Hochschule Konstanz
Fakultät Informatik

Prof. Dr. Oliver Dürr

Data Analysis:: Course Logistics and Introduction

Logistics

- In Moodle 8:45 – 11:15 / 15 Minutes breaks
- O-008

Proposed

- 8:45 – 9:30
- 9:35 – 10:20
- 10:30 – 11:15

One short break 5 Minutes
One longer break 10 Minutes

Modul ITM02	**Data Analysis**			
Modul-Koordination	**Start**	**Modul-Kürzel/-Nr.**	**ECTS-Punkte**	**Arbeitsaufwand**
Prof. Dr. O. Dürr	SS	DTAN/ITM02	5	150 h
	Dauer	**SWS**	**Kontaktzeit**	**Selbststudium**
	1 Semester	3	45 h	105 h

https://www.htwg-konstanz.de/master/business-information-technology/dokumente/alles-auf-einen-blick

Today

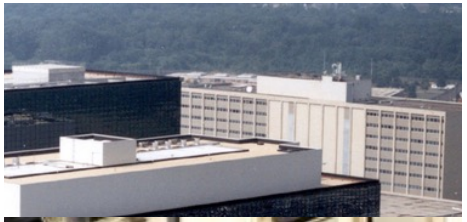
- Logistics
- What is Data Analysis
- Recap Statistics

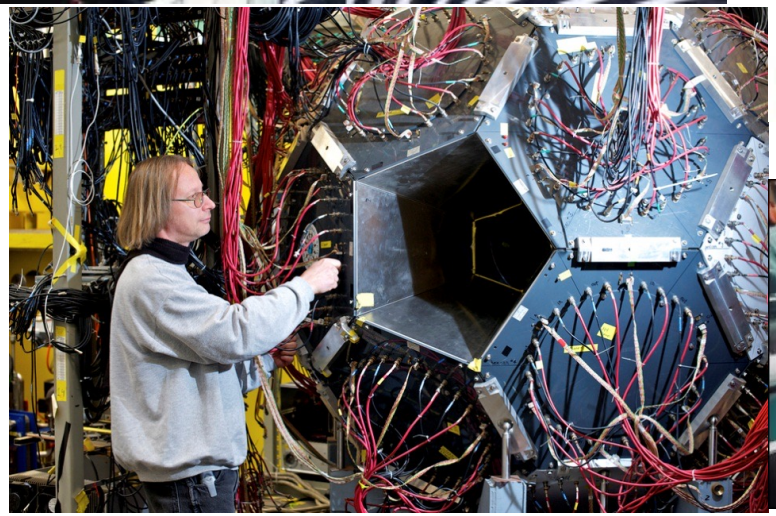
Prerequisites: Please refresh your math and stats skills!

- Specifically I assume the following Prerequisites
  - Probability rules (basic)
    - You need to know how to calculate basic probabilities (e.g., if I have 5 red and 3 blue balls in a bucket, what is the probability of picking a red ball?)
    - Are able to apply the conditional rule of probability (i.e., P(A, B) = P(A | B) P(B)).
  - Statistics notions (intermediate)
    - You need to know the following statistics: Mean (average), (empirical) Variance, (empirical) Standard Deviation.
    - Given a list of data, you must know how to calculate the above statistics on it.
    - You need to know the concept of estimation
      - The average as estimator for the expected value should not sound too scary
    - You should know about confidence intervals and (ideally) tests
  - Probability distributions (basic)
    - You need to understand probability distributions on a conceptional level. E.g. you should know that probabilities sum up (or integrate) to 1. You need to be able to locate the expected value and estimate the probability for certain outcomes given the graph of a distribution.
    - A detailed knowledge of specific distributions is not needed.
  - Calculus (basic)
    - You need know how to calculate derivatives, including applying the sum, product, as well as the chain rule.
    - You must be well versed in the following mathematical functions: logarithm, exponential, power functions.
    - You must be familiar with the summation and product sign
  - Matrix Algebra (basic)
    - You must know how to multiply matrices and vectors.

What is Data Analysis (@ HTWG)

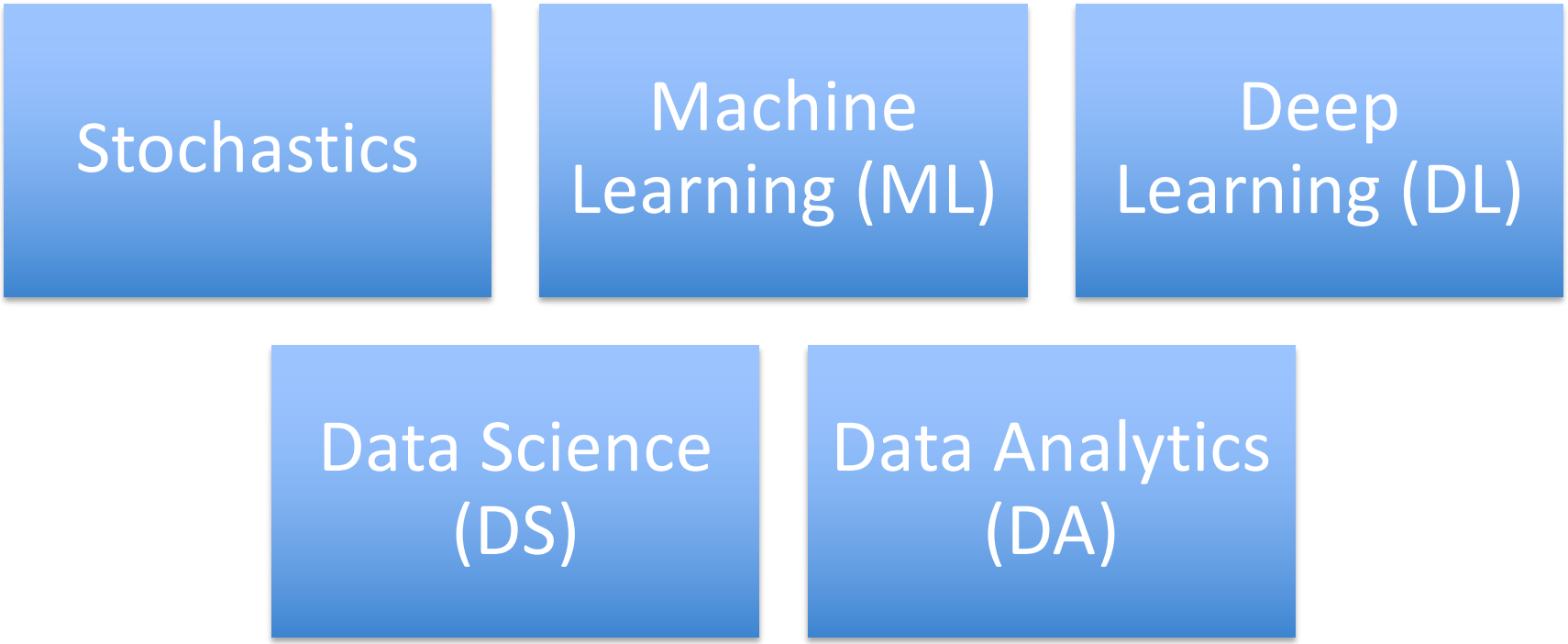
Examples of Data





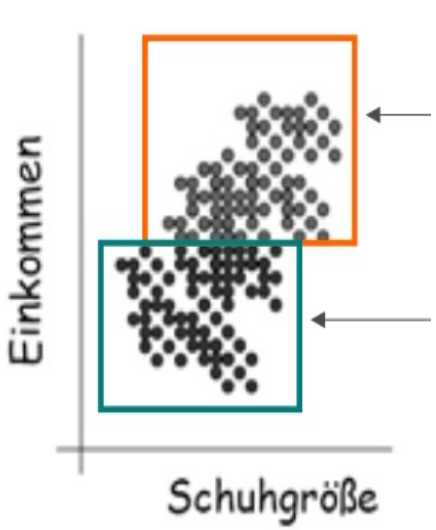
The Analysis of Data is known under many names: Statistics, Machine Learning, Statistical Learning, Data Science, Data Analysis …

Analysis of Data at HTWG



How does this all relate? All these field addresses different aspects.

Aspect 1::Motivation to Analyze data



- **Predict**
  - Given an unknown example of x what will be y?
  - Subtle difference
    - The example is like the training data (observation)
    - I forcefully change x (intervention)

- **Explain / interpret**
  - What is the connection between x and y?
  - Subtle difference
    - In the observed data
    - When you forcefully change x (intervention)

Different methods are used to predict and explain and also if this is done forcefully

Aspect 2:: Types of Data

 Text files and documents	 Server, website and application logs	 Sensor data	 Images
 Video files	 Audio files	 Emails	 Social media data

Aspect 2:: Types of Data (Structured / Tabular Data)

- Data objects and their attributes (like Excel Sheet)
- A variable, often also called a feature, is a property of an object
  - Example: eye color, sepal width, etc.
  - For classification there is often a distinguished attribute (class attribute)
- A collection of attributes describes an object

p Features

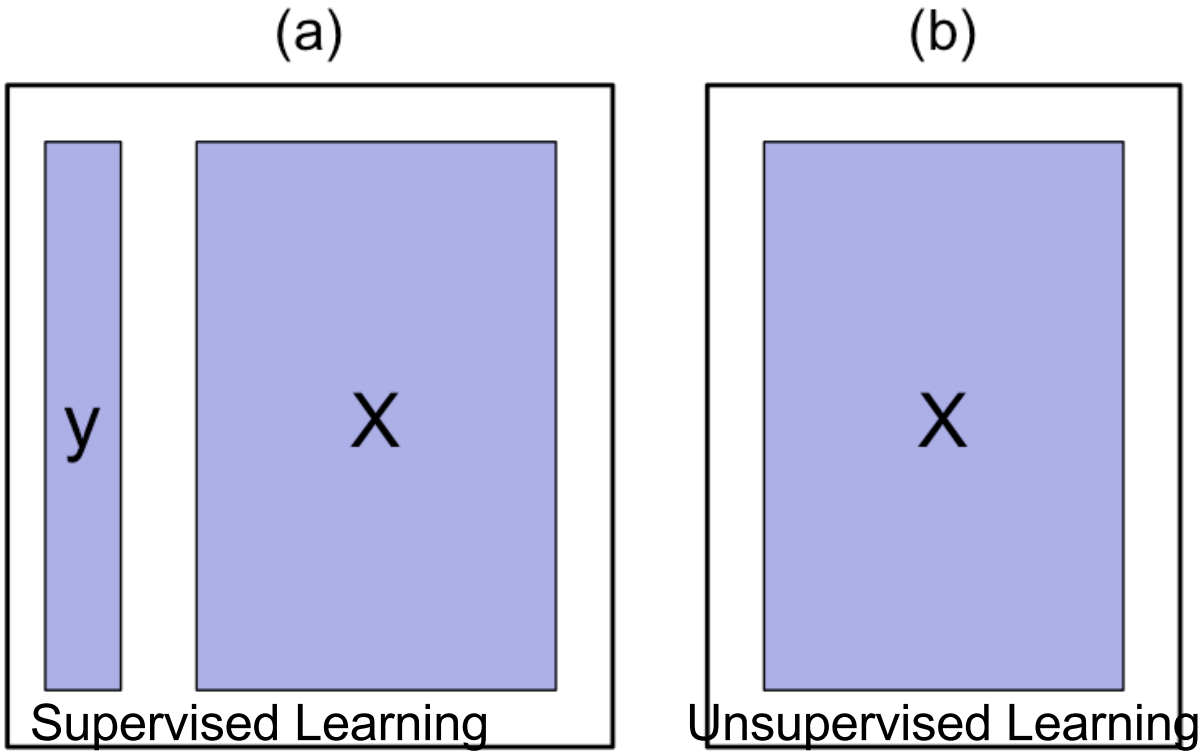
n Objects

Blume	Type	Sepal.Length	Sepal. Width	Petal.Length	Petal. Width
1	setosa	5.1	3.5	1.4	0.2
2	setosa	4.9	3	1.4	0.2
3	virinica	3.3	3.2	1.6	0.5
4	setosa	5.1	3.5	1.4	0.2
...	...	...	...	...	...
150	virinica	4.9	3	1.4	0.2

Aspect 3:: Supervised vs. Unsupervised Learning

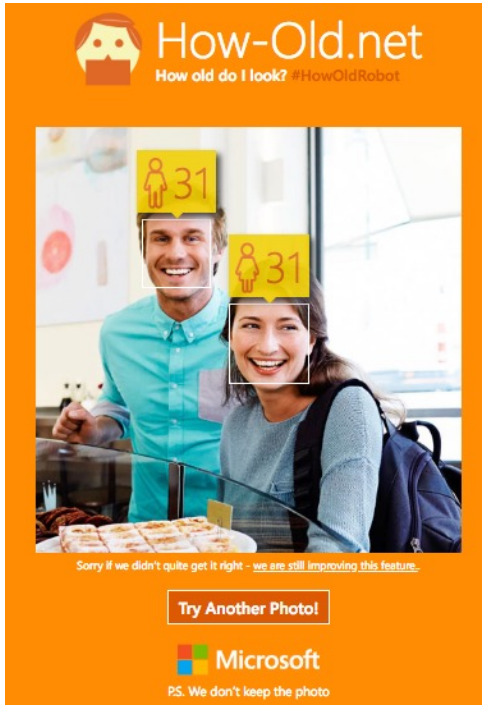
- Supervised Learning: both X and Y are known
  - Based on X make Prediction on Y

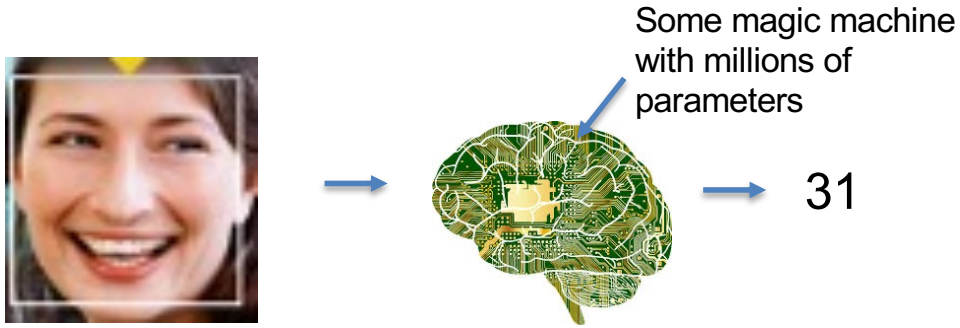
- Unsupervised Learning: only X



Slide credit: Abbas Al Sharif IOM 530: Intro. to Statistical Learning

Aspect 4:: Complexity of the model (Complex)





https://www.how-old.net/

Here (HTWG): Part of Deep Learning

Aspect 4:: Complexity of the model (Simple)

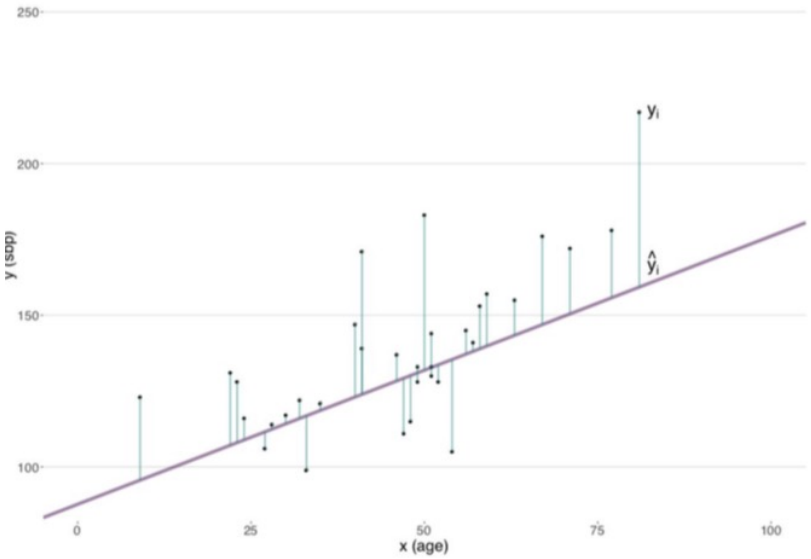
Model:

$$\hat{y} = \mathrm{a} \cdot x + b$$

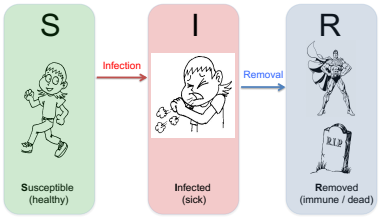
Simple model to

predict SBP from age

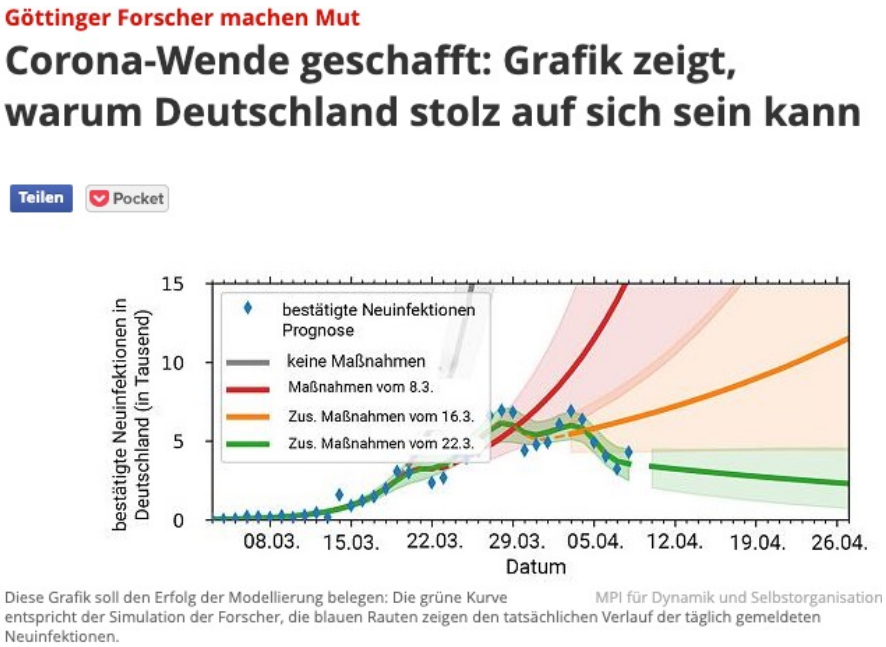
describe effect of age on SBP



Model



The output is modeled with some parameters, having unknown values and fitted to the data



Here (HTWG): The heart of data analysis

Analysis of Data at HTWG (simplified)

- Stochastics
  - In-depth treatment rather small data sets, **interpretation**
- Machine Learning
  - Tabular (structured) data, **prediction**
- Deep Learning
  - High dimensional complex unstructured data (often images, text,…), **prediction. Complex models**
- Data Science
  - Working with data on a more practical level (w/o too much theoretical understanding)
- Data Analytics
  - Learning with **simple models**, tabular data, prediction and interpretation

Data Science vs Data Analytics at HTWG

- Data-Science
  - R the tool
    - Basic R
    - Visualization (using ggplot)
    - Data Wrangling dyplr
  - Unsupervised Learning
    - Visualization of HD data
    - Clustering
  - Supervised Learning
    - Classification
      - Classifiers Lin Reg, SVM,…
      - Bias Variance
      - X-Val

Working with Data

- Data-Analytics
  - Basic Concepts of statistics
    - Recap of basic statistical terms
    - Linear Regression
  - Introduction to causal reasoning
    - Directed Acyclical Graphs (DAGs) and the Backdoor criterium
    - Counterfactuals
  - Traditional Data Analysis using the Likelihood Principle
    - Distributional Regression
  - Bayesian Data Analysis
    - The Bayesian Concept of Probability
    - Bayesian Modeling (posterior predictive distribution)
    - Markov Chain Monte Carlo
    - Modeling using Stan
    - Evaluation of the predictive performance
    - Hierarchical Bayesian Models

Modelling the Data

Literature

Statistical Rethinking

A Bayesian Course with Examples in R and Stan (& PyMC3 & brms & Julia too, see links below)

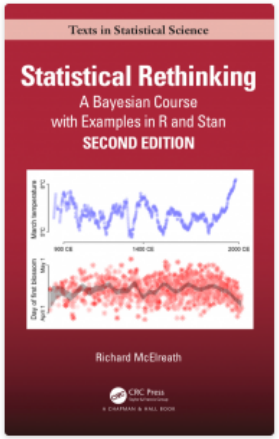
Second Edition

The second edition is now out in print. Publisher information on the CRC Press page. For more detail about what is new, look here.

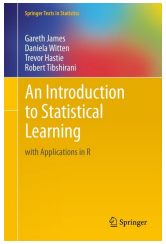
Materials

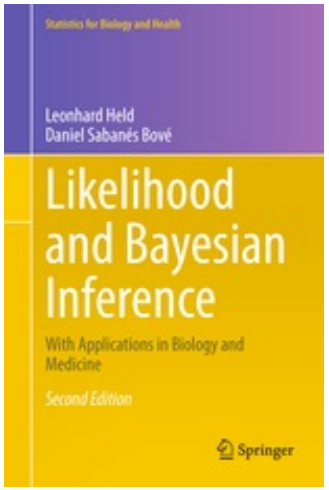
2nd Edition

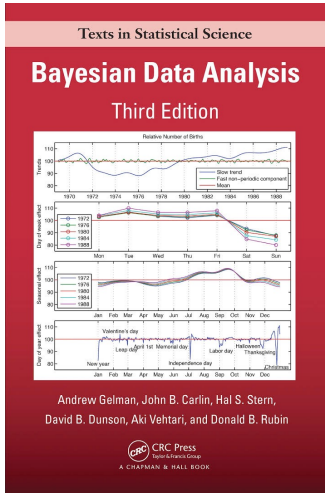
- Book: CRC Press
- Book sample: Chapters 1 and 2 (2MB PDF)
- Lectures and slides:
  * Winter 2019 materials
- Code and examples:
  * R package: rethinking (github repository)
  * R code examples from the book: code.txt
  * Book examples in Stan+tidyverse
  * brms + tidyverse conversion here
  * PyMC3 code examples: PyMC repository
  * NumPyro!
  * TensorFlow Probability notebooks
  * Julia & Turing examples (both 1st and 2nd edition)
  * R-INLA examples



Other Books



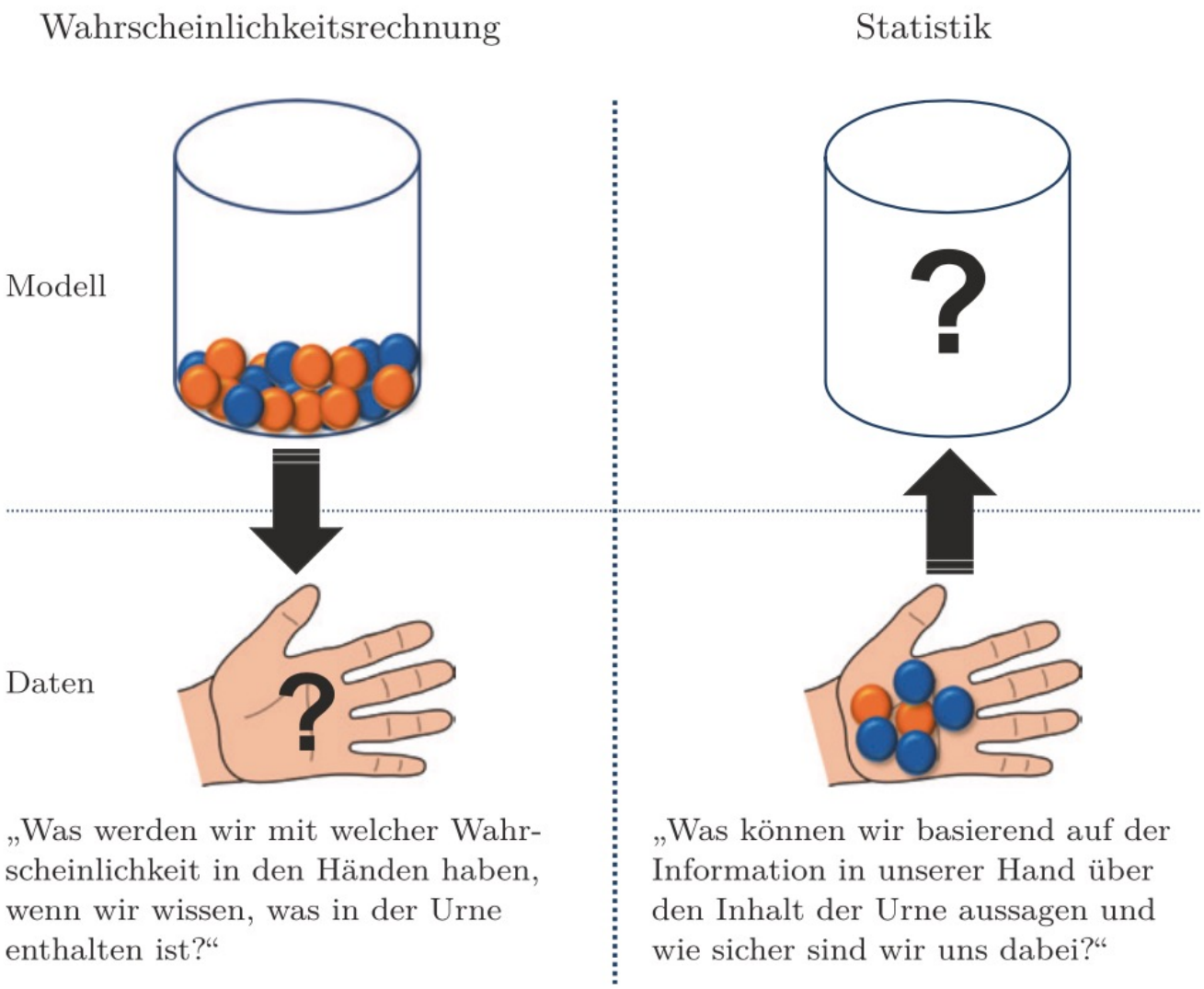




General Disclaimer: The material of this course is based on many slides developed with Prof. Dr. Beate Sick (UZH/ZHAW) for various courses at the ZHAW

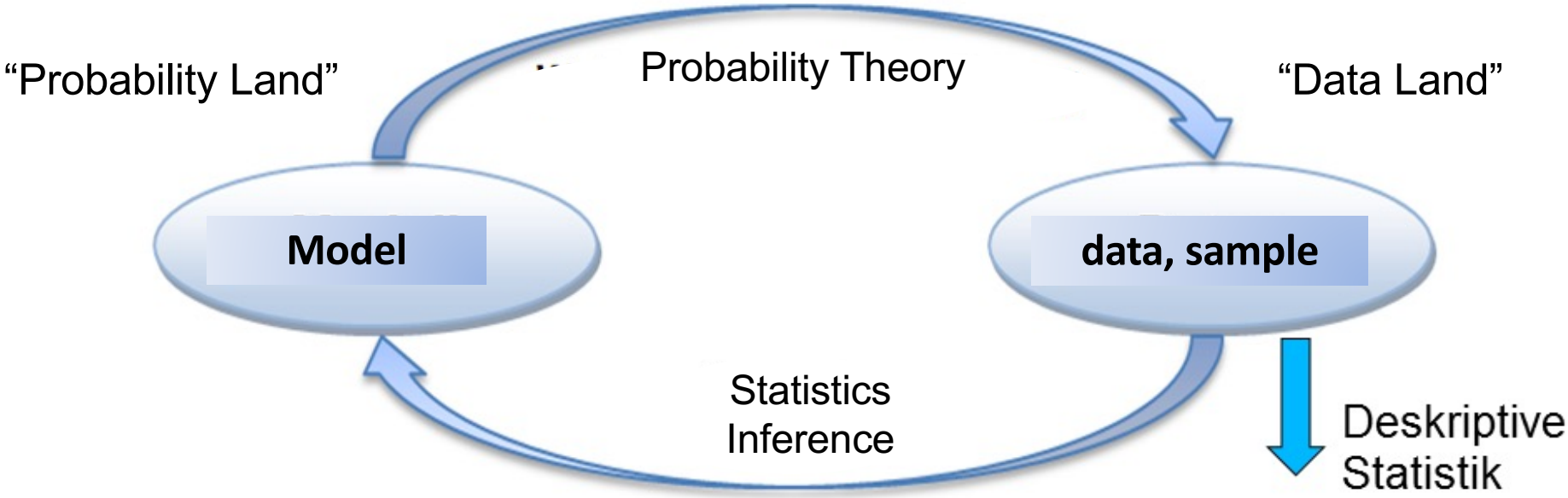
Short Recap of Statistics

Statistik vs W'keitsrechnung



Aus Lukas Meier W‘keit und Statistik

Statistics connects data with models



describe data
Visualize & Summarize

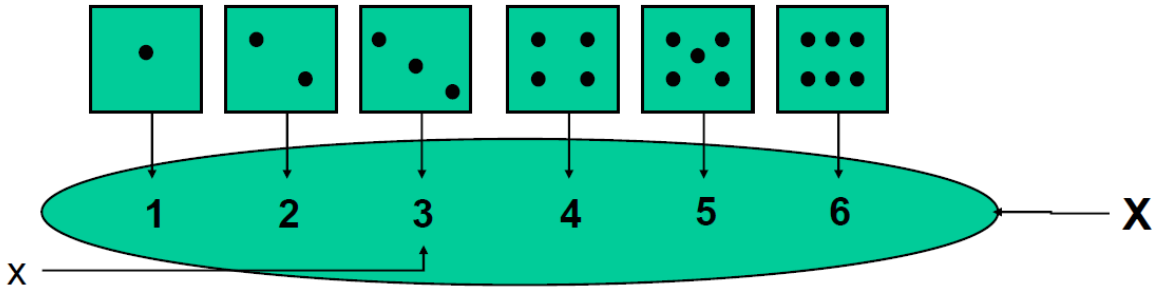
Statistical inference

- Parameter estimation
  - Regression
  - Classification
- Confidence intervals
- Tests

Data Land

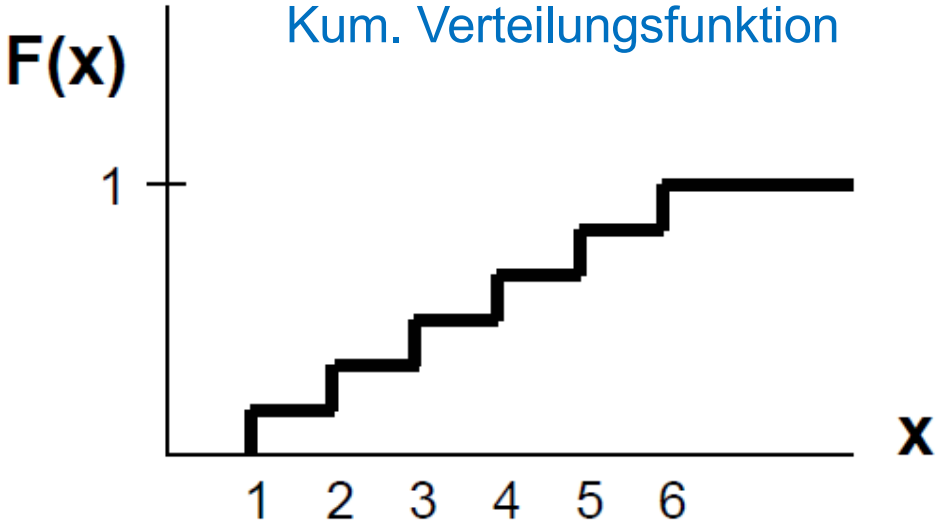
- Tables / Histograms **mean and empirical variances**
- Data (Dice rolls): 1,1,2,1,3,5,3,4,6,…
- Data: 0.1, 0.2, 0.23, 0.32, 0.123123, 0.3

Probability Land



Diskrete
W'keitsverteilung

Kum. Verteilungsfunktion

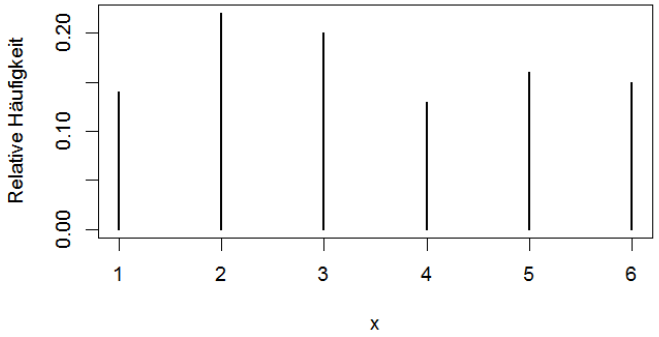


P(x)=P(X=x) W'keit, dass ZV X den Wert x annimmt.
F(x) = W'keit dass ein Wert <= x angenommen wird

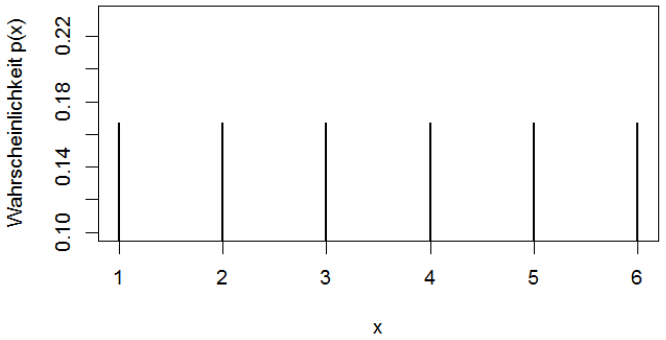
Interpretation of the propability density function (pdf)

Discrete

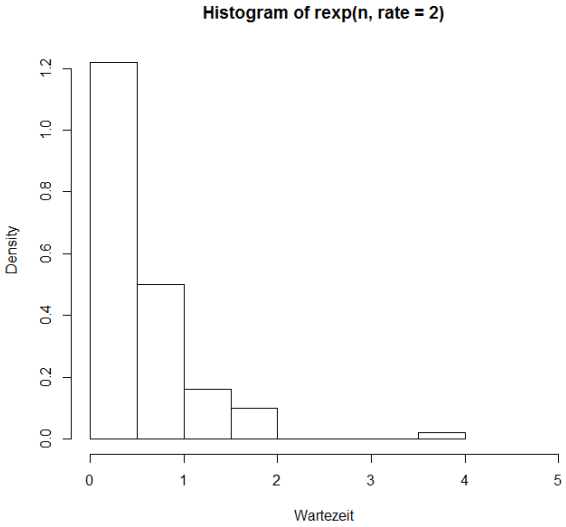
Hist. of relative frequency



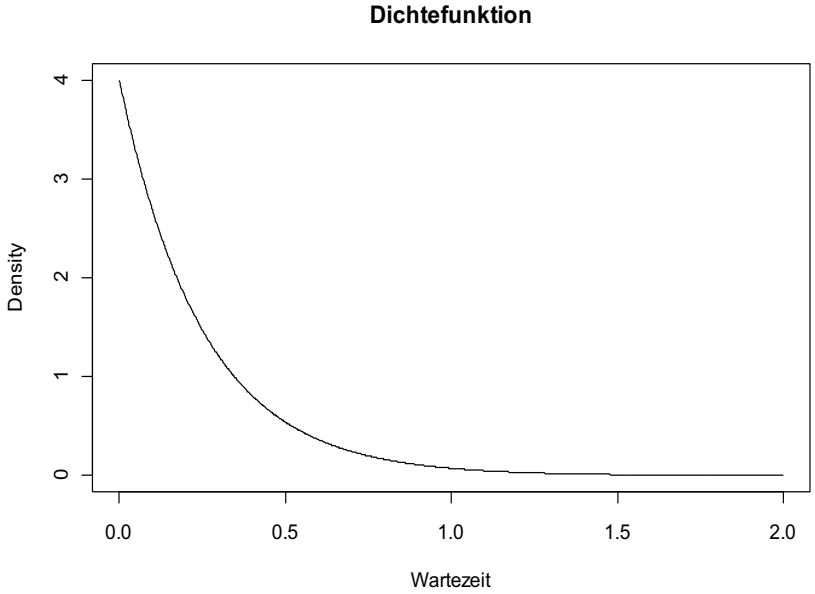
Probability



Continious

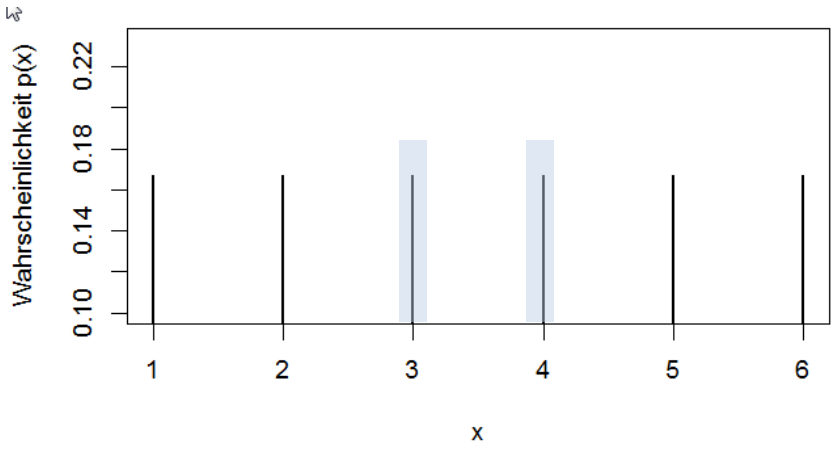


Probability?



Density vs. Probability

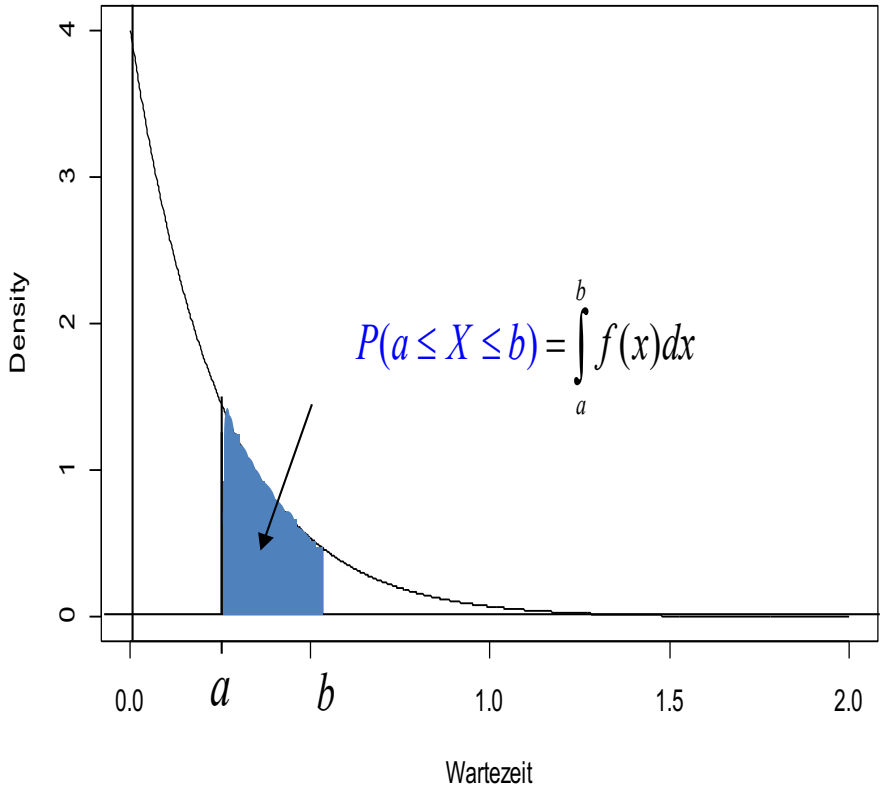
Diskret



$$P(a \le X \le b) = \sum p(X = x_i)$$

Continious

Dichtefunktion



The probability of obtaining a result between a and b is the sum of the probabilities.

The probability of obtaining a result between a and b is the integral of the density function over a to b.

Expectation

- What is expected value of the probability density.
- Integral Formula
- Draw a pdf and explain center-of-mass

Homework (Task 1, probability densities)

Excercise 1 (Probability Densities / Expectations)

Assume the following pdf:

$$f(x) = \begin{cases} c \cdot x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a) What is the value of the constant c?

b) What is probability that x is in between 0 and 0.5

c) What is the expectation $E(X)$

d) What is the expectation of $E(X^2)$

Generally: Try to do a little calculation as possible, at least for a) and b) you do not need to do a calculation but can argue graphically.

Start der übung 10:00

Probability: model → data (sampling)

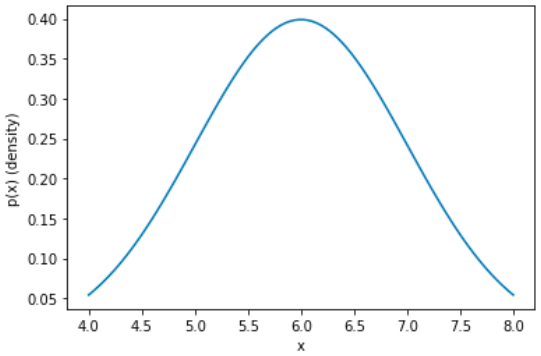
- Probability Theory
  - A coin is thrown $\boldsymbol{n = 20}$ times, the probability for one head is $\boldsymbol{\pi = 0.2}$ how probable is it to have k heads? T**he parameters are known**

$$k \sim Binom(\mathrm{n} = 20, \pi = 0.2)$$

  - You drive 100 km, how many liters *x* of gas do you need? We assume that the data **is sampled/drawn from or generated by a** normal distribution with $\mu = 6, \sigma^2 = 1$ (**the parameters are known**)

$$x \sim N(\mu = 6, \sigma^2 = 1)$$

  - Properties of pmf, pdf
    - sum / integrate to one
    - Expectation is center of mass



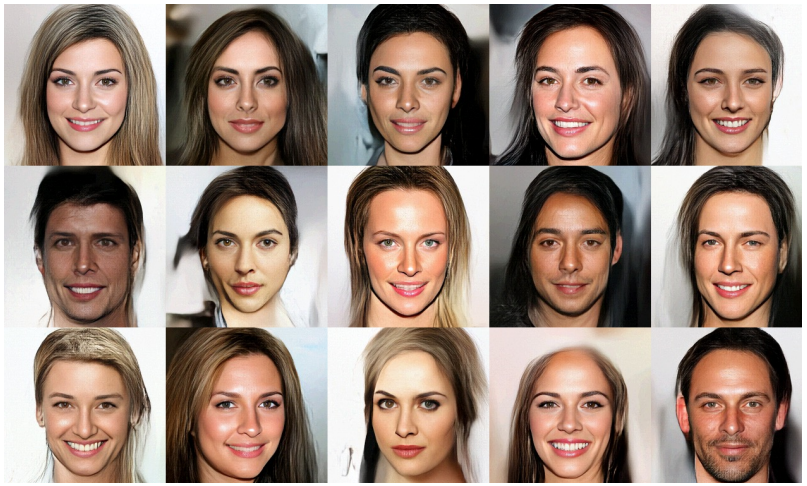
We here drop the distinction between X (RV) and x (sample)

Probability: model → data (sampling)

- Data varies
  - ipython example of car
    - np.random.normal(loc=6,scale=1,size=10)
  - Model with some parameters (knobs are μ and σ)

$x \sim$ 

  - x can also be high dimensional data (256x256x3)



15 different x sampled from a deep learning model called glow (more than 100 mio. Parameters, 800 MB)

These are not real people!

For glow see: arXiv:1807.03039

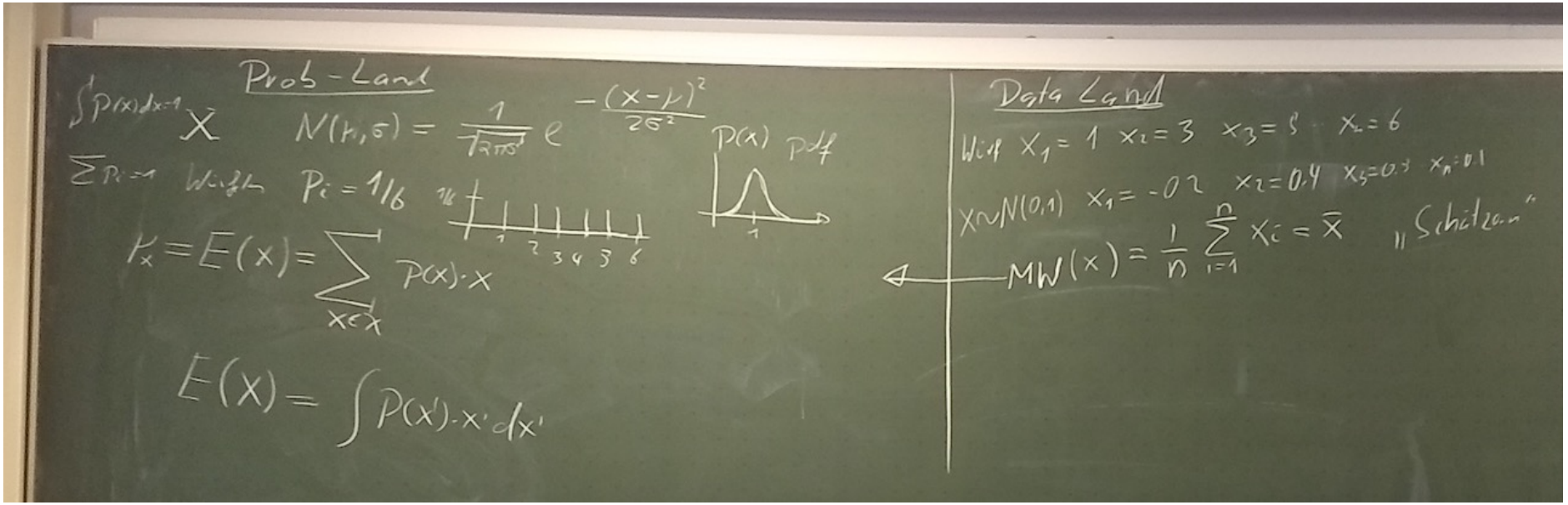
Inference: data → model (training)

- Parameter Estimation:
  - A coin is thrown n=20 times, you observe 12 heads, what is your best guess for pi?
  - A car is driven n=10 times 100 km you observe the following 10 values:
    - [6.27045683, 5.94976189, 5.76105195, 5.09243634, 5.42322867, 6.75539123, 6.50091719, 5.02244476, 6.09933231, 6.75138712]
  - **You assume** it is a Normal Distribution

$$x \sim N(\mu, \sigma^2)$$

  - What is your best guess μ and σ ? Any idea?

Probability Land / Data Land



Cook book

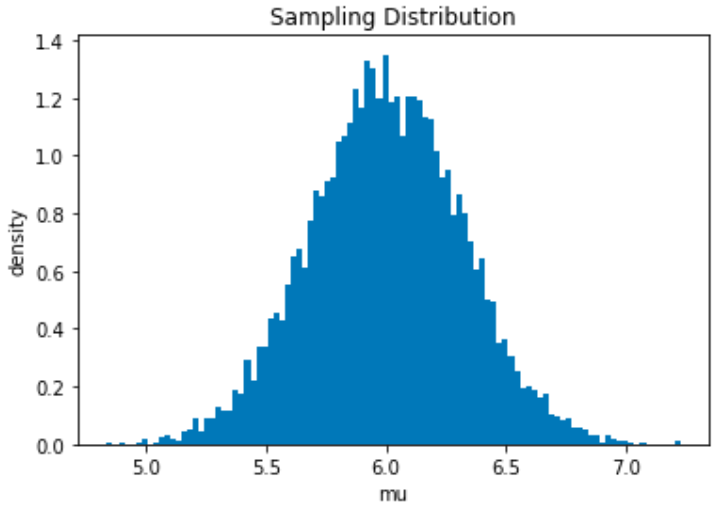
Try to connect probability land and data land

expectation < == > mean

variance < == > (empirical variance)

Goodness of estimation

- Is taking the mean of the data good estimate for the expectation?



- The expectation of estimate should be the real value (**unbiasedness**)
- Sample size $n \to \infty$ estimate should be true value (**consistent**)
  - a bit complicated to formulate that correctly (we don't care here)
- The difference between the estimated and true value should be small (**efficiency**)

Homework (Task 2, Estimation)

Excercise 2 (Estimation)

Assume that you know that your data comes from a Gaussian, with

$$Y \sim N(\mu, \sigma^2 = 1)$$

For the exercise you need to draw samples from a Gaussian. You can do this in R with `rnorm()`. You should draw samples of size 6, assume that $\mu = 42$.

a) Suppose you don't know the parameter $\mu = 42$. Use the mean and the median to estimate the unknown parameter μ given a sample of size 6?

b) Repeat a) 10'000 times, how is the mean distributed. You might want to use a for loop.

c) Discuss b) in terms of the sampling distribution (https://www.youtube.com/watch?v=uPX0NBrJfRI). Is the mean an unbiased estimator of μ?

d) Compare the sampling distribution of the mean and the median. Display the densities (`plot(density(res_mean))`) and calculate the standard deviations. Which estimator would you prefer?

e) [Optional] Is the mean a consistent estimator? Try to understand the term: 'consistent estimator' Do a small simulation to show this fact.

Übung am Ende der Stunde
Besprechung Woche2

Homework (Task 3, Sampling instead of intergration)

Excercise 3 (Sampling instead of integration)

You often have samples from the pdf (much of the Bayesian Statics in the second part is based on this). Instead of calculating expectations of a function $g(y)$ via integration over the pdf $f(y)$ pdf , you can use the $n-$samples[1] $y_i \sim f(y)$ and calculate the mean:

$$E(f(y)) = \frac{1}{n}\sum_{i=1}^{n} f(y_i)$$

Assume that the data comes from standard normal $y_i \sim N(0,1)$.

a) Estimate $E(y)$ using $n = 1000$ samples

b) Estimate $E(y^2)$ using $n = 1000$ samples

c) [Optional] Do the integration over the pdf $N(y;0,1)$ of a standard normal (no solution provided). Hint (probably partial integration will do the trick)

$$E[y^2] = \int_{-\infty}^{\infty} y^2 N(y;0,1)dy$$