How to Conduct a Machine Learning Project for Clinical Medicine

Wei-Hung Weng, MD, MMSc
MIT CSAIL

TMU
Sep 27, 2019

Massachusetts Institute of Technology

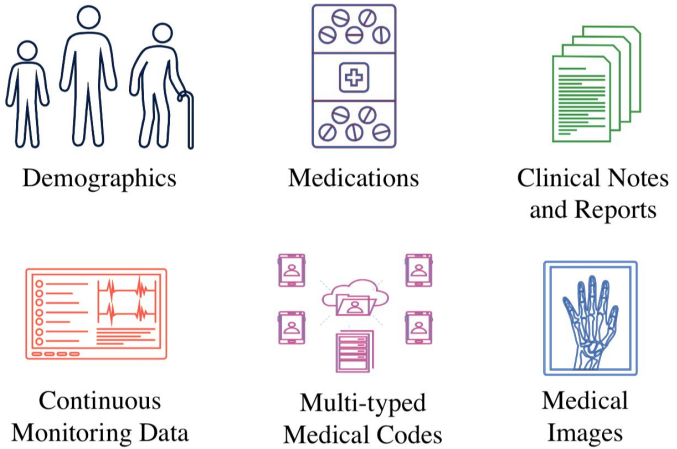
CSAIL

Why ML?

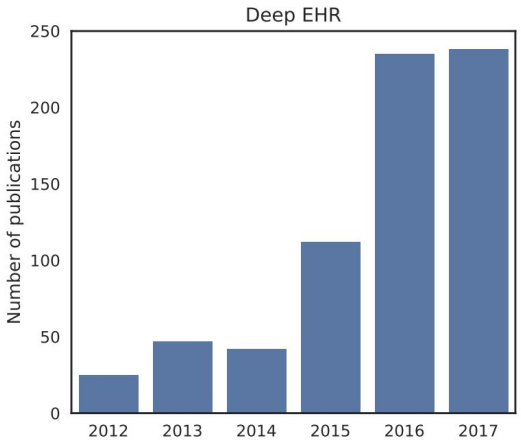
- Failure (?) of classical symbolic AI
  - A lot of knowledge is intuitive, difficult to put in rules and facts, not consciously accessible
- Principles giving rise to intelligence via learning
- ***Get knowledge directly from data and experience***
- To facilitate learning higher levels of abstraction [Bengio 2013, LeCun 2015]
  - Deep learning
- Can be described compactly since our intelligence is not just the result of a huge bag of tricks and pieces of knowledge, but of general mechanisms to acquire knowledge [Bengio 2013]

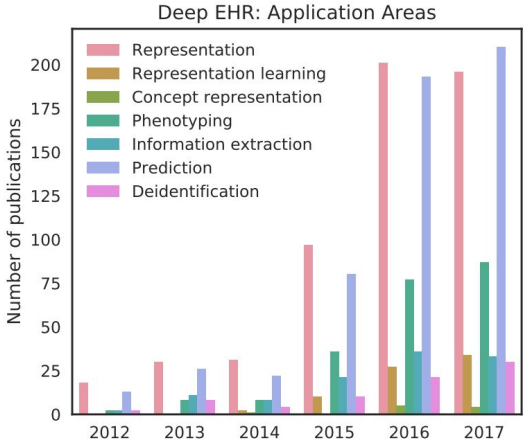
Why ML/DL for EHR?

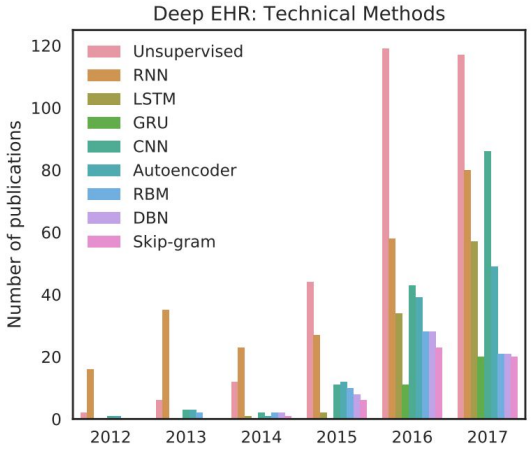
- Demographics, diagnoses, laboratory test results, medication prescriptions, clinical notes, medical images, ...
- Challenging
  - data quality (noisy, biased, ...)
  - data and annotation availability
  - heterogeneity of data types
- ***Traditional modeling → feature engineering***
  - labor intensive efforts
  - expert-defined phenotyping
  - ad-hoc feature engineering
  - limited generalizability across datasets or institutions
- ***Deep learning → learning hidden representations***
  - expert-driven feature engineering to data-driven feature construction



Deep! (2017/06)


Deep EHR
Number of publications
250
200
150
100
50
0
2012
2013
2014
2015
2016
2017

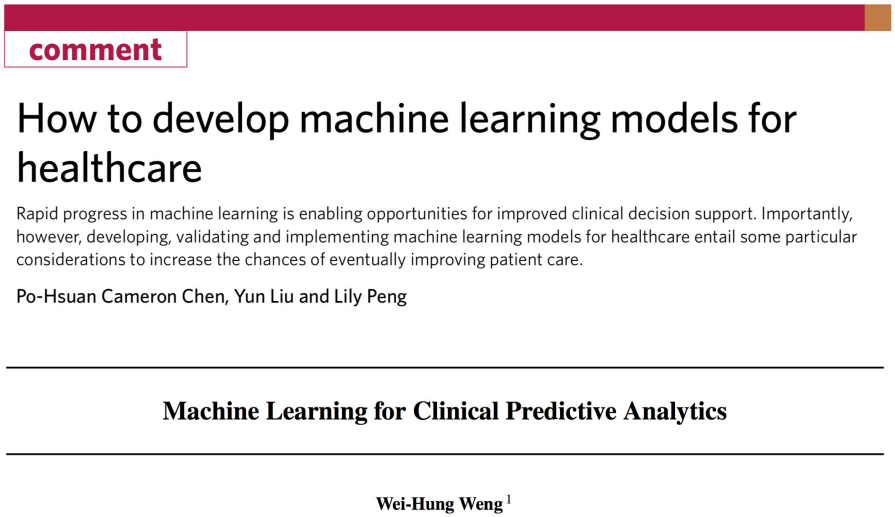

Deep EHR: Application Areas
Number of publications
200
175
150
125
100
75
50
25
0
Representation
Representation learning
Concept representation
Phenotyping
Information extraction
Prediction
Deidentification
2012
2013
2014
2015
2016
2017


Deep EHR: Technical Methods
Number of publications
120
100
80
60
40
20
0
Unsupervised
RNN
LSTM
GRU
CNN
Autoencoder
RBM
DBN
Skip-gram
2012
2013
2014
2015
2016
2017

Process

- How to develop machine learning models for healthcare [Chen 2019]
- Machine Learning for Clinical Predictive Analytics [Weng 2019]

- Problem definition
- Data curation
- ML model development
- Validation
- Assessment of clinical impact
- (Deployment and monitoring)



comment

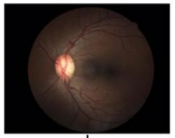
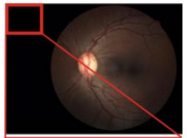
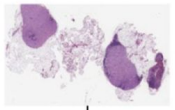
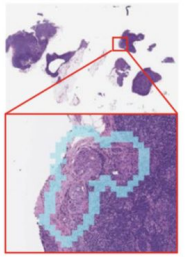
How to develop machine learning models for healthcare

Rapid progress in machine learning is enabling opportunities for improved clinical decision support. Importantly, however, developing, validating and implementing machine learning models for healthcare entail some particular considerations to increase the chances of eventually improving patient care.

Po-Hsuan Cameron Chen, Yun Liu and Lily Peng

Machine Learning for Clinical Predictive Analytics

Wei-Hung Weng [1]

Problem selection	Data collection	ML development	Validation	Assessment of impact	Deployment and monitoring
	Development dataset		Retrospective		
	128,175 images		Sensitivity: 0.90 Specificity: 0.98		
			Gulshan et al. (2016)	Model predictions None Mild Moderate Severe Proliferative	Future work
	Validation dataset		Prospective		
Referable diabetic retinopathy?	11,711 images		Sensitivity: 0.87 Specificity: 0.91	40% reduction in false negatives	
Gulshan et al. (2016) Ting et al. (2017)	Gulshan et al. (2016)	Gulshan et al. (2016)	Abramoff et al. (2018)	Sayres et al. (2018)	
	Development dataset		Retrospective		
	270 images (millions of patches)		AUC: 0.99 Sens@1*: 0.77		
			Bejnordi et al. (2017)		Future work
	Validation dataset		Retrospective		
Metastatic breast cancer?	129 images		AUC: 0.99 Sens@1*: 0.86	2× review speed ½ false negatives	
Bejnordi et al. (2017)	Bejnordi et al. (2017)	Bejnordi et al. (2017) Liu et al. (2018)	Liu et al. (2018) *: tumour detection sensitivity at one false positive per slide	Steiner et al. (2018)	

[1] Appropriate Problem Definition

- Problem selection
  - Make a meaningful impact in patient care by ***providing actionable insights***
  - ML model should only leverage input data that are available at the time when the proposed clinical decision is being made
- Defining prediction task
  - ***Learning from humans***
    - Removing human-factor bottleneck (availability and fatigue of human raters)
    - ***E.g. use fundus imaging to classify DR [Gulshan 2016]***
  - ***Enabling extraction of previously unknown insights***
    - Detection of novel signals has the potential to improve diagnosis or prognosis using cheaper and scalable modalities
    - ***E.g. use fundus imaging to predict CV risk [Poplin 2018]***
    - Must be taken to ensure that ‘novel signals’ found are not the result of confounding factors or random chance

[1] Clinical Perspective

- Cost / Risk assessment and adjustment
  - Insurance
  - Resource redistribution
- Precision / Personalized medicine
  - For oncology / rare diseases / mental disorders / …
  - Applications
    - Clinical decision support
    - Drug discovery
    - Outcome prediction
      - Lifespan prediction / Disease progression
    - Chronic disease management
      - Early prediction of blood glucose for self-management

[1] ML Perspective

- Risk stratification
- Causal inference
- Bias
- Time-series
- Modeling unstructured data
- Interpretability and explainability
- Disease progression modeling
- Reasoning and decision making (current ML probably not enough)

[1] Framing into the ML Scenario

- ***Supervised learning***
  - Regression
  - Classification
    - Linear
    - Non-linear (e.g. deep learning, SVM, decision tree, ...)
  - Structured learning
- Unsupervised learning
  - Clustering
  - Dimensionality reduction
- Transfer learning
- Reinforcement learning
- ...

[1] Appropriate Problem

- ***Verification / Evaluation method***
  - ***Independent dataset***
  - Saliency ‘heatmaps’ for ***interpretability***
  - ***Ask clinicians*** (who were blinded to the prediction task) for quantitative, unbiased evaluation
- Data availability
  - Lack of digitization (pathology slides)
  - Inaccessible because of patient privacy or commercial concerns
  - Lacking because the disease of interest is too rare

[2] Curating Datasets

- ***Data split***
  - ***Should be 'clean' with respect to patients***
  - Merging from multiple sources / patient-level overlap → image similarity to detect duplication
- The size of the validation set, information from ***clinical trials*** may be helpful
  - ***Power calculations*** helps determine the sample size required to confidently evaluate the model performance
  - All primary / secondary analyses should be pre-specified, avoiding 'post-hoc' analysis
  - Only perform exploratory analyses on the training set, and validate the hypotheses on the validation set
- Class imbalance
  - Data augmentation
  - Additional steps to ensure proper model calibration or adjustments in the evaluation metric

[2] Curating Datasets

- Data Quality
  - The notion of ‘good quality’ may be disease-specific, for example the same fundus image may be of sufficient quality for assessing glaucomatous nerve head features but not for diabetic retinopathy
- ***Reference / Ground truth***
  - Determination of the reference truth often involves subjective judgement, introducing systematic errors, random errors or both
  - ***Adjudication*** by a panel of experts may be helpful but can be slow and expensive
  - Adjudication of only a subset of the data
    - Testing dataset for final evaluation
    - Validation dataset for hyperparameter optimization during the modeling

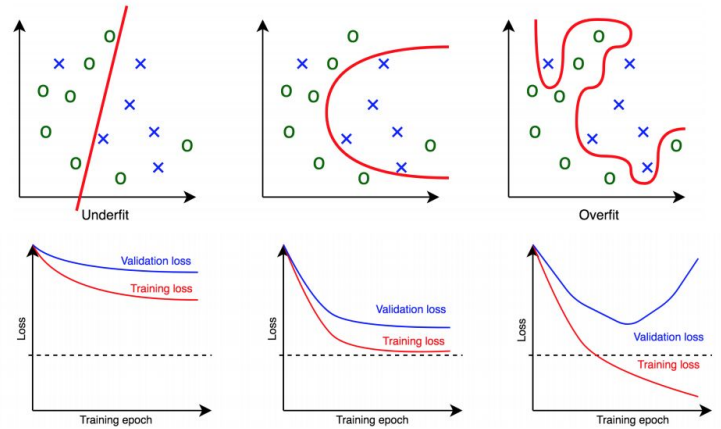
[3] Modeling

- Several considerations for model architecture design
  - Data modality and volume, model interpretability, model inference time, balancing model overfitting and underfitting, ...
- End-to-end? ***Data size & the value of intermediate outputs***
  - Better when large datasets are available and if the final performance is the primary metric of interest (not the case in healthcare)
- ***Decomposing the model***
  - Intermediate output may be useful (e.g. interpretability of the prediction)
  - Healthcare data can differ substantially across data sources → easier generalization
  - Models with a large number of parameters usually have better predictive performance
    - May require minutes to hours for inference, which may hinder intraoperative usage where time is of the essence
    - ***Real-time need***

[3] Model Generalizability

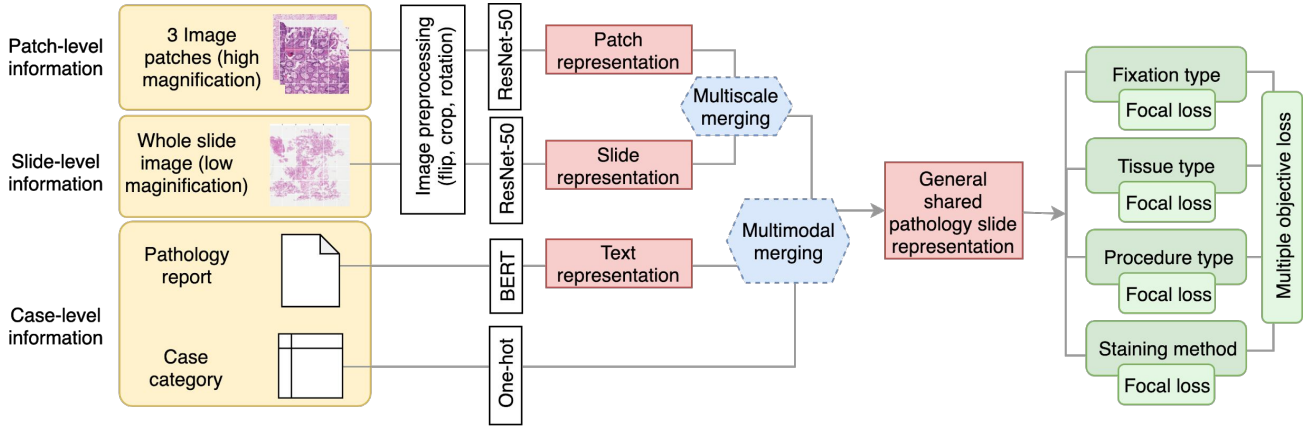
- Bias-variance trade-off
  - Balancing model underfitting/overfitting
- ***Regularization***
  - Data augmentation: particularly useful in the data-limited healthcare regime
  - Inject prior knowledge
  - Perturbations that may not correspond to real-world changes may still be helpful in learning
- ***Train-validation-test split must be carefully preserved***
  - Any violation of train-validation-test hygiene can result in ungeneralizable performance

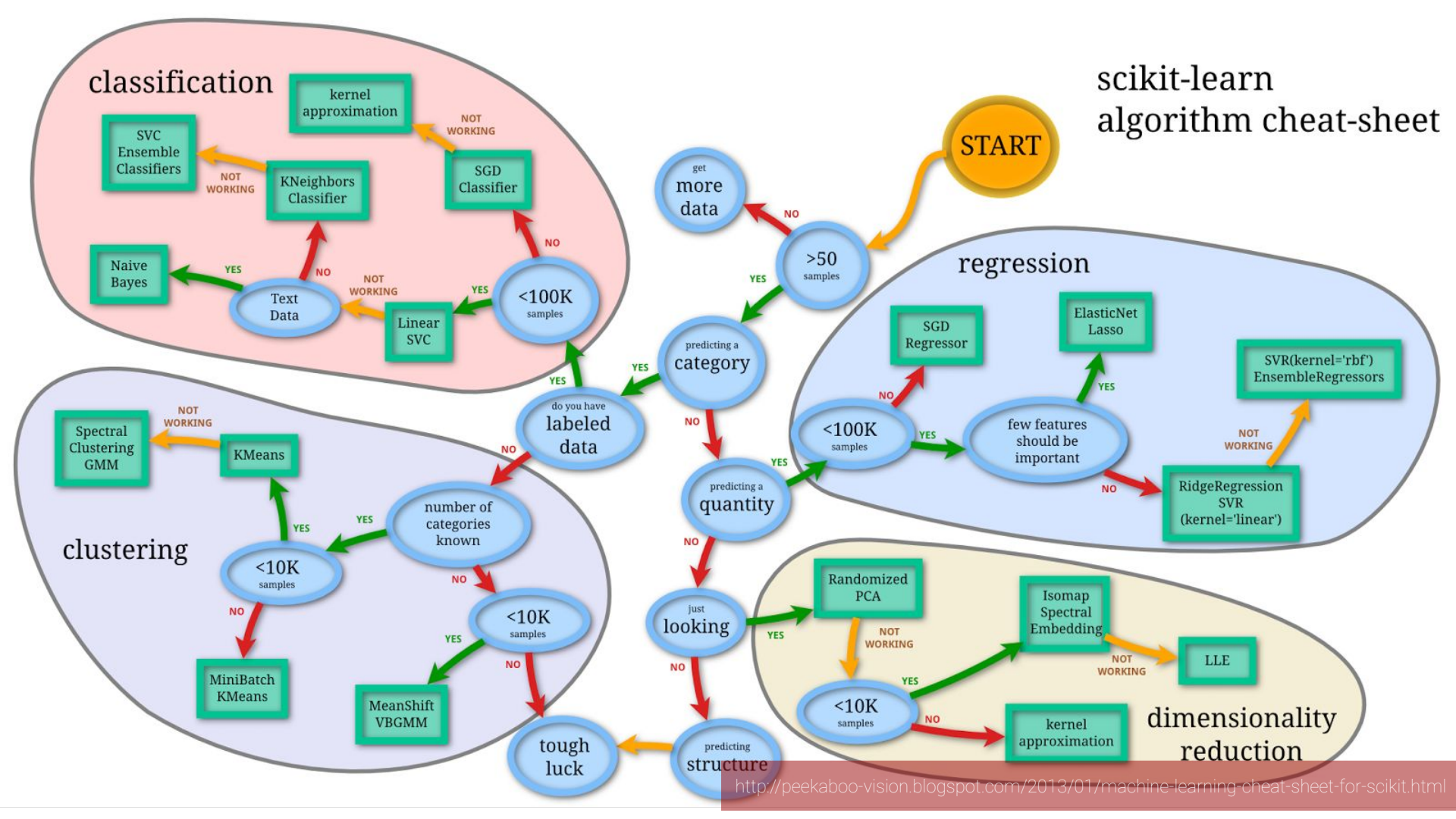
	Training error	Validation error	Approach
High bias	High	Low	Increase complexity
High variance	Low	High	Decrease complexity Add more data



[3] An Example

- Learning general pathology data representation [Weng 2019]
  - Limited pathology data, extremely class imbalance
  - Multimodal (image + text + structured data), multitask, data augmentation, neural network regularization, prior knowledge from pretraining networks, loss function design, …




scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NO
SVC Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
NO
tough luck
NO
<10K samples
YES
MeanShift VBGMM
YES
number of categories known
NO
<10K samples
NO
MiniBatch KMeans
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering GMM
clustering
http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Predicting categories

Predicting values

Three or more

Two

ANOMALY DETECTION

- **One-class SVM** — >100 features, aggressive boundary
- **PCA-based anomaly detection** — Fast training

CLUSTERING

- **K-means**

MULTI-CLASS CLASSIFICATION

- Fast training, linear model — **Multiclass logistic regression**
- Accuracy, long training times — **Multiclass neural network**
- Accuracy, fast training — **Multiclass decision forest**
- Accuracy, small memory footprint — **Multiclass decision jungle**
- Depends on the two-class classifier, see notes below — **One-v-all multiclass**

REGRESSION

- **Ordinal regression** — Data in rank ordered categories
- **Poisson regression** — Predicting event counts
- **Fast forest quantile regression** — Predicting a distribution
- **Linear regression** — Fast training, linear model
- **Bayesian linear regression** — Linear model, small data sets
- **Neural network regression** — Accuracy, long training time
- **Decision forest regression** — Accuracy, fast training
- **Boosted decision tree regression** — Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

- **Two-class SVM** — >100 features, linear model
- **Two-class averaged perceptron** — Fast training, linear model
- **Two-class logistic regression** — Fast training, linear model
- **Two-class Bayes point machine** — Fast training, linear model

- Accuracy, fast training — **Two-class decision forest**
- Accuracy, fast training, large memory footprint — **Two-class boosted decision tree**
- Accuracy, small memory footprint — **Two-class decision jungle**
- >100 features — **Two-class locally deep SVM**
- Accuracy, long training times — **Two-class neural network**

Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

[3] Modeling

- ***Always start from simpler algorithms / models***
  - e.g. LR, then random forest

[4] Evaluating Model Performance

- Evaluation metrics should be consistent with the ones used in the community!
- ***Discrimination metric***
  - ***Threshold*** selection tends to play a critical role in healthcare because clinical applications commonly involve binary decisions
  - High sensitivity for screening / high specificity for diagnosis
  - Resource constraints (time, labor effort, cost limitations for screening)
- ***Calibration metric***
  - Evaluate how well the predicted probabilities match the actual probabilities
  - Although under-reported, calibration metrics (e.g., the Hosmer-Lemeshow statistic) are crucial for real-world use because these probabilities are used for expected cost-benefit analysis
- Validation should be done using large, heterogeneous datasets to ensure generalization to diverse patient populations

[4] Evaluating Model Performance

- Subgroup analysis /cluster analysis
- Sensitivity analysis
- Data augmentation for class imbalance
  - Validation set collected may have a different distribution of disease subtypes relative to real-world populations → evaluation should be adjusted according to realistic prevalence distributions
- ***Baseline comparison***
  - Comparison with a 'human baseline'
  - Comparison with a baseline (e.g. LR) ***based on variables that are readily available in the clinic (e.g. APACHE, KDIGO, CHADS2, ...)*** may be useful to evaluate the added value of the proposed novel association

[3+4] Keys

- ***Loss (Objective) function***
- ***Metrics***
- ***Quantitative***
- ***Qualitative***
- ***Baseline***
- ***Error analysis***
- ***Ablation analysis***

Task	Error type	Loss function	Note
Regression	Mean-squared error	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Easy to learn but sensitive to outliers (MSE, L2 loss)
	Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Robust to outliers but not differentiable (MAE, L1 loss)
Classification	Cross entropy = Log loss	$-\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$ $= -\frac{1}{n}\sum_{i=1}^{n} p_i \log q_i$	Quantify the difference between two probability distributions
	Hinge loss	$\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i\hat{y}_i)$	For support vector machine
	KL divergence	$D_{KL}(p\|\|q) = \sum_i p_i(\log \frac{p_i}{q_i})$	Quantify the difference between two probability distributions

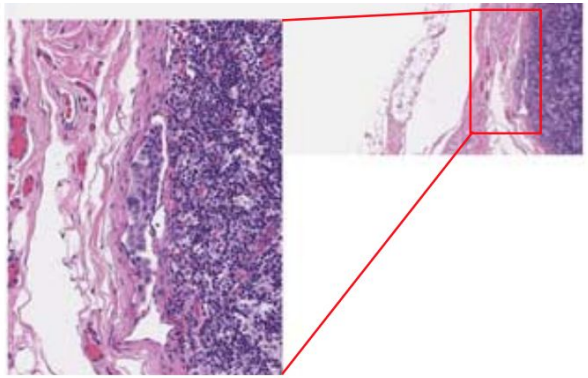
		Predicted		
		True	False	
Actual	True	True positive (TP)	False negative (FN) Type II error	Recall = Sensntivity = $\frac{TP}{TP+FN}$
	False	False positive (FP) Type I error	True negative (TN)	Specificity = $\frac{TN}{TN+FP}$
		Precision = $\frac{TP}{TP+FP}$		Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ F1 = $\frac{2\times Precision \times Recall}{Precision+Recall}$

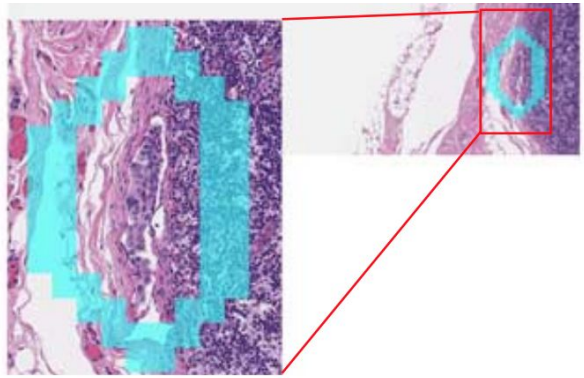
[5] Potential Clinical Impact

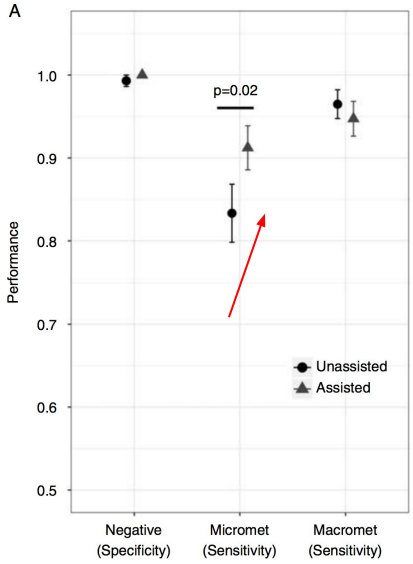
- A performant model alone is insufficient to create clinical impact
- The system has to be designed to be useful even in cases of failure such as false positives
- ‘Pre-diagnosis’ to pre-screen images and draw attention to images of high or uncertain risk, ‘peri-diagnosis’ to highlight lesions on the image during the regular image grading process, or ‘post-diagnosis’ to resurface images that may have been graded wrongly due to inexperience or fatigue
- ***User trust → human-in-the-loop development***
    - The degree of reliance can be measured by the user-model agreement rate, comparing that with the ML prediction accuracy
- ***‘Alarm fatigue’***

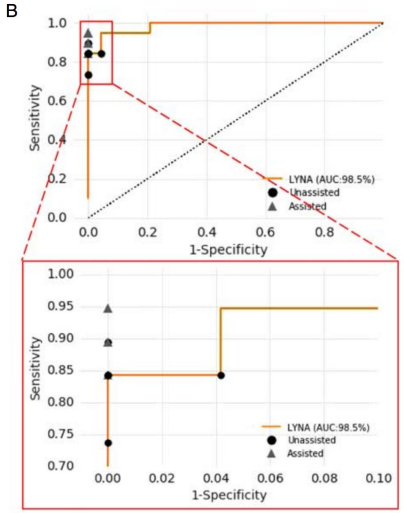
[5] An Example

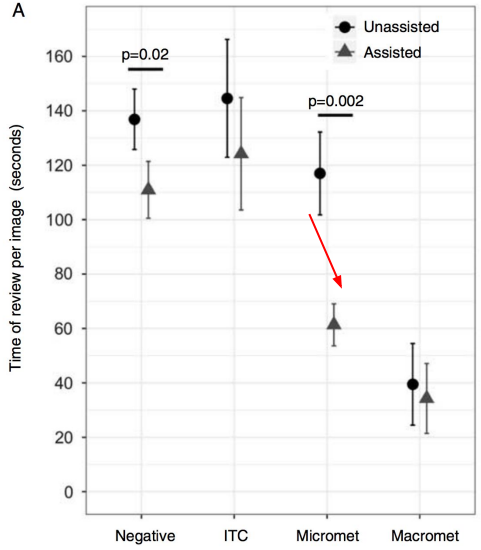
- Steiner 2018

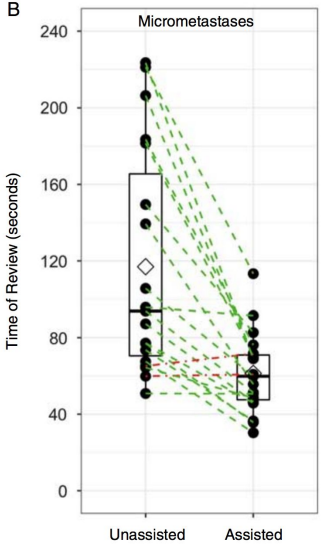




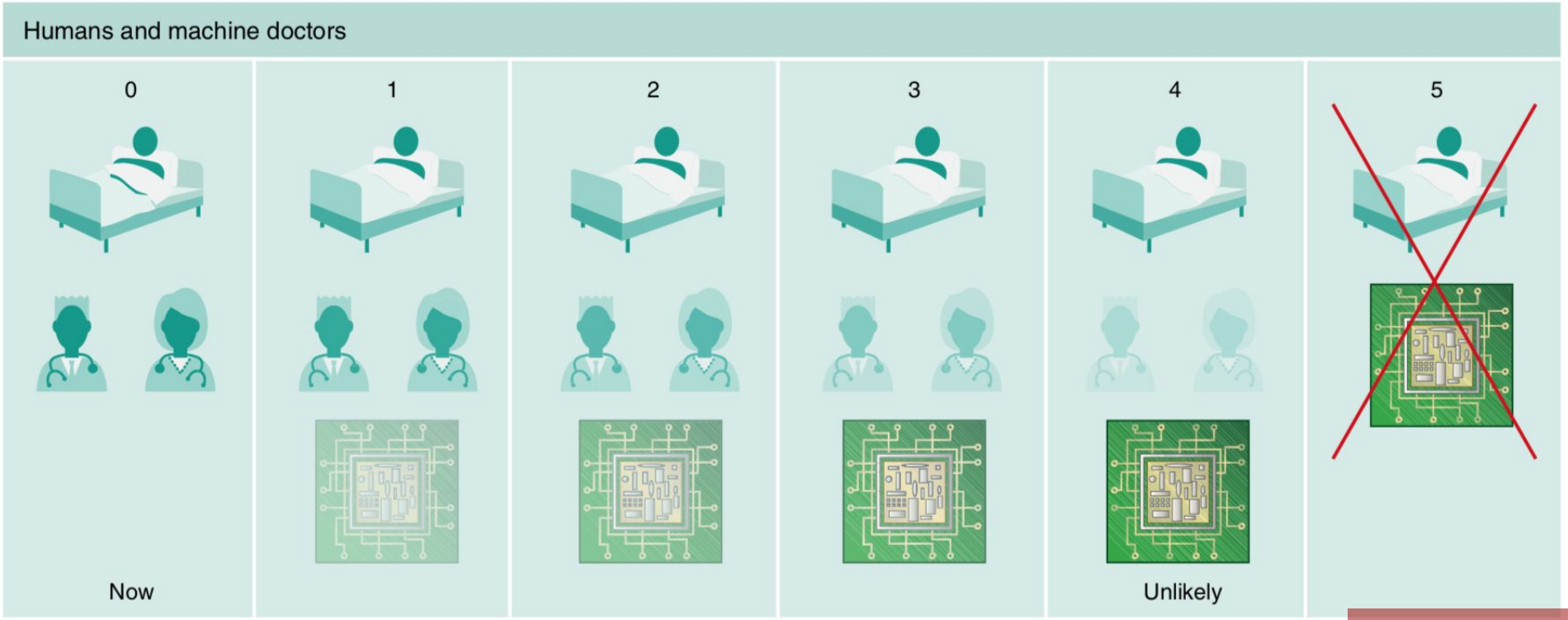








Human-in-the-loop

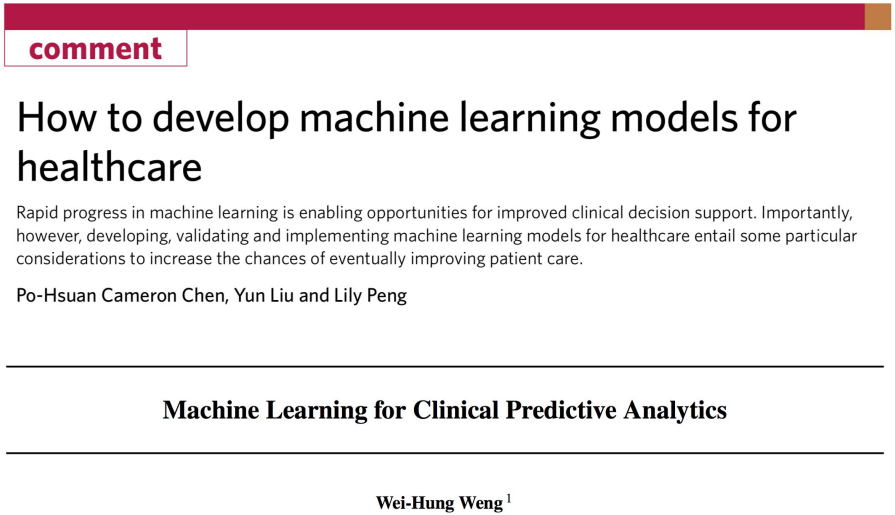


Topol. Nat Med. 2019

tl;dr

- How to develop machine learning models for healthcare [Chen 2019]
- Machine Learning for Clinical Predictive Analytics [Weng 2019]

- ***Problem definition***
- ***Data curation***
- ***ML model development***
- ***Validation***
- ***Assessment of clinical impact***
- ***(Deployment and monitoring)***



comment

How to develop machine learning models for healthcare

Rapid progress in machine learning is enabling opportunities for improved clinical decision support. Importantly, however, developing, validating and implementing machine learning models for healthcare entail some particular considerations to increase the chances of eventually improving patient care.

Po-Hsuan Cameron Chen, Yun Liu and Lily Peng

Machine Learning for Clinical Predictive Analytics

Wei-Hung Weng [1]

- ckbjimmy@mit.edu / Wei-Hung Weng (LinkedIn)