

Module II: Mathematical Foundations on Exploratory Data Analysis

On averages, variations, correlations and the rest...

This module covers...

- Statistical Moments 1 - 4
 - 1st: mean / average / median
 - 2nd: standard deviation / variance
 - 3rd: skewness
 - 4rd: kurtosis
- Covariance, covariance matrices and correlation
- Multidimensional vector spaces

In this Video you will learn...

Statistical Moments

The 1st moment

- known as “average” or “measure of central tendency”
- but there are more ways of doing it
 - mean
 - median
 - mode

mean

- easy to calculate

$$AM = \frac{1}{n} \sum_{i=1}^n a_i = \frac{1}{n} (a_1 + a_2 + \cdots + a_n)$$

- “sum up all values and divide by the number of values”

median

- more complex to calculate
- “sort the list and pick the middle element”

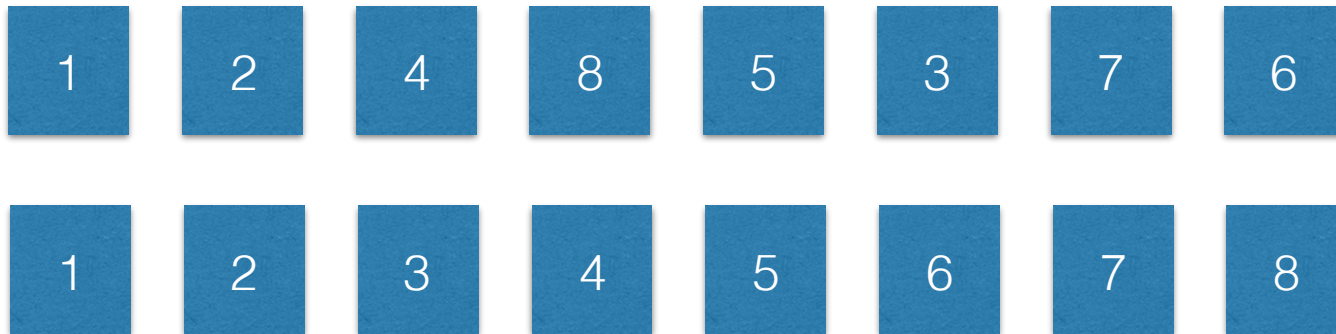
median

- more complex to calculate
- “sort the list and pick the middle element”



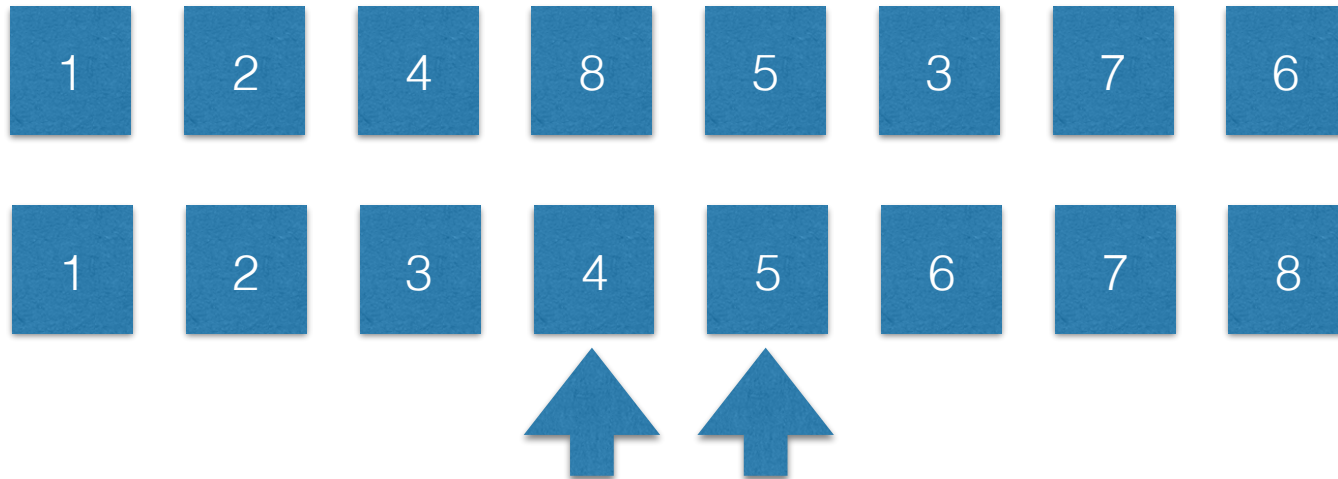
median

- more complex to calculate
- “sort the list and pick the middle element”



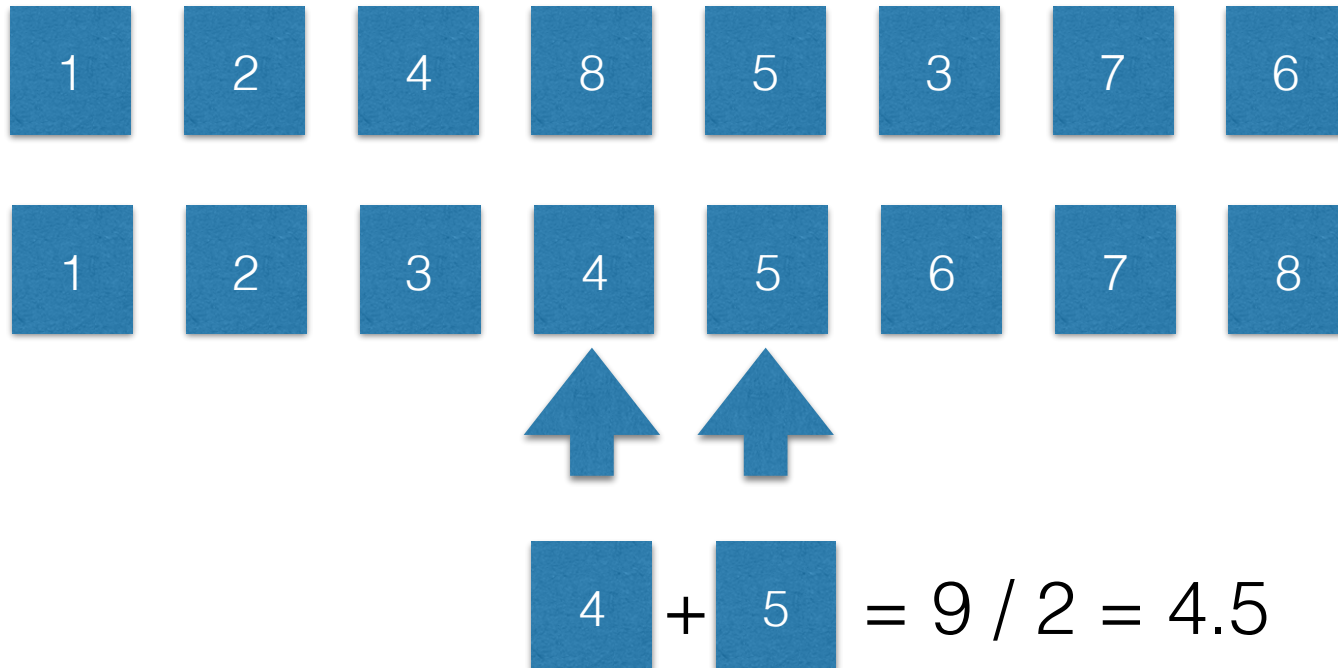
median

- more complex to calculate
- “sort the list and pick the middle element”



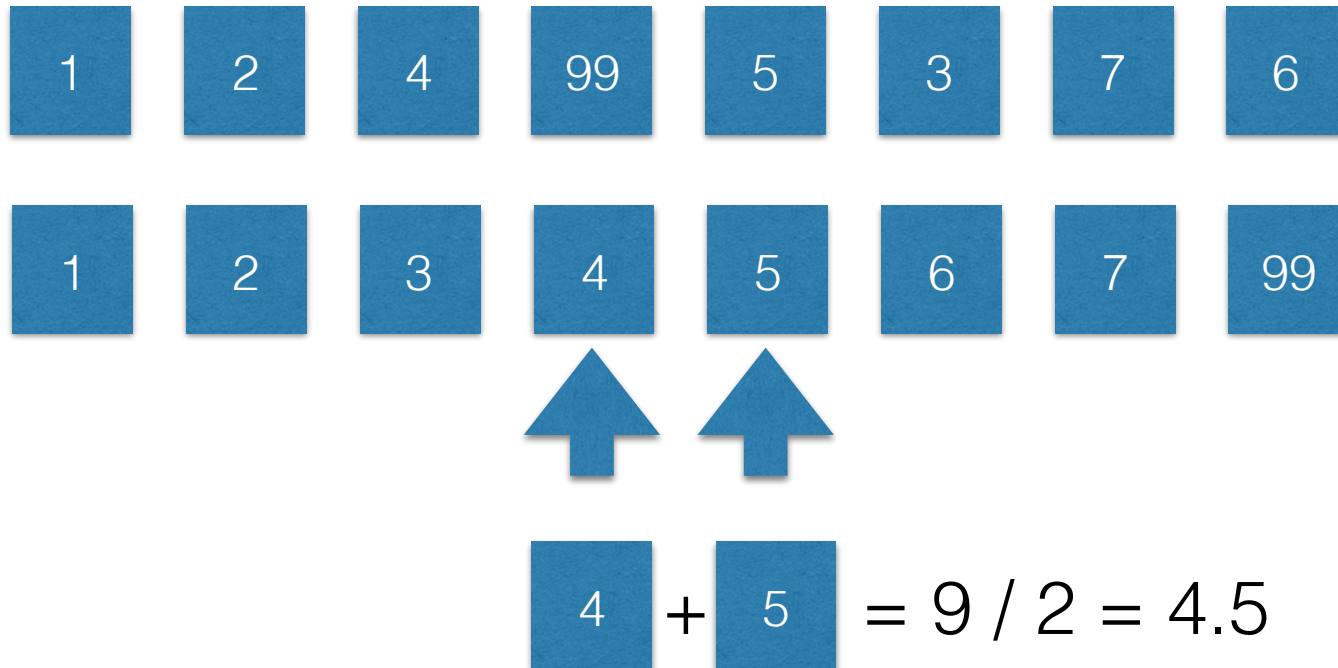
median

- more complex to calculate
- “sort the list and pick the middle element”



median

- more complex to calculate
- “sort the list and pick the middle element”



median


- more complex to calculate
- “sort the list and pick the middle element”

mean()

median

- more complex to calculate
- “sort the list and pick the middle element”



mean()=127

The text shows the word "mean(" followed by a row of eight blue squares containing the numbers 1, 2, 3, 4, 5, 6, 7, and 99 in order, followed by a closing parenthesis and "=127".

median

- more complex to calculate
- “sort the list and pick the middle element”



mean(

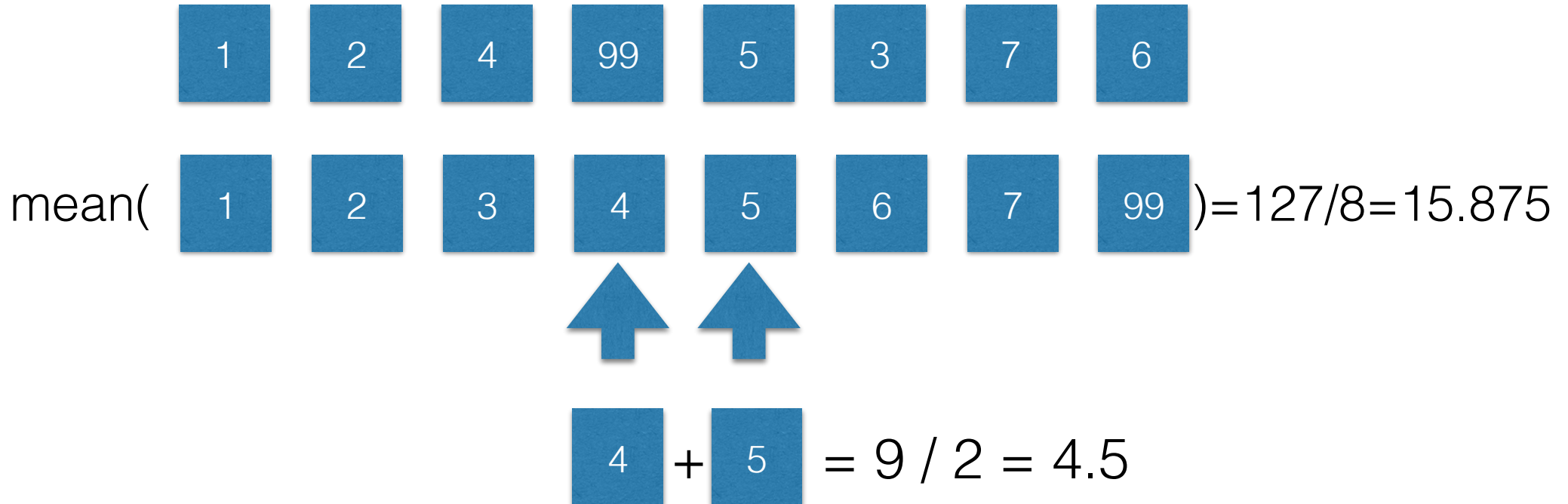
A horizontal row of eight blue squares, each containing a white number. The numbers from left to right are: 1, 2, 3, 4, 5, 6, 7, and 99.

1	2	3	4	5	6	7	99
---	---	---	---	---	---	---	----

)=127/8

median

- more complex to calculate
- “sort the list and pick the middle element”



median

- more complex to calculate
- “sort the list and pick the middle element”
- median is out-lier resistant version of mean

Summary

- mean and median are the 1st moment of a statistical distribution
- give insights on central tendency of a sensor value
- ApacheSpark parallelises it's computation using the lambda calculus
- The same program can be run on a couple of bytes as well as on multiple tera bytes of data

The next video covers...

Standard Deviation