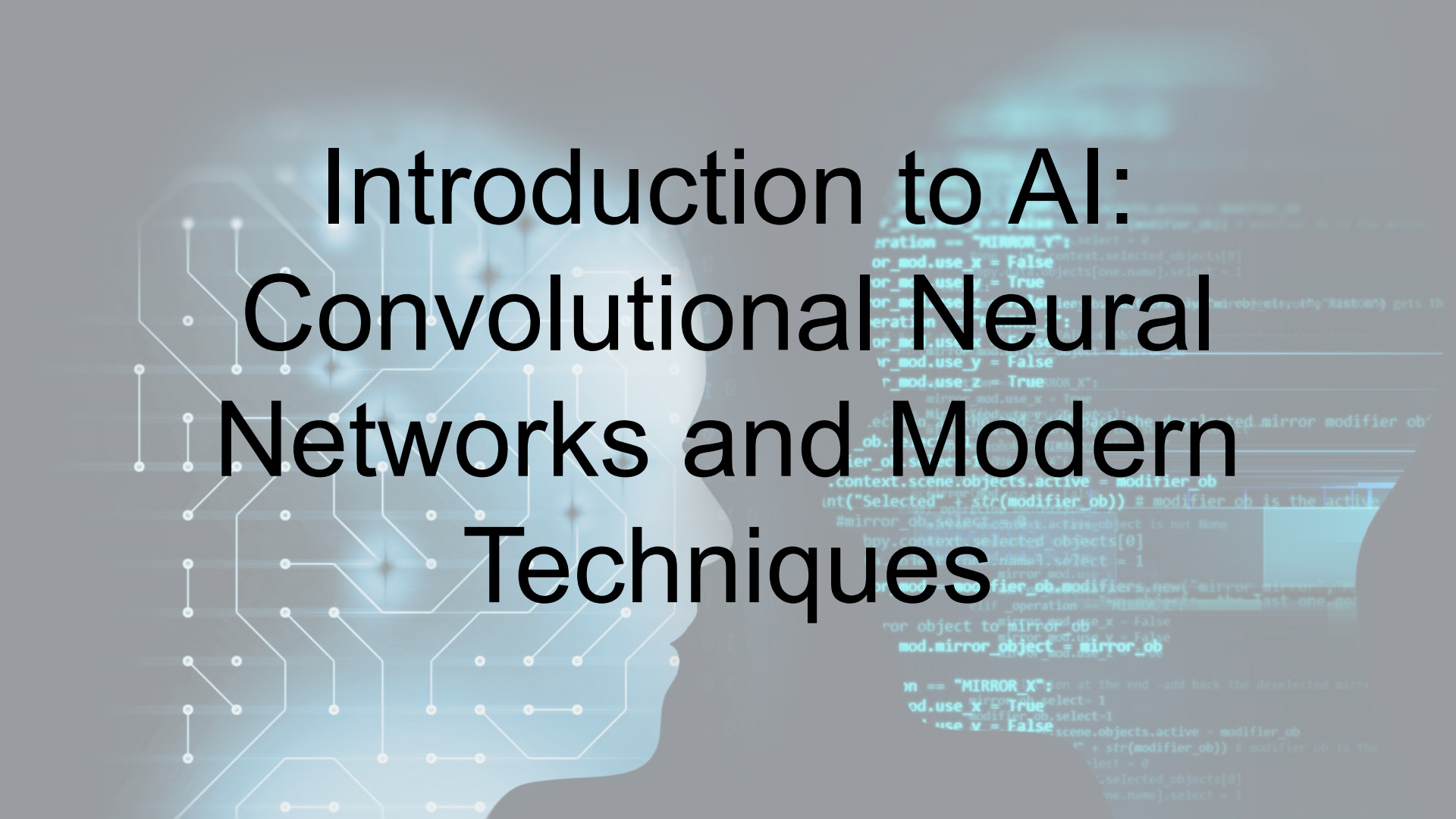
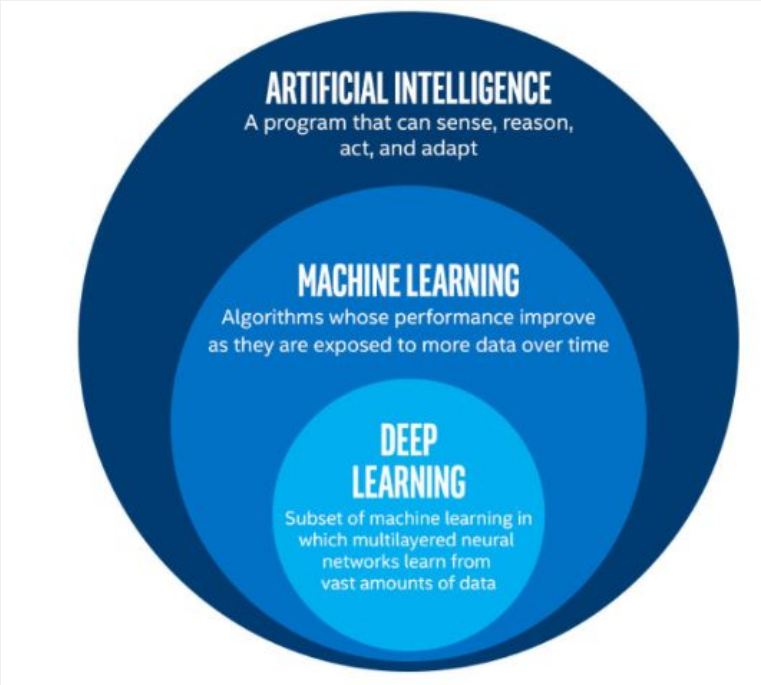
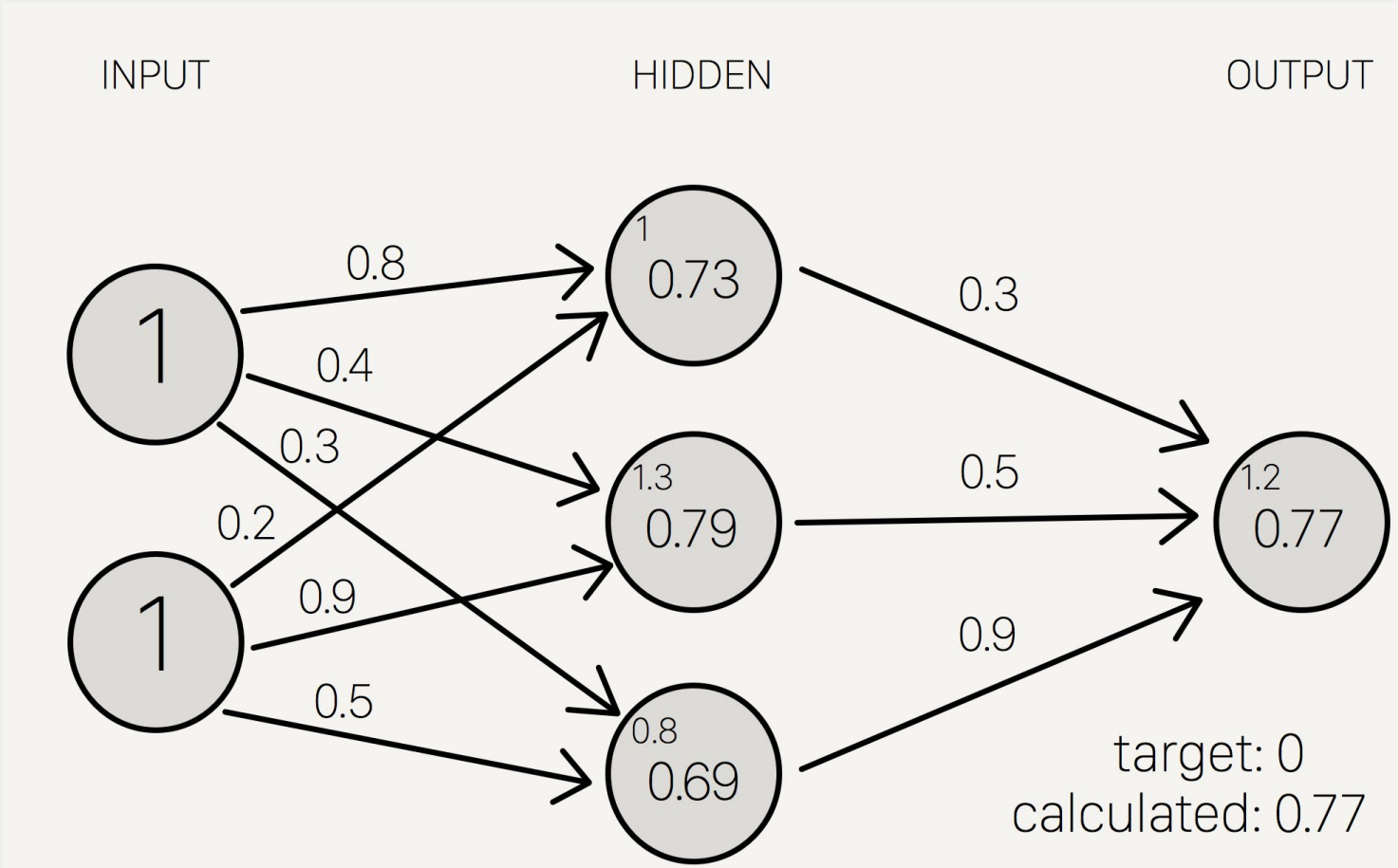


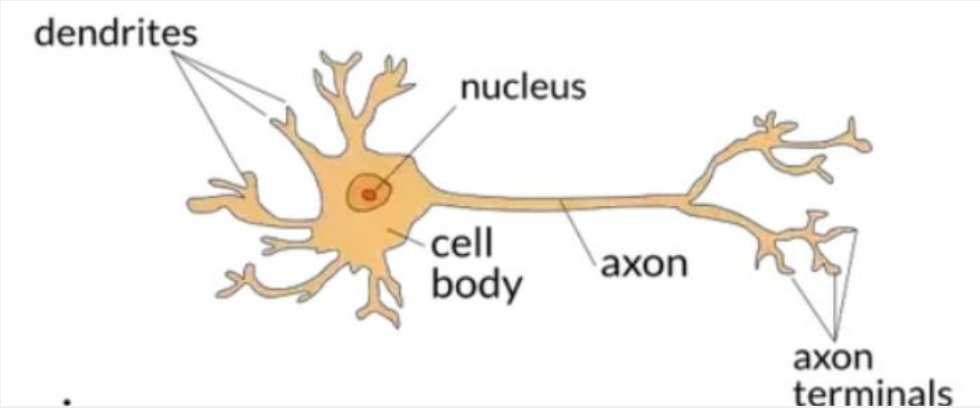
Introduction to AI: Convolutional Neural Networks and Modern Techniques

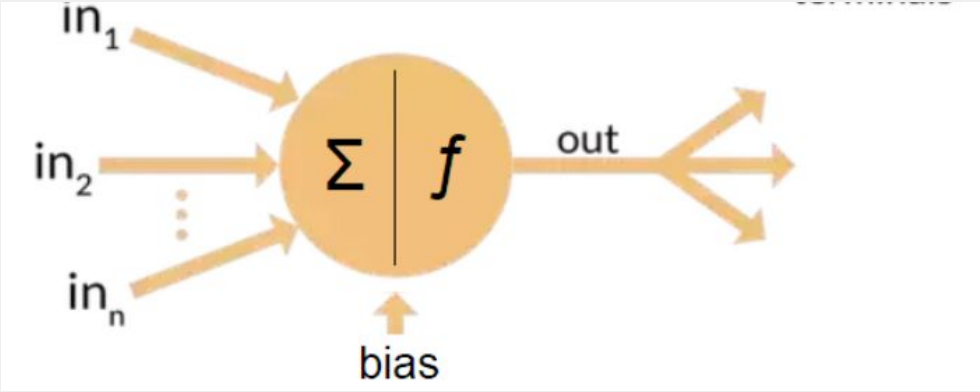
Previously in Session 1...



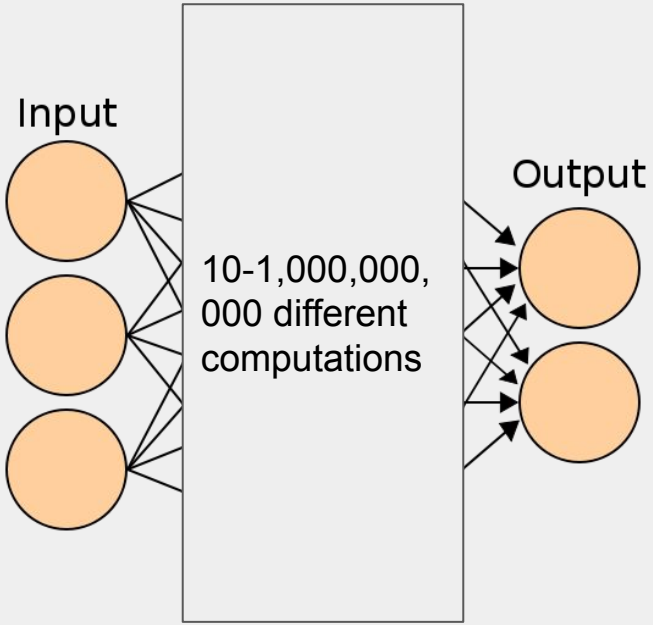

INPUT
HIDDEN
OUTPUT
1
1
0.8
0.4
0.3
0.2
0.9
0.5
1
0.73
1.3
0.79
0.8
0.69
0.3
0.5
0.9
1.2
0.77
target: 0
calculated: 0.77

Analogy

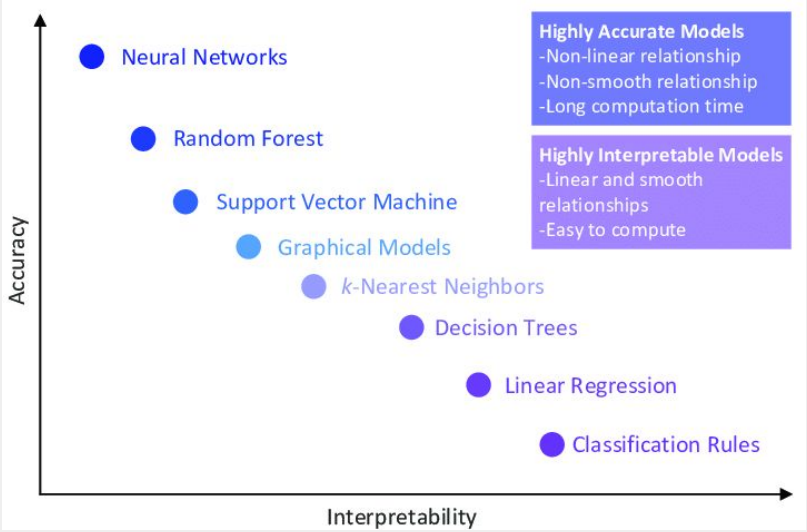

dendrites
nucleus
cell body
axon
axon terminals

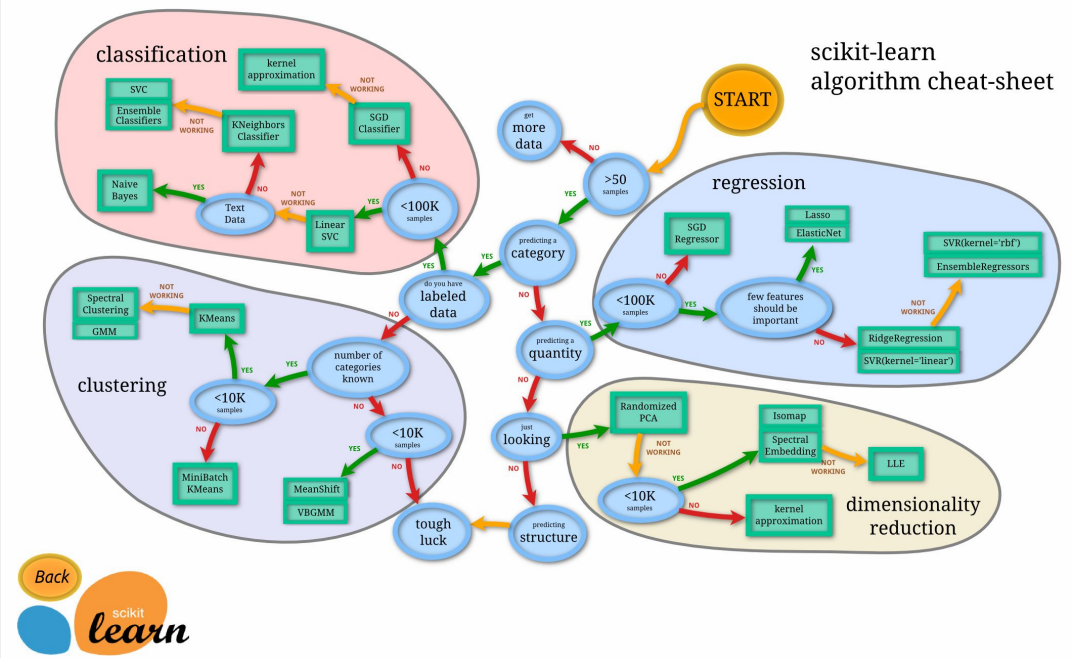

in1
in2
inn
Σ
f
out
bias

Hard to interpret hidden layers, “black box” effect



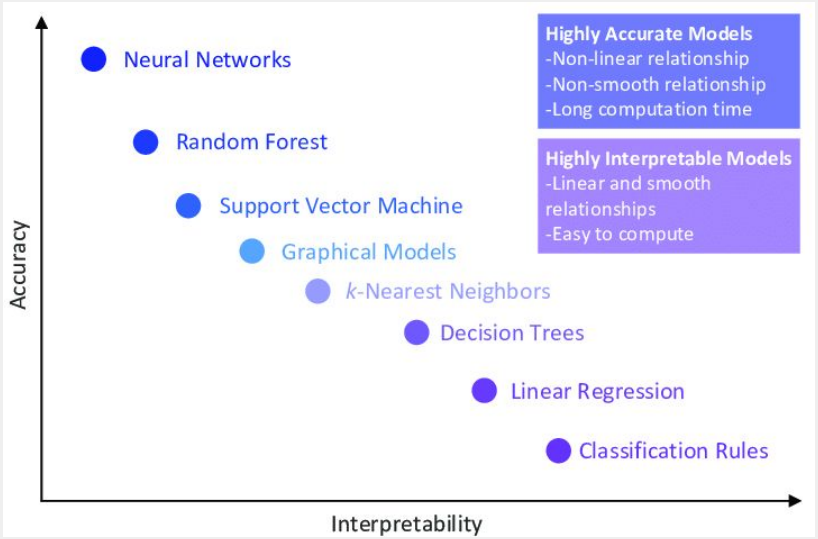
ML Models Cheat Sheet



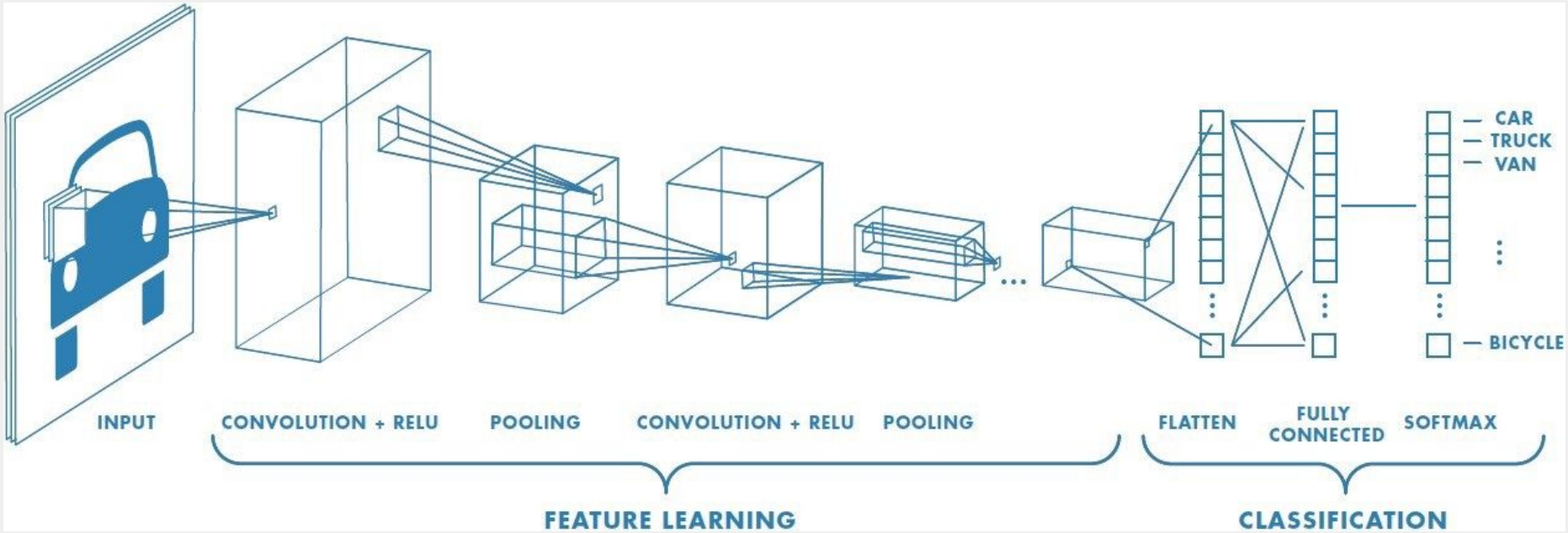


Alternatives to NNs - Demo

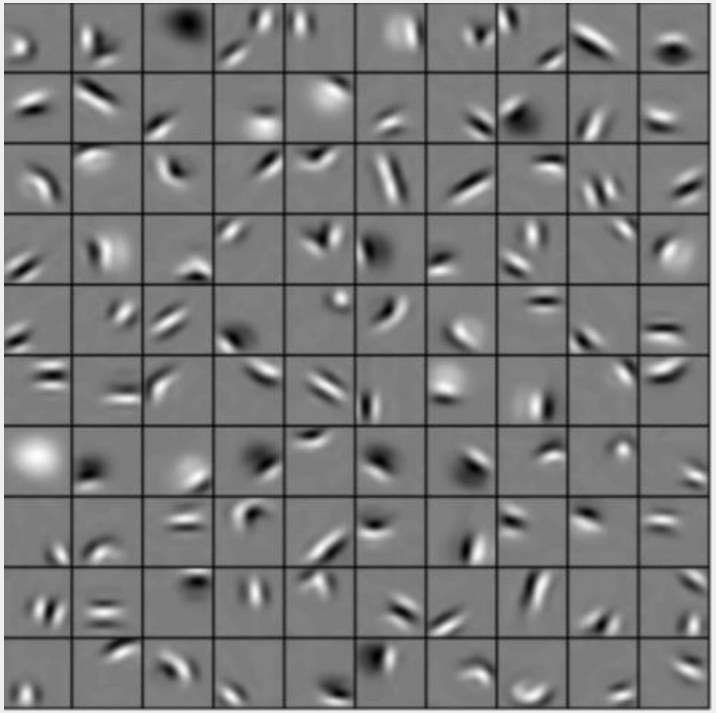
www.ml-playground.com



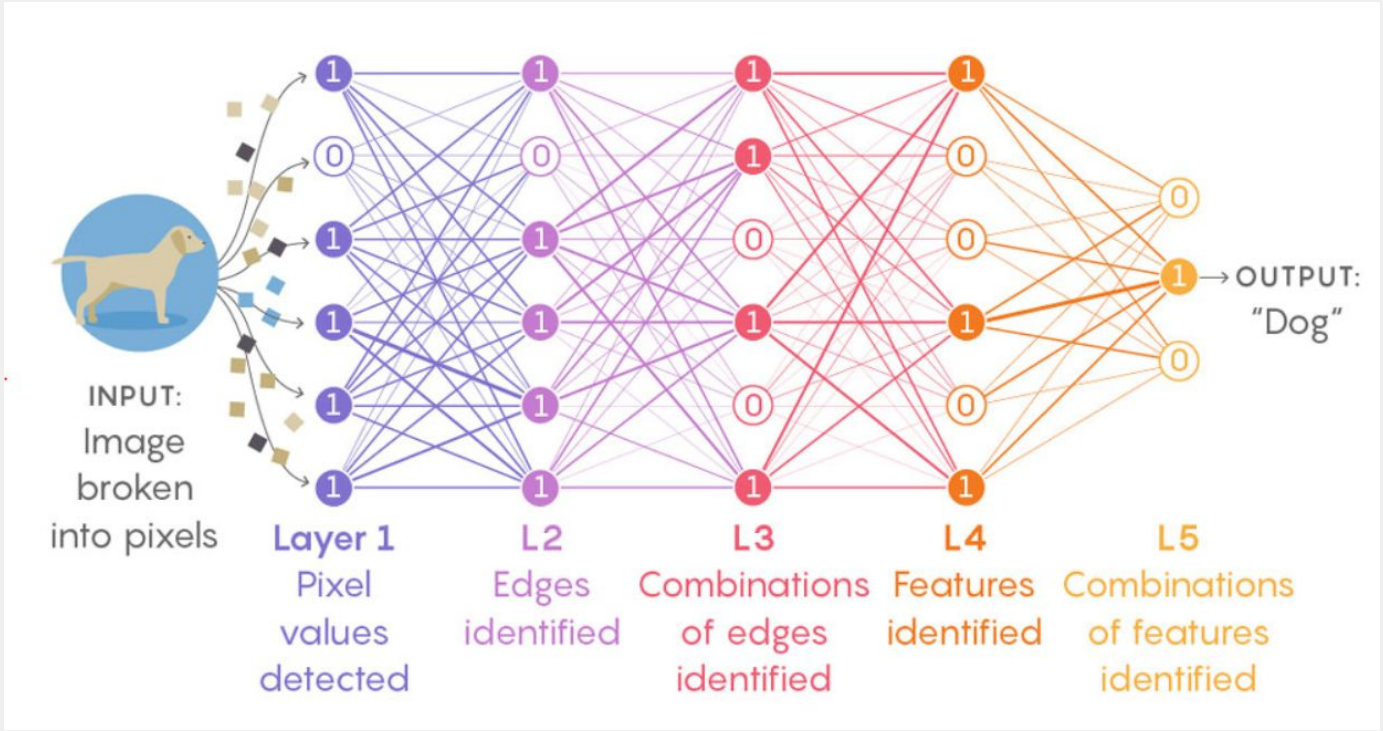
Convolutional Neural Networks



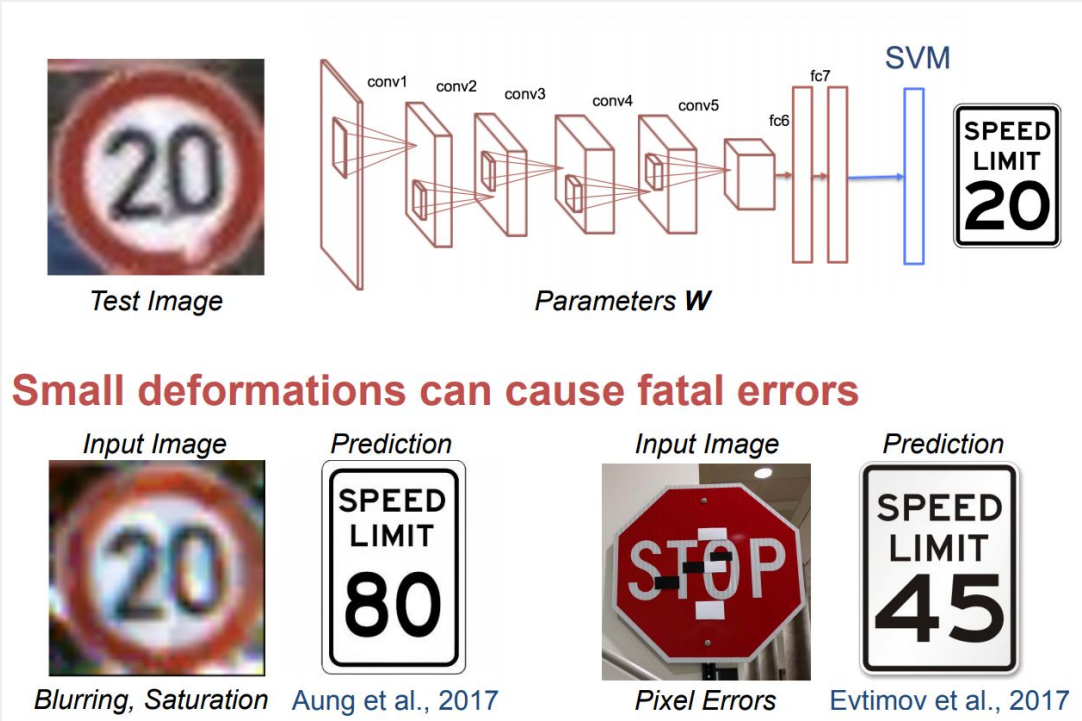
What does a neuron look like?



Generally more detailed when deeper, but lower res



Neural Networks are very sensitive



Slide reproduced from 'C19 - Machine Learning', M. Pawan Kumar, University of Oxford

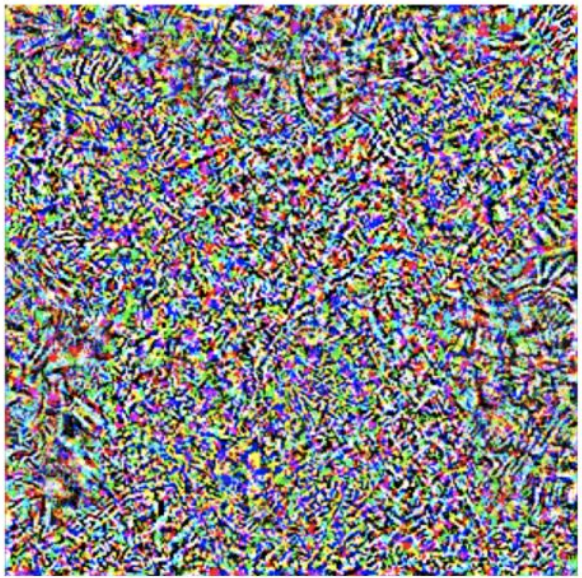
Spot the difference

Slide reproduced from ‘C19 - Machine Learning’, M. Pawan Kumar, University of Oxford

Spot the difference

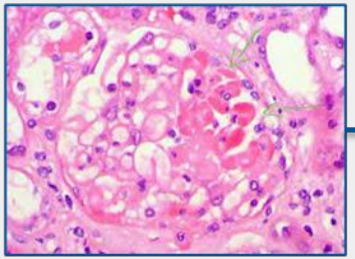
The difference

0.005 x



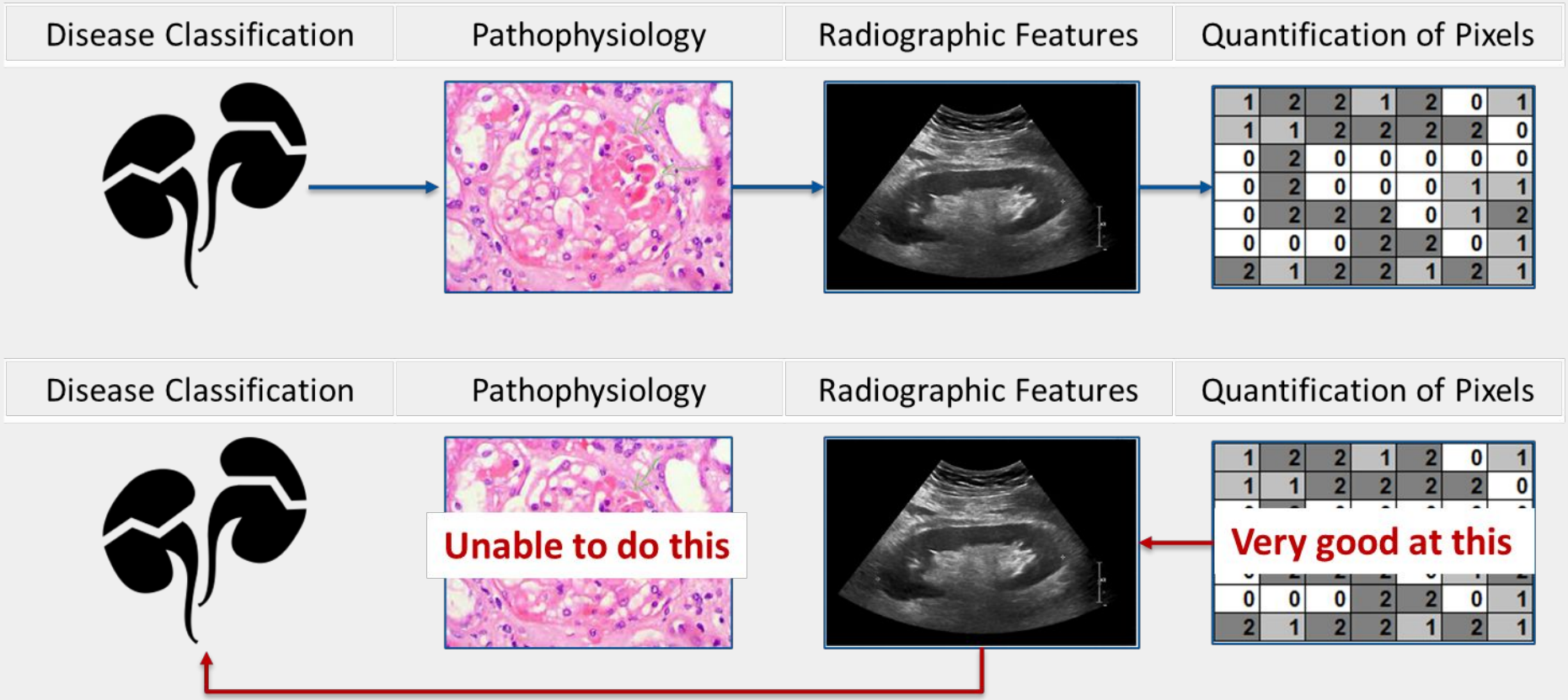
Slide reproduced from ‘C19 - Machine Learning’, M. Pawan Kumar, University of Oxford

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

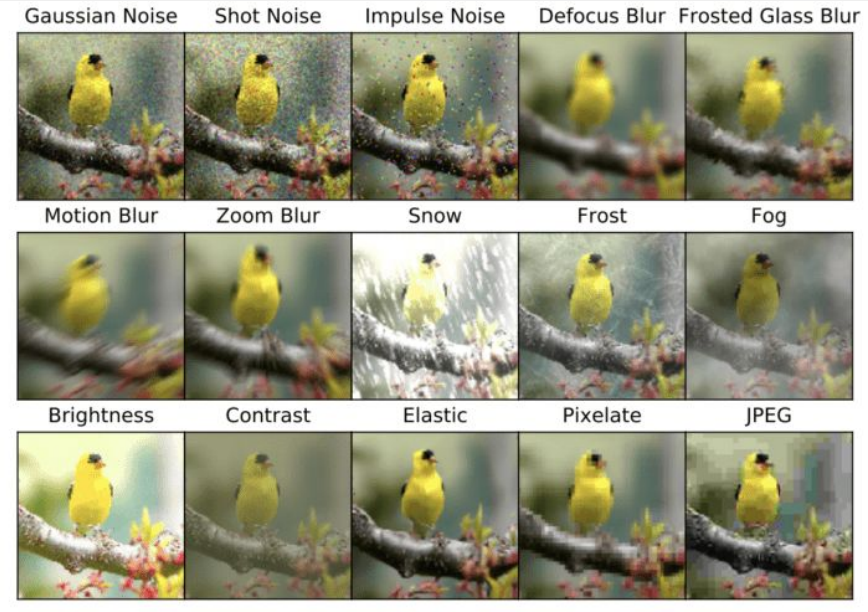
Humans: first principles “bottom up” approach



AI: “top down” approach

Training CNNs to be more Robust

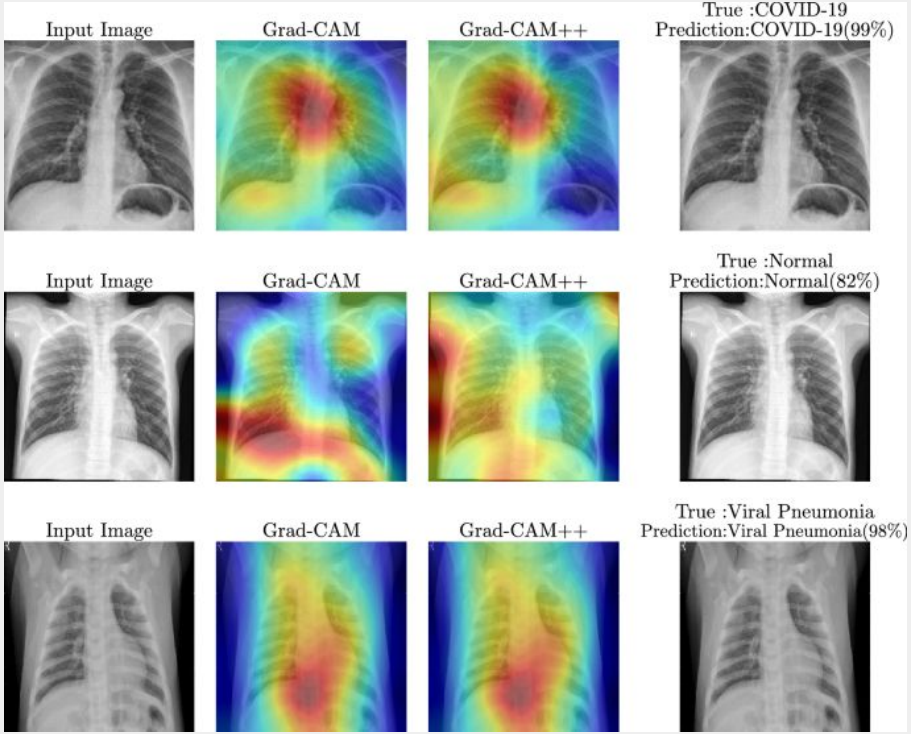
Data Augmentation



Training CNNs to be more Robust

Data Augmentation

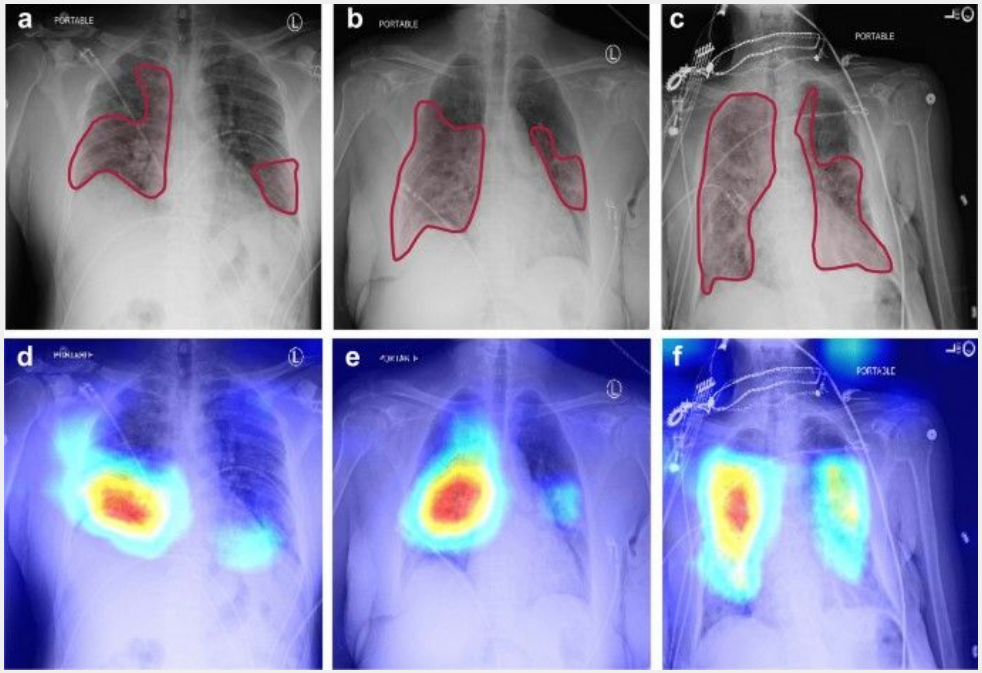
Interpretability - gradient class activation maps



Last layer - where is the most detailed neurons?

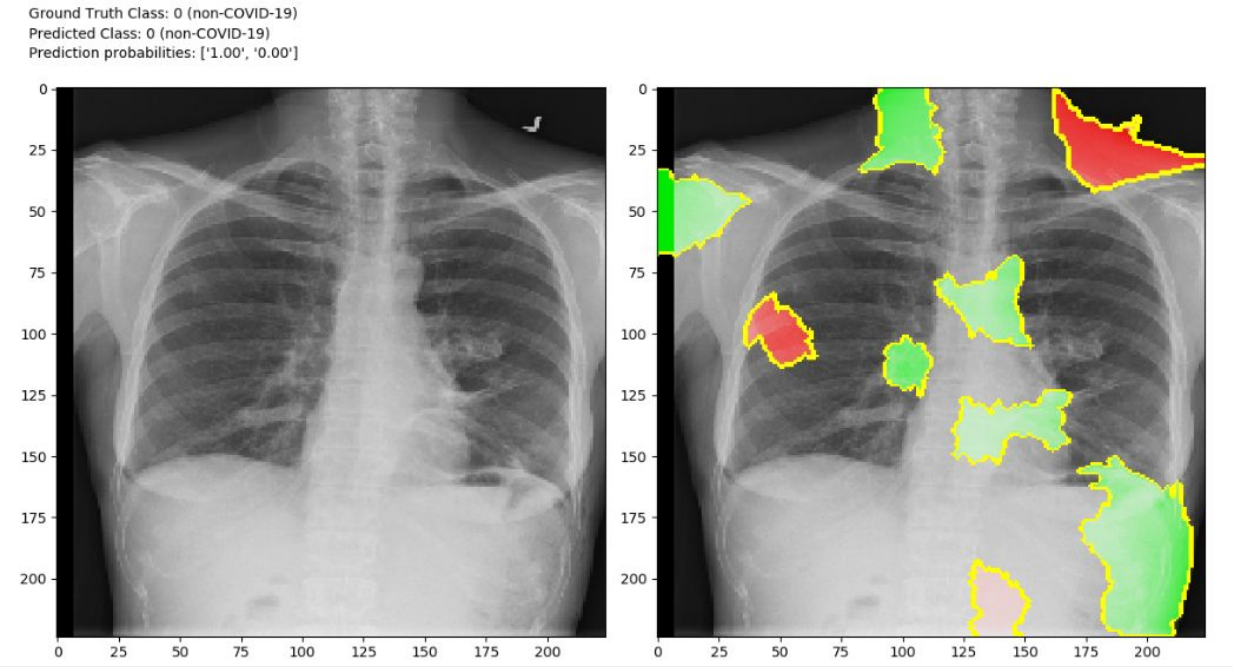
Caveat: **LOW resolution** (last layer usually 24x24, likely artificially upsampled)

Interpretability - saliency map



Similar, but uses deconvolution. Works better for **earlier layers** (high resolution, low complexity)

Interpretability - Local Interpretable Model-Agnostic Explanations



Fits linear model to approximate local regions, “local sampling” to see what’s contributory

Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy

Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, Kunlin Cao, Daliang Liu, Guisheng Wang, ... Show all authors

Table 2: Summary of Diseases in the Training and Independent Testing Datasets

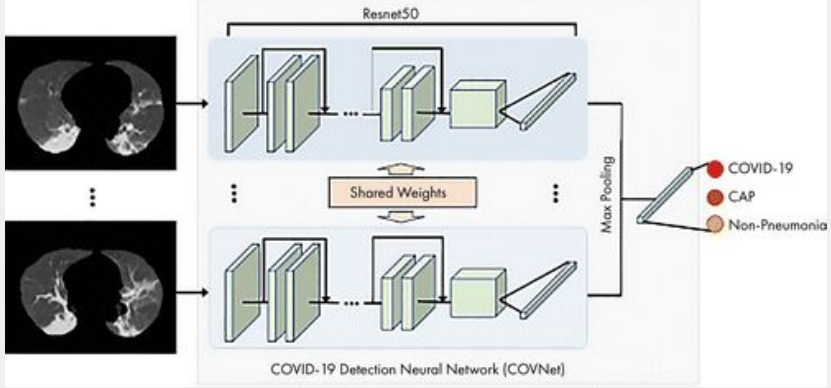
Dataset and Category	No. of Patients	Bacterial Culture Positive	Bacterial Culture Negative	Bacterial Culture Status Unknown	Normal or Noninfectious Lung Disease	Other Infectious Lung Disease
Training set						
COVID-19	400 (1165 scans)	...	...	...	...	...
CAP	1396	Bacterial pneumonia, *n* = 72; bacterial and *Mycoplasma* pneumonia, *n* = 26	Viral pneumonia, *n* = 24; *Mycoplasma* pneumonia, *n* = 60; *Pneumocystis carinii* pneumonia, *n* = 2	1212	...	...
Non-pneumonia	1173	...	...	...	Normal, *n* = 459; nodule, *n* = 500; COPD, *n* = 107; other, *n* = 13; CHF, NA; drug reaction, NA	Chronic inflammation, *n* = 235
Testing set						
COVID-19	68 (127 scans)	...	...	...	...	...
CAP	155	Bacterial pneumonia, *n* = 13; bacterial and *Mycoplasma* Pneumonia, *n* = 1	Viral pneumonia, *n* = 7; *Mycoplasma* pneumonia, *n* = 5	129	...	...
Non-pneumonia	130	...	...	...	Normal, *n* = 51; nodule, *n* = 52; COPD, *n* = 8; other, *n* = 1; CHF, NA; drug reaction, NA	Chronic inflammation, *n* = 29

Note.—Two-hundred ten patients with CAP received laboratory confirmation of the disease cause. CAP = community-acquired pneumonia, CHF = congestive heart failure, COPD = chronic obstructive pulmonary disease, COVID-19 = coronavirus disease 2019, NA = not available, RT-PCR = reverse-transcription polymerase chain reaction.

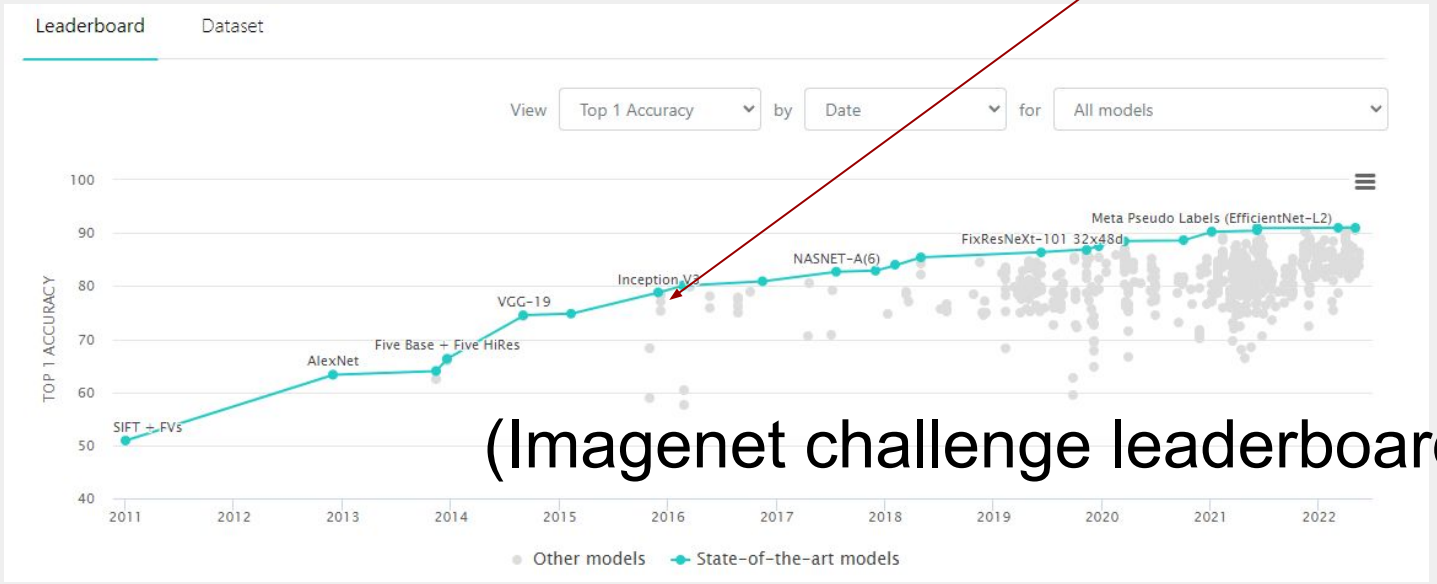
Multiple centers (nice)

No resampling (CV or bootstrap) (not nice)

No subgroup analysis



It's a CNN classifier based on ResNet-50



(Imagenet challenge leaderboards)

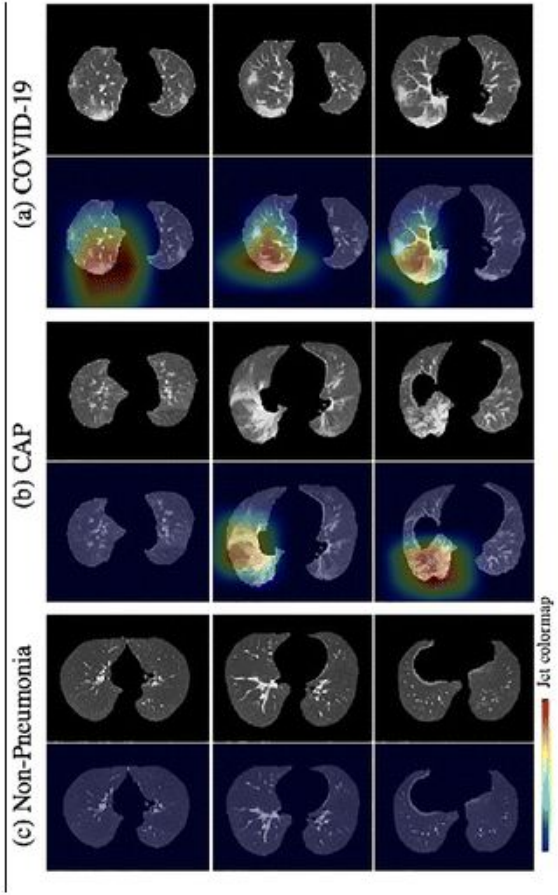


Figure 4: Representative examples of attention heatmaps generated by using the gradient-weighted class activation mapping, or Grad-CAM, method for, *A*, coronavirus disease 2019 (COVID-19), *B*, community-

Their gradCAM

Table 3: Performance of Deep Learning Framework COVNet on the Independent Testing Set

Group	Sensitivity (%)	Specificity (%)	AUC	*P* Value
COVID-19	90 (114 of 127) [83, 94]	96 (294 of 307) [93, 98]	0.96 [0.94, 0.99]	<.001
CAP	87 (152 of 175) [81, 91]	92 (239 of 259) [88, 95]	0.95 [0.93, 0.97]	<.001
Non-pneumonia	94 (124 of 132) [88, 97]	96 (291 of 302) [94, 98]	0.98 [0.97, 0.99]	<.001

Note.—Values in parentheses are the numbers of scans for the percentage calculation. Values in brackets are 95% confidence intervals. AUC = area under the receiver operating characteristic curve, CAP = community-acquired pneumonia, COVID-19 = coronavirus disease 2019, COVNet = COVID-19 detection neural network.

Overfitted? Likely. CT = even more complexity than CXR

Is there a signal? Likely. Need translational studies to assess if added value over radiologists

Original Research

External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review

Alice C. Yu, Bahram Mohajer, John Eng

Author Affiliations

Published Online: May 4 2022 | https://doi.org/10.1148/ryai.210064

Results

Eighty-three studies reporting 86 algorithms were included. The vast majority (70 of 86, 81%) reported at least some decrease in external performance compared with internal performance, with nearly half (42 of 86, 49%) reporting at least a modest decrease (≥0.05 on the unit scale) and nearly a quarter (21 of 86, 24%) reporting a substantial decrease (≥0.10 on the unit scale). No study characteristics were found to be associated with the difference between internal and external performance.

CNNs in Medicine - CheXNeXt

PLOS | MEDICINE

RESEARCH ARTICLE

Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists

Pranav Rajpurkar[1‡*], Jeremy Irvin[1‡], Robyn L. Ball[2], Kaylie Zhu[1], Brandon Yang[1], Hershel Mehta[1], Tony Duan[1], Daisy Ding[1], Aarti Bagul[1], Curtis P. Langlotz[3], Bhavik N. Patel[3], Kristen W. Yeom[3], Katie Shpanskaya[3], Francis G. Blankenberg[3], Jayne Seekins[3], Timothy J. Amrhein[4], David A. Mong[5], Safwan S. Halabi[3], Evan J. Zucker[3], Andrew Y. Ng[1☯], Matthew P. Lungren[3☯]

CNNs in Medicine - CheXNeXt

Methods - Data

- Data: 112,120 frontal-view (both posteroanterior and anteroposterior) chest radiographs of 30,805 unique patients.

- Each image annotated with up to 14 different thoracic pathology labels

CNNs in Medicine - CheXNeXt

Methods - Data Splitting

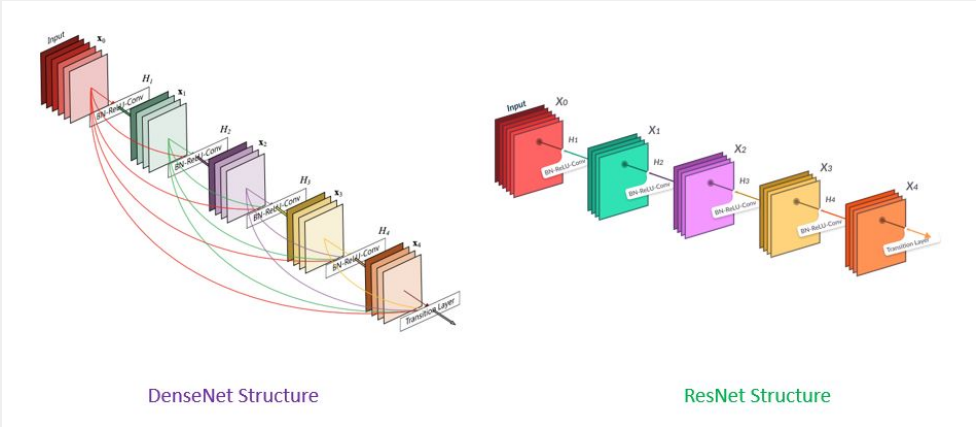
S1 Table. Summary Statistics of Training, Tuning, and Validation Datasets.

Pathology	Train ChestX-ray14 Labels No. (%)	Tuning ChestX-ray14 Labels No. (%)	Validation Labels No. (%)
Atelectasis	10053 (10.2)	671 (10.6)	184 (43.8)
Cardiomegaly	2293 (2.3)	110 (1.7)	141 (33.6)
Consolidation	3869 (3.9)	262 (4.1)	106 (25.2)
Edema	1820 (1.8)	130 (2.0)	66 (15.7)
Effusion	11326 (11.5)	752 (11.8)	129 (30.7)
Emphysema	2011 (2.0)	135 (2.1)	12 (2.9)
Fibrosis	1414 (1.4)	89 (1.4)	24 (5.7)
Hernia	110 (0.1)	4 (0.1)	31 (7.4)
Infiltration	16947 (17.2)	1186 (18.7)	53 (12.6)
Mass	4878 (4.9)	347 (5.5)	61 (14.5)
Nodule	5437 (5.5)	374 (5.9)	71 (16.9)
Pleural Thickening	2802 (2.8)	175 (2.8)	83 (19.8)
Pneumonia	1107 (1.1)	65 (1.0)	40 (9.5)
Pneumothorax	4360 (4.4)	237 (3.7)	45 (10.7)
Total No. of Images[a]	98637	6351	420
Total No. of Patients	28744	1672	389

CNNs in Medicine - CheXNeXt

Methods - CNN Model

- The neural network used in this study is a 121-layer DenseNet architecture



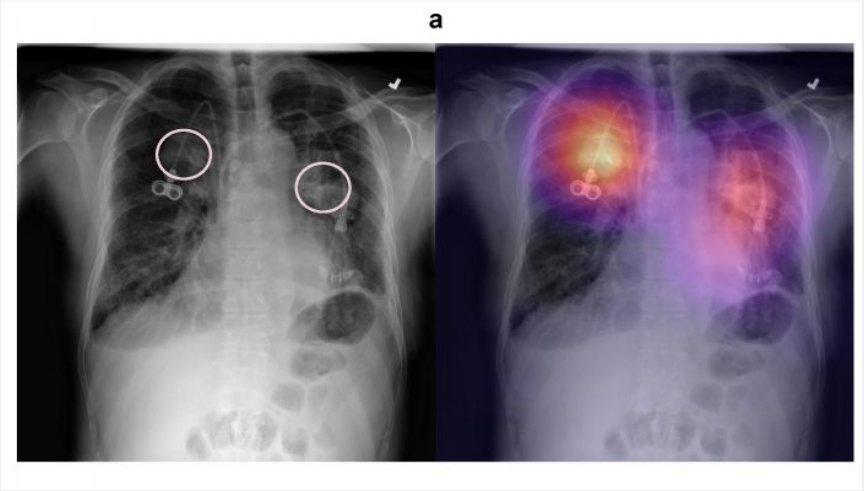
CNNs in Medicine - CheXNeXt

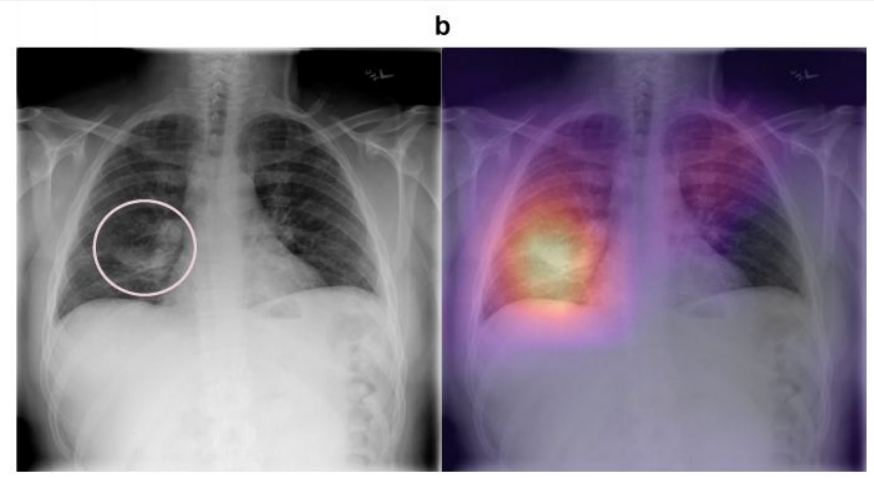
Methods - Designing the Test/Validation

- Validation set of 420 frontal-view chest radiographs
- 50+ cases of each pathology
- Annotated by 3 independent radiologists for each of the 14 pathologies

CNNs in Medicine - CheXNeXt

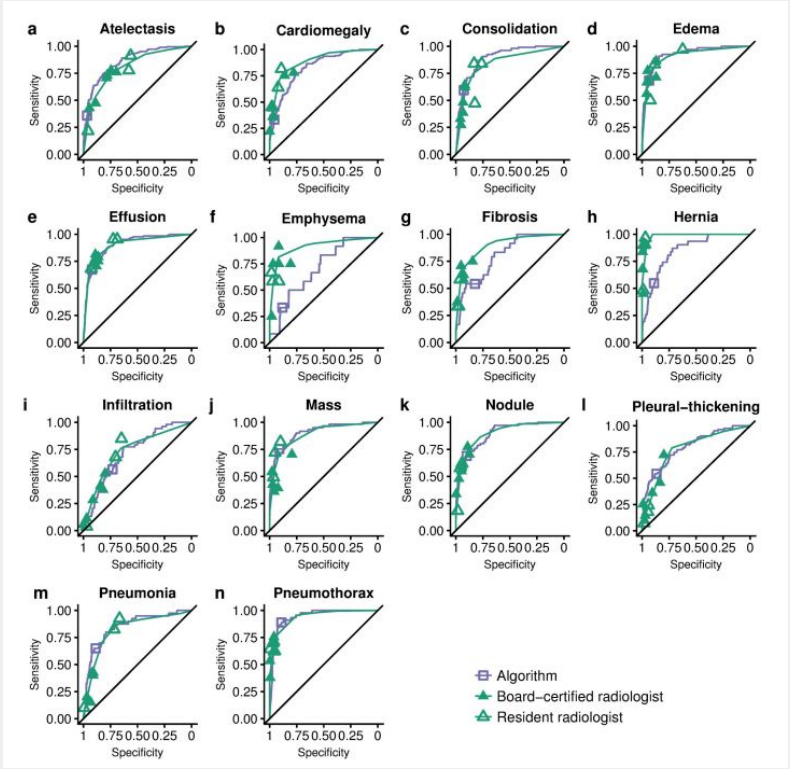
Results - Visual





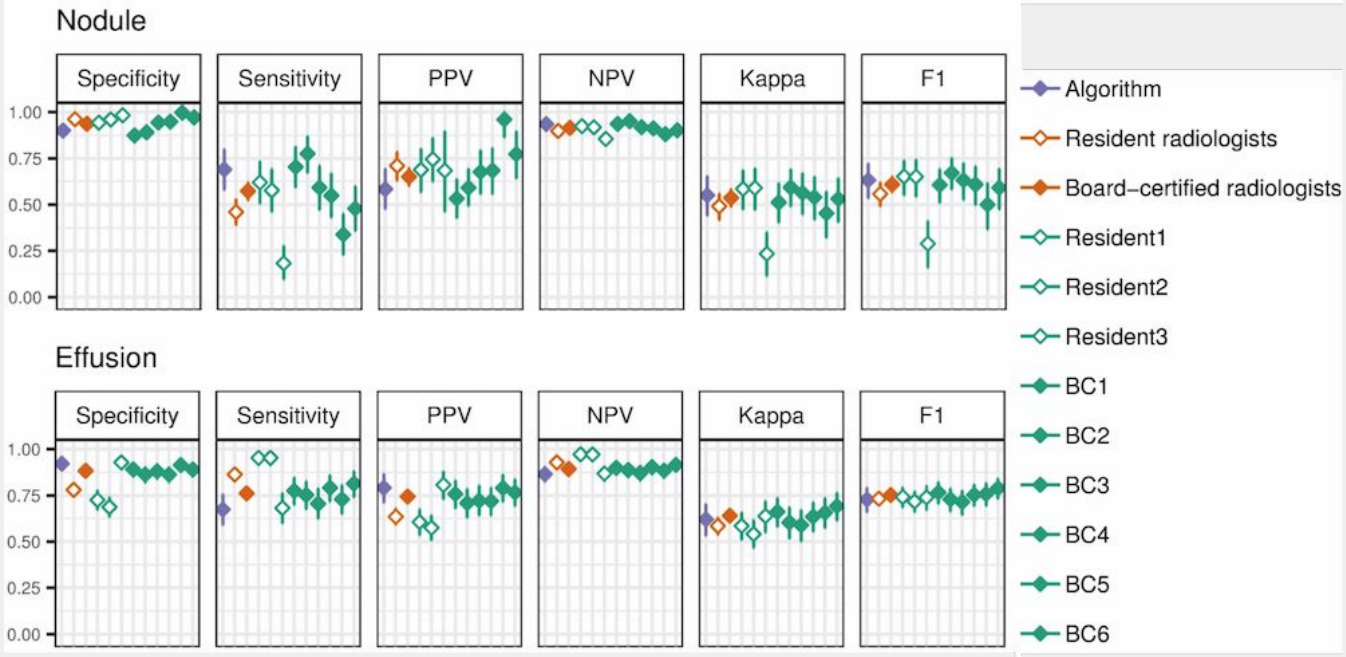
CNNs in Medicine - CheXNeXt

Results - Quantitative



CNNs in Medicine - CheXNeXt

Results - Quantitative



CNNs in Medicine - CheXNeXt

Discussion - A win for AI?

- **No** - Not as accurate as a panel of radiologists
- **Yes?** - Mostly the same as experts
- **Yes** - 420 X-rays in 90 seconds, radiologists = mean of 4 hours
- **Yes -** Does not require an expert, increasing efficiency

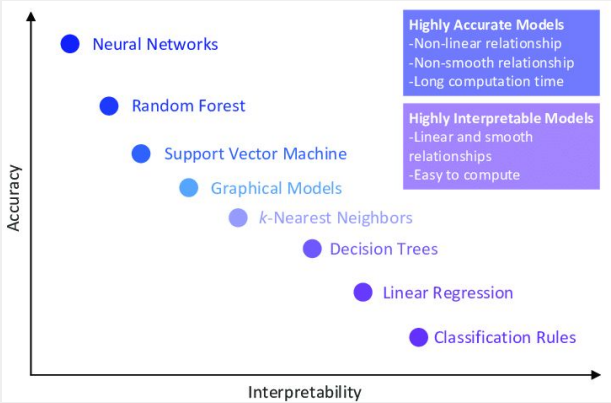
Proof of concept, potential screening tool

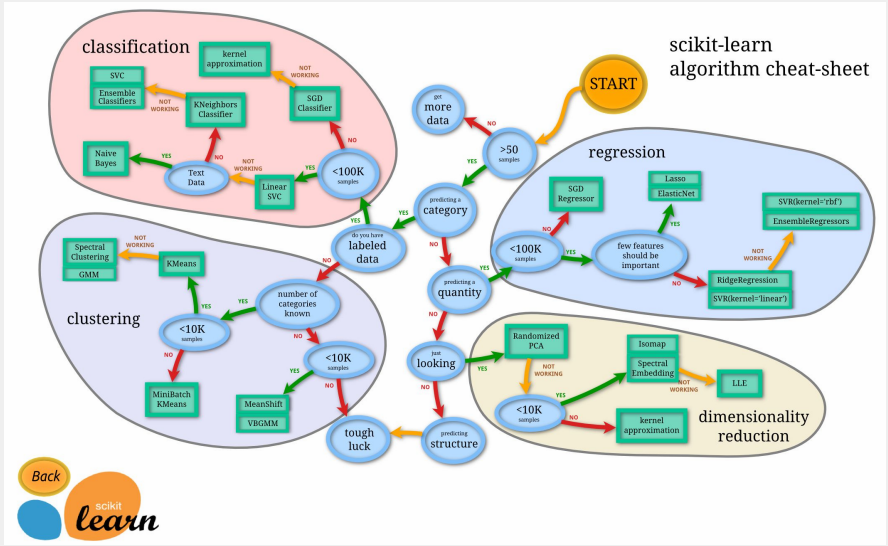
Summary

- AI: automatic task; ML: iteratively improving model; DL; lots of intermediate layers
- Steps in ML:
  - Define task, select variables, clean data, preprocess (feature select, dim transform), select model(s), train model, test
- Task definition:
  - Classification vs. regression; ask what is gold standard/comparator?
- Data requirements:
  - Heuristic: samples = 10*number of degrees of freedom (e.g. 10*# of feats in logreg), not hard rule. General gestalt: proof of concept ~75-300 samples, clinical trial/validation: 1000s-10000s
  - Impossible to get 10000 for uncommon events!
- Data cleaning/preprocessing:
  - Remove outliers, irrelevant features - too many feats = **overfitting!**
  - Feature select options: simple rank by odds ratios, PCA, collinearity filter (e.g. VIF), or repeated logreg/RF and remove low importance features
- Model selection:
  - Point-data models: classification rule, logreg, decision tree, KNN, SVM, XGB, RF, NN (increasing complexity)
  - For images / multidimension signals: convolutional neural network
  - Feature transforming image: radiomics, then point data model
- Validation:
  - Cross validate or bootstrap! Can do holdout-set with CV. Usually 80/20 train/test ratio
  - Stratify train/test to be balanced
- Interpretation:
  - Odds ratios/feature importances for point-data models
  - Top-down methods for CNNs: saliency map, gradCAM, LIME

Good references

- AI replacing rads?
  - Not likely, only CXR and mammo have good accuracies (Rodriguez-Ruiz et al. (2019)), others not enough data and too complex; bottom-up (rads) vs. top-down (AI) paradigm
  - Very few FDA approved technologies, all have limited indications (Leeuwen et al. (2021))
- Old vs. new paradigm
  - Manual scoring systems: logreg/hazard regression, then nomogram/classification rule, now ML does the “nomogram” part too
- Checklists for “screening” AI:
  - TRIPOD: checklist for predictive analysis
  - PROBAST: checklist for risk of bias
  - RQS: checklist for radiomic feature extraction
- Good engineering journals/conferences
  - Foundational science: CVPR (top), NeuroIPS, ICML, ICCV
  - Specific to med imaging: MICCAI (top conference), IEEE TMI
- Model cheatsheet:





Special thanks

Content creators:
Ricky Hu, Zoe Hu, Arsalan Rizwan, Patrick Wang, Tony Li

Supervisors / Faculty Sponsors:
Dr. Kwan, Dr. Chung