ABCs of Machine Learning for Epidemiology

SER WORKSHOP 2022

ERIC LOFGREN AND JEANETTE STINGONE

Contact us:

Eric Lofgren
eric.lofgren@wsu.edu
@GermsAndNumbers

Jeanette Stingone
j.stingone@columbia.edu
@jstingone

Schedule for the 4-hour Workshop

0:00-0:50	Introduction and General Concepts
0:55-1:45	Evaluation: Understanding bias, fairness and error in the context of Machine Learning
1:45-2:00	15 minute Break
2:00-2:50	Implementation in R: The Caret Package
2:55-3:45	Machine Learning beyond Prediction and The Role of Epidemiology
3:45-4:00	Wrap-Up and Questions

Start a discussion, post Q&A, etc on the Slack Channel shorturl.at/fiTY0

All materials available at https://github.com/jstingone/mlworkshop2022

Introduction and General Concepts

Everyone’s Talking about Machine Learning

American Journal of Epidemiology

Vol. 190, No. 8
https://doi.org/10.1093/aje/kwab047
Advance Access publication:
March 6, 2021

Special Article

Thirteen Questions About Using Machine Learning in Causal Research (You Won’t Believe the Answer to Number 10!)

JAMA Network | Open

Original Investigation | Statistics and Research Methods

Use of Machine Learning to Estimate the Per-Protocol Effect of Low-Dose Aspirin on Pregnancy Outcomes

A Secondary Analysis of a Randomized Clinical Trial

Journal of Traumatic Stress
Published by the International Society for Traumatic Stress Studies
ISTSS
WILEY

Research Article | Full Access

Gender Differences in Machine Learning Models of Trauma and Suicidal Ideation in Veterans of the Iraq and Afghanistan Wars

American Journal of Epidemiology

https://doi.org/10.1
Advance A

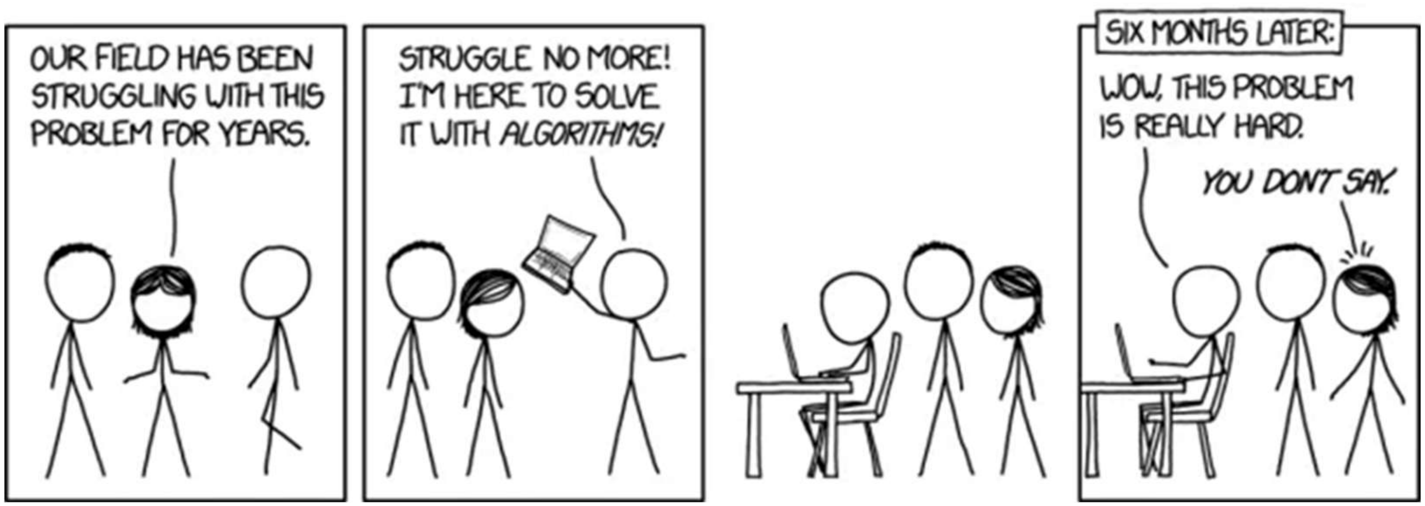
Invited Commentary

Invited Commentary: Machine Learning in Causal Inference—How Do I Love Thee? Let Me Count the Ways

Annual Review of Public Health

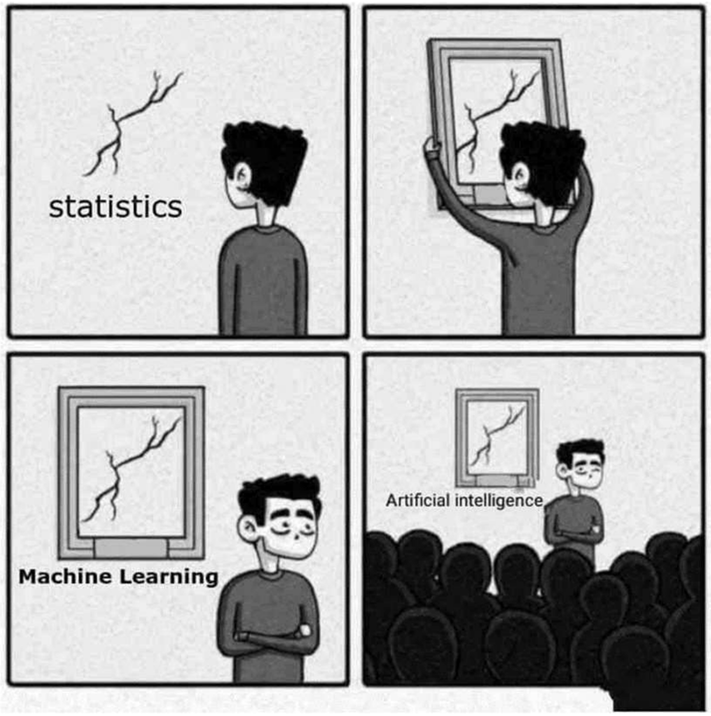
Machine Learning in Epidemiology and Health Outcomes Research

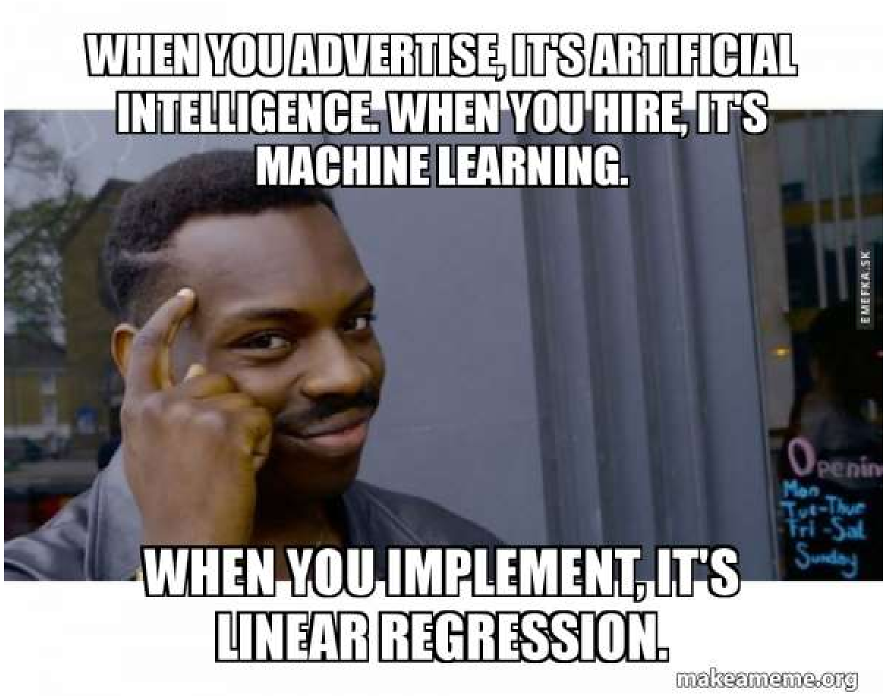
Machine Learning is not Magic



Here to Help: https://xkcd.com/1831

On the flip side, are some too cynical?





Epidemiologists use tools for different purposes

Questionnaire Development

CLINICAL TEST PROTOCOLS

Biological Assays

Propensity Scores

Community Engagement

EXPOSURE MODELING

Regression

AGENT-BASED MODELS

Machine Learning??

Utility of the Tool Depends upon the Problem

JAMA Network | Open

Original Investigation | Cardiology

Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes

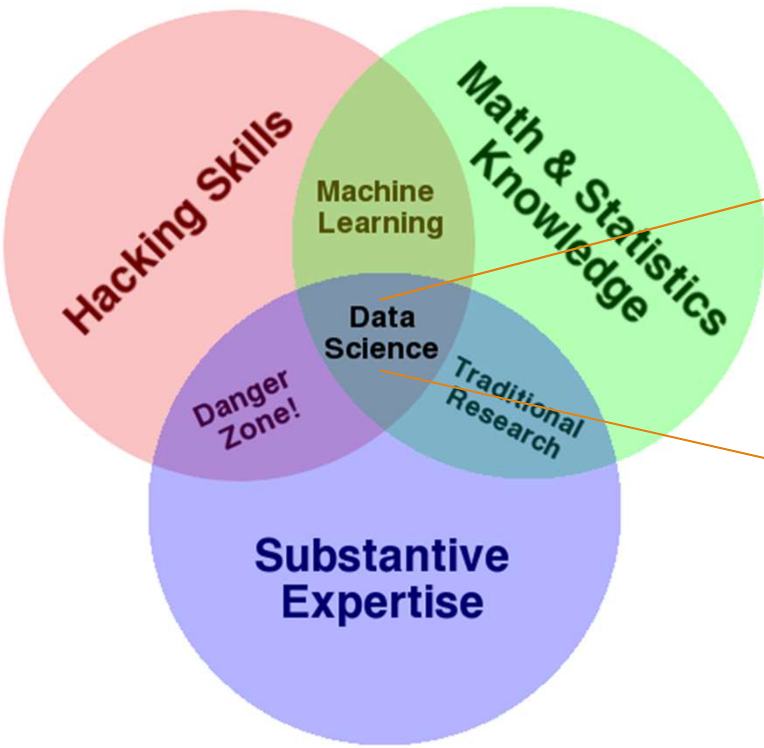
6113 obs in training
3389 obs for testing

54 variables from Medicare claims
8 variables from EHR

"In our study, we observed that when using only claims-based predictors, many of which are binary variables indicating presence or absence of medical conditions or use of specific medications, the performance improvement with machine learning approaches was minimal for prediction of most outcomes. However, when the predictor set was expanded to include EMR-based information, which included numerous laboratory test results as continuous variables, we noted that machine learning approaches generally fared better than logistic regression. This observation follows the intuition that, because tree-based machine learning approaches, such as GBM or random forests, are nonparametric and do not assume linearity for a predictor-outcome association, they are usually more adept at generating predictions based on continuous variables."

What is Machine Learning and Why should Epidemiologists Care about it?

Machine Learning: Intersection between Computational and Mathematical/Statistical Knowledge



Machine learning combined with substantive expertise can allow us to advance our research and gain more from our scientific data

The utility of machine learning (or any complex method) will depend upon our research question and the structure/content of our data

To Explain or To Predict: What is the question…and what is the difference?

Explanatory Modeling: use of statistical models to test (or estimate) hypothesized causal associations; requires pre-existing causal model

Predictive Modeling: use of data to develop model that can predict new or future observations

Machine learning approaches traditionally used **AND** developed for prediction goals.

- Note there are questions of prediction within explanatory modelling
  - construction of propensity scores
  - use of risk scores to account for confounding
  - predicting the counterfactual
- If goal is not prediction, do we need to adapt machine learning approaches for our goal?

But what if my goal is explanation, but I don't have a good pre-existing causal model…..

- "By capturing underlying complex patterns and relationships, predictive modeling can suggest improvements to existing explanatory models" ---Shmueli 2010

Can machine learning help us navigate our new data reality?

How can Epidemiologists Benefit from Training in Machine Learning?

- Facilitate use of large and/or complex data where relationships cannot be easily visualized
  - Use of ML approaches can identify patterns in data; potentially generate hypotheses, refine metrics of exposure and/or outcome

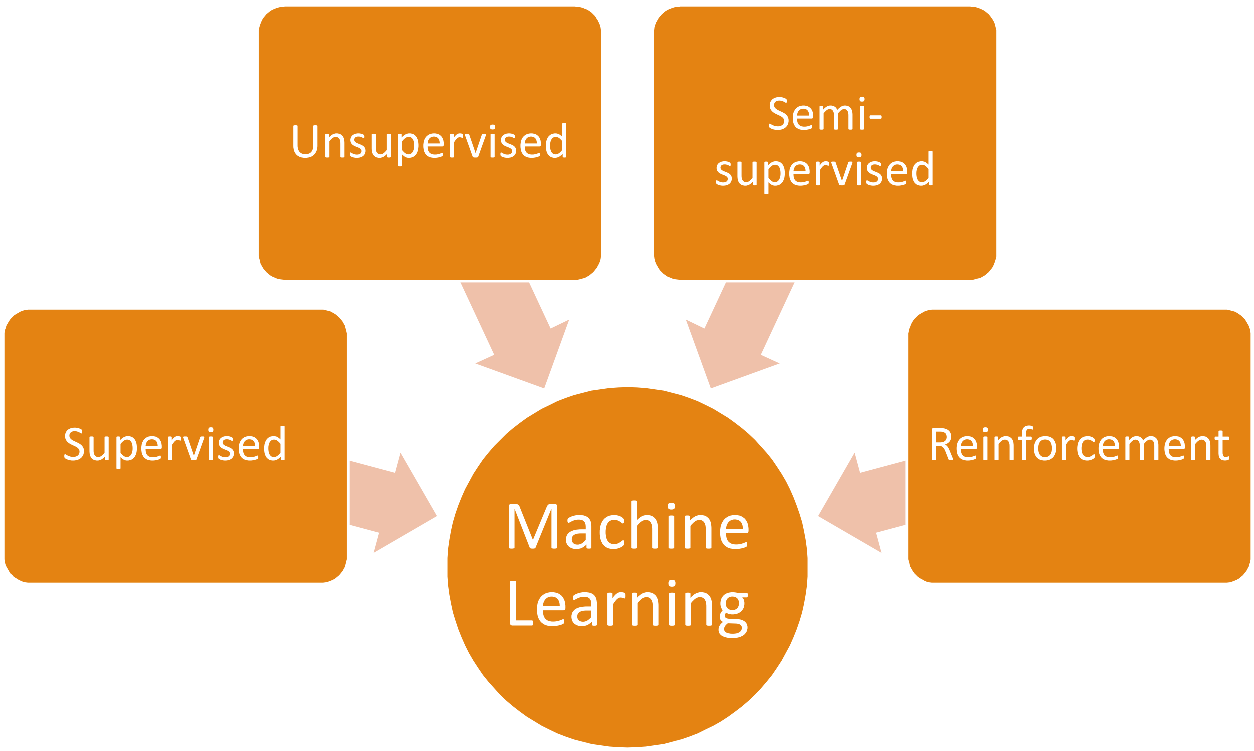
- Make exploratory data analysis and model selection more formal
  - Similar to use of DAGs to explicitly represent assumptions of relationships between variables
  - Don’t just publish the final model, show how you arrived there.

- Greater consideration of questions of prediction and how they can benefit public health

- Improve methods for causal inference

What are the different types of machine learning?

Types of Machine Learning



Unsupervised

Context: for each observation of the inputs (predictor/exposure/independent variables), there is no associated output (response measurement); also described as data are “unlabeled”

Algorithm identifies patterns within the vector of inputs and generates an output that seeks to understand or represent the relationships between variables and/or observations.

Addresses: Clustering and Dimension Reduction Problems

Clustering to refine the outcome classification

International Journal of GYNECOLOGY & OBSTETRICS

CLINICAL ARTICLE | Open Access

Cluster analysis identifying clinical phenotypes of preterm birth and related maternal and neonatal outcomes from the Brazilian Multicentre Study on Preterm Birth

Clustering for Exposure Assessment

Vol. 127, No. 10 | Research

Air Pollution, Clustering of Particulate Matter Components, and Breast Cancer in the Sister Study: A U.S.-Wide Cohort

Alexandra J. White, Joshua P. Keller, Shanshan Zhao, Rachel Carroll, Joel D. Kaufman, and Dale P. Sandler

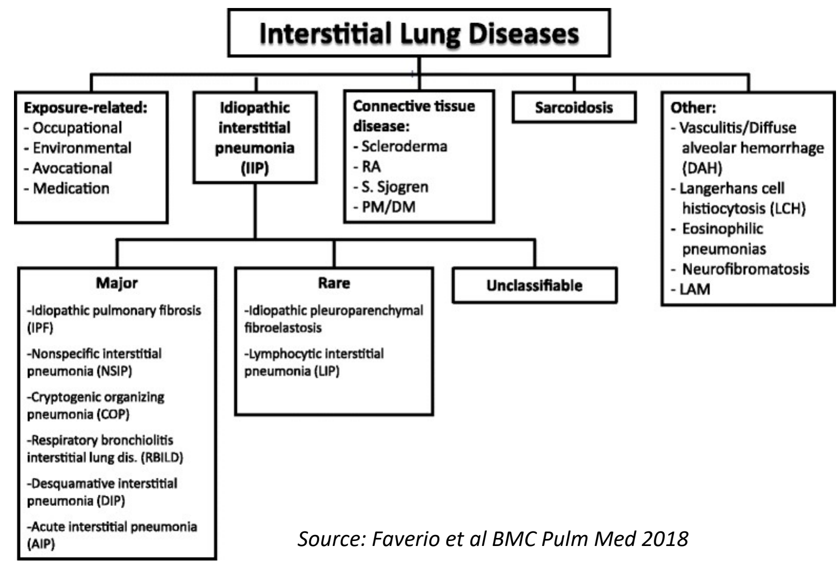
Published: 9 October 2019 | CID: 107002 | https://doi.org/10.1289/EHP5131 | Cited by: 12

Example: Phenotypic Subtypes

ELSEVIER
Chest
Volume 153, Issue 2, February 2018, Pages 349-360
CHEST
Original Research: Diffuse Lung Disease

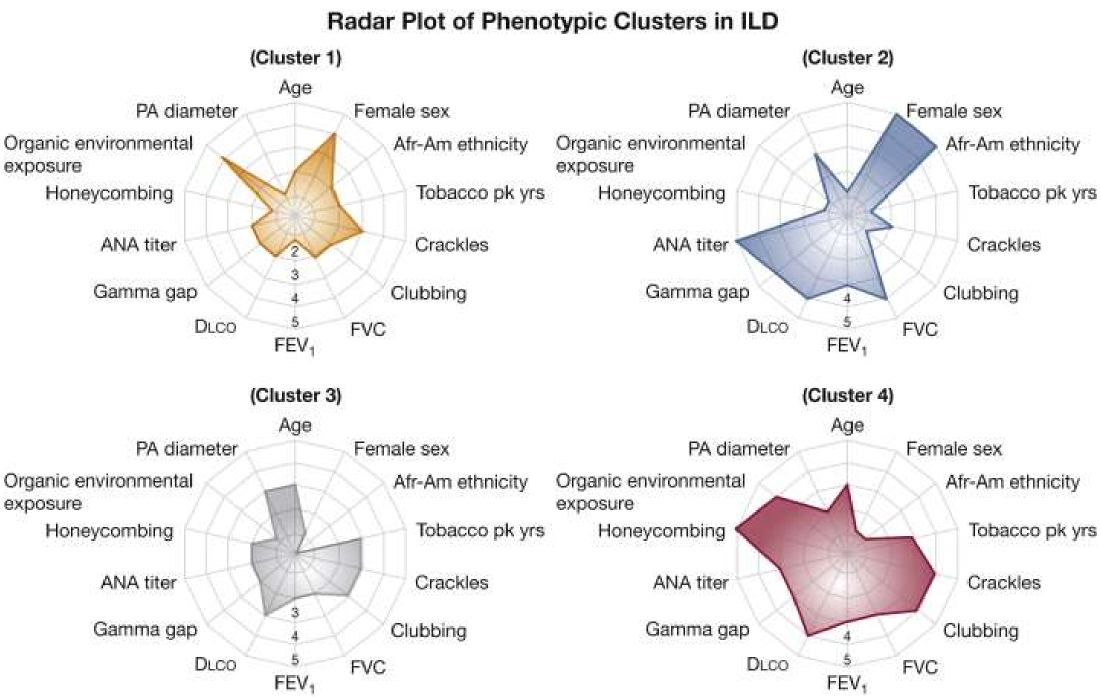
Phenotypic Clusters Predict Outcomes in a Longitudinal Interstitial Lung Disease Cohort

Goal: "Identify distinct clinical phenotypes in heterogeneous diseases"



Source: Faverio et al BMC Pulm Med 2018

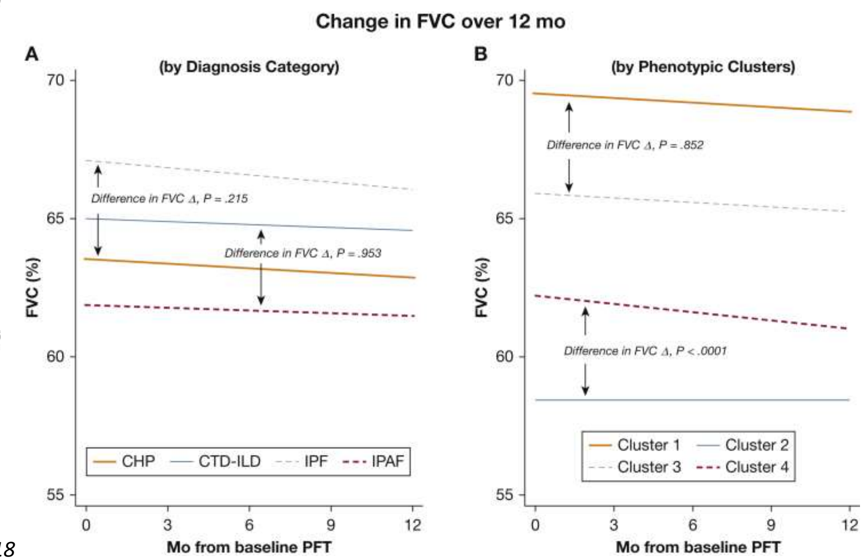
Method: Partitioning around Medoids (PAM)



Why PAM?

Similar to K-means, but more robust to outliers

Relies on median distances rather than means



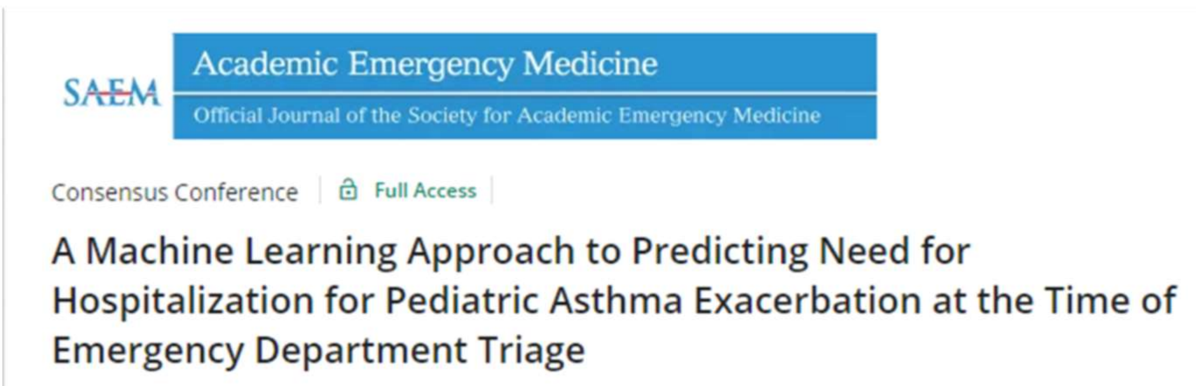
Source: Adegunsoye et al Chest 2018

Supervised

Context: for each observation of the inputs (predictor/exposure/independent variables), there is an associated output (response measurement); also described as data are “labeled”

Algorithm learns how to use inputs to generate outputs through training and receives feedback by looking at actual outcomes; process is “supervised”

Addresses: Regression, Classification and Estimation Problems



SAEM Academic Emergency Medicine

Official Journal of the Society for Academic Emergency Medicine

Consensus Conference | Full Access

A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage

Example Applications of Supervised ML

Traditionally…..Questions of Prediction

- ❑ Identify individuals/communities most in need of treatment or intervention
- ❑ Forecast future observations for planning/resource allocation

Structured Analytic Pipeline

- ❑ Train a model to predict some outcome then test it on "unseen" data to evaluate performance

More Recently…Integrated with Other Methods to Advance Causal Inference

- ❑ Generate propensity scores or IPTW to improve exchangeability
- ❑ "Predicting" the Counterfactual

Commonly-used Algorithms

Unsupervised

K-Means

Hierarchical clustering

PCA

Self Organizing Maps

Gaussian Mixture Models

Supervised

Support Vector Machines

Naïve Bayes

K-Nearest Neighbors

Regularized Regression

Decision Trees

Neural Networks

Ensemble

- Bagging
    - Random Forest
- Boosting
    - XGBoost
- Stacking
    - SuperLearner

Critical thinking is not optional….

Credit: XKCD

Consider needs of research question….

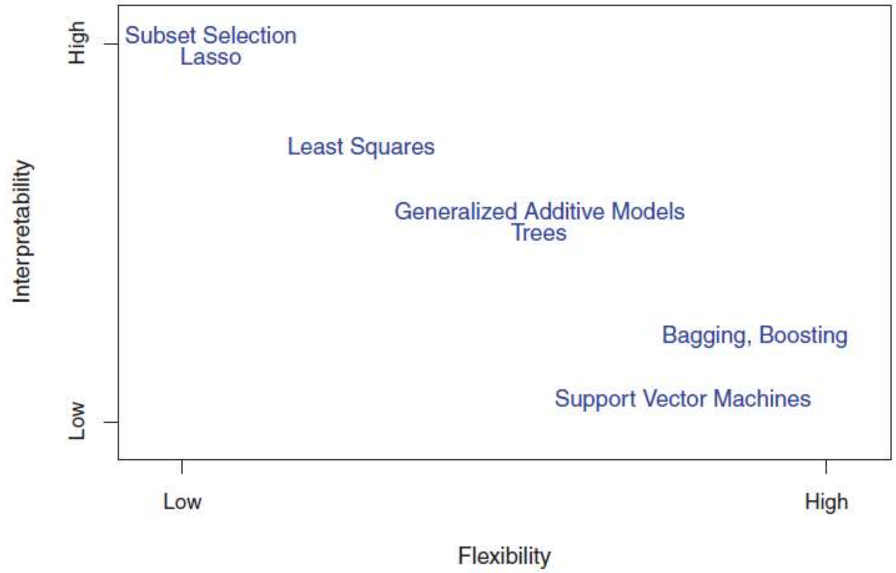


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Source: ISLR

What types of epidemiologic questions/tasks benefit from high flexibility?

What types require more interpretability?

What does interpretability mean in the context of machine learning?

Consider particulars of the data and your question

Are data highly correlated?

Do you anticipate non-linear effects?

Are you interested in interactions between features/exposures?

Are you using predictions as an intermediate result for an epi analysis?

Key Terms in Machine Learning

Knowing and Using Key Terms Facilitates Communication

- Different fields have different vocabularies…Collaboration requires we learn how to speak each other's language.

- Many terms used interchangeably, sometimes incorrectly.

- Sometimes differences in language based on substantive field that is utilizing machine learning. Get comfortable with the language used in your area by reading the literature, attending talks, etc.

Algorithm vs Model

Often used interchangeably

Model: a mathematical representation of a real-world process; given an input, a model will provide an output

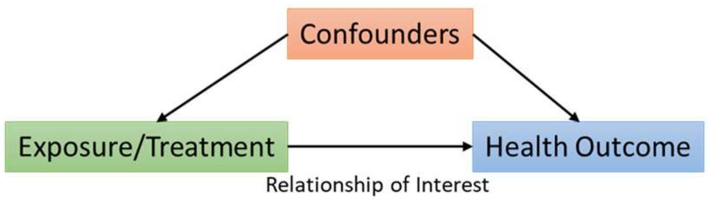
Algorithm: a step-by-step procedure for solving a problem or accomplishing a task

In context of machine learning, algorithms are used to train a model which can then be applied to new, unseen data.

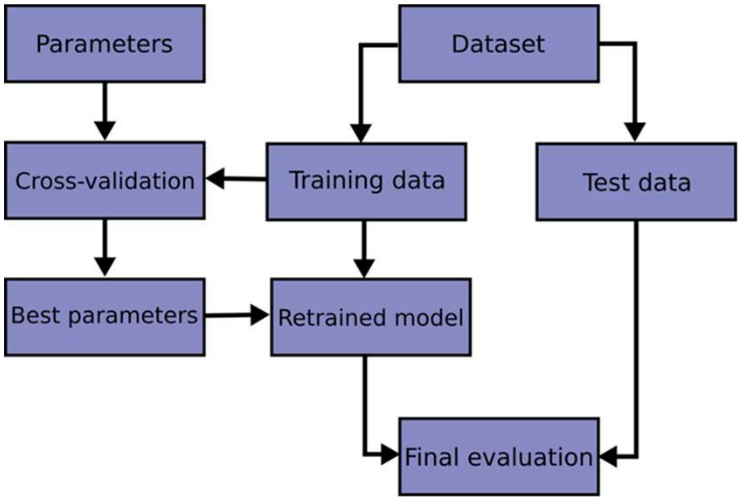
For many machine learning applications, the model ***is*** is the output.

What is different about model building process?

Traditional epidemiologic approach



Traditional Machine learning application



From sci-kit learn

Features and Feature Engineering

Features: Data representing various dimensions of the input observations

- Synonymous with exposures, predictors, inputs, measurements, attributes, independent variables
- Examples: demographics, measurements from an environmental sensors, census data of neighborhood of residence

Feature Engineering: Creating new features from available data to capture latent effects

- Examples include: taking the logarithm of a continuous variable; principal components analysis

Feature Selection: common application of machine learning to select the inputs that are most important for predicting or understanding the outcome of interest.

- Synonymous with variable selection

Feature Reduction: application of reducing the number of features without losing information, typically by trying to construct new features that represent shared information

- Synonymous with dimensionality reduction

Labels and Labeling

Label:
- Synonymous with outcome of interest; the observed or computed value or classification associated with an individual observation
- Examples: breast cancer vs no breast cancer, IQ Score, Frequency of substance use in a 30 day period

Labeling: the process of recording labels (i.e. the classification or value of the outcome) for observations
- Synonymous with obtaining outcome data on participants

Key consideration when discussing supervised vs unsupervised vs semi-supervised methods

How much effort/resources are required to obtained labeled training data?

Descriptions of data and algorithms

Small n, large-p vs Small p, large-n

- n-number of individuals in dataset, p-number of features for each individual
- Refers to shape of dataset (wide vs long) with each having specific set of challenges

Parameters

- a variable, internal to the model, and derived from the data; often saved as part of final model
- Example: β in a regression model

Hyperparameters

- a variable, external to the model and often set by the programmer/analyst; used to estimate model parameters or to optimize the algorithm; can also be called tuning parameter
- Example: number of trees in a random forest

Tuning

- Customization of a model by varying the hyperparameters to determine the values that provide the optimal performance

Tidying

- Structuring data to facilitate analysis
- Similar to data cleaning but has specific rules/guidelines

Descriptions of data and algorithms (2)

Class Balance

- Proportion of cases/non-cases; if outcome is multi-categorical, proportion of cases at each level of outcome
- Data are *imbalanced* if distribution across outcome classes is not equal; can be slight or severe

Majority Class

- The class with the largest proportion of observations

Minority Class

- The class with the smallest proportion of observations

Training, Validation and Testing

Data Partitioning

- Splitting a dataset into random subsets for use in either training, validating or testing the machine learning model
- Use of different subsets
  - Training: used by algorithm to learn the resulting model
  - Validation: used to compare performance of models produced by different algorithms, hyperparameters, …
  - Test/Hold Out: used to obtain final metrics of performance and results of the model

Sample size typically dictates how data are partitioned.
More data used for training than testing in the context of prediction.
Also includes creation of K-folds for cross-validation. Folds are equal sized.

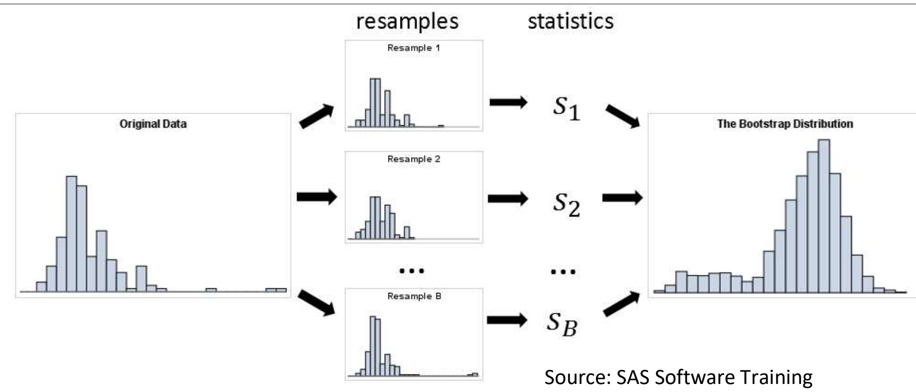
Resampling Methods

Bootstrapping

- Iteratively sampling with replacement
- Used to estimate parameters and draw inferences on a population
- Used in ensemble methods e.g. bagging

Cross-validation

- Validation technique
- Partition data into k non-overlapping subsets
- Estimate model parameters on k-1 subsets (training) then apply model in the held-out subset for evaluation metrics
- Repeat k times
- Similar Approach for Leave-one-out Cross-Validation



Source: SAS Software Training

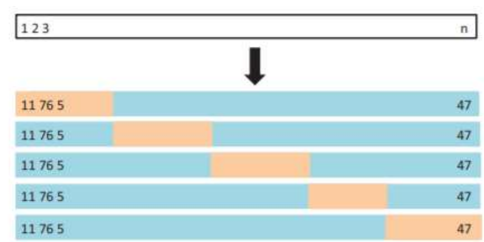


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Source: Introduction to Statistical Learning in R

General Evaluation Terms

Accuracy
- Proportion of results correctly classified
- Reported for classification problems

Precision
- Synonymous with Positive Predictive Value

Recall
- Synonymous with Sensitivity

Mean Square Error
- Reported for regression problems
- Average Squared difference between observed and predicted values

Overfitting
- Model describes random error in individual dataset rather than relationships that are transportable across datasets

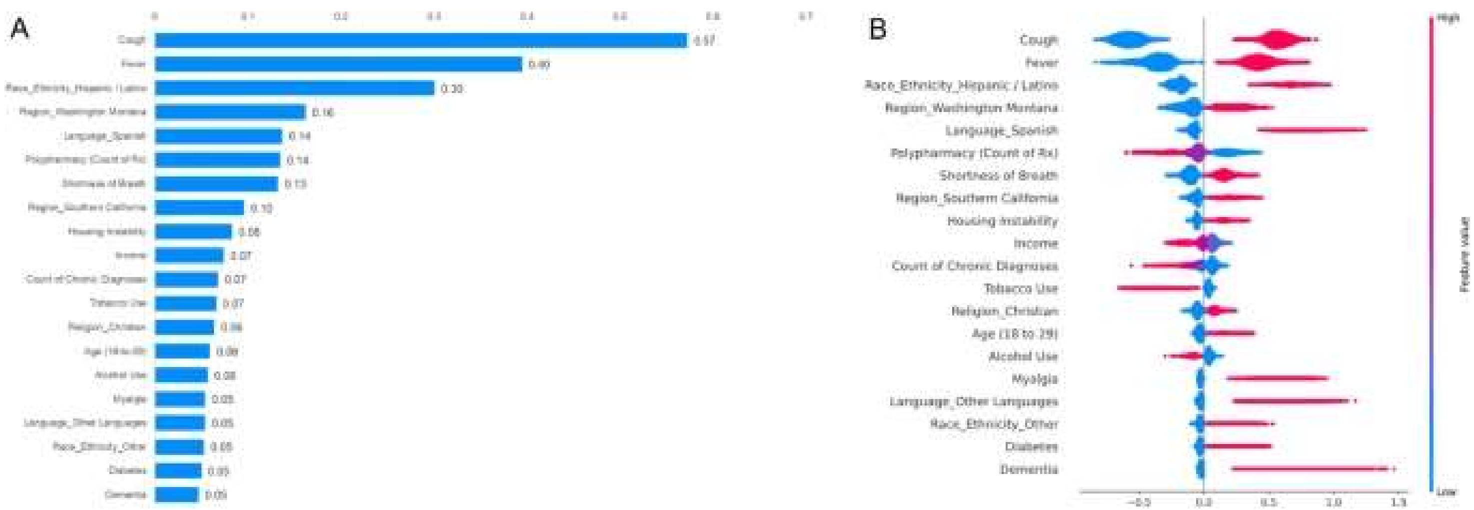
		Predicted	
		+	-
Observed	+	True Pos	False Neg
	-	False Pos	True Neg

Confusion Matrix

Variable Importance Factors

Measure of Individual Variable Contribution to the Overall Prediction

- Calculation varies based on algorithm and/or specific software package
  - E.g. Tree-based approaches like random forest have accuracy-based and node purity based variable importance metrics
- Shapley Values
  - *Marginal contribution of a feature value across all of the possible combinations of features*



Source: Adeoye, E.A., Rozenfeld, Y., Beam, J. *et al. Med Biol Eng Comput* **60,** 2039–2049 (2022). https://doi.org/10.1007/s11517-022-02549-5

Practical Considerations: Software and Resources for Continued Learning

Helpful Textbooks

An Introduction to Statistical Learning

with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Home

About this Book

R Code for Labs

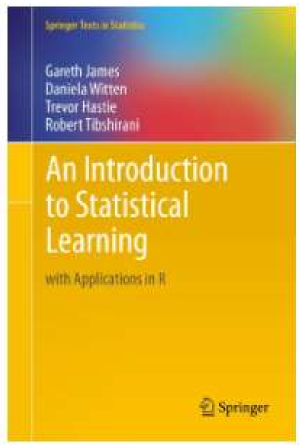
Data Sets and Figures

ISLR Package

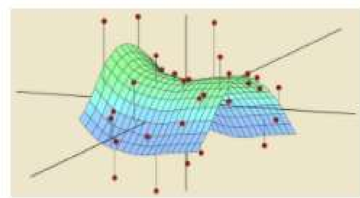
Get the Book

Author Bios

Errata



Download the book PDF
(corrected 7th printing)



Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students, masters students and Ph.D. students in the non-mathematical sciences. The book also contains a number of R labs with detailed explanations on how to implement the various methods in real life settings, and should be a valuable resource for a practicing data scientist.

Multiple Software Options for Analytics

Open-source and Commercial Available

- **R and R Studio**
- Python
- TensorFlow
- SAS Viya
- Stata

Considerations when choosing analytic environment

- Programming Ability, Experience and Enjoyment
- Cost and Availability
- Availability of Support within and external to your substantive field

Introduction to R and R Studio

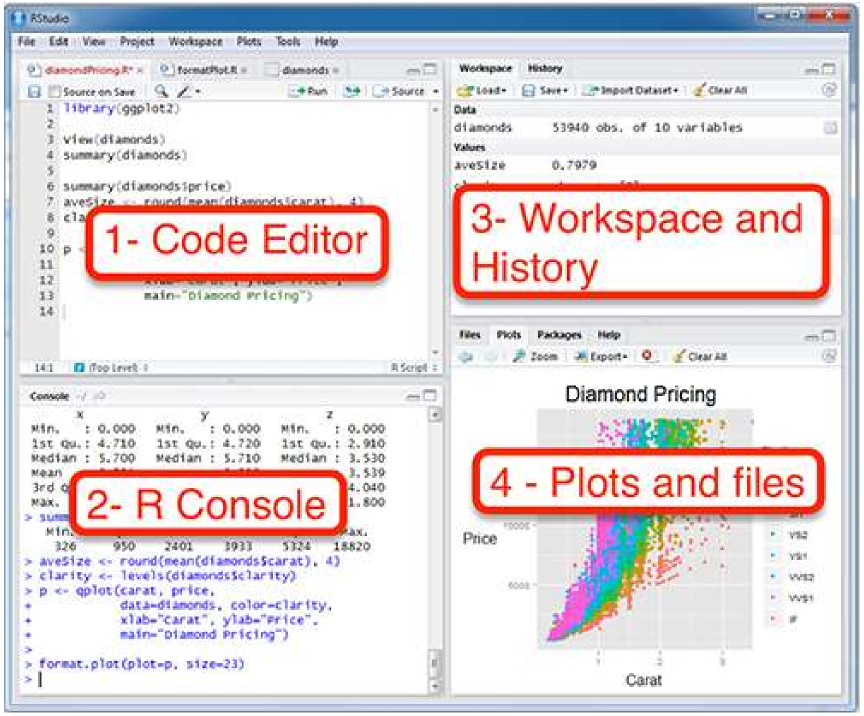
R

Open-source software environment for statistical computing and graphics

Need to download and install individual packages in addition to main environment

RStudio

IDE: integrated development environment



Tutorial on using R Studio
https://datacarpentry.org/R-ecology-lesson/01-intro-to-r.html

R Markdown and R Notebooks

Promotes reproducibility in research

- Ability to save and execute code
- Generates high-quality reports for sharing and distribution in a variety of formats
  - HTML, PDF, MS Word, etc.
- Multiple support documents to facilitate use

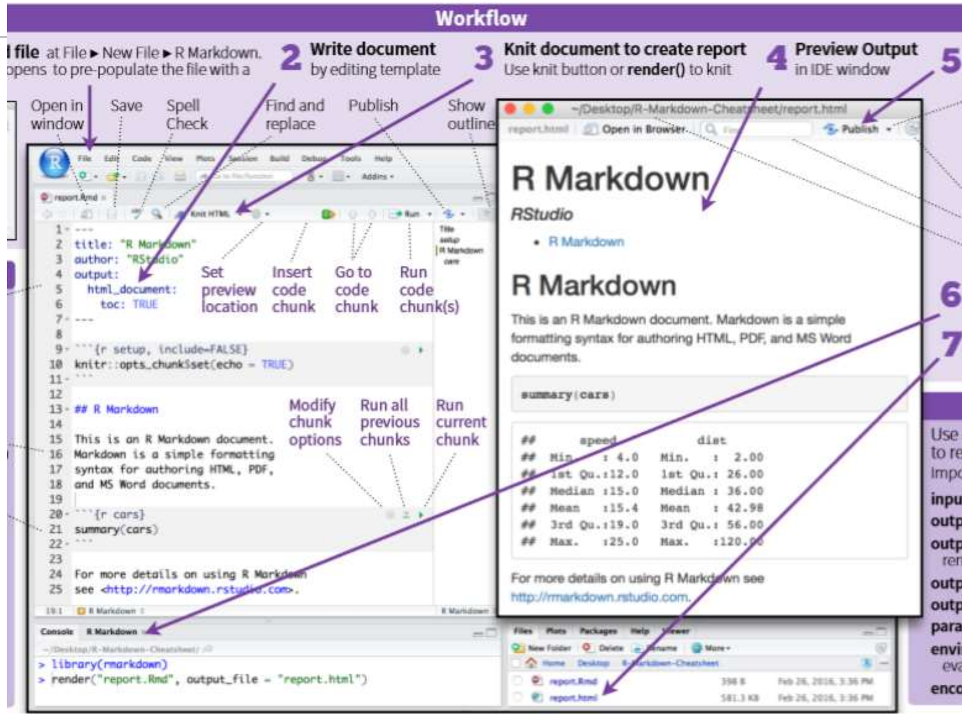
R Markdown CheatSheet

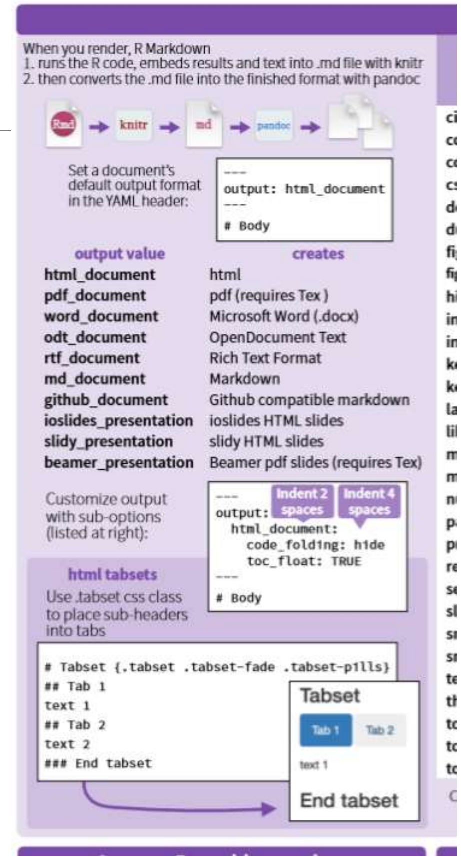
https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdowr

Pandoc's Markdown

Write with syntax on the left to create effect on right (after render)

`Plain text`	Plain text
`End a line with two spaces to start a new paragraph.`	End a line with two spaces to start a new paragraph.
`*italics* and **bold**`	*italics* and **bold**
`` `verbatim code` ``	`verbatim code`
`sub/superscript^2^~2~`	sub/superscript$^2{}_2$
`~~strikethrough~~`	~~strikethrough~~
`escaped: \* \_ \\`	escaped: * _ \
`endash: --, emdash: ---`	endash: –, emdash: —
`equation: $A = \pi*r^{2}$`	equation: $A = \pi * r^2$
`equation block:`	equation block:
`$$E = mc^{2}$$`	$$E = mc^2$$
`> block quote`	block quote
`# Header1 {#anchor}`	Header1
`## Header 2 {#css_id}`	Header 2
`### Header 3 {.css_class}`	Header 3
`#### Header 4`	Header 4
`##### Header 5`	Header 5
`###### Header 6`	Header 6
`<!--Text comment-->`	
`\textbf{Tex ignored in HTML}` `<em>HTML ignored in pdfs</em>`	*HTML ignored in pdfs*
`<http://www.rstudio.com>`	http://www.rstudio.com
`[link](www.rstudio.com)`	link
`Jump to [Header 1](#anchor)`	Jump to Header 1
`image:`	image:
`![Caption](smallorb.png)`	Caption
`* unordered list` `+ sub-item 1` `+ sub-item 2` `- sub-sub-item 1`	• unordered list ◦ sub-item 1 ◦ sub-item 2 ▪ sub-sub-item 1
`* item 2` `Continued (indent 4 spaces)`	• item 2 Continued (indent 4 spaces)
`1. ordered list` `2. item 2` `i) sub-item 1` `A. sub-sub-item 1`	1. ordered list 2. item 2 i. sub-item 1 A. sub-sub-item 1





When you render, R Markdown
1. runs the R code, embeds results and text into .md file with knitr
2. then converts the .md file into the finished format with pandoc

Set a document's default output format in the YAML header:

```
---
output: html_document
---
# Body
```

output value	creates
html_document	html
pdf_document	pdf (requires Tex)
word_document	Microsoft Word (.docx)
odt_document	OpenDocument Text
rtf_document	Rich Text Format
md_document	Markdown
github_document	Github compatible markdown
ioslides_presentation	ioslides HTML slides
slidy_presentation	slidy HTML slides
beamer_presentation	Beamer pdf slides (requires Tex)

Customize output with sub-options (listed at right):

```
---
output:
  html_document:
    code_folding: hide
    toc_float: TRUE
---
# Body
```

Indent 2 spaces Indent 4 spaces

html tabsets

Use .tabset css class to place sub-headers into tabs

```
# Tabset {.tabset .tabset-fade .tabset-pills}
## Tab 1
text 1
## Tab 2
text 2
### End tabset
```

Tabset
Tab 1 Tab 2
text 1
End tabset

Resources for Finding Packages in R

https://cran.r-project.org/web/views/MachineLearning.html

CRAN Task View: Machine Learning & Statistical Learning

Maintainer:	Torsten Hothorn
Contact:	Torsten.Hothorn at R-project.org
Version:	2020-10-28
URL:	https://CRAN.R-project.org/view=MachineLearning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this fi to as machine learning. The packages can be roughly structured into the following topics:

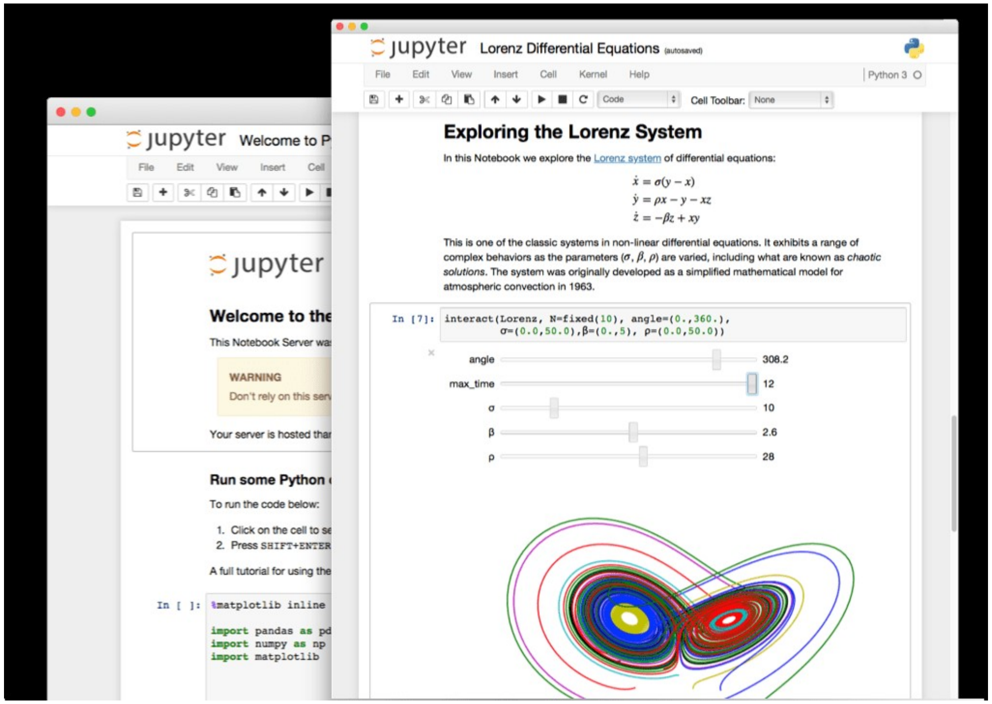
- *Neural Networks and Deep Learning* : Single-hidden-layer neural network are implemented in package nnet (shipped with ba interface to the Stuttgart Neural Network Simulator (SNNS). Packages implementing deep learning flavours of neural netwo neural network, restricted Boltzmann machine, deep belief network, stacked autoencoders), RcppDL (denoising autoencoder, restricted Boltzmann machine, deep belief network) and h2o (feed-forward neural network, deep autoencoders). An interface tensorflow.
- *Recursive Partitioning* : Tree-structured models for regression, classification and survival analysis, following the ideas in the rpart (shipped with base R) and tree. Package rpart is recommended for computing CART-like trees. A rich toolbox of partitic Weka , package RWeka provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The Cubist p (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The C50 pack trees, rule-based models, and boosted versions of these.
  Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in p Function `ctree()` is based on non-parametric conditional inference procedures for testing independence between response an `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the re party and partykit as well.
  Graphical tools for the visualization of trees are available in package maptree

Jupyter Notebooks

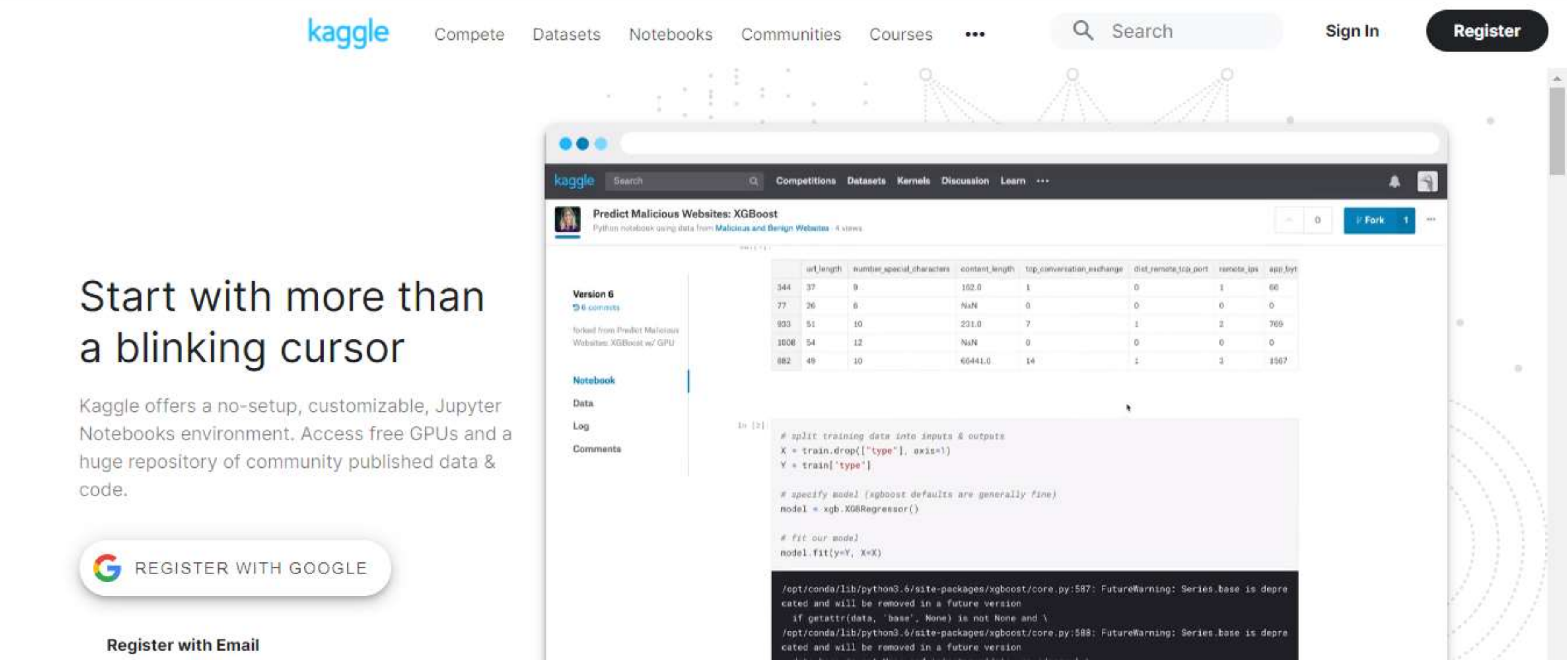
Open-source web application that allows for documents that contain live code, equations, visualizations, etc.

Promotes reproducibility and sharing

Supports over 40 programming languages including Python and R



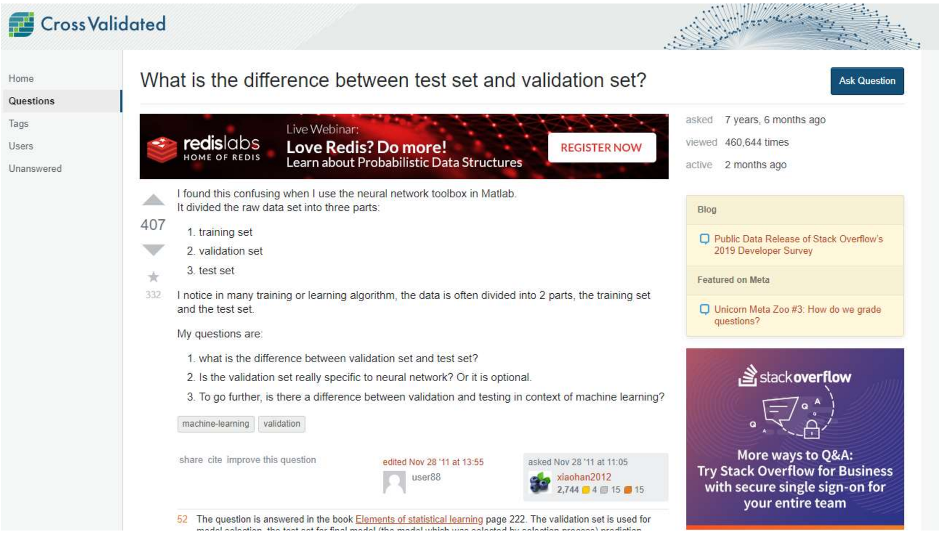
Online Resources



Online Support

Stack Exchange Q&A Communities: collection of “expert” communities that compile Q & A

- Relevant communities: stackoverflow-programming; cross validated: statistics and machine learning
- Community norms on how to post and answer questions
- Typically top answers when googling



https://stackoverflow.com/

https://stats.stackexchange.com/

Recap

- Machine learning is not magic, but it’s not all hype.
- Critical thinking is not optional
- Four types but this workshop will focus on supervised and unsupervised
- Utility of machine learning depends upon the research question and nature of your data
- Need for epidemiologists to have basic understanding of these methods
  - Enhance their own research
  - Critically review others research
- Lots of practical resources, many of them free