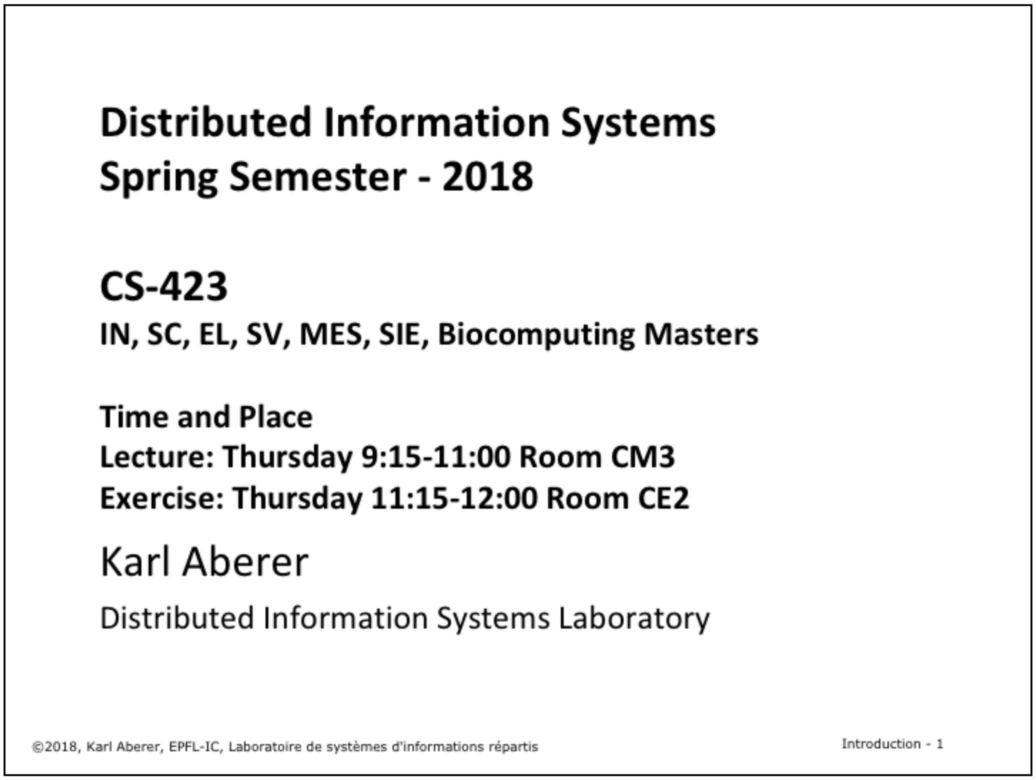


Distributed Information Systems
Spring Semester - 2018

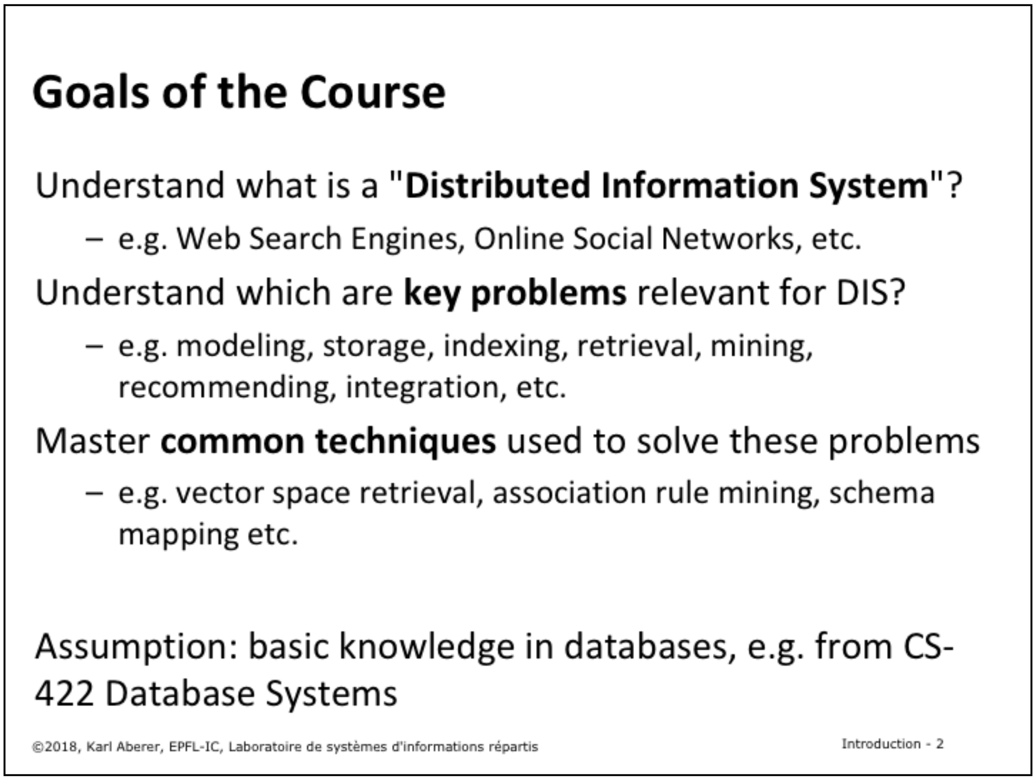
CS-423

IN, SC, EL, SV, MES, SIE, Biocomputing Masters

Time and Place
Lecture: Thursday 9:15-11:00 Room CM3
Exercise: Thursday 11:15-12:00 Room CE2

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.

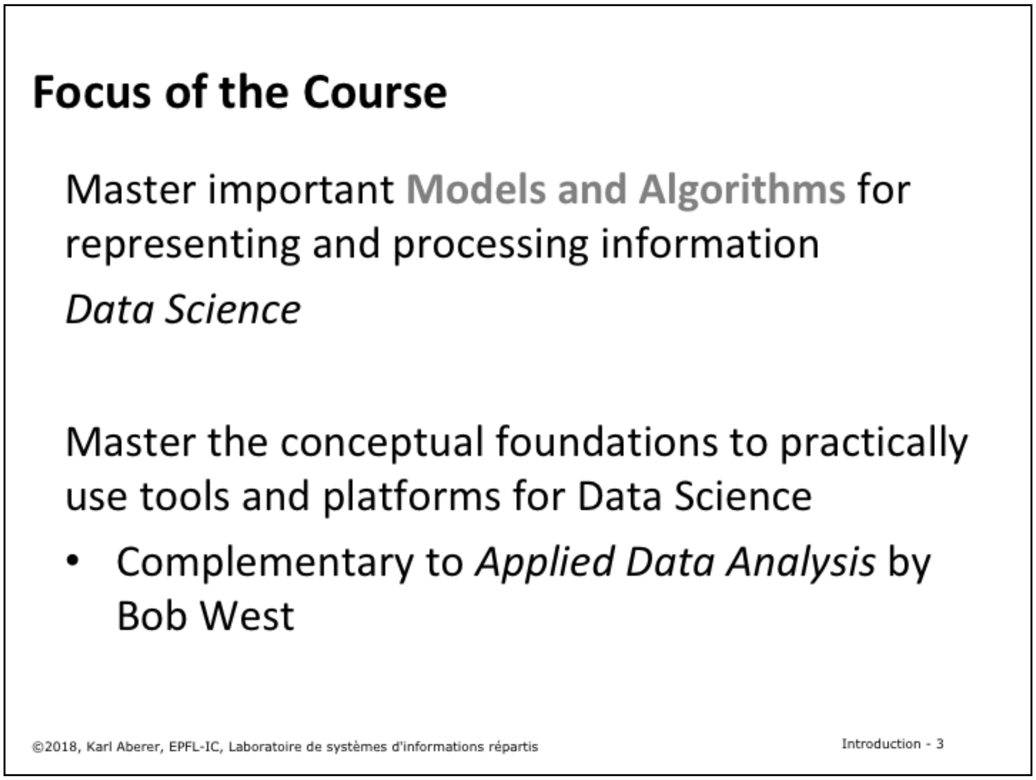
Understand which are **key problems** relevant for DIS?

- 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



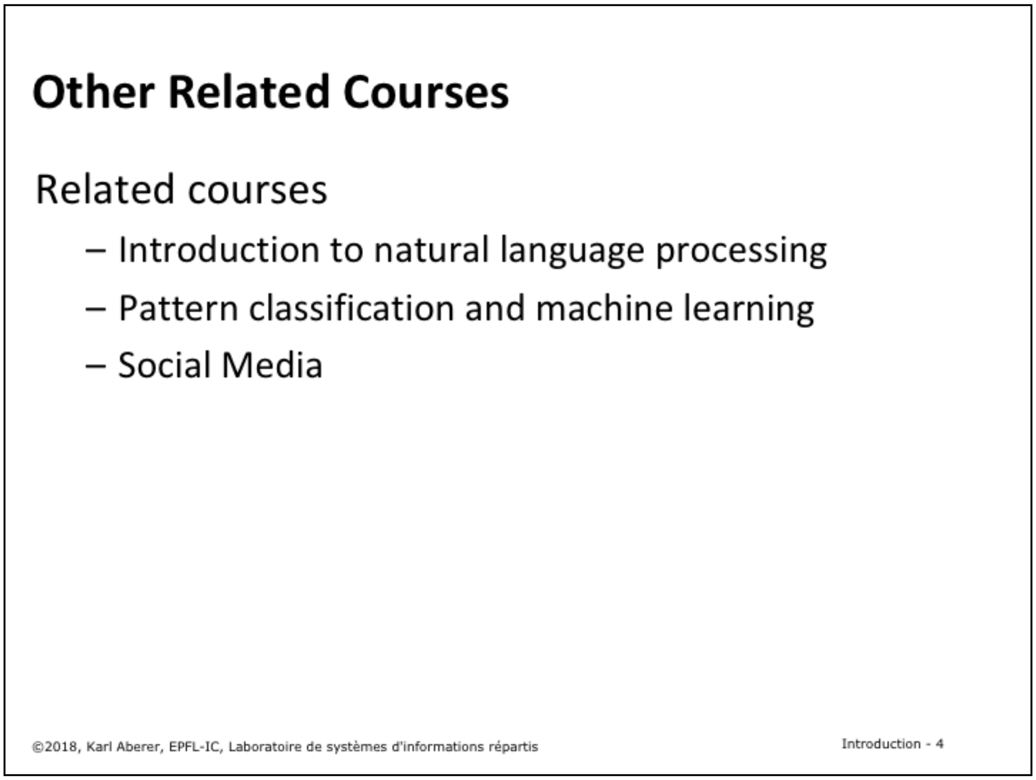
Focus of the Course

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

Data Science

Master the conceptual foundations to practically use tools and platforms for Data Science

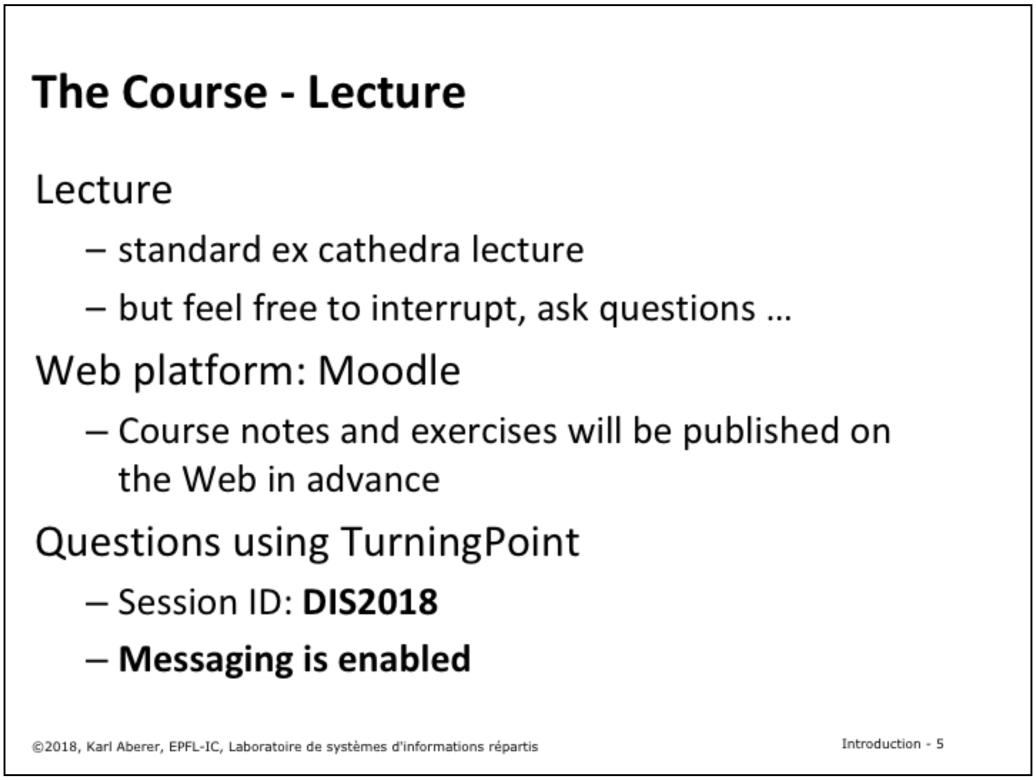
- Complementary to *Applied Data Analysis* by Bob West



Other Related Courses

Related courses

- Introduction to natural language processing
- Pattern classification and machine learning
- Social Media



The Course - Lecture

Lecture

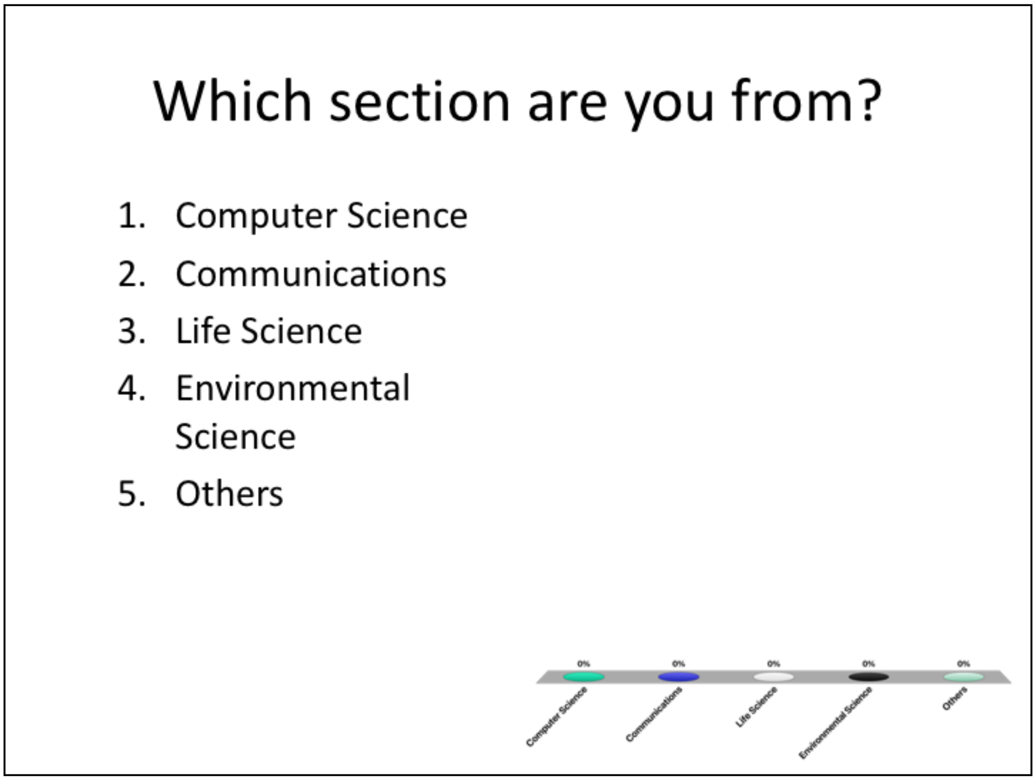
- standard ex cathedra lecture
- but feel free to interrupt, ask questions …

Web platform: Moodle

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

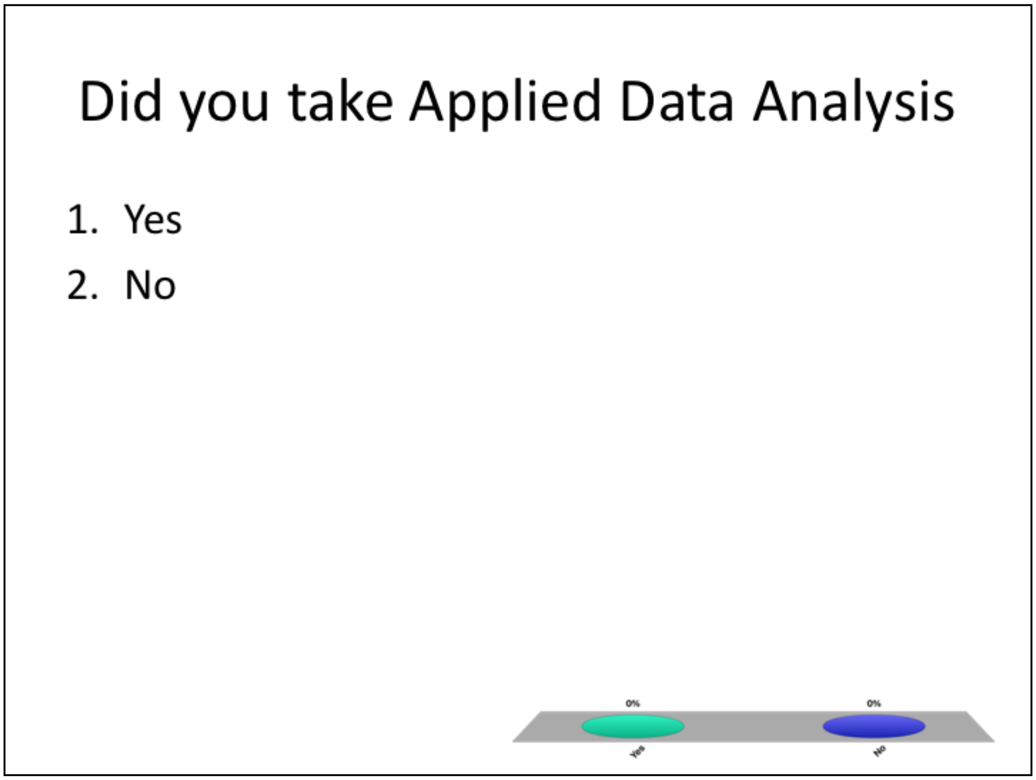
Questions using TurningPoint

- Session ID: **DIS2018**
- **Messaging is enabled**



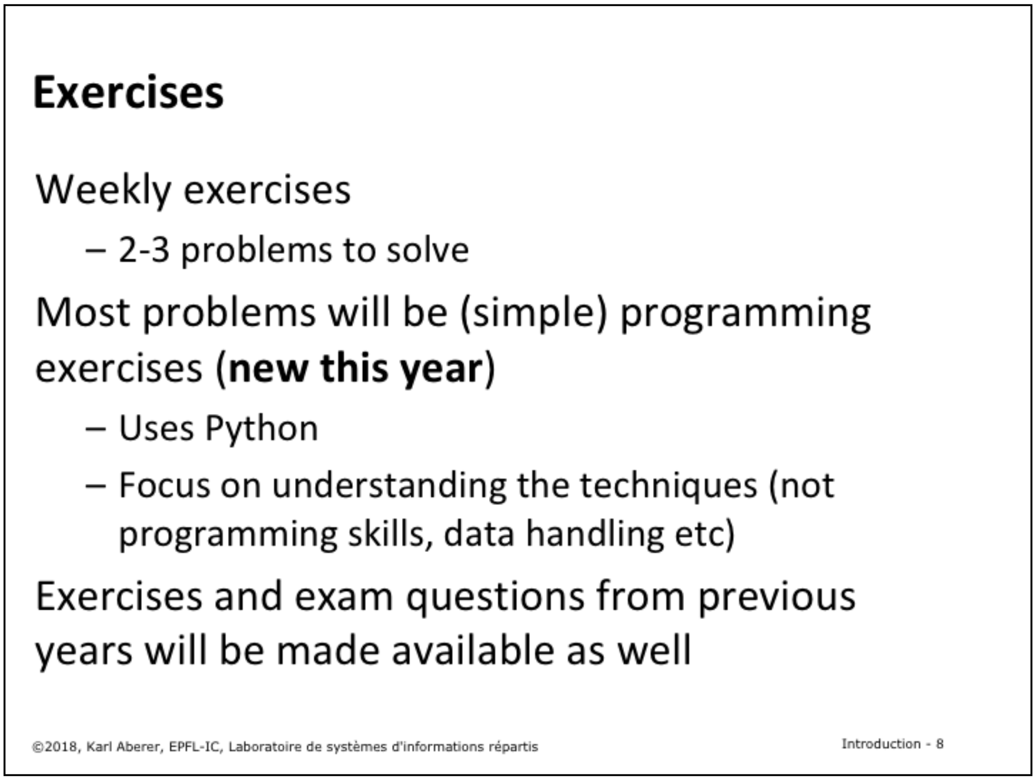
Which section are you from?

1. Computer Science
2. Communications
3. Life Science
4. Environmental Science
5. Others



Did you take Applied Data Analysis

1. Yes
2. No



Exercises

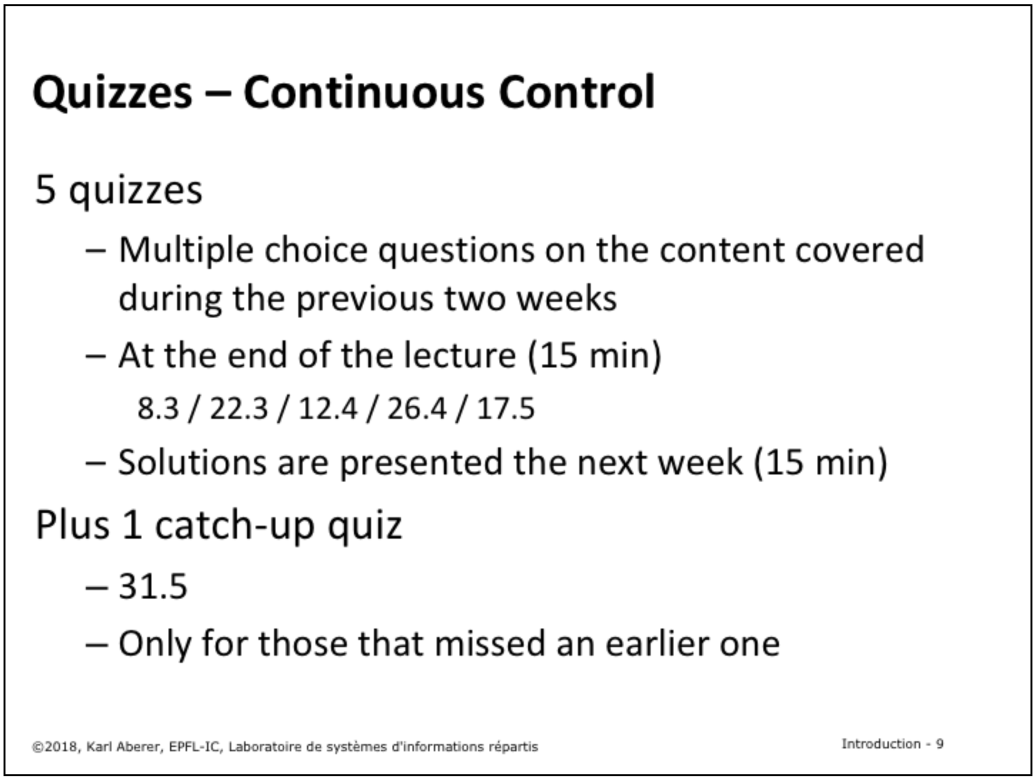
Weekly exercises

- 2-3 problems to solve

Most problems will be (simple) programming exercises (**new this year**)

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

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



Quizzes – Continuous Control

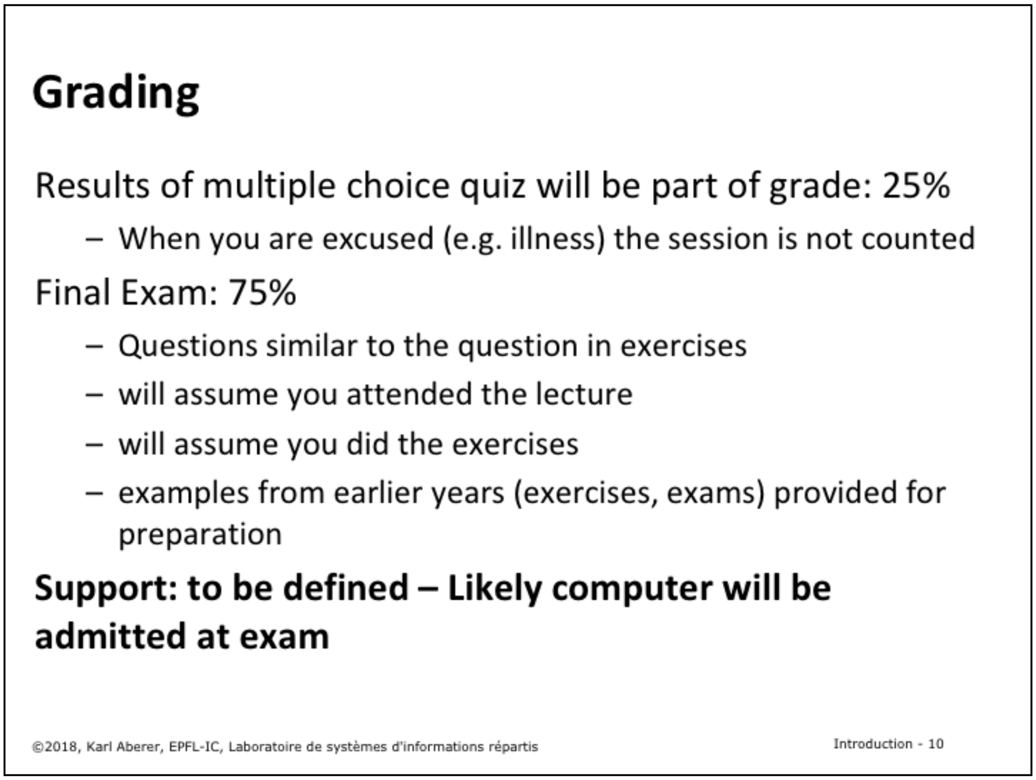
5 quizzes

- Multiple choice questions on the content covered during the previous two weeks
- At the end of the lecture (15 min)

  8.3 / 22.3 / 12.4 / 26.4 / 17.5
- Solutions are presented the next week (15 min)

Plus 1 catch-up quiz

- 31.5
- Only for those that missed an earlier one



Grading

Results of multiple choice quiz will be part of grade: 25%

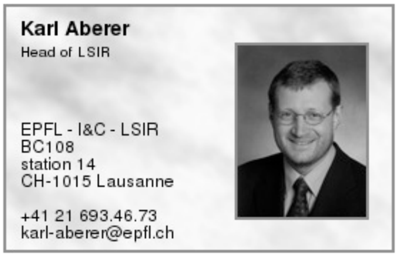
- When you are excused (e.g. illness) the session is not counted

Final Exam: 75%

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

Support: to be defined – Likely computer will be admitted at exam

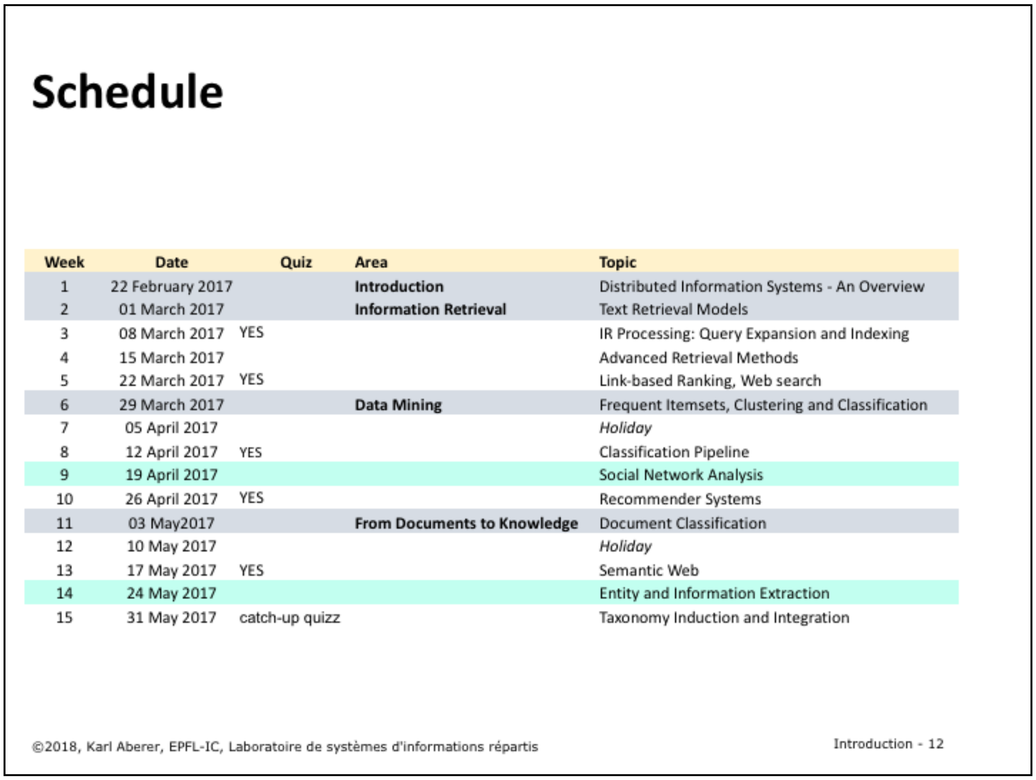
Lecturer



Karl Aberer
Head of LSIR

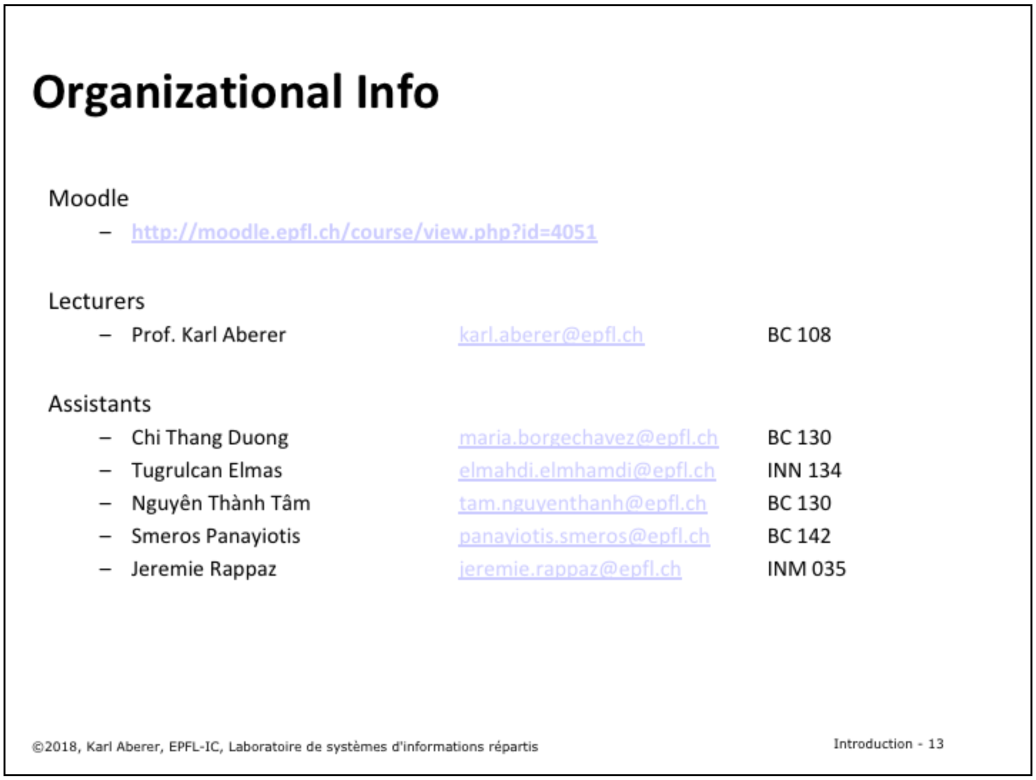
EPFL - I&C - LSIR
BC108
station 14
CH-1015 Lausanne

+41 21 693.46.73
karl-aberer@epfl.ch



Schedule

Week	Date	Quiz	Area	Topic
1	22 February 2017		**Introduction**	Distributed Information Systems - An Overview
2	01 March 2017		**Information Retrieval**	Text Retrieval Models
3	08 March 2017	YES		IR Processing: Query Expansion and Indexing
4	15 March 2017			Advanced Retrieval Methods
5	22 March 2017	YES		Link-based Ranking, Web search
6	29 March 2017		**Data Mining**	Frequent Itemsets, Clustering and Classification
7	05 April 2017			*Holiday*
8	12 April 2017	YES		Classification Pipeline
9	19 April 2017			Social Network Analysis
10	26 April 2017	YES		Recommender Systems
11	03 May2017		**From Documents to Knowledge**	Document Classification
12	10 May 2017			*Holiday*
13	17 May 2017	YES		Semantic Web
14	24 May 2017			Entity and Information Extraction
15	31 May 2017	catch-up quizz		Taxonomy Induction and Integration



Organizational Info

Moodle

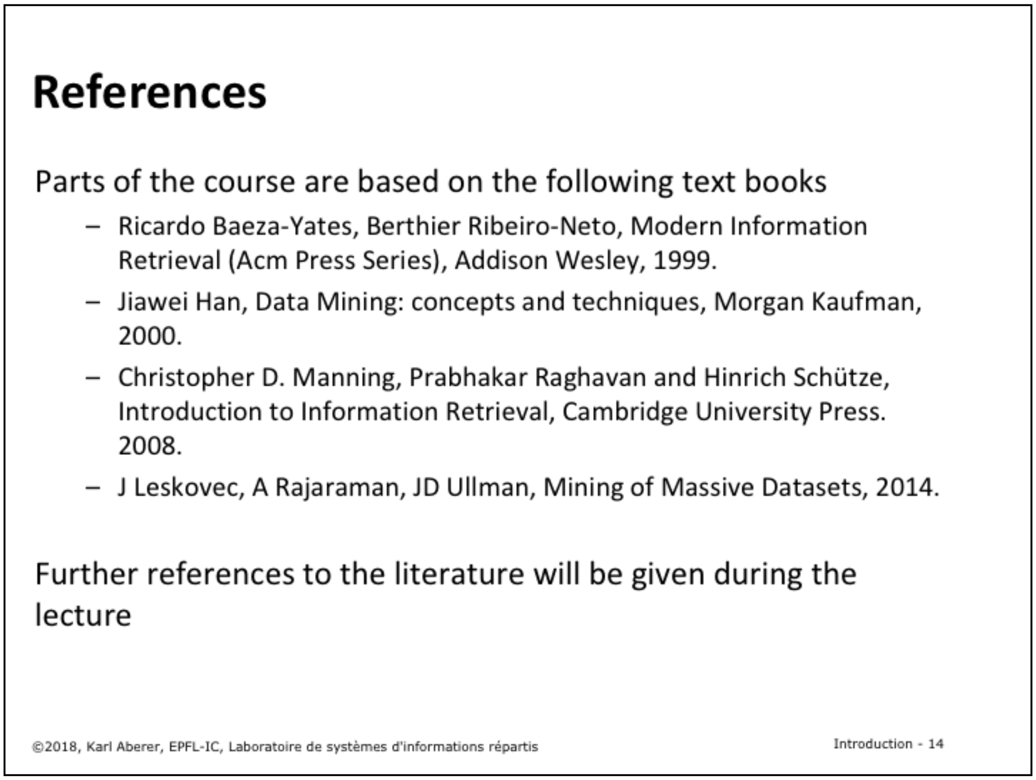
- http://moodle.epfl.ch/course/view.php?id=4051

Lecturers

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

Assistants

– Chi Thang Duong	maria.borgechavez@epfl.ch	BC 130
– Tugrulcan Elmas	elmahdi.elmhamdi@epfl.ch	INN 134
– Nguyên Thành Tâm	tam.nguyenthanh@epfl.ch	BC 130
– Smeros Panayiotis	panayiotis.smeros@epfl.ch	BC 142
– Jeremie Rappaz	jeremie.rappaz@epfl.ch	INM 035



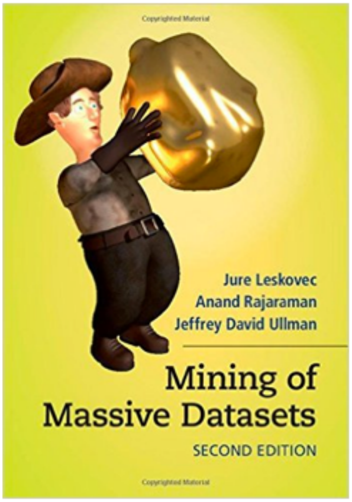
References

Parts of the course are based on the following text books

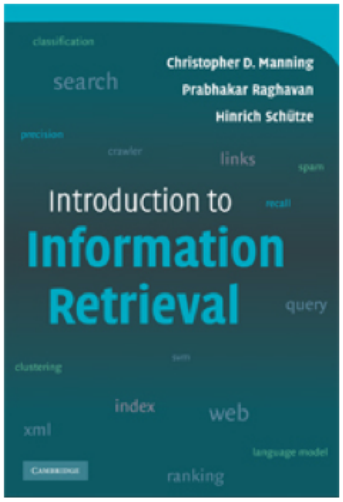
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge 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/