Noticing

Instructor: Rob Faludi
Class 2

Plan for Today

- Mysteries Presentations
- Data Exercise
- Scrappy Science, ethics
- EDA
- Exercise: Coins
- Exercise: Data Observation
- Readings & Assignments

Exercise One: Data

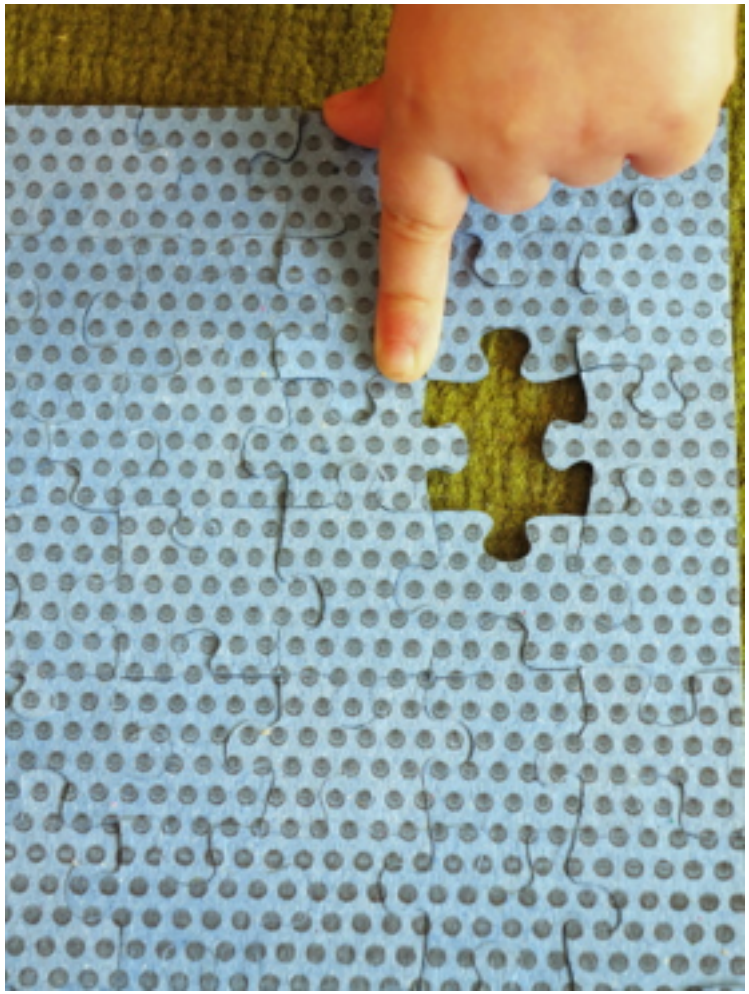
- Get some data!
  - You have 10 minutes to collect some data and return with it
  - Use your own definitions of data, we'll talk about that after your return

Mysteries Presentations

- Where you went
- What you found and what it might mean

Data Gathering Results

- What is your data?
- How did you get it?
- What does it tell and what does it conceal?
- What questions does it raise?
- Other data that might help?
- Anything that might give you more insight?



Scrappy Science

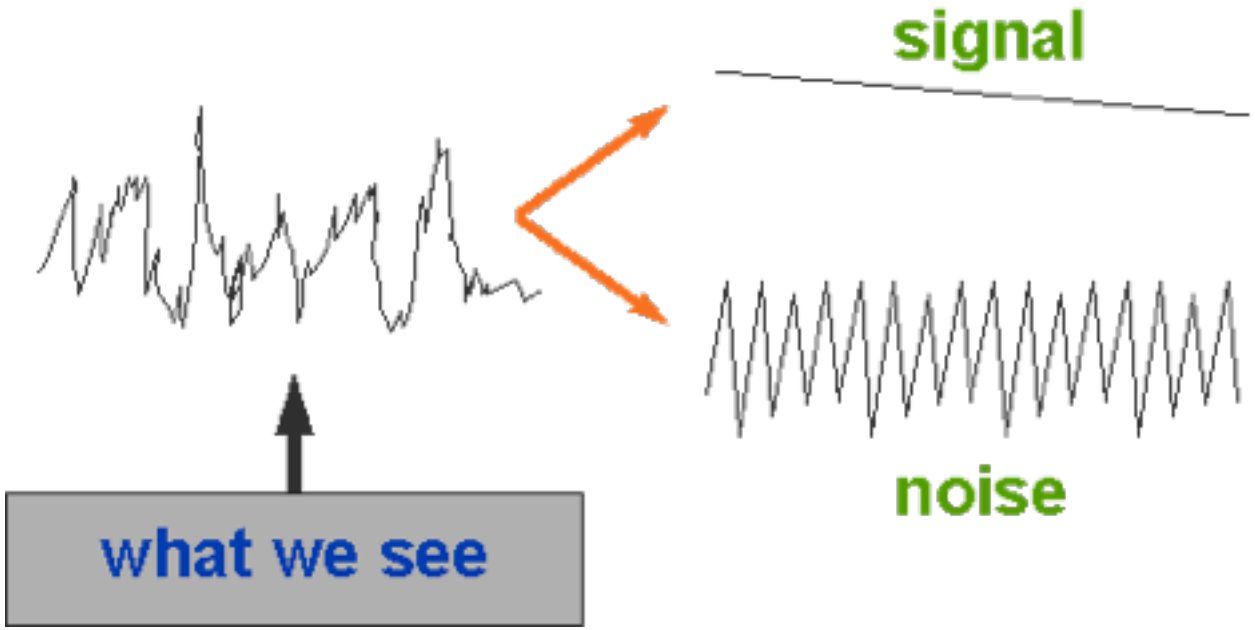
Seeing with Data

- Unmasking the invisible
- Manipulating time, space
- Comparisons across variables
- Thinking mode

Signal vs Noise

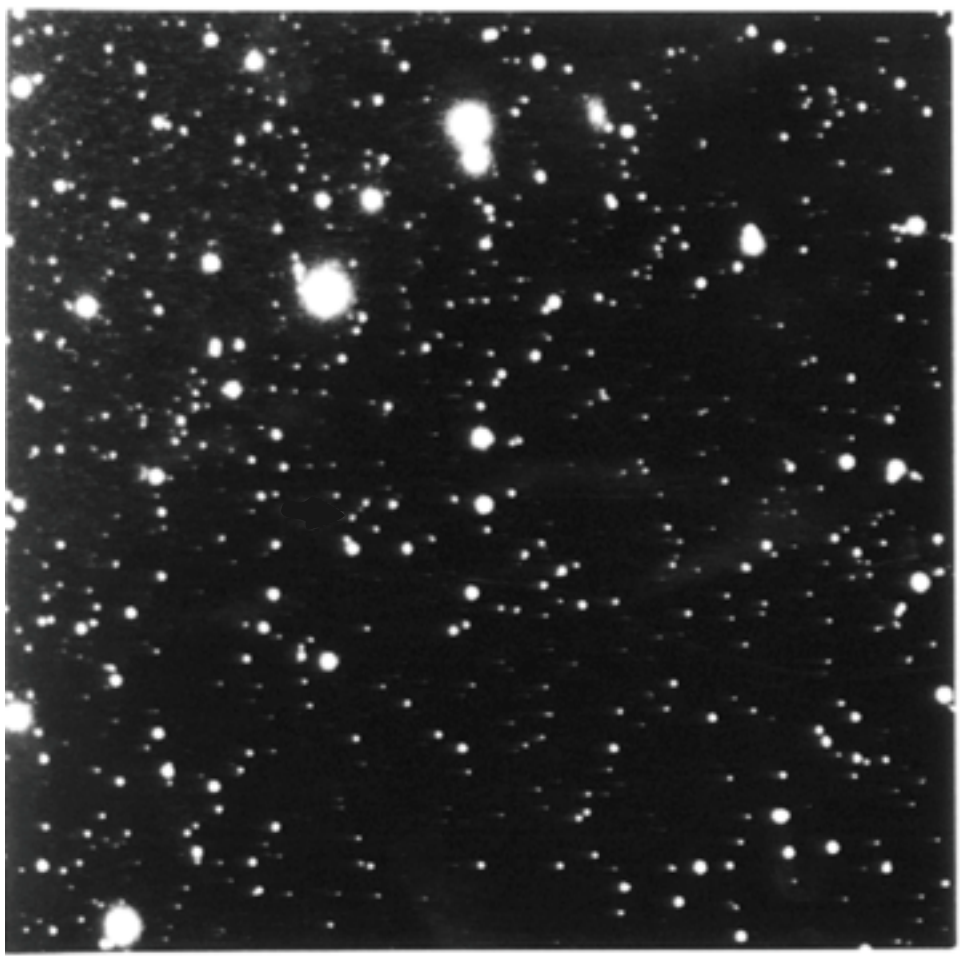
- What is signal?
- What is noise?
- When does that shift?

What we observe can be divided into:

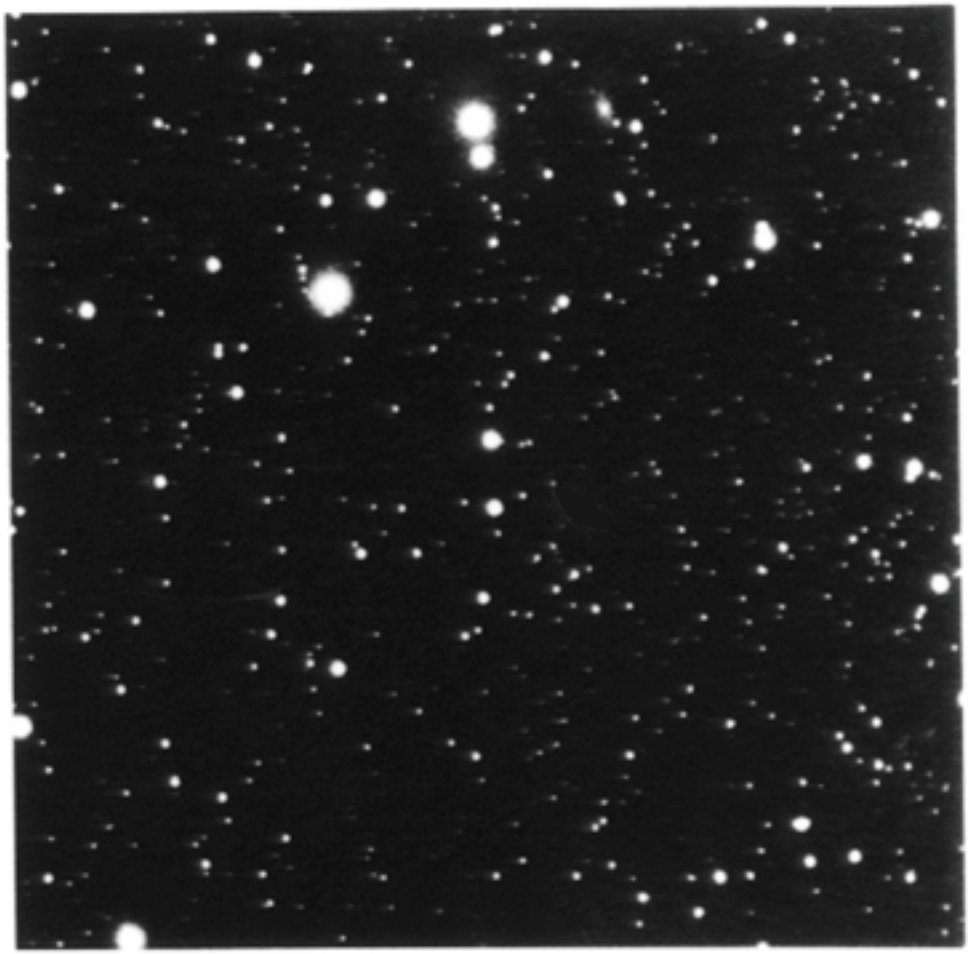


Mostly Noise

DISCOVERY OF THE PLANET PLUTO



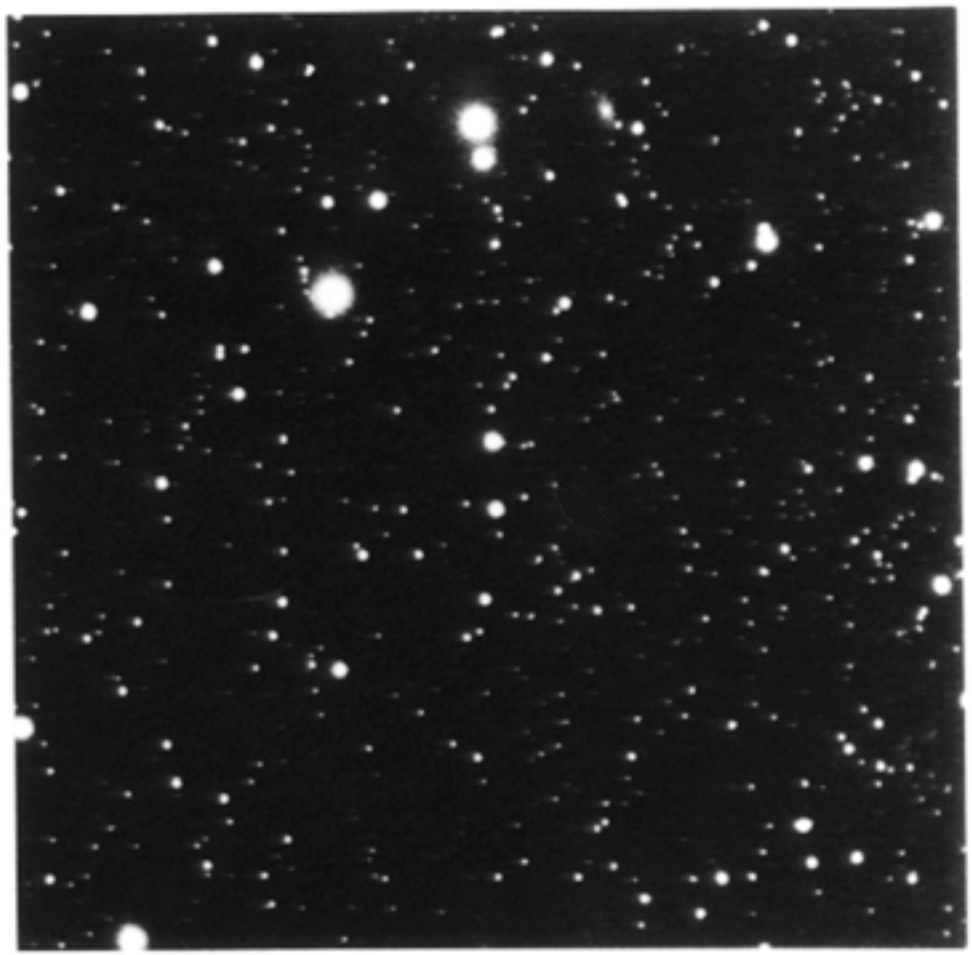
January 23, 1930



January 29, 1930

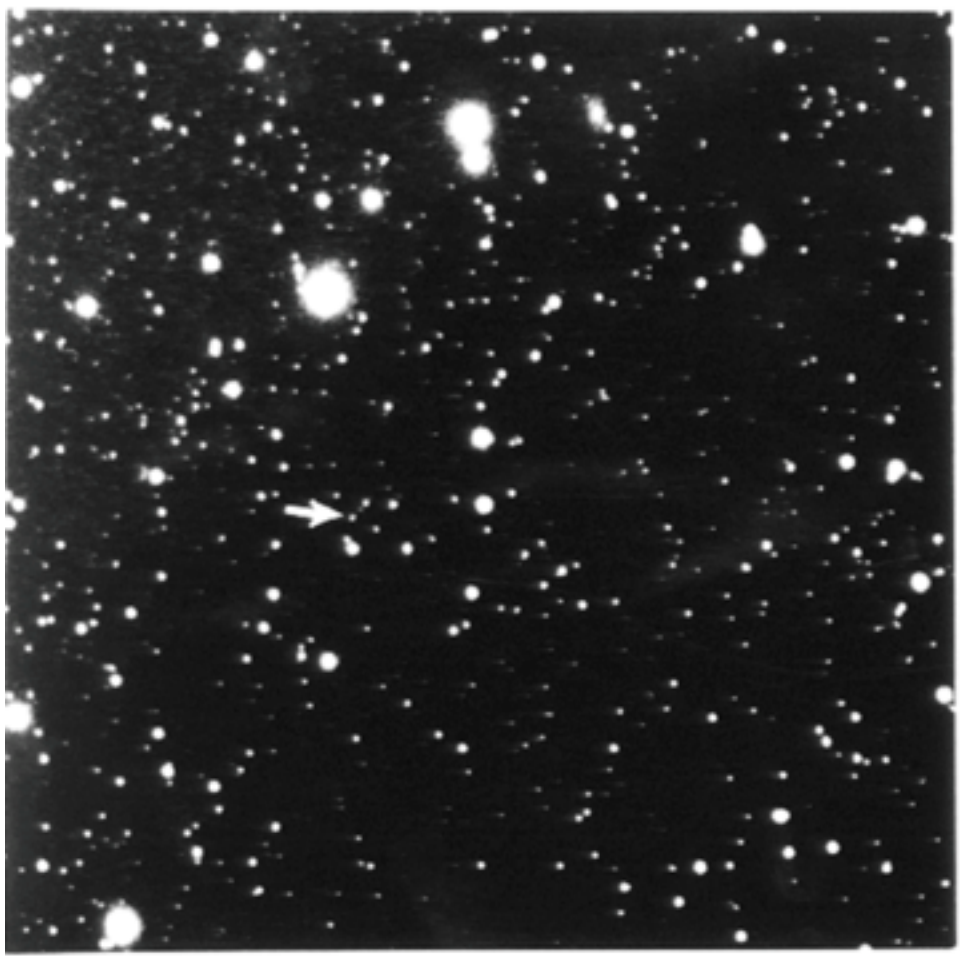
Compare

Compare



Signal

DISCOVERY OF THE PLANET PLUTO

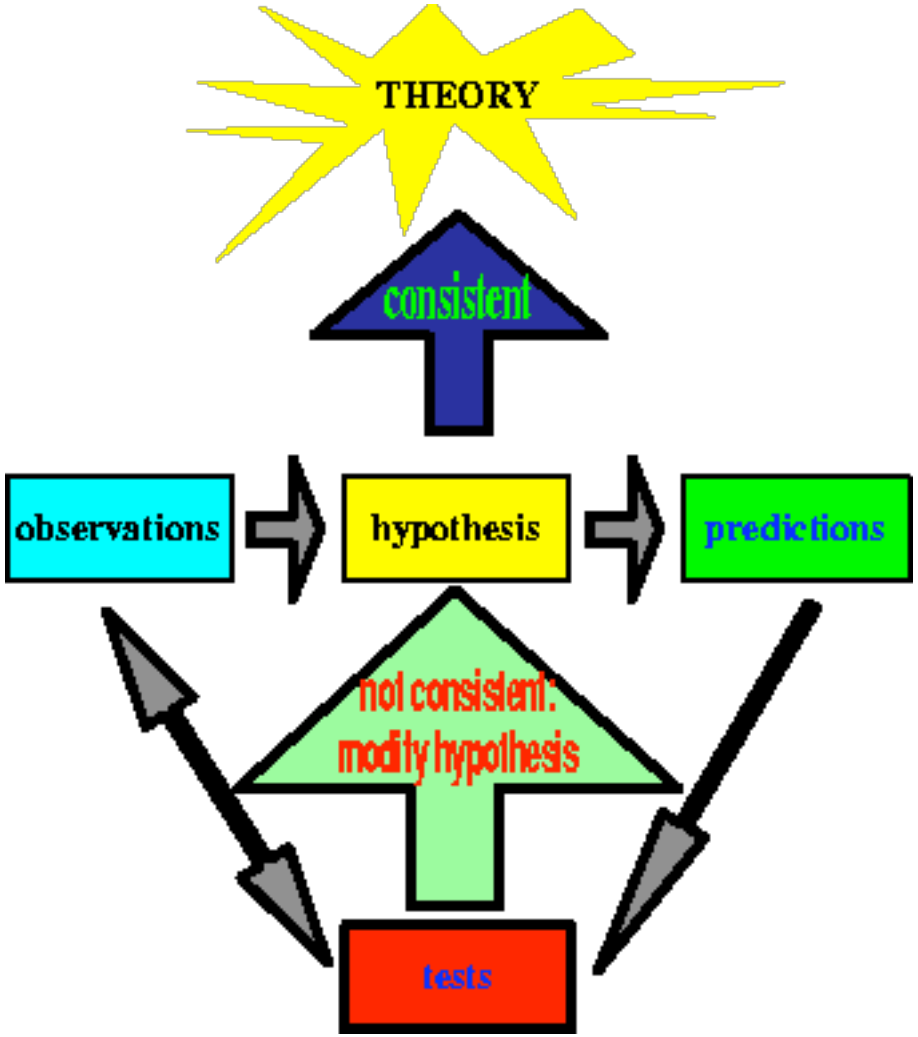


January 23, 1930

January 29, 1930

Scientific Method

1. Observation
2. Hypothesis
3. Predictions
4. Experimental Testing
5. ...repeat as necessary



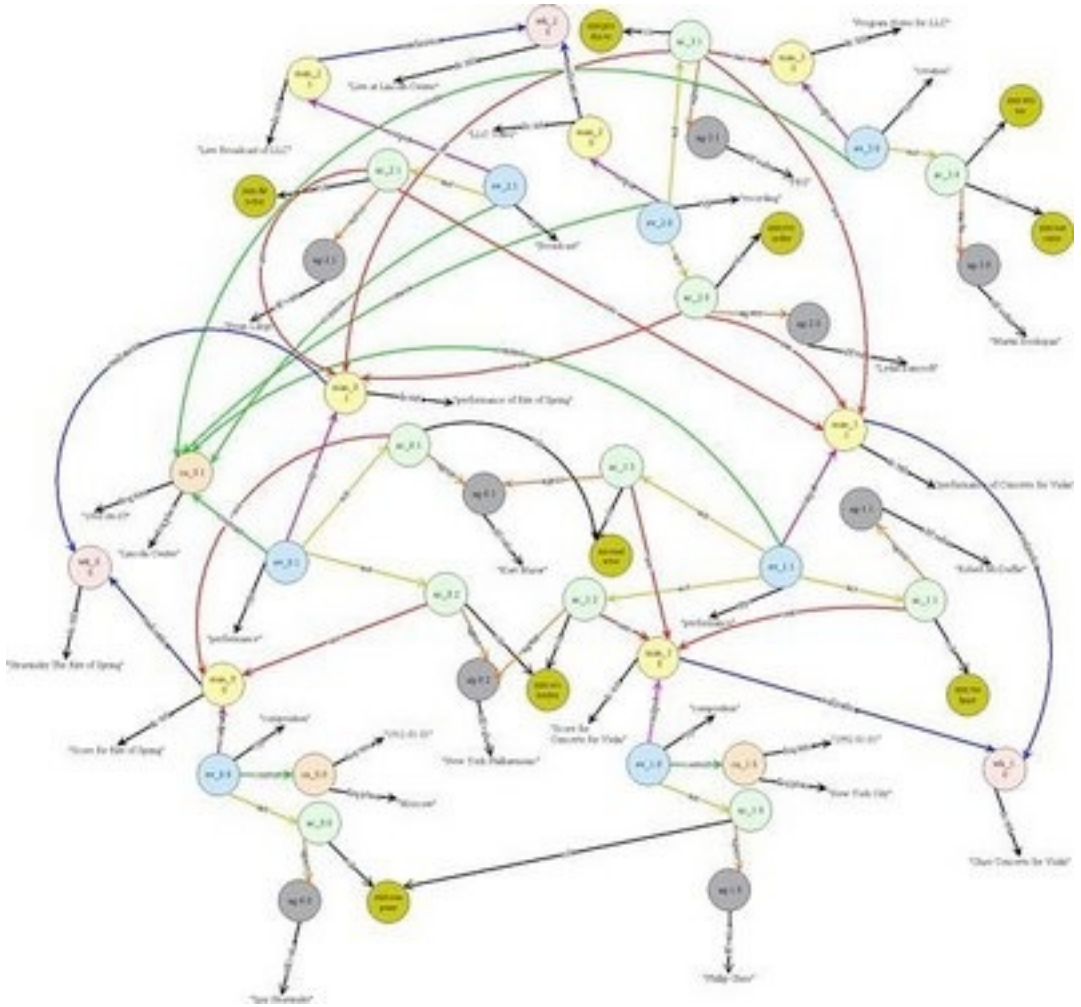
- results can be replicated
- hypotheses can be falsified

Scientific Method...also

- document your methods
- randomize your sample
- try some pilot studies
- get some experts to review your work
- get some peers to review your findings
- disclose your methods
- share your critical thinking about projects (not people)

There’s more

- Reading
- Discussion
- Fantasies
- Procrastination
- Confusion
- Dejection
- Excuses
- Paranoia
- Excitement
- Overcommitment
- Elation
- Debate
- Boredom
- Construction
- Selling
- Resolution



Statistics are Awesome

- Confirmations
- Predictions
- Definite statements
- Evidential focus
- Explorations
- Precision & rigor

Statistics Suck

- Summary masking
- Correlations
- Effect size issues
- Significance issues
- Proof issues
- Lies

Getty Image Search: “Bad Science”

HOW TO LIE WITH STATISTICS

Darrell Huff

Illustrated by Irving Geis

Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

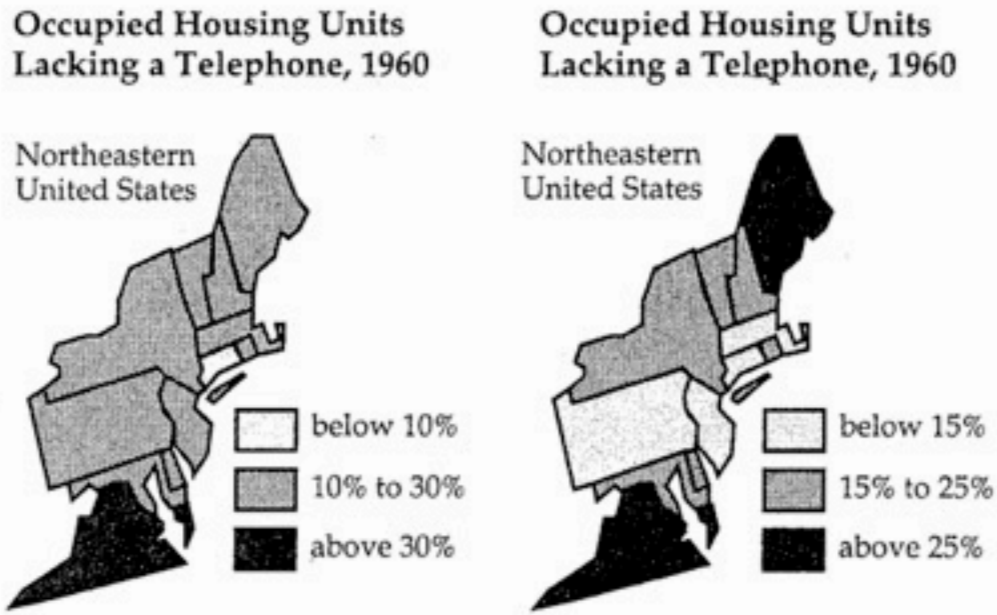


FIGURE 3.10. Different sets of class breaks applied to the same data yield different-looking choropleth maps.

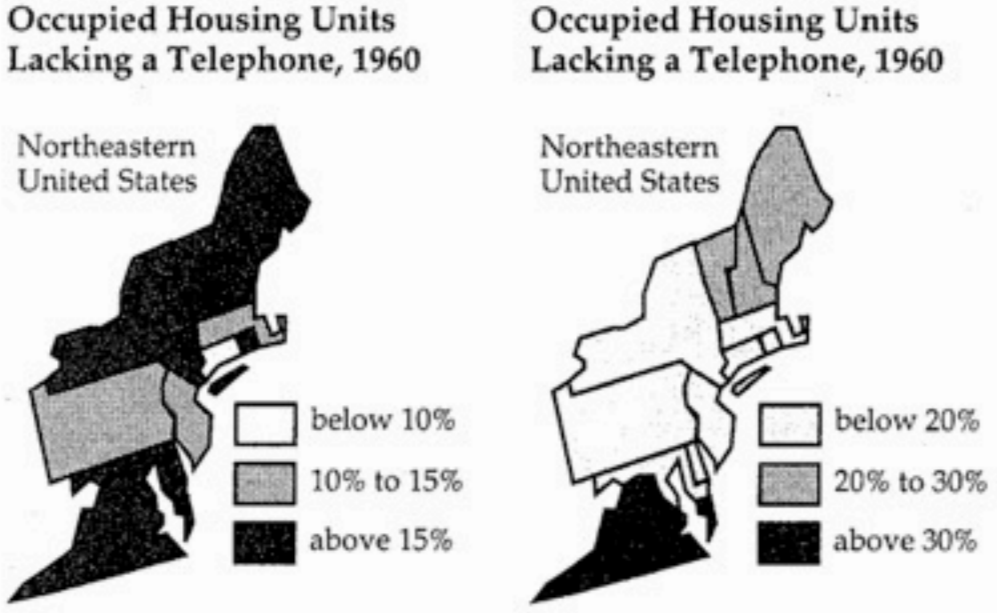


FIGURE 3.11. Class breaks can be manipulated to yield choropleth maps supporting politically divergent interpretations.

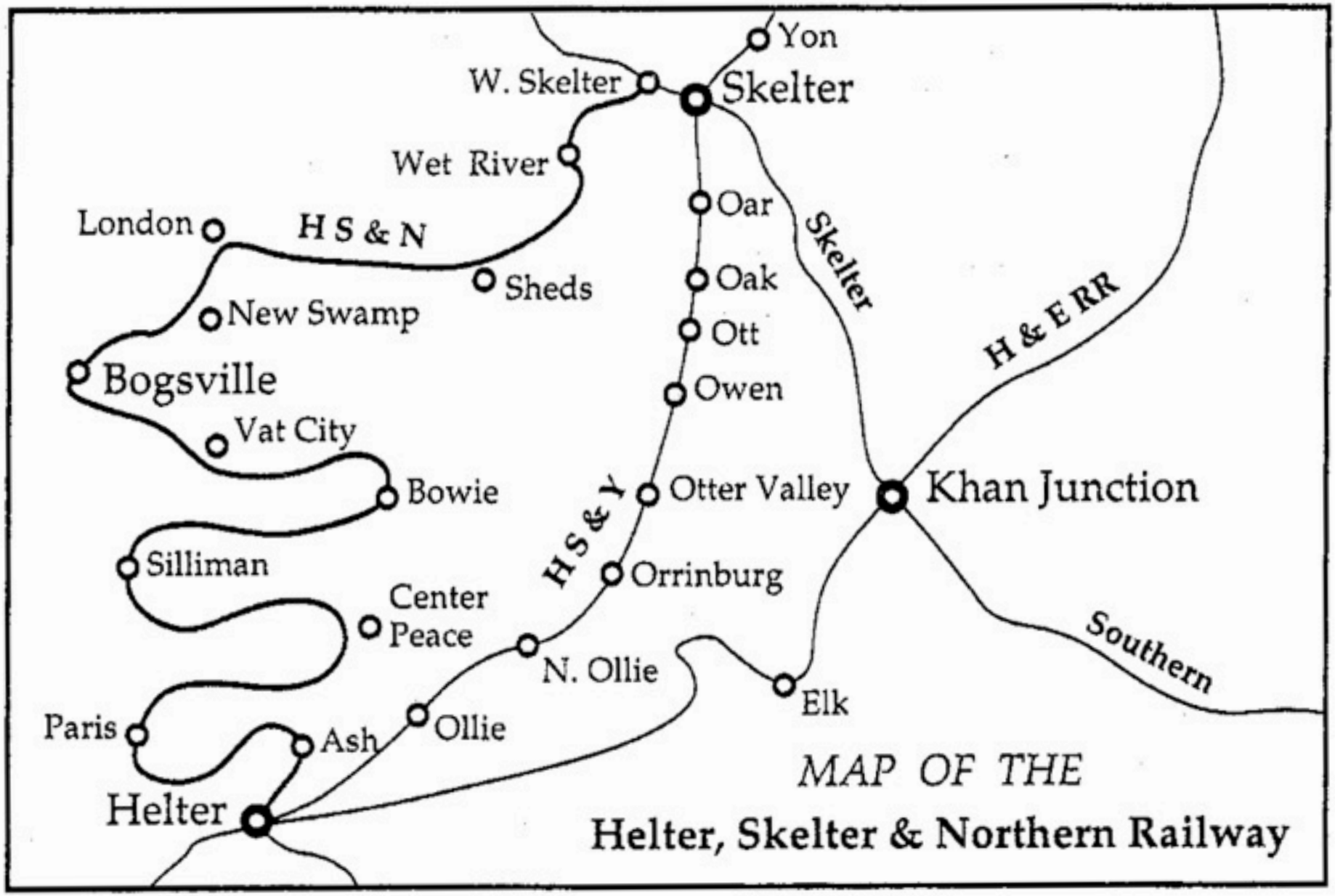


FIGURE 5.1. Engineering department's map of the Helter, Skelter and Northern Railway.

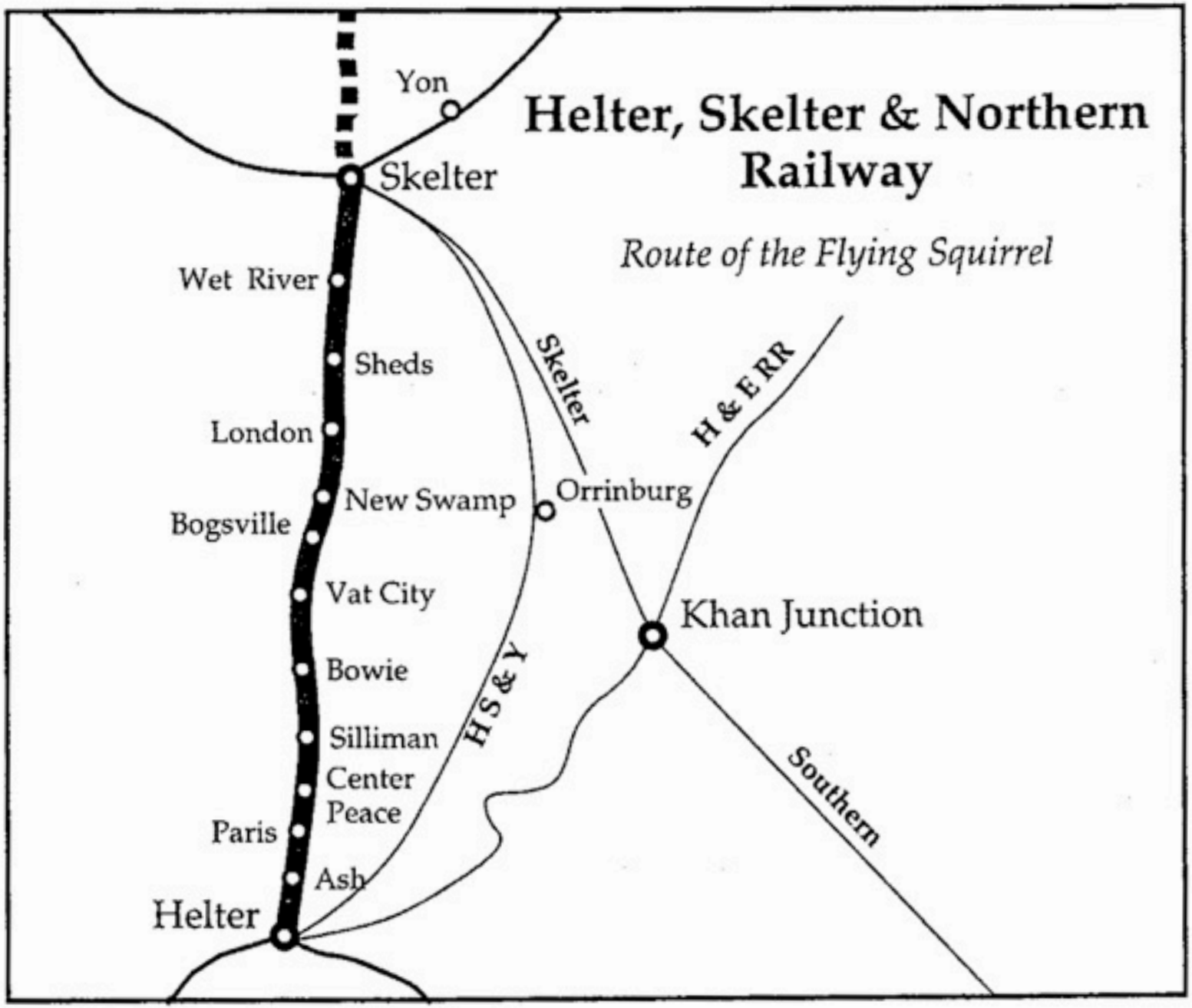


FIGURE 5.2. Advertising and timetable map of the Helter, Skelter and Northern Railway.

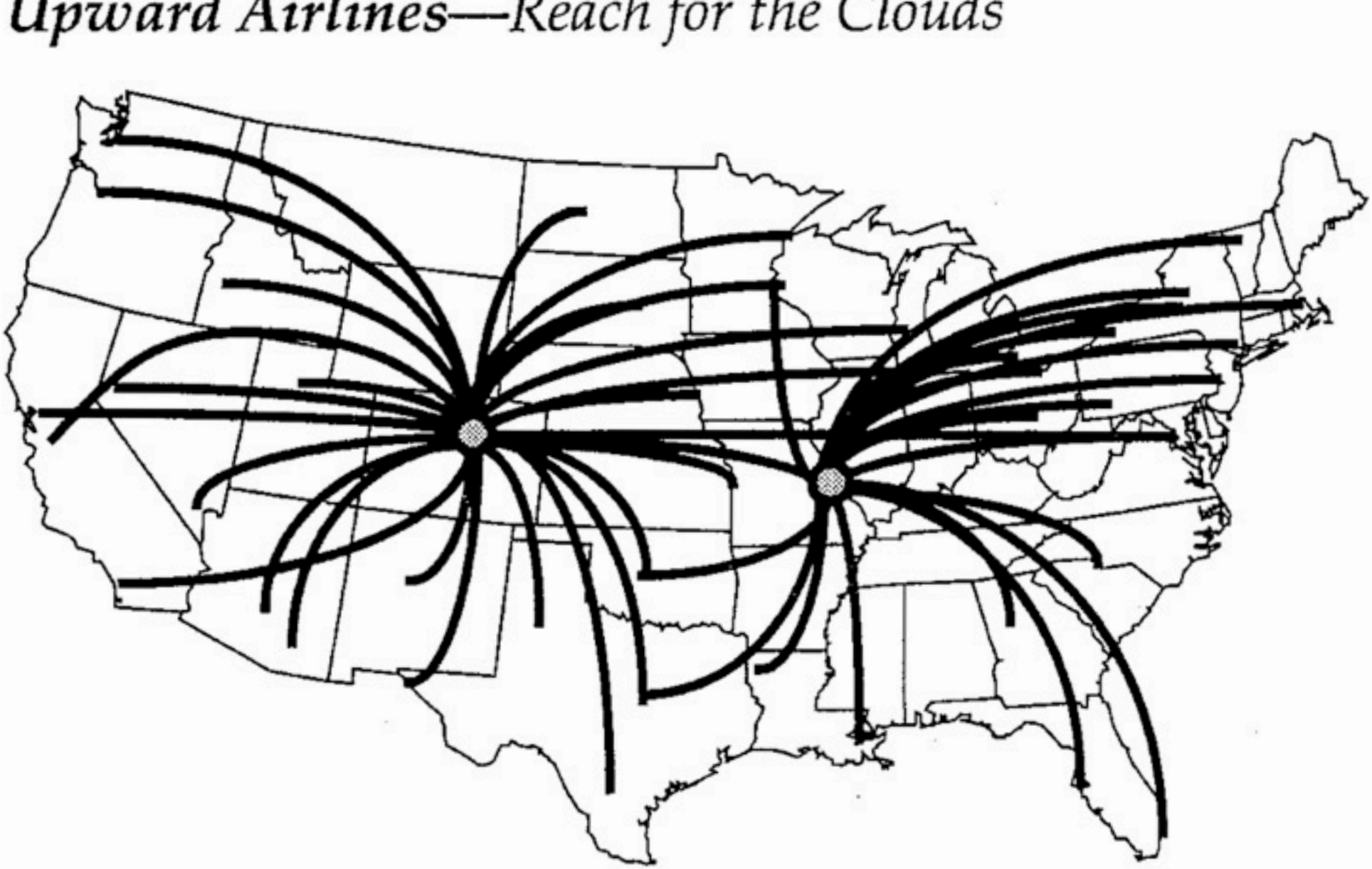
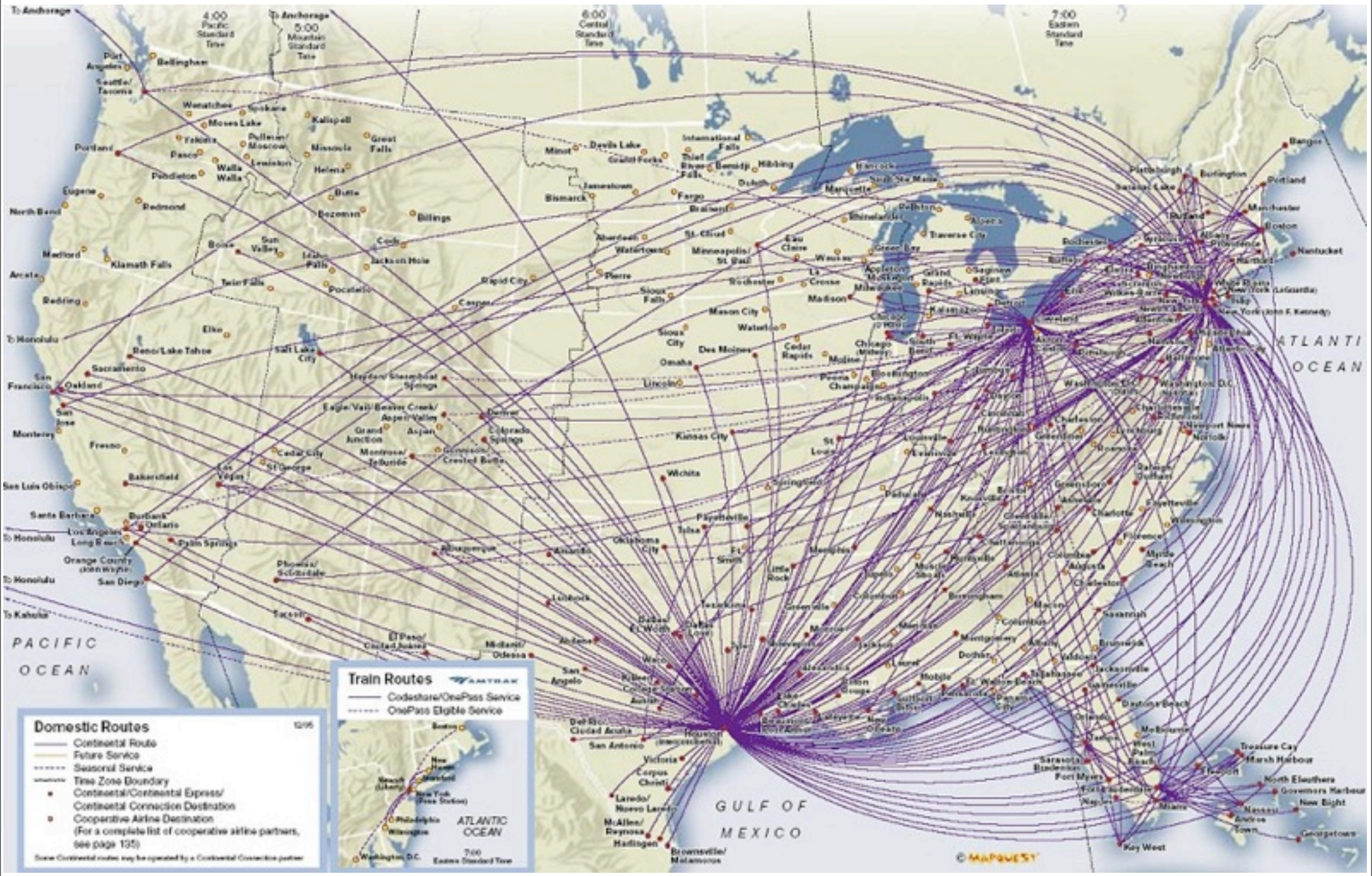
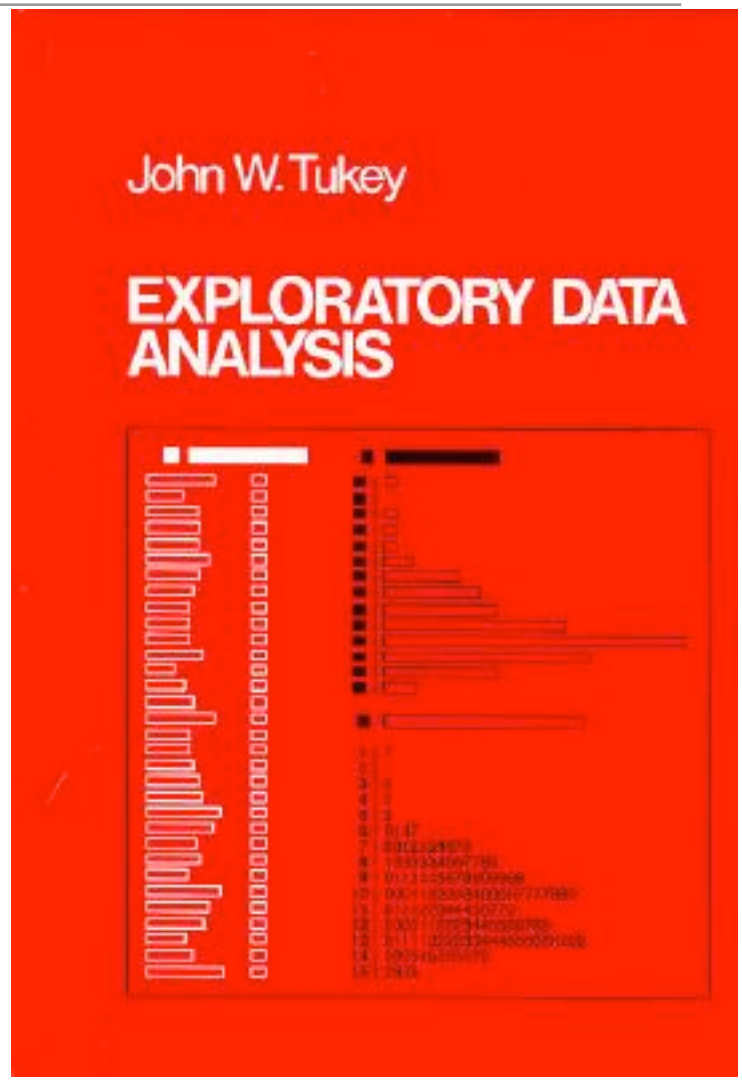


FIGURE 5.3. Advertising map emphasizing Upward Airlines' service area and connections.


4:00 Pacific Standard Time
5:00 Mountain Standard Time
6:00 Central Standard Time
7:00 Eastern Standard Time
PACIFIC OCEAN
ATLANTIC OCEAN
GULF OF MEXICO
Train Routes
Codeshare/OnePass Service
OnePass Eligible Service
Domestic Routes
Continental Route
Future Service
Seasonal Service
Time Zone Boundary
Continental/Continental Express/ Continental Connection Destination
Cooperative Airline Destination
(For a complete list of cooperative airline partners, see page 135)
ATLANTIC OCEAN
© MAPQUEST

Exploratory Data Analysis

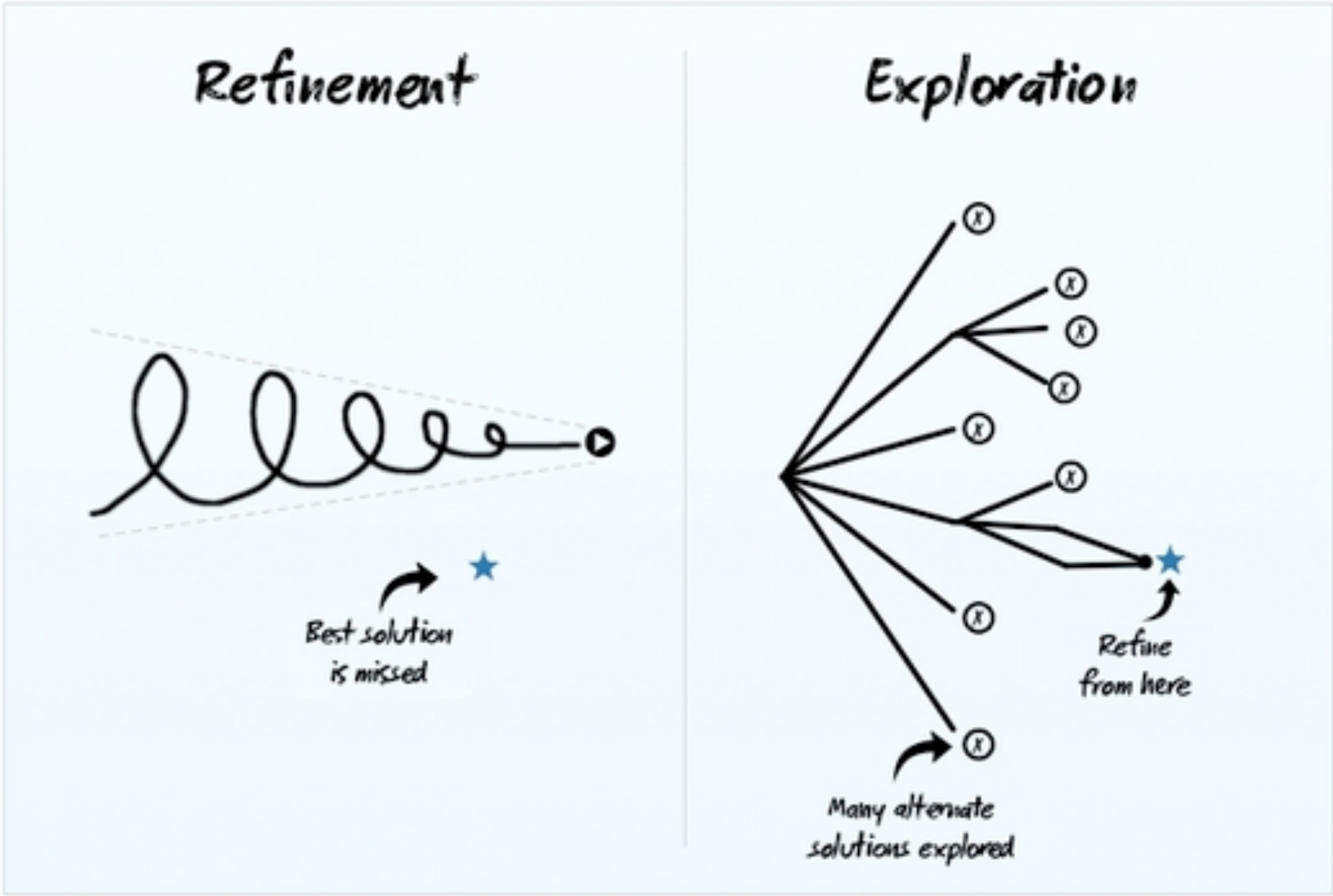
- many wells, many routes
- state of mind
- data as a source of new ideas
- skepticism for summaries
- openness to new patterns and views
- not by statistical summary alone, that's not open



EDA

- Smooth vs. Rough
- inductive methods
- re-expression, scales
- visualization
- resistant or robust statistics
- outliers, include or exclude?

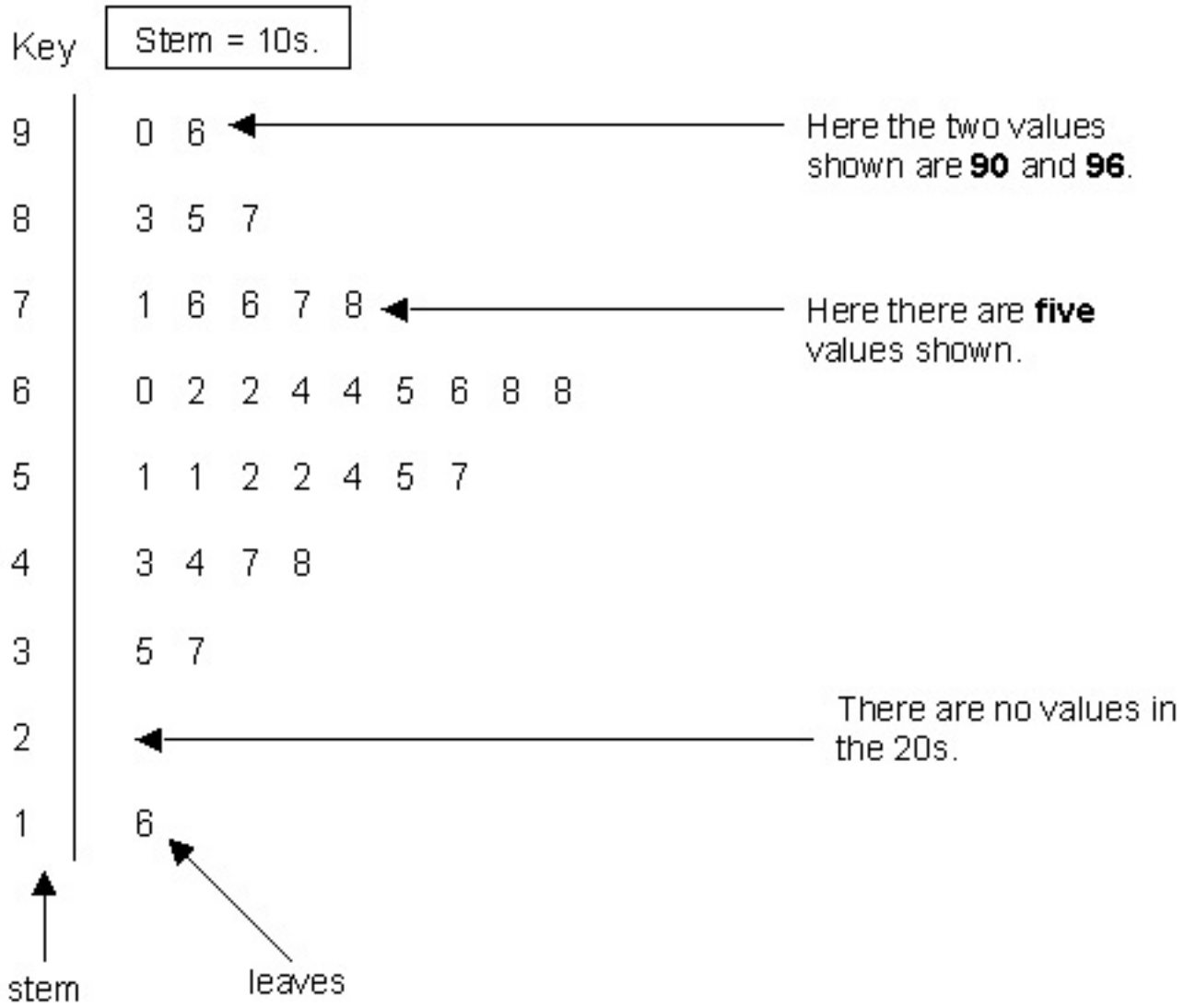
Why Explore?


Refinement
Best solution is missed
Exploration
Refine from here
Many alternate solutions explored

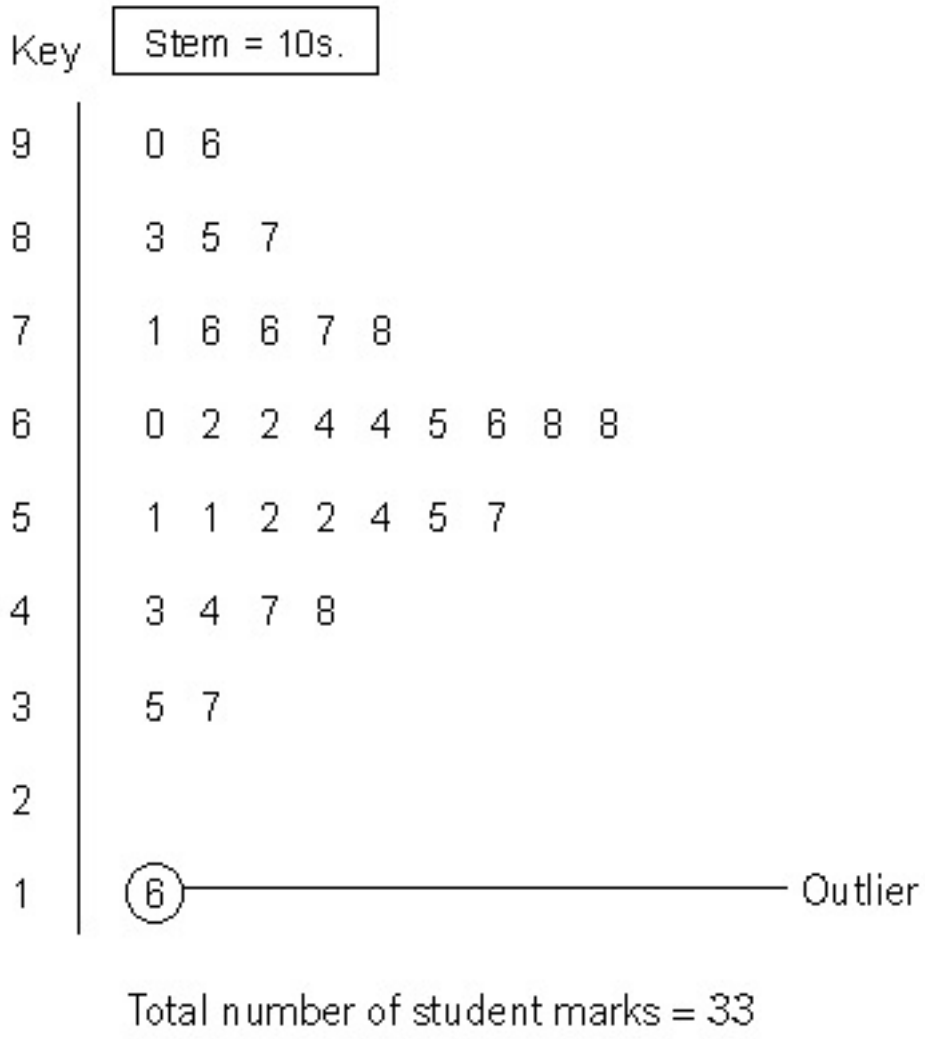
Visualization: Stem & Leaf

Key	Stem = 10s.
9	0 6
8	3 5 7
7	1 6 6 7 8
6	0 2 2 4 4 5 6 8 8
5	1 1 2 2 4 5 7
4	3 4 7 8
3	5 7
2	
1	6

Visualization: Stem & Leaf

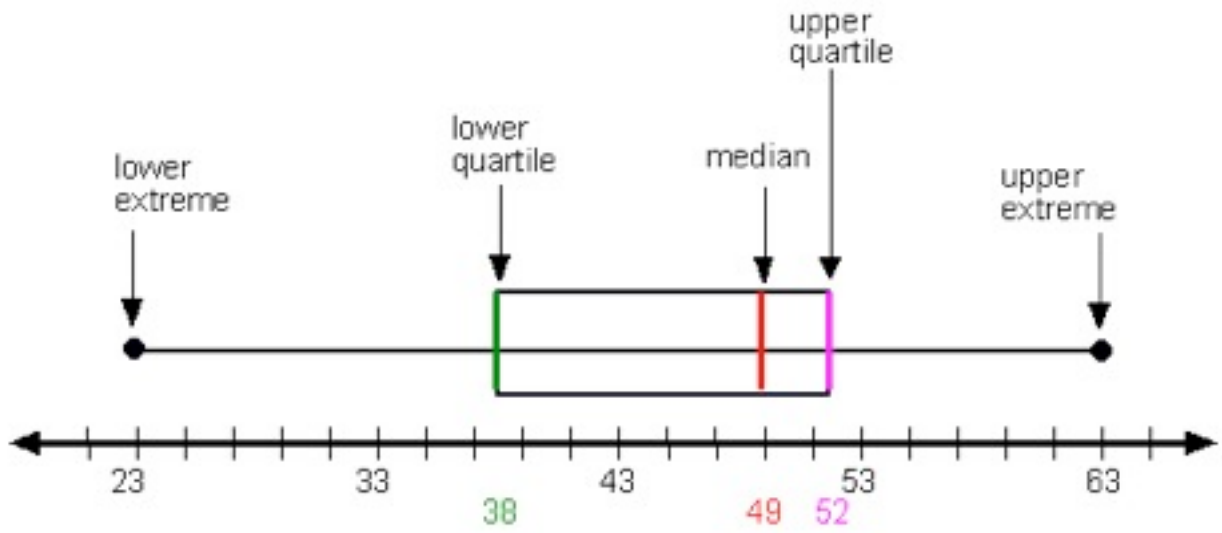


Visualization: Stem & Leaf

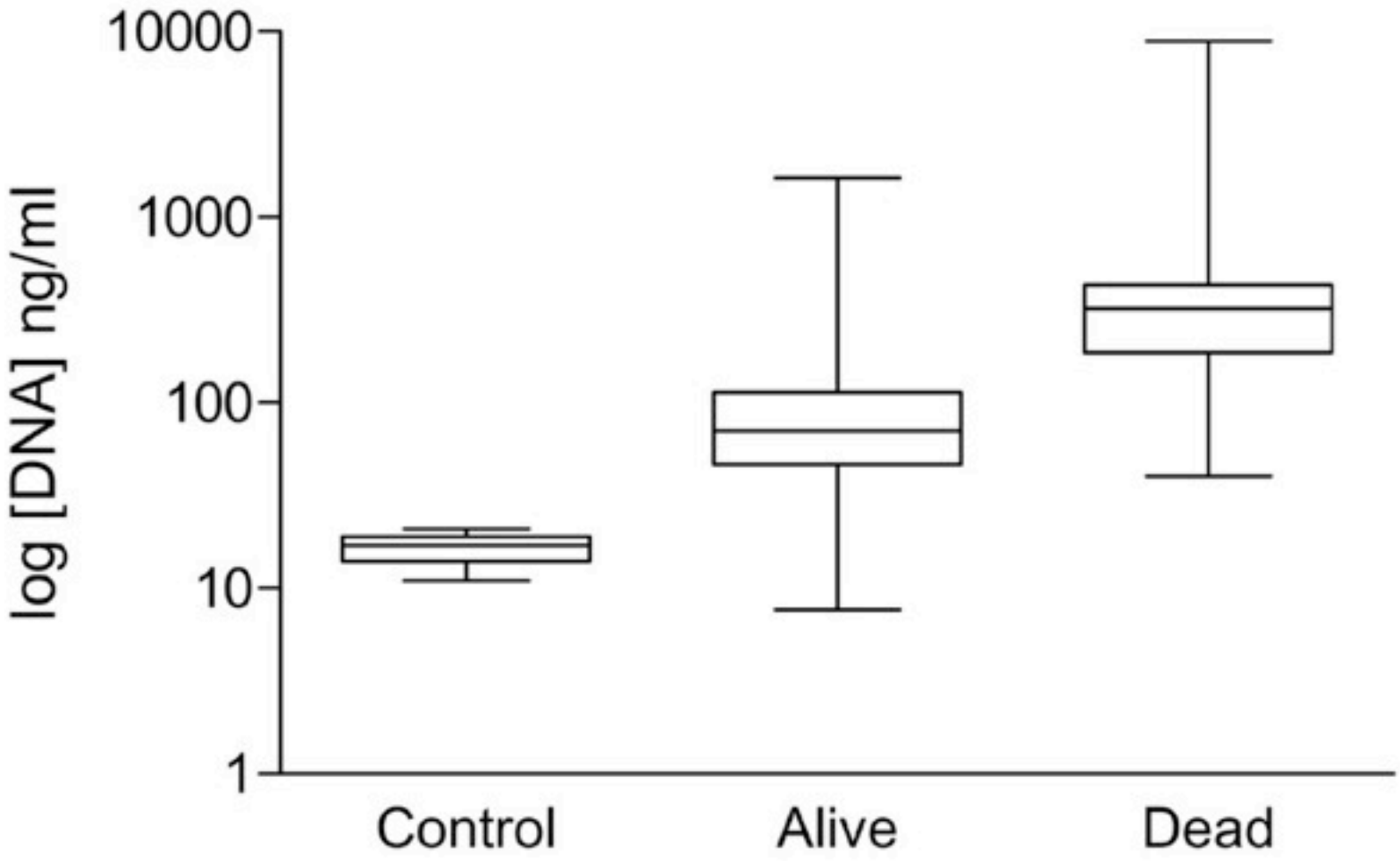


Total number of student marks = 33

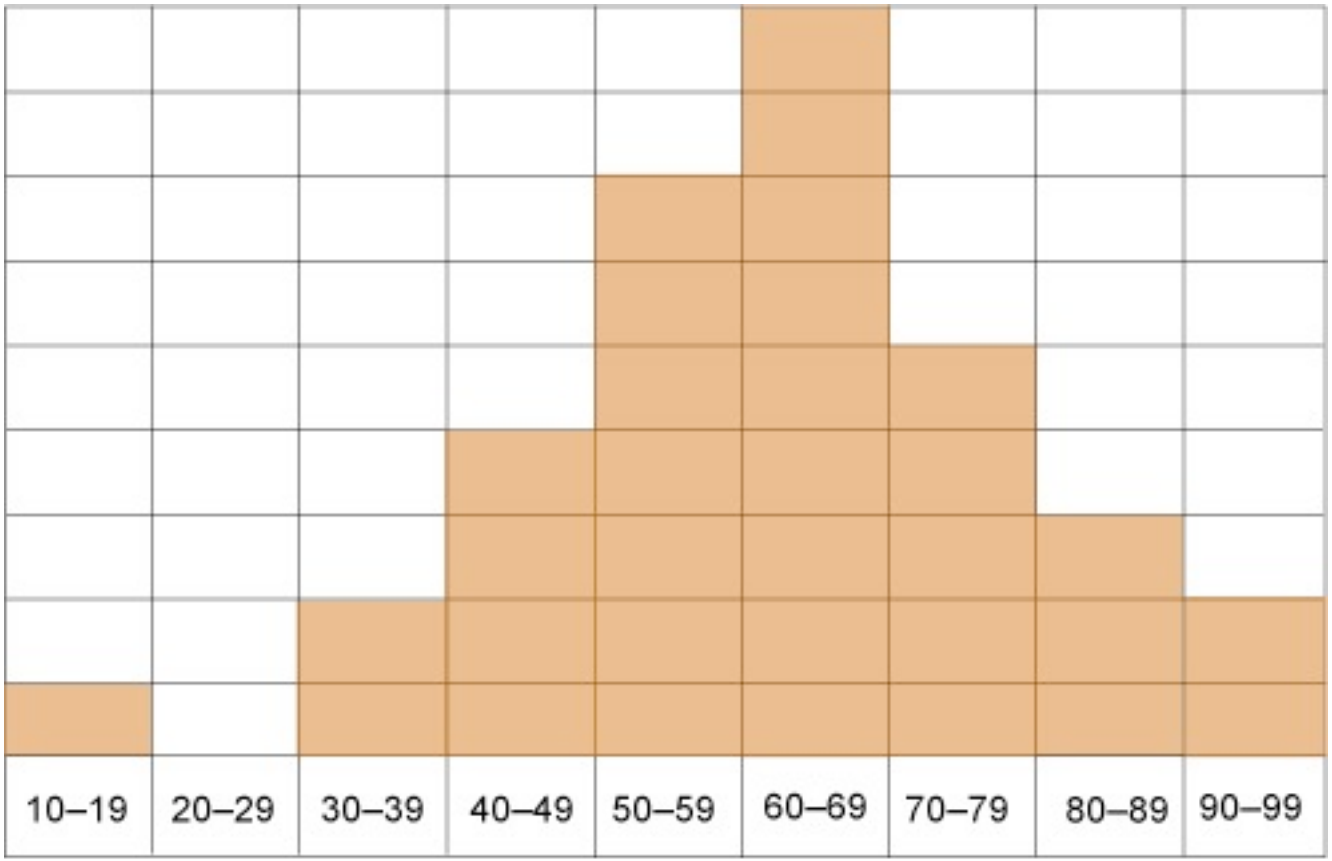
Visualization: Box & Whiskers



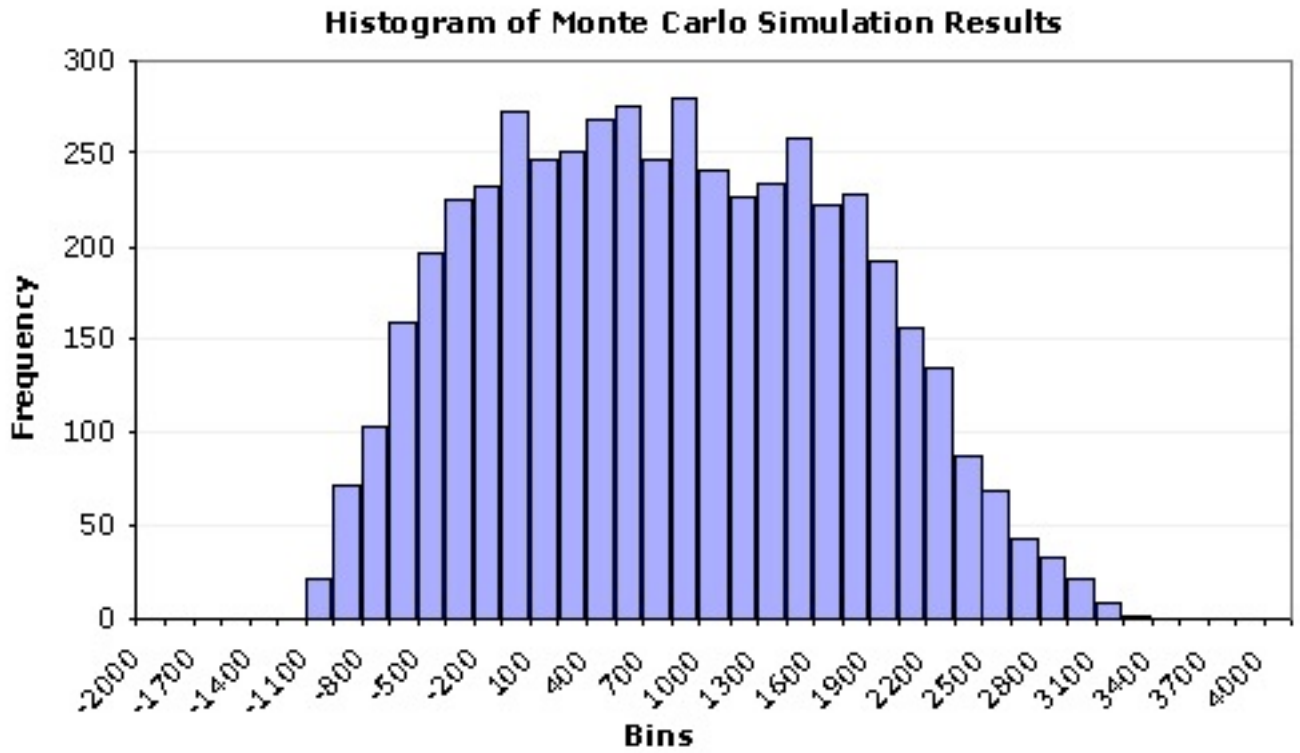
Visualization: Box & Whiskers



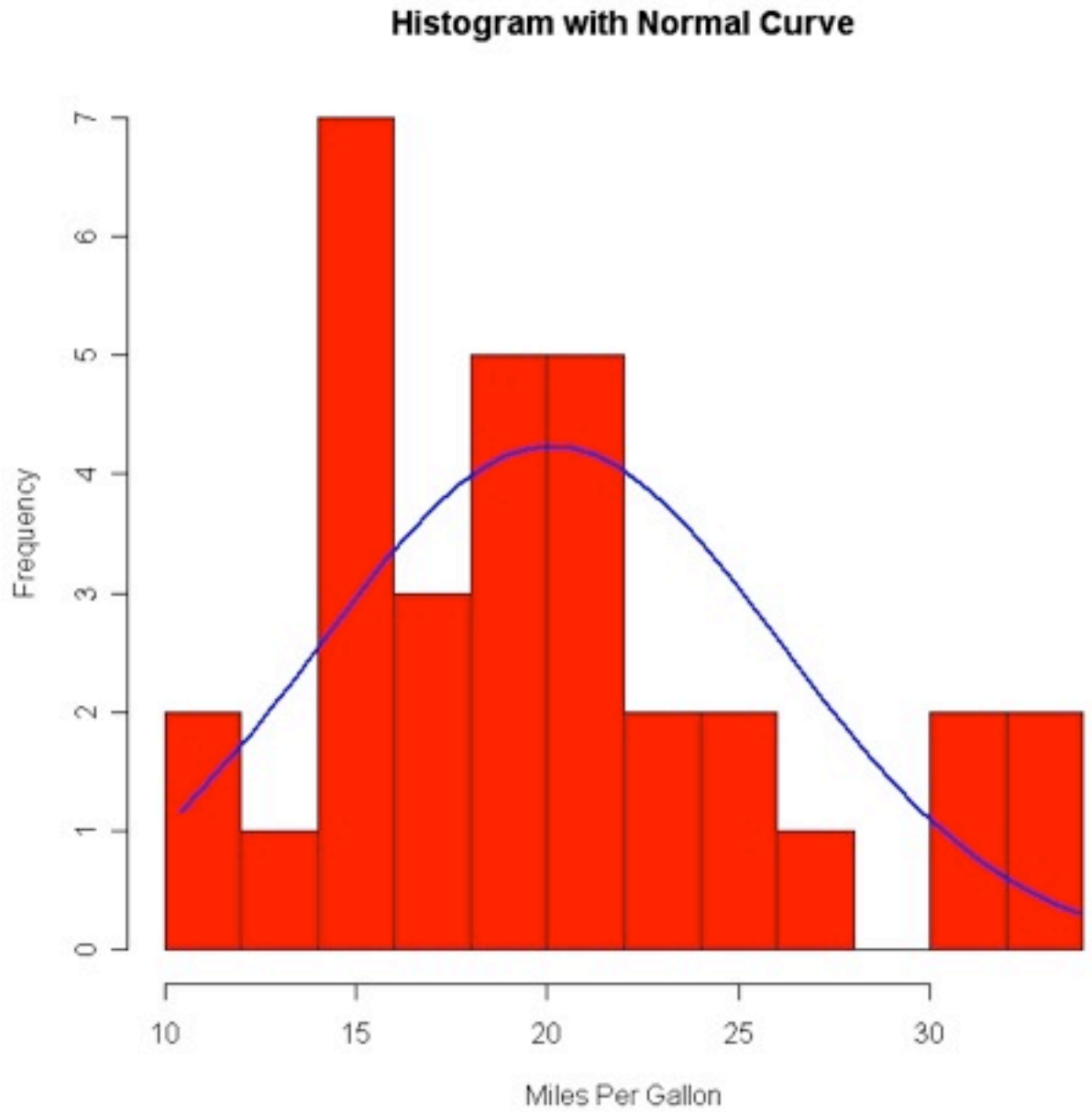
Visualization: Histograms



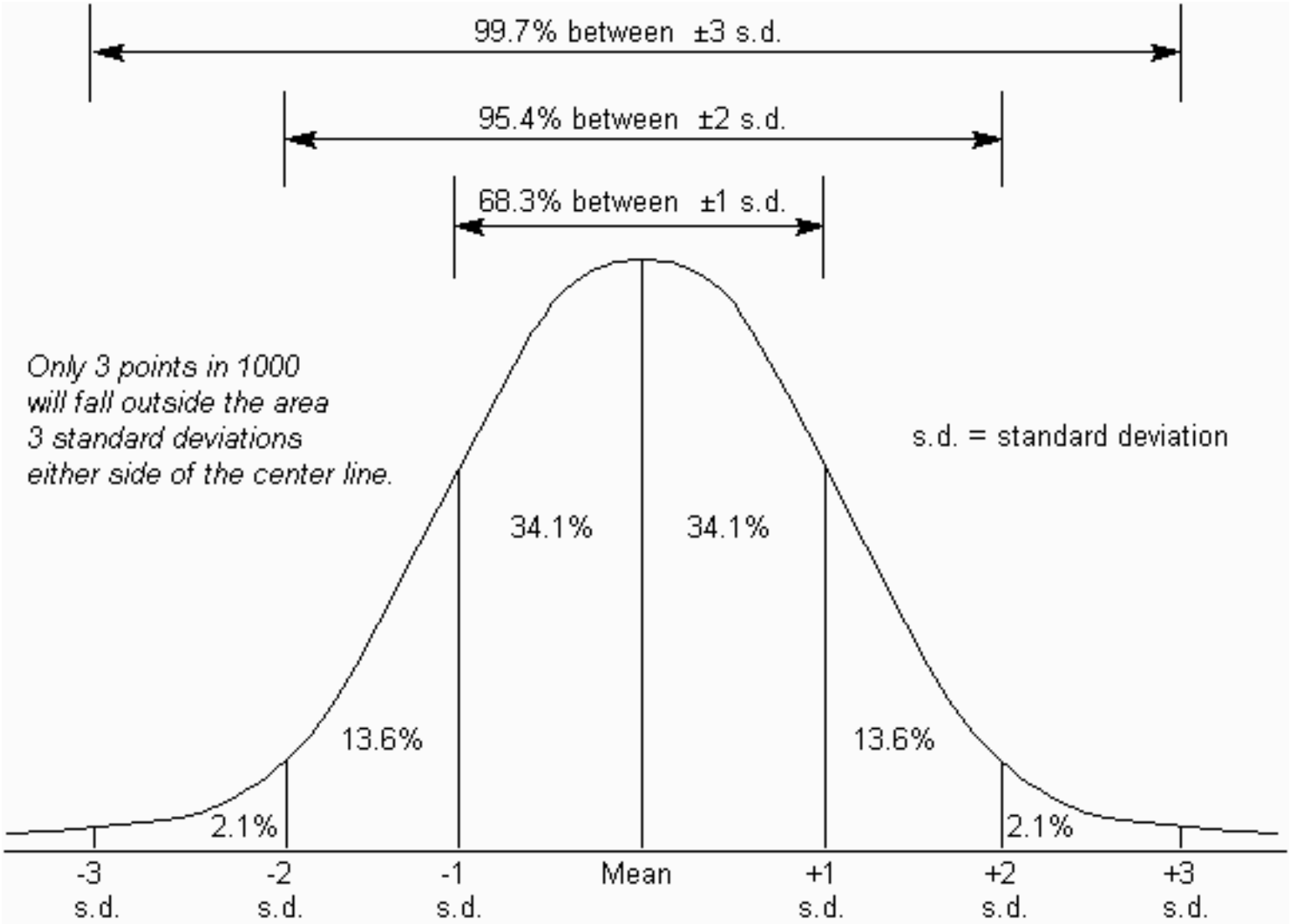
Visualization: Histograms



Visualization: Histograms



Parameters

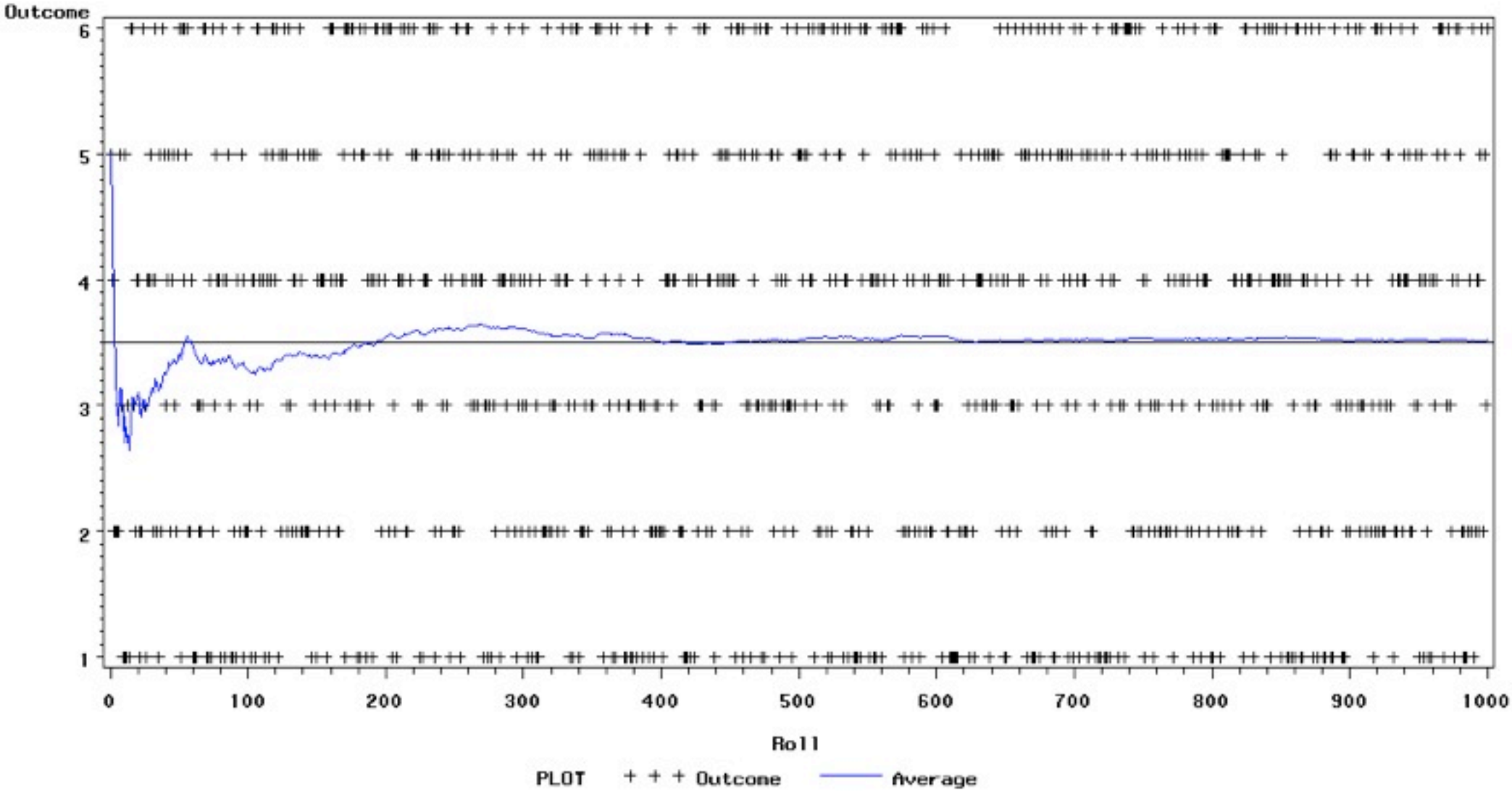

99.7% between ±3 s.d.
95.4% between ±2 s.d.
68.3% between ±1 s.d.
Only 3 points in 1000
will fall outside the area
3 standard deviations
either side of the center line.
s.d. = standard deviation
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
-3
s.d.
-2
s.d.
-1
s.d.
Mean
+1
s.d.
+2
s.d.
+3
s.d.

One-in-a-million

- happens
- always happens
- streaks
- predictable chaos—regularity in randomness

LAW OF LARGE NUMBERS IN AVERAGE OF DIE ROLLS

AVERAGE CONVERGES TO EXPECTED VALUE OF 3.5

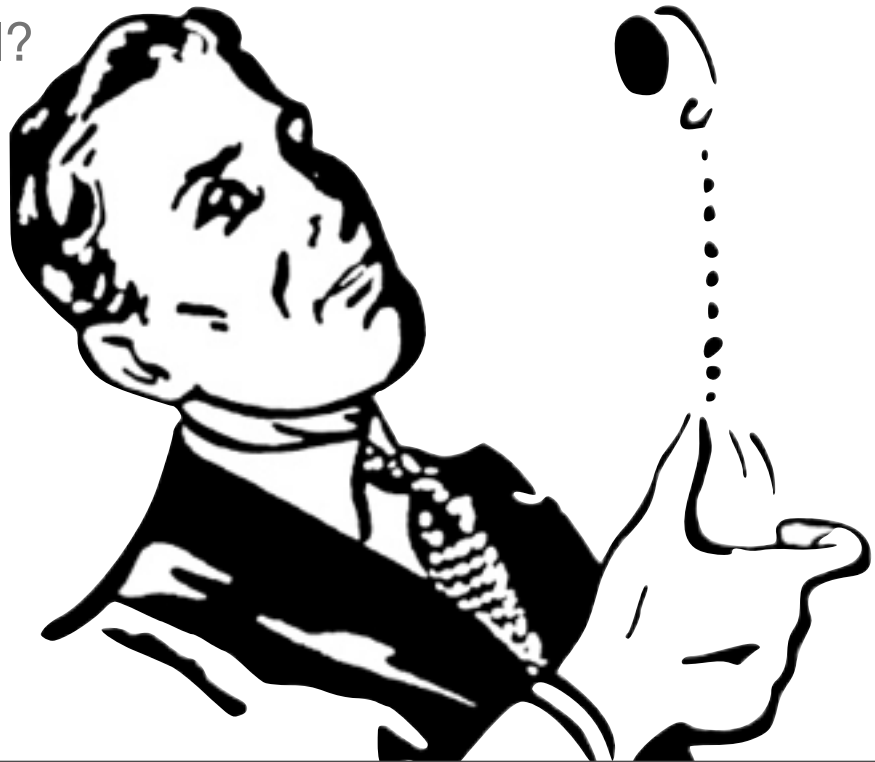


Exercise One: Flipping and Spinning Coins

- Using pennies, gather the results of 30 coin spins each, spinning the coin on edge on a table, achieving a stable spin and waiting for it to fall over without hitting anything.

- Enter the results in the class Google Spreadsheet: http://bit.ly/VvNsMi

- Are the results biased or fair? How can we tell?



Results!

- BINOMDIST(number_s, trials, probability_s ,cumulative)
  - number heads, total trials, 0.5, TRUE
- *Extra Credit: try standing some coins on edge and then banging the table.*

Binomial Distribution

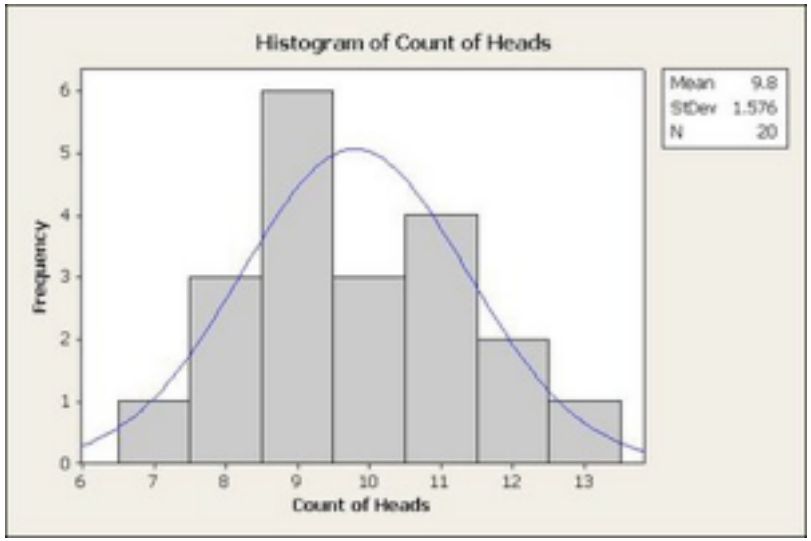
- The probability that a random variable X with binomial distribution B(n,p) is equal to the value k, where k = 0, 1,....,n. p is the probability of success

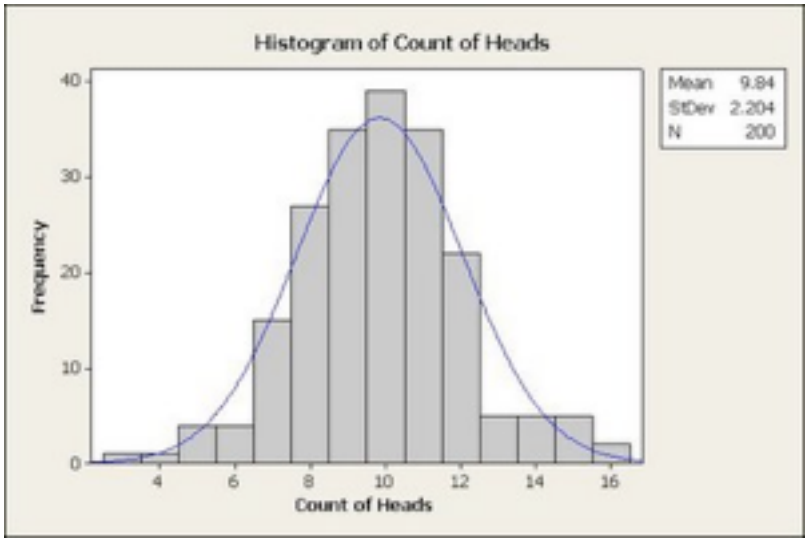
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

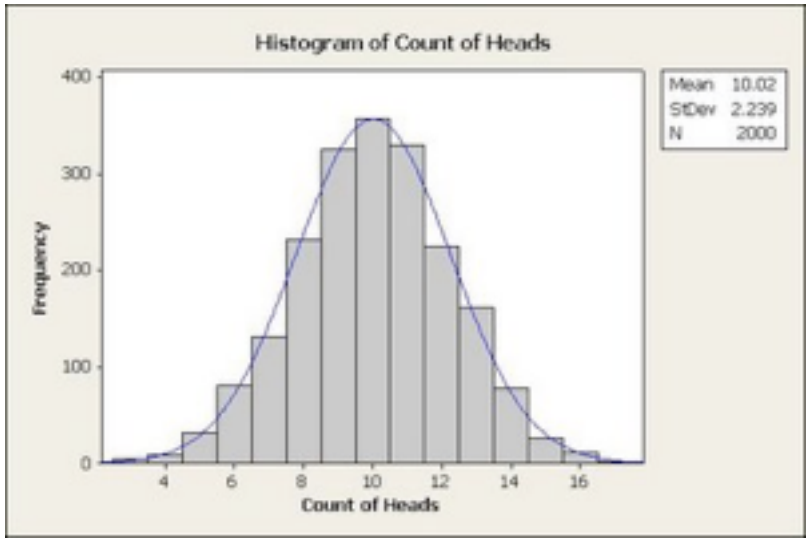
- *n choose k (number of ways to choose k successes from n variations, like five heads in ten coin tosses)*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial Distribution

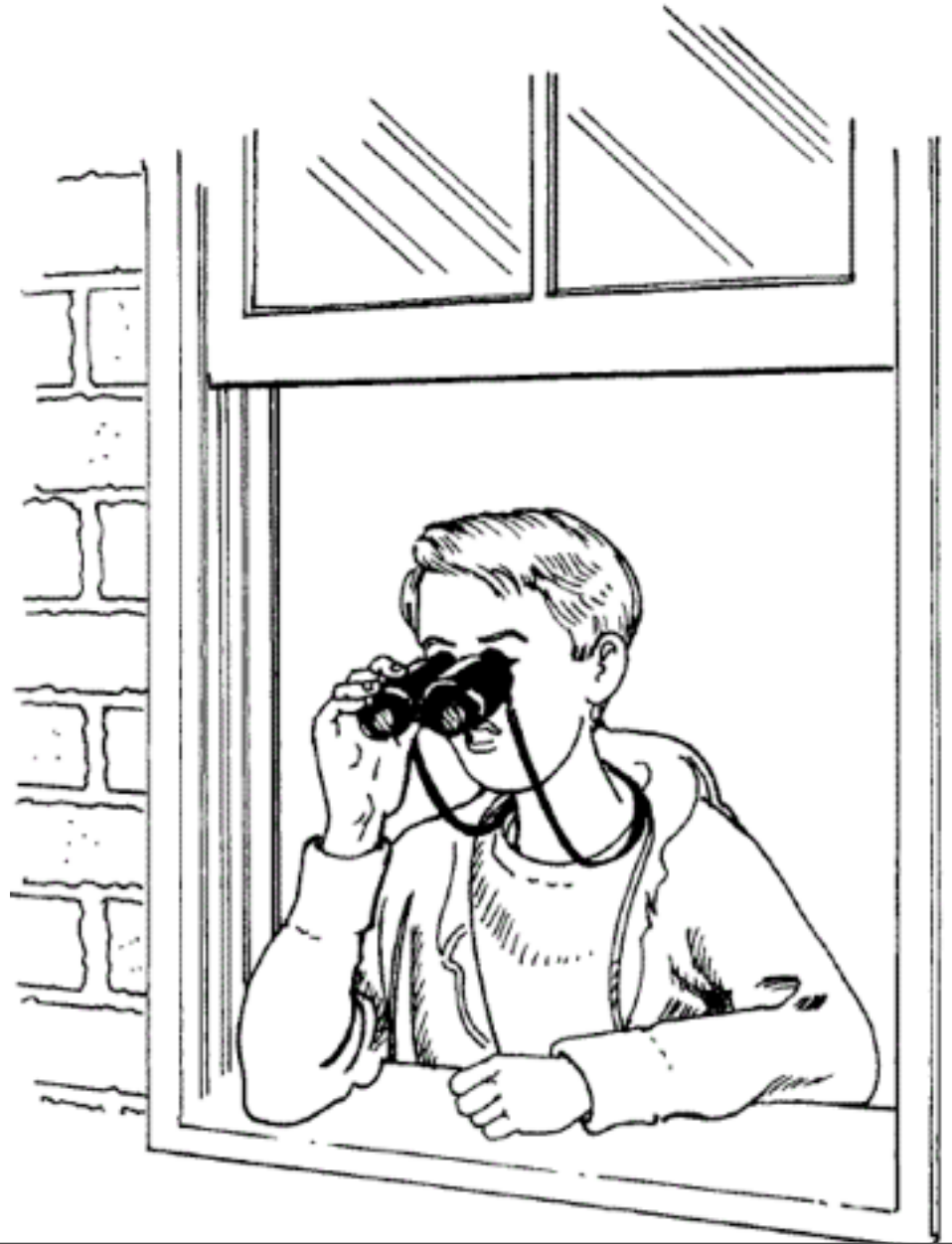






http://stattrek.com/online-calculator/binomial.aspx

Observation Assignment



Readings and Assignments

- Readings
  - Read Zen of Seeing; Cracking Creativity introduction & chapter 1; Drawing on the Right Side of the Brain, chapters 1, 3, 4.
- Assignment
  - Observation Assignment