In [Vapnik–Chervonenkis theory](https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_theory" \o "Vapnik–Chervonenkis theory), the **Vapnik–Chervonenkis (VC) dimension** is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification) [algorithm](https://en.wikipedia.org/wiki/Algorithm" \o "Algorithm).It was originally defined by [Vladimir Vapnik](https://en.wikipedia.org/wiki/Vladimir_Vapnik) and [Alexey Chervonenkis](https://en.wikipedia.org/wiki/Alexey_Chervonenkis" \o "Alexey Chervonenkis).[[1]](https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_dimension#cite_note-vc-1)

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

\*  Partition into (equi-depth) bins:

      - Bin 1: 4, 8, 9, 15

      - Bin 2: 21, 21, 24, 25

      - Bin 3: 26, 28, 29, 34

\*  Smoothing by bin means:

      - Bin 1: 9, 9, 9, 9

      - Bin 2: 23, 23, 23, 23

      - Bin 3: 29, 29, 29, 29

\*  Smoothing by bin boundaries:

      - Bin 1: 4, 4, 4, 15

      - Bin 2: 21, 21, 25, 25

      - Bin 3: 26, 26, 26, 34

**Bias Varience**

The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn the algorithm should achieve good prediction performance.

You can see a general trend in the examples above:

* **Linear** machine learning algorithms often have a high bias but a low variance.
* **Nonlinear** machine learning algorithms often have a low bias but a high variance.

**Application**

Image Recognition

### Speech Recognition

### Medical Diagnosis

**Photo tagging Applications:**

**Spam Detector:**

**Web Search Engine:**

ISSUSESS

## Data Quality

## Transparency

## Manpower

## Manpower

1. The data needs preprocessing. Often the data comes from different sources, has missing data, has noise.
2. Figure out exactly what you are trying to predict.
3. Figure out what assumptions can be safely made about the data and the underlying system.