Lecture 1: Data, Lists and Models

Information is described by **data**. While data entries are often not numerical values initially, they can always be encoded in a numerical manner so that we can look at **numerical data**. Here are some examples: stock data, weather data, networks, vocabularies, address books, encyclopedias, surfaces, books, the human genome, pictures, movies or music pieces. Is there any information which is not described by data?

While data describe complicated objects, they can be organized in **lists**, lists of lists, or lists of lists of lists etc. Stock market or weather data can be represented as finite sequences of numbers, pictures are arrays of numbers, a movie is an array of pictures. A book is a list of letters. A graph is given by a list of nodes and connections between these nodes. A movie is a list of pictures and each picture is a list of list of pixels and each pixel is a list of red, green and blue values. If you think about these examples, you realized that **vectors**, **matrices** ore higher dimensional arrays are helpful to organize the data. We will define these concepts later but a vector is just a finite list of things and a matrix is a list of lists, a **spreadsheet**. In the case of pictures, or music pieces the translation into an array is obvious. Is it always possible to encode data as sequences of numbers or arrays of numbers or lists of arrays of numbers etc? Even for complicated objects like networks, one can use lists. A network can be encoded with an array of numbers, where we put a 1 at the node (i, j) if node i and j are connected and 0 otherwise. This example makes clear that we need a **mathematical language** to describe data. It looks like a difficult problem at first because data can appear in such different shapes and forms. It turns out that we can use vectors to describe data and use matrices to describe relations between these data. The fact that most popular databases are **relational databases** organized with tables vindicates this point of view. In this first lecture, we also want to see that linear algebra is a tool to organize and work with data. Even data manipulation can be described using linear algebra. Data of the same type can be added, scaled. We can mix for example two pictures to get a new picture. Already on a fundamental level, nature takes linear algebra seriously: while classical mechanics deals with differential equations which are in general nonlinear and complicated, quantum mechanics replaces this with a linear evolution of functions. Both in the classical and quantum world, we can describe the evolution of observables with linear laws.

Here are four important uses of linear algebra to describe data:

Tool	Goal	Using	Example
Databases	Describe the data	Lists	Relational database, Adjacency matrix
Modeling	Model the data	Probability	Markov process, Filtering, Smoothing
Fitting	Reduce the data	Projections	Linear regression.
Computation	Manipulate the data	Algebra	Fourier theory.

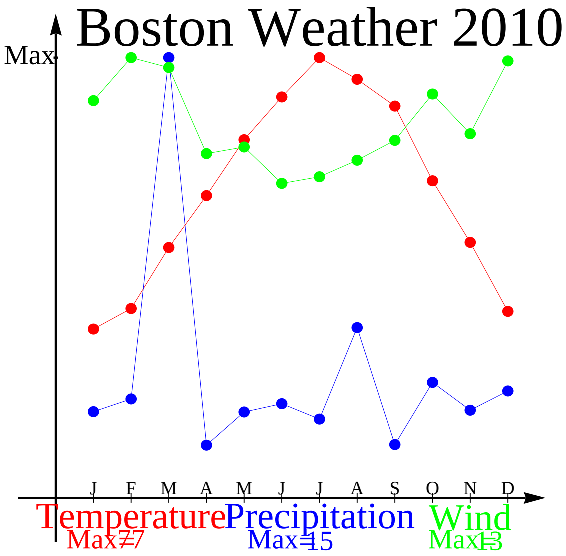
We have seen that a fundamental tool to organize data is the concept of a **list**. Mathematicians call this a **vector**. Since data can be added and scaled, data can be treated as vectors. We can also look at lists of lists. These are called **matrices**. Matrices are important because they allow to describe **relations** and **operations**. Given a matrix, we can access the data using coordinates. The entry $(3, 4)$ for example is the forth element in the third row. Having data organized in lists, one can manipulate them more easily. One can use **arithmetic** on entire lists or arrays. This is what spreadsheets do.

Observed quantities are functions defined on lists. One calls them also **random variables**. Because observables can be added or subtracted, they can be treated as vectors. If the data themselves are not numbers like the strings of a DNA, we can add and multiply numerical functions on these data. The function $X(x)$ for example could count the number of A terms in a genome sequence x with letters A,G,C,T which abbreviate Adenin, Guanin, Cytosin and Tymin. It is a fundamental and pretty modern insight that all mathematics can be described using algebras and operators. We have mentioned that data are often related and organized in relational form and that an array of data achieves this. Lets look at weather data accessed from http://www.nws.noaa.gov on January 4'th 2011, where one of the row coordinates is "time". The different data vectors are listed side by side and listed in form of a **matrix**.

Month	Year	Temperature	Precipitation	Wind
01	2010	29.6	2.91	12.0
02	2010	33.2	3.34	13.3
03	2010	43.9	14.87	13.0
04	2010	53.0	1.78	10.4
05	2010	62.8	2.90	10.6
06	2010	70.3	3.18	9.5
07	2010	77.2	2.66	9.7
08	2010	73.4	5.75	10.2
09	2010	68.7	1.80	10.8
10	2010	55.6	3.90	12.2
11	2010	44.8	2.96	11.0
12	2010	32.7	3.61	13.2

To illustrate how linear algebra enters, lets add up all the rows and divide by the number of rows. This is called the **average**. We can also look at the average squre distance to the mean, which is the variance. Its square root is called the **standard deviation**.

Month	Year	Temperature	Precipitation	Wind
6.5	2010	53.7667	4.13833	11.325


Boston Weather 2010
Max
J
F
M
A
M
J
J
A
S
O
N
D
Temperature
Max=77
Precipitation
Max=15
Wind
Max=13