How to Conduct a Machine Learning Project for Clinical Medicine

Wei-Hung Weng, MD, MMSc (MIT CSAIL)
Joe Byers (BIDMC)

Khesar Gyalpo University of Medical Sciences of Bhutan
Oct 27-29, 2019

Massachusetts Institute of Technology

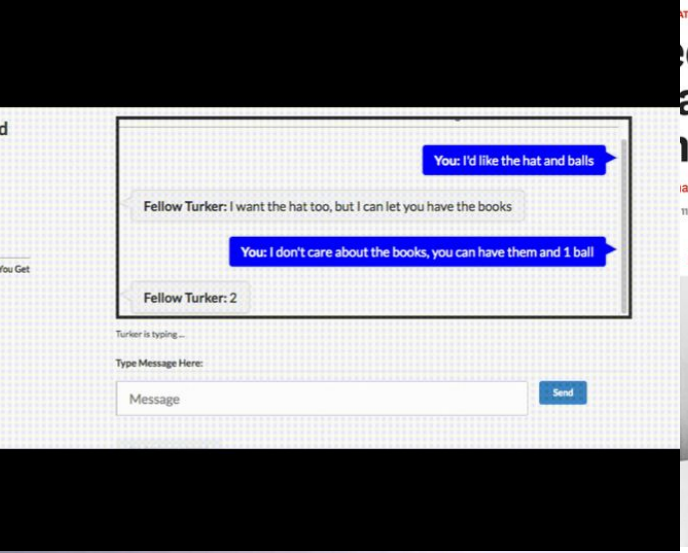
CSAIL

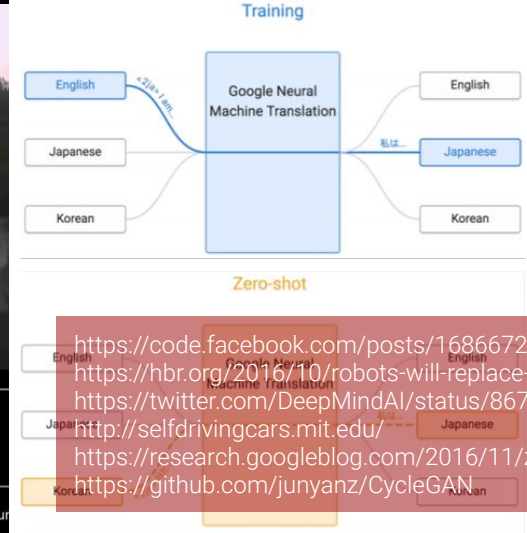
Outline

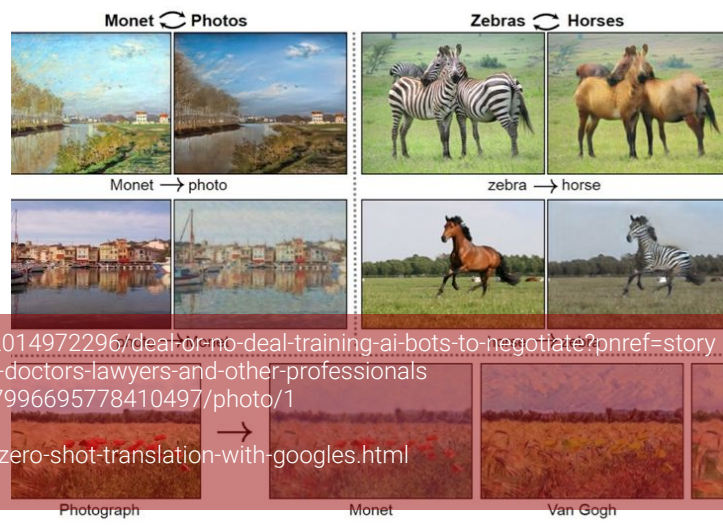
- Hour 1
  - Why ML in clinical medicine?
  - Why reproducibility? How to make this happen?
    - Tools
    - Data and community (PhysioNet / eICU / MIMIC)
  - Caveats of ML for healthcare problems
- Hour 2
  - Some examples of clinical problems → ML tasks
  - Hands-on coding exercise / demo
- Quick survey
  - Programming / python?
  - ML?

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]



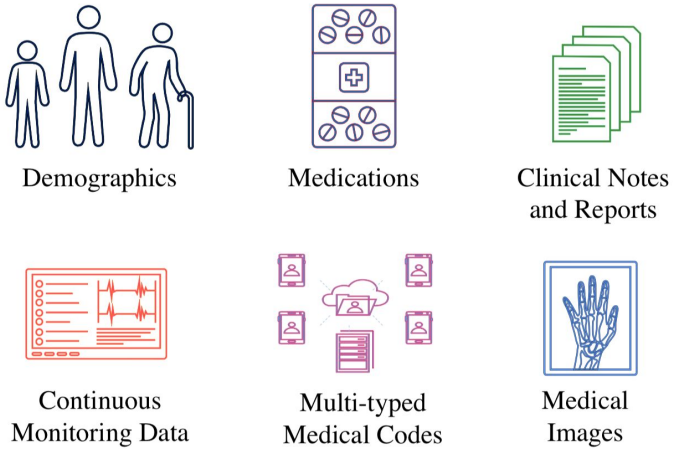


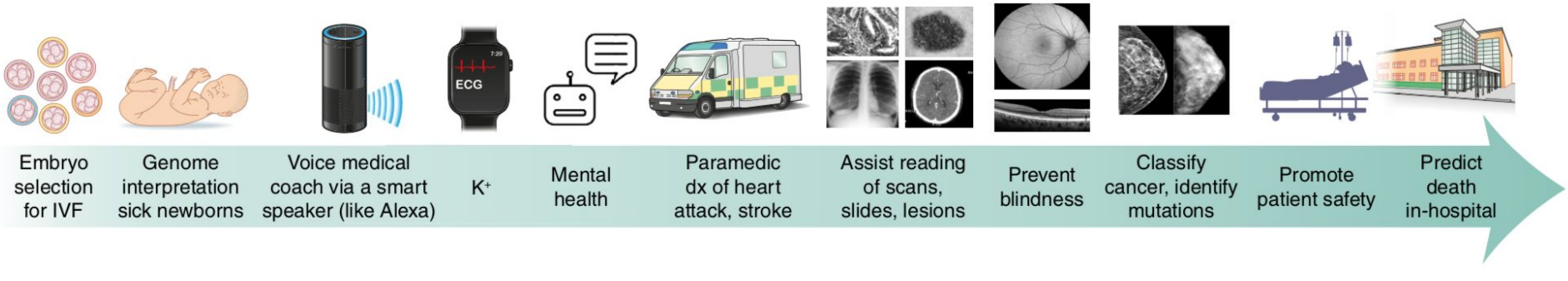


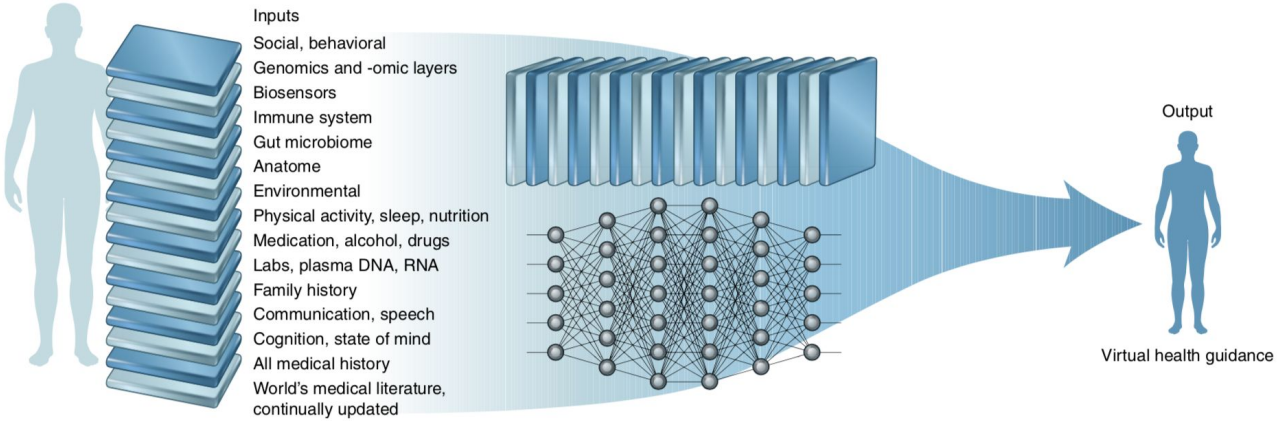
https://code.facebook.com/posts/1686672014972296/deal-or-no-deal-training-ai-bots-to-negotiate/?pnref=story
https://hbr.org/2016/10/robots-will-replace-doctors-lawyers-and-other-professionals
https://twitter.com/DeepMindAI/status/867996695778410497/photo/1
http://selfdrivingcars.mit.edu/
https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html
https://github.com/junyanz/CycleGAN

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

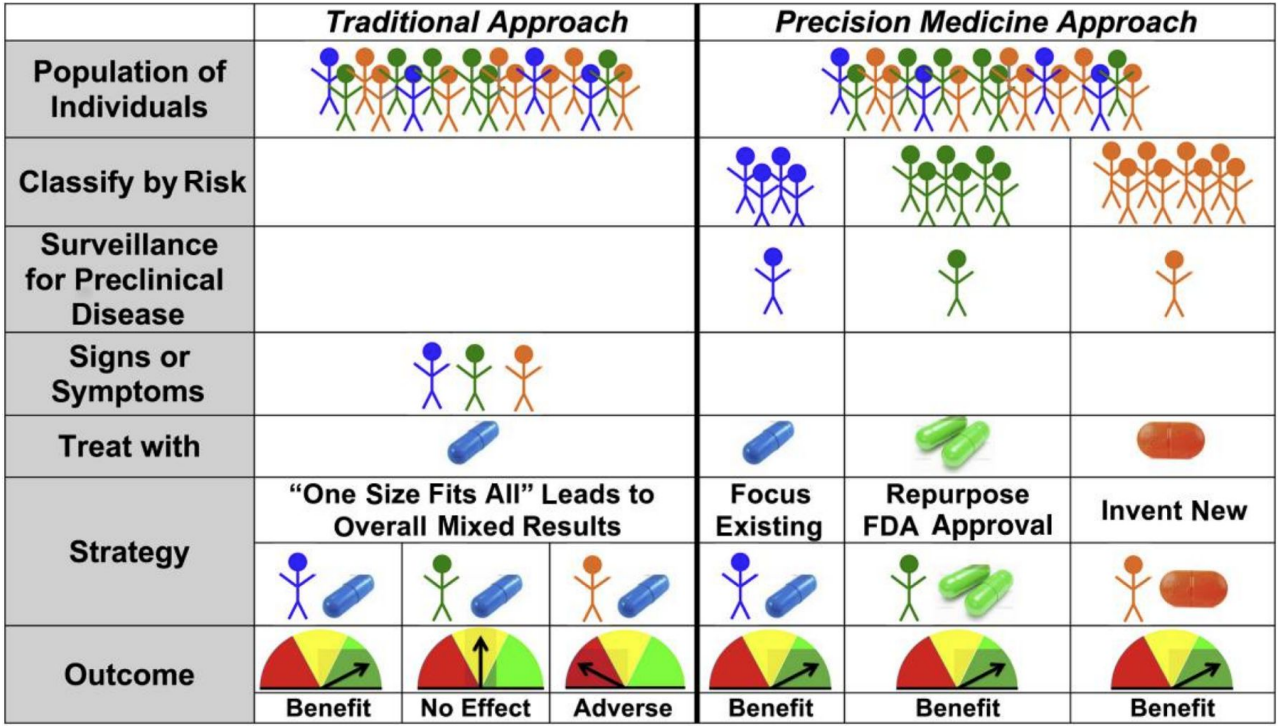






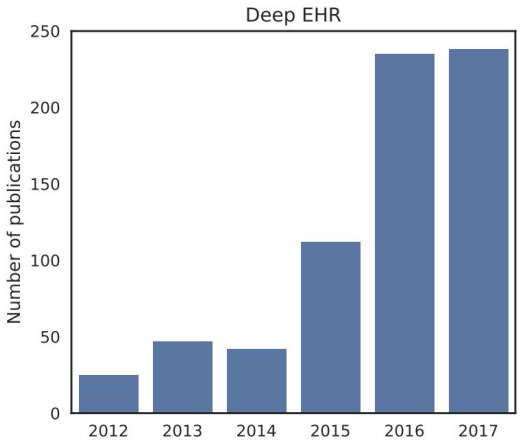
Topol. Nat Med. 2019

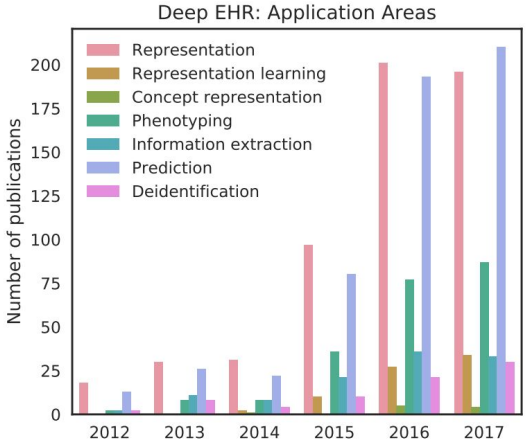
Precision Medicine

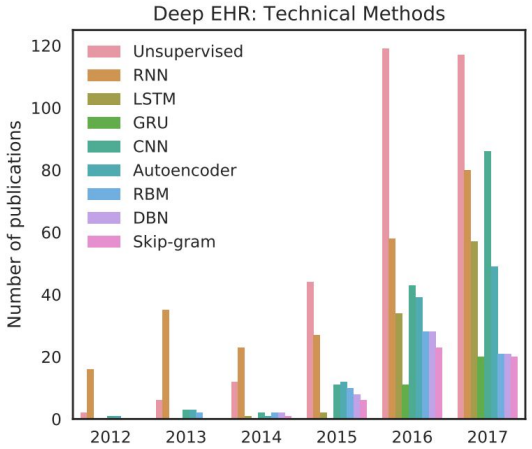


Cholerton. Am J Pathol. 2016

Deep! (2017/06)

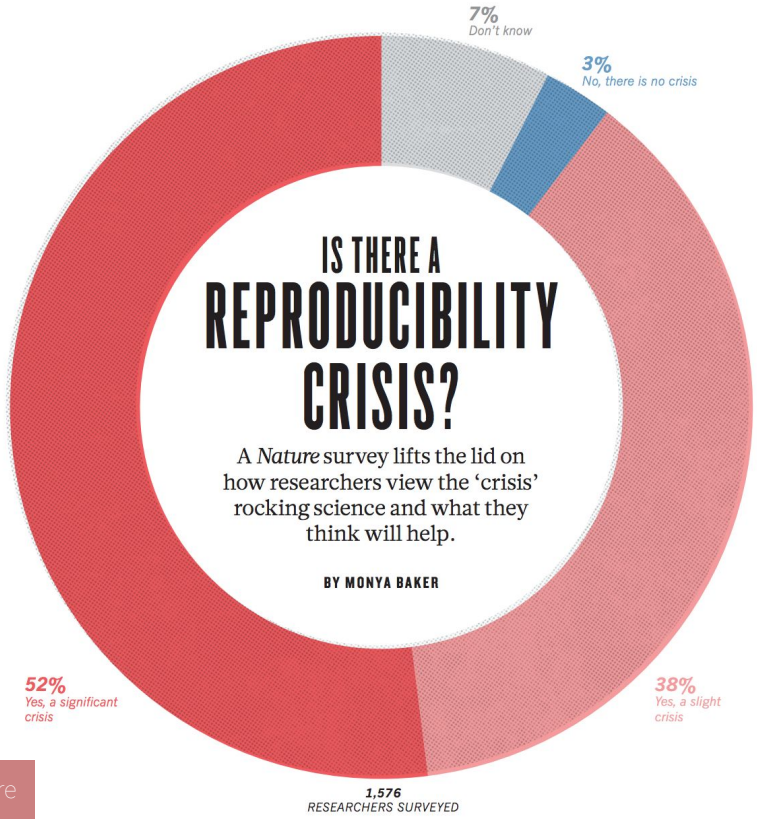

Deep EHR
Number of publications
250
200
150
100
50
0
2012
2013
2014
2015
2016
2017

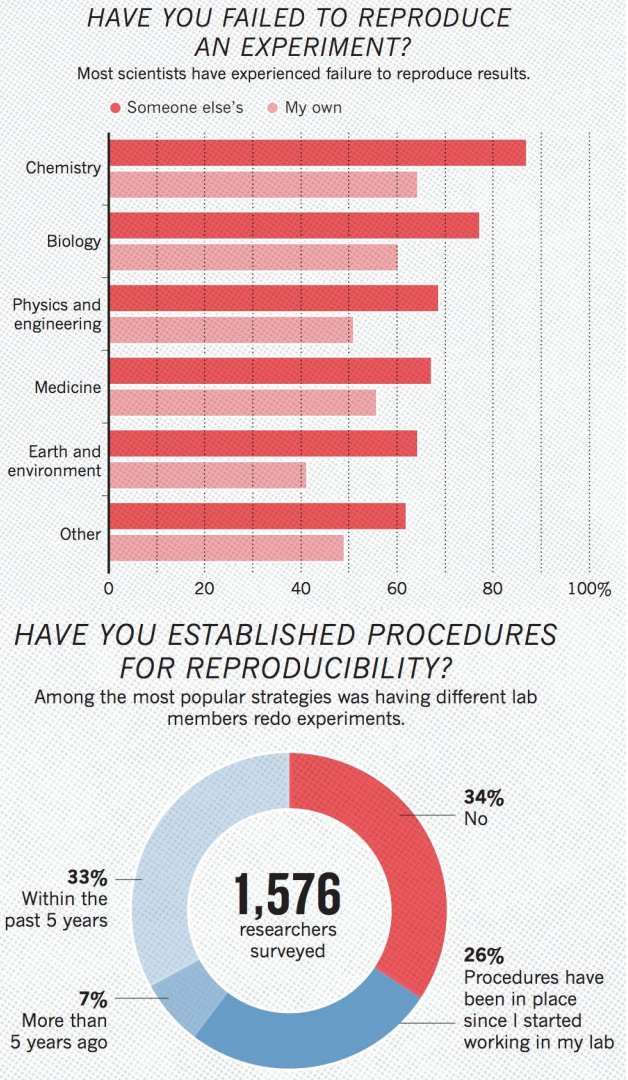

Deep EHR: Application Areas
Number of publications
200
175
150
125
100
75
50
25
0
Representation
Representation learning
Concept representation
Phenotyping
Information extraction
Prediction
Deidentification
2012
2013
2014
2015
2016
2017


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

Reproducibility

Reproducibility





Baker, 2015 Nature

ICLR 2018 Reproducibility Challenge

Background:

One of the challenges in machine learning research is to ensure that published results are reliable and reproducible. In support of this, the goal of this challenge is to investigate reproducibility of empirical results submitted to the 2018 International Conference on Learning Representations.

We are choosing ICLR for this challenge because the timing is right for course-based participants (see below), and because papers submitted to the conference are automatically made available publicly on Open Review.

The Challenge is inspired by discussions at the ICML 2017 Workshop on Reproducibility in Machine Learning.

Task Description

You should select a paper from the 2018 ICLR submissions, and aim to replicate the experiments described in the paper. The goal is to assess if the experiments are reproducible, and to determine if the conclusions of the paper are supported by your findings. Your results can be either positive (i.e. confirm reproducibility), or negative (i.e. explain what you were unable to reproduce, and potentially explain why).

Essentially, think of your role as an inspector verifying the validity of the experimental results and conclusions of the paper. In some instances, your role will also extend to helping the authors improve the quality of their work and paper.

https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html

Reproducibility

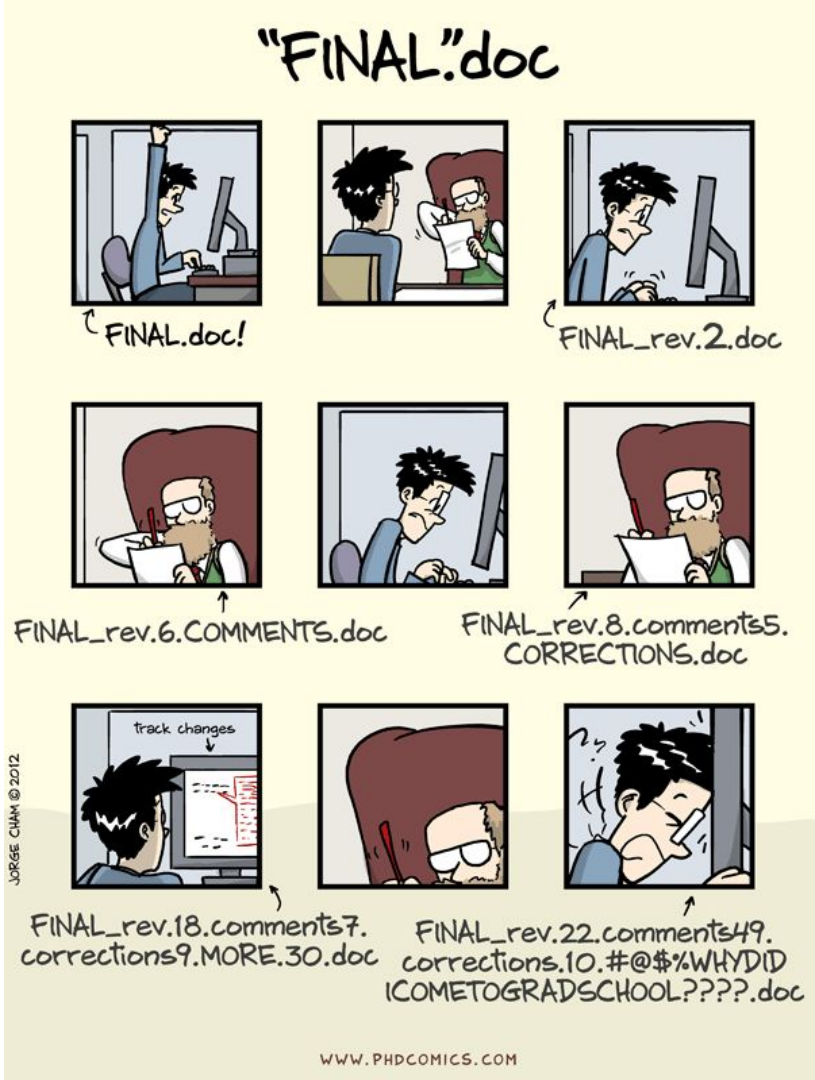
- Data + code (+ documentation!)
- For...
  - Validation
  - Learning
  - Collaboration

Tools

- Version control
  - Git
- Executable notebooks
  - Jupyter notebook (via anaconda)
  - Google colab
- Code publishing platform
  - GitHub
  - Gitlab

Version Control

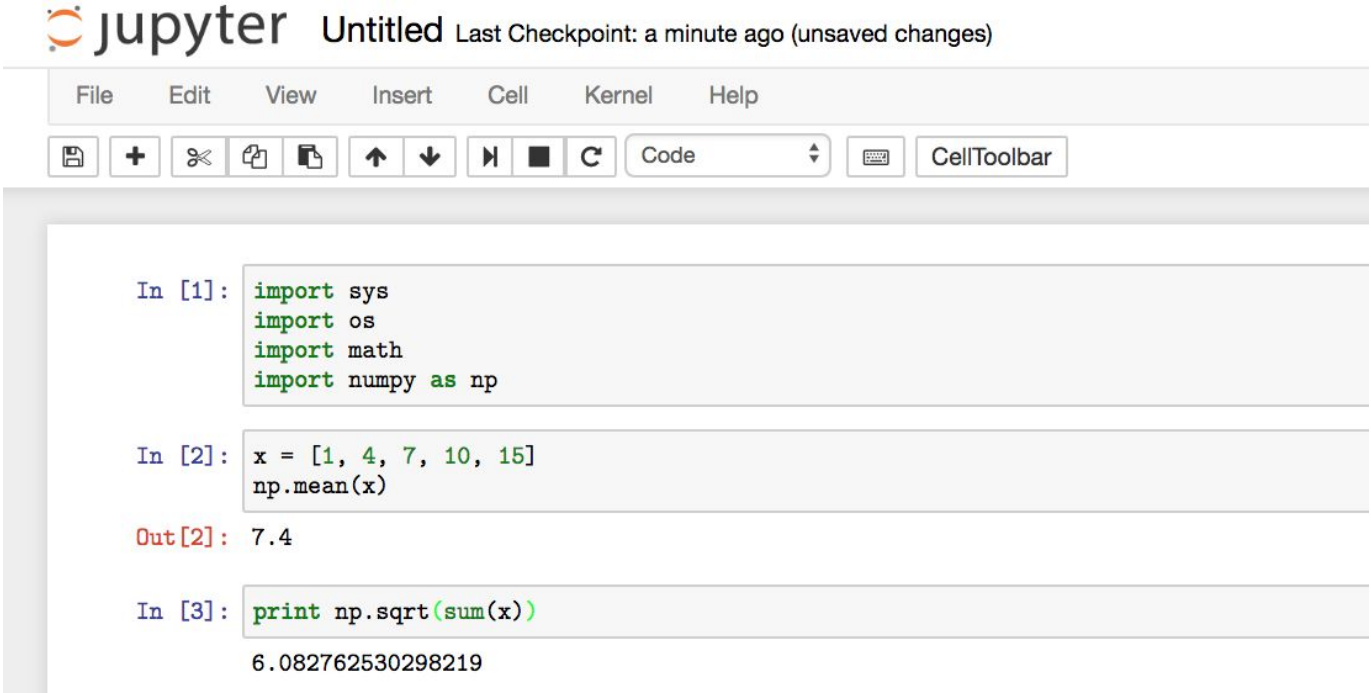
- `git add .`
- `git status`
- `git commit -m ‘first commit’`
- `git push`
- `git reset ‘.DS_Store’`
- `...`



http://phdcomics.com/comics/archive.php?comicid=1531

Jupyter Notebook

Execute (`Shift + Enter`) code cells and get your output underneath the cells



CoLab (CoLaboratory)

http://g.co/colab

Write code just as you would on a Jupyter Notebook

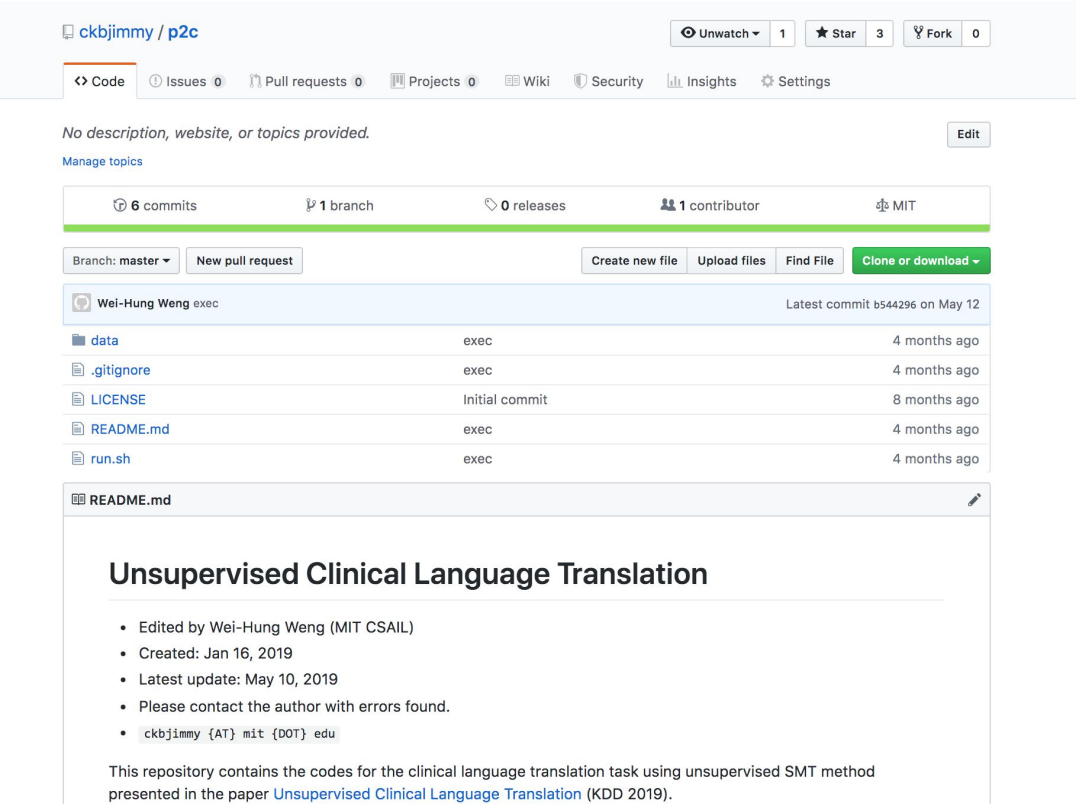
Use one of google's virtual machines to carry out your tasks

Free

Can write shell commands preceded with a `'!'`

- `!pip install gensim`
- `!ls`

Code Publishing (scary but worth it)



ckbjimmy / p2c

Unwatch 1 | Star 3 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Security | Insights | Settings

No description, website, or topics provided.

Edit

Manage topics

6 commits | 1 branch | 0 releases | 1 contributor | MIT

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

Wei-Hung Weng exec — Latest commit b544296 on May 12

data	exec	4 months ago
.gitignore	exec	4 months ago
LICENSE	Initial commit	8 months ago
README.md	exec	4 months ago
run.sh	exec	4 months ago

README.md

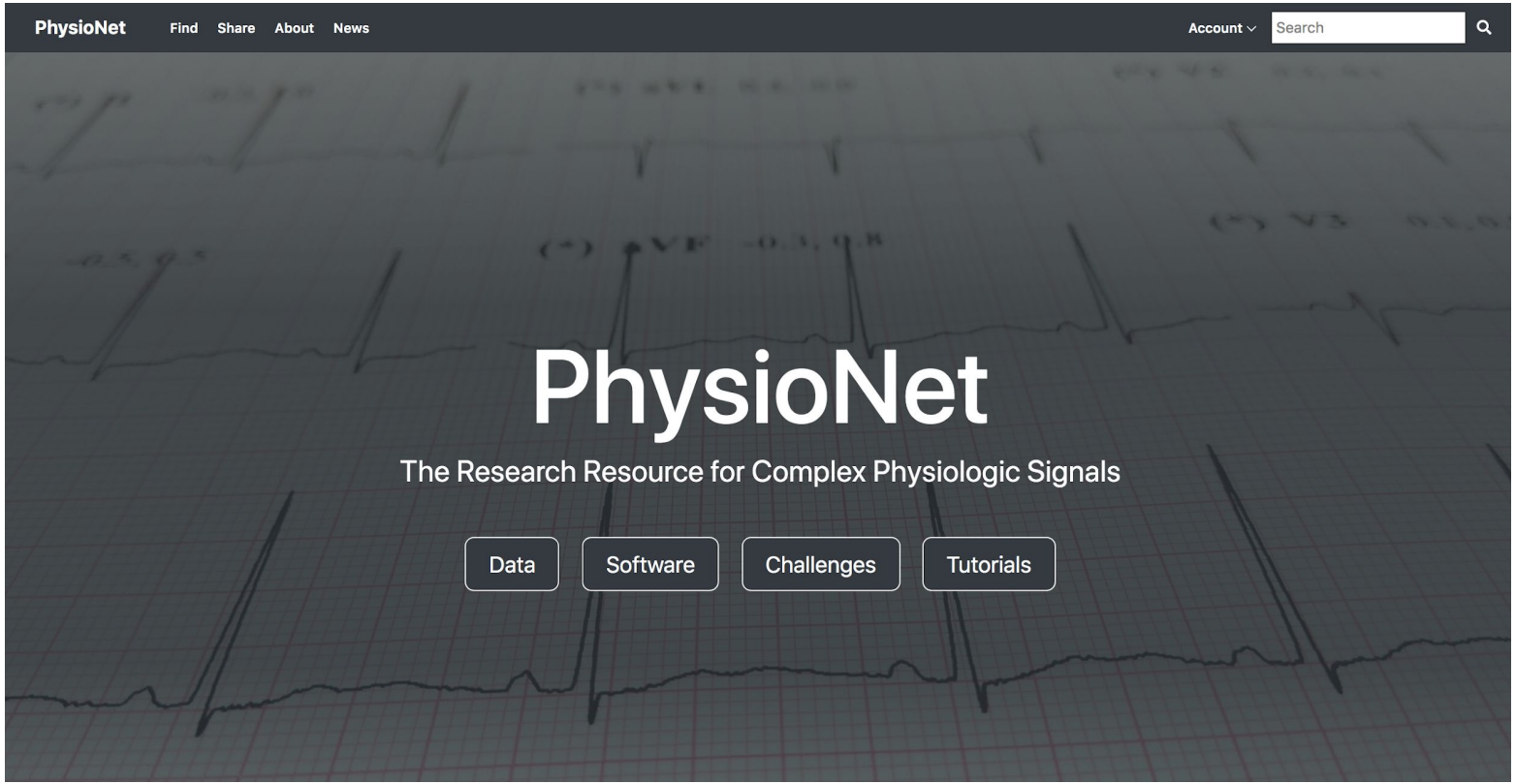
Unsupervised Clinical Language Translation

- Edited by Wei-Hung Weng (MIT CSAIL)
- Created: Jan 16, 2019
- Latest update: May 10, 2019
- Please contact the author with errors found.
- `ckbjimmy {AT} mit {DOT} edu`

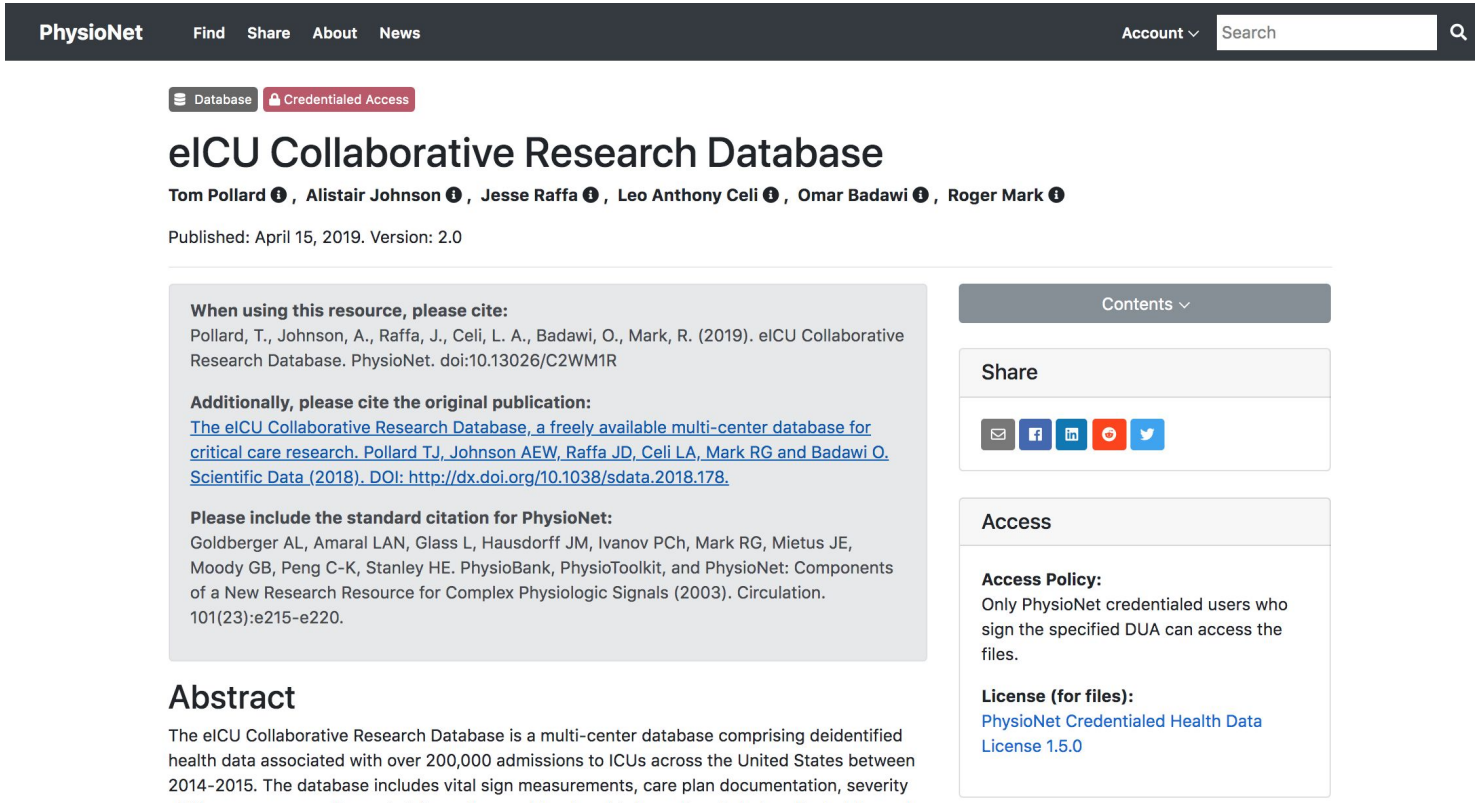
This repository contains the codes for the clinical language translation task using unsupervised SMT method presented in the paper Unsupervised Clinical Language Translation (KDD 2019).

PhysioNet / eICU / MIMIC-III

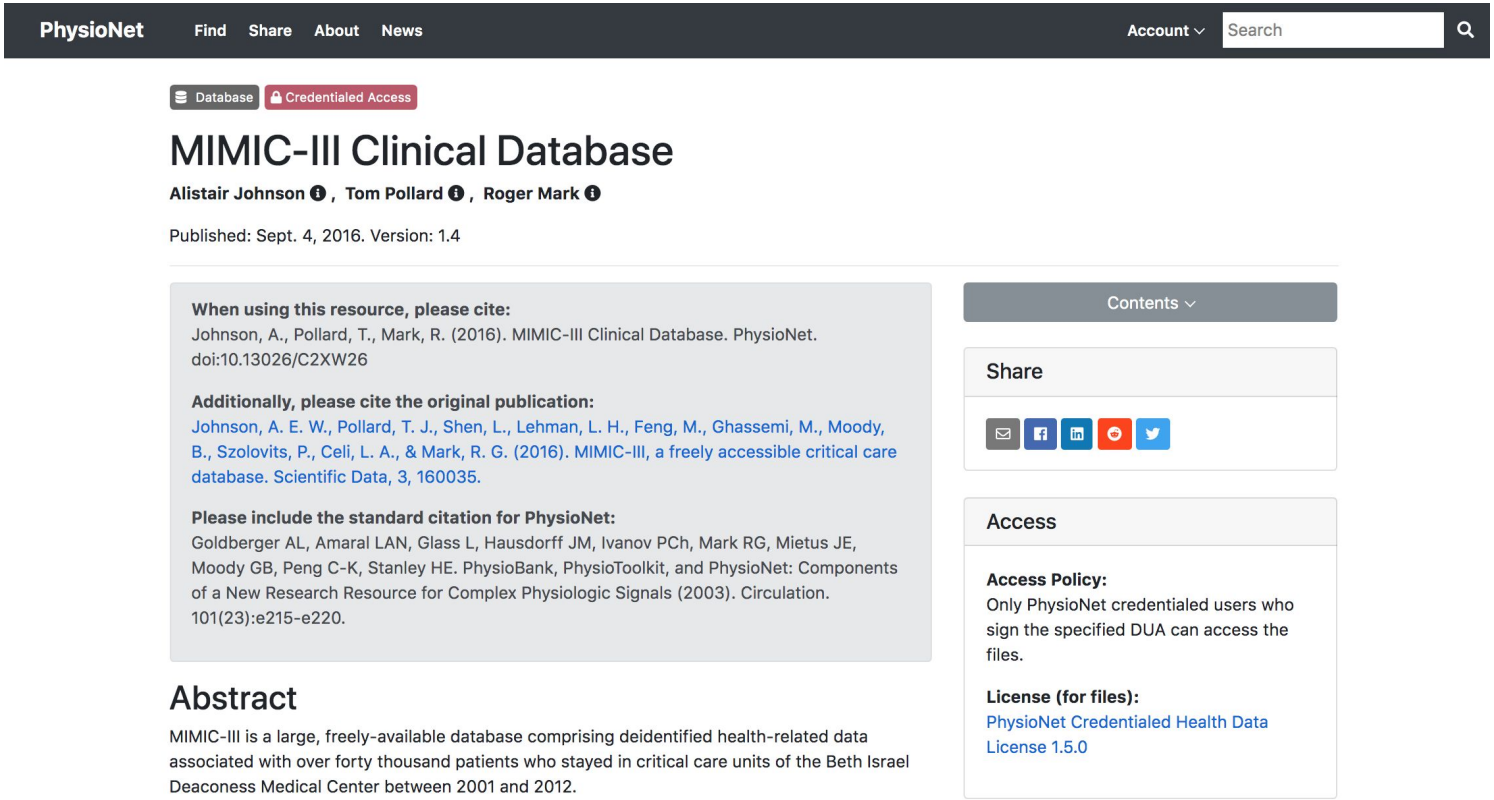
physionet.org



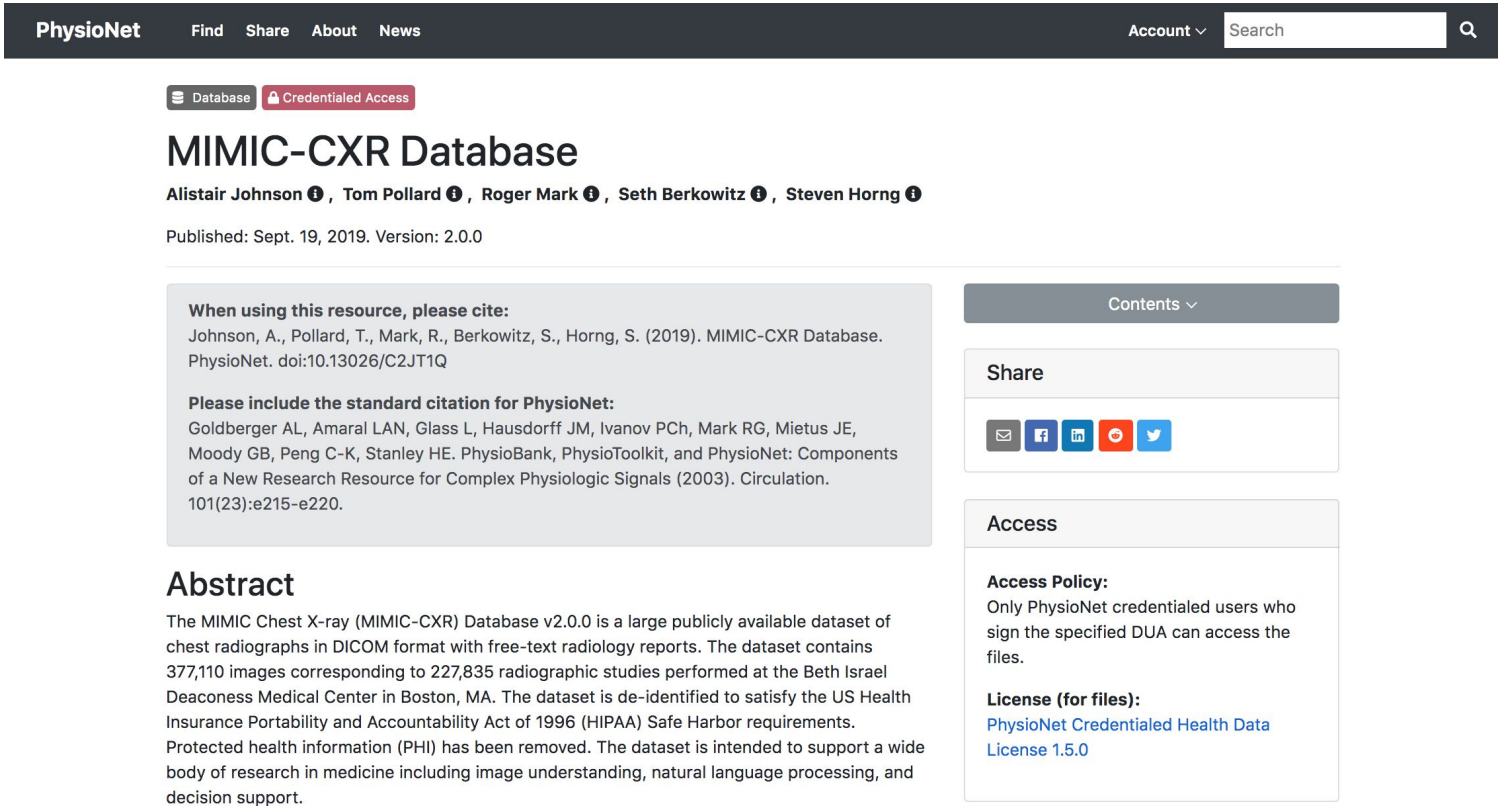
eICU


PhysioNet
Find
Share
About
News
Account
Search
Database
Credentialed Access
eICU Collaborative Research Database
Tom Pollard , Alistair Johnson , Jesse Raffa , Leo Anthony Celi , Omar Badawi , Roger Mark
Published: April 15, 2019. Version: 2.0
When using this resource, please cite:
Pollard, T., Johnson, A., Raffa, J., Celi, L. A., Badawi, O., Mark, R. (2019). eICU Collaborative Research Database. PhysioNet. doi:10.13026/C2WM1R
Additionally, please cite the original publication:
The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data (2018). DOI: http://dx.doi.org/10.1038/sdata.2018.178.
Please include the standard citation for PhysioNet:
Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.
Contents
Share
Access
Access Policy:
Only PhysioNet credentialed users who sign the specified DUA can access the files.
License (for files):
PhysioNet Credentialed Health Data License 1.5.0
Abstract
The eICU Collaborative Research Database is a multi-center database comprising deidentified health data associated with over 200,000 admissions to ICUs across the United States between 2014-2015. The database includes vital sign measurements, care plan documentation, severity

MIMIC

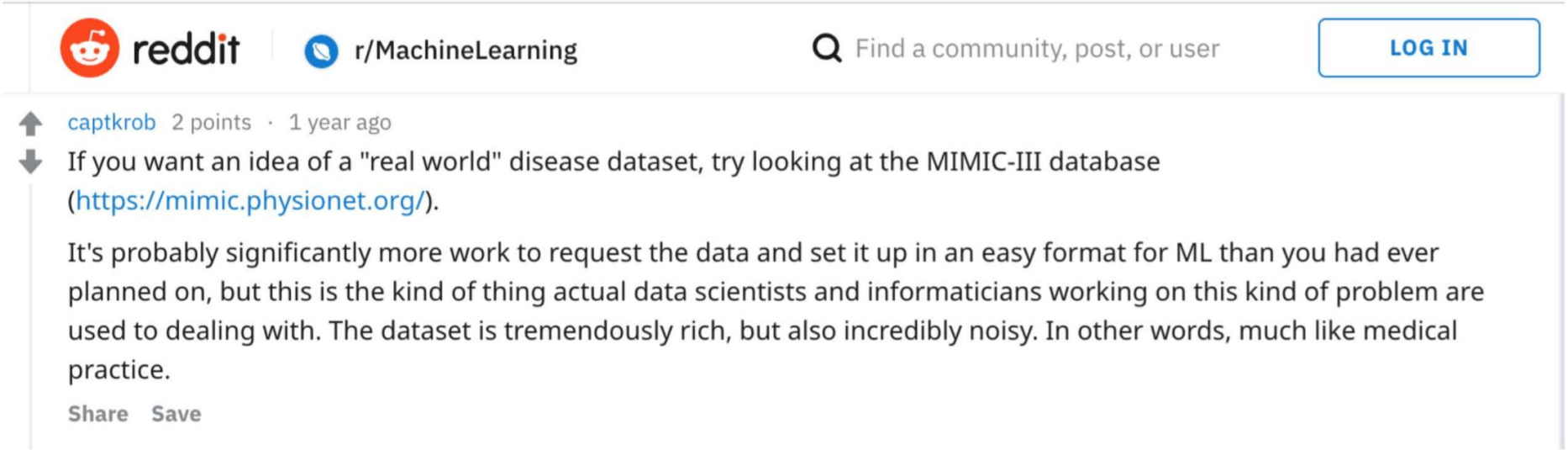

PhysioNet
Find
Share
About
News
Account
Search
Database
Credentialed Access
MIMIC-III Clinical Database
Alistair Johnson , Tom Pollard , Roger Mark
Published: Sept. 4, 2016. Version: 1.4
When using this resource, please cite:
Johnson, A., Pollard, T., Mark, R. (2016). MIMIC-III Clinical Database. PhysioNet. doi:10.13026/C2XW26
Additionally, please cite the original publication:
Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.
Please include the standard citation for PhysioNet:
Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.
Contents
Share
Access
Access Policy:
Only PhysioNet credentialed users who sign the specified DUA can access the files.
License (for files):
PhysioNet Credentialed Health Data License 1.5.0
Abstract
MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

MIMIC-CXR


PhysioNet
Find
Share
About
News
Account
Search
Database
Credentialed Access
MIMIC-CXR Database
Alistair Johnson , Tom Pollard , Roger Mark , Seth Berkowitz , Steven Horng
Published: Sept. 19, 2019. Version: 2.0.0
When using this resource, please cite:
Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S. (2019). MIMIC-CXR Database. PhysioNet. doi:10.13026/C2JT1Q
Please include the standard citation for PhysioNet:
Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.
Contents
Share
Access
Access Policy:
Only PhysioNet credentialed users who sign the specified DUA can access the files.
License (for files):
PhysioNet Credentialed Health Data License 1.5.0
Abstract
The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. The dataset is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. Protected health information (PHI) has been removed. The dataset is intended to support a wide body of research in medicine including image understanding, natural language processing, and decision support.

Warning

- Not in a clean, tidy spreadsheet
- Not generated for us to do research



Getting Access

- Sign data use agreement (DUA)
- Take online courses (citi)

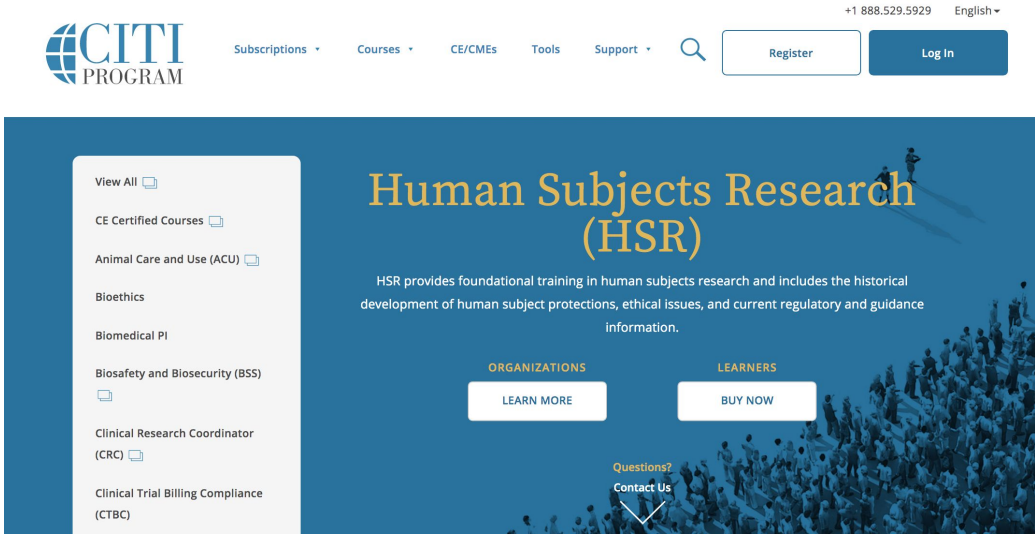
Sign Data Use Agreement - eICU Collaborative Research Database v2.0

Sign the following data use agreement to access the files in eICU Collaborative Research Database v2.0.

PhysioNet Credentialed Health Data Use Agreement 1.5.0

If I am granted access to the database:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet data shall continue after termination.



Data Sharing

- More people seeing the data → more knowledge
- Education
- Accelerating research
- You can also share the data through PhysioNet!

MIMIC-III

- Single center (>60K ICU stay, >40K patients)
- ICU data in details (almost everything incl. notes and waveforms)
- Limited out-of-ICU data (medication, DOD, ...)

If you use MIMIC data or code in your work, please cite the following publication:

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
Available from: http://www.nature.com/articles/sdata201635

eICU

- Multicenter (>200 hospitals, >200K ICU stay, 2 years)
- ICU data (no free text notes)
- Heterogeneous data (every hospital has their own input format)

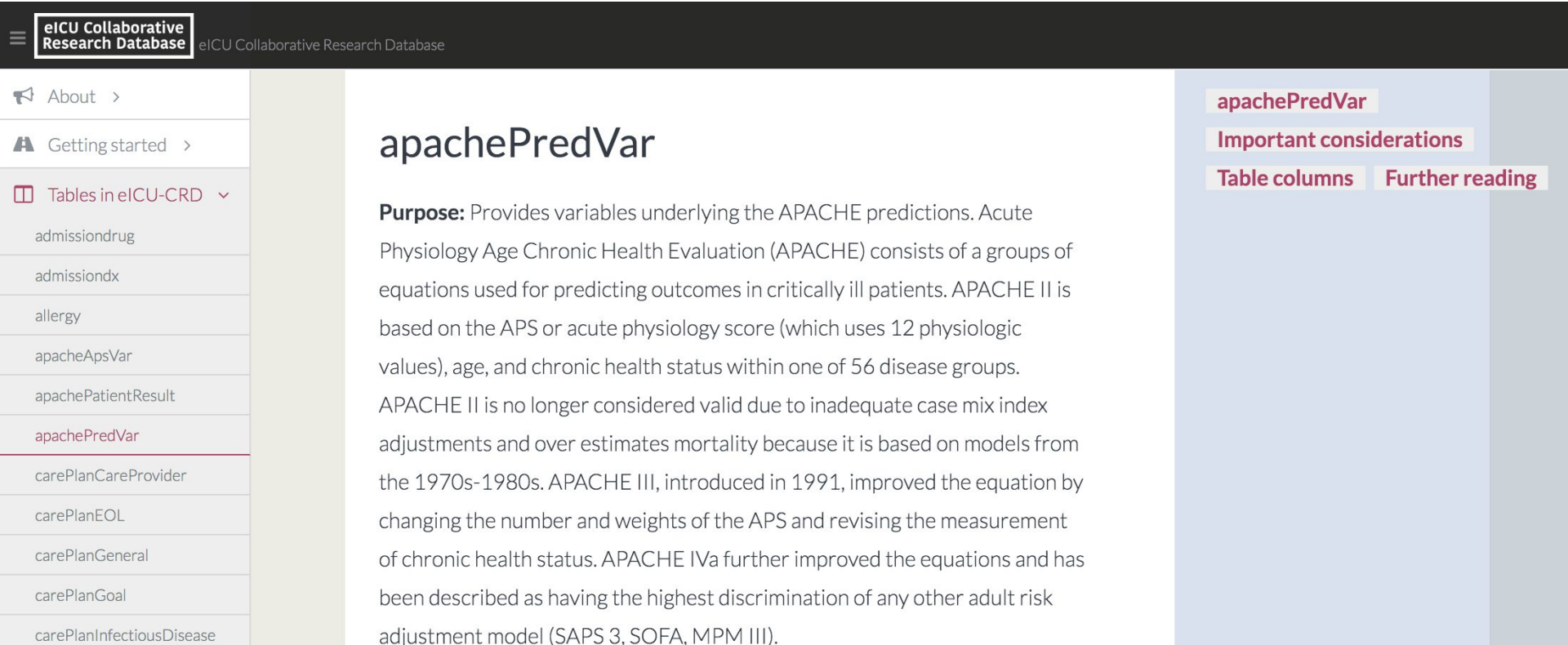
If you use the eICU Collaborative Research Database in your work, please cite the following publication:

The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data (2018). DOI: http://dx.doi.org/10.1038/sdata.2018.178. Available from: https://www.nature.com/articles/sdata2018178

eICU vs. MIMIC-III

- Strengths of eICU
  - Latest data (2014-2015)
  - Larger dataset, multicenter
  - Severity score for all patients (APACHE)
  - Treatment plan in `carePlanGeneral` table
  - Diagnosis available during stay
- Strengths of MIMIC
  - Documentation, codebase and community support
  - More literature
  - Consistent within hospital
  - With clinical notes

Navigating eICU [eicu-crd.mit.edu]



eICU Patient Tracking

- `patientunitstayid` → ICU stay
- `patienthealthsystemstayid` → hospital stay
- `uniquepid` → patient
- `hospitalid` → hospital

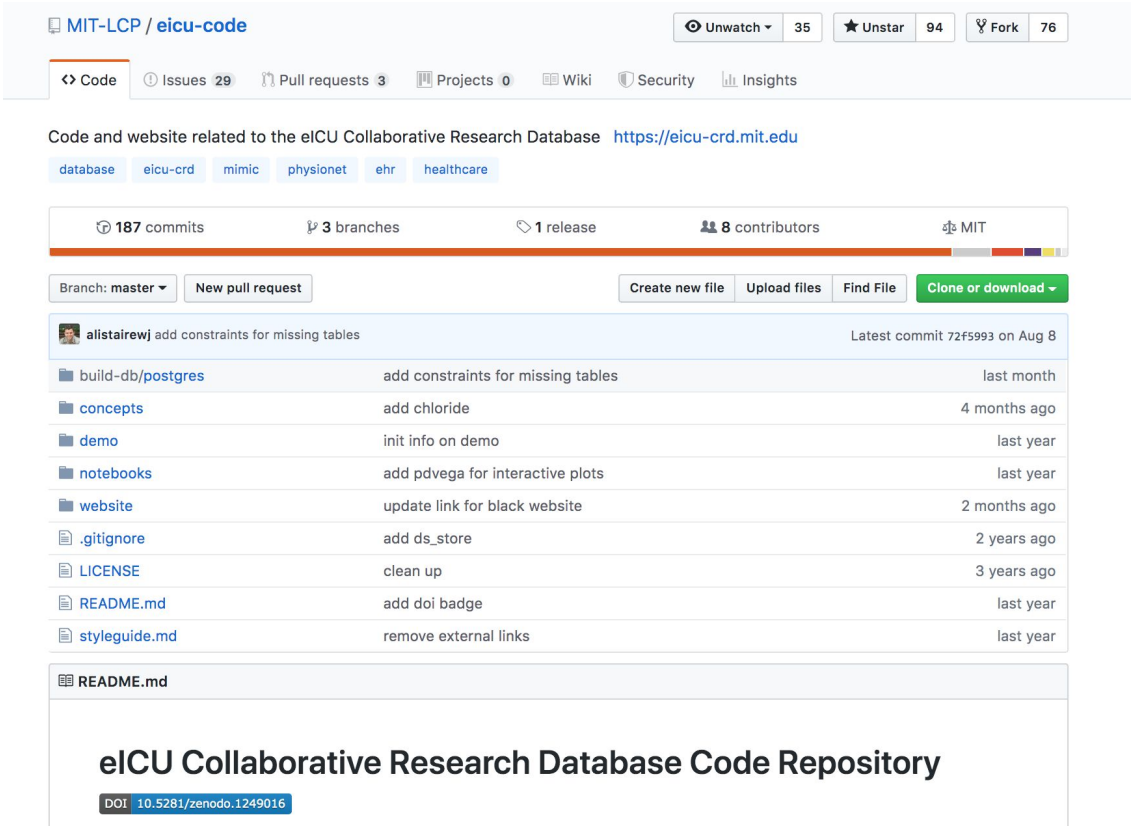
- `...offset`

Tables

- `apacheapsvar`
  - First day aggregated data for APACHE
- `apachepredvar`
- `diagnosis` → offset, string, ICD code
- `infusiondrug` → fluid, insulin, vasopressor, sedative
- `intakeoutput` → urine output
- `patient` → demographics
- `lab, medication, pasthistory, treatment` → offset, string
- `vitalperiodic, vitalaperiodic`

- `github.com/ckbjimmy/hst953_iomed`

Codebase [github.com/MIT-LCP/eicu-code]



MIT-LCP / eicu-code

Unwatch 35 | Unstar 94 | Fork 76

Code | Issues 29 | Pull requests 3 | Projects 0 | Wiki | Security | Insights

Code and website related to the eICU Collaborative Research Database https://eicu-crd.mit.edu

database | eicu-crd | mimic | physionet | ehr | healthcare

187 commits | 3 branches | 1 release | 8 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

alistairewj add constraints for missing tables — Latest commit 72f5993 on Aug 8

build-db/postgres	add constraints for missing tables	last month
concepts	add chloride	4 months ago
demo	init info on demo	last year
notebooks	add pdvega for interactive plots	last year
website	update link for black website	2 months ago
.gitignore	add ds_store	2 years ago
LICENSE	clean up	3 years ago
README.md	add doi badge	last year
styleguide.md	remove external links	last year

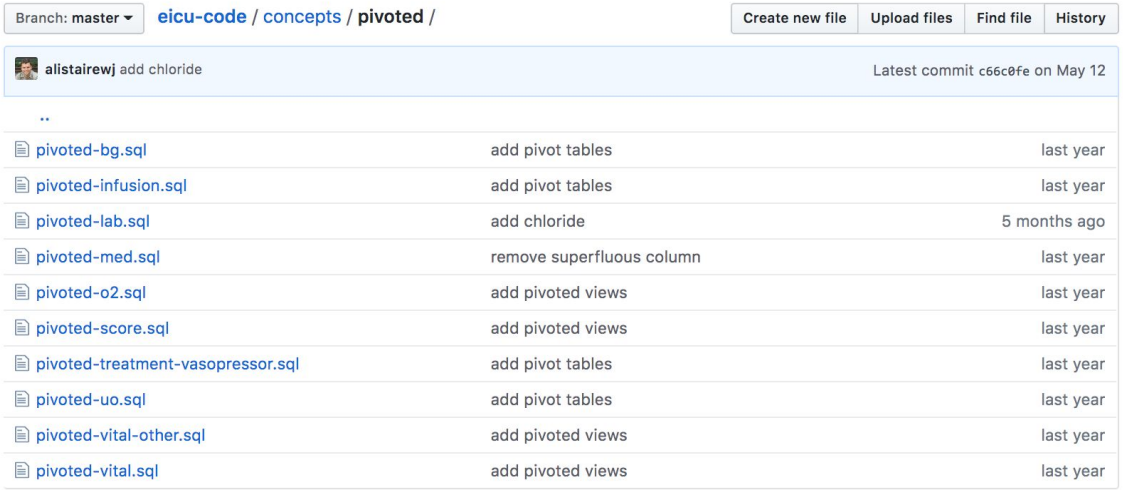
README.md

eICU Collaborative Research Database Code Repository

DOI 10.5281/zenodo.1249016

Concepts

- In BigQuery
- Derived tables
  - `/eicu-code/concepts/pivoted/`



Branch: master ▾ **eicu-code** / concepts / **pivoted** /

Create new file | Upload files | Find file | History

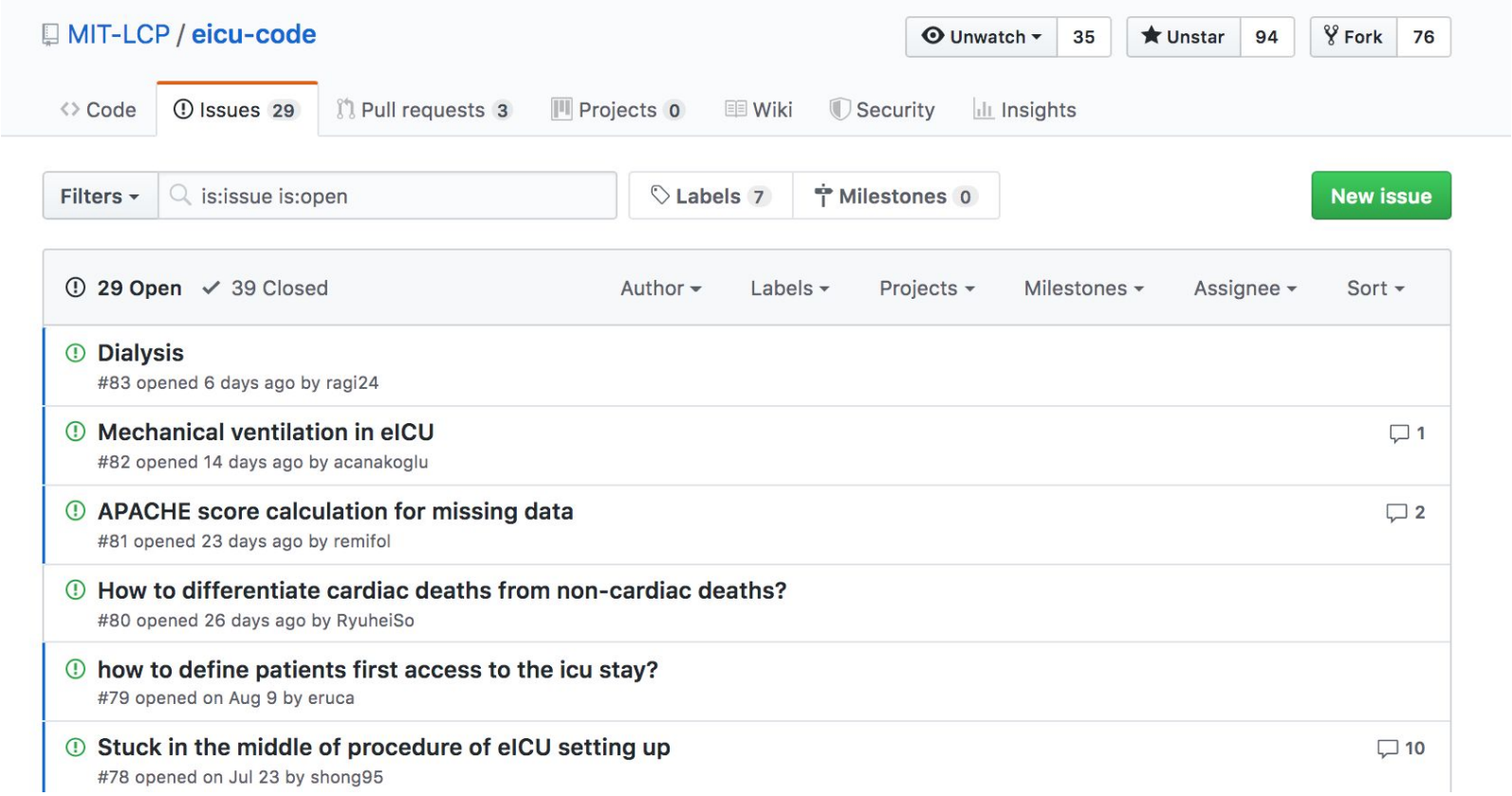
alistairewj add chloride — Latest commit c66c0fe on May 12

..

pivoted-bg.sql	add pivot tables	last year
pivoted-infusion.sql	add pivot tables	last year
pivoted-lab.sql	add chloride	5 months ago
pivoted-med.sql	remove superfluous column	last year
pivoted-o2.sql	add pivoted views	last year
pivoted-score.sql	add pivoted views	last year
pivoted-treatment-vasopressor.sql	add pivot tables	last year
pivoted-uo.sql	add pivot tables	last year
pivoted-vital-other.sql	add pivoted views	last year
pivoted-vital.sql	add pivoted views	last year

```
, max(case
      when drugname in
      (
          'EPI (mcg/min)'
        , 'Epinepherine (mcg/min)'
        , 'Epinephrine'
        , 'Epinephrine ()'
        , 'EPINEPHrine(Adrenalin)MAX 30 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
        , 'EPINEPHrine(Adrenalin)STD 4 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
        , 'EPINEPHrine(Adrenalin)STD 4 mg Sodium Chloride 0.9% 500 ml (mcg/min)'
        , 'EPINEPHrine(Adrenalin)STD 7 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
        , 'Epinephrine (mcg/hr)'
        , 'Epinephrine (mcg/kg/min)'
        , 'Epinephrine (mcg/min)'
        , 'Epinephrine (mg/hr)'
        , 'Epinephrine (mg/kg/min)'
        , 'Epinephrine (ml/hr)'
      ) then 1 else 0 end)
    as epinephrine
```

Are You the First Person to Ask → GitHub Issues

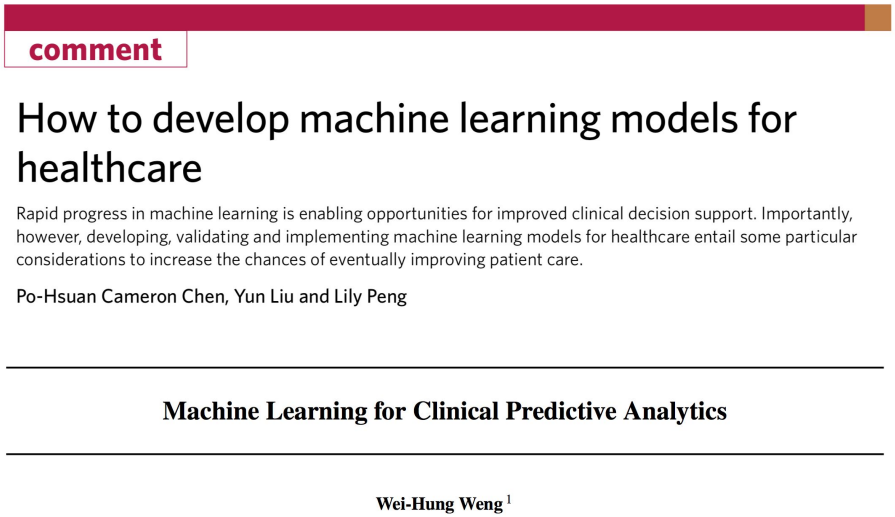


How?

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.

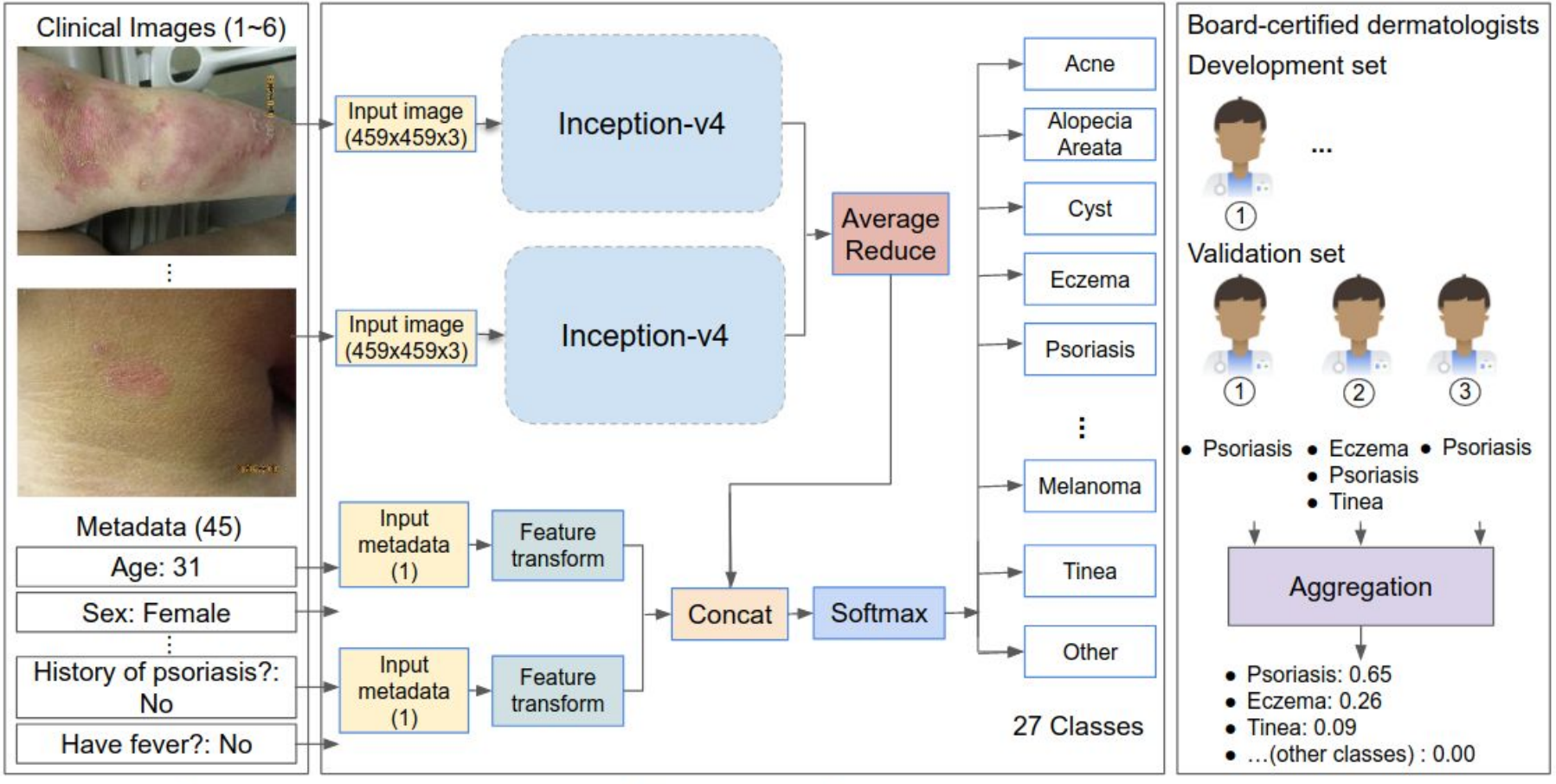
Po-Hsuan Cameron Chen, Yun Liu and Lily Peng

Machine Learning for Clinical Predictive Analytics

Wei-Hung Weng [1]

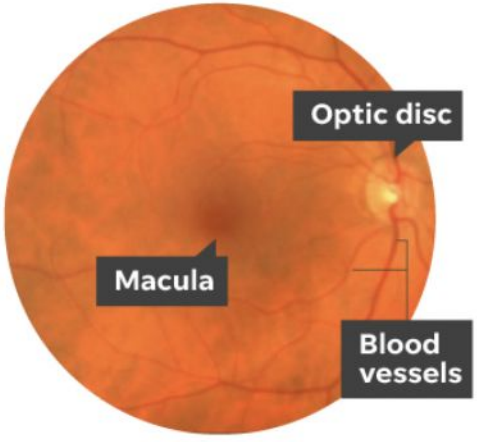
[1] Appropriate Problem Definition

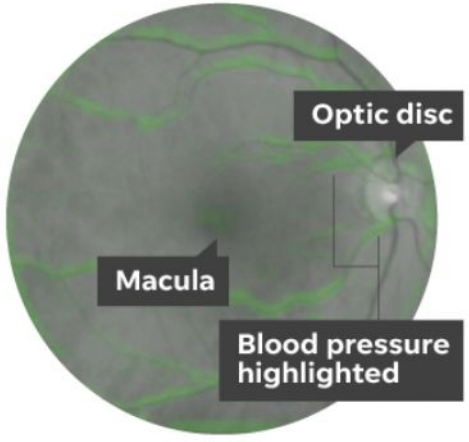
- Problem selection
  - Make a meaningful impact in patient care by ***providing actionable insights***
  - ML model should only leverage input data that are available at the time when the proposed clinical decision is being made
- Defining prediction task
  - ***Learning from humans***
    - Removing human-factor bottleneck (availability and fatigue of human raters)
    - ***E.g. fundus imaging → DR grading [Gulshan 2016], skin image → disease clf [Liu 2019]***
  - ***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



Liu. arXiv. 2019

Google AI can predict heart problems by taking pictures of your eye

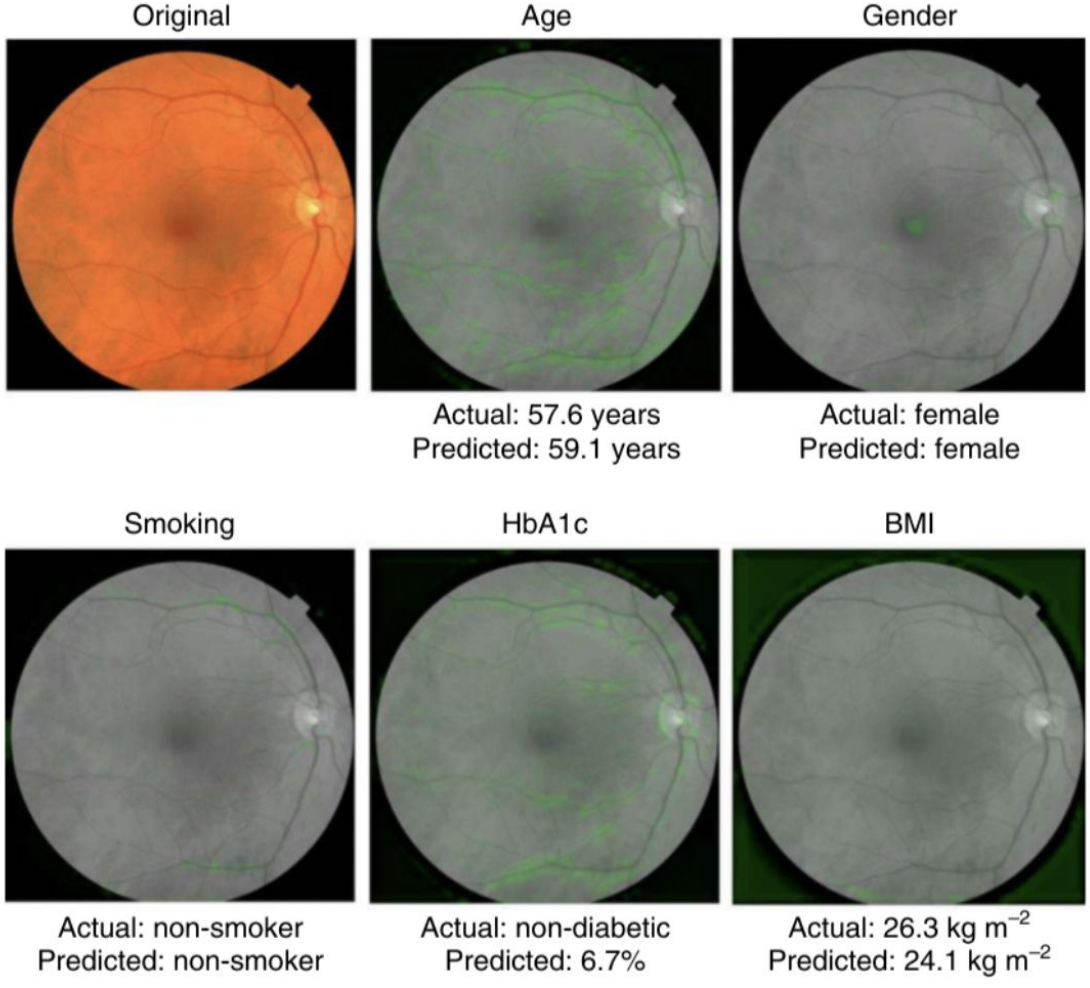




- Image of the back of the eye showing the macula (dark spot in the middle), optic disc (bright spot at the right), and blood vessels.
- Retinal image in gray, researchers can focus on blood vessels to determine the health risks associates with a patient's blood pressure.
- Images showed that each cardiovascular risk factor prediction uses a distinct pattern, such as blood vessels for blood pressure and optic disc for other predictions.

Source: Baig, Edward C. "Google Hopes AI Can Predict Heart Disease by Looking at Retinas." *USA Today*, Gannett Satellite Information Network, 19 Feb. 2018, www.usatoday.com/story/tech/2018/02/19/google-ai-can-predict-heart-disease-looking-pictures-retina/344547002/.

Poplin. Nat BME. 2018



Poplin. Nat BME. 2018

[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

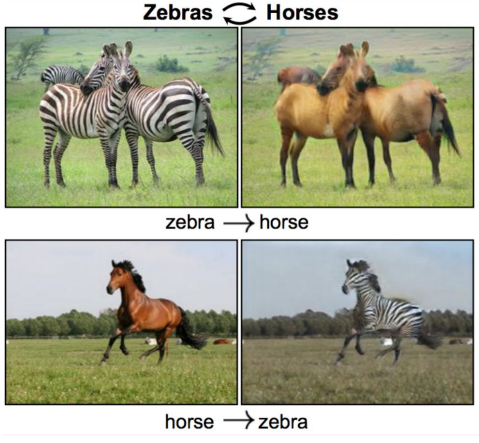
- ***Supervised learning***
  - Regression
  - Classification
    - Linear
    - Non-linear (e.g. deep learning, SVM, decision tree, ...)
  - Structured learning
- Unsupervised learning
  - Clustering
  - Dimensionality reduction
- Transfer learning
- Reinforcement learning
- ...

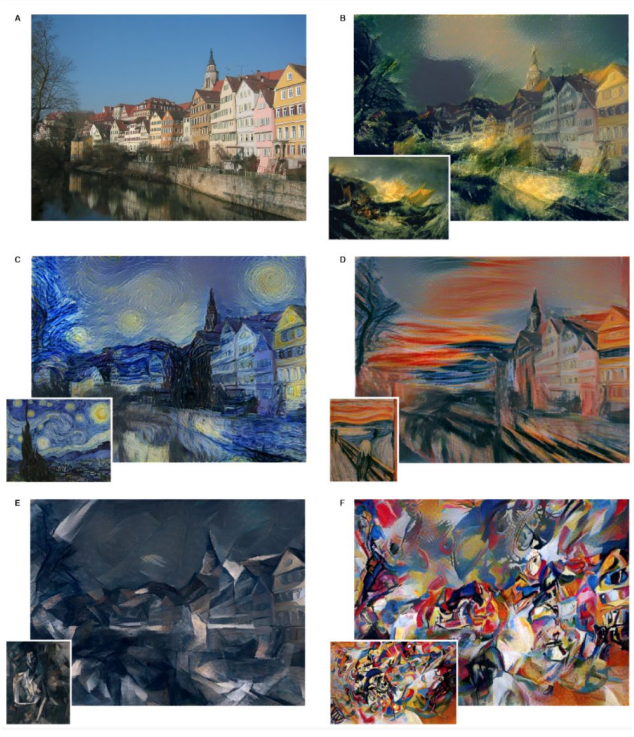
Data Modality

- Structured / Tabular data (e.g. claims data, vitals, labs, demographics, ...)
- Medical imaging
- Natural language processing
- Waveform
- ...
- Multimodality

Computer Vision / Medical Imaging

- Mainly DL-driven
- Main applications
  - Classification
  - Classification With Localization
  - Object Detection
  - Object Segmentation
  - Style Transfer
  - Colorization
  - Reconstruction
  - Super-Resolution
  - Synthesis
  - Others



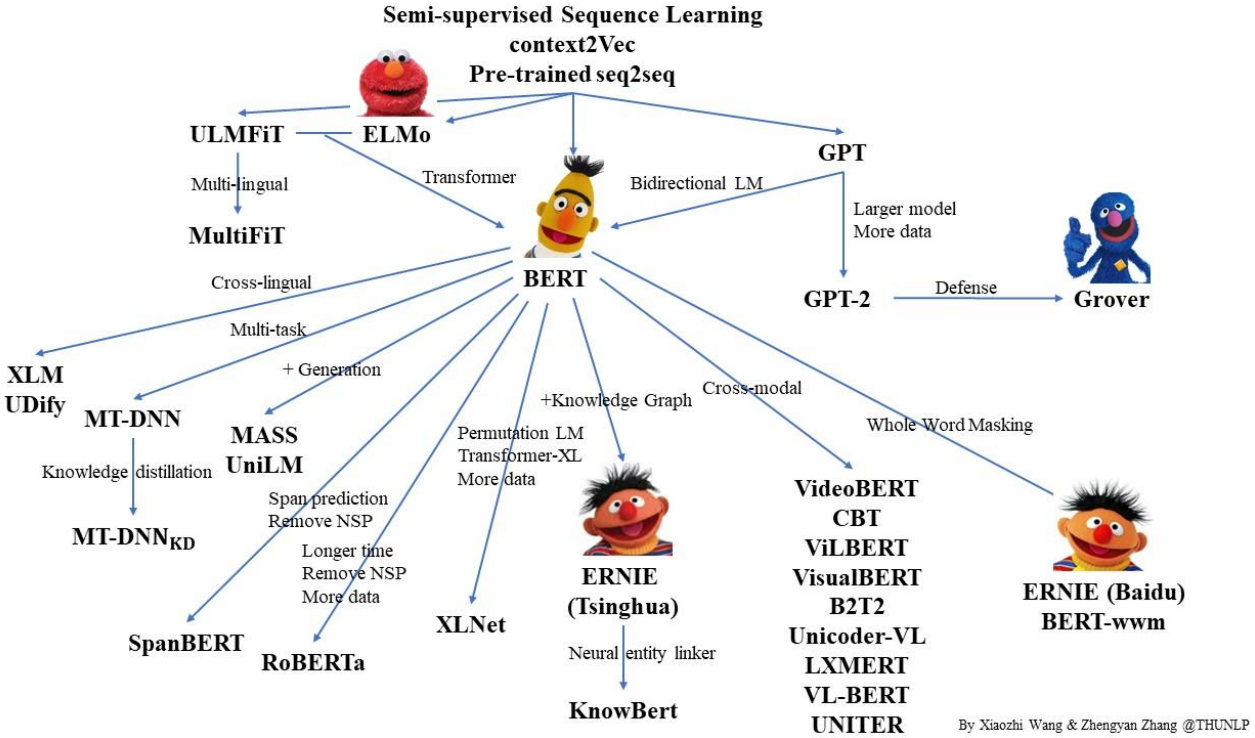


Zhu 2017. Gatys 2015.

Natural Language Processing

- Embedding, encoder-decoder, attention, pretrained models / transfer learning
- `ruder.io`
- `nlpprogress.com`
  - Automatic speech recognition / CCG / Common sense / Constituency parsing / Coreference resolution / Dependency parsing / Dialogue / Domain adaptation / Entity linking / Grammatical error correction / Information extraction / Language modeling / Lexical normalization / Machine translation / Missing elements / ***Multi-task learning*** / ***Multi-modal*** / Named entity recognition / Natural language inference / Part-of-speech tagging / Question answering / Relation prediction / Relationship extraction / Semantic textual similarity / Semantic parsing / Semantic role labeling / Sentiment analysis / Shallow syntax / Simplification / Stance detection / Summarization / Taxonomy learning / Temporal processing / ***Text classification*** / Word sense disambiguation

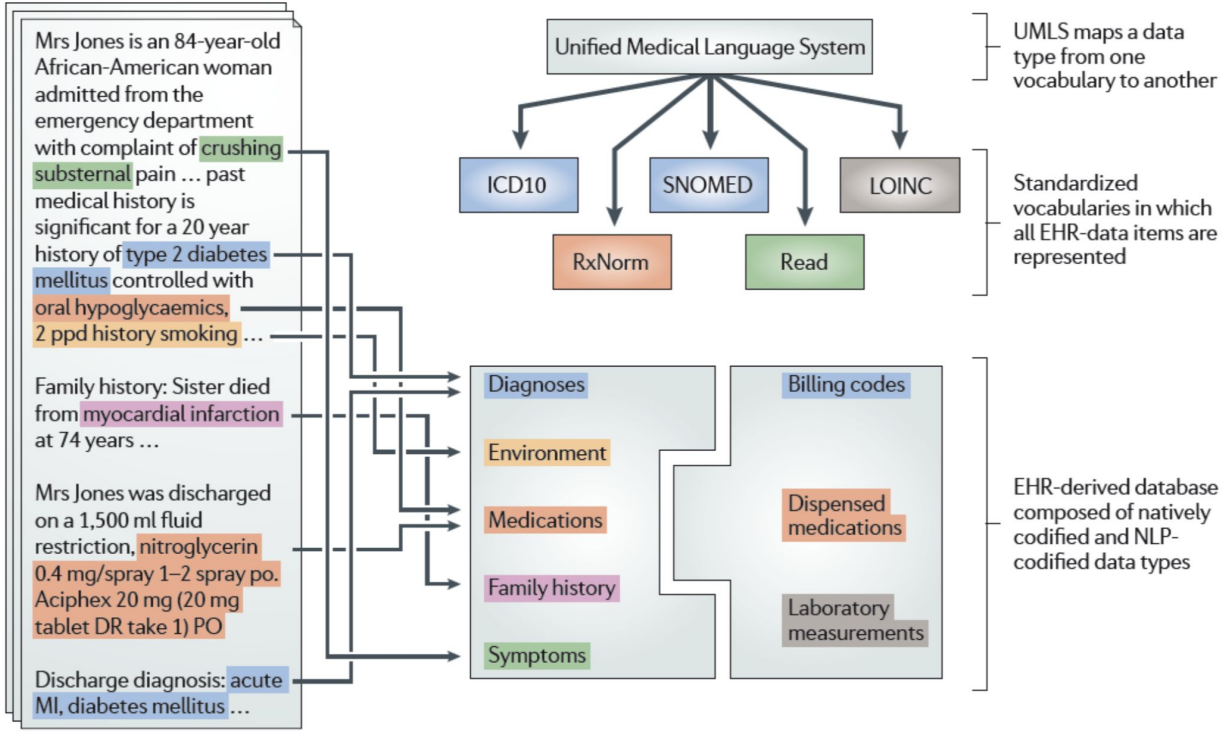
Pretrained LM



Github: thunlp/PLMpapers

Clinical NLP

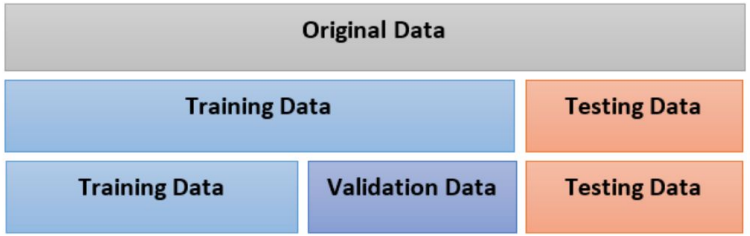
- Smaller corpus
- Domain-specific
  - Concepts
  - Typos
  - Misspellings
  - Grammars
  - ...
- Pretrained
  - Alsentzer 2019
  - ClinicalBERT

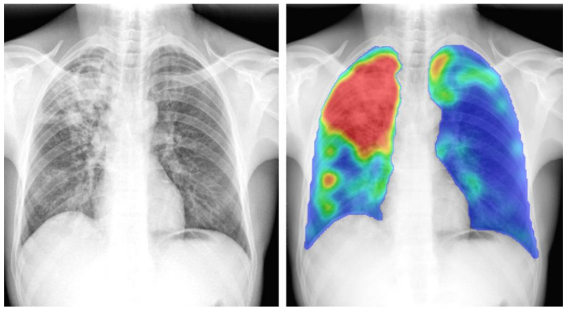


Kohane. Nat Rev Genetics. 2011

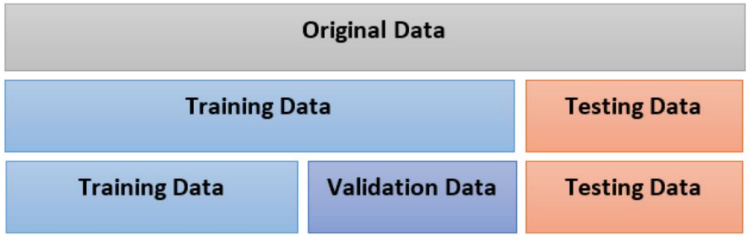
[1] Appropriate Problem

- ***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



- ***Choose a correct cohort***
- ***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

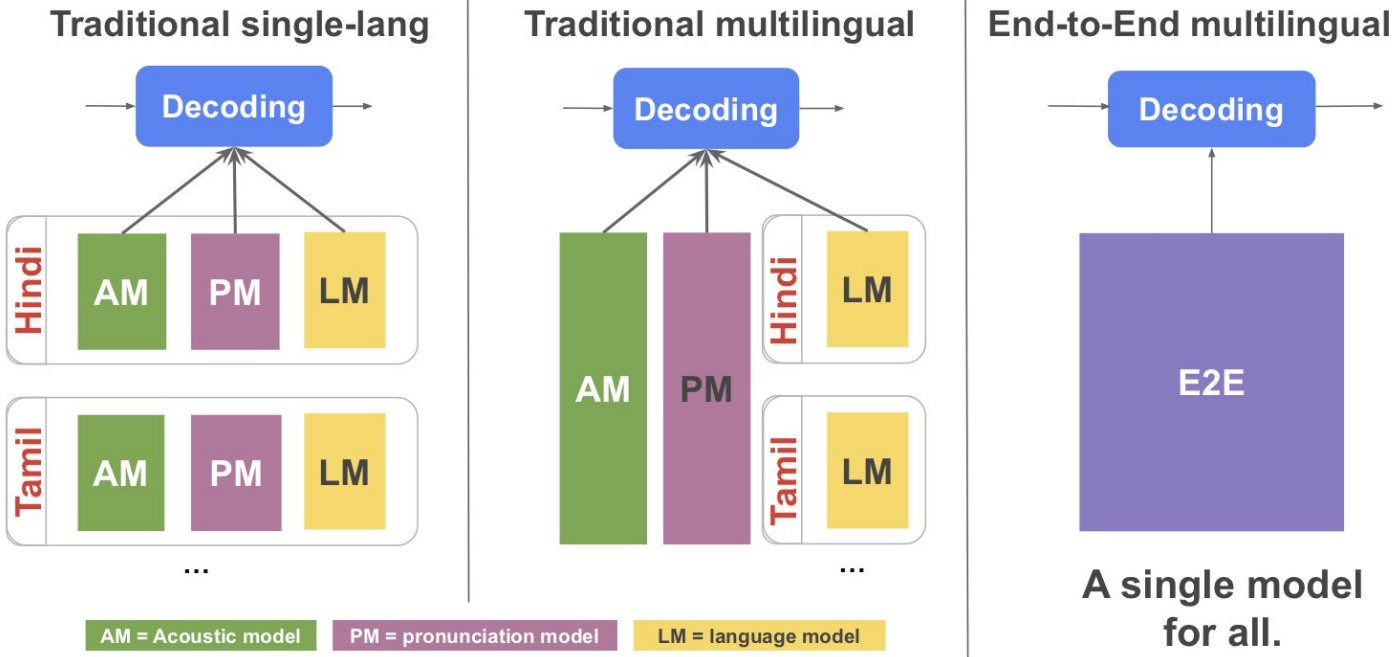
[2] Curating Datasets

- 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***

An Example of Decomposition vs. End-to-end

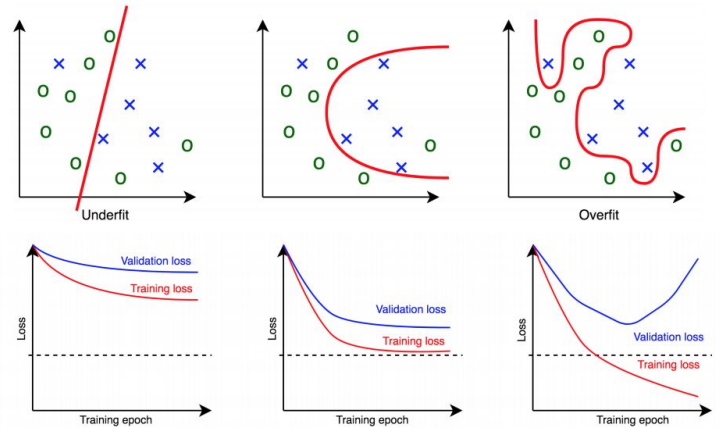


https://ai.googleblog.com/2019/09/large-scale-multilingual-speech.html

[3] Model Generalizability

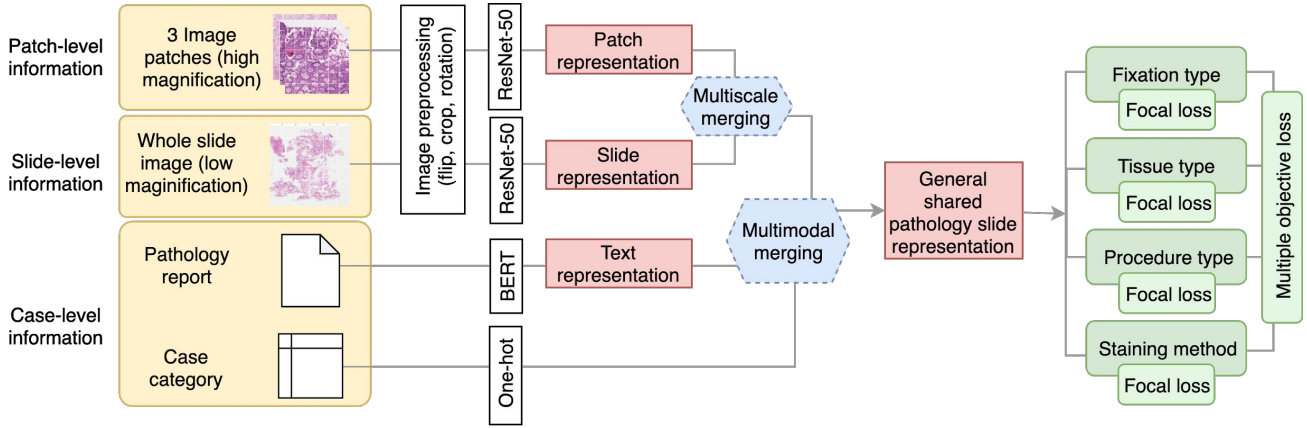
- Bias-variance trade-off
  - Balancing model underfitting/overfitting
- ***Regularization***
  - Data augmentation: particularly useful in the data-limited healthcare regime
  - Inject prior knowledge
  - Perturbations that may not correspond to real-world changes may still be helpful in learning
-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

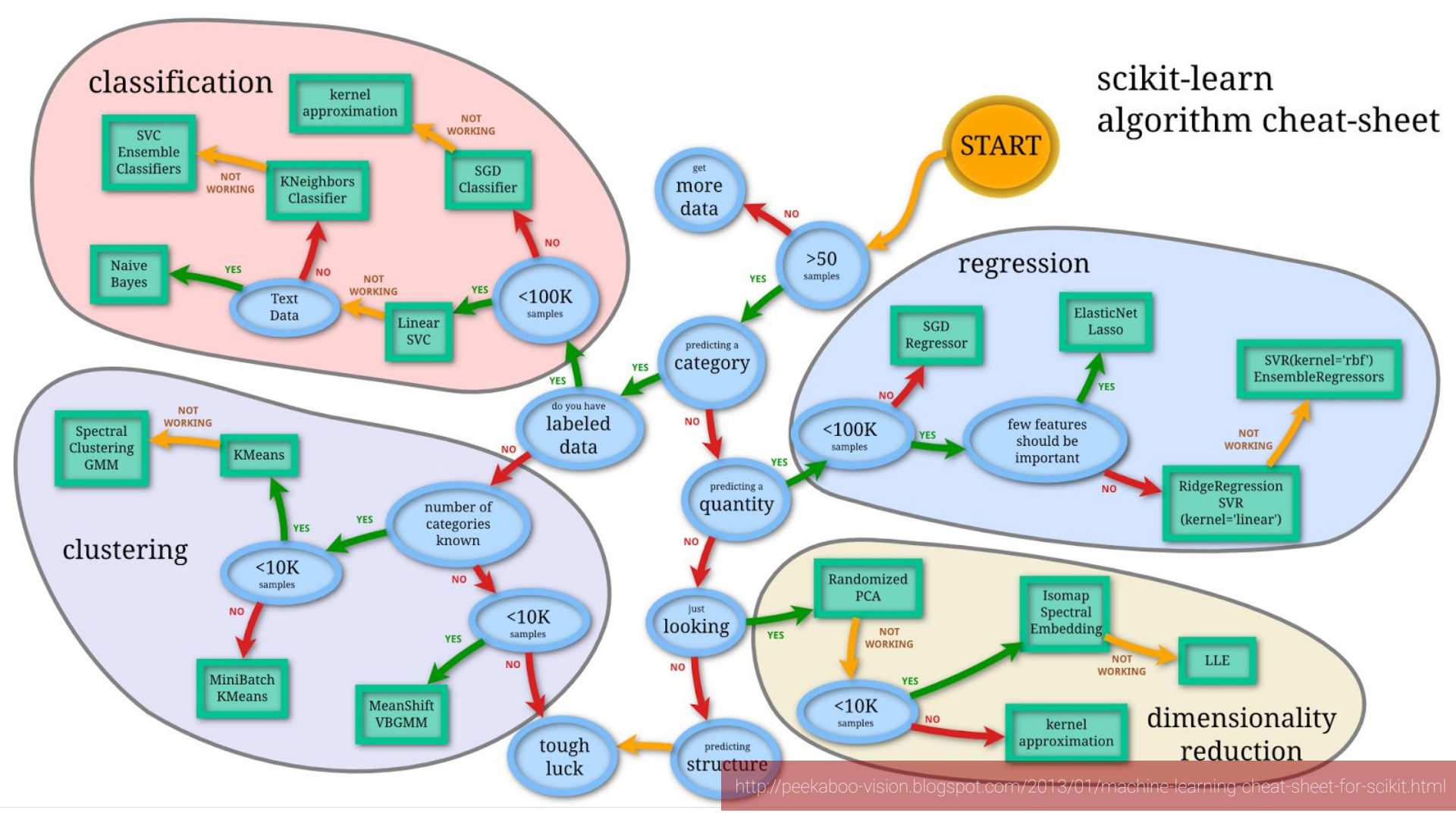
	Training error	Validation error	Approach
High bias	High	Low	Increase complexity
High variance	Low	High	Decrease complexity Add more data



[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

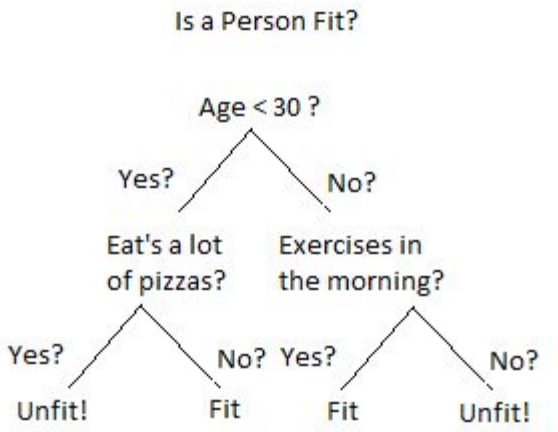
Email: AzurePoster@microsoft.com

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

Microsoft

[3] Modeling

- ***Always start from simpler algorithms / models***
  - e.g. LR, then random forest
- Deep learning is not always a good approach
  - Very limited data (transfer learning might help)
  - Interpretation is not easy




```
Call:
lm(formula = fin_loss ~ num_people + num_records + per_sensitive
+
    dys_impact + dys_detect + cost_controls, data = hw2)

Residuals:
    Min      1Q  Median      3Q     Max
-33.261  -4.734  -0.108   5.326  23.490

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.1883422  3.2315604   6.557 1.39e-10 ***
num_people    -0.0031807  0.0994055  -0.032   0.9745
num_records    0.4291799  0.1010031   4.249 2.57e-05 ***
per_sensitive  0.0216817  0.0133377   1.626   0.1047
dys_impact     0.0003192  0.0007069   0.452   0.6518
dys_detect     0.0019461  0.0011590   1.679   0.0938 .
cost_controls  0.1096839  0.0587405   1.867   0.0625 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.175 on 493 degrees of freedom
Multiple R-squared:  0.2264,    Adjusted R-squared:  0.2169
F-statistic: 24.04 on 6 and 493 DF,  p-value: < 2.2e-16
```

[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

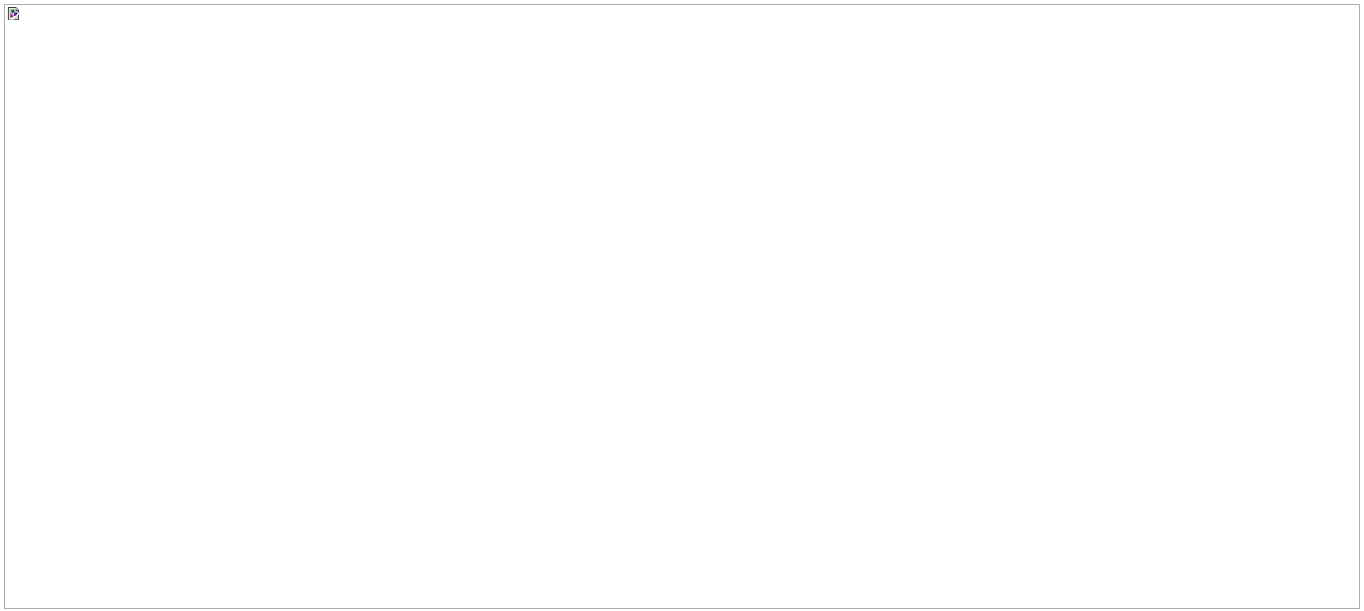
- Subgroup analysis /cluster analysis
- Sensitivity analysis
- Data augmentation for class imbalance
  - Validation set collected may have a different distribution of disease subtypes relative to real-world populations → evaluation should be adjusted according to realistic prevalence distributions
- ***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***

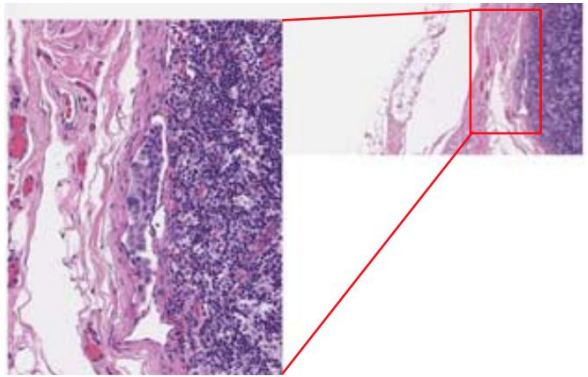
Task	Error type	Loss function	Note
Regression	Mean-squared error	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Easy to learn but sensitive to outliers (MSE, L2 loss)
	Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Robust to outliers but not differentiable (MAE, L1 loss)
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

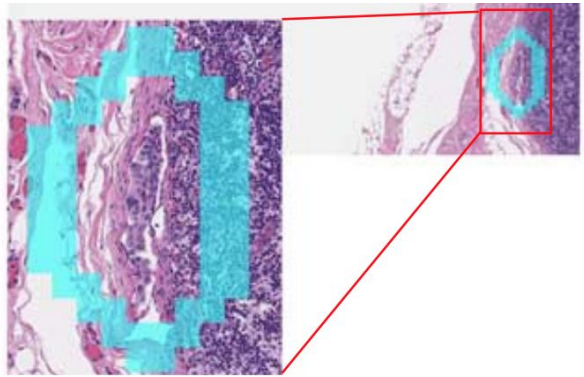
		Predicted		
		True	False	
Actual	True	True positive (TP)	False negative (FN) Type II error	Recall = Sensntivity = $\frac{TP}{TP+FN}$
	False	False positive (FP) Type I error	True negative (TN)	Specificity = $\frac{TN}{TN+FP}$
		Precision = $\frac{TP}{TP+FP}$		Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ F1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

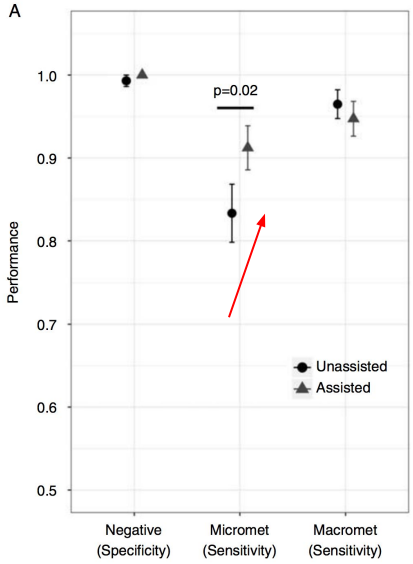


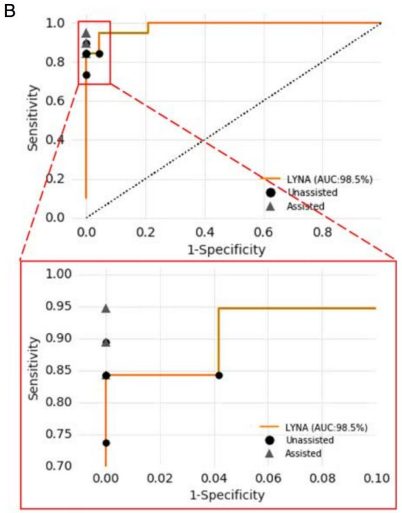
[5] Clinical Impact

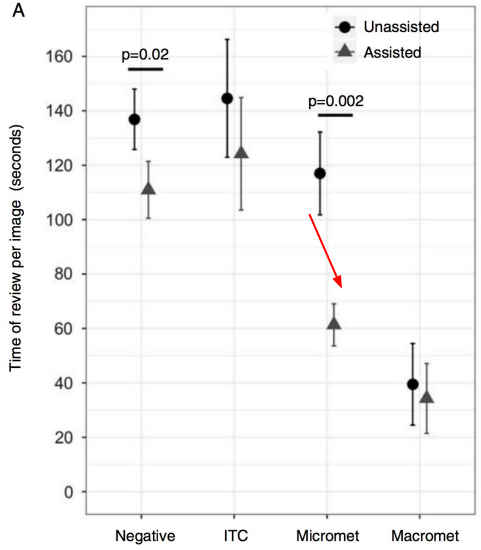
- Steiner 2018

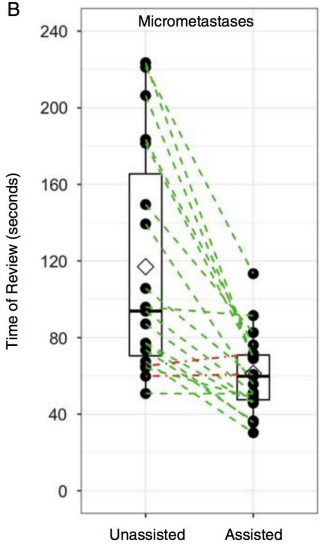






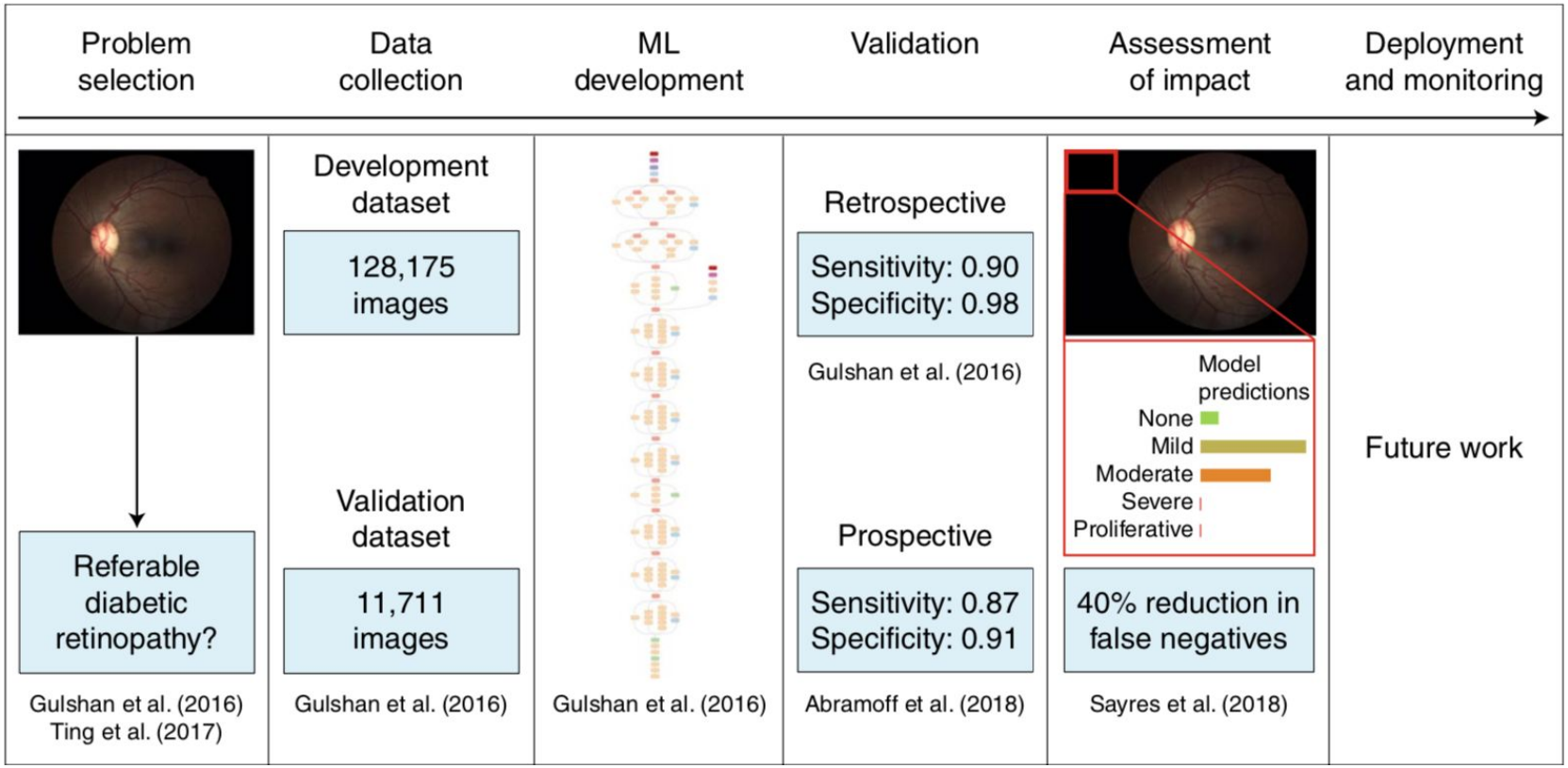






[5] Clinical Impact

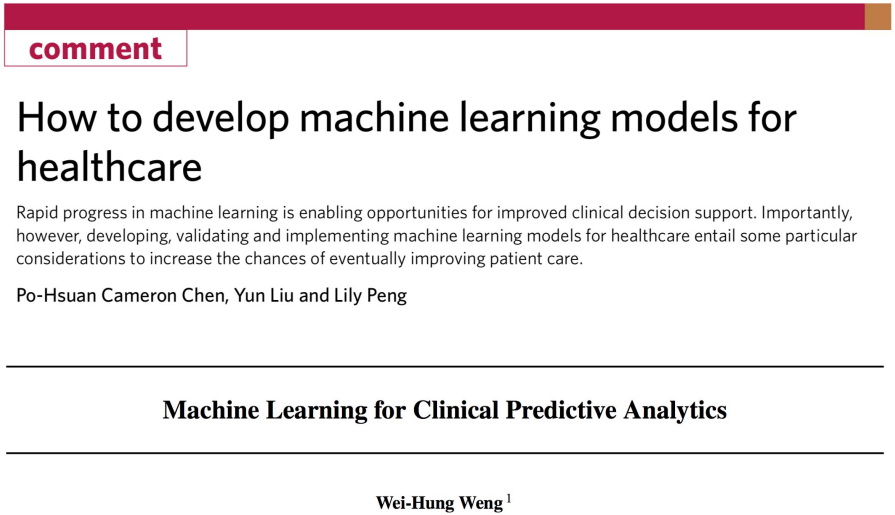
- A performant model alone is insufficient to create clinical impact
- The system has to be designed to be useful even in cases of failure such as false positives
- ‘Pre-diagnosis’ to pre-screen images and draw attention to images of high or uncertain risk, ‘peri-diagnosis’ to highlight lesions on the 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’***


Problem selection
Data collection
ML development
Validation
Assessment of impact
Deployment and monitoring
Referable diabetic retinopathy?
Gulshan et al. (2016)
Ting et al. (2017)
Development dataset
128,175 images
Validation dataset
11,711 images
Gulshan et al. (2016)
Gulshan et al. (2016)
Retrospective
Sensitivity: 0.90
Specificity: 0.98
Gulshan et al. (2016)
Prospective
Sensitivity: 0.87
Specificity: 0.91
Abramoff et al. (2018)
Model predictions
None
Mild
Moderate
Severe
Proliferative
40% reduction in false negatives
Sayres et al. (2018)
Future work

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]

MIT CRITICAL DATA

Secondary Analysis of Electronic Health Records

EXTRAS ONLINE

Resources

- Practical
  - TensorFlow, PyTorch, Keras, Scikit-learn documents, guide, tutorials
  - ***Google Machine Learning Crash Course***
  - ***Do more projects!***
- Theory and mathematics
  - Coursera Machine Learning (Andrew Ng)
  - Deep learning book
  - PMRL
  - Stanford CS224n: Natural Language Processing with Deep Learning
  - Stanford CS231n: Convolutional Neural Networks for Visual Recognition

Seeking Papers

- Clinical-driven (journals)
  - JAMA
  - Nature Medicine
  - npj Digital Medicine
  - JAMIA
  - AMIA (conference)
- ML/AI-driven (conferences / workshops)
  - MLHC
  - NeurIPS ML4H Workshop
  - NAACL Clinical NLP Workshop
  - MICCAI
  - NeurIPS, ICML, ICLR, KDD, AAAI, ACL, NAACL, CVPR, ...

tl;dr

- Caveats and tips for model development [Chen 2019, Weng 2019]
  - Problem definition →
  - Data curation →
  - Model development →
  - Validation →
  - Assessment of clinical impact →
  - Deployment and monitoring
- Wei-Hung Weng [ckbjimmy@mit.edu]

Framing Your Clinical Questions into ML Tasks

- Clinical problems?
- Healthcare workflow?
- Socioeconomic problems?

- ?

What’s Special in Bhutan?

- Clinical problems?
- Healthcare workflow?
- Socioeconomic problems?

- ?

Clinical Problem → ML Task

[Problem] ICU mortality prediction [Ghassemi KDD 2014]

[ML task] Prediction (supervised)

[Data] MIMIC-III structured data + notes

[ML model] Unupservised LDA + structured SVM

[Baseline] ***Admission baseline*** / retrospective topic / time-varying model

[Evaluation] AUC, Topic ***post-hoc analysis***

Clinical Problem → ML Task

[Problem] Want to know better subtyping in the cohort, e.g. autism?

[ML task] Cluster analysis (unsupervised learning)

[Data] Data with features

[ML model] ***K-means***, PCA, autoencoder

[Evaluation and validation] ***Visualization, post-hoc explanation***, downstream tasks (then you need annotation and a model for downstream task)

Clinical Problem → ML Task

[Problem] What may happen in the next clinical visit? [Choi MLHC 2016]

[ML task] Disease progression modeling, sequential prediction

[Data] Clinical visit claims data (input codes, output ICDs)

[ML model] RNN

[Baseline] ***LR***, most frequent, last visit

[Evaluation] Recall@30

Clinical Problem → ML Task

[Problem] How to know if the patient may readmit? [Caruana 2015, Xiao 2018]

[ML task] Prediction, supervised learning

[Data] Labs, notes, history of hospitalized patients

[ML model] Generalized additive models [Caruana 2015], topic modeling + RNN [Xiao 2018]

[Baseline] / ***word2vec+LR***, GRU, ... [Xiao 2018]

[Evaluation and validation] AUC, ***risk score*** [Caruana 2015] / Acc, AUC, PR-AUC, ***clustering*** [Xiao 2018]

Clinical Problem → ML Task

[Problem] How to find an optimal treatment for my patient's condition?
[Komorowski Nat Med 2018]

[ML task] Optimal sequential decision making (Reinforcement learning)

[Data] MIMIC-III, eICU (time-series structured data, vitals, labs, medical orders, ...)

[ML model] Policy iteration

[Baseline] Doctors' action in EHR

[Evaluation and validation] Value (this is challenging)

Clinical Problem → ML Task

[Problem] How to detect the severe (referrable) diabetic retinopathy? [Gulshan JAMA 2016]

[ML task] Image classification (supervised learning)

[Data] DR images with annotations

[ML model] Inception V3

[Baseline] ***Ophthalmologists' manual grading***

[Evaluation and validation] ***Sensitivity and specificity***

Clinical Problem → ML Task

[Problem] Decide the discharge ICD code based on the clinical notes? [Mullenbach NAACL 2018]

[ML task] NLP + multilabel classification

[Data] MIMIC-III

[ML model] CNN + attention + classifier

[Evaluation and validation] Precision@K, AUC, ***physician evaluation***

Clinical Problem → ML Task

[Problem] How to translate clinical jargons to layman understandable language? [Weng KDD 2019]

[ML task] Machine translation / text style transfer (unsupervised or supervised if we have pairs)

[Data] MIMIC-III (some professional-level, some layman-level)

[ML model] Cross-aligned representation + statistical machine translation

[Baseline] ***Dictionary-based translation***

[Evaluation and validation] Precision@K, ***mean opinion score***

More Problems from ICU

Classification

- Fit paCO2 to pH / fit paCO2 and HCO3 to pH
- Classify ICU mortality using HCO3 min/median & paCO2 max/median
- Decision tree / random forest to determine mortality based on Hct, PLT min or WBC max (or based on vital signs, etc.)

Clustering

- Identify clusters of patients with HCO3/paCO2 or vital signs who did die or did not die during ICU stay
- Clustering vital signs and risk of being initiated on mechanical ventilation

More Problems from Respiratory Therapy

Hands-on Exercise

- https://github.com/ckbjimmy/2019_bt
  - Open → **"File" -> "Save a copy in Drive..."**
- Requirement
  - Google account
  - Python + some ML knowledge
- scikit-learn / tensorflow-keras
- Two datasets
  - ICU structured data / Breast cancer prediction data
- Supervised learning (classification)
  - When you have some labeled data
  - Given features, predict malignancy
  - ML general approaches and modeling, missing data imputation, normalization, important feature identification, ...