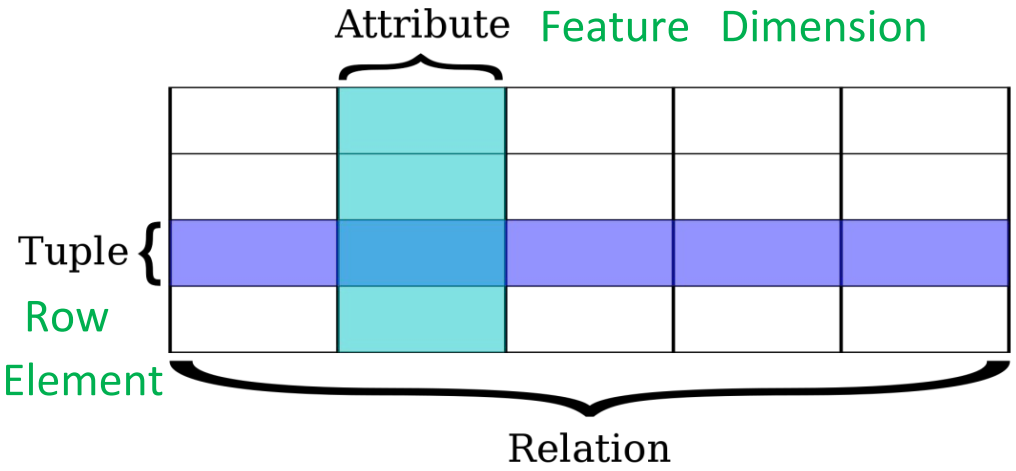
Unsupervised Learning

Praphul Chandra

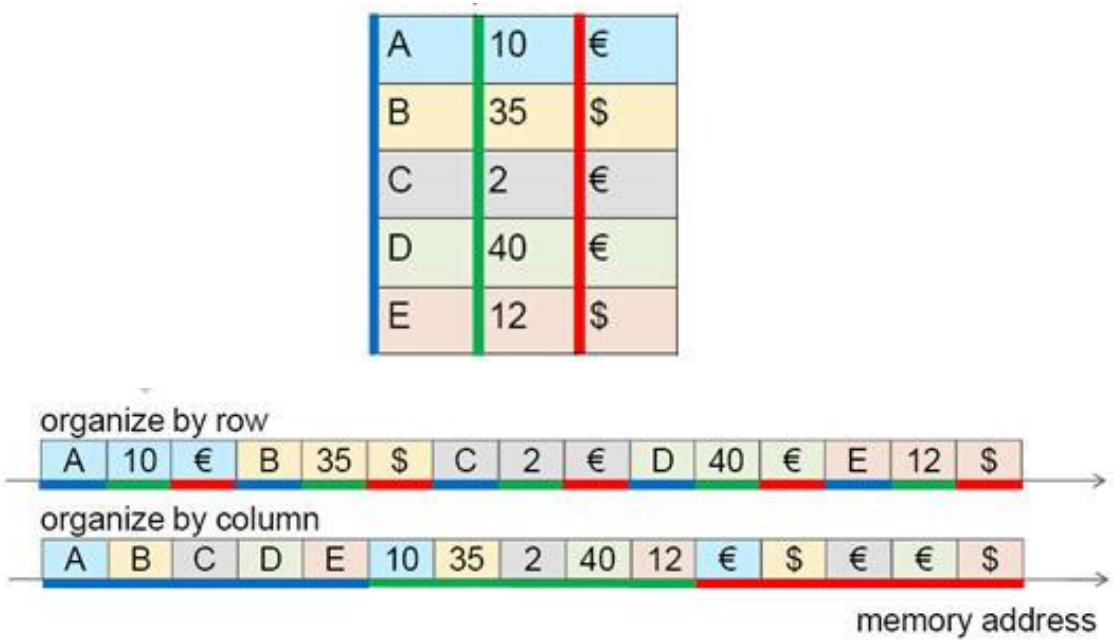
1. James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Kuhn, Max, and Kjell Johnson. *Applied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

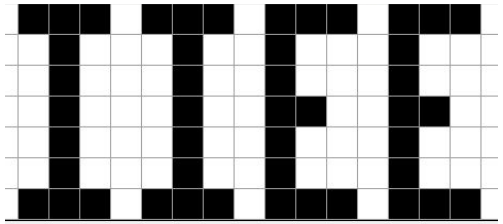
- Row store
  - At creation
- Columnar store
  - At analysis

Relational Data Model

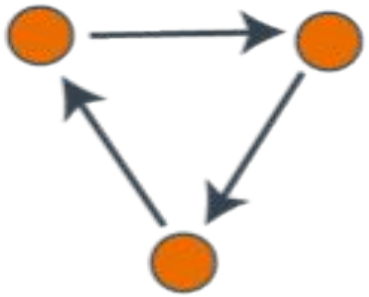
- Pretty powerful
  - RDBs
  - Spreadsheets
  - Matrices
  - Very often the data view
  - Brittle : Schema exists before data

Relational data model

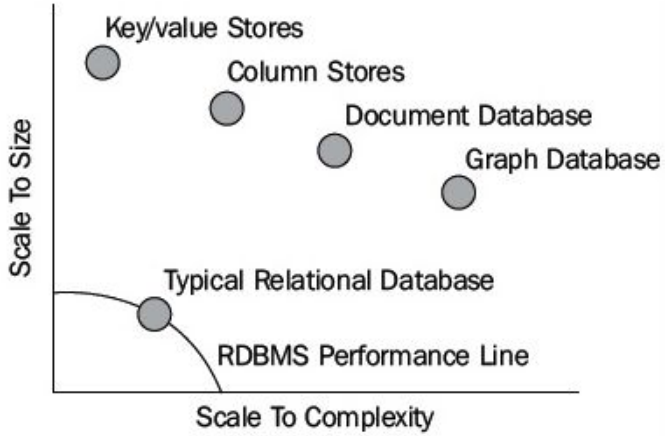
$$\begin{array}{r|cccc} & s_1 & s_2 & s_3 & s_4 \\ \text{how} & 1 & 0 & 0 & 0 \\ \text{much} & 1 & 1 & 0 & 0 \\ \text{wood} & 2 & 2 & 0 & 2 \\ \text{would} & 1 & 1 & 0 & 1 \\ \text{a} & 2 & 2 & 0 & 1 \\ \text{woodchuck} & 2 & 3 & 1 & 2 \\ \text{chuck} & 2 & 3 & 1 & 2 \\ \text{if} & 1 & 1 & 0 & 1 \\ \text{could} & 1 & 2 & 1 & 1 \\ 35 & 0 & 0 & 1 & 0 \\ \text{cubic} & 0 & 0 & 1 & 0 \\ \text{feet} & 0 & 0 & 1 & 0 \\ \text{of} & 0 & 0 & 1 & 1 \\ \text{dirt} & 0 & 0 & 1 & 0 \\ 700 & 0 & 0 & 0 & 1 \\ \text{pounds} & 0 & 0 & 0 & 1 \end{array} \rightarrow A_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 2 & 2 & 0 & 1 \\ 2 & 3 & 1 & 2 \\ 2 & 3 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$



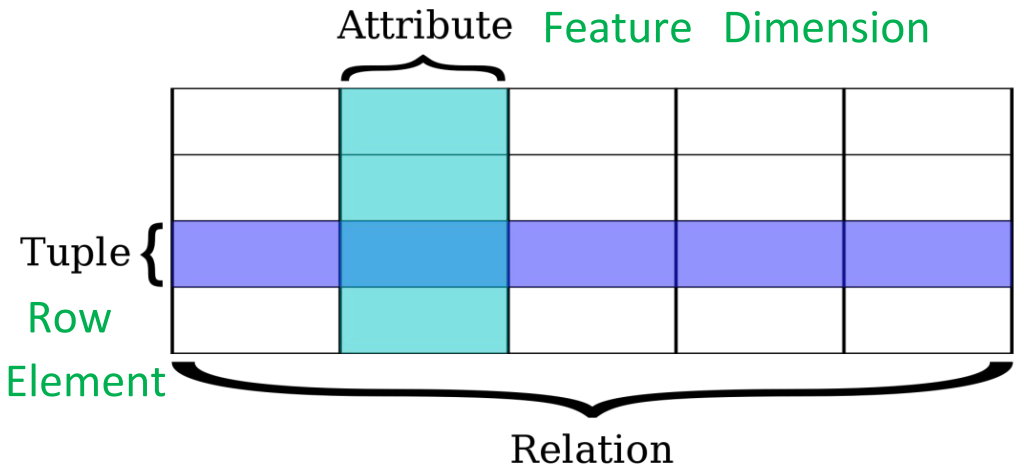
- Alternate
  - Unstructured data
  - Structure on Read (Delay Structure)
  - Non-relational data models



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

Unsupervised Learning

Finding patterns in data.

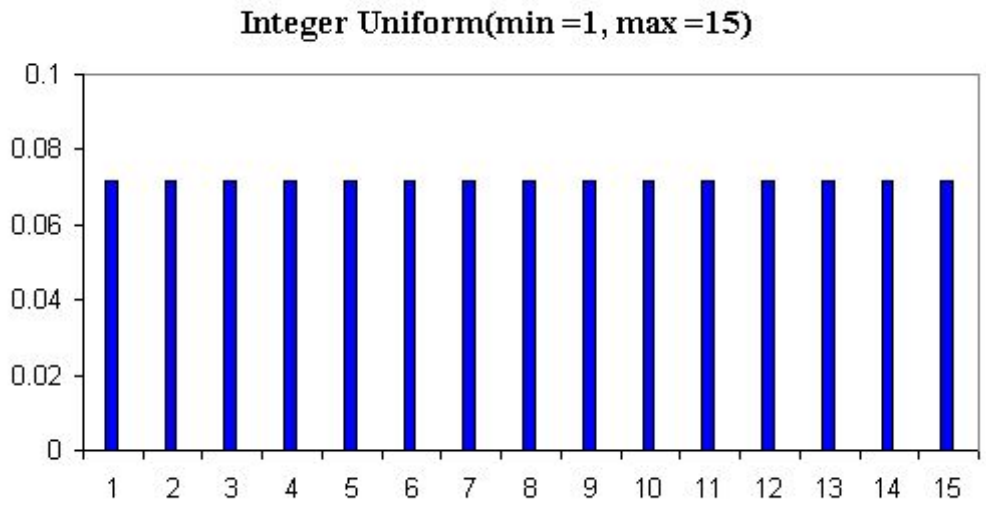
Definitions

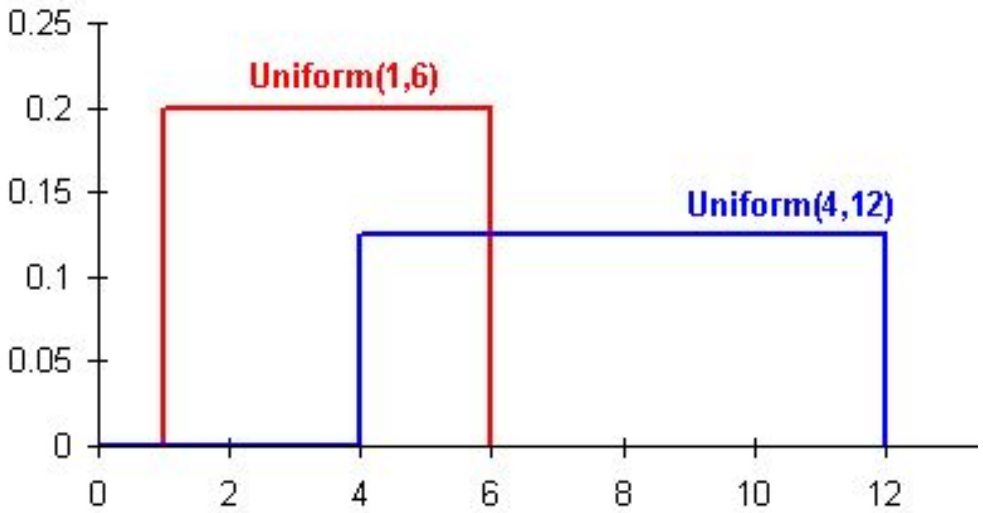
- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.

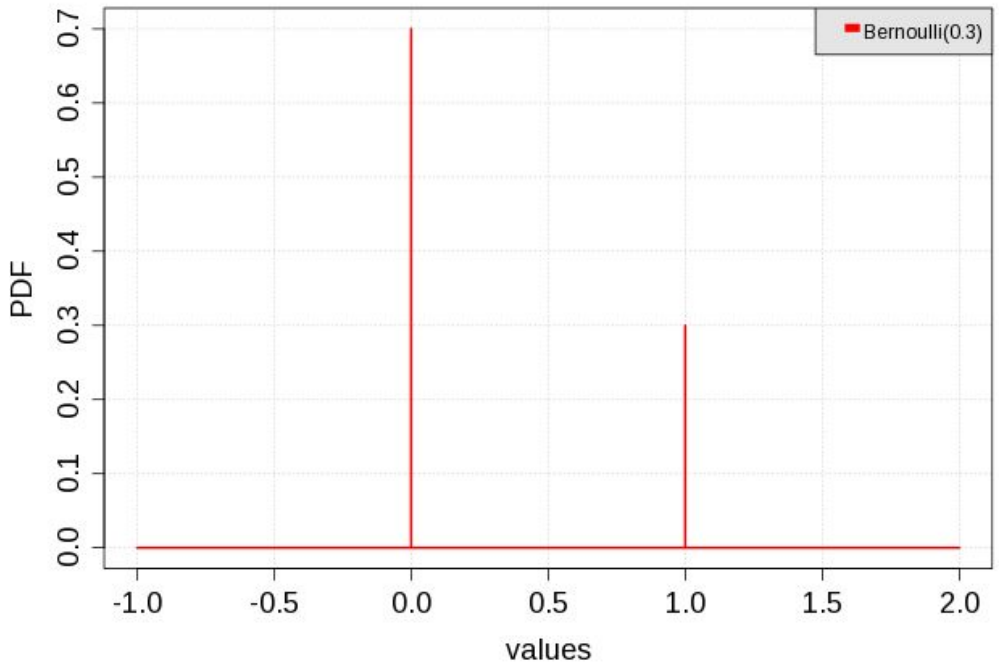
- “Unsupervised”
  - Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution – this distinguishes unsupervised learning from supervised learning and reinforcement learning.

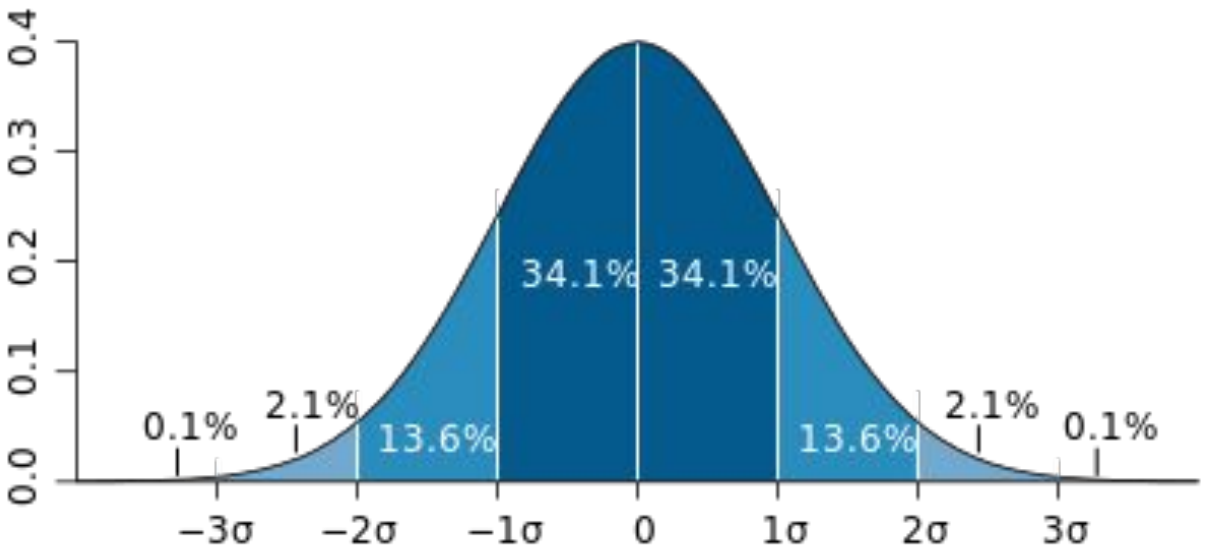
- … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density
  - Summary statistics

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

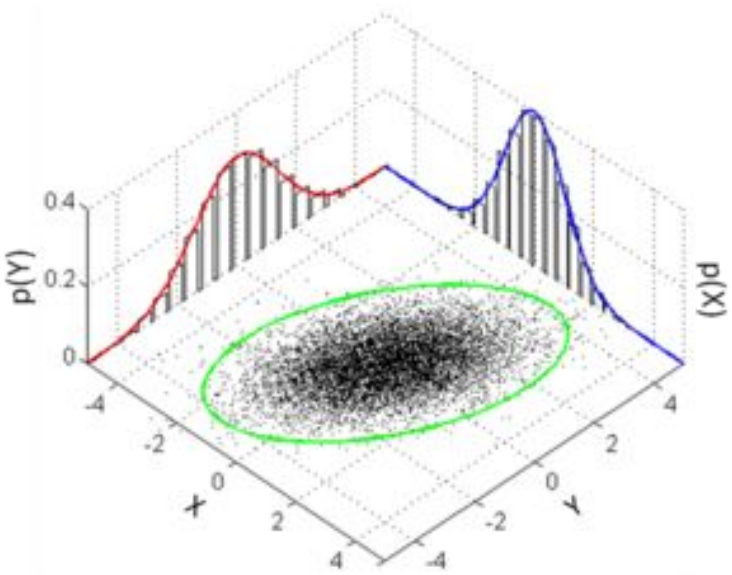


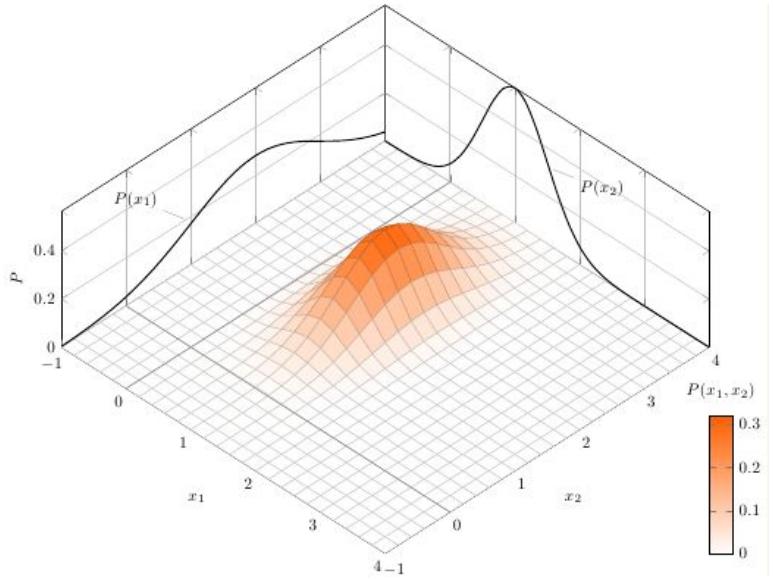


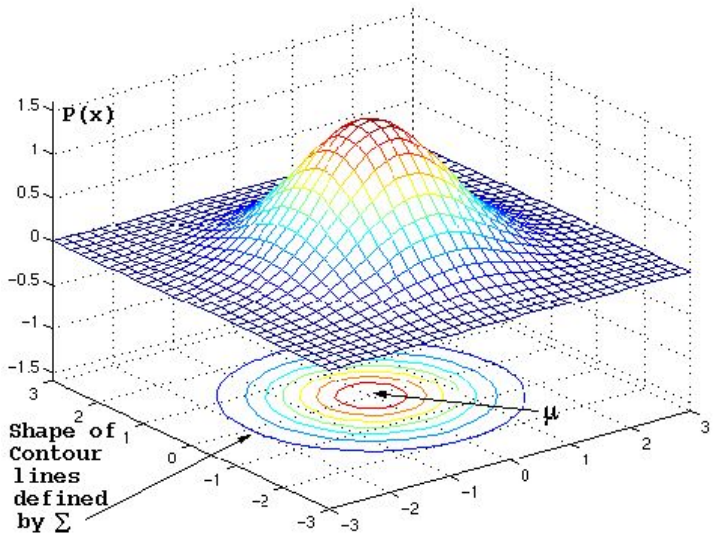




What does a Distribution look like? (p=2)







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.

Q?

Praphul Chandra

Insofe