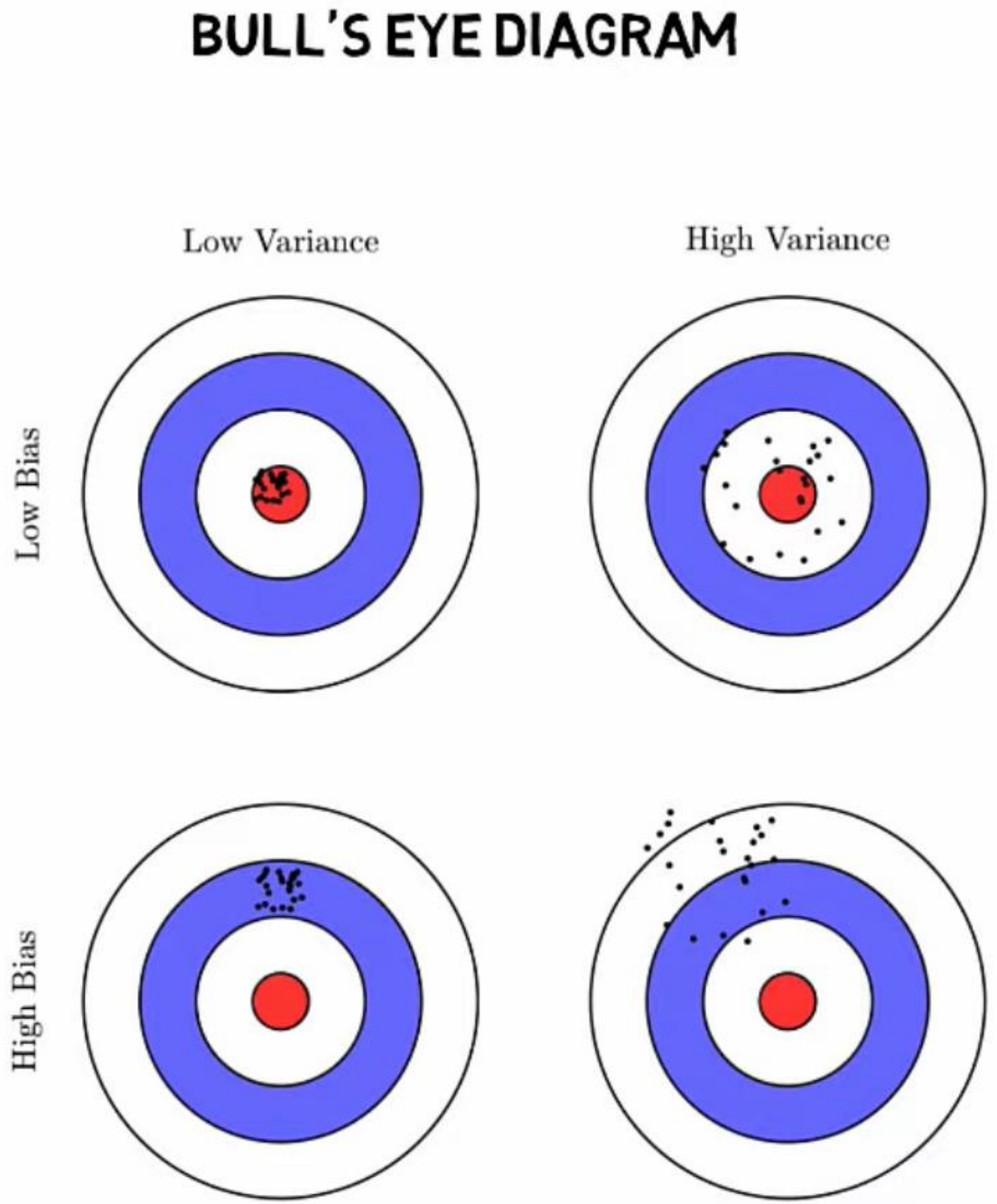
Unione europea
Fondo sociale europeo

RegioneEmilia-Romagna

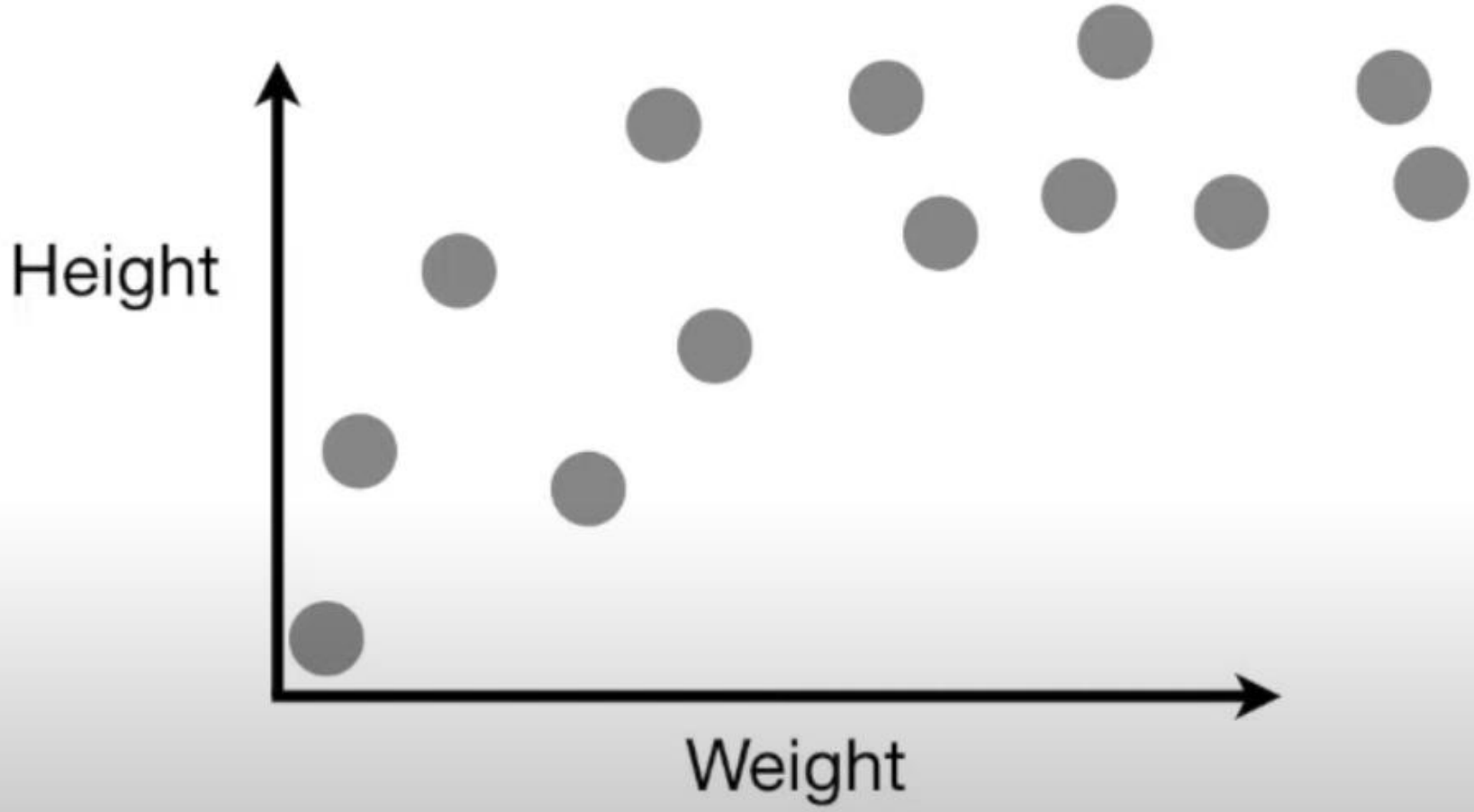
RETE POLITECNICA
La tecnica per crescere

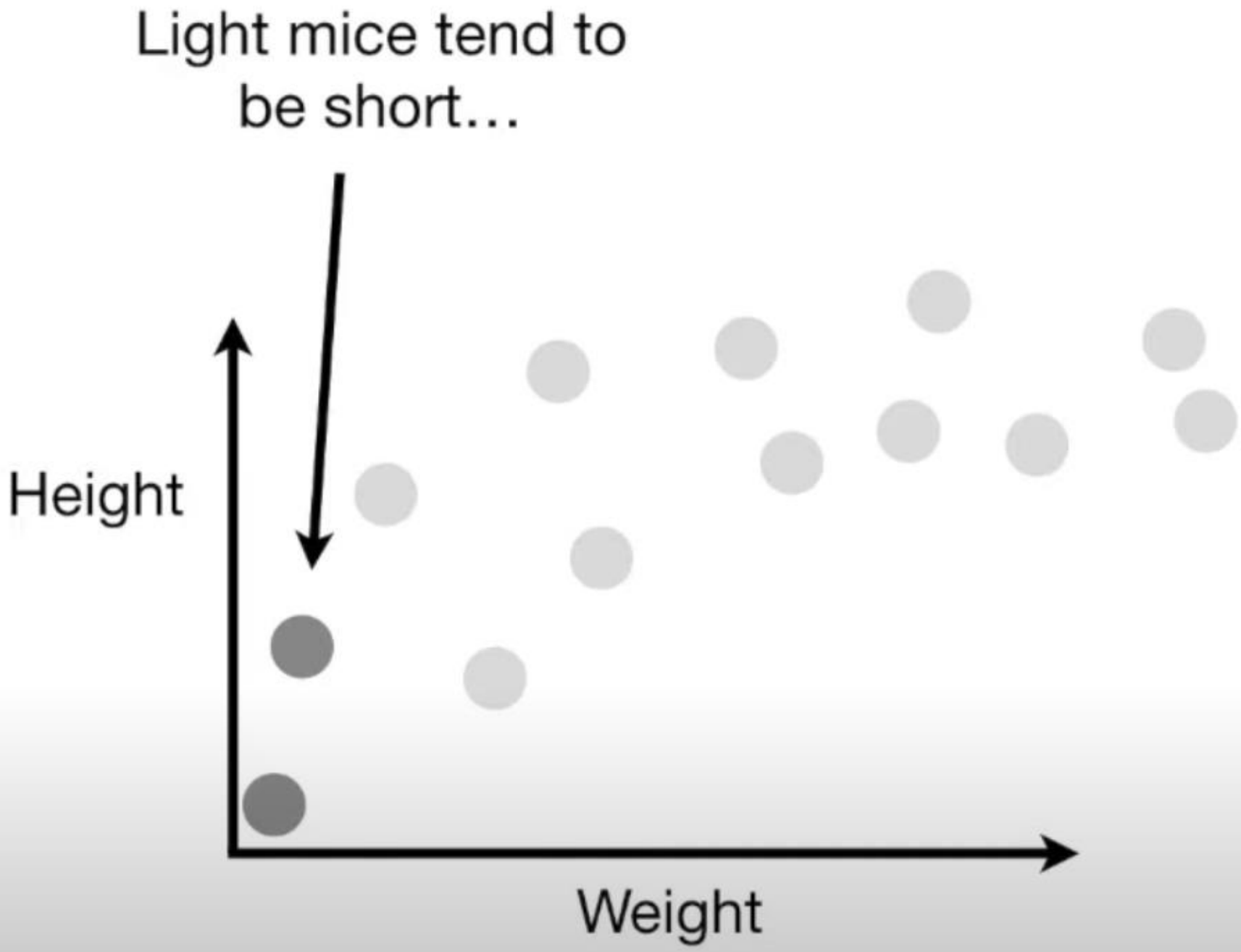
Materiale didattico per partecipante al corso "**TECNICO ESPERTO NELL'ANALISI E NELLA VISUALIZZAZIONE DEI DATI**" – Rif.P.A. 2021-15998/RER – approvata con DGR n. 1263 del 02/08/2021 di IFOA – Istituto Formazione Operatori Aziendali

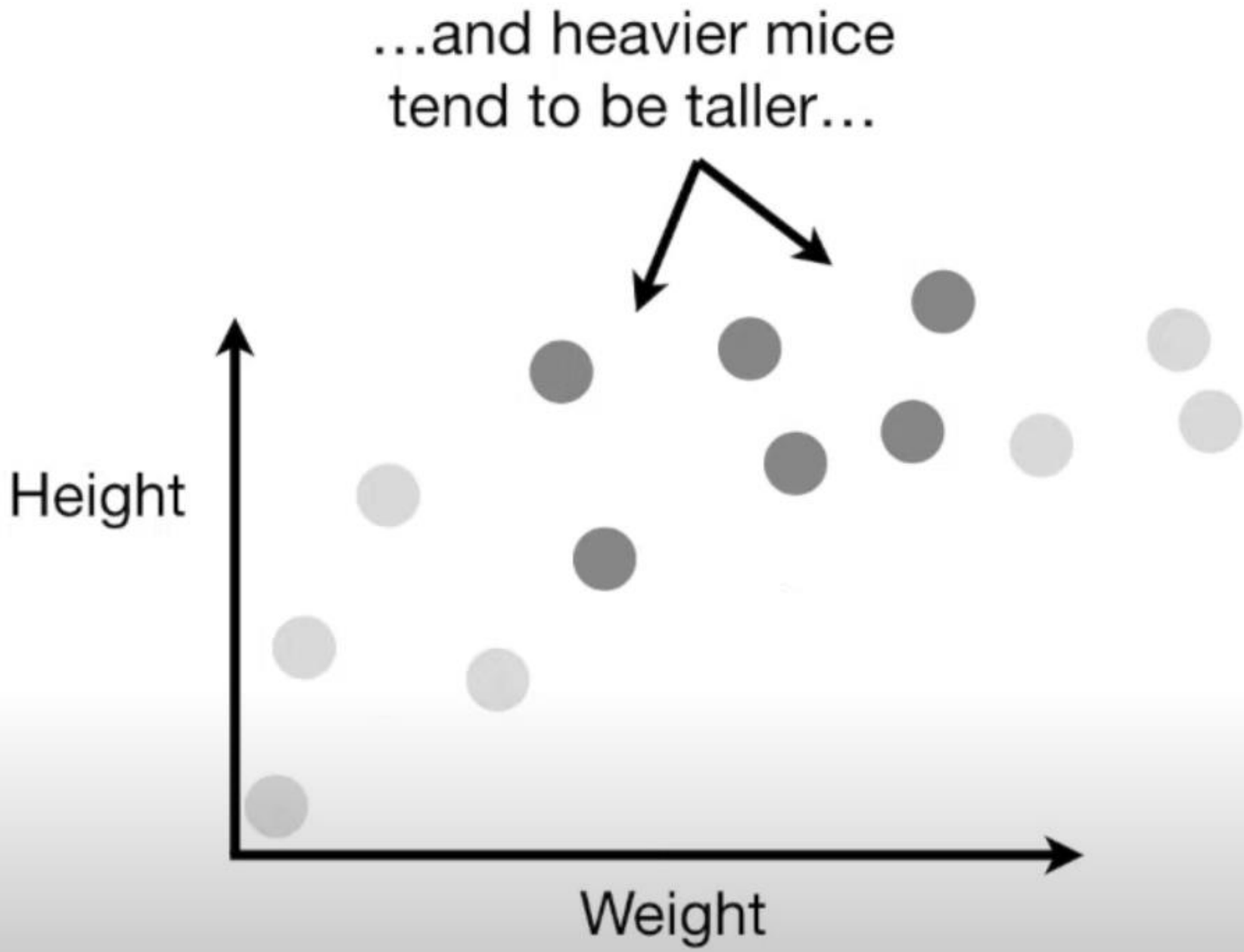
Bias vs Variance



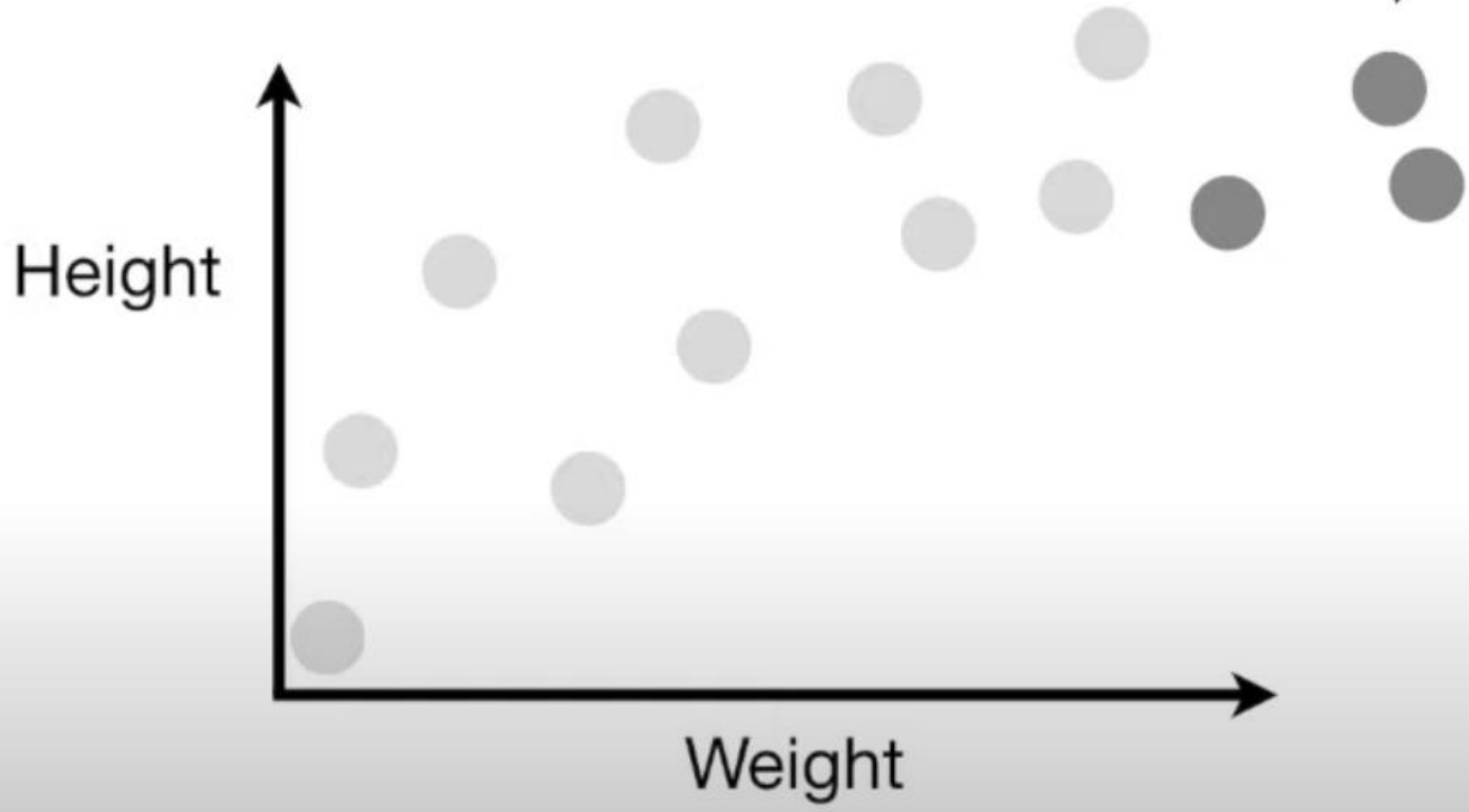
Imagine we measured the weight and height of a bunch of mice and plotted the data on a graph…



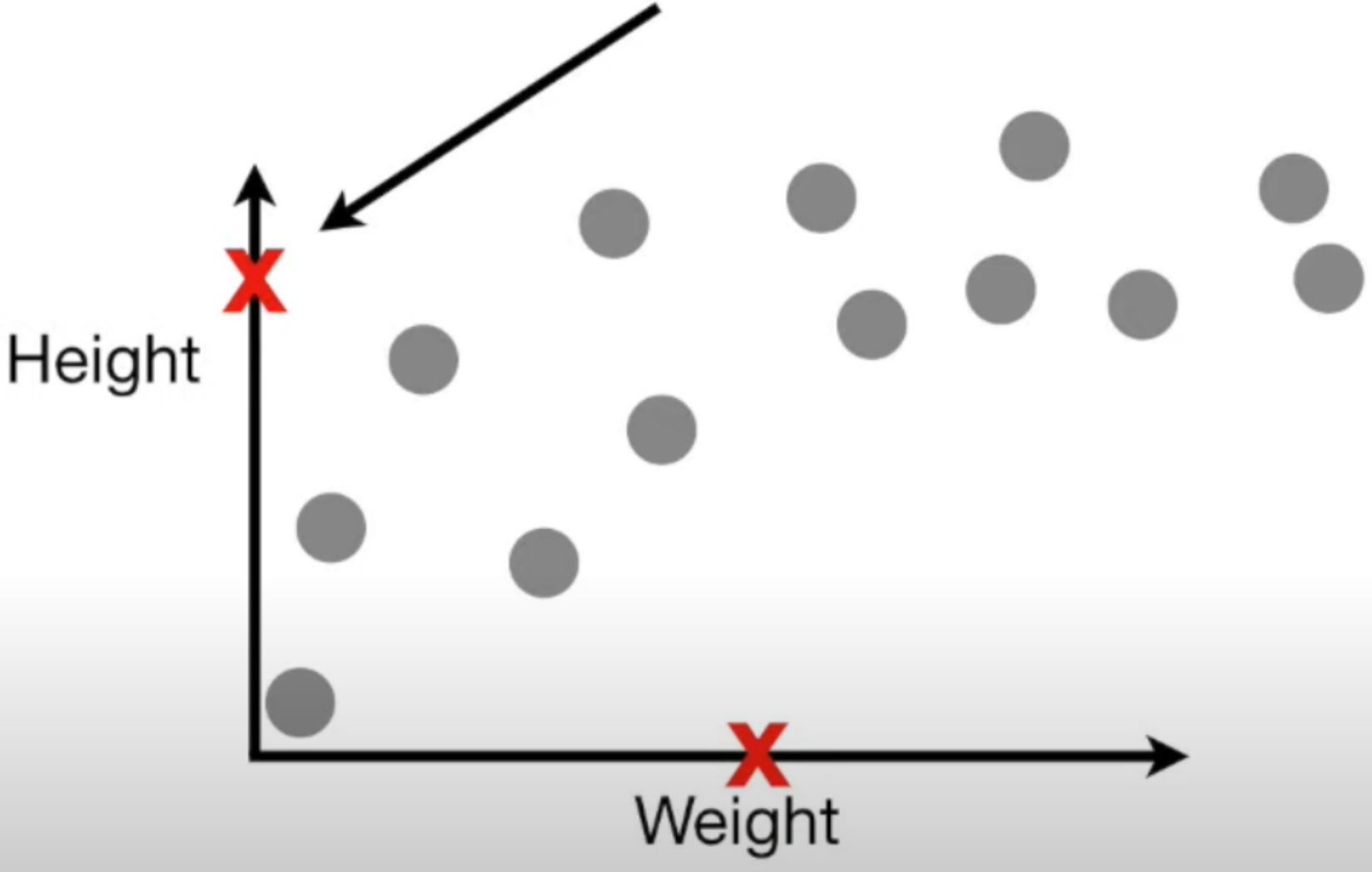

Light mice tend to be short…
Height
Weight


…and heavier mice
tend to be taller…
Height
Weight

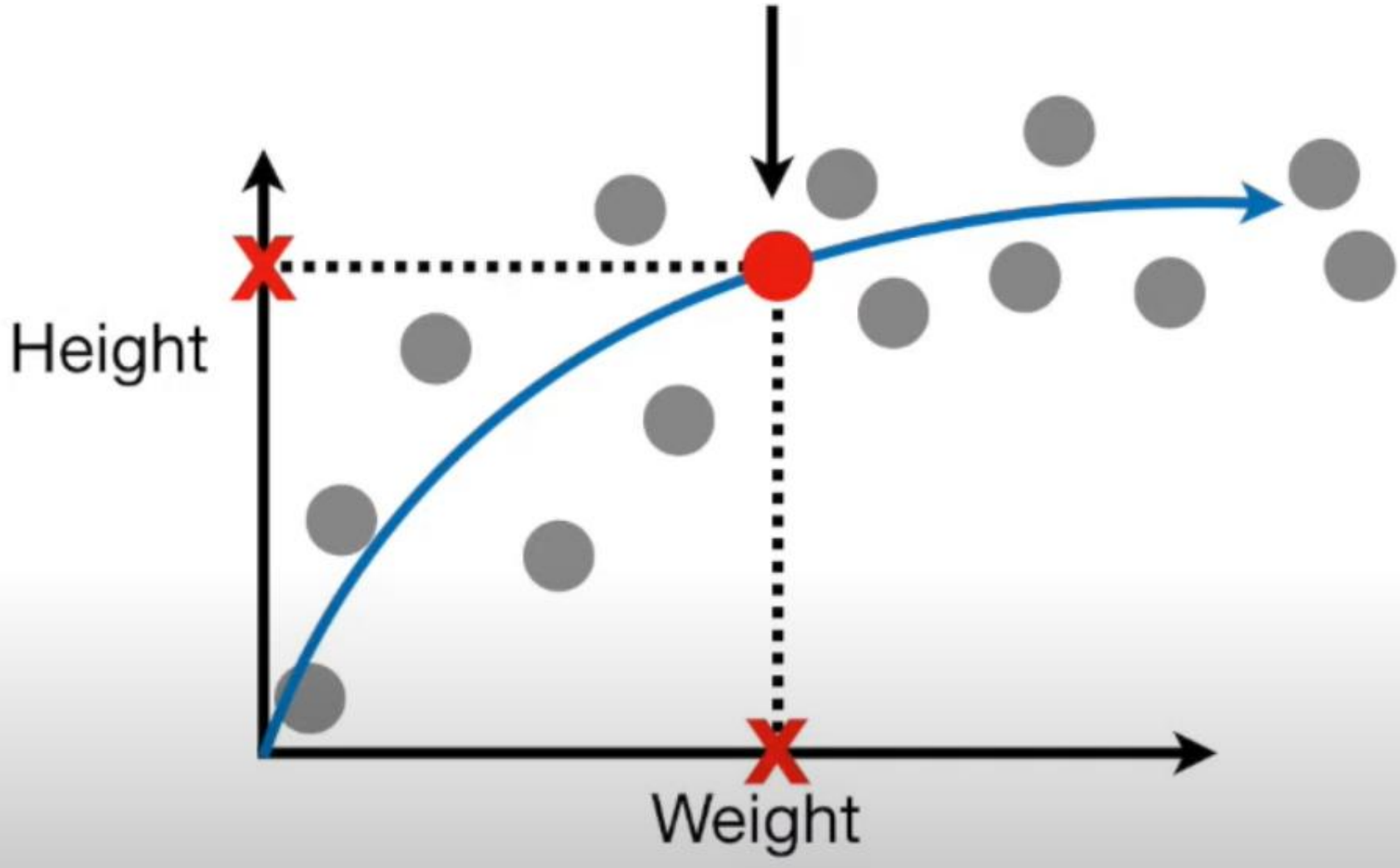
…but after a certain weight, mice don't get any taller, just more obese.



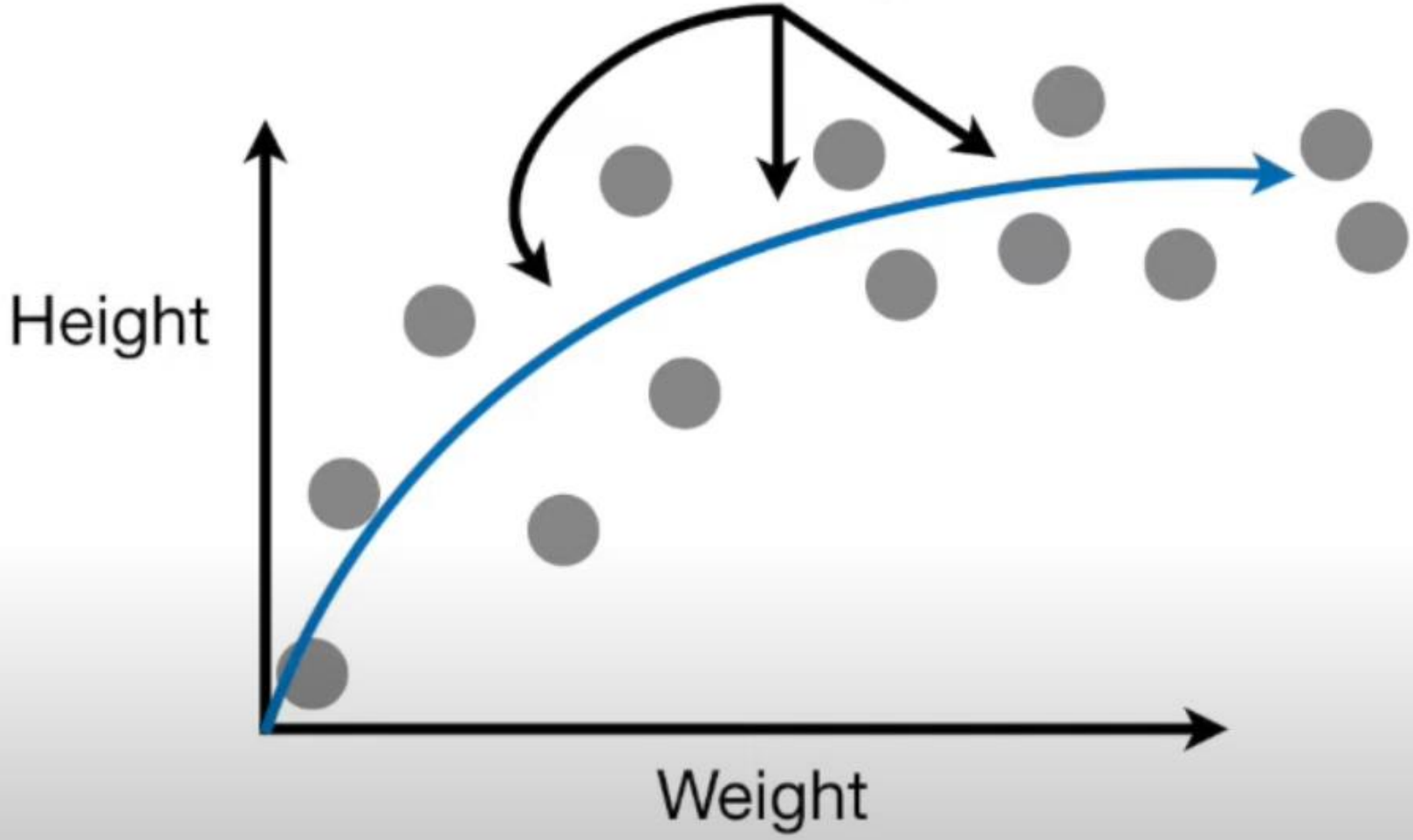
…then we might predict that the mouse is this tall….



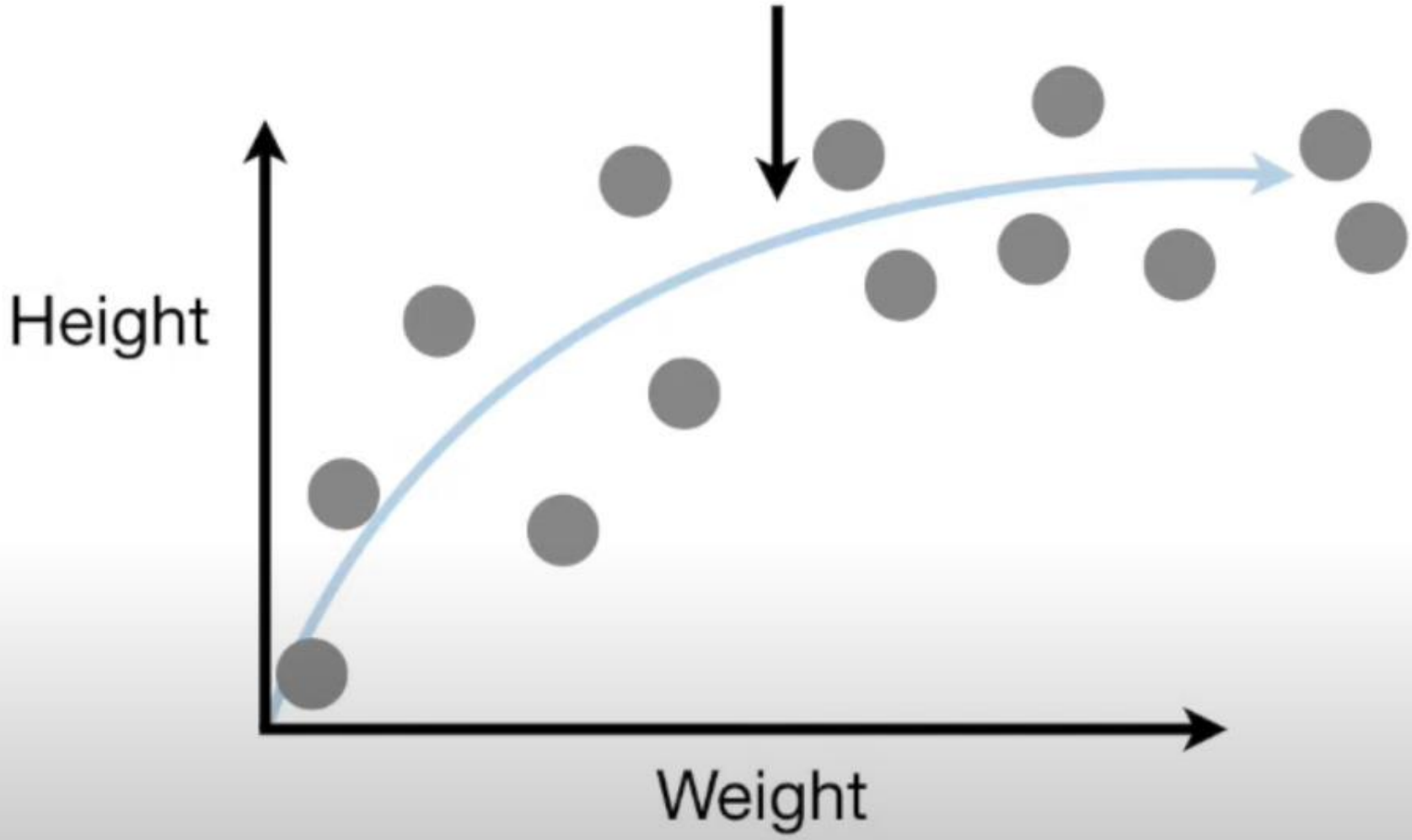
Ideally, we would know the exact mathematical formula that describes the relationship between weight and height…



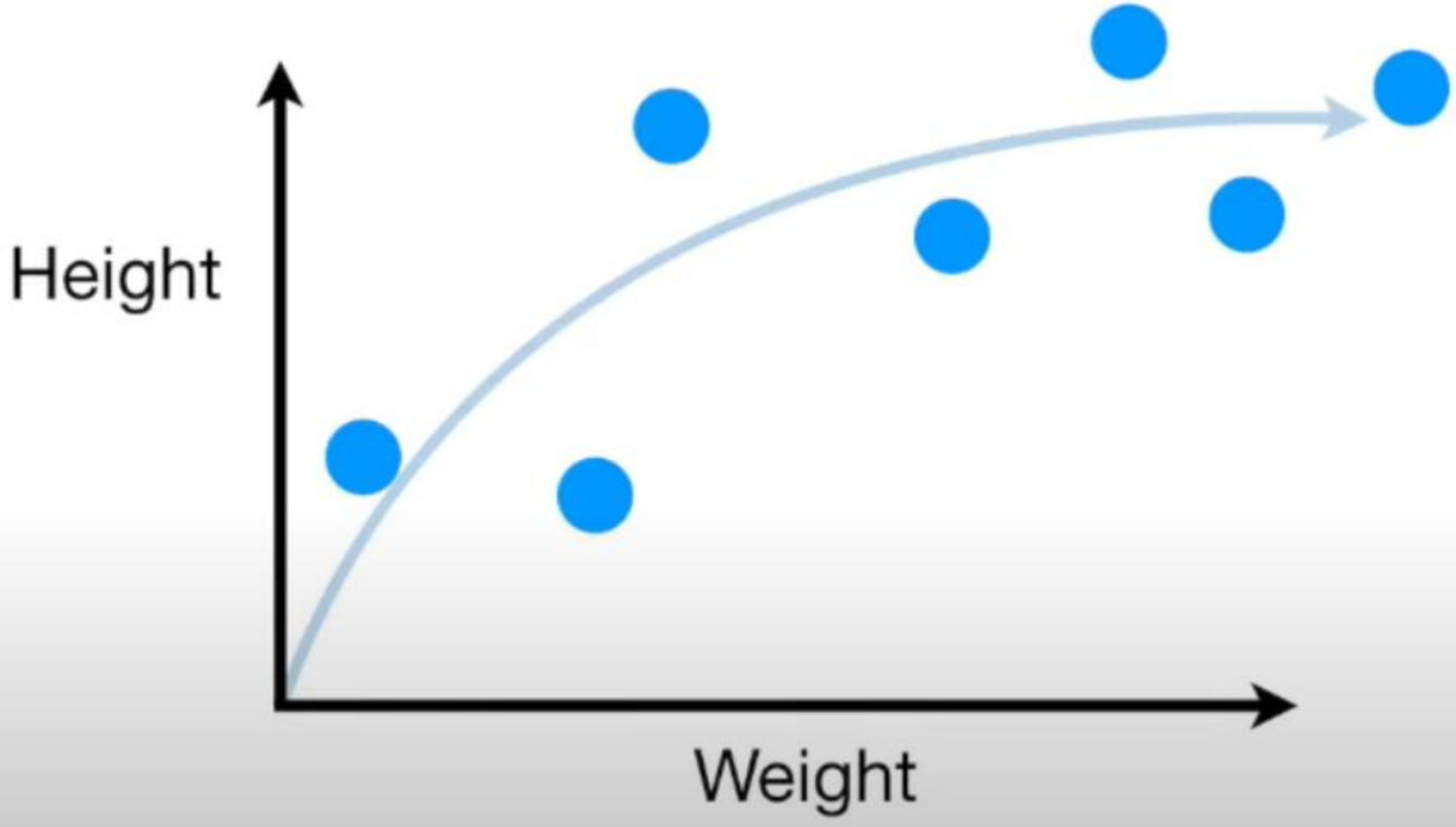
…but, in this case, we don't know the formula, so we're going to use two machine learning methods to approximate this relationship.



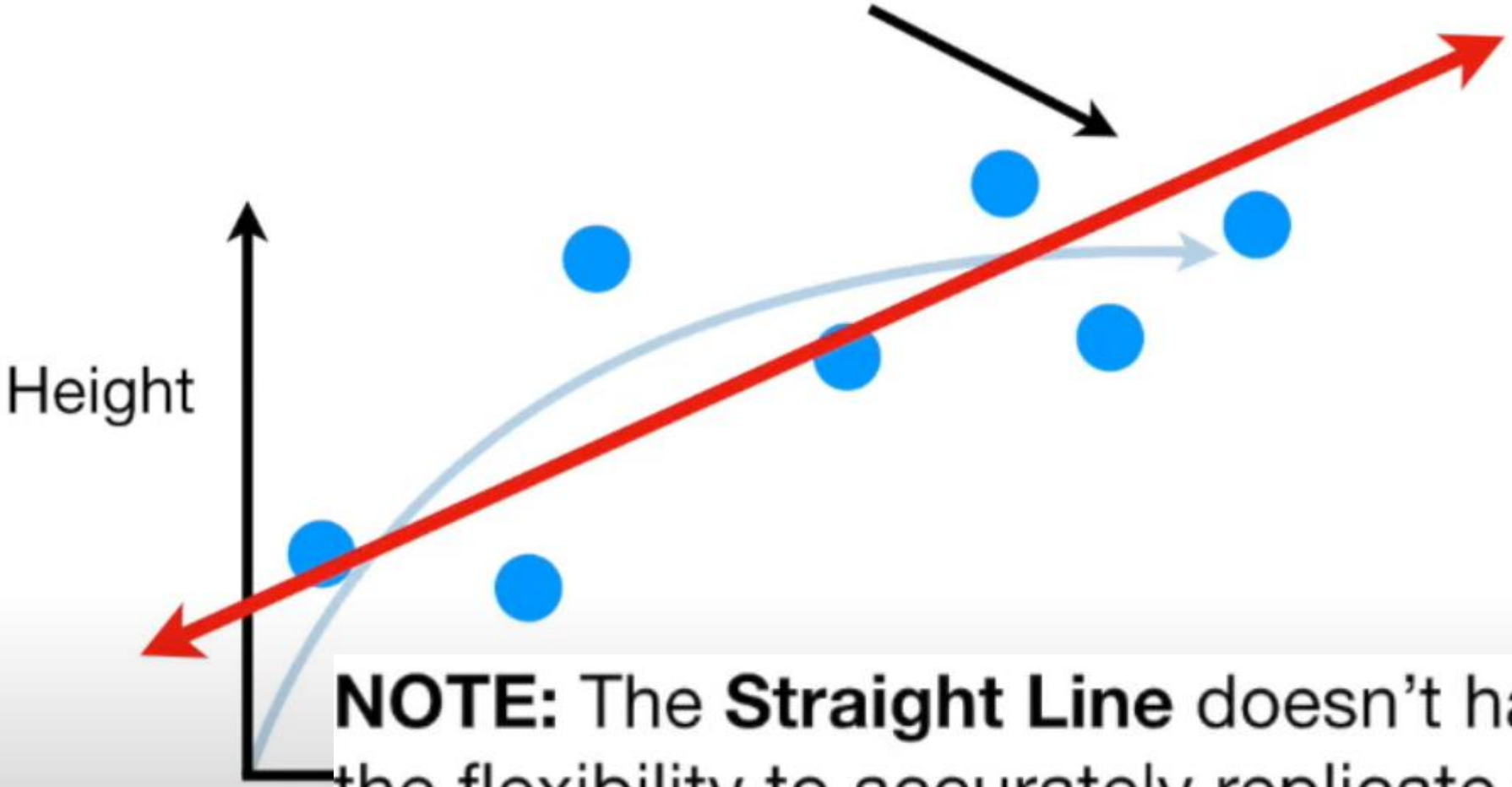
However, I'll leave the "true" relationship curve in the figure for reference.



Here is just the **training set**…

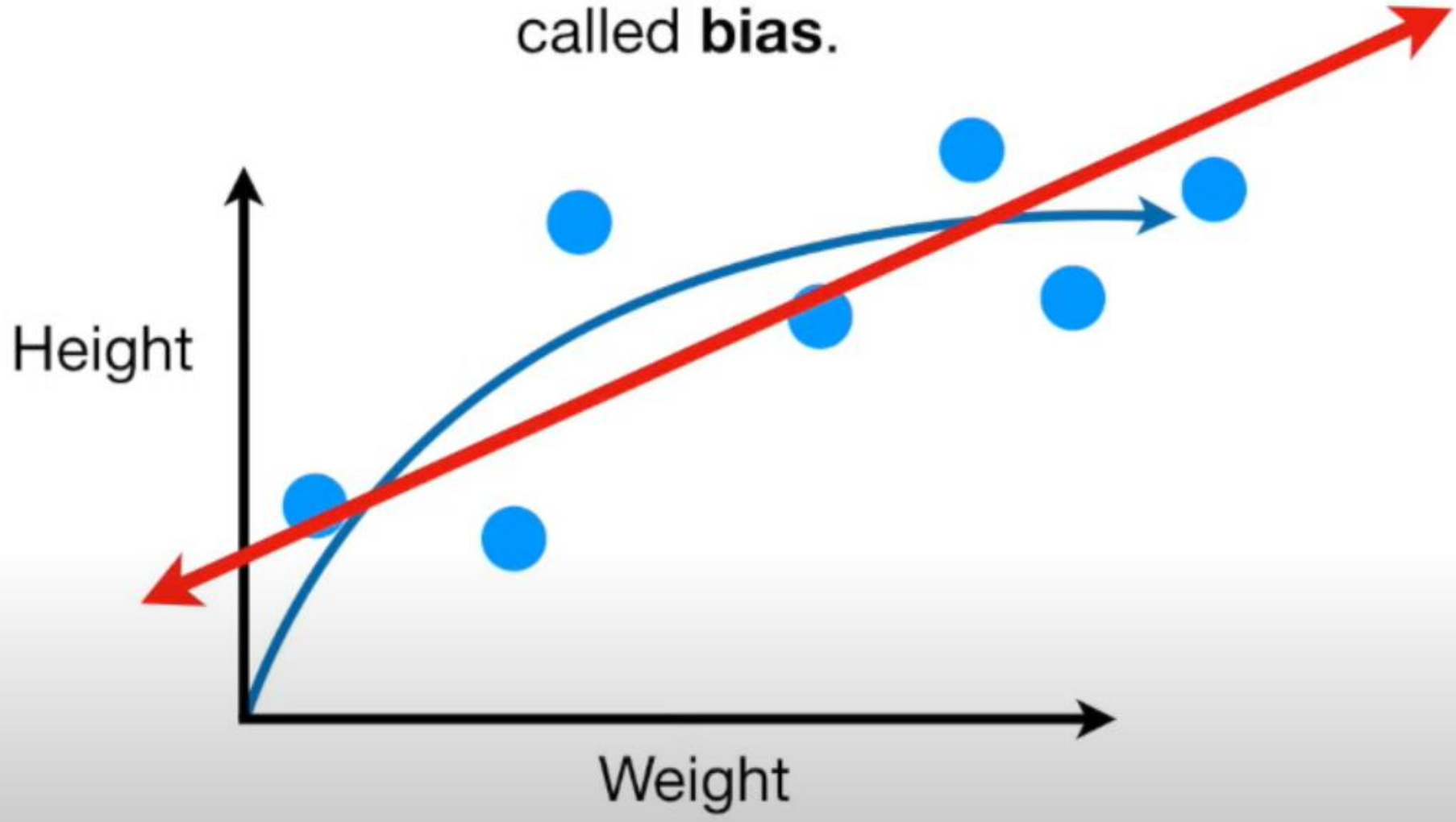


The first machine learning algorithm that we will use is Linear Regression (aka “Least Squares”).

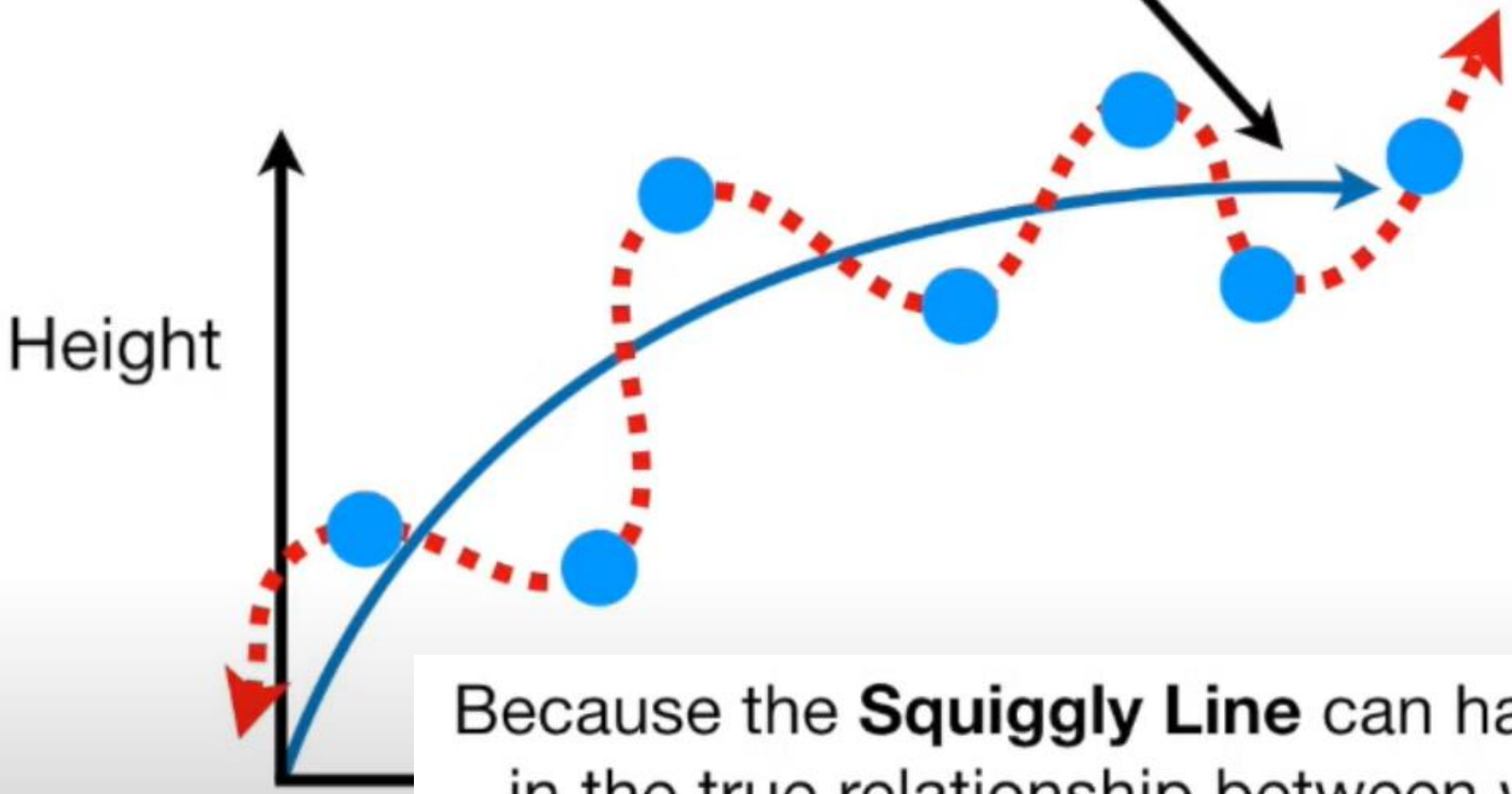


NOTE: The **Straight Line** doesn’t have the flexibility to accurately replicate the arc in the “true” relationship.

The inability for a machine learning method (like linear regression) to capture the true relationship is called **bias**.

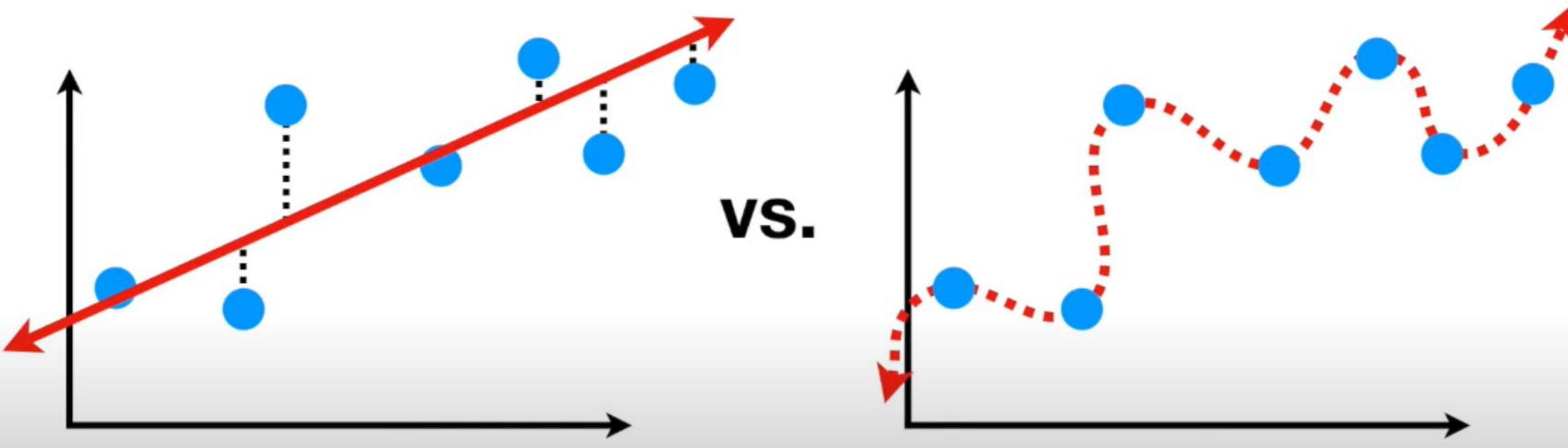


The **Squiggly Line** is super flexible and hugs the **training set** along the arc of the true relationship.

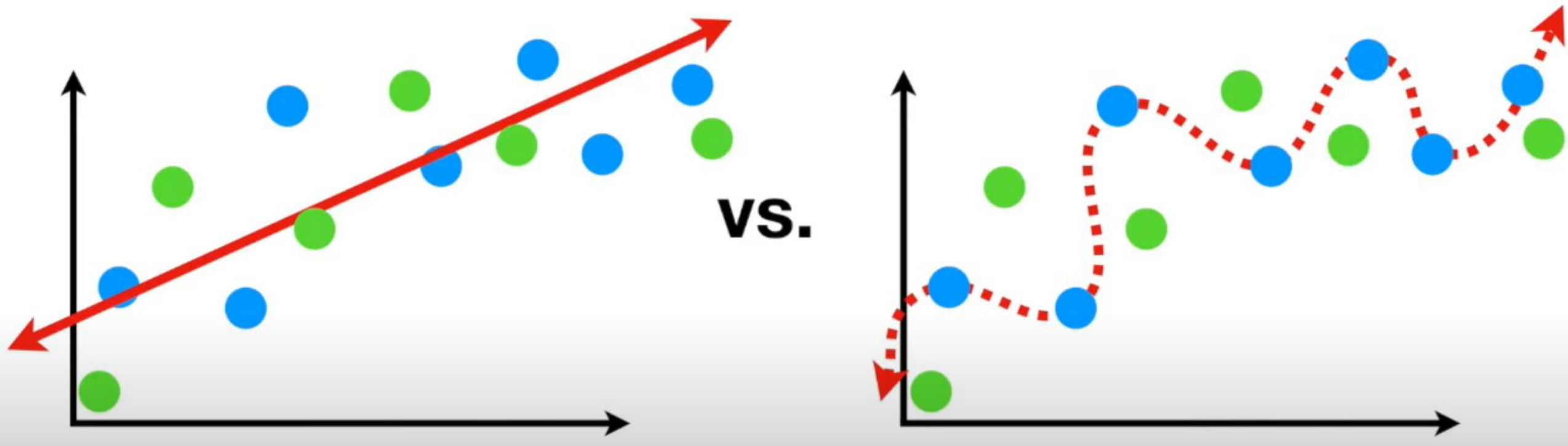


Because the **Squiggly Line** can handle the arc in the true relationship between weight and height, it has very little **bias**.

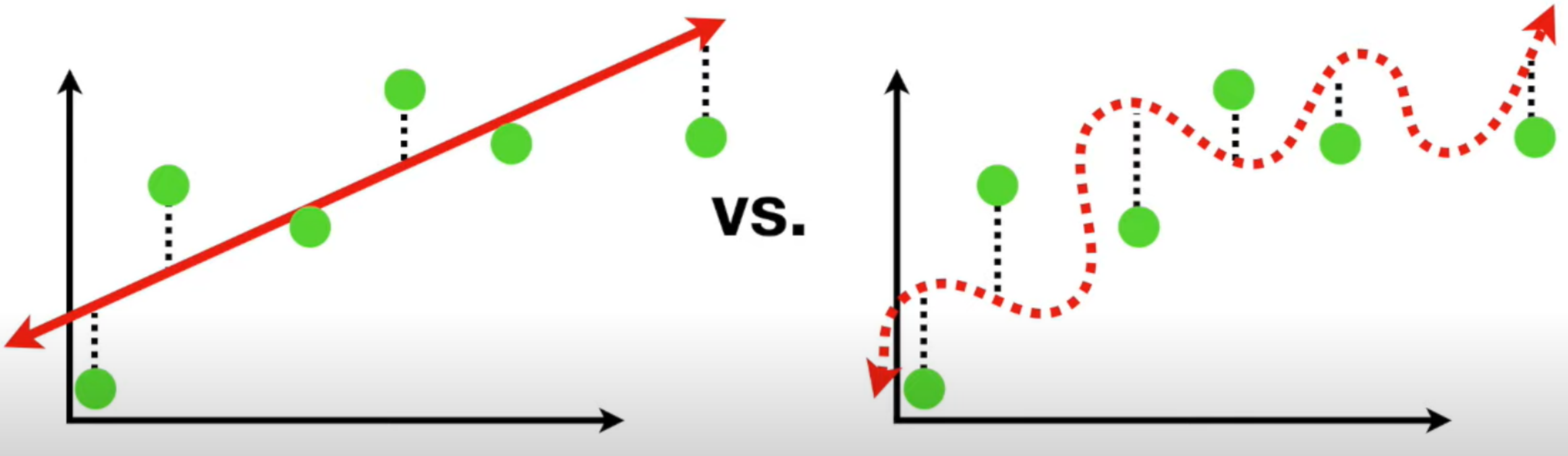
The **Squiggly Line** wins.

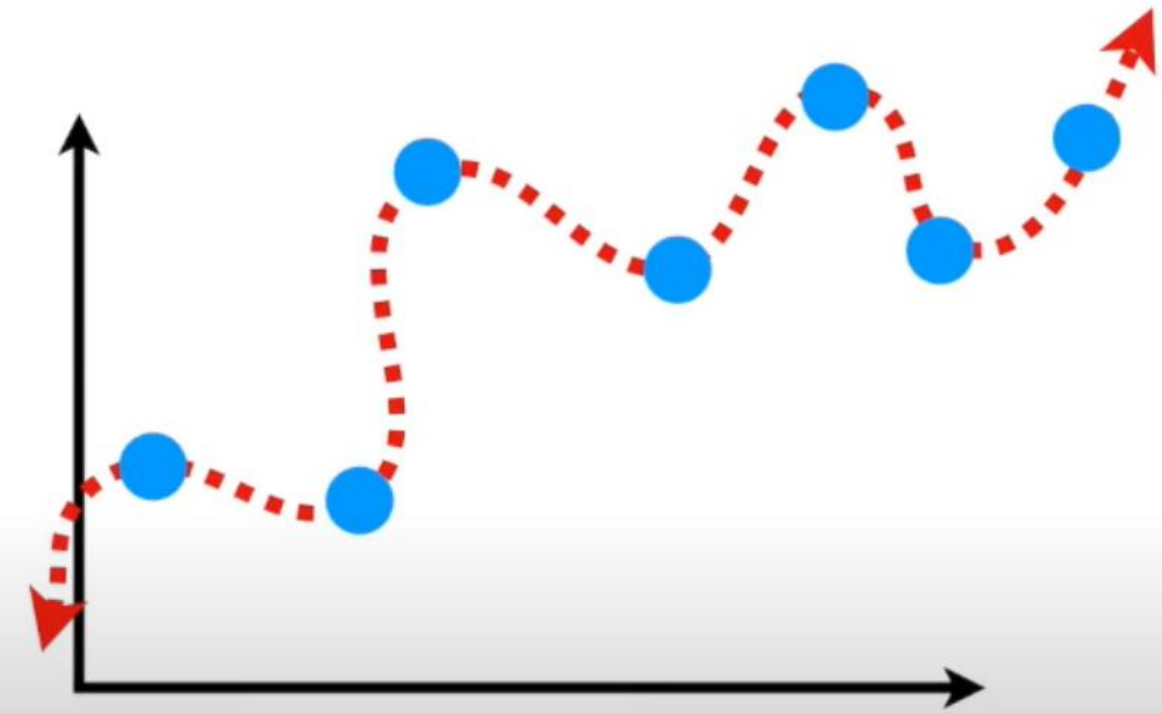


We also have a **testing set**…

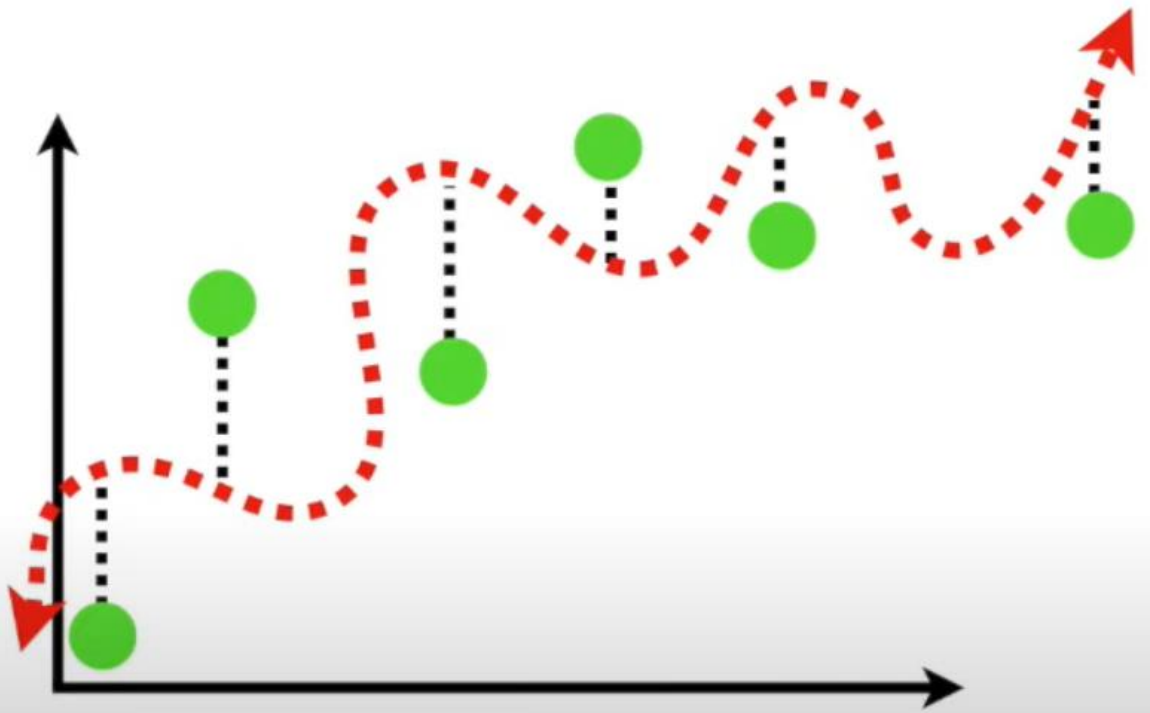


In the contest to see whether the **Straight Line** fits the **testing set** better than the **Squiggly Line**…

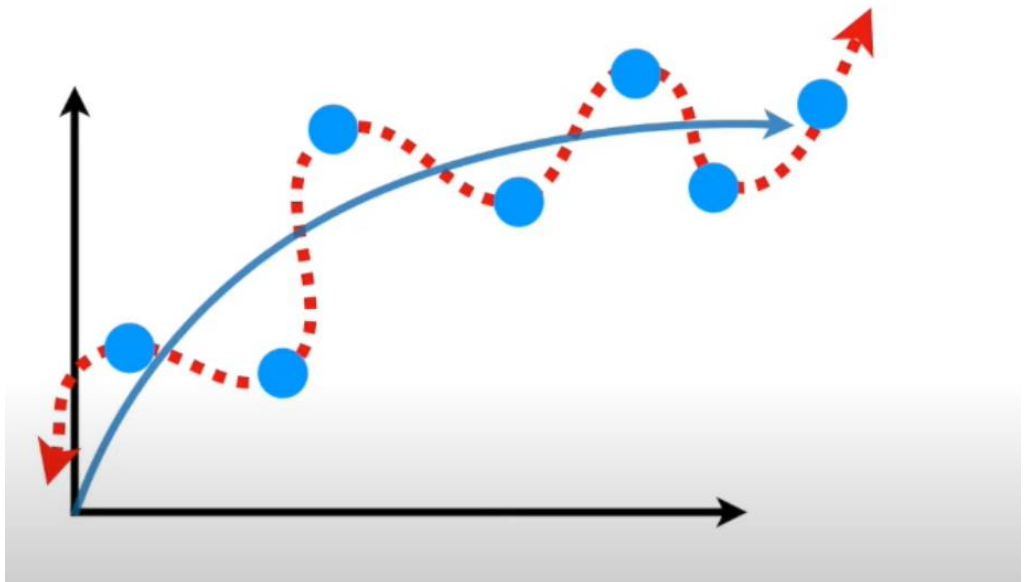




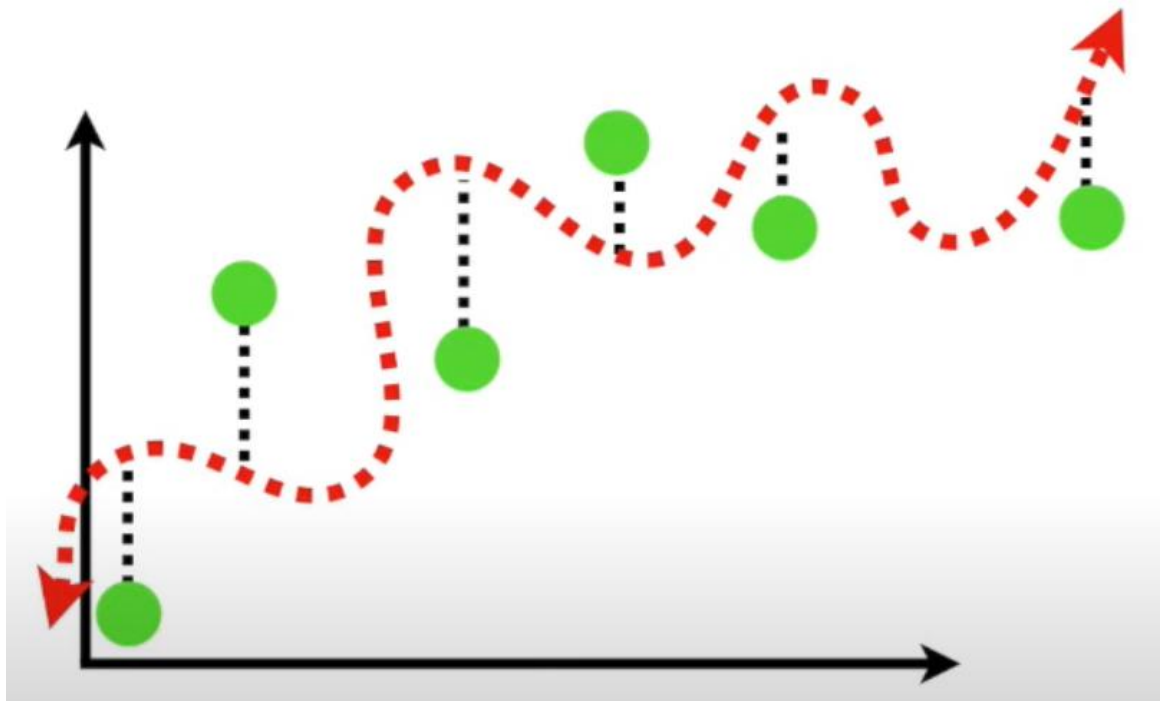
…it did a terrible job fitting the **testing set**…



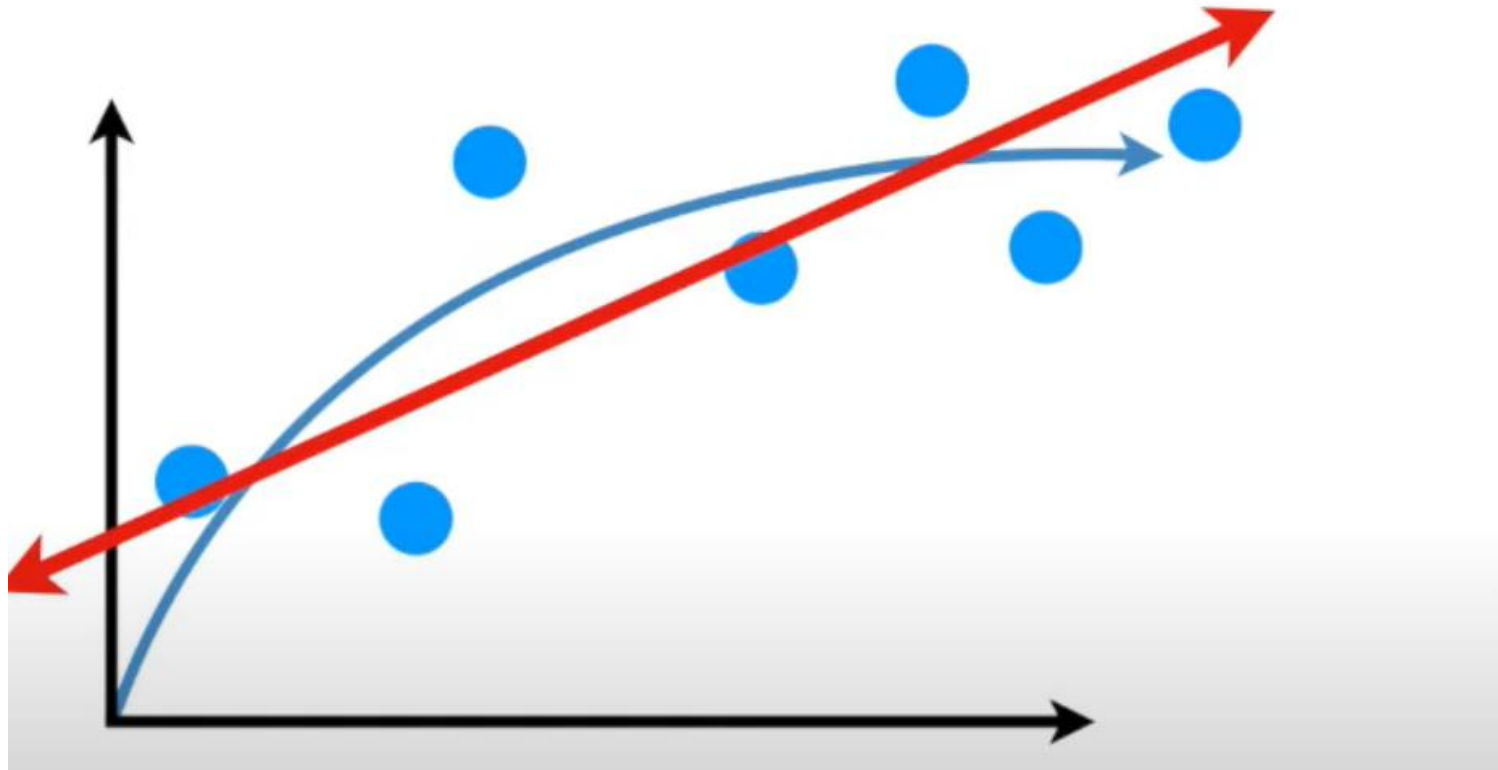
The **Squiggly Line** has **low bias**, since it is flexible and can adapt to the curve in the relationship between weight and hight…



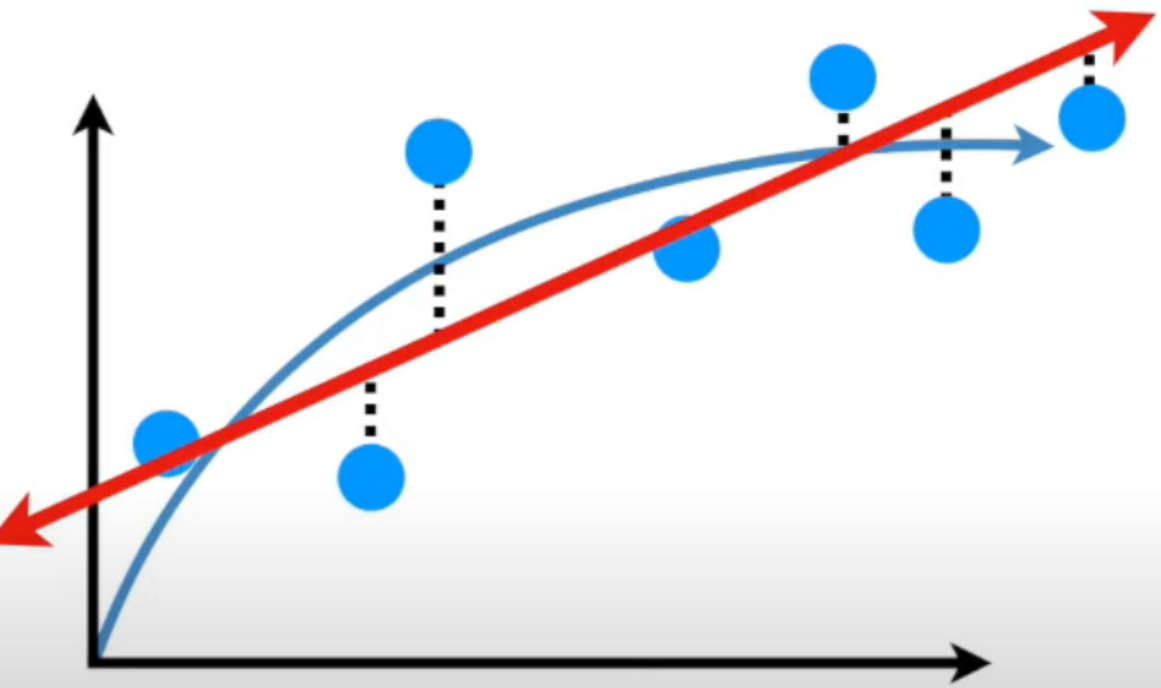
…but the **Squiggly Line** has **high variability**, because it results in vastly different Sums of Squares for different datasets.

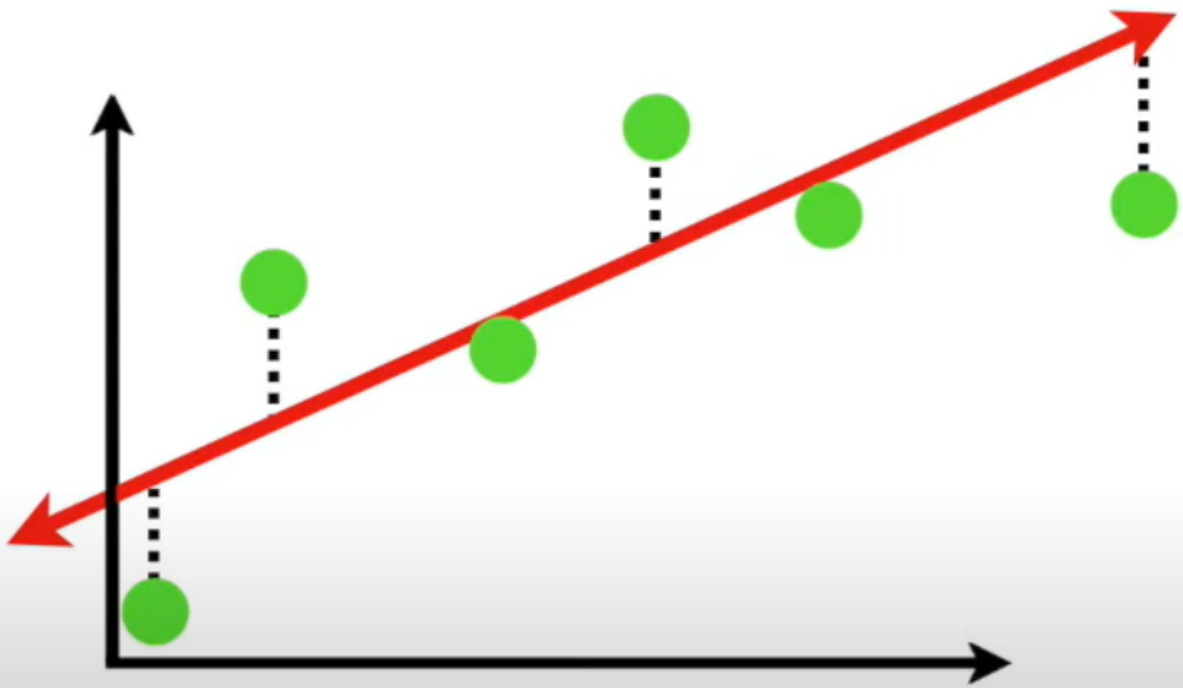


In contrast, the **Straight Line** has relatively **high bias**, since it can not capture the curve in the relationship between weight and height…

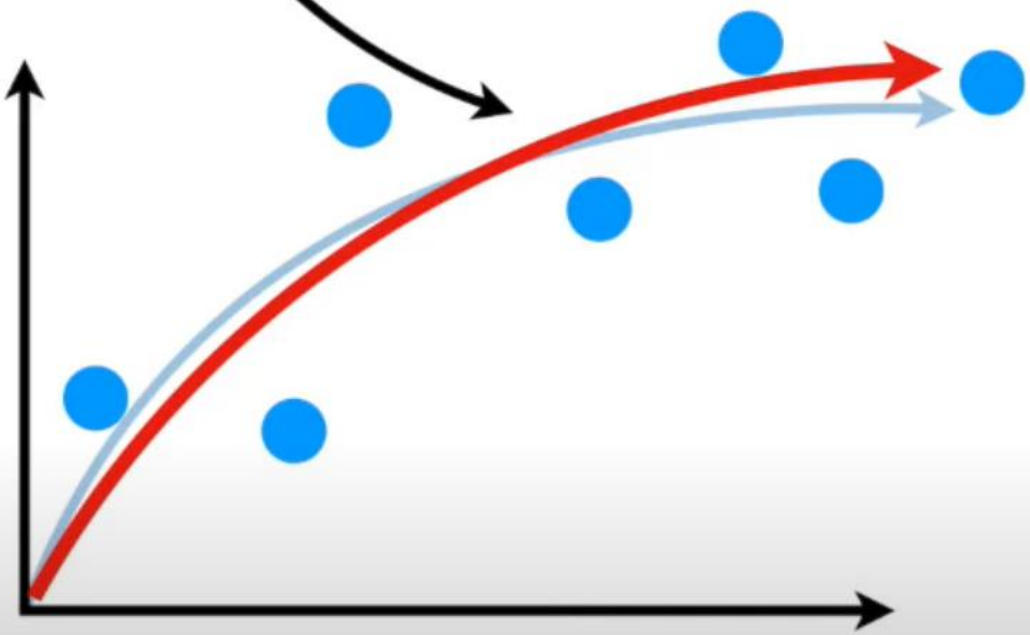


…but the **Straight Line** has relatively **low variance**, because the Sums of Squares are very similar for different datasets.

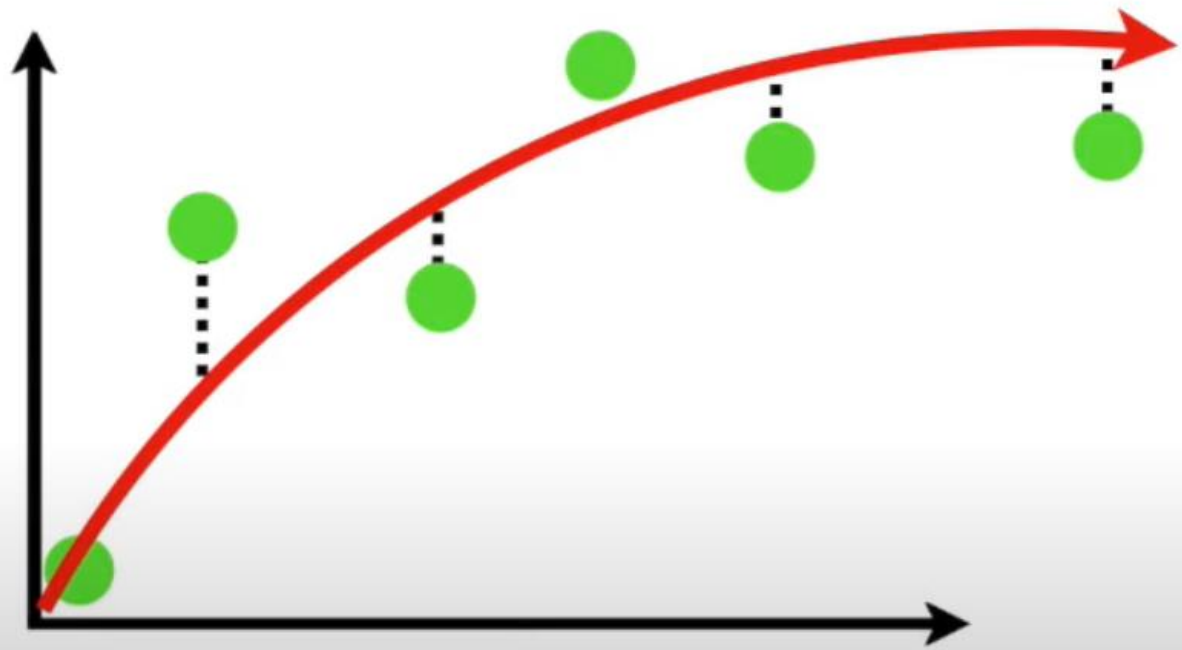


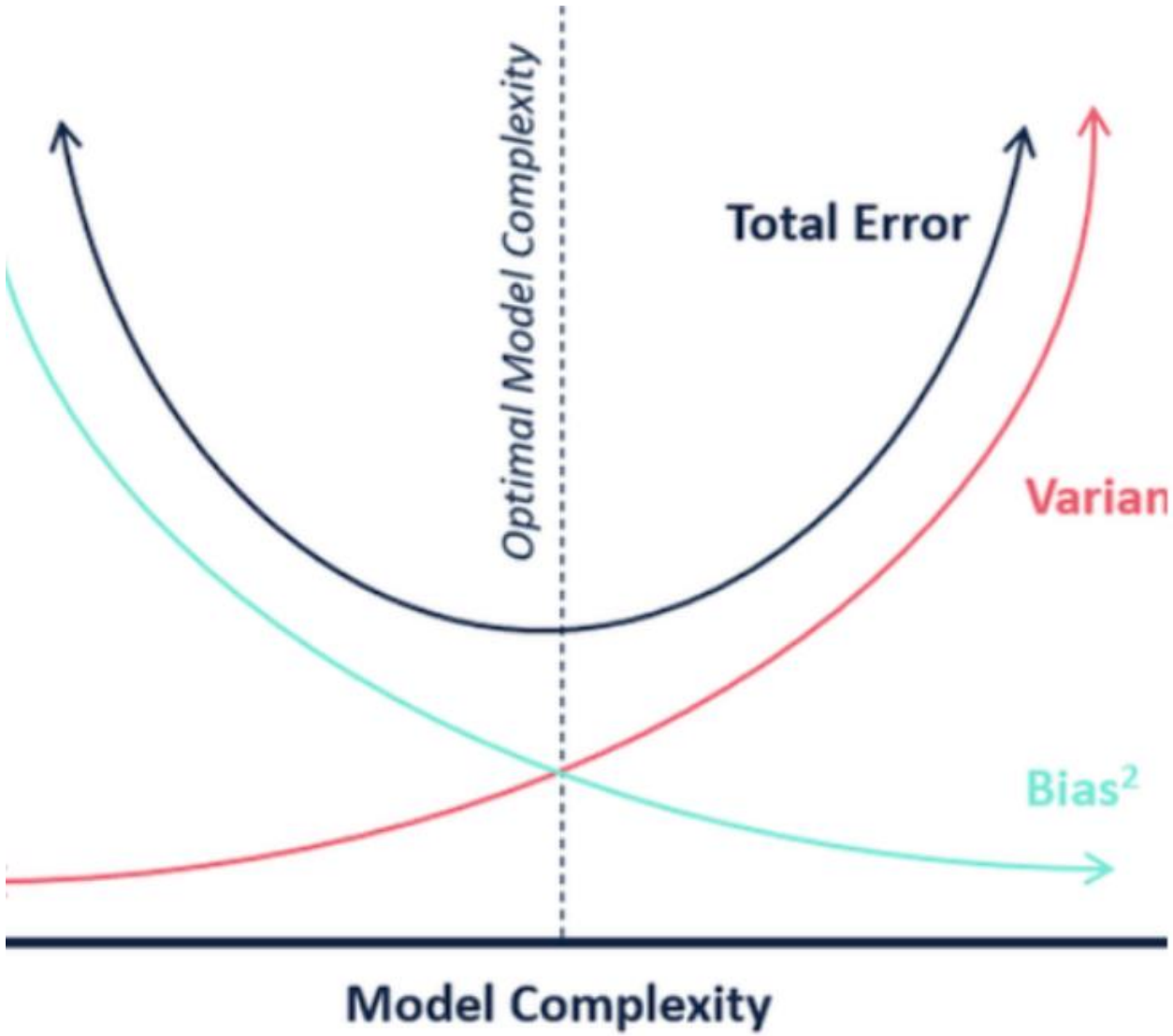


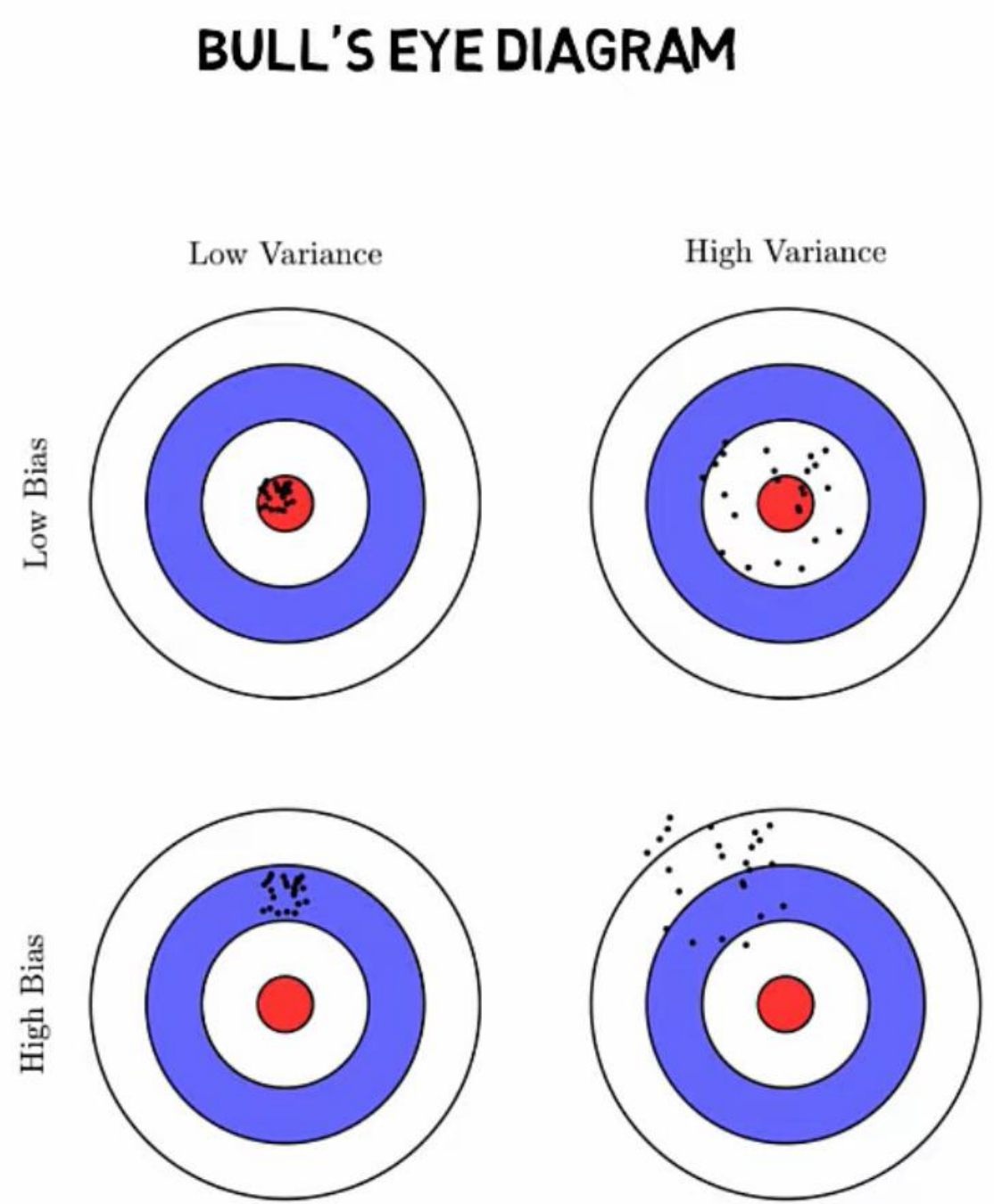
In machine learning, the ideal algorithm has **low bias** and can accurately model the true relationship…



…and it has **low variability**, by producing consistent predictions across different datasets.




Optimal Model Complexity
Total Error
Varian
Bias²
Model Complexity


BULL'S EYE DIAGRAM
Low Variance
High Variance
Low Bias
High Bias