Data Aggregation and Sampling (DAS)

2IAB1

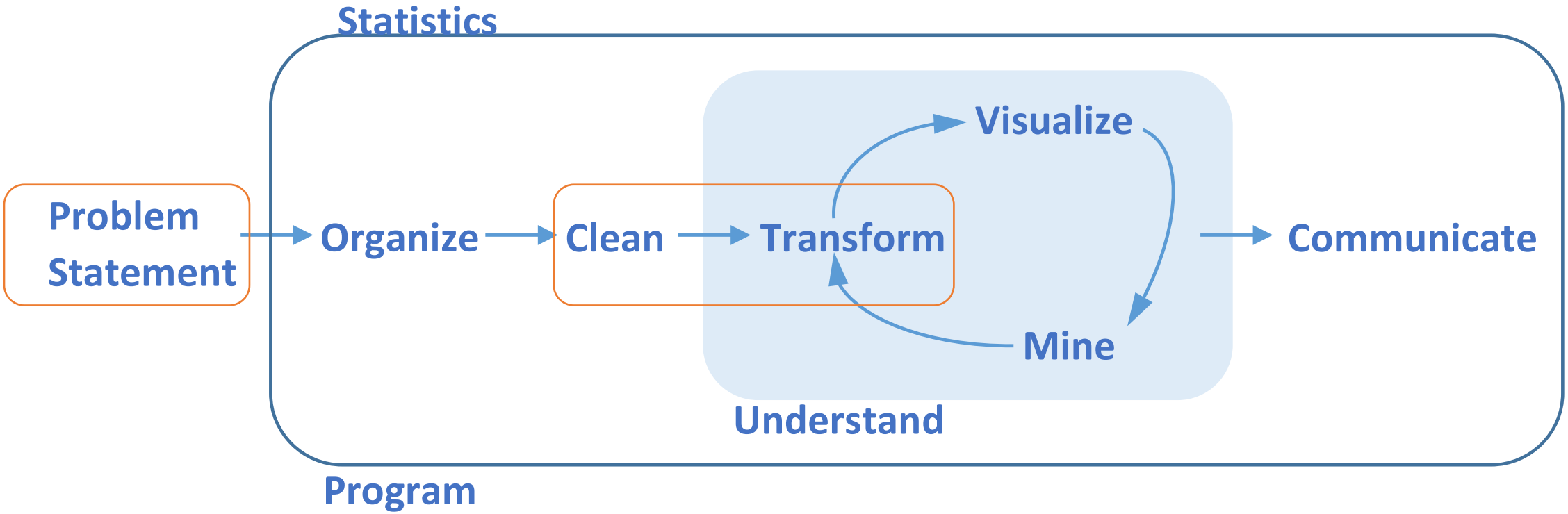
Week 5

Foundations of Data Analytics

2023-2024

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Today


Statistics
Problem Statement
Organize
Clean
Transform
Visualize
Mine
Understand
Communicate
Program

Overview of the lecture & learning objectives

1. Scientific method for studying a problem statement, or research question (on illustrative examples)
2. Data sampling
3. Data cleaning and preprocessing
4. Features for time series

Case study

K. Nieuwenhuizen, D. Aliakseyeu and J.B. Martens, Insight into Goal-directed Movement Strategies, Computer-Human Interaction (CHI) 2010 Conference, 883-886
See https://doi.org/10.1145/1753326.1753457

Target size matters when (reaction) time is important

Source: https://www.angieslist.com/research/car-brakes/

brake pedal

accelerator pedal

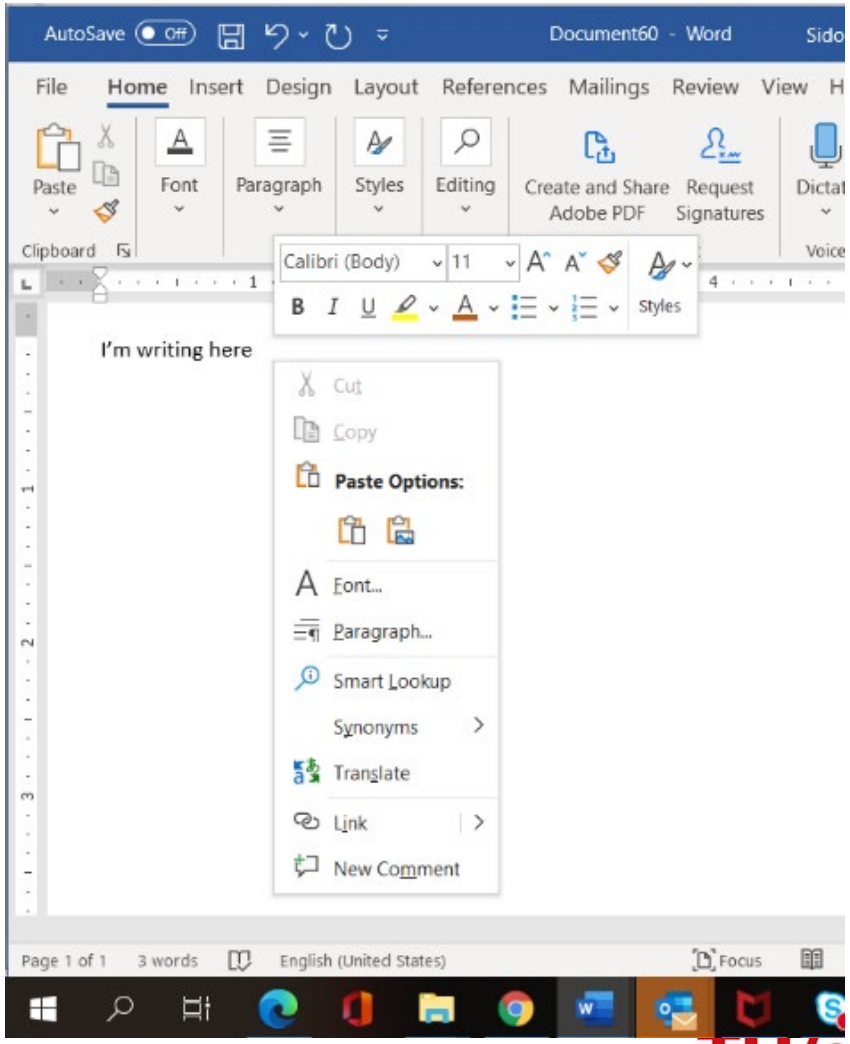
Source: https://www.dreamstime.com

emergency stop

restart

Consequences for computer interfaces

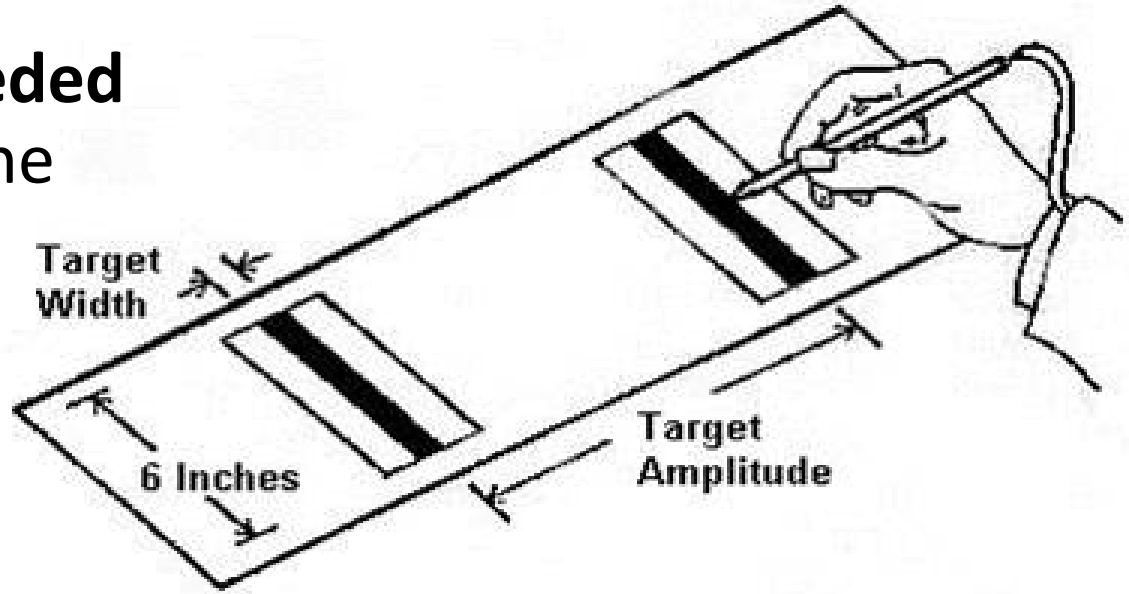
- Choosing the button size:
  - Make important buttons large
  - Use corners as they are equivalent to large buttons (you cannot escape)
- Reduce the distance:
  - Put important menus under your mouse (right-click context menu)
- Most people move the mouse from top left to bottom right → let the user use that movement



Original (1D) Fitts' law experiment

- Classic paper (from 1954) in the field of movement science;

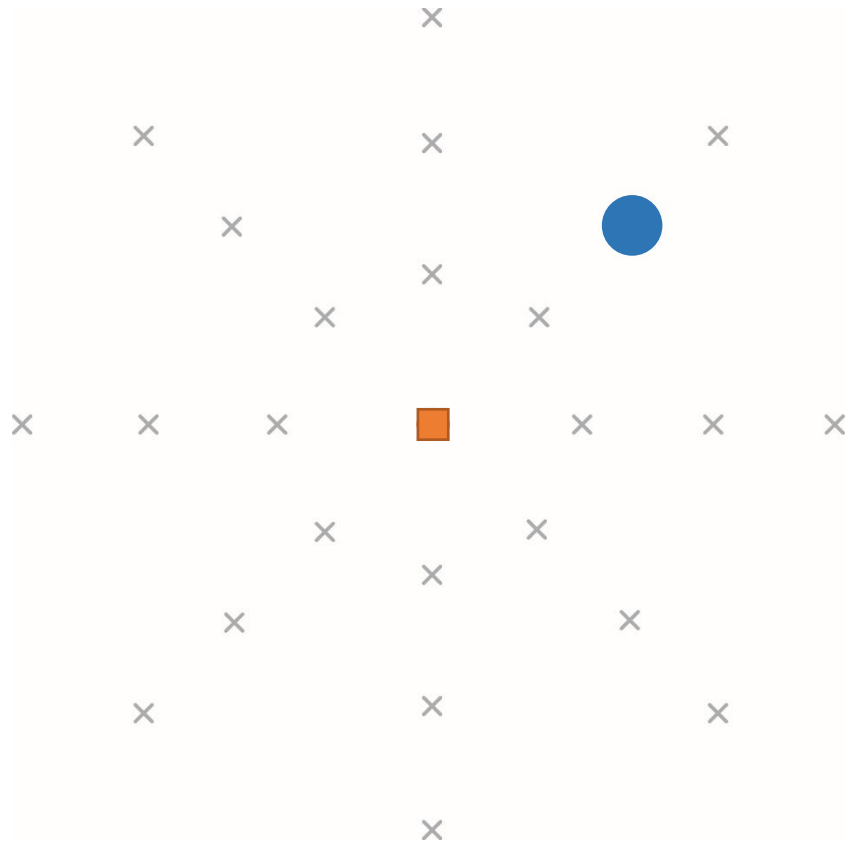
  Fitts proposed to use a standard task to measure (human) motoric skill
- Interaction task: select target with **width W** positioned at a **distance D** from the current position
- Fitts' law states that the **average time needed** to select the target is **linearly related to** the index-of-difficulty $ID = log\left(1 + \frac{D}{W}\right)$
- Supported by empirical evidence



A mouse experiment

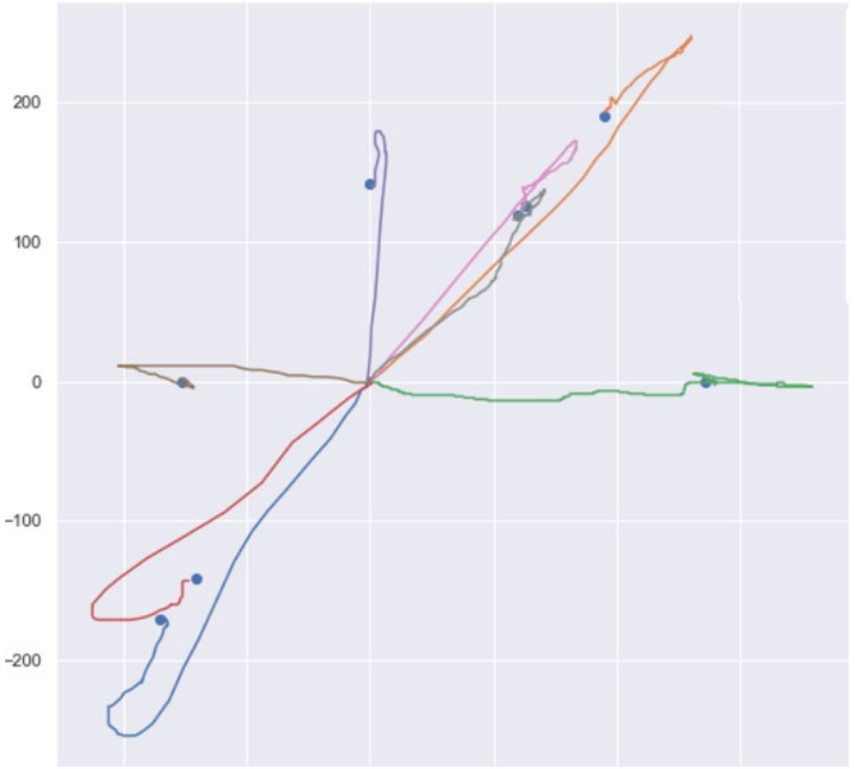
Based on a standardized task in Human-Computer interaction

- Using a mouse or a trackpad, move onto a square shown in the center of the window
- After a random delay of 2 to 4 seconds a target disc appears
  - in one of the eight possible directions (-135^0, -90^0, …, 135^0, 180^0 to the X-axis)
  - at a distance chosen from at random from 100 to 290 pixels
  - the target radius is 3, 6 or 9 pixels (chosen randomly)
- Move towards the blue disc and click on it



Data collected

- User characteristics (right-handed?, the major study program,…)
- Characteristics of the mouse, operating system, etc.
- Trial characteristics, like
  - the trial number for this user (there might be a learning curve)
  - delay
  - target radius
  - target position
  - …
- Paths
  - timestamp with the corresponding position for each trial



From research question to testable hypotheses

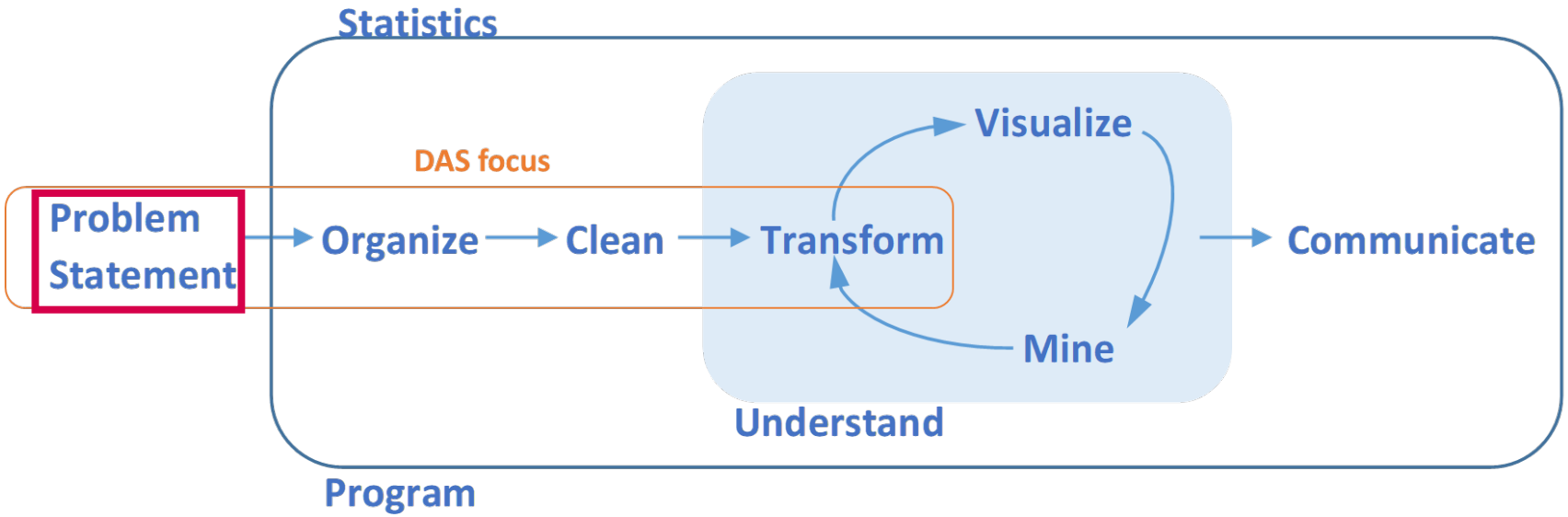
Research question examples

Air quality and traffic data

- *What is the relationship between air quality and traffic?*

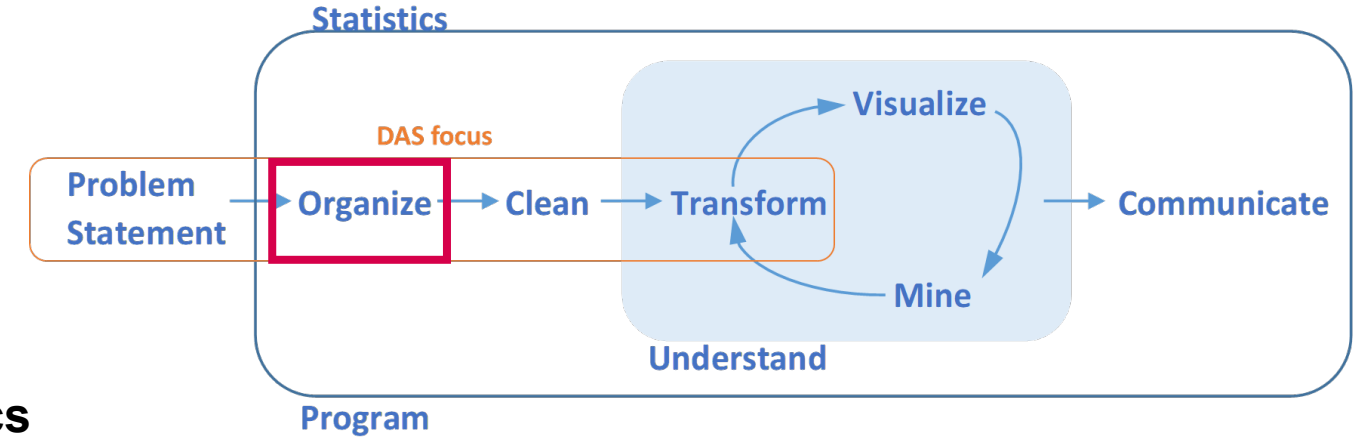
Mouse experiment

- *How does the task difficulty influence the reaction time of the user?*



What data can we collect / do we get to answer the research question?

- **Primary data:** collected by you / your team (like in the mouse experiment)
- **Secondary data:** collected by others (like in the air quality – traffic data)
- Database schema for the mouse experiment:
  - paths(user, trial, t, x, y), where t is a time stamp
  - trials(user, trial, delay, input_method, mouse_acceleration, target_radius, target_x, target_y, …)
  - users(user, major, gender, right_handed, use_tue_laptop, …)



Formulate a hypothesis

First attempt:

- The reaction time is larger if the task is more difficult.

Questions about it:

1. *What is exactly meant with the reaction time?*
2. *How is the task difficulty defined/measured?*

Reaction time and *task difficulty* are examples of **features**

Feature

- An **object** can be represented in data **as a vector of features**
- Feature is a **measurable** property or characteristic
- Some features are **directly gathered** during an experiment,
  e.g. *target width, dominant hand*
- Some features are **computed** based on the collected data,
  e.g. *reaction time, length of a user's path from a source to a target*

Hypothesis refinement process:
1. Ask questions

- paths(user, trial, t, x, y)
- trials(user, trial, delay, input_method, mouse_acceleration, …, target_radius, target_x, target_y,…)
- users(user, major, gender, right_handed, use_tue_laptop, …)

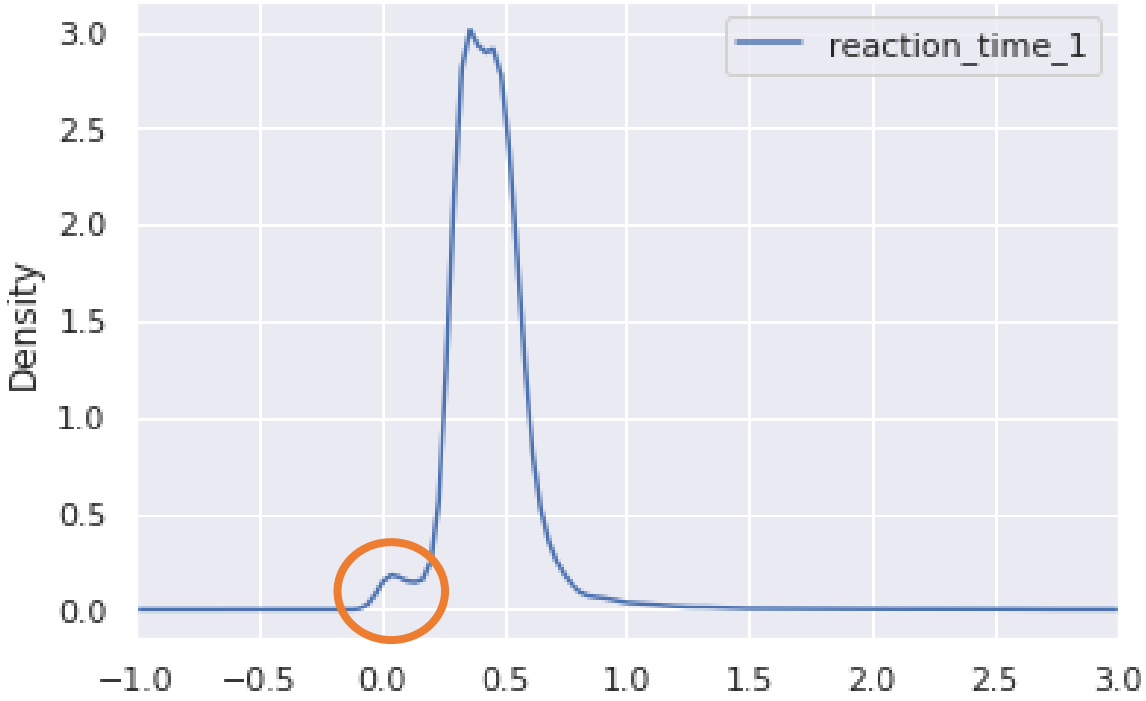
Ask questions! (data-driven and/or based on domain knowledge)

- *Could the task difficulty be related to the moment of target appearance?*
- *Could the task difficulty depend on the target radius? If so, how can we describe this dependency?*
- *Could the position of the target contribute to task difficulty? If so, how?*
- *Could the task difficulty depend on the dominant hand? If so, how?*
- …

Hypothesis refinement process:
2. Answer questions (use data and domain knowledge)

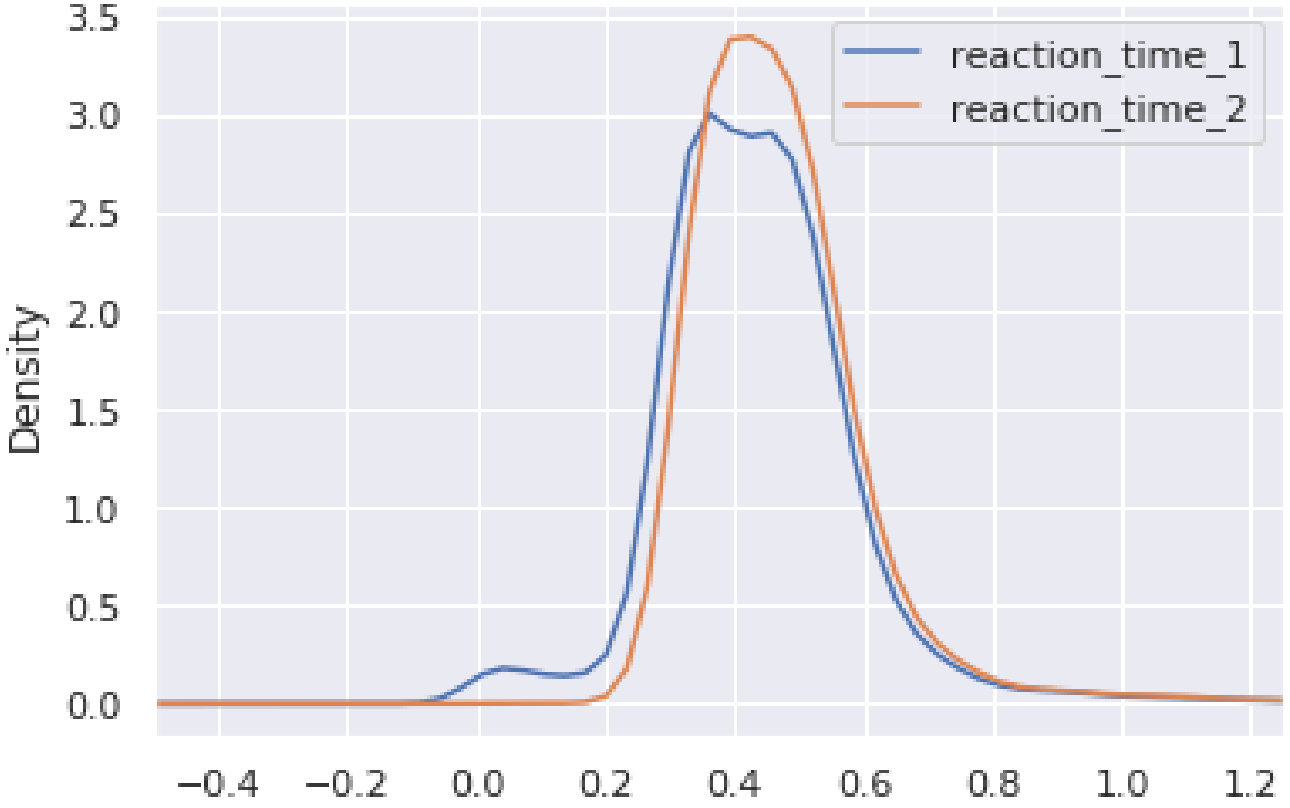
Reaction time definition?

- Definition 1: the reaction time for a trial is the time between target appearance and the first nonzero time stamp in the paths table.
  - *Can reaction times be so close to 0?*
  - *Look at some concrete trials with very small reaction times. What could be an explanation?*



Reaction time: definition attempt 2

- The time from the moment when the target appears (t=0) until the moment the first point outside the circle with radius 10 pixels around the origin (x=0,y=0) is reached.
  - No reaction times close to 0
  - Reaction time according to definition 2 is slightly larger than the one based on definition 1. Possible consequences?



- **Choose your definitions wisely!**
- **Check them on the data!**

Generating features through data aggregation

Air quality and weather data (secondary data)

- You got data that was **generated from raw data** using data aggregation:
  - Hourly precipitation amount (in mm)
  - Mean wind speed (in m/s) during the 10-minute period preceding the time of observation
  - …
- Possible further aggregations:
  - Daily precipitation amount
  - …

Mouse experiment data (primary data)

- You get **raw data that can be aggregated**. Compute e.g.:
  - The length of the path in a trial
  - The length factor: $\frac{path\ length}{distance\ from\ origin\ to\ target}$
  - The average reaction time per user
  - The median reaction time per user
  - …

Summary so far

- There are choices to make on the path from **a high-level research question** to a testable and meaningful hypothesis. Use domain knowledge and data!
- Data can be **primary** or **secondary**
- Features are **measurable** properties or characteristics.
  Clear definitions of features are important!

Scientific Method

Scientific method: a way of thinking

- **Scientific method** is a systematic approach to conducting *empirical research* to obtain *sound* answers to questions
- The scientific (research) method is used to
  - to collect **valid data** for exploratory or confirmatory analysis
  - test a hypothesis or theory (collect **evidence** for a **claim**)
- Empirical (experimental) research characteristics:
  - procedures, methods and techniques that have been tested for **validity** and **reliability**
  - **unbiased** and **objective** (***research attitude***)

Tools for scientific method

- Deduction
- Induction
- Verification by others
- Occam’s razor / Parsimony principle
- Statistics

Deduction vs Induction

Deductive reasoning

Premise 1: Rain droplets collide with airborne particles during free fall.

Premise 2: Today, there has been heavy rainfall in the region.

Premise 3: The more rain, the more traffic.

Premise 4: Cars emit airborne particles.

Conclusion from Premise 1 + Premise 2: The concentration of airborne particles went down today due to the rain.

Conclusion from Premise 3 + Premise 4: There was more cars than usual on the road due to the rainfall, and since they emit airborne particles, the concentration of airborne particles should have increased

If the premises are true, the conclusion is valid.

Have you combined them correctly? Have you derived contradictory statements? Have you missed some pieces of information?

What if one of the premises is not true? (e.g. cars do not use fuel *anymore*)

Inductive reasoning

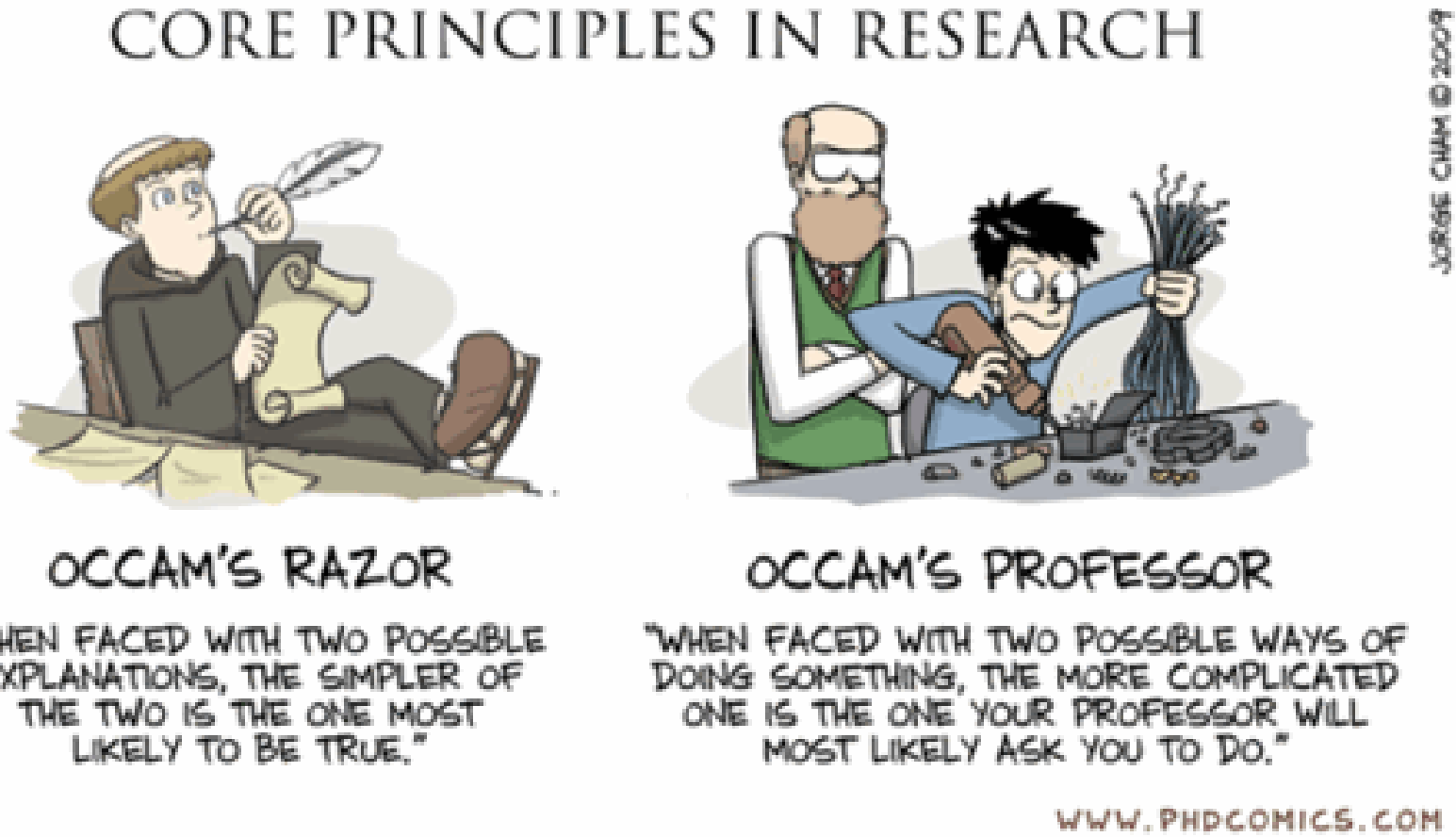
inferences from particular cases to the general case, or from a finite sample of data to a generalization about a whole population

Collect data about the concentration of airborne particles, rain and traffic.

Conclusion: make it based on data! Then see:

1. Do we have enough observations to generalize?
2. Were the observations made under a wide variety of conditions?
3. Are there observations in conflict with a derived law?

Occam's Razor (Parsimony)



Occam's Razor / Parsimony principle

- A simpler explanation of the phenomenon is to be preferred
  - A model with less variables is preferred, if it fits the data "equally well".
  - Newton's laws are simpler than Einstein's laws and they are still used for situations when they explain the phenomenon equally well (the difference in outcomes would be negligible).
- If you can explain the concentration of NO_2 using just traffic flow intensity and the wind flow well enough, you go for such a simple model. Otherwise, you have to involve more/other variables.

Summary

- **Scientific method** is a systematic approach to conducting *empirical research* to obtain *sound* answers to questions
- It makes use of deduction, induction, Occam's razor / Parsimony principle, verification by others, and statistics

Validity, reliability and reproducibility

Validity, reliability and reproducibility

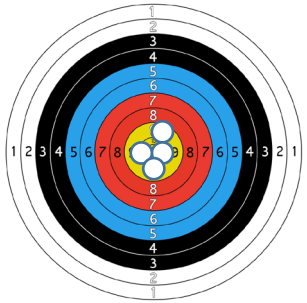
We discuss trustworthiness of data, data collection process, data measurement and data analytics methods using concepts of

- validity
- reliability, and
- reproducibility

Validity and reliability

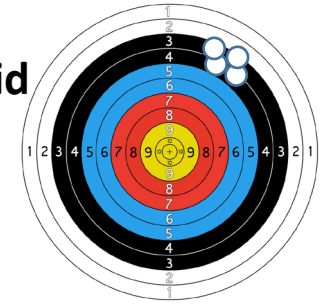
- Measurements or conclusions are **valid** when they accurately describe the real world
- Measurements or conclusions are **reliable** when the similar results are obtained under similar conditions
- Suppose the sensors used at a traffic measurement station malfunction and the number of cars per hour is always two times lower than the real number
  - our measurements are not valid – they do not describe the reality accurately
  - they are reliable – they would consistently show the same numbers in repeated experiments
- Reliability does not imply validity, nor does validity imply reliability!

Valid and reliable



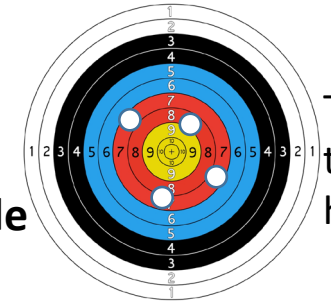
The mean is close to the target, low σ

Not valid but reliable



The mean is off, but low σ

Valid but not reliable



The mean is close to the target, but high σ

Internal validity versus external validity

Internal validity: are the conclusions valid **within the study**?

- Is there a strong justification for the conclusion about a causal relationship or **can there be an alternative explanation**?
  - i.e. is it possible that weather conditions lead to both a higher traffic intensity and a larger concentration of some chemical compound in the air?

External validity: can the conclusions of a scientific study be applied **beyond the context of the study**?

- Will the reaction time in the mouse experiment depend on the task difficulty in the same way if you consider people of different ages?

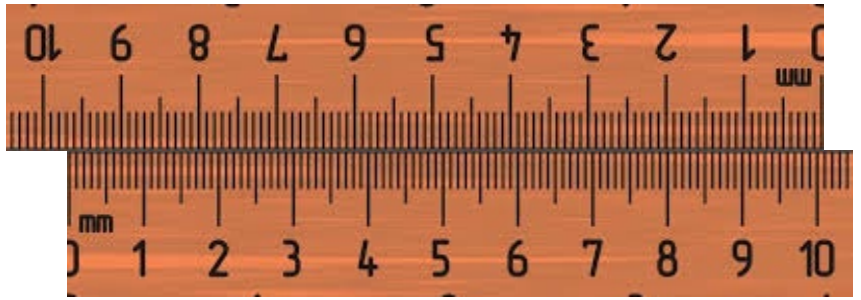
Reproducibility, verification by others

- Refers to the ability of others to **replicate** the findings
  - If others do not get similar measurements/conclusions, analysis results will not be accepted

- Example: traffic
  - Imagine you found that the amount of NO_2 in the air is proportional to the number of cars passing nearby, while the others do not see this relationship in their experiments. – so the results are not reproducible!
  - This can happen because of:
    - **Implicit assumptions**, e.g. the number of cars is an appropriate as an indicator for the amount of car emissions, assuming that all cars use fuel (think of electric cars)
    - your data or their data was collected in different, very **specific conditions** (e.g. a power plant near their measurement station)
    - the **measurements were incorrect** (e.g. malfunctioning measuring instruments)
    - you omitted some facts/results when drawing conclusions
    - …

Measurements: precision and accuracy

- In a **measurement process**
  - Is the measurement process actually measuring what it is intended to measure?
  - It involves both the **definition** of a measurement and the **instruments** used
  - Example: mouse experiment
    - What does reaction time mean? How is it measured?
- **Precision** of the instrument and **accuracy** of the measurements
  - **Precision** refers to **errors introduced by the measuring instrument**, i.e. ±0.5 mm (random errors)
  - **Accuracy** refers **to deviations from real values** (systematic errors)

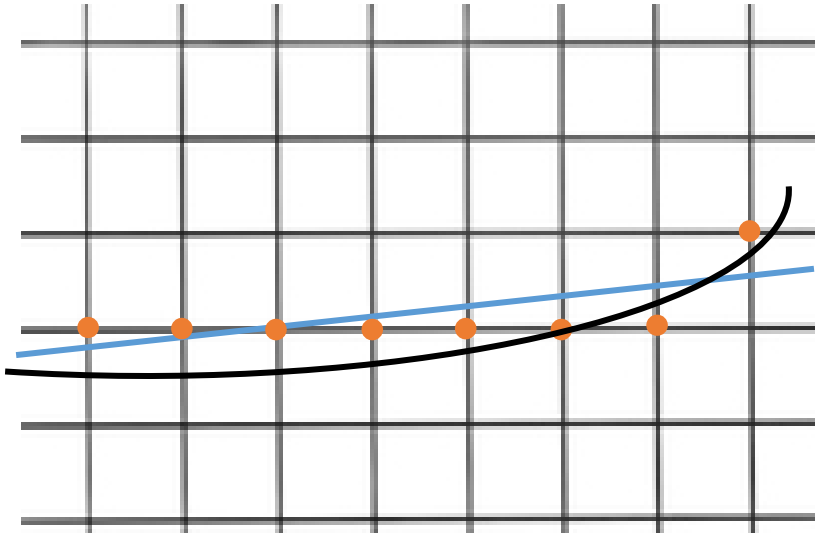


Possible threats to validity in measurements

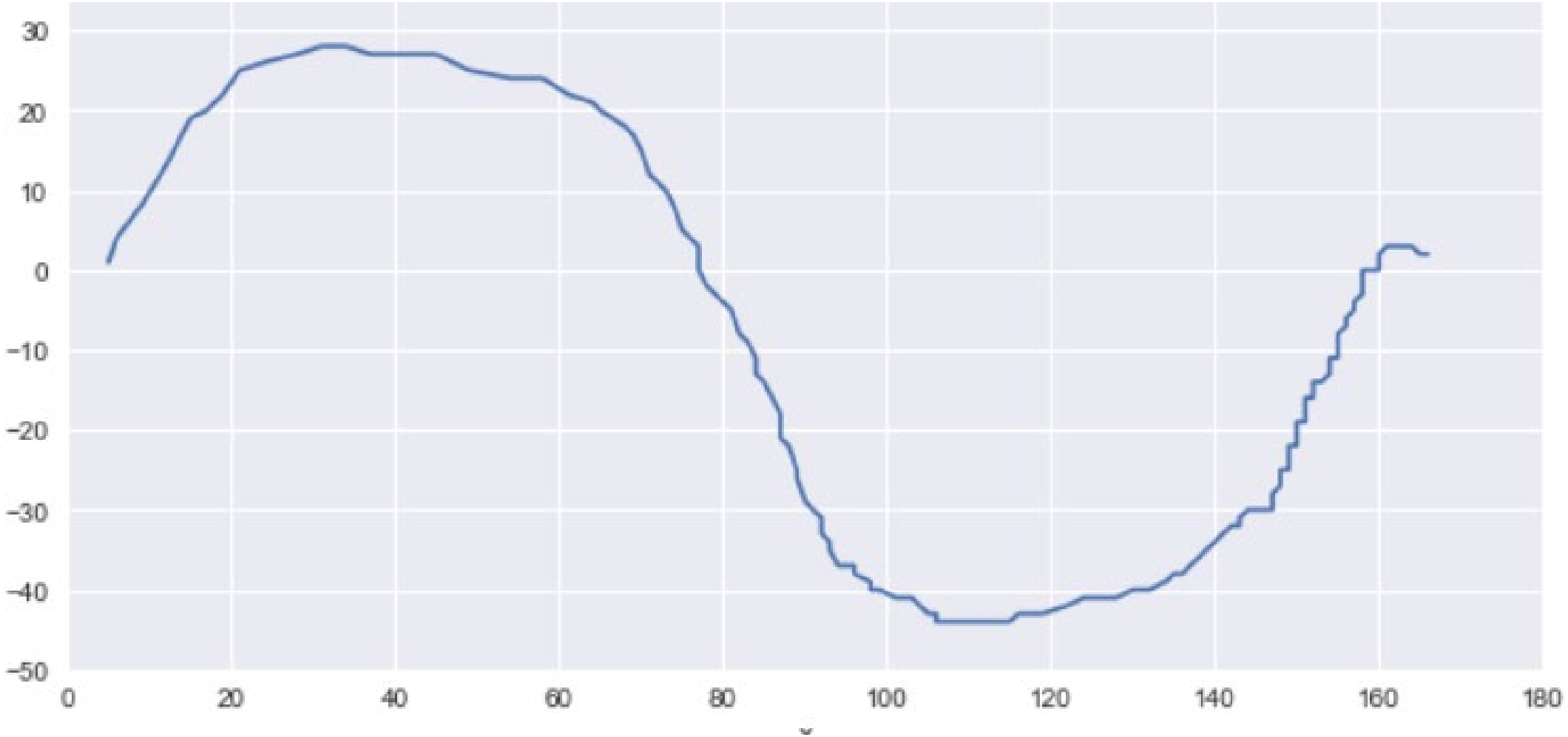
- **Sources of errors**:
  - Inadequacies of the technology employed
  - Measurement process
  - The definition of what is being measured
- **Random** errors (not forming any pattern) vs. **systematic** errors (consistent errors, like **offset** errors or **scale** errors)
  - e.g., every second car is measured whenever the traffic is intensive, so the numbers are too low compared to reality – **scale error** with the factor 0.5
  - The thermometer always showing the temperature that is 2^{o}C higher that the real temperature has an **offset error**

Studying path properties

- **Two real paths: the black one and the blue one**
- **Data collected**: for each path, we get coordinates of orange dots
- Which questions can be answered with this data and which not?
- How to compute the direction of the movement?



Mouse trajectory in the mouse experiment



Summary

- **Validity** refers to the ability to accurately **describe the real world**
- **Reliability** refers to the ability to obtain **consistent results** when being in **similar conditions**
- **Reproducibility** refers to the ability of others to **replicate** findings
- **Precision** refers to **errors related to the characteristics of the measuring instrument**
- **Accuracy** refers **to deviations from real values**
- **Random errors** vs. **systematic errors**
  - **Scale** errors and **offset** errors

Data Sampling

Population and sample

- **Population** is a **complete set of all elements**, like traffic and air quality data for every single moment of time at every location in the Netherlands
  - collect the data for the whole population is too expensive, or even physically impossible
  - it might be not feasible to analyze the data for the whole population (when it is available) because of the data size and the complexity of algorithms.
- A sample is any part of the population (whether it is representative or not)
  - the data from the mouse experiment in GA2 is a sample, containing data collected by the students following the DAE course
- A sample is ***biased*** if some part of the population is overrepresented compared to others
  - young people and males are overrepresented in the data collected in the mouse experiment, since the data is collected by students following the DAE course

Limitations due to sampling in data collection

Traffic and air quality

- Observations for one month only
  - Is it enough to build models?
  - Is it representative for the whole year?
- The measurements were made only at some locations in the Netherlands – a *sample of all possible locations*

Mouse experiment

- The users are 1st year TU/e students: a *non-representative sample* of all computer users w.r.t.
  - their age
  - educational background
  - …

What sampling methods are used and

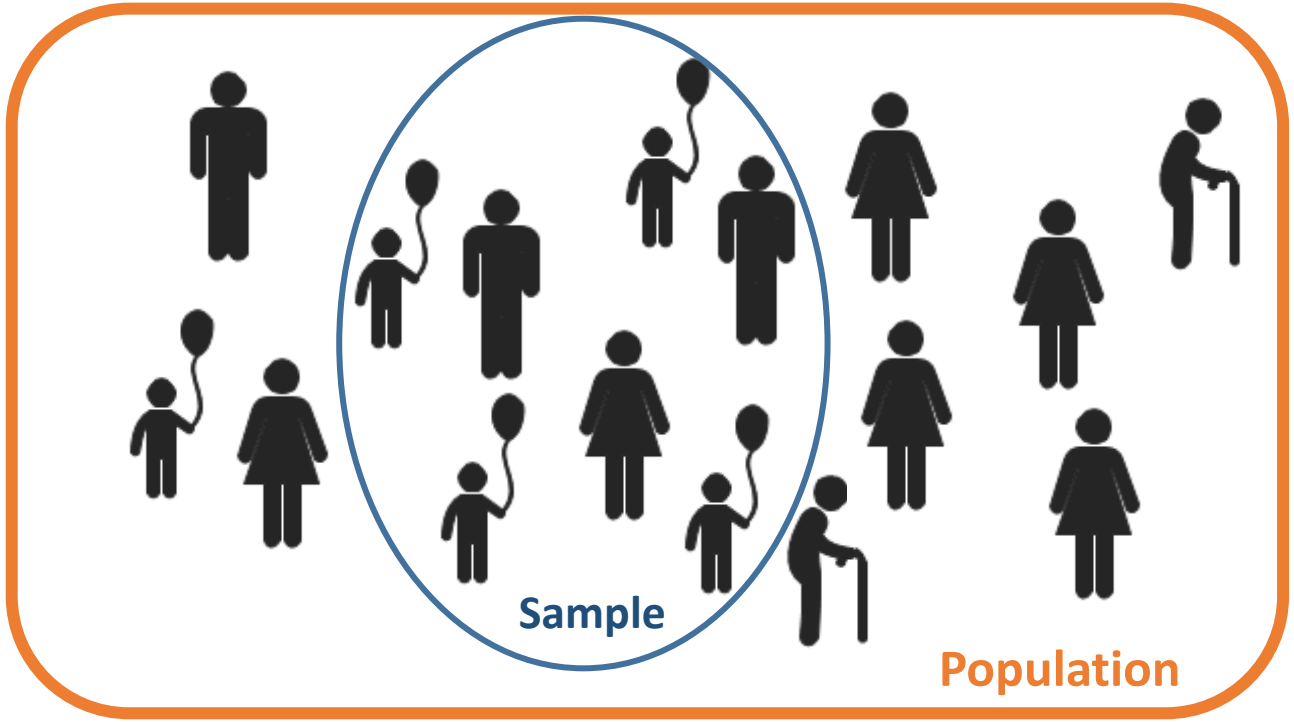
how do they influence the validity, reliability and reproducibility of conclusions?

Convenience sampling

- Going for data that is easier to collect
  - DAE students for the mouse experiment
- Advantages: saving time, effort, money, …
- Disadvantages: possible bias that is a threat to external validity
  - Would the reaction time in the mouse experiment depend on the target size in the same way for older people?

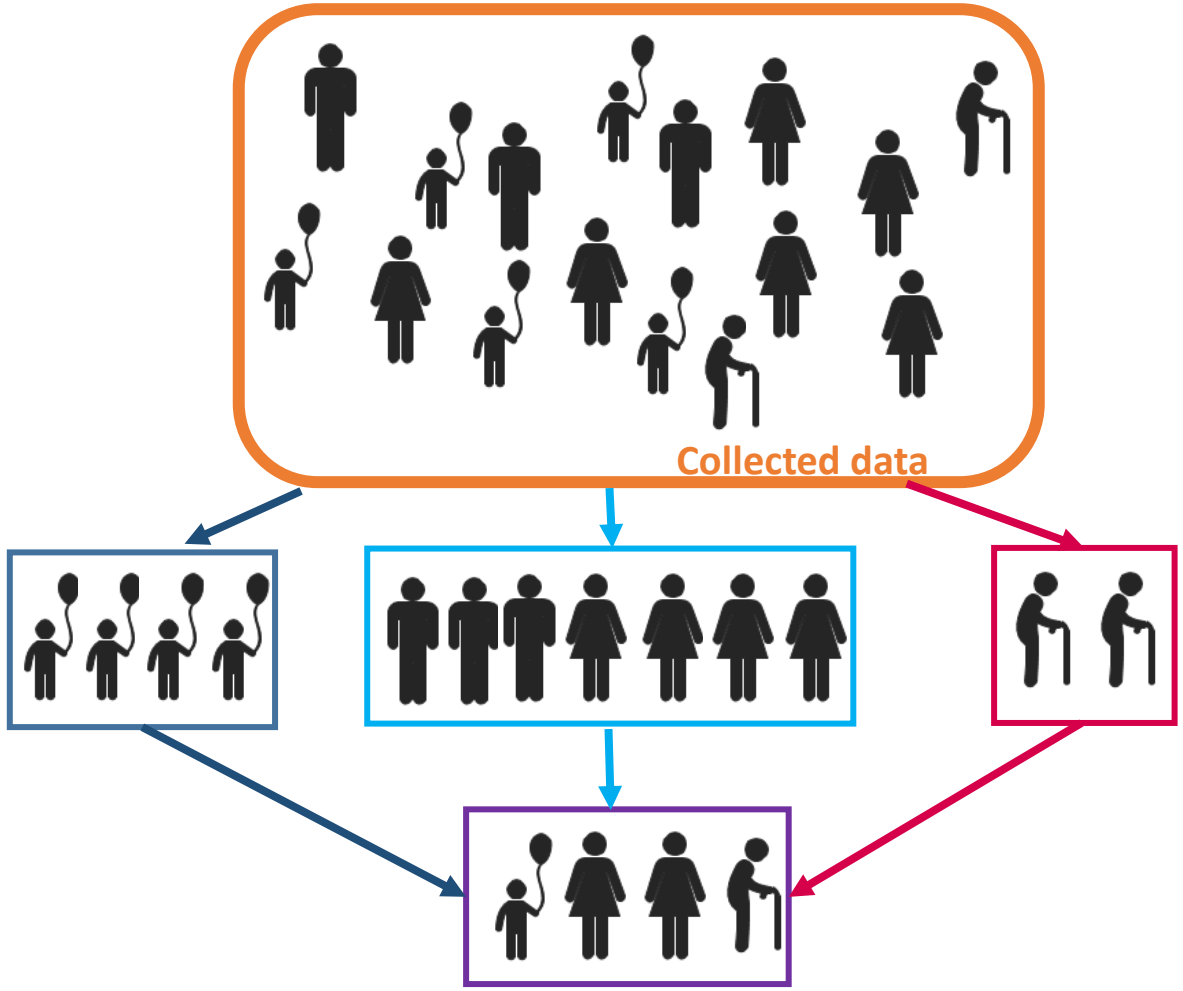
Random sampling

- Each individual is equally likely to be included into the sample
- Ignoring any knowledge of the population



Stratified random sampling

- Defining **strata** - disjoint parts forming the whole target population
  - e.g. based on age– see the example
- Define your sample size, e.g. N
- **Proportionate** stratified random sampling: take a random sample from every stratum in the proportion equal to **proportion of this stratum in the population**
- **Disproportionate** stratified random sampling: when you want to over-represent a particular stratum in the sample
  - a study of people with a rare medical condition: a sample should over-represent these people compared to the whole population



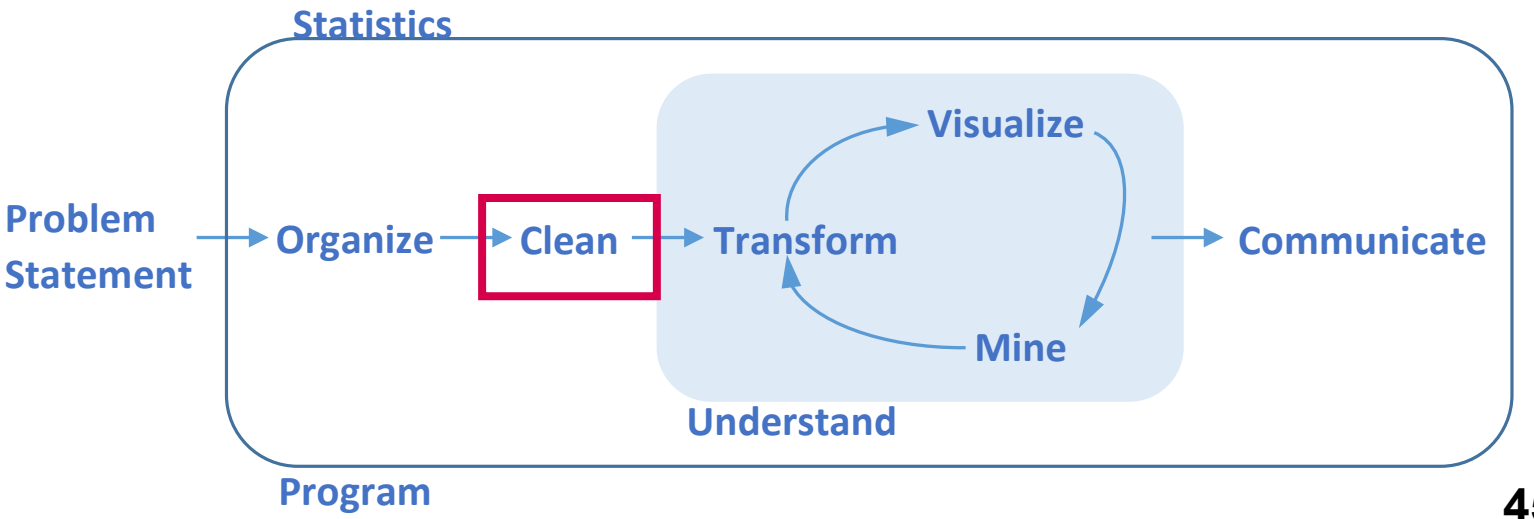
Voluntary sampling

- individuals select themselves
  - e.g. course evaluation questionnaires
- self-selection bias the effects of which are difficult to measure
  - Would more engaged students fill in the questionnaire?
  - more unsatisfied students?

Summary

- Difference between population and sample
- Types of sampling:
  - Convenience sampling
  - Random sampling
  - Stratified random sampling
    - Proportionate and disproportionate
  - Voluntary sampling

Data Cleaning and Preprocessing



Can we fully trust our data?

Traffic and air quality

- Constant speed of 144 km/h all the time for a traffic measurement station ?
- Strange values for air quality measurements ?
- Missing values for some time periods ?

Mouse experiment

- Reaction time = 5 minutes ?
- The path length is 3 times longer than the distance from the origin to the target ?
- Atypical mouse settings ?

Need for data cleaning and filtering!

Data cleaning

- is a process of detecting, diagnosing, and editing faulty data
  - **Incorrect data**
  - **Incomplete (missing) data**

Sources of problems with data collection

- Equipment or data transmission errors
  - Malfunctioning of sensors
  - Lost Internet connection – lost records
- Data collection circumstances
  - Weather conditions, resulting in bad visibility and as consequence too low values for the number of cars per hour
- Manual data entry procedures
  - Typos in entries, e.g.1890 instead of 1980
- Non-optimal data collection protocol / study setup
  - No air quality measurement stations in the proximity of a traffic measurement station
- …

Minimal checks of problems with data

- Incomplete (missing) data
  - No air quality measurements for certain time moments at some measurement stations
- Out-of-range data
  - The car speed of 500 km/h
  - Negative values for the concentration of NO_2
- Inconsistent data
  - Very different air quality measurements at two stations in a close proximity to each other: which one is right?
  - A period with high intensity of traffic in combination with a high deviation in the car speed values

Handling inconsistent, invalid or missing data

- **Discard** all records (rows) with at least one inconsistent, invalid or missing value or discard an attribute (column) with lots of such values
  - potentially loosing a lot of data
  - potentially introducing a **bias** in data: suppose no car speed data could be gathered when it is foggy and we remove all such periods from the data $\Rightarrow$ our conclusions on the remaining data set might not hold
- **Impute** values: fill in estimated values in place of inconsistent, invalid or missing values
  - replace all missing values with a single constant value based on the domain knowledge or an estimate, like a feature mean/median computed
    - risks: the estimate can be off, e.g. the car speed in a foggy period is likely to be far below the mean value
  - use data mining methods to estimate the likely value for the missing value on the values of other features, a feature mean for the given class (e.g. obtained with clustering), time series analysis, …
    - risks: the estimates introduce a bias in the data
- **Work in the presence** of missing data

In practice: try different options and see what works in your case

Noise reduction in time series data

- **Time series** is a sequence of pairs (t_n, x_n), where t_n is the observation time and x_n is the observed value, such that $t_n < t_{n+1}$.
- For **equispaced** time series, the spacing of observation times $(t_{n+1} - t_n)$ is constant and it is called **sampling frequency**.
- **Noise** is an *unwanted disturbance* in time-series data.
  - It can be in the form of small fluctuations present in the data, which you want to remove to see trends
  - or in the form of wrong measurements which you want to suppress

Summary

- Data cleaning is a process of detecting, diagnosing, and editing faulty data
- It is important to understand the source of faults in order to choose an appropriate data cleaning strategy.
- You should always perform at least some minimal checks of data quality (missing data, out-of-range data, inconsistent data)
- The main strategies are to discard missing data, impute data values or to work in the presence of missing data
- There are specific methods for noise removal in time series

Filtering time series data

Median filter

- Choose the **window size**, e.g. 3
- Position the window at the beginning of the time series and compute the median.
- Move the window by 1 and compute the next median value.
- Proceed till the end of time series.
- The filtered time series consists of the computed window medians.

t	0	1	2	3	4	5	6
v	7	8	2	6	9	0	7
v_{mf}		7	6	6	6	7	

window size = 3

t	0	1	2	3	4	5	6
v	7	8	2	6	9	0	7
v_{mf}			7	6	6		

window size = 5

Median filter application

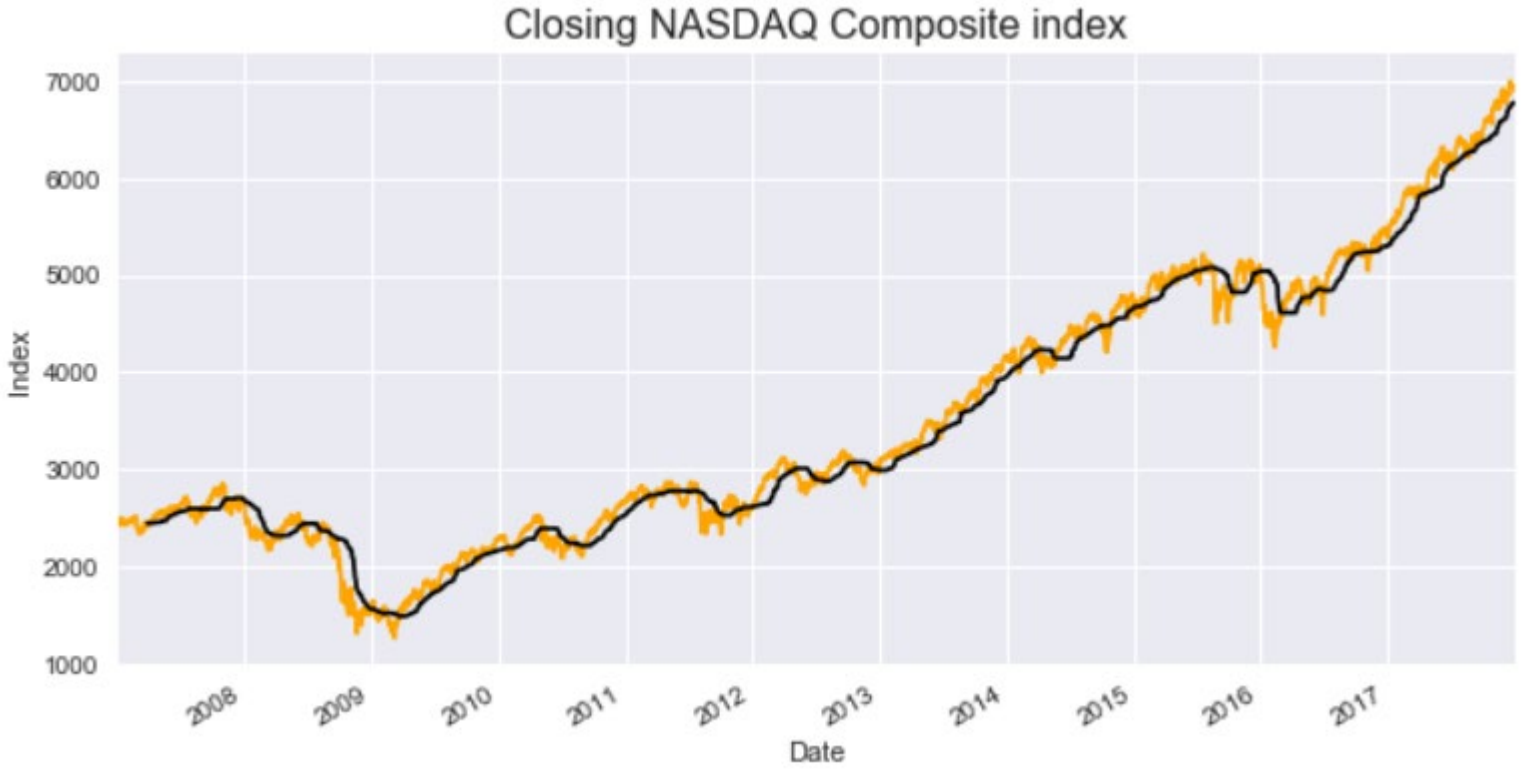
- Example from the DAS Programming exercises
- Window size = **10**

Median filter application

- Example from the DAS Programming exercises
- Window size = **30**

Median filter application

- Example from the DAS Programming exercises
- Window size = **60**



original data
filtered data

Notice that for a window size k the filtering function implementation takes the window from i to i + k – 1 and assigns this value to an element with time stamp t_{i+k-1} , which explains the shift to the right in the filtered data

Mean filter (moving average)

- Choose the window size, e.g. 3
- Position the window at the beginning of the time series and compute the mean.
- Move the window by 1 and compute the next mean value.
- Proceed till the end of time series.
- The filtered time series consists of the computed window means.

t	0	1	2	3	4	5	6
v	7	8	2	6	9	0	7
v_{mf}		5.7	5.3	5.7	5	5.3	

window size = 3

- Mean filter is more sensitive to outliers than median filter.
- Values far away have the same influence as values nearby

t	0	1	2	3	4	5	6
v	7	8	2	6	9	0	7
v_{mf}			6.2	5	4.8		

window size = 5

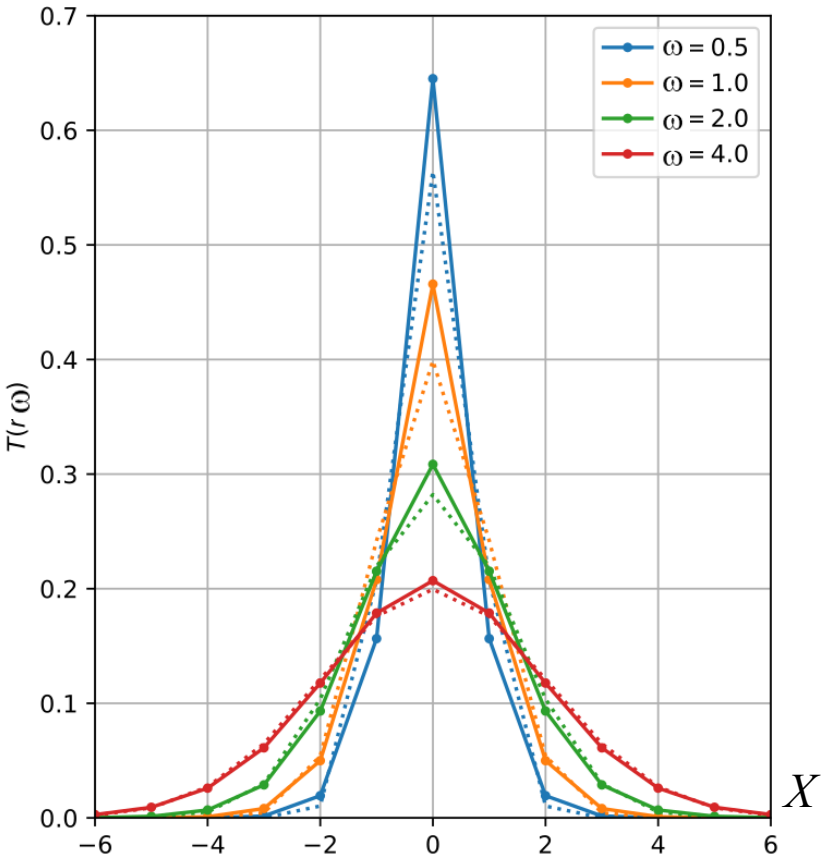
Gaussian filter

- Choose the filter width ω
- Use the Gaussian kernel T for the value ω
- Consider a pair (t_n , x_n) from the time series
- Assign weights $w_i = T(t_n - t_i, \omega)$ to each value x_i,
- The further away, the lower the weight
- Compute the value for the filtered data as:

$$\bar{x}_n = \sum_i w_i \, x_{n-i}$$

- Note that $\sum_i w_i = 1$

Next lecture: Gaussian function, on basis of which the Gaussian kernel is constructed



discrete Gaussian kernel (solid line)
sampled Gaussian kernel (dashed line)

Gaussian filter application example

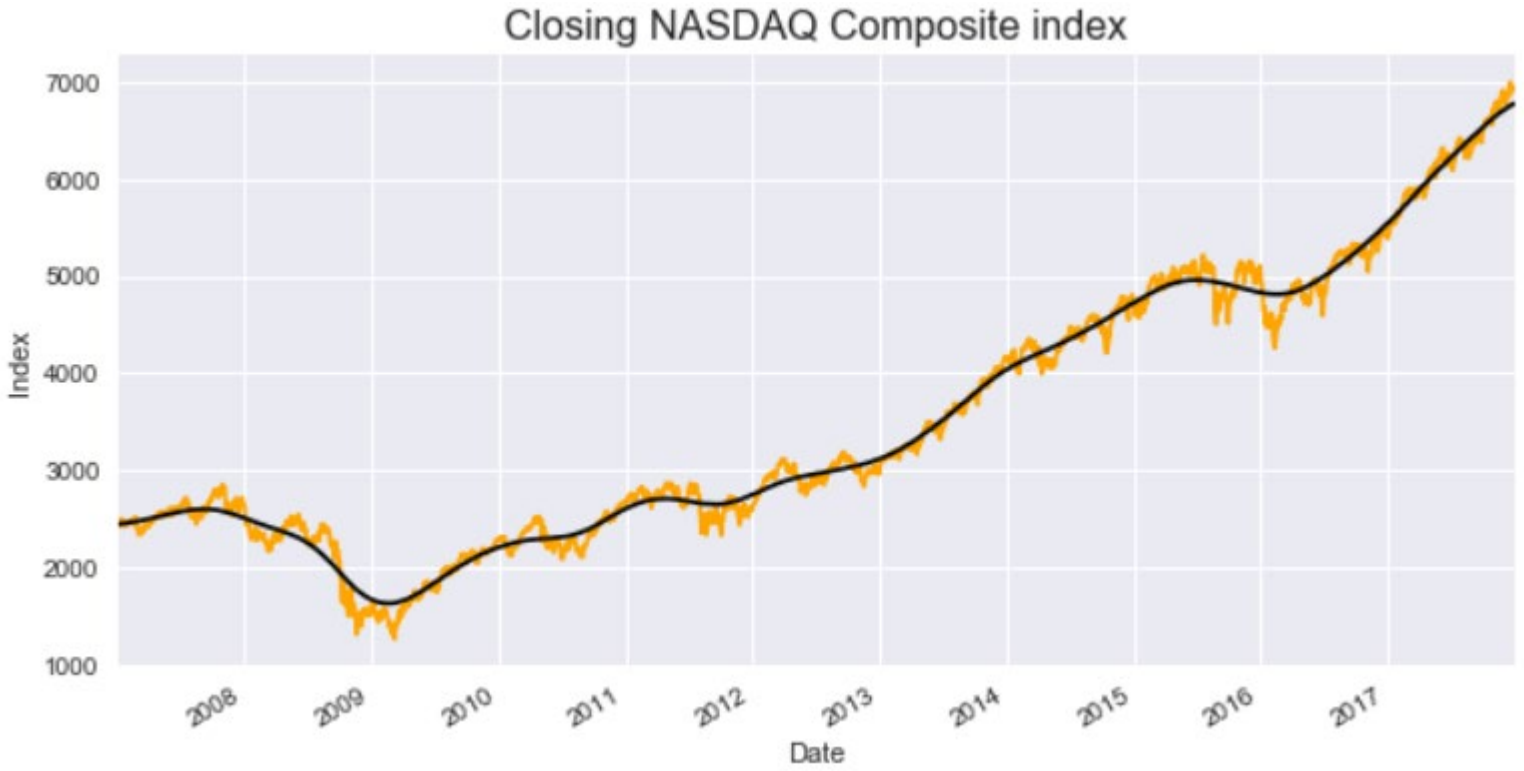
- ω = 10 days

Gaussian filter application example

- ω = 30 days

Gaussian filter application example

- ω = 60 days

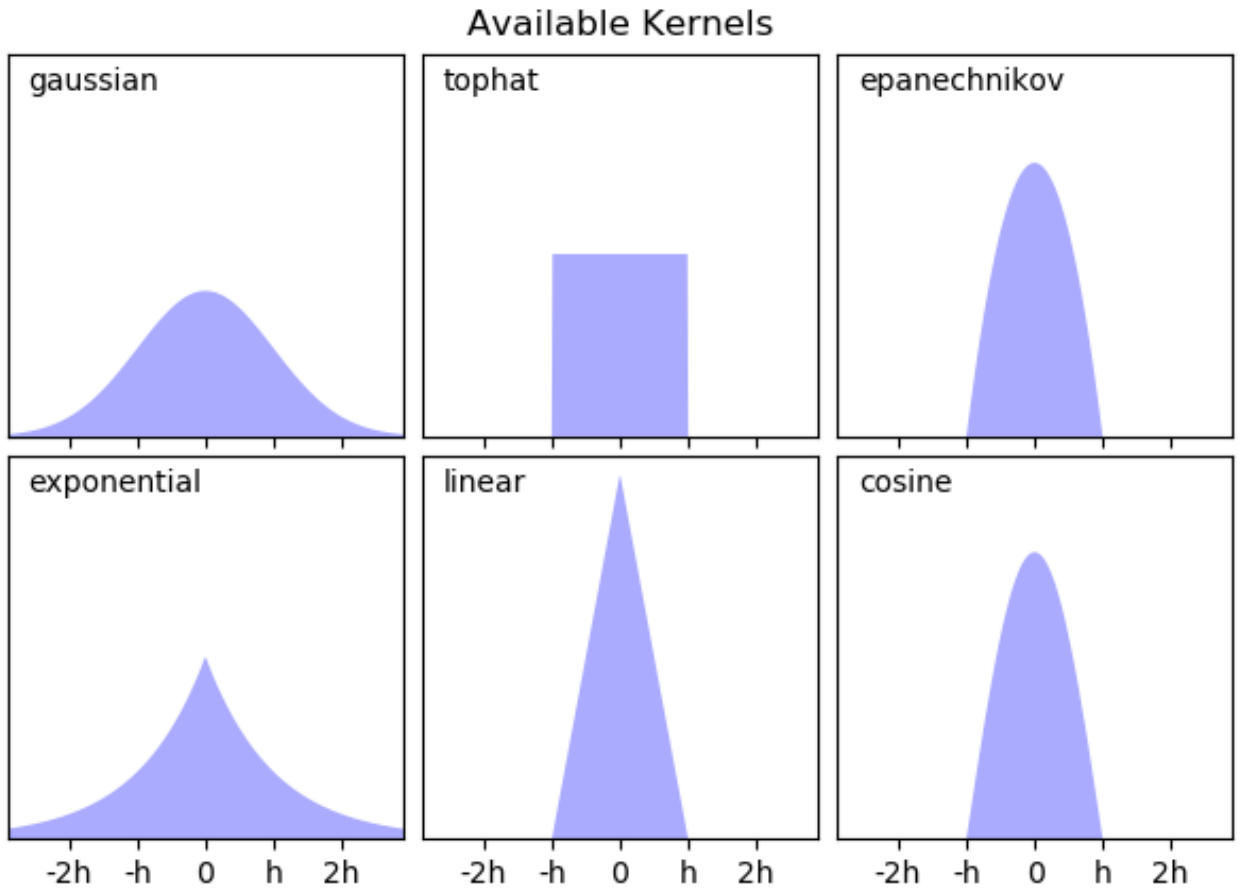


Convolution filters

- The same idea as Gaussian filter, but using differently shaped functions to generate weights
  - You can also define your own weights, but make sure that

$$\sum_i w_i = 1$$

- Which kernel corresponds to the moving average (mean) filter?



Applying a convolution filter

- Let's consider a convolution filter with weights [0.25, 0.5, 0.25]
  (i.e. $w_{-1} = 0.25$, $w_0 = 0.5$ and $w_1 = 0.25$)
- $\bar{v}_3 = \sum_i w_i \, v_{n-i} = w_{-1}v_2 + w_0 v_3 + w_1 v_4 = 5.75$

- Compute the other values yourself!

t	0	1	2	3	4	5	6
v	7	8	2	6	9	0	7
v_{mf}		6.25	4.5	5.75	6	4	

window size = 3

One more example

Suppose we have the convolution filter [0.1, 0.2, 0.4, 0.2, 0.1]

(the sum of weights is 1)

We will compute the value of the 10th element in the filtered time series.

								0.1	0.2	0.4	0.2	0.1								
weights:																				
t	1	2	3	4	5	6	7	8	9	**10**	11	12	13	14	15	16	17	18	19	20
v	4	3	5	6	5	5	30	29	28	4	5	4	5	6	3	4	5	6	6	4
v'										11.5										

$$0.1 \cdot 29 + 0.2 \cdot 28 + 0.4 \cdot 4 + 0.2 \cdot 5 + 0.1 \cdot 4 = 11.5$$

Summary

- Filters can be used to reduce noise in time series data
- Median, mean and Gaussian filter are popular simple filters
- You can define your own convolution filter
  - Usually, the sum of the weights should be 1

Variables (features)

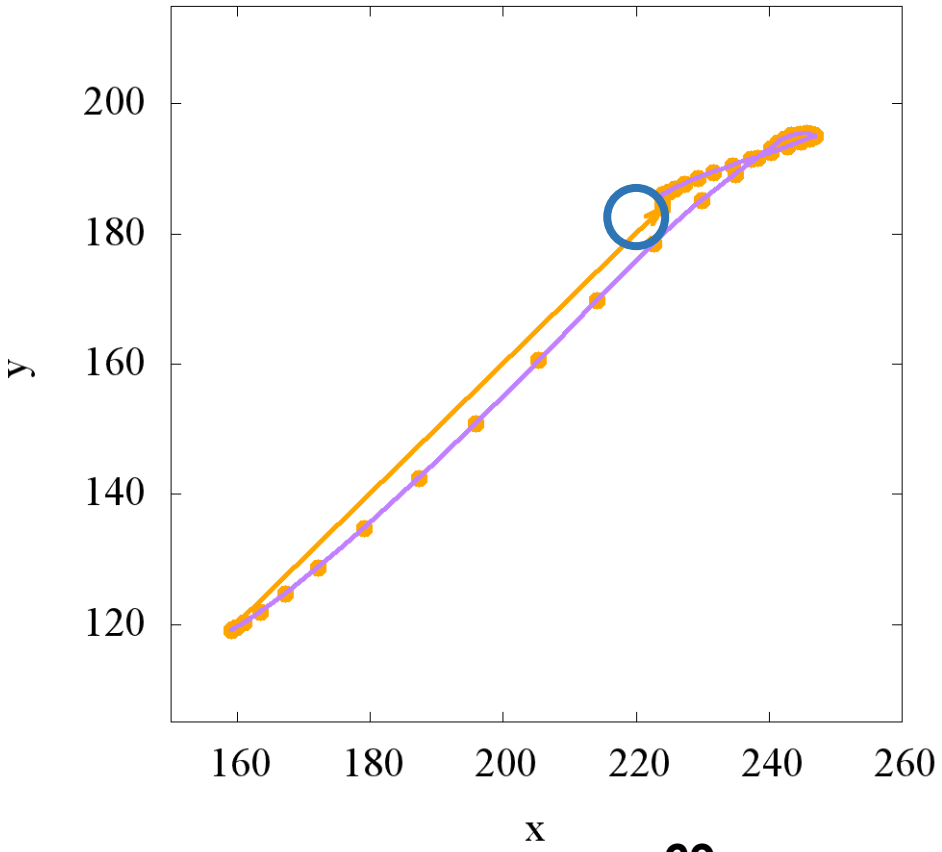
Dependent, independent and confounding variables

- Variables are opposite to constants: variables change their value
- **Independent variables** are used to see how the change of their value will be reflected in a change of value of a **dependent variable**
  - Independent variables can sometimes be actively controlled by the experimenter (the size of the target generated) and sometimes not (the traffic intensity)
- Statistical analysis might indicate that an independent variable appears to be **correlated** with variation in the dependent variable, but the dependency is **not necessarily causal**
- **Confounding variable** is a variable that is not taken into account and that can provide an explanation to the observed effect of the independent variable on the dependent variable
  - Suppose you mined a model prediction PM 2.5 concentration based on weather conditions. Wind, temperature and rain can lead to a decrease of the concentration of PM 2.5 in the air, but they could also trigger an increase of traffic intensity, thus making the effect of weather conditions on air pollution less visible
  - We often do not know all potential confounding variables at the beginning of the study

Feature generation: identifying *informative* variables

- Goal: reduce the number of data/variables in the dataset by creating new, more informative variables from existing ones
- mouse trajectory:
  - use the coordinates of the origin and the target to compute
    - the distance between them
    - the direction of the target w.r.t. the origin
    - i.e. compute the polar coordinates for the target
  - use the trace coordinates to split the path into segments
    - can you propose an explanation for the shape of the last part of the path?

a path from center to target



Mouse experiment: path features

Data contain coordinates (*X*,*Y*) over time *T* (the **path** followed) but for problem statements related to overshoot and correction movements we need additional measures ("features") that take the time aspect into account

Examples of such features:

- *PL* "length of the path travelled"
- *DT* "distance to target" as a function of time *T* (Euclidean distance)
- **ratio** of *PL* and the Euclidean distance to the starting point (the minimal possible travel distance)

Independent variables for the mouse experiment problem

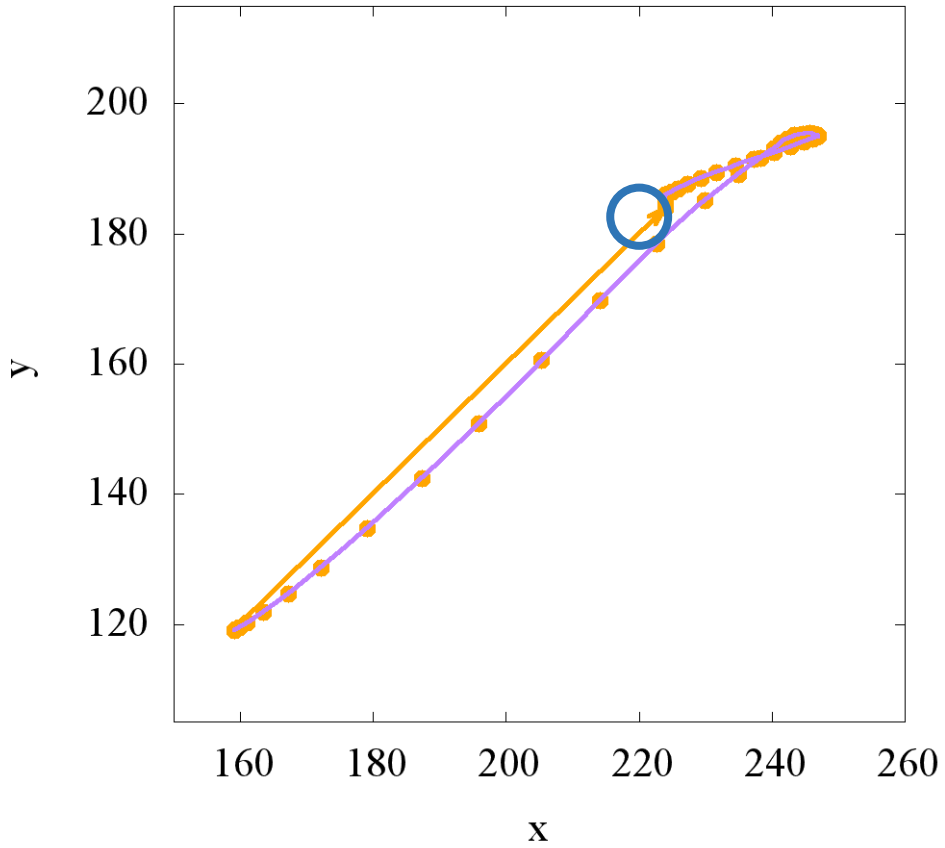
- The angle to the target
- A ratio between some variables
- …

paths(user, trial, t, x, y)
trials(user, trial, delay, input_method,
mouse_acceleration, target_radius, target_x, target_y, …)
users(user, major, right_handed, use_tue_laptop, …)

Candidates for dependent variables

- Time to target
- Reaction time
- …

a path from center to target



Confounding variables

- Gaming experience?

Summary

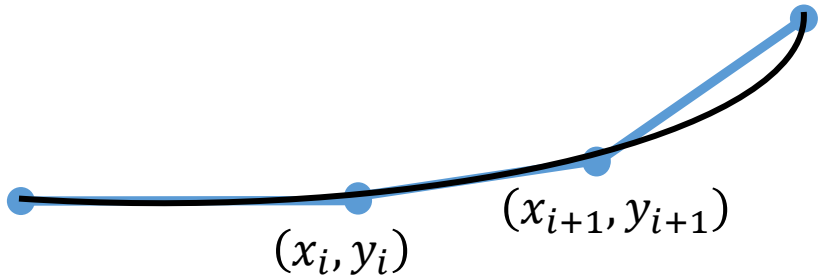
- Remember which variables you use as independent variables and which ones are dependent.
- Be careful with confounding variables.

Extracting features from time signals

Path length by linear approximation

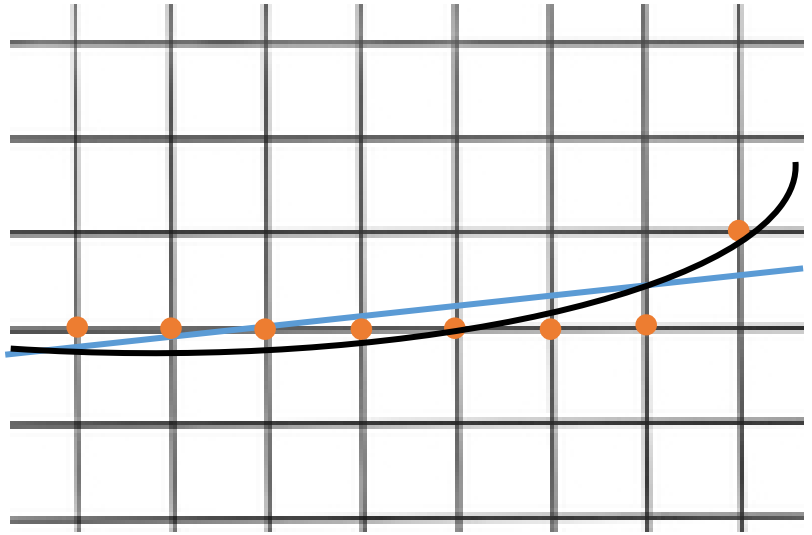
Distance between $(\boldsymbol{x_i}, \boldsymbol{y_i})$ and $(\boldsymbol{x_{i+1}}, \boldsymbol{y_{i+1}})$ can be approximated as Euclidean distance between the two points:

$$d_{i,i+1} = \sqrt{(\boldsymbol{x_{i+1}} - \boldsymbol{x_i})^2 + (y_{i+1} - \boldsymbol{y_i})^2}$$



Path $\boldsymbol{p_{k,l}}$ from point k to point l can be computed as sum

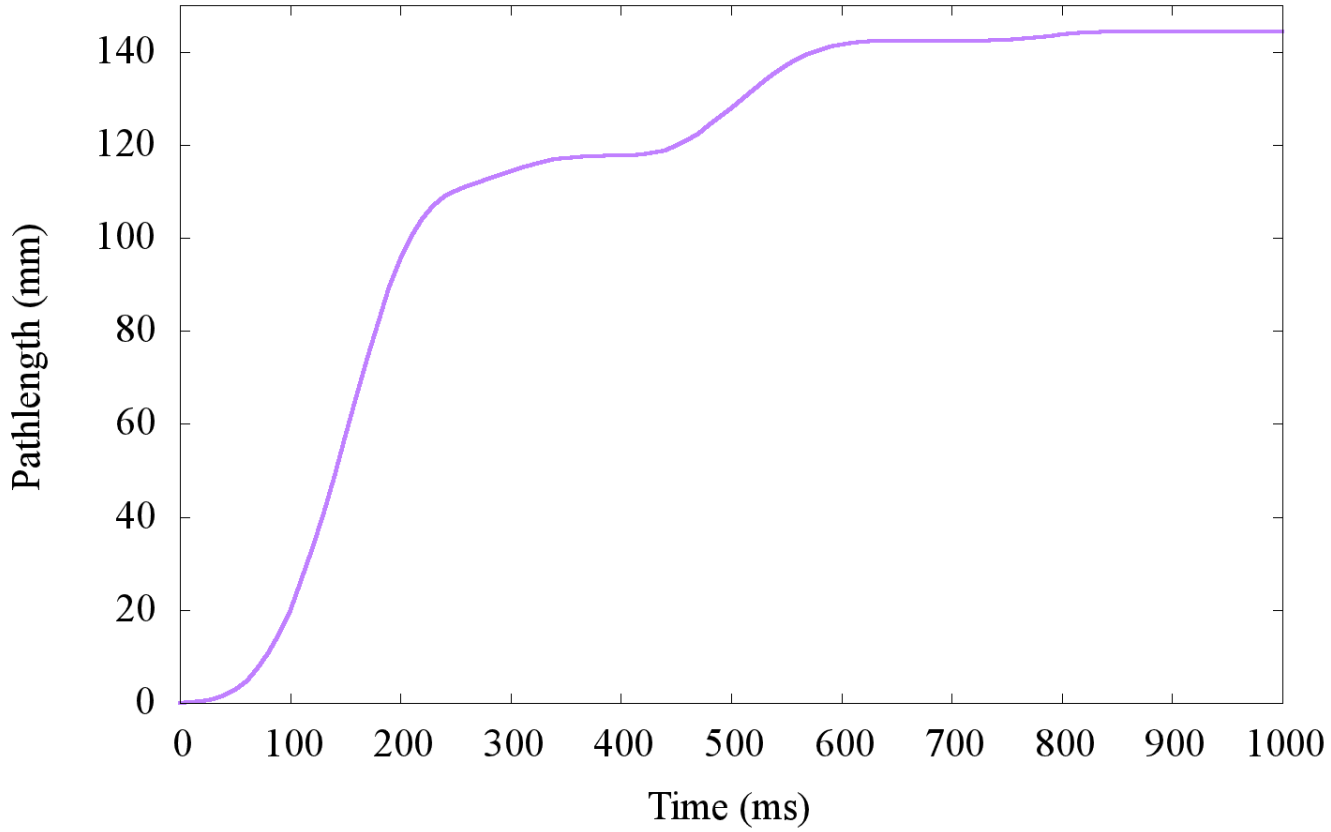
$$\boldsymbol{p_{k,l}} = \sum_{i=k}^{i=l-1} \boldsymbol{d_{i,i+1}}$$



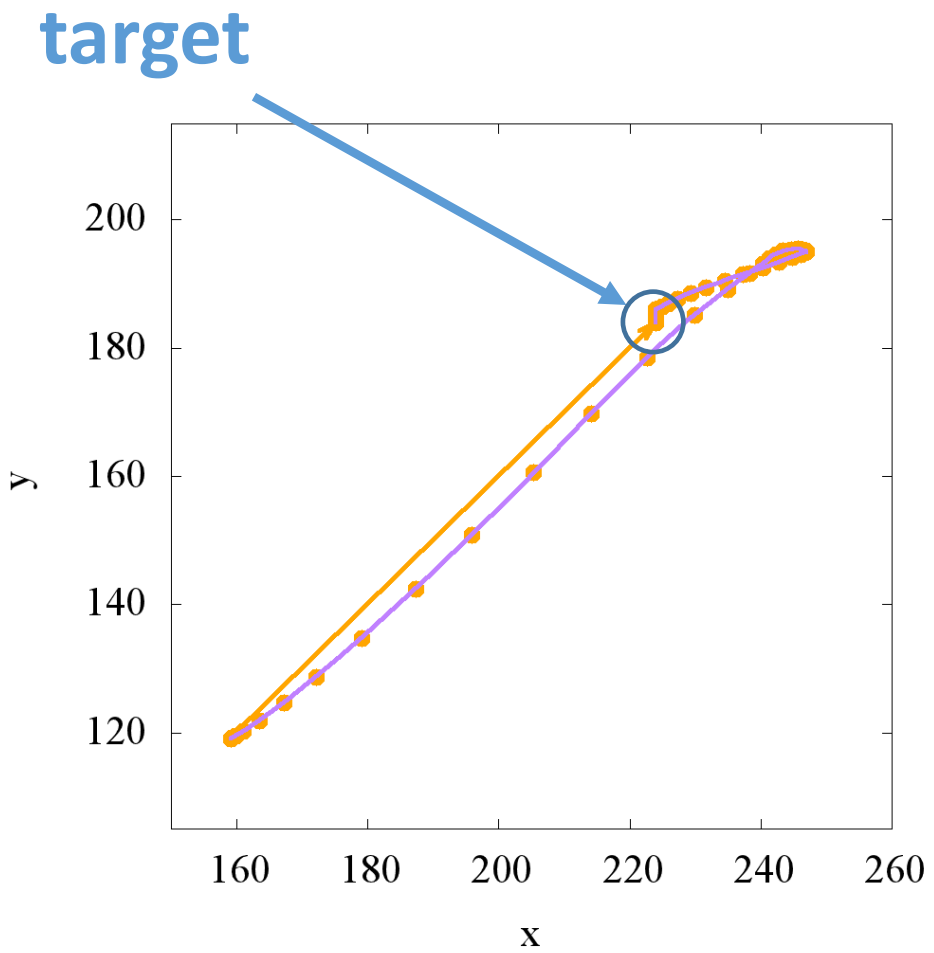
Remember that the precision of the approximation will be influenced by the precision of the measurements!!!

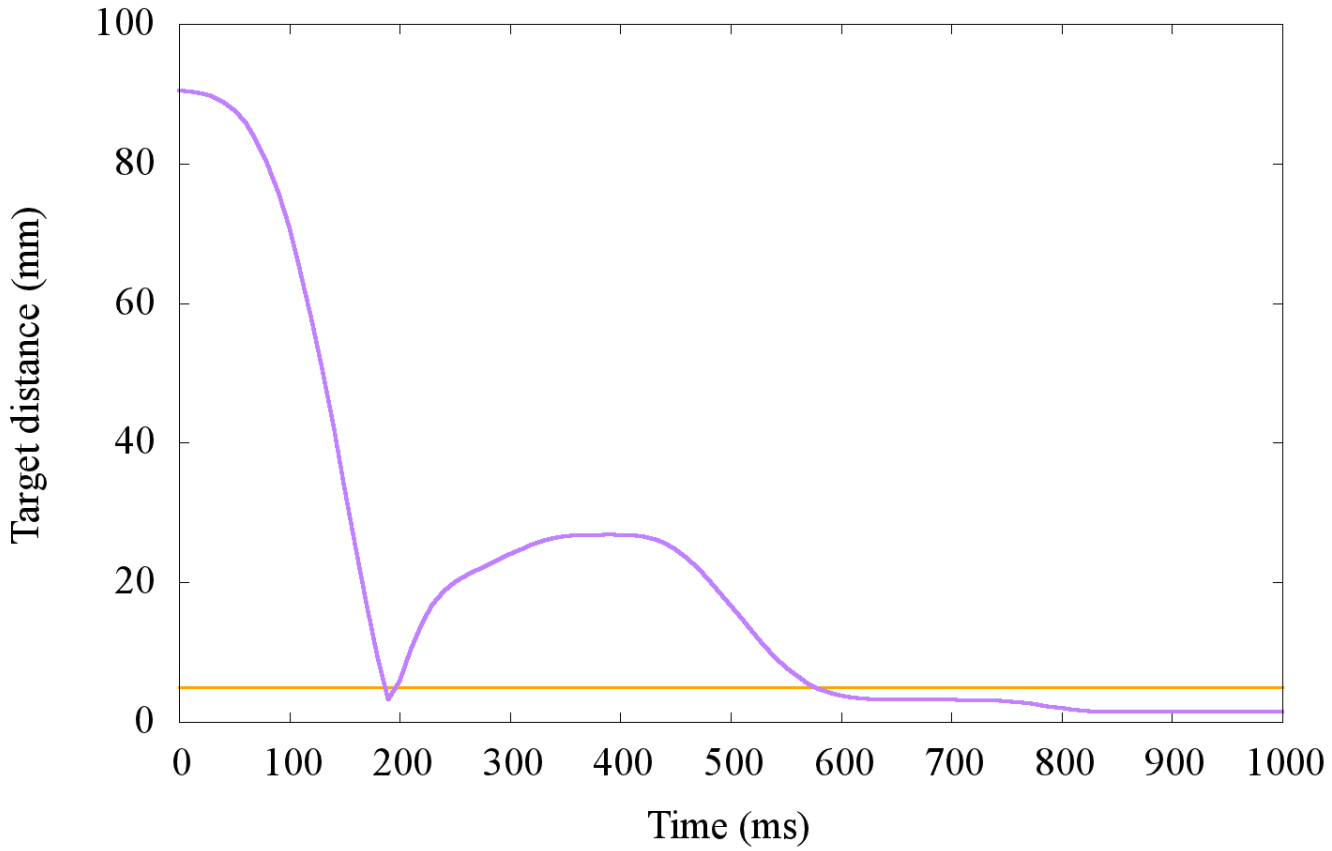
Mouse experiment: computing path length from the source - $p_{0,l}$

Euclidean distance is 90 mm, actual distance travelled is 144 mm



Mouse experiment: distance to target (DT)





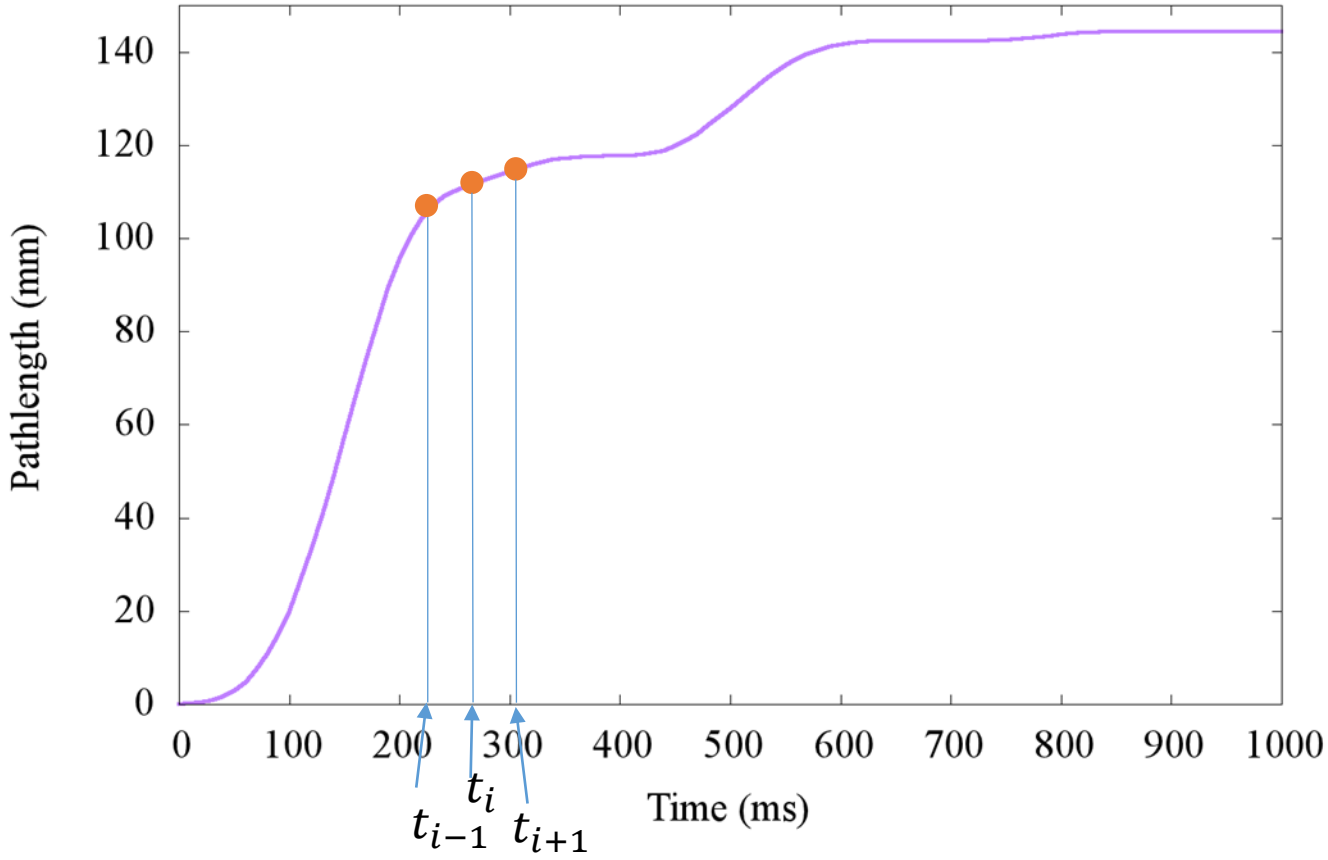
Numerical differentiation – computing speed

We constructed a new time series $(\boldsymbol{i}, \boldsymbol{p_i})$, where $\boldsymbol{p_i}$ is the length of the path from the origin at moment $\boldsymbol{t_i}$.

Now we want to compute the speed at moment $\boldsymbol{t_i}$.

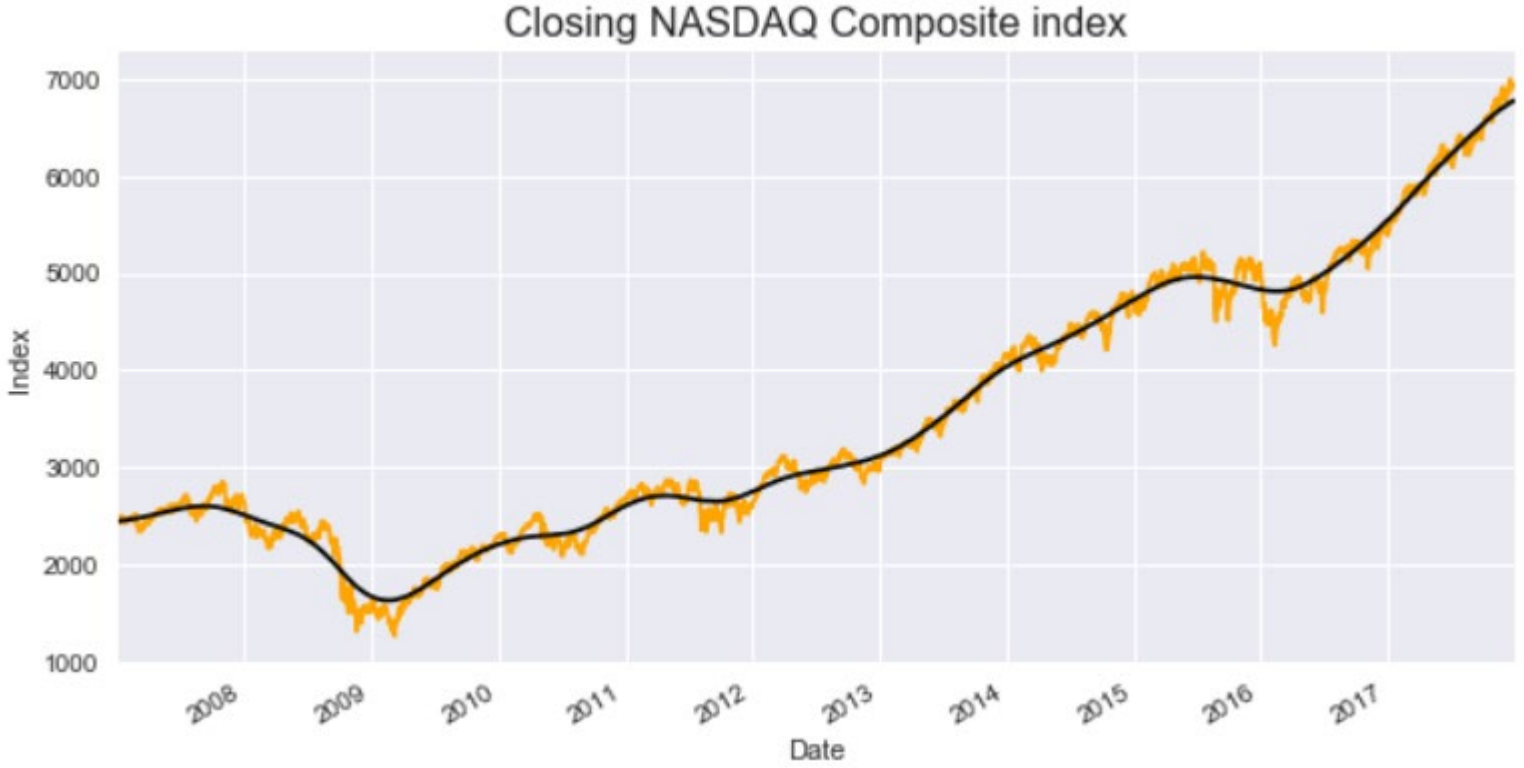
Approximate the derivative:

$$\boldsymbol{p'(t_i) = \frac{p_{i+1} - p_{i-1}}{t_{i+1} - t_{i-1}}}$$



Numerical differentiation in the presence of noise

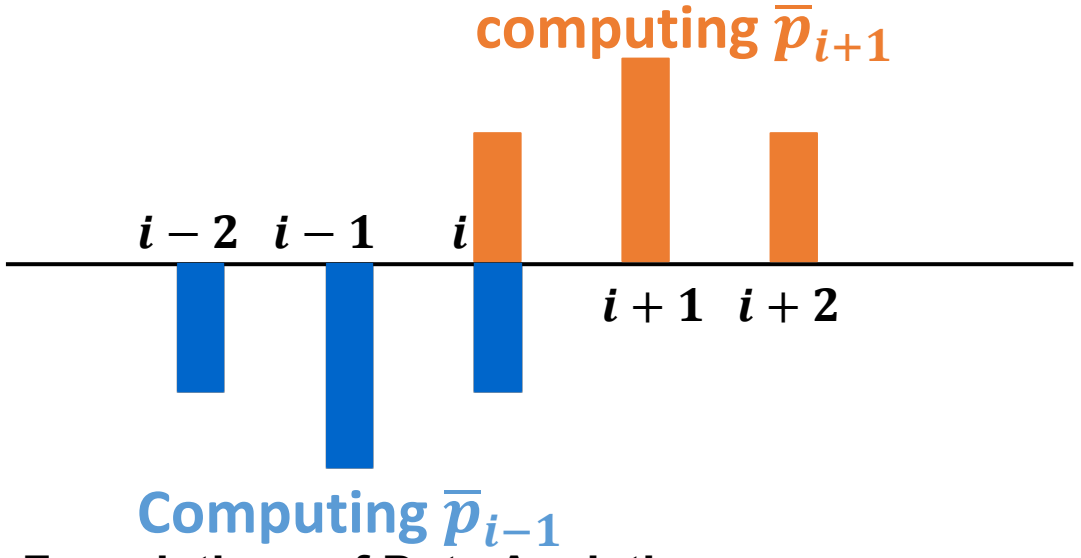
- Smooth the data first using filtering techniques
- Compute the derivative on the smoothed time series

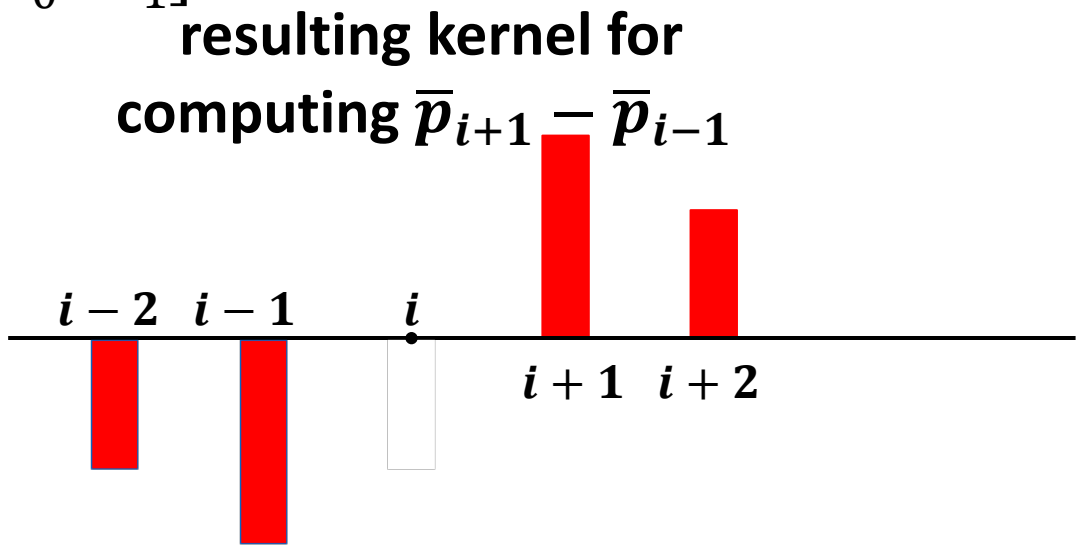


Numerical differentiation in the presence of noise

- Compute a kernel for computing derivatives! $\boldsymbol{p'(t_i) = \frac{p_{i+1} - p_{i-1}}{t_{i+1} - t_{i-1}}}$
- Consider a smoothing kernel with weights $[w_1, w_0, w_1]$, $w_0 + 2w_1 = 1$
- The direct filter to compute filtered $\boldsymbol{\overline{p}_{i+1} - \overline{p}_{i-1}}$ is

$$[-w_1, -w_0, 0, w_0, w_1]$$





Computing the kernel for differentiation

- Consider a kernel with weights $[w_1, w_0, w_1]$ $(w_0 + 2w_1 = 1)$

$$\bar{\boldsymbol{p}}_{\boldsymbol{i+1}} - \bar{\boldsymbol{p}}_{\boldsymbol{i-1}} =$$

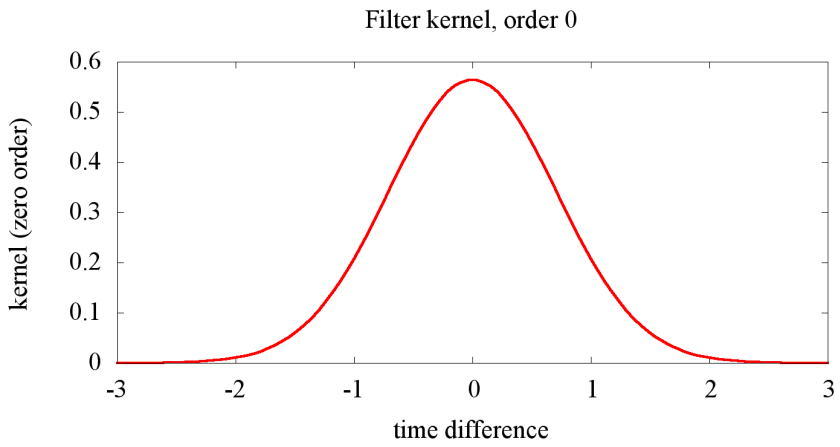
$$(\boldsymbol{w_1 p_i} + \boldsymbol{w_0 p_{i+1}} + \boldsymbol{w_1 p_{i+2}}) - (\boldsymbol{w_1 p_{i-2}} + \boldsymbol{w_0 p_{i-1}} + \boldsymbol{w_1 p_i}) =$$

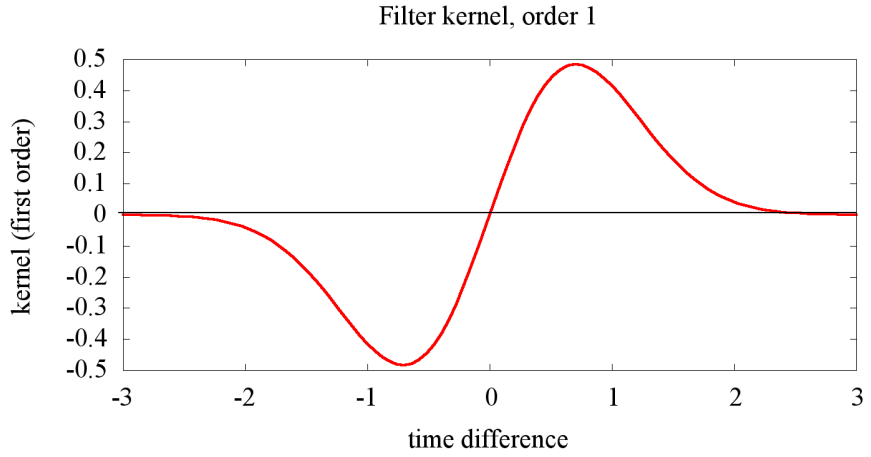
$$-\boldsymbol{w_1 p_{i-2}} - \boldsymbol{w_0 p_{i-1}} + \mathbf{0}\,\boldsymbol{p_i} + \boldsymbol{w_0 p_{i+1}} + \boldsymbol{w_1 p_{i+2}}$$

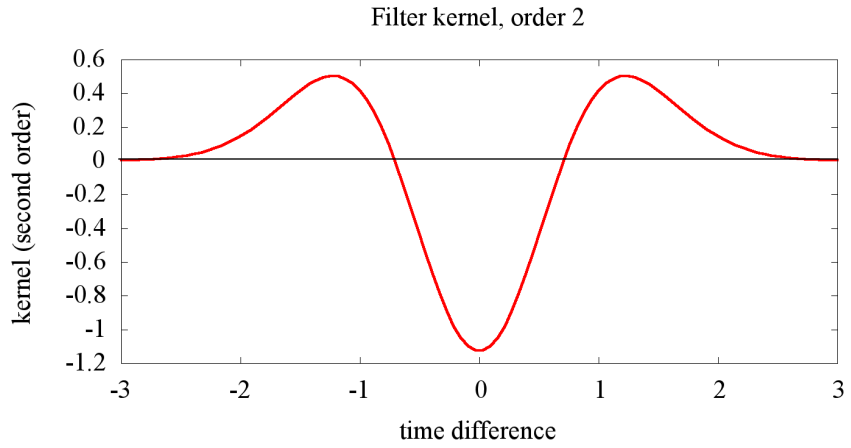
- So the direct filter is $[-w_1, -w_0, 0, w_0, w_1]$

computing $-p_{i-1}$

Gaussian kernel functions for computing first and second derivatives



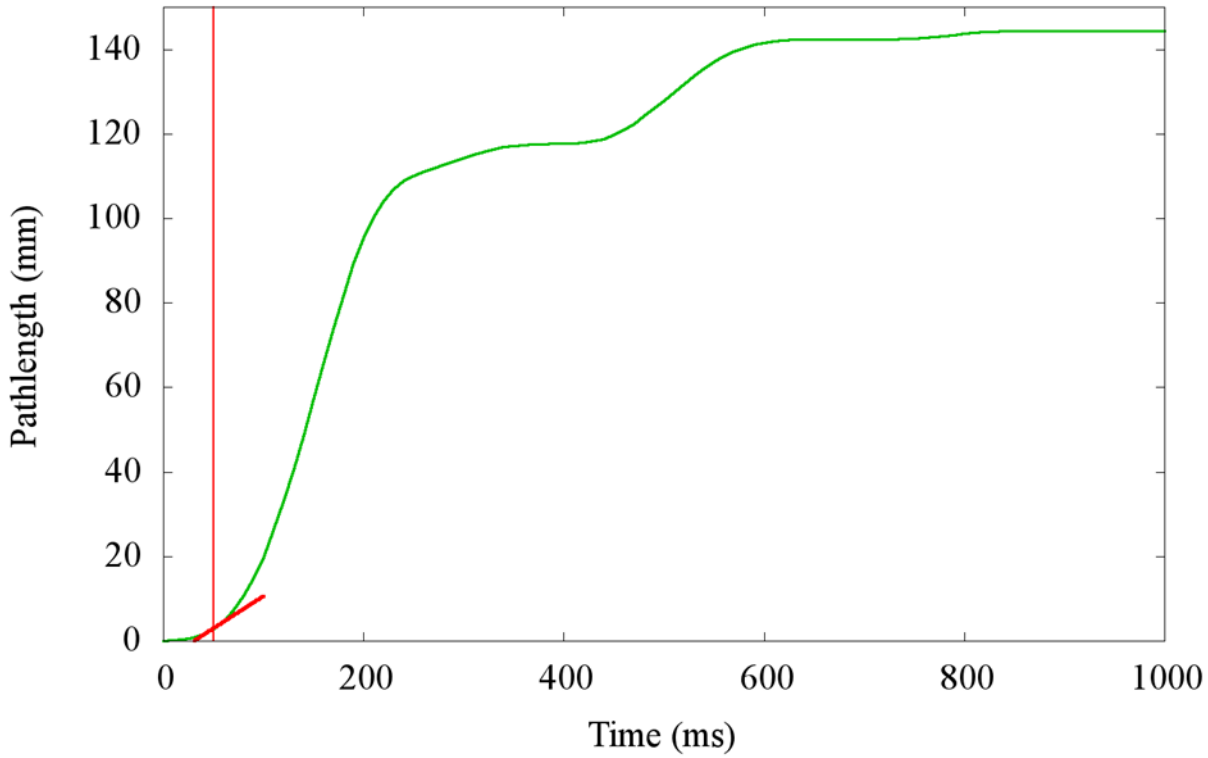


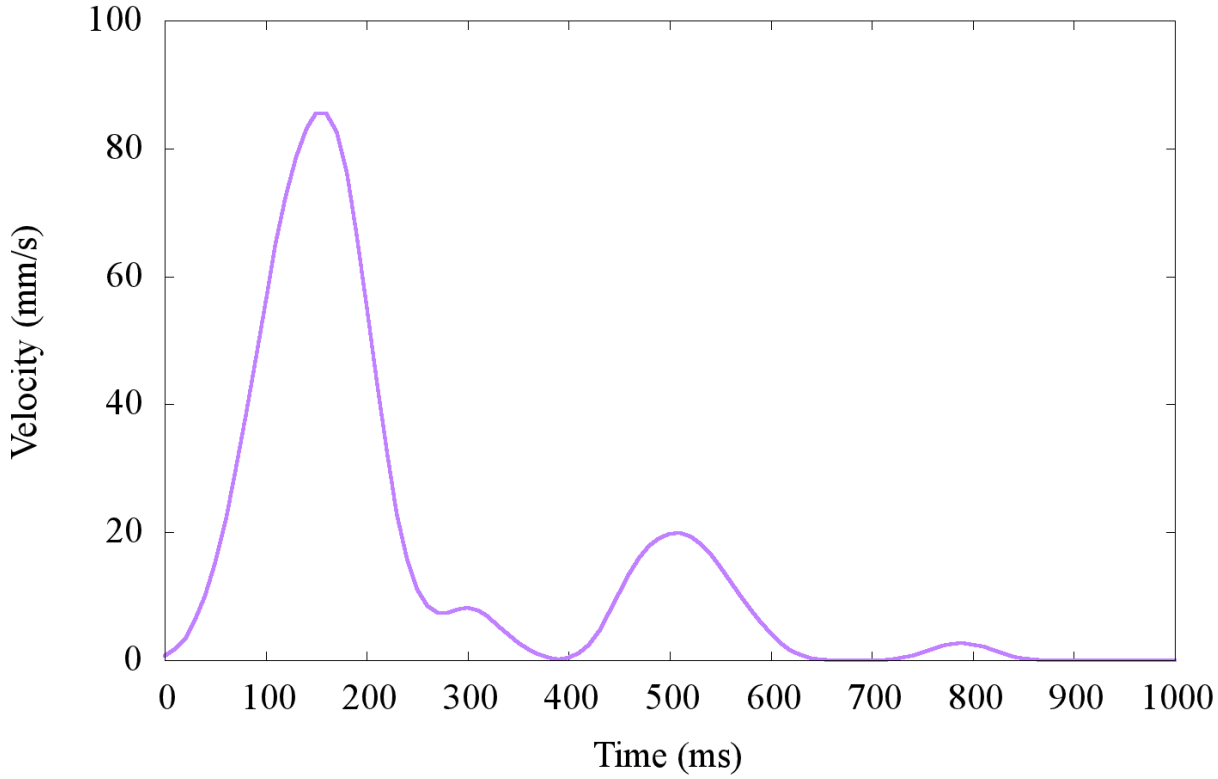


Number of zero crossings of a kernel is equal to the derivative order

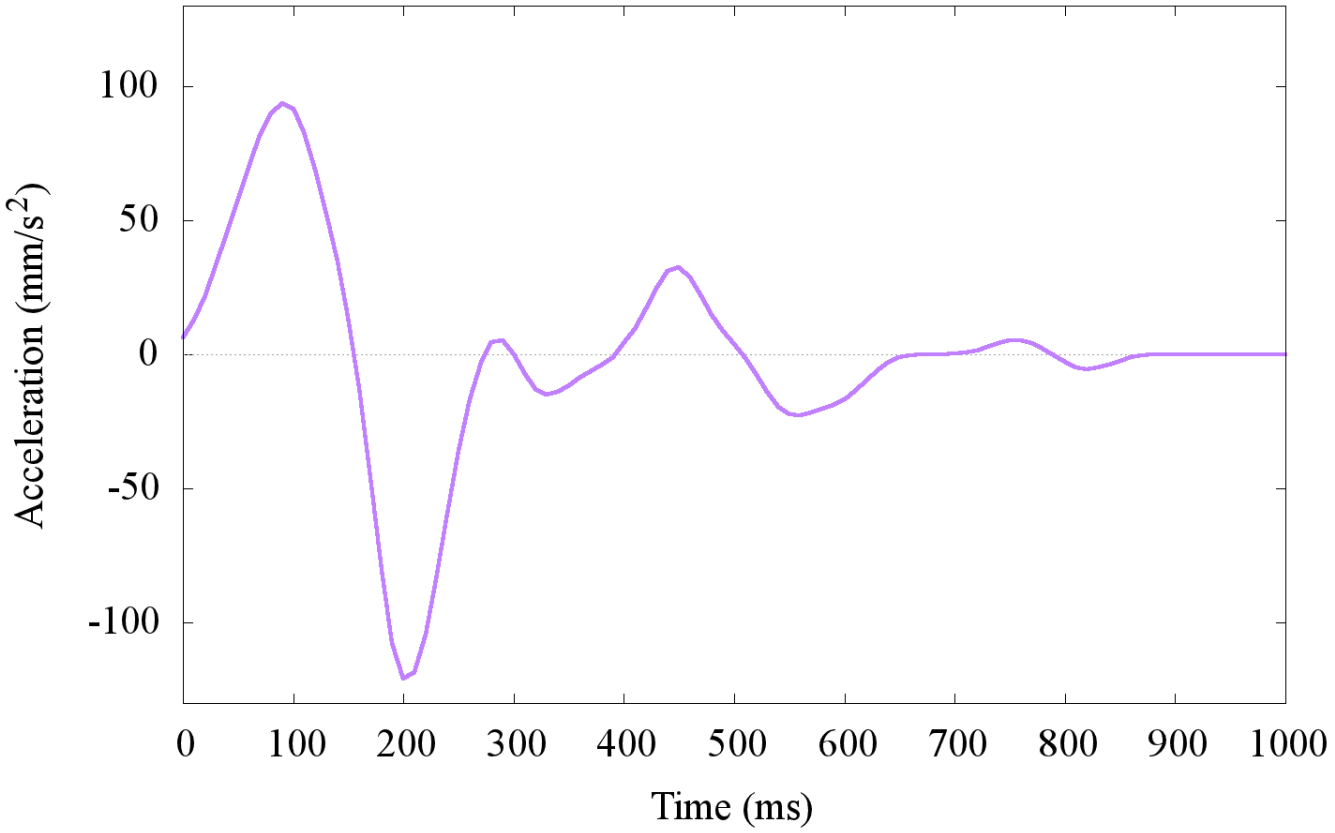
Even order: symmetric
Odd order: asymmetric

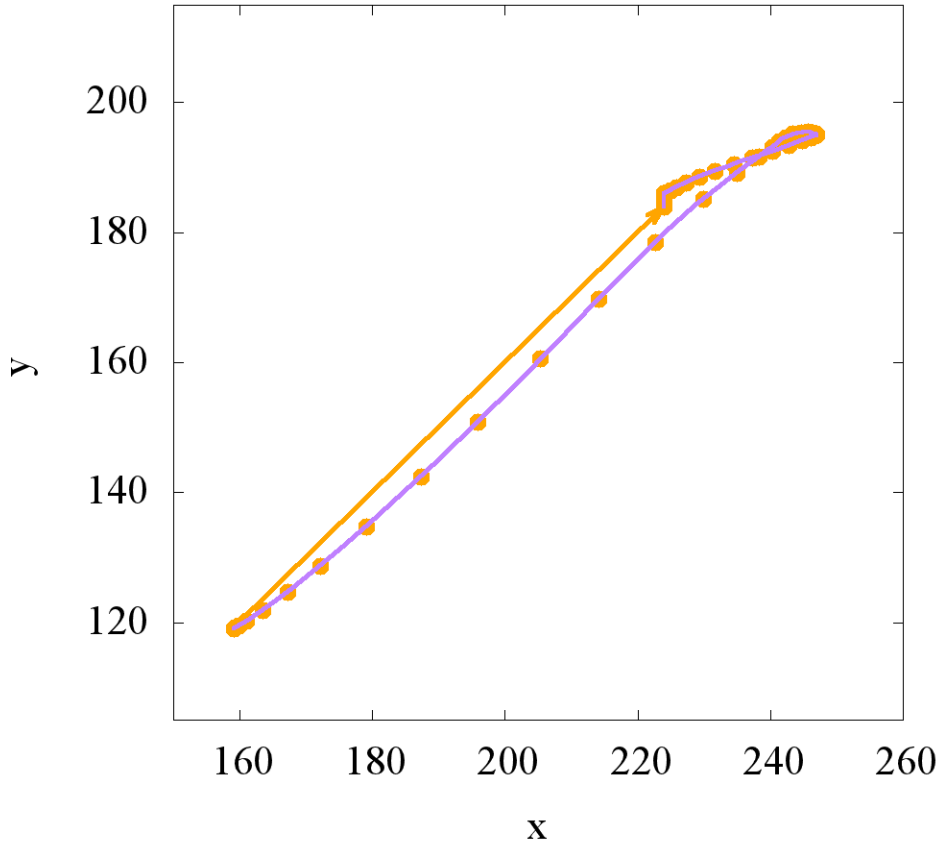
Mouse experiment: speed



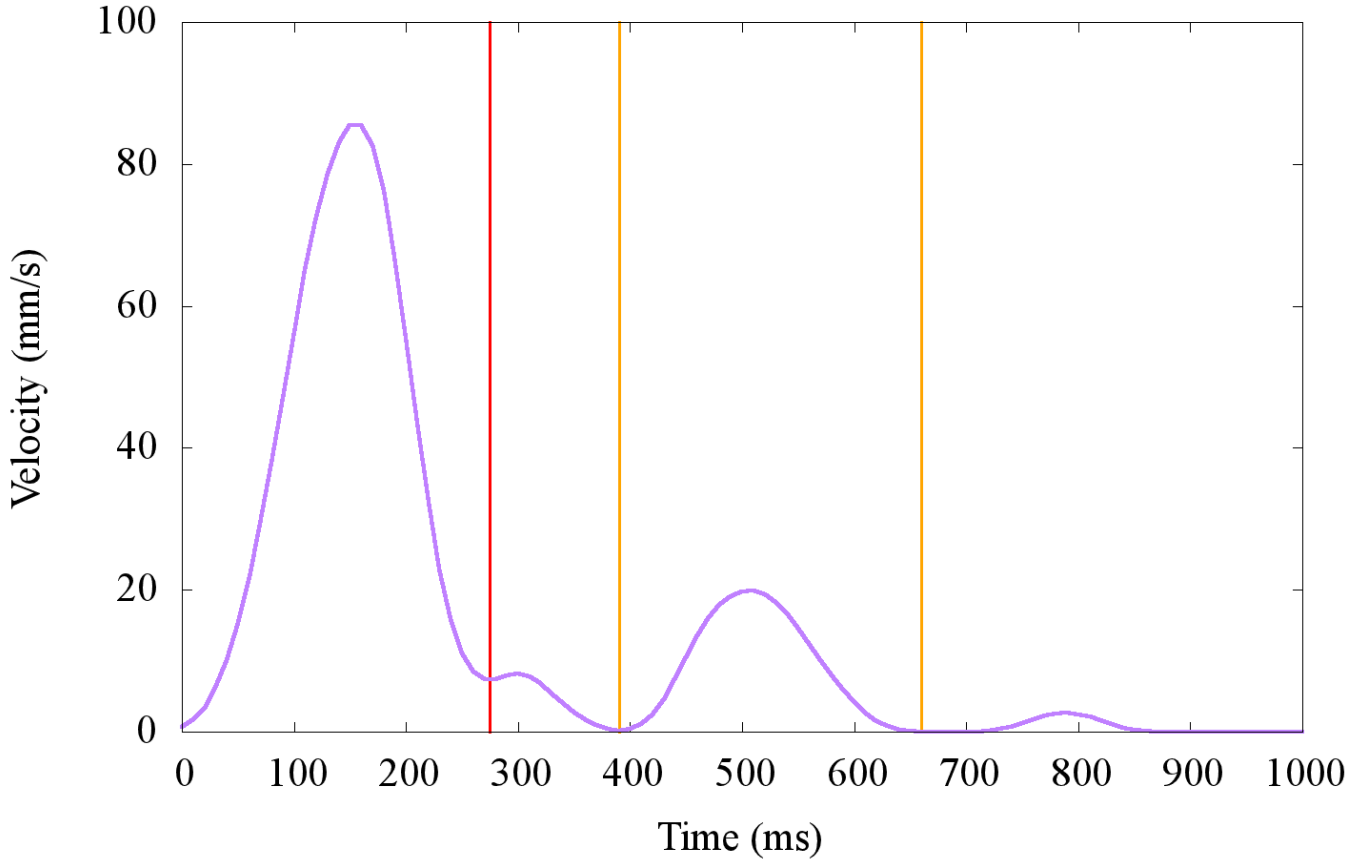


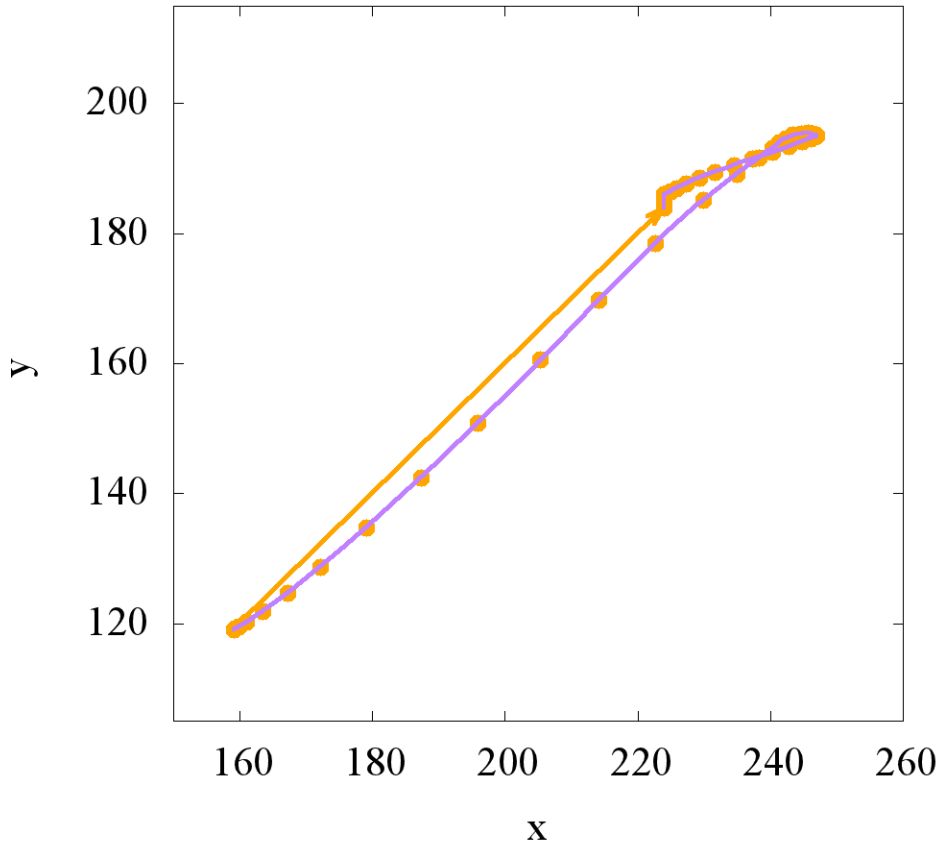
Acceleration (second derivative)





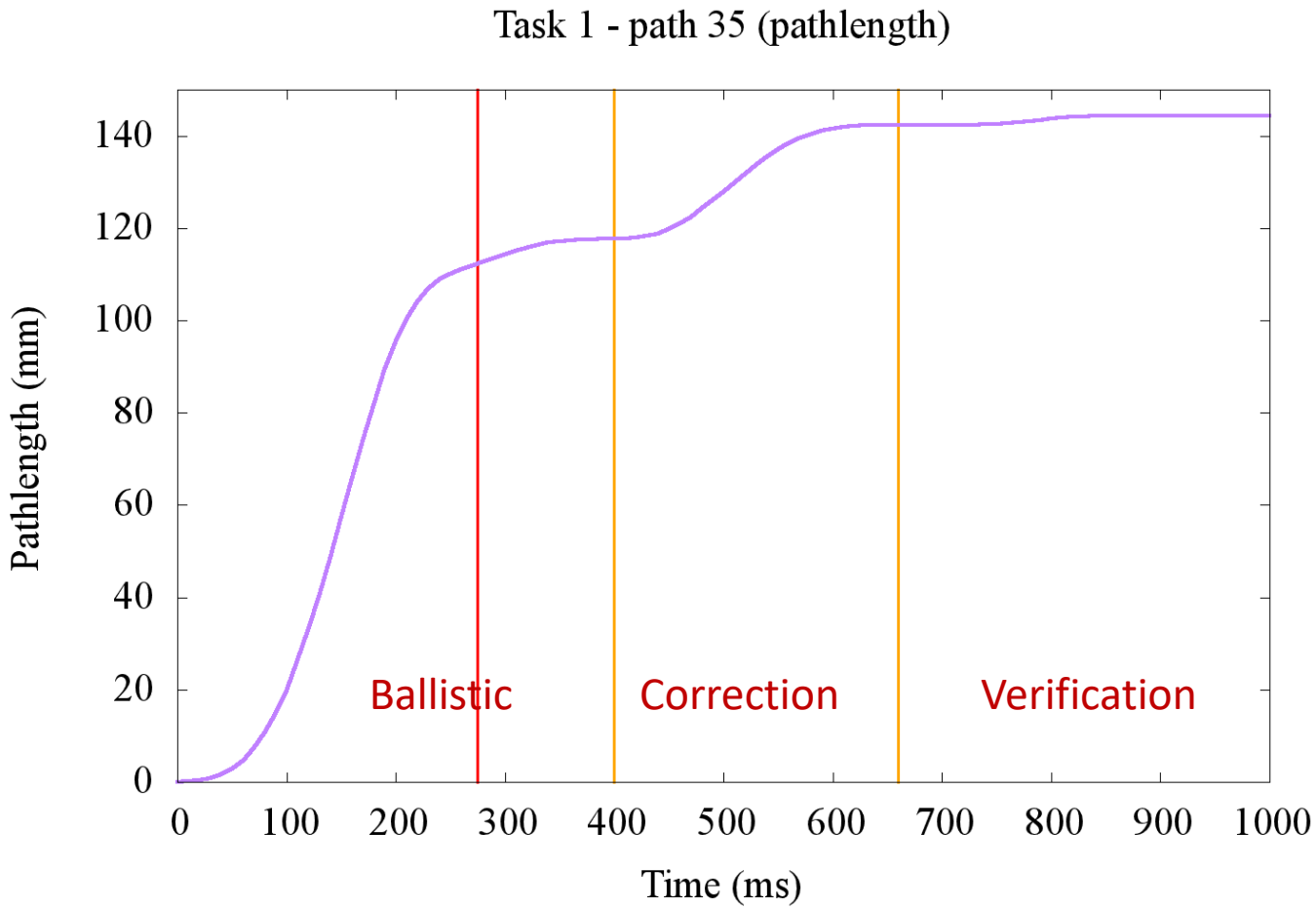
Sub-movement segmentation (on minima of speed)





Intuitive observation: first minimum (in red) is of a different type than the other two (in orange)

Path length with ballistic and correction movement



Apply additional segmentation rules to combine one or more sub-movements into ballistic movement, correction movement and verification phase

Summary

- All in all, we see that there is a huge potential for deriving different features.
- Let your research question be central when selecting which features to generate.

Summary of lecture

What should you know

- Scientific method and its elements
- Reliability, validity and threats to them
- Reasons and methods for data cleaning
- Filtering techniques
- Sampling methods
- Dependent, independent and confounding variables
- Numerical differentiation

What should you be able to do?

- Apply scientific method in practical settings
- Analyse threats to reliability and validity
- Choose and apply methods for data cleaning
- Choose and apply filtering techniques
- Choose and apply sampling methods
- Define dependent, independent and confounding variables in practical settings
- Perform numerical differentiation

Key insights

- The use of the scientific method enables obtaining valid and reliable results
- Data cleaning is time-consuming but indispensable
- An appropriate choice of a sampling method is very important
- Numerical methods can be used to compute many useful features

Next week

- basic concepts of probability distributions
- expressing uncertainty through confidence intervals
- performing and interpreting hypothesis tests for one-sample problems and two-sample problems

Literature

- Peter Pruzan Research Methodology: the aims, practices and ethics of science, Springer 2016
  - Sections 4.1, 5.1-5.3, 6.4
- Karin Nieuwenhuizen, Jean-Bernard Martens, Advanced modeling of selection and steering data: beyond Fitts' law, International Journal of Human-Computer Studies, Volume 94, 2016, Pages 35-52,