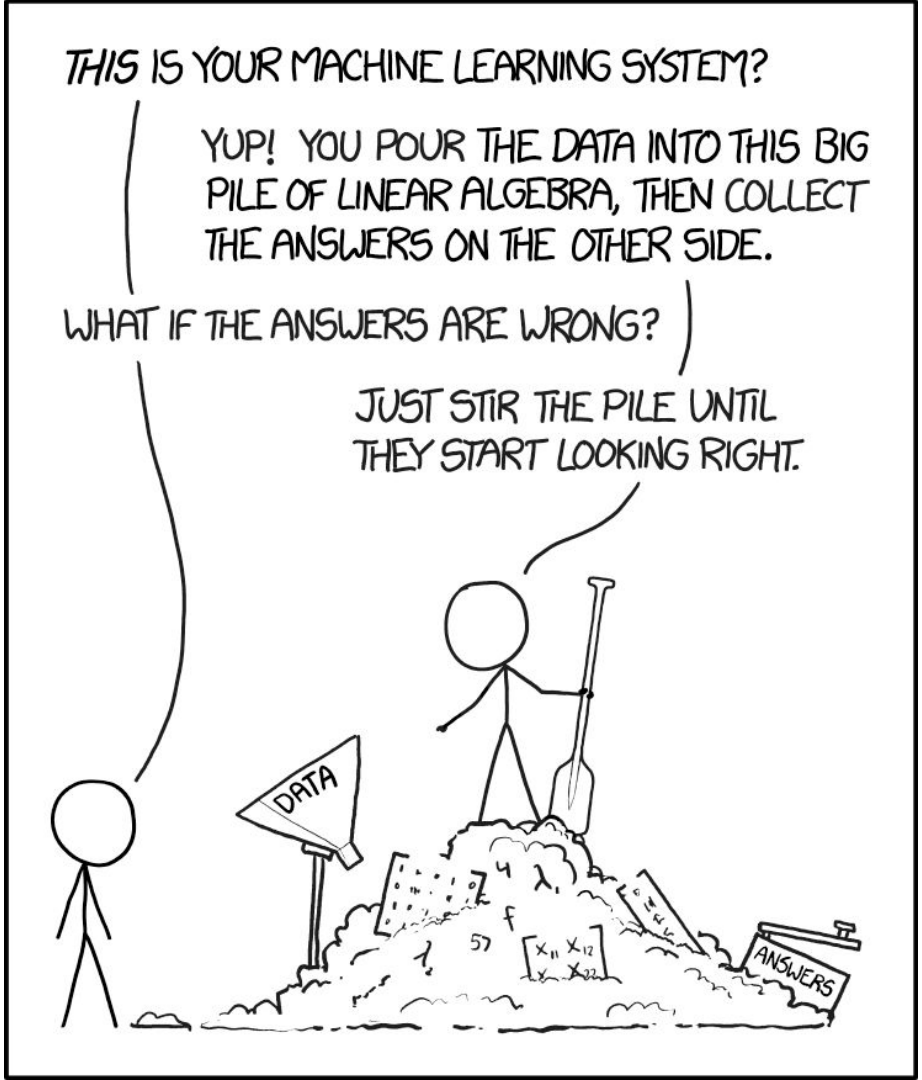
INTRO TO DECISION TREES AND BEYOND



[xkcd.com/1838]


Artificial Intelligence
HEY WHY DO YOU ALWAYS WEAR THAT MASK?
Machine Learning
Machine Learning
Machine Learning
Statistics
LET'S KEEP THIS ON.
Machine Learning

[reddit.com]

STATISTICS vs MACHINE LEARNING

What do we want?

- DESCRIBE past trends/effects OR PREDICT “future”?
- Testing HYPOTHESIS OR finding new PATTERNS?

What kind of data is available?

CLASSICAL STATISTICS

- HYPOTHESIS TESTING

SUPERVISED LEARNING

- REGRESSION
- CLASSIFICATION (BINARY vs MANY)

OTHER

- ANOMALY DETECTION
- SIMULATIONS
- OPTIMIZATION
- …and more

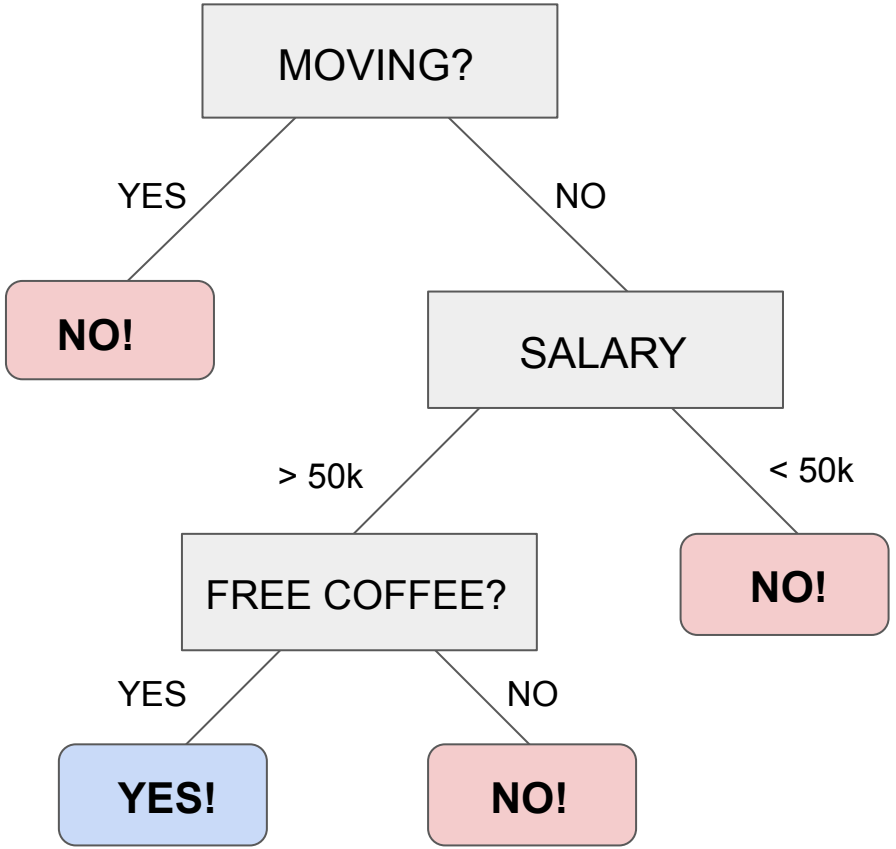
UNSUPERVISED LEARNING

- CLUSTERING
- HIDDEN PATTERNS
- DIMENSIONALITY REDUCTION

DECISION TREE

- classification algorithm
- tree-like structure to model relationships between the data features and possible outcomes
- branching decisions based on conditions
- split based on minimizing entropy in each node
- leaf nodes contain the decision results

Example: Should I take a job offer?



DECISION TREE

PROS

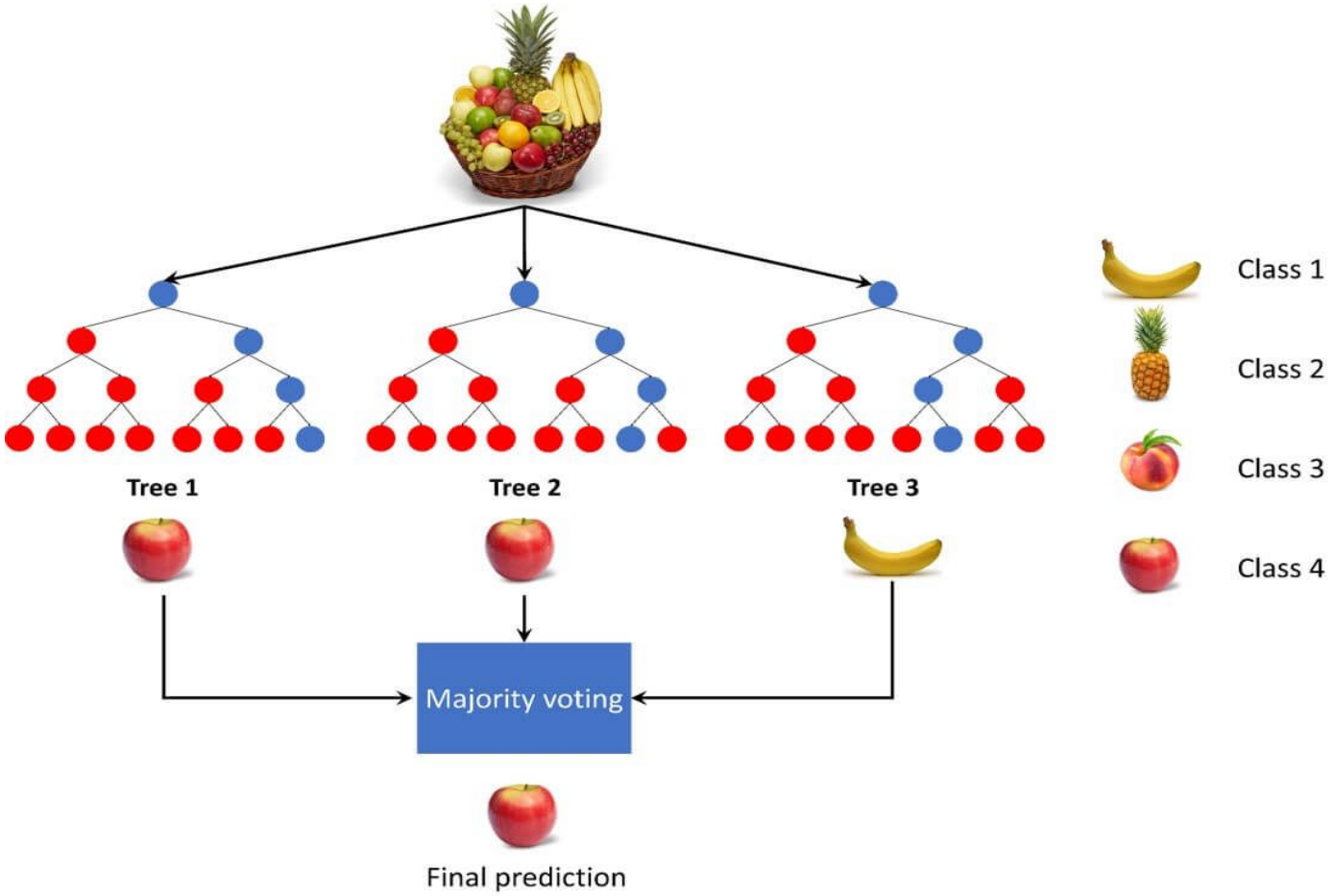
- classification & regression
- intuitive & interpretable
- handling both numerical and categorical data
- access to feature importance
- robust to outliers
- can handle missing data
- relatively fast

CONS

- prone to overfitting
- bias towards features with high cardinality (many unique values)
- data imbalance (accurate predictions for minority class)
- high variance for different subsets of data

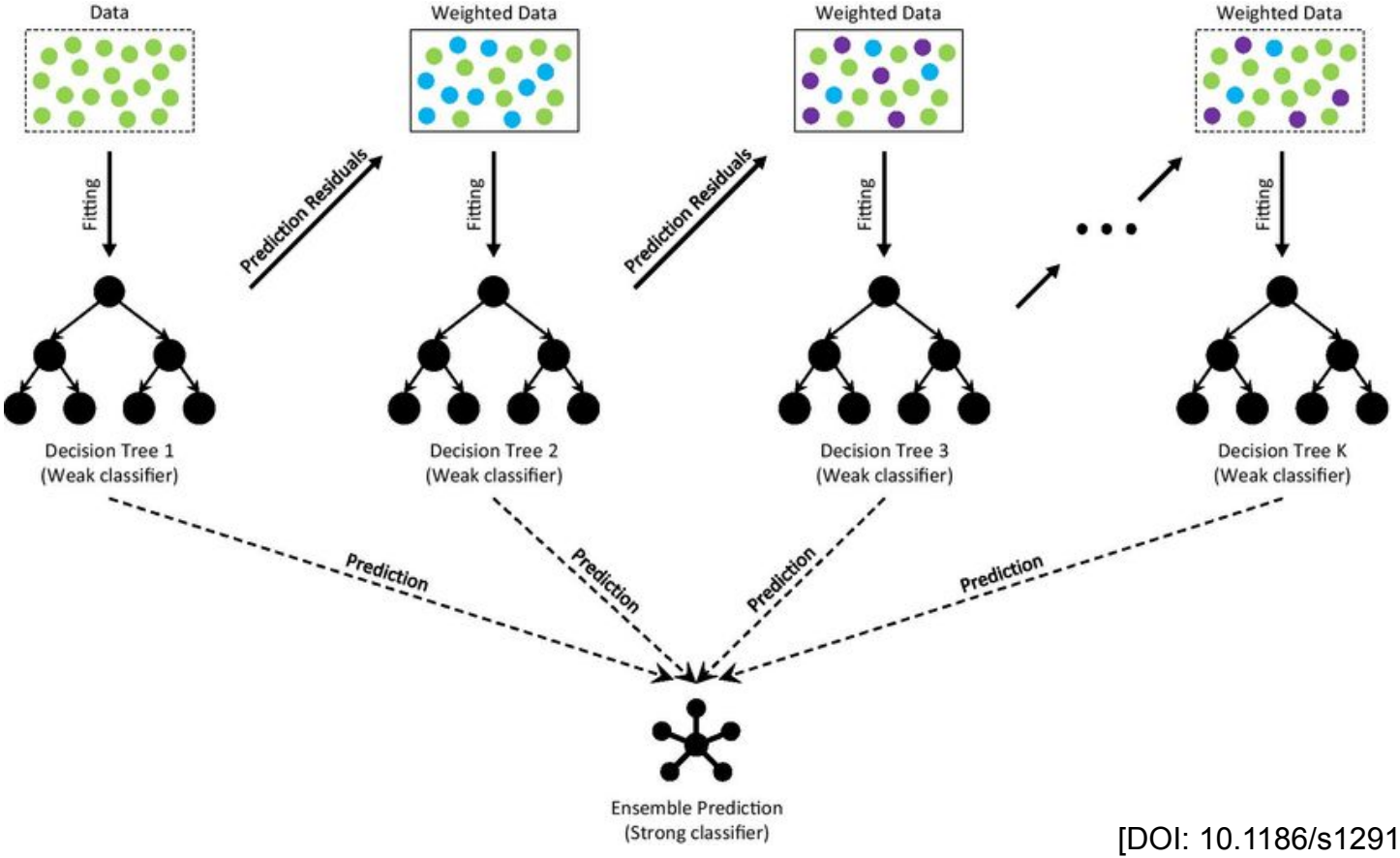

FROM SINGLE DT
TO MORE ROBUST MODELS

BAGGING – RANDOM FOREST



[datahacker.rs]

BOOSTED DECISION TREES



[DOI: 10.1186/s12911-021-01701-9]

ML (TREES & FORESTS) IN HEP

- **Problems of interest:**
  - classification - signal vs background, jet finding & identification, regression
  - HL-LHC era - high pile-up, demanding on computational resources
  - reduce complexity, find new patterns in data
- Most popular - **BDT** and **Neural Networks**
- **ROOT TMVA package -** https://root.cern/manual/tmva
  - supervised learning for multivariate classification and regression
  - takes a “signal” and “background” (ROOT) data tree for training (usually from MC)
  - trained model applied to data to distinguish class of interest
- **Be cautious!**
  - imbalanced problems (e.g. rare signal with large background)
  - train/test data selection (size, representativeness, NO mixing)
  - overfitting

SOME EXAMPLES

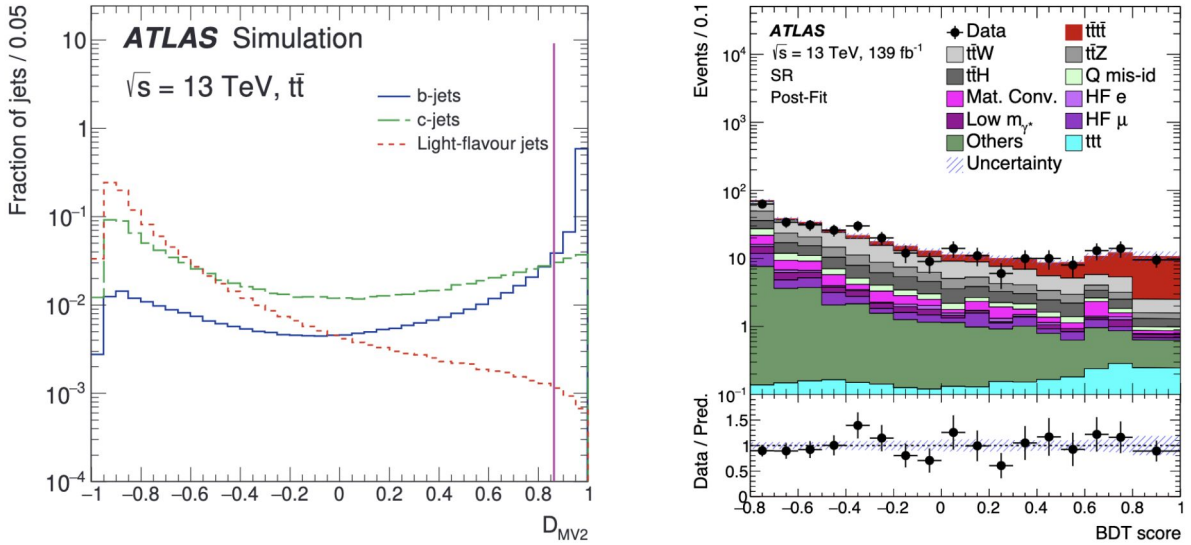


Fig. 1. (a) Output of the boosted decision tree used to identify jets originating from b-quarks in ATLAS [7]. (b) Boosted decision tree output used in a fit between data and physics model to extract the $t\bar{t}t\bar{t}$ signal [8].

[arXiv:2206.09645]

BONUS: References, NN (in particle physics)

- A Living Review of ML for Particle Physics: https://github.com/iml-wg/HEPML-LivingReview
- Inter-Experimental LHC Machine Learning Working Group: https://iml.web.cern.ch/homepage
- Nice illustrative intro to neural networks: https://aegeorge42.github.io/
- Scikit-learn: https://scikit-learn.org/stable
- Intro to TensorFlow: https://www.tensorflow.org/tutorials/keras/classification
- Intro to PyTorch: https://pytorch.org/tutorials/beginner/basics/intro.html