

Lecture 0

Praphul Chandra

Why are you here?

Indian Folk Lore : Translated in English by John Saxe



*It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind*



*The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a **wall**"*



*The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a **spear**!"*



*The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a **snake**!"*



*The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
" 'Tis clear enough the Elephant
Is very like a **tree**!"*



*The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a **fan**!"*



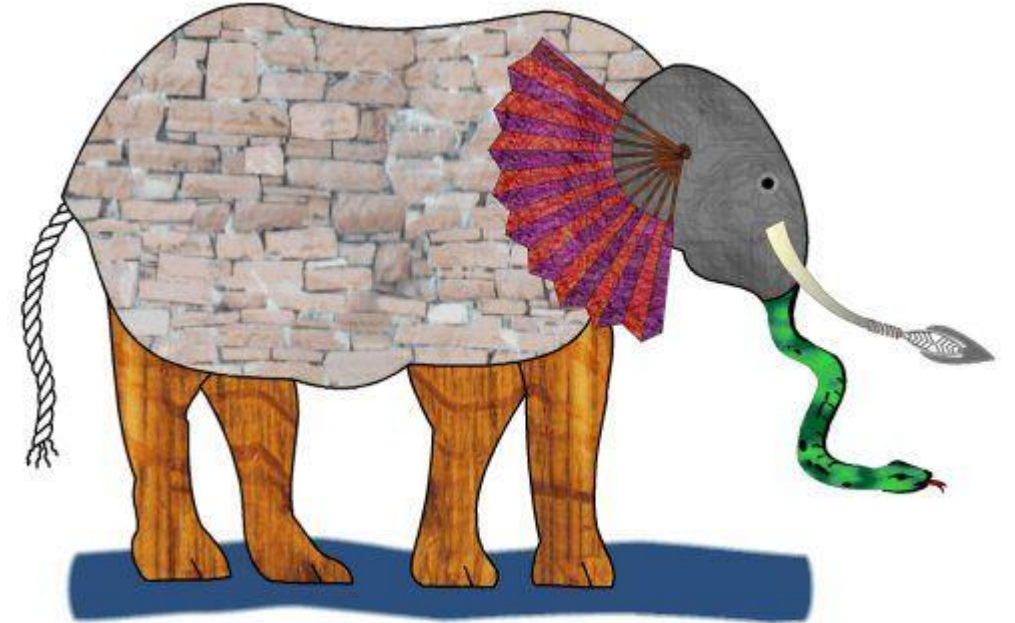
*The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a **rope**!"*



*And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!*

The elephant of Machine Learning

- Symbolists : Derive rules from data
 - Decision Trees
- Connectionists : Brain
 - Neural Networks w/ Backpropagation, Deep Learning
- Evolutionaries : Natural Selection
 - Genetic Programming
- Bayesians : Bayes Theorem
 - Naive Bayes, HMM, Bayesian Networks, MCMC
- Analogizers : Reason with similarities
 - knn, SVM



Other ways of splitting the pie (or the elephant?)

- Supervised vs. Unsupervised
 - Is there a variable of interest? Labelled data? Ground truth?
 - Do you know what you are looking for? Predict?
- Parametric vs. Non-parameteric
 - Does the algorithm have parameters that need to be specified by a human?
 - Hyper-parameter search
- Data Science vs. Artificial Intelligence
 - Is the end-goal to imitate human intelligence? vision? Speech? Natural Language understanding?
 - Is the end-goal custom
- Batch vs. Stream
 - Is all the data from which to learn available?
 - Or will it keep flowing?

Objective

- Learn the tools
- Appreciate each tool
- Develop an instinct for which tool to use when
- Understand the relationship among tools

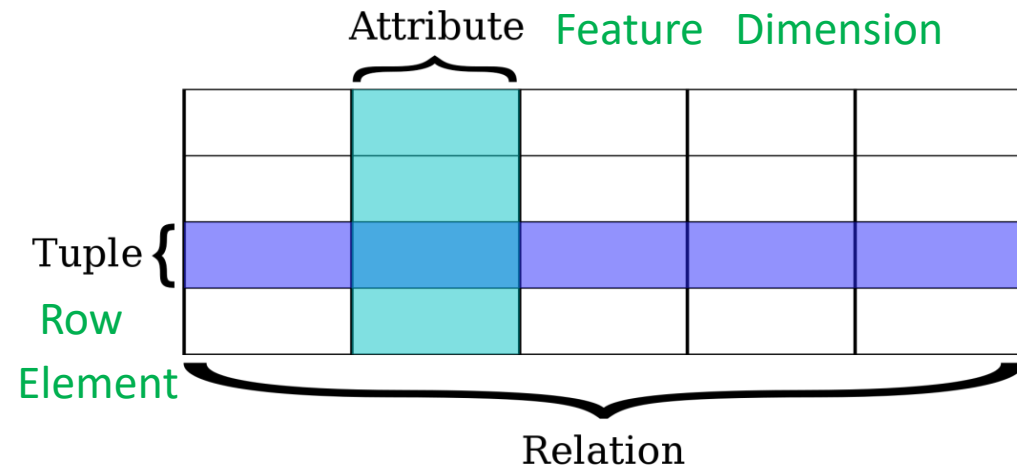


Unsupervised Learning

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?

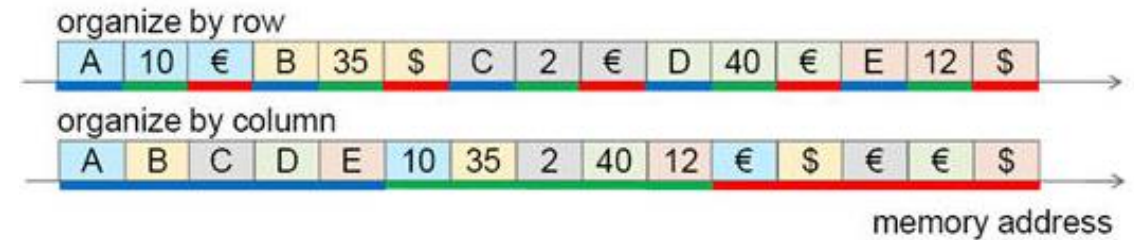


$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$$

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

- 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

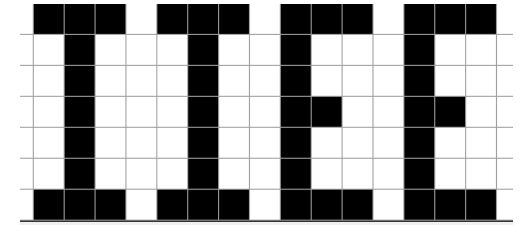
- RDBs
- Spreadsheets
- Matrices
- Very often the data view
- Brittle : Schema exists before data



Relational data model

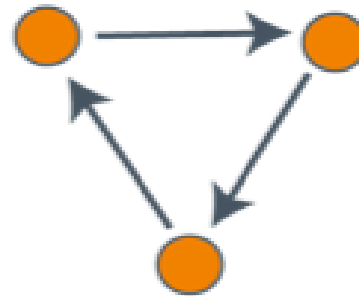
	s_1	s_2	s_3	s_4
how	1	0	0	0
much	1	1	0	0
wood	2	2	0	2
would	1	1	0	1
a	2	2	0	1
woodchuck	2	3	1	2
chuck	2	3	1	2
if	1	1	0	1
could	1	2	1	1
35	0	0	1	0
cubic	0	0	1	0
feet	0	0	1	0
of	0	0	1	1
dirt	0	0	1	0
700	0	0	0	1
pounds	0	0	0	1

$$\rightarrow A_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 2 & 2 & 0 & 1 \\ 2 & 3 & 1 & 2 \\ 2 & 3 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

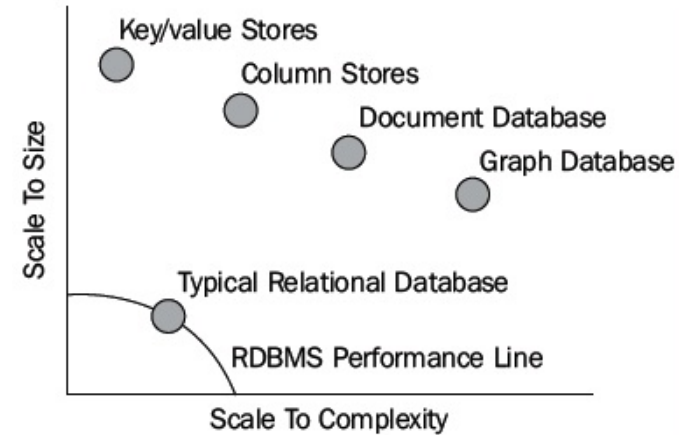


- Alternate

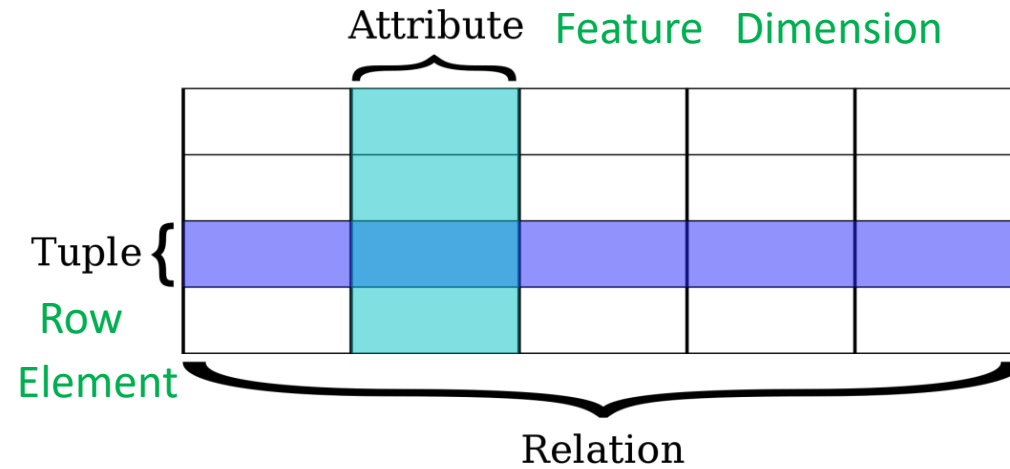
- Unstructured data
- Structure on Read (Delay Structure)
- Non-relational data models



Document data model



What does data look like?



$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$$

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

What does data “really” look like?



If you look carefully, data has patterns.



Unsupervised Learning is about finding patterns in data.

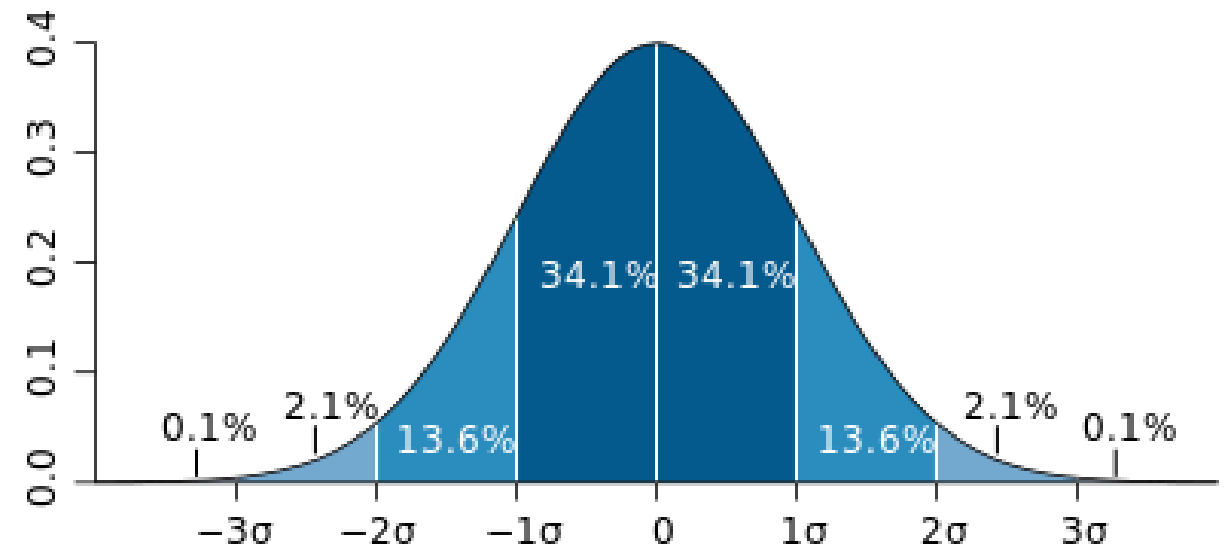
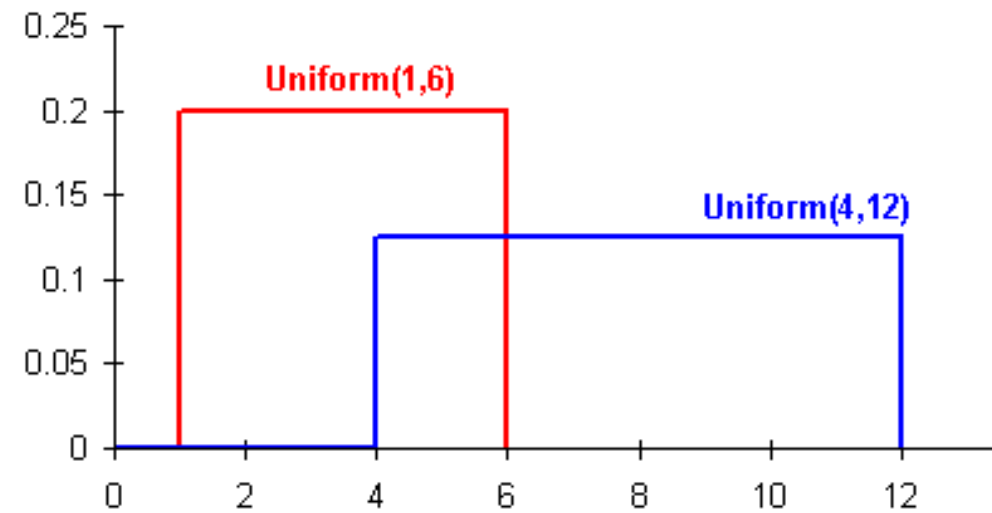
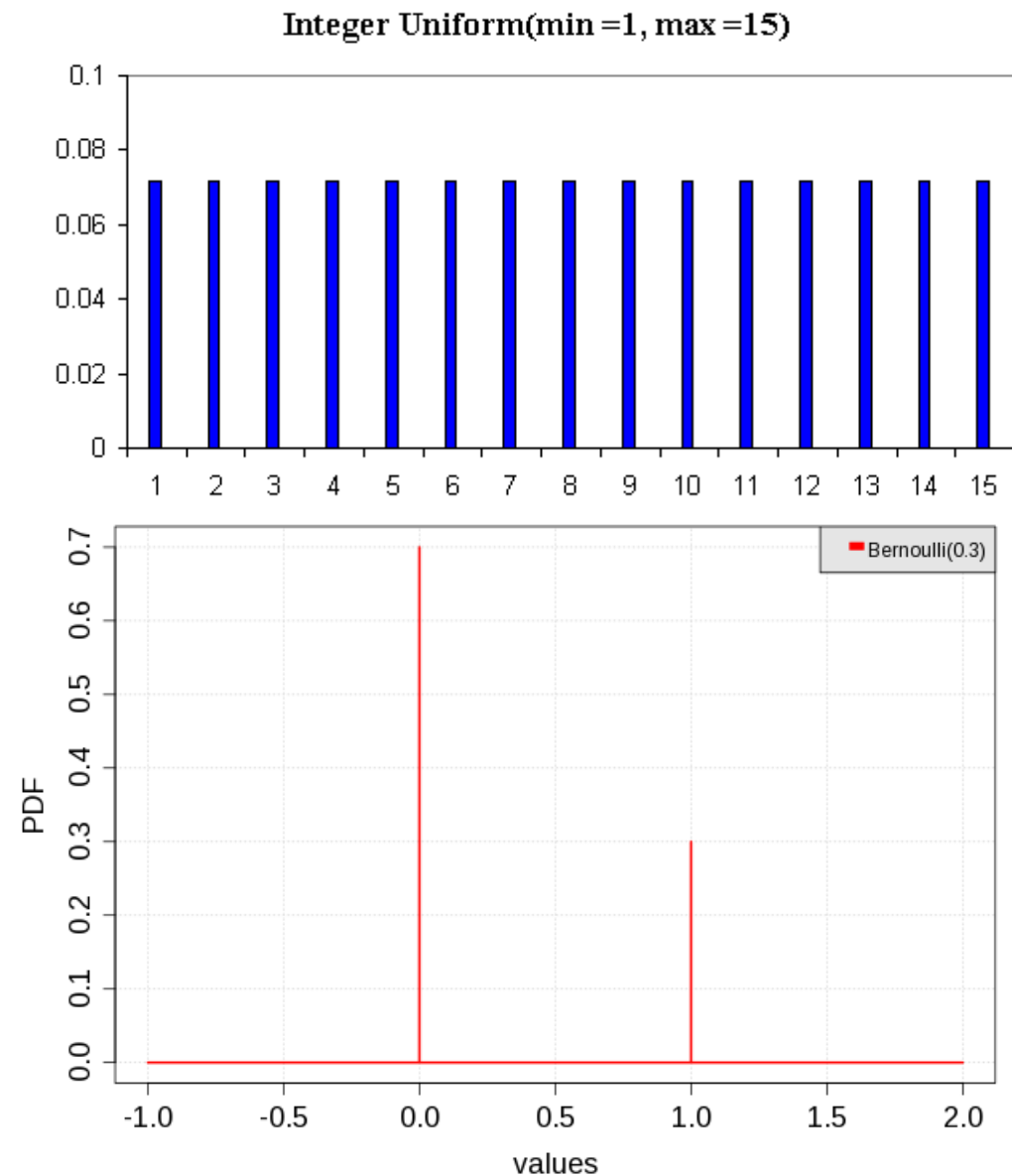
Unsupervised Learning

Finding patterns in data.

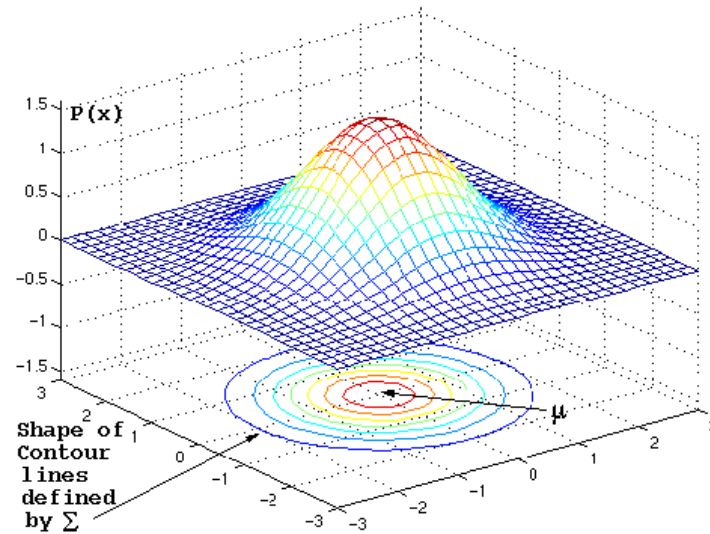
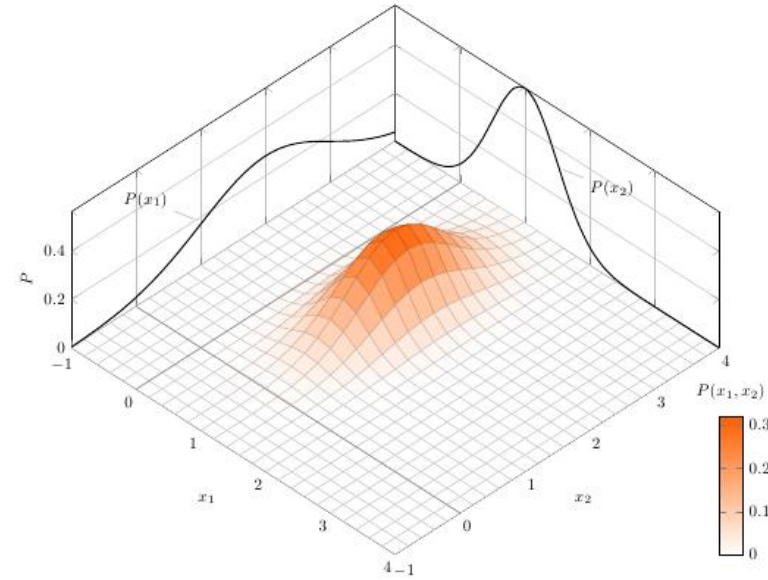
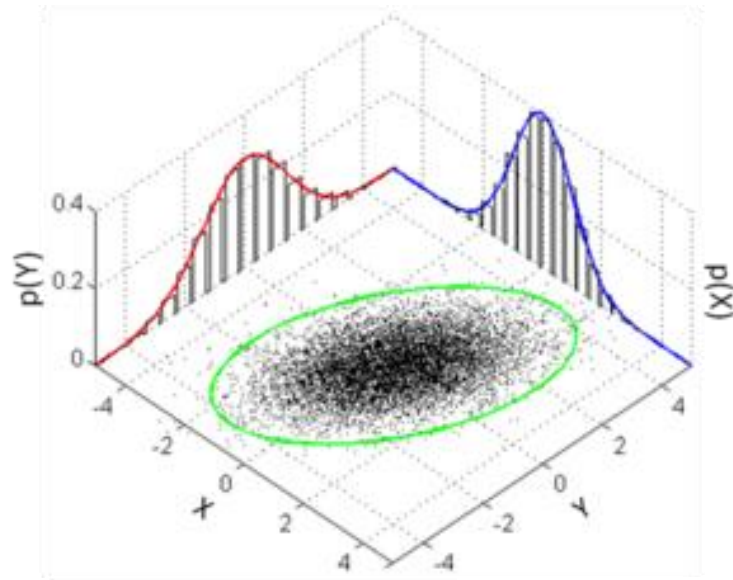
Definitions

- ... 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.

Q?

Praphul Chandra

Insofe