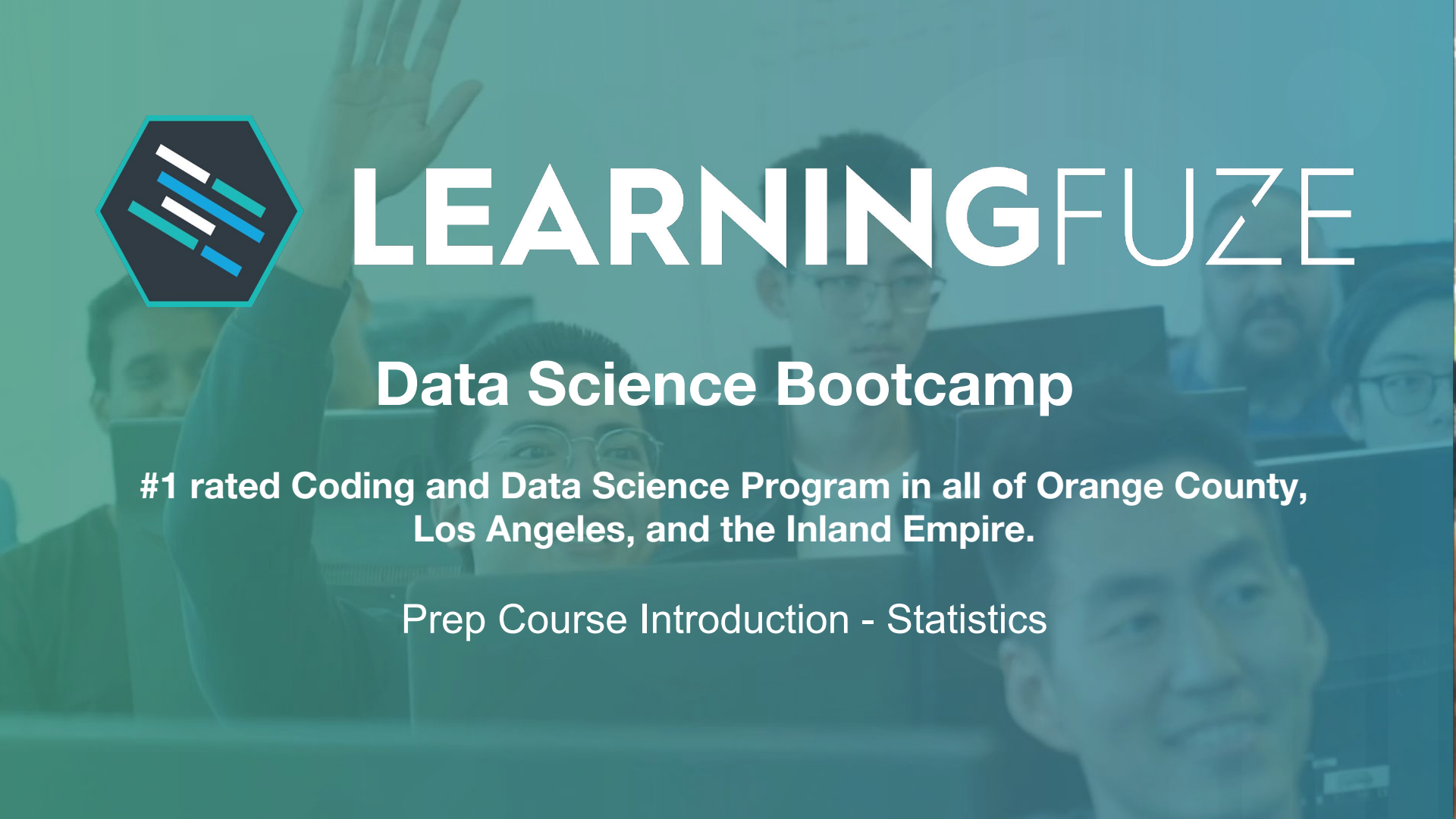


LEARNINGFUZE

Data Science Bootcamp

#1 rated Coding and Data Science Program in all of Orange County, Los Angeles, and the Inland Empire.

Prep Course Introduction - Statistics

LEARNINGFUZE

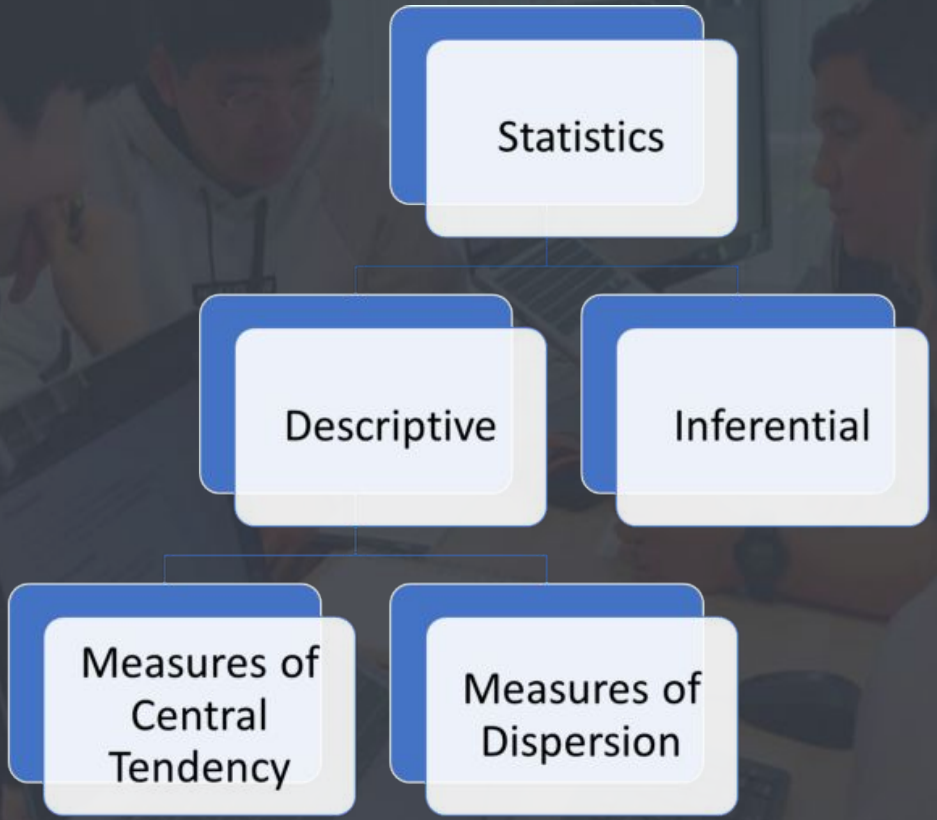
Statistics

Statistics is the discipline that is concerned with the collection, organization, analysis, interpretation, and presentation of data.

Source: https://en.wikipedia.org/wiki/Statistics

Statistics is a science of collecting and analyzing data taken from a sample population

Statistics


Statistics
Descriptive
Inferential
Measures of Central Tendency
Measures of Dispersion

Population vs Sample

- **Population**

Population is the broader group of people to whom you intend to generalize the results

- **Sample**

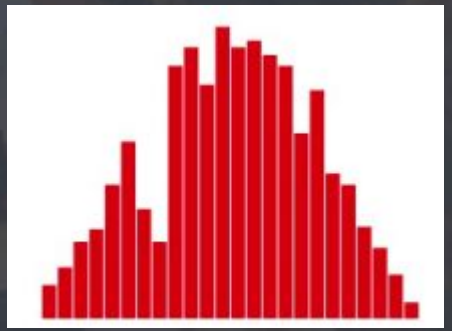
Sample is the specific group that you collect data from. To be a truly random sample, every subject in your target population must have an equal chance of being selected in your sample.

Population vs Sample

Population

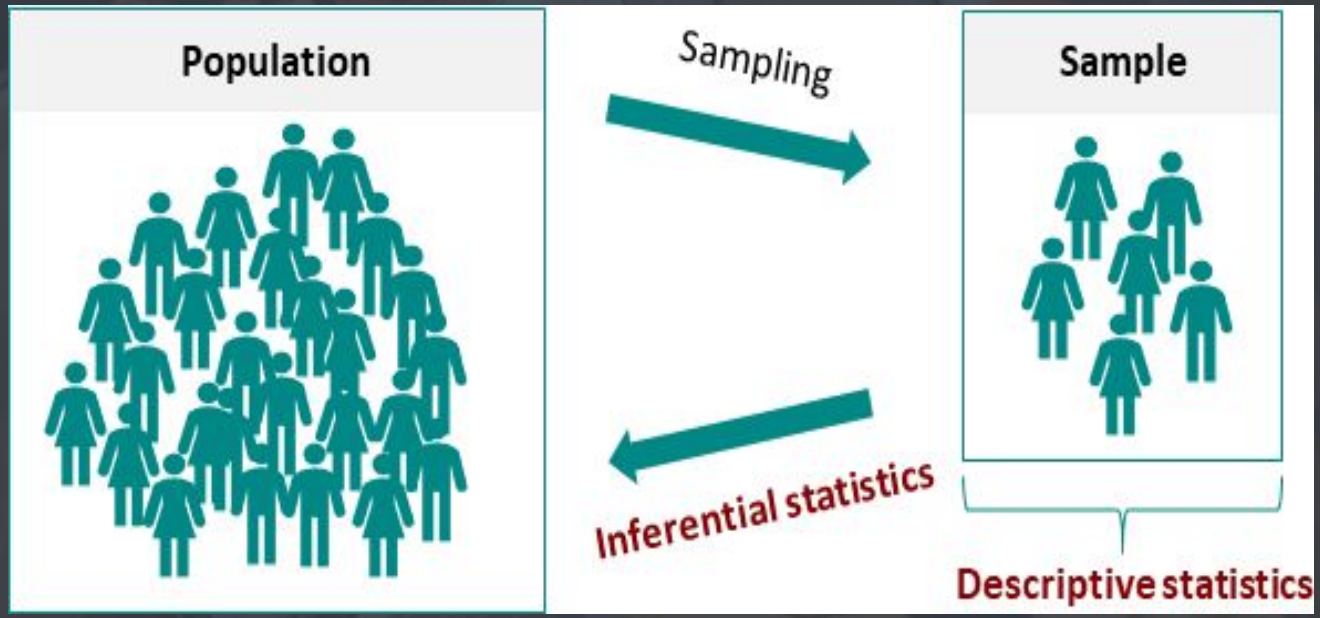
Sample

Population



Sample

Statistics

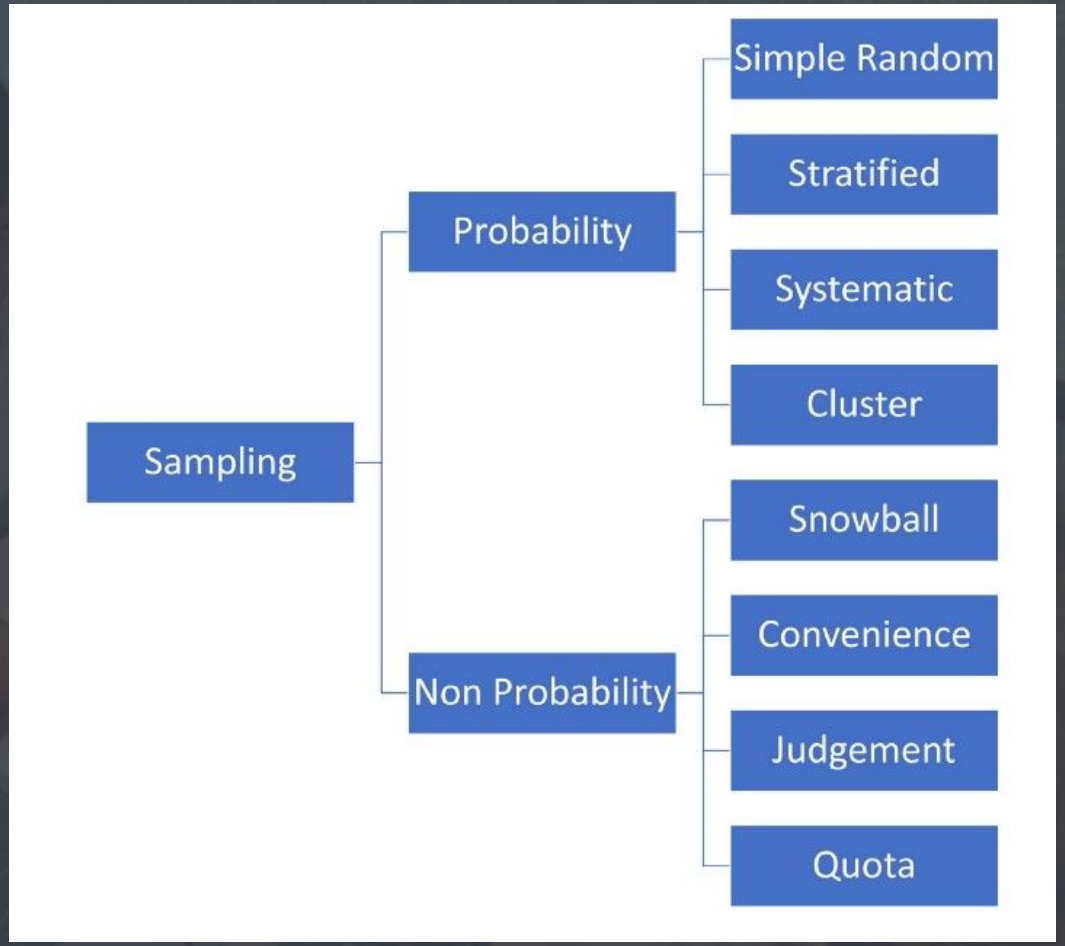

Population
Sampling
Sample
Inferential statistics
Descriptive statistics

Population vs Sample

- N = Total population

- n = Total sample from population

Statistics


Sampling
Probability
Simple Random
Stratified
Systematic
Cluster
Non Probability
Snowball
Convenience
Judgement
Quota

Population vs Sample

Sample A subset from a larger data set.

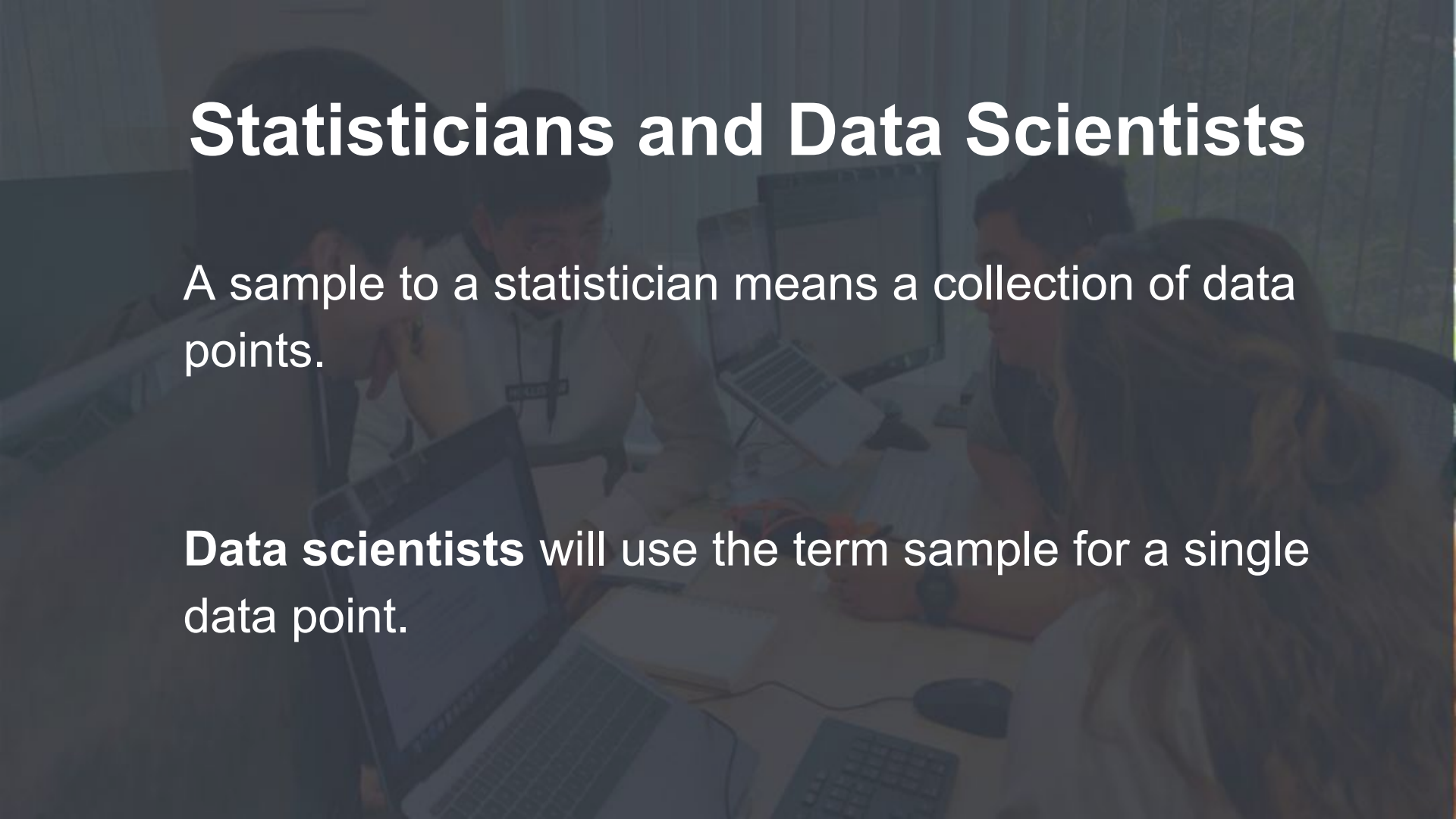
Population The larger data set or idea of a data set.

N (n) The size of the population (sample).

Random sampling Drawing elements into a sample at random.

Simple random sample The sample that results from random sampling without stratifying the population.

Sampling with replacement and without replacement



Statisticians and Data Scientists

A sample to a statistician means a collection of data points.

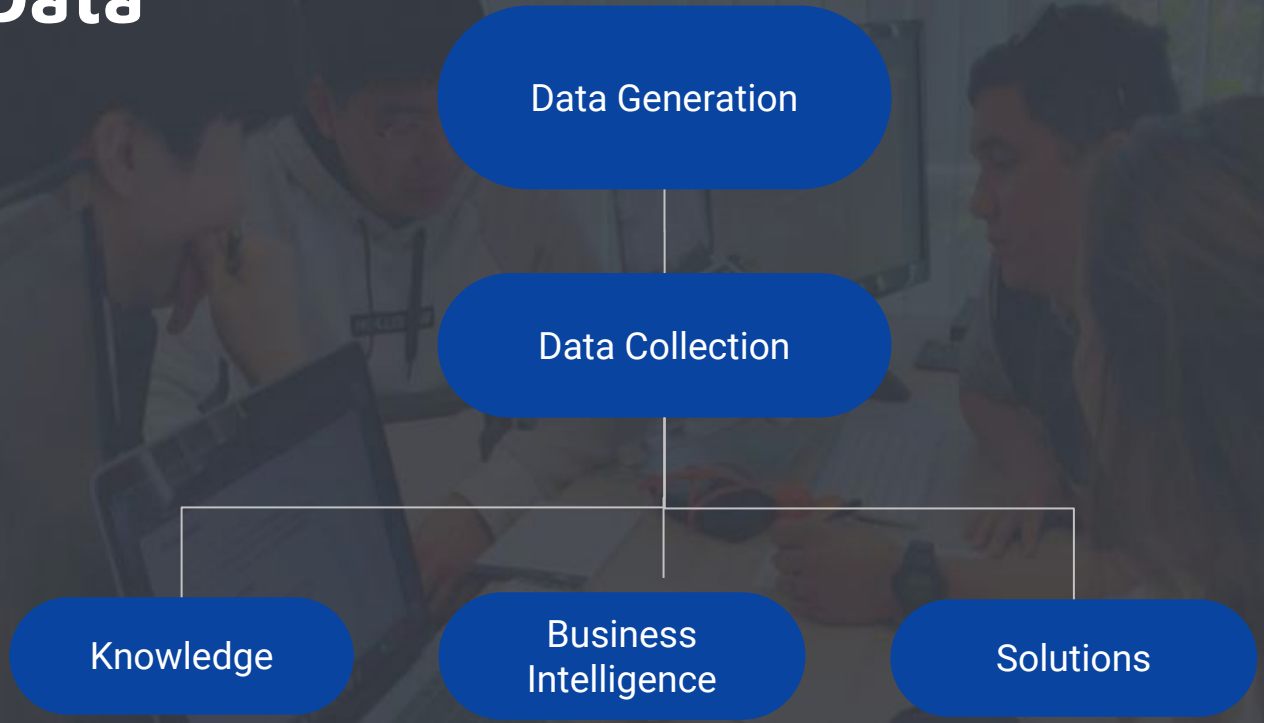
Data scientists will use the term sample for a single data point.

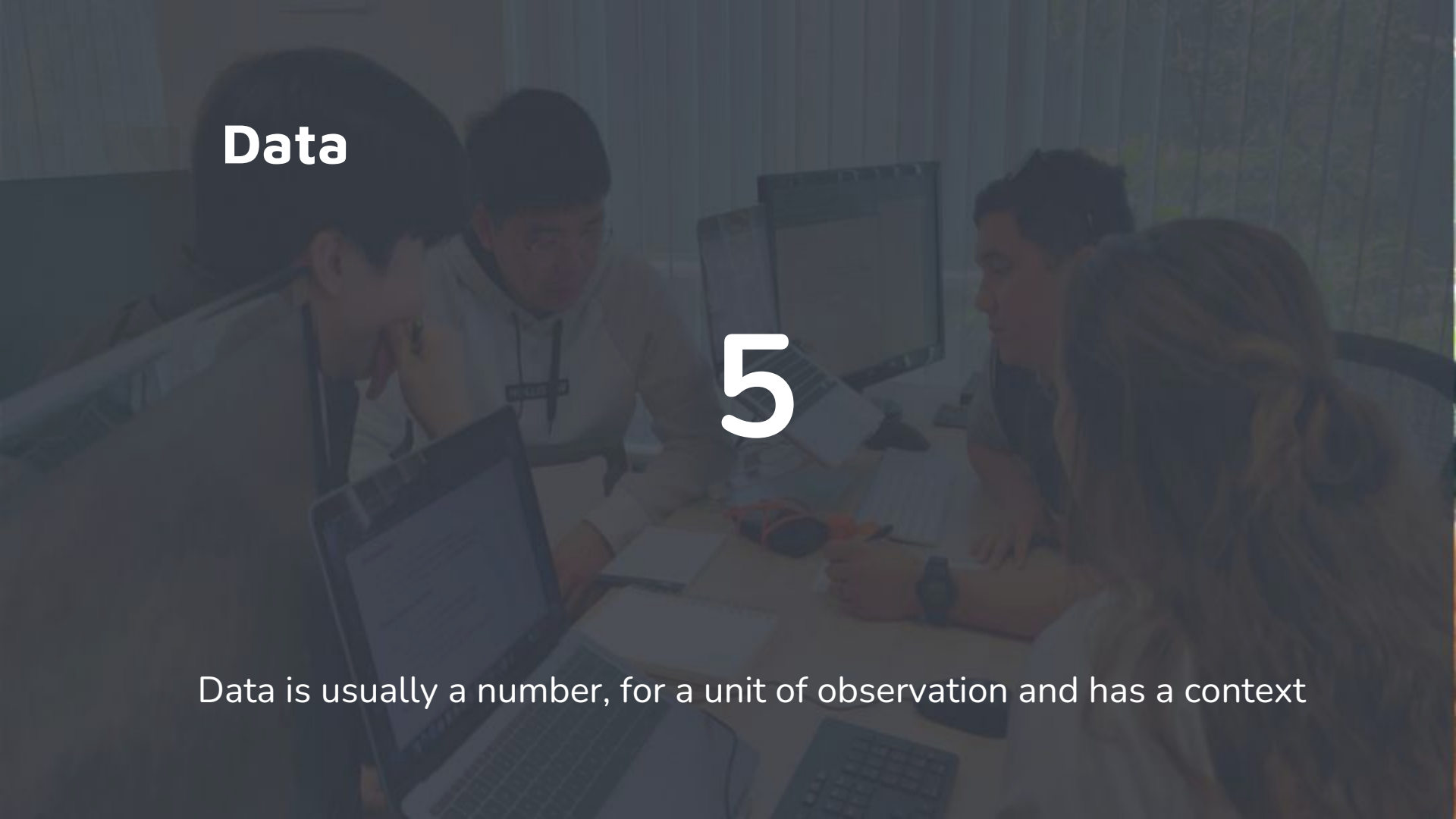
Statisticians and Data Scientists

Statistician, predictor variables are used in a model to predict a response or dependent variable.

For a data scientist, features are used to predict a target.

Data


Data Generation
Data Collection
Knowledge
Business Intelligence
Solutions



Data

5

Data is usually a number, for a unit of observation and has a context

Type of data

Types of variable

- Quantitative variables
- Qualitative variables
- Date and time

Quantitative Variable

Types of Quantitative variable

- Continuous variables
- Discrete variables

Qualitative Variable

Types of Qualitative variable

- Ordinal variables
- Nominal variables

Fundamental Ideas

- Estimates of Location
  - Variance
- Uncertainty
  - Probability

Centrality Measures/ Descriptive statistics

- Central tendency measures, which capture the center around which the data is distributed.

- Variation or variability measures, which describe the data spread, i.e. how far the measurements lie from the center.

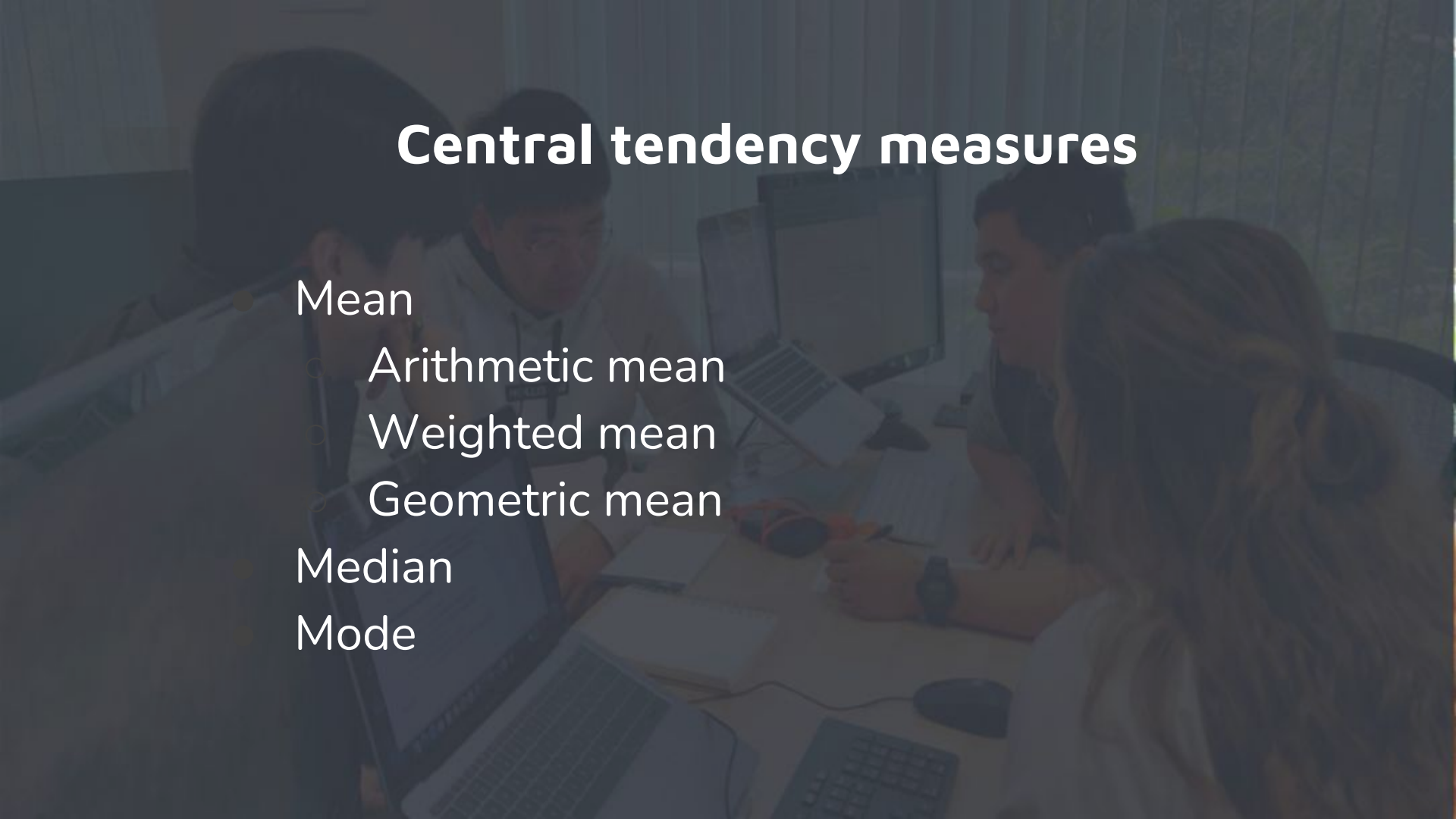
Estimates of Location

Mean: The sum of all values divided by the number of values. Synonyms average

Median: The value such that one-half of the data lies above and below. Synonyms 50th percentile

Mode: The most frequent value

Outlier: A data value that is very different from most of the data. Synonyms extreme value



Central tendency measures

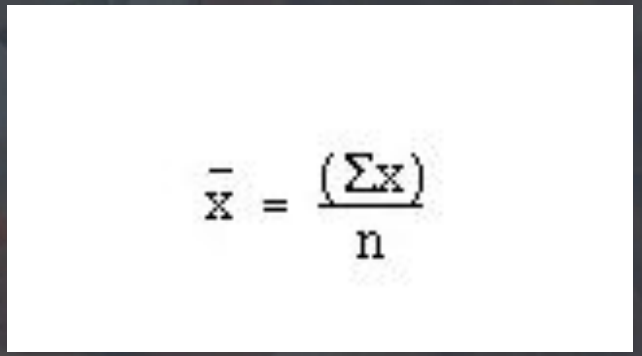
- Mean
  - Arithmetic mean
  - Weighted mean
  - Geometric mean
- Median
- Mode

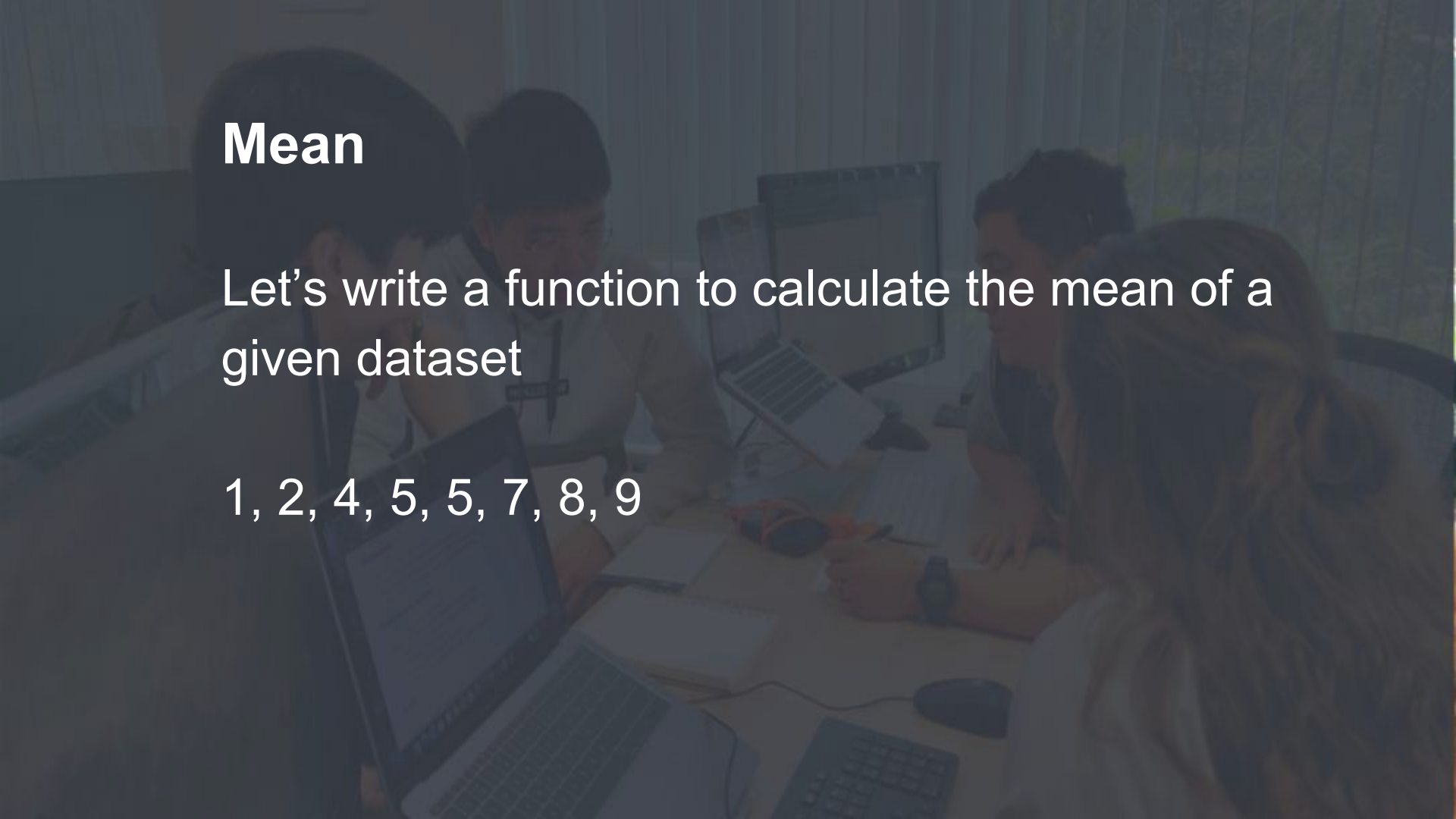
Mean

Average

μ = Mean of population

x = Mean of sample





Mean

Let’s write a function to calculate the mean of a given dataset

1, 2, 4, 5, 5, 7, 8, 9

Weighted Mean

μ = Mean of population

x = Mean of sample

$$= \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Weighted Mean

Let’s write a function to calculator mean of a given dataset

Weighted Values

HW	15%
Quiz	15%
Mid term	35%
Final	35%

Student 1

HW	78
Quiz	70
Mid term	89
Final	94

Students 2

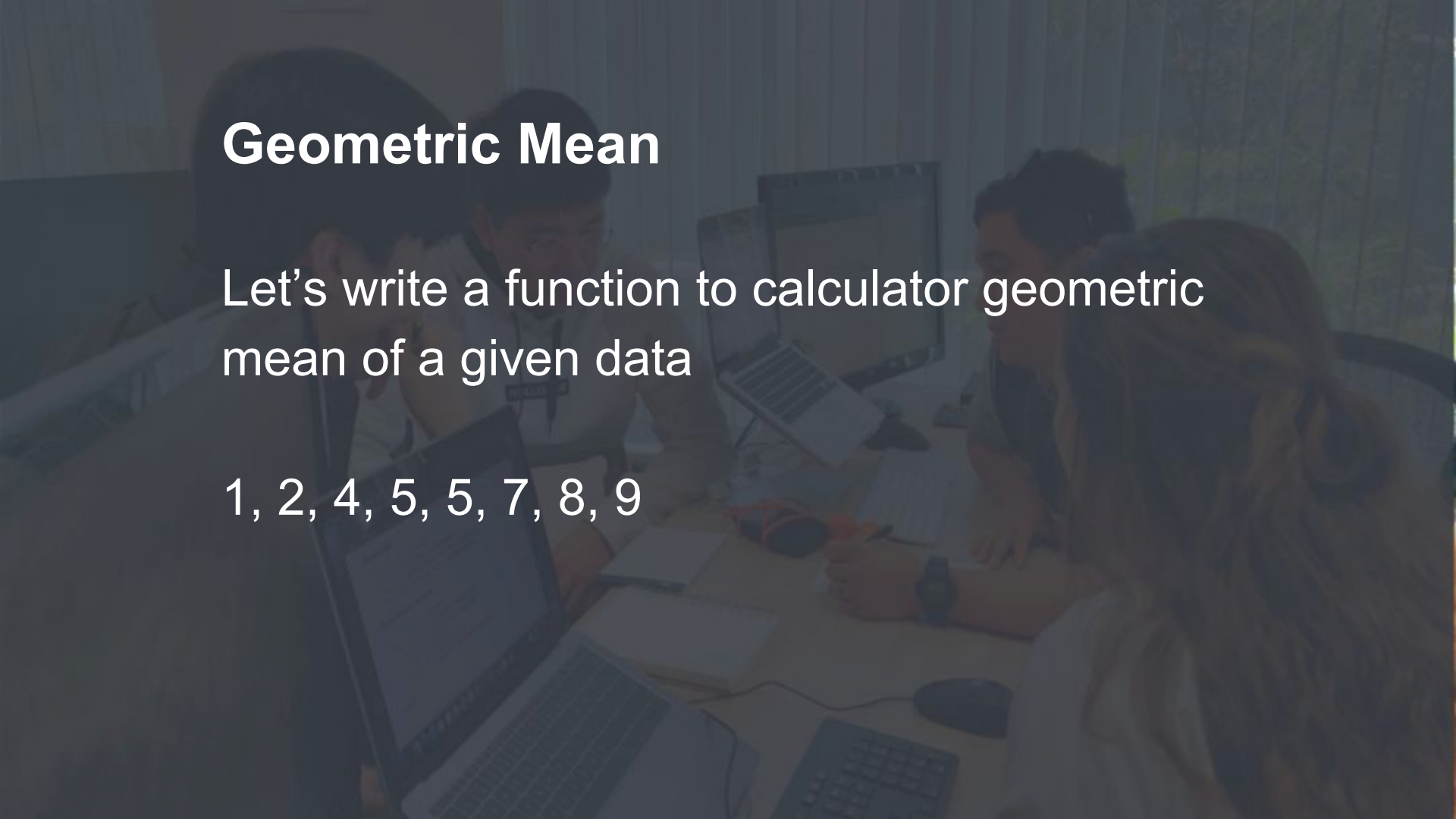
HW	100
Quiz	97
Mid term	76
Final	79

Geometric Mean

μ = Mean of population

$$= \sqrt[n]{x_1 x_2 \cdots x_n}$$

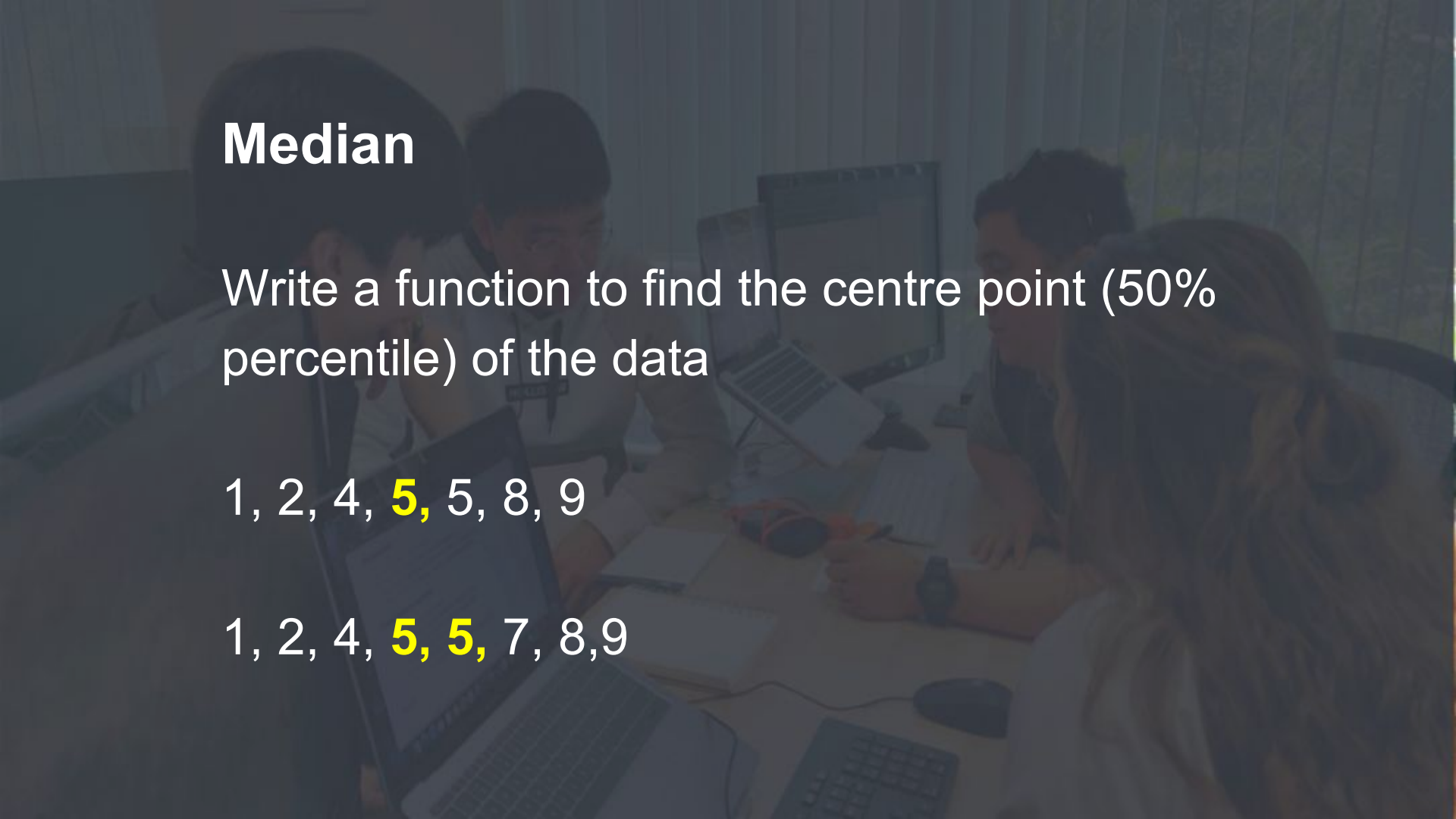
x = Mean of sample



Geometric Mean

Let’s write a function to calculator geometric mean of a given data

1, 2, 4, 5, 5, 7, 8, 9

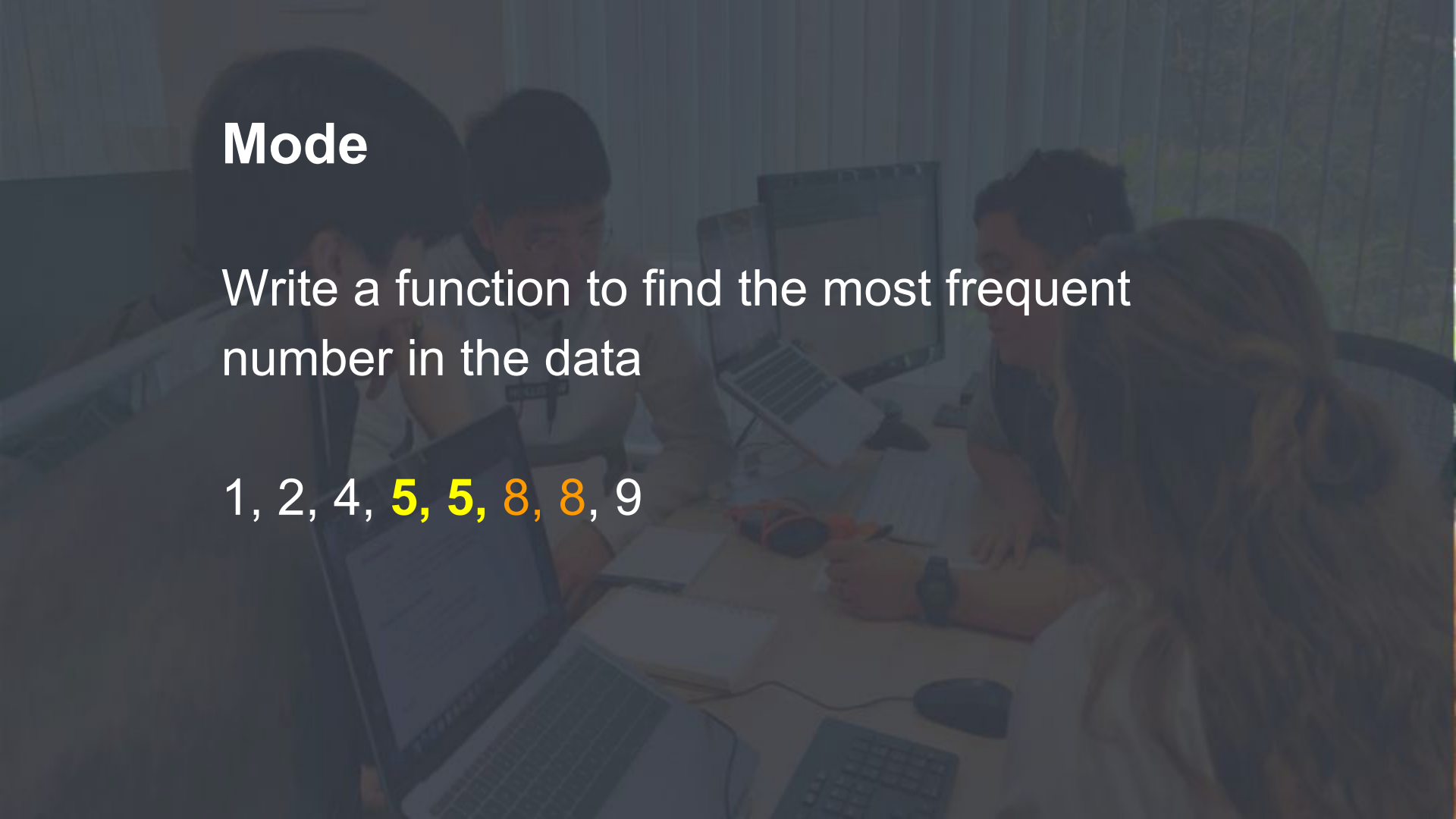


Median

Write a function to find the centre point (50% percentile) of the data

1, 2, 4, **5,** 5, 8, 9

1, 2, 4, **5, 5,** 7, 8,9



Mode

Write a function to find the most frequent number in the data

1, 2, 4, **5, 5,** 8, 8, 9

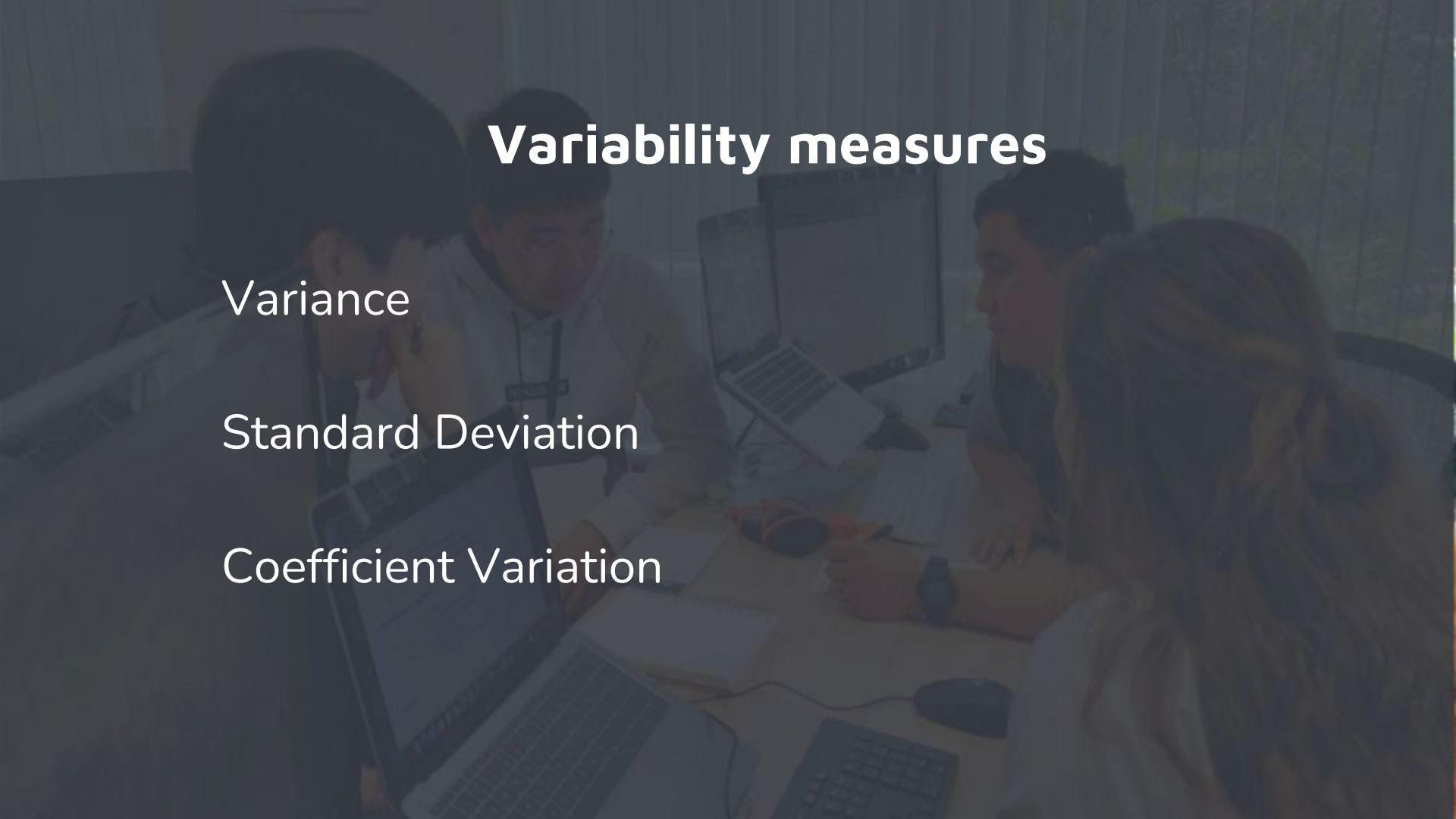
Estimates of Location

Mean: The sum of all values divided by the number of values. Synonyms average

Weighted mean: The sum of all values times a weight divided by the sum of the weights. Synonyms weighted average

Geometric mean is calculated by multiplying all numbers and then taking the nth root of that number (where n=number of numbers).

Trimmed mean: The average of all values after dropping a fixed number of extreme values. Synonyms truncated mean



Variability measures

Variance

Standard Deviation

Coefficient Variation

Variance

Spread of data around the mean

σ^2 = Variance of population

$$\text{Variance} = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

S^2 = Variance of sample

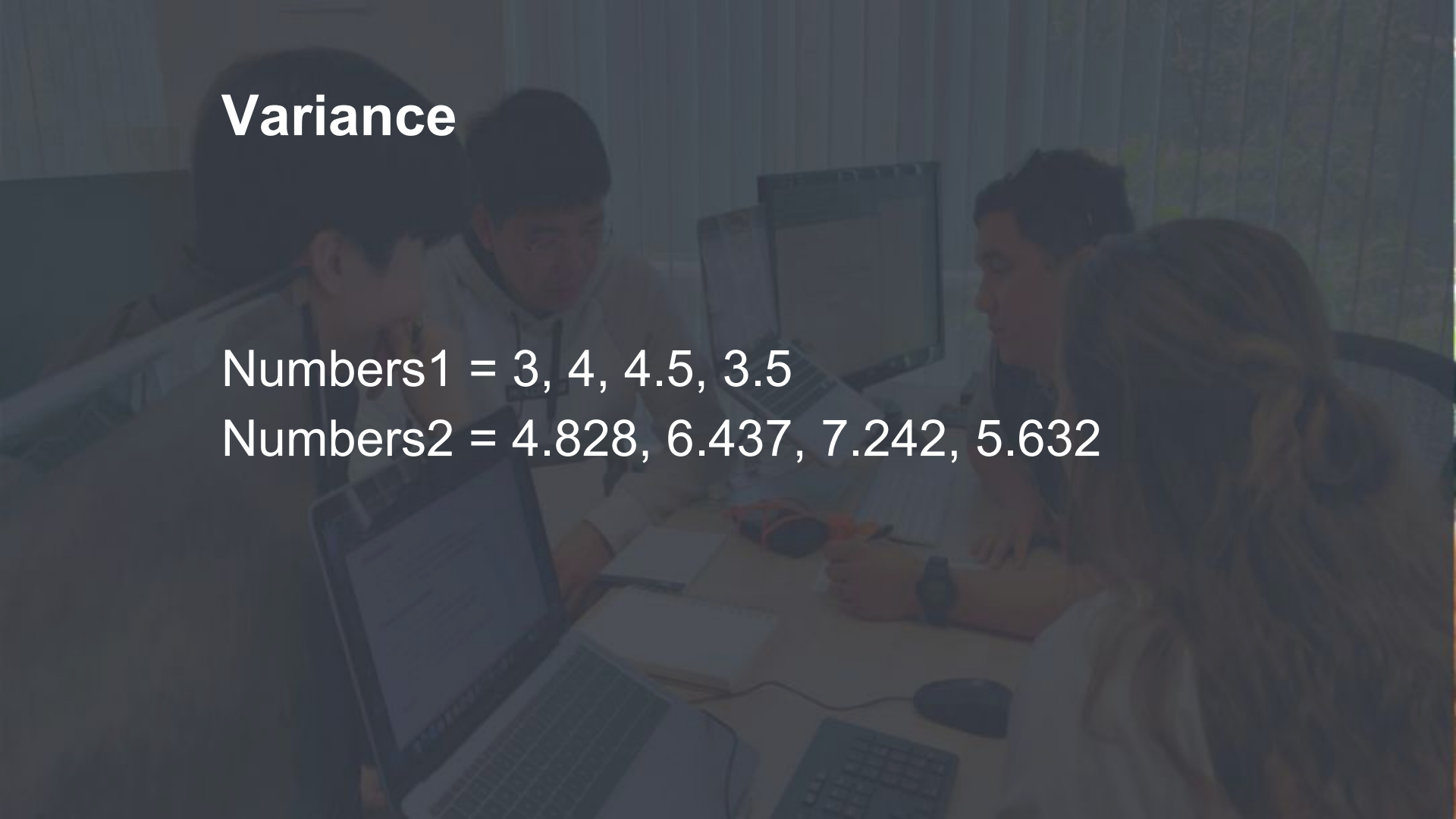
Standard deviation

Square root of variance

$$SD = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

σ = Standard deviation of population

s = Standard deviation of sample



Variance

Numbers1 = 3, 4, 4.5, 3.5

Numbers2 = 4.828, 6.437, 7.242, 5.632

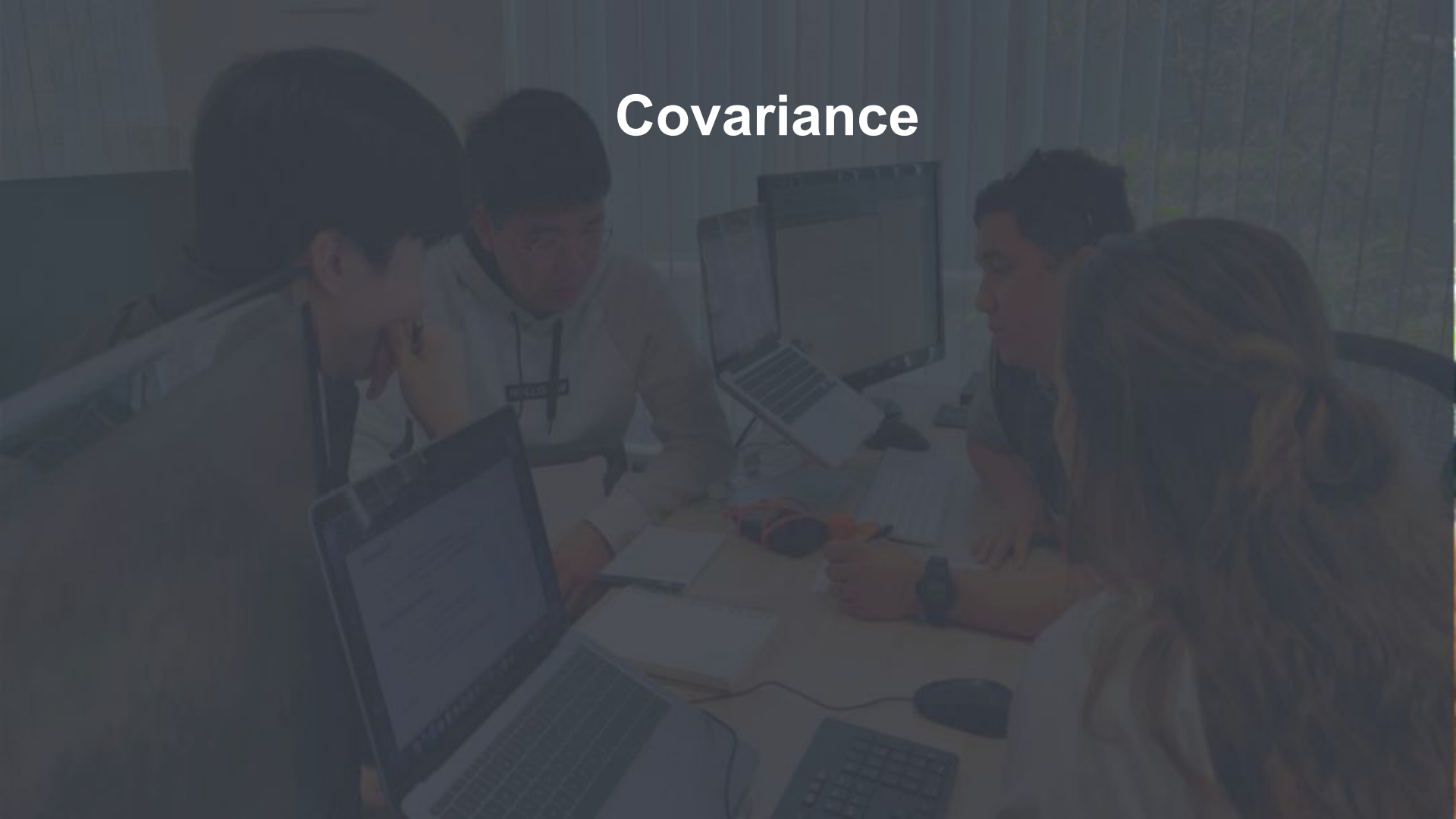
Coefficient of variation

Compare two datasets which operate on different scales i.e miles and kilometers

Coefficient of variation = Standard deviation / Mean

Miles = 3, 4, 4.5, 3.5
Kms = 4.828, 6.437, 7.242, 5.632

Covariance



Covariance

Covariance if two values are moving in the same direction

Age	Internet use
18	125
25	120
40	100
50	70

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable a and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values