# Homework #4

# Due March 25, 2024, 11:59pm

***Problem 4.1.***

Read the `sidneybw1000.png` gray-scale image, 1000x750 pixels, dtype='uint8'. Convert the image to a 2D numpy array. Then, compute the first 200 eigenvalues and eigenvectors.

(i) Plot the spectrum of the eigenvalues as a function of their rank.

(ii) Also plot the cumulative fraction of the variance in each mode.

(iii) Finally, use only the N largest eigenvalues, reconstruct the truncated image, and display, for different values of N, from 10 to 100. Determine, which gives an adequate quality reconstruction of the image. For each value of N, compute the amount of storage needed and compare it to the original image size to get the compression ratio.

(iv) Repeat the exercise with the `einstein.png` image.

*Hints:*
*(a) for reading the image use the `imageio` package.*
*(b) For a truncated SVD, use `scipy.sparse.linalg.svds`.*
*Beware that in svds the largest eigenvalues are last.*
*(c) For computing the dot products of two matrices A,B in **numpy** use the operator A@B*
*(d) The eigenvalues from the image SVD need to be squared to represent the variance*

***Problem 4.2.***

The `temperature.csv` text file contains the daily mean temperature in F° for the cities Helsinki and Melbourne for the years 2013 and 2014. The first column is the day measured from 01-01-2013. Build a linear model that fits the temperature variations with a linear combination of sin and cos functions. The fundamental period should be 1 year (365 days), use up to the third harmonic. Plot the best fit solutions on top of the data.

*Hint: Watch out for the header line in the text file*

***Problem 4.3.***

A die is rolled 24 times. Use the Central Limit Theorem analytically to estimate the probability that

a. The sum is greater than 84
b. The sum is equal to 84
c. Perform 10,000 numerical realizations to illustrate the result

***Problem 4.4.***

The file `atacama-2012-sample.csv` contains hourly measurements from various sensors from the Atacama desert in Chile. The sensors `c3` and `c4` measure the $CO_2$ concentration in part per million (ppm), uncorrected for the high altitude (Atacama is at 16,000 ft, and the air pressure is about half of the sea-level one). The columns `t5` and `t6` are the outside temperature from two sensors in $^oC$. The time is displayed in different granularities (hours from the beginning of the experiment, hours within each day (`dhours`), days from the beginning of the experiment. There is a glitch in the $CO_2$ sensor values on day 70, ignore those values (set them to zero).

The expression below defines the cross-correlation between two different time-series *a* and *b*.

$$C_{ab}(\tau) = \frac{1}{N}\sum_t \left(a(t) - \langle a \rangle\right)\left(b(t+\tau) - \langle b \rangle\right)$$

Here *N* is the number of measurements included in the sum, <*a*> and <*b*> are the averages of the two series.

a. Consider the time series of the two temperature sensors. Break these into daily vectors, and compute the top 3 principal components. Guess the meaning of each component. Expand each vector on the basis of the top 3 components. Estimate the fraction of variance contained in the three components. Estimate the truncation error due to using three components only. Display the amplitudes of the components as a function of time during the observations.
b. Repeat the above with 5 components and compare.
c. Compute the temporal autocorrelation function of both the average temperature and the average $CO_2$ concentration, out to 48 hours. Interpret the result.
d. Compute the temporal cross-correlation function between the average temperature and the average CO2 concentration, out to 48 hours. Discuss the meaning of the result.

*Hint: look out for missing or erroneous values in the data, often marked with NaN (not a number).*