Machine Learning for Clinical Predictive Analytics

Workshop

Wei-Hung Weng

BIG DATA FOR HEALTH WORKSHOPS AND CONFERENCE
Jul 10, 2018

Massachusetts Institute of Technology

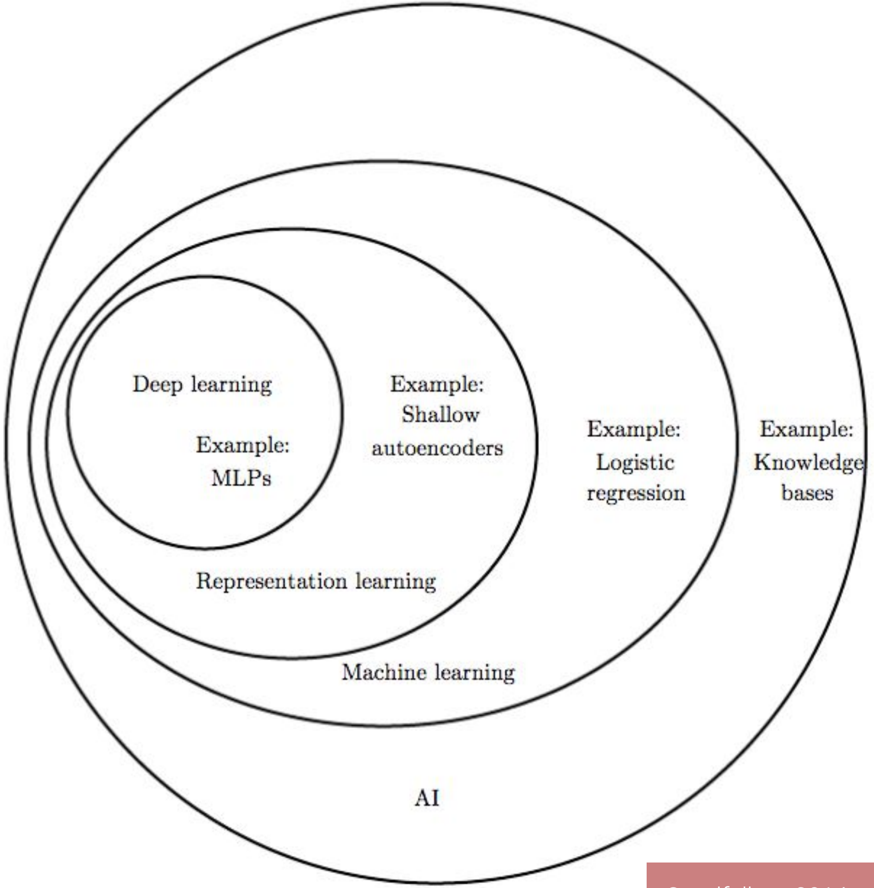
CSAIL

Artificial intelligence (AI) is defined by
the American cognitive scientist Marvin Minsky
as **the science of making machines do things
that would require intelligence if done by man**

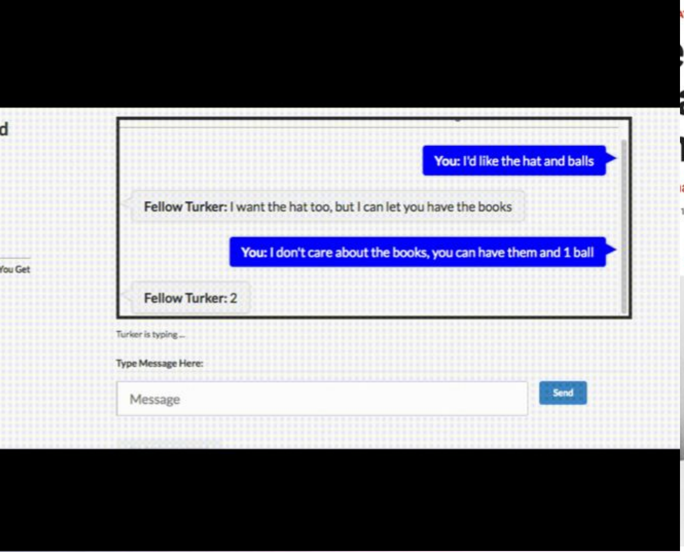
AI?

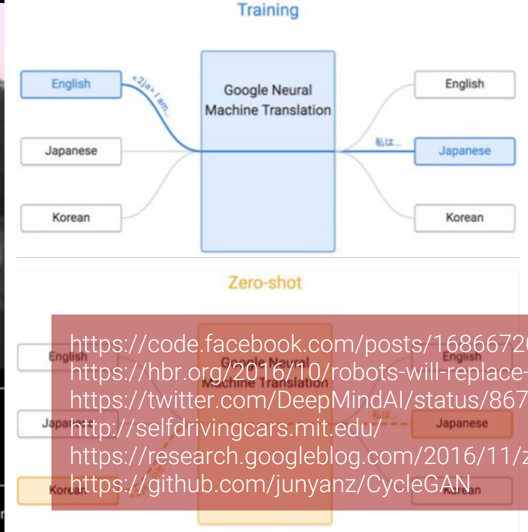
Human intelligence as a **goal**

Algorithms / Data science as an **approach**



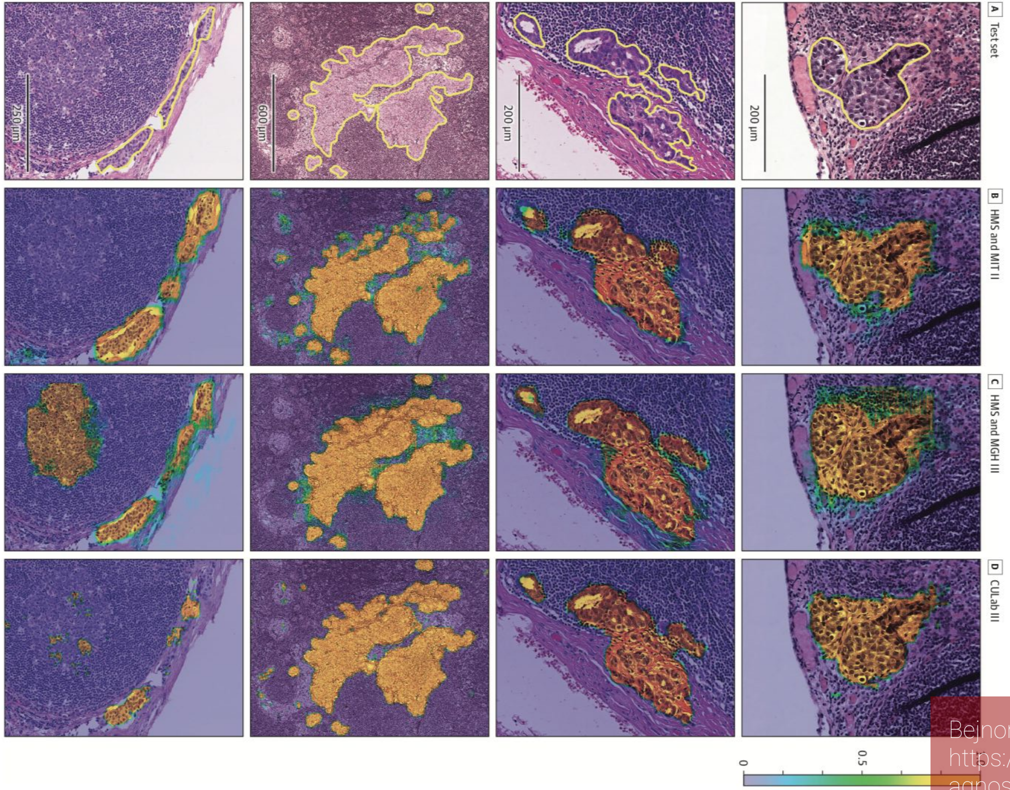
Goodfellow, 2014





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

Biomedical Imaging Informatics



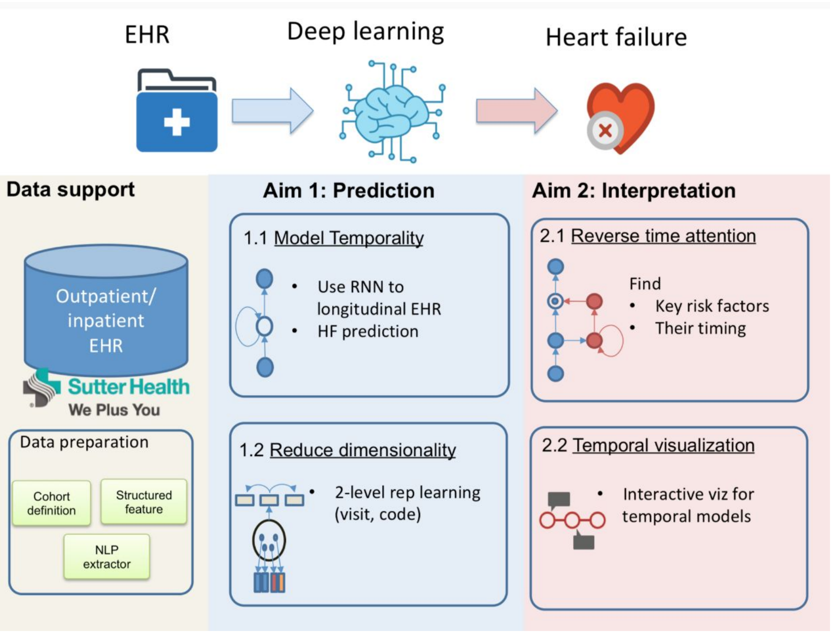
FDA approves first AI-powered diagnostic that doesn't need a doctor's help

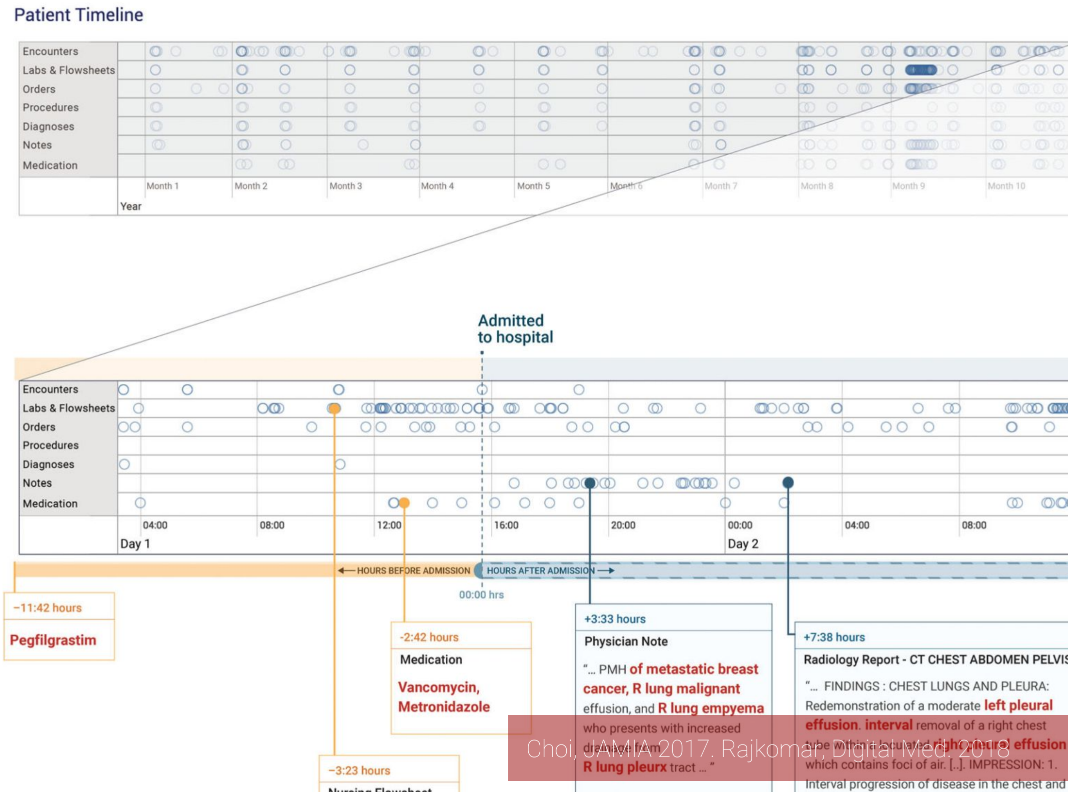
Marking a new era of "diagnosis by software," the US Food and Drug Administration on Wednesday gave permission to a company called IDx to market the first AI-powered diagnostic device.

What it does: The software is designed to detect greater than a mild level of diabetic retinopathy, which causes vision loss and affects 30 million people in the US. It occurs when high blood sugar damages blood vessels in the retina.

Bejnordi et al. JAMA 2017
https://www.technologyreview.com/the-download/610853/fda-approves-first-ai-powered-diagnostic-that-doesnt-need-a-doctors-help/

Clinical / Public Health Informatics





Choi, JAMIA 2017. Rajkomar, Digital Med. 2018

Sciences Behind

Data organization / structuring

- Database
- Knowledge representation / Ontology
- Visualization

Learning from data (Algorithms)

- Statistics, probability, information theory, ...
- **Machine learning**
  - Computer vision, natural language processing, signal processing, …

Machine Learning

Optimize a performance criterion using example data or past experience

- Given data X, we want to learn a function mapping f(X) for certain purpose
  - Patient age, gender, vitals, labs, … → mortality Y/N
- ML - Given objective and evaluation metrics, and get high quality f(X)

Three steps

1. Modeling
   - Choose functions, features, (hyper)parameters
2. Evaluating the function
   - Define loss function, optimizer
3. Choosing the best one
   - Decide bias/variance trade-off

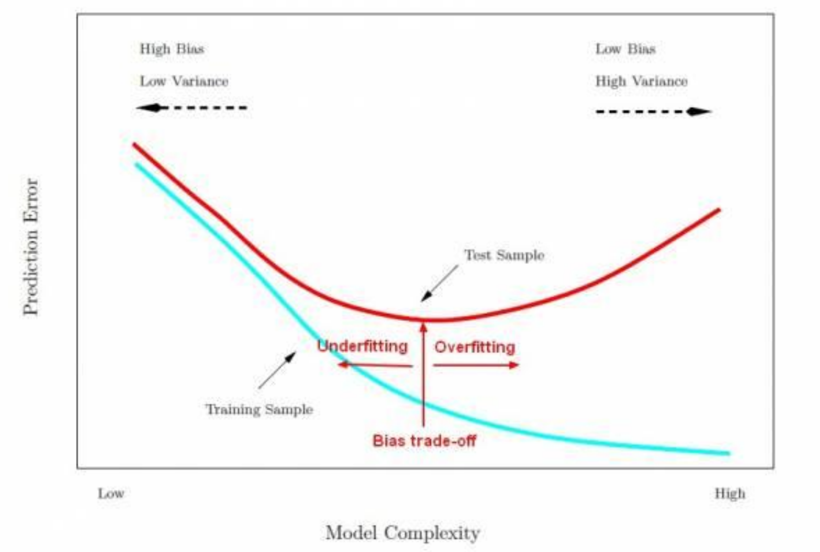
Bias vs. Variance

Bias (underfitting)

- High training/validation error
- Train more, increase model complexity, decrease regularization, add features

Variance (overfitting)

- Low training error but high validation/testing error
- More data, reduce model complexity, add regularization, reduce features

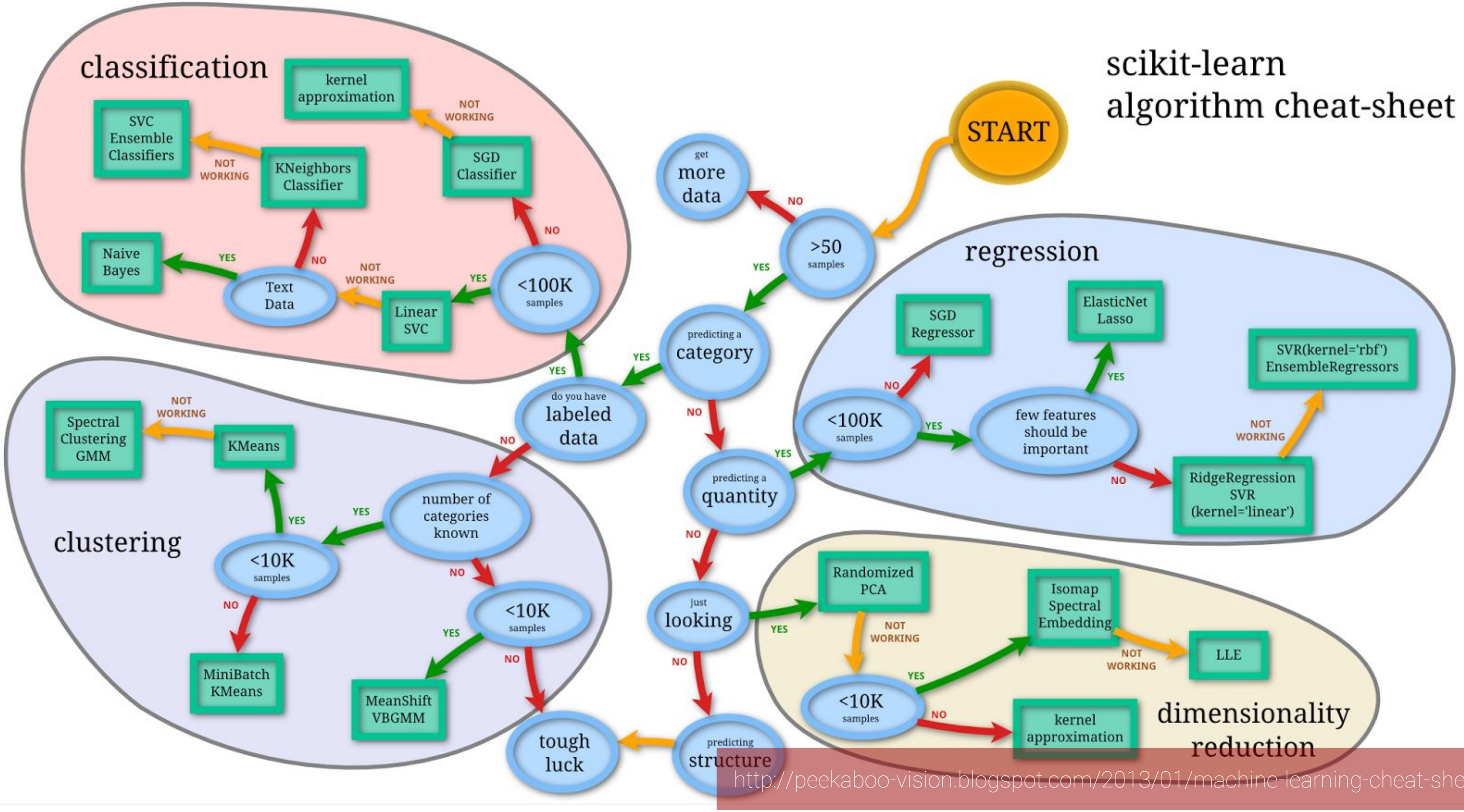


https://gerardnico.com/_media/data_mining/model_complexity_error_training_test.jpg?w=600&tok=9ed6da

Scenario

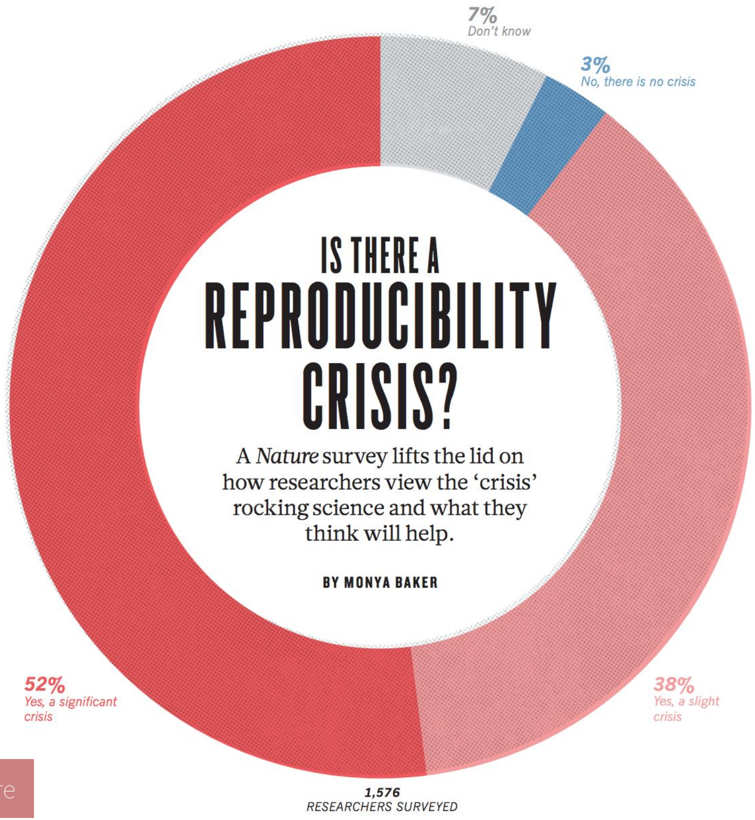
- Supervised learning
  - Regression
  - **Classification**
    - Linear
    - Non-linear (e.g. **deep learning**, SVM, decision tree, ...)
  - Structured learning
- Unsupervised learning
  - **Clustering**
  - **Dimensionality reduction**
- Transfer learning
- Reinforcement learning

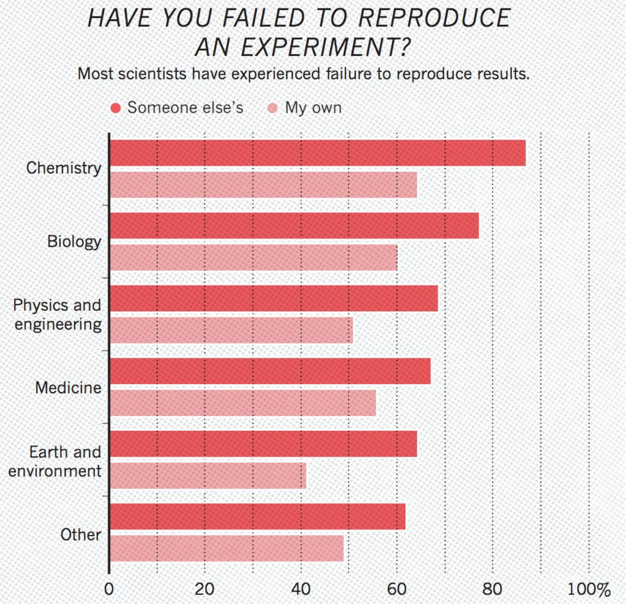
Algorithms

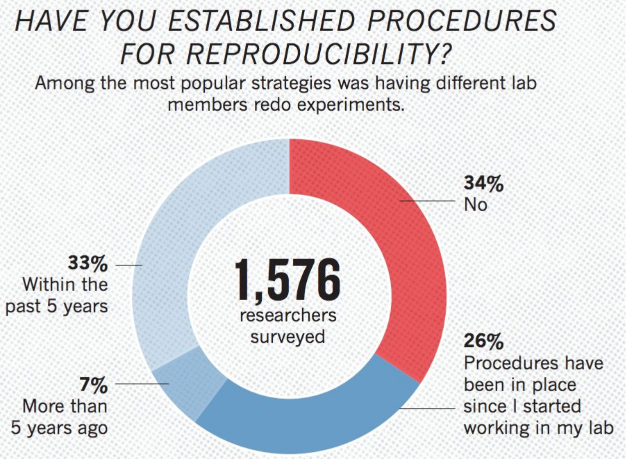


http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html

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

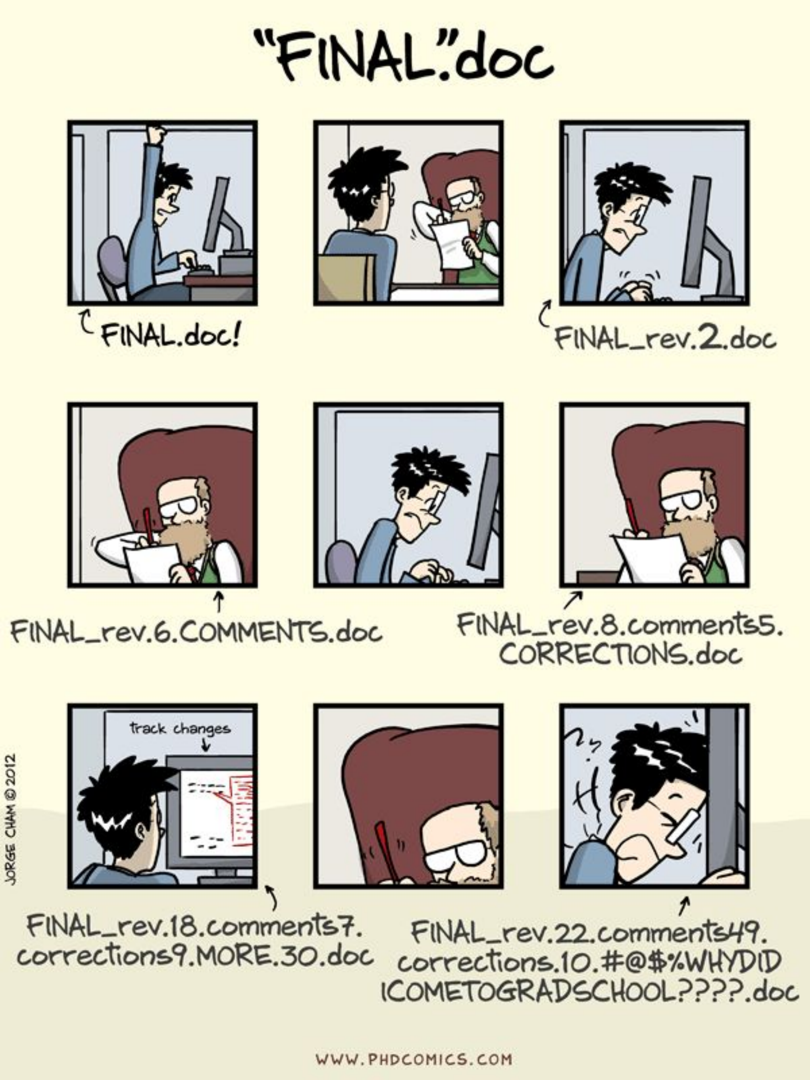
Reproducibility

What may be useful...

- Executable notebooks - `.ipynb`
- Code publishing platform - GitHub
- Version control - Git

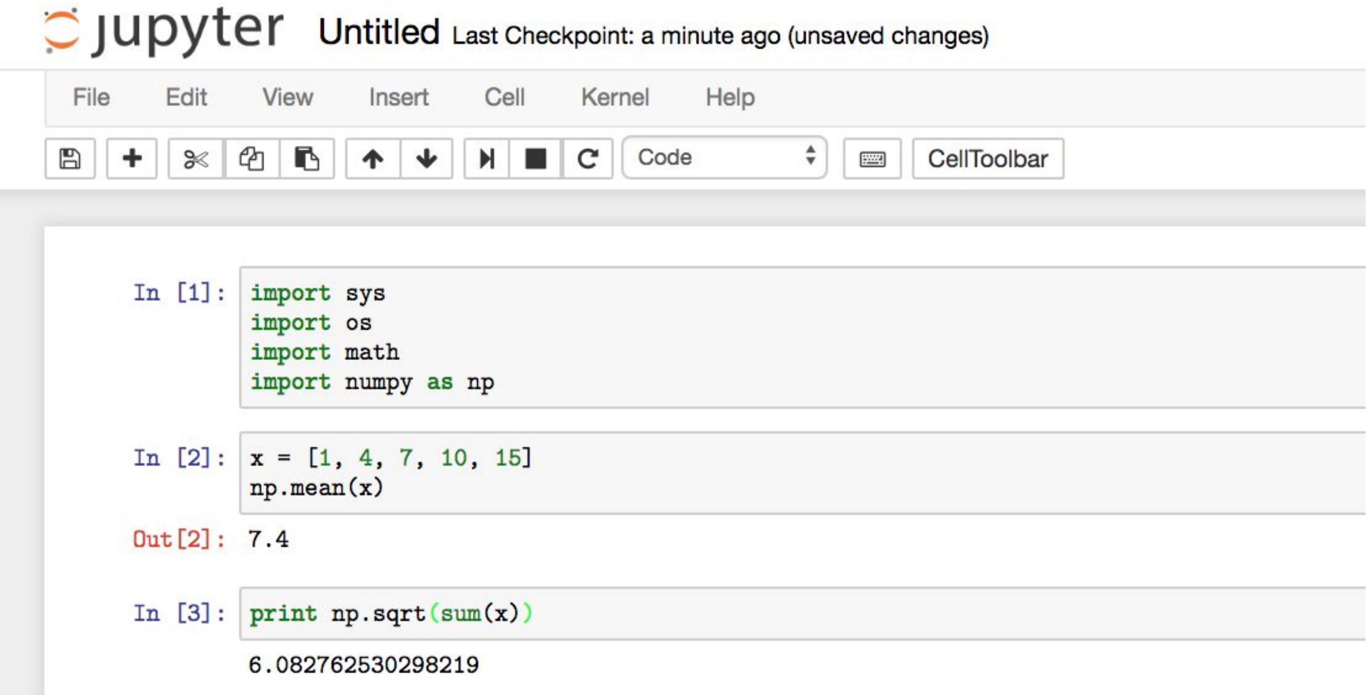
Google colab

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

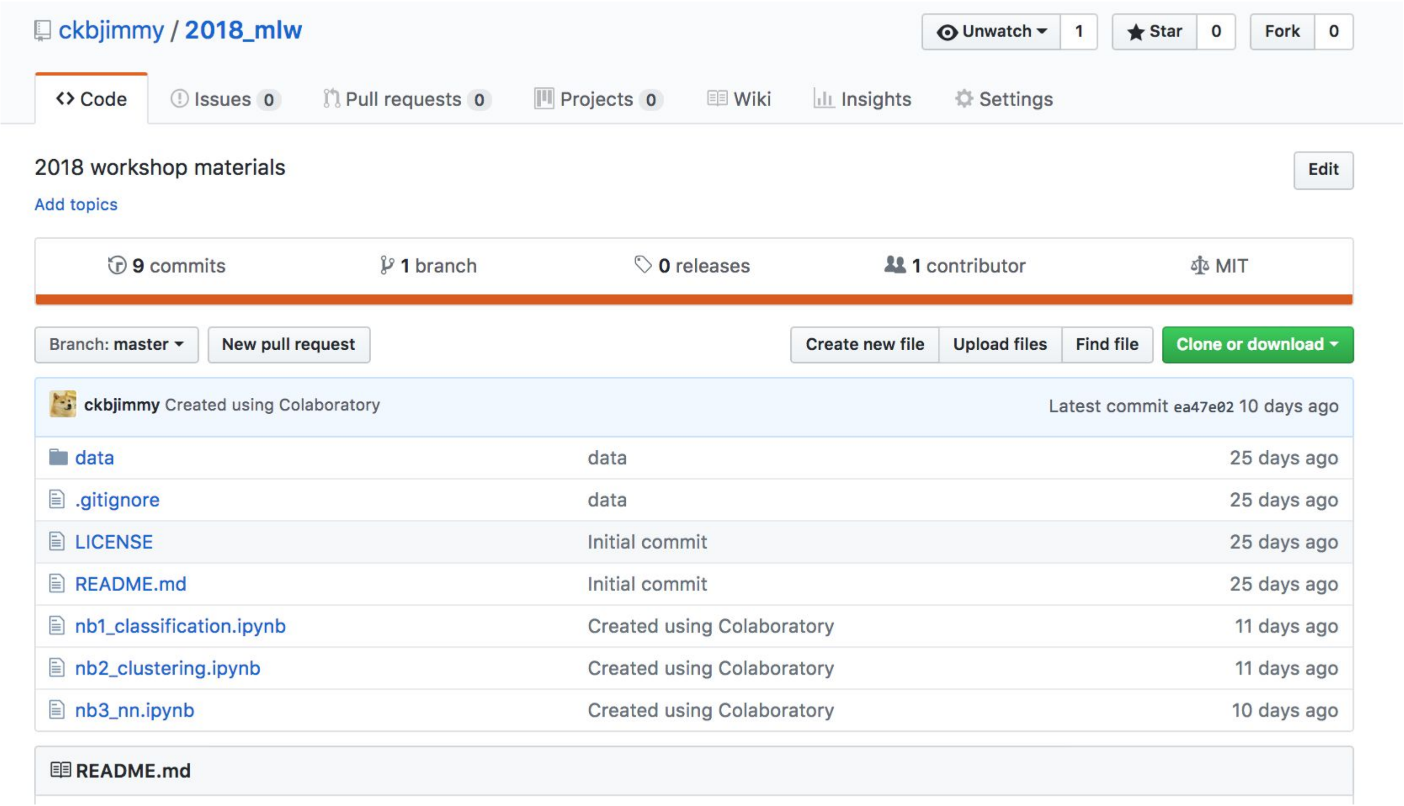


Jupyter Notebook

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



Code Publishing Platform



Version Control

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

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`

Tutorial

- https://github.com/ckbjimmy/2018_mlw
    - Open → **"File" -> "Save a copy in Drive..."**
- Use python scikit-learn / keras
- Two datasets
- Part 1 - Supervised learning (classification)
    - When you have some labeled data
    - Given features, predict malignancy
    - ML general approaches, missing data imputation, normalization, important feature identification, …
- Part 2 - Unsupervised learning (clustering / dimensionality reduction)
    - When you don't have labeled data
    - Grouping the similar cases
    - K-means / PCA

Tutorial

- Part 3 - Neural network
  - Deep feedforward neural network
    - ICU structured data
    - Breast cancer prediction data
  - Convolutional neural network (CNN) for image (MNIST)
  - Recurrent neural network (RNN) for text (IMDB reviews)

More ICU Problems

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

Further Readings

Theory and mathematics

- Coursera Machine Learning (Andrew Ng)
- Deep learning book
- Stanford CS224n: Natural Language Processing with Deep Learning
- Stanford CS231n: Convolutional Neural Networks for Visual Recognition

Practical

- TensorFlow, PyTorch, Keras, Scikit-learn documents, guide, tutorials
- Google Machine Learning Crash Course
- Coursera Deep Learning Specialization
- Coursera Machine Learning with TensorFlow on Google Cloud Platform