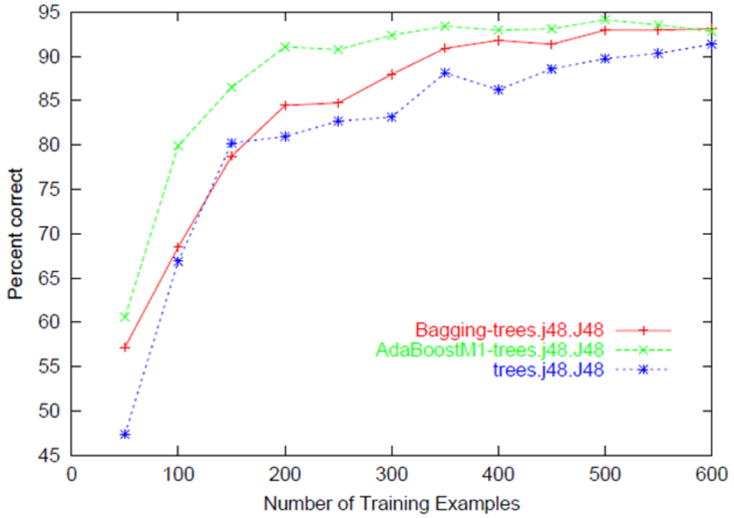
Frequent problems in ML

DATA ANALYTICS | IRONHACK

INSUFFICIENT DATA

- How much data is enough?
- **Learning curves** answer this question. They plot how an error metric improves with the dataset size.
- At some point, the performance of the model will stabilize showing that the dataset size is big enough.
- Even so, you may have error (bias) because of lack of relevant data in your dataset.



HARDWARE LIMITATIONS

- Can your computer train a model with 1,000,000 observations and 25 features? **Try it! (save everything important first)**

- Consider cloud computing

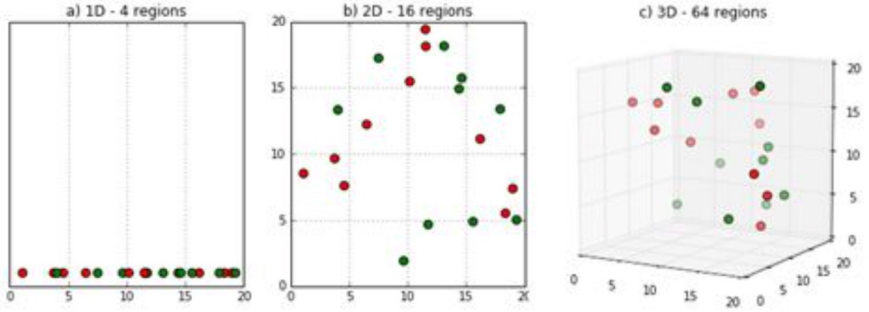
THE CURSE OF DIMENSIONALITY (I)

- We want as much information as possible for our algorithms to find patterns, but more attributes mean:
  - More training time
  - Difficult interpretation
  - Possibly worse performance

- Problem common in:
  - Computer vision (each pixel is a data point)
  - NLP (each word + each pair of words + each triplet of words + …)

THE CURSE OF DIMENSIONALITY (II)

- Problem common in:
  - Computer vision (each pixel is a data point)
  - NLP (each word + each pair of words + each triplet of words + ..
  - High dimensionality = lower density
  - With more “regions” and the same number of observations, we have fewer observations per region



BIAS VS VARIANCE TRADEOFF

- **Bias:** the difference between the average prediction and the real value.
- **Variance:** how much variability our model predictions have.
- Having both small is not possible, we need to find a tradeoff.

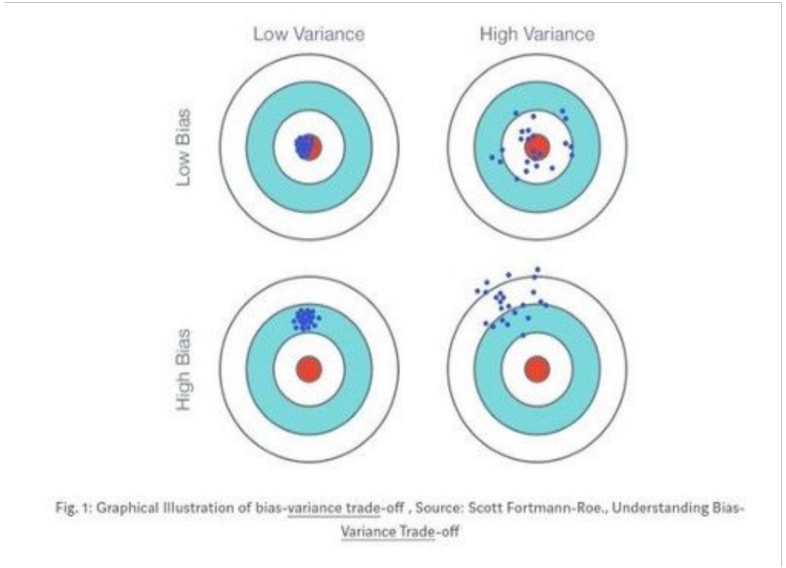
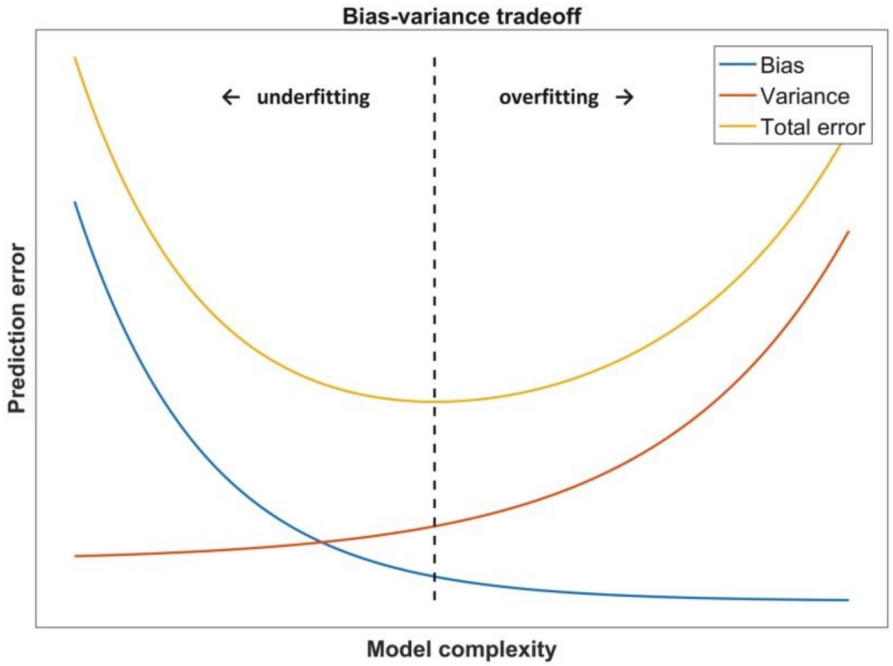


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

BIAS VS VARIANCE TRADEOFF



THANKS !