Multi-Disciplinary Teams for ML in Healthcare Applications

BIODS388/BIOMED388

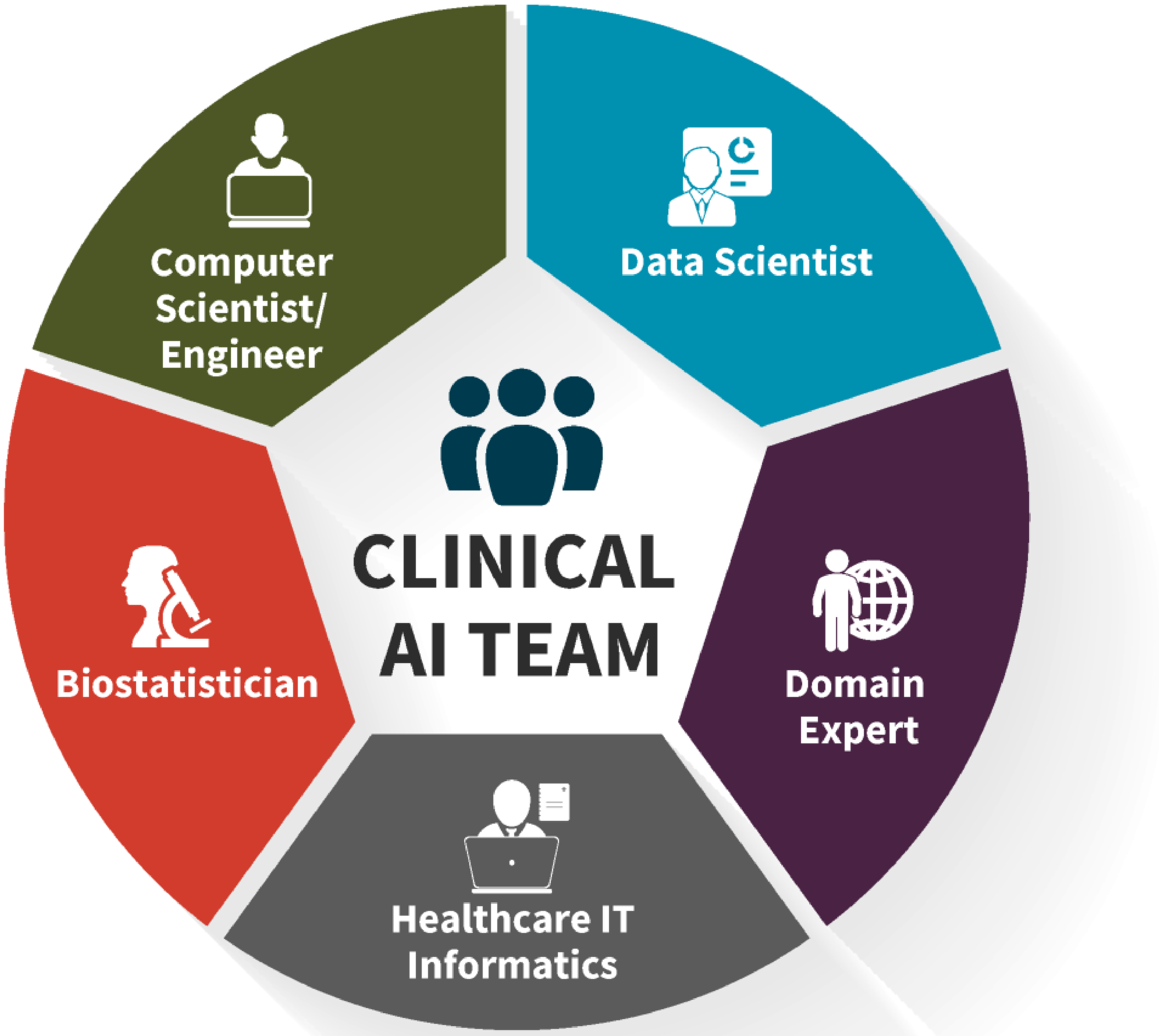
Matthew Lungren MD MPH

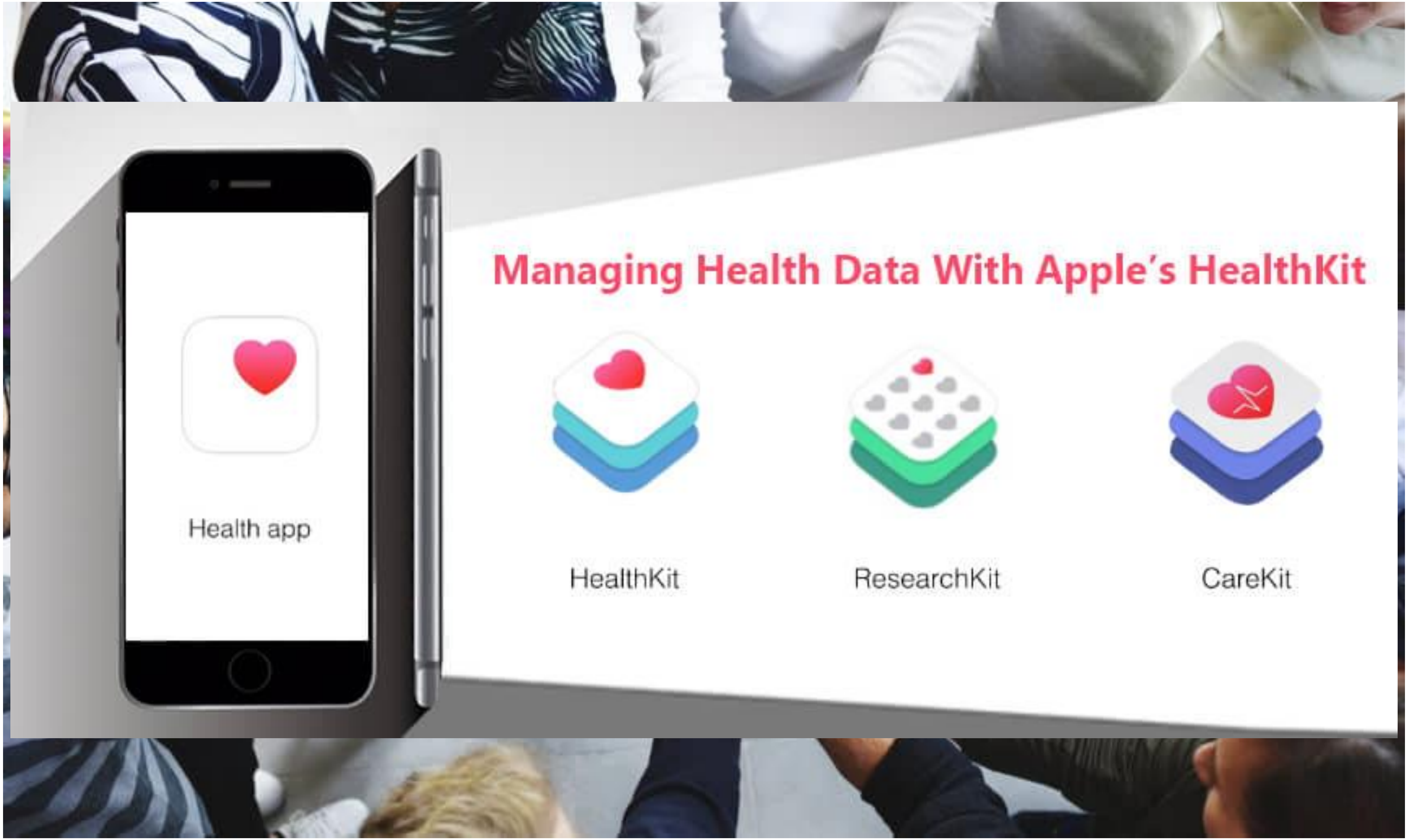
11/12/2020

- Axess is now open for students to provide course feedback until 11:59 PM on Mon, Nov 23, 2020 PST
- Complete feedback at Axess > Student > Course and Section Evaluations

- Instructors only see aggregated, anonymous responses

THER TEAM


CLINICAL AI TEAM
Computer Scientist/ Engineer
Data Scientist
Domain Expert
Healthcare IT Informatics
Biostatistician


Managing Health Data With Apple's HealthKit
Health app
HealthKit
ResearchKit
CareKit


HEALTHCARE ECOSYSTEM
AI developer
COMPANY
tech company
COMPANY
policy maker
GOVERNMENT
regulator
GOVERNMENT
health care system leadership
pharma
COMPANY
medical device
COMPANY
frontline clinician
HEALTHCARE PROVIDER
ethicist
HEALTHCARE PROVIDER
patients
patient caregiver

Ingredient for Success	AI Startups	Established Companies	Healcare Delivery Systems	Professional Societies	Academic Medical Centers	Pharma
Deep technical knowledge	✓	✓	●	✓	✓	✓
High performance computing	✓	✓	●	●	✓	✓
Interdisciplinary teams	●	●	●	✓	✓	●
Ongoing source of labeled images	✗	✗	✓	●	✓	✓
Infrastructure for propective evaluation	●	●	●	●	✓	✓
Market dissemination channel	✓	✓	✗	✗	✓	✓

✓ Available ● Can acquire ✗ Difficult to acquire

Role confusion…?

- *Analytics Data Scientist*
- *Machine Learning Data Scientist*
- *Data Science Engineer*
- *Data Analyst/Scientist*
- *Machine Learning Engineer*
- *Applied Scientist*
- *Machine Learning Scientist…*

Terminology

Data Scientist

Machine Learning Engineer

Terminology

Data Scientist

- Data Curation/mining
- Feature Engineering
- Analytics
- Preliminary model development
- *Biostatistics

Machine Learning Engineer

Terminology

Data Scientist

- Data Curation/mining
- Feature Engineering
- Analytics
- Preliminary model development
- *Biostatistics

Machine Learning Engineer

- Computer science
- Formal code
- Development workflow
- Pipeline development
- Advanced approaches/architectures

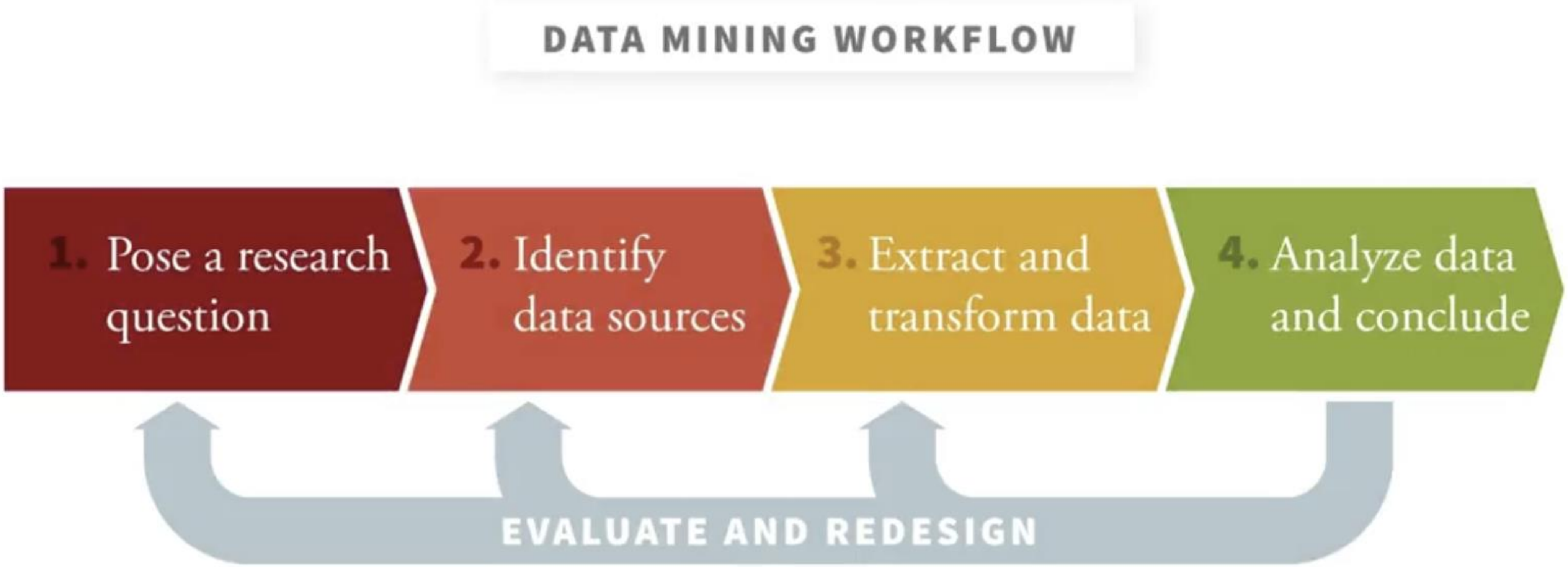
In practice

Data scientists can grab data, throw together an algorithm, and show that it works. But when they hack together a demo, they may take shortcuts. They may create their solution under idealized assumptions about the data inputs and algorithmic outputs.

Machine Learning Engineers ensure it can be packaged and deployed into production and set the infrastructure, get the data pipeline in place, and ensure the data scientists have everything they need to focus on the models

Domain Experts	Category	Examples of Applications
Device product developers, cliniciancs, end users (patients and families)	Health monitoring Benefit/risk assessment	Devices and wearables Smartphone and tablet apps, websites
	Disease prevention and management	Obesity reduction Diabetes prevention and management Emotional and mental health support
	Medication management	Medication adherence
	Rehabilitation	Stroke rehabilitation using apps and robots
Clinician care teams	Early detection, prediction, and diagnostics tools	Imaging for cardiac arrhythmia detection, retinopathy Early cancer detection (e.g., melanoma)
	Surgical proce- dures	Remote-controlled robotic surgery AI-supported surgical roadmaps
	Precision medicine	Personalized chemotherapy treatment
	Patient safety	Early detection of sepsis
Public health program managers	Identification of individuals at risk	Suicide risk identification using social media
	Population health	Eldercare monitoring
	Population health	Pollution epidemiology Water microbe detection

Healthcare administrators	Cybersecurity	Protection of personal health information
Healthcare administrators	Physician management	Assessment of quality of care, outcomes, billing
Geneticist	Genomics	Analysis of tumor genomics
Pharmacologist	Drug Discovery	Drug discovery and design


DATA MINING WORKFLOW
1. Pose a research question
2. Identify data sources
3. Extract and transform data
4. Analyze data and conclude
EVALUATE AND REDESIGN

FINDING PROBLEMS WORTH SOLVING

	SCIENCE	PRACTICE	DELIVERY
CLASSIFY	Finding sybtypes of heart failure with preserved injection fraction	Who might be at high risk for a thromboembolism?	Who is burnt out?
PREDICT	Estimating the disease risk conferred by genetic variations	Which patients are at risk of dying in the next 3-12 months?	Who will be a no show?
ACT/TREAT	XYZ solid tumors can be treated by allogeneic chimeric antigen receptor T-cell By	What is a good second line drug to manage diabetes after metformin?	Request four back up nurses on Wed, for the Ortho OR.

Example questions to address before project start	Considerations
What will the downstream interventions be?	
Who is the target user of the model’s output?	
What are the mechanics of executing the intervention?	
What is the capacity to intervene given existing resources?	
What accuracy is needed and are false positives or negatives less desirable?	
What is the risk of failure and adverse events?	
What is the desired outcome change following intervention?	

Prioritizing projects

Ideal: project has high impact and high feasibility.

- Look for places where prediction drives clinical value
- Look for complicated rule-based scoring systems where we can use ML to learn rules instead of programming them (*and compare to baseline or standard of care)*

Feasibility

- Cost/time for data acquisition
  - How hard is it to acquire data?
  - How much data will be needed?
  - *How expensive is data labeling?*

Metrics

- Cost of wrong predictions
  - How frequently does the system need to be right to be useful?
- Availability of good published work about similar problems
  - Has the problem been reduced to practice?
  - Is there sufficient literature on the problem?
- *Computational resources available both for training and inference*
  - Will the model be deployed in a resource-constrained environment?

Establish a single value optimization metric for the project. Can also include several other metrics (ie. performance thresholds) to evaluate models, but can only ***optimize*** a single metric.

Example:

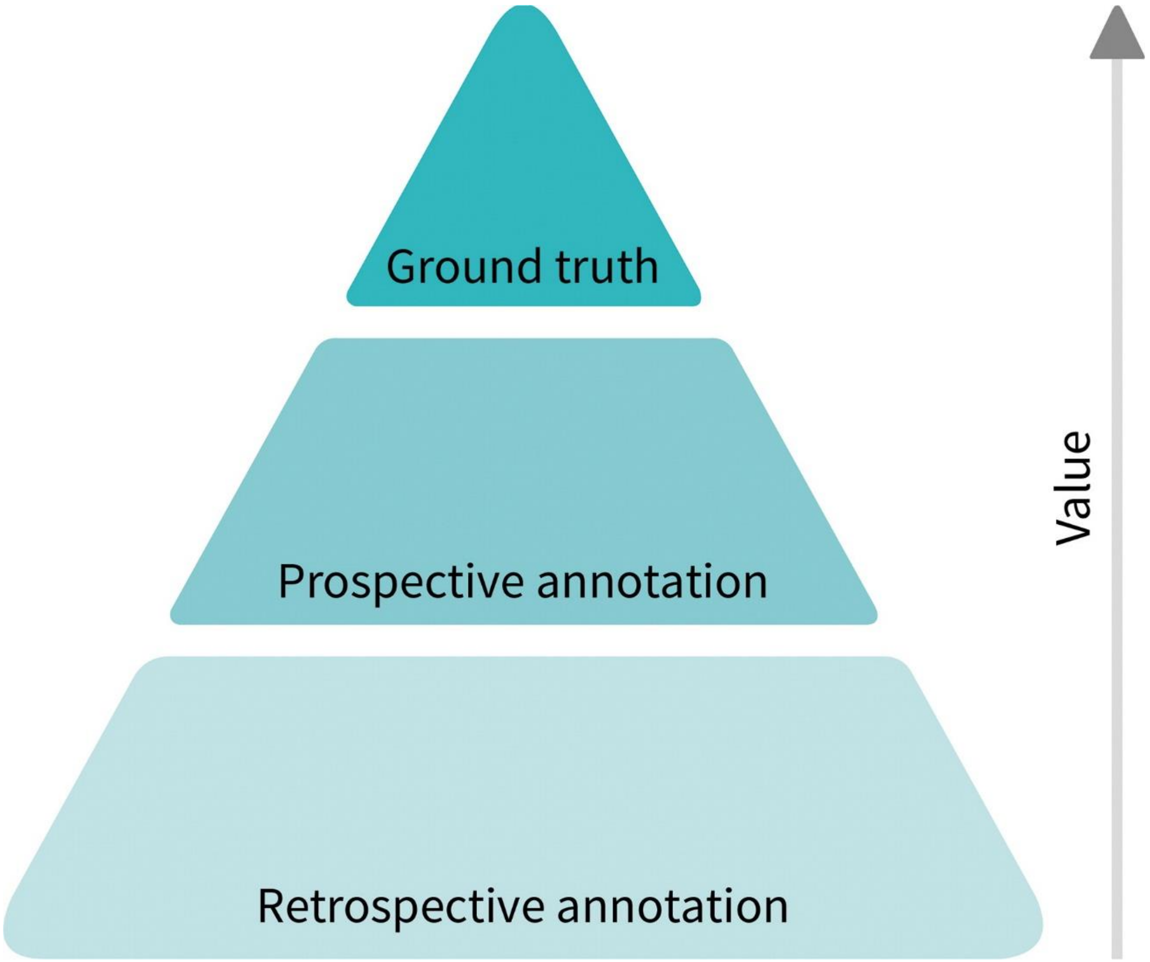
- Optimize for sensitivity (screening)
- Prediction latency under 10 ms
- Model requires no more than 1gb of memory
- 90% coverage (model confidence exceeds required threshold to consider a prediction as valid)

Some teams aim for a “neutral” first launch: a first launch that explicitly deprioritizes machine learning gains, to avoid getting distracted.

— Google Rules of Machine Learning

Label the lane lines

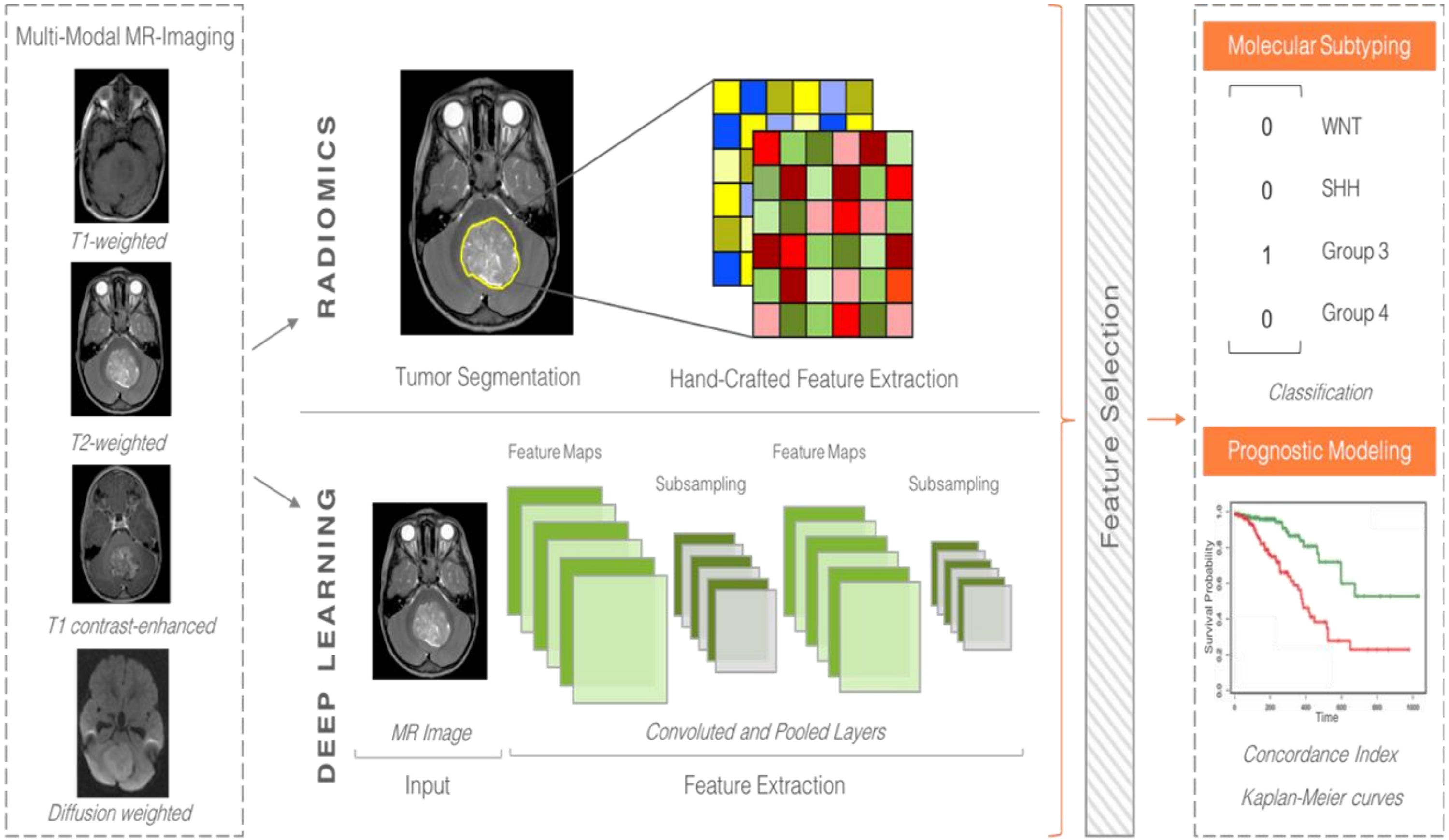
Labeling Medical Data



Preparing Medical Imaging Data for Machine Learning.
Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP.
Radiology. 2020 Apr;295(1):4-15

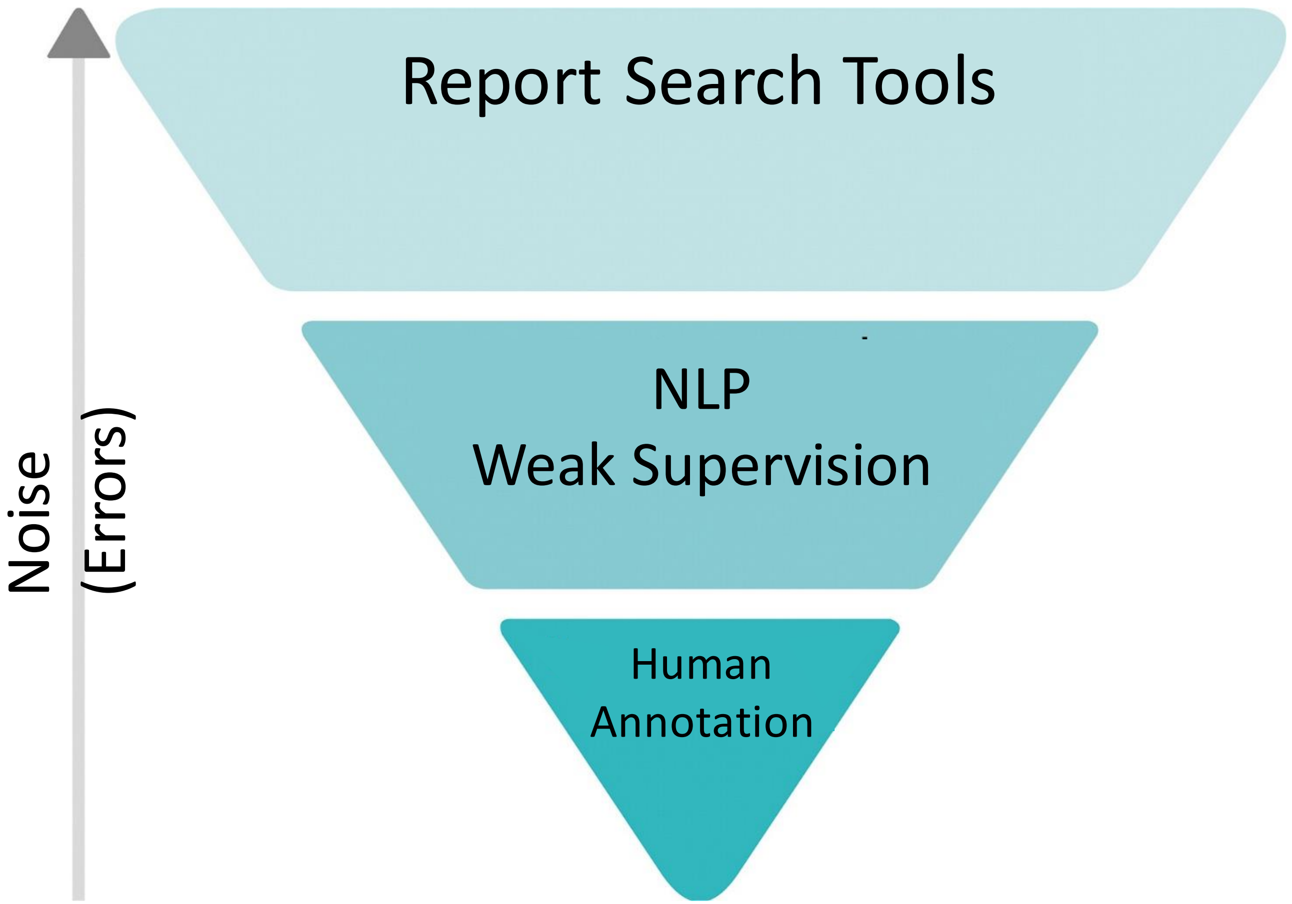
Ground truth labels

- Pathology/Genomics
- Outcomes
- Confirmatory Imaging
- Future diagnoses



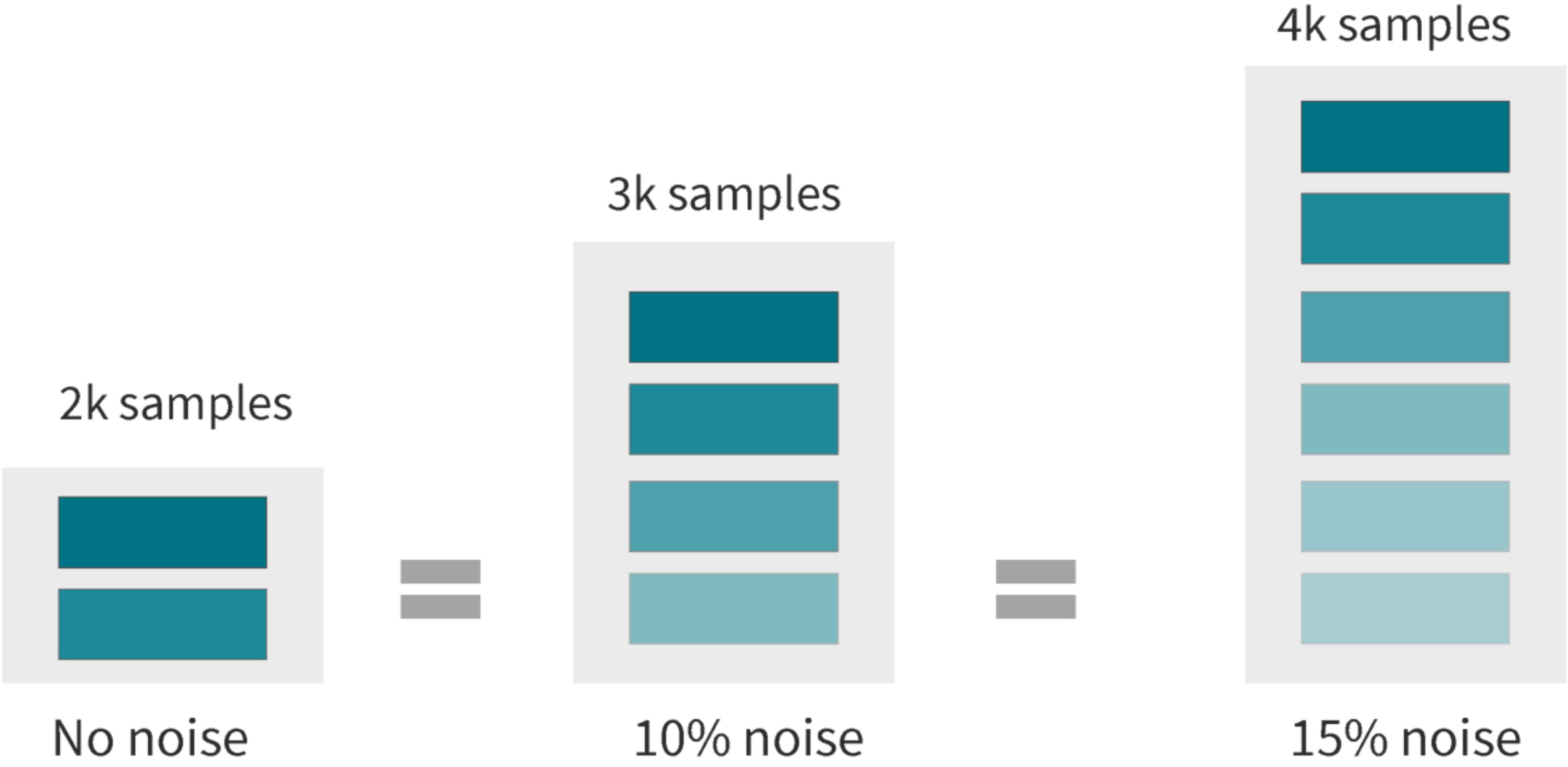
Credit: Dr. Kristen Yeom

Annotation


Report Search Tools
NLP
Weak Supervision
Human
Annotation
Noise
(Errors)

Can you think of ways we could use machine learning to make labeling easier?

OVERCOMING DATA LABEL NOISE WITH DATA VOLUME



Rule of thumb result was with 10% noise you need 50% more data and with 15% noise you need to double the data

Labeling “cheats”

Active learning (augmented labeling)

General approach:

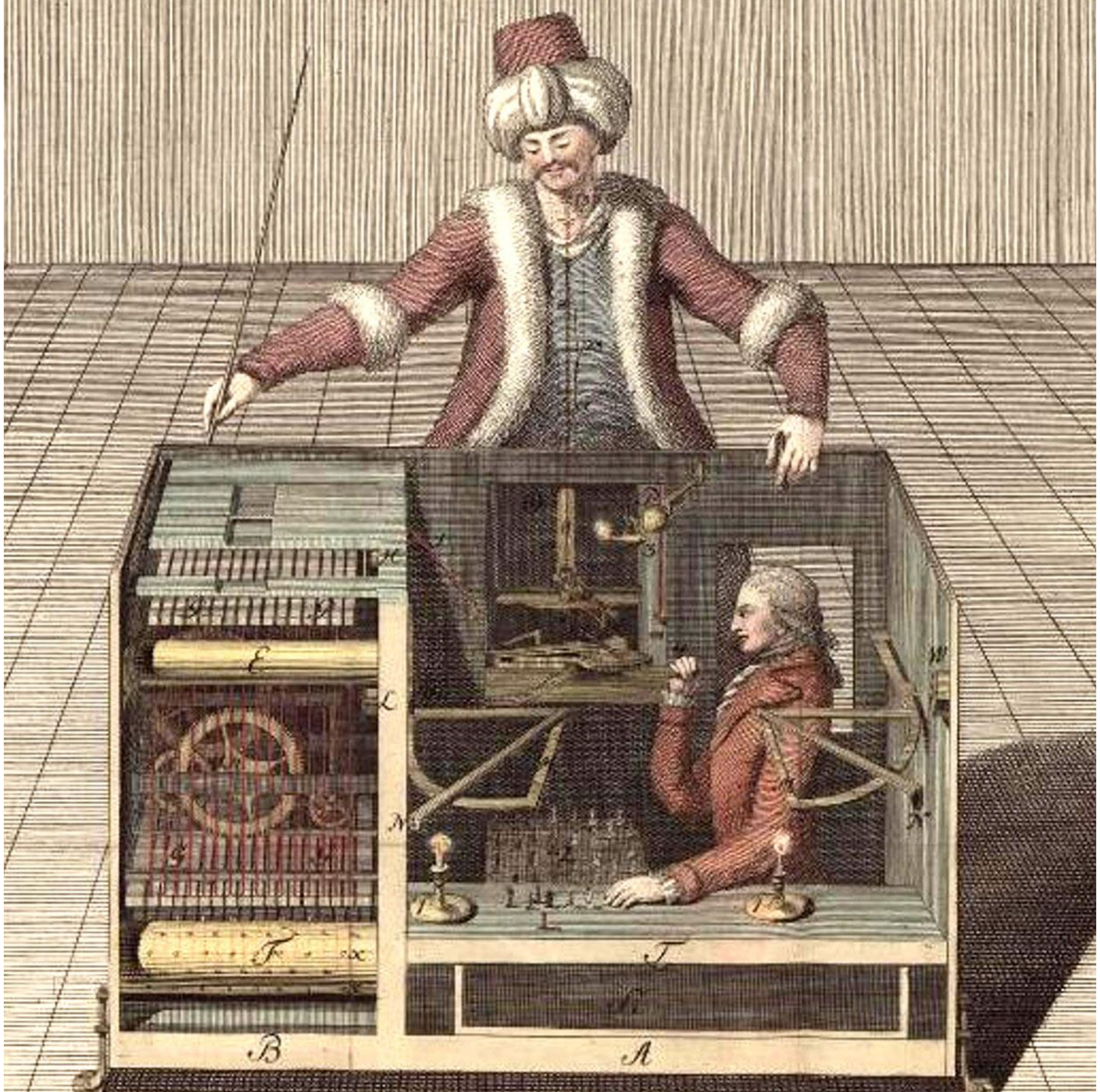
1. Starting with an unlabeled dataset, bu
   acquiring labels for a small subset of i
2. Train initial model on the seed datase
3. Predict the labels of the remaining un
4. Use the uncertainty of the model's pr
   labeling of remaining observations

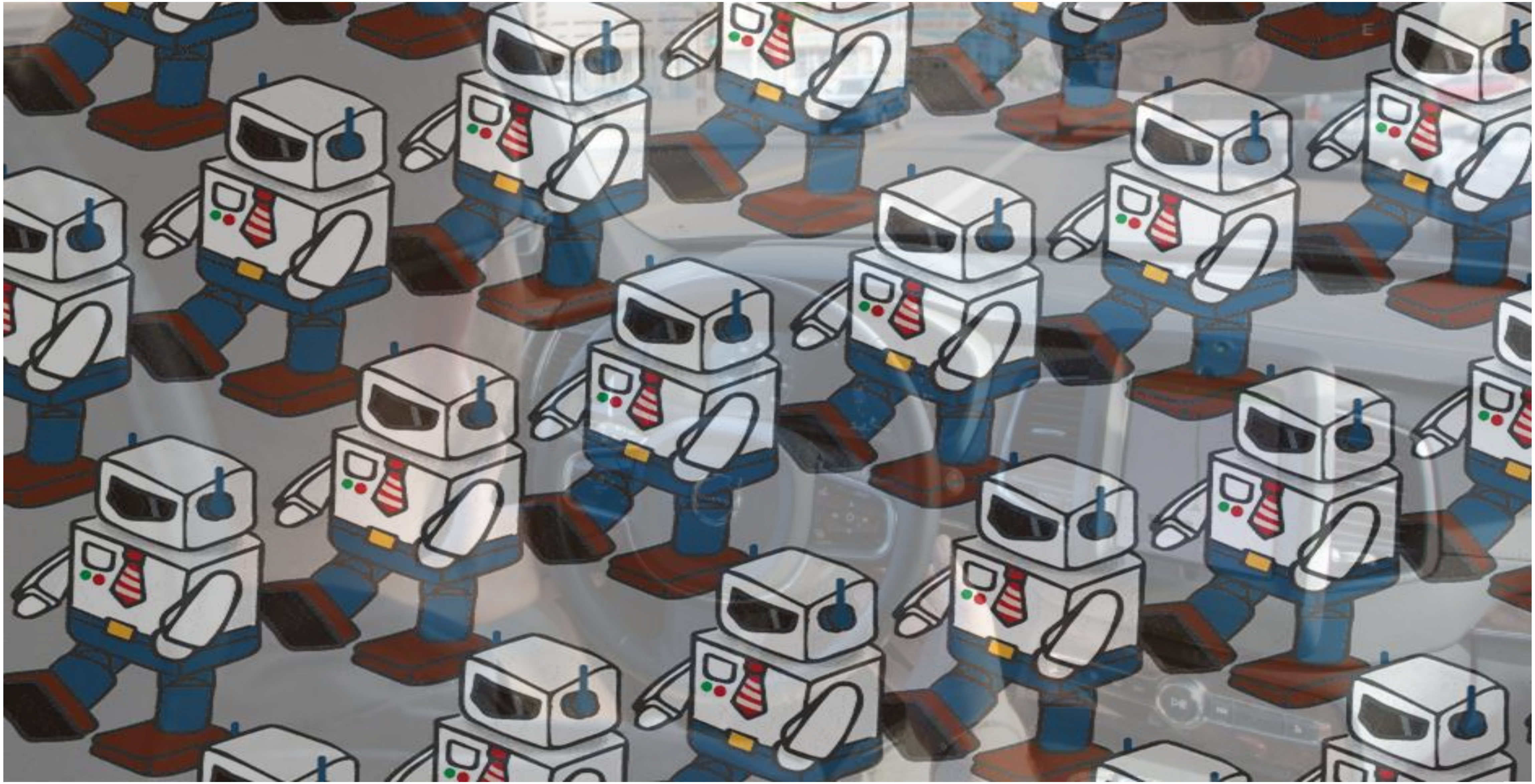
Labeling "cheats"

Leveraging weak labels

Data (or automated labeling methods) can have information which provides a noisy estimate of the ground truth.

Snorkel is an interesting project produced by the Stanford DAWN (Data Analytics for What's Next) lab which formalizes an approach towards combining many noisy label estimates into a probabilistic ground truth.


E
F
T
B
A



Centaur Labs

Home Our Work Blog About Contact Get Started

Accurate and scalable medical data labeling

Leverage our network of medical experts to annotate your text, image and video data

Try it out | Contact us

20,000,000
total labels

50,000
labels per day

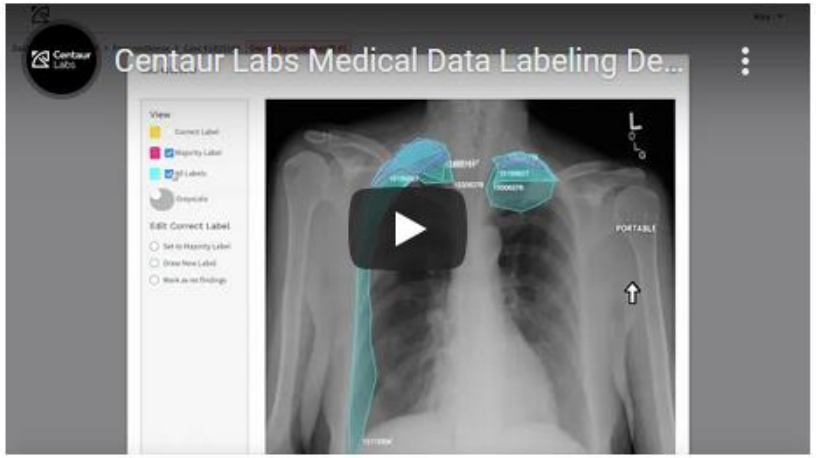
Thousands
of medical experts

How it works

Upload your dataset to our secure cloud and create labeling tasks. When you're ready, launch your tasks to our network of medical experts

See results within days, not weeks or months.

Try it out

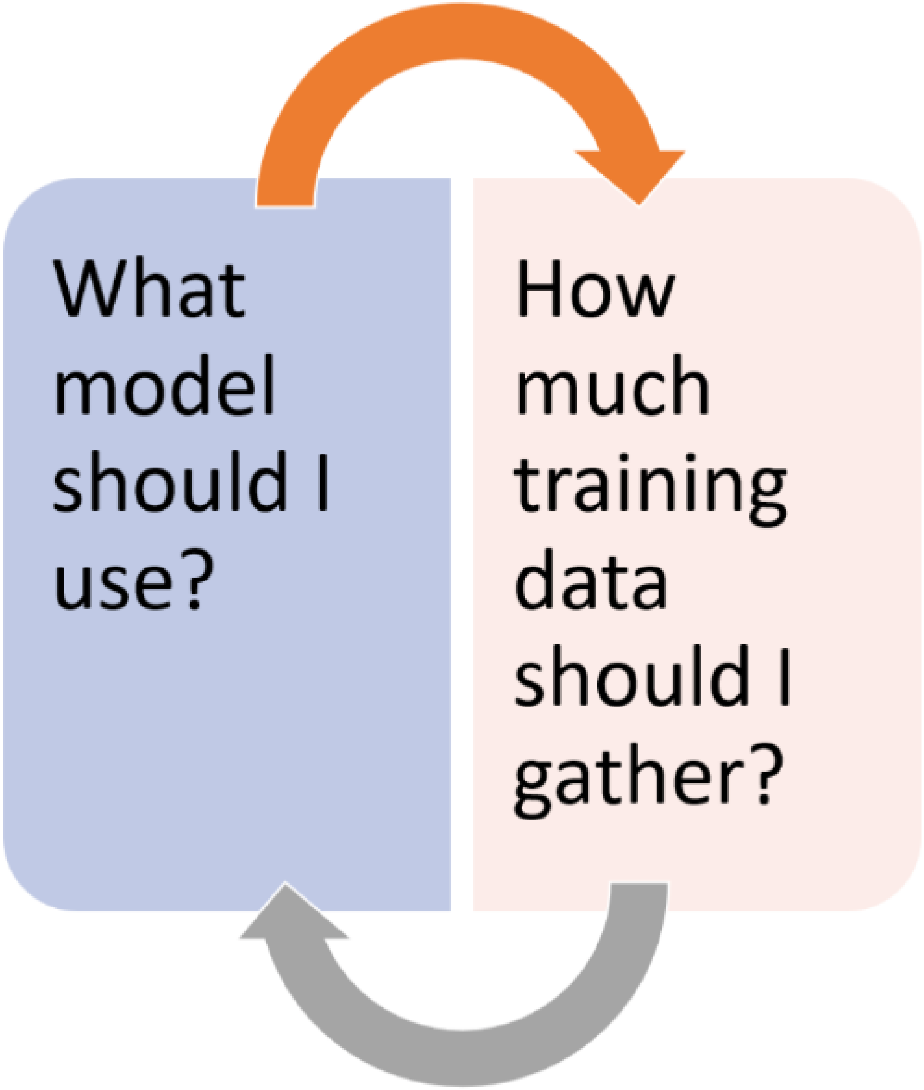


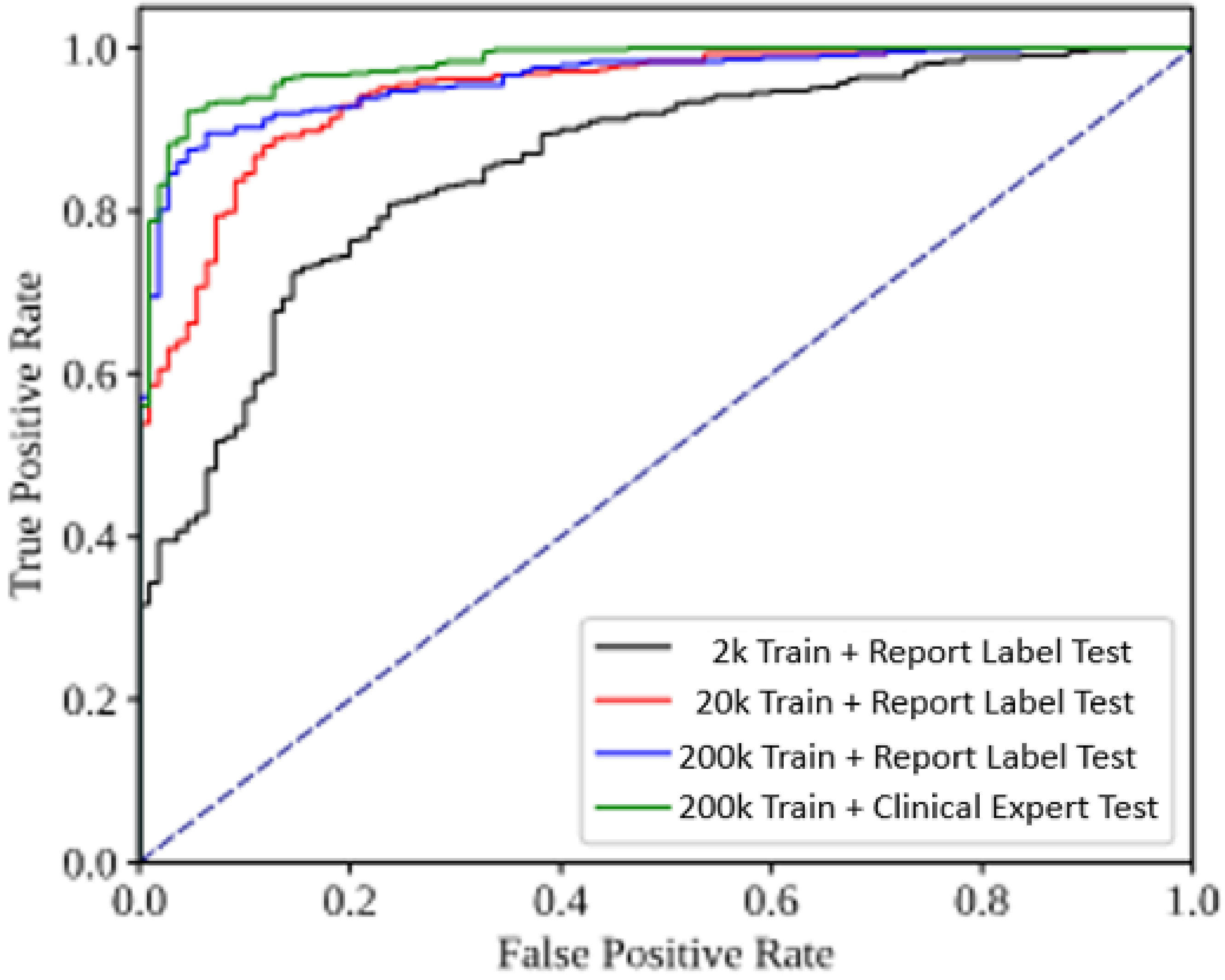
We guarantee satisfaction.

We won't rest until you're 100% satisfied with your labeled datasets.

Start simple..

- Simple baselines tc el complexity.
- If your problem is v to approximate a baseline based on asks/datasets.
- Try to estimate hur :he given task.
- Check to see if moc
- Understand how m




1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
2k Train + Report Label Test
20k Train + Report Label Test
200k Train + Report Label Test
200k Train + Clinical Expert Test

Troubleshooting

Why is your model performing poorly?

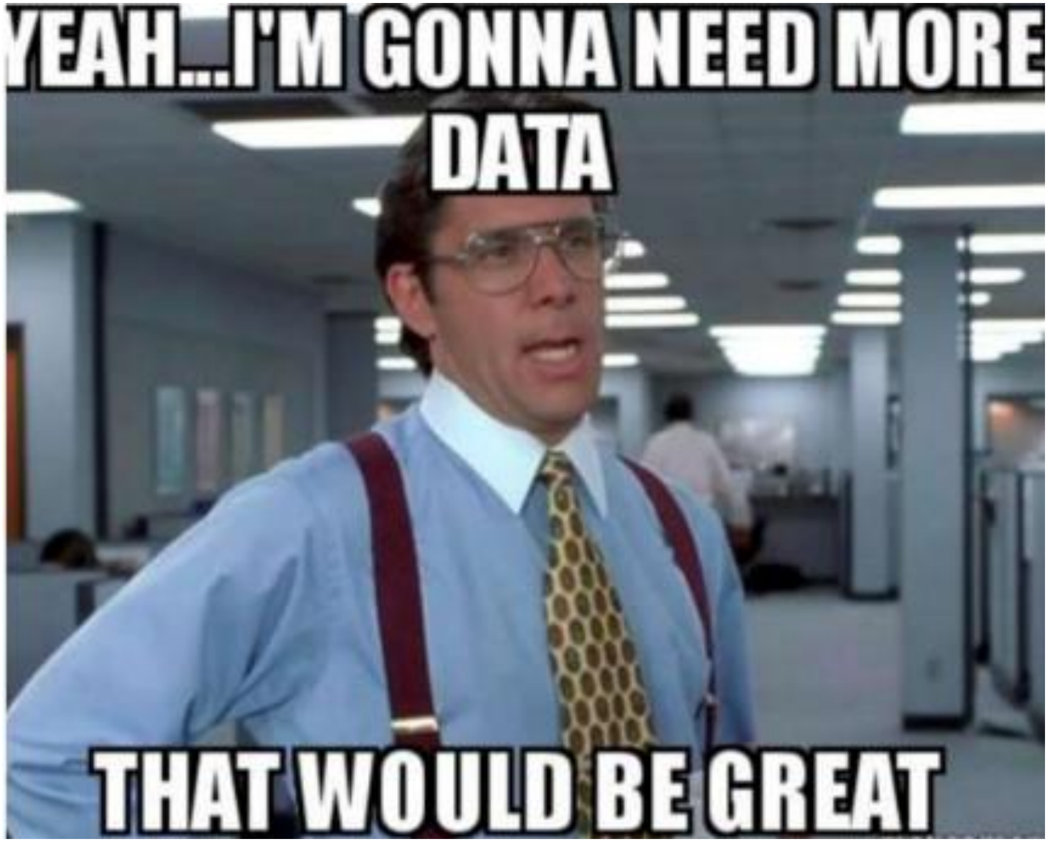
- Manual review of all discrepancies (false positive and false negatives)
- User adoption and potential automation bias or lack of trust
- Distribution shift?

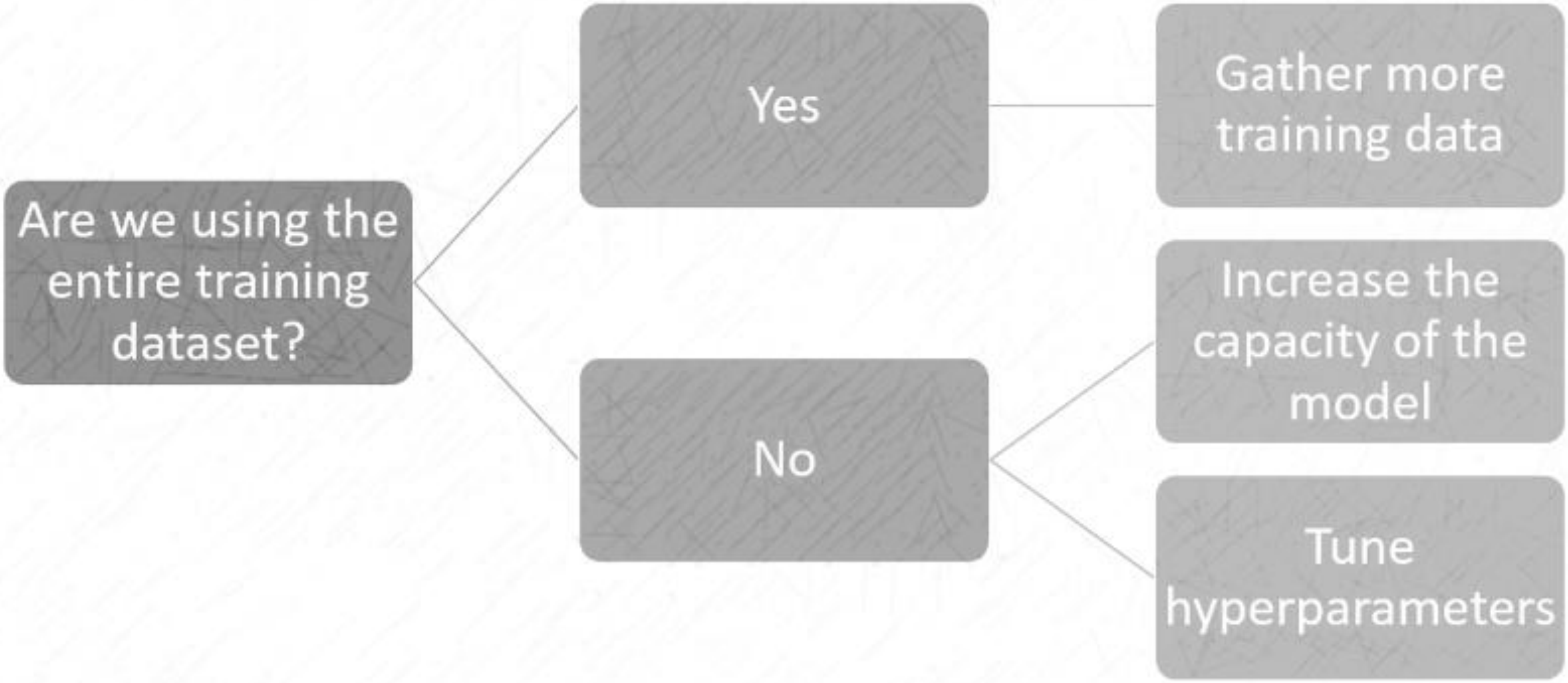
Underfitting

- Increase model capacity
- Reduce regularization
- Error analysis
- Choose a more advanced architecture (closer to state of art)
- Tune hyperparameters
- Add features

Overfitting

- Add more training data
- Add regularization
- Add data augmentation
- Error analysis
- Tune hyperparameters
- Reduce model size





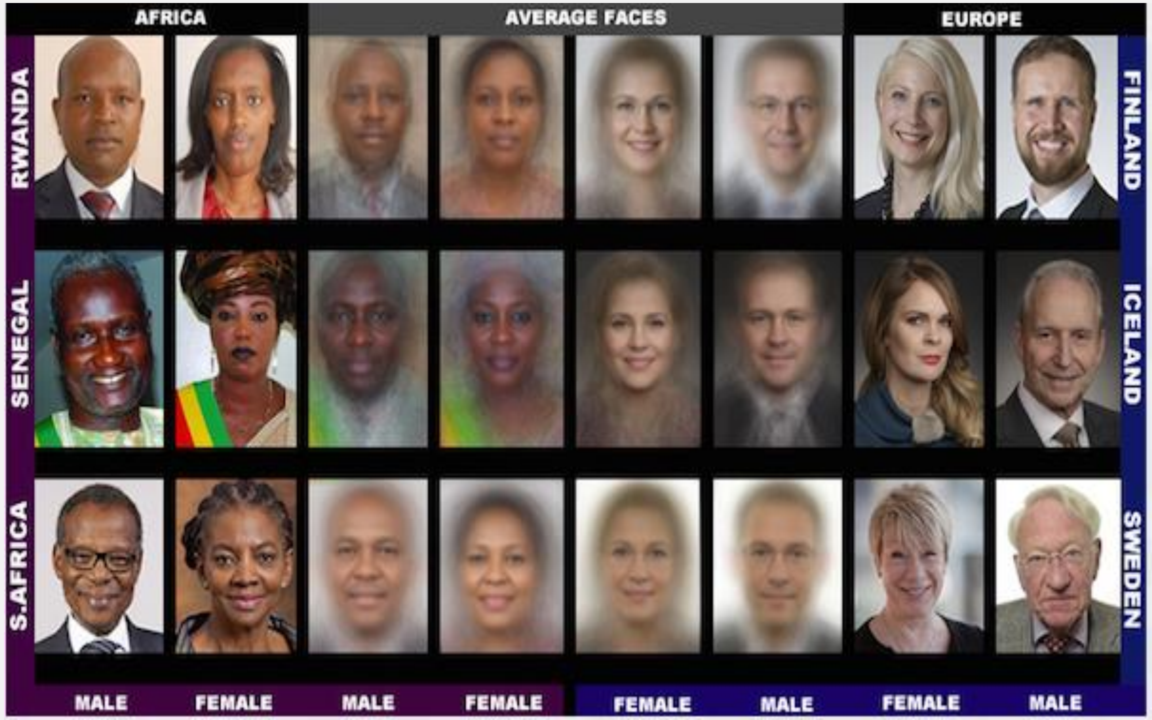
Are we using the entire training dataset?
Yes
No
Gather more training data
Increase the capacity of the model
Tune hyperparameters

Unconscious

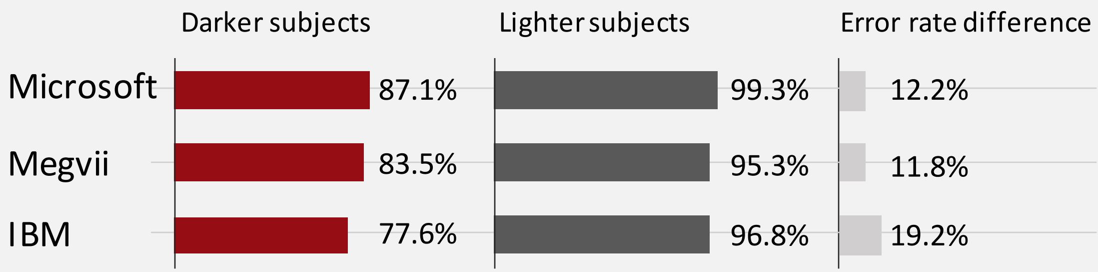
Computational

Cognitive

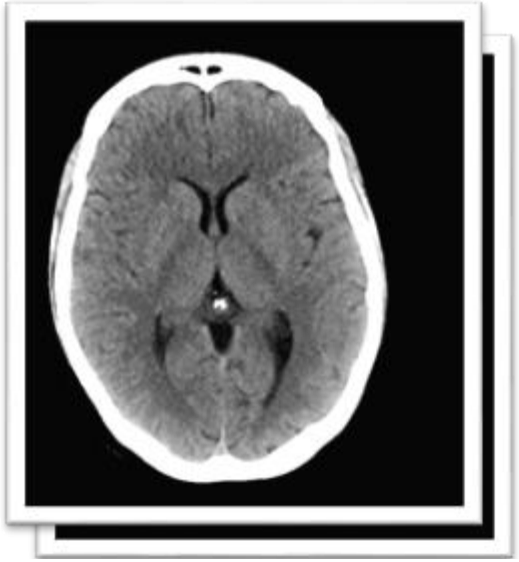
Automated systems are not inherently neutral because data is not inherently neutral



Accuracy rate for gender identification, by skin color



Bleed



No Bleed

the AI Radiology Assistant

About

NOT FOR MEDICAL USE. This is a prototype system for diagnosing chest x-rays using neural networks. All processing is done on your device and images are not sent to the server. If you continue you assume all liability when using the system. A neural network model (~150mb) will be downloaded to your browser.

By Joseph Paul Cohen, Paul Bertin, and Vincent Frappier 2019

Made by Mila Medical

Done in 6951ms

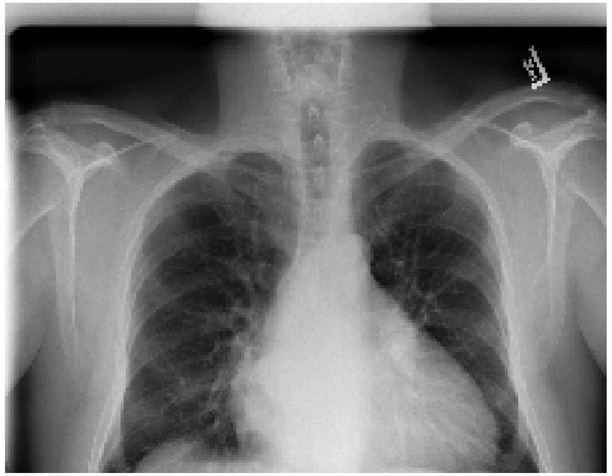
Process an image locally:

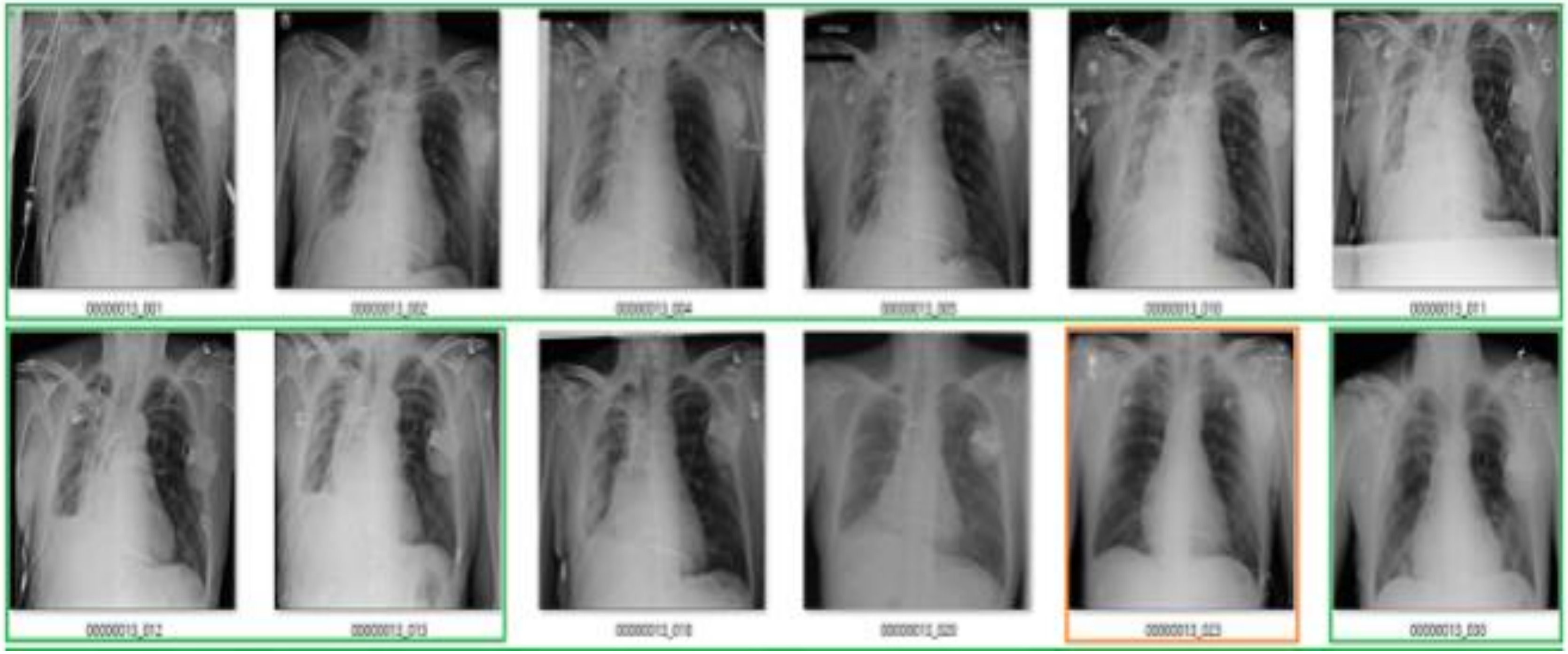
Choose File no file selected

Download example files

Example Image (00000001_001-Cardiomegaly-Emphysema.png)

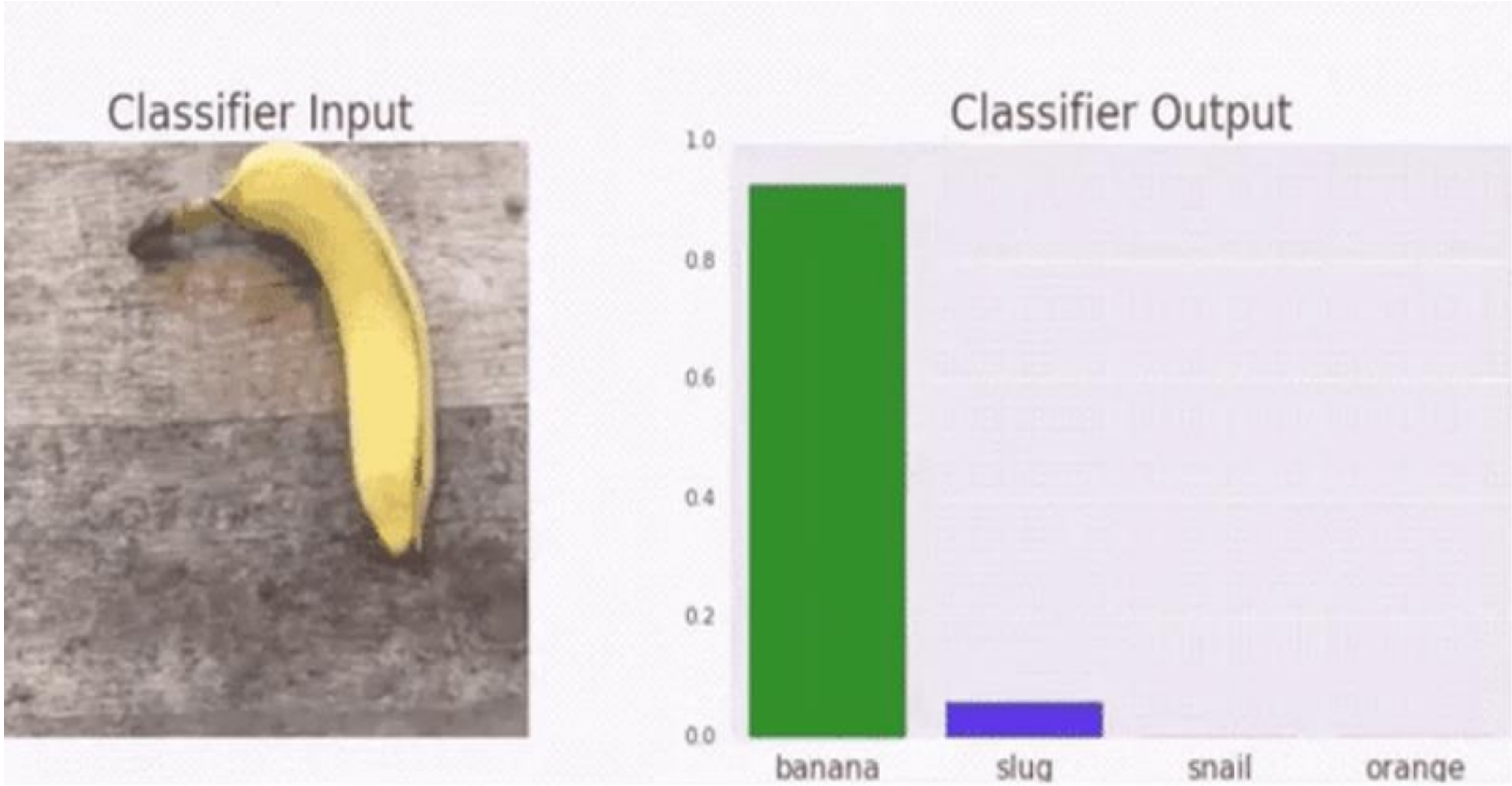
Input Image	Predictive image regions	Disease Predictions
	Heatmap of image regions which influence the prediction.	Risk of a disease.





> AI system will identify **chest drains** instead of **pneumothoraces**

Source: https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/



Robust Physical-World Attacks on Machine Learning Models, Evtimov et al. https://arxiv.org/abs/1707.08945

Normal

Pneumothorax

Human behavior is related to the perceived chance of error …. and over time observe a natural “risk homeostasis”

In 1967 Sweden changed from left hand to right hand traffic which led to **reduction** in the traffic accident rate.

About a year and a half later the accident rate returned to the **same rate** as before the changeover.

After “childproof” medicine vials were introduced the annual number of accidental poisoning went down

But within a few years the number **increased** therefore the change and has steadily risen since


0.1
LIMIT 50
Speed Limit 500
EXIT 46A
Central Expwy
LEFT EXIT
LIMIT 55
Speed
40
I-30
Arrival
12:51
GARMIN

Coachclass
82609
CHIEF

Driver follows GPS into sand

AN 80-year-old driver has crashed off a motorway into a huge pile of sand, ignoring several warning signs because his car's GPS told him to keep going.

Reuters MARCH 17, 2009 1:04AM

Automation bias - humans trust output of computer automated systems and adjust behavior to assume risk is lower

NJ Ma

By Brian Thompson

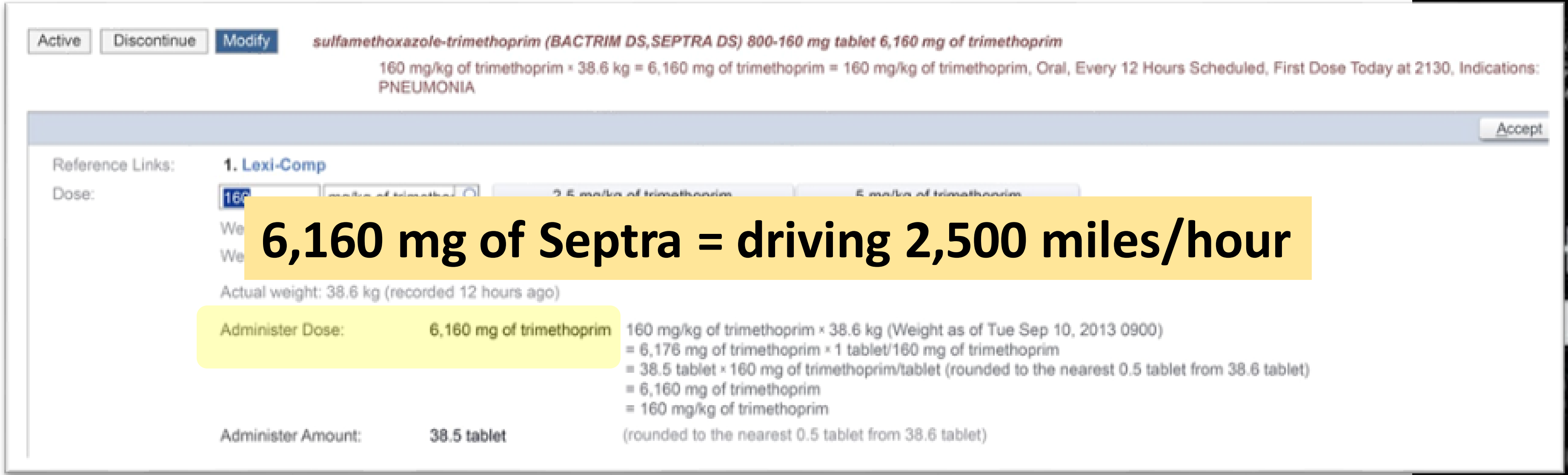
Man follows GPS directions onto train tracks, car

More errors occur with automated systems (when wrong) vs without systems due to the lower perceived chance for error

Death by GPS

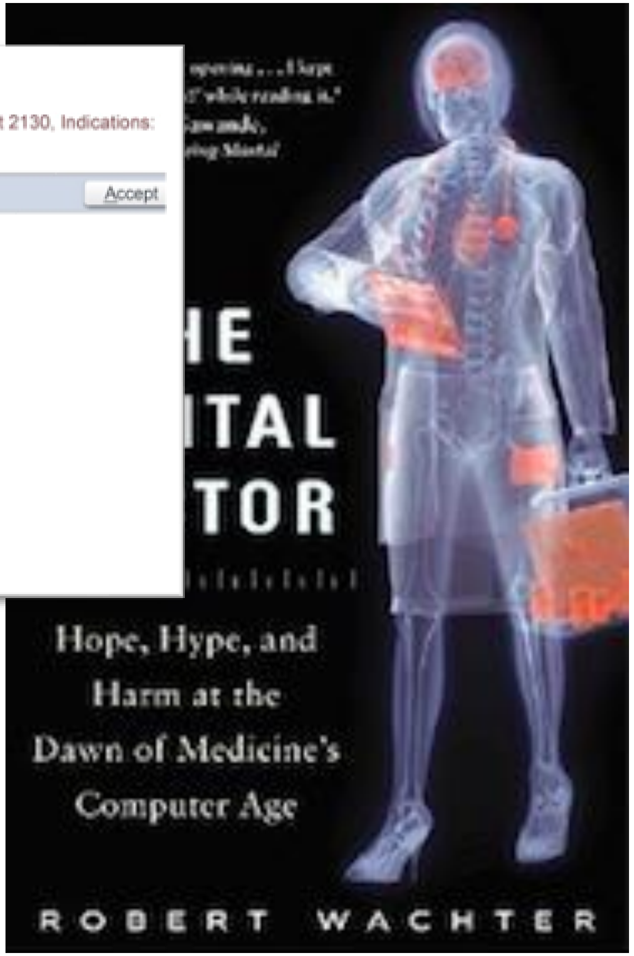
Why do we follow digital maps into dodgy places?

GREG MILNER - 5/3/2016, 4:00 AM



6,160 mg of Septra = driving 2,500 miles/hour

UCSF School of Medicine



Automation bias can be mitigated by providing ”confidence” with recommendation systems

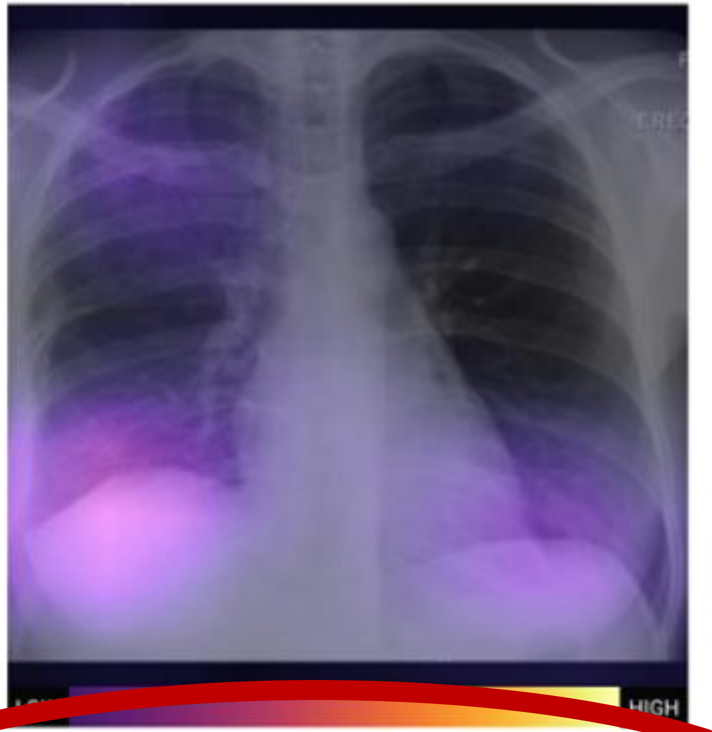
Automation Bias in Intelligent Time Critical Decision Support Systems M.L. Cummings https://arxiv.org/abs/1804.05296
Presenting System Confidence Information to Support Trust Calibration and Adaptive Function Allocation

Patient's Clinical Information

Variable	Value	Reference Range
Age	31	NA
Sex	Male	NA
Temperature (Celsius)	38.8 ↑	36.1-37.2
Oxygen Saturation (Percent)	100	95-100
Haemoglobin	8.6 ↓	13.5-17.5
WBC Count	9.08	4.5-11
CD4 Count	390 ↓	500-1500
Previous TB	no	NA
HIV status	Yes	NA
Current ART Status	Yes	NA
Cough	Yes	NA
Cough Duration (day(s))	Unknown	NA

Patient's X-ray

Regions Consistent with TB



Positive

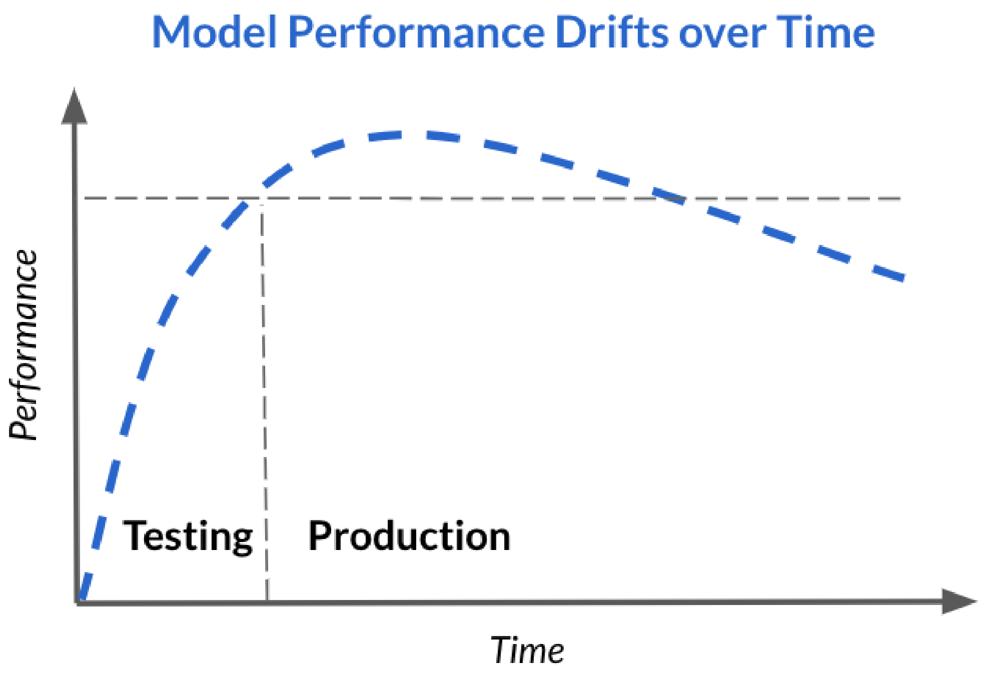
So you think you are ready to deploy….?

Hidden Technical Debt of Machine Learning Systems

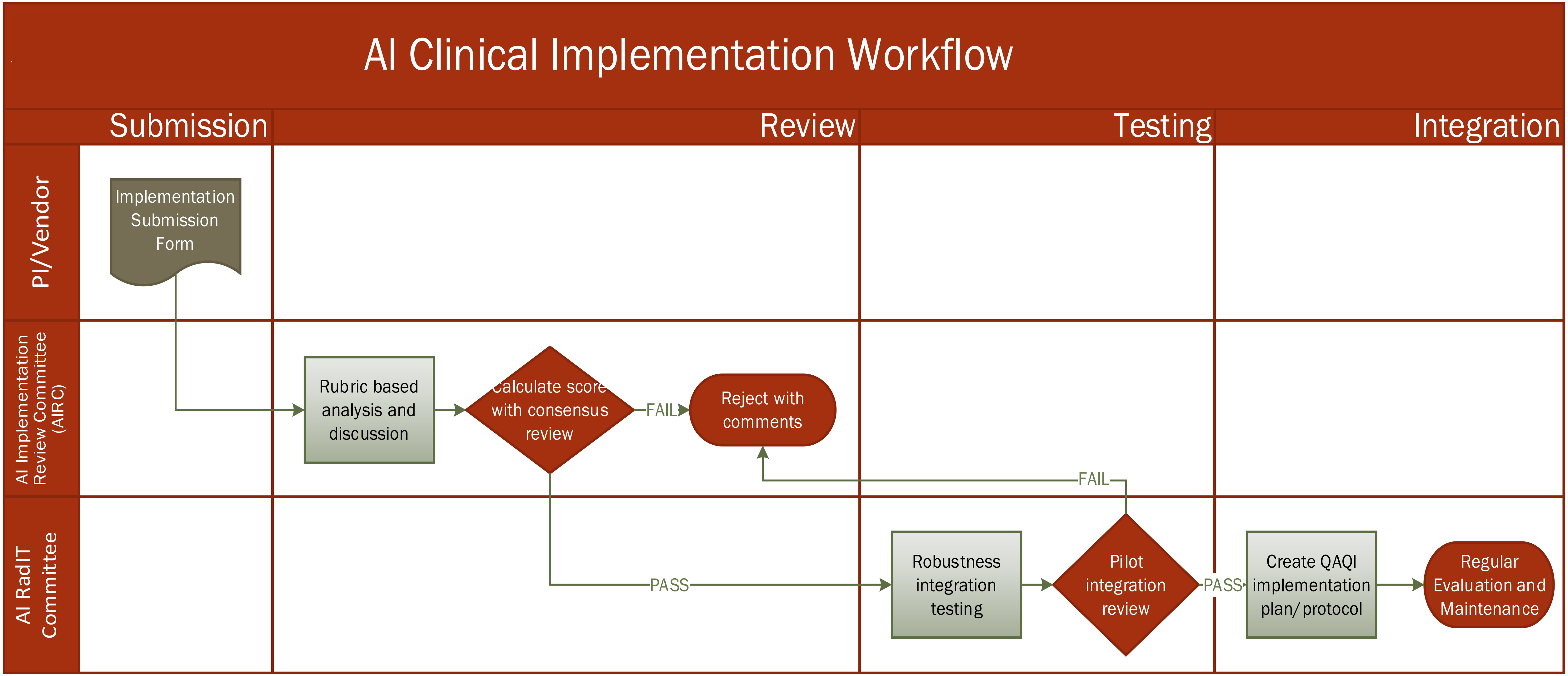
"CACE" principle

Access control

Model performance will decline over time.



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf



AI Clinical Implementation Workflow
Submission
Review
Testing
Integration
PI/Vendor
AI Implementation Review Committee (AIRC)
AI RadIT Committee
Implementation Submission Form
Rubric based analysis and discussion
Calculate score with consensus review
FAIL
Reject with comments
PASS
Robustness integration testing
Pilot integration review
FAIL
PASS
Create QAQI implementation plan/protocol
Regular Evaluation and Maintenance

AI Implementation Review Committee (AIRC)

Stanford AIMI

- Scoring Rubric for Evaluation
- Utility Analysis
- Risk Estimation
- Evidence Review
- Technical Readiness
- Clinical Readiness
- Value/Economic Review

Role	Expertise
Clinical	Radiologist
Technical	Rad IT
Clinical Practice Manager	Practice Manager
AI Technical Lead	Computer Science
Quality Assurance	Safety/Quality
Operations Coordinator	Hospital IT
Finance Manager	Accounting

Intake Submission Form

Stanford AIMI

Data Source	Deployment	Origin	Purpose	Automation Level
Raw/Source	Standalone System • on prem • Cloud • Hybrid • Edge	Vendor	Clinical Trial	1. Data Presentation (image reconstruction, worklist, measurement, annotation)
Pixel Data	Integration with existing platform • PACS • EMR • Powerscribe • AI platform • Edge	Internal Research	Clinical Integration	2. Decision Support (risk score, abnl highlight)
EMR Data		External Research	Other	3. Conditional Automation (prelim diagnosis/report)
Other/Hybrid		Other		4. Full Automation (bypass rad, direct to clinician result)

Scoring Rubric

Stanford AIMI

Clinical Readiness	Technical Readiness	Evidence
0 – No existing clinical workflow	0 – No existing technical infrastructure	0 – none / pilot data
1 – Major clinical workflow modification	1 – Major technical modification	1 – Level III/IV limited cohort
2 – Minor clinical workflow modification	2 – Minor technical modification	2 – Level II prospective clinical evaluation
3 – No significant clinical workflow modification	3 – All technical infrastructure in place	3 – Level I evidence for primary outcome

Value	Data Security Risk	Clinical Utility
0 – Costs unknown	0 – unknown	0 – Low volume low acuity
1 – Net negative	1 – High	1 – moderate volume or moderate acuity
2 – Net neutral	2 – Low	2 – high volume or high acuity
3 – Billing code	3 – Full compliance / Lowest risk	3 – high volume and high acuity