San José State University

Math 250: Mathematical Data Visualization

Data sets and their visualization in 3D

Dr. Guangliang Chen

Outline:

- Various kinds of matrix-type data
- Selected benchmark data sets
- Loading data into MATLAB
- Plotting and visualizing data
- In-class demonstrations

Data types

Data exists (or is collected) in various forms, such as

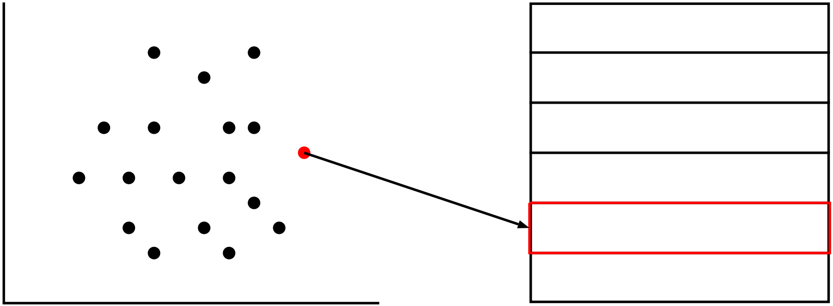
- Numerical / categorical vectors
- Images (gray-scale, color)
- Text documents
- Graphs (networks)
- Videos
- Hyperspectral images

Storing data as matrices

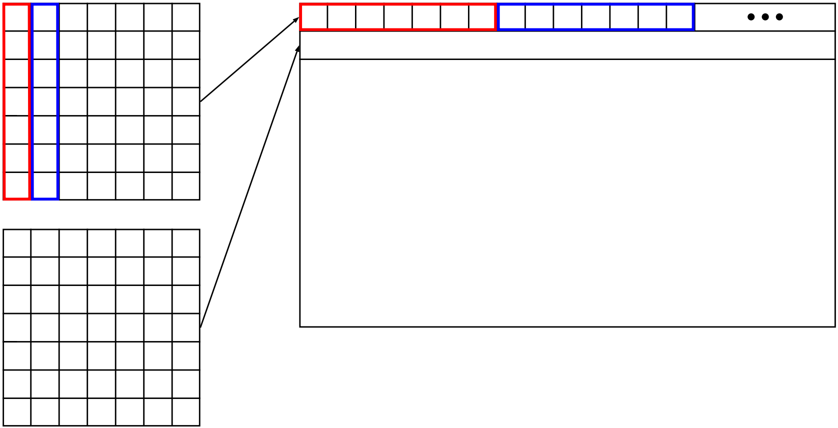
The following data objects can all be conveniently represented as matrices:

- Vectors in Euclidean spaces
- Digital images and their collections
- Text corpus (collections of text documents)
- Graph/network data

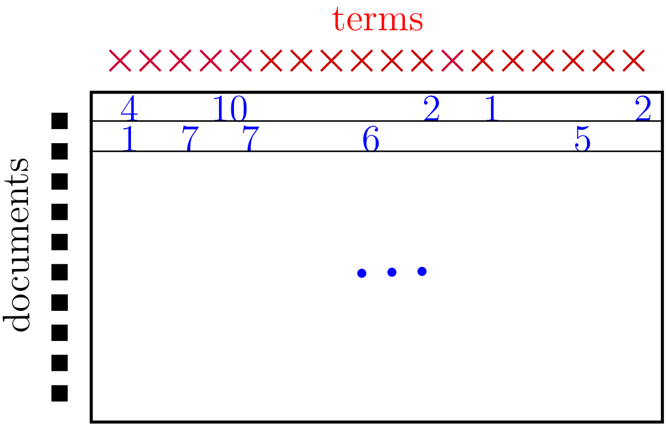
Data points in Euclidean spaces as matrices



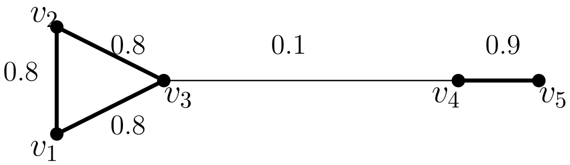
Digital images as matrices



Collections of text documents as matrices



Networks (graphs) as matrices

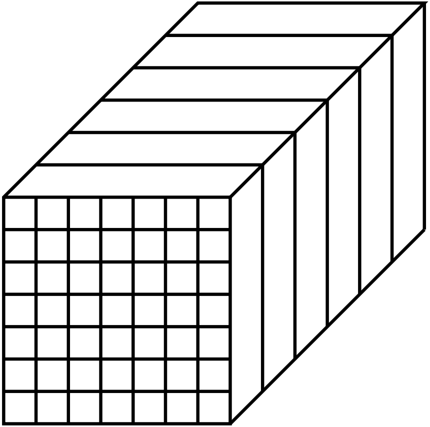


$$\mathbf{W} = \begin{pmatrix} & .8 & .8 & & \\ .8 & & .8 & & \\ .8 & .8 & & .1 & \\ & & .1 & & .9 \\ & & & .9 & \end{pmatrix}$$

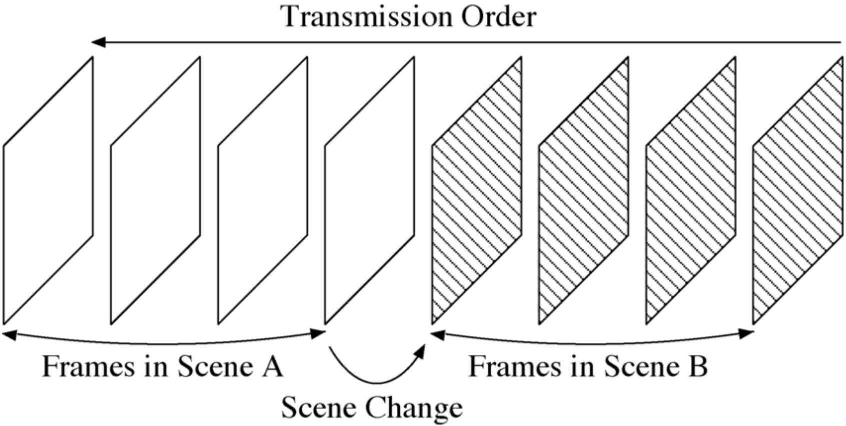
Storing data as tensors (3D arrays)

- Collection of images of the same size
- Videos
- Hyperspectral images

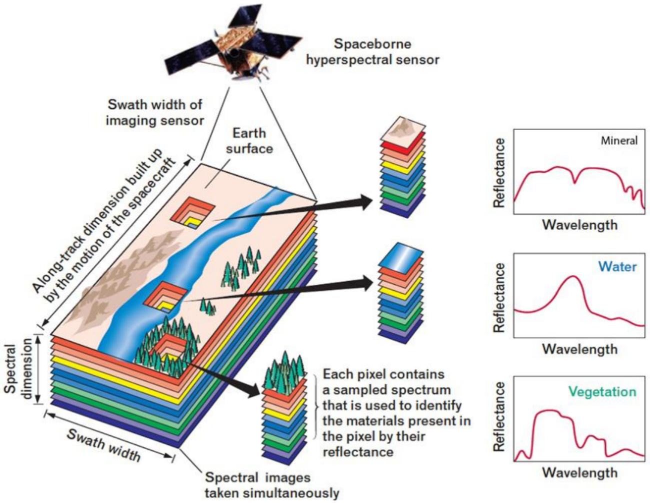
Data sets and their visualization in 3D



Data sets and their visualization in 3D



Data sets and their visualization in 3D


Spaceborne hyperspectral sensor
Swath width of imaging sensor
Earth surface
Along-track dimension built up by the motion of the spacecraft
Spectral dimension
Swath width
Spectral images taken simultaneously
Each pixel contains a sampled spectrum that is used to identify the materials present in the pixel by their reflectance
Mineral
Reflectance
Wavelength
Water
Reflectance
Wavelength
Vegetation
Reflectance
Wavelength

Data sets to be used in this course

- **Image collections**: *MNIST handwritten digits**, *Fashion MNIST**, *USPS handwritten digits**
- **Text corpus***: *20 newsgroups**
- **UCI Machine Learning Repository**[1]: smaller data sets such as *iris*, and *wine quality*

*Available in Canvas.

Let me know if you have a good data set for visualization in mind!

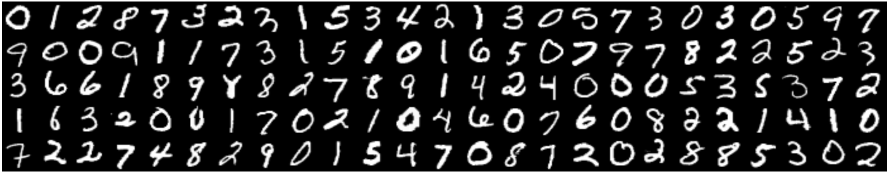
[1] http://archive.ics.uci.edu/ml/

MNIST Handwritten Digits

`http://yann.lecun.com/exdb/mnist/`

70,000 digital images of size 28x28 of handwritten digits 0. . . 9 collected from about 250 people

A benchmark data set used for machine learning classification

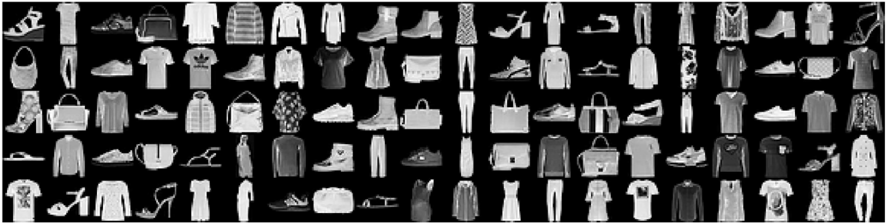


Fashion-MNIST

`https://github.com/zalandoresearch/fashion-mnist`

Same size and format with MNIST, but the contents are clothes instead

The data set is harder than MNIST.

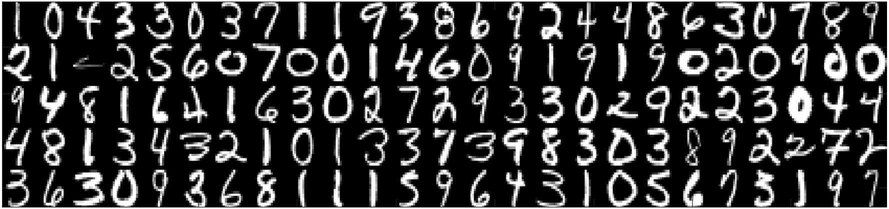


USPS Zip Code Data

`http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`

9,298 size 16×16 grayscale images of handwritten digits scanned from envelops

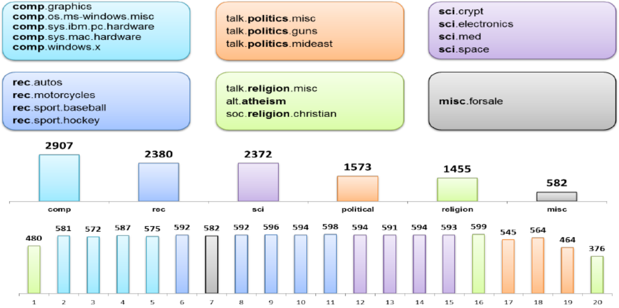
Smaller than MNIST but more noisy



20 Newsgroups Data

`http://qwone.com/~jason/20Newsgroups/`

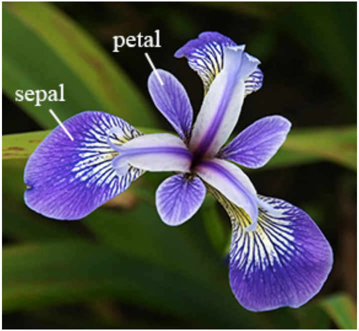
18,824 text documents, divided into 20 news groups with different topics



UCI Machine Learning Repository - iris

`https://archive.ics.uci.edu/ml/datasets/iris`

- **150 instances**
- **4 numerical attributes**
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
- **1 categorical**: Iris type



UCI Machine Learning Repository - wine quality

`https://archive.ics.uci.edu/ml/datasets/wine+quality`

- **4,898 instances** (two datasets are combined, related to red and white vinho verde wine samples from north of Portugal)
- **11 numerical attributes**
- **1 response variable: quality** (score between 0 and 10)

Data plotting

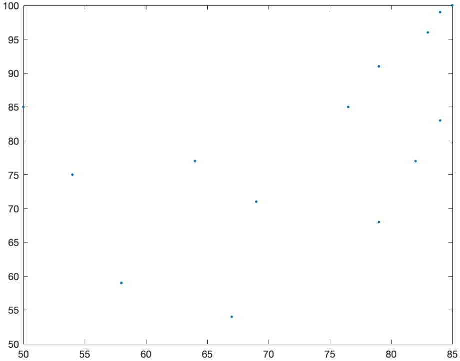
Example: *A scatterplot of Math 250 (spring 2021) test scores*

```
% midterm 1 (out of 90)
x = [85 83 84 84 79 76.5 82 79 64 69 50 54 58 67];

% midterm 2 (out of 105)
y = [100 96 99 83 91 85 77 68 77 71 85 75 59 54];

figure; plot(x,y, ’.’)
```

A low quality plot



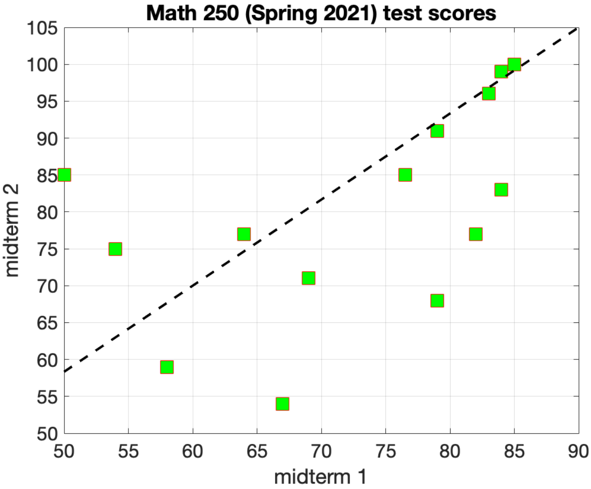
Things to keep in mind when plotting data

- Symbol (marker) type and size
- Font sizes (title and axis labels)
- Color contrast
- Line styles
- Legend
- Aesthetics

How to plot the data (elegantly)

```
figure;
plot(x, y, 'rs', 'MarkerSize', 14, ...
'MarkerEdgeColor','r', 'MarkerFaceColor', 'g')
hold on
plot([50 90],[50*105/90 105], 'k-', 'linewidth', 2)
xlabel('midterm 1', 'fontsize', 16)
ylabel('midterm 2', 'fontsize', 16)
xlim([50 90]); ylim([50 105])
set(gca, 'fontsize', 16)
grid on
title('Math 250 (Spring 2021) test scores', 'fontsize', 18)
```

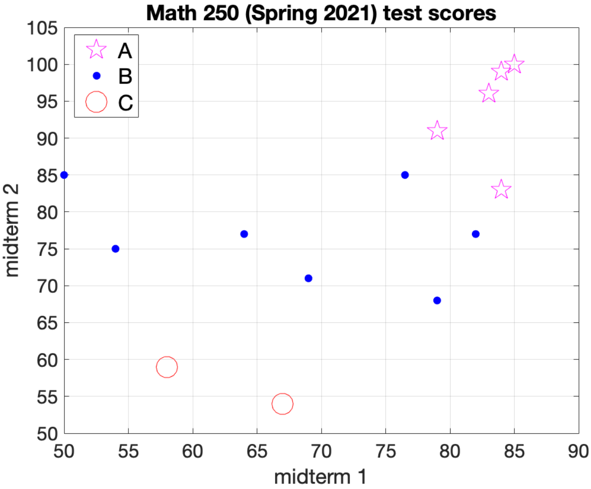
Data sets and their visualization in 3D



Data sets and their visualization in 3D

```
grades={'A','A','A','A','A','B','B','B','B','B','B','B','C','C'};
grades=categorical(grades);
figure;
gscatter(x, y, grades, 'mbr', 'p.o', 18)
legend(categories(grades), 'fontsize', 18)
box on
xlabel('midterm 1', 'fontsize', 16)
ylabel('midterm 2', 'fontsize', 16)
xlim([50 90]); ylim([50 105])
set(gca, 'fontsize', 16)
grid on
title('Math 250 (Spring 2021) test scores', 'fontsize', 18)
```

Data sets and their visualization in 3D



What to look for in a plot

- range of each dimension
- general pattern and trend
- center, peaks, symmetry, etc
- clusters (if any)
- outliers (peculiar points)

Data visualization

Goals: For each data set, we will focus on both of the following tasks:

- data plotting (with publication quality)
- data exploration (for getting insights)

Strategy: We will examine the variables in the following ways:

- Single variable:
  - Numerical: 1-D scatterplot, histogram, boxplot, bar graph (if frequency data)
  - Categorical: bar graph, pie chart

Data sets and their visualization in 3D

- Two variables:
  - Both numerical: 2-D scatterplot
  - Both categorical: stacked bar plot
  - Mixed: side-by-side boxplot
- Three variables:
  - All numerical: 3-D scatterplot, scatterplot matrix
  - Two numerical and 1 categorical: 2-D scatterplot with groups
  - One numerical and two categorical: heatmap, 3D bar plot

In-class demonstrations

See scripts from instructor in class.

The case of high dimensional data

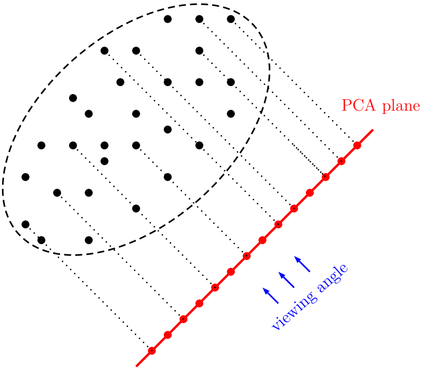
High dimensional data sets are hard to visualize due to physical limitations.

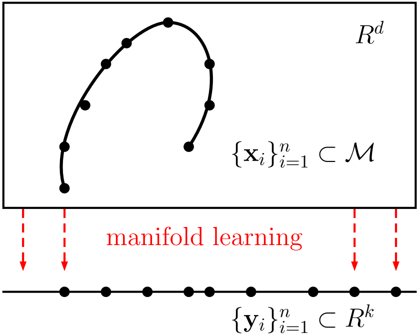
The best we can do is to find a proper angle to peek into the data in order to understand its structure that is relevant to the given task.

Later in this course, we will cover the following methods:

- **Linear projection methods**: PCA, LDA
- **Nonlinear embedding methods**: MDS, ISOMap, LLE, Laplacian Eigenmaps

Data sets and their visualization in 3D




R^d
$\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{M}$
manifold learning
$\{\mathbf{y}_i\}_{i=1}^n \subset R^k$