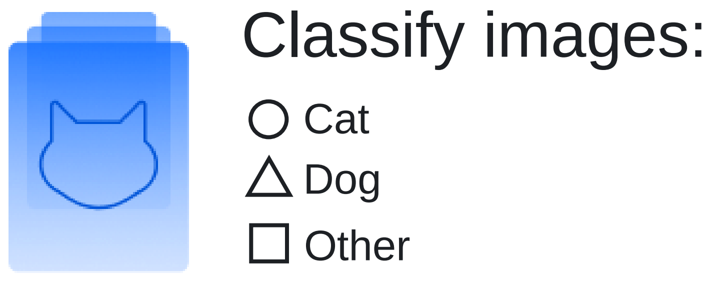
Decides:
(1) which objects are labeled
(2) which objects to relabel

Repeat until all tasks are labeled

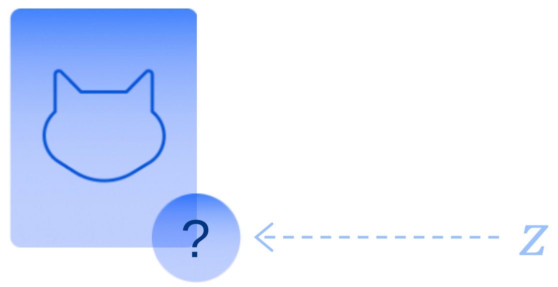


Notations

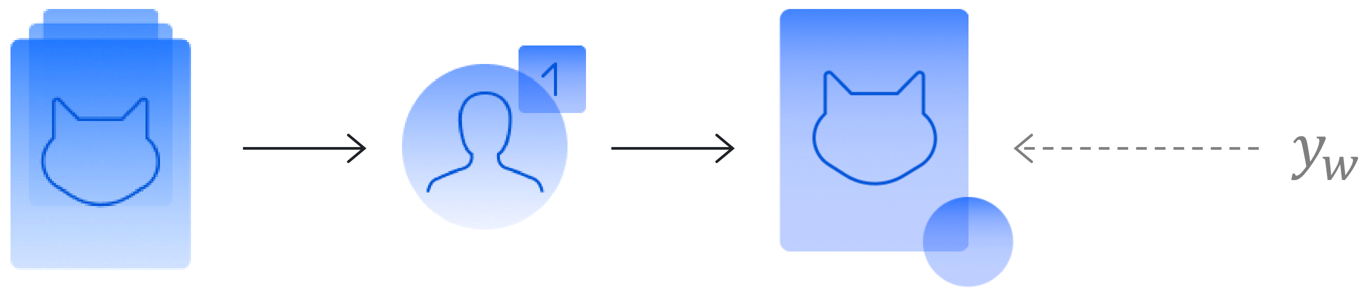
- Consider one object



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



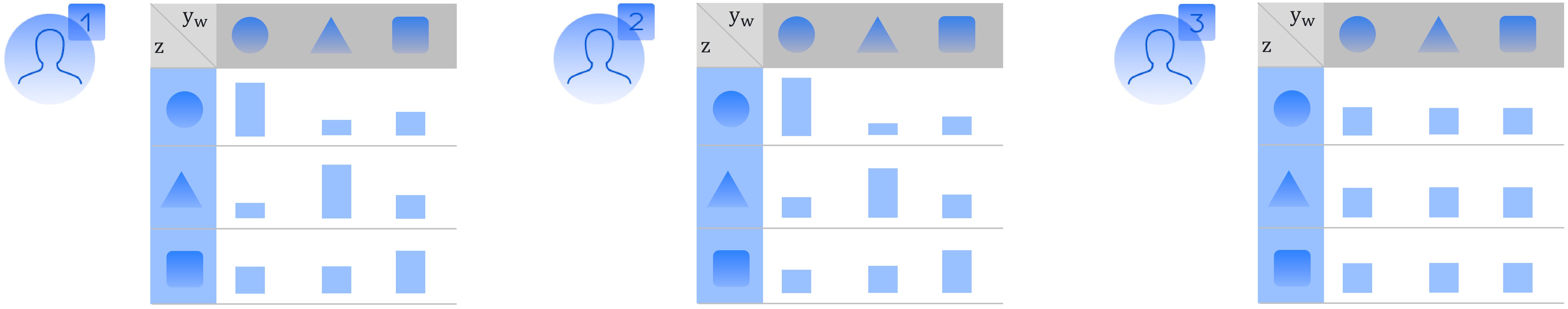
- $y_w \in \{1, \dots, K\}$ — observed noisy label from worker w:



Notations

- Noisy label model for worker w:

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



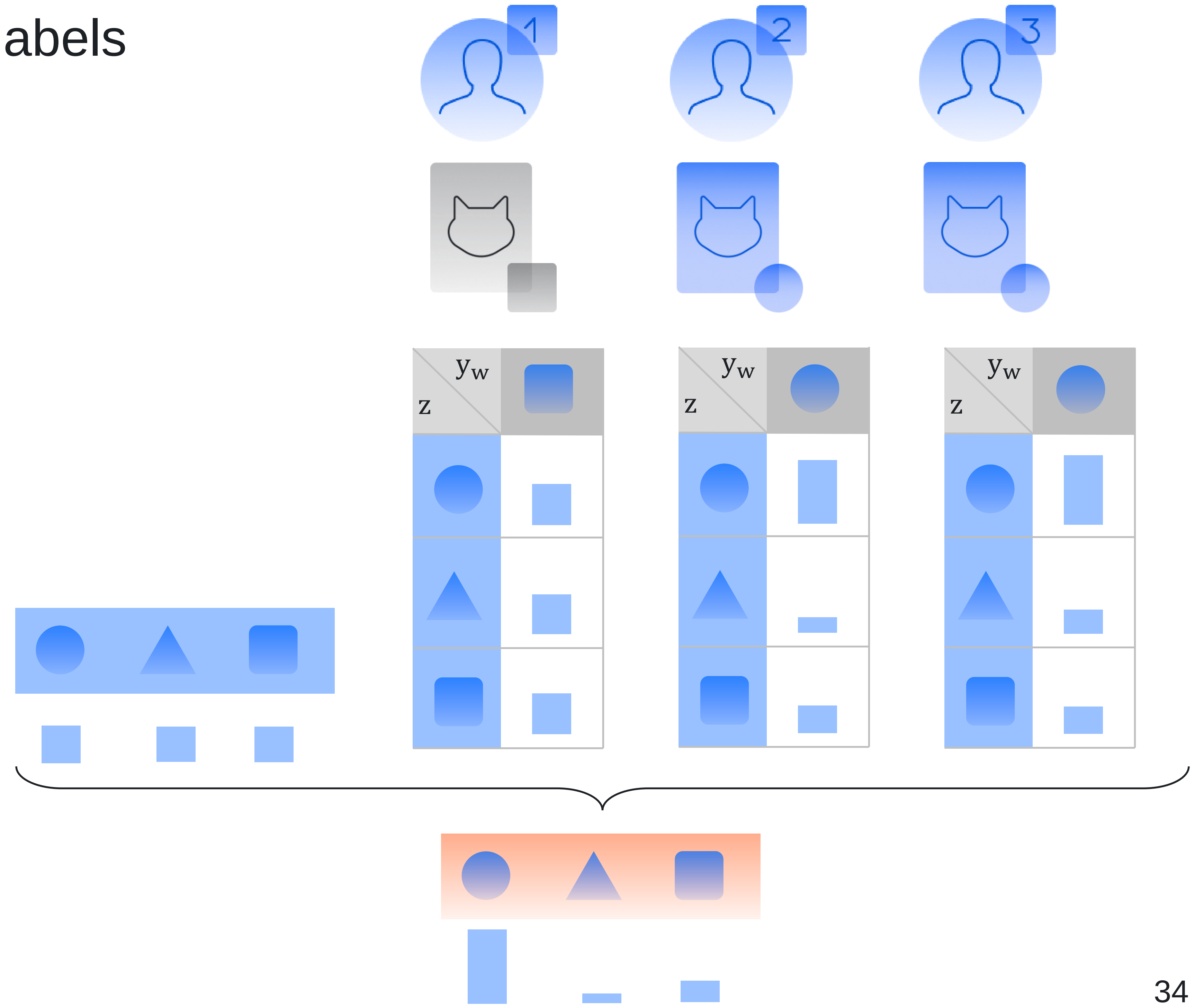
- Prior distribution: $\Pr(Z = k) = p_k$

Posterior distribution

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

- Using Bayes rule:

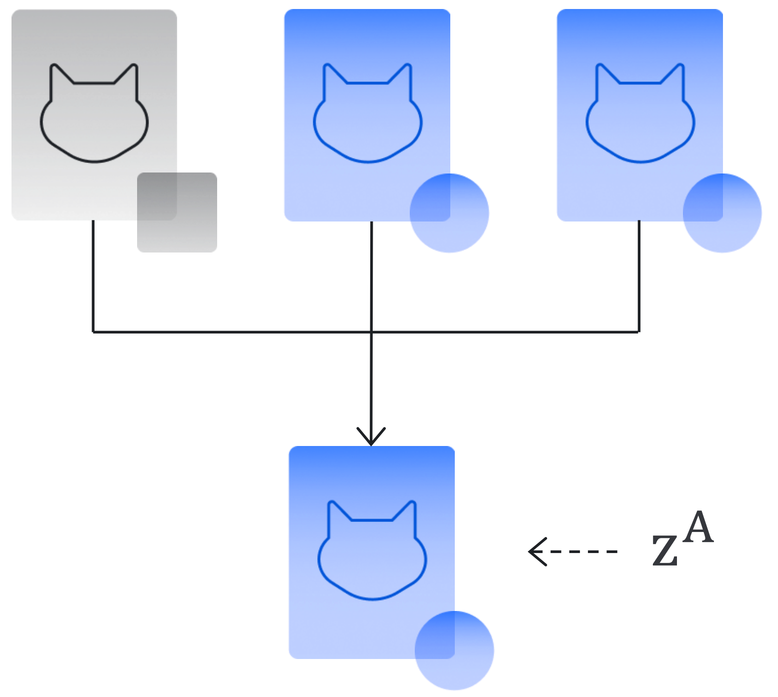
$$\Pr(Z = k \mid \{y_{w_1}, \dots, y_{w_n}\}) = \frac{\Pr(Z = k)\Pr(\{y_{w_1}, \dots, y_{w_n}\} \mid Z = k)}{\Pr(\{y_{w_1}, \dots, y_{w_n}\})} = \frac{p_k \prod_{i=1}^{n} M_{w_i}[k, y_{w_i}]}{\sum_{t=1}^{K} p_t \prod_{i=1}^{n} M_{w_i}[t, y_{w_i}]}$$



Expected accuracy of aggregated labels

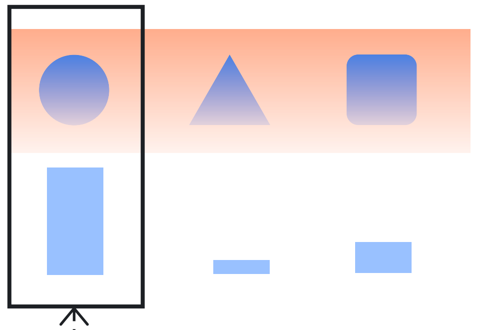
- Let A be an aggregation model, e.g. MV, DS, GLAD,…
- Denote aggregated label $z^A = A(\{y_{w_1}, \ldots, y_{w_n}\})$
- Expected accuracy of aggregated labels given noisy labels is

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



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

Parameter



Incremental relabelling algorithm

Input: $\cup_{t=1}^{T-1} Y^t$ — previous labels till step T

Y^T — new labels

Output: R — objects to relabel

For each object j with a label in Y^T: ← Object with a new label

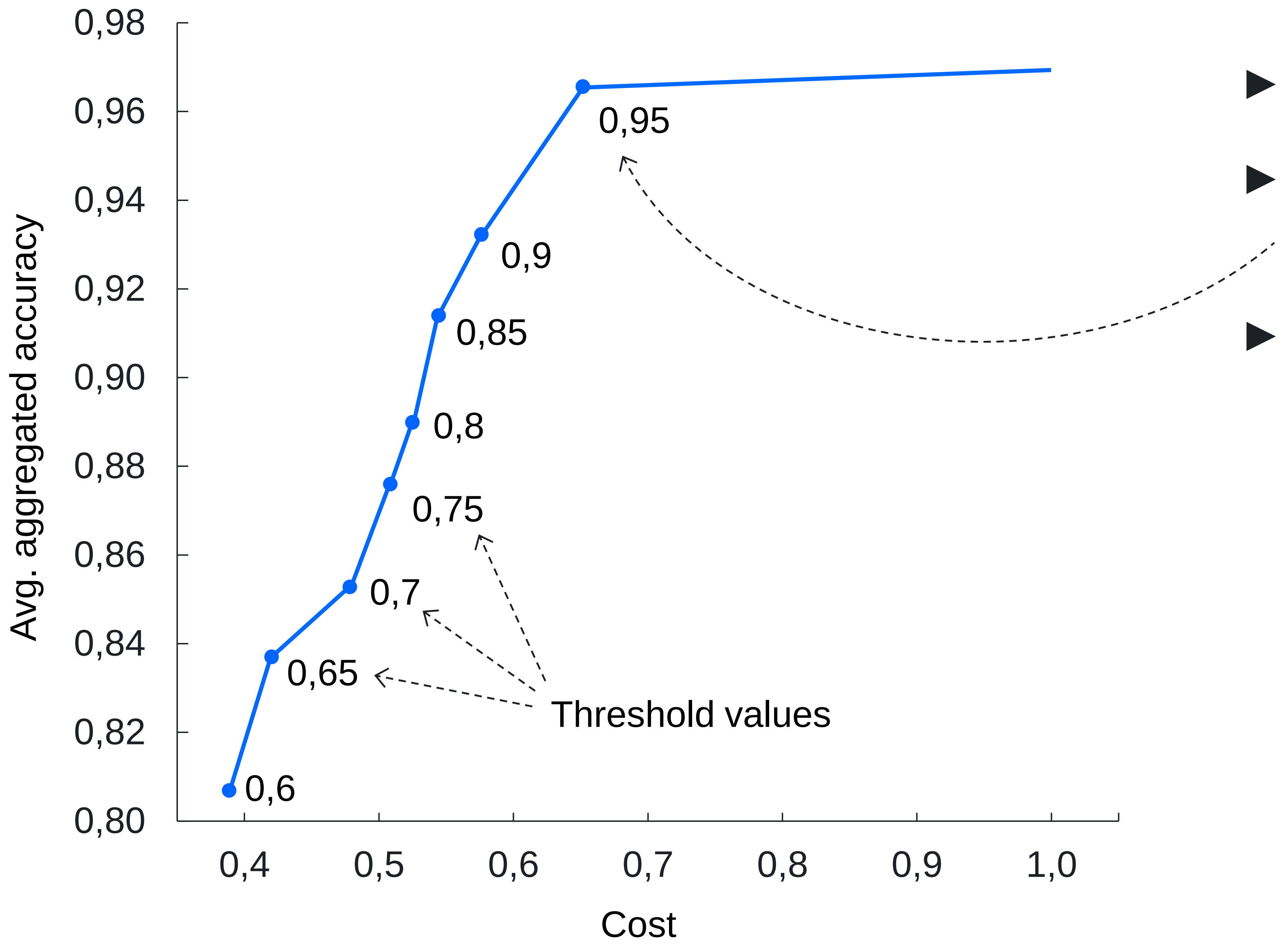
$z_j^M = M(\cup_{t=1}^{T} Y^t)$ ← Current aggregated label

$c_j = E(z_j = z_j^M | \cup_{t=1}^{T} Y^t)$ ← Expected accuracy for the current aggregated label

If $c_j < c$, then $R = R \cup j$

Parameter: c — threshold for expected accuracy

Threshold in IRL: cost – accuracy trade-off



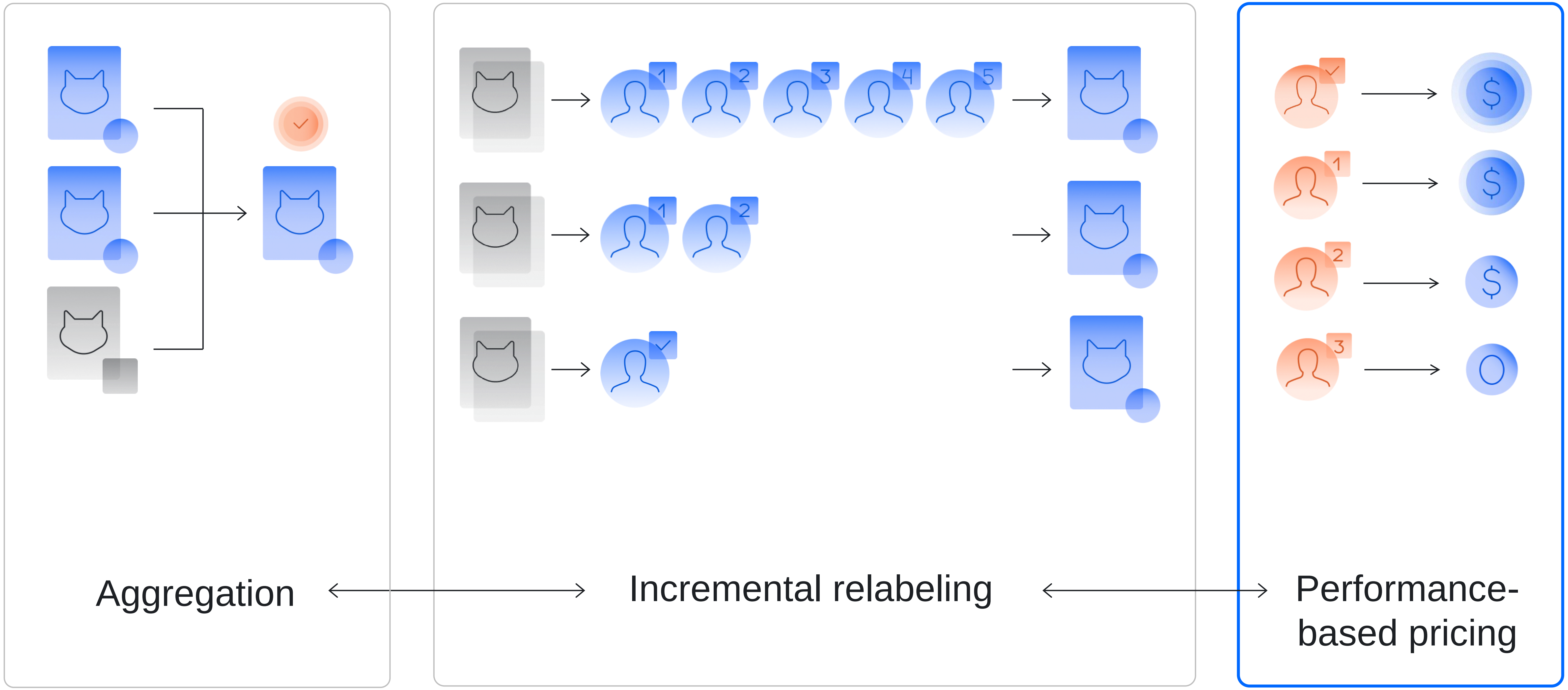
- Optimal threshold $c = 0.95$
- A higher c does not increase accuracy
- Saving ≈ 35% of noisy labels

How to obtain a cost-accuracy plot

Data for the plot:
- ► Label a pool of objects with a redundant overlap (e.g., 10)
- ► Obtain ground truth labels for the objects (e.g., expert labels or MV labels)

Simulate IRL with different thresholds using the data:
- ► For each threshold c from a grid $0 < c_0 < \dots < c_m \leq 1$
- ► Repeat N times:
  1. Shuffle noisy labels and fix the order of labels
  2. Draw labels sequentially and test the IRL condition after each label
  3. Once the IRL condition for an object is met, discard unused labels for the object
  4. When all objects are labelled calculate
     - accuracy of aggregated labels
     - cost as the fraction of used noisy labels
- ► Average N values of aggregated accuracy and N values of cost for each value of threshold c

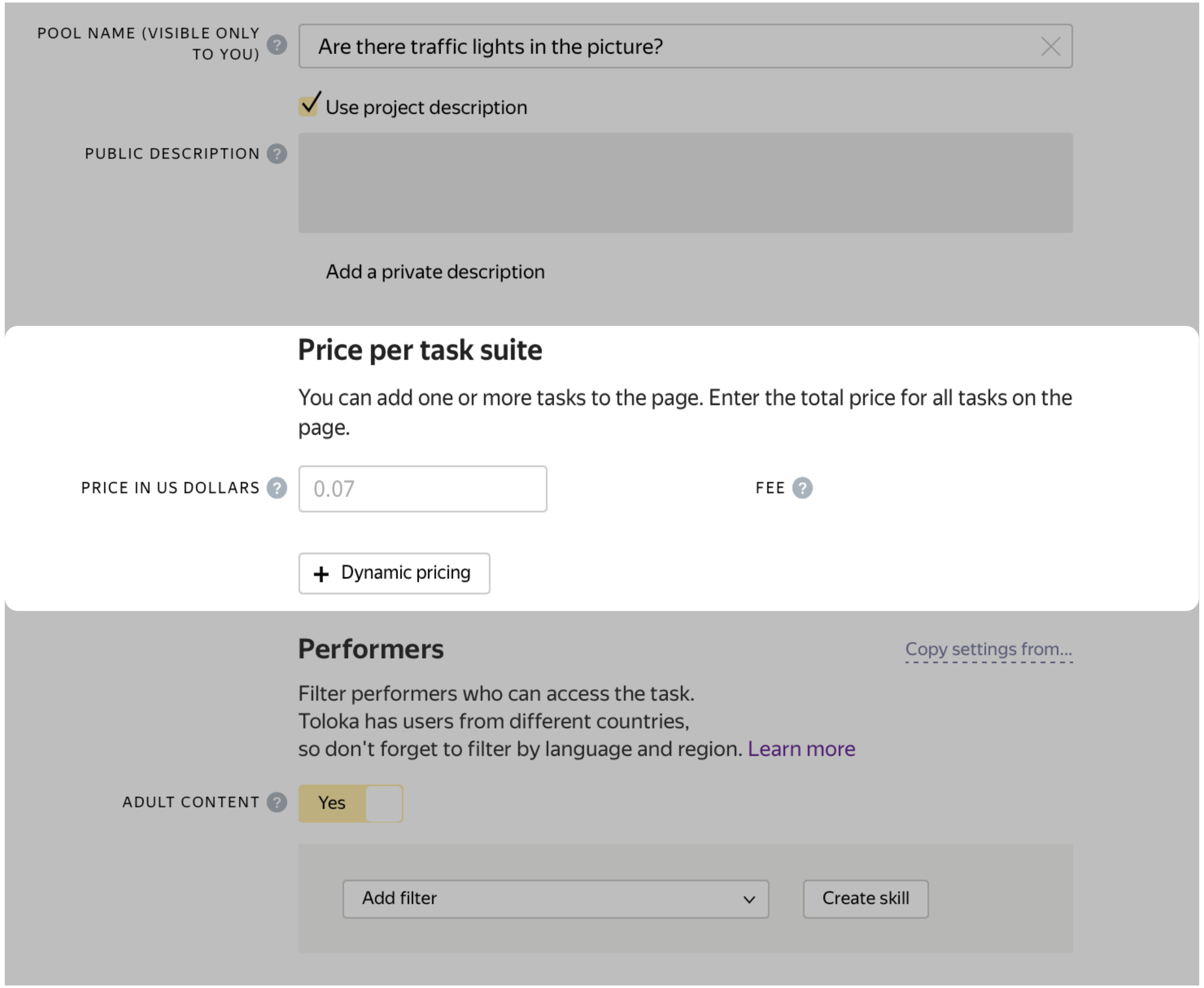
Key components of labeling with crowds



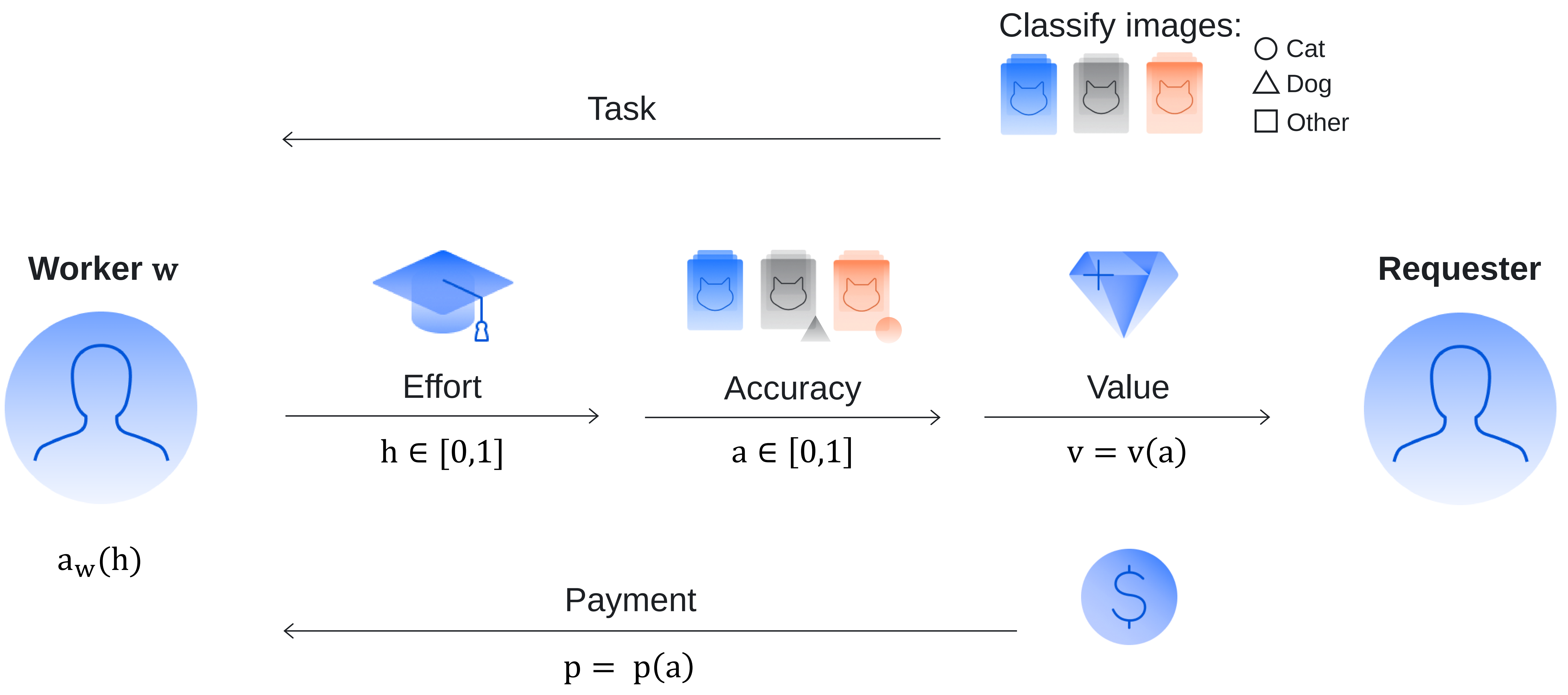
Performance-based pricing

aka dynamic pricing

Pool settings: dynamic pricing

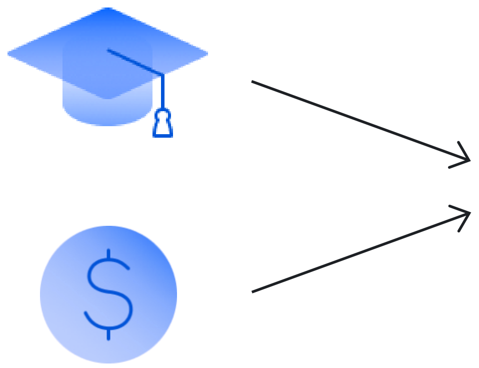


Labeling as a game: notation



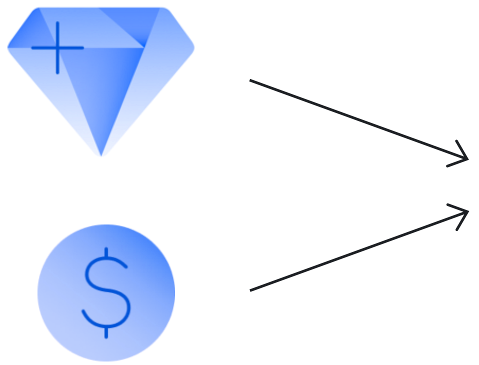
Labeling as a game: formalization

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



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

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



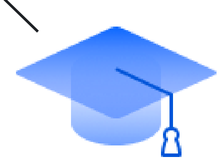
$$\frac{v(a)}{p(a)} \rightarrow \max_{a \in [0,1]}$$

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



Theorem: the requester and workers 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.:

p = 0.3$ a_0 = 0.99

- $\hat{q}_w = \Pr(y^w = z)$ — estimated quality level of worker 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 $\hat{z}_j^{wMV}$

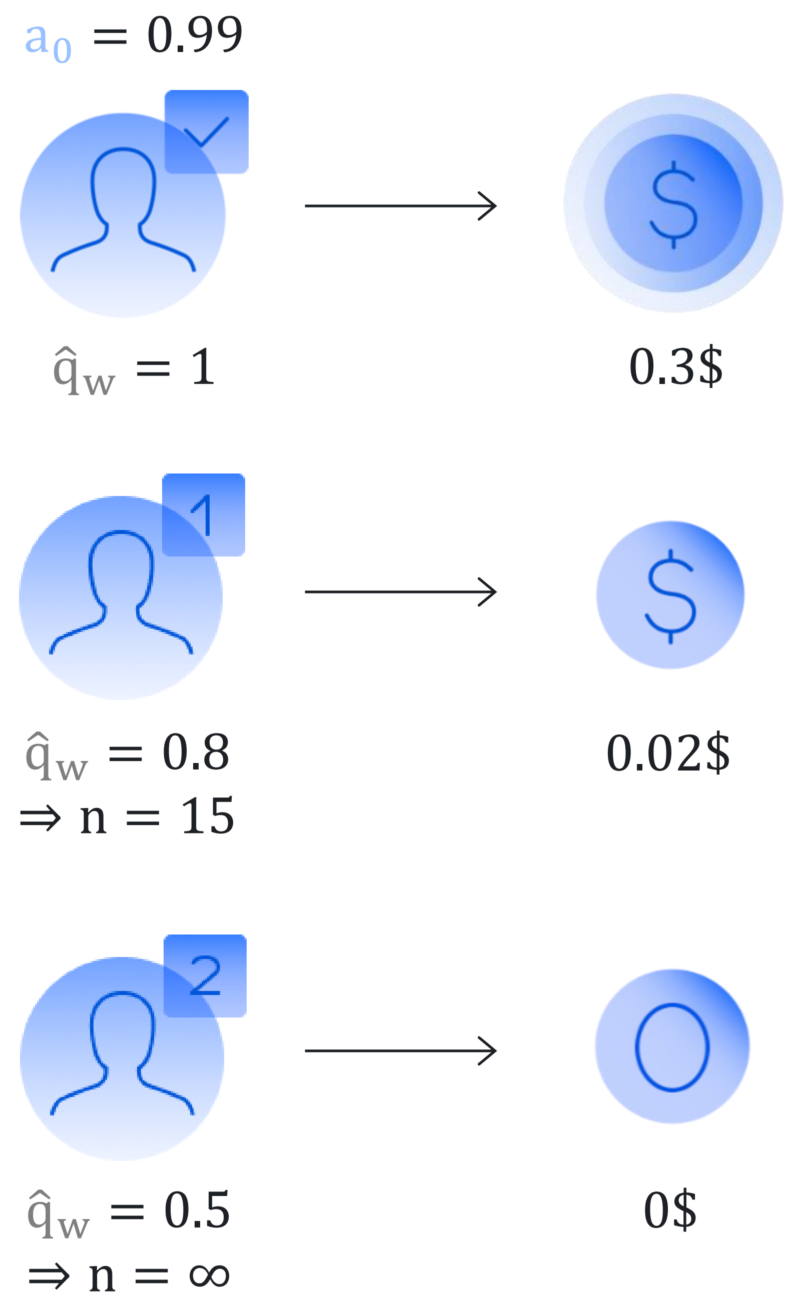
Performance-based pricing in practice

- Pricing rules

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

$$\underbrace{\sum_{k=0}^{n/2} \binom{n}{k} \hat{q}_w^{n-k} (1 - \hat{q}_w)^k}_{\text{Expected accuracy for MV}} \geq a_0$$

The price is p/n



Key components of labeling with crowds