Mining Time Series

Mining Massive Datasets

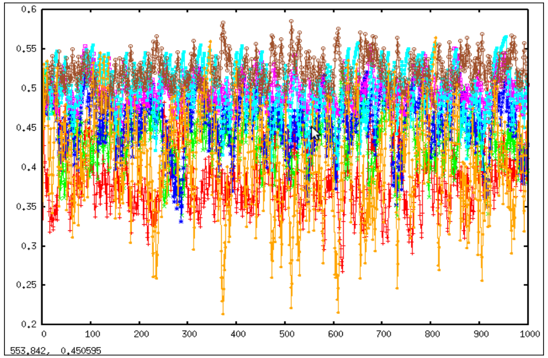
Prof. Carlos Castillo

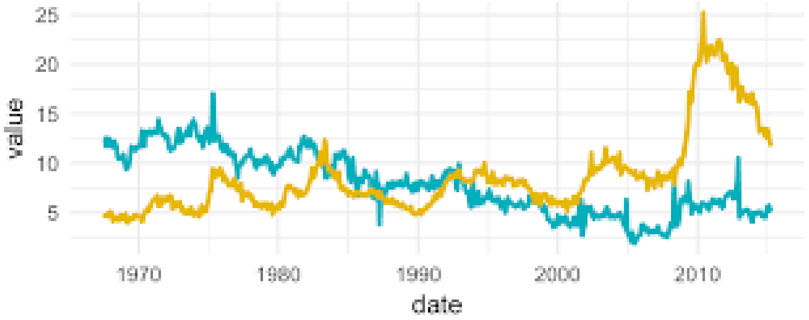
Topic 27

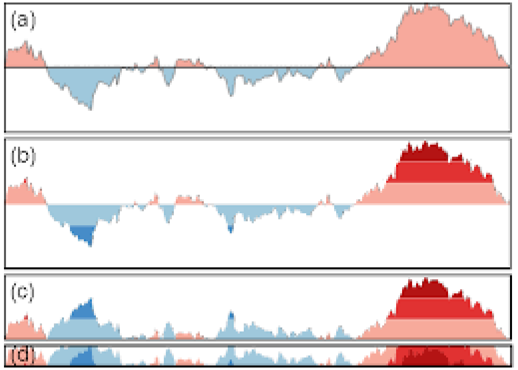
Universitat Pompeu Fabra Barcelona

Sources

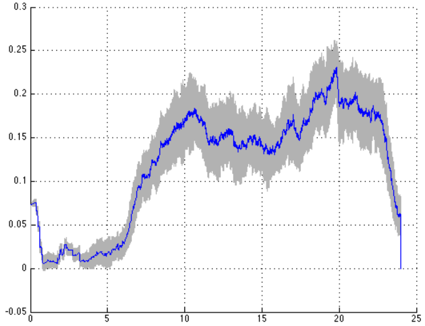
- Data Mining, The Textbook (2015) by Charu Aggarwal (chapter 14)
- Introduction to Time Series Mining (2006) tutorial by Keogh Eamonn [alt. link]
- Time Series Data Mining (2006) slides by Hung Son Nguyen

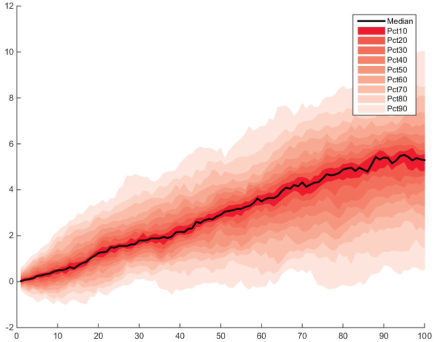


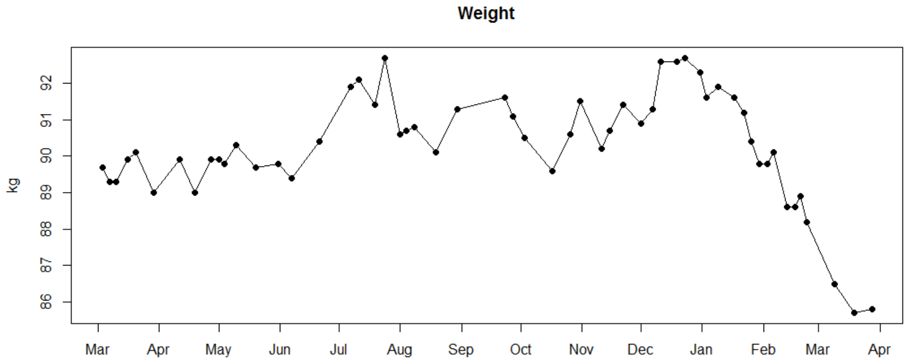




Why do we mine time series? Examples

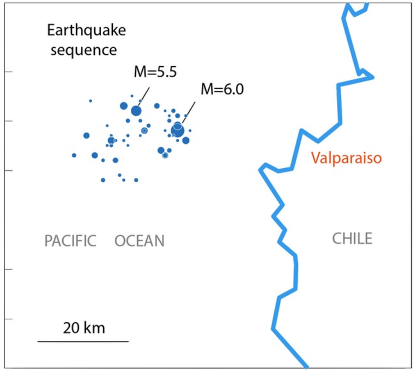


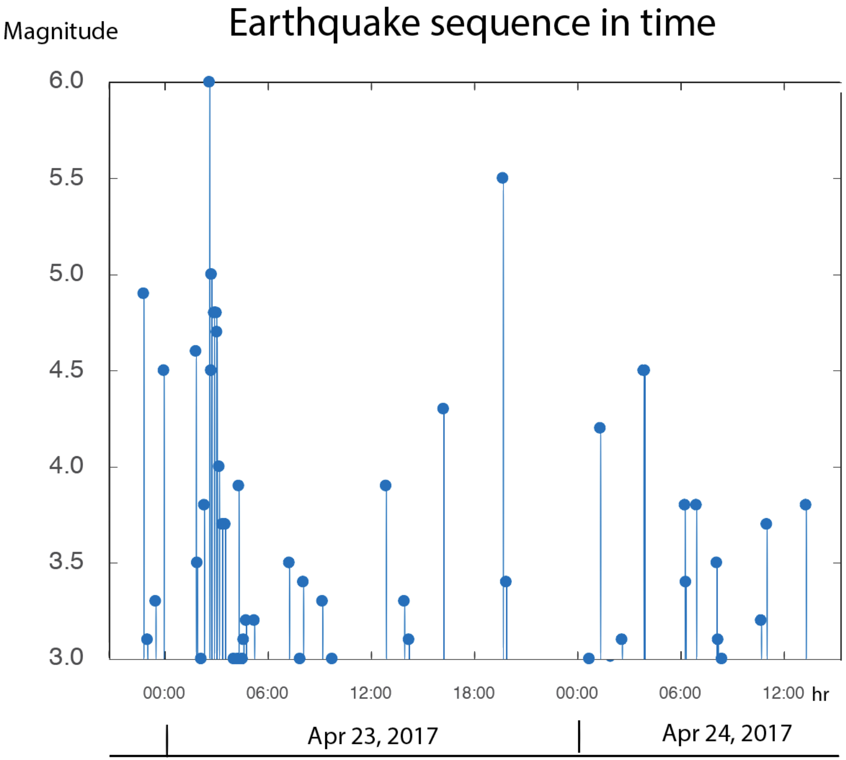




Seismic data

- Observations = earthquakes
- Goal: characterize when peeks occur



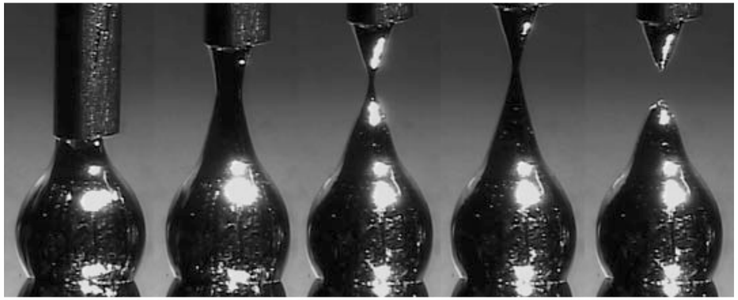


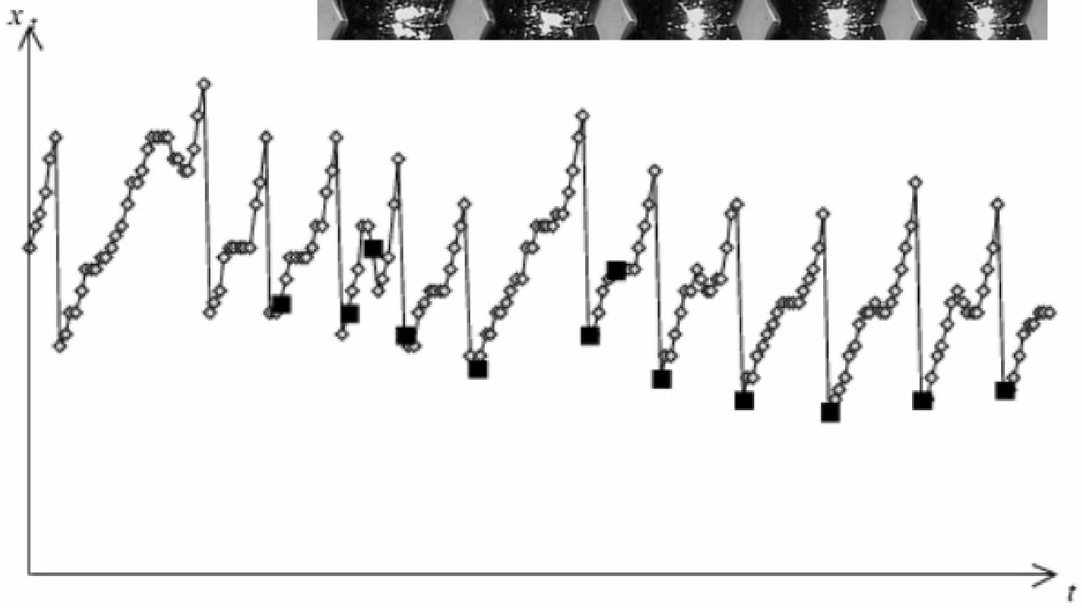
Liquid metal droplets

◇ = length of hot metal droplet

■ = droplet release (chaotic, noisy)

Goal: prediction of release





Stock prices

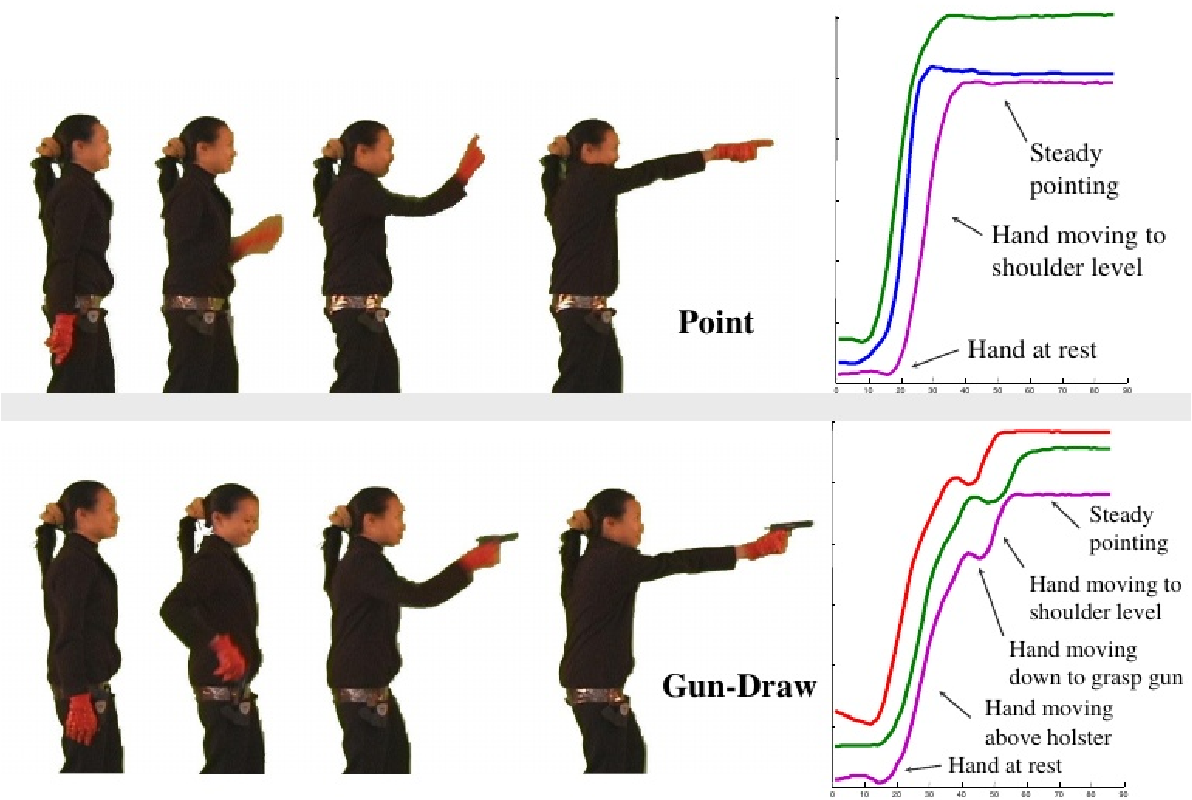
Price

Volume traded

Goal: find hidden patterns providing an advantage

Video data / gestures

- Series of angles of articulations in the body
- Temporal patterns can reveal gestures



Introduction to Time Series Mining (2006) tutorial by Keogh Eamonn [alt. link]

Applications

- Clustering
- Classification
- Motif discovery
- Event detection
- …

1) All require a reasonable definition of the **similarity** between two time series

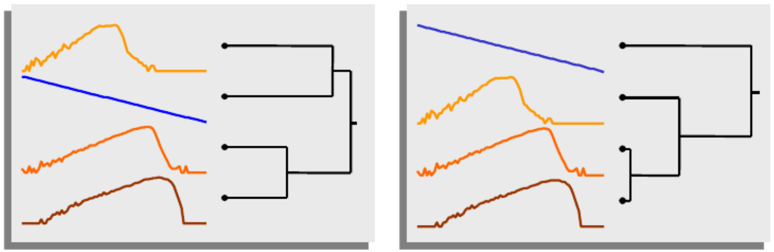
2) All can be done in **real-time** or **retrospectively**

Context vs Behavior

- **Contextual attribute(s)**
  - $x(i) = t_i$ = timestamp is the typical one
  - Sometimes other attributes providing context
- **Behavioral attribute(s)**
  - $y^j(i)$ = temperature, angle, price, sensor reading, …
    $j \in 1 \ldots d$

What are the difficulties?

- High sampling rate of many series over extended periods of time means …
  - Tons of data
  - Things are bound to fail at several points (missing data, noisy data)
- Subjectivity



Preparing a time series

Notation: multivariate time series

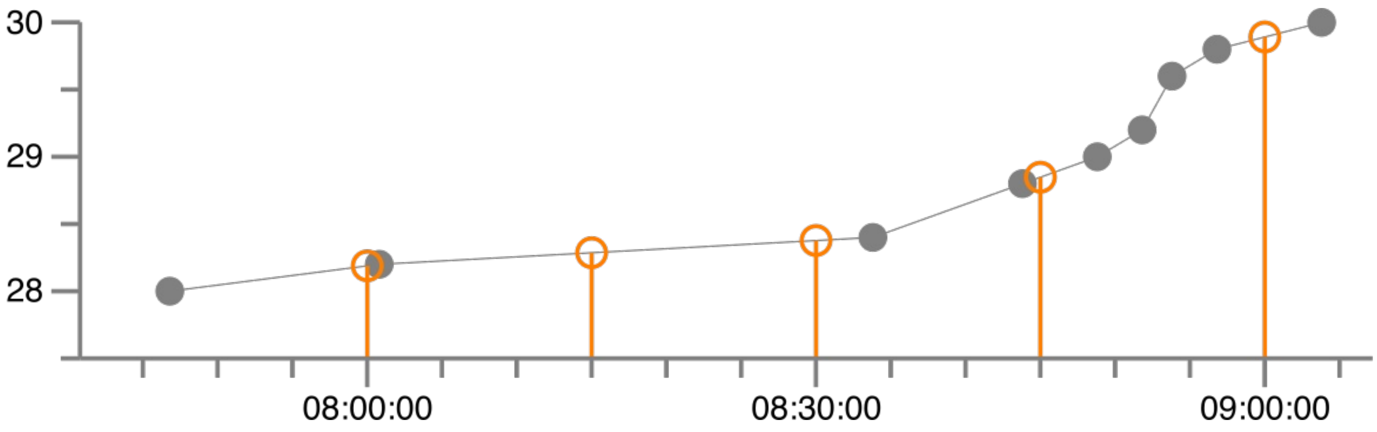
- Length n, timestamps $t_1, t_2, \ldots, t_n$
- Values at time t_i : $(y_i^1, y_i^2, \ldots, y_i^d)$
- If series is univariate we drop the superscript

Missing values: linear interpolation

- Let $t_i < t_x < t_j$

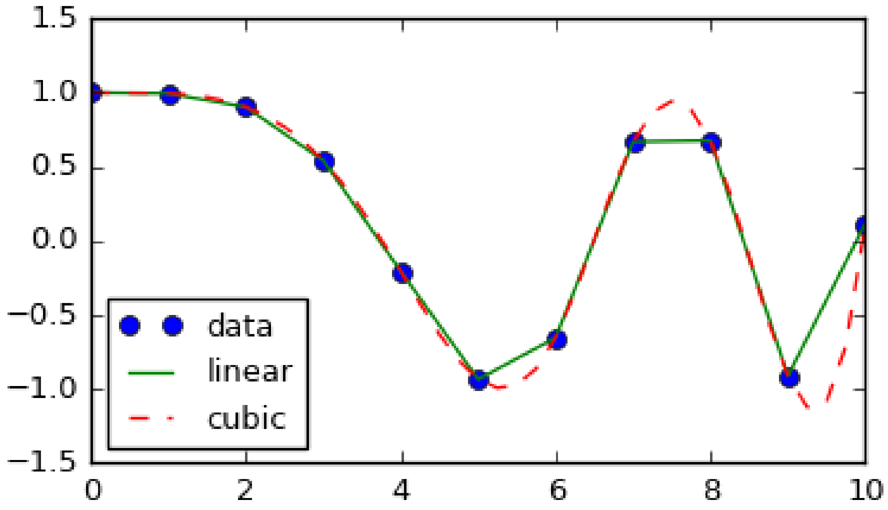
$$y_x = y_i + \left(\frac{t_x - t_i}{t_j - t_i}\right) \cdot (y_j - y_i)$$

- Example: make an irregular series regular



Missing values: splines

Cubic polynomials between y_i, y_{i+1} that have the same slope at those points as the original curve.

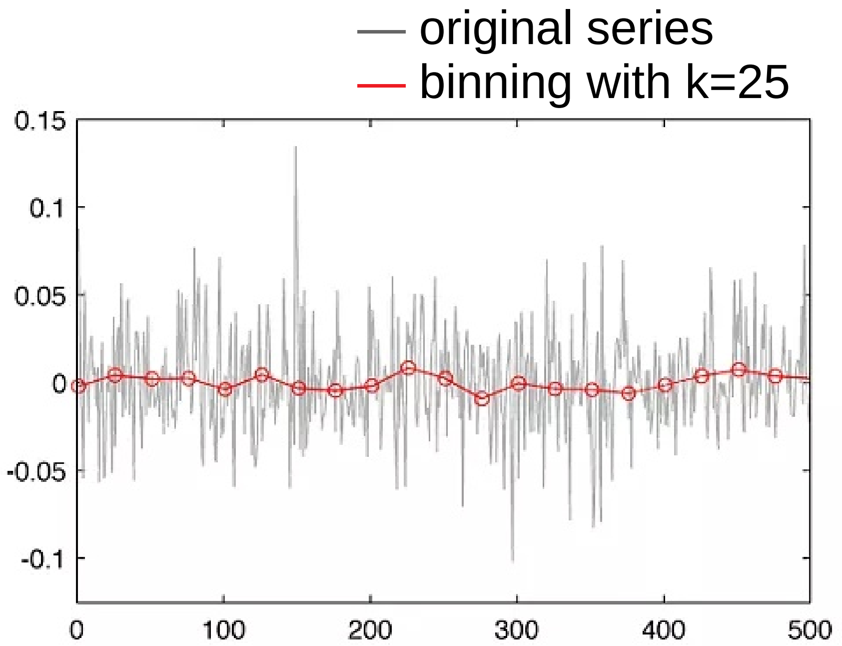


http://lagrange.univ-lyon1.fr/docs/scipy/0.17.1/tutorial/interpolate.html

Noise removal: binning

- Replace series by average of values in bins (subsequences) of length k

$$y'_{i+1} = \frac{1}{k} \sum_{r=1}^{k} y_{i \cdot k + r}$$



http://www.quantatrisk.com/2013/03/22/rebinning-of-financial-time-series/

Noise removal: moving average smoothing

- Equivalent to overlapping bins

$$y'_i = \frac{1}{k}\sum_{r=1}^{k} y_{i-r+1}$$

- Larger k leads to smoother series, but losses more information
- Use smaller k for first k-1 items

https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/sma

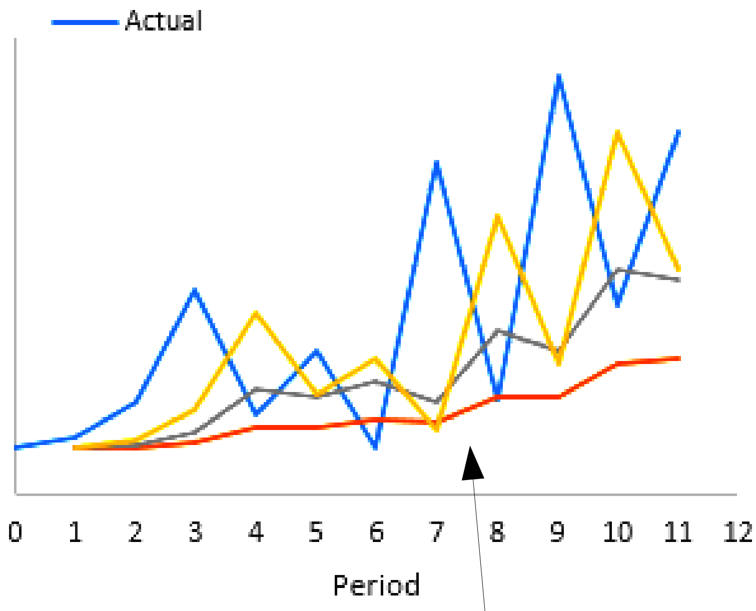
Noise removal: exponential smoothing

- Combine previously smoothed point with current point

$$y'_i = \alpha \cdot y_i + (1-\alpha) \cdot y'_{i-1}$$

- Recursively substituting

$$y'_i = (1-\alpha)^i \cdot y'_0 + \alpha \sum_{j=1}^{i} y_j \cdot (1-\alpha)^{i-j}$$



Which y' has the larger alpha?

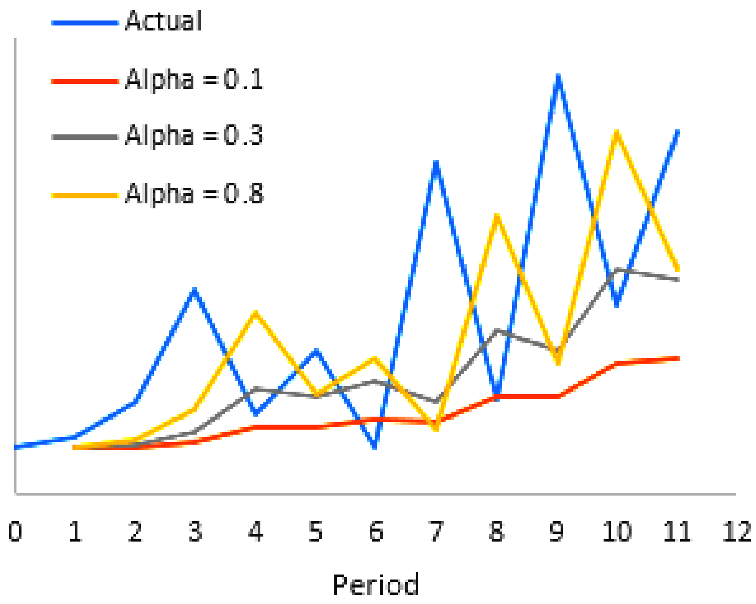
Noise removal:
exponential smoothing

- Combine previously smoothed point with current point

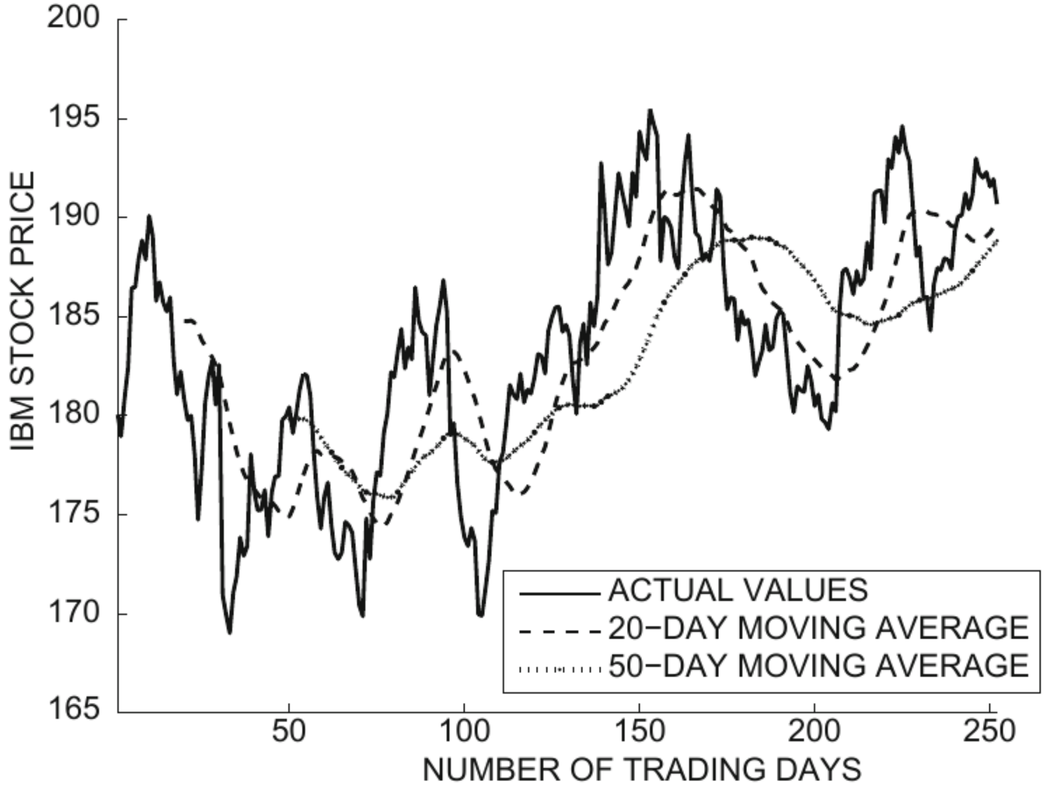
$$y'_i = \alpha \cdot y_i + (1-\alpha) \cdot y'_{i-1}$$

- Recursively substituting

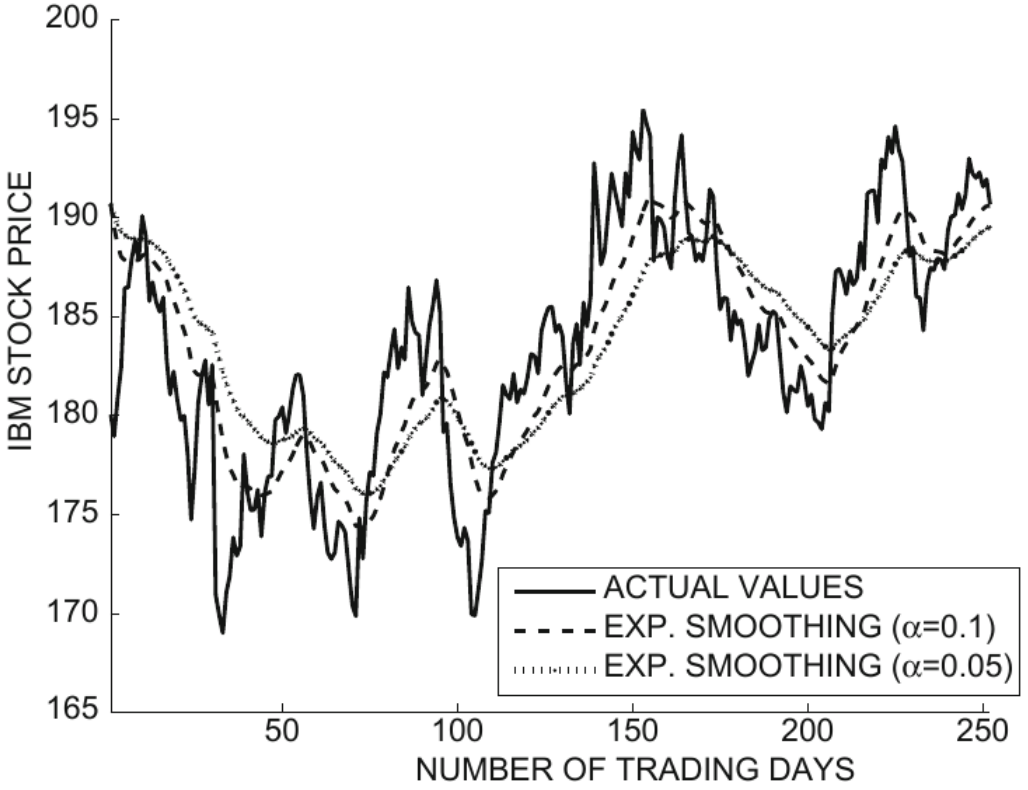
$$y'_i = (1-\alpha)^i \cdot y'_0 + \alpha \sum_{j=1}^{i} y_j \cdot (1-\alpha)^{i-j}$$



Moving average vs exponential smoothing



(a) Moving average smoothing



(b) Exponential smoothing

Exercise

Answer in
Google Spreadsheet

- Given the following series:

t	1	2	3	4	5	6	7	8	9	10
y(t)	2	4	12	2	1	-2	0	15	3	3
1. y'(t)										
2. y'(t)										

- 1. Moving average with k=3
- 2. Exponential average with alpha=0.5

Summary

Things to remember

- Series preparation
  - Interpolation
  - Smoothing

Exercises for TT27-TT29

- Data Mining, The Textbook (2015) by Charu Aggarwal
  - Exercises 14.10 → 1-6