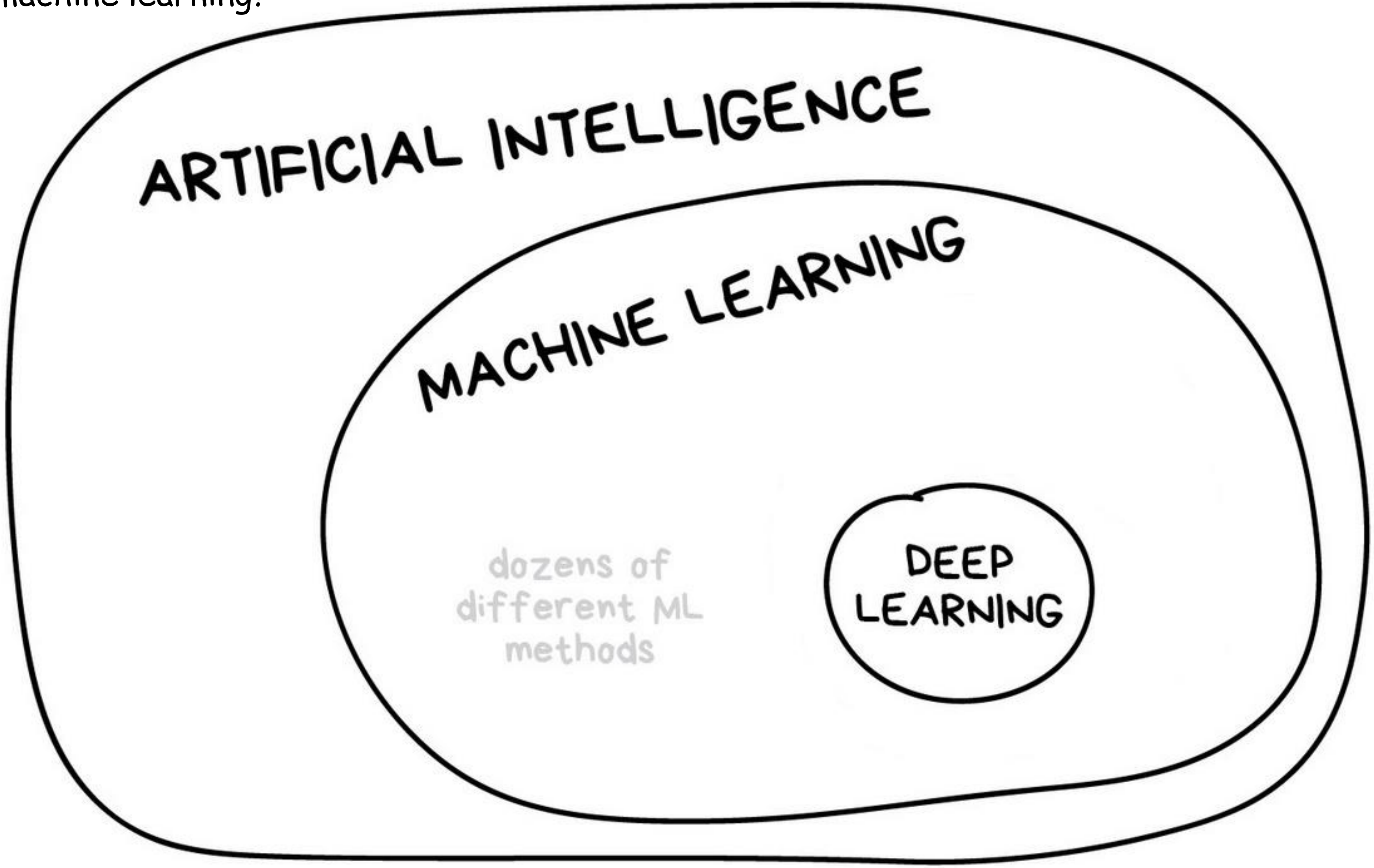
MACHINE LEARNING

What is machine learning?

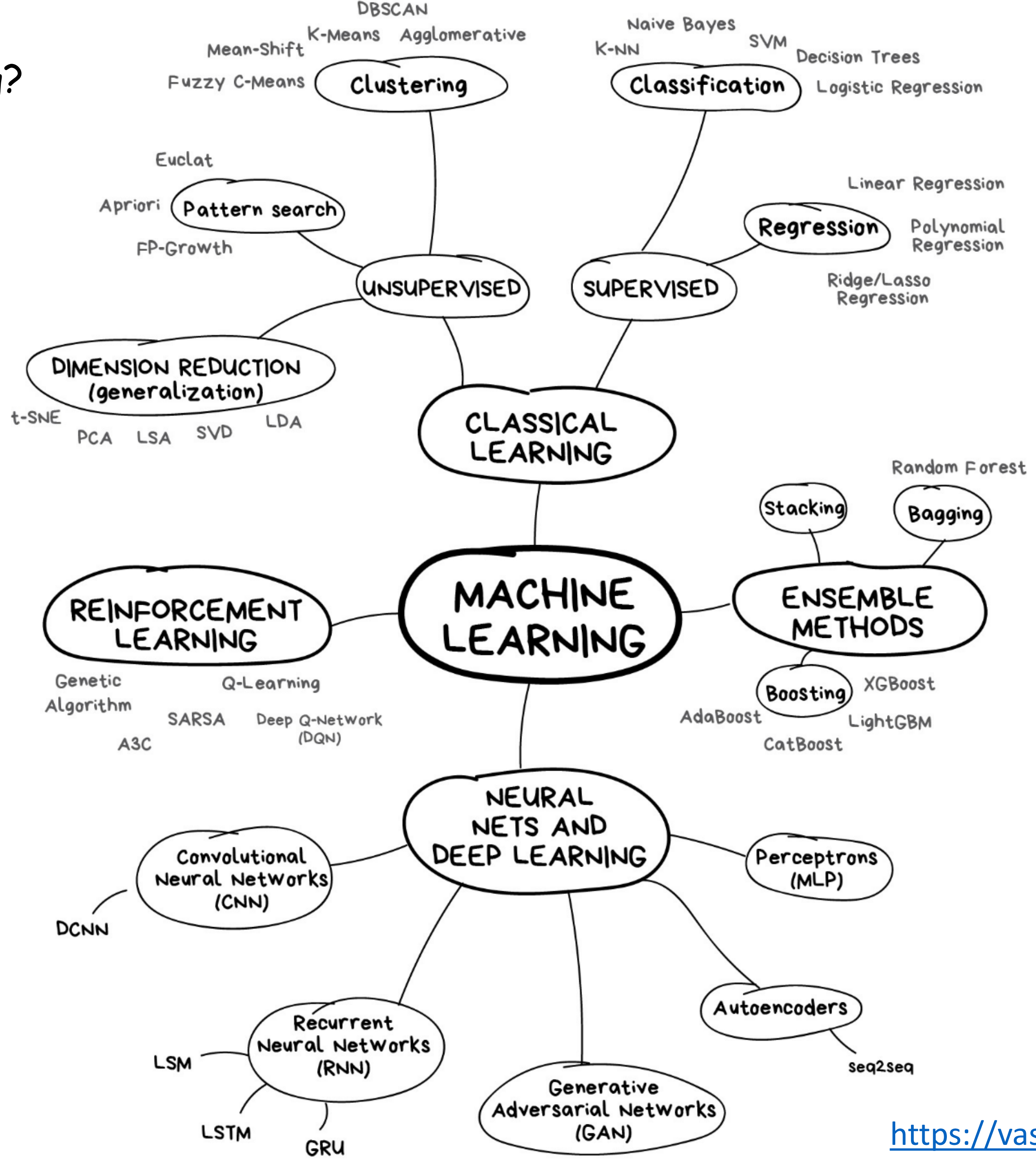
https://xkcd.com/1838/

What is machine learning?



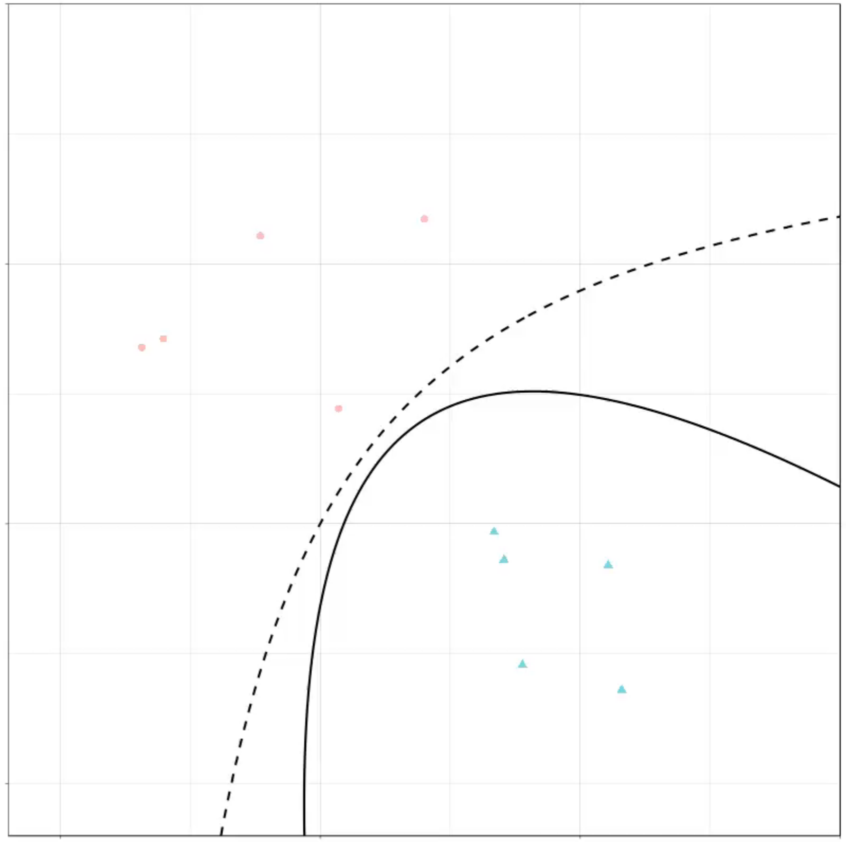
https://vas3k.com/blog/machine_learning/

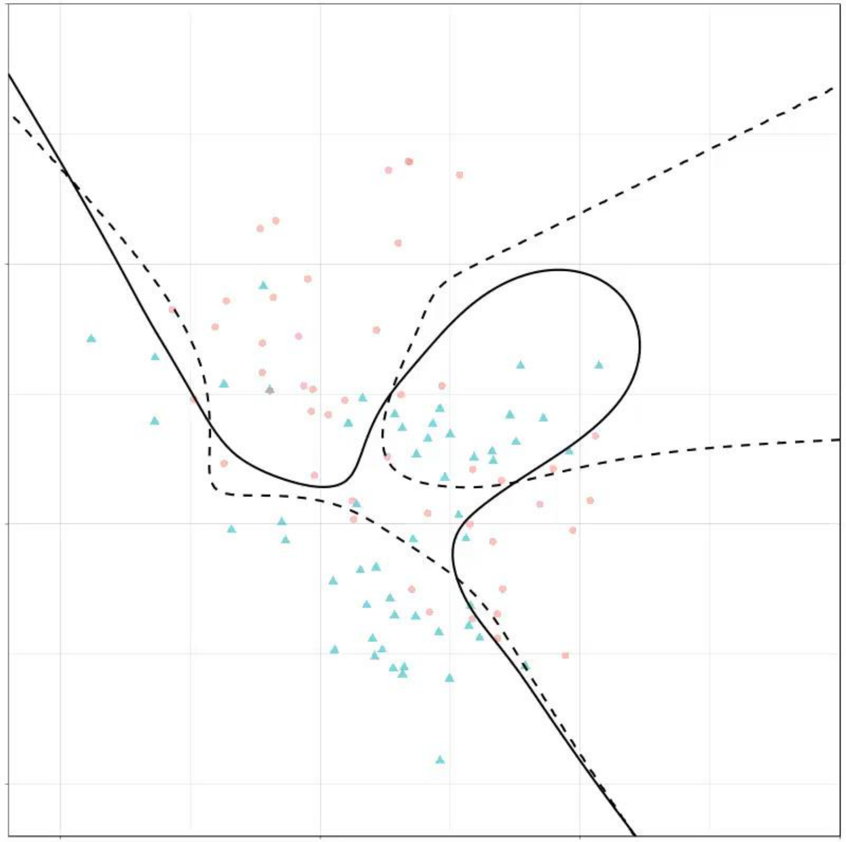
What is machine learning?



https://vas3k.com/blog/machine_learning/

But what **is** machine learning?

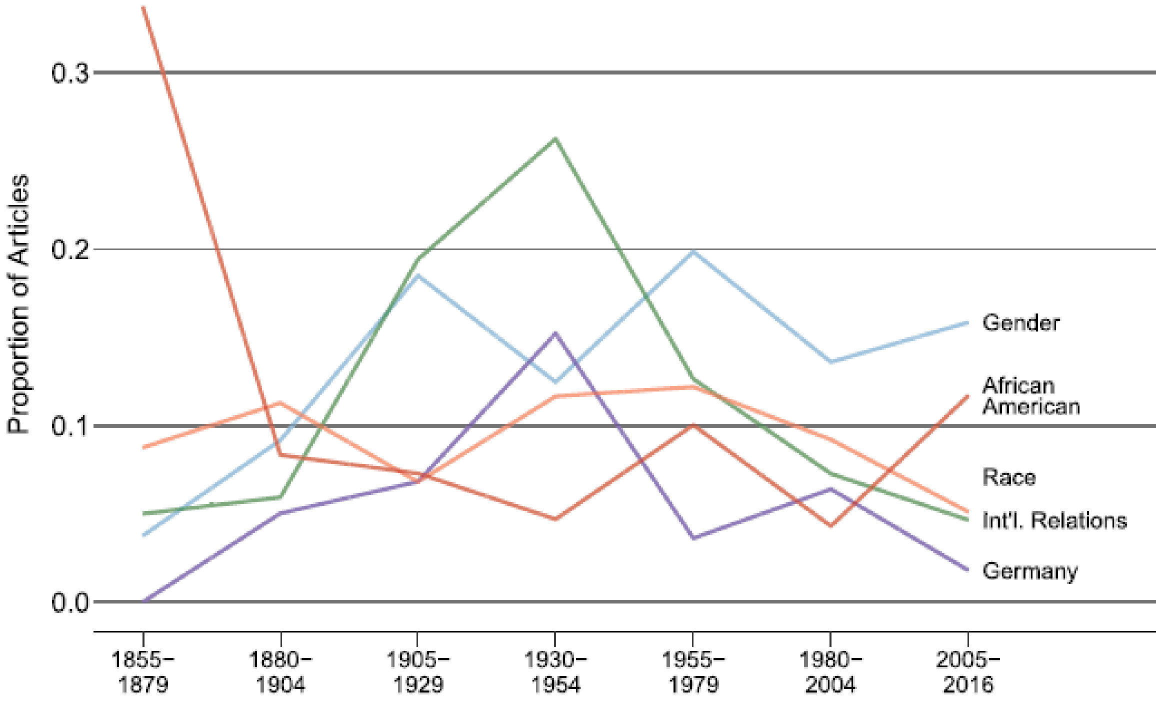




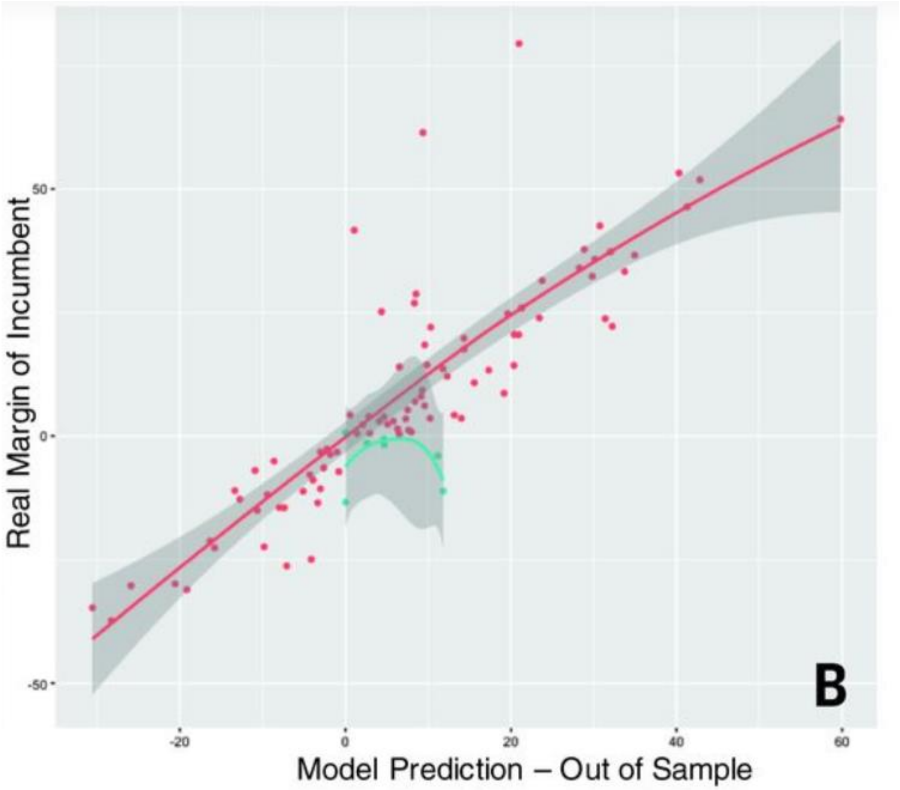
“All the impressive achievements of deep learning [and machine learning] amount to just curve fitting.” – Judea Pearl

https://twitter.com/ryanpholbrook/status/1218526189410824193
https://mindmatters.ai/2020/12/ai-still-just-curve-fitting-not-finding-a-theory-of-everything/

What can we do with this in social science?

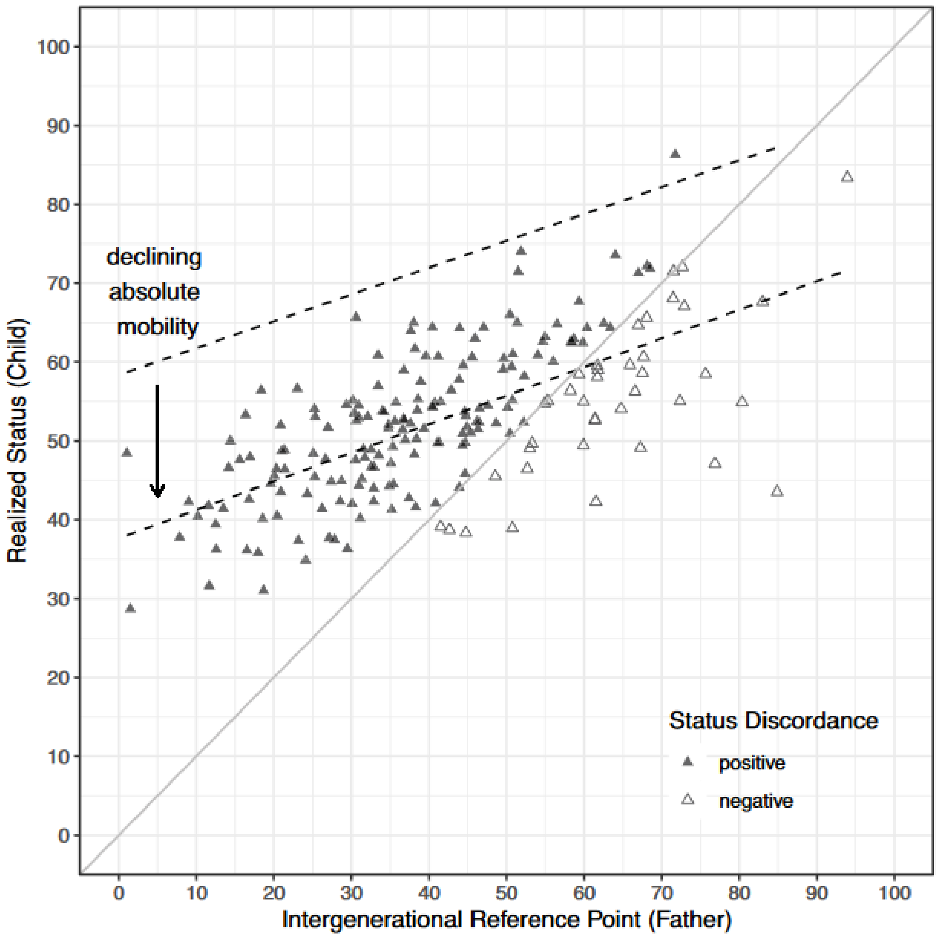


Rodman (2020)

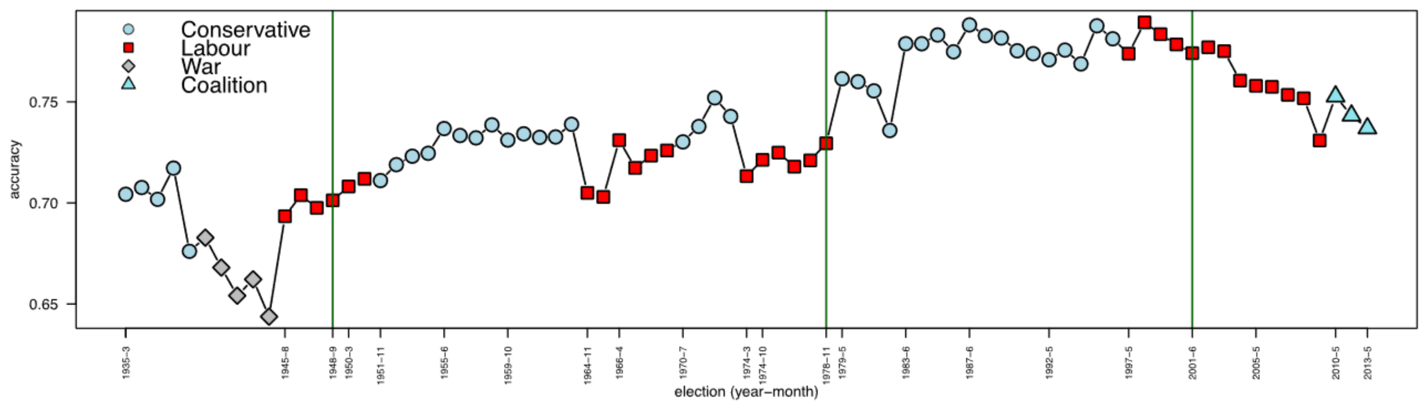


Kennedy et al. (2017)

What can we do with this in social science?



Kurer and van Staalduinen (2022)



Peterson and Spirling (2018)

What to expect

Day One:

What is machine learning?

Machine learning in social science

Basic machine learning vocabulary

The tidymodels package

Day Two:

Supervised machine learning algorithms:

Naive bayes

Support vector machine

Random forest

Day Three:

Neural networks:

Frank Rosenblatt’s perceptron

Multi-layer perceptron

Automated text analysis with neural networks:

Word2Vec

Word vectors

Bring your own data!

What to expect

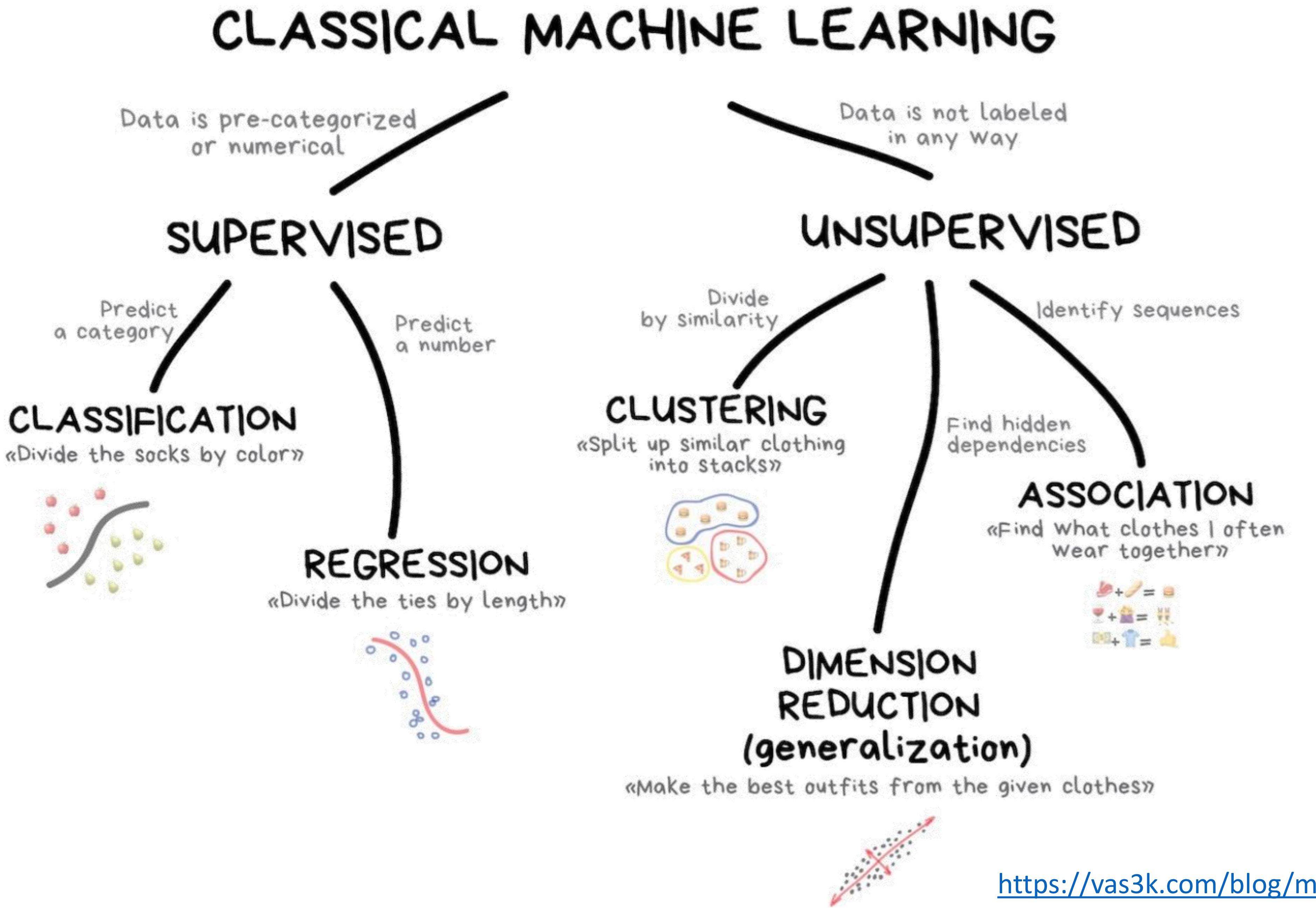
Understanding of **basic principles and vocabulary** of machine learning

Theoretical idea of most common **machine learning algorithms**

Basic command of the **tidymodels** package in R

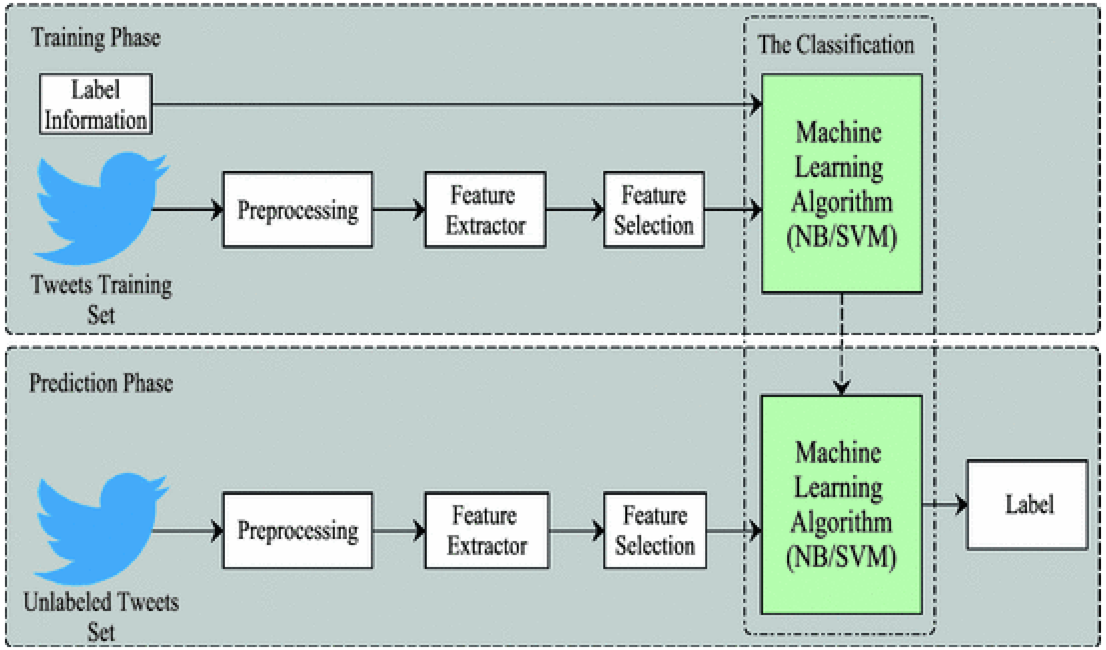
Ability to **understand and use tidymodels code**

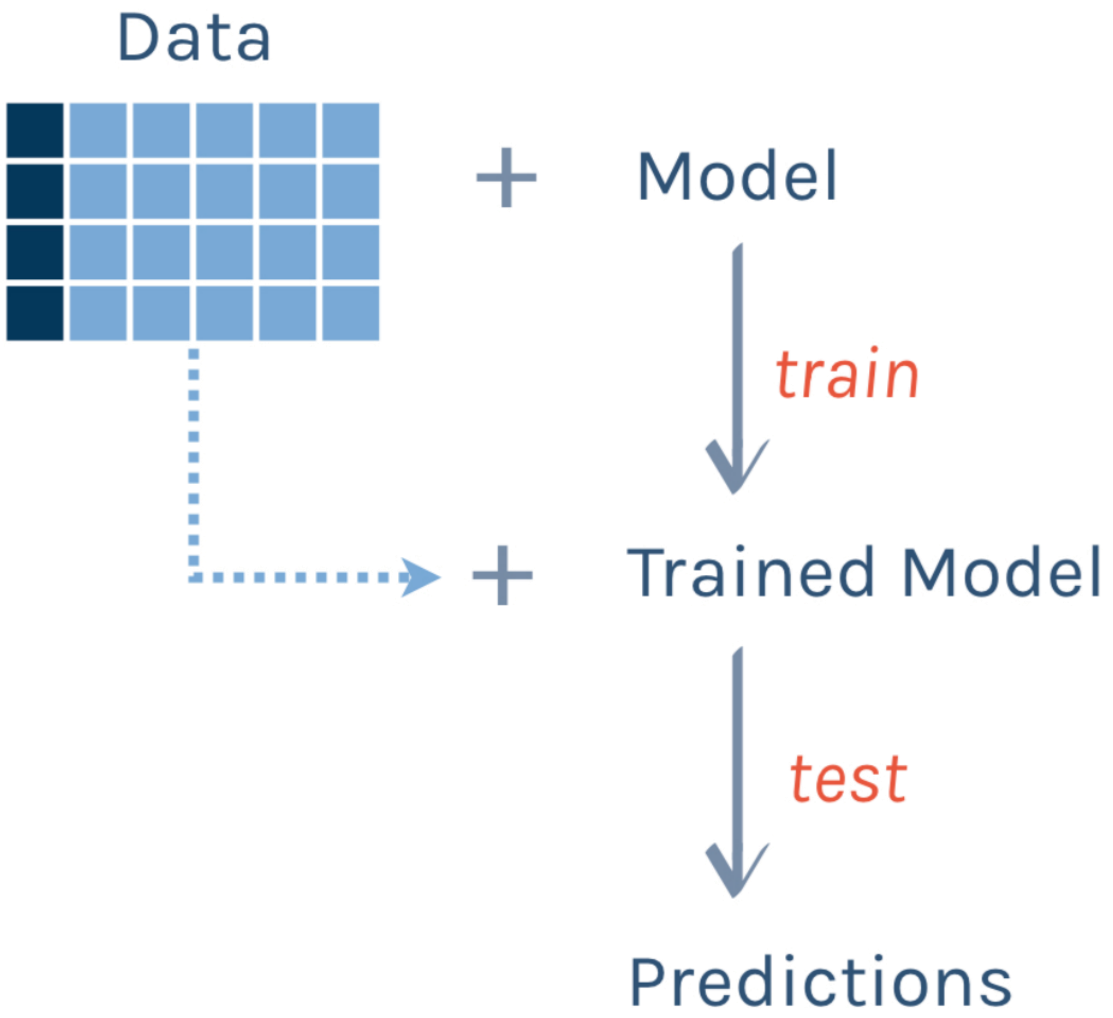
Basic ML principles and vocabulary – supervised and unsupervised machine learning



https://vas3k.com/blog/machine_learning/

Basic ML principles and vocabulary – training and testing

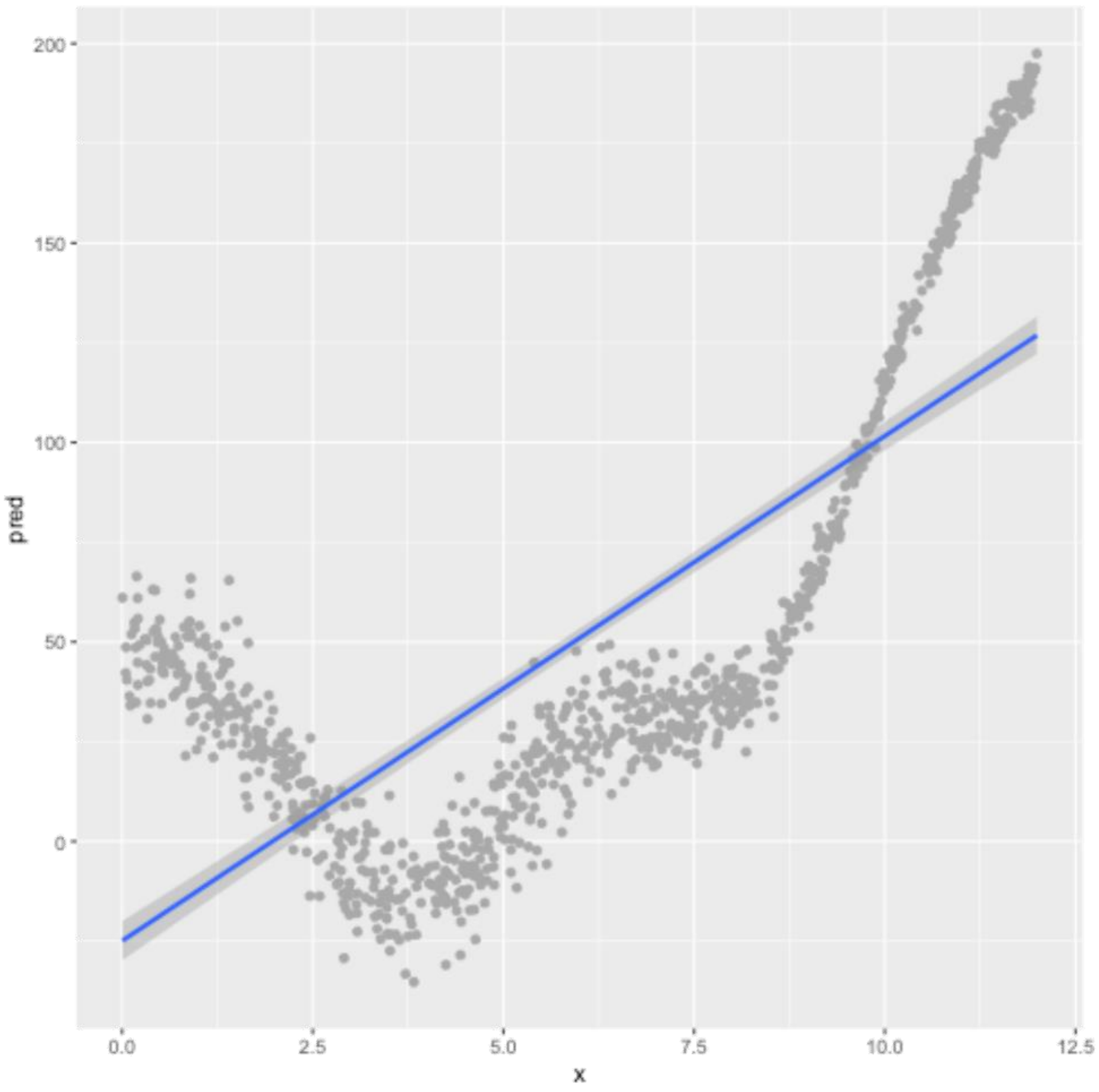




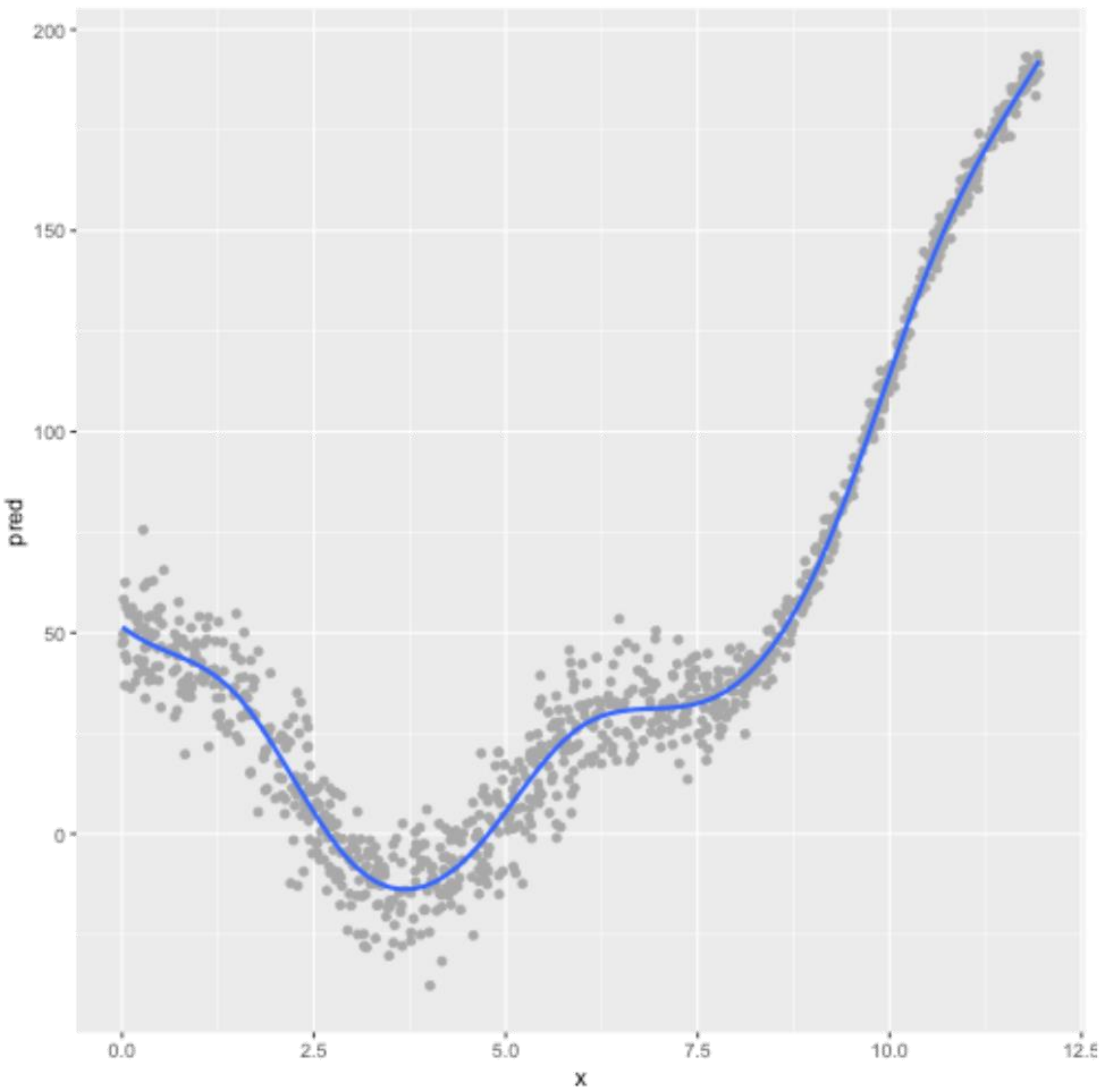
https://conf20-intro-ml.netlify.app/slides/01-predicting.html#35
https://conf20-intro-ml.netlify.app/slides/01-predicting.html#82
https://doi.org/10.1007/978-3-319-92058-0_71

Basic ML principles and vocabulary – the bias-variance trade-off

High bias

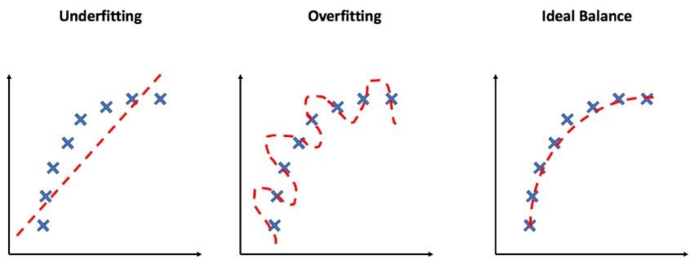


High variance



https://jrosen48.github.io/ML-in-Science-Education-Workshop-Materials/#25
https://jrosen48.github.io/ML-in-Science-Education-Workshop-Materials/#26

Basic ML principles and vocabulary – the bias-variance trade-off



Bias

- How (in-)correct our models' predictions are

- Models with high bias can fail to capture important relationship, they can be under-fitted to the data

- In short, how well our model reflects the patterns in the data

Variance

- How sensitive our predictions are to the specific sample on which we trained the model

- Models with high variance can fail to predict different data well, they can be over-fitted to the data

- In short, how stable the predictions of our model are when applied to new data

Regardless of the specific algorithm used, we often wish to balance between bias and variance. This is to balance between under- and over-fitting a model to the data at hand

https://jrosen48.github.io/ML-in-Science-Education-Workshop-Materials/#30
http://astrowizici.st/teaching/phs5000/10/

Basic ML principles and vocabulary – hyperparameters

Hyperparameters might address model design questions such as:

- What **degree of polynomial features** should I use for my linear model?
- What should be the **maximum depth** allowed for my decision tree?
- What should be the minimum number of **samples required at a leaf node** in my decision tree?
- **How many trees** should I include in my random forest?
- **How many neurons** should I have in my neural network layer?
- **How many layers** should I have in my neural network?
- What should I set my **learning rate** to for gradient descent?

https://www.jeremyjordan.me/hyperparameter-tuning/
https://giphy.com/gifs/fallontonight-tonight-show-im-sorry-what-jimmy-falllon-jUbCZEdpoNB3OJu1TT

The tidymodels package universe

tidymodels

tidymodels is a meta-package that installs and load the core packages listed below that you need for modeling and machine learning. Go to package ...

rsample

rsample provides infrastructure for efficient data splitting and resampling. Go to package ...

parsnip

parsnip is a tidy, unified interface to models that can be used to try a range of models without getting bogged down in the syntactical minutiae of the underlying packages. Go to package ...

recipes

recipes is a tidy interface to data pre-processing tools for feature engineering. Go to package ...

workflows

workflows bundle your pre-processing, modeling, and post-processing together. Go to package ...

tune

tune helps you optimize the hyperparameters of your model and pre-processing steps. Go to package ...

yardstick

yardstick measures the effectiveness of models using performance metrics. Go to package ...

broom

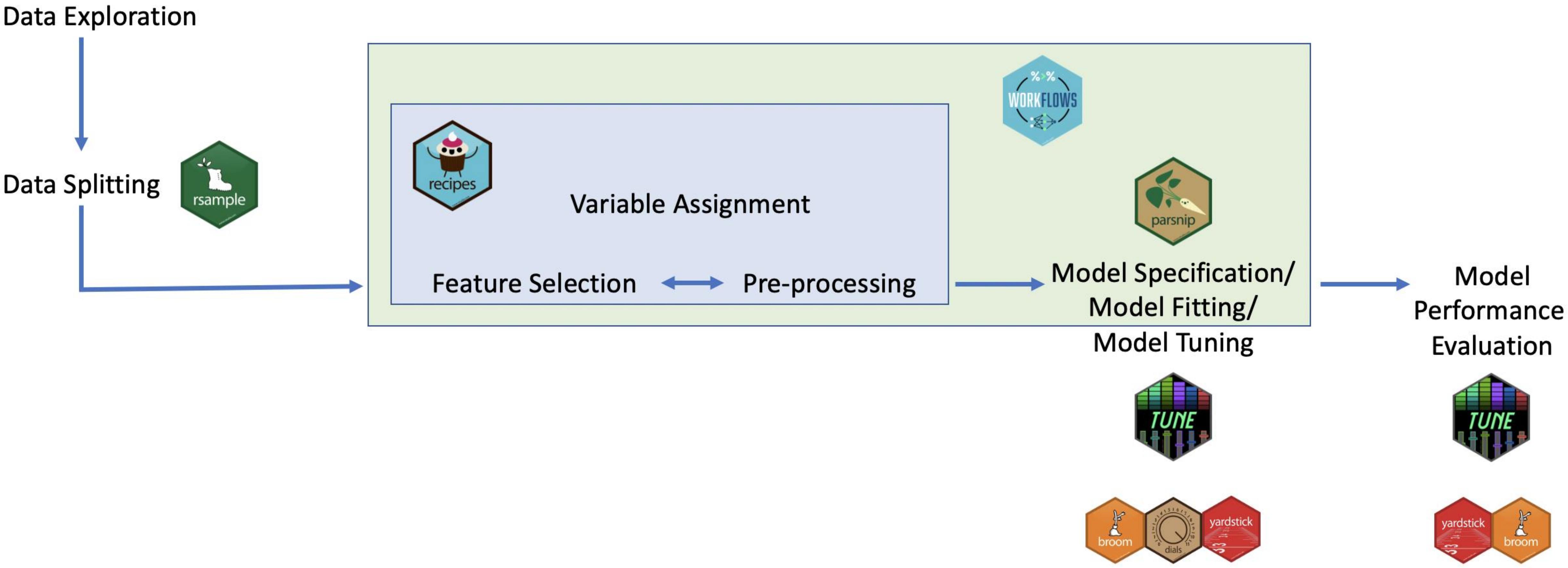
broom converts the information in common statistical R objects into user-friendly, predictable formats. Go to package ...

dials

dials creates and manages tuning parameters and parameter grids. Go to package ...

https://www.tidymodels.org/packages/

The tidymodels package universe



https://jhudatascience.org/tidyversecourse/model.html

The tidymodels package universe

Overview of *tidymodels* Basics		
Package	**Step**	**Functions**
rsample	**1.** Split into testing and training sets	initial_split() training() testing()
recipes	**2.** Create recipe + assign variable roles	recipe() update_role()
parsnip	**3.** Specify model, engine, and mode	parsnip function for specifying model (ex. decision_tree()) (https://www.tidymodels.org/find/parsnip/) set_engine() set_mode()
WORKFLOWS	**4.** Create workflow, add recipe, add model	workflow() add_recipe() add_model()
parsnip	**5.** Fit workflow	fit()
parsnip	**6.** Get predictions	predict()
yardstick	**7.** Use predictions to get performance metrics	rmse() (continuous outcome) accuracy() (categorical outcome) metrics() (either type of outcome)

https://jhudatascience.org/tidyversecourse/model.html

The tidymodels package universe

IDE: https://colab.to/r
Our course folder in Google Drive:
https://drive.google.com/drive/folders/13ZpamWwL10L45q9W6z3JuBlfYs0952Dp?usp=sharing
Website: www.tidymodels.org
Book: www.tmwr.org

tidymodels

rsample — Splitting data

recipes — Preprocessing

parsnip — Model specifications

yardstick — Model performance

broom — View models and metrics in a tidy way

WORKFLOWS — Make modeling workflow

TUNE — Tune hyperparameters and get performance metrics

dials — Tune hyperparameters

https://jhudatascience.org/tidyversecourse/model.html