What is a random forest:

- universal ML prediction technique
  - ↳ categorical (dog/cat)
  - ↳ continuous (sales price).
- handle many different types of data
  - pixels
  - zip
  - revenue
  - ordinal data.
- scalable – almost linearly with the number of CPUs.
- generally doesn't overfit
- generally doesn't need a validation set
  - ↳ can glean how much the random forest generalizes.
- few statistical assumptions/dependencies.

Curse of Dimensionality

- the more columns you have, the more probable the matrix will have empty values (sparse)

* additionally the more dimensions → the higher likelyhood data will be "on the edge" of that space. because there is a higher probability of data being ommitted than having various combinations of the data when you have it.
  - ↳ distance is overvalued because all the data is predisposed to be close together.
    - ↳ for columns/features which cause this edge affinity → to use them will deteriorate your performance.
      - ↳ According to FastAI this is false.
        - distance is relative → and will still work.

No Free Lunch Theorem

Claim: there is no model that works well for all datasets.

In real life data is made on a lower dimensional manifold — the data is produced by some causality and has some inherent relationships and qualities. It is not random. Thus in practice certain technique work better than others.