

**Dipartimento di Matematica e Applicazioni (DMA) “Renato Caccioppoli”**  
**Università degli Studi di Napoli Federico II**

Il **Prof. Erkki Somersalo** (Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University) terrà un corso dal titolo

*Mathematics of data mining and pattern recognition*

secondo il calendario seguente

- 2 Maggio 2018, 9.30-12.30 (Aula G-DMA)
- 8 Maggio 2018, 11.00-13.00 (Aula G-DMA)
- 9 Maggio 2018, 9.30-12.30 (Aula G-DMA)
- 14 Maggio 2018, 9.00-11.00 (Aula G-DMA)
- 15 Maggio 2018, 11.00-13.00 (Aula G-DMA)
- 16 Maggio 2018, 9.30-12.30 (Aula G-DMA)
- 21 Maggio 2018, 11.00-13.00 (Aula F-DMA)
- 21 Maggio 2018, 14.30-16.30 (Aula F-DMA)

## **Title:** Mathematics of Data Mining and Pattern Recognition

The increasing capacity of computers and devices harvesting and transporting information makes it necessary to develop fast and efficient tools to extract pertinent information out of tremendous data flows. Wikipedia defines data mining as “the process of extracting hidden patterns from data”. Human sensory system and brain are notably effective in extracting patterns from visual or auditory data, but it remains a challenge to device a machine to do similar pattern recognition from digital data in an unsupervised fashion.

This course discusses some of the commonly used algorithms and principles on which the modern data mining and pattern recognition are based. The course is algorithm-oriented, and the functionality of the algorithms is demonstrated first with well-controlled toy problems, and subsequently applied to real data sets. Data mining can be taught with different flavors: the emphasis in this class will be on linear algebraic methods and their implementation, and less on statistical interpretations, which would be another versatile view. Examples of data mining problems considered in this course include

- Automatic classification of hand written digits
- Classification of medical patient data
- Segmentation of medical images
- Page ranking of search engine data
- Sentence extraction and key words
- Target identification in remote sensing applications

Tentatively, the course covers the following list of methods:

- Data reduction and visualization by principal component analysis (PCA)
- Clustering:  $k$ -means and  $k$ -medoids.
- Cluster visualization, Linear Discriminant Analysis (LDA)
- Non-negative Matrix Factorization (NMF)
- Self-organizing Maps (SOM)
- Classifiers: Nearest neighbors, Learning Vector Quantization (LVQ), SVD classifiers
- Text Mining: Query matching, Latent Semantic Indexing (LSI)
- Page ranking
- Classification trees, random forest algorithm.