

Rail-Data

Prof. Dr. Raphael Pfaff

Fachhochschule Aachen

21. Februar 2019

Subsection 1

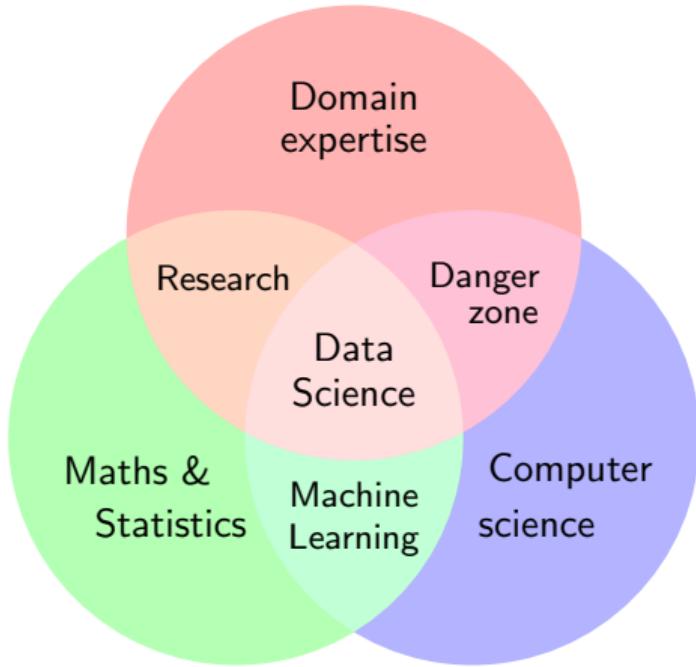
Brief Introduction to Data Science

What is Data Science?

Data science is our means of taming unstructured information and gathering insight. - Matthew Mayo, KDnuggets

- Interdisciplinary field of
 - Systems,
 - Methods and
 - Processs to extract insight or knowledge from data.
- Term coined in 2001, gained popularity in 2010
- Integrates:
 - Data Engineering
 - Scientific Method
 - Mathematics
 - Statistics
 - Advanced Computational Methods
 - Visualisation
 - Hacker Mindset
 - Domain Expertise

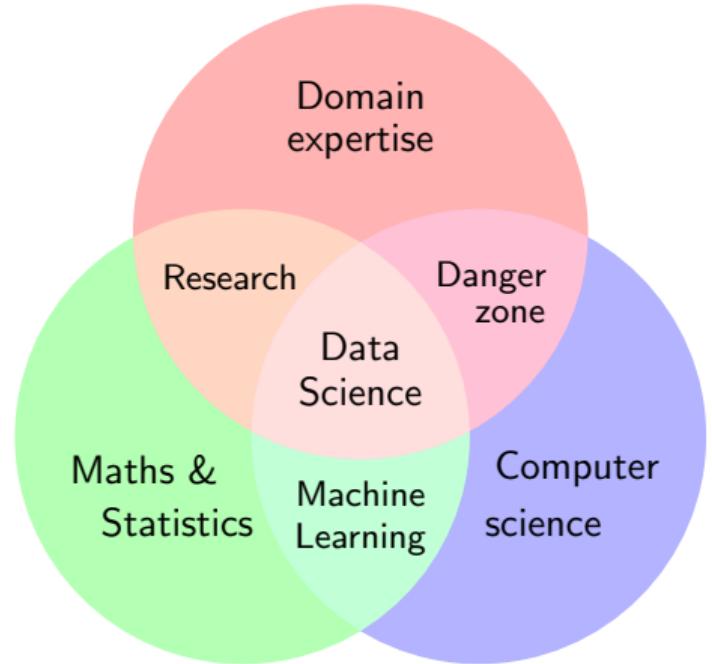
How does Data Science integrate to Mechanical Engineering?



Why Data Sciene?

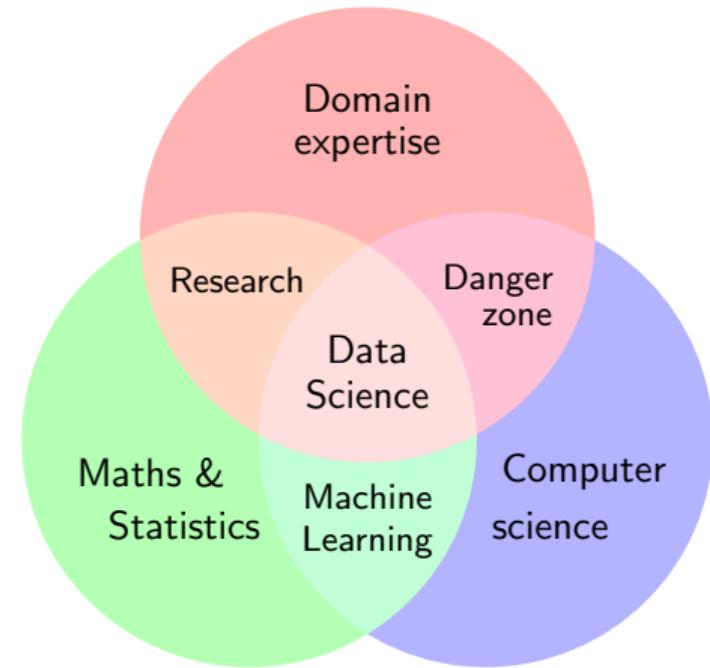
Applying the right tools and techniques, you bring more value!

- Turn data to information
 - Inform decisions
 - Increase insight
- Companies:
 - Collect large amounts of data
 - Do rarely integrate them
 - Frequently decide based on the “gut”



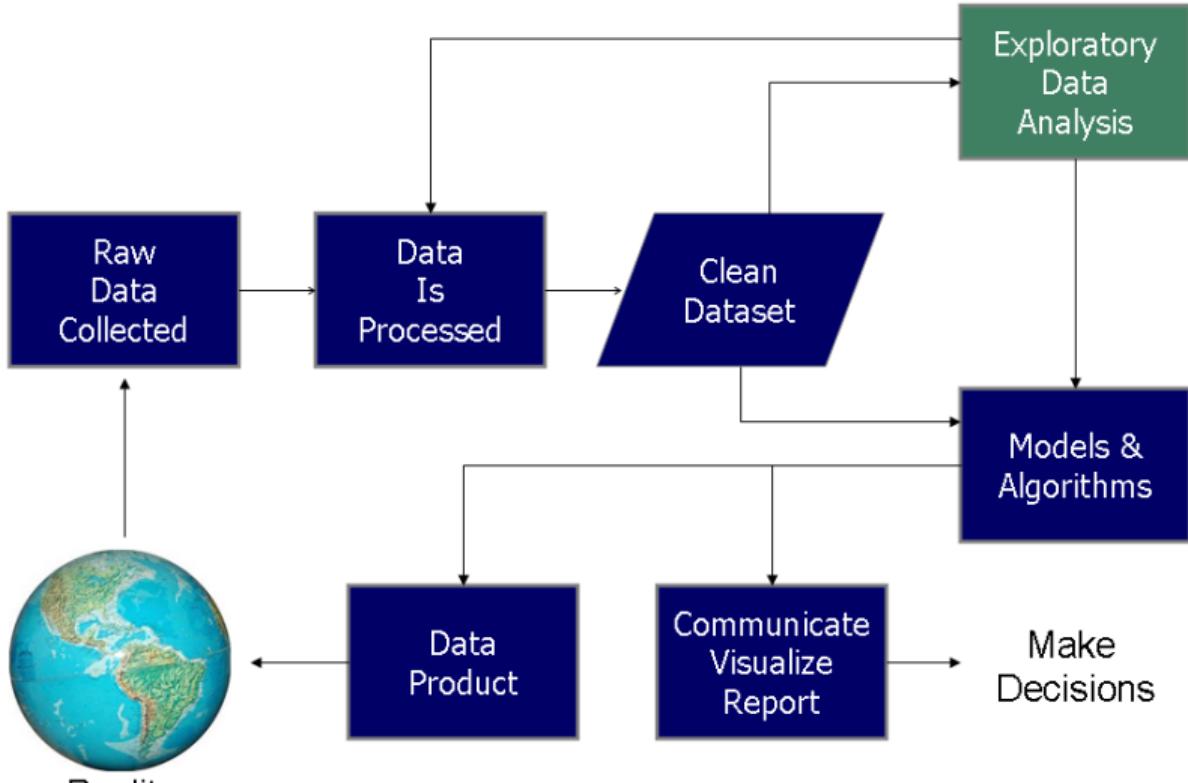
How do you increase Value with Data Science?

- Improve decision making
 - Empower management
 - Supply data driven evidence
- Identify trends and bring to action
- Challenge your colleagues
- Find opportunities for improvement
- Test decisions
- Understand customers



The data science process

Data Science Process



How to get started

- Set up your system.
 - Install Anaconda to obtain Python/Jupyter
 - Set up a free education account with github.com
 - Install the app for your operating system
- Acquire data: start with popular open data sets. Get your company to make data accessible.
- Ingest and transform: figure out the formats and sizes of your data. Find appropriate ways to import or access them.
- Explore the data. Do you already find patterns from just plotting them?
- Try your “toolbox” of methods (or a subset of it that sounds promising).
- Visualise the results. Make your findings convincing to others: colleagues, managers, customers etc.

Section 1

Data Science: Tools and Techniques

Tools for Data Scientist

- Programming languages:

- R
- Python
- SAS
- ...

- Visualisation app:

- Tableau

- Development Environment (IDE):

- Jupyter
- Spyder

- Also potentially:

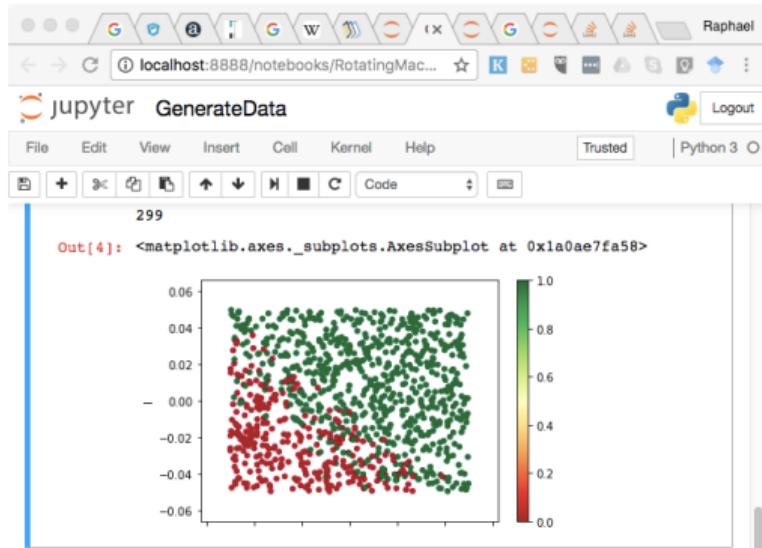
- Matlab
- Scilab
- ...

For installation, visit:

<https://www.anaconda.com/>

Data folder:

<https://bit.ly/2Jf12Q1>



```
Out[4]: SVC(C=inf, cache_size=200, class_weight=None, coef0=0.0,
decision_function_shape=None, degree=3, gamma='auto', kerne
l='linear',
max_iter=-1, probability=False, random_state=None, shrinkin
g=True,
tol=0.001, verbose=False)
```

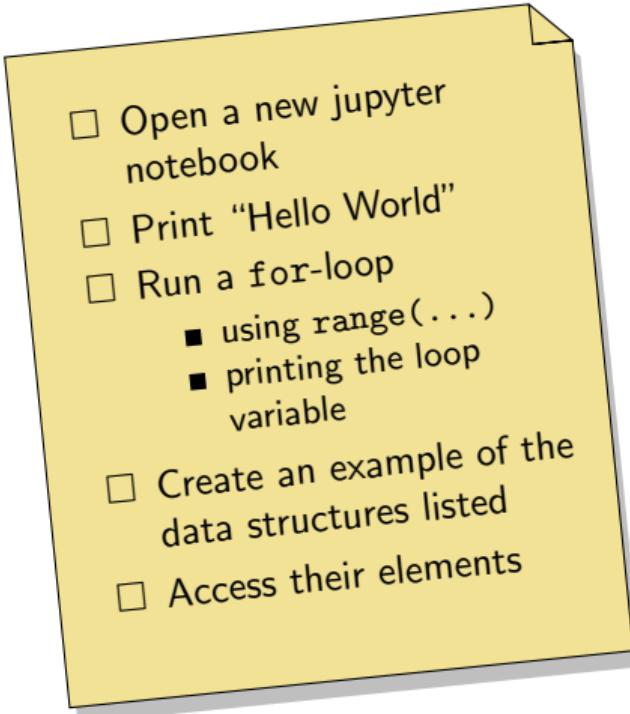
Python Introduction

Get started

Create your first jupyter notebook

- Python
 - is interpreted
 - can be run on terminal or IDE
- Indentation is strict, defines program structures (no braces)
- Dynamically typed
 - No variable declaration required
- Syntax
 - Refer to example worksheets
- Data structures
 - List: `l = [1, 2, 'a']`
 - Tuples: `t = (1, 2, 'a')`
 - Dictionaries: `d = {'a':1, 'b':2}`
 - Sets: `s = set()`

1, 2, 3, 4



Pandas Introduction

- Pandas is a framework for manipulation and handling of series and rectangular data
- Extremely efficient and powerful
- Syntax

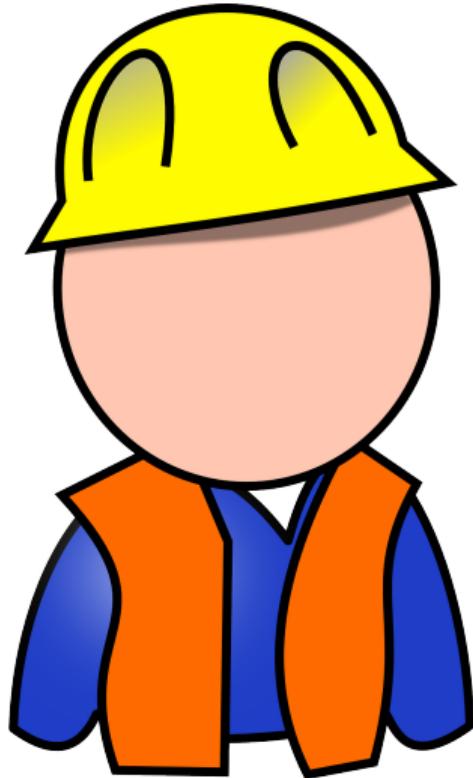
- ```
import pandas as pd
```
- ```
df = pd.DataFrame(...)
```
- ```
df.head()
```
- ```
df[‘Column’]
```
- ```
df[df[‘Column’] == 0]
```

- Data structures:
  - Series (`pd.Series`(from e.g. array))
  - DataFrame (`pd.DataFrame`(from e.g. `dicts`))

- Import Pandas
- Create a DataFrame (`pd.DataFrame`) from a dictionary
- Show the head of the DataFrame
- Show only certain parts of the DataFrame, e.g. all non-zeros entries

# Get to work!

Load and try 0-Basics.ipynb (<http://bit.ly/ifv122019> )



# Selected Techniques applied in Data Science

- **Visualisation**

- Regression: Linear, Logistic

- **Density Estimation**

- Confidence Intervals

- Test of Hypotheses

- Pattern Recognition

- Time Series

- **Unsupervised Learning (Clustering)**

- Supervised Learning

- Decision Trees

- **Monte-Carlo-Simulation**

- Bayesian Statistics

- **Principal Component Analysis**

- **Support Vector Machines**

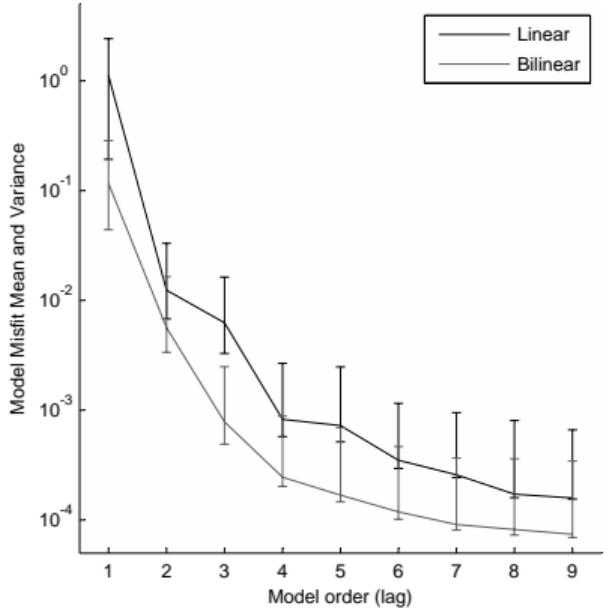
## Subsection 1

### Visualisation

# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

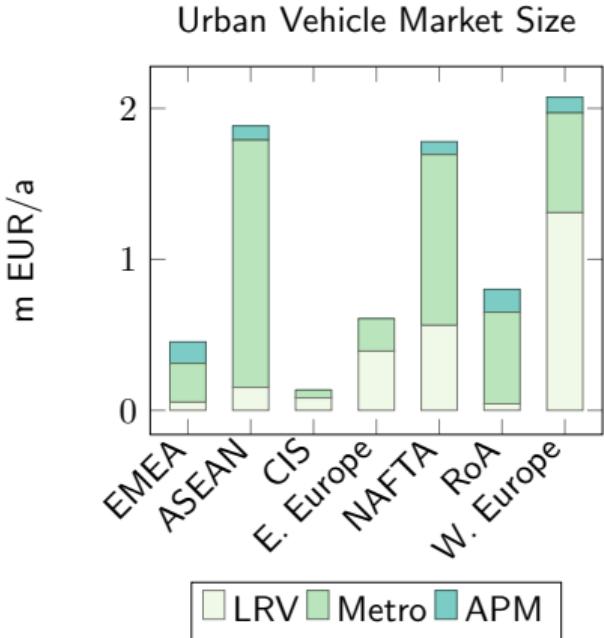
- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

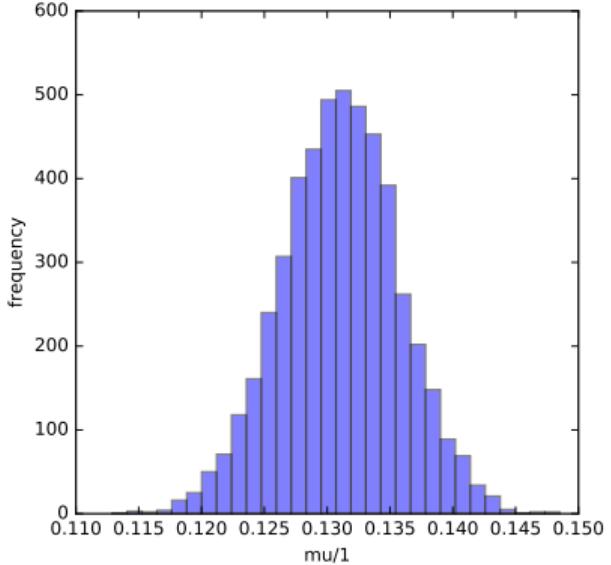
- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

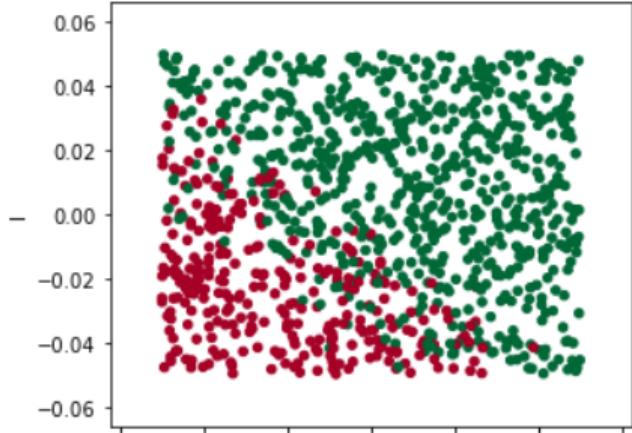
- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

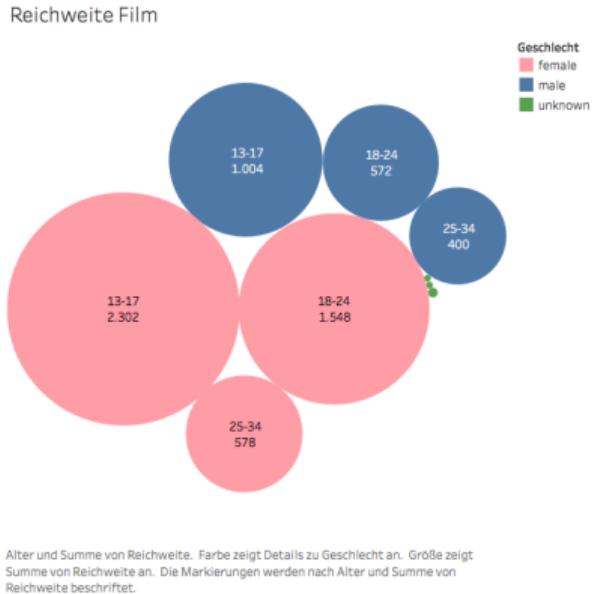
- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map

Intentionally left blank

# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

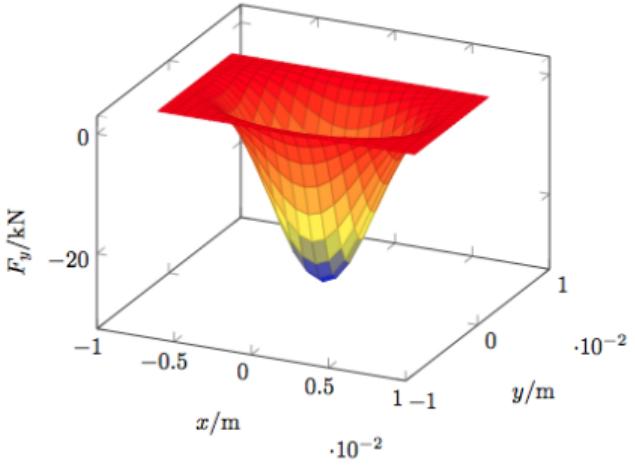
- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



# Visualisation approaches

Likely, you will all know Excel plots - there is much more to it...

- Line plot with / w/o confidence
- Bar plots (stacked/grouped)
- Histogram
- Scatter plot
  - Blob sizes
  - Colors
- Box whisker
- Bubble chart
- Geospatial plots
- Surface plot
- Heat map
- Tree map



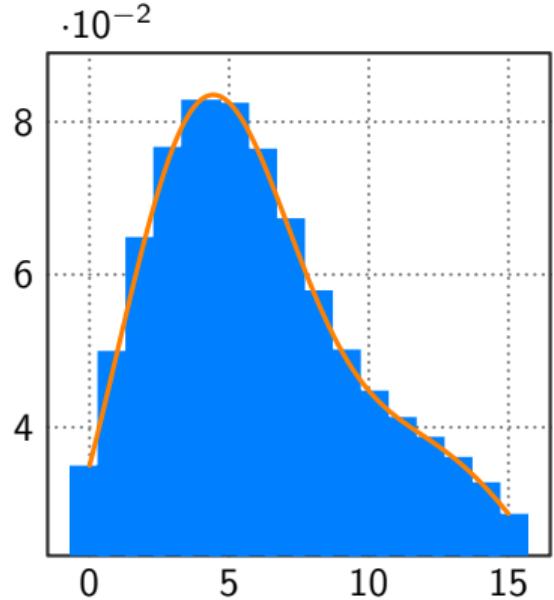
## Subsection 2

### Density Estimation

# Density estimation

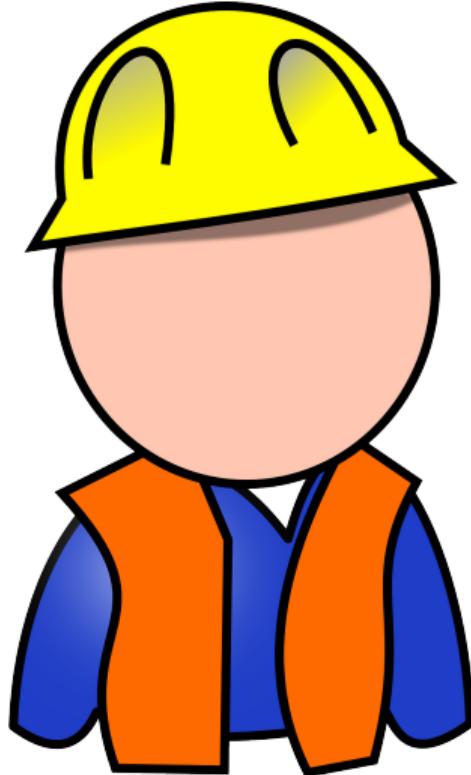
Estimate underlying probability density function from observed data

- Analysis of distribution properties from sampled data
- Fit an appropriate kernel to the samples



# Get to work!

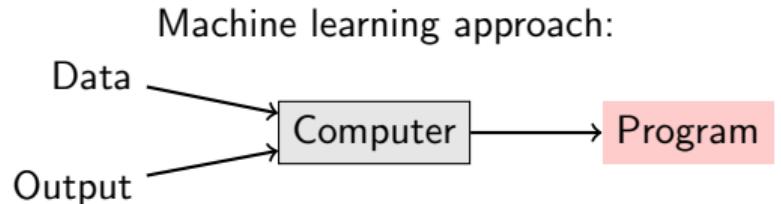
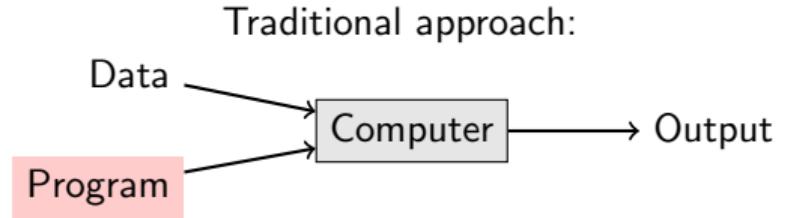
Load and try 1-Visualisation.ipynb (<http://bit.ly/ifv122019> )



# What is machine learning?

ML is a field of study that gives computers the ability to learn without being explicitly programmed.

- Different paradigm:
  - Derive rules from observations
- Typical procedure:
  - **Train:** Observe set of examples “training data”
  - **Fit:** Infer information of process behind data
  - **Test:** Make prediction on unseen data “Test data”
- Variations:
  - Supervised learning: provide labels with data
  - Unsupervised learning: cluster data based on patterns



## Subsection 3

### Unsupervised Learning (Clustering)

# Clustering

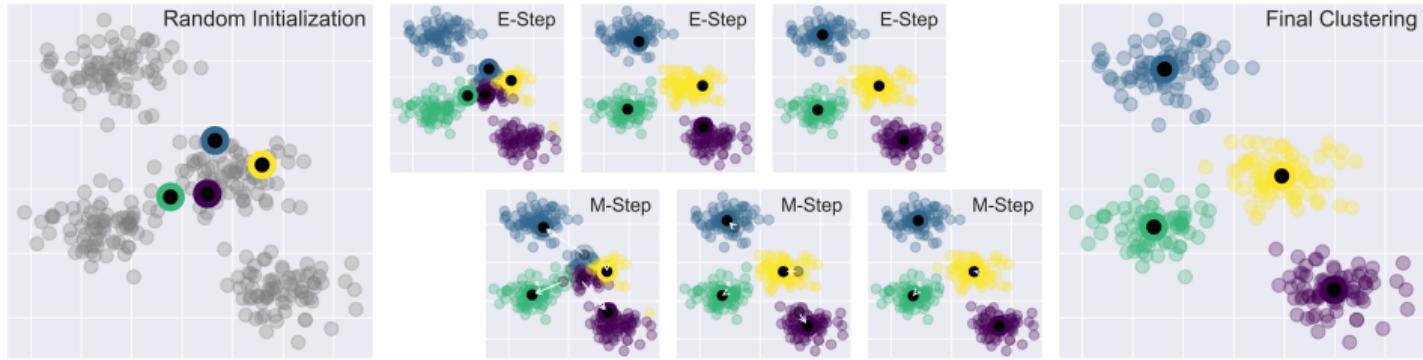
Two fundamental approaches,  $k$  number of clusters defined *a priori*

## Hierarchical clustering

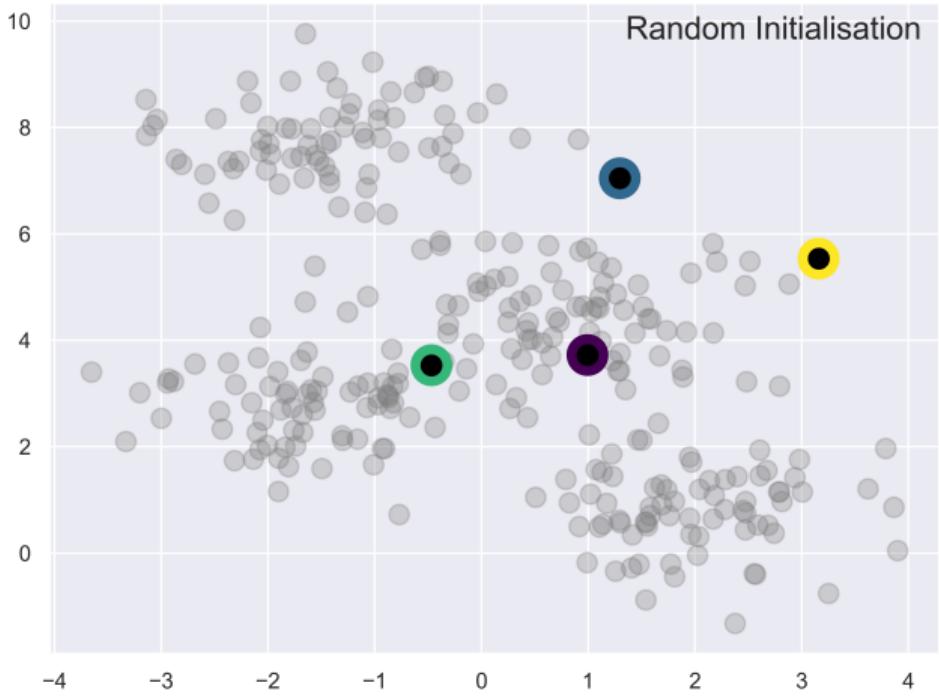
- 1 Assign each item to one cluster
- 2 Find closest (most similar cluster): merge them
- 3 Continue until desired number of clusters  $k$  is reached

## K-means clustering

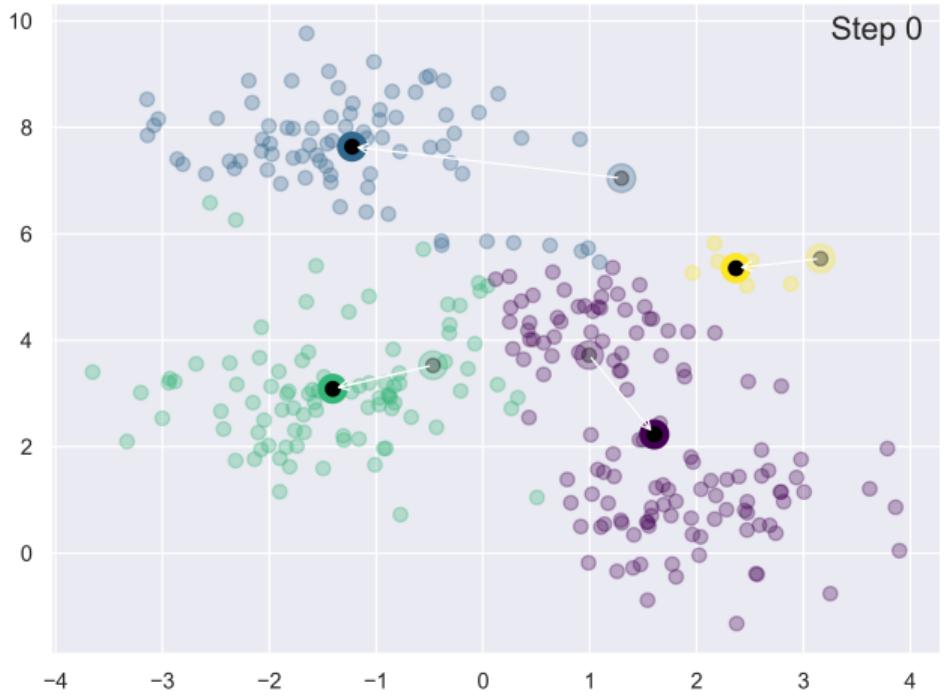
- 1 Randomly choose  $k$  initial centroids
- 2 While centroids don't change
  - E: Create  $k$  clusters by assigning each item to closest centroid
  - M: Compute  $k$  new centroids by averaging items in each cluster



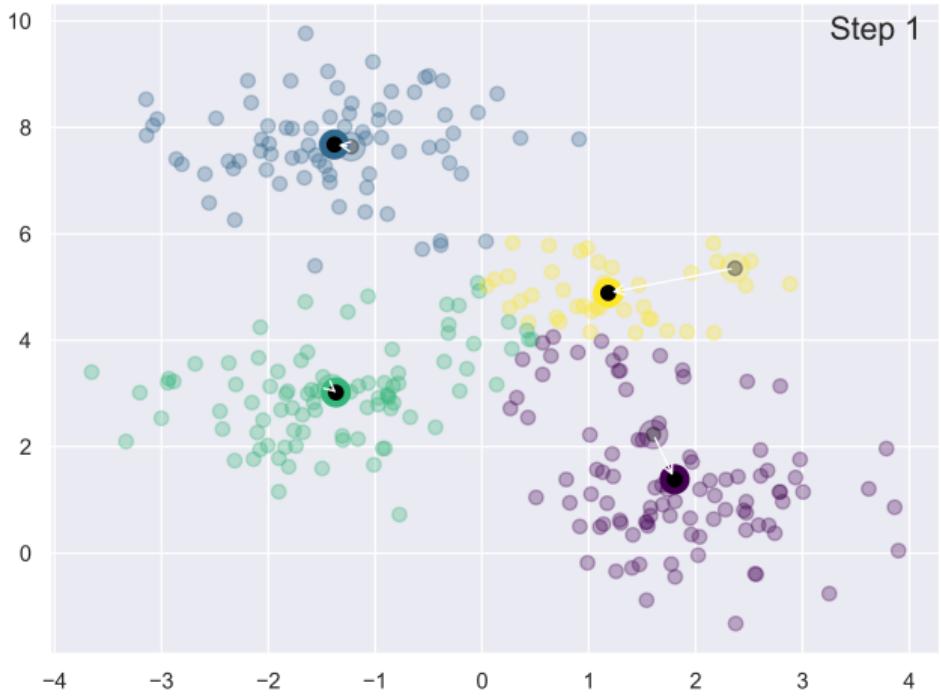
# K-means clustering procedure



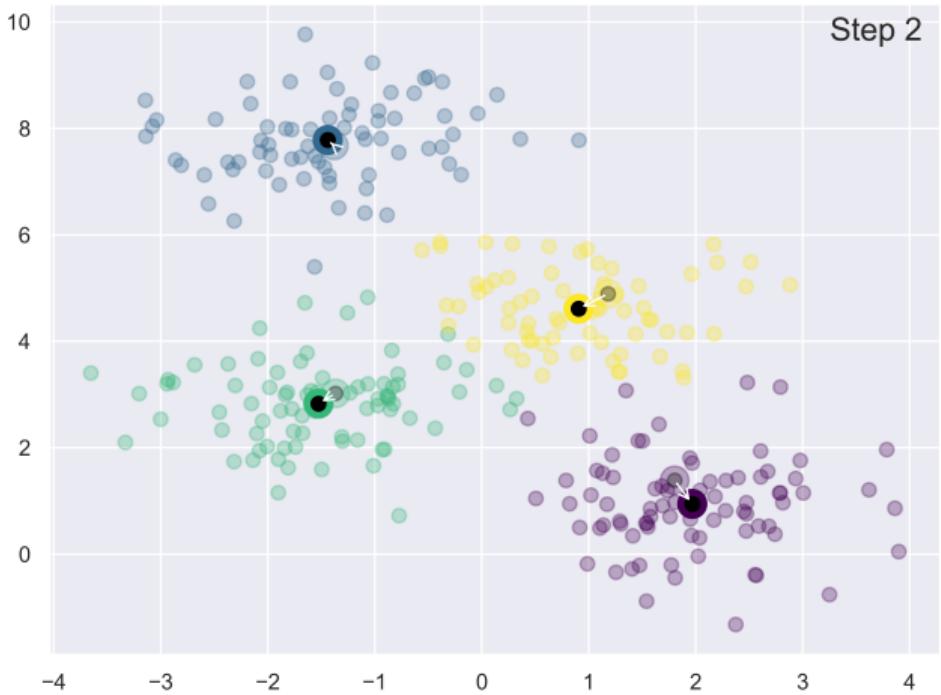
# K-means clustering procedure



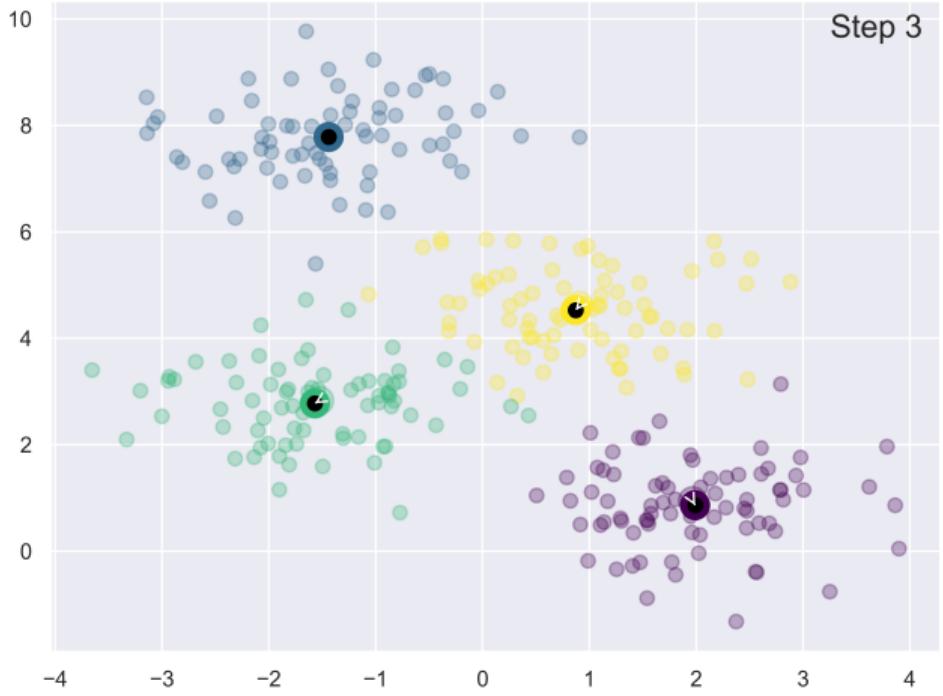
# K-means clustering procedure



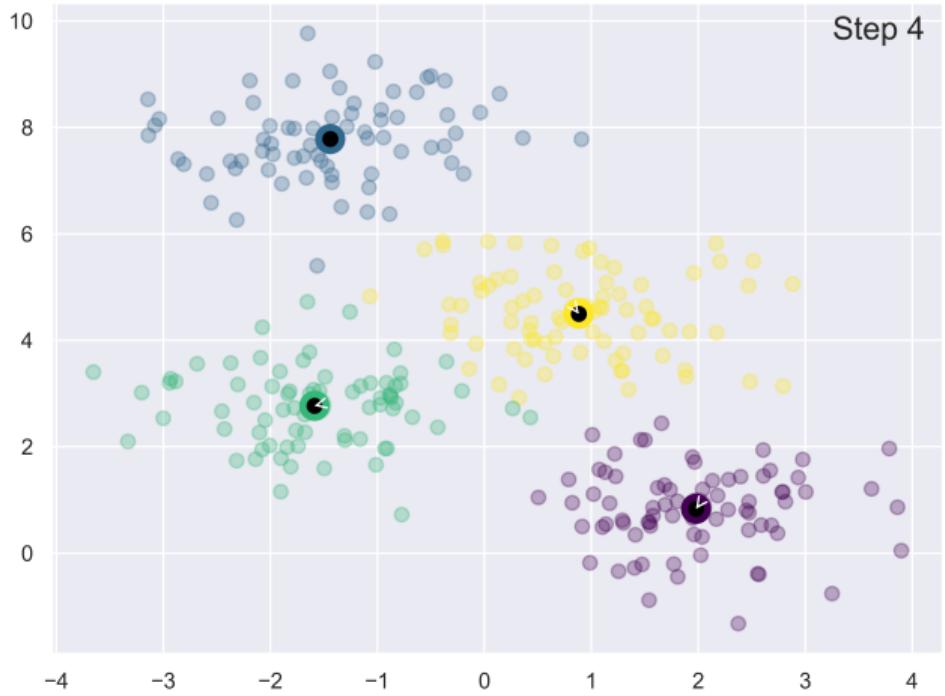
# K-means clustering procedure



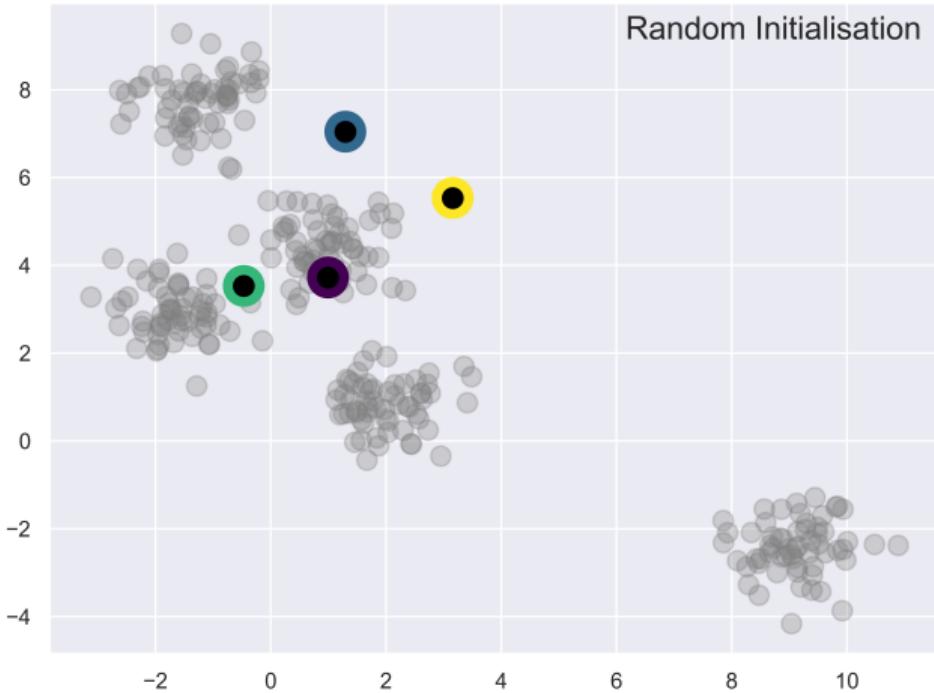
# K-means clustering procedure



# K-means clustering procedure



# K-means inappropriate cluster number



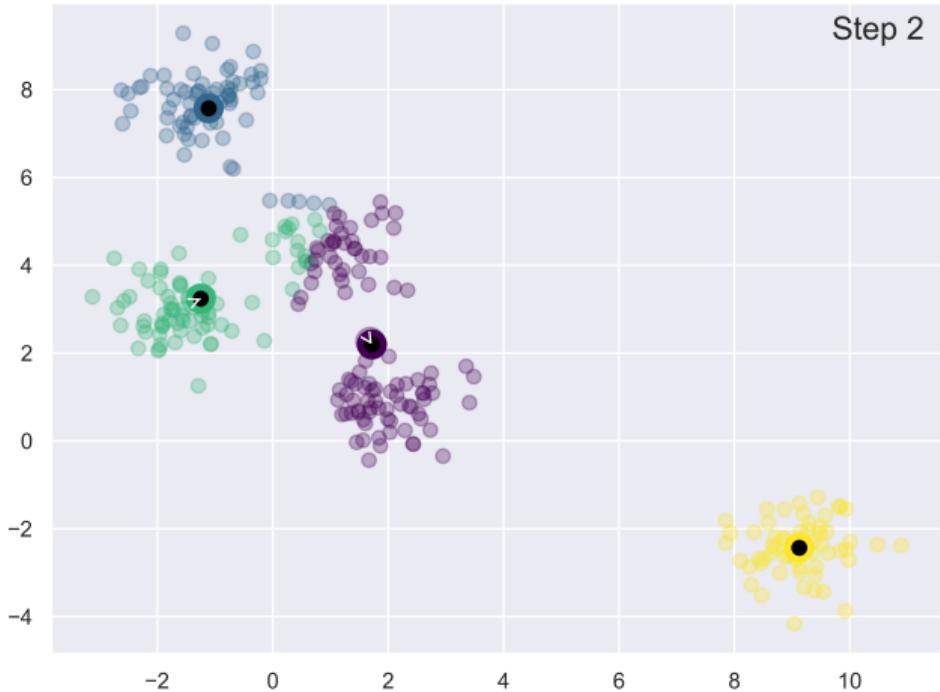
# K-means inappropriate cluster number



# K-means inappropriate cluster number



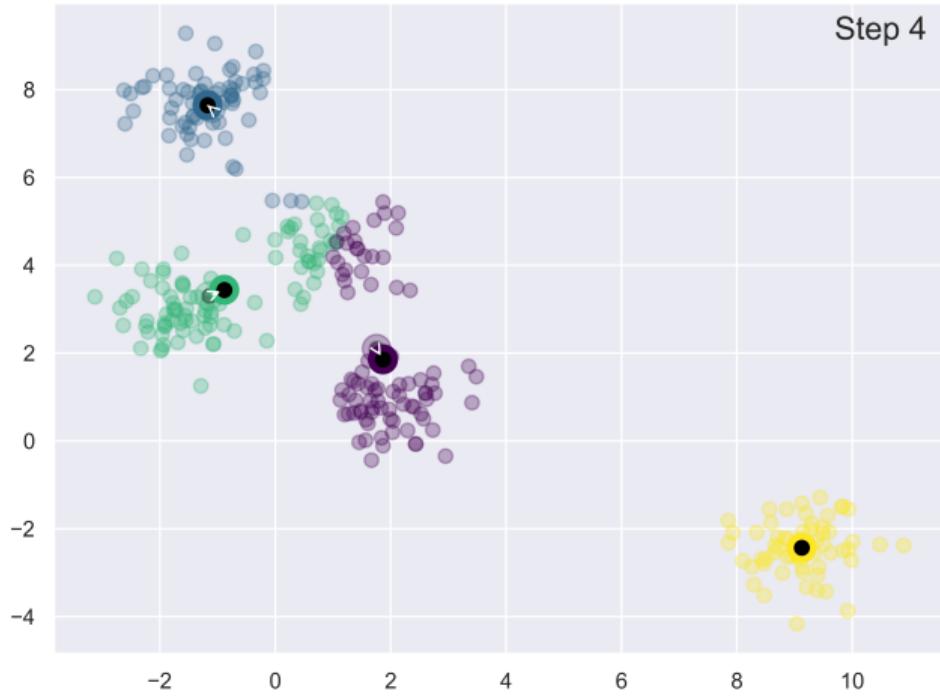
# K-means inappropriate cluster number



# K-means inappropriate cluster number



# K-means inappropriate cluster number



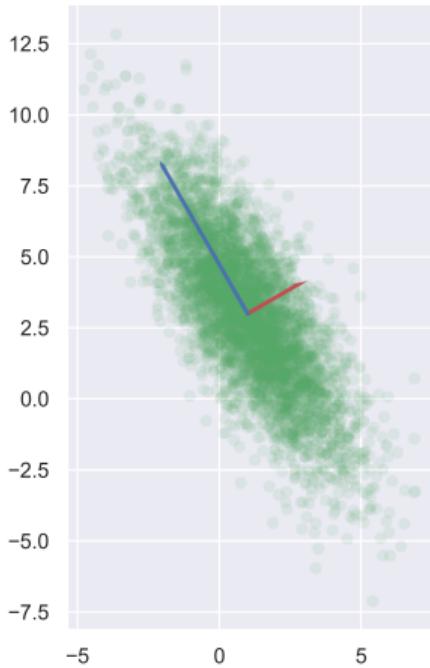
## Subsection 4

# Principal Component Analysis

# Principal Component Analysis (PCA)

Reduce dimensionality of data

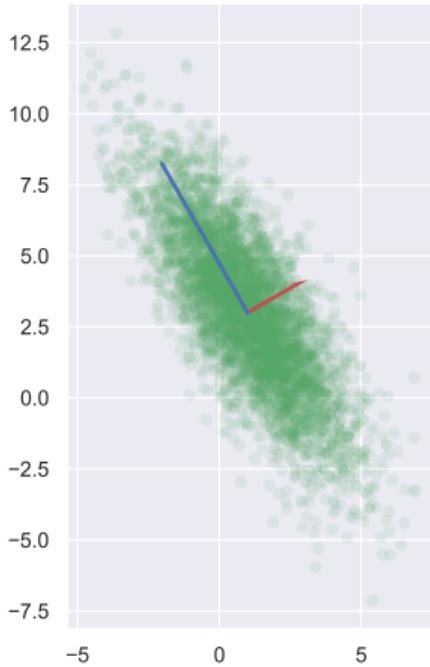
- Sample  $X_1, \dots, X_n$  is a cloud of points in  $\mathbb{R}^d$
- Typically,  $d$  is large
- **Question:** can we project the point cloud onto a linear subspace of dimension  $d' < d$  while keeping as much of the information as possible?



# Principal Component Analysis (PCA)

Reduce dimensionality of data

- Sample  $X_1, \dots, X_n$  is a cloud of points in  $\mathbb{R}^d$
- Typically,  $d$  is large
- **Question:** can we project the point cloud onto a linear subspace of dimension  $d' < d$  while keeping as much of the information as possible?
- **Answer:** PCA achieves this by keeping important aspects of the covariance structure in orthogonal directions.



## Subsection 5

### Support Vector Machines

# Support Vector Machine

Separate data into subsets according to their  $nD$ -coordinates.

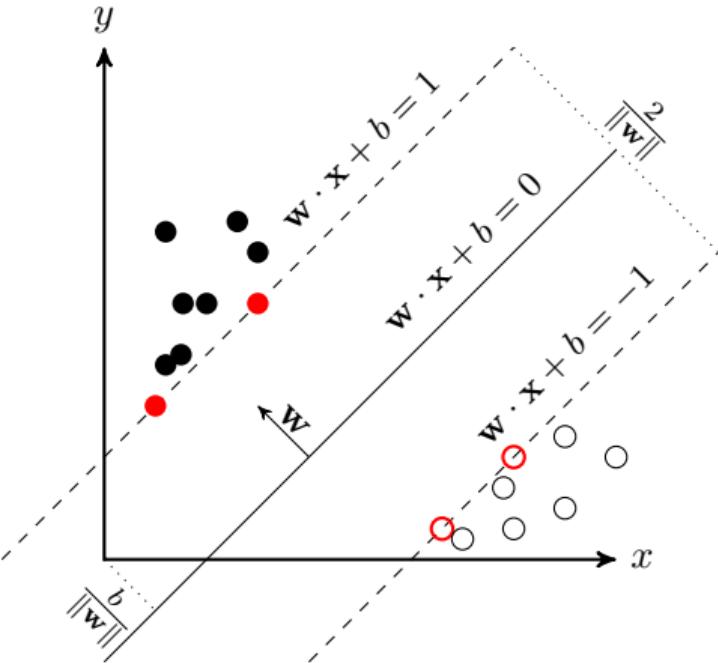
## ■ Idea:

- Find separating hyperplane maximising the distance between borderline instances
- For data mixed in feature space, slack parameter (and penalty  $C$ ) is used
- If no separating hyperplane can be found in feature space, dimension is increased "Kernel trick"

## ■ Robust:

- High dimensionality
- Small datasets

## ■ Simple to complex models



# Support Vector Machine

Separate data into subsets according to their  $nD$ -coordinates.

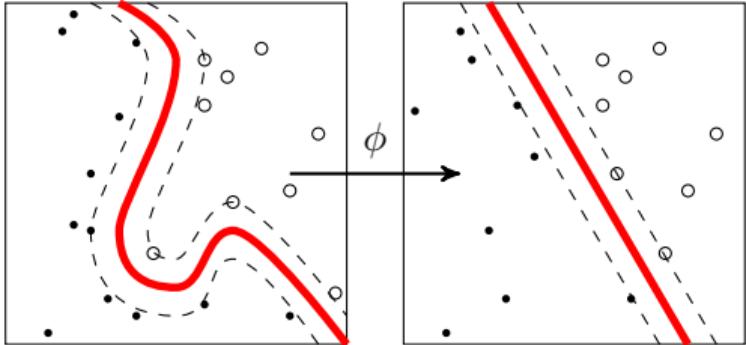
- Idea:

- Find separating hyperplane maximising the distance between borderline instances
- For data mixed in feature space, slack parameter (and penalty  $C$ ) is used
- If no separating hyperplane can be found in feature space, dimension is increased "Kernel trick"

- Robust:

- High dimensionality
- Small datasets

- Simple to complex models



## Subsection 6

### Monte-Carlo-Simulation

# Monte-Carlo-Simulation (MC-Simulation)

- Method of estimating value of unknown quantity using inferential statistics
- Inferential statistics terms:
  - Population: set of examples
  - Sample: Proper subset of population
- Random sample tends to exhibit same properties as population it is drawn from

## Example (Flip coins)

Let's estimate the probabilities of heads vs. tails for an infinite number of coin flips:

- One flip (heads): 100% heads?
- Two flips (h, h): still 100 % heads? Confidence level?
- 100 flips (52 h, 48 t): Probability of next coin coming up heads 52/100.

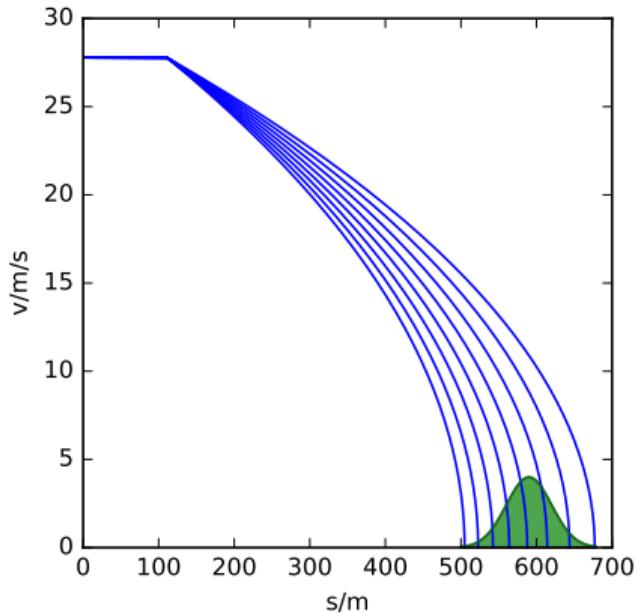
# Key aspects to be asked in MC-Simulations

- Never possible to guarantee perfect accuracy through sampling
- Never assume that an estimate is precisely correct
- How many samples do we need to look at before we can have justified confidence in our answers?
  - Answer depends on underlying distribution
  - Especially hard for defects “Rare event simulation”

# MC-Simulation application example: Braking curve

CCS systems rely on braking curves to describe the train's braking capability.

- To supervise train velocity, CCS systems predict the future braking capability of the train
- However, there is not *the* braking capability
- Braking curves exhibit a randomised behaviour



## Subsection 7

### Practical hints

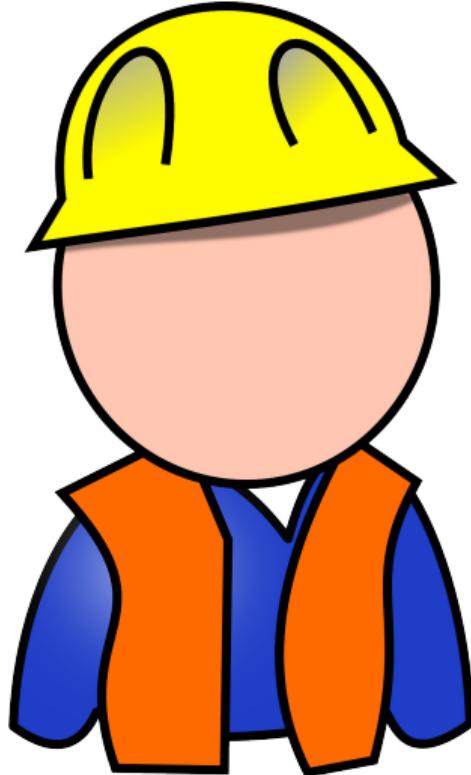
# Practical hints for application of data science methods

- Get to know your data
- Use training and validation sets
- For clustering and ML:
  - Regularise
  - Remove outliers, NaN
- When building models:
  - Don't expect perfect fit
  - Inspect confusion matrix

|              |    | Prediction outcome |                | total |
|--------------|----|--------------------|----------------|-------|
|              |    | p                  | n              |       |
| actual value | p' | True Positive      | False Negative | P'    |
|              | n' | False Positive     | True Negative  | N'    |
| total        |    | P                  | N              |       |

# Get to work!

Load and try 2-Container-SVM.ipynb (<http://bit.ly/ifv122019> )



## Subsection 8

### Recap Fertigungsmesstechnik

# Definition Fertigungsmesstechnik

[...] Oberbegriff für alle mit Mess- und Prüfaufgaben verbundenen Tätigkeiten, die beim industriellen Entstehungsprozess eines Produktes zu erbringen sind. [2, S. 1]

- Im Entstehungsprozess
- Messen und Prüfen
- Aspekte der Definition:
  - Geringe Fertigungstiefe
  - Automatisierung
  - gestiegene Qualitätsforderungen
- Von Kontrollinstanz zu Komponente des QM

# Frühe Entwicklung und Rolle der Fertigungsmesstechnik

- Messtechnik seit ca. 4000 v. Chr.: Vergleich mit natürlichen Maßen
- Festlegung Urmeter 1799
- Beschleunigt durch Austauschbau und Massenfertigung
- Elektronische Messtechnik ca. seit 1970

## ■ Rolle der Messtechnik:

- 20er Jahre: Sortierung
- 30er Jahre:  
Prozessüberwachung und  
Regelung
- 80er Jahre: Planung zur  
Fehlervermeidung
- 90er Jahre:  
Gesamtheitliches  
Qualitätsdenken



# Aufgaben und Ziele der Fertigungsmesstechnik

- Erfassung von Qualitätsmerkmalen an Messobjekt
  - Werkstoffeigenschaften (z.B. Gefüge, Härte)
    - Als Eingangsprüfung
    - Nach thermischer Behandlung
    - auch für Schweißnahtgüte
  - Geometrie (z.B. Maß, Form, Lage)
    - Dominierende Prüfung
    - Gestalt des Werkstücks
    - Oberflächeneigenschaft
  - Funktion (z.B. Kraft, Geschwindigkeit)
    - Produkt oder Baugruppe
    - Von Sichtprüfung bis zur vollautomatisierten Funktionsprüfung

# Größenordnungen und Definitionen

- Messgröße Länge: üblicherweise  $(10^{-9} \dots 10^2)$  m
- Toleranzen zunehmend enger

## Definition (Messgröße)

Die Messgröße ist die physikalische Größe, der die Messung gilt.

## Definition (Messen)

Messen ist das Ausführen von geplanten Tätigkeiten zum Vergleich der Messgröße mit einer Einheit.

## Definition (Prüfen)

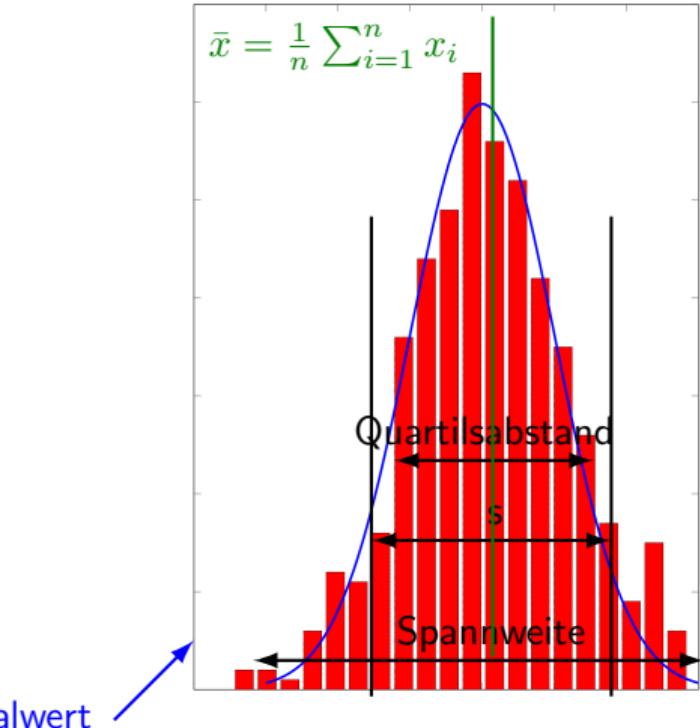
Prüfen heißt feststellen, inwieweit ein Prüfobjekt eine Forderung erfüllt.

## Subsection 9

### Prüfdatenauswertung

# Beschreibung

Beschreibung der Verteilung durch Lage und Streuungsparameter verdichtet die information.



## ■ Lageparameter:

- Arithmetisches Mittel  $\bar{x}$
- Modalwert: am häufigsten angenommene Klasse
- Median: je 50% der Messwerte größer bzw. kleiner

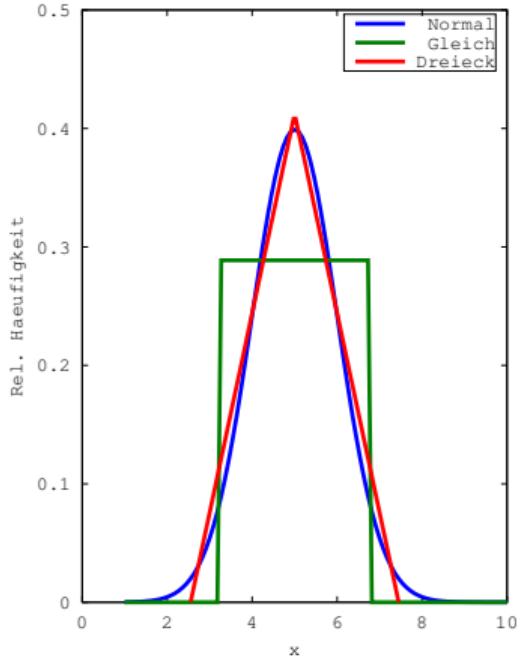
## ■ Streuungsparameter

- Spannweite
- Quartilsabstand: Spannweite der zentralen 50% der Messwerte
- Standardabweichung  $s$
- Quantile: Merkmalsausprägung, für die ein Anteil  $\alpha$  kleiner ist

# Verteilungen

Zufallsmodelle können helfen, beobachtete Phänomene zu beschreiben und über die Modellbildung vorhersagen zu treffen

- Näherungsweise Beschreibung
- I.d.R. Konvergenz für große Stichproben
- Wahrscheinlichkeit kann als relative Häufigkeit interpretiert werden
- Verteilung stetiger und diskreter Merkmale unterschiedlich zu modellieren



# Verteilungen diskreter Merkmale

- Poisson-Verteilung:

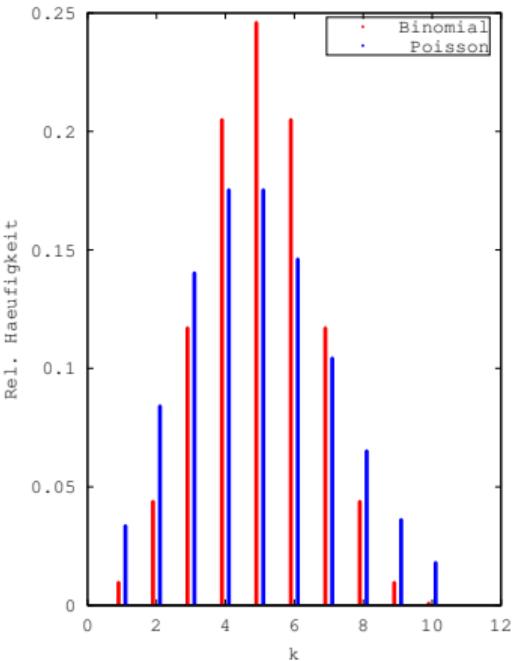
$$p_P(k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

- Geringe Wahrscheinlichkeiten, z.B.  
 $p \leq 0,1$

- Binomial-Verteilung:

$$p_B(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Höhere Wahrscheinlichkeiten, z.B.  
 $p > 0,1$



# Verteilungen kontinuierlicher Merkmale

- Normalverteilung:

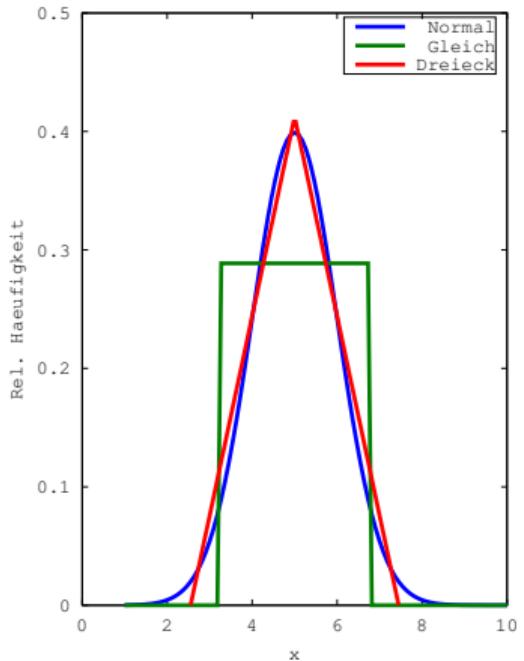
$$p_N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

→ Faltung vieler Verteilungen

- Gleichverteilung

- Dreieckverteilung

→ Faltung zweier  
Rechteckverteilungen



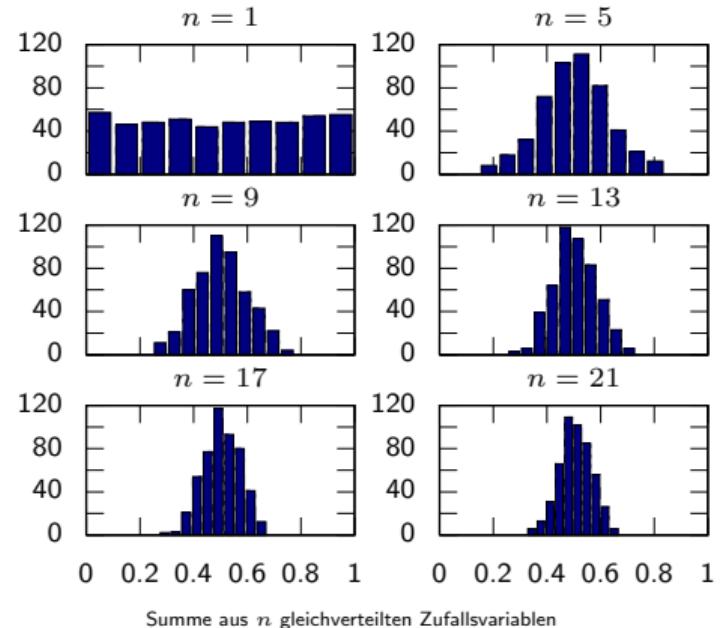
# Zentraler Grenzwertsatz

Konvergenz der Summe von Zufallsvariablen gegen die Standardnormalverteilung.

- Sei  $X_1, X_2, X_3, \dots$  eine Folge von Zufallsvariablen, die auf demselben Wahrscheinlichkeitsraum unabhängig und identisch verteilt sind.
- Sei weiterhin  $S_n = X_1 + X_2 + \dots + X_n$  die  $n$ -te Teilsumme, eine Zufallsvariable mit  $E(S_n) = n\mu$  und  $\text{Var}(S_n) = n\sigma^2$
- Dann konvergiert die Verteilungsfunktion der standardisierten Zufallsvariablen

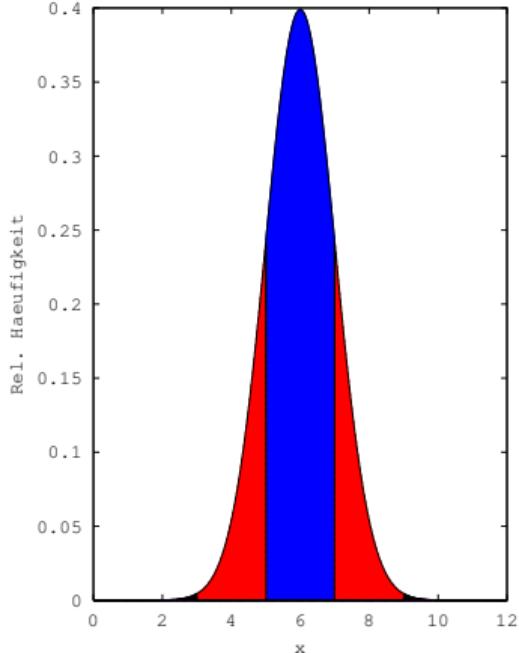
$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

für  $n \rightarrow \infty$  punktweise gegen die Standardnormalverteilung  $\mathcal{N}(0, 1)$ .



# Eigenschaften der Normalverteilung

Viele Phänomene in Technik und Naturwissenschaft lassen sich durch eine Normalverteilung annähern.



- Mittelwert  $\mu$
- Standardabweichung  $\sigma$
- $\sigma$ -Umgebungen:

| $k$ | % der Realisierungen |
|-----|----------------------|
| 1   | 68,3                 |
| 2   | 95,5                 |
| 3   | 99,4                 |

- Quantile:

| % der Realisierungen | $k$   |
|----------------------|-------|
| 50                   | 0,675 |
| 90                   | 1,65  |
| 95                   | 1,96  |
| 99                   | 2,58  |

# Vertiefung: Konfidenzintervall, Bestimmung $u$

Konfidenzintervalle sind Intervalle in denen der wahre Wert einer Merkmalsausprägung mit Wahrscheinlichkeit  $p$  liegt.

- Größe und Lage abhängig von

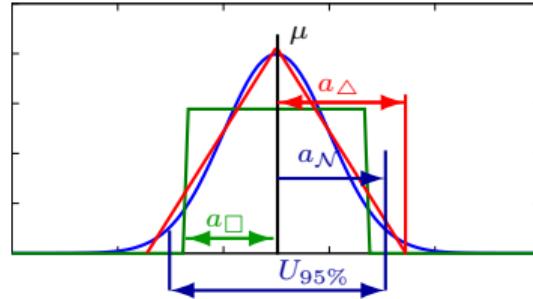
- Verteilung
- Konfidenzniveau (z.B. 95%)

- Für Normalverteilung:

| $k$ | % der Realisierungen |
|-----|----------------------|
| 1   | 68,3                 |
| 2   | 95,5                 |
| 3   | 99,4                 |

- Fertigungsmesstechnik:

95%-Konfidenzintervall  $U_{95\%}$  für  
 $k = 2$



$$u = \frac{a}{K}$$

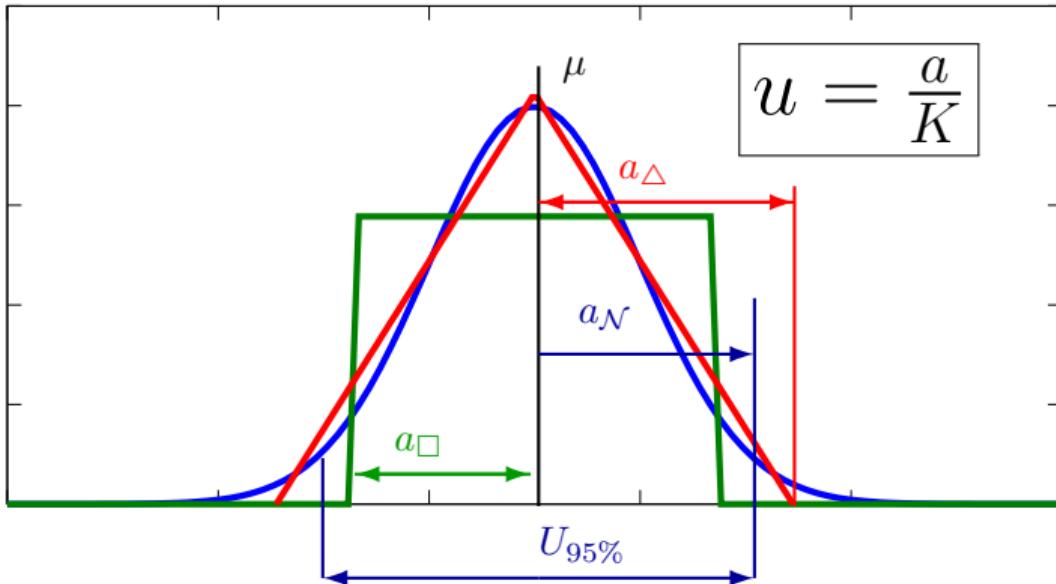
Normalverteilung:  $k = \sqrt{4}$ ,  $a$

Abmaße von  $U_{95\%}$

Gleichverteilung:  $k = \sqrt{3}$

Dreiecksverteilung:  $k = \sqrt{6}$

# Vertiefung: Konfidenzintervall, Bestimmung $u$



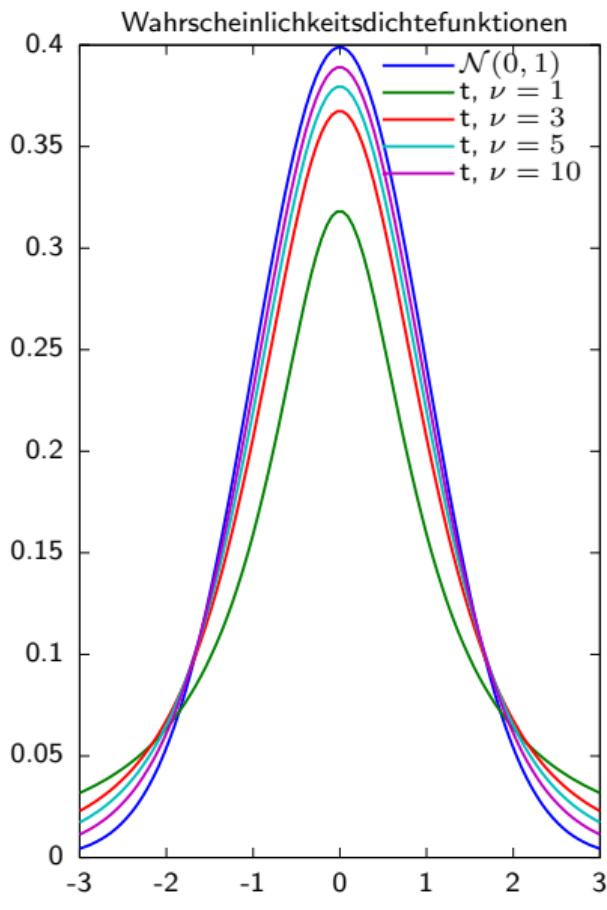
- Normalverteilung:  $k = \sqrt{4}$ ,  $a$  Abmaße von  $U_{95\%}$
- Gleichverteilung:  $k = \sqrt{3}$
- Dreiecksverteilung:  $k = \sqrt{6}$

# Student'sche t-Verteilung

- Normalverteilung (Hypothese):
  - Endliche Stichprobe
  - Wahrscheinlichkeit, "seltene" Werte zu realisieren?
- Stichprobenvarianz fällt zu klein aus
- "schlanke" Normalverteilung nicht konservativ
- Abhilfe: Student'sche t-Verteilung:

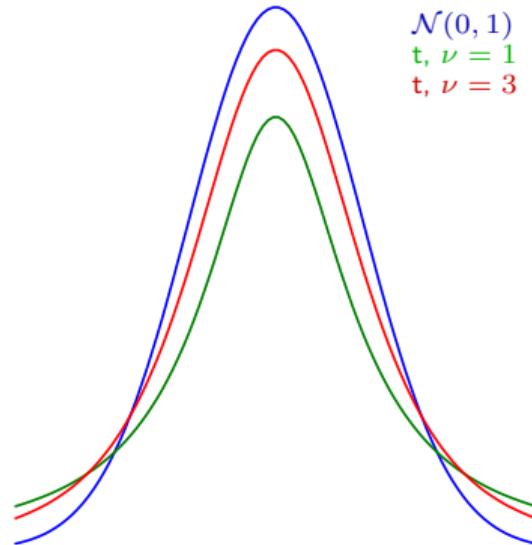
$$p_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{n+1}{2}}$$

- für  $n$  Freiheitsgrade



# Praktische Anwendung der Student'schen t-Verteilung

- t-Verteilung unterstellt "breitere" Wahrscheinlichkeitsdichte
- Schätzwert der Stichprobenvarianz wird korrigiert
- Bestimmung Korrekturfaktor aus Konfidenz und Stichprobengröße  $n$
- Freiheitsgrade  $\nu = n - m$  für  $m$  zu schätzende Parameter

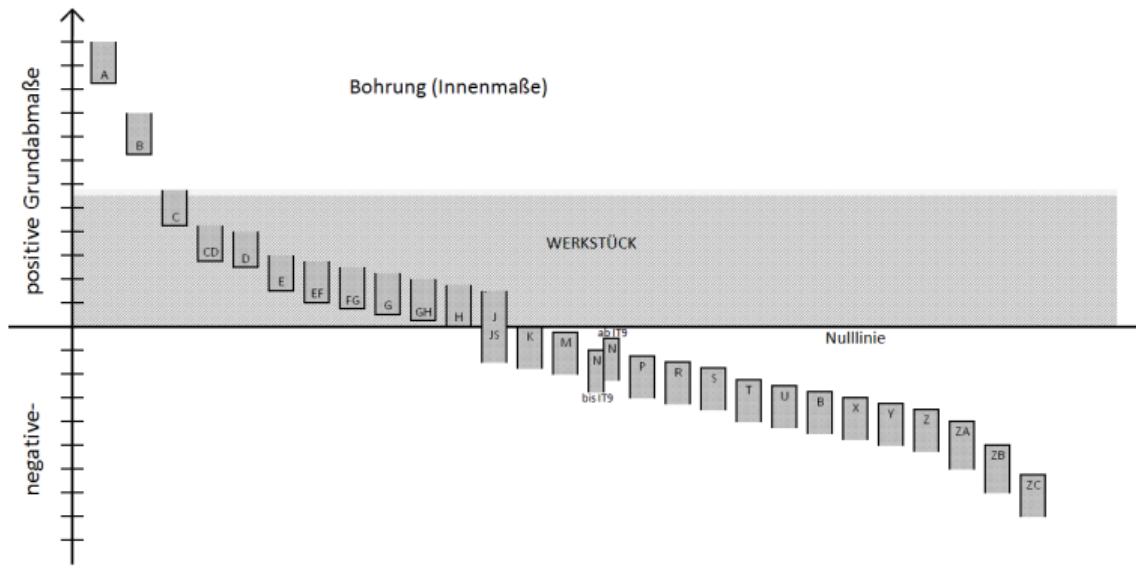


| $\nu$ | 1     | 2    | 3    | 4    | 5    | 10   | 20   | 50   | $\infty$ |
|-------|-------|------|------|------|------|------|------|------|----------|
| $k$   | 13,97 | 4,53 | 3,31 | 2,87 | 2,65 | 2,28 | 2,13 | 2,05 | 2,00     |

$k$ -Werte zur Bestimmung der erweiterten Messunsicherheit  $U_{95\%}$

# Definitionen

- Maß: Bestimmung einer Länge
- Nennmaß: Zeichnungangabe
- Istmaß: tatsächliches Maß
- Oberes/unteres Abmaß: zulässige Abweichung - Achtung Vorzeichen!
- Mindestmaß: Nennmaß - unteres Abmaß
- Höchstmaß: Nennmaß + oberes Abmaß



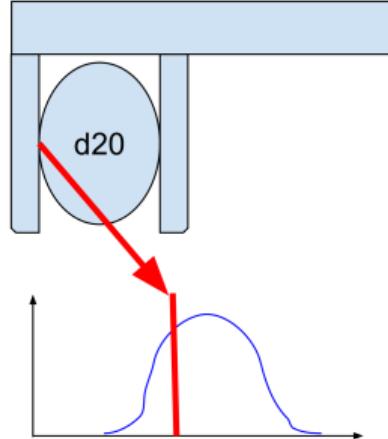
## Subsection 10

### Messunsicherheit

# Messergebnis und Messunsicherheit

Der Vergleich einer Messgröße mit einer Einheit gelingt nicht fehlerfrei.

- Messgeräteabweichungen
- Instabilität der Messgröße
- Umwelteinflüsse (z.B. Temperatur)
- Beobachtereinflüsse
- Es werden unterschieden:
  - Systematische Messabweichungen
    - bekannt o. unbekannt
  - Zufällige Messabweichungen



## Definition (Messergebnis)

Das Messergebnis ist der Schätzwert des wahren Wertes einer Messgröße.

# Verfahren zur Abschätzung der Messunsicherheit

## Verfahren A

- Abschätzung aus vorliegender Stichprobe
- $n$  Messwerte liegen vor und können statistisch ausgewertet werden
- Die Messwerte sind näherungsweise normalverteilt

$$u = \frac{s}{\sqrt{n}}$$

$s$ : Standardabweichung der Stichprobe

## Verfahren B

- Ermittlung der minimalen/maximalen Messabweichung  $a$  durch Modellbildung
- Ermittlung von Grenzen und Form der Verteilung der Messabweichungen

$$u = \frac{a}{k}$$

$k = \sqrt{4}$  für Normalverteilung

$k = \sqrt{3}$  für Gleichverteilung

$k = \sqrt{6}$  für Dreiecksverteilung

# Schätzfunktionen

Eine Schätzfunktion (Schätzer) dient zur Ermittlung eines Parameter-Schätzwertes aus empirischen Daten

- Grundlage: endlich viele Beobachtungen (Stichprobe)
  - Schätzer selbst fehlerbehaftet
  - Häufig Zufallsvariable
- Schluß auf Grundgesamtheit
- Schätzen einzelner Parameter der Verteilung
  - Mittelwert
  - Median
  - Standardabweichung

## Definition (Zufallsvariable)

Als Zufallsvariable bezeichnet man eine messbare Funktion von einem Wahrscheinlichkeitsraum in einen Messraum.

## Definition (Schätzfunktion)

Eine Schätzfunktion dient dazu, aufgrund von empirischen Daten einer Stichprobe einen Schätzwert zu ermitteln und dadurch Informationen über unbekannte Parameter einer Grundgesamtheit zu erhalten.

# Schätzfunktionen und Eigenschaften

Gängige Schätzfunktionen und wünschenswerte Eigenschaften

## ■ Mittelwert

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

## ■ Varianz

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\hat{\sigma}^2 = s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

## ■ Erwartungstreue:

- Erwartungswert der Schätzfunktion gleich wahrem Parameter
- Kein systematischer Fehler (Bias).

## ■ Konsistenz:

- Unsicherheit des Schätzers nimmt für  $n \rightarrow \infty$  ab

## ■ Effizienz:

- Minimale Varianz des Schätzers

## ■ BLUE: Best Linear Unbiased Estimator

# Bestimmung der Messunsicherheit durch Modellbildung

- Ermitteln der Einflussgrößen:
  - Ermittlung systematischer Abweichungen, ggf. kompensieren.
  - Einfluss des Normals
  - Wiederholpräzision (Messunsicherheit bei wiederholter Messung an einem Messobjekt)
  - Temperatureinfluss
- Modellbildung Messgröße  $y = f(x_1, x_2, \dots, x_n)$
- Bestimmung kombinierte Standardunsicherheit  $u_c$

$$u_c = \sqrt{\left(\frac{\partial f}{\partial x_1} u_{x_1}\right)^2 + \left(\frac{\partial f}{\partial x_2} u_{x_2}\right)^2 + \dots + \left(\frac{\partial f}{\partial x_n} u_{x_n}\right)^2}$$

- Vereinfachung für Linearität von  $f$  und gleichem Gewicht der  $x_i$

$$u_c = \sqrt{u_{x_1}^2 + u_{x_2}^2 + \dots + u_{x_n}^2}$$

# Monte-Carlo-Simulation: Motivation

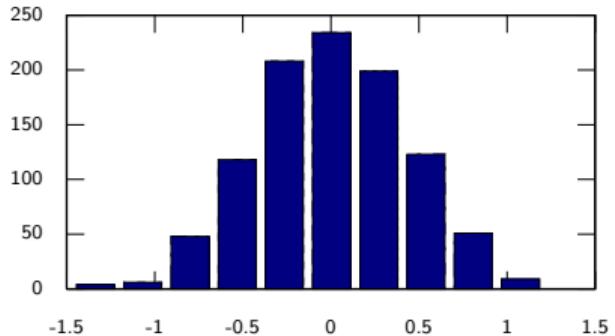
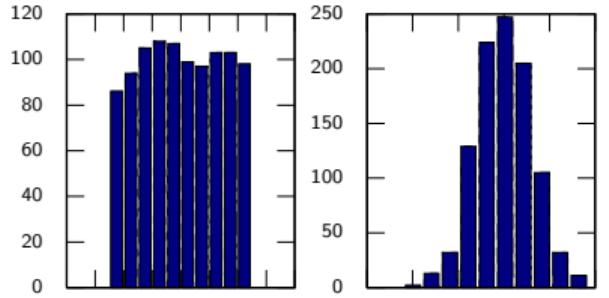
Monte-Carlo-Simulationen (MC-Simulationen) können zur Ermittlung kombinierter Wahrscheinlichkeiten eingesetzt werden.

- Statistische Verfahren:

- Hohe Stichprobenzahlen für Validität nötig
- Zeitaufwändig
- Teuer

- Fehlerfortpflanzung:

- Analytisch teils komplex
- Konservativ bei nicht normalverteilten Zufallsvariablen:
  - Annäherung Gleichverteilung durch Normalverteilung



# Monte-Carlo-Simulation: Grundlagen

MC-Simulationen basieren auf der häufig wiederholten Simulation von Zufallsexperimenten.

## ■ Grundlage:

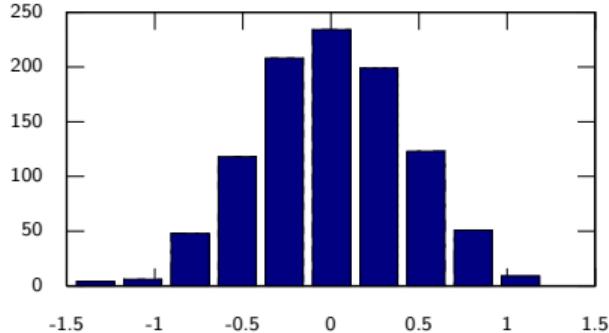
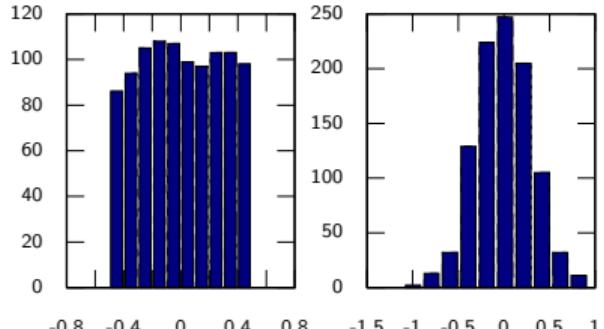
- Gesetz der großen Zahlen
- Pseudozufallsverteilungen der Parameter erzeugen
- Resultierendes Ergebnis bilden
- Analyse der Lage- und Streuparameter

## ■ Vorteile:

- Einfache Modellierung
- Einfache Durchführung

## ■ Nachteil:

- Korrektheit der Lösungen nicht immer einfach zu beurteilen



# Vergleich der Verfahren des GUM und MC-Simulationen

## MC-Simulation

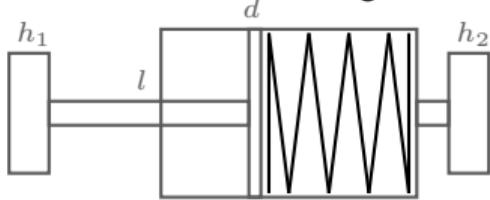
- Eingangsgrößen  $x_i$  werden explizit W-Dichtefunktionen zugeordnet
- Keine partiellen Ableitungen benötigt, können bereitgestellt werden
- Keine Einschränkung in Form der Verteilung von  $y$
- Allgemeine, kürzest mögliche, Konfidenzintervalle

## GUM Verfahren A/B

- Eingangsgrößen  $x_i$  werden Schätzwert und Standardunsicherheit zugeordnet
- Empfindlichkeitskoeffizienten und Taylor-Approximation benötigt
- Beschränkung auf Gauß- oder t-Verteilung
- Um den Schätzwert symmetrische Konfidenzintervalle

# Monte-Carlo-Simulation: Beispiel

Parameter eines Produktes variieren häufig in Form von Gleich- oder Dreiecksverteilungen.

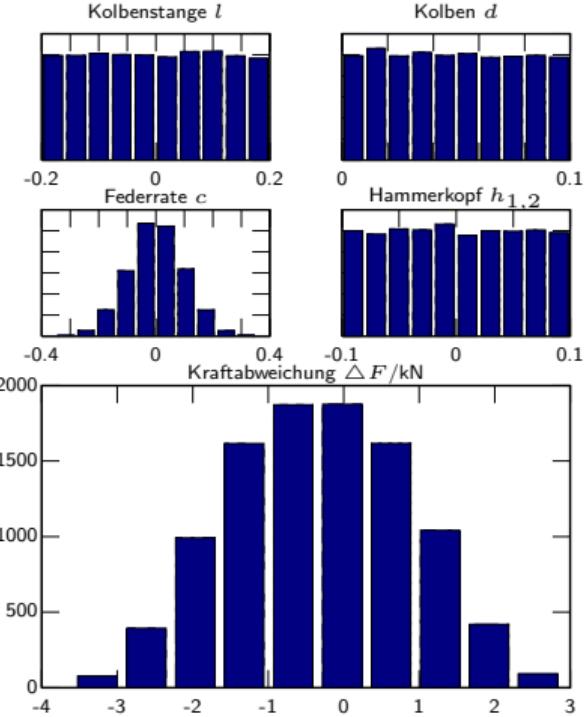


- Kraft eines Federspeicherzylinders

$$\Delta F =$$

$$c(\Delta h_1 + \Delta l + \Delta d + \Delta h_2)$$

- MC-Simulation mit  $N = 10000$



# MC-Simulation zur Bestimmung der Messunsicherheit

Nicht normatives Beiblatt 1 zu DIN V ENV 13005

- MC-Simulation als Ergänzung zu Verfahren A und B des “GUM”, z.B. falls
  - Linearisierung des Modells zu unangemessener Darstellung führt, z.B. durch Sensibilität
  - Partielle Ableitungen schwierig oder unmöglich zu finden sind
  - die Wahrscheinlichkeitsdichte merklich von Gauß- oder  $t$ -Verteilung abweicht , z.B. durch Asymmetrie
  - Wiederholte Versuche zur Bestimmung der Messunsicherheit nicht möglich sind
  - Modell des Messprozesses nicht in explizite Form gebracht werden kann
  - Unsicherheitsbeiträge nicht näherungsweise von der gleichen Größenordnung sind
- MC-Simulation ist im Einklang mit GUM
  - Ermittlung von Erwartungswert, Standardunsicherheit und Überdeckungsintervall der Messgröße $y$

# Ablauf Erstellung einer MC-Simulation

## 1) Modellierung des Systems:

- a) Ausgangsgröße  $y$
- b) Eingangsgrößen  $\mathbf{x} = (x_1, \dots, x_i)^T$
- c) Modell als Beziehung zwischen  $y$  und  $\mathbf{x}$ , nicht zwingend explizit
- d) Zuordnung von Wahrscheinlichkeitsdichtefunktionen zu den  $x_i$ 
  - $x_i$  unabhängig: individuelle W-Verteilungen, z.B. Gauß-, Gleichverteilung etc.
  - $x_i, x_j$  abhängig: Gemeinsame W-Verteilung

## 2) Fortpflanzung: Simulation des Systems

- a) Bestimmung des Wertes  $y$  aus den  $x_i$
- b) Werte der  $x_i$  als Pseudozufallszahlen

## 3) Zusammenfassung:

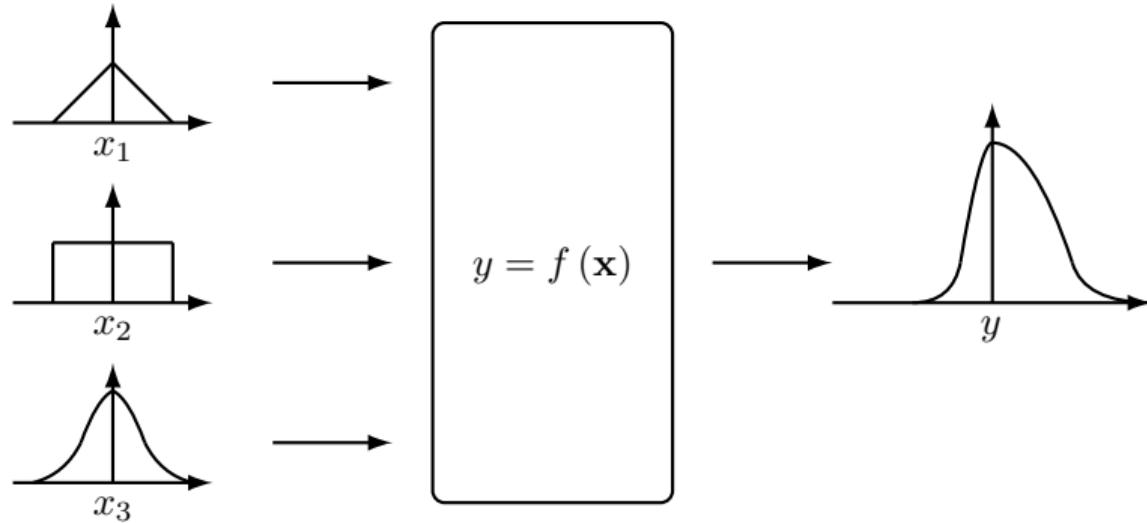
- a) Bestimmung des Erwartungswerts  $E(y)$ <sup>1</sup>
- b) Bestimmung der Standardabweichung  $\sigma_y$ <sup>2</sup>
- c) Bestimmung des Konfidenzintervalls (Überdeckungsintervalls), das  $y$  mit einer festgelegten Wahrscheinlichkeit enthält

<sup>1</sup>Nicht alle W-Verteilungen weisen einen Erwartungswert auf.

<sup>2</sup>Nicht alle W-Verteilungen weisen eine Standardabweichung auf.

# Bestimmung der Fortpflanzung

Neben analytischen und statistischen Methoden sind MC-Simulationen zulässig und effizient. Verschiedene W-Verteilungen und nichtlineare Systemfunktionen machen sie notwendig.



# W-Dichtefunktionen für die Eingangsgrößen

- Allgemein:  $\mathbf{x}$  wird gemeinsame PDF zugeordnet
- Unabhängigen  $x_i$  werden einzelne PDF zugewiesen
- Allgemein: Bestimmung PDF gemäß Bayes-Theorem oder Prinzip der maximalen Entropie
- Praktisch häufig auftretende Fälle: [1, S. 34f.]
  - Rechteckverteilung: Toleranzbänder
  - arcsin-Verteilung (U-förmig): Sinusförmige periodische Schwingungen
  - Gauß-Verteilung: Schwankung durch viele Einflussfaktoren (Zentraler GWS)
  - $t$ -Verteilung: Modellierung aus endlich vielen Anzeigewerten
  - Exponential-Verteilung: Linkssteile Verteilung für positive Größen

# Voraussetzungen für Anwendbarkeit und Gültigkeit der MC-Simulation

- Es ist  $f$  stetig bezüglich der  $x_i$  in der Nachbarschaft der besten Schätzwerte  $\hat{x}_i$
- Die Verteilungsfunktion für  $y$  ist streng wachsend
- Die Wahrscheinlichkeitsdichtefunktion (PDF) für  $y$  ist
  - stetig über dem Intervall, für das die PDF streng positiv ist
  - unimodal
  - streng wachsend (oder 0) links vom Modalwert und streng fallend (oder 0) rechts vom Modalwert
- Es existieren  $E(y)$  und  $\text{Var}(y)$ 
  - Gewährleistet stochastische Konvergenz mit steigendem  $M$
- Es ist die Zahl der Monte-Carlo-Versuche  $M$  ausreichend groß gewählt
  - Gewährleistet Zuverlässigkeit der stochastischen Information

# Praktische Umsetzung einer MC-Simulation

Die hohe Anzahl Versuche  $M$  macht effiziente Software nötig.

## 1 $M$ wählen:

- Da MC-Simulationen Zufallsexperimente sind, kann kein festes  $M$  Korrektheit des Algorithmus garantieren
- Für Signifikanzniveau  $\alpha$ :  $M \gg \frac{1}{1-\alpha}$ , z.B.  $M \geq 10^4 \frac{1}{1-\alpha}$
- Alternativ: Adaptive Verfahren

## 2 Erzeugung von Zufallszahlen für $\mathbf{x}_r$ aus den gewählten PDF

- Pseudozufallszahlengenerator muss über gewisse Eigenschaften verfügen, z.B. Sequenzlänge

## 3 Bestimmung der Werte $y_r = f(\mathbf{x}_r)$ für $r = 1, \dots, M$

## 4 Diskrete Darstellung der Verteilungsfunktion für die Ausgangsgröße

## 5 Schätzung der Ausgangsröße und der beigeordneten Standardunsicherheit

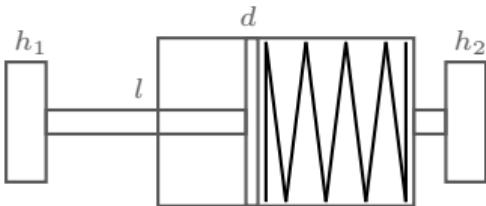
- $\bar{y} = \frac{1}{M} \sum_{i=1}^M y_i$
- $s^2(\bar{y}) = \frac{1}{M-1} \sum_{i=1}^M (y_i - \bar{y})^2$

# Section 2

## Manufacturing data

# Scenario

Please assume that you are working as a freelance quality consultant to a manufacturing company in the railway sector. Your client manufactures spring parking brake cylinders as depicted below.



A spring parking brake is used to maintain a rail vehicle in a stationary position, for movements, the brake is released by applying compressed air to the cylinder. For this purpose, one of the determining parameters of a spring parking brake the force is the force applied to the so called hammerheads of the brake cylinder. This force has to maintain the vehicle stationary. For the product under consideration, the required minimum force is 18 kN.

# Tasks

Data import and graphical representation

You receive a call from your client because in today's production shift, more than 10% of the production had to be scrapped due to insufficient force. The client sends you a .csv file containing the quality records of the shift.

The file contains the following data:

| Time, Date | F/N | $\Delta d/\text{mm}$ | $\Delta h_1/\text{mm}$ | $\Delta h_2/\text{mm}$ | $\Delta l/\text{mm}$ | $x/\text{mm}$ |
|------------|-----|----------------------|------------------------|------------------------|----------------------|---------------|
|------------|-----|----------------------|------------------------|------------------------|----------------------|---------------|

$$x = h_1 + h_2 + d + l$$

In order to start your analysis, you import it and inspect it qualitatively for randomness and systematic influences. Your customer is rather fluent in Python, so you decide to exchange Jupyter Notebooks highlighting your findings.

# Tasks

Optional: Monte-Carlo simulation

The client considers a systematic behaviour of the spring stiffness as a possible root cause of the force loss. Each cylinder contains 24 springs of three types, each being supplied with a  $\pm 10\%$  tolerance on the nominal value:

- 8 small springs:  $c_{1,\text{nom}} = 100 \text{ N/mm}$
- 8 medium springs:  $c_{2,\text{nom}} = 150 \text{ N/mm}$
- 8 large springs:  $c_{3,\text{nom}} = 200 \text{ N/mm}$

Perform a Monte Carlo Simulation of the combined spring stiffness  $c$  and its variation. Assume two cases:

- $C_1$  The supplier ships springs equally distributed within the tolerance range.
- $C_2$  The supplier ships large springs within tolerance, however the distribution is skewed towards the lower end of the spectrum, yielding forces with  $c_3 \in [180, 190] \text{ N/mm}$ . All other springs are as described above.

Hint: remember to draw the individual springs of a set from an independent and identically distributed (i.i.d.) random variable.

# Tasks

Using support vector machine to predict an unobserved product property

esting of the cylinders for their effective force is a costly task. The customer would prefer to sort the components such that only capable cylinders are assembled. They propose to use the hammerhead dimensions  $h_1$ ,  $h_2$  as well as the length  $l$  to predict the force based on learned data. For this purpose:

- Import the existing dataset
- Scale the data using a standard scaler
- Train a support vector machine based on this data
- Use the new data set (unknown to the SVM) “SpringPBDataValidation.csv” to test the performance
  - You may also generate test data using a Monte-Carlo simulation
- Plot a confusion matrix of the results

How do you like the performance of your classifier? Can you manufacture without end of line testing? Is your data basis sufficient to exclude any non-conforming cylinders from being shipped to the customer?

# Literatur I

- [1] DIN V ENV 13005: Beiblatt 1 - Fortpflanzung von Verteilungen unter Verwendung einer Monte-Carlo-Methode.  
DIN Deutsches Institut für Normung e.V., 2012.
- [2] T. Pfeifer and R. Schmitt.  
*Fertigungsmesstechnik*.  
Oldenbourg Verlag, München, 2010.