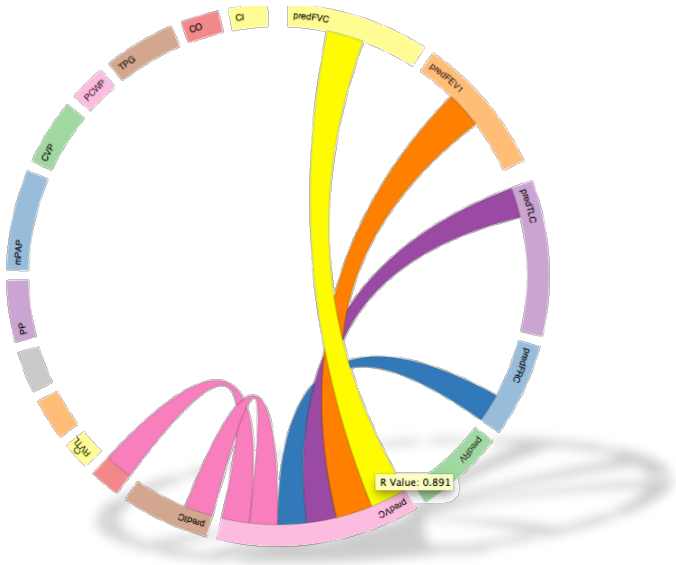
Module #1:
Introduction to Data Visualization



EN 605.462

Objectives

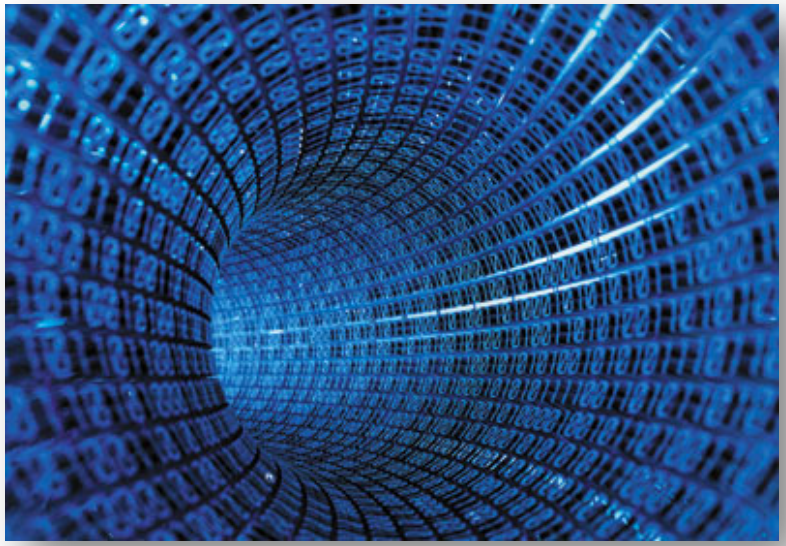
- Introduce basic concepts of Data Visualization
- Describe the history of Data Visualization
- Justify the purpose of data visualization
- Demonstrate misrepresentation of data using visualization techniques
- Understand general topics to be discussed in class

Data Growth

- The volume of digital data is exploding
  - more data has been created in the past two years than in the entire previous history of the human race

- Data is growing 40% a year[1]

- By 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes (i.e. 44 trillion gigabytes.) [1]



[1]www.cisco.com

How much data?

- Facebook:
  - over 1.79 billion active monthly users
  - Users like over 4 million posts every minute which adds up to 250 million posts per hour
  - Five new profiles every second
  - 300 millions pictures uploads per day

- Instagram:
  - 500 million monthly users in 2016
  - 1,736,111 likes on photos each minute of the day
  - 100 million likes per hour

- Twitter
  - 325 million active monthly users
  - Over 347,222 Tweets each minute – or 21 million Tweets per hour.

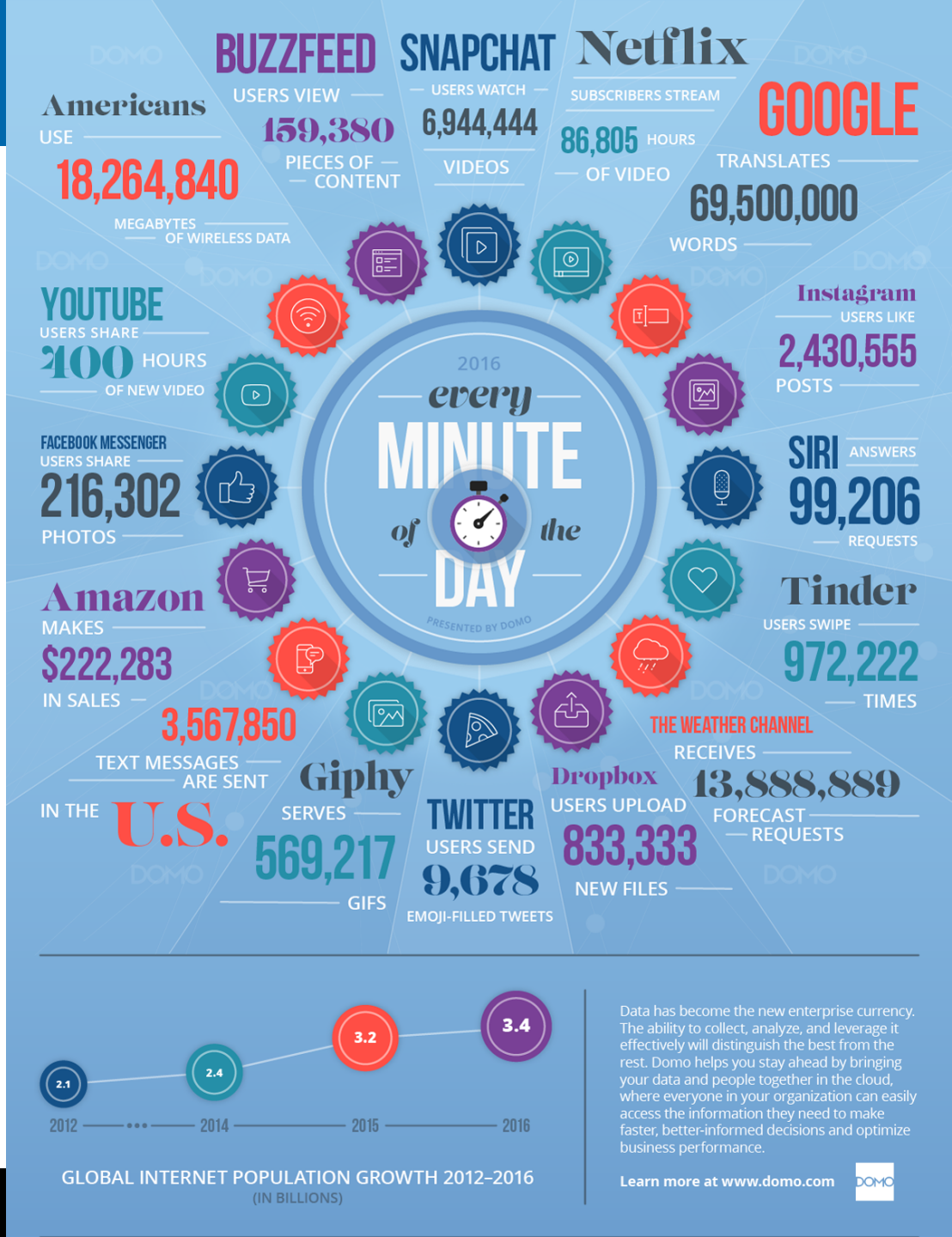
- Google
  - over 40,000 search queries every second
  - 3.5 billion searches per day

Source: Internet Live Stats

How much data?

- Netflix streams 87K hrs of video every minutes
- Over 3.5M text messages are sent
- 6.9M videos watched in Snapchat
- 69.6M words translated by Google

Source: www.domo.com



HOW CAN WE EFFECTIVELY EXPLORE COMPLEX DATASETS?

Sample Dataset

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.7	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.8	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Mean

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Variance

$$\text{Variance} = s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

Correlation

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Sample Dataset

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.7	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.8	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

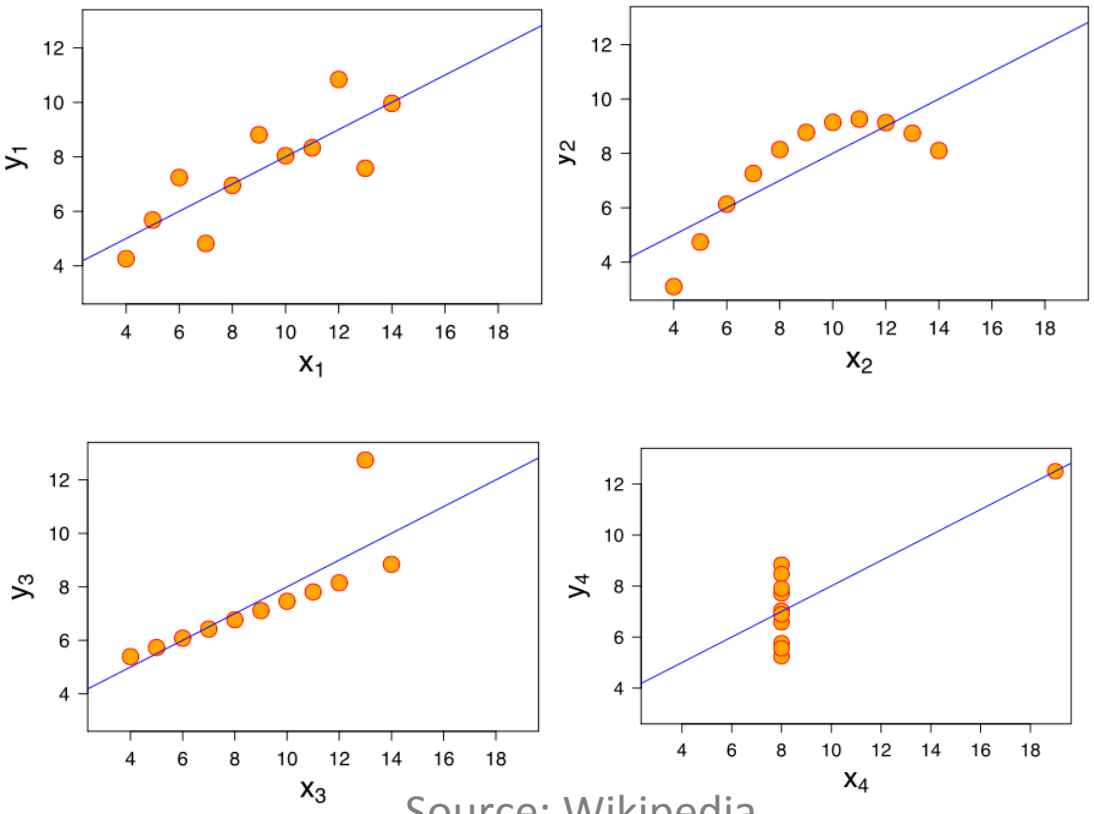
Descriptive Statistics	
N	11
Mean *x*	9
Mean *y*	8
Variance *x*	11
Variance *y*	4.12
Correlation *x&y*	0.816
Linear regression	y=3.00 + 0.500x

Anscombe's quartet

- Given the uncertainty and non-regular properties of some data, traditional data analysis techniques have many limitations.

Descriptive Statistics	
N	11
Mean *x*	9
Mean *y*	8
Variance *x*	11
Variance *y*	4.12
Correlation *x&y*	0.816
Linear regression	y=3.00 + 0.500x

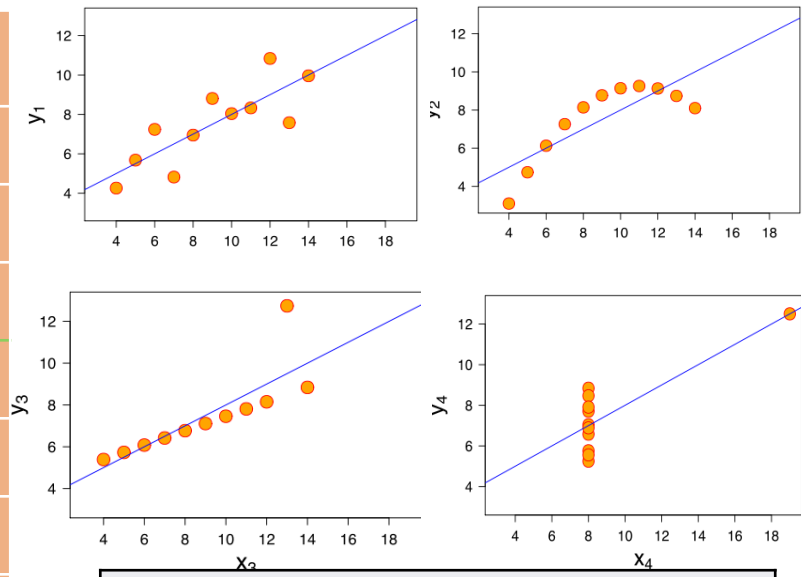
Anscombe's quartet



Source: Wikipedia

Anscombe's quartet

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.7	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.8	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

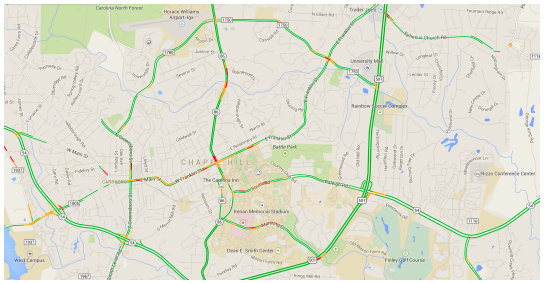


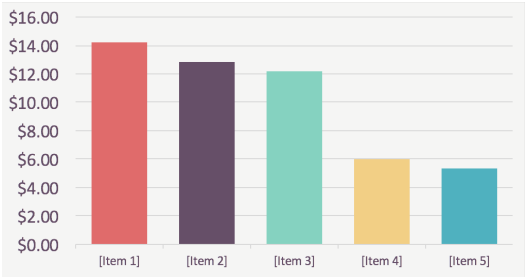
Descriptive Statistics	
N	11
Mean *x*	9
Mean *y*	8
Variance *x*	11
Variance *y*	4.12
Correlation *x&y*	0.816
Linear regression	y=3.00 + 0.500x

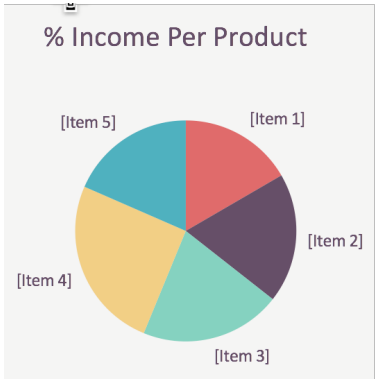
What's Data Visualization?

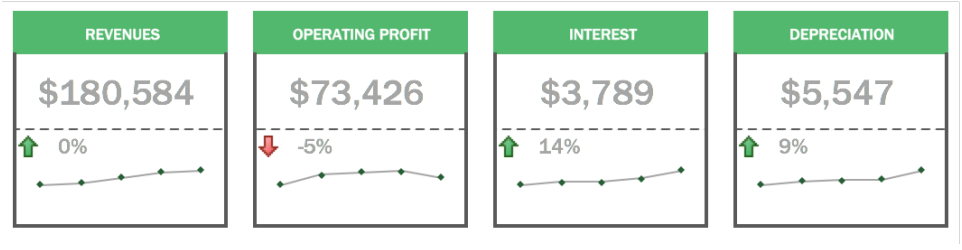
- "**Visualization** is the communication of information using graphical representations."
  - Ward et al., "Interactive Data Visualization"

		North	South	East	West
2011	1Q	$54,423.00	$51,234.00	$59,732.00	$58,534.00
	2Q	$51,345.00	$55,398.00	$57,423.00	$48,423.00
	3Q	$49,123.00	$46,245.00	$49,356.00	$49,976.00
	4Q	$45,923.00	$45,912.00	$54,989.00	$53,234.00
2012	1Q	$56,263.00	$87,690.00	$48,123.00	$63,343.00
	2Q	$52,103.00	$47,233.00	$49,325.00	$78,054.00
	3Q	$54,423.00	$52,344.00	$51,484.00	$53,012.00
	4Q	$51,345.00	$68,453.00	$53,323.00	$52,432.00
2013	1Q	$49,123.00	$45,234.00	$51,376.00	$49,643.00
	2Q	$52,103.00	$46,342.00	$34,376.00	$47,032.00
	3Q	$45,923.00	$35,432.00	$41,234.00	$45,123.00
	4Q	$56,263.00	$34,632.00	$44,532.00	$40,995.00



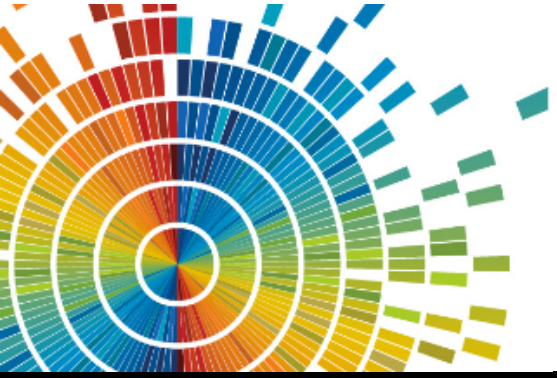






What's Data Visualization?

- Transformation of the symbolic into the geometric" [McCormick et al. 1987]
- "... finding the artificial memory that best supports our natural means of perception." [Bertin 1967]
- "The use of computer-generated, interactive, visual representations of data to amplify cognition." [Card, Mackinlay, & Shneiderman 1999]



First, A Test….

		North	South	East	West
2011	1Q	$54,423.00	$51,234.00	$59,732.00	$58,534.00
	2Q	$51,345.00	$55,398.00	$57,423.00	$48,423.00
	3Q	$49,123.00	$46,245.00	$49,356.00	$49,976.00
	4Q	$45,923.00	$45,912.00	$54,989.00	$53,234.00
2012	1Q	$56,263.00	$87,690.00	$48,123.00	$63,343.00
	2Q	$52,103.00	$47,233.00	$49,325.00	$78,054.00
	3Q	$54,423.00	$52,344.00	$51,484.00	$53,012.00
	4Q	$51,345.00	$68,453.00	$53,323.00	$52,432.00
2013	1Q	$49,123.00	$45,234.00	$51,376.00	$49,643.00
	2Q	$52,103.00	$46,342.00	$34,376.00	$47,032.00
	3Q	$45,923.00	$35,432.00	$41,234.00	$45,123.00
	4Q	$56,263.00	$34,632.00	$44,532.00	$40,995.00

Find the single highest sales figure.
How about the top 3? Or bottom 3?

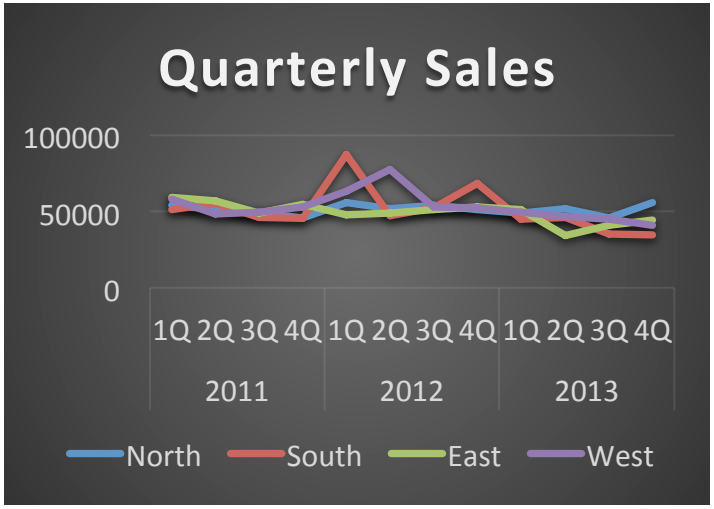
First, A Test….

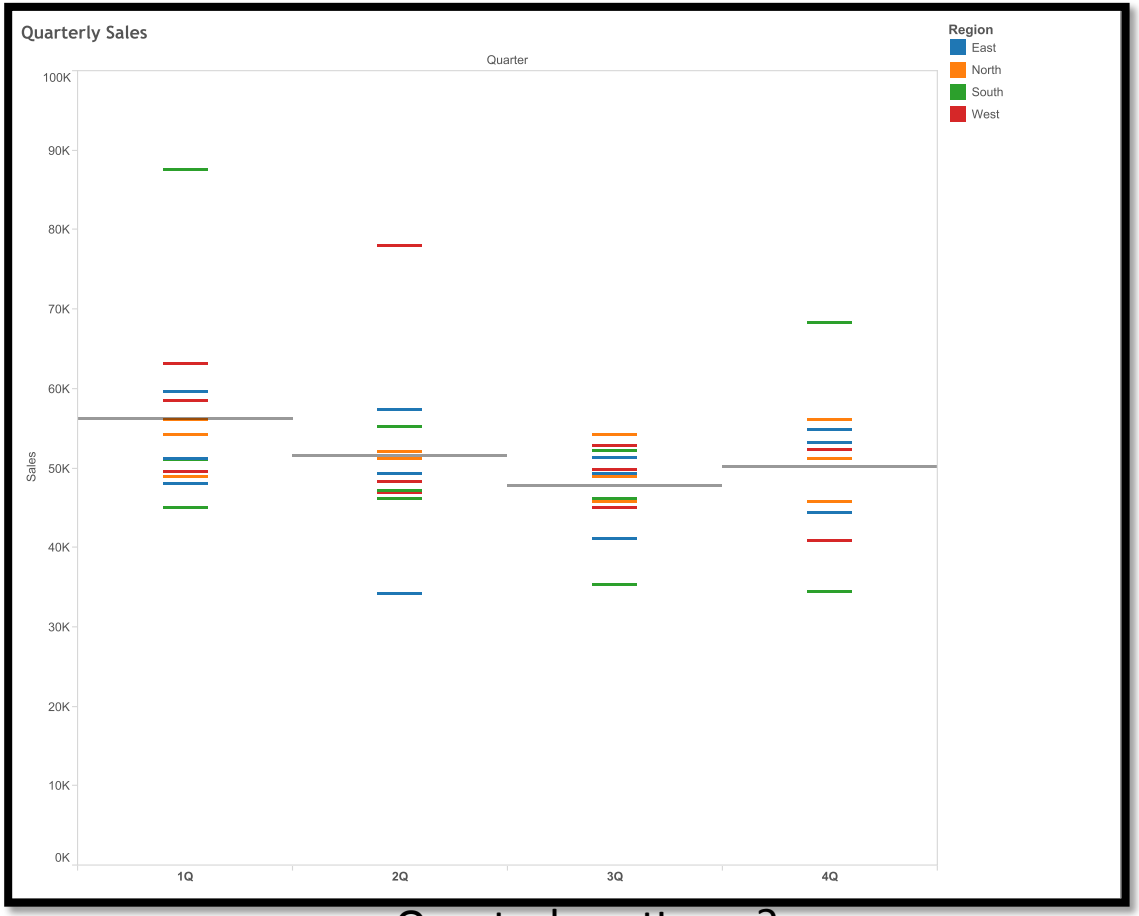
		North	South	East	West
2011	1Q	$54,423.00	$51,234.00	$59,732.00	$58,534.00
	2Q	$51,345.00	$55,398.00	$57,423.00	$48,423.00
	3Q	$49,123.00	$46,245.00	$49,356.00	$49,976.00
	4Q	$45,923.00	$45,912.00	$54,989.00	$53,234.00
2012	1Q	$56,263.00	$87,690.00	$48,123.00	$63,343.00
	2Q	$52,103.00	$47,233.00	$49,325.00	$78,054.00
	3Q	$54,423.00	$52,344.00	$51,484.00	$53,012.00
	4Q	$51,345.00	$68,453.00	$53,323.00	$52,432.00
2013	1Q	$49,123.00	$45,234.00	$51,376.00	$49,643.00
	2Q	$52,103.00	$46,342.00	$34,376.00	$47,032.00
	3Q	$45,923.00	$35,432.00	$41,234.00	$45,123.00
	4Q	$56,263.00	$34,632.00	$44,532.00	$40,995.00

Multiple Views of the Same Data

Lookup values; Identify Outliers

		North	South	East	West
2011	1Q	$54,423.00	$51,234.00	$59,732.00	$58,534.00
	2Q	$51,345.00	$55,398.00	$57,423.00	$48,423.00
	3Q	$49,123.00	$46,245.00	$49,356.00	$49,976.00
	4Q	$45,923.00	$45,912.00	$54,989.00	$53,234.00
2012	1Q	$56,263.00	$87,690.00	$48,123.00	$63,343.00
	2Q	$52,103.00	$47,233.00	$49,325.00	$78,054.00
	3Q	$54,423.00	$52,344.00	$51,484.00	$53,012.00
	4Q	$51,345.00	$68,453.00	$53,323.00	$52,432.00
2013	1Q	$49,123.00	$45,234.00	$51,376.00	$49,643.00
	2Q	$52,103.00	$46,342.00	$34,376.00	$47,032.00
	3Q	$45,923.00	$35,432.00	$41,234.00	$45,123.00
	4Q	$56,263.00	$34,632.00	$44,532.00	$40,995.00





Quarterly patterns?

Why Data Visualization?

- Human Pattern Recognition
  - Identifying Outliers
- Summary Statistics Are Only Part of Story

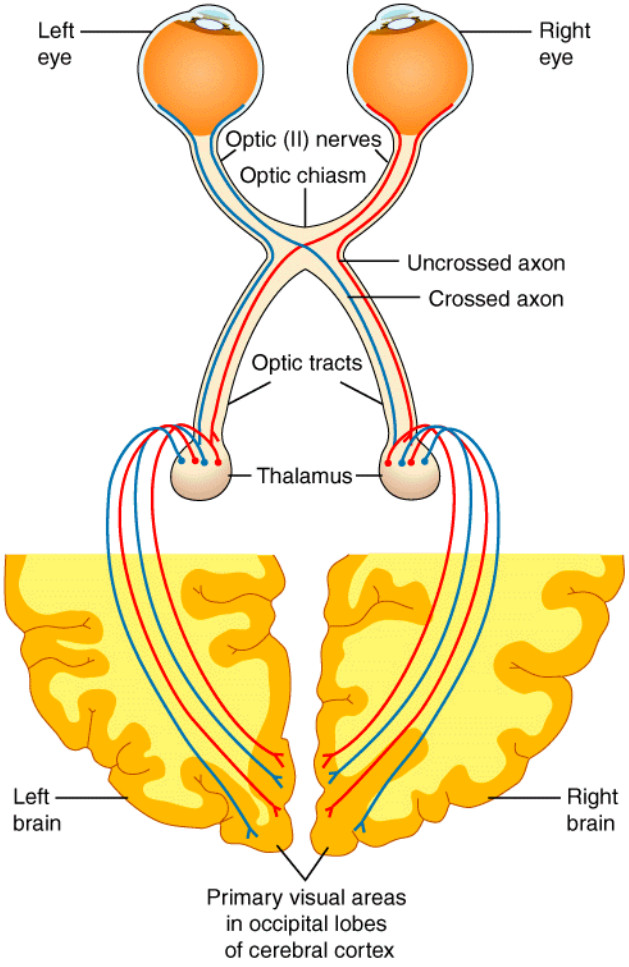
Why Does Visualization Work?

- Why is **a good** visualization easier to "see" than tables of numbers?

		North	South	East	West
2011	1Q	$54,423.00	$51,234.00	$59,732.00	$58,534.00
	2Q	$51,345.00	$55,398.00	$57,423.00	$48,423.00
	3Q	$49,123.00	$46,245.00	$49,356.00	$49,976.00
	4Q	$45,923.00	$45,912.00	$54,989.00	$53,234.00
2012	1Q	$56,263.00	$87,690.00	$48,123.00	$63,343.00
	2Q	$52,103.00	$47,233.00	$49,325.00	$78,054.00
	3Q	$54,423.00	$52,344.00	$51,484.00	$53,012.00
	4Q	$51,345.00	$68,453.00	$53,323.00	$52,432.00
2013	1Q	$49,123.00	$45,234.00	$51,376.00	$49,643.00
	2Q	$52,103.00	$46,342.00	$34,376.00	$47,032.00
	3Q	$45,923.00	$35,432.00	$41,234.00	$45,123.00
	4Q	$56,263.00	$34,632.00	$44,532.00	$40,995.00

Why Does Visualization Work?

- Why is **a good** visualization easier to "see" than tables of numbers?
- Our visual systems have tremendous power to:
  - See **patterns**
  - Identify **Trends**
  - Locate **Outliers and anomalies**
- Much of that power is **precognitive**
  - Fast
  - Efficient



Why do we create visualizations?

1. Answer questions (or discover them)
2. Make decisions
3. See data in context
4. Expand memory
5. Support graphical calculation
6. Find patterns
7. Present argument or tell a story
8. Inspire

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING