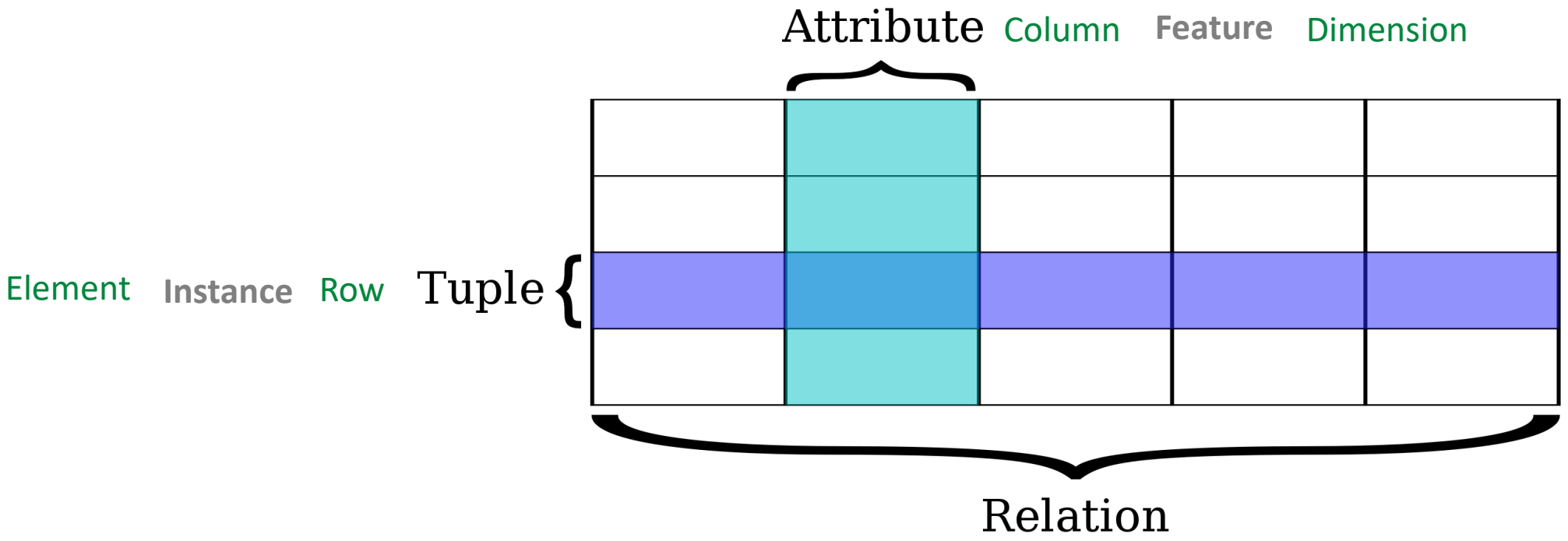
Machine Learning

Praphul Chandra

1. James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. New York: Springer, 2013.

Focus on Relational Data Model



$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

Need to describe (not only visualize) the data

Patterns in Data

- Visualization
- Description
- The idea of a (summary) statistic

Statistic

- is a single measure of some attribute of a given data set
- It is calculated by applying a function to the given data set.
- Example : Average (Add each data element; Divide by total number of elements)
- a.k.a. Summary statistic (since it summarizes some attribute of the data)

Descriptive Analytics using statistics

- Centrality : *Mean, Median, Mode*
- Spread : *Variance, Standard Deviation, Min, Max, Outliers*
- Symmetric : *Skew*
- Cleanliness : *Missing Values*

The notion of a random variable

Random Variable

- Intuitively: Header / Column-name of a relational table
- More generally: A ‘variable’ that can take different values for different instances (rows)
- *e.g. age of a person, presence of a word in text document, value of a pixel, sentiment of a tweet*

Type of a Random Variable : What values are possible?

- Numeric Values : Discrete, Continuous, [0,1]
- Categorical : Nominal, Ordinal

Range of a Random Variable

- Intuitively: The range of valid values allowed in a column
- Answers a key Question: What values are possible?
- *e.g. Age > 0, Income ≥ 0, Sentiment ∈ {+,-,0}, Probability ∈ [0,1]*
- Range of a random variable depends on the units of measurement!
- Interpreting Invalid Values requires domain knowledge *(0 age, -ve Income,)!*

The notion of a random variable (cont'd)

Probability

- Intuitively: How often does a particular value occur in the data set (column) ?
- What is the probability that an instance (row) in the data set has a given value?
- *e.g. How many people are aged 35 in the data? How many adults?*

$$P(x = 35)$$

$$P(x > 18)$$

Distribution of a Random Variable

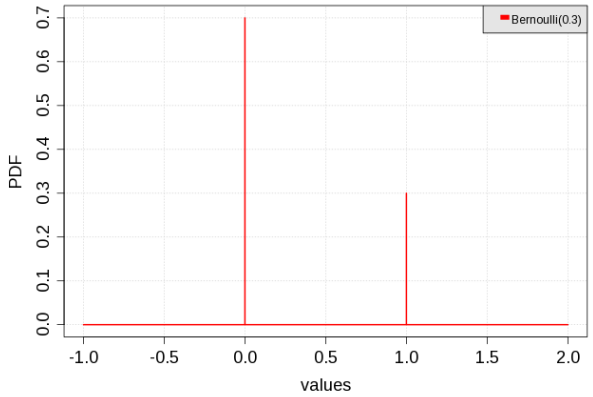
- Intuitively: For each possible value in the Range, how often does it occur?
- What is the probability that an instance (row) in the data set takes each of the possible values?
- *Note: We are now looking for patterns in a data set!*

$$P(sex = Male), P(sex = Female)$$

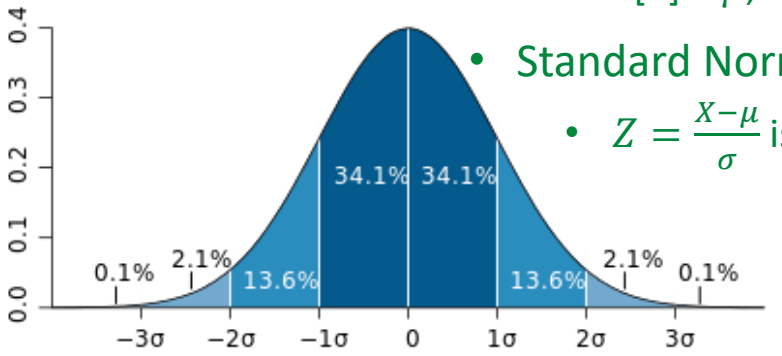
$$P(x = 1), P(x = 2), P(x = 3), \ldots, P(x = 100)$$

Probability Mass Function

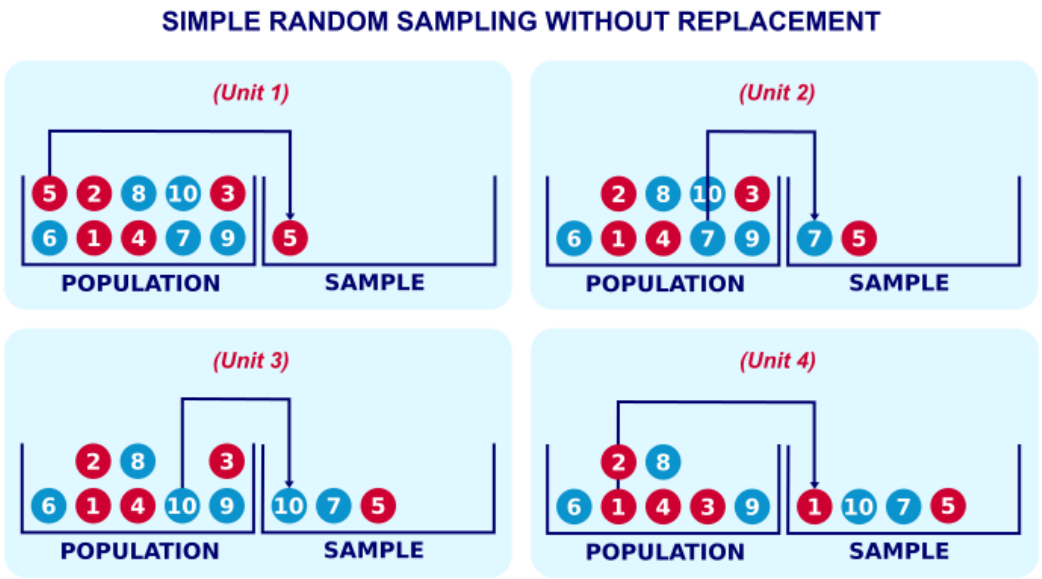
- For Discrete Random Variables
  - Random variable X takes a finite number of values
- Specify P(X=x) for all possible x
- Given a PMF
  - Expected Value (Mean) : $\sum x_i P(x_i)$
  - Variance : $\sum (x_i - \mu)^2 P(x_i)$



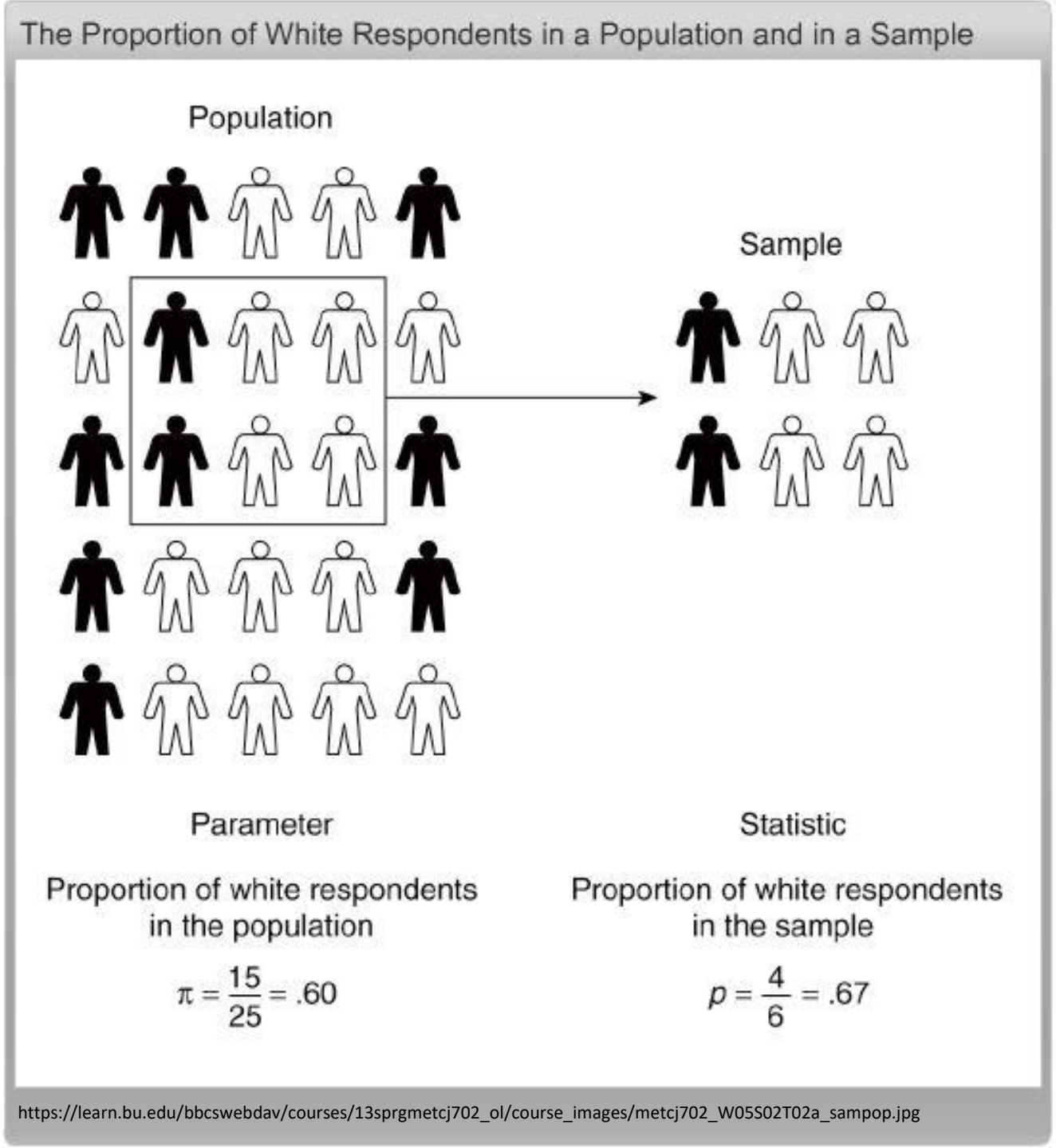
- For Continuous random variable
  - Specify P(X=x) for all possible x?
  - Continuous r.v. take infinitely many values➔ Probability of a particular value → 0
- Specify P(x_1 < X < x_2) for all possible x
  - Specify probability density P(x_1 < X < x_2) /(x_2 - x_1)
- Given PDF
  - Expected Value (Mean): $\int x_i f(x_i)$
  - Variance: $\int (x_i - \mu)^2 f(x_i)$
- X ~ N(μ, σ^2)
  - E[X] = μ; Var(X) = σ^2; 68-95-99.7 empirical rule
- Standard Normal Random Variable
  - $Z = \frac{X-\mu}{\sigma}$ is called the Standard Score or the z-score.



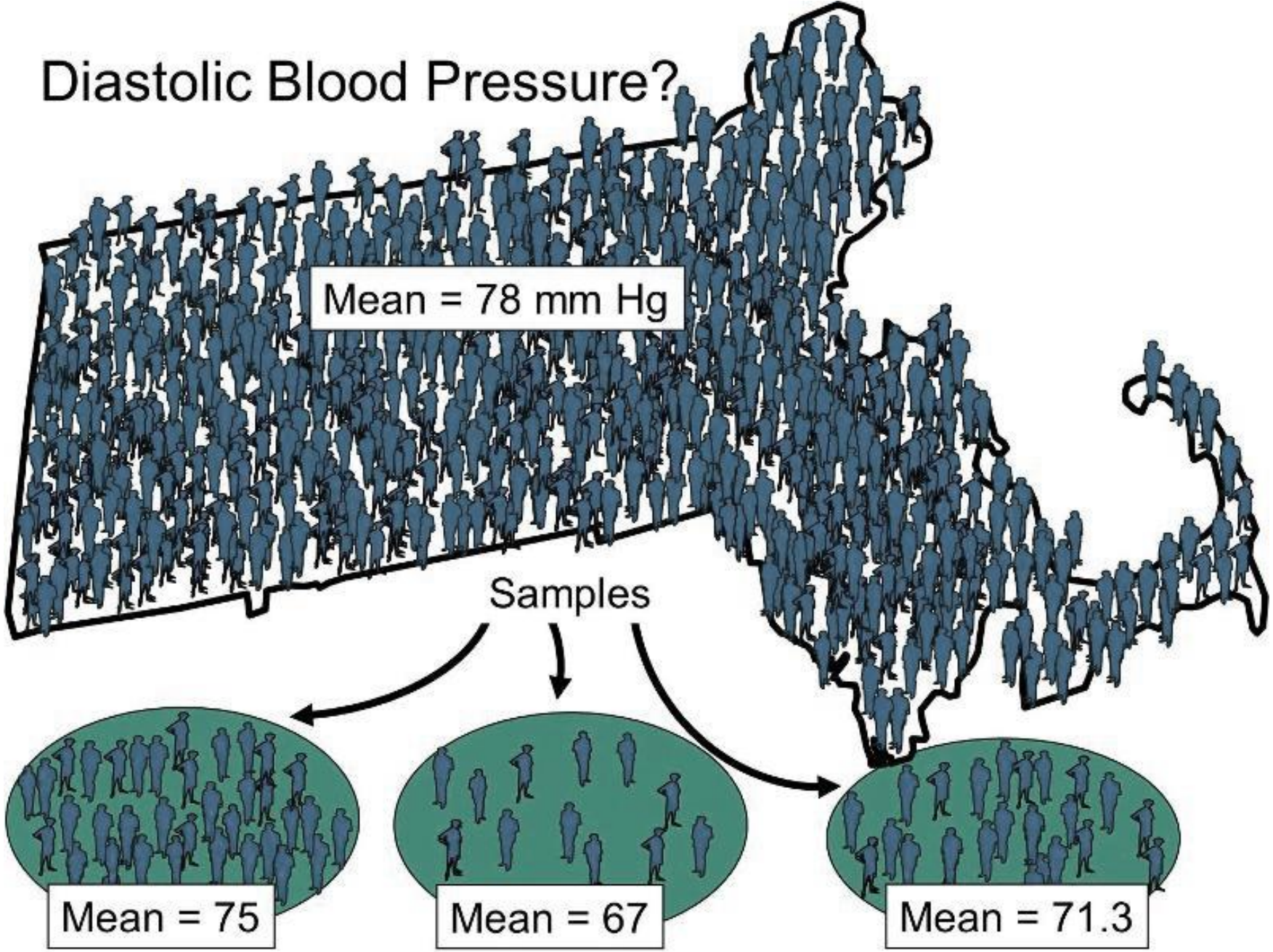
Population & Sample



https://spss-tutorials.com/img/simple-random-sampling-without-replacement.png



https://learn.bu.edu/bbcswebdav/courses/13sprgmetcj702_ol/course_images/metcj702_W05S02T02a_sampop.jpg



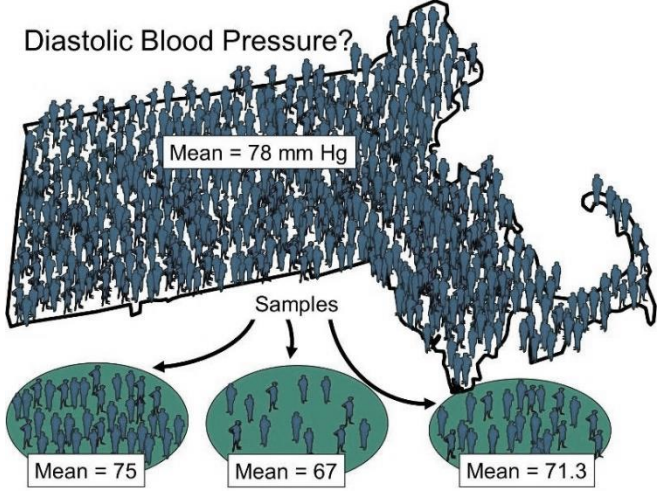
http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_BiostatisticsBasics/Sampling3.jpg

Sample vs. Population : Significance

- Find patterns in your data which hold across “data”.
  - Find patterns in the given data which hold in other data coming from the same process.
  - Find patterns in your sample data which hold in the population.
  - Find patterns in your training data which hold in the test data.

- When can you generalize from sample?
  - i.e. When can you make predictions?
  - Is your sample representative?
  - Is past a good predictor of future?
  - *e.g. Were there systemic changes? Was there a rare event in the past? Did you sample randomly?*

Inferential Statistics



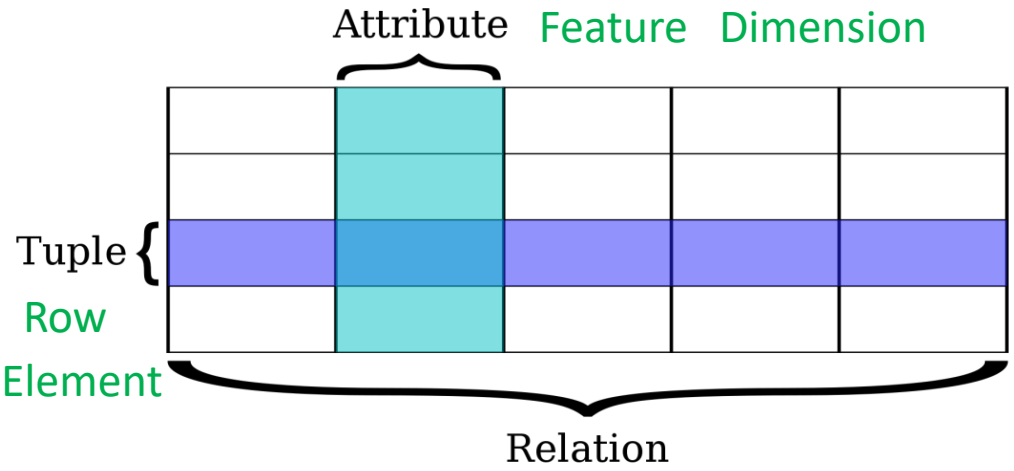
- What can we infer about the **population** from the (observed) **sample**?
  - The objective of inferential statistics is to use sample data to obtain results about the whole population.
- What can we infer about the **population parameter** from the (observed) **sample statistic**?
- What can we infer about the **population mean** from the (observed) **sample mean**?
  - Can the sample mean be very far away from the population mean?
  - Can one sample mean be very far away from the another sample mean?
  - Do these depend on the sample size?
    - Sample size = 1?
    - Sample size = population?
  - Do these depend on the underlying distribution?

Unsupervised Learning

Praphul Chandra

1. James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. New York: Springer, 2013.

What does data look like?

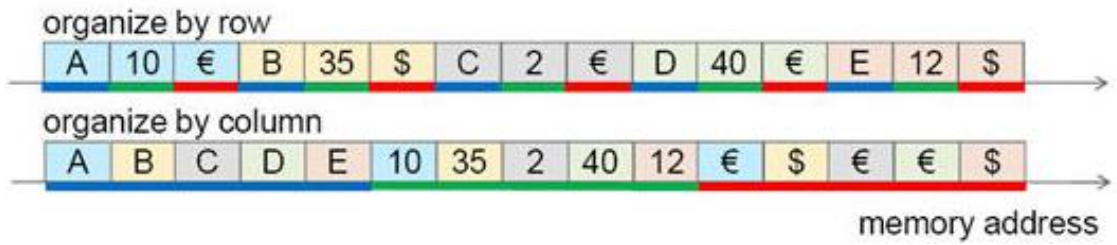


$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

- Number of rows = n
  - Large n : Big Data
- Number of column = p
  - Large p : High dimensional data

A	10	€
B	35	$
C	2	€
D	40	€
E	12	$

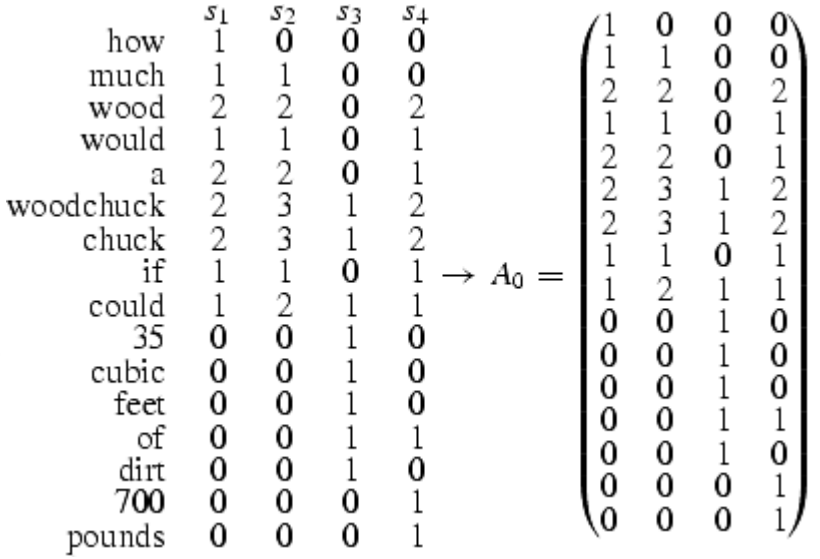


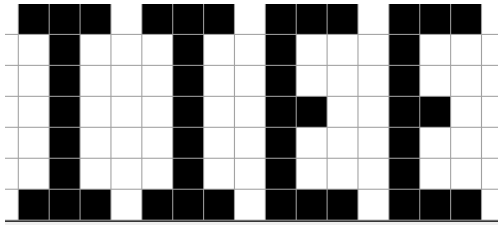
- Row store
  - At creation
- Columnar store
  - At analysis

Relational Data Model

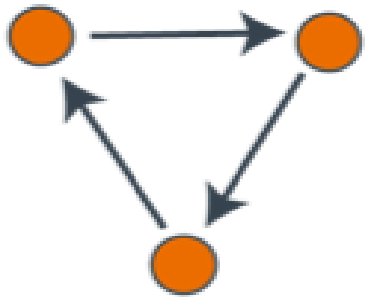
- Pretty powerful
  - RDBs
  - Spreadsheets
  - Matrices
  - Very often the data view
  - Brittle : Schema exists before data

Relational data model

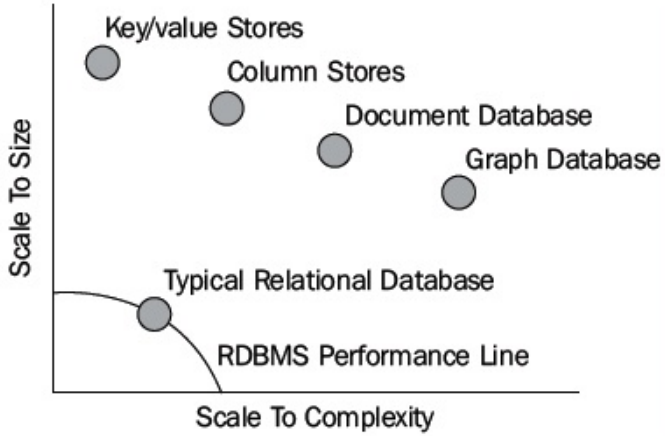




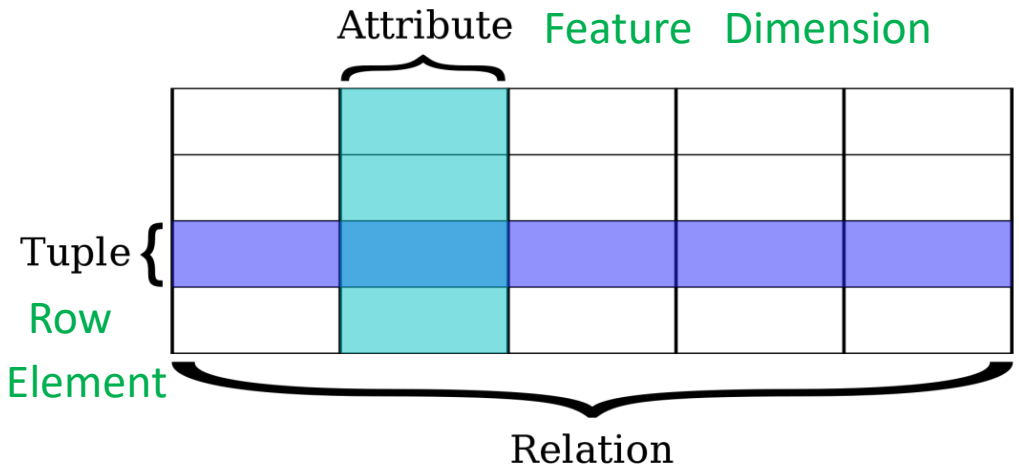
- Alternate
  - Unstructured data
  - Structure on Read (Delay Structure)
  - Non-relational data models



Document data model



What does data look like?



$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

What does data “really” look like?

If you look carefully, data has patterns.

Unsupervised Learning is about finding patterns in data.

Unsupervised Learning

Finding patterns in data.

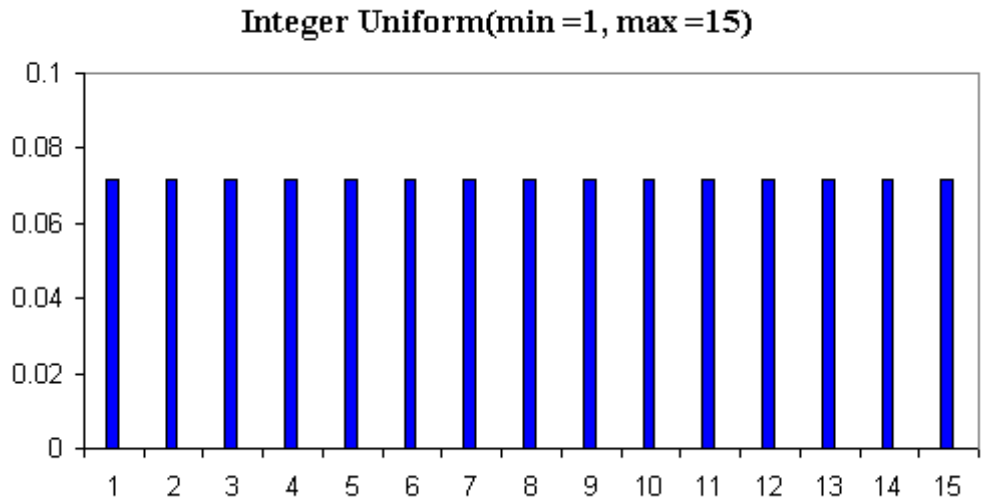
Definitions

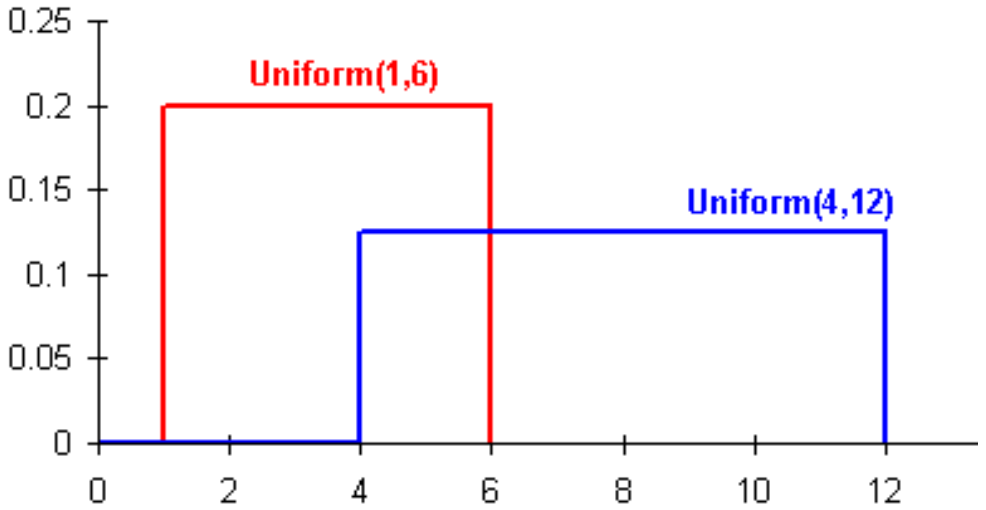
- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.

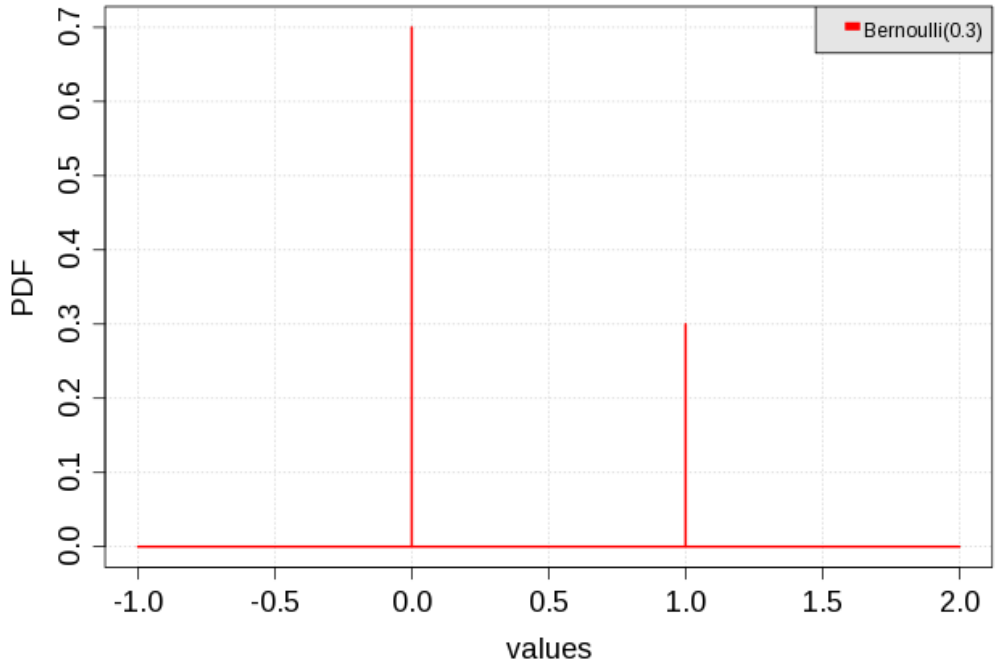
- “Unsupervised”
  - Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution – this distinguishes unsupervised learning from supervised learning and reinforcement learning.

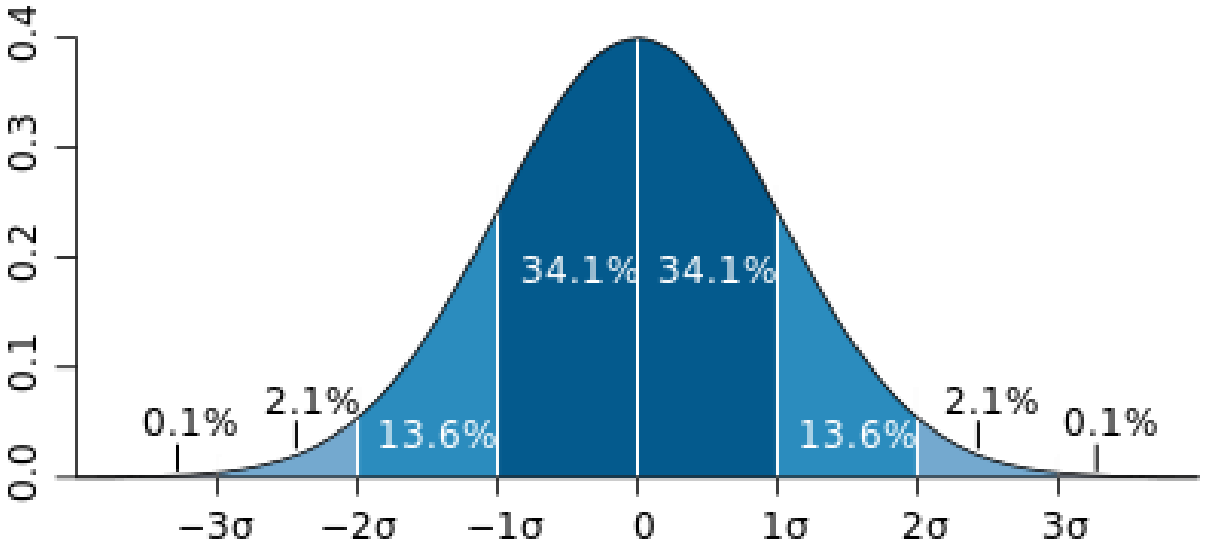
- … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density
  - Summary statistics

What does a Distribution look like? (p=1)

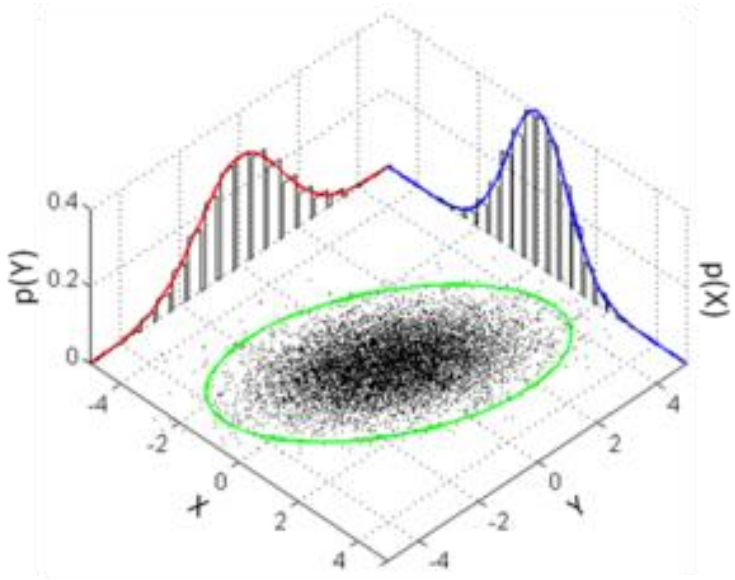


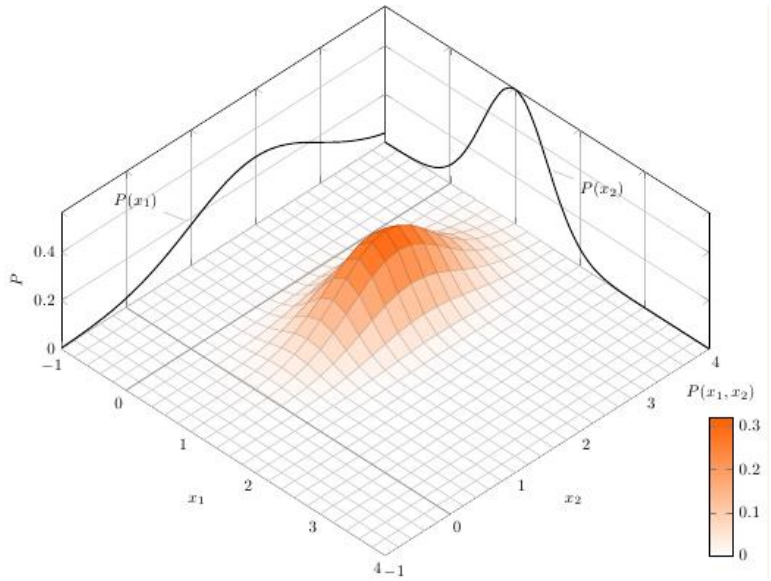


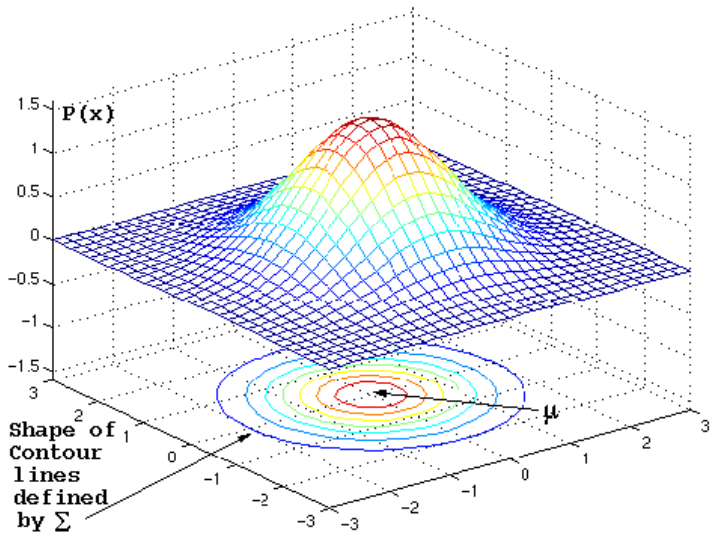




What does a Distribution look like? (p=2)







Patterns in data

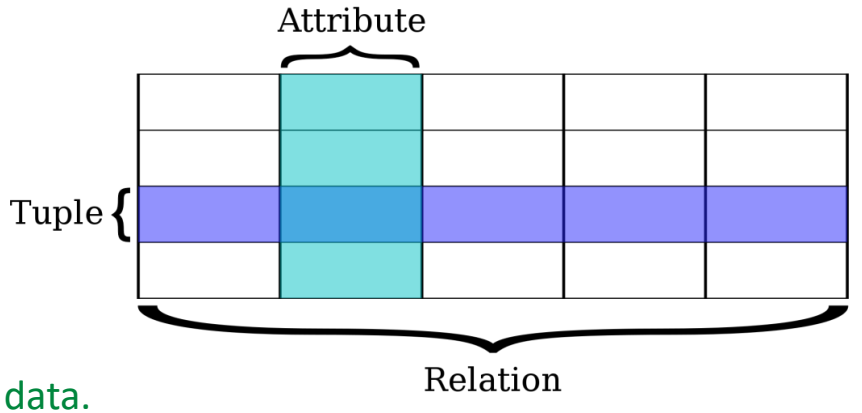
- They describe structure (patterns) in the data
  i. Which value(s) occur most frequently?
  ii. How much does the data vary?
  iii. How symmetrically does data vary around center?
  iv. Is data clustered around value(s)?
  v. Sub-space where data is “concentrated”
- Summary statistics
  i. Median
  ii. Variance, Standard Deviation
  iii. Skewness, Kurtosis
  iv. Mode
- Multiple dimensions
  i. Are two features / dimensions correlated

- Clustering
  - Find data elements which are similar.
  - Finding “areas” in space where data is concentrated
- Dimensionality Reduction
  - Find smaller dimensional representations of the data which preserve it’s essential structure.
  - Find subspaces where data varies the most.
- Remember
  - The Elephant
  - Both are tools : Learn when to use what.

Supervised Learning

Modelling

~~Un~~Supervised Learning



- Unsupervised Learning
  - Given X
  - … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density, Summary statistics, Clustering, Association Rules, Dimensionality Reduction

- Supervised Learning
  - Given X & y (a **particular** random variable)
  - Find what is the relation between the particular random variable and other random variables
    - What if we are only interested in identifying customers who bought Milk?
  - Find how the value of the dependent variable depends on the value of others
  - Find how the outcome is related to the features
  - Key Variations: Type of outcome / dependent r.v.
    - Numeric (Discrete, Continuous, [0,1])
    - Categorical : Nominal, Ordinal

The idea of a Model

- Physical
  - a physical copy of an object such as a globe
- Computer
  - a simulation to reproduce behavior of a system
- Scientific
  - a simplified & idealized understanding of physical systems
  - Newton's Law model the physical universe
- Conceptual
  - a representation of a system using general rules & concepts
- Mathematical
  - a representation of a system using mathematical concepts

$$y = 3x + 4$$
$$y = x^2$$
$$y = e^x$$
$$y = \log(x)$$
$$y = \sin(x)$$

- Statistical
  - a parameterized set of probability distributions

All models are false. Some models are useful.

The idea of a Statistical / ML Model

- Model
  - A function relates two (or more) variables
  - Captures the relation between x and y
  - For every value of x, there must be a unique value of y
  - Data looks like $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots (x_n, y_n)\}$

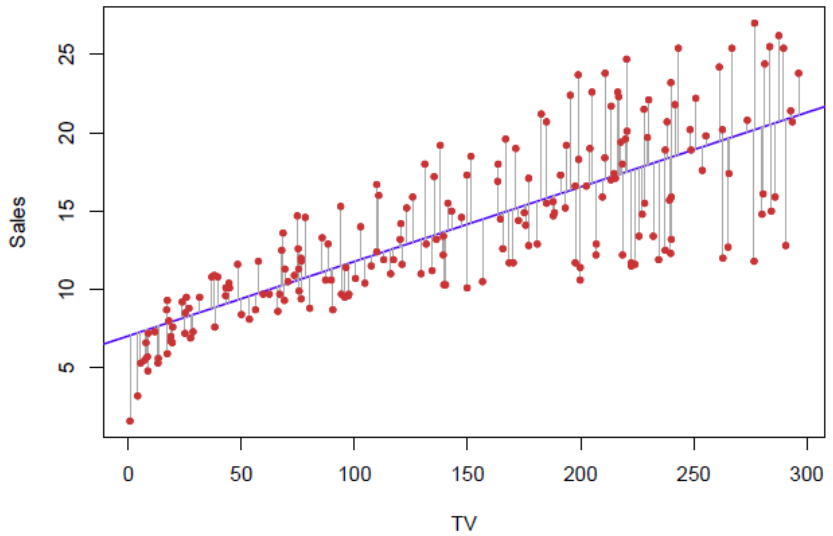
$$y = 3x + 4$$
$$y = x^2$$
$$y = e^x$$
$$y = \log(x)$$
$$y = \sin(x)$$

$$y = f(x)$$

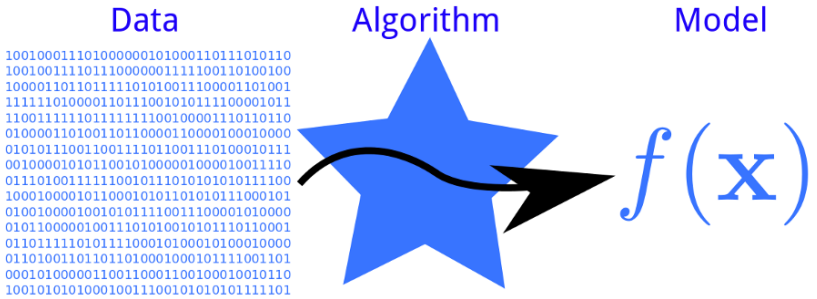
- Statistical Model
  - Real world data looks like $\{(x_1, y_1), (x_1, y_2), \ldots, (x_n, y_n)\}$
  - Multiple values of y for a single value of x
  - In expectation (on average), "model" captures the relationship between variables
  - Effects due to unobserved variables / Errors in measurements : capture by ε
  - Randomness / Stochasticity / Noise : Zero-mean; Normal distribution
  - Violations of Assumption is an indication of systemic errors

$$y = f(x) + \varepsilon$$
$$\varepsilon \sim \mathrm{N}(0, \sigma)$$

$$\hat{y} = \hat{f}(x) + 0$$
$$\mathrm{P}(y \mid x)$$

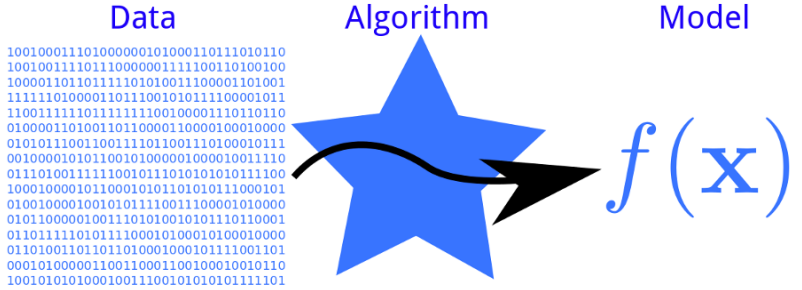


Un/Supervised **Learning**

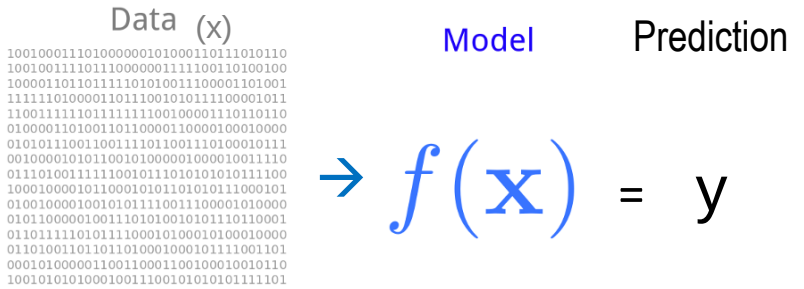


- Given X
  - … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density, Summary statistics, Clustering, Association Rules, Dimensionality Reduction

$$y = f(x)$$



- Given X & y (a **particular** random variable)
  - Find what is the **relation** between the particular random variable and other random variables
  - Find how the value of the **dependent (particular)** variable depends on the value of others
  - Find how the outcome is related to the **features**
  - Generalize : Make **predictions** about new data



Q?

Praphul Chandra

Insofe