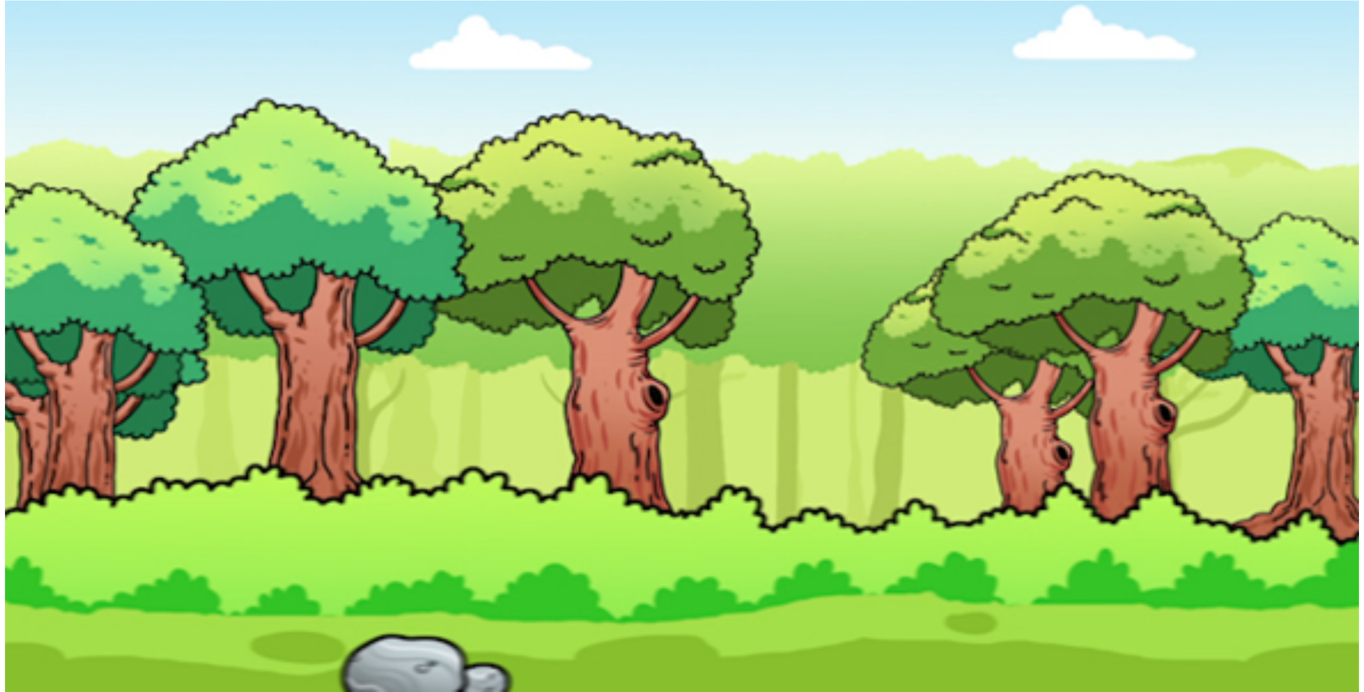
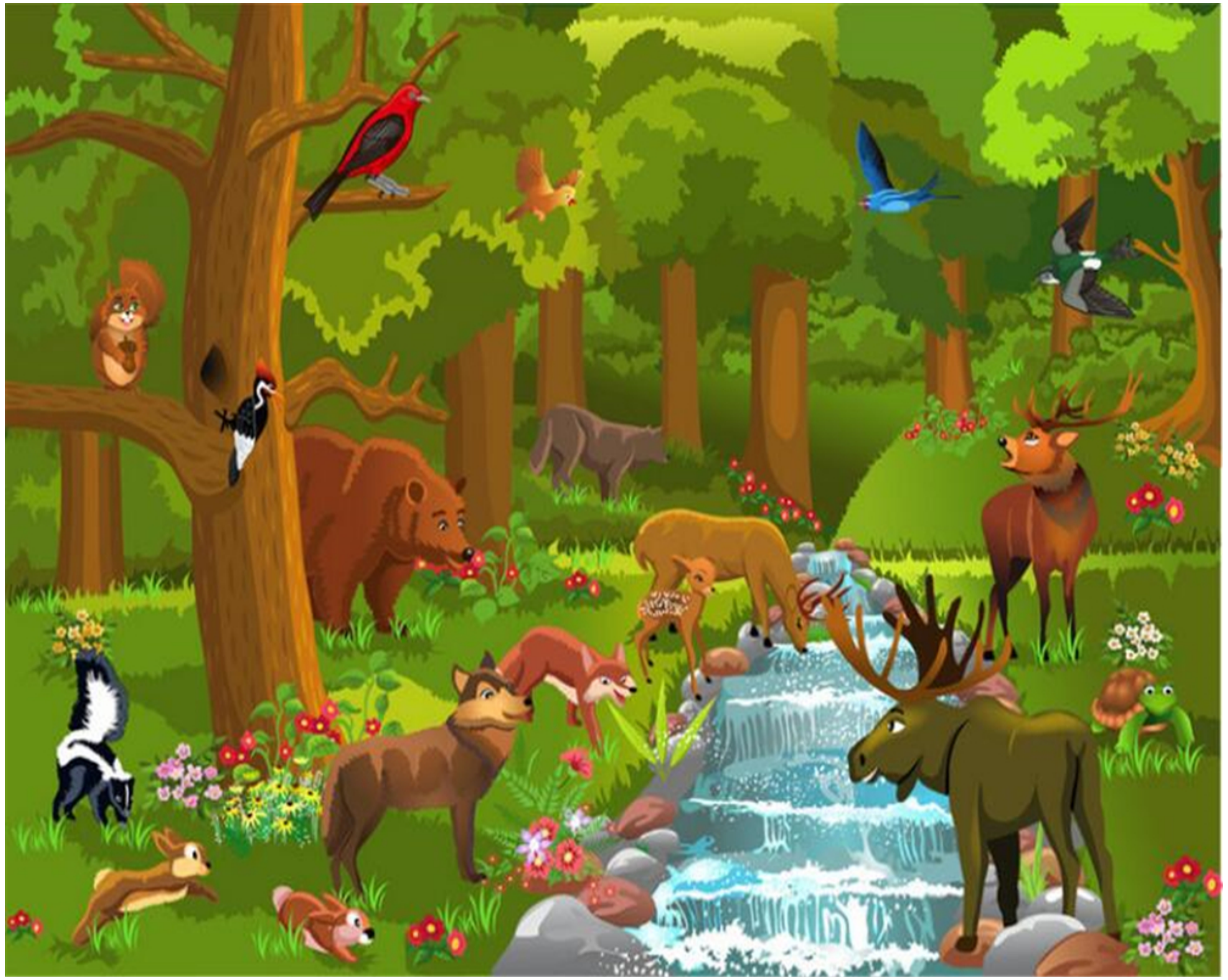
A Deep Dive into Random Forests

By: Ari Silburt



Random Forests May Seem Scary…

But They’re Actually Not Too Bad!



Plan

- Decision Tree
- Random Forests (+ Bagging)
- Optimization (Gini Criterion, Regularization)
- Closing Notes

Decision Tree

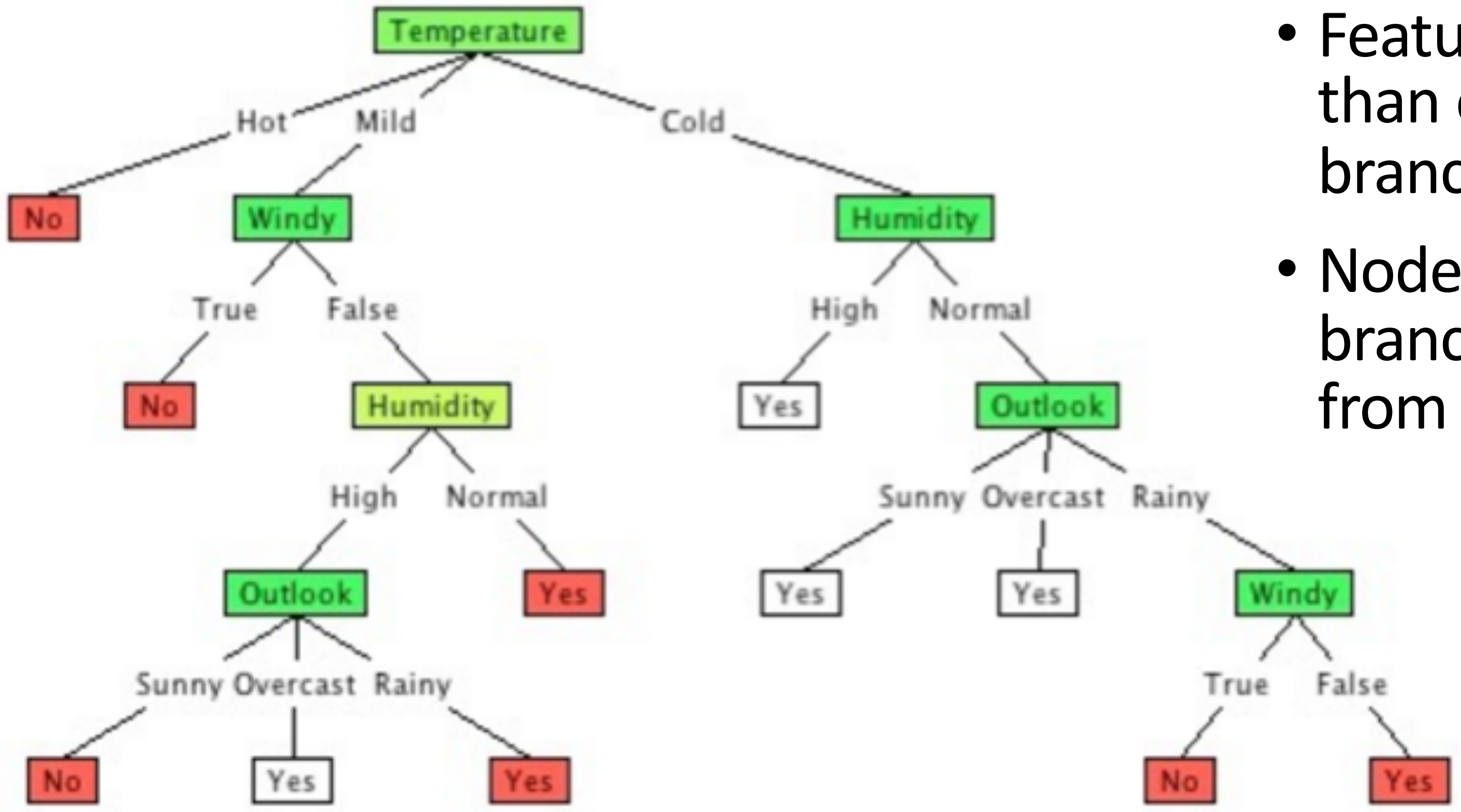
Decision Tree

Example: Should We Play Tennis?

Play Tennis	Outlook	Temperature	Humidity	Windy
No	Sunny	Hot	High	No
No	Sunny	Hot	High	Yes
Yes	Overcast	Hot	High	No
Yes	Rainy	Mild	High	No
Yes	Rainy	Cold	Normal	No

- If temperature is not hot
  - Play
- If outlook is overcast
  - Play tennis
- Otherwise
  - Don't play tennis

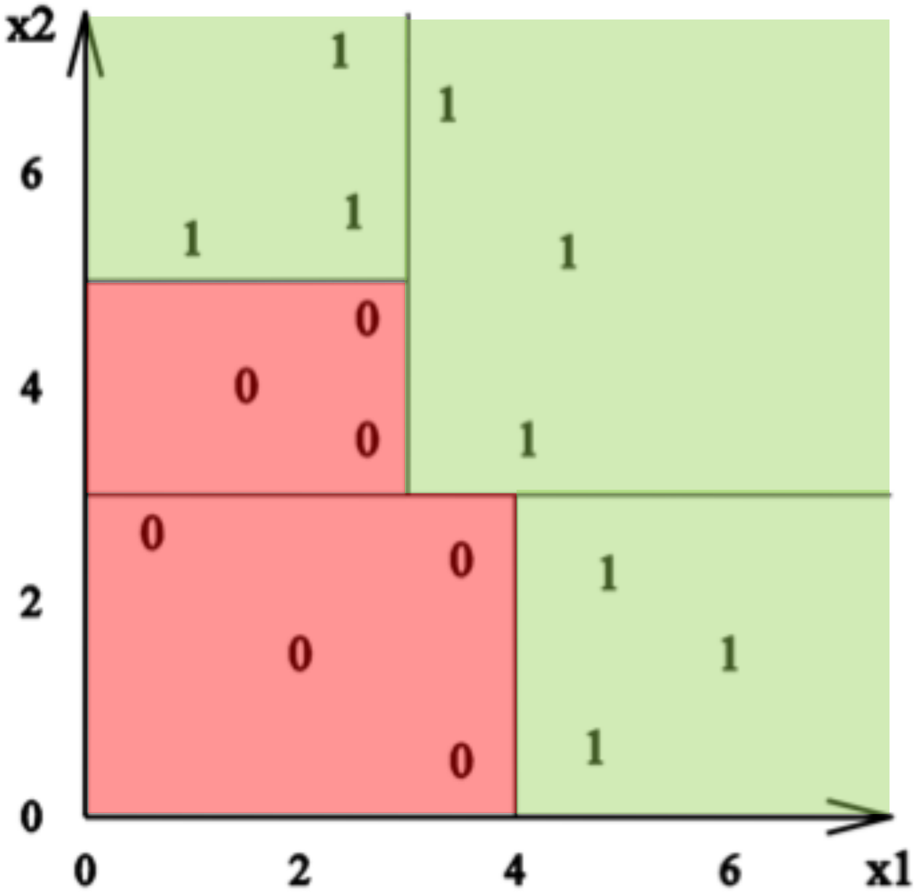
Decision Tree

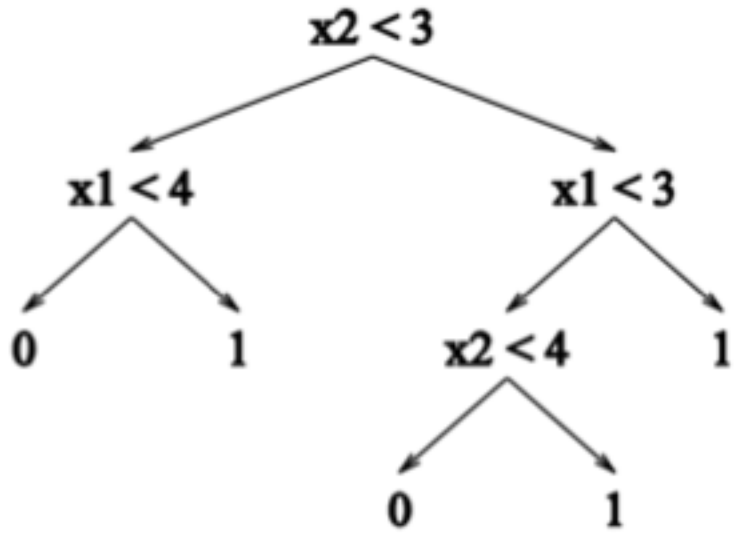


<u>Properties</u>

- Feature can show up more than once in different branches (e.g. windy).
- Node can have both a branch and leaf stemming from it.

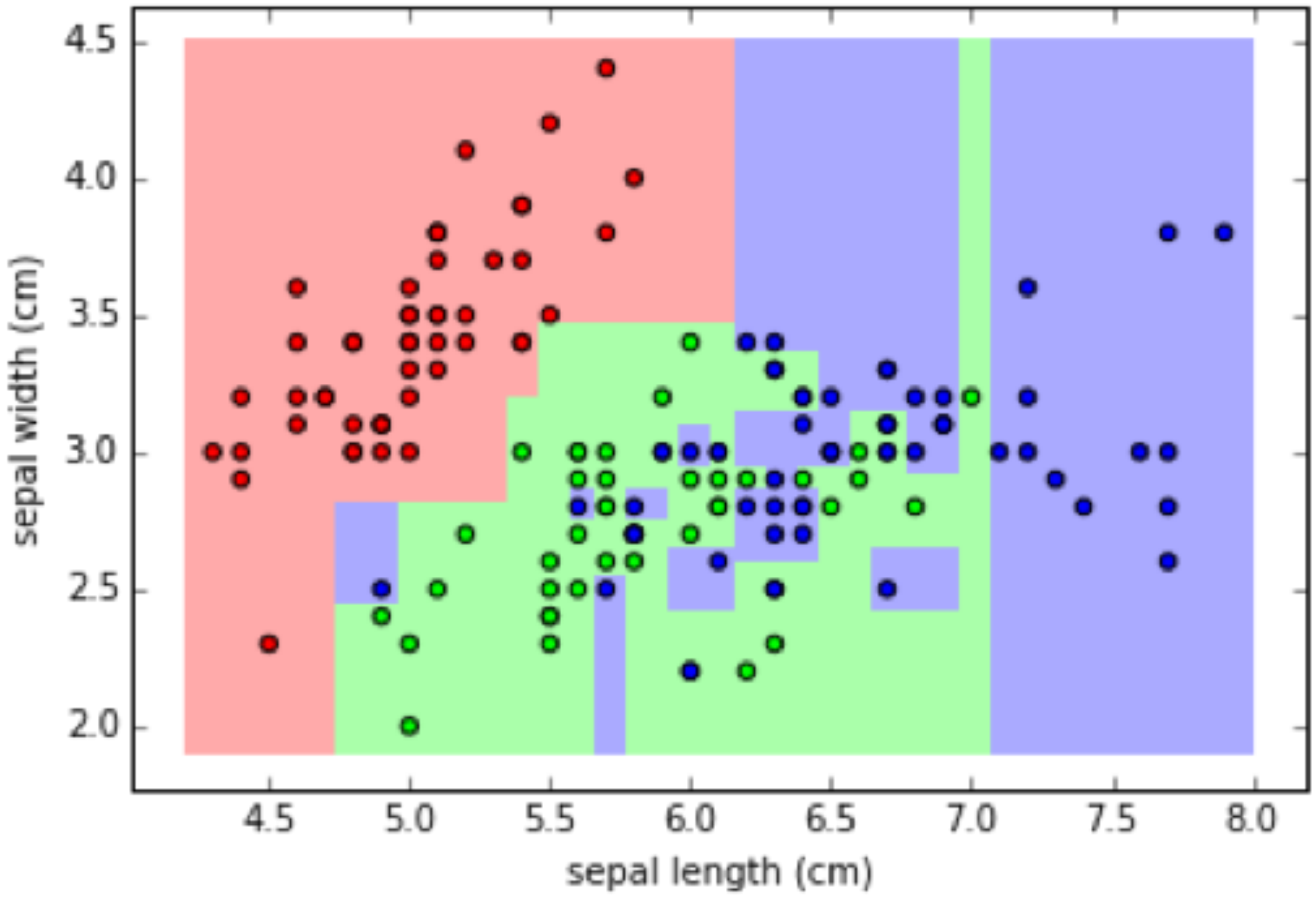
Decision Tree – Pros and Cons





Pros:
- Non-linear decision boundaries
- Easy to interpret
- Numerical & Categorical Data

Decision Tree – Pros and Cons



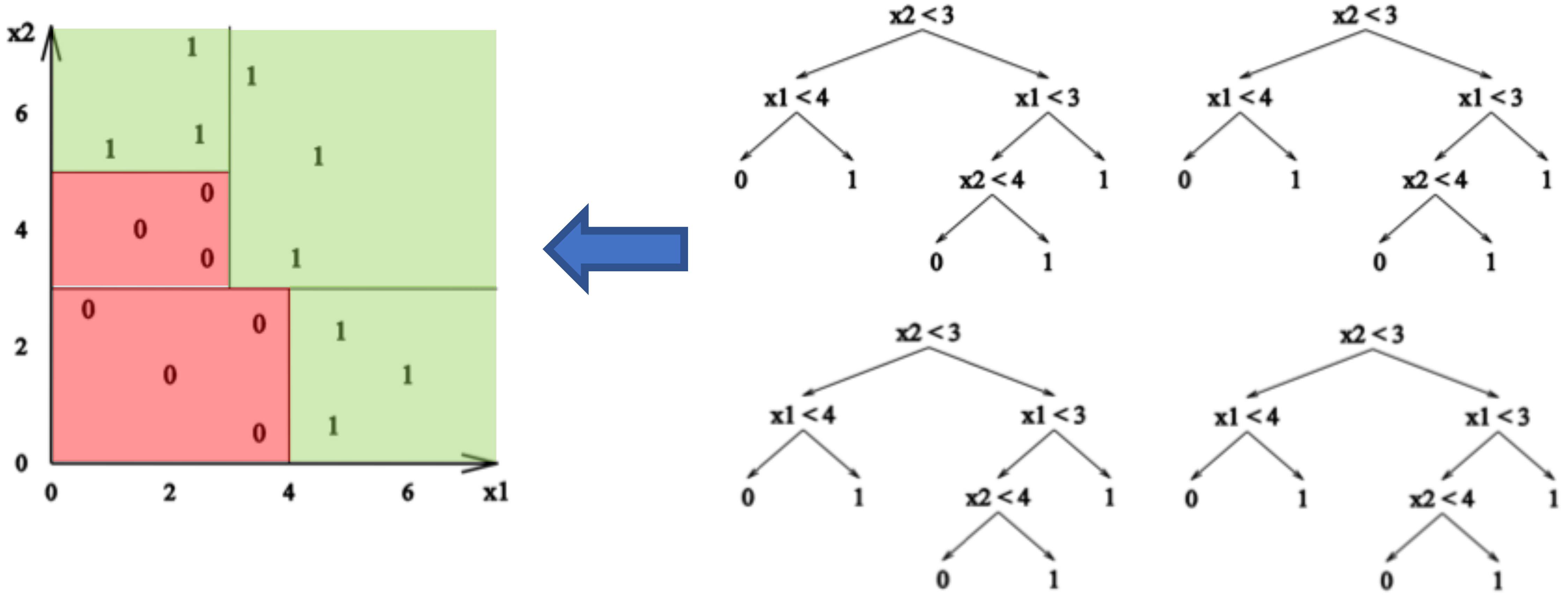
Cons:

- Easy to Overfit
- High Variance (i.e. unstable).

Random Forests

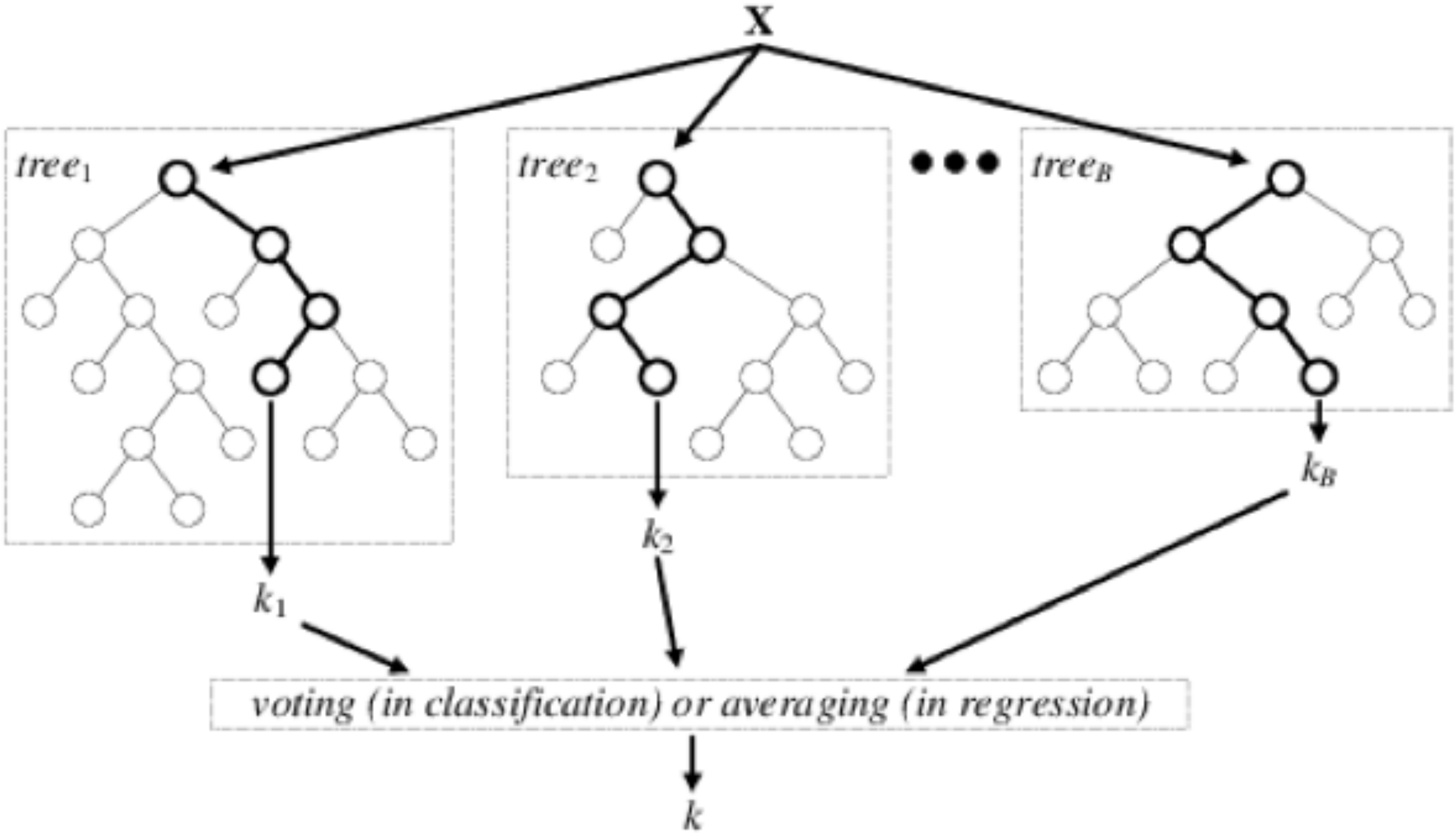
Random Forests – Many Decision Trees

- We can guess that a Random Forest = many decision trees. But how?
- Many copies of the exact same tree is useless…



Random Forests – Many Decision Trees

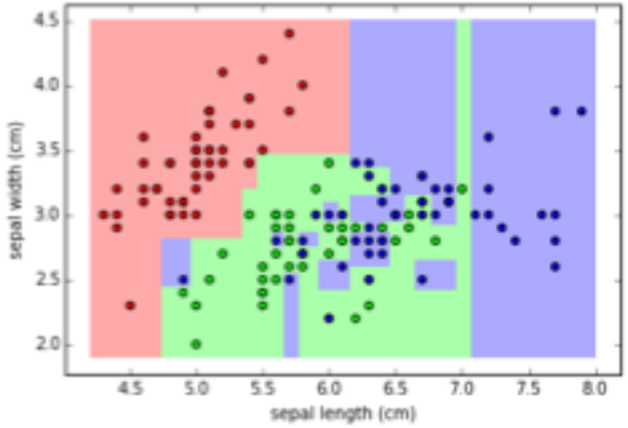
- OK, so we want some tree variation, but how…



Random Forests – Many Decision Trees

- We want tree variation, but how…
- **Vary trees such that overall variance is reduced:**

Remember: Decision Tree

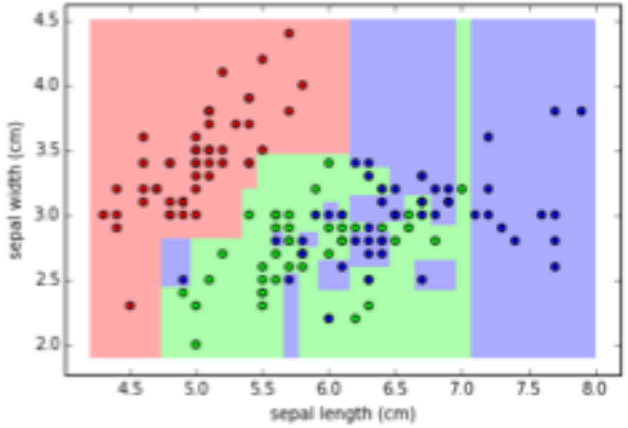


Random Forests – Many Decision Trees

- We want tree variation, but how…
- **Vary trees such that overall variance is reduced:**
- STATS101:

Given a set of **independent, uncorrelated observations** $Z_1, Z_2, \ldots, Z_n$ each with variance σ^2, the variance of Z is $\boldsymbol{\frac{\sigma^2}{n}}$.

Remember: Decision Tree



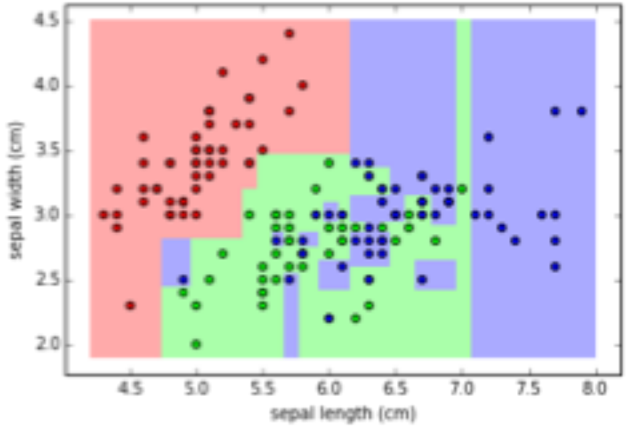
Random Forests – Many Decision Trees

- We want tree variation, but how…
- **Vary trees such that overall variance is reduced:**
- STATS101:

Given a set of **independent, uncorrelated observations** $Z_1, Z_2, \ldots, Z_n$ each with variance σ^2, the variance of Z is $\boldsymbol{\frac{\sigma^2}{n}}$.

This is why a forest of identical trees is useless.

Remember: Decision Tree



Random Forests – Many Decision Trees

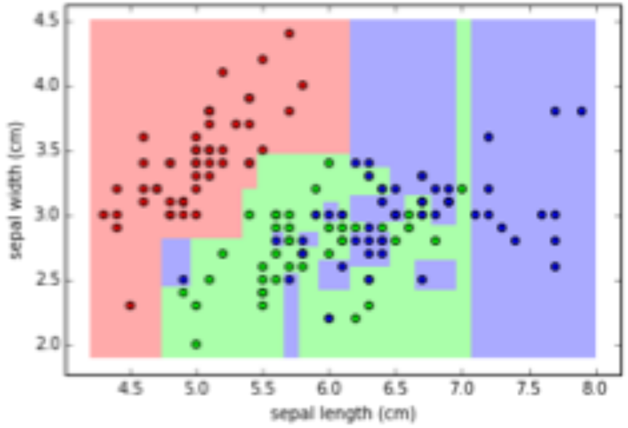
- We want tree variation, but how…
- **Vary trees such that overall variance is reduced:**
- STATS101:

Given a set of **independent, uncorrelated observations** $Z_1, Z_2, \dots, Z_n$ each with variance σ^2, the variance of Z is $\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}$.

This is why a forest of identical trees is useless.

This is why ensembling many models together always improves results.

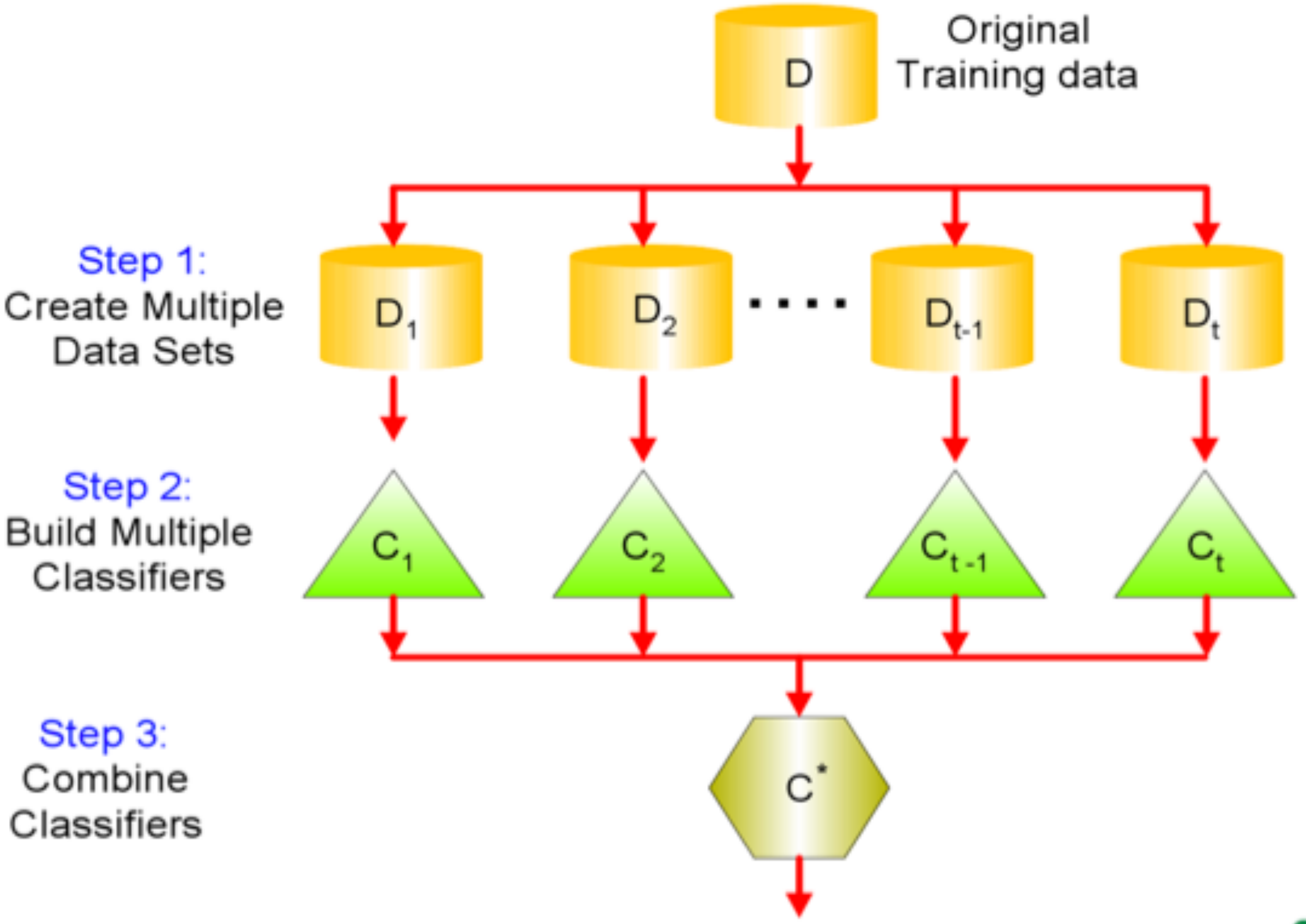
Remember: Decision Tree



Random Forests – Randomize Data

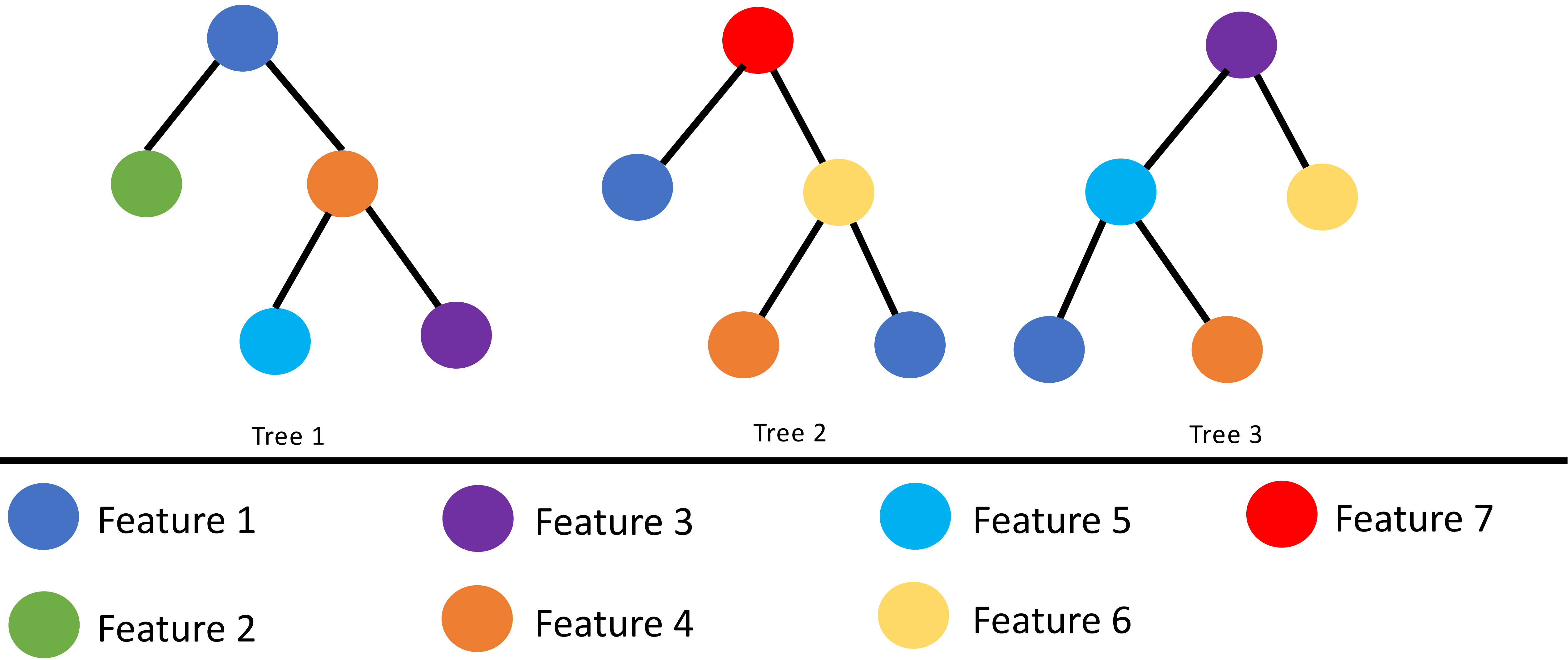
- **Bootstrap sampling**: Given set D containing N training examples, create D' by drawing N examples at random **with replacement** from D
- **Bagging**
  - Create k bootstrap samples $D_1, \ldots, D_k$
  - Train distinct classifier on each D_i
  - Classify new instance by majority vote / average

Random Forests – Randomize Data



[image from the slides of David Sontag]

Random Forests – Randomize Features



Random Forests – Intuition Check

- <u>What happens if you assign more/less data per tree</u>?

- <u>What happens if you select more/less of the total features per tree</u>:

Random Forests – Intuition Check

- <u>What happens if you assign more/less data per tree</u>?

Less: Trees more uncorrelated, but at some point too little data hurts training.

More: Trees become more correlated, training of each tree improved.

- <u>What happens if you select more/less of the total features per tree</u>:

Less: Trees more uncorrelated, but at some point many trees become “dead”, i.e. fitting entire trees on unimportant features.

More: Trees become more correlated, training of each tree improved.

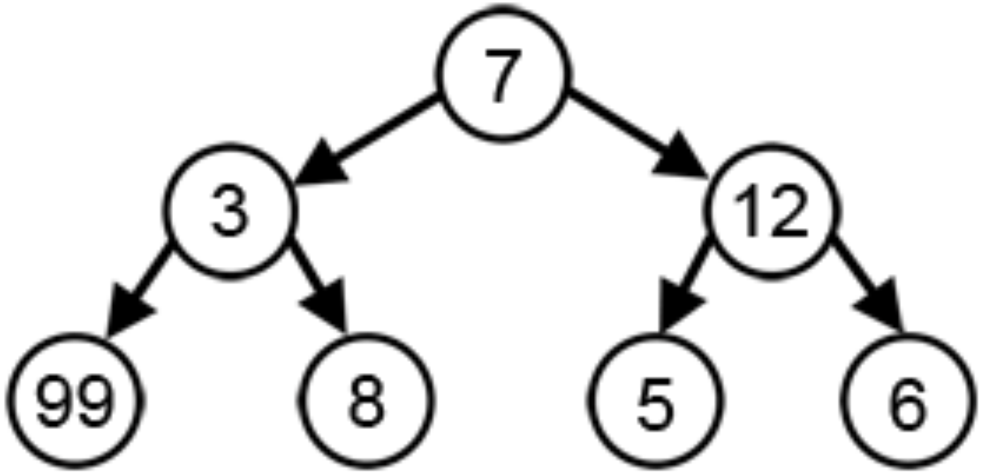
Tree/Model Optimization

Tree Optimization – Greedy Criterion

- Trees grown according to *what the local best option is*.
- Criterion: Gini, Information Gain.

Short Aside - Greedy Algorithm Example

Example: Find largest path.



Tree Optimization – Greedy Criterion

- Note: The criterion governing tree growth is *different* than your global cost function (e.g. precision-recall, accuracy, etc.), which determines how well your entire model is doing.

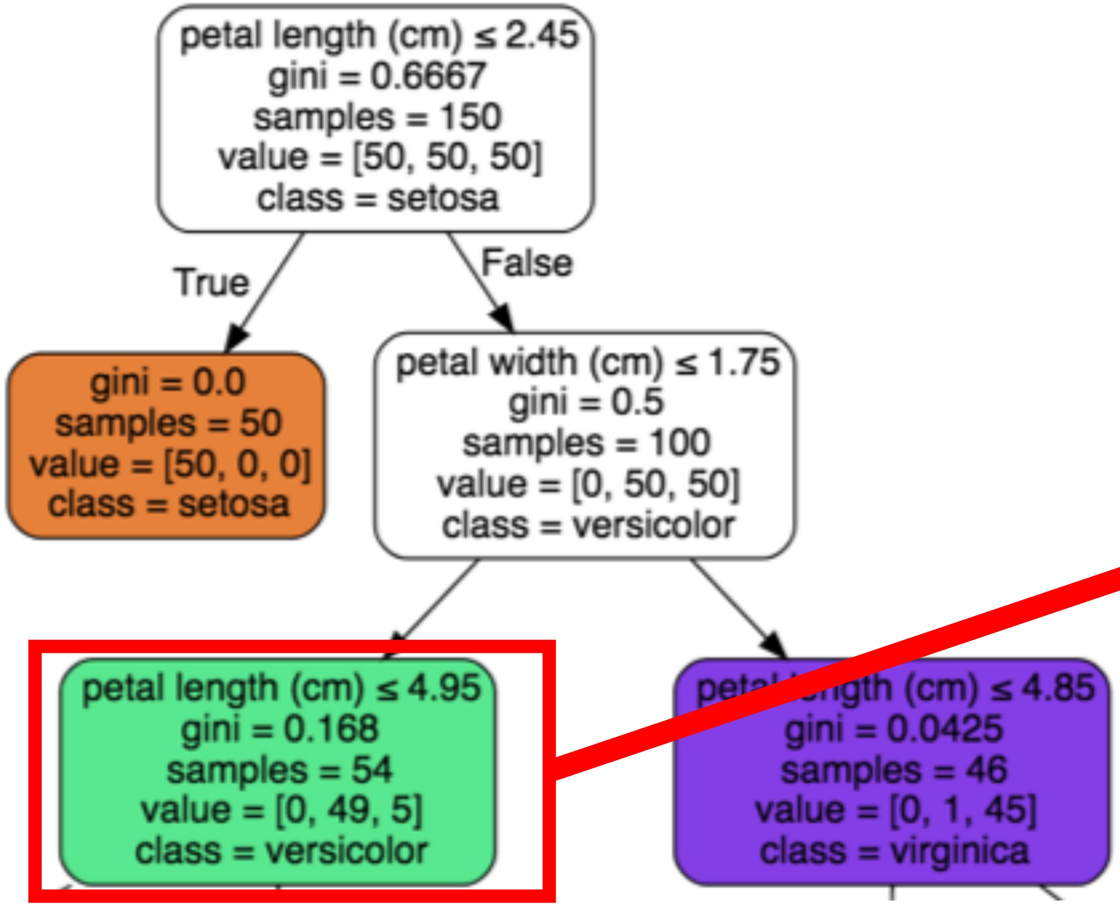
Tree Optimization – Gini Impurity

"Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset."

$$G = 1 - \sum_{i=0}^{N_{class}} \left(\frac{N_i}{T}\right)^2$$

G = Gini Impurity of node
N_{class} = total number of classes in node
N_i = Number of items in node from class i in sample
T = Total samples in node

Tree Optimization – Gini Impurity Example



$$G = 1 - \sum_{i=0}^{N_{class}} \left(\frac{N_i}{S}\right)^2$$

$$G = 1 - \left[\left(\frac{0}{54}\right)^2 + \left(\frac{49}{54}\right)^2 + \left(\frac{5}{54}\right)^2\right]$$

$$G = 0.168$$

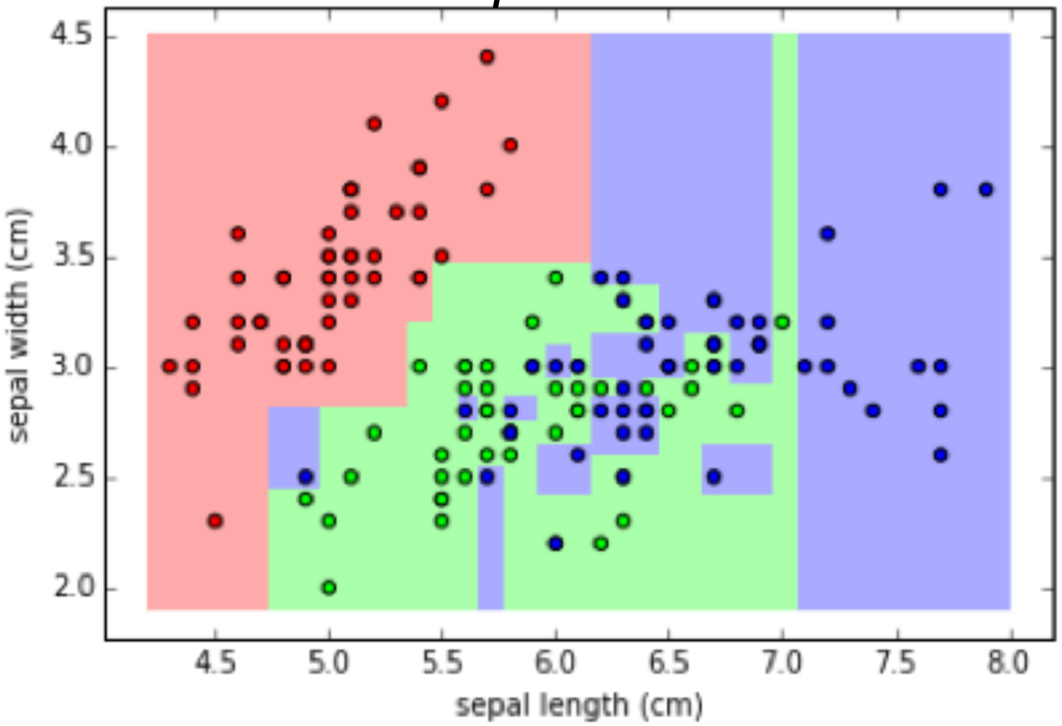
Splits decided such that the gini impurity is minimized

Model Optimization – Regularization

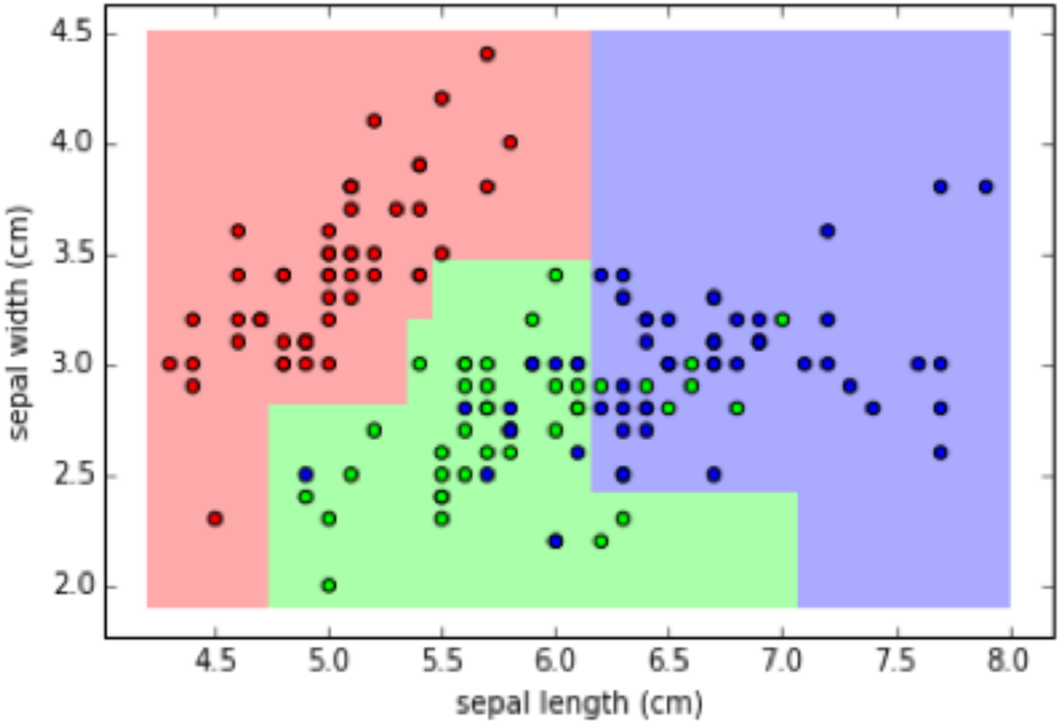
- Main complexity parameter is *max_depth* of the tree.
- Deep trees can split the data up more, leading to overfitting.

Model Optimization – Regularization

Deeper Tree



Shallower Tree



Great Demo

https://cs.stanford.edu/people/karpathy/svmjs/demo/demoforest.html

Final Nice Attributes - Proof Left as an Exercise to the Student ;)

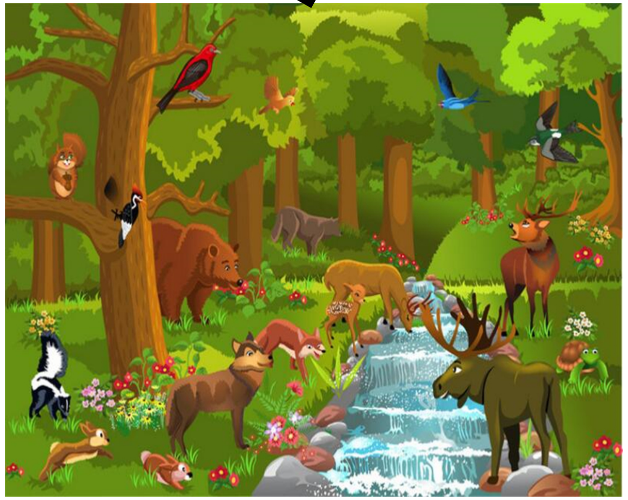
- Get Out Of Bag (OOB) error rate for free, which is equivalent to leave-one-out cross validation.
- In theory, Random Forest cannot overfit on the number of trees, i.e. you can't have too many trees, it's just more expensive to train more.
- Deals with missing values through "surrogate splits".
- Efficiently ignores redundant/irrelevant variables.

The End

Insight_Fellow_i

Watched the clock/dozed off

Paid attention the entire time



Links

- https://www2.isye.gatech.edu/~tzhao80/Lectures/Lecture_6.pdf
- http://scikit-learn.org/stable/modules/tree.html
- https://stackoverflow.com/questions/20224526/how-to-extract-the-decision-rules-from-scikit-learn-decision-tree
- http://www.utdallas.edu/~nrr150130/cs7301/2016fa/lects/Lecture_10_Ensemble.pdf
- http://www2.stat.duke.edu/~rcs46/lectures_2017/08-trees/08-tree-advanced.pdf