# Big Data

Chapter 2: MORE

Ibrahim Olanigan

# Shift of Mindset

- Use of full datasets as opposed to smaller sets
- Embracing the real-word messiness of Data
- Growing respect for correlations

# Shift of Mindset

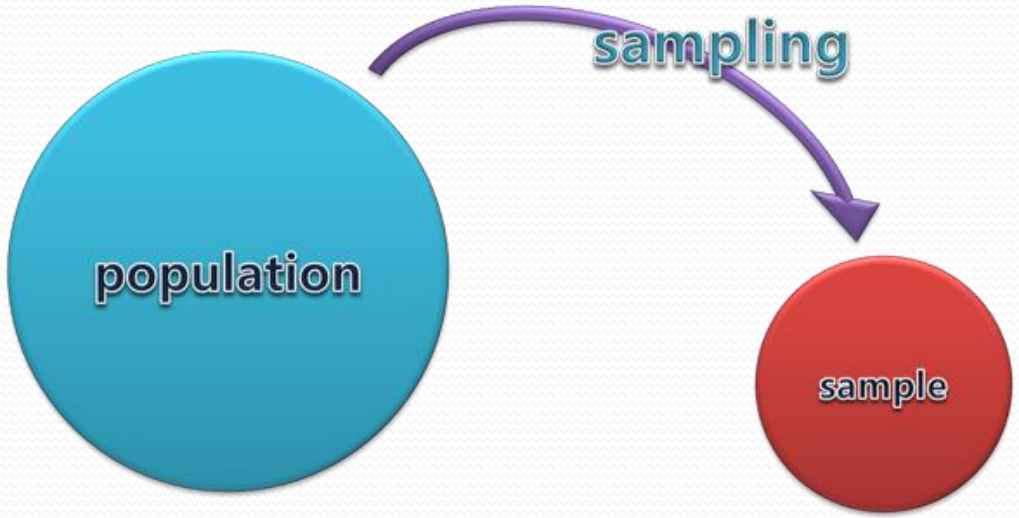
- **Use of full datasets as opposed to smaller sets**
- Embracing the real-word messiness of Data
- Growing respect for correlations

# 'Big' data problems

- Inadequacy of tools
- Census
  - Costliness
  - Processing time
  - Herman Hollerith (punch cards and tabulation machines)

# Data Sampling

- John Graunt and advent of statistics
- Importance of Randomness
- Solution to information overload



# Data Sampling: Limitations

- Second-best alternative.
- Dependency of randomness
- Increased inaccuracy with subcategories

# DNA Gene Sequencing

- 23AndMe
  - Cost-effective solution
  - Target sampling
- Steve Jobs & Cancer
  - Entire DNA Sequencing
  - More effective treatment

# Dataset, N= all

- Sampling in the era of information-processing constraints
- More effecient and adequate technical tools.
- Case studies
  - Google Flu Trends
  - Xoom and Discover Card fraud
  - Match-fixing in Japan's sumo wrestling

# Conclusion

- Usefulness of Sampling in the past and present
- Full dataset
  - In-depth exploration and analysis.
  - Reusability

# Questions? Comments?