Yandex

Yandex

SIGMOD 2020 hands-on tutorial

Crowdsourcing Practice for Efficient Data Labeling: Aggregation, Incremental Relabeling, and Pricing

Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zerminova, Daria Baidakova

Part V:

Theory on Aggregation

Dmitry Ustalov and Valentina Fedorova,
Data Analysis and Research Group, Yandex

Tutorial schedule

Introduction: 15 min

Part I: 20 min
Main Components

Part II: 10 min
Introduction to Crowd Platform

Part III: 15 min
Brainstorming pipeline

Part IV: 60 min
Set & Run Projects

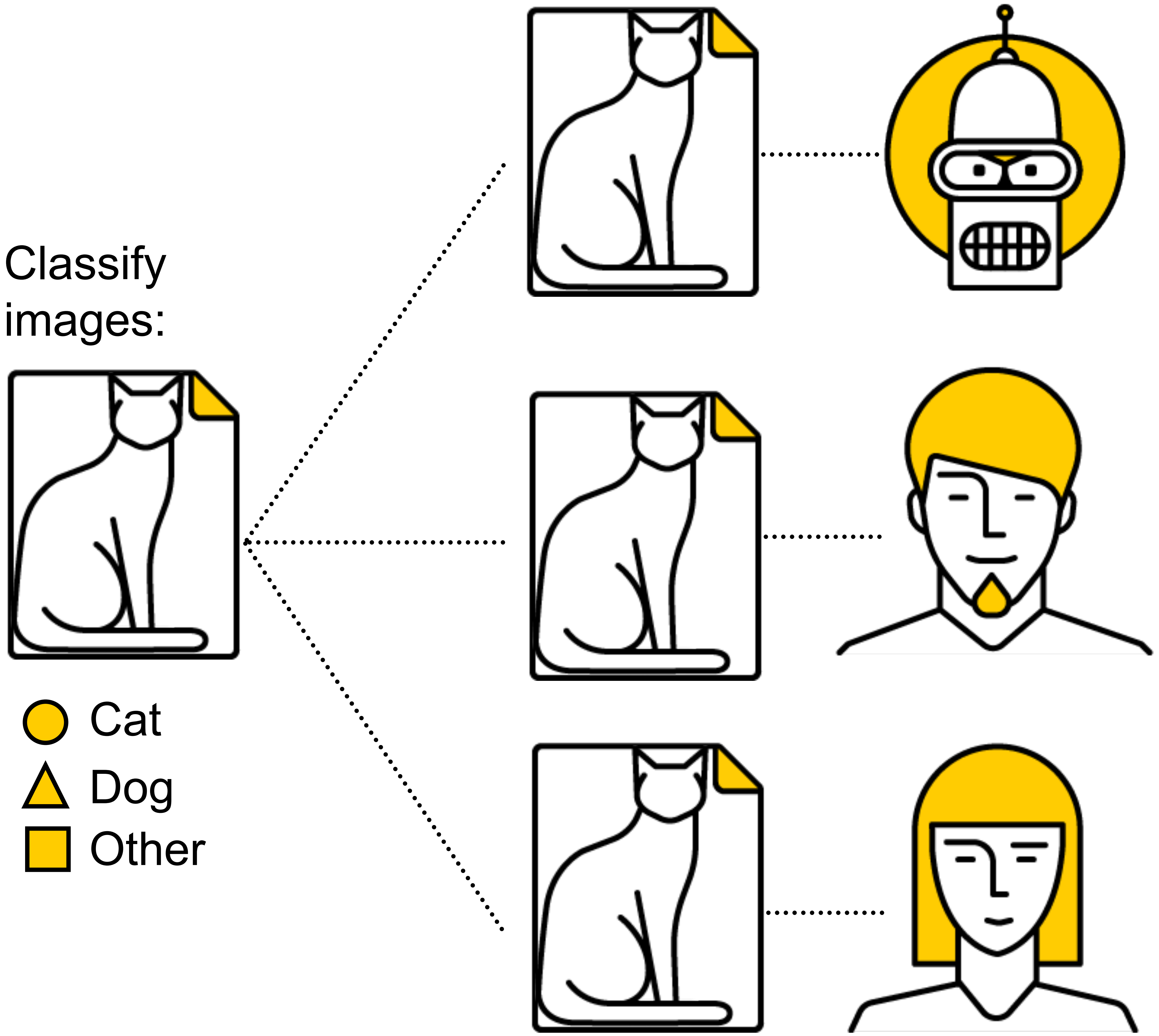
Part V: 25 min
Theory on Aggregation

Break:
30 min

Part VI: 20 min
Set & Run Projects cont.

Part VII: 10 min
Results & Conclusions

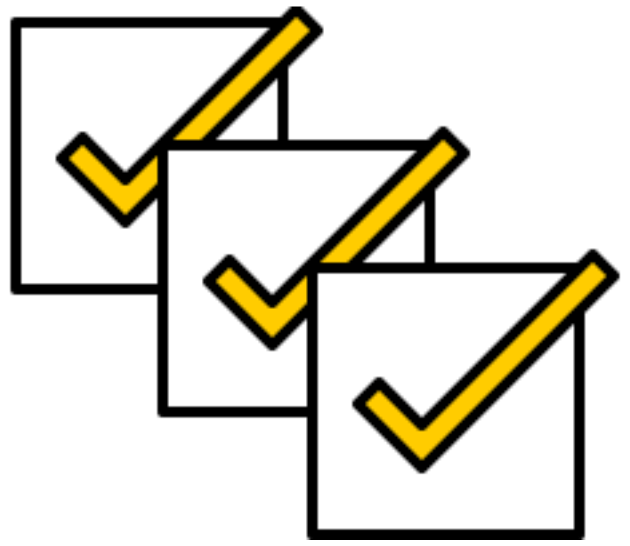
Labeling data with crowdsourcing



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

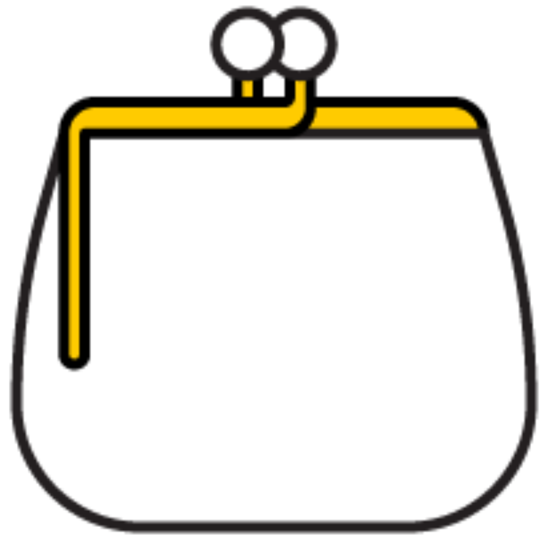
Evaluation of labeling approaches

Accuracy



vs

Cost

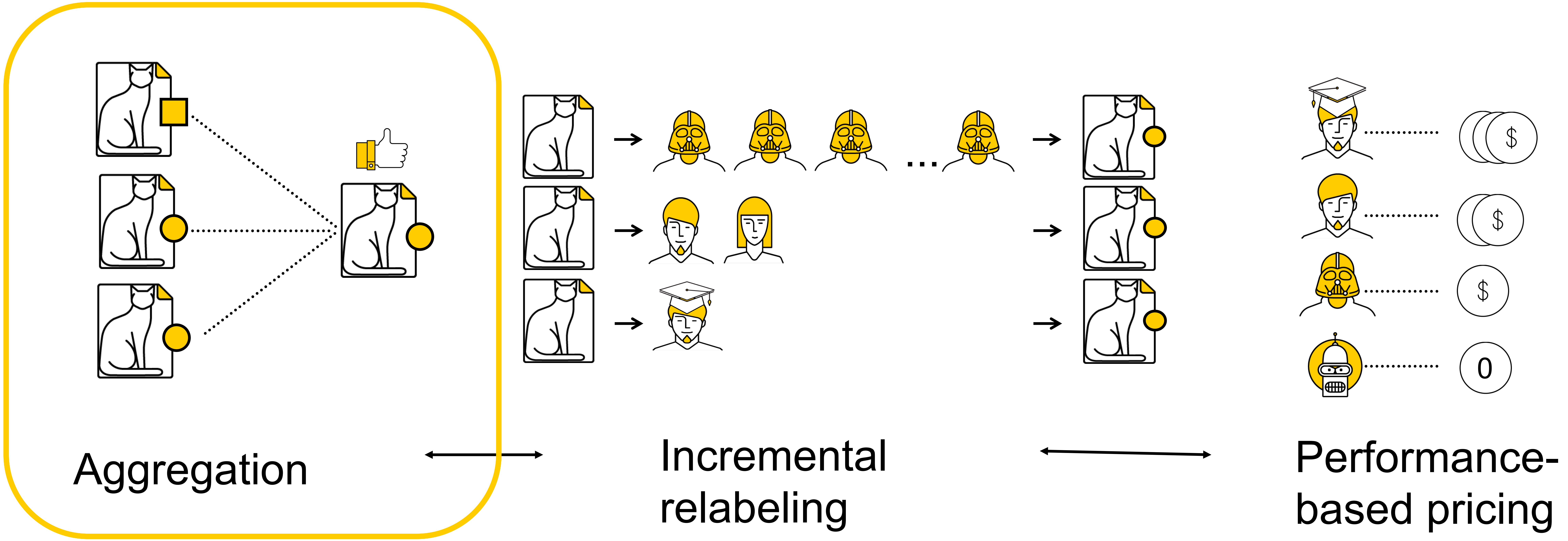


Labels with a maximal level of accuracy for a given budget

or

Labels of a chosen accuracy level for a minimal budget

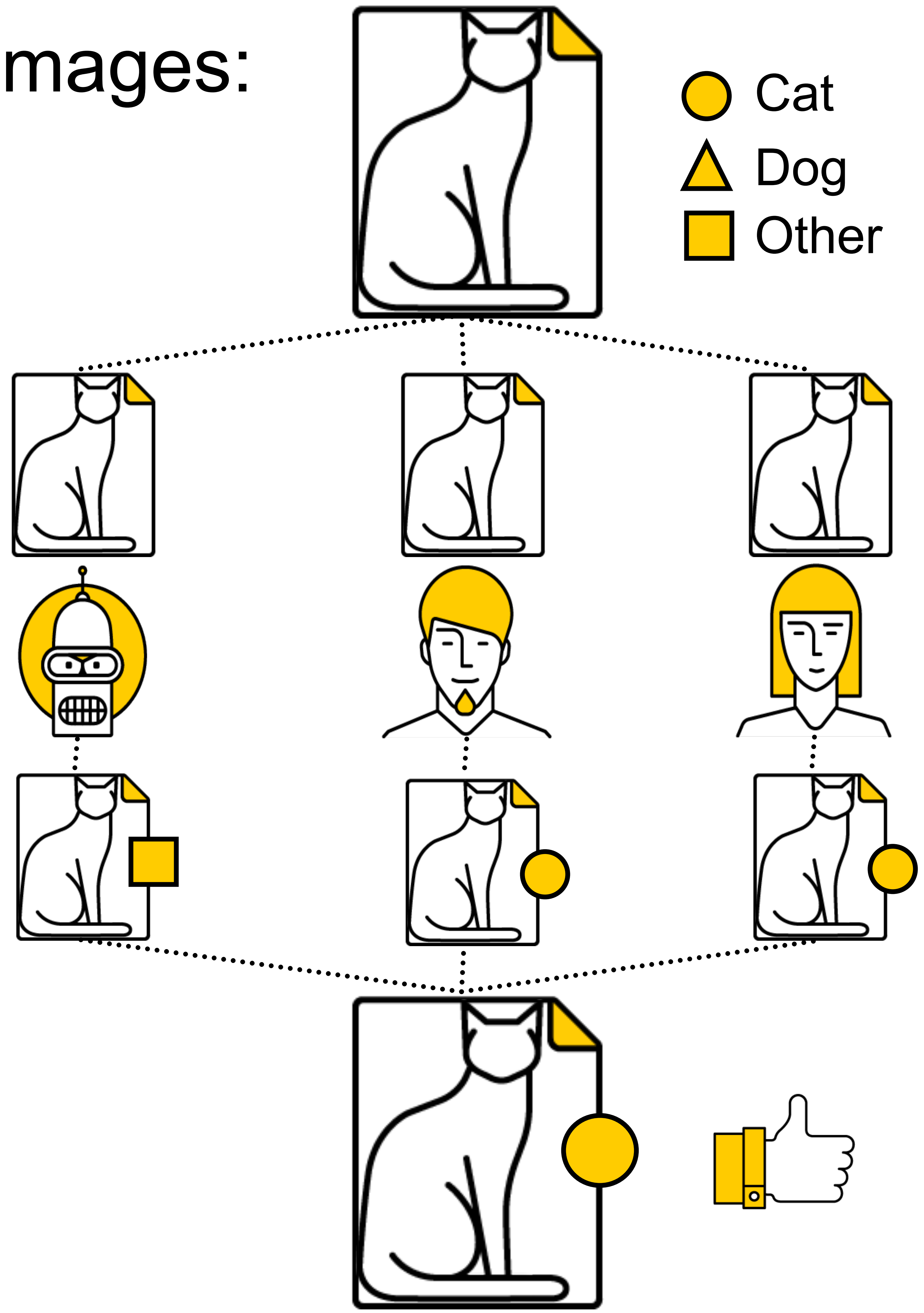
Key components of labeling with crowds



Aggregation

Labeling data with crowds

Classify images:

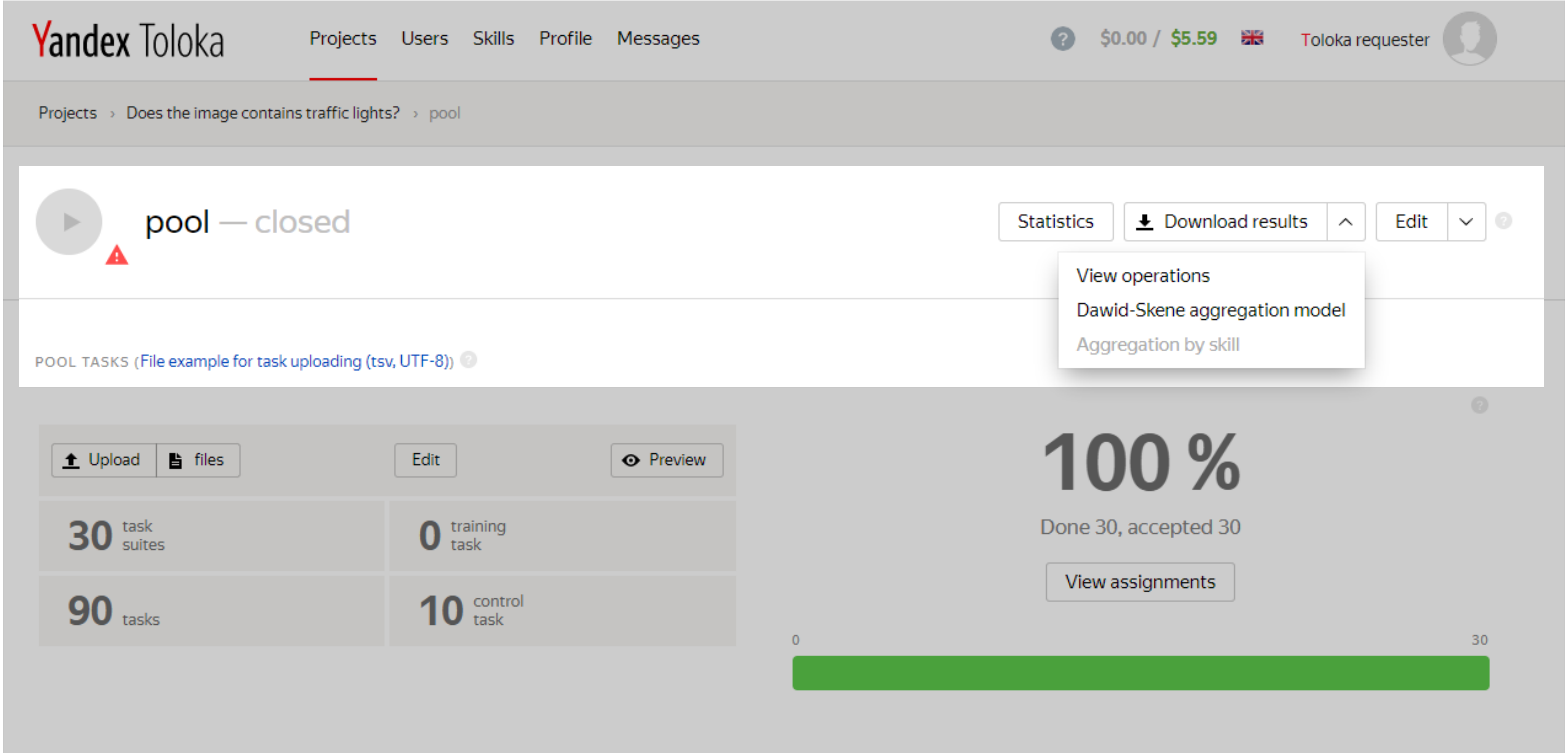


Upload multiple copies of each object to label

Performers assign noisy labels to objects

Aggregate multiple labels for each object into a more reliable one

Process results



Multiclass labels

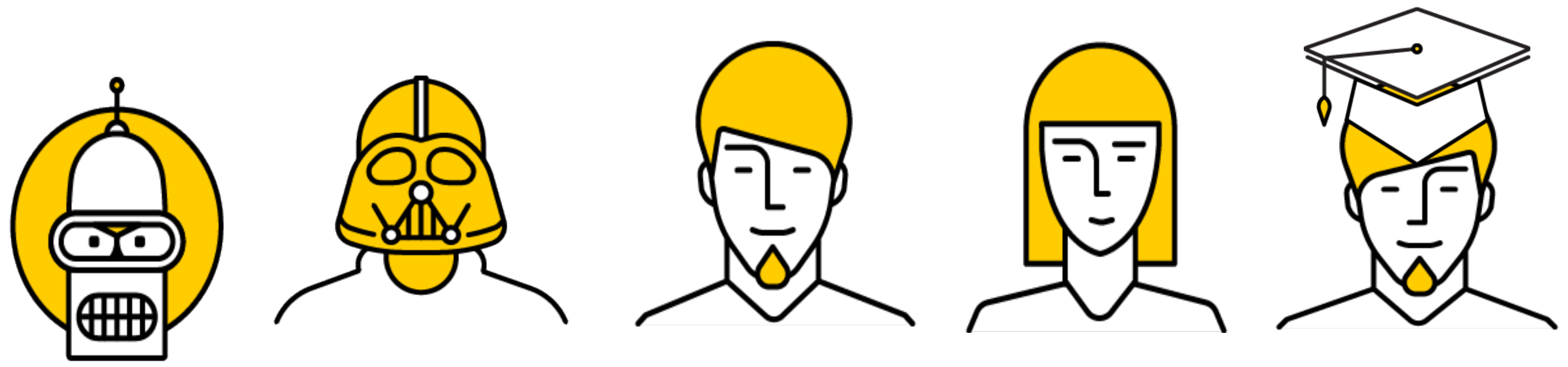
Project 1: Filter images

Are there shoes in the picture?

Yes

No

Notation

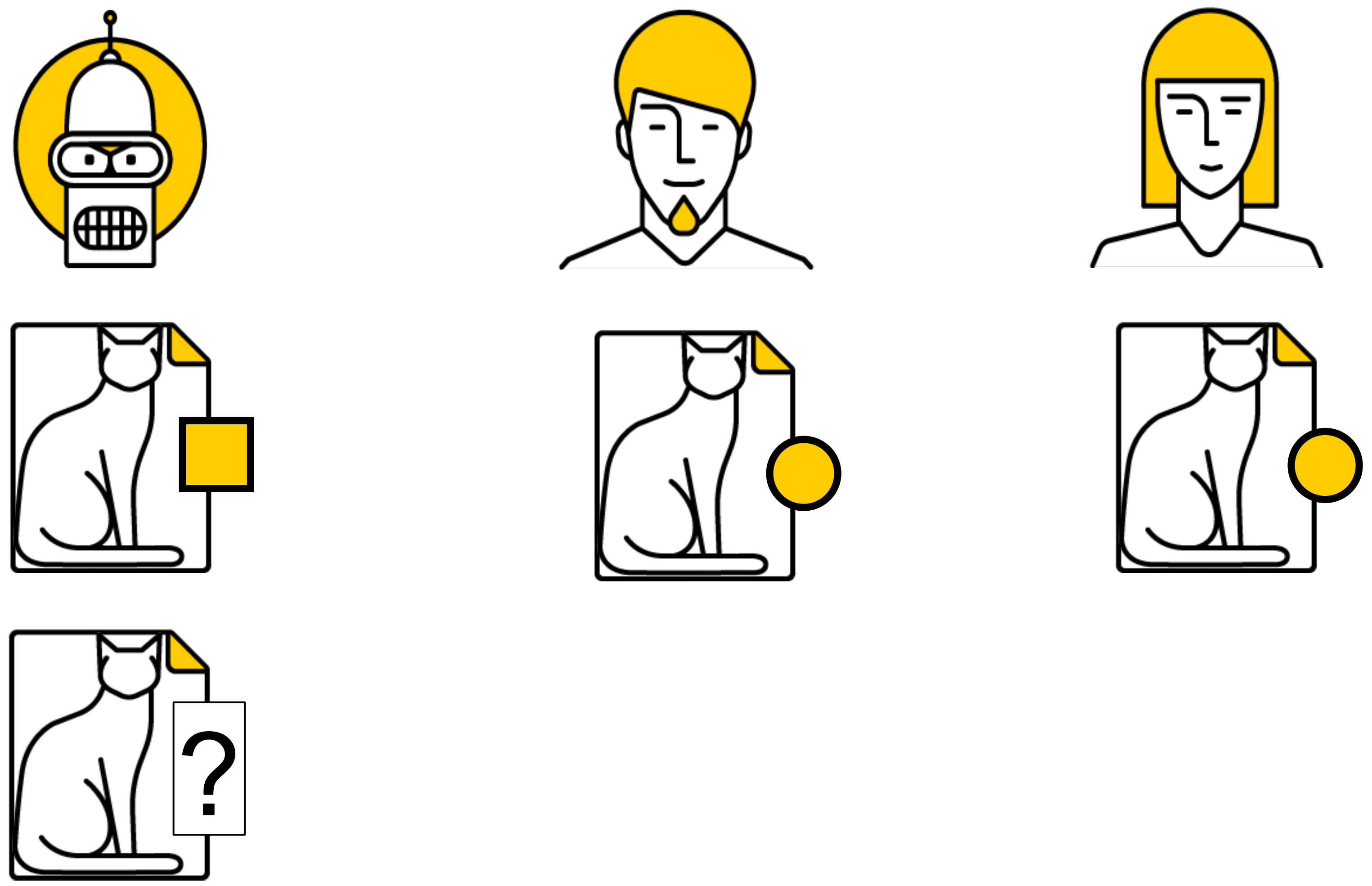
- Categories $k \in \{1, \dots, K\}$. E.g.: Cat Dog Other
- Objects $j \in \{1, \dots, J\}$. E.g.: 
- Performers: $w \in \{1, \dots, W\}$. E.g.: 
  - $W_j \subseteq \{1, \dots, W\}$ - performers labeled object j

The simplest aggregation: Majority Vote (MV)

The problem of aggregation:

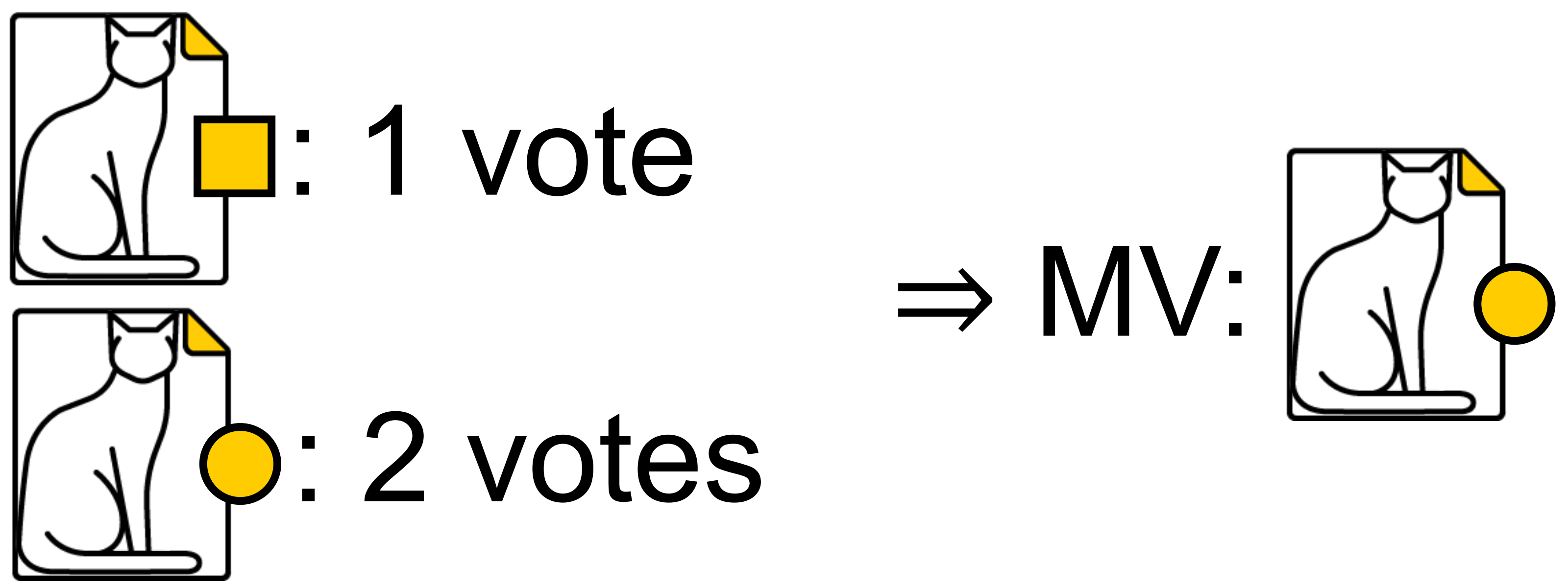
- Observe noisy labels

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



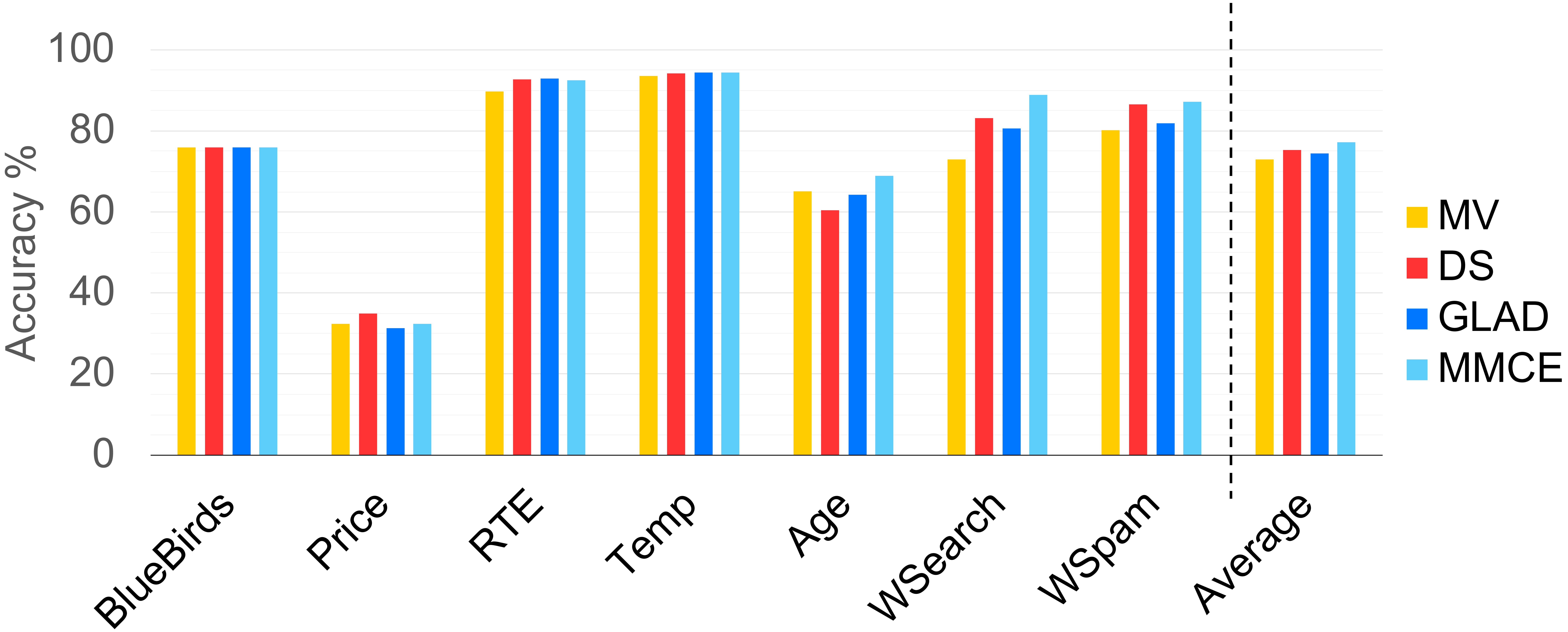
- Recover true labels $\mathbf{z} = \{z_j \mid j = 1, \dots, J\}$

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

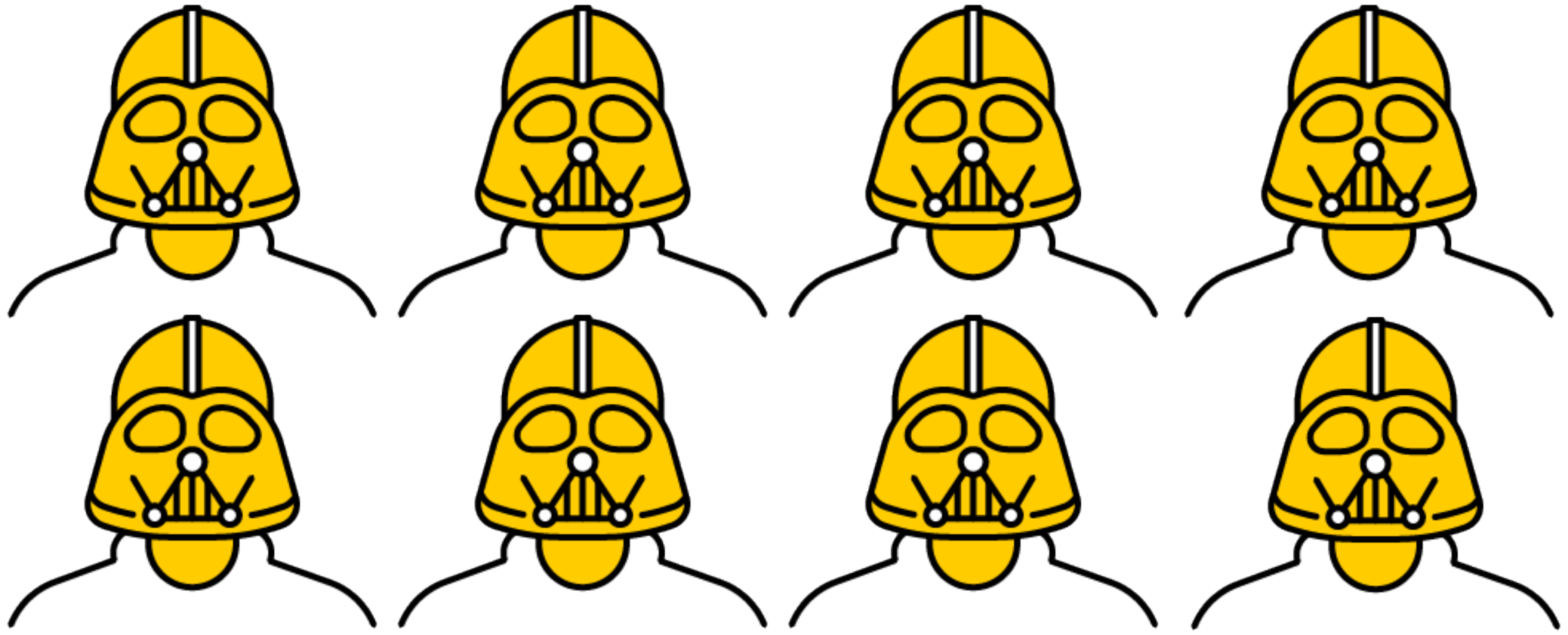
Performance of MV vs other methods



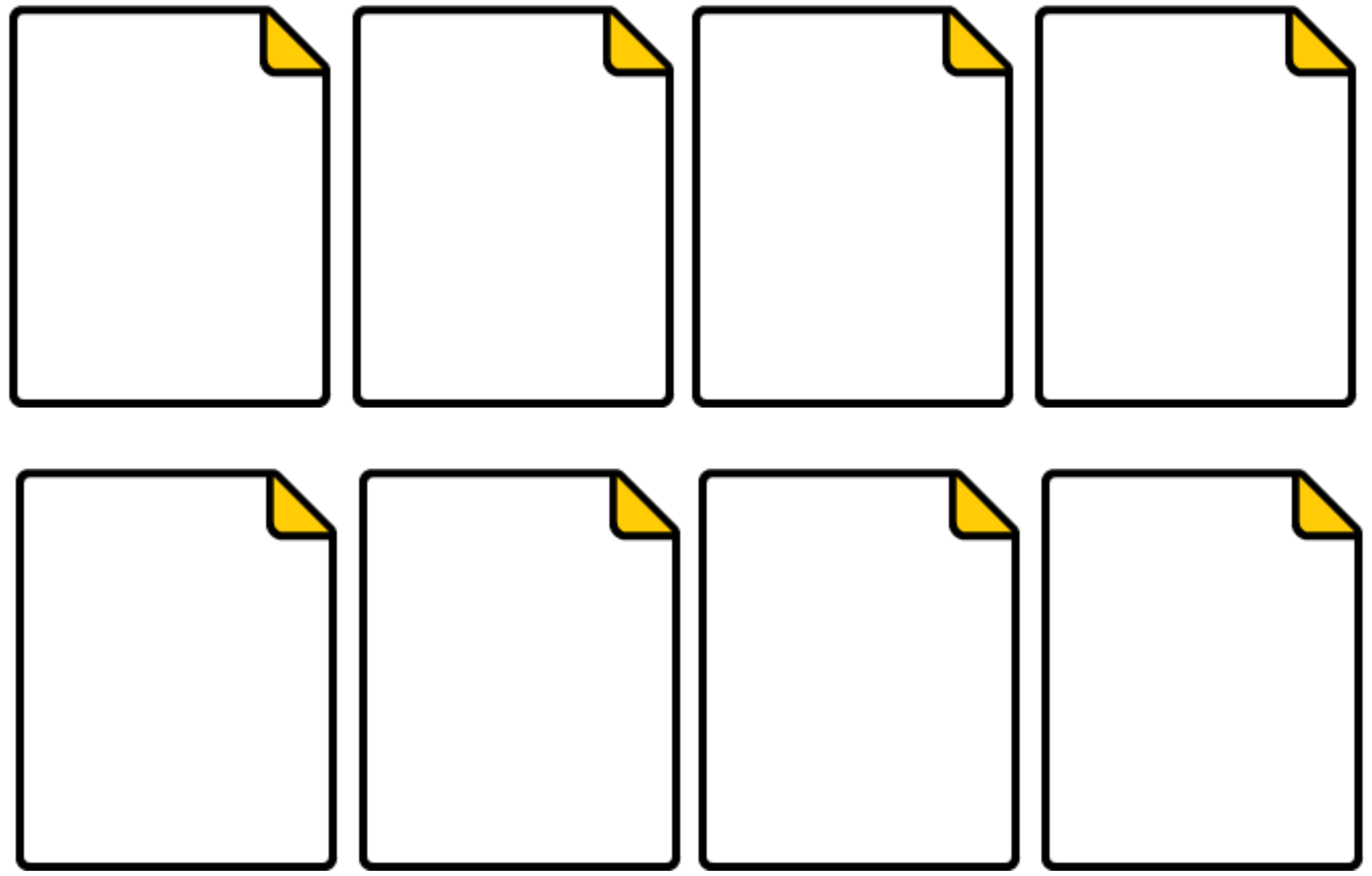
Zhou D. et al. Regularized minimax conditional entropy for crowdsourcing. 2015

Properties of MV

- All performers are treated similarly

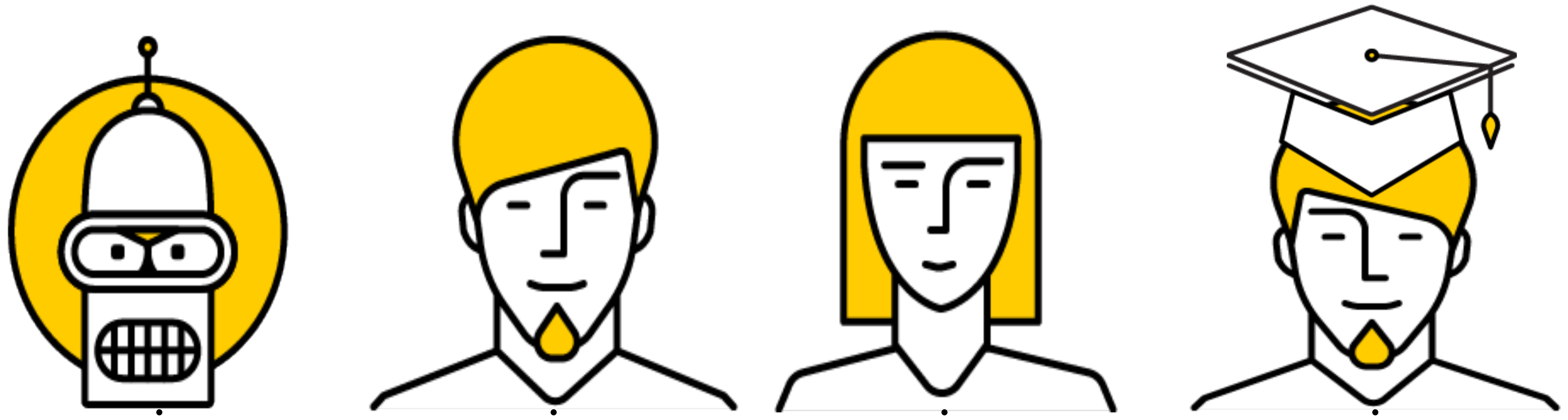


- All objects are treated similarly



Advanced aggregation: performers and objects

- Parameterize expertise of performers by e^w

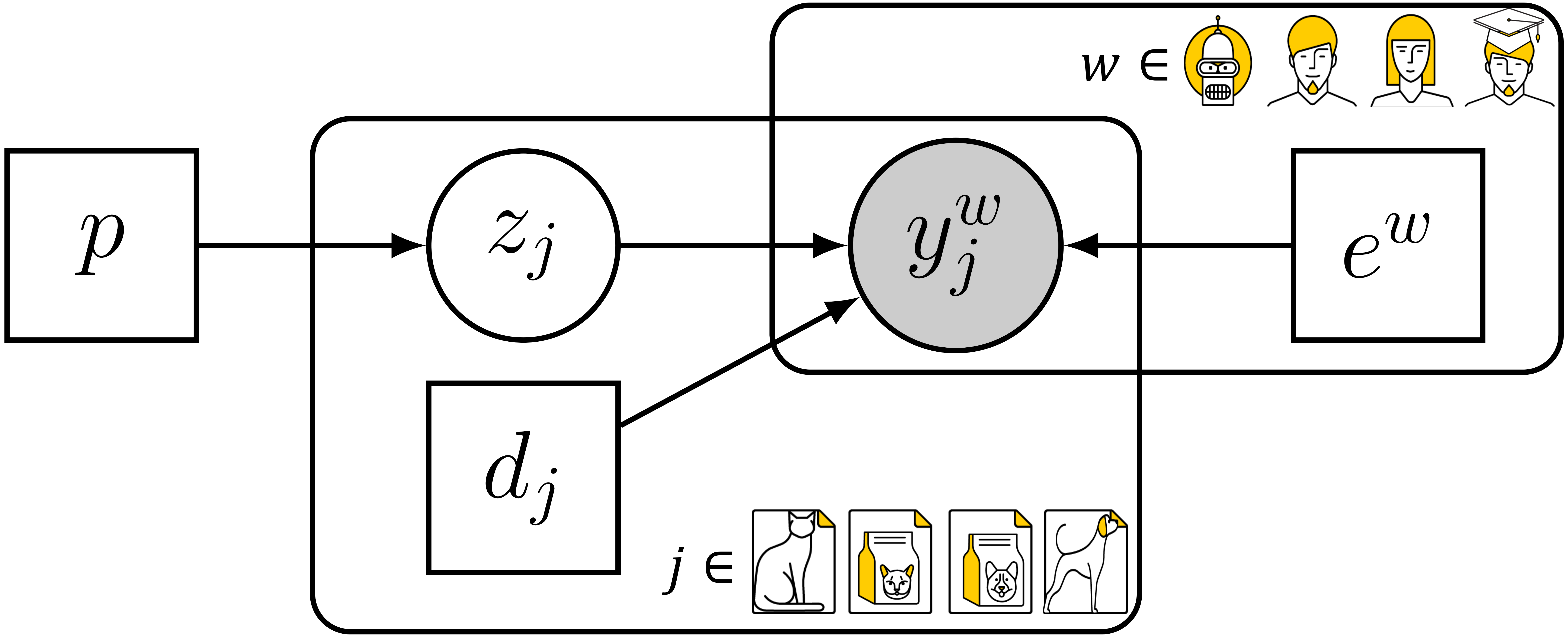


e^{w_1} e^{w_2} e^{w_3} e^{w_4}

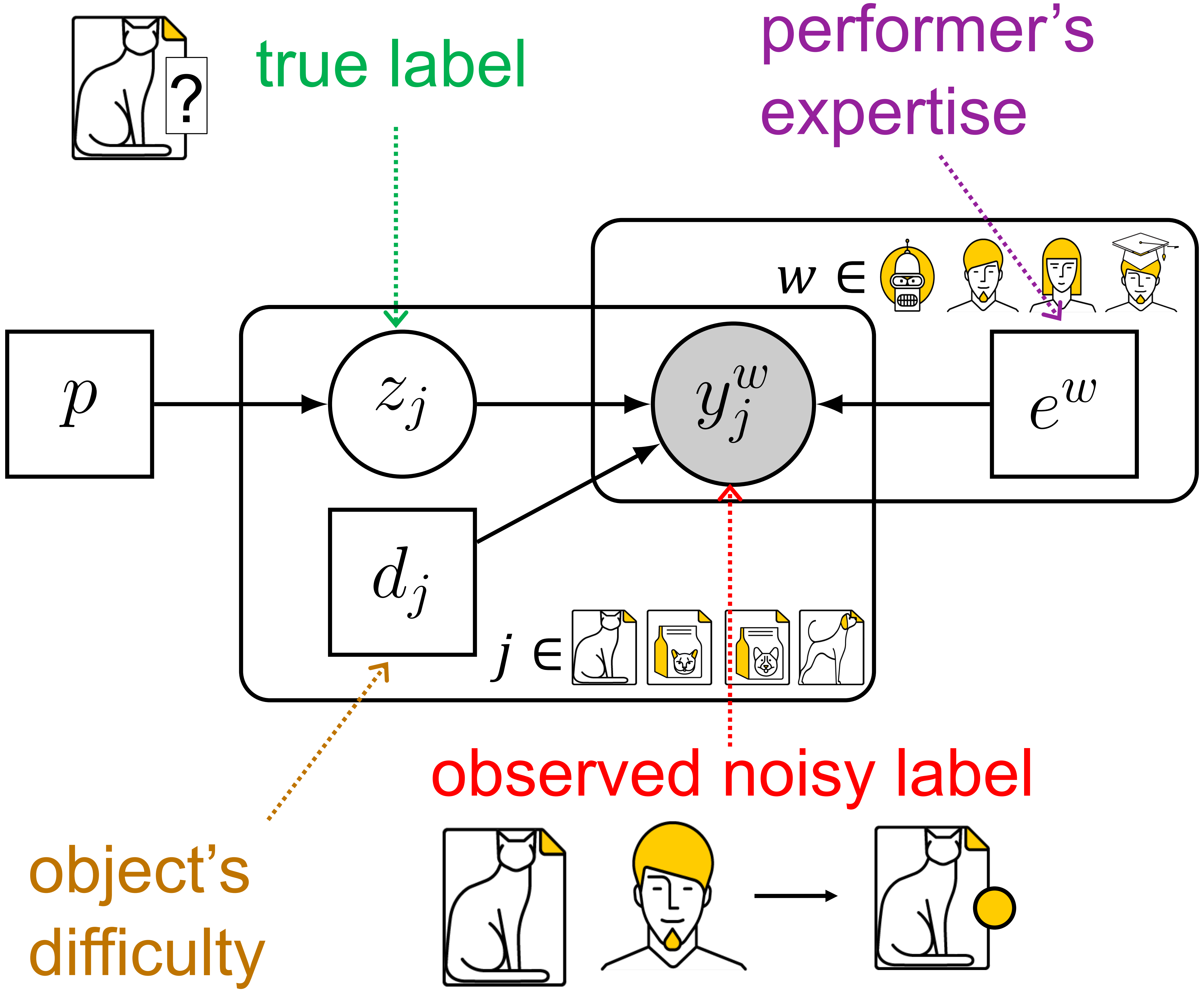
- Parameterize difficulty of objects by d_j

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

Advanced aggregation: latent label models

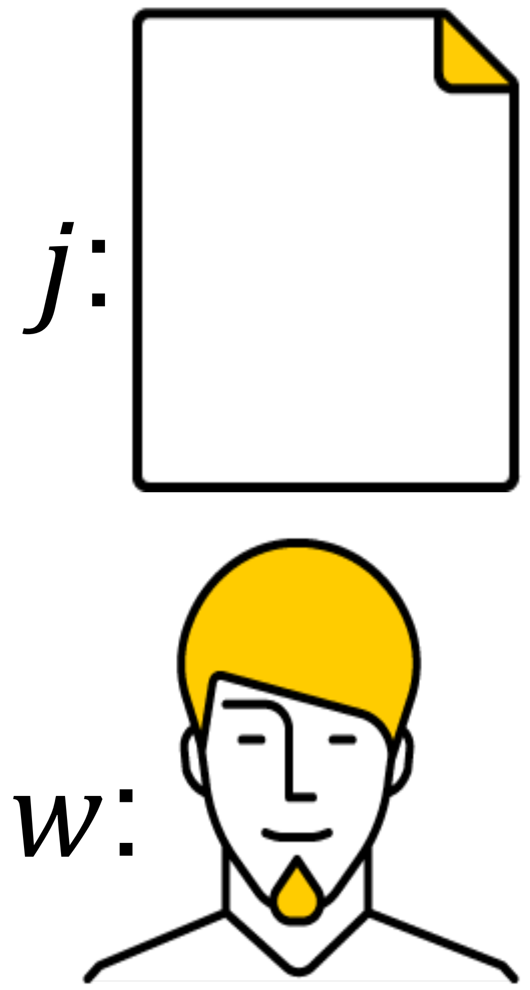


Latent label models: noisy label model



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

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

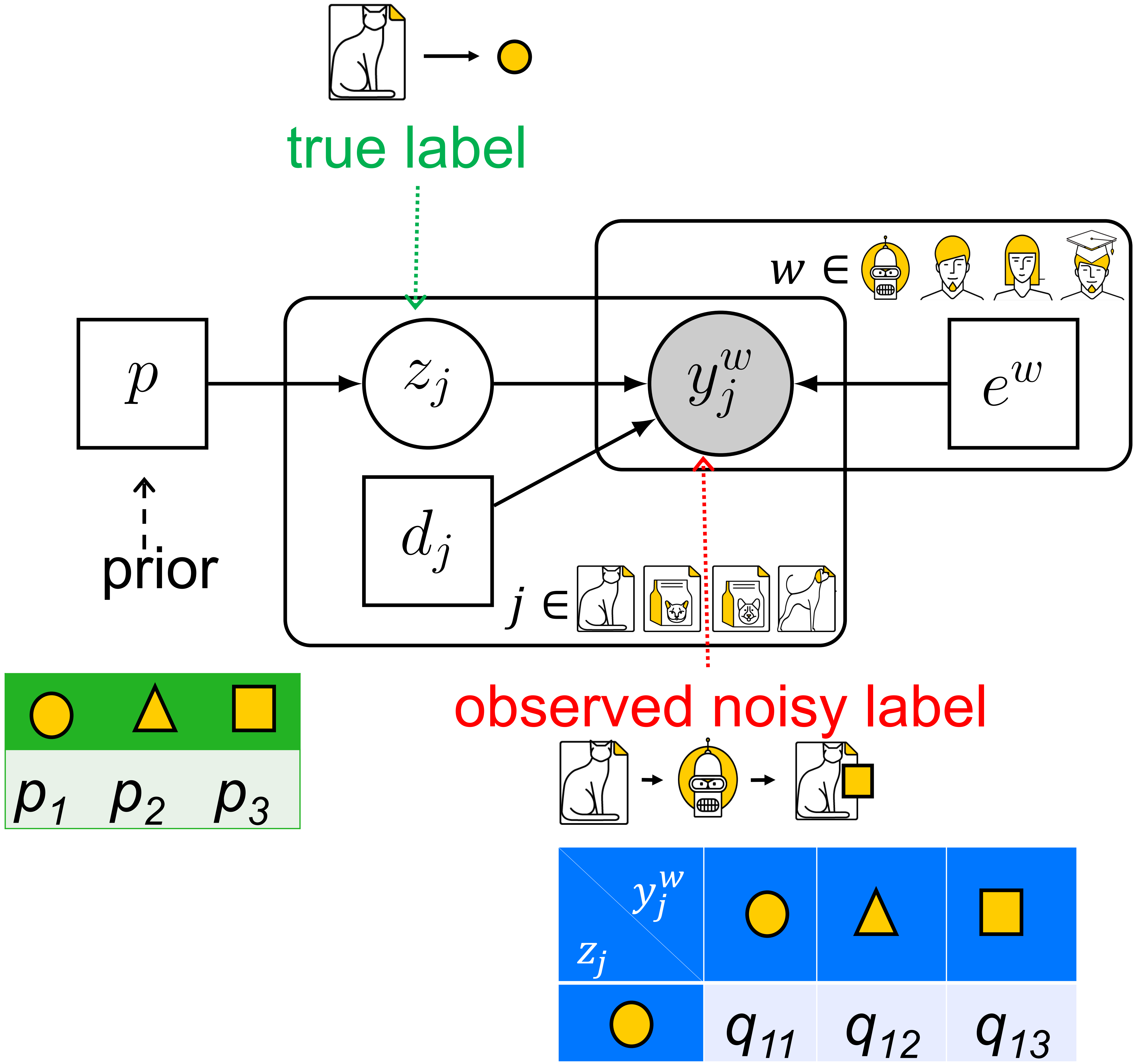


M_j^w:

True \ Noisy	○	△	□
○	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:

1. Sample z_j from a distribution $P_Z(p)$
2. 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 $\mathrm{Mult}(\cdot)$

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 $\mathbf{y}$ and $\mathbf{z}$ under the latent label model:

observed noisy label

$$L\left(\{z_j\}_{j=1}^{J}, p, \{d_j\}_{j=1}^{J}, \{e^w\}_{w=1}^{W}\right) = \prod_{j\in J} \underbrace{\sum_{z_j\in\{1,\dots,K\}} \Pr\left(z_j|p\right) \prod_{w\in W_j} \Pr\left(y_j^w|z_j, d_j, e^w\right)}_{\text{likelihood of noisy and true labels for object } j}$$

latent true label

latent parameters

- 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**

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

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

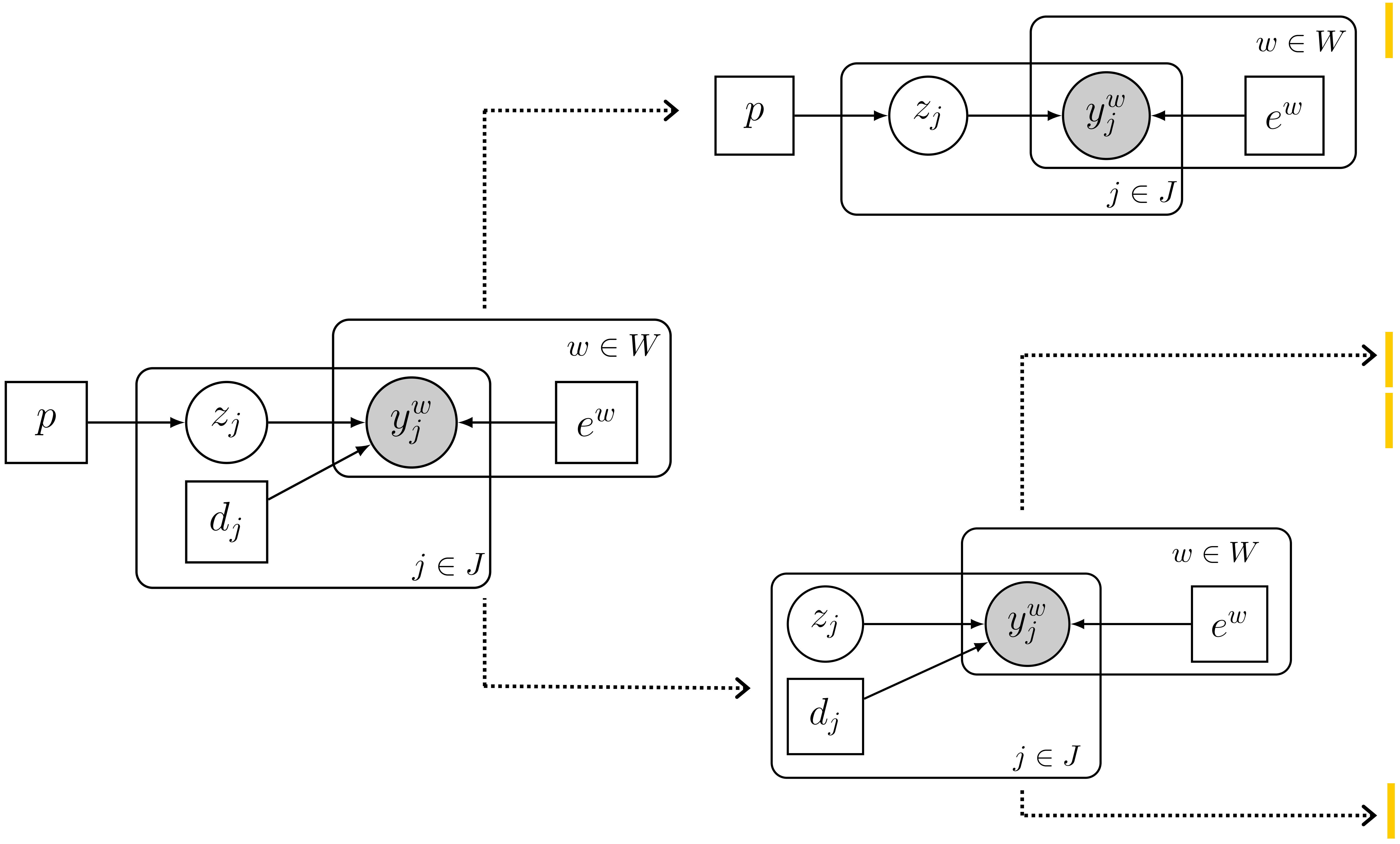
$$\hat{z}_j[c] = \Pr(Z_j = c | \mathbf{y}, p, \mathbf{d}, \mathbf{e}) \propto \Pr(Z_j = c | p) \prod_{w \in W_j} \Pr(y_j^w | Z_j = c, d_j, e^w)$$

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

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

  - Analytical solutions
  - Gradient descent

Latent label model (LLM): special cases



Dawid and Skene model (DS):

- categories are different
- objects are similar
- performers are different

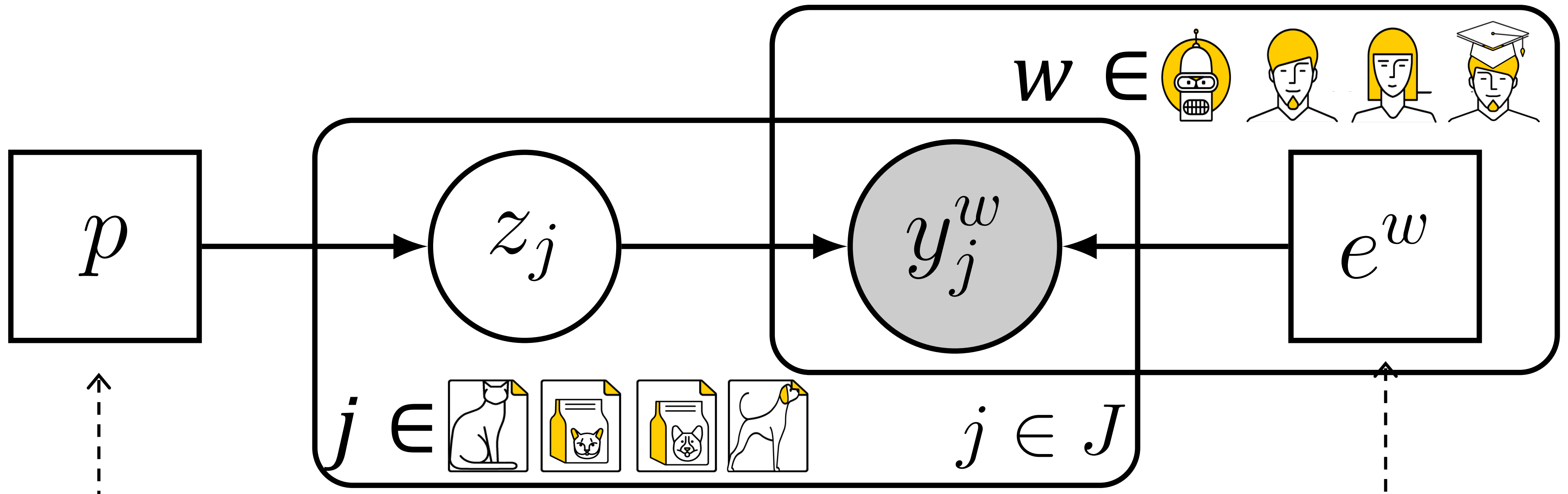
Generative model of labels, abilities, and difficulties (GLAD):

- 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)

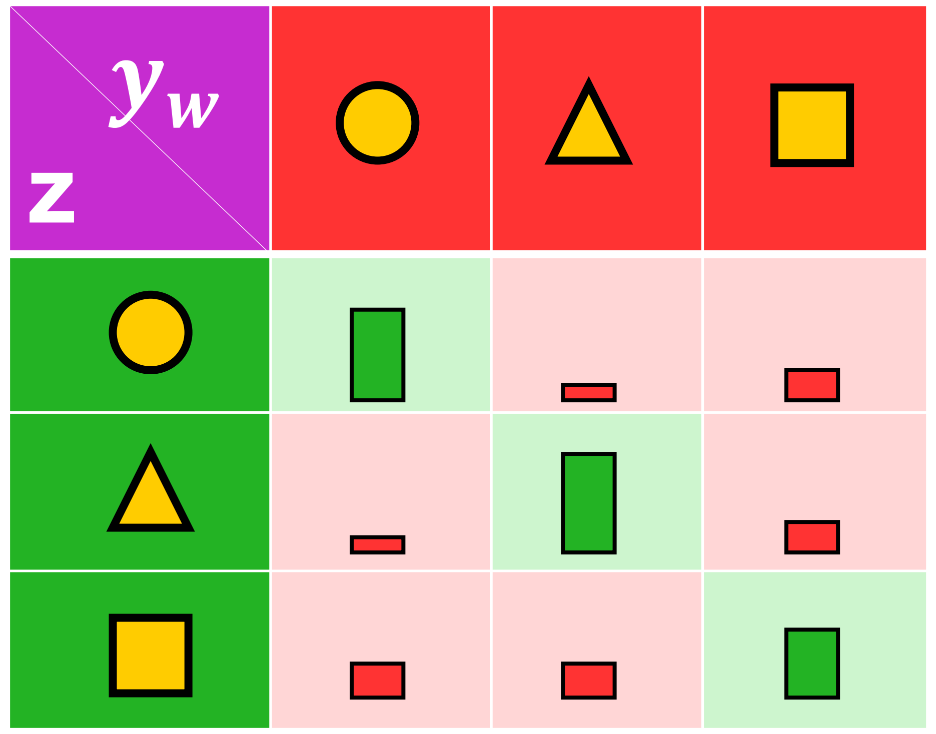


class prior

performer's confusion matrix

LLM with parameters:

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



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

DS: parameters optimization

- **E-step**:

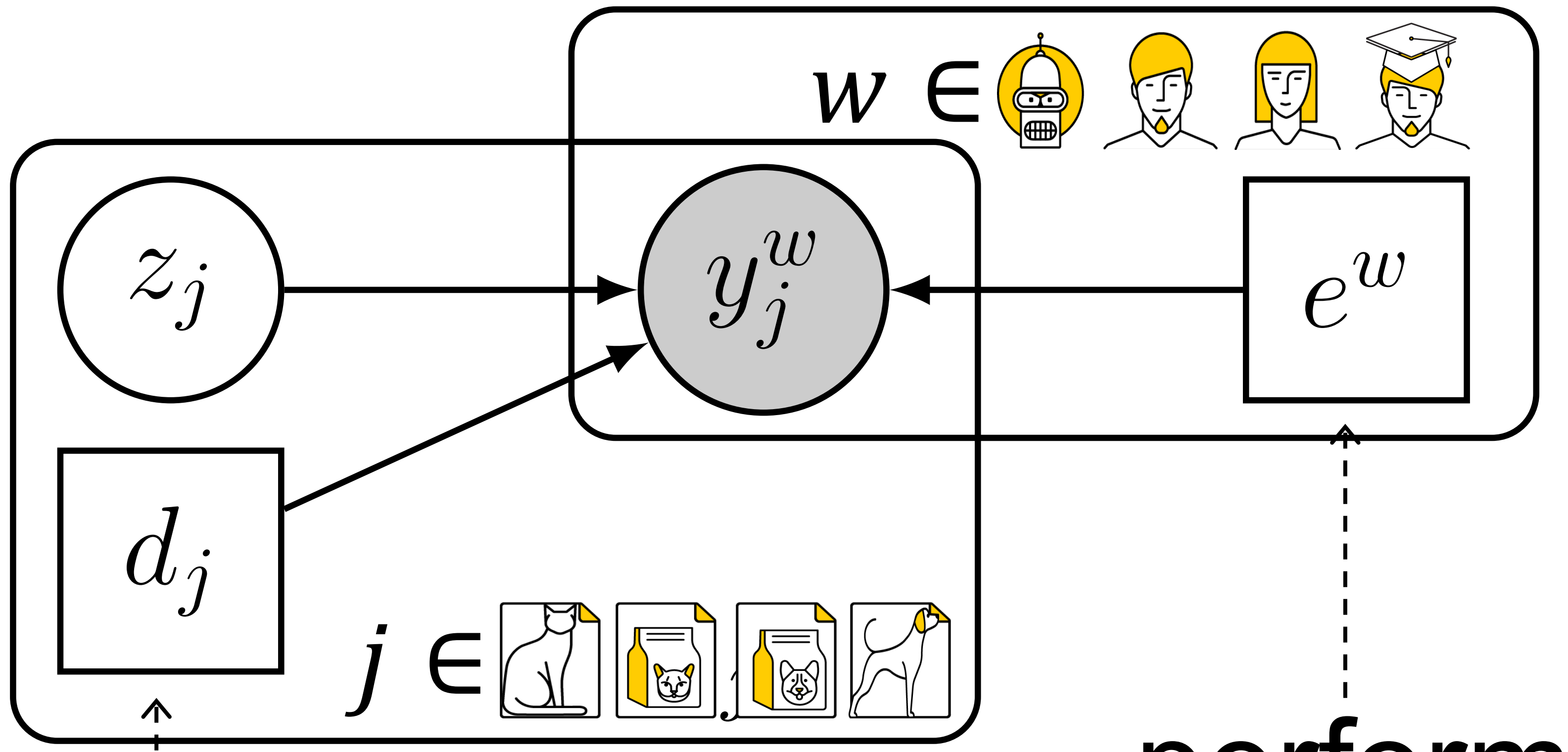
$$\hat{z}_j[c] = \frac{p[c]\ \prod_{w \in W_j} e^w\left[c, y_j^w\right]}{\sum_k p[k]\ \prod_{w \in W_j} e^w\left[k, y_j^w\right]}, \qquad c = 1, \ldots, K$$

- **M-step**: Analytical solution

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

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

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



object's inverse difficulty

performer's ability

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

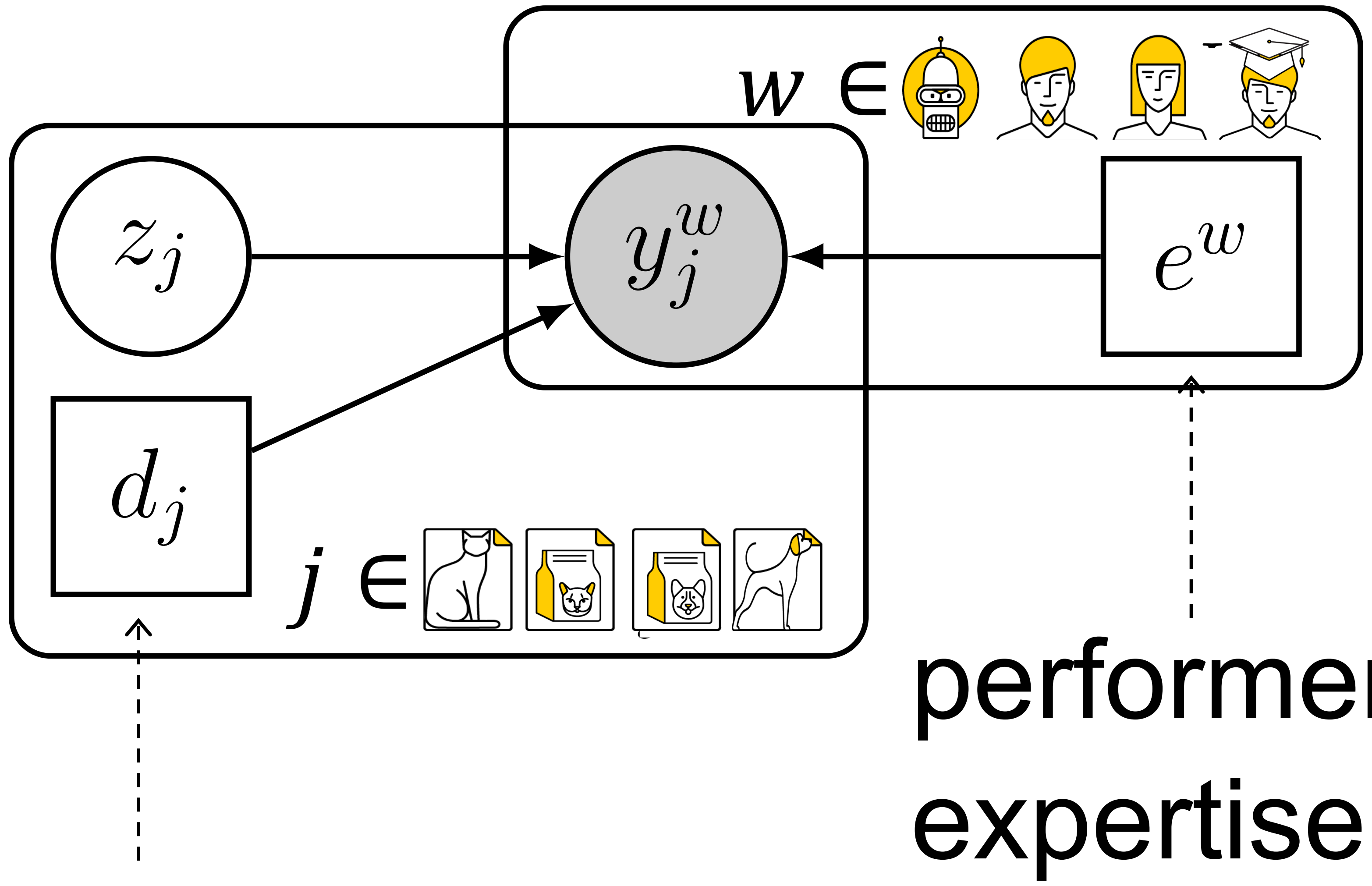
- Let $a(w,j) = \frac{1}{1+\exp(-e^w d_j)}$ and $\mathrm{P}(z_j)$ be a predefined prior (e.g., $\mathrm{P}(z_j) = {}^{1}/_{K}$)
- **E-step**:

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

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

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

MiniMax Conditional Entropy model (MMCE)



object's confusability matrix

performer's expertise matrix

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

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

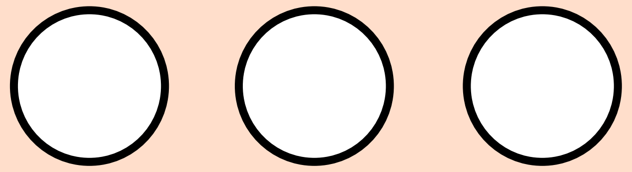
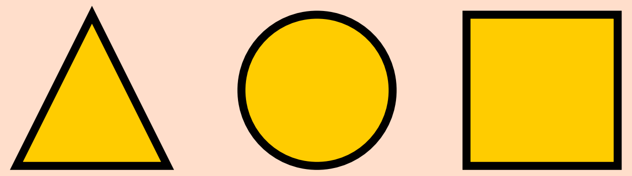
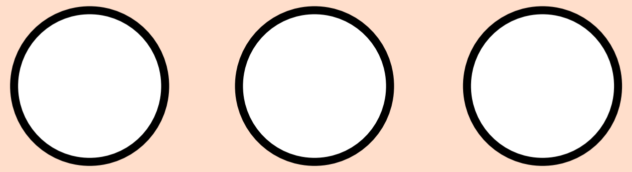
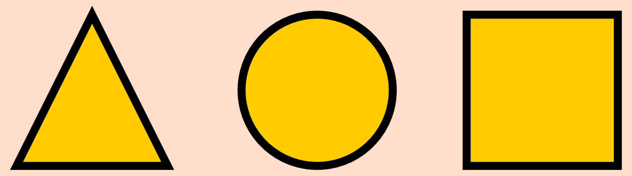
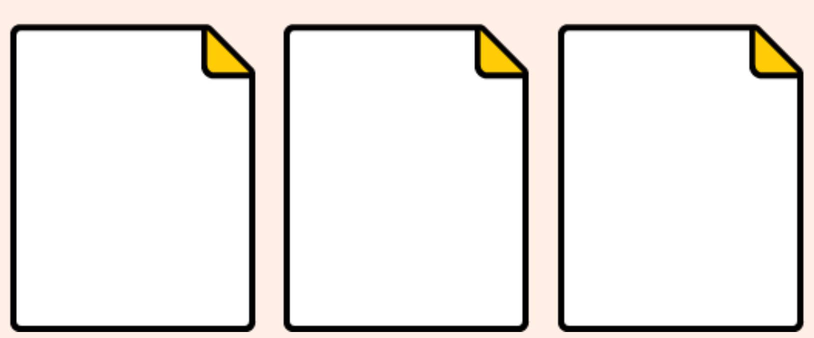
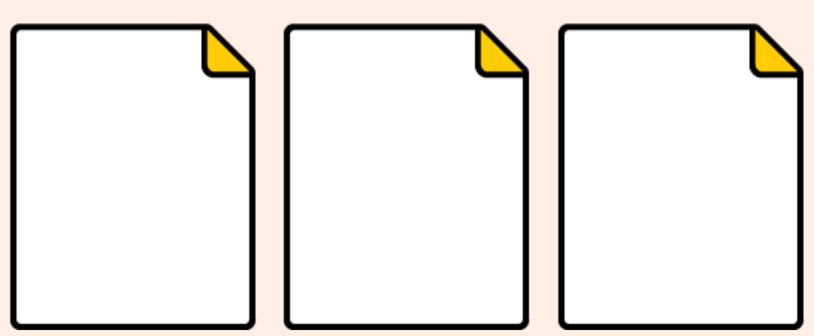
LLM with parameters:

- d_j – matrix of size $K \times K$
- e^w – matrix of size $K \times K$
- Noisy label model:

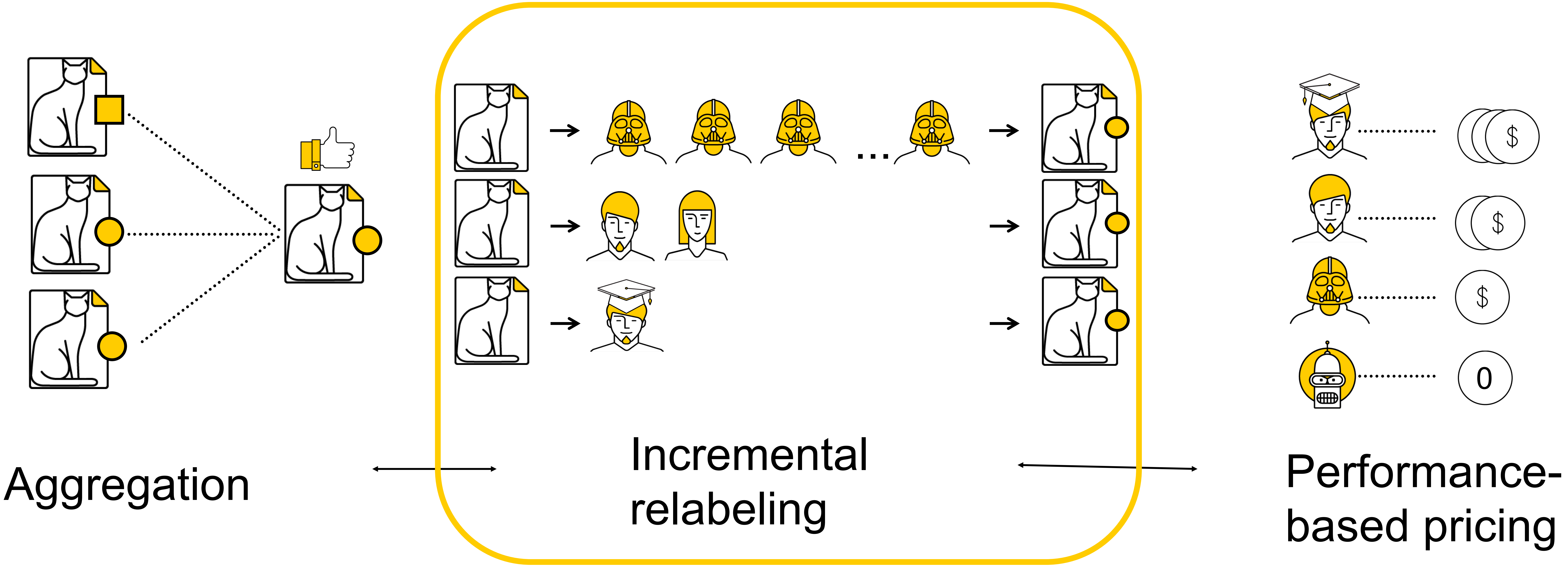
$$\Pr(Y_j^w = k | Z_j = c) = \exp(d_j[c,k] + e^w[c,k])$$

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

Summary of aggregation methods

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

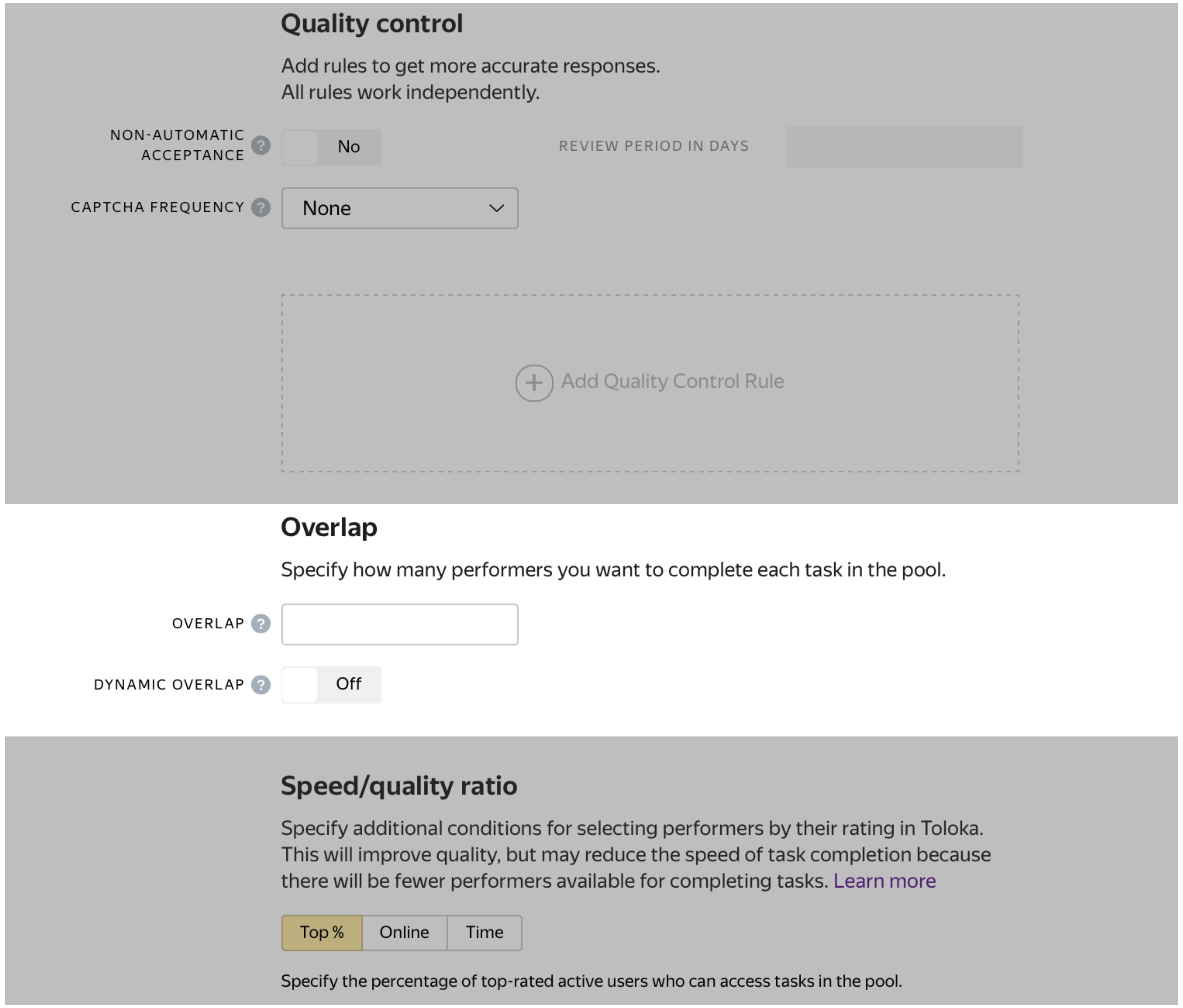
Key components of labeling with crowds



Incremental relabeling

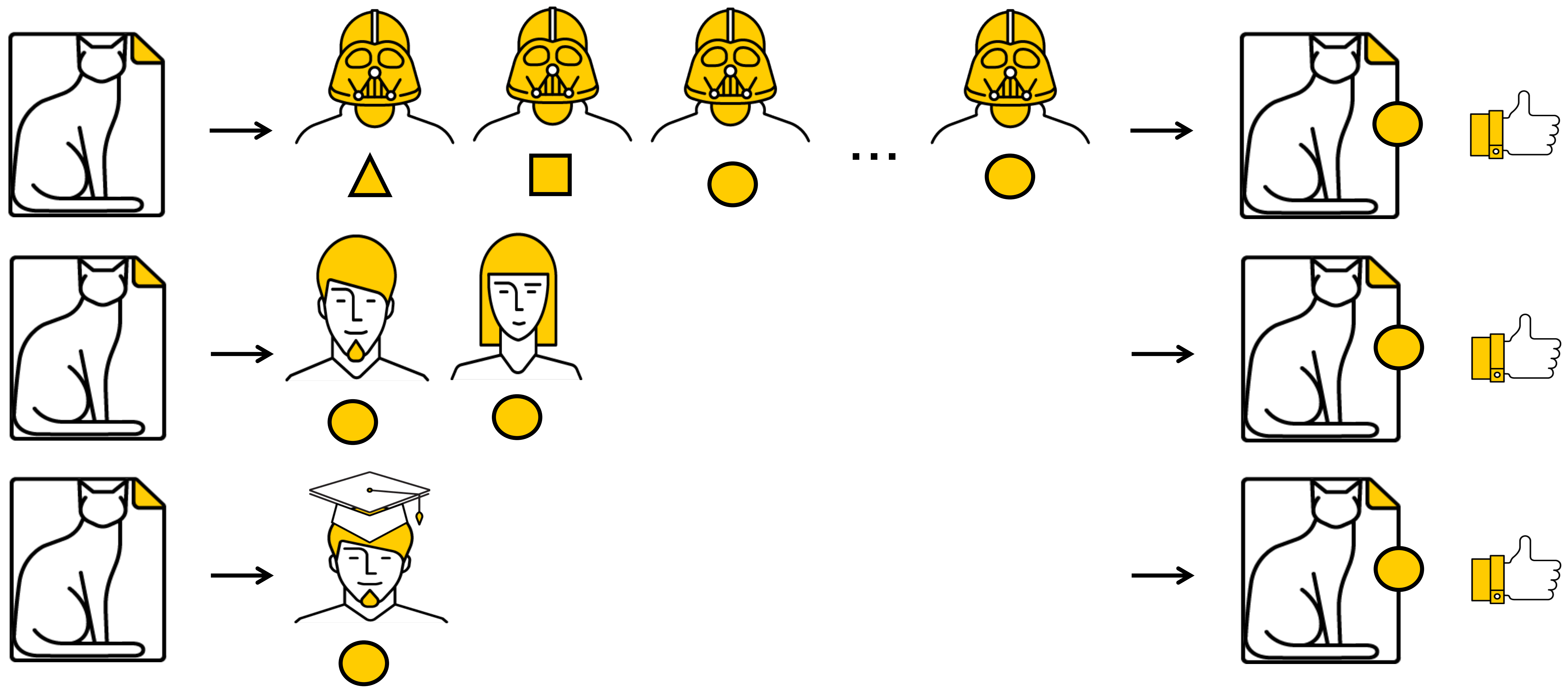
aka dynamic overlap

Pool settings: dynamic overlap



Incremental relabeling problem

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



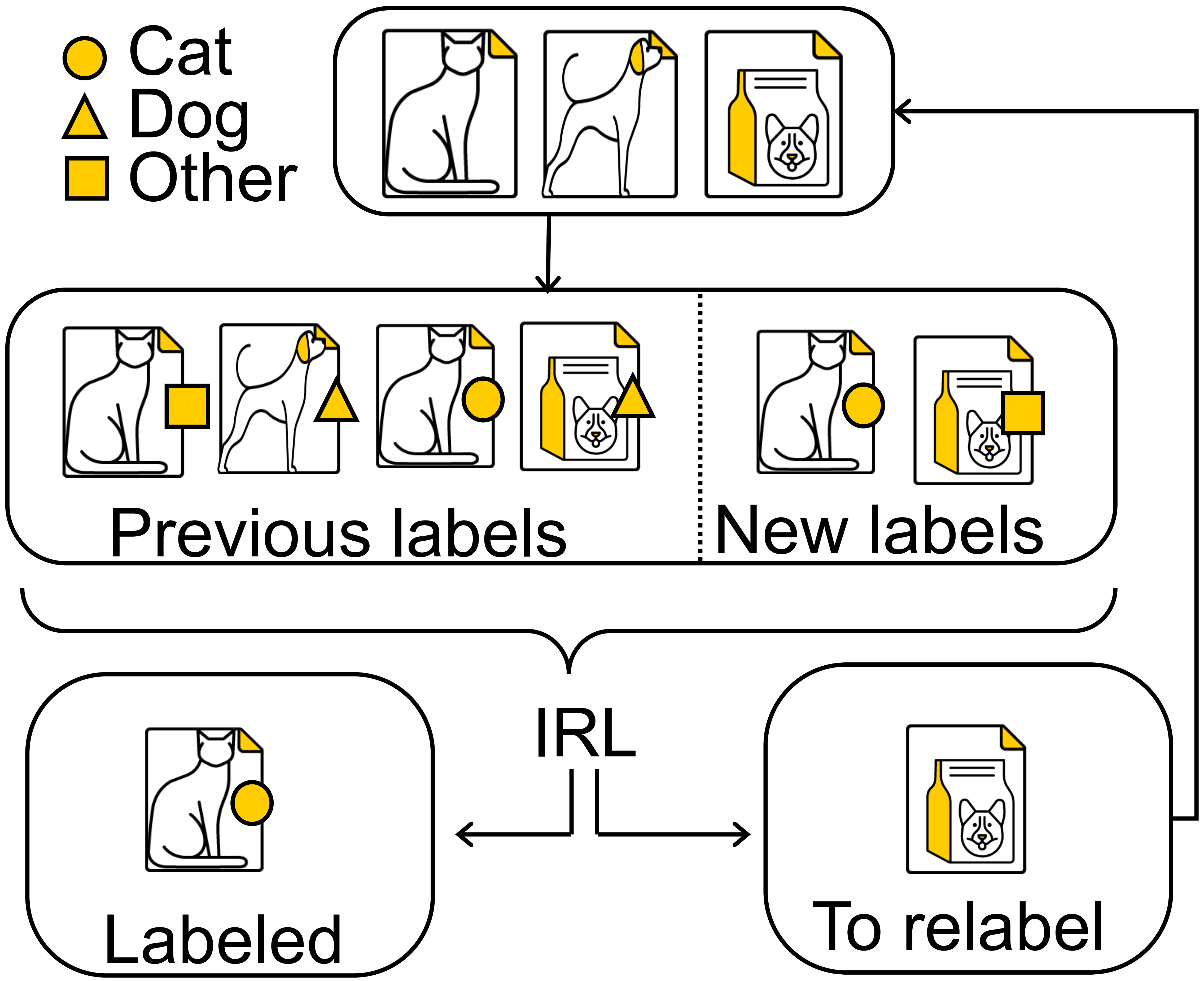
Incremental relabeling scheme (IRL)

Request a label for each object

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

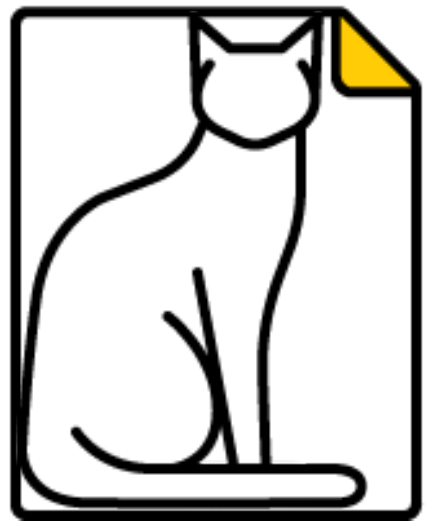
Decides:
(1) which objects are labeled
(2) which objects to relabel

Repeat until all tasks are labeled



Notations

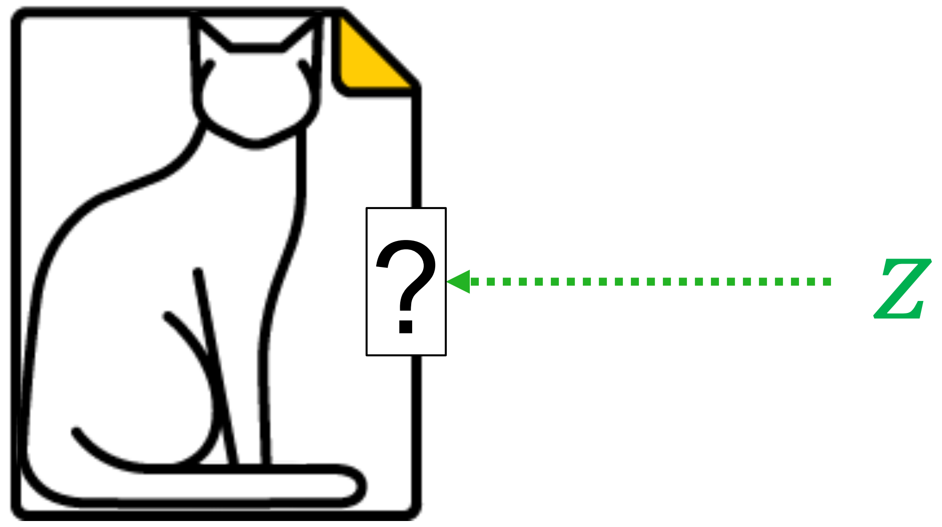
- Consider one object



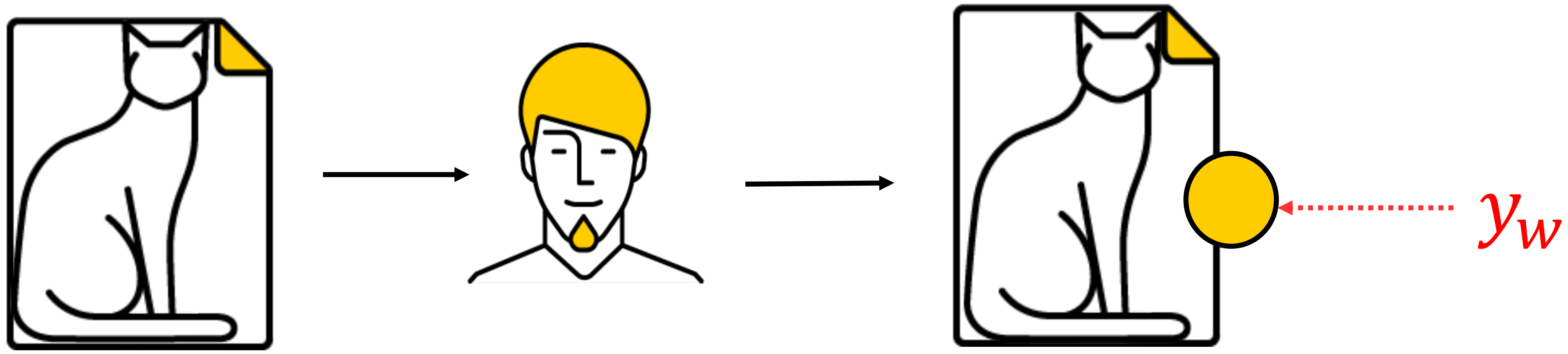
Classify images:

- Cat
- Dog
- Other

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



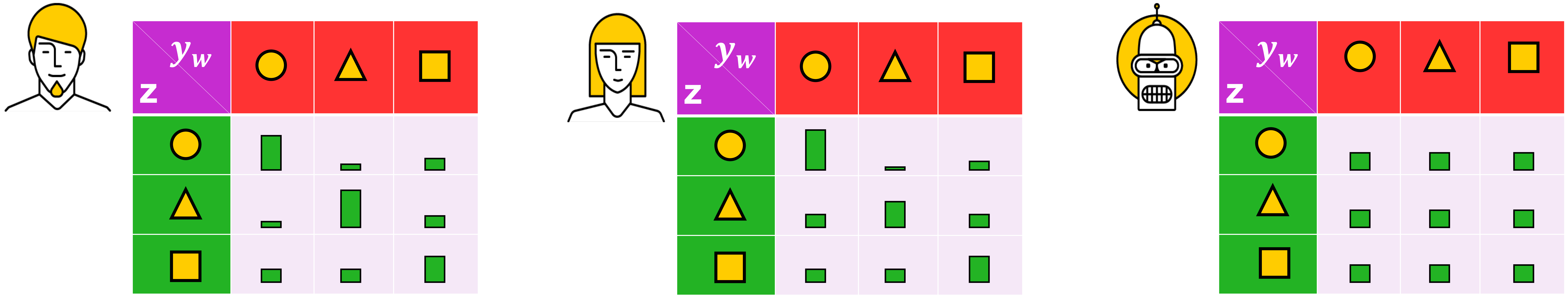
- $y_w \in \{1, \dots, K\}$ - observed noisy label from performer w:



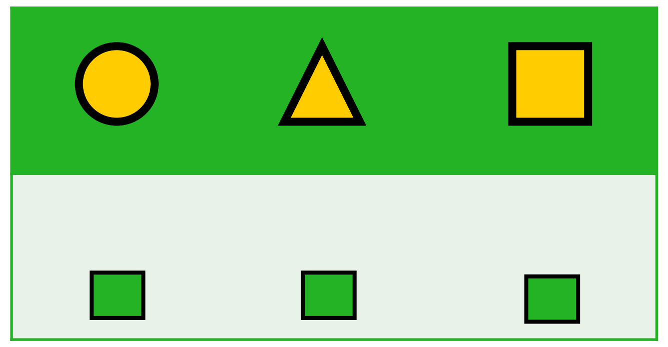
Notations

- Noisy label model for performer w:

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



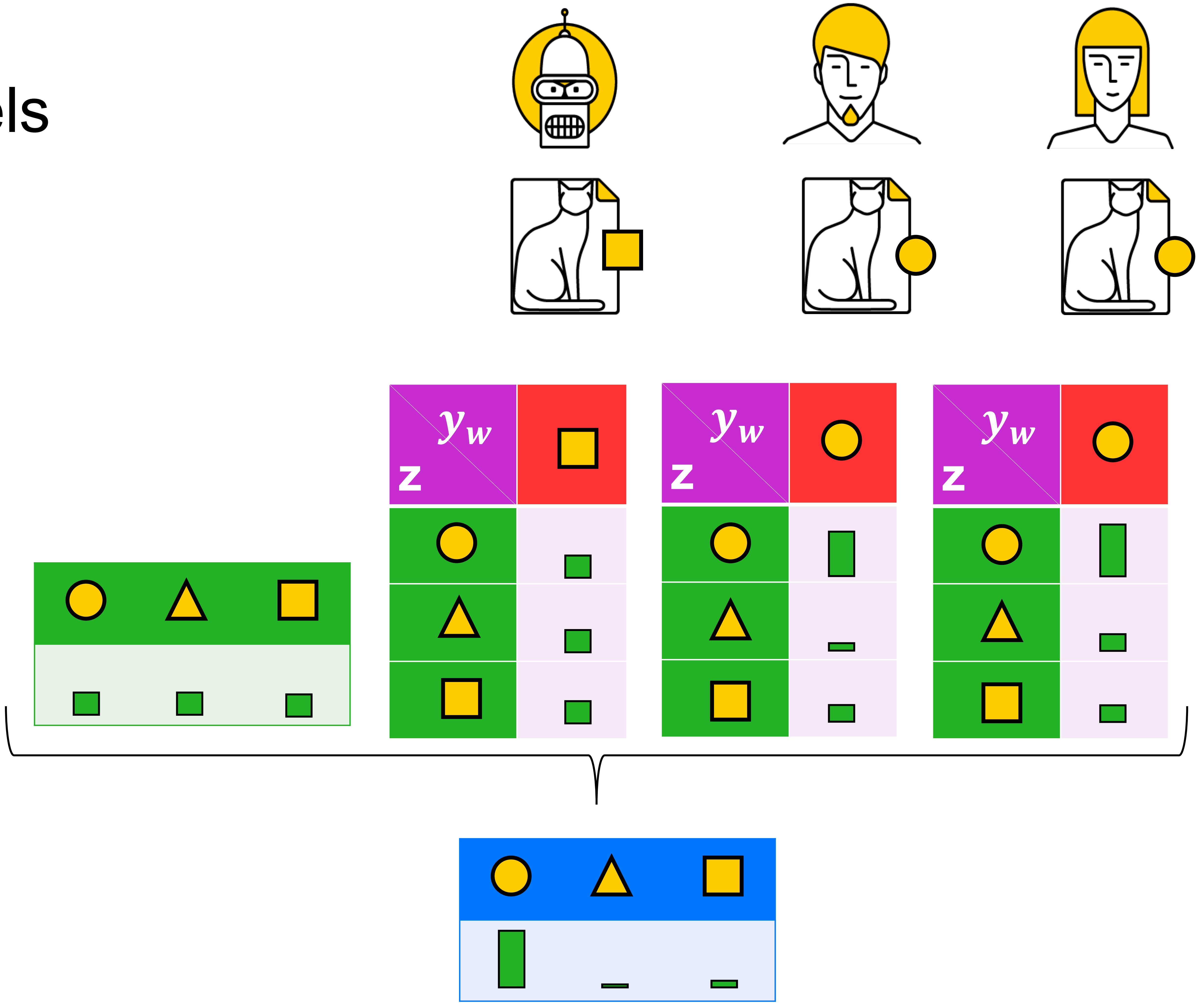
- Prior distribution: $\Pr(Z = k) = \ p_k$



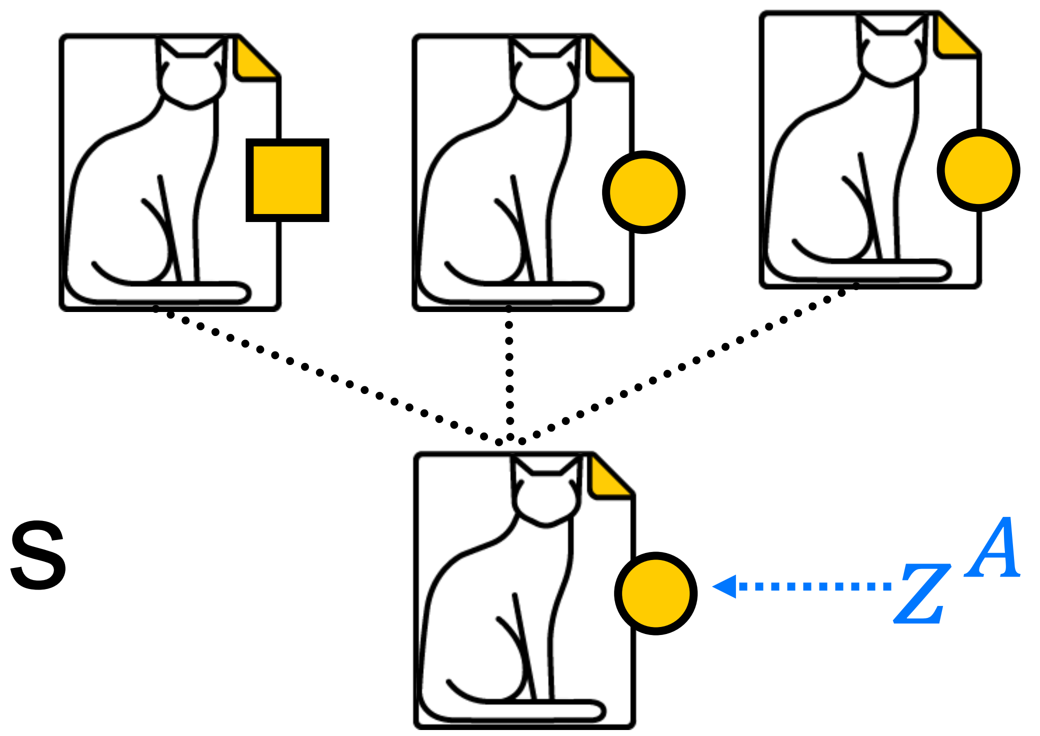
Posterior distribution

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

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



Expected accuracy of aggregated labels

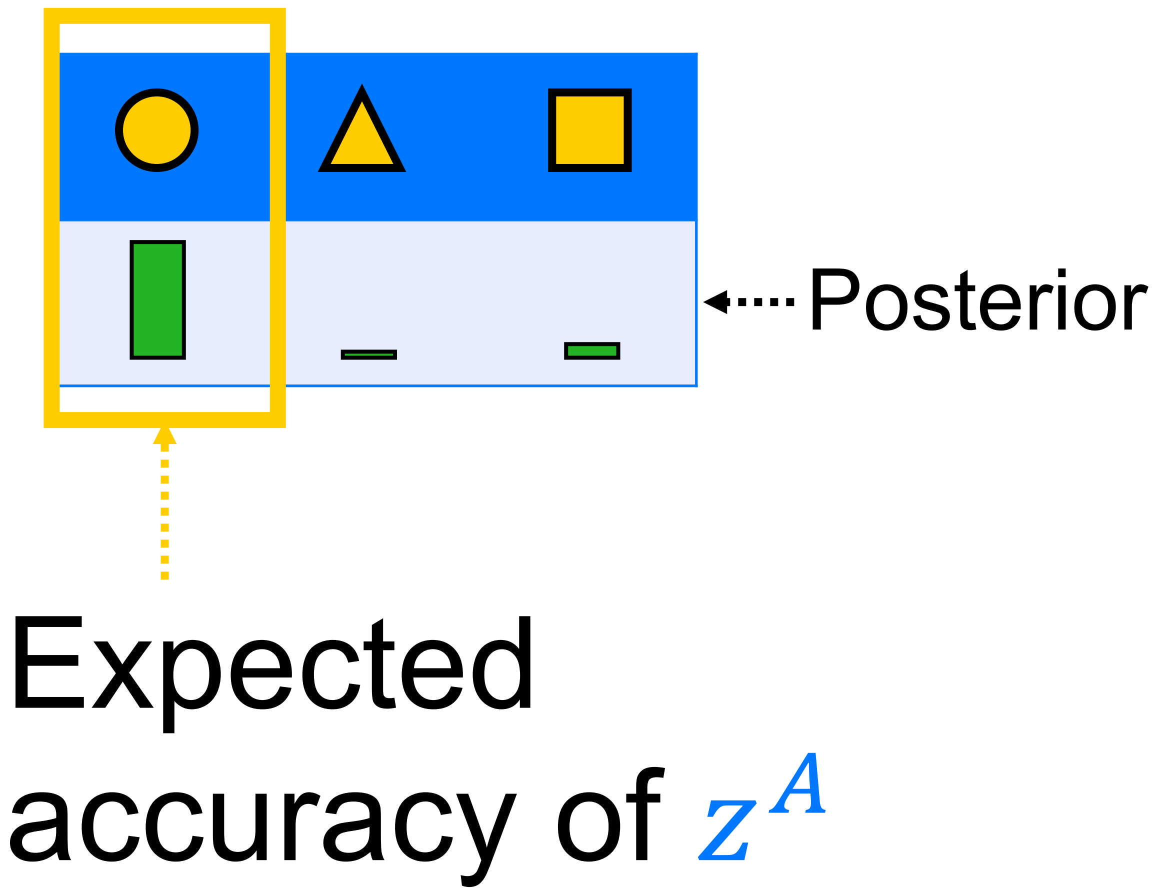


- Let A be an aggregation model, e.g. MV, DS, GLAD,…
- Denote aggregated label $z^A = A(\{y_{w_1}, \dots, y_{w_n}\})$
- Expected accuracy of aggregated labels given noisy labels is

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

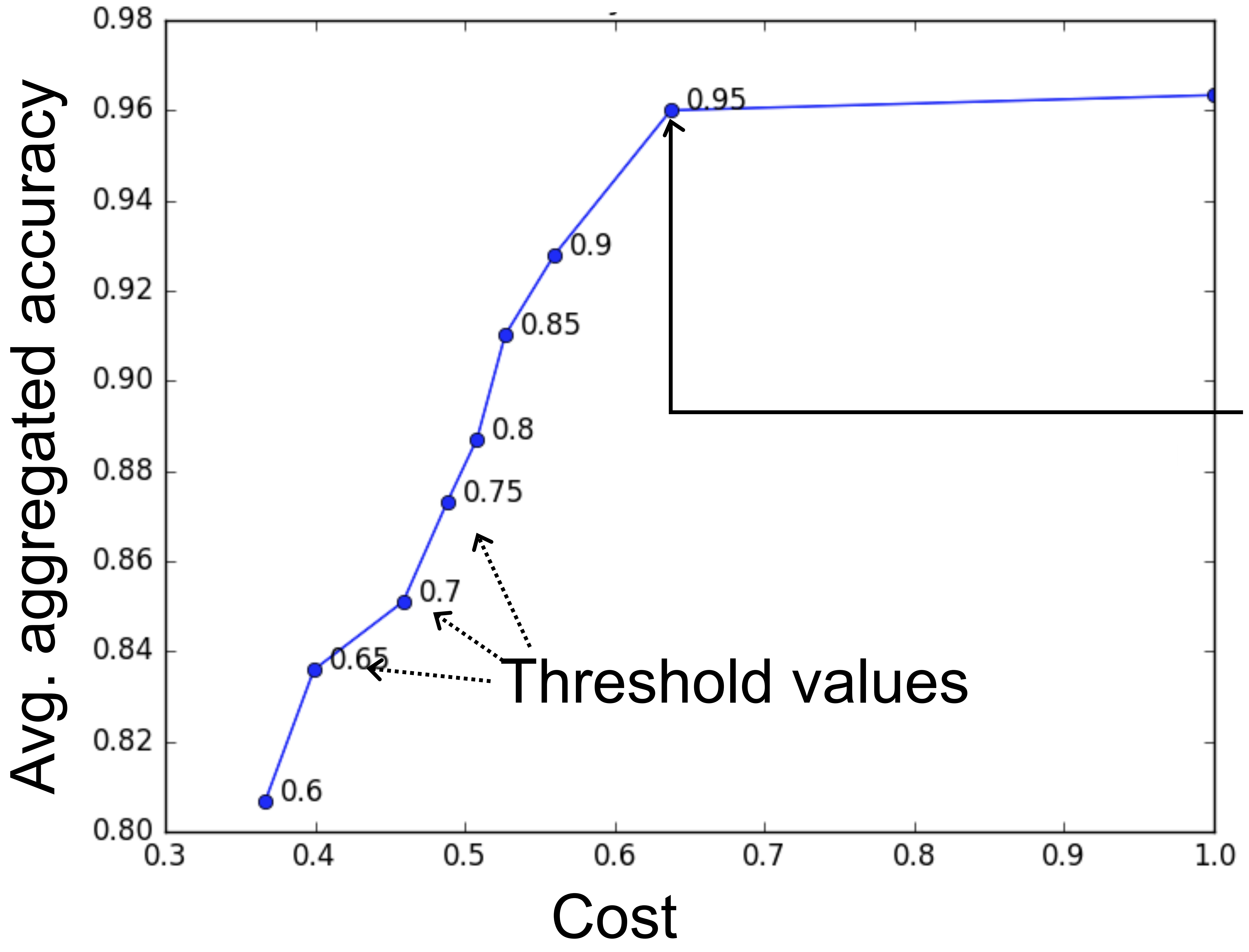
- Stop labeling if $E\left(\delta(z = z^A) \middle| \{y_{w_1}, \dots, y_{w_n}\}\right) \geq C$

parameter



Sheng VS, Provost F, Ipeirotis PG. Get another label? improving data quality and data mining using multiple, noisy labelers. KDD 2008

Threshold in IRL: cost – accuracy trade-off

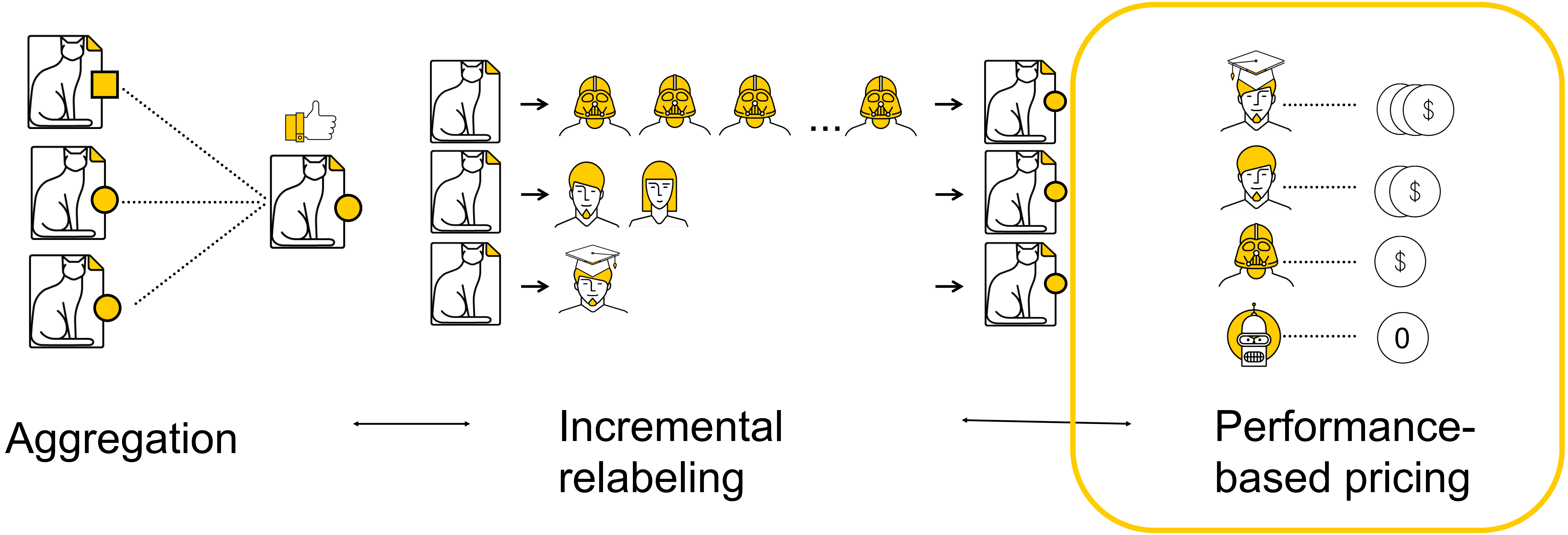


Optimal threshold $c = 0.95$

A higher c does not increase accuracy

Saving ≈ 35% of noisy labels

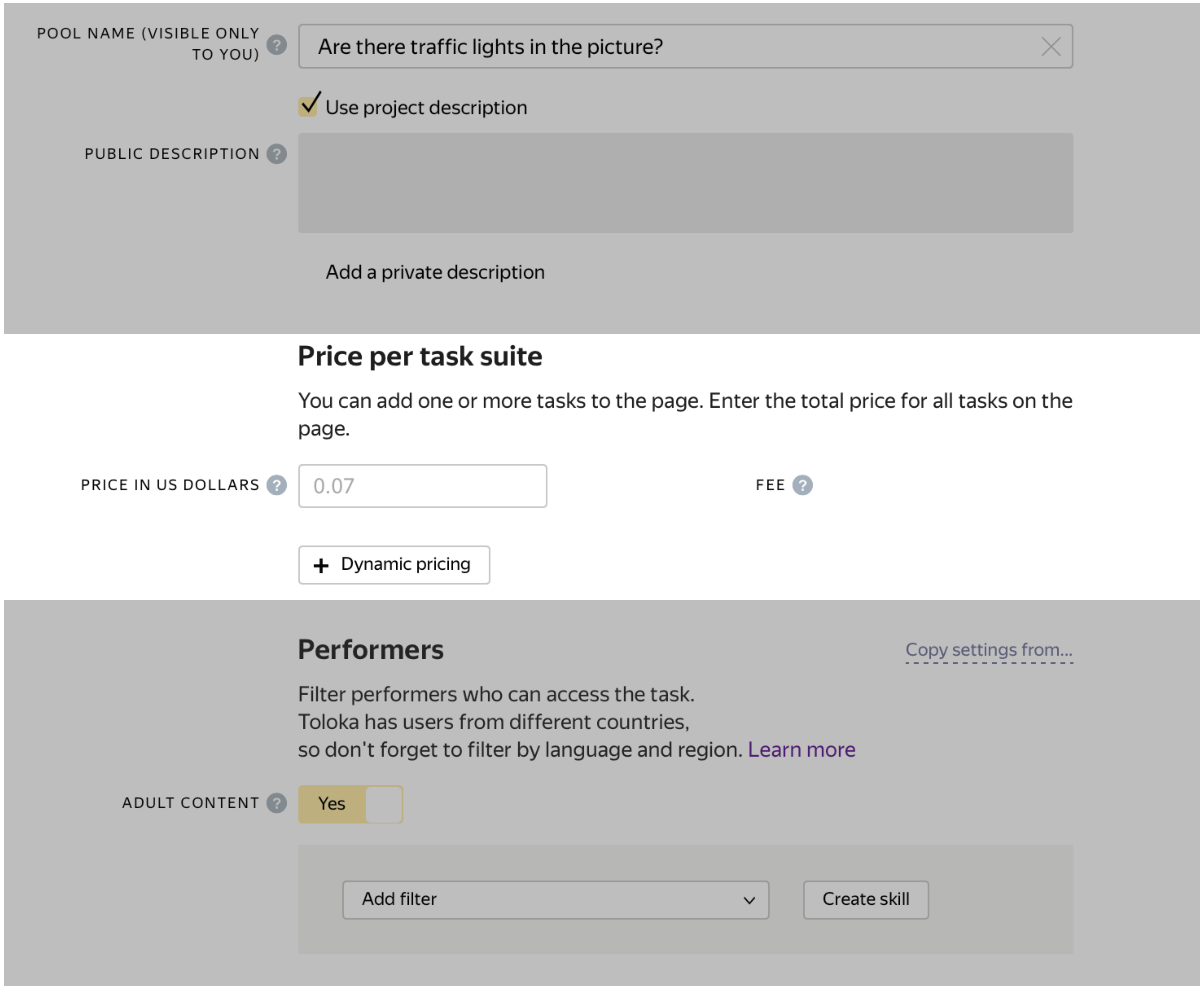
Key components of labeling with crowds



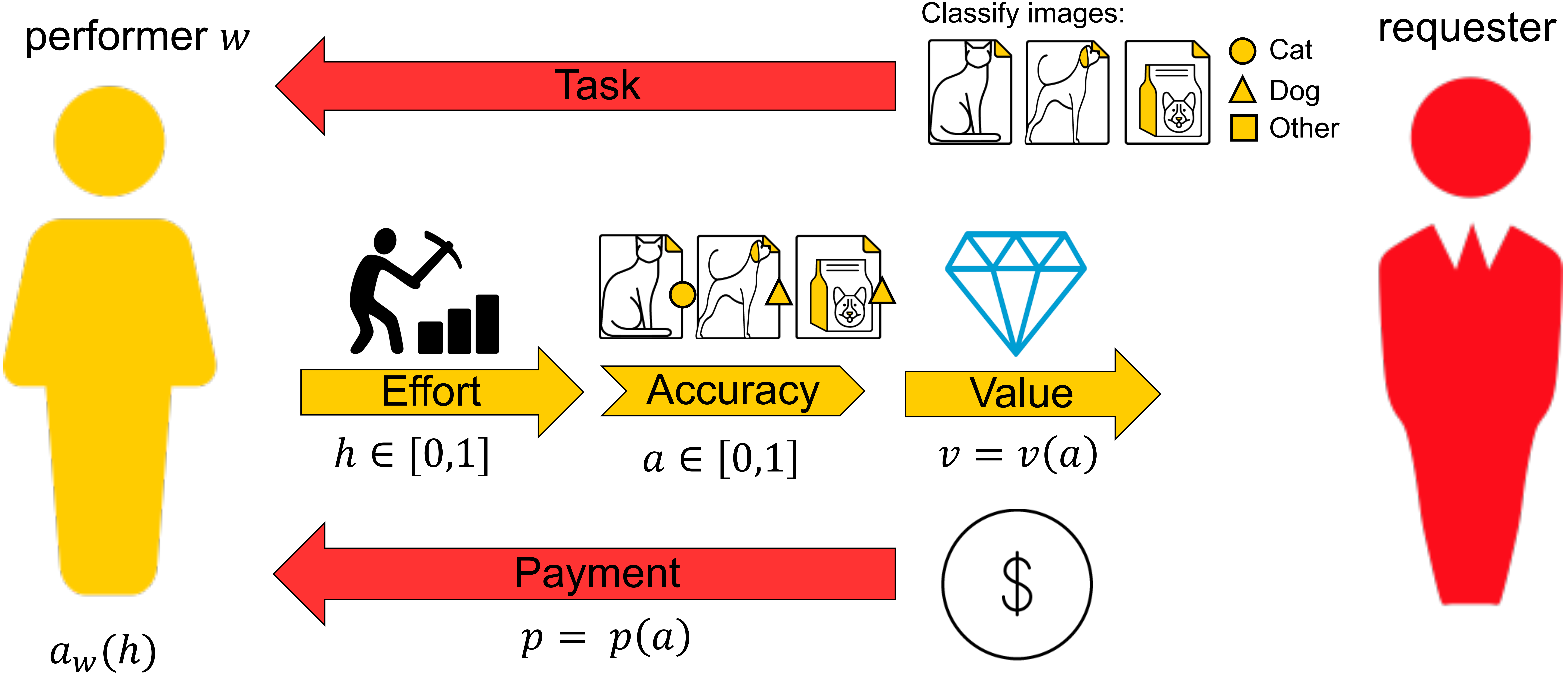
Performance-based pricing

aka dynamic pricing

Pool settings: dynamic pricing

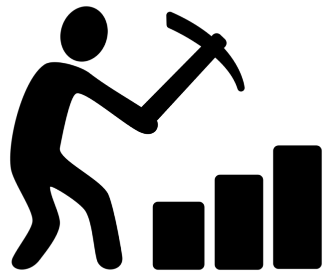


Labeling as a game: notation



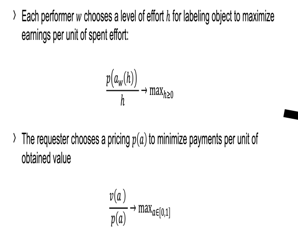
Labeling as a game: formalization

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



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

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



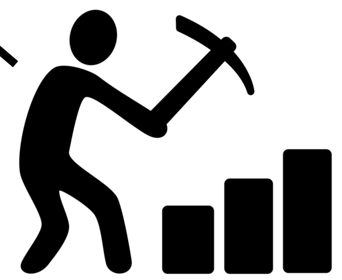
$$\frac{v(a\ )}{p(a)} \to \max_{a\in[0,1]}$$

Labeling as a game: incentive compatible pricing

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

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

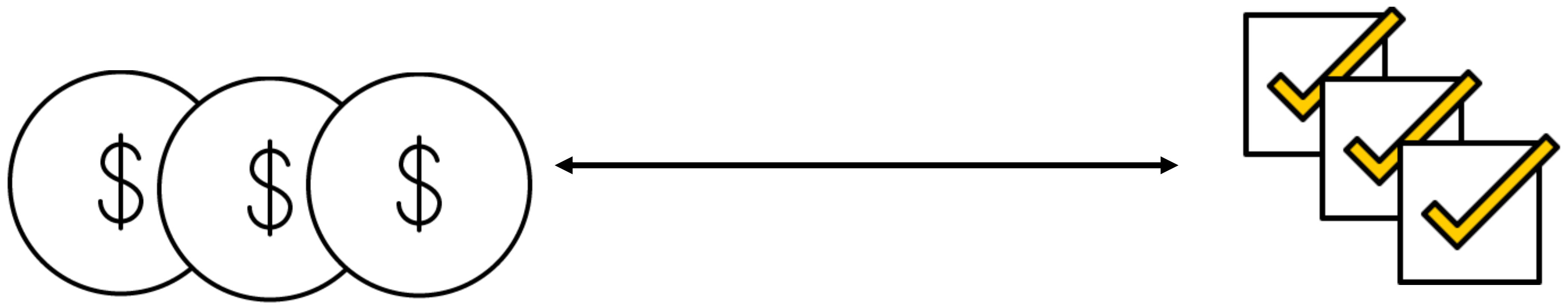
Accuracy



The requester and 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

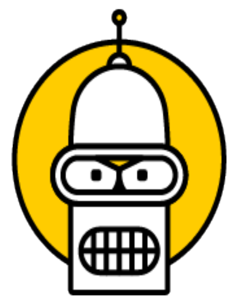
- Price p for the level of accuracy a_0 : $\Pr(\hat{z} = z) \geq a_0$ E.g.:



p = 0.3$ a_0 = 0.99

- $\hat{q}_w = \Pr(y^w = z)$ - 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

$\hat{q}_w = 0.8$

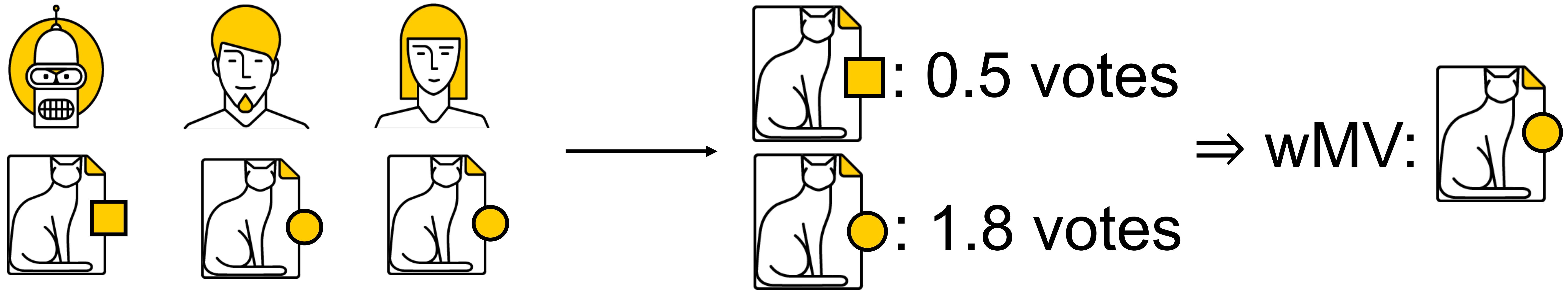
100 correct GS among 100

$\hat{q}_w = 1$

Wang, Ipeirotis, and Provost, Quality-Based Pricing for Crowdsourced Workers, 2013

Performance-based pricing in practice: settings

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



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

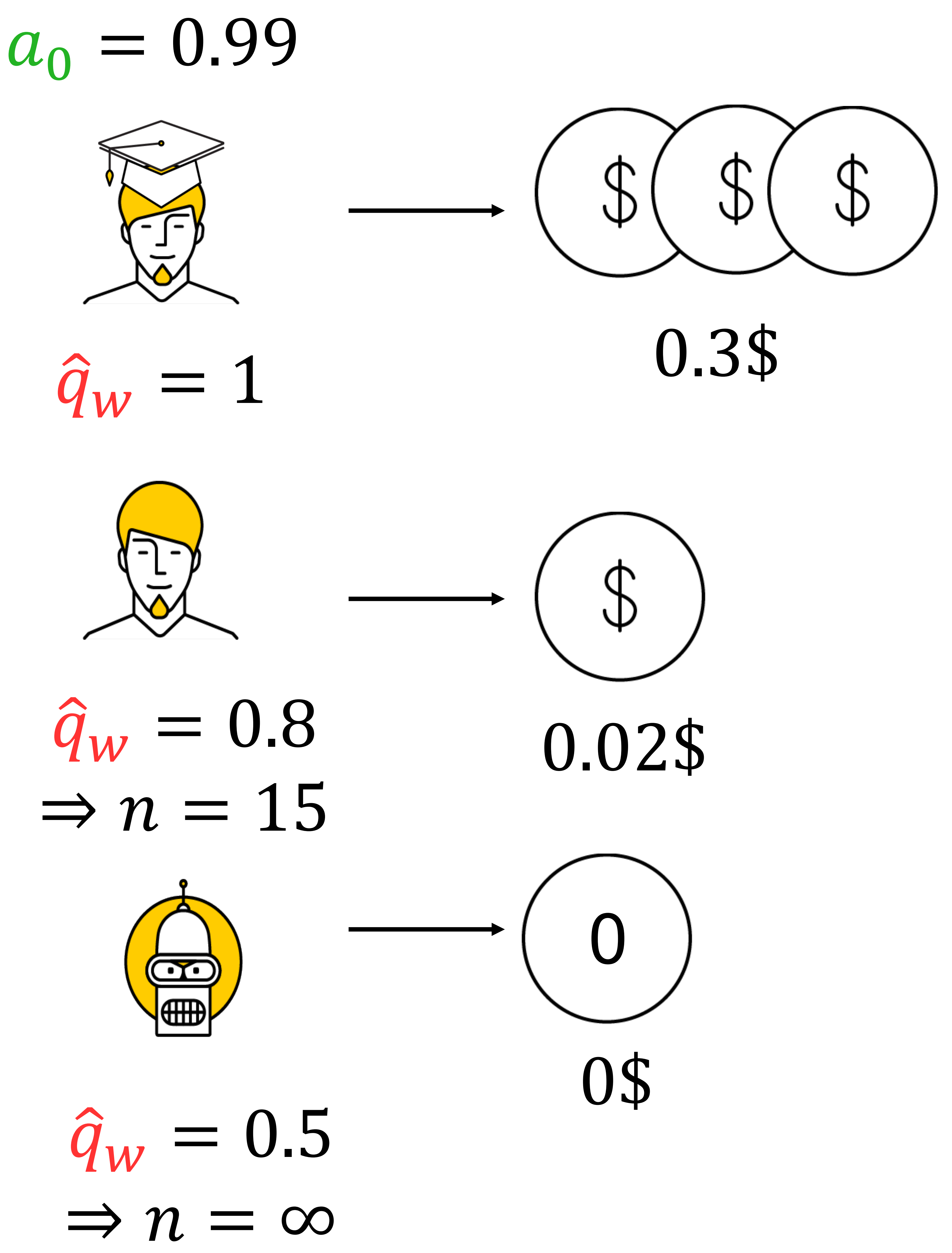
Performance-based pricing in practice

Pricing rules

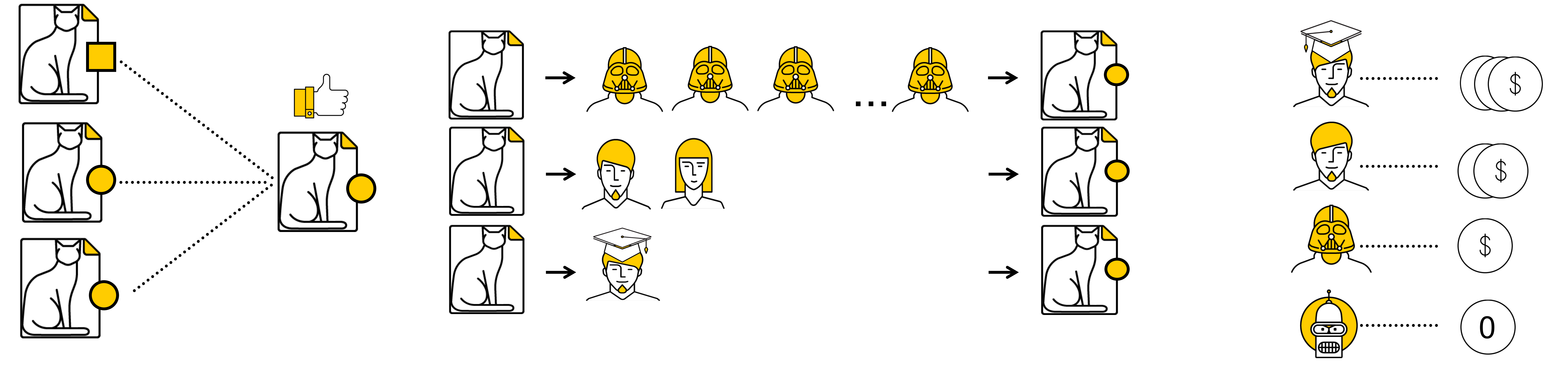
1. If $\hat{q}_w \geq a_0$, then the price is p
2. Else find n:

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

The price is p/n



Key components of labeling with crowds



Aggregation

Incremental relabeling

Performance-based pricing

Yandex

Thank you!
Questions?

Dmitry Ustalov

Analyst at Data Analysis and Research Group

dustalov@yandex-team.ru

https://research.yandex.com/tutorials/crowd/sigmod-2020