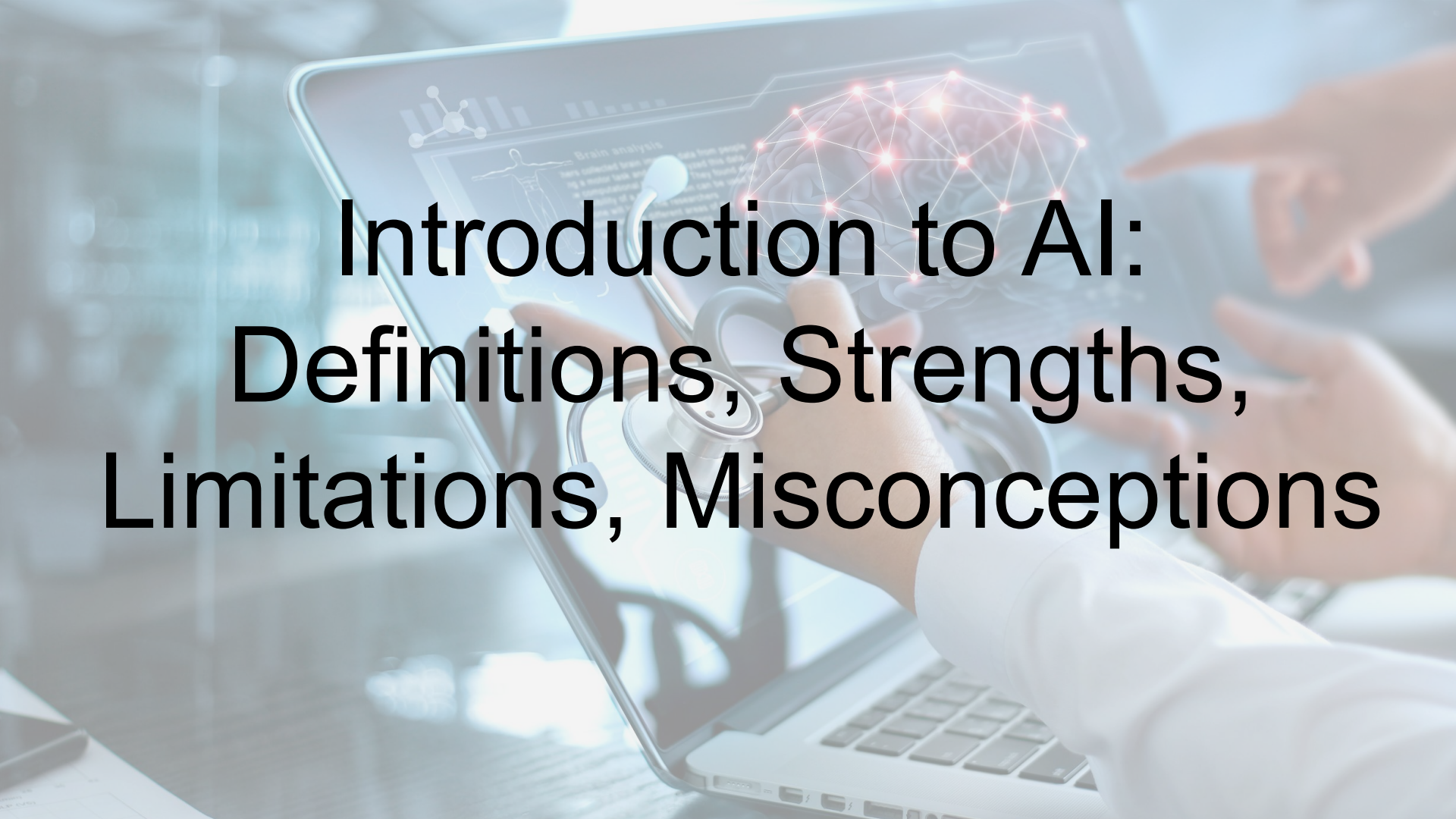
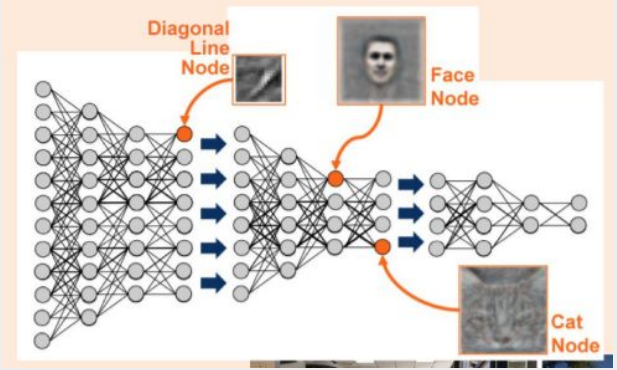



# Introduction to AI: Definitions, Strengths, Limitations, Misconceptions

# Objectives

To provide medical residents with knowledge in AI to

- Describe fundamental concepts in machine learning
- Understand clinical implications of AI literature
- Identify strengths and limitations of AI in radiology




(you should be able to interpret this after)

Content creators: Ricky Hu, Arsalan Rizwan, Zoe Hu

Thanks to Dr. Kwan and Dr. Chung for supervising

# Lay Terms

- Be comfortable with reading AI papers

- Have content to reply questions on AI in radiology (e.g. will AI replace physicians???)

- Apply detailed data science best practices to any data-based project!

**Feel free to bring any questions/challenges!**

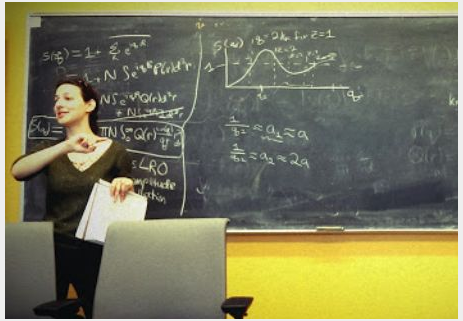
| | Session 1 | Session 2 | Session 3 |
|---|---|---|---|
| **Didactic** | Definitions, Myths, ML Pipeline, Training and Testing | Data Preprocessing, ML Models (LogReg, Random Forest, comparison of models) | Strengths + Limitations of Models, Neural Networks, Modern Techniques |
| **Case Study** | Case Studies:<br>1. AI vs. Rads performance<br>2. AI Usage in ED (TBI Algorithm) | Case Studies:<br>1. MIT COVID “Cough” Algorithm<br>2. Radiomics + Machine Learning | Case Studies:<br>1. ChexNet<br>2. U-net, nnU-net<br>3. Modern CNNs |
| **Programming** | Programming:<br>Exploratory Data Analysis, Feature Selection | Programming:<br>Decision Tree, Cross Validation | Programming:<br>Neural Networks, Convolutional Neural Networks |
| | Q&A | Q&A | Q&A |

# Potential Roles

## User

## Analyst

## Developer

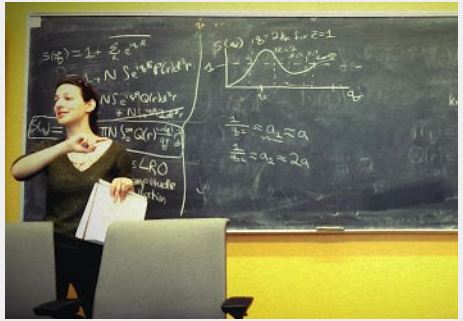


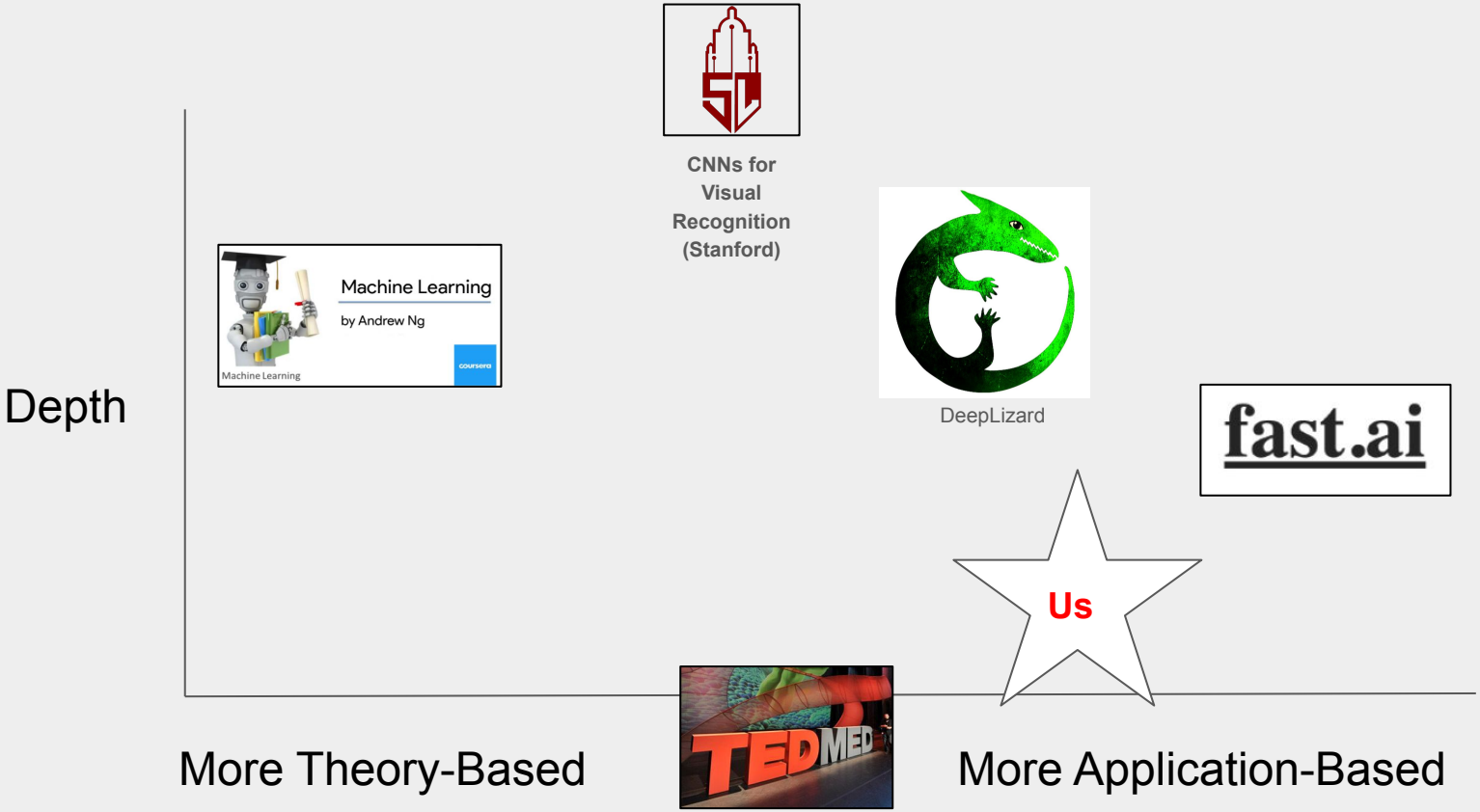
# Potential Roles

## User

## Analyst

## Developer

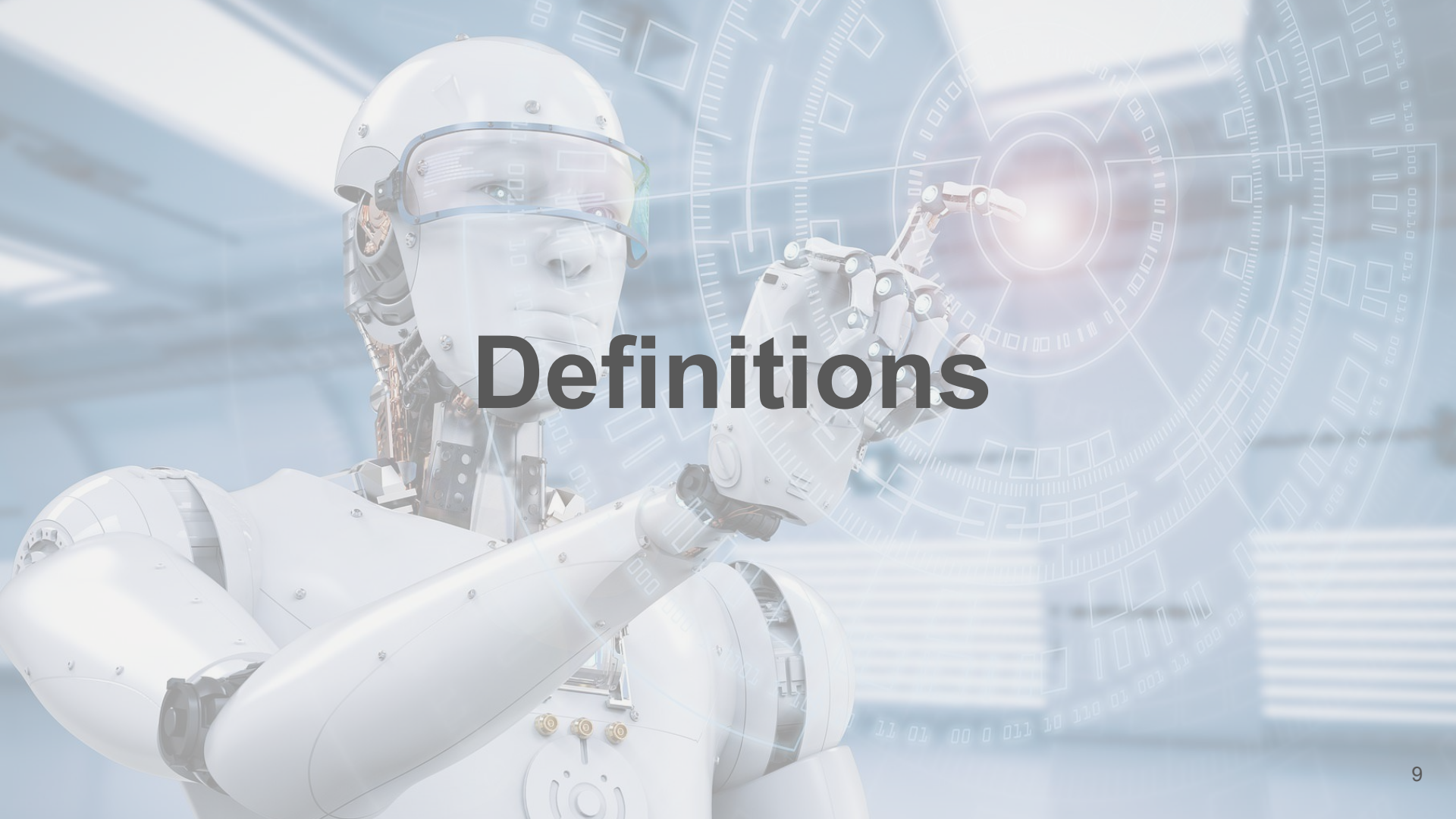



CNNs for Visual Recognition (Stanford)
Machine Learning
by Andrew Ng
Machine Learning
coursera
DeepLizard
fast.ai
Us
Depth
More Theory-Based
TEDMED
More Application-Based

# Hopefully you feel comfortable discussing

- Neural Network
- Activation map/neuron/kernel/state (whatever they decide to call it)
- Hidden layers
- Supervised/unsupervised learning
- Convolutional neural network
- KNN, SVM, Logistic Regression, Decision trees, Random Forest
- Hyperparameters
- Data Augmentation
- **Cross Validation**

  etc...

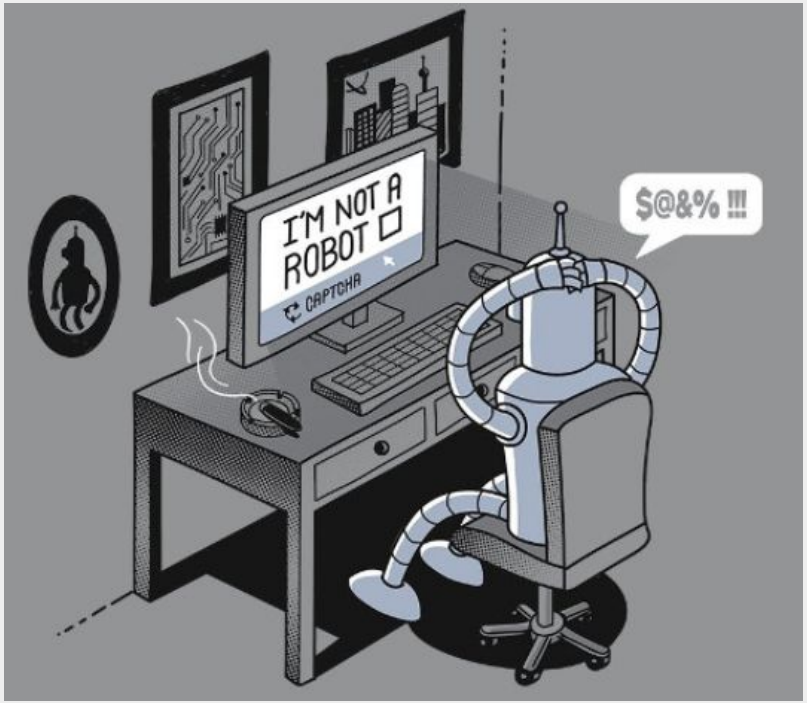


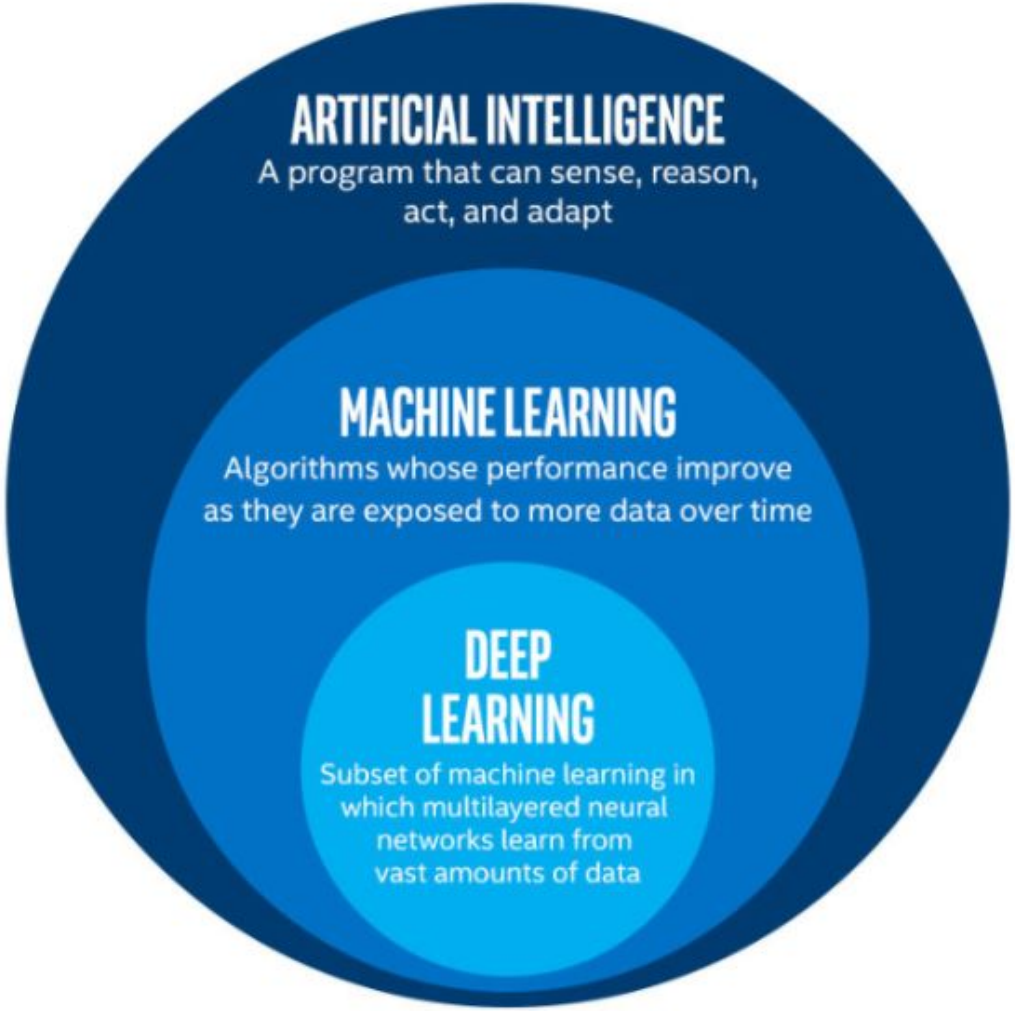
# Definitions

# What is AI?

A machine that can

- Analyze environment
- Complete a task
- Exhibit “natural intelligence”
  - Subjective




ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

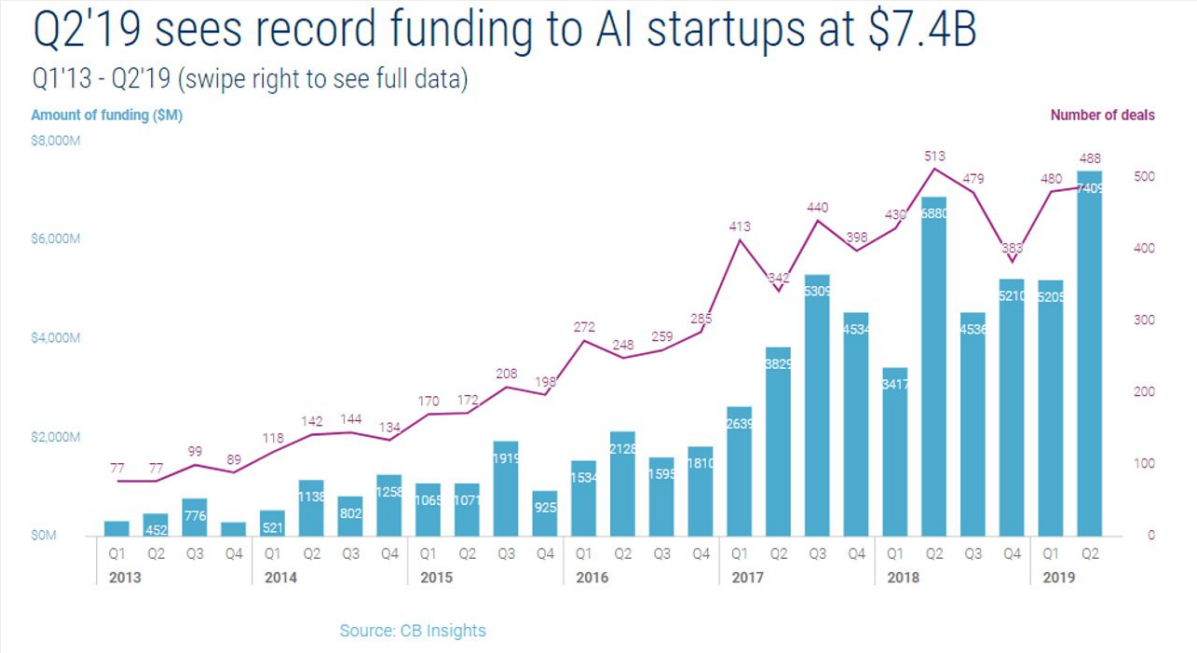
# Misconceptions

- AI doesn’t need humans
  - At this moment, few 100% unsupervised AI exist for high-stakes tasks
- AI is 100% objective
  - Bias from programming, not 100% neutral
- AI can just “figure out” your data
  - Still need defined problem space and data type

# What is the “value added” of AI?

Accomplishing tasks:

- Faster
- Automatically
- With higher accuracy (?)
- With greater complexity



# “With greater complexity”

AI defeats human
(Deep Blue vs. Kasparov)

# “With greater complexity”

AI-generated “advantage” score

AI defeats human
(Deep Blue vs. Kasparov)

↓

Humans study AI
(Magnus Carlsen)

# “With greater complexity”

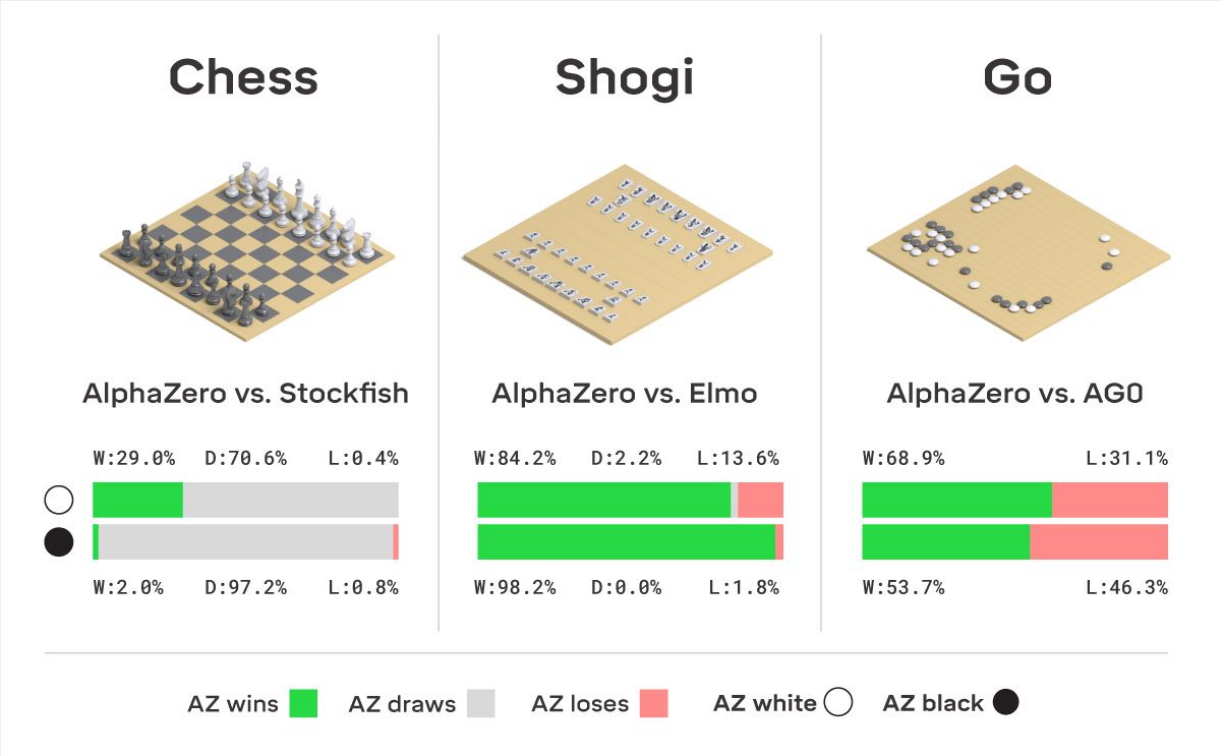
AI defeats human
(Deep Blue vs. Kasparov)

↓

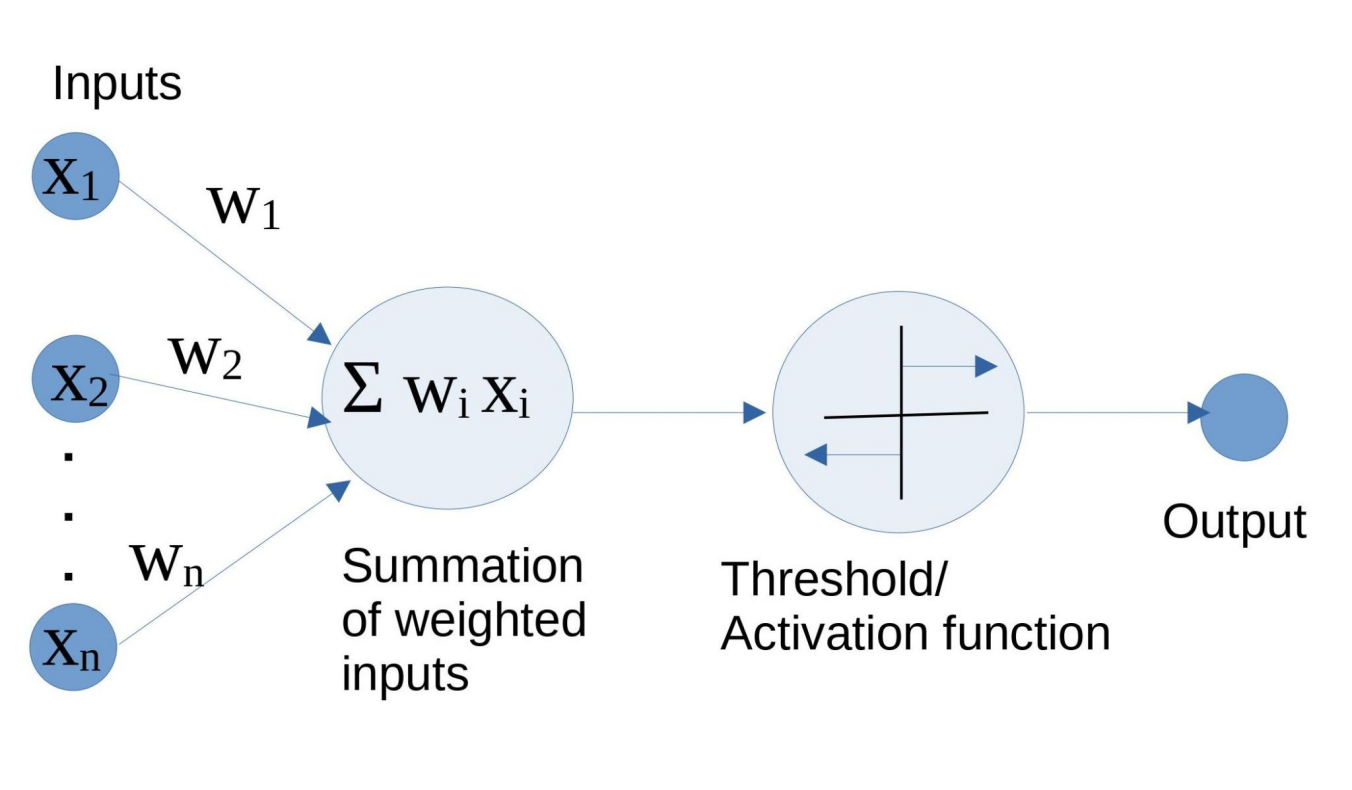
Humans study AI
(Magnus Carlsen)

↓

Stronger AI built
(Stockfish 12,
AlphaZero)



# Universal Approximation Theorem: Neural Networks can Model Anything (hmm…)



# Definitions

e.g. predict if hockey players will have >100 points this year

| ID | Division | Age | Minutes Played | Points Last Year | Jersey Color | Points > 100? |
|---|---|---|---|---|---|---|
| Matthews | Atlantic | 23 | 20:52 | 86 | Blue/white | Yes |
| Stamkos | Atlantic | 30 | 22:01 | 91 | Blue/white | No |
| McDavid | Pacific | 23 | 30:34 | 120 | Orange/blue | No |
| Petersson | Pacific | 20 | 21:46 | 90 | Blue/whute | No |
| Gaudreau | Pacific | 67 | 22:42 | 99 | Red/black | Yes |

# Definitions

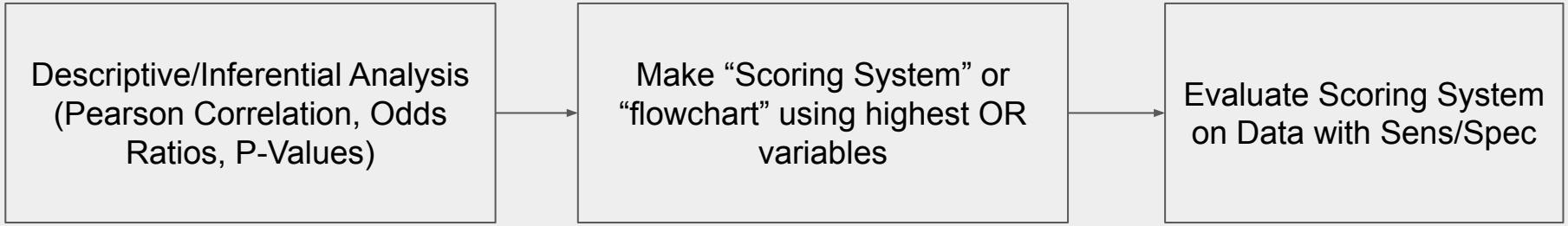
Outcome (or ground truth)

Features (or variables, characteristics, descriptors (don't like))

| ID | Division | Age | Minutes Played | Points Last Year | Jersey Color | Points > 100? |
|---|---|---|---|---|---|---|
| Matthews | Atlantic | 23 | 20:52 | 86 | Blue/white | Yes |
| Stamkos | Atlantic | 30 | 22:01 | 91 | Blue/white | No |
| McDavid | Pacific | 23 | 30:34 | 120 | Orange/blue | No |
| Petersson | Pacific | 20 | 21:46 | 90 | Blue/whute | No |
| Gaudreau | Pacific | 67 | 22:42 | 99 | Red/black | Yes |

Sample (or examples or items (don't like))

# Classical Method for Predictive Analysis

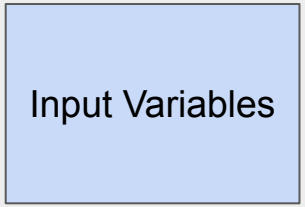


| CURB-65 | Clinical Feature | Points |
|---|---|---|
| C | Confusion | 1 |
| U | Urea > 7 mmol/L | 1 |
| R | RR ≥ 30 | 1 |
| B | SBP ≤ 90 mm Hg OR<br>DBP ≤ 60 mm Hg | 1 |
| 65 | Age > 65 | 1 |

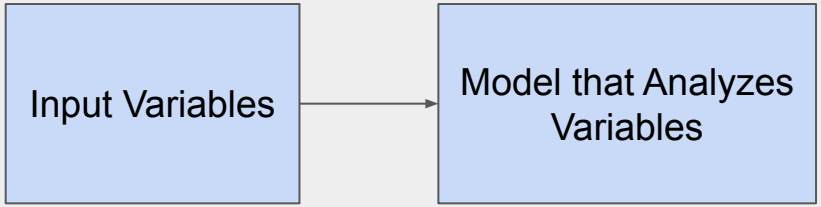
| CURB-65 Score | Risk group | 30-day mortality | Management |
|---|---|---|---|
| 0 -1 | 1 | 1.5% | Low risk, consider home treatment |
| 2 | 2 | 9.2% | Probably admission vs close outpatient management |
| 3-5 | 3 | 22% | Admission, manage as severe |

Hand selected low p, high OR variables

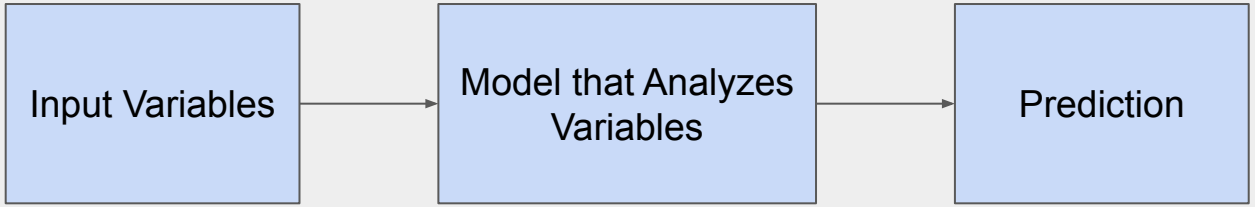
# Machine Learning



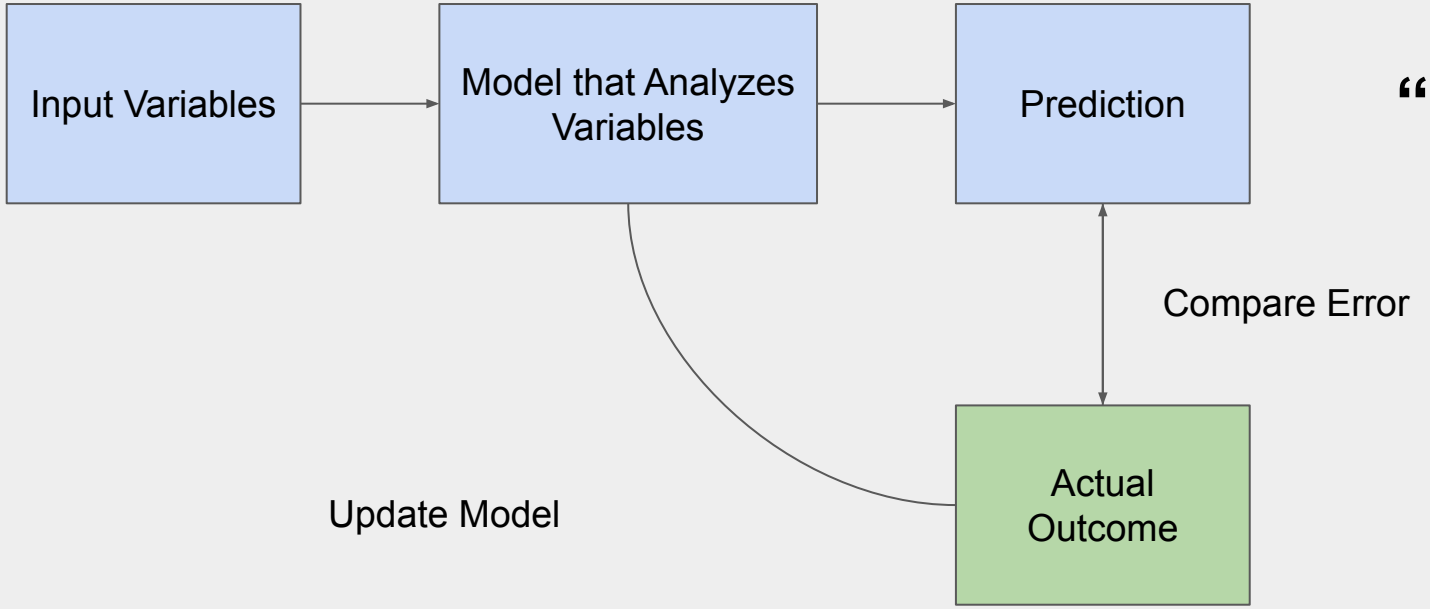
# Machine Learning



# Machine Learning

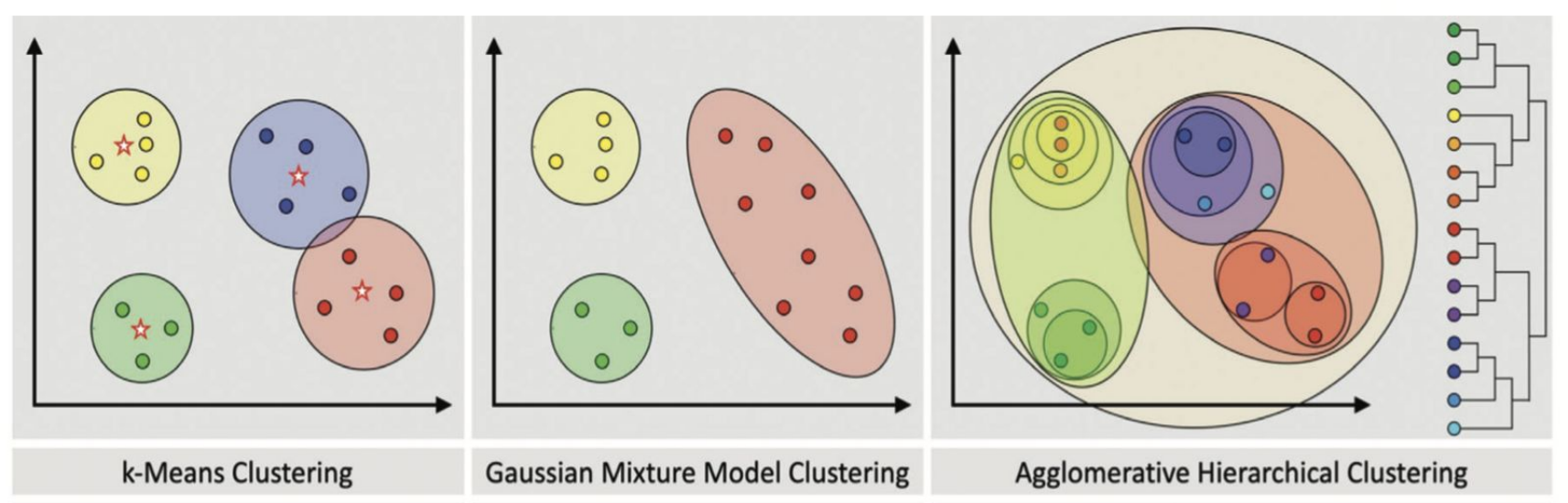


# Machine Learning (supervised learning)




**“Training”**

# Unsupervised Machine Learning



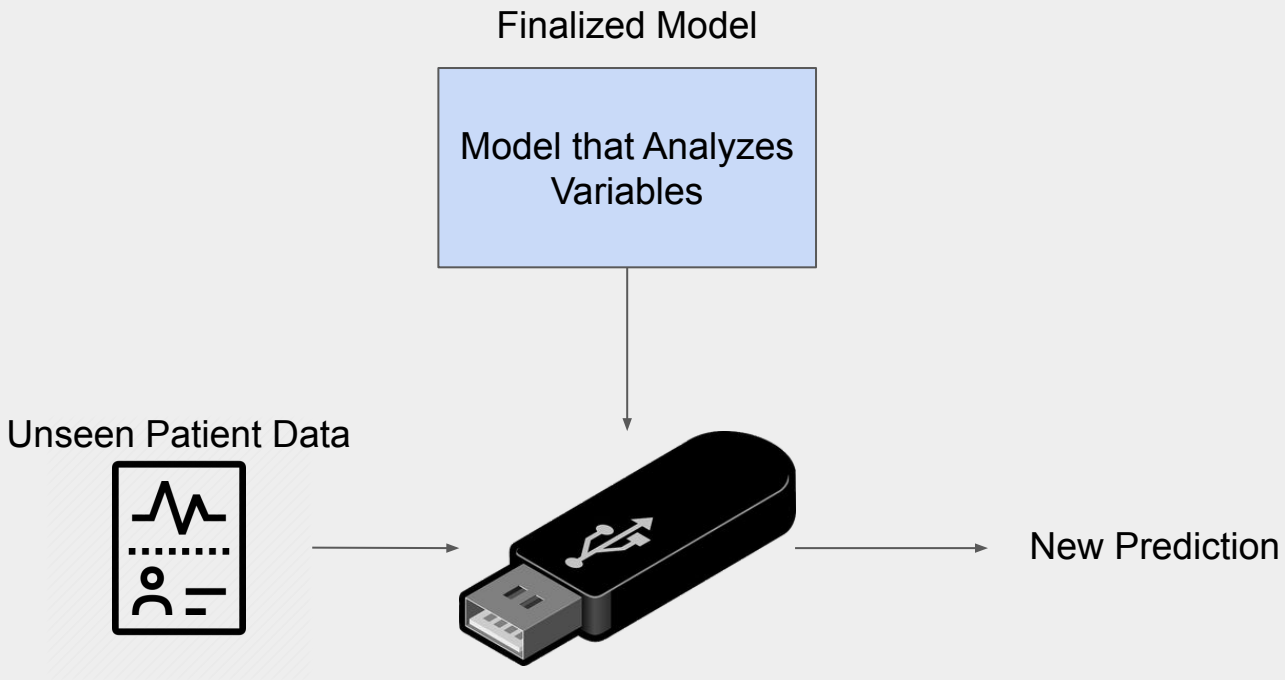

Find patterns without knowing ground truth

# Machine Learning

Finalized Model

Model that Analyzes Variables

# Machine Learning




**“Testing”**

# Important Questions:

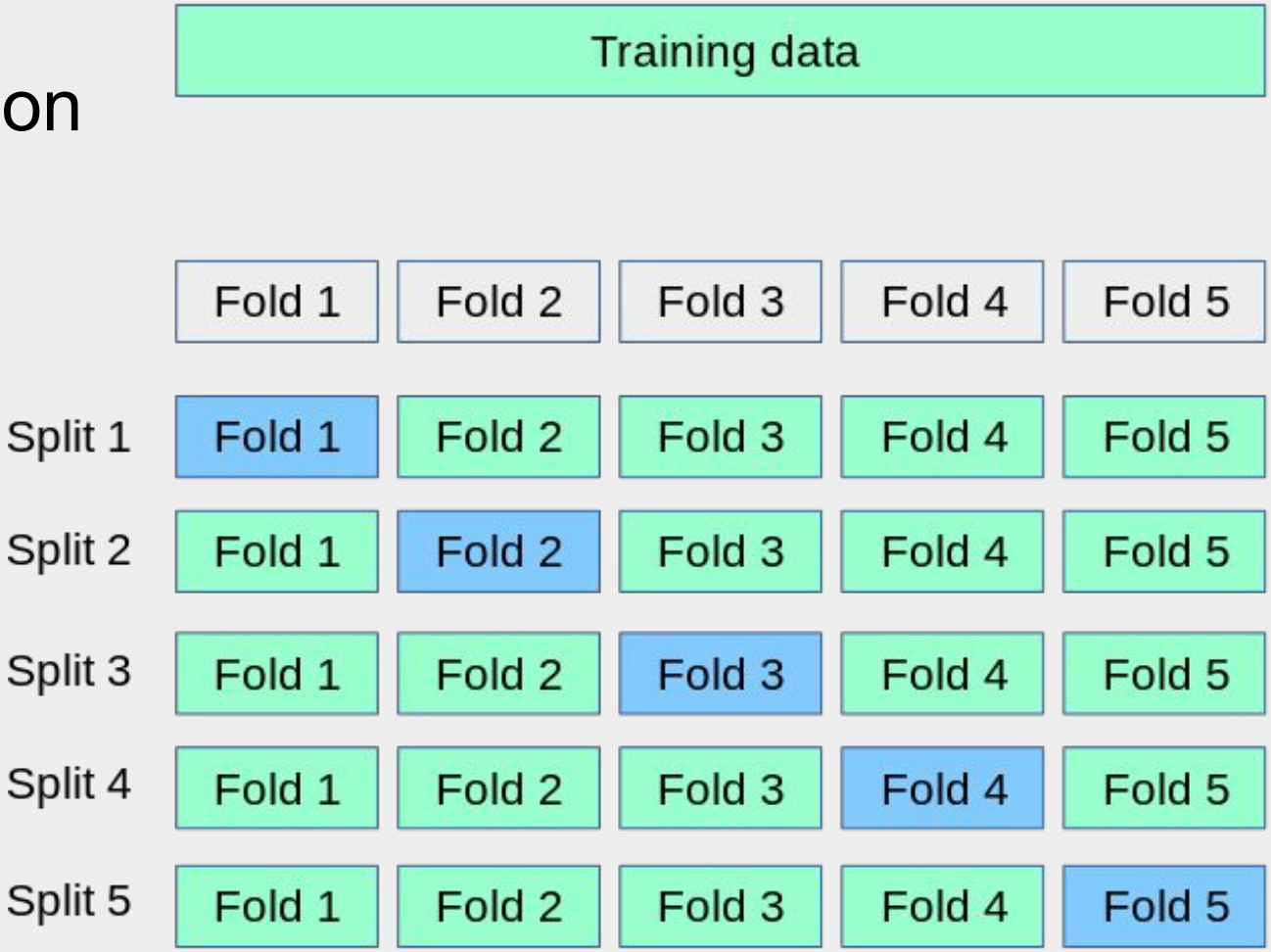
1. What model to pick?
2. How much data is needed?
3. What variables to include?
4. How to “clean data”
5. How to split train/validate/test?
6. How to optimize parameters during training?
7. What metric to validate model?
8. How good is “good enough”?
9. How thorough do comparison tests need to be?
10. How do I interpret the model/

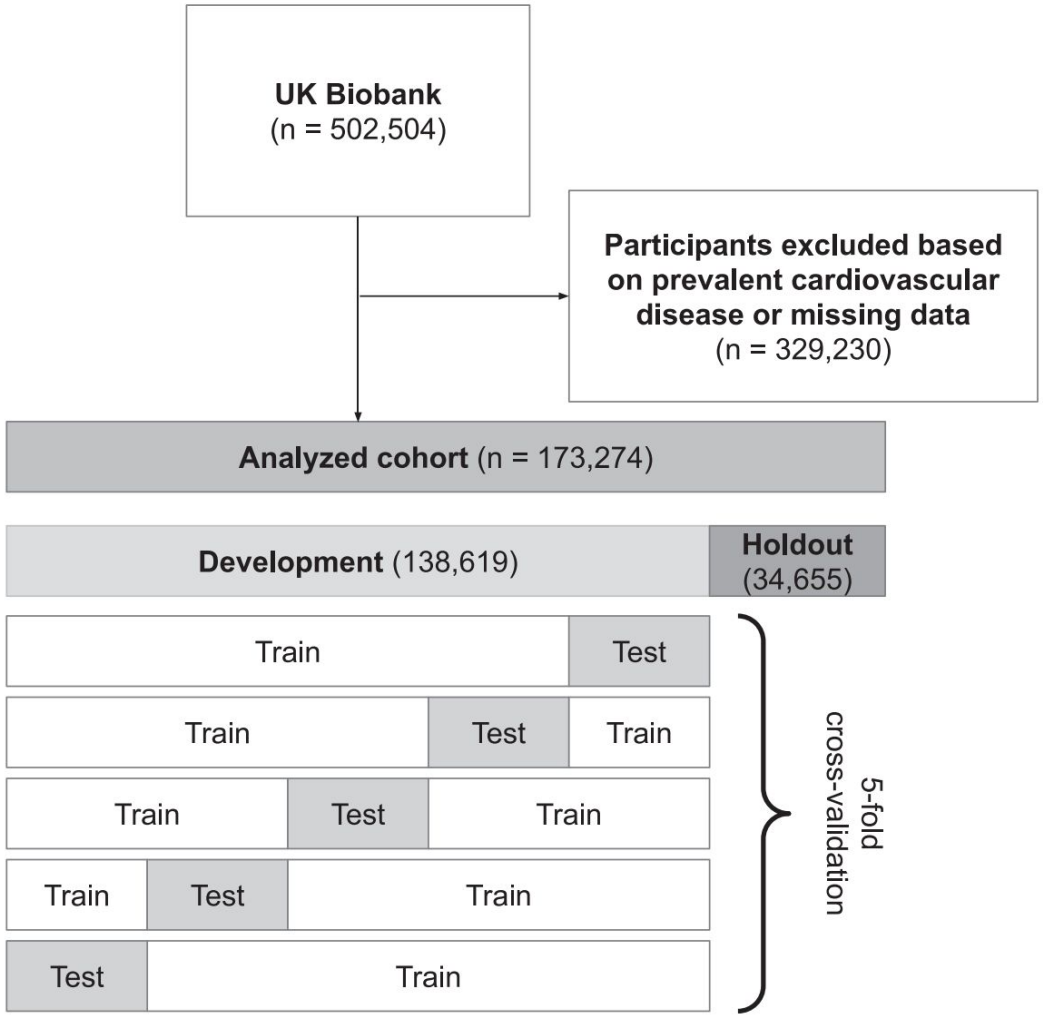
# Standard answers (not exhaustive!):

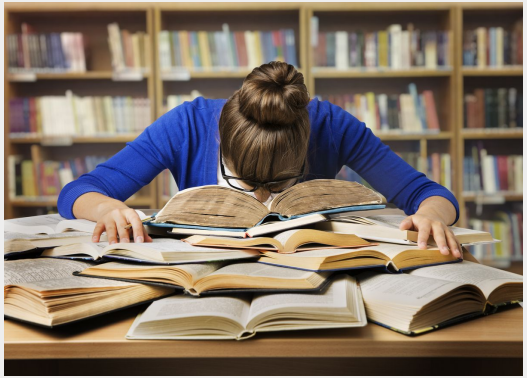
1. **What model to pick?**
    a. Classifiers (less to more complex): classification rule, linear/logistic regression, decision trees, k-nearest neighbors, support vector machine, random forest, neural network
2. **How much data is needed?**
    a. As much as possible, generally >100 samples, and >10 samples per feature
3. **What features to include?**
    a. Generally, as many as you think are biologically plausible
4. **How to "clean data"**
    a. Initial "biological relevance sweep", decide if you want to censor outliers, decided if you want sample with missing features or otherwise impute, remove collinears, (optional) rank by feature selection algorithm (e.g. mutual info ranking, pearson coefficient rank, or do backwards feature select after first machine learning model fit
5. **How to split train/validate/test?**
    a. Generally 60/20/20 standard, or 75/25 or 70/30 if no holdout test set. Best is if holdout set is external institution data (see TRIPOD classes)
6. **How to optimize parameters during training?**
    a. Can initialize to "common values", then gridsearch is common (e.g. sweep through a bunch of different combinations of parameters)
7. **What metric to validate model?**
    a. Classification: usually auc/acc/sens/spec
    b. Regression: mean distance error
    c. Segmentation: Dice coefficient, Hausdorff distance
8. **How good is "good enough"?**
    a. Depends on clinical question. Generally above ~0.92 get diminishing returns in medicine. In aerospace, then need 0.999999 etc.
    b. Also consider to inter-observer error of the gold standard
9. **How thorough do comparison tests need to be?**
    a. Want to test against current standard
    b. For completeness of model, ablation analysis (remove parts of the algorithm e.g. remove feature selection and re-assess performance)
    c. For comparison to different models, run more and less complex models. No hard rule, publications can get through with 2 or 10 comparison models.
10. **How do I interpret the model?**
    a. Really a problem for neural networks, methods include GradCAM, saliency maps, and LIME.

# Validation

# Cross-validation

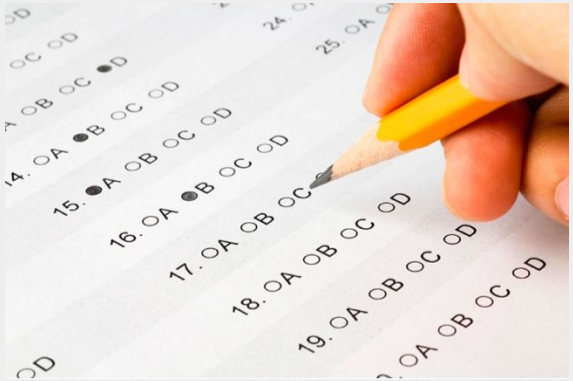



UK Biobank
(n = 502,504)
Participants excluded based on prevalent cardiovascular disease or missing data
(n = 329,230)
Analyzed cohort (n = 173,274)
Development (138,619)
Holdout (34,655)
Train
Test
Train
Test
Train
Train
Test
Train
Train
Test
Train
Test
Train
5-fold cross-validation



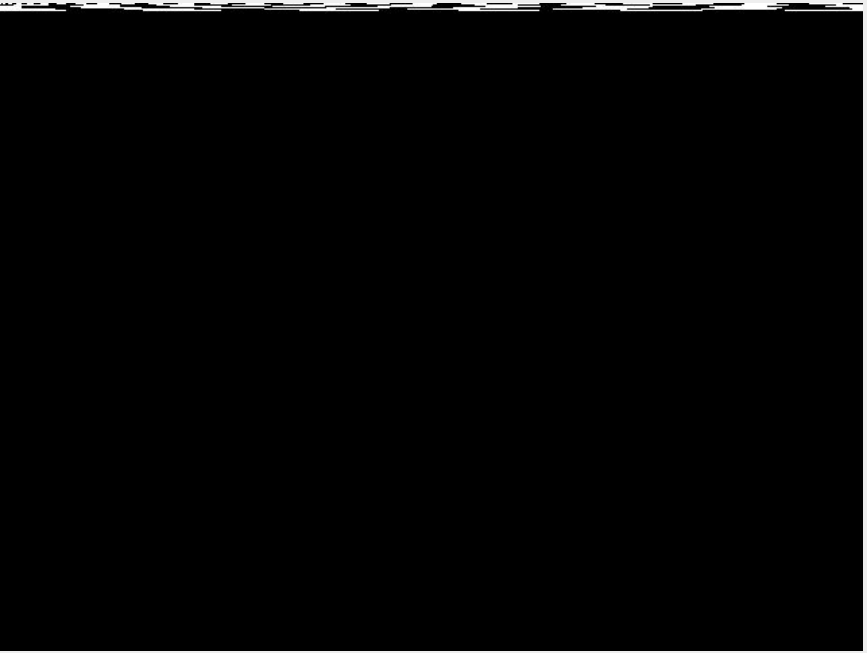
Revision
(Training)

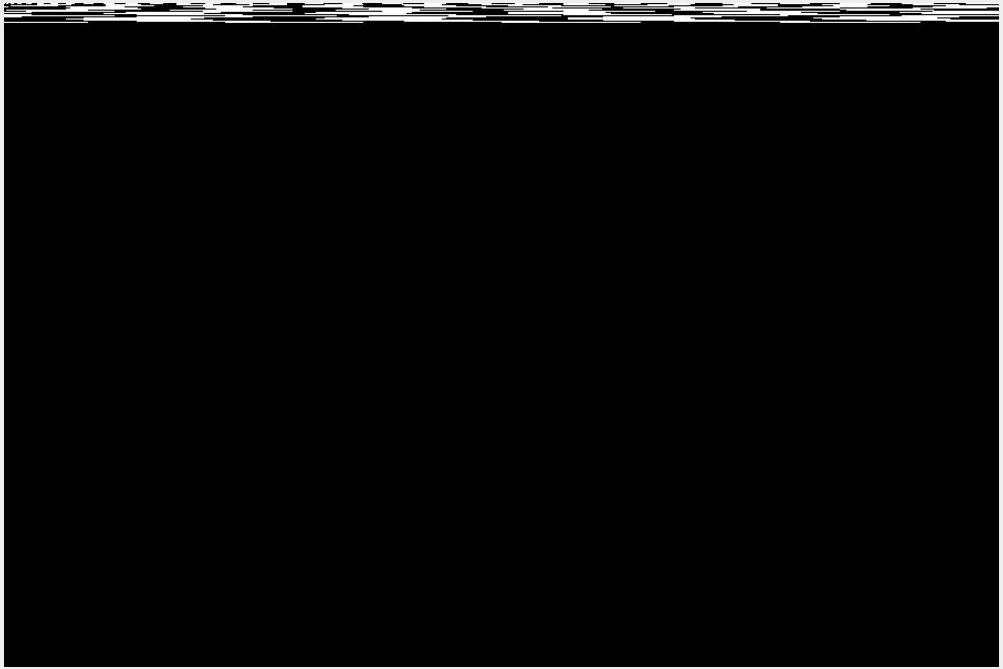
Practice Tests
(Validation)



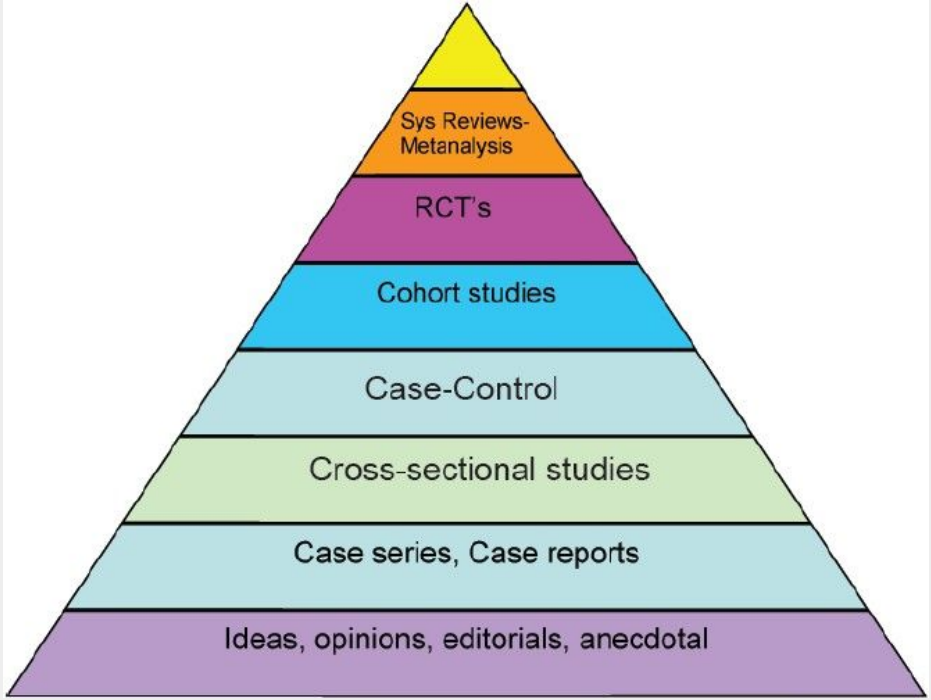
Final Exam
(Test)

# What does “learning” look like?

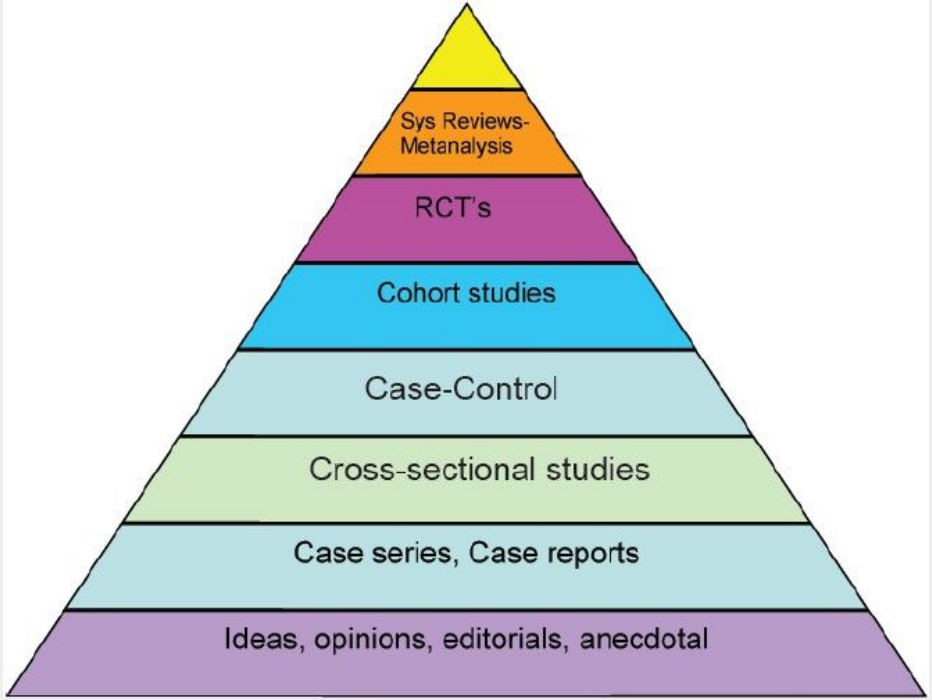




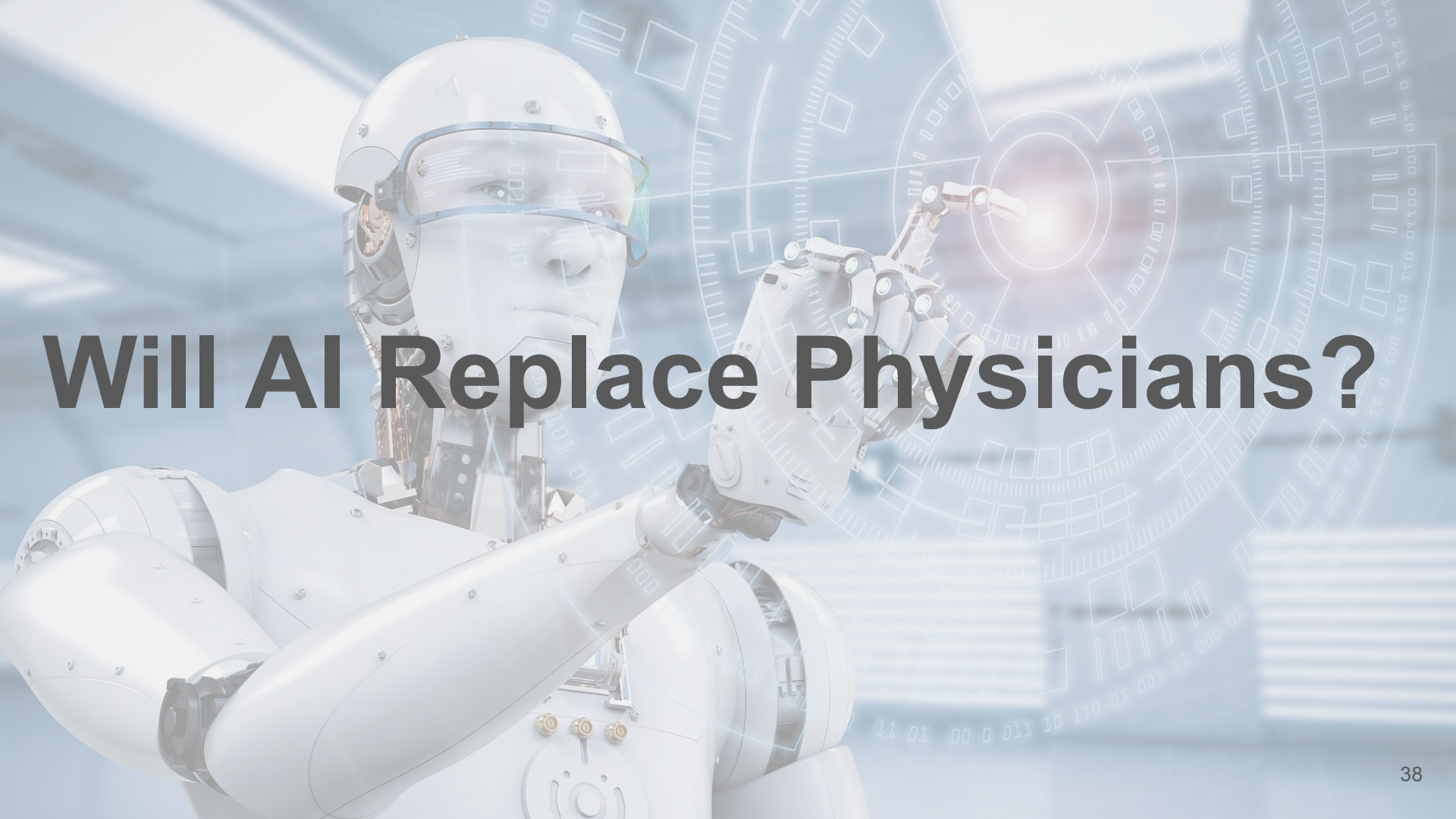
# Hierarchy of Evidence



# Hierarchy of Evidence




ML isn't a "traditional" control vs. comparator study, where does it stand?



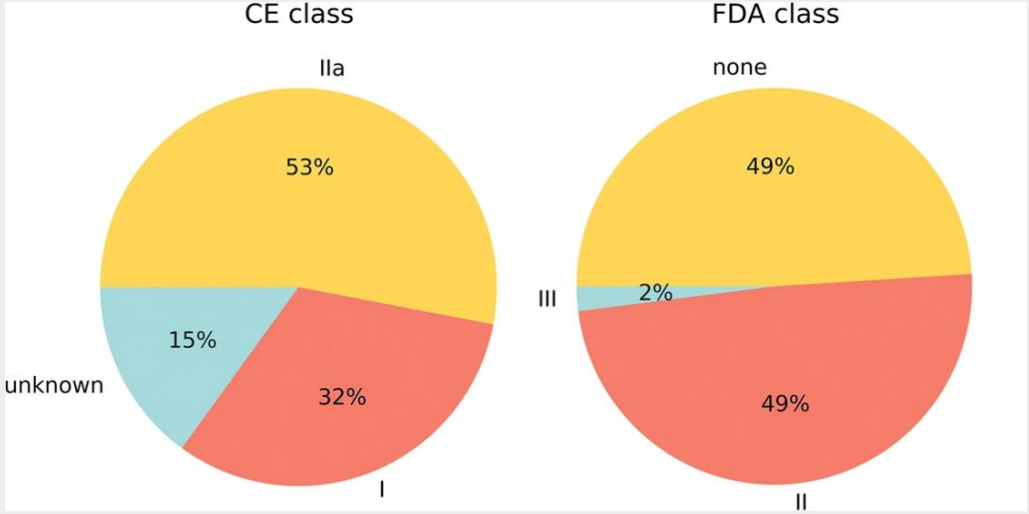
# Will AI Replace Physicians?

# Unlikely anytime soon

Many hospitals still use paper records!

AI needs databases & trained personnel

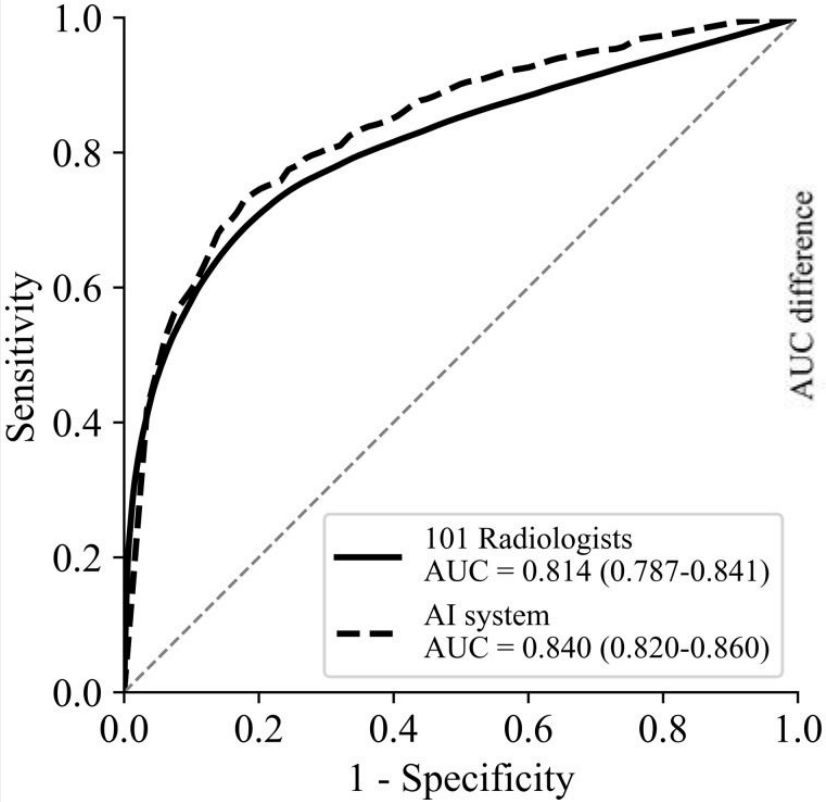
# Current tech is “physician assist”, not replacement

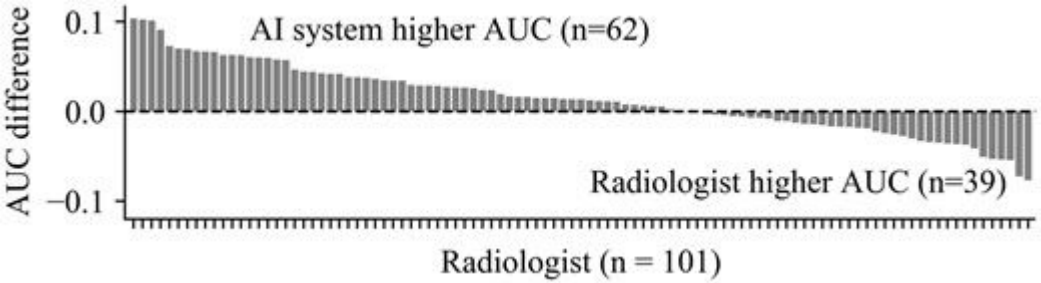



Nearly all AI are non-invasive and require physician supervision (< Class III)

Leeuwen et al. (2021)

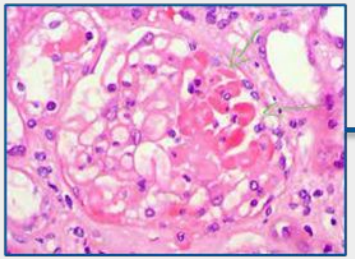
# AI rarely outperforms radiologists in aggregate

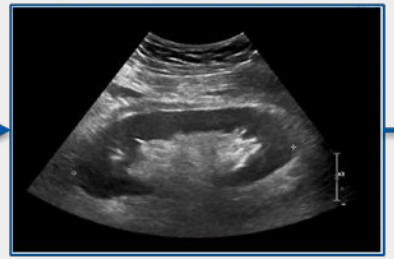



Rodriguez-Ruiz et al. (2019)

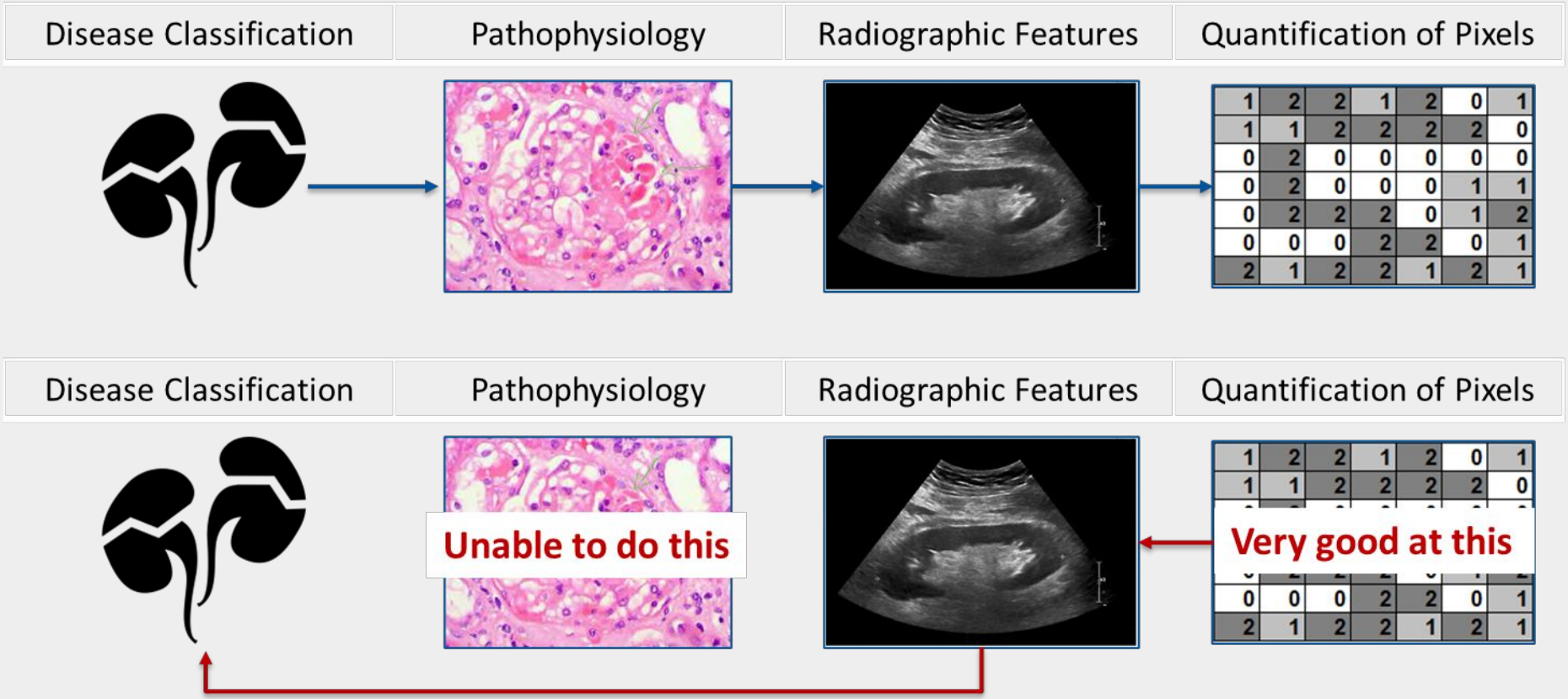
| Disease Classification | Pathophysiology | Radiographic Features | Quantification of Pixels |
|---|---|---|---|





| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 1 | 2 | 0 | 1 |
| 1 | 1 | 2 | 2 | 2 | 2 | 0 |
| 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 0 | 2 | 0 | 0 | 0 | 1 | 1 |
| 0 | 2 | 2 | 2 | 0 | 1 | 2 |
| 0 | 0 | 0 | 2 | 2 | 0 | 1 |
| 2 | 1 | 2 | 2 | 1 | 2 | 1 |

Radiologists: experts at identifying pathology from gross imaging features




AI: “top down” approach

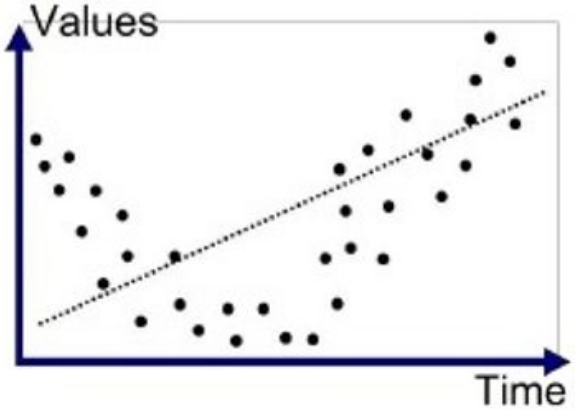
# AI Accuracies look good in single-center studies

**CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning**

| Pathology | Wang et al. (2017) | Yao et al. (2017) | CheXNet (ours) |
|---|---|---|---|
| Atelectasis | 0.716 | 0.772 | **0.8094** |
| Cardiomegaly | 0.807 | 0.904 | **0.9248** |
| Effusion | 0.784 | 0.859 | **0.8638** |
| Infiltration | 0.609 | 0.695 | **0.7345** |
| Mass | 0.706 | 0.792 | **0.8676** |
| Nodule | 0.671 | 0.717 | **0.7802** |
| Pneumonia | 0.633 | 0.713 | **0.7680** |
| Pneumothorax | 0.806 | 0.841 | **0.8887** |
| Consolidation | 0.708 | 0.788 | **0.7901** |
| Edema | 0.835 | 0.882 | **0.8878** |
| Emphysema | 0.815 | 0.829 | **0.9371** |
| Fibrosis | 0.769 | 0.767 | **0.8047** |
| Pleural Thickening | 0.708 | 0.765 | **0.8062** |
| Hernia | 0.767 | 0.914 | **0.9164** |

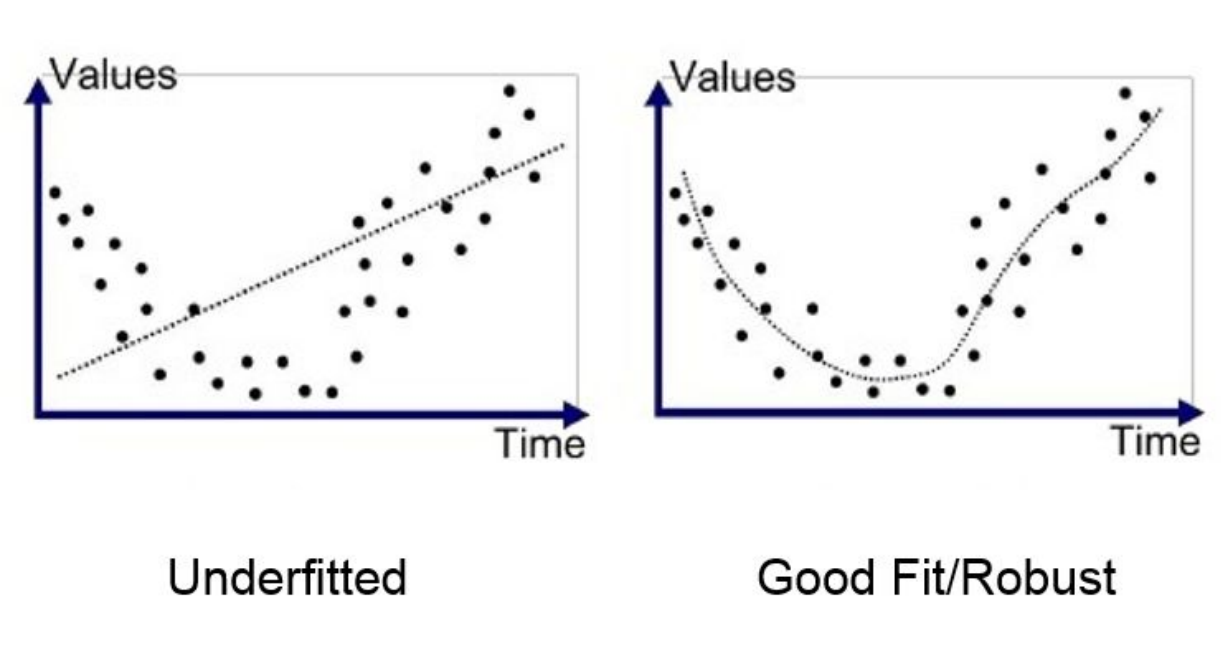
Rajpurkar et al. (2017)

# “I can force a correlation with anything”

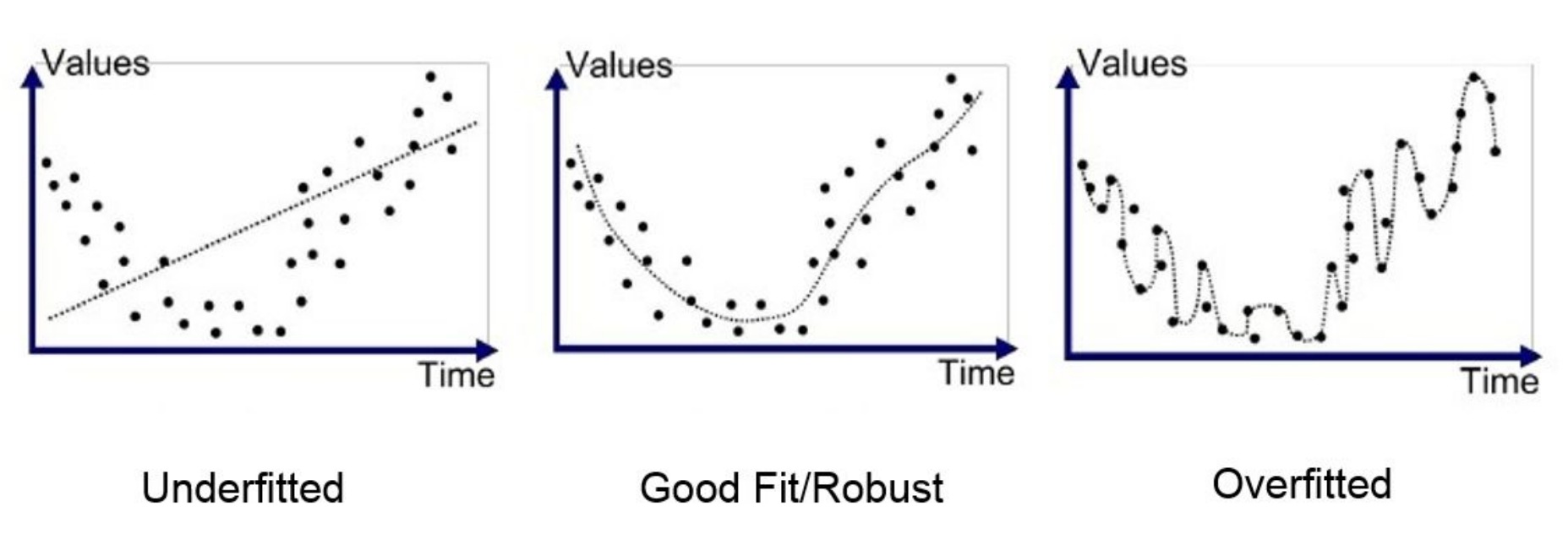



Underfitted

# “I can force a correlation with anything”



# Overfit models will perform poorly on new data.




**In medicine, this has important equity implications!**

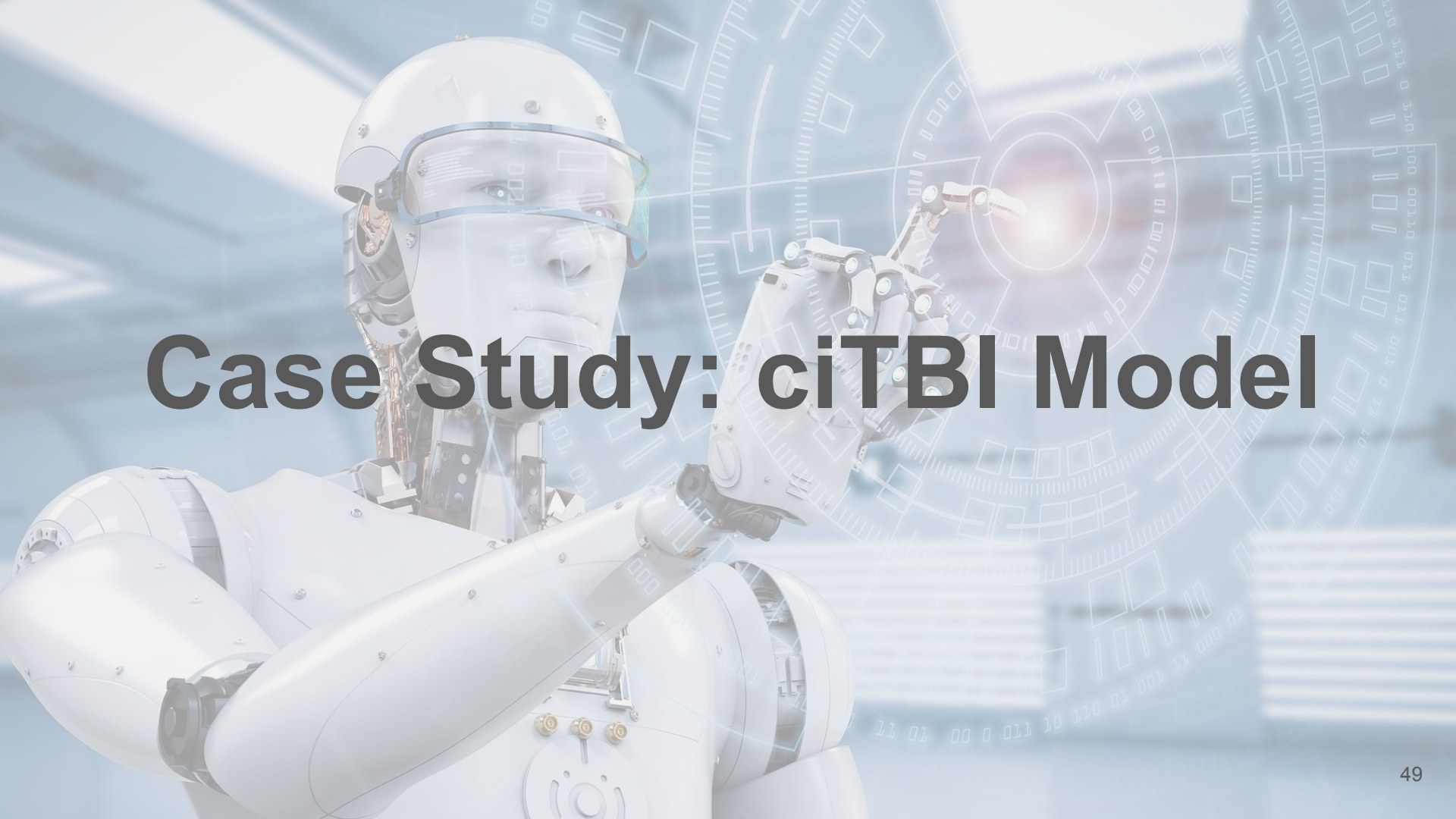
# Overfitting signalling questions:

1. Are there too few samples:features (generally want 10:1, but not hard rule)

2. Are there enough positive/negative events?

3. Was the model external validated?

4. Was the model validated/tested multiple times (e.g. CV/LOOCV)



# Case Study: ciTBI Model

Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study

Prof Nathan Kuppermann, MD • Prof James F Holmes, MD • Peter S Dayan, MD • John D Hoyle, MD • Shireen M Atabaki, MD • Richard Holubkov, PhD • et al. Show all authors • Show footnotes

Published: September 15, 2009 • DOI: https://doi.org/10.1016/S0140-6736(09)61558-0 • Check for updates

Developing algorithm to predict if CT is needed for clinically important traumatic brain injury (ciTBI)

# Let’s try to PICO

**P:** Patients <18 with GCS14-15 within 24h of head trauma. N=42,412

**I:** …CT scan? (14,969 had CTs)

**C:** …No CT scan?

**O:** TBI

Not a traditional intervention/comparator study!

# Let’s try to PICO

**P:** Patients <18 with GCS14-15 within 24h of head trauma. N=42,412

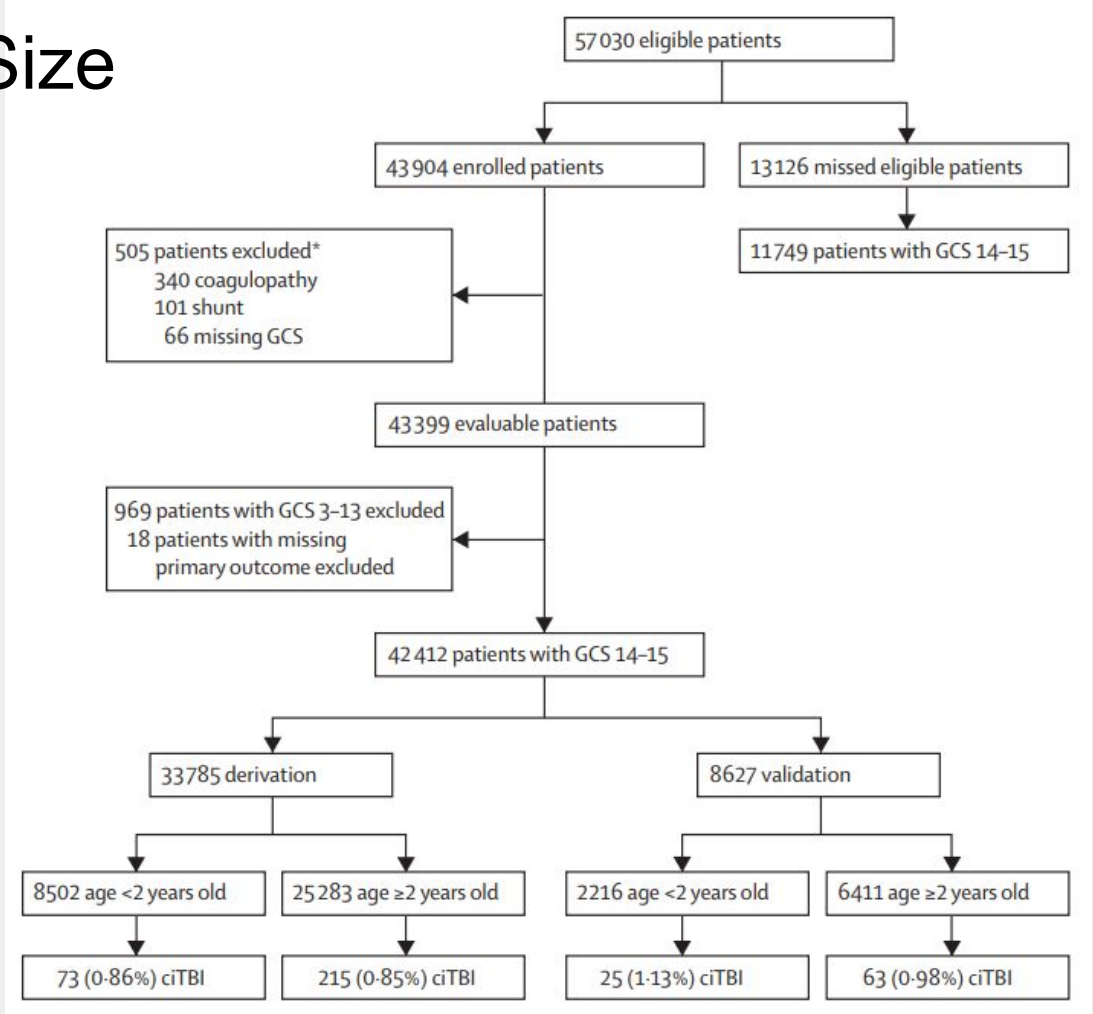
**I:** Prediction of ciTBI with new algorithm

**C:** Prediction of ciTBI with comparator algorithm (physician judgment, CATCH etc.)

**O:** Prediction accuracy/sens/spec/auc

Now you can do t-test between intervention and comparator

# Assessing Sample Size



# Assessing Model

(features selected)

| | Age <2 years (n=10 718) | | Age ≥2 years (n=31 694) | |
|---|---|---|---|---|
| | Derivation (n=8502) | Validation (n=2216) | Derivation (n=25 283) | Validation (n=6411) |
| Severity of injury mechanism* | | | | |
| Mild | 1262/8424 (15·0%) | 309/2186 (14·1%) | 4505/25 128 (17·9%) | 1030/6361 (16·2%) |
| Moderate | 5317/8424 (63·1%) | 1383/2186 (63·3%) | 17 598/25 128 (70·0%) | 4432/6361 (69.7%) |
| Severe | 1845/8424 (21·9%) | 494/2186 (22·6%) | 3025/25 128 (12·0%) | 899/6361 (14·1%) |
| History of LOC | | | | |
| Known or suspected | 425/8179 (5·2%) | 116/2119 (5·5%) | 4701/24 275 (19·4%) | 1044/6120 (17·1%) |
| LOC duration | | | | |
| No LOC | 7754/8113 (95·6%) | 2003/2102 (95·3%) | 19574/22489 (87·0%) | 5076/5706 (89·0%) |
| <5 s | 61/8113 (0·8%) | 20/2102 (1·0%) | 679/22489 (3·0%) | 147/5706 (2·6%) |
| 5–60 s | 173/8113 (2·1%) | 46/2102 (2·2%) | 1331/22489 (5·9%) | 272/5706 (4·8%) |
| 1–5 min | 79/8113 (1·0%) | 24/2102 (1·1%) | 781/22489 (3·5%) | 181/5706 (3·2%) |
| >5 min | 46/8113 (0·6%) | 9/2102 (0·4%) | 124/22489 (0·6%) | 30/5706 (0·5%) |
| Headache | .. | .. | 10296/21997 (46·8%) | 2379/5498 (43·3%) |
| Severity of headache | | | | |
| No headache | .. | .. | 11701/21193 (55·2%) | 3119/5301 (58·8%) |
| Mild | .. | .. | 4262/21193 (20·1%) | 986/5301 (18·6%) |
| Moderate | .. | .. | 4572/21193 (21·6%) | 1050/5301 (19·8%) |
| Severe | .. | .. | 658/21193 (3·1%) | 146/5301 (2·8%) |
| History of vomiting | 1271/8446 (15·0%) | 294/2190 (13·4%) | 3236/25102 (12·9%) | 756/6374 (11·9%) |
| Number of vomiting episodes | | | | |
| 0 | 7175/8389 (85·5%) | 1896/2178 (87·1%) | 21866/24964 (87·6%) | 5618/6328 (88·8%) |
| 1 | 548/8389 (6·5%) | 128/2178 (5·9%) | 1144/24964 (4·6%) | 268/6328 (4·2%) |
| 2 | 241/8389 (2·9%) | 67/2178 (3·1%) | 661/24964 (2·6%) | 139/6328 (2·2%) |
| >2 | 425/8389 (5·1%) | 87/2178 (4·0%) | 1293/24964 (5·2%) | 303/6328 (4·8%) |
| Acting abnormally according to parent | 1166/8142 (14·3%) | 273/2152 (12·7%) | 3792/23177 (16·4%) | 966/5935 (16·3%) |
| GCS score | | | | |
| 14 | 366/8502 (4·3%) | 92/2216 (4·2%) | 720/25283 (2·8%) | 163/6411 (2·5%) |
| 15 | 8136/8502 (95·7%) | 2124/2216 (95·8%) | 24563/25283 (97·2%) | 6248/6411 (97·5%) |
| Altered mental status† | 978/8444 (11·6%) | 232/2205 (10·5%) | 3427/25083 (13·7%) | 850/6364 (13·4%) |
| Signs of basilar skull fracture | 42/8408 (0·5%) | 15/2187 (0·7%) | 179/25052 (0·7%) | 51/6344 (0·8%) |

# Assessing Model

## Statistical analysis

Preverbal (<2 years of age) and verbal (2 years and older) children were analysed separately because of young patients' greater sensitivity to radiation, minimal ability to communicate, and different mechanisms and risks for traumatic brain injury.[9,15,31,32] Because the main goal of these analyses was to identify children at very low risk of ciTBI in whom CT can be avoided, we aimed to maximise the negative predictive value and sensitivity of the prediction rules. We regarded a child to be at very low risk of ciTBI if none of the predictors in the derived rules was present. We derived the rules with binary recursive partitioning (CART PRO 6.0; San Diego, CA, USA, Salford Systems).[33] We used ten-fold cross validation to create stable prediction trees, and standard Gini splitting rules.[33] To keep risks of misclassification of patients with ciTBIs to a minimum,

# Assessing Results

| | Derivation | | | Validation | | |
|---|---|---|---|---|---|---|
| | ciTBI | No ciTBI | Total | ciTBI | No ciTBI | Total |
| Any predictor present | 72 | 3903 | 3975 | 25 | 1016 | 1041 |
| No predictor present | 1 | 4526 | 4527 | 0 | 1175 | 1175 |
| **Total** | **73** | **8429** | **8502** | **25** | **2191** | **2216** |

| | Derivation | Validation |
|---|---|---|
| Prediction rule sensitivity (95% CI) | 98·6% (92·6–99·97) | 100·00% (86·3–100·00) |
| Prediction rule specificity (95% CI) | 53·7% (52·6–54·8) | 53·6% (51·5–55·7) |
| Negative predictive value (95% CI) | 99·9% (99·88–99·999) | 100·00% (99·7–100·00) |
| Positive predictive value (95% CI) | 1·8% (1·4–2·3) | 2·4% (1·6–3·5) |
| Negative likelihood ratio (95% CI) | 0·03 (0·001–0·14) | 0·0 (0–0·26) |

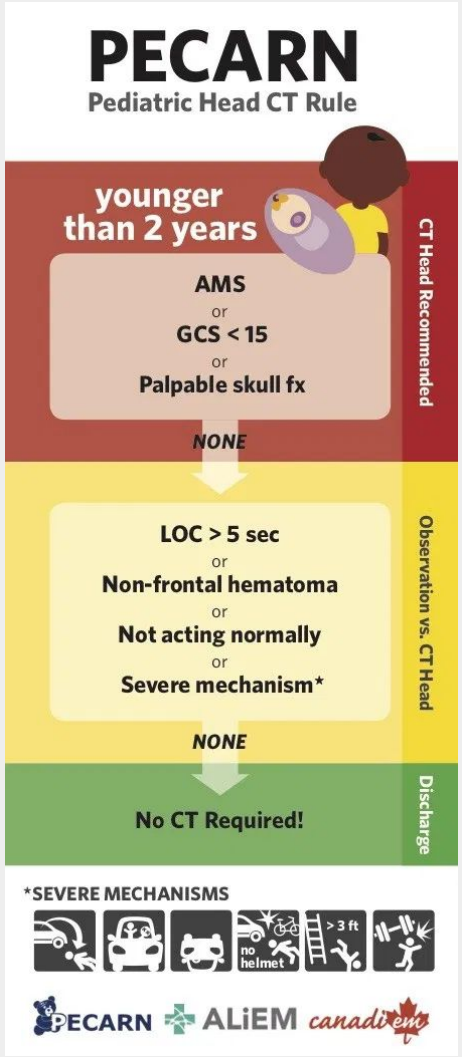
| | Derivation | | | Validation | | |
|---|---|---|---|---|---|---|
| | ciTBI | No ciTBI | Total | ciTBI | No ciTBI | Total |
| Any predictor present | 208 | 10 635 | 10 843 | 61 | 2652 | 2713 |
| No predictor present | 7 | 14 433 | 14 440 | 2 | 3696 | 3698 |
| **Total** | **215** | **25 068** | **25 283** | **63** | **6348** | **6411** |

| | Derivation | Validation |
|---|---|---|
| Prediction rule sensitivity (95% CI) | 96·7% (93·4–98·7) | 96·8% (89·0–99·6) |
| Prediction rule specificity (95% CI) | 57·6% (57·0–58·2) | 58·2% (57·0–59·4) |
| Negative predictive value (95% CI) | 99·95% (99·9–99·98) | 99·95% (99·80–99·99) |
| Positive predictive value (95% CI) | 1·9% (1·7–2·2) | 2·2% (1·7–2·9) |
| Negative likelihood ratio (95% CI) | 0·06 (0·03–0·12) | 0·05 (0·01–0·19) |

<2y

>=2y

(would prefer AUC and comparison to other methods)




# This clinical guideline was made with ML!

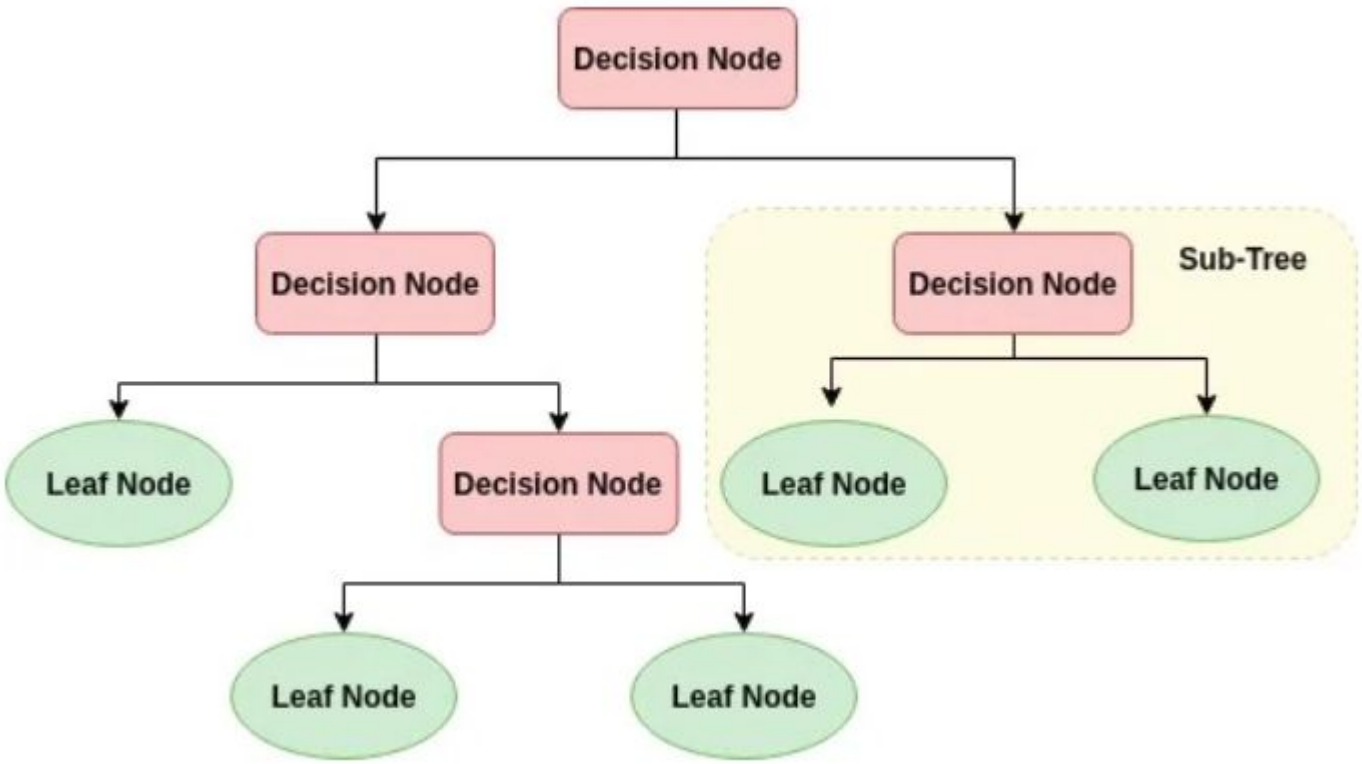
Take lots of relevant variables

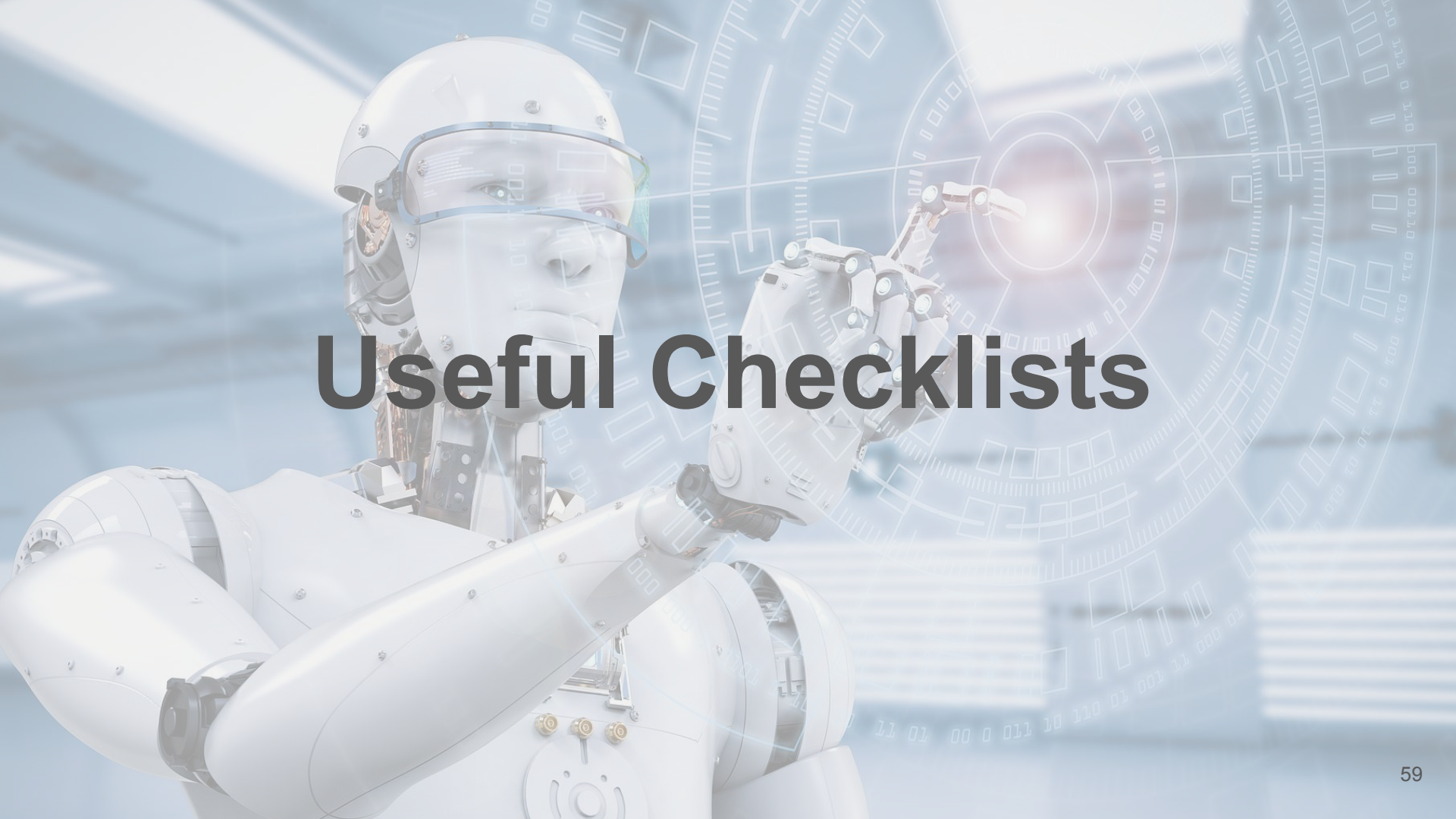
↓

Train a machine learning model (CART decision tree)

↓

Use model to evaluate future patients


Decision Node
Decision Node
Decision Node
Sub-Tree
Leaf Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Leaf Node



# Useful Checklists

Annals of Internal Medicine®

Search Journ

LATEST ISSUES IN THE CLINIC JOURNAL CLUB MULTIMEDIA CME / MOC AUTHORS / SUBMIT

Research and Reporting Methods | 6 January 2015

# Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement FREE

Gary S. Collins, PhD, Johannes B. Reitsma, MD, PhD, Douglas G. Altman, DSc, and Karel G.M. Moons, PhD

Author, Article and Disclosure Information

https://doi.org/10.7326/M14-0697

TRIPOD Checklist: Prediction Model Development

| Section/Topic | Item | Checklist Item | Page |
|---|---|---|---|
| **Title and abstract** | | | |
| Title | 1 | Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. | |
| Abstract | 2 | Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | |
| **Introduction** | | | |
| Background and objectives | 3a | Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. | |
| | 3b | Specify the objectives, including whether the study describes the development or validation of the model or both. | |
| **Methods** | | | |
| Source of data | 4a | Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. | |
| | 4b | Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up. | |
| Participants | 5a | Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres. | |
| | 5b | Describe eligibility criteria for participants. | |
| | 5c | Give details of treatments received, if relevant. | |
| Outcome | 6a | Clearly define the outcome that is predicted by the prediction model, including how and when assessed. | |
| | 6b | Report any actions to blind assessment of the outcome to be predicted. | |
| Predictors | 7a | Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured. | |
| | 7b | Report any actions to blind assessment of predictors for the outcome and other predictors. | |
| Sample size | 8 | Explain how the study size was arrived at. | |
| Missing data | 9 | Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method. | |
| Statistical analysis methods | 10a | Describe how predictors were handled in the analyses. | |
| | 10b | Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. | |
| | 10d | Specify all measures used to assess model performance and, if relevant, to compare multiple models. | |
| Risk groups | 11 | Provide details on how risk groups were created, if done. | |
| **Results** | | | |
| Participants | 13a | Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. | |
| | 13b | Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. | |
| Model development | 14a | Specify the number of participants and outcome events in each analysis. | |
| | 14b | If done, report the unadjusted association between each candidate predictor and outcome. | |
| Model specification | 15a | Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). | |
| | 15b | Explain how to the use the prediction model. | |
| Model performance | 16 | Report performance measures (with CIs) for the prediction model. | |
| **Discussion** | | | |
| | | Discuss any limitations of the study (such as nonrepresentative sample, few events | |

https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf

| Analysis Type | Description |
|---|---|
| Type 1a | Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance). |
| Type 1b | Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as 'internal validation', are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15). |
| Type 2a | The data are randomly split into two groups: one to develop the prediction model, and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16). |
| Type 2b | The data are nonrandomly split (e.g., by location or time) into two groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a, because it allows for nonrandom variation between the 2 data sets (6, 13, 17). |
| Type 3 | Development of a prediction model using one data set and an evaluation of its performance on separate data (e.g., from a different study). |
| Type 4 | The evaluation of the predictive performance of an existing (published) prediction model on separate data (13). |

Types 3 and 4 are commonly referred to as 'external validation studies.' Arguably, type 2b is as well, although it may be considered an intermediary between internal and external validation.

Annals of Internal Medicine®

Search Journal

LATEST ISSUES IN THE CLINIC JOURNAL CLUB MULTIMEDIA CME / MOC AUTHORS / SUBMIT

Research and Reporting Methods | 1 January 2019

# PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies FREE

Robert F. Wolff, MD*, Karel G.M. Moons, PhD*, Richard D. Riley, PhD, ... View all authors +

Author, Article and Disclosure Information

https://doi.org/10.7326/M18-1376

| 1. Participants | 2. Predictors | 3. Outcome | 4. Analysis |
|---|---|---|---|
| **Signaling questions** | | | |
| 1.1. Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data? | 2.1. Were predictors defined and assessed in a similar way for all participants? | 3.1. Was the outcome determined appropriately? | 4.1. Were there a reasonable number of participants with the outcome? |
| 1.2. Were all inclusions and exclusions of participants appropriate? | 2.2. Were predictor assessments made without knowledge of outcome data? | 3.2. Was a prespecified or standard outcome definition used? | 4.2. Were continuous and categorical predictors handled appropriately? |
| – | 2.3. Are all predictors available at the time the model is intended to be used? | 3.3. Were predictors excluded from the outcome definition? | 4.3. Were all enrolled participants included in the analysis? |
| – | – | 3.4. Was the outcome defined and determined in a similar way for all participants? | 4.4. Were participants with missing data handled appropriately? |
| – | – | 3.5. Was the outcome determined without knowledge of predictor information? | 4.5. Was selection of predictors based on univariable analysis avoided?† |
| – | – | 3.6. Was the time interval between predictor assessment and outcome determination appropriate? | 4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately? |
| – | – | – | 4.7. Were relevant model performance measures evaluated appropriately? |
| – | – | – | 4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?† |
| – | – | – | 4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?† |
| **ROB** | | | |
| Selection of participants | Predictors or their assessment | Outcome or its determination | Analysis |
| **Applicability** | | | |
| Included participants or setting does not match the review question | Definition, assessment, or timing of predictors does not match the review question | Its definition, timing, or determination does not match the review question | – |