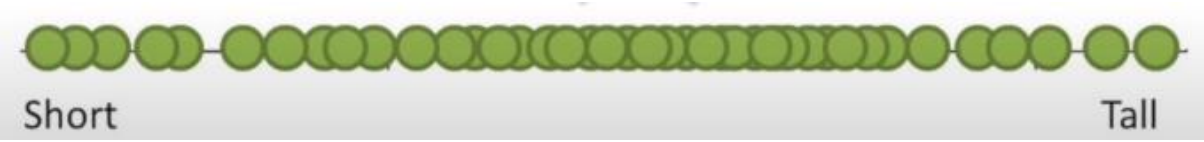
Histograms

Imagine we went out and measured height of someone.

And the we measured someone else

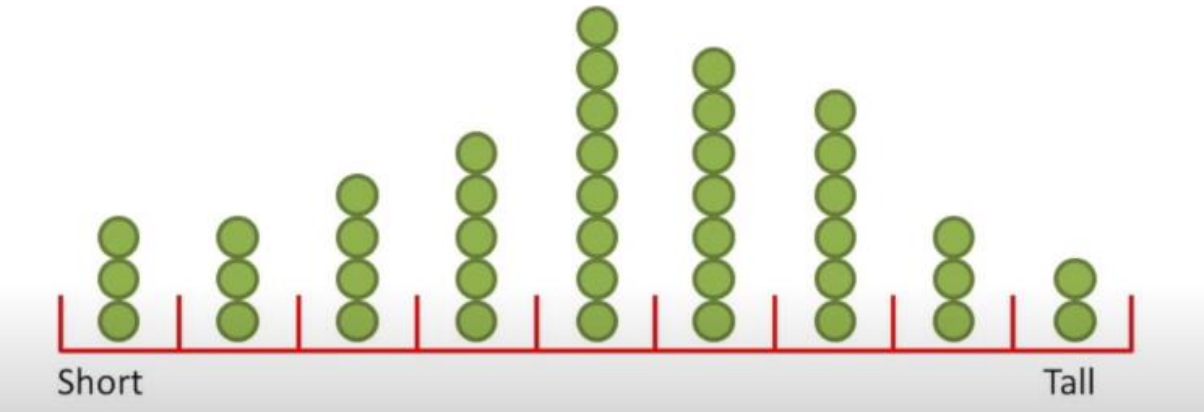
And the we measured whole bunch of people, we've measured so many people that the dots overlap, some dots are completely hidden.



We could try to make it easier to see the hidden measurements by stacking any that are exactly same.

But the measurements that are the exactly same are rare, and a lot of hidden details are still hidden.

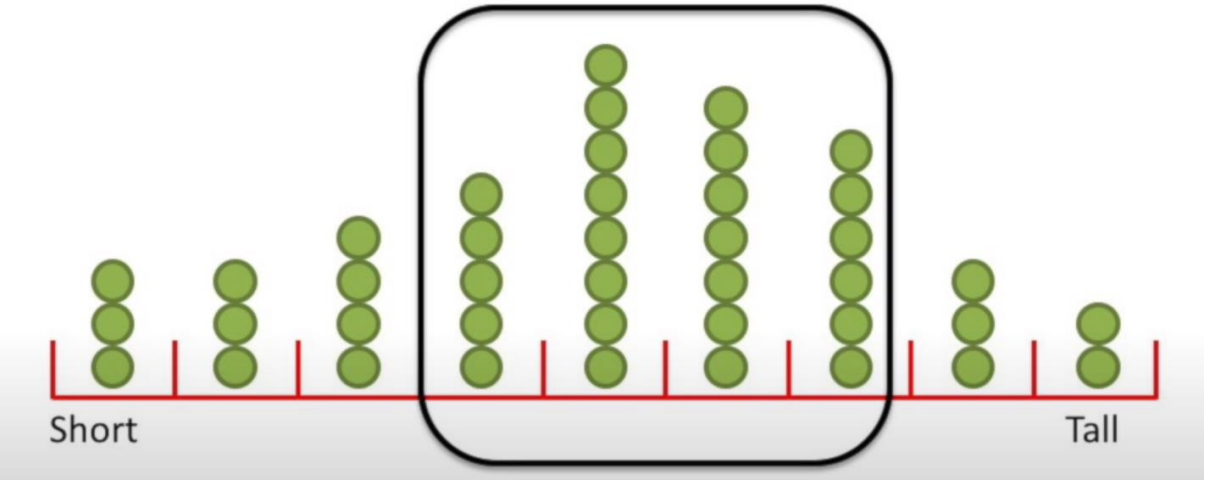
So, instead of stacking measurements that are exactly same, we divide the range of values into bins and stack the measurements that fall into same bin.



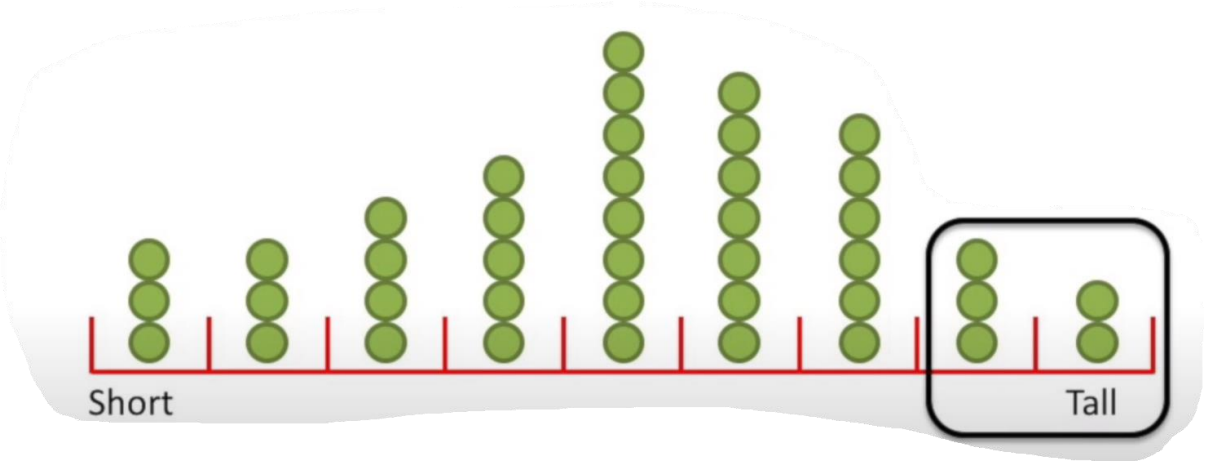
This is a **histogram.**

We can use the histogram to predict the probability of getting future measurements.

I would be willing to bet that the next measurement we make is somewhere in this range.

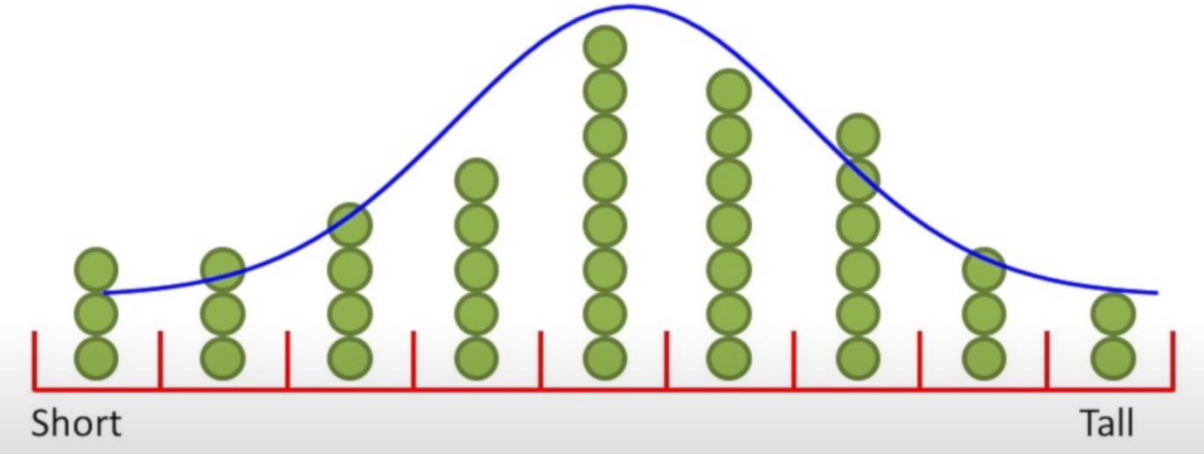


Measurements out here are rare, and less likely to happen in the future.

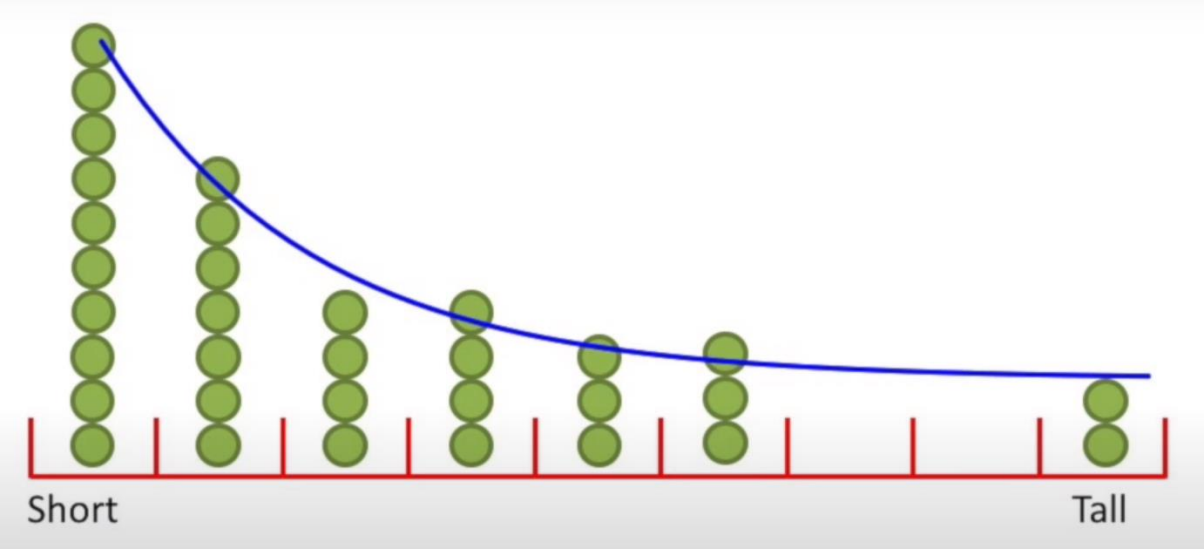


If you want to use a **distribution** to approximate your data (or future measurements). Histograms are a good way to justify your decision.

In this case, we might use a normal distribution to approximate this data and future measurements.



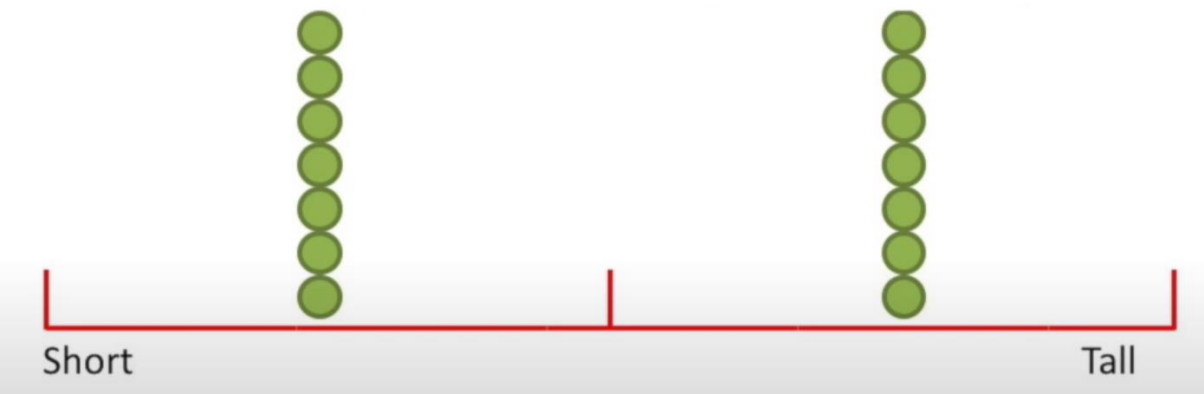
For a data like this we might use an exponential distribution to approximate this data and future measurements



Note: Figuring out how wide to make the bins is tricky

If bins are too narrow, then they are not much help.

And if the bins are too wide, they are not much help.



Sometimes you have to try a bunch of different bin widths before you get a clear picture. In other words, don't rely on the default setting of whatever program you are using to draw the histogram