Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Lehrstuhl für
INFORMATIK I
Effiziente Algorithmen und
wissensbasierte Systeme

Institut für Informatik

Seminar :
Mathematical Foundations of Data Science

Summer Term 2021

Introductory Class on April, 14, 2021

Chair of Computer Science I - Algorithms and Complexity

Kamyar Khodamoradi **Joachim Spoerhase** Alexander Wolff

Agenda

1. Organization
2. Concept of the Seminar
3. Topic Assignment
4. Introduction to IPE

Agenda

1. Organization
2. Concept of the Seminar
3. Topic Assignment
4. Introduction to IPE

All classes via ZOOM.
Switch on cameras!

Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic

Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic
  (ca 5 min., ca. 3 slides)

Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic (ca 5 min., ca. 3 slides)

  Content:

  - outlook to the talk
  - problem motivation
  - presenting key results

Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic
  (ca 5 min., ca. 3 slides)

  Content:
  - outlook to the talk
  - problem motivation
  - presenting key results

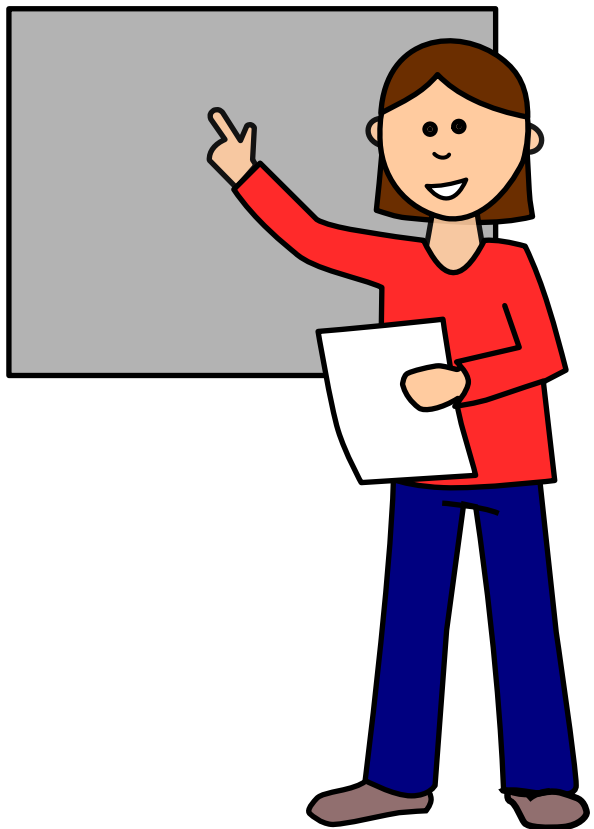
  Goal:
  - getting started quickly
  - select key parts, results (agree with supervisor)
  - synchronize prerequisites with other talks!
  - practice talking
  - getting feedback without grading

Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic
  (ca 5 min., ca. 3 slides)
- Wed, April, 29: **Primer on Tail Bounds** (Kamyar Khodamoradi)

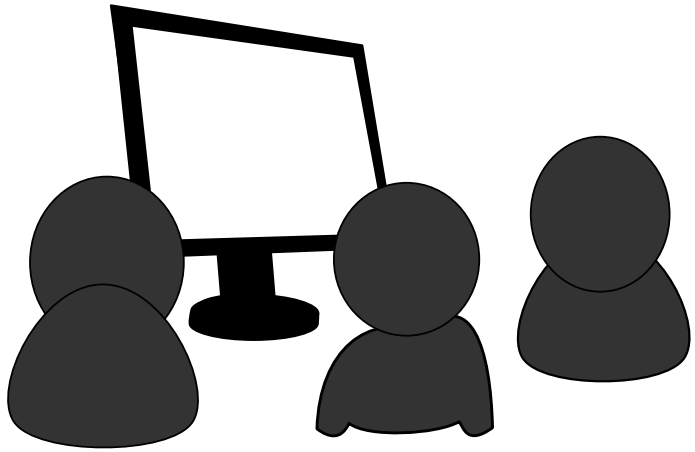
Organization

- Wed, April, 14, 2021: **Introduction**
- Wed, April, 21: **Short Talks** to every topic (ca 5 min., ca. 3 slides)
- Wed, April, 29: **Primer on Tail Bounds** (Kamyar Khodamoradi)
- Wed, May, 19–July, 14: **Talks** (one per week)
- Fr, July, 16: hand in **Reports**



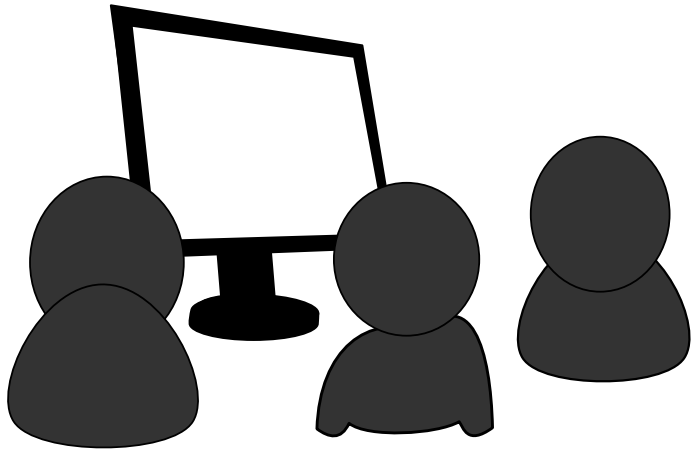
Talks

- roughly 45 minutes **talk**
  (60 minutes for groups of two)



Talks

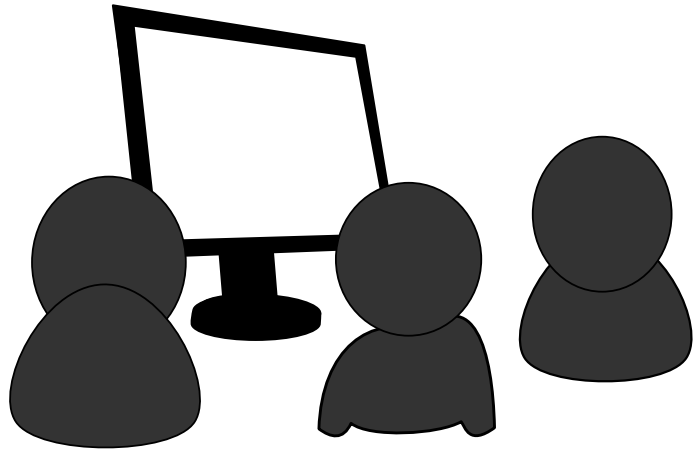
- roughly 45 minutes **talk**
  (60 minutes for groups of two)



This is not enough to cover a full book chapter!

$\rightarrow$ identify most important results, treat essential parts comprehensively, outline less essential parts

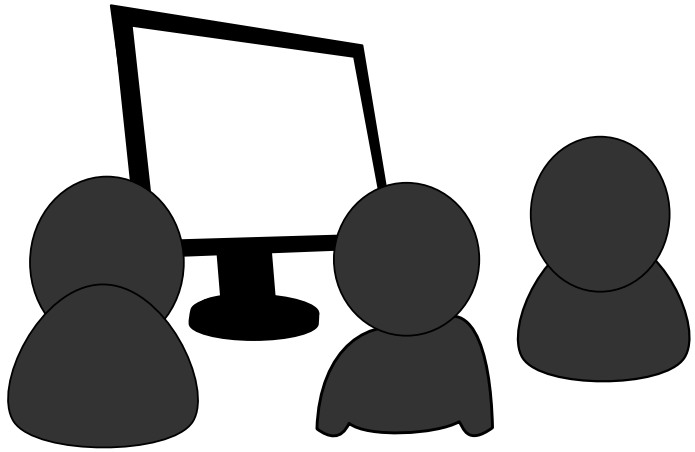
Talks



- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

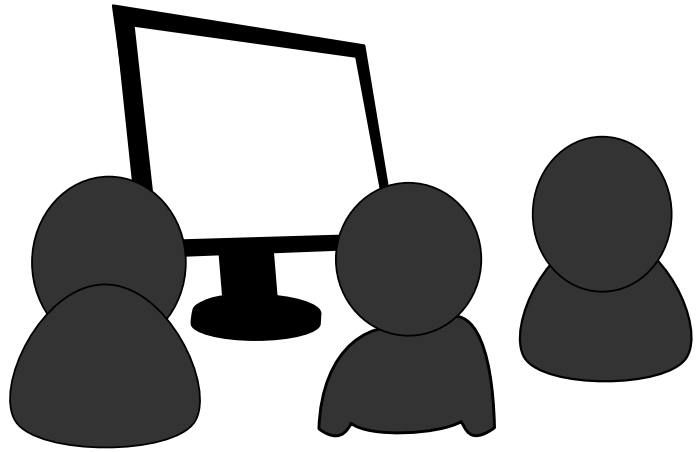
Talks



- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!
- two quizzes (comprehension questions) during each talk, e.g. via PINGO: `pingo.upb.de`

Talks

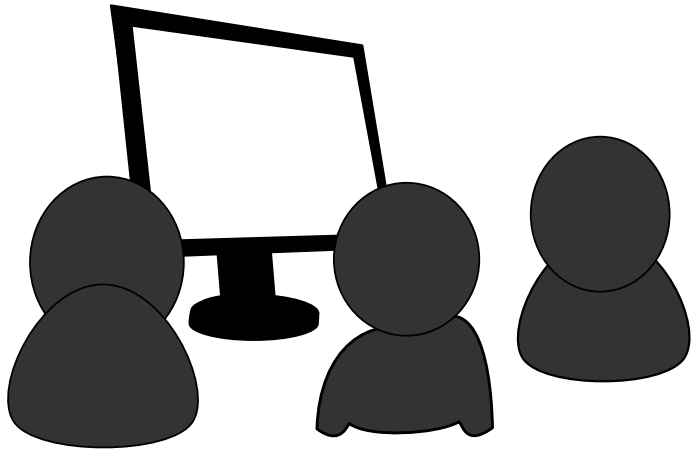


- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

- two quizzes (comprehension questions) during each talk, e.g. via PINGO: `pingo.upb.de`
- at the end, presenters pose a question that the other participants discuss via the think-pair-share method

Talks

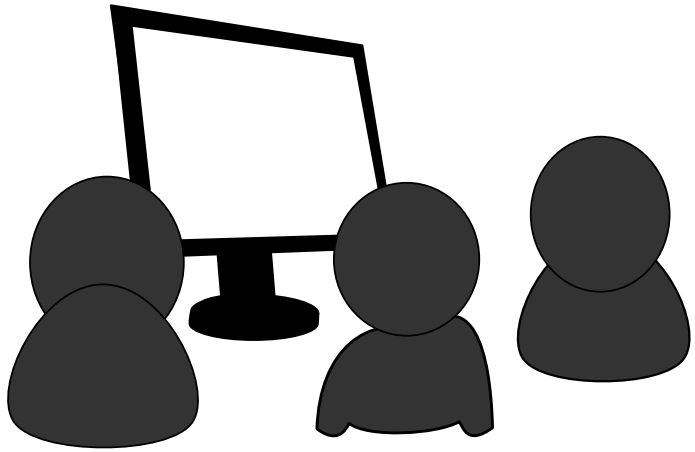


- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

- two quizzes (comprehension questions) during each talk, e.g. via PINGO: `pingo.upb.de`
- at the end, presenters pose a question that the other participants discuss via the think-pair-share method
- every participant must share her/his thoughts at least once over the seminar

Talks



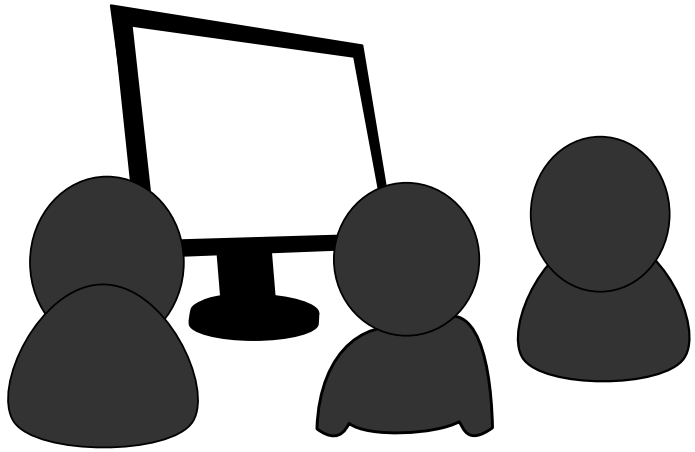
- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

Preliminary discussion:

- **three** weeks before the talk:
  discuss the **structure of your talk** with your supervisor
- **two** weeks before your talk:

Talks



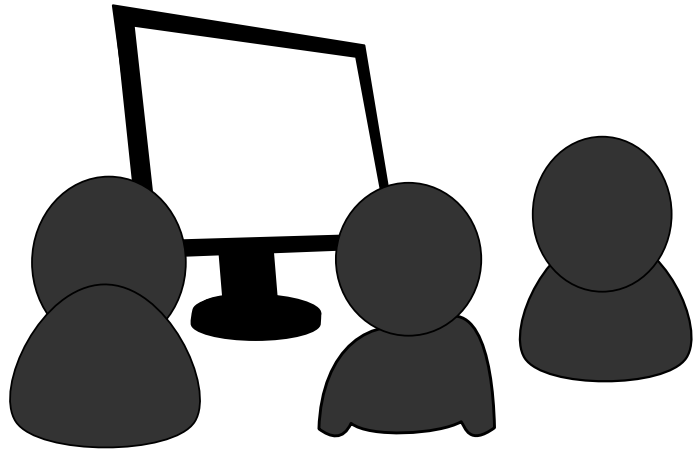
- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

Preliminary discussion:

- **three** weeks before the talk:
  discuss the **structure of your talk** with your supervisor
- **two** weeks before your talk:
  Discuss your **slides** with your supervisor

Talks



- roughly 45 minutes **talk**
  (60 minutes for groups of two)
- during/after **interaction / discussion**
  (interactive examples, quizzes, exercises, related topics, etc.) (does not contribute to time)

  incorporate ideas from discussion to reports!

These deadlines are strict (except for 1. talk)!

Preliminary discussion:

- **three** weeks before the talk:
  discuss the **structure of your talk** with your supervisor
- **two** weeks before your talk:
  Discuss your **slides** with your supervisor

Reports

- roughy 10 pages for two ($\sim$ 8 for one)

Reports

- roughy 10 pages for two ($\sim$ 8 for one)
- not only a summary of the topic; e. g. adding further details or intuition for proofs

Reports

- roughy 10 pages for two ($\sim$ 8 for one)
- not only a summary of the topic; e. g. adding further details or intuition for proofs
- relation to other topics

Reports

- roughy 10 pages for two ($\sim$ 8 for one)
- not only a summary of the topic; e. g. adding further details or intuition for proofs
- relation to other topics
- LaTeX template

Reports

- roughy 10 pages for two ($\sim$ 8 for one)
- not only a summary of the topic; e. g. adding further details or intuition for proofs
- relation to other topics
- LaTeX template
- **Preliminary version** of report two weeks after talk, but at the latest until Mon, July, 5

Passing & Grading

Requirements for Passing the Seminar

- giving a talk to the selected topic including the required interactive parts (two quizes, question/discussion)
- creating a report
- presence at the talks
- absence at most once
- participation in the discussions (share your thoughts at least once in the plenum)

Passing & Grading

Requirements for Passing the Seminar

- giving a talk to the selected topic including the required interactive parts (two quizes, question/discussion)
- creating a report
- presence at the talks
- absence at most once
- participation in the discussions (share your thoughts at least once in the plenum)

Grading

- Talk (content, design of the slides, comprehensibility)
- Report (content, language, spelling, connection to other topics)
- 50 : 50

Concept of the Seminar

A Short History of Computer Science

Application	Theory	Start
programming, compilers, operating systems	computability, automata theory, formal languages	1960s

A Short History of Computer Science

Application	Theory	Start
programming, compilers, operating systems	computability, automata theory, formal languages	1960s
perform well-defined tasks (e.g. sorting, searching, optimization)	algorithms, complexity theory	1970s

A Short History of Computer Science

Application	Theory	Start
programming, compilers, operating systems	computability, automata theory, formal languages	1960s
perform well-defined tasks (e.g. sorting, searching, optimization)	algorithms, complexity theory	1970s
extracting information and learning from massive data (for user applications)	data science and machine learning (mathematical foundations)	recent years

Book: Foundations of Data Science

computability,
automata theory,
formal languages

algorithms,
complexity theory

Book: Foundations of Data Science

computability,
automata theory,
formal languages

algorithms,
complexity theory

traditional TCS curriculum

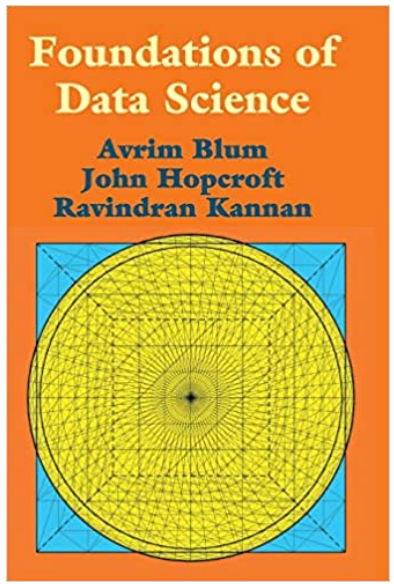
Book: Foundations of Data Science

computability,
automata theory,
formal languages

algorithms,
complexity theory

traditional TCS curriculum

add



data science and
machine learning
(mathematical
foundations)

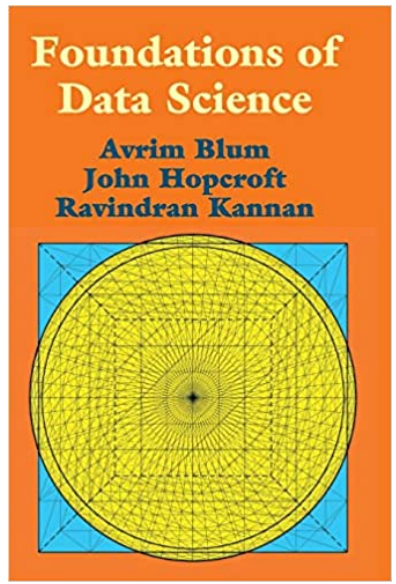
Book: Foundations of Data Science

computability,
automata theory,
formal languages

algorithms,
complexity theory

traditional TCS curriculum

add



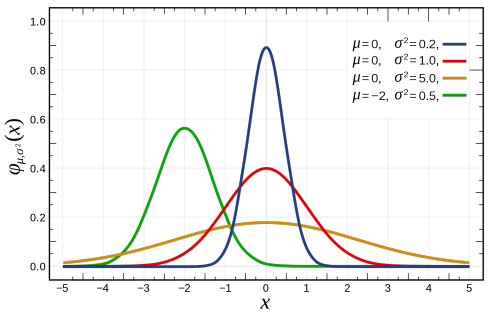
data science and
machine learning
(mathematical
foundations)

"[. . .] With this in mind we have written this book to cover the theory we expect to be useful in the next 40 years, just as an understanding of automata theory, algorithms, and related topics gave students an advantage in the last 40 years. [. . .]"

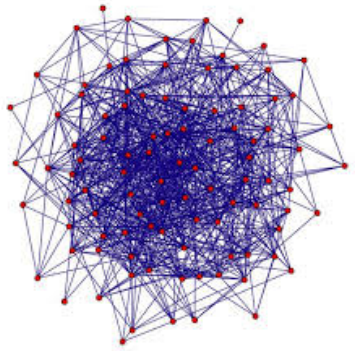
Key Elements and Tools

assumptions on the input

(rather than analyzing worst case)



probability distributions

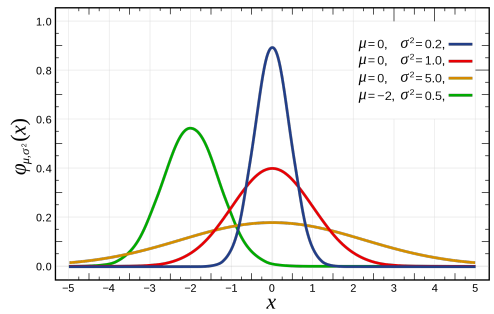


models

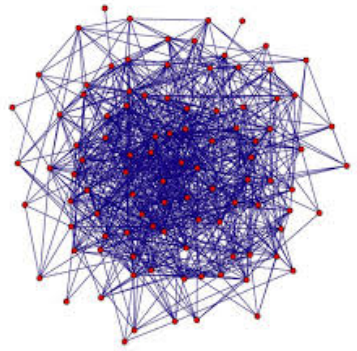
Key Elements and Tools

assumptions on the input

(rather than analyzing worst case)



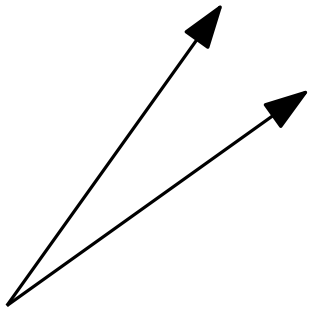
probability distributions



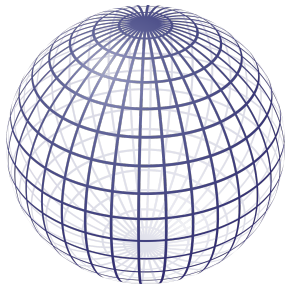
models

high-dimensional geometry

$d \to \infty$



	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2



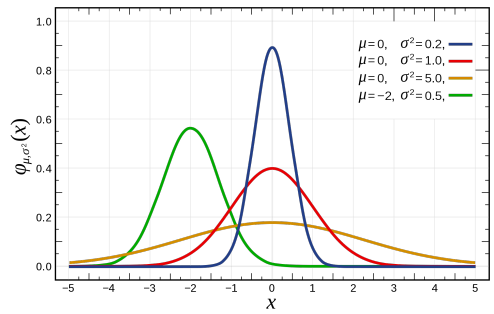
document term matrix

adjacency matrix of web graph

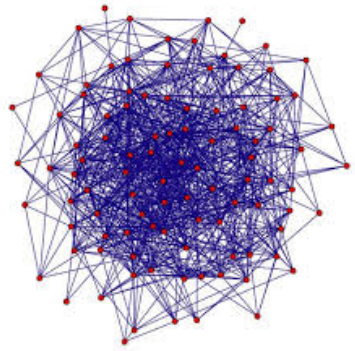
Key Elements and Tools

assumptions on the input

(rather than analyzing worst case)



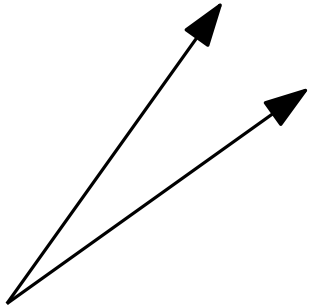
probability distributions



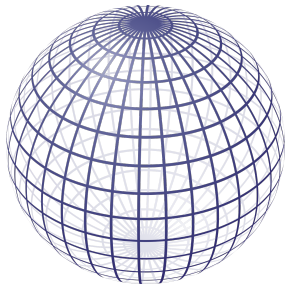
models

high-dimensional geometry

$d \to \infty$



	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2



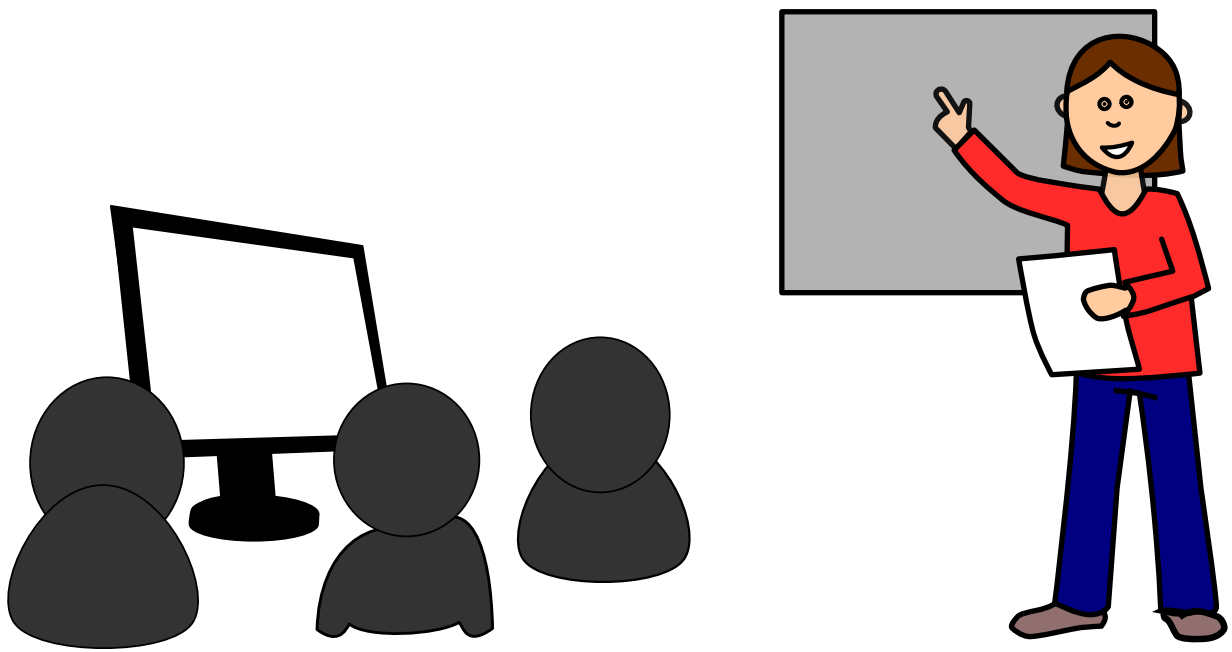
document term matrix

adjacency matrix of web graph

mathematical tools: probability, statistics, linear algebra, (multivariate) analysis

Concept of the Seminar

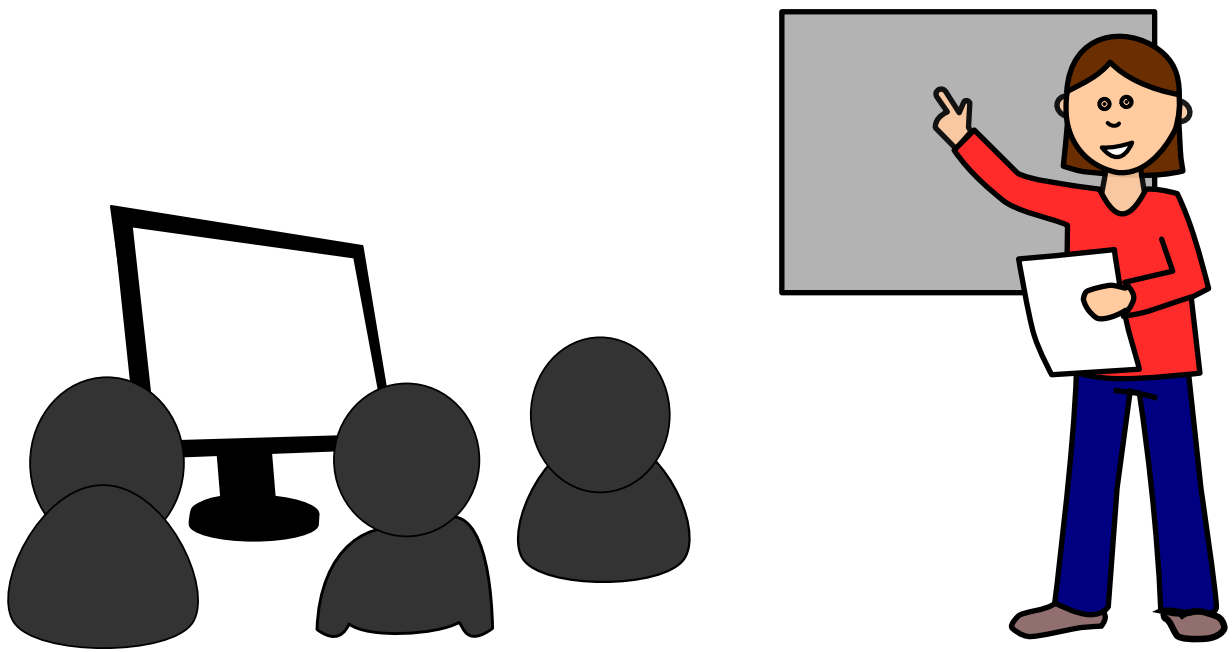
Consider this a reading group were we jointly learn the mathematical foundations of this subject



Concept of the Seminar

Consider this a reading group were we jointly learn the mathematical foundations of this subject

Disclaimer: We do not cover applied aspects such as software, libraries, specific applications

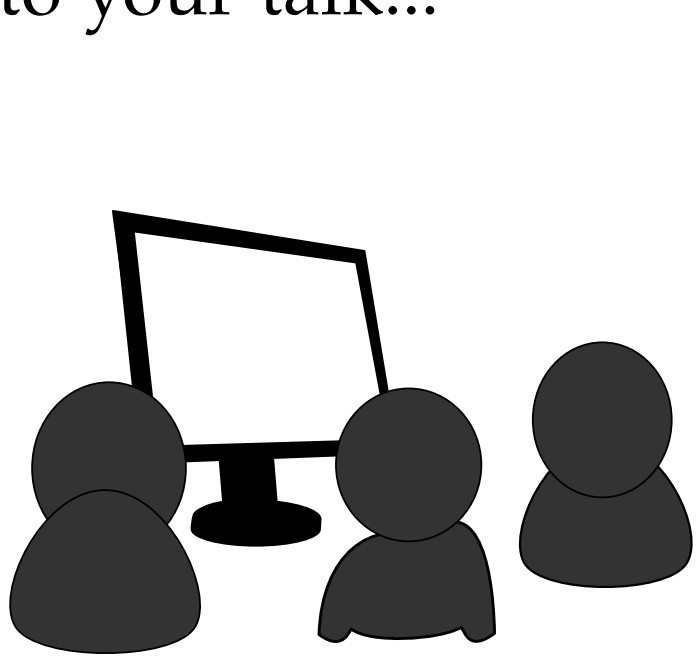


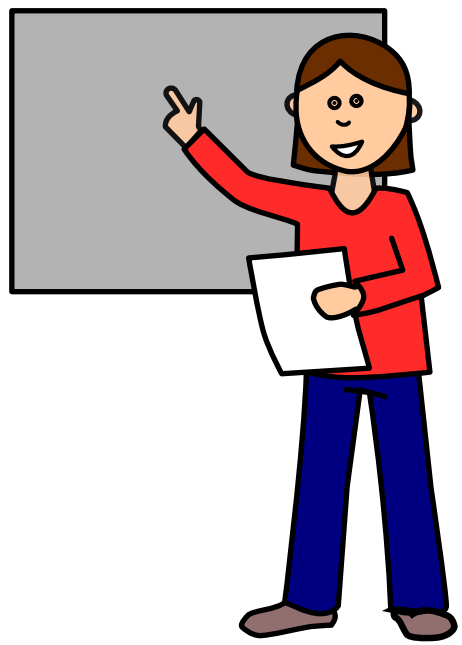
Concept of the Seminar

Consider this a reading group were we jointly learn the mathematical foundations of this subject

Disclaimer: We do not cover applied aspects such as software, libraries, specific applications

Feel free to add potential applications as illustrative motivation to your talk...





Seminar Topics

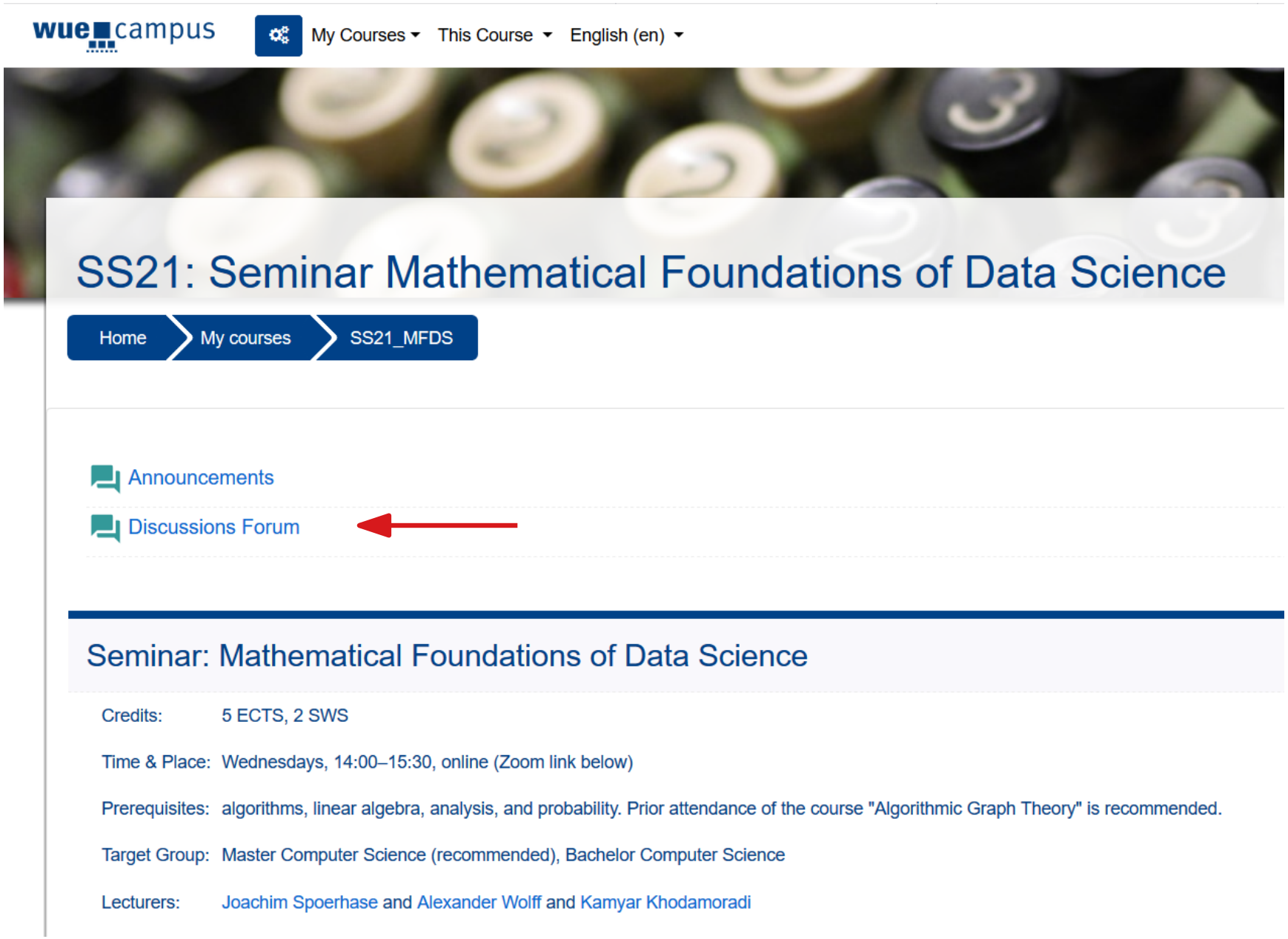
General Information

- You will be assigned one section or chapter from the book: *Foundations of Data Science*.
- You should select the key parts/results from these sections for your presentation.
- Check with your supervisor (as mentioned before).

List of Topics

- High-Dimensional Geometry
- Best-Fit Subspaces and Singular Value Decomposition (SVD)
- Random Walks and Markov Chains
- Machine Learning
- Algorithms for Massive Data
- Clustering Techniques
- Analysis of Random Graphs
- Social Choice
- Compressed Sensing

Discussions Forum

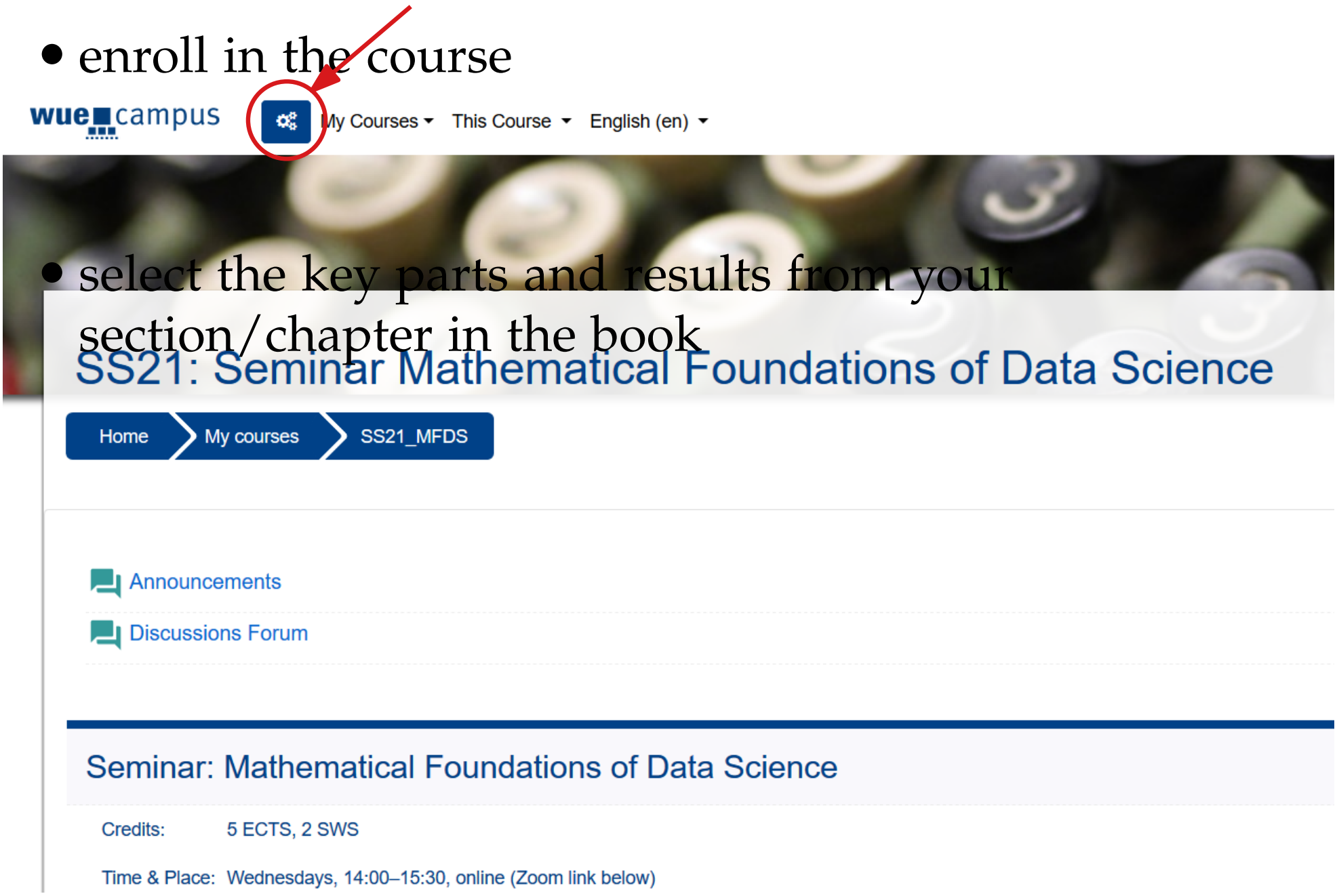


Next Steps

- enroll in the course

Next Steps

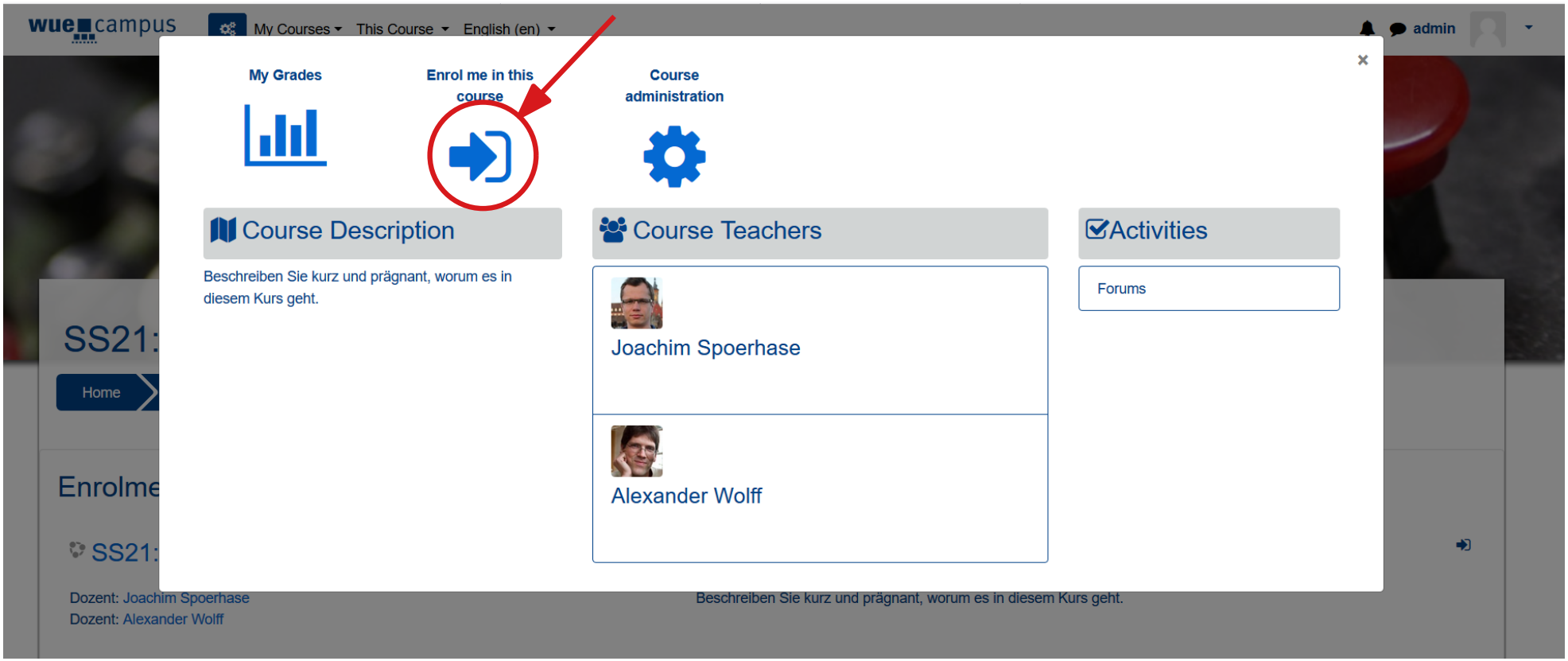
- enroll in the course



- select the key parts and results from your section/chapter in the book

Next Steps

- enroll in the course



Next Steps

- enroll in the course
- get an overview over the book section/chapter assigned to you

Next Steps

- enroll in the course
- get an overview over the book section/chapter assigned to you
- select the key parts and results from your section/chapter in the book

Next Steps

- enroll in the course
- get an overview over the book section/chapter assigned to you
- select the key parts and results from your section/chapter in the book
- identify potential prerequisites from previous sections

Next Steps

- enroll in the course
- get an overview over the book section/chapter assigned to you
- select the key parts and results from your section/chapter in the book
- identify potential prerequisites from previous sections
- double check with your supervisor and synchronize with other students (presenting the prerequisites); send by Monday, April, 19!

Next Steps

- enroll in the course
- get an overview over the book section/chapter assigned to you
- select the key parts and results from your section/chapter in the book
- identify potential prerequisites from previous sections
- double check with your supervisor and synchronize with other students (presenting the prerequisites); send by Monday, April, 19!
- prepare a short presentation for next week (April 21st).

Presentation Software IPE

Finally:

Demonstration of the IPE program
for creating images and slides
http://ipe.otfried.org/