

COM-308

**Internet analytics**

Grossglauser Matthias

| <b>Cursus</b>               | <b>Sem.</b> | <b>Type</b> |
|-----------------------------|-------------|-------------|
| Communication systems minor | E           | Opt.        |
| Communication systems       | BA6         | Opt.        |
| Computer science            | BA6         | Opt.        |

|                            |                 |
|----------------------------|-----------------|
| Language of teaching       | English         |
| Credits                    | 5               |
| Session                    | Summer          |
| Semester                   | Spring          |
| Exam                       | Written         |
| Workload                   | 150h            |
| Weeks                      | 14              |
| <b>Hours</b>               | <b>5 weekly</b> |
| Courses                    | 2 weekly        |
| Exercises                  | 1 weekly        |
| Project                    | 2 weekly        |
| <b>Number of positions</b> |                 |

**Summary**

Internet analytics is the collection, modeling, and analysis of user data in large-scale online services, such as social networking, e-commerce, search, and advertisement. This class explores a number of the key functions of such online services that have become ubiquitous over the past decade.

**Content**

The class seeks a balance between foundational but relatively basic material in algorithms, statistics, graph theory and related fields, with real-world applications inspired by the current practice of internet and cloud services.

Specifically, we look at social & information networks, recommender systems, clustering and community detection, search/retrieval/topic models, dimensionality reduction, stream computing, and online ad auctions. Together, these provide a good coverage of the main uses for data mining and analytics applications in social networking, e-commerce, social media, etc.

The course is combination of theoretical materials and weekly laboratory sessions, where we explore several large-scale datasets from the real world. For this, you will work with a dedicated infrastructure based on Hadoop & Apache Spark.

**Keywords**

data mining; machine learning; social networking; map-reduce; hadoop; recommender systems; clustering; community detection; topic models; information retrieval; stream computing; ad auctions

**Learning Prerequisites****Required courses**

Stochastic models in communication (COM-300)

**Recommended courses**

Basic linear algebra  
Algorithms & data structures

**Important concepts to start the course**

Graphs; linear algebra; Markov chains; Java

**Learning Outcomes**

By the end of the course, the student must be able to:

- Explore real-world data from online services
- Develop frameworks and models for typical data mining problems in online services
- Analyze the efficiency and effectiveness of these models
- data-mining and machine learning techniques to concrete real-world problems

### Teaching methods

Ex cathedra + homeworks + lab sessions

### Expected student activities

Lectures with associated homeworks explore the basic models and fundamental concepts. The labs are designed to explore very practical questions based on a number of large-scale real-world datasets we have curated for the class. The labs draw on knowledge acquired in the lectures, but are hands-on and self-contained.

### Assessment methods

Project 20%, midterm 30%, final exam 50%

### Resources

#### Bibliography

- C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
- A. Rajaraman, J. D. Ullman: Mining of Massive Datasets, 2012
- M. Chiang: Networked Life, Cambridge, Cambridge, 2012
- D. Easley, J. Kleinberg: Networks, Crowds, and Markets, Cambridge, 2010
- Ch. D. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval, Cambridge, 2008
- M.E.J. Newman: Networks: An Introduction, Oxford, 2010

#### Websites

- <http://icawww1.epfl.ch/ix/>