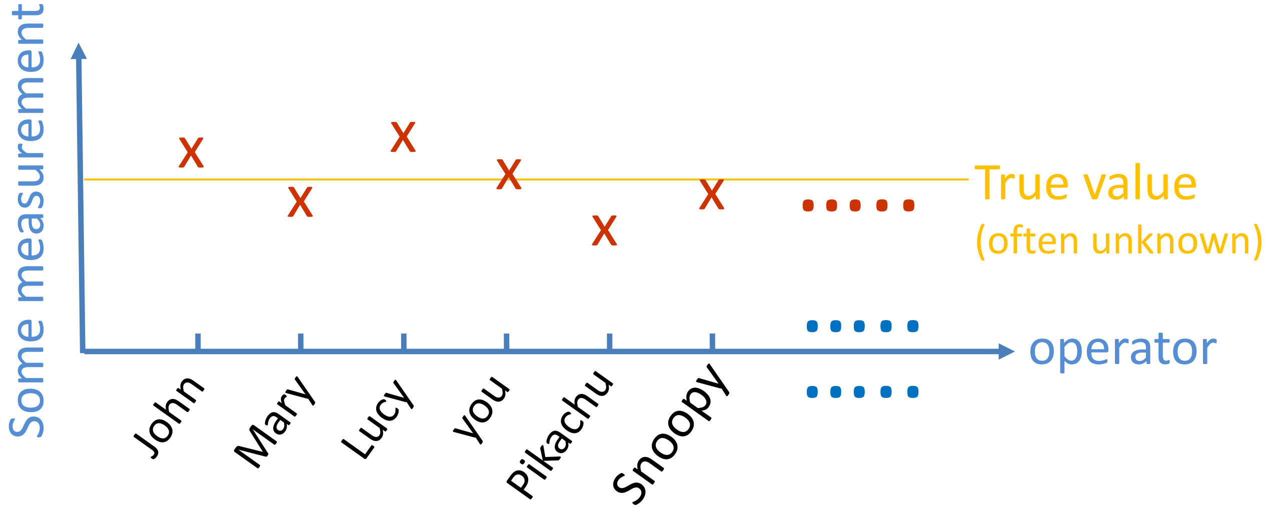
# Week 1 Lecture:

# Data Processing and Analysis

Last update: August 18, 2022

# Data Analysis: Basic Concepts

Suppose you and your classmates are tasked to measure something …



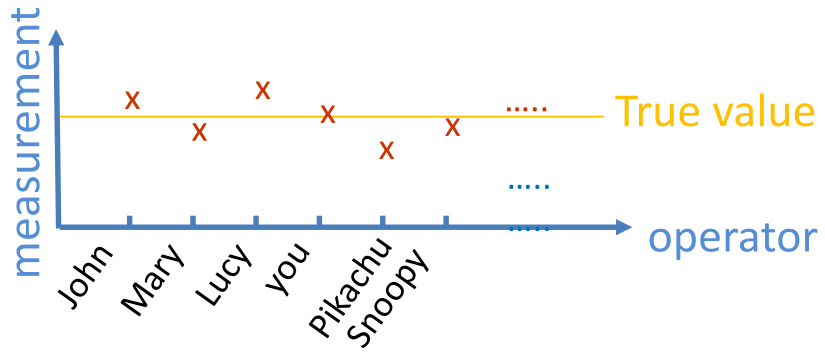

**Data (Measurements) = true value + error**

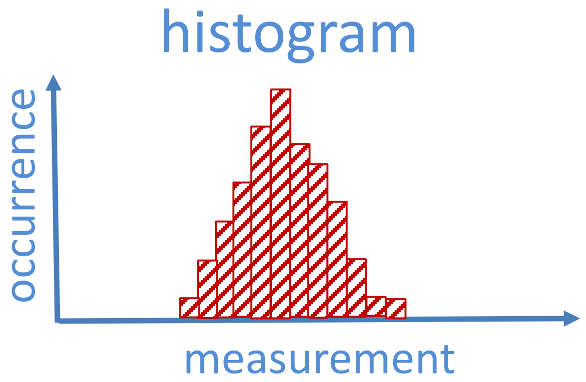
# Data Analysis: Basic Concepts (cont'd)

## Random error

Random error, often called "noise" and "fluctuation", is almost always incurred in measurement, causing the data to spread around a central average value (hopefully the true value) in a unpredictable way

There are a variety of random error types and sources:
thermal noise, electronic noise, human interpretation error, quantization/truncation/rounding error, etc.



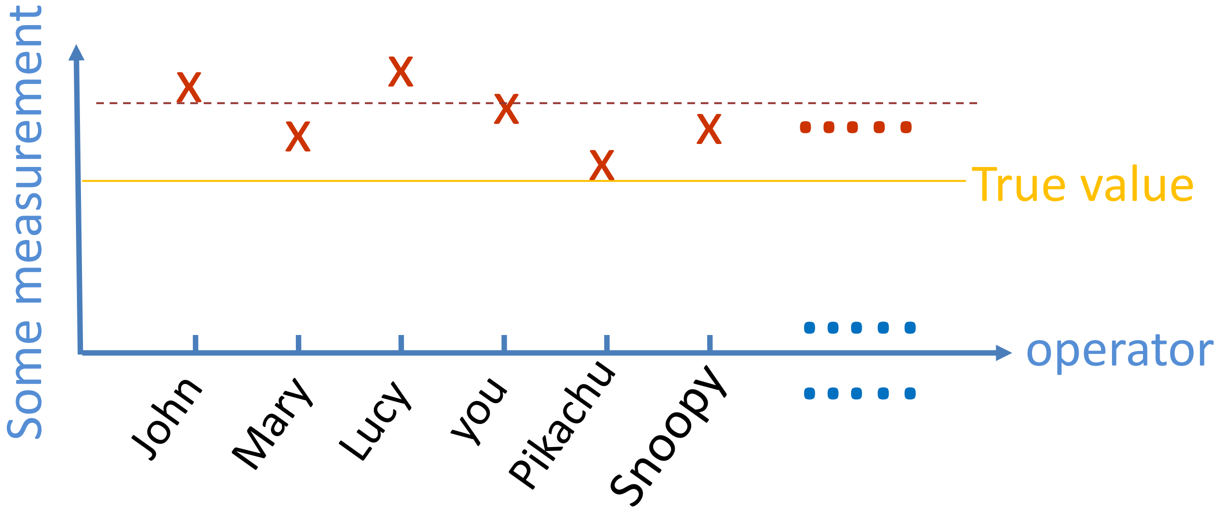



# Data Analysis: Basic Concepts (cont'd)

## Systematic error (bias)

Caused by imperfect measurement methods and/or calibrations, environmental interferences, etc. in a predictable way

Say you and your classmates use the same ruler which is inaccurate with a constant offset or bias …



# Data Analysis: Basic Concepts (cont'd)

## Level of Accuracy, Precision, Resolution

The *accuracy* of measurement is often dictated by measuring device's *resolution*. E.g. your ruler's smallest marking is only 1/8 inch (so don't bother to take data in 1/64 inch).

Sometimes the environment (e.g. vibration) limits the data accuracy, even your measuring device has high *precision*

## Sampling, Population, Sampling Rate

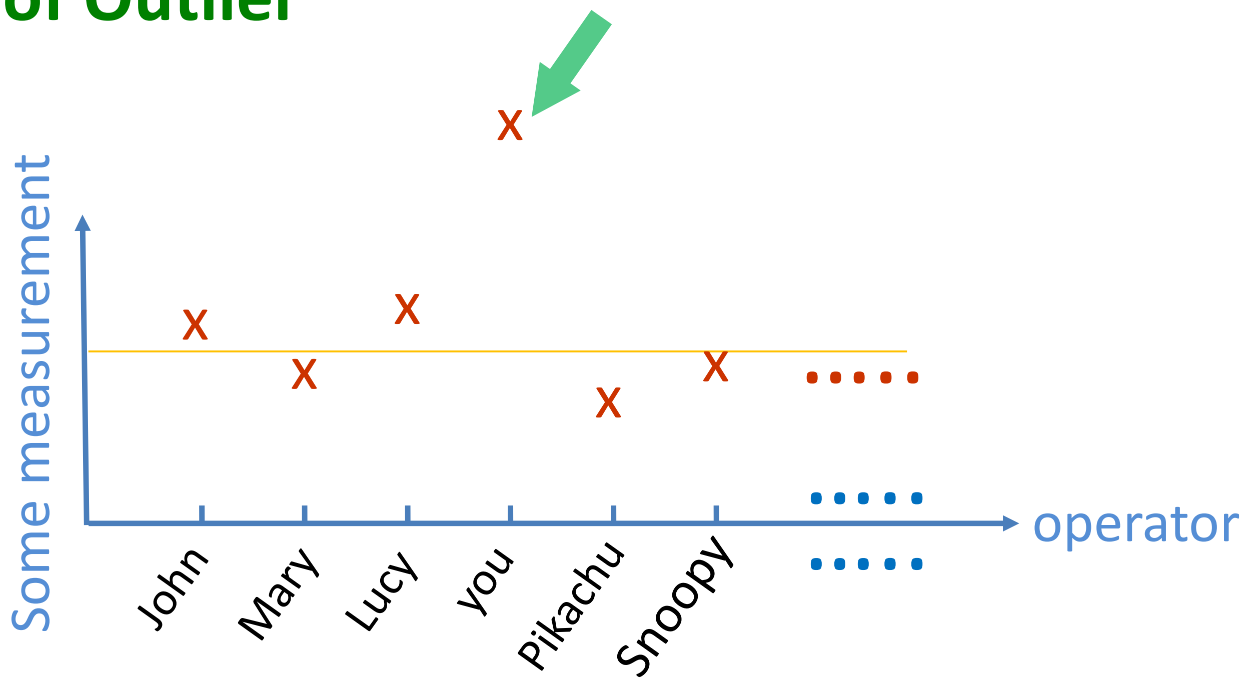
*Sampling* describes the process of data collection from a sample group, organization or *population*. *Sampling rate* determines how the data are collected in time (frequency) or in space (spacing)

# Data Analysis: Basic Concepts (cont'd)

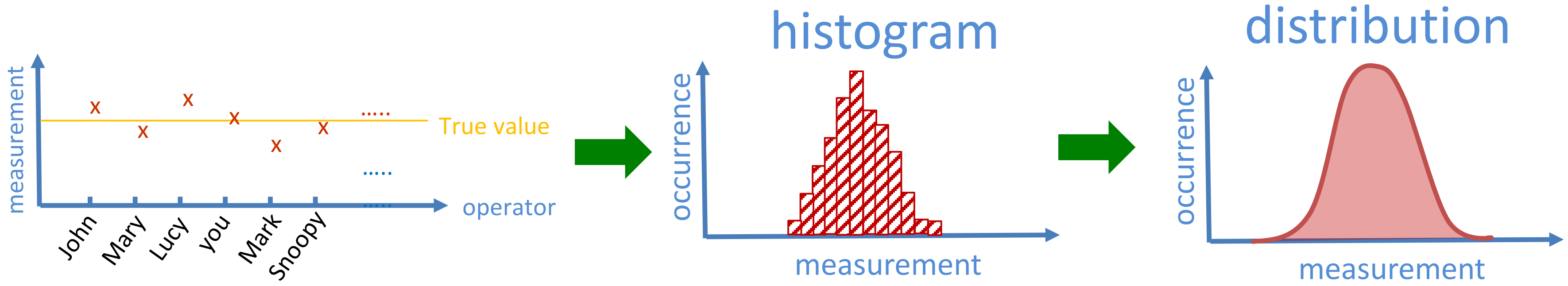
## Reproducibility

Is your Nobel-prize-winning discovery a one-time deal?  Can you repeat the same measurement with the same accuracy?

## Identification of Outlier



# Statistics 101: Quantifications of Data




## Mean (central tendency)

Arithmetic mean $\mu = \dfrac{x_1 + x_2 + x_3 + \cdots}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

## Standard Deviation (variability)

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Variance = $\sigma^2$

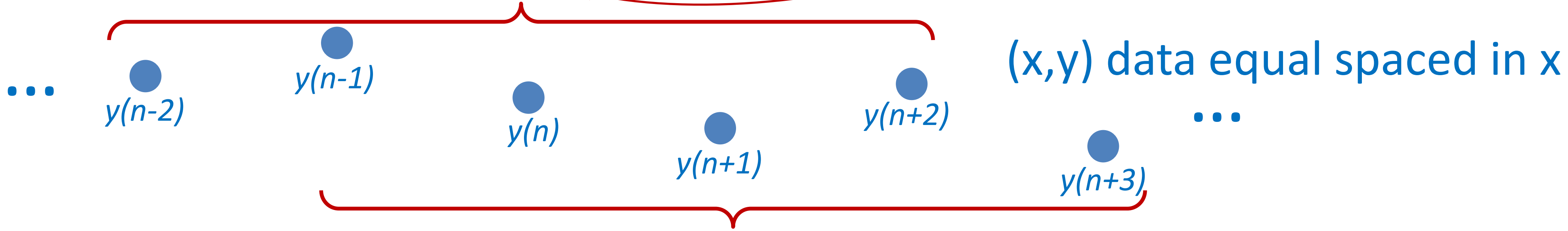
# Data Processing: Smoothing

Smoothing operations are commonly applied to the data for suppressing interference or noise while preserving important features. Popular algorithms include running average and Savitzky-Golay filter

**Running average** (or moving average) works by continuously taking average of neighboring data on the run

$$y'(n) = \frac{y(n-2)+y(n-1)+y(n)+y(n+1)+y(n+2)}{5},$$

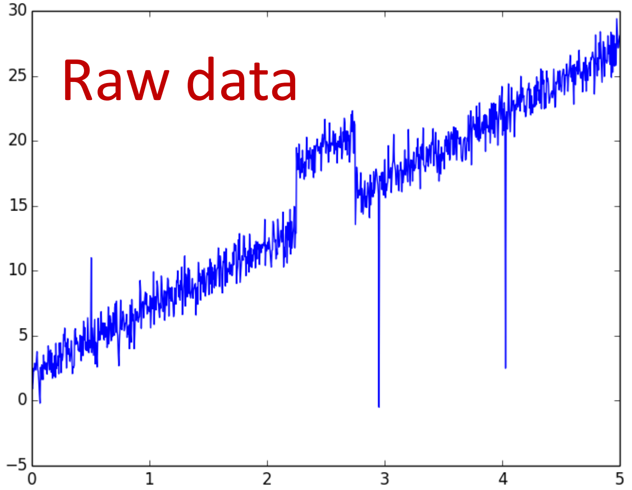
here average span = 5

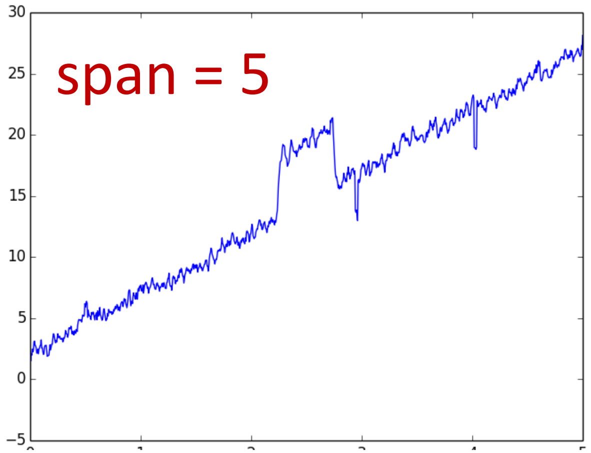



$$y'(n+1) = \frac{y(n-1)+y(n)+y(n+1)+y(n+2)+y(n+3)}{5}, etc.$$

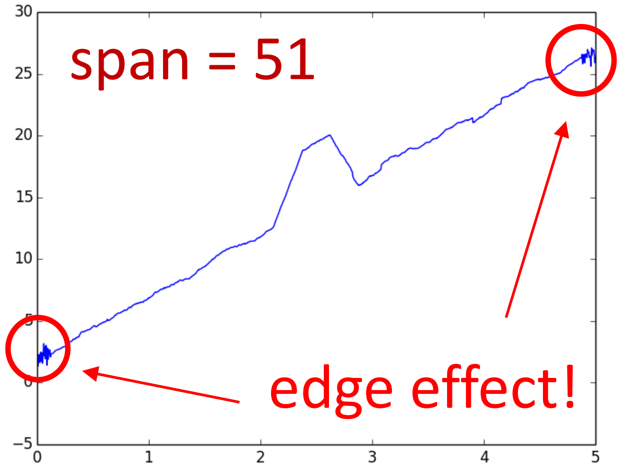
# Data Processing: Smoothing (cont'd)

***A priori*** information is important! Your ***prior*** knowledge about the data dictate how you will process the data.




Short span to preserve feature



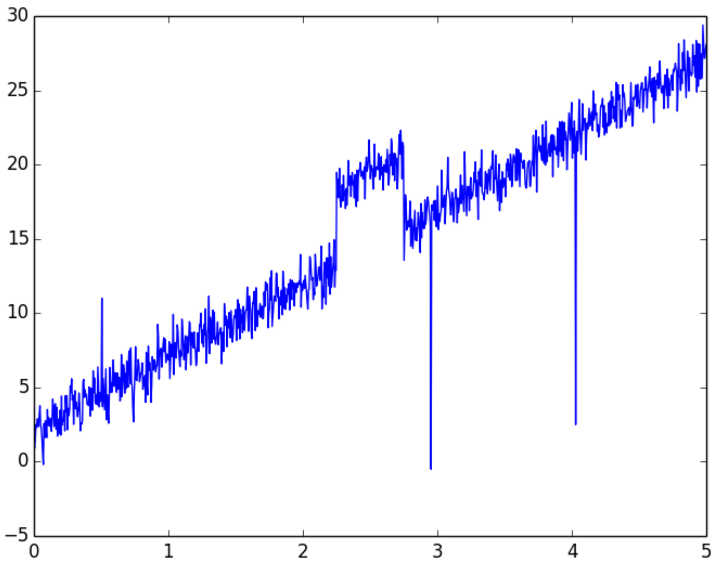

long span to maximize smoothing

Edge effect is caused by the fact the first and last span/2 data points are left unprocessed. It can be reduced by using short span in later multi-pass (see next slide)

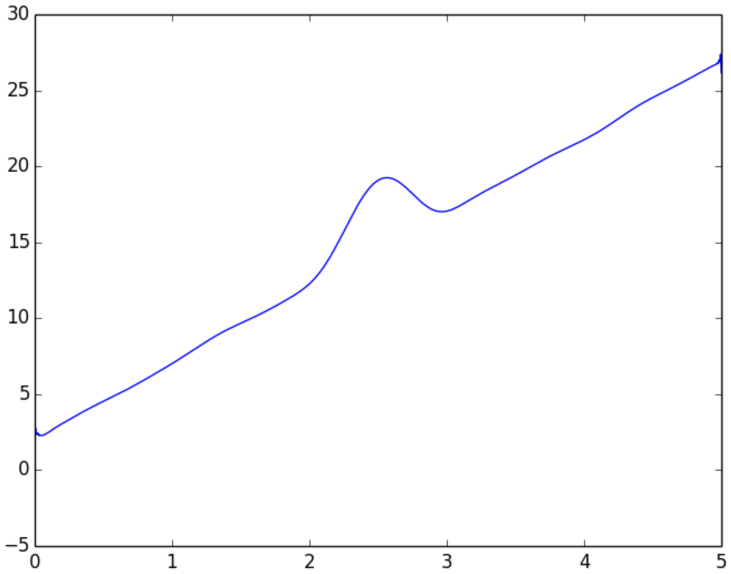
# Data Processing: Smoothing (cont'd)

Often the running average is applied iteratively using multiple passes with different average spans - output of previous run become the new data for next run

Raw data

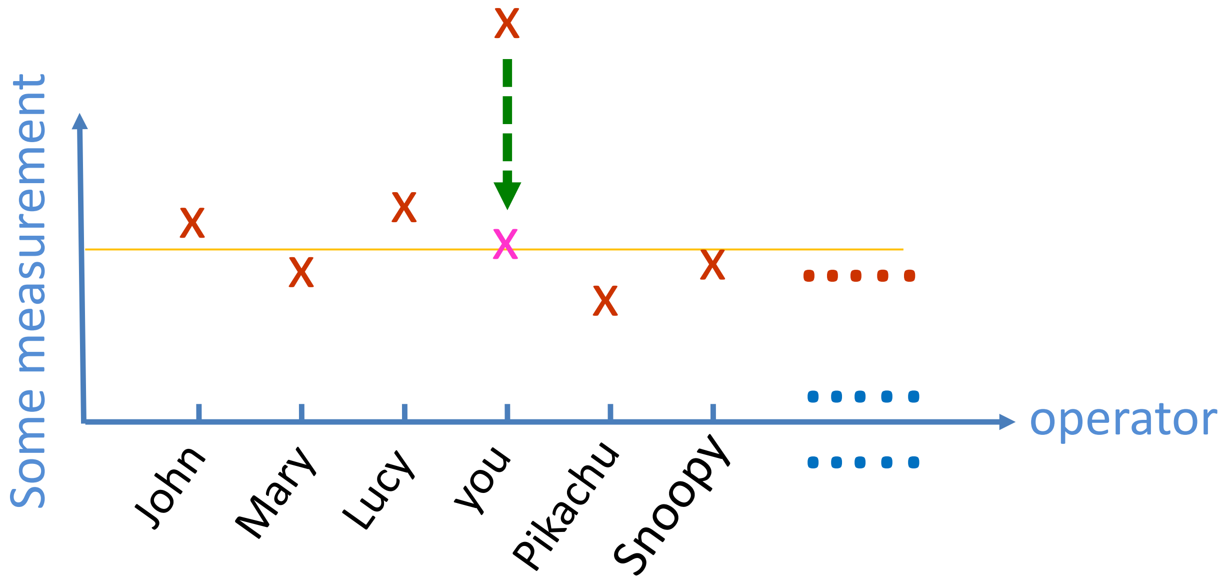



span combinations of 11,31,51,71,31,21,11,5



# Data Processing: Outlier Removal

Brute force approach: remove them "by hand" if you can guess where the data suppose to be located

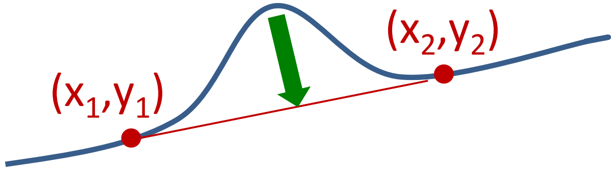


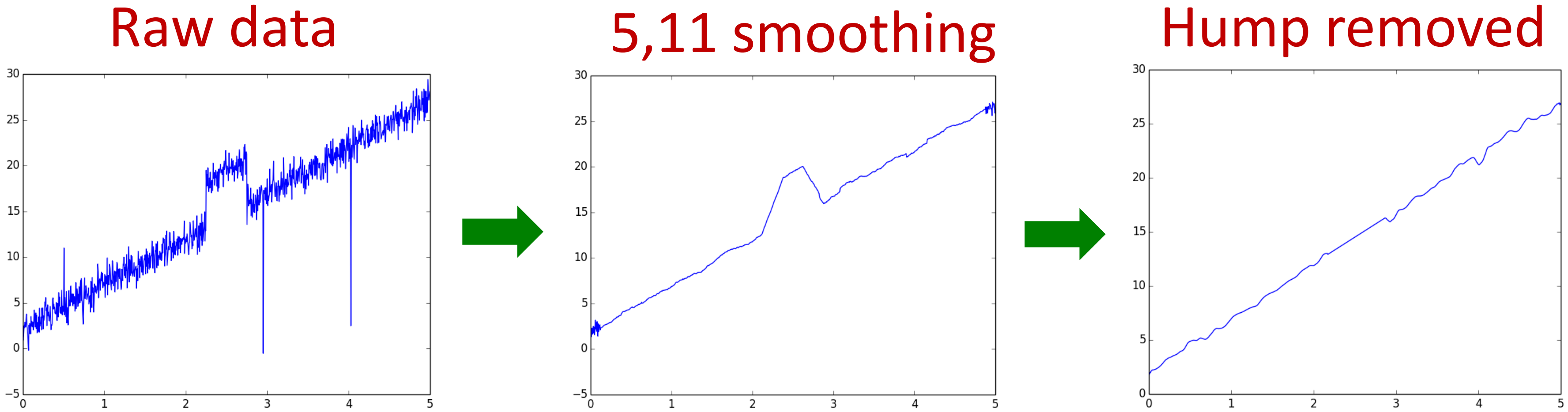

Or you can continue using running average in multiple passes to smooth out the outliers. But don't overdo it – otherwise you may ruin the underlying features you wish to preserve

# Data Processing: Outlier Removal (cont'd)

For more elaborate data, the following simple formula allows you to remove a "hump" from a slope by using "good" neighboring data

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x_1}$$



# Data Processing: Curve Fitting

Once you have done enough data smoothing, you may want to continue fitting the smoothed data to a model for gaining more insight to the underlying trend of the data and even backing out key parameters of the trend

As mentioned, *A priori* information is important. Your knowledge about the data will help you to decide how you will fit the data: the choice of the fitting functions and the parameters within, e.g. straight line vs. polynomials, first order vs. second order, etc.

# Data Processing: Curve Fitting (cont'd)

## Least squares fit

A popular and robust method for general functional fit: linear or non-linear. Suppose you wish to fit your ($\underline{x}$,$\underline{y}$) data to a model function *f* with parameters $\underline{a}$:

$$\underline{y} = f(\underline{a}\,;\underline{x}) \qquad \text{e.g.} \quad \underline{y} = a_1 + a_2\underline{x} + a_3\underline{x}^2$$

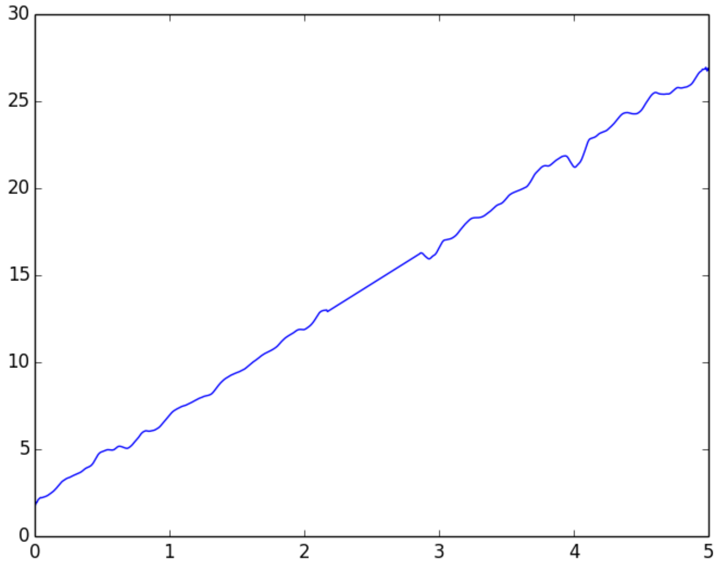
You can use a minimization software to obtain the best-fit parameters $\underline{a}$ by minimizing the least squares sum of errors

$$\text{Min.} \ \sum_{i=1}^{N}\left[\hat{y}_i - f(\underline{a}\,;\,x_i)\right]^2 \quad \Rightarrow \quad \text{best-fit parameters } \underline{a}$$

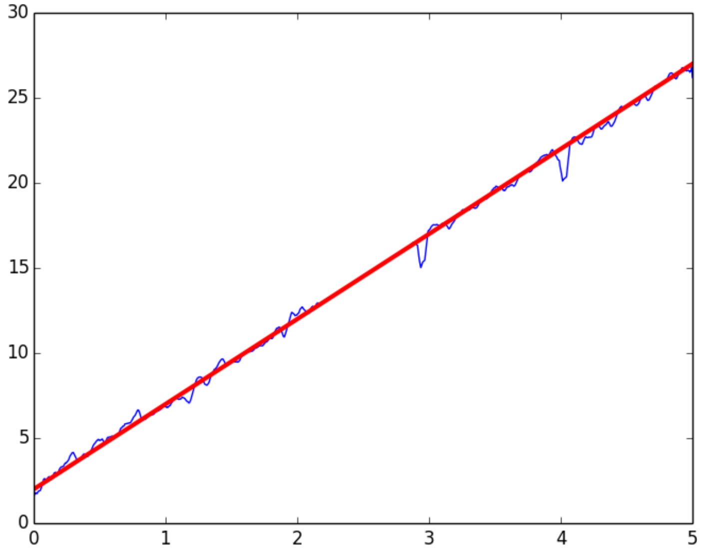
$\hat{y}_i$ : measured data

# Data Processing: Curve Fitting (cont'd)

Data after 5,11 smoothing and hump removed




Fitted with linear line




Original underlying baseline:

$$\underline{y} = 2 + 5\underline{x}$$

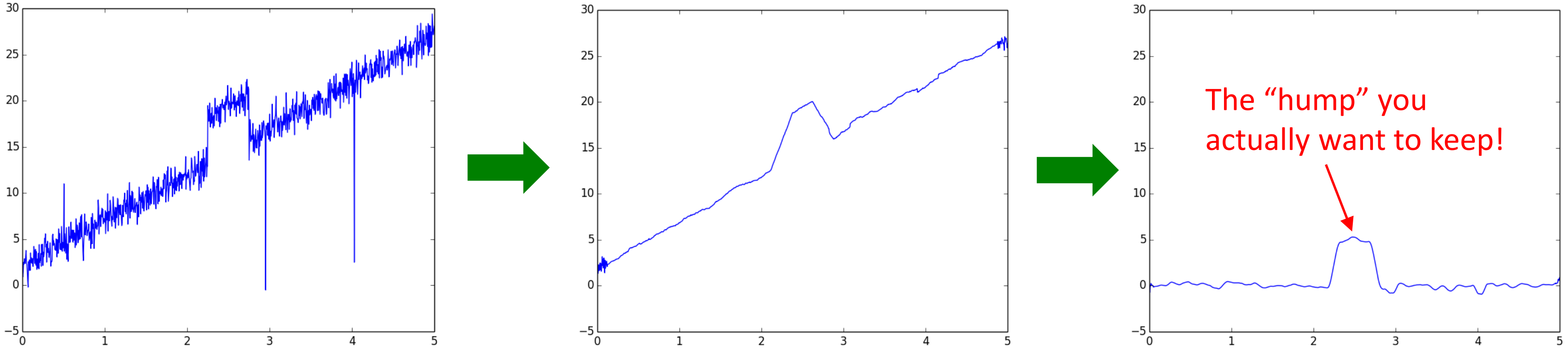
Best-fit baseline:

$$\underline{y} = 2.11 + 4.93\underline{x}$$

# Data Processing: Baseline Removal

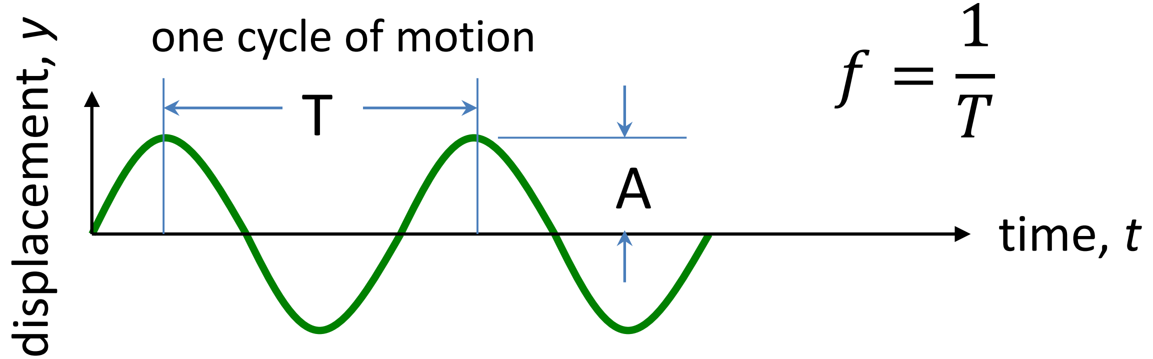
**Warning!** Once again, your prior knowledge about the data dictate how you should process the data.  Suppose your data actually came from a communication channel in which some non-zero signals were transmitted. Then those non-zero outliers are the features you need to keep and the random noise and the baseline are the ones you should remove!

Raw data | 5,11 smoothing | baseline removed



# Periodic Motion 101

Time measurements in aircraft structure are often oscillatory and in many cases sinusoidal. As you have learned from elementary physics, a sinusoidal motion is periodic with period T, frequency f and maximum amplitude A:




$$f = \frac{1}{T}$$

The sinusoidal motion above can be expressed as:

$$y = A\sin(\omega t + \theta)\ ;\ \omega = 2\pi f$$

ω: angular frequency
θ: phase (motion offset in time)

Unit: time $t$ in second, microsecond (μs, 1μs=$10^{-6}$ second), etc.; frequency $f$ in Hz (1 Hz = 1 cycle/second), kilohertz (kHz, 1 kHz = $10^3$ Hz), etc.; phase $\theta$ in radian ($360^o$ = $2\pi$ radian).

# Data Processing: Useful MATLAB Functions

**Running Average:**
Matlab function: smooth
yy = smooth(y,span)
y = original data
span = running average span
yy = output of smoothed data
The default span is 5.

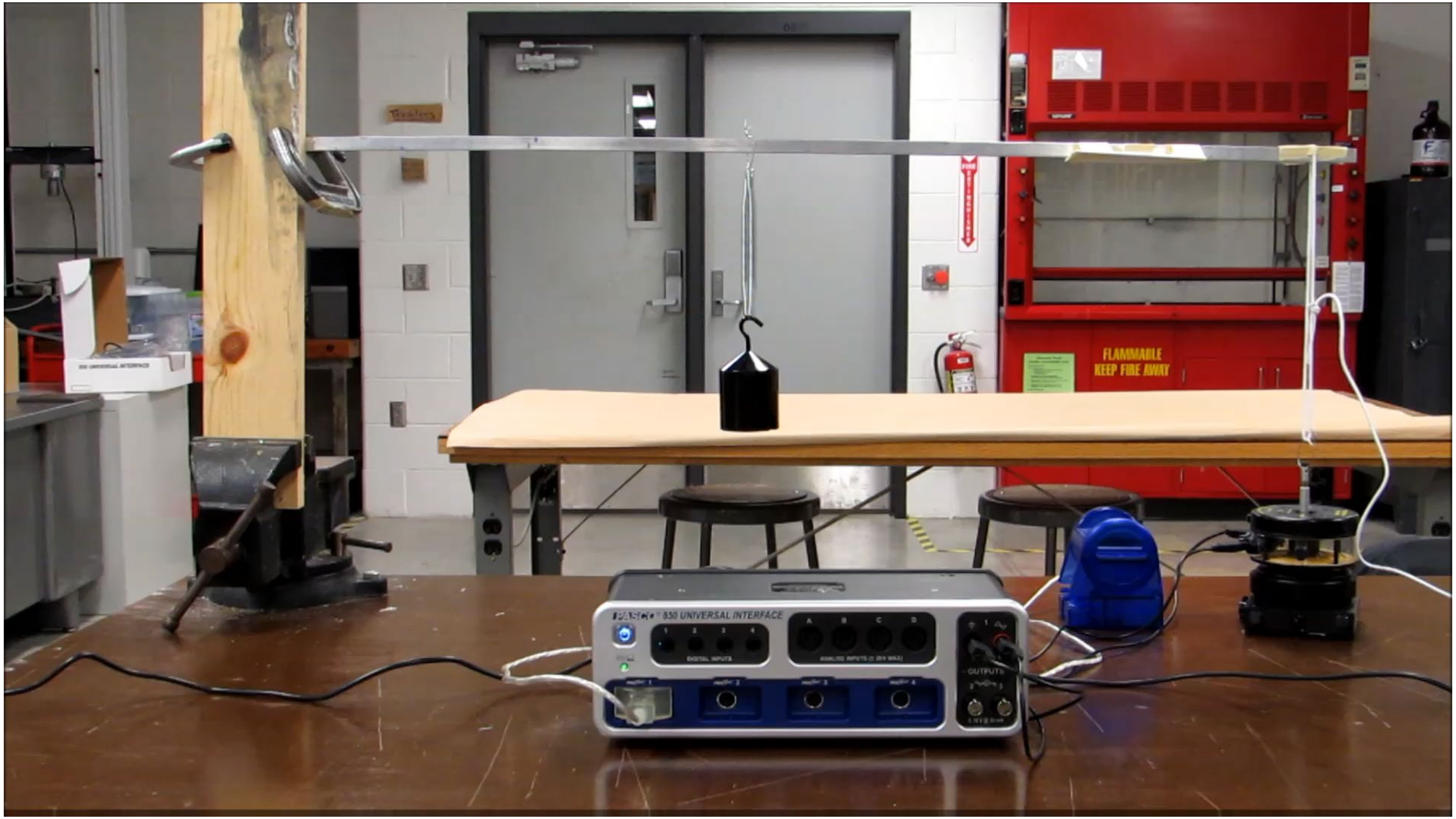
**Polynomial Curve Fitting:**
Matlab function: polyfit
p = polyfit(x,y,n)
x = original x-array
y = original y-array
n = degree of polynomial
p = array of the polynomial's coefficients

**Polynomial Curve Evaluation:**
y = polyval(p,x)
p = array of polynomial's coefficients
x = array of x-data
y = array of polynomial's y-data

**General Curve Fitting:**
Matlab function: lsqcurvefit
x = lsqcurvefit(fun,x0,xdata,ydata)
fun = function format
x0 = original guess of the coefficients
xdata = x-array of data
ydata = y-array of data
x = coefficients of the curve

example of fun and x0:
fun = @(x,t) x(1)+x(2)*t+x(3)*t.*t
x0 = [1,5,10]

# Lab 1: Practice Experiment and Data Analysis



!!! Remember to apply what you learned today in data analysis and processing to all other labs in this course (and your future measurement work as well) !!!