感測資料分析面面觀—從災難預報、交通預測、到行爲鑑定

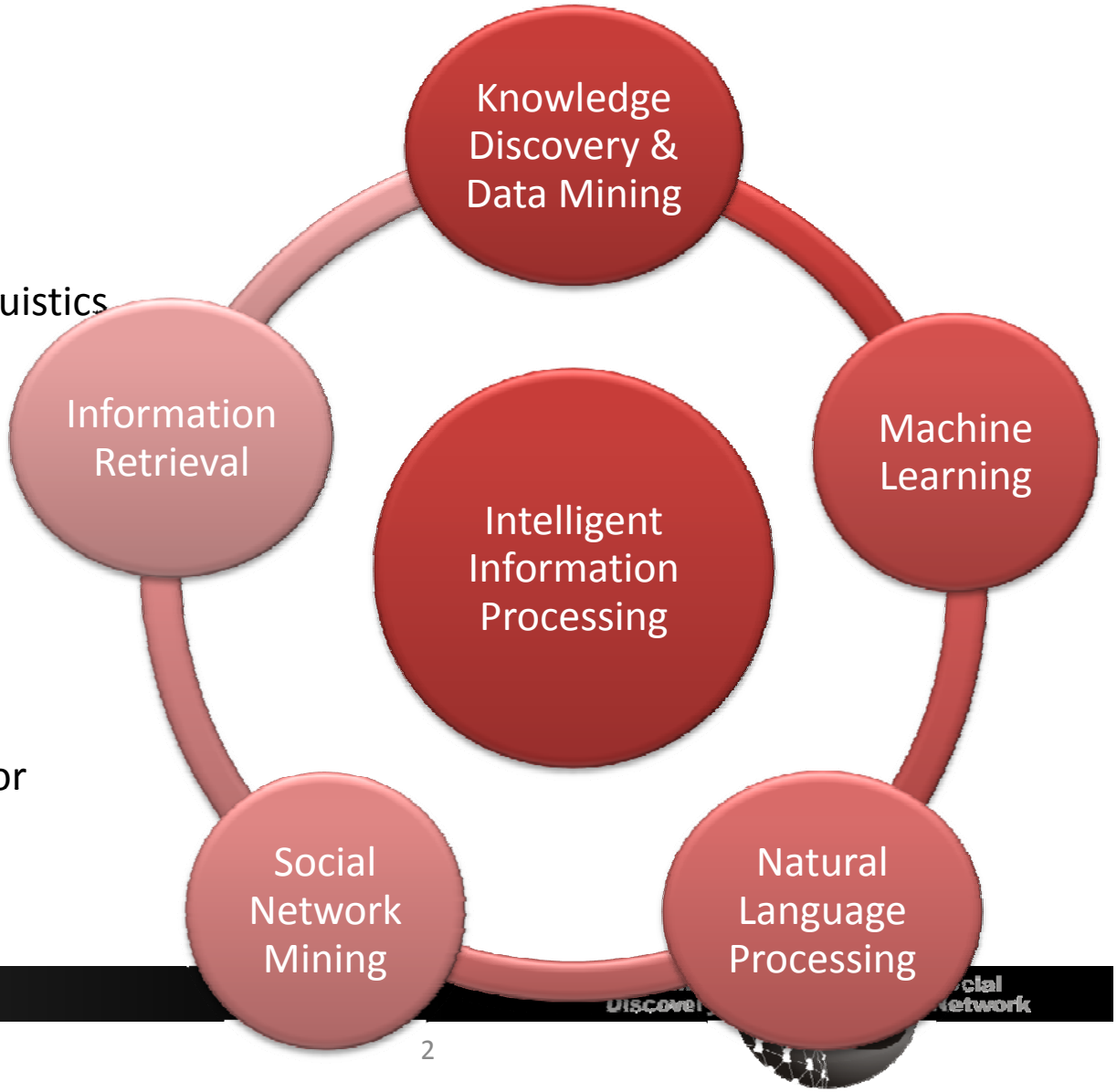
林守德教授

CSIE/GINM, NTU

sdlin@csie.ntu.edu.tw

About Myself

- **PI:** Shou-de Lin
  - B.S. in NTUEE
  - M.S. in EECS, UM
  - M.S. in Computational Linguistics USC
  - Ph.D. in CS, USC
  - Postdoc in LANL
- **Members:**
  - 6 Ph.D. students
  - 14 MS students
  - 5 undergraduate students
- **Position in NTU-INTEL Lab**
  - PI for Heterogeneous Sensor Network Analysis Project



資料分析與處理在物聯網中扮演的角色

- 當Sensors都佈建好了，資料的傳送都沒有問題了，物聯網是否就成功了？ Maybe not!!
- 如果不能對資料作智慧型的處理，用資料來預測即將發生的事情，物聯網就無法發揮功效。
- M2M資料分析可能的應用
  - 健康照護
  - 交通狀況預測
  - 節能省電
  - 災難監控

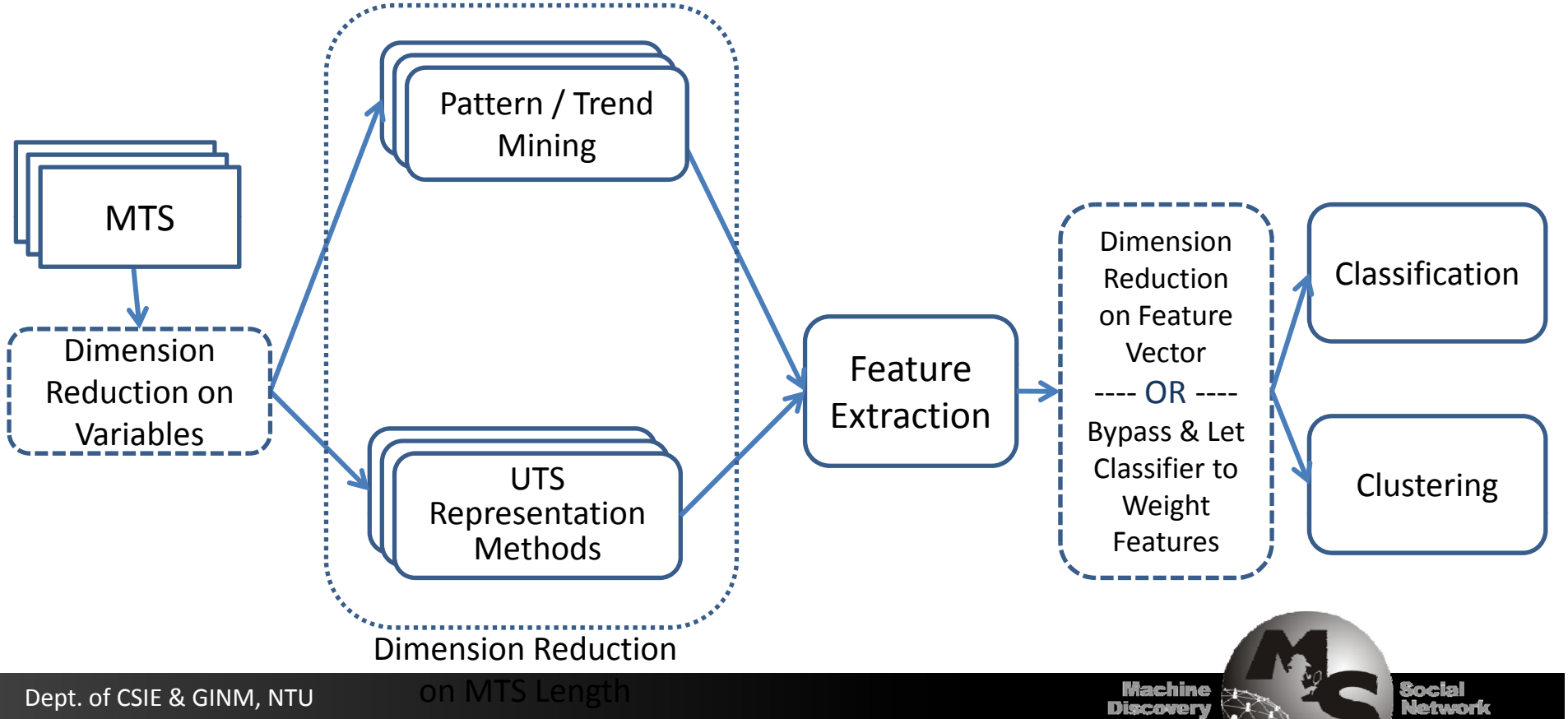
Information without processing is garbage!!

- M2M framework creates the paradigm shift for data analysis research and applications:
  - Single stream Data → Information → Knowledge → intelligent tasks (e.g. recognition, decision making)

What we can do with M2M data?

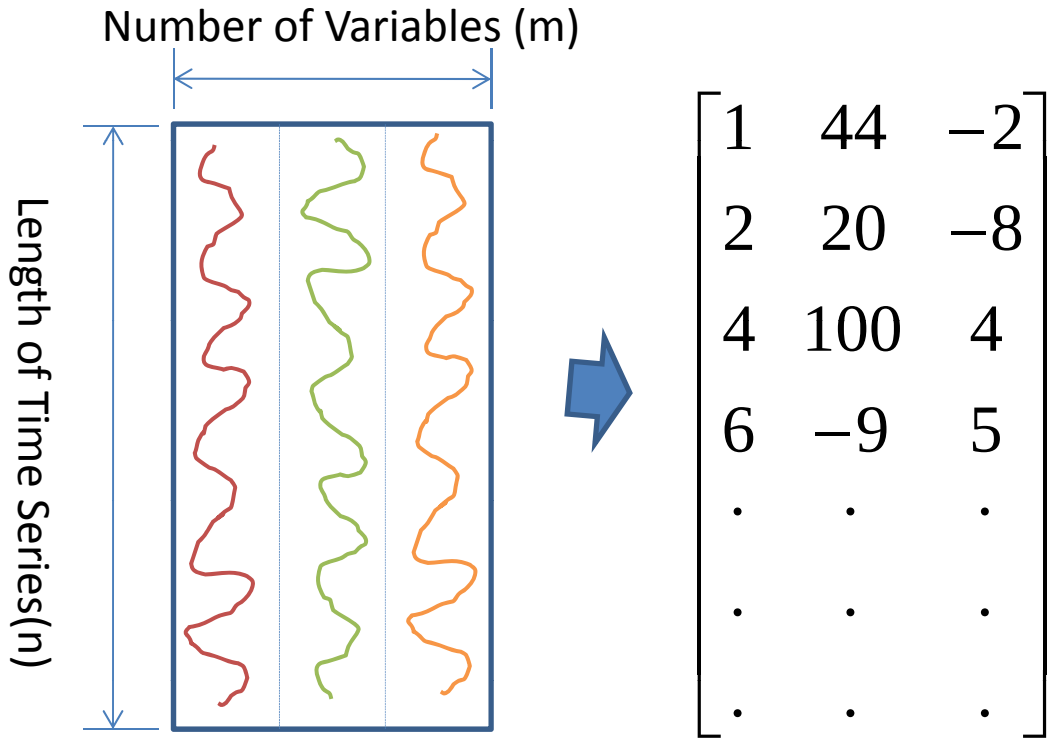
- Pattern discovery
- Segmentation
- Classification
- Anomaly detection
- Event prediction
- …

M2M Data 分析流程



MTS Representation

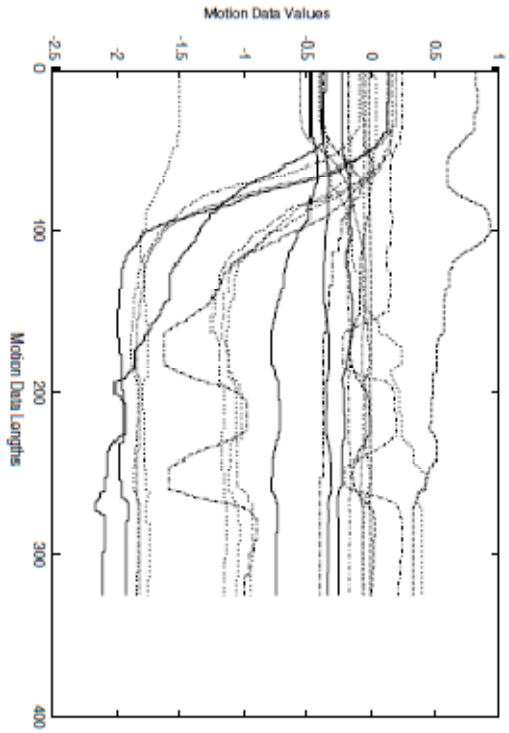
- **MTS Feature Extraction**



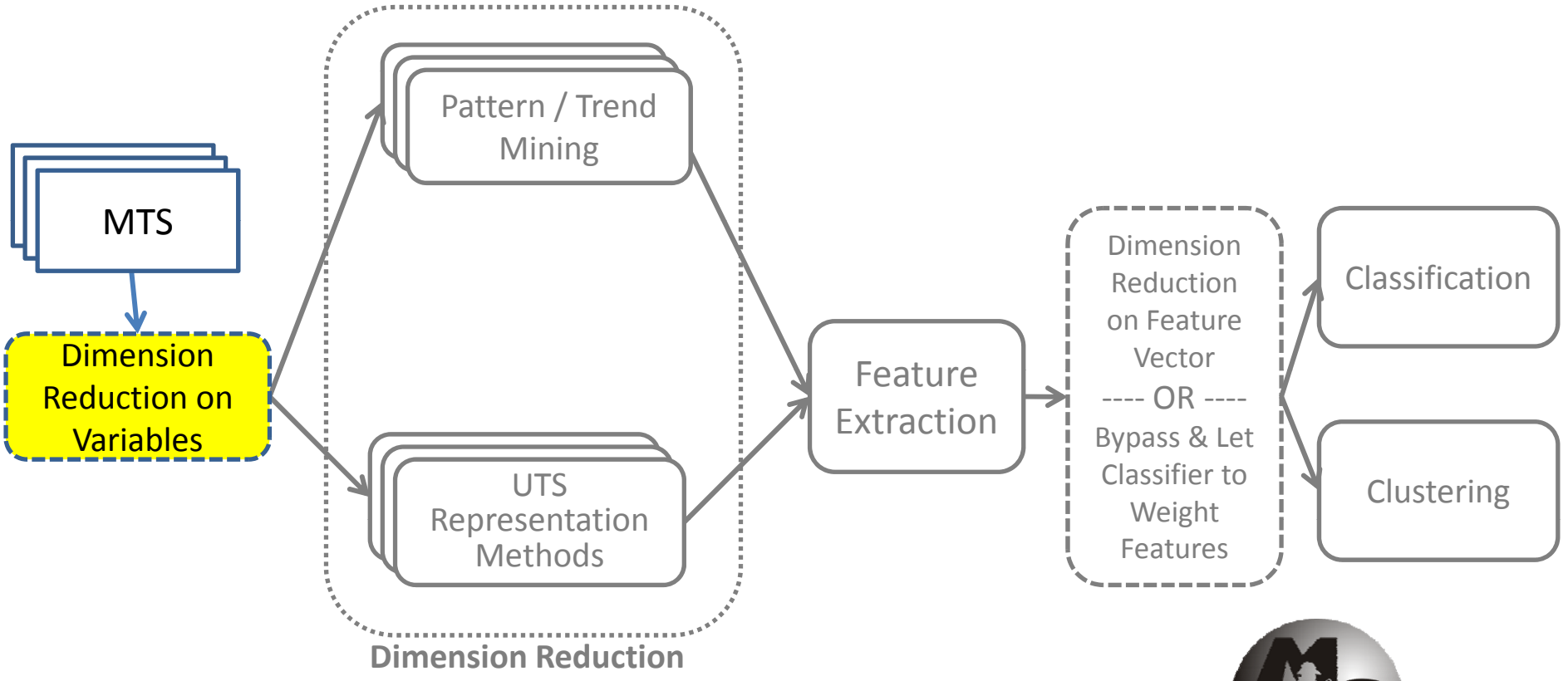
$$\begin{bmatrix} 1 & 44 & -2 \\ 2 & 20 & -8 \\ 4 & 100 & 4 \\ 6 & -9 & 5 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$$

A Matrix Representation (X)

- **Two MTS Examples:**



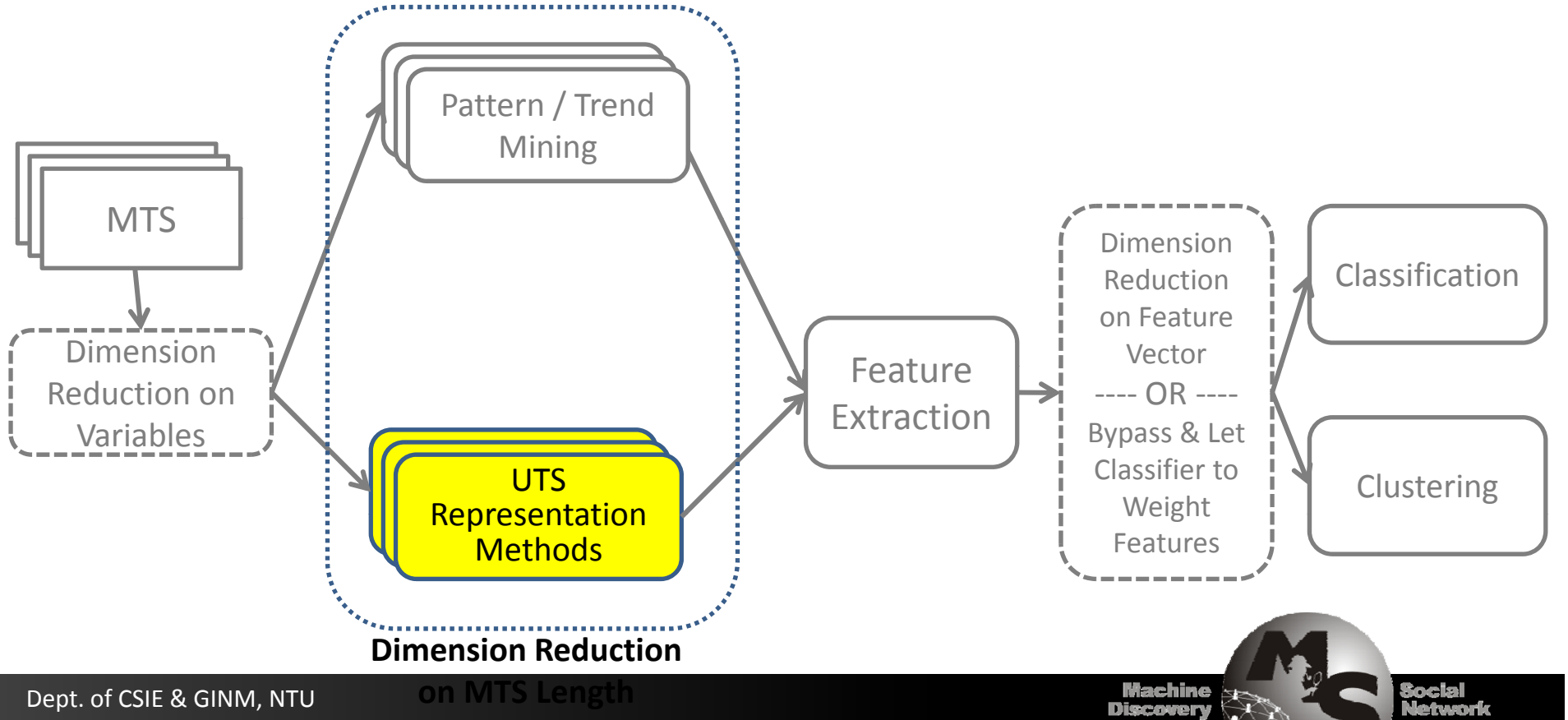
UTS Representation Methods



Dimension Reduction on Variables

- Sensors data can be dependent, and we do not have to store/process all of them
- There are lots of dimension reduction techniques we can apply here.

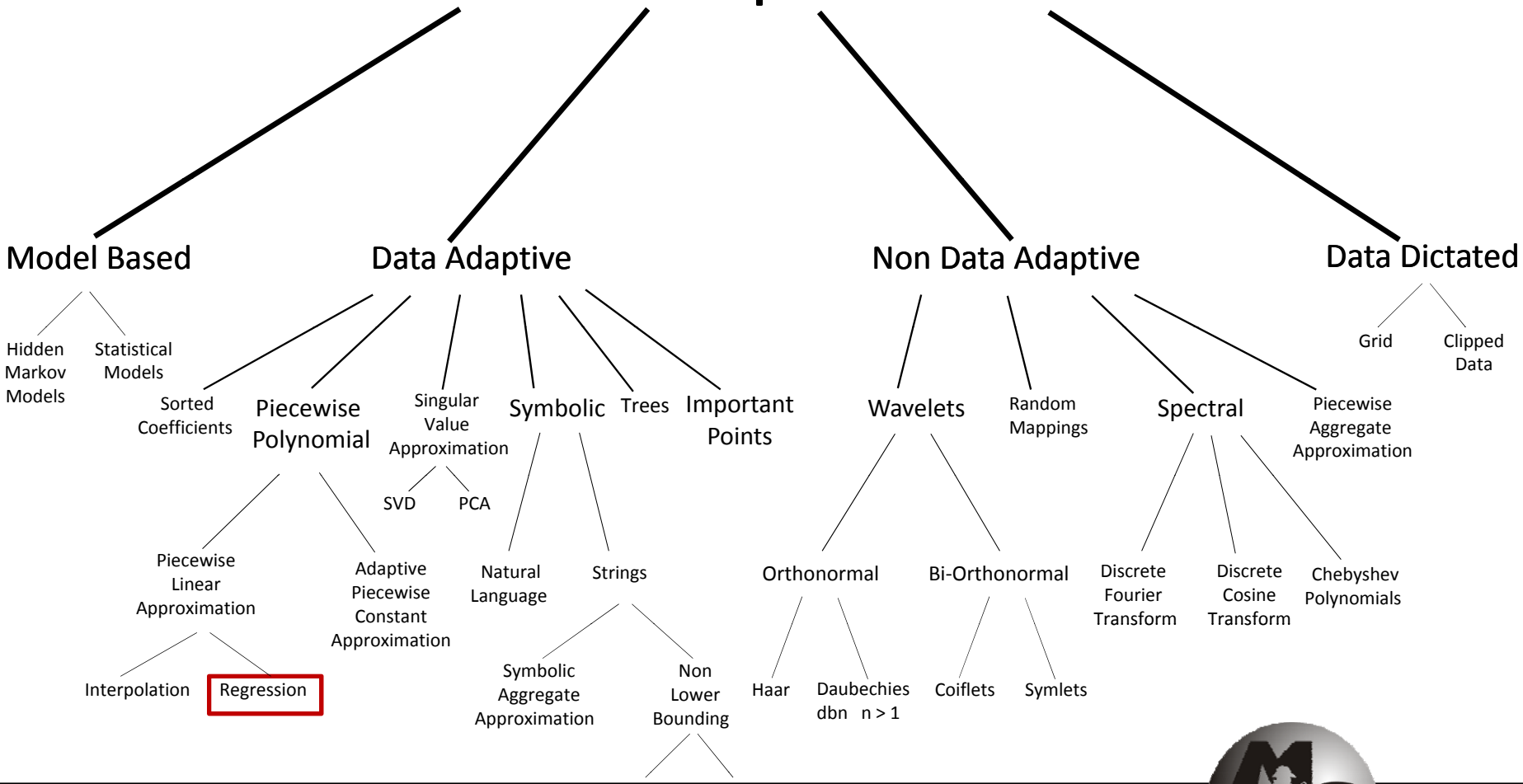
UTS Representation Methods



UTS Representation Methods

- UTS Representation Methods
  - Non-Data Adaptive Representation Methods
  - Data Adaptive Representation Methods
  - Model-based Representation Methods
  - Data-dictated Representation Methods

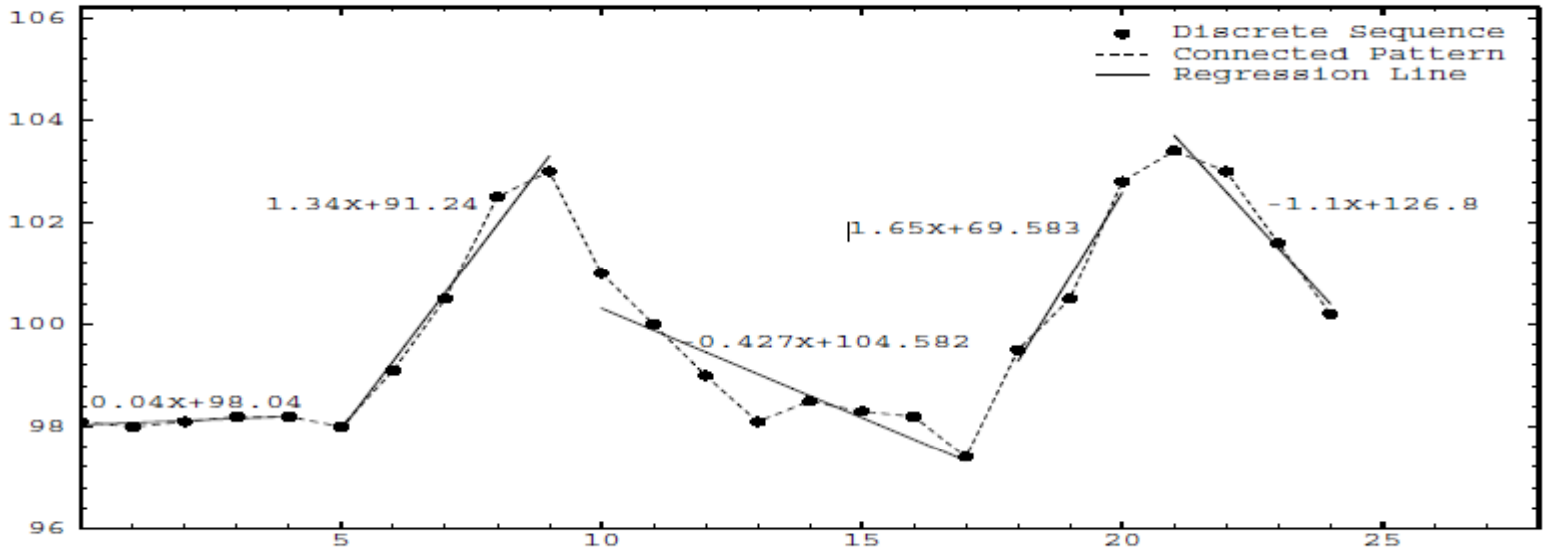
Time Series Representations



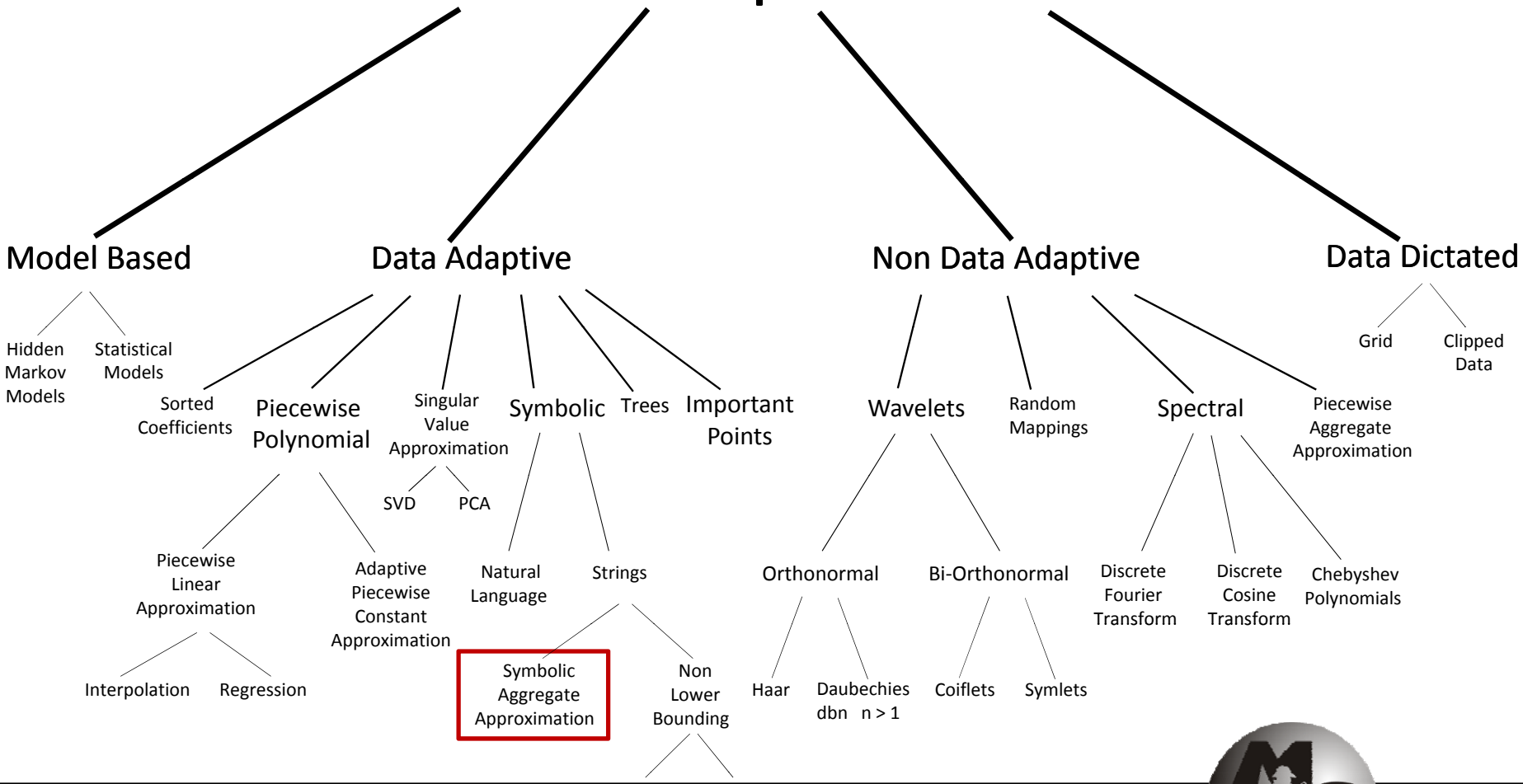
Machine Discovery Social Network

Piecewise Linear Approximation - Regression

- Window-based Method
  - (*H. Shatkay, "Approximate queries and representations for large data sequences", 1996*)
  - Fit points in the window by a D-degree polynomial



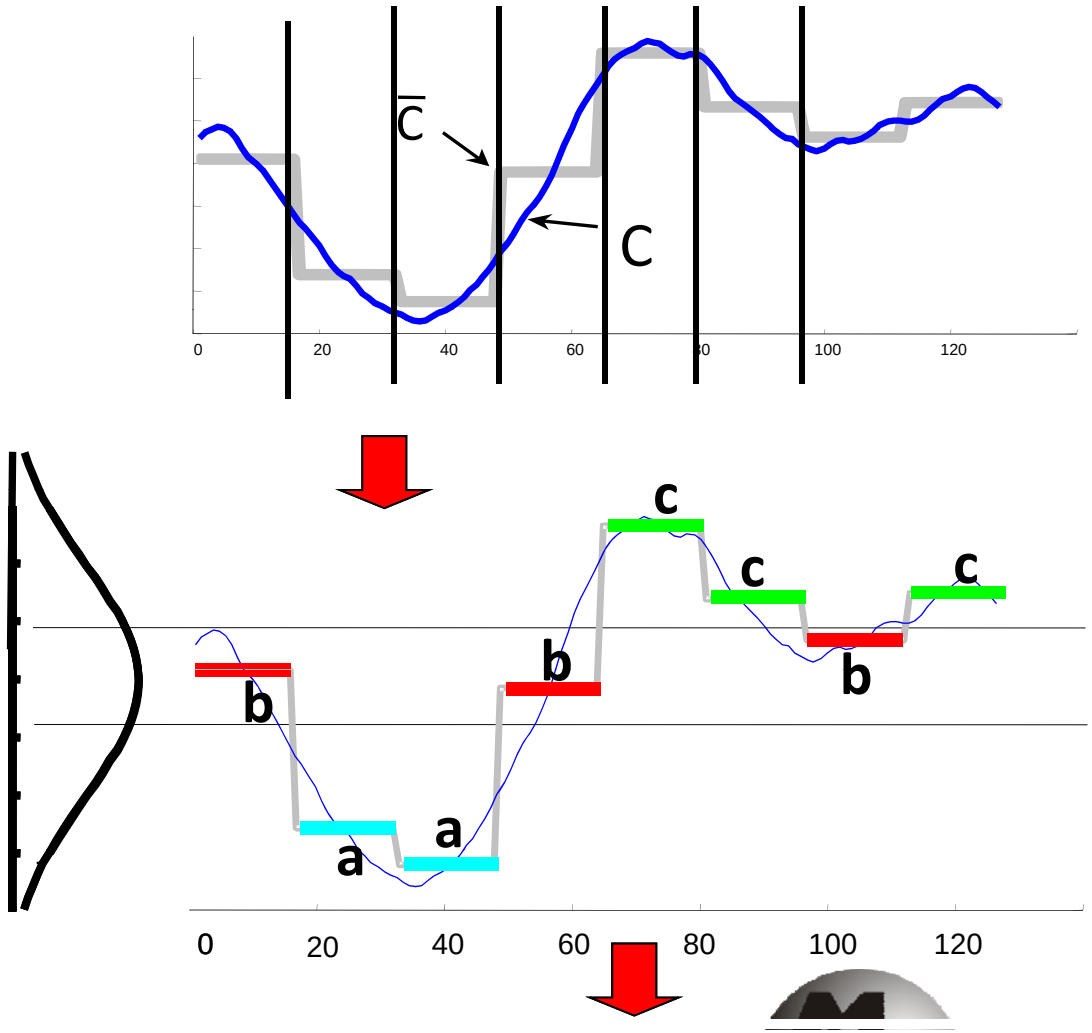
Time Series Representations



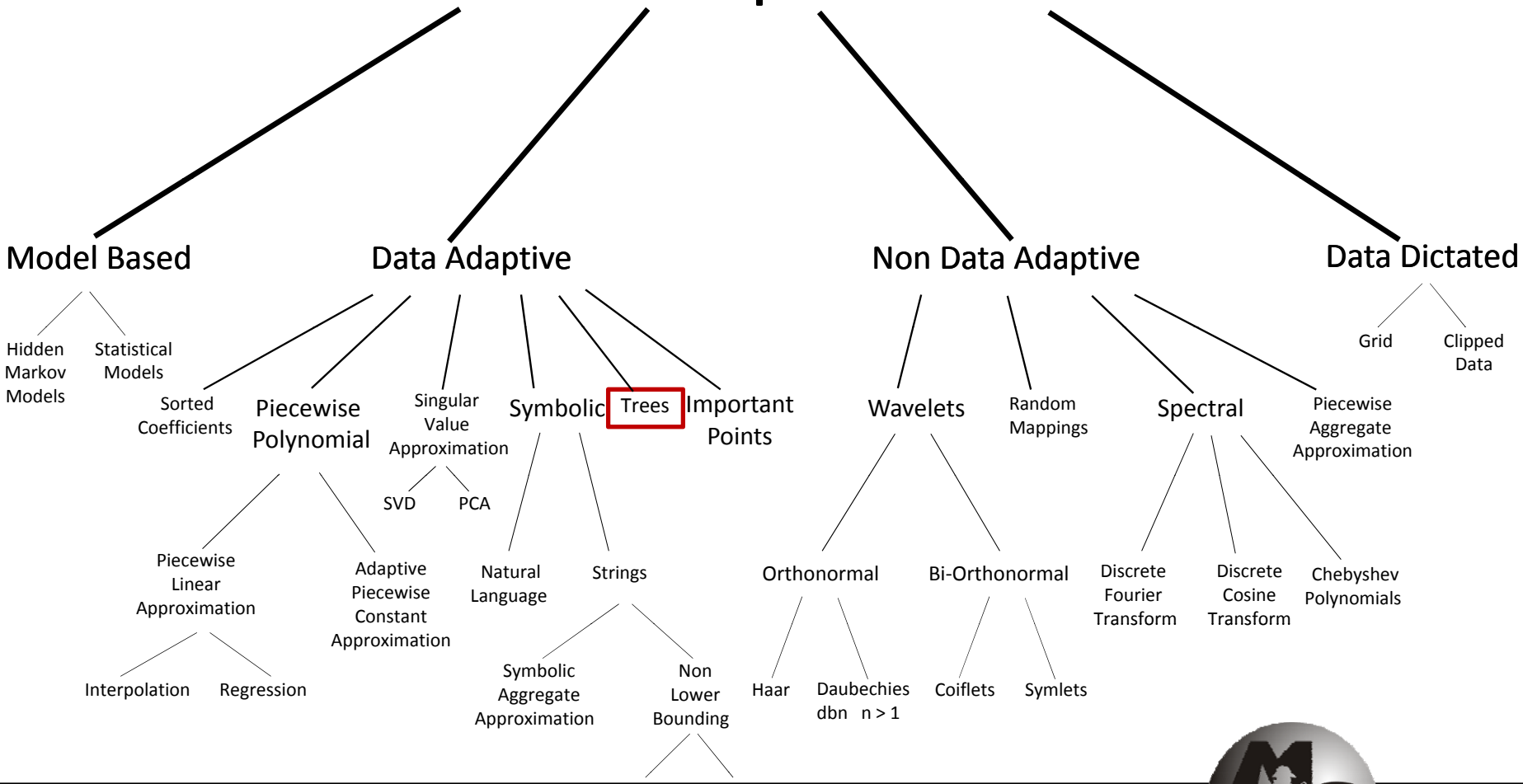
Machine Discovery Social Network

Symbolic Aggregate Approximation (SAX)

- First convert the time series to PAA (i.e. Segmented Mean) representation
- Then convert it to symbolic string
- It take linear time



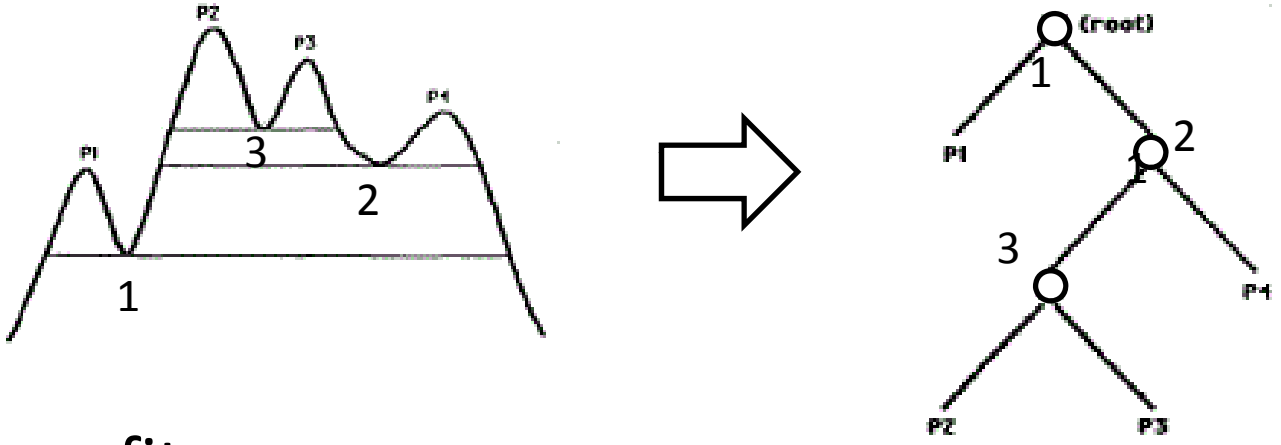
Time Series Representations



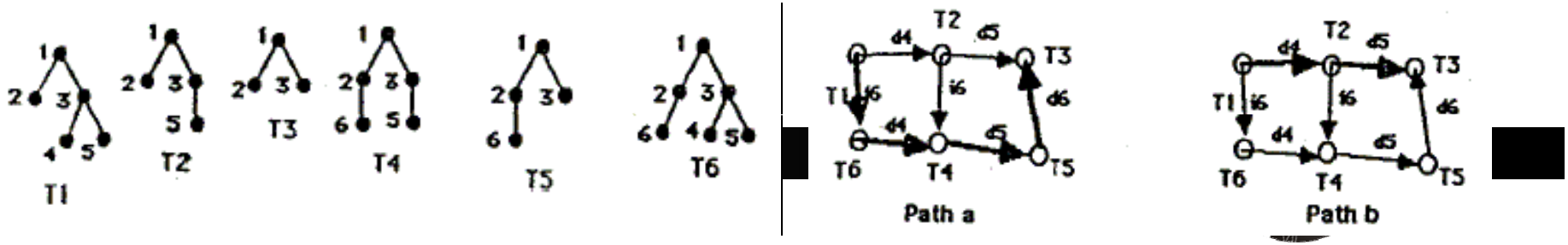
Machine Discovery Social Network

Relational Tree Representation

(Scott W. Shaw et al, "Structural Processing of Waveforms as Trees", IEEE Transactions on Acoustics, Speech and Signal Processing, 1990)

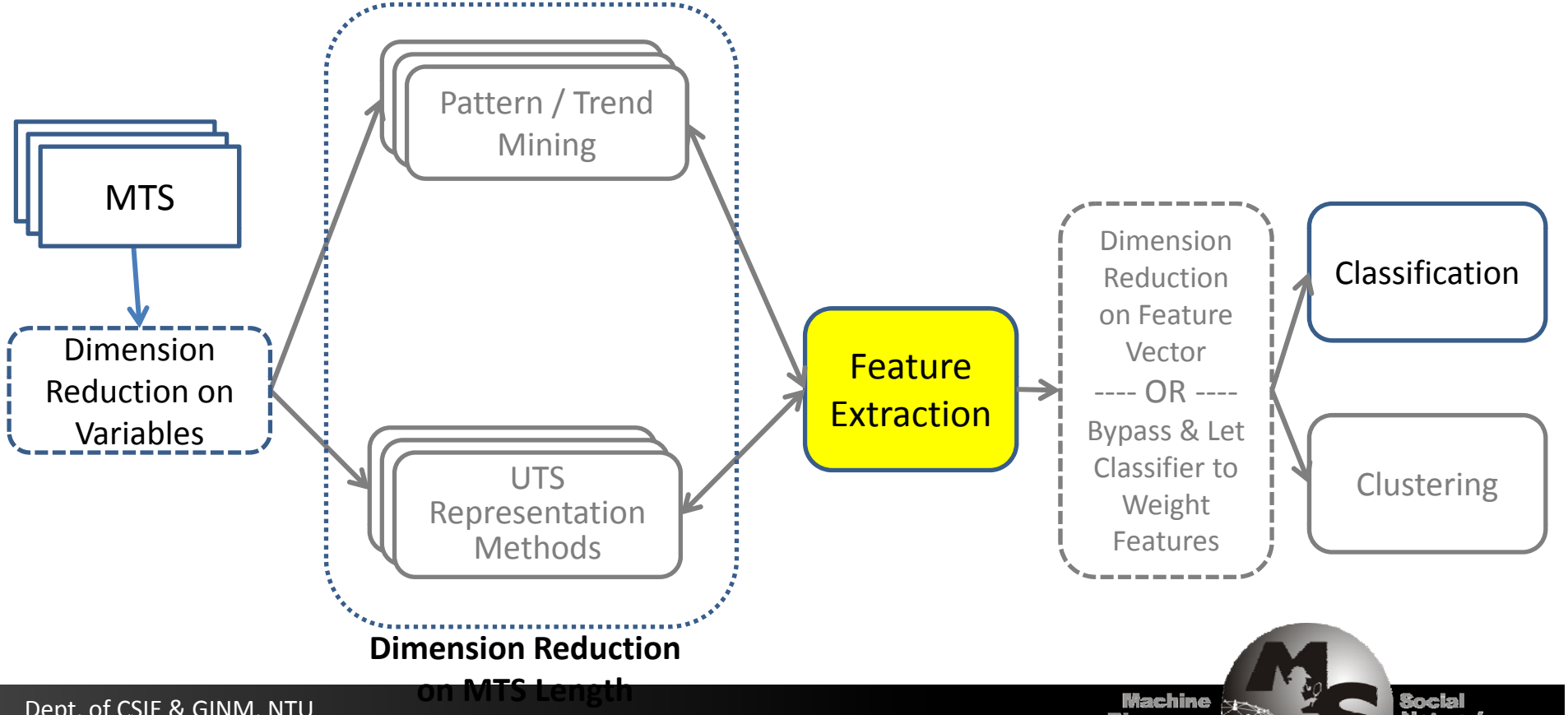


- Benefits
  - Invariant to monotonic scaling along the time or domain axis
- Distance Measurement
  - The distance between two trees is the minimum path length on the directed graph from one tree to another tree



Feature Extraction from Signal Processing (1/2)

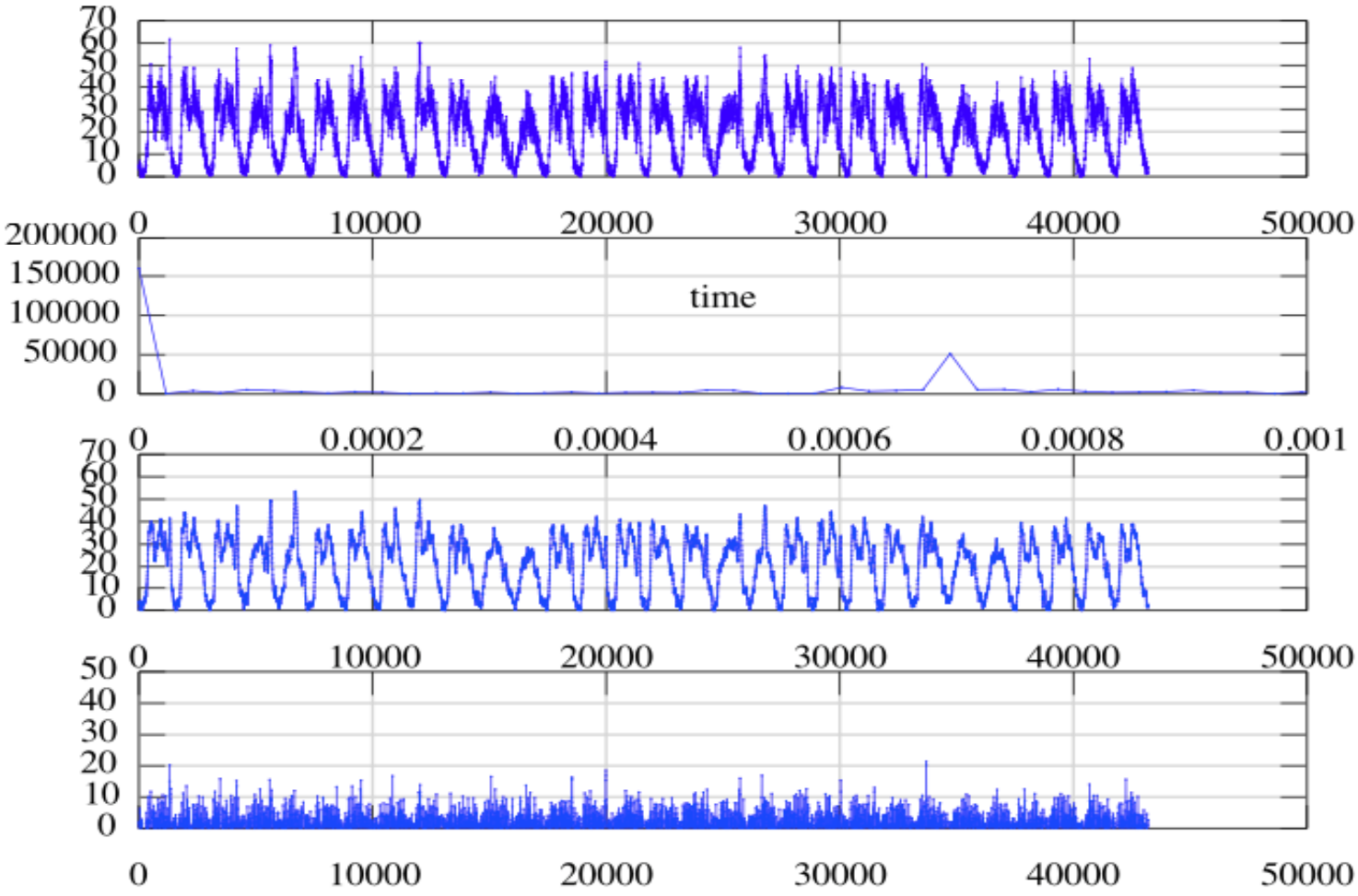
(T. N. Lal et al, "Support Vector Channel Selection in BCI", IEEE Transactions on Biomedical Engineering, 2004)



Dodgers Loop Data: Data Description

- Detect game event basic on traffic data
- Collected in 2005(4/10~8/24)
- Data:
  - Traffic data
    - Number of Observations: 50400. 25 weeks, 288 time slices per day (5 minute count aggregates).
  - Event data
    - Date: MM/DD/YY
    - Begin event time: HH:MM:SS (military)
    - End event time: HH:MM:SS (military)
    - Game attendance
    - Away team
    - W/L score

Dodgers Loop Data:
30 days since 4/12



Time domain

Freq domain

IFFT(F(f))
with
threshold = 1000

X[t]-IFFT(F(f))

Sitex02 dataset

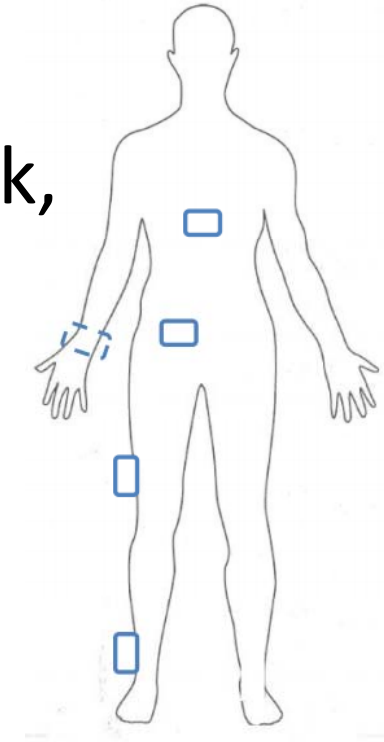
- Vehicle classification:
  - collected during a real world WDSN experiment carried out at Twenty-nine Palms, CA in November 2001
  - Using acoustic/seismic time series to classify the types of moving vehicles
  - http://www.ece.wisc.edu/~sensit/publications/ClassificationFusion.pdf

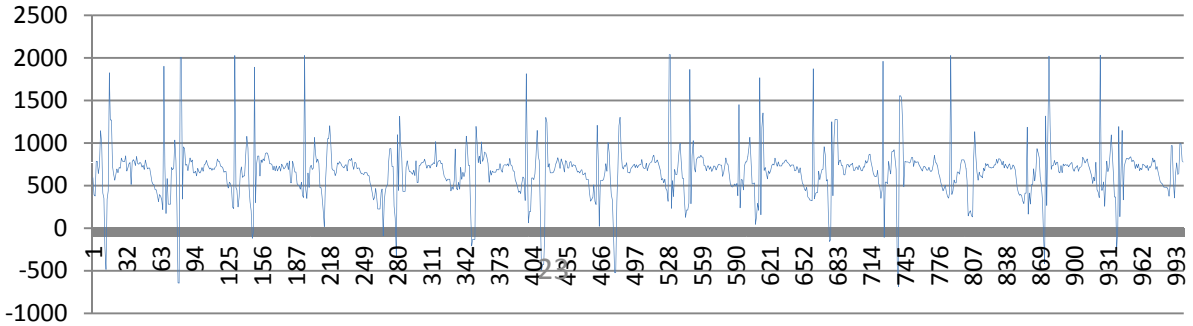
Sitex02 dataset: collection

Photo from http://www.seapahn.us/gal/sitex02/dcp_0911.html

Body Sensor Network

- Sensors (accelerometers, gyroscopes) record activities such as shelving a book, walking, sitting up, etc.
- Multiple datasets, each with unique sensor configurations
- Tasks include such things as classifying activity type and determining average stride time.

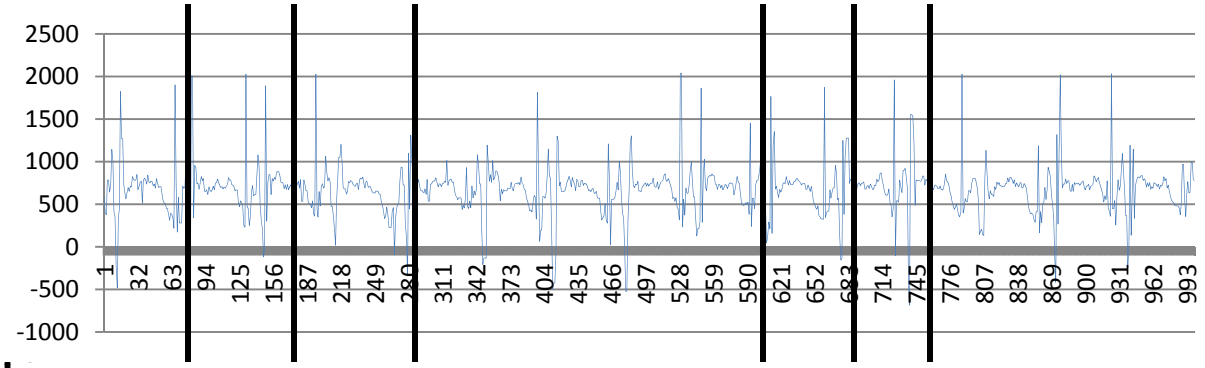




Body Sensor Network Contest

- Task 3: action classification
- Task 1: strike duration estimation
- Task 2: action segmentation and classification

Task 3: B

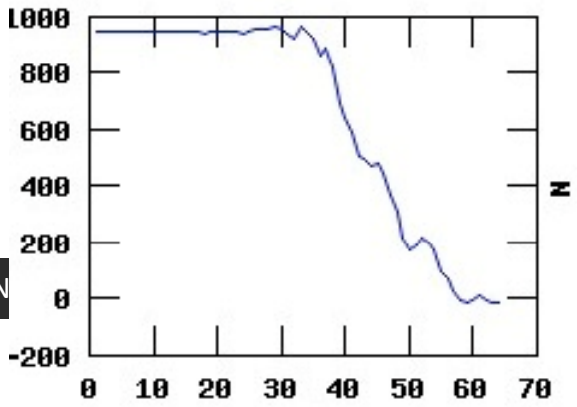


- Sit to stand detection
  - For this task, the algorithm needs to be able to distinguish trials of “sit to stand” from trials of different types
- Training & testing
  - The data is accelerometer and gyroscope reading over time (roughly 20 sensor data)
  - Segmentation is known
- Evaluation:
  - For each of the testing trials, the algorithm must classify an action as “sit to stand” or “not sit to stand”
  - Objective function to maximize: number of correct classifications

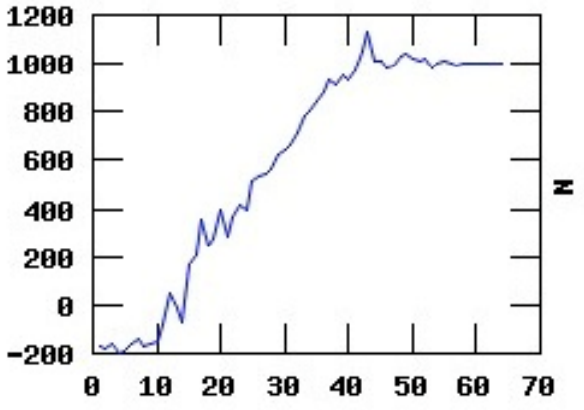
Analysis

- Size of training dataset is small, but feature size is large (avg of 60 for each class)
  - We need to reduce the dimensions to avoid overfitting
- There only a few subjects in training data, but there are many subjects in test set
- The length of raw features is not fixed
- Different action have unique waveforms, some complementary

Sit To Stand



Stand To Sit



Our Approach

- Raw features : the sensor values collected in time intervals
- Linear scale the raw feature
  - Actions of the same type may be performed at different speeds
  - Data must be scaled to the same absolute length in order to perform feature-based learning
- Perform Fast Fourier transform on the data from a single sensor
  - Filter high-frequency noise
- Training model & predict by Support Vector Machine
  - Used linear kernel

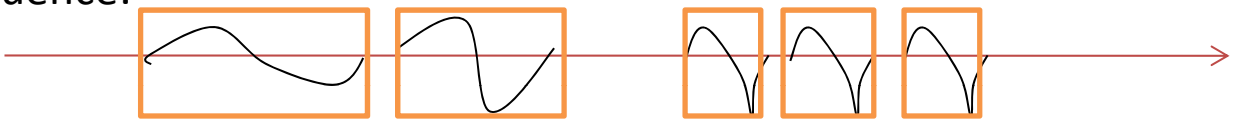
Results

- Testing error using each of the three datasets provided achieves an accuracy of 97~98%
  - FFT , wavelet or raw data were used as features.
  - FFT features proved to be most useful for this task

Task 1: Multi-Class Action Segmentation/Classification

- 4 sensor nodes attached to each test subject, each with 5 sensors (measure acceleration, rotation).
- Goal : Given a testing sequence, detect 9 kinds of actions along with begin and end times of the detected action within a tolerance of 0.5 seconds

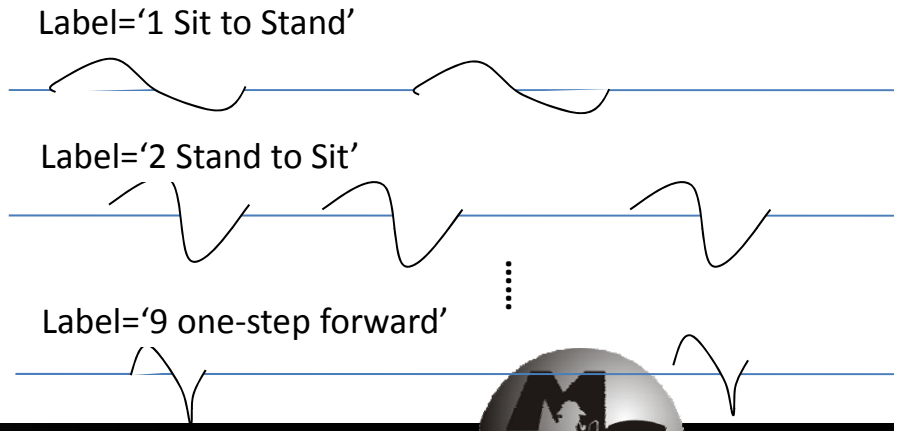
Testing Sequence:



prediction: 1 2 9 9 9

- Provided Training Data:

In all, there were 9 actions performed by 3 test subjects.



Analysis

- Since the test subjects used in the creation of the testing and training were different, high generality of the model was required.
- For each label, we had (4 sensor)*(5 readings/sensor)*Length features, while there were only (3 person)*(10 example/person) training examples. Overfitting may be an issue.
- There may be unknown actions performed in the testing sequence. Thus, negative examples in testing can be much more diverse than in training.
  - But after we studied the testing sequence, we found there were actually no unknown actions performed

Our Approach

- Model Selection:
  - After trying different classifiers, we found SVM with a linear kernel and large margin can yield stable and generalizable performance compared with other approaches we tried.
- Feature Extraction:
  - Raw Features: Scale windows to length of 64 data points
  - FFT Features: Transform sequence into Fourier Coefficients, use low frequency portion (top 16), e.g. low-pass filter.
  - Wavelet Features: Transform sequence into wavelet coefficients.
- Window Selection:
  - Strategy 1: Move sliding window with different sizes, classify each window, and select those with most confidence.
  - Strategy 2: Find those segments with significant vibration and use the classifier to determine which actions they are. Here we need to assume vibration is not caused by unknown actions.

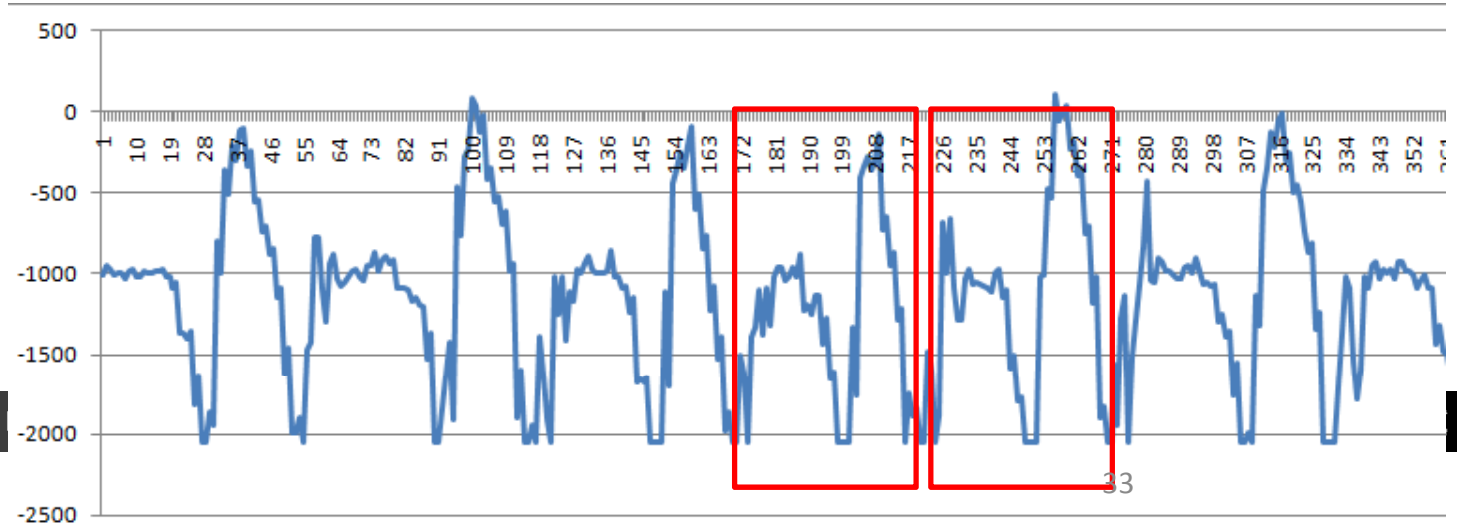
Results

- Competition-driven:
  - If the sensors are in a stable state most of the time, or for a dataset like the one in this competition, strategy 2 is better for its simplicity and efficiency.

- Research value:
  - Strategy 1 is closer to real world applications.
  - How to design a algorithm that deals with segmentation and classification at the same time is a valuable research topic.

BSN: Average Stride Time Calculation

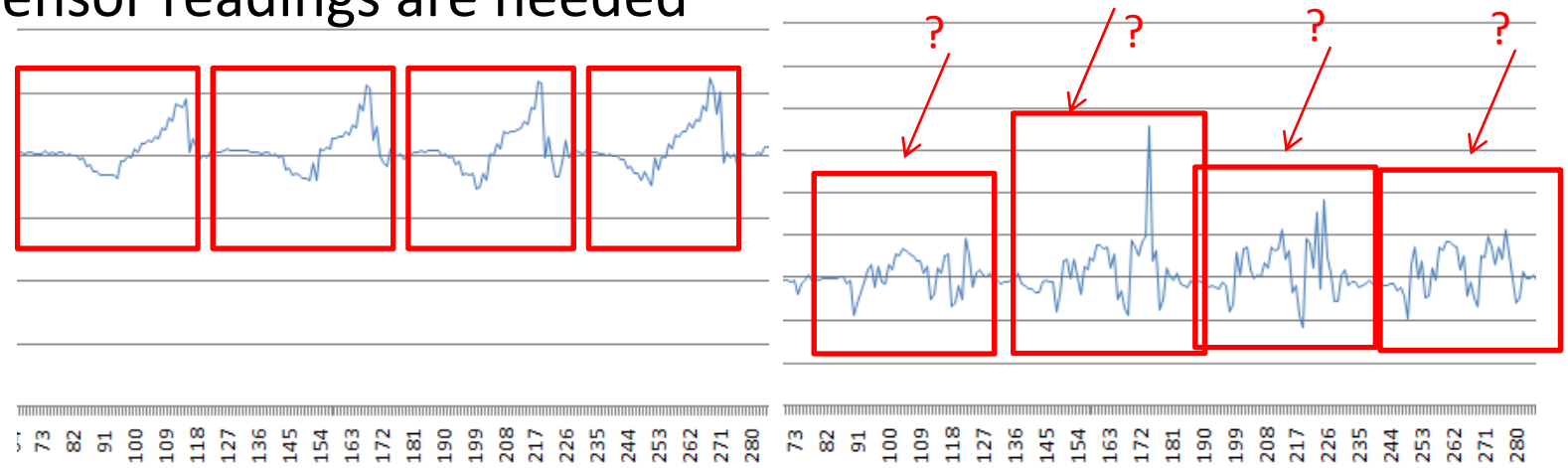
- Find the average stride time = $Time_{Total}/Num_{Stride}$
- Details:
  - Three different datasets, each having a different sensor configuration and with different test subjects at distinct locations
  - One of the datasets included trials at different walking speeds and inclinations
  - Calculate the full cycle of sensor reading on one leg



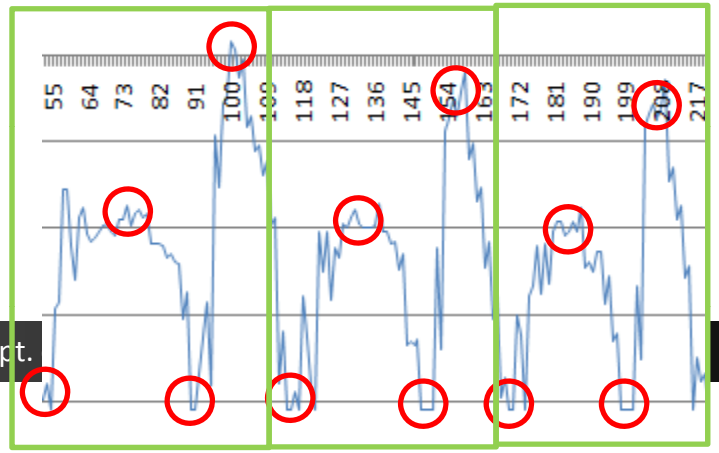
Right leg sensor, x-acc reading

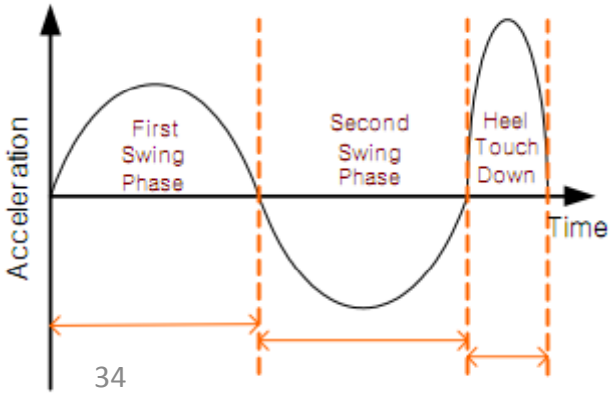
Analysis

Data: not all sensor readings are distinguishable, only certain sensor readings are needed



Detecting peaks and valleys for cycle recognition





Our Method

1. Perform a moving average to smooth the raw time series data
2. Estimate the probable range of the periodical interval R_{pi} according to the most frequent distance between peak points
3. Count stride number N_s and accumulate total stride time T_{total} by searching peak points in the sliding window with the range of size R_{pi}
4. Average Stride Time = T_{total} / N_s

Conclusion & Possible Improvements

- The estimated range of periodical interval may be hard to detect and become lost in the noise.
- A possible solution is to convert the raw time series into a different representation form to filter out the highest-frequency peaks before beginning the estimation process.
- It is crucial to automatically detect and focus on the most significant set of time series data in the sensor network with respect to a desired event of interest

BSN Contest Overall Conclusion

The First Body Sensor Network Contest
In Conjunction with BSN 2011

Second Place

is awarded to

National Taiwan University Team
National Taiwan University

On Behalf of the Organizing Committee
Roozbeh Jafari, John Lach

Conclusion

- 資料處理、探勘、預測在M2M中扮演核心的角色
- 學會運用智慧型的方法處理「大量、平行」的時間序列資料不是件簡單的事，卻是很重要的事。
  - 電機（訊號處理）以及資工（資料探勘，機器學習）的知識都需要
- 如果各位對如何M2M資料處理有興趣，歡迎跟我聯絡。