Decision Trees

Part 2) Bias & Variance

Prof. Dr. Jan Kirenz
HdM Stuttgart

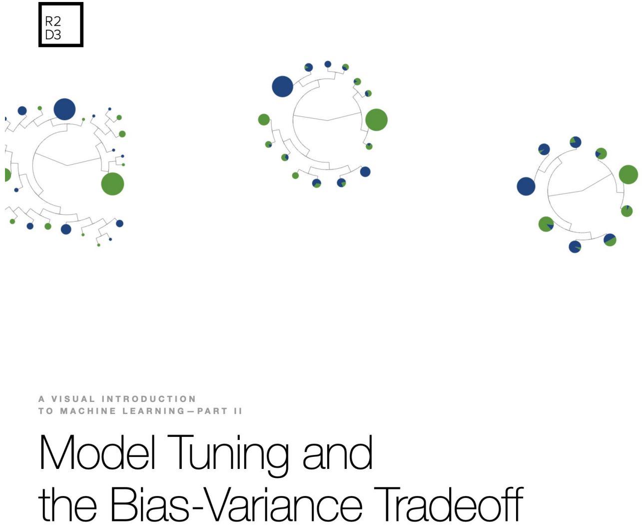
Decision trees part 2

The following slides are based on the interactive tutorial

“Model Tuning and

the Bias-Variance Tradeoff”

by R2D3



Bias-Variance tradeoff

- Goal of modeling:
    - approximate real-life situations by identifying and encoding rules in data.
- Models make **mistakes** if those patterns are
    - overly simple or
    - overly complex.

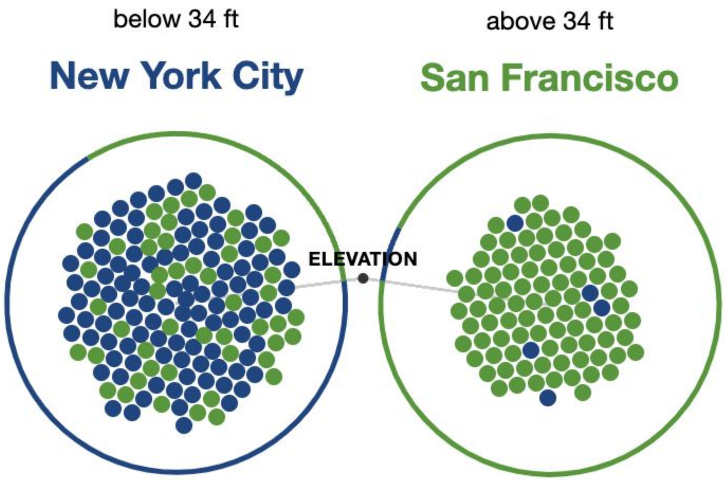
- In Part 1, we created a model that distinguishes homes in San Francisco from those in New York.
- Now, we'll talk about **tuning** and the **Bias-Variance tradeoff**.

Model parameters

- Models can be adjusted to change the way they fit the data.
- These 'settings' are called (**hyper**-) **parameters**.
- An example of a decision-tree parameter is the **minimum node size,** which regulates the creation of new splits.
  - A node will not split if the number of data points it contains is below the minimum node size.

- The tree from Part 1 had a **minimum node size of one.**
- It was very **complex**, had lots of splits, and **overfit** the data.
- To see why, let's revisit how the decision tree was trained.

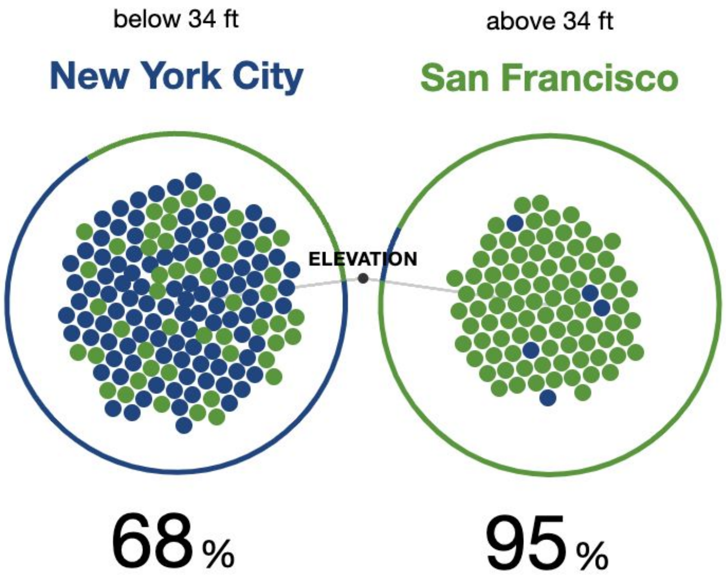
Overly simple decision tree: a stump



- The simplest version of a decision tree is called a **stump**.
- Comprised of a single split, stumps are comprised of a single rule, such as
    - “Every house whose elevation is above 34 feet is in San Francisco,

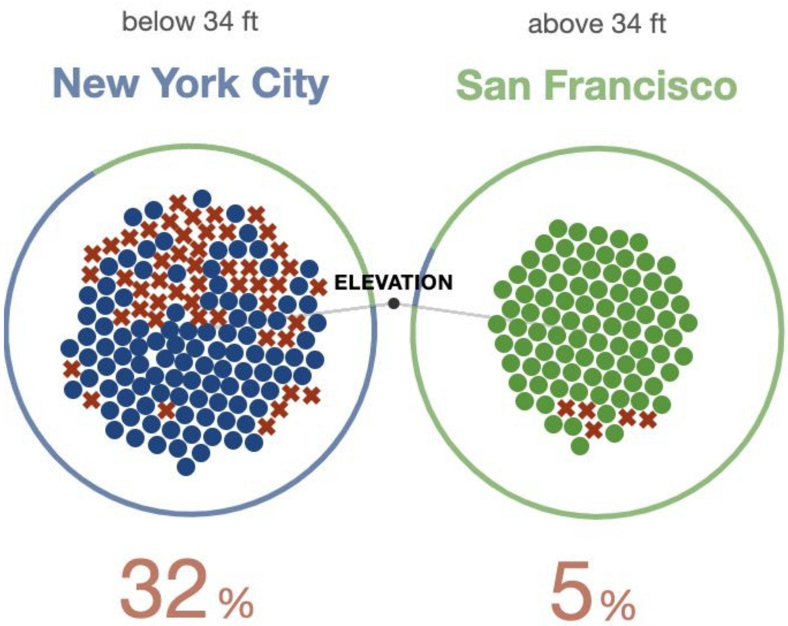
      and all others are in New York.”

Overly simple models suffer from bias



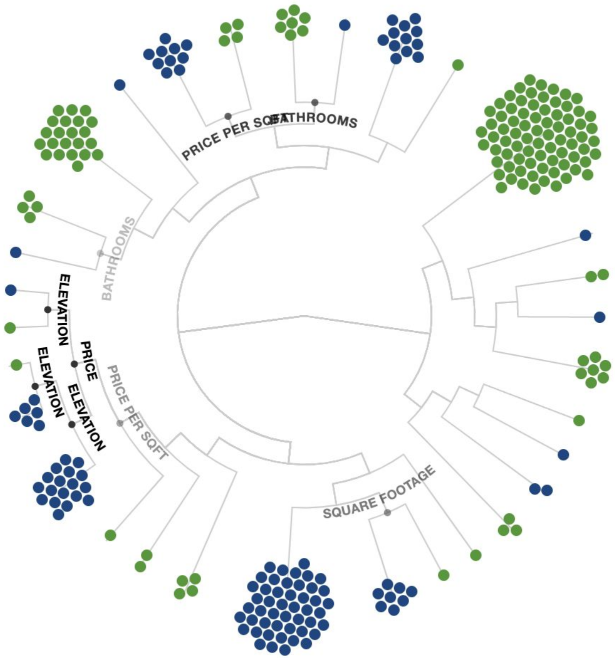
- **Stumps** take a **binary view** to the world and ignore complexity and nuance in the training data.
- This black-and-white interpretation of the world is prone to errors due to **bias**.

Overly simple



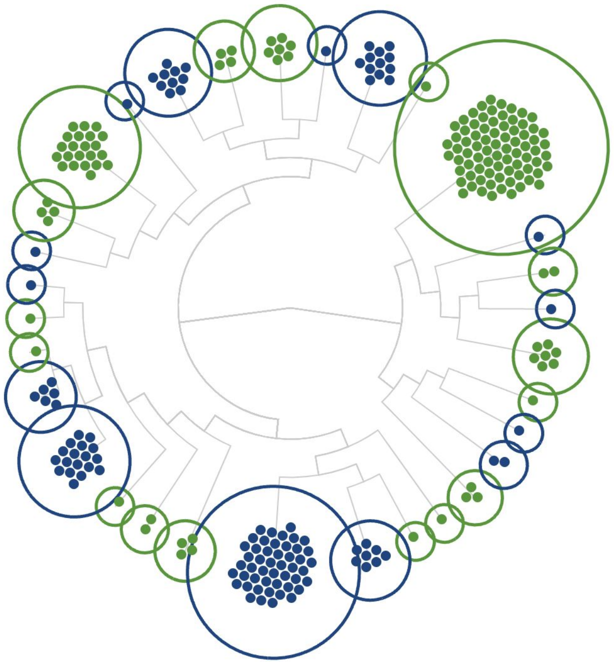
- A model with **too much bias** systematically ignores relevant details and is **wrong** in **consistent** ways.
- The stump incorrectly classifies all lower-elevation homes in San Francisco.

A decision tree with many splits



- To decrease the error due to bias, you can add **additional splits** to the tree

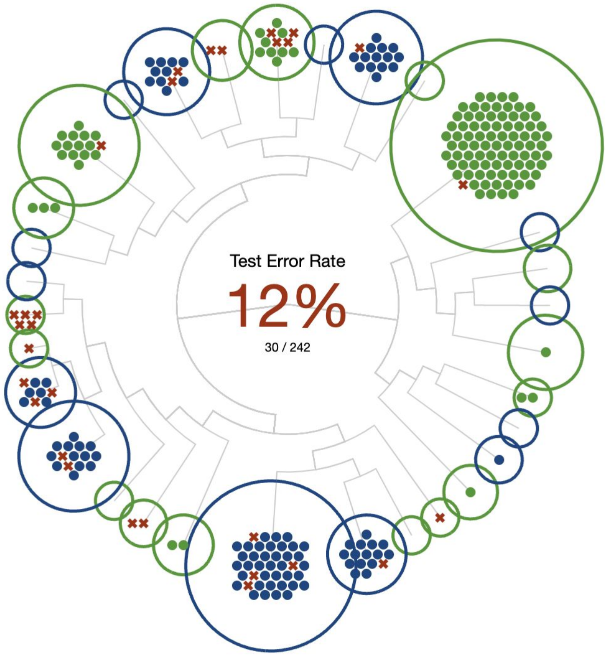
Overly complex?



- Additional splits allow the tree to take into account **more complexity**.
- You can add splits until a tree's leaf nodes contain only homes in either San Francisco or New York.

The question is,
how does it
perform on the
test data?

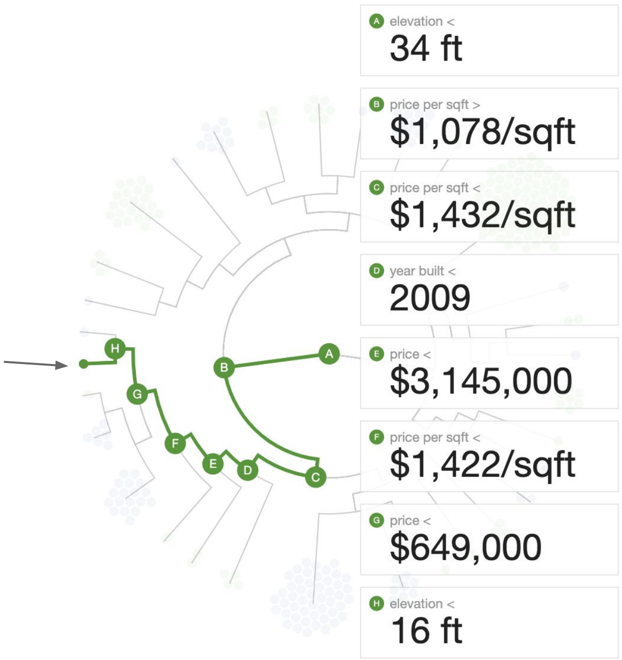
Overly complex: **high variance**



Test-error rate: 12%.

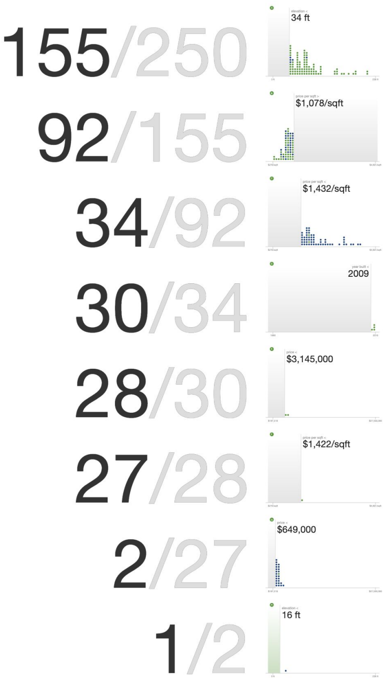
- **Overly-complex** trees suffer from errors due to **variance**.
- **High-variance** models make mistakes by **overfitting** to the idiosyncrasies of the training data.
- They tend to be **wrong** in **inconsistent ways**

A tangible example of variance



- Follow the creation of a **single leaf node**:
  - This leaf node is the result of eight separate forks (A to H).
  - Each fork divides the data set into **smaller subsets**, until the leaf node contains *a single San Francisco home*.

Overfitting leads to bad generalization



- If **terminal nodes** were made using **very little data**
- It's no surprise that the **generalizations** they make are **incorrect**.
- **Patterns** drawn from two homes are more likely to be flukes than anything real

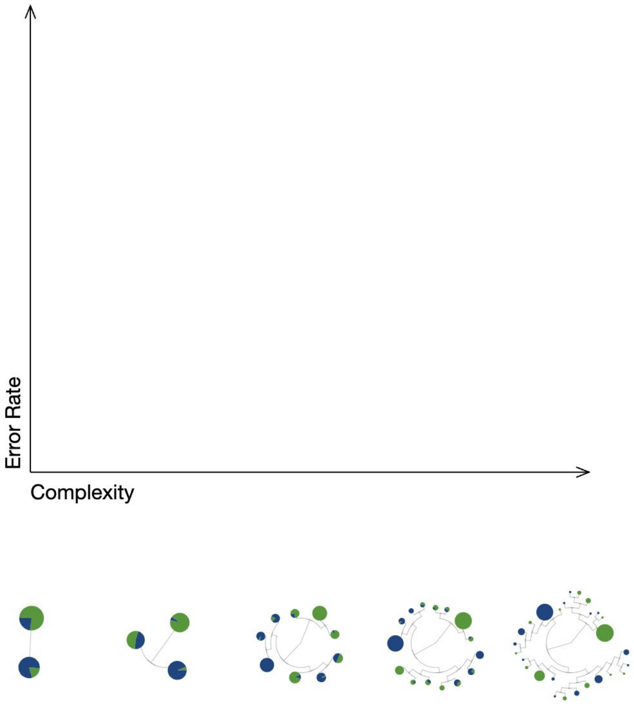
How to address overfitting?

- We could impose **limits** on how a **tree grows** by changing the **minimum-node-size threshold.**

As the **minimum-node-size threshold increases**, there are fewer splits.

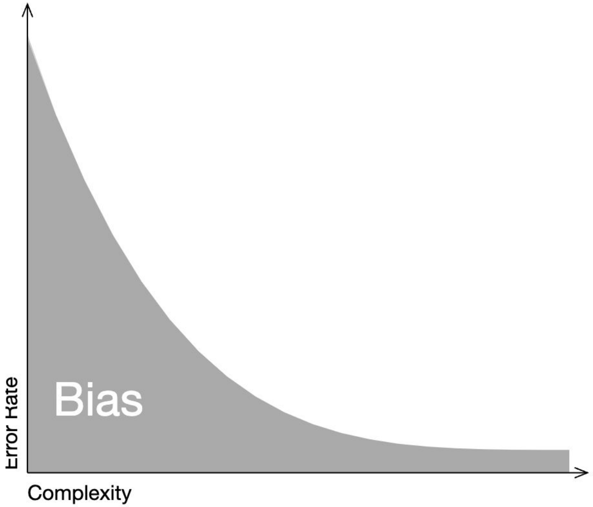
The trees get **less bushy**.

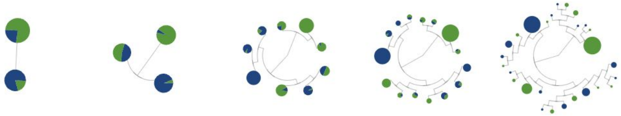
Model complexity and model error



- The relationship between a parameter like **minimum node size** and **model error** illustrates the tradeoff between **bias and variance** more explicitly.

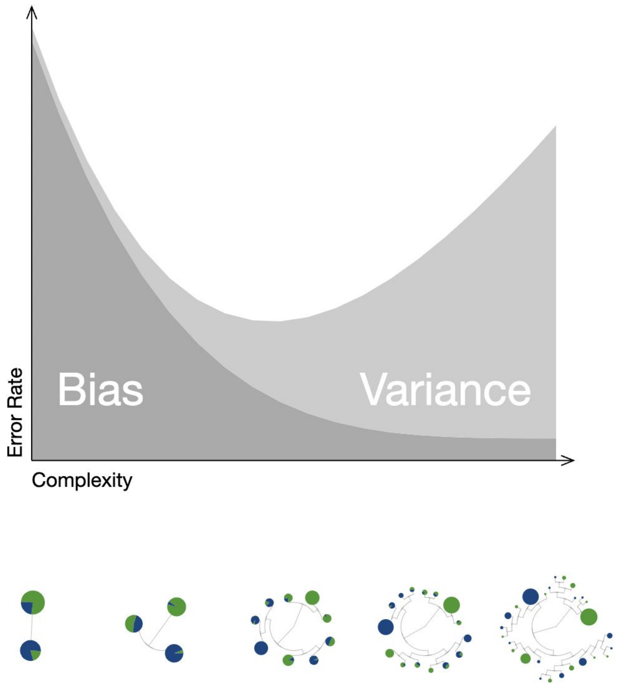
Model with low complexity and high **bias**





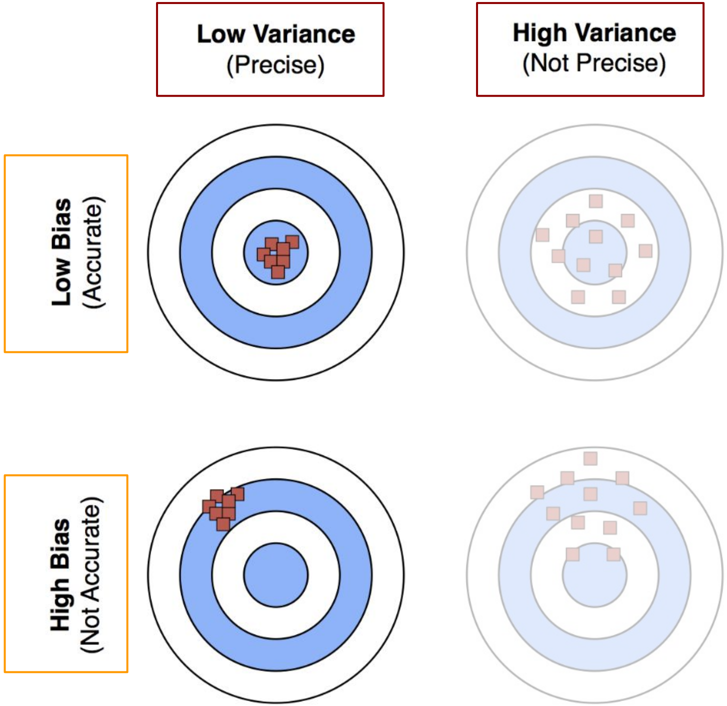
- When a model is **less complex**, it ignores relevant information, and **error due to bias is high.**
- As the model becomes **more complex**, error due to **bias decreases**.

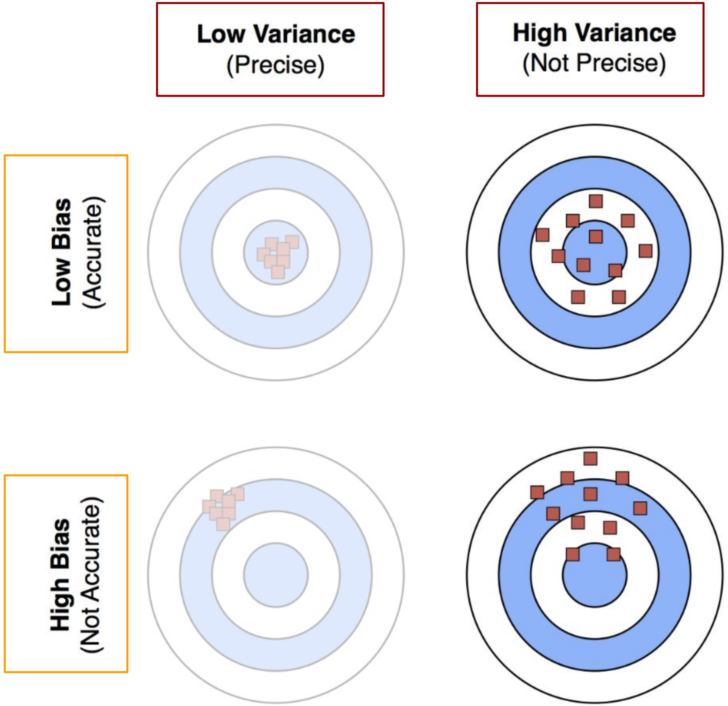
Bias VS variance



- When a model is **less complex**, error due to **variance is low.**
- Error due to **variance increases** as **complexity increases**.

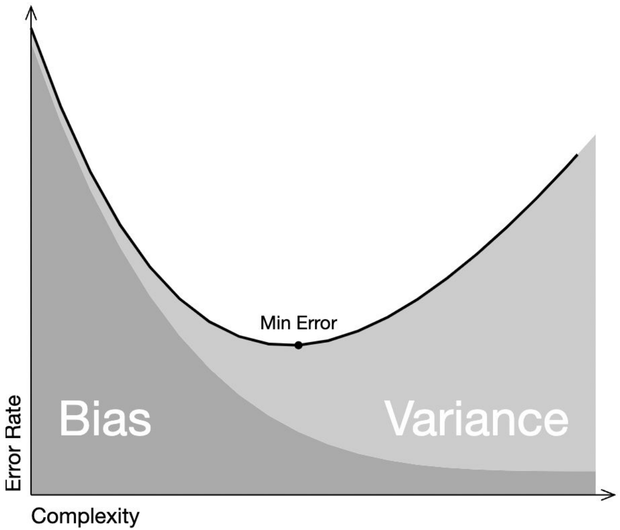
Bias and variance

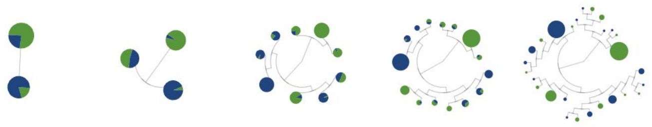




Source: Raschka, S. (2016): [Model evaluation, model selection, and algorithm selection in machine learning]

Overall model error





- **Overall model error** is a function of error due to
    - **bias** plus error due to **variance**.
- The ideal model **minimizes error from each.**

Single decision trees are not ideal

- Even at their **optimal depth,** single decision trees aren't the best performing models.
- While trees are very **easy to understand**, the world is more complex than a bunch of **if-then statements.**
- Nevertheless, decision trees can be used in **aggregate** (as so called **ensembles**) to yield very strong results.