Exploring Data

1.2 Describing Distributions with Numbers

YMS3e

AP Stats at LSHS

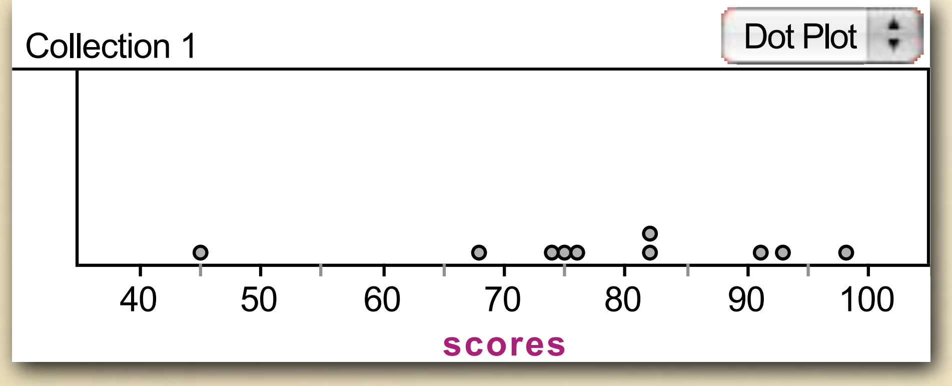
Mr. Molesky

Sample Data

- Consider the following test scores for a small class:

75	76	82	93	45	68	74	82	91	98

Plot the data and describe the SOCS:



Shape?

Outliers?

Center?

Spread?

What number best describes the "center"?

What number best describes the "spread'?

Measures of Center

- **Numerical descriptions of distributions begin with a measure of its "center".**
  - **If you could summarize the data with one number, what would it be?**

Mean: $\bar{x}$ **The "average" value of a dataset.**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \bar{x} = \frac{\sum x_i}{n}$$

Median: **Q2 or M** **The "middle" value of a dataset.**

Arrange observations in order min to max
Locate the middle observation, average if needed.

Mean vs. Median

- The *mean* and the *median* are the most common measures of center.
  - If a distribution is perfectly symmetric, the *mean* and the *median* are the same.
  - The *mean* is **not resistant to outliers**.
- *You* must decide which number is the most appropriate description of the center...

MeanMedian Applet

Measures of Spread

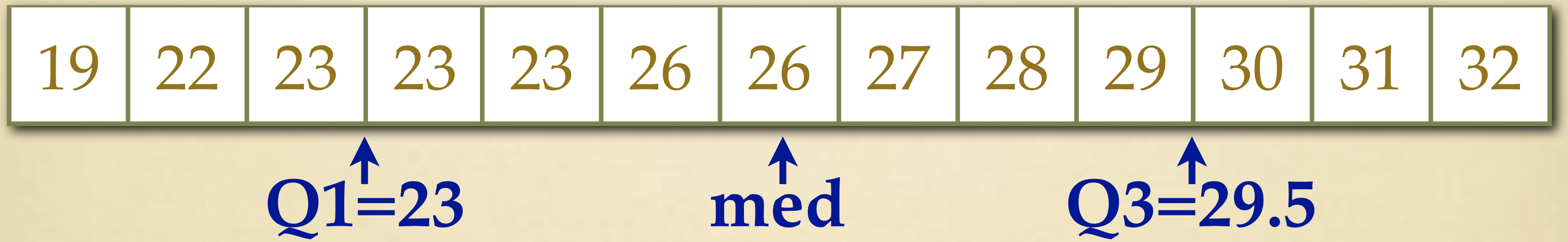
- **Variability is the key to Statistics. Without variability, there would be no need for the subject.**
  - **When describing data, *never* rely on center alone.**

- **Measures of Spread:**
  - **Range - {*rarely* used...why?}**
  - **Quartiles - InterQuartile Range {IQR=Q3-Q1}**
  - **Variance and Standard Deviation {var and s_x}**

- **Like Measures of Center, *you* must choose the most appropriate measure of spread.**

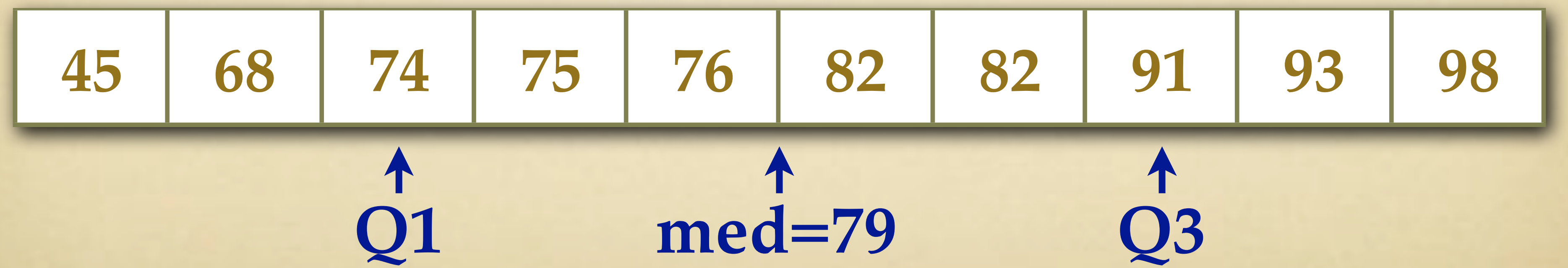
Quartiles

- **Quartiles Q1 and Q3** represent the 25th and 75th percentiles.
  - ☑ To find them, order data from min to max.
  - ☑ Determine the **median** - average if necessary.
  - ☑ The **first quartile** is the middle of the 'bottom half'.
  - ☑ The **third quartile** is the middle of the 'top half'.



19	22	23	23	23	26	26	27	28	29	30	31	32

Q1=23 **med** **Q3=29.5**



45	**68**	**74**	**75**	**76**	**82**	**82**	**91**	**93**	**98**

Q1 **med=79** **Q3**

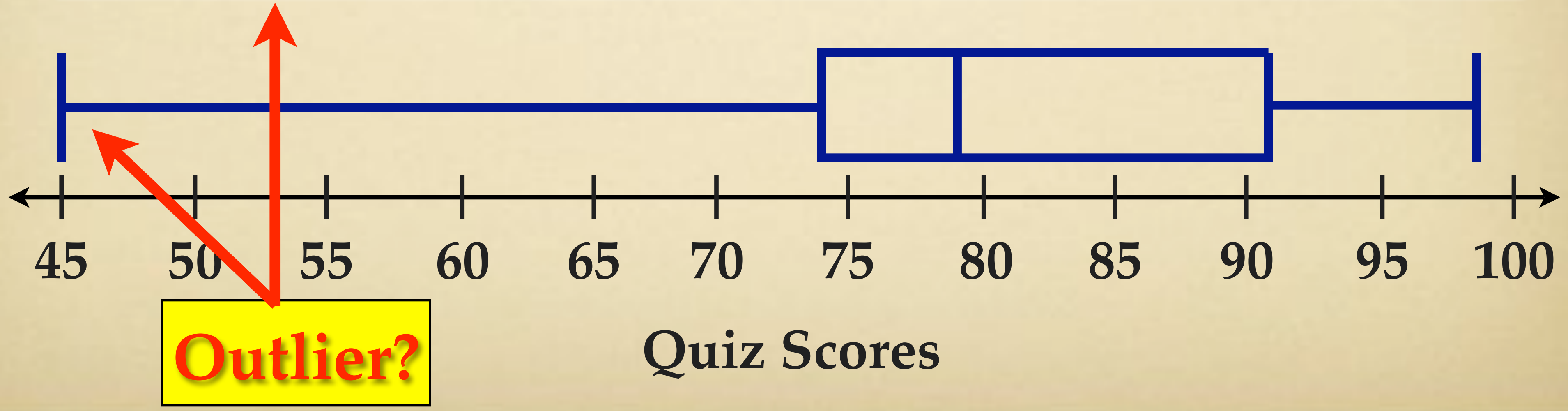
5-Number Summary, Boxplots

- The **5 Number Summary** provides a reasonably complete description of the center and spread of distribution

MIN	Q1	MED	Q3	MAX

- We can visualize the 5 Number Summary with a **boxplot**.

min=45	Q1=74	med=79	Q3=91	max=98



Outlier?

Determining Outliers “1.5 • IQR Rule”

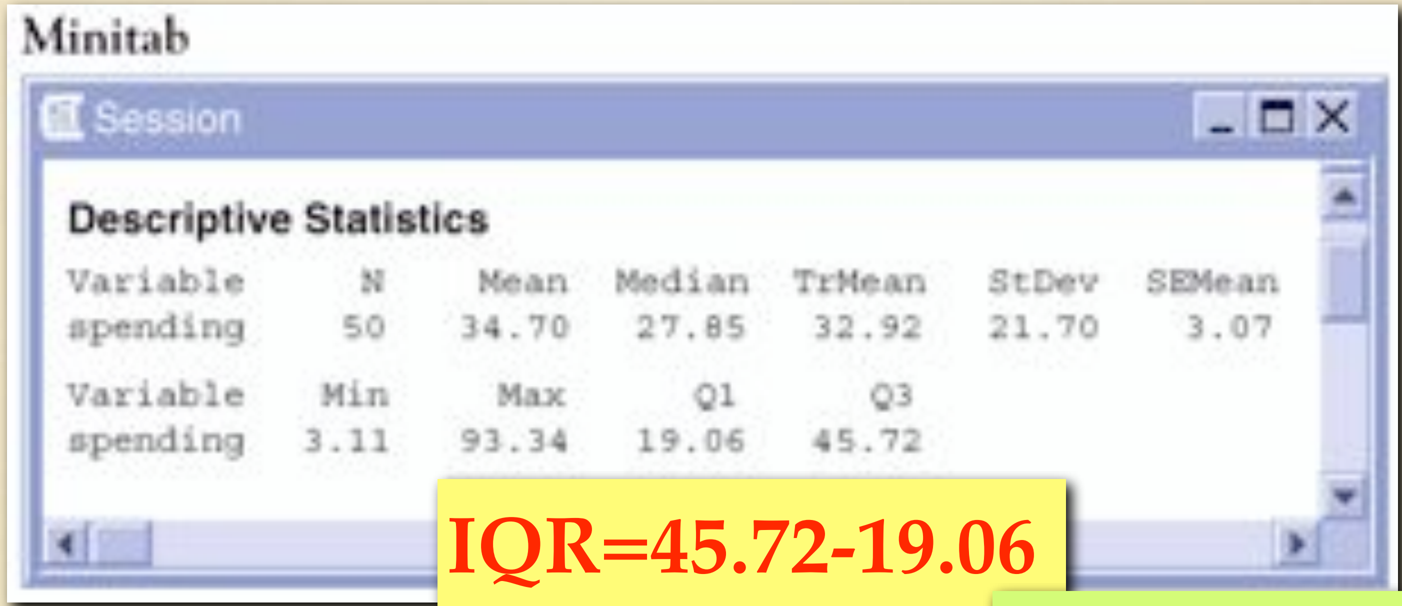
- **InterQuartile Range “IQR”:** Distance between Q1 and Q3. Resistant measure of spread...only measures middle 50% of data.
  - **IQR = Q3 - Q1** **{width of the “box” in a boxplot}**
- **1.5 IQR Rule:** If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an **outlier**.

Why 1.5? According to John Tukey, 1 IQR seemed like too little and 2 IQRs seemed like too much...

1.5 • IQR Rule

- To determine outliers:
  - ☑ Find 5 Number Summary
  - ☑ Determine IQR
  - ☑ Multiply 1.5xIQR
  - ☑ Set up “fences” Q1-(1.5IQR) and Q3+(1.5IQR)
  - ☑ Observations “outside” the fences are outliers.

Outlier Example



```
Descriptive Statistics

Variable        N     Mean   Median   TrMean    StDev   SEMean
spending       50    34.70    27.85    32.92    21.70     3.07

Variable      Min      Max       Q1       Q3
spending     3.11    93.34    19.06    45.72
```

All data on p. 48.

IQR=45.72-19.06

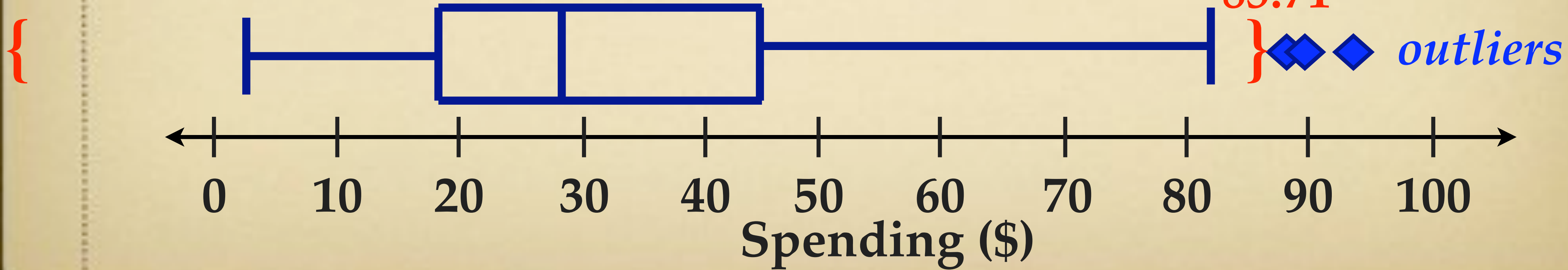
IQR=26.66

1.5IQR=1.5(26.66)

1.5IQR=39.99

fence: 19.06-39.99
= -20.93

fence: 45.72+39.99
= 85.71



Standard Deviation

- Another common measure of spread is the **Standard Deviation**: a measure of the "*average*" deviation of all observations from the mean.

- To calculate **Standard Deviation**:

  ☑ Calculate the **mean**.

  ☑ Determine each observation's **deviation (x - xbar).**

  ☑ "Average" the ***squared*-deviations** by dividing the total *squared* deviation by **(n-1)**.

  ☑ This quantity is the **Variance**.

  ☑ Square root the result to determine the **Standard Deviation.**

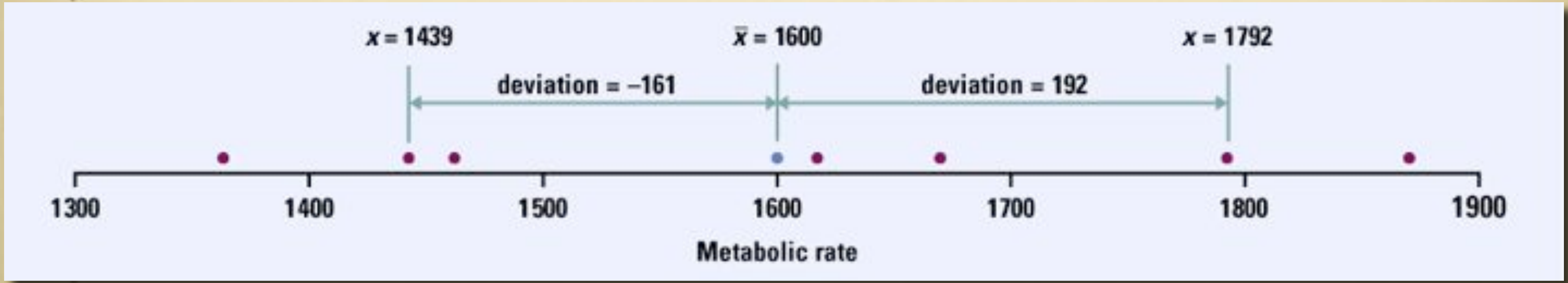
Standard Deviation

- **Variance:** $$\text{var} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

- **Standard Deviation:** $$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

  - **Example 1.16 (p.85): Metabolic Rates**

1792	1666	1362	1614	1460	1867	1439



Standard Deviation

1792	1666	1362	1614	1460	1867	1439

Metabolic Rates: mean=1600

x	$(x - \bar{x})$	$(x - \bar{x})^2$
1792	192	36864
1666	66	4356
1362	-238	56644
1614	14	196
1460	-140	19600
1867	267	71289
1439	-161	25921
Totals:	0	214870

Total Squared Deviation	214870

Variance	var=214870/6
	var=35811.66

Standard Deviation	$s=\sqrt{35811.66}$ s=189.24 cal

What does this value, s, mean?

Linear Transformations

- Variables can be measured in different units (feet vs meters, pounds vs kilograms, etc)
- When converting units, the measures of center and spread will change.
- **Linear Transformations (x_{new}=a+bx)** do not change the shape of a distribution.
    - ☑ Multiplying each observation by *b* multiplies both the measure of center and spread by *b*.
    - ☑ Adding *a* to each observation adds *a* to the measure of center, but does not affect spread.

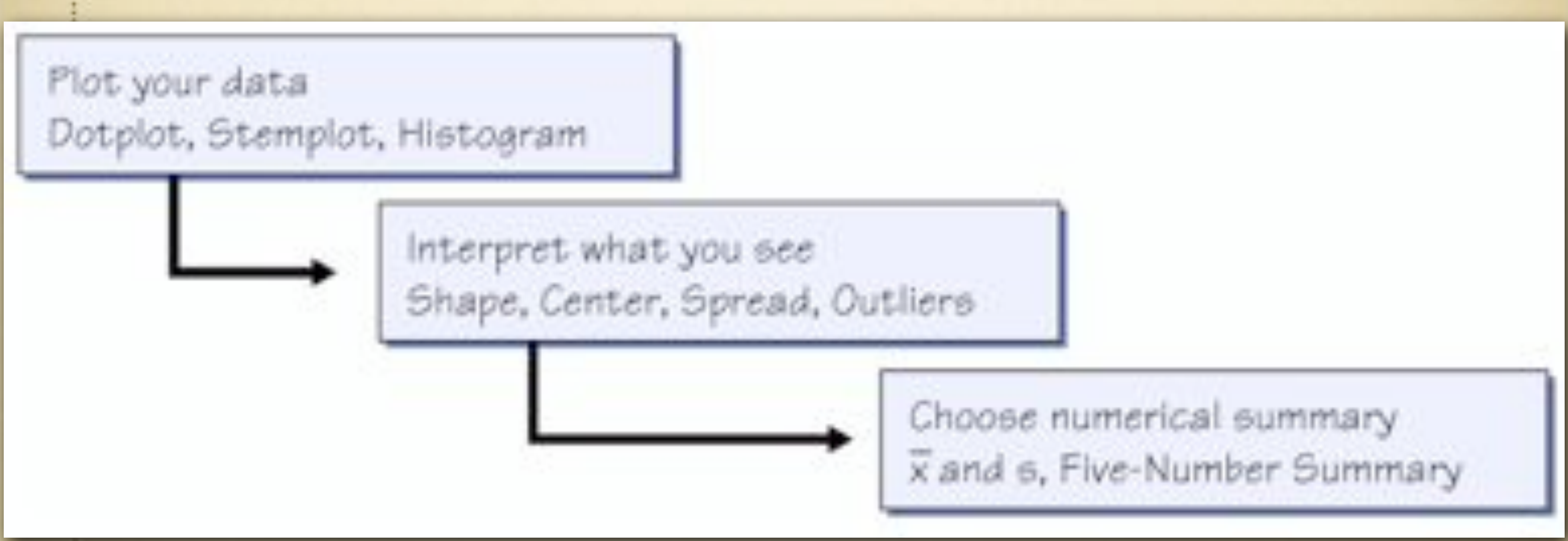
Data Analysis Toolbox

To answer a statistical question of interest:

- **Data**: Organize and Examine
  - **Who** are the individuals described?
  - **What** are the variables?
  - **Why** were the data gathered?
  - **When,Where,How,By Whom** were data gathered?
- **Graph**: Construct an appropriate graphical display
  - Describe **SOCS**
- **Numerical Summary**: Calculate appropriate center and spread (**mean and s *or* 5 number summary**)
- **Interpretation**: Answer question **in context**!

Chapter 1 Summary

- Data Analysis is the art of describing data in context using graphs and numerical summaries. The purpose is to describe the most important features of a dataset.



Plot your data
Dotplot, Stemplot, Histogram

Interpret what you see
Shape, Center, Spread, Outliers

Choose numerical summary
$\bar{x}$ and s, Five-Number Summary