Statistical Learning- Classification

STAT 441/ 841, CM 764

Ali Ghodsi

Department of Statistics

and

Actuarial Science

University of Waterloo

aghodsib@uwaterloo.ca

Two Paradigms

Classical Statistics

- Infer information from small data sets (Not enough data)

Machine Learning

- Infer information from large data sets (Too many data)

We are drowning in information and starving for knowledge.

Rutherford D. Roger

Fundamental problems

- Classification
- Regression
- Clustering
- Representation Learning (Feature extraction, Manifold learning, Density estimation)

Applications

Machine Learning is most useful when the structure of the task is not well understood but can be characterized by a dataset with strong statistical regularity.

- Search and recommendation (e.g. Google, Amazon)
- Automatic speech recognition and speaker verification
- Text parsing
- Face identification
- Tracking objects in video
- Financial prediction, fraud detection (e.g. credit cards)
- Medical diagnosis

Applications

Machine Learning is most useful when the structure of the task is not well understood but can be characterized by a dataset with strong statistical regularity.

- Search and recommendation (e.g. Google, Amazon)
- Automatic speech recognition and speaker verification
- Text parsing
- Face identification
- Tracking objects in video
- Financial prediction, fraud detection (e.g. credit cards)
- Medical diagnosis

More Applications

More science and technology applications:

- handwritten identification
- drug discovery

  (to identify the biological activity of chemical compounds using features describing the chemical structures)
- Gene expression analysis (thousands of features with only dozens of observations)

Classification

Classification:

Predicting a discrete random variable Y from another random variable X.

Classification

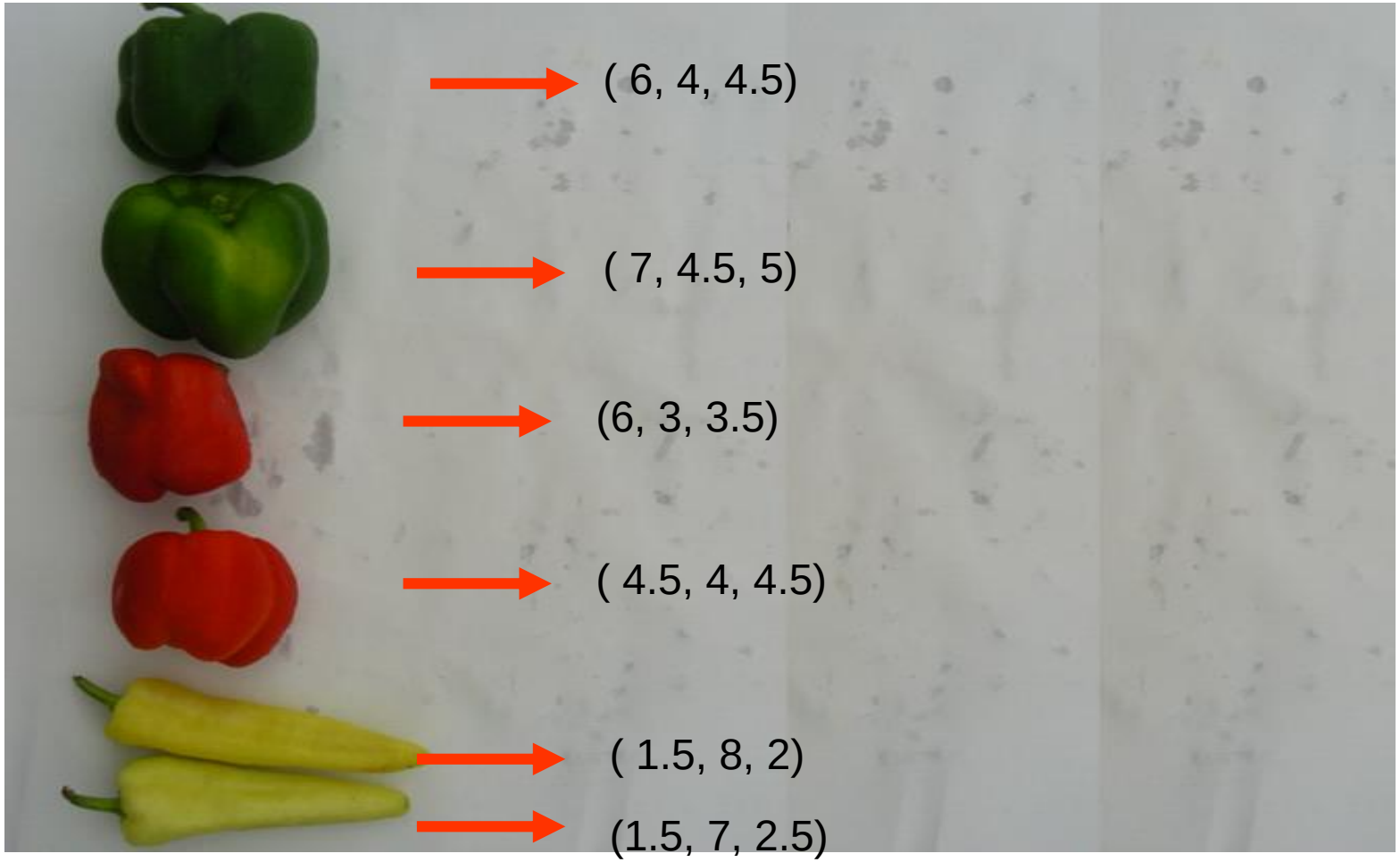
Consider data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$X_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X} \subset \mathbb{R}^d$$

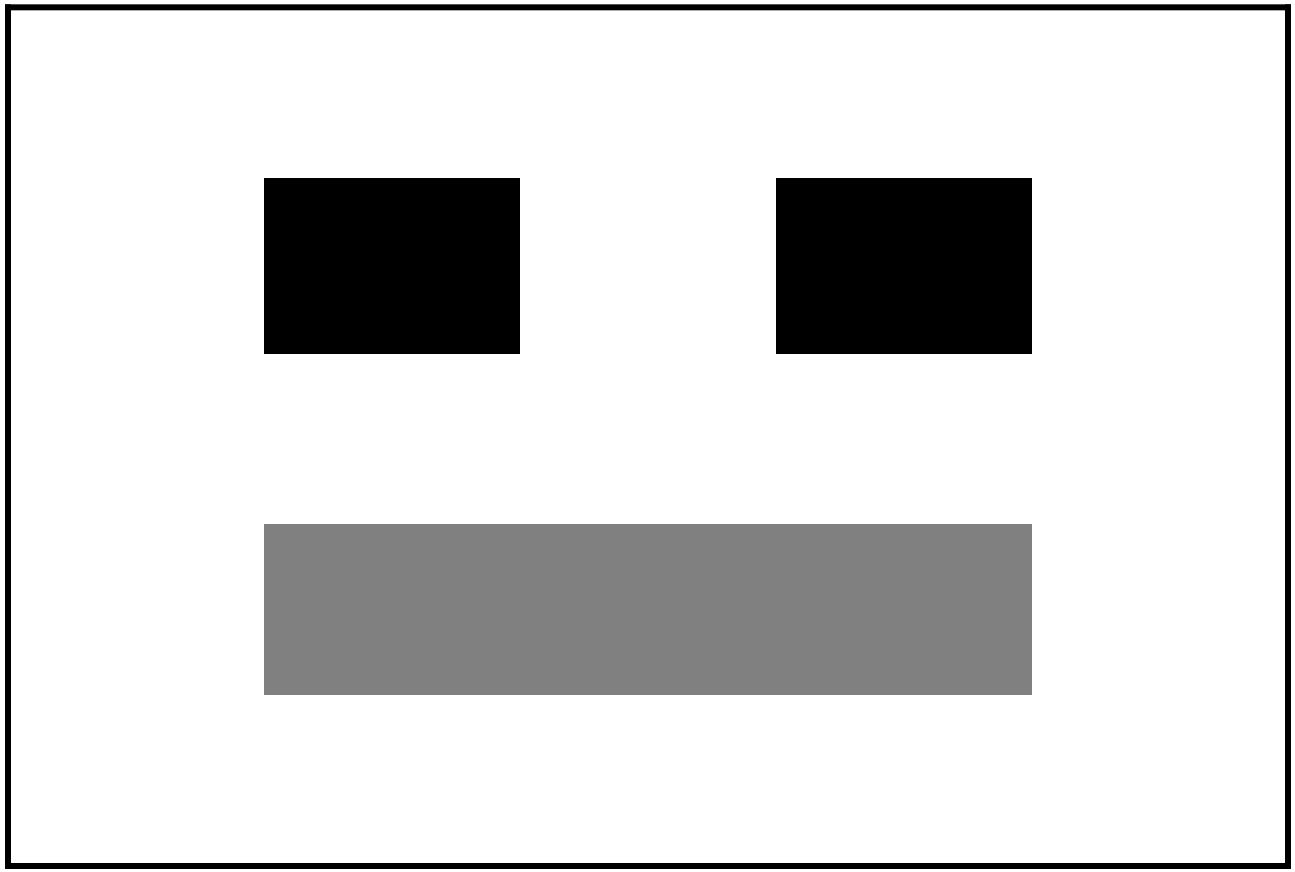
is a d-dimensional vector and Y_i takes values in some finite set $\mathcal{Y}$. A **classification rule** is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new X, we predict Y to be $h(X)$.

Data

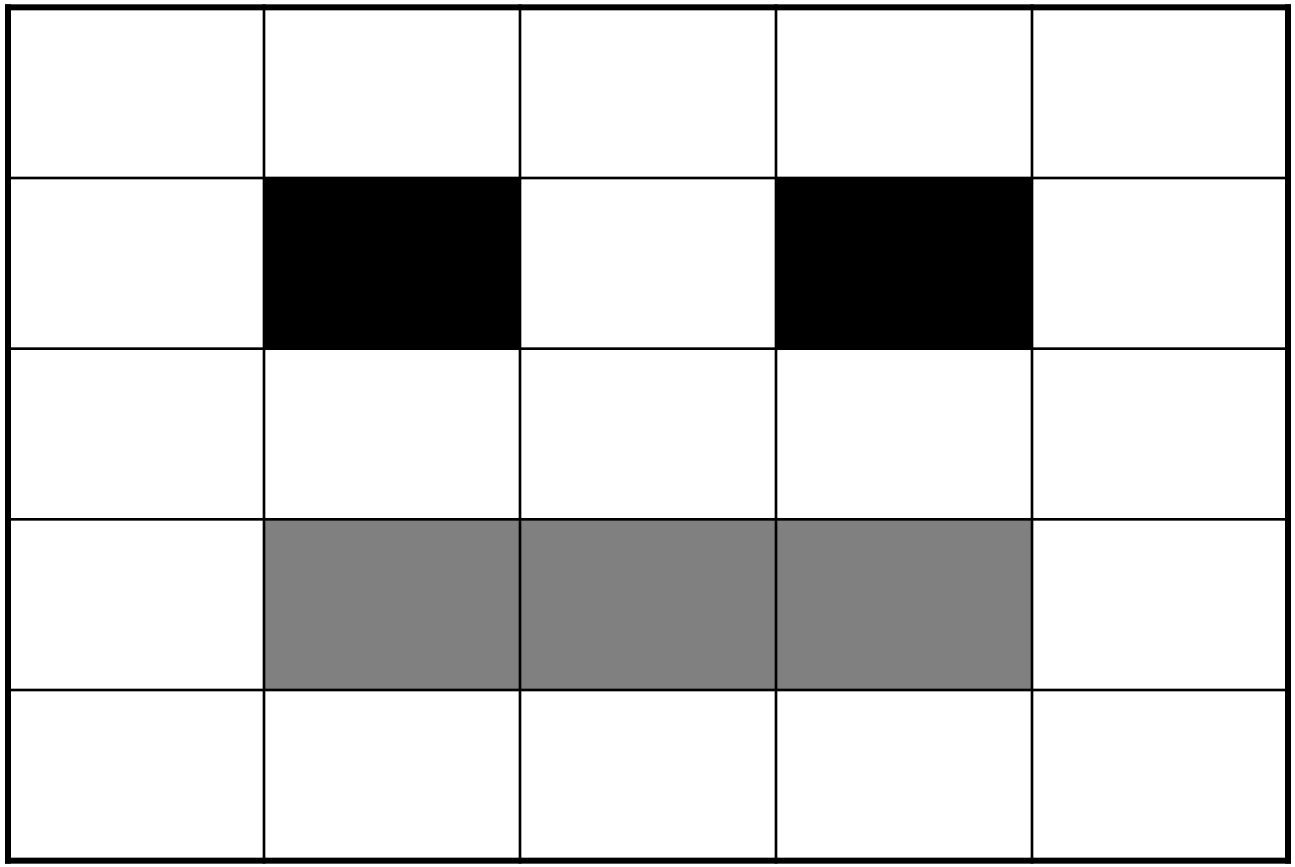
Features (X)



Data Representation



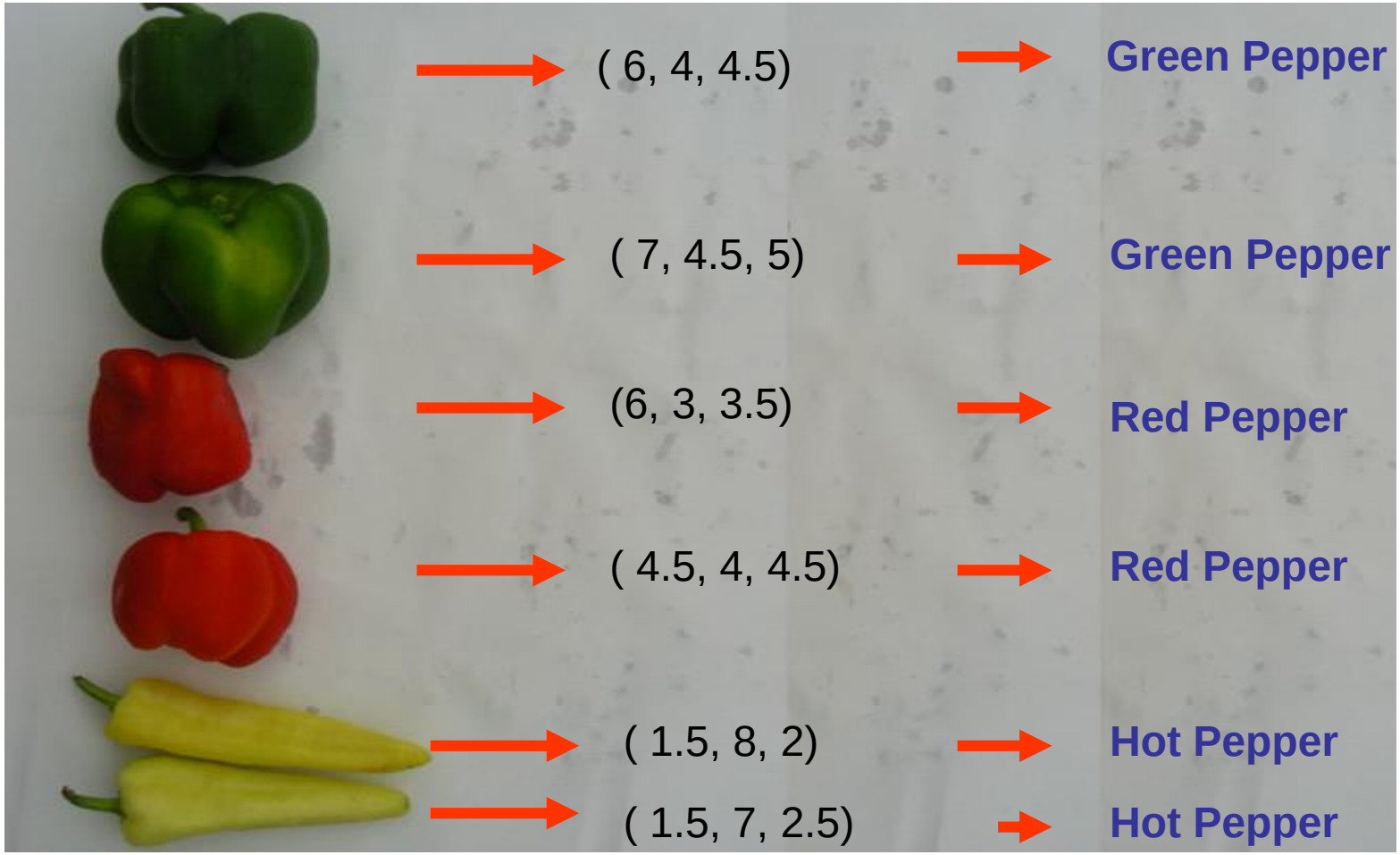
Data Representation



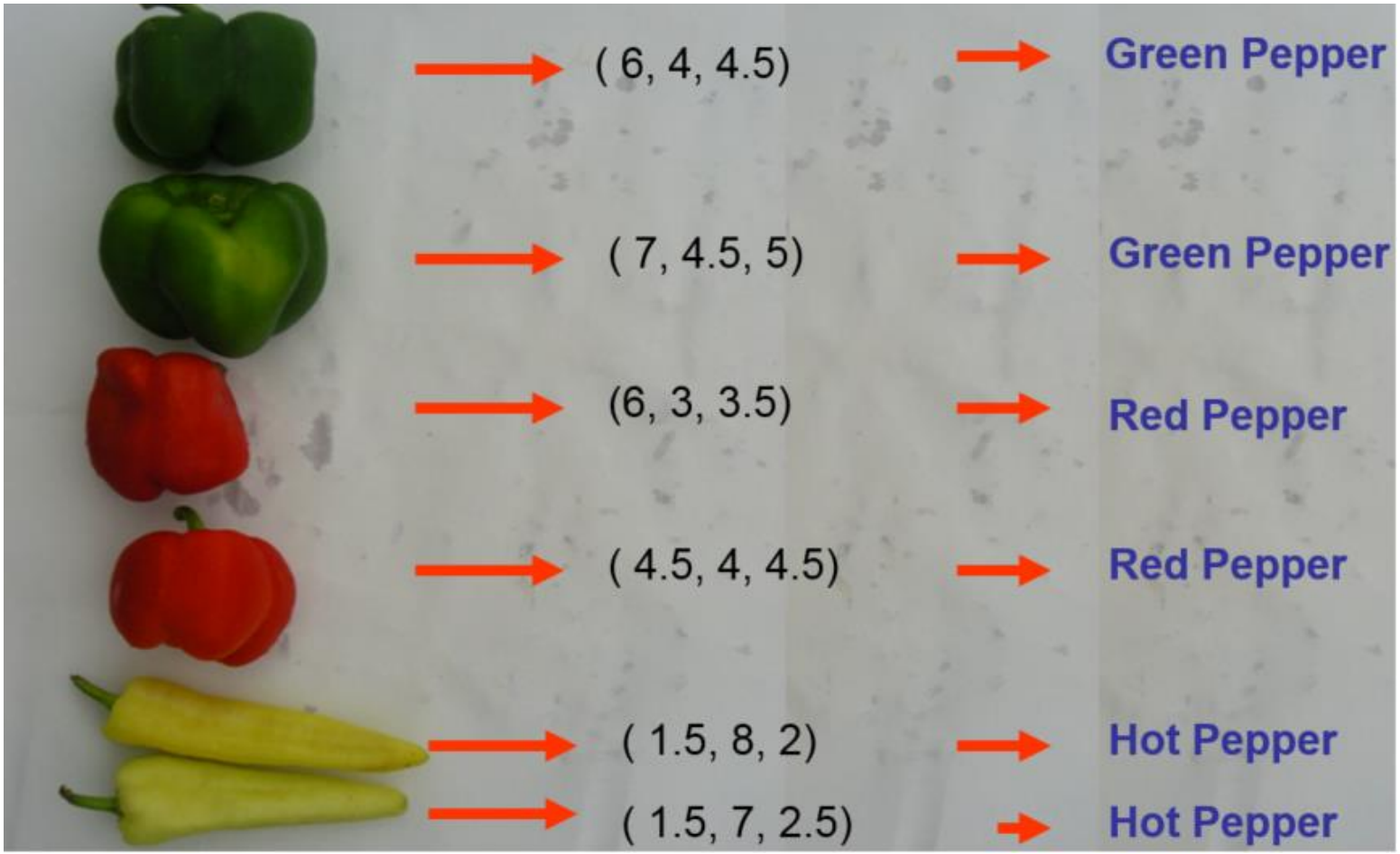
Data Representation

1	1	1	1	1
1	0	1	0	1
1	1	1	1	1
1	0.5	0.5	0.5	1
1	1	1	1	1

Features and labels



Features and labels



Classification (New point)

(7, 4, 4.5)

h(7, 4, 4.5)

?

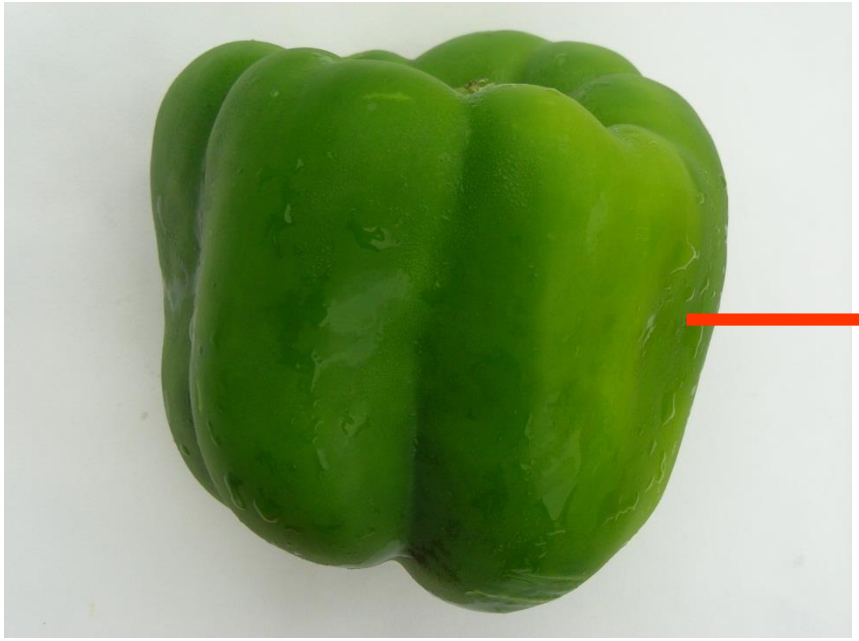
Classification (New point)

(5, 3, 4.5)

h(5, 3, 4.5)

?

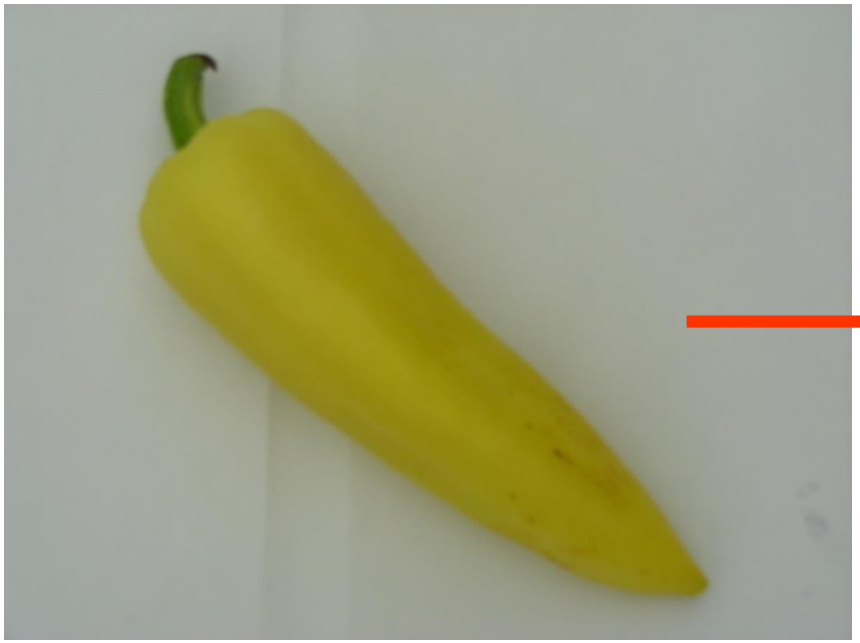
Classification (New point)



(6, 4, 4.5)

h(6, 4, 4.5) → ?

Classification (New point)



(2, 10, 1.2)

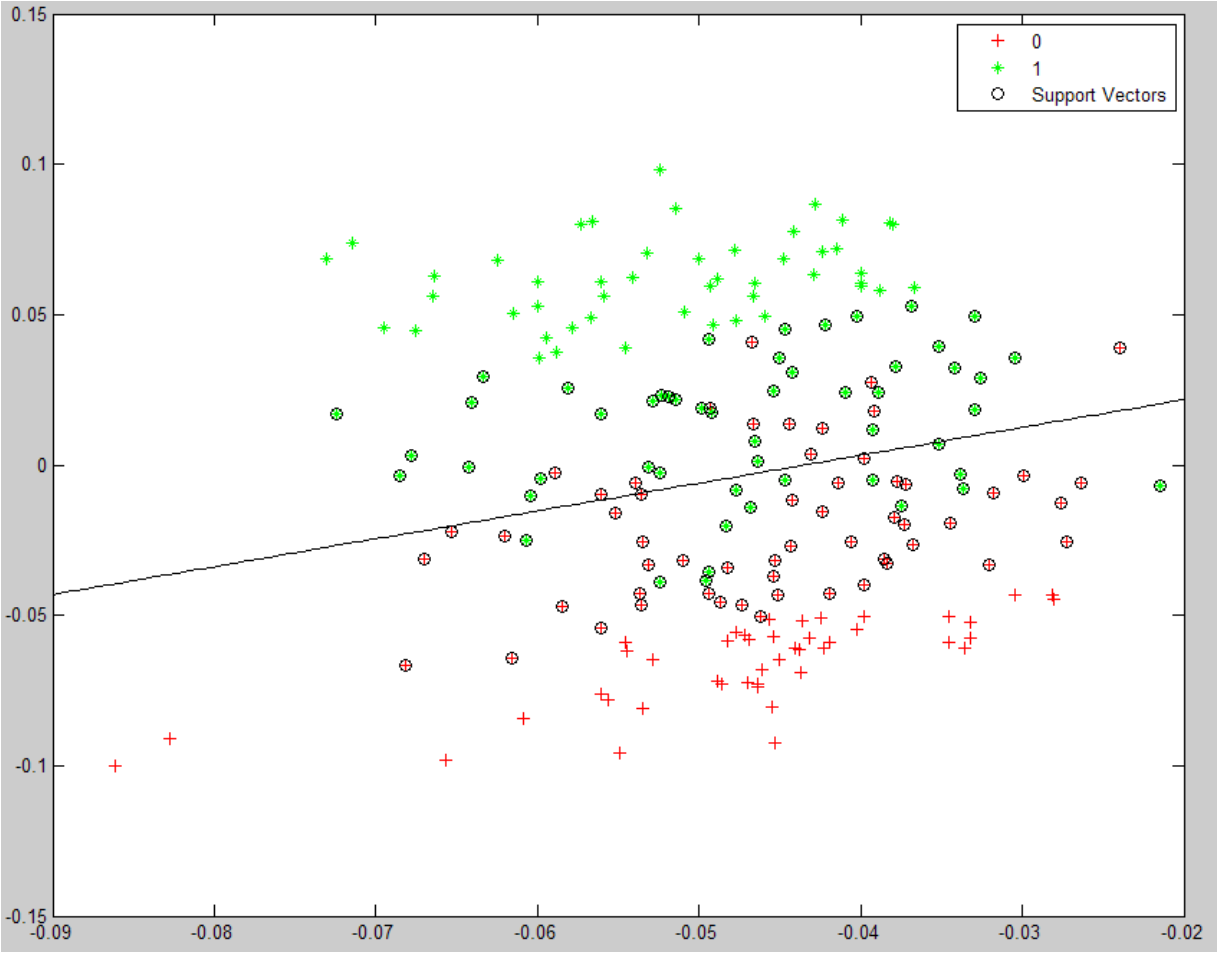
h(2, 10, 1.2)

?

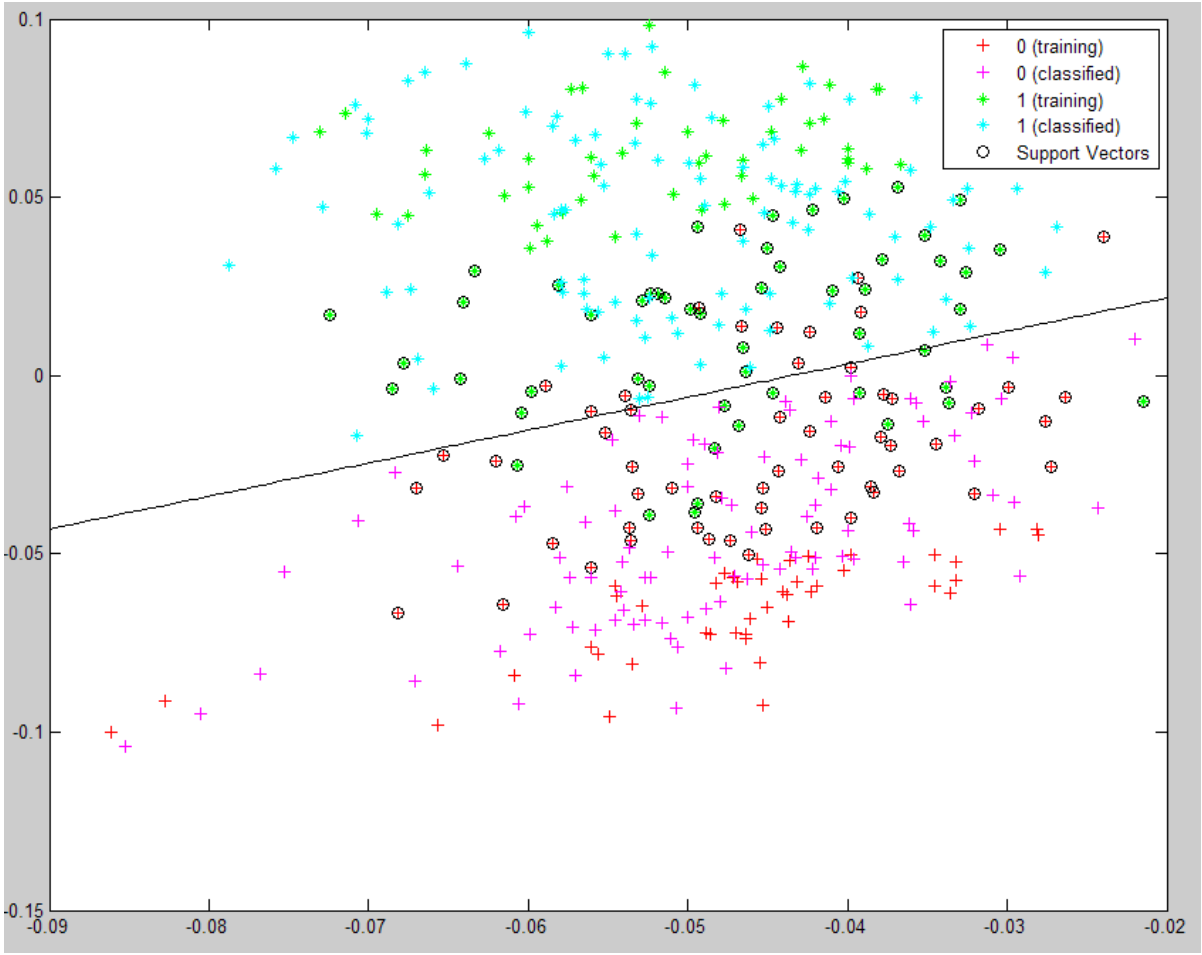
Face Identification

Face Identification

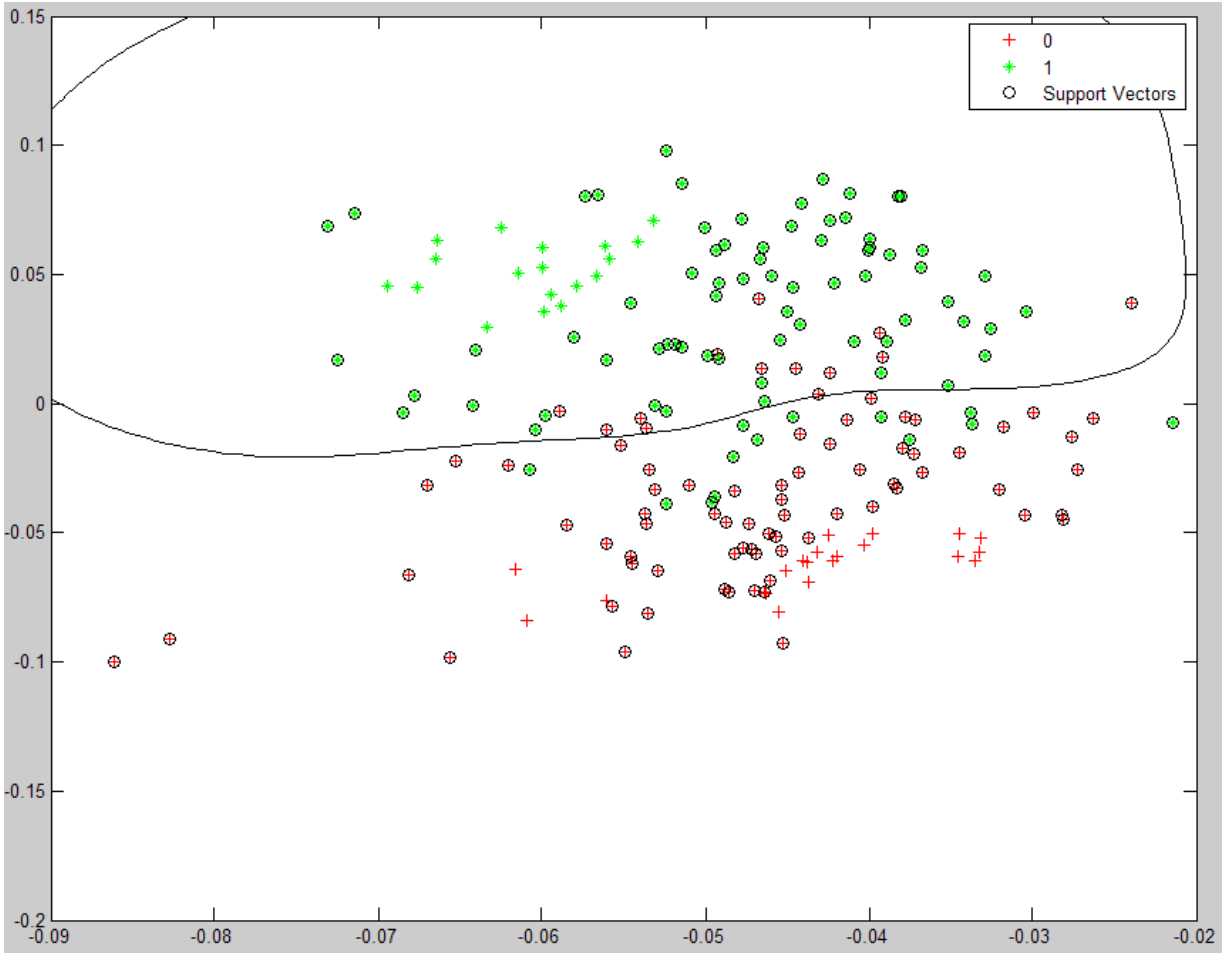
Classification



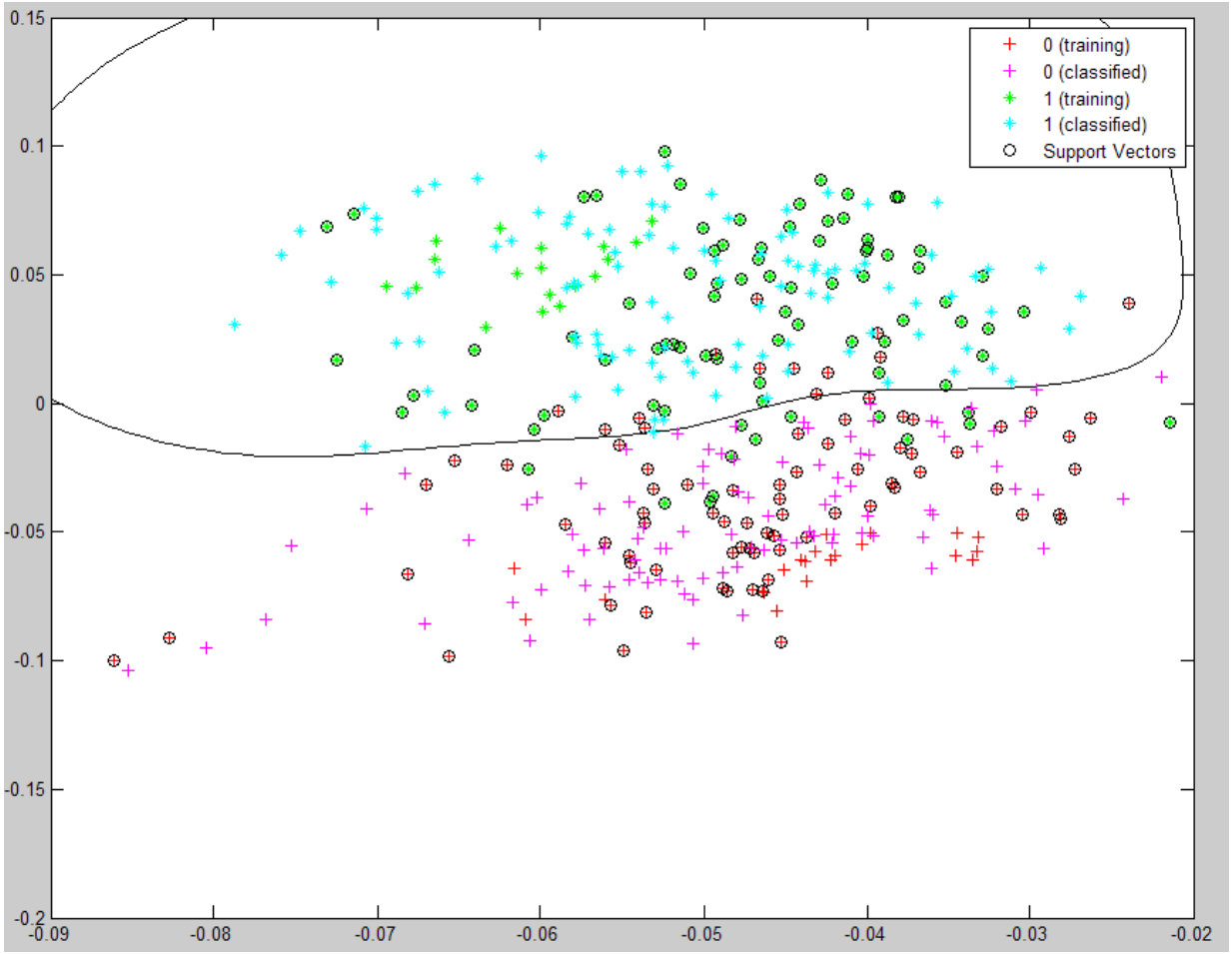
Classification



Classification



Classification



Digit Recognition

The Matlab data file 2_3.mat contains 200 handwritten 2's and 3's images scanned from postal envelopes, like the ones shown below.

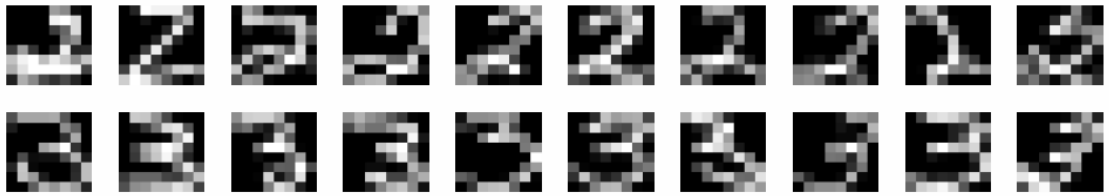
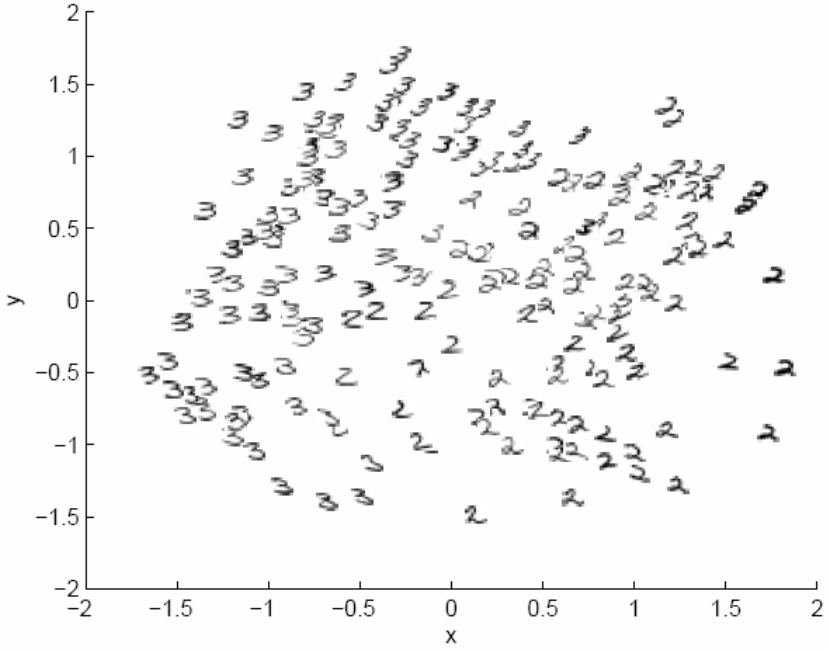
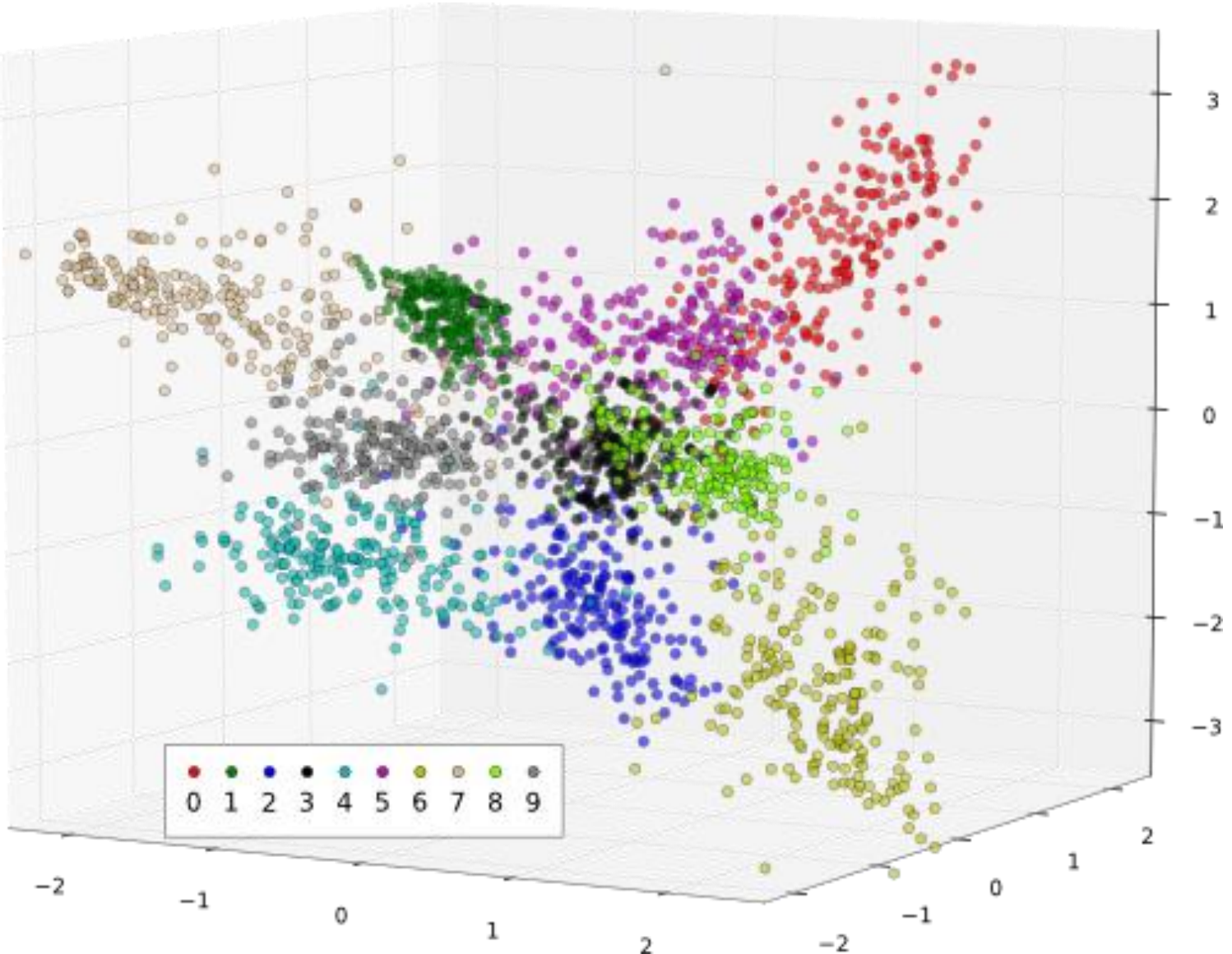


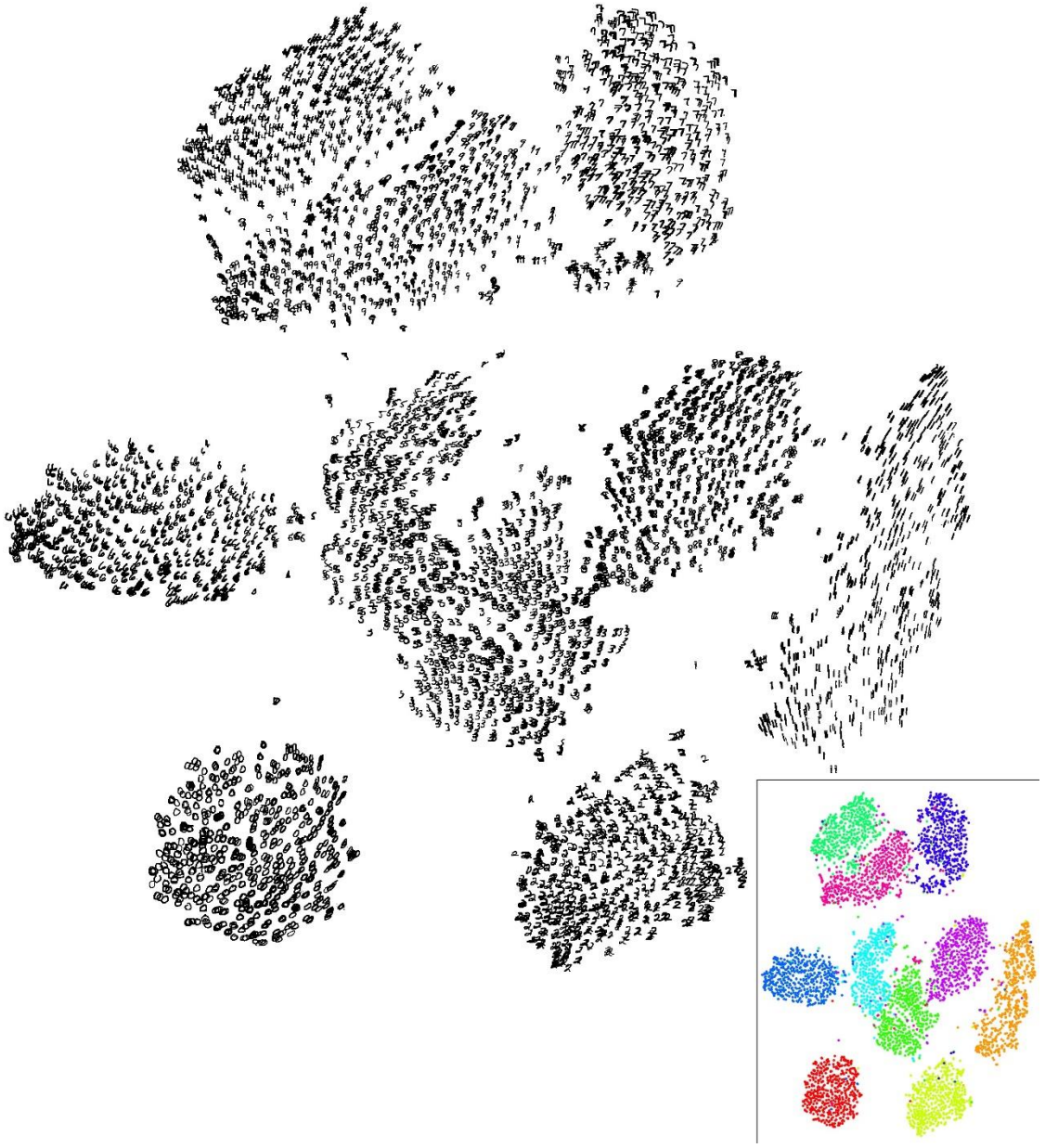
Figure 1:

These images are stored as a 64×400 matrix. Each column of the matrix is an 8×8 greyscale image (the pixel intensities are between 0 and 1).

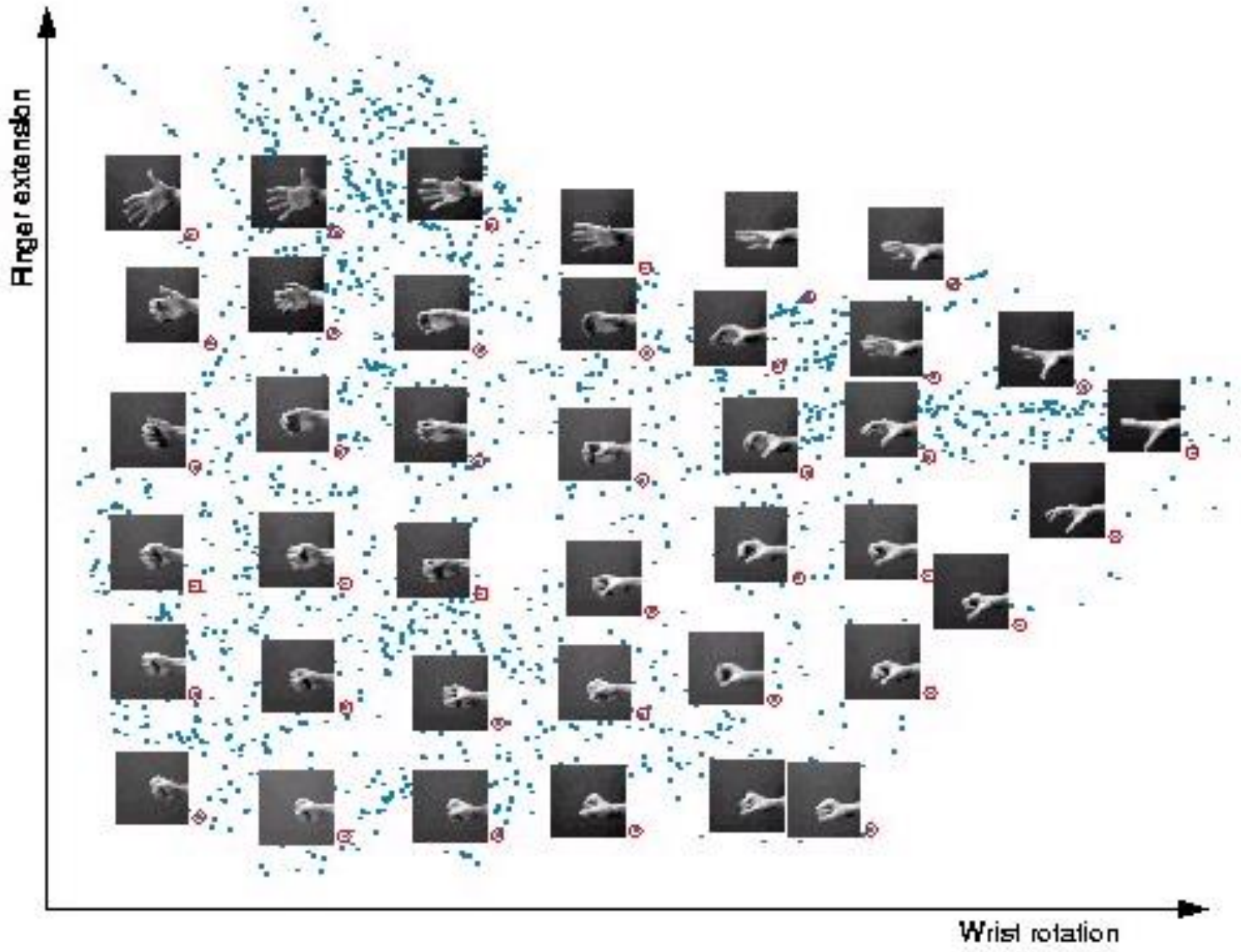


A canonical dimensionality reduction problem from visual perception. The input consists of a sequence of 64-dimensional vectors, representing the brightness values of 8 pixel by 8 pixel images of digits 2 and 3. Applied to $n = 400$ *raw images. A two-dimensional projection is shown, with the original input images.*

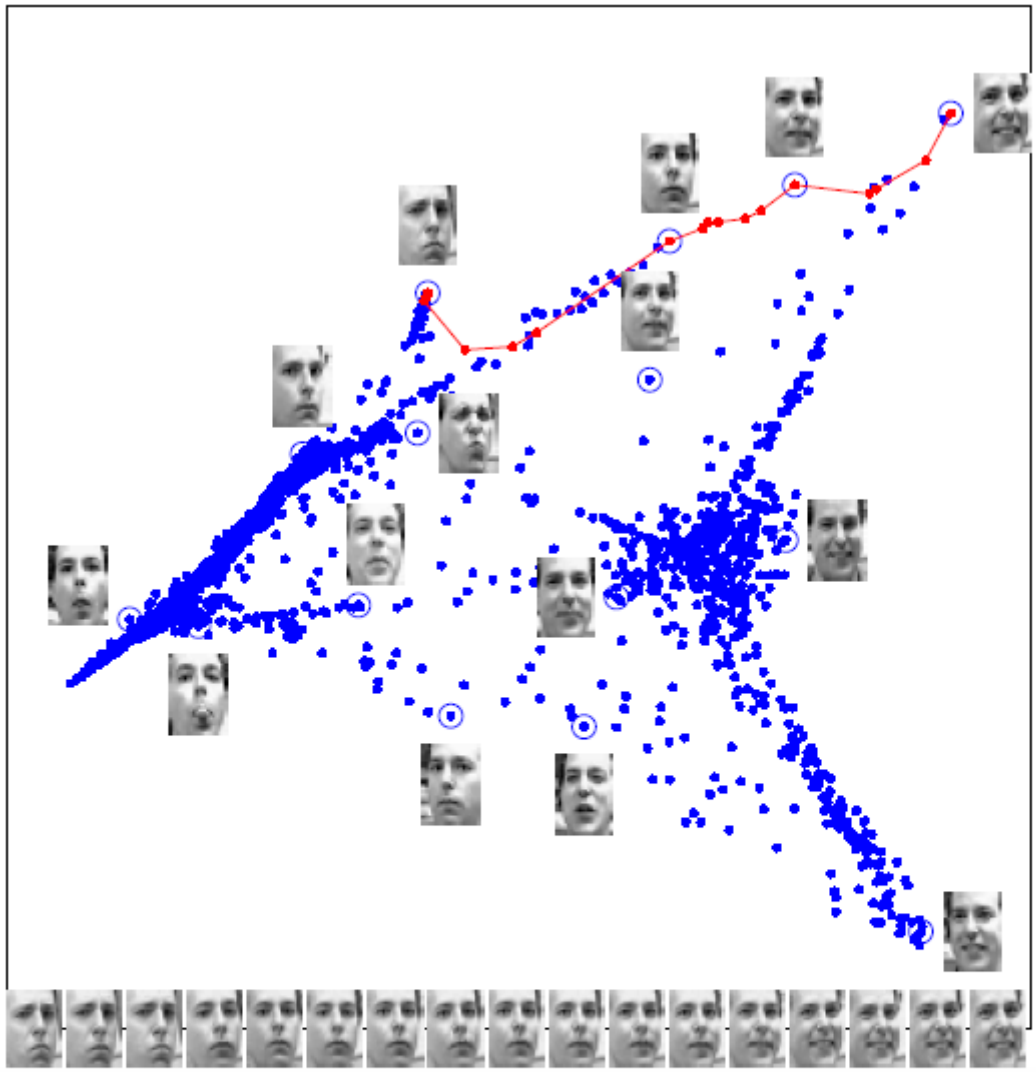

0 1 2 3 4 5 6 7 8 9
−2
−1
0
1
2
3
2
1
0
−1
−2
−3
2
1
0
−1
−2



t-SNE: most images of faces were clustered in the bottom. Most images of airplanes were clustered on the right.



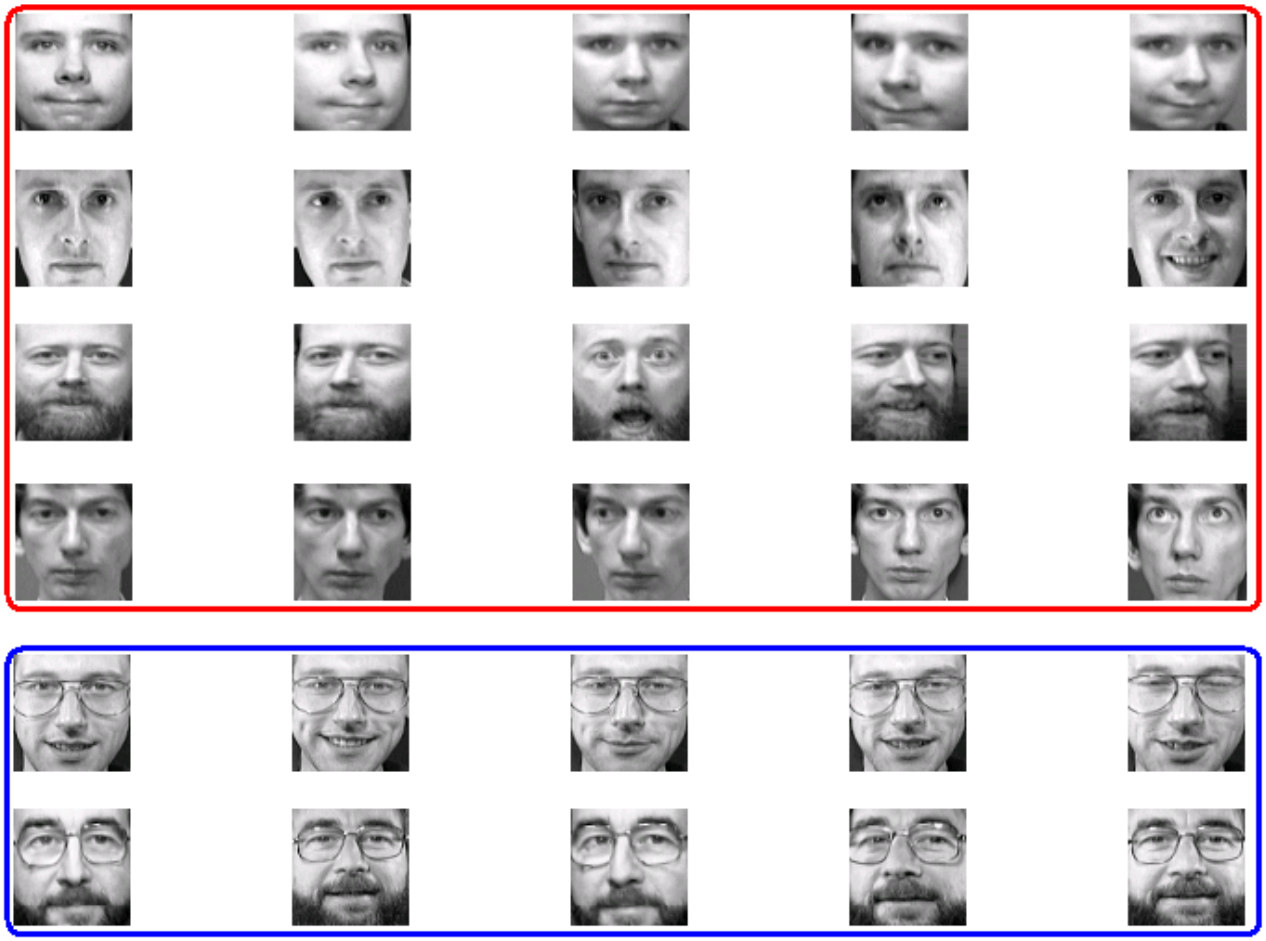
Example from (Tenenbaum 2000)



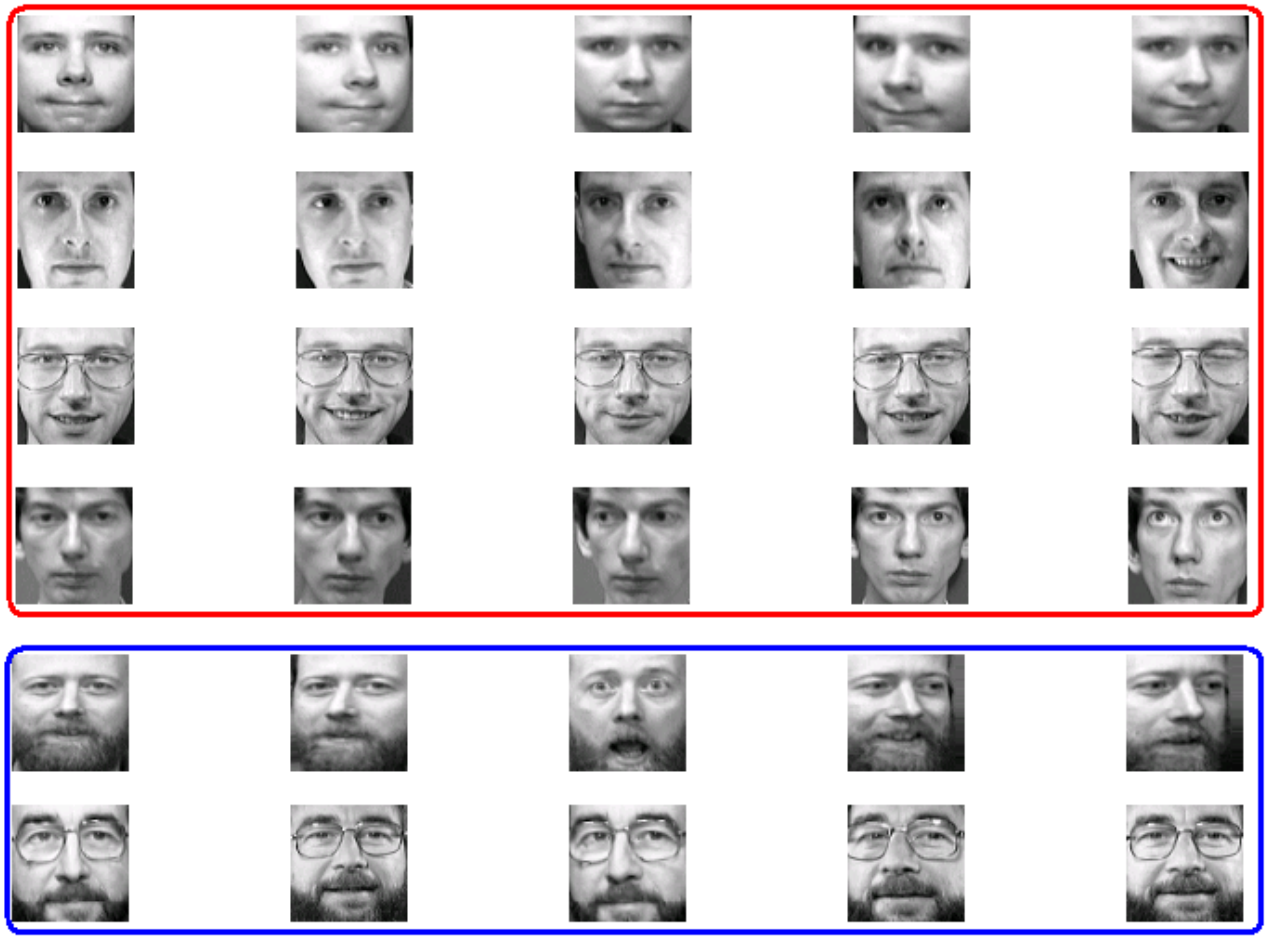
Images of faces mapped into the embedding space described by the first two coordinates of LLE, using $K = 12$ nearest neighbors. Representative faces are shown next to circled points at different points of the space. The bottom images correspond to points along the top-right path, illustrating one particular mode of variability in pose and expression. The data set had a total of $N = 1965$ grayscale images at 20×28 resolution ($D = 560$).

Different Features

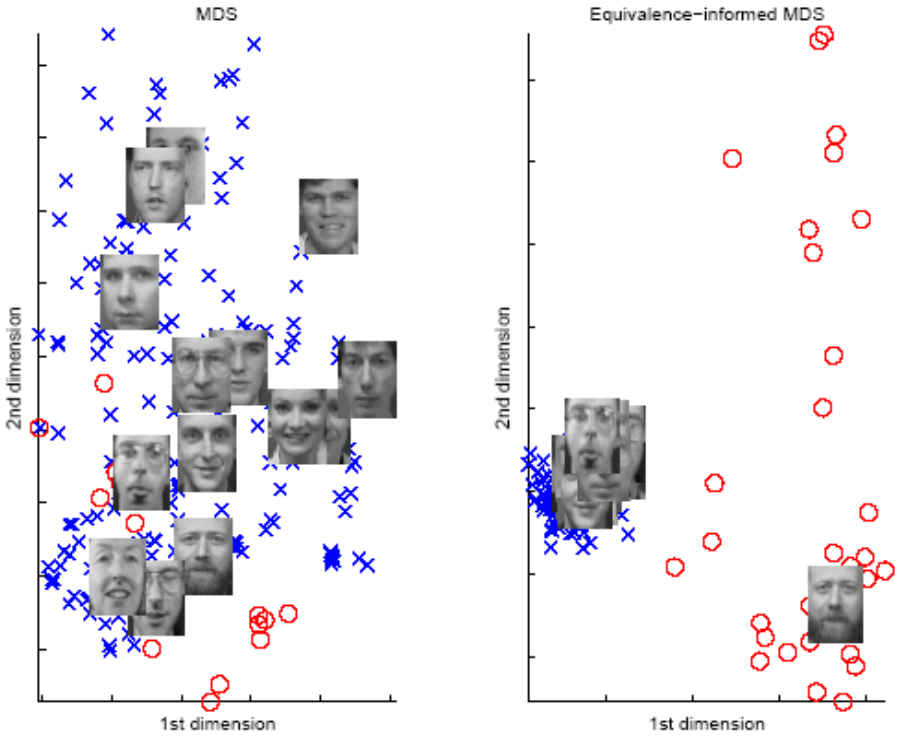
Glasses vs. No Glasses



Beard vs. No Beard

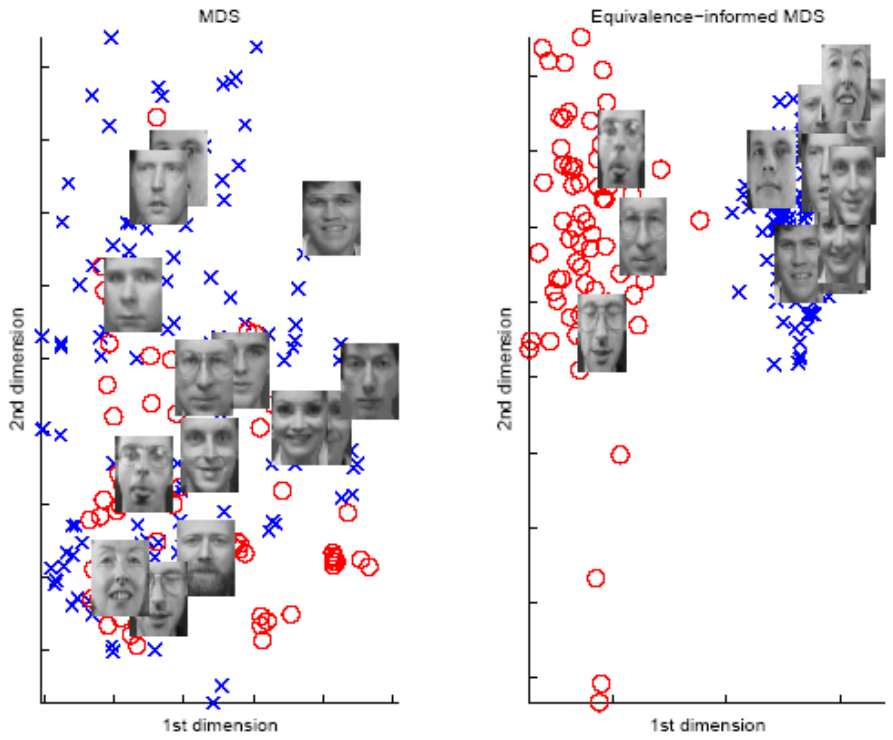


Beard Distinction



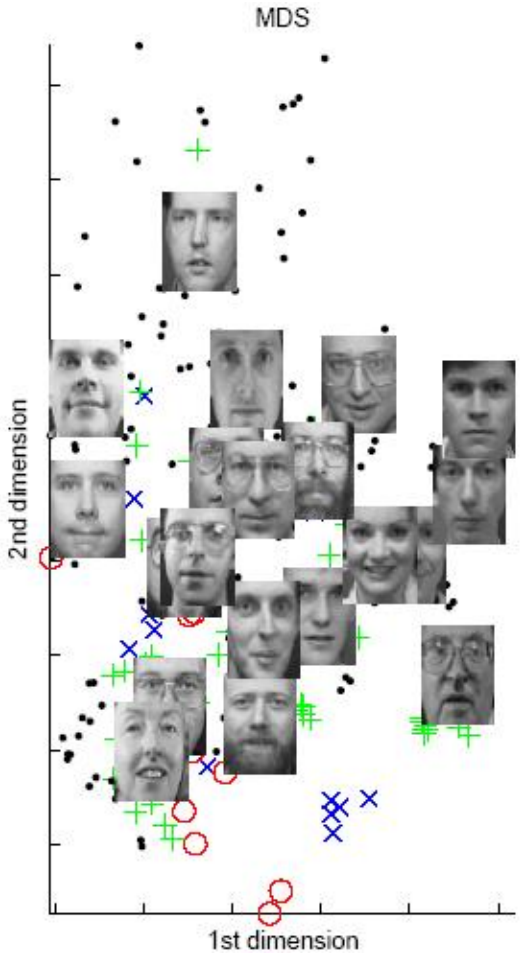
MDS and Equivalence-Informed MDS with bearded/unbearded distinction (all class-equivalent pairs)

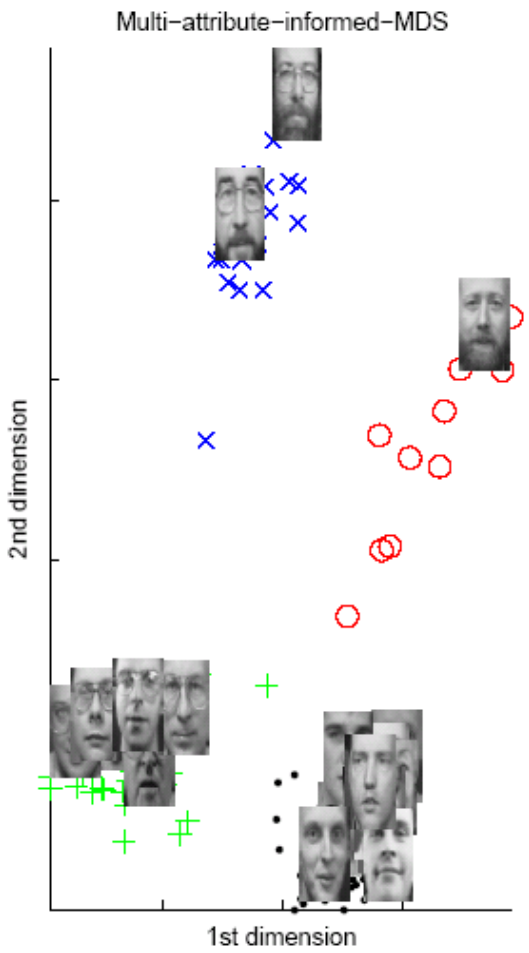
Glasses Distinction



MDS and Equivalence-Informed MDS with glasses/no glasses distinction (all class-equivalent pairs)

Multiple-Attribute Metric





Textbook

- There is no required textbook for the class. Some classic papers will be assigned as readings.

- Three recommended books that cover the similar material are:

Hastie, Tibshirani, Friedman

Elements of Statistical Learning.

Bishop

Pattern Recognition and Machine Learning.

Murphy

Machine Learning: a Probabilistic Perspective

Course Evaluation (tentative)

- Assignment 50% ,
- Group Project 50%

Project

- Final group project (presentation and reports up to 7 pages of PDF) are worth 50% of your final grade .

- Right Whale Recognition

kaggle competition as your final project.

Project

- We will find out if teaching environment computational resources are adequate.

- If it turns out that you don't have access to adequate computational resources, you may chose other possible types of projects as follows:

- Another active kaggle completion.

The basic types of projects

- Develop a new algorithm. In this case, you will need to demonstrate (theoretically and/or empirically) why your technique is better (or worse) than other algorithms.

Note: A negative result does not lose marks, as long as you followed proper theoretical

- and/or experimental techniques.

- Application of classification to some domain. This could either be your own research problem, or you could try reproducing results of someone else's paper.

- The project is a chance to learn more about some sub-area of classification that you might be most interested in.

- You may benefit more from implementing an algorithm and doing some simulations rather than trying to read and summarize some state-of-the-art papers.

- Final project reports will be checked by Turnitin (Plagiarism detection software).

Communication

- All communication should take place using the *Piazza* discussion board.

- Piazza is a good way to discuss and ask questions about the course materials, including assignments, in a public forum

Communication

- It enables you to learn from the questions of others, and to avoid asking questions that have already been asked and answered.

- Students are expected to read Piazza on a regular basis.

Enrolling in Piazza

- You will be sent an invitation to your UW email address. It will include a link to a web page where you may complete the enrollment process.

Piazza Guidelines

- In any posts you make, do not give away any details on how to do any of the assignments. This could be construed as cheating, and you will be responsible as the poster.

Course Website

- We will mostly use Piazza for communication.

- Assignments and grades will be handled through Learn.
- Please log on frequently. You are responsible for being aware of all material, information and email messages found on Piazza and Learn..

Reading

- Journals: Neural Computation, JMLR, ML, IEEE PAMI
- Conferences: NIPS, UAI, ICML, AI-STATS, IJCAI, IJCNN
- Vision: CVPR, ECCV, SIGGRAPH
- Speech: EuroSpeech, ICSLP, ICASSP
- Online: citesser, google
- Books:
  - Elements of Statistical Learning, Hastie, Tibshirani, Friedman
  - Pattern Recognition and Machine Learning, Bishop
  - Pattern Classification, Duda, Hart, Strok
  - Machine Learning an Algorithmic Perspective, Marsland

Important Dates

- Oct 6: Final project proposal due (Use the link posted on Learan)

- Nov 17: Presentation begin (tentative)

Prerequisite

- Grads: none for STATS/CS/ECE/SYDE grad students , instructor permission otherwise
- Undergrads: CM 361/STAT 341 or (STAT 330 and 340)

Tentative topics

- Feature extraction
- Error rates and the Bayes classifier
- Gaussian and linear classifier
- Linear regression and logistic regression
- Neural networks
- Deep Learning
- Radial basis function networks
- Density estimation and Naive Bayes
- Trees

Tentative topics

- Assessing error rates and model selection
- Support vector machines
- Kernel methods
- k-nearest neighbors
- Bagging
- Boosting
- Semi-supervised learning for classification
- Metric learning for classification