Behavioral Science Contributions to Medical Image Decision-making

Jennifer Trueblood
Vanderbilt University

Acknowledgements

VU

Eeshan Hasan

William Holmes

Wenrui Huang

Payton O'Daniels

Megan Woodruff

VUMC

Margaret Compton

Jonathan Douds

Quentin Eichbaum

Adam Seegmiller

Charles Stratton

Eszter Szentirmai

Alfred P. Sloan Foundation

ASCP STRONGERTOGETHER

Vanderbilt University Medical Center

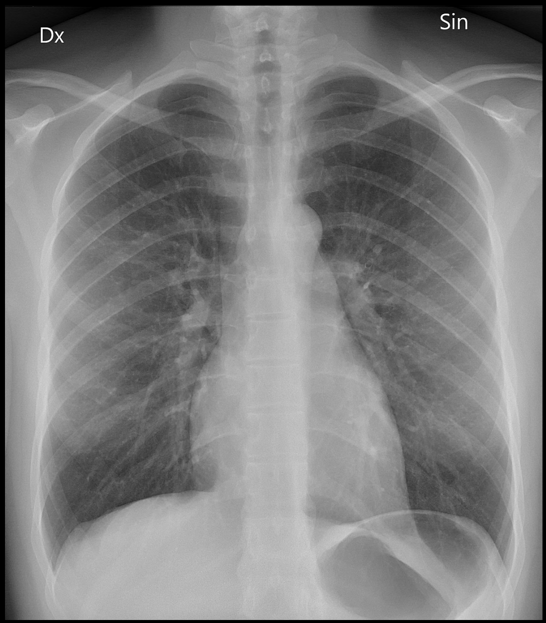
Outline

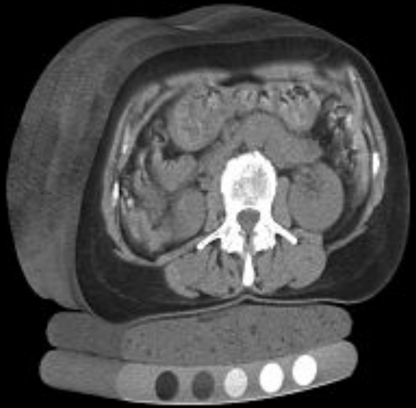
- Introduction to Cognitive and Perceptual work in Medical Image Decision-making
- Examples
  - Factors that affect Pathology Decisions
  - Strategies to Reduce Errors
- Research Gaps and Open Problems
- Resources

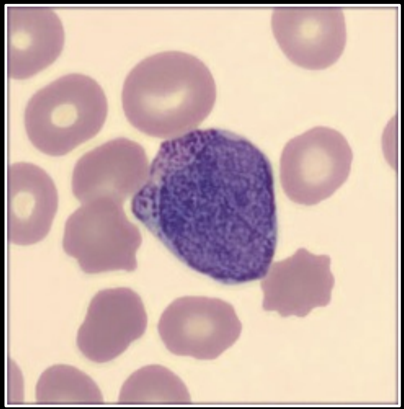
What is Medical Image Decision-making?

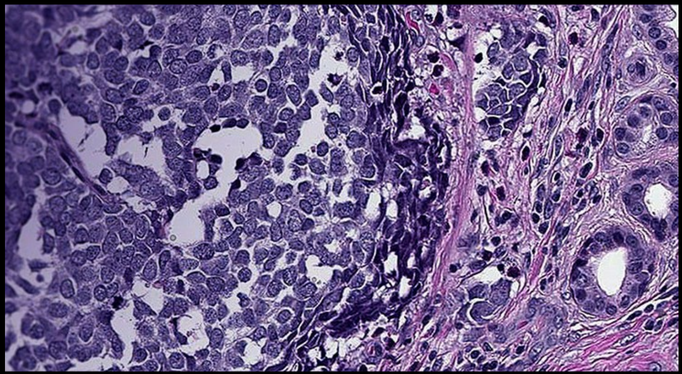
- Detecting and diagnosing diseases from medical images typically by identifying abnormalities
  - All forms of image acquisition from light microscopy to magnetic resonance
  - All forms of presentation (e.g., viewing glass slides through a microscope to radiographs on a computer)
  - Generally the fields of radiology, pathology, and dermatology

Examples of Medical Images







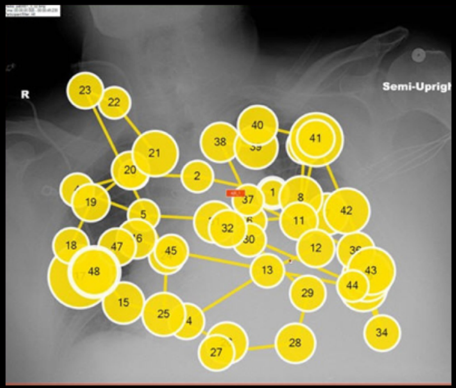


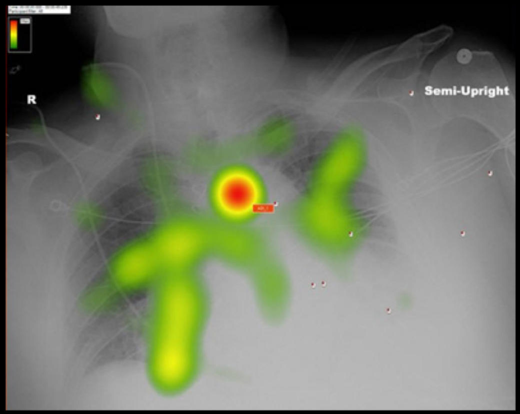
Diagnostic Errors

- How can we reduce **diagnostic errors?**
  - Improve imaging
  - Construct better Computer-Aided Detection (CAD) systems
  - Understand the cognitive and perceptual processes that lead to errors and improve clinical practices

Psychology and Medical Image Interpretation

- Growing interest in using cognitive and vision sciences to study medical image observers
  - NCI special funding opportunity (2017-present)

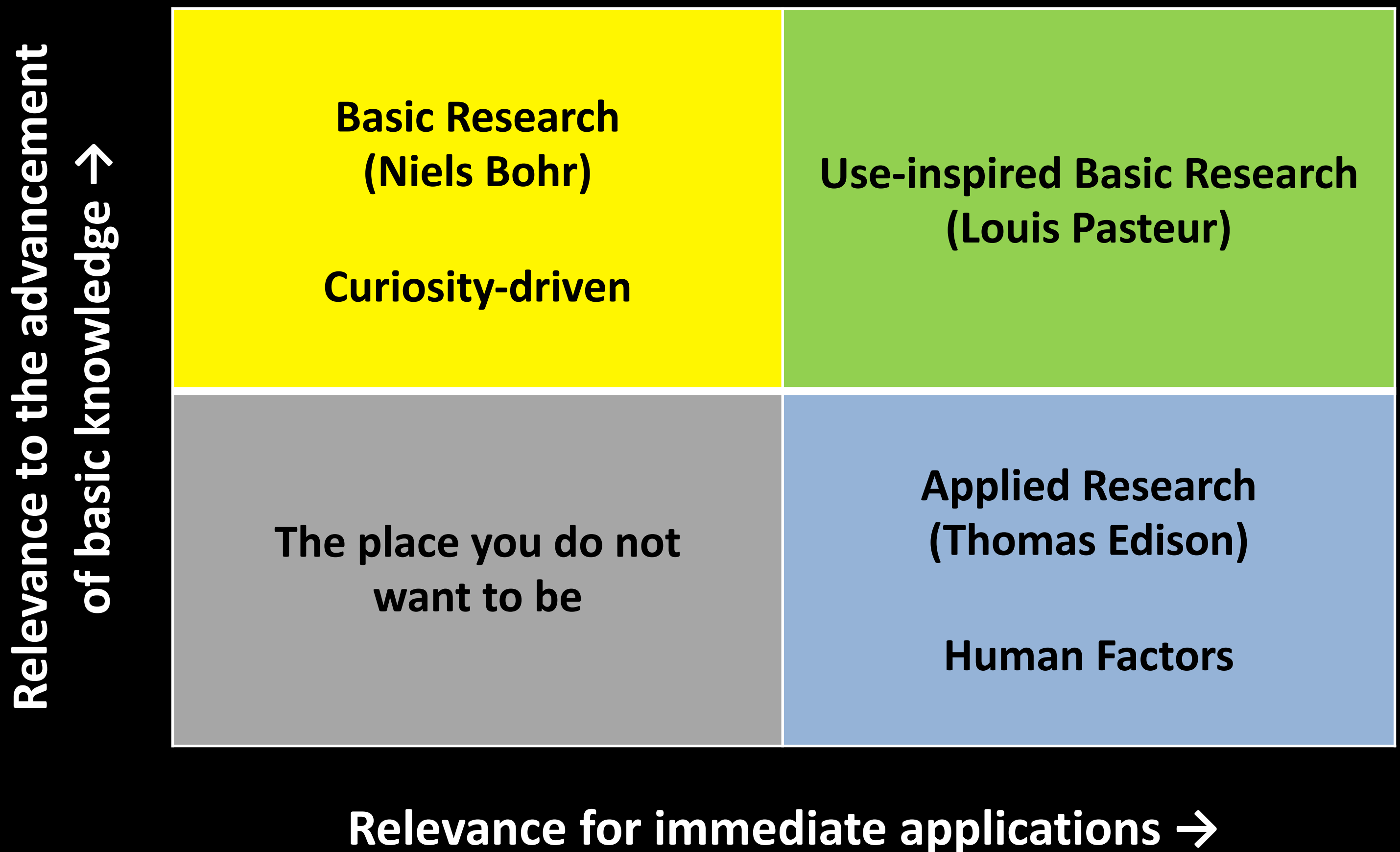




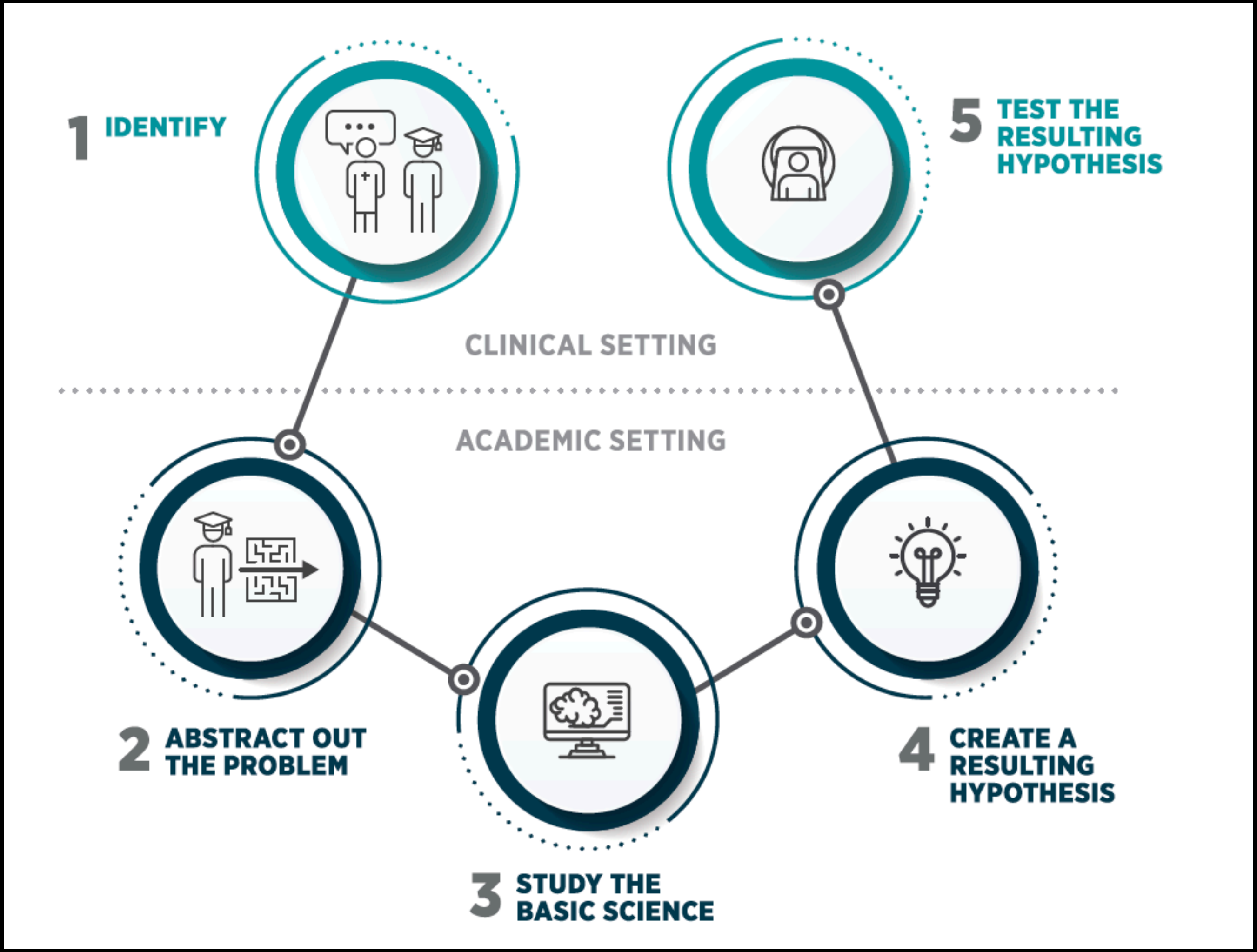
Challenges with Studying Clinical Questions

- Cognitive and Perceptual Studies typically involve:
  - Nonexpert populations
  - Artificial tasks and stimuli
  - Controlled testing environment with few distractions
  - Low stakes settings
- In the Clinic:
  - Expert populations
  - Complex stimuli
  - Busy environment with many distractions
  - High stakes (life or death) settings

Use-inspired Basic Research



Reverse Translation



Treviño et al. (2021) *JNCI Cancer Spectrum*

Examples of Use-inspired Basic Research in Pathology

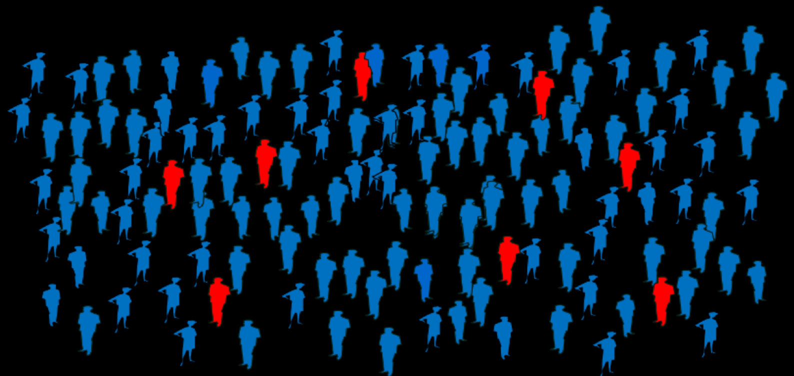
1. Understanding the external factors than can cause errors in Pathology image-based decisions
2. Developing strategies to reduce errors

Factors that affect Pathology Decisions

Two External Factors that can Influence Pathology Decisions

Prevalence

- When targets (abnormalities) are very rare or very common



Time Pressure

- Increasing work load demands

Blast Identification Task

- Distinguish between normal white blood cells and abnormal cancer cells (“blast” cells, associated with acute leukemia)

“Is this a blast cell?”

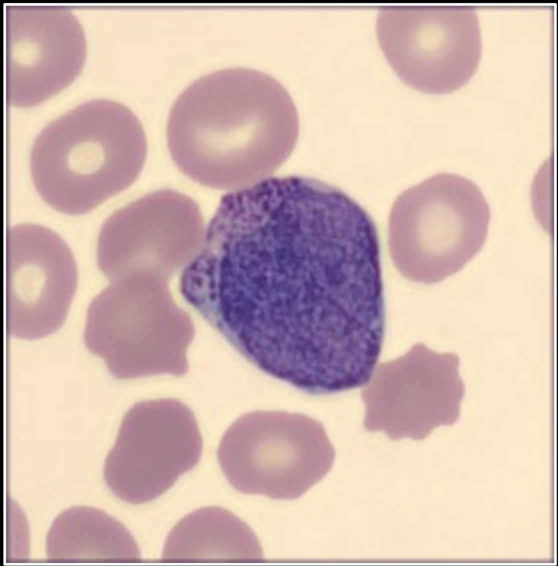


Image Curation

- Ratings Panel of three hematopathology faculty from VUMC
  - Identified each image as a blast or non-blast
  - Provided a rating of difficulty

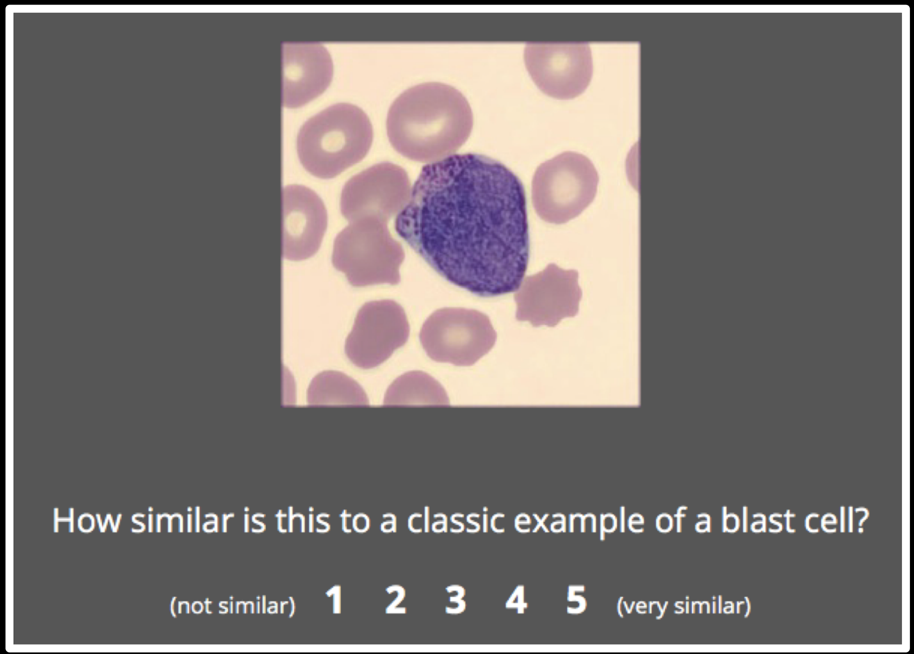
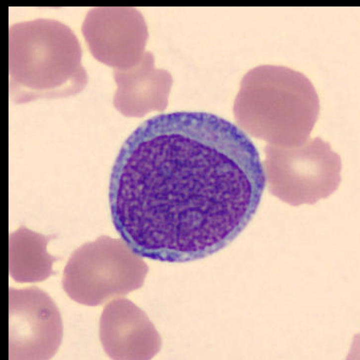


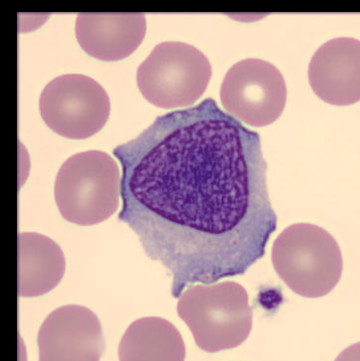
Image Categories



Blast Easy

Blast Hard

Non-blast Easy



Non-blast Hard

Prevalence

The Prevalence Effect in Medical Image Decision-making

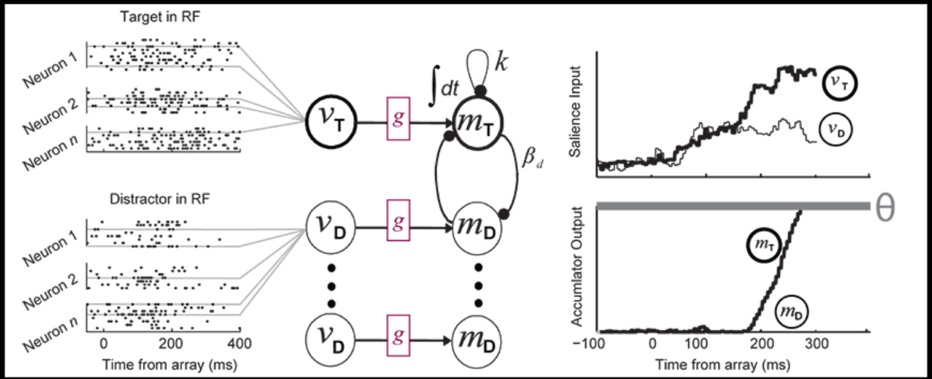
- Pathologists rarely see abnormal and normal cells at equal prevalence
- Extreme prevalence rates result in different types of errors (Wolfe & Van Wert, 2010; Horowitz, 2017)
  – Low prevalence → increase in misses
  – High prevalence → increase in false alarms

Why Does Prevalence Effect Occur?

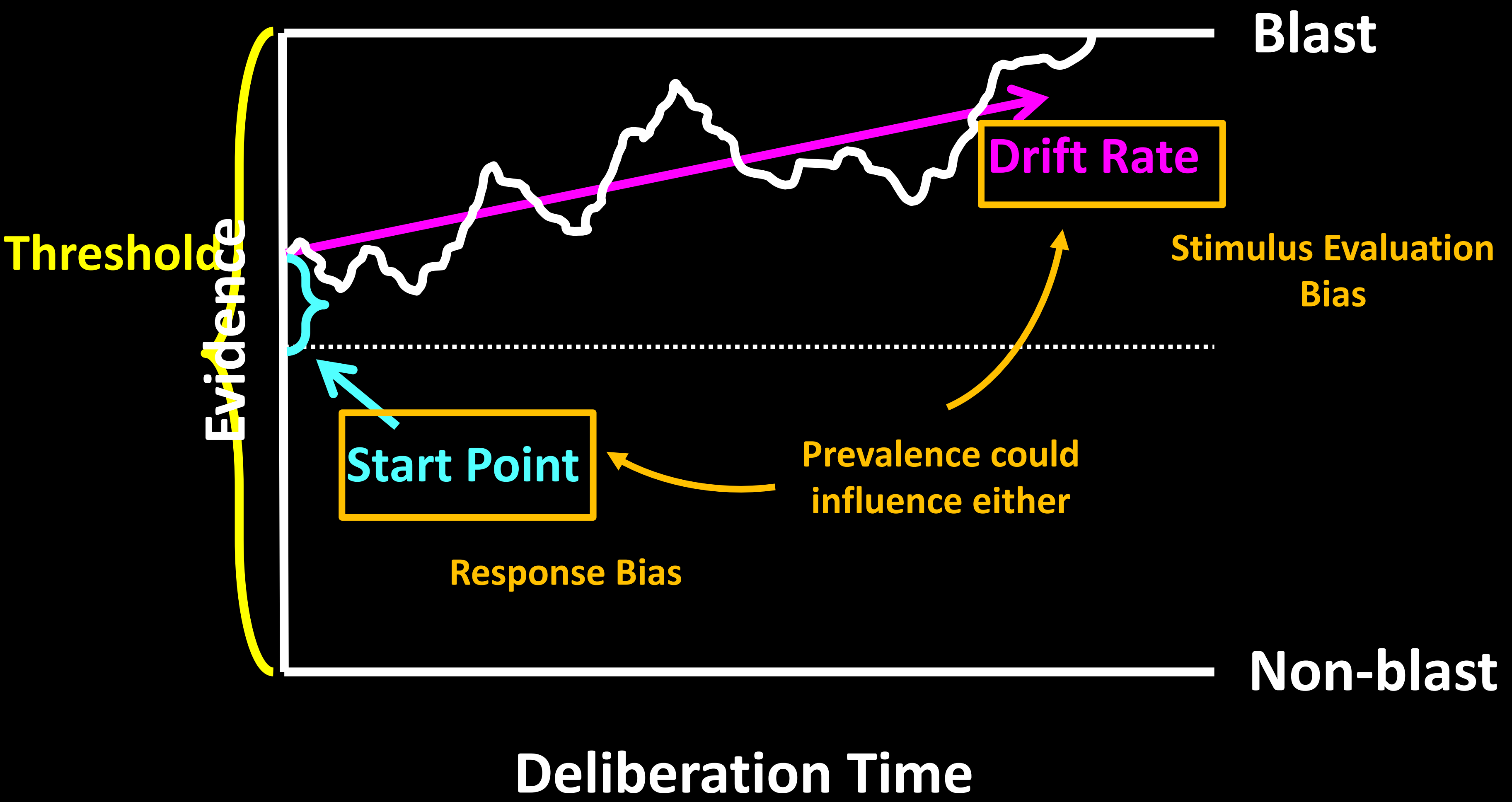
- Two possible cognitive biases:
  - Response bias
  - Stimulus evaluation bias
- Model both biases using Evidence Accumulation Models (EAMs)

Evidence Accumulation Models

- **Decisions are made by sequentially sampling information over time until an internal decision criterion is met**
- **Applied in almost every area of cognitive psychology: memory, perception, categorization, and decision-making**
- **Linked to neural processing in the brain**



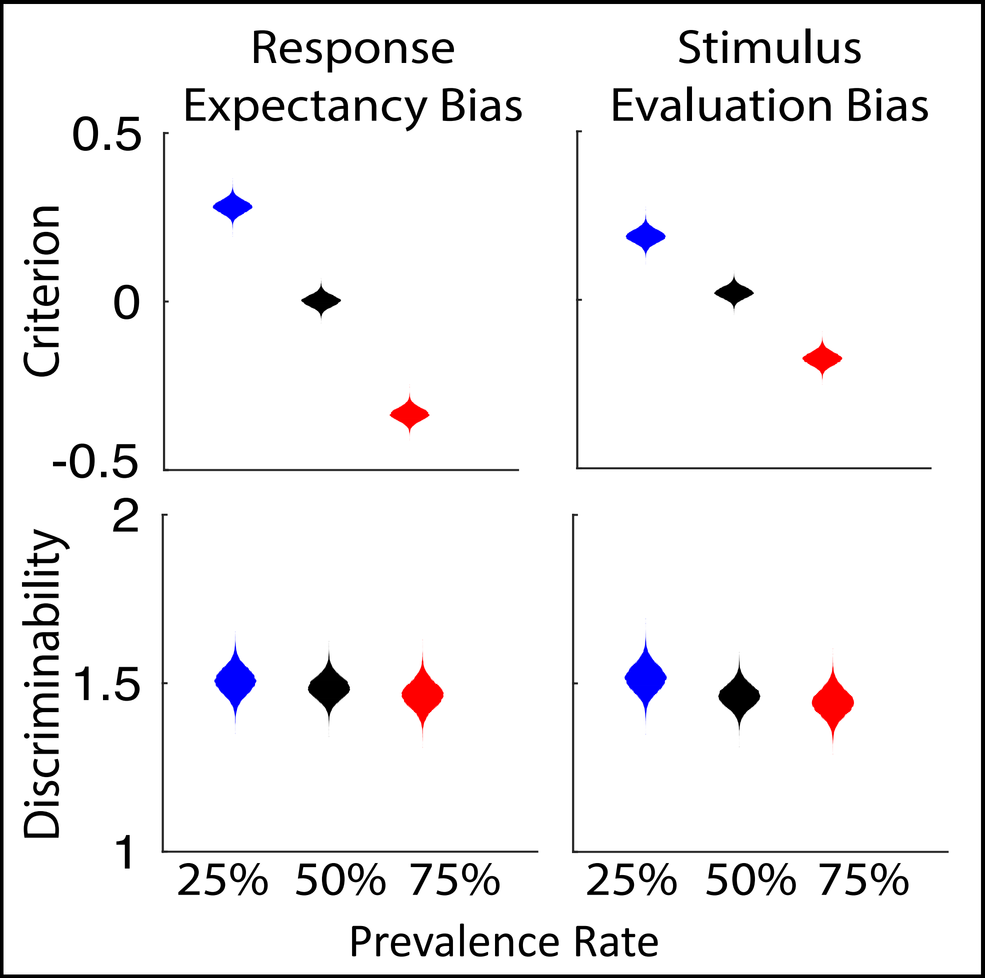
Diffusion Decision Model (DDM)



Ratcliff, 1978

Signal Detection Theory can't Distinguish between Biases

- **A response bias and stimulus evaluation bias both influence the criterion in SDT**
- **Simulated data from the DDM and fit with SDT**



Three Prevalence Studies

1. Novice: 25/50/75% prevalence
2. Novice: 10/50/90% prevalence
3. Expert: 50/90% prevalence

Prevalence: Experiment 1a

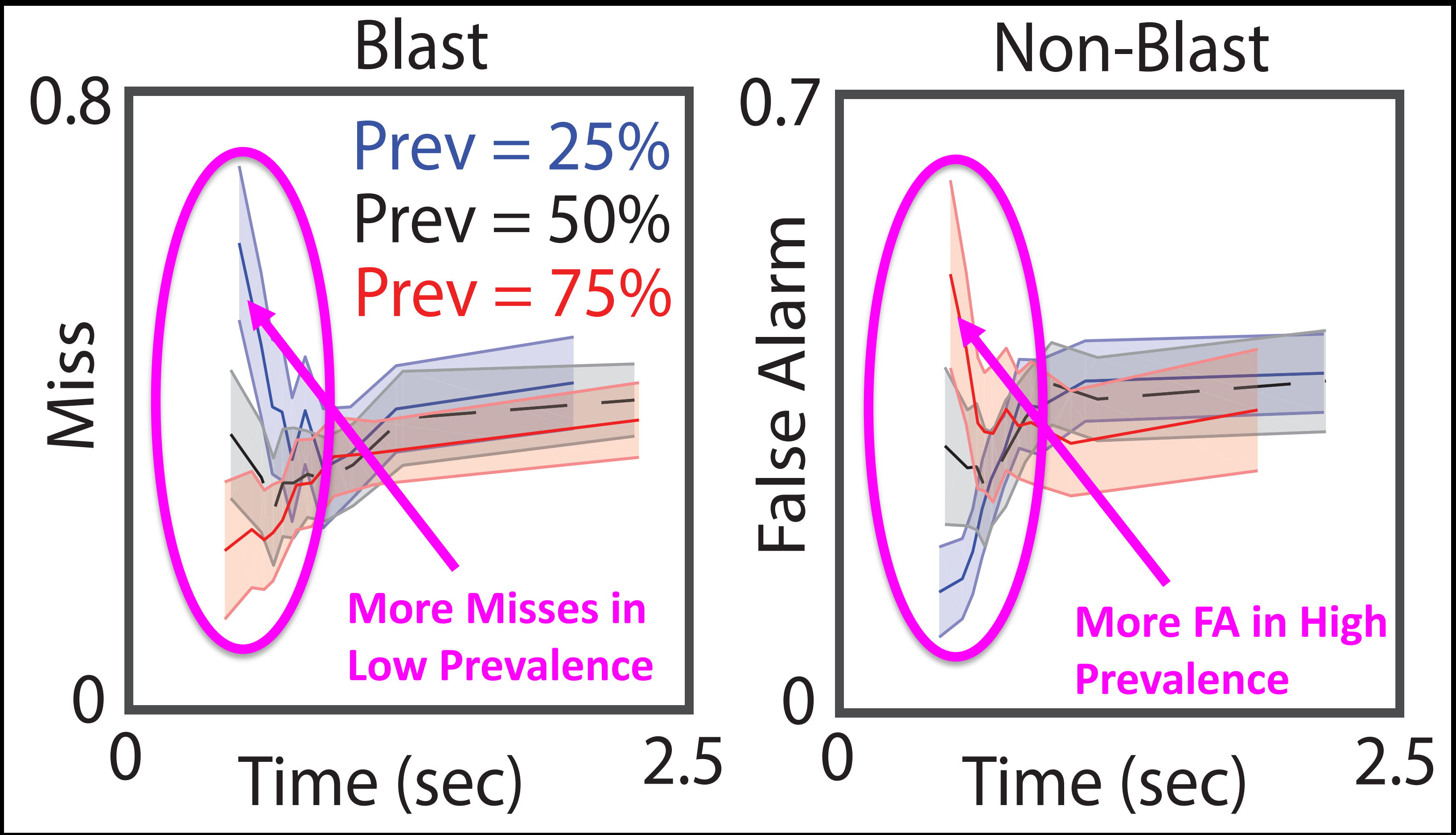
Novice: 25/50/75% prevalence (within-subjects)

- 39 VU undergrads
- Procedure
  1. **Learning phase**: single image + label
  2. **Training phase**: select the image that matches the label
  3. **Practice phase**: 3 blocks of 48 trials at each prevalence rate (25% blast, 50% blast, 75% blast)
  4. **Main task:** 21 blocks of 48 trials (7 blocks at each prevalence level)

Trueblood et al. (2021) *Cognition*

Results Exp 1a: Error Rates

Novice: 25/50/75% prevalence



Prevalence: Experiment 1b

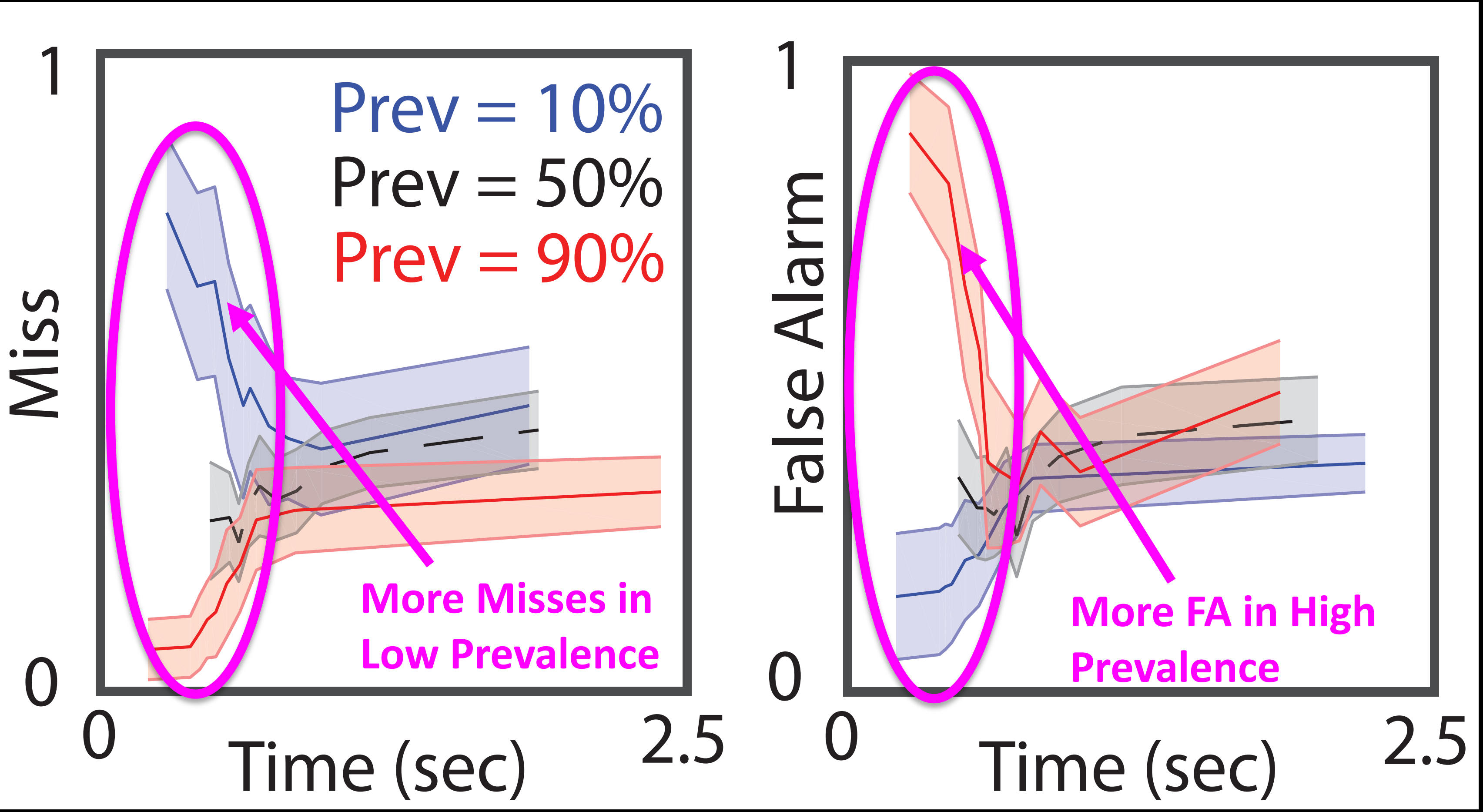
Novice: 10/50/90% prevalence (between-subjects)

- 57 VU undergrads
- Procedure
  1. **Learning phase**: single image + label
  2. **Training phase**: select the image that matches the label
  3. **Practice phase**: 1 block of 80 trials at 50%
  4. **Main task:**
     - 2 blocks of 80 trials at 50%
     - **High prevalence group**: 12 blocks of 80 trials at 90% prevalence
     - **Low prevalence group**: 12 blocks of 80 trials at 10% prevalence

Trueblood et al. (2021) *Cognition*

Results Exp 1b: Error Rates

Novice: 10/50/90% prevalence



Prevalence: Experiment 2

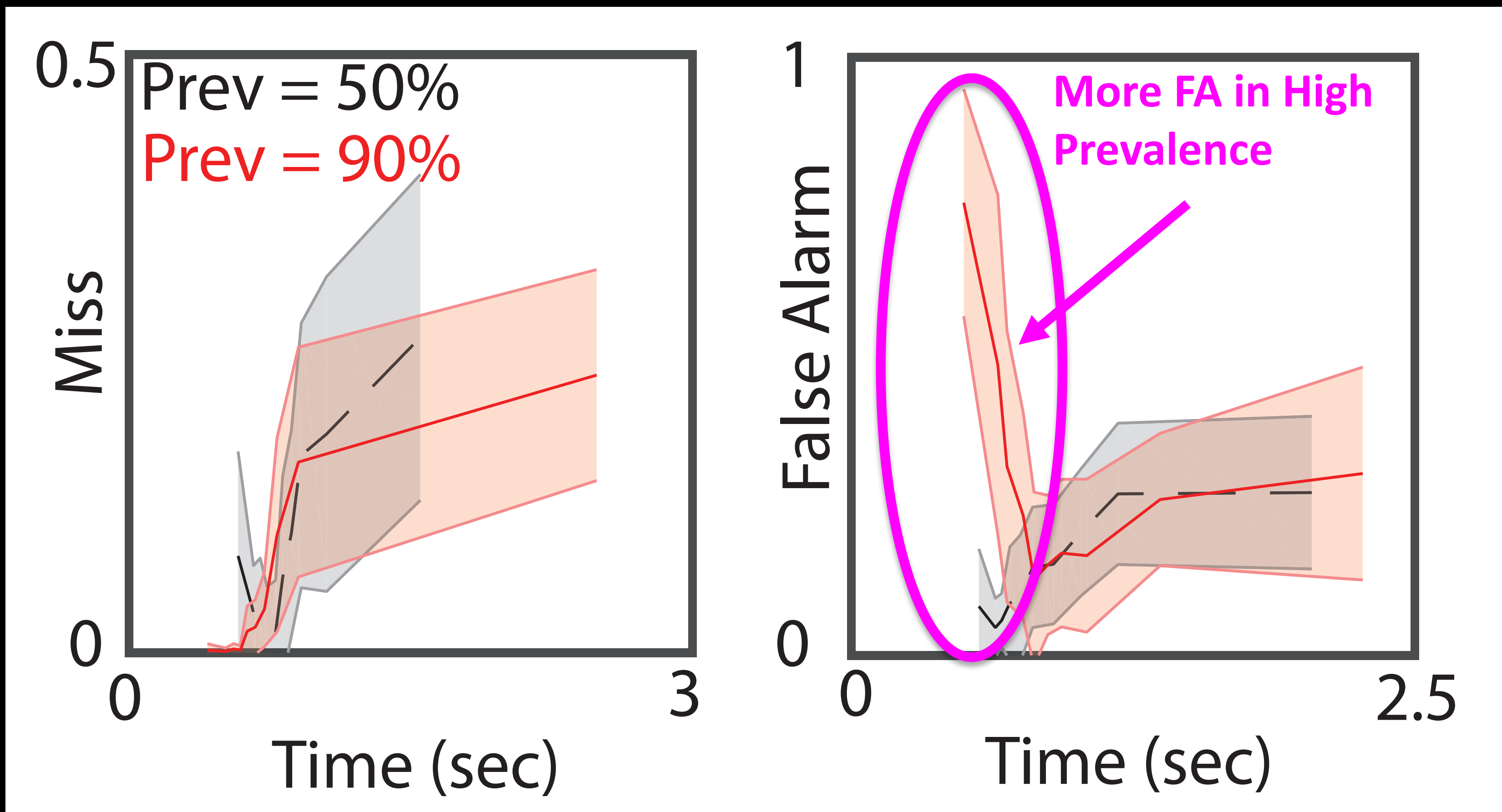
Expert: 50/90% prevalence

- 19 medical laboratory professional from Vanderbilt Medical Center
- Procedure
  1. **Same training** as Experiment 1 (no learning)
  2. **Practice phase:** 1 block of 40 trials at 50%
  3. **Main task:**
     - 2 blocks of 80 trials at 50%
     - 8 blocks of 80 trials at 90% prevalence

Trueblood et al. (2021) *Cognition*

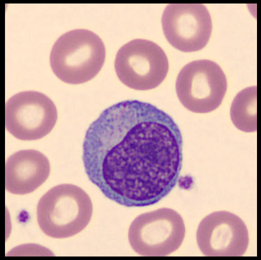
Results Exp 2: Error Rates

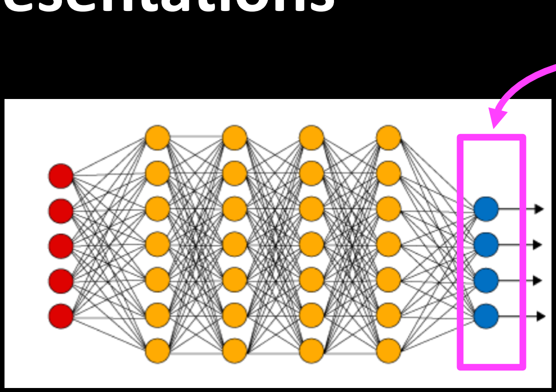
Expert: 50/90% prevalence



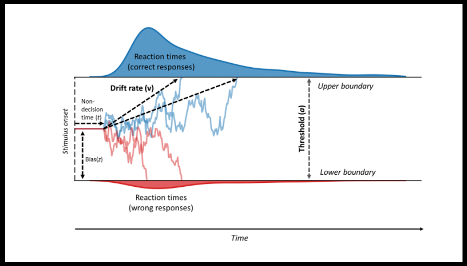
Challenges with "Naturalistic Images"

- Naturalistic stimuli typically have latent features
  - Problem: Cognitive models require numeric representations of stimuli
  - Solution: Use machine learning tools to generate numeric representations


Stimulus



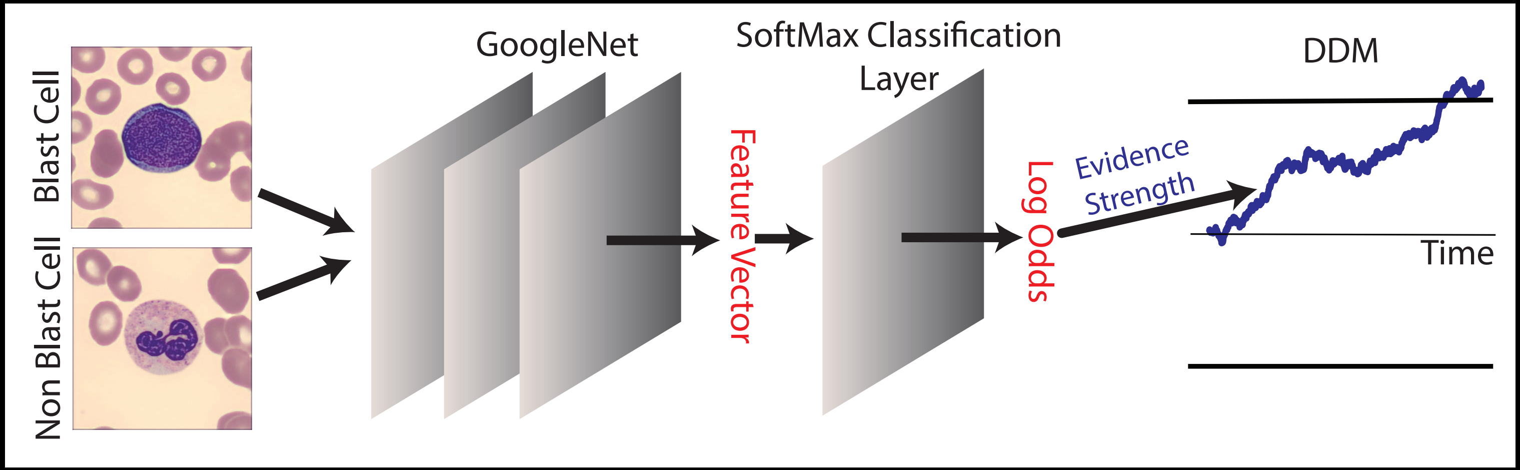
Machine Learning Model



Cognitive Model

Sanders & Nosofsky, 2018, 2020; Battleday et al., 2020; Holmes, O'Daniels, & Trueblood, 2020

Convolutional Neural Net + DDM



$$d_i = \boxed{u} + v * o_i$$

Drift rate for image i

Varies with prevalence rate

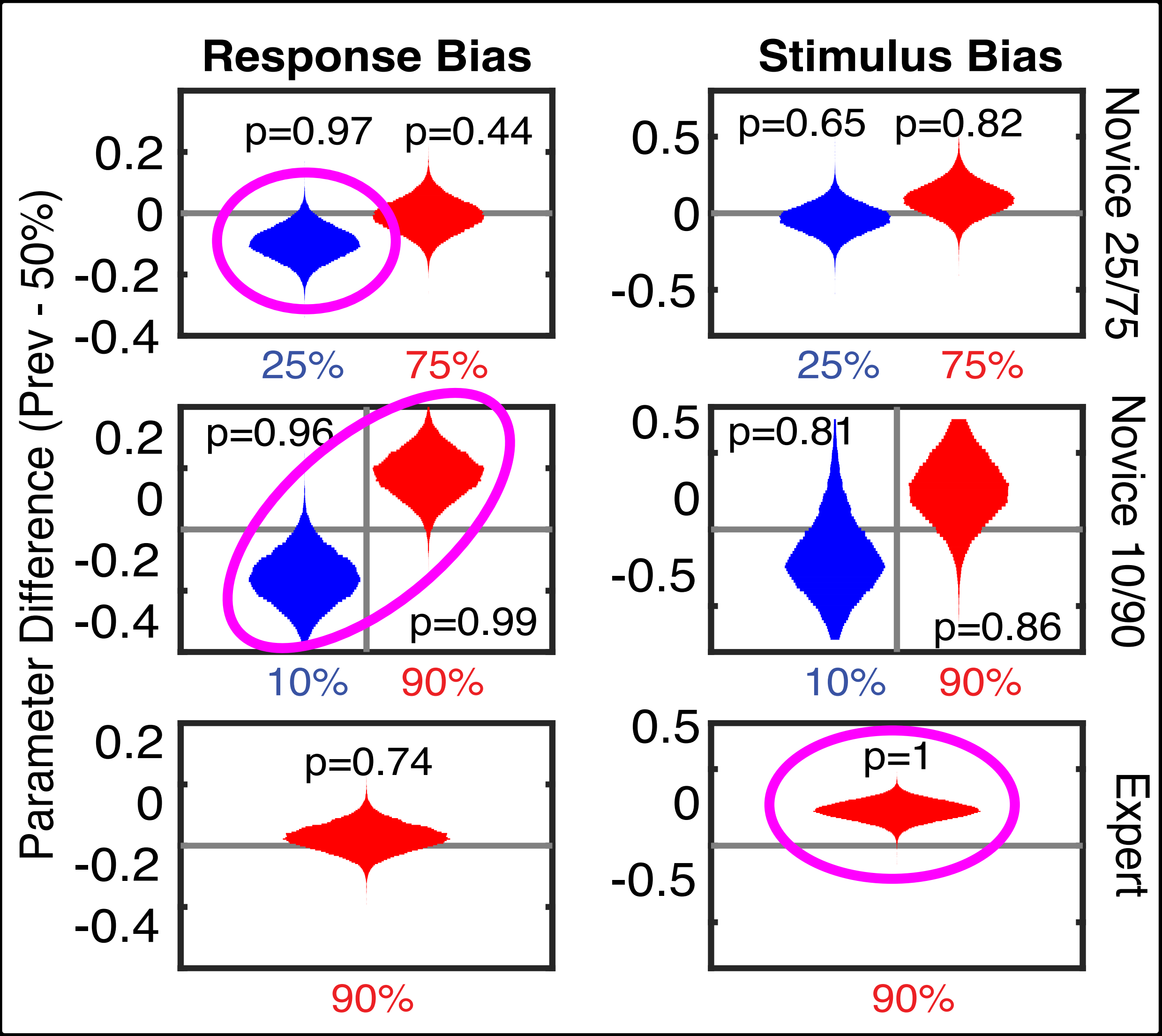
Stimulus Evaluation bias (White & Poldrack, 2014)

Weight on log odds

Log odds from convolutional neural net for image i

Holmes et al. (2020) *Computational Brain & Behavior*

CNN + DDM Modeling Results



Interim Conclusions

- Prevalence influences novices and experts differently
- A strong response bias in novices suggest a strategy of responding more often for the high base-rate category
- A strong stimulus bias in experts suggest that the evaluation of cell images changes with the base-rate

Time Pressure

Time Pressure

- Time pressure can lead to a speed / accuracy tradeoff
- In Pathology, time pressure occurs because of
  - Current and projected shortages of medical image observers
  - Increases in workload due to the introduction of AI (e.g., FDA increase in workload of cytotechnologists from 100 to 200 slides per day if using ThinPrep)

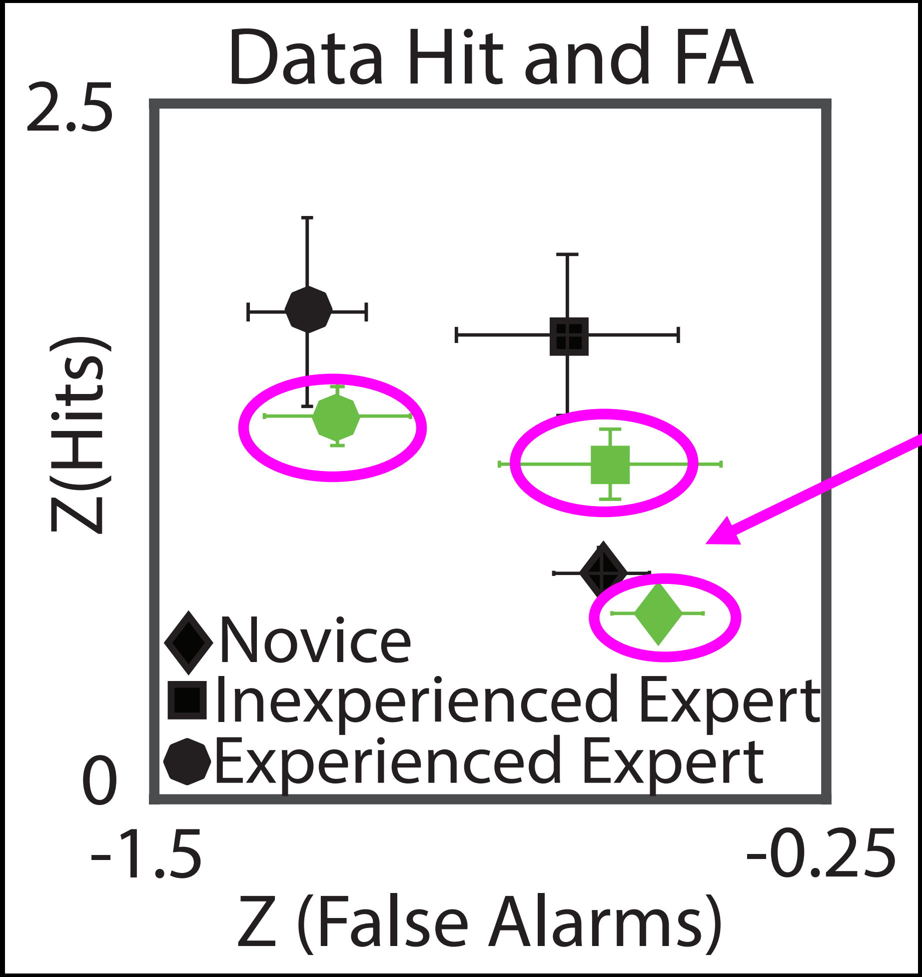
Time Pressure Studies

1. **Novice**
   - 35 VU undergrads
   - Within-subjects: different blocks for time pressure / no time pressure
     - No time pressure: instructed to be as accurate as possible
     - Time pressure: only 1 second to respond

2. **Expert**
   - 18 pathologists from VUMC (ranging from first year residents to faculty members)
     - 8 participants who had completed all four mandatory hematopathology rotations
     - 10 participants who had not
   - Same time pressure conditions as novices

Trueblood et al., 2018, CR:PI

Behavioral Results Time Pressure

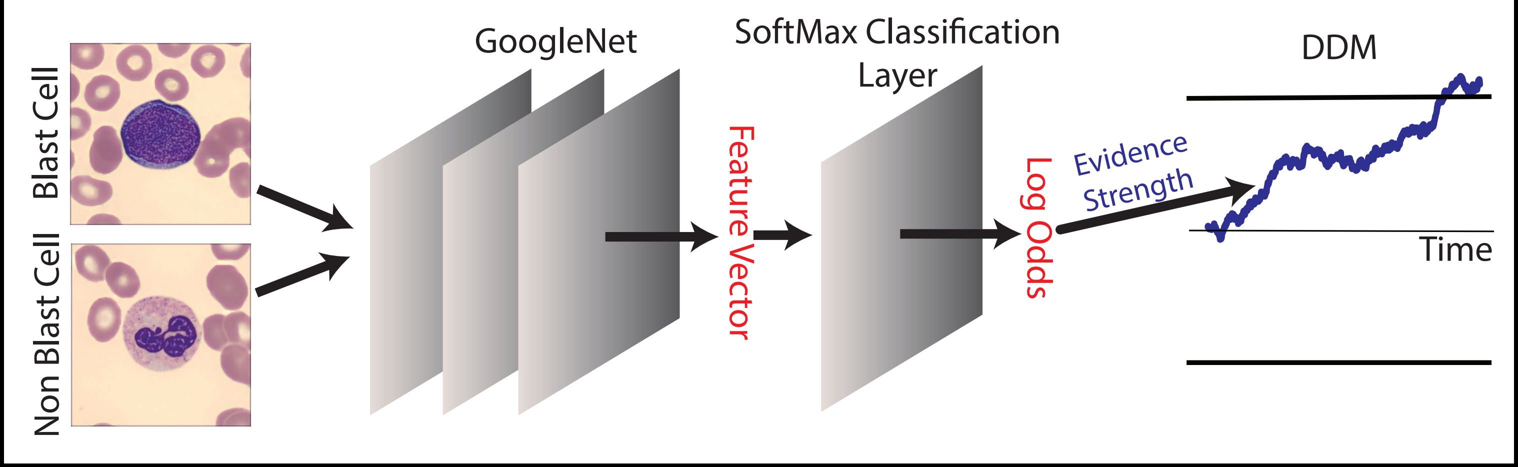


Both experts and novices are worse under time pressure

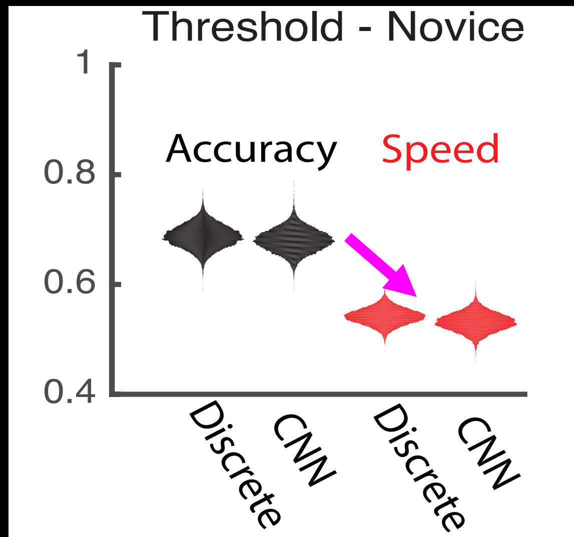
Cognitive Modeling

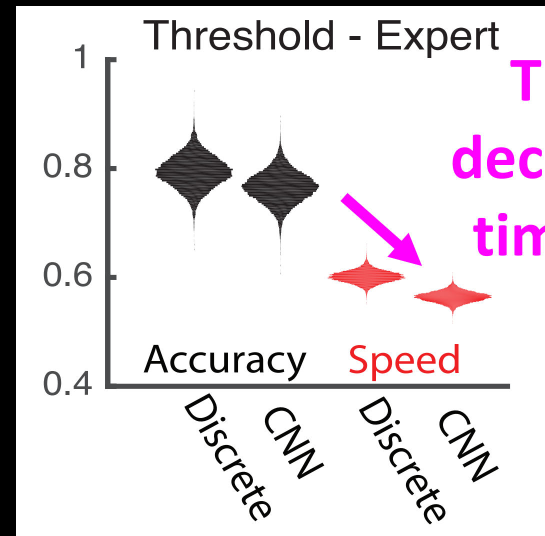
Fit two versions of the DDM:

1. DDM with a separate drift rate for each of the four image categories (images in the same category are treated the same)
2. CNN + DDM with a different drift rate for each image



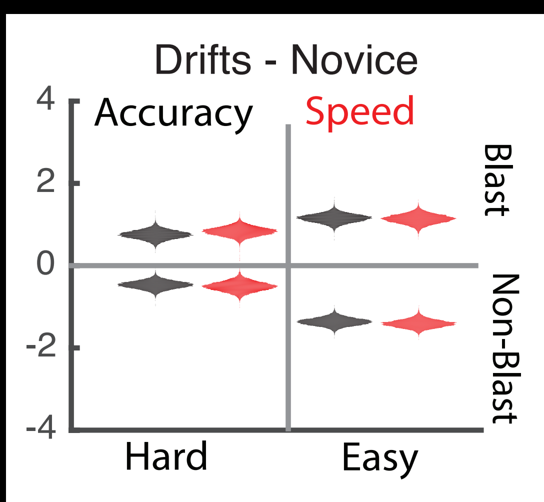
Modeling Results

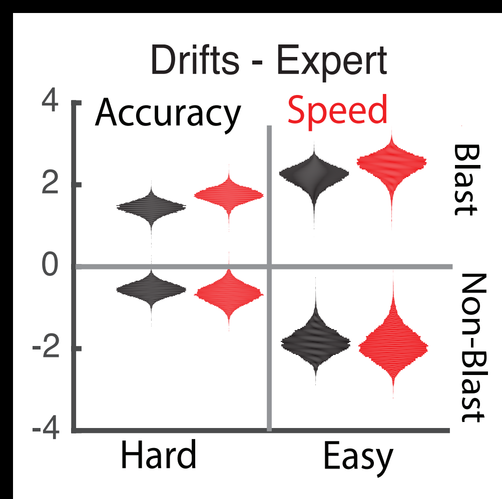




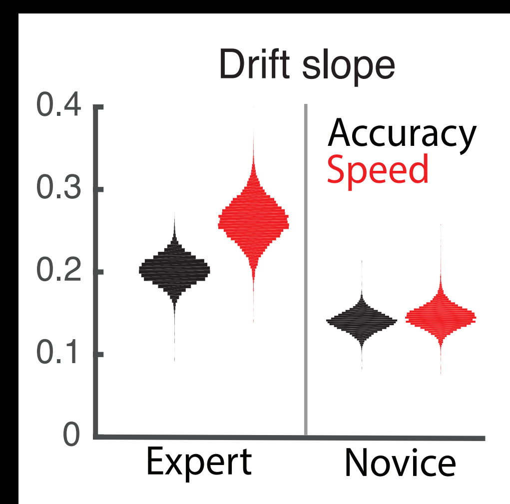
Thresholds decrease under time pressure

Discrete Model

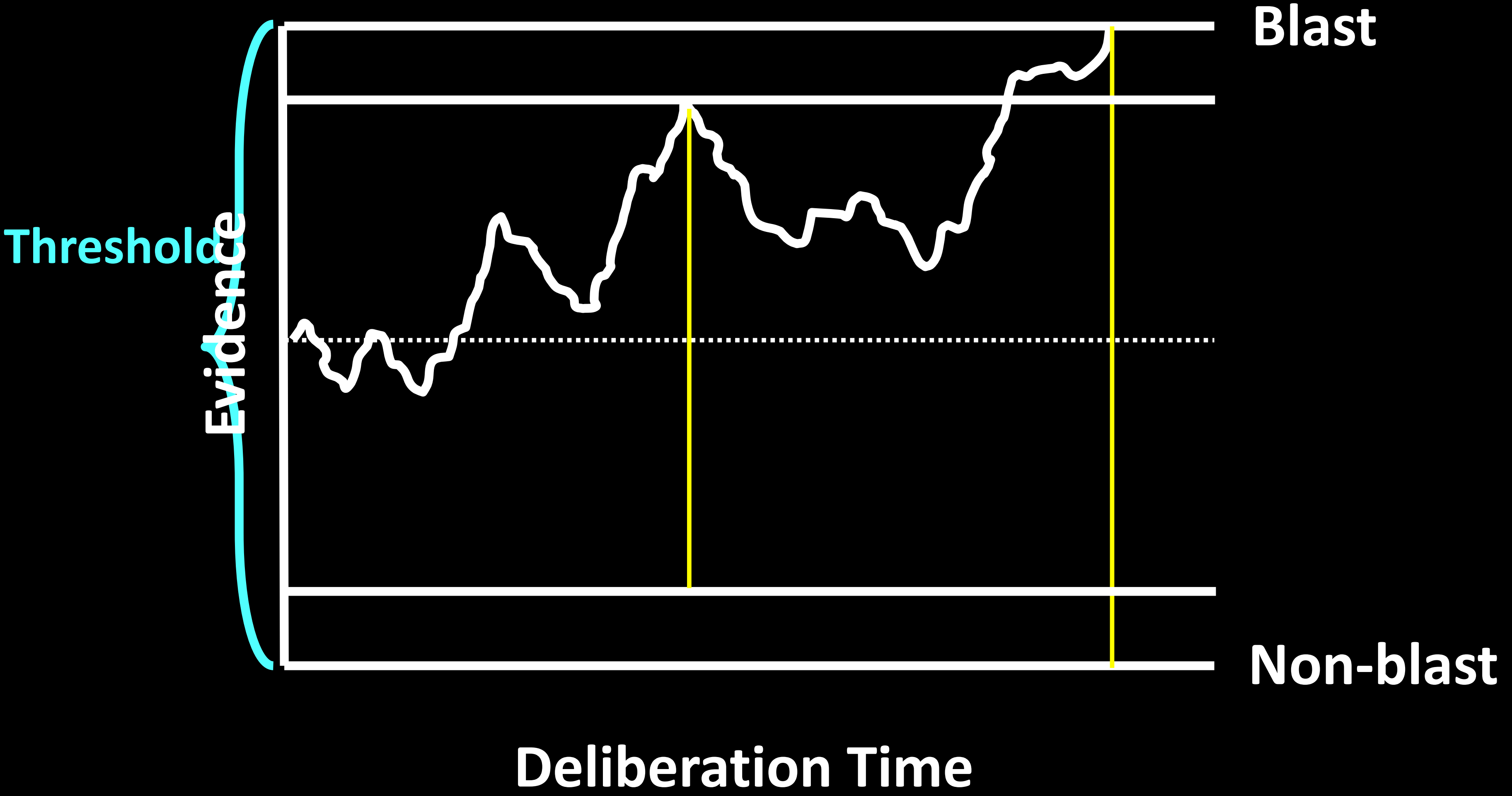




CNN Model



Key Result: Time Pressure Reduces Caution



Interim Summary

- Experts and novices are *similarly* influenced by time pressure
  - Reduced response caution under time pressure

- However, prevalence impacts experts and novices differently
- Critically important to study both populations
  - All expert medical image observers are novices at some point
  - Implications for training and error migration strategies

Strategies to Reduce Errors

Two Approaches to Reducing Errors

Simple Techniques to Improve Performance

- Wisdom of the Crowd Within

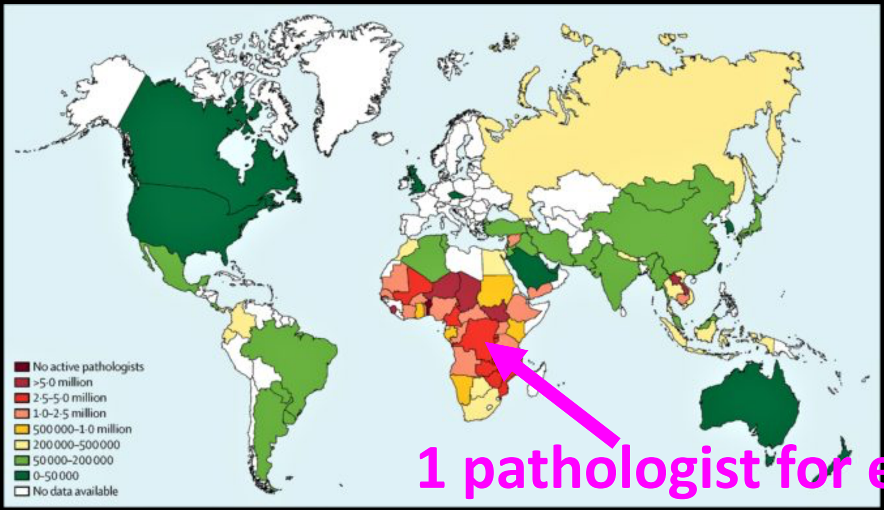
Using AI to assist Humans

- First step is to train medical AI systems
- Strategies for generating labeled image sets

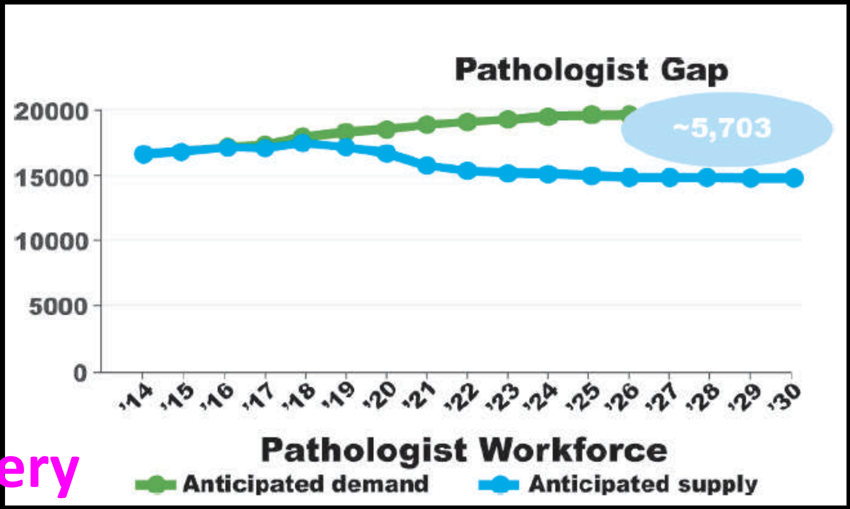
Wisdom of the Crowd Within

Double Readings

- Second opinions can significantly improve diagnostic accuracy
  - Misclassification rate decreased from 24.7% to 18.1% in breast histopathology (Elmore et al., 2016)
- But, multiple readings are not always possible



1 pathologist for every million people



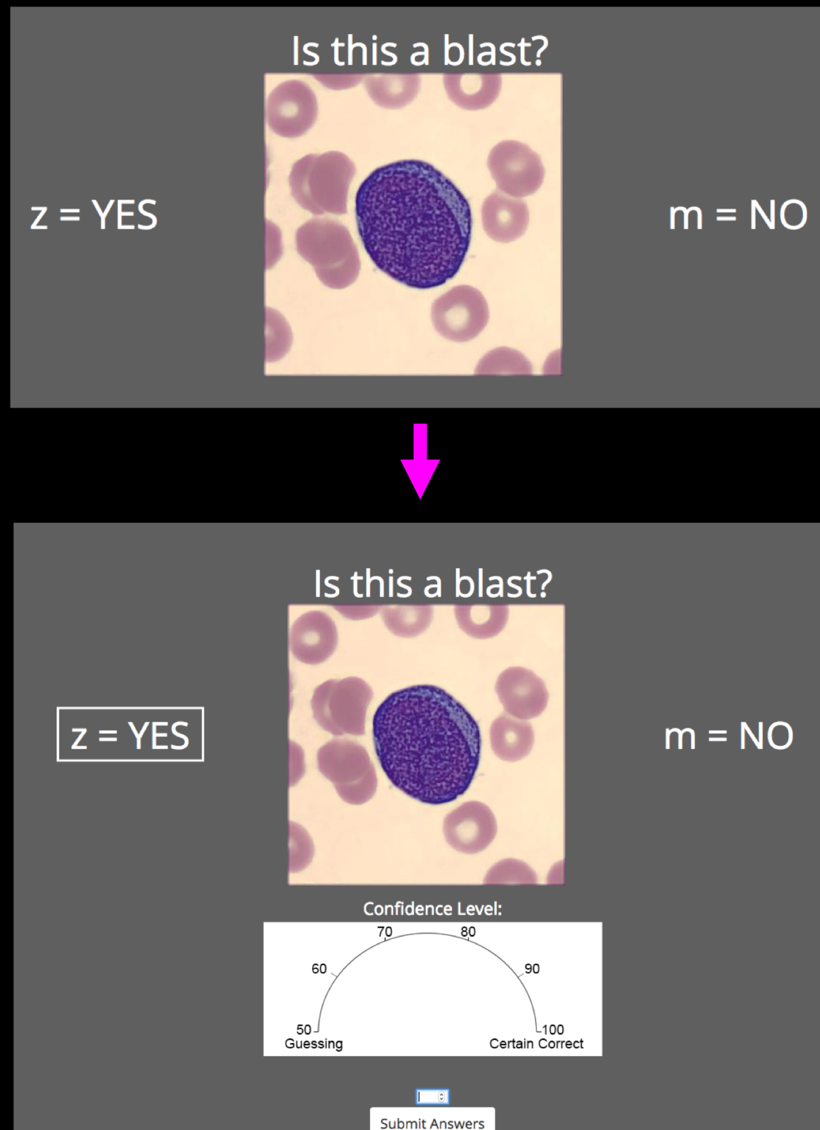
Can we reduce errors by having the same person do multiple readings?

- "Wisdom of the crowd within" (Vul & Pashler, 2008; Herzog & Hertwig, 2009)
- Consider the opposite technique (Lord et al., 1984; Hirt & Markman, 1995)
  - Example (Soll & Klayman, 2004):

    "I am 90% sure that Oscar Wilde was born *after*…"

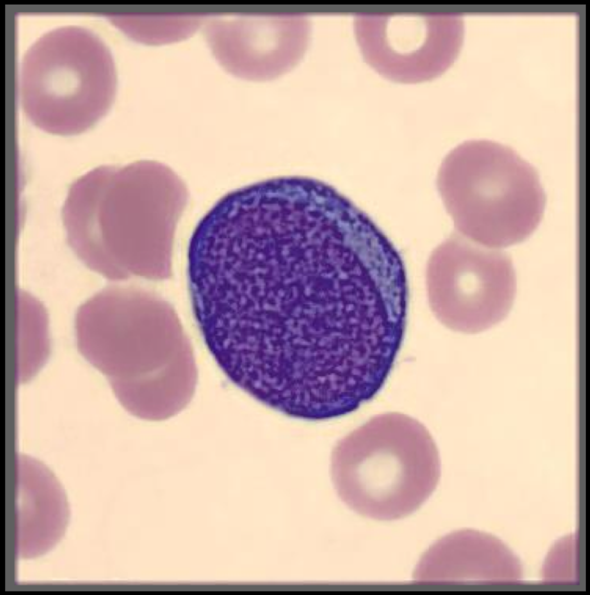
    "I am 90% sure that Oscar Wilde was born *before*…"

Experimental Task

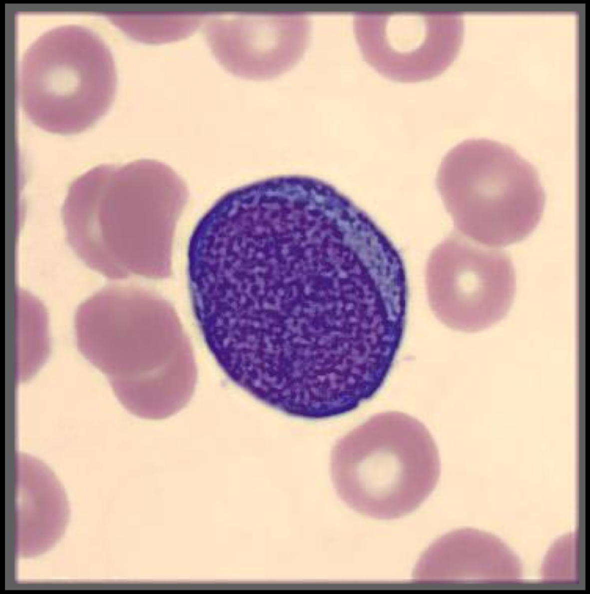


Implementing "Consider the Opposite"

First Presentation:
Is this a blast?

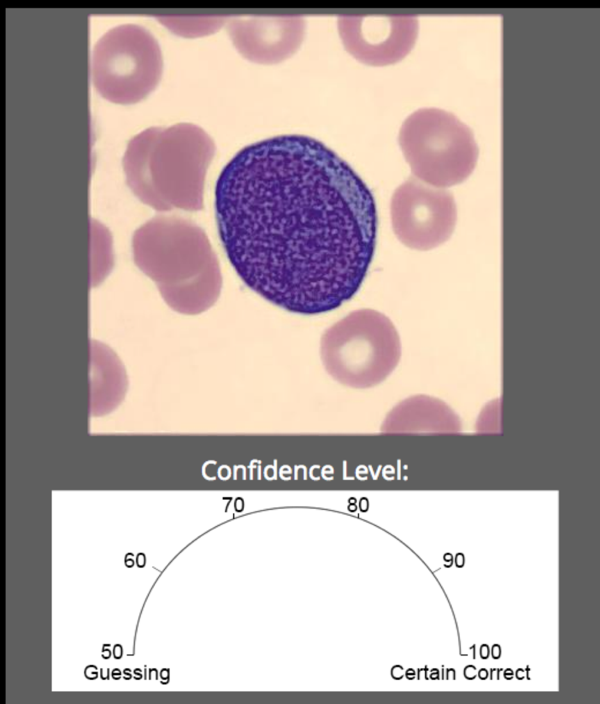


Second Presentation:
Is this a non-blast?



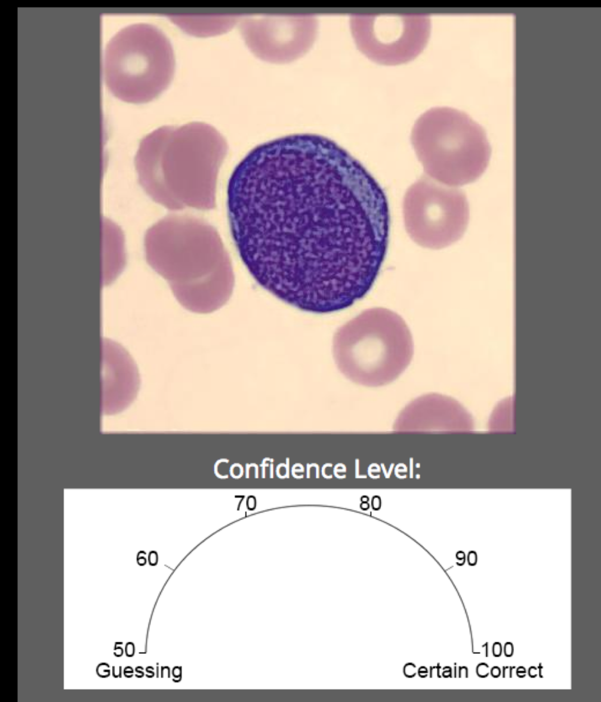
Aggregating Responses: Max Confidence Slating

Is this a blast?



Response: Yes
Confidence: 70

Is this a non-blast?



Response: Yes
Confidence: 60

Maximum Confidence Slating

Mismatch in response

Koriat, 2012; Bahrami, 2010

Confidence Slating Studies

1. Novice (two experiments)
   - 45 VU undergrads in Exp 1a; 42 in Exp 1b
   - 300 images viewed twice
   - “Is this a blast” blocks presented before “Is this a non-blast” blocks in Exp 1a
   - No change in prompt in Exp 1b

2. Expert
   - 23 pathologists and laboratory professionals recruited at the ASCP conference
   - 60 images viewed twice (only hard images)
   - “Is this a blast” blocks presented before “Is this a non-blast” blocks

Hasan et al. (2022) *TopiCS*

Results

Compared mean accuracy across both responses with the accuracy from Maximum Confidence Slating

Experiment	Average Accuracy	Max Conf. Slating	Difference
Exp 1a (novices)	66.1%	67.4%	1.3%*
Exp 1b (novices)	66.5%	67.4%	0.9%*
Exp 2 (experts)	71.6%	73.8%	2.2%*

$^*p < .01$

Hasan et al. (2022) *TopiCS*

Interim Summary

- It is possible to improve accuracy through multiple readings by the same individual
- Improvements were small
- Importantly, confidence is associated with accuracy
  - Could investigate other ways of using confidence to flag images for further review

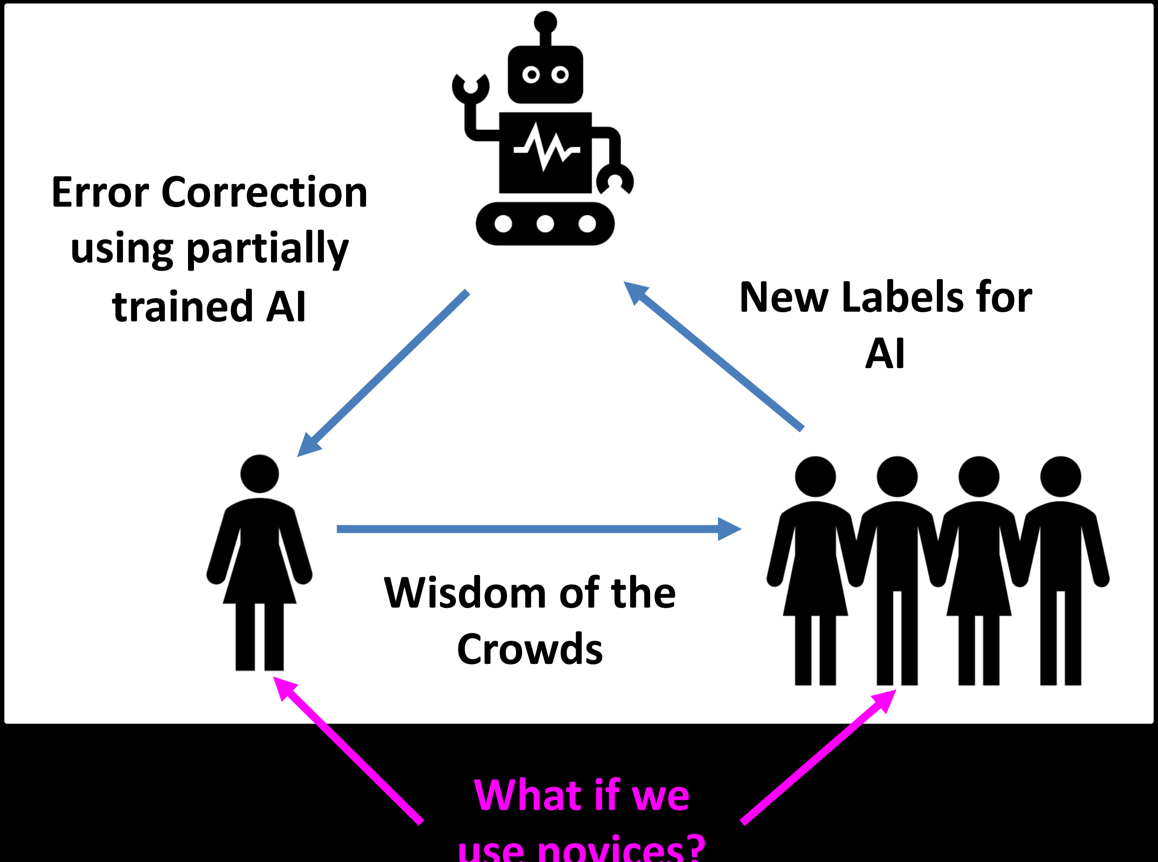
Humans Supporting Medical AI

Leveraging Human Decision-making to Support AI-Enabled Diagnosis

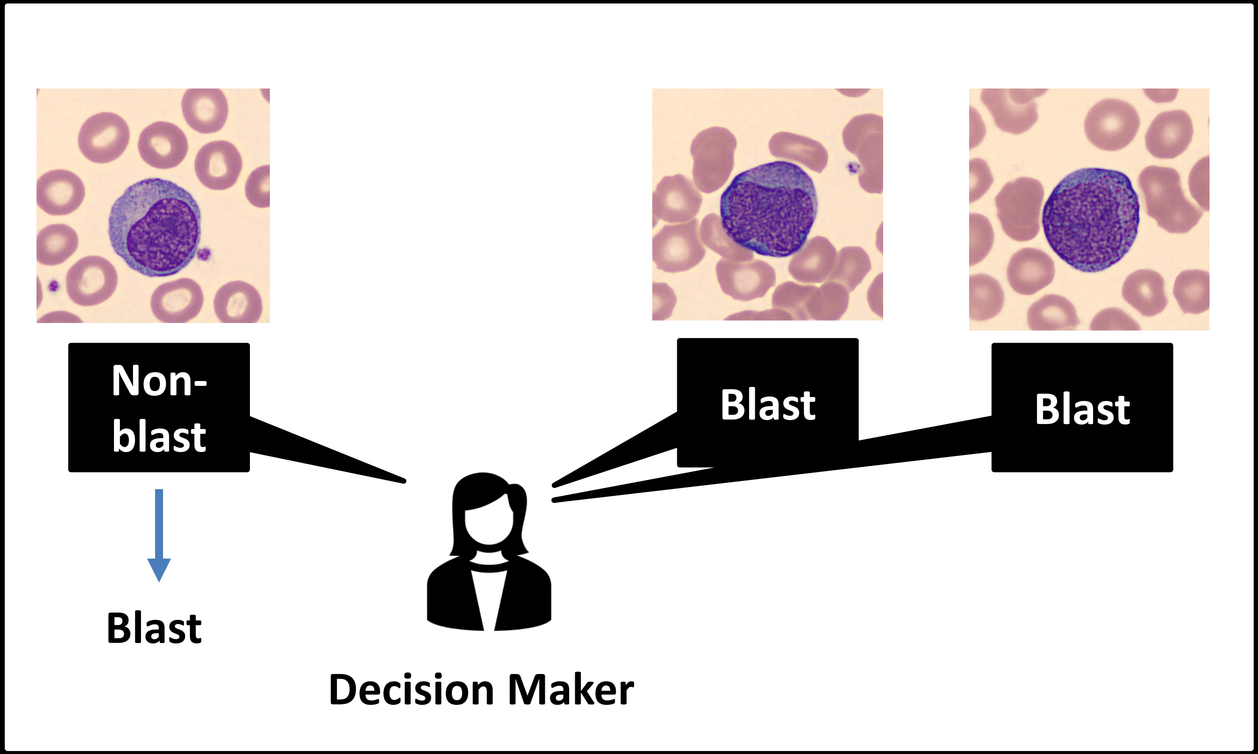
- In 2017, The FDA approved the first whole slide digital imaging system for pathology, opening the door to AI based diagnosis
- Medical AI is only as good as the data it's trained on
- Better Labels -> Better AI

How to generate accurately labeled images?

- Error correction at the individual-level + Wisdom of the Crowds



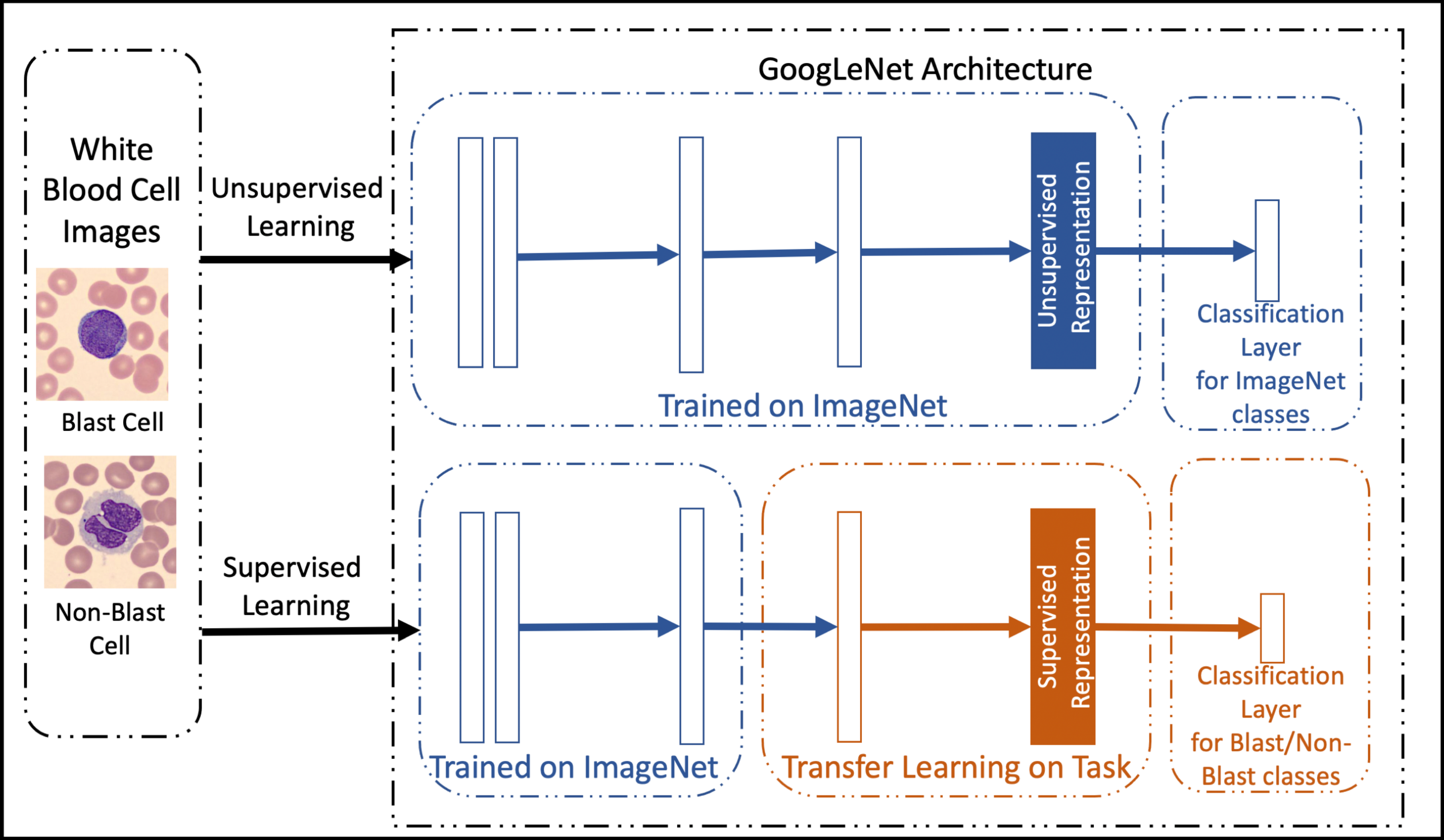
Similarity Based Error Correction



Two Questions:

1. How to determine similarity?
2. How many images to aggregate over?

CNN Based Similarity Representations



Hasan et al. (2022) *TopiCS*

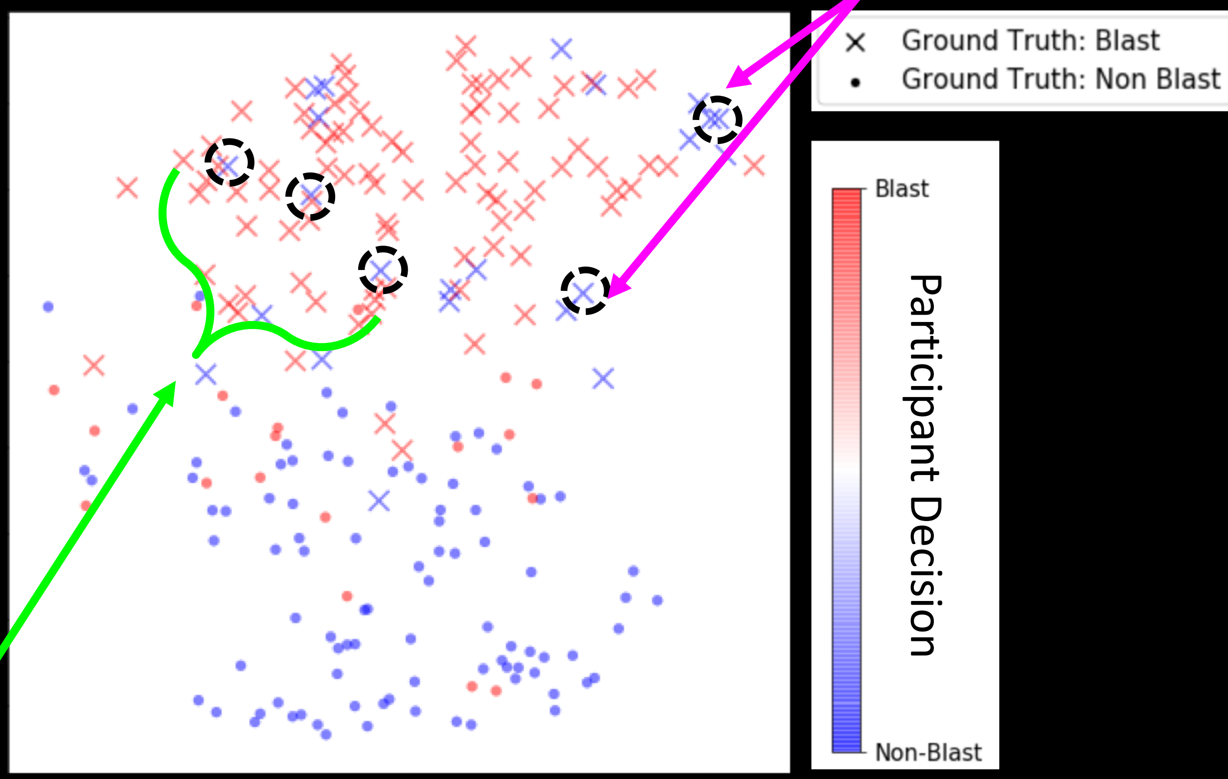
Application to Prevalence Experiments

Three prevalence studies (Trueblood et al., 2021):

- Novice: 25/50/75% prevalence
- Novice: 10/50/90% prevalence
- Expert: 50/90% prevalence

Error Correction with Supervised Representation

Incorrectly labeled Non-blast unchanged



Incorrectly labeled Non-blast changed to Blast

Error Correction for Novices

	Experiment 1a			Experiment 1b		
Blast Prevalence	**50%**	**75%**	**25%**	**50%**	**90%**	**10%**
Average Accuracy	0.69	0.71	0.70	0.68	0.70	0.67
Supervised k=3	**0.71**	**0.75**	**0.72**	**0.70**	**0.72**	**0.68**
Supervised k=7	**0.74**	**0.77**	**0.74**	**0.72**	**0.73**	**0.70**
Unsupervised k=3	**0.64**	**0.68**	**0.70**	**0.63**	**0.68**	**0.64**
Unsupervised k=7	**0.63**	**0.67**	**0.65**	**0.62**	**0.67**	**0.64**

up to 6% increase in accuracy

Error Correction Experts

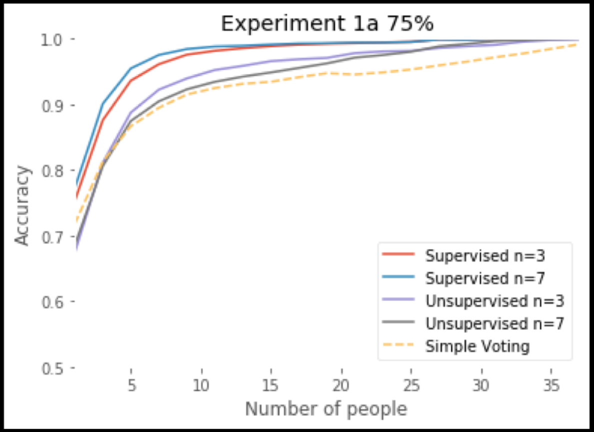
	Experiment 2	
Blast Prevalence	50%	90%
Average Accuracy	0.90	0.88
Supervised k=3	0.86	0.89
Supervised k=7	0.84	0.89
Unsupervised k=3	0.76	0.81
Unsupervised k=7	0.71	0.77

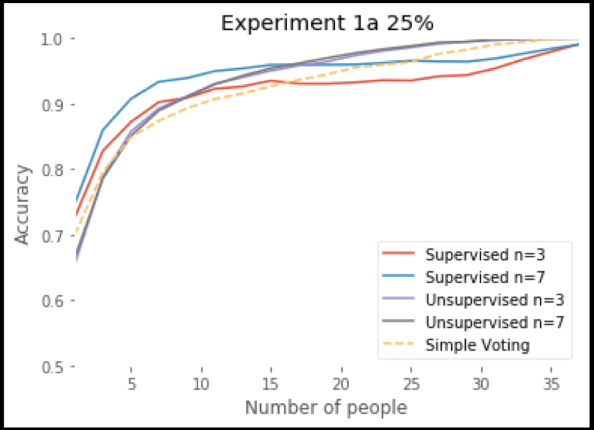
Why does it work for novices, but not experts?

- **Novices exhibit a response bias, possibly leading to more random errors**
- **Experts show a stimulus bias, suggesting systematic biases in image evaluation**

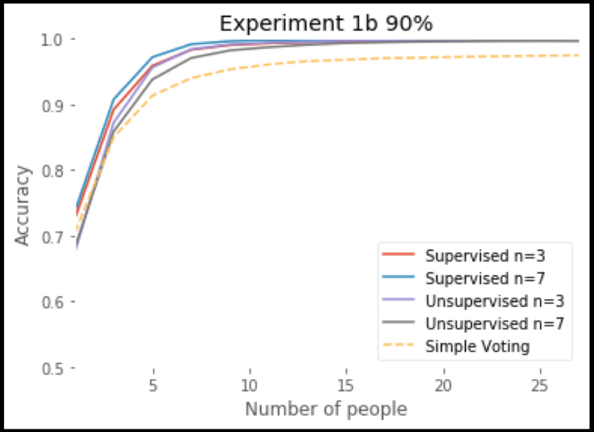
Wisdom of the Crowds

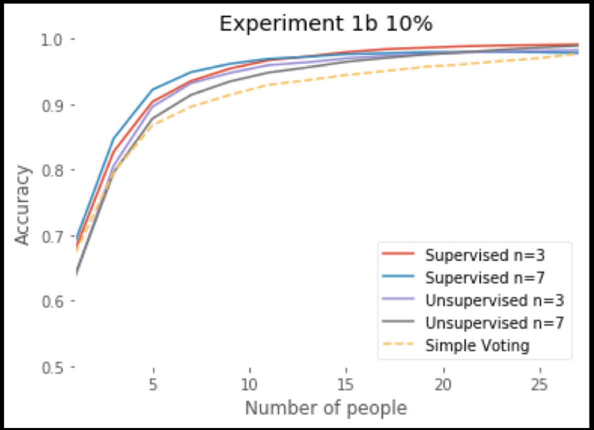
Exp 1a (Novice)



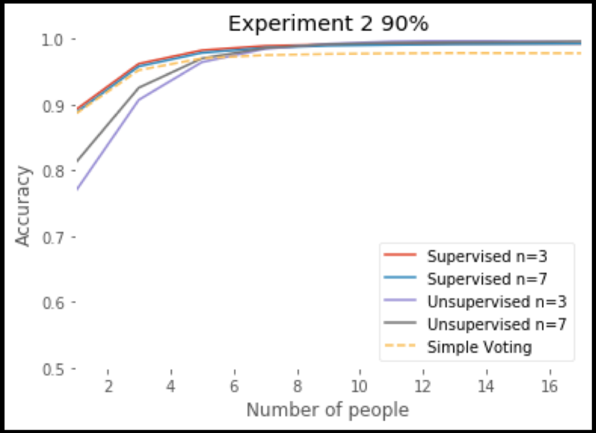


Exp 1b (Novice)





Exp 2 (Expert)



Almost perfect accuracy with a group size of about 10

Group Size

Interim Summary

- **Dramatically improve accuracy in medical image decision-making through neural network based error correction and Wisdom of the Crowds**
- **Applications to building better AI-based diagnosis tools**
- **Differences in cognitive biases help explain variations in algorithm performance across experts and novices**

Research Gaps and Open Problems

Information Overload in the Digital Era

- Advances in imaging technology have resulted in more complex images to be analyzed
  - Shift from 2D images to 3D images in Radiology
  - Shift from glass slides to digital images in Pathology
- Nonimaging information has to be incorporated into the decision-making process
  - Health records
  - Genetic panels

Treviño et al. (2021) *JNCI Cancer Spectrum*

Workflow Difficulties

- Different cases might involve
  - Different organs
  - Different image modalities
- Many forms of distractions and interruptions
  - Texts
  - Email
  - Pager messages
  - Telephone calls
  - Colleagues and trainees dropping by

Treviño et al. (2021) *JNCI Cancer Spectrum*

Fatigue and Workload

- **Diagnosticians' workload has increased** (Bhargavan et al., 2009)

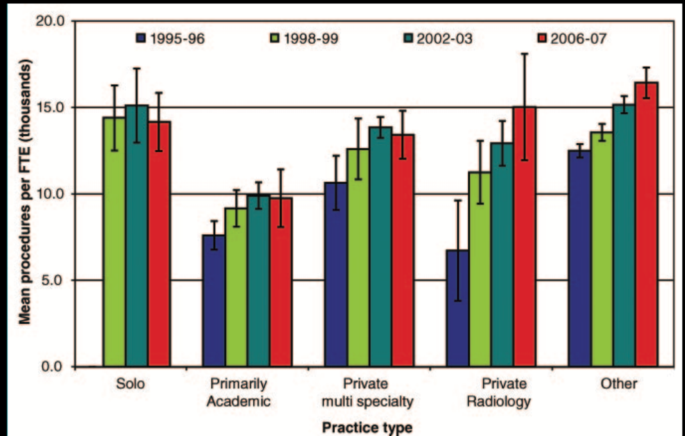


Figure 2: Bar graph shows trends in number of procedures per FTE radiologist according to practice type. Error bars = 95% confidence intervals.

- **Mental fatigue is at unprecedented levels**
- **Direct relationship between detection accuracy and fatigue**

Treviño et al. (2021) *JNCI Cancer Spectrum*

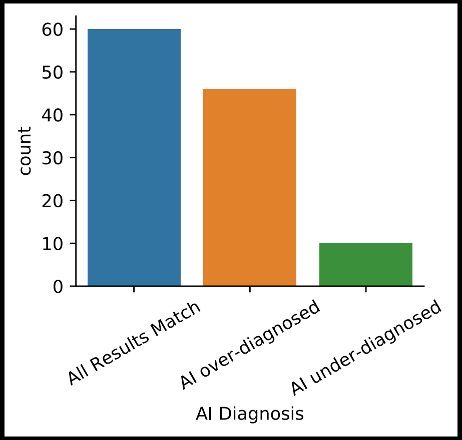
Artificial Intelligence

- AI algorithms that might perform well in a lab setting may not perform well in real-world clinical practice
- Diagnosticians may over-rely on AI or, conversely, learn to ignore them
- The impact of AI training biases on human observers

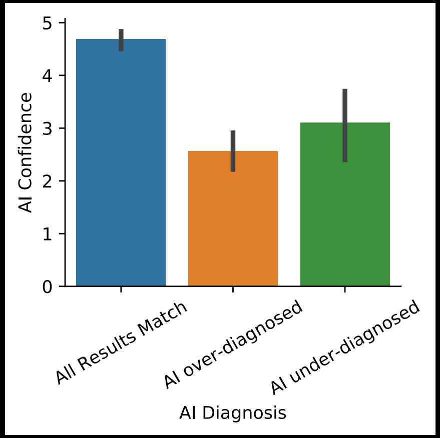
Treviño et al. (2021) *JNCI Cancer Spectrum*

Understanding the Interaction of Humans and Machines

- Examine AI-induced biases in medical image decision-making
- Collaborating with PathNet, a digital Pathology company



AI was more often seen as over-diagnosing as compared to under-diagnosing

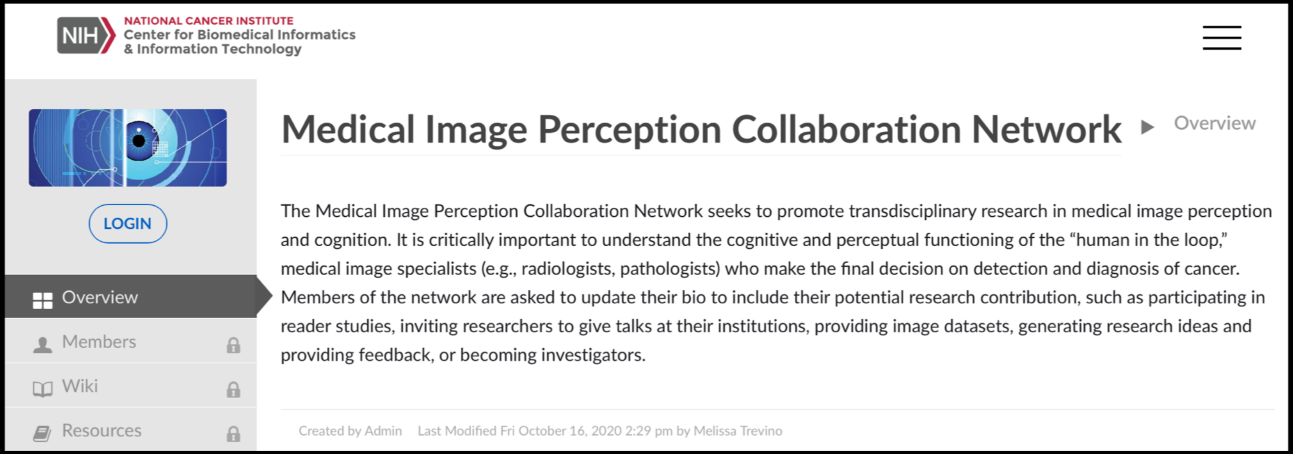


Confidence in the AI was lower when there was disparity in the diagnosis

Resources

Developing Collaborations

- **Medical Image Perception Collaboration Network:**

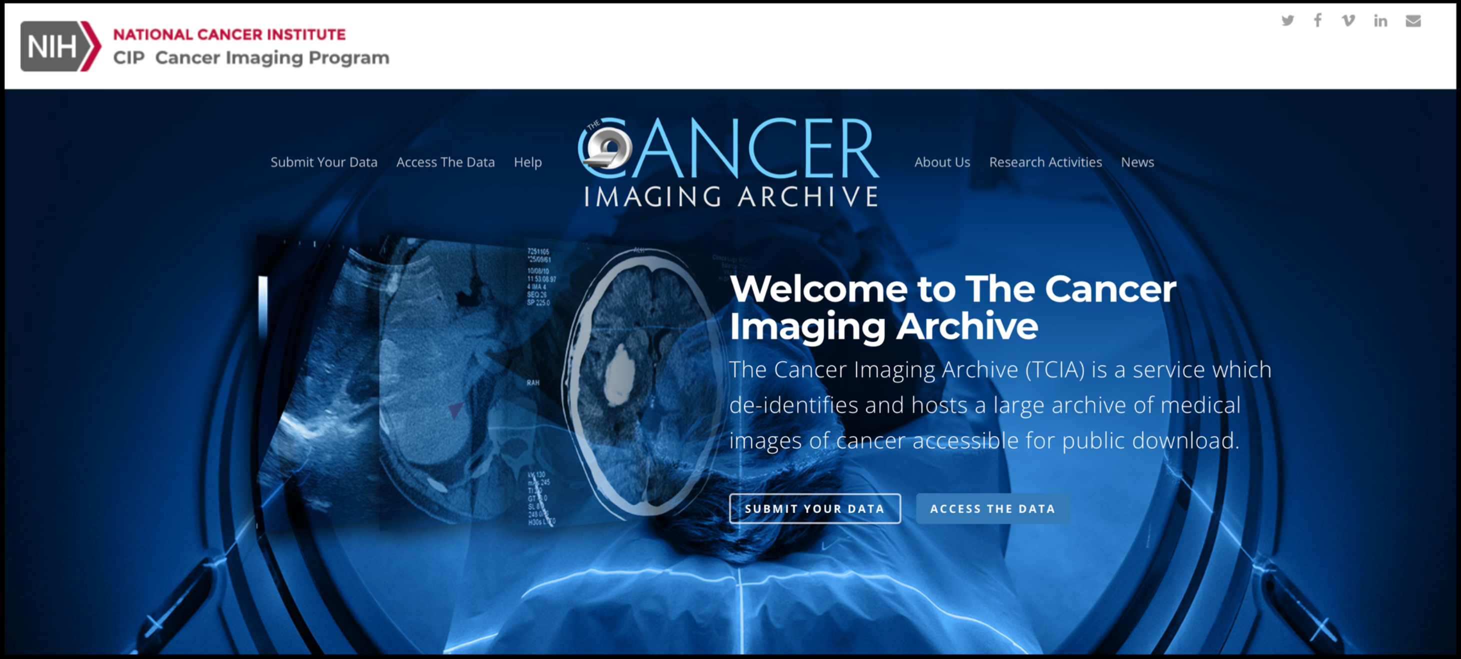


- **https://ncihub.org/groups/medicalimageperception/overview**

Treviño et al. (2021) *JNCI Cancer Spectrum*

Images

- **NCI's Cancer Imaging Archive**



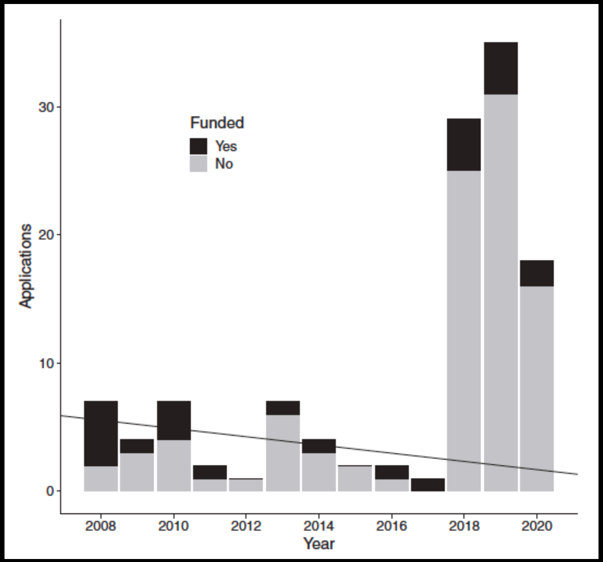
Treviño et al. (2021) *JNCI Cancer Spectrum*

Data Collection

- Pop-up labs at the Radiological Society of North America and American Society for Clinical Pathology
- Email blasts to society listserv

Funding

- **Perception and Cognition Research to Inform Cancer Image Interpretation**
  - https://grants.nih.gov/grants/guide/pa-files/par-19-387.html



Treviño et al. (2021) *JNCI Cancer Spectrum*

Thank you