Research Methods

EDA for Time Series & Experiment Design

Dr. Sven Magg, Prof. Dr. Stefan Wermter

Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. EDA with time series data
   - visualising time series
   - correlations
   - trend & periodic features
   - cross- & autocorrelation

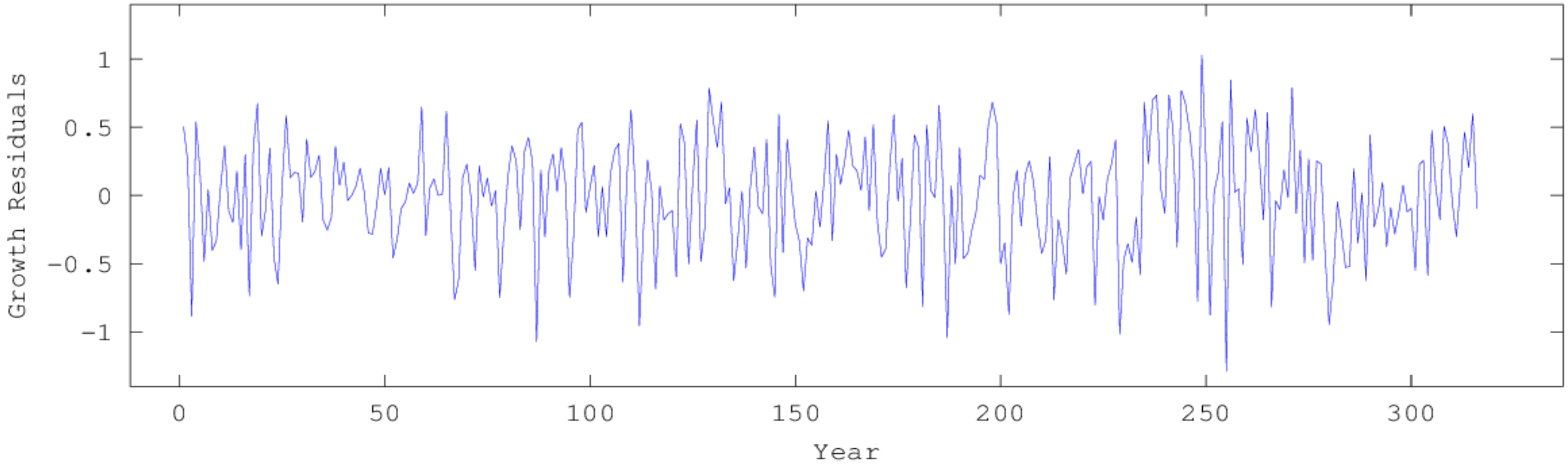
- Experiment Design
   - The first steps

Time Series Data

- Values recorded over time
  - values separated by constant time interval, or
  - data points are pairs of value and time of recording

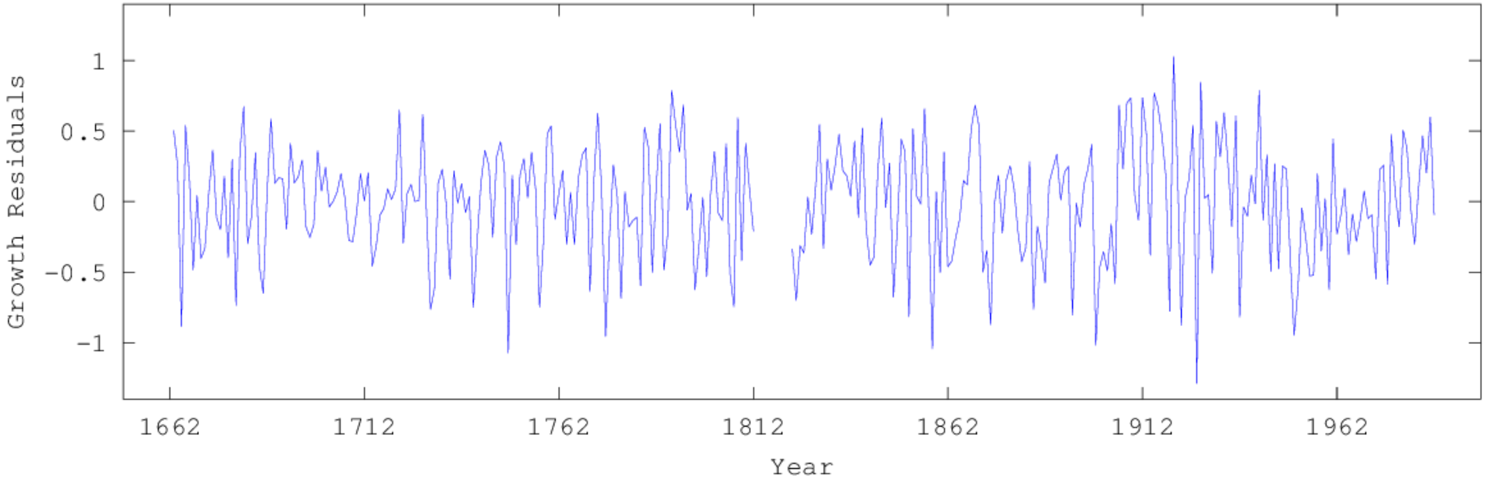
⇒ **Time series data is 2-dimensional!**

- Time series of 325 years of tree growth:



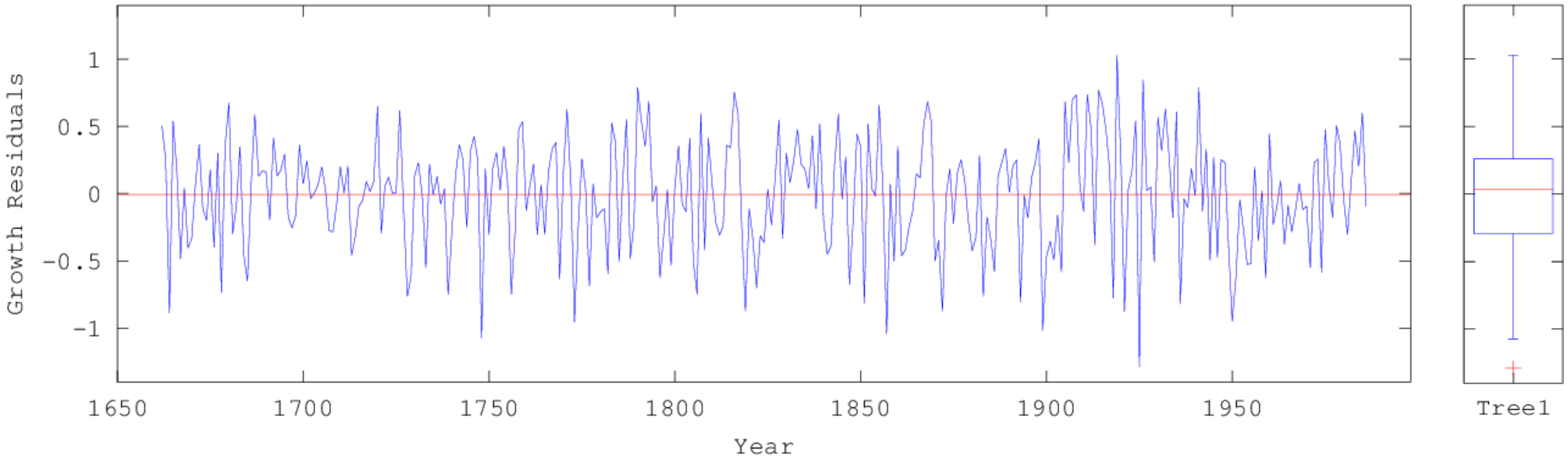
Visualising Time Series

- First question: Do we have gaps?
  - If data is one vector, how can we be sure?
  - Always record the time as well and plot both!



- Plotting over both helps to avoid errors and makes interpretation easier

Visualising Time Series



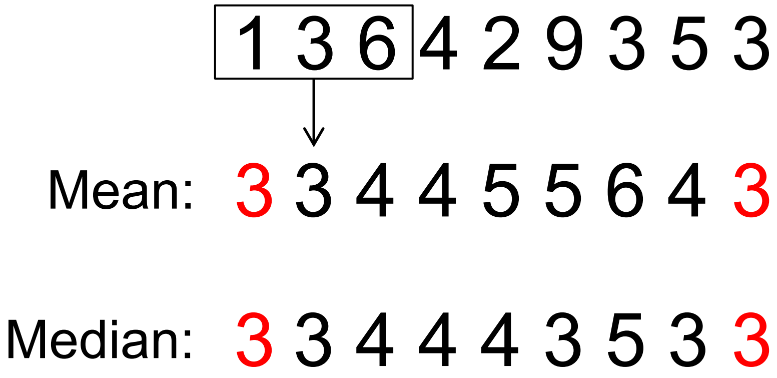
- Time series often have high variation
- Difficult to spot general trends or areas of interest

- What is of interest?
  - Trend, periodic events, areas with a consistent deviation of the average over a period of time

Smoothing

- Use neighbourhood information
- Replace value with average of neighbourhood
  - Different averages can be used: mean, median, hanning,…
  - Size of neighbourhood: Window size
  - Beginning and end of the series are handled separately

- Mean/Median smoothing
  - n-smooth (window size = n)
  - $x_i = f(x_{i-\lfloor \frac{n}{2} \rfloor}, \cdots, x_i, \cdots, x_{i+\lfloor \frac{n}{2} \rfloor})$
  - f either $mean()$ or $median()$



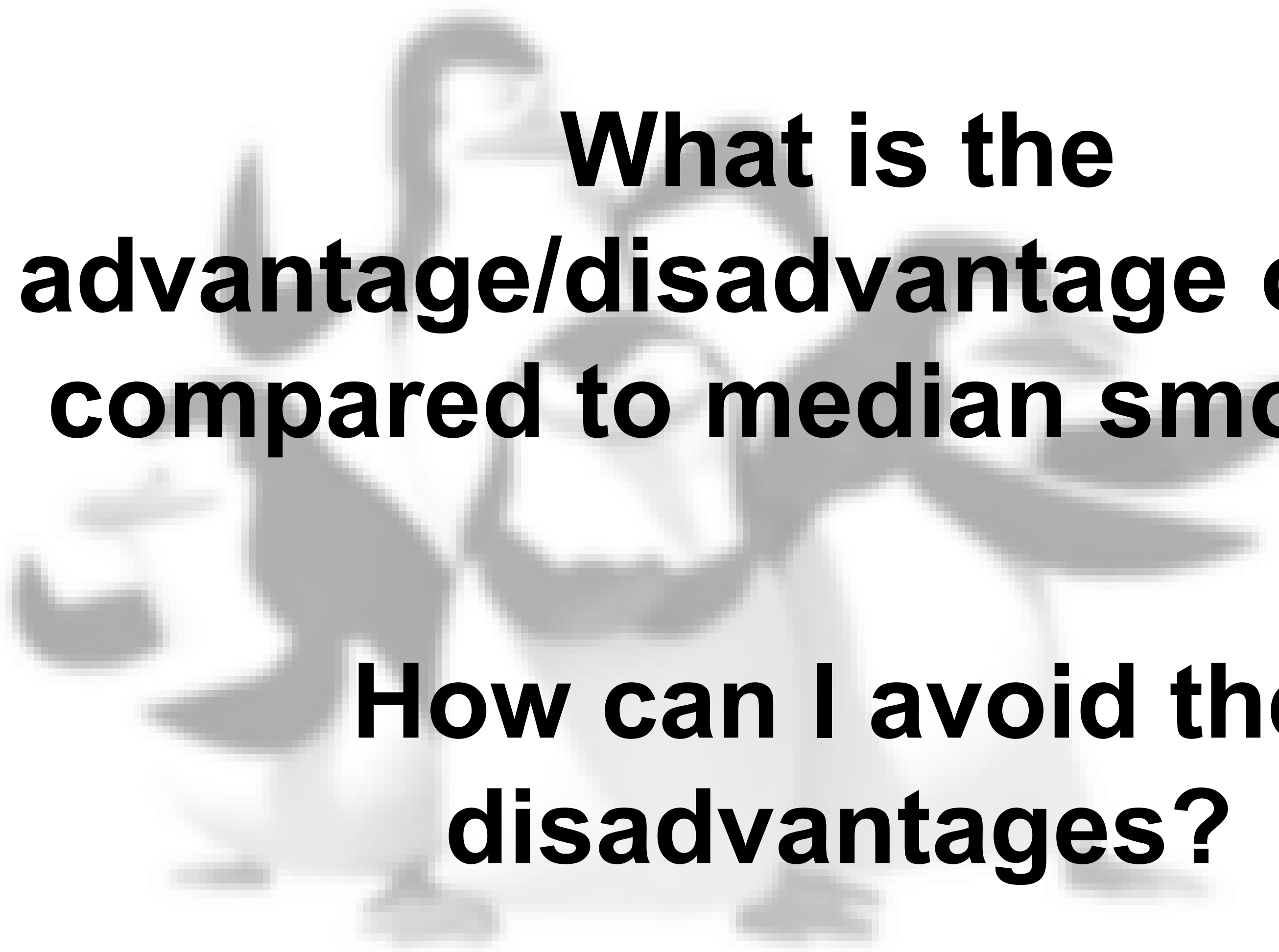
$\boxed{1\ 3\ 6}$ 4 2 9 3 5 3

Mean: 3 3 4 4 5 5 6 4 3

Median: 3 3 4 4 4 3 5 3 3

Group Task!

3 5



What is the advantage/disadvantage of mean compared to median smoothing

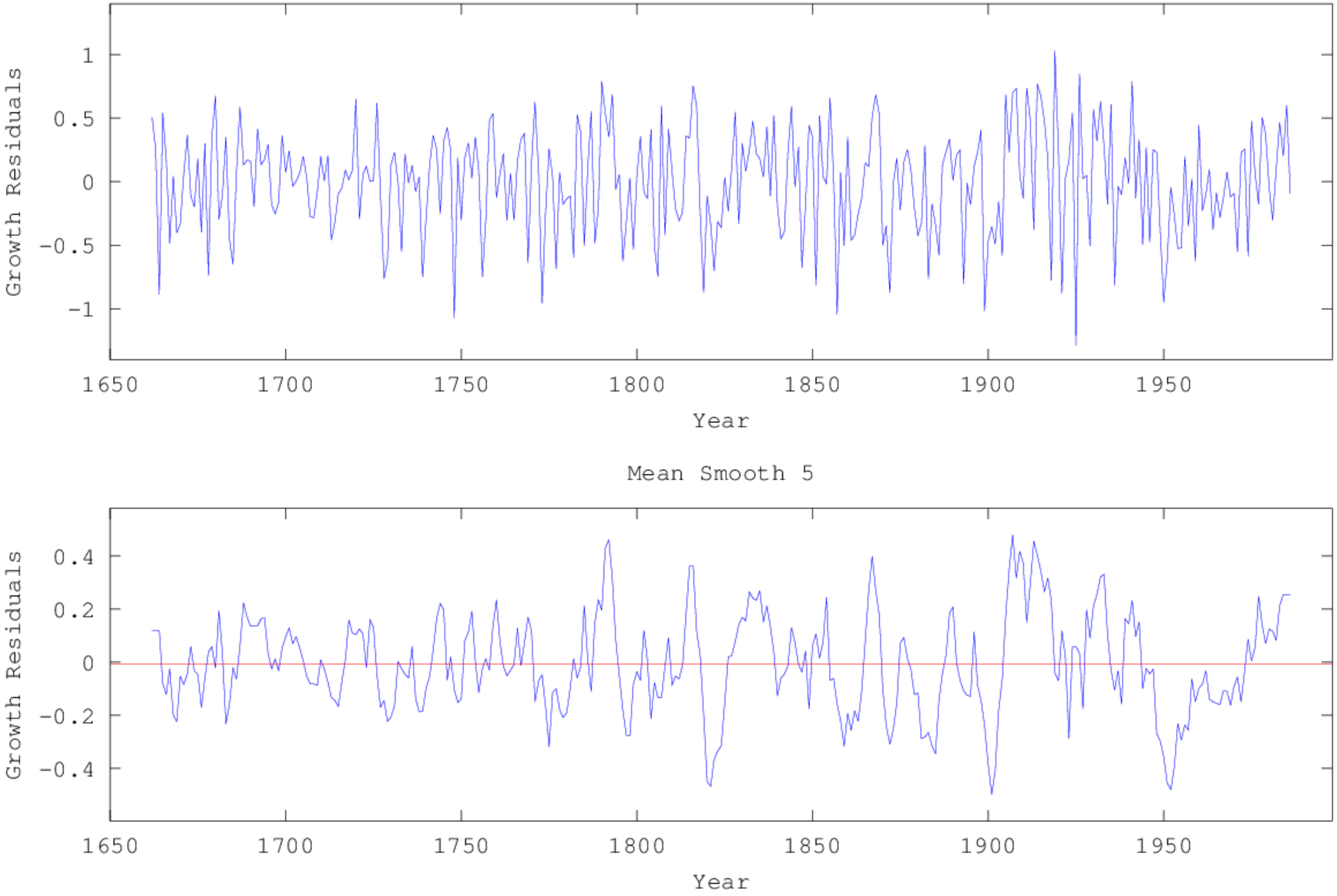
How can I avoid the disadvantages?

Smoothing

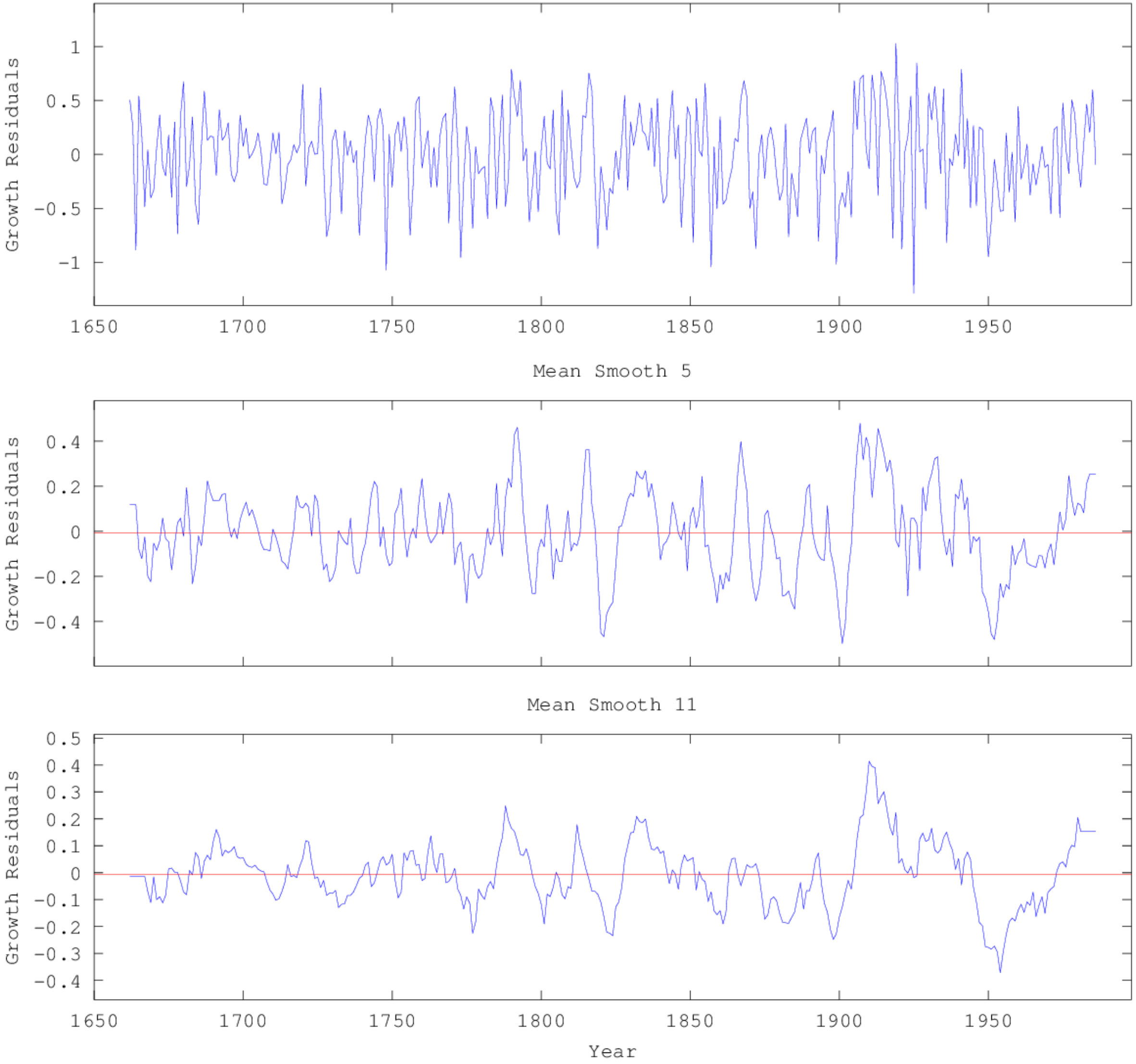
- Mean smoothing
  - sensitive to outliers, which affect nearby values
  - creates “smoother” graphs than median smoothing
- Median smoothing
  - ignores single outliers
  - produces *mesas* (areas with same values)

- *Mesas* can be handled by re-smoothing with mean smoothing
- Sequences can be smoothed several times with different smoothing techniques

Smoothing Example



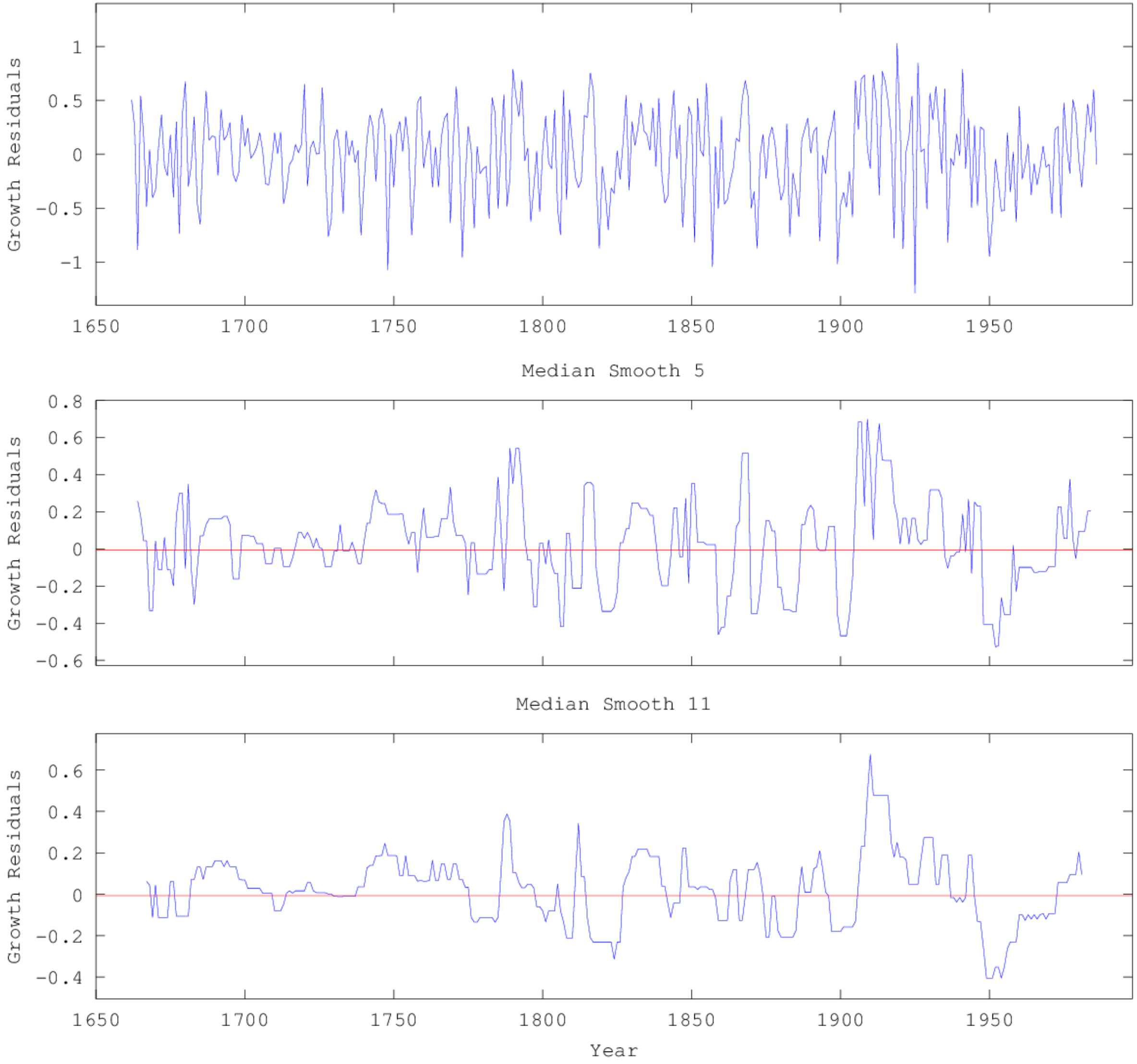
Smoothing Example



Bigger window = more general features become visible

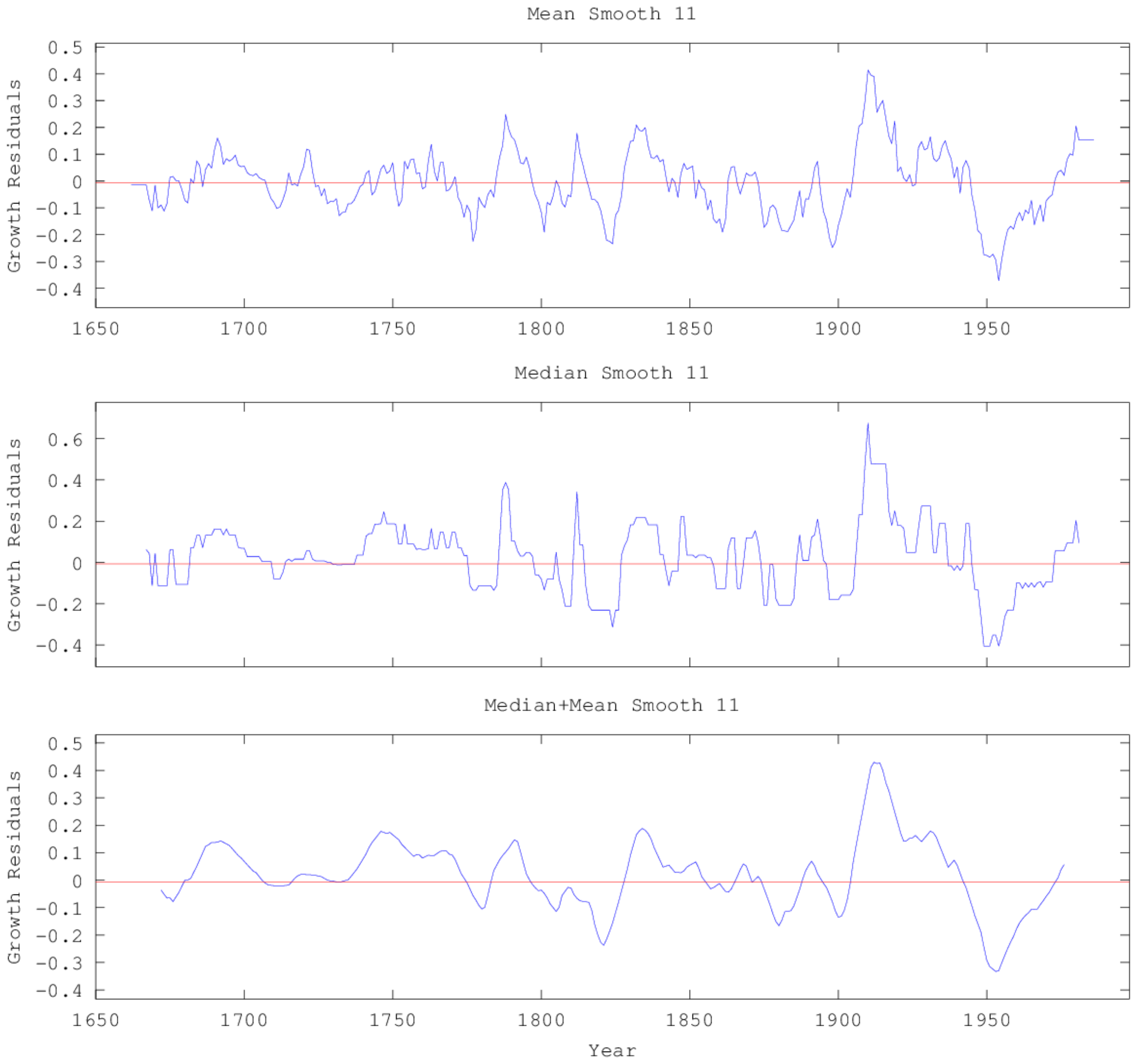
but:
more local features are lost

Smoothing Example



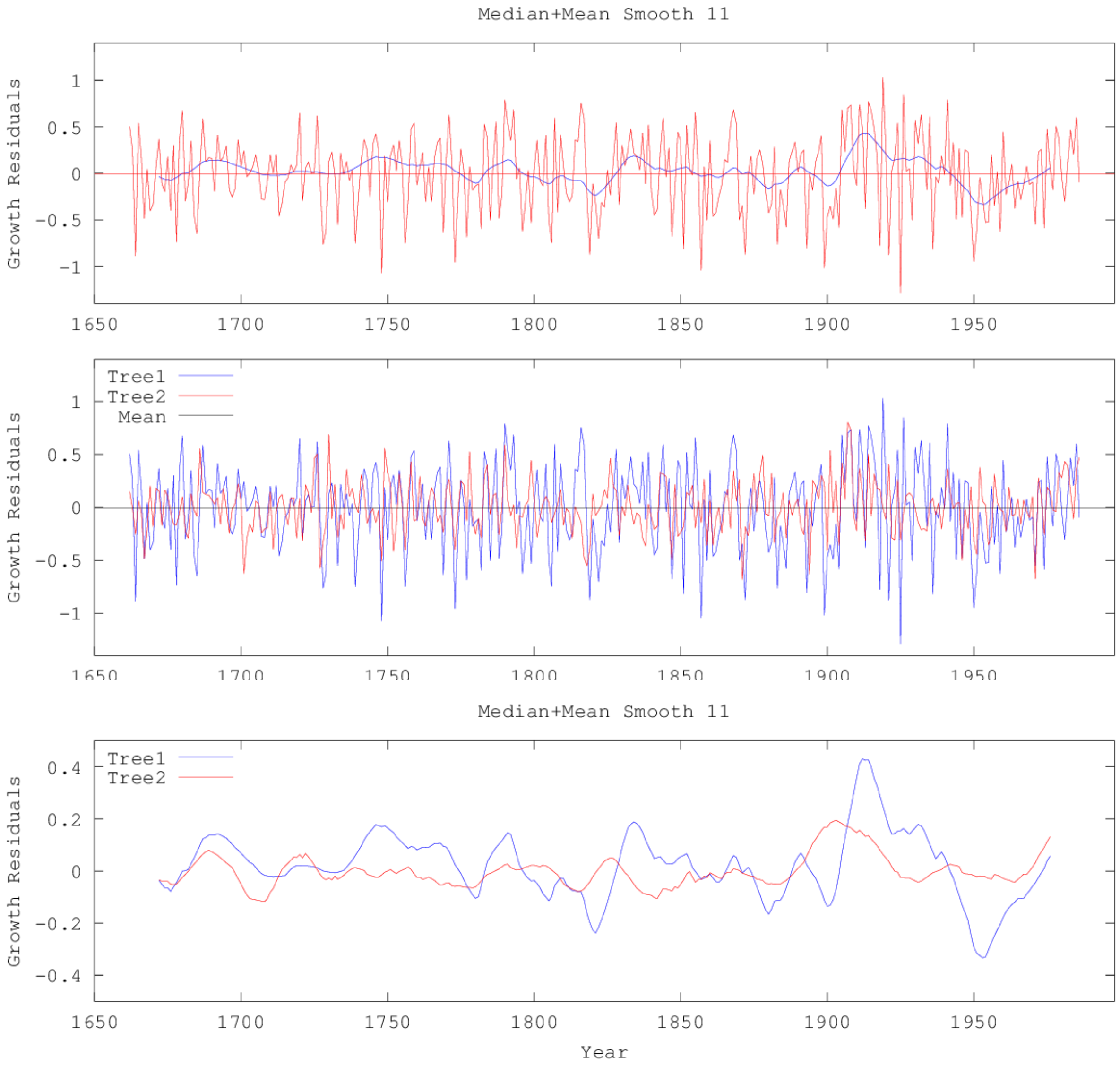
Bigger window = mesas become larger

Smoothing Example



Re-smoothing with mean “cleans” up the graph

Smoothing Example



Noisy time series difficult to compare!

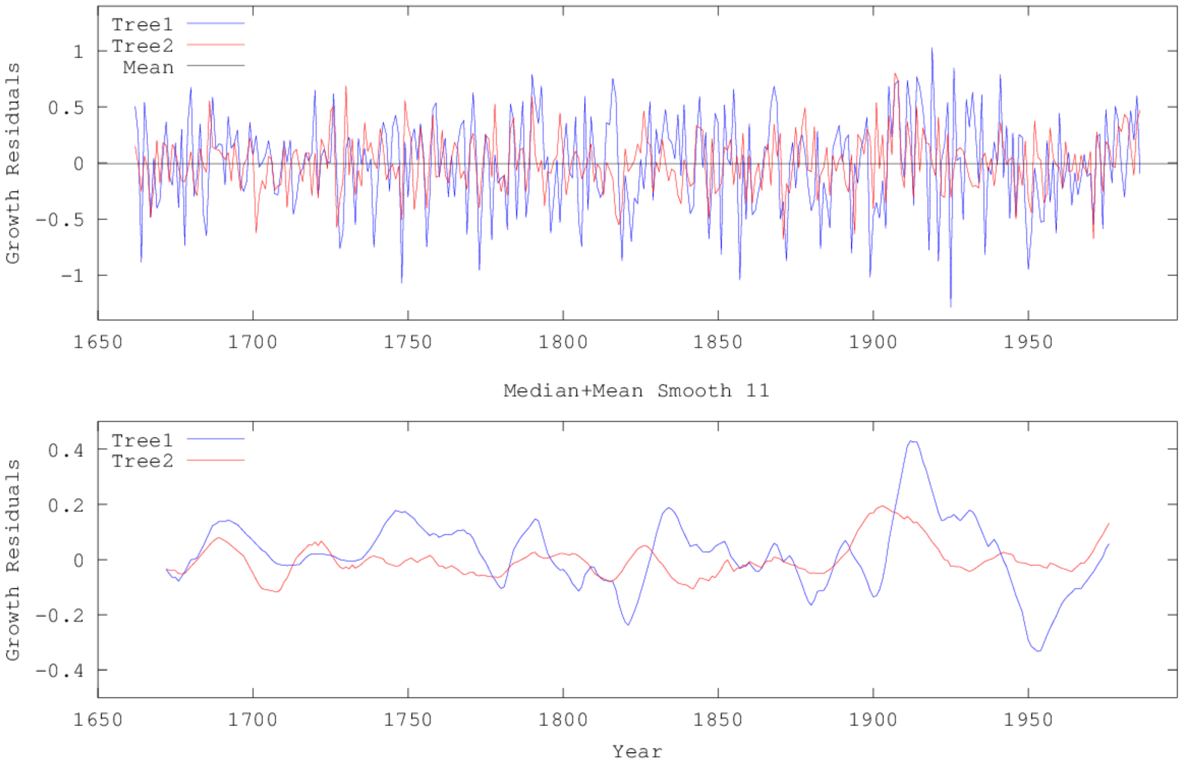
Smoothing enables us to compare noisy time series visually!

Statistics for Time Series

- For one time series:
  - Correlation between value and time:
    Positive (or negative) linear trend
  - Correlation between different time steps with a given lag
    = evidence of recurring or periodic events

- What do we want to see when dealing with two or more time series?
  - Cross-Correlation (Correlation at each time step)
  - Correlation with lag $\Rightarrow$ One series is indicator or predictor of the other

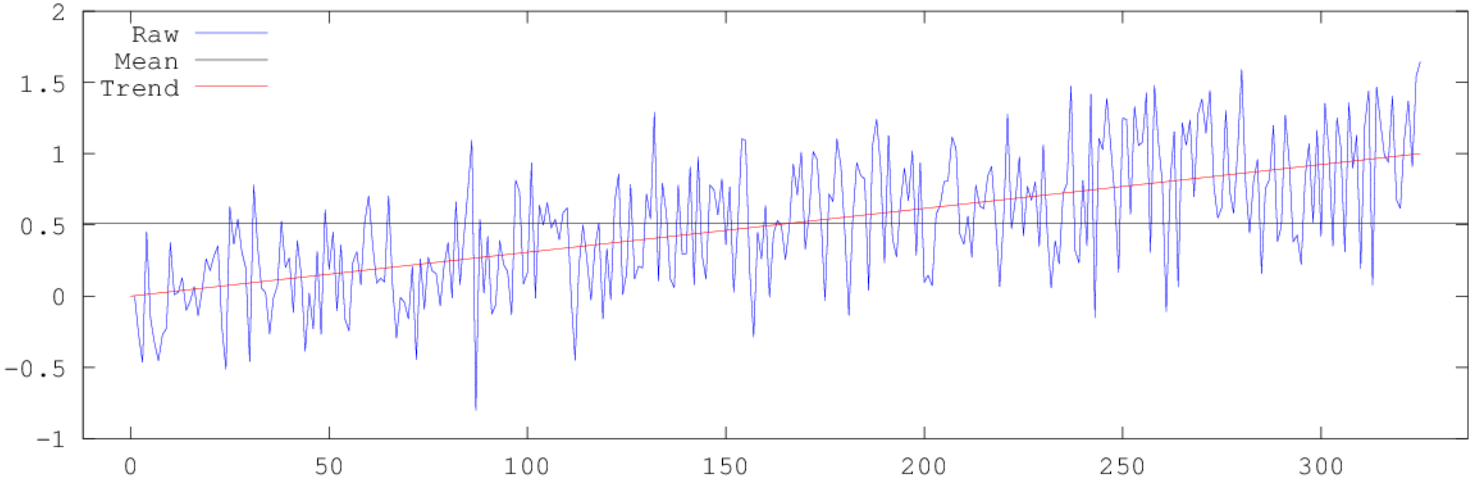
Correlations Between Time Series



- There seems to be some correlation
- Pearson’s correlation coefficient: 0.33
- Not a perfect correlation, but evidence

- Remember: Pearson’s coefficient measures linear correlation!
- Correlation means there is evidence that one influences the other, or that both are affected by a set of other factors!

Trend

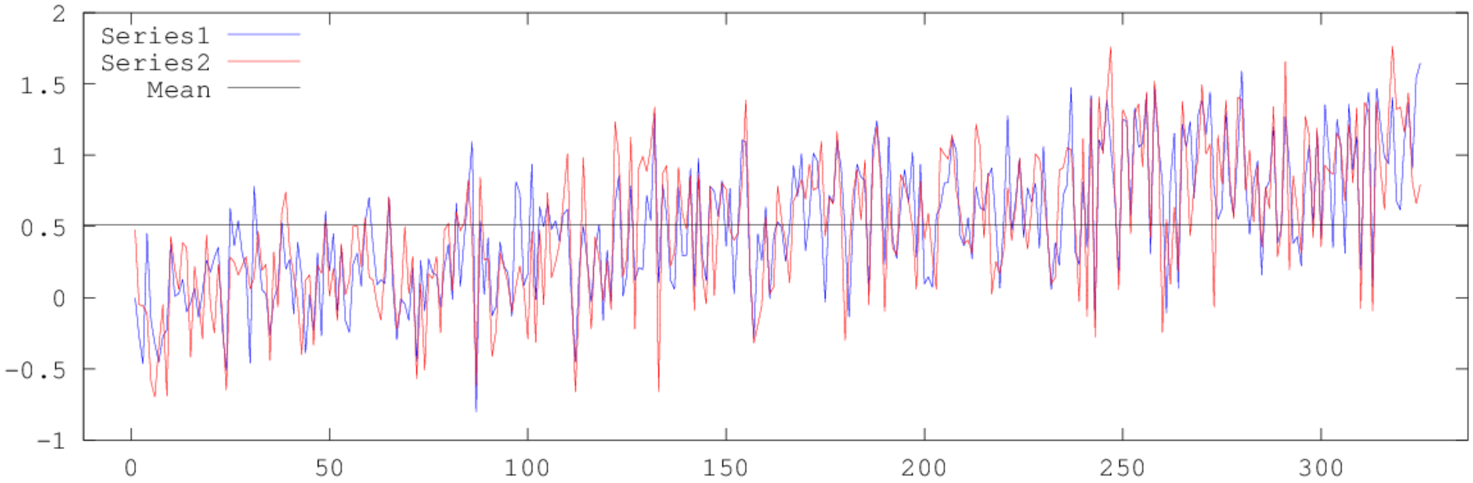


- Positive correlation between values and time
- Pearson's correlation coefficient for x and y: 0.64

⇒ positive linear trend

- But: Trend does not have to be linear!

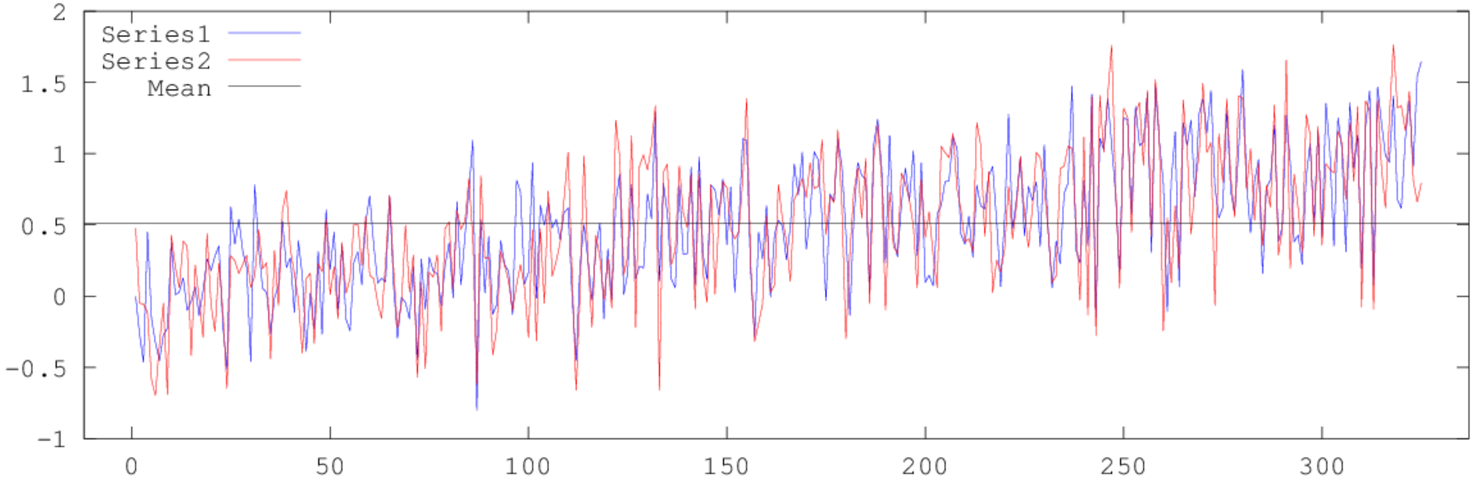
Correlations Between Time Series



- Two comparable time series:
  - Both individually have positive trend: Coefficients: 0.64 and 0.59
- Do they correlate?
  - Coefficient Series1 vs. Series2: 0.79

Group Task!

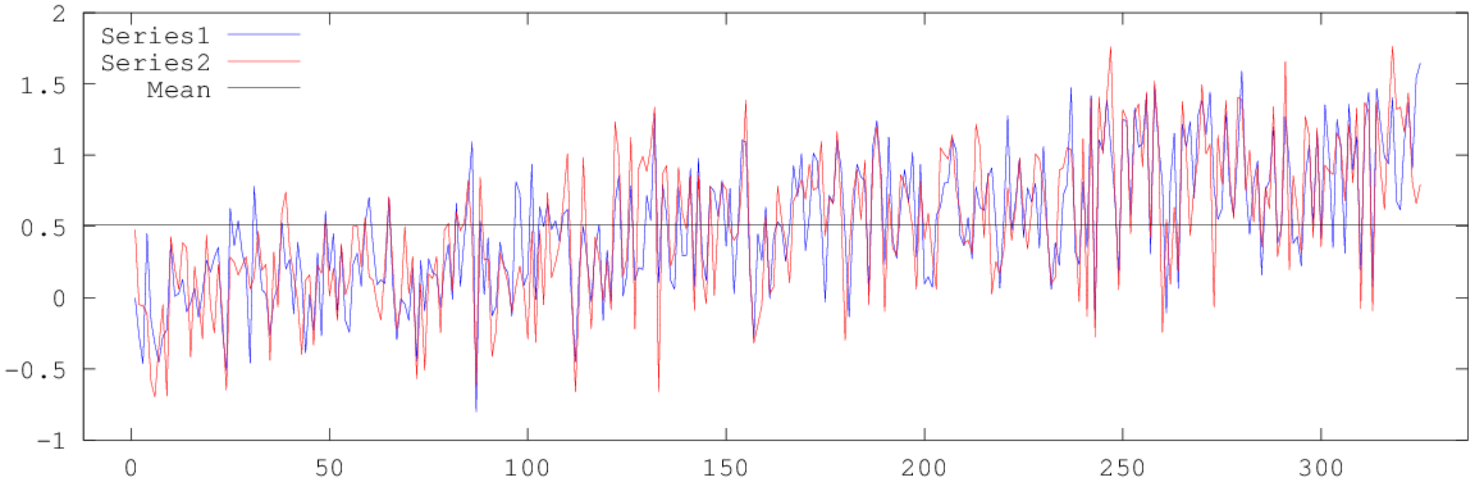
4 5



Both series correlate (0.79)

Any doubts or questions?

Correlations Between Time Series



- Two comparable time series:
  - Both individually have positive trend: Coefficients: 0.64 and 0.59
- Do they correlate?
  - Coefficient Series1 vs. Series2: 0.79
- Trend obscures correlation! Coefficient without trend: 0.67

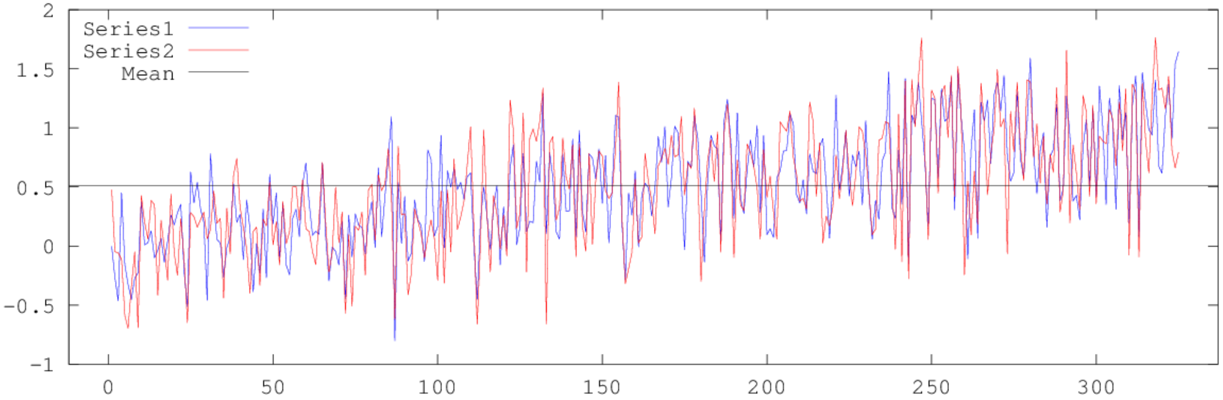
De-Trending

- To see the features that are superimposed on the trend more clearly, we have to remove the trend!
- De-Trending in general: Once we have a fit for our data, remove this fit from our time series

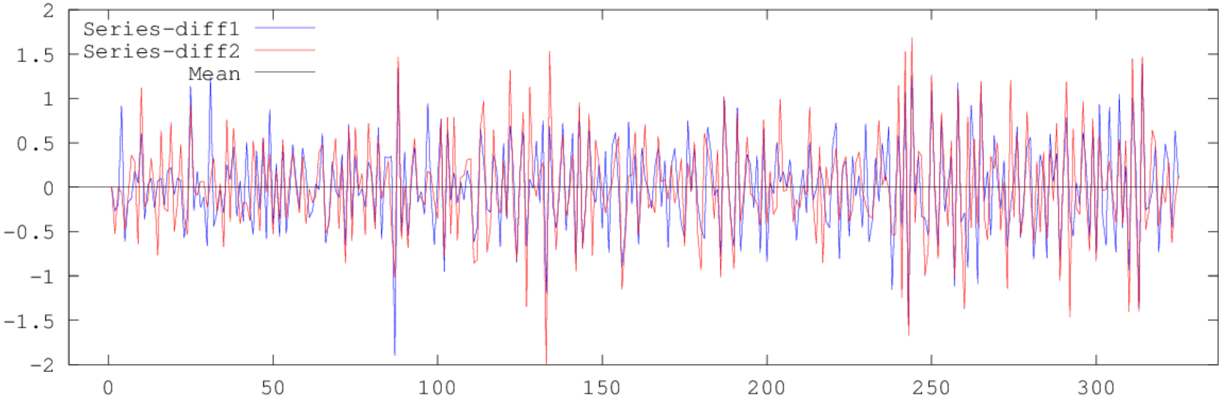
- Simple version: Differencing $x_{i\_diff} = x_i - x_{i-1}$
  - Discrete version of subtracting the first derivative (=trend)
  - If the trend is of higher order: successive differencing

- Other common forms of de-trending
  - Removing a linear (or polynomial) fit

De-trending



- Correlation: 0.79



- Correlation: 0.7

- Before I said that the correlation coefficient of the data (before adding an artificial trend) was 0.67….
- Did we do something wrong?

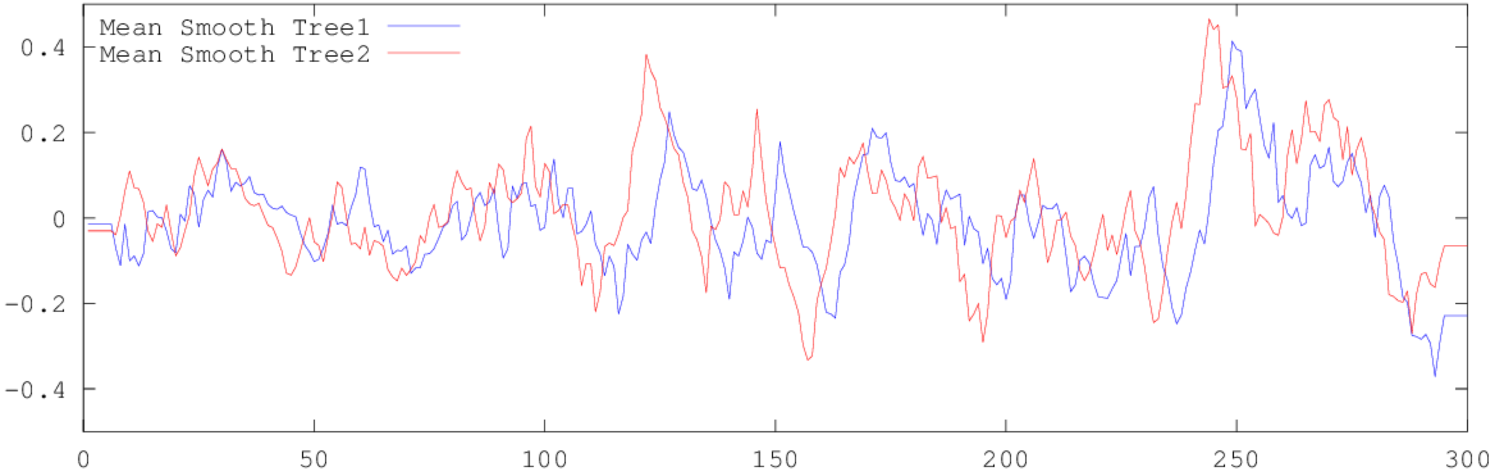
Correlations with Lag

- Often events that correlate at different time steps are interesting:
  - Events that predict another event
  - Periodically occurring features
- How can we capture that statistically?

- Cross-Correlation between two series a and b
  - Calculate the correlations $corr(a_i, b_{i+lag})$ for
    $$lag = -n, \cdots, -1,0,1, \cdots n$$
  - Result is a vector of correlations

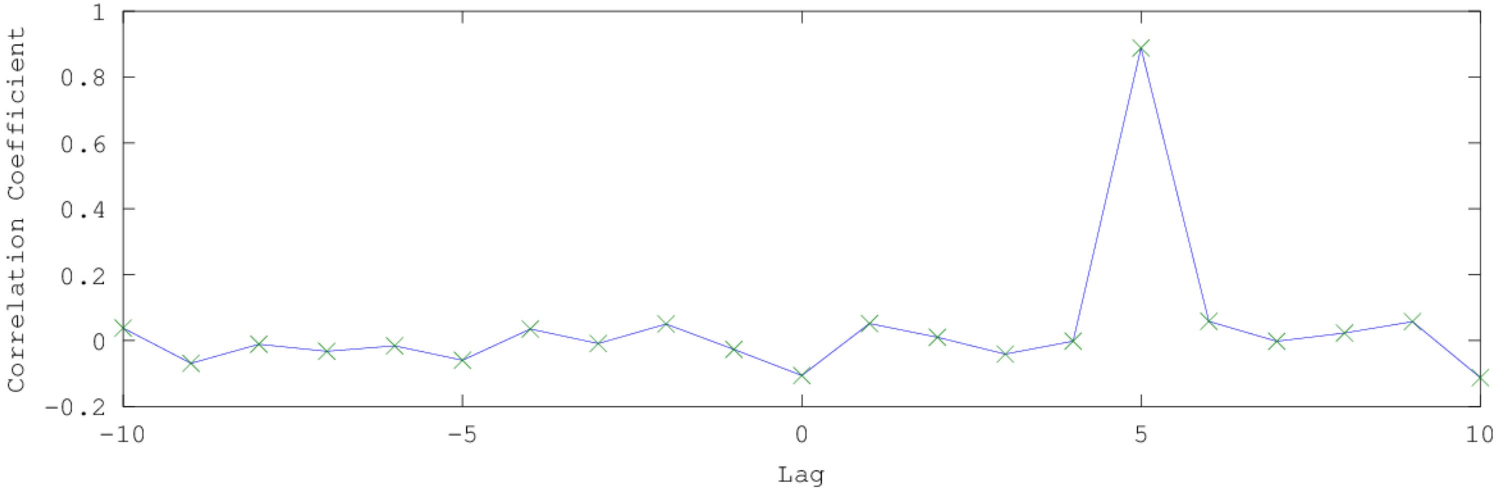
Cross-Correlation

- Example with some fabricated data:



- Tree2 seems to be a predictor of Tree1 in the smoothed series
- Calculate cross-correlation up to a lag of 10 to check

Cross-Correlation

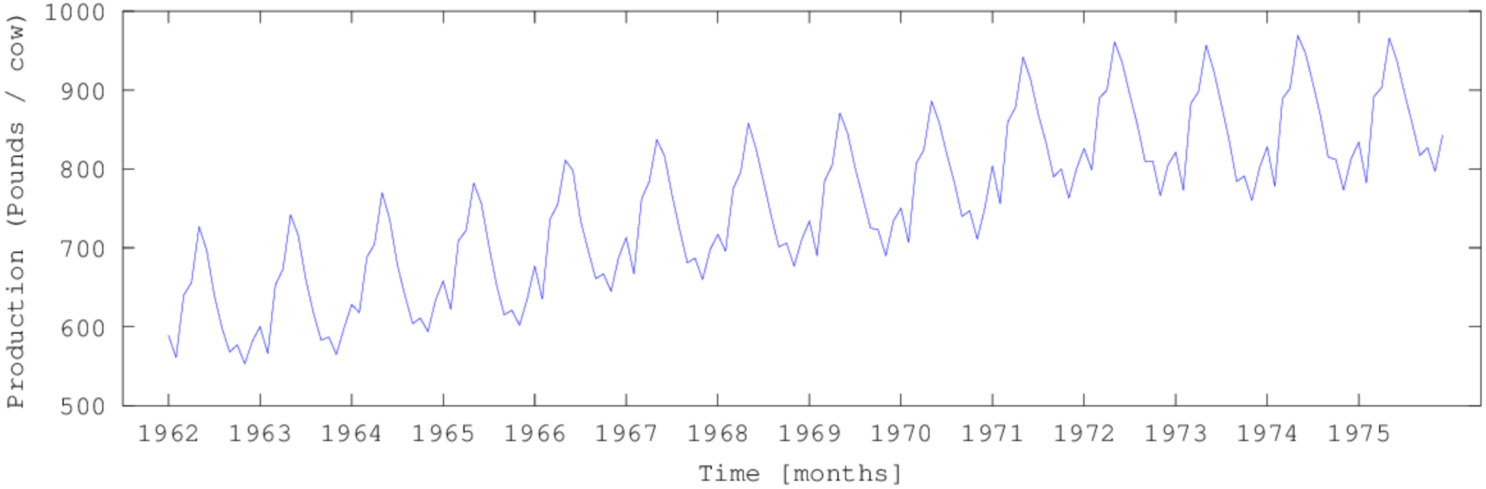


- There is a high correlation at lag +5
- This would be good evidence that factors affecting Tree2, affect Tree1 with a delay of 5 years.
- Tree2 could be used to make predictions about Tree1

- Again: Be carful with trend!

Periodic Series

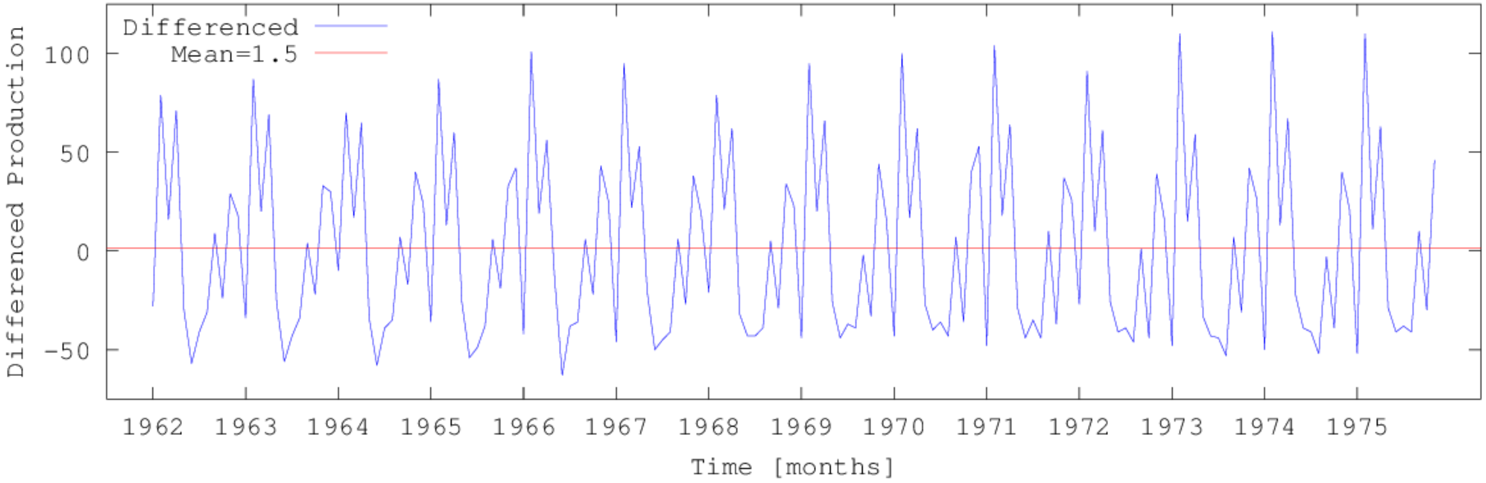
- Cross-Correlation can be used to find periodic events (e.g. seasonal influences) in data



- What can we see?
  - Trend superimposed by a periodic cycle
  - Periodic peaks always in May

Periodic Series

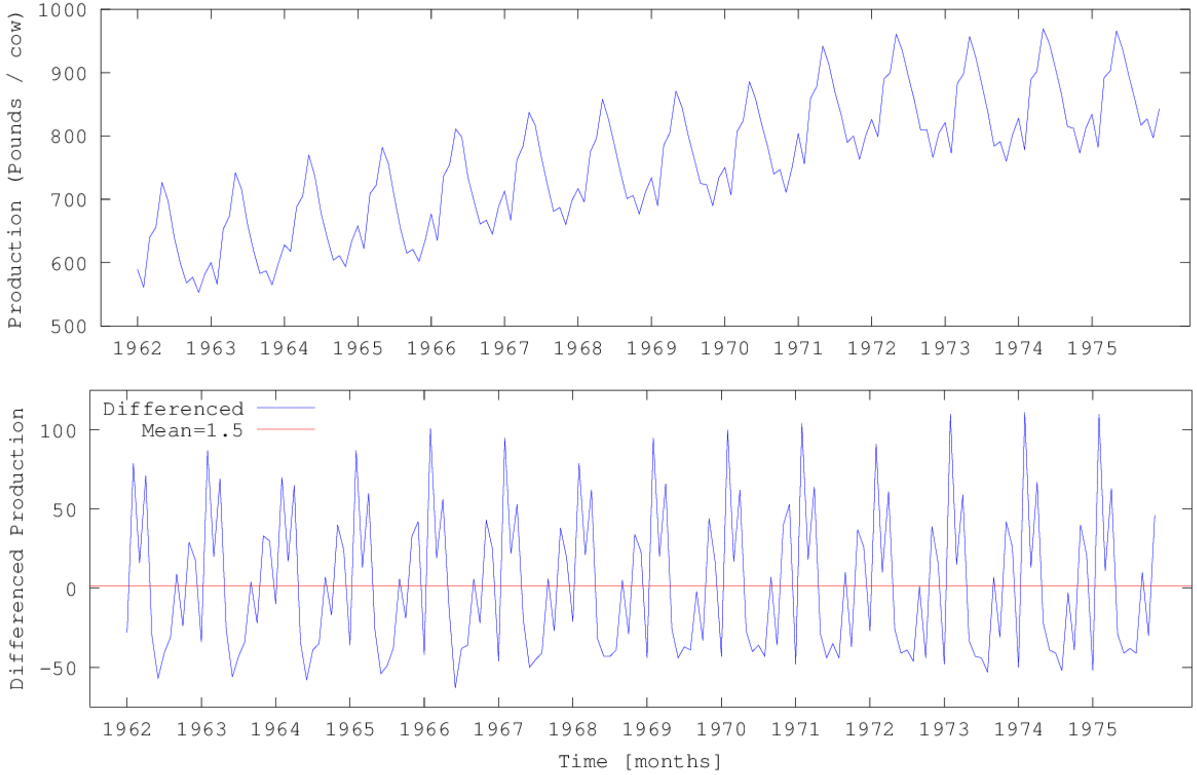
- First remove the trend by differencing



- Suddenly we have more peaks? What does that mean?

Group Task!

3 5

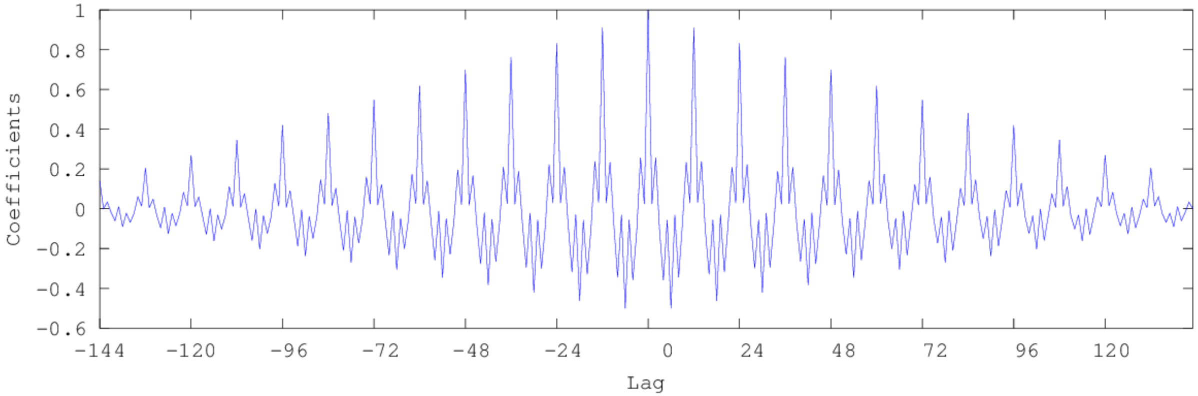


First one peak, now many?

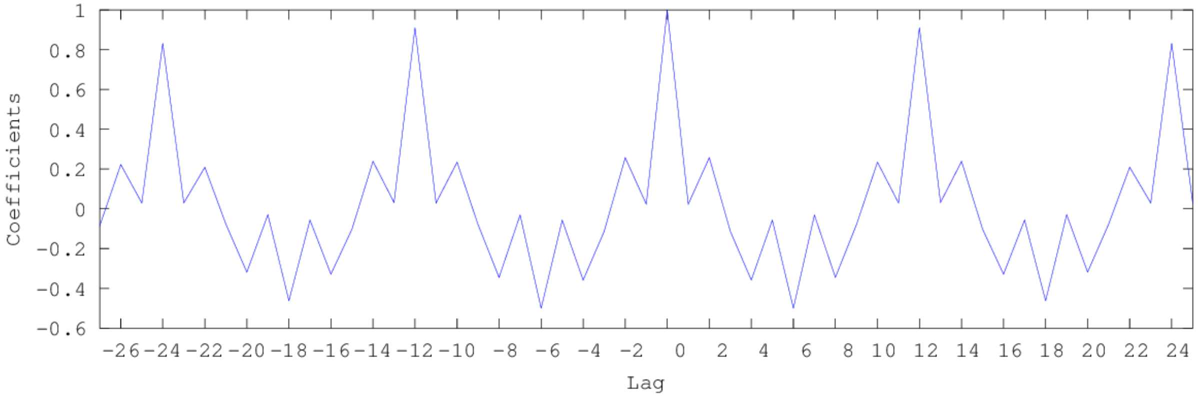
Is that correct?

Periodic Series

- Calculate cross-correlation with itself: Autocorrelation!

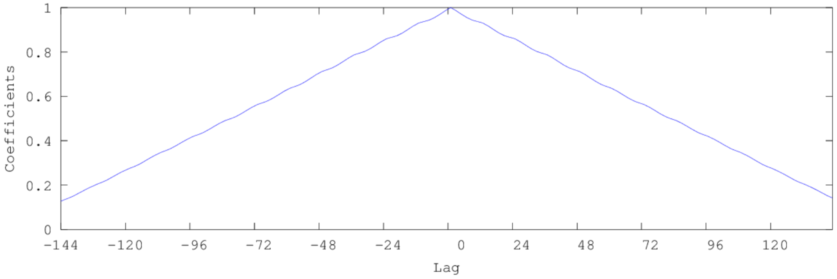


- Zoom to +/- 24 lag



Time Series …

- …are 2-dimensional data
  - Check for gaps and record and plot explicitly over time
- …often show high-frequency fluctuations (e.g. noise)
  - Remove by smoothing (and maybe repeated re-smoothing)
- …can have a trend
  - Fit functions to estimate trend or calculate correlation with time
  - Remove by differencing or subtracting fitted function
- Cross/Auto-correlation can reveal
  - predictors or indicators
  - periodic features (e.g. season)
  - Beware of trend in series!



EDA Summary

- With EDA we search for patterns and structure in data to
  - learn and understand the behaviour of our system
  - find factors that influence our outcome
  - find interactions between factors
  - form hypotheses about the behaviour and the causal connections in our system

$$y = f(x, \varepsilon)$$

- We try to find factors x that influence y in our model
  - Try to identify how the combine/interact: $f()$
  - Look for evidence of hidden factors in ε

EDA Summary

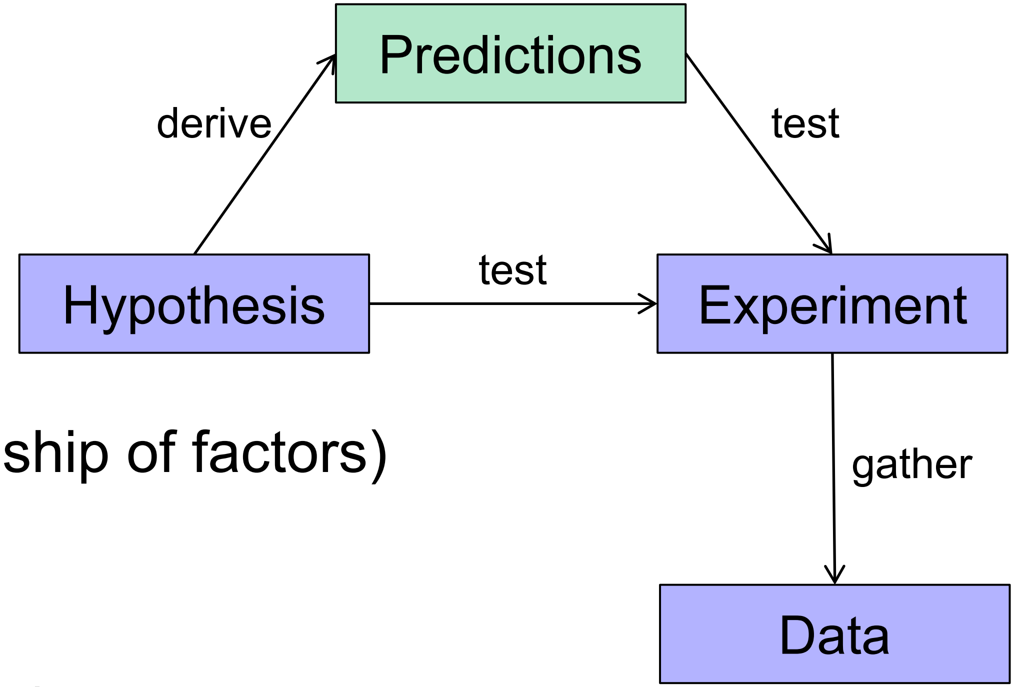
- Visualisations to exploit human pattern recognition abilities
  - Frequency diagrams and boxplots
  - Scatterplots and line plots
  - Contingency tables and proportion charts
- Different measures in our toolset:
  - Central tendency (Mean, Median, Mode)
  - Dispersion (variance, standard deviation, range, IQR)
  - Shape (skew, kurtosis)
  - Association (chi-square, covariance, correlation coefficients)
  - Time series (trend, cross/auto-correlation)

EDA Summary

- Allowed is everything that helps you understand the data, but
  - be careful to remember what the graph/number represents, e.g. after transformation and smoothing
  - you only find evidence for causal relations, interactions, dependencies
- EDA can be useful to
  - build up a preliminary causal model
  - check early hypotheses
  - define and refine hypotheses for experiments

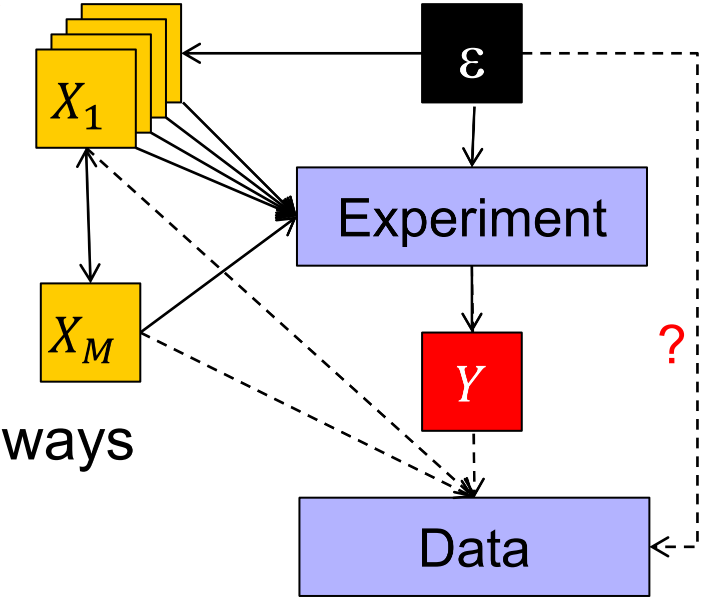
Experiment Design

- Why experiments?
  - To answer a question!
  - To test a hypothesis (about a causal relationship of factors)



- General Hypothesis: Factor X affect behaviour/outcome Y
- We have preliminary ideas of X, Y, and the effect, through EDA, a model, maybe just from an idea, ….

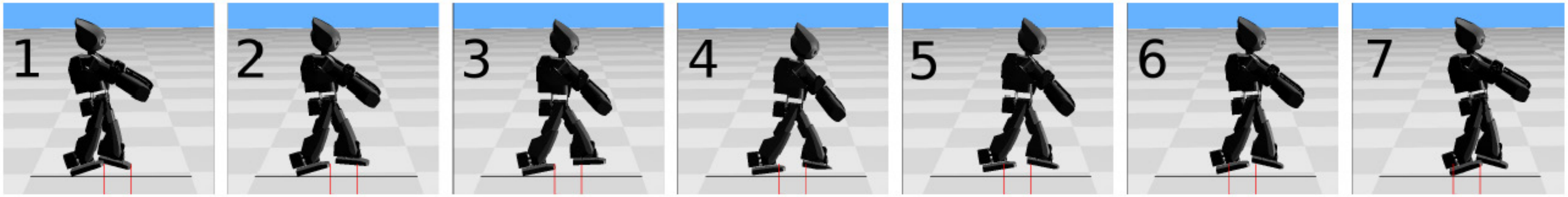
Finding X



- Step 1:
  - Define factors X_M and $X_{1 \ldots n}$
  - Define outcome Y
  - One by one, find valid and reliable ways to measure Xs and Y

- Example: Evolving a neural controller for a simulated robot:
  - Leg servos are controlled by neural network
  - Simulated flat environment without obstacles but linear slope
  - Question: How does network layout affect the robustness of the walking gait

Finding X and Y



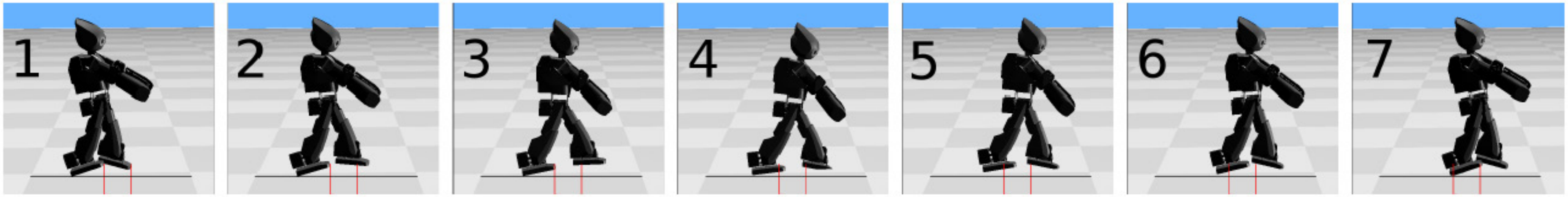
Factors X

- Neural Network
  - Layout, Parameters, Weights
- Evolution
  - Parameters, Operators, Fitness Function
- Environment
  - Slope
- Agent
  - Initial Condition

Factor Y

- Walking gait
- Robustness
  - Stability, Distance

Finding X and Y



Measuring X

- NN:
  - Layout: #hidden neurons
  - Parameters: Snapshot of all values
  - Weights: double over generations
- Environment:
  - Slope as angle relative to initial direction
- Agent: Initial servo values
- ….

Measuring Y

- Walking gait
  - servo values over t and generations
- Distance
  - Euclidian distance to end point
  - Integrated path?
  - Position over time?
- Stability?
- …..

Finding X and Y

- Outcome of step 1:
  - A tree for X and Y, listing all factors and their measurements (=variables)
  - Decision of which variable(s) to manipulate and control
  - Optimal: A diagram of interactions between variables (model)
- The diagram and the tree of factors are your thinking aids!
  - Aid as a representation of your thinking process and progress
- Pilot study + EDA to check the validity and reliability of the chosen measurements

What have we learned?

1. Time series are two-dimensional data
2. Visualise them over time and use smoothing to reveal trends and general features
3. Be careful with trend and correlations
4. Cross-Correlations with lags can reveal interactions with time delay or periodic features
5. Step one of experimental design: Lay out all factors in front of you and have a long and close look at them!
6. The better you do this step, the better the experiment will be