# Demystifying Random Forest

Henn Lab Meeting

Feb. 24, 2022

# Why learn about random forest?

# 1. Random forest is trendy and useful

## It shows up in helpful methodologies …

RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference

Brian K. Maples,[1,2] Simon Gravel,[1,3] Eimear E. Kenny,[1,4,5,6,7,8] and Carlos D. Bustamante[1,8,*]

**Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest**

François-David Collin, Ghislain Durif, Louis Raynal, Eric Lombaert, Mathieu Gautier, Renaud Vitalis, Jean-Michel Marin, Arnaud Estoup

First published: 05 May 2021

## And relevant research / modeling papers

**An RNA-seq Based Machine Learning Approach Identifies Latent Tuberculosis Patients With an Active Tuberculosis Profile**

*Olivia Estévez[1,2], Luis Anibarro[2,3,4], Elina Garet[1,2], Ángeles Pallares[5], Laura Barcia[3], Laura Calviño[3], Cremildo Maueia[6], Tufária Mussá[6,7], Florentino Fdez-Riverola[1,2,8], Daniel Glez-Peña[1,2,8], Miguel Reboiro-Jato[1,2,8], Hugo López-Fernández[1,2,8], Nuno A. Fonseca[9,10], Rajko Reljic[11] and África González-Fernández[1,2*]*

**A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer**

**Ashraf Abou Tabl[1*], Abedalrhman Alkhateeb[2*], Waguih ElMaraghy[1], Luis Rueda[2] and Alioune Ngom[2]**

[1]Department of Mechanical, Automotive and Materials Engineering, University of Windsor, Windsor, ON, Canada
[2]School of Computer Science, University of Windsor, Windsor, ON, Canada

# 2. ML needs more diversity, less gatekeeping

TOM SIMONITE BUSINESS AUG 17, 2018 7:00 AM

## AI Is the Future—But Where Are the Women?

**Just 12 percent of machine learning researchers are women—a worrying statistic for a field supposedly reshaping society.**

## The artificial intelligence field is too white and too male, researchers say

*A new report explores AI's 'diversity crisis'*

By Colin Lecher | @colinlecher | Apr 16, 2019, 8:00pm EDT

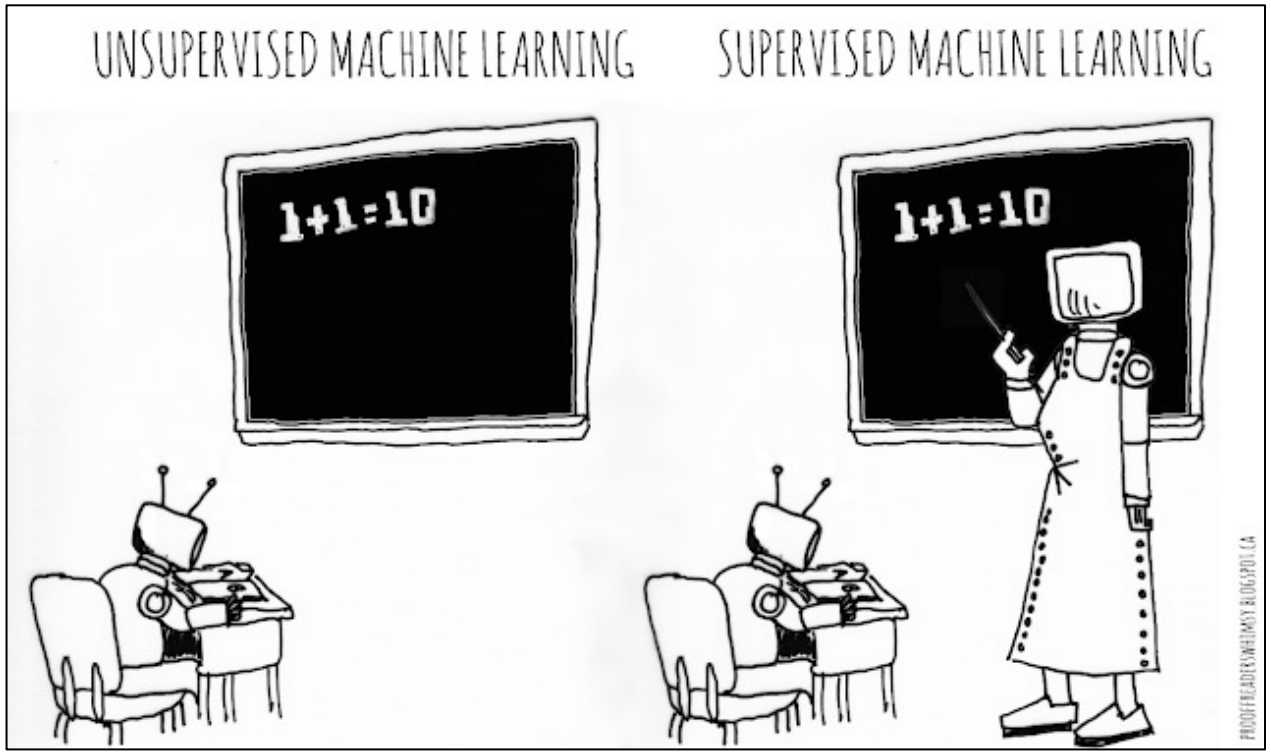
RESEARCH-ARTICLE

## A Survey on Bias and Fairness in Machine Learning

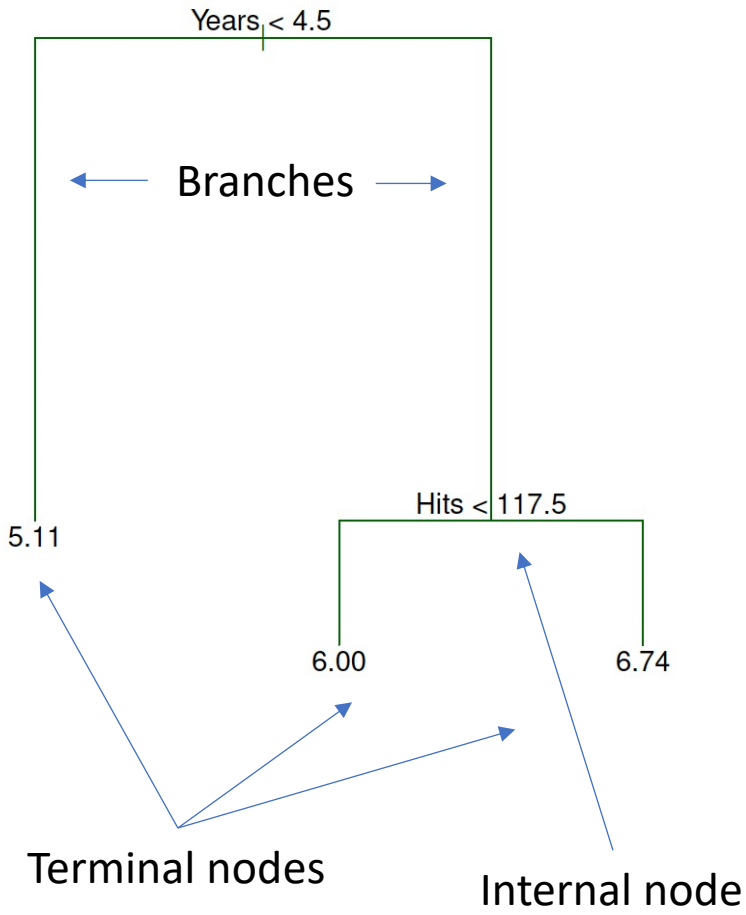
**Authors:** Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan

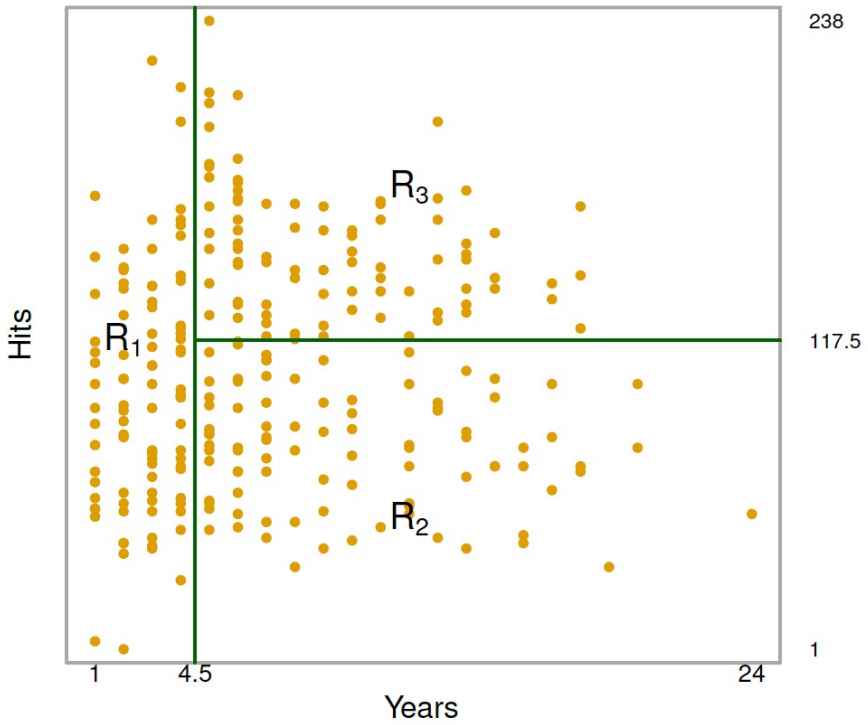
# Plan

- Review decision trees
- Bagging
- Random forest
- Pros & Cons
- Hands-on practice



# Decision trees



# Creating a decision tree

1. **Recursive binary splitting**
   - Grow large tree based on training data
   - Top-down, greedy approach
   - Makes a complex tree that over-fits the data
2. **Tree pruning**
   - Cost complexity pruning based on $\alpha$
3. **K-fold cross-validation**
   - To choose best $\alpha$

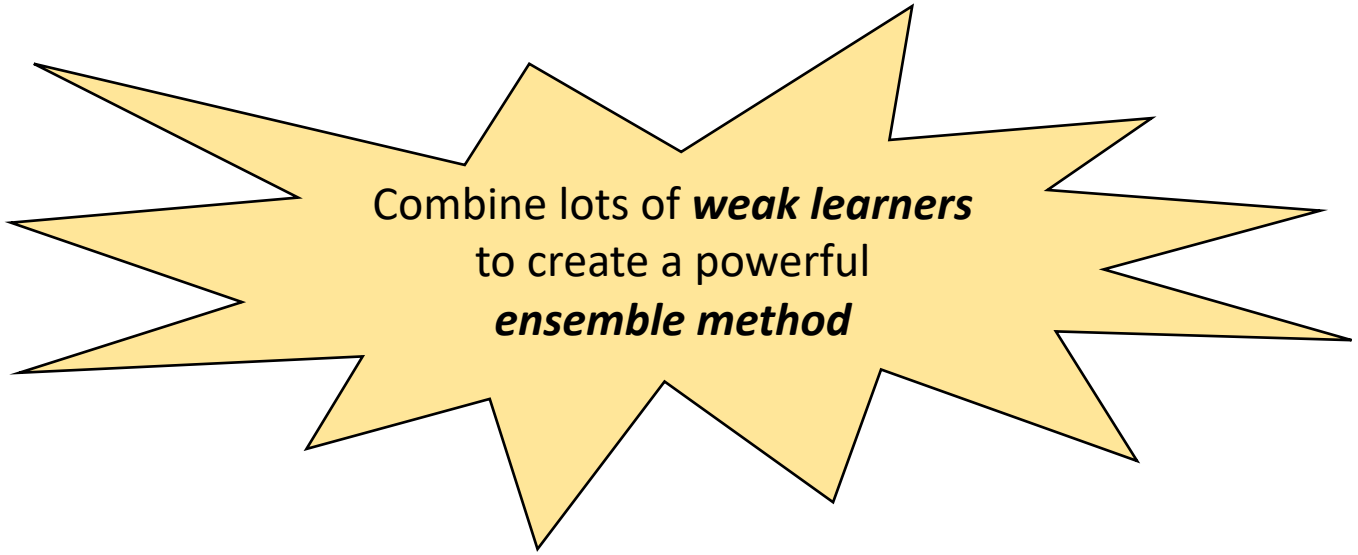
# Decision Trees: Pros and Cons

## Pros

- Trees are easy to explain
- Trees may mirror human decision-making
- Trees can be displayed graphically and are easily interpretable

## Cons

- They don't work very well!
- Not as accurate as other methods
- Highly variant, not robust



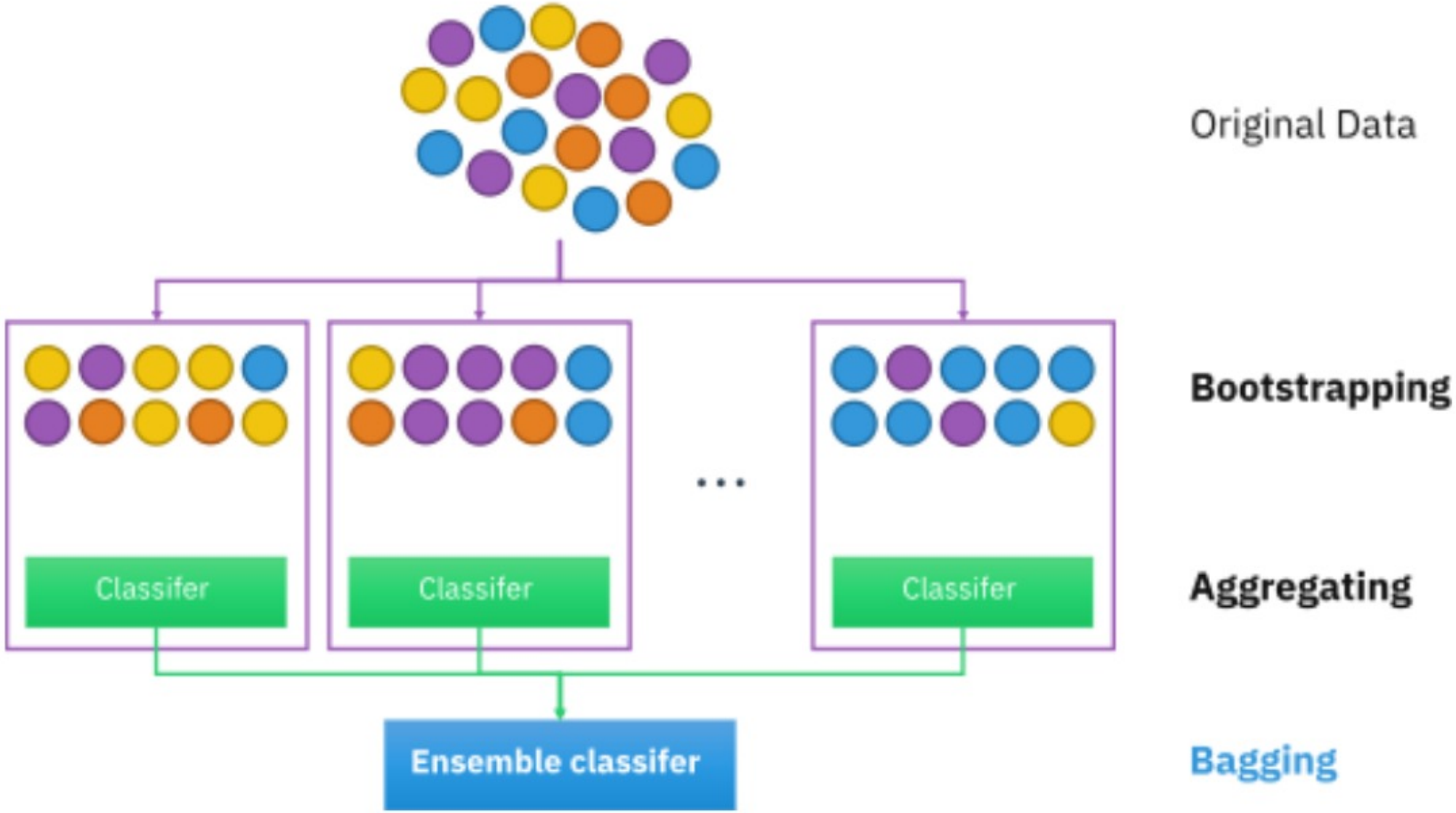
# Bagging

- Averaging a set of observations reduces variance (!)
- If we had B separate training sets:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x).$$

- Instead, we can take random repeated samples from the dataset, build models, and average all the predictions ("aggregated bootstrap"):

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$



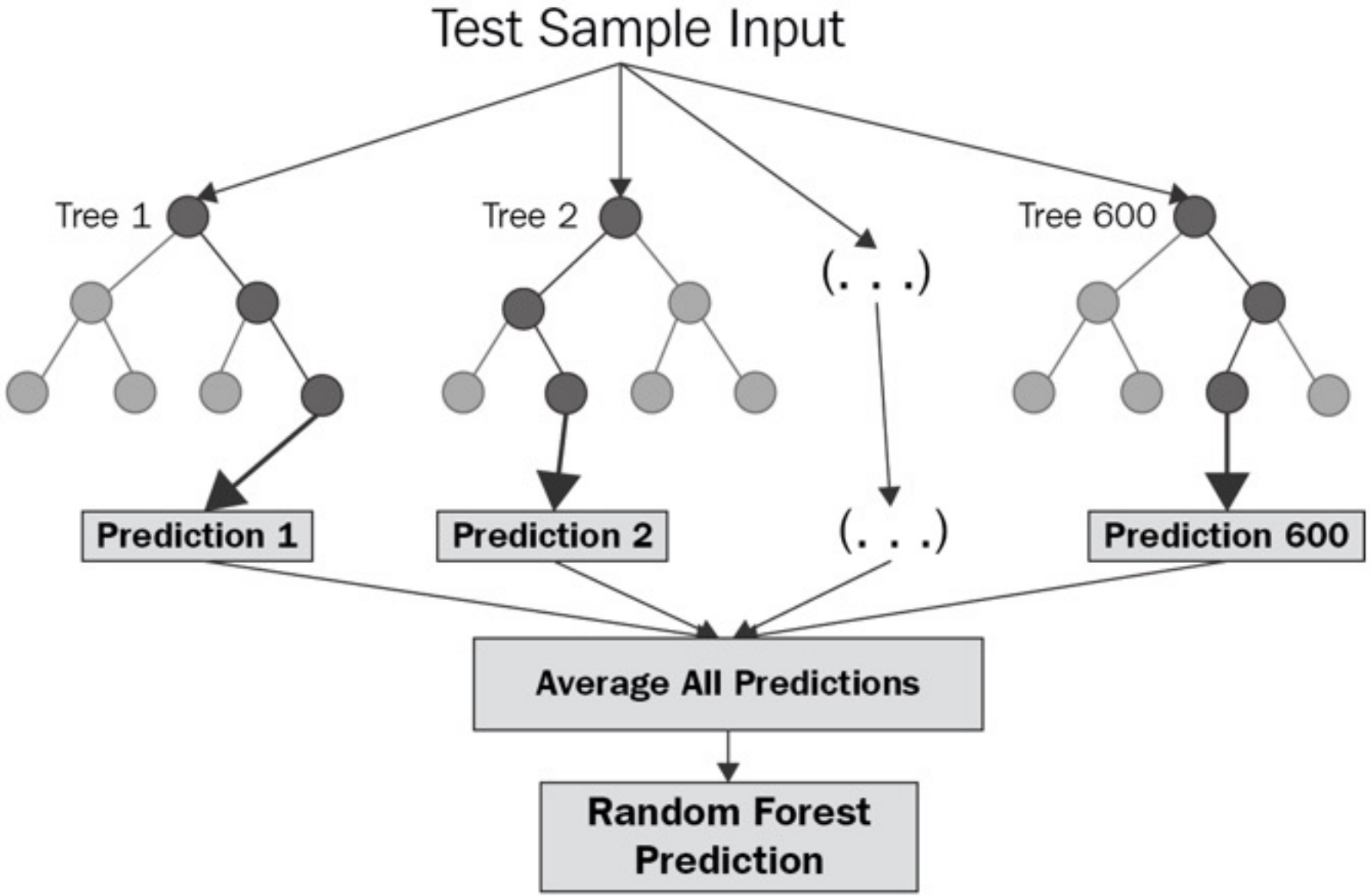

https://www.analyticsvidhya.com/blog/2020/02/what-is-bootstrap-sampling-in-statistics-and-machine-learning/

# Perks of bagging

- Each bagged tree makes use of around two-thirds of the observations
  - prevents overfitting
- We can calculate Out-Of-Bag error estimate
  - similar to cross-validation
- We can calculate variable importance based on how much the error decreases due to splits for a given predictor (averaged over all the trees)

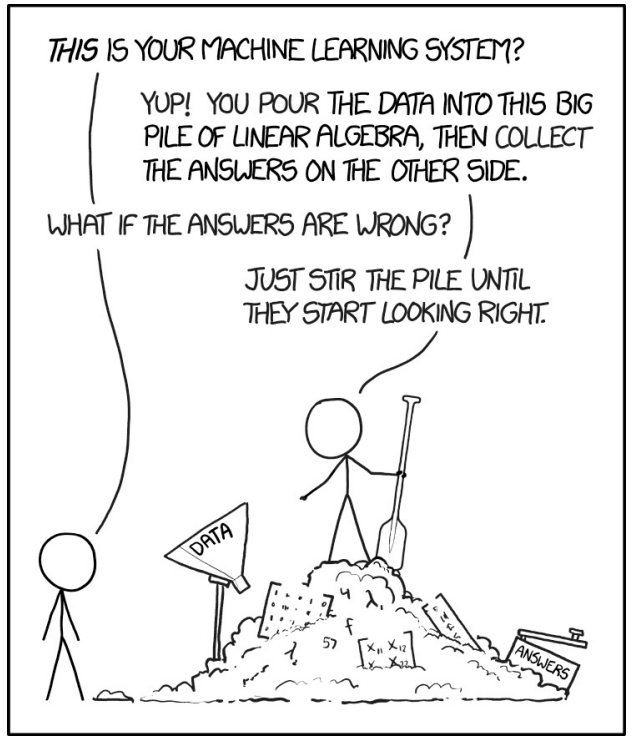
# Random Forest

- Similar to bagging: we are still building lots of trees using bootstrapped training samples
- This time **the trees are decorrelated**
- At each split in the decision tree, a random sample of m predictors is chosen as split candidates from the full set of p predictors
  - Generally m = sqrt(p)
- Prevents one important variable from dominating every tree




https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

# Random forest: Pros and Cons




## Pros

- Relationship between variables does not have to be linear or consistent
- Doesn’t overfit data
- Can included correlated variables
- Can use continuous or categorical variables
- Easy to implement

## Cons

- Hard to interpret model
- Needs large dataset

# Want to learn more?

- Check out Introduction to Statistical Learning (https://hastie.su.domains/ISLR2/ISLRv2_website.pdf) chapter 8