Lecture 1:

Definition of Machine Learning:

The field of machine learning studies the design of computer programs able to induce patterns, regularities, or rules from past experiences. Learner (a computer program) processes data **D** representing past experiences and tries to either develop an appropriate response to future data, or describe in some meaningful way the data seen.

Example:

Learner sees a set of patient cases (patient records) with corresponding diagnoses. It can either try:
– to predict the presence of a disease for future patients
– describe the dependencies between diseases, symptoms

Types of Learning

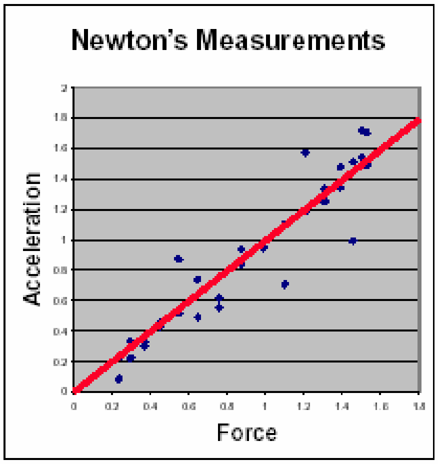
• **Supervised learning**
– Learning a mapping between an input **x** and a desired output **y**
• **Unsupervised learning**
– Understanding the relationships between data components
• **Reinforcement learning**
– Learning to act in the environment based on the delayed rewards

Examples of ML Applications

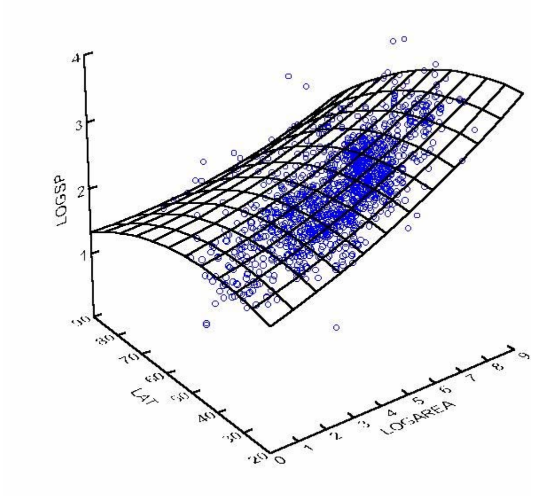
Supervised Learning

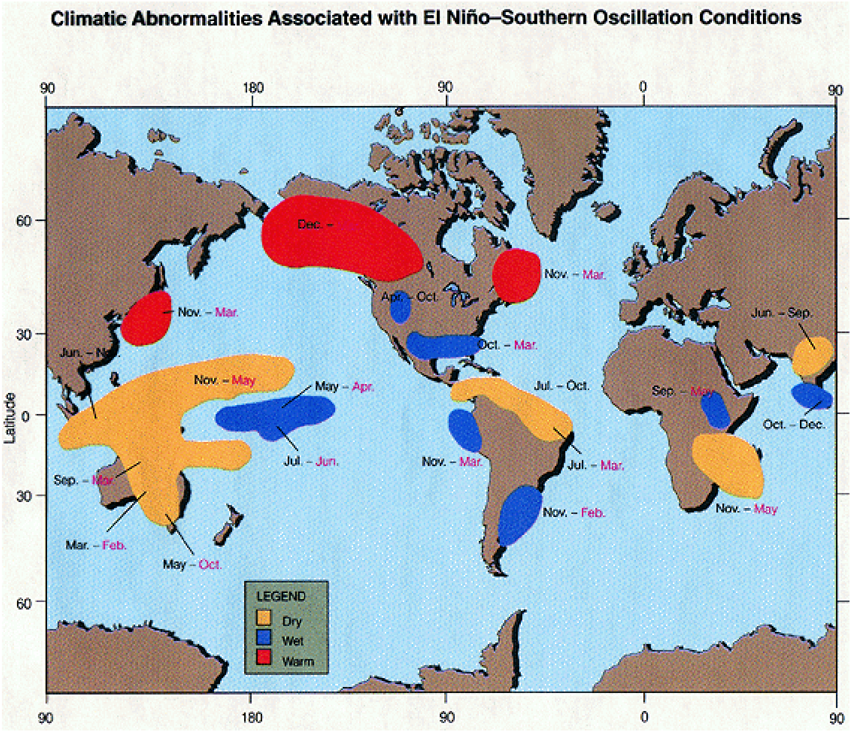
Regression: Continuous or Discrete Input, Continuous Output

1. (Interest Rates, Previous Stock Prices) → Future Stock Prices
2. (Southern Oscillation Index) → Inches Rain
3. (Inches Rain) → Corn production
4. (Force) → Acceleration



Newton's Measurements

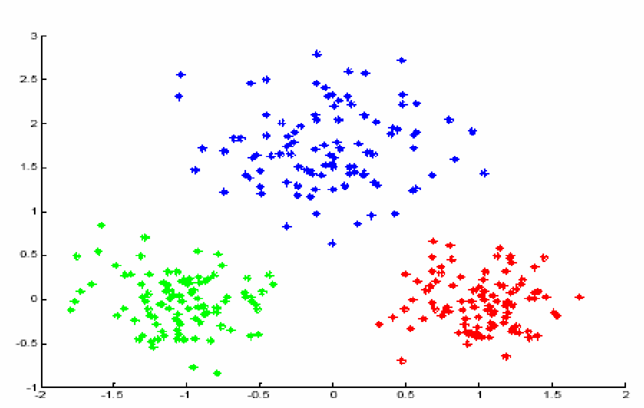


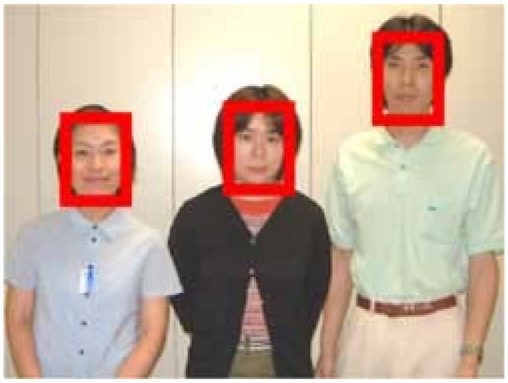


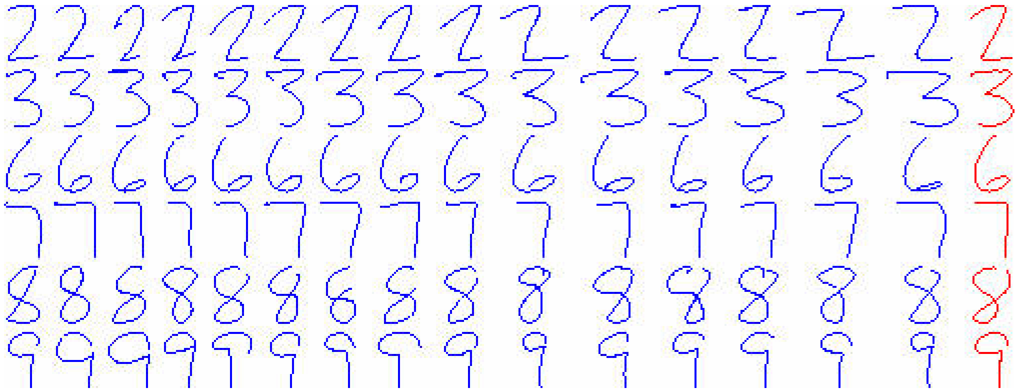
Climatic Abnormalities Associated with El Niño–Southern Oscillation Conditions

Classification: Continuous or Discrete Input, Discrete Output

1. (Color, Shape, Seed Size) → Fruit Name
2. Alarm → BreakIn, Alarm & Earthquake → NoBreakIn
3. (Midterm, Homework) → Final Grade
4. (Income, Current Debt) → Loan Repayment
5. (Pixel Values) → Character Recognition
6. (Text) → Spam Email

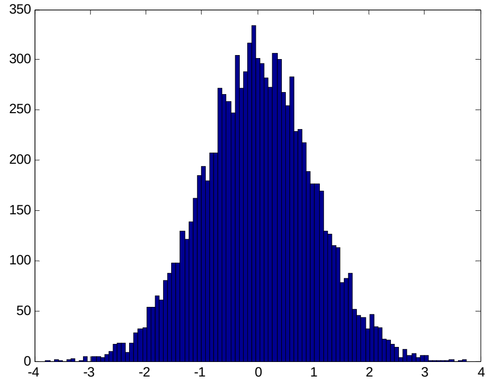


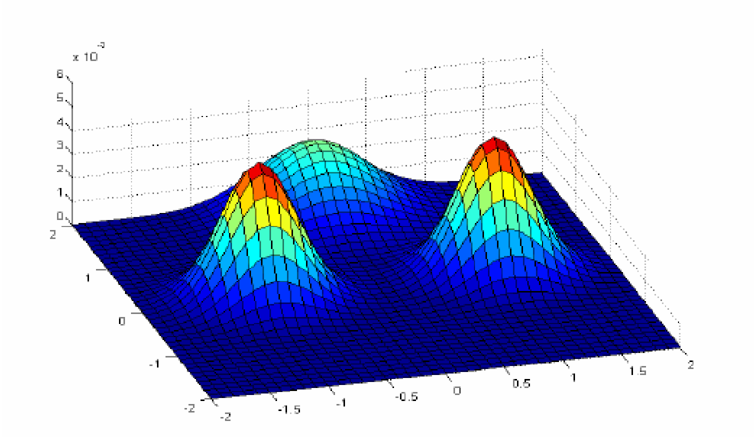




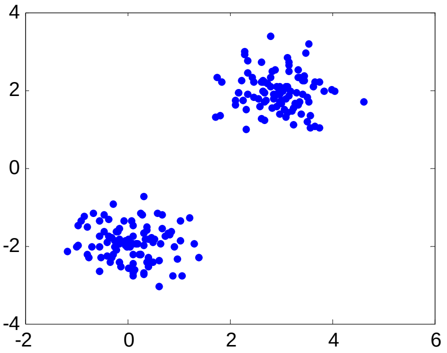
Unsupervised Learning: Continuous or Discrete Input, No Output

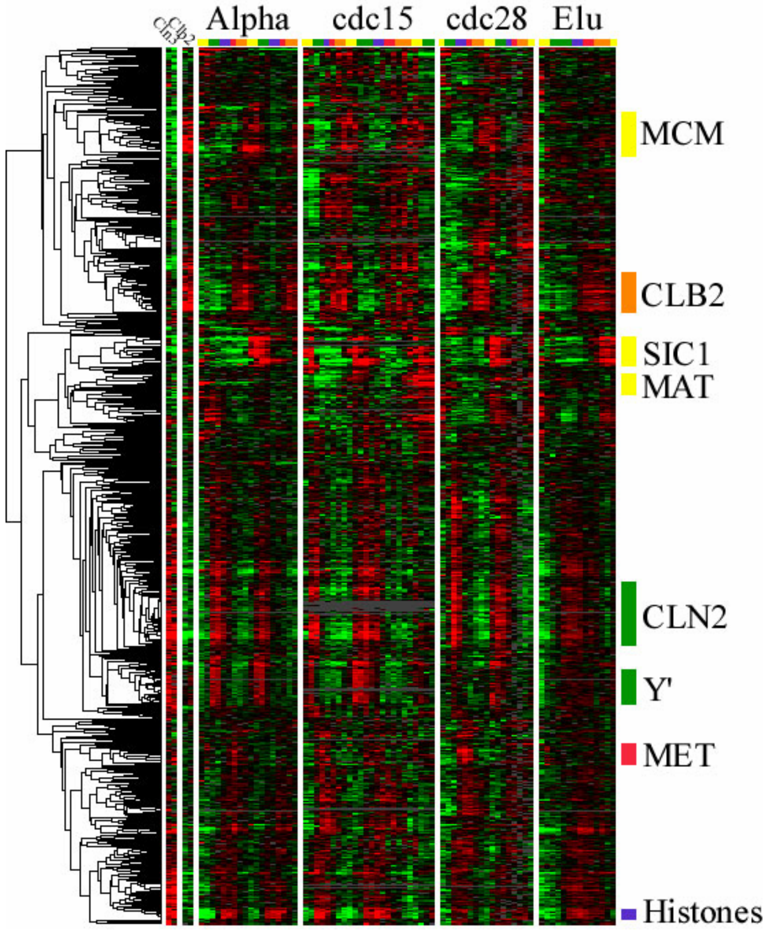
Density Estimation: Understand the data distribution

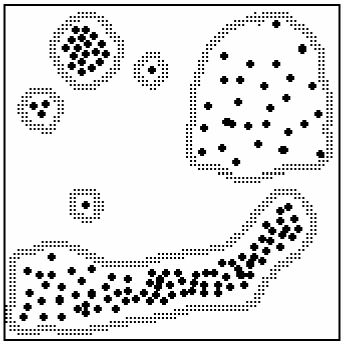


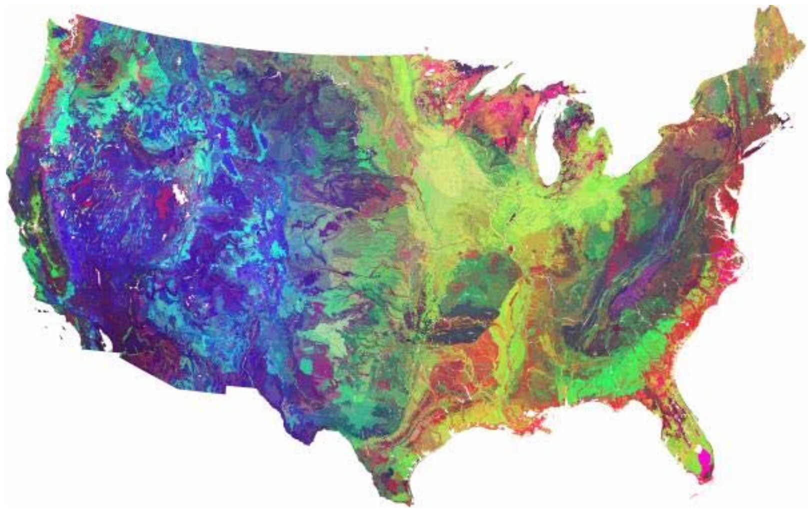


Clustering: Find groups of similar objects

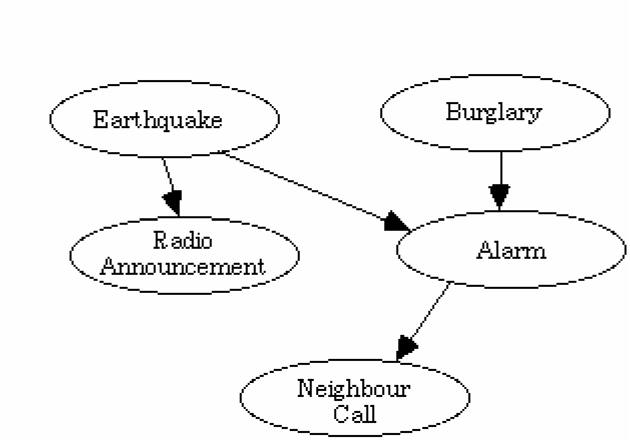


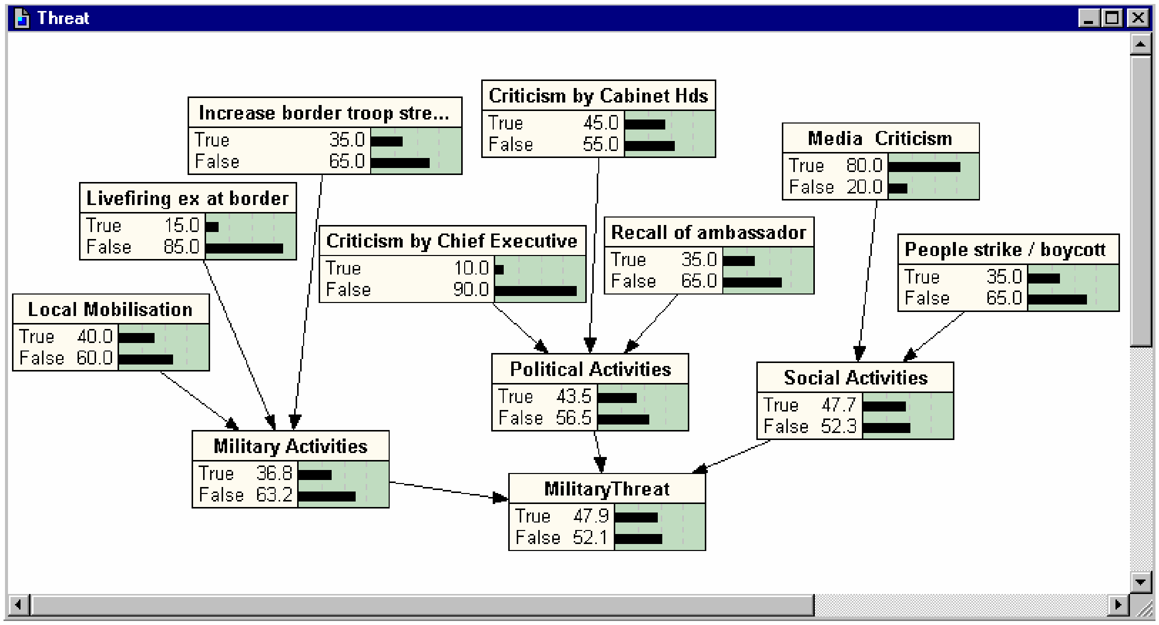




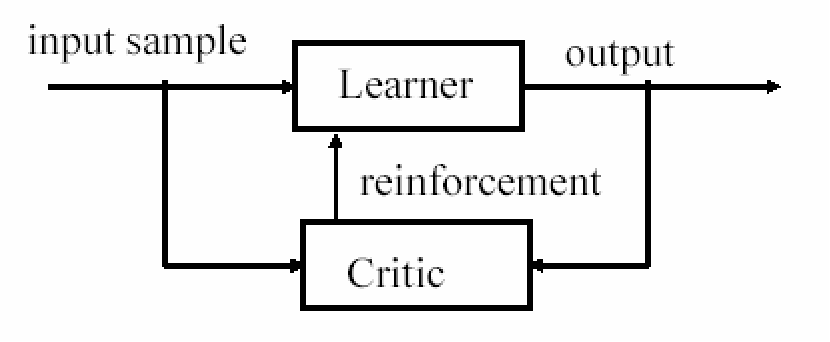


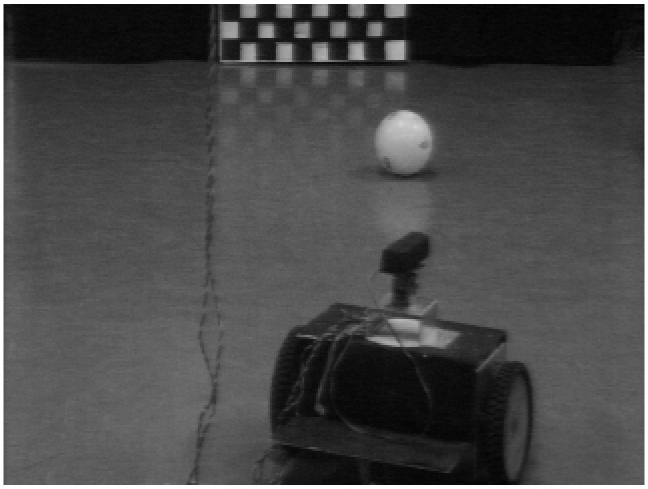
Bayesian Networks: Find dependencies between events

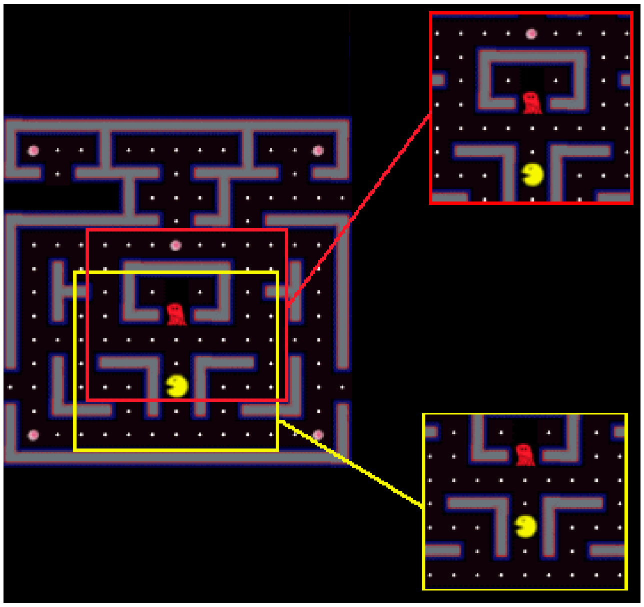




Reinforcement Learning







Notation: Data Sets for Supervised Learning

Data set is a set of tuples:

$$\mathbf{D} = \{(x_{i1}, x_{i2}, \ldots, x_{iM}, y_i),\ i = 1, 2, \ldots N\} = \{(\mathbf{x}_i, y_i),\ i=1, 2, \ldots N\}$$

where:

$(\mathbf{x}_i, y_i)$ is example (also called pattern, data point)
$X_1, X_2, \ldots, X_M$ are attributes (also called features, inputs, variables)
Y is response (also called target, dependent variable)
M is number of attributes
N is number of examples (or size of data set)

Example: *Snapshot of Census Income Dataset (N=23, M=7)*

13th example

x_{23} = 'Exec-managerial'

x_{46} = 40

Age	Education	Occupation	Race	Gender	Hours/Week	Origin	Salary
39	Bachelors	Adm-clerical	White	Male	40	United-States	<=50K
50	Bachelors	Exec-managerial	White	Male	13	United-States	<=50K
38	HS-grad	Handlers-cleaners	White	Male	40	United-States	<=50K
53	11th	Handlers-cleaners	Black	Male	40	United-States	<=50K
28	Bachelors	Prof-specialty	Black	Female	40	Cuba	<=50K
37	Masters	Exec-managerial	White	Female	40	United-States	<=50K
49	9th	Other-service	Black	Female	16	Jamaica	<=50K
52	HS-grad	Exec-managerial	White	Male	45	United-States	>50K
31	Masters	Prof-specialty	White	Female	50	United-States	>50K
42	Bachelors	Exec-managerial	White	Male	40	United-States	>50K
37	Some-college	Exec-managerial	Black	Male	80	United-States	>50K
30	Bachelors	Prof-specialty	Asian-Pac-Islander	Male	40	India	>50K
23	Bachelors	Adm-clerical	White	Female	30	United-States	<=50K
32	Assoc-acdm	Sales	Black	Male	50	United-States	<=50K
40	Assoc-voc	Craft-repair	Asian-Pac-Islander	Male	40	?	>50K
34	7th-8th	Transport-moving	Amer-Indian-Eskimo	Male	45	Mexico	<=50K
25	HS-grad	Farming-fishing	White	Male	35	United-States	<=50K
32	HS-grad	Machine-op-inspct	White	Male	40	United-States	<=50K
38	11th	Sales	White	Male	50	United-States	<=50K
43	Masters	Exec-managerial	White	Female	45	United-States	>50K
40	Doctorate	Prof-specialty	White	Male	60	United-States	>50K
54	HS-grad	Other-service	Black	Female	20	United-States	<=50K
35	9th	Farming-fishing	Black	Male	40	United-States	<=50K

Missing data

Attribute X_2 = 'Education' is discrete (or categorical, or nominal)

Attribute X_6 = 'Hour/Week' is continuous (or numerical)

Response Y = 'Salary' is binary variable ⇒ **Classification**