Educational Data Mining: Möglichkeiten und Unmöglichkeiten

Stefan Kramer

Institut für Informatik

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

Begriffsklärung und Hintergrund zur EDM und Lernanalytik

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

EDM vs. LA

- Educational Data Mining (EDM)
  - Anwendung von Data Mining auf Daten aus dem Bildungsumfeld
  - „*Human judgement is a tool to accomplish automated discovery*."
  - Wohldefinierte Probleme wie das sog. Bayesian Knowledge Tracing
- Learning Analytics / Lernanalytik (LA)
  - Sammeln, Analysieren und Berichten von Daten über Lernende und deren Kontexte, um das Lernen und die Umgebungen, in denen es stattfindet, zu verstehen und zu optimieren.
  - „*Automated discovery is a tool to accomplish human judgement*."

LA vs. AA

- Learning Analytics / Lernanalytik (LA)
  - Sammeln, Analysieren und Berichten von Daten über Lernende und deren Kontexte, um das Lernen und die Umgebungen, in denen es stattfindet, zu verstehen und zu optimieren.
  - „*Automated discovery is a tool to accomplish human judgement*.“
- Academic Analytics (AA)
  - Anwendung von Data Mining Tools und Strategien, um die Entscheidungspraxis in Bildungseinrichtungen zu steuern, so dass Stärken und Schwächen von Betrieb, Programm und Studierenden identifiziert werden können.

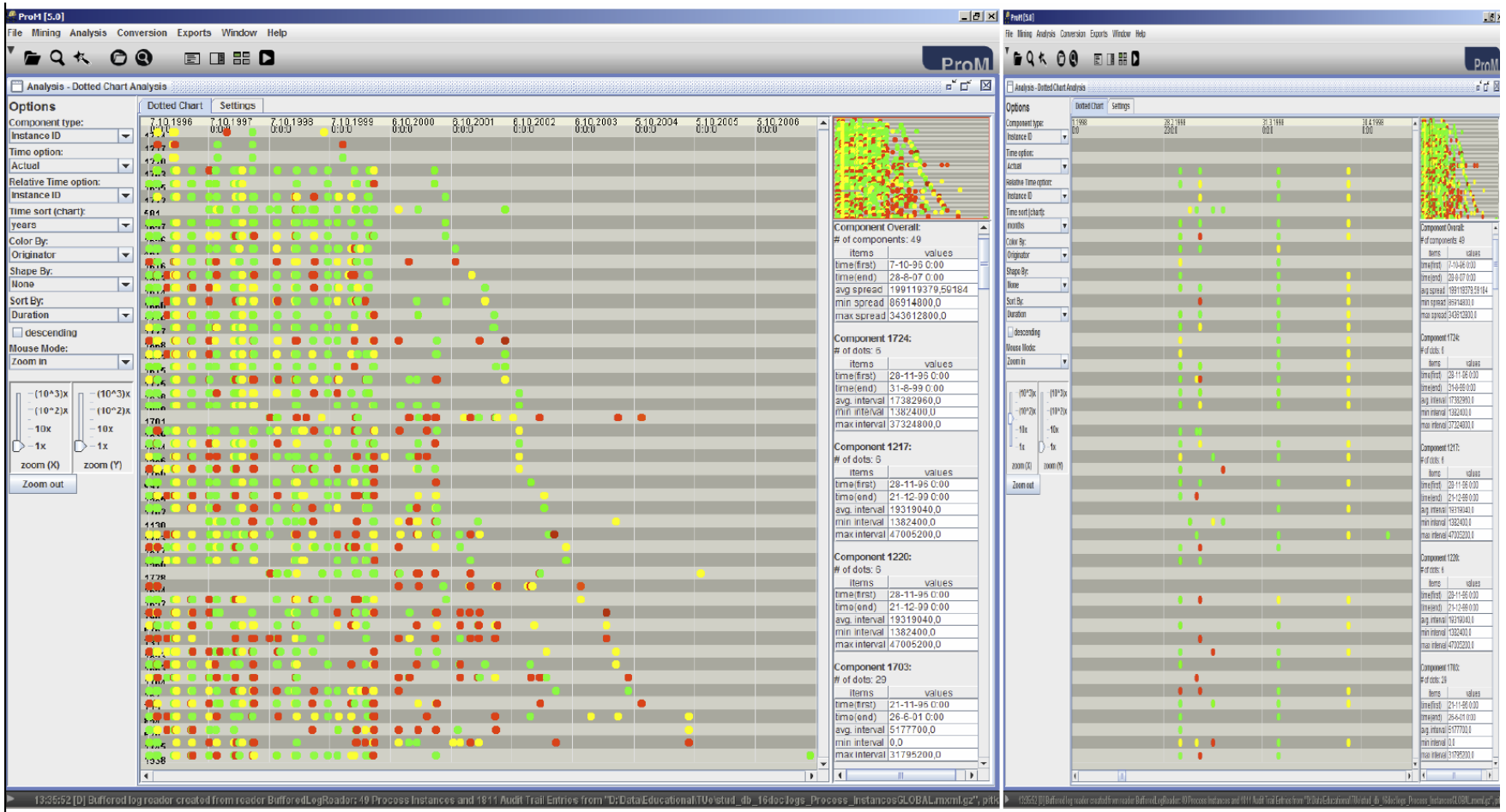
Data-Driven Education Communities: Gesellschaften, Konferenzen und Journale

- Educational data mining (EDM / JEDM / IEDMS)
- Learning analytics (LAK = learning analytics and knowledge / JLA / SoLAR)
- Learning at scale (L@S)
- Intelligent tutoring systems (ITS)
- AI in education (AIED)
- Learning sciences (LS)

Vielfältige Aufgaben und Ziele

- Wie lernen Studierende? Wie können wir ihnen helfen?
- Erfolgsfaktoren verstehen
- Diagnostizieren von Missverständnissen, Lernfähigkeiten, etc.
- Vorhersage und Vermeidung von Abbrüchen
- Verstehen, wie Studierende gemeinsam lernen
- Datengetriebene Gestaltung von Kursen und Studiengängen

Beispiel: Überblick/Visualisierung von Leistungen von Studierenden



Probleme in den Anfängen des EDM

- Lyrics by Jack Mostow, CMU
  - *All my data from my tutor,*
    I put on a USB
    That I accidentally swallowed --
    Now my study's history.
  - *All my data,*
    Educational data that I mine,
    Now is lost and gone forever --
    Dreadful sorry, data mine

... Song sang at one of the EDM conference

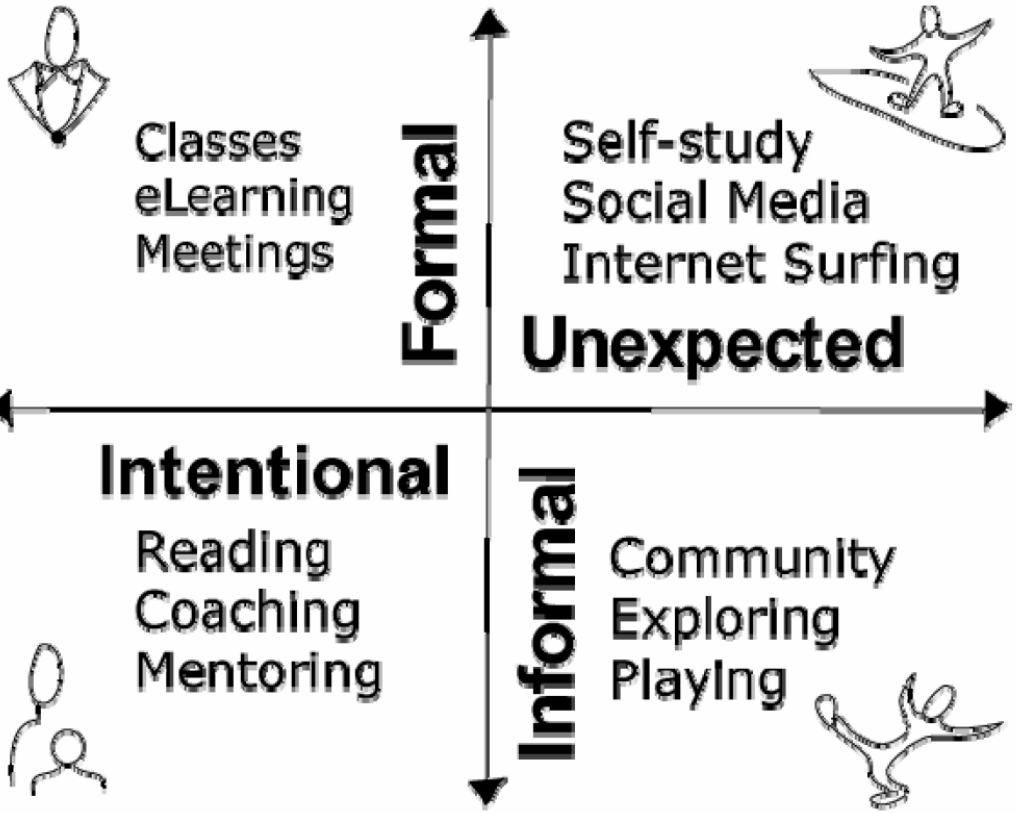
EDM Geschichte

- Phase I: Bildungsforschung
  - Hypothese, Datenerhebung, Hypothesentest; Publikation
- Phase II: Analytik
  - Daten werden gesammelt, weil man es kann; Data Mining auf Bildungsdaten; Publikation
- Phase III: getrieben durch Bedarf an Bildung
  - Technologie und Wissen vorhanden
  - Kommerzialisierung und Vermarktung
- Phase IV: „Learning at Scale“
  - Synergie von Forschung und Entwicklung. Effiziente Art der Organisation von A/B-Tests. Möglicher Datenmissbrauch und neue Welle des Datenschutzes.

Vier Arten von Lernen und wo EDM unterstützen kann

How to (re)organize the classes, or assessment, or placement of materials based on usage and performance data

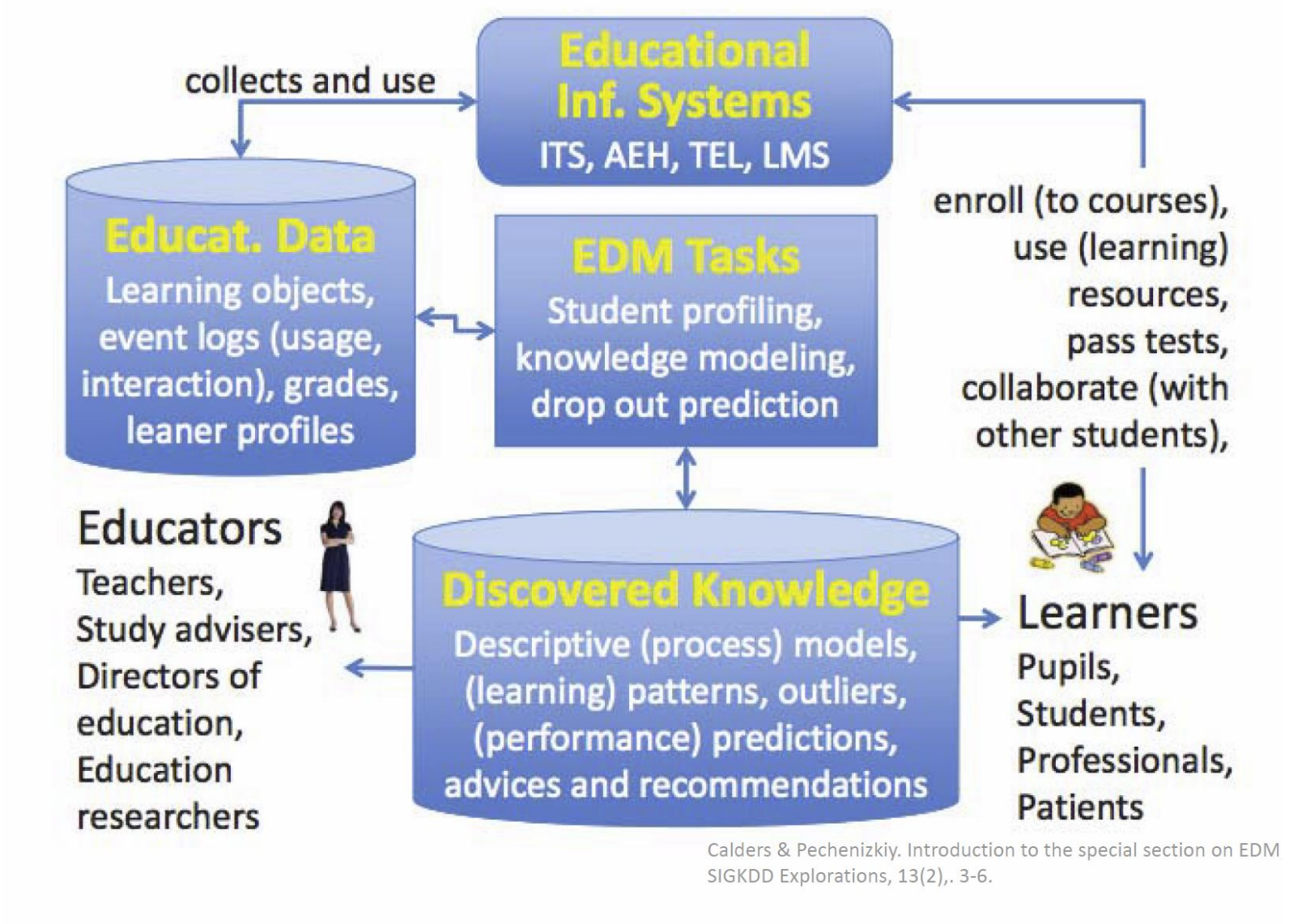
How to identify those who would benefit from provided feedback, study advice or other help; How to decide which kind of help would be most effective?



How to help learners in (re-) finding useful material, done whether individually or collaboratively with peers

How to help learners in (re-) finding useful material, done whether individually or collaboratively with peers

Educational Data Mining: Möglichkeiten und Unmöglichkeiten



Calders & Pechenizkiy. Introduction to the special section on EDM SIGKDD Explorations, 13(2),. 3-6.

Arten von erfassten Daten

- Verwaltungsdaten
  - Wer wo eingeschrieben ist, wer welchen Kurs belegt
  - Anmeldungen für eine (Zwischen-)Prüfung, Wiederholungsprüfungen
  - Demographie, Schulnoten, etc.
- MOOC und LMS
  - Ressourcennutzung
  - Bewertungsdaten (Online-Tests, …)
  - Foren, Zusammenarbeit, Feedback/Hilfeanfragen
  - Bewertung von Lernressourcen durch die Studierenden
- ITS, Lernspiele, e-Health, Simulatoren, ...
- Spielen, Surfen, Email, Facebook, Twitter, etc.

Daten, Methoden und Ziele

Interactions data
- Usage logs & contexts

Administrative data
- Enrolments
- Results
- Payments
- Graduation
- Employment

"Feedback" data
- Opinions
- Preferences
- Needs

Descriptive data
- Demographics
- Characteristics

Classification

Categorizing students

Clustering

Grouping similar students

Association Analysis, Sequence mining

Find courses taken together or Popular (parts of) study programs

Visual Analytics

Facilitate reasoning about the process or results via interactive data/model visualization

Process mining

Understanding study curricular

Goals

- Identify high risk students
- Predict new student application rates
- Predict students retention/dropout
- Course planning & scheduling
- Faculty teaching load estimation
- Predict demand for resources (library, cafeteria, housing)
- Predict alumni donation

Educational Data Mining als Teil des LOB Projekts

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

Das Wichtigste vorab

- Vorhersage des Erfolgs von Studenten
- Nur akademische Daten (Studienleistungen und Prüfungsleistungen)
- Technischer Beitrag: Neuer Ansatz für das Lernen von Aggregationsfunktionen

Studienberatung - aktueller Stand

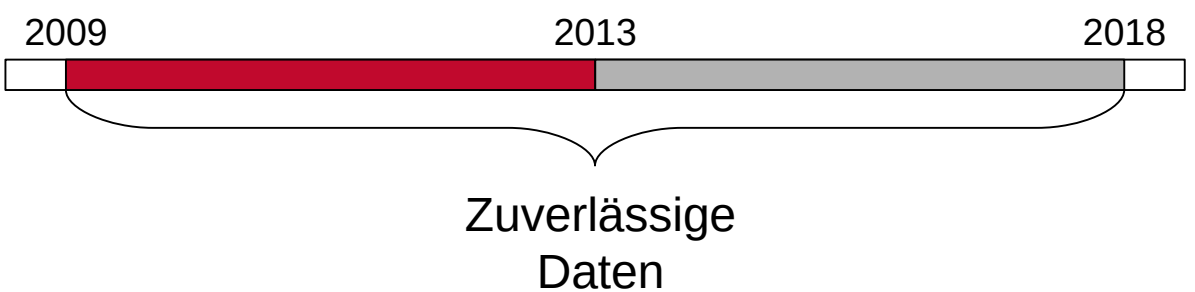
- Freiwillige Studienberatung
  - Basierend auf der Selbsteinschätzung der Studenten
  - Häufig in Anspruch genommen wegen dringender Probleme
  - Manchmal zu spät konsultiert, um diese Probleme zu lösen
- Obligatorische Studienberatung
  - Basierend nur auf ECTS-Leistungspunkteschwellen
  - Beispiel: 15 Leistungspunkte nach dem ersten Jahr (25% des erwarteten Betrages)

Studienberatung - Ziele

- Freiwillige Studienberatung
  - Verbesserung der Selbsteinschätzung der Studenten
  - Einblicke in mögliche Risikofaktoren geben
  - Bereitstellung eines Frühwarnsystems
- Obligatorische Studienberatung
  - Verbesserung der Auswahlregeln
    - höhere Aussagekraft
    - einfach und verständlich (erklärbar)

Daten

- Bachelor of Science: Informatik, Mathematik, Meteorologie, Physik, Physik
- Mindestens eine abgelegte Prüfung
- Entweder erfolgreich abgeschlossenes Studium oder Studienabbruch
- Erste Einschreibung zwischen 2009 und 2013

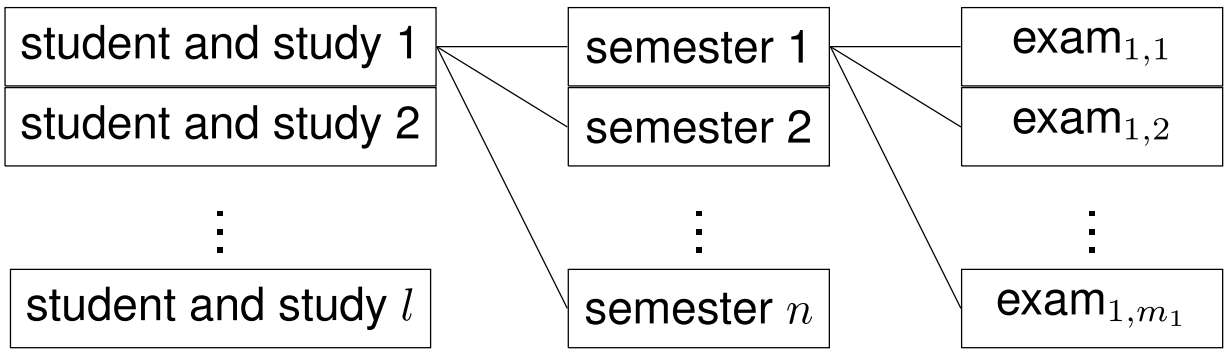


Information

- Student und Studium:
  - Schulabschluss, Hochschulzugangsberechtigung
- Semester:
  - ECTS-Leistungspunkte, Immatrikulationsstatus
- Prüfung:
  - Art der Prüfung, Bestehensstatus, Note
  - Keine demographischen oder sonstige Informationen

Datenstruktur

- Repräsentiert in einer relationalen Datenbank



- Herausforderung, die Struktur in das Modell zu integrieren
  - *Relationales Data Mining*

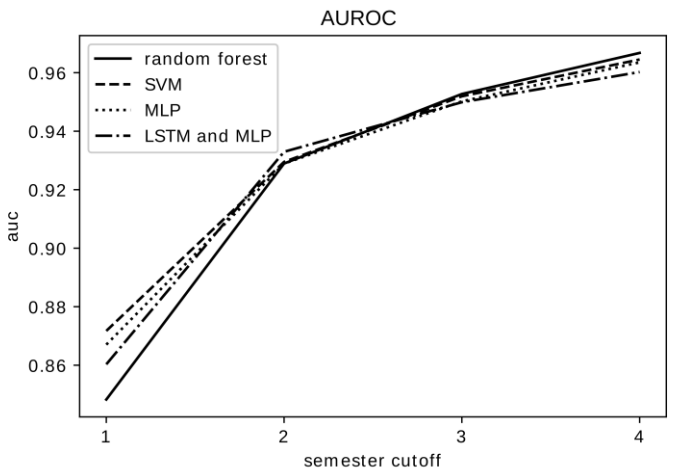
Klassifikation

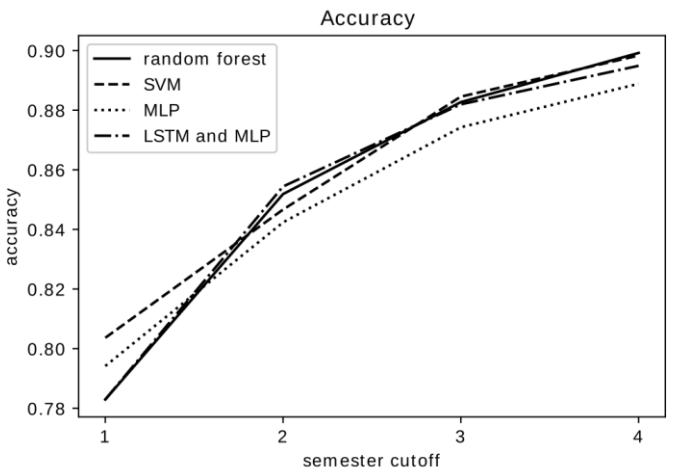
- Klassifizieren von Studierenden als Absolventen (y=1) oder Studienabbrecher (y=0)
- Methoden:
  - RandomForests
  - Lineare Stützvektormaschinen (SVM)
  - Multilayer Perceptrons (MLP)

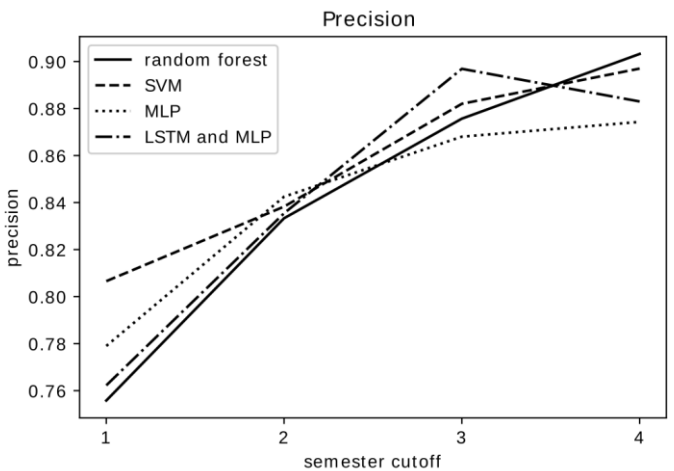
Experimentelles Set-Up

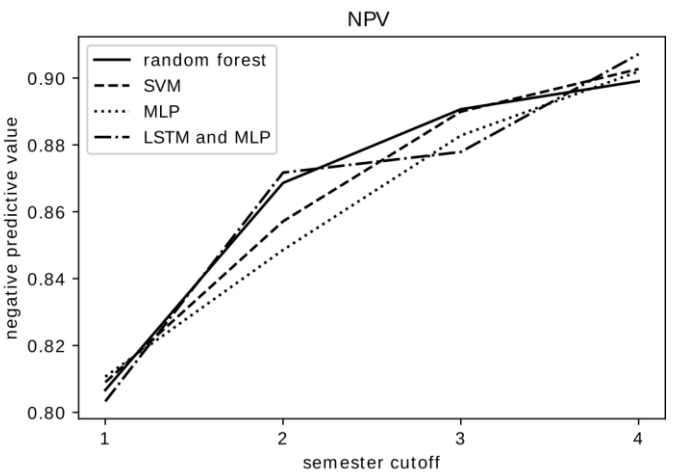
- Unterschiedliche Semesterschranken (z.B. Vorhersage nach einem Semester, Vorhersage nach zwei Semestern, …)
- 10-fache Kreuzvalidierung mit interner Hyperparameter-optimierung
- Verwendete Leistungskennzahlen:
  - Area Under ROC Curve (AUROC)
  - Accuracy
  - Precision
  - Negative Predictive Value (NPV)

Experimentelle Resultate









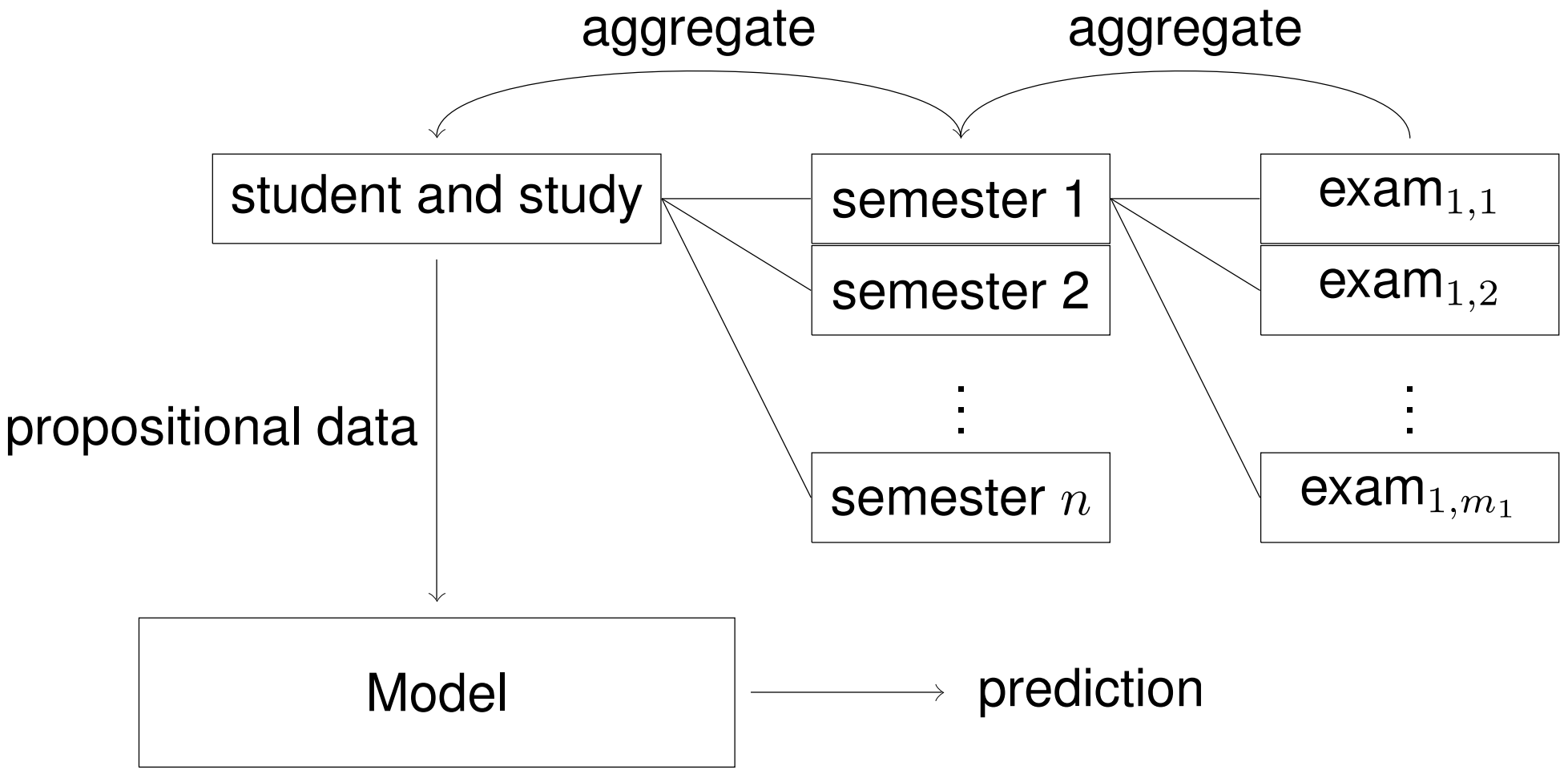
Experimentelle Resultate

Semesters	Top features
1	Grade
	Cumulative credit points
	Credit points per semester
	Type of exam: written exam
	Passing status
1,2	Cumulative credit points
	Credit points per semester
	Type of exam: written exam
	Grade
	Passing status

Semesters	Top features
1,2,3	Cumulative credit points
	Credit points per semester
	Type of exam: written exam
	Grade
	Passing status
1,2,3,4	Cumulative credit points
	Credit points per semester
	Grade
	Type of exam: written exam
	Type of exam: active participation

Abgeleitete Methodische Fragestellungen

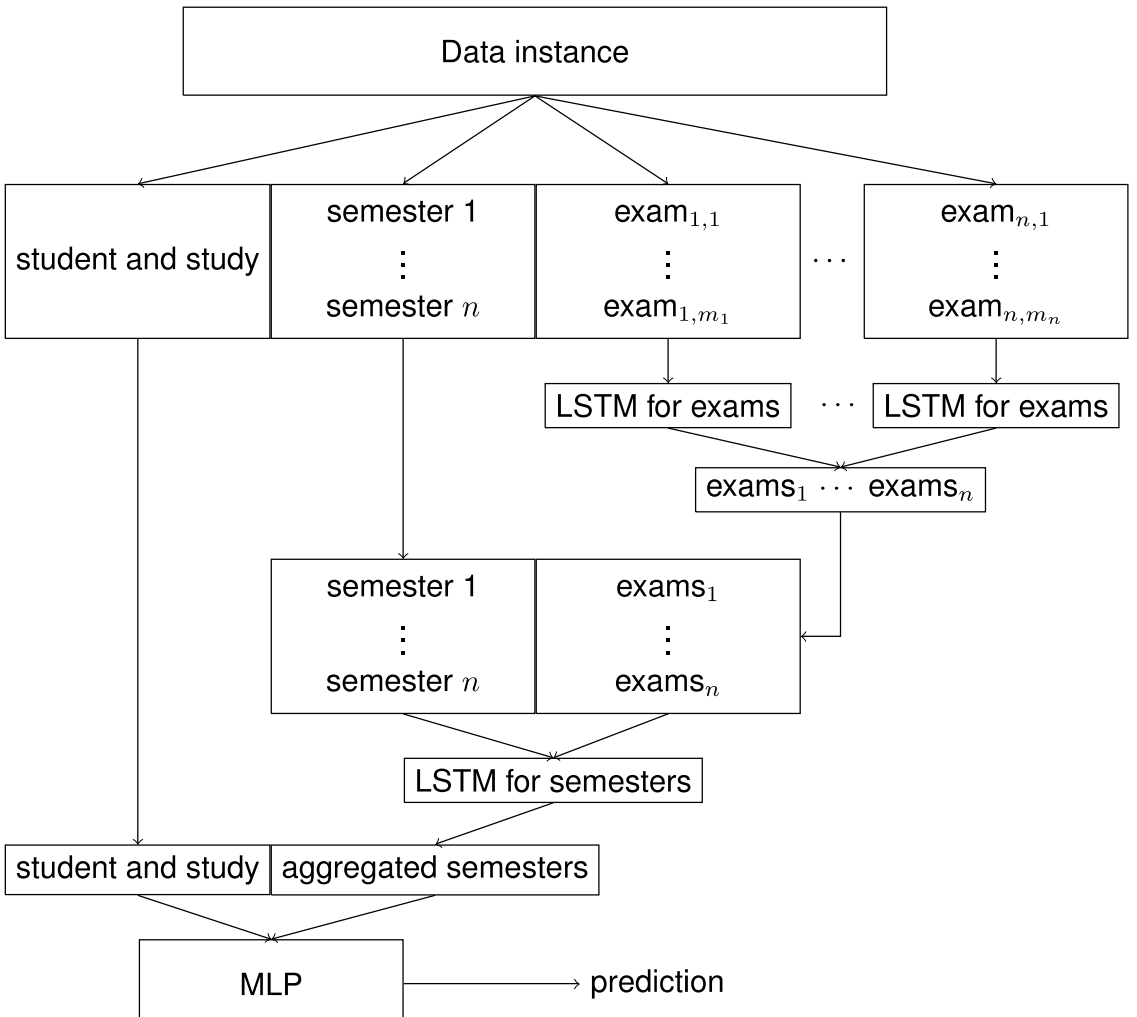
Relationale Daten und Aggregationen



Datenvorverarbeitung

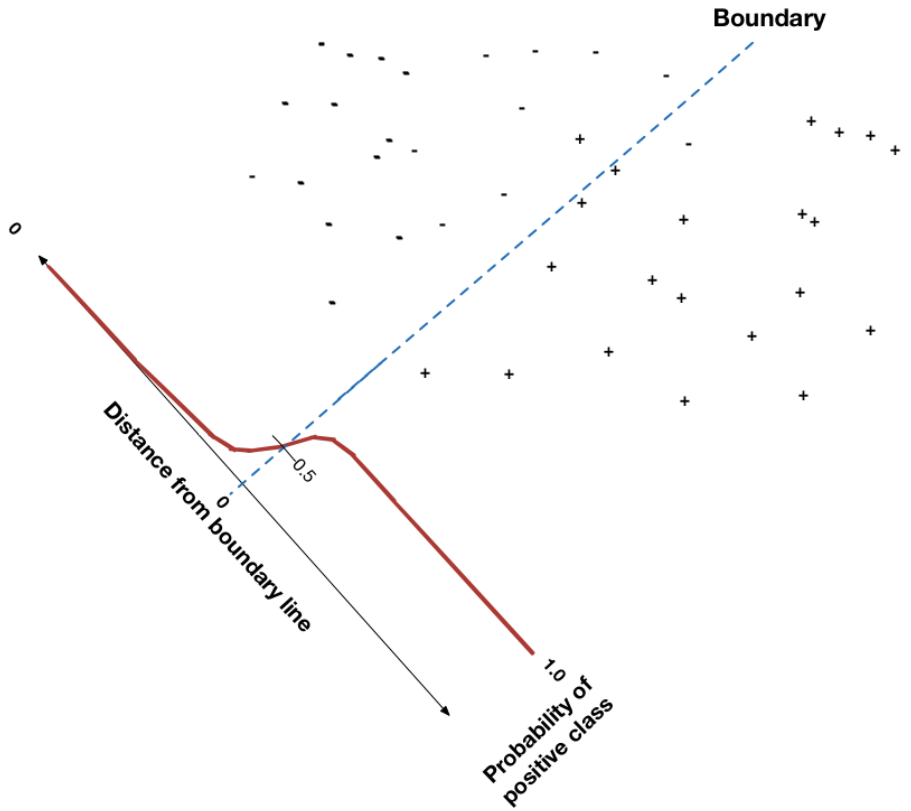
- Propositionalisierung durch Aggregation
- Relationale Aggregationen (RELAGGS)
  - Mittelwert, Minimum, Maximum, Standardabweichung, Summe
- Aggregationsfunktionen lernen
  - Nutzung von Long Term Memory (LSTM)-Netzwerken

Aggregationsnetzwerk mithilfe von LSTMs



Weitere Machine Learning Szenarien

- Lernen von Risikofaktoren



- Lernen aus Permutationen

Unmöglichkeiten

Unmöglichkeiten

- *Man wird niemals perfekt Studienverlauf und -erfolg statistisch oder mit KI vorhersagen und/oder optimieren können (oder wollen!)*

- Phil McRae: “Adaptive learning systems … primarily attend to those things that can be easily digitized and tested (math, science and reading). They fail to recognize that *high quality learning environments* are deeply relational, humanistic, *creative*, *socially constructed*, active and *inquiry-oriented*.”

Aus Artikel von chinesischen Autoren …

Student daily behavior. Students' daily behavior can clearly reflect their academic effort. Just studying in class only goes so far. If the student wants to get good academic performance, he/she must also work hard after class. So student daily behavior is important for accurately predicting student performance. But some students prefer studying in the classroom; some prefer studying in the library; some prefer studying in the dorm. And some students prefer work hard in the beginning; some work hard in the late; some work hard in the mid. The situation is very complicated, summary quantities can not work very well. These information should be combined together and every day during the semester is important. We count the number of visiting library during 2017 spring semester for every student and plot Figure 6. The shape is like a Poisson distribution and this means that most students do not go to the library to study. This may result from that libraries are not sufficient for students to use. Table III shows some average frequencies about entering the dorm and eating meals in 2017 spring semester. From the Table III, we can see that larger students come back to dorm between 22:00:00 and 23:00:00; the number of students eating lunch on time is larger than the number of students eating dinner on time; the number of students eating breakfast on time is obviously small. We did a survey about eating meals and find that most students can not get up early, they have no time eating breakfast; some girls will choose not to eat dinner due to dieting.

TABLE III
SOME AVERAGE FREQUENCIES IN 2017 SPRING SEMESTER.

Item	Value
#Coming back to dorm after 19:00:00	42
#Coming back to dorm after 20:00:00	37
#Coming back to dorm after 21:00:00	28
#Coming back to dorm after 22:00:00	19
#Coming back to dorm after 23:00:00	8
#Eating breakfast on time	48
#Eating lunch on time	68
#Eating dinner on time	60

Angst vor Datenschutzverletzungen

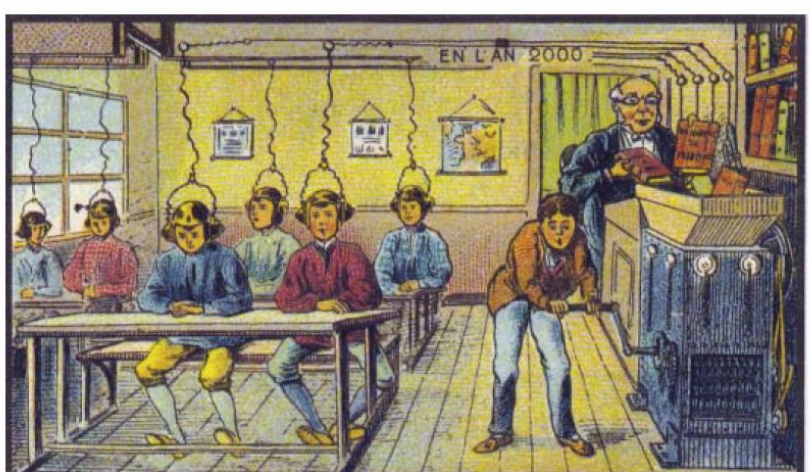
corpwatch.org/img/original/google.jpg

"Many companies are looking to profit from student and teacher data that can be easily collected, stored, processed, customized, analyzed, and then ultimately resold".

Philip McRae (Alberta Teachers' Association)

Neue (?) Ängste vor Personalisierung

- "When Personalization Goes Bad"
  http://www.portical.org/blog/when-personalization-goes-bad
- "Rebirth of the Teaching Machine through the Seduction of Data Analytics: This Time It's Personal"
  http://www.philmcrae.com/2/post/2013/04/rebirth-of-the-teaching-maching-through-the-seduction-of-data-analytics-this-time-its-personal1.html
- "This time it is Personal and Dangerous"
  http://barbarabray.net/2013/12/30/this-time-its-personal-and-dangerous/



Postcard (World's Fair, Paris 1899) predicting what learning will be like in France in the year 2000

© Pawel Kuczynski

Student Privacy Laws in den USA

- "Die Gesetzgebung [wird] … die Kontrolle über Bildungsdaten an Schüler/Studierende und LehrerInnen/ProfessorInnen zurückgeben.“ David Vitter
- Op-in-Anforderung für alle Studien (Einwilligung)
- Verbot der Sammlung "jeglicher Art von psychologischen Daten, einschließlich der Bewertung von nicht-kognitiven Fähigkeiten oder Attributen“
- Identifizierbare Datensätze dürfen nicht mit anderen Datenquellen verknüpft werden.
- Datenquellen von Bund und Bundestaaten dürfen nicht mit den Verwaltungsdaten der Schulen/Universitäten verknüpft sein.

EU DSGVO (GDPR)

- (Neu-)Definition personenbezogene Daten: Entscheidend ist alleine die Tatsache, inwieweit es gelingen kann, die Daten mit **vertretbarem Aufwand** einer bestimmten Person zuzuordnen
- „Privacy-by-Default“ und „Privacy-by-Design“
- Einwilligung und Zweckbindung
- Grundsatz der Datensparsamkeit
- Hohe Strafen
- Erste Gerichtsurteile

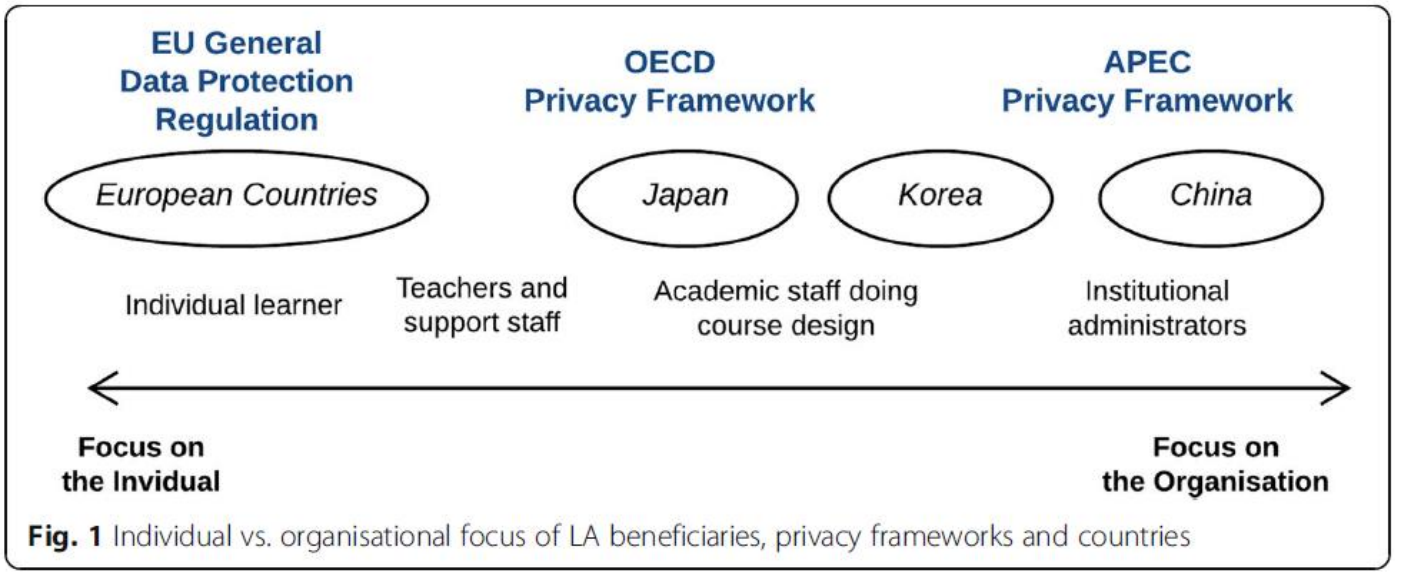


Fig. 1 Individual vs. organisational focus of LA beneficiaries, privacy frameworks and countries

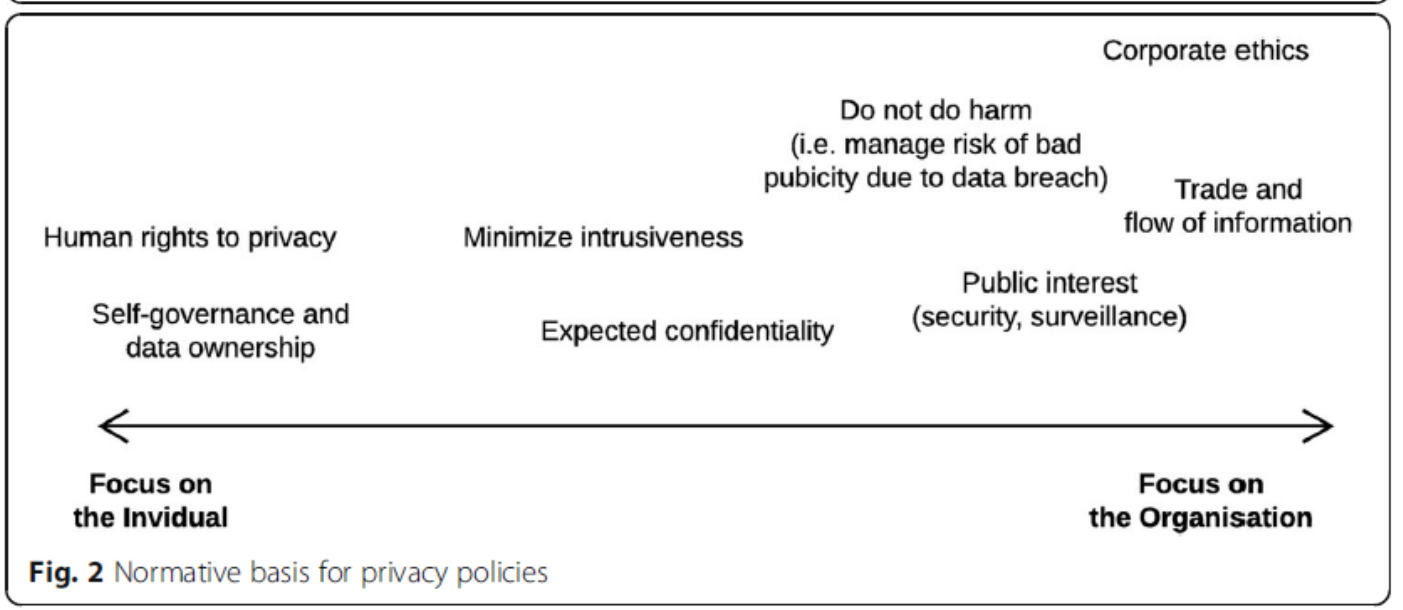


Fig. 2 Normative basis for privacy policies

Educational Data Mining: Möglichkeiten und Unmöglichkeiten

Hoel and Chen *Research and Practice in Technology Enhanced Learning* (2018) 13:20
https://doi.org/10.1186/s41039-018-0086-8

Research and Practice in Technology Enhanced Learning

RESEARCH Open Access

Privacy and data protection in learning analytics should be motivated by an educational maxim—towards a proposal

Tore Hoel* and Weiqin Chen

* Correspondence:
tore.hoel@oslomet.no
Oslo Metropolitan University,
Postboks 4 St. Olavs plass, 0130
Oslo, Norway

Abstract

Privacy and data protection are a major stumbling blocks for a data-driven educational future. Privacy policies are based on legal regulations, which in turn get their justification from political, cultural, economical and other kinds of discourses. Applied to learning analytics, do these policies also need a pedagogical grounding? This paper is based on an actual conundrum in developing a technical specification on privacy and data protection for learning analytics for an international standardisation organisation. Legal arguments vary a lot around the world, and seeking ontological arguments for privacy does not necessarily lead to a universal acclaim of safeguarding the learner meeting the new data-driven practices in education. Maybe it would be easier to build consensus around educational values, but is it possible to do so? This paper explores the legal and cultural contexts that make it a challenge to define

Empfehlungen von Hoel und Chen

1. “Privacy and data protection in LA are achieved by negotiating data sharing with each student.”

2. “Openness and transparency are essential and should be an integral part of institutional policies. How the educational institution will use data and act upon the insights of analysis should be clarified in close dialogue with the students.”

3. “Big data will impact all society. Therefore, in negotiating privacy and data protection measures with students, schools and universities should use this opportunity to strengthen their personal data literacies.”

Zusammenfassung und Schlussfolgerungen

Zusammenfassung und Schlussfolgerungen

- Eine Vielzahl von Lernsettings:
  Educational Data Mining, Learning Analytics, Academic Analytics, Data Driven Education
- Learning Science <-> Data Mining
- Zwischenstand einer Fallstudie an der JGU: erfolgreiche Identifikation „gefährdeter Studierender“ ausschließlich auf der Grundlage von Studienleistungen, ohne zusätzliche (beispielweise: demographische) Daten
- Verwendung einfacher und erlernter Aggregationsfunktionen
- Verantwortungsbewusste Analyse und Nutzung der Daten

Vielen Dank für Ihre Aufmerksamkeit!

FACT und FAIR Daten Prinzipien

FACT

– Fairness / Discrimination-Awareness
– Accountability / Accuracy
– Confidentiality / Privacy
– Transparency / Interpretability

→ *Erfordert multidisziplinäre Forschung!*

FAIR

- Findable
- Accessible
- Interoperable
- Reusable.

Zusammenfassung und Schlussfolgerungen

- Besseres Verständnis der *Trade-offs*, z.B.:
  - *Datenschutz-Personalisierung*
  - *Personalisierung-Diskriminierung*
- Bessere Werkzeuge für datengesteuerte Entscheidungsfindung:
  - *Vertrauen, Transparenz, Zuverlässigkeit*
- Information und mögliche Reduzierung der Angst vor Big Data Technologien mit Hinblick auf die Öffentlichkeit, Regulierungsbehörden und die politischen Entscheidungsträger

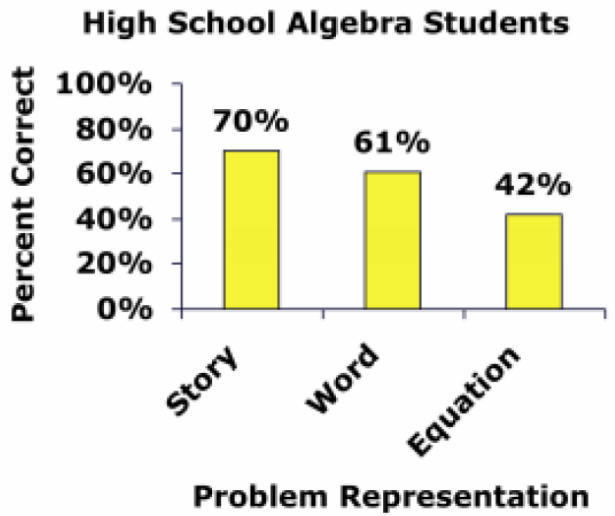
Fragestellungen und Themen

- Ist unsere Problemformulierung richtig?
- Sind die Erfolgskriterien klar?
- Kostensensitivität
- Falsch-positive Raten?
- Wie werden Prognosen verwendet (Entscheidungshilfe?)
- Black-Box vs. Transparent
- Wie einfach es ist, zu testen, zu debuggen, zu tunen?
  Nicht nur für die Richtigkeit der Umsetzung, sondern auch für die Leistung
- Wie können wir die Ergebnisse richtig interpretieren?
- (Statistische vs. praktische) Signifikanz der Ergebnisse

Fragestellungen und Themen

- Geben prädiktive Modelle Garantien?
- Ist die Genauigkeit hoch genug?
- Bieten prädiktive Modelle Einblicke? Interpretierbarkeit.
- Korrelation ist nicht Kausation
- Sind die Entscheidungen basierend auf prädiktiven Modellen immer ethisch einwandfrei?
  - Sowohl global (A/B-testbasiert) als auch personalisierte Entscheidungen können für eine bestimmte Gruppe ungerecht sein.
  - Wenn dies mit manchen sensiblen Attributen zusammenhängt, kann das nicht nur ethisch problematisch sein, sondern auch rechtliche Konsequenzen nach sich ziehen.

Which way the students learn better?



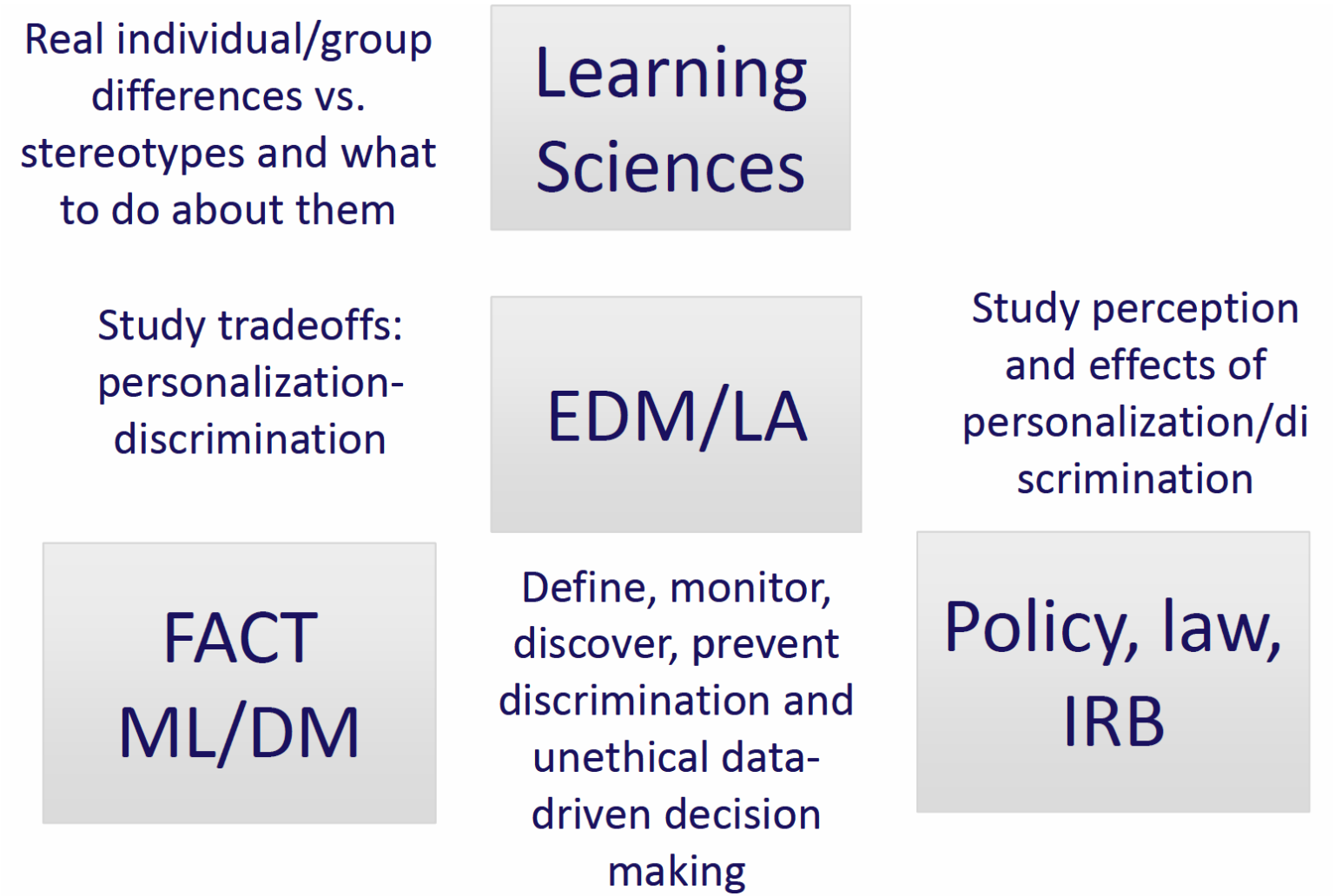
How these averages could possibly differ per

- student's learning style,
- major, background,
- country they studied,
- ethnicity, gender, ...

MOOC, ITS, L@S: A/B testing becomes ubiquitous

How to automatically choose the best path/treatment for every student? Realizing the data-driven education (personalization) dream based on mining data from massive A/B testing

Multidisciplinary R&D Landscape



Daten

- Aus dem Studentenmanagementsystem
- Bachelor of Science-Studenten
- Erste Einschreibung zwischen 2009 und 2013
- Mindestens eine besuchte Prüfung
- Ausgewählte Hauptfächer:
  - Informatik
  - Mathematik
  - Meteorologie
  - Physik

Major	**Total number**	**Graduates**	**Dropouts**
Computer science	339	106	233
Math	380	186	194
Meteorology	59	29	30
Physics	383	219	164
Combined	**1161**	**540**	**621**

Struktur der Studien

- Einblicke hinsichtlich „gefährdeter“ Studierender
  - Allgemeine Studienmuster
  - Faktoren mit hoher Vorhersagekraft
- Erstellen einer Liste von strukturellen Risikofaktoren
  - Verbesserungspotenziale
    - für das Studium
    - für Kurse

Quellen

- Introduction to the special section on educational data mining (of the ACM SIGKDD Explorations) by T. Calders & M. Pechenizkiy
  https://dl.acm.org/citation.cfm?doid=2207243.2207245

- Learning analytics and educational data mining: Towards communication and collaboration by G. Siemens & R. Baker
  https://www.researchgate.net/publication/254462827_Learning_analytics_and_educational_data_mining_Towards_communication_and_collaboration

- Forecast of Study Success in the STEM Disciplines Based Solely on Academic Records by L. Pensel & S. Kramer
  https://drive.google.com/file/d/17RGxIWdrVwcF6BvhwBVepeiKDMv45-5A/view

- Rebirth of the Teaching Machine through the Seduction of Data Analytics: This Time It's Personal by P. McRae & J. Bower
  https://nepc.colorado.edu/blog/rebirth-teaching-machine-through-seduction-data-analytics-time-its-personal

- Privacy and data protection in learning analytics should be motivated by an educational maxim—towards a proposal by T. Hoel & W. Chen
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6294277/pdf/41039_2018_Article_86.pdf