DATA MINING 2
Maximum Likelihood Estimation

Riccardo Guidotti

a.a. 2024/2025

Contains edited slides from StatQuest

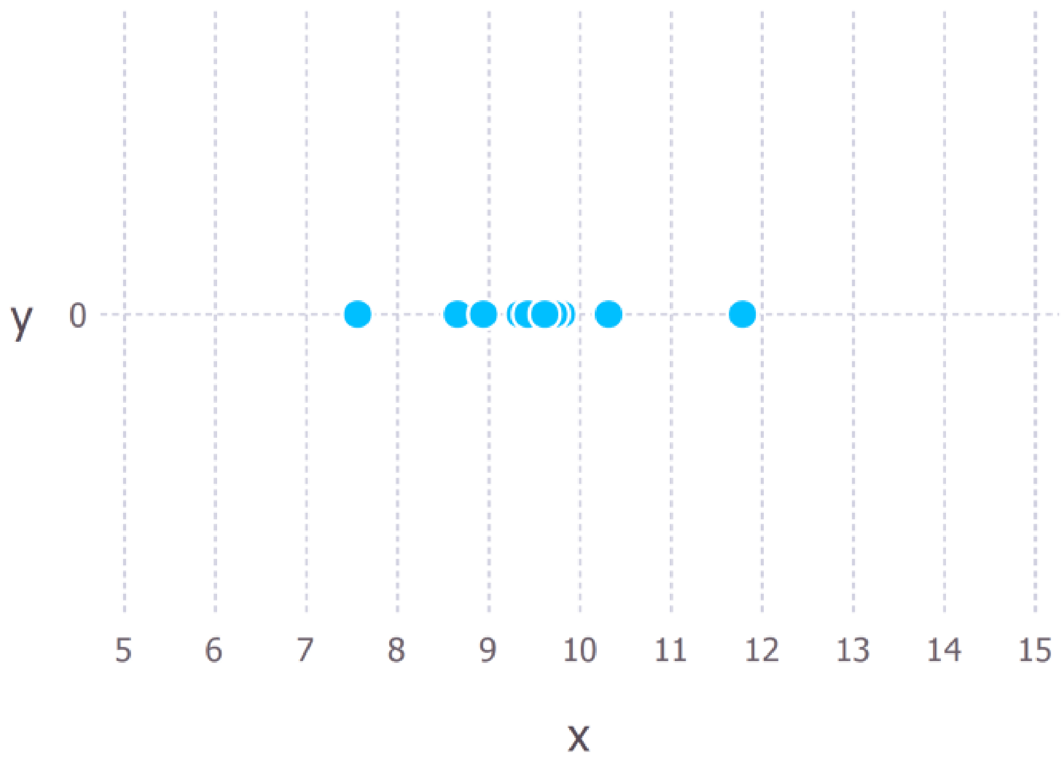
UNIVERSITÀ DI PISA

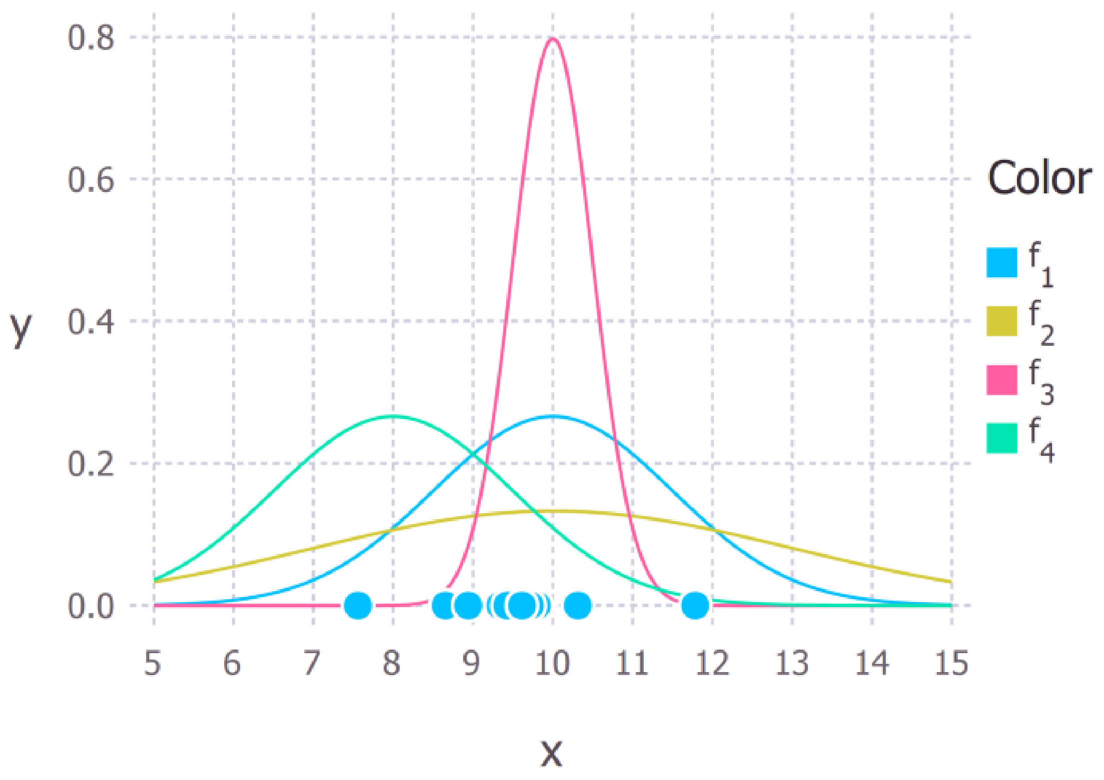
Intuition

- Maximum Likelihood Estimation (MLE) is a method that determines values for the parameters of a model.
- The parameter values are found such that they maximize the **likelihood** that the process described by the model produced the data that were actually observed.

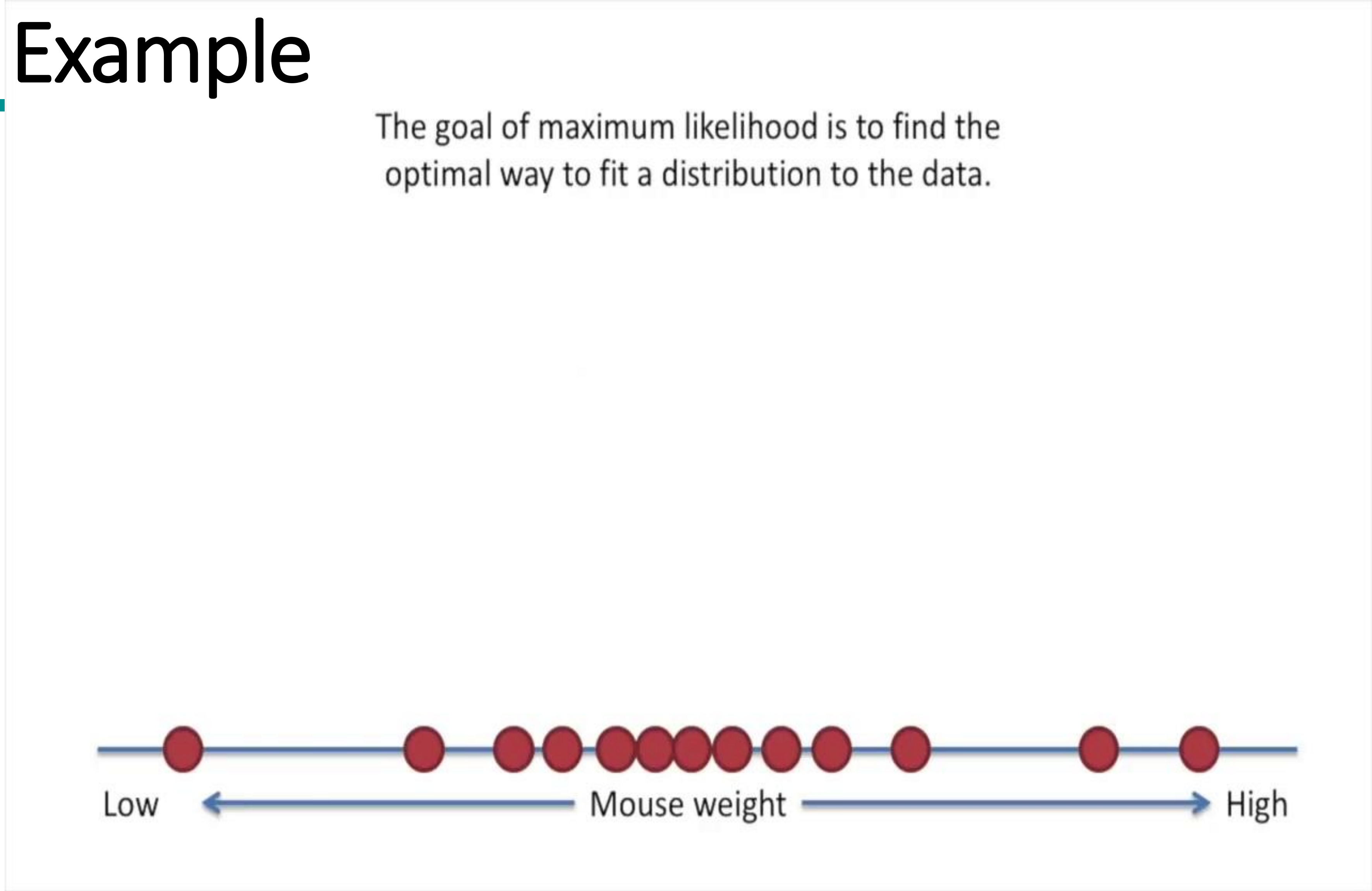
Which model fit best?

- Normal Gaussian distribution
- Parameters: *mean* and *standard deviation*

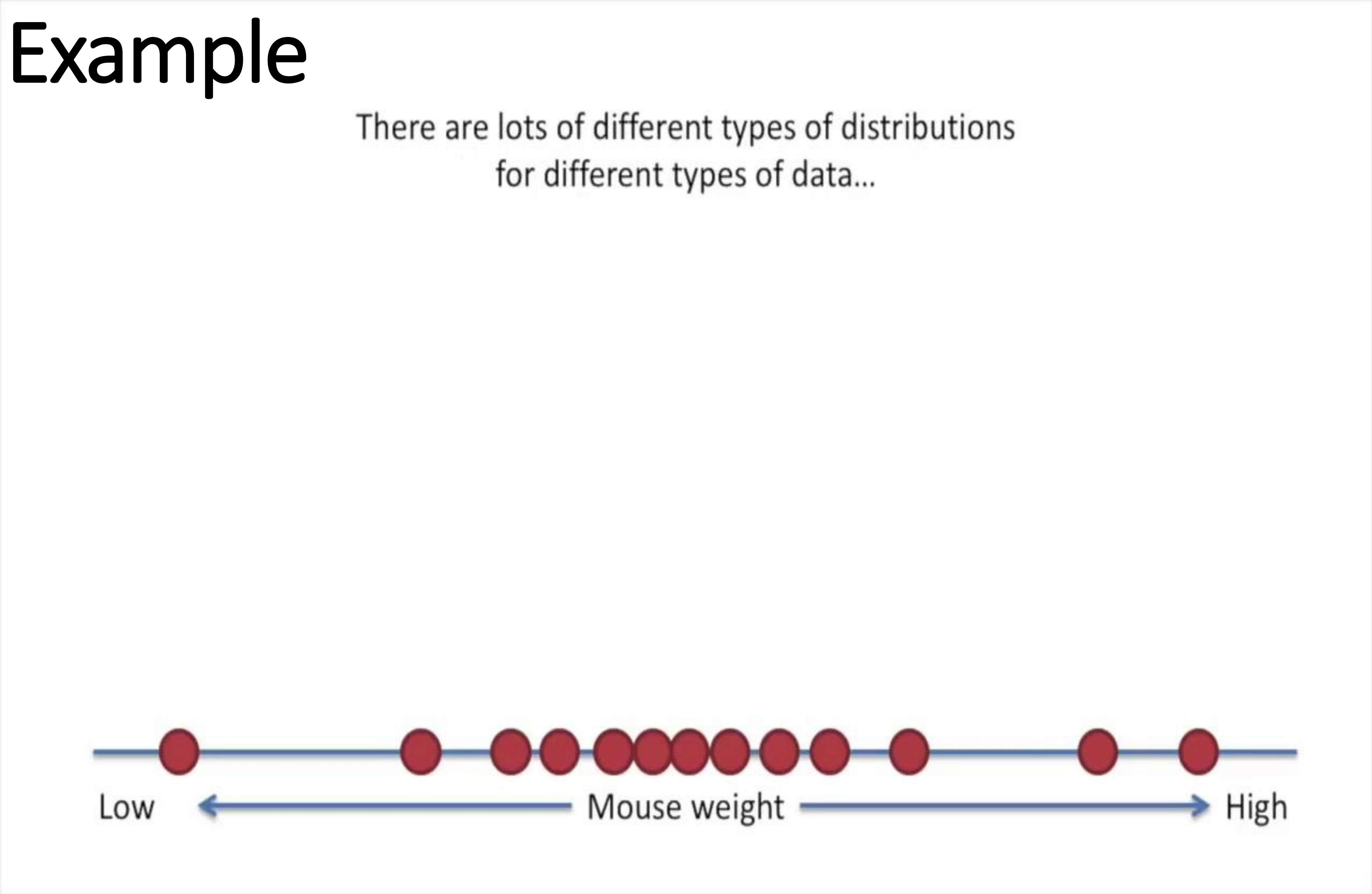




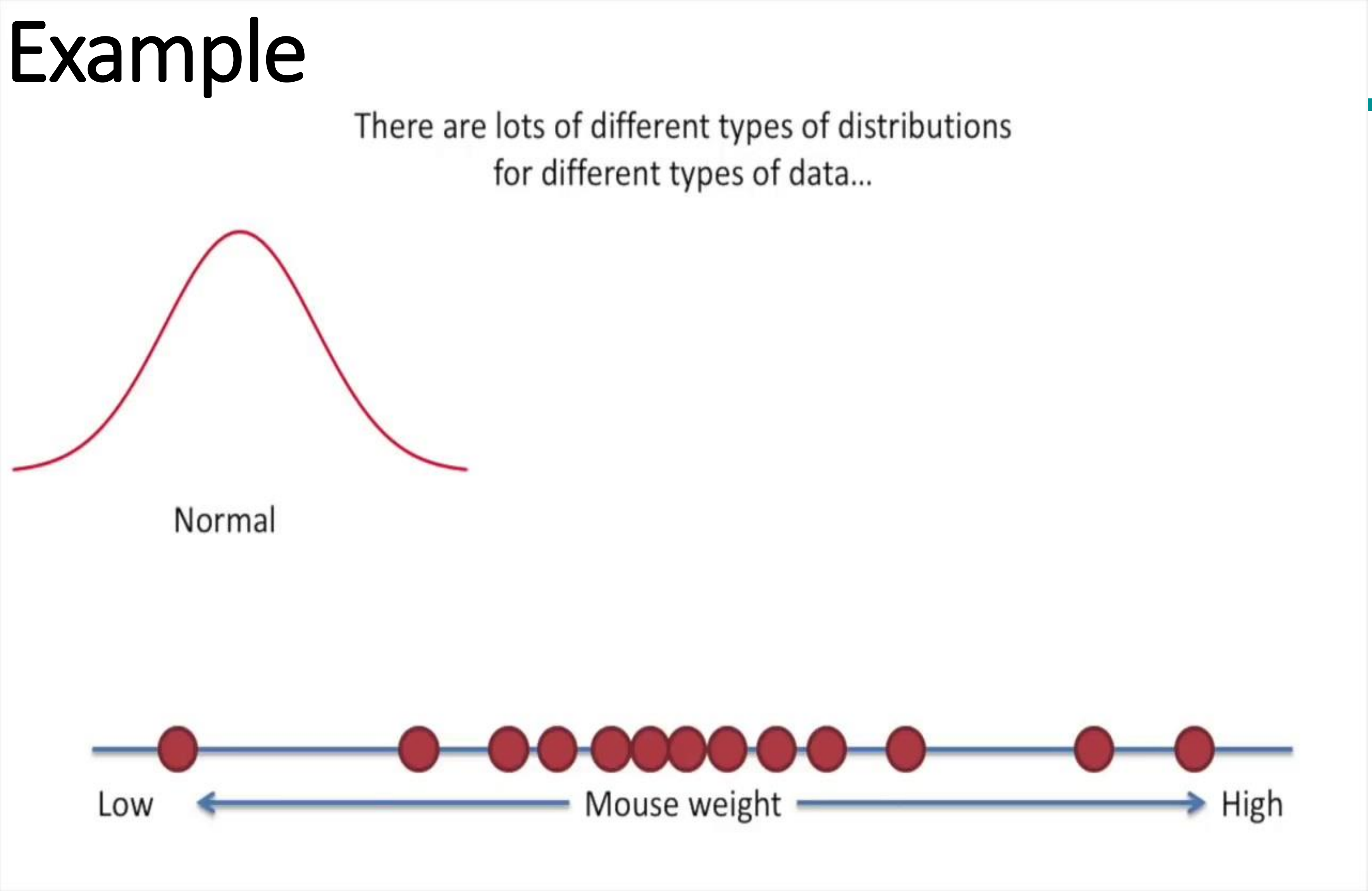
MLE Example


The goal of maximum likelihood is to find the
optimal way to fit a distribution to the data.
Low
Mouse weight
High

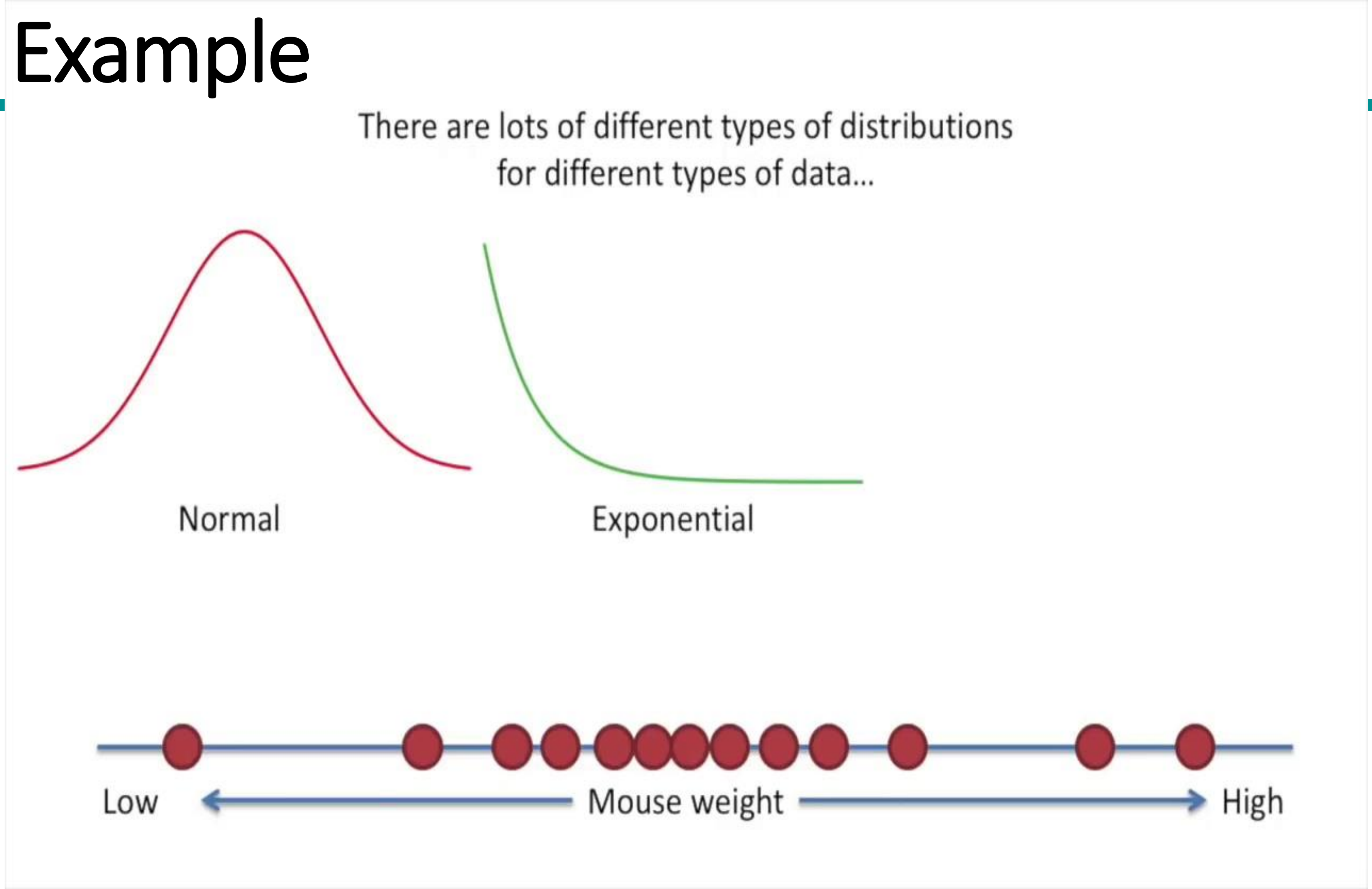
MLE Example


There are lots of different types of distributions
for different types of data...
Low
Mouse weight
High

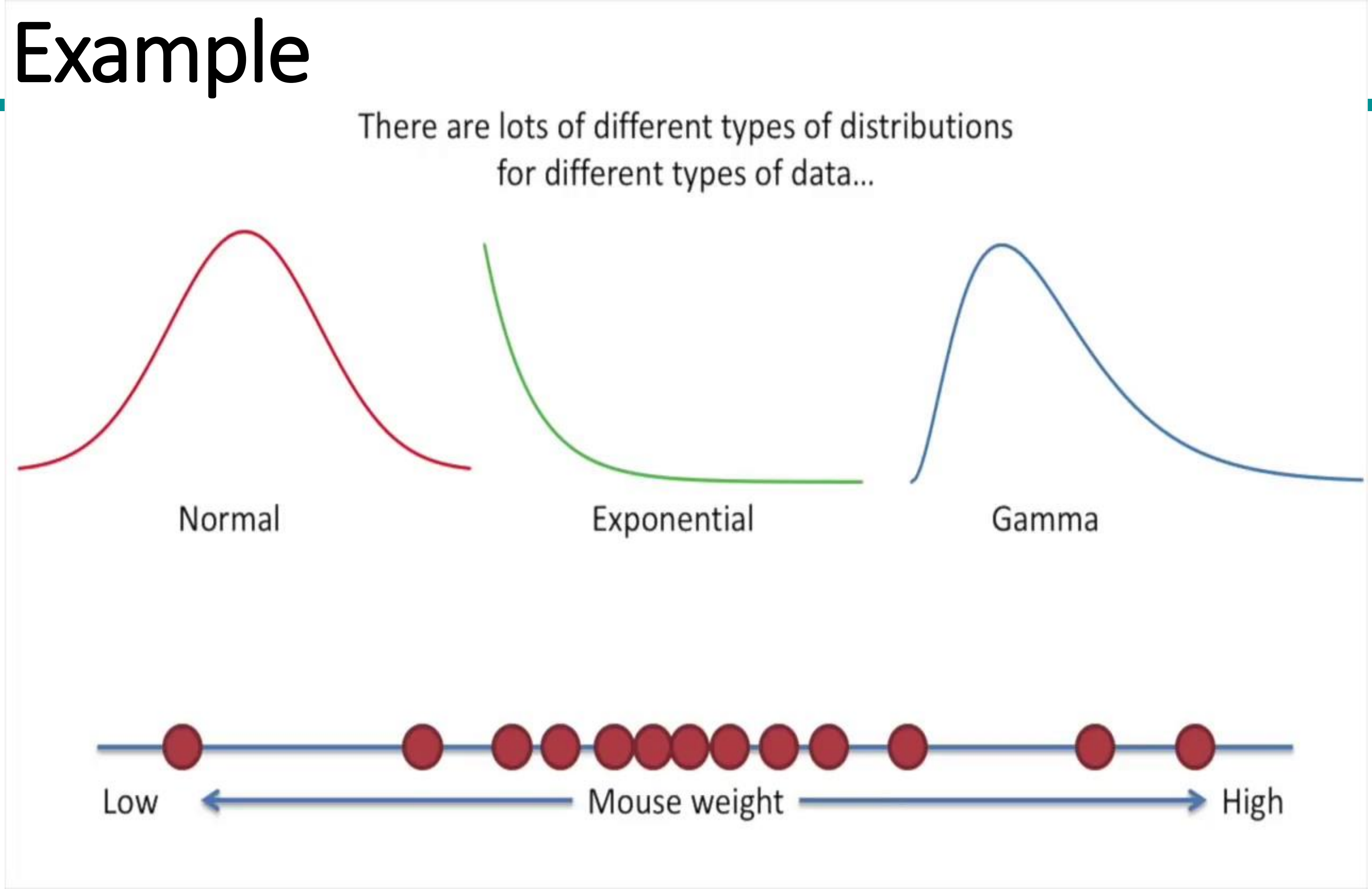
MLE Example


There are lots of different types of distributions
for different types of data...
Normal
Low
Mouse weight
High

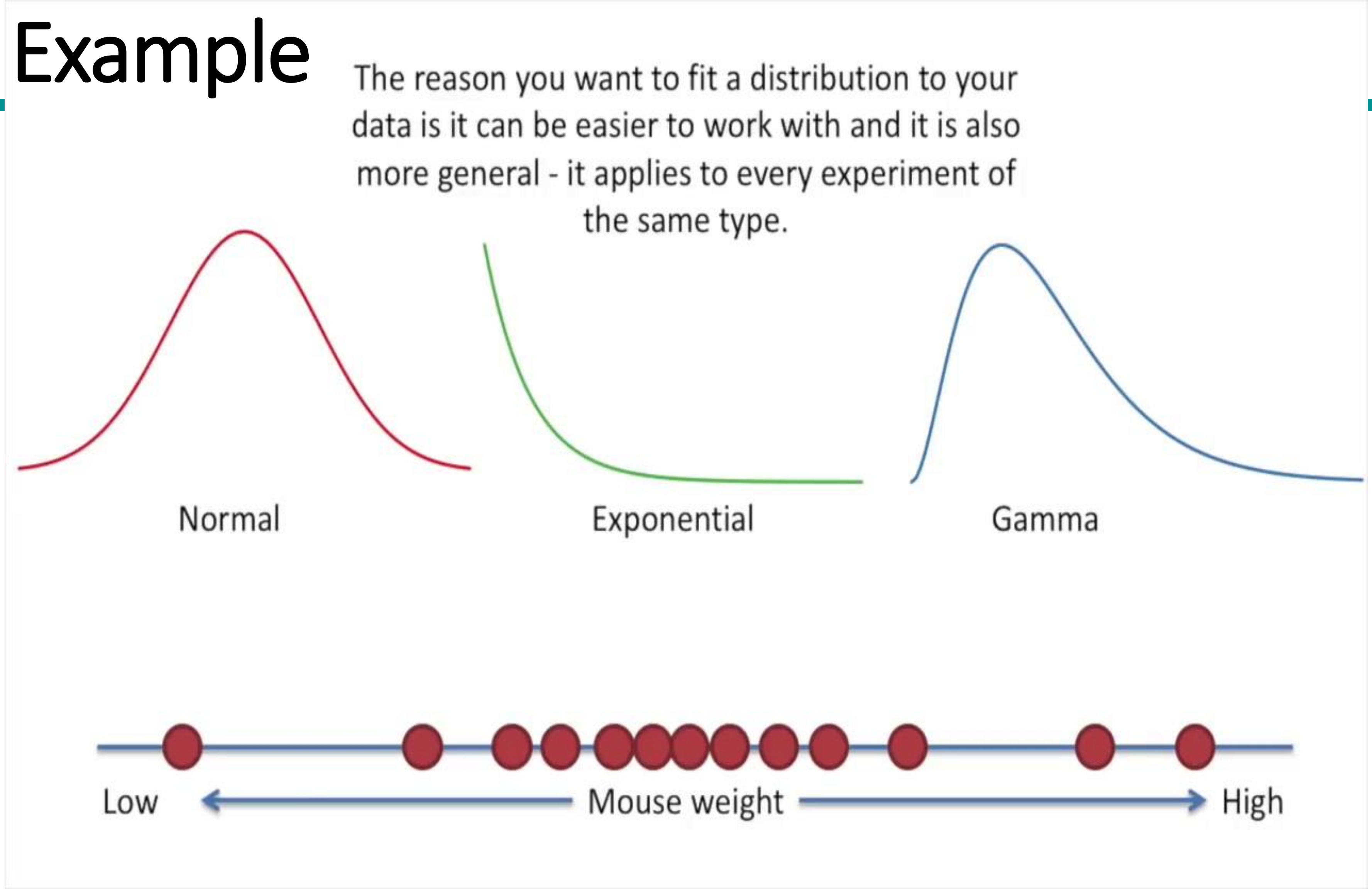
MLE Example


There are lots of different types of distributions
for different types of data...
Normal
Exponential
Low
Mouse weight
High

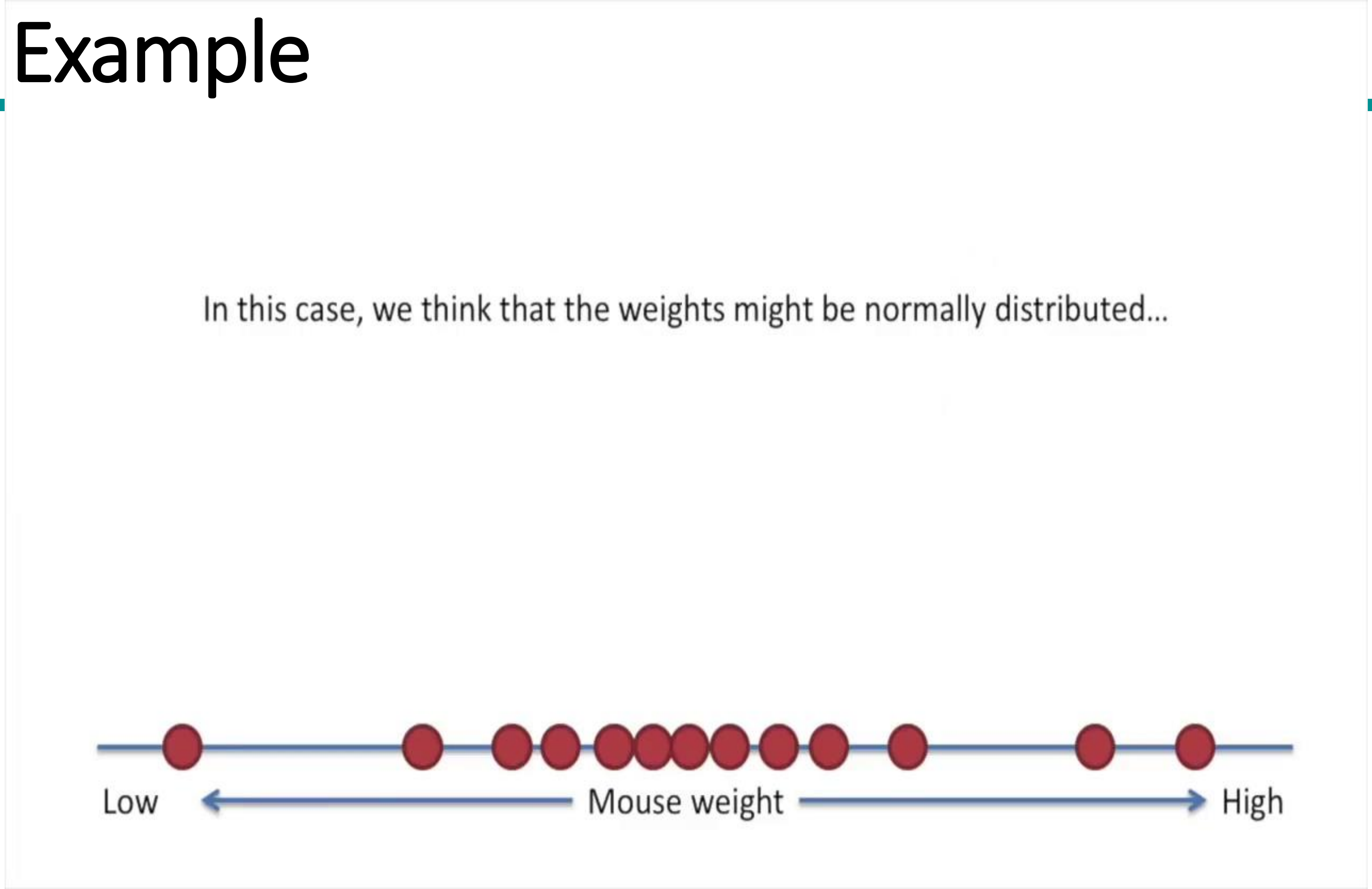
MLE Example


There are lots of different types of distributions
for different types of data...
Normal
Exponential
Gamma
Low
Mouse weight
High

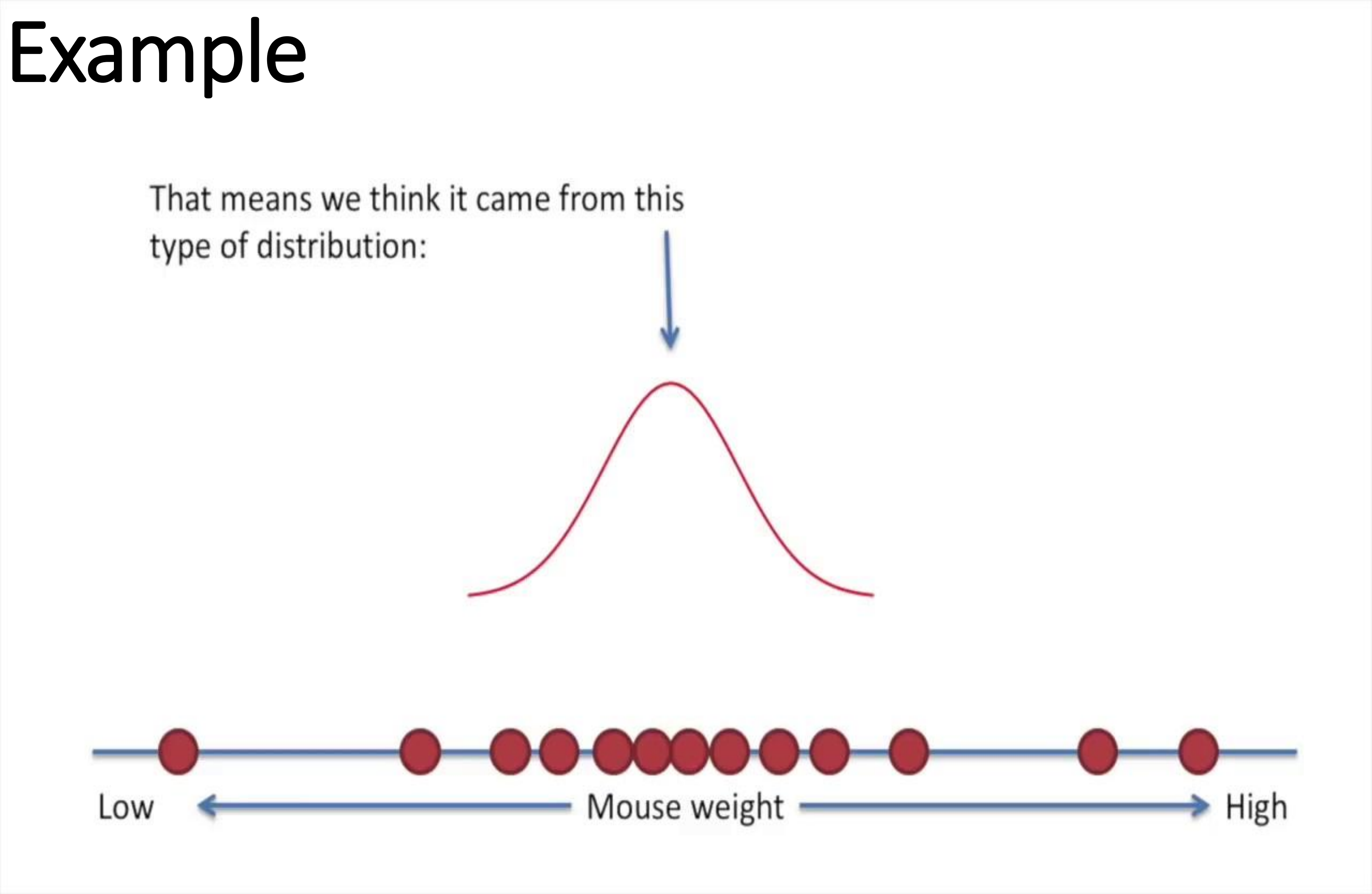
MLE Example


The reason you want to fit a distribution to your data is it can be easier to work with and it is also more general - it applies to every experiment of the same type.
Normal
Exponential
Gamma
Low
Mouse weight
High

MLE Example

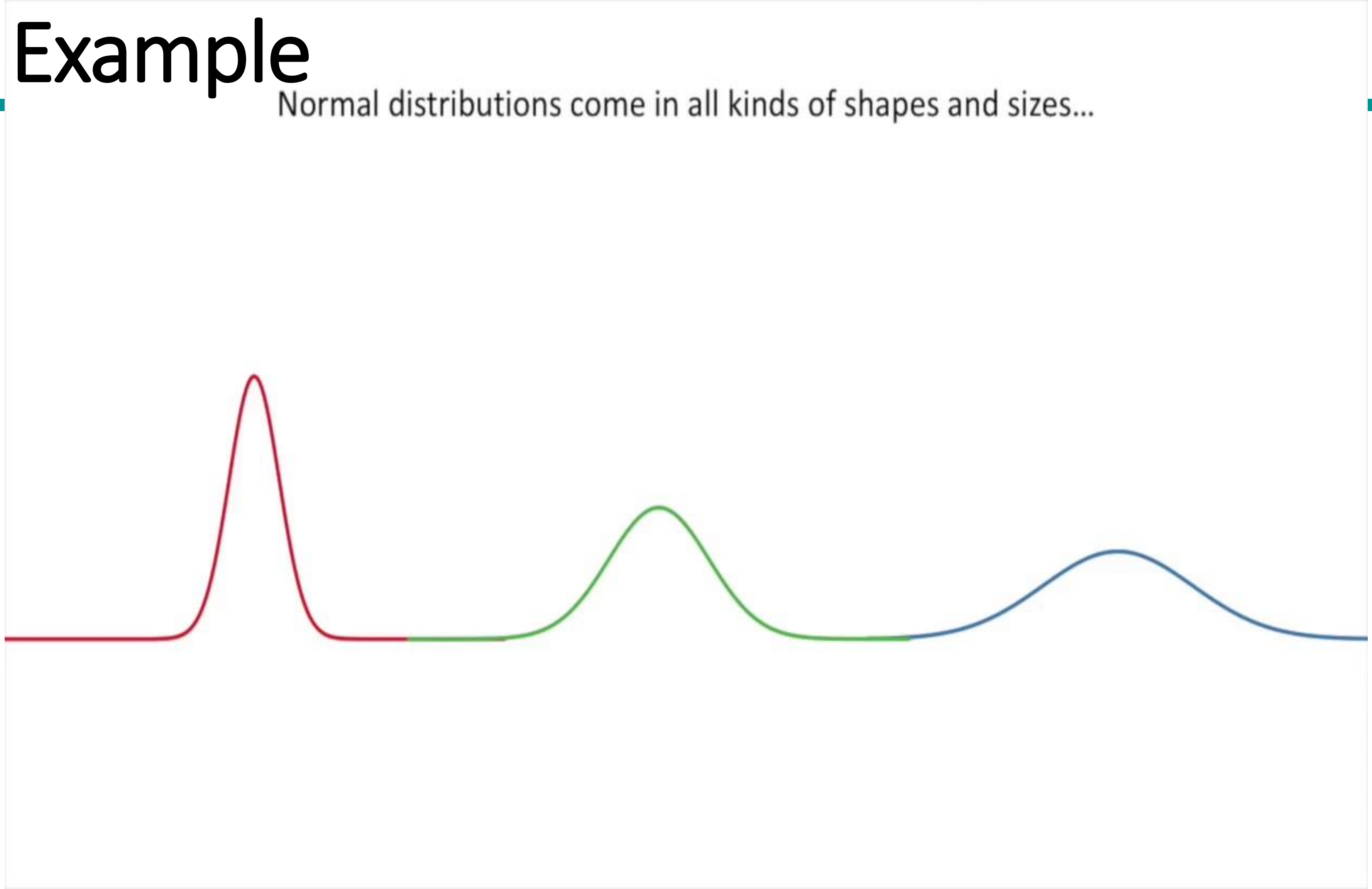

In this case, we think that the weights might be normally distributed...
Low
Mouse weight
High

MLE Example

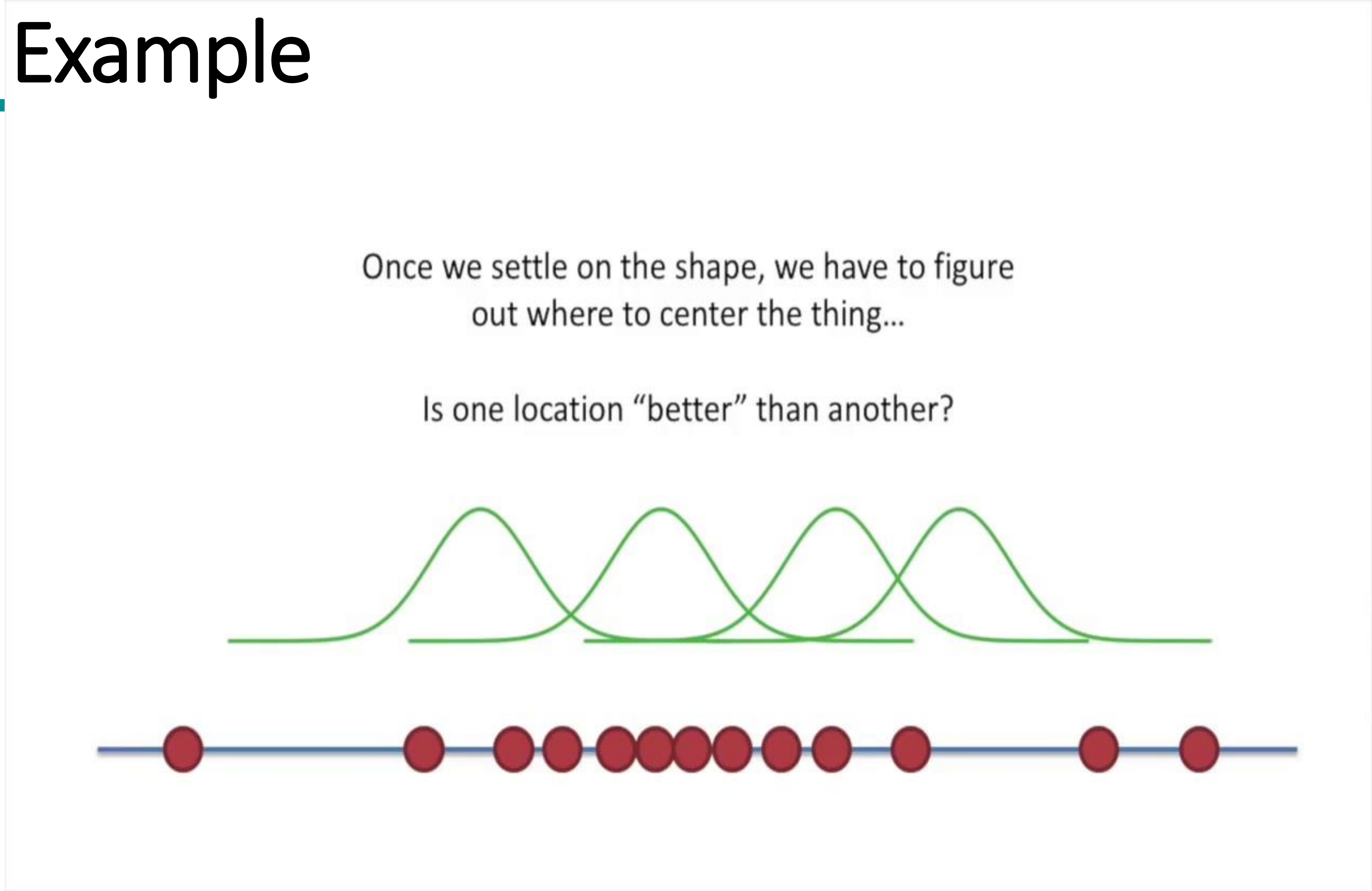


That means we think it came from this type of distribution:
Low
Mouse weight
High

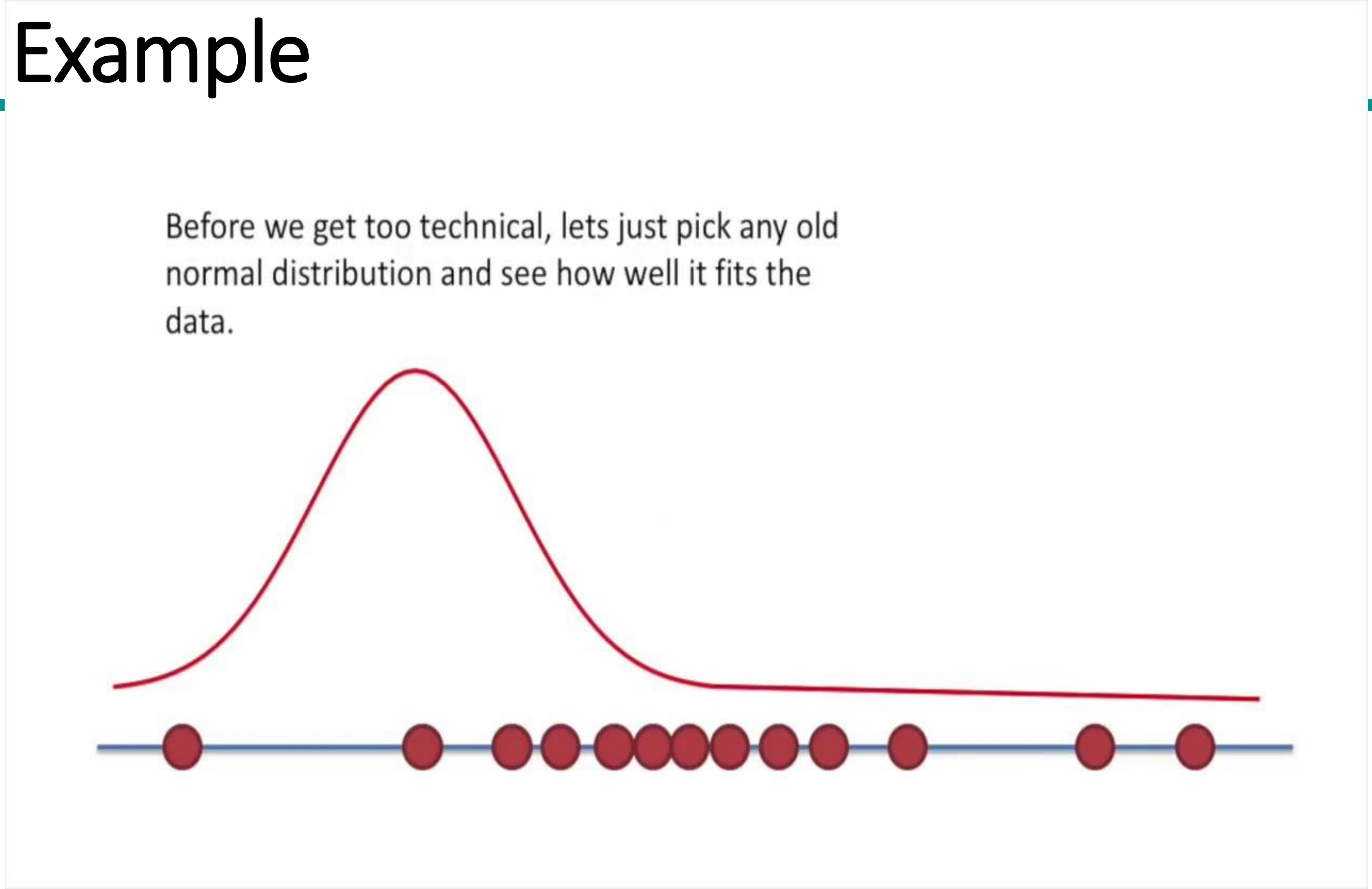
MLE Example



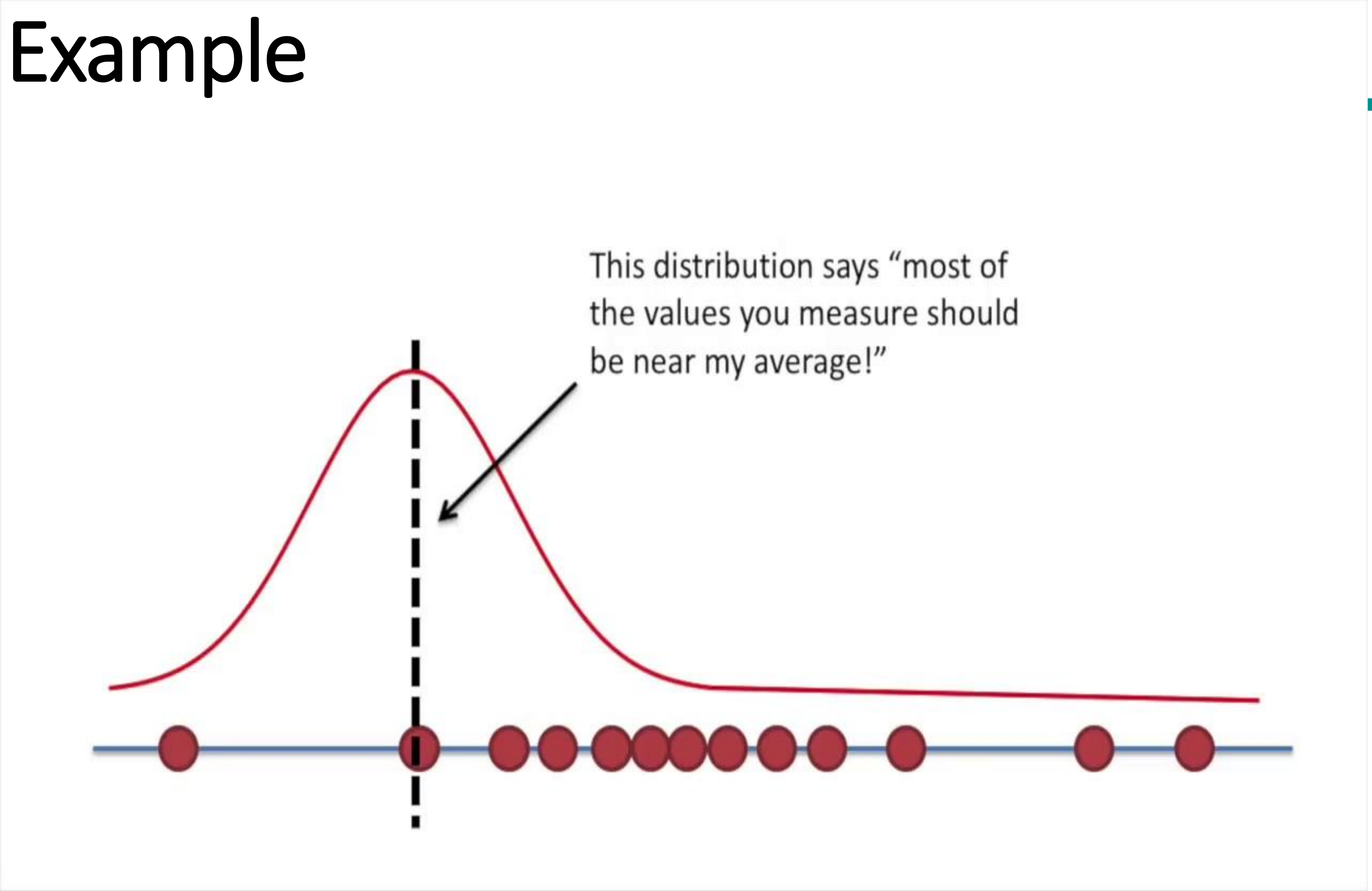
MLE Example


Once we settle on the shape, we have to figure out where to center the thing...
Is one location "better" than another?

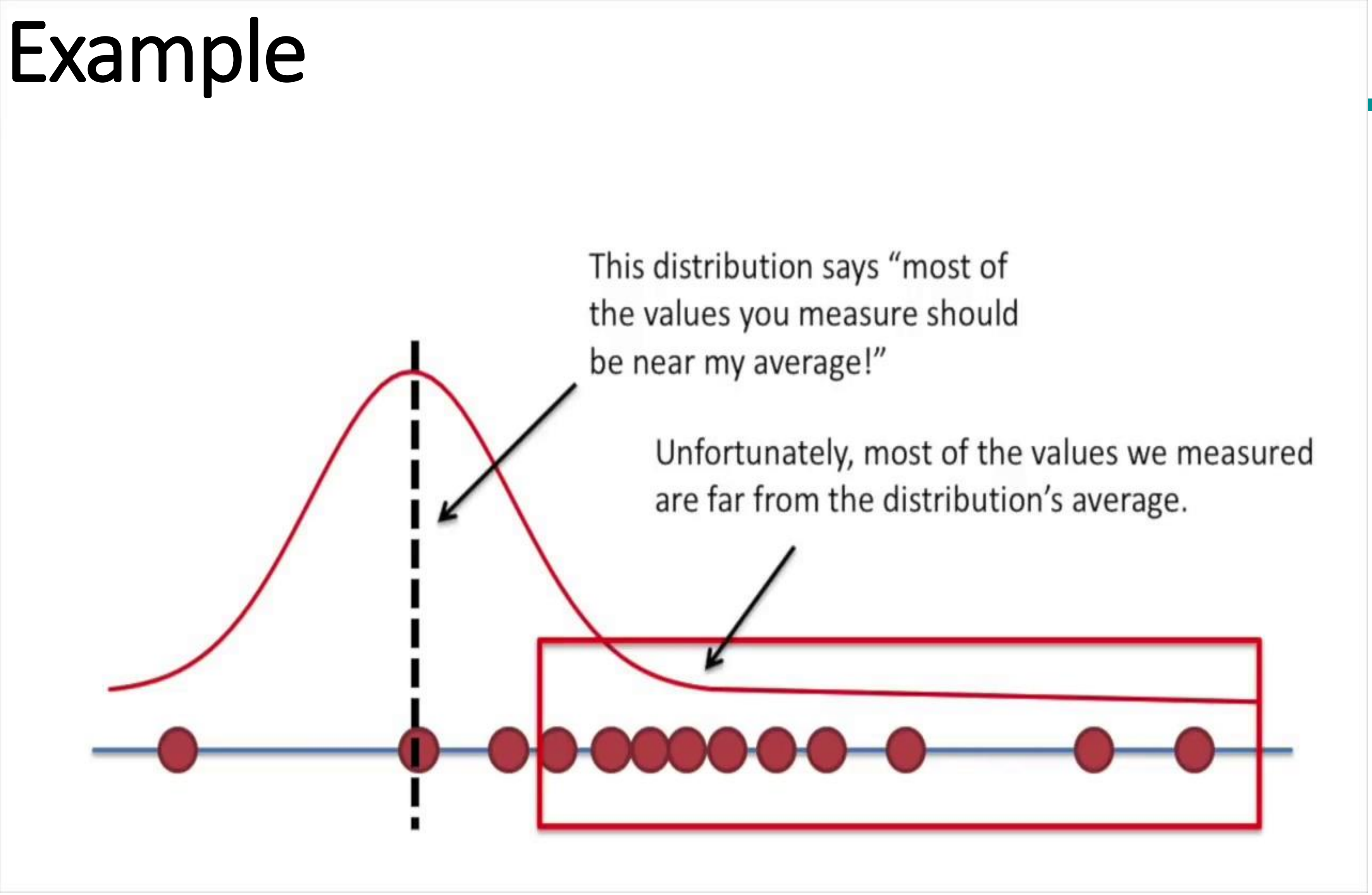
MLE Example


Before we get too technical, lets just pick any old normal distribution and see how well it fits the data.

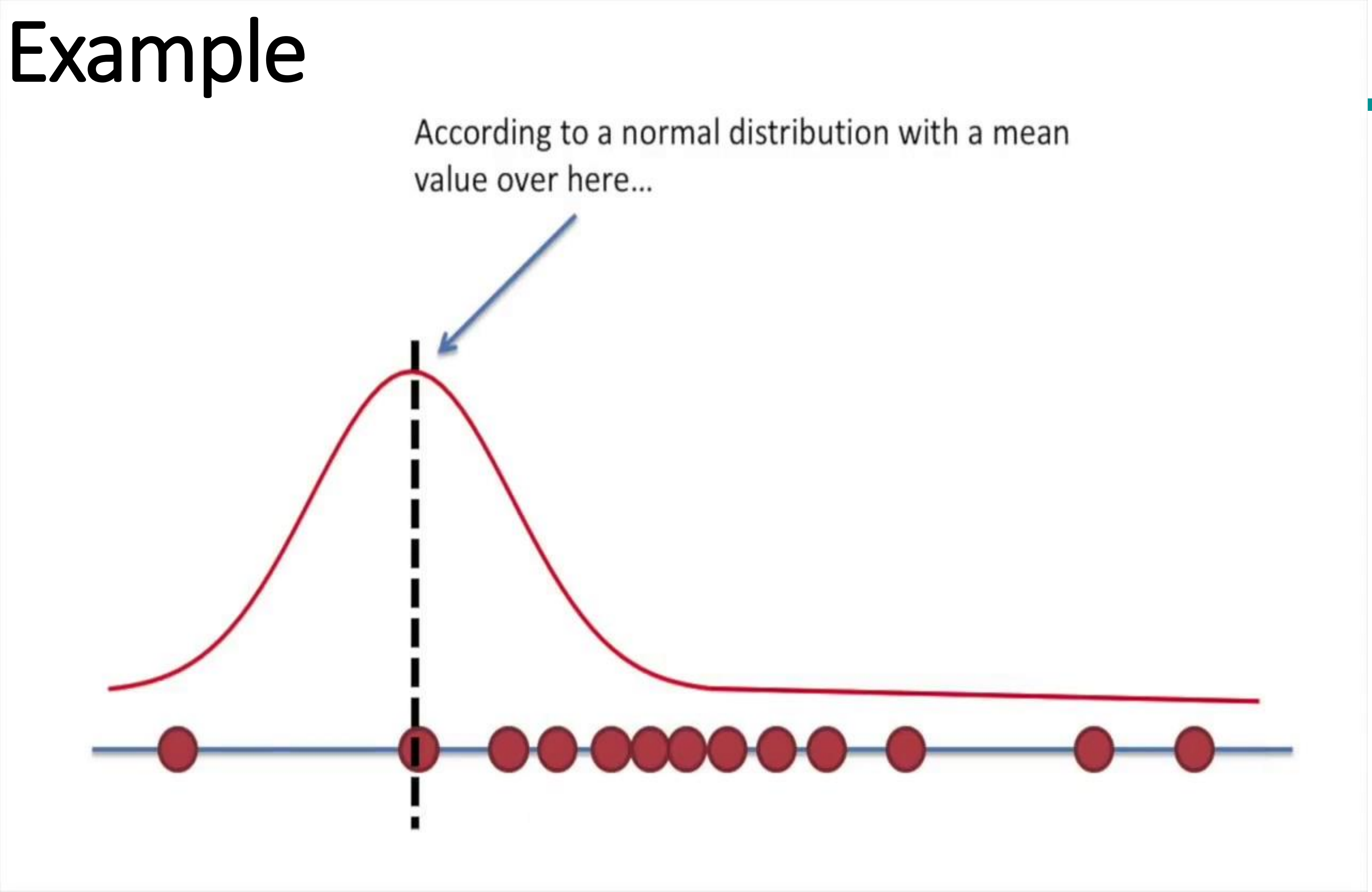
MLE Example


This distribution says "most of the values you measure should be near my average!"

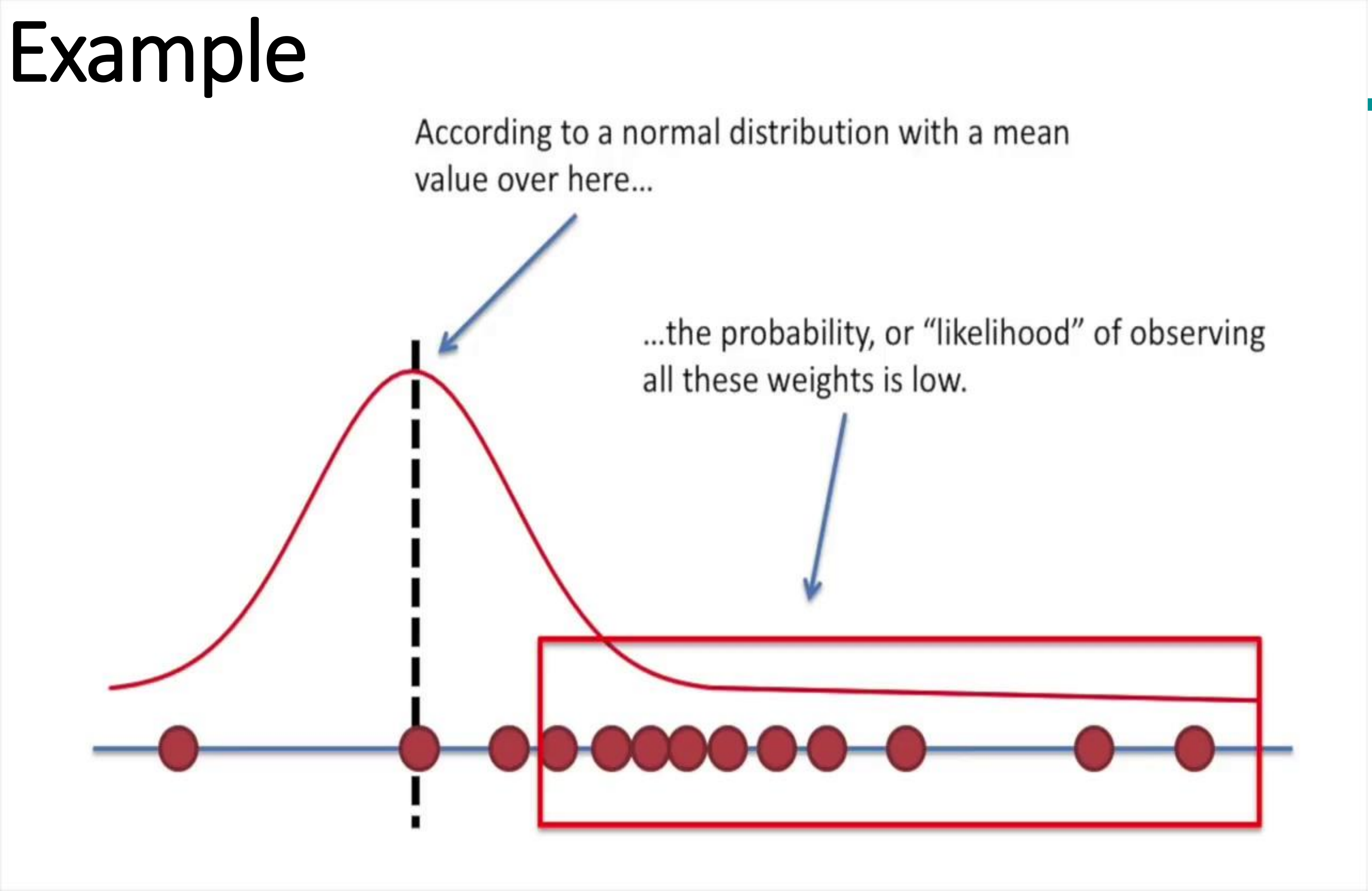
MLE Example


This distribution says "most of the values you measure should be near my average!"
Unfortunately, most of the values we measured are far from the distribution's average.

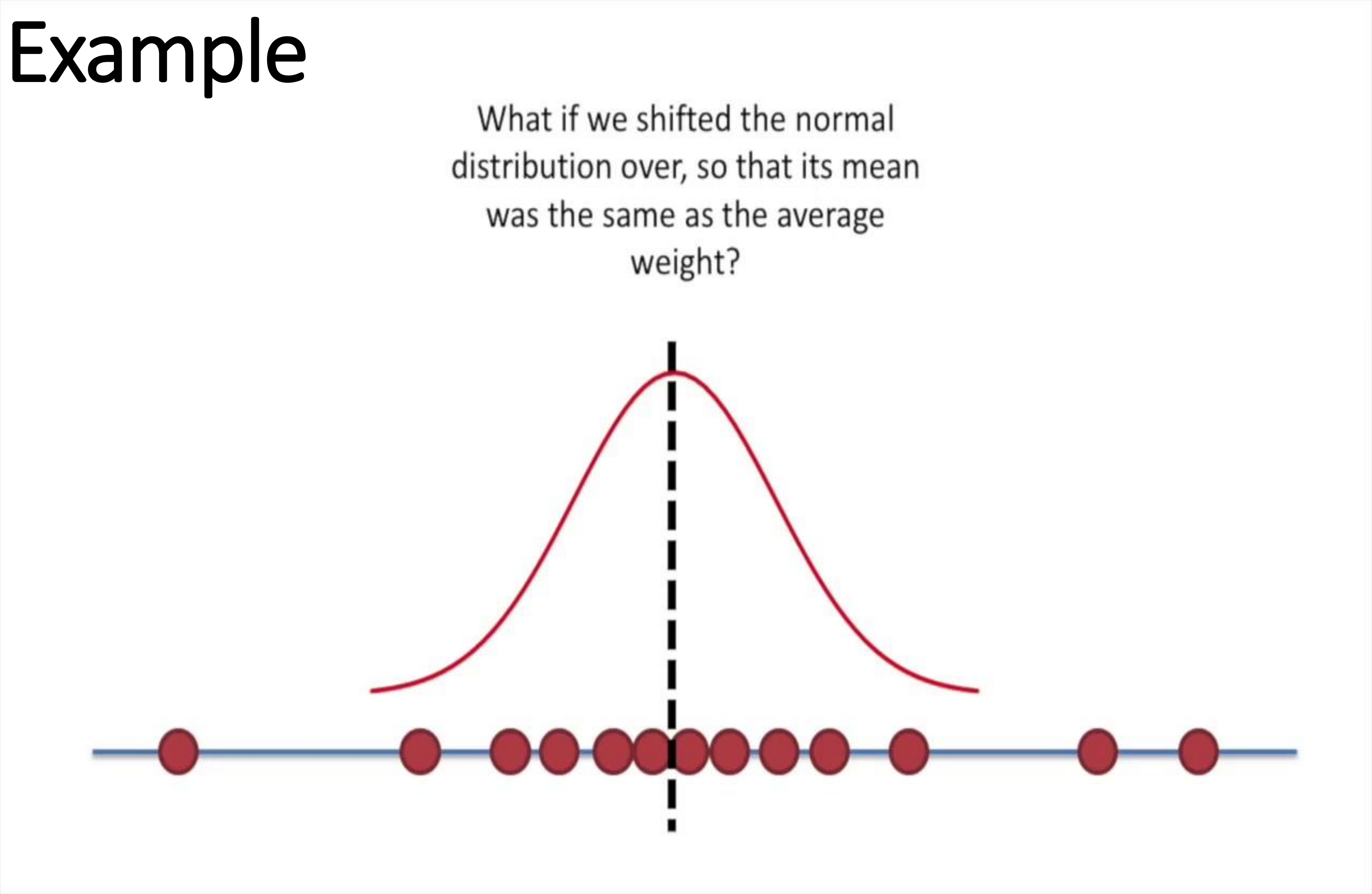
MLE Example


According to a normal distribution with a mean
value over here...

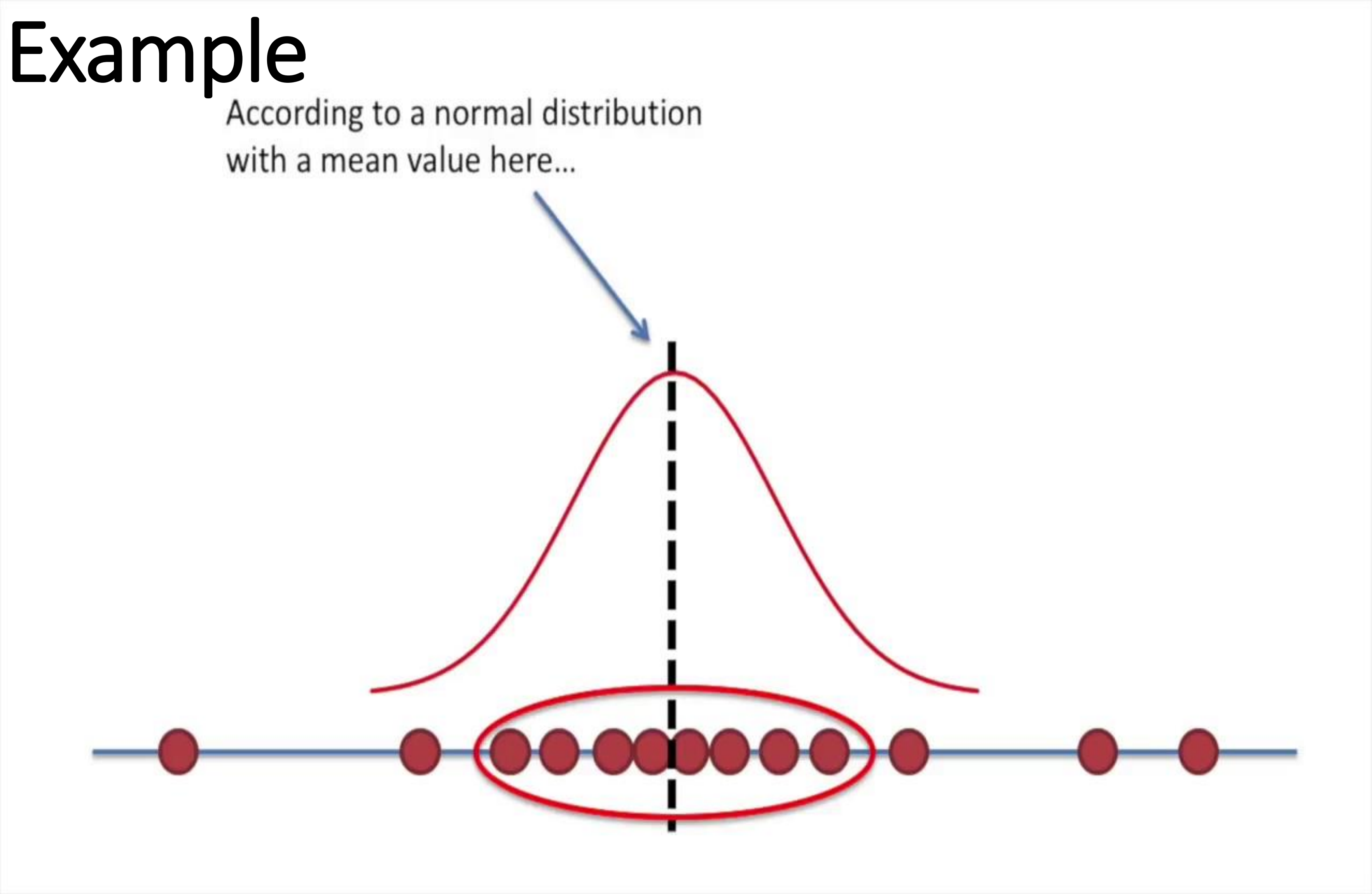
MLE Example


According to a normal distribution with a mean value over here...
...the probability, or "likelihood" of observing all these weights is low.

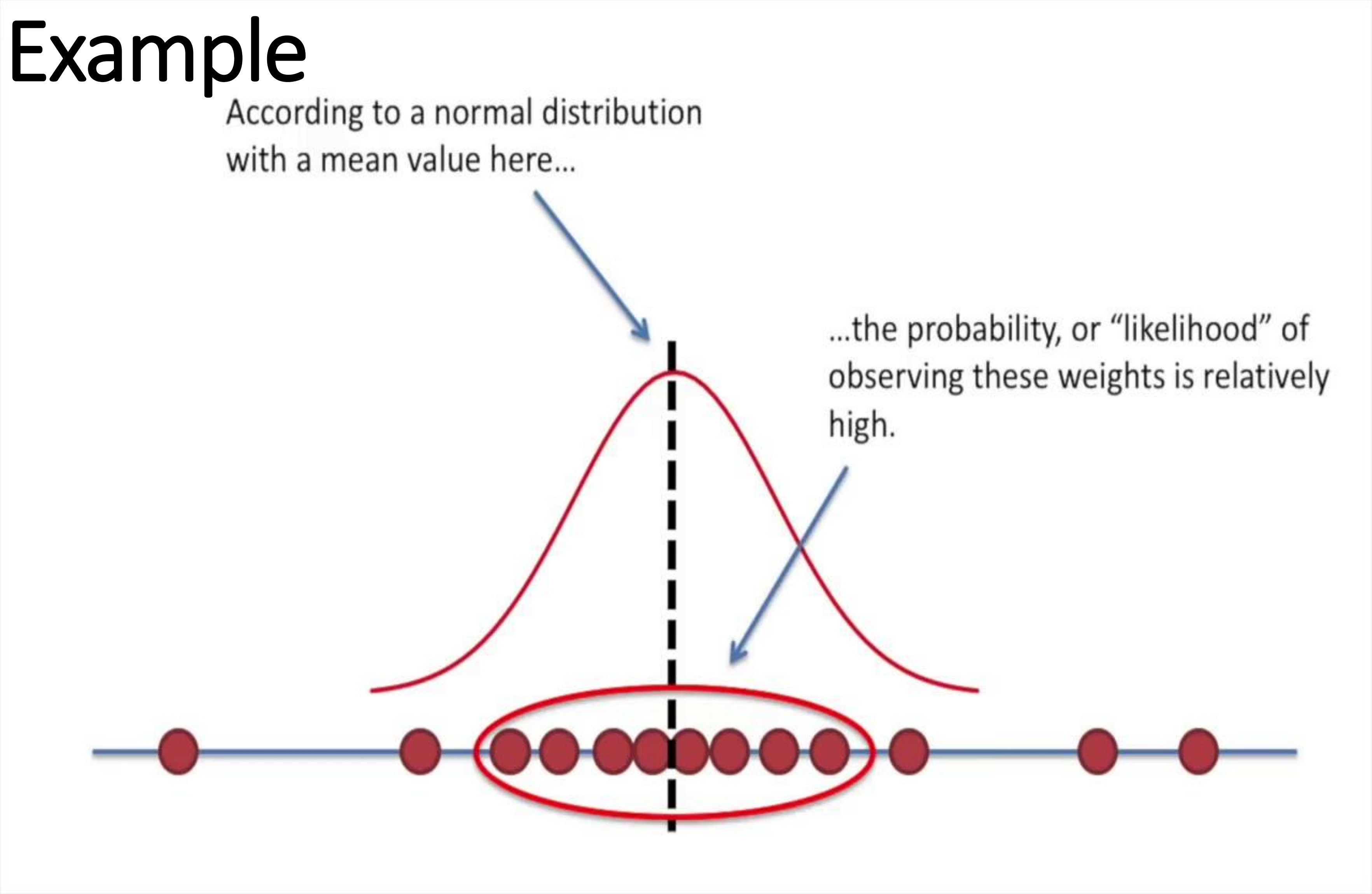
MLE Example


What if we shifted the normal distribution over, so that its mean was the same as the average weight?

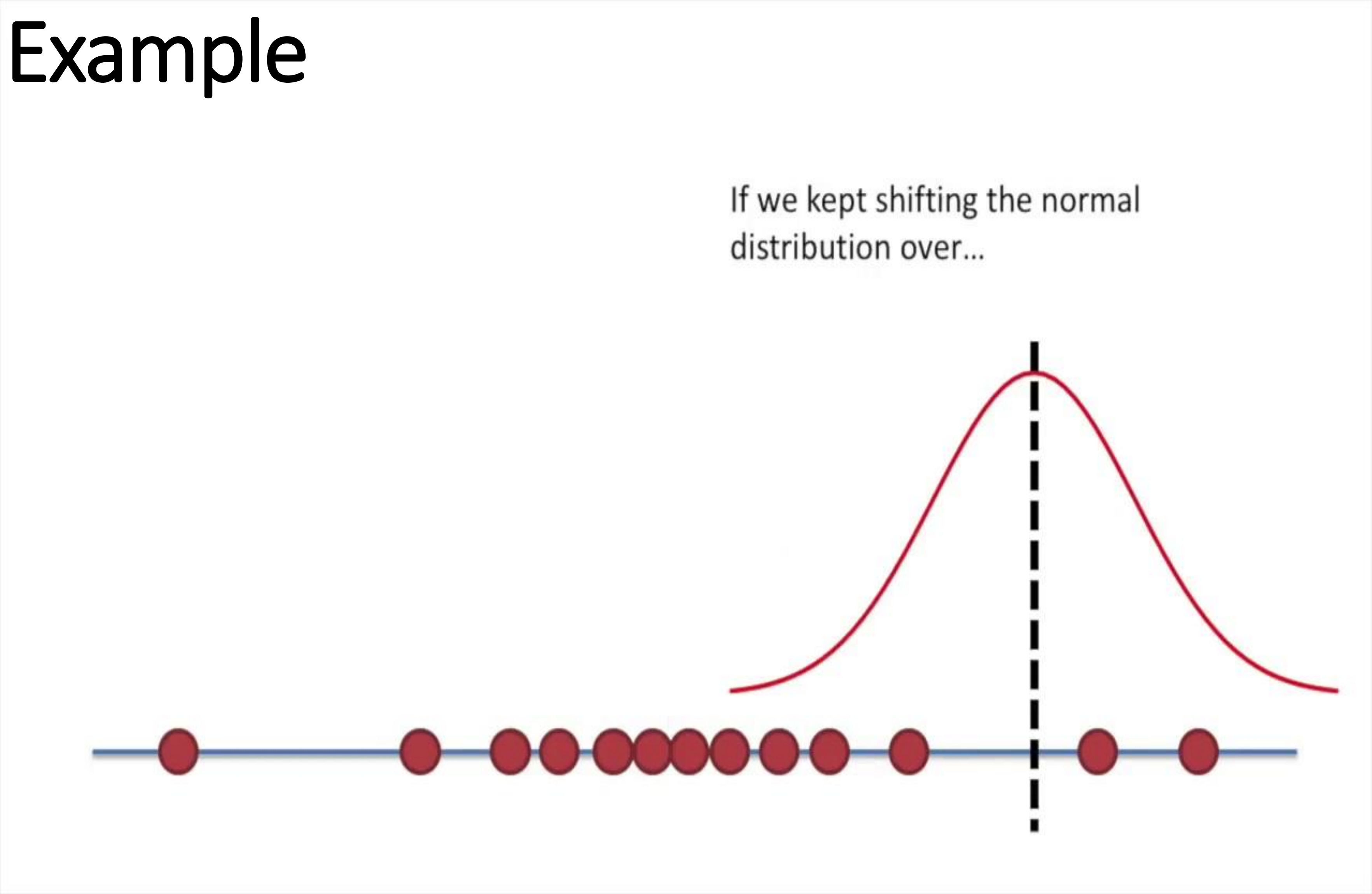
MLE Example



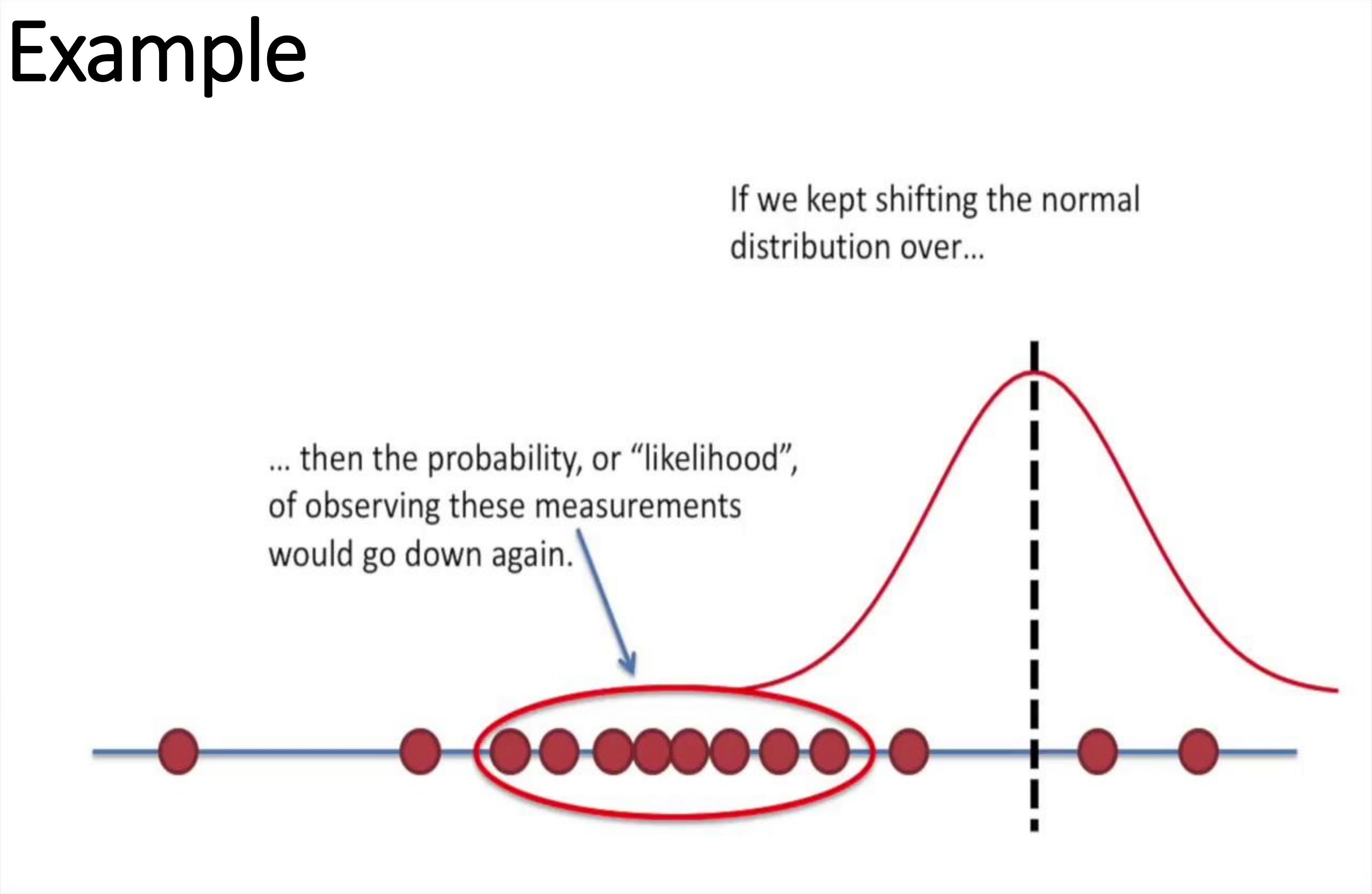
MLE Example


According to a normal distribution
with a mean value here...
...the probability, or "likelihood" of
observing these weights is relatively
high.

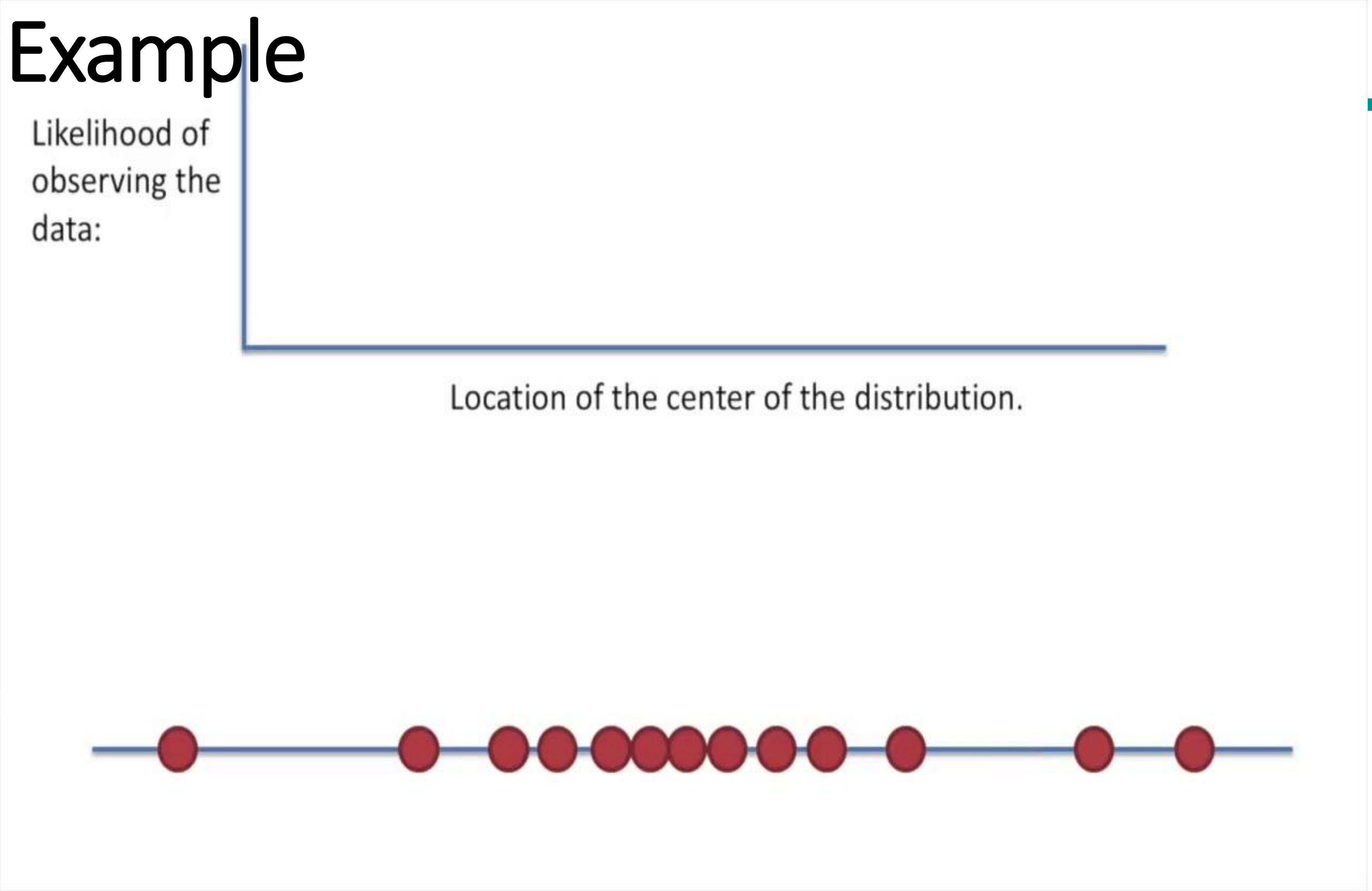
MLE Example



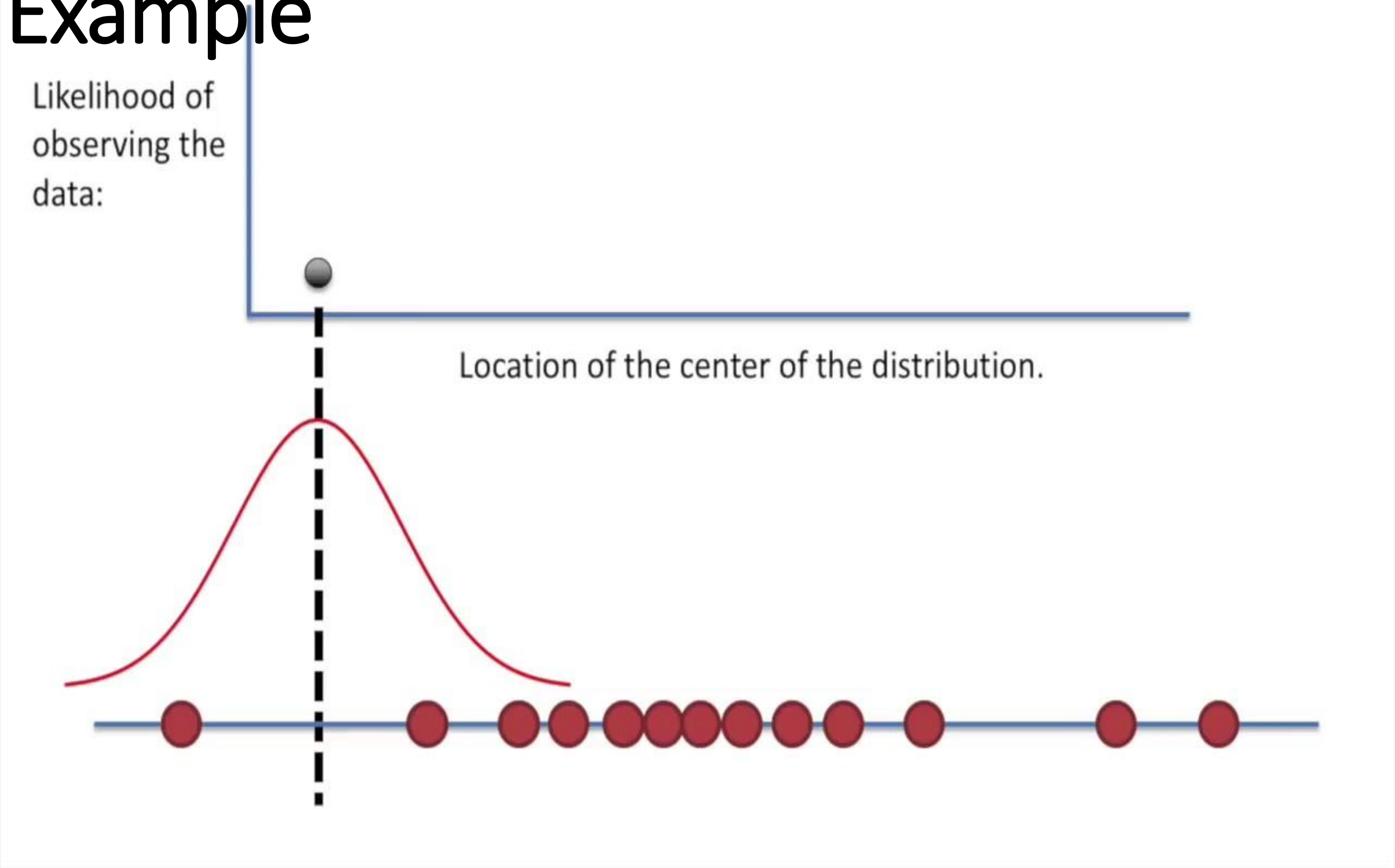
MLE Example


If we kept shifting the normal distribution over...
... then the probability, or "likelihood", of observing these measurements would go down again.

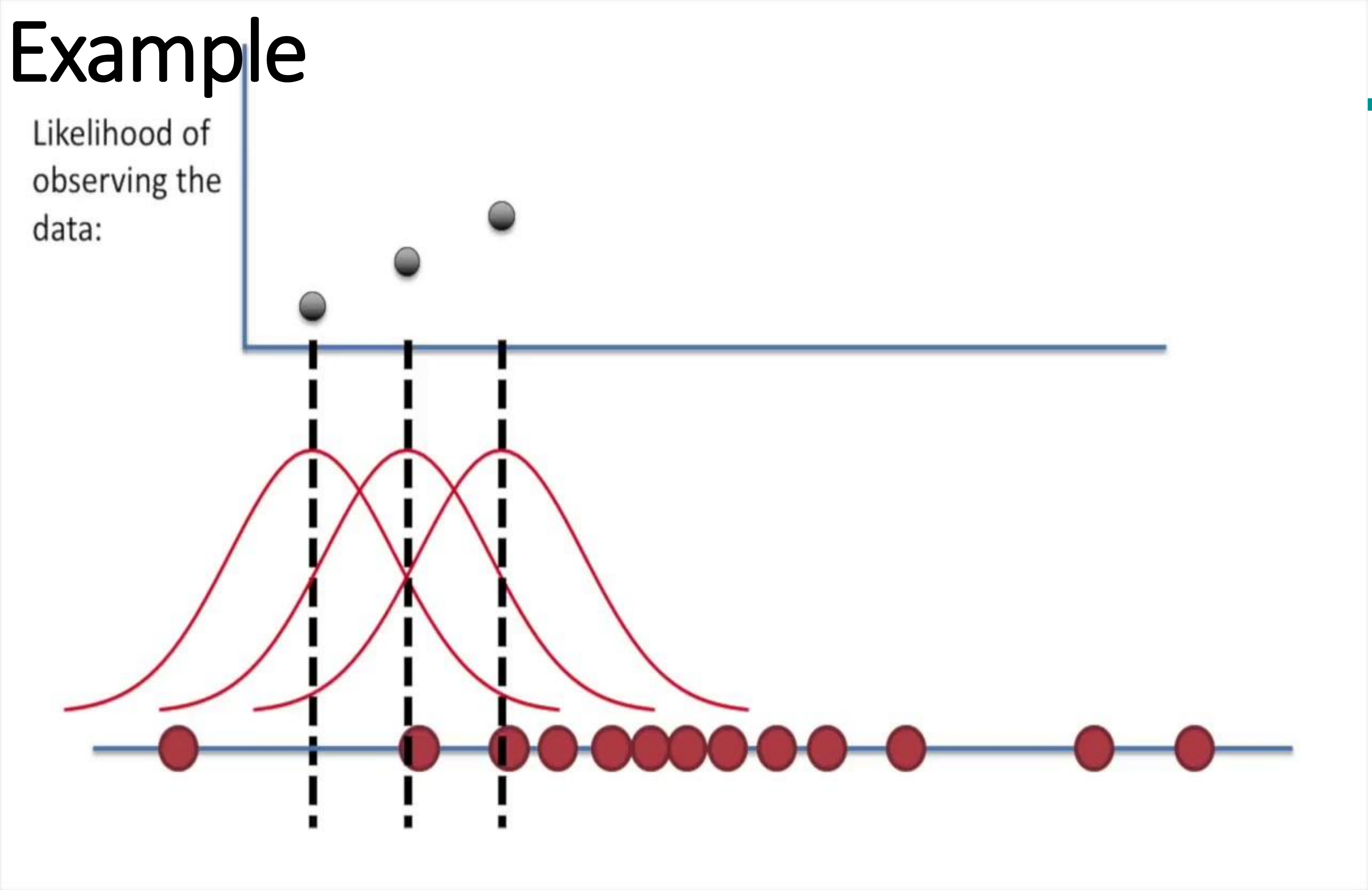
MLE Example


Likelihood of observing the data:
Location of the center of the distribution.

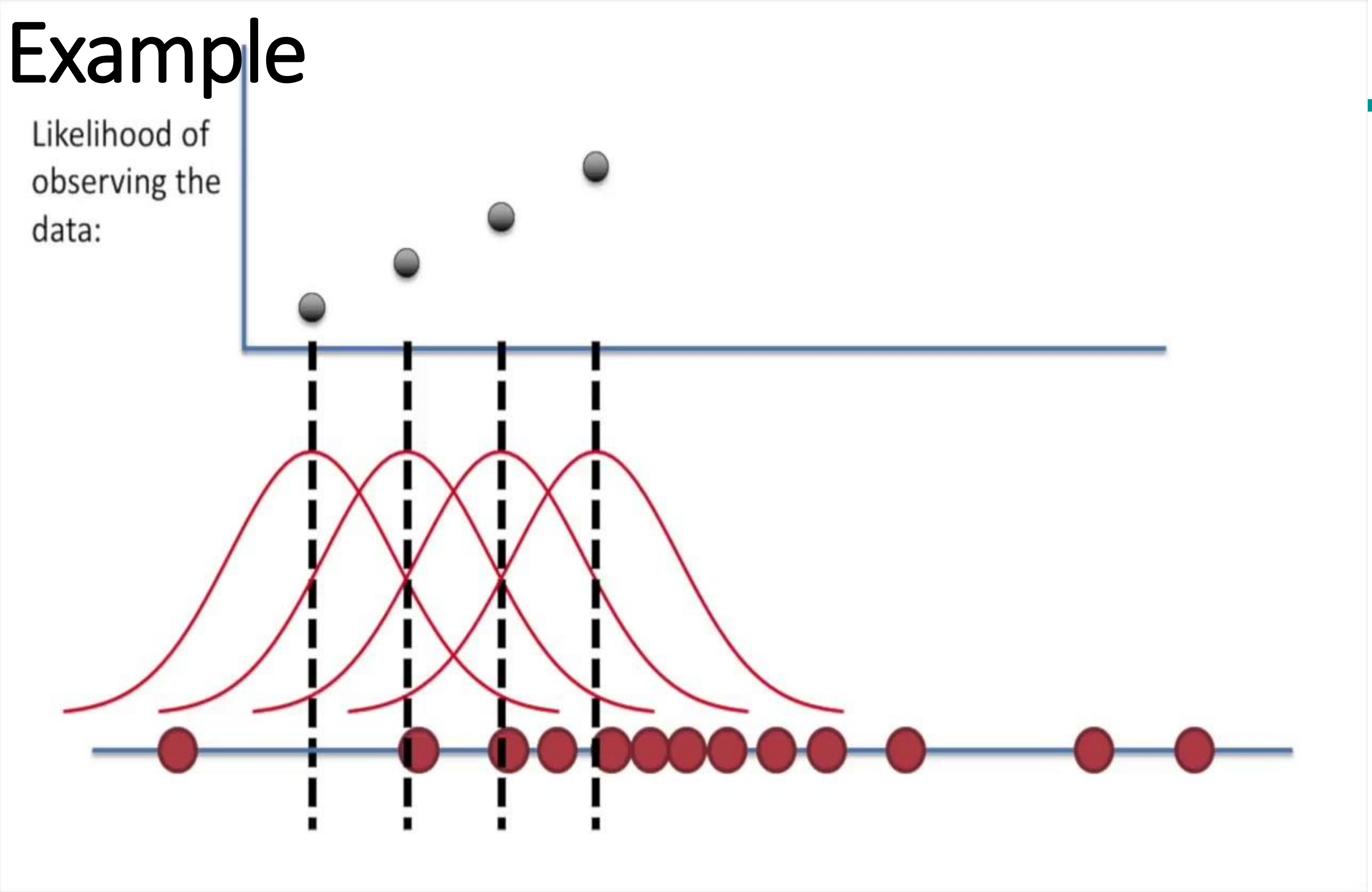
MLE Example


Likelihood of observing the data:
Location of the center of the distribution.

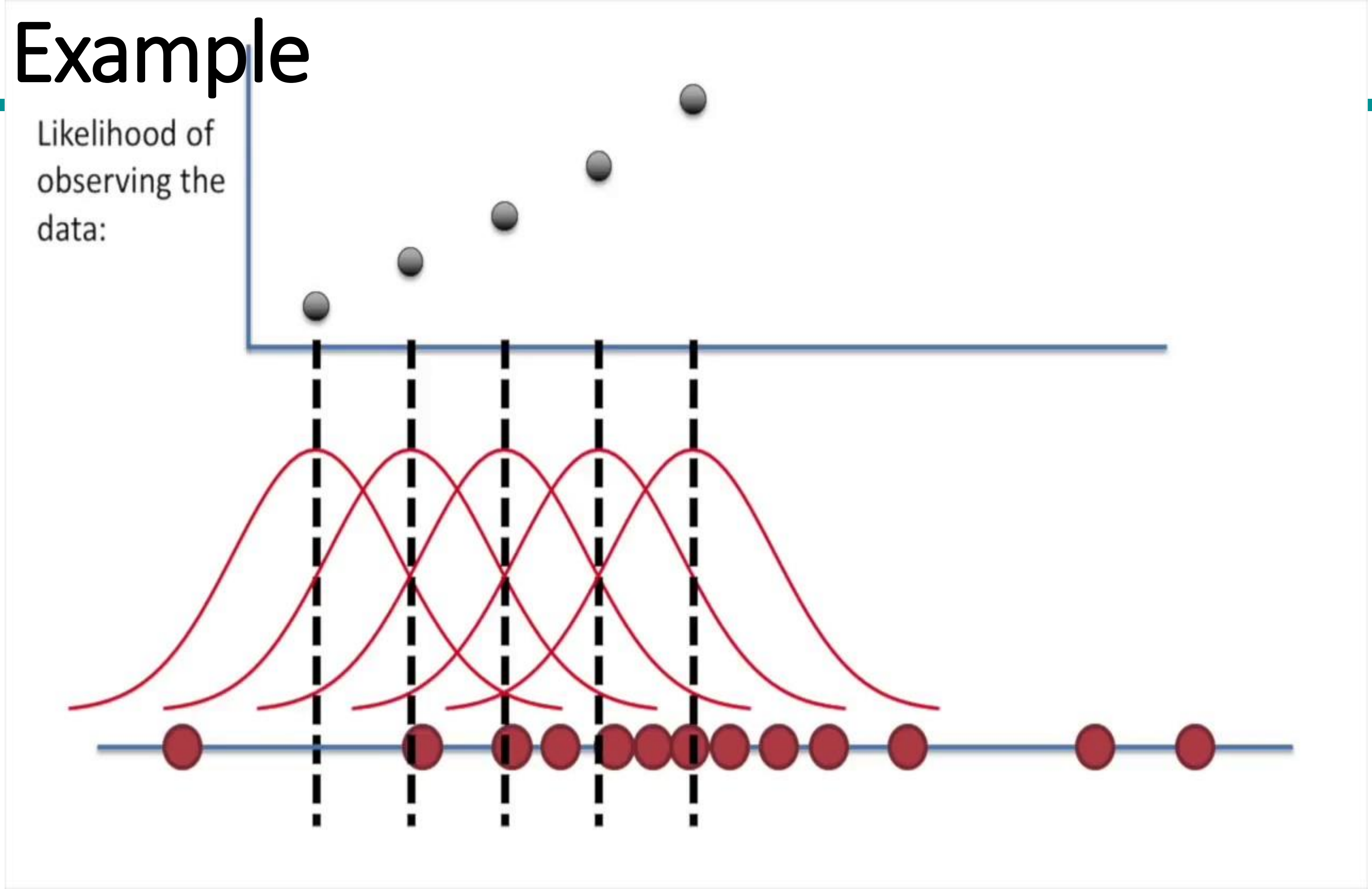
MLE Example



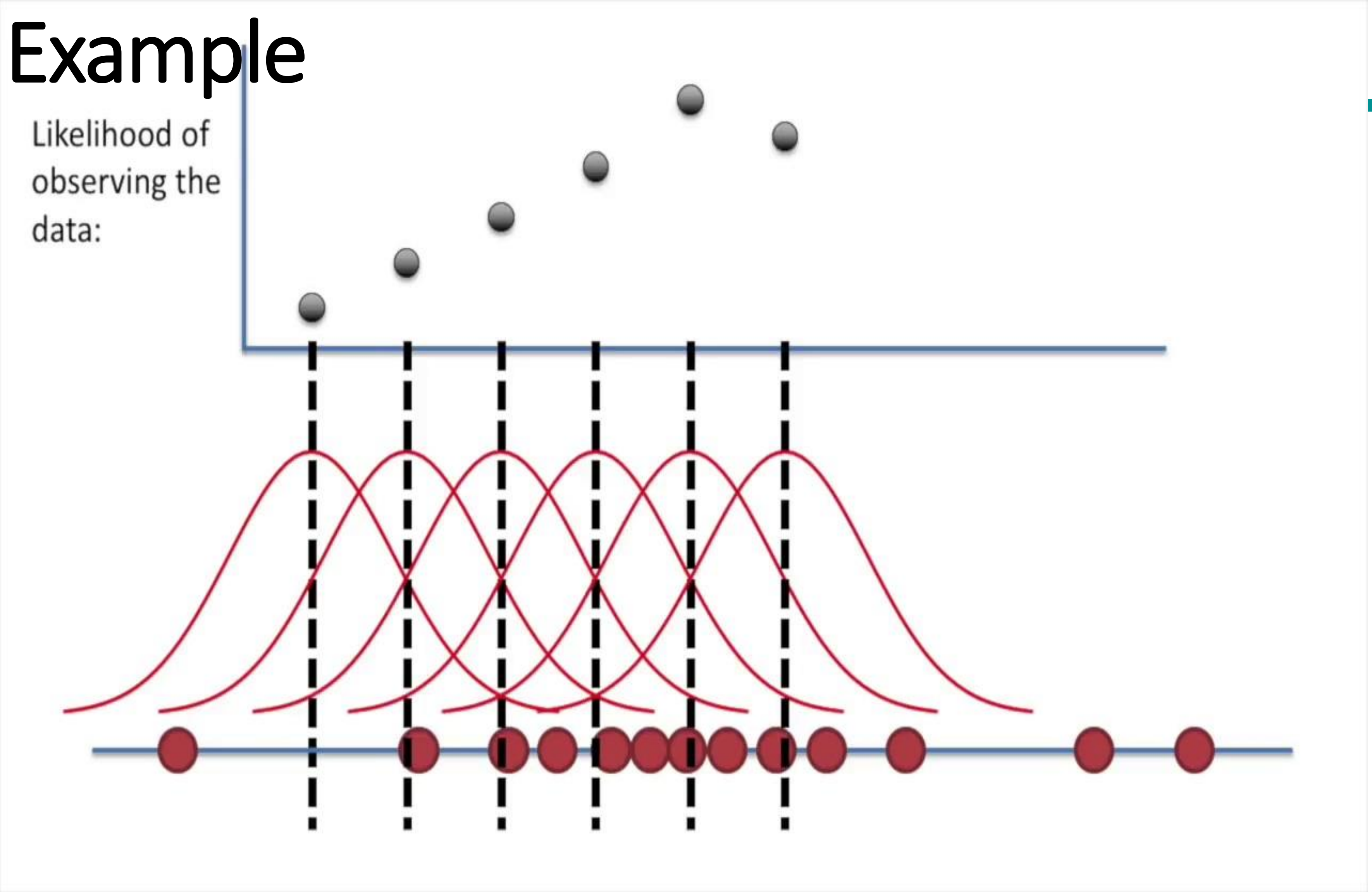
MLE Example



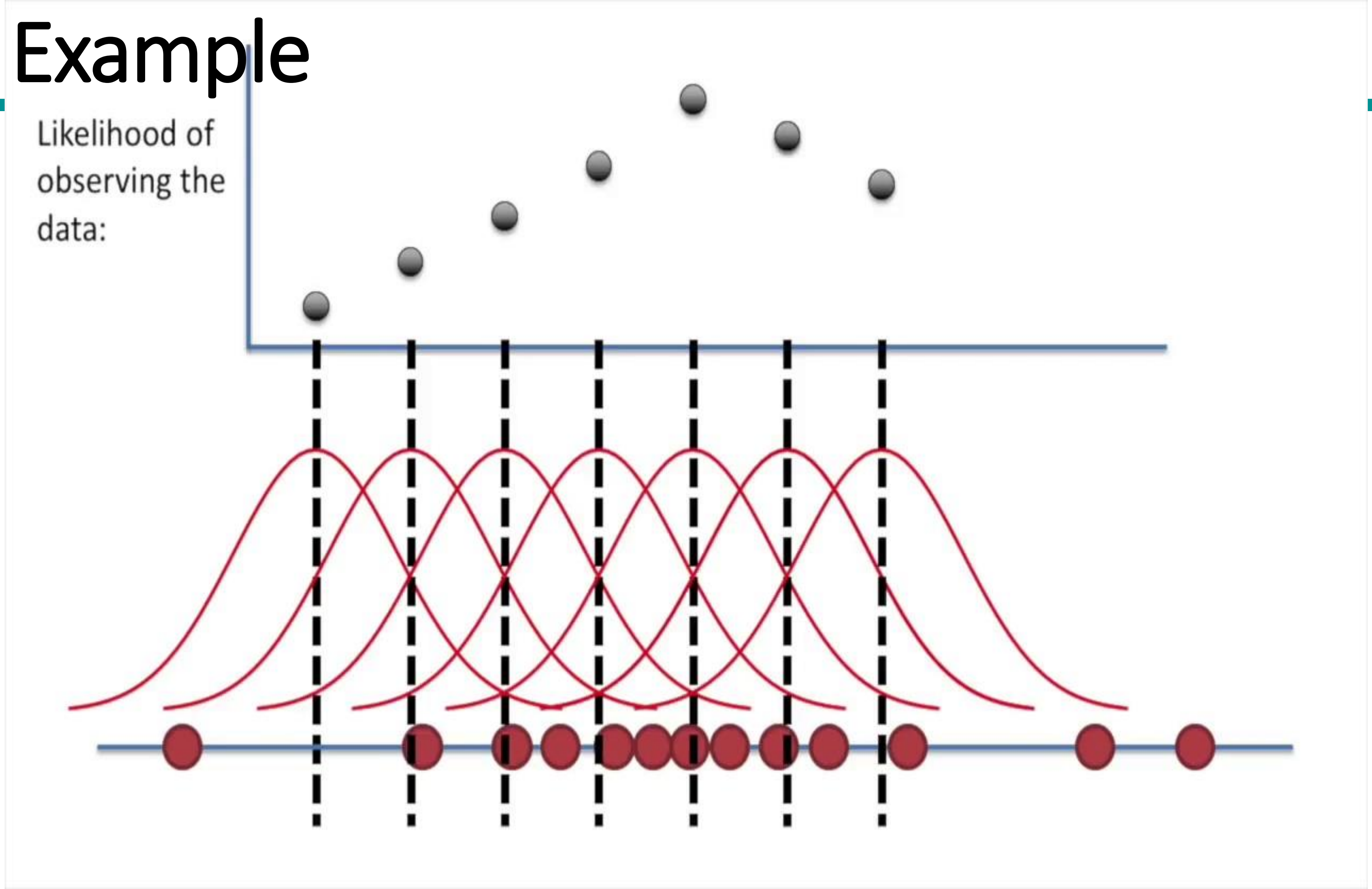
MLE Example



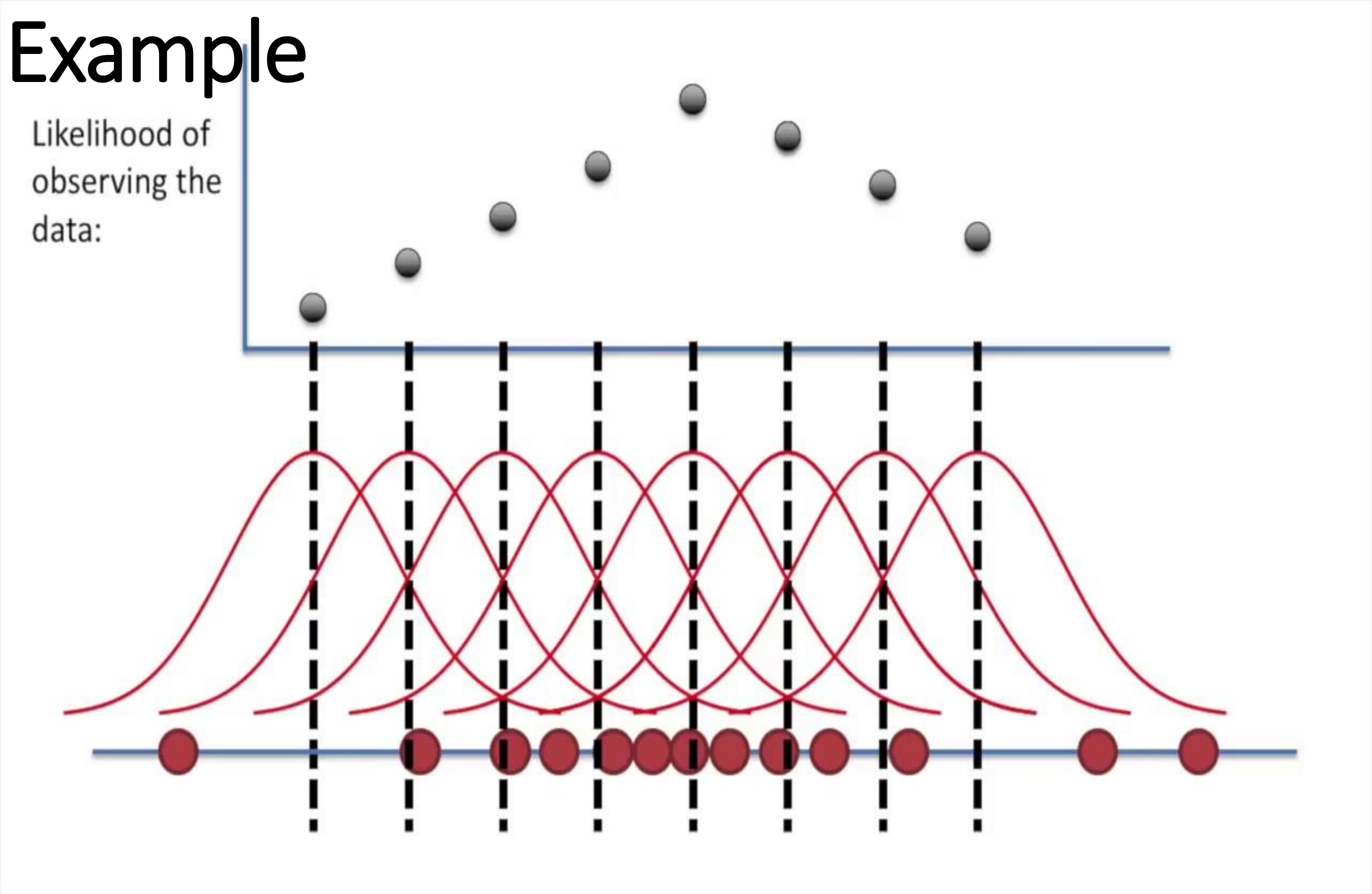
MLE Example



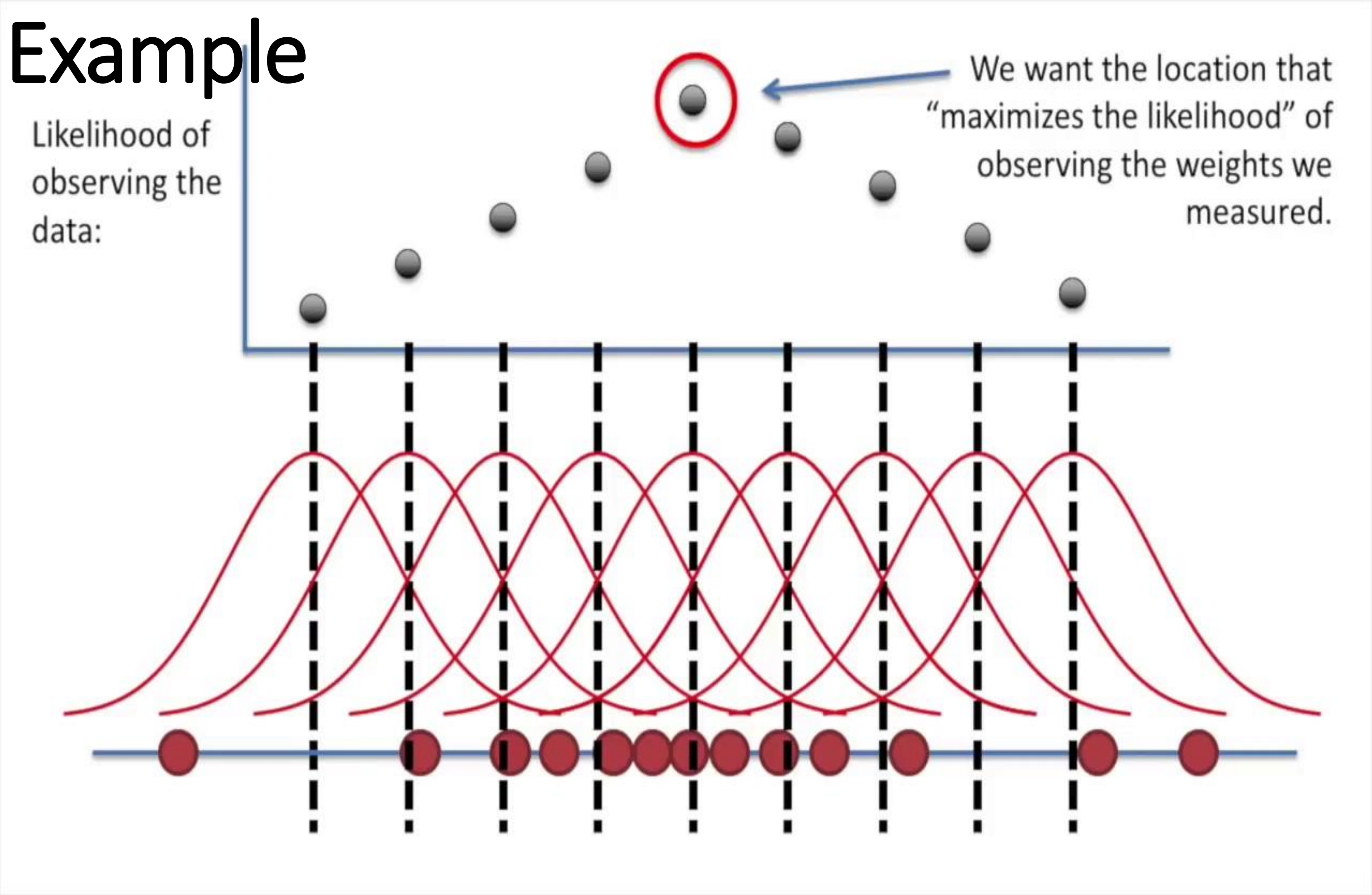
MLE Example



MLE Example

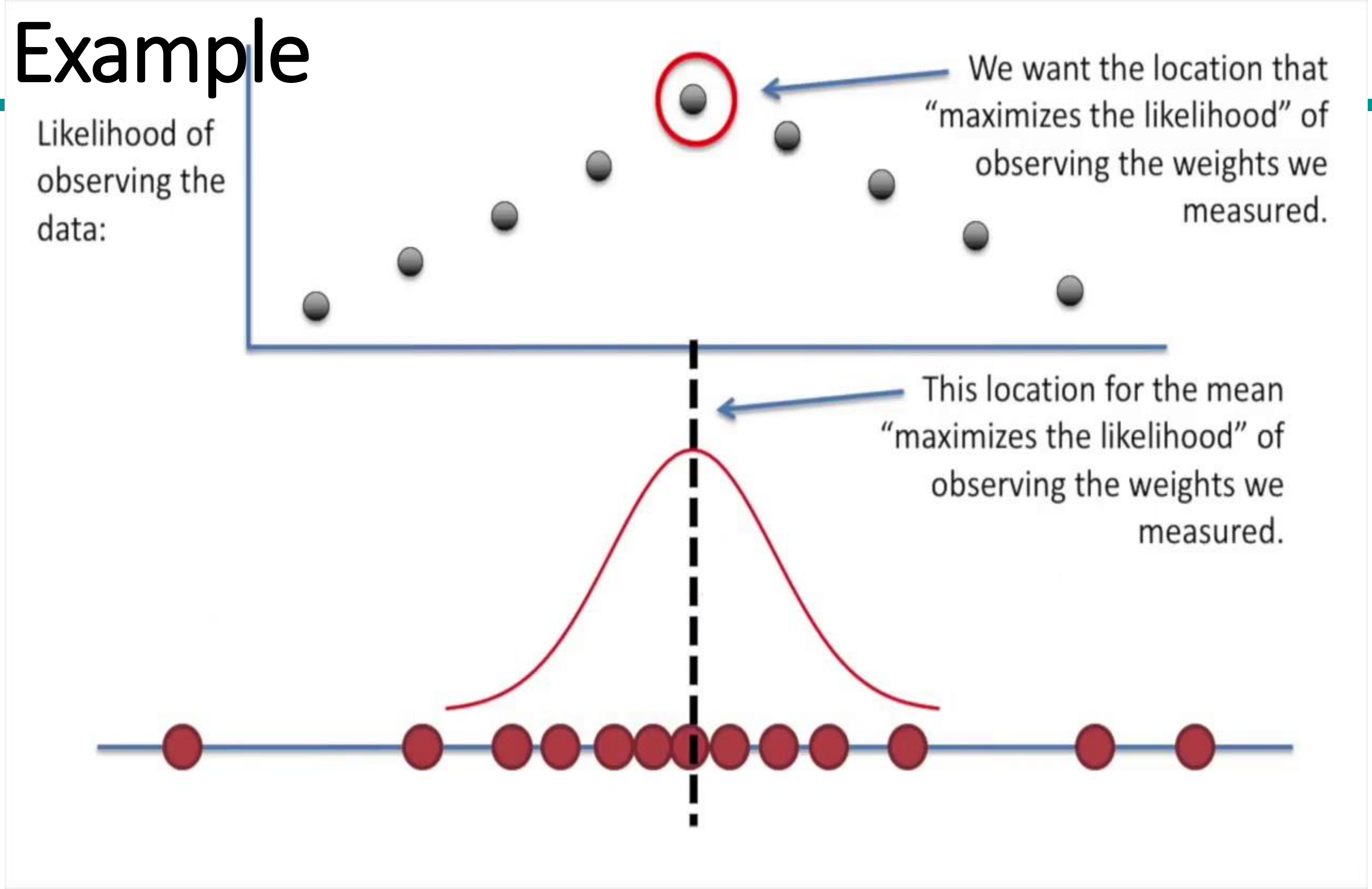


MLE Example



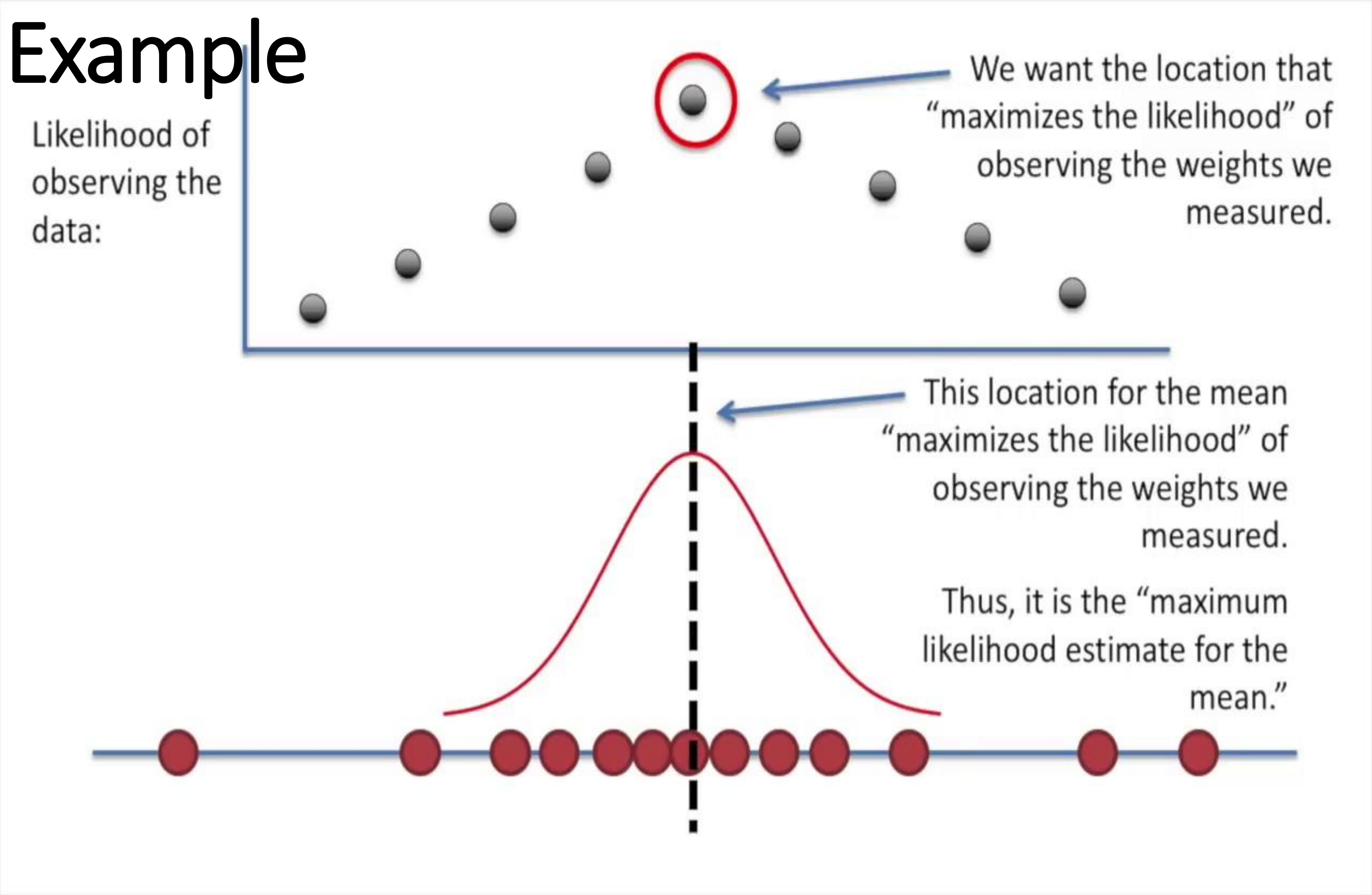
Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.

MLE Example



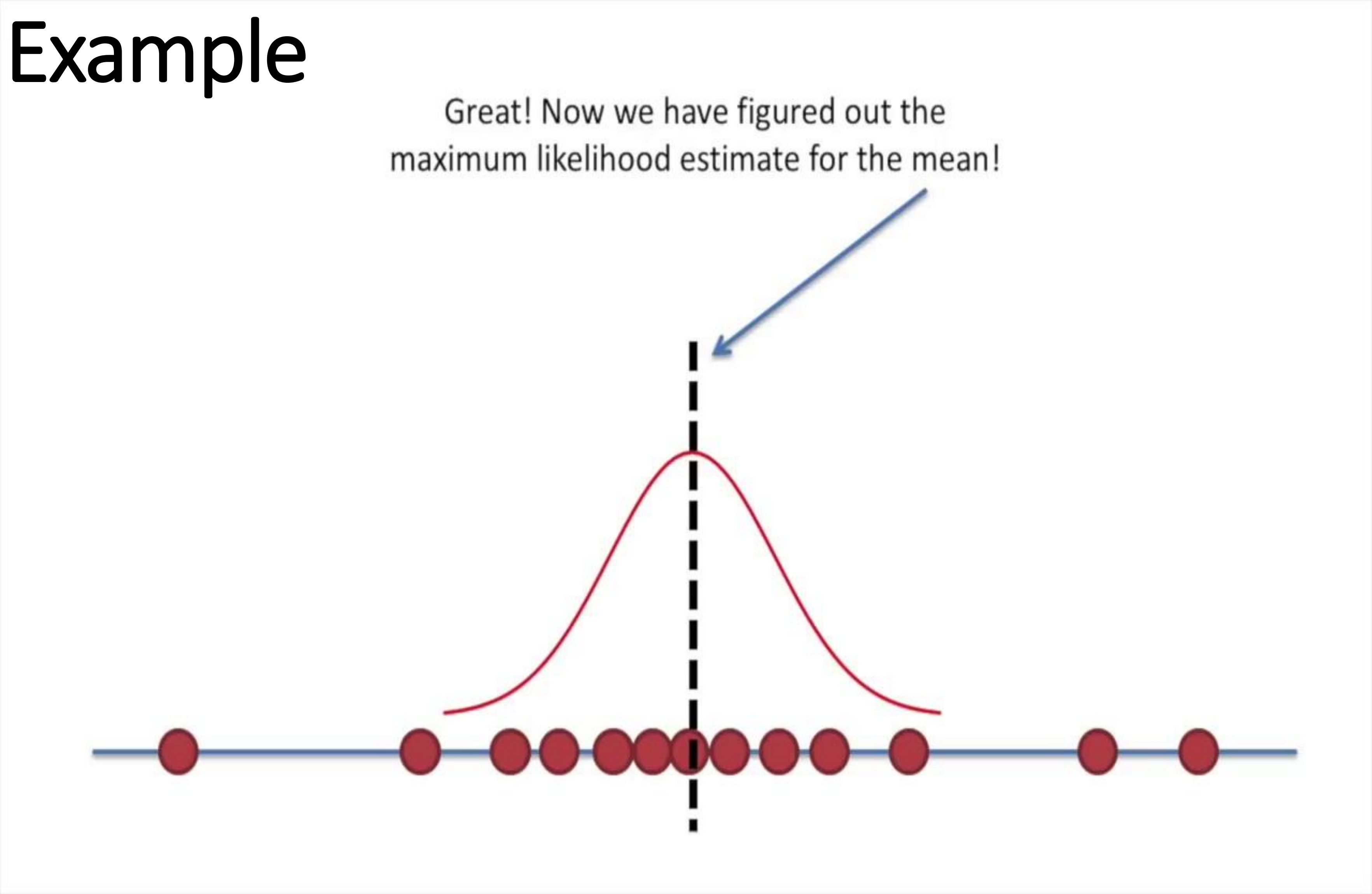
Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.
This location for the mean "maximizes the likelihood" of observing the weights we measured.

MLE Example

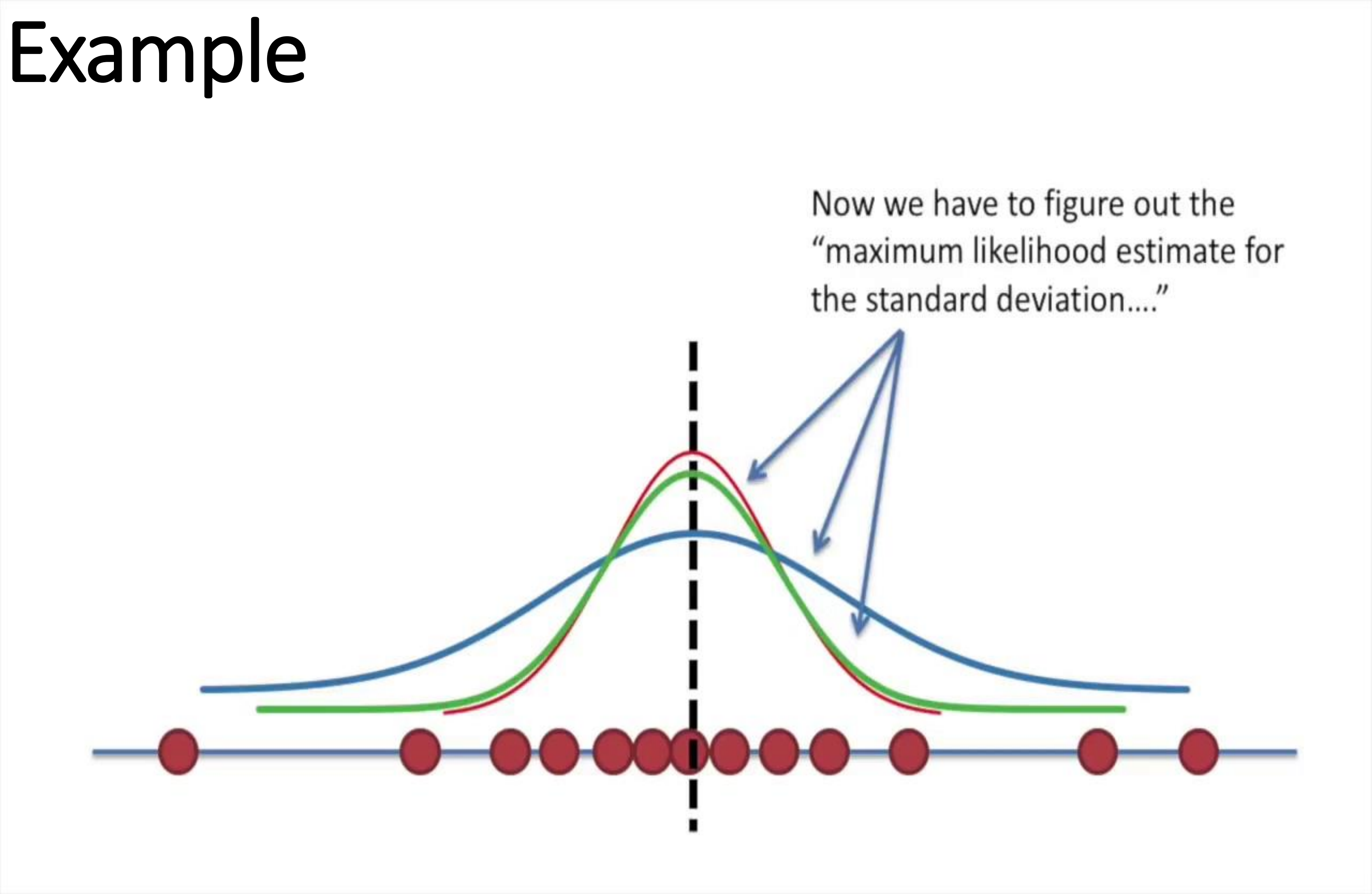


Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.
This location for the mean "maximizes the likelihood" of observing the weights we measured.
Thus, it is the "maximum likelihood estimate for the mean."

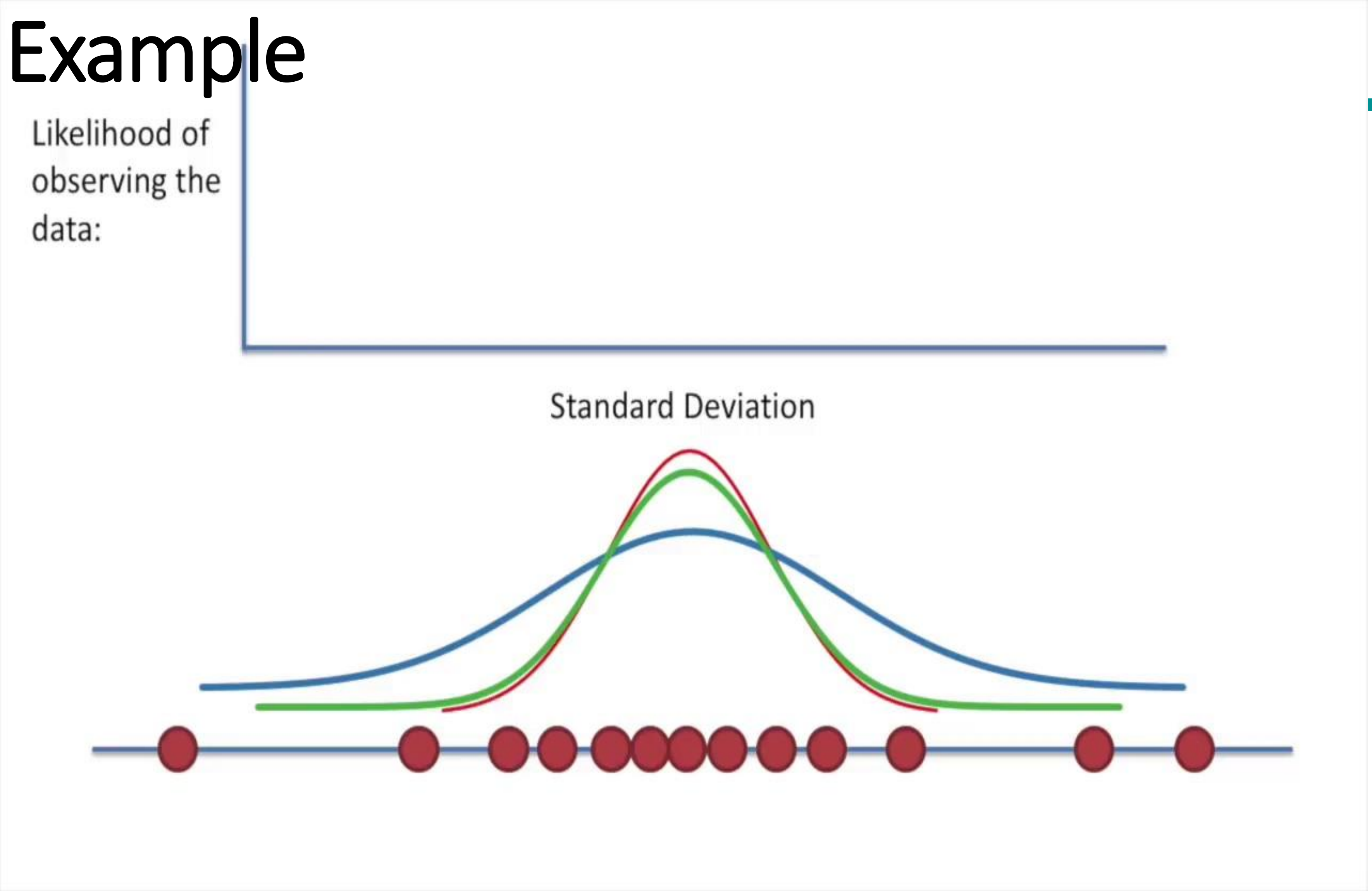
MLE Example


Great! Now we have figured out the
maximum likelihood estimate for the mean!

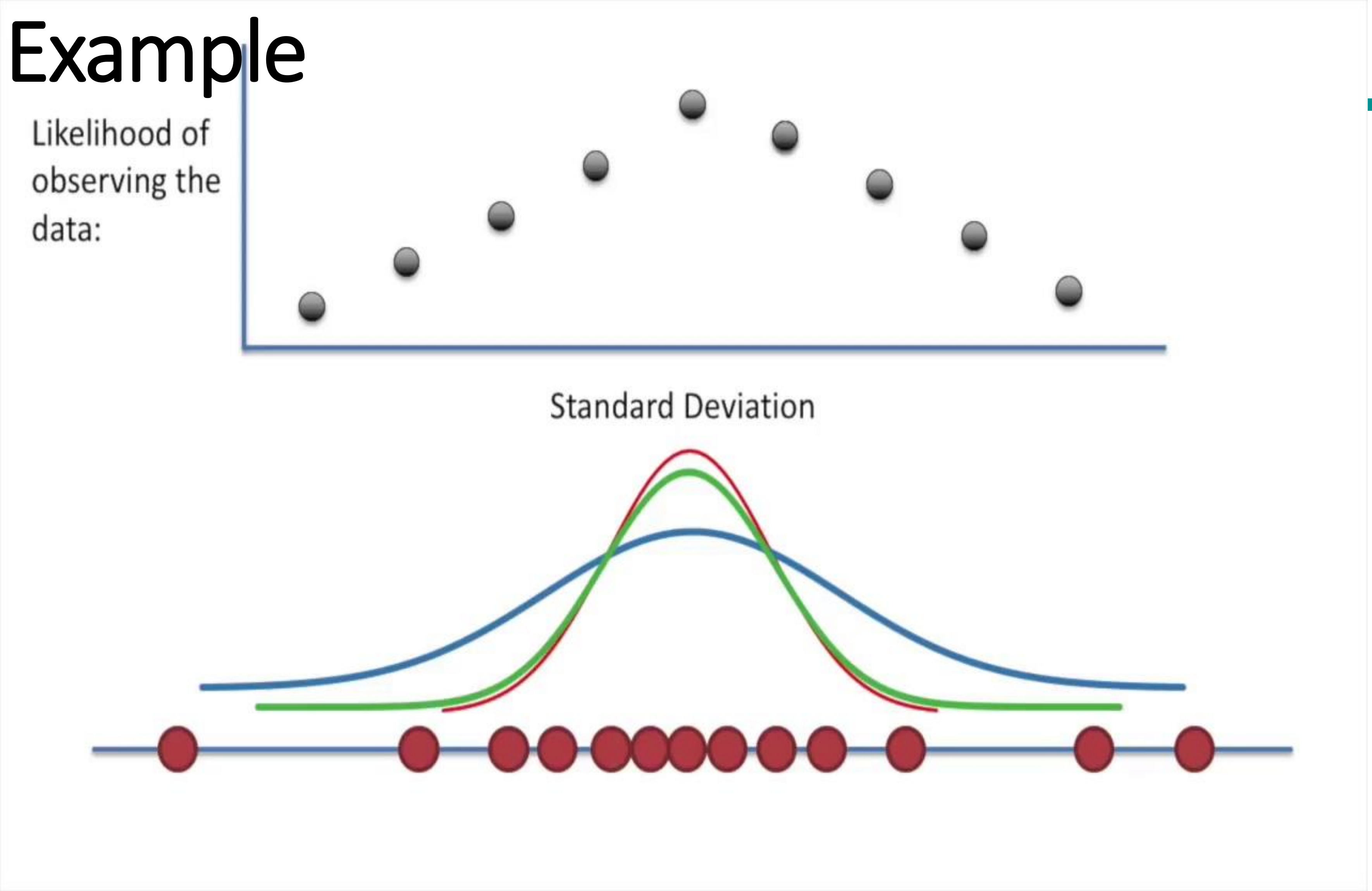
MLE Example


Now we have to figure out the
“maximum likelihood estimate for
the standard deviation….”

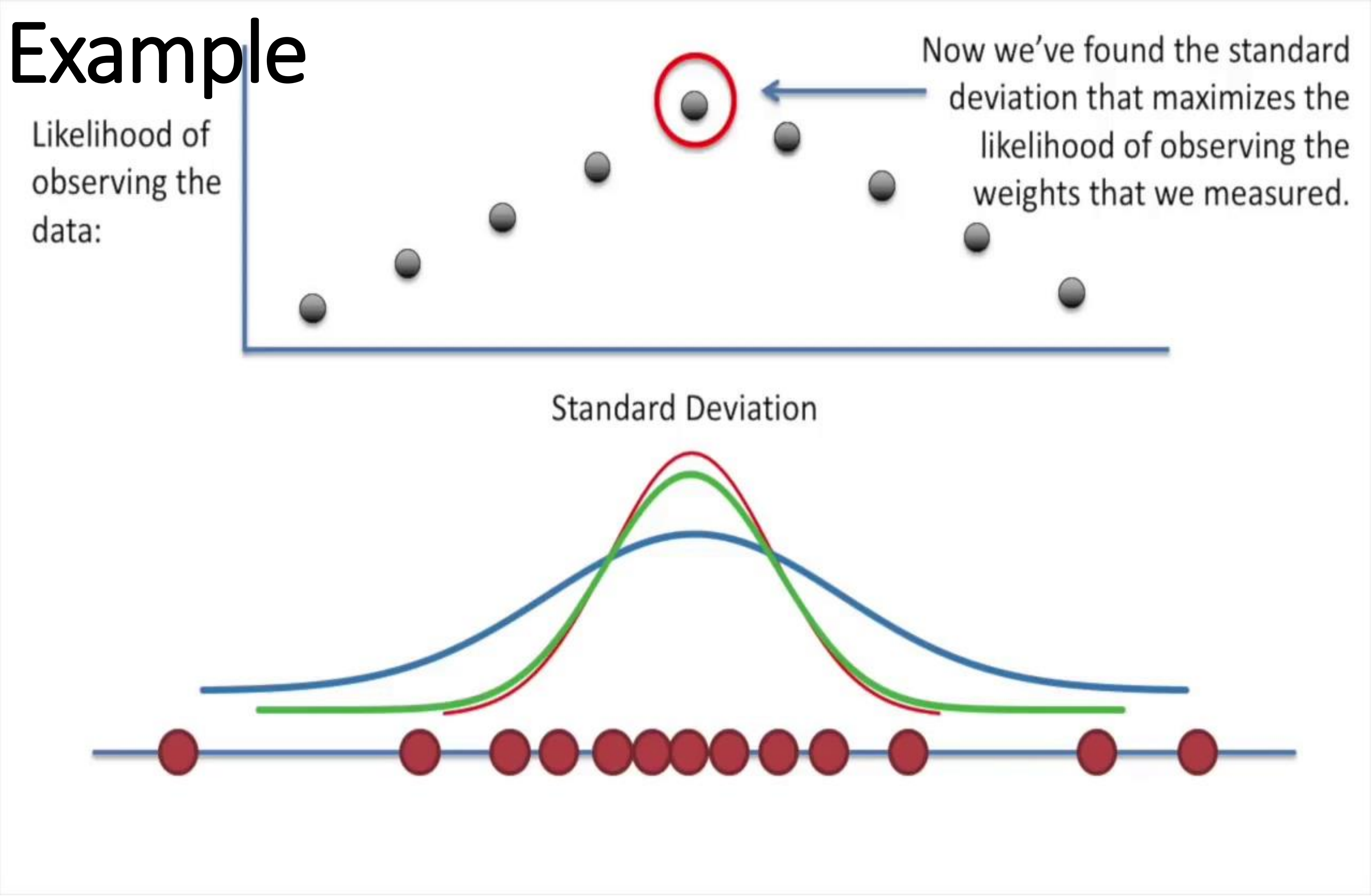
MLE Example



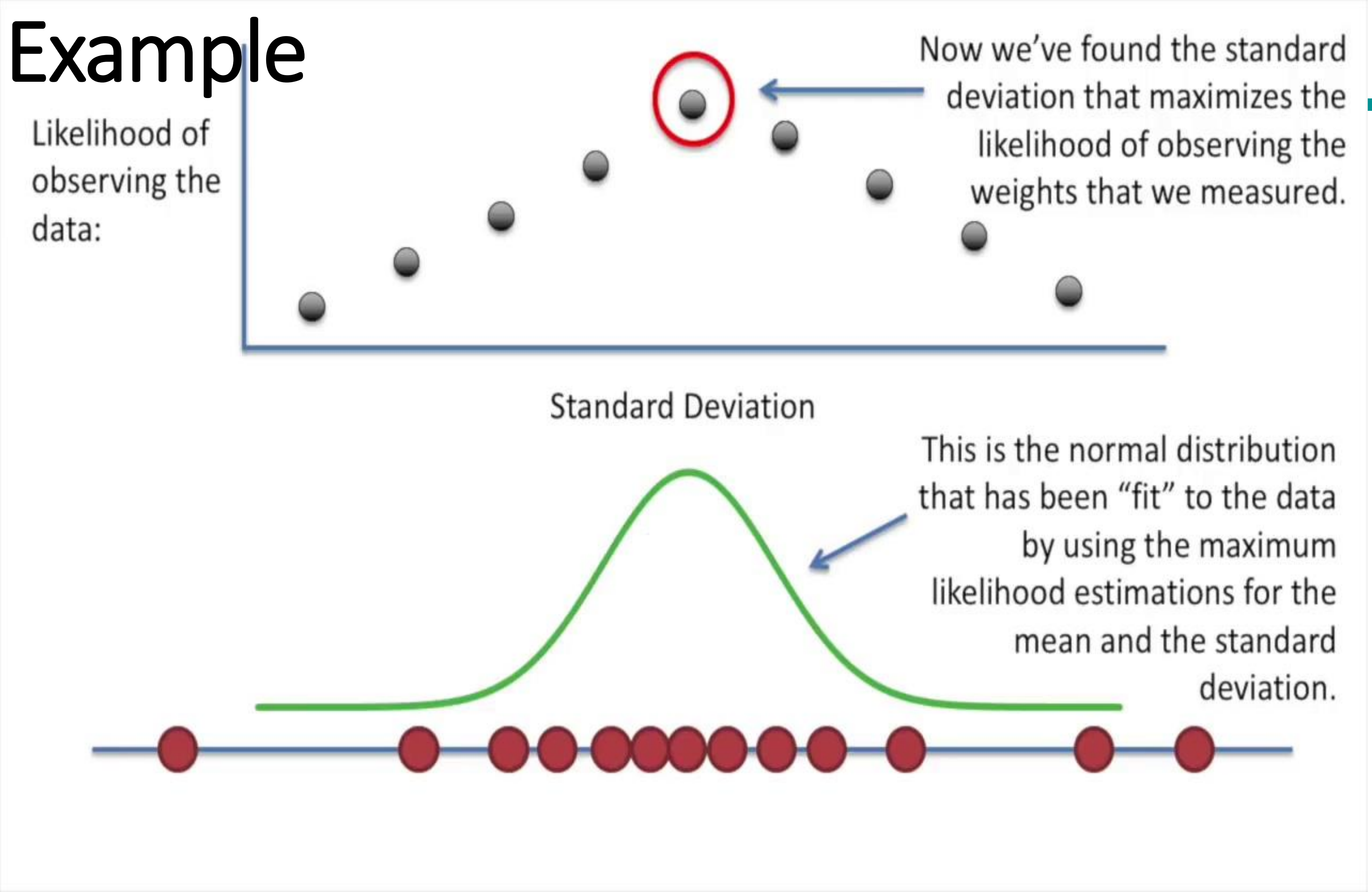
MLE Example



MLE Example

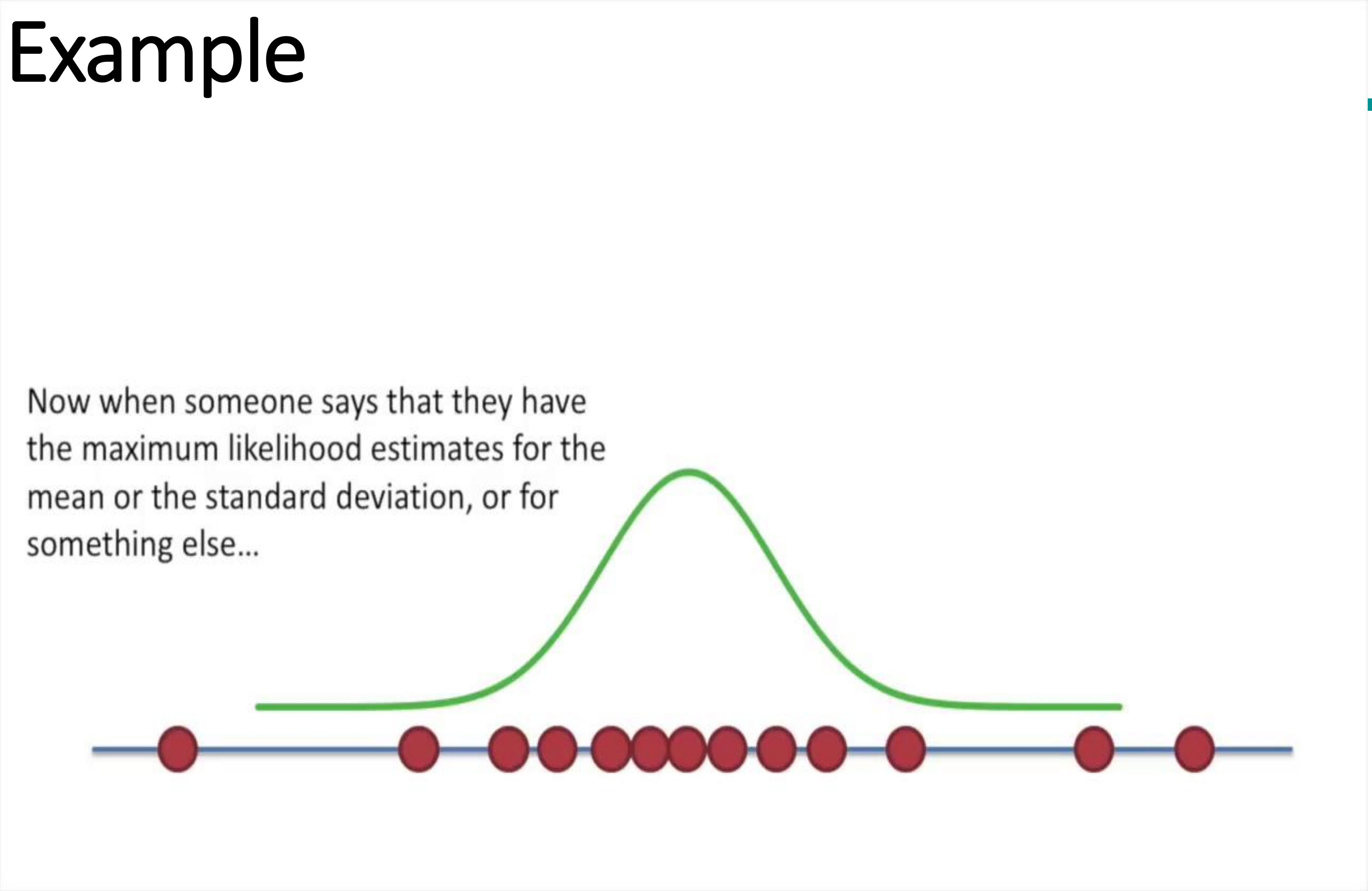

Likelihood of observing the data:
Now we've found the standard deviation that maximizes the likelihood of observing the weights that we measured.
Standard Deviation

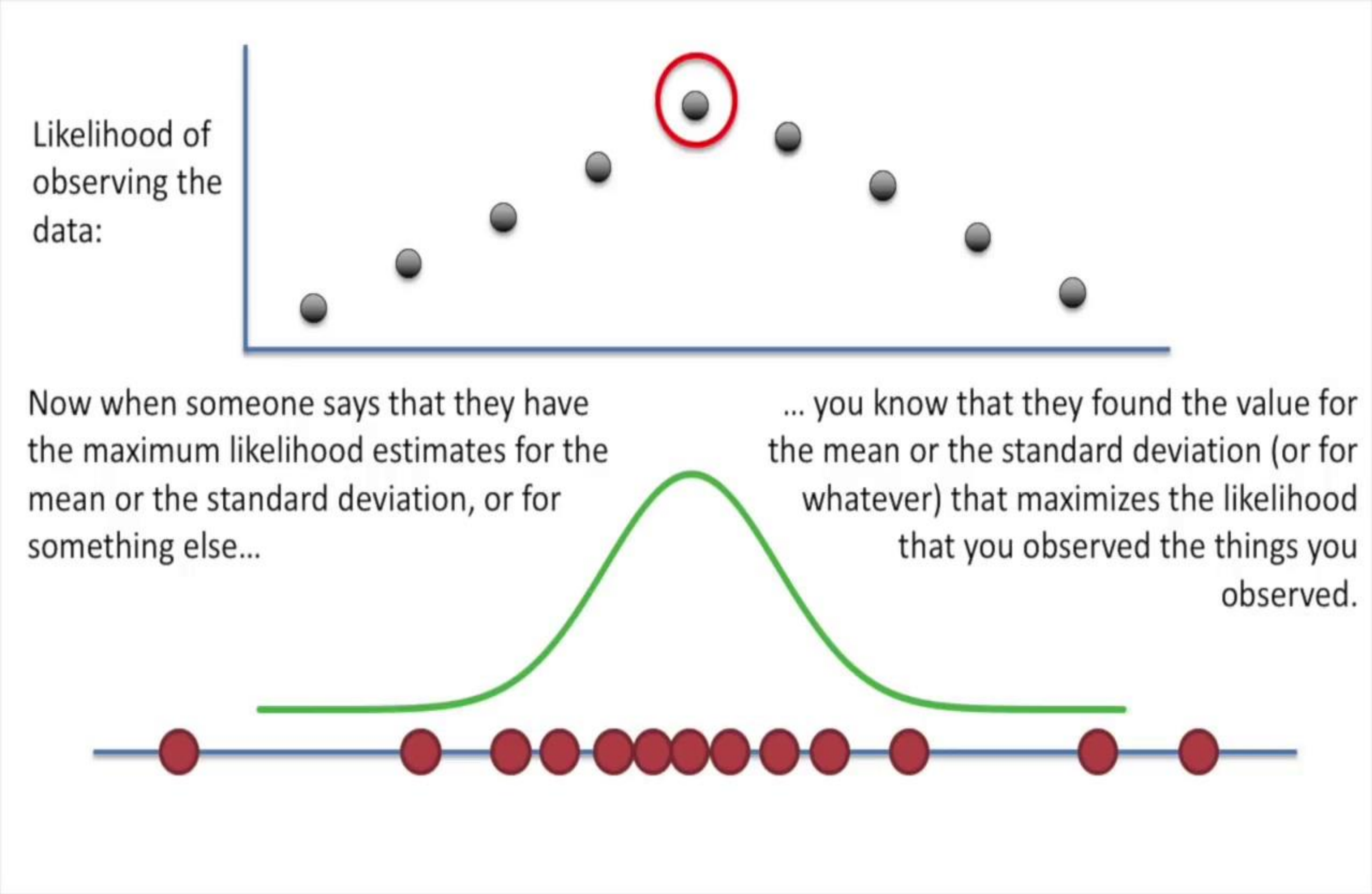
MLE Example



Likelihood of observing the data:
Now we've found the standard deviation that maximizes the likelihood of observing the weights that we measured.
Standard Deviation
This is the normal distribution that has been "fit" to the data by using the maximum likelihood estimations for the mean and the standard deviation.

MLE Example


Now when someone says that they have the maximum likelihood estimates for the mean or the standard deviation, or for something else...


Likelihood of observing the data:
Now when someone says that they have the maximum likelihood estimates for the mean or the standard deviation, or for something else...
... you know that they found the value for the mean or the standard deviation (or for whatever) that maximizes the likelihood that you observed the things you observed.

Calculating the MLE

- Example: we have three data points 9, 9.5, 11
- We want to calculated the total probability of observing all the data, i.e. the joint probability distribution of all observed data points.
- Assumption: each data point is generated independently from the others.
- If the events are independent, then the total probability of observing all the data is the product of observing each data point individually (i.e. the product of the marginal probabilities).

Calculating the MLE

- Probability of observing a single data point *x*

$$P(x; \underbrace{\mu, \sigma}_{\text{Parameters}}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Example:

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

The Log Likelihood

- Maximum is found by differentiation, i.e., find the derivative of the function w.r.t. a variable, set it to zero and find the required value.
- Since the previous expression is not easy to differentiate, we simplify the calculus considering the natural logarithm of the expression.

$$\ln(P(x;\mu,\sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

$$\ln(P(x;\mu,\sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

The Log Likelihood

- This expression can be easily differentiated to find the maximum.

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

- The same can be done for the standard deviation.

MLE Summary

- MLE is a general approach to estimating parameters by maximizing the likelihood function defined as
  - $L(\theta|X) = f(X|\theta)$
- that is the probability of obtaining X given the parameters θ.
- Knowing the likelihood function L you can look for θ that maximizes the probability of obtaining the data you have.
- Sometimes we have known estimators, e.g. arithmetic mean is a MLE estimator for μ parameter for normal distribution
- In other cases, you can obtain the best parameter values using different methods that include using optimization algorithms.

MLE and GD

- You can obtain MLE using different methods.
- Using an optimization algorithm like GD is one of them.
- On the other hand, GD can also be used to maximize functions other than likelihood function.