Understanding Experimental Data

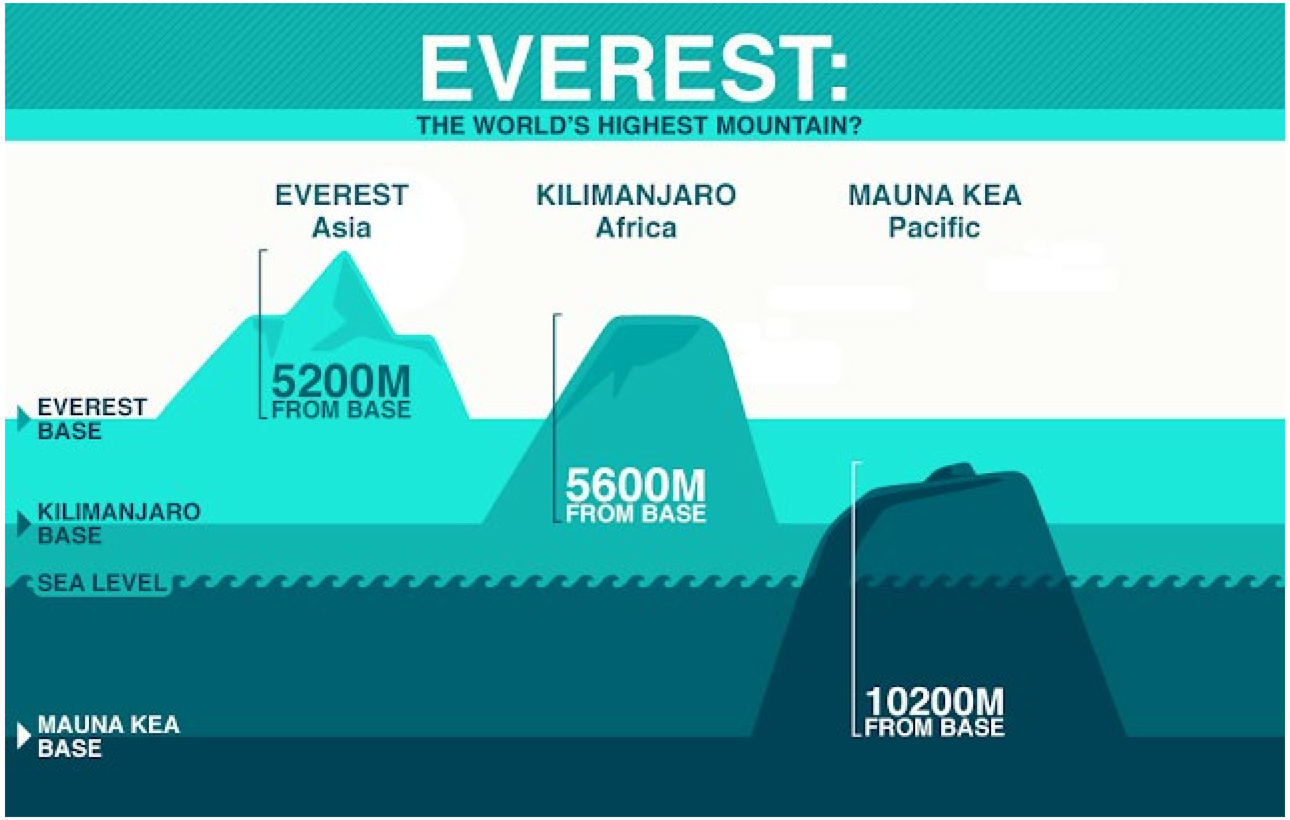
Chapter 20

Understanding Data

- Before we can utilize data for relations we must first understand the data itself
  - What are the characteristics of the data
  - Is it continuous or discrete?
  - Does it cluster or continue to trend?
  - Do two or more variables seem to be related
  - What is the scale?
  - What is the baseline?
  - Visualization is important!

What is the tallest mountain in the world?

- How do you measure that?



http://www.geologyin.com/2017/07/mount-everest-is-not-tallest-mountain.html

Experiment #1 – Spring has sprung

- In 1676 English physicist Robert Hooke formulated a *law of elasticity*
  - *Ut tensio, sic vis*
  - As the tension, so the extension!
  - F = -kx
  - *k* is **Spring constant**
  - Unit the spring reaches an **elastic limit**

Determining *k*

- Not a universal constant
- Constant for a spring
  - Slinky and motorcycle springs have different *k*s
  - All Slinkies have the same *k*

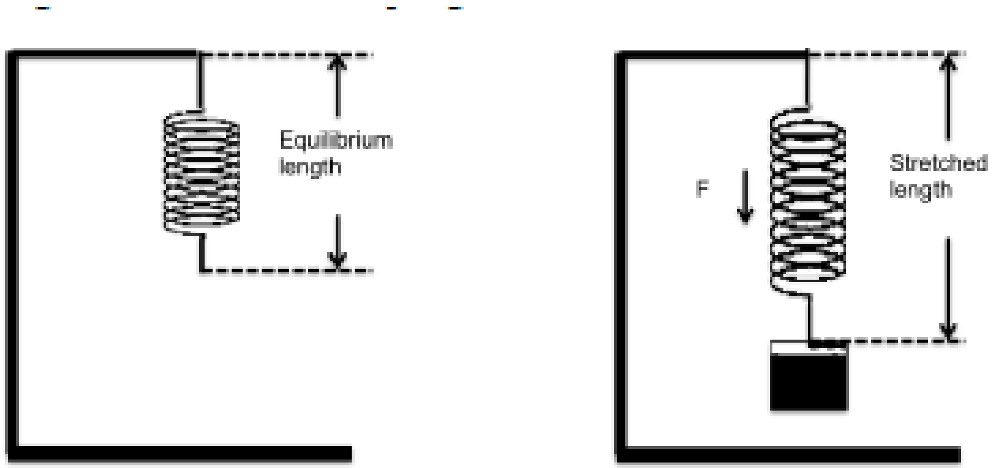
- *k* must be calculated for a spring

Taking measurements

- Test a variety of inputs
- Measure the outcomes

Calculating the Spring Constant (*k*)

- $F = -kx$
- $k = -\frac{F}{x}$
  - $F = ma$
  - $a = g = \frac{9.81m}{s^2}$



- $k = -\frac{m * {}^{9.81m}/_{s^2}}{x}$
  - m is the mass hung on the spring
  - x is the distance traveled

Linear regression

- Hooke’s Law tells us that Force and Mass are linearly related
- We should be able to **fit** a straight line to the data
- Define an **objective function**
  - Determine the “goodness” of fit
  - Minimize the error (optimization problem – chapter 14)

Least Squares

- Most commonly used objective function
- $\sum_{i=1}^{len(observed)-1}(observed[i] - predicted[i])^2$

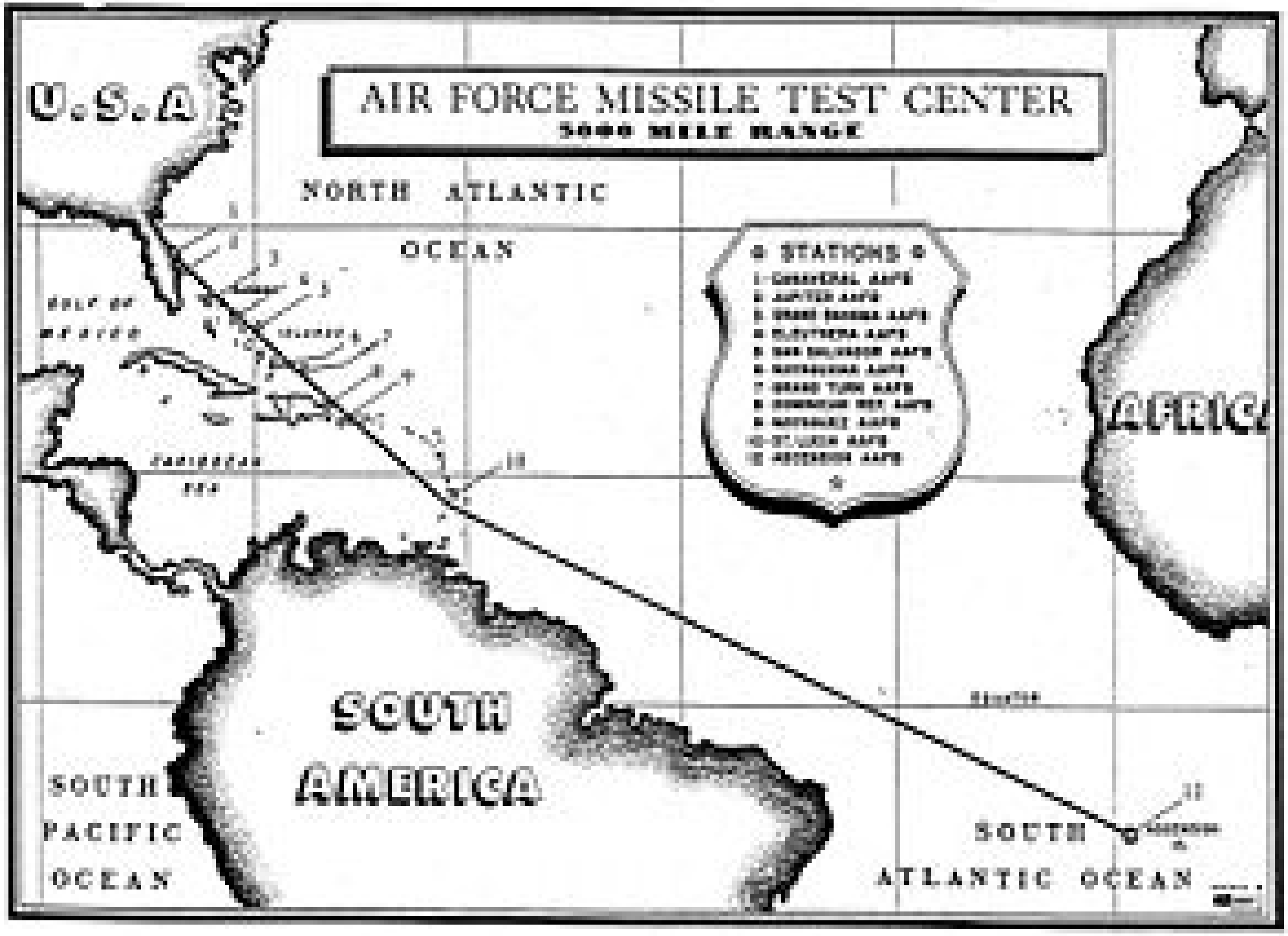
- We need the same number of predictions and observations
  - Squaring emphasizes large differences
  - Squaring washes out sign differences
- How do we make the predictions?

Numpy and linear regression

- **polyfit()** performs our curve fitting
  - Arrays for x and y values
  - In our case, measured distances and calculated forces
  - The degree of the polynomial
  - Returns coefficients

- **polyval()** then applies a fit to a set of values
  - Coefficients
  - Set of values
  - Returns predicted values

Experiment #2 – Behavior of Projectiles



Which model fits better

- Fit gives us a function that relates an independent variable to a dependent variable
- What were they in our examples?

- **Goodness of fit** tells us how good our predictions are
- Linear regression minimizes Mean Square Error (MSE)

```
Linear MSE:    616.381680147
Quadratic MSE: 8.93245704374
```

- Is 8.932 good?

Coefficient of Determination R^2

- We need to *normalize* MSE
- $R^2 = 1 - \frac{\sum_i (y_i - p_i)^2}{\sum_i (y_i - \mu)^2}$
    - Specific to linear regression

- $R^2 = 1$ gives a perfect fit
- $R^2 = 0$ means there is absolutely no fit

Using a computational model

- How fast is our projectile moving when it hits the ground?
  - Distance is predicted by $ax^2 + bx + c$
  - Maximum height (peak altitude) is reached at the midpoint
    - At that time momentum is overcome by gravity
    - So, at the midpoint the projectile falls with an acceleration of $9.81\ {}^{m}/_{s^2}$
    - Peak altitude is $aMidPoint^2 + bMidPoint + c$
  - Time to fall from peak altitude $\sqrt{{}^{(2\ *peak\ altitude)}/_{g}}$
  - Vertical velocity at time of impact = g*t
  - Horizontal velocity = t / distance

Fitting exponential Data

- Discussed in code review

When we can’t measure

- Collect data
- Divide onto training and test sets

Summary

- We can use linear regression to fit a curve to data
  - Mapping a dependent variable to an independent variable
- The fit can be used to predict additional values
- R^2 can be used to evaluate a model
- We need to be cautious about overfitting the data
  - Remember the system we are modeling
- Choose complexity based on
  - Theory about the data
  - Cross validation
  - KISS