Skoltech
Skolkovo Institute of Science and Technology

“Space Data Processing:
Making Sense of Experimental Data”

# Topic 1
# "Introduction to statistical analysis”

**Tatiana Podladchikova   Rupert Gerzer**
**Term 4, March 28 – May 27, 2016**
**t.podladchikova@skoltech.ru**

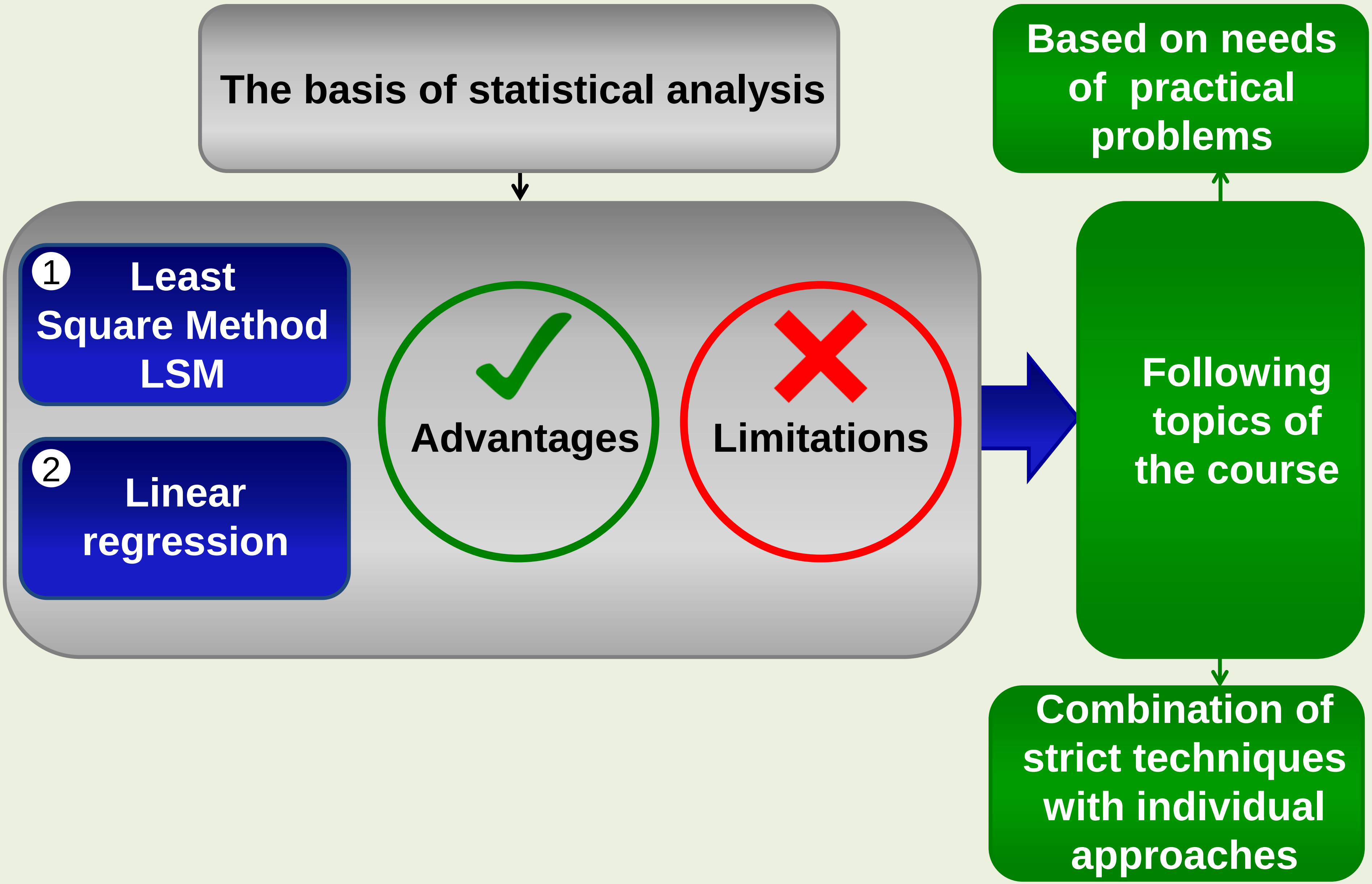
# The basis of statistical analysis

1. Least Square Method LSM
2. Linear regression

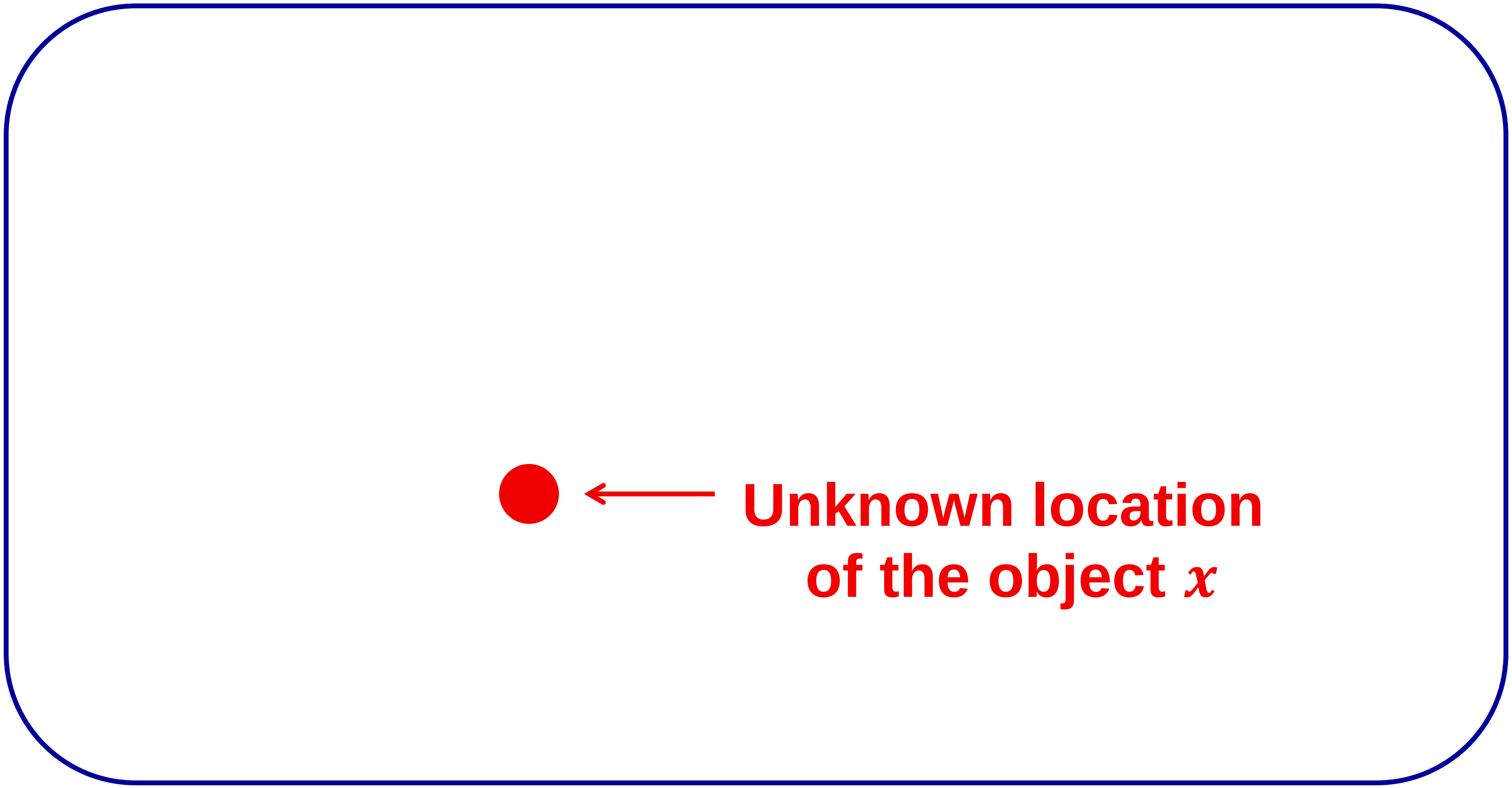
Advantages

Limitations

# Transform theoretical knowledge into useful skills

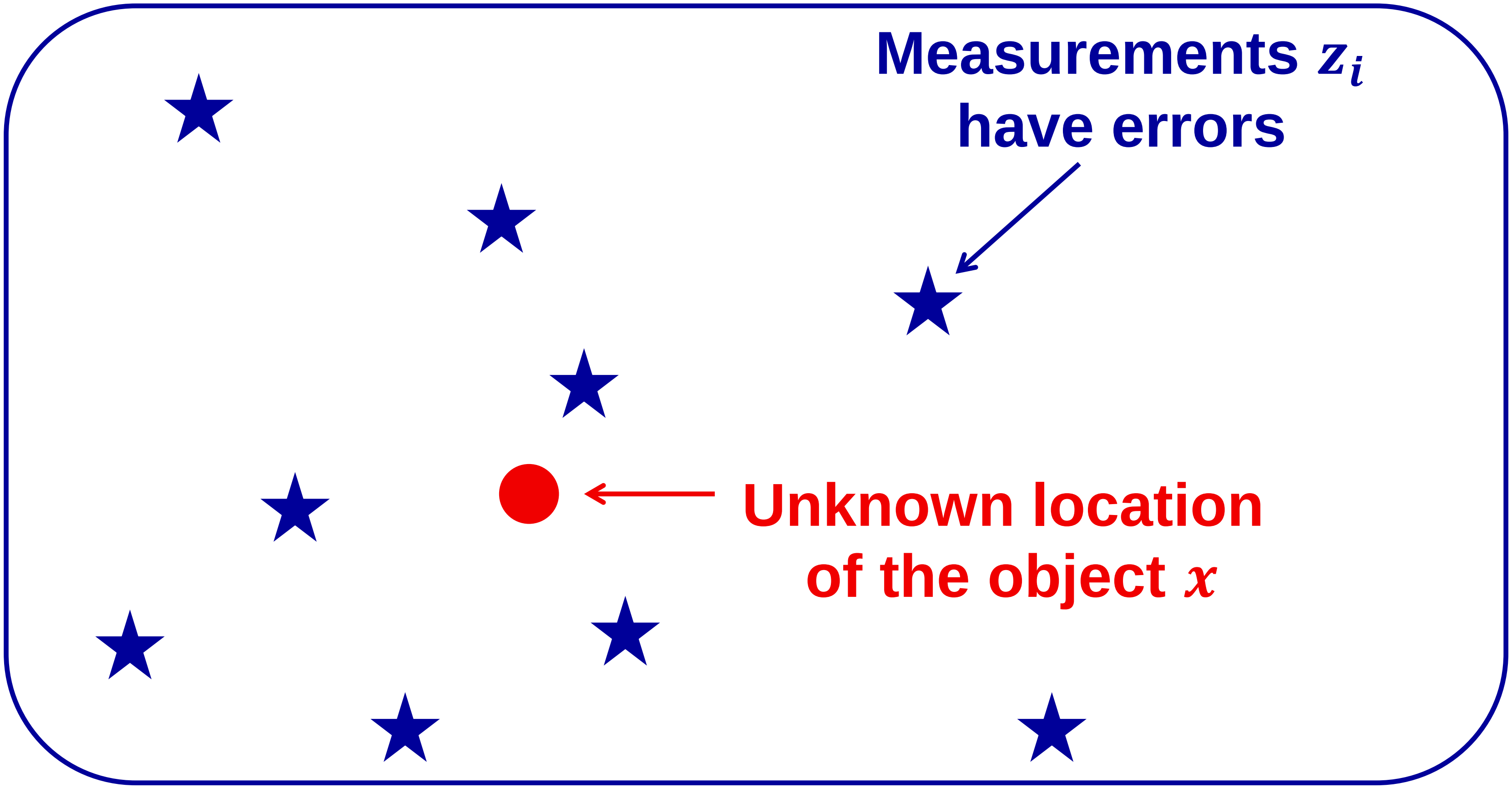


# Estimate the location of an unmoving object




LSM. Example 1

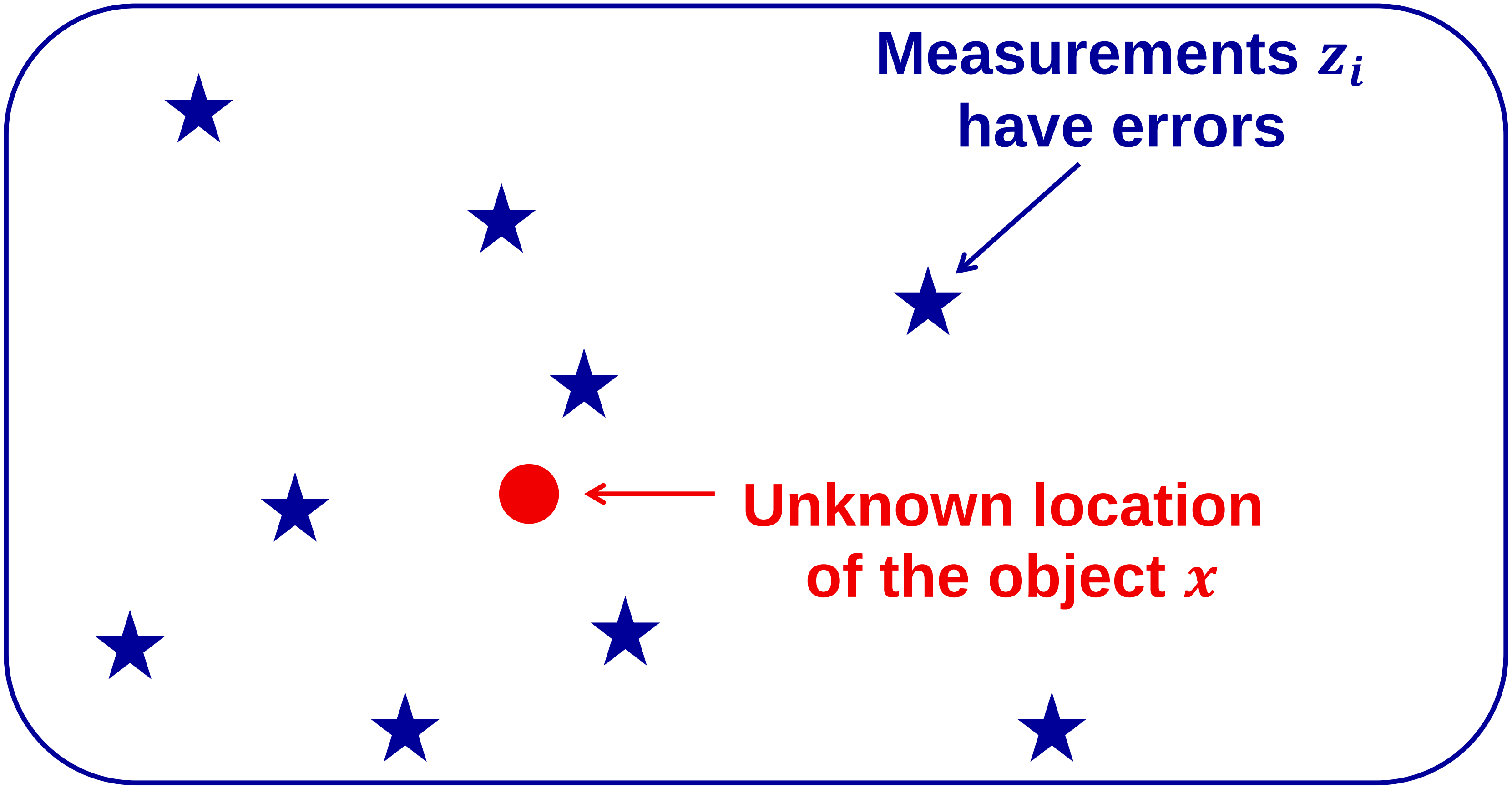
# Estimate the location of an unmoving object




**LSM. Example 1**

# Estimate the location of an unmoving object

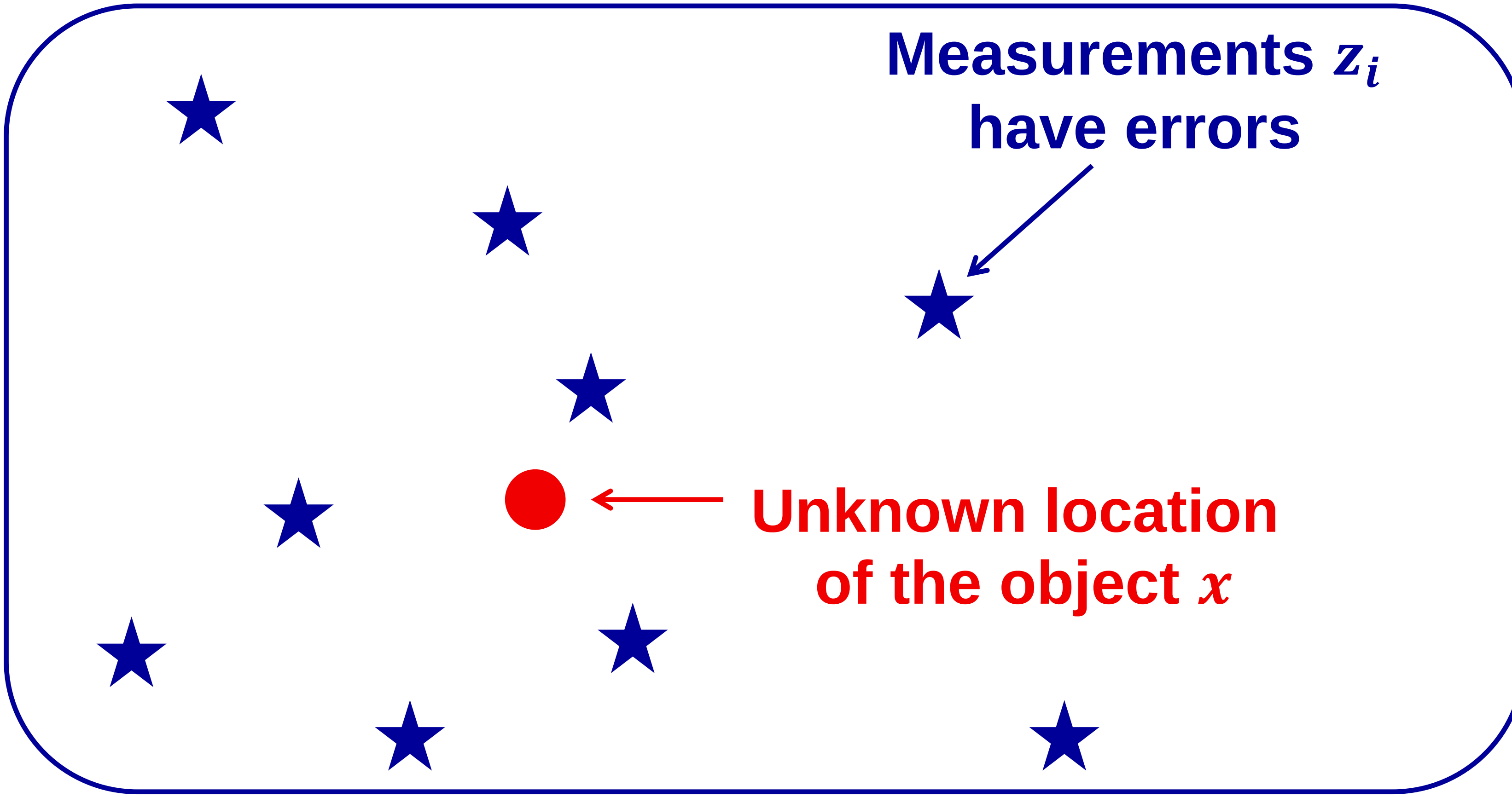
## First use common sense!




LSM. Example 1

# Estimate the location of an unmoving object

## First use common sense!




**LSM. Example 1**

**Solution using common sense: to average measurements**

# Estimate the location of an unmoving object

**Let's find the theoretical ground of this solution using least-square method (LSM)**

**Solution using common sense: to average measurements**

## LSM. Example 1

# Estimate the location of an unmoving object

1 Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (z_i - x)^2$$

$z_i$ – measurements
$x$ – object location
$N$ – number of measurements

## LSM. Example 1

**Solution using common sense: to average measurements**

# Estimate the location of an unmoving object

1 Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (z_i - x)^2$$

$z_i$ – measurements
$x$ – object location
$N$ – number of measurements

2 First derivative in respect to $x$

$$\frac{dJ}{dx} = -2\sum_{i=1}^{N} (z_i - x) = 0$$

## LSM. Example 1

Solution using common sense: to average measurements

# Estimate the location of an unmoving object

① Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (z_i - x)^2$$

$z_i$ – measurements
$x$ – object location
$N$ – number of measurements

② First derivative in respect to $x$

$$\frac{dJ}{dx} = -2\sum_{i=1}^{N} (z_i - x) = 0$$

③ The estimate of location

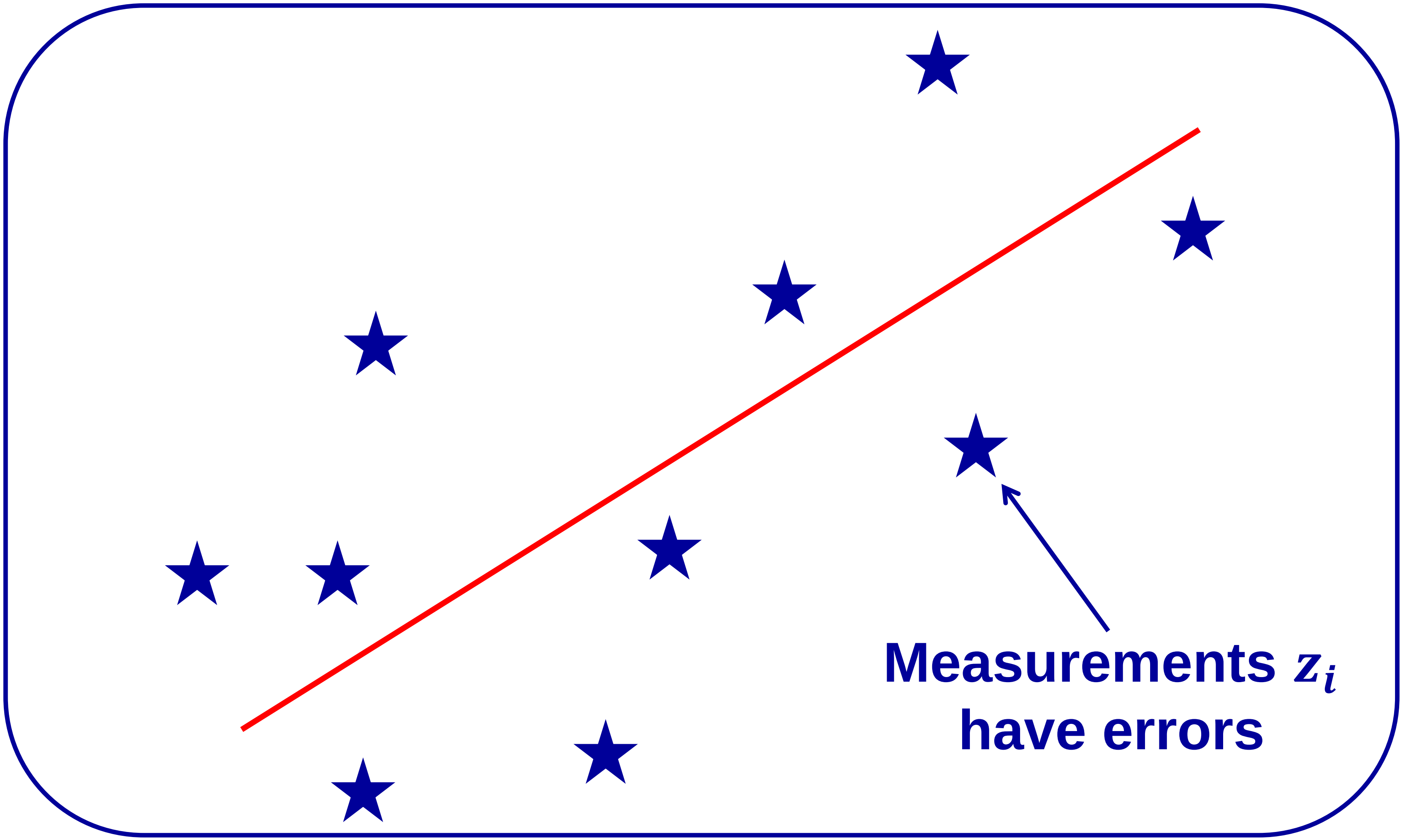
$$\hat{x} = \frac{1}{N}\sum_{i=1}^{N} z_i$$

LSM provides BLUE estimate: Best Linear Unbiased Estimator

Solution using common sense: to average measurements

**LSM. Example 1**

# Estimate the location of an moving object




**LSM. Example 2**

**Uniform and linear movement**

# Estimate the location of an unmoving object

1

**Motion equation**

$$x_i = x_0 + V \cdot iT$$

$x_0$ – initial position
$V$ – velocity
$i$ – time step
$T$ – time interval between measurements

**LSM. Example 2**

**Uniform and linear movement**

# Estimate the location of an unmoving object

1 Motion equation

$$x_i = x_0 + V \cdot iT$$

$x_0$ – initial position
$V$ – velocity
$i$ – time step
$T$ – time interval between measurements

2 Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (z_i - x_0 - V \cdot iT)^2$$

**LSM. Example 2**

**Uniform and linear movement**

# Estimate the location of an unmoving object

$x_0$ – initial position
$V$ – velocity
$j$ – time step
$T$ – time interval between measurements

① Motion equation

$$x_i = x_0 + V \cdot iT$$

② Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (z_i - x_0 - V \cdot iT)^2$$

③ The estimate of location

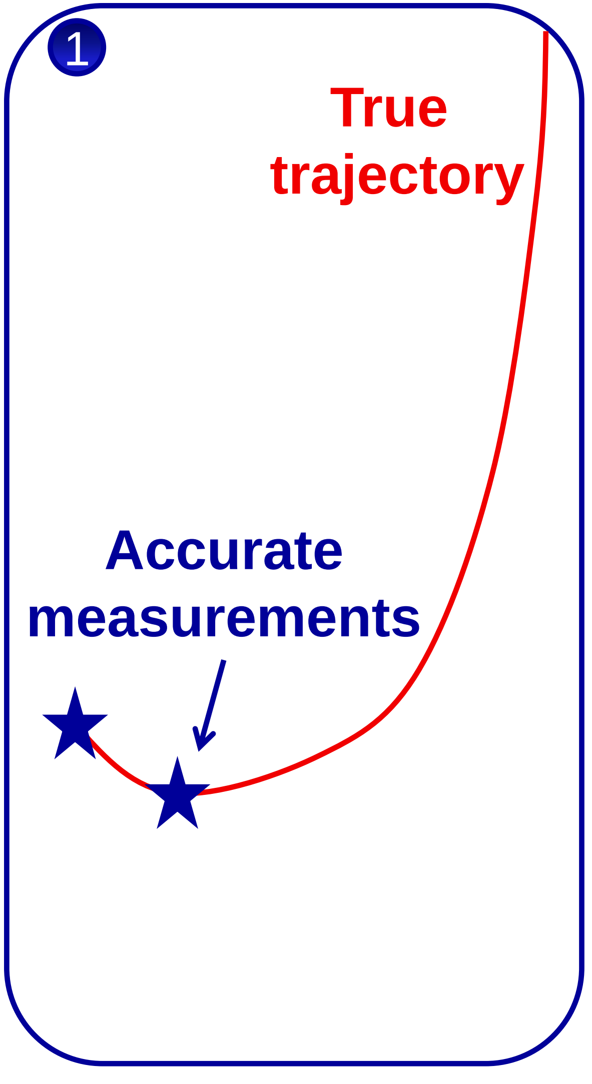
$$\frac{dJ}{dx_0} = 0; \ \frac{dJ}{dV} = 0$$

From these equations we find $x_0$ and $V$

**LSM. Example 2**

**Uniform and linear movement**

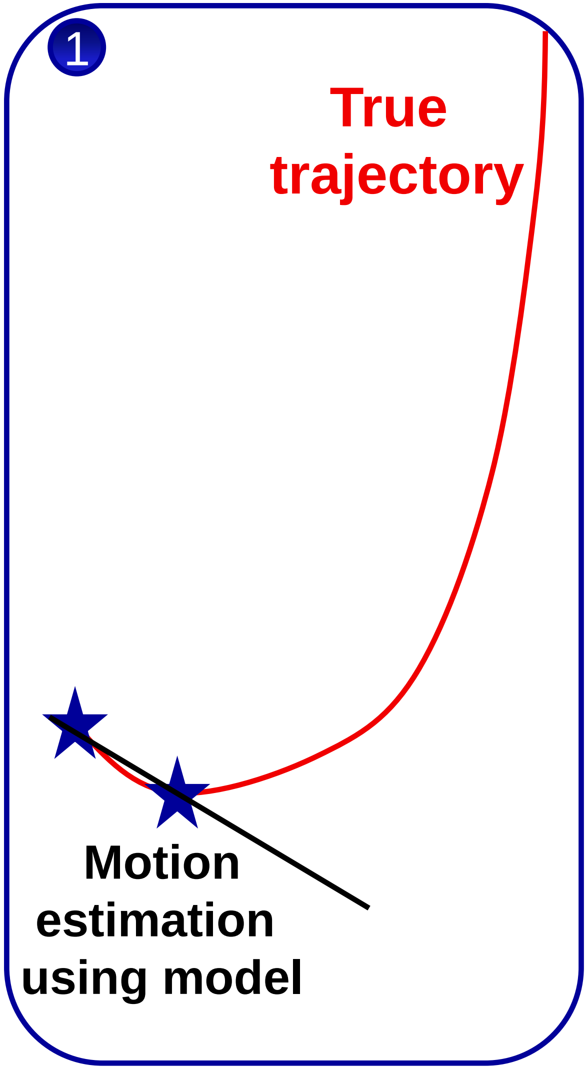
## What happens if we assume the object to move along the line, but in fact it moves along the parabola?




**LSM. Example 3**

**Unknown object dynamics**

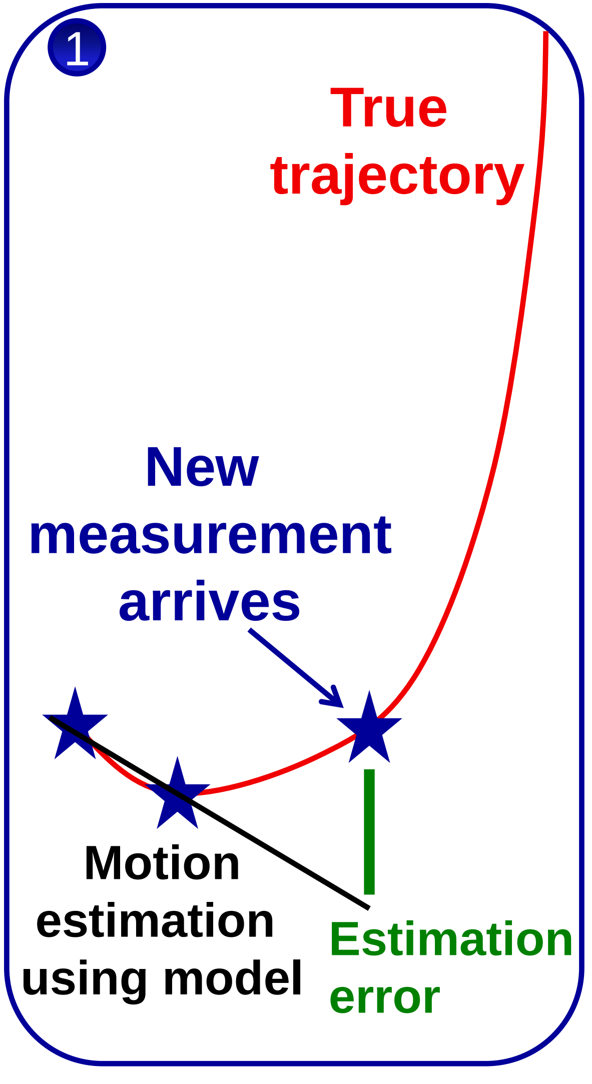
# What happens if we assume the object to move along the line, but in fact it moves along the parabola?




**LSM. Example 3**

**Unknown object dynamics**

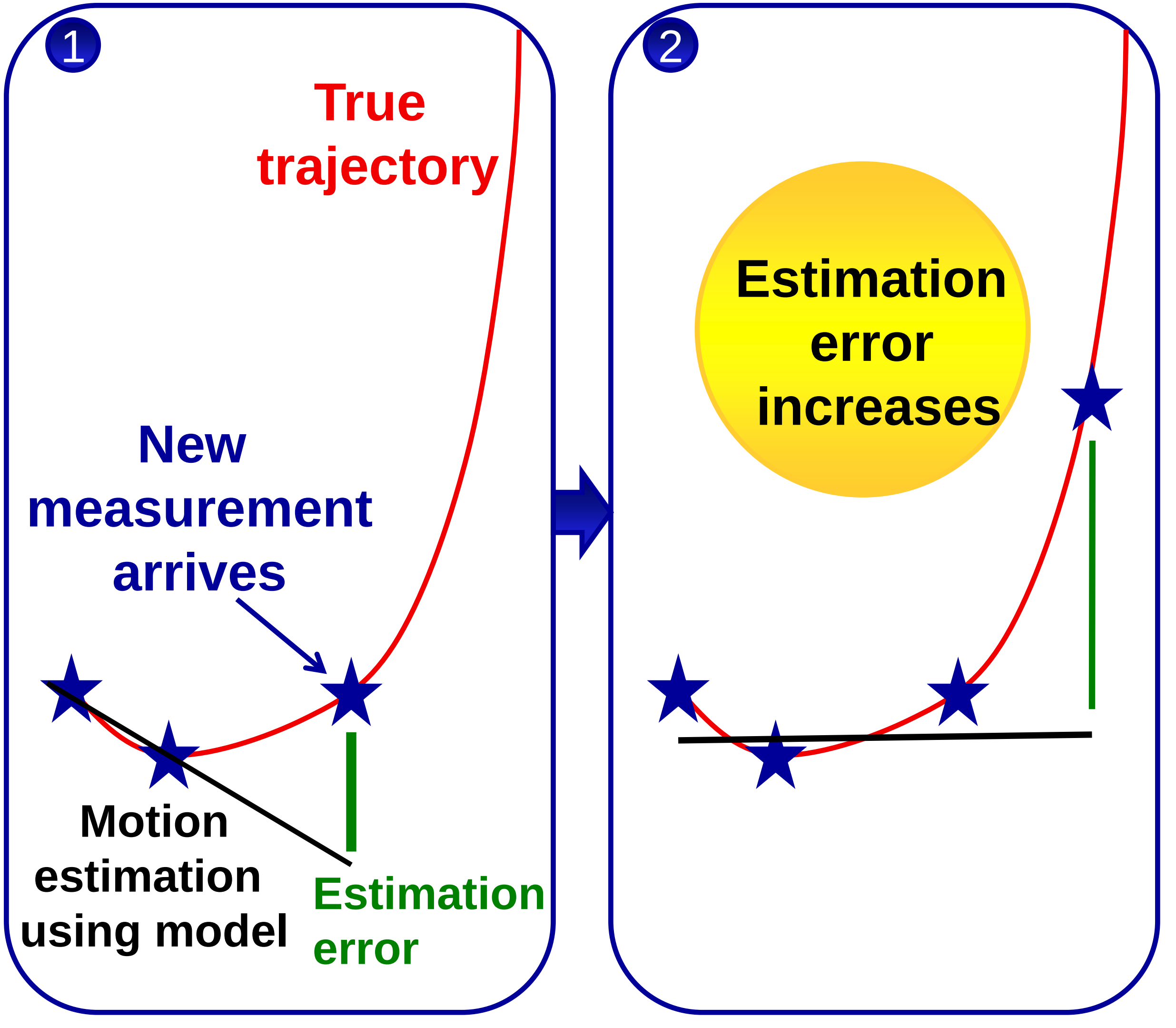
# What happens if we assume the object to move along the line, but in fact it moves along the parabola?




**LSM. Example 3**

**Unknown object dynamics**

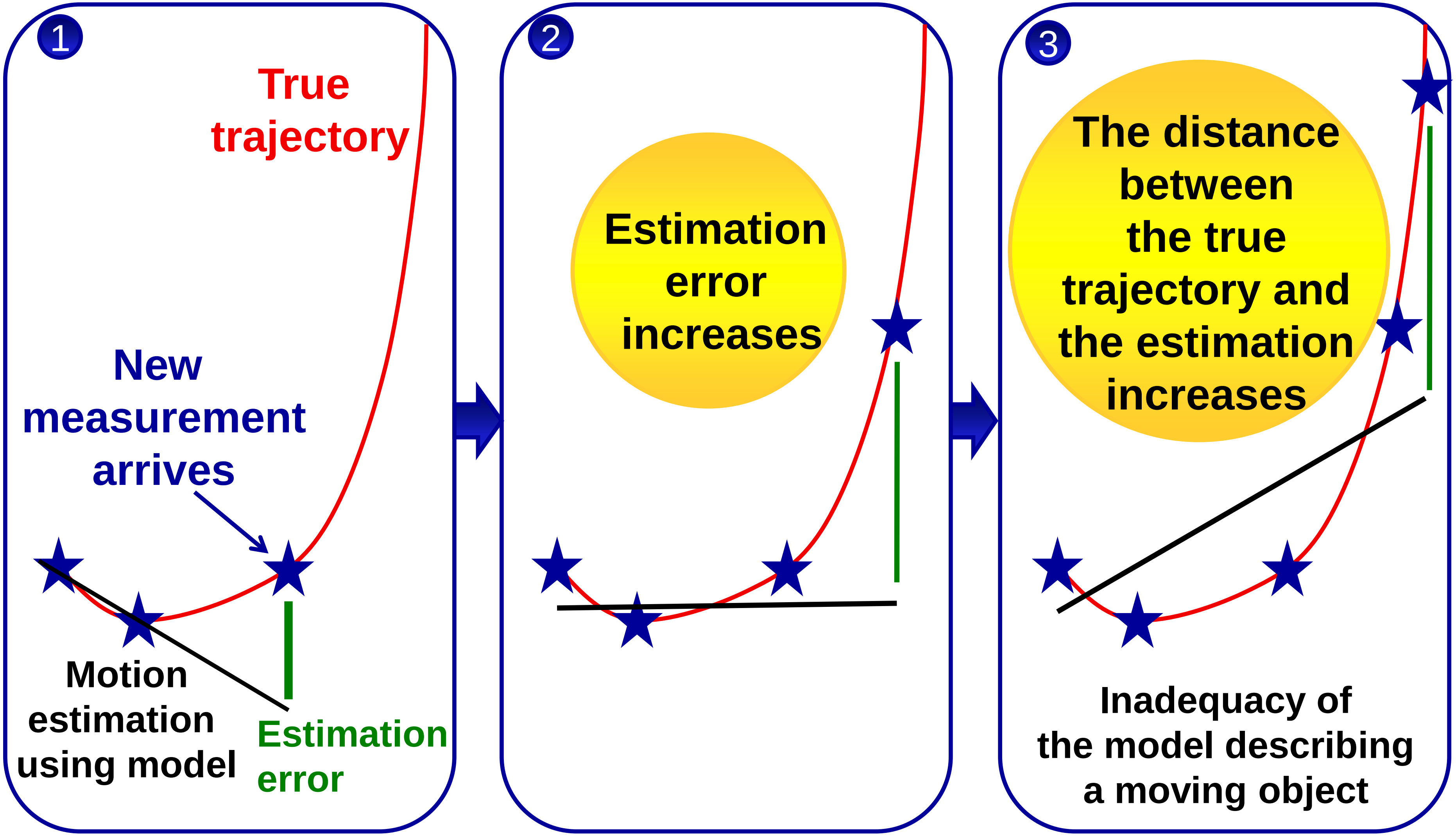
# What happens if we assume the object to move along the line, but in fact it moves along the parabola?




**LSM. Example 3**

**Unknown object dynamics**

## What happens if we assume the object to move along the line, but in fact it moves along the parabola?




# LSM. Example 3

**Unknown object dynamics**

# Conclusion. Advantages and limitations of LSM applications

**Process dynamics is known**

**Advantages**

**LSM provides BLUE estimate: Best Linear Unbiased Estimator**

**Process dynamics is unknown**

**Limitations**

**The LSM method leads to divergence and loses its practical value**

# Conclusion. Advantages and limitations of LSM applications

**Process dynamics is known**

**Advantages**

LSM provides BLUE estimate: Best Linear Unbiased Estimator

**Process dynamics is unknown**

**Limitations**

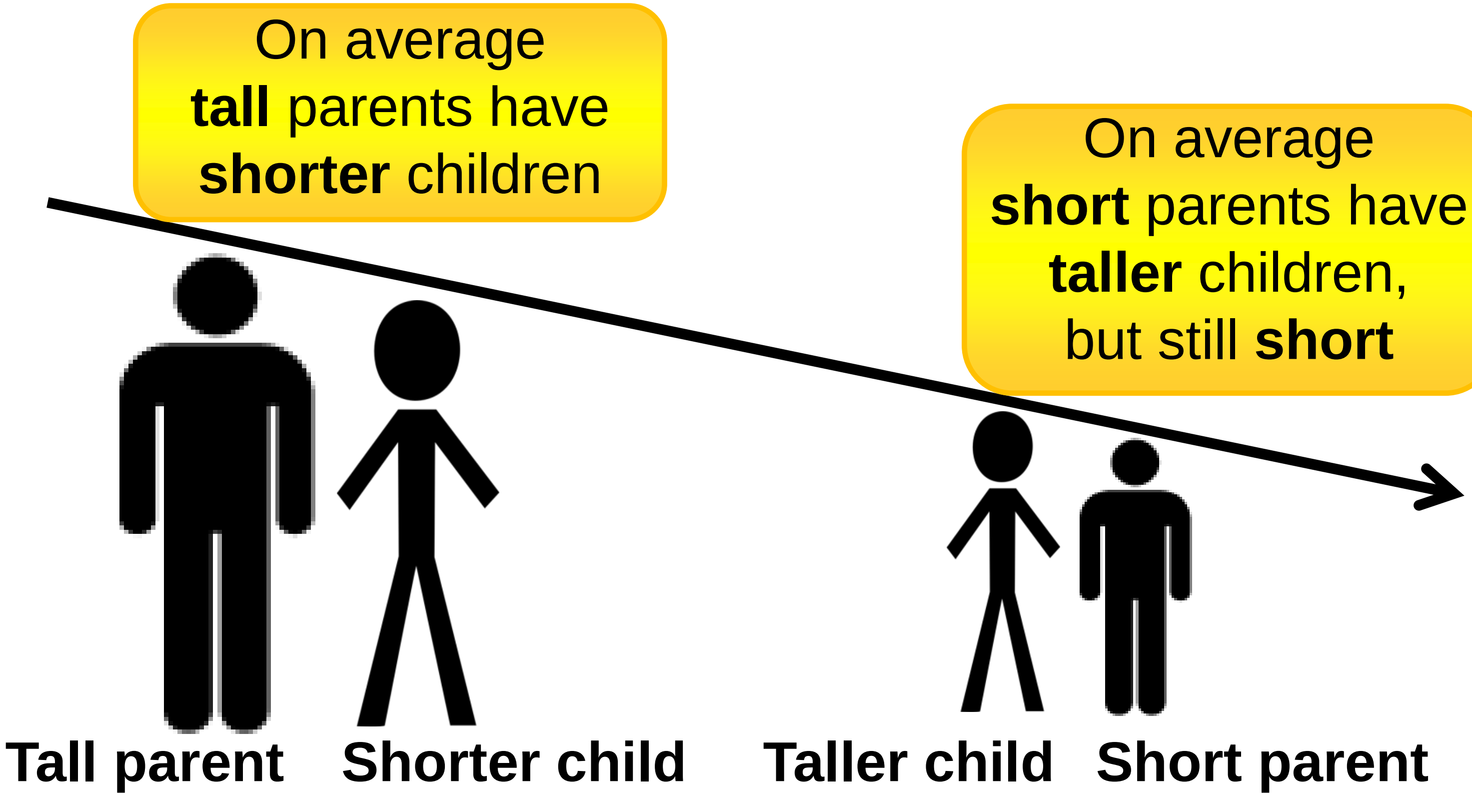
The LSM method leads to divergence and loses its practical value

Therefore in parallel to optimal methods the robust quasi optimal methods are developed

Such methods can be applied in conditions of uncertainty of process dynamics

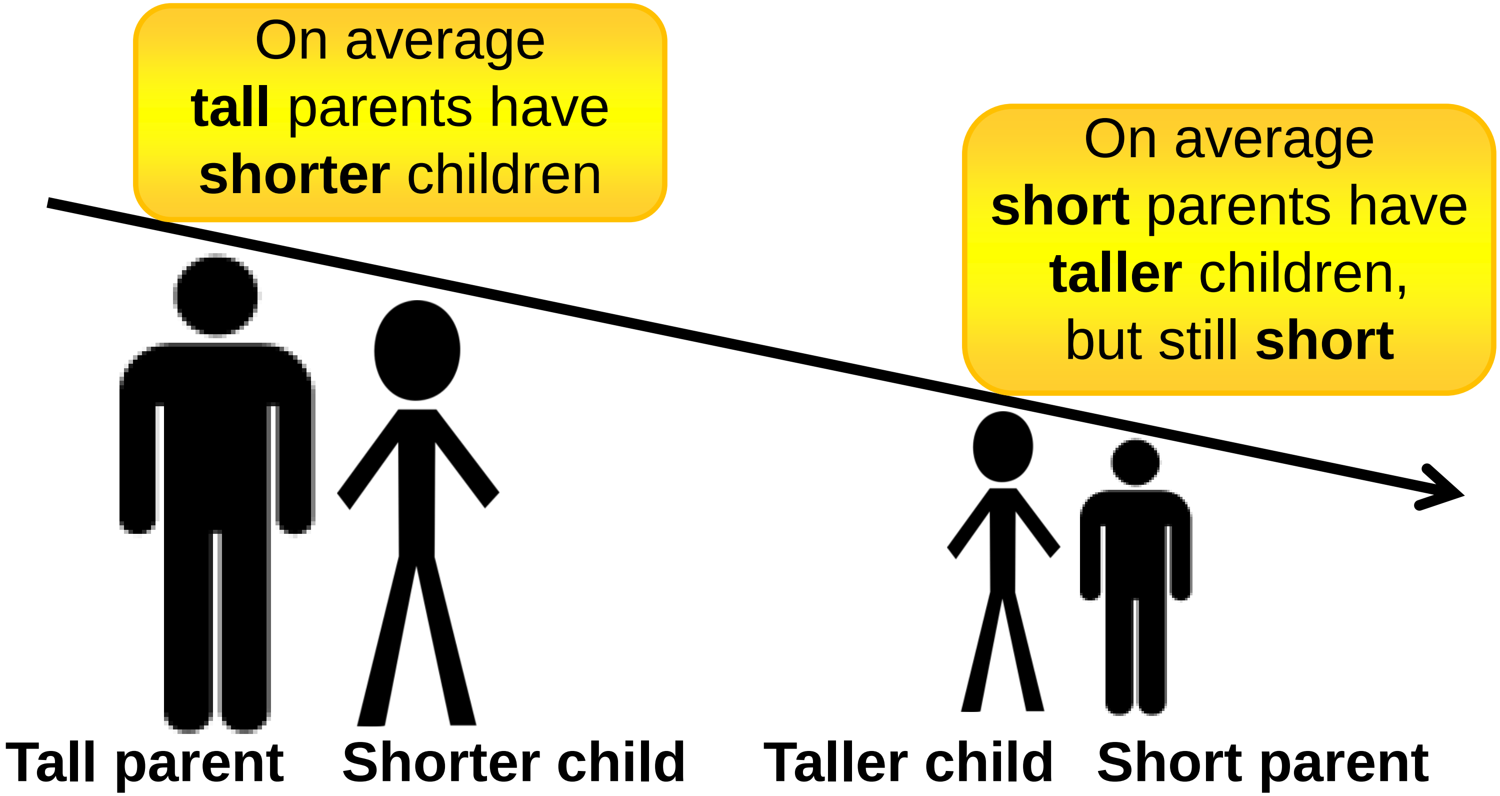
Following topics of the course

# Is there any relationship between the height of parents and children?




**Sir Francis Galton, 1822 –1911**

# Is there any relationship between the height of parents and children?

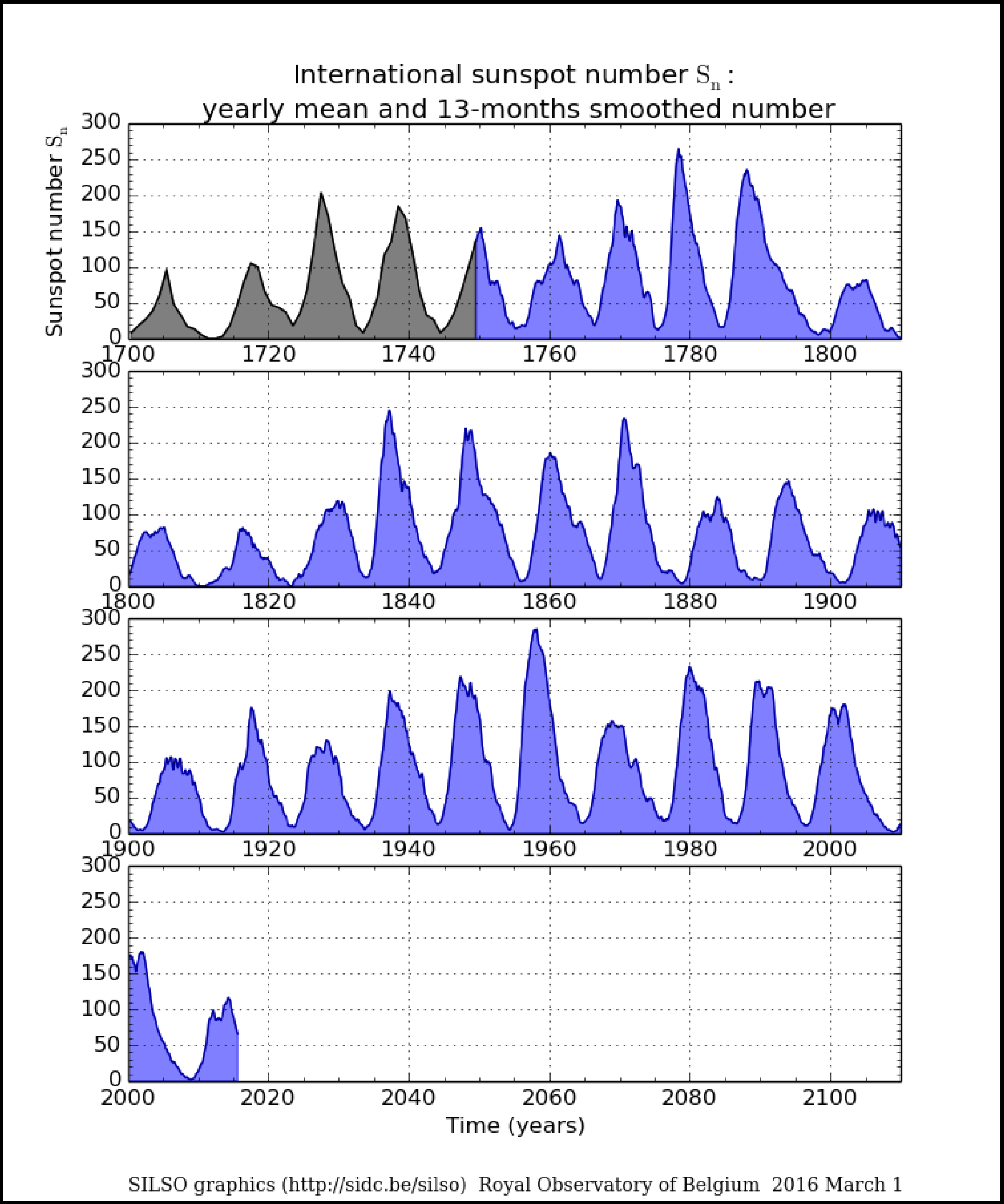



**Height of children approaches to average height of humans**

**REGRESSION**

**Sir Francis Galton, 1822 –1911**

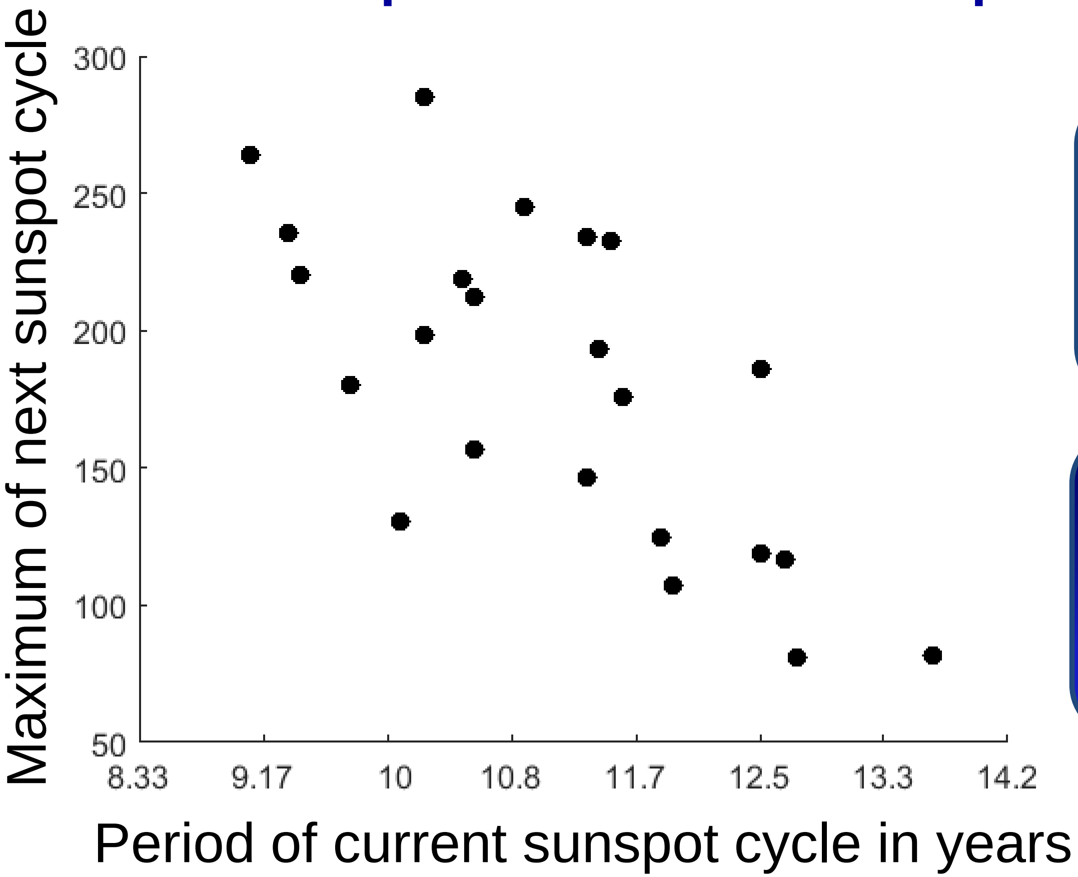
# Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?




**On average a period of sunspot cycle is 11 years. It can be shorter or greater.**

# Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?

**Scatter plot maximum versus period**




**Correlation coefficient -0.7**

**Greater the period of cycle – less the cycle maximum**

# One-dimensional linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$i = 1, \dots . N$

$y_i$
Dependent variable
Regressand

$\beta_0, \beta_1$
Coefficients
$\beta_0$ - intercept
$\beta_1$ - slope

$x_i$
Independent variable
Regressor

$\varepsilon_i$
Unbiased uncorrelated Gaussian noise with constant variance

# One-dimensional linear regression

1 **Motion equation**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$y_i$ – the maximum of next sunspot cycle

$x_i$ – the period of current sunspot cycle

Let's minimize $\sum_{i=1}^{N} \varepsilon_i^2$

$\beta_0, \beta_1$ unknown

2 **Let's minimize functional $J$**

$$J = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

3 **Estimation of $\beta_0, \beta_1$**

$$\frac{dJ}{d\beta_0} = 0; \quad \frac{dJ}{d\beta_1} = 0$$

# Determine coefficients $\beta_0$ and $\beta_1$ using LSM

Let's minimize functional $J$

$$J = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

1

$$\frac{dJ}{d\beta_0} = -2\sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) = 0$$

2

$$\frac{dJ}{d\beta_1} = -2\sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0$$

1

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) = 0$$

2

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0$$

Excursus: solution

# Determine coefficients $\beta_0$ and $\beta_1$ using LSM

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i) = 0$$ 1.1

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \beta_0 - \sum_{i=1}^{N} \beta_1 x_i = 0$$ 1.2

$$\frac{dJ}{d\beta_0} = \sum_{i=1}^{N} y_i - N\beta_0 - \beta_1 \sum_{i=1}^{N} x_i = 0$$ 1.3

$$\beta_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i\right)$$ 1.4

Excursus: solution

# Determine coefficients $\beta_0$ and $\beta_1$ using LSM

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \quad (2.1)$$

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \beta_0 x_i - \sum_{i=1}^{N} \beta_1 x_i^2 = 0 \quad (2.2)$$

$$\frac{dJ}{d\beta_1} = \sum_{i=1}^{N} y_i x_i - \beta_0 \sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \quad (2.3)$$

**Let's substitute value of $\beta_0$ from Equation 1.4 to Equation 2.3**

**Excursus: solution**

# Determine coefficients $\beta_0$ and $\beta_1$ using LSM

$$\sum_{i=1}^{N} y_i x_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i\right)\sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \qquad 2.4$$

$$\sum_{i=1}^{N} y_i x_i - \frac{1}{N}\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i + \frac{1}{N}\beta_1\left(\sum_{i=1}^{N} x_i\right)^2 - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \qquad 2.5$$

$$\beta_1\left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right] = \frac{1}{N}\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i \qquad 2.6$$

$$\beta_1 = \left[\frac{1}{N}\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i\right] \Big/ \left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right] \qquad 2.7$$

Excursus: solution

# Determine coefficients $\beta_0$ and $\beta_1$ using LSM

$$\beta_1 = \left[\frac{1}{N}\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i\right] \Big/ \left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right]$$

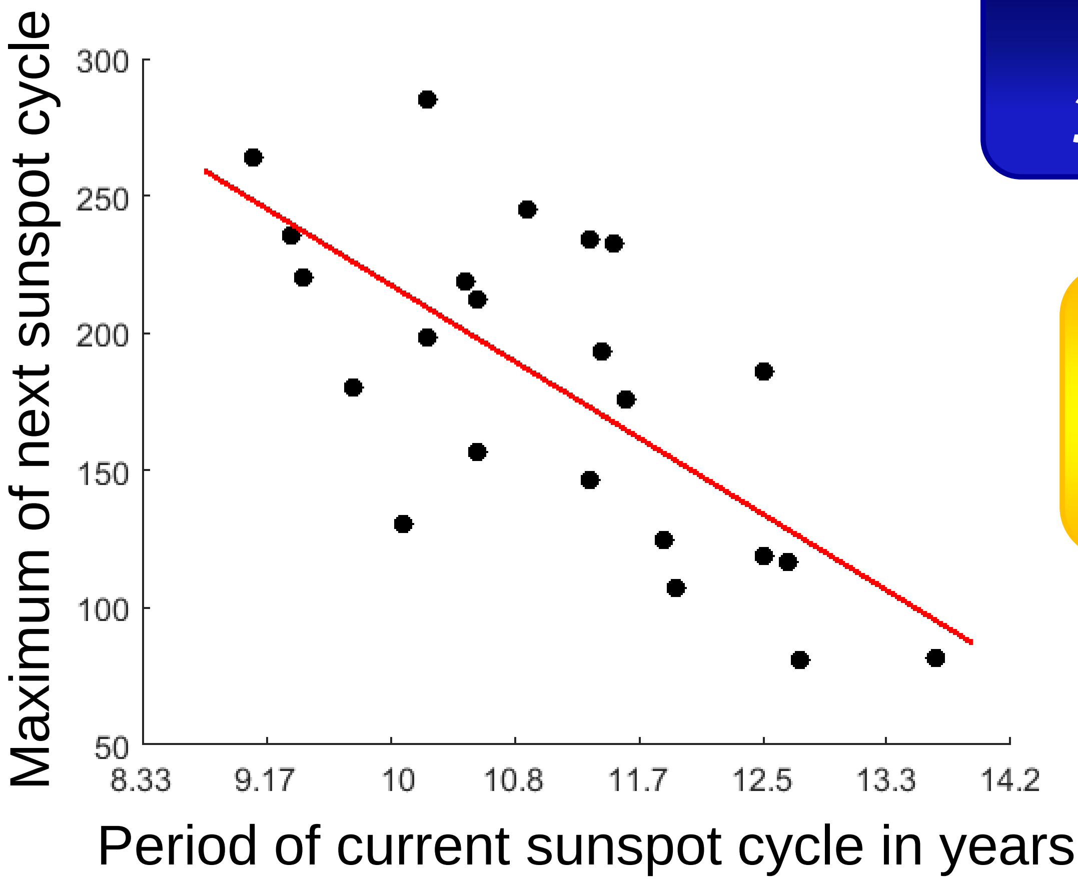
$$\beta_0 = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i\right)$$

$\beta_0 =$ 551.4

$\beta_1 =$ -2.8

Excursus: solution

**Is there any relationship between period of current sunspot cycle and the maximum of next sunspot cycle?**




**Linear regression**

$$y_i = 551.4 - 2.8x_i$$

**Thus, current cycle contains information on future sunspot activity**

# Precursor techniques to predict the next 11-year sunspot cycle strength

Extraction of useful knowledge from current sunspot cycle to predict future sunspot activity

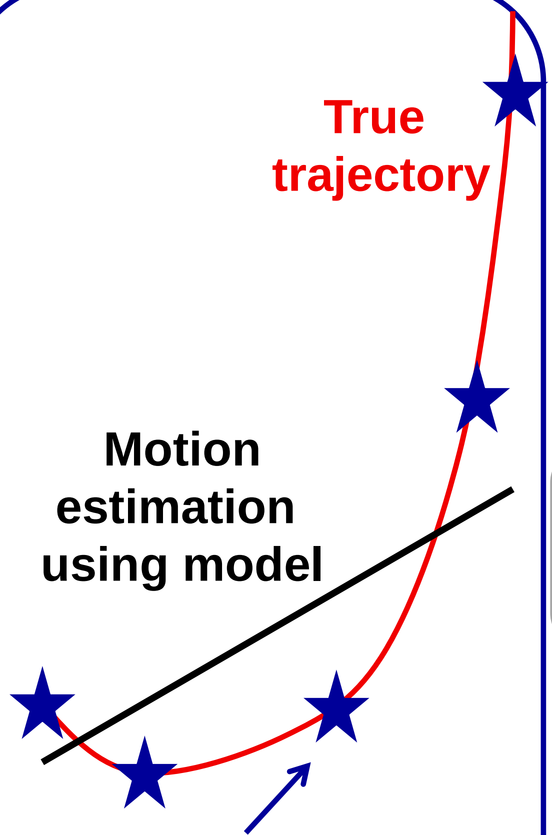
Example

Regressand:
Next sunspot cycle maximum

Regressors:

- Cycle minimum
- Geomagnetic index around its minimum
- Number of geomagnetically disturbed days
- Reversed polar field

# Example of inadequate regression model




**Measurements**

$$z_i = x_i + \varepsilon_i$$

$$x_i = x_{i-1} + VT$$

We assume the object to move uniformly

$$x_i = x_{i-1} + V_{i-1}T + \frac{aT^2}{2}$$

But in fact it is uniformly accelerated motion

$$z_i = x_0 + V \cdot iT + \varepsilon_i$$

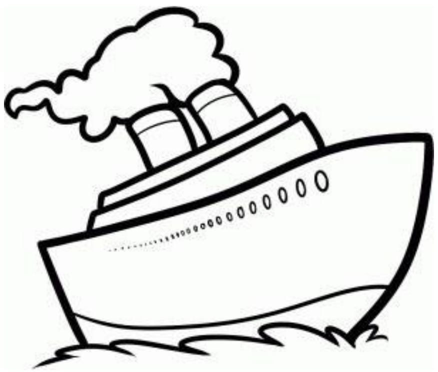
No practical value

Thus, regression model is inadequate

# How to take into account the unintentional maneuver?

**Motion model**

$$x_i = x_{i-1} + VT$$



**However how to take into account unintentional maneuver to track moving object (i.e., airplane turbulence, ship pitching or undercurrents)?**

# How to take into account the unintentional maneuver?

**Motion model**

$$x_i = x_{i-1} + V_{i-1}T + \frac{a_{i-1}T^2}{2}$$

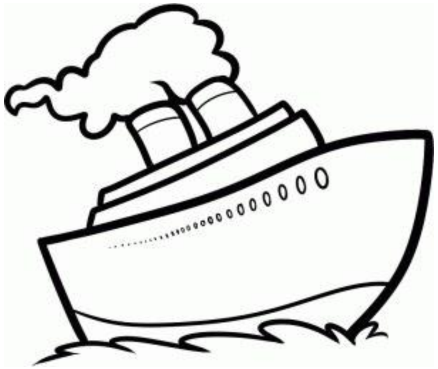
**However how to take into account unintentional maneuver to track moving object (i.e., airplane turbulence, ship pitching or undercurrents)?**

**Unintentional maneuver can be described by random acceleration $a$**

# Process noise should not be filtered

Motion model

$$x_i = x_{i-1} + V_{i-1}T + \frac{a_{i-1}T^2}{2}$$



However how to take into account unintentional maneuver to track moving object (i.e., airplane turbulence, ship pitching or undercurrents)?

Unintentional maneuver can be described by random acceleration $a$

$$\frac{a_i T^2}{2}$$

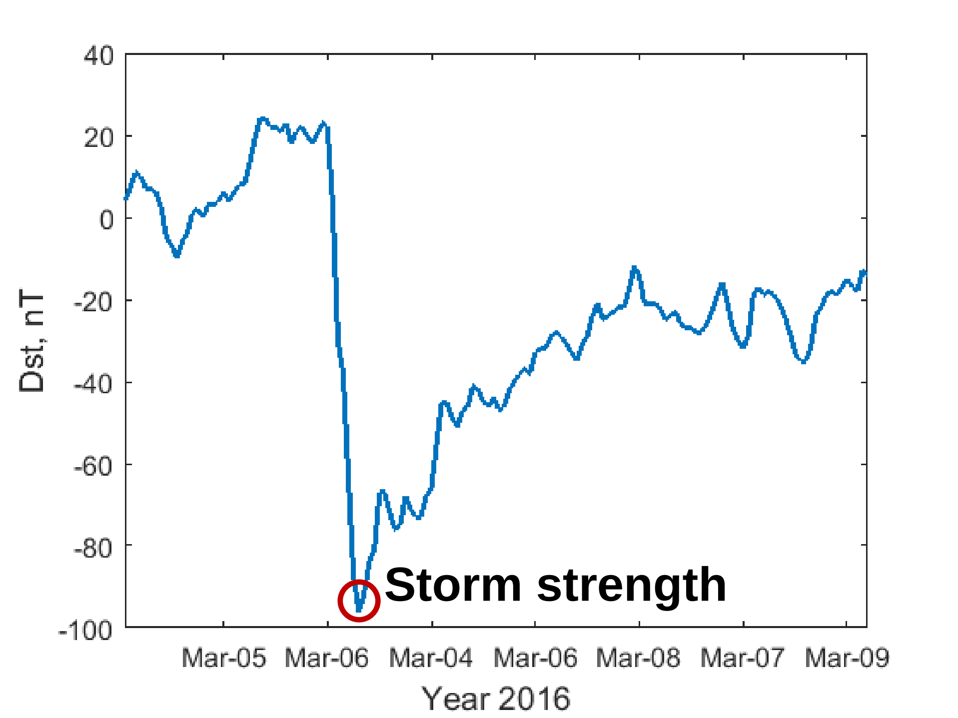
Noise intrinsic to the process itself that should not be filtered

However linear regression doesn't separate process noise and measurement noise and thus loses its practical value

# Multi-dimensional linear regression

**Dependence on multiple regressors**

**Example: Disturbance storm time (Dst) geomagnetic index**




**Dst is sensitive to:**

1. **Solar wind speed**
2. **Southward component of Interplanetary magnetic field (IMF)**
3. **Previous Dst values**
4. **Solar wind density and pressure**

# Multi-dimensional linear regression

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{k-1} x_{i,k-1} + \varepsilon_i$$

$$i = 1, \ldots . N$$

$y_i$
**Dependent variable Regressand**

$\beta_j$
**Coefficients of regression**

$x_{i,j}$
**Independent variable Regressor**

$\varepsilon_i$
**Unbiased uncorrelated Gaussian noise with constant variance**

**Coefficients $\beta_j$ are determined by LSM**

$$\sum_{i=1}^{N} \varepsilon_i^2 \to min$$

# Multi-dimensional linear regression

$$Y = \begin{vmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{vmatrix}$$ → Vector of dependent variables

$$\beta = \begin{vmatrix} \beta_0 \\ \beta_k \\ \dots \\ \beta_{k-1} \end{vmatrix}$$ → Vector of coefficients

$$X = \begin{vmatrix} 1 & x_{1,1} & \dots & x_{1,k-1} \\ 1 & x_{2,1} & \dots & x_{2,k-1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{N,1} & \dots & x_{N,k-1} \end{vmatrix}$$ → Matrix of independent variables

$$\varepsilon = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{vmatrix}$$ → Vector of random errors

**Linear regression in matrix form** → $\boldsymbol{Y = X \cdot \beta + \varepsilon}$ → $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Y}$

**Linear Regression Analysis,**
**G.A.F. Seber and J. Lee, Wiley, N.Y., 2003**

# Estimation error of coefficients $\beta$

**Covariance matrix of estimation error**

$$cov(\hat{\beta}) = \sigma_\varepsilon^2 (X^T X)^{-1}$$

**Variance estimation of random noise**

$$\sigma_\varepsilon^2 = \frac{1}{N-k} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

**Weighted LSM**

$$\sum_{i=1}^{N} \frac{\varepsilon_i^2}{\sigma_{\varepsilon_i}^2} \rightarrow min$$

# Main problems of applying LSM and linear regression

**LSM**
**If process dynamics is unknown**

The LSM method leads to divergence and loses its practical value

**Linear regression**

No separation between process noise and measurement noise and thus no practical value

**Following topics of the course overcome these problems**