Predictive Analytics in Higher Education

Hans Pongratz
Senior Vice President & Chief Information Officer
pongatz@tum.de

EUNIS BITF Conference, March 7th 2014, Paris
13 Faculties
~ 400 Buildings
154 Studiengänge
~ 36 000 Students 33% female Students
20% int. Students
~ 500 Professors
~ 10 000 Staff members
13 Nobel Prizes
15 Leibniz-Prizes
4 Humboldt Professors
#53 2013 Academic Ranking of World Universities
IT-Strategy: The Digital University

Leitmotif since 2002

Completed IT-Projects:
SAP@TUM
IntegraTUM
elecTUM
mediaTUM
Data Warehouse
Corporate Design
CM@TUM

7.3.2014, Hans Pongratz
Data Warehouse: BW@TUM

Data collection

Loading process

Other operational data sources

Campus Manag., SAP R/3, Diapers,...

External sources e.g. state

Data management

Extraction
Transformation
Loading

TUM Data Warehouse

Data provision

Data provisioning

OLAP

Metadata

Monitoring and Administration tools

Queries/Reports

Data Mining

TUM Data Warehouse

Analyzes

Analyses

Data Warehouse

Metadata

Other operational data sources

Campus Manag., SAP R/3, Diapers,...

External sources e.g. state

Loading process

Extraction
Transformation
Loading

BW@TUM, H. Vogg
## Data sources at TUM

<table>
<thead>
<tr>
<th>Domain</th>
<th>InfoProvider</th>
<th>Operatives System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>Personnel Administration</td>
<td>SAP R/3 HR</td>
</tr>
<tr>
<td></td>
<td>Organization Management</td>
<td>SAP R/3 HR</td>
</tr>
<tr>
<td>Student &amp; Exam Data</td>
<td>Students</td>
<td>TUMonline</td>
</tr>
<tr>
<td></td>
<td>Tests / Exams</td>
<td>TUMonline</td>
</tr>
<tr>
<td></td>
<td>Applicants</td>
<td>TUMonline</td>
</tr>
<tr>
<td>Accounting</td>
<td>Financial Planing</td>
<td>SAP R/3 FI</td>
</tr>
<tr>
<td></td>
<td>Funds Management</td>
<td>SAP R/3 FM</td>
</tr>
<tr>
<td></td>
<td>Controlling</td>
<td>SAP R/3 CO</td>
</tr>
<tr>
<td>Integrationdomain</td>
<td>various</td>
<td>various</td>
</tr>
</tbody>
</table>
2008: Bachelor admission and enrolment statistic
Self-service rankings for students per test

Grade distribution: Percent % / grade
K. = Number of candidates

Exam date: WZ0152 13S 4SWS L Technology and Utilization of Non-Wood Biogenic Resources on 02.10.2013 for Schieder

Registered: 88
Attempted: 71
Not present: 17
Withdrawal with approved reasons: 3
Not valid/cheating: 0
Rejection: 0
Percent. of exams assessed as failed: 26.761%
Average total: 3.516
Average (assessed as passed): 3.12
How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target data scientists Pole and his team for Target about how they did it all.

"[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole’s colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date."
New idea? No, e.g. former corner shops

What Happens in an Internet Minute?

639,800 GB of global IP data transferred
20 New victims of identity theft
204 million Emails sent
47,000 App downloads
$83,000 In sales
61,141 Hours of music
20 million Photo views
3,000 Photo uploads
320+ New Twitter accounts
100,000 New tweets
6 million Facebook views
2+ million Search queries
277,000 Logins
30 Hours of video uploaded
1.3 million Video views
1,300 New mobile users
100+ New LinkedIn accounts
135 Botnet infections
6 New Wikipedia articles published

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second

Academic & Learning Analytics

- Business Analytics for HEIs
- Goals, e.g. prediction of room allocation, study success, enrollment, ...

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 401</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>201-400</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51-200</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>10-50</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>2-9</td>
<td>90</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>4</td>
</tr>
</tbody>
</table>

7.3.2014, Hans Pongratz
First experiences. Modul assignments (1/2)

7.3.2014, Hans Pongratz

Source: A. Baumann, TUM
First experiences, Modul assignments (2/2)

Statistics

Module Attendance by Subjecttype

<table>
<thead>
<tr>
<th>Subject Type</th>
<th>Number of</th>
<th>MIN (Attendance)</th>
<th>AVG (Attendance)</th>
<th>MAX (Attendance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>60</td>
<td>1</td>
<td>16</td>
<td>302</td>
</tr>
<tr>
<td>Compulsory subjects</td>
<td>12</td>
<td>18</td>
<td>325</td>
<td>518</td>
</tr>
<tr>
<td>Elective subjects</td>
<td>73</td>
<td>1</td>
<td>63</td>
<td>347</td>
</tr>
</tbody>
</table>

Attendance Percentage

- 0% - 25%: 72 students
- 25% - 50%: 36 students
- 50% - 75%: 48 students
- 75% - 100%: 18 students

Source: A. Baumann, TUM
Frist experiences, Elective Modules

7.3.2014, Hans Pongratz

Source: A. Baumann, TUM
Conceptual Framework
Proposal by EDUCAUSE

## EDUCAUSE: Proposed Definitions

<table>
<thead>
<tr>
<th>Term</th>
<th>Proposed Definition</th>
<th>Proposed Level of Focus</th>
<th>Sample Projects (see below for links)</th>
</tr>
</thead>
</table>
| Analytics          | An overarching concept that is defined as data-driven decision making (from Ravishanker). | All levels              | • M-Reports Dashboard  
• Learning and Career Outcomes                                               |
| Academic Analytics | A process for providing higher education institutions with the data necessary to support operational and financial decision making (adapted from Goldstein and Katz). | Institution             | • Effectiveness Sources Portal (ESP)                                     
• Sponsored Project Excellence Achieved through Redesign (SPEAR)                |
| Learning Analytics | The use of analytic techniques to help target instructional, curricular, and support resources to support the achievement of specific learning goals (adapted from Bach). | Department/learner       | • Course Signals  
• Check My Activity                                                             |
| Predictive Analytics | An area of statistical analysis that deals with extracting information using various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behavior and events (adapted from Eckersen). | All levels              | • Student Success Plan  
• Student Readiness Inventory                                                  |

Purdue University’s Course Signals project

Examples & Projects (1/5)

Based on the dissertation of Dr. John Campbell a tool analyzes the learning behavior of students and can identify students who are not expected to successfully complete on the basis of their learning activity the course.

2009: first release

2010: „Users scored up to 26 % more As or Bs. Earned up to 12 % fewer Cs and up to 17 % fewer Ds & Fs”.

2011: 7k of 20k student used tool

now: added value, e.g. Workshops, Consultation, Advisors, …

Further details: www.itap.purdue.edu/studio/signals/
University of Maryland, Baltimore County (UMBC)

Examples & Projects (2/5)

„Check My Activity (CMA)“ allows self-comparison between students with regard to their activity within the LMS and achieved course score.

Report Code is open, see „Get The Code“

Further details: www.umbc.edu/blackboard/reports

7.3.2014, Hans Pongratz
University of Kentucky: Academic Health student app
Examples & Projects (3/5)

WAKE UP! GET TO CLASS!

- Who sets alarms for themselves?
- Why not automatically set alarms for students around their schedule?
- Why not have automated wake-up calls?
- Why not suggest wake up times based on class attendance?
- Why not consider manipulation of reminders as a form of engagement?
- Can we ascertain student prospective memory capability and personalize based on it?

Source:
University of Kentucky: Framework

Overview: EDUCAUSE review Academic Analytics
Examples & Projects (4/5)

• Enrollment Predictive Modeling at Baylor University
• Predicting and Improving Student Retention at the University of Alabama
• Developing a Student Success Plan and Early Alert System at Sinclair Community College
• Connecting Resource Utilization, Risk Level, and Outcomes at Northern Arizona University

Further details: http://www.educause.edu/ero/article/academic-analytics-new-tool-new-era
Open Academic Analytics Initiative (OAAI)

Examples & Projects (5/5)

Goal: open-source “early alert” system, which predicts “at risk” students within the first 2-3 weeks of a course and provides assistance for them.

Further details: www.educause.edu/events/educause-learning-initiative-2012-annual-meeting/open-academic-analytics-initiative-leveraging-openness-improve-learne
### Open Academic Analytics Initiative (OAAI)

First results seem very promising! See:


---

#### Examples & Projects (5/5)

<table>
<thead>
<tr>
<th>College</th>
<th>AAR run</th>
<th>#Students</th>
<th>Accuracy</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Savannah</td>
<td>AAR1</td>
<td>504</td>
<td>76.26%</td>
<td>35.36%</td>
<td>61.48%</td>
<td>70.54%</td>
</tr>
<tr>
<td>Ceritos</td>
<td>AAR1</td>
<td>502</td>
<td>61.95%</td>
<td>43.69%</td>
<td>47.41%</td>
<td>72.32%</td>
</tr>
<tr>
<td>AAR2</td>
<td>601</td>
<td>71.88%</td>
<td>27.49%</td>
<td>59.62%</td>
<td>70.78%</td>
<td></td>
</tr>
<tr>
<td>AAR3</td>
<td>649</td>
<td>75.19%</td>
<td>25.12%</td>
<td>60.60%</td>
<td>75.76%</td>
<td></td>
</tr>
<tr>
<td>Redwoods</td>
<td>AAR1</td>
<td>195</td>
<td>67.69%</td>
<td>40.48%</td>
<td>51.78%</td>
<td>82.61%</td>
</tr>
<tr>
<td>AAR2</td>
<td>195</td>
<td>78.97%</td>
<td>13.90%</td>
<td>71.88%</td>
<td>65.22%</td>
<td></td>
</tr>
<tr>
<td>AAR3</td>
<td>195</td>
<td>77.95%</td>
<td>14.29%</td>
<td>70.97%</td>
<td>63.77%</td>
<td></td>
</tr>
</tbody>
</table>

---

Table 1 - Correlations between course grades and CMS

<table>
<thead>
<tr>
<th>Undergraduate CMS event frequencies</th>
<th>Course Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marist Fall 2010</td>
<td>Campbell (2007)</td>
</tr>
<tr>
<td>N=18968</td>
<td>N=27276</td>
</tr>
</tbody>
</table>

- **Sessions Opened**
  - Correlation: 0.147
  - Significance: 0.000(***)
  - N: 11195

- **Content Viewed**
  - Correlation: 0.098
  - Significance: 0.000(***)
  - N: 7651

- **Discussions Read**
  - Correlation: 0.133
  - Significance: 0.000(***)
  - N: 19205

- **Discussions Posted**
  - Correlation: 0.233
  - Significance: 0.000(***)
  - N: 7667

- **Assign. Submitted**
  - Correlation: 0.146
  - Significance: 0.000(***)
  - N: 4309

- **Assmnts Submitted**
  - Correlation: 0.101
  - Significance: 0.000(***)
  - N: 4085

(***) Significant at the 0.01 level (2-tailed)

Marist data uses ratios over course mean instead of frequencies.

---

Table 3 - Prediction analysis on spring and fall pilot data

See:

https://confluence.sakaiproject.org/download/attachments/5671025/OAAI%20Final%20Progress%20Report.pdf?version=1&modificationDate=1391705397000&api=v2
Talking about laws (1/2)

• EU Directive 95/46/EC on data protection
• EU Directive requires member states to achieve result by not dictating
• Terms and conditions

Personal data: any information concerning the personal or material circumstances of identified or identifiable natural person (concerned). Under personal data thus fall details of name, student number, degree, address, affiliations associations, email, etc. (http://www.bfdi.bund.de/bfdi_wiki/index.php/3_BDSG_Kommentar_Absatz_1_Beispiele)

Anonymized in the sense of the Bavarian Laws (quite similar to German & EU): Personal data will be considered anonymous if the data has been modified so that the reference to individuals cannot or only under extremely difficult conditions are restored (see also Article 4, Section 8 BayDSG, http://byds.juris.de/byds/009_1.1_DSG_BY_1993_Art4.html)

7.3.2014, Hans Pongratz
Talking about laws (2/2)


Member States shall provide that personal data must be:

- Processed fairly and lawfully.
- Collected for specified, explicit and legitimate purposes and used accordingly.
- Appropriate and relevant in relation to the purpose for which they are processed.
- Accurate and kept up to date.
- Kept no longer than the time necessary for the purpose for which they are processed (30).

Personal data can be processed if:

- The data subject has been adequately informed and has given unambiguously his consent for the collection and further use of his data.
- Processing is necessary to perform a contract having as a party the data subject or to enter into a contract requested by the data subject.
- A legal obligation requires the processing of personal data.
- Processing data is necessary in order to ensure the essential interests of the data subject;
- Processing is necessary to perform tasks of public interests or carried out by an official authority.
- The data controller has a legitimate interest in processing the personal data of the data subject; this interest, however, has to be necessary balanced with the right to privacy of the data subject (51).
Questions to address & action items

- Root Questions vs. Research Approach
- Data sources
- Tools to use
- Anonymization of data
- Information about stored data
- Deleting stored data after X month/years
- Policy for data analytics?
- How gets access to data at which level?

Get in touch with data protection officer of your organization!
Exchange ideas, tools, results, approaches! -> pongratz@tum.de 😊