Κεφάλαιο 5

Μέθοδοι Πρόγνωσης και Πρόβλεψης

Ι. Βλαχάβας, καθηγητής
Τμήμα Πληροφορικής, ΑΠΘ

Forecasting and Prediction[1]

Πρόγνωση και Πρόβλεψη

- Prediction (πρόβλεψη) is a general term, which includes Forecasting.
  - Αναφέρεται στην πρόβλεψη μιας τιμής ή τον χαρακτηρισμό ενός γεγονότος που συμβαίνει τώρα
  - Π.χ. η αναγνώριση μιας τραπεζικής συναλλαγής ως νόμιμης ή όχι (fraud detection)
  - Βασίζεται σε δεδομένα του παρελθόντος, στατικά ή δυναμικά (χρονοσειρές)
- Forecasting (πρόγνωση) is the process of making predictions about the future based on past and present data
  - Αργότερα αυτές οι προβλέψεις συγκρίνονται με το τι πραγματικά συνέβη
  - Για παράδειγμα μια εταιρεία προβλέπει τα κέρδη της επόμενης χρονιάς και στο τέλος τα συγκρίνει με τις πραγματικές τιμές.
  - Βασίζεται σε δυναμικά δεδομένα που αναπαριστώνται ως χρονοσειρές (Timeseries)
- They are powerful tools for many kinds of decision making.
- Risk and uncertainty are central to forecasting and prediction
  - it is generally considered a good practice to indicate the degree of uncertainty attaching to forecasts.
  - In any case, the data must be up to date in order for the forecast/prediction to be as accurate as possible.

[1] Forecasting: Principles and Practice (3rd ed): Link
Rob J Hyndman and George Athanasopoulos, Monash University, Australia

- Ανάλογα με το πρόβλημα η ακρίβεια διαφοροποιείται σημαντικά
  - If the factors that relate to what is being forecast are known and well understood and there is a significant amount of data that can be used, it is likely the final value will be close to the forecast.
- Δεν υπάρχει μια μοναδική μέθοδος πρόγνωσης (forecasting method)
  - Selection of a method should be based on our objectives and conditions (data etc.).
- Ανάλογα με το είδος της προβλεπόμενης τιμής διακρίνονται 2 είδη:
  - Ταξινόμηση (classification) αφορά στην πρόβλεψη διακριτών τάξεων (κλάσεων/κατηγοριών, είτε ονομαστικών (nominal) ή βαθμωτών (ordinal), (π.χ. η πρόβλεψη για έγκριση ή όχι δανείου, η προβλεπτική συντήρηση).
  - Παρεμβολή (regression) αφορά στην πρόβλεψη συνεχών αριθμητικών τιμών (π.χ. πρόβλεψη/πρόγνωση ισοτιμίας νομισμάτων ή της τιμής μιας μετοχής ή της θερμοκρασίας).

Τύποι Δεδομένων

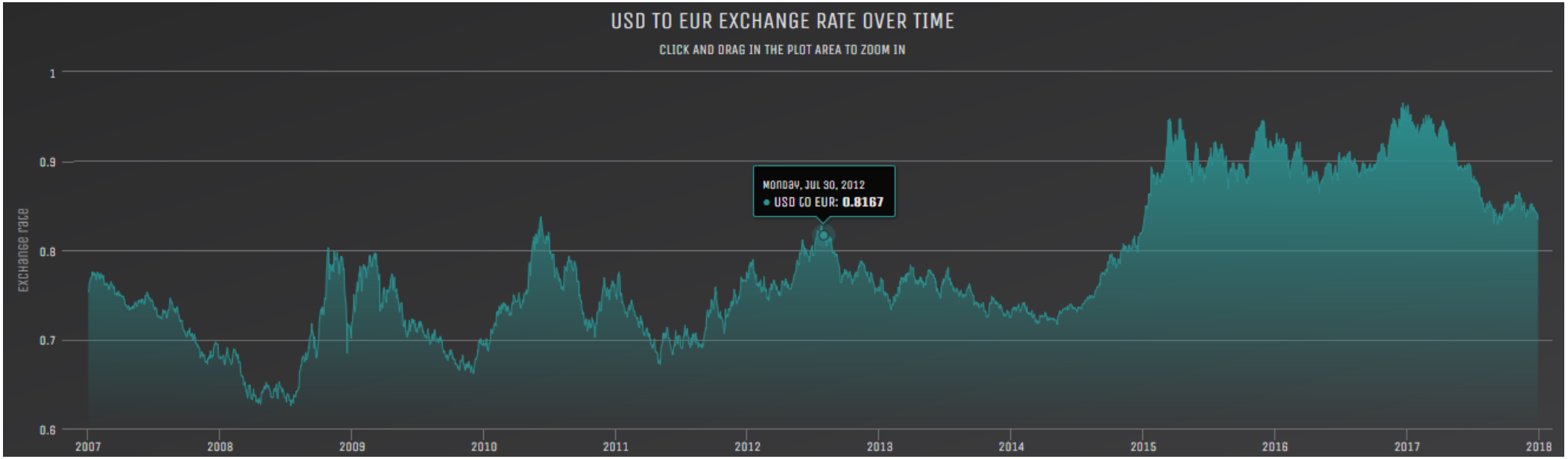
- Static data does not mention the time being recorded. It is a fixed data set.
  - π.χ. η πιστοληπτική ικανότητα πελατών τράπεζας, οι τιμές ακινήτων μιας περιοχής, η πιθανότητα εμφάνισης μιας ασθένειας, fraud detection, κλπ
  - Usually, these tasks are mentioned as Prediction (πρόβλεψη)
- Dynamic data may change after it is recorded and has to be continually updated.
  - π.χ. η τιμή μιας μετοχής, η ισοτιμία νομισμάτων, οι ημερήσιες πωλήσεις ενός αγαθού, κλπ
  - Usually, these tasks are mentioned as Forecasting (πρόγνωση) and the data are represented as Timeseries (χρονοσειρές)

Applications

- Forecasting has applications in a wide range of fields where estimates of future conditions are useful.
  - Supply chain management and customer demand planning
    - ✓ To ensure that the right product is at the right place at the right time.
    - ✓ Accurate forecasting will help retailers reduce excess inventory and thus increase profit margin.
  - Economic forecasting
  - Earthquake forecasting
  - Energy forecasting for renewable power integration
  - Land use forecasting
  - Player and team forecasting performance in sports
  - Product forecasting (its success)
  - Political forecasting
  - Sales forecasting
  - Weather forecasting, flood forecasting and meteorology

Χρονοσειρές (Time Series)

- Τι είναι χρονοσειρά;
  - Το σύνολο των δεδομένων, τα οποία συλλέγονται διαχρονικά και εκφράζουν την εξέλιξη των τιμών μιας μεταβλητής κατά τη διάρκεια ίσων διαδοχικών χρονικών περιόδων
  - Μια χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων μιας μεταβλητής, οι τιμές της οποίας είναι ιεραρχημένες με βάση τη χρονική περίοδο στην οποία αναφέρονται, π.χ. έτος, τρίμηνο, μήνας κ.α.
  - Παράδειγμα: μεταβολή ισοτιμίας USD/EUR τα τελευταία έτη



Ανάλυση Χρονοσειρών

- Χρήση παρελθοντικών δεδομένων για την πρόβλεψη μελλοντικών τιμών (Forecasting)
  - Υπόθεση: Όλοι οι παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή θα παραμείνουν σταθεροί
  - Παράδειγμα:

Έτος	2015	2016	2017	2018	2019	2020
Τιμή	10.5	12	10.1	11.5	9.6	??

- Ιδιότητες χρονοσειράς X_t
  - Τάση (trend): Μοτίβο ανοδικής/καθοδικής πορείας της χρονοσειράς σε συγκεκριμένο χρονικό διάστημα
  - Εποχικότητα (seasonality): Συνηθισμένα-κανονικά-προβλέψιμα ανεβοκατεβάσματα των τιμών σε συγκεκριμένα χρονικά διαστήματα **με σταθερή περιοδικότητα μεταξύ των εναλλαγών** (π.χ. θερμοκρασία μέσα στο έτος - εποχές).
    - ✓ Εξαρτώνται μόνο από το χρόνο
  - Κυκλικότητα (cyclicality): Ανεβοκατεβάσματα τιμών σε συγκεκριμένα χρονικά διαστήματα **χωρίς σταθερή περιοδικότητα μεταξύ των εναλλαγών** (π.χ. εναλλαγές μεταξύ υψηλών/χαμηλών τιμών αγοράς στο χρηματιστήριο).
    - ✓ Δεν εξαρτώνται (μόνο) από το χρόνο, αλλά και από άλλους παράγοντες
  - Τυχαιότητα/μη κανονικότητα (irregularity): Απρόβλεπτες αυξομειώσεις των τιμών της χρονοσειράς (π.χ. θόρυβος)

Ανάλυση Χρονοσειρών (συνεχ.)

- Συνήθως μια χρονοσειρά X_t περιγράφεται ως εξής:

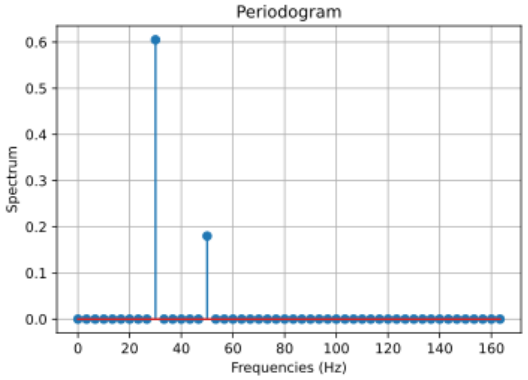
$$X_t = T_t + S_t + R_t$$

$$\text{όπου } T\text{: τάση}, S\text{: εποχικότητα}, R\text{: τυχαιότητα}, t\text{: χρονική στιγμή}$$

  - ✓ Αν ξέρουμε τις 3 αυτές συνιστώσες, μπορούμε να μοντελοποιήσουμε τη χρονοσειρά!

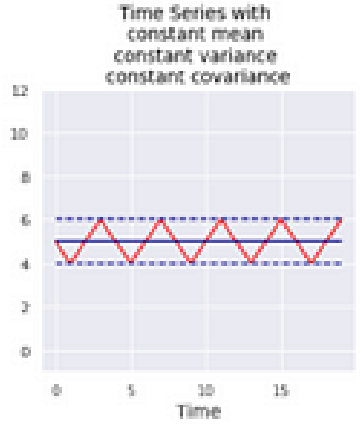
- Εύρεση *τάσης* χρονοσειράς:
  - *Μέθοδος διαφορών (differencing)*. Αφαίρεση προηγούμενης τιμής μεταβλητής από την επόμενη
  - Γραμμική/πολυωνυμική παρεμβολή (linear/nonlinear regression)

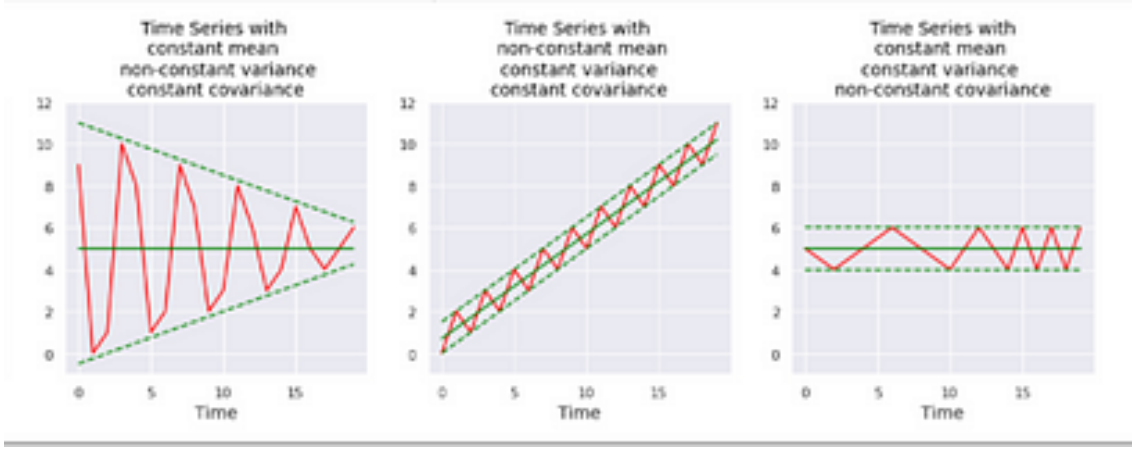
- Εύρεση *εποχικότητας* χρονοσειράς:
  - *Μέθοδος διαφορών*. Αφαίρεση των τιμών μιας μεταβλητής μεταξύ σταθερών χρονικών διαστημάτων
    - ✓ π.χ. αφαίρεση μέσης μηνιαίας θερμοκρασίας για τον κάθε μήνα μέσα στα έτη και παρατήρηση αν η διαφορά είναι κοντά στο 0
  - *Οπτικοποίηση Περιοδογράμματος (periodogram)*
    - ✓ *Εφαρμογή Διακριτού Μετασχηματισμού Fourier* (Discrete Fourier Transform - DFT) στη συνάρτηση αυτοσυνδιακύμανσης (Autocovariance Function - ACF) της χρονοσειράς
    - ✓ Μόνο αφού έχει αφαιρεθεί η τάση!



Ανάλυση Χρονοσειρών (συνεχ.)

- Μόλις βρεθεί η *τάση* και η *εποχικότητα*, μπορούν να αφαιρεθούν για την εκτίμηση της *τυχαιότητας* με στατιστικές μεθόδους
  - Η αφαίρεση αυτών των συνιστωσών μετατρέπει τη χρονοσειρά από μη-στάσιμη (non-stationary) σε στάσιμη (stationary)[2]
  - A time series has stationarity when its statistical properties will not change with time thus they will have constant mean, variance, and covariance.
    - ✓ A time series has stationarity when the observations are not dependent on the time.
  - Cyclic behavior and white noise in time series are stationary
    - ✓ The cyclic behavior of time series will be stationary because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.
  - Non Stationary time series will have nonconstant mean, variance, or covariance.






```
from statsmodels.tsa.stattools import adfuller
result = adfuller(series, autolag='AIC')
print(f'ADF Statistic: {result[0]}')
print(f'n_lags: {result[1]}')
print(f'p-value: {result[1]}')
```

[2] What is Stationarity in Time Series

```
for key, value in result[4].items():
    print('Critial Values:')
    print(f'   {key}, {value}')
```

p-value > 0.05: The data is non-stationary.

p-value <= 0.05: The data is stationary.

Μοντελοποίηση χρονοσειράς

- AutoRegressive (AR) model τάξης *p*
  - ✓ The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term)
  - ✓ Thus the model is in the form of a stochastic difference equation
  - ✓ [Μοντελοποίηση x_t ως γραμμικά εξαρτώμενη από όλες τις τιμές της μεταβλητής μέχρι **p** χρονικές στιγμές πίσω] **+** [θόρυβος [3](υπόθεση κανονικής κατανομής)]
  - ✓ $x_t = w_t + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} \quad (\varphi_i: \beta\acute{\alpha}\rho\eta, w_t: \theta\acute{o}\rho\upsilon\beta o\varsigma\ \tau\eta\ \chi\rho o\nu\iota\kappa\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta}\ t)$
- Moving Average (MA) model τάξης *q*
  - ✓ The moving-average model specifies that the output variable is cross-correlated with a non-identical to itself random-variable.
  - ✓ [Μοντελοποίηση τιμής x_t ως γραμμικά εξαρτώμενη από όλες τις τιμές θορύβου μέχρι **q** χρονικές στιγμές πίσω]
  - ✓ $x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} \quad (\theta_i: \beta\acute{\alpha}\rho\eta, w_t: \theta\acute{o}\rho\upsilon\beta o\varsigma\ \tau\eta\ \chi\rho o\nu\iota\kappa\acute{\eta}\ \sigma\tau\iota\gamma\mu\acute{\eta}\ t)$
- AutoRegressive Moving Average (ARMA) model
  - ✓ Given a time series of data x_t, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series.
    - The AR part involves regressing the variable on its own lagged (i.e., past) values.
    - The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past.

[3] white noise

- ✓ Συνδυασμός μοντέλων AR και MA
- ✓ Ένα μοντέλο ARMA(*p,q*) ορίζεται ως εξής:

$$x_t = w_t + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}$$

- ❑ AutoRegressive Integrated Moving Average (ARIMA) model
  - ✓ Γενίκευση του μοντέλου ARMA για μη-στάσιμες χρονοσειρές με τάση
  - ✓ Ένα μοντέλο ARIMA(*p,d,q*) επεξεργάζεται τη χρονοσειρά με τη μέθοδο διαφορών d φορές και τροφοδοτεί ένα μοντέλο ARMA(p,q)
  - ✓ Το αποτέλεσμα του μοντέλου ακολουθεί την αντίστροφή διαδικασία για να μεταπέσει στην αρχική μορφή της χρονοσειράς (πριν την επεξεργασία της)
- ❑ …και διάφορα άλλα μοντέλα, όπως SARIMA (seasonal-ARIMA) κα GARCH (κυρίως για χρηματοοικονομικά δεδομένα)

❖ Αξιολόγηση μοντέλων:

- ❑ Akaike Information Criterion (AIC)
  - ✓ Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.
- ❑ Bayesian Information Criterion (BIC)
  - ✓ Is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred.
  - ✓ It is closely related to the Akaike information criterion (AIC)
  - ✓ Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model

Η μηχανική μάθηση σε σχέση με τη στατιστική

- Η Μηχανική Μάθηση:
  - Δεν απαιτεί πρότερη γνώση για την πιθανή υποκείμενη σχέση μεταξύ των μεταβλητών.
  - Το μόνο που χρειάζεται είναι να εισαχθούν όλα τα δεδομένα στον αλγόριθμο ο οποίος τα επεξεργάζεται ανακαλύπτοντας μοντέλα/πρότυπα που μπορεί να χρησιμοποιηθούν για προβλέψεις σε μελλοντικά δεδομένα.
- Η στατιστική
  - Πρέπει να γνωρίζει επακριβώς τι αναζητείται και να γίνει η επιλογή των κατάλληλων παραμέτρων που θα βοηθήσουν σε αυτό.
  - Συνήθως εφαρμόζεται σε δεδομένα λίγων διαστάσεων

Γενική μοντελοποίηση πρόβλεψης χρονοσειρών

- Για την πρόβλεψη χρονοσειράς χρησιμοποιούνται κυρίως:
  - Ιστορικές τιμές της ίδιας χρονοσειράς
  - Γνωστές μελλοντικές τιμές άλλων χαρακτηριστικών που επηρεάζουν την χρονοσειρά.
- Παράδειγμα από την περιοχή πρόβλεψης πωλήσεων
  - *Sales(t+1)*: Η άγνωστη μεταβλητή στόχος που θέλουμε να προβλέψουμε
  - *Sales(t' ≤ t)*: Ιστορικές τιμές πωλήσεων, γνωστές τη χρονική στιγμή t
  - *Promotions(t+1)*: Μελλοντική πληροφορία που όμως είναι γνωστή τη χρονική στιγμή t
  - *F:* Το μοντέλο πρόβλεψης
  - Sales(t+1) = F(Sales (t' ≤ t) , Promotions(t+1))

Πρόβλεψη με μεθόδους Μηχανικής Μάθησης

- Οι ίδιοι αλγόριθμοι μπορούν να εφαρμοστούν και στα 2 είδη δεδομένων (static, dynamic)
  - Για την εφαρμογή των κλασσικών αλγορίθμων Μηχανικής Μάθησης σε δυναμικά δεδομένα (χρονοσειρές), θα πρέπει τα δεδομένα να μετατραπούν σε στατικά (μη χρονικά) διανύσματα.
  - Συνήθεις αλγόριθμοι: Linear Regression, Decision Tree, Random Forest, Gradient Boost, Support Vector Machine.
- Τα αναδρομικά νευρωνικά δίκτυα (RNN) εφαρμόζονται μόνο σε δεδομένα χρονοσειρών και δεν απαιτούν κάποιο είδος μετατροπής τους
  - Προηγμένα αναδρομικά νευρωνικά δίκτυα: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)

Πρόβλεψη Χρονοσειρών με Μηχανική Μάθηση

- Η Μηχανική Μάθηση εφαρμόζεται στην ανάλυση χρονοσειρών συνήθως για προβλήματα πρόβλεψης (forecasting), anomaly detection, predictive maintenance, κ.α.
- Τεχνικές εφαρμογής κλασσικής Μηχανικής Μάθησης σε χρονοσειρές
  - «Απαλοιφή» της έννοιας του χρόνου με κυλιόμενο παράθυρο, δηλ. μετατροπή της χρονοσειράς σε στατικό σύνολο δεδομένων με χαρακτηριστικά.
  - Δημιουργία μοντέλου για πρόβλεψη της χρονοσειράς (Y1, Y2, …, Y6).

Input lagged (i.e., prior) values	Output
Y1, Y2	Y3
Y2, Y3	Y4
Y3, Y4	Y5
Y4, Y5	Y6

    - ✓ Τα χαρακτηριστικά μπορεί να είναι χρονικά (π.χ. ημέρα, μήνας, κλπ.) ή περιγραφικά μιας περιόδου (π.χ. εάν υπήρχε έλλειψη προϊόντος τις προηγούμενες 7 μέρες)
    - ✓ Επιλογή χαρακτηριστικών
    - ✓ Εφαρμογή παραδοσιακών τεχνικών Μηχανικής Μάθησης στη συνέχεια για τους σκοπούς μας

Ομαδοποίηση

- Static Data
  - Οι περισσότεροι αλγόριθμοι ομαδοποίησης χρησιμοποιούν συνάρτηση απόστασης
    - ✓ k-Means με ευκλείδεια απόσταση ή απόσταση Manhattan
  - Ιεραρχικοί αλγόριθμοι ομαδοποίησης με ευκλείδεια απόσταση ή απόσταση Manhattan
  - Αλγόριθμοι ομαδοποίησης βασισμένοι στην πυκνότητα (dbscan)
- Timeseries data
  - Μπορούν να εφαρμοστούν οι γνωστοί αλγόριθμοι ομαδοποίησης
    - ✓ k-Means ή Ιεραρχικοί αλγόριθμοι σε συνδυασμό με DTW.
  - Οι κλασσικές μετρικές ομοιότητας δε λαμβάνουν υπόψιν τις χρονικές μετατοπίσεις (shifts) μεταξύ των χρονοσειρών
  - Μέθοδος υπολογισμού ομοιότητας: Dynamic Time Warping (DTW)
    - ✓ Υπολογισμός ομοιότητας μεταξύ χρονοσειρών
    - ✓ DTW has been applied to temporal sequences of video, audio, and graphics data
    - ✓ Any data that can be turned into a one-dimensional sequence can be analyzed with DTW
    - ✓ Κύριο χαρακτηριστικό: μέθοδος αναλλοίωτη ως προς τη μετατόπιση (shift invariant)

…the end

Questions?