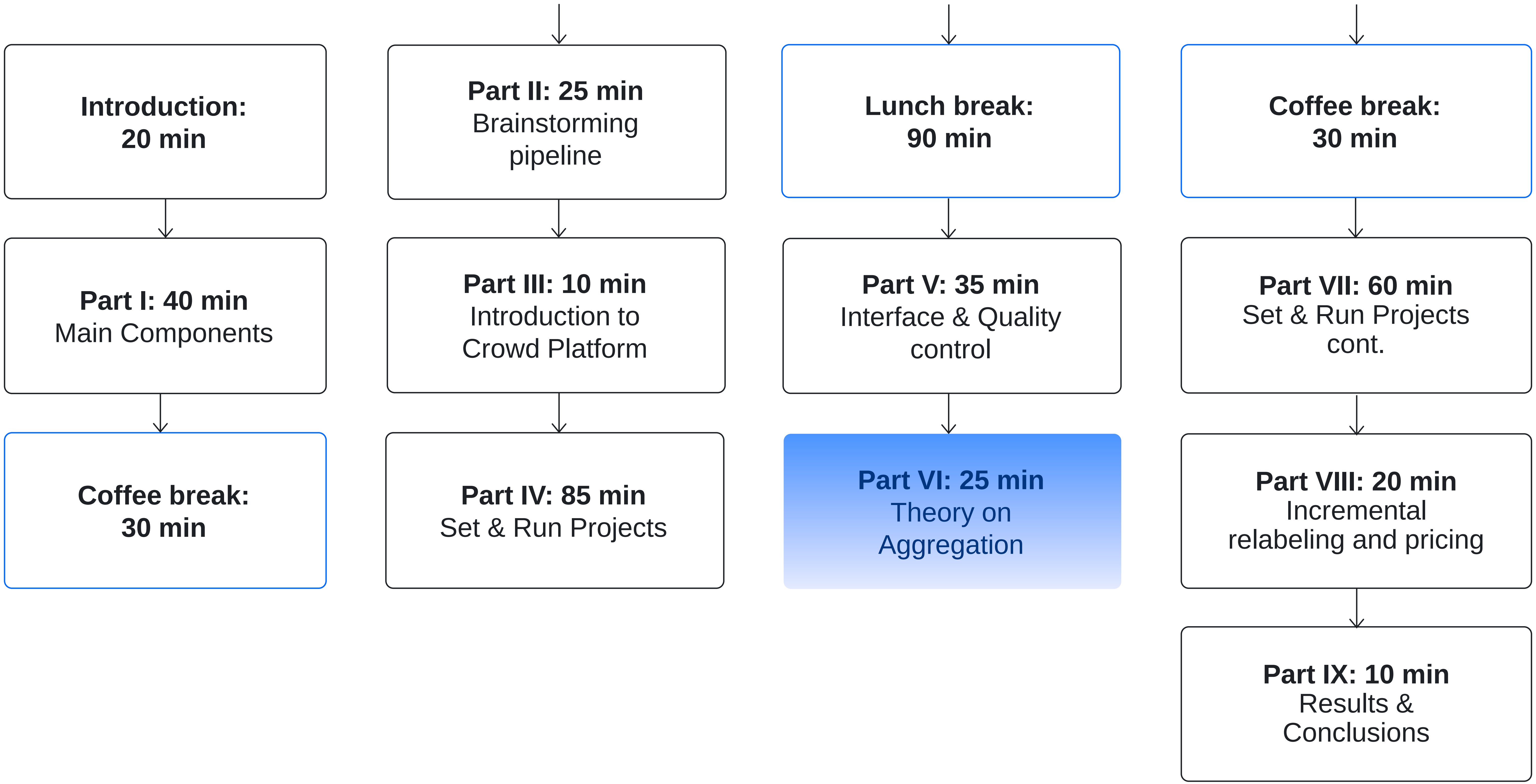
Part VI

Theory on Aggregation

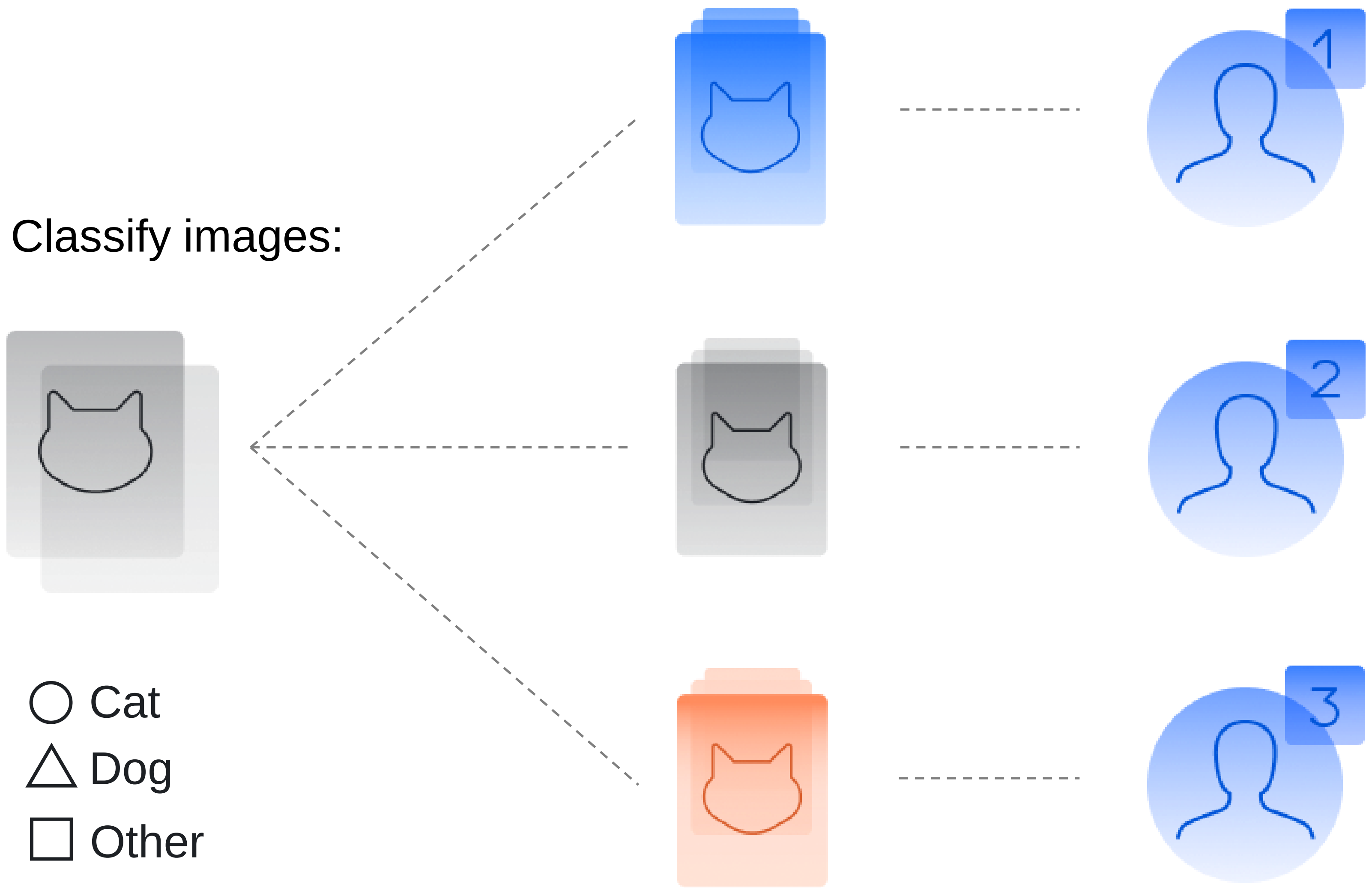
Valentina Fedorova,
Research analyst

Toloka

Tutorial schedule



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

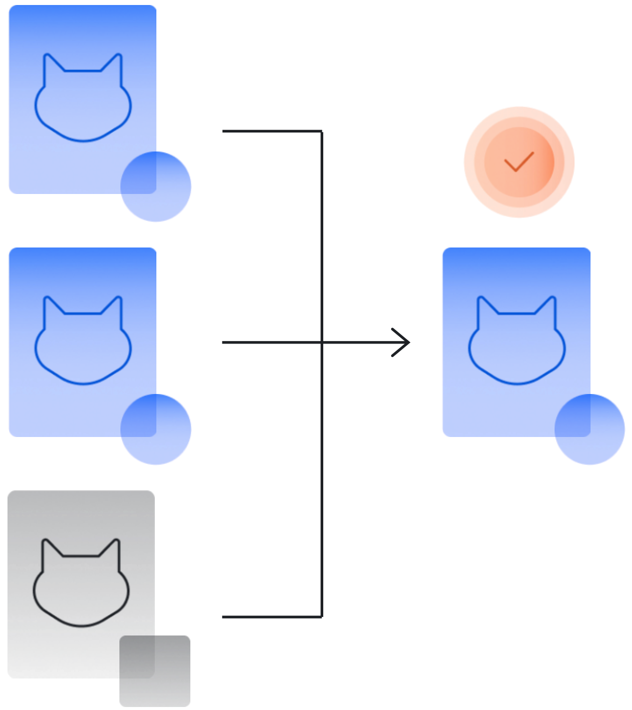
Accuracy vs Cost

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

or

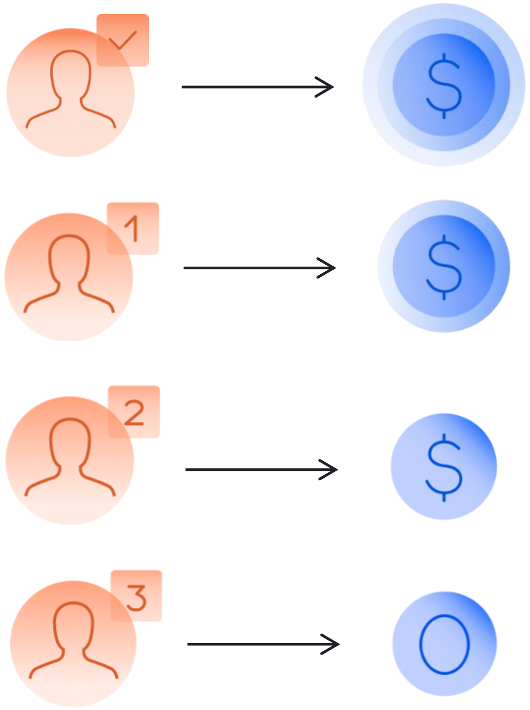
- Labels of a chosen accuracy level for a **minimal budget**

Key components of labeling with crowds



Aggregation

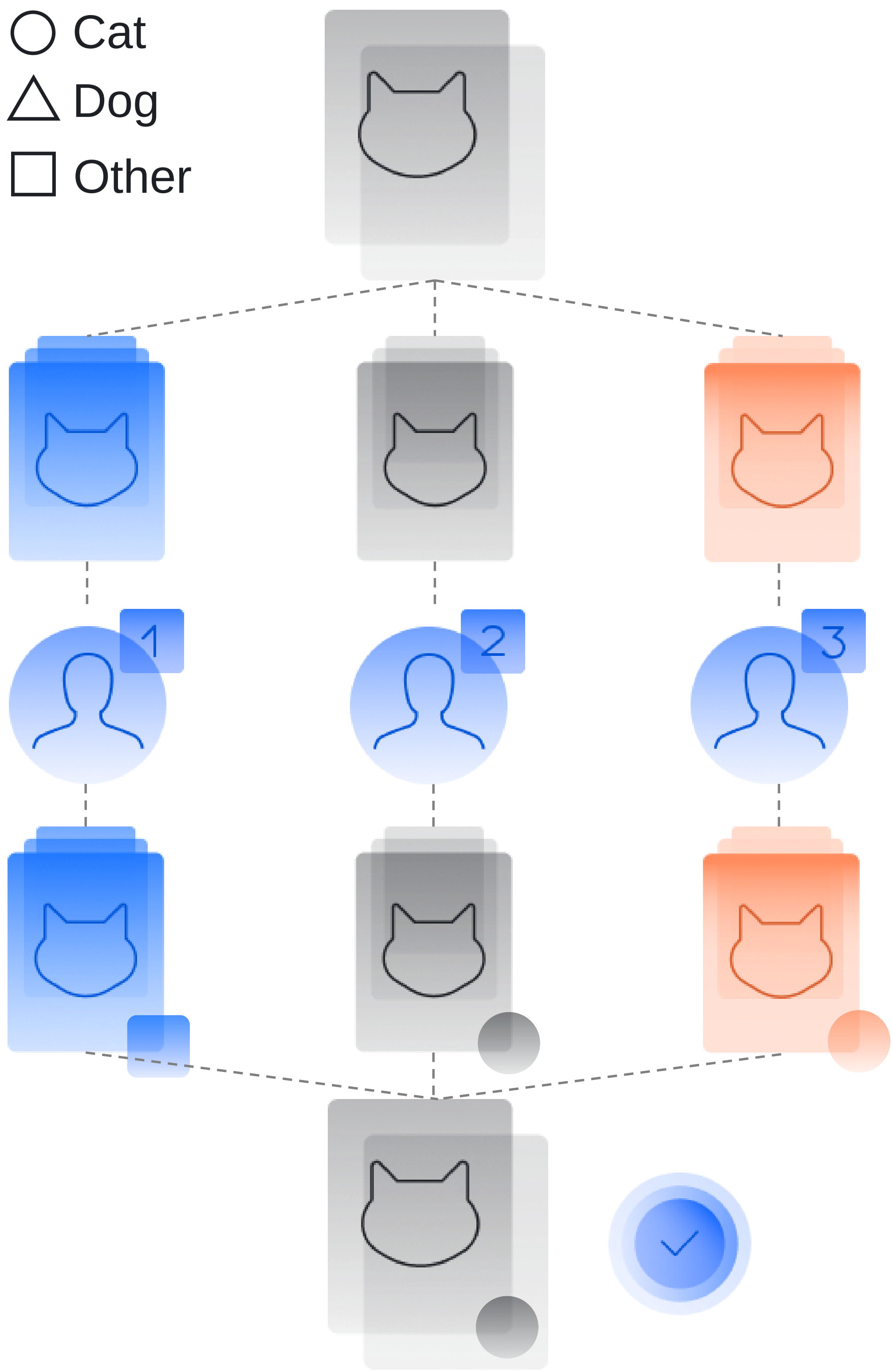
Incremental relabeling



Performance-based pricing

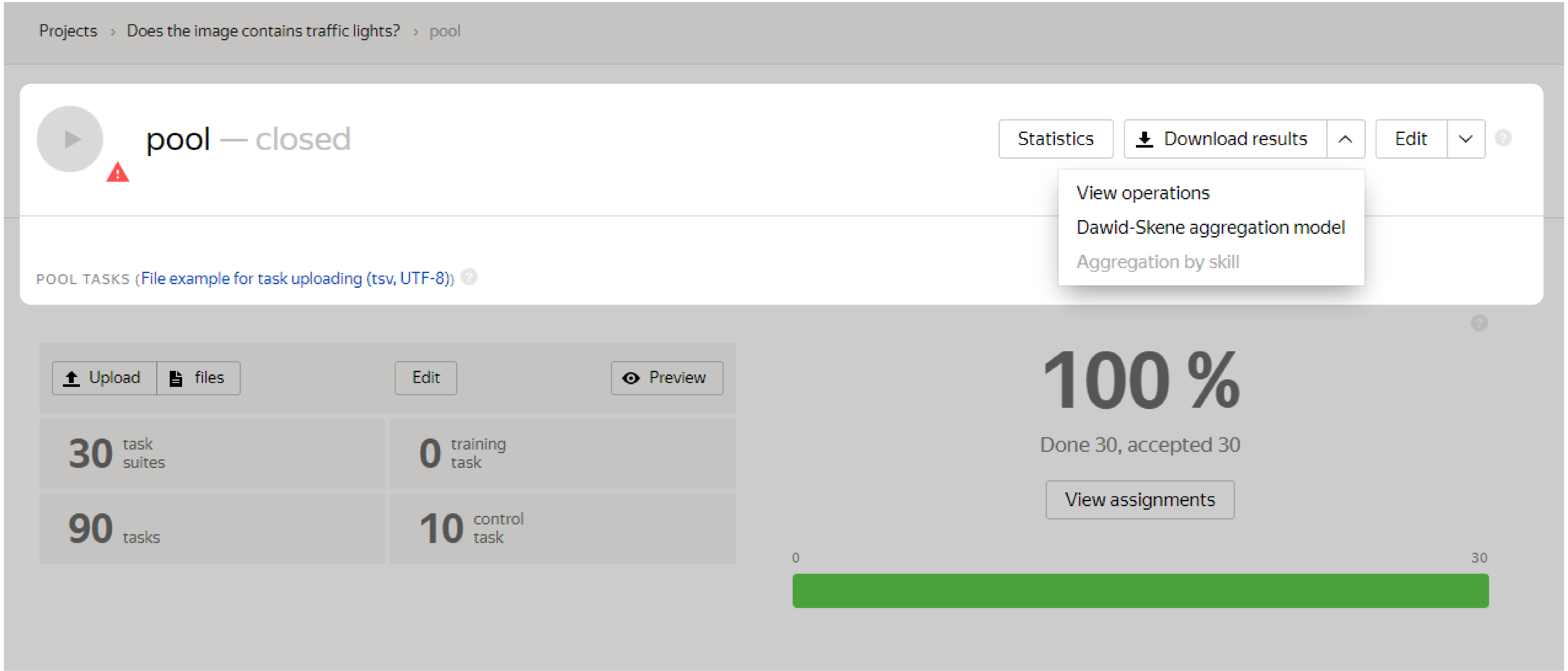
Aggregation

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



Multiclass labels

Project 1: Filter images

Are there shoes
in the picture?

Yes

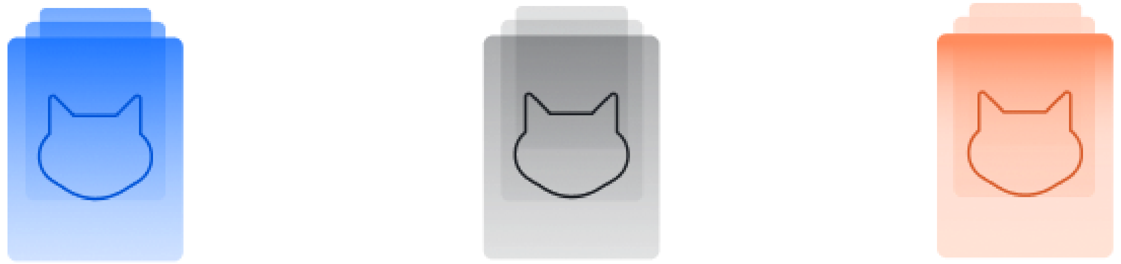
No

Notation

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

○ Cat △ Dog ☐ Other

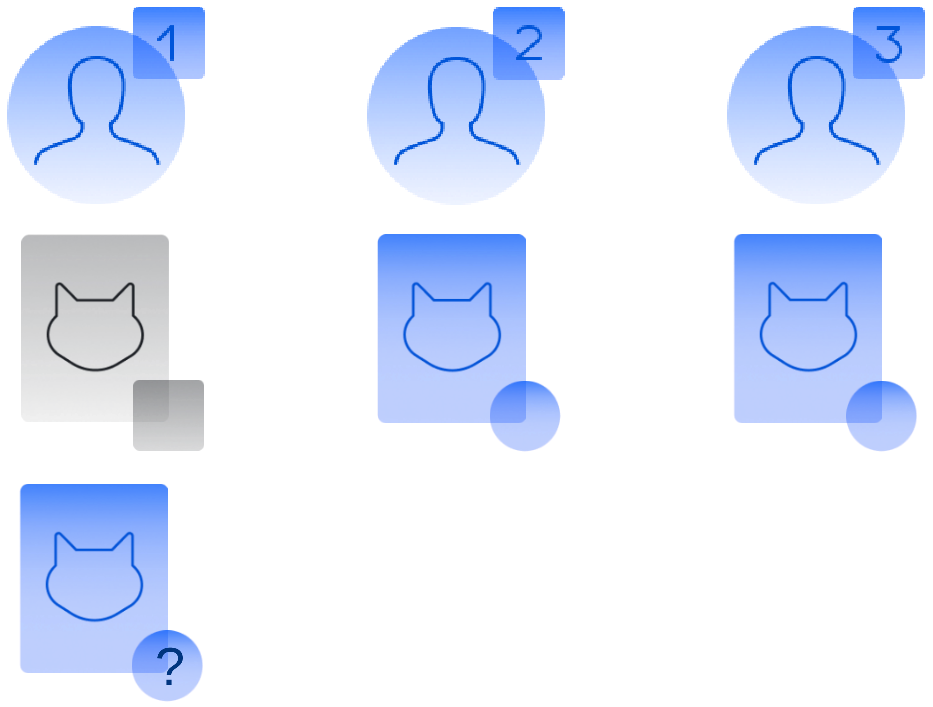
- Objects $j \in \{1,\dots,J\}$. E.g.:



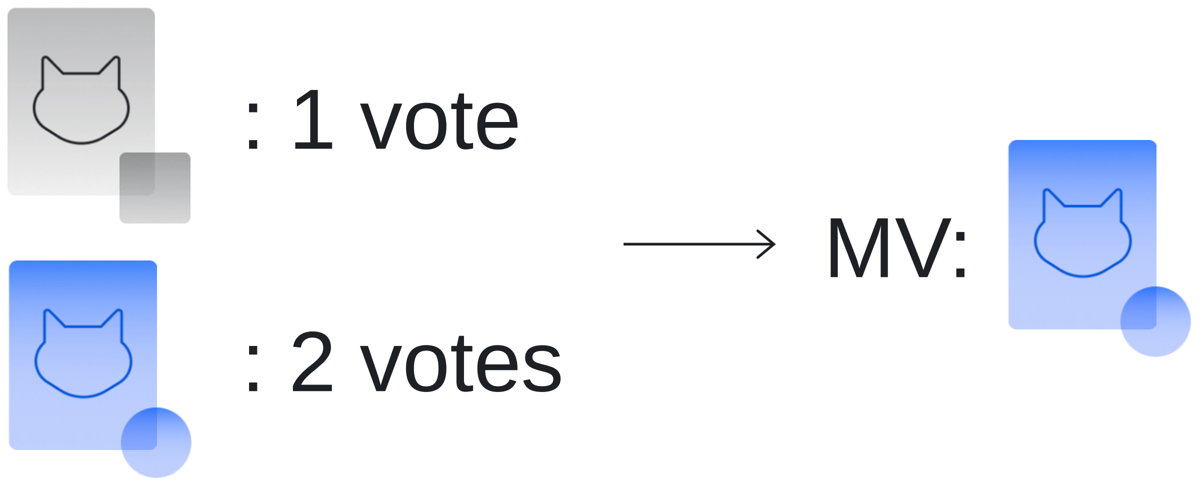
- Workers: $w \in \{1,\dots,W\}$. E.g.:
  - $W_j \subseteq \{1,\dots,W\}$ — workers labeled object j

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 $z = \{z_j \mid j = 1, \dots, J\}$

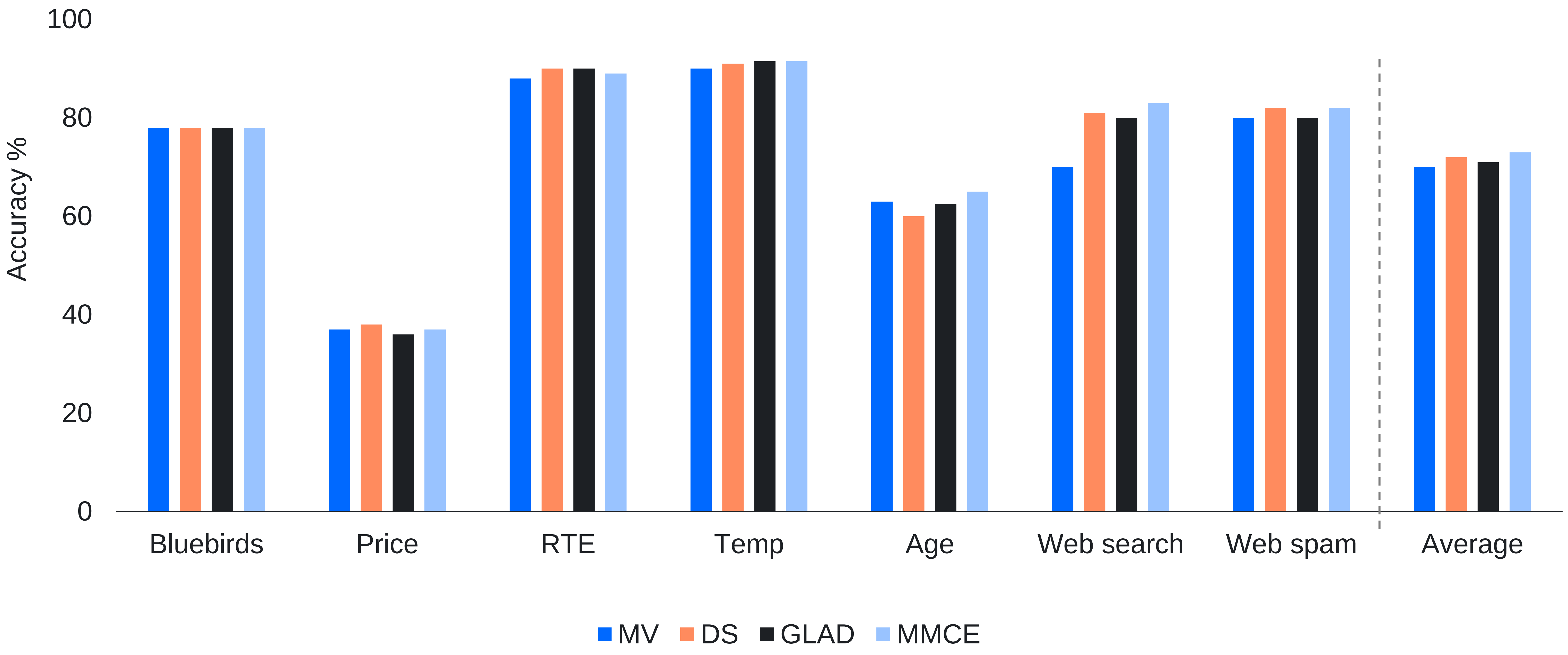


- A straightforward solution:



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

Performance of MV vs other methods

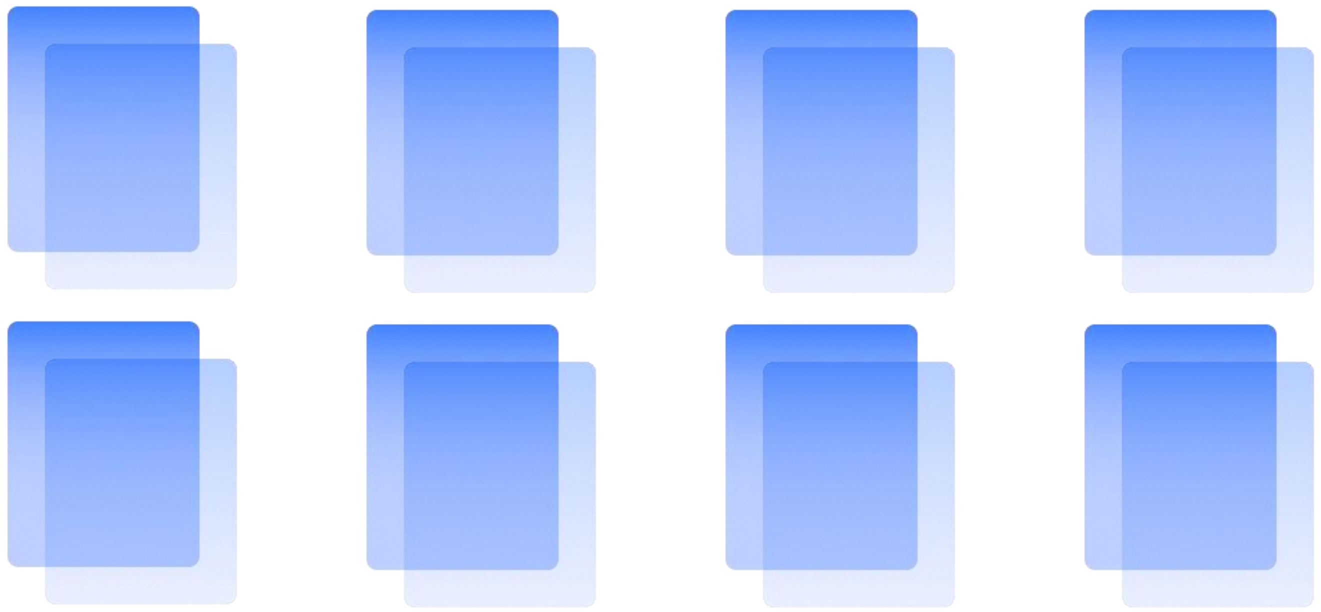


Zhou D. et al. Regularized minimax conditional entropy for crowdsourcing. 2015

Properties of MV

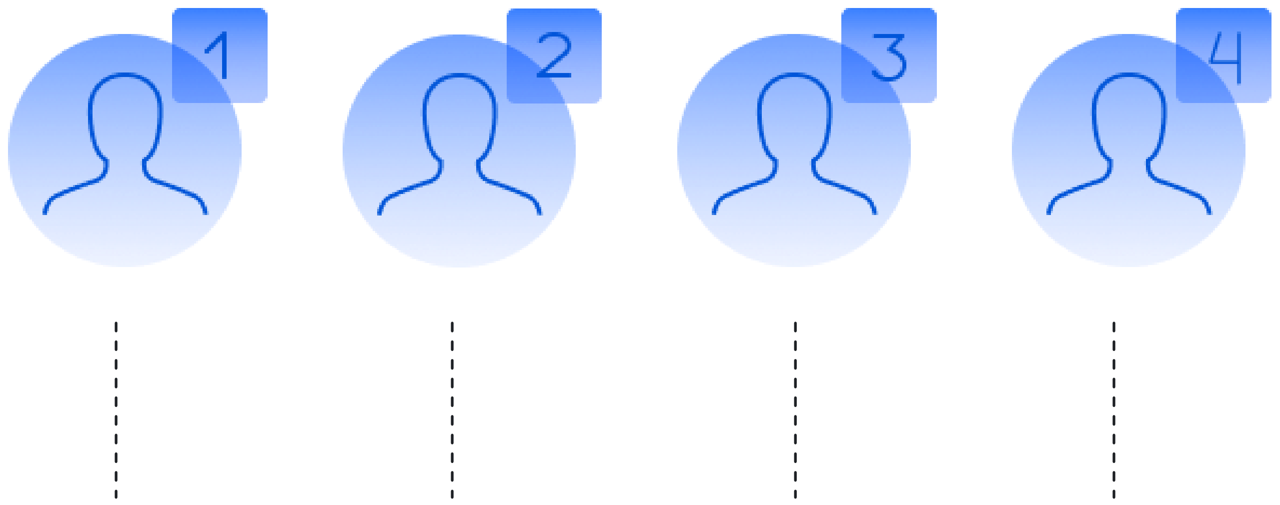
All workers are
treated similarly

All objects are
treated similarly



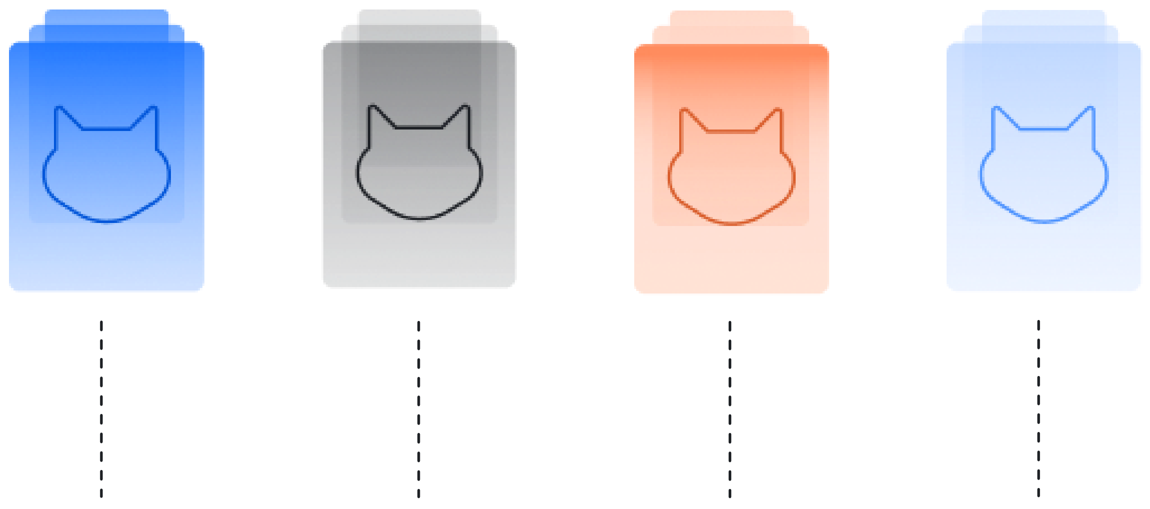
Advanced aggregation: workers and objects

Parameterize expertise of workers by e^{w}



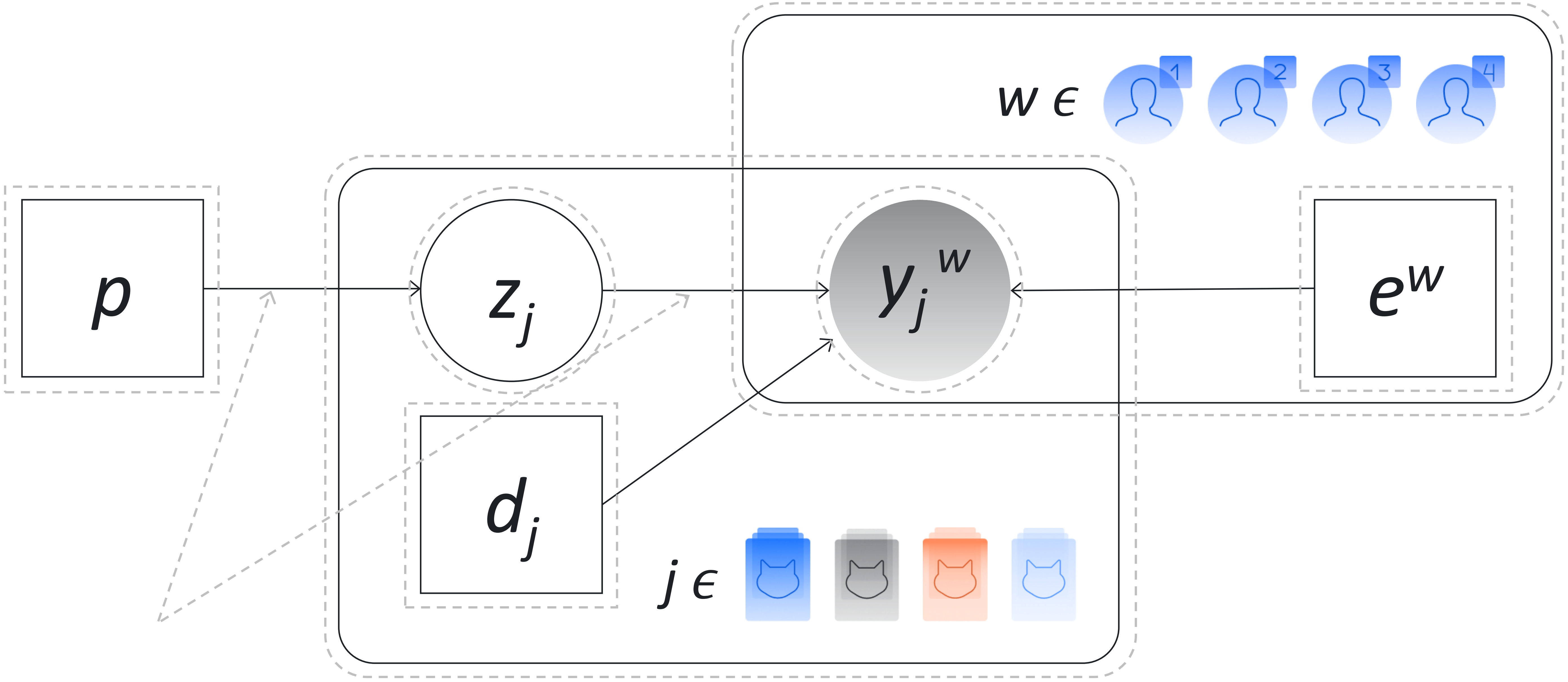
e^{w_1} e^{w_2} e^{w_3} e^{w_4}

Parameterize difficulty of objects by d_j

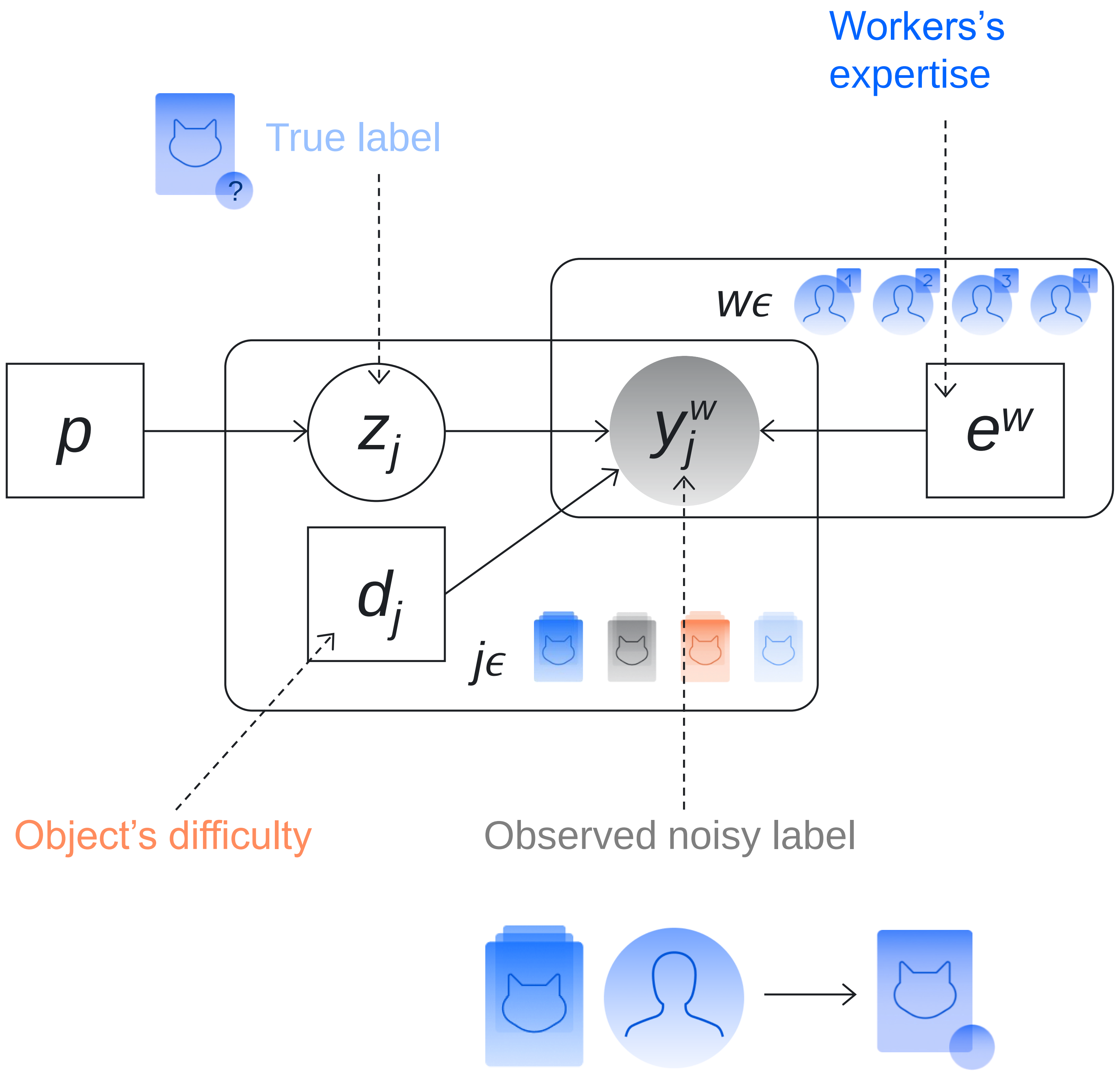


d_{j_1} d_{j_1} d_{j_1} d_{j_1}

Advanced aggregation: latent label models

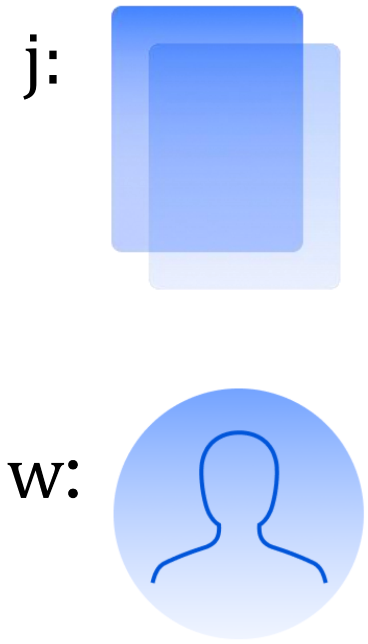


Latent label models: noisy label model



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

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

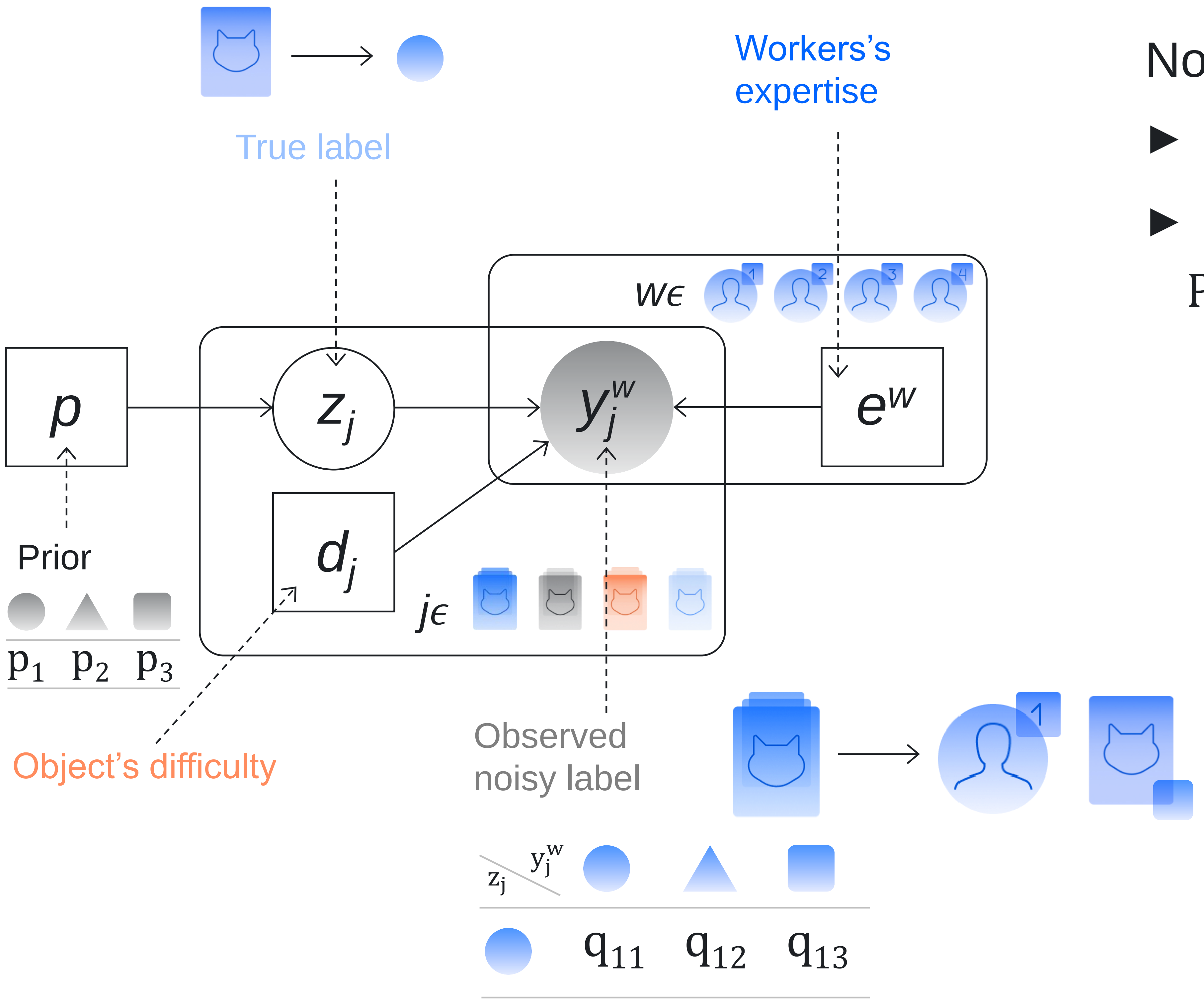


M_j^w:

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 $P_Y(M_j^w[z_j, \cdot])$

In multiclassification, a standard choice for $P_Z(\cdot)$ and $P_Y(\cdot)$ is a Multinomial distribution 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)*

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

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

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

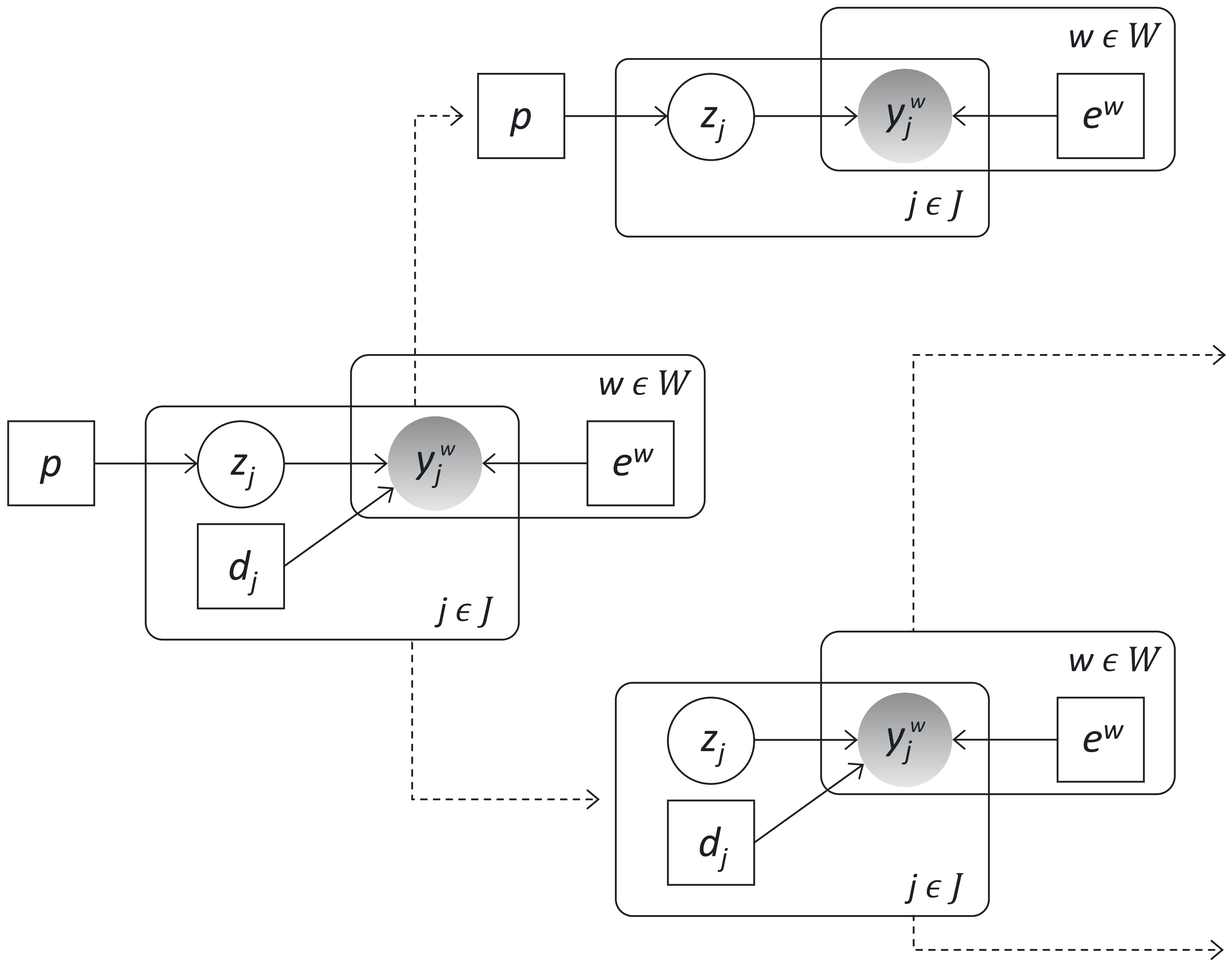
- **M-step:** Maximize the expectation of LL with respect to the posterior distribution of $\hat{z}$:

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

  - Analytical solutions
  - Gradient descent

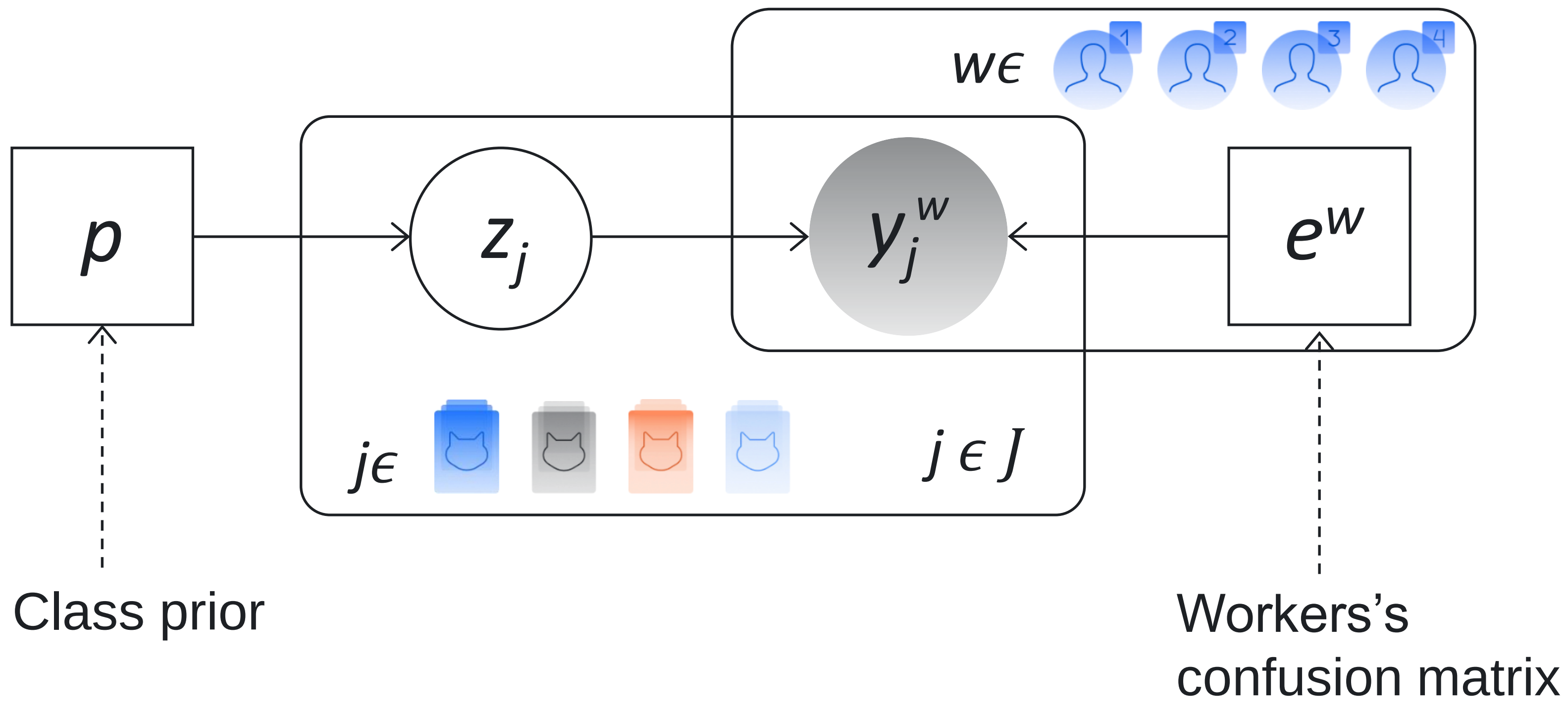
* it is a lower bound on LL of y and z

Latent label model (LLM): special cases



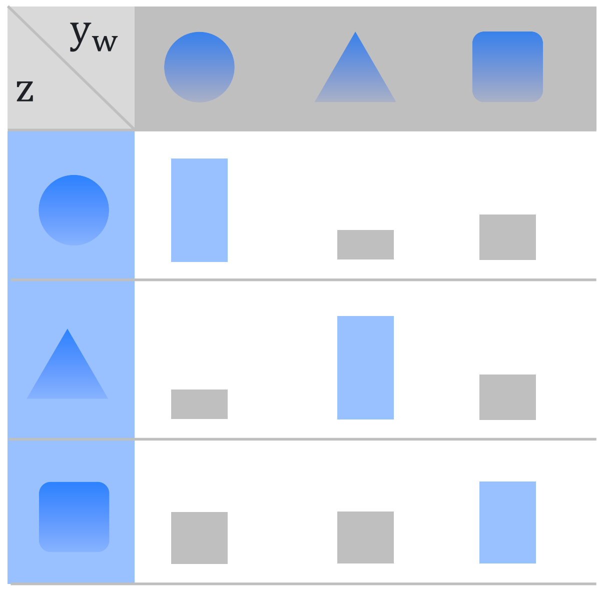
- Dawid and Skene model (DS):
  - Categories are different
  - Objects are **similar**
  - Workers are different
- Generative model of labels, abilities, and difficulties (GLAD):
  - Categories are **similar**
  - Objects are different
  - Workers are different
- Minimax conditional entropy model (MMCE):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



DS: parameters optimization

- **E-step:**

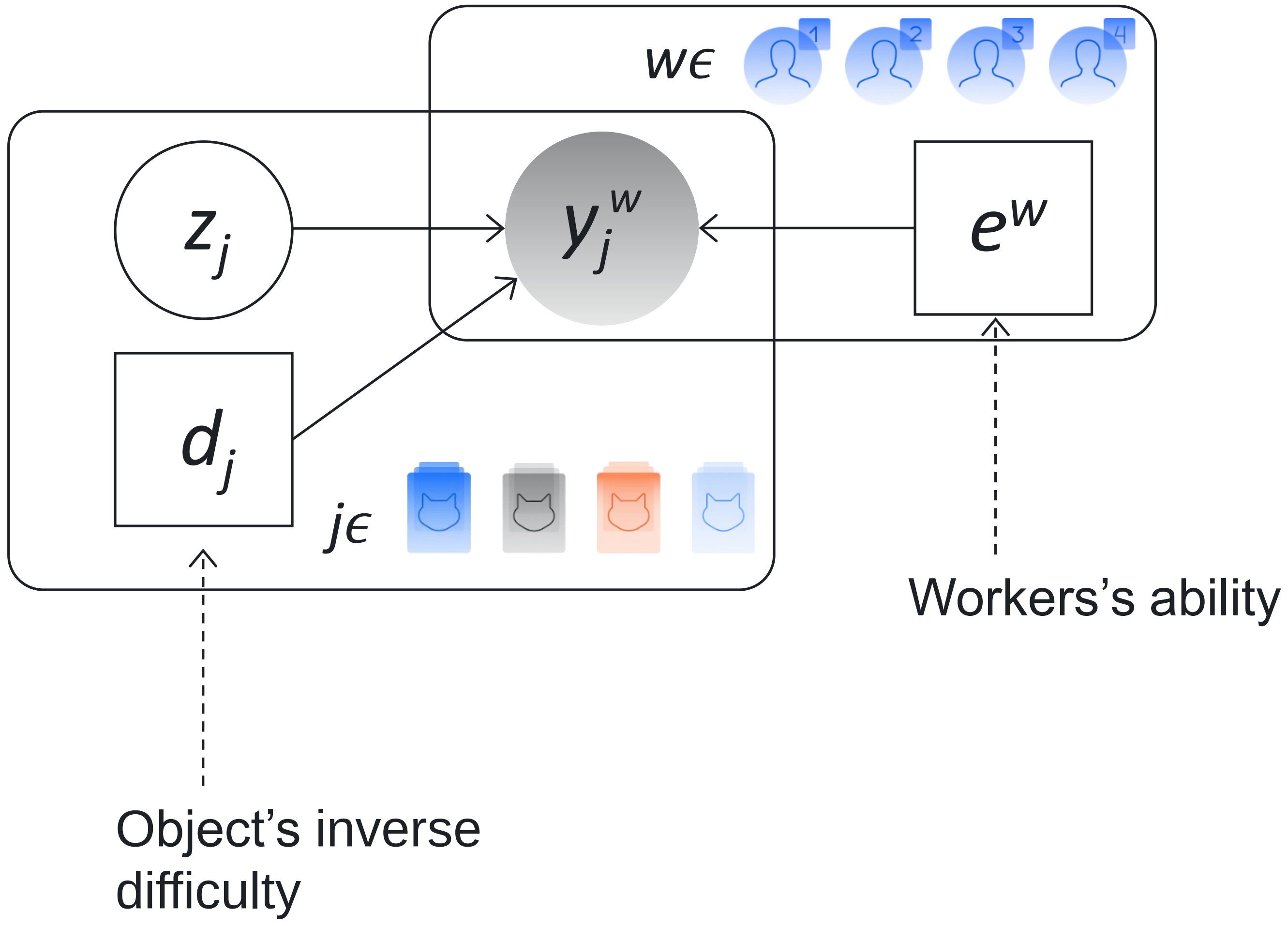
$$\hat{z}_j[c] = \frac{p[c]\ \prod_{w \in W_j} e^w[c, y_j^w]}{\sum_k p[k]\ \prod_{w \in W_j} e^w\left[k, y_j^w\right]}, \qquad c = 1, \dots, K$$

- **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:

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

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

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., $P(z_j) = {}^1/_K$)

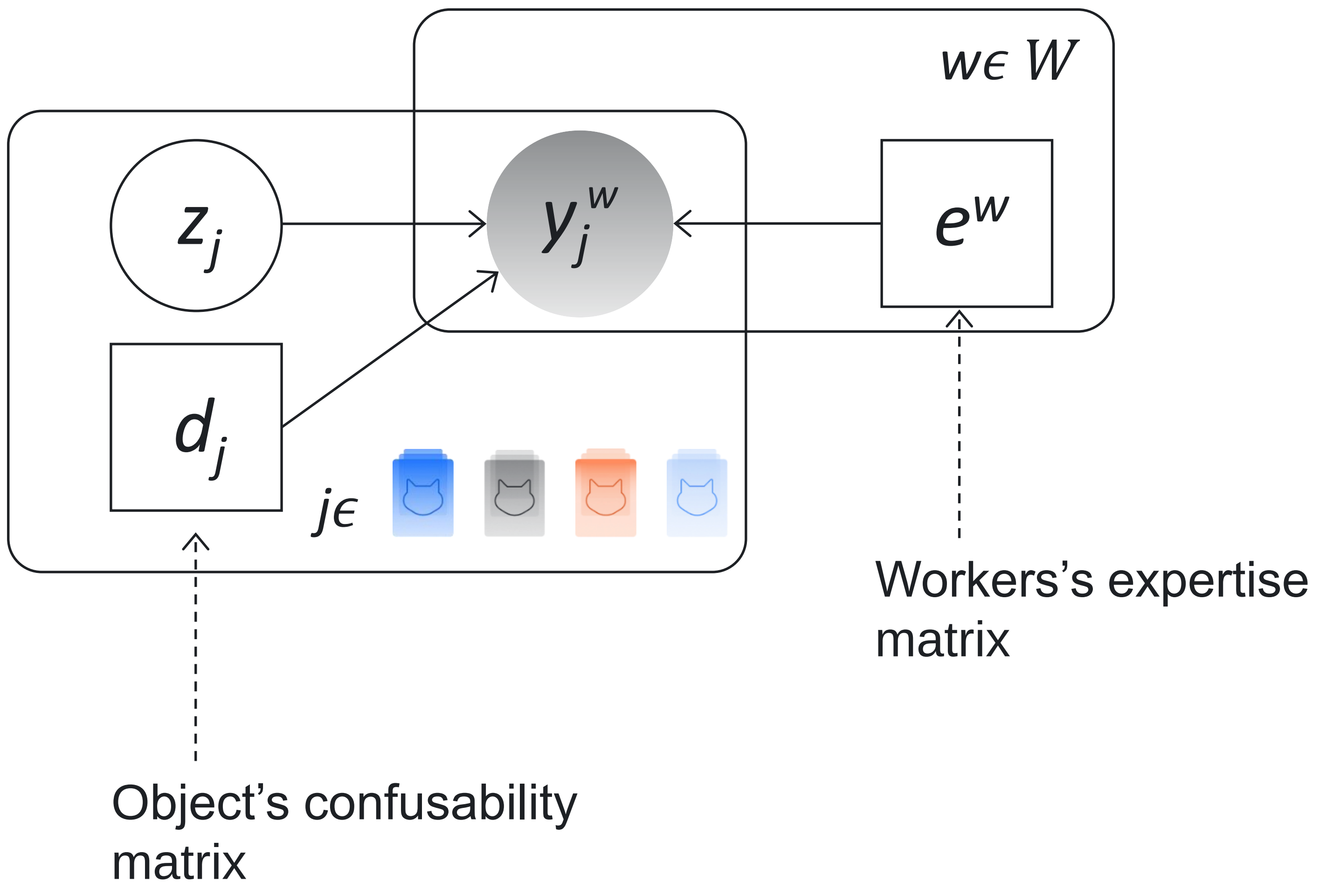
- **E-step:**

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

- **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)



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

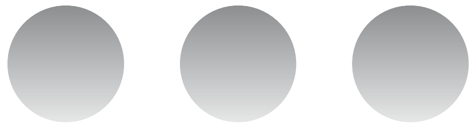
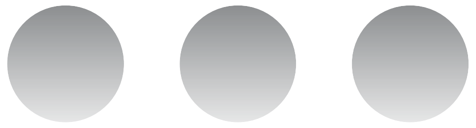
$$\min_Q \max_P - \sum_{\substack{j \in J \\ c \in \{1,\dots,K\}}} Q(Z_j = c) \sum_{\substack{w \in W \\ k \in \{1,\dots,K\}}} P(Y_j^w = k | Z_j = c) \log P(Y_j^w = k | Z_j = c)$$

- 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

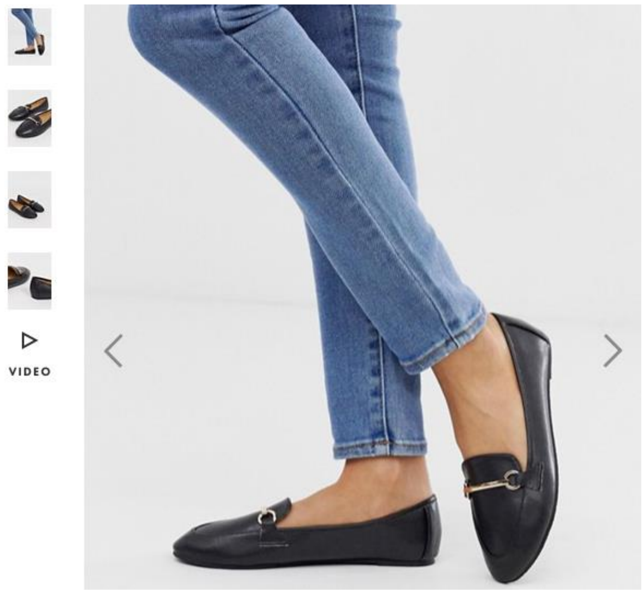
Summary of aggregation methods

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

Pairwise comparisons

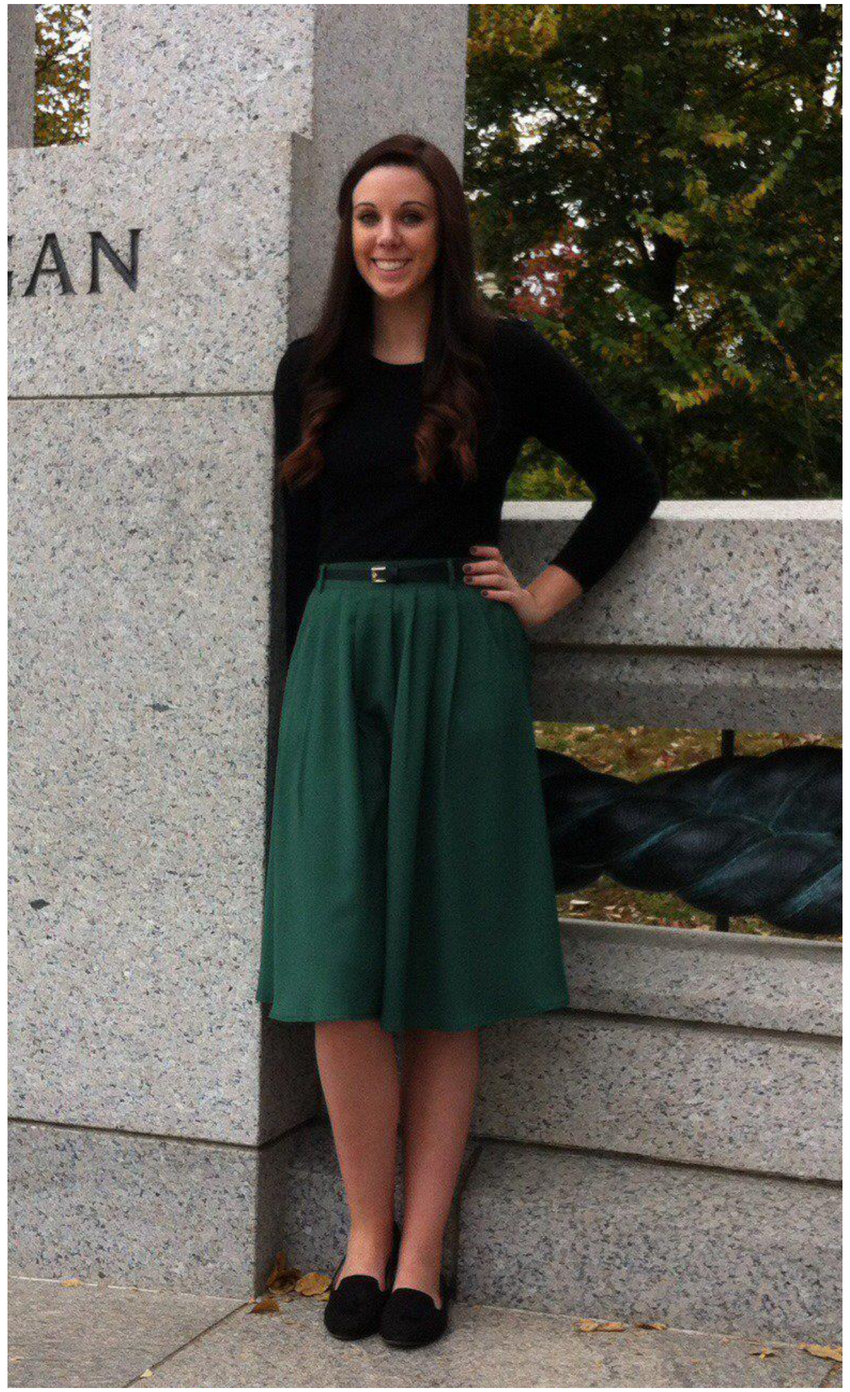
Project 4: Compare items

Which shoes look more similar to the one in the picture?



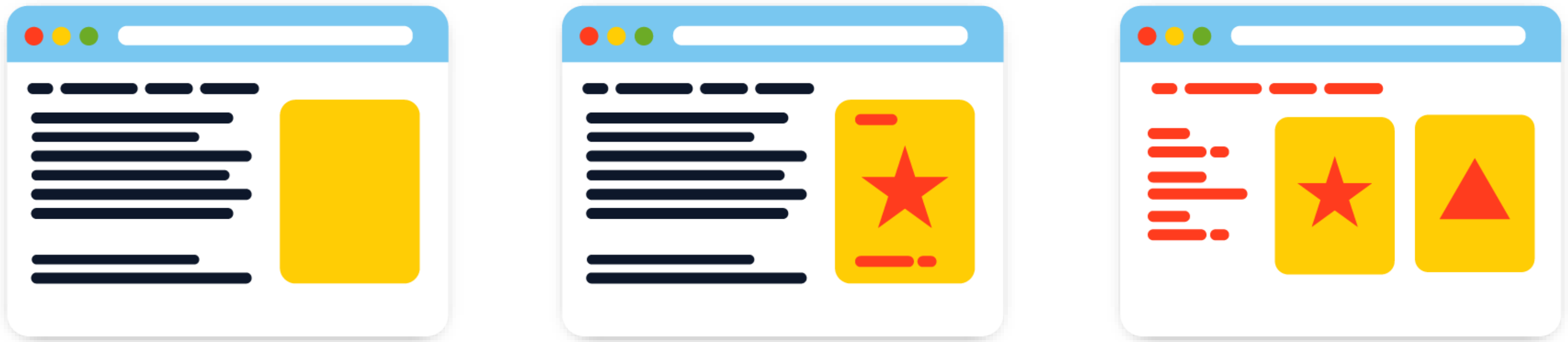
Left

Right

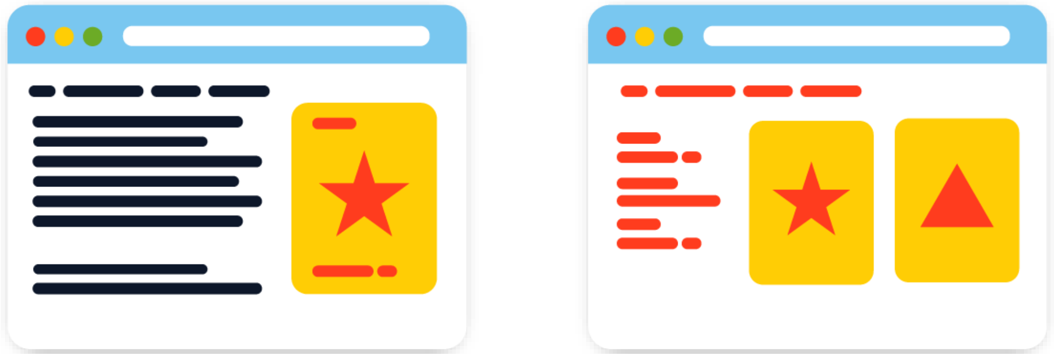


Notation

- Answers: Left or Right
- Items $d_j \in \{1, \ldots, N\}$ E.g.:

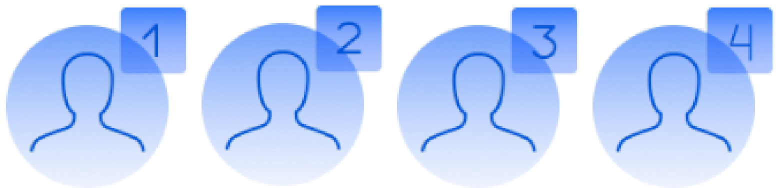


- Tasks:



Choose a better item:
Left
Right

- Workers $w \in \{1, \ldots, W\}$ E.g.:



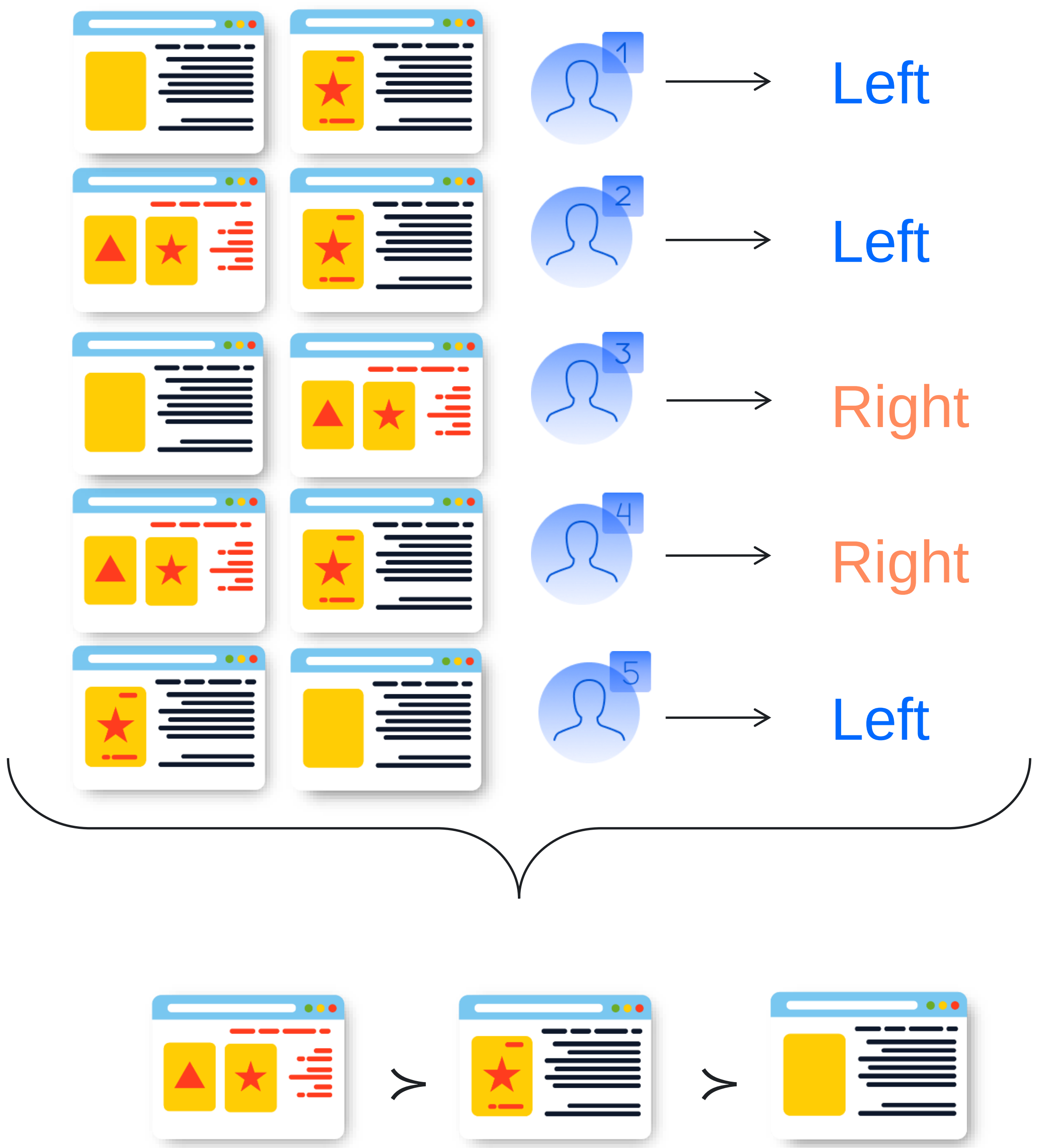
Formalization

Ranking from pairwise comparisons:

- Given pairwise comparisons for items in D:

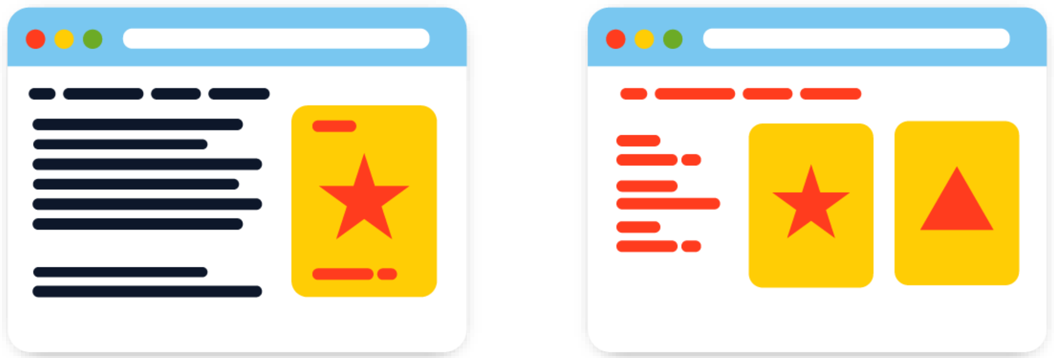
$$P = \{(w_k, d_i, d_j) : i \succ_k j\}$$

- Obtain **a ranking** π over items $D \rightarrow \{1, \ldots, N\}$ based on answers in P



Difference from multiclassification

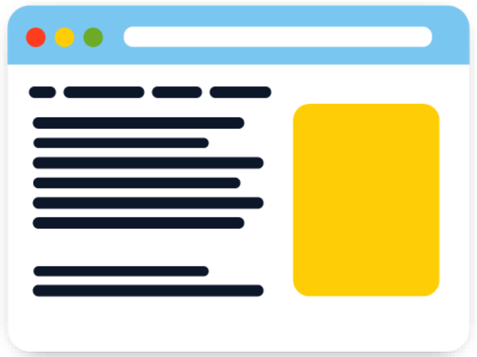
- The latent label assumption is not satisfied when comparing complex items



- Different tasks may contain common items

Bradley and Terry model (BT)

- Assume that each item $d_i \in D$ has a latent "quality" score $s_i \in \mathbb{R}$

d_i  $\longleftrightarrow$ s_i

- The probability that $d_i \in D$ will be preferred in a comparison over $d_j \in D$

$$\Pr(i \succ j) = f\left(s_i - s_j\right), \text{ where } f(x) = {}^{1}/_{1+e^{-x}}.$$

- The model assumes that all workers are equally good and truthful

Bradley, R. A. and Terry, M. E. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". 1952

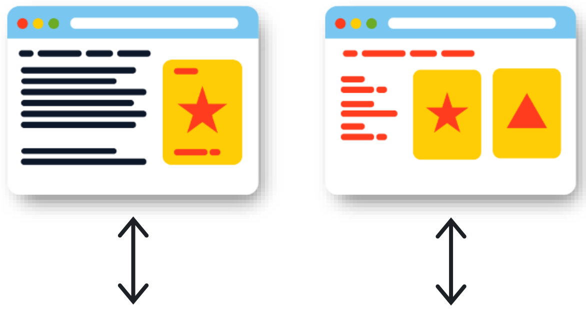
NoisyBT model: parameterization of workers

w_k  $\longleftrightarrow$ "reliability" γ_k and "bias" q_k

- The probability that w reads task is

$$\Pr(w_k \text{ reads a task}) = f(\gamma_k) \longleftarrow \text{Logistic function}$$

- If w_k reads a task, she answers according to scores:



$$\underbrace{(f(s_i - s_j)}_{\text{Probability to choose Left if compares items}}, f(s_j - s_i))$$

Probability to choose Left if compares items

- If w_k does not read a task, she answers according to her bias

$$\underbrace{f(q_k)}, f(-q_k)$$

Probability to choose Left if answers randomly

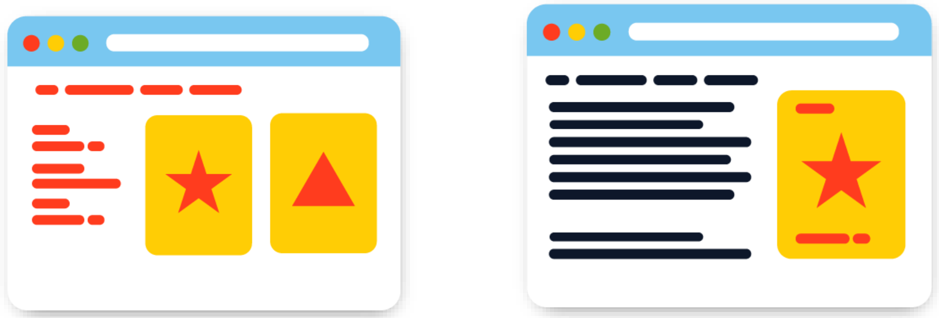
NoisyBT: likelihood of workers' answers

The likelihood of $i \succ_k j$ is

$$\Pr(i \succ_k j) = \underbrace{f(\gamma_k) f\left(s_i - s_j\right)}_{\text{Truthful answer}} + \underbrace{\left(1 - f(\gamma_k)\right) f\left((-1)^{(1-\mathbb{I}(d_i \text{ was left}))} q_k\right)}_{\text{Random answer}},$$

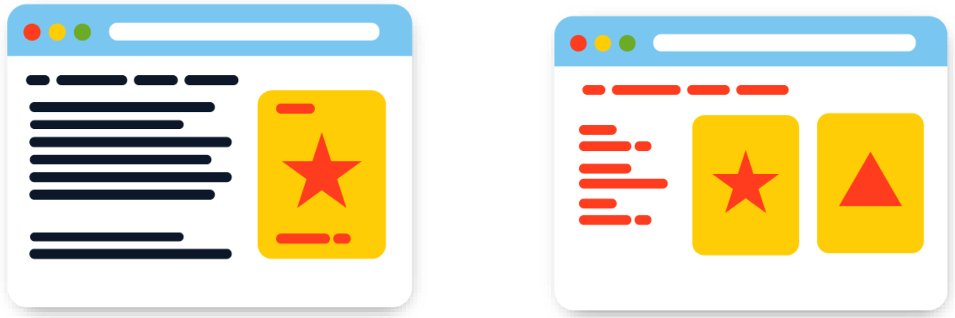
where $\mathbb{I}(d_i \text{ was left})$ is the indicator for the order of d_i and d_j

d_i d_j



$\mathbb{I}(d_i \text{ was left}) = 1$

d_j d_i



$\mathbb{I}(d_i \text{ was left}) = 0$

NoisyBT: parameters optimization

Likelihood of observed comparisons:

$$T(s, q, \gamma) = \sum_{(w_k, d_i, d_j) \in P} \log \Pr(i \succ_k j) =$$

$$\sum_{(w_k, d_i, d_j) \in P} \log\left[f(\gamma_k) f\left(s_i - s_j\right) + \left(1 - f(\gamma_k)\right) f\left((-1)^{(1 - \mathbb{I}(d_i \text{ was left}))} q_k\right)\right]$$

- $\{s_i\}_{i=1,\dots,N}$ and $\{\gamma_k, q_k\}_{k=1,\dots,W}$ are inferred by maximizing the log-likelihood:

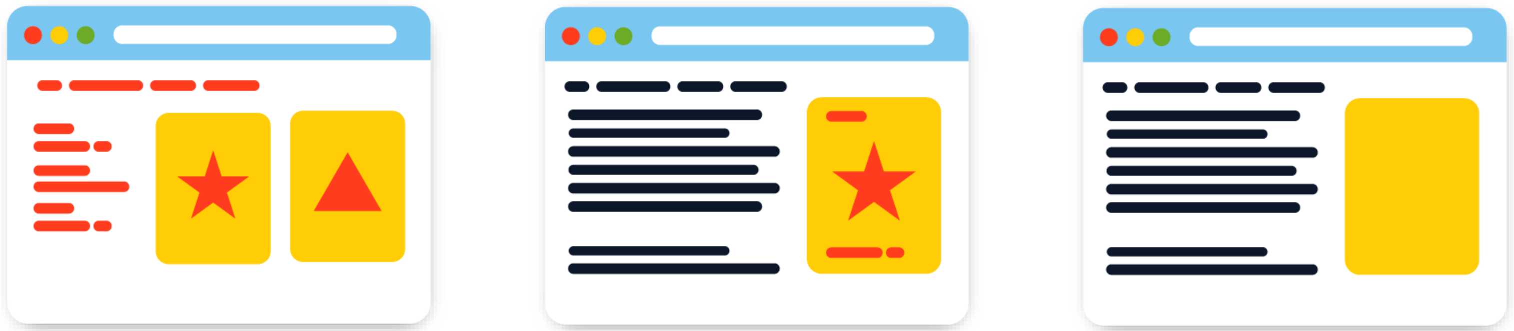
$$T(s, q, \gamma) \to \max_{\{s_i, \gamma_k, q_k\}}$$

- To obtain a **ranking** π over items, **sort** items according to their **scores**

Summary about pairwise comparisons

- Latent scores models for ranking from pairwise comparisons:

$$s_A \geq s_B \geq s_C$$



- To reduce bias from unreliable answers parameterize workers

w_k  $\longleftrightarrow$ "reliability" γ_k and "bias" q_k