Databases & Data Mining

Erwin M. Bakker & Stefan Manegold

e.m.bakker
s.manegold
@liacs.leidenuniv.nl

https://homepages.cwi.nl/~manegold/DBDM/
http://liacs.leidenuniv.nl/~bakkerem2/dbdm/
DBDM: “Registration”

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

– Your full name
– Your email address
– Your student ID
– Your affiliation (university)
– Your program / subject

By Sunday 16 September 2018, 23:59 CEST.
DBDM: Overview

Period: September 11th - December 4th 2018 (Tuesdays)
Place: Room 312 (LIACS, Snellius building, Niels Bohrweg 1, 2333 CA Leiden)
Time: 15.30 - 17.15
ECTS: 6

Description:
The course Databases & Data Mining consists of a series of lectures in which advanced
database and data mining techniques will be discussed, with applications to
bioinformatics.

Grading:
There will be 2 database and 2 data mining assignments, i.e., 4 assignments in total,
and a final exam (open book). The final grade will be based on a weighted average of
the grades obtained for assignments P1, P2, P3, P4 and the Exam (E >5):
Final Grade = (0.5*P1 + P2 + 0.5*P3 + P4 + 3*E)/6.

## DBDM: (tentative) Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Room</th>
<th>Subject (tentative)</th>
<th>Topic &amp; Lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-09</td>
<td>312</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>18-09</td>
<td>312</td>
<td>Database Technology</td>
<td>Databases and Data Management for Data Mining</td>
</tr>
<tr>
<td>25-09</td>
<td>312</td>
<td>Database Technology</td>
<td>Stefan Manegold</td>
</tr>
<tr>
<td>02-10</td>
<td>312</td>
<td>Data Preprocessing</td>
<td></td>
</tr>
<tr>
<td>09-10</td>
<td>312</td>
<td>No class</td>
<td></td>
</tr>
<tr>
<td>16-10</td>
<td>312</td>
<td>Data Warehousing and OLAP</td>
<td></td>
</tr>
<tr>
<td>23-10</td>
<td>312</td>
<td>Data Cube Technology</td>
<td></td>
</tr>
<tr>
<td>30-10</td>
<td>312</td>
<td>Basic Data Mining Algorithms I</td>
<td>Data Mining Techniques and Applications</td>
</tr>
<tr>
<td>06-11</td>
<td>312</td>
<td>Basic Data Mining Algorithms II</td>
<td>Erwin Bakker</td>
</tr>
<tr>
<td>13-11</td>
<td>312</td>
<td>Advanced Data Mining Algorithms</td>
<td></td>
</tr>
<tr>
<td>20-11</td>
<td>312</td>
<td>Mining in Bio-Data</td>
<td></td>
</tr>
<tr>
<td>27-11</td>
<td>312</td>
<td>Graph Mining I</td>
<td></td>
</tr>
<tr>
<td>04-12</td>
<td>312</td>
<td>Graph Mining II</td>
<td></td>
</tr>
</tbody>
</table>

DBDM: Assignments

• 2 database assignments & 2 data mining assignments

• Will be announced individually during lectures and posted on website
DBDM: Exam

- **open book exam**: you can take with you your book, and printed course notes (slides). *No electronic equipment is allowed, though.*

- **Materials to be studied:**
  - All content covered and discussed during lectures (slides will be shared).
  - More to be announced.

- **Date**: *Monday, January 7, 2019*
- **Time**: *14:00 - 17:00*
- **Place**: *Room F104, Van Steenisgebouw, Einsteinweg 2, 2333 CC Leiden*

https://homepages.cwi.nl/~manegold/DBDM/  
http://liacs.leidenuniv.nl/~bakkerem2/dbdm/
DBDM: Recommended Books

• Data Mining:

• Database systems (e.g.):

DBDM: “Registration”

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

– Your full name
– Your email address
– Your student ID
– Your affiliation (university)
– Your program / subject

By Sunday 16 September 2018, 23:59 CEST.
Databases & Data Mining

Stefan Manegold

Group leader Database Architectures
Centrum Wiskunde & Informatica (CWI)
Amsterdam
http://homepages.cwi.nl/~manegold/

http://www.monetdb.org/

Prof. Data Management (0.2 fte)
LIACS & LCDS
Faculty of Science, Leiden University

Data

Data Management

Database

Data Mining
The age of Big Data

- 1500TB/min = 1000 full drives per minute
- = a stack of 20 meter high

- 4000 million TeraBytes = 3 billion full disk drives

And Future Growth is Staggering

- By 2017, mobile traffic will have grown 13X in just 5 years
- In 2017, there will be 3X more connected devices than people on Earth
- All digital data created reached 4 zettabytes in 2013
The Data Economy
Disruptions by the Data Economy

UBER

EVERYONE’S PRIVATE DRIVER™
DBDM: Selected Challenges

GIS (LIDAR):
Massive point clouds: 640 Billion (x,y,z) points / 15 TB
=> spatial joins between point cloud and polygons

Logistics:
> 5 trillion ($10^{12}$) GPS points (grows with >60k points/sec)

Seismology:
~ 4 M files, ~ 500 GB (10x compressed)
=> Transparent data ingestion: Data Vault

Remote sensing:
~2 PB satellite image data
=> Array data processing: SciQL

Astronomy:
Raw data: 25 TB / hour; derived data: 100 TB / year
=> Transient detection inside DBMS
Earth Observation:
DLR Satellite Data Acquisition [in TeraByte]

~2000 TB
Low Frequency Array for Radio Astronomy

International LOFAR Telescope (ILT)

Raw data: ~25 TB / hour
Derived data: ~100 TB / year
Data Disrupting Science: Paradigm Shift in Scientific Research

1st observing empirical

2nd modeling theoretical

3rd simulating computational

collecting & analyzing data
data exploration (eScience)

Jim Gray (1944 - 2007)
Data Management & Data Mining