

Distributed Information Systems Spring Semester - 2019

CS-423

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

Time and Place

Lecture: Monday 10:15-12:00 Room INF1

Exercise: Thursday 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

Explain the content of those courses



Career-Changing Courses Start Monday

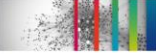
Learn in-demand data science skills and earn credentials in 2019 with the world's best online learning experience.



Advanced Business Analytics
University of Colorado, Boulder

Skills you'll learn: SQL, Decision-Making, Communication, Presentation


[Enroll Now](#) 5 courses



Applied Data Science with Python
University of Michigan

Skills you'll learn: Python Scripting, Machine Learning, Text Mining


[Enroll Now](#) 5 courses



Business Analytics
The Wharton School of the University of Pennsylvania

Skills you'll learn: Demand Forecasting, Big Data Analytics, Regression Analysis


[Enroll Now](#) 5 courses



Advanced Machine Learning
National Research University Higher School of Economics

Skills you'll learn: Linear Modeling, Bayesian Methods, Deep Learning


[Enroll Now](#) 7 courses



Big Data for Data Engineers
Yandex

Skills you'll learn: Predictive Modeling, MapReduce, Spark, Hive, NoSQL, databases


[Enroll Now](#) 5 courses



Business Statistics and Analysis
Rice University

Skills you'll learn: Excel Functions, Statistical Distribution, Hypothesis Testing


[Enroll Now](#) 5 courses



Data Mining
University of Illinois at Urbana-Champaign

Skills you'll learn: Pattern Discovery, Text Mining, Cluster Analysis


[Enroll Now](#) 6 courses



Probabilistic Graphical Models
Stanford University

Skills you'll learn: Bayesian Networks, Probabilistic Inference, Parameter Estimating


[Enroll Now](#) 8 courses



Statistics with R
Duke University

Skills you'll learn: Statistical Inference, Variance Analysis, Model Selection


[Enroll Now](#) 5 courses



Deep Learning
deeplearning.ai

Skills you'll learn: Neural Network, Optimization, Tensorflow, Error Analysis


[Enroll Now](#) 5 courses



Recommender Systems
University of Minnesota

Skills you'll learn: Collaborative Filtering, Personalized Ranking, Matrix Factorization

[Enroll Now](#) 5 courses



Survey Data Collection and Analytics
University of Michigan and University of Maryland, College Park

Skills you'll learn: Audience Research, Questionnaire Design, Weighting Data

[Enroll Now](#) 7 courses

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

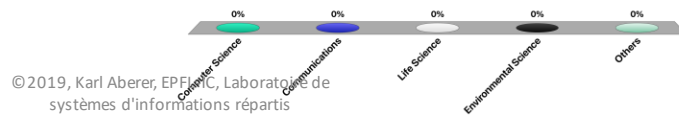
- Course notes and exercises will be published on the Web in advance

Questions using TurningPoint

- Session ID: **DIS2019**
- **Messaging is enabled**

Which section are you from?

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



© 2019, Karl Aberer, EPFL, Laboratoire de systèmes d'informations répartis

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 18

- Evaluate your programming skills (for yourself)

2 quizzes: April 22 and May 20

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

All during exercise session

Grading

Results of continuous 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



Schedule

Week	Date	Cont. Eval.	Area	Topic
1	18 February 2019		Introduction	Distributed Information Systems - An Overview
2	25 February 2019		Information Retrieval	Basic Text Retrieval Models
3	04 March 2019			Indexing and Probabilistic Retrieval
4	11 March 2019			Advanced Retrieval Methods
5	18 March 2019	Prog. Midterm		Relevance Feedback and Link-based Retrieval
6	27 March 2019		Data Mining	Frequent Itemset Mining
7	01 April 2019			Clustering and Classification
8	08 April 2019			Classification Methodology
9	15 April 2019	Quiz		Document Classification and Recommender
10	22 April 2019			<i>Holiday</i>
11	29 May 2019			Social network mining
12	06 May 2019		From Documents to Knowledge	Semantic Web
13	13 May 2019			Entity and Information Extraction
14	20 May 2019	Quiz		Data Integration
15	27 May 2019			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
- Nguyễn Thành Tâm ta.m.nguyenthanh@epfl.ch BC 130
- Smeros Panayiotis panayiotis.smeros@epfl.ch BC 142
- Jeremie Rappaz jeremie.rappaz@epfl.ch INM 035

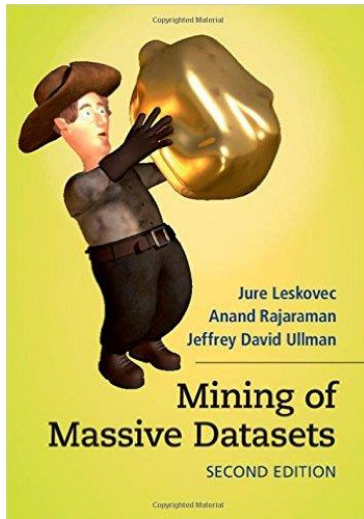
References

Parts of the course are based on the following text books

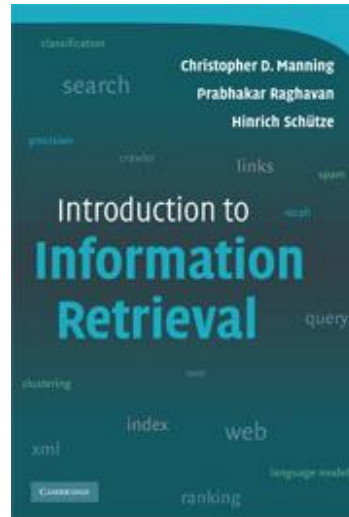
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (Acm Press Series), Addison Wesley, 1999.
- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge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

Exam Date