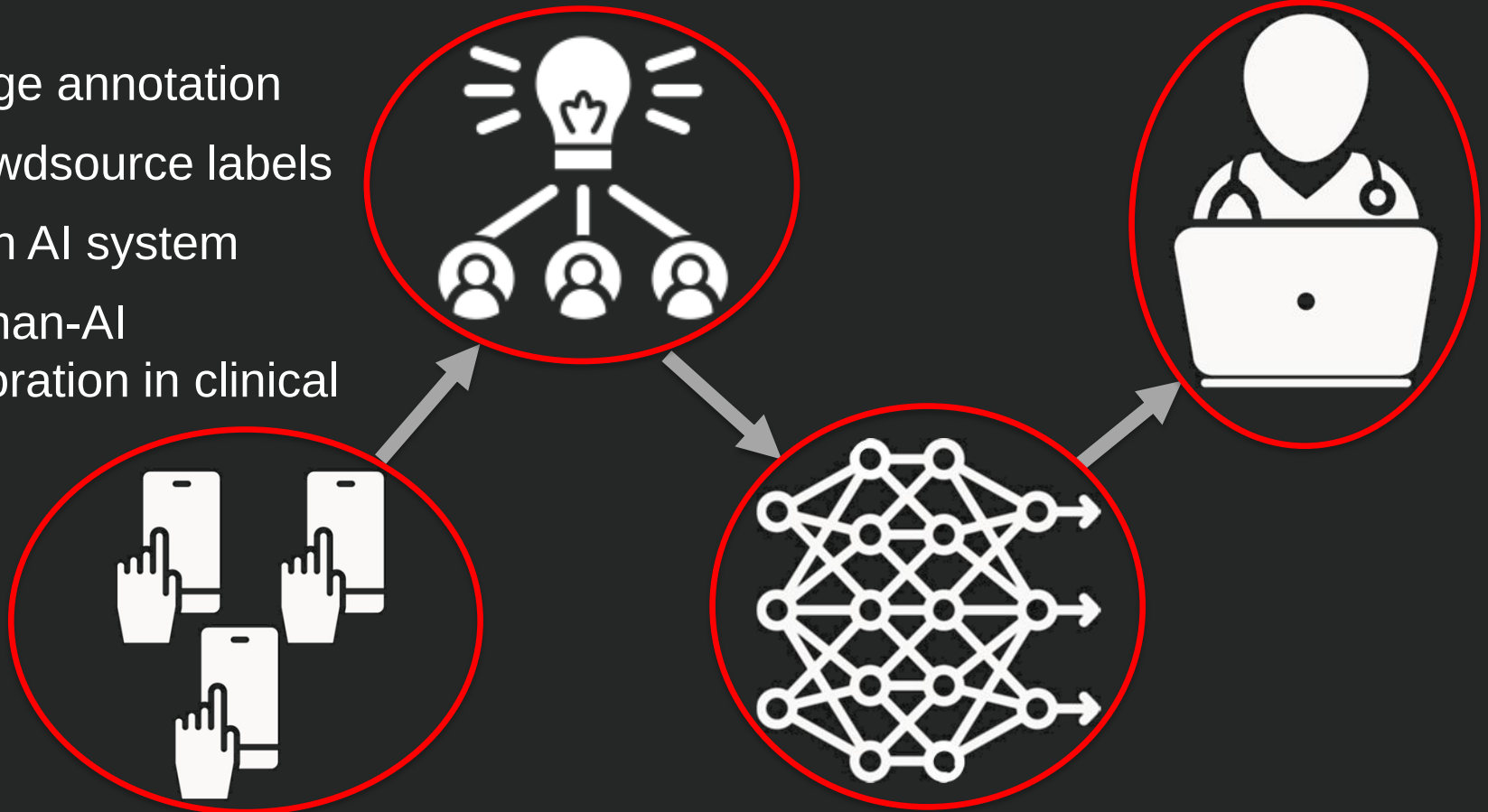
Crowdsourcing a labeled dataset using binary choices versus probability judgments

Gunnar Epping
INDIANA UNIVERSITY BLOOMINGTON
Andrew Caplin, William Holmes, Daniel Martin, and Jennifer Trueblood

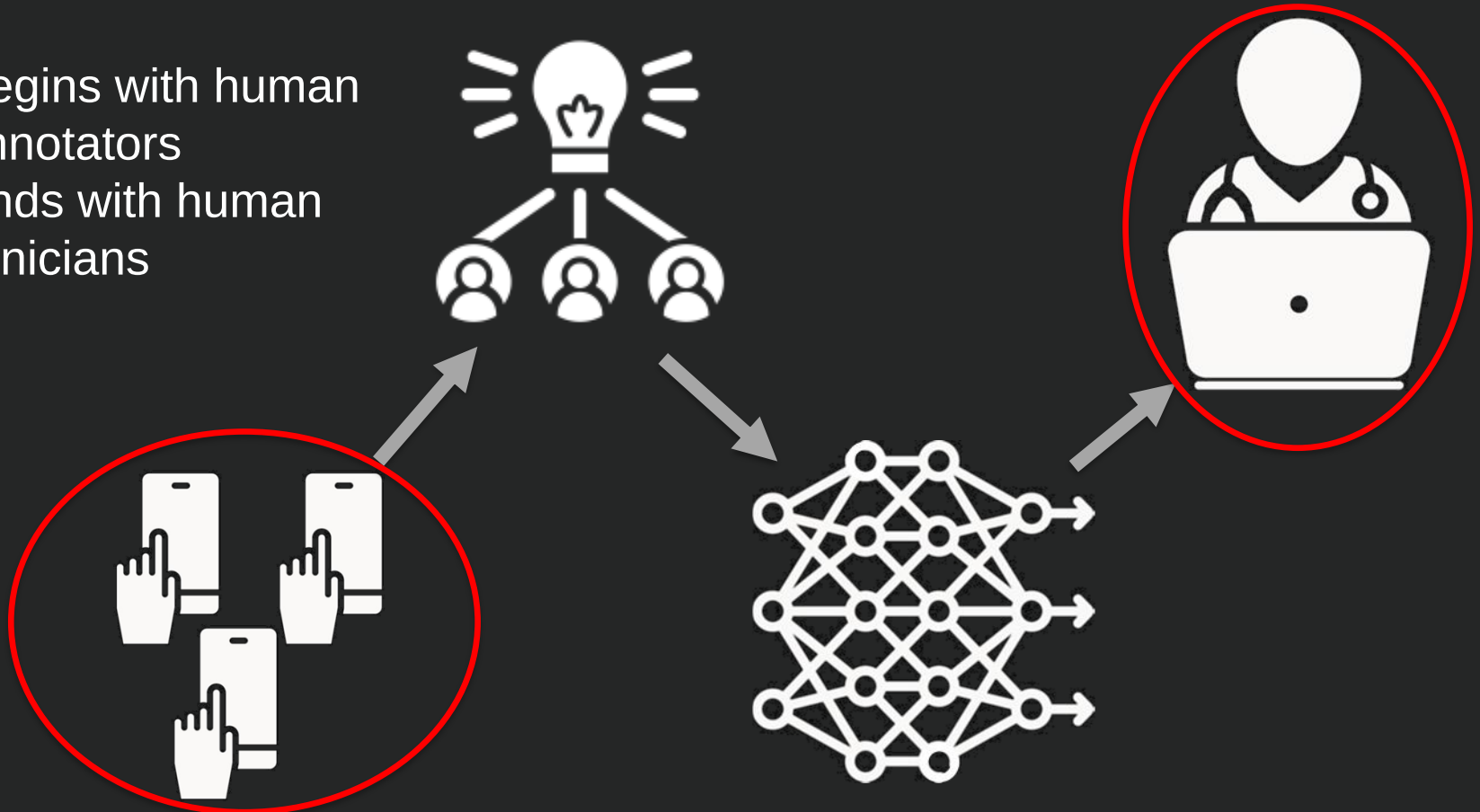
Medical AI information value chain

1. Image annotation
2. Crowdsource labels
3. Train AI system
4. Human-AI collaboration in clinical setting



Humans are critical to the value chain

- Begins with human annotators
- Ends with human clinicians



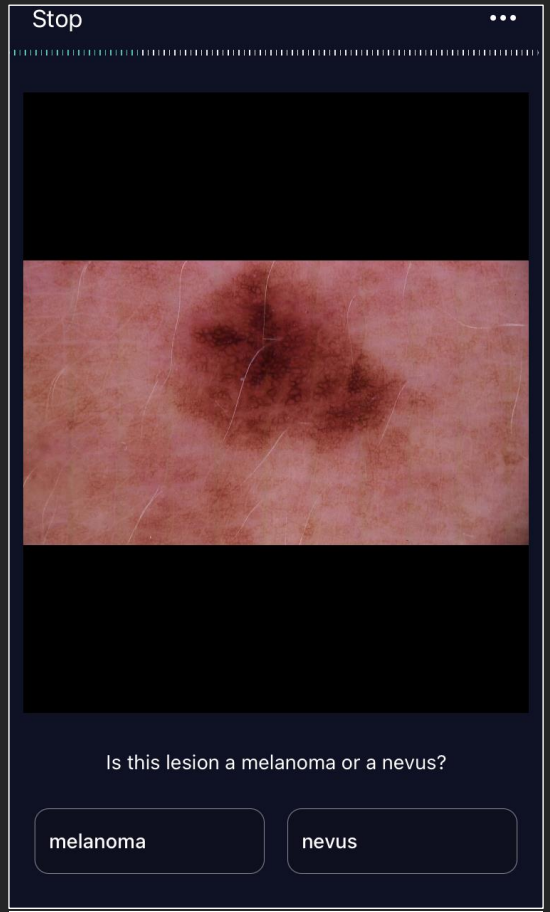
Improving medical AI by understanding the people involved

- Our focus: data annotation
  - Impacts entire chain
  - Major bottleneck for medical AI

- Can cognitive psychology and experimental economics improve the data annotation process?

Current approaches to annotation

1. Elicit binary choices (classifications) from skilled annotators



Centaur Labs

Current approaches to annotation

1. Elicit binary choices from skilled annotators

2. Aggregate binary choices into a single label using Wisdom of the Crowd (WoC)

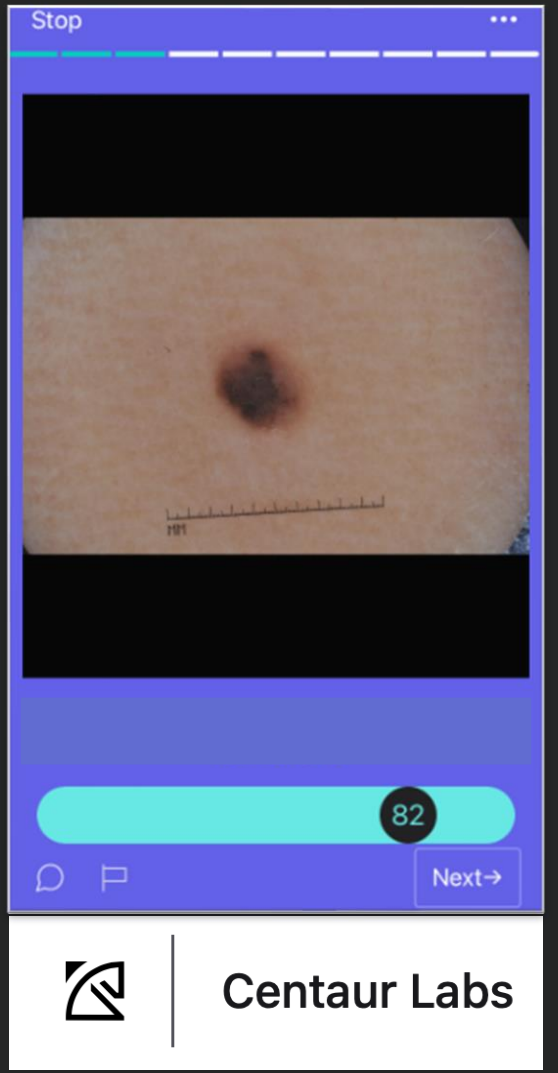
67% nevus
33% melanoma

WoC label:
nevus

Centaur Labs

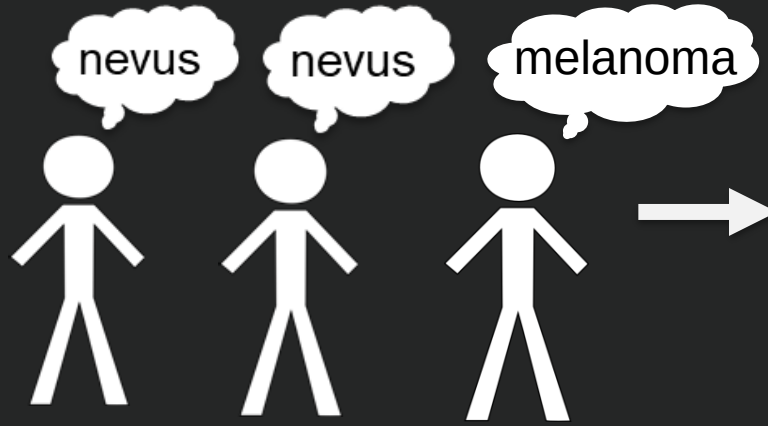
How can we improve the annotation process?

- Probability judgments allow participants to express uncertainty
- Binary choices throw away information



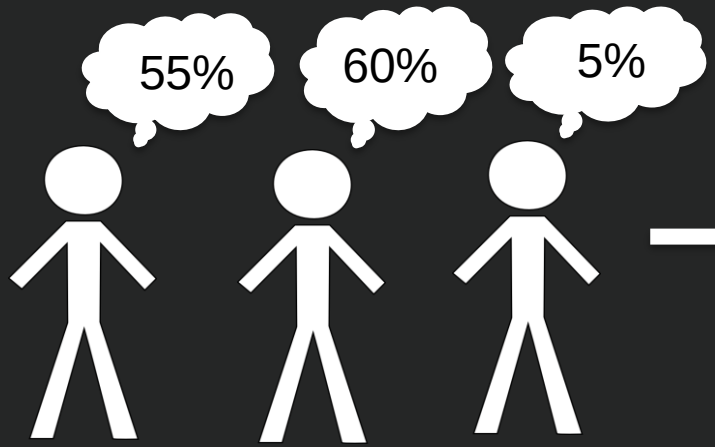
What are the most useful annotations?

"What is the probability that this lesion is a nevus?"



67% nevus
33% melanoma

WoC label:
nevus



40% nevus
60% melanoma

WoC label:
melanoma

Research questions

- How does annotation mode (binary choices versus probability judgments) impact the accuracy of Wisdom of the Crowd (WoC) labeling approaches?

- How does annotation mode impact the accuracy and calibration of models trained on WoC labels?

Study overview

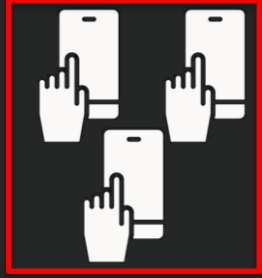
1. Collect annotations from non-experts, both binary choices and probability judgments

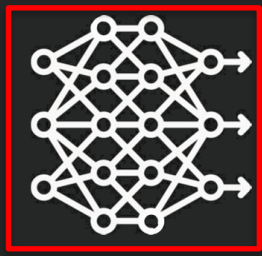
2. Aggregate annotations into crowdsourced labels using WoC

3. Evaluate the accuracy of WoC labeled datasets

4. Train machine learning models on WoC labeled datasets

5. Evaluate the accuracy and calibration of the models

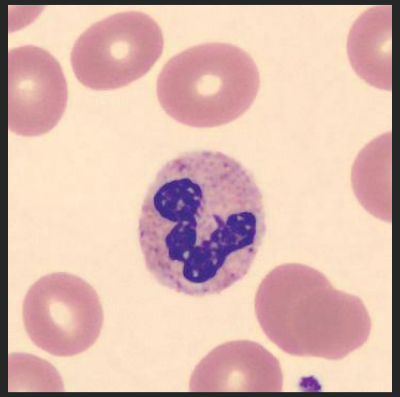


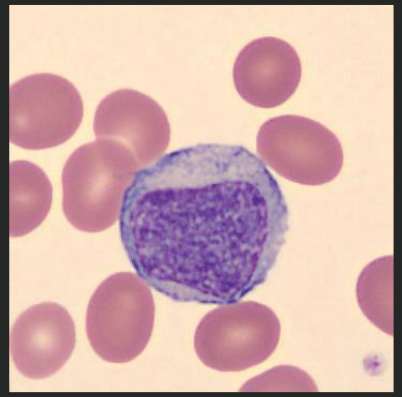


Stimuli

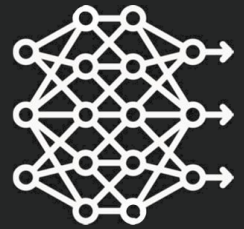
- White blood cell images showing either a **blast** (cancer) cell or **non-blast** (non-cancer) cell
- Image Curation: Trueblood et al. (2018) had three Vanderbilt pathology faculty (experts) independently classify 840 images
- Out of 840 images, there were 633 images with three-way agreement
  - We used a subset of 433 images
- We take expert agreement to be ground truth

Non-blast



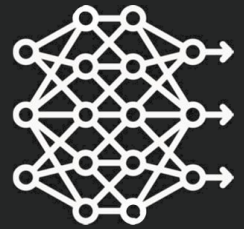


Blast



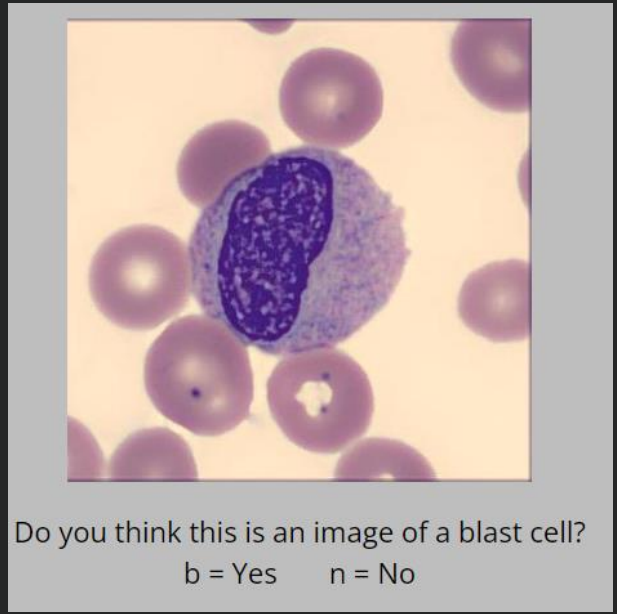
Participants

- 400 total participants from MTurk
  - 200 provided binary choice annotations and other 200 provided probability judgment annotations
  - $1 base pay
  - Potential $5 bonus
- Participants were trained before providing annotations during a testing phase
  - First, they looked at several images from each class
  - Second, they practiced classifying images with feedback
- Study was pre-registered (AsPredicted #122152)



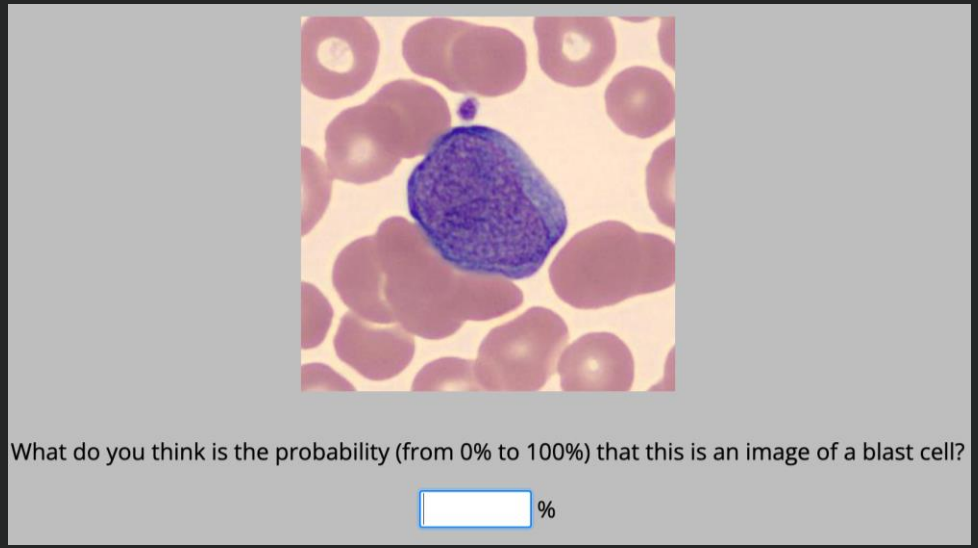
Testing phase for binary choices

- 100 trials
- No feedback
- One trial randomly chosen for $5 bonus



Testing phase for probability judgments

- 100 trials
- No feedback
- One trial randomly chosen for $5 bonus
- Leveraged proper scoring rules to incentivize truthful reporting of probabilities



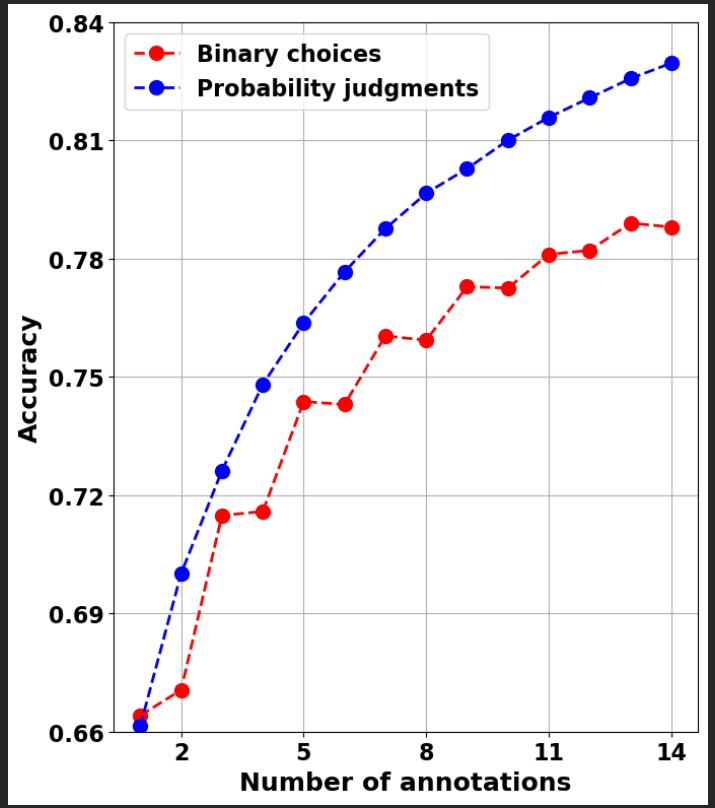
Wisdom of the Crowd accuracy results

Annotation mode	Mean individual accuracy	WoC label accuracy
Binary choice	.663	.829
Probability judgment	.663	**.882**

- When computing accuracy, probabilities are binarized
  - p(blast) > 50% → blast classification
  - p(blast) ≤ 50% → non-blast classification
- Individuals are equally accurate, on average
- WoC probability judgments are much more accurate

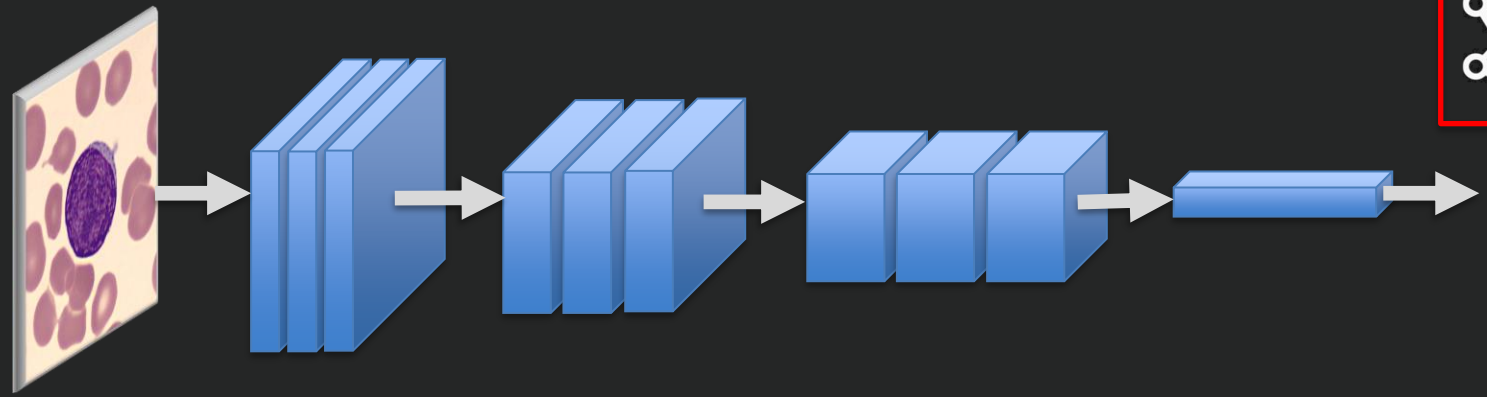
How does the number of annotations impact WoC accuracy?

- Between 14 and 64 annotations per image
- How does the number of annotations impact accuracy?
  - Randomly select a subset of the annotations
- Fewer probability judgments are needed to reach an arbitrary level of accuracy

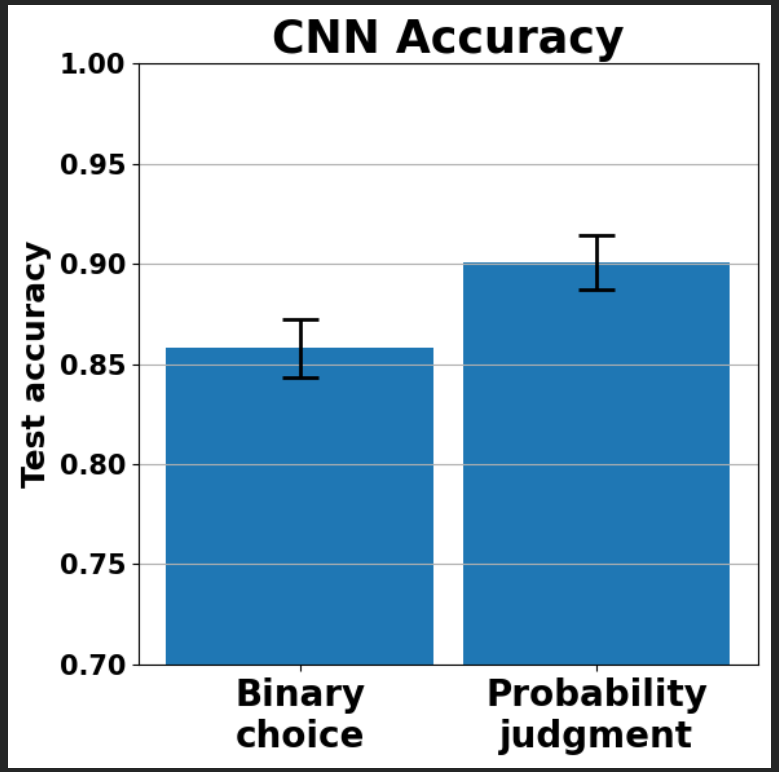


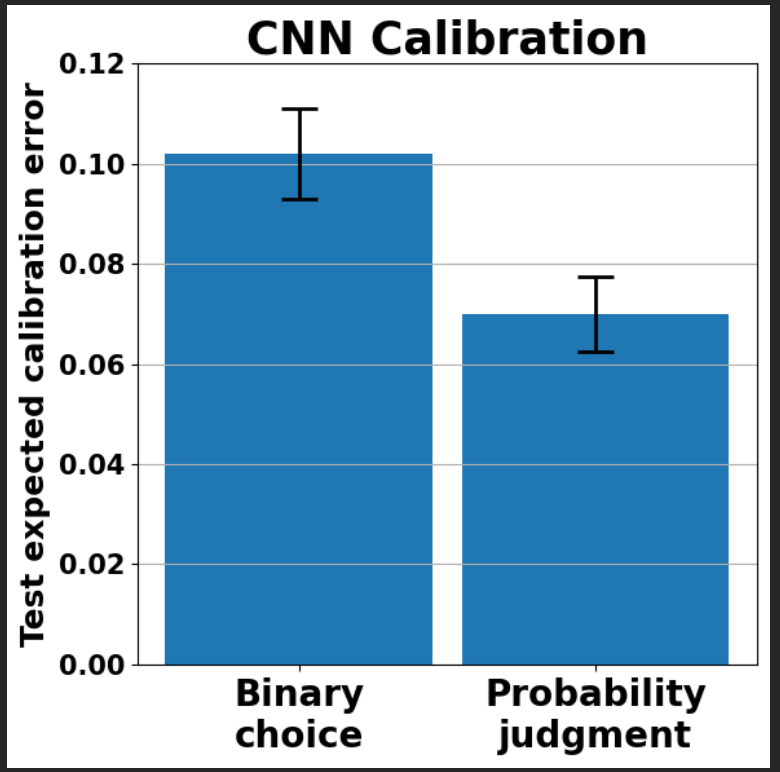
Machine learning model training

- Trained a convolutional neural network on the WoC labeled datasets
- We used 30 random testing/training splits
  - 80% of images/labels were used for training, other 20% were used for evaluating the models



Model accuracy and calibration results





Conclusions

Q1: How does annotation mode (binary choices versus probability judgments) impact the accuracy of WoC labeling approaches?

C1: Probability judgments lead to more accurate labels
Bonus: Fewer probability judgments are required

Q2: How does annotation mode impact the accuracy and calibration of models trained on WoC labels?

C2: Model trained on labels obtained via probability judgments are more accurate and better calibrated

Our team

Indiana University,
Cognitive Science

Gunnar Epping

Jennifer Trueblood

William Holmes

New York University,
Economics

Andrew Caplin

University of California,
Santa Barbara,
Economics

Daniel Martin

Supported by the Sloan Foundation grant “Cognitive Economics at Work”

Discussion

Thank you for attending this talk!

Any questions?

If you have any lingering questions, feel free to talk with me afterwards or email me at gepping@iu.edu

Calibration curves