A Bayesian Framework for Modeling Human Evaluations

Himabindu Lakkaraju[1] Jure Leskovec[1] Jon Kleinberg[2] Sendhil Mullainathan[3]

[1]Stanford University

[2]Cornell University

[3]Harvard University

SIAM International Conference on Data Mining
Apr. 30th – May 2nd, 2015

Goal: Evaluating the Evaluators

Goal: Evaluating the Evaluators

How good are evaluators?
What mistakes are they making?

The setting: Human Evaluations

Evaluator *j*

Items

.
.
.

The setting: Human Evaluations

Evaluator *j* **Decisions of *j*** **Items**

Gray catbird →

Green tanager →

Yellow black bird →

Yellow black bird →

.
.
.

The setting: Human Evaluations

Evaluator *j*	Decisions of *j*	Items	True Labels
	Gray catbird		Gray catbird
	Green tanager		Green tanager
	Yellow black bird		Yellow black bird
	Yellow black bird		Green tanager
		⋮	

How good are evaluators?

Evaluator *j*

Decision of evaluator *j* on item *i*

Assigned Label $a_{j,i}$

Gray catbird

Item *i*

True Label t_i

Gray catbird

How good are evaluators?

Evaluator *j*

Decision of evaluator *j* on item *i*

Item *i*

True Label t_i

Gray catbird

Assigned Label $a_{j,i}$

Gray catbird

How do we evaluate the quality of evaluator j?

Quality of evaluator $j \approx P(a_{j,i} == t_i)$

What mistakes are they making?

Evaluator *j*

Decision of evaluator *j* on item *i*

Item *i*

True Label t_i

Gray catbird

What mistakes are they making?

Evaluator *j*

Decision of evaluator *j* on item *i*

Item *i*

True Label t_i

Gray catbird

Assigned label

True label

$p_{1,1}$	$p_{1,2}$	$p_{1,3}$
$p_{2,1}$	$p_{2,2}$	$p_{2,3}$
$p_{3,1}$	$p_{3,2}$	$p_{3,3}$

What mistakes are they making?

Evaluator *j*

Decision of evaluator *j* on item *i*

Item *i*

True Label t_i

Gray catbird

Assigned label

True label

$p_{1,1}$	$p_{1,2}$	$p_{1,3}$
$p_{2,1}$	$p_{2,2}$	$p_{2,3}$
$p_{3,1}$	$p_{3,2}$	$p_{3,3}$

Probability that j assigns an item that 'truly' belongs to class 1 to class 3

Our Goals

- Discover interesting patterns in the collective evaluation process
  - Ex: People with color-blindness often confuse between green tanager and blue catbird

Our Goals

- Discover interesting patterns in the collective evaluation process
  - Ex: People with color-blindness often confuse between green tanager and blue catbird
- Model individual evaluator behavior and decisions
  - Ex: Evaluator j is likely to label a green tanager as yellow blackbird with a probability of 0.6

Challenges & Considerations

- **Challenges**:
  - True labels are hard to obtain

Challenges & Considerations

- **Challenges**:
  - True labels are hard to obtain
- **Considerations:**
  - Certain evaluators have similar decision making styles
  - Certain items might be confused in similar ways

Problem Setting

- **Given:**
  - J: Set of evaluators
  - I: Set of items
  - D: Decisions made by evaluators on items
  - K: Set of class labels to be assigned to items
  - **a**: Attributes of evaluators
  - **b**: Attributes of items
  - **z**: Small fraction of true labels of items

Problem Setting

- **Output:**
  - Discover groups of evaluators and items that share similar decision patterns
  - For each such group, infer the corresponding confusion matrix
  - Infer true labels of (remaining) items in I

Our Approach

- A Bayesian framework of a series of models that balances the trade-off between modeling individual and collective behavior

Our Approach

- A Bayesian framework of a series of models that balances the trade-off between modeling individual and collective behavior

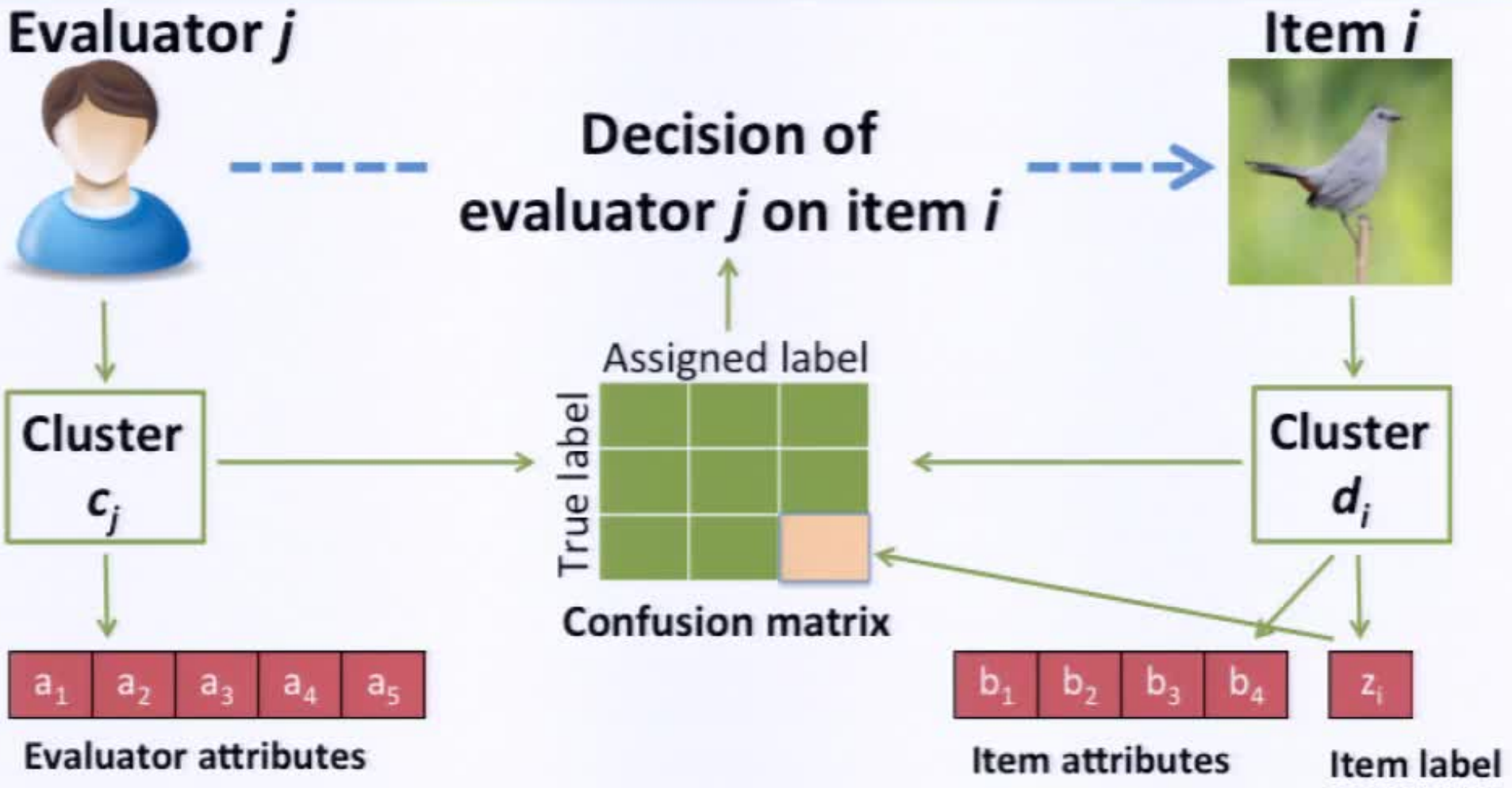
- We present three models:

  1: Joint Confusion: Joint inference of latent groups of evaluators and items

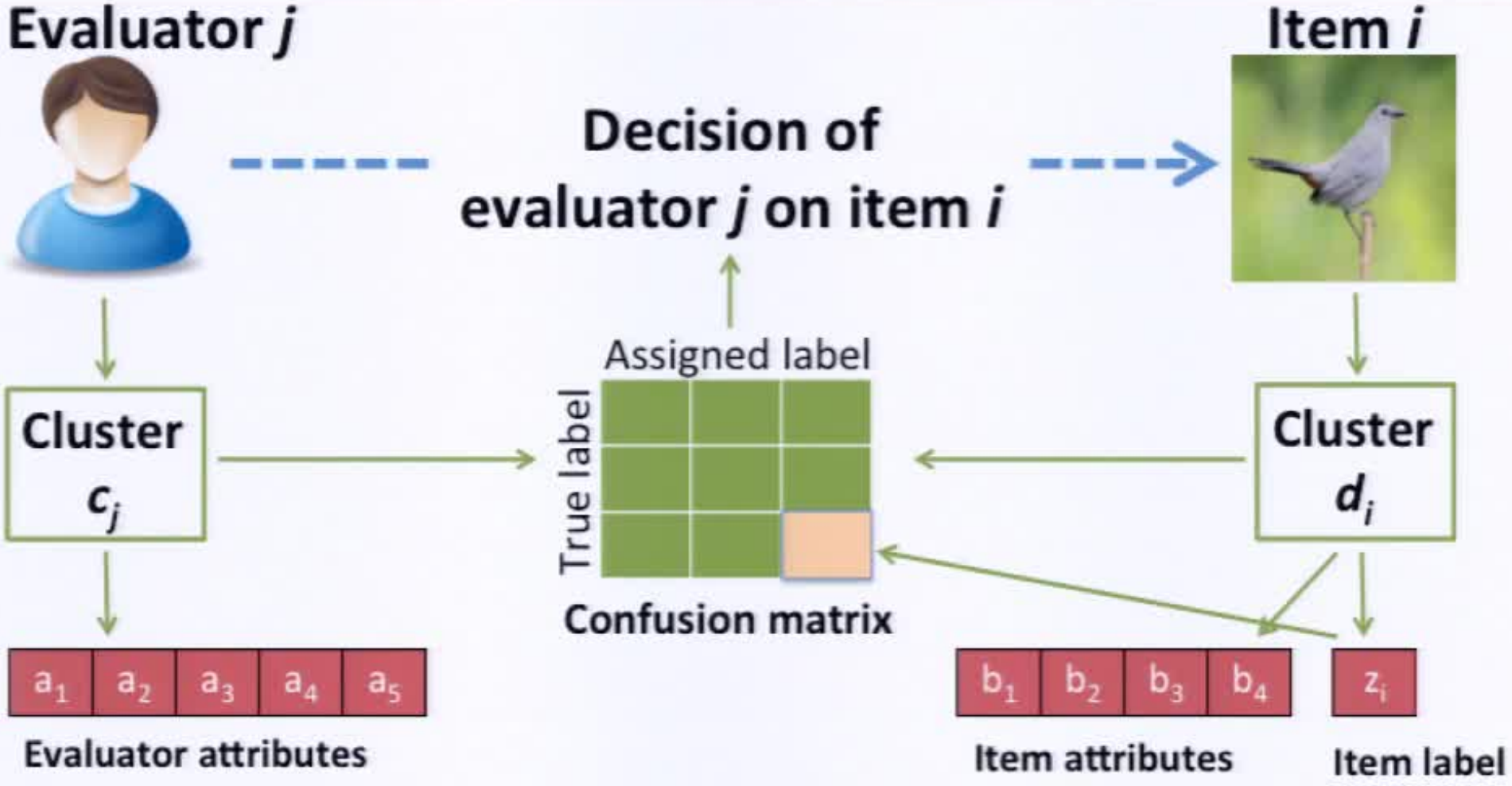
  2: Evaluator Confusion: Infer latent groups of evaluators

  3: Item Confusion: Infer latent groups of items

1) Joint Confusion Model

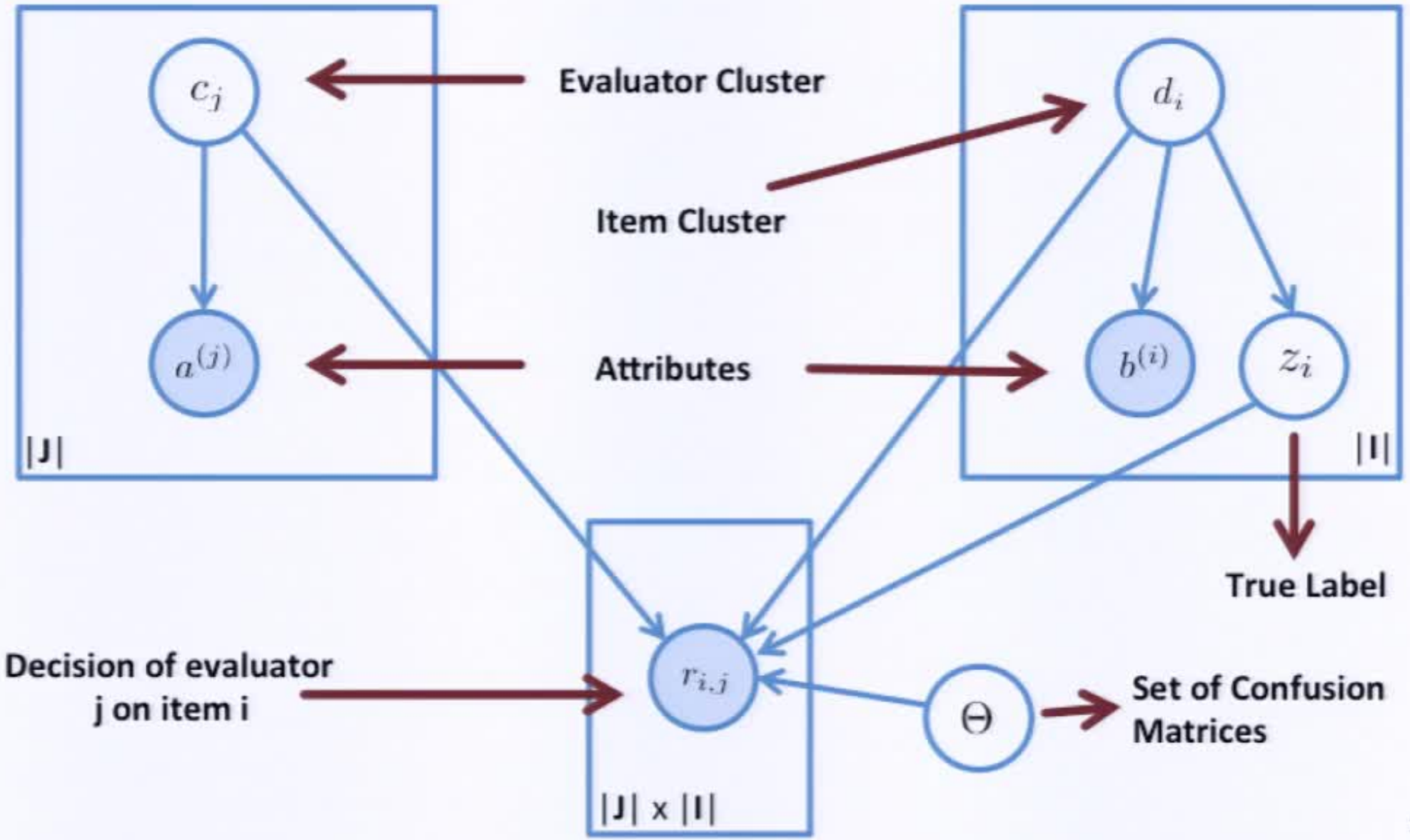


1) Joint Confusion Model



Similar evaluators share confusion matrices when deciding on similar items

Joint Confusion Model: Plate notation



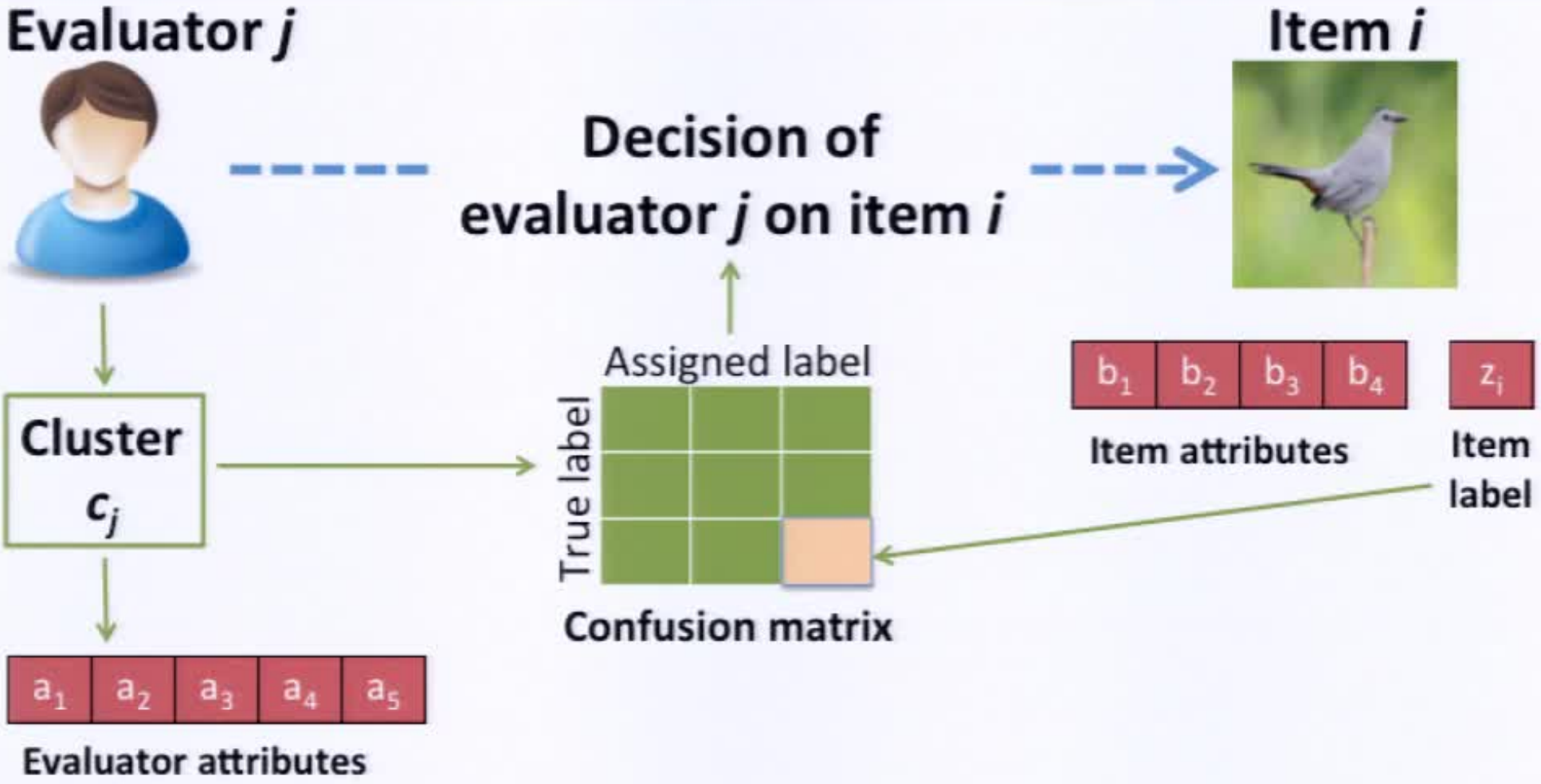
Inference of Joint Confusion Model

- Approximate inference using Collapsed Gibbs sampling
  - Integrate out all the intermediate latent variables
  - We only sample for c_j, z_i, d_i
- Conditional distribution for cluster assignment c_j of evaluator j can be computed as:

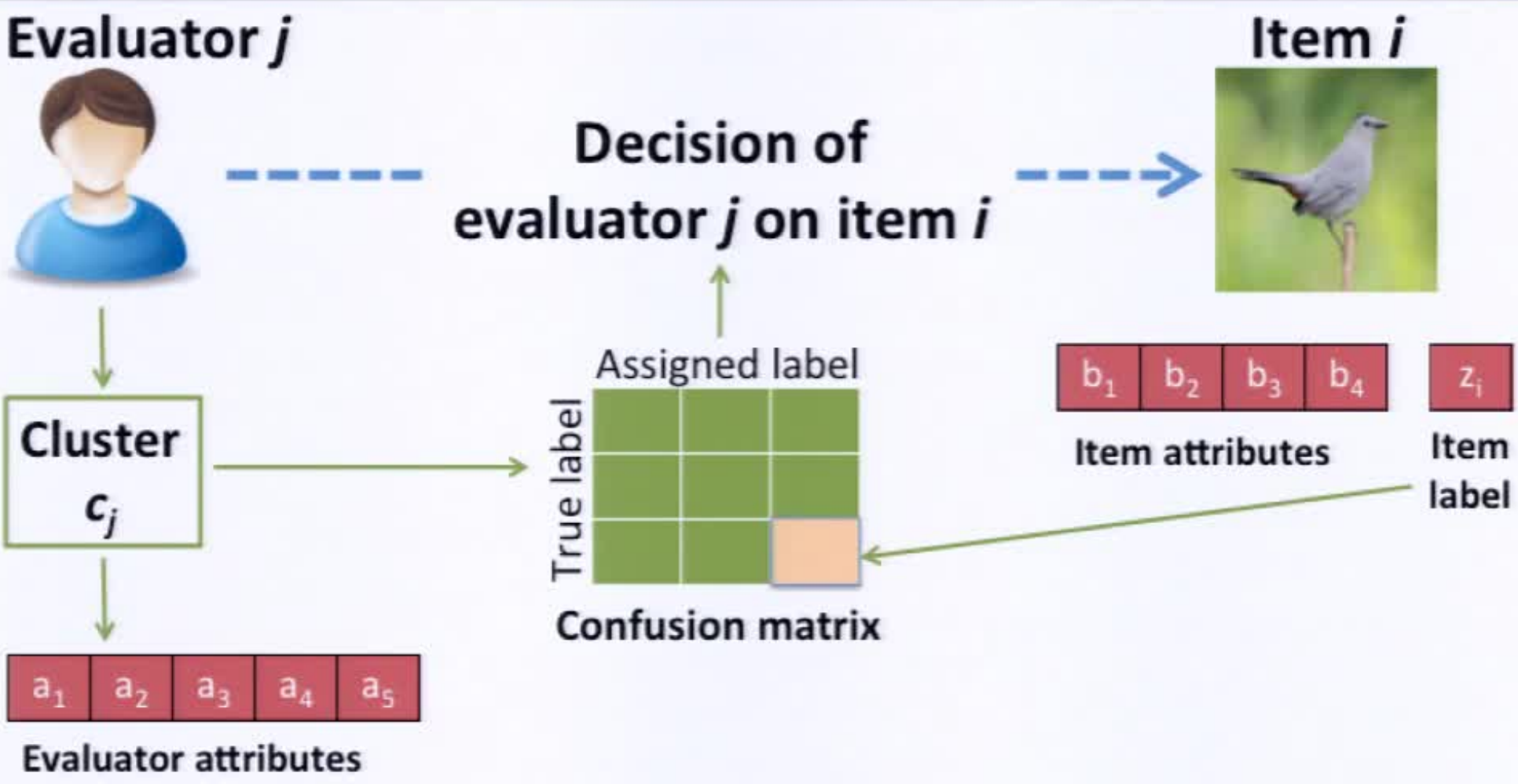
$$P(c_j = c|\boldsymbol{c}^{-j}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{a}) \propto P(c_j = c|\boldsymbol{c}^{-j})$$

$$\times \prod_{items\ i\ labeled\ by\ j} P(r_{i,j}|\boldsymbol{r}^{-j}, \boldsymbol{c}, \boldsymbol{z}) \times \prod_{n=1}^{N} P(a_n^{(j)}|\boldsymbol{a}^{-j}, \boldsymbol{c})$$

2) Evaluator Confusion Model

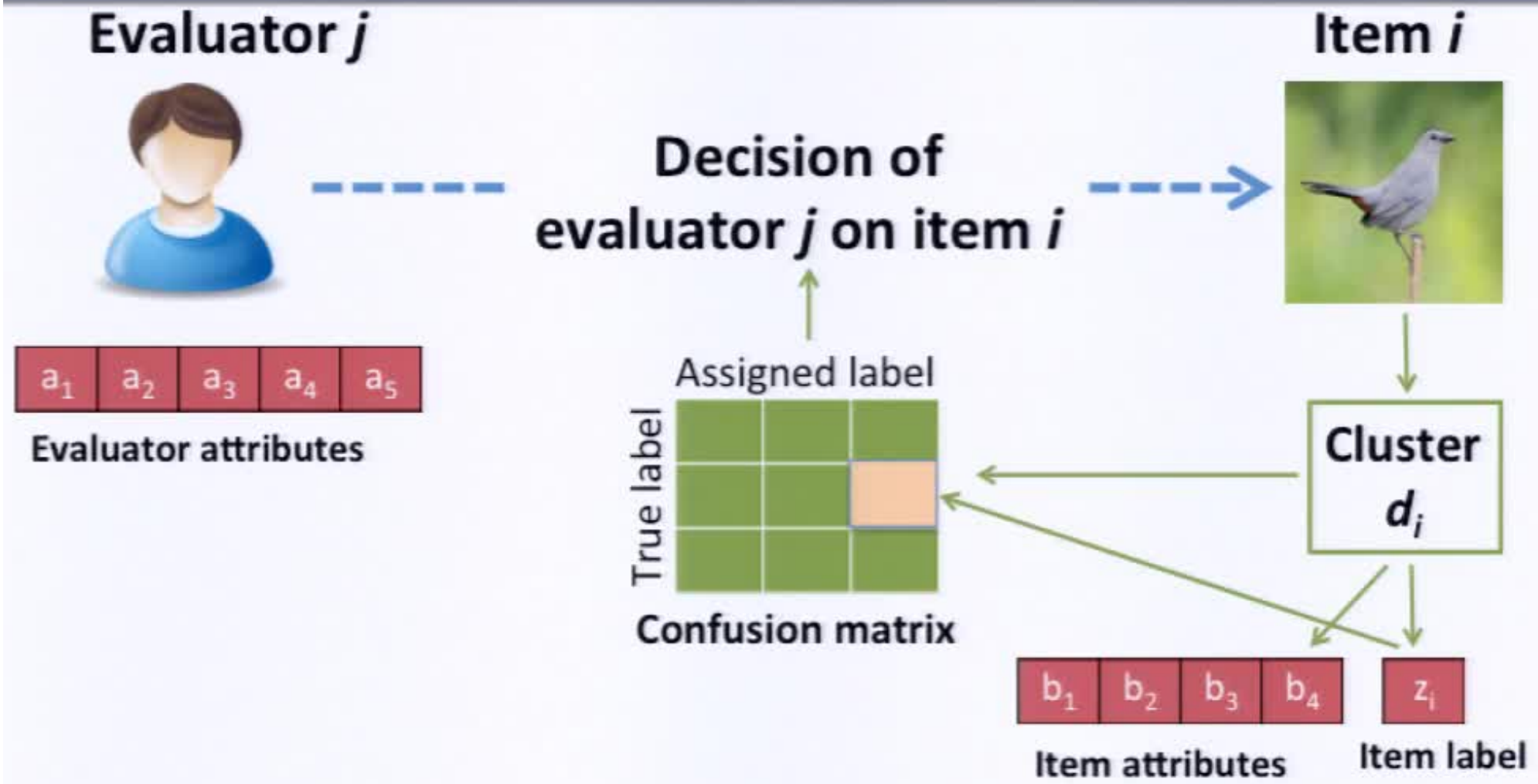


2) Evaluator Confusion Model

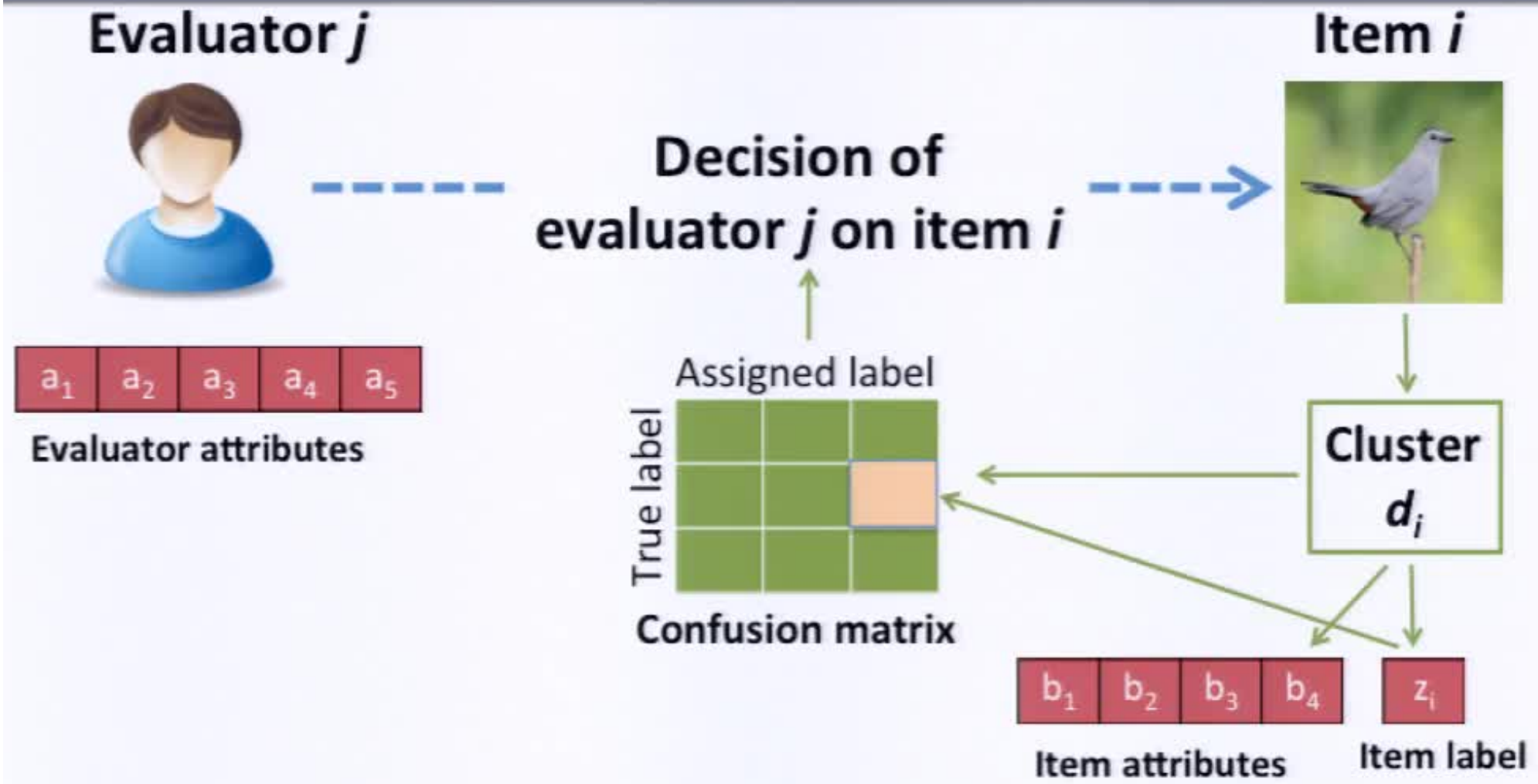


Similar evaluators share confusion matrices

3) Item Confusion Model



3) Item Confusion Model



All evaluators share confusion matrices for similar items

Experimental Evaluation

- Quantitative Evaluation:
  1) Estimating confusion matrices
  2) Predicting true labels of items
  3) Predicting evaluator decisions

- Qualitative Analysis:
  – Insights into patterns of evaluation

Dataset Description

- **Real world datasets:**

Dataset	# of Evaluators	# of Items	# of Decisions
Student Exams	4000	107	214000
Peer Grading	5000	6224	19208
Text Labeling	152	4000	11400
Image Labeling	101	450	3915

- **Attributes:**
  - Evaluator properties: age, gender, occupation etc.
  - Item properties: topic, length, color etc.

Experimental Setting

- Weakly supervised setting:
  - True labels of only 15% of the items are available to the model
- Inference process is executed till the approximate convergence of log-likelihood
- Number of clusters:
  - Bayesian Information Criterion (BIC)
  - Non-parametric versions of the models

Baselines

- Dawid-Skene Model [JRSS, 1979]:
  - One confusion matrix per evaluator
  - Evaluator confusions are independent
- Single Confusion Model [ICML, 2012]:
  - One confusion matrix shared by all the evaluators
- Other baselines:
  - Logistic Regression for predicting item labels and evaluator decisions

1) Estimating Confusion Matrices

- **Metric:** Mean absolute error computed across all the entries in all the confusion matrices

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.25	0.26	0.23	0.25
Evaluator Confusion	0.32	0.28	0.24	0.27
Item Confusion	0.34	0.34	0.29	0.29
Baseline (Dawid-Skene)	0.33	0.35	0.36	0.32
Baseline (Single Confusion)	0.46	0.42	0.48	0.41

Lower values are better

1) Estimating Confusion Matrices

- **Metric:** Mean absolute error computed across all the entries in all the confusion matrices

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.25	0.26	0.23	0.25
Evaluator	0.32	0.28	0.24	0.27
Iter				9
(Da				2
Baseline (Single Confusion)	0.46	0.42	0.48	0.41

Joint Confusion achieves a gain of about 27% over the baselines

Lower values are better

2) Predicting Labels of Items

- **Metric:** Accuracy of predicting item labels

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.68	0.69	0.69	0.70
Evaluator Confusion	0.65	0.66	0.65	0.60
Item Confusion	0.64	0.64	0.65	0.60
Baseline (Dawid-Skene)	0.60	0.62	0.65	0.60
Baseline (Single Confusion)	0.56	0.57	0.56	0.51
Baseline (Logistic Regression)	0.38	0.53	0.51	0.57

Higher values are better

2) Predicting Labels of Items

- **Metric:** Accuracy of predicting item labels

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.68	0.69	0.69	0.70
Evaluator Confusion	0.65	0.66	0.65	0.60
Item C				60
Bas (Dawi				60
Bas (Single Confusion)				51
Baseline (Logistic Regression)	0.38	0.53	0.51	0.57

Joint Confusion achieves a gain of about 12% over the baselines

Higher values are better

3) Predicting Evaluator Decisions

- So far, evaluator decisions are observed

3) Predicting Evaluator Decisions

- So far, evaluator decisions are observed
- We can predict evaluator decisions using a supervised setting:
  - Infer latent clusters and matrices using 90% of the data
  - Use inferred cluster assignments and matrices to predict residual 10% of evaluator decisions
  - 10-fold cross validation

3) Predicting Evaluator Decisions

- **Metric:** Accuracy of predicting decisions

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.74	0.70	0.71	0.71
Evaluator Confusion	0.72	0.70	0.66	0.69
Item Confusion	0.72	0.69	0.64	0.66
Baseline (Dawid-Skene)	0.61	0.64	0.60	0.64
Baseline (Single Confusion)	0.58	0.53	0.51	0.56
Baseline (Logistic Regression)	0.63	0.66	0.68	0.68

Higher values are better

3) Predicting Evaluator Decisions

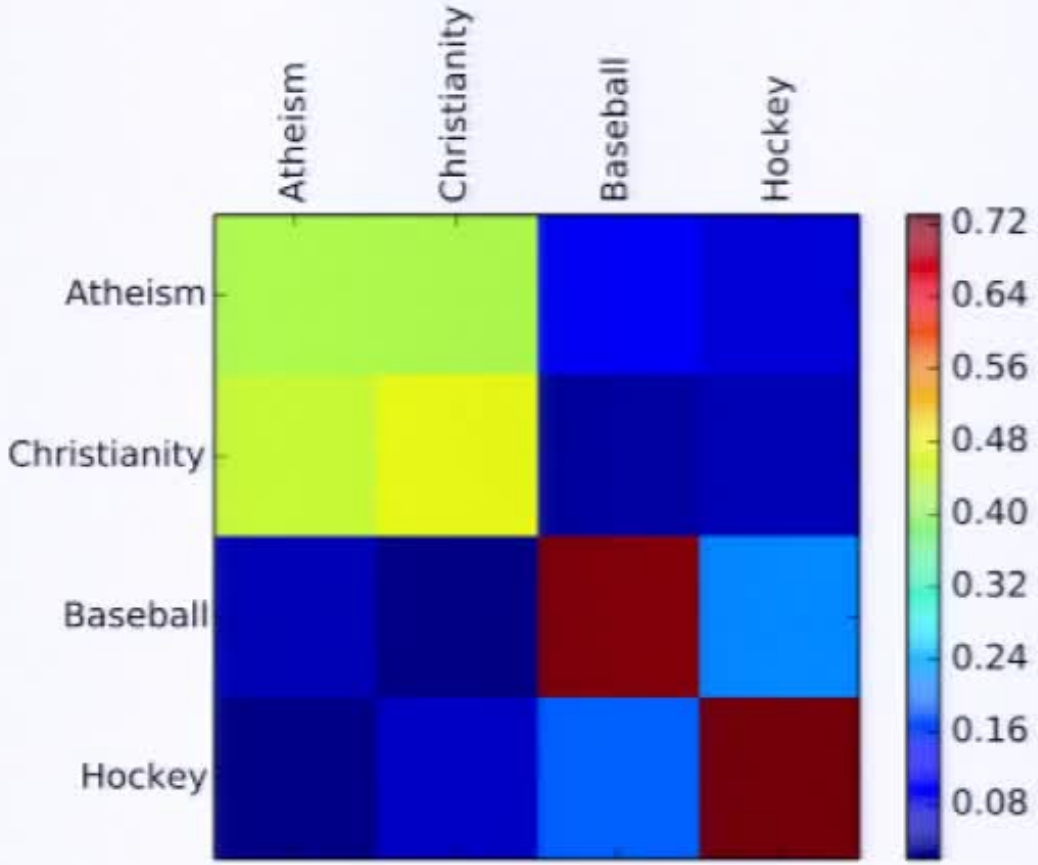
- **Metric:** Accuracy of predicting decisions

Model	Student Exams	Peer Grading	Text Labeling	Image Labeling
Joint Confusion	0.74	0.70	0.71	0.71
Evaluator Confusion	0.72	0.70	0.66	0.69
Item C				66
Bas (Dawi				64
Bas (Single Confusion)				66
Baseline (Logistic Regression)	0.63	0.66	0.68	0.68

Joint Confusion achieves a gain of about 8% over the baselines

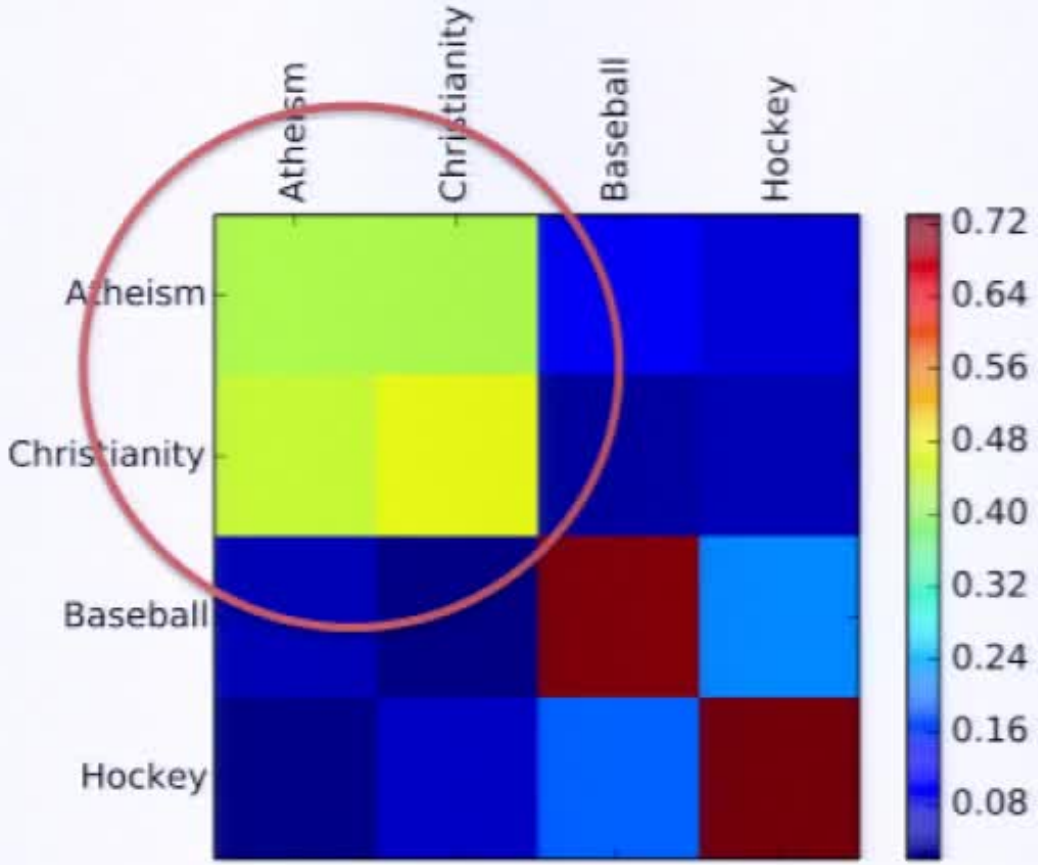
Higher values are better

Qualitative Insights – Joint Confusion



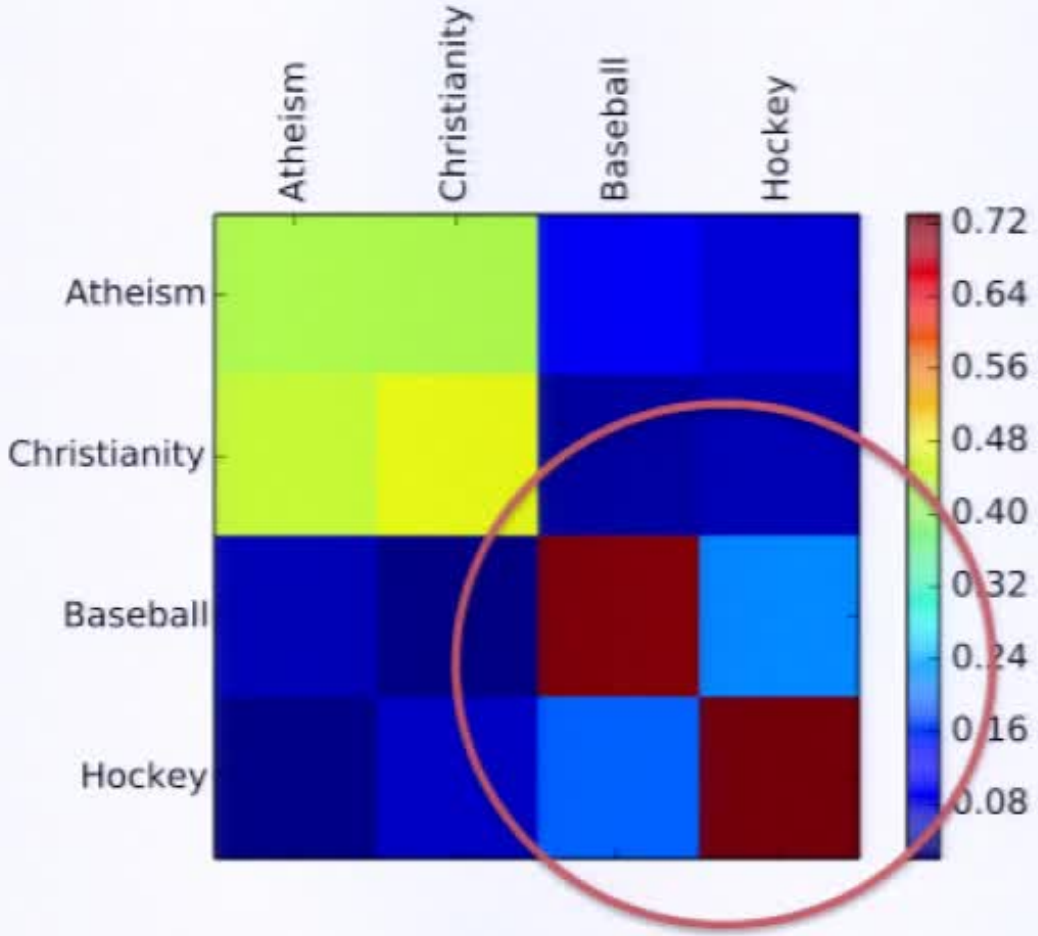
Document Labeling Task: Male evaluators < 23 years old and documents with length < 20 words

Qualitative Insights – Joint Confusion



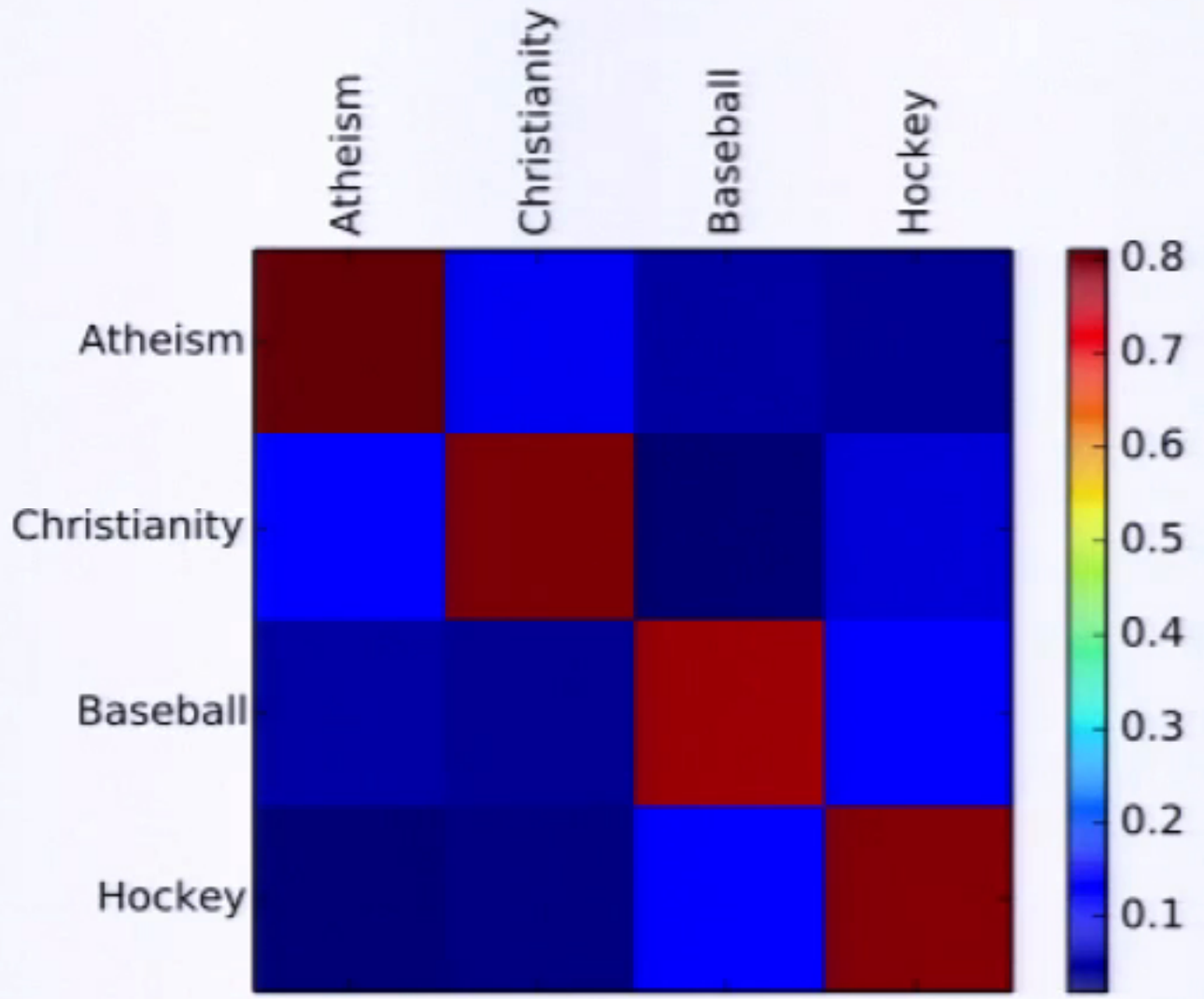
Document Labeling Task: Evaluators are unable to distinguish between atheism and Christianity when documents are short

Qualitative Insights – Joint Confusion



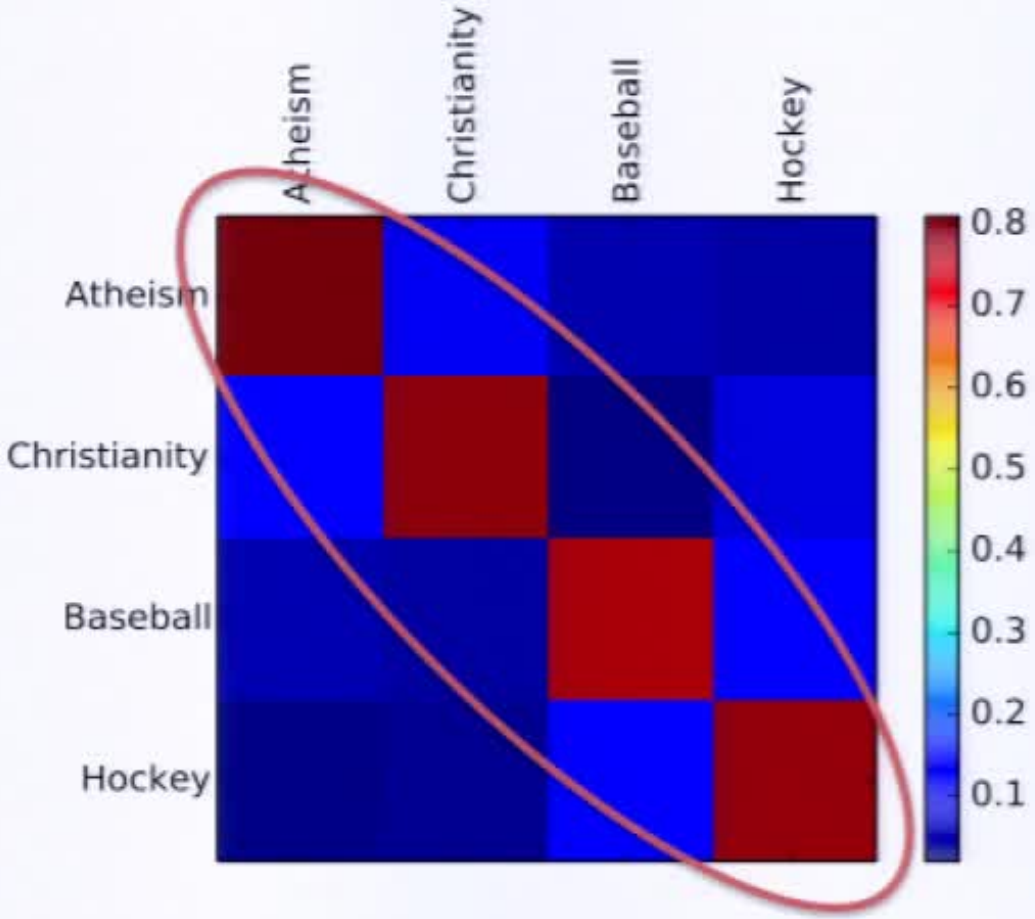
Document Labeling Task: Evaluators are able to differentiate between documents on hockey and baseball even with short document lengths

More Qualitative Insights



Document Labeling Task: Female evaluators with low self-reported confidence scores

More Qualitative Insights



Document Labeling Task: Female evaluators with low self-reported confidence scores are highly accurate

Summary

- A Bayesian framework of a series of models which:
  - Identifies latent groupings of evaluators and items
  - Infers corresponding confusion matrices
  - Infers true labels of items

Summary

- A Bayesian framework of a series of models which:
  - Identifies latent groupings of evaluators and items
  - Infers corresponding confusion matrices
  - Infers true labels of items

- Our framework
  - Facilitates a fine-grained analysis of evaluator quality
  - Provides aggregate insights into patterns of evaluation
  - Mimics real world settings where true labels are hard to obtain

Summary

- A Bayesian framework of a series of models which:
  - Identifies latent groupings of evaluators and items
  - Infers corresponding confusion matrices
  - Infers true labels of items
- Our framework
  - Facilitates a fine-grained analysis of evaluator quality
  - Provides aggregate insights into patterns of evaluation
  - Mimics real world settings where true labels are hard to obtain
- Applications:
  - Recommending evaluators based on item characteristics
  - Recommending training programs based on evaluator skill