Modelling Human Uncertainty

How to teach machines when experts disagree with each other

Ryutaro Tanno
University College London, UK

“Academic” Deep Learning Setup

“Academic” Deep Learning Setup

- Predict labels (e.g. dog or cat) from the given input (e.g. pictures).

“Academic” Deep Learning Setup

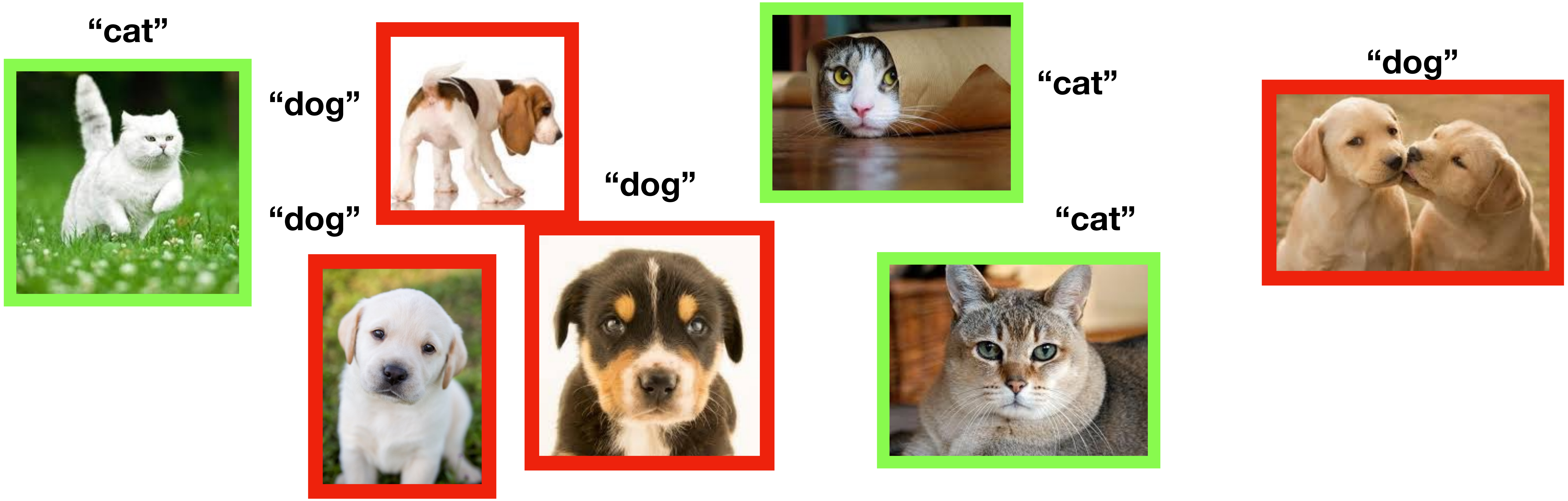
- Predict labels (e.g. dog or cat) from the given input (e.g. pictures).
- Trained with **many** examples of inputs and labels

“Academic” Deep Learning Setup

- Predict labels (e.g. dog or cat) from the given input (e.g. pictures).
- Trained with **many** examples of inputs and labels

“Academic” Deep Learning Setup

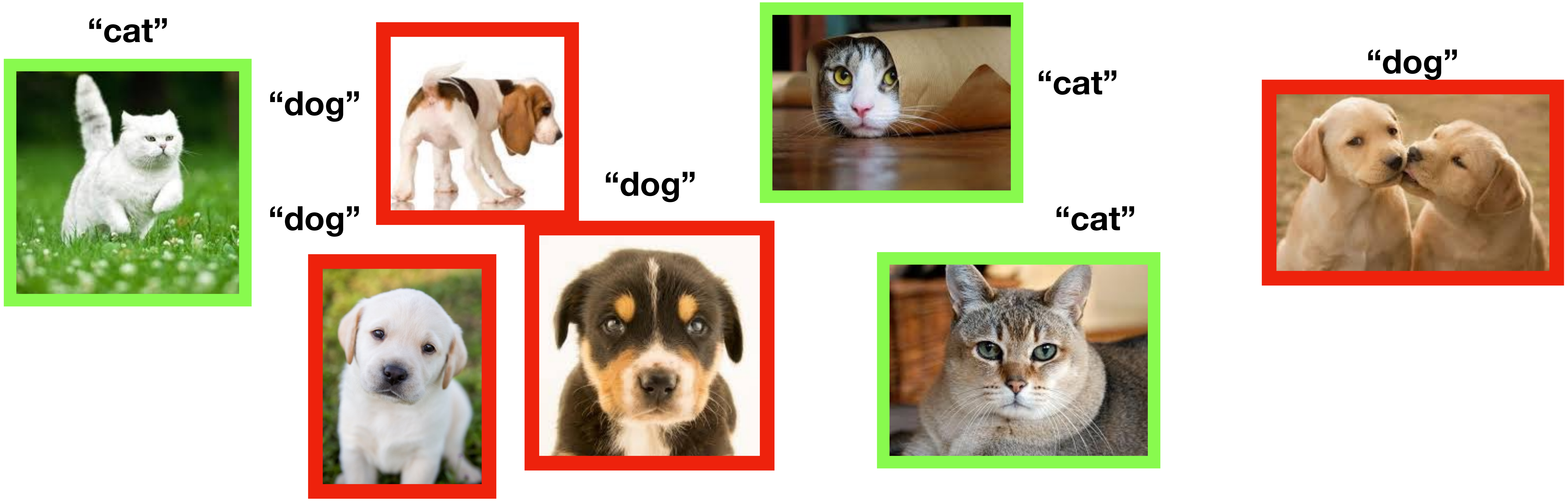
- Predict labels (e.g. dog or cat) from the given input (e.g. pictures).
- Trained with **many** examples of inputs and labels



“Academic” Deep Learning Setup

- Predict labels (e.g. dog or cat) from the given input (e.g. pictures).

- Trained with **many** examples of inputs and labels



- **Clean** data => great performance!

Deep Learning in the “wild”

Deep Learning in the “wild”

- Practical applications require learning from **noisy labels**

Deep Learning in the “wild”

- Practical applications require learning from **noisy labels**
- Multiple annotators of different **skill levels** and **biases**

Deep Learning in the “wild”

- Practical applications require learning from **noisy labels**
- Multiple annotators of different **skill levels** and **biases**



Deep Learning in the "wild"

- Practical applications require learning from **noisy labels**
- Multiple annotators of different **skill levels** and **biases**

David
(bird expert)

"Canada Goose"

"Red-necked Grebe"

"Am. Black Duck"

Deep Learning in the “wild”

- Practical applications require learning from **noisy labels**
- Multiple annotators of different **skill levels** and **biases**

	David (bird expert)	**Hannah (amateur bird watcher)**
	“Canada Goose”	“Am. Black Duck”
	“Red-necked Grebe”	“Red-necked Grebe”
	“Am. Black Duck”	“Canada Goose”

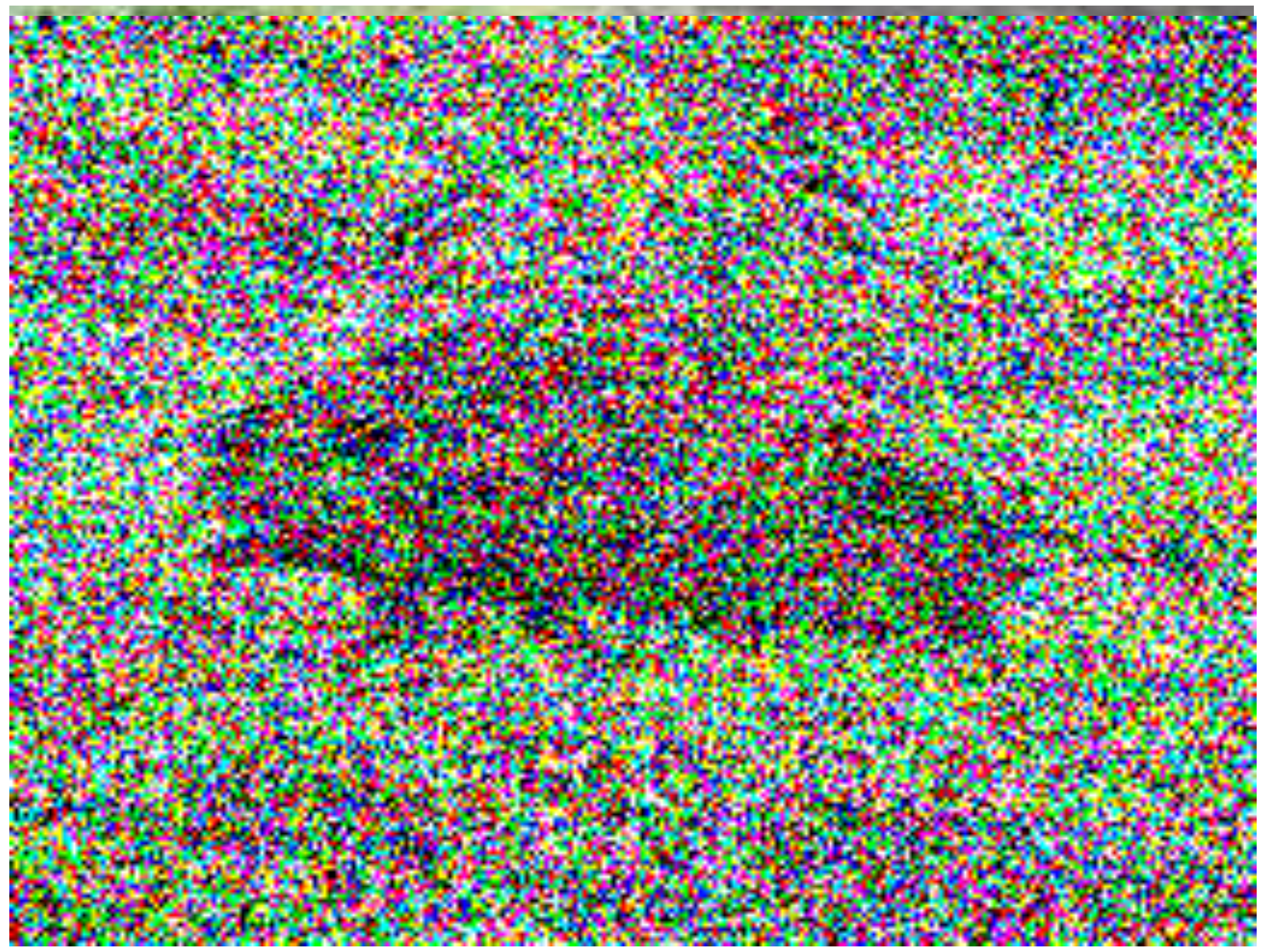
Deep Learning in the "wild"

- Practical applications require learning from **noisy labels**
- Multiple annotators of different **skill levels** and **biases**

	David (bird expert)	**Hannah (amateur bird watcher)**	**Alex (engineer)**
	"Canada Goose"	"Am. Black Duck"	"Bird"
	"Red-necked Grebe"	"Red-necked Grebe"	"Bird"
	"Am. Black Duck"	"Canada Goose"	"Bird"

Deep Learning in the “wild”

- **Input** can also be noisy! e.g. hard to interpret / nebulous images

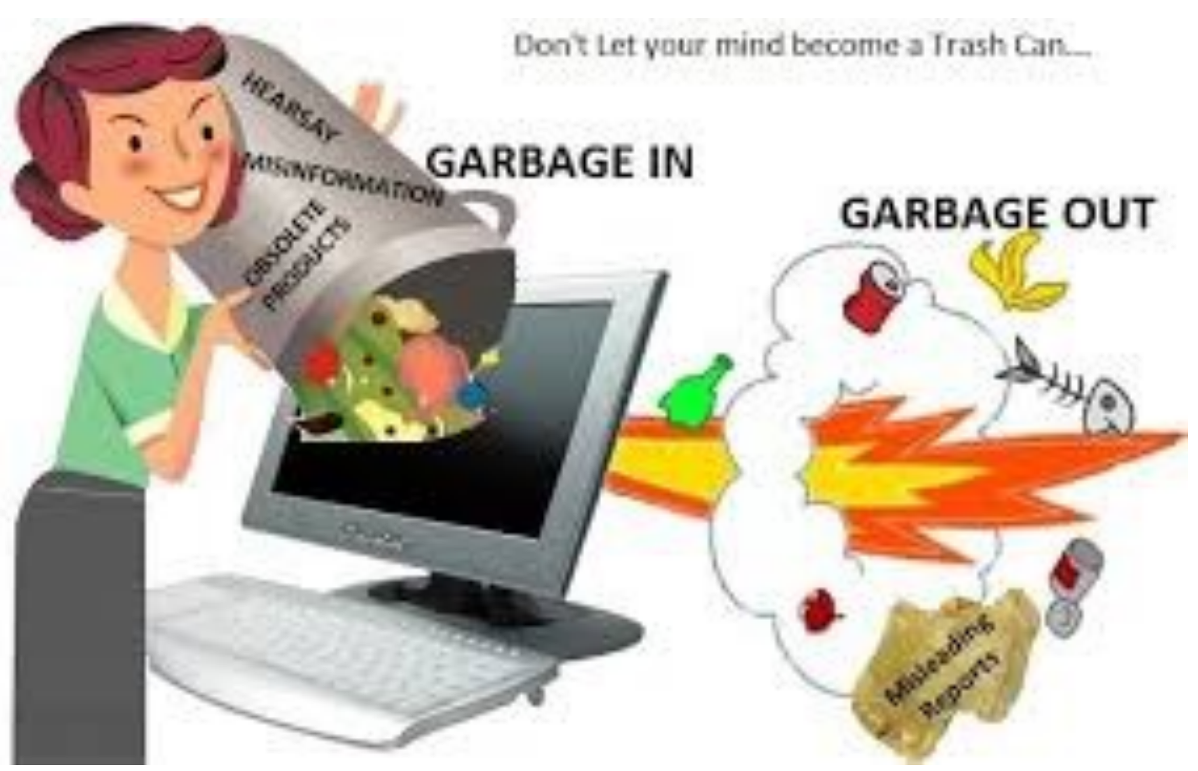


- **But, not the focus of this talk.**

Problems

Problems

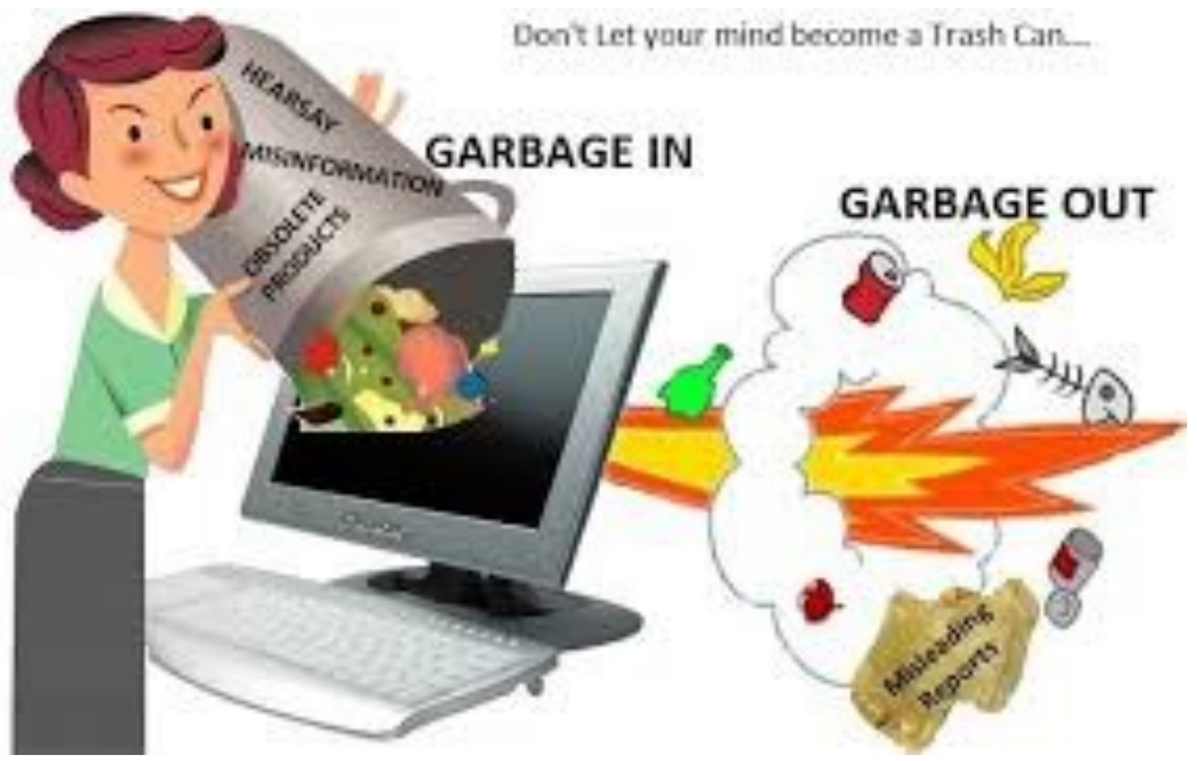
- “Garbage in, garbage out”



Problems

- “Garbage in, garbage out”

- Data curation is time-consuming and suboptimal

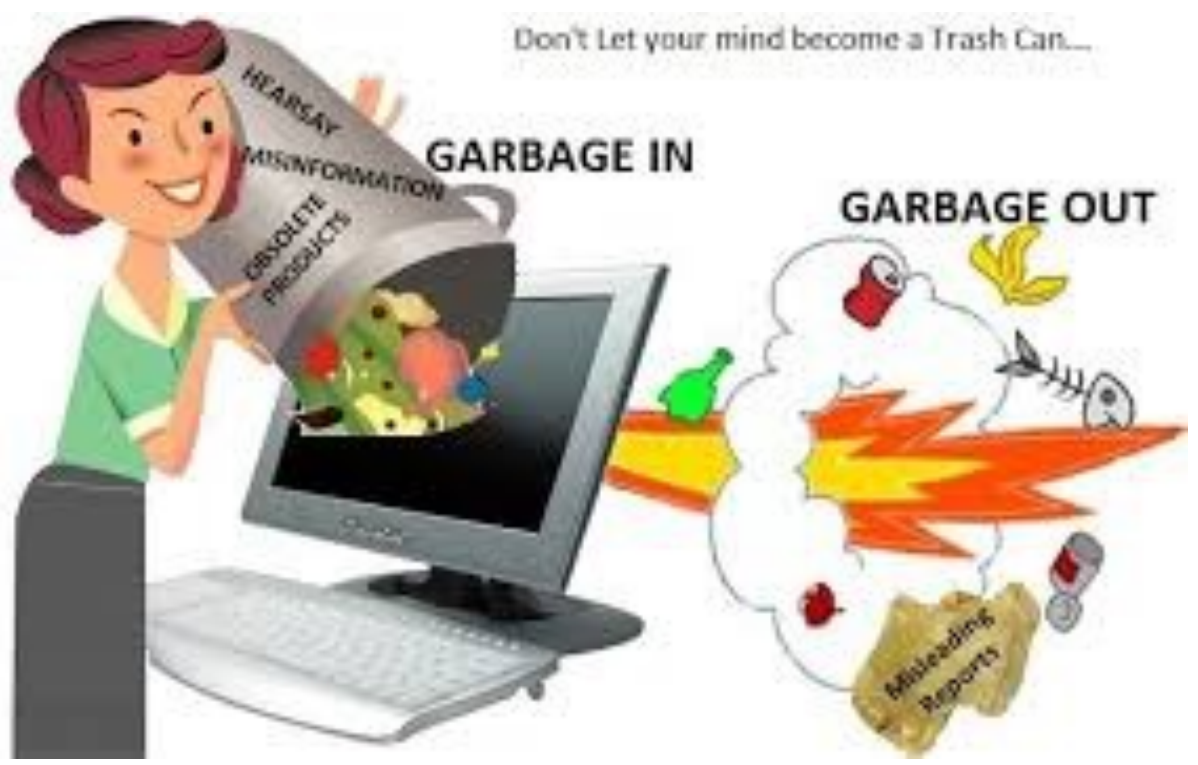


Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press Contributor
I write about technology, entrepreneurs and innovation.

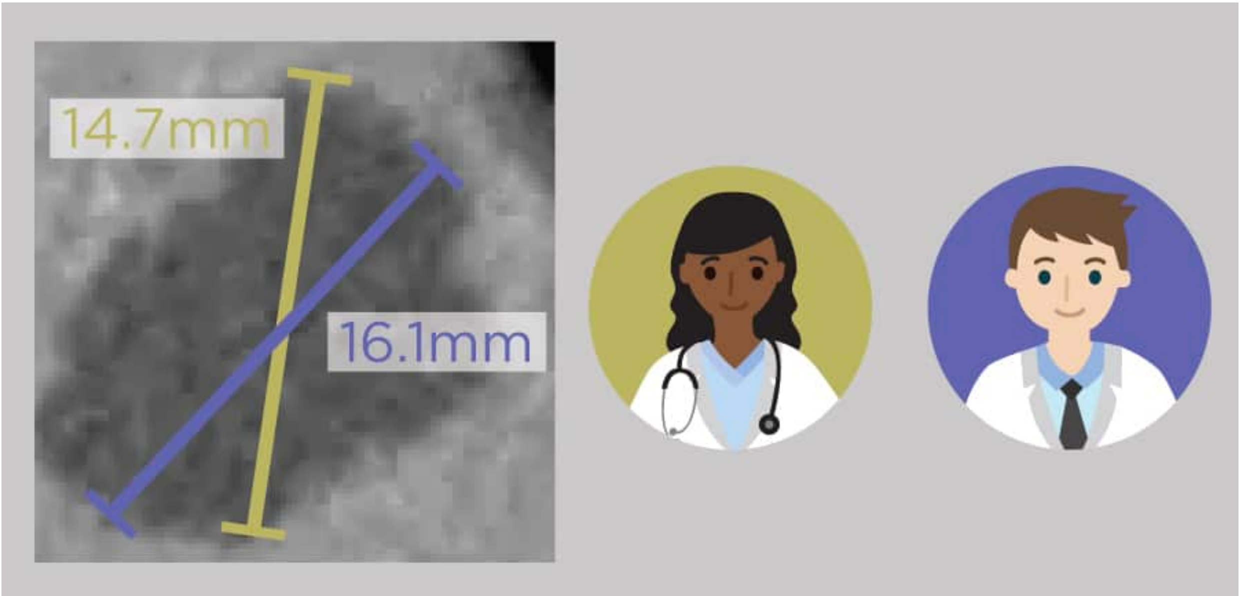
Problems

- “Garbage in, garbage out”



- Data curation is time-consuming and suboptimal

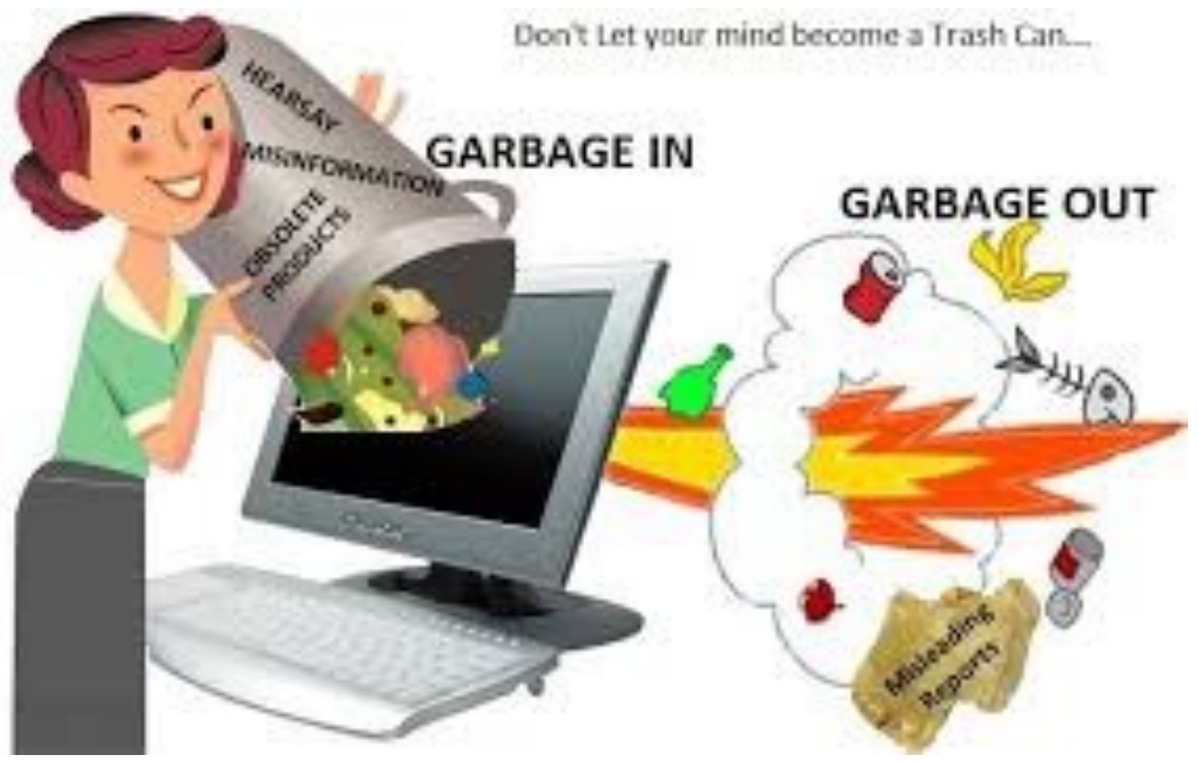
- High inter-reader variability in radiology



(Watadani et al., Radilogy 2013),
(Lazarus et al., Radiology 2006),
(Warfield et al., TMI 2004), many others

Problems

- "Garbage in, garbage out"



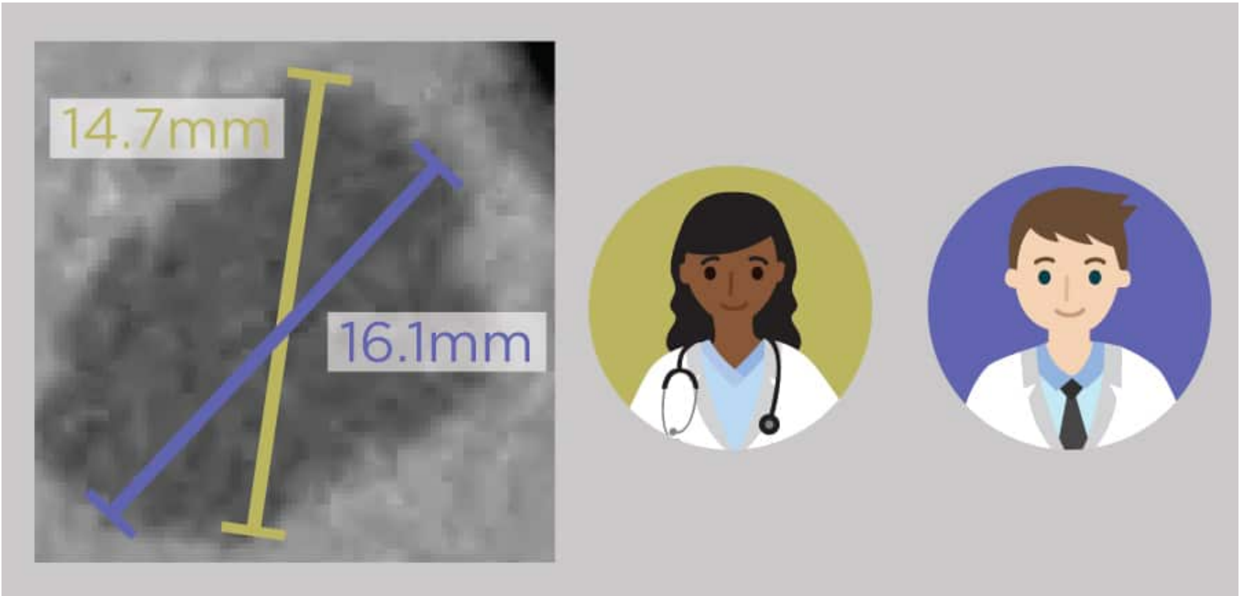
- Data curation is time-consuming and suboptimal


Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press Contributor

I write about technology, entrepreneurs and innovation.

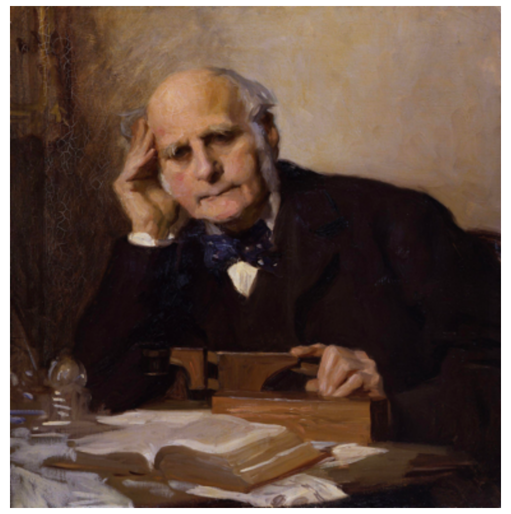
- High inter-reader variability in radiology



(Watadani et al., Radilogy 2013), (Lazarus et al., Radiology 2006), (Warfield et al., TMI 2004), many others

- Majority vote ("Wisdom of Crowds") is **not** always a solution!

  (1) **Expensive,** (2) **Rare experts**




Francis Galton, 1907

My Goal

Simultaneously model **uncertainty of annotators** & **true label distribution**.

My Goal

Simultaneously model **uncertainty of annotators** & **true label distribution**.

=> Automate data curation

My Goal

Simultaneously model **uncertainty of annotators** & **true label distribution**.

=> Automate data curation

=> Improve future label acquisition

David (bird expert) > **Hannah (amateur bird watcher)** > **Alex (engineer)** ?

Set-up

- Multiple annotators
- At least 1 label per image
- No meta-information e.g. expert level, reviews, etc
- No “golden” data
- **Task:** classification

Our Model

Our Model

- *“Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion”*, CVPR 2019

Our Model

- *“Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion”*, CVPR 2019

- An extension of **“Whom to trust when everyone lies a bit”**, [Rayker, ICML 2009]
    - Multi-class & integrates CNN as a component
    - Simpler optimisation, amenable to sparse labels

Our Model

- *“Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion”*, CVPR 2019

- An extension of **“Whom to trust when everyone lies a bit”**, [Rayker, ICML 2009]
    - Multi-class & integrates CNN as a component
    - Simpler optimisation, amenable to sparse labels

- Models the uncertainty of each annotator with a **confusion matrix**.

Our Model

- *“Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion”*, CVPR 2019

- An extension of **“Whom to trust when everyone lies a bit”**, [Rayker, ICML 2009]
  - Multi-class & integrates CNN as a component
  - Simpler optimisation, amenable to sparse labels

- Models the uncertainty of each annotator with a **confusion matrix**.

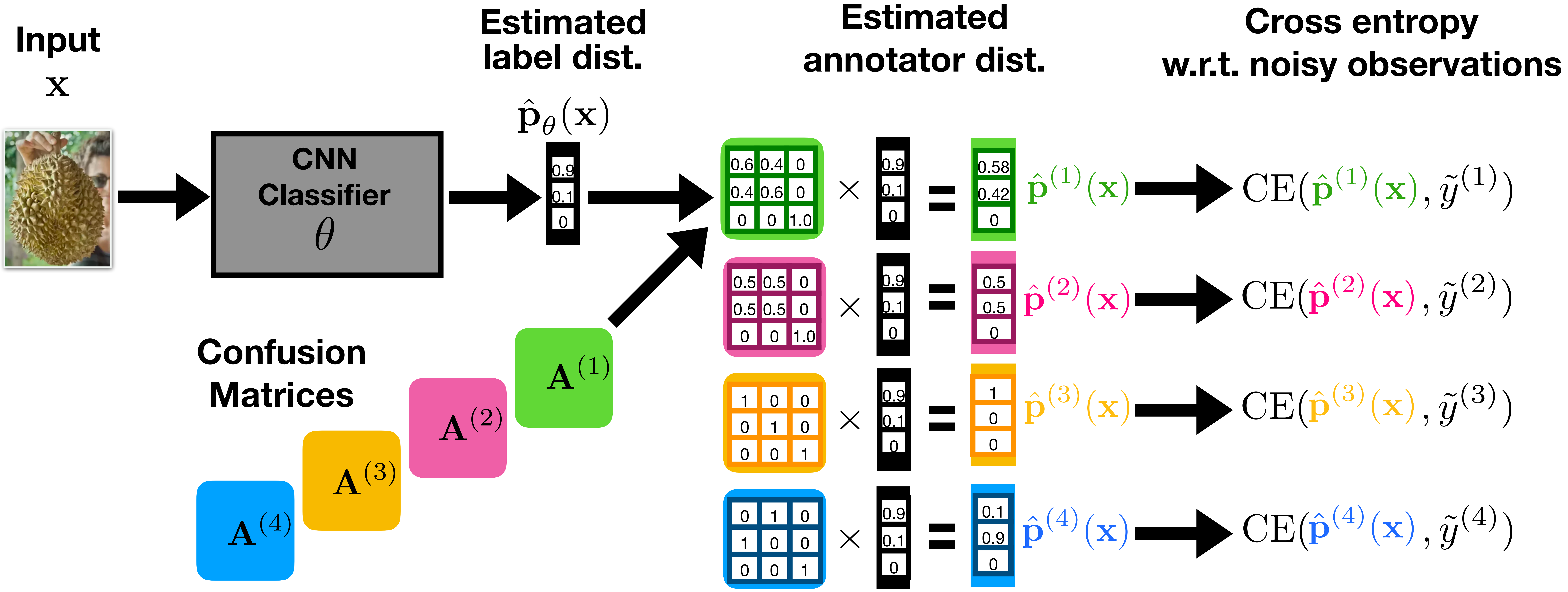
- Use this **confusion matrix** to “correct” noisy labels to learn **true label distribution.**

What is a confusion matrix?

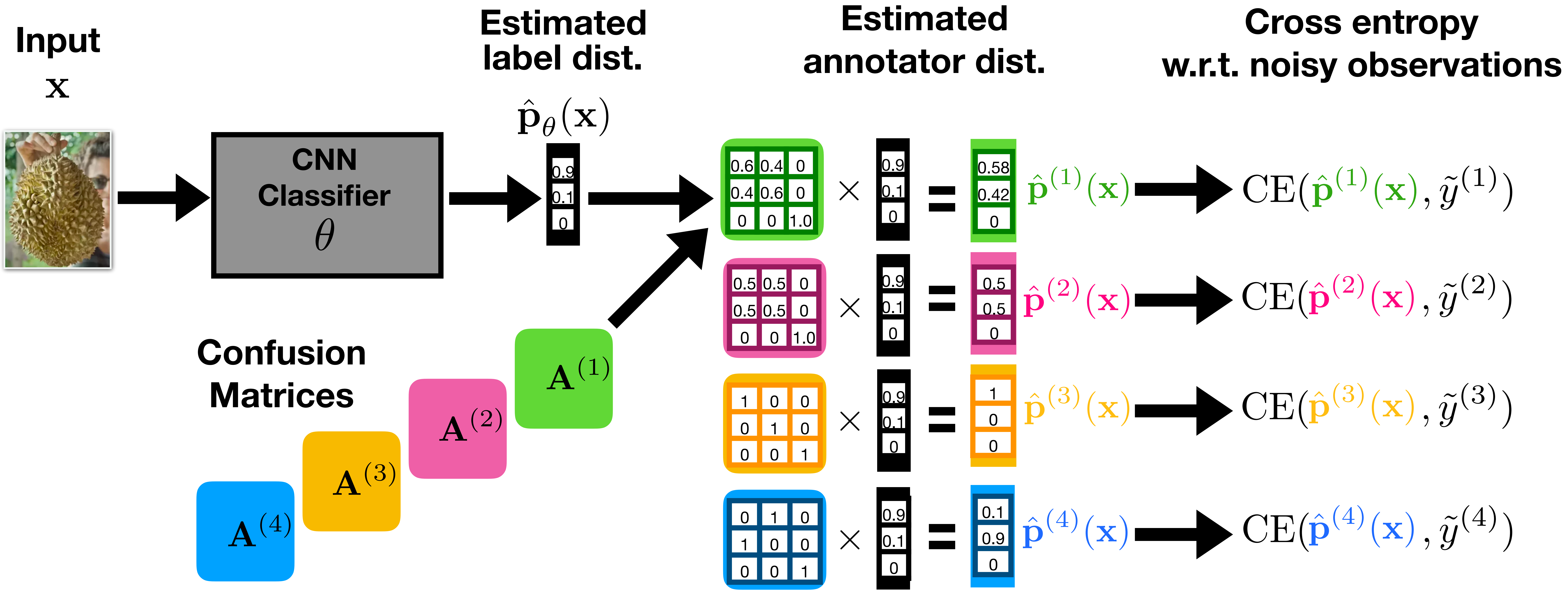
Correct	Durian	Jack fruit	Apple
Durian	0.6	0.4	
Jack fruit	0.5	0.5	
Apple			1.0
	Predictions		

?

Model Schematic



Model Schematic



Loss Function

$$\sum_{i=1}^{N}\sum_{r=1}^{R} \mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i))\cdot \mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda \sum_{r=1}^{R} \mathrm{tr}(\hat{\mathbf{A}}^{(r)})$$

where:

R = no. of annotators

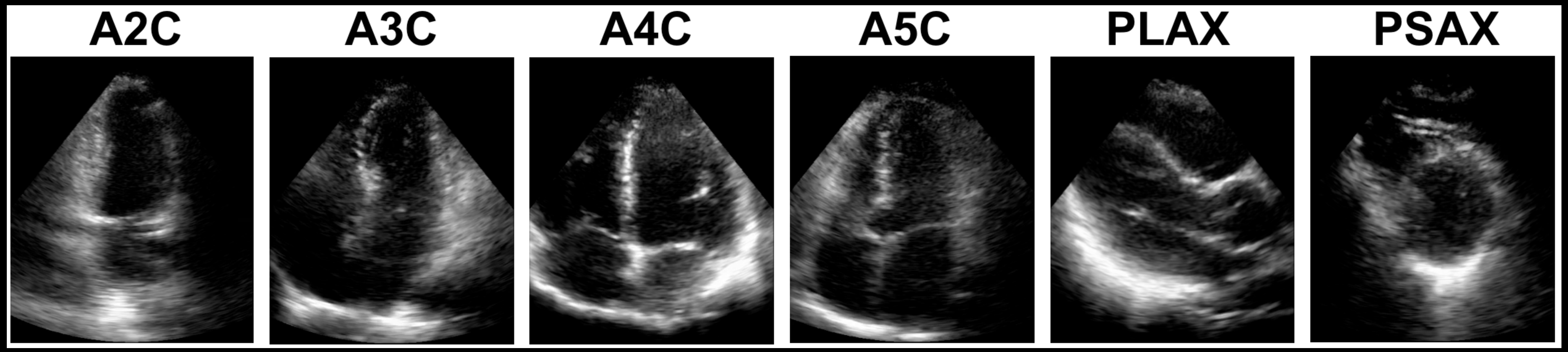
N = no. of samples

S(**x**) = set of available labels for **x**

Experiments

- MNIST digit classification dataset

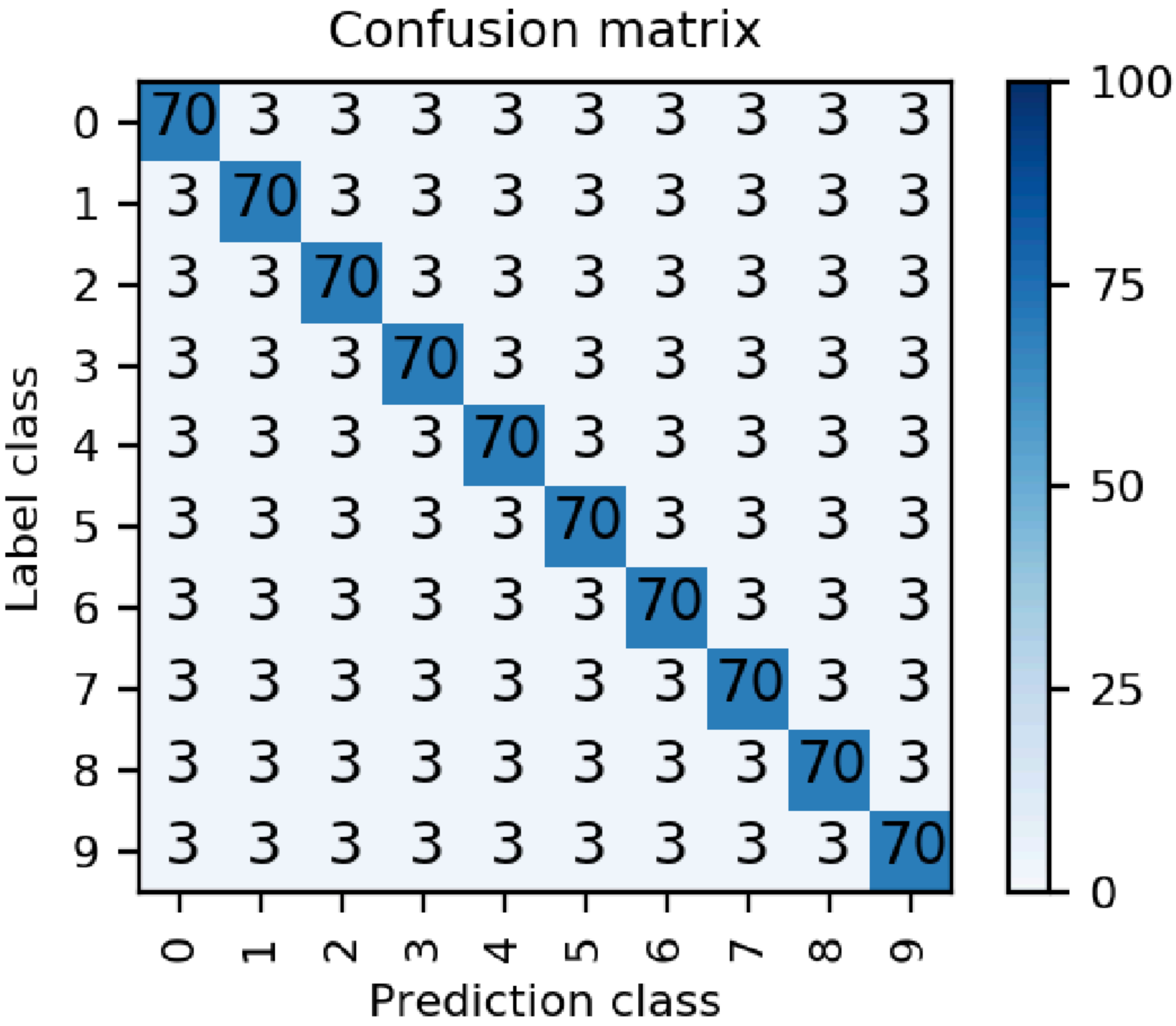
- Ultrasound Cardiac View Classification



Can the model curate and learn simultaneously?

Experiment 1: demo on a diverse annotator group

Experiment 1: demo on a diverse annotator group

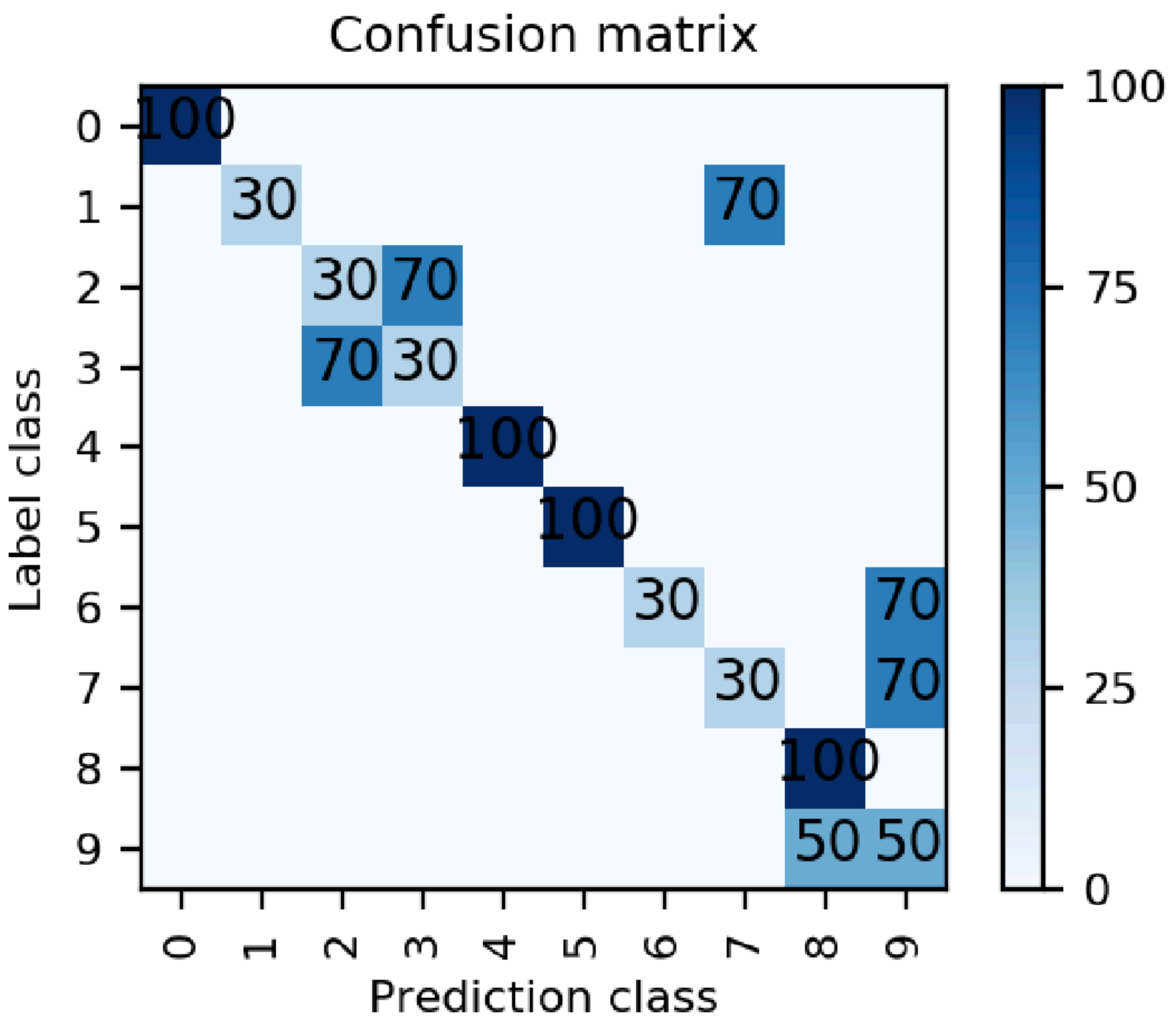


Name: A+ Alice

Accuracy: 70 %

Characteristics:
she whimsically assigns random labels 30% of the time.

Experiment 1: demo on a diverse annotator group



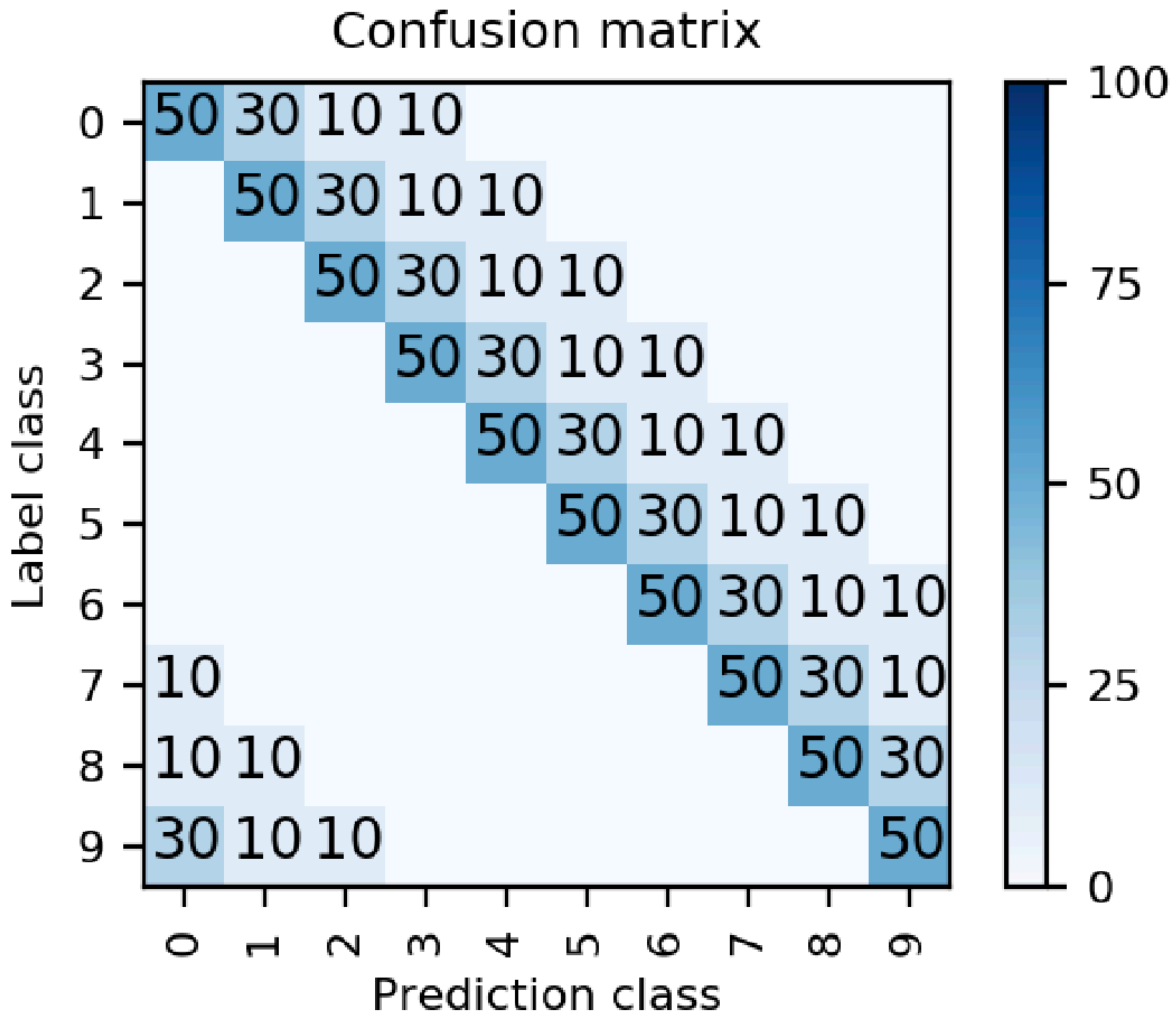
Name: A- Andy

Accuracy: 60 %

Characteristics:
He is not very good at discriminating similar looking numbers.

Flips labels as follows:
1 => 7, 2 <=> 3 ,
6 => 9, 7=>9, 9 => 8

Experiment 1: demo on a diverse annotator group



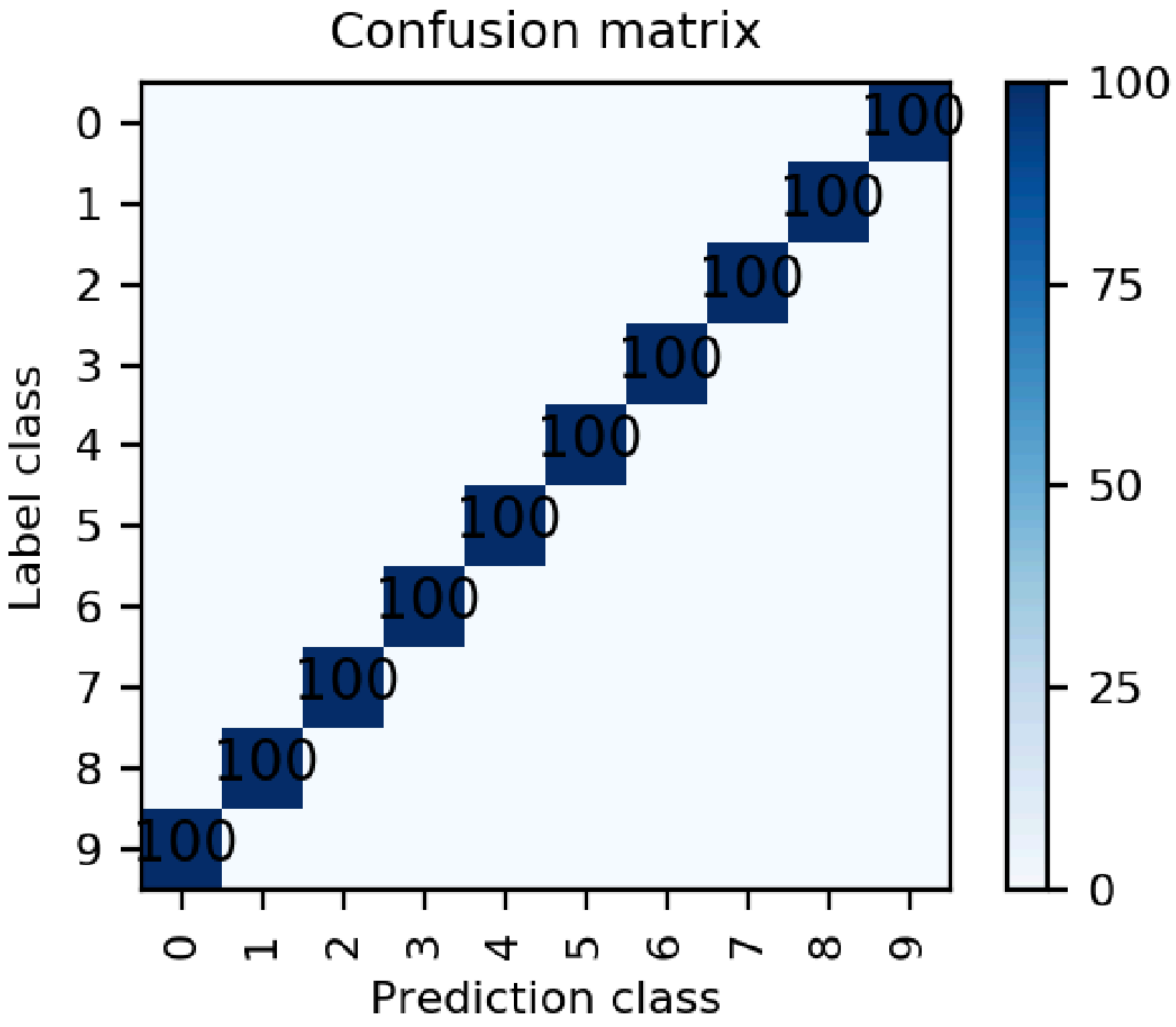
Name: Solid C, Carla

Accuracy: 50 %

Characteristics:
He is not very good at discriminating neighboring digits.

E.g. 1 and 2, 2 and 3, etc

Experiment 1: demo on a diverse annotator group



Name: Failing Frank

Accuracy: 0 %

Characteristics:
In his head,
1 is 9
2 is 8
3 is 7
…
9 is 1.

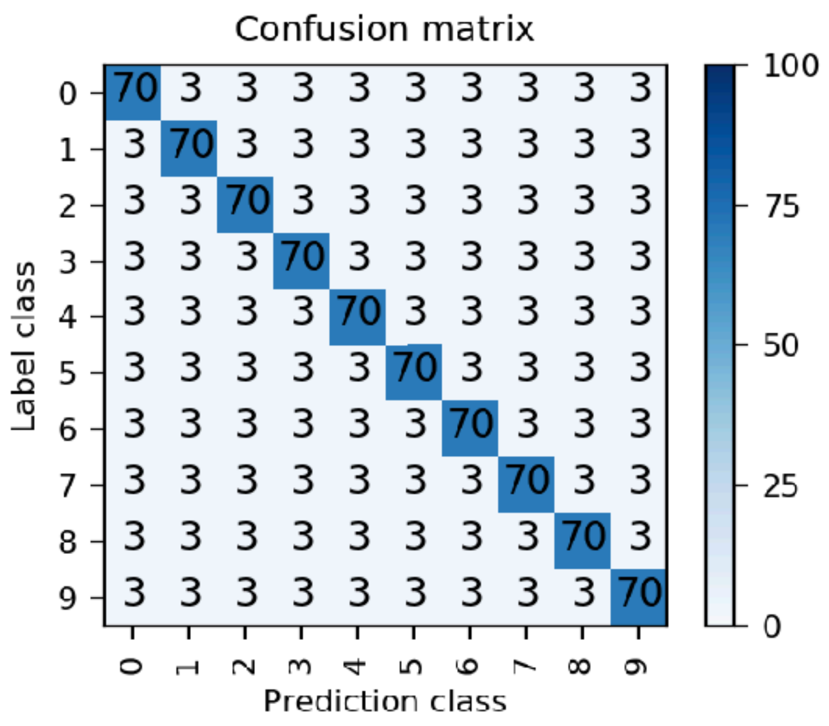
Curation Results

Experiment 1: demo on a diverse annotator group

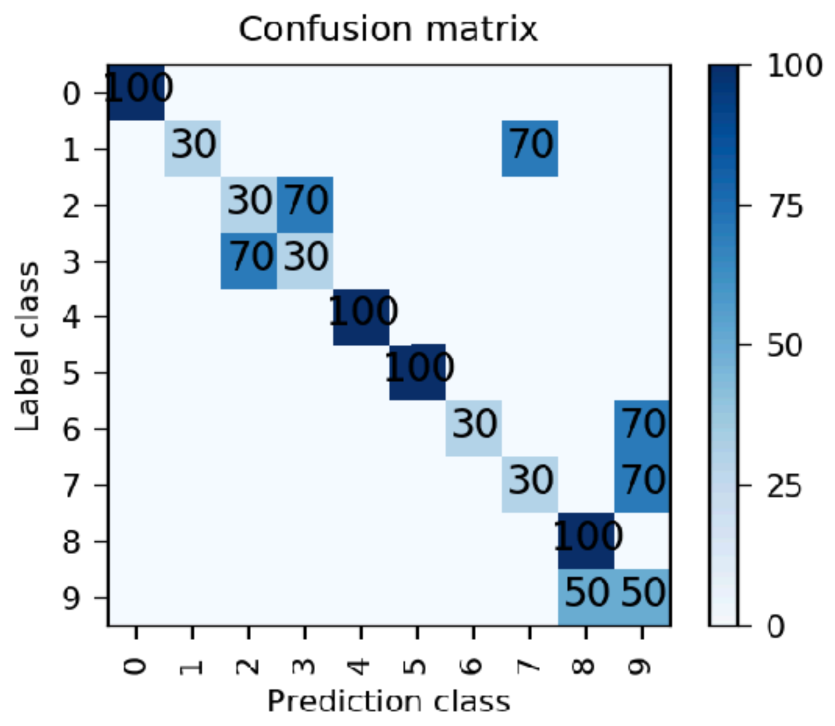
- Now train our model on labels obtained from these people …

Ground Truth

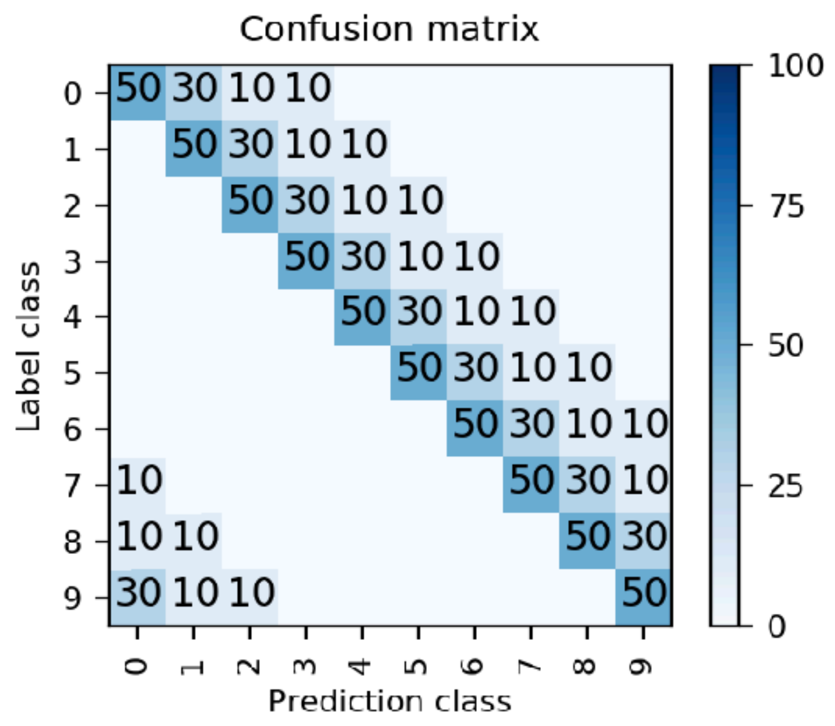
A+ Alice



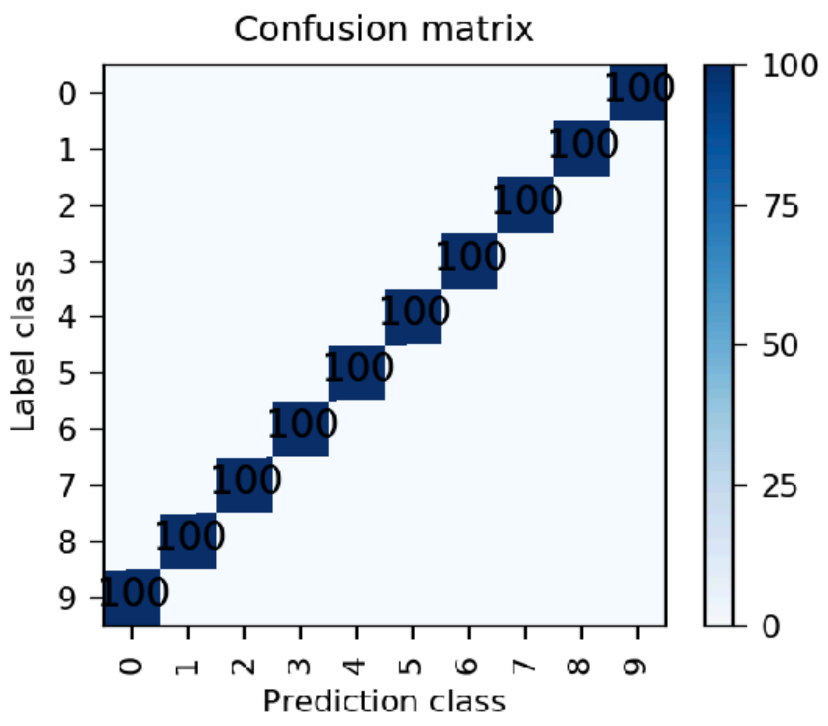
A- Andy



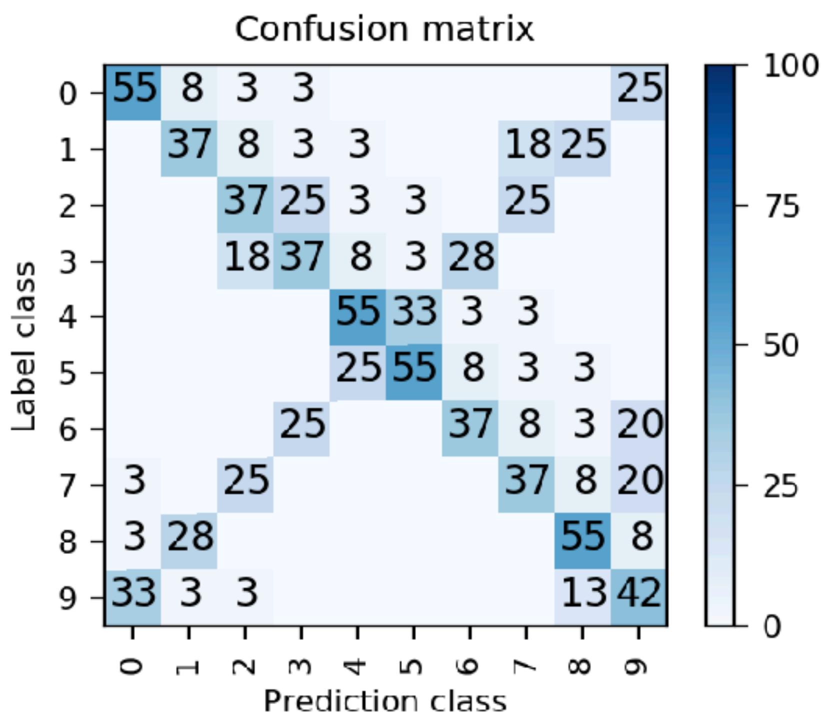
Solid C, Carl



Failing Frank



Average

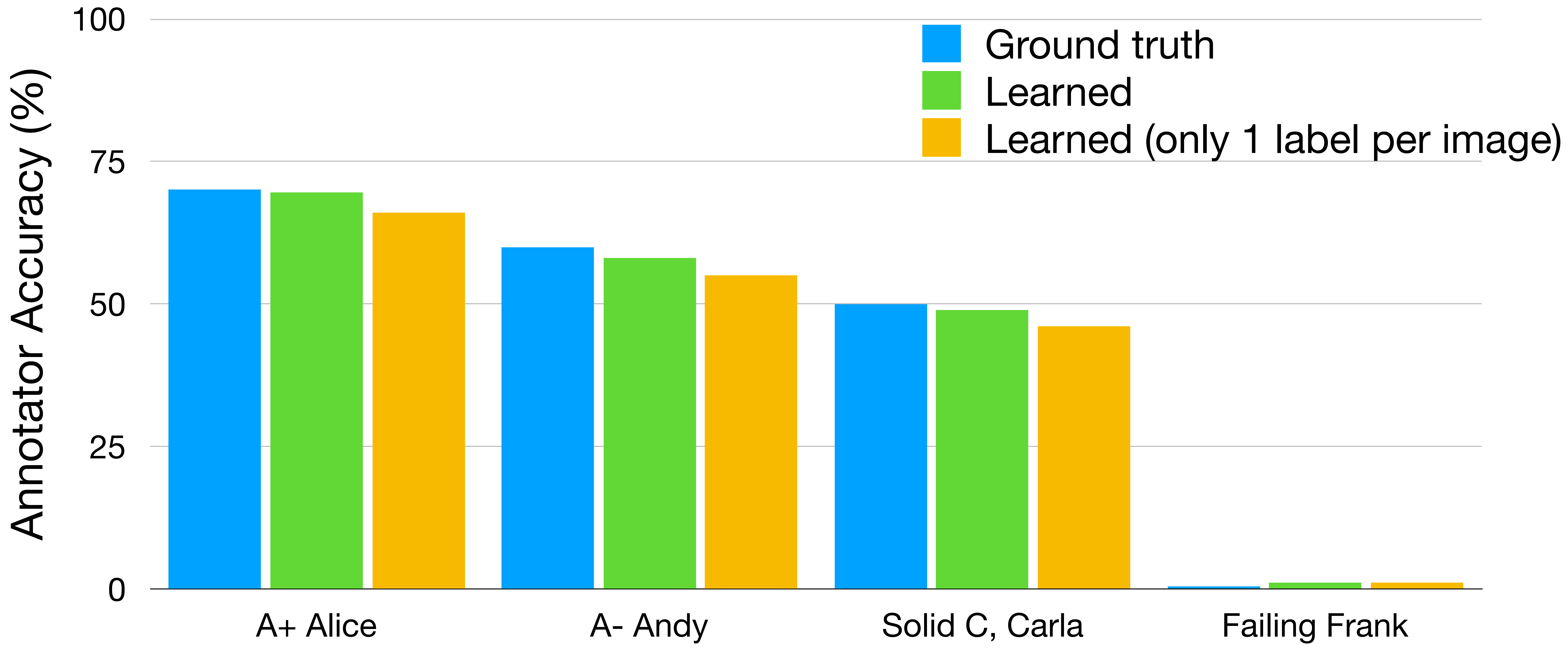


Experiment 1: demo on a diverse annotator group

- Confusion matrices are successfully recovered!

Experiment 1: demo on a diverse annotator group

- Annotator accuracy are well estimated! Useful for ranking.



Model Prediction Results

Model Performance

- > 99 % classification accuracy, outperforms other models.

Oracle (known confusion matrices)
Ours
Naive CNN
Who Said What, Guan et al., AAAI 2018

When does it work (or fail)?

When does it work (or fail)?

Theorem (motivation for the trace term).

If the average confusion matrix of annotators is **diagonally dominant (D.D.)**, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated confusion matrices **uniquely** recover the true confusion matrices.

When does it work (or fail)?

Theorem (motivation for the trace term).

If the average confusion matrix of annotators is **diagonally dominant (D.D.)**, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated confusion matrices **uniquely** recover the true confusion matrices.

Proposed Loss Function

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i))\cdot\mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda\sum_{r=1}^{R}\mathrm{tr}(\hat{\mathbf{A}}^{(r)})$$

where:

R = no. of annotators

N = no. of samples

S(**x**) = set of available labels for **x**

When does it work (or fail)?

Theorem (motivation for the trace term).

If the average confusion matrix of annotators is **diagonally dominant (D.D.)**, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated confusion matrices **uniquely** recover the true confusion matrices.

Proposed Loss Function

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i))\cdot\mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda\sum_{r=1}^{R}\mathrm{tr}(\hat{\mathbf{A}}^{(r)})$$

where:

R = no. of annotators

N = no. of samples

S(**x**) = set of available labels for **x**

What is diagonal dominance?

Every diagonal entry is larger than any other element in the same row.

When does it work (or fail)?

Theorem (motivation for the trace term).

If the average confusion matrix of annotators is **diagonally dominant (D.D.)**, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated confusion matrices **uniquely** recover the true confusion matrices.

Proposed Loss Function

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i))\cdot\mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda\sum_{r=1}^{R}\mathrm{tr}(\hat{\mathbf{A}}^{(r)})$$

where:

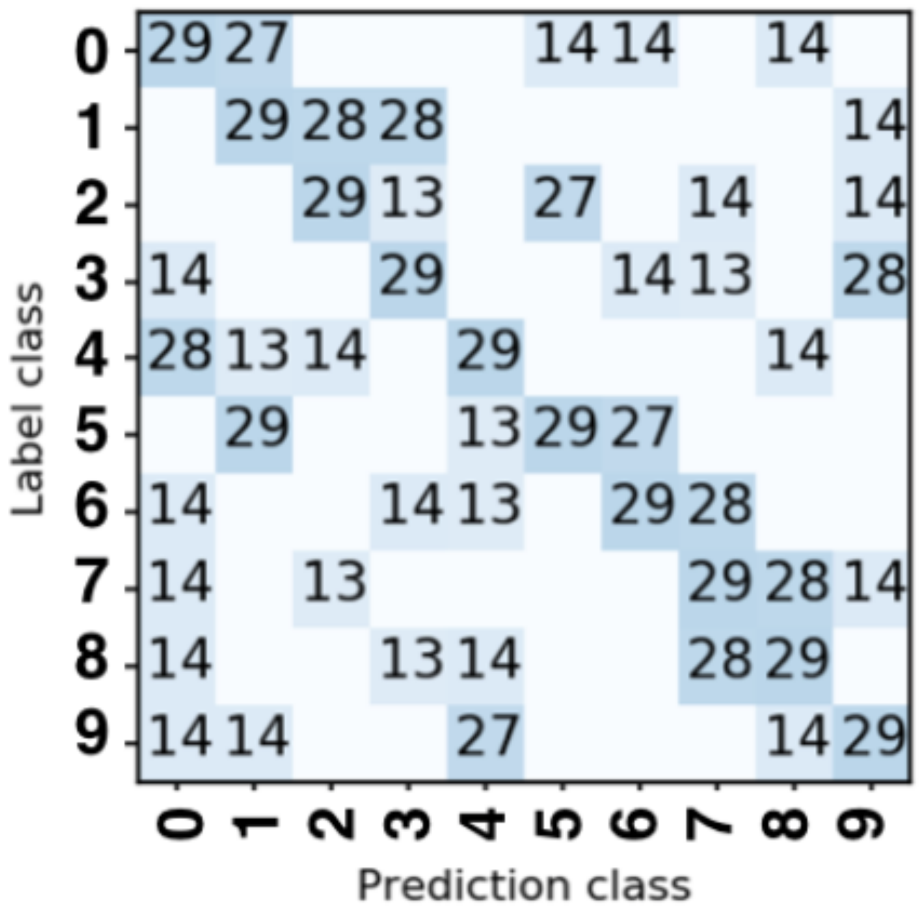
R = no. of annotators

N = no. of samples

S(**x**) = set of available labels for **x**

What is diagonal dominance?

Every diagonal entry is larger than any other element in the same row.



diagonally dominant

When does it work (or fail)?

Theorem (motivation for the trace term).

If the average confusion matrix of annotators is **diagonally dominant (D.D.)**, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated confusion matrices **uniquely** recover the true confusion matrices.

Proposed Loss Function

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbb{1}(\tilde{y}_i^{(r)} \in \mathcal{S}(\mathbf{x}_i))\cdot\mathrm{CE}(\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda\sum_{r=1}^{R}\mathrm{tr}(\hat{\mathbf{A}}^{(r)})$$

where:

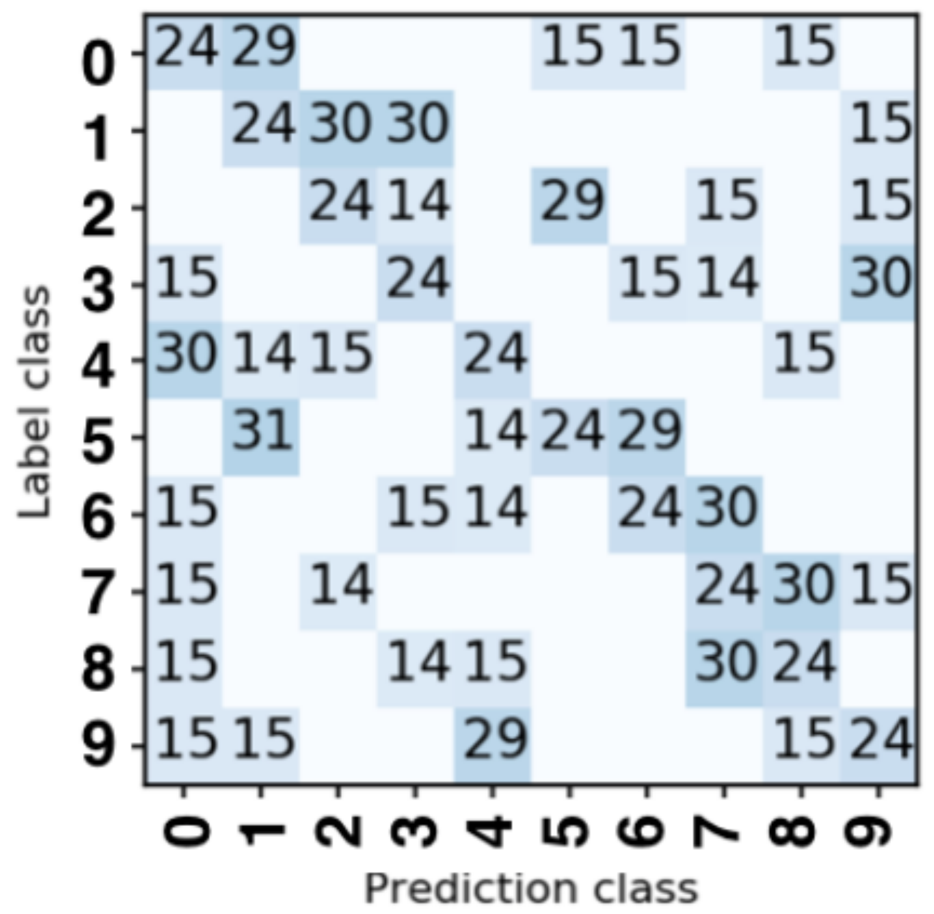
R = no. of annotators

N = no. of samples

S(**x**) = set of available labels for **x**

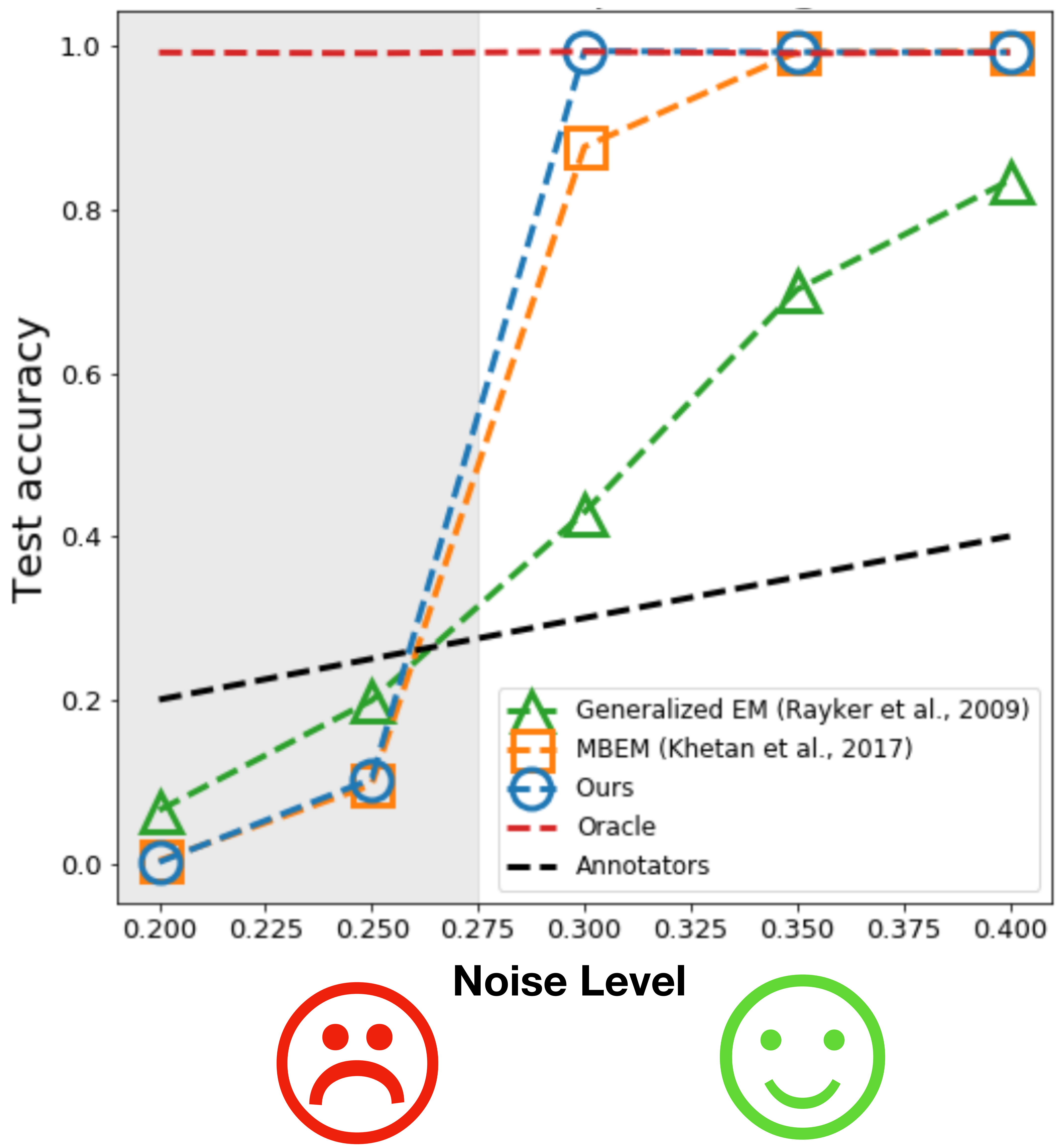
What is diagonal dominance?

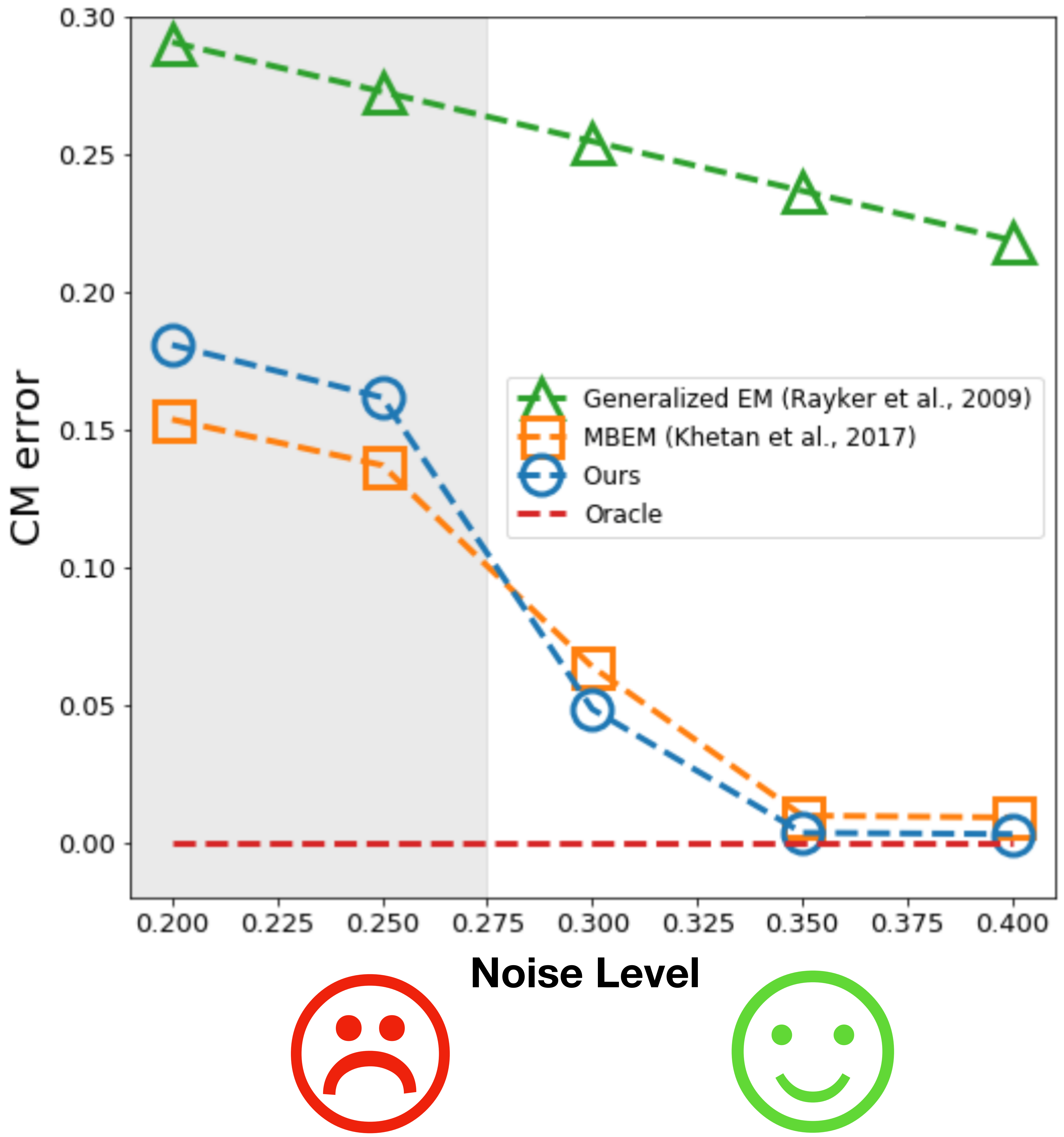
Every diagonal entry is larger than any other element in the same row.



not diagonally dominant

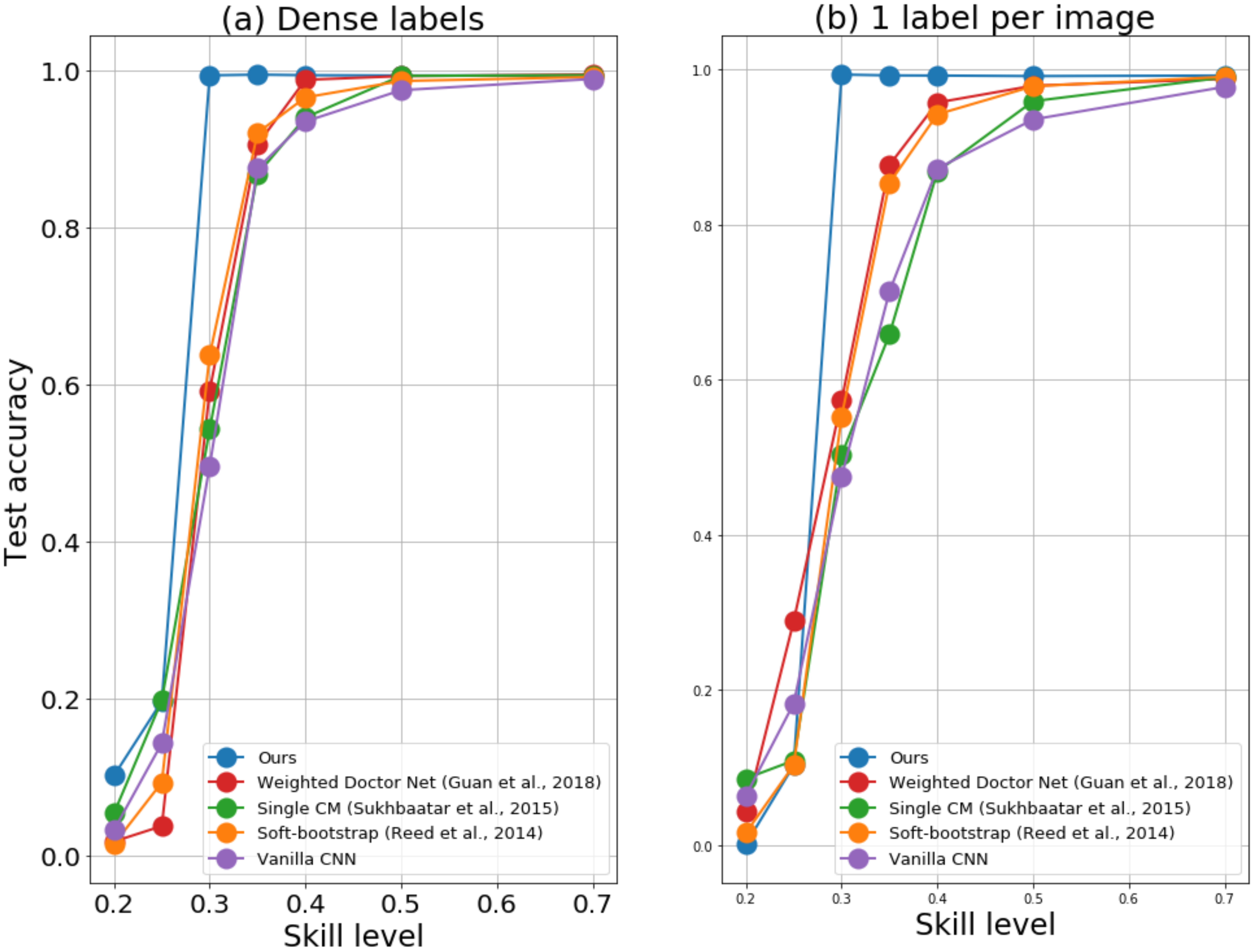
When does it work (or fail)?





Is it important model individual annotators?

Yes!



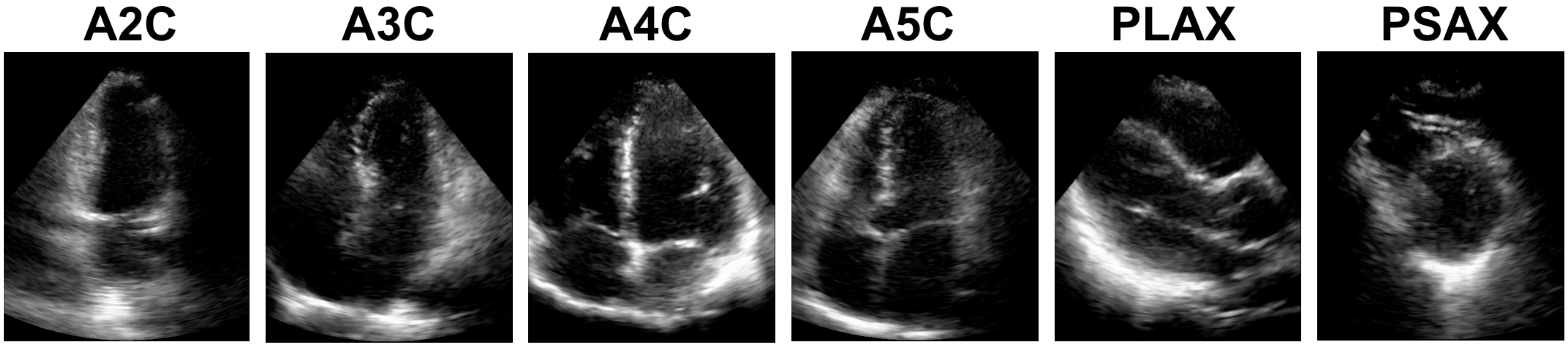
Is it important model individual annotators?

Method	Accuracy
Our method	$\mathbf{81.23 \pm 0.21}$
Single CM [22]	68.82 ± 2.27
Weighted Doctor Net [24]	60.11 ± 1.80
Soft-bootstrap [21]	54.73 ± 1.33
Vanilla CNN [21]	52.33 ± 0.31

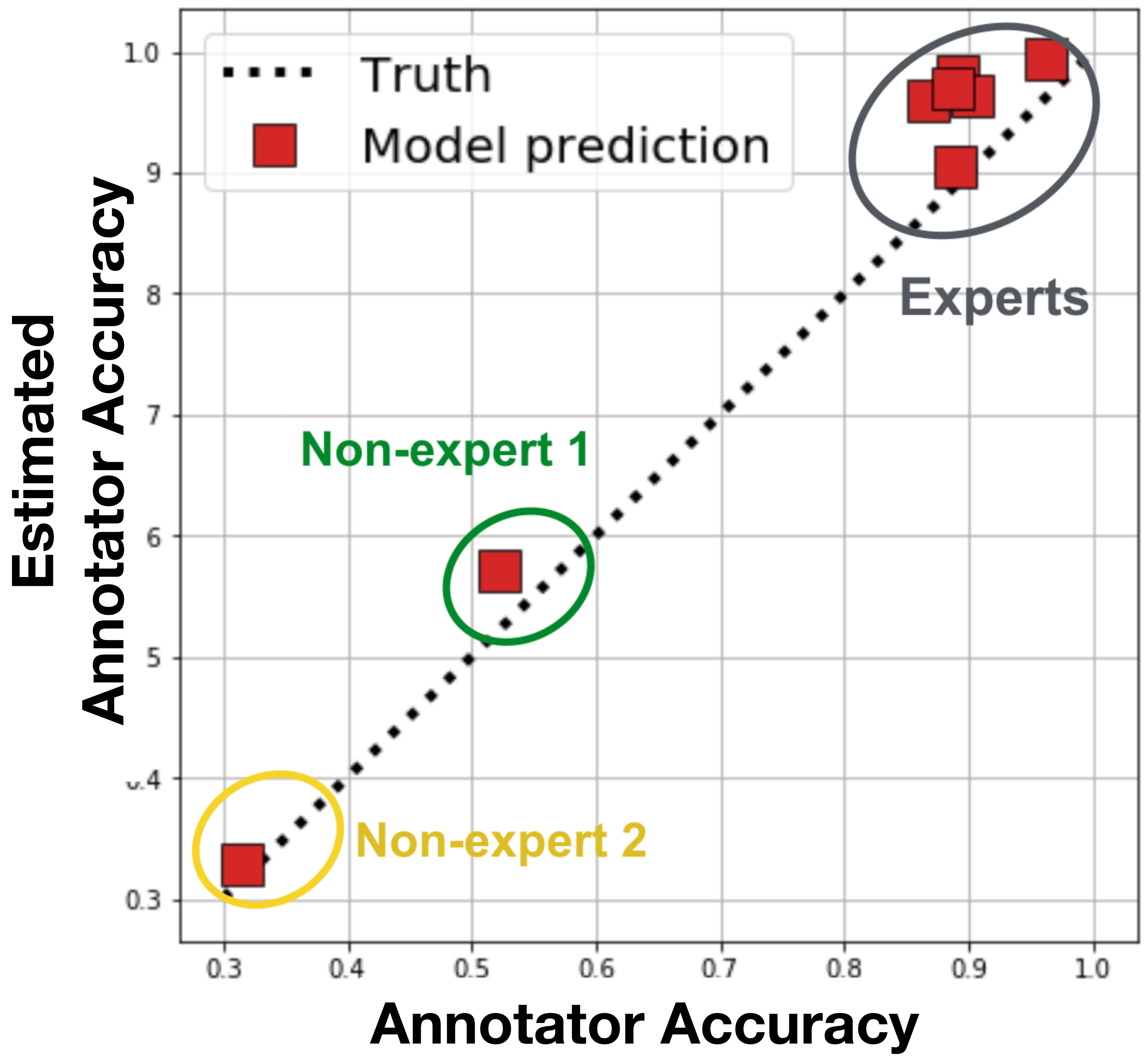
Test on ultrasound data

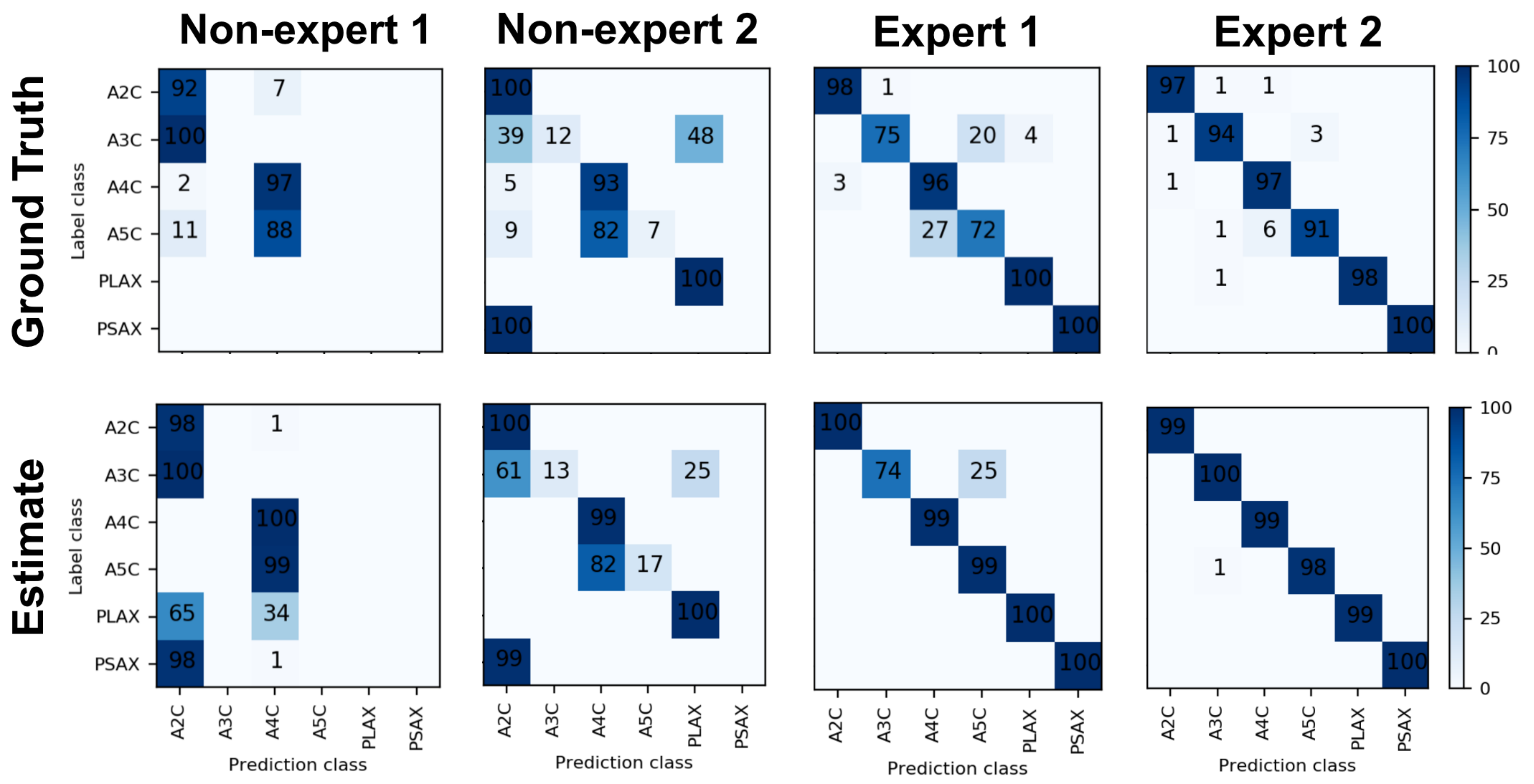
Ultrasound Cardiac View Classification

- 6 classes
- 240,000 training images and 20, 000 test images
- Sparsely labelled by 9 experts + 2 engineers
- Ground truth generated as the unanimous labels from top 3 experts



Ultrasound Cardiac View Classification





	Accuracy (%)
Our method	75.57 ± 0.16
Naive CNN	70.95 ± 0.44

Summary

- One model can simultaneously curate and learn from noisy data, performing better than the state-of-the-art in a very noisy mix of annotators with different skill levels.
- Successful recovery of confusion matrices, can visualise annotator mistakes.
- Robust performance with sparse labels (which is cheaper)

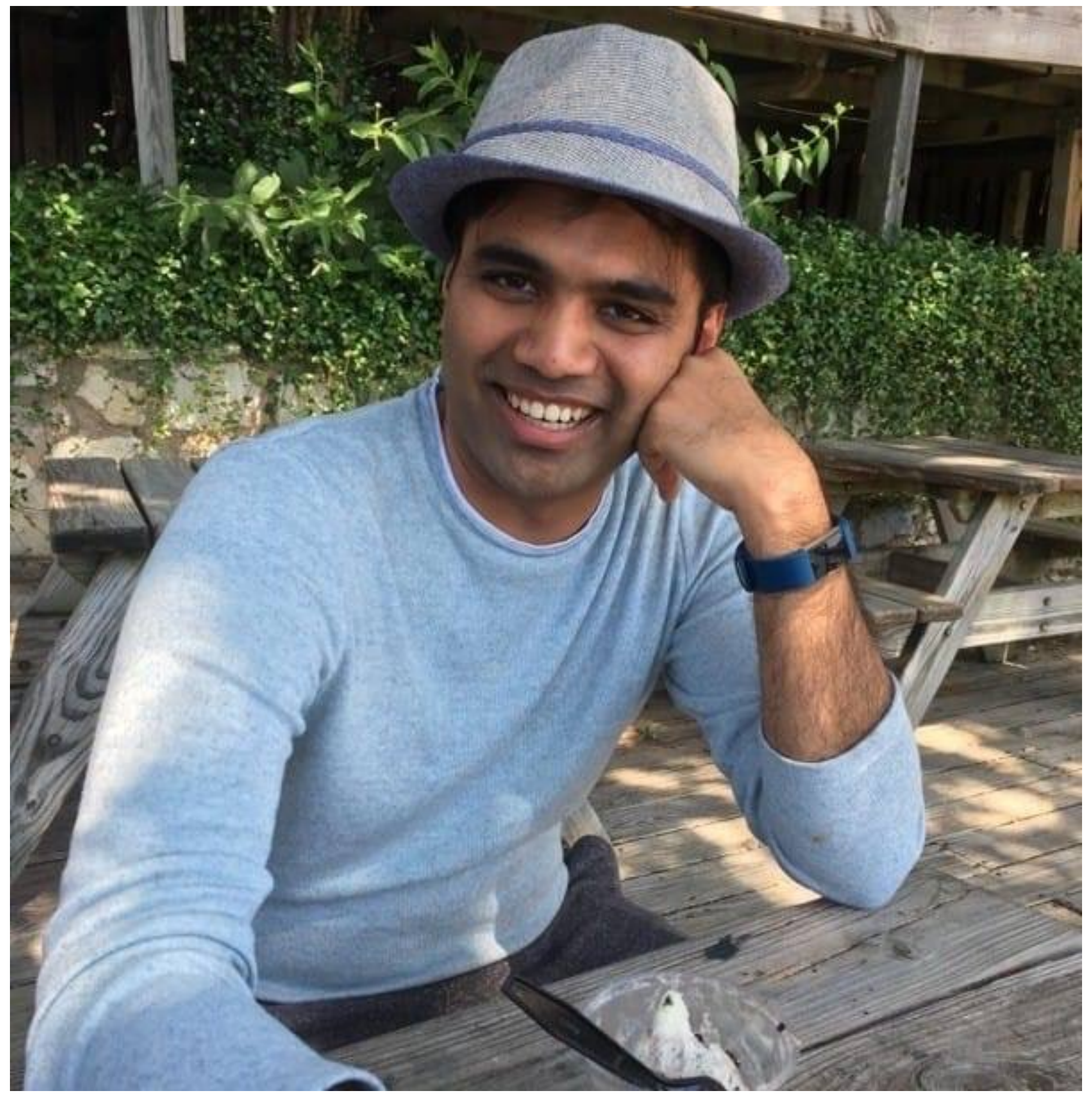
Next Steps

- Account for prior knowledge e.g. expert levels
- Model image dependence of annotators
- Trying to infer different “schools of thoughts”
- Active Learning
- Extend to other tasks e.g. structured prediction?
    - Segmentation errors
    - Geometric errors e.g. misalignment
    - Artefacts in data (e.g. PVEs, motion, etc)

Acknowledgements

Nathan Silberman

Ardavan Saeedi



Swami Sankaranarayanan

Danny C. Alexander

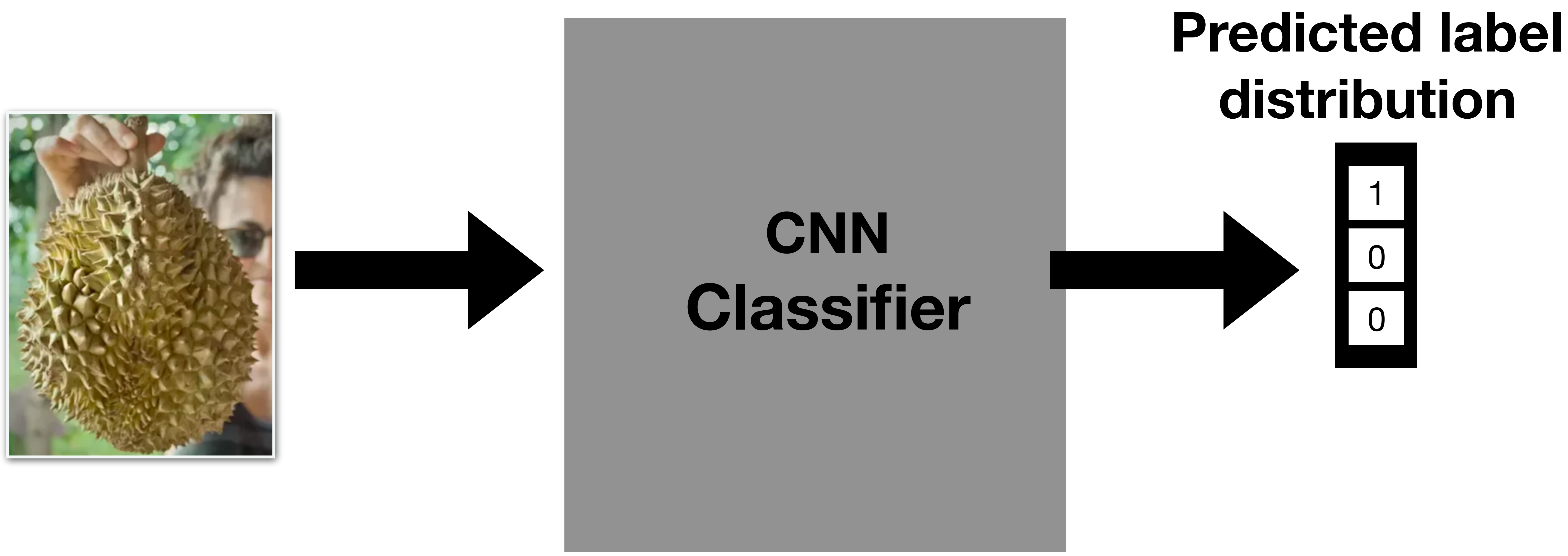
Butterfly™ Network

cmic Centre for Medical Image Computing

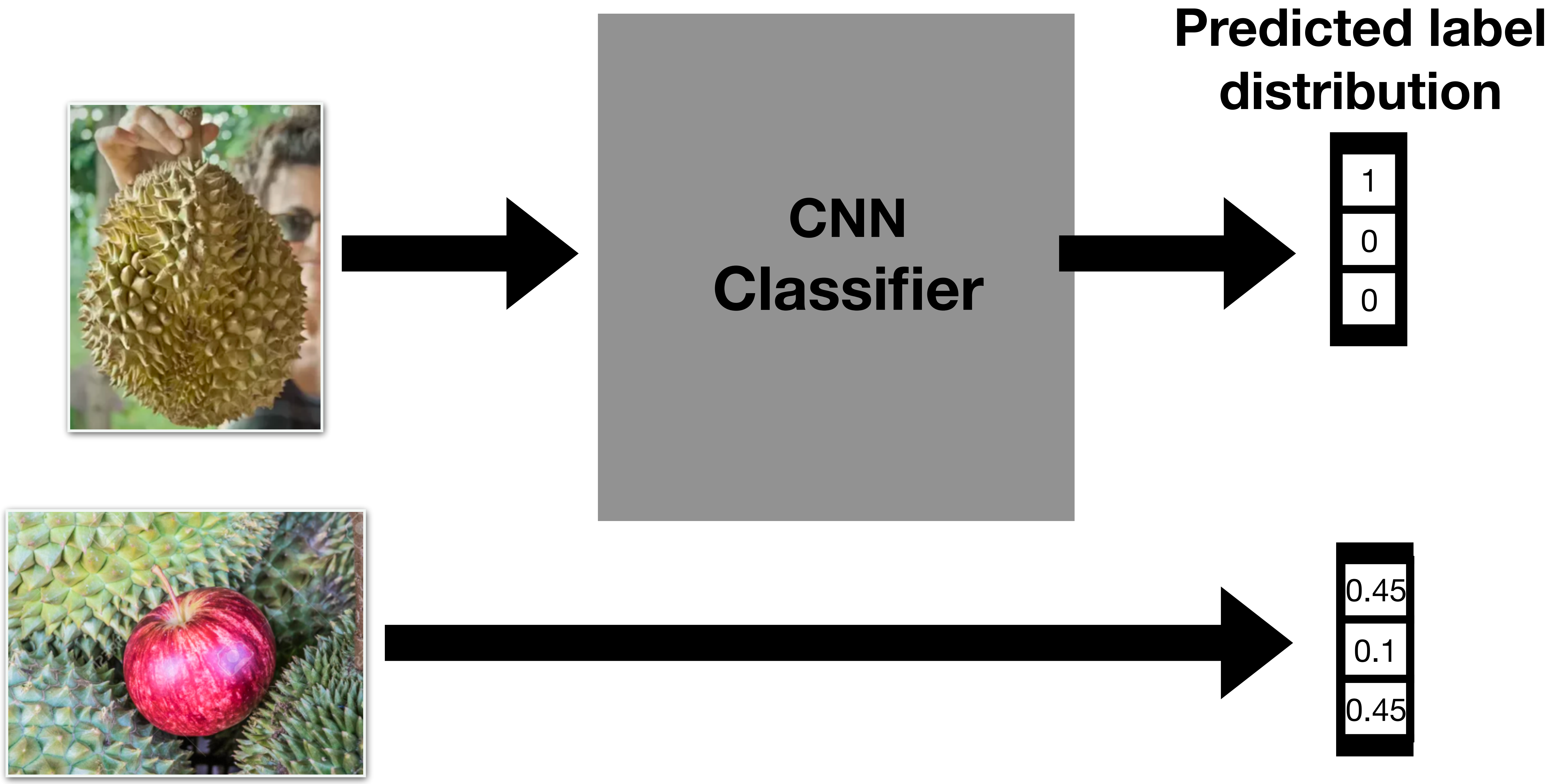

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
gracias
tapadh leat
bedankt
hvala
maururu
dziękuję
mochchakkeram
go raibh maith agat
obrigado
sagolun
sukriya
kop khun krap
grazie
arigatô
takk
dakujem
мерси
terima kasih
감사합니다
ευχαριστώ
merci

Can we gauge the difficulty of images?

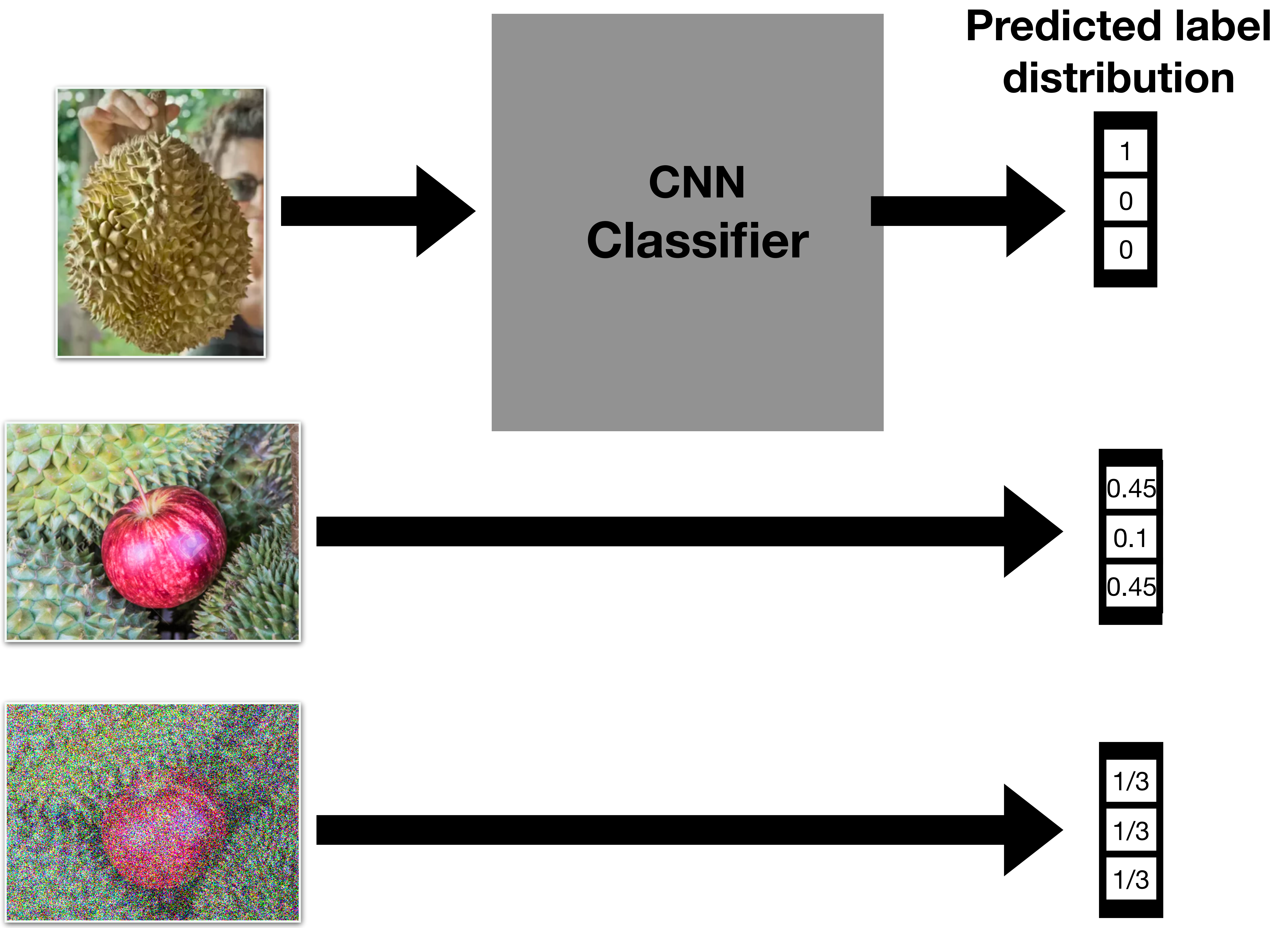
Quantifying image difficulty



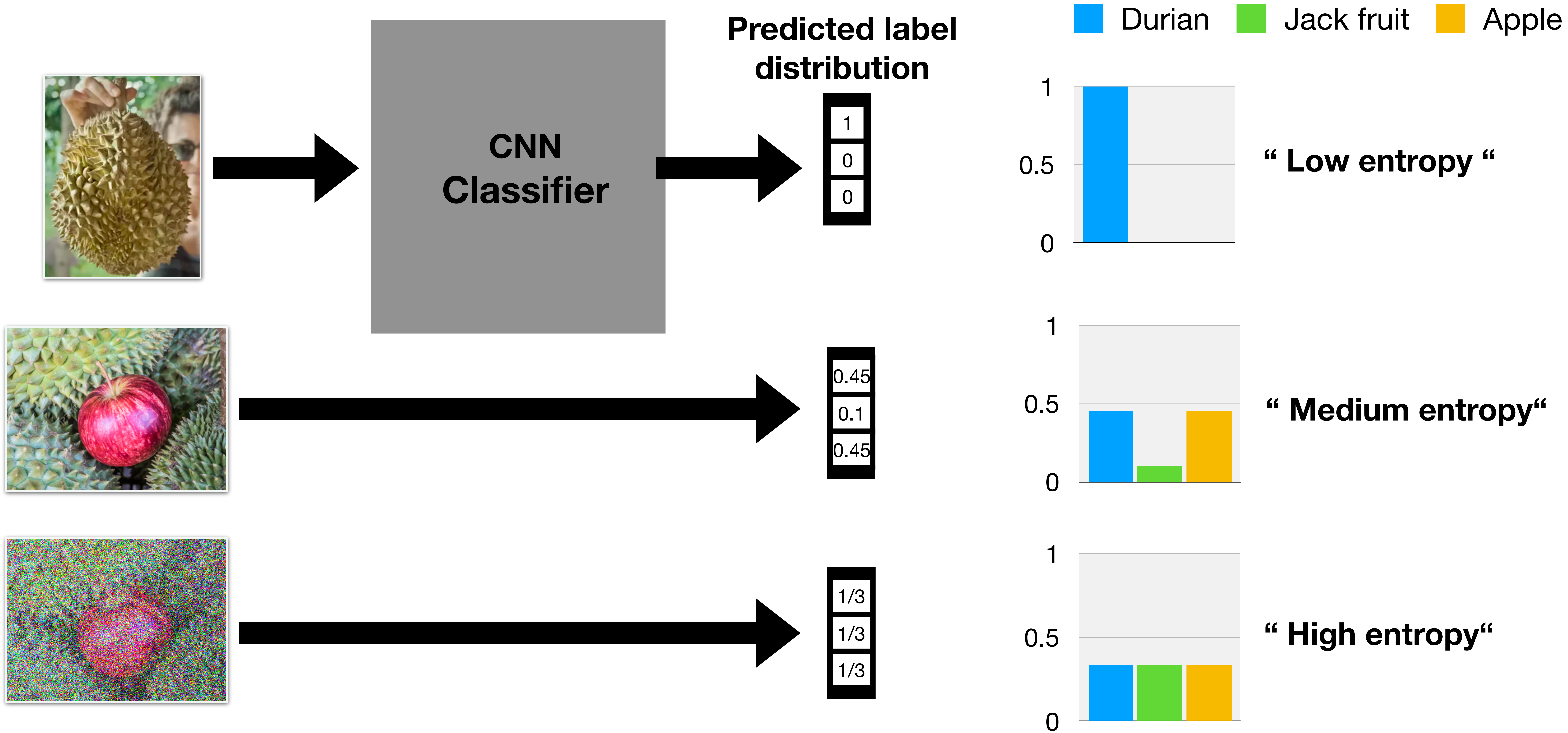
Quantifying image difficulty



Quantifying image difficulty

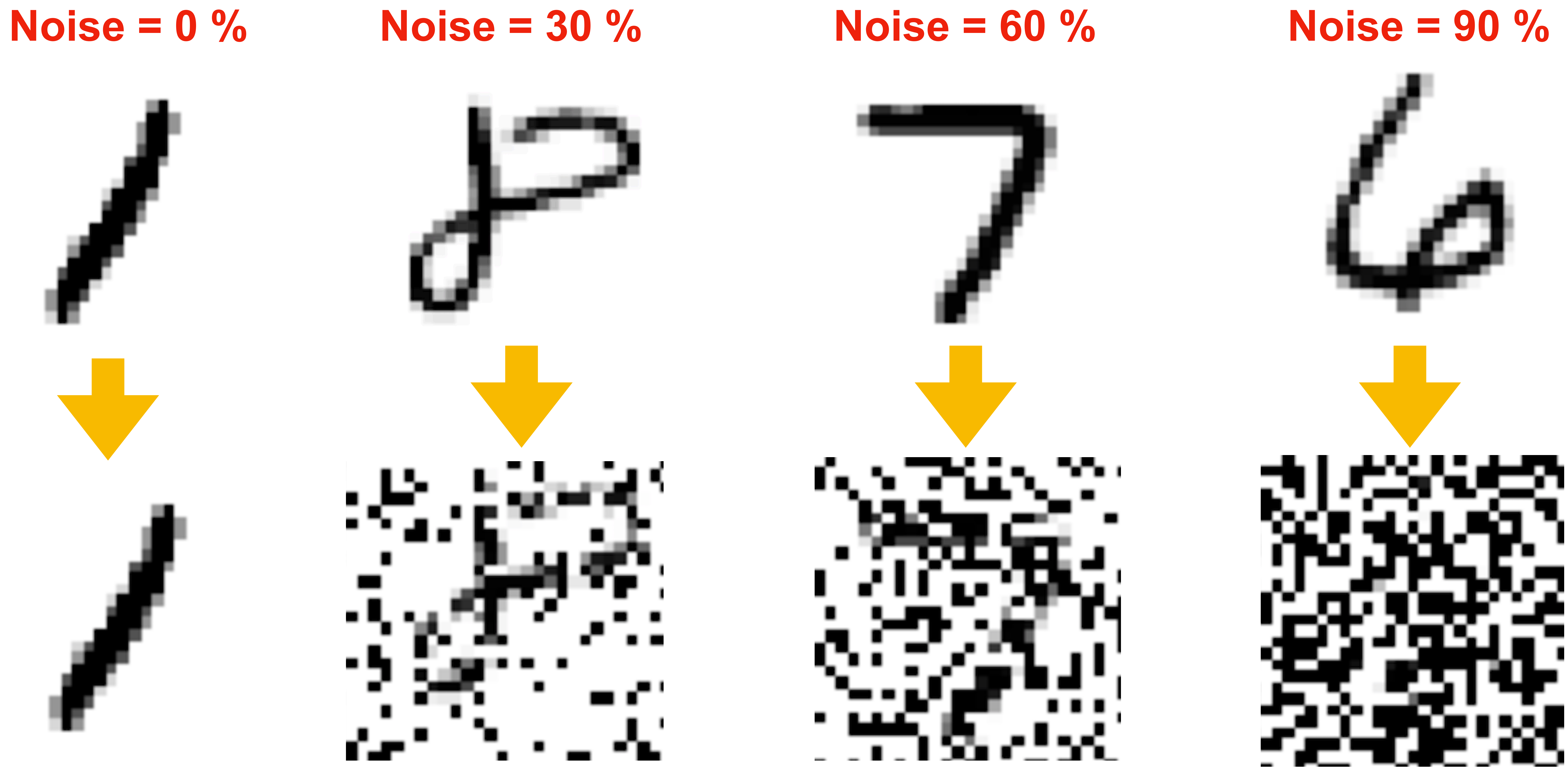


Quantifying image difficulty



Quantifying image noise

- Make the labelling task more difficult by corrupting images



Quantifying image noise

- Labels are obtained from A+ Alice, A- Andy, Solid C Carl, Failing Frank
- Compare the correlation between **image noise level** & **entropy of label distribution**

	Naive softmax	Logit Noise	Loss Attenuation
Ours	0.72	**0.83**	0.77
Sukhbaatar et al., ICLR'15	0.80	0.81	**0.85**
Naive CNN	0.79	**0.87**	0.81