*Date or other relevant info*

# Generating Simulated Data

// FLATIRON SCHOOL

# Agenda

- What is simulated data?
- Key Ideas
- Why use it? (and when not to)
- Review foundational concepts
- Generating simulated data in Python

# What is simulated data?

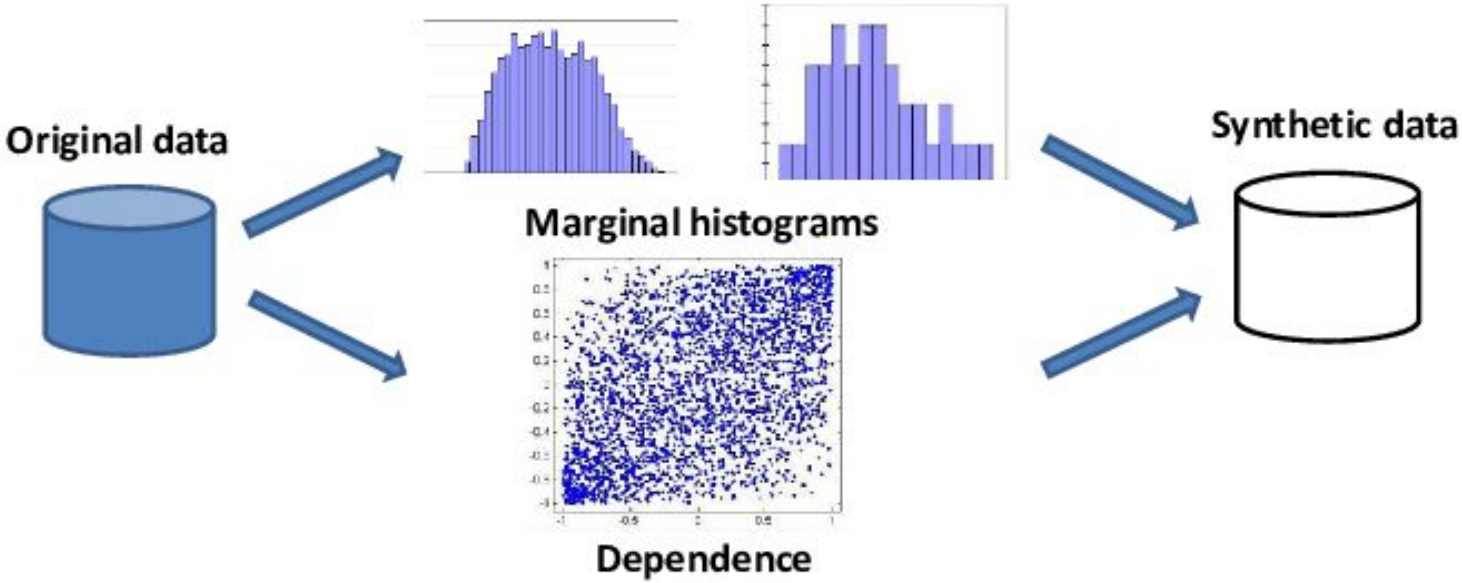
Date, subhead or category

// FLATIRON SCHOOL

# Simulated data…

- “Fake” or synthetic data created to reflect some real world data or system
- Generated by a computer
- Similar characteristics to real world data
  - Univariate distributions
  - Structure (multivariate relationships)

# Generation Process



Original data

Marginal histograms

Dependence

Synthetic data

# Key Ideas

Date, subhead or category

// FLATIRON SCHOOL

# Terminology

- Data can be *structured* or *unstructured*
    - **Structured** - tabular format
    - **Unstructured** - images, video, text, etc
        - OpenAI GPT-3
        - OpenAI Dall-E 2
- ***Utility*** - how accurately simulated data reflects real data
    - Required utility depends on the use case

# Generation Methods

- Real data
  - High utility
  - More resources required
- Existing model or knowledge
  - Lower utility
  - Fewer resources
- **Reproducibility** - generation process can be replicated

# Simulated Data Metrics

- Measure how well simulated data reflects real data
- Squared error
- Statistical Tests
    - Kolmogorov-Smirnov (KS) test
    - Chi-squared goodness of fit test
- Machine learning

# Why use it?
# (and when not to)

Date, subhead or category

// FLATIRON SCHOOL

# Efficient Data Access

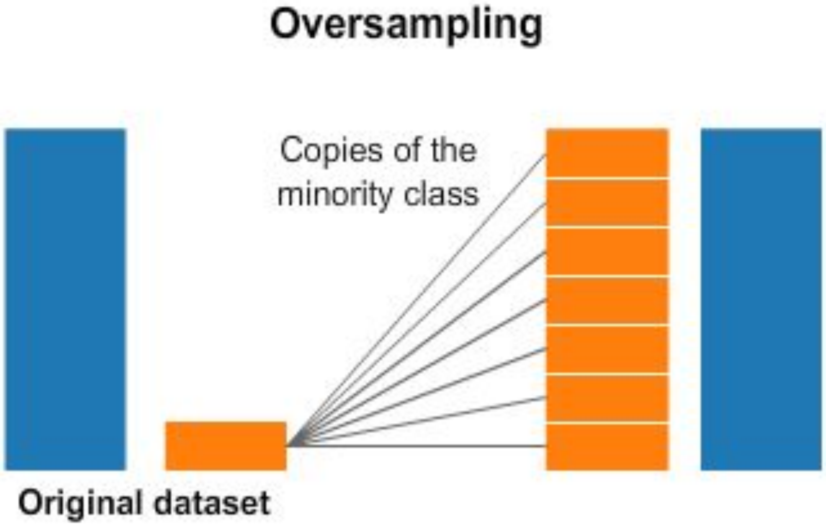
- Simulated data is cheaper and faster
- Restrictions prevent access to data
- Solve privacy concerns
- Open data sources lead to:
    - Reproducibility
    - Innovation

# Improve Analytics

- Make data open
- Test hypotheses
- Increase data size for modeling
- Account for edge cases and rare events

# Machine Learning

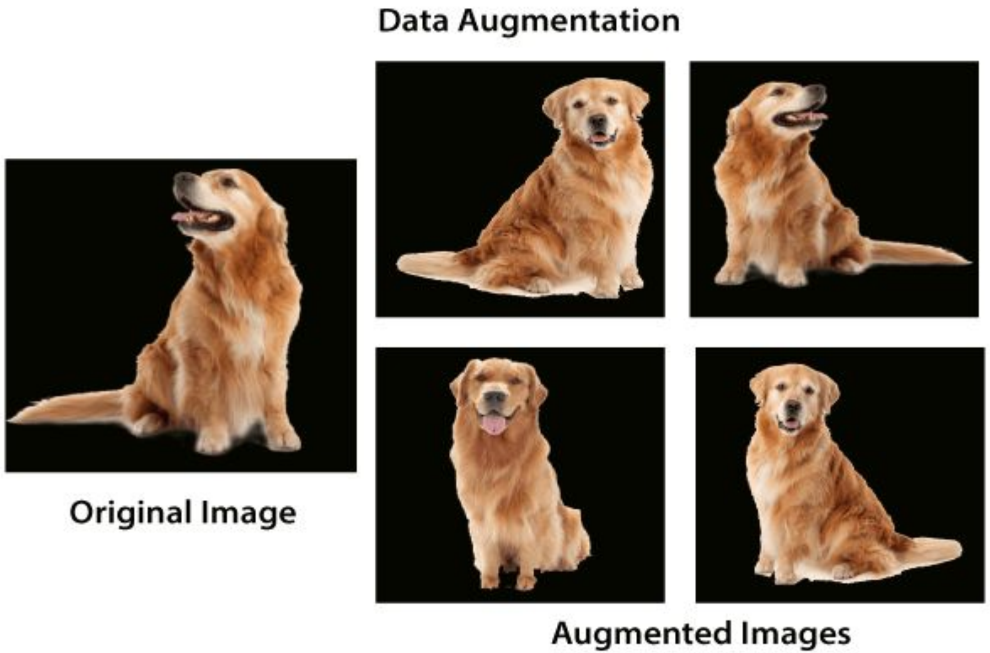
- Synthetic data to solve for class imbalance
  - Fraud
  - Car crashes
  - Medical scans
- More data for training



// FLATIRON SCHOOL

# Computer Vision

- Requires immense amount of data
- Impractical and expensive to collect
- Data augmentation creates synthetic samples for model

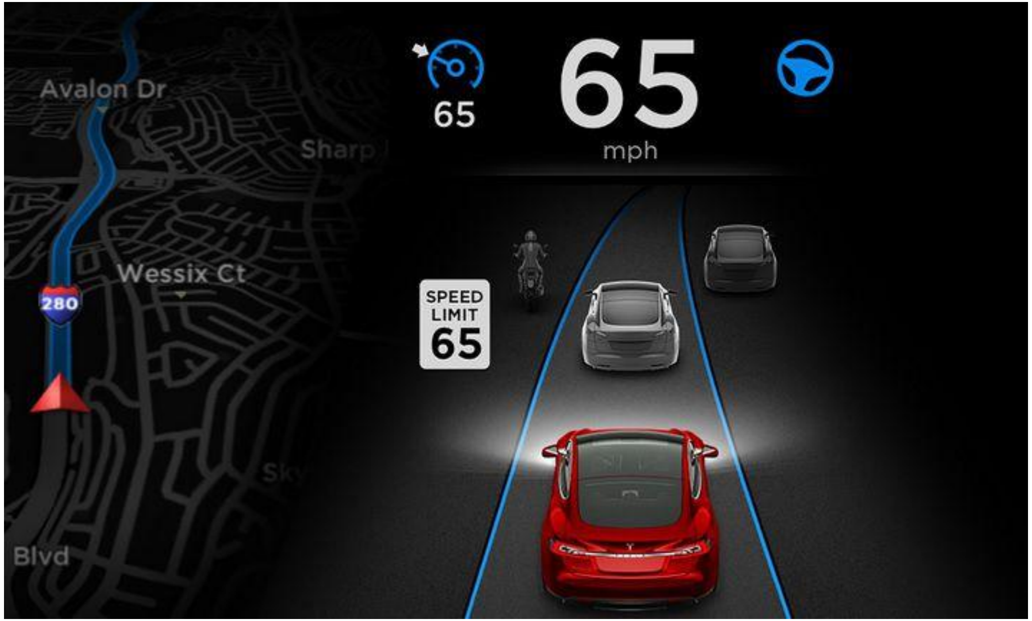


# Healthcare

- Challenges
  - Health data is often private
  - Expensive to collect real data
- Simulated data solution
  - Explore new digital health technologies
  - Sharing data open source

# Autonomous Vehicles

- Limited data on edge cases
- Requires lots of training data
- Run simulations for rare events



# Why NOT to use it

- Lack of resources
- Limited understanding of data and/or process
- Concern of privacy breach
- Misinterpreted as real data
    - Control narrative with misinformation
    - Always communicate data was simulated

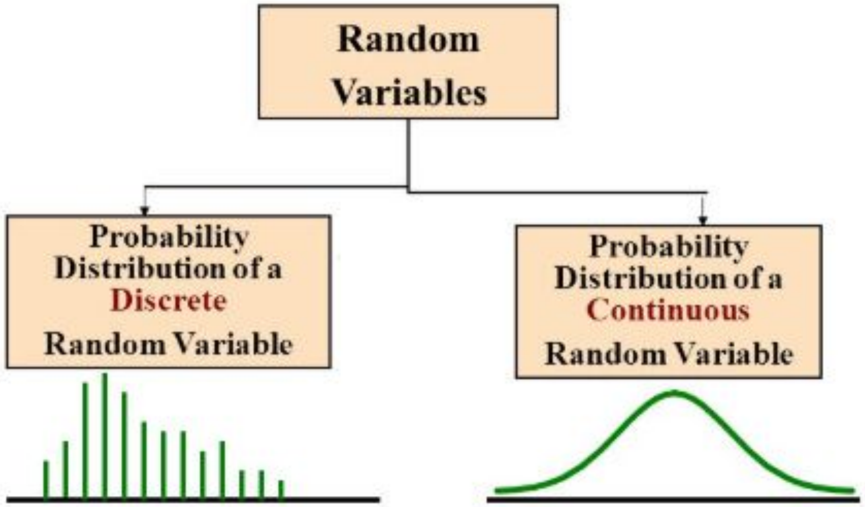
# Review foundational concepts

Date, subhead or category

// FLATIRON SCHOOL

# Discrete vs Continuous

- Discrete
  - Finite number of values
- Continuous
  - Infinite number of values within range



Random Variables

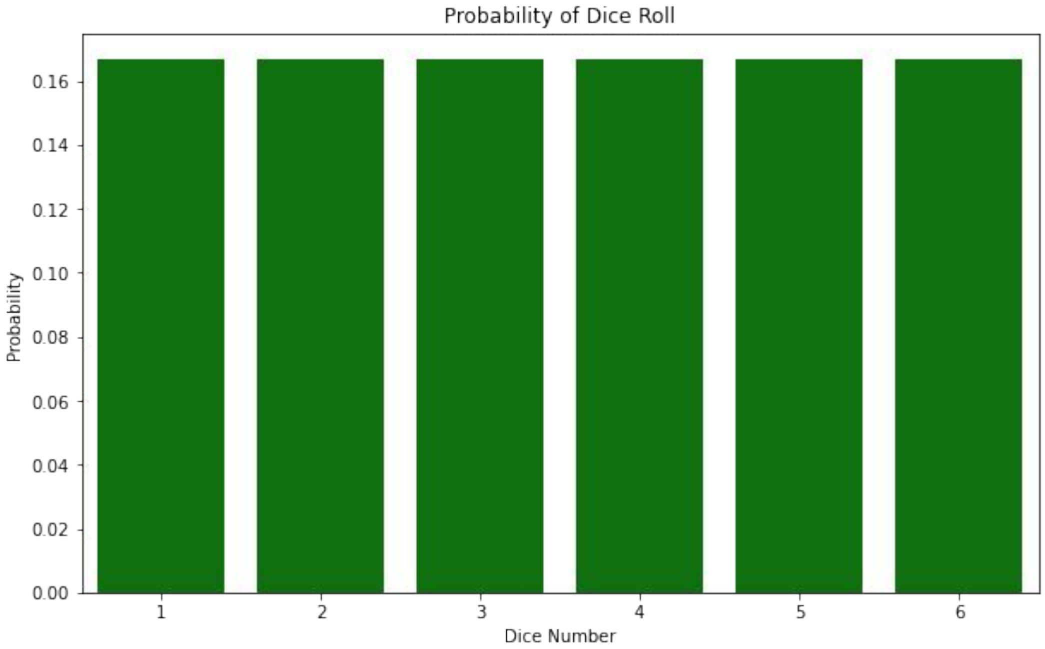
Probability Distribution of a Discrete Random Variable

Probability Distribution of a Continuous Random Variable

# Discrete Probabilities

- Likelihood of specific outcome
- P(x=2) = .16

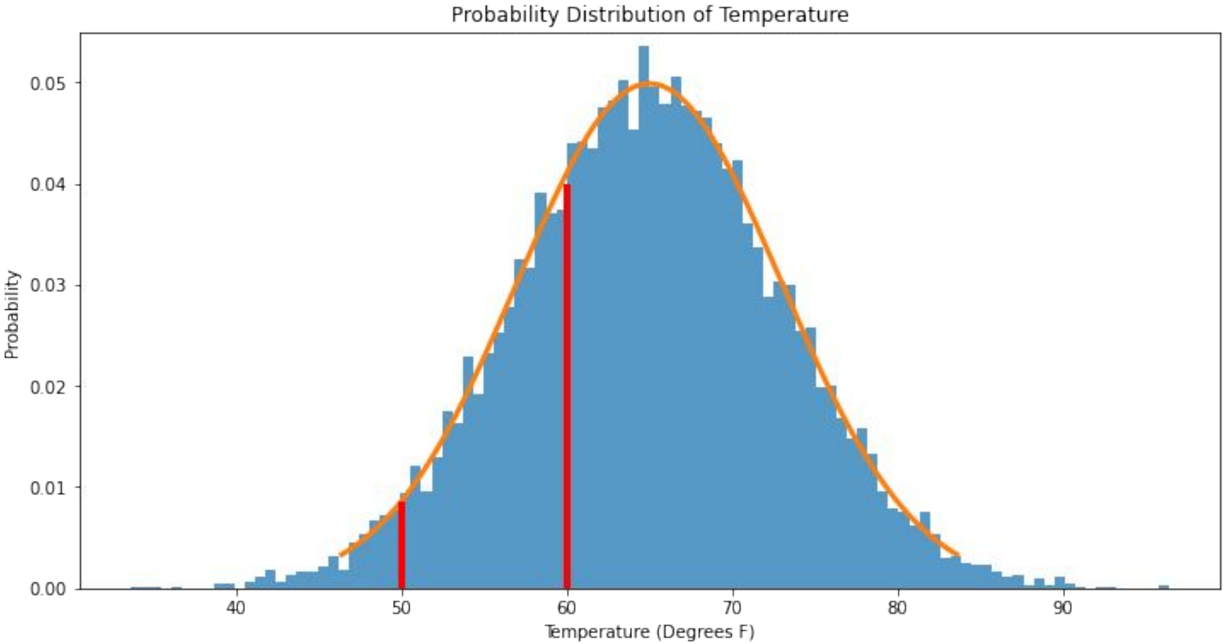
| Roll | P(X) |
|---|---|
| 1 | .16 |
| 2 | .16 |
| 3 | .16 |
| 4 | .16 |
| 5 | .16 |
| 6 | .16 |



# Continuous Probabilities

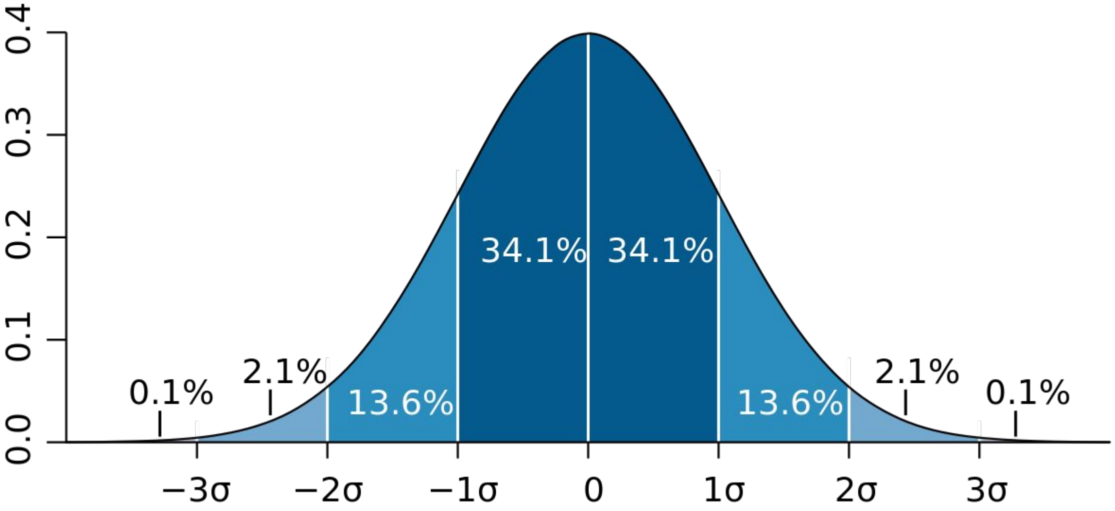
- Probability of range
- P(50 <= x <= 60) = .235

| Temperature |
|---|
| Mean = 65 |
| Standard Deviation = 8 |

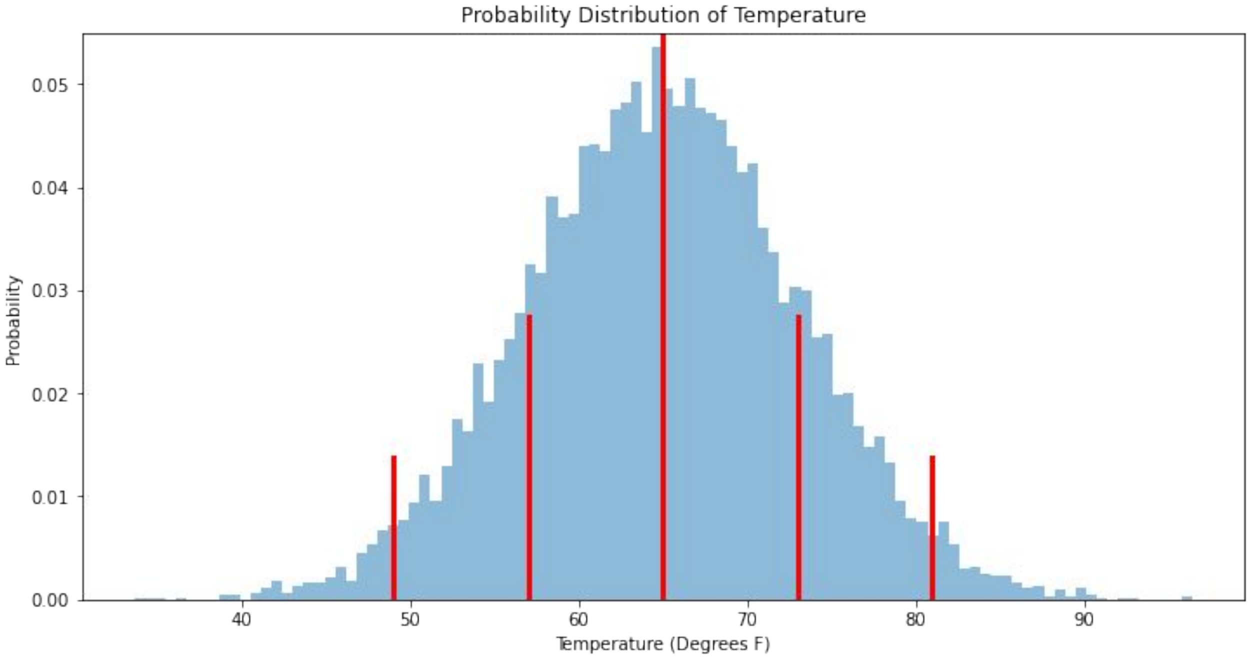


# Normal Distribution

- Mean - center
- Standard deviation - spread
- Empirical rule

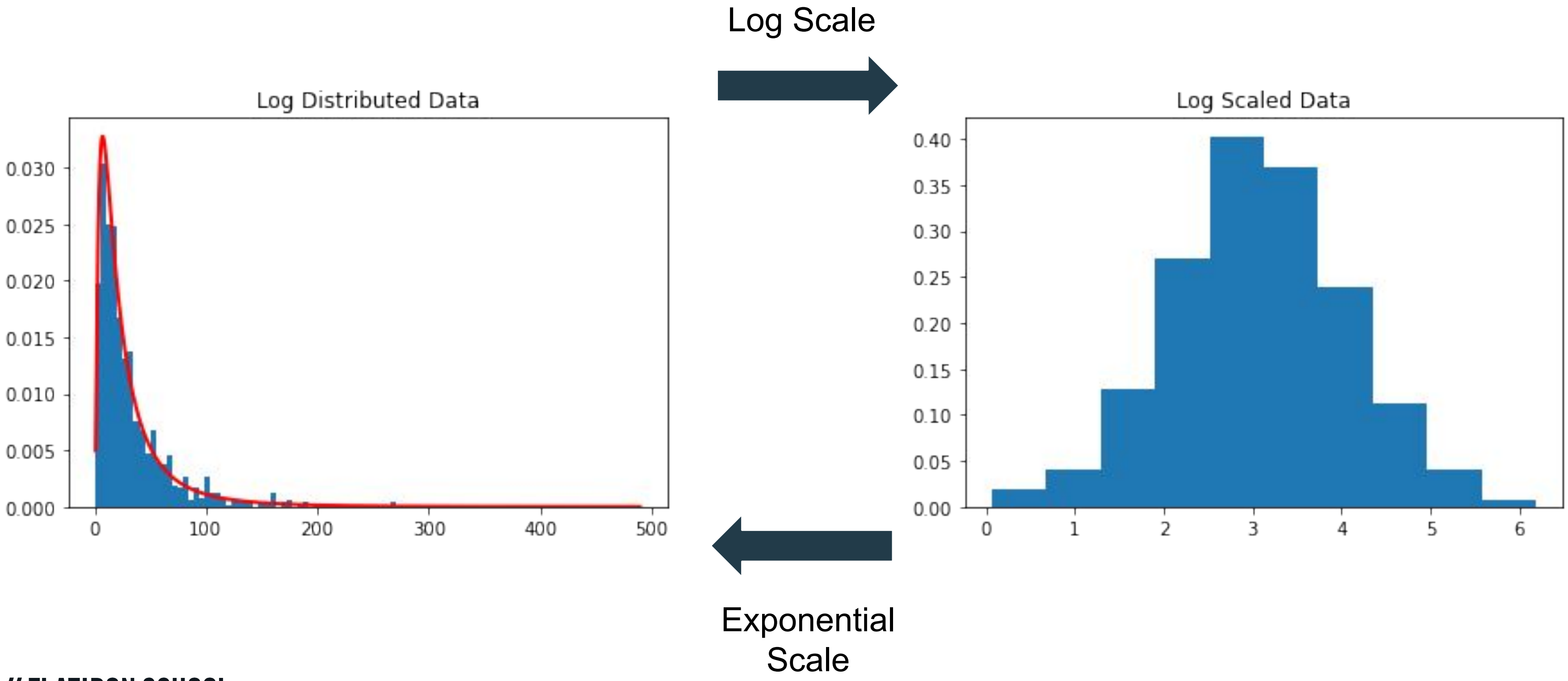


# Normal Distribution



| Temperature |
|---|
| Mean = 65 |
| Standard Deviation = 8 |

# Scaling Data

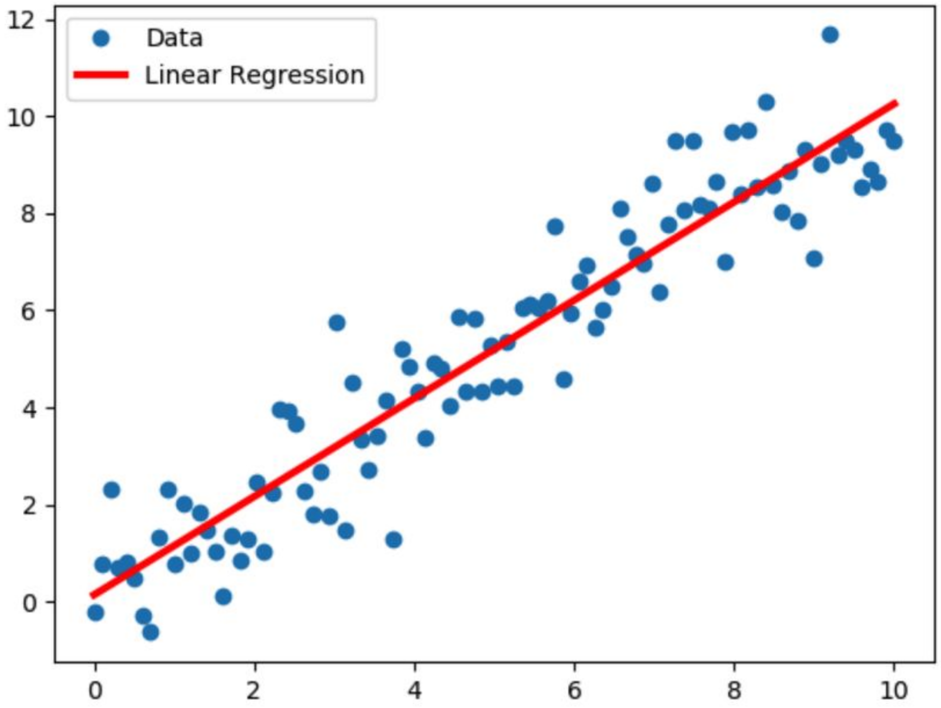


Log Scale

Exponential Scale

# Linear Regression

- Model relationship between 2 variables
- Noise
  - Inherent error in the data



# References

Date, subhead or category

// FLATIRON SCHOOL

# References

- Practical Synthetic Data Generation - OReilly
- Webinar: What is Synthetic Data?
- We need Synthetic Data - Medium Article