# A Decade of Progress in Indexing and Mining Large Time Series Databases

Eamonn Keogh
Computer Science & Engineering Department
University of California, Riverside
Riverside, CA 92521
*eamonn@cs.ucr.edu*

## ABSTRACT

Time series data is ubiquitous; large volumes of time series data are routinely created in scientific, industrial, entertainment, medical and biological domains. Examples include gene expression data, electrocardiograms, electroencephalograms, gait analysis, stock market quotes, space telemetry etc. Although statisticians have worked with time series for more than a century, many of their techniques hold little utility for researchers working with massive time series databases.

A decade ago, a seminal paper by Faloutsos, Ranganathan, Manolopoulos appeared in SIGMOD. The paper, *Fast Subsequence Matching in Time-Series Databases*, has spawned at least a thousand references and extensions in the database/ data mining and information retrieval communities. This tutorial will summarize the decade of progress since this influential paper appeared.

## 1. INTRODUCTION

Time series data is ubiquitous; large volumes of time series data are routinely created in scientific, industrial, entertainment, medical and biological domains. Examples include gene expression data, electrocardiograms, electroencephalograms, gait analysis, stock market quotes, space telemetry etc. This tutorial will summarize the progress made in indexing and mining this rich data source. A brief outline of the tutorial includes:

- Introduction, Motivation
  - The ubiquity of time series
  - Converting text/DNA/shapes/video to time series
- The Utility of Similarity Measurements
  - Properties of distance measures
  - The Minkowski metrics
  - Preprocessing the data
  - Time warped and time scaled measures
- Indexing Time Series
  - Spatial Access Methods and the curse of dimensionality
  - The GEMINI Framework
  - Dimensionality reduction
    (Discrete Fourier Transform, Discrete Wavelet Transform, Singular Value Decomposition, Piecewise Linear Approximation, Chebyshev Polynomials, Symbolic Approximation, Symbolic Aggregate Approximation Piecewise Aggregate Approximation, Adaptive Piecewise Constant Approximation, Random mappings)
  - Empirical Comparison of all methods
- Mining Time Series
  - Clustering short time series (shape based clustering)
  - Clustering long time series (model based clustering)
  - Novelty/anomaly/interestingness detection in time series
  - Burst discovery in time series
  - Summarizing time series with text/graphics (visualization)
  - Finding repeated patterns (motifs) in time series
  - Late breaking topics from this years conferences
  - Top 10 problems to solve
  - Future directions
- Summary, Conclusions


*VLDB '06*, September 12–15, 2006, Seoul, Korea.


## 2. BIOGRAPHY OF SPEAKER

Dr. Keoghs research interests are in Data Mining, Machine Learning and Information Retrieval. He has published papers on time series in the SIGMOD, SIGKDD, SIGIR, SIGGRAPH, VLDB, EDBT, ICML, PKDD, PAKDD, IEEE ICDM, IEEE ICDE, SIAM SDM, IDEAL, FQAS, SSDM, AI and INTERFACE conferences and in the TODS, DMKD, KAIS, INFORMATION VISUALIZATION and IJTAI journals. Several of his papers have won "best paper" awards and in addition he has won several teaching awards. He is the recipient of a 5-year NSF Career Award for "*Efficient Discovery of Previously Unknown Patterns and Relationships in Massive Time Series Databases*" and a grant from Aerospace Corp to develop a time series visualization tool for monitoring space launch telemetry. His papers on time series data mining have been referenced well over 2,000 times [2].

## 3. ACKNOWLEDGMENTS

I gratefully acknowledge Dr. Chotirat Ratanamahatana, Dr. Jessica Lin, Li Wei and Xiaopeng Xi for help with the slides.

This tutorial was partly funded by the National Science Foundation under grant IIS-0237918.

## 4. REFERENCES

[1] Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in timeseries databases, In Proc. of the ACM SIGMOD Int.Conf. on Management of Data, Minneapolis, MN (1994) 419-429

[2] www.cs.ucr.edu/~eamonn/selected_publications.htm