# Quantitative Cyber-Security

## Colorado State University

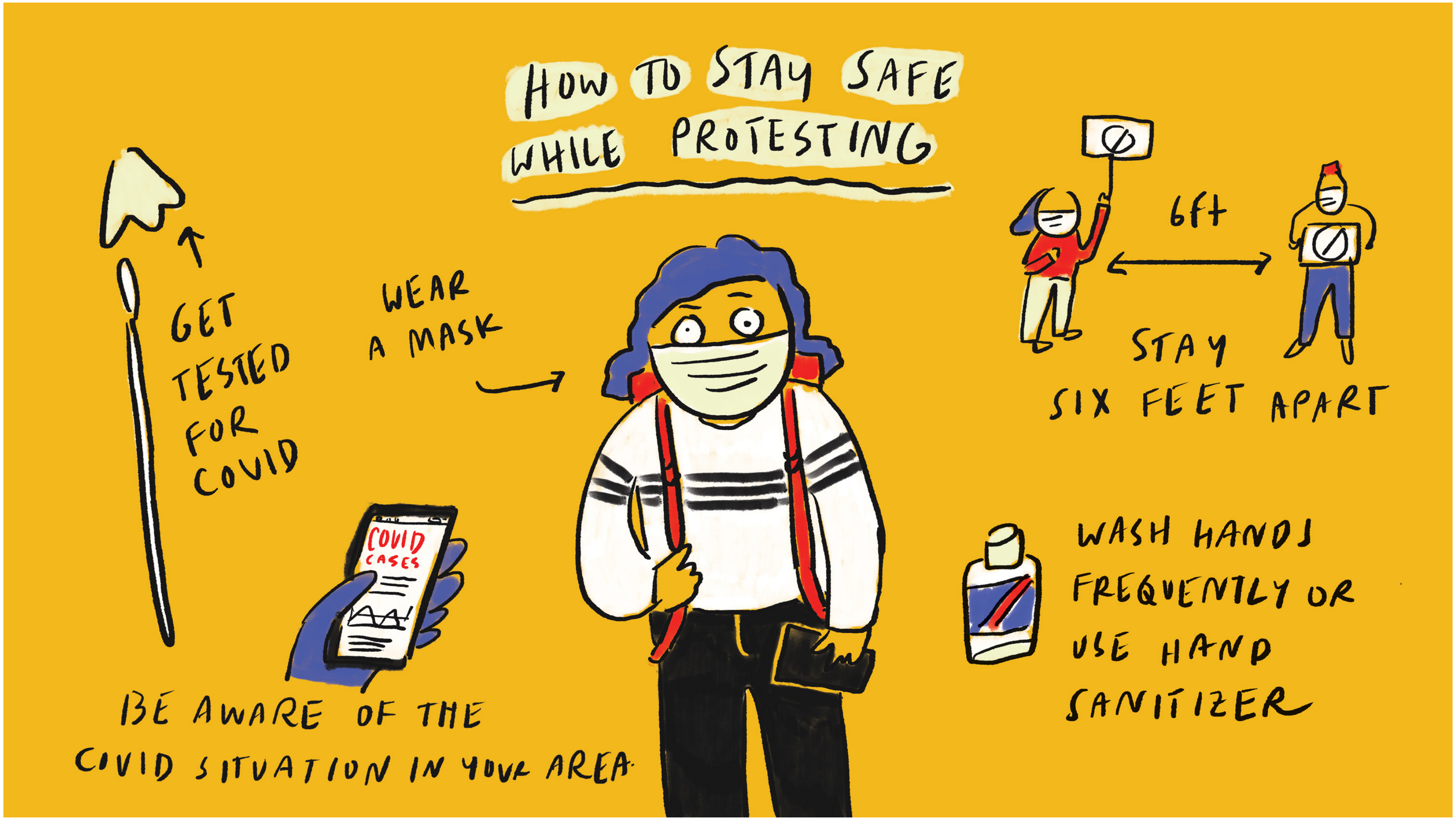
## Yashwant K Malaiya

## CS559

## Quick Research Presentations Tu c

CSU Cybersecurity Center
Computer Science Dept

# Stay Safe!


HOW TO STAY SAFE WHILE PROTESTING
GET TESTED FOR COVID
WEAR A MASK
6ft
STAY SIX FEET APART
COVID CASES
BE AWARE OF THE COVID SITUATION IN YOUR AREA.
WASH HANDS FREQUENTLY OR USE HAND SANITIZER

Colorado State University

# Tuesday

- Presenters: limit yourself to 5 minutes, 1 minute for q/c
  - Upload your slides and be ready to present
- The Peer Review Form (Canvas Assignments) due tomorrow
  - Novelty/ Interest, Technical/ Research, Presentation
- 7 **Annual security breach costs incurred to society/government/nations**

  Sarah Houlton
- 3 **Quant. Examination of schemes for discovering previously unknown vulnerabilities**

  Don Neumann
- 5 **Assessing probability of security breaches**

  Siddhi Kotian

  Dhruv Padalia

# Quantitative Cyber-Security

## Colorado State University

## Yashwant K Malaiya

## CS559

## Quick Research Presentations Tu c

CSU Cybersecurity Center
Computer Science Dept

# FAQ

- Note that the algebra involving the probabilities is often simple. It is **going from a verbal description to an algebraic representation** of the problem, where the challenge often lies.

**Example: A drug test produces 99% true positive and 99% true negative results. 0.5% are drug users. If a person tests positive, what is the probability he is a drug user?**

- 99% true positive means 99% of the actual drug users will test positive.
- 99% true negative means 99% of the non-drug users will test negative.
- We know P{P/DU} = 0.99. But we need to find P{DU/P}. Thus you need **Bayes' rule**.
- Numerator is 0.99x(0.5/100)
- The second term of the denominator is (1-0.99)x(99.5/100)
- See the rest posted in Discussions L6,L7.

**Short videos**: Firewall, Access control, Binomial dist., Bayes’ theorem

# FAQ

**Mitre ATT&K framework**

- Very detailed, result of a lot of work
- Organizes numerous attack techniques into 11 tactics (Enterprise), from Initial Access to Exfiltration and Impact.
- Can be used to plan attacks and defensive mechanisms
- Several tools are based on ATT&CK
- Detailed documentation

# Quantitative Security

## Colorado State University

## Yashwant K Malaiya

## CS 599

## Modeling and Regression

CSU CyberCenter
*Course Funding Program* – 2019

# Modeling & Regression

# Model vs Reality

- Reality is complex, often too complex
- A model represents the actual behavior of interest with acceptable accuracy
  - System design & evaluation of alternatives
  - System simulation
  - Predictions
- When there are multiple models, how do you choose one?
  - Familiarity, convenience?
  - Goodness of fit?
  - Predictive capabilities?

# Quantitative models

- **Derived from first principles**:
  - Arguments are actual things measured
  - Example: monthly mortgage payment

$$c = \frac{rP}{1 - (1 + r)^{-N}}$$

- **Empirical**
  - Arguments are just parameters
  - Example Ideal body weight in kg = a+b (height in in. - c)

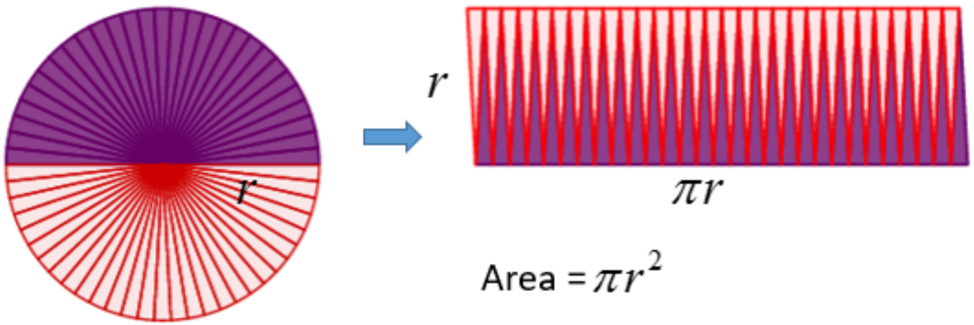
    Men: a=52 kg, b= 1.9, c = 60 in.

- **Combined**
  - Empirical with interpretation of parameters
  - From first principles, adjusted to fit

# Derived using first principles

- Obtain model
  - Understand how things work
  - Use approximations
  - Derive formulas
- Validate:
  - Get real data
  - See if it fits. If not,
    - Make adjustments
    - Get alternative models
- Use the model
  - Plug in values to estimate parameters
  - Make projections

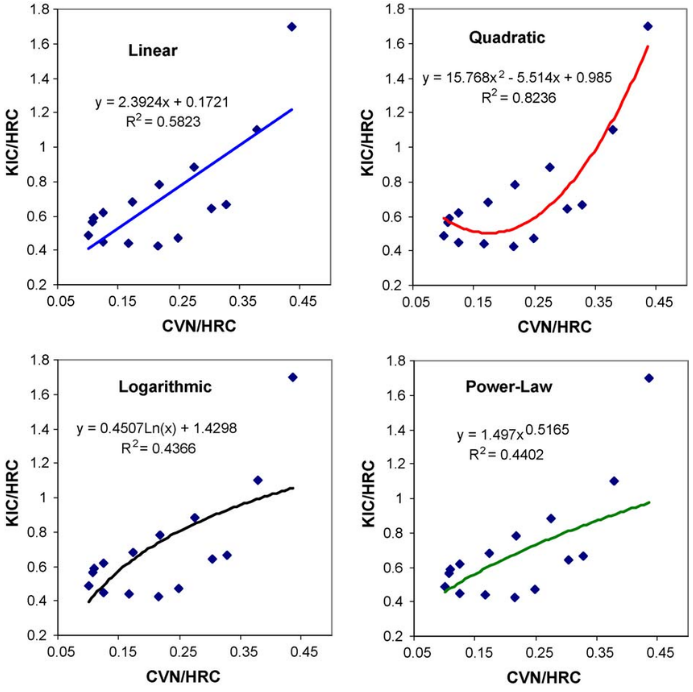
Proof of Area of Circle Formula



Area = $\pi r^2$

# Empirical models

- Look at data
- See if it resembles a function
  - Linear, quadratic, logarithmic, exponential..
  - Involving 1, 2 or more parameters
- See if it fits
  - If not try something more complex
- If it fits, see if an interpretation of the parameters is possible
  - Not necessary  but will be good.

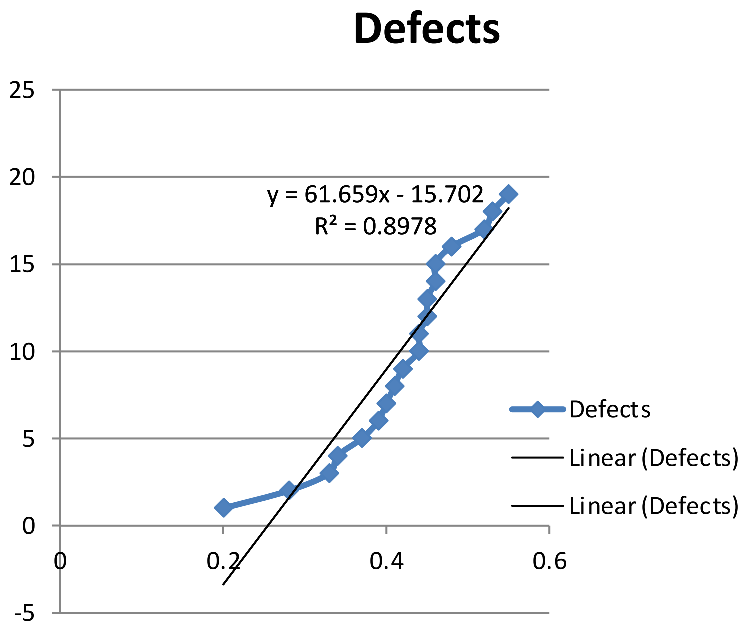


# Curve fitting

- Get x-axis and y-axis numbers ($x_i$, $y_i$).
- Draw a scatter plot.
- Fit (i.e. estimate parameters):
  - Use formulas for parameters
  - linest( ) or logest( ) functions in excel
  - Select plot, rc, add trendline, select display options in excel
  - Use Solver for general fitting for excel
- Open Excel sheet Examples.xlsx

# Example

| Branch Cov | Defects |
|---|---|
| 0.2 | 1 |
| 0.28 | 2 |
| 0.33 | 3 |
| 0.34 | 4 |
| 0.37 | 5 |
| 0.39 | 6 |
| 0.4 | 7 |
| 0.41 | 8 |
| 0.42 | 9 |
| 0.44 | 10 |
| 0.44 | 11 |
| 0.45 | 12 |
| 0.45 | 13 |
| 0.46 | 14 |
| 0.46 | 15 |
| 0.48 | 16 |
| 0.52 | 17 |
| 0.53 | 18 |
| 0.55 | 19 |



# Curve fitting vs Predictive Capability

- A good model has good predictive capabilities.
- Curve fitting partial data may not necessarily identify the best model.

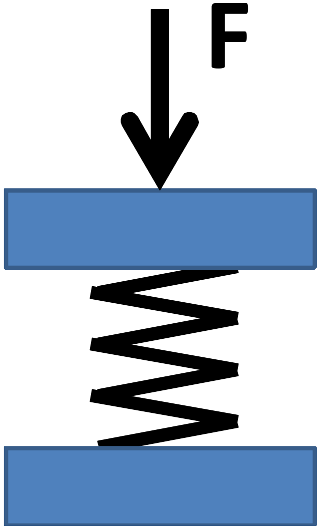
# Curve Fitting

From

*Engineering Computation: An Introduction Using MATLAB and Excel, Ch 5*

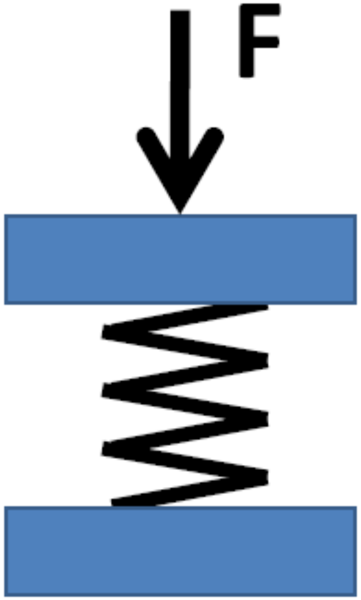
- Often, we have data points and we want to find an equation that "fits" the data
- Simplest equation is that of a straight line

- A spring is placed between two flat plates and force is slowly applied to the upper plate, causing the spring to compress. When the force reaches a pre-load value of 25 N, the location of the upper plate is recorded. As the upper plate continues to move, the distance that the top plate moves, *d*, and the force required to move the plate, *F*, are recorded. As soon as the upper plate has moved 10 mm, the test is ended.



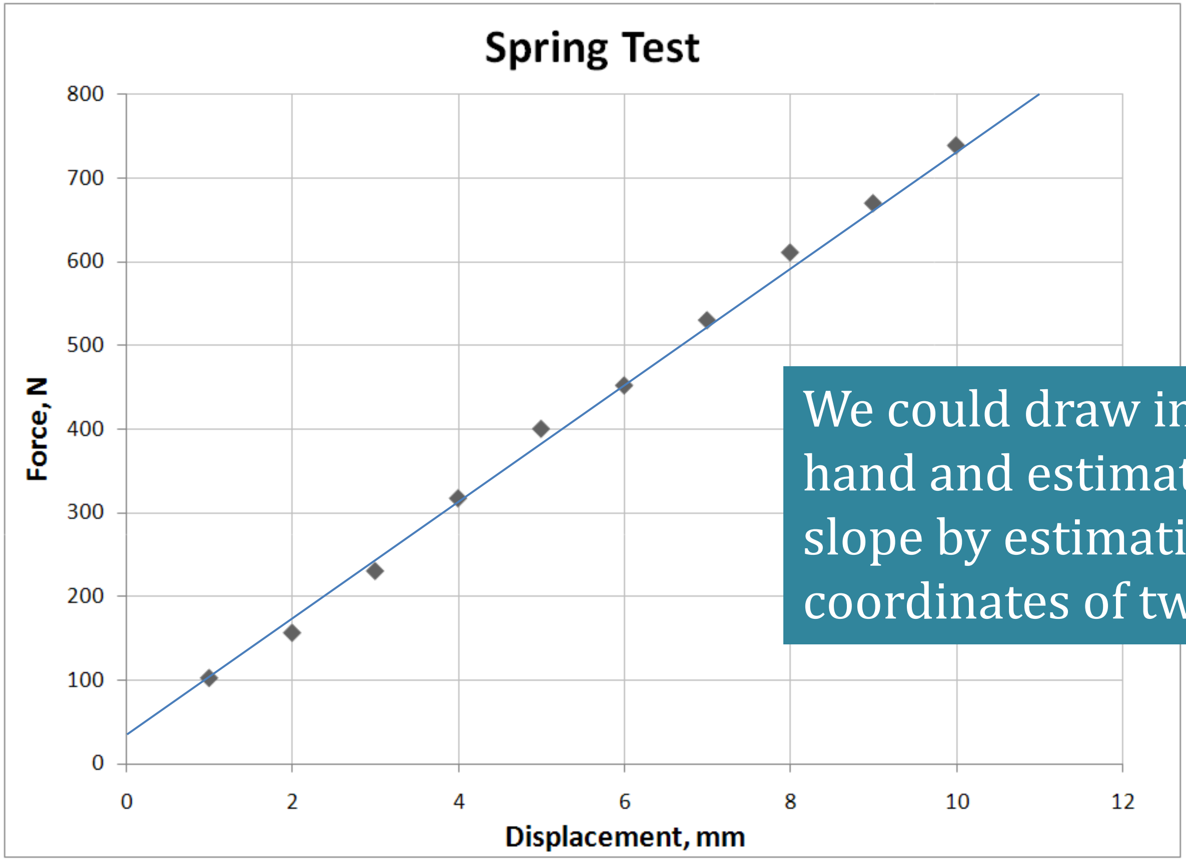
# Curve Fitting Example

- Data from the test is shown here. The slope of the load-displacement curve is called the spring constant.



| d, mm | F, N |
|---|---|
| 1 | 102 |
| 2 | 156 |
| 3 | 230 |
| 4 | 317 |
| 5 | 400 |
| 6 | 452 |
| 7 | 530 |
| 8 | 611 |
| 9 | 670 |
| 10 | 739 |

- Data from a test – how do we find the slope?



We could draw in a line by hand and estimate its slope by estimating the coordinates of two points

# Curve Fitting Example

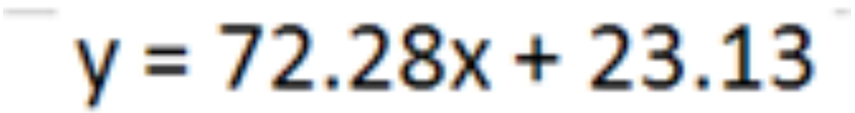
- How do we find the line that *best* fits the data?
- Find the equation of the line  that minimizes the sum of the *squares* of the differences between the values predicted from the equation and the actual data values
- Why do we square the differences?  Because we are interested in the magnitudes of the differences.  If one point is above the line and other is below it, we don’t want these difference to cancel each other

- In Excel, we can choose to have the equation of the best-fit line (the trendline) displayed in the form

$$y = mx + b$$

where m = the slope of the line

and b = the intercept of the line with the y-axis

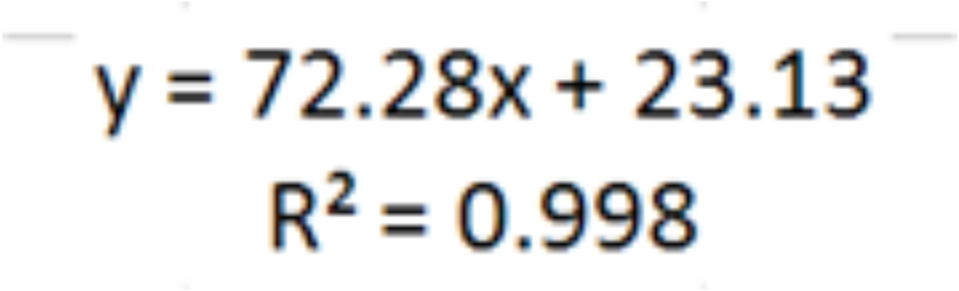


$$y = 72.28x + 23.13$$

- In this example, the slope of the line (the stiffness of the spring) equals 72.3 N/mm
- When the deflection (x) equals zero, the force (y) equals 23.1 pounds. This is consistent with our nominal pre-load of 25 pounds

# Correlation Coefficient

- The *correlation coefficient* ($R^2$) is a measure of how well the trendline fits the data
- A value of one represents a perfect fit
- In our example, the line fit is very good

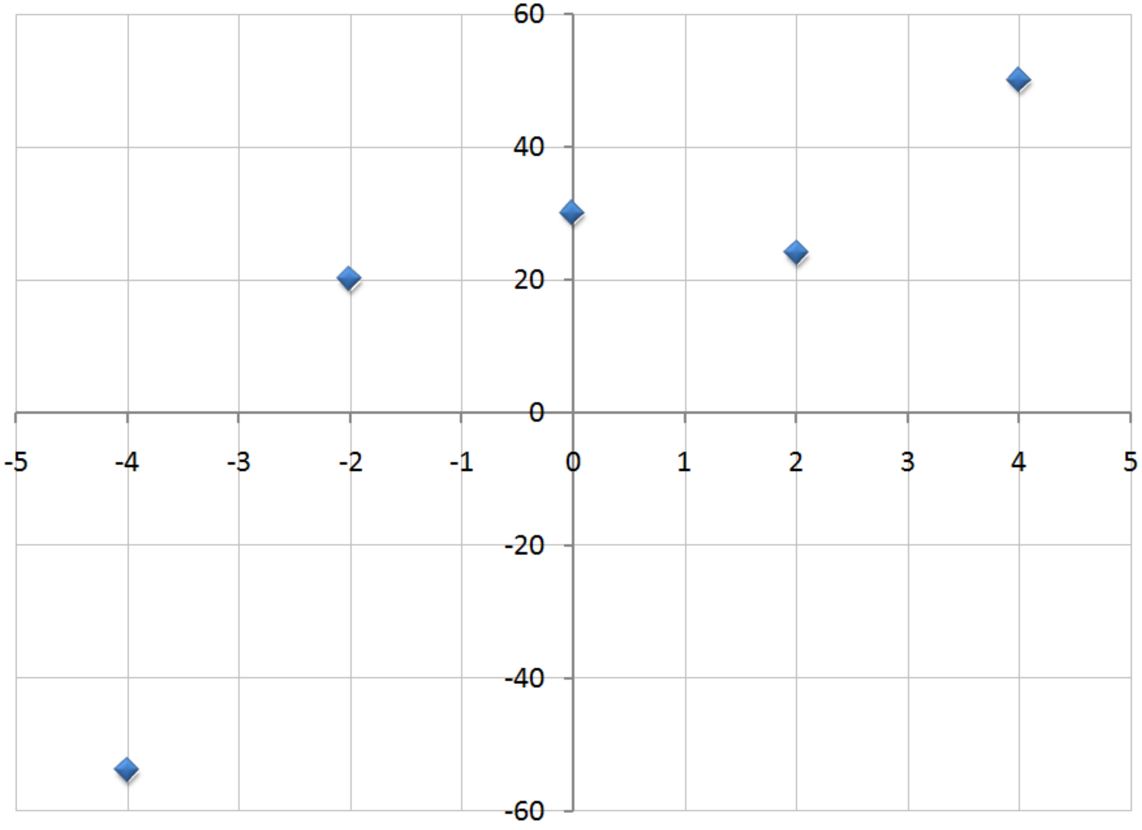


$y = 72.28x + 23.13$

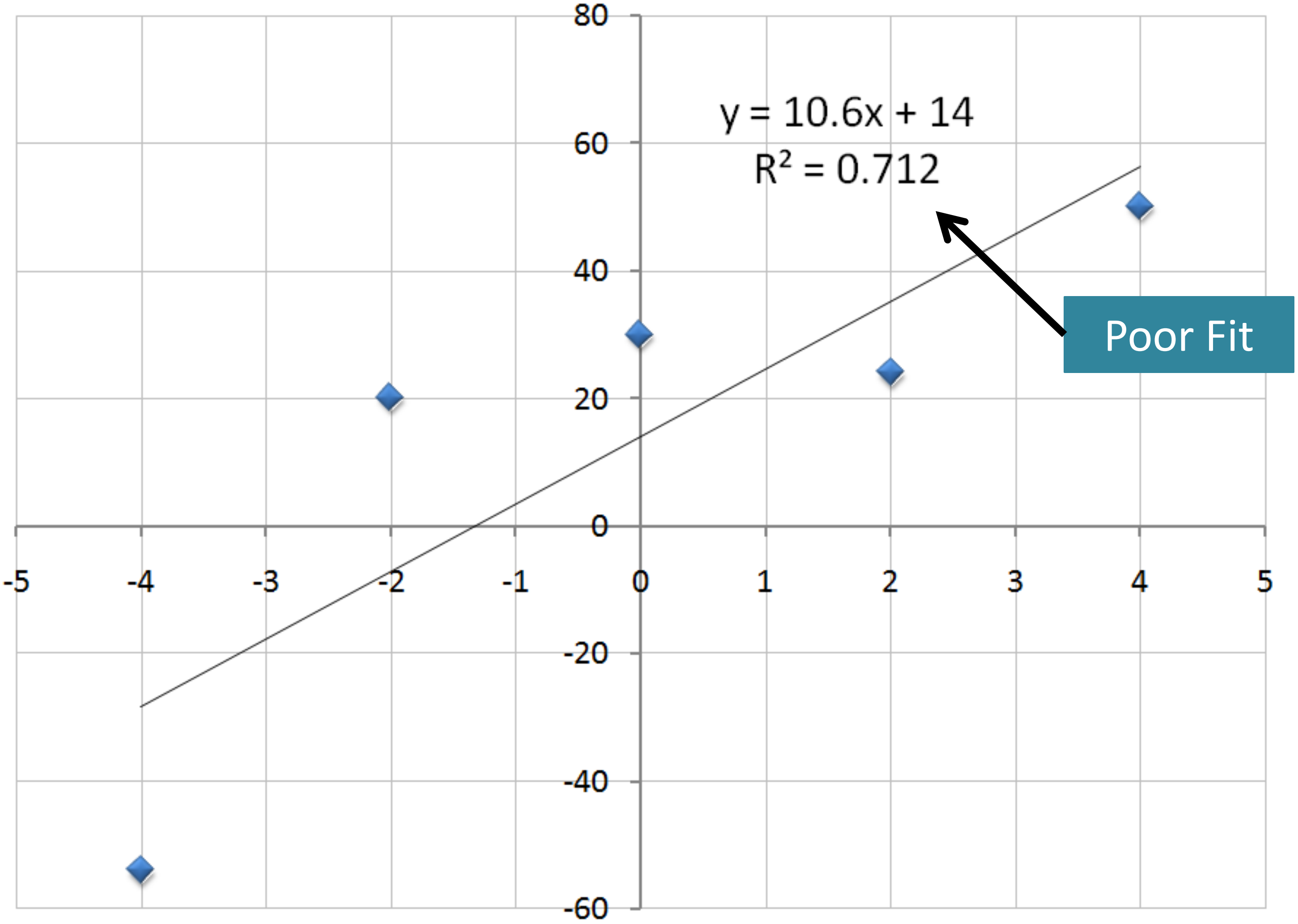
$R^2 = 0.998$

- Consider these five data points:

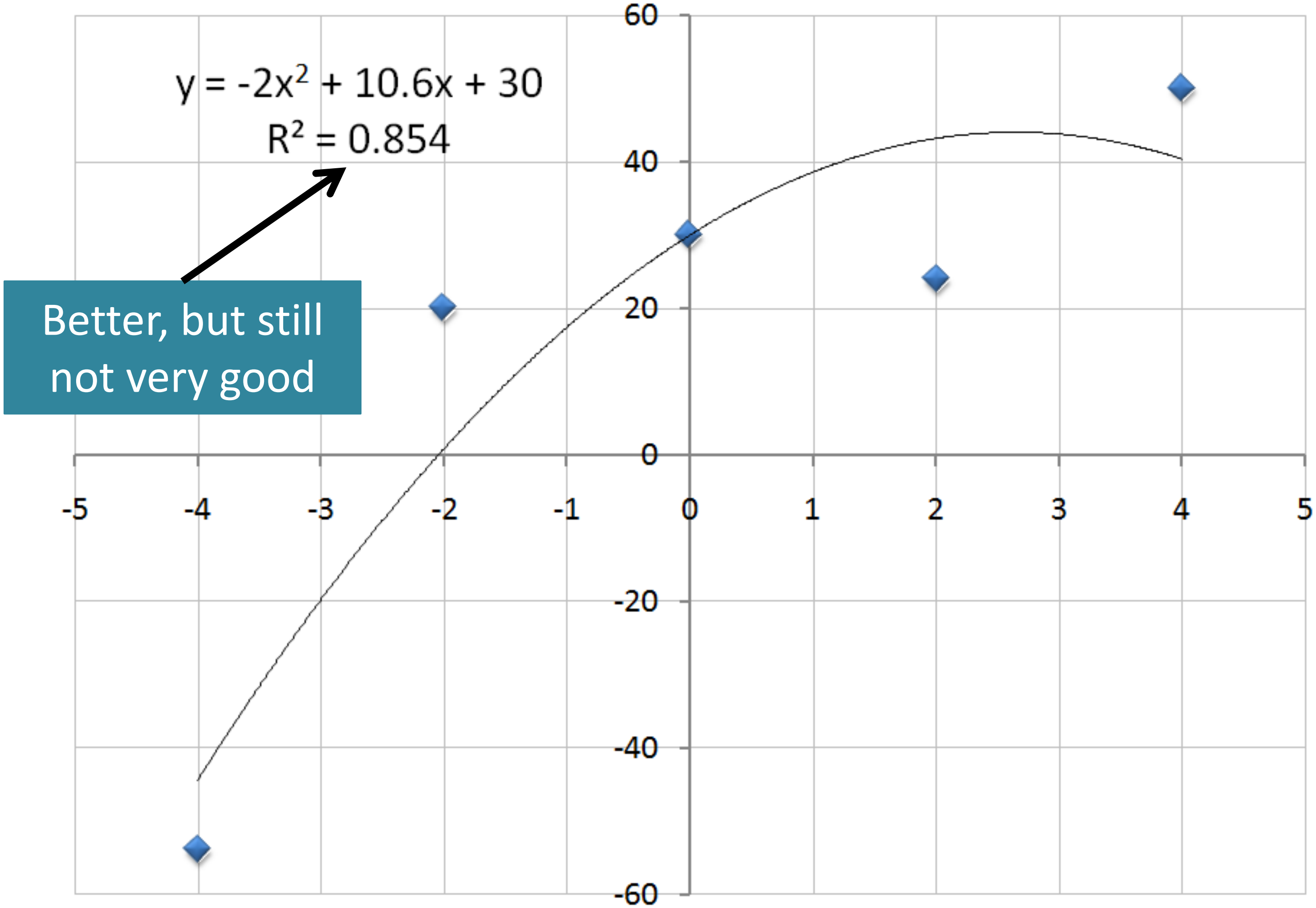
| x | y |
|---|---|
| -4 | -54 |
| -2 | 20 |
| 0 | 30 |
| 2 | 24 |
| 4 | 50 |



# Linear Curve Fit

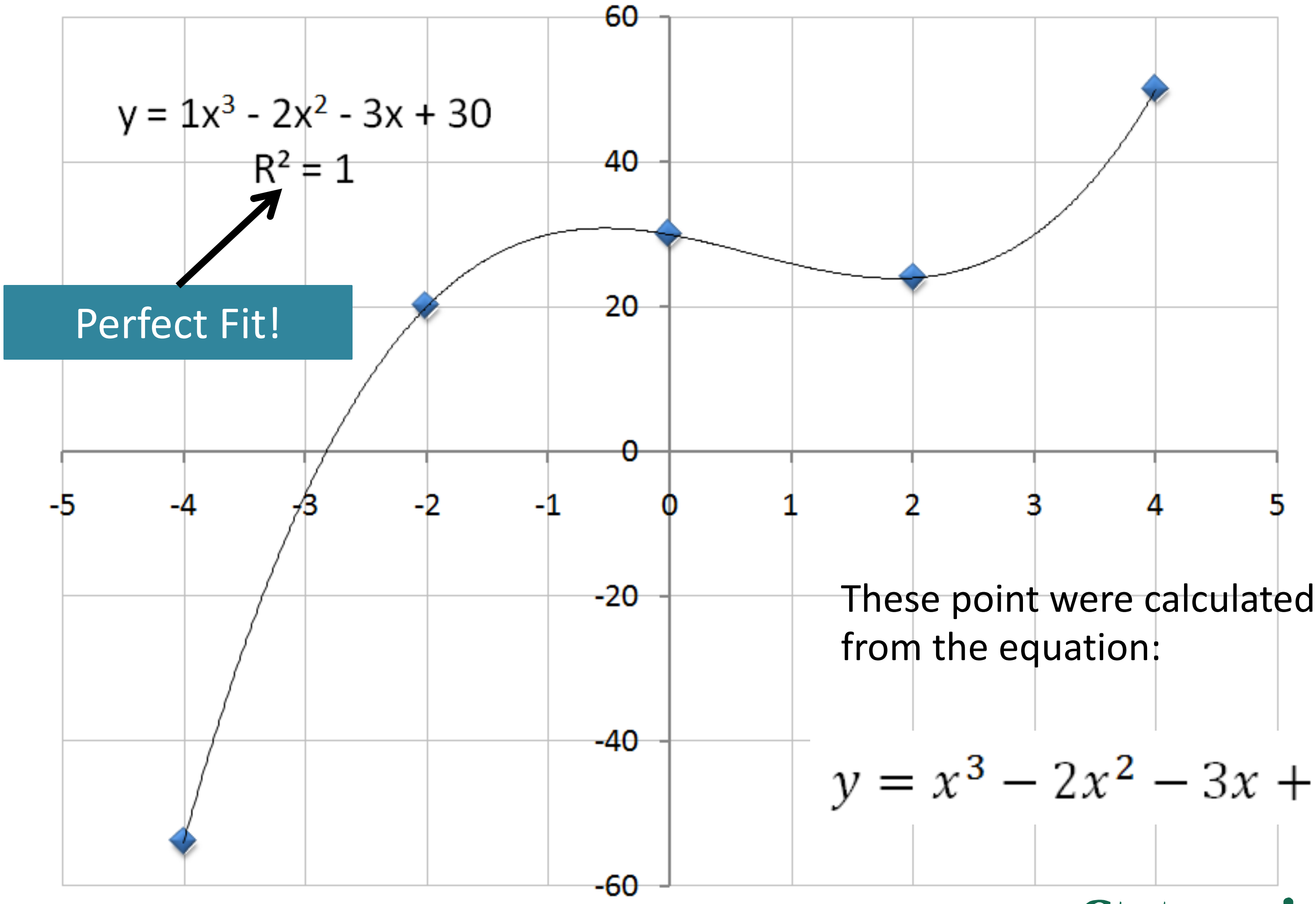


# Try Second–Order Polynomial Fit



Better, but still not very good

# Third-Order Polynomial Fit



Perfect Fit!

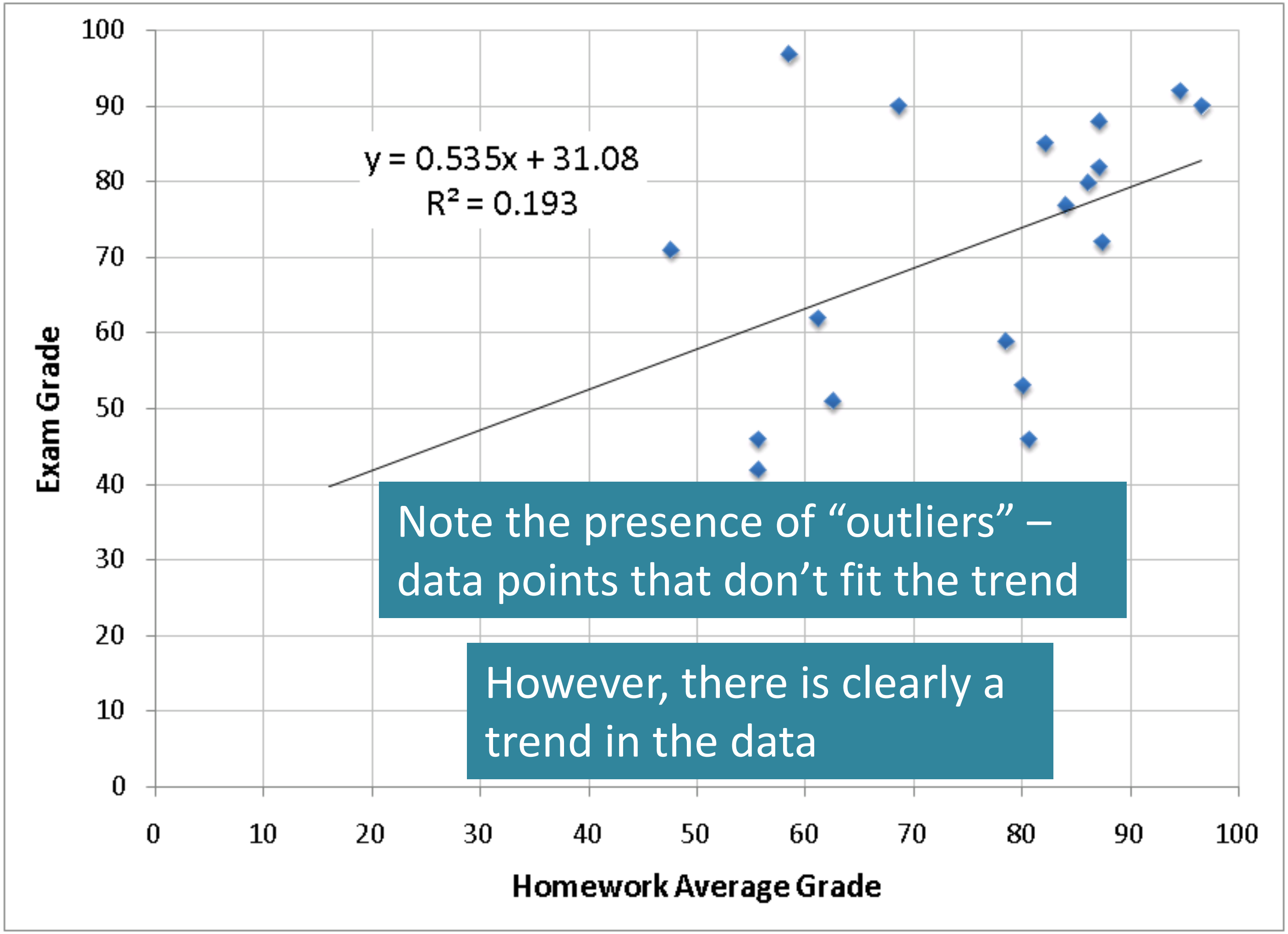
These point were calculated from the equation:

$$y = x^3 - 2x^2 - 3x + 30$$

# Correlation Coefficient

- A “good” fit to data is relative
- In the case of the spring example, the data should fit a mathematical model, and so an $R^2$ value of close to one is expected
- For other cases, a much lower $R^2$ value is expected
- Consider a comparison of final exam scores in a class vs. homework averages
- We would expect that students who do well on HW will generally do well on the final exam, but there will be exceptions

# Exam vs. HW Example



Note the presence of “outliers” – data points that don’t fit the trend

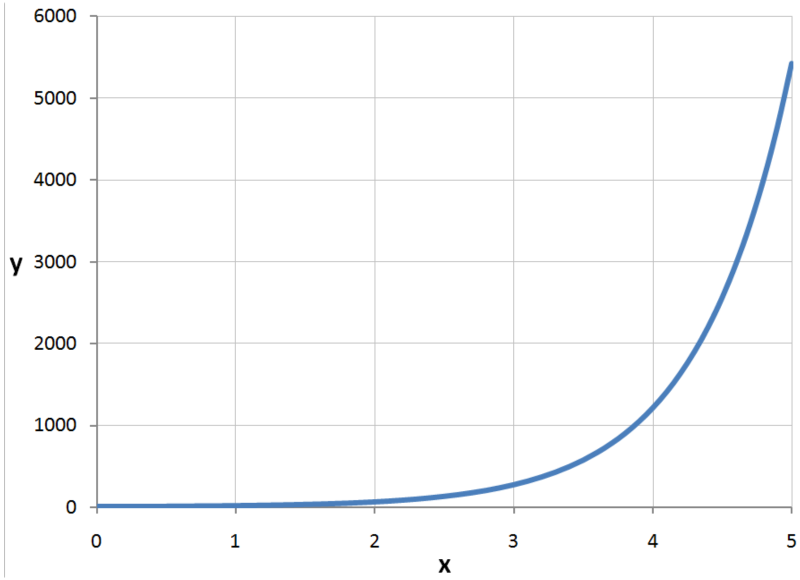
However, there is clearly a trend in the data

# Exponential and Power Equations

- Consider this equation:

- Here is a graph of the equation:

$$y = 3e^{1.5x}$$



Engineering Computation: An Introduction Using MATLAB and Excel

Colorado State University

$$y = 3e^{1.5x}$$

$$\ln(y) = \ln(3e^{1.5x})$$

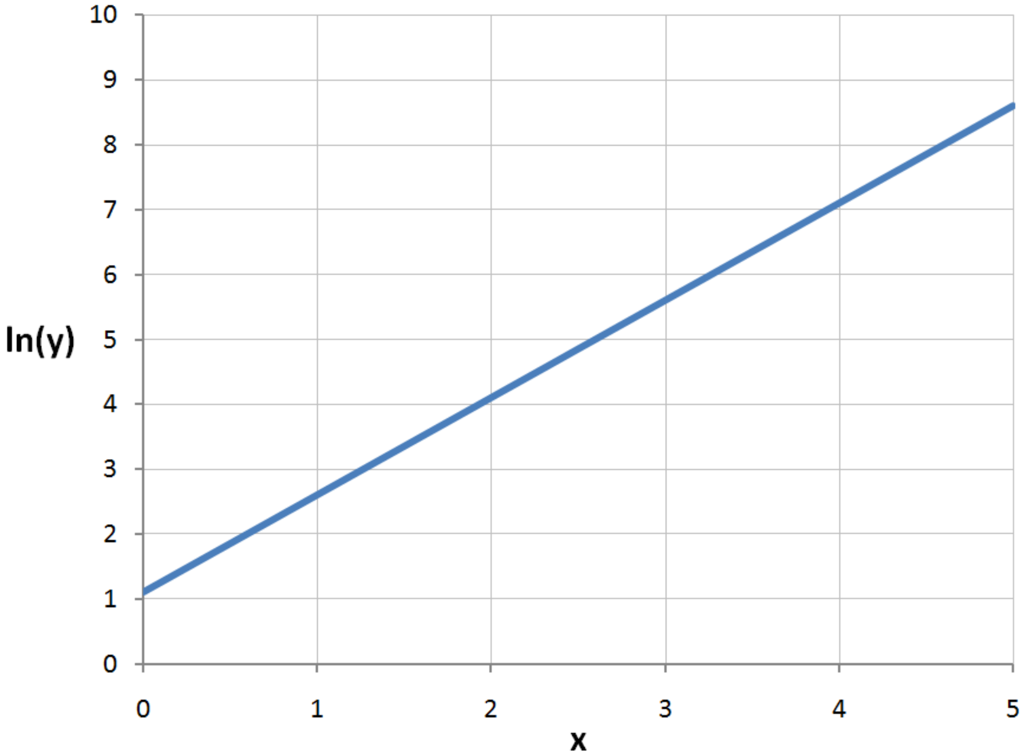
$$\ln(y) = \ln(3) + \ln(e^{1.5x})$$

$$\ln(y) = 1.5x + \ln(3)$$

If we plot the ln(*y*) instead of *y*, then we have the equation of a straight line

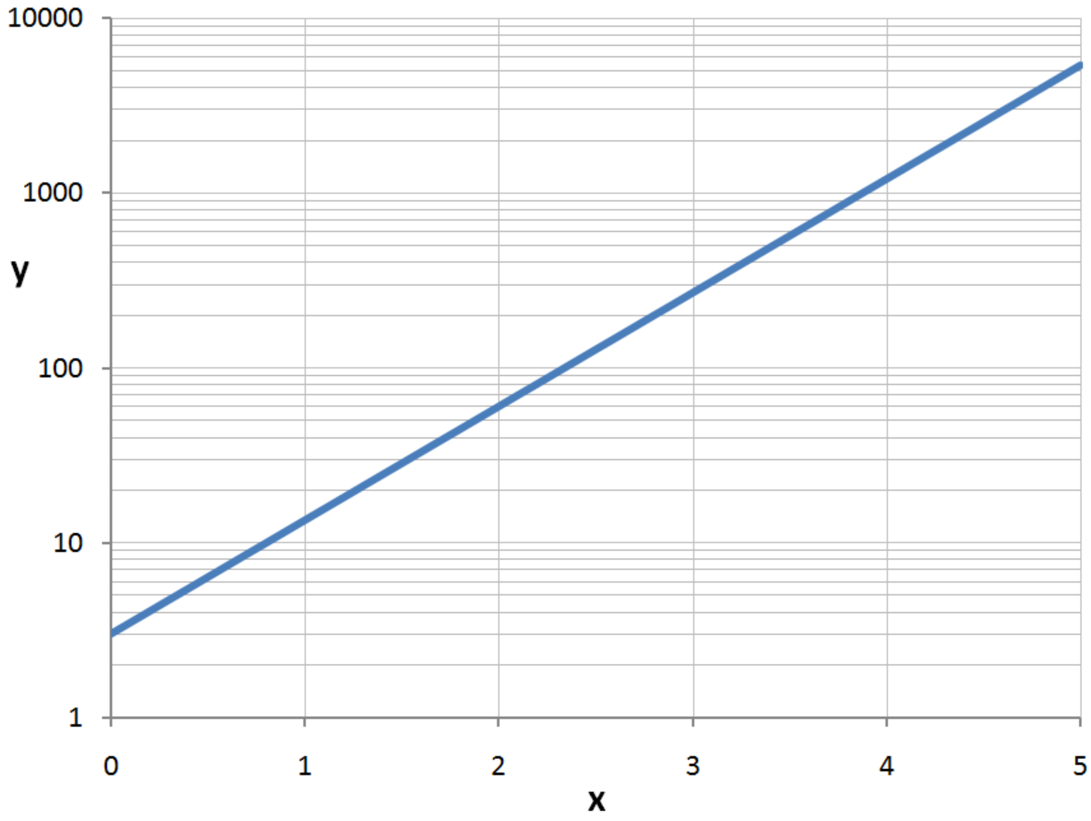
# Exponential Equation

- This plot is not particularly useful, since it requires us to read the ln of the dependent variable *y*, instead of *y* itself



# Exponential Equation

- A better way is to display the *y* values on a *logarithmic scale*:



Colorado State University

# Exponential Equation

- We call this a *semi-log* plot since one of the axes is logarithmic
- Note that we have used a base-10 scale: remember that a logarithm can be converted to a logarithm of another base by multiplying by a constant:

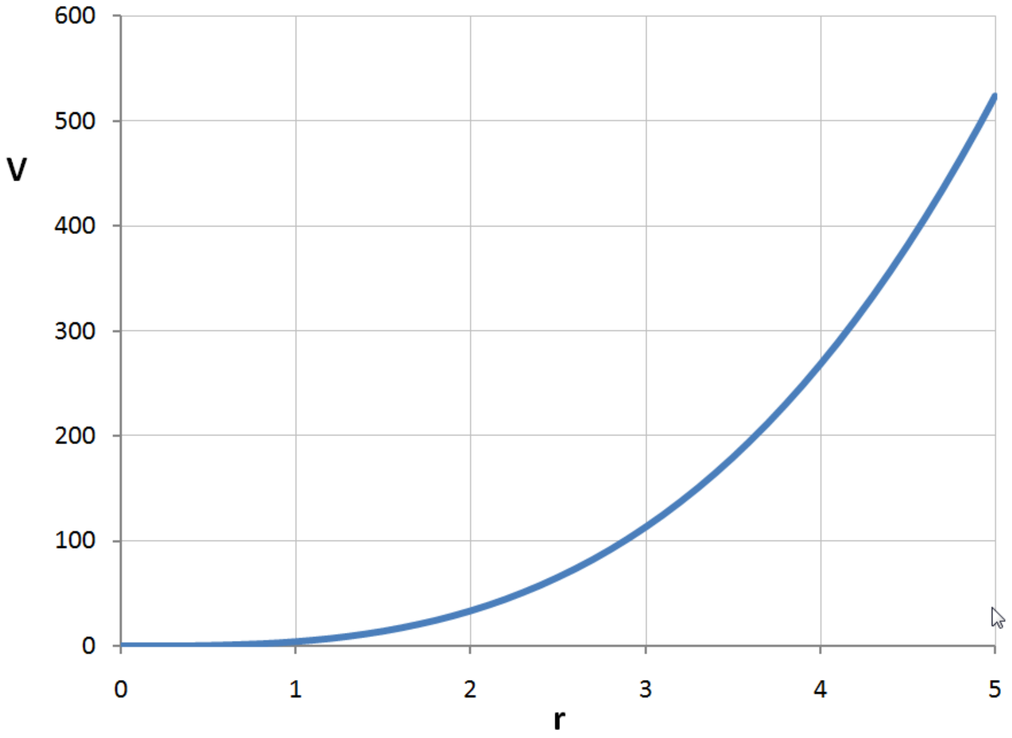
$$\log_{\mathrm{a}}(x) = \log_{\mathrm{b}}(x) * \log_{\mathrm{a}}(\mathrm{b})$$

- Therefore, the equation will produce a straight line on semi-log axes regardless of the base of the logarithmic scale.  Base 10 normally used – easiest to read
- Logarithmic scales can be specified in Excel and MATLAB

- Consider the equation for the volume of a sphere:

$$V = \frac{4}{3}\pi r^3$$

Plot:



# Power Equations

$$V = \frac{4}{3}\pi r^3$$

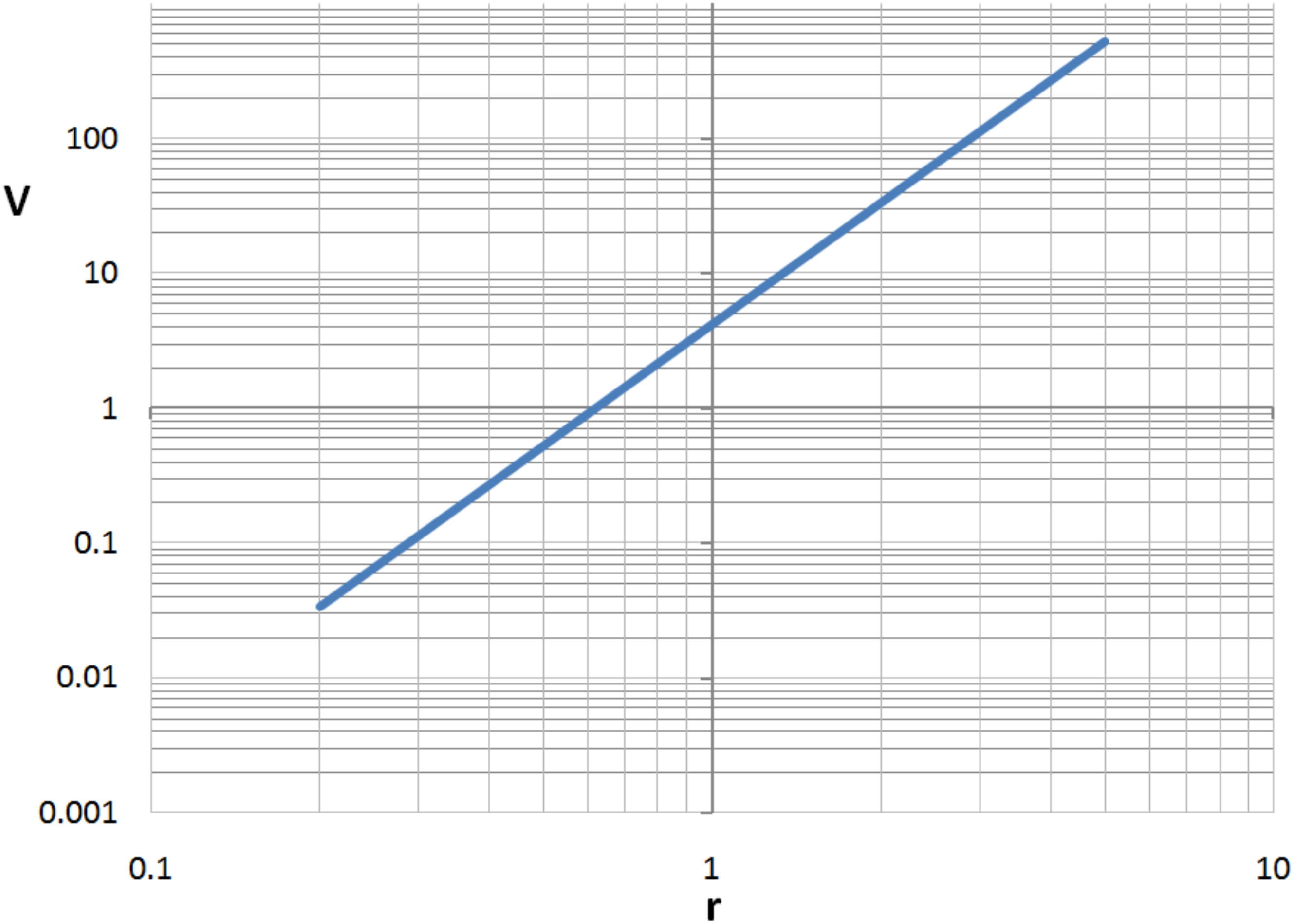
$$\log(V) = \log\left(\frac{4}{3}\pi r^3\right)$$

$$\log(v) = \log\left(\frac{4}{3}\pi\right) + \log(r^3)$$

$$\log(V) = 3 * \log(r) + \log\left(\frac{4}{3}\pi\right)$$

If we plot the log(y) *and* log(x), then we have the equation of a straight line

- This is called a *log-log* plot, since both axes are logarithmic


V
100
10
1
0.1
0.01
0.001
0.1
1
10
r