TADA
practicalities & more
on DM

24 April 2014
More on Data Mining as a Science
DM as method development

• Data mining develops methods for scientists
  • C.f. mathematics or statistics
• The research of DM in universities doesn’t follow the scientific paradigm
  • But that doesn’t make it a voodoo science
• ...the applications of DM are another story
Of DM, ML, and Stat

• One trichotomy:
  • Statistics studies how reliable inferences can be drawn from imperfect data
  • ML develops technology of automated induction
  • DM is the art of extracting useful patterns from large bodies of data

Data Mining success stories
Bioinformatics

- BioGraph provides automated inference of functional hypotheses
  - E.g. which genes are most potential to be associated with certain diseases

Making money

• “Recommended for you”
• “Others often bought also”
• All of modern targeted advertisement is based on some type of data mining
Obama’s re-election

- Data of electorate was used to target the campaigning efforts where they count
- DM was also used to optimize fund-raising from small donations
Church uses Big Data

- Evangelical Lutheran Church of Finland uses data mining to study its parishes
- What type of people live in which geographical areas?

http://www.hs.fi/talous/iso+data+auttaa+pappia+saarnassa/a1397539201451
Space program safety

- ORCA searches outliers from sensor readings by comparing parameter-value vectors to their neighbors.
- IMS builds a model of normal variance of sensor readings to detect anomalies.
More on IMS

- In early January 2007, ISS Early External Thermal Control System developed an ammonia gas bubble
- Bubble noted by ISS controllers only ~9 hours before it “burst” and dissipated back into liquid

Ashok N. Srivastava: Data Mining at NASA: from Theory to Applications, KDD 2009
Practicalities
<table>
<thead>
<tr>
<th>Month</th>
<th>Day</th>
<th>Lecture topic</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>April</td>
<td>17</td>
<td>Intro</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Practicalities &amp; where DM is used</td>
<td>1st assignment given out</td>
</tr>
<tr>
<td>May</td>
<td>1</td>
<td>No lecture (First of May)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Intro to Tensors</td>
<td>1st assignment DL, 2nd assignment given out</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>Tensors in DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>Special topics in tensors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>No lecture (Ascension day)</td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>5</td>
<td>MDL for pattern mining</td>
<td>2nd assignment DL, 3rd assignment given out</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Maximum entropy &amp; iterative data mining</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>No lecture (Corpus Christi)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>Kolmogorov complexity, cumulative entropy, and causality</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>3</td>
<td>Graphs I</td>
<td>3rd assignment DL, 4th assignment given out</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Graphs II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>Graphs III</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Wrap-up</td>
<td>4th assignment DL</td>
</tr>
</tbody>
</table>
On Exam

• Day and place TBA
  • Most likely in early September
• Type TBA
• Final grade is based on the final exam and the assignments
  • Assignments also determine the eligibility to sit the final exam
On assignments: general

- 4 assignments
- Grading: fail, pass, excellent
- You can fail one assignment
  - 2 fails $\Rightarrow$ course failed
- Every excellent gives $1/3$ point improvement on the final exam grade
  - But maximum of 1 full point (3 ex’s)

- You must pass the final exam to pass the course
On assignments: requirements

• Assignments are to be written in proper academic-style English

• **Proper citations**

• You are given sources, but you can also use outside sources
  • Naturally must be mentioned
  • Plagiarism ⇒ failed assignment
On assignments: format

• Assignments need to be returned as PDF files by email
  • No .doc(x), .odt, .rtf, .txt, .xml, .html, .pages, .ps, .wp, or anything else
• No length limits — use the space you need
  • Probably most will need 3–4 pages...
• All PDFs must have name, matriculation number, email address, and clearly state the topic
On assignments: returning

• The assignments are returned by email to tadal4@mpi-inf.mpg.de

• DL is 1600 hours on the stated day
  • No delays, no excuses, time based on the mail time stamp

• We’ll acknowledge the submission that we receive before the lecture on the DL day
On assignments: grading

- Assignments are **not** for repeating what the papers say
  - We’ve read the papers already
- We expect you to discuss **and criticize** the sources, build connections, point out differences, provide new insights, etc.
- Some assignments are marked *hard*
  - This is taken into account when grading
First assignments

1. Did Tukey invent Data Mining?

2. (Don’t) Believe the Hype

3. Big Data: The Best Thing since Sliced Bread or just Another Bottle of Snake Oil?

4. Where did the Candidates Go? (Hard)

http://resources.mpi-inf.mpg.de/d5/teaching/ss14/tada/assignments/1.html