Google DeepMind

Evaluation and calibration of AI models with *uncertain* ground truth

David Stutz

Jul 11 2023

Outline

- ❑ Motivation through annotator disagreement
- ❑ Statistical model
- ❑ Measuring uncertainty
- ❑ Evaluating AI models
- ❑ Case study in dermatology:
  - ❑ Results
  - ❑ Bonus: calibration
- ❑ Conclusion and outlook

Paper: arxiv.org/abs/2307.02191

`Bird", "cat", or "frog"?

"Hemangioma" or "Melanoma"?
Benign or cancer?

Google DeepMind

2023-7-6

Evaluating AI systems under uncertain ground truth: a case study in dermatology

David Stutz[*,@,1], Ali Taylan Cemgil[*,@,1], Abhijit Guha Roy[*,@,2], Tatiana Matejovicova[*,1], Melih Barsbey[*,1,3], Patricia Strachan[2], Mike Schaekermann[2], Jan Freyberg[2], Rajeev Rikhye[2], Beverly Freeman[2], Javier Perez Matos[2], Umesh Telang[2], Dale R. Webster[2], Yuan Liu[2], Greg S. Corrado[2], Yossi Matias[2], Pushmeet Kohli[1], Yun Liu[2,+], Arnaud Doucet[1,+] and Alan Karthikesalingam[2,+]

[1]Google DeepMind, [2]Google, [3]Bogazici University, [*]Equal first authors, [@]Corresponding authors, [+]Equal last authors

Standard evaluation of supervised models

Observation

Correct/good prediction?

AI prediction

"bird"

Standard evaluation of supervised models

Unknown
true label

?

Observation

Correct/good prediction?

AI prediction

“bird”

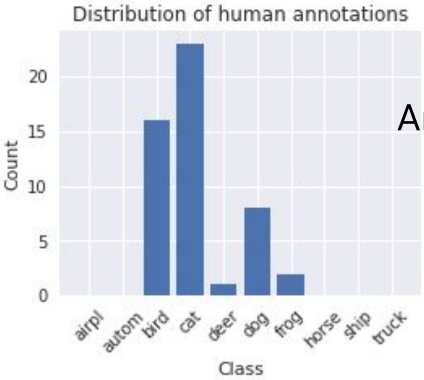
Standard evaluation of supervised models

Unknown
true label

?

Observation

Annotations



?

Annotators disagree!

AI prediction

“bird”

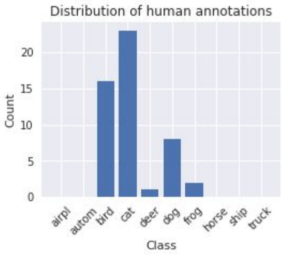
Standard evaluation of supervised models

Unknown true label

?

Observation

Annotations



Majority vote

“cat”

Incorrect!

AI prediction

“bird”

Standard evaluation of supervised models

Unknown true label

?

Observation

Annotations

$\mathbf{b}^1$: {*Pyogenic granuloma* (Low)} {*Hemangioma* (Med)} {*Melanoma* (High)}
$\mathbf{b}^2$ {*Angiokeratoma of skin* (Low)} {*Atypical Nevus* (Med)}
$\mathbf{b}^3$: {*Hemangioma* (Med)} {*Melanocytic Nevus* (Low), *Melanoma* (High), *O/E – ecchymoses present* (Low)}
$\mathbf{b}^4$: {*Hemangioma* (Med), *Melanoma* (High), *Skin Tag* (Low)}
$\mathbf{b}^5$: {*Melanoma* (High)}
$\mathbf{b}^6$: {*Hemangioma* (Med)} {*Melanoma* (High)} {*Melanocytic Nevus* (Low)}

Conditions, Low/Med/High risk conditions

Majority vote non-trivial

?

AI prediction

“Hemangioma”

Standard evaluation of supervised models

Unknown true label

?

Observation

Annotations

b¹: {*Pyogenic granuloma* (Low)} {*Hemangioma* (Med)} {*Melanoma* (High)}
b² {*Angiokeratoma of skin* (Low)} {*Atypical Nevus* (Med)}
b³: {*Hemangioma* (Med)} {*Melanocytic Nevus* (Low), *Melanoma* (High), *O/E – ecchymoses present* (Low)}
b⁴: {*Hemangioma* (Med), *Melanoma* (High), *Skin Tag* (Low)}
b⁵: {*Melanoma* (High)}
b⁶: {*Hemangioma* (Med)} {*Melanoma* (High)} {*Melanocytic Nevus* (Low)}

?

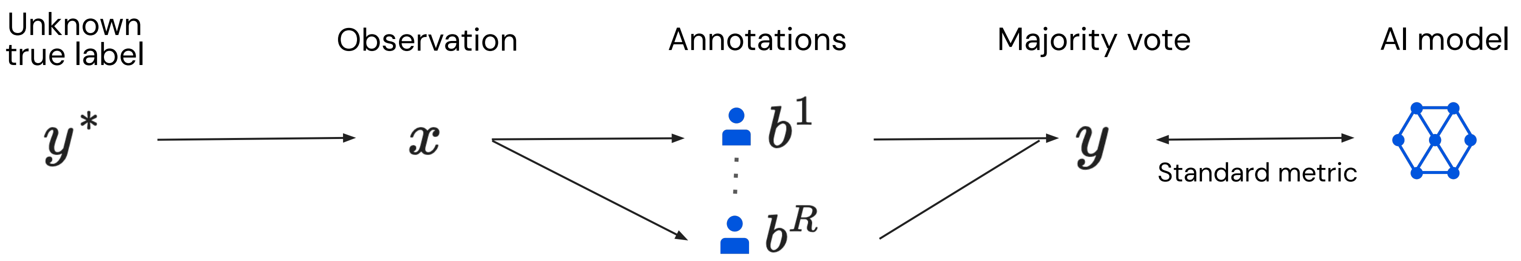
AI prediction *set*

"Hemangioma"

"Atypical Nevus"

"Melanocytic Nevus"

Standard evaluation of supervised models

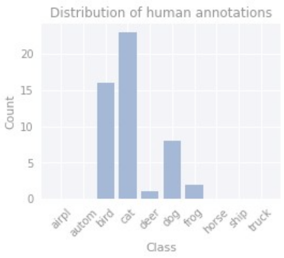


Unknown true label

?

Observation

Annotations



Majority vote

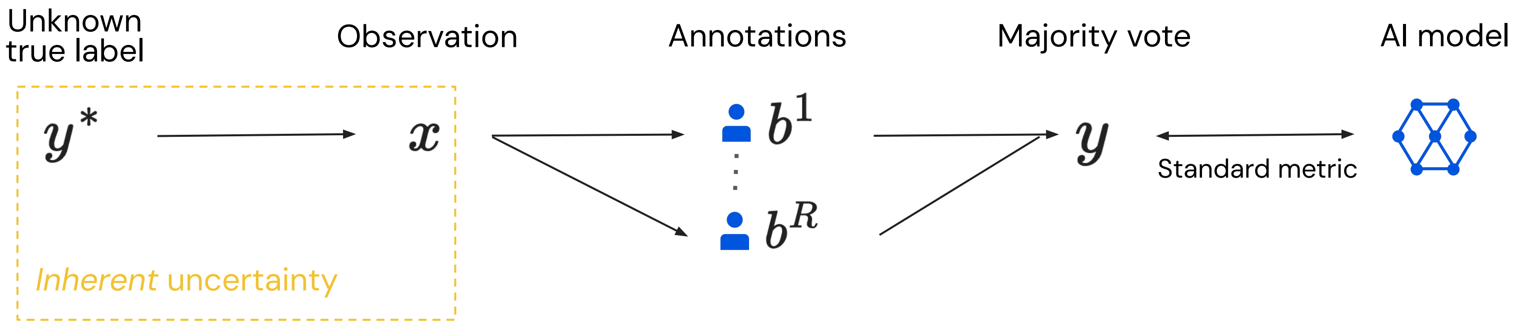
"cat"

Incorrect!

AI prediction

"bird"

Inherent uncertainty

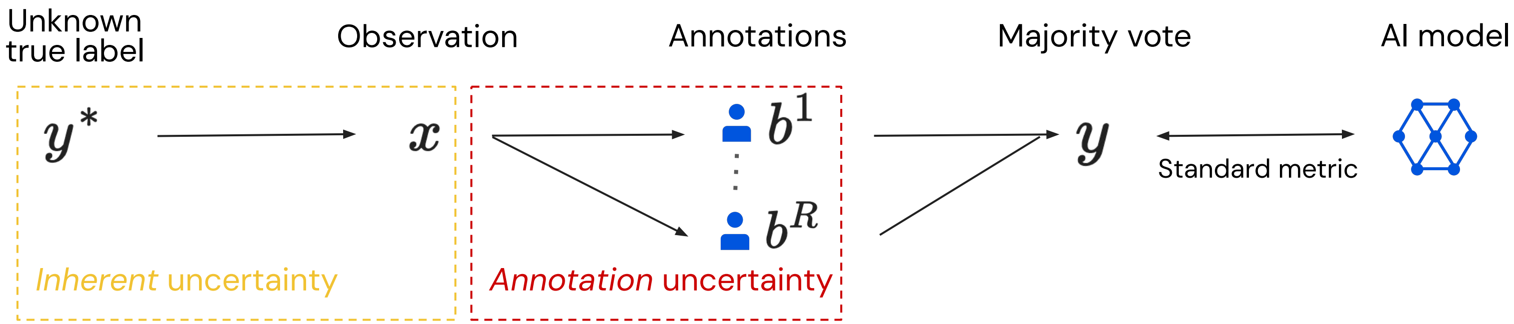


Inherent uncertainty = limited observational information:
(typically called data uncertainty)

- Low-resolution images in image recognition (e.g., CIFAR10)
- Single 2D view in 3D reconstruction
- Missing meta information or no option to question the patient in health
- ...

TL;DR: $p(y^*|x)$ is not one-hot and has high entropy!

Annotation uncertainty

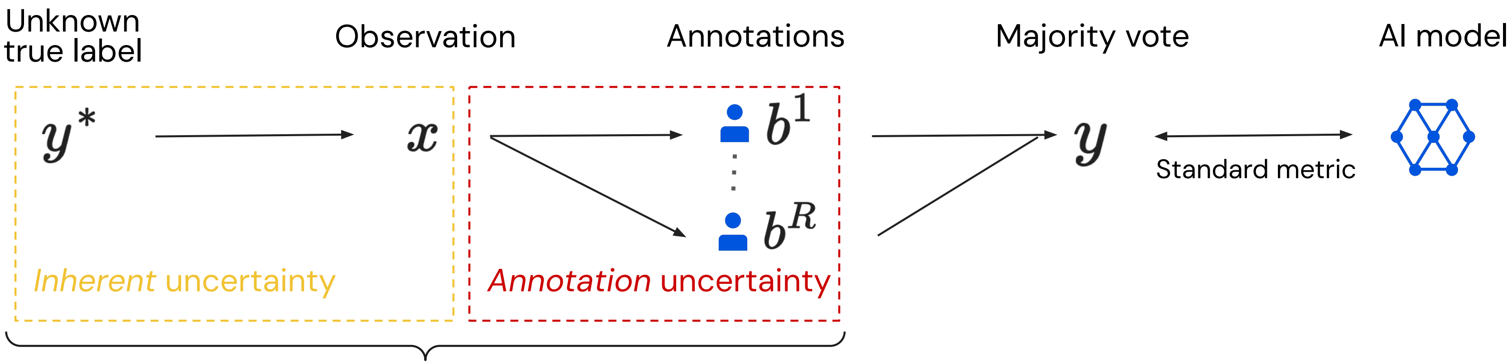


Annotation uncertainty = uncertainty induced through human annotators:

- Subjective tasks
- Inexperience of annotators
- Insufficient training of annotators
- Inappropriate annotation tool
- Different biases or background from annotators

TL;DR: annotation is difficult.

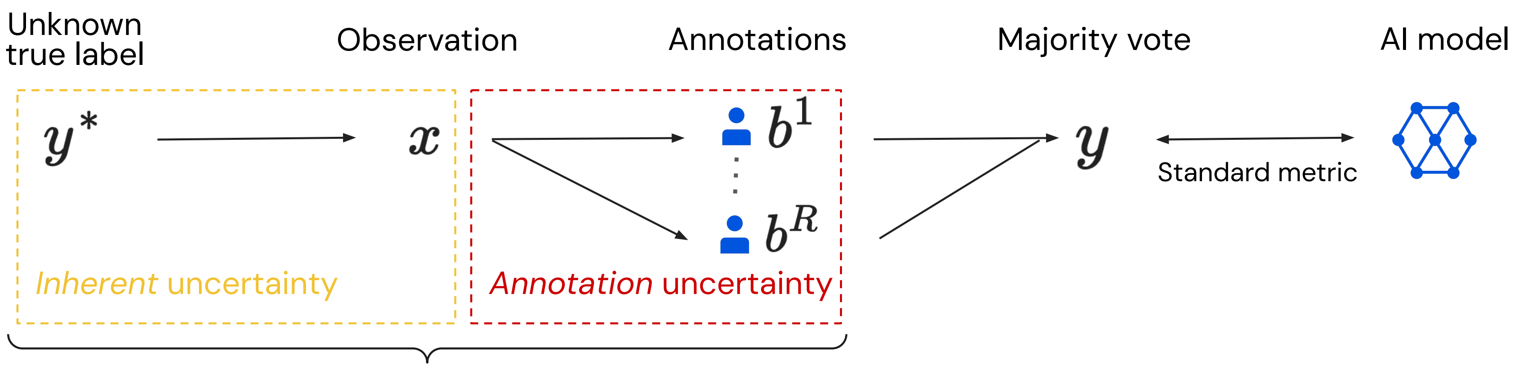
Ground truth uncertainty



Ground truth uncertainty = inherent + annotation uncertainty

- We observe both through annotation **disagreement**

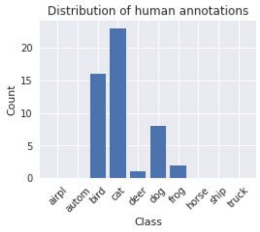
Ground truth uncertainty



Ground truth uncertainty = inherent + annotation uncertainty

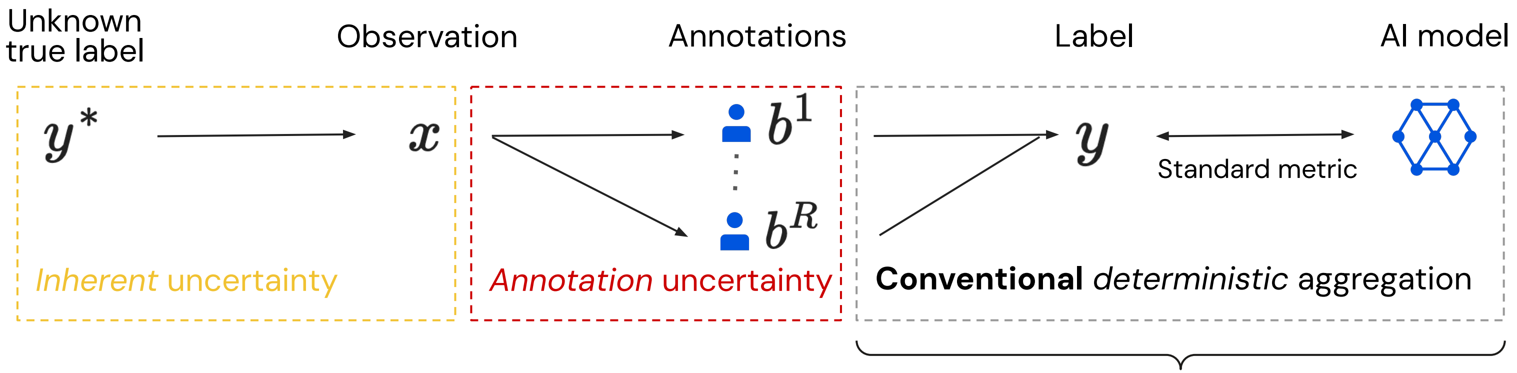
- We observe both through annotation disagreement
- Usually we cannot disentangle between inherent and annotation uncertainty

Inherent uncertainty →



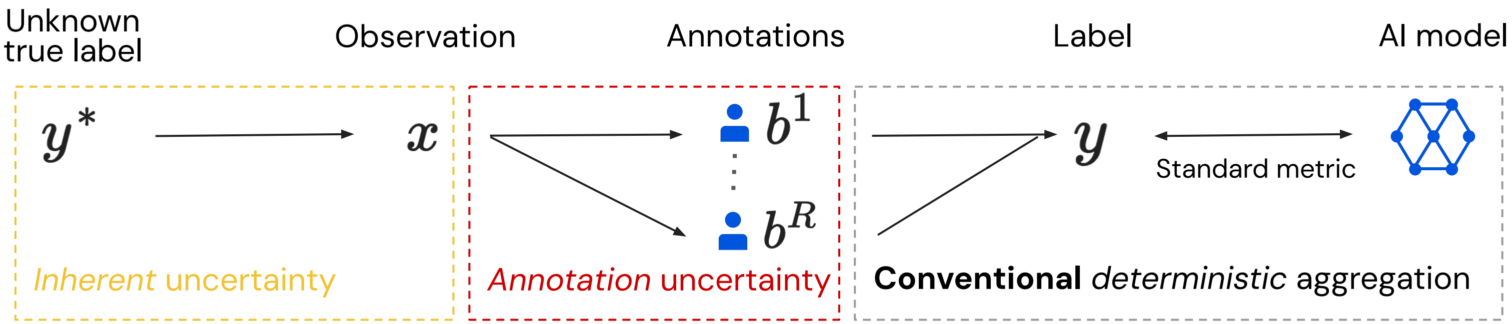
← Annotation uncertainty

Deterministic aggregation ignores uncertainty

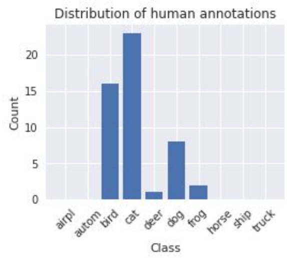


Simple majority voting or deterministic aggregation ignores this uncertainty

Deterministic aggregation ignores uncertainty



Deterministic aggregation:
- ➔ Might evaluate against the wrong labels
- ➔ Ignores large parts of the annotators
- ➔ Does not quantify uncertainty on top of metrics



Introducing *plausibilities*

Observation

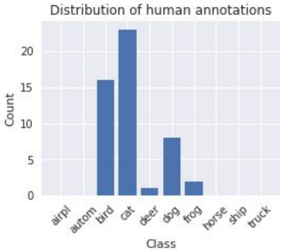
Annotations

b^1

$\vdots$

b^R

Plausibilities $\hat{\lambda}$



argmax

Label

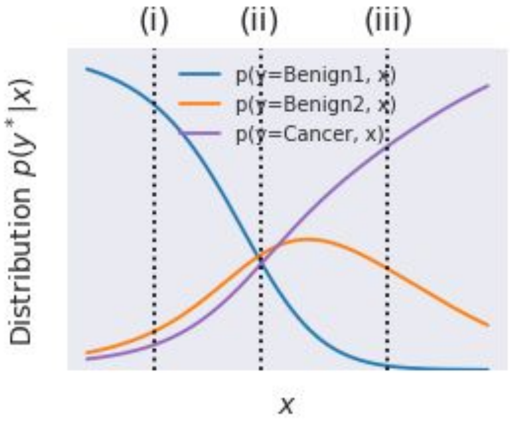
"cat"

Deterministic aggregation approximates posterior $p(y^*|x)$
using a point estimate $\hat{\lambda}$

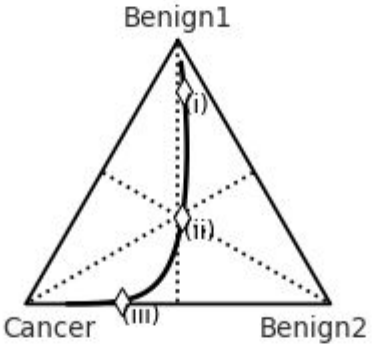
- "Plausibilities" = how *plausible* is a label given the annotations
- In this talk: categorical distributions over classes

Plausibilities on one-dimensional toy example

Example cases

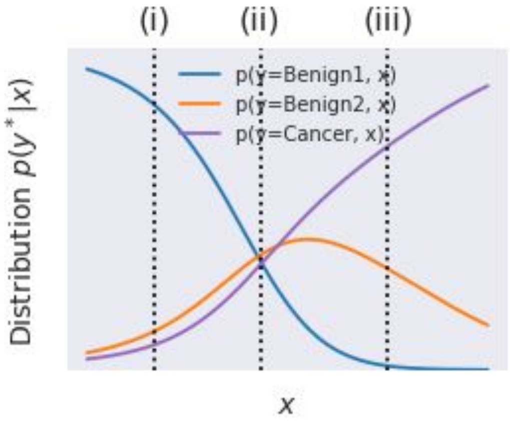


True posterior on
3-simplex

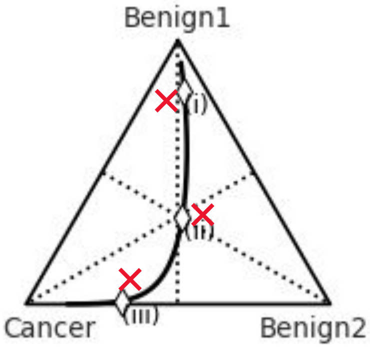


Inherent uncertainty =
location on simplex

Point estimates from deterministic aggregation



True posterior
◇ = $\lambda^* = p(y^*|x)$
on 3-simplex



Annotations

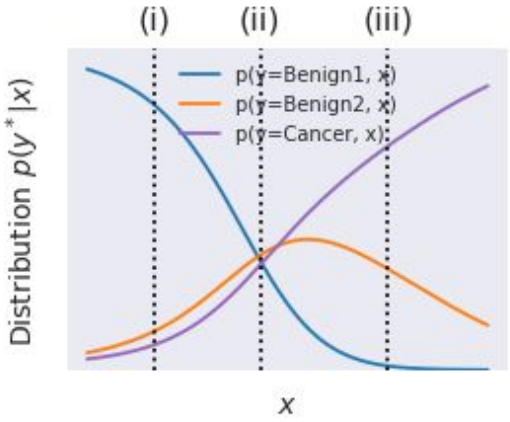
b^1
⋮
b^R

Deterministic aggregation

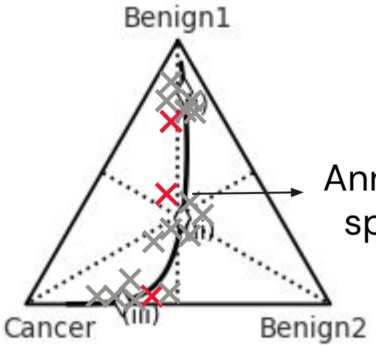
× = $\hat{\lambda}$

"Annotators see $p(y^*|x)$ more or less accurately"

Variation in plausibilities through re-annotating



True posterior
◊ = $\lambda^* = p(y^*|x)$
on 3-simplex



Annotation uncertainty =
spread of plausibilities

Annotations

b^1

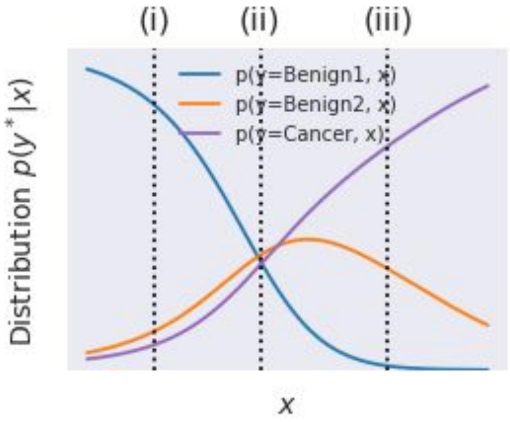
b^R

Deterministic
aggregation

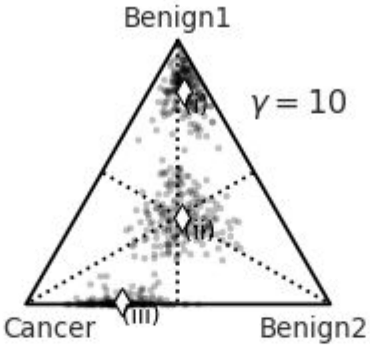
× = $\hat{\lambda}$

Different annotators yield
different plausibilities $\hat{\lambda}$

Statistical aggregation



True posterior
◇ = $\lambda^* = p(y^*|x)$
on 3-simplex



$b^r \sim p(b|\lambda^*)$

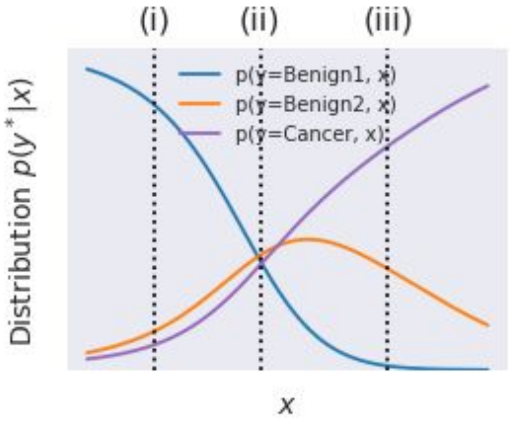
Annotation model

b^1
⋮
b^R

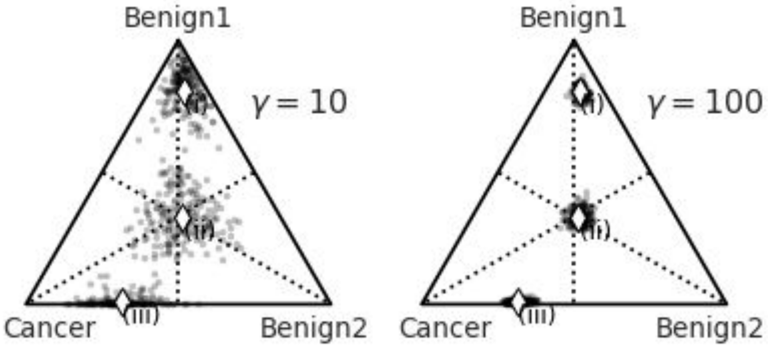
$\lambda^m \sim p_\gamma(\lambda|b, x)$

Statistical aggregation model

Annotator *reliability* in statistical aggregation



True posterior
◇ = $\lambda^* = p(y^*|x)$
on 3-simplex



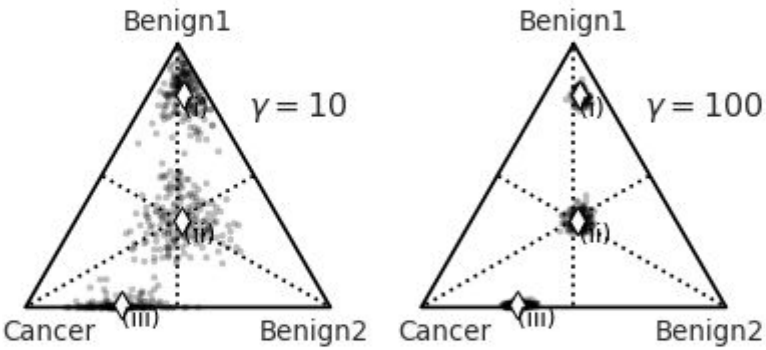
$$b^r \sim p(b|\lambda^*)$$

b^1
⋮
b^R

$$\lambda^m \sim p_\gamma(\lambda|b, x)$$

Reliability γ = lower or higher
prior trust in annotators

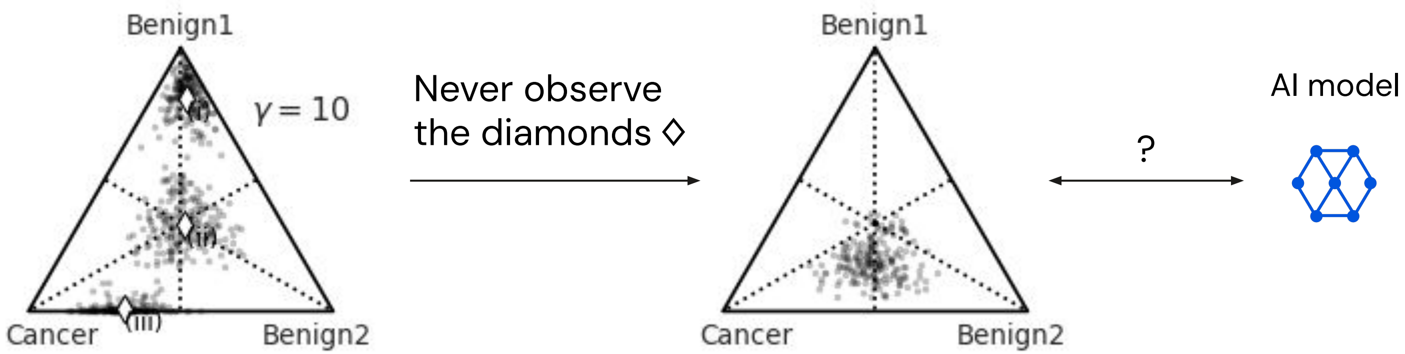
Conclusion: plausibilities on toy example



Ground truth uncertainty on the simplex:

- Location of plausibilities on simplex = inherent uncertainty
- Spread of plausibilities = annotation uncertainty

Conclusion: plausibilities on toy example



In practice, we only observe annotation disagreement:

- Cannot disentangle annotation and inherent uncertainty
- Often impossible to deterministically derive a good approximation $\hat{\lambda}$

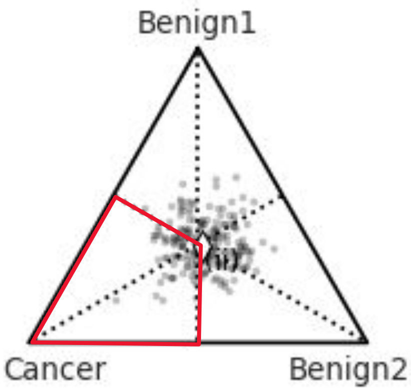
➔ How to *measure* this uncertainty and to *evaluate* AI models?

Measuring annotation uncertainty

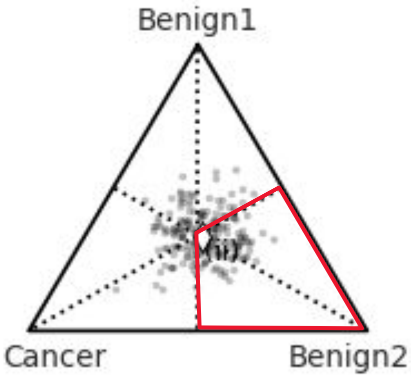
- How *certain* is it that y is the top-1 label?

$$\text{Certainty}(y; b, x) = \mathbb{E}_{p(\lambda|b,x)}\left[\delta[y = \arg\max_j \lambda_j]\right]$$

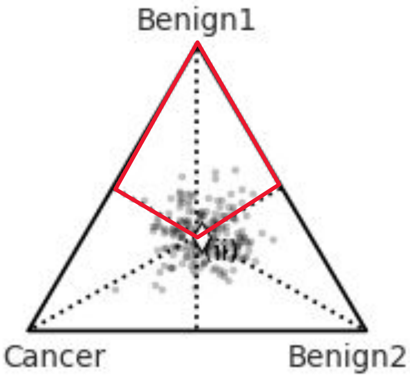
Certainty(Cancer)



Certainty(Benign2)



Certainty(Benign1)

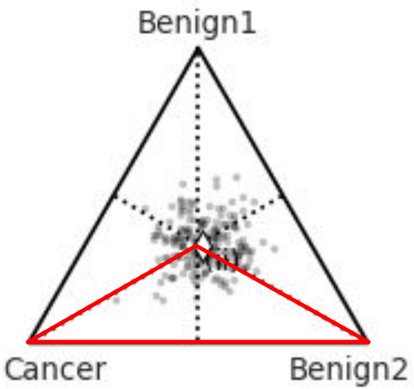


Measuring annotation uncertainty

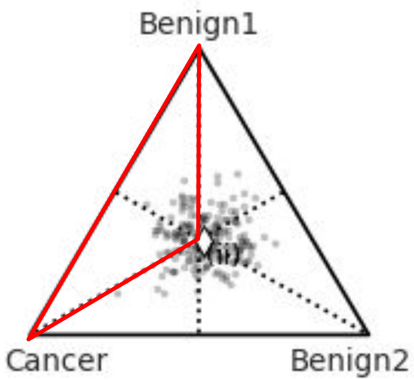
- Can also quantify certainty of label sets Y :

$$\mathrm{Certainty}(Y; b, x) = \mathbb{E}_{p(\lambda|b,x)}\left[\delta[Y = \mathrm{top\_k}(\lambda)]\right]$$

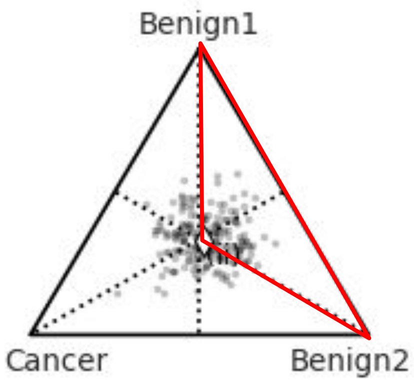
Certainty({Cancer, Benign2})



Certainty({Cancer, Benign2})



Certainty({Benign1, Benign2})



Measuring annotation uncertainty

- How *certain* is it that y is the top-1 label?

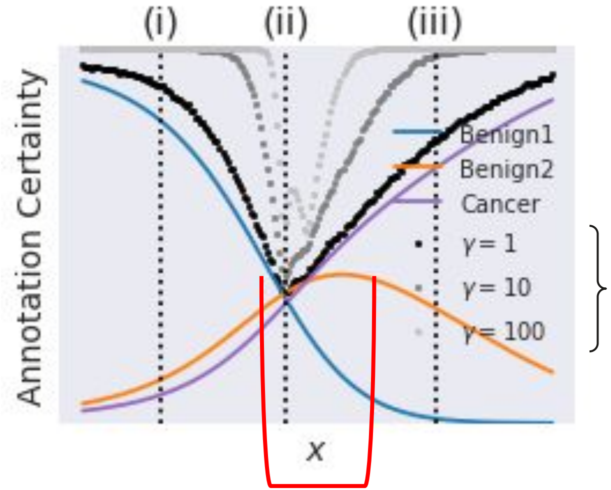
$$\mathrm{Certainty}(y; b, x) = \mathbb{E}_{p(\lambda|b,x)}\left[\delta[y = \arg\max_j \lambda_j]\right]$$

- What is the highest certainty across labels?

$$\mathrm{AnnotationCertainty}(b, x) = \max_y \mathrm{Certainty}(y; b, x)$$

Measuring annotation uncertainty

- Annotation certainty on toy example for different reliabilities γ :

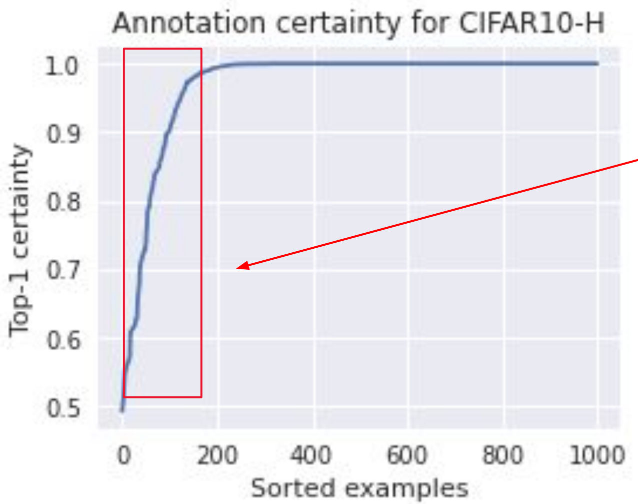


Different reliabilities

Top-1 label unclear uncertain irrespective of how much we trust our annotators

Measuring annotation uncertainty

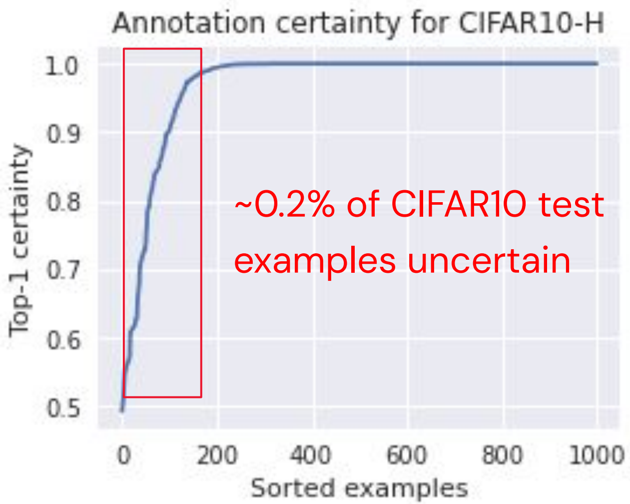
- Annotation certainty on CIFAR10 using annotations from CIFAR10-H:



- 178 examples with annotation certainty < 0.99
- This is ~0.2% of all CIFAR10 test examples

Measuring annotation uncertainty

- Annotation certainty on CIFAR10 using annotations from CIFAR10-H:



Papers with Code leaderboard:

μ2Net	99.49	2022
ViT-L/16	99.42	2020
CaiT-M-36 U 224	99.4	2021
CvT-W24	99.39	2021
BiT-L	99.37	2019
ViT-B	99.3	2022

Improvements within 0.2%

Uncertainty-adjusted (top-k) accuracy

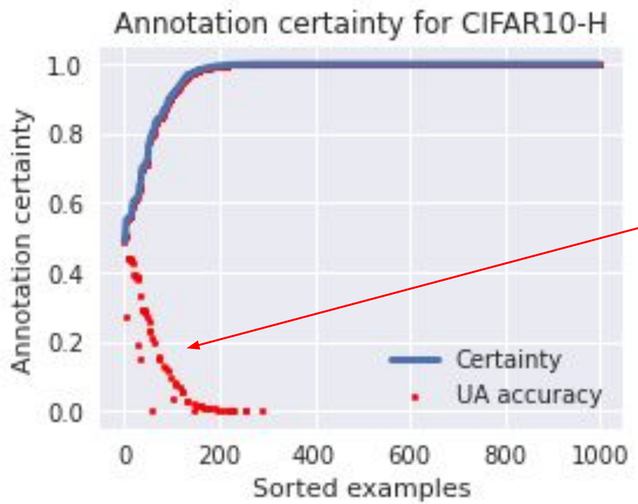
Given a model that yields a top-k prediction set $C_{\text{top-}k}(x)$:

$$\text{UA-Accuracy}_{\text{top-}k} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\lambda|b,x)}\left[\delta[\arg\max_j \lambda_j \in C_{\text{top-}k}(x)]\right]$$

Uncertainty-adjusted (top-k) accuracy

Given a model that yields a top-k prediction set $C_{\text{top-}k}(x)$:

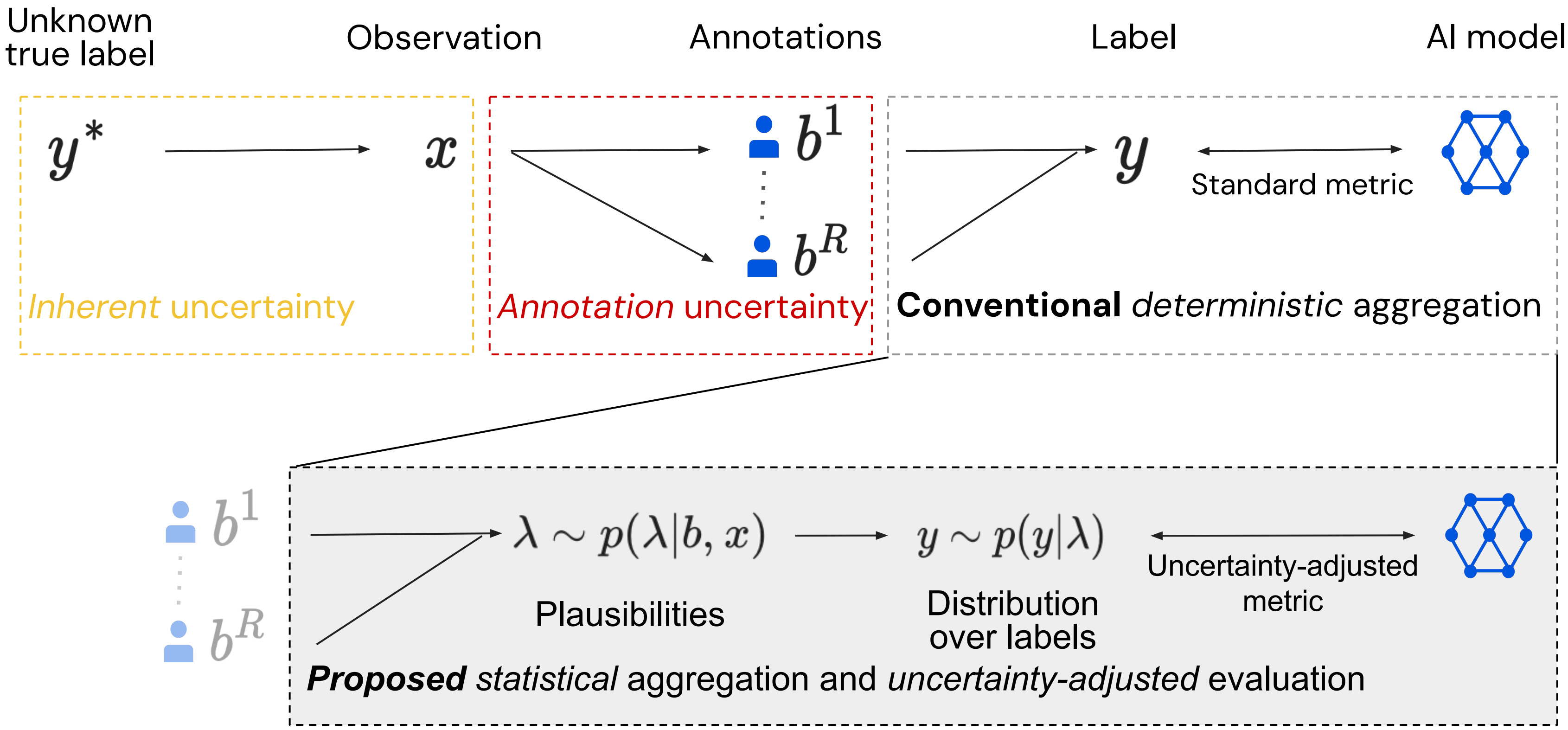
$$\text{UA-Accuracy}_{\text{top-}k} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\lambda|b,x)}\left[\delta[\arg\max_j \lambda_j \in C_{\text{top-}k}(x)]\right]$$



- $C_{\text{top-}k}(x)$ = original CIFAR10 labels (k = 1)
- Annotations taken from CIFAR10-H

➔ Even CIFAR10 labels perform poorly on uncertain examples!

Summary: proposed statistical framework



Case study: dermatology

Task: predict dermatological conditions from (multiple, consumer-grade) images.

Observation

Annotations

b^1: {*Pyogenic granuloma* (Low)} {*Hemangioma* (Med)} {*Melanoma* (High)}
b^2 {*Angiokeratoma of skin* (Low)} {*Atypical Nevus* (Med)}
b^3: {*Hemangioma* (Med)} {*Melanocytic Nevus* (Low), *Melanoma* (High), *O/E - ecchymoses present* (Low)}
b^4: {*Hemangioma* (Med), *Melanoma* (High), *Skin Tag* (Low)}
b^5: {*Melanoma* (High)}
b^6: {*Hemangioma* (Med)} {*Melanoma* (High)} {*Melanocytic Nevus* (Low)}

Conditions, Low/Med/High risk conditions

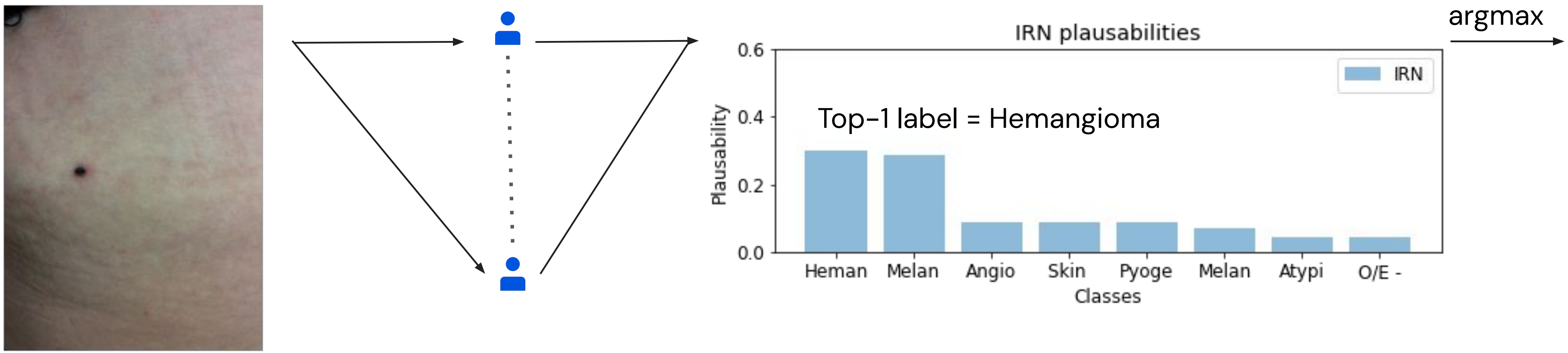
Partial rankings to model differential diagnoses

Case study: deterministic aggregation using IRN

Task: predict dermatological conditions from images.

- Inverse rank normalization (IRN) to aggregate annotators' differential diagnoses.

Observation Annotations Plausibilities $\hat{\lambda}$



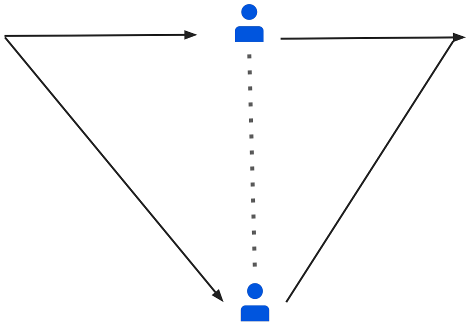
Case study: statistical aggregation using PrIRN

Task: predict dermatological conditions from images.

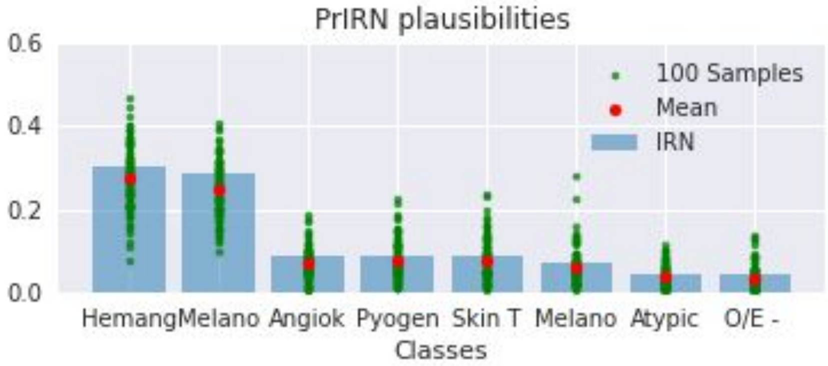
- Plackett-Luce or probabilistic IRN (PrIRN) to model $p(\lambda|b)$

Observation

Annotations



Plausibilities $\lambda^m \sim p_\gamma(\lambda|b)$



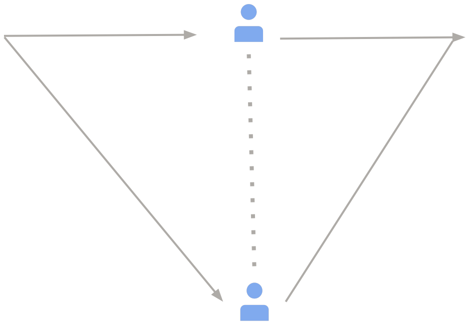
Case study: statistical aggregation using PrIRN

Task: predict dermatological conditions from images.

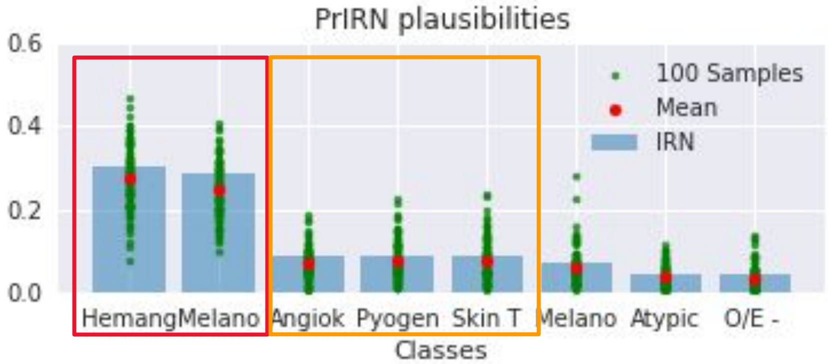
- Plackett-Luce or probabilistic IRN to model $p(\lambda|b)$

Observation

Annotations



Plausibilities $\lambda^m \sim p_\gamma(\lambda|b)$

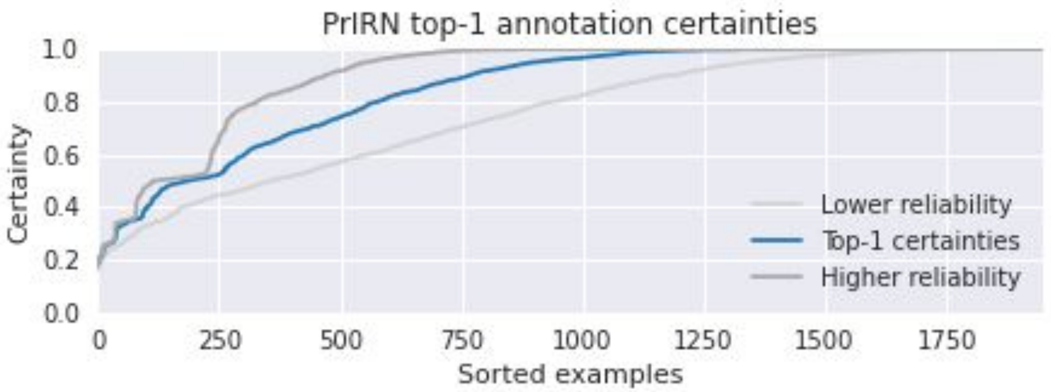


Top-1 condition changes easily = low annotation certainty

3rd, 4th, 5h conditions also change easily

High annotation uncertainty

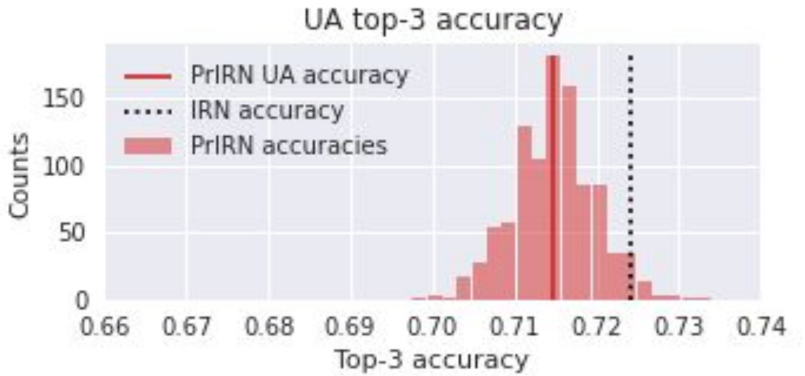
- Significant portions of cases with high annotation uncertainty:



➔ In discussions with dermatologists often attributed to inherent uncertainty

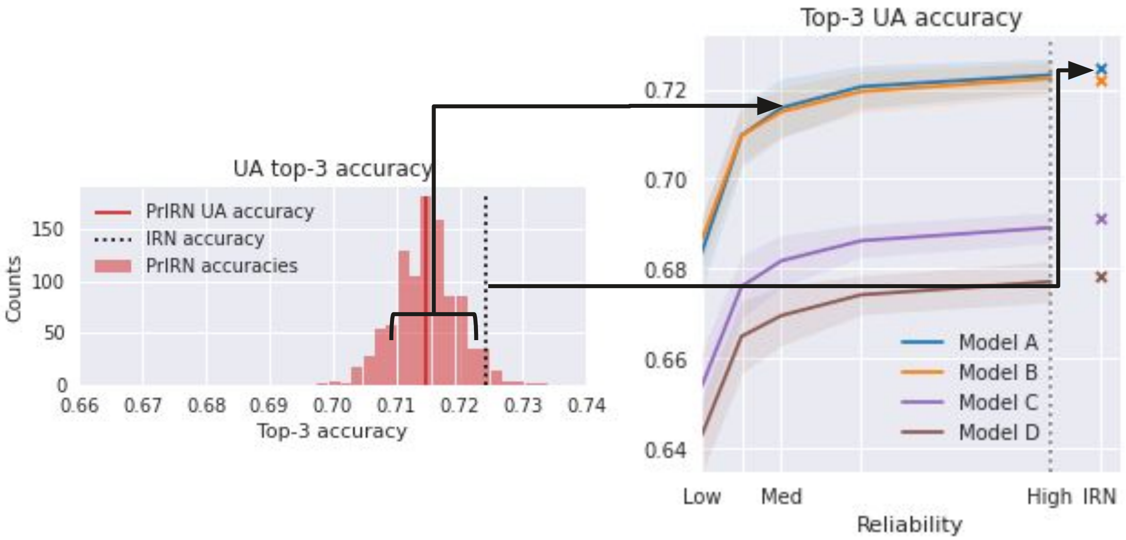
Uncertainty-adjusted top-3 accuracy

- Across cases / per plausibility:



➔ Significant variation in top-3 accuracy

Evaluation across annotator reliabilities

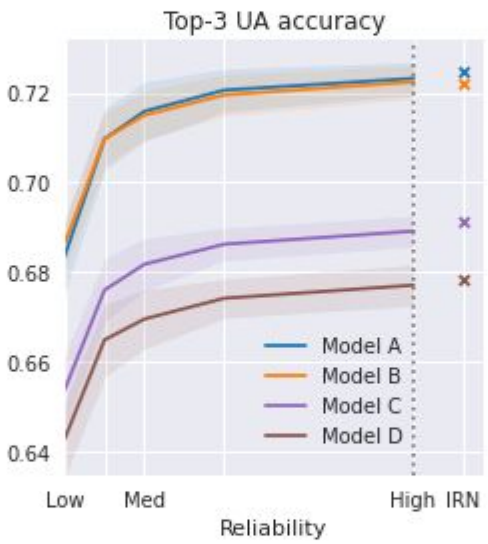


- UA accuracy varies significantly by reliability
- IRN implicitly evaluates infinite annotator reliability
- Large spread/uncertainty in accuracies (shaded)

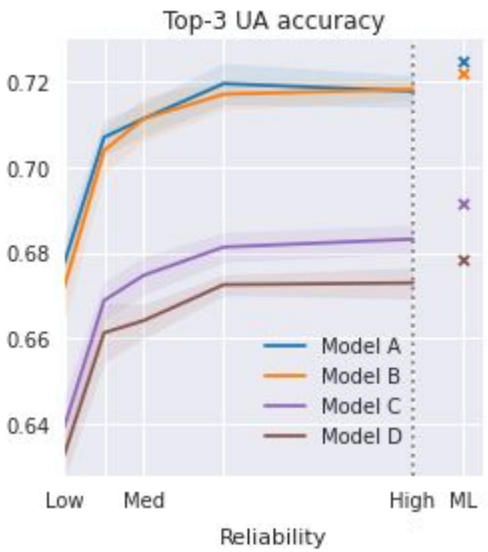
Alternative statistical aggregation methods

- Alternative statistical aggregation models exhibit different results:

PrIRN



Plackett-Luce



➔ Aggregation is a mode choice usually not made explicit!

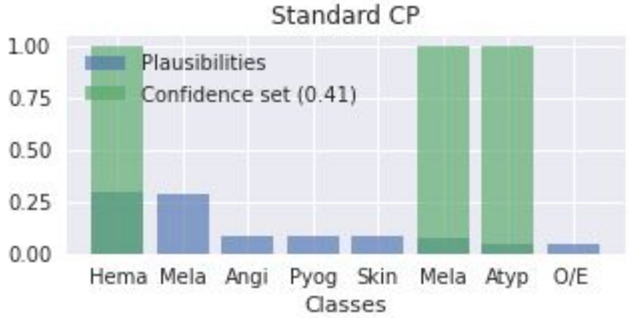
Bonus: calibration with uncertain ground truth

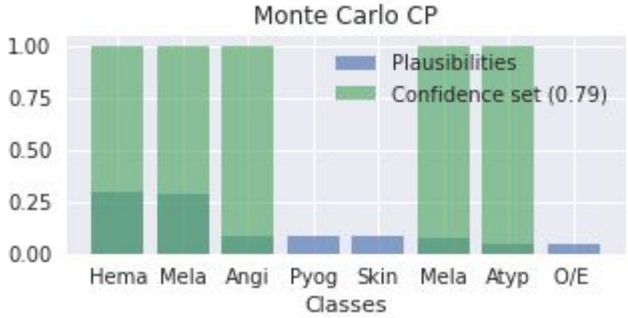
Calibration usually based on ground truth labels on a calibration/validation set:

- Conformal prediction uses ground truth labels to calibrate a softmax threshold τ
- Threshold used to predict confidence sets of classes at test time instead of the top-k:

$$C_{\text{top-}k}(x) \longrightarrow C_{\text{CP}}(x) := \{k \in [K] : k - \text{th softmax} \geq \tau\}$$

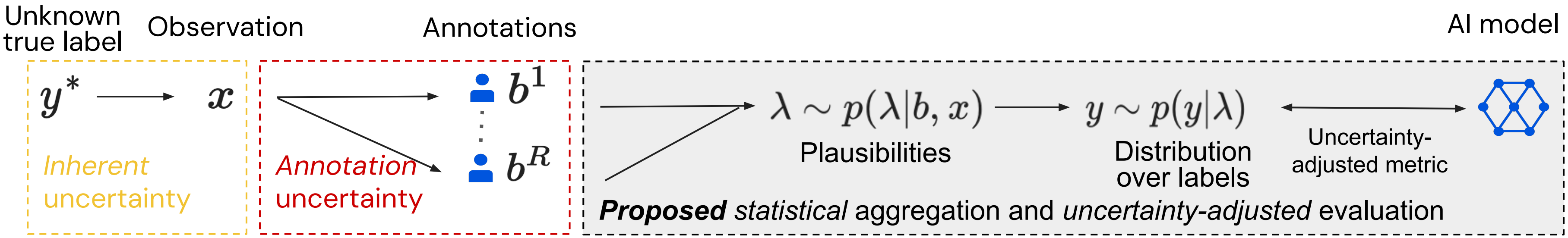
- We propose *Monte Carlo* conformal prediction to address this issue, improving uncertainty-adjusted performance



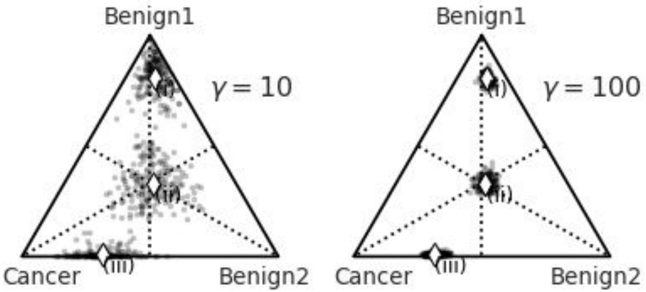


Conclusion

Proposed a statistical framework for dealing with ground truth uncertainty:



- ➔ Ground truth uncertainty = inherent + annotation uncertainty (location + spread of plausibilities)
- ➔ *Annotation certainty* explicitly measures annotation uncertainty
- ➔ Uncertainty-adjusted metrics to evaluate and evaluate models



More details: arxiv.org/abs/2307.02191 | dstutz@google.com