Data Science and Machine Learning

Antony Ross

Environment Set-up

Anaconda 3

Create course folder

```
$ jupyter notebook
```

Data Science Python Libraries

Numpy

Pandas

Scikit-Learn

Matplotlib

Seaborn

Data Science Python Libraries

Numpy

Pandas

Scikit-Learn

Matplotlib

Seaborn

The Data Science Process

1.) Identify a useful question

2.) Acquire the data

3.) Clean the data

4.) Explore the data

5.) Model the data

6.) Communicate the results

The Data Science Process

1.) Identify a useful question

2.) Acquire the data

3.) Clean the data

4.) Explore the data

5.) Model the data

6.) Communicate the results

Identify a Useful Question

Acquire the Data

Datasets

- kaggle.com/datasets
- https://registry.opendata.aws
- https://cloud.google.com/bigquery/public-data/
- data.gov
- archive.ics.uci.edu/ml/
- https://github.com/fivethirtyeight/data
- https://www.quandl.com/search
- public APIs (e.g., Twitter, Facebook, Spotify)
- web scraping
- your company

Datasets

Google Dataset Search

toolbox.google.com/datasetsearch

ProPublica Data Store

propublica.org/datastore

NASA’s Open Data Portal

data.nasa.gov

World Bank Open Data

data.worldbank.org

Descriptive Statistics Review

Descriptive Measures

Central Tendency

Variation

Relative Standing

Central Tendency

Mean

Median

Mode

Mean

[3, 5, 5, 1, 7, 2, 6, 7, 0, 4]

Feature 1
3
5
5
1
7
2
6
7
0
4

Sum = 40

40/10 = **4**

Median

	Feature 1
	0
	1
	2
	3
	4
	5
	5
	6
	7
	7
Sum =	40

Put the numbers in order

Half of measures are above

= 4.5

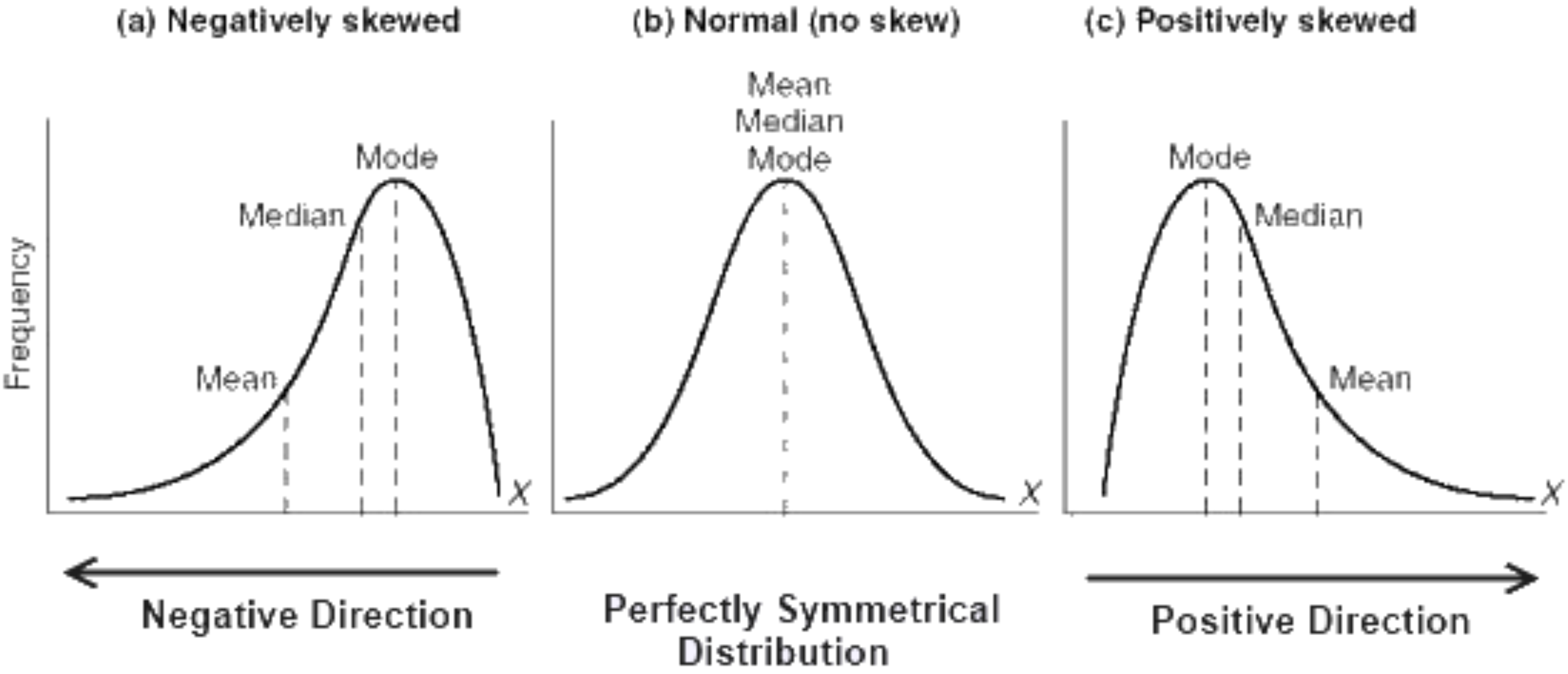
Half of measures are above

Mode

	Feature 1
	0
	1
	2
	3
	4
	5
	5
	6
	7
	7
Sum =	40

Number(s) which appears most often

= 5 and 7



Negative Direction

Perfectly Symmetrical Distribution

Positive Direction

Variation

Variance

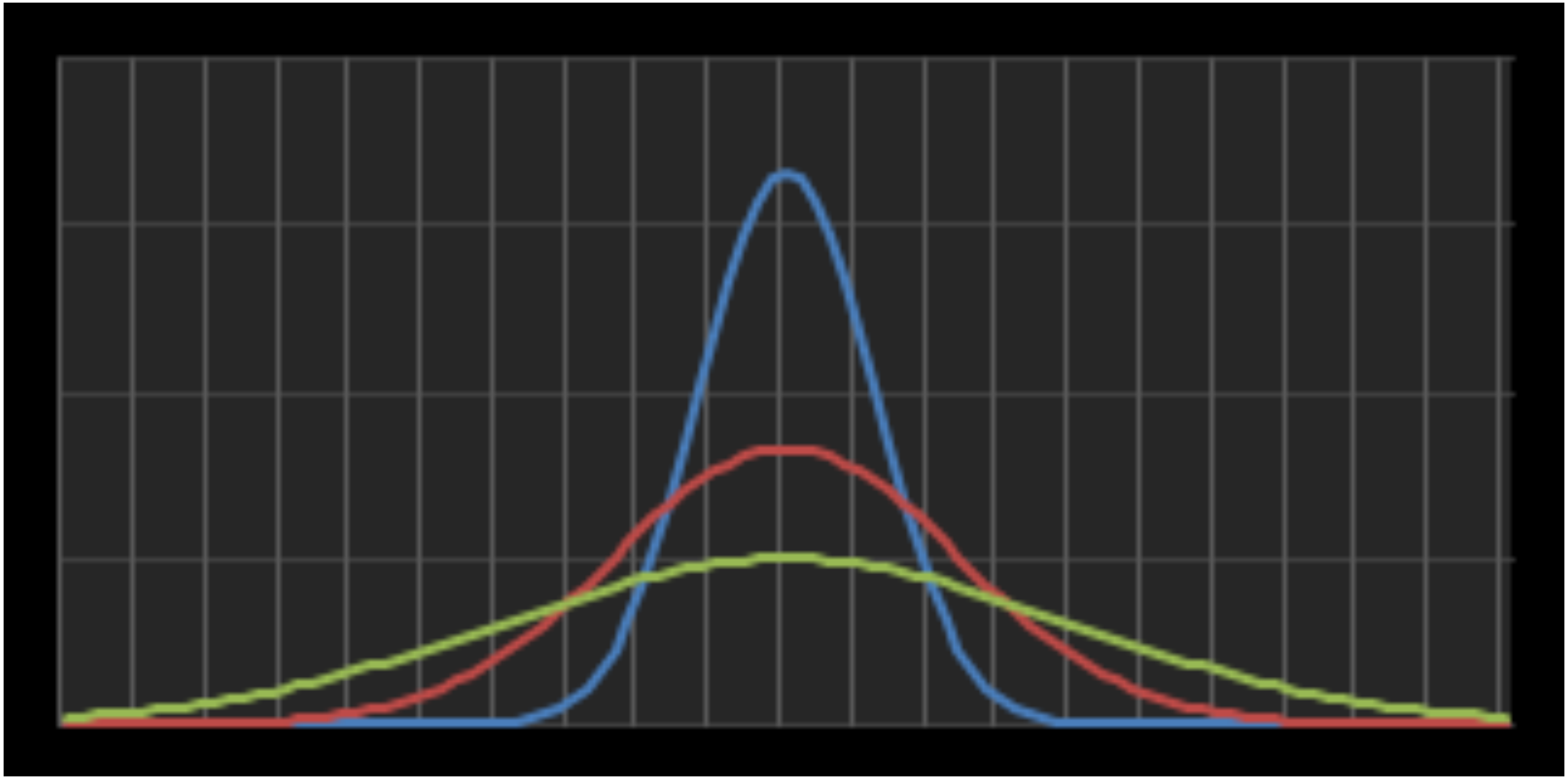
Standard Deviation

Range

Quartiles

Interquartile Range

The spread of the data



Variance

	Feature 1	Deviations	Squared Deviations
	0	-4	16
	1	-3	9
	2	-2	4
	3	-1	1
	4	0	0
	5	1	1
	5	1	1
	6	2	4
	7	3	9
	7	3	9
Mean = 4	40	0	54

54/9 = **6**

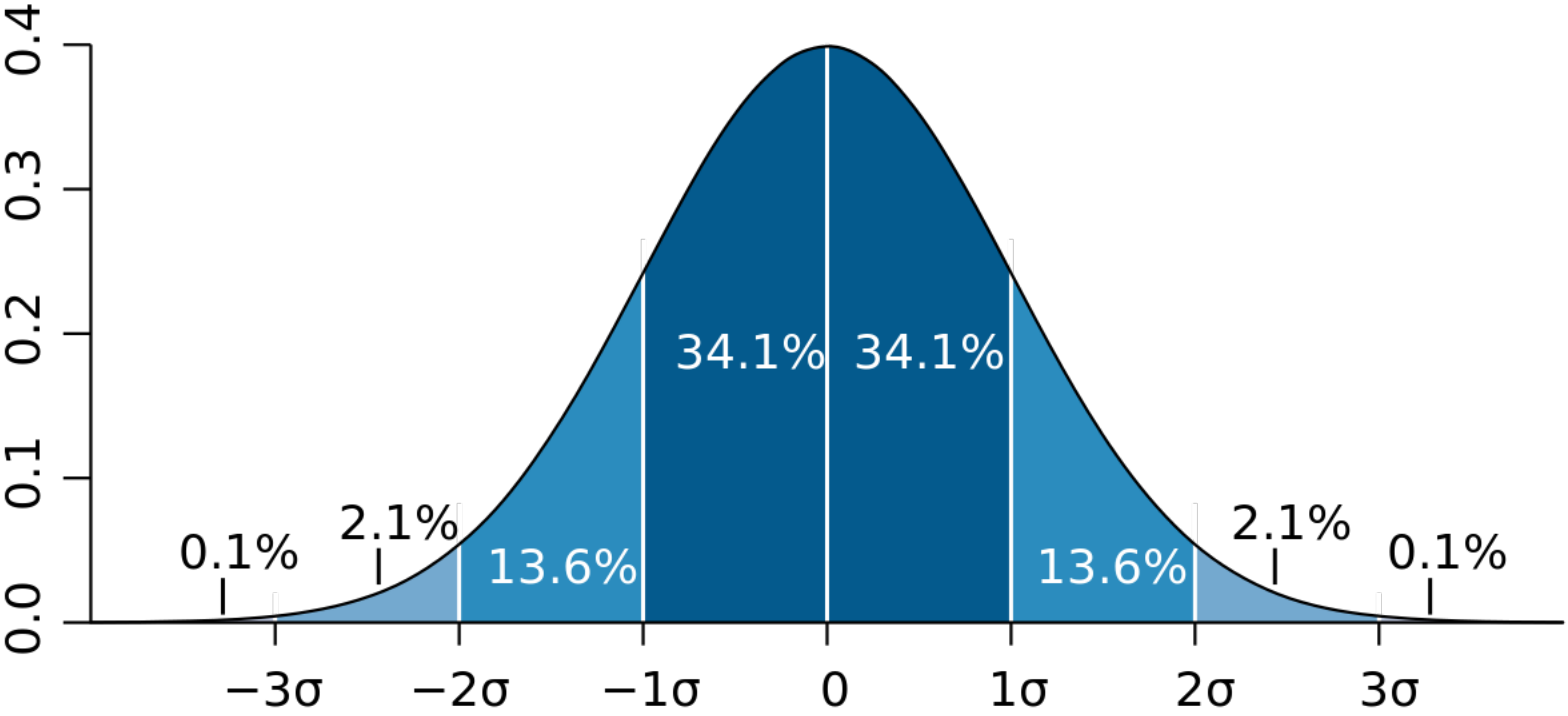
Standard Deviation

	Feature 1	Deviations	Squared Deviations
	0	-4	16
	1	-3	9
	2	-2	4
	3	-1	1
	4	0	0
	5	1	1
	5	1	1
	6	2	4
	7	3	9
	7	3	9
Mean = 4	40	0	54

$$\sqrt{6} = \mathbf{2.45}$$

Standard Score
(standardization)

z-score = (x - mean)/std


0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Range

Feature 1	Deviations	Squared Deviations
0	-4	16
1	-3	9
2	-2	4
3	-1	1
4	0	0
5	1	1
5	1	1
6	2	4
7	3	9
7	3	9
40	0	54

Max value = 7

Min value = 0

7-0 = **7**

Percentiles

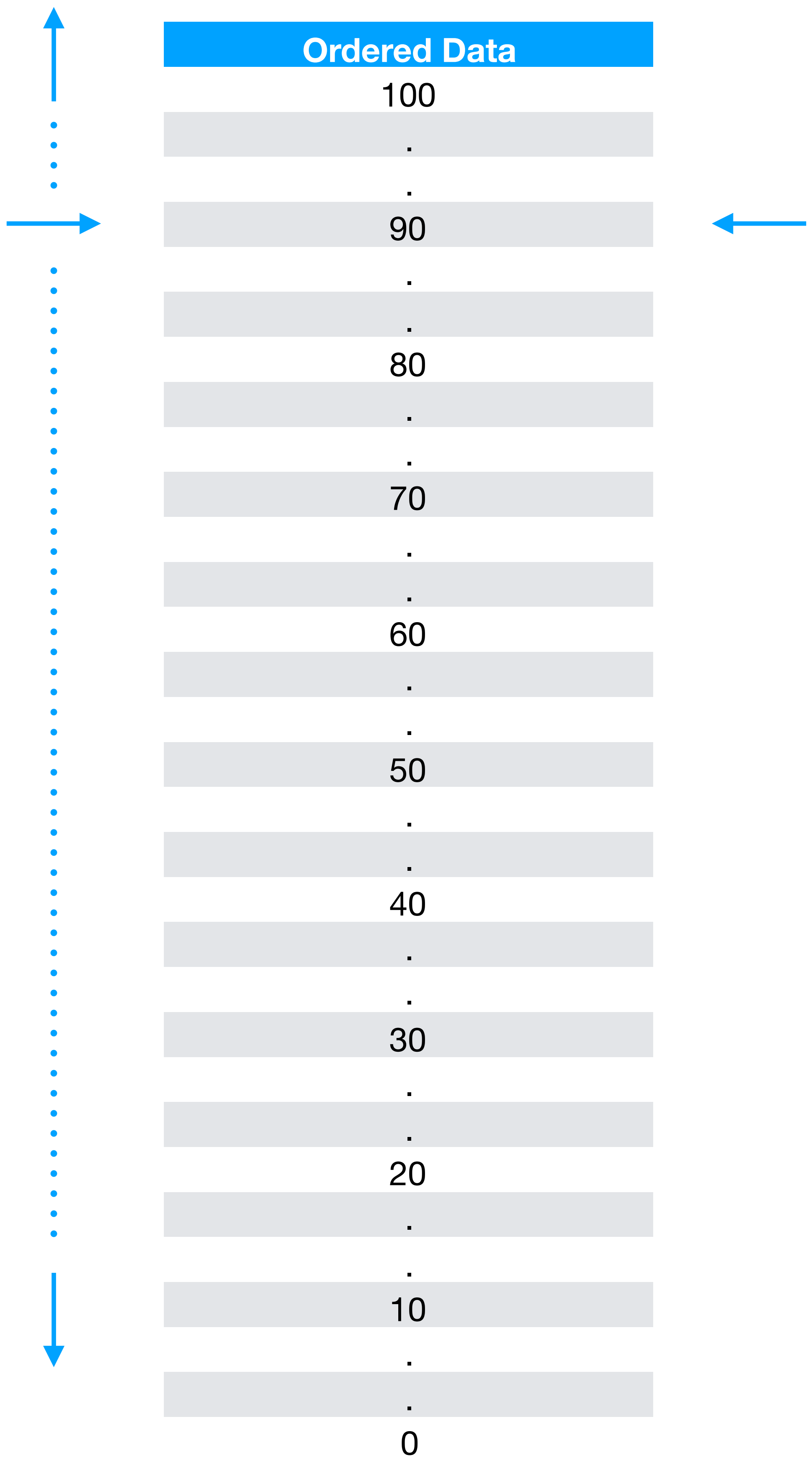
Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0

Percentiles

Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0

← 90th percentile

Percentiles



10% of values above

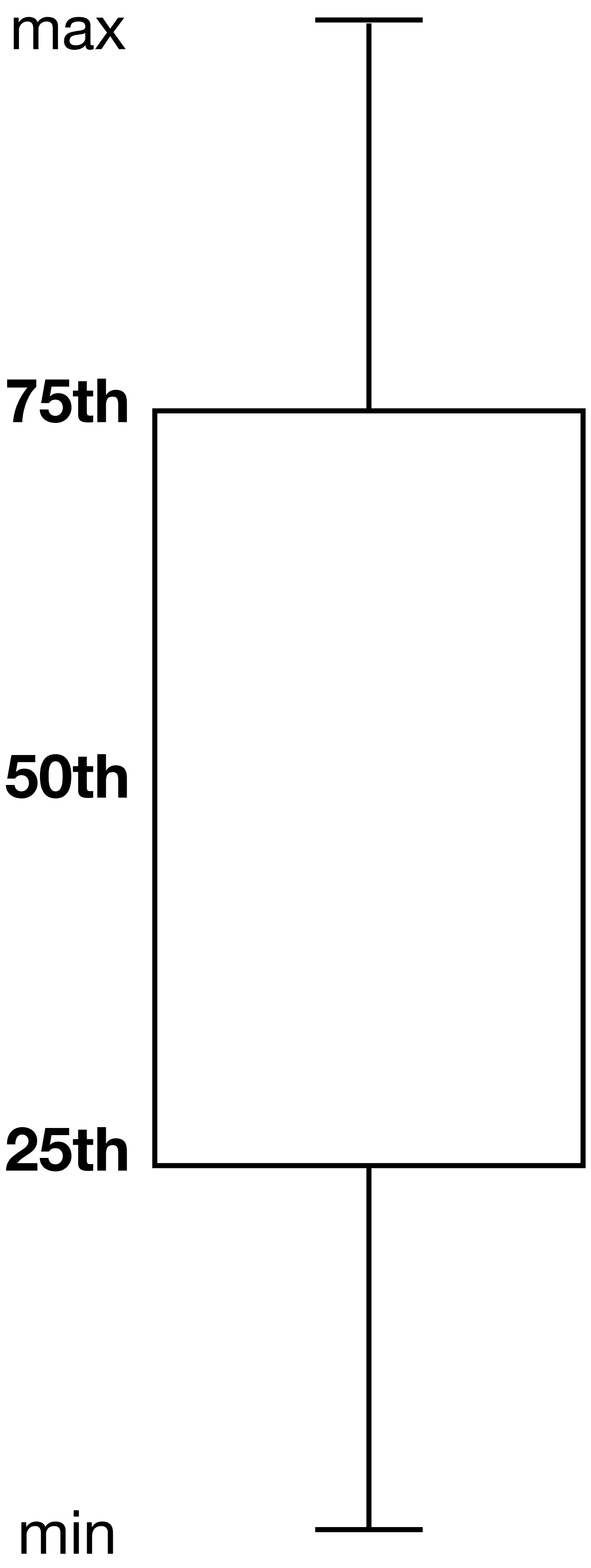
Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0

90th percentile

90% of values below

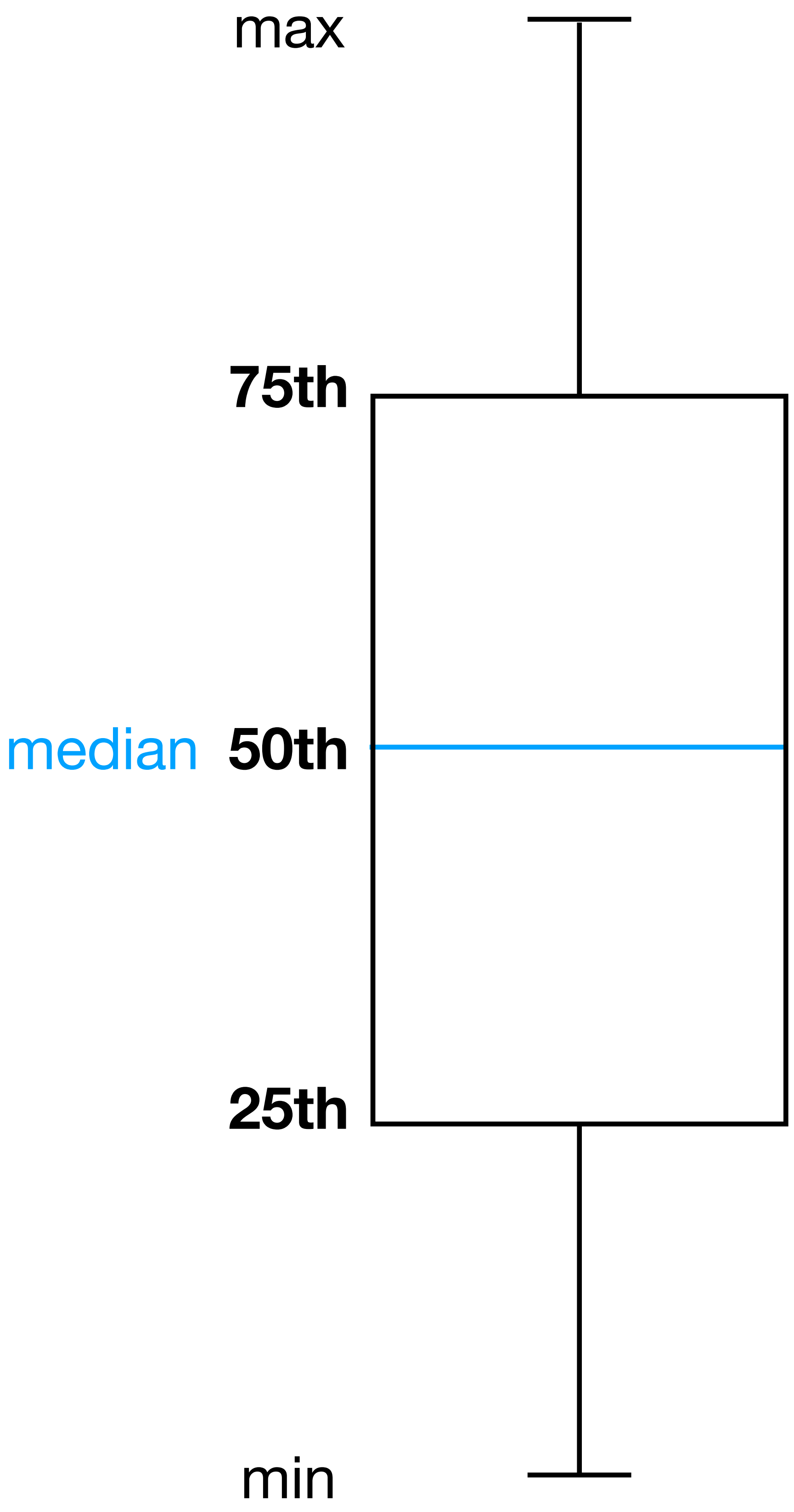
Quartiles

Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0



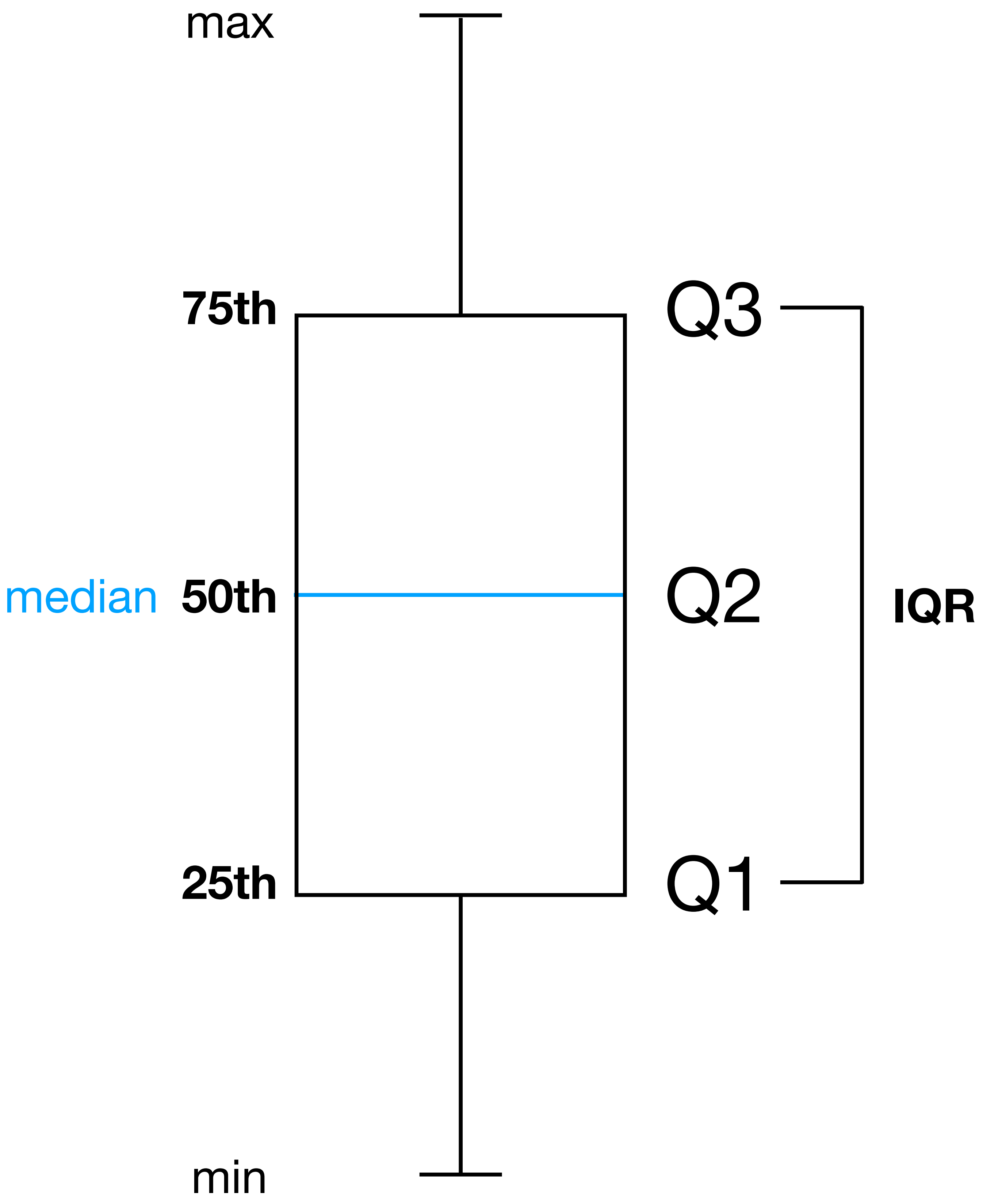
Quartiles

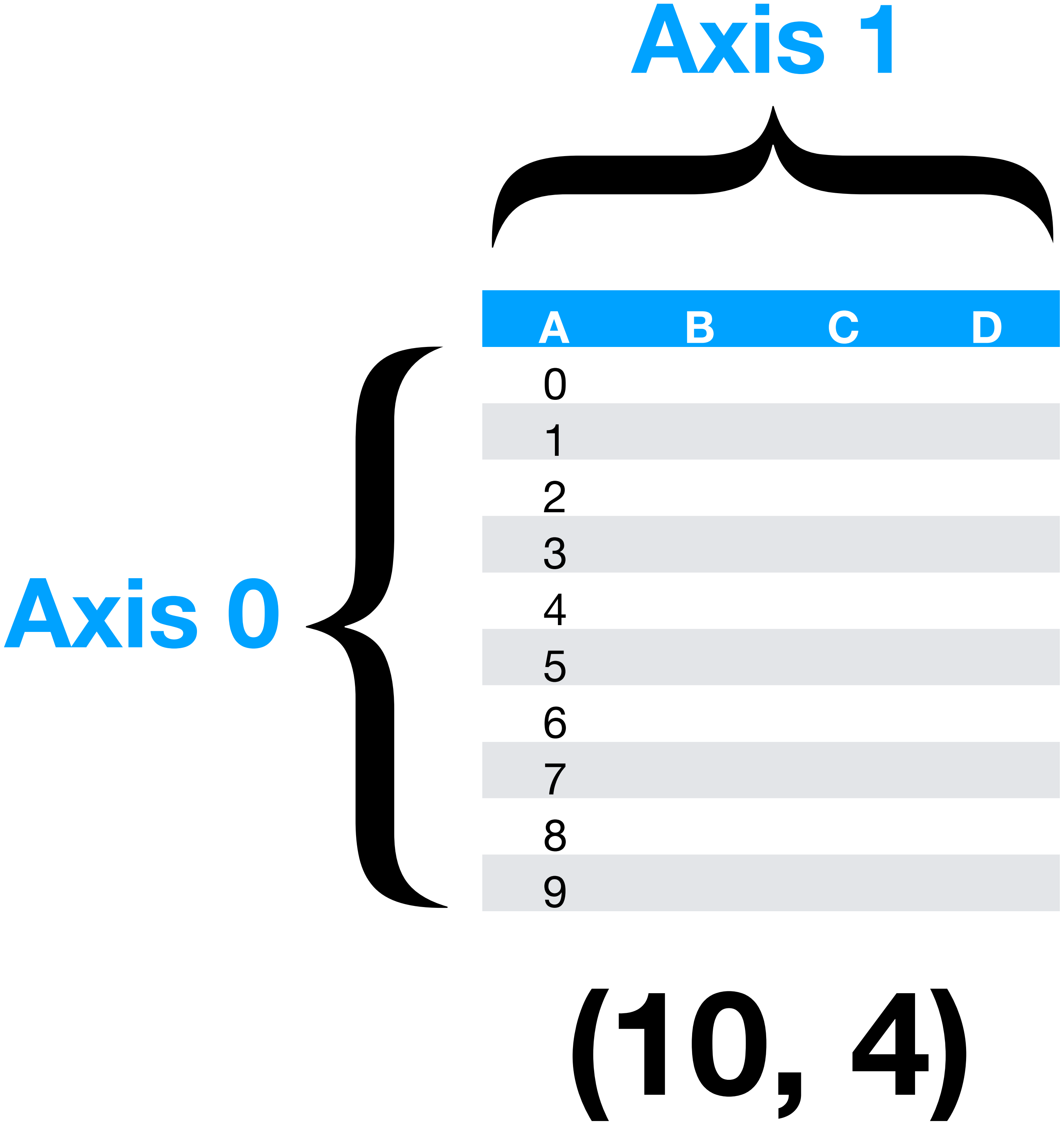
Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0



Quartiles

Ordered Data
100
.
.
90
.
.
80
.
.
70
.
.
60
.
.
50
.
.
40
.
.
30
.
.
20
.
.
10
.
.
0





Axis 1

Axis 0

	A	B	C	D
0				
1				
2				
3				
4				
5				
6				
7				
8				
9				

(10, 4)