Artificial Intelligence and Machine Learning for Intestinal Parasite Detection: Machines Helping Make Humans Better Since 2019

Marc Roger Couturier Ph.D., D(ABMM)

Medical Director of Parasitology/Fecal Testing, Infectious Disease Antigen Testing, Bacteriology
ARUP Laboratories
Associate Professor of Pathology, University of Utah

Disclosures

Relevant

- Apacor – research reagents, travel honorarium
- Techcyte – research reagents

Other

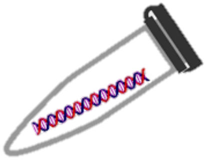
- Biofire Diagnostics – spousal income
- Biomerieux – stock ownership

Objectives

- Understand the theory of AI and how models are trained.
- Recognize the role of AI in stool parasite detection for trichrome stains.
- Describe the future applications of AI in parasitology.

How do we conventionally detect intestinal parasites?

- NEWER:
  - Multiplex PCR



- OLDER:
  - Antigen detection

- ANCIENT RELICS of TIMES OLD:

Pros/Cons of These Methods

- Multiplex PCR
  - Pros: High sensitivity, detect only pathogens
  - Cons: Expensive, limited targets
- Antigen detection
  - Pros: Rapid, inexpensive, detect only pathogens
  - Cons: Limited targets, nonspecific/insensitive (?)
- Microscopy
  - Pros: Detect anything you can see
  - Cons: Insensitive, requires well-trained personnel, difficult to maintain competence, time-consuming, utilizes highly trained/expensive technologists, scope fatigue/burnout, retiring workforce…lack of new interest
    [POOR MARGIN: Growth = pains]

Microscopy and The Ova and Parasite Exam

UNDERSTANDING THE INSANITY OF THE METHOD

Ova and parasite exam

- Fixed stool
  Specimen is concentrated (↑ sensitivity)
  - 2 Components of an O&P
    - **Iodine stain:** specimen added to slide, mixed with iodine and visualized as wet mount
    - **Trichrome stain**: specimen smeared on slide & stained

O&P Recommended Use

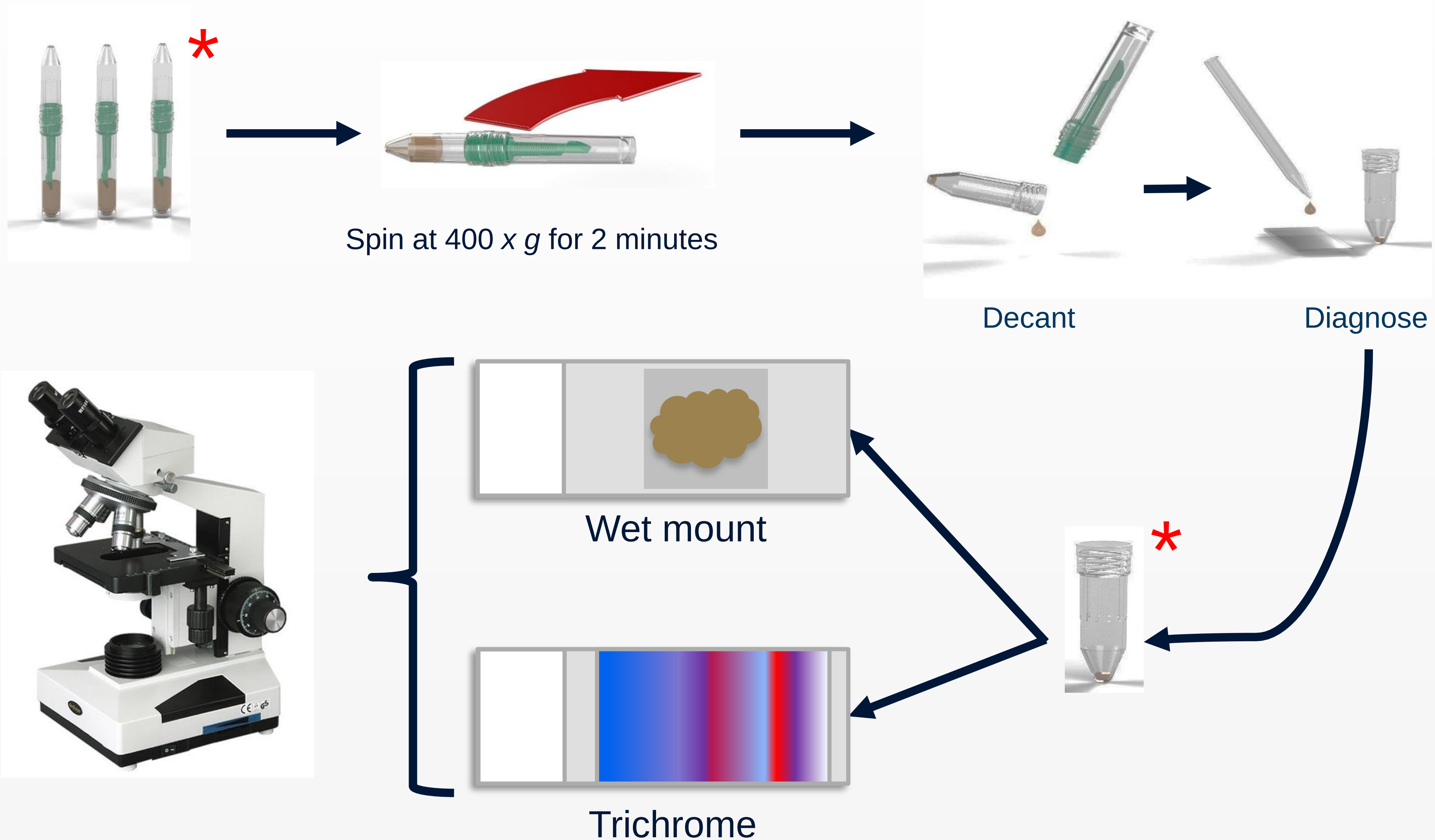
- 3+ unique specimens/patient
- Not recommended for patients with hospital onset diarrhea
- Only for patients with high pre-test probability
  - Immunocompromised patients
  - Pertinent exposure history (immigrants, hikers, splash parks)
  - Pertinent travel history
  - Persistent (>15d)/chronic(>30d) diarrhea with no alternative Dx

O&P ACTUAL Use

- EVERYONE with diarrhea (exaggeration)
- Most unique patients only have one specimen tested
- Unexplained peripheral eosinophilia/allergy workup

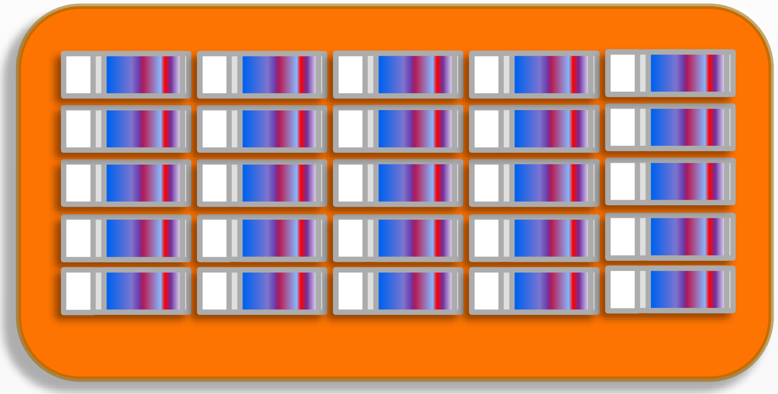
- 65-75k orders
  - ~150k preps/examinations

What goes into an O&P



*Previous automation work

Reading an O&P Run



Run tray

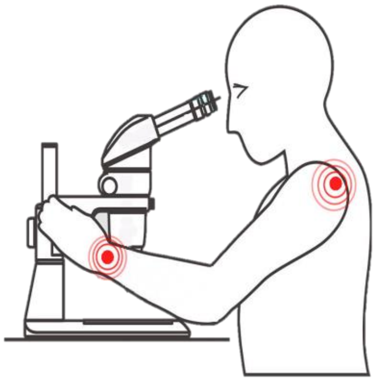
- 30 trichrome slides
- 30 wet mounts

✓ ~2.5 – 3 hours/run
✓ ~98% negative
✓ Positives back-read

Technologist scans specimens looking for parasites

- Anywhere from 2-5 min/slide (technologist variable)
    - “Questionable Negatives” can take longer

Concerns for O&P Reading

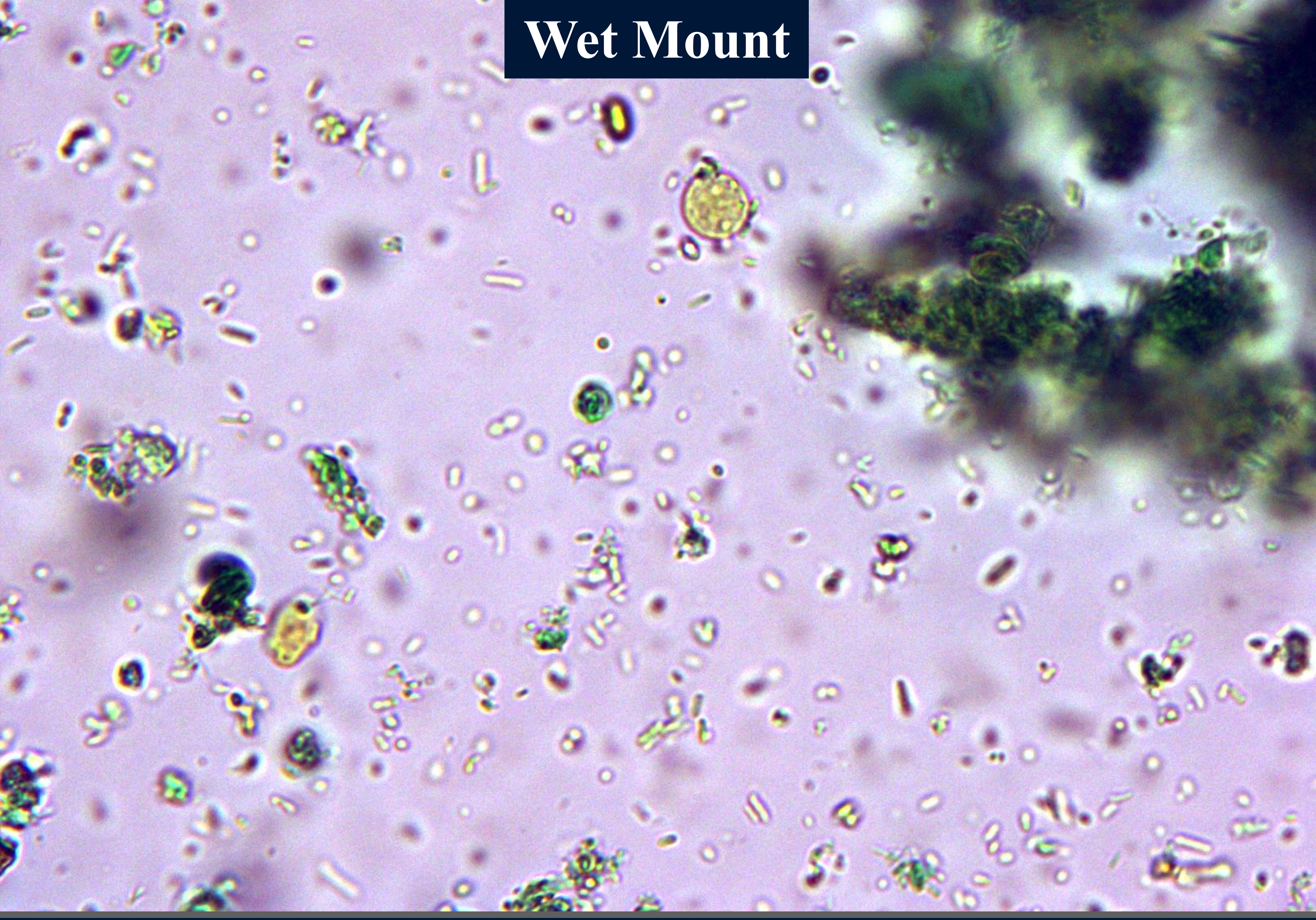


- ✓ Eye strain
- ✓ Neuromuscular strain
- ✓ Burnout/Low Satisfaction
- ✓ Accuracy
  - – Technologist (experience, rest, distractions, *etc*)
  - – AM vs PM
  - – Run 1 vs Run 2 vs Run 3
  - – Low parasite burden challenges interpretation
    - Bias, perceptions over time (searching over time)

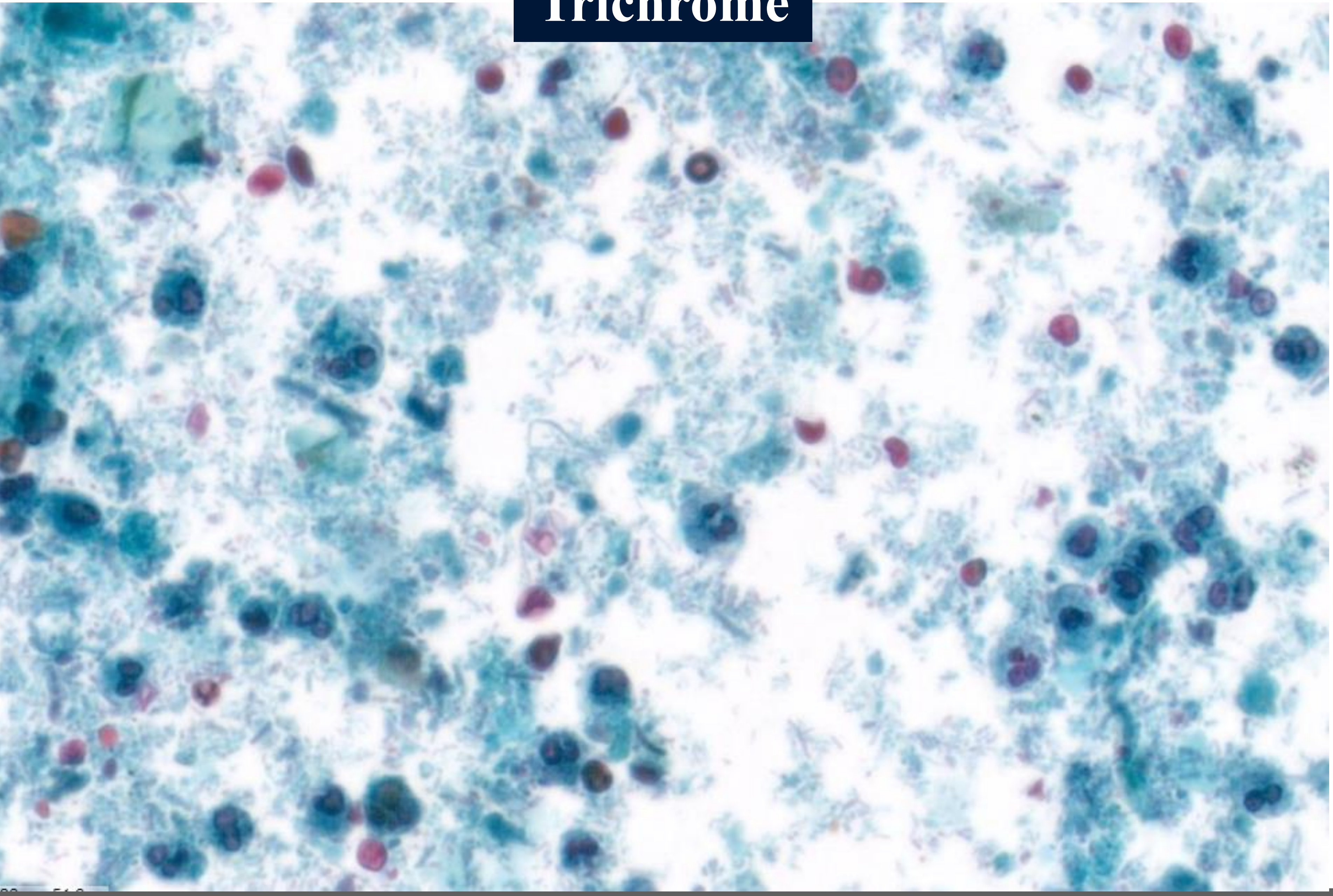
So…

WHAT GOES ON UNDER THE SCOPE?

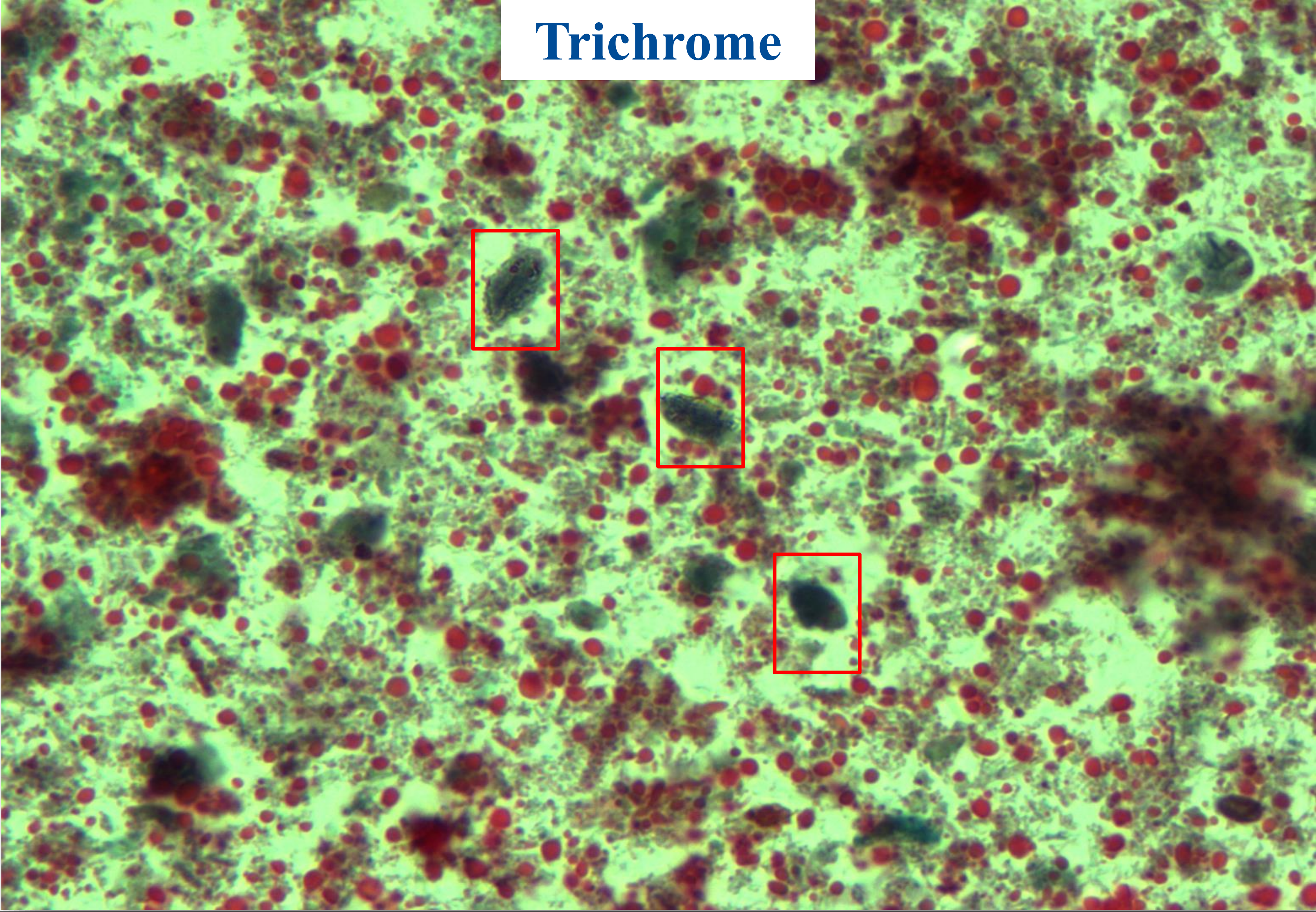
Wet Mount



Trichrome



Trichrome



How can we make this process more efficient and accurate?

Digital imaging and machine learning

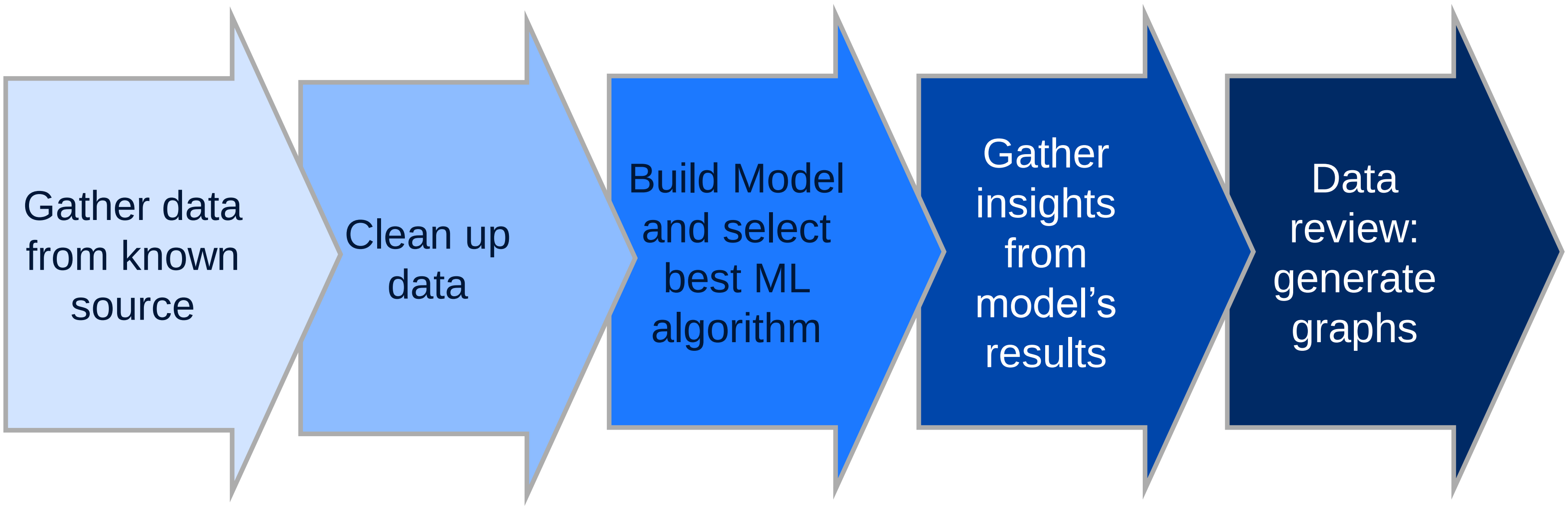
Digital Imaging

Capture images as seen in a microscope and “thread” into a virtual slide for machine or human evaluation

- ✓ Must be high resolution for fine detail determination
- ✓ Must improve ease of review
- ✓ Must be time-effective for scan time considering test volumes
- ✓ Must be user friendly
- ✓ Must be equal or better than what is seen through an eyepiece

Machine Learning – Artificial Intelligence

Machine learning: a computer program determines a solution to a problem without being given an explicit set of commands on how to solve it. Once developed, algorithms are available to use for predictions with new data.



Machine Learning – Simplified

Convolutional Neural Network

- A class of deep neural network models used primarily for image recognition
  - Multiple layers of objects are created from a single input layer
  - “Textures” extracted & analyzed in a 3-Dimensional virtual space
  - Reassembled into an output layer.

What I see…

…What the CNN sees

12	25	31	46	2	11	32	16	13
28	30	11	32	16	13	28	30	12
25	31	46	10	5	45	33	11	32
16	41	7	19	32	17	21	37	1
04	31	19	25	28	30	11	32	16
13	28	30	12	25	31	46	10	5
16	13	28	30	12	16	13	28	30

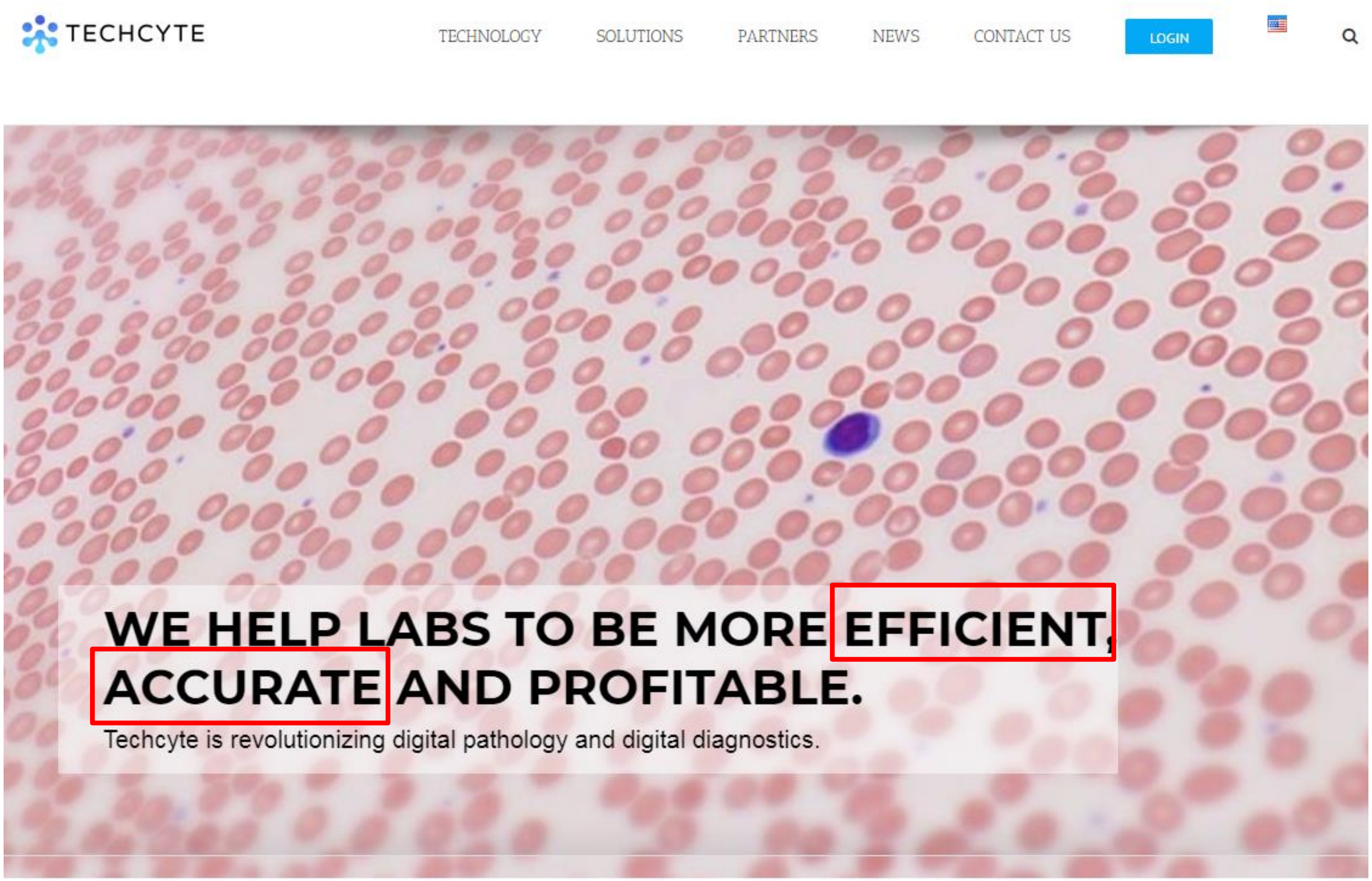
Important Concepts

- Class – a group of images for which a known true identity has been defined
  - *e.g. Giardia duodenalis*

- Class confusion – Assignment of a target to an incorrect class
  - *e.g.* Detection of *Giardia* that is *Dientamoeba*
- Exemplar – a data point that is representative of a group of datapoints
  - *e.g. Giardia* with textbook morphology will allow cluster identification of similar objects going forward. Variations that are common can serve as additional examplars
    - Poor quality examples should not be used as exemplars, but those may be detected by training on diverse exemplars

Techcyte

Located in Lindon, Utah



Proof of Concept with TechCyte

- Too exploratory. Can it even detect *Giardia*?
  - Circa late 2016

Fast forward to 2018…*Giardia* worked

End of summer 2019, technology integrated into routine workflow

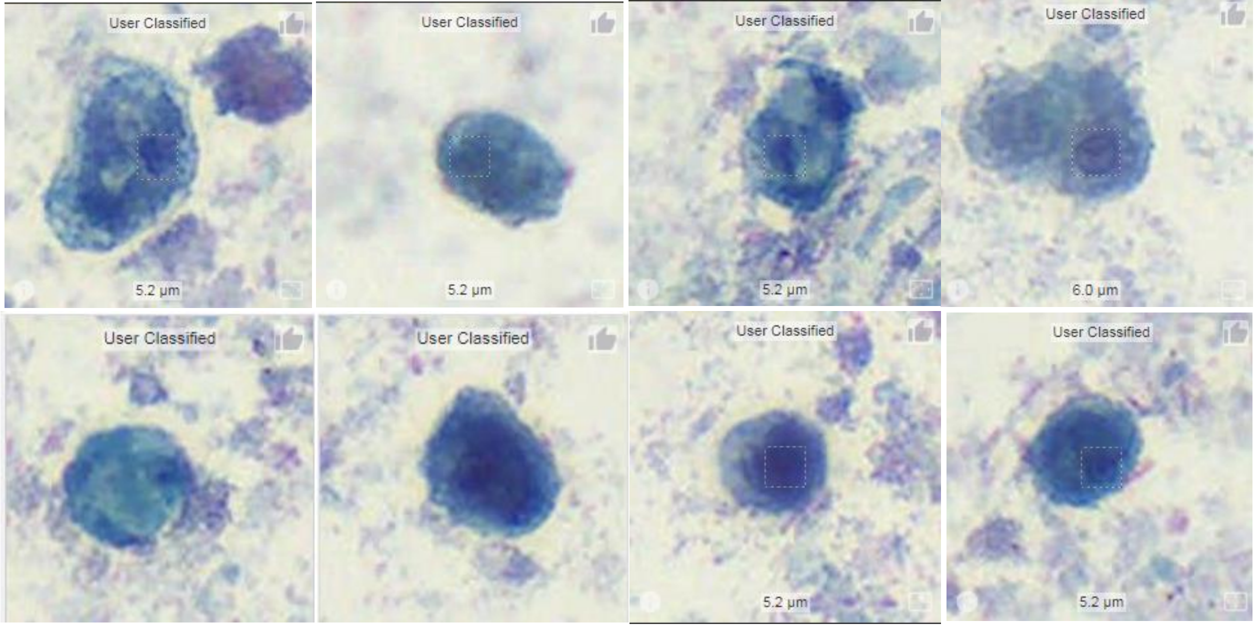
…but how did we get there?

Unsupervised vs Supervised Learning

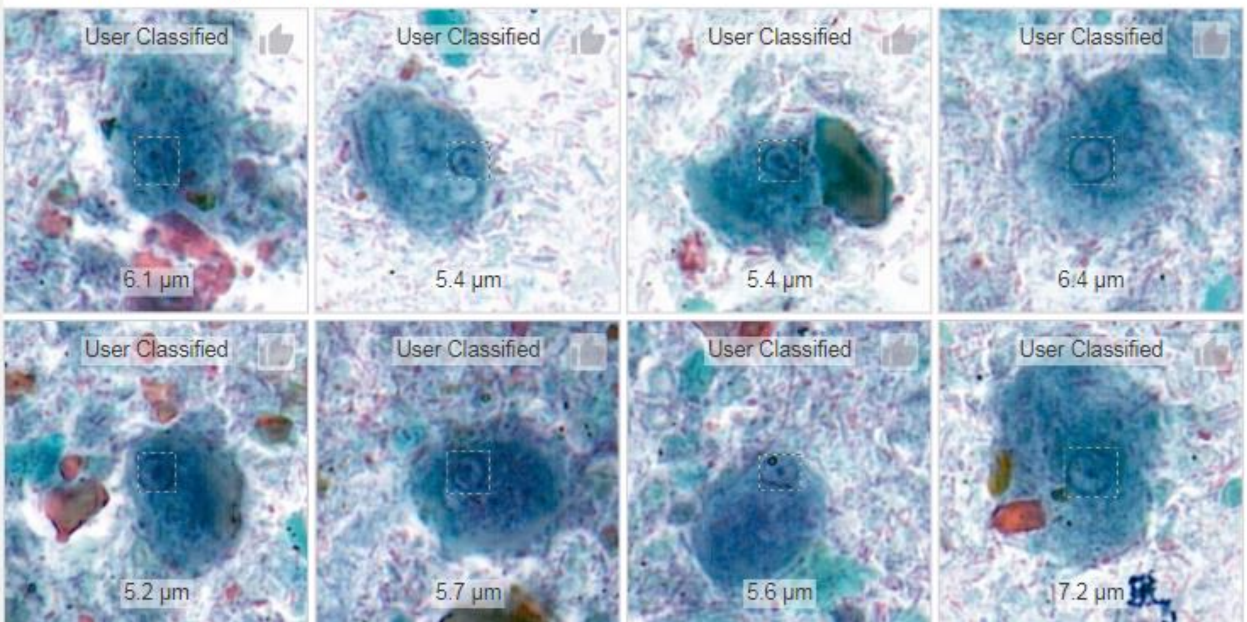
- Slides positive for *Giardia* collected for scanning
- Scanned several slides with *Giardia* (Jess Kohan)
- Allowed TechCyte software to "box" suspicious objects (Unsupervised)
  - Messy! - STOP
- "Expert" teaches software by boxing exemplars (Blaine Mathison)
  - *e.g.* True organisms
  - Supervised (more work up front…better end-product)

Supervised Learning Steps for CNN

- Boxed organisms become “truth”, software finds them again on the same slide
  - Expert can also box “garbage” to teach incorrect objects
- Software eventually can be allowed to predict organisms
  - New objects are boxed by software → confirmed or denied by expert
- Software will continue to run reiterations and will identify more correct and incorrect organisms
  - Expert remediates software…software learns
- After >1000 remediated examples of a class, the predictions become very accurate…BUT…



Entamoeba Troph Nucleus



Scanner Quality Matters

Inexpensive scanner

- Cheaper optics
- Lower throughput/ mechanics

Expensive scanner

- Higher quality optics
- Better throughput/ automation

Wish list metrics for success (circa 2018)

- ✓ Improve speed of review
- ✓ Improve ease of review
- ✓ Reduce/remove the human from the process

CNN Model

Goal: 70% of negative specimens will be screened out by the software with no human review. Remaining 30% (false positives [~28%] and true positives [~2%]) will be read manually.

Development Plan

- Teach the software all necessary organisms/objects (“Classes”) from stool to gain equivalence to trichrome stain
  - Class list:
    - *Giardia duodenalis* trophozoites
    - *Giardia duodenalis* cysts
    - *Entamoeba* species, non-hartmanni trophozoites
    - *Entamoeba hartmanni* trophozoites
    - *Iodamoeba/Endolimax/Dientamoeba* trophozoites
    - *Blastocystis* species
    - *Chilomastix mesnili* trophozoites
    - RBC
    - WBC

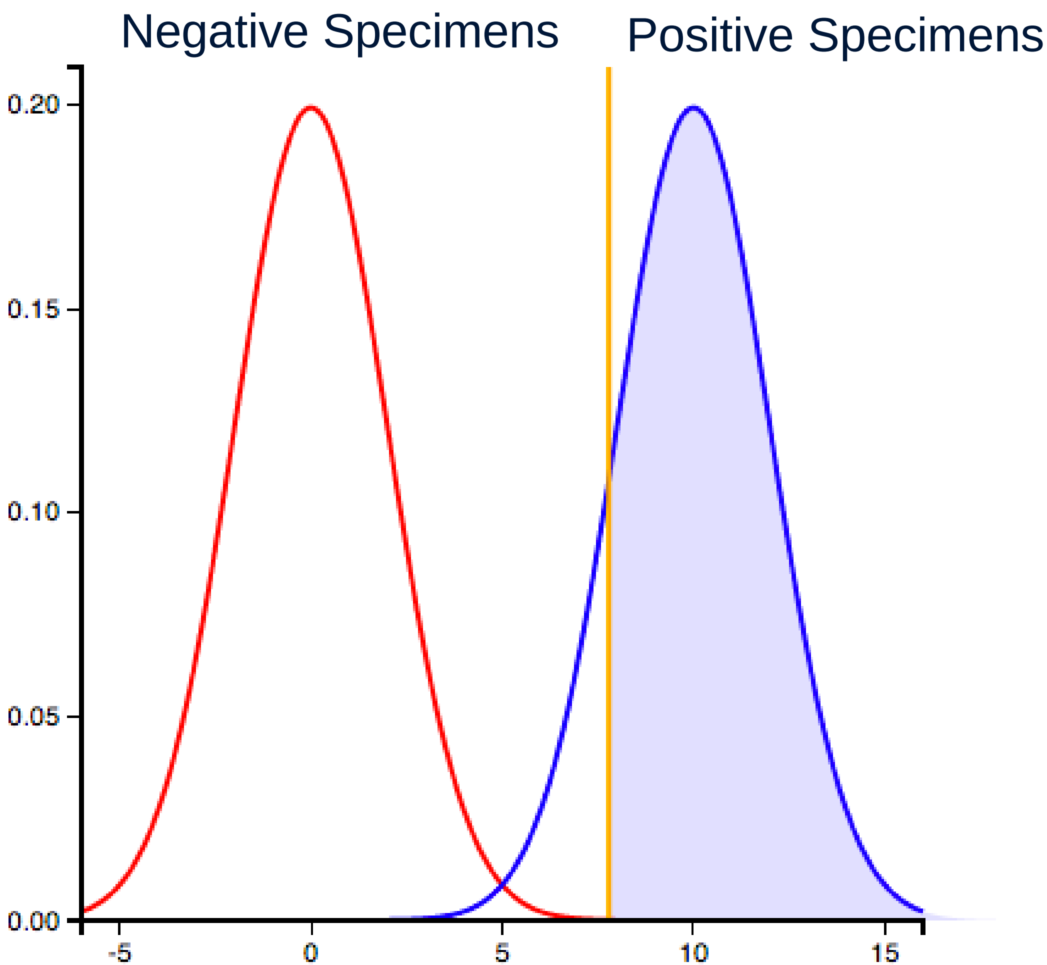
Even with a Great Scanner…Training Material can Enable Machine ~~Learning~~ “Cheating”

- CNN Model recognized patterns and textures features…not organisms
  - The same slide scanned twice cannot fail…WORST DATA
  - Different scan area of same slide previously used (organism similar, background texture similar) – OK
  - Mix positive 1:1 with negative (organism similar, background texture different, dilute target) - BETTER
  - Unique patient specimens (organism and background are unique) – BEST
    - Unfortunately for some organisms…finding 50+ unique positives is tough

What is success? It is complicated…

- Sensitivity/specificity values become subjective without context
  - Slide-level specificity/sensitivity?
    - Human sees the whole story – Final answer “X”
    - CNN Model captures everything it sees and documents it – No final answer, just body of evidence for a human to use
    - If classic test characteristics are applied…specificity is 0.00%
  - Organism-level? – Unrealistic
    - To be 100% specific at the individual organism level, sensitivity at the slide level would suffer
    - Not our goal to be perfectly specific…a human isn’t

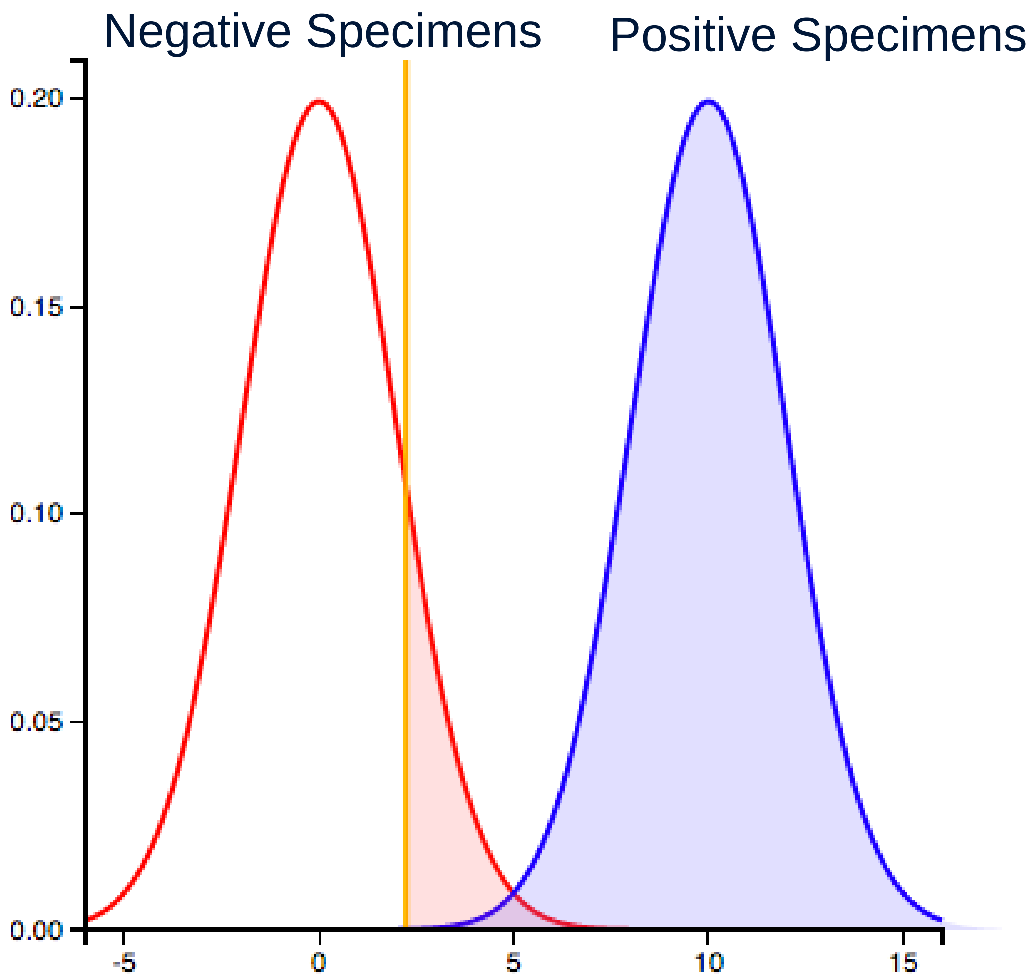
Perfect Specificity, Lower Sensitivity…



- Everything identified is true
- Will miss some positives

Courtesy Rick Smith, Techcyte

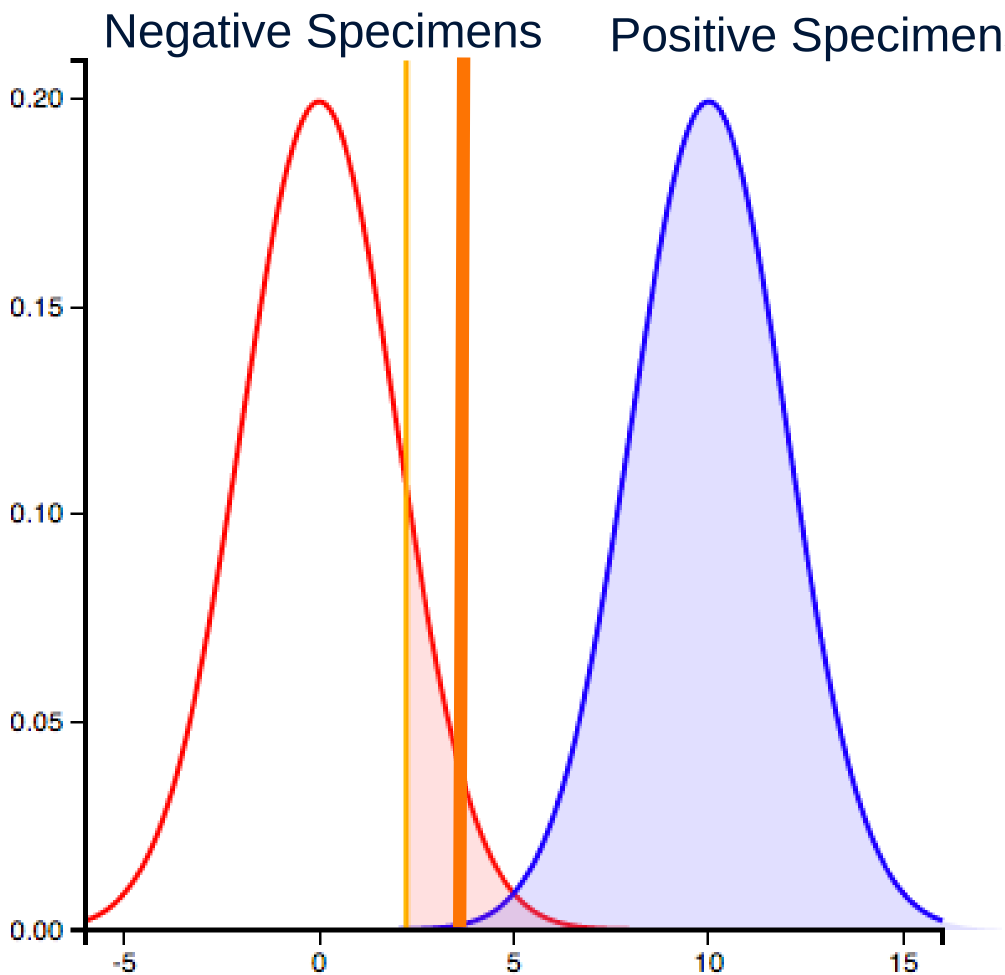
Perfect Sensitivity, Lower Specificity…



- Identifies everything
- Also shows you some junk, but the human arbitrates it
- Find a sweet spot that catches all, but limits junk!

Courtesy Rick Smith, Techcyte

Perfect Sensitivity, Lower Specificity…



- Identifies everything
- Also shows you some junk, but the human arbitrates it
- Find a sweet spot that catches all, but limits junk!

Courtesy Rick Smith, Techcyte

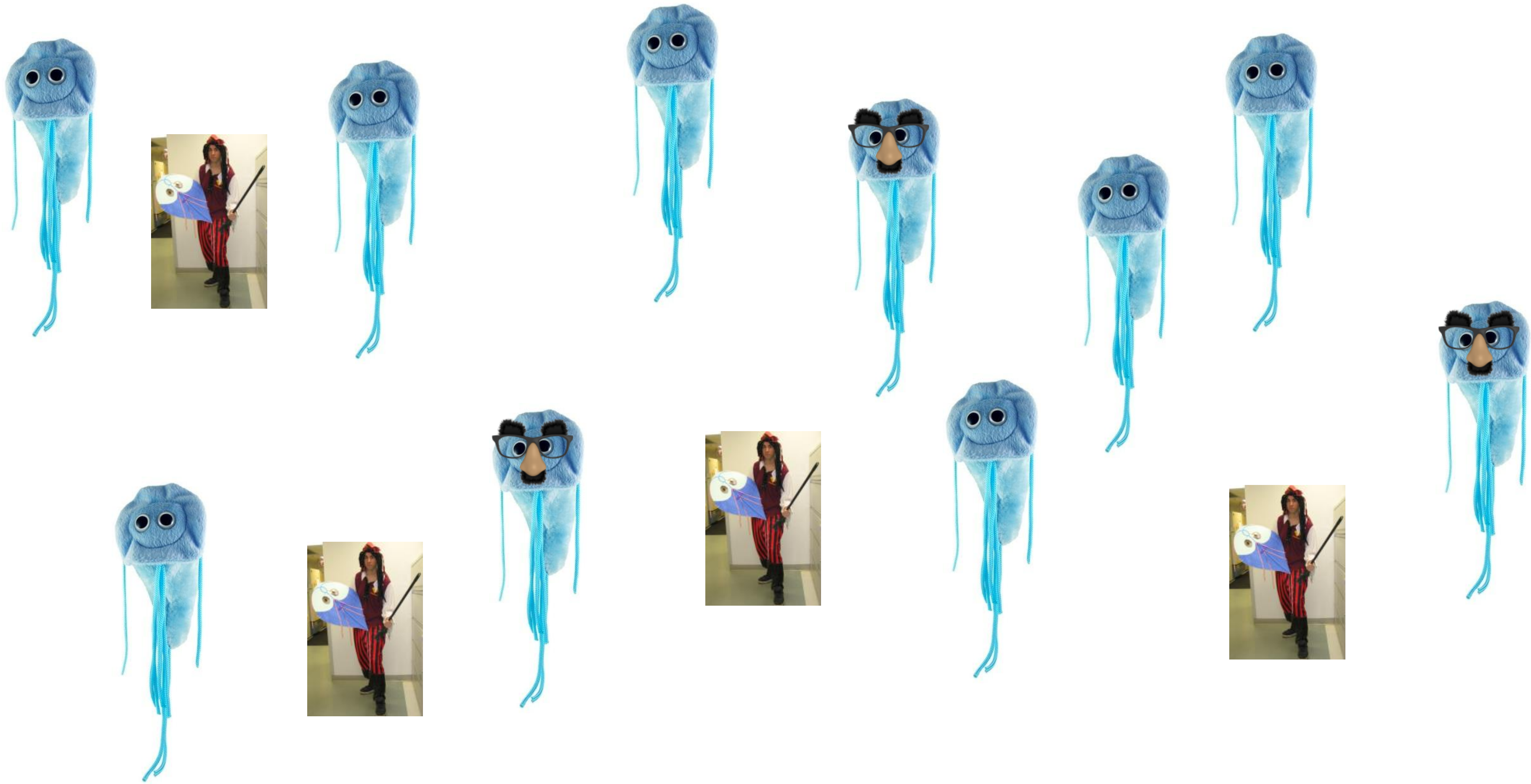
Slide Level Detection

Giardia

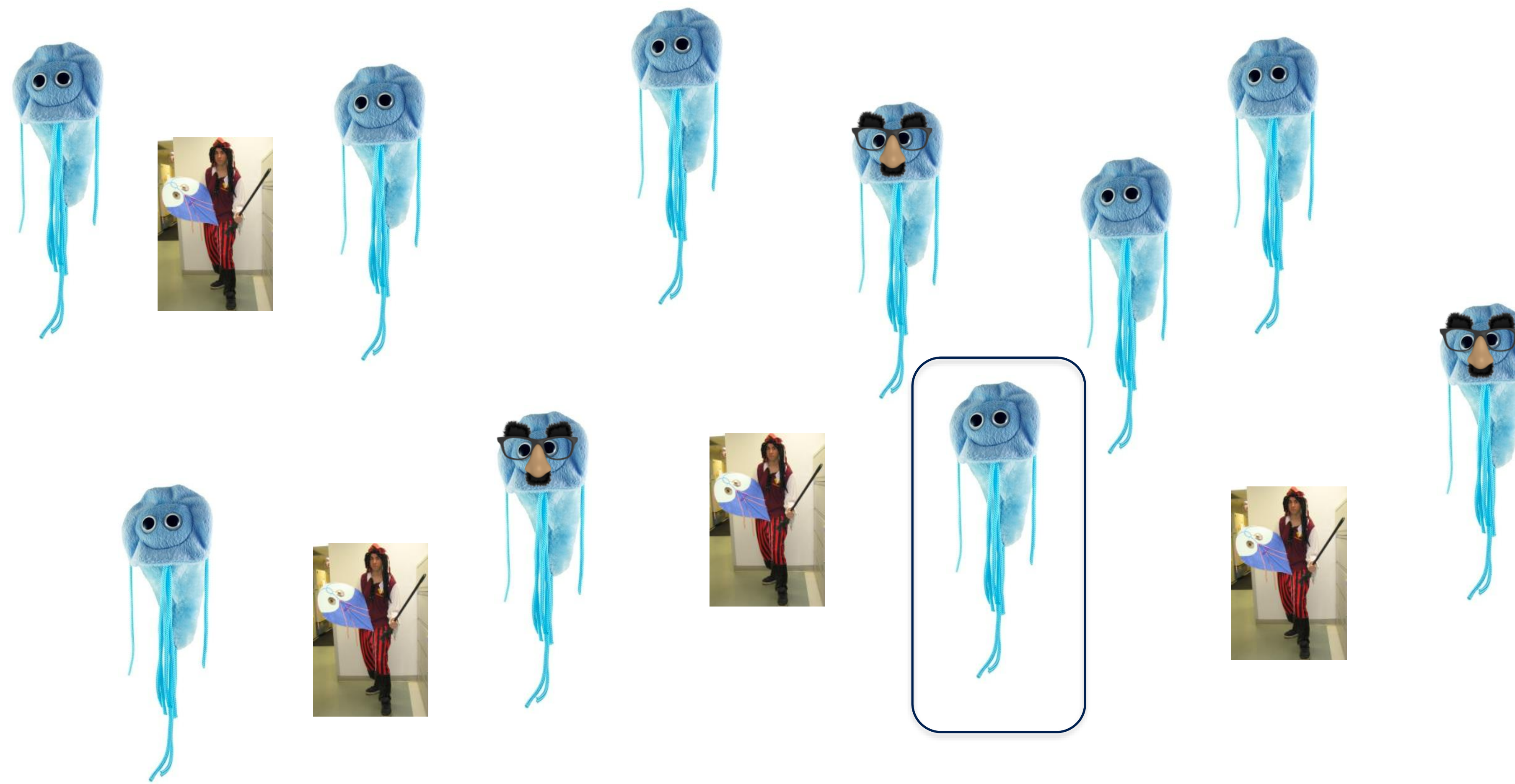
Unusual Giardia
or different parasite

Not Giardia…
no idea what this is

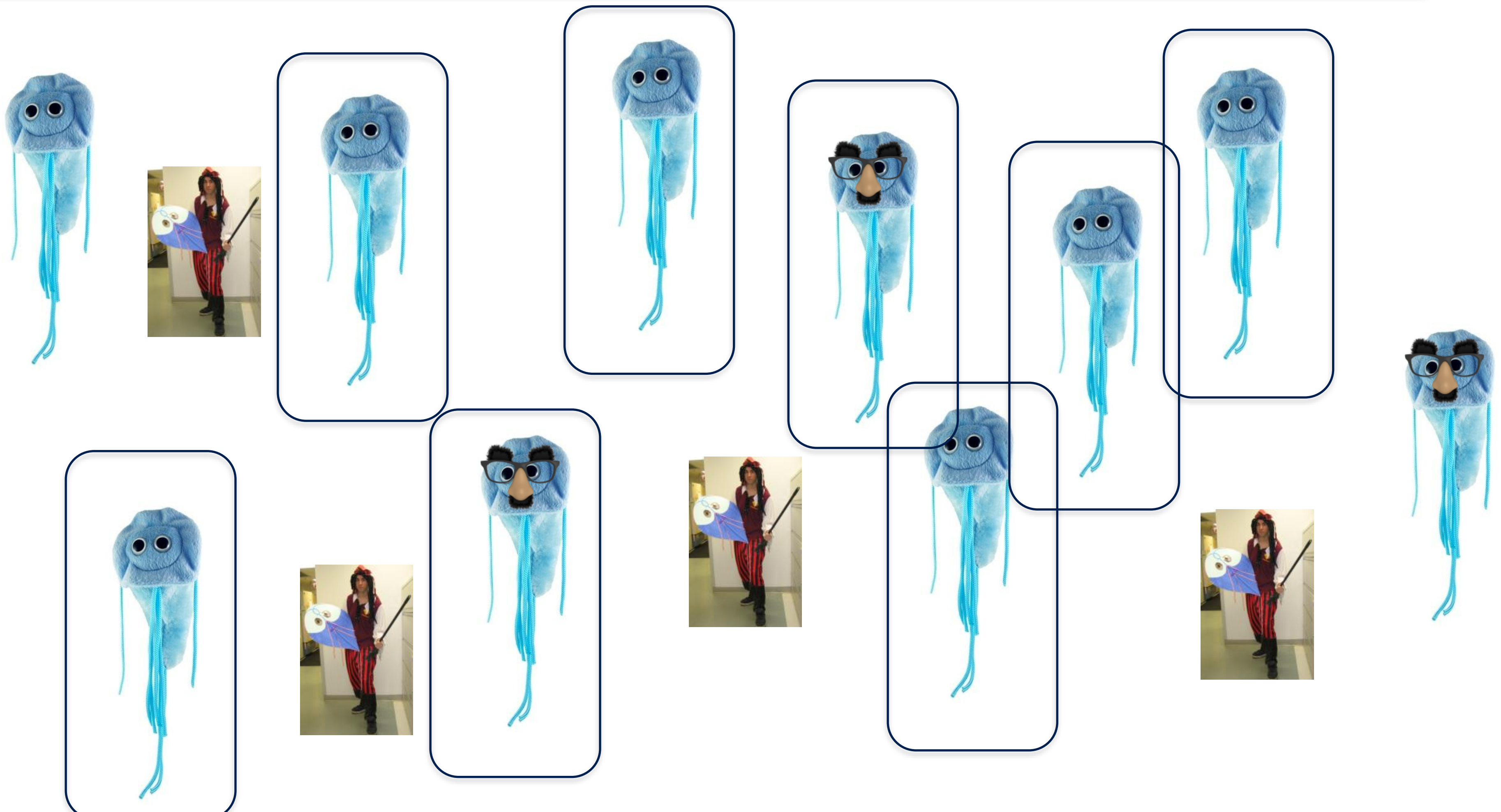
Slide Level Detection = Find 1, you win (same as human)



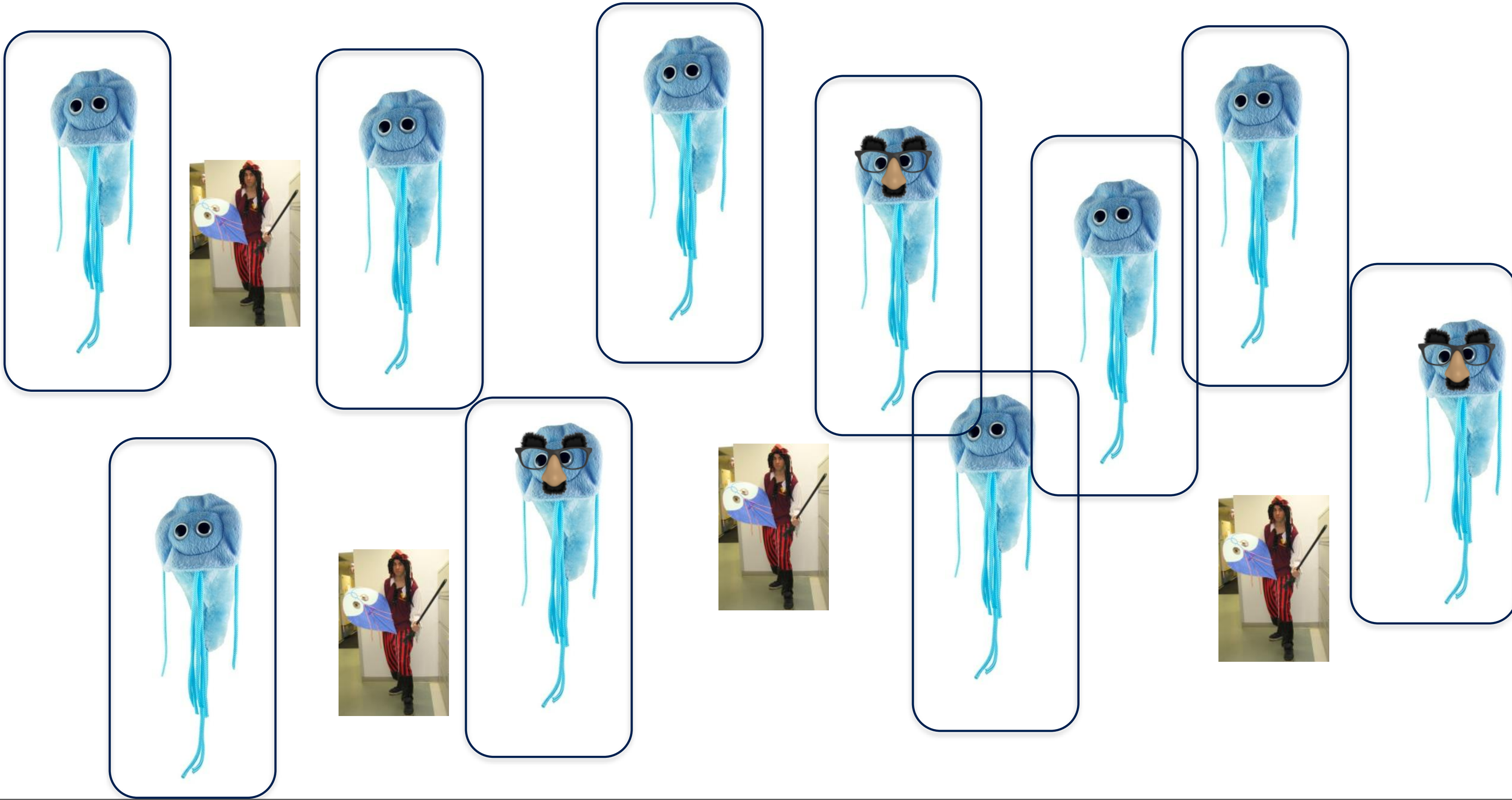
10% sensitivity = Still win, but risky



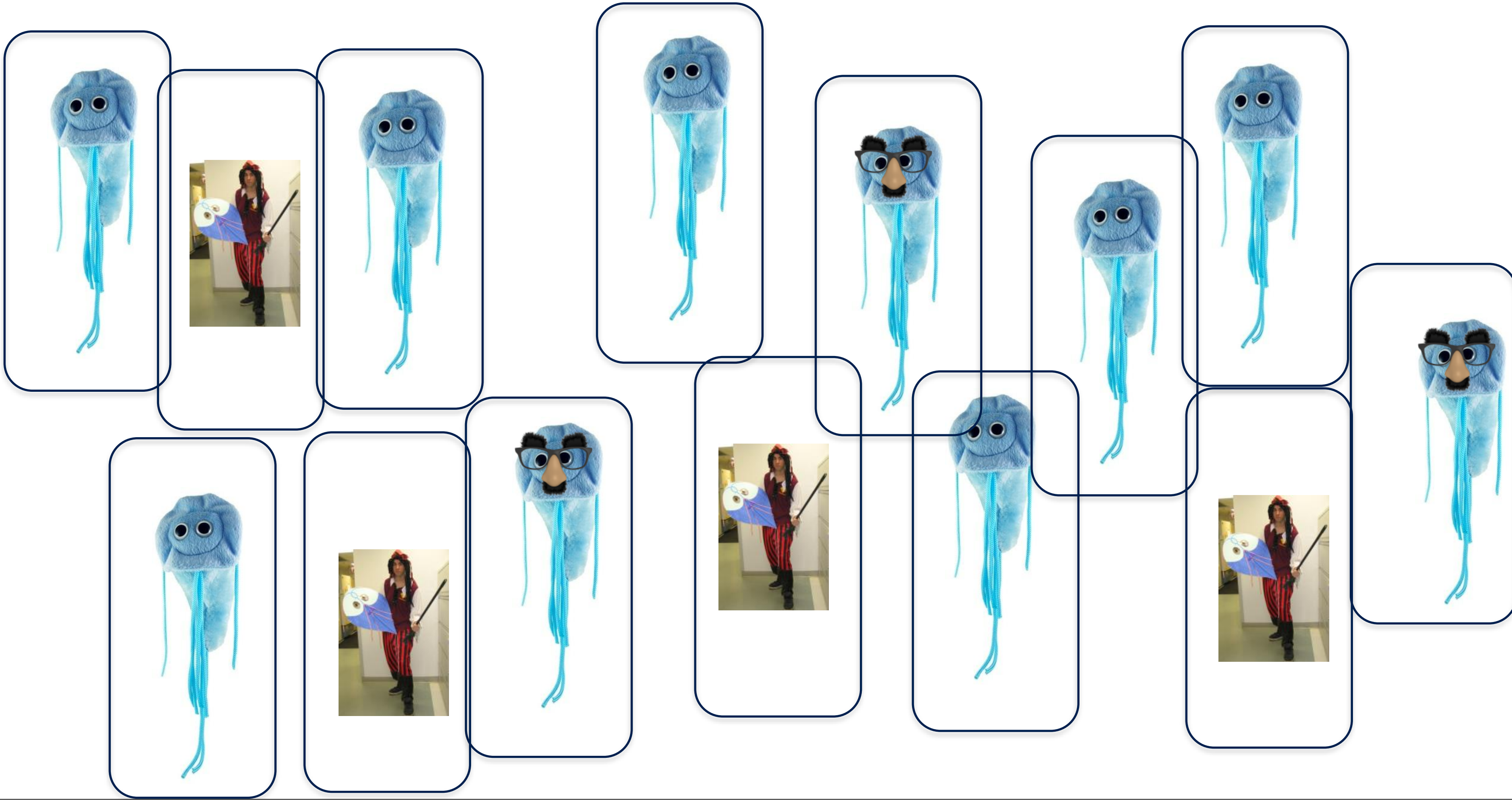
80% sensitivity = Better, but more challenging



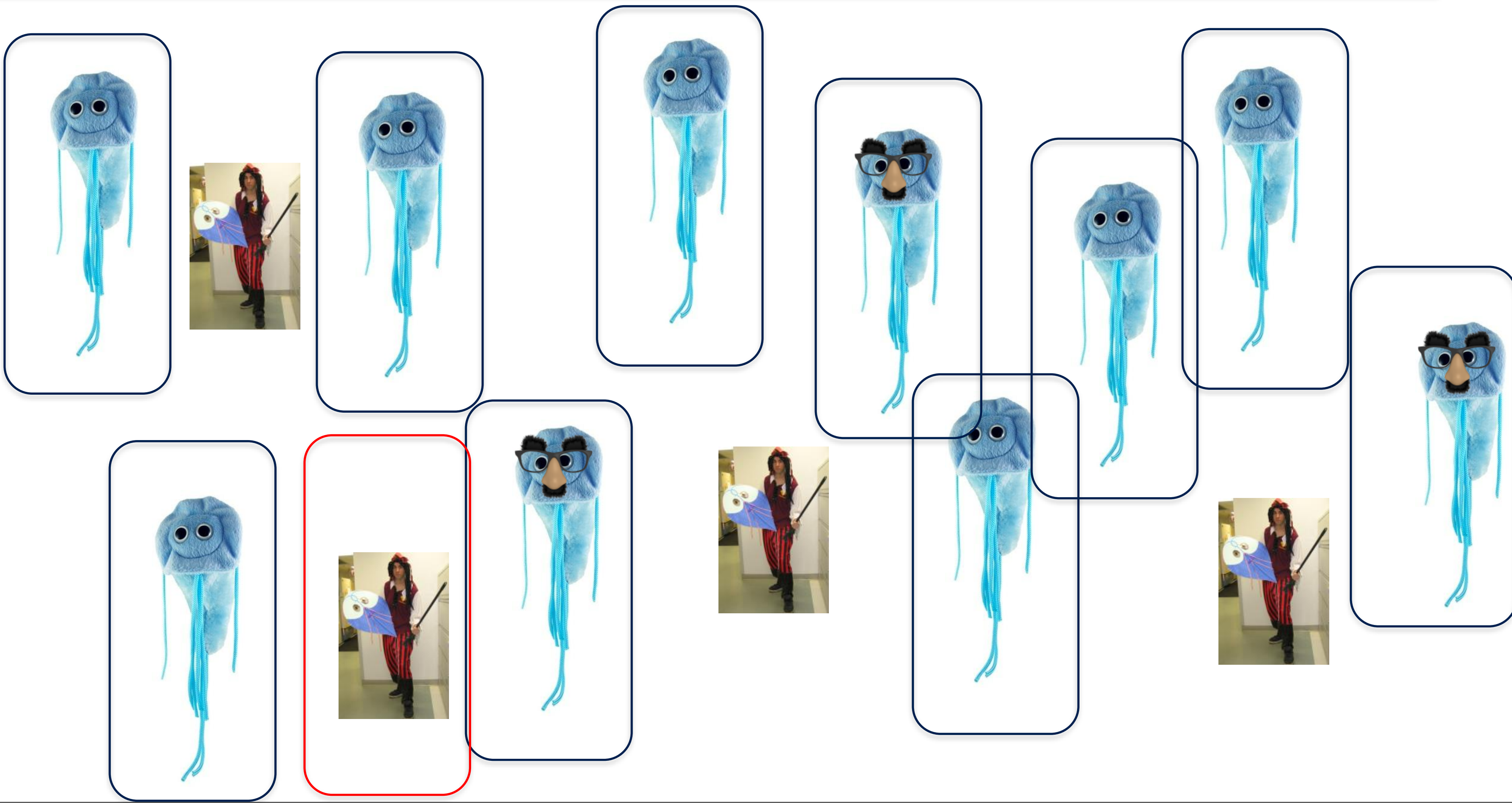
100% sensitivity = Great but…



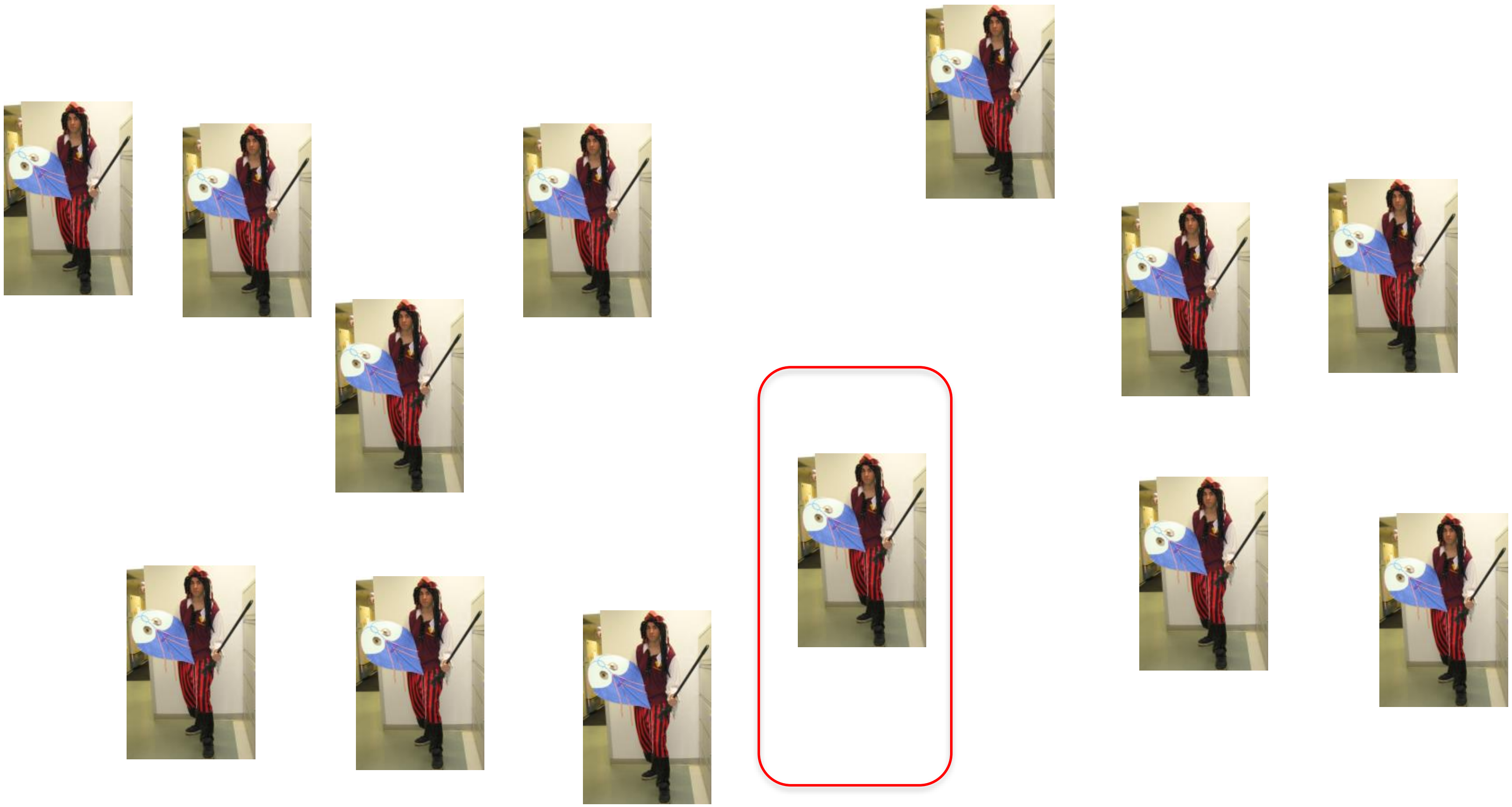
100% sensitivity = May get to this…



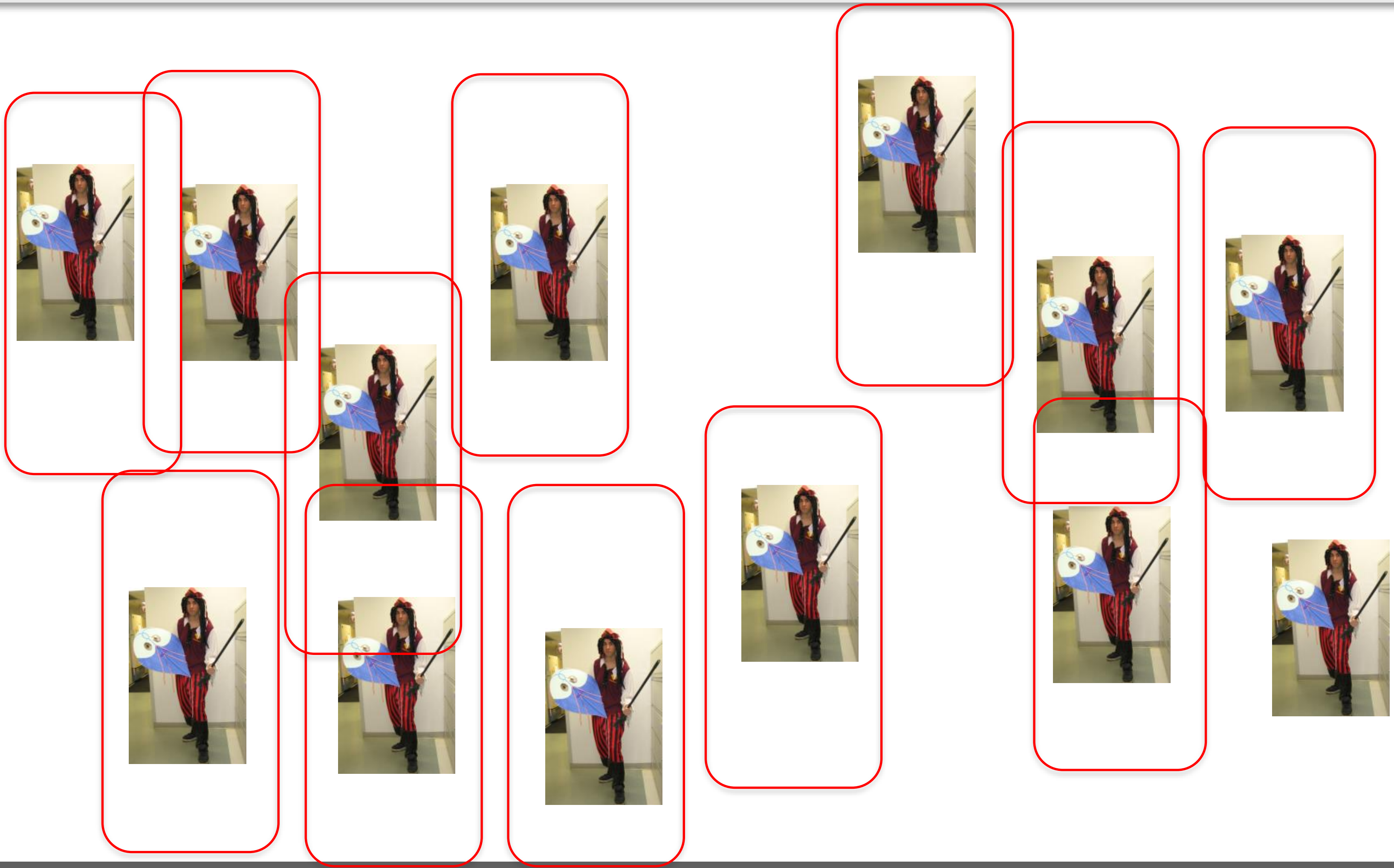
Is this a problem? No…human arbitrates



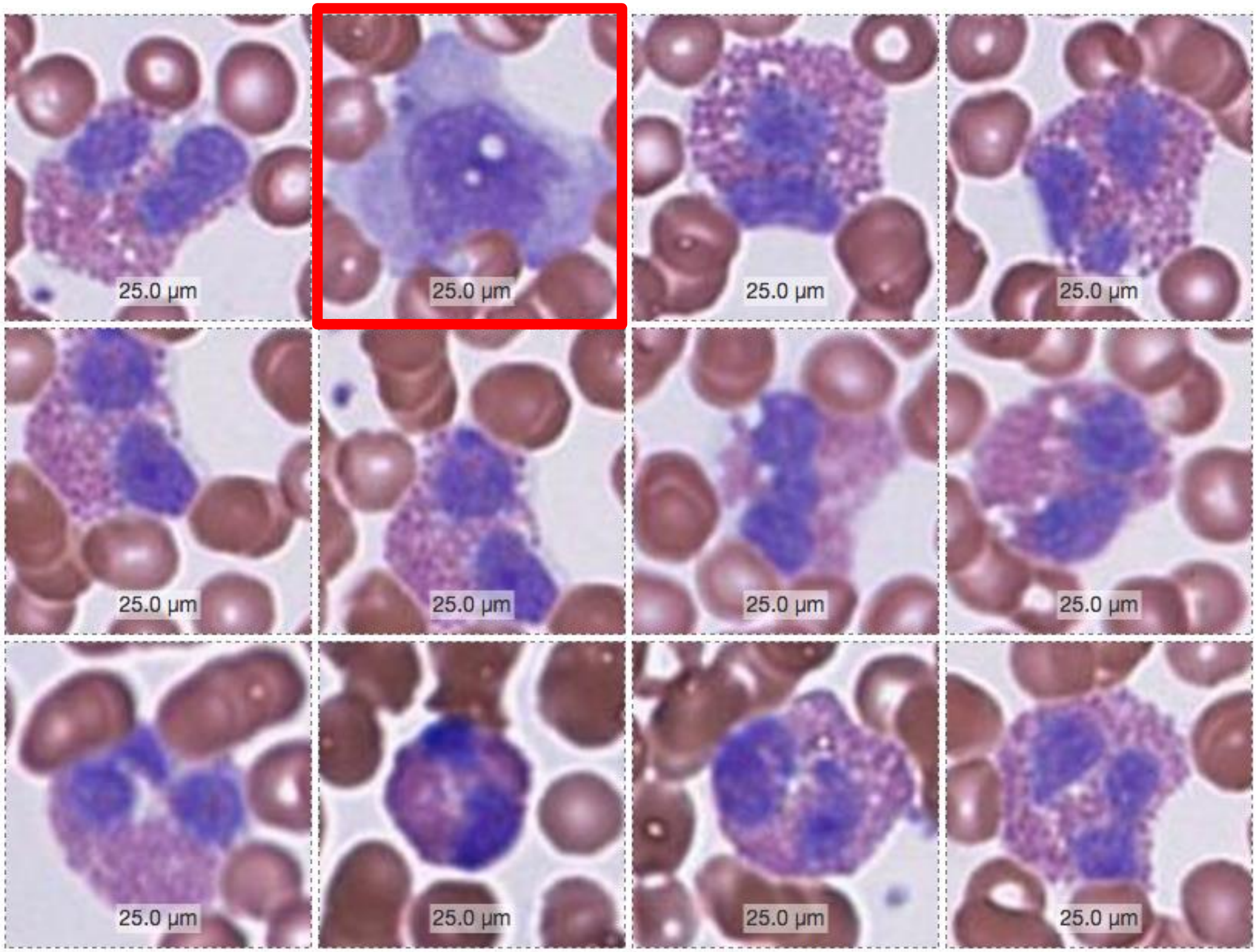
Is this a problem? No…human arbitrates



Is this a problem? Yes…time waste



Using blood as an example



Courtesy Rick Smith, Techcyte

Forget the classic metrics, 

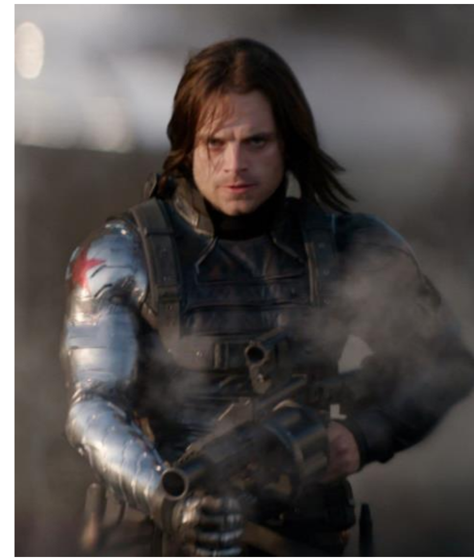
& accuracy are key!

REVISED GOAL: CNN model will detect stool parasites equal to or better than a human. CNN model will result in gains in efficiency and accuracy.

Replacement vs Tool

Robot that learns, eliminates its teacher, and takes over the job.

Human that uses a tool to do its job faster and more efficient



CNN Model does not need to replace a human

- **<u>Augmentation!!</u>**

CNN Model helps a human be:

- More efficient
- More accurate
- Suffer lower burnout…

Training View of CNN Model for a User Look

How does CNN Model compare to humans?

- Accuracy on positives:
  - ~120 total positives scanned for training
    - 15 specimens contained additional organisms not originally identified by the human
      - ~12% of positive specimens were inaccurately identified by humans
        - » CNN Model would provide guidance for manual review
    - 1 specimen contained organism identified by human: CNN model missed
      - *e.g*. 0.8% of positive specimens scanned was missed by CNN model
        - » But it did catch other organisms in the slide, which would prompt manual review

DISCLAIMER: <u>Full</u> slide scan versus human scanning SOP

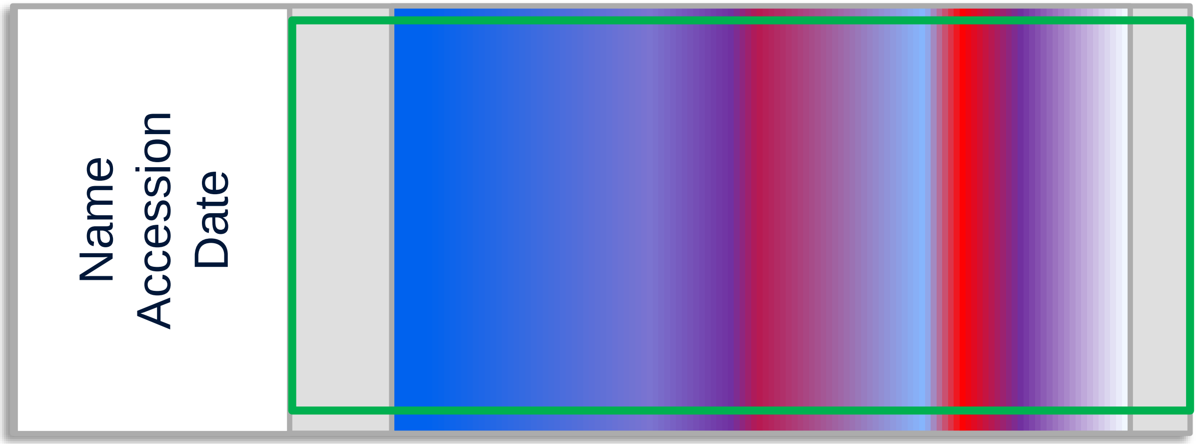
Where is the Sweet Spot?

- Confidence Class Chart
  - Apply a filter cut-off for confidence
    - Model, only show me ">XX%"
  - Maximize True Positives
  - Minimize False Negatives
  - Minimize False Positives

		Blastocystissp	ChilomastixTroph	Dientamoebafragilis	EndolimaxIodamoebaTroph	EntamoebaTrophNucleus	EntamoebahartmanniTroph	GiardiaCyst	GiardiaTroph	Redbloodcells	Whitebloodcells	Yeast
	TP	2205	269	1033	727	405	101	575	191	804	196	1351
0.05	FP	1003	85	230	810	187	91	214	55	654	118	1092
	FN	51	84	285	53	11	27	11	6	51	19	75
	TP	2174	269	1033	715	401	99	569	191	789	194	1323
0.1	FP	798	85	230	626	134	70	174	42	521	96	784
	FN	82	84	285	65	15	29	17	6	66	21	103
	TP	2121	269	1033	693	397	97	556	186	768	185	1272
0.2	FP	620	85	230	422	96	54	147	35	431	76	564
	FN	135	84	285	87	19	31	30	11	87	30	154
	TP	2051	269	1033	666	387	96	538	180	746	179	1221
0.3	FP	517	85	230	339	79	45	129	29	355	63	431
	FN	205	84	285	114	29	32	48	17	109	36	205
	TP	1979	269	1033	635	379	92	521	179	716	177	1147
0.4	FP	421	85	230	270	63	34	115	26	297	57	359
	FN	277	84	285	145	37	36	65	18	139	38	279
	TP	1882	269	998	596	375	89	501	177	681	170	1084
0.5	FP	353	85	201	219	50	31	99	25	259	48	288
	FN	374	84	320	184	41	39	85	20	174	45	342
	TP	1755	269	926	534	365	86	481	176	626	169	1033
0.6	FP	280	85	144	174	42	29	82	21	225	44	239
	FN	501	84	392	246	51	42	105	21	229	46	393
	TP	1582	259	830	480	356	84	455	175	569	166	945
0.7	FP	218	80	101	140	37	26	62	19	186	37	180
	FN	674	94	488	300	60	44	131	22	286	49	481
	TP	1352	247	701	423	343	79	418	169	497	161	827
0.8	FP	150	62	63	105	33	23	46	17	145	33	132
	FN	904	106	617	357	73	49	168	28	358	54	599
	TP	942	219	495	320	320	73	352	164	371	147	652
0.9	FP	71	50	34	64	26	14	35	15	70	27	72
	FN	1314	134	823	460	96	55	234	33	484	68	774
	TP	572	195	316	217	292	70	285	159	252	137	458
0.95	FP	39	43	19	32	22	10	25	12	35	21	36
	FN	1684	158	1002	563	124	58	301	38	603	78	968
	TP	202	160	136	103	251	63	206	152	121	117	241
0.98	FP	9	30	7	8	16	5	14	9	11	15	13
	FN	2054	193	1182	677	165	65	380	45	734	98	1185
	TP	53	129	48	50	223	58	156	146	67	110	125
0.99	FP	2	16	4	1	12	4	10	7	1	9	1
	FN	2203	224	1270	730	193	70	430	51	788	105	1301

Whole Slide Scanning

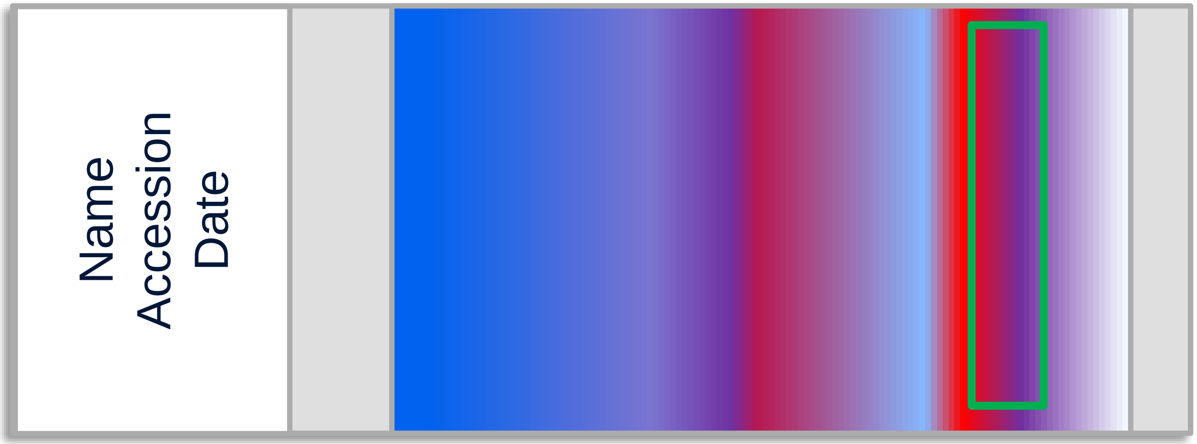
- Not time effective
- Determine slide scan area necessary to minimize scan time and maintain equal or better accuracy than human



~25-30 min

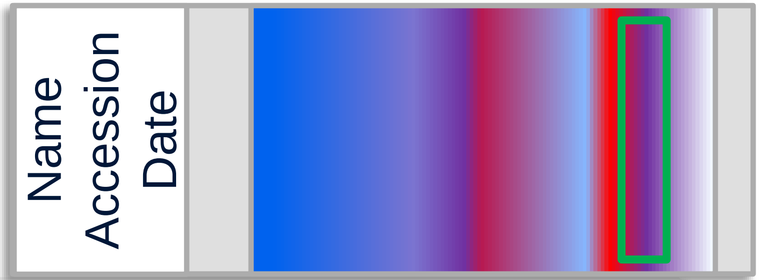
Whole Slide Scanning

- Not time effective
- Determine slide scan area necessary to minimize scan time and maintain equal or better accuracy than human



~4-5 min

Are We Still Sensitive? (@ slide level)



- Identified positive stool specimen containing *Giardia* & *Blastocystis*
- Serially diluted in negative stool
- Prepared duplicate slides of each dilution
  - Manual read in lab (blindly integrated into run)
  - Scanned and read by CNN Model
- Compare Analytical sensitivity with new scan area

Limit of Detection

Dilution	Technologist read	CNN Model read
Neat	*Giardia* + *Blastocystis*	*Giardia* (276) + *Blastocystis* (129)
1:1	*Giardia* + *Blastocystis*	*Giardia* (95) + *Blastocystis* (19)
1:2	*Giardia* + *Blastocystis*	*Giardia* (68) + *Blastocystis* (17)
1:4	*Giardia* + *Blastocystis*	*Giardia* (79) + *Blastocystis* (46)
1:8	Negative	*Giardia* (70) + *Blastocystis* (13)
1:16	*Giardia* + *Blastocystis* (rare)	*Giardia* (12) + *Blastocystis* (10)
1:32	Negative	*Giardia* (16) + *Blastocystis* (5)
1:64	Negative	*Giardia* (15) + *Blastocystis* (2)
1:128	Negative	*Giardia* (9) + *Blastocystis* (1)
1:256	Negative	*Giardia* (15) + *Blastocystis* (1)

CNN Model (with constrained scan region) was 4-6 fold more sensitive than a human BUT…remember, this is a tool…so human still wins!

CNN Model enters validation

- Software is locked down: no further learning
- Modified slide prep and autocoverslipper validated
- TechCyte uses “holdout” slides to validate final software performance
  - ARUP uses unique validation slide set to internally validate performance of software
    - Development = >12 months, validation = < 2 weeks
- Production lab trains on new process
- Go Live (invisible to anyone external)

Validation Slide Set

Category (Class)	Unique Slides per Class	Examples per Class
Giardia duodenalis cyst	23	6,499
Giardia duodenalis trophozoite	21	2,191
Blastocystis sp.	61	23,566
Dientamoeba fragilis	29	12,764
Entamoeba non-hartmanni trophozoite	34	4,307
Entamoeba hartmanni trophozoite	10	1,394
Chilomastix mesnili trophozoite	15	4,064
Endolimax nana/Iodamoeba buetschlii trophozoite	36	7,914
Red Blood Cells	18	8,482
White Blood Cells	31	2,099
Yeast	94	13,450

Limit of Detection

- 4 serial dilution slide sets blindly run through lab
- 1 additional set tested by CNN model

- No dilutions series (n=4) was detected below 1:16 dilution (human)
- CNN model detected to 1:256

Slide-Level Accuracy

- Consider definitions:
- False Positive: CNN model presents images to an expert that cannot be excluded as "False" without review of physical slide.
  - Shows me 80, 10 are wrong…why consider that false positive?

O&P Examination

CNN Model Analysis		Positive	Negative
	Positive	86	2
	Negative	1	104

Positive percent agreement: 98.88% [95% CI 93.76% to 99.98%]

Negative percent agreement: 98.11% [95% CI 93.35% to 99.77%]

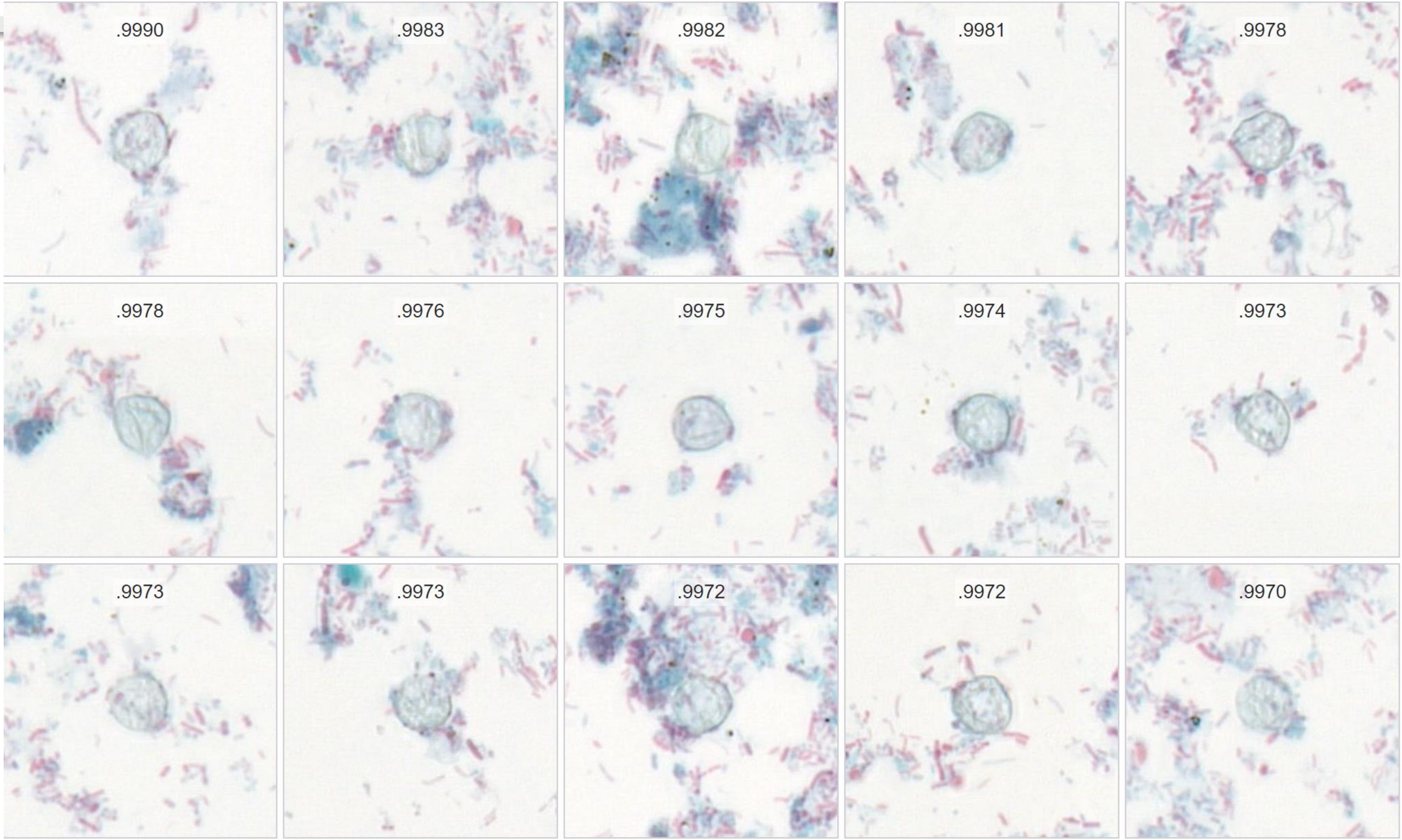
The Future?

- Continuing to teach CNN model with run data (in training environment), validate future iterations of software
  - Add new targets: *Cyclospora* (a GOOD class confusion)

- Wet mounts
  - Second component of O&P more challenging

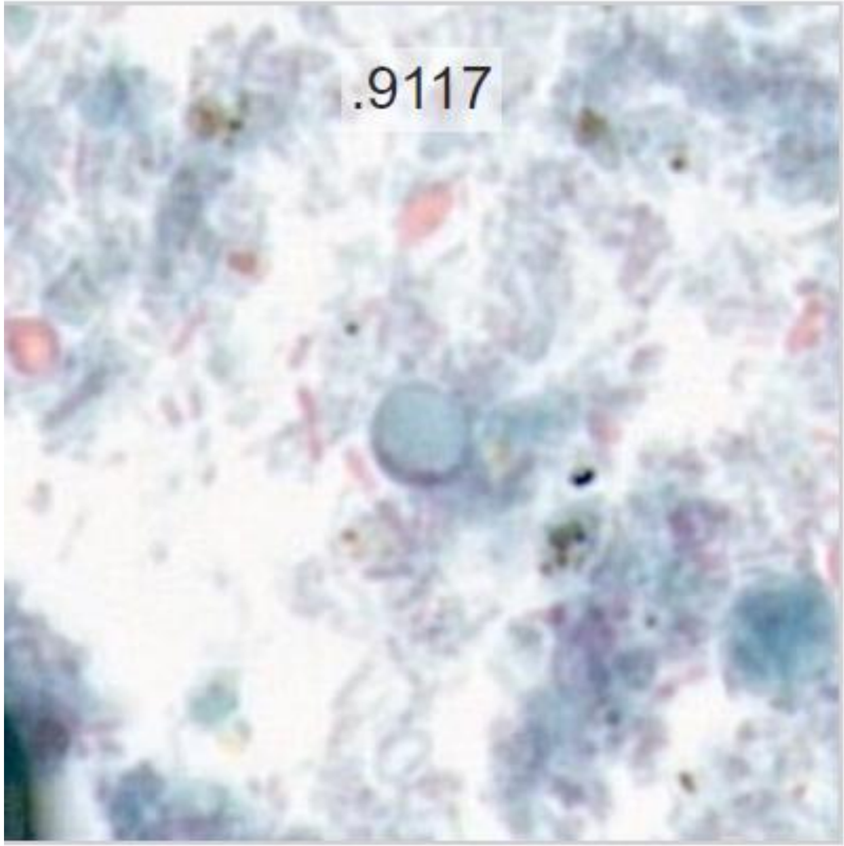
- Modified acid fast stain: MAF (*Cryptosporidium & Cyclospora*)

Cyclospora in trichrome NOT *Blastocystis*

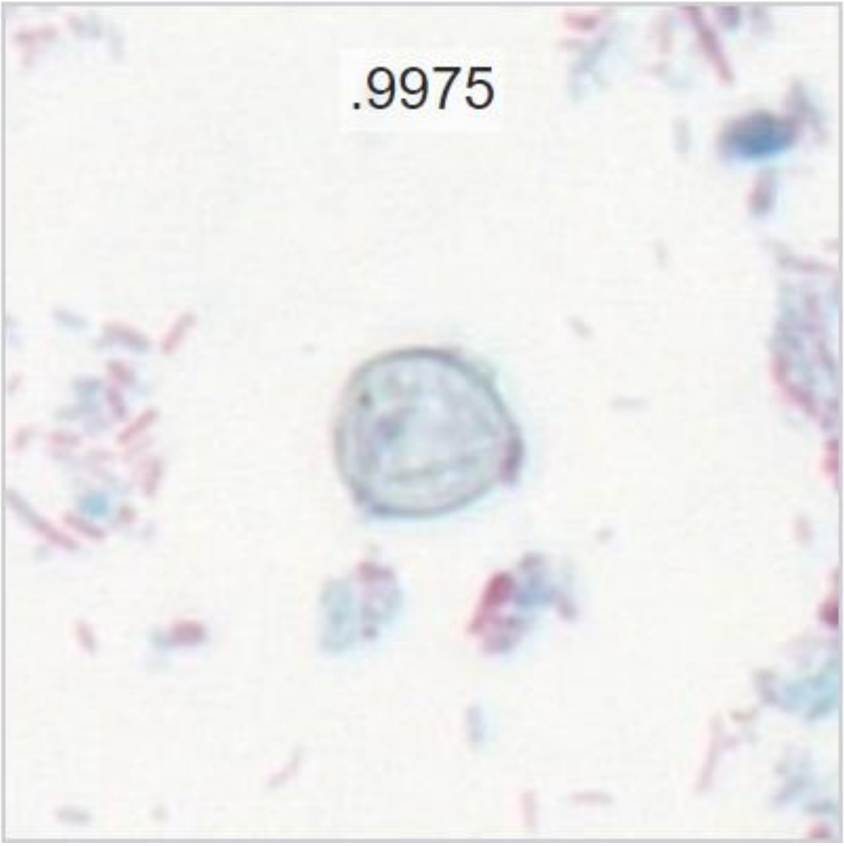


Cyclospora in trichrome NOT *Blastocystis*

Blastocystis



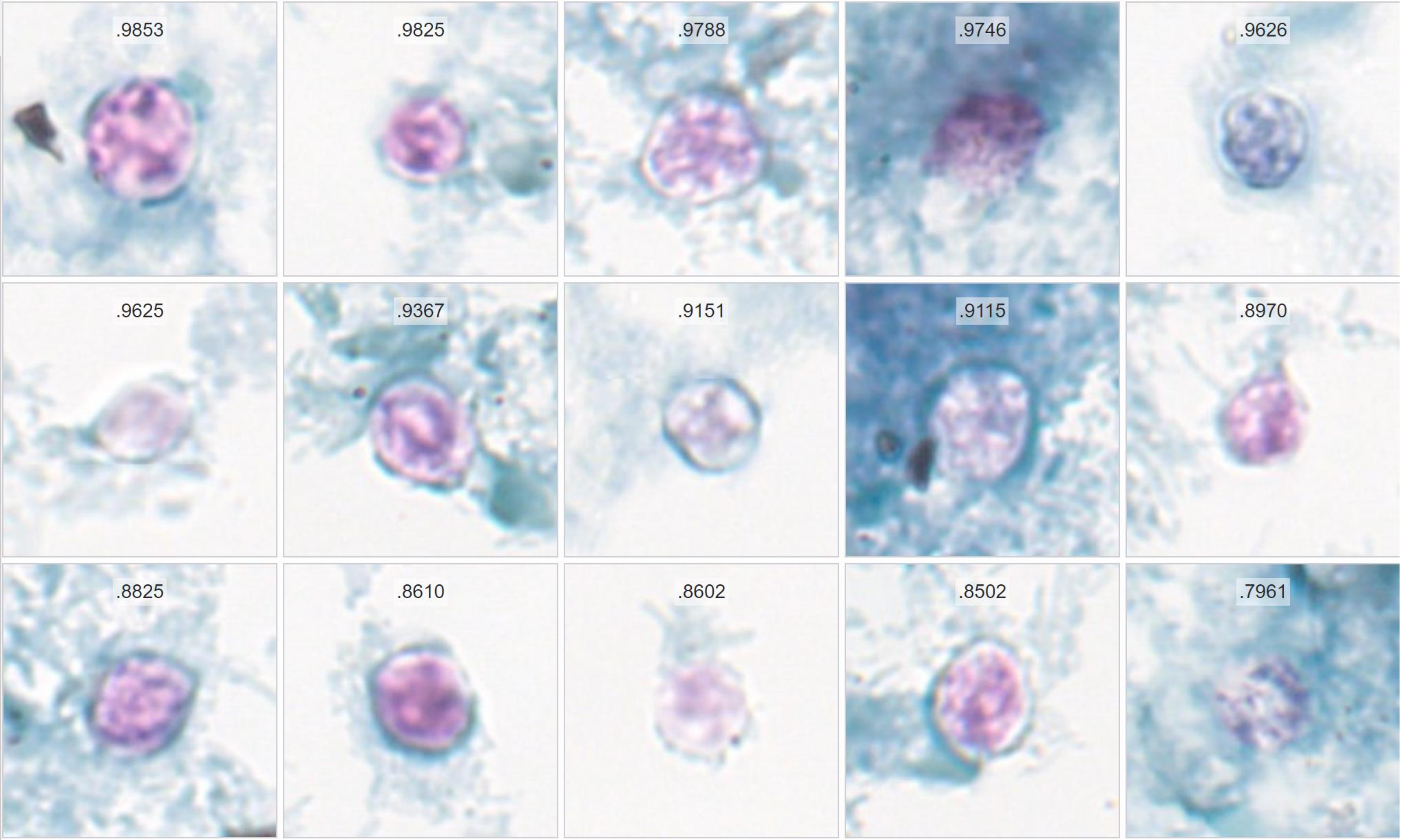
Cyclospora



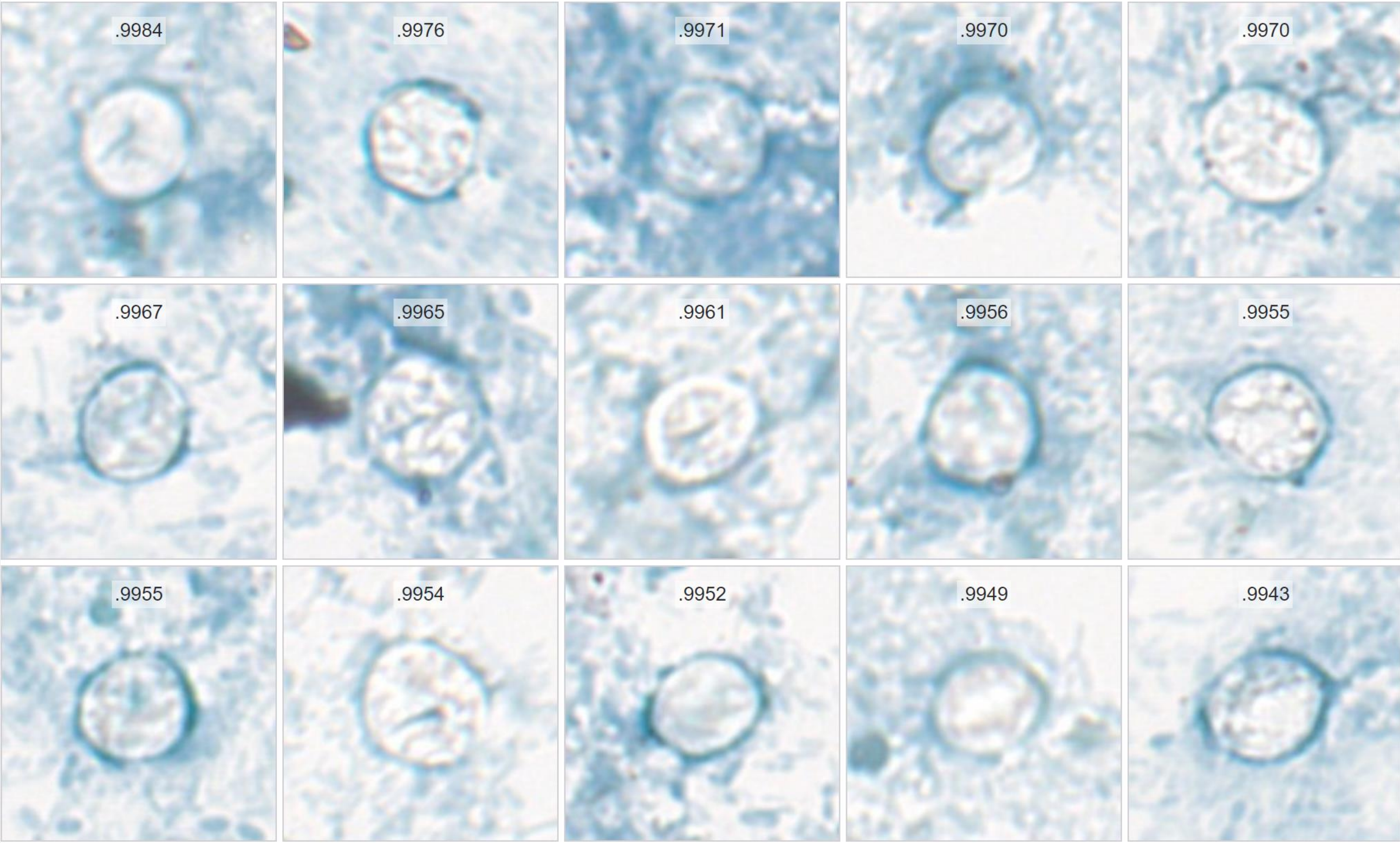
Modified Acid Fast – Future is now

- Traditional stool stain for *Cryptosporidium* and *Cyclospora*
  - Neither retain trichrome stain well
  - “Ghost” forms can be detected by human…what about a model?

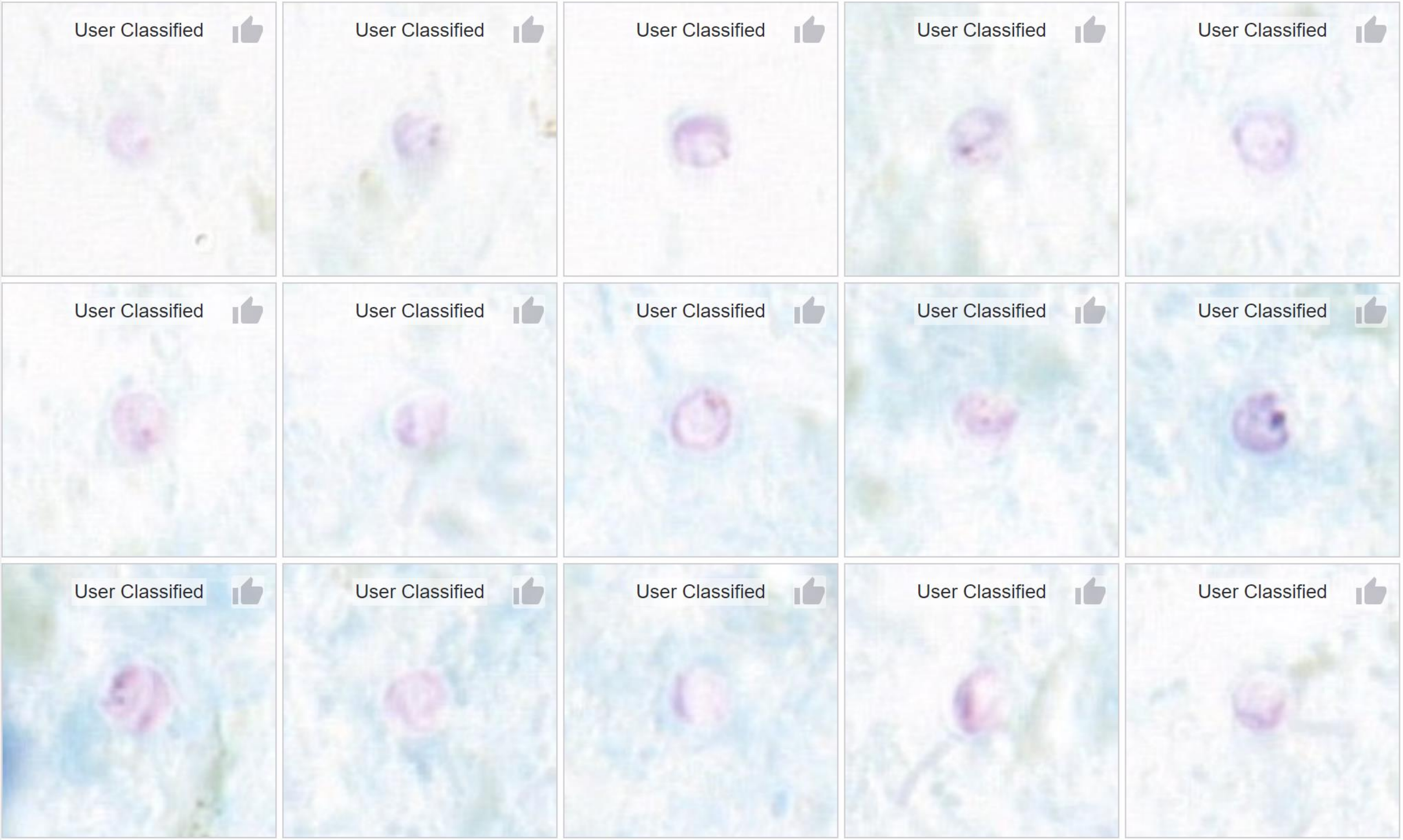
Cyclospora cayetanensis Stained



Cyclospora cayetanensis Ghost



Cryptosporidium sp. Stained



IN CLOSING

Is it better to be efficient or accurate?

Is it too much to ask for both?

With a CNN model, we can be both

Acknowledgments

R&D

- Blaine Mathison
- Jessica Kohan
- Mark Astill
- Dr. Adam Barker
- Dr. Orly Ardon

TechCyte

- Ben Martin
- John Walker
- Ben Cahoon
- Rick Smith
- Rob Fotheringham

ARUP Parasitology Lab

Thank You

ARUP LABORATORIES

Questions?

marc.couturier@aruplab.com