Machine Learning Foundations

Introduction, Terminology and Setup

Harish Guruprasad Ramaswamy
IIT Madras

Outline

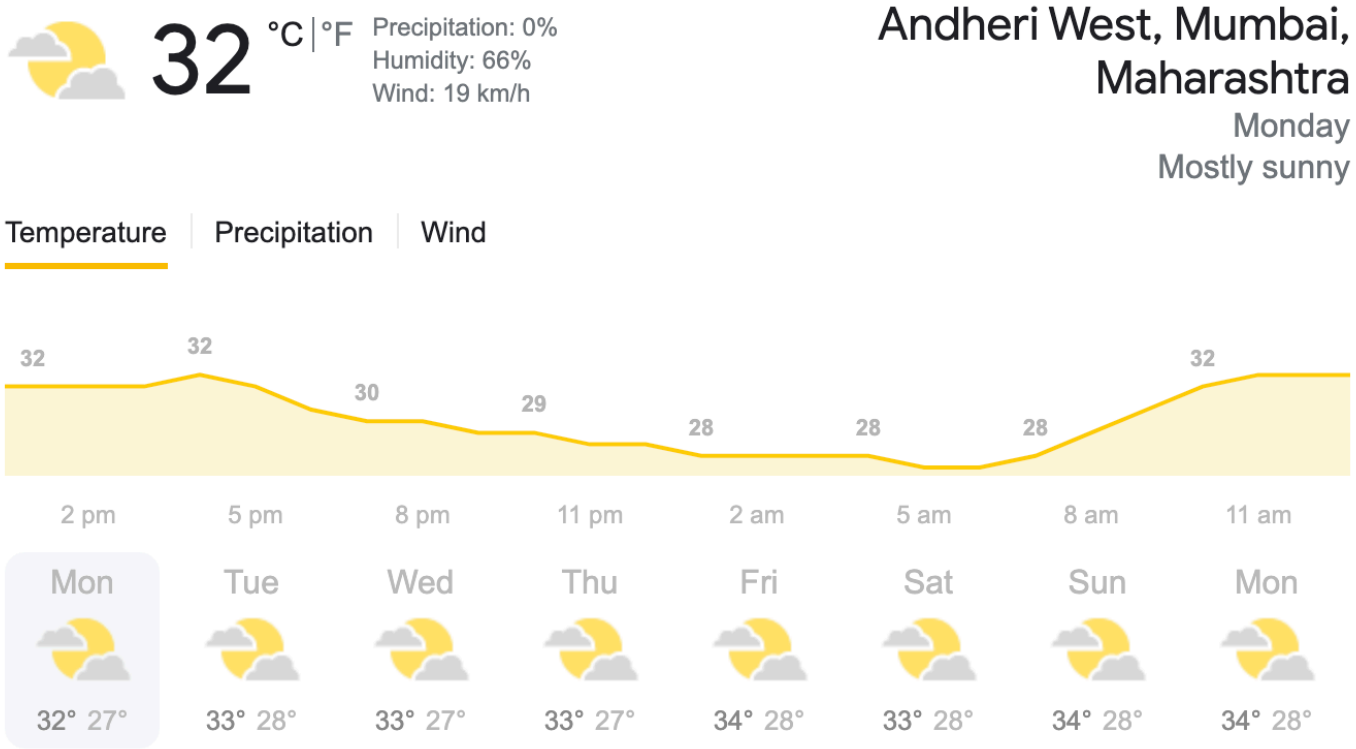
1. **What is Machine Learning?**
2. The Wonders of Machine Learning
3. Data, Models and ML Tasks
4. Supervised Learning
    1. Regression
    2. Classification
5. Unsupervised Learning
    1. Dimensionality Reduction
    2. Density Estimation

Machine Learning Definition

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data.

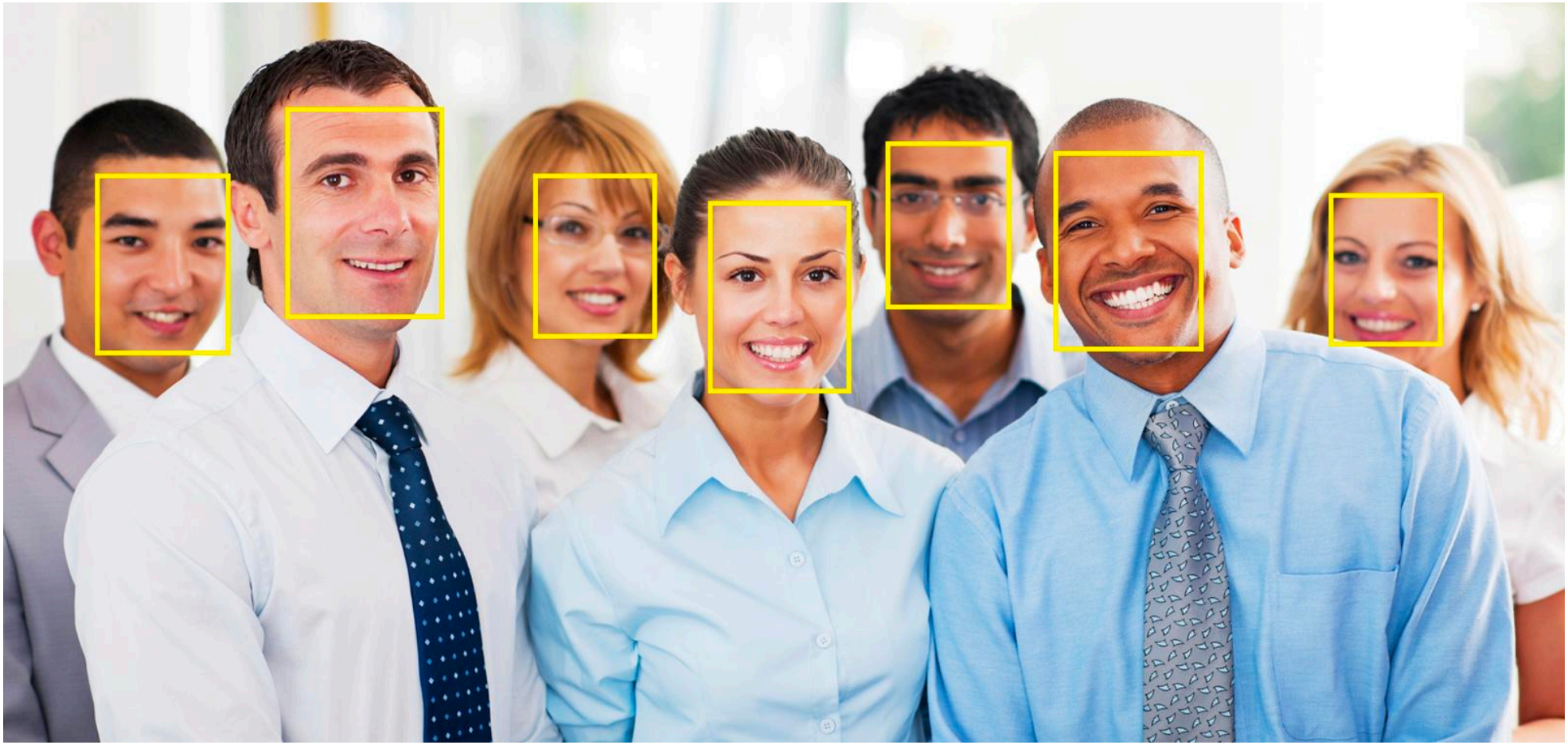
ML Tasks You (Might Have) Performed Today

Weather prediction



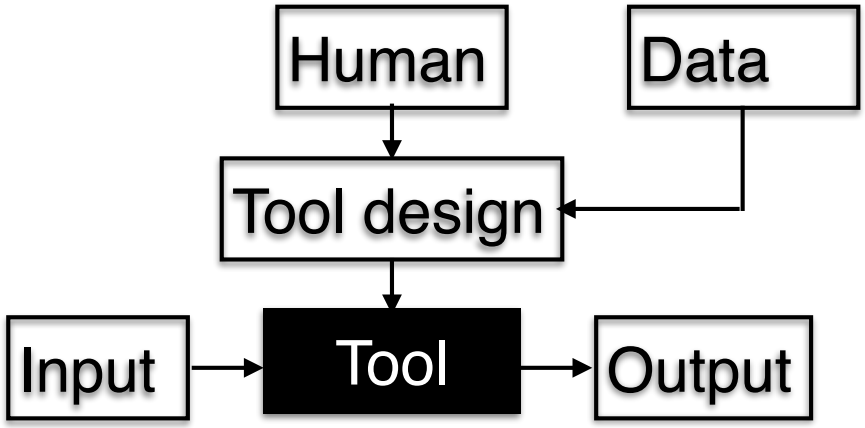
ML Tasks You (Might Have) Performed Today

Face Detection



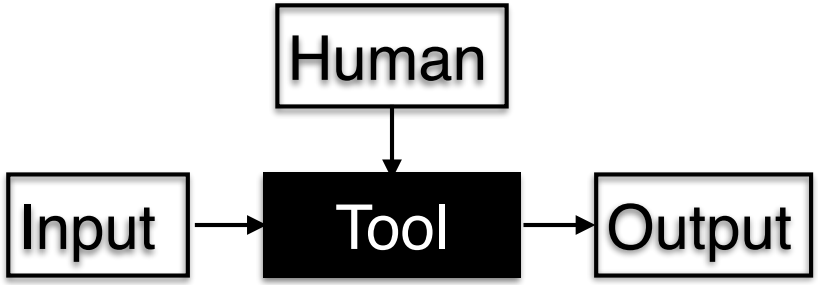
Task Hierarchy

Machine Learning



Programming

Tool usage



Manual Labour

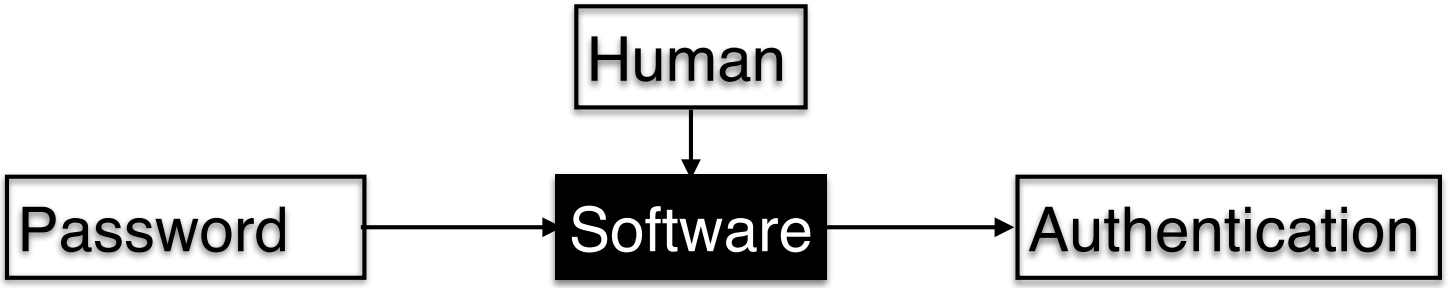
Why and When Machine Learning?

1. Programming/Human Labour Fails
    1. Scale/Speed/Cost of human labor
    2. Inability to express rules using language.
    3. Don’t know the exact rules transforming input to output.
2. Machine Learning can succeed.
    1. Have lots of example data
    2. Have some structural idea on the rules

Task Analysis : Password Verify

Manual labour

Programming



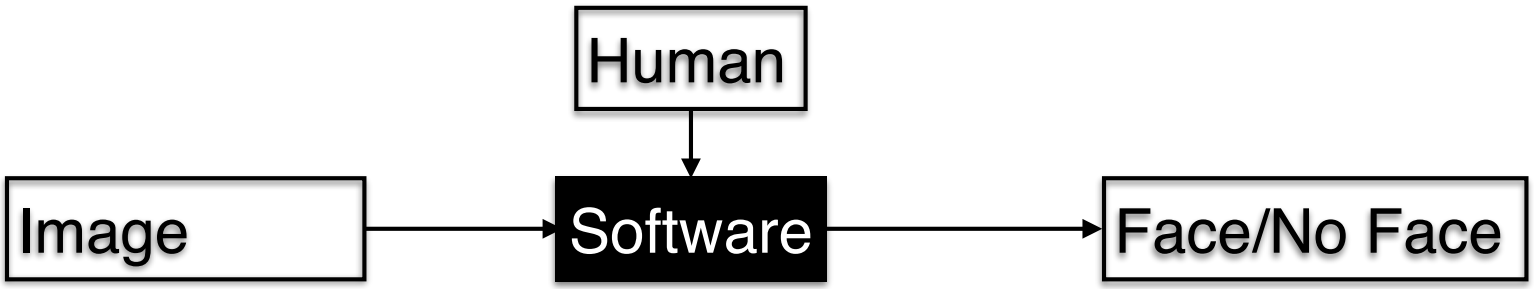
Task Analysis : Password Verify

- Problems with Manual Labour
    - Scale: Having humans check login details of every login is impractical.
- Problems with programming
    - None.
- Machine Learning not required.

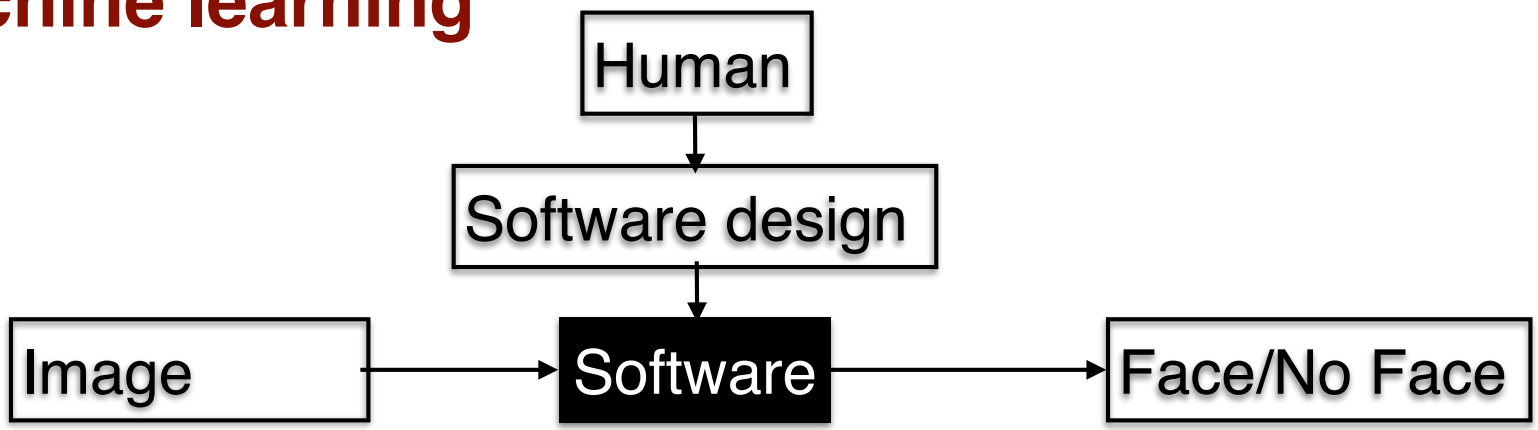
Task Analysis : Face Detection

Manual labour

Programming



Machine learning



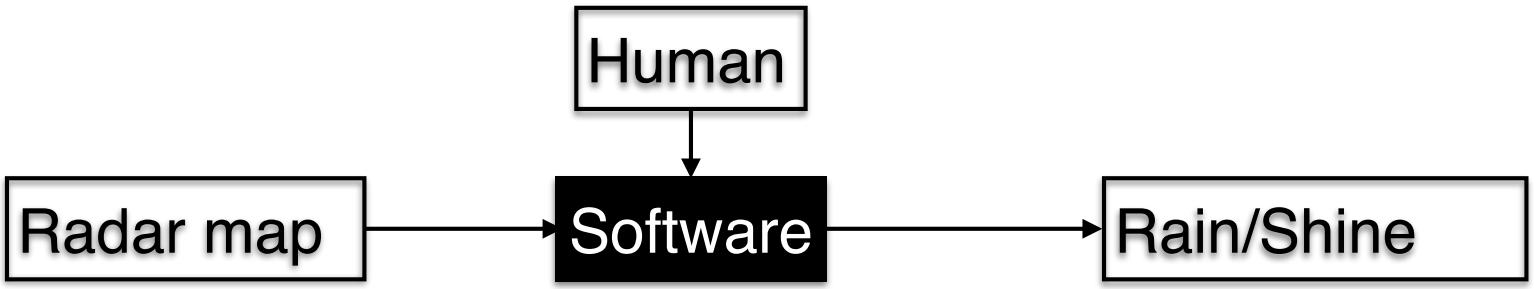
Task Analysis : Face Tagging

- Problems with Manual Labour
    - Scale: Having humans check all faces of every image is impractical.
- Problems with programming
    - Expressing face/not face in code is impossible.
- Case for Machine Learning
    - Lots of images available.

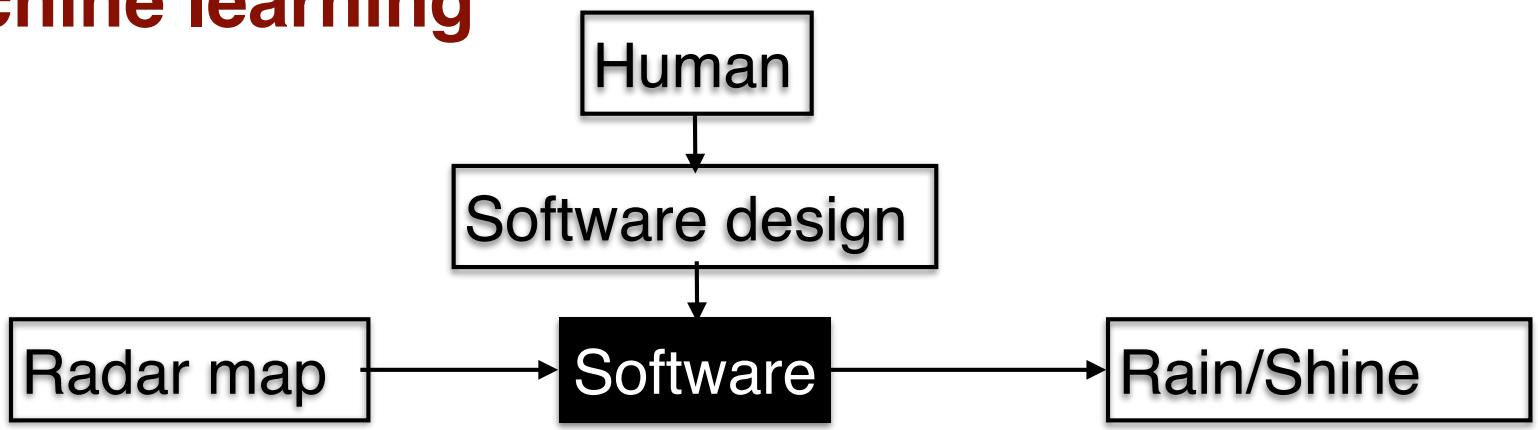
Task Analysis : Weather Prediction

Manual labour

Programming



Machine learning



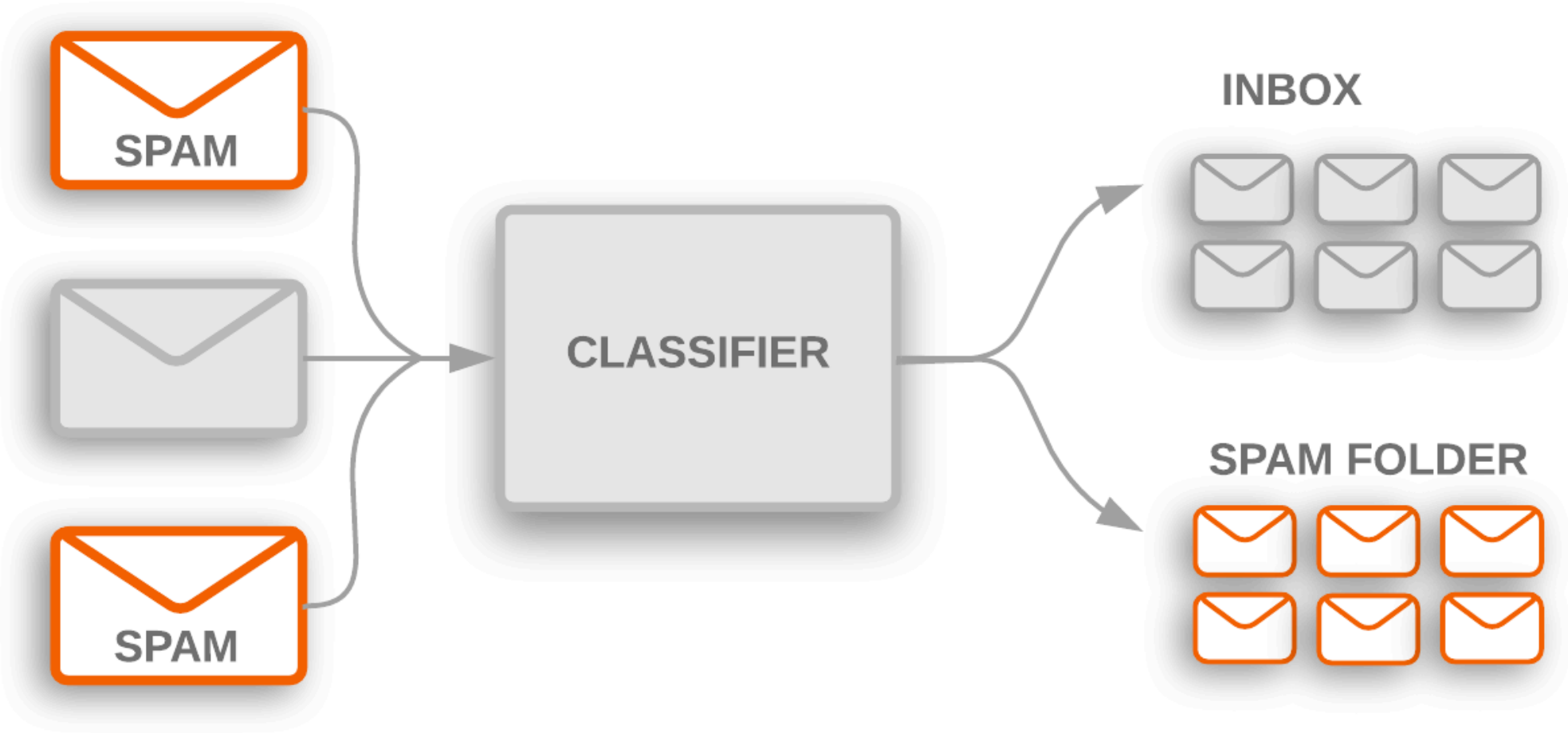
Task Analysis : Weather Prediction

- Problems with Manual Labour
    - Humans just do not know the full rules and can't process that much information.
- Problems with programming
    - Cannot code unknown rules.
- Case for Machine Learning
    - Lots of weather data available.

Outline

1. What is Machine Learning??
2. **The Wonders of Machine Learning**
3. Data, Models and ML Tasks
4. Supervised Learning
    1. Regression
    2. Classification
5. Unsupervised Learning
    1. Dimensionality Reduction
    2. Density Estimation

Machine Learning in your Inbox



Machine Learning in your Shopping Cart

Frequently Bought Together

Price For All Three: $258.02

Add all three to Cart

- **This item:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie
- Pattern Recognition and Machine Learning (Information Science and Statistics) by Christopher M. Bishop
- Pattern Classification (2nd Edition) by Richard O. Duda

Customers Who Bought This Item Also Bought

All of Statistics: A Concise Course in Statist... by Larry Wasserman

(8) $60.00

Pattern Classification (2nd Edition) by Richard O. Duda

(27) $117.25

Data Mining: Practical Machine Learning Tools an... by Ian H. Witten

(29) $41.55

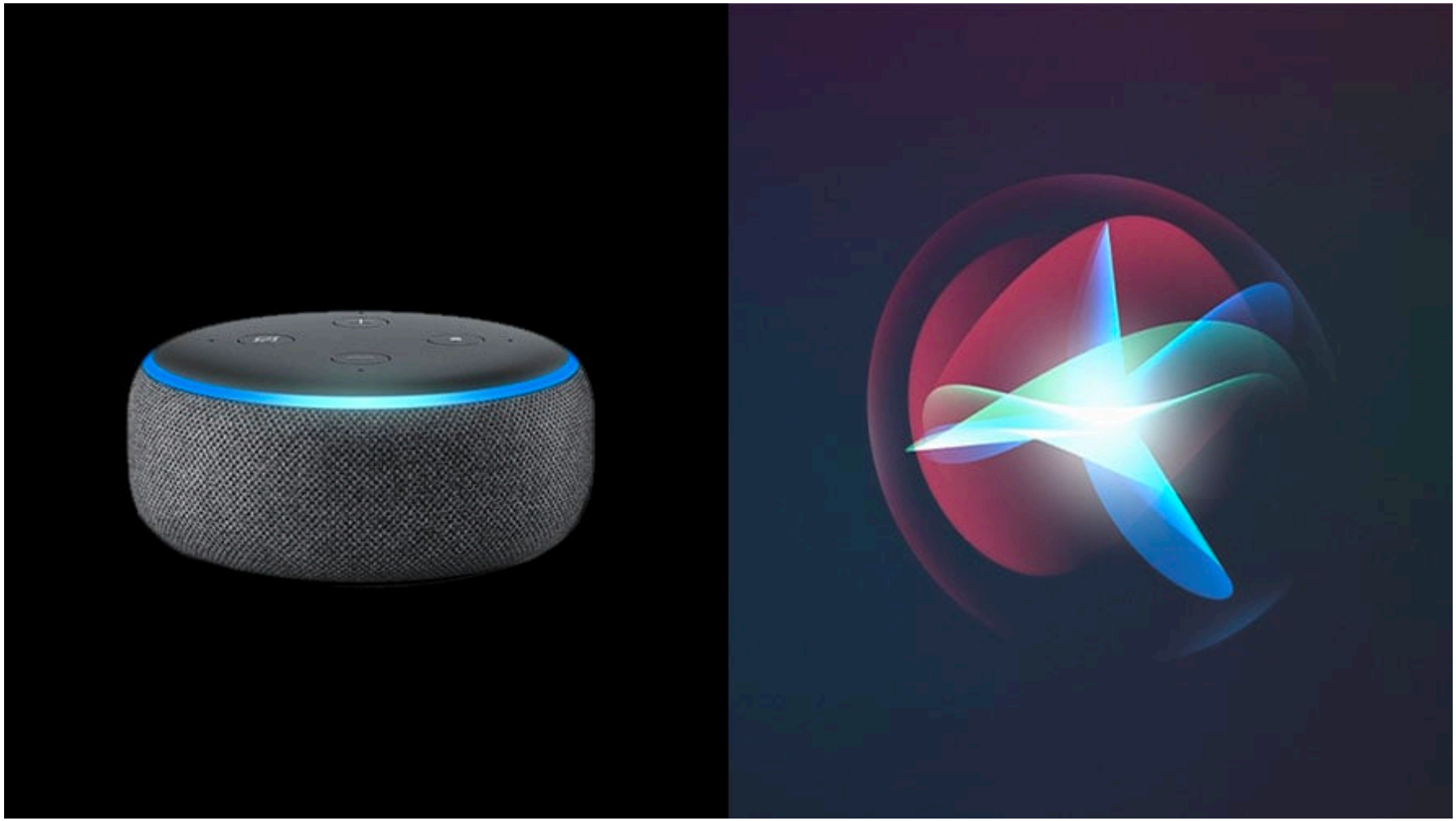
Bayesian Data Analysis, Second Edition (Texts in... by Andrew Gelman

(10) $56.20

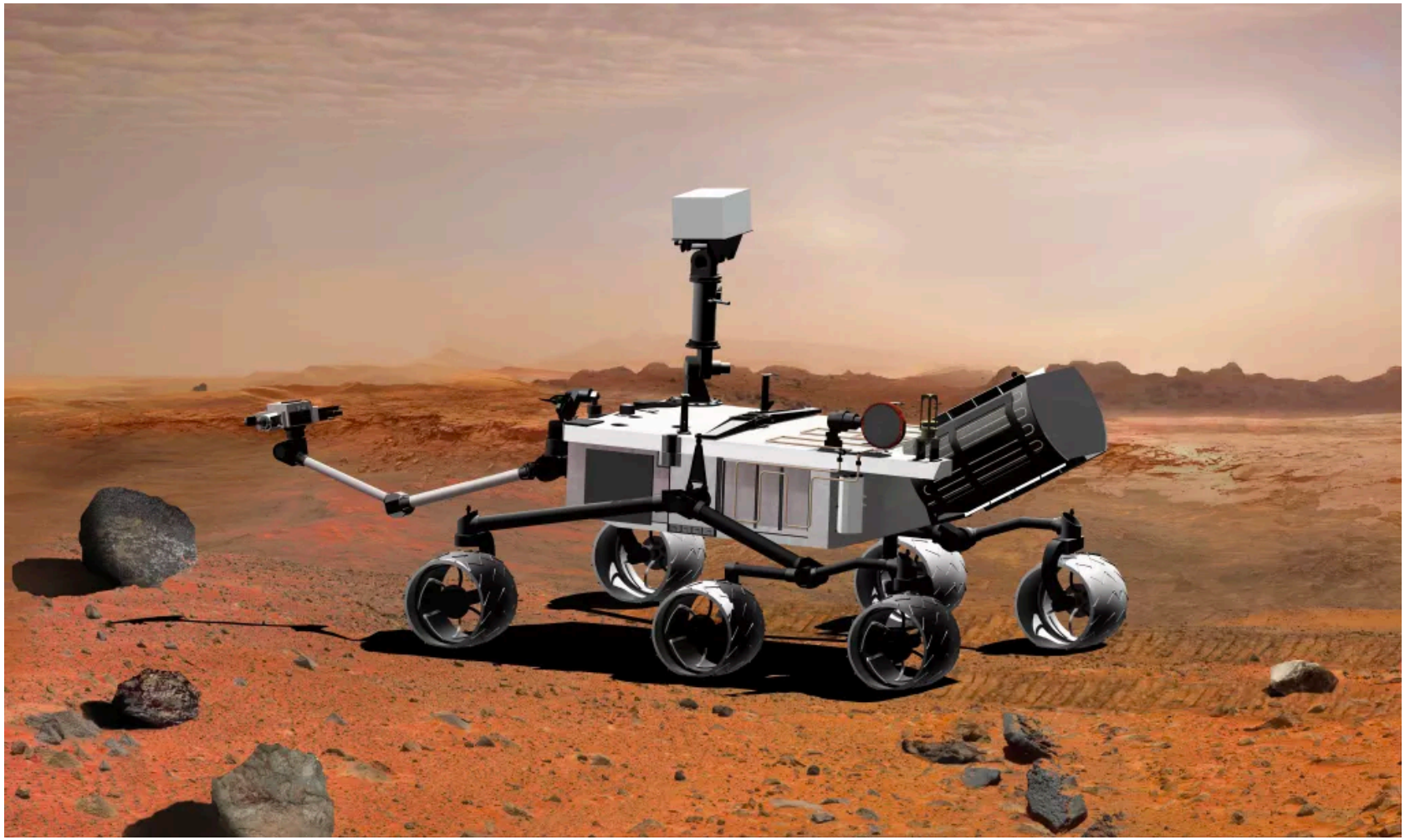
Data Analysis Using Regression and Multilevel /... by Andrew Gelman

(13) $39.59

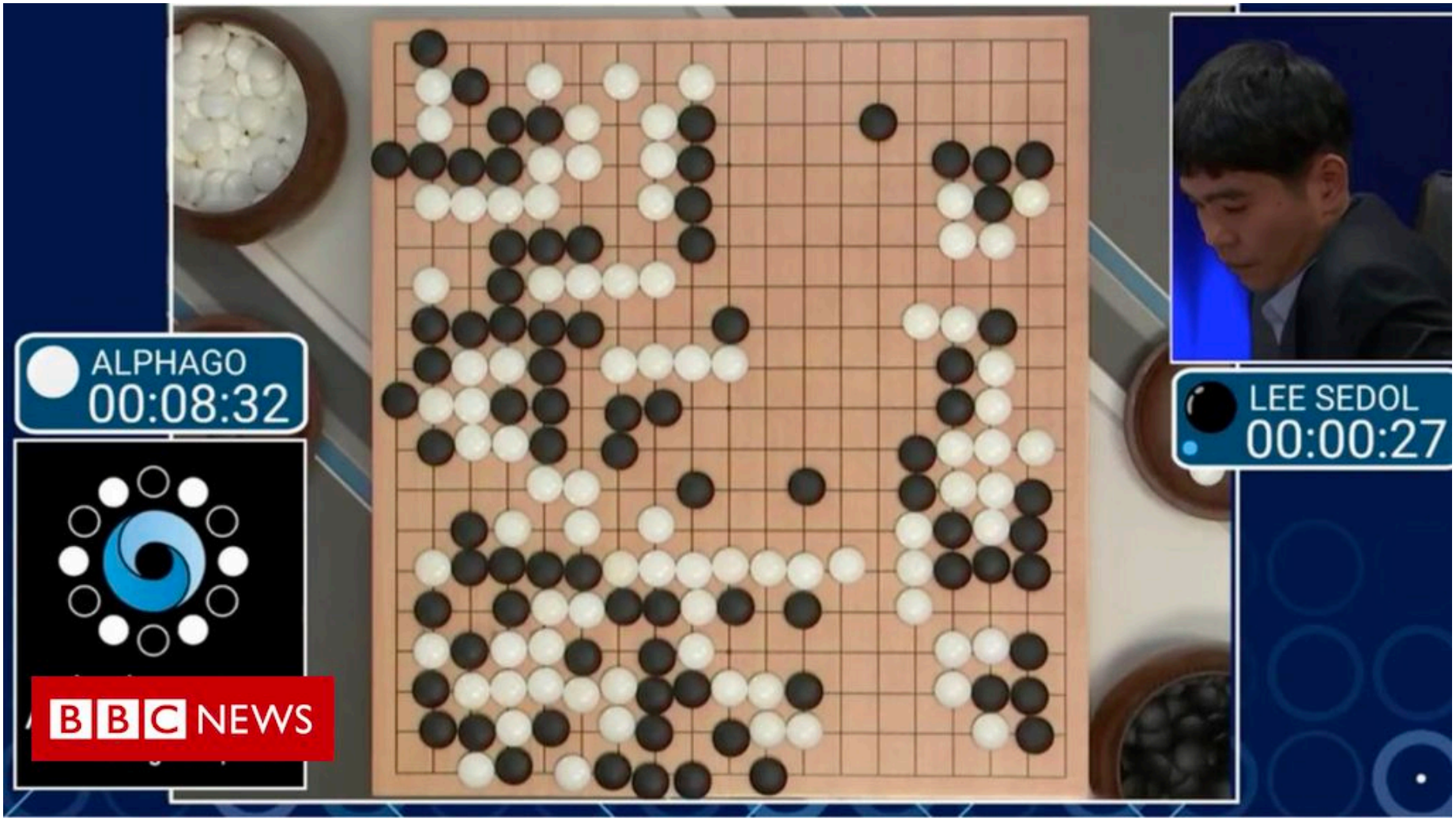
Machine Learning in your Smart Assistant



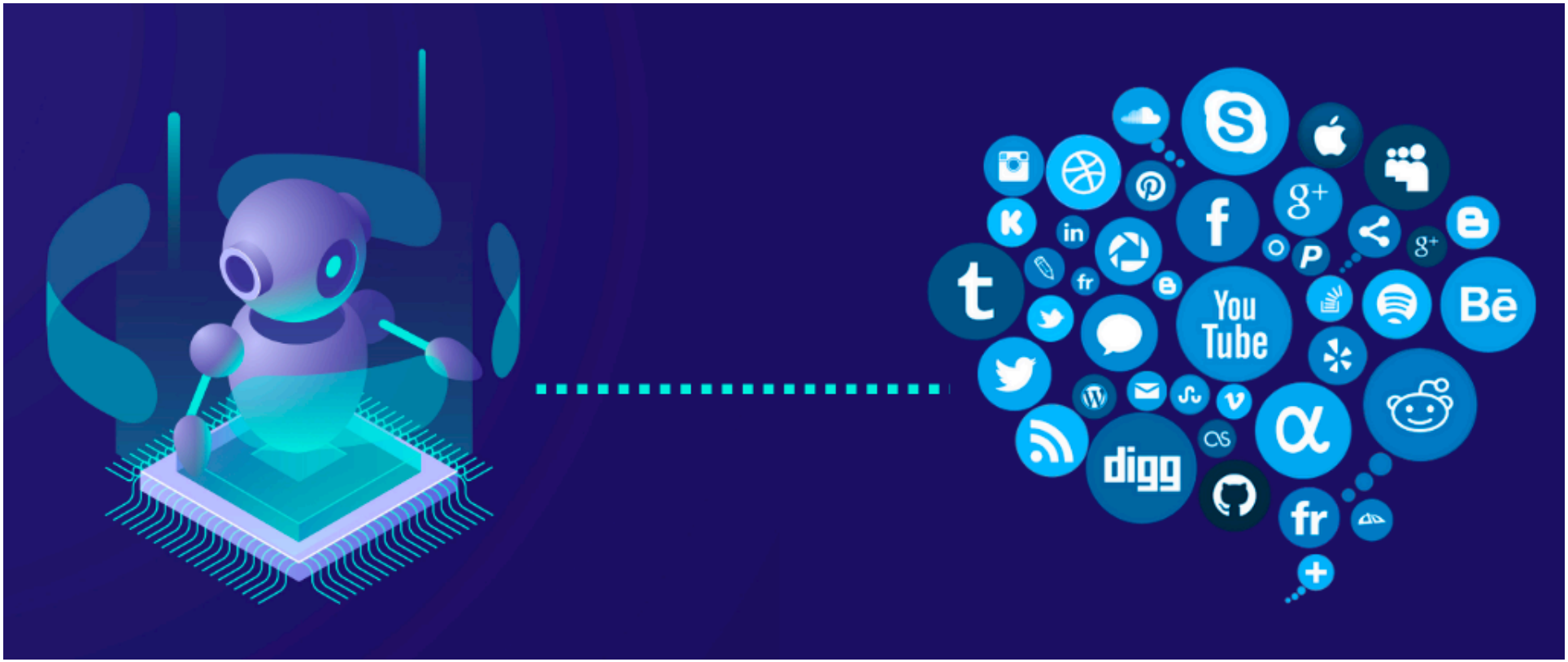
Machine Learning in Robot AIs



Machine Learning in Games



Machine Learning in Marketing



Outline

1. What is Machine Learning??
2. The Wonders of Machine Learning
3. **Data, Models and ML Tasks**
4. Supervised Learning
    1. Regression
    2. Classification
5. Unsupervised Learning
    1. Dimensionality Reduction
    2. Density Estimation

What is Data?

Data is a collection of vectors.

E.g.

3	9	1.9	5.0	House 1
2	7	2.1	3.2	House 2
4	12	2.8	6.6	House 3
5	16	0.9	9.8	House 4
5	15	3.1	8.5	House 5
4	11	1.6	6.9	House 6

Metadata is information on the data.

E.g. : (# rooms, Area in 100 sq.ft, Distance to metro in km, Price in 10 lakhs)

What is a Model?

A model is a mathematical simplification of reality.

Some examples:

The Ideal Gas model

Inverse square law for gravitational attraction

Moore's Law for semiconductors

Cobb–Douglas model in Economics

"All models are wrong, but some are useful"

George Box

Types of Models in ML

- Predictive Model
  - Regression Model
  - Classification Model
  - …
- Probabilistic Model
  -

Predictive Models

Regression Model

Model the price of a house based on its area and distance to metro.

Example good model:

Price = 0.5*Area - Distance

Predictive Models

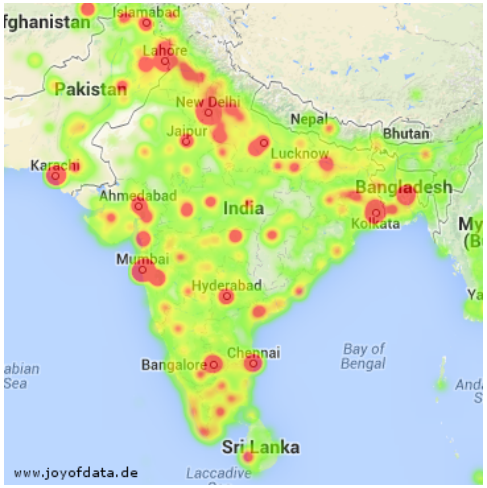
Classification Model

Model whether a house is closer than 2kms to a metro based on price and area

Example good model:

Answer = Close if 2*ROOMS-PRICE < 1
Far otherwise

Probabilistic Models



What is the probability that a randomly chosen person is in lat-long : (25N,30E) ?

"A formless void transforms total mysteries"

RECEIVE MORE WISDOM...

Tweet the wisdom

What is the probability that a given tweet was generated by Mr. Chopra?

Learning Algorithms

Learning Algorithms: Data → Models

Choose from a collection of models, with same structure but different **parameters**.

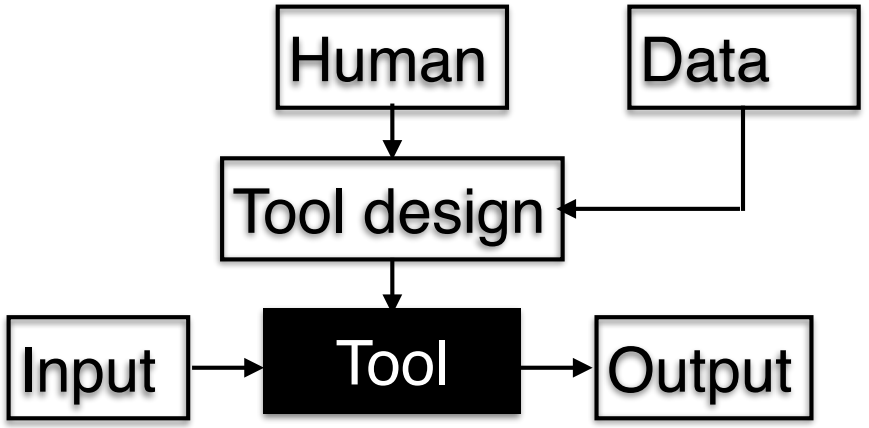
E.g.
Price = a*(area) + b*(# rooms) + c*(distance to metro)

Parameters: a,b,c

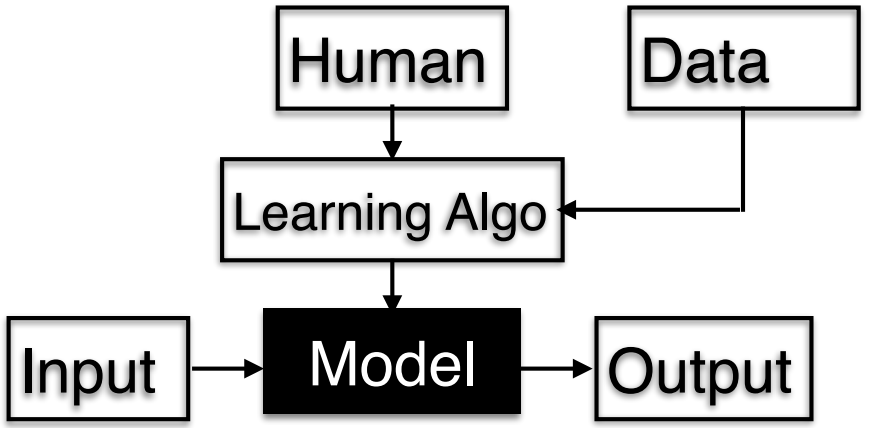
Use data to get the "**best**" parameters

Machine Learning Tasks Revisited

Machine Learning



Machine Learning



Outline

1. What is Machine Learning??
2. The Wonders of Machine Learning
3. Data, Models and ML Tasks
4. **Supervised Learning**
    1. Regression
    2. Classification
5. Unsupervised Learning
    1. Dimensionality Reduction
    2. Density Estimation

Notation

$x^1 = [1, 2, 3]$

$x^2 = [7, 8, 9]$

$x_2^2 = 8$

$\begin{pmatrix} 1.3 \\ -7.6 \\ 5.9 \end{pmatrix} \in \mathbb{R}^3$

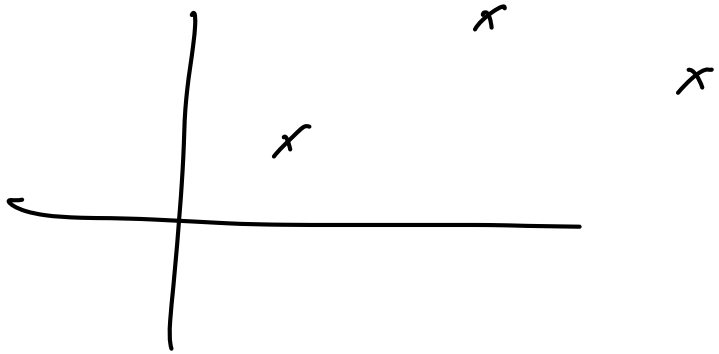
- $\mathbb{R}$: real numbers, $\mathbb{R}_+$: Positive reals, $\mathbb{R}^d$: d-dimensional vector of reals.
- $\mathbf{x}$: vector. x_j: j^{th} co-ordinate. $\|\mathbf{x}\|$: Length of vector $\mathbf{x}$.
- $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$: Collection of n vectors.
- x_j^i : j^{th} co-ordinate of i^{th} vector.
- $(x_1)^2$: Square of the first co-ordinate of the vector $\mathbf{x}$
- $\mathbf{1}(2 \text{ is even}) = 1$, $\mathbf{1}(2 \text{ is odd}) = 0$.

$x = \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix}$

$x_3 = 7$

$\|x\|^2 = x_1^2 + x_2^2 + \ldots + x_d^2$

Supervised Learning



- Supervised learning is curve-fitting.
- Given $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$ $\mathbf{x}^i \in \mathbb{R}^d$
- Find a model f such that $f(\mathbf{x}^i)$ is 'close' to y^i

f is a function

Regression

- E.g. Predict house price from room, area, distance.
- Training data: $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$ ($\mathbf{x} \in \mathbb{R}^3$, $y \in \mathbb{R}$)
- $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$
- Algorithm outputs a model $f : \mathbb{R}^d \to \mathbb{R}$
- Loss[f] $= \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}^i) - y^i)^2$ = Squared loss
- $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = \sum_{j=1}^{d} w_j x_j + b$

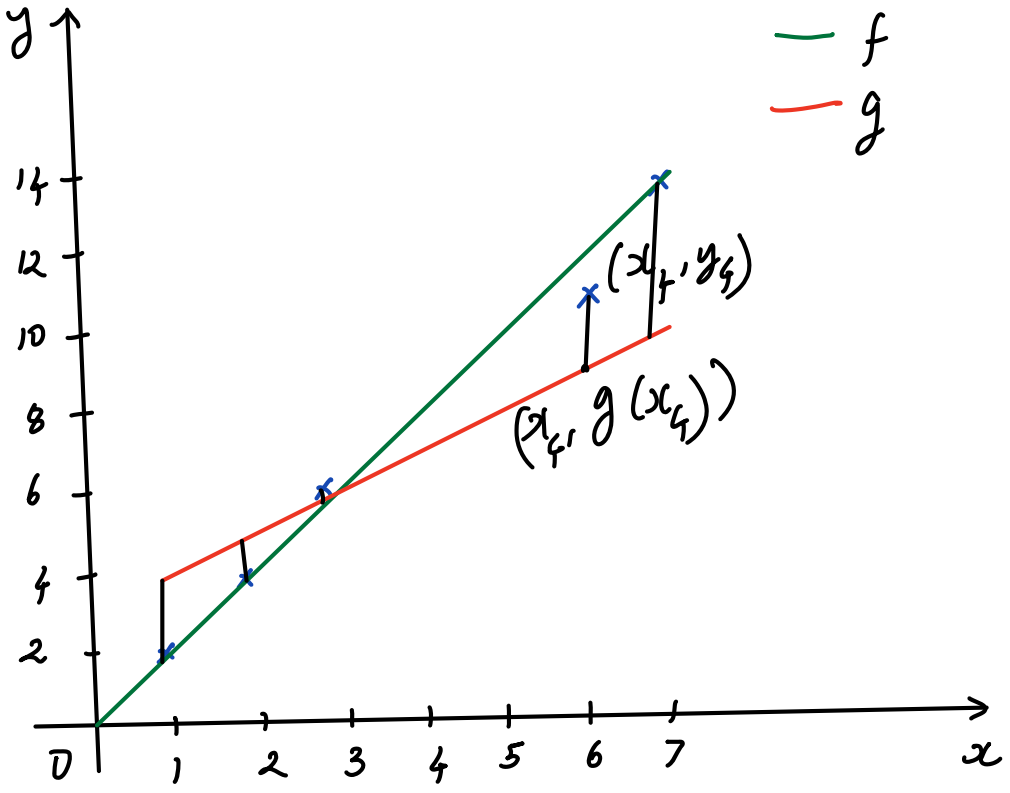
Linear Parameterisation

$= w_1$ (# rooms) + w_2 (area) + w_3 (distance) + b

Regression Illustration 1

$d = 1$

x	y	f	g
[1]	2.1		
[2]	3.9		
[3]	6.2		
[6]	11.5		
[7]	13.9		



$f(x) = 2x_1$

$g(x) = x_1 + 3$

Loss [f] :

=

Loss [g] .

Regression Illustration 2

Rooms	Area	Distance	Price
3	9	1.9	**5.0**
2	7	2.1	**3.2**
4	12	2.8	**6.6**
5	16	0.9	**9.8**
5	15	3.1	**8.5**
4	11	1.6	**6.9**

$f = 2 * \text{Rooms} - 0.5 * \text{dist}$

$g = \text{Rooms} + 2 * \text{distan}$

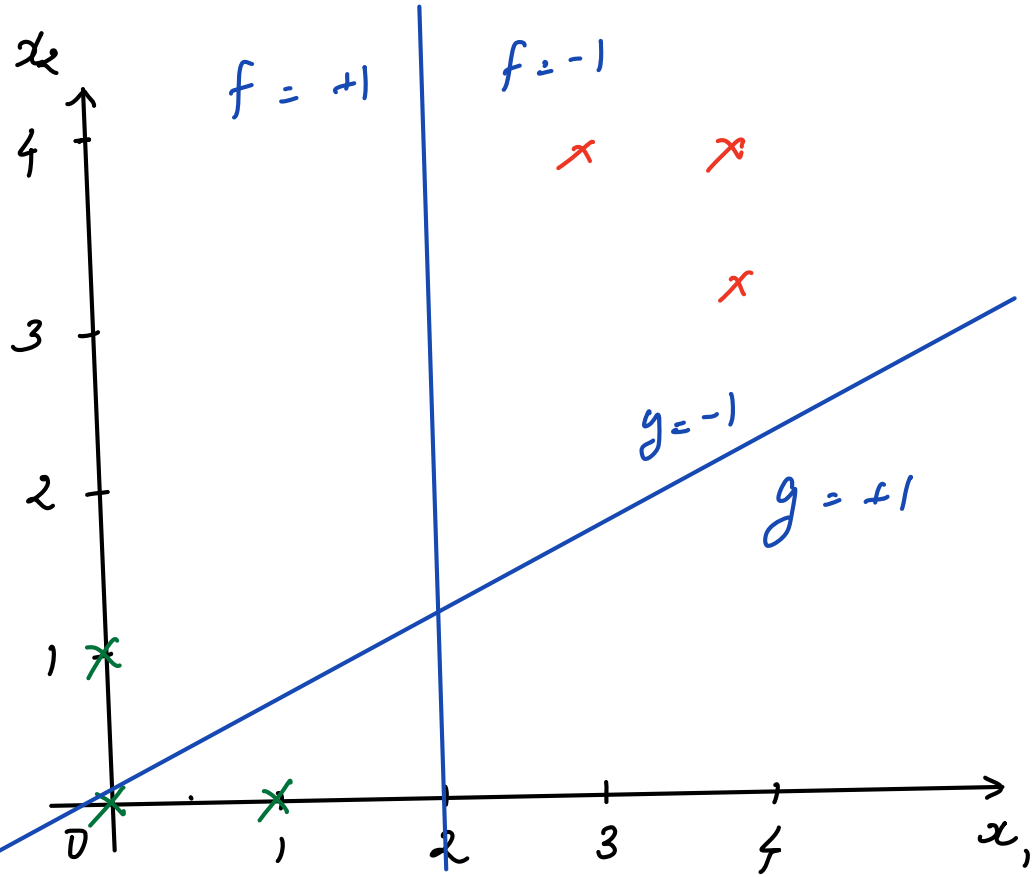
Classification

- E.g. Predict if rooms>3 from area and price.
- Training data: $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$
- $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \{+1, -1\}$
- Algorithm outputs a model $f : \mathbb{R}^d \rightarrow \{+1, -1\}$
- Loss[f] $= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(f(\mathbf{x}^i) \neq y^i)$ = Fraction of training data classified wrongly by f
- $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$

Linear separator

Classification Illustration 1

x	y	f	g
[0, 0]	+1		
[0, 1]	+1		
[1, 0]	+1		
[4, 4]	-1		
[3, 4]	-1		
[4, 3]	-1		



$f(x) = \text{sign}(2 - x_1)$

$g(x) = \text{sign}(x_1 - 2x_2)$

$\text{Loss}[f] = \frac{1}{6}(0) = 0$

$\text{Loss}[g] = \frac{1}{6}(1) = 1/6$

Classification Illustration 2

Area	Price	Rooms
9	5.0	-1
7	3.1	-1
12	6.9	+1
16	9.7	+1
15	8.5	+1
11	7.1	+1

Rooms=1 or 2 or 3 is encoded as -1.
Rooms > 3 is encoded as +1.

f	g	h
-1	-1	-1
-1	-1	-1
+1	+1	-1
+1	+1	+1
+1	+1	-1
+1	+1	-1

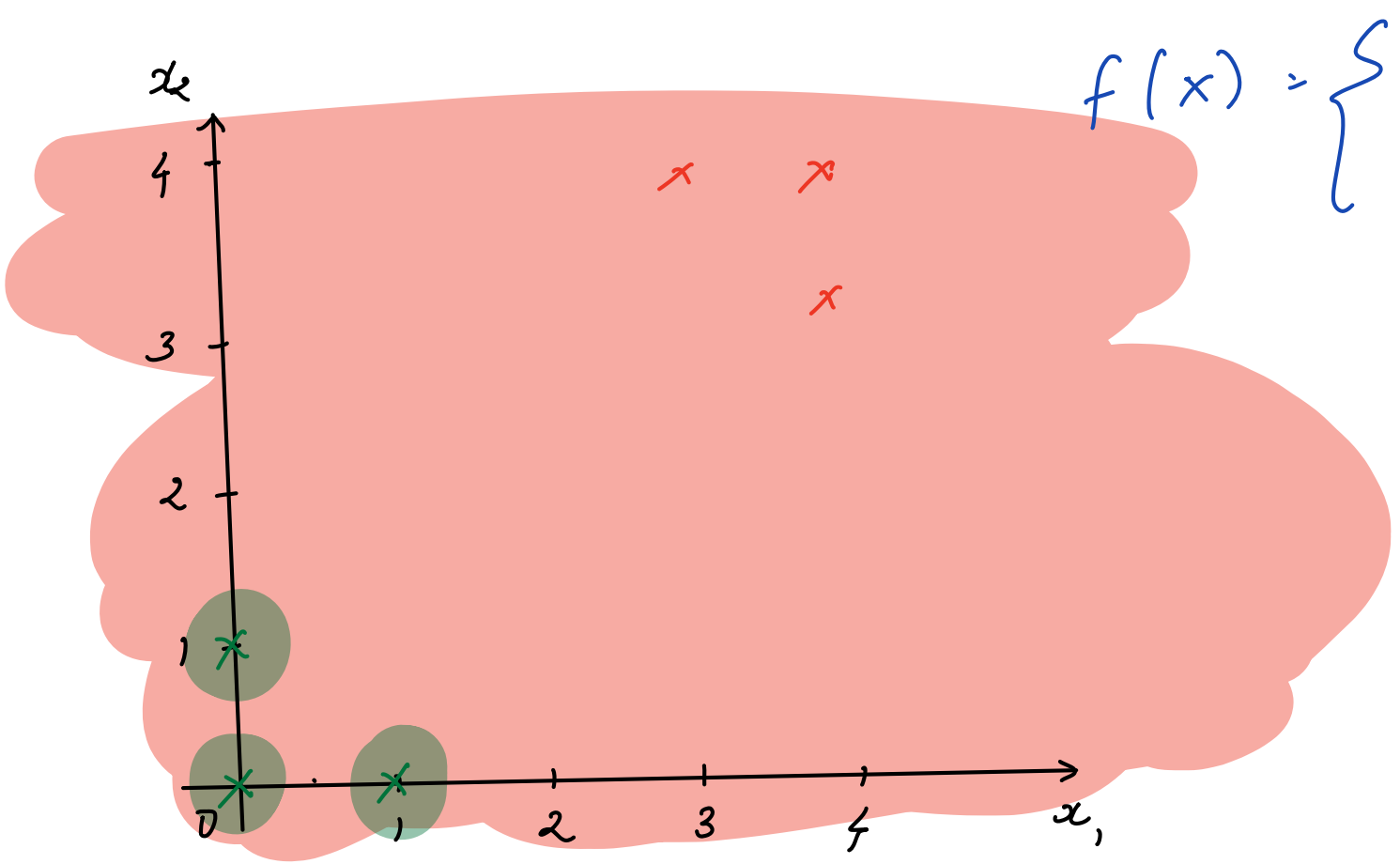
$f(x)$: sign (area - 10)

$g(x)$: sign (price - 6)

$h(x)$: sign (Price - 9)

Evaluating Learned Models : Test Data

- Learning algorithm uses training data $(x^1, y^1), \ldots, (x^n, y^n)$ to get model f.
- But evaluating the learned model must **not** be done on the training data itself.
- Use test data that is **not** in the training data for model evaluation.



$$f(x) = \begin{cases} +1 & \text{if } x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ +1 & \text{if } x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ +1 & \text{if } x = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ -1 & \text{otherwise} \end{cases}$$

Model Selection : Validation Data

- Learning algorithms just find the “best” model in the collection of models given by the human.
- How to find the right collection of models?
- This is called model selection, and it is done by using another subset of data called **validation data** that is distinct from train and test data.

$$Price = w_1 * (\#rooms) + w_2 (area) + w_3 (distance) + b$$

Outline

1. What is Machine Learning??
2. The Wonders of Machine Learning
3. Data, Models and ML Tasks
4. Supervised Learning
    1. Regression
    2. Classification
5. **Unsupervised Learning**
    1. Dimensionality Reduction
    2. Density Estimation

Unsupervised Learning

- Unsupervised learning is ‘understanding data’
- Data: $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$
- $\mathbf{x}^i \in \mathbb{R}^d$
- Build models that compress, explain and group data.

Unsupervised Learning Application

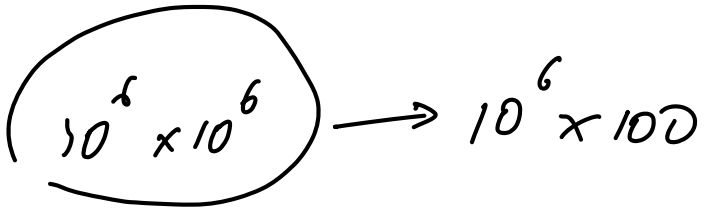
Tweet 1

⋮

Tweet 999999

Group the million tweets into 10 manageable groups

Dimensionality Reduction

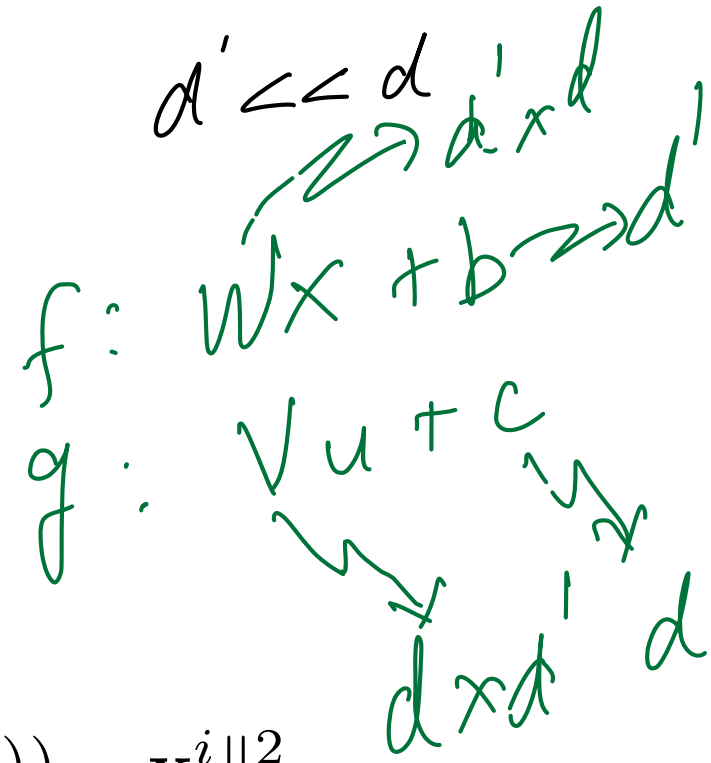


E.g.: Represent a million gene expression levels of a million people, using just 100 numbers per person.

Dimensionality reduction: compression and simplification.

Dimensionality Reduction

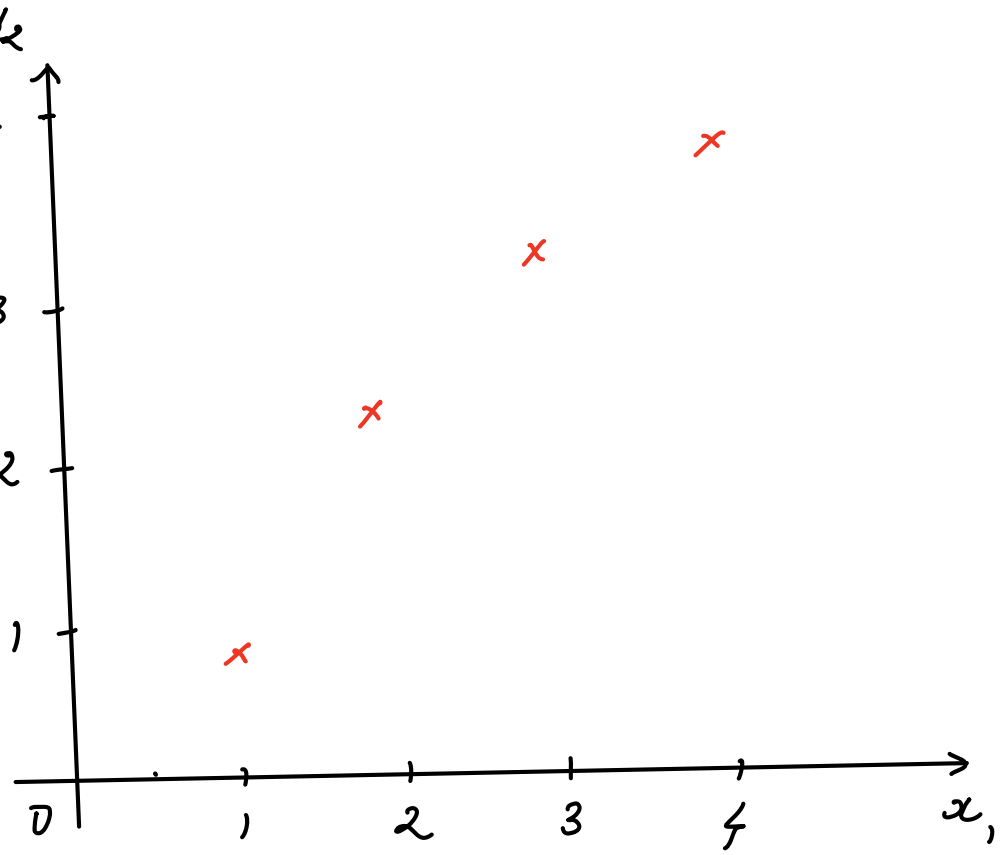
- Data: $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$
- $\mathbf{x}^i \in \mathbb{R}^d$
- Encoder $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$
- Decoder $g : \mathbb{R}^{d'} \rightarrow \mathbb{R}^d$
- Goal : $g(f(\mathbf{x}^i)) \approx \mathbf{x}^i$
- Loss $= \frac{1}{n} \sum_{i=1}^{n} \|g(f(\mathbf{x}^i)) - \mathbf{x}^i\|^2$



Dimensionality Reduction Illustration

$d = 2$, $d' = 1$, $n = 4$

	f	g
$[1, 0.8]$	0.2	$[0.2, 0.2]$
$[2, 2.2]$	-0.2	$[-0.2, -0.2]$
$[3, 3.2]$	-0.2	$[-0.2, -0.2]$
$[4, 3.8]$	0.2	$[0.2, 0.2]$



$f(x) = x_1 - x_2$

$g(u) = [u, u]$

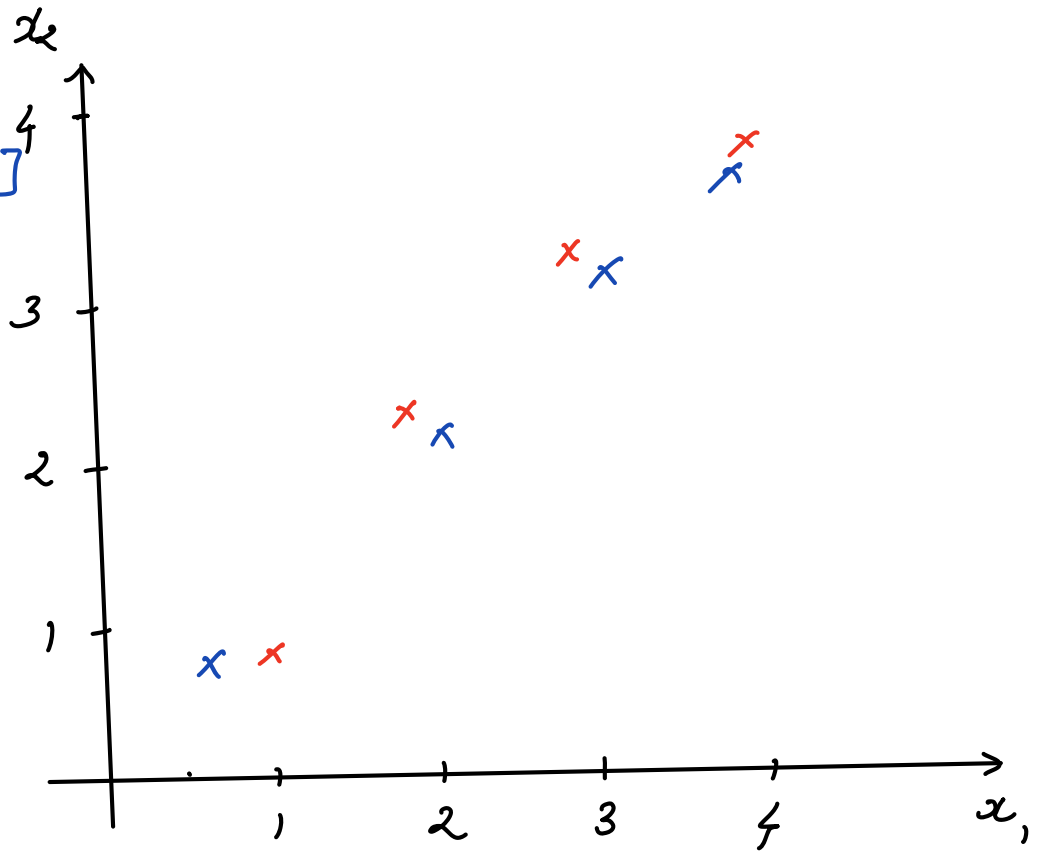
$\check{f}(x) = \frac{x_1 + x_2}{2}$

$\check{g}(u) = [u, u]$

Dimensionality Reduction Illustration

$d = 2$, $d' = 1$, $n = 4$

	$\tilde{f}$	$\tilde{g}$
$[1, 0.8]$	0.9	$[0.9, 0.9]$
$[2, 2.2]$	2.1	$[2.1, 2.1]$
$[3, 3.2]$	3.1	$[3.1, 3.1]$
$[4, 3.8]$	3.9	$[3.9, 3.9]$



$f(x) = x_1 - x_2$

$g(u) = [u, u]$

$\tilde{f}(x) = \frac{x_1 + x_2}{2}$

$\tilde{g}(u) = [u, u]$

$10^6 \times 100$

$\hat{f}$, $\tilde{g}$

Density Estimation

E.g.: Assuming tweets from an account are independently generated randomly. Create a robot account that generates more such tweets.

$f(\text{Tweet})$ = Score of the tweet

wisdomofchopra.com

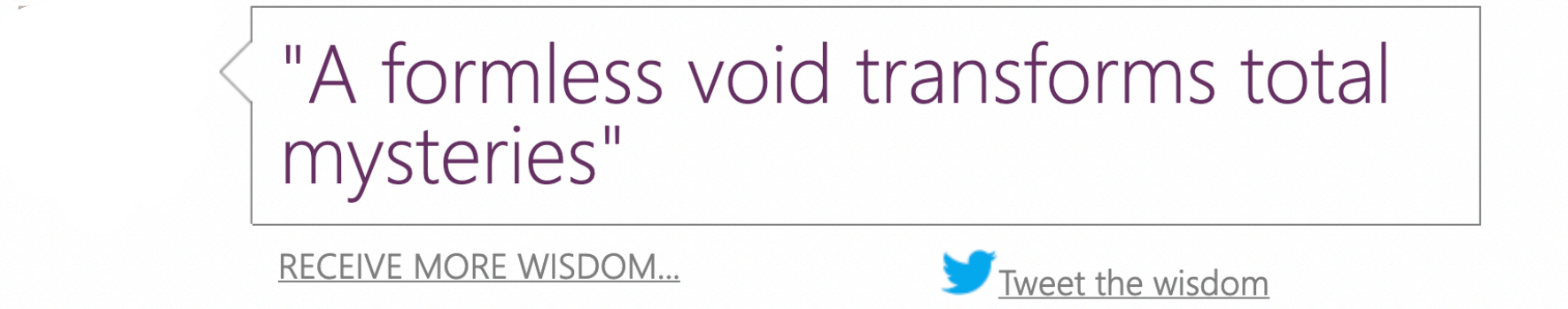
It has been said by some that the thoughts and tweets of Mr. Chopra are indistinguishable from a set of profound sounding words put together in a random order, particularly the tweets tagged with "#cosmisconciousness". This site aims to test that claim! Each "quote" is generated from a list of words that can be found ir Chopra's Twitter stream randomly stuck together in a sentence.

"A formless void transforms total mysteries"

RECEIVE MORE WISDOM... Tweet the wisdom

Disclaimer: This is intended for entertainment purposes only. It in no way reflects the thoughts of any real person.

Density Estimation



To generate such sentences randomly, we need to be able to assign a probability score to every possible 128 character sentence, giving high scores to those that are likely to be from the original source.

A density estimation model takes in several samples from a random source, and outputs a model that assigns a probability score to every possible instance.

Density Estimation

- Data: $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$
- $\mathbf{x}^i \in \mathbb{R}^d$
- Probability mapping $P : \mathbb{R}^d \to \mathbb{R}_+$ that 'sums' to one.
- Goal : $P(\mathbf{x})$ is large if $\mathbf{x} \in$ Data, and low otherwise.
- Loss $= \frac{1}{n} \sum_{i=1}^{n} -\log\left(P(\mathbf{x}^i)\right)$

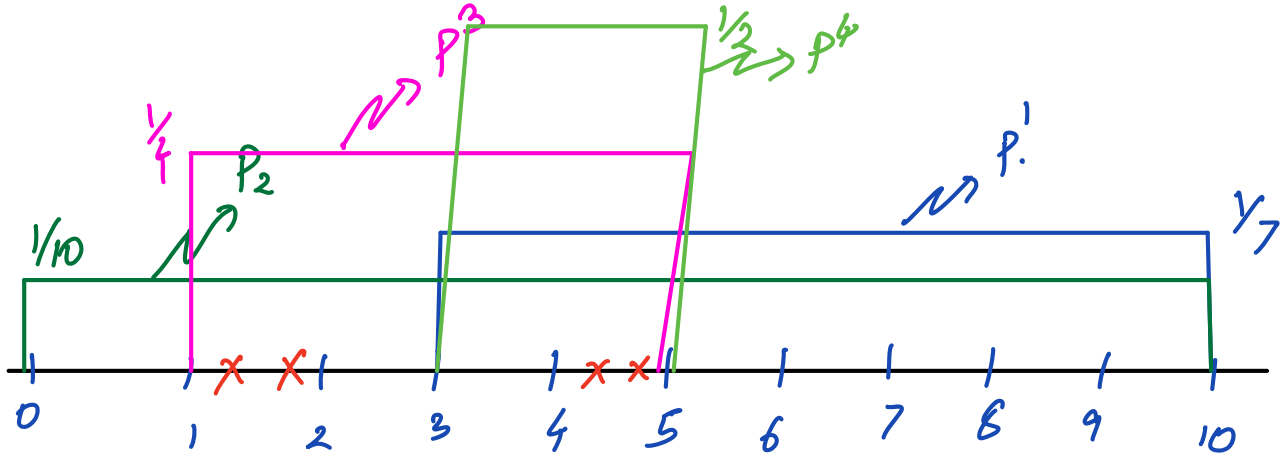
Σ sum over all tweets $\{a..z\}^{128}$

$P(\text{tweet}) = 1$

$P(x^i)$ is large.

$P(\text{anything}) = 10^{10}$

Density Estimation Illustration 1



x^1: [1.2]

x^2: [1.9]

x^3: [4.3]

x^4: [4.8]

p^1 = Uniform in [3,10] 0, 0, 1/7, 1/7

p^2 = Uniform in [0,10] 1/10, 1/10, 1/10, 1/10

p^3 = Uniform in [1,5] 1/4, 1/4, 1/4, 1/4

p^4 = Uniform in [3,5] 0, 0, 1/2, 1/2

$loss[p^4] : loss[p^1] : \infty > loss[p^2] > loss[p^3]$

Density Estimation Illustration 2

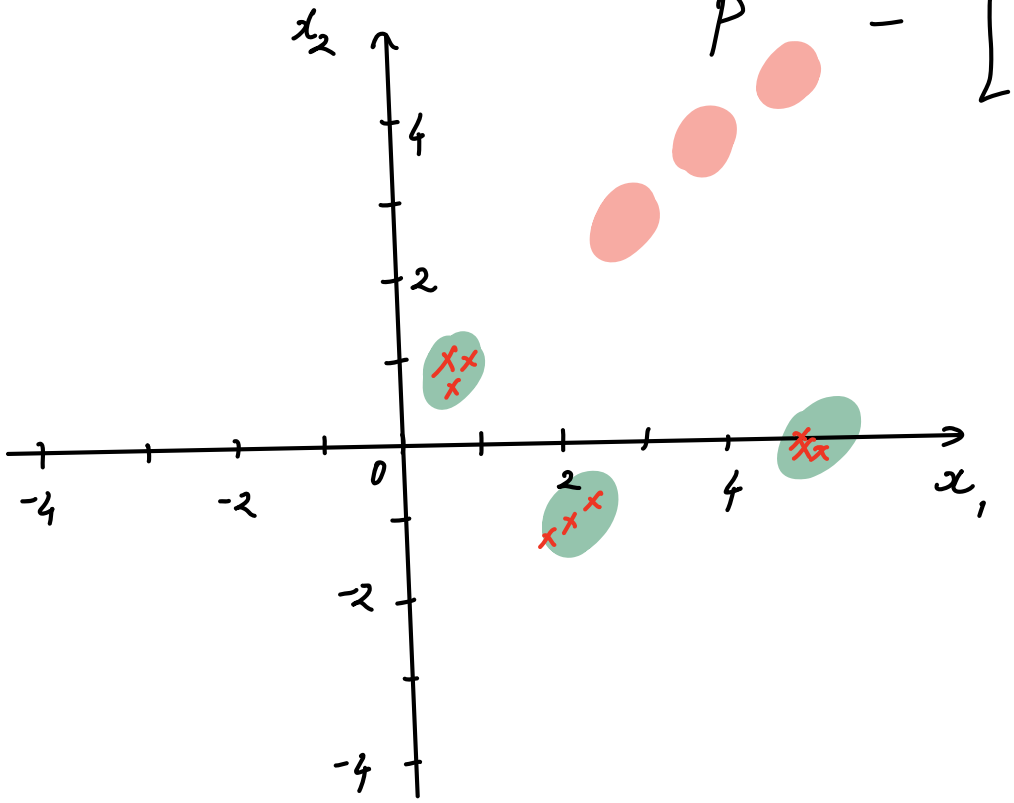
$d = 2$

(1.1, 1.3)
(0.9, 0.7)
(2.1, -1)
(5.1, 0.1)
(2.2, -0.9)
(5.1, 0.0)
(0.9, 1.2)
(1.9, -1.1)
(4.8, -0.1)

Gaussian Mixture Model

$p^1 - \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 5 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$p^2 - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix}$



Clustering

Density Estimation Illustration 2