Crowdsourcing with Contextual Uncertainty

Viet-An Nguyen, Peibei Shi, Jagdish Ramakrishnan,

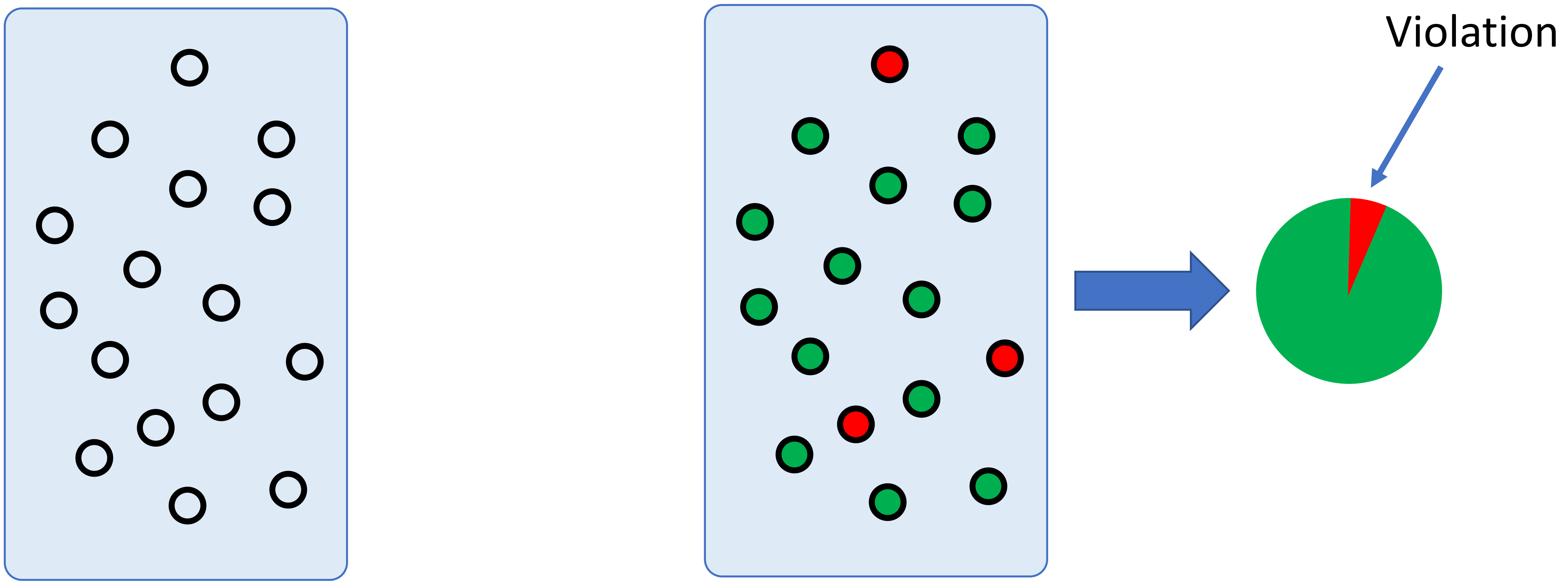
Narjes Torabi, Nimar S. Arora, Udi Weinsberg, Michael Tingley

Meta

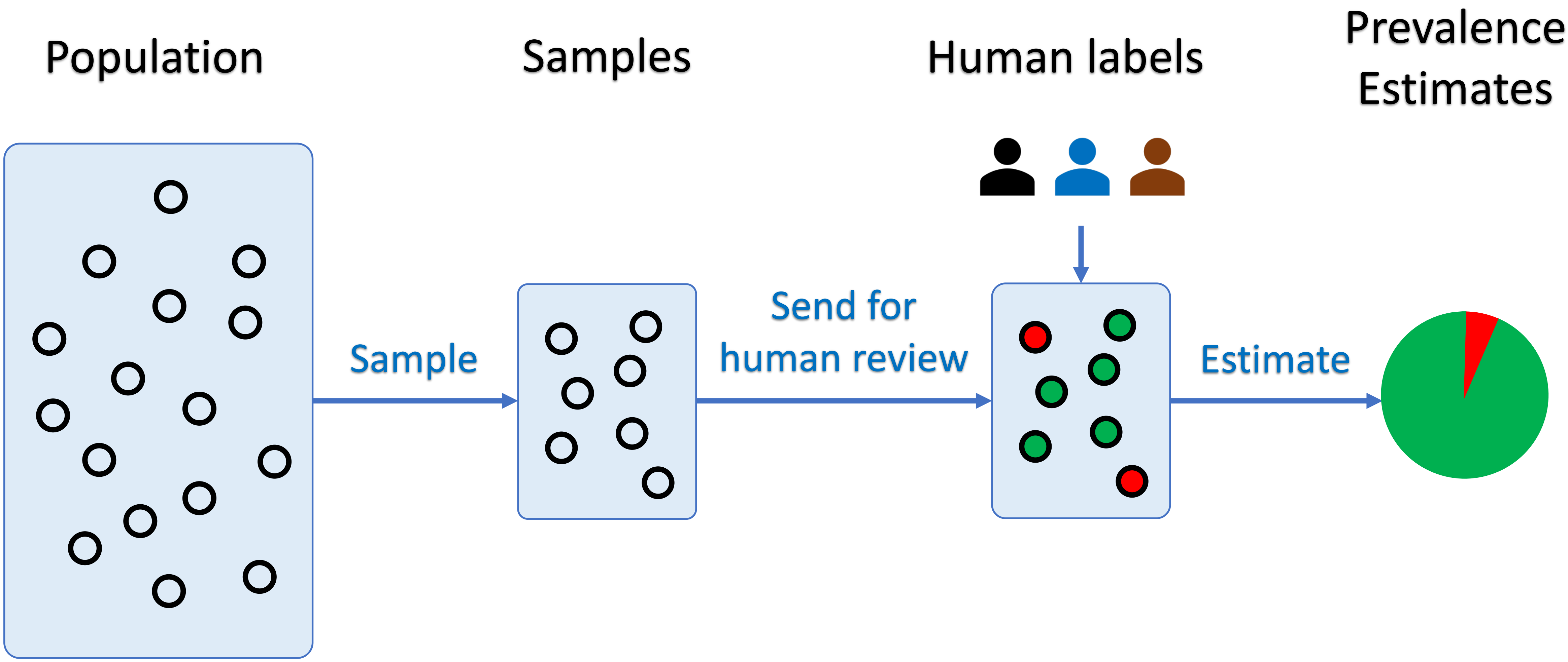
KDD 2022

Prevalence of violations

- Prevalence of violations: the rate at which policy violations occur



Estimating the prevalence



Two problems in practice

Low prevalence

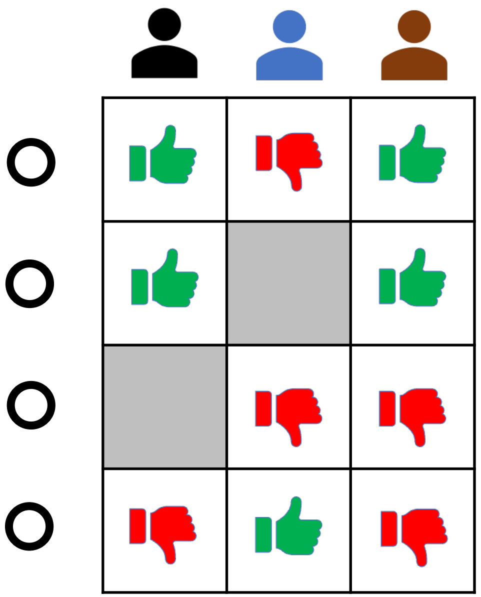
- Uniform sampling will result in too few violating examples
- Solution:
  - Sampling non-uniformly based on item’s features (i.e., context) which correlate with the likelihood of being a violation
  - In practice, there is a classifier which converts items’ features into a single score

Labeling mistakes

- Human labelers make mistakes
- Solution:
  - Send each item to multiple labelers and use statistical models to infer the true labels based on the observed labels

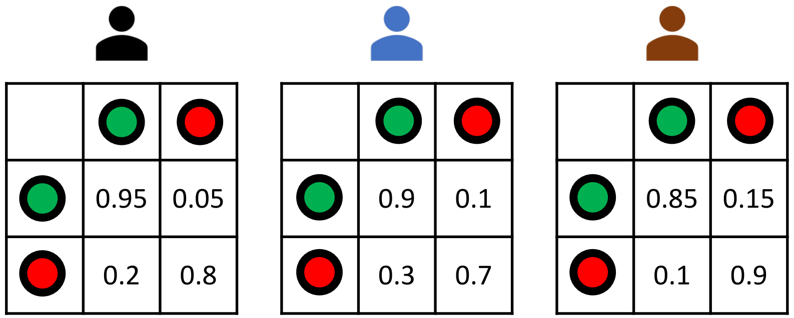
Modeling crowdsourced labels

- The "classic" Dawid-Skene (D&S) model

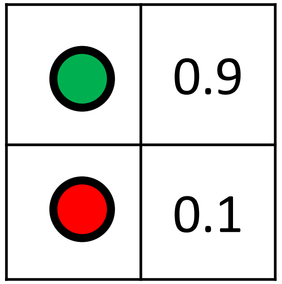


Crowdsourced data

Per-labeler Confusion Matrix



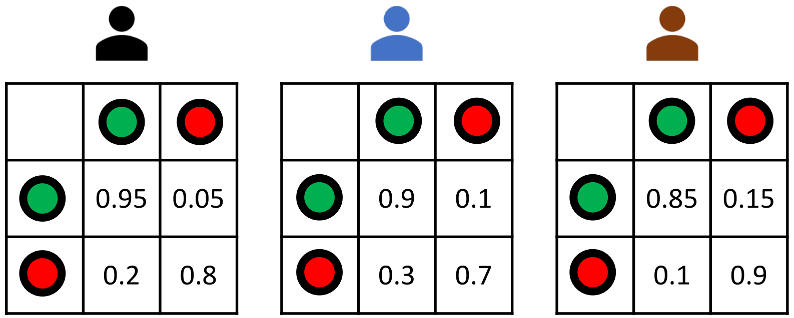
Overall Prevalence

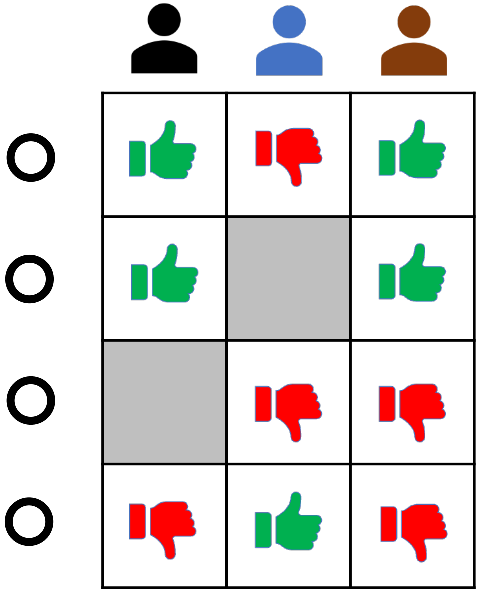


Modeling crowdsourced labels

- The “classic” Dawid-Skene (D&S) model

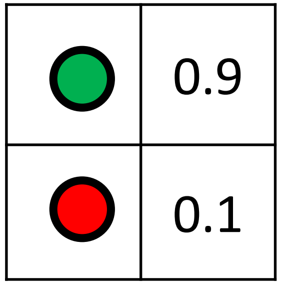
Per-labeler
Sensitivity (TPR) and
Specificity (TNR)
for binary label



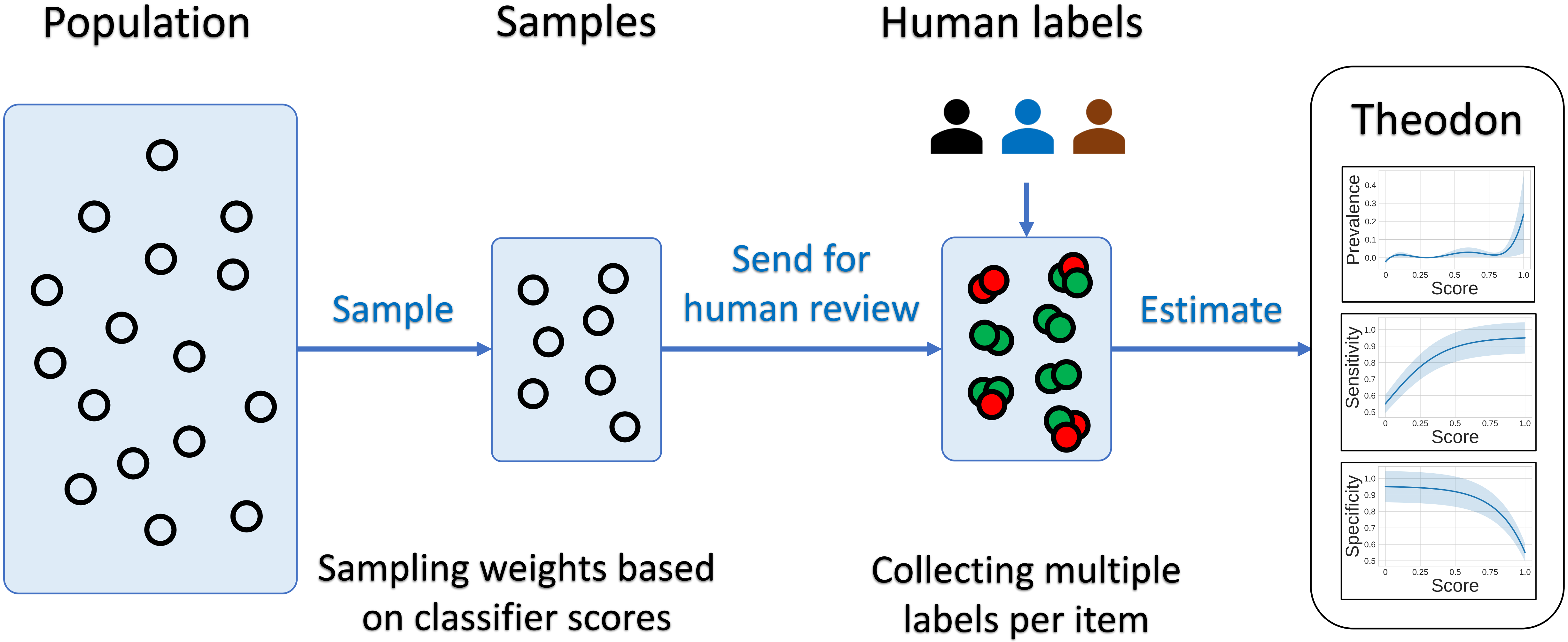


Crowdsourced
data

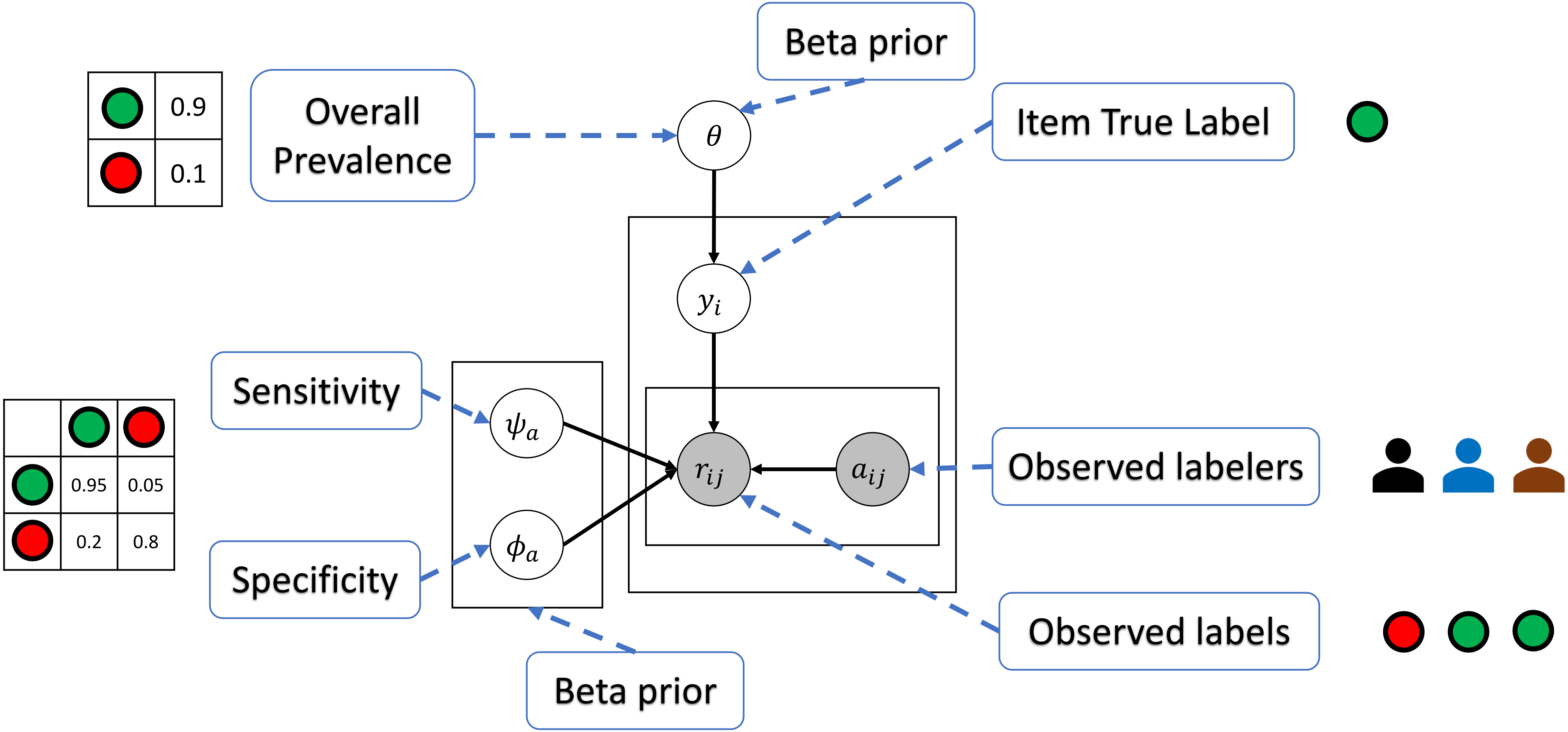
Overall
Prevalence



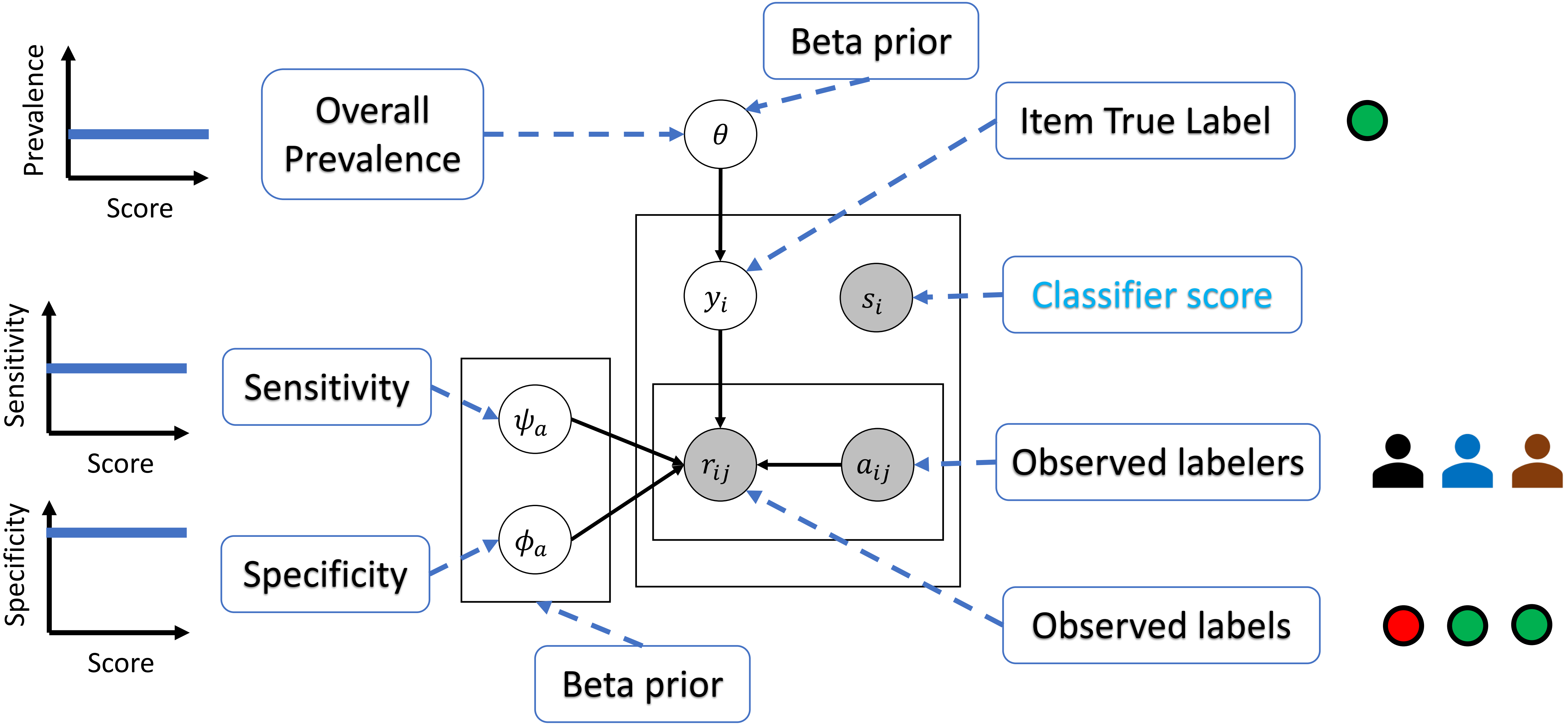
Theodon


Population
Samples
Human labels
Sample
Send for
human review
Estimate
Theodon
Prevalence
Score
Sensitivity
Score
Specificity
Score
Sampling weights based
on classifier scores
Collecting multiple
labels per item

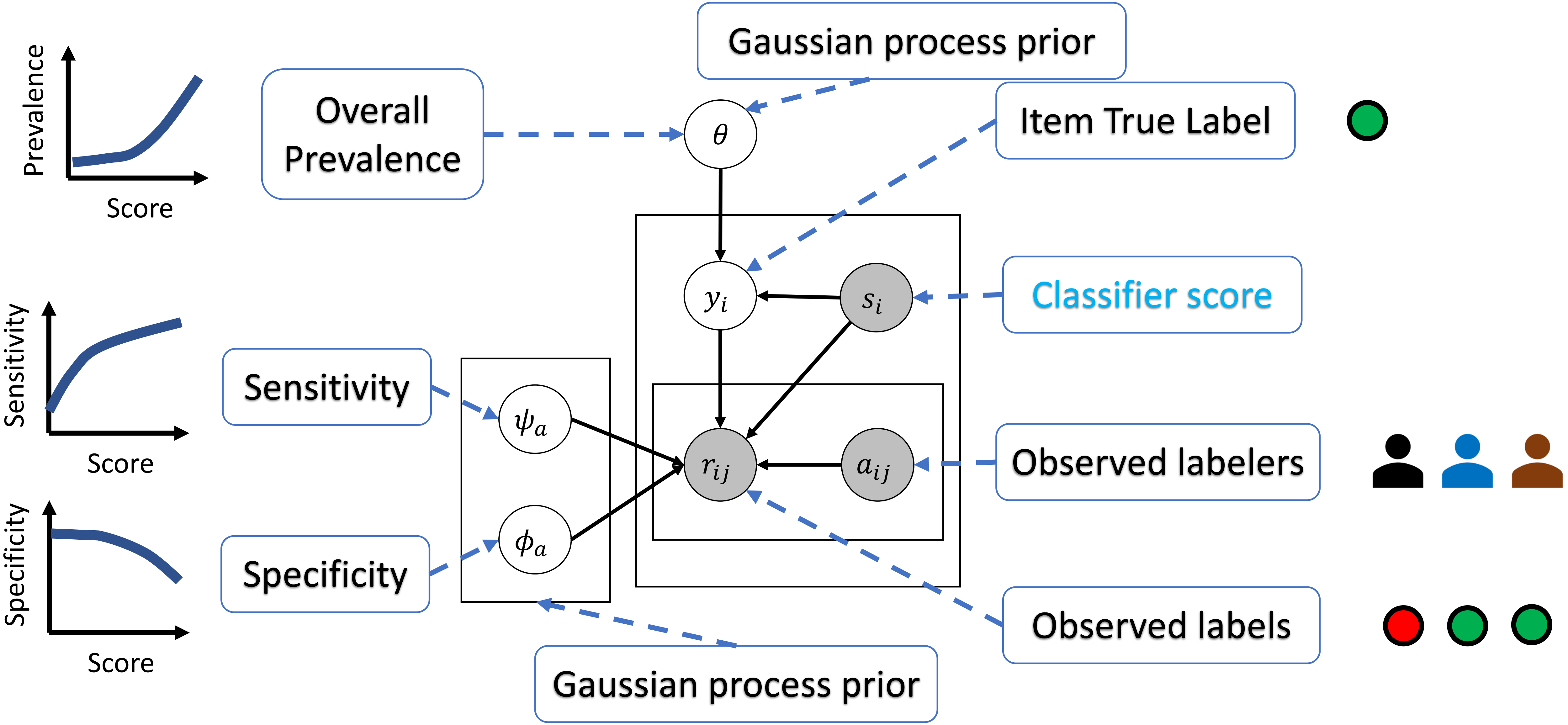
Generative process: D&S model



Generative process: D&S model



Generative process: Theodon



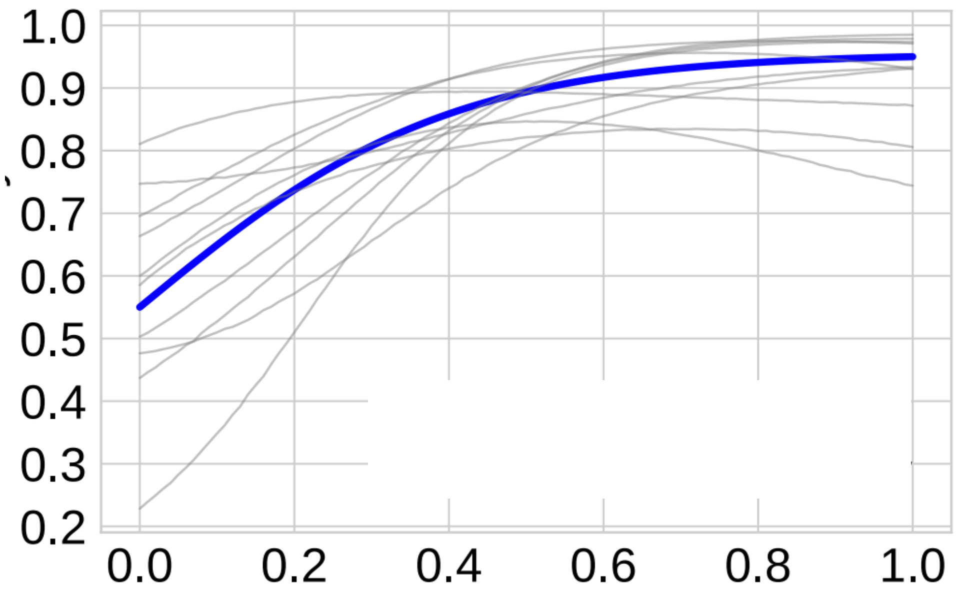
Gaussian processes (GPs)

A GP is a stochastic process which defines a probability distribution over the function

$$p(f \mid x) = \mathcal{GP}\left(m(x), K(x, x')\right)$$

- m(x): the mean function
- K(x, x'): the covariance function

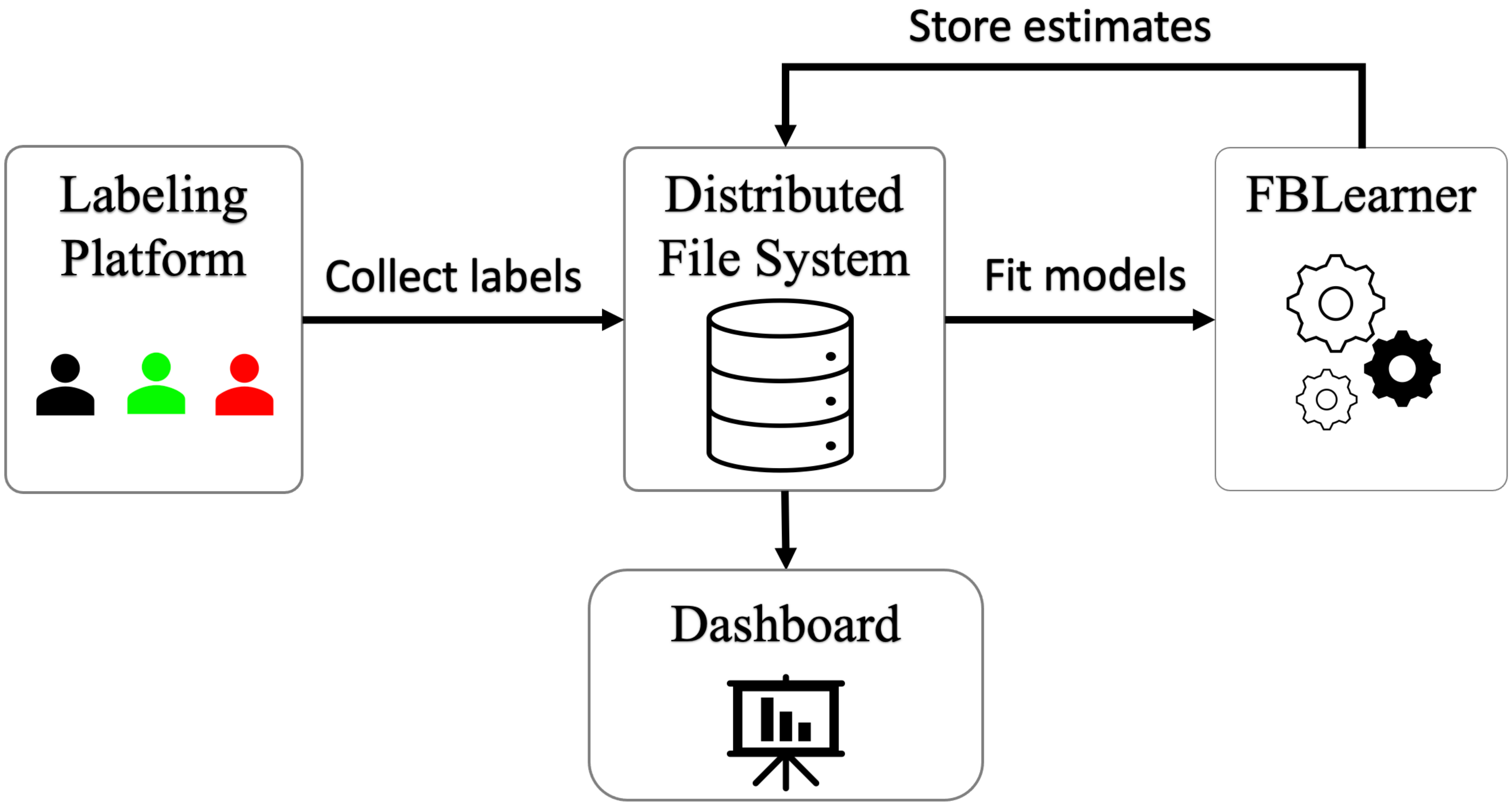
$$K_{\alpha,\rho}(x, x') = \alpha^2 \exp\left(-(x - x')^2 / 2\rho^2\right)$$



Related models

Baseline	Prevalence	Sensitivity & Specificity	References
FL-FL	Flat	Flat	Dawid & Skene (D&S) (1979)
LR-FL	Logistic Regression	Flat	Raykar et al. (ICML 2009, JMLR 2010)
GP-FL	Gaussian Process	Flat	Rodrigues et al. (ICML 2014)
LR-LR	Logistic Regression	Logistic Regression	Yan et al. (AISTATS 2010, MLJ 2014)
Theodon	Gaussian Process	Gaussian Process	Our work

Deployment at Meta



Empirical results

Applications

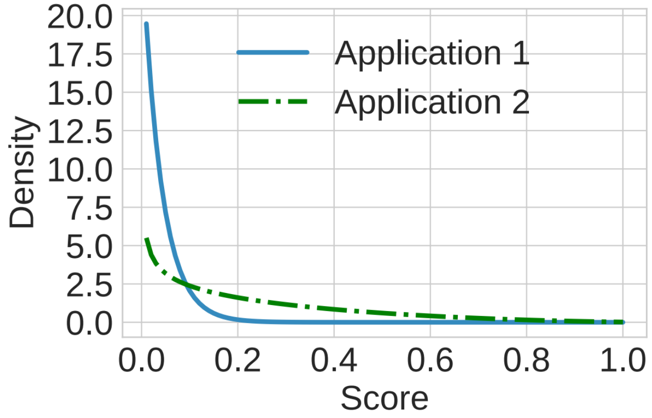
- Prevalence measurement
- Labeler performance measurement
- Classifier calibration
- Per-item label aggregation

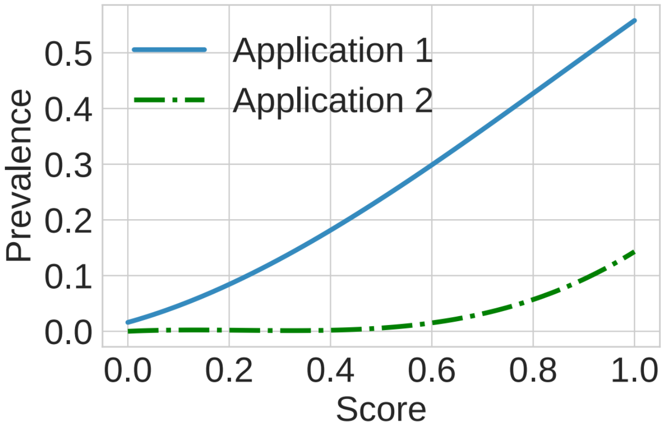
Datasets

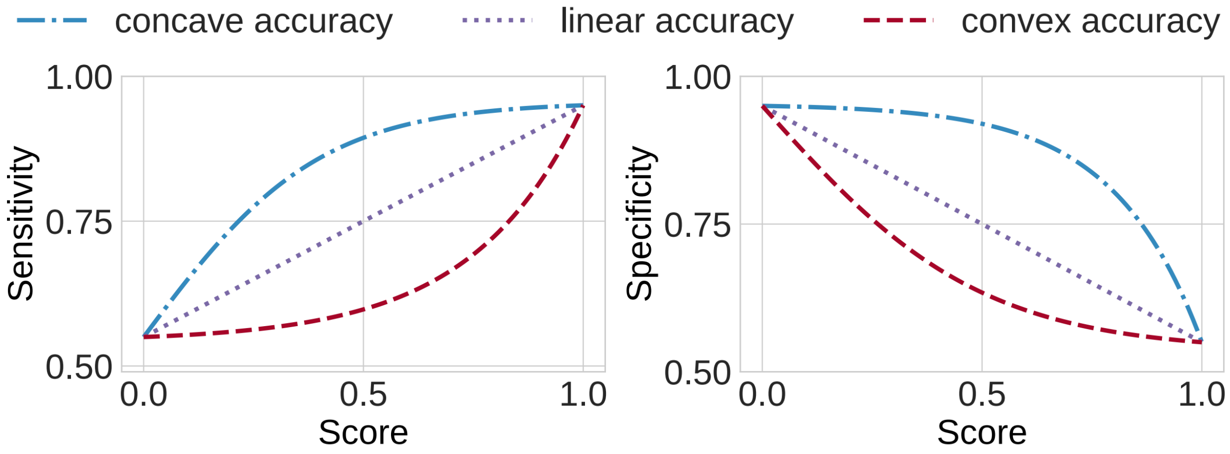
- Data generated from crowdsourcing applications at Meta
- Public crowdsourcing datasets: Music and Sentiment

Crowdsourced data for prevalence measurement at Meta

- Simulate data based on crowdsourcing applications at Meta







Experimental setup

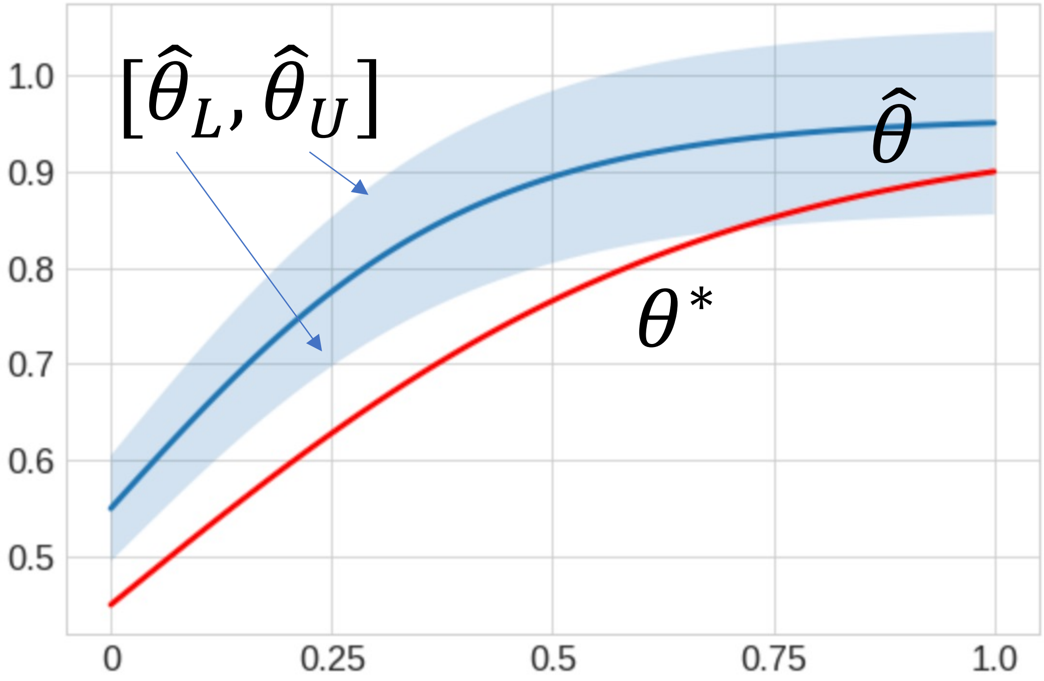
- Tasks
  - Prevalence measurement: estimating the **prevalence function**
  - Labeler performance measurement: estimating the **sensitivity and specificity functions** for each labeler
- Evaluation
  - Comparing the function estimate (mean $\hat{\theta}$ with 95%-CI $[\hat{\theta}_L, \hat{\theta}_U]$) with the true function θ^*

Mean absolute error (MAE) of the mean $\hat{\theta}$:

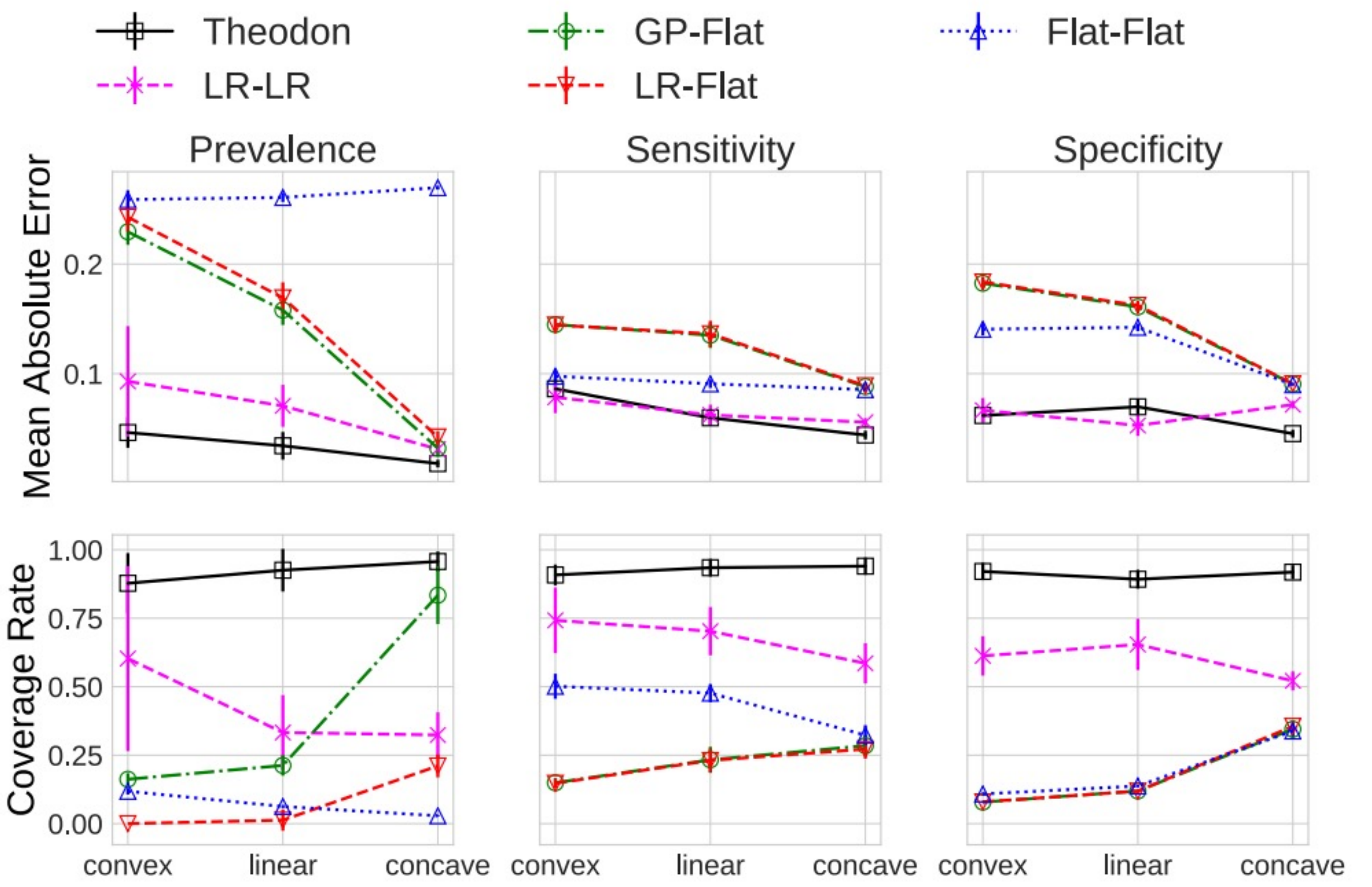
$$\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left| \hat{\theta}(s) - \theta^*(s) \right|$$

Coverage rate of the confidence interval (CI):

$$\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1} \left\{ \theta^*(s) \in [\hat{\theta}_L(s), \hat{\theta}_U(s)] \right\}$$



Results: prevalence & labeler performance measurement



Theodon consistently provides low Mean Absolute Error (MAE) while achieving high coverage rate compared to other baselines

Public crowdsourcing datasets

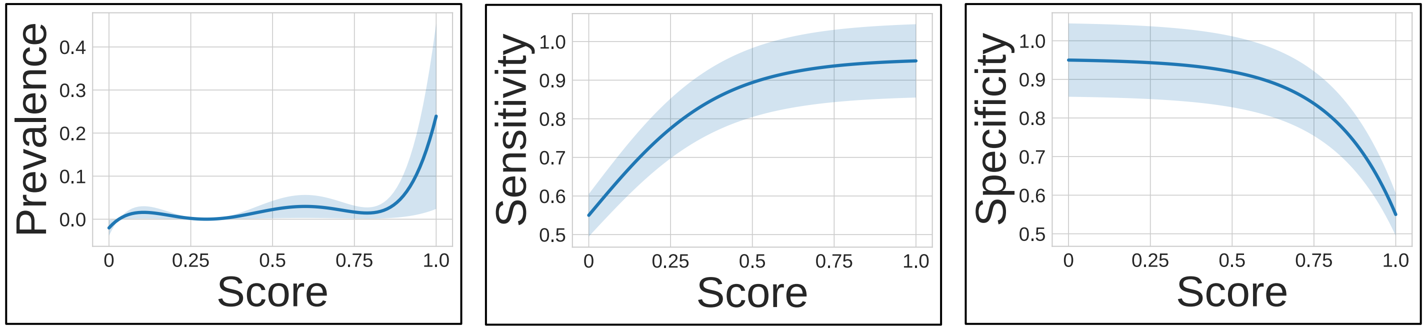
- Two public datasets by Rodrigues et al.: ***Music*** and ***Sentiment***
  - Each item has both **crowdsourced labels** and **a ground truth label**
- Label aggregation
  - Goal: inferring the ground truth label from crowdsourced labels
  - Metric: Area under the PR curve (AUC-PR)
- Classifier calibration
  - Goal: transforming the raw classifier scores into the true correctness probabilities using crowdsourced labels
  - Metric: Expected calibration error (ECE)

Results: label aggregation & classifier calibration

	Metric	l	Base	MV	FV	SNORKEL	FL-FL	LR-FL	GP-FL	LR-LR	THEODON
Sentiment	ECE	3	0.2304	0.1360	0.0754	0.0989	0.0952	0.0855	0.0814	0.0729	**0.0661**
		9	0.2394	0.1116	0.0887	0.0765	0.0842	0.0725	0.0726	0.0704*	**0.0656**
	AUC-PR	3	0.9210	0.8931	0.9169	0.9440	0.9283	0.9578	0.9604*	0.9641*	**0.9649**
		9	0.9345	0.9151	0.9498	0.9694	0.9536	0.9701	0.9716	0.9718	**0.9771**
Music	ECE	3	0.1816	0.0580	0.0514	0.0649	0.0487	0.0424*	**0.0403**	0.0470	0.0413*
		7	0.1835	0.0570	0.0635	0.0582	0.0448	0.0420*	**0.0413**	0.0443	0.0423*
	AUC-PR	3	0.7245	0.7139	0.7094	0.7808	0.7719	0.8276	0.8513*	0.8276	**0.8619**
		7	0.7660	0.7261	0.7399	0.8132	0.7906	0.8378*	0.8474*	0.8329*	**0.8515**

Conclusion

- Theodon: a system developed and deployed at Meta to model crowdsourced labels by capturing the dependencies of **label's prevalence** and **labelers' performance** on the input classifier score using Gaussian Processes



- Extensive empirical results on Meta's and public datasets

Thank you!