

---

# ABCs of Machine Learning for Epidemiology

---

SER WORKSHOP 2021

ERIC LOFGREN AND JEANETTE STINGONE

**Contact us:**

**Eric Lofgren**

✉ [eric.lofgren@wsu.edu](mailto:eric.lofgren@wsu.edu)

🐦 [@GermsAndNumbers](https://twitter.com/GermsAndNumbers)

**Jeanette Stingone**

✉ [j.stingone@columbia.edu](mailto:j.stingone@columbia.edu)

🐦 [@jstingone](https://twitter.com/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	Review of Algorithms and Implementation in R
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



All materials available at <https://github.com/jstingone/mlworkshop2021>

---

# Introduction and General Concepts

---

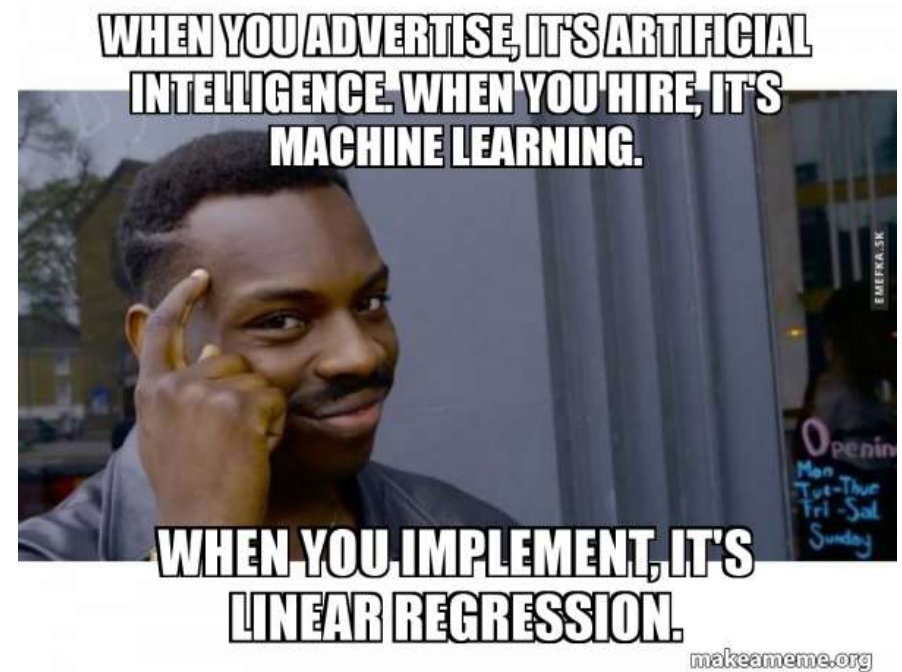
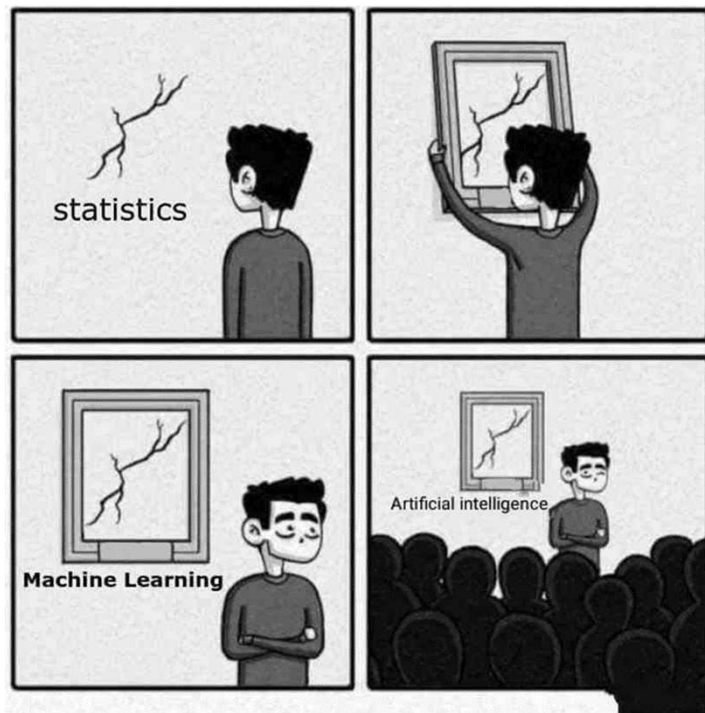
# Machine Learning is not Magic

---



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

**EXPOSURE MODELING**

Community  
Engagement

**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?





# How “machine learning” is defined often depends on who you ask.....

Computational methods **using experience to improve performance or to make accurate predictions.** Here, experience refers to the past information available to the learner, which typically takes the form of electronic data....In all cases, its quality and size are crucial to the success of the predictions made by the learner.  
-Foundations of Machine Learning Mohri, Rostamizadeh, Talwalkar, The MIT Press

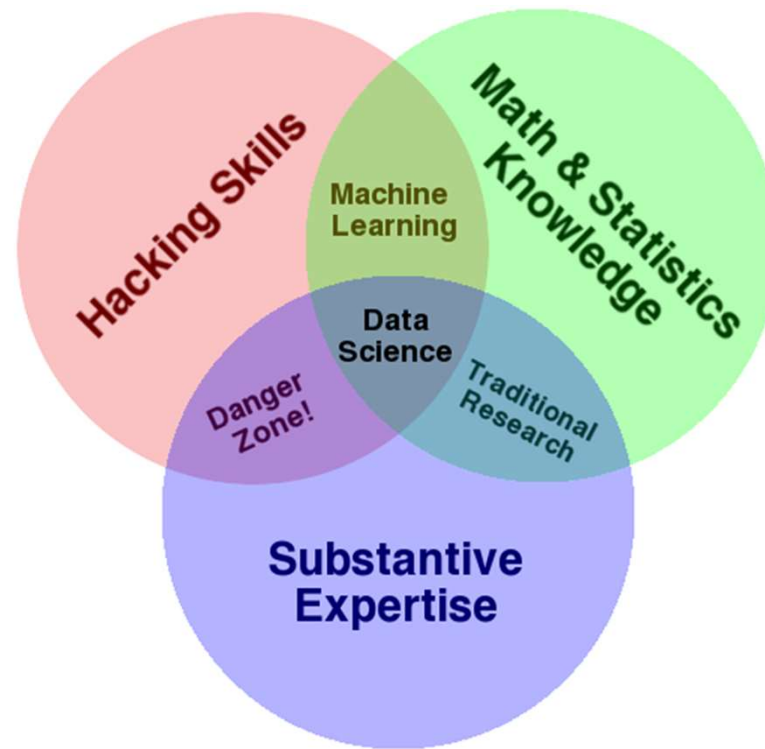
A program or system that builds (trains) **a predictive model from input data.** The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model. Machine learning also refers to the field of study concerned with these programs or systems.  
-Google

..an umbrella term for techniques that **fit models algorithmically by adapting to patterns in data**  
-Mooney and Pejaver, Annual Review of Public Health

Scientific study of algorithms and statistical models that computer systems use to effectively **perform a specific task without using explicit instructions, relying on patterns and inference instead.** -Wikipedia

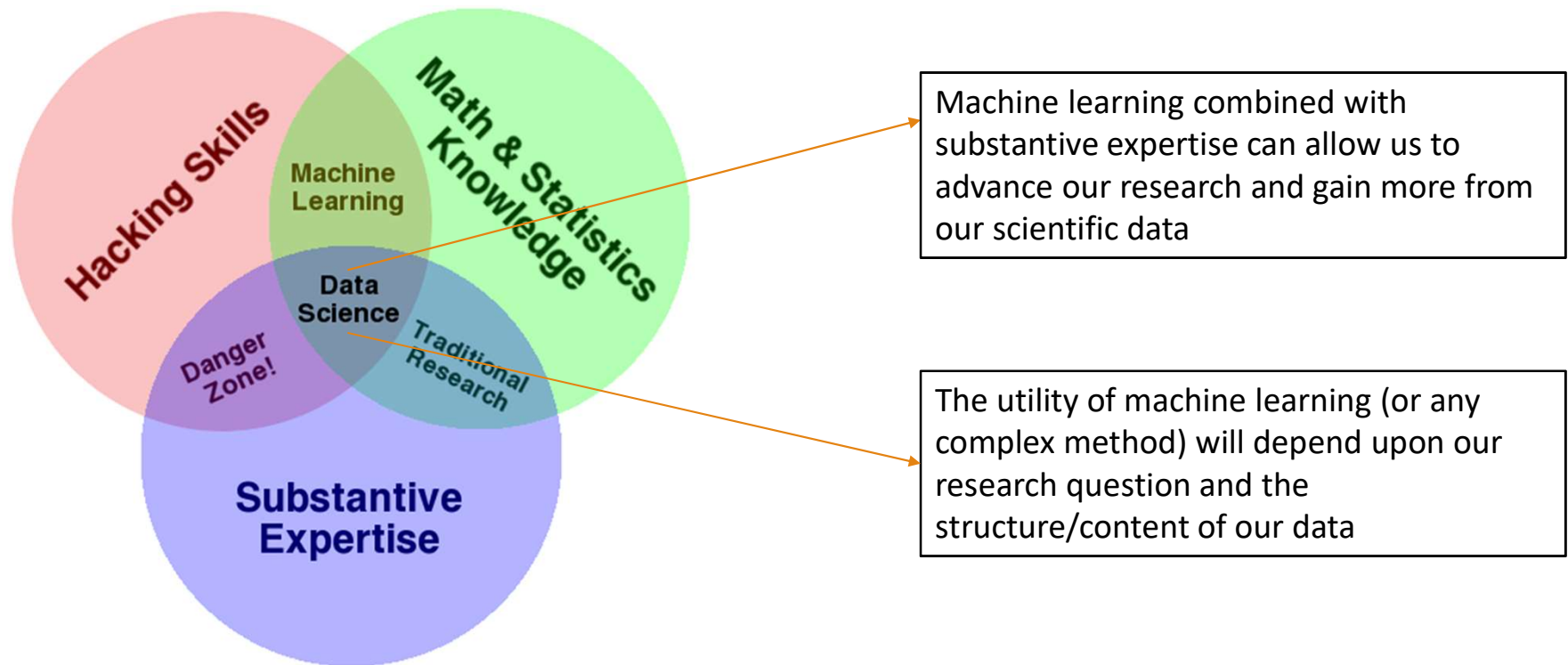
# Machine Learning: Intersection between Computational and Mathematical/Statistical Knowledge

---



# Machine Learning: Intersection between Computational and Mathematical/Statistical Knowledge

---



# 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

# Identifying “Predictors” using machine learning

---

## **Factors Related to Pediatric Unintentional Burns: The Comparison of Logistic Regression and Data Mining Algorithms**

Abbas Aghaei, PhD, Hamid Soori, PhD, Azra Ramezankhani, PhD,  
Yadollah Mehrabi, PhD ✉

*Journal of Burn Care & Research*, irz066, <https://doi.org/10.1093/jbcr/irz066>

**Published:** 27 April 2019

**Excerpt from the Abstract:** The majority of the burn-related variables were related to individuals' social welfare status and their environments. Lessening the effects of these factors could reduce the incidence of pediatric burns.

...Reliant on the assumption that a good predictor is a good explainer.....

# 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

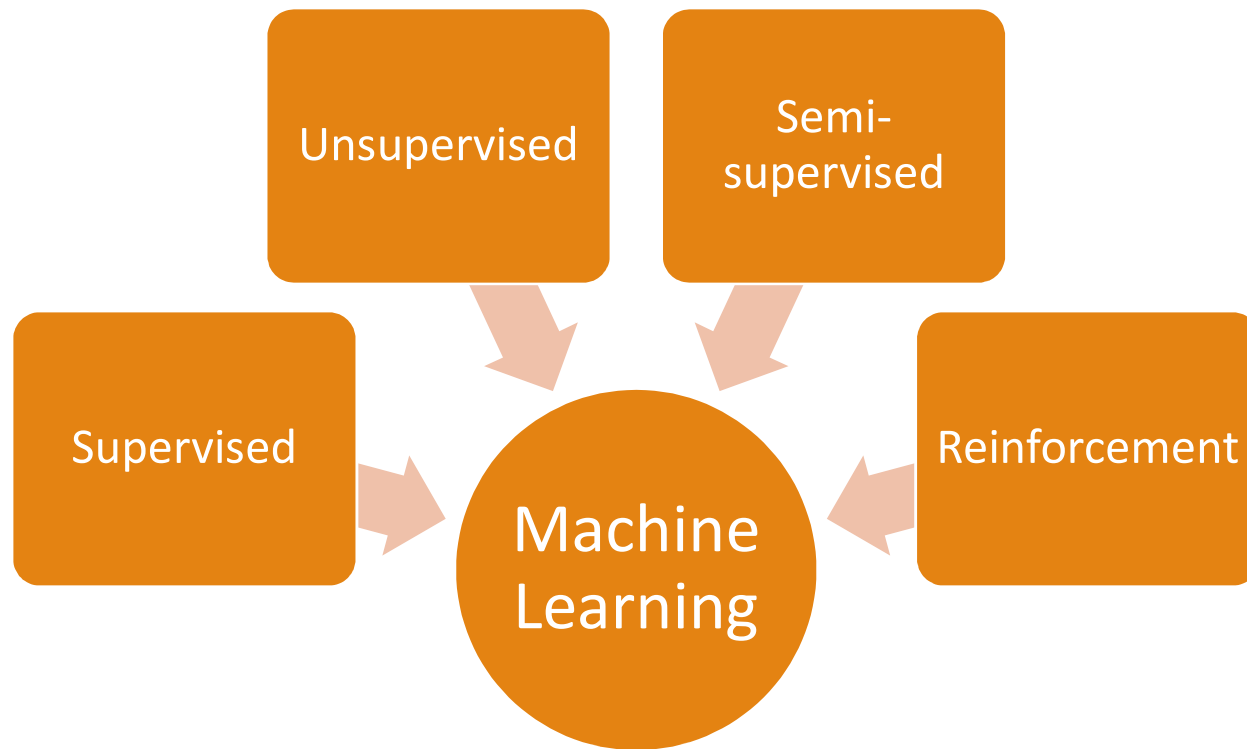
---

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

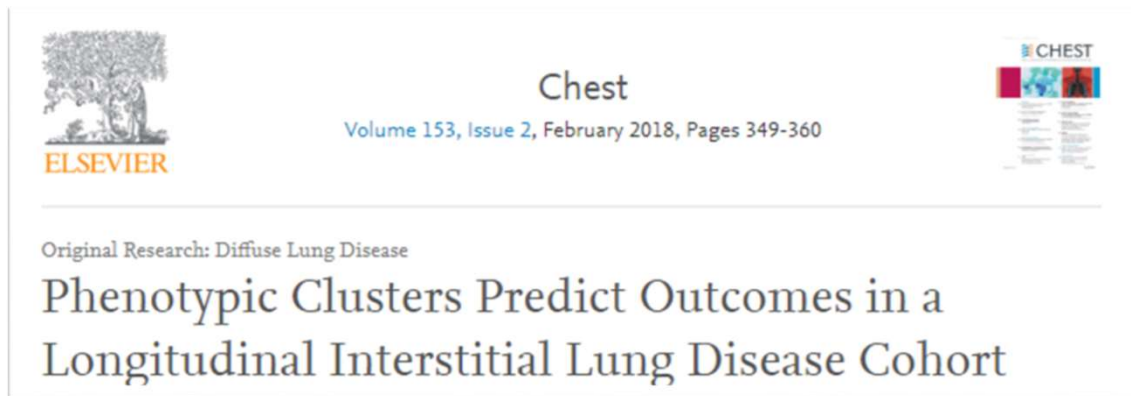
Vol. 127, No. 10 | Research

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

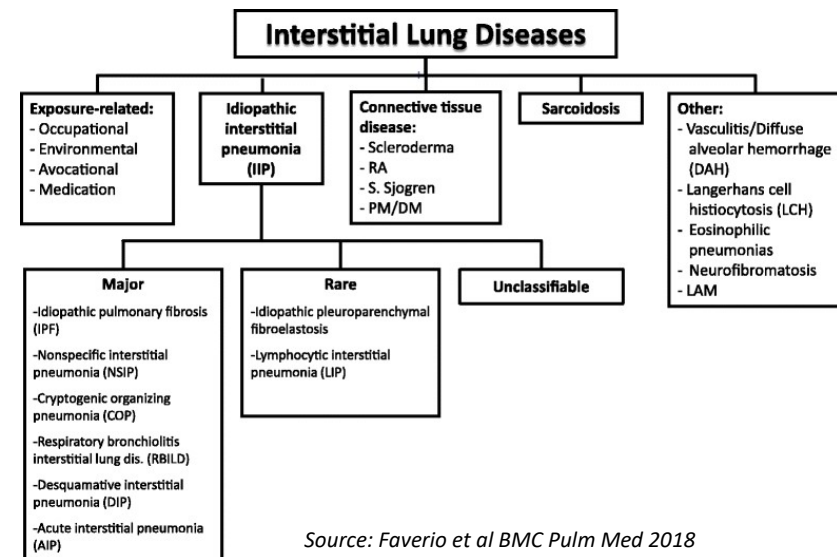
Alexandra J. White , Joshua P. Keller, Shanshan Zhao, Rachel Carroll, Joel D. Kaufman, and Dale P. Sandler

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

# Example: Phenotypic Subtypes



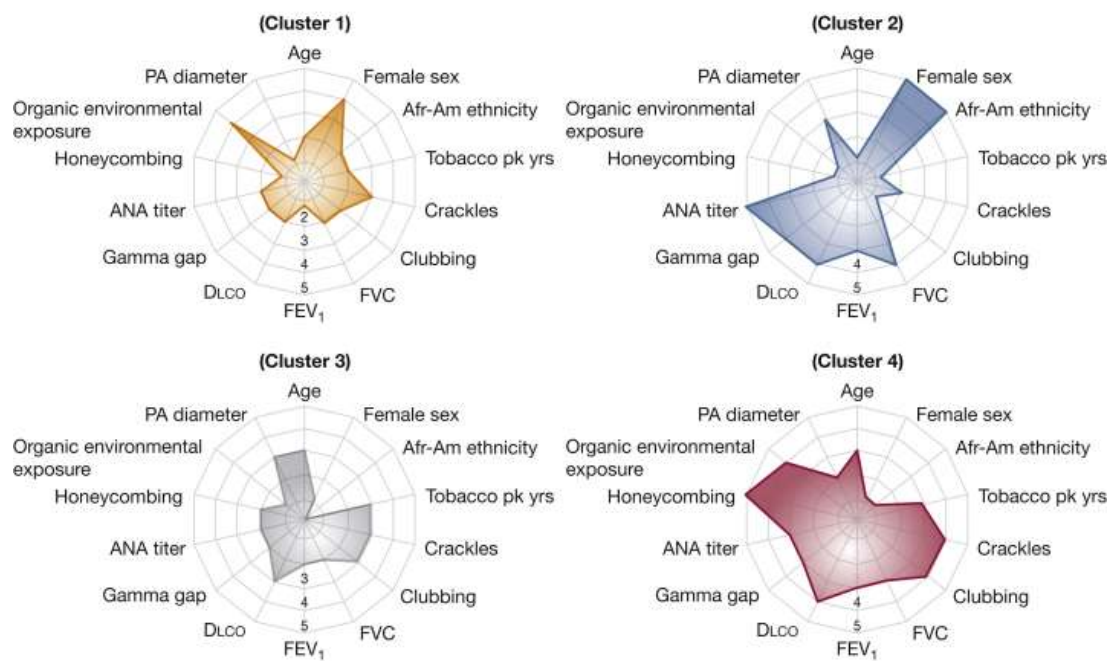
Goal: “Identify distinct clinical phenotypes in heterogeneous diseases”



Source: Faverio et al BMC Pulm Med 2018

# Method: Partitioning around Medoids (PAM)

Radial Plot of Phenotypic Clusters in ILD

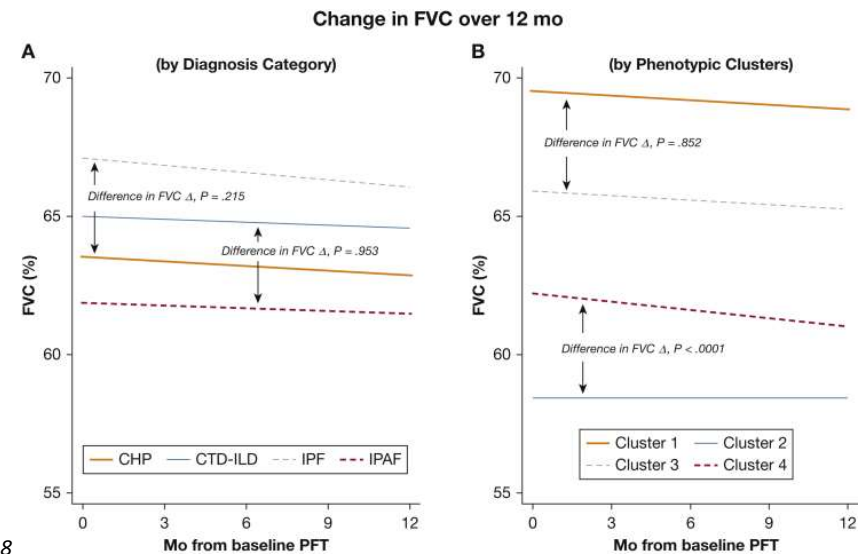


Source: Adegunsoye et al Chest 2018

## Why PAM?

Similar to K-means, but more robust to outliers

Relies on median distances rather than means



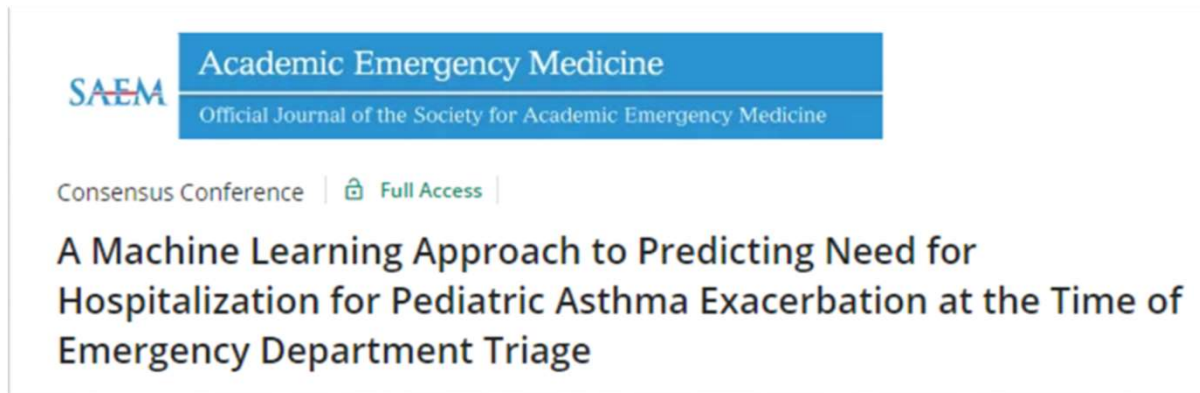
# Supervised

---

**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



# 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

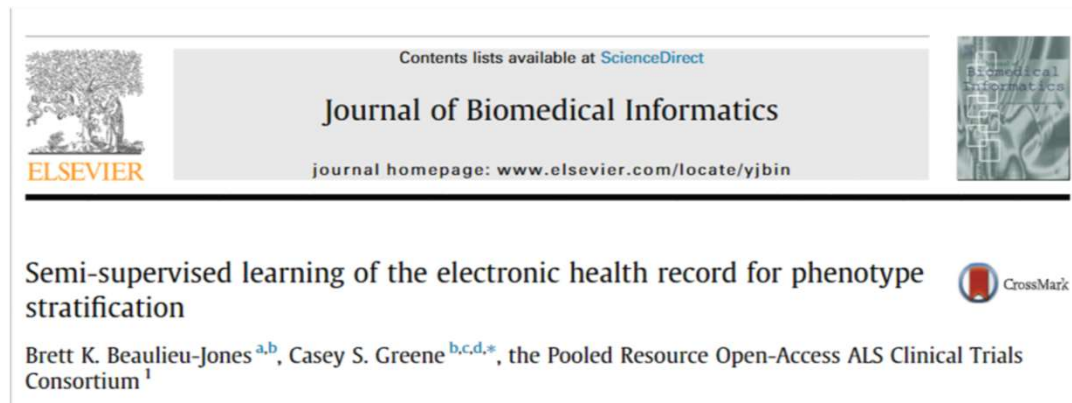
- ❑ Generate propensity scores or IPTW to improve exchangeability
  - ❑ “Predicting” the Counterfactual
- 

# Semi-supervised

**Context:** for each observation of the inputs (predictor/exposure/independent variables), only a (typically) small subset has an associated output (response variable); combination of labeled and unlabeled data

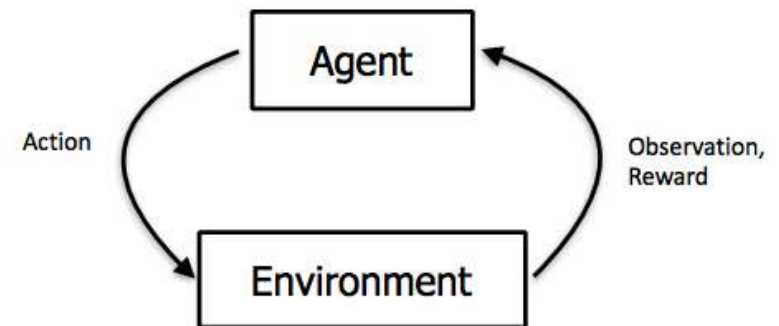
Algorithm learns from subset of labeled data and then applies what it has learned to the unlabeled data.

**Addresses:** Useful when it is overly human or resource intensive to obtain sufficient amounts of labeled data to train a supervised algorithm



# Reinforcement

Algorithm learns how to act in a given environment through maximization of reward; needs to anticipate future rewards from short-term actions



---

# 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.

# 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 obtain 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:  $\beta$  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

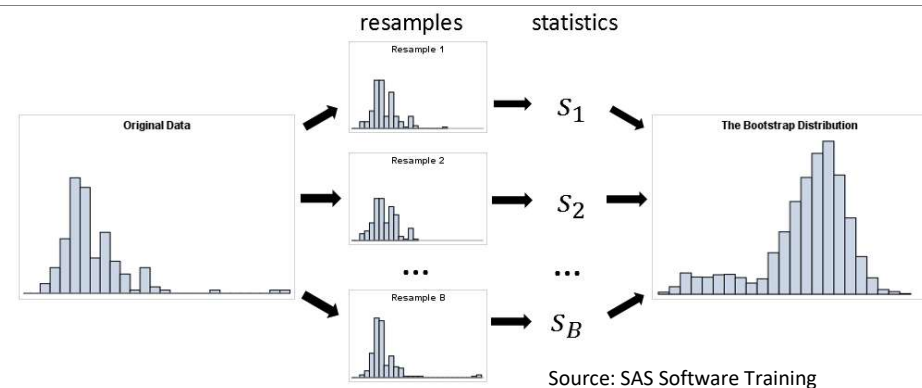
- Splitting a dataset into random subsets for use in either training, validating or testing the machine learning model
- Use of different subsets
  - Training: used by algorithm to learn the resulting model
  - Validation: used to compare performance of models produced by different algorithms, hyperparameters, ...
  - 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

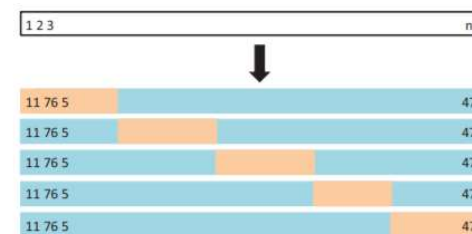
## Bootstrapping

- 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



**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

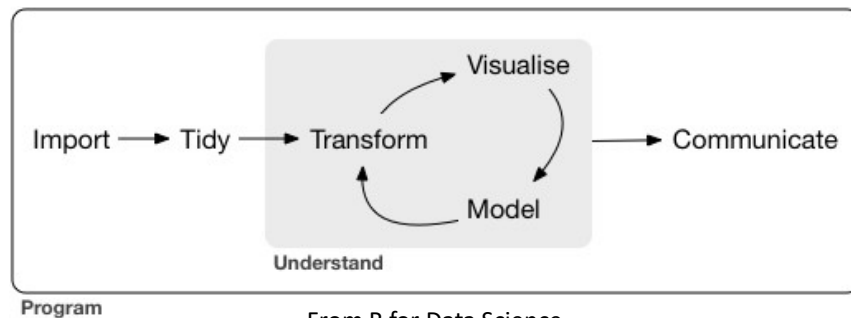
	Predicted		
		+	-
Observed	+	True Pos	False Neg
	-	False Pos	True Neg

*Confusion Matrix*

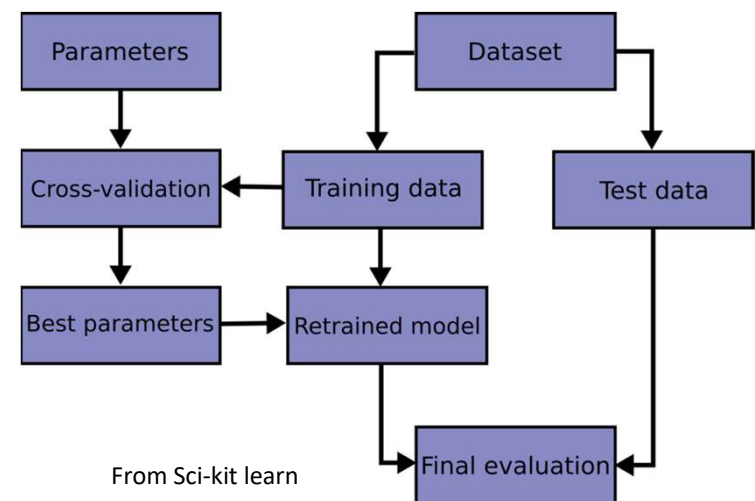
# Pipelines

Ordered set of tasks to accomplish a specific task or goal; Visual study protocol

Specific definition varies by field and their perspective on data analytics



From R for Data Science  
<https://r4ds.had.co.nz/>



From Sci-kit learn

---

## Practical Considerations: Software and Resources for Continued Learning

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# 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](#)

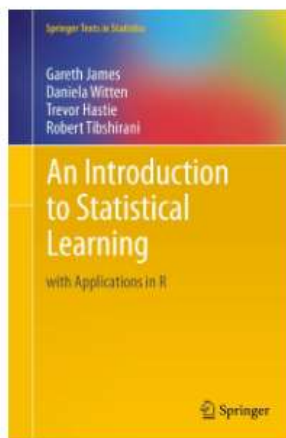
[Data Sets and Figures](#)

[ISLR Package](#)

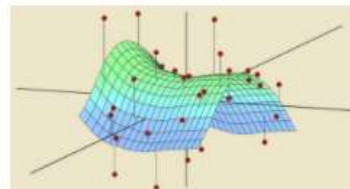
[Get the Book](#)

[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

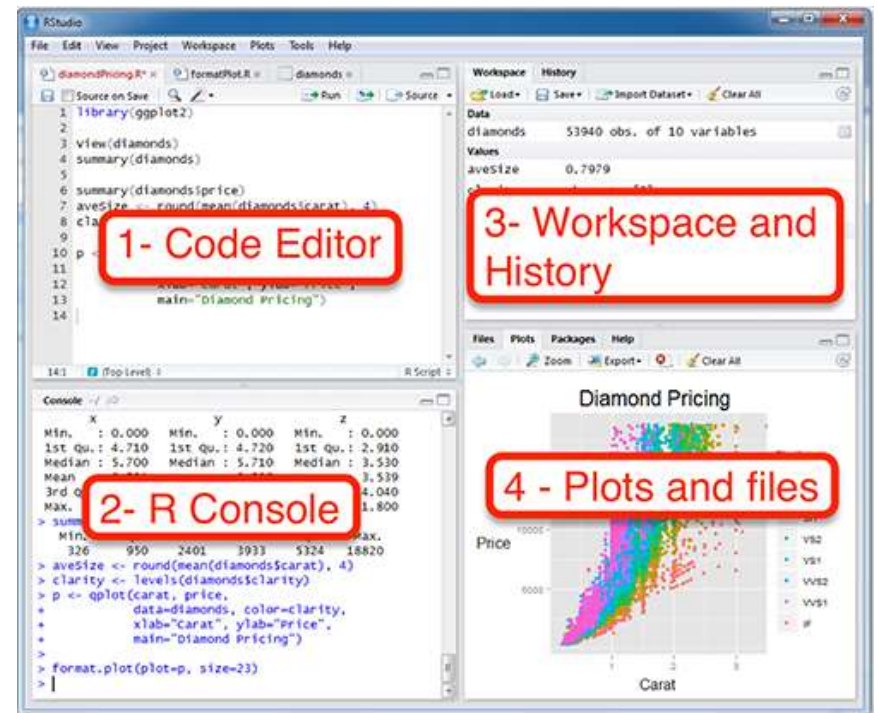
Need to download and install individual packages in addition to main environment

## 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/rmarkdown>

## Pandoc's Markdown

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

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

**Plain text**  
End a line with two spaces to start a new paragraph.  
*italics* and **bold**  
`verbatim code`  
`sub/superscript^2`  
`strikethrough`  
escaped: `\_ \\`  
endash: `--`, emdash: `---`  
equation:  $A = \pi * r^2$   
equation block:  
$$E = mc^2$$
  
block quote  
**Header1**  
**Header 2**  
**Header 3**  
**Header 4**  
**Header 5**  
**Header 6**  
Header 6  
HTML ignored in pdfs  
<http://www.rstudio.com>  
link  
Jump to Header 1  
image:  
Caption  
+ unordered list  
+ sub-item 1  
+ sub-item 2  
+ sub-sub-item 1  
+ item 2  
Continued (indent 4 spaces)  
1. ordered list  
2. item 2  
i. sub-item 1  
A. sub-sub-item 1  
+ A. first sub-item in continuation

## Workflow

1. file at File > New File > R Markdown. opens to pre-populate the file with a
2. Write document by editing template
3. Knit document to create report Use knit button or render() to knit
4. Preview Output in IDE window
5. Publish
6. Open in Browser
7. Show outline

Open in window Save Spell Check Find and replace Publish Show outline

Set preview location Insert code chunk Go to code chunk Run code chunk(s) Modify chunk options Run all previous chunks Run current chunk

R Markdown  
RStudio  
• R Markdown

R Markdown  
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.

```
summary(cars)
```

```
##      speed      dist
## Min       4.0   Min.    2.00
## 1st Qu.:12.0   1st Qu.: 26.00
## Median :15.0   Median : 36.00
## Mean   :15.4   Mean   : 42.98
## 3rd Qu.:19.0   3rd Qu.: 56.00
## Max.   :25.0   Max.   :120.00
```

For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

Console  
R Markdown  
~/Desktop/R-Markdown-CheatSheet/jd  
> library(rmarkdown)  
> render("report.Rmd", output\_file = "report.html")

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
---
```

output value

html\_document

pdf\_document

word\_document

odt\_document

rtf\_document

md\_document

github\_document

ioslides\_presentation

slidy\_presentation

beamer\_presentation

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

output:

html\_document:

code\_folding: hide

toc\_float: TRUE

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

ci  
cc  
cc  
cc  
d  
fi  
fi  
hi  
in  
in  
ko  
ko  
la  
li  
m  
m  
n  
n  
p  
p  
p  
re  
sa  
sl  
sl  
te  
te  
te  
C



# 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 field is known as machine learning. The packages can be roughly structured into the following topics:

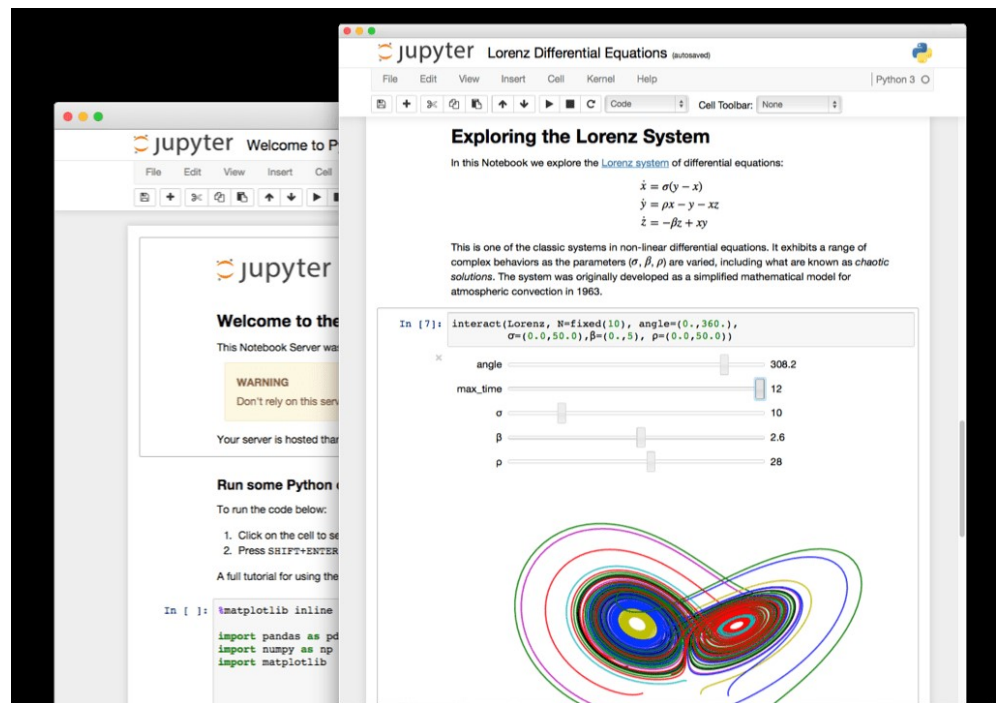
- *Neural Networks and Deep Learning* : Single-hidden-layer neural network are implemented in package [nnet](#) (shipped with base R) and [neuralnet](#) (interface to the Stuttgart Neural Network Simulator (SNNS)). Packages implementing deep learning flavours of neural network (restricted Boltzmann machine, deep belief network, stacked autoencoders), [RcppDL](#) (denoising autoencoder, restricted Boltzmann machine, deep belief network) and [h2o](#) (feed-forward neural network, deep autoencoders). An interface to [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 partitioning algorithms is available in [Weka](#), package [RWeka](#) provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The [Cubist](#) package (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The [C50](#) package implements trees, rule-based models, and boosted versions of these. Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in package [party](#). Function `ctree()` is based on non-parametric conditional inference procedures for testing independence between response and predictors. `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the recursive partitioning algorithms are available in [party](#) and [partykit](#) as well. Graphical tools for the visualization of trees are available in package [mantree](#).

# 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

## Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email

kaggle

Compete

Datasets

Notebooks

Communities

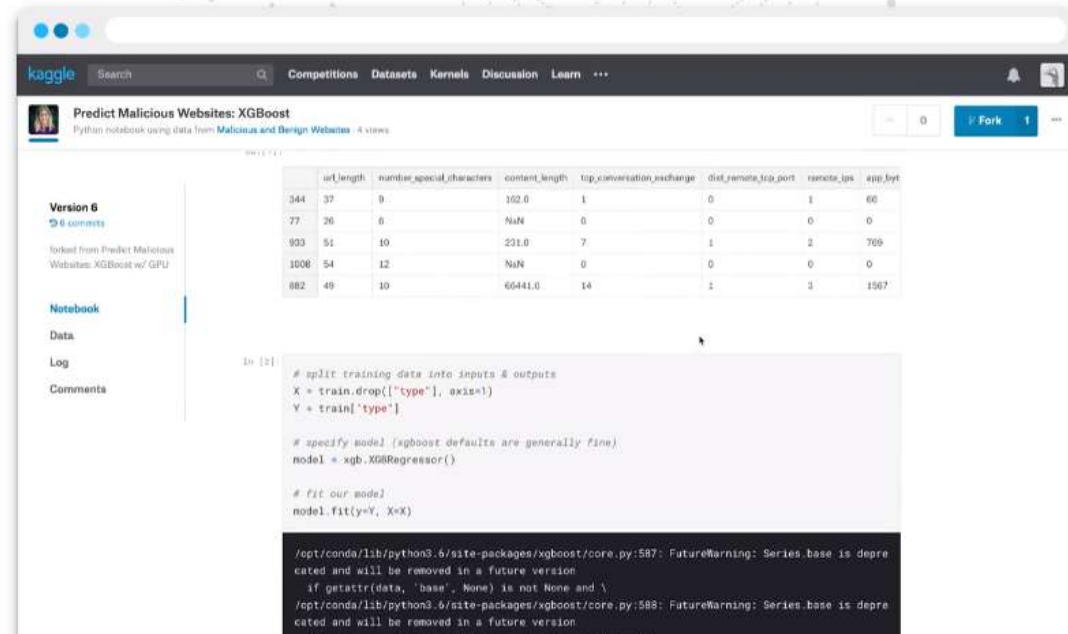
Courses

...

Search

Sign In

Register



The screenshot shows a Kaggle notebook interface. At the top, there's a navigation bar with links to Competitions, Datasets, Kernels, Discussion, and Learn. Below this, the notebook title 'Predict Malicious Websites: XGBoost' is displayed, along with a description 'Python notebook using data from Malicious and Benign Websites' and a 'Fork' button. The notebook content is divided into sections: Version 6 (with 6 commits), Notebook, Data, Log, and Comments. The 'Data' section shows a table with columns: url\_length, number\_special\_characters, content\_length, top\_conversation\_exchange, dist\_remote\_tcp\_port, remote\_ips, and app\_byte. The 'Log' section shows a Jupyter code cell with the following code:

```
In [2]: # split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor()

# fit our model
model.fit(y=Y, X=X)
```

Below the code, there are two FutureWarning messages from XGBoost:

```
/opt/conda/lib/python3.6/site-packages/xgboost/core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version.
if getattr(data, 'base', None) is not None and \
/opt/conda/lib/python3.6/site-packages/xgboost/core.py:588: FutureWarning: Series.base is deprecated and will be removed in a future version
```

# 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

The screenshot shows a question on the Cross Validated site. The question title is "What is the difference between test set and validation set?". The question body states: "I found this confusing when I use the neural network toolbox in Matlab. It divided the raw data set into three parts: 1. training set, 2. validation set, 3. test set." The question is tagged with "machine-learning" and "validation". The question was asked 7 years, 6 months ago and has been viewed 460,644 times. The top answer, by user88, explains that the validation set is used for model selection and the test set is used for model evaluation. The answer is highly upvoted and includes a reference to the book "Elements of statistical learning".

<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