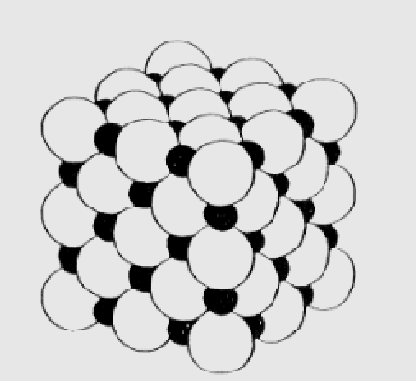
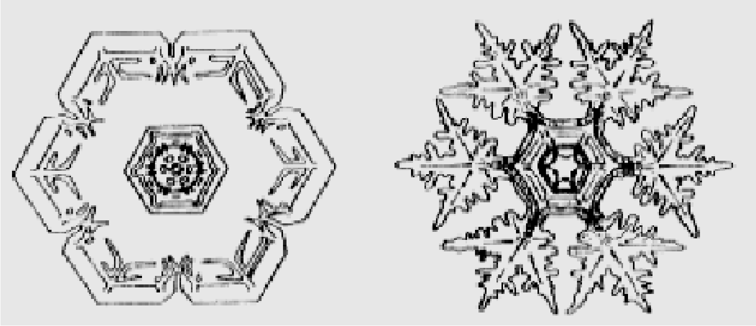
# *Pattern Recognition Artificial Neural Networks, and Machine Learning*

Yuan-Fang Wang
Department of Computer Science
University of California
Santa Barbara, CA 93106, USA

# "Pattern Recognition"
# What is a Pattern?

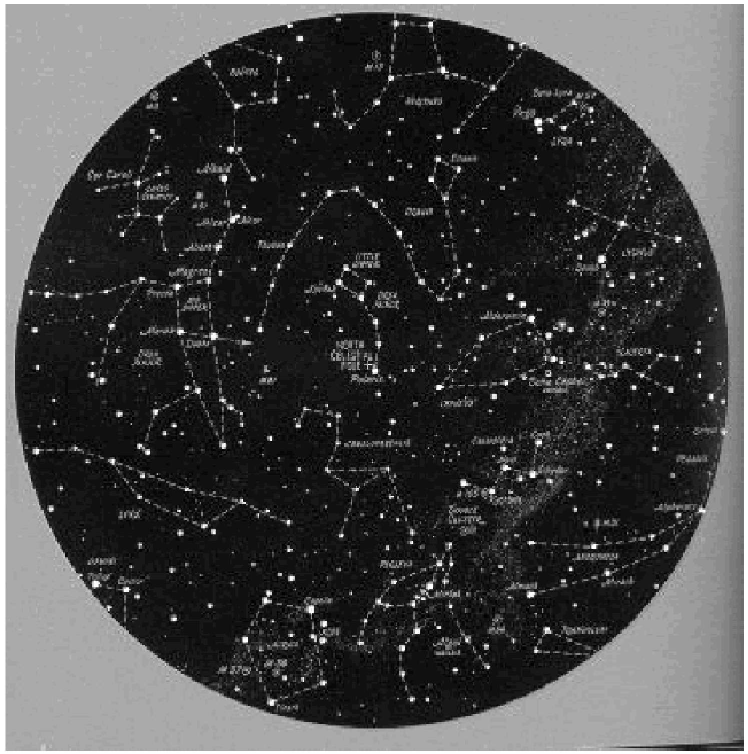
**Crystal Patterns:**

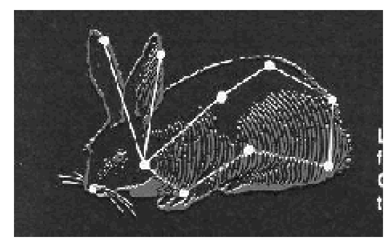




The crystal structures are represented by 3D graph, and they can be described by deterministic grammars or formal languages.

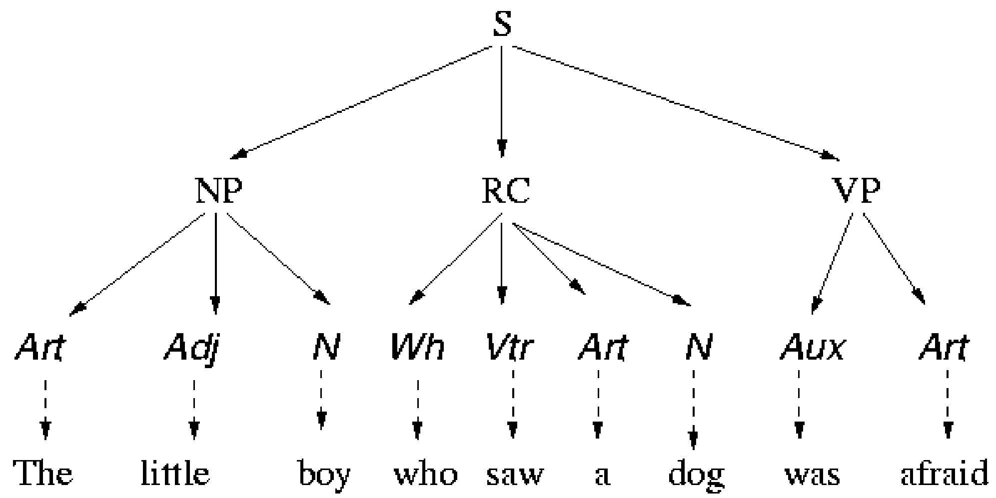
# Constellation Patterns:





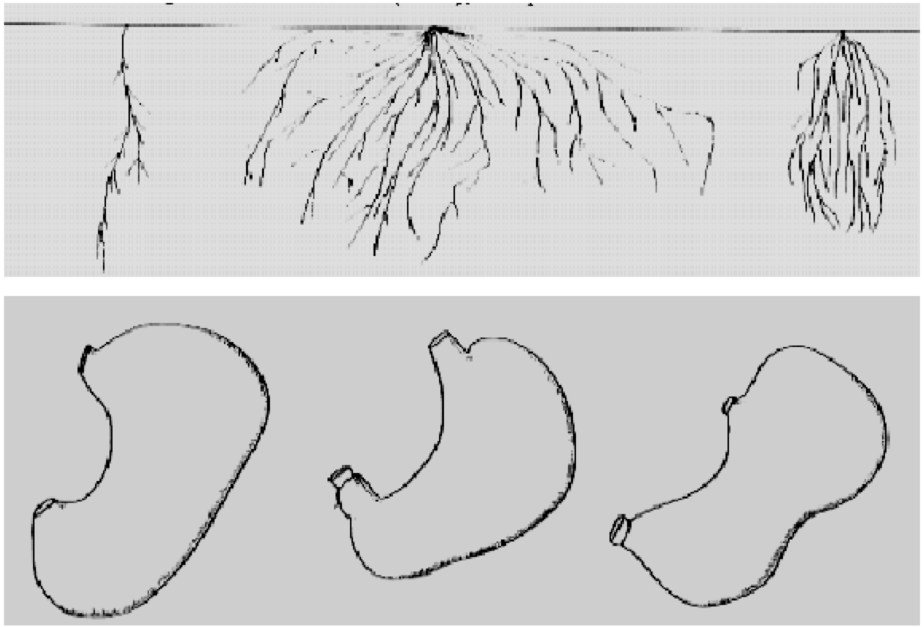
Each constellation could be represented by a planar graph, which maintains a certain regular shape with slight deformation during a season.

# English Pattern:



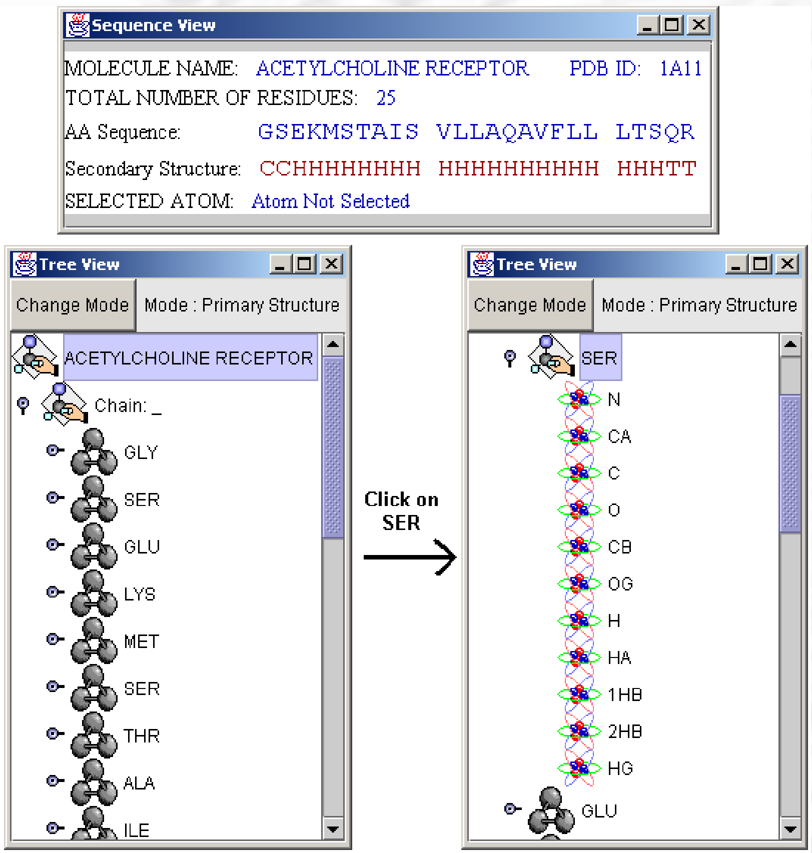

English sentences are patterns governed by English grammar and some stochastic process of the semantics.

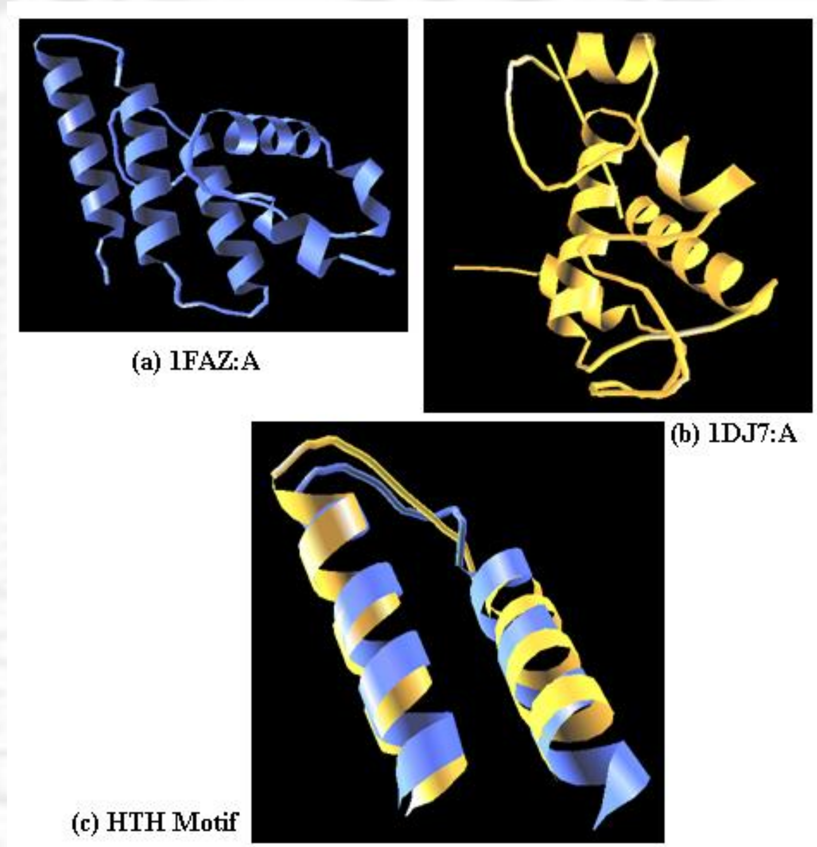
## **Biology Patterns:** — Root of plant and Human Stomach



Like English sentences, biology organs present regularities in their shape – governed by the genetic codes as well as non-deterministic appearance – influenced by the stochastic environment.

- DNA patterns
  - AGCTCGAT
- Protein Patterns
  - 20 amino acids



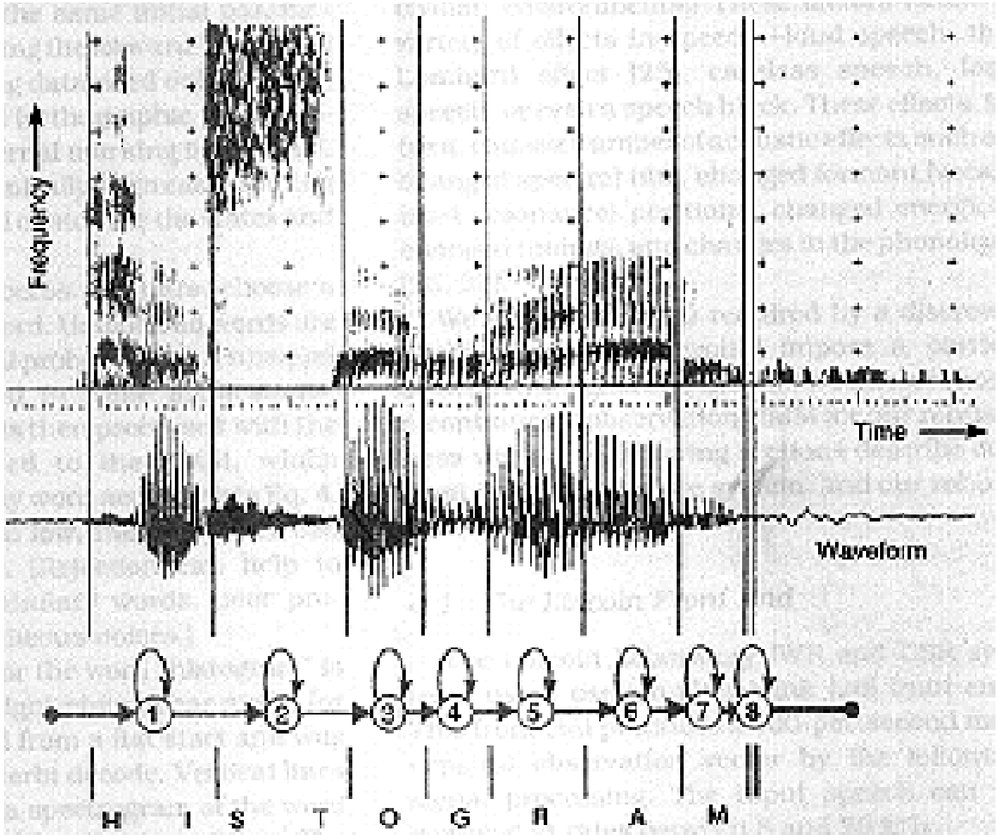



(a) 1FAZ:A

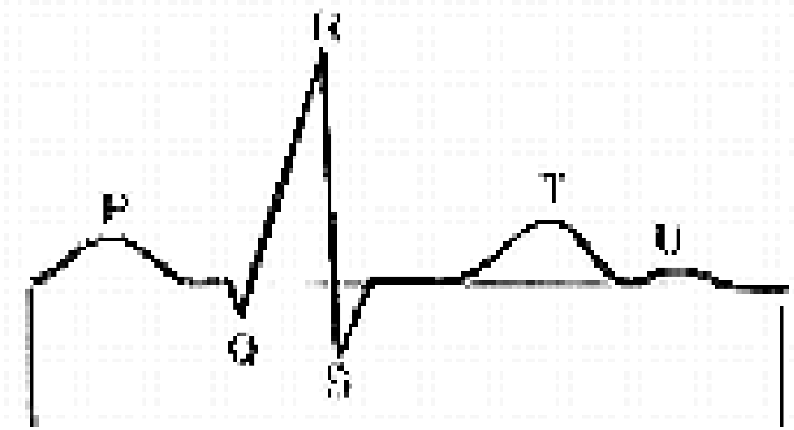
(b) 1DJ7:A

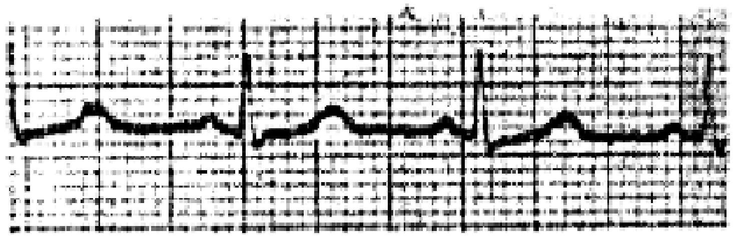
(c) HTH Motif

## Speech Signal:

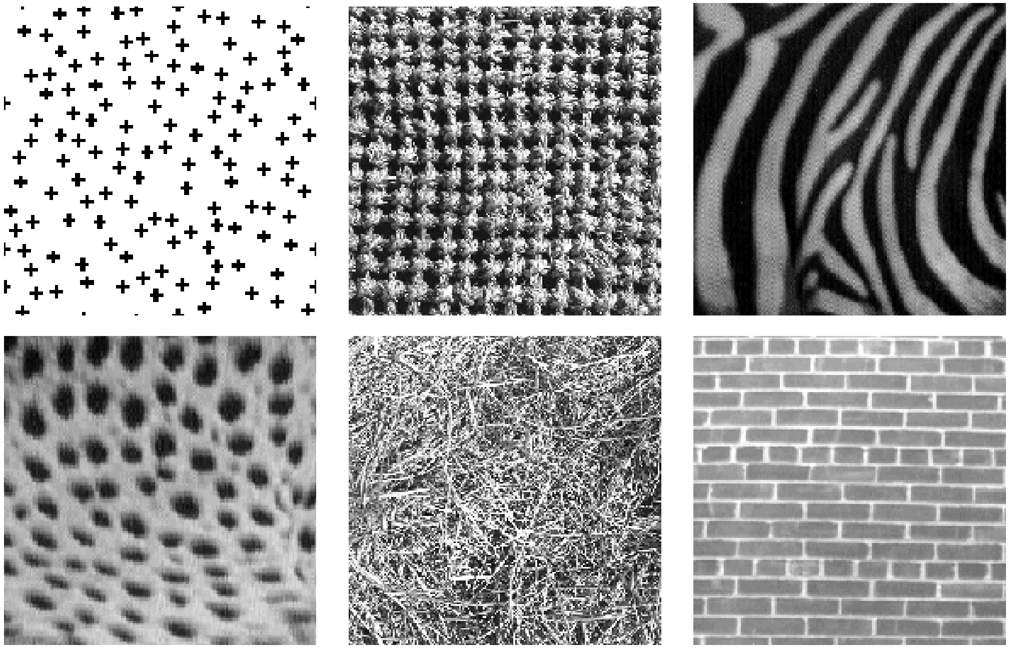


# EGK signal for diagnosing heart diseases:



## Texture Patterns:



Textures are the richest pattern created in nature, perceptually each class of texture has some common features–regularities, and it also contains non-deterministic characteristics.

Faces

Finger prints

# Other Patterns

- Insurance, credit card applications
  - applicants are characterized by a pattern
    - # of accidents, make of car, year of model
    - income, # of dependents, credit worthiness, mortgage amount
- Dating services
  - Age, hobbies, income, etc. establish your "desirability"

# Other Patterns

- Web documents
  - Key words based description (e.g., documents containing War, Bagdad, Hussen are different from those containing football, NFL, AFL, draft, quarterbacks)
- Intrusion detection
  - Usage and connection patterns
- Cancer detection
  - Image features for tumors, patient age, treatment option, etc.

# *Other Patterns*

- Housing market
  - Location, size, year, school district
- University ranking
  - Student population, student-faculty ratio, scholarship opportunities, location, faculty research grants, etc.
- Too many
  - E.g., http://www.ics.uci.edu/~mlearn/MLSummary.html

# *What is a pattern?*

- A pattern is a set of objects, processes or events which consist of both deterministic and stochastic components
- A pattern is a record of certain dynamic processes influenced both by deterministic and stochastic factors

# *What is a Pattern? (cont.)*

Constellation patterns, texture patterns, EKG patterns, etc.

Completely regular, deterministic

(e.g., crystal structure)

Completely random

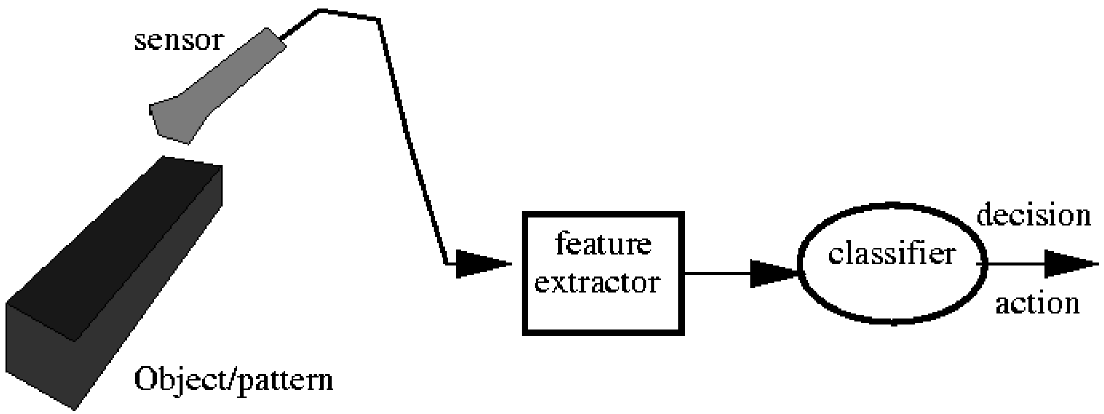
(e.g., white noise)

# What is Pattern Recognition?

- Classifies "patterns" into "classes"
- Patterns (x)
  - have "measurements", "traits", or "features"
- Classes ( $\varpi_i$)
  - likelihood (a prior probability $P(\varpi_i)$)
  - class-conditional density $p(x|\varpi_i)$
- Classifier (f(x) -> $\varpi_i$)
- An example
  - four coin classes: penny, nickel, dime, and quarter
  - measurements: weight, color, size, etc.
  - Assign a coin to a class based on its size, weight, etc.

We use *P* to denote probability ***mass*** function (*discrete*) and
*p* to denote probability ***density*** function (*continuous*)

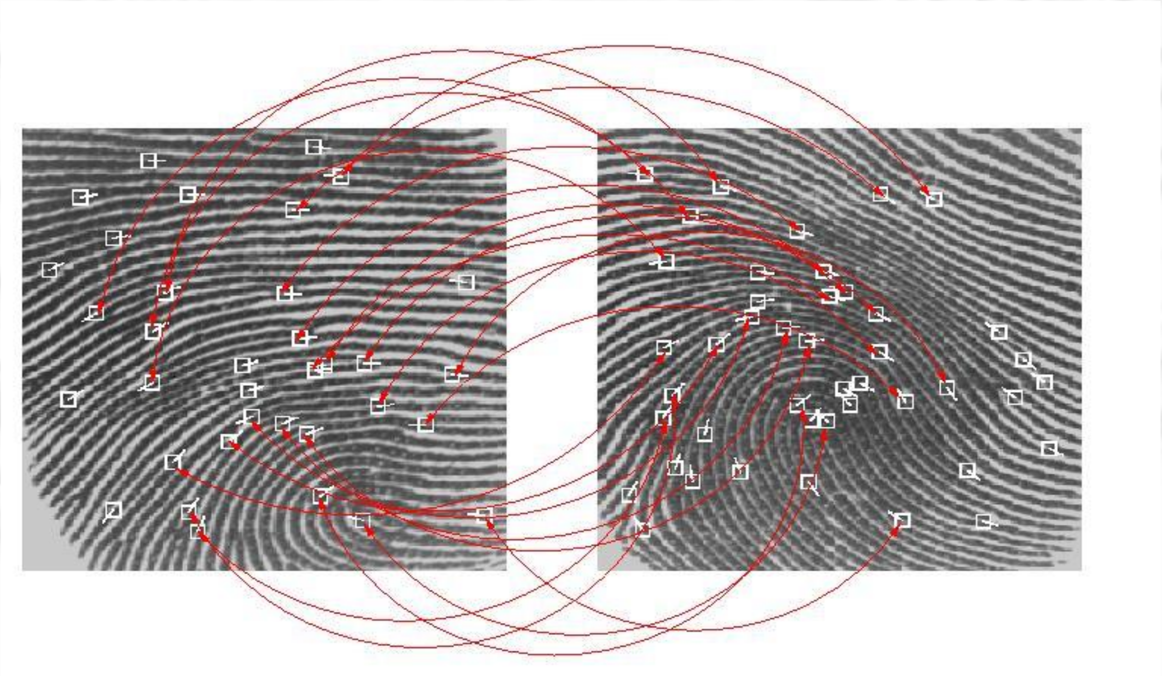
# An Example




Such system works in limited situations at a very fast speed.

Many visual inspection systems are like this:
Circuit board, fruit, OCR, etc.

# Another Example



# *Features*

- The intrinsic traits or characteristics that tell one pattern (object) apart from another
- Features extraction and representation allows
  - Focus on relevant, distinguishing parts of a pattern
  - Data reduction and abstraction

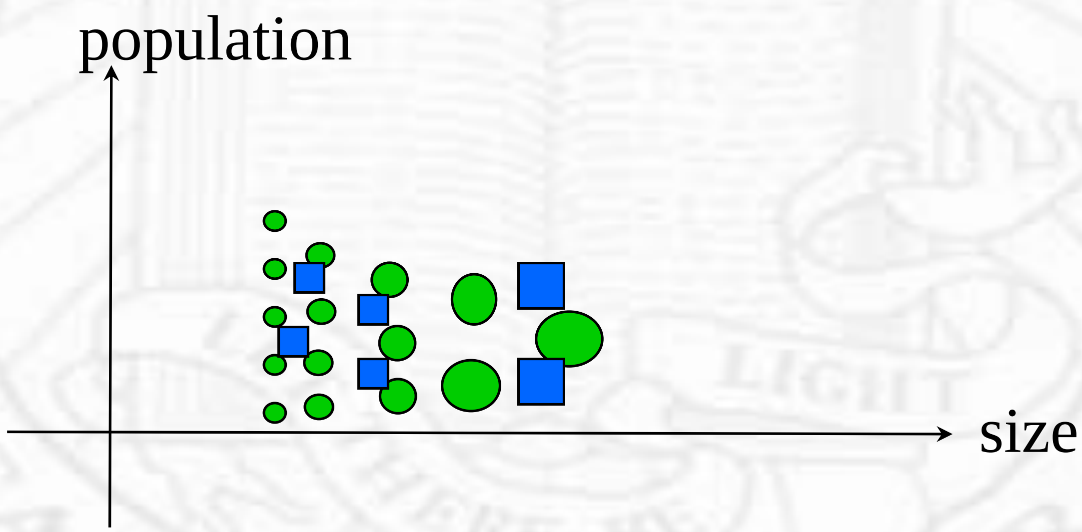
# Detection vs. Description

- Detection: something happened
- Heard noise
- Saw something interesting
- Non-flat signals

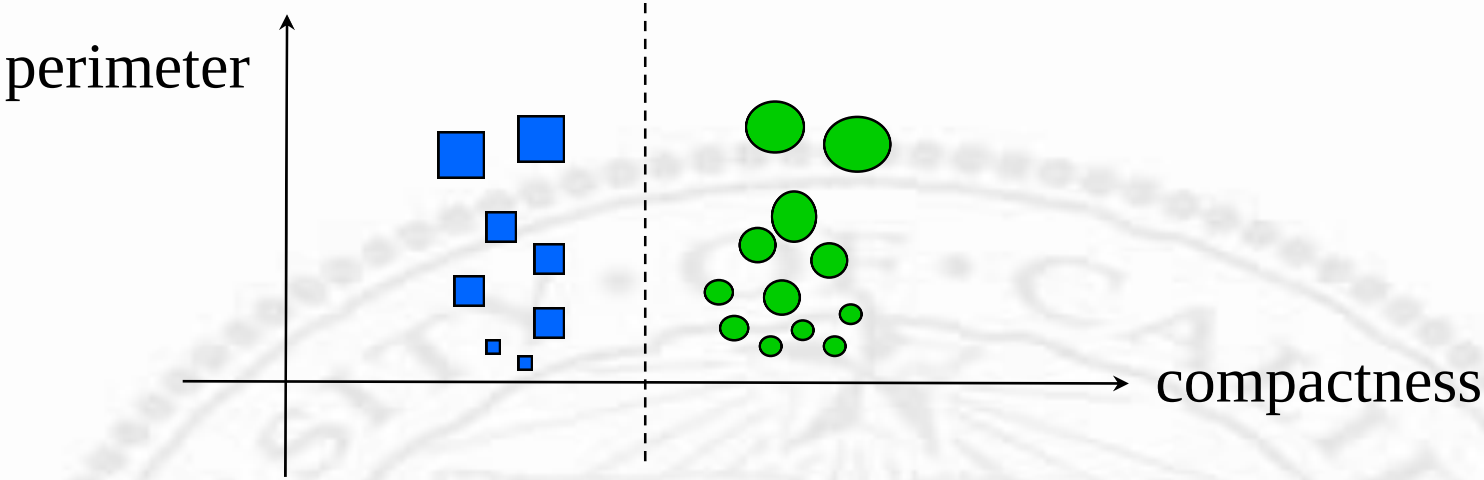
- Description: what has happened?
- Gun shot, talking, laughing, crying, etc.
- Lines, corners, textures
- Mouse, cat, dog, bike, etc.

# Feature Selection

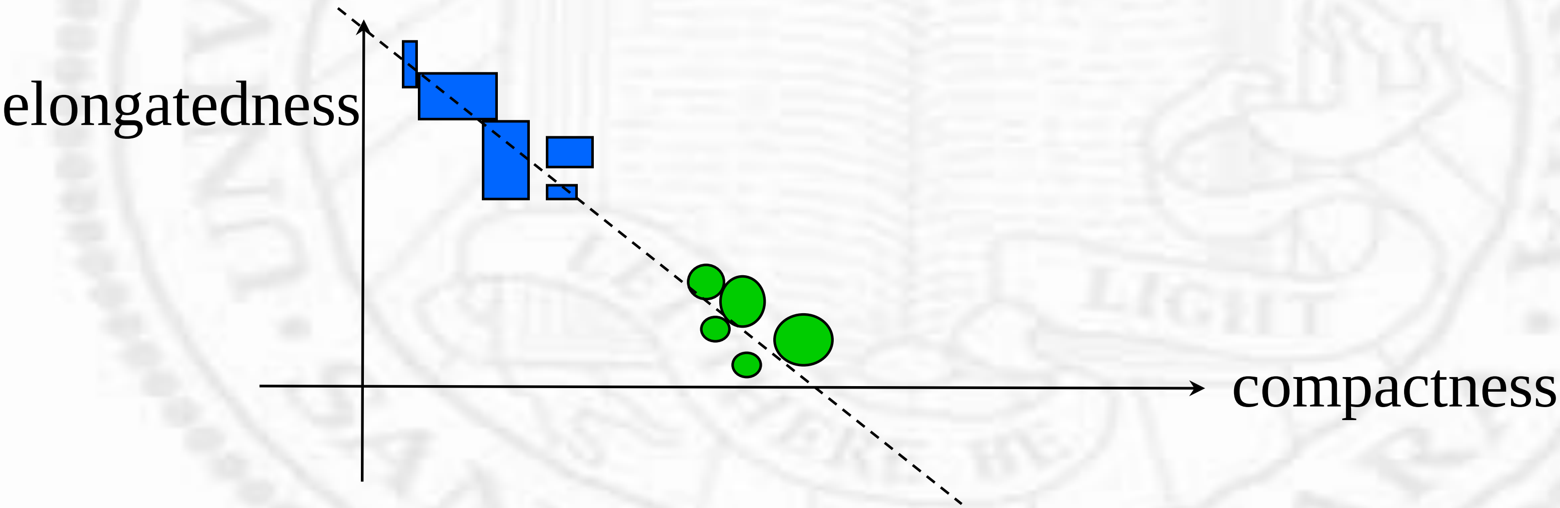
- More an art than a science
- Effectiveness criteria:




*Size* alone is not effective



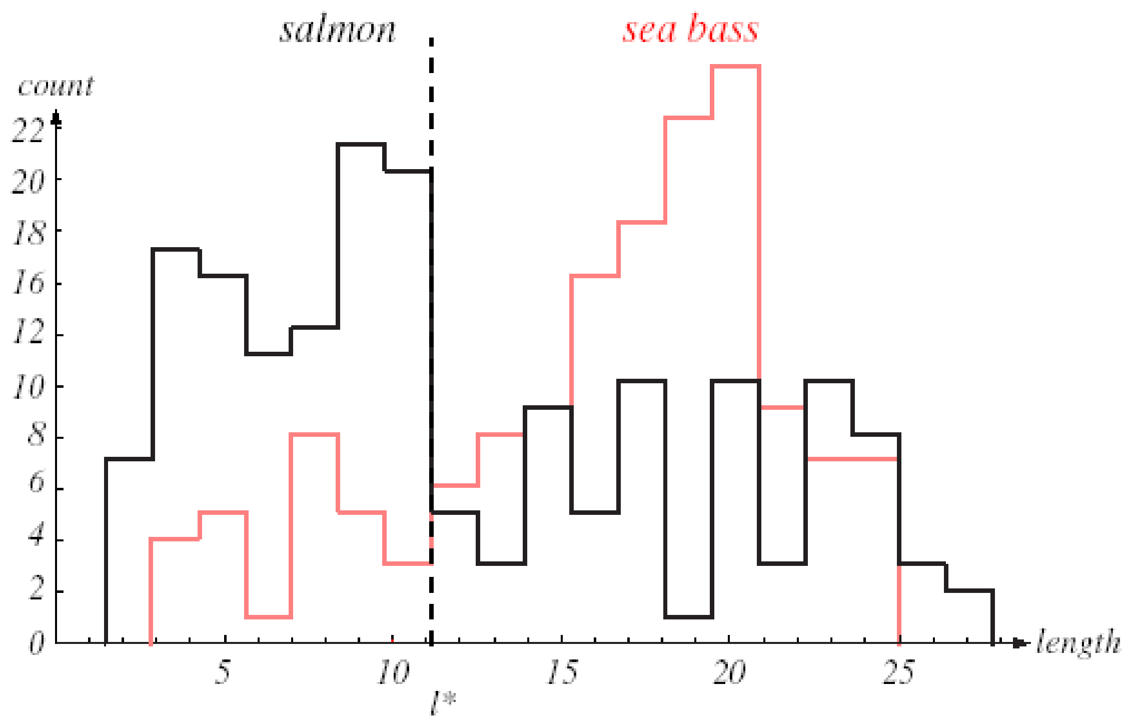

*Perimeter* is not effective
Discrimination is accomplished by *compactness* alone



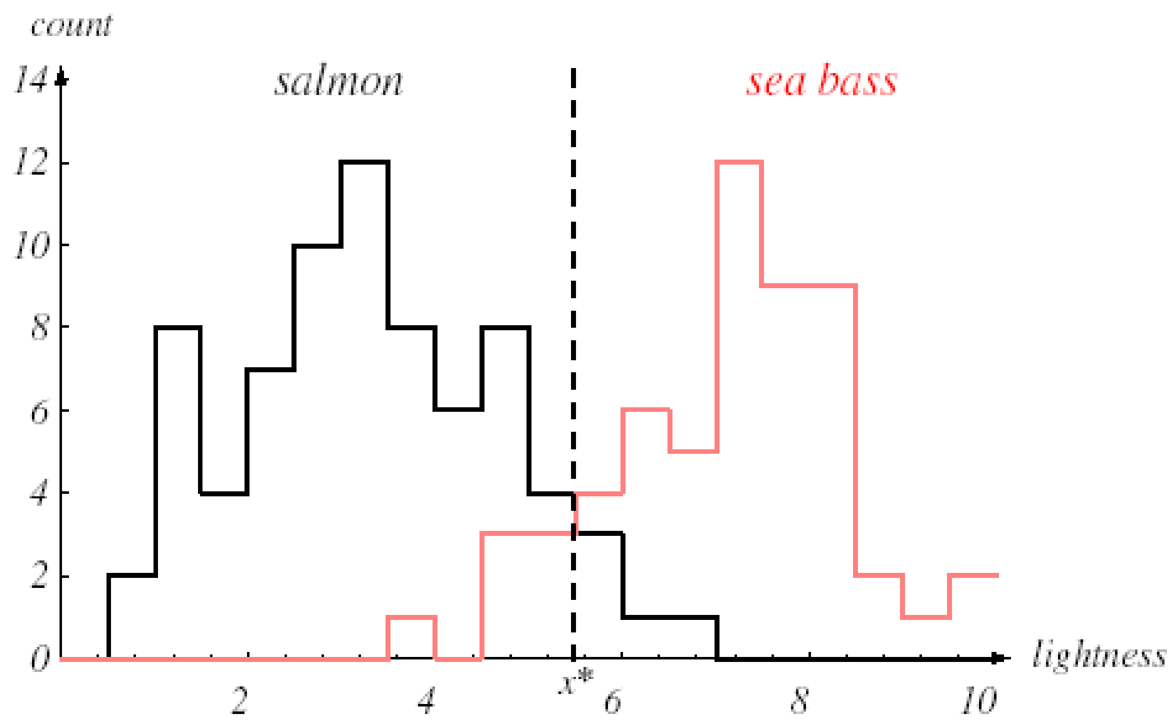

The two feature values are correlated, only one of them is needed

An example of fish classification

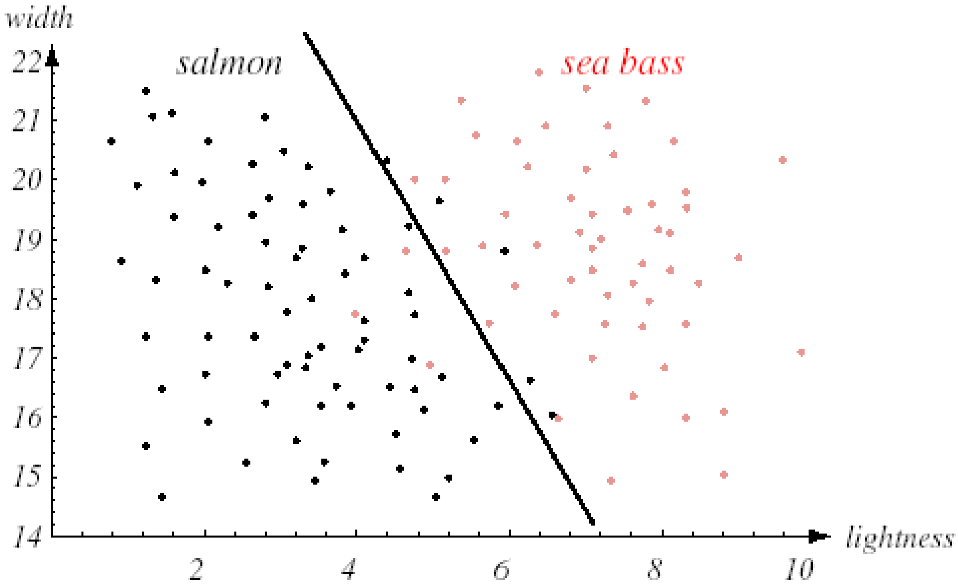
Salmon Vs. Sea Bass – histogram of fish length

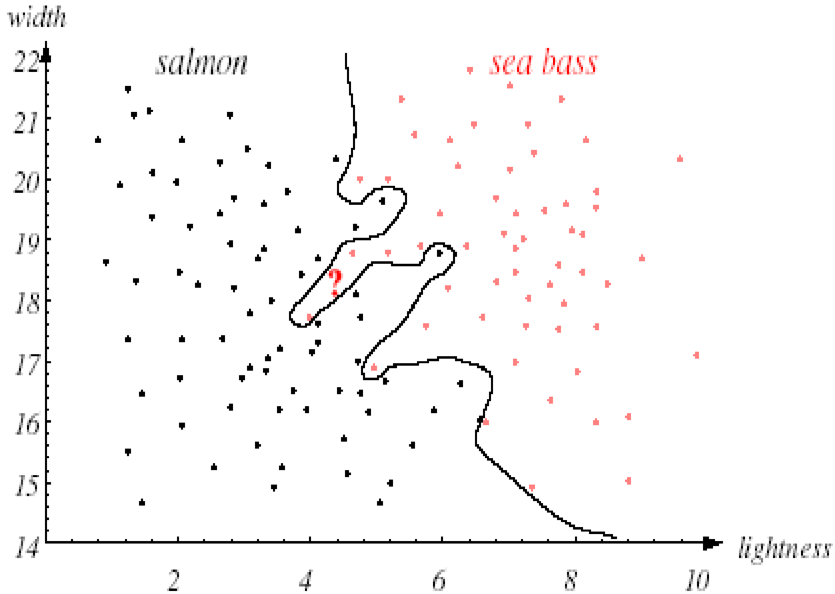


# Salmon Vs. Sea Bass – histogram of fish lightness



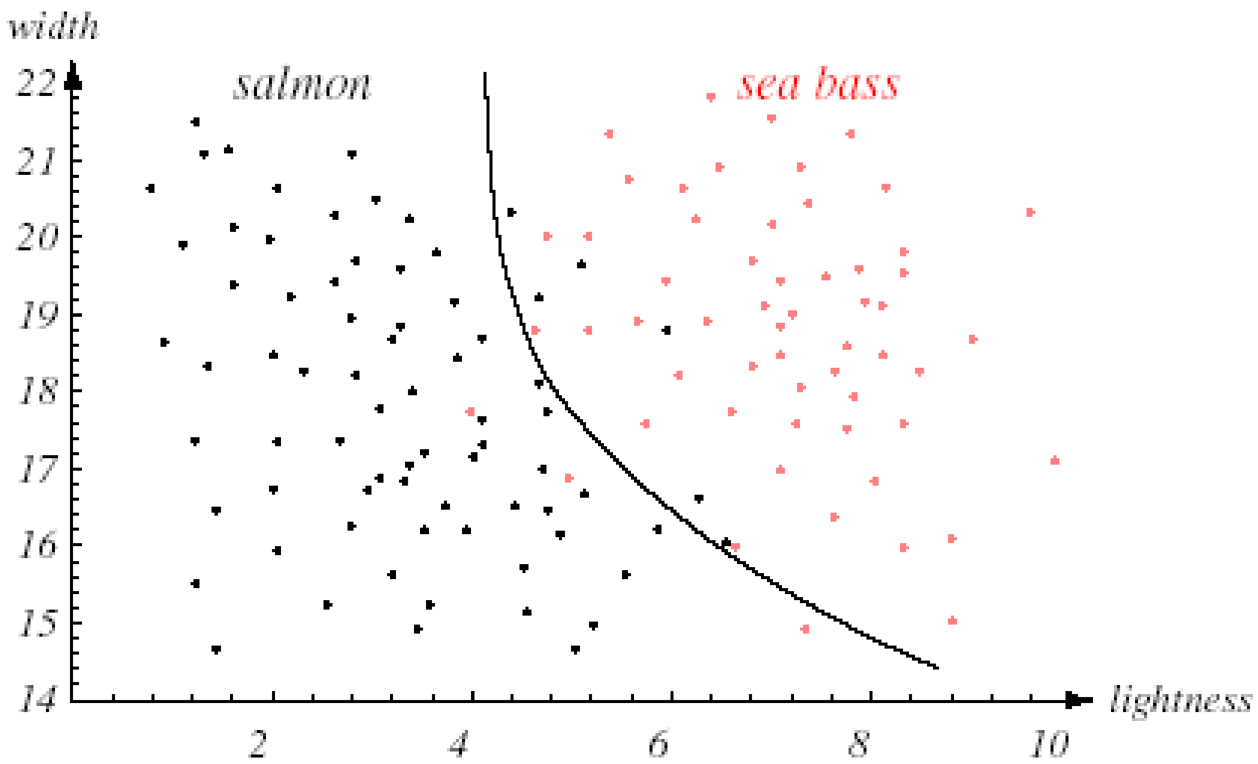
Salmon Vs. Sea Bass – Using two dimensional feature $x = (x_1, x_2)$




Too simple

Too complicated




Optimal tradeoff between performance and generalization

# *Importance of Features*

- Cannot be over-stated
- We usually don't know which to select, what they represent, and how to tune them (face, gait recognition, tumor detection, etc.)
- Classification and regression schemes are mostly trying to make the best of whatever features are available

# Features

- One is usually not descriptive (no silver bullet)
- Many (shotgun approach) can actually hurt
- Many problems:
  - Relevance
  - Dimensionality
  - Co-dpendency
  - Time and space varying characteristics.
  - Accuracy
  - Uncertainty and error
  - Missing values

# *Feature Selection (cont.)*

- Q: How to decide if a feature is effective?
- A: Through a training phase
  - Training on typical samples and typical features to discover
    - Whether features are effective
    - Whether there are any redundancy
    - The typical cluster shape (e.g., Gaussian)
    - Decision boundaries between samples
    - Cluster centers of particular samples
    - Etc.

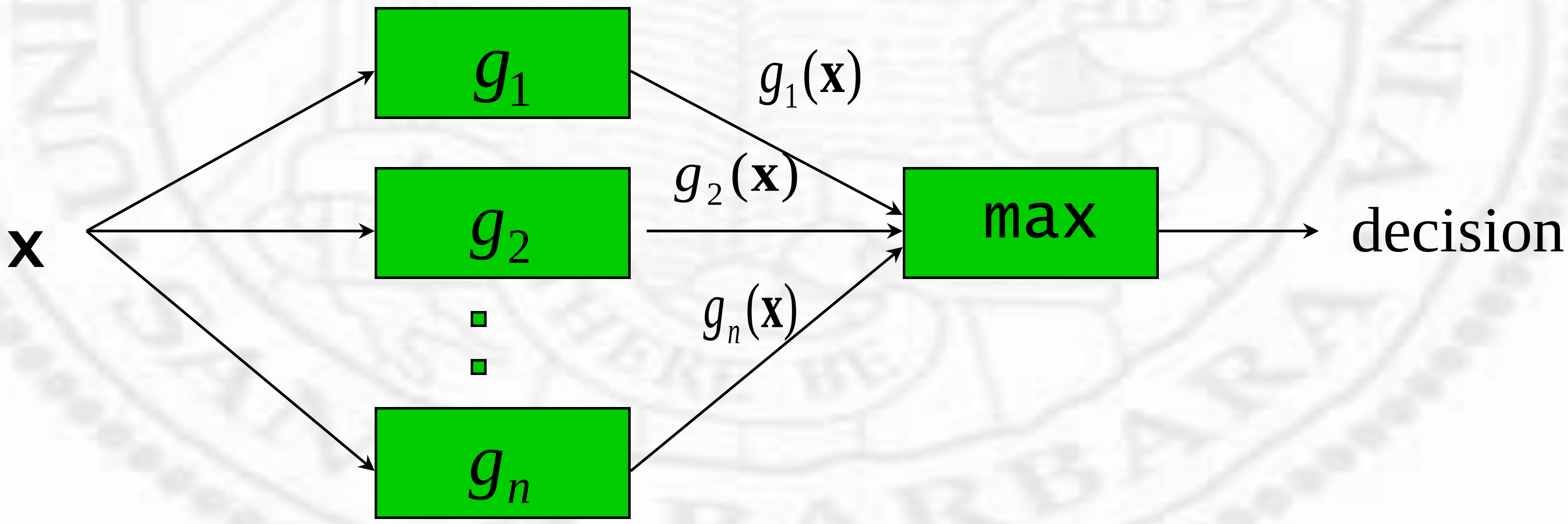
# Classifiers

$$\varpi_i \quad \text{if} \;\; g_i(x) > g_j(x) \;\; \text{for all} \;\; j \neq i$$

$$g_i(x) = P(\varpi_i)$$ if no measurements are made

$$g_i(x) = P(\varpi_i | x)$$ minimize misclassification rate

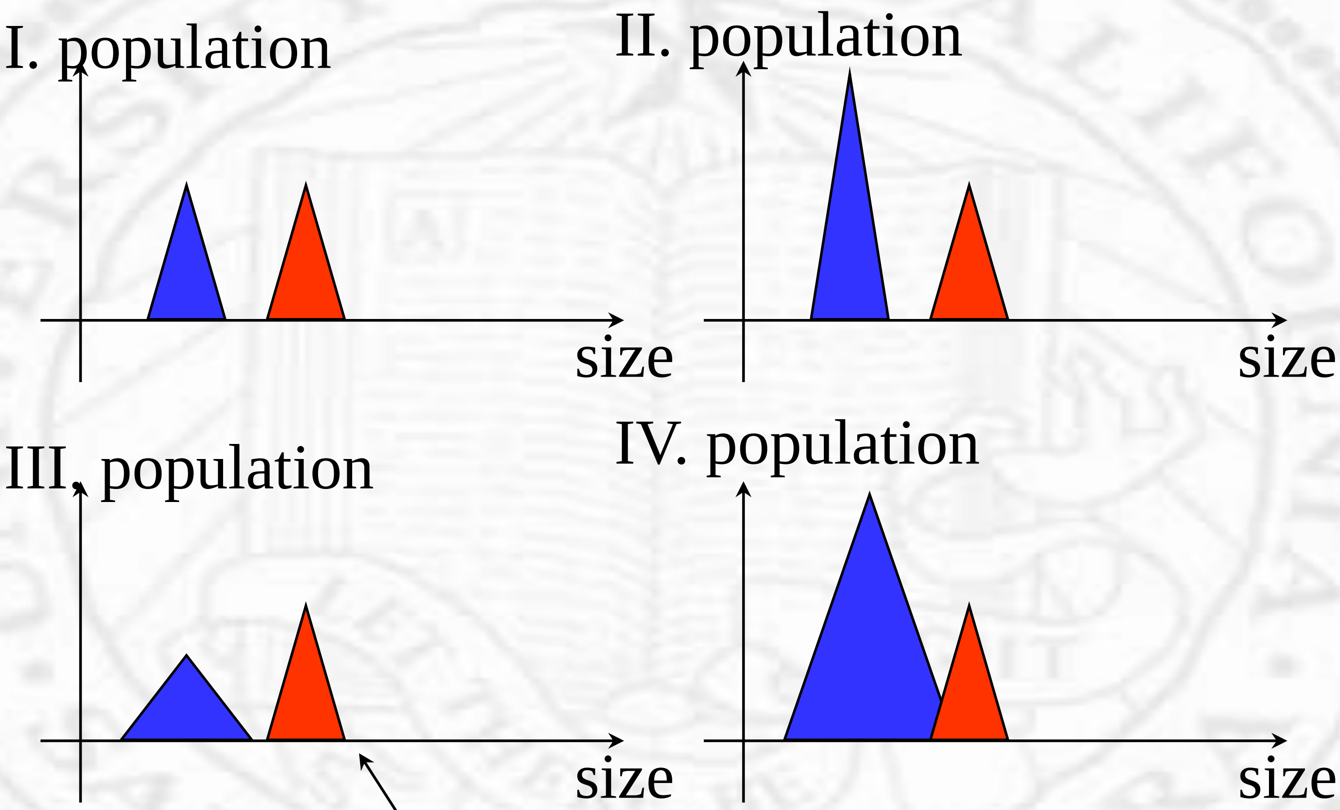
$$g_i(x) = R(\alpha_i | x)$$ minimize associated risk



# *Traditional Pattern Recognition*

- Parametric methods
  - Based on class sample exhibiting a certain parametric distribution (e.g. Gaussian)
  - Learn the parameters through training
- Density methods
  - Does not enforce a parametric form
  - Learn the density function directly
- Decision boundary methods
  - Learn the separation in the feature space

# *Parametric Methods*




$$\frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2}\frac{|x-\bar{x}|^2}{\sigma^2}}$$

# Density Methods

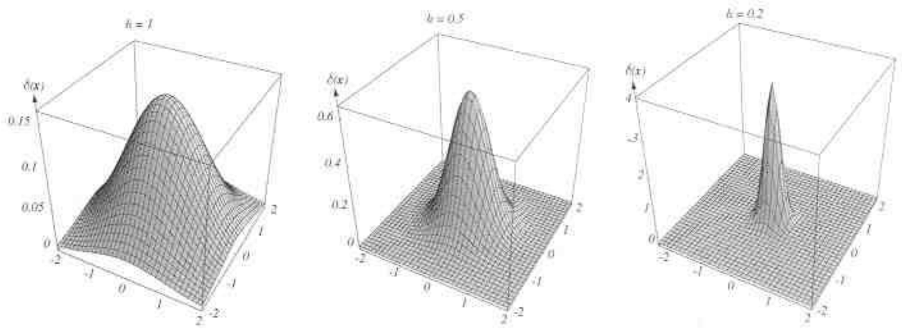
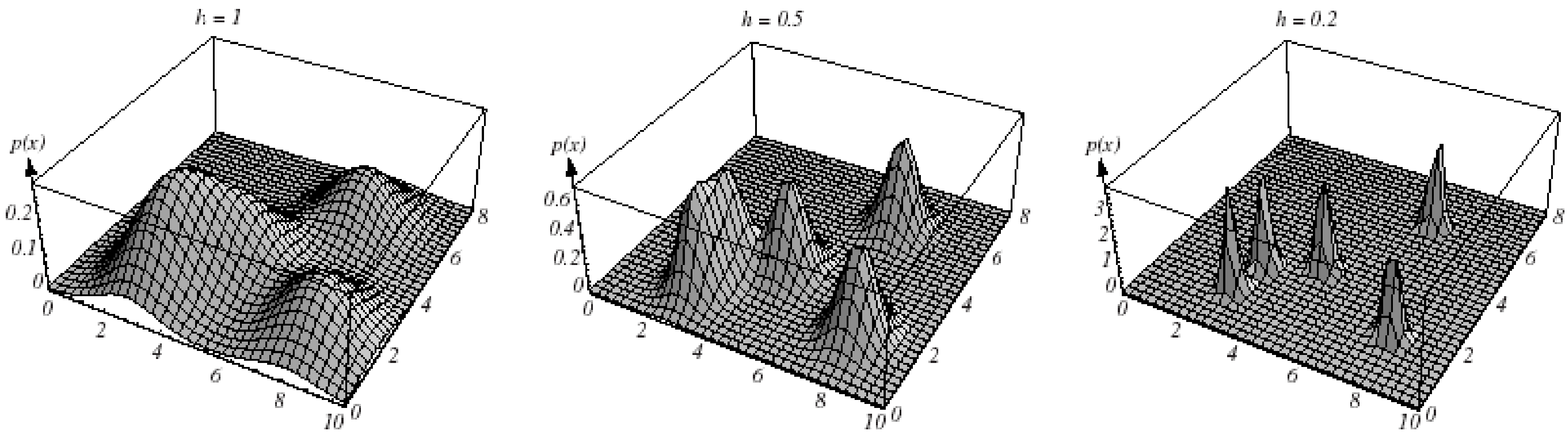
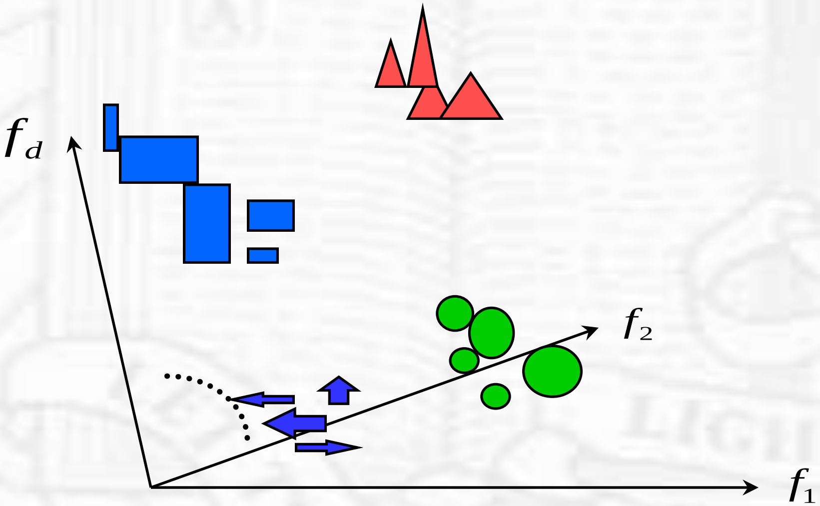



FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of $h$. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure.

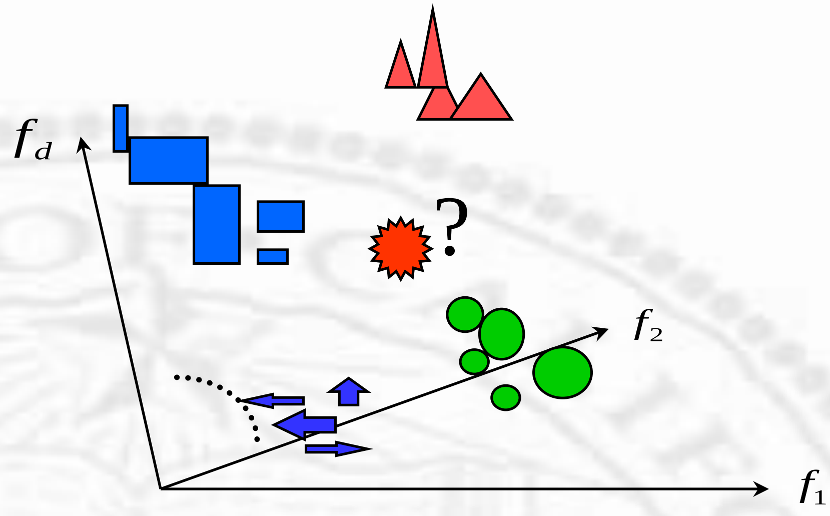


# Feature space

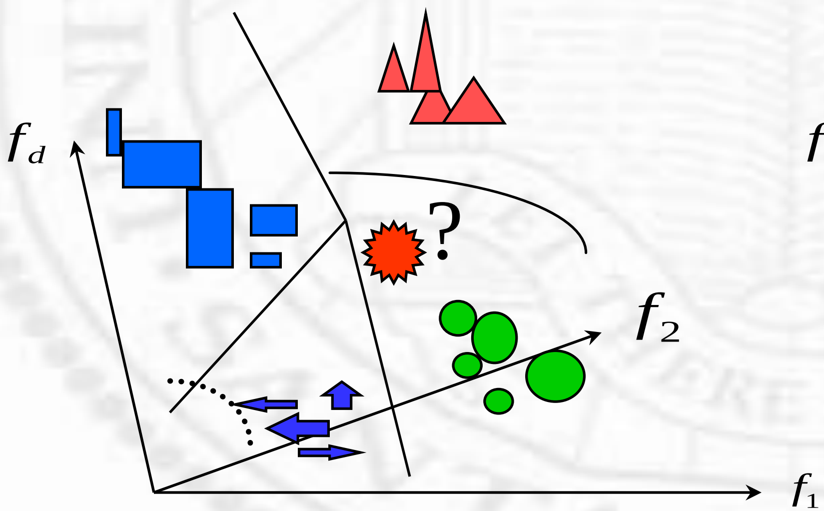
- d dimensional (d the number of features)
- populated with features from training samples



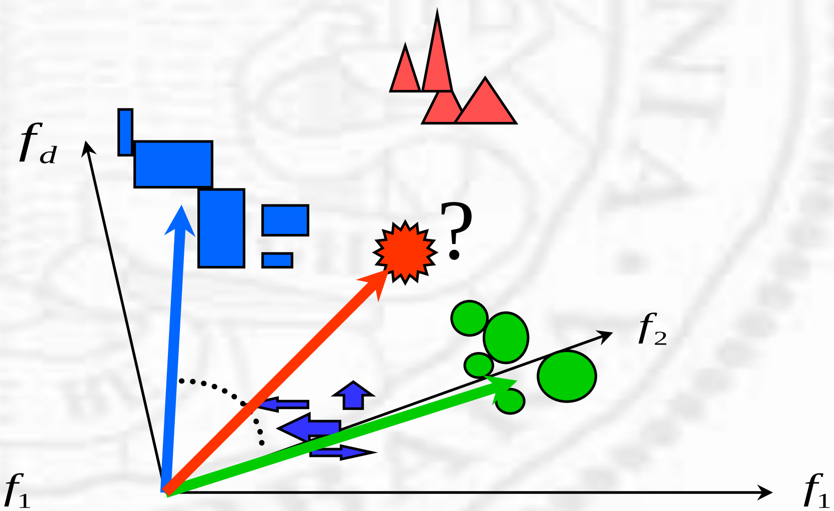
# Decision Boundary Methods




- Decision surfaces




- Cluster centers



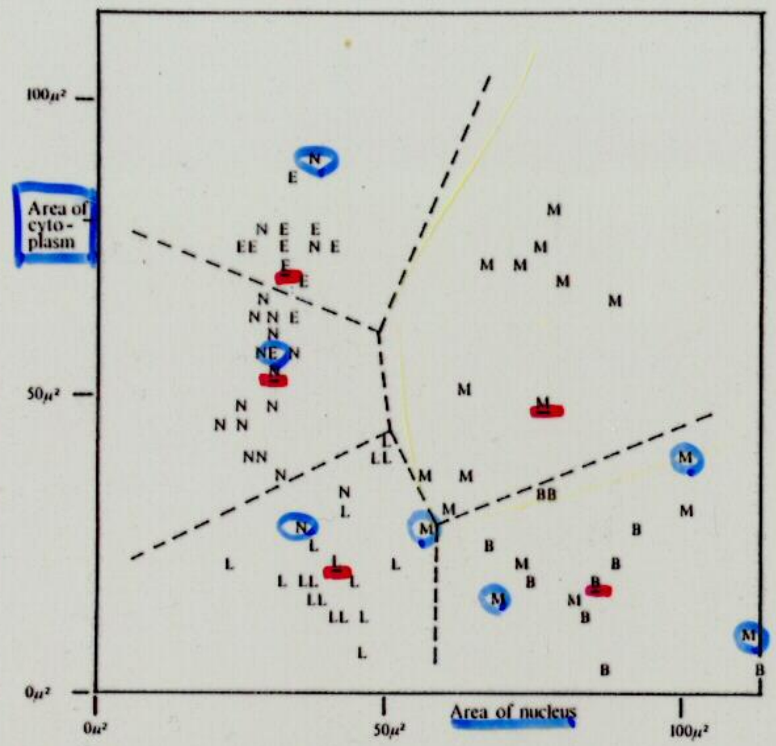



**Figure 14-1.** Scattergram of cytoplasm area versus nuclear area for five different common types of white blood cells. The letters denote the different classes, with the centroids underlined. The dashed lines show linear boundaries that best separate the classes. Several samples are misclassified. (Plotted from data in "Automated Leukocyte Recognition" by I.T. Young, Ph.D. thesis, MIT, Cambridge, Massachusetts, 1969.)

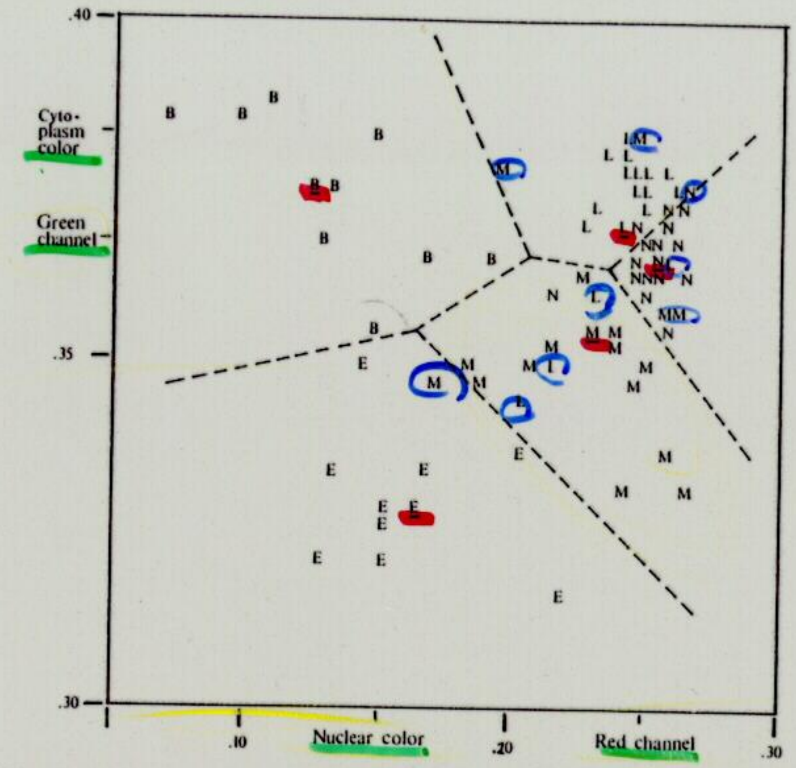


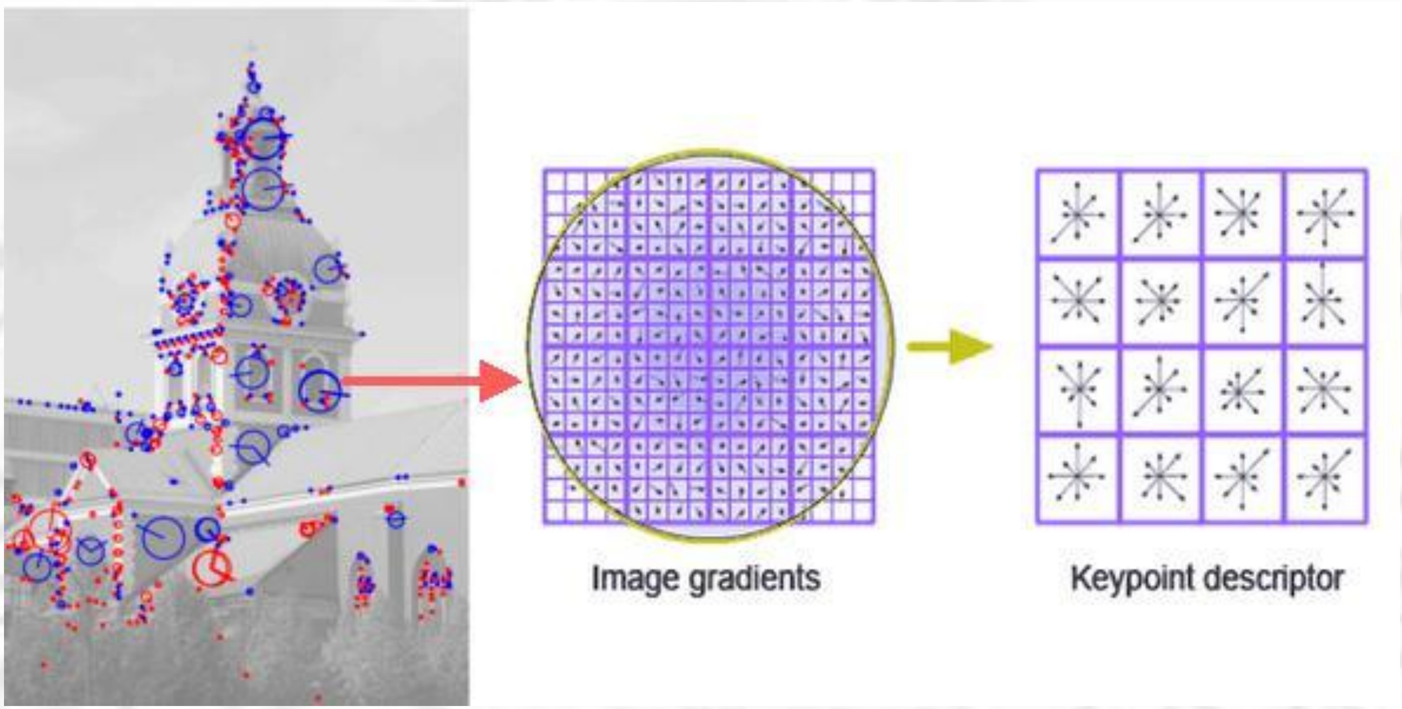

**Figure 14-2.** Scattergram of brightness of the cytoplasm and the nucleus measured through two different filters. The centroids are indicated by underlining, and the dashed lines are the linear boundaries that best separate the classes. It is clear that reliable classification using just these two features is not possible. (Plotted from data in "Automated Leukocyte Recognition" by I.T. Young, Ph.D. thesis, MIT, Cambridge, Massachusetts, 1969.)

# *"Modern" vs "Traditional" Pattern Recognition*

- Hand-crafted features
- Simple and low-level concatenation of numbers or traits
- Syntactic
- Feature detection and description are separate tasks from classifier design

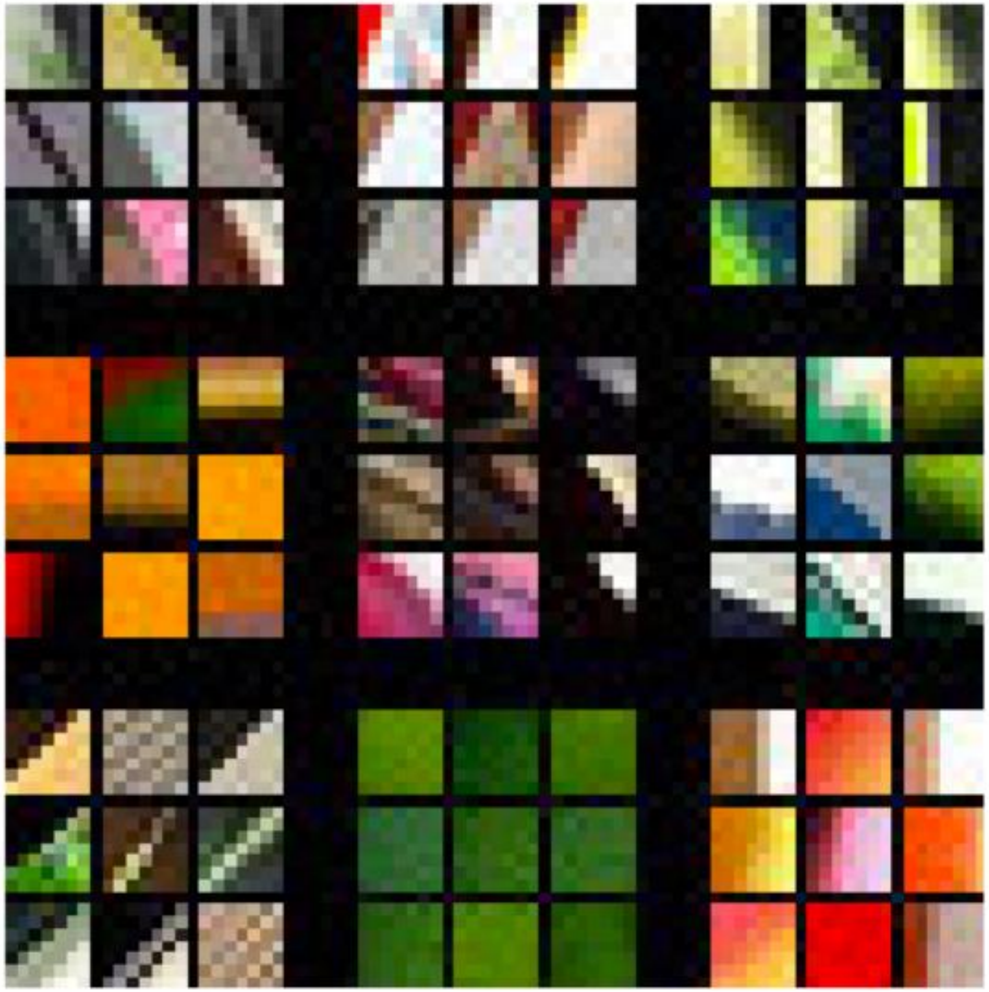
- Automatically learned features
- Hierarchical and complex
- Semantic
- Feature detection and description are not jointly optimized with classifiers

# Traditional Features



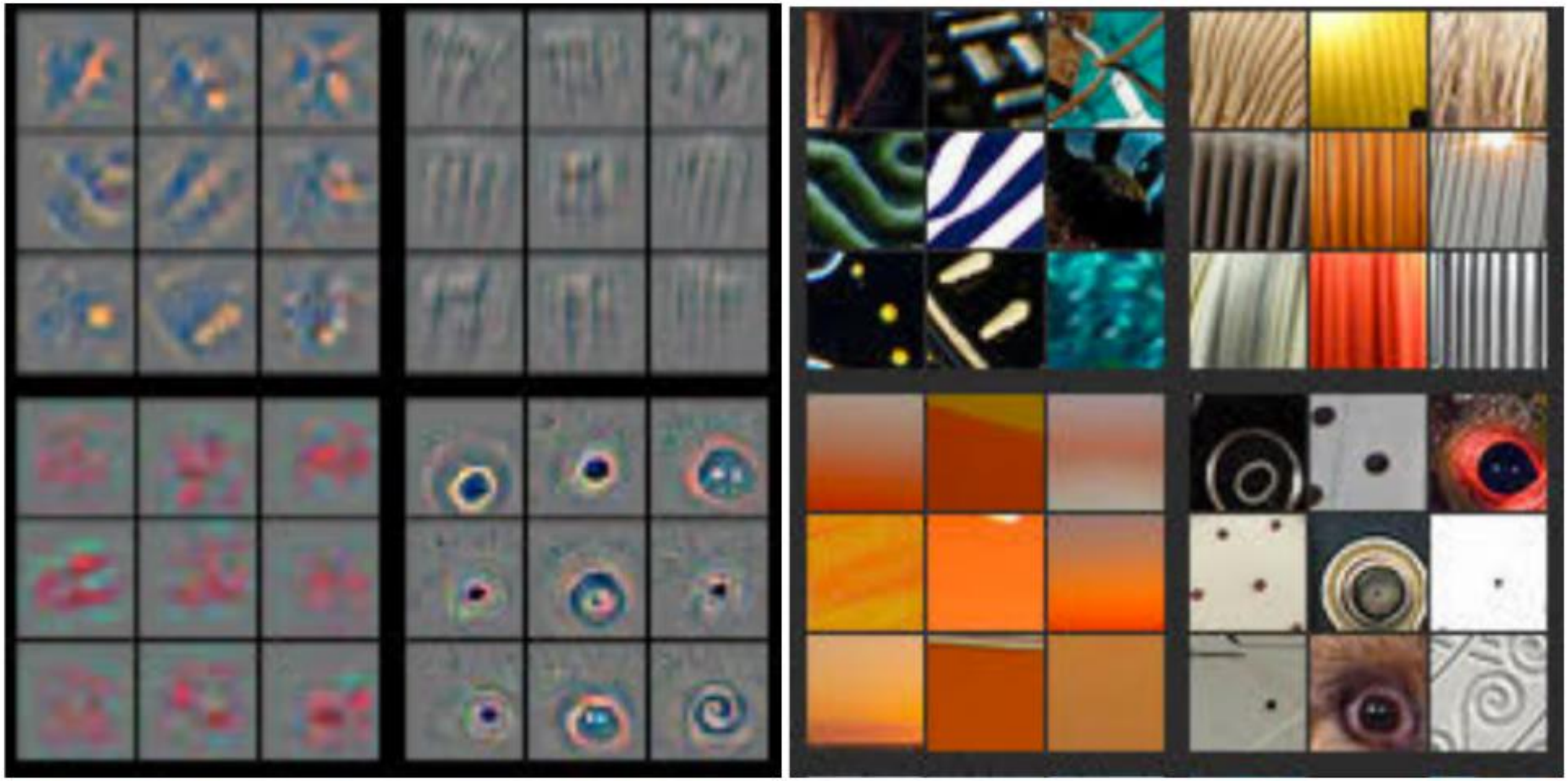

# Modern Features

- Layer1



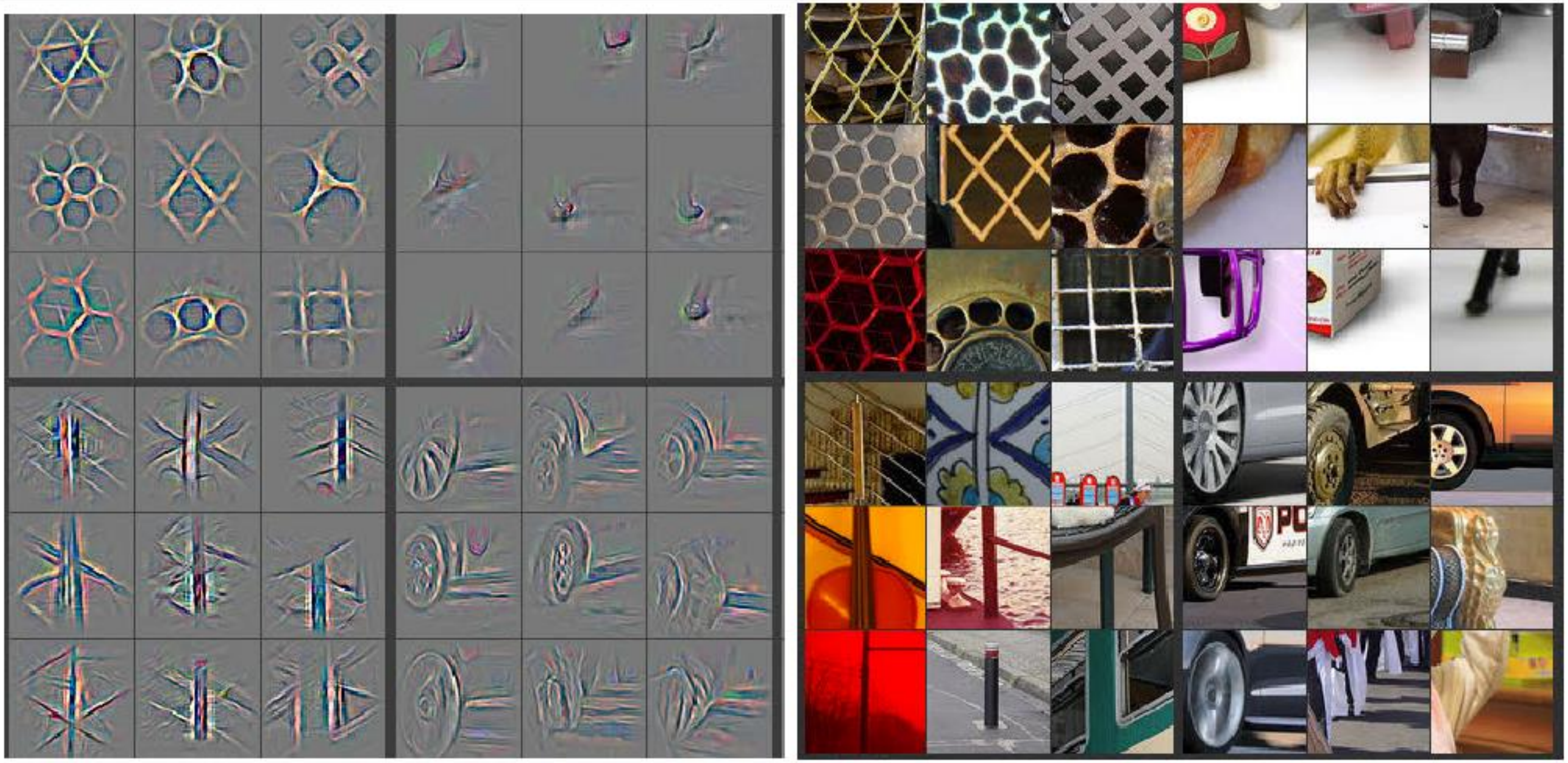
# Modern Features

- Layer2



# Modern Features

- Layer3

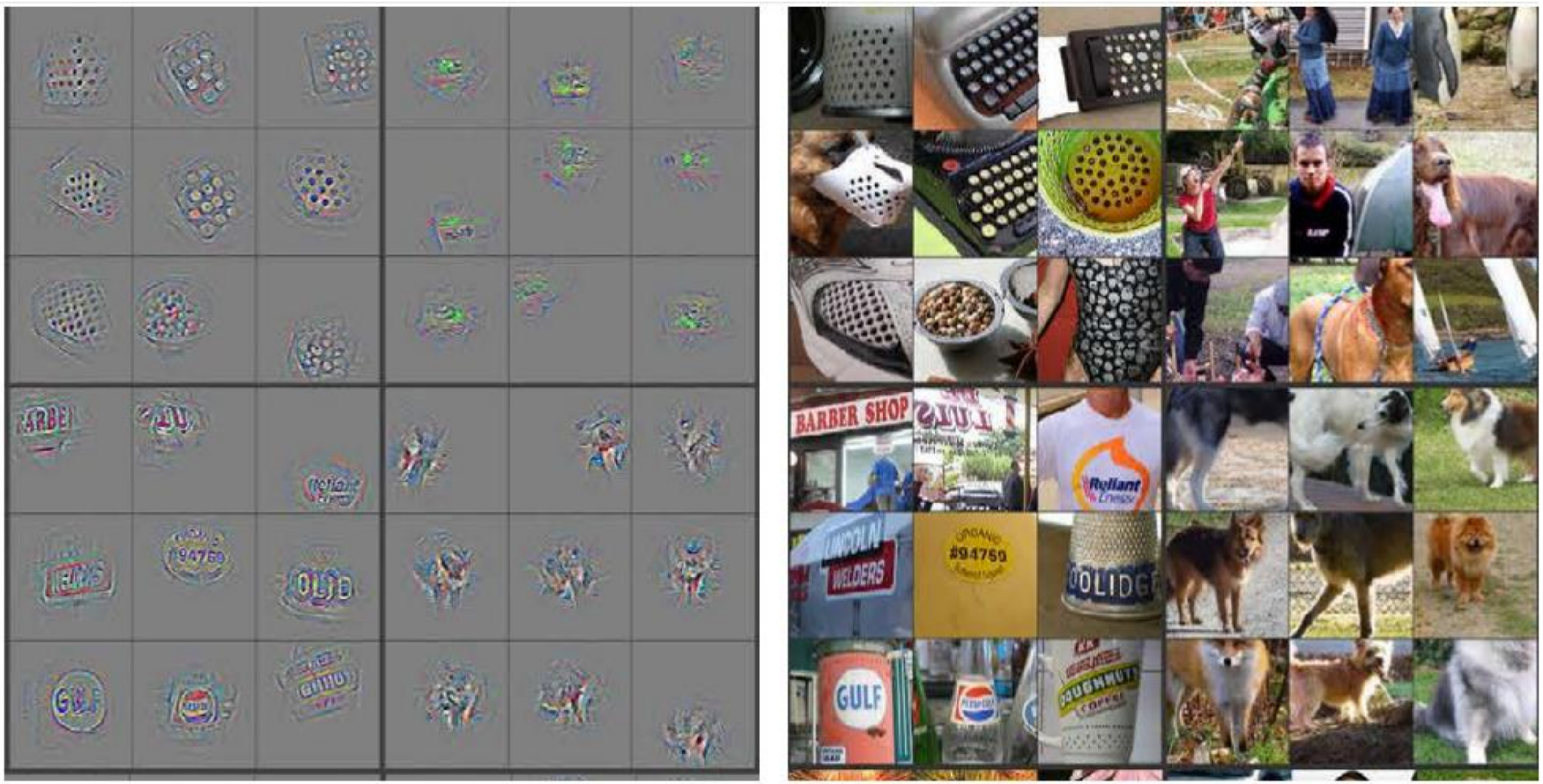


# Modern Features

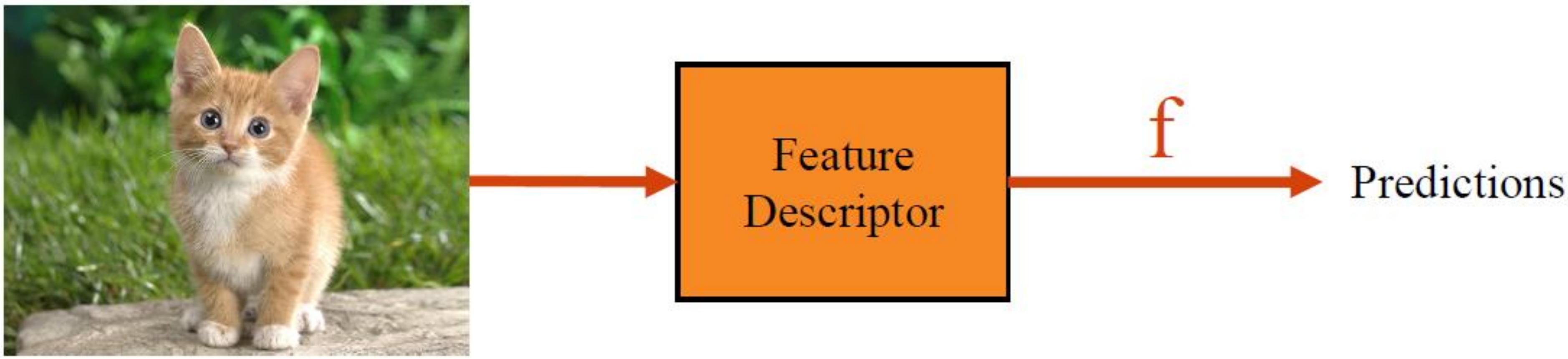
- Layer4

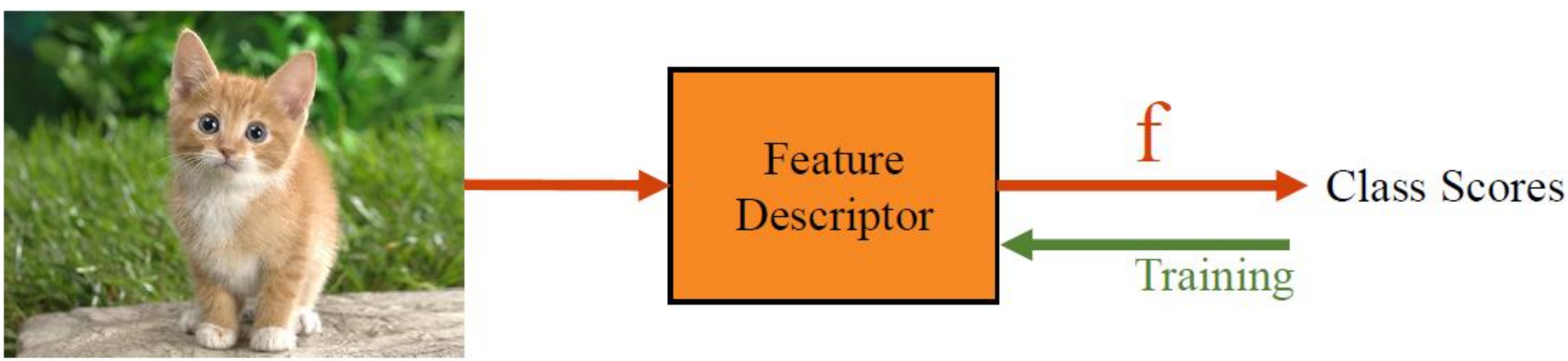
# Modern Features

- Layer5



# *"Modern" vs "Traditional" Pattern Recognition*



# *Mathematical Foundation*

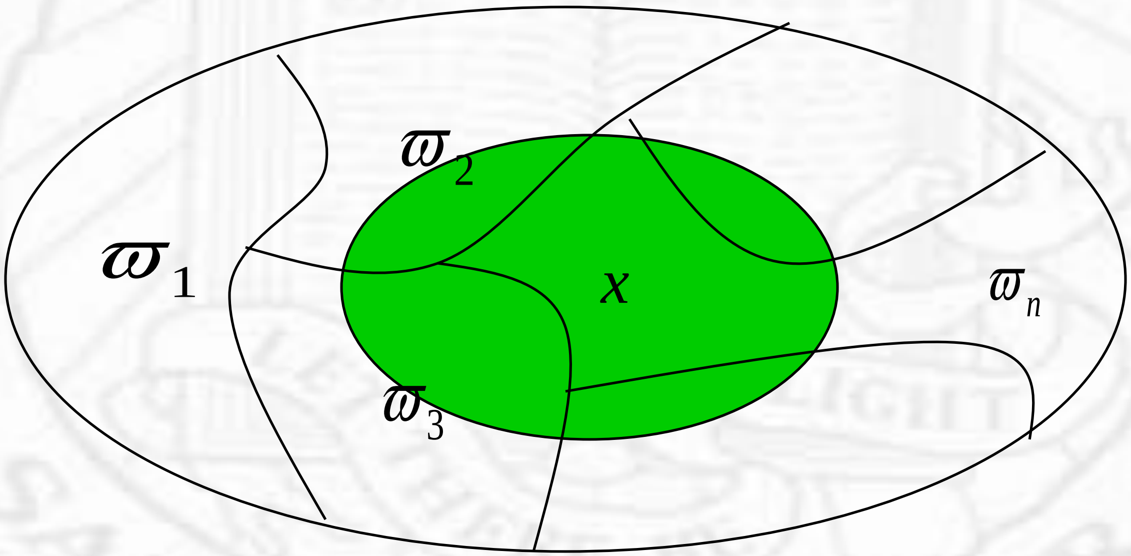
- Does not matter what methods or techniques you use, the underlying mathematical principle is quite simple
- Bayesian theory is the foundation

# Review: Bayes Rule

- Forward (*synthesis*) route:
  - From class to sample in a class
    - Grammar rules to sentences
    - Markov chain (or HMM) to pronunciation
    - Texture rules (primitive + repetition) to textures
- Backward (*analysis*) route:
  - From sample to class ID
    - A sentence parsed by a grammar
    - A utterance is "congratulations" (not "constitution")
    - Brickwall vs. plaid shirt

# Review: Bayes Rule

- Backward is always harder
  - Because the interpretation is not unique
  - Presence of x has multiple possibilities

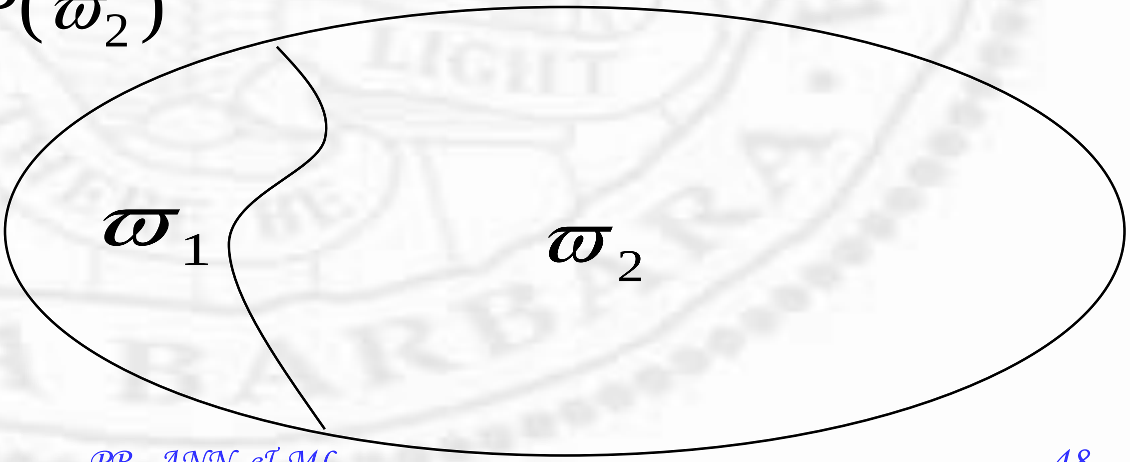


# The simplest example

- Two classes: pennies and dimes
- No measurements
- Classification:
  - based on the a prior probabilities
- Error rate:

$$
\begin{array}{ll}
\varpi_1 & \text{if } P(\varpi_1) > P(\varpi_2) \\
\varpi_2 & \text{if } P(\varpi_1) < P(\varpi_2) \\
\varpi_1 \text{ or } \varpi_2 & \text{otherwise}
\end{array}
$$

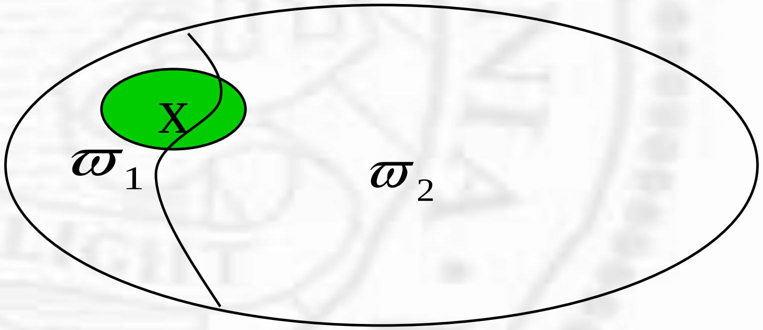
$$\min(P(\varpi_1), P(\varpi_2))$$



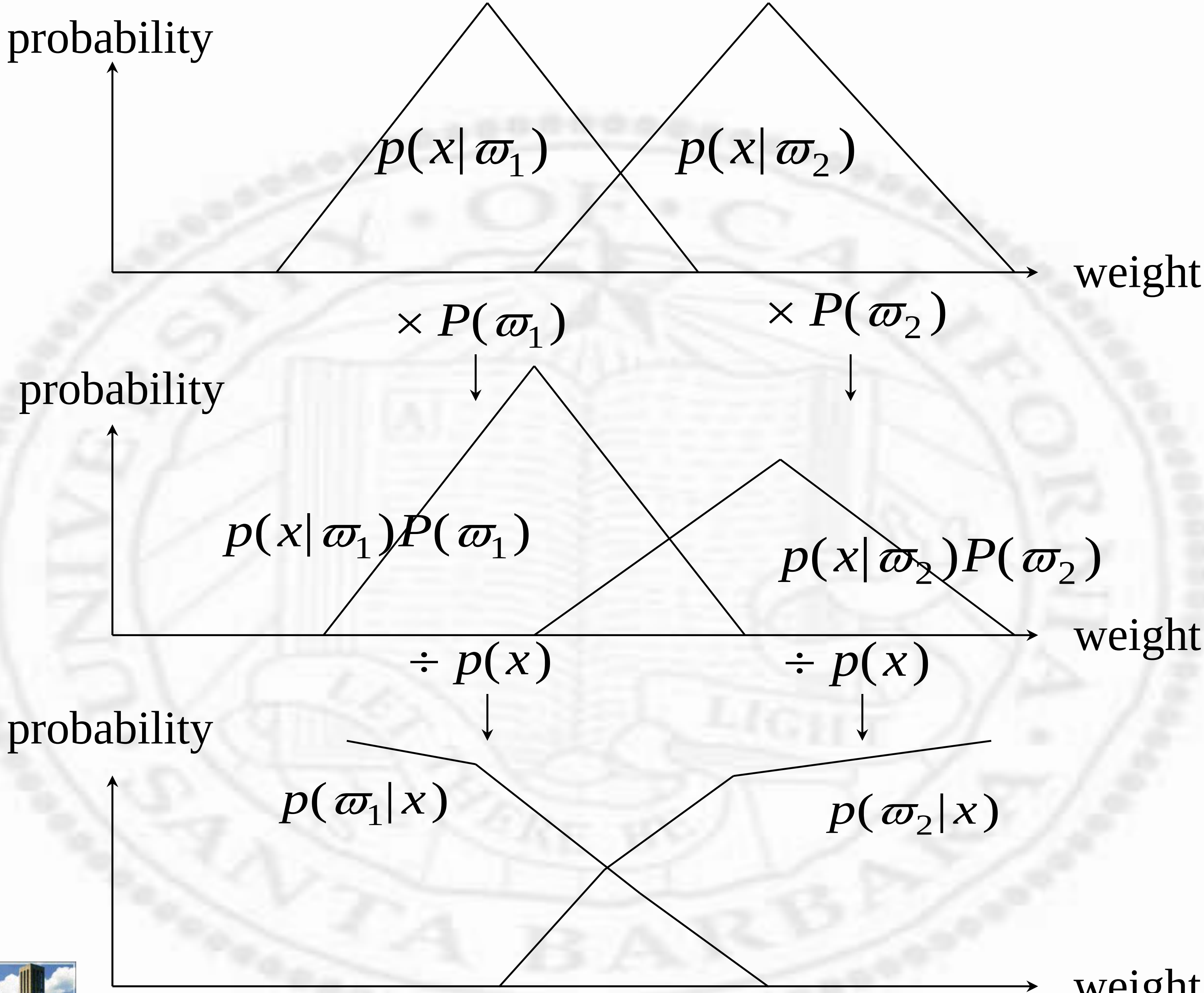
# A slightly more complicated example

- Two classes: pennies and dimes
- A measurement x is made (e.g. weight)
- Classification
  - based on the a posterior probabilities with Bayes rule

$$\begin{array}{ll} \varpi_1 & if\ \ P(\varpi_1|x) > P(\varpi_2|x) \\ \varpi_2 & if\ \ P(\varpi_1|x) < P(\varpi_2|x) \\ \varpi_1\ or\ \varpi_2 & otherwise \end{array}$$




$$P(\varpi_i|x) = \frac{p(x,\varpi_i)}{p(x)} = \frac{p(x|\varpi_i)P(\varpi_i)}{p(x)}$$


probability
$p(x|\varpi_1)$
$p(x|\varpi_2)$
weight
$\times P(\varpi_1)$
$\times P(\varpi_2)$
probability
$p(x|\varpi_1)P(\varpi_1)$
$p(x|\varpi_2)P(\varpi_2)$
weight
$\div p(x)$
$\div p(x)$
probability
$p(\varpi_1|x)$
$p(\varpi_2|x)$
weight

# Why Both?

$$p(x \mid \varpi_i) \quad \& \quad P(\varpi_i)?$$

- In the day time, some animal runs in front of you on the bike path, you know exactly what it is (p(x|w) is sufficient)
- In the night time, some animal runs in front of you on the bike path, you can hardly distinguish the shape (p(x|w) is low for all cases, but you know it is probably a squirrel, not a lion because of p(w))

# *Essence*

- Turn a backward (analysis) problem into several forward (synthesis) problem
- Or analysis-by-synthesis
- Whichever model has a highly likelihood of synthesizing the outcome wins
- The formula is not mathematically provable

# Error rate

- Determined by
  - The likelihood of a class
  - The likelihood of measuring x in a class

$$\min(P(\varpi_1|x), P(\varpi_2|x)) \quad or$$

$$\frac{1}{p(x)}\min(p(x|\varpi_1)P(\varpi_1), p(x|\varpi_2)P(\varpi_2))$$

# Error Rate (cont.)

- Bayes Decision Rule minimizes the average error rate:

$$error = \int p(error \mid x)\, p(x)\, dx$$

$$p(error \mid x) = \sum_{\varpi_i \neq \varpi_{(x)}^*} p(\varpi_i \mid x) = 1 - p(\varpi_{(x)}^* \mid x)$$

*where*

$$\varpi_{(x)}^* = \arg\max_i \; p(\varpi_i \mid x)$$

# Various types of errors

| | | Condition (as determined by "Gold standard") | | |
|---|---|---|---|---|
| | | Condition positive | Condition negative | |
| Test outcome | Test outcome positive | True positive | False positive (Type I error) | Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$ |
| | Test outcome negative | False negative (Type II error) | True negative | Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$ |
| | | Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ | Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$ | Accuracy |

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

# Key quantities as fractions

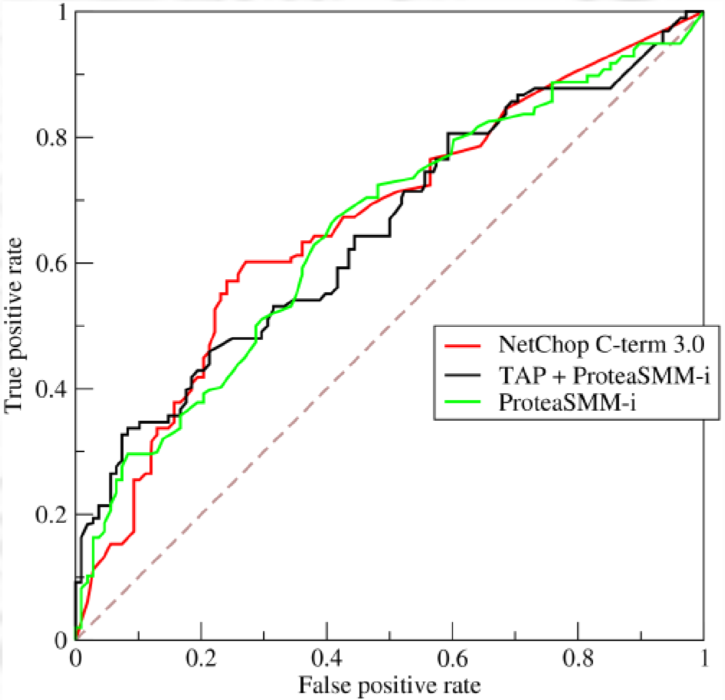
| | | DISEASE | |
|---|---|---|---|
| | | + | − |
| TEST | + | TP | FP |
| | − | FN | TN |

| | | |
|---|---|---|
| Sensitivity | → | TP / (TP+FN) |
| Specificity | → | TN / (FP+TN) |
| Positive Predictive Value | → | TP / (TP+FP) |
| Negative Predictive Value | → | TN / (FN+TN) |
| Accuracy | → | (TP+TN) / (TP+FP+FN+TN) |

# Precision vs. Recall

- A very common measure used in PR and MI community
- One goes up and the other HAS to go down
- A range of options (Receiver operating characteristic curves)
- Area under the curve

as a goodness measure



# *Various ways to measure error rates*

- Training error
- Test error
- Empirical error
- Some under your control (training and test)
- Some not (empirical error)
- How: n-fold validation
- Why: Overfitting and underfitting problems

# An even more complicated example

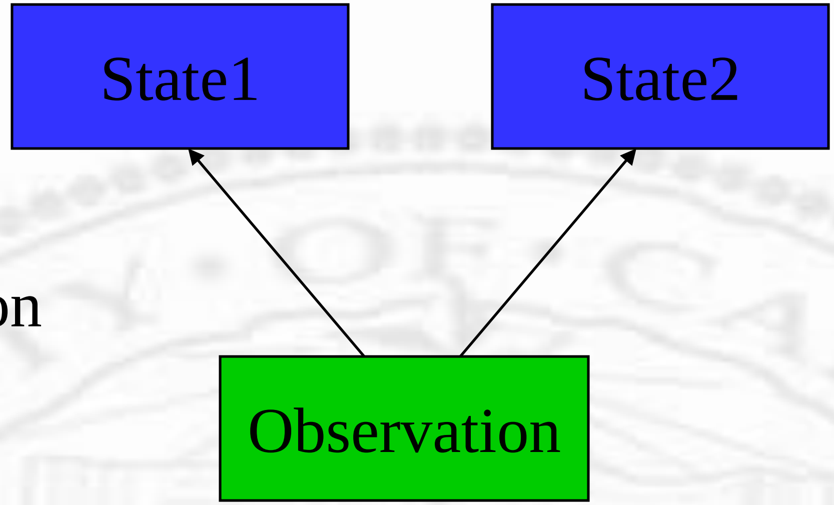
- Two classes: pennies or dimes
- A measurement x is made
- Risk associated with making a wrong decision
- Based on the a posterior probabilities with Bayesian risk

$$R(\alpha_1|x) = \lambda_{11}P(\varpi_1|x) + \lambda_{12}P(\varpi_2|x)$$

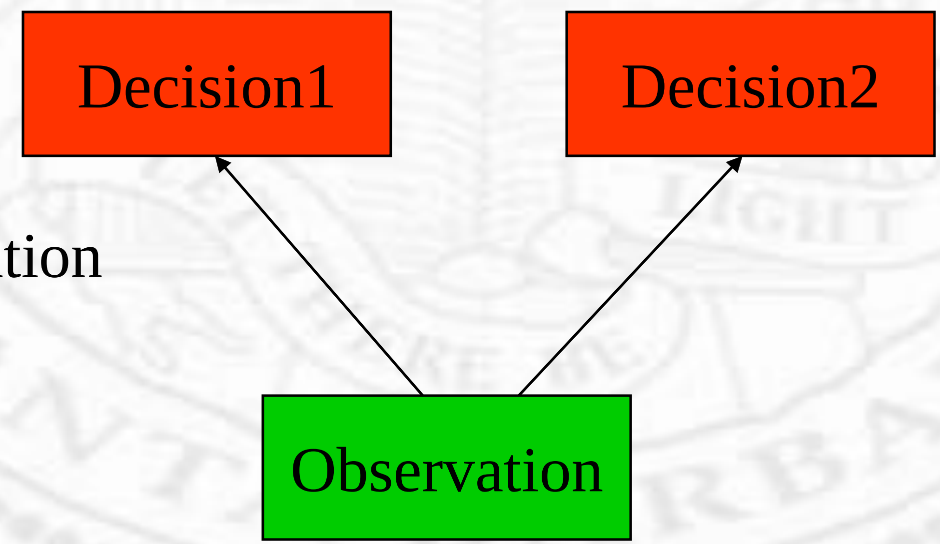
$$R(\alpha_2|x) = \lambda_{21}P(\varpi_1|x) + \lambda_{22}P(\varpi_2|x)$$

$$\lambda_{ij}: the\ loss\ of\ action\ \alpha_i\ in\ state\ \varpi_j$$

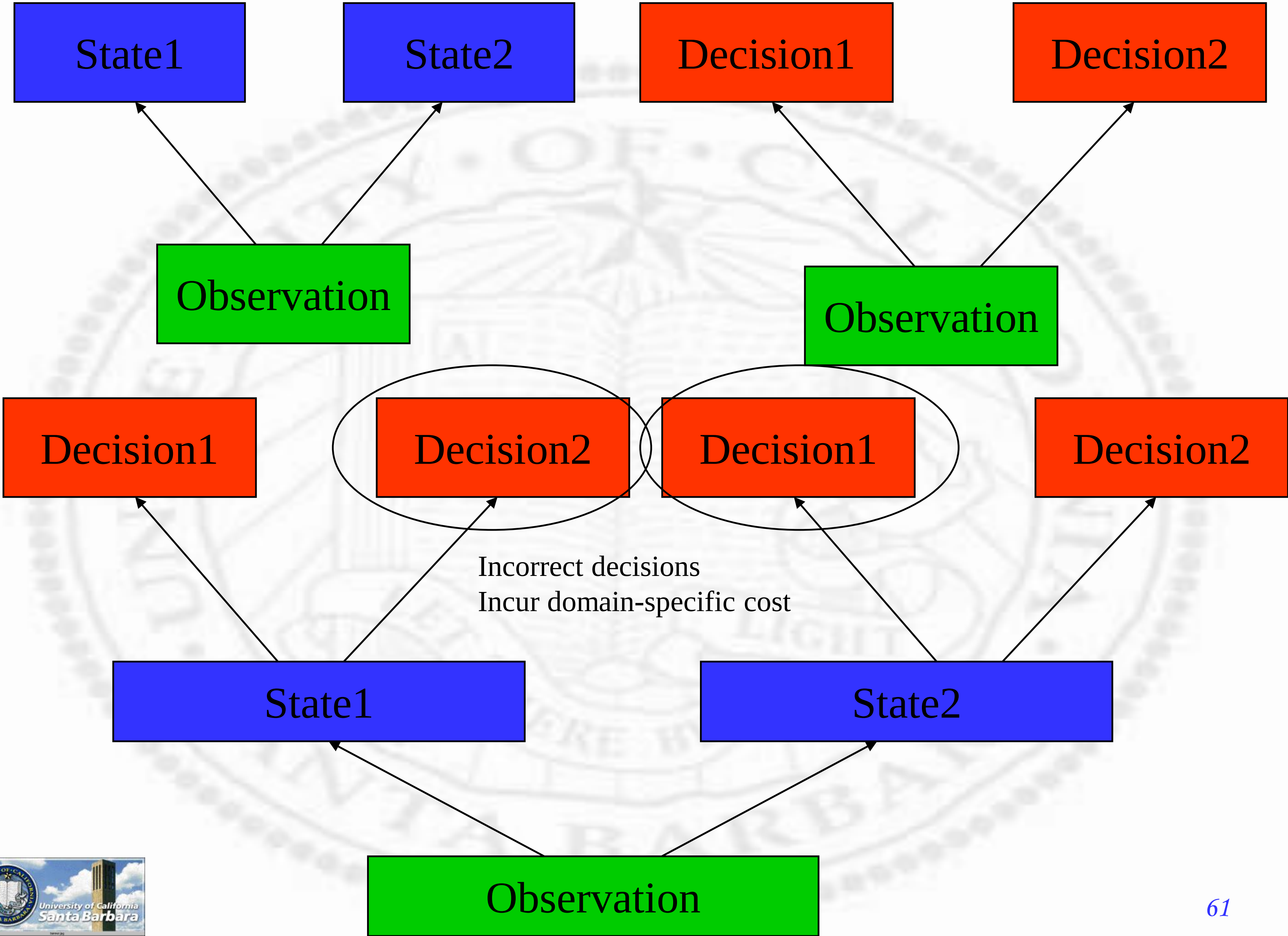
$$R(\alpha_i|x): the\ conditional\ risk\ of\ action\ \alpha_i\ with\ x$$




Mis-classification
Math




Mis-interpretation
Human factor


State1
State2
Decision1
Decision2
Observation
Observation
Decision1
Decision2
Decision1
Decision2
Incorrect decisions
Incur domain-specific cost
State1
State2
Observation

# *An even more complicated example*

R(used as pennies | x) =

$\boxed{\text{p(x|pennies)P(pennies)}}$

r(pennies used as pennies) * $\boxed{\text{P(pennies | x)}}$ +

r(dimes used as pennies) * $\boxed{\text{P(dimes | x)}}$

$\boxed{\text{p(x|dimes)P(dimes)}}$

R(used as dimes | x) =

r(pennies used as dimes) * P(pennies | x) +

r(dimes used as dimes) * P(dimes | x)

# *A more credible example*

R(call FD|smoke) =

r(call,fire)*P(fire|smoke) +

r(call, no fire)*P(no fire|smoke)

False positive

R(no call FD|smoke)=

r(no call, no fire)*P(no fire|smoke) +

r(no call, fire)*P(fire|smoke)

False negative

- The risk associated with false negative is much higher than that of false positive

# A more credible example

R(attack|battle field intelligence) =

r(attack,<50%)*P(<50%|intelligence) +

$\boxed{\text{r(attack,>50\%)}}$*P(>50%|intelligence)

False positive

R(no attack|battle field intelligence)=

r(no attack, >50%)*P(>50%|intelligence) +

$\boxed{\text{r(no attack, <50\%)}}$*P(<50%|intelligence)

False negative

# Baysian Risk

- Determined by
  - likelihood of a class
  - likelihood of measuring x in a class
  - the risk of making a wrong action
- Classification
  - Baysian risk should be minimized

$$
\begin{aligned}
&\min(R(\alpha_1 \mid x), R(\alpha_2 \mid x)) or \\
&\min(\lambda_{11}P(\varpi_1 \mid x) + \lambda_{12}P(\varpi_2 \mid x), \lambda_{21}P(\varpi_1 \mid x) + \lambda_{22}P(\varpi_2 \mid x)) \quad or \\
&R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \Rightarrow \varpi_1 \\
&(\lambda_{21} - \lambda_{11})P(\varpi_1 \mid x) > (\lambda_{12} - \lambda_{22})P(\varpi_2 \mid x)
\end{aligned}
$$

# Bayesian Risk (cont.)

- Again, decisions depend on
  - likelihood of a class
  - likelihood of observation of x in a class
  - Modified by some positive risk factors
- Why?
  - Because in the real world, it might not be the misclassification rate that is important, it is the action you assume

$$(\lambda_{21} - \lambda_{11}) P(\varpi_1 \mid x) > (\lambda_{12} - \lambda_{22}) P(\varpi_2 \mid x)$$

# *Other generalizations*

- Multiple classes
  - n classes

$$\sum_{i=1}^{n} P(\varpi_i) = 1$$

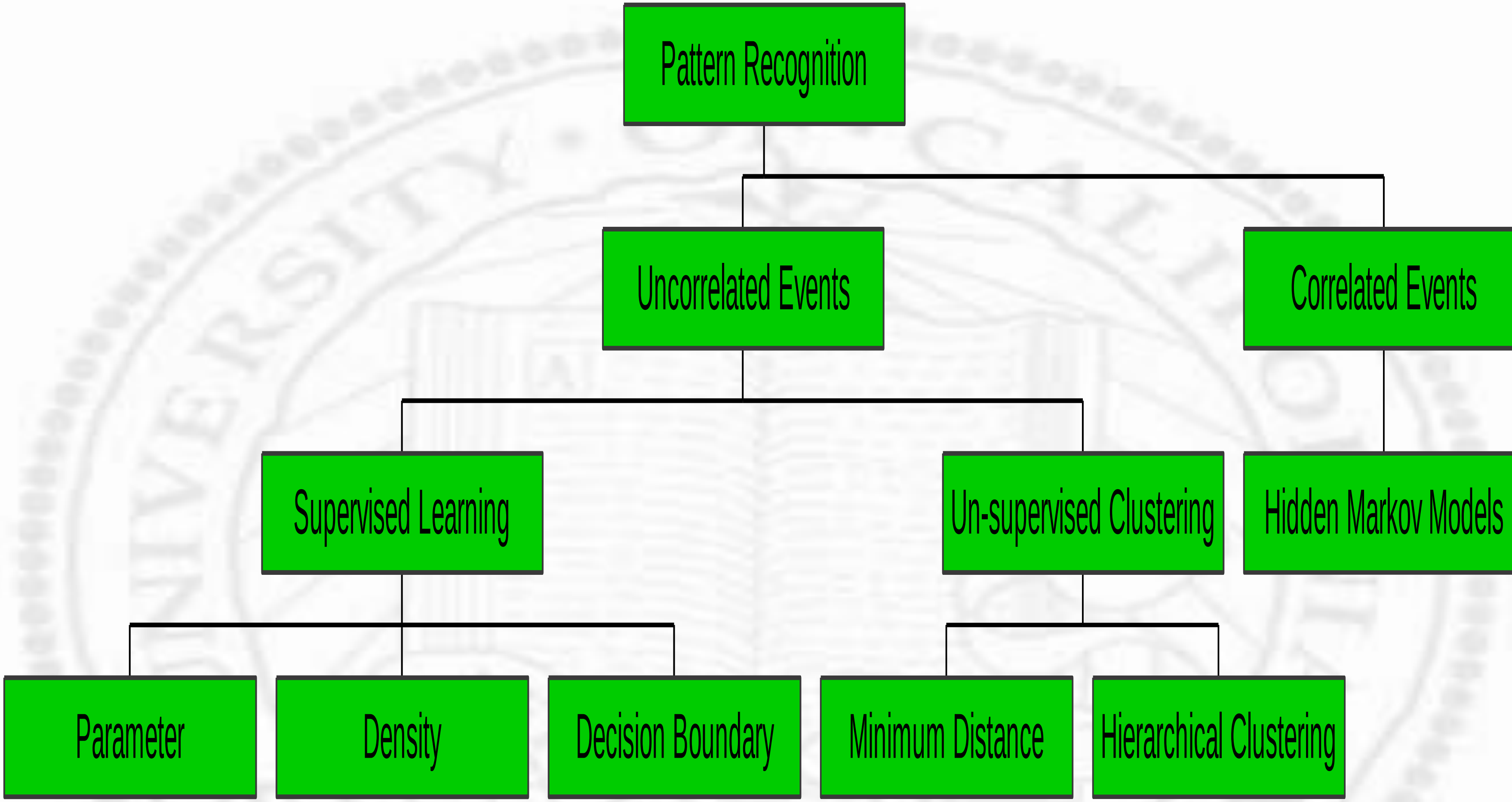
- Multiple measurements
  - X is a vector instead of a scalar
- Non-numeric measurements
- Actions vs. decisions
- Correlated vs. independent events
  - speech signals and images
- Training allowed or not
- Time-varying behaviors

# Difficulties

- What features to use
- How many features (the curse of dimensionality)
- The a prior probability $P(\varpi_i)$
- The class-conditional density $p(x|\varpi_i)$
- The a posterior probability $P(\varpi_i \mid x)$

# *Typical Approaches*

- Supervised (with tagged samples x):
  - parameters of a probability function (e.g. Gaussian ) $p(x|\varpi_i) = N(\mu_i, \Sigma_i)$
  - density functions (w/o assuming any parametric forms)
  - decision boundaries (classes are indeed separable)
- Unsupervised (w/o tagged samples x):
  - minimum distance
  - hierarchical clustering
- Reinforced (with hints)
  - Right or wrong, but not correct answer
  - Learning with a critic (not a teacher as in supervised)


Pattern Recognition
Uncorrelated Events
Correlated Events
Supervised Learning
Un-supervised Clustering
Hidden Markov Models
Parameter
Density
Decision Boundary
Minimum Distance
Hierarchical Clustering

# *Applications*

- DNA sequence
- Lie detectors
- Handwritten digits recognition
- Classification based on smell
- Web document classification and search engine
- Defect detection
- Texture classification
- Image database retrieval
- Face recognition
- etc.

# Other formulations

- We talked about 1/3 of the scenarios – that of classification (discrete)
- Regression – continuous
  - Extrapolation and interpolation
- Clustering
  - Similarity
  - Abnormality detection
  - Concept drift (discovery), etc.

# Classification vs. Regression

- Classification
- Large vs. small hints on category
- Absolute values does not matter as much (can actually hurt)
- Normalization is often necessary
- Fitting order stays low

- Regression
- Large means large, small means small
- Absolute values matter
- Fitting orders matter

# Recent Development

- Data can be "massaged" Surprisingly, massaging the data and use simple classifiers is better than massaging the classifiers and use simple data (for simple problems & small data sets)
- Hard-to-visualize concept
  - Transform data into higher dimensional space (e.g., infinite dimensional) has a tendency to separate data and increase error margin
- Concept of SVM and later kernel methods

# More Recent Development

- Think about fitting linear data with a model
  - Linear, quadratic, cubic, etc.
- Higher the order, better the fit
  - n data points can be perfectly fit by an (n-1) order polynomial
- However
  - Overfitting is likely
  - No ability to extrapolate
- “Massage” the classifiers (using deep networks)
  - Feature detection and description
  - Classification
  - Jointly optimization