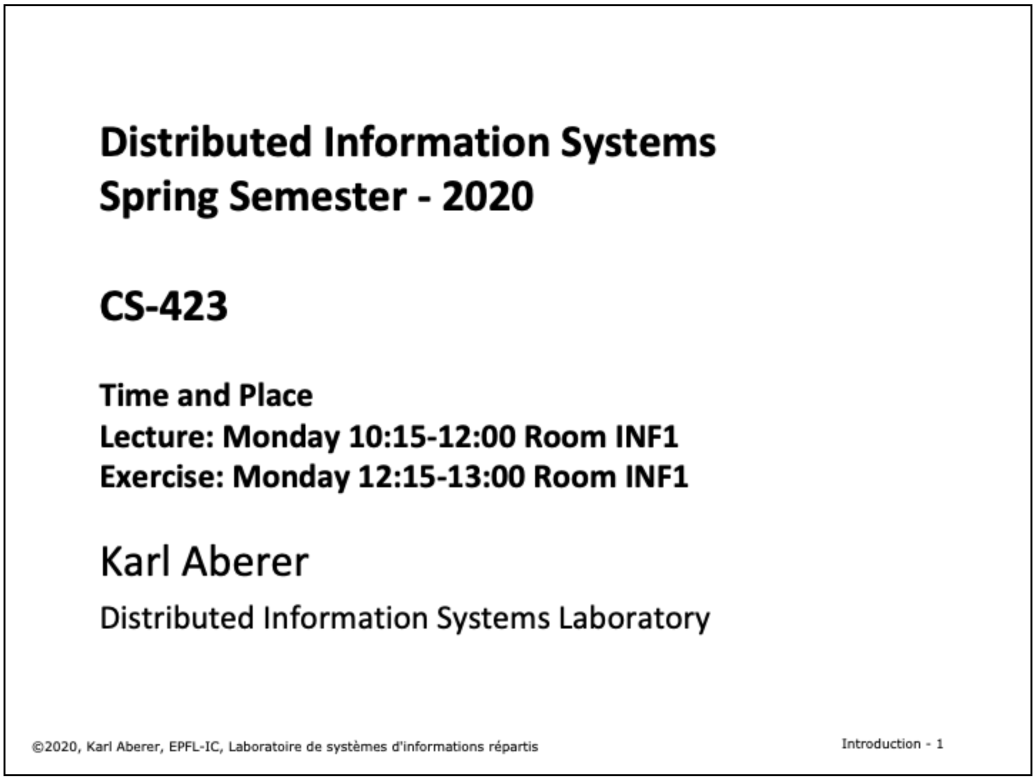


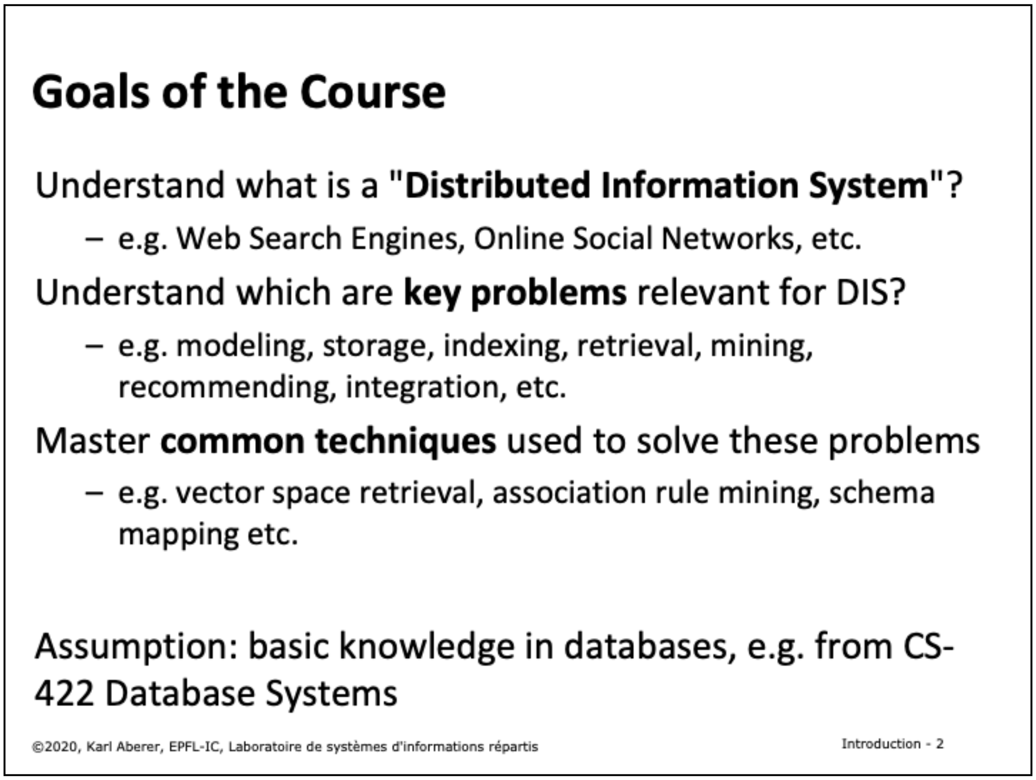
Distributed Information Systems
Spring Semester - 2020

CS-423

Time and Place
Lecture: Monday 10:15-12:00 Room INF1
Exercise: Monday 12:15-13:00 Room INF1

Karl Aberer

Distributed Information Systems Laboratory



Goals of the Course

Understand what is a "**Distributed Information System**"?

- e.g. Web Search Engines, Online Social Networks, etc.

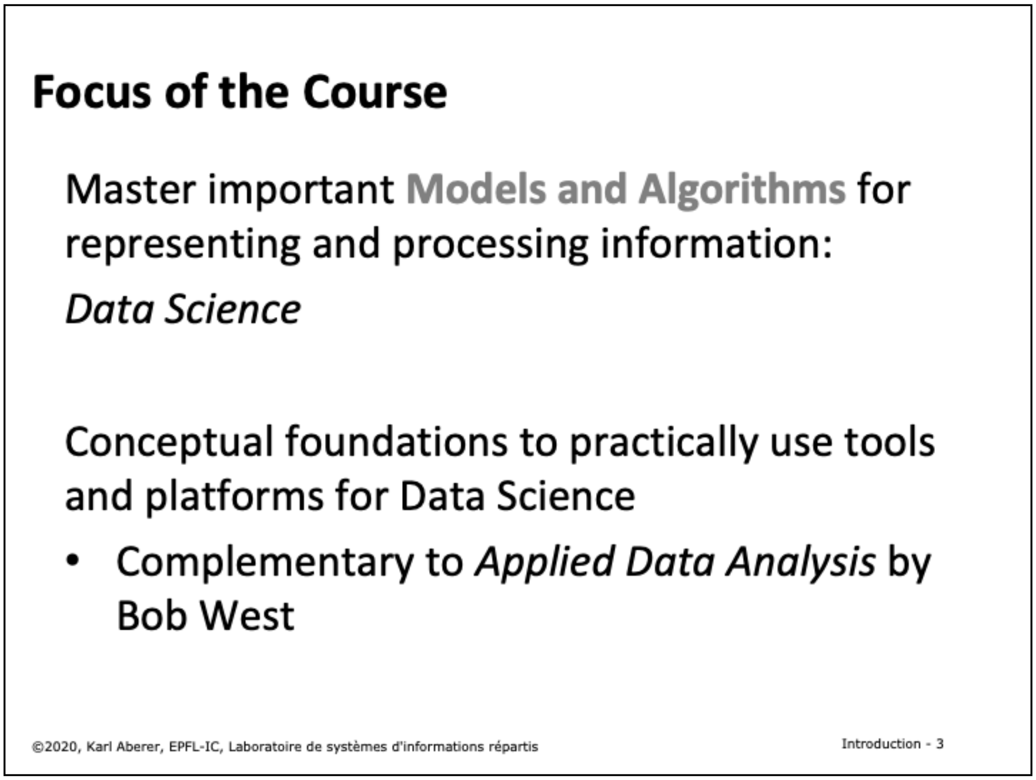
Understand which are **key problems** relevant for DIS?

- e.g. modeling, storage, indexing, retrieval, mining, recommending, integration, etc.

Master **common techniques** used to solve these problems

- e.g. vector space retrieval, association rule mining, schema mapping etc.

Assumption: basic knowledge in databases, e.g. from CS-422 Database Systems



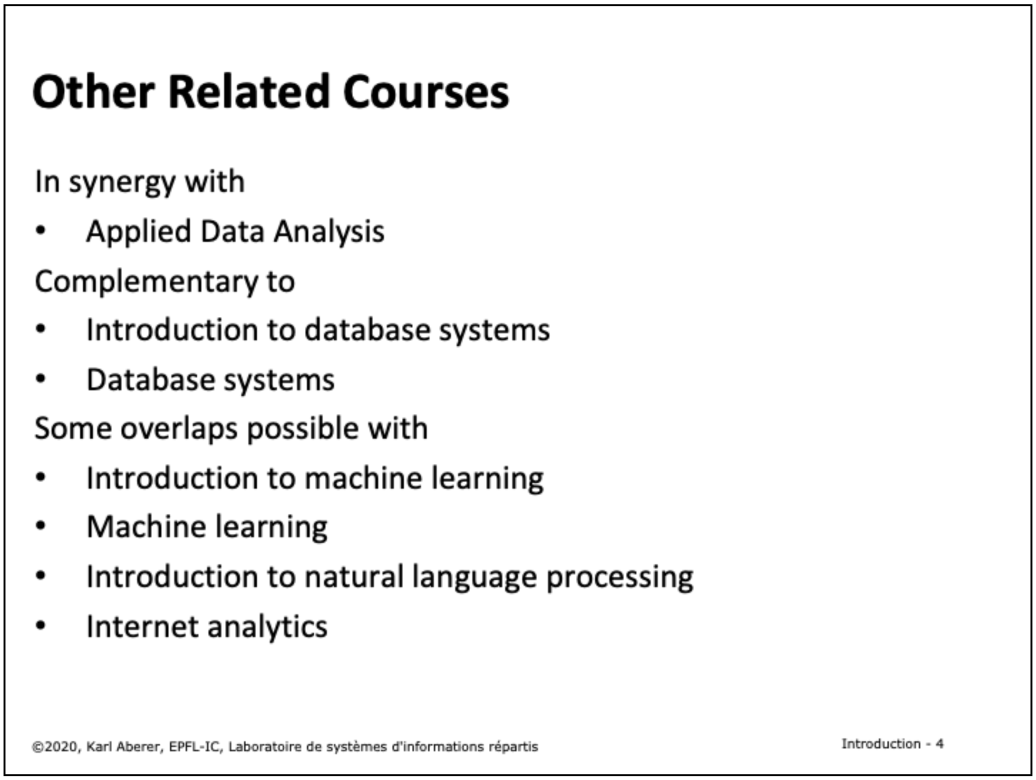
Focus of the Course

Master important **Models and Algorithms** for representing and processing information:

Data Science

Conceptual foundations to practically use tools and platforms for Data Science

- Complementary to *Applied Data Analysis* by Bob West



Other Related Courses

In synergy with

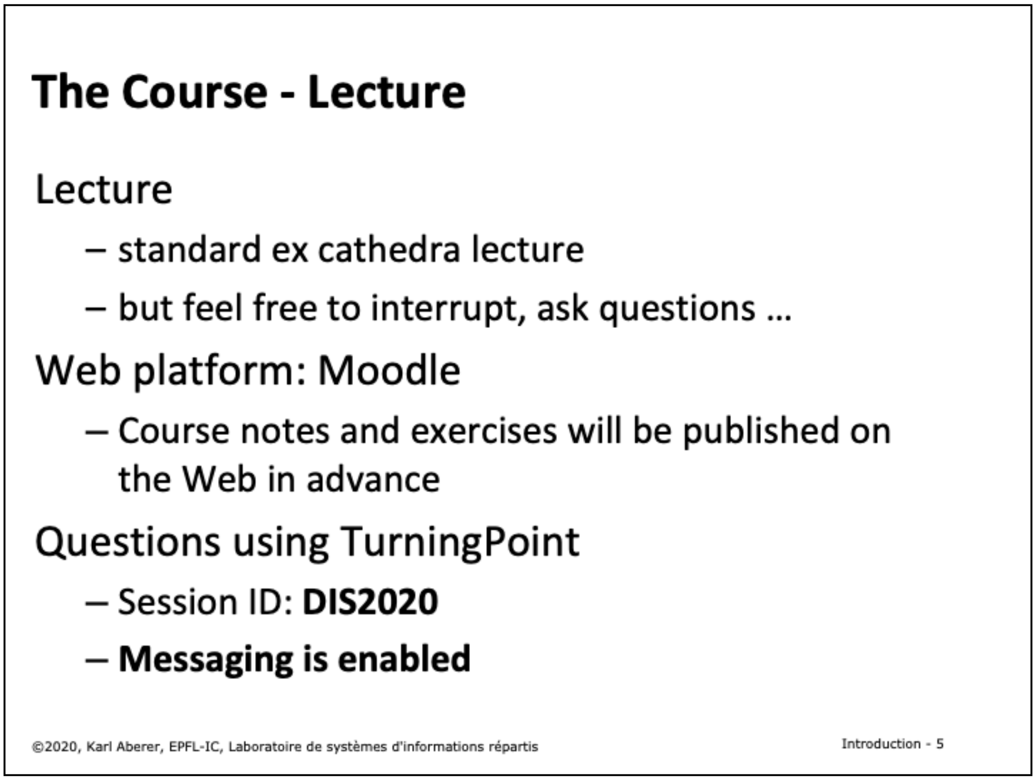
- Applied Data Analysis

Complementary to

- Introduction to database systems
- Database systems

Some overlaps possible with

- Introduction to machine learning
- Machine learning
- Introduction to natural language processing
- Internet analytics



The Course - Lecture

Lecture

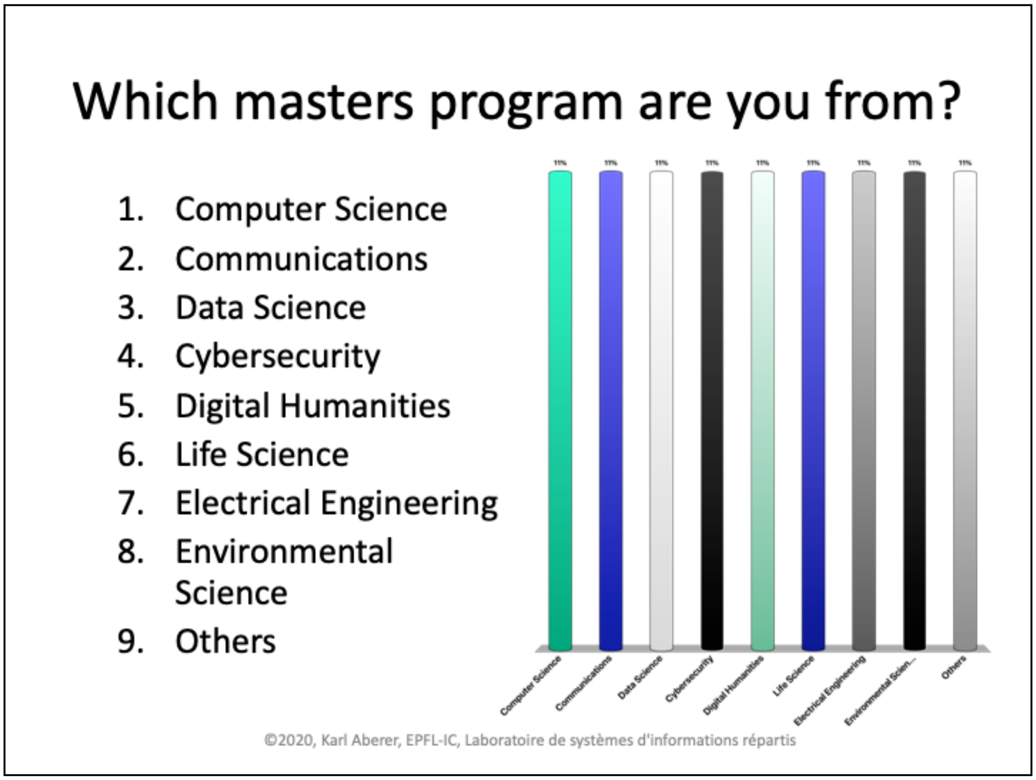
- standard ex cathedra lecture
- but feel free to interrupt, ask questions …

Web platform: Moodle

- Course notes and exercises will be published on the Web in advance

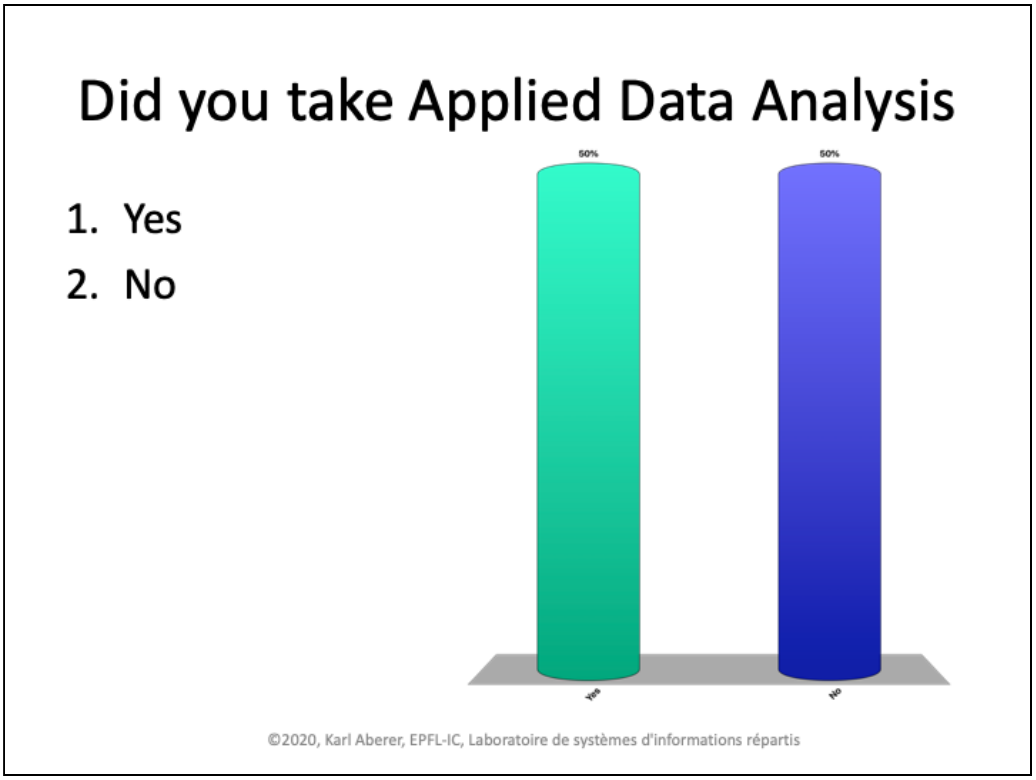
Questions using TurningPoint

- Session ID: **DIS2020**
- **Messaging is enabled**



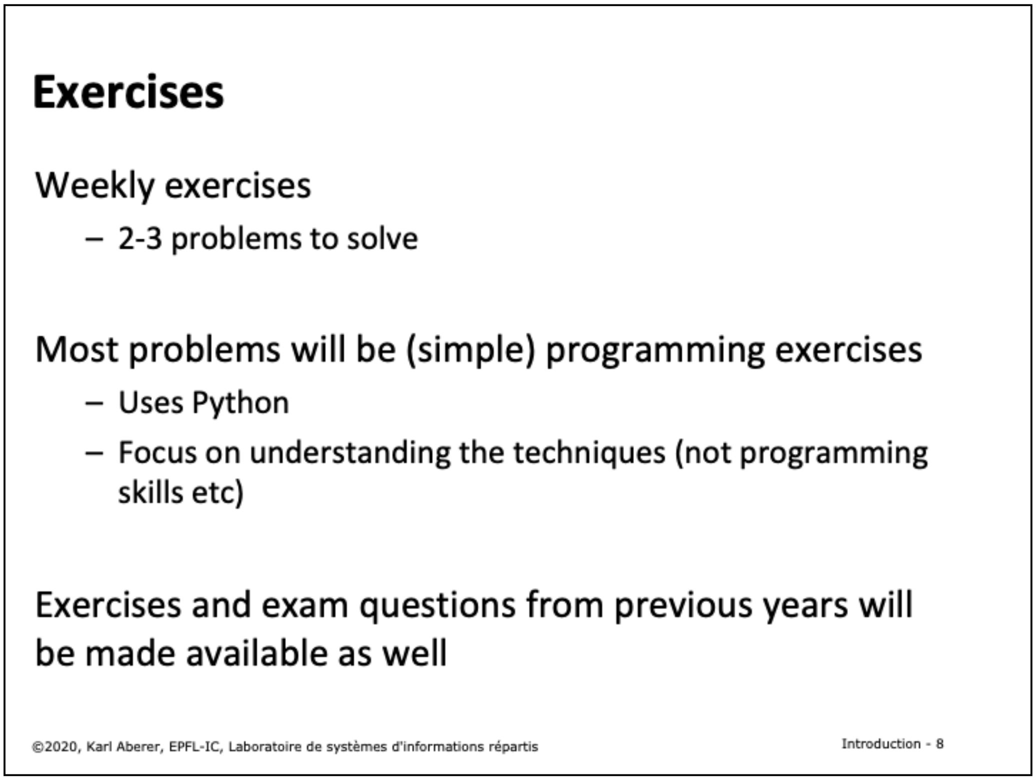
Which masters program are you from?

1. Computer Science
2. Communications
3. Data Science
4. Cybersecurity
5. Digital Humanities
6. Life Science
7. Electrical Engineering
8. Environmental Science
9. Others



Did you take Applied Data Analysis

1. Yes
2. No



Exercises

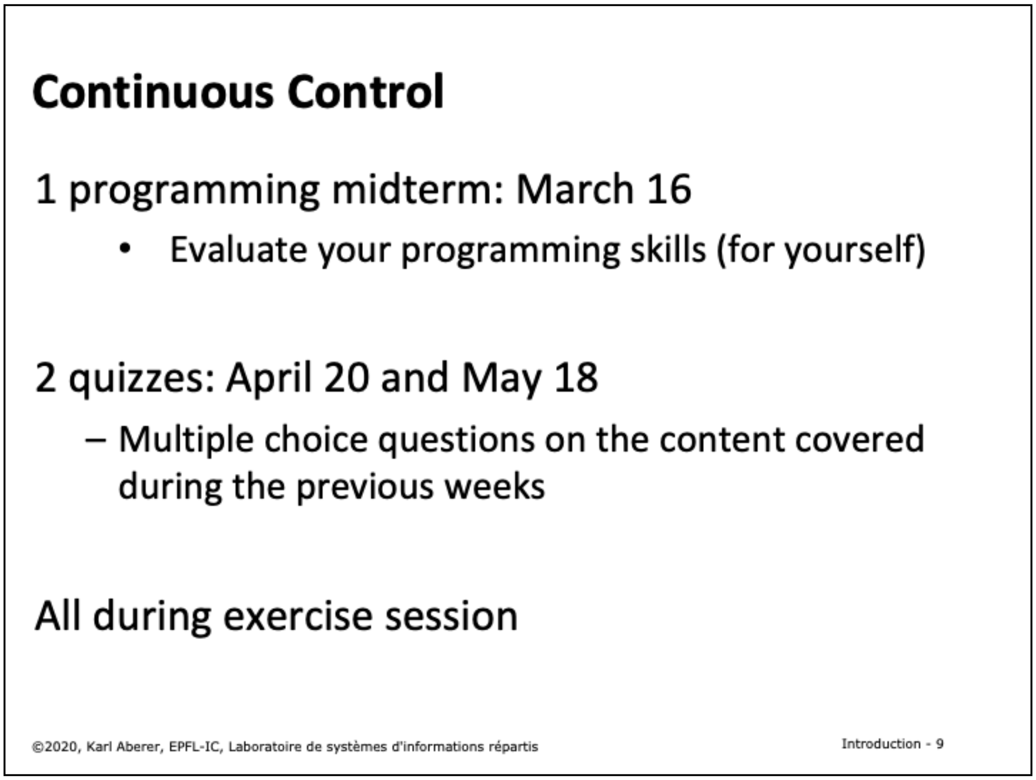
Weekly exercises

- 2-3 problems to solve

Most problems will be (simple) programming exercises

- Uses Python
- Focus on understanding the techniques (not programming skills etc)

Exercises and exam questions from previous years will be made available as well



Continuous Control

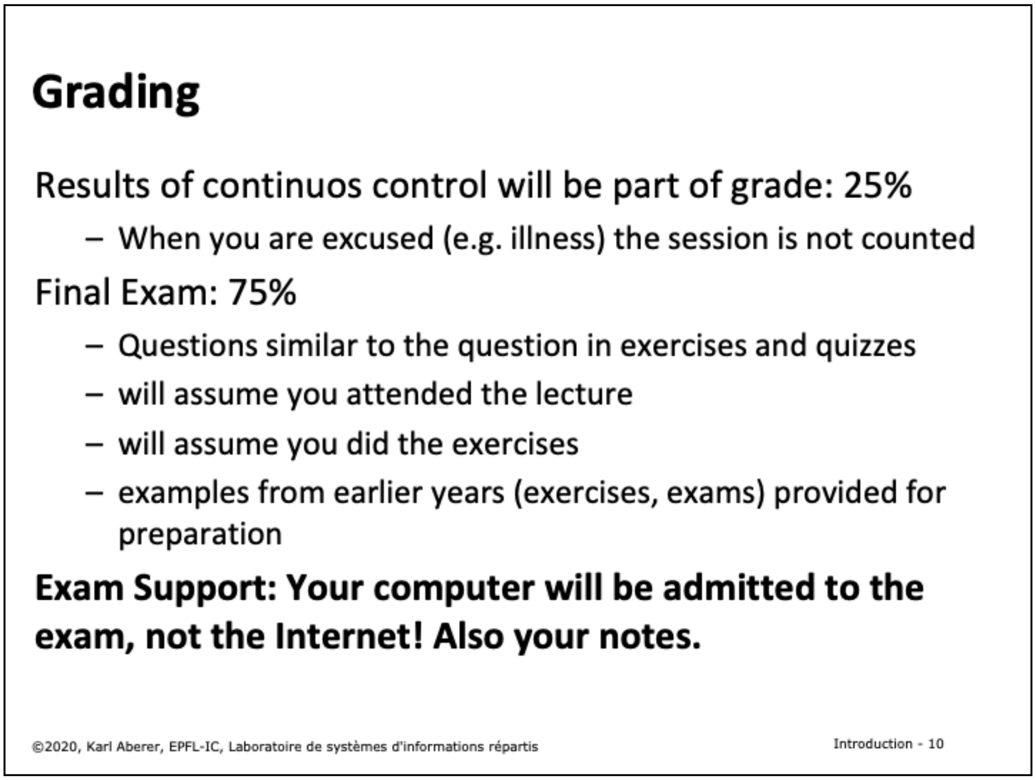
1 programming midterm: March 16

- Evaluate your programming skills (for yourself)

2 quizzes: April 20 and May 18

- Multiple choice questions on the content covered during the previous weeks

All during exercise session



Grading

Results of continuos control will be part of grade: 25%

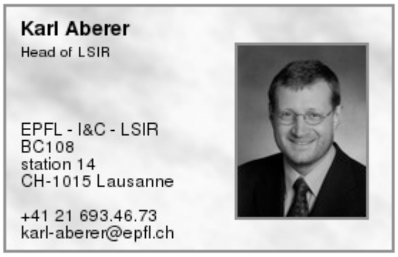
- When you are excused (e.g. illness) the session is not counted

Final Exam: 75%

- Questions similar to the question in exercises and quizzes
- will assume you attended the lecture
- will assume you did the exercises
- examples from earlier years (exercises, exams) provided for preparation

Exam Support: Your computer will be admitted to the exam, not the Internet! Also your notes.

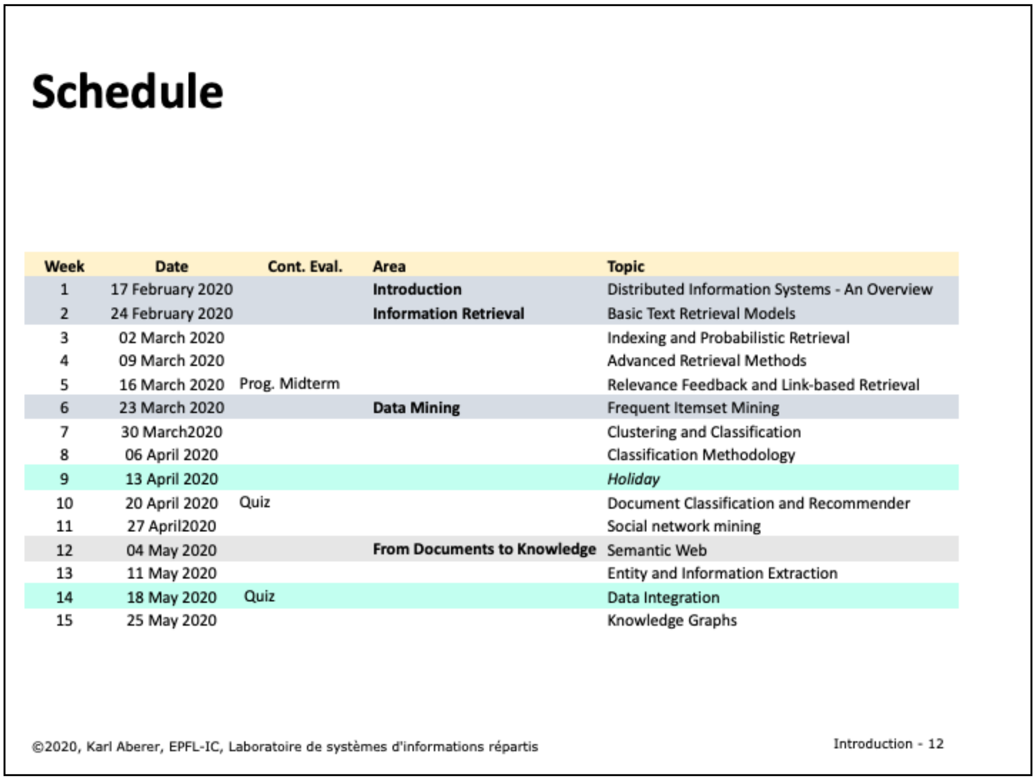
Lecturer



Karl Aberer
Head of LSIR

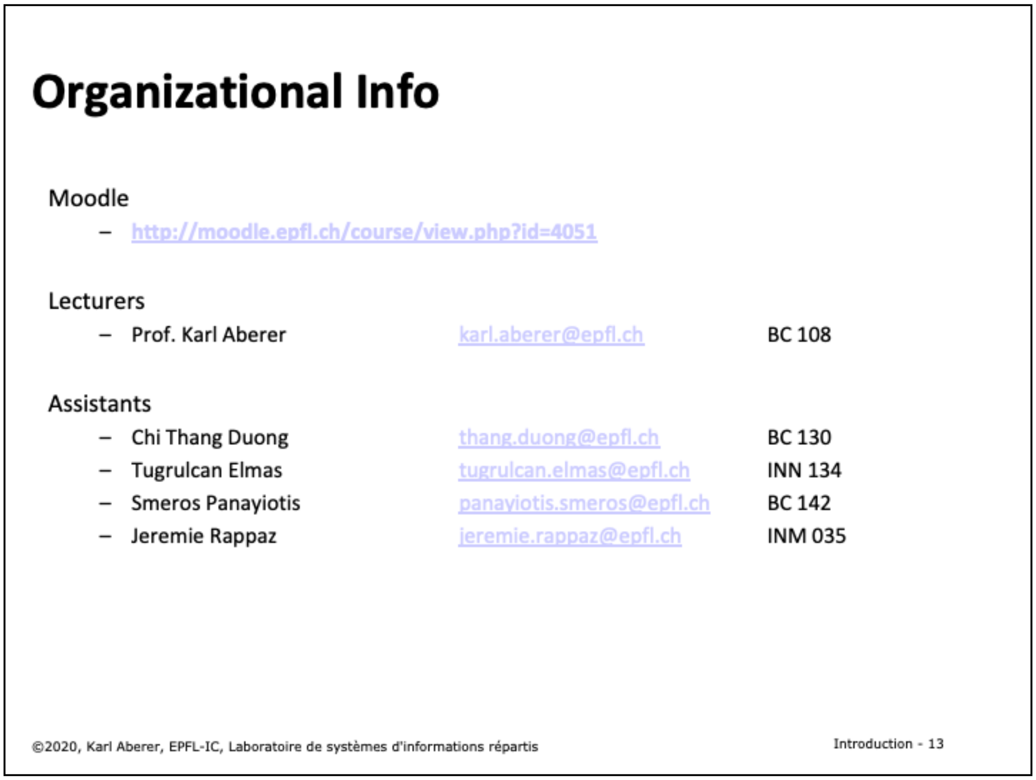
EPFL - I&C - LSIR
BC108
station 14
CH-1015 Lausanne

+41 21 693.46.73
karl-aberer@epfl.ch



Schedule

Week	Date	Cont. Eval.	Area	Topic
1	17 February 2020		**Introduction**	Distributed Information Systems - An Overview
2	24 February 2020		**Information Retrieval**	Basic Text Retrieval Models
3	02 March 2020			Indexing and Probabilistic Retrieval
4	09 March 2020			Advanced Retrieval Methods
5	16 March 2020	Prog. Midterm		Relevance Feedback and Link-based Retrieval
6	23 March 2020		**Data Mining**	Frequent Itemset Mining
7	30 March2020			Clustering and Classification
8	06 April 2020			Classification Methodology
9	13 April 2020			*Holiday*
10	20 April 2020	Quiz		Document Classification and Recommender
11	27 April2020			Social network mining
12	04 May 2020		**From Documents to Knowledge**	Semantic Web
13	11 May 2020			Entity and Information Extraction
14	18 May 2020	Quiz		Data Integration
15	25 May 2020			Knowledge Graphs



Organizational Info

Moodle

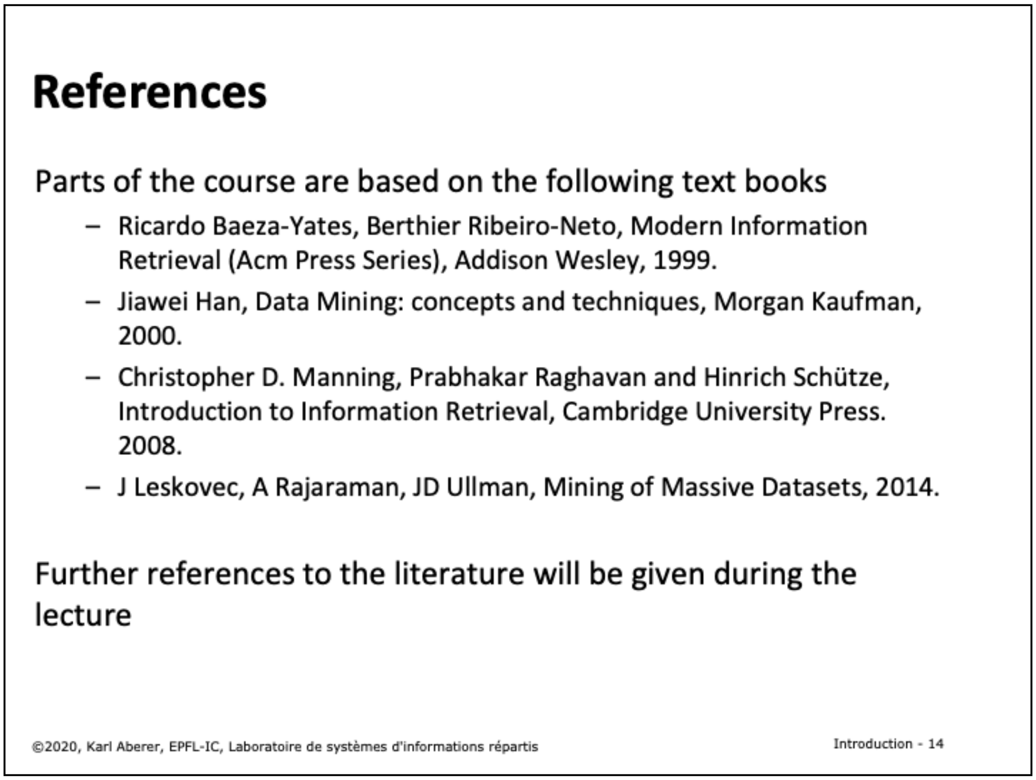
- http://moodle.epfl.ch/course/view.php?id=4051

Lecturers

– Prof. Karl Aberer	karl.aberer@epfl.ch	BC 108

Assistants

– Chi Thang Duong	thang.duong@epfl.ch	BC 130
– Tugrulcan Elmas	tugrulcan.elmas@epfl.ch	INN 134
– Smeros Panayiotis	panayiotis.smeros@epfl.ch	BC 142
– Jeremie Rappaz	jeremie.rappaz@epfl.ch	INM 035

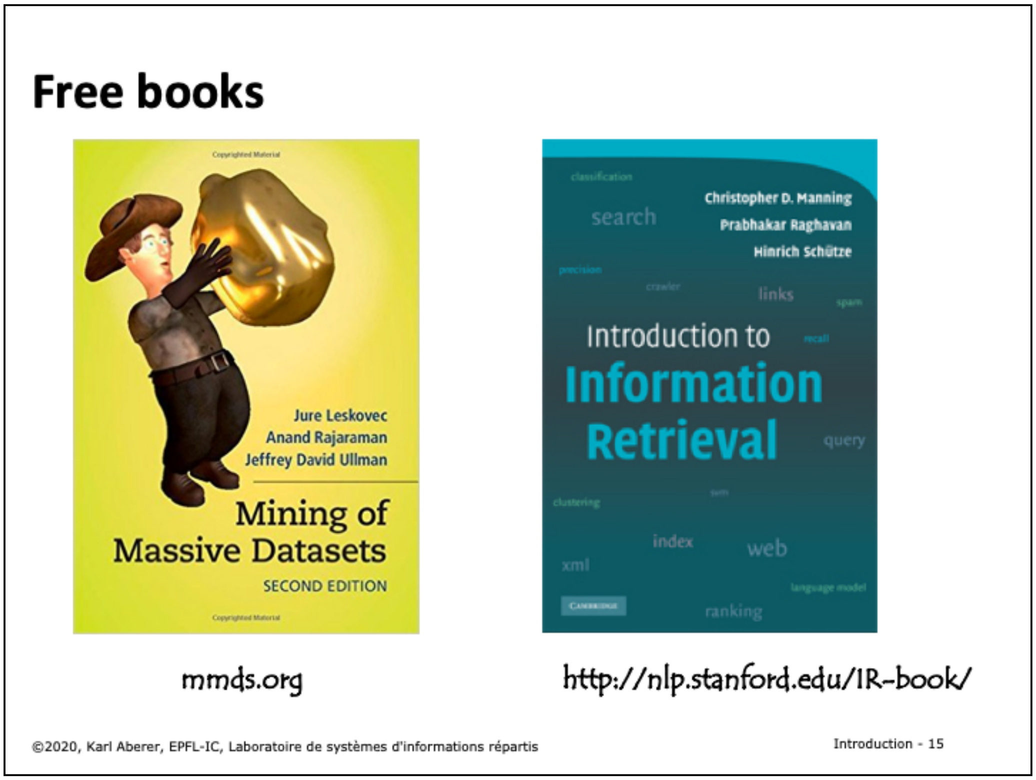


References

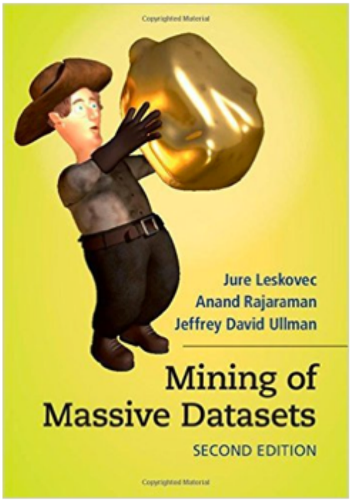
Parts of the course are based on the following text books

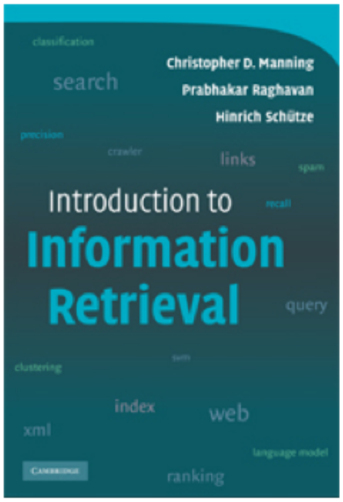
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- J Leskovec, A Rajaraman, JD Ullman, Mining of Massive Datasets, 2014.

Further references to the literature will be given during the lecture



Free books





mmds.org

http://nlp.stanford.edu/IR-book/

Exam Date