Data Transformation

De-trending

Time series data

- Shows what happens as time passes.
- Each data point is made up of 3 components:
  - Trend
  - Seasonal
  - Irregular.

  For an **additive** series:

  Data value = Trend + Seasonal + Irregular

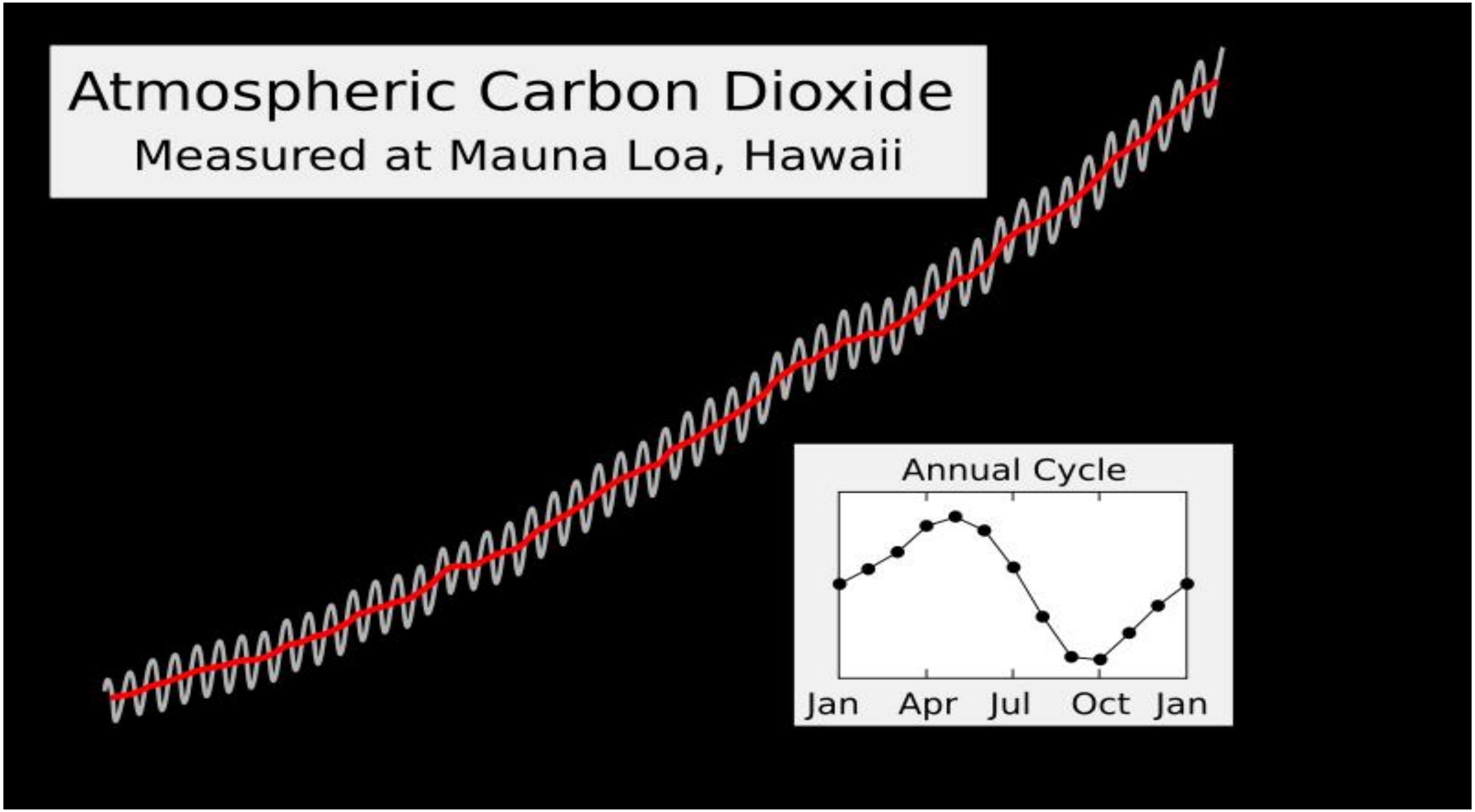
Beginnings

- It is important to look at the series you are analysing before you start.
- Draw a graph.
- Look for the different components.
- Think about what might be the best way of analysing it.

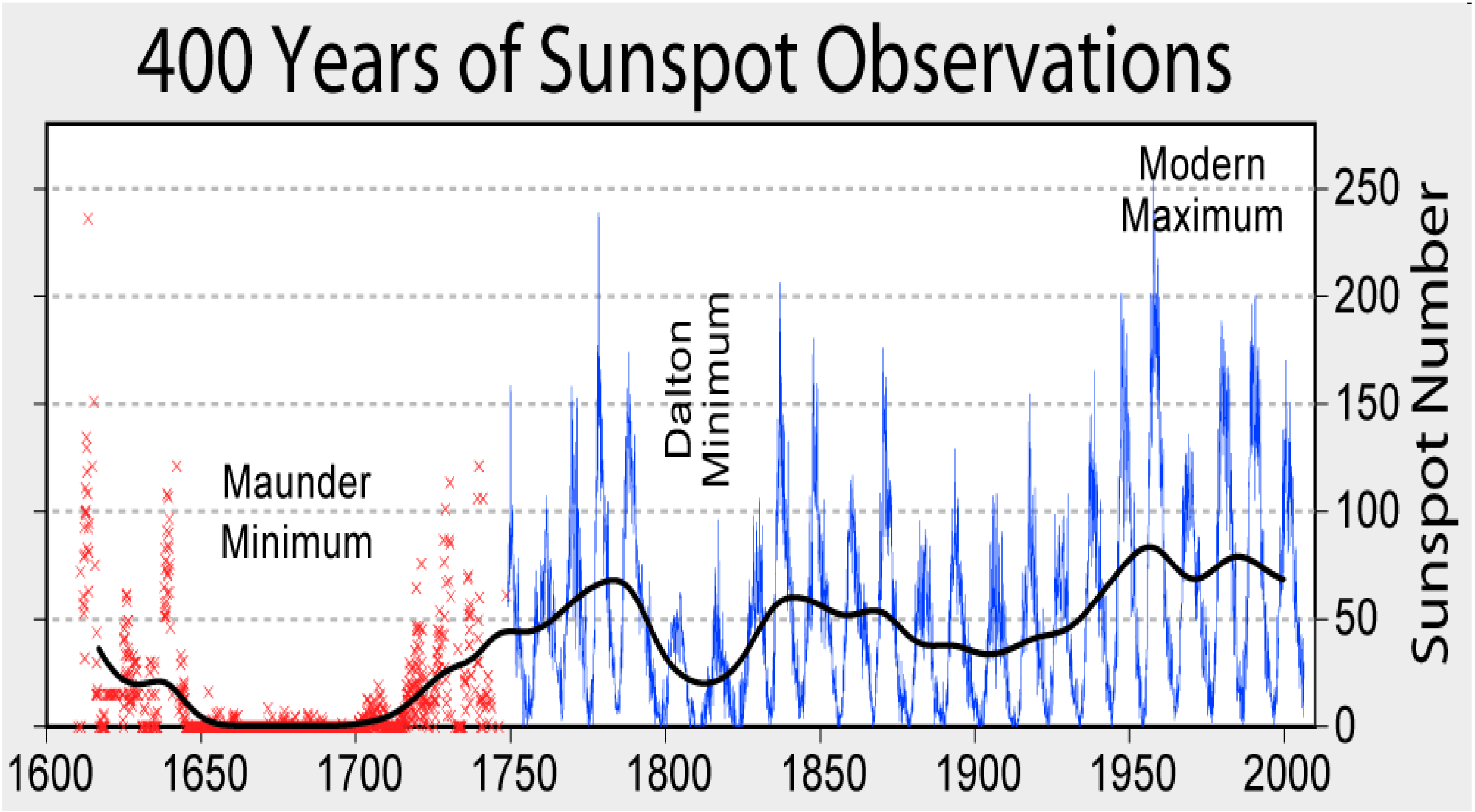
Always look at the data!

or else!

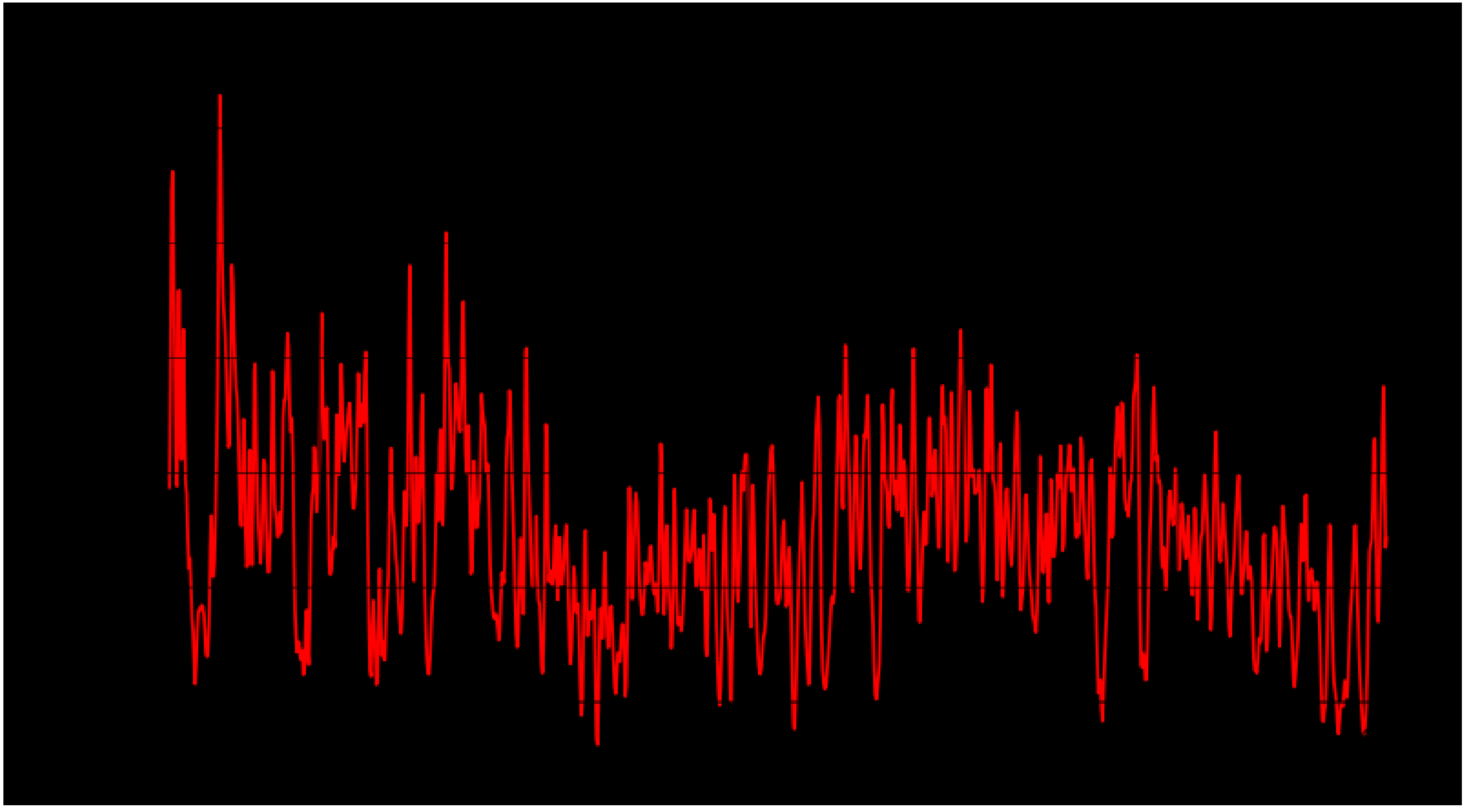
Atmospheric CO_2



Years: 1958 to now; vertical scale 300 to 400ish


400 Years of Sunspot Observations
Maunder
Minimum
Dalton
Minimum
Modern
Maximum
Sunspot Number
0
50
100
150
200
250
1600
1650
1700
1750
1800
1850
1900
1950
2000

Ancient sunspot data



Detrending: deterministic trend

- Fit a plain linear regression, then subtract it out:
  - Fit $Y_t = m*t + b$,
  - New data is $Z_t = Y_t - m*t - b$
  - Or use quadratic fit, exponential fit, etc.

Detrending: stochastic trend

- Differencing
  - For linear trend, new data is $Z_t = Y_t – Y_{t-1}$
  - To remove quadratic trend, do it again:
    - $W_t = Z_t – Z_{t-1} = Y_t – 2Y_{t-1} + Y_{t-2}$
  - Like taking derivatives

  - Thought Exercise: what is the equivalent if you think the trend is exponential, not linear?

- Hard to decide: Which of the two should be applied: regression or differencing?

Cycles/Seasonality

- Suppose an annual cycle
- Sample quarterly: 3-mo, 6-hi, 9-med, 12-lo
- Sample every 6 months: 3-med, 9-med
  - Or 6-hi, 12-lo
- To detect a cycle, must sample at least twice its frequency

The basic problem

- We have data, want to find
  - Cycle length (e.g. monsoon), or
  - Strength of seasonal components
- Simple Initial Idea:
  - use sine waves as explanatory variables
- If a sine wave at a certain frequency explains things well, then there's a lot of strength.
  - Could be our cycle's frequency
  - Or strength of known seasonal component
- Add physical insights and domain understanding