Statistics for Data Science

MSc Data Science WiSe 2020/21

Univ.-Prof. Dr. Dirk Ostwald

(1) Introduction

Introduction

- Data science
- Statistics
- Statistics for Data Science
- Exercises

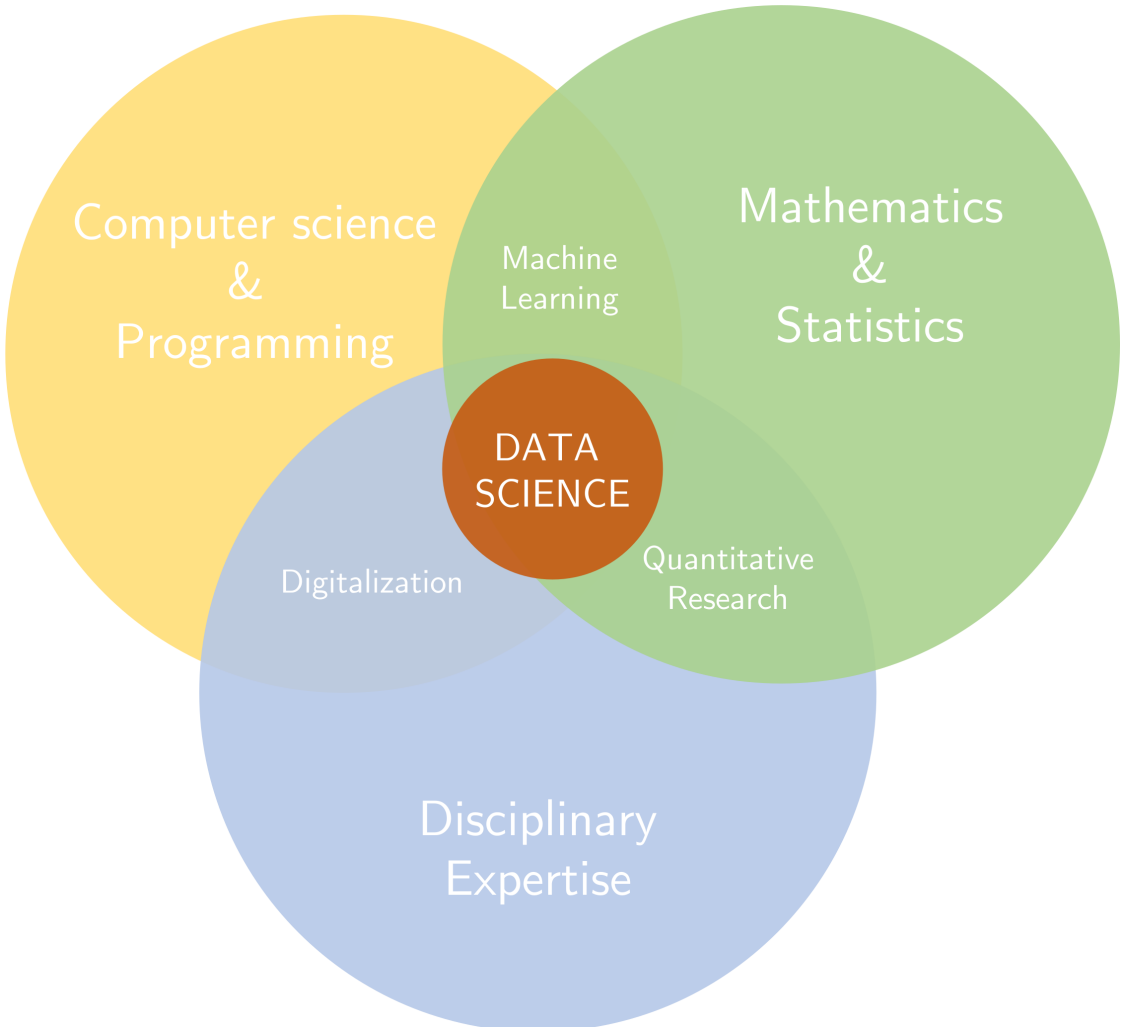
Introduction

- **Data science**
- Statistics
- Statistics for Data Science
- Exercises

Data science

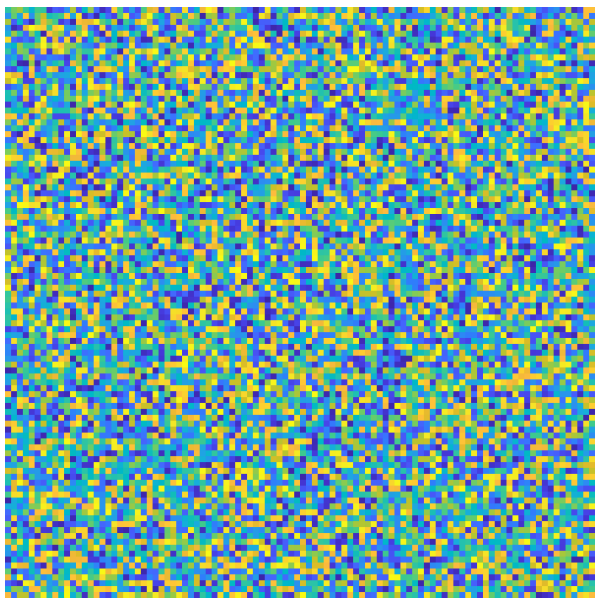
The art of creating meaning from data

Data science

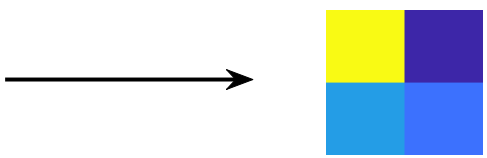


Data analysis is data reduction

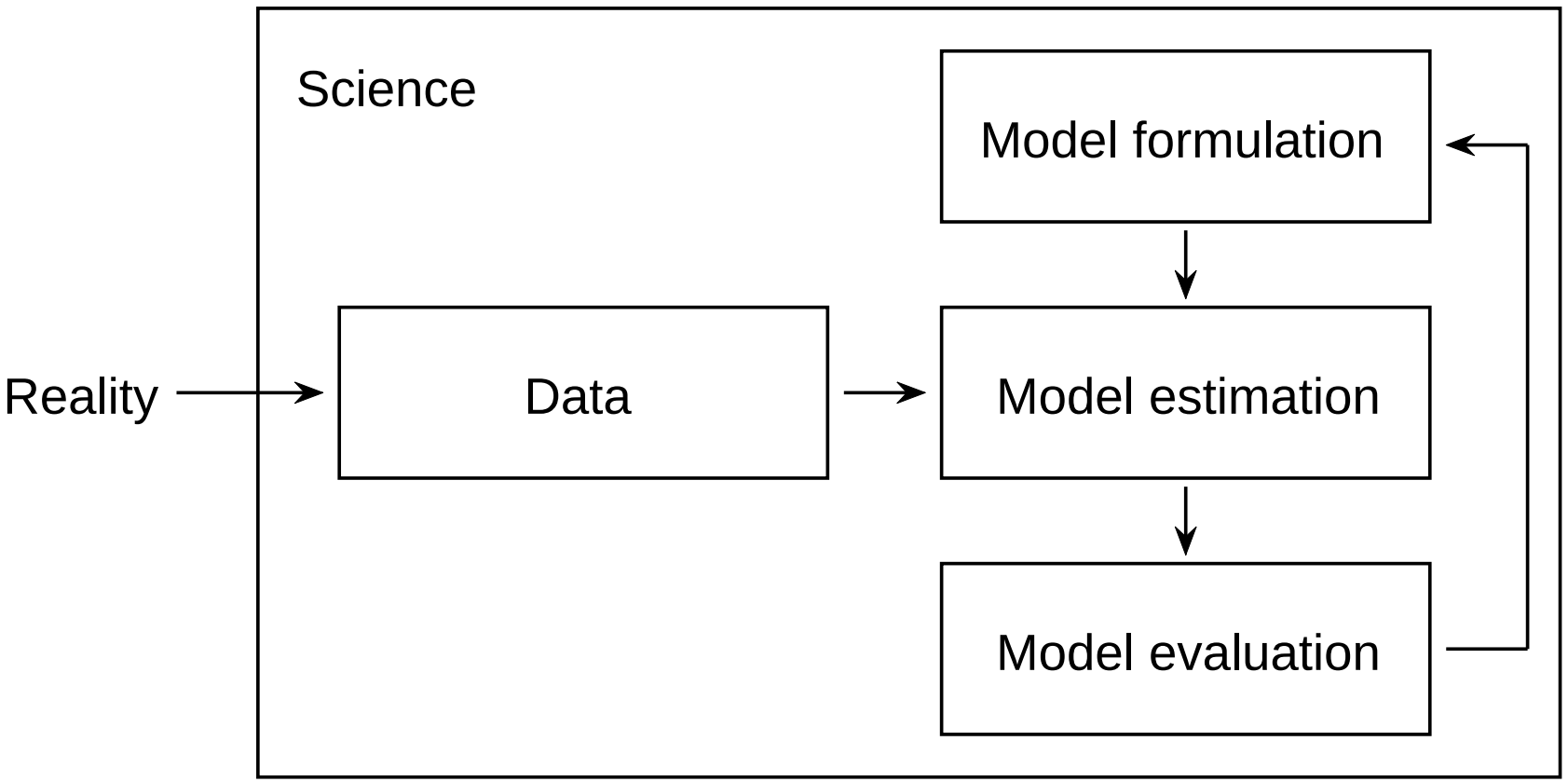
Raw Data



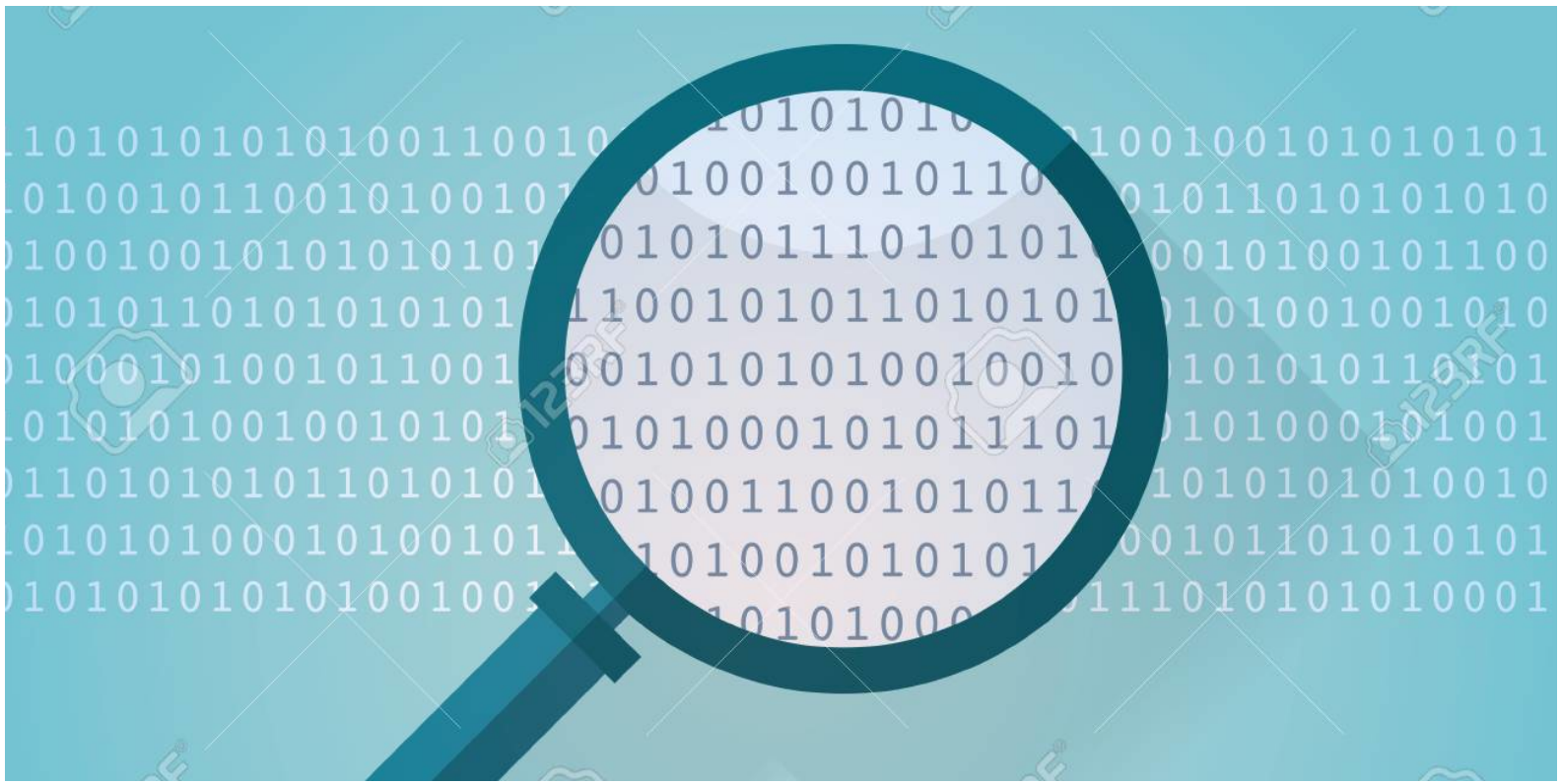
Reduced Data



Data analysis is model-based



Data analysis is an interpretative device



Data science = Statistics = Machine learning = Artificial intelligence

Statistics

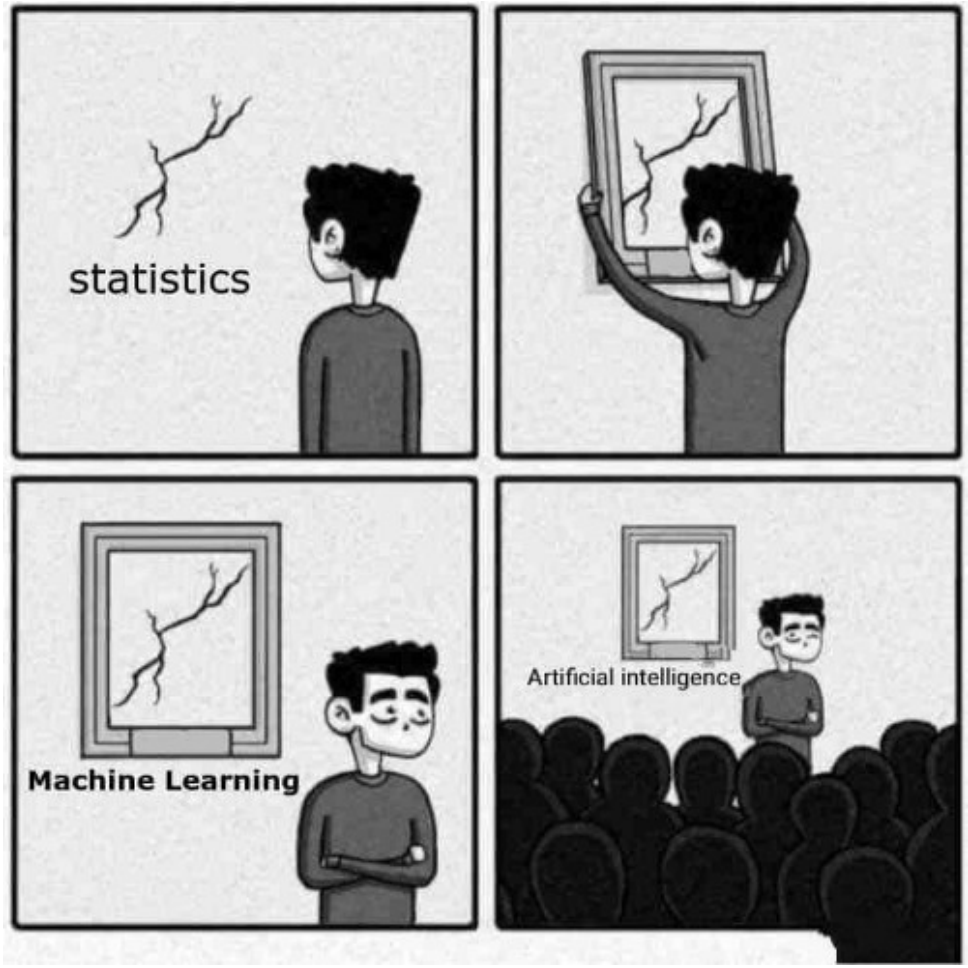
- Probabilistic models
- Theoretical analysis
- Optimality
- Asymptotics
- Science philosophy

Machine learning

- Deterministic models
- Classification
- Bayesian models
- Benchmarking
- Applications

Artificial intelligence

- Deep learning
- Reinforcement
- Decisions
- Data analysis
- Hype



www.instagram.com/sandserifcomics/

www.oak-tree.tech/blog/ml-models

Introduction

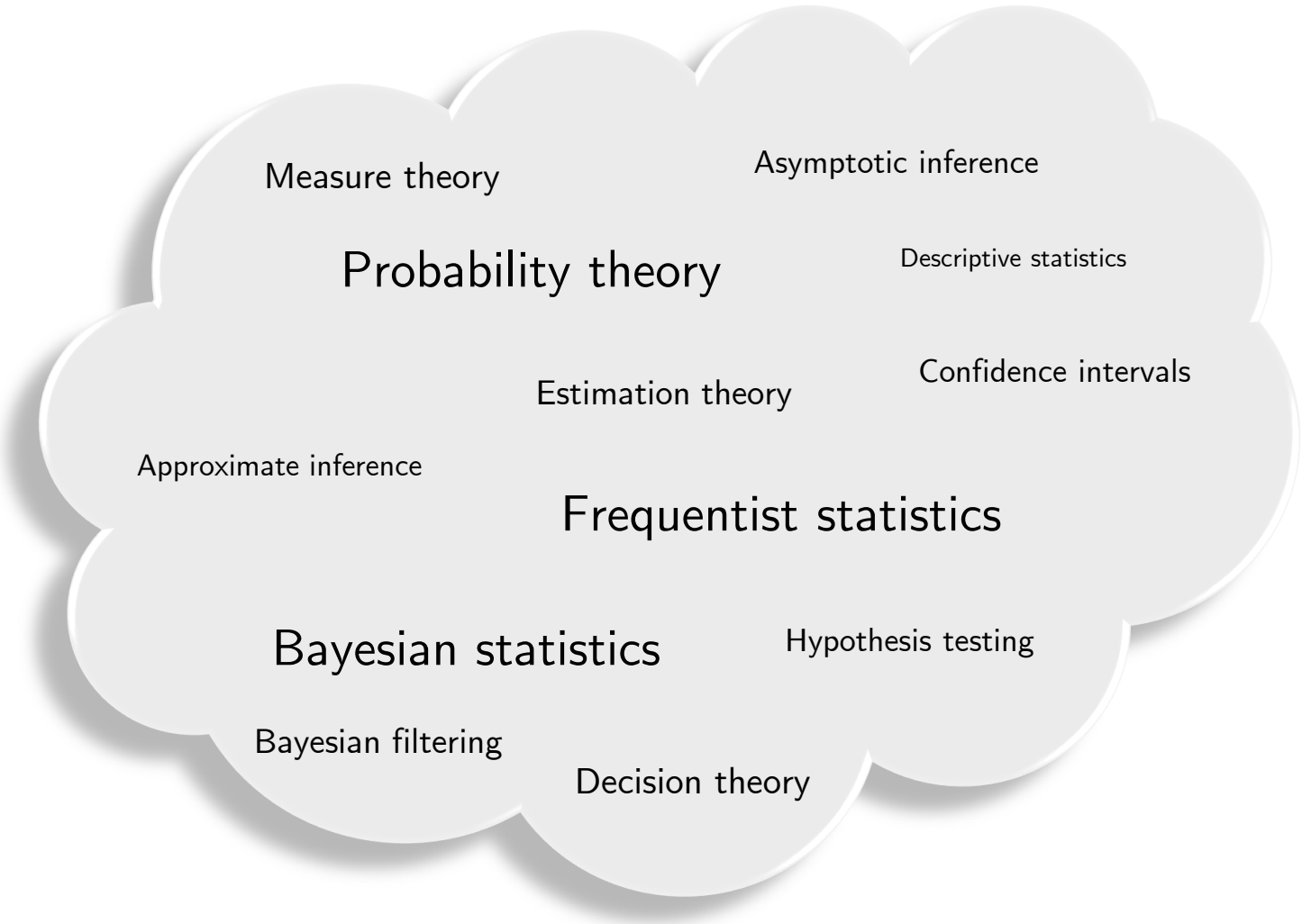
- Data science
- **Statistics**
- Statistics for Data Science
- Exercises

Statistics

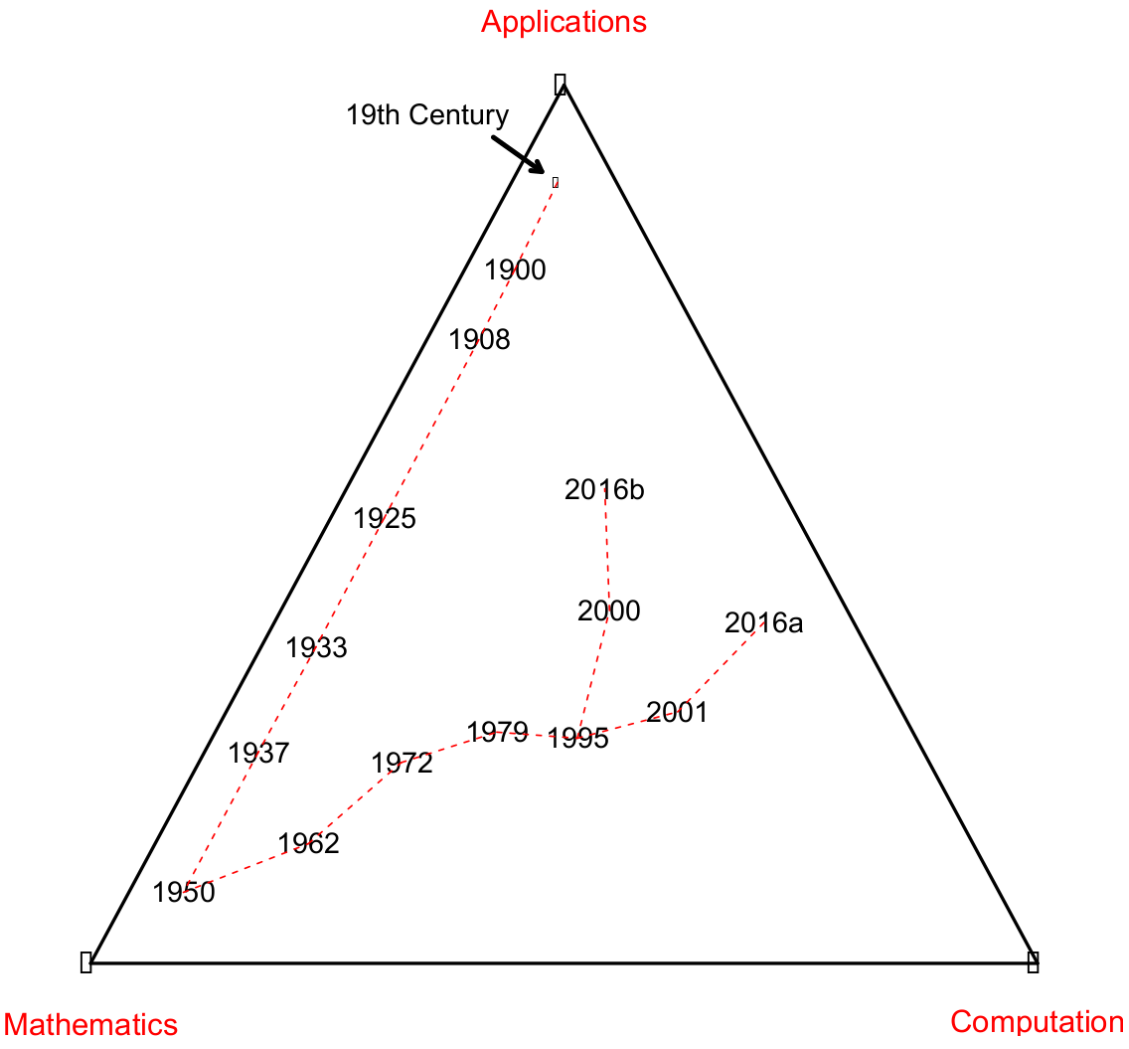
The art of creating meaning from data

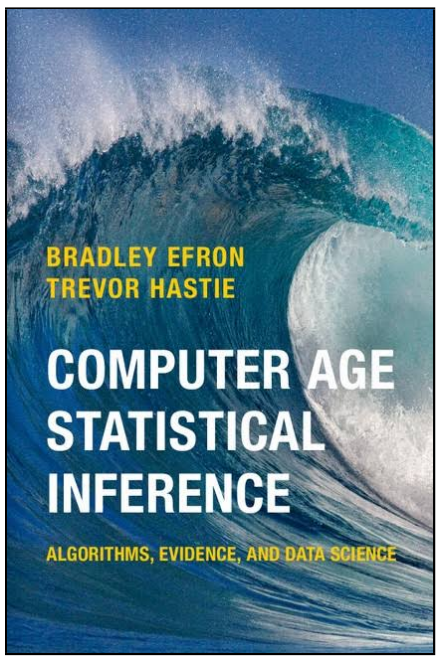
and quantifying its associated uncertainty

Statistics


Measure theory
Asymptotic inference
Probability theory
Descriptive statistics
Confidence intervals
Estimation theory
Approximate inference
Frequentist statistics
Bayesian statistics
Hypothesis testing
Bayesian filtering
Decision theory

Historical development of statistics





Historical development of statistics

1900	Karl Pearson's chi-square test
1908	Student's t statistic
1925	Fisher's Statistical Methods for Research Workers
1933	Neyman and Pearson's optimal hypothesis testing
1937	Neyman's confidence intervals
1950	Wald's statistical decision theory
1950	Savage's & de Finetti's Bayesian decision theory
1961	Raiffa & Schlaifer's Applied statistical decision theory
1962	Tukey's The future of data analysis
1971	Lindley's Bayesian statistics
1972	Cox's proportional hazards
1979	Bootstrap and MCMC
1995	False discovery rate and LASSO
1996	Support vector machines
2000	Microarray and neuroimaging multiple testing
2010	Resurgence of neural networks as deep learning
2015	Data science

Central postulates of Probability theory

- Chance processes can be described mathematically.
- Mathematics can be used to make predictions about random events.
- Reasoning about uncertain events is naturally related to measuring volumes.

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.

- Parameters are fixed, unknown constants, referred to as *true, but unknown* values. No probability statements are made about parameters.

- Statistical procedures are designed to have good long run frequency properties and are typically assessed by studying their sampling distributions.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like "the probability that it will rain this afternoon is 0.5" are meaningful.

- Parameters are fixed, unknown constants, about which probabilistic statements quantifying our uncertainty about their value can be made.

- Probabilistic statements about parameters are made with the help of probability distributions, from which further inferences, such as point or interval estimates, can be derived.

Statistics known as Machine learning

- Principal and independent component analysis
- Logistic regression and linear discriminants
- Support vector machines, kernel methods
- Latent variable and graphical models
- Generalized linear models, neural networks, deep learning
- Gaussian process regression

Statistics known as Artificial intelligence

- Markov decision processes
- Partially observable Markov decision processes
- Reinforcement learning

Examples of community nomenclatures

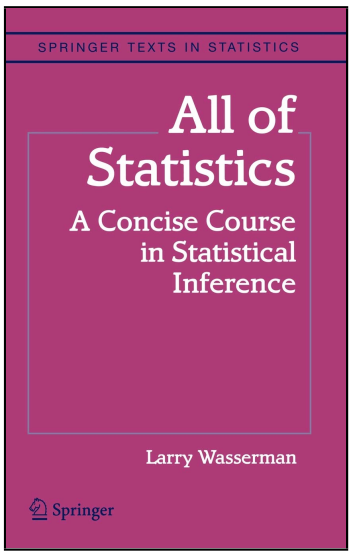
Statistics	Machine Learning	Meaning
Data	Training data	Data
Estimation	Learning, Training	Using data to estimate parameters
Frequentist inference	-	Optimal many samples methods
Bayesian inference	Bayesian inference	Data-based uncertainty updating
Covariates	Features	Structural and known data predictors

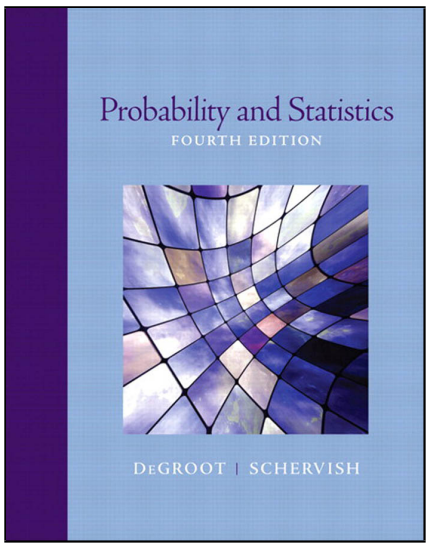
Introduction

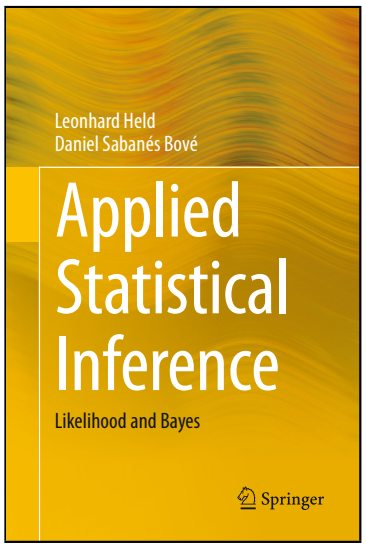
- Data science
- Statistics
- **Statistics for Data Science**
- Exercises

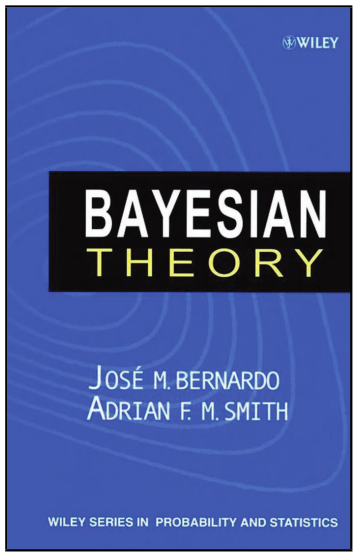
Unit	Date	Theme
(1) Introduction	06.11.2020	
(2) Probability spaces	13.11.2020	Probability theory
(3) Random variables	20.11.2020	Probability theory
(4) Joint distributions	27.11.2020	Probability theory
(5) Transformations	04.12.2020	Probability theory
(6) Expectation and covariance	11.12.2020	Probability theory
(7) Inequalities and limits	18.11.2020	Probability theory
(8) Foundations and maximum likelihood	08.01.2021	Frequentist inference
(9) Finite-sample estimator properties	15.01.2021	Frequentist inference
(10) Asymptotic estimator properties	22.01.2021	Frequentist inference
(11) Confidence intervals	29.01.2021	Frequentist inference
(12) Hypothesis testing	05.02.2021	Frequentist inference
(13) Foundations and conjugate inference	12.02.2021	Bayesian inference
(14) Variational inference	19.02.2021	Bayesian inference
(15) Bayesian estimator properties	26.02.2021	Bayesian inference

Key references









Course components

Component	Aims
Vorlesung	Core course content, knowledge acquisition, breadth
Übung	Active participation, knowledge consolidation, depth
Study questions	Focus, memorization support, examination
Theoretical exercises	Theoretical depth, self-study
Programming exercises	Intuition, Python training, application
Exam	Knowledge reproduction

Course requirements for MSc Data Science students

Vorlesung

- Written exam, pass or fail, 30 questions, 90 min
- 2 questions for each of the 15 units, 1 point per question
- Exam question pool := Study questions
- The study questions pool provided with the lecture slides is final
- $\geq$ 15 points and ≥ 1 point on the 2 question of each unit to pass
- Exam date 05.03.2021, exam resit date 26.03.2021

Übung

- Presentation of one theoretical or programming exercise in class
- Python script with honest solution attempts of all programming exercises
- Programming exercises honest solutions attempts deadline 26.03.2021

Course requirements for Non-MSc Data Science students

Vorlesung

- Graded exam
- 2 questions for each of the 15 units, 1 point per question
- 30 questions, 90 min
- Exam question pool := Study questions
- The study questions pool provided with the lecture slides is final
- $\geq$ 15 points to pass.
- Exam date 05.03.2021, exam resit date 26.03.2021

Übung

- Python script with honest solution attempts of all programming exercises
- Programming exercises honest solutions attempts deadline 26.03.2021

Übung active participation criteria

- 15 min presentation of one theoretical or programming exercise in class
  - Theoretical exercise $\Rightarrow$ Beamer presentation
  - Programming exercise $\Rightarrow$ Jupyter Notebook
- Except on 13.11.2020, active participation can be failed
- Think of the presentation and subsequent discussion as a mini oral exam
- Binary feedback will be provided only on the Tuesday before the presentation
- The link to the exercise pool is provided by email
- The link to the exercise sign up form is provided by email
- The programming exercise pool provided with the lecture slides is final

Introduction

- Data science
- Statistics
- Statistics for Data Science
- **Exercises**

Exercises

Study questions

1. Give a definition of Data Science.
2. Give a definition of Statistics.
3. Name three central postulates of Probability theory.
4. Name three central postulates of Frequentist inference.
5. Name three scientists involved in the development of Frequentist statistics.
6. Name three central postulates of Bayesian inference.
7. Name three scientists involved in the development of Bayesian statistics.
8. Name five typical topics in Statistics.
9. Name three topics commonly discussed in Machine Learning.
10. Name three topics commonly discussed in Artificial Intelligence.

Exercises

Programming exercises

1. Sample a univariate Gaussian using scipy.stats.
2. Evaluate the PDF of a univariate Gaussian using scipy.stats.
3. Visualize the PDF of a univariate and a normalized sample histogram of samples from a univariate Gaussian with identical parameters on top of each other using Matplotlib.