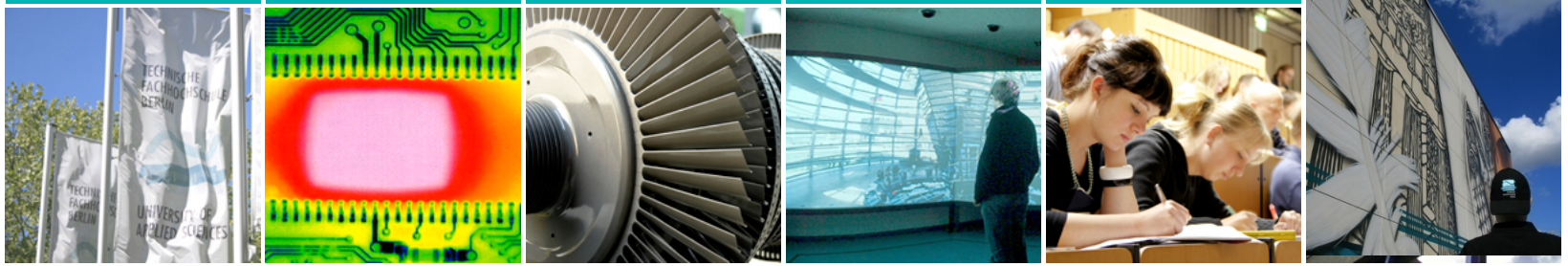
BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

EDM: Methods, Tasks and Emerging Trends

Agathe Merceron

Beuth Hochschule für Technik, Berlin



Outline

- Introduction: Big Data in Education
- Methods and Tasks:
  - Prediction
  - Clustering
  - Relationship Mining
  - Distillation of Data for Human Judgment
  - Discovery with Models
- Current Trends
- Conclusions
- *List of References*

Introduction

- “Big Data in Education” MOOC by Ryan Baker
  - Coursera 2013
  - EdX 2015
- Any idea of what big data can mean?
- Illustration with 3 main sources of data in education (categorization by Romero & Ventura 2010).

Introduction

- Data from administration:
  - Small-medium university of 12 000 students
  - Each student is enrolled in 6 courses (of 5 ECT each)
  - Each semester 60 000 new marks to store.

Introduction

- Data from Learning Management Systems (LMS):
  - Small-medium university of 12 000 students.
  - 40 degree-programs with 15 courses per degree.
  - Each course taught for 60 students over 12 weeks.
  - Each course has each week 1 set of slides and 1 quiz + 1 forum for the semester.
  - Each student access weekly twice the set of slides and the quiz, and 3 times the forum during the whole semester.
  - 1 836 000 access-interactions stored by the LMS each semester.

Introduction

- Data from dedicated software such as Intelligent Tutoring Systems, Serious Games etc. :
  - DataShop (Ködinger et al. 2010)
  - Hubble (Luengo 2014)
- Other sources of data:
  - Social media
  - Questionnaires
  - Forums
  - Etc.
- These data are big enough to be analysed by algorithms: the core of fields like Educational Data Mining and Learning Analytics.

Introduction

- **Educational Data Mining** is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. http://www.educationaldatamining.org/

Introduction

- **Learning analytics** is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.
  https://tekri.athabascau.ca/analytics/

Introduction

- Research is vivid in both fields!

LAK 15

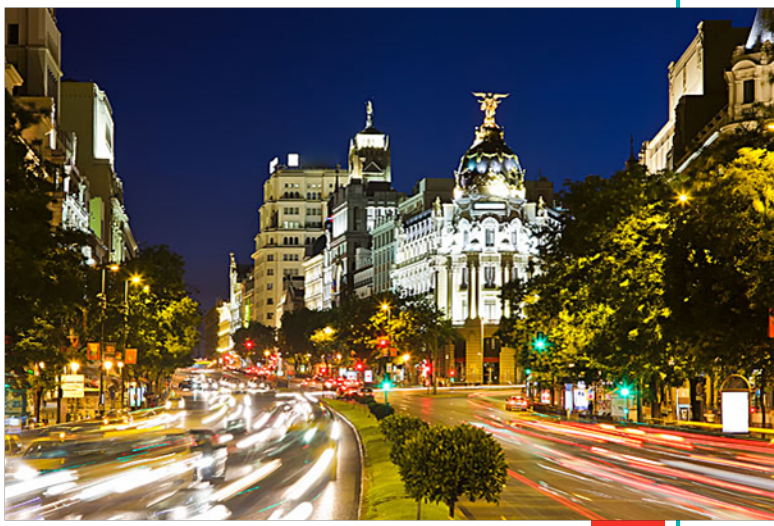
The 5th International
Learning Analytics & Knowledge Conference
Scaling Up: Big Data to Big Impact
Marist College, Poughkeepsie, NY, USA March 16 - 20, 2015

SOLAR
SOCIETY for LEARNING ANALYTICS RESEARCH
acm In-Cooperation
sig web

EDM 2015

The 8th International Conference on Educational Data Mining

26-29 June 2015
Madrid - Spain



Introduction

- Methods (Baker & Yacef 2009) come mainly from data mining, machine learning, to some lesser extend from *classical artificial intelligence*, and increasingly from natural language processing.
  - The construction of a Q-Matrix from students answers to questions uses a hill-climbing algorithm (Barnes 2005).

	q1	q2	q3	q4	q5
conc1	1	0	0	0	0
conc2	0	1	0	0	0
conc3	0	0	1	1	0
conc4	0	0	0	0	1

Outline

- Introduction: Big Data in Education
- Methods and Tasks:
  - Prediction
  - Clustering
  - Relationship Mining
  - Distillation of Data for Human Judgment
  - Discovery with Models
- Current Trends
- Conclusions

Prediction

- Major task: predict performance.
- Different levels of granularity:
  - Drop-off (Wolff & al. 2013)
  - Pass/fail, mark in a degree (Zimmerman & al. 2011)
  - Pass/fail, mark in a course (Lopez & al. 2012)
  - Skill mastery in a tutoring system (Pardos & al. 2007).

Prediction: Mark in a degree

- Many works show that pass or fail, or even the interval of a mark in a degree or a course can be predicted with an accuracy of 70% or higher.
- No classifier that works best in all contexts.
- No set of features that work best in all contexts, though some works to predict the interval of the mark for a university degree suggest that including marks is essential (Golding & al. 2006, Zimmerman & al. 2011).

Prediction: Mark in a degree

- Predict the interval of the degree mark: A, B, C. D or E (Asif & al. 2014).
- 4-years Bachelor Computing and IT in a well known (there) technical university of Pakistan.
- Competitive: selection on the marks in the High School Certificate (HSC) and entrance exam.
- Conjecture: academic records (no socio-economic feature) might be enough to predict the final mark with a reasonable accuracy: better than the baseline of predicting the majority interval C, 51.92%.

Prediction: Mark in a degree

Prediction: Mark in a degree

- Which features? HSC marks, marks of all modules from 1st and 2nd year and number of attempts.
- Which classifiers? Try all the well-known ones.
- Validation: one cohort as training set and the next cohort as test set (needs some stability in the curriculum) for generalization and pragmatic policy. Different from other works which mostly use cross-validation.
  - Cohort 1: 105 students graduated in 2012
  - Cohort 2: 104 students graduated in 2013

Prediction: Mark in a degree

Classifier	Accuracy / Kappa
Decision Tree with Gini Index	68.27% / 0.493
Decision Tree with Information Gain	69.23% / 0.498
Decision Tree with Accuracy	60.58% / 0.325
Rule Induction with Information Gain	55.77% / 0.352
1- Nearest Neighbors	**74.04% / 0.583**
Naives Bayes	**83.65% / 0.727**
Neural Networks	62.50% / 0.447
Random Forest with Gini Index	**71.15% / 0.543**
Random Forest with Information Gain	69.23% / 0.426
Random Forest with Accuracy	62.50% / 0.269

Clustering

- Big variety of tasks.
- Variety of algorithms.
- Two works:
  - Clustering students to find out typical behaviours in a forum (Cobo & al. 2012)
  - Clustering utterances to find out speech acts or dialog acts (Ezen-Can & al. 2015).

Clustering: behaviours in forum

- 8 features: 4 for writing, 4 for reading:
  - Number of initiated threads, number of reply posts, number of students replied, number of days with writing.
  - All calculated as ratio.
- Hierarchical agglomerative clustering:
  - 2 clusterings: writing features and reading features.
  - Normalized Euclidean distance.
  - Complete link.
  - Adaptation of inconsistency criterion to isolate the best clusters.
  - Clusters from the 2 clusterings are combined.

Clustering: behaviours in forum

- Find known results: less students write than read.
- The smaller the reading, the higher the drop-off rate and fail rate.
- results

Clustering: dialog acts

- Dialog acts:
  - Question: “What is an anonymous class?”.
  - Answer: “An anonymous class is a class without name.”.
  - Issue, problem.
  - Statement.
  - Reference: “An interesting video about Bubblesort.”.
  - Positive, negative acknowledgment: “Thanks, I got it”, “I am still confused”.
- Problem: classify automatically sentences in forums or tutorial dialogs in dialog acts.

Clustering: dialog acts

- Classical approach is supervised:
  - Annotate manually a large corpus (bottle neck).
  - Identify cues or features: punctuation, unigram, bigram, position of unigram in the sentence, preceding dialog act, etc. (Kim & al. 2010) .
  - Train a classifier. Support Vector Machine (Kim & al. 2010):
    - Positive_ack: F-Score 0.54 (9.20% of the sentences).
    - Questions: F-Score 0.95 (55.31% of the sentences).

Clustering: dialog acts

- Unsupervised approach(Ezen-Can & al. 2015). Dialogues come from a computer mediated environment to tutor students on programming. Students recorded by Kinect cameras.
- Features to describe sentences:
  - Lexical features: unigram, word ordering, punctuation.
  - Dialog-context features: position in the dialog, length, author of previous message (tutor, student), dialog act of previous message.
  - Task features: task before the utterance (writing, compiling), status of most recent coding action etc.
  - Posture features: head distance, torso distance.
  - Gesture features: one hand and two hands to head. Posture and gesture features not trivial to calculate.

Clustering: dialog acts

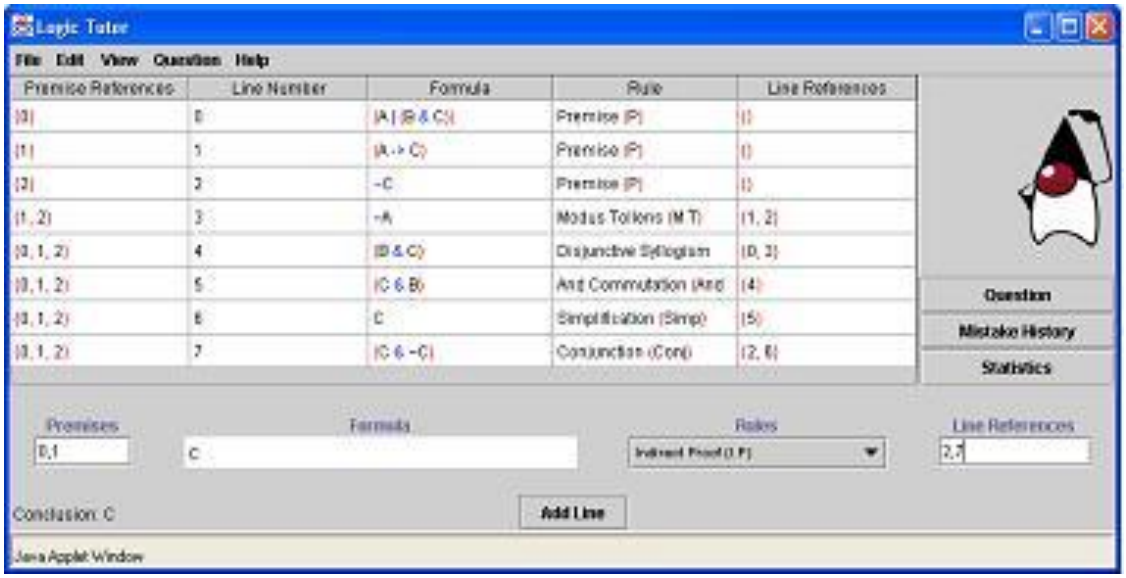
- K-Medoids algorithm with Bayesian Information Criterion (BIC) to infer the optimal number of clusters.
- Distance between utterances: cosine + longest common subsequence for lexical features.
- 7 clusterings according to the previous dialog act of tutors.
- The majority vote in each cluster gives the dialog act.
- A new utterance is predicted according to the cluster with the nearest center.
- Leave-on-Student-out validation: 67% average accuracy, 61% without posture and gesture features.

Relationship Mining

- Two sub-categories:
  - Association rules mining.
  - Correlation mining.

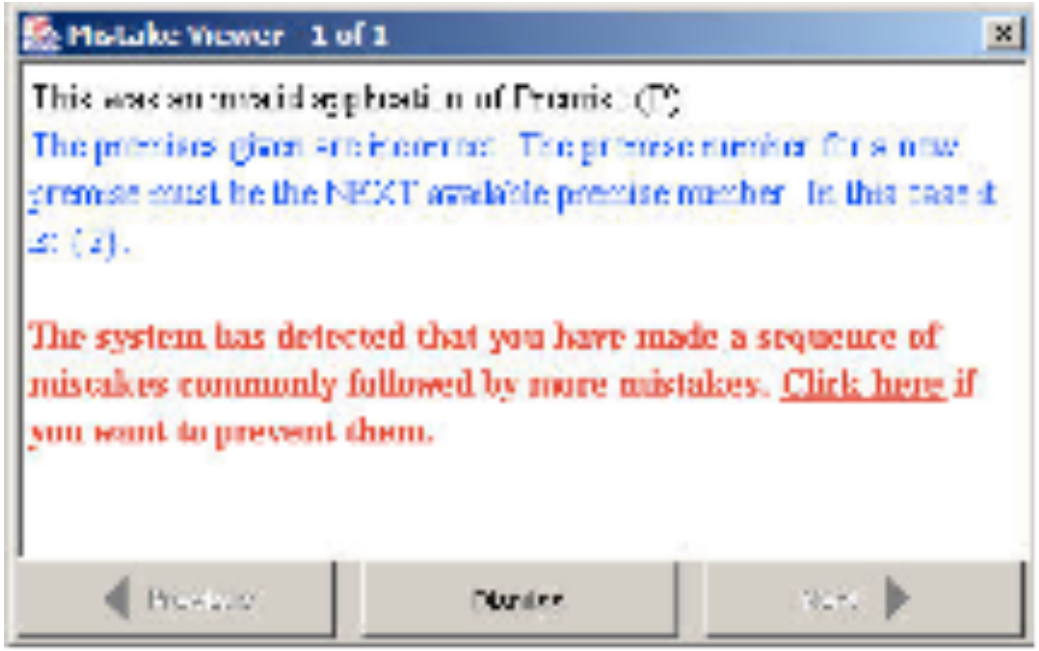
Association Rule Mining

- (Merceron & Yacef 2005) uses the apriori Algorithm to find mistakes often made together while students solve logical proofs exercises in a Logic Tutor:



Association Rule Mining

- Association of mistakes found:
  - Wrong set of premisses -> wrong deduction.
- Enhance the tutor with proactive feedback:



Correlation Mining

- In (Baker & al. 2004) observers code students while using a cognitive tutor:
  - On-task.
  - Off-Task: conversation, something else, inactive or gaming the system.
- Largest correlation found: -0.38 correlation between gaming the system and post-test.

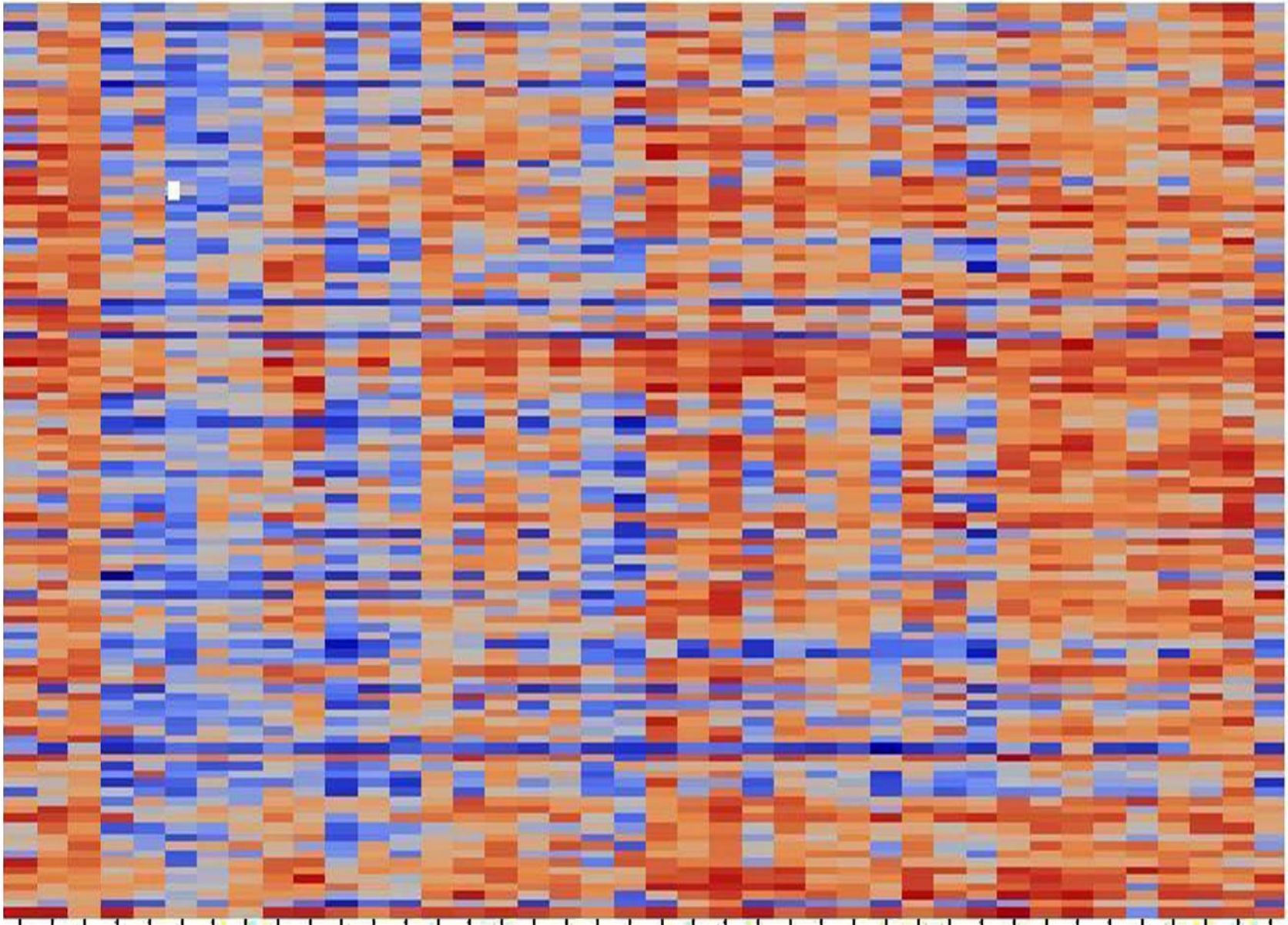
Distillation of Data for Human Judgment

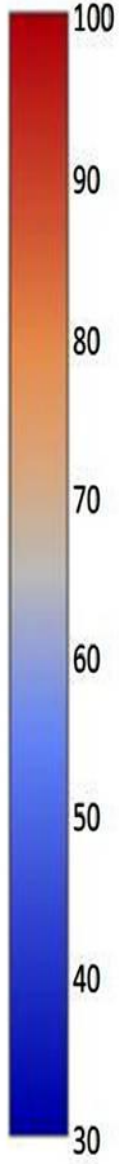
- Preliminary statistics.
- Visualizations. Here too data preparation is crucial.
  - LeMo project (Fortenbacher & al. 2013)

Distillation of Data for Human Judgment

- Heatmap: marks of all students in all courses of a 4 years Bachelor degree, technical university Pakistan:
  - First year courses on the left, then 2nd year courses, 3rd year courses and on the right 4th year courses.

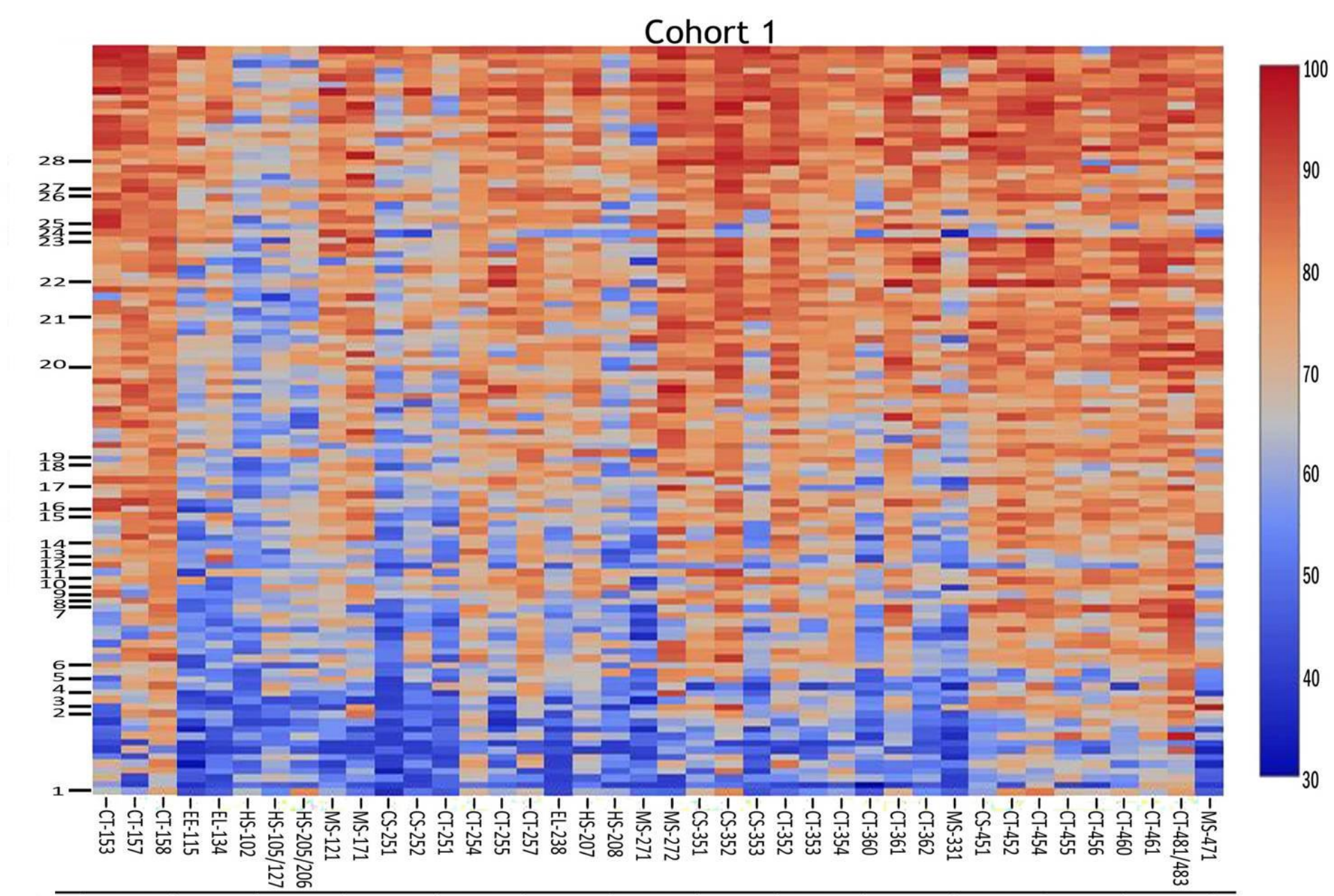
Cohort 1 Heat map with unsorted students





Distillation of Data for Human Judgment

- X-means clustering year wise:
  - Euclidean distance.
  - Tool: Rapid Miner.
- For each year, clusters of students with low marks in all courses, intermediate marks in all courses and high marks in all courses. No cluster with students having high marks in courses A, B, C, intermediate marks in course D and low marks in courses E, F (Asif & al. 2015).
- Heatmap shows now the groups of students with low marks, average marks and high marks, and give hints about courses that could act as detectors.



1- (60,52,54,65)	**2-** (60,52,66,65)	**3-** (60,52,66,77)	**4-** (60,62,54,65)	**5-** (60,62,66,65)	**6-** (60,62,66,77)	**7-** (60,62,66,85)
8- (60,62,75,77)	**9-** (60,73,66,65)	**10-** (60,73,66,77)	**11-** (60,73,75,77)	**12-** (71,52,54,65)	**13-** (71,62,66,65)	**14-** (71,62,66,77)
15-(71,62,66,85)	**16-**(71,62,75,77)	**17-** (71,73,66,77)	**18-** (71,73,75,65)	**19-** (71,73,75,77)	**20-** (71,73,75,85)	**21-** (71,73,83,77)
22- (71,73,83,85)	**23-** (78,62,66,65)	**24-** (78,73,75,65)	**25-** (78,73,75,77)	**26-** (78,73,75,85)	**27-** (78,73,83,77)	**28-** (78,73,83,85)

Discovery with Models

- Building on (Baker & al. 2004), (Baker & al. 2006) proposes a model for gaming the system. Features include:
  - Number of times a specific problem is wrong across all problems.
  - Probability that a student knows a skill.
  - Various times: time taken for the last 3 actions, 5 actions
  - Etc…
- Latent Response Models as statistical basis.
- Generalize to new lessons and new students.
- This detector is used with new data to discover more patters such as in (SanPedro & al. 2015): What happens to students who game the system?

Outline

- Introduction: Big Data in Education
- Methods and Tasks:
  - Prediction
  - Clustering
  - Relationship Mining
  - Distillation of Data for Human Judgment
  - Discovery with Models
- Current Trends
- Conclusions

Current Trends

- Natural Language Processing: tutorial dialogues, essays, forums.
- Multimodal Analysis: data from the educational system + data from camera, from EEG etc.
- Multilevel Analysis: different levels of analysis with the data recorded by the system.

Current Trends

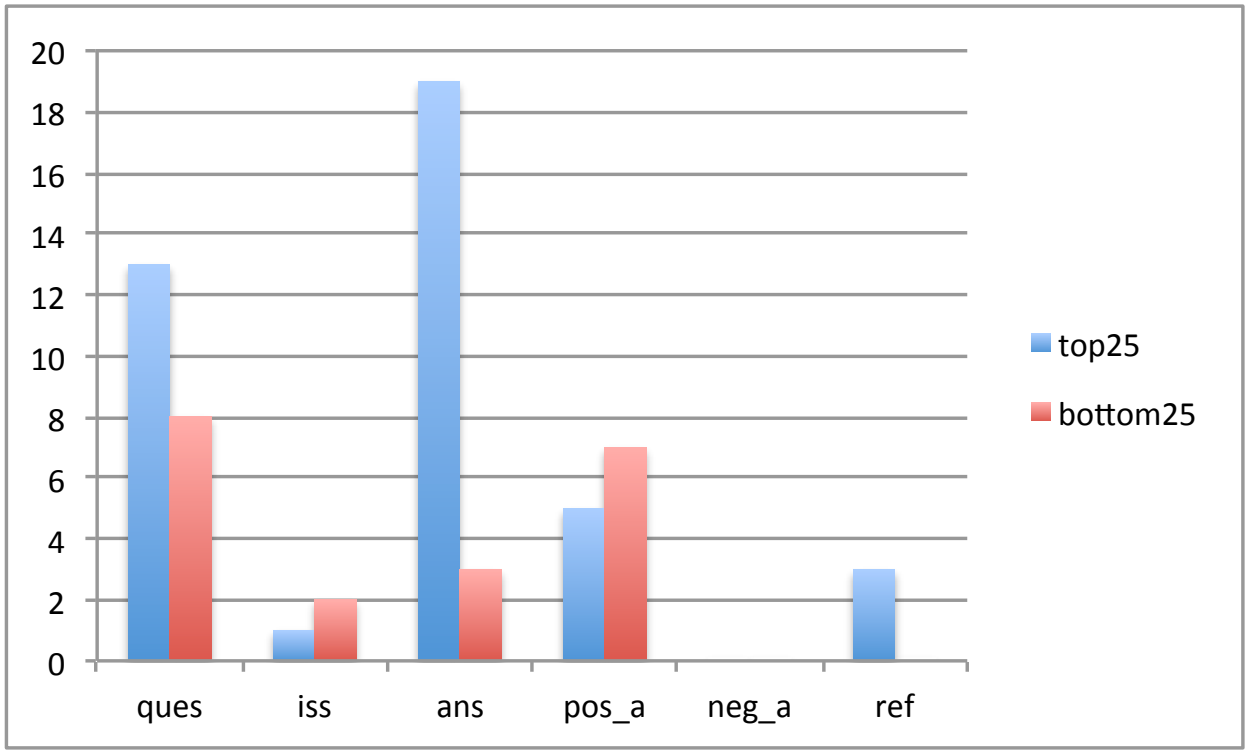
- Multilevel Analysis with Traces (Suthers 2015):
  - Learning platform with chats, forums, file upload and calendar.
  - Contingency graphs show the likelihood that events are related: proximal contingency when 2 events like uploading a file and writing a message occur close in time; lexical contingency when 2 messages have overlap in their vocabulary; etc.
  - Graphs are abstracted and folded, the most abstract level are sociograms representing relations between actors through their contributions.

Current Trends

- On a lower scale: relating level forum and performance level (Merceron 2014) in a programming course of a LMS over 4 years:
  - Posts manually labelled with dialog acts: questions, issues, answers, references, positive and negative acknowledgments.
  - Hypothesis: questions and issues come preliminary from low achieving students.

Current Trends

- After removing an outlier: high achieving students had more questions (and much more answers) than low achieving students.



Outline

- Introduction: Big Data in Education
- Methods and Tasks:
  - Prediction
  - Clustering
  - Relationship Mining
  - Distillation of Data for Human Judgment
  - Discovery with Models
- Current Trends
- Conclusions

Conclusions

- Big data in education is a reality.
- Numerous approaches.
- Numerous tasks.
- Numerous findings.

- What is not a reality yet is the analysis of educational data on a *routine basis* to understand learning and teaching better and to improve them.

Conclusions

- Challenges:
  - Privacy: Opt-in. Limit the available data, hence the findings and validity of the results.
  - Generalizability: is a classifier to predict performance still valid 2 years later, or in another degree? Probably not. Data scientists needed.

Comments? Ideas? Questions?

Thank you for your attention!

References

- (Asif & al. 2014) Asif, R., Merceron, A. and Pathan, M. 2015. Predicting student academic performance at degree level: a case study. In International Journal of Intelligent Systems and Applications (IJSA), Vol. 7(1), 49-61. DOI: 10.5815/ijisa. 2015.01.05.
- (Asif & al. 2015) Asif, R., Merceron, A. and Pathan, M. 2015. Investigating Performance of Students: a Longitudinal Study. In LAK'15, March 16 - 20, 2015, Poughkeepsie, NY, USA. ACM, 108-112.
- (Baker & al. 2004) Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., and Wagner, A.Z. (2004). Off-task behavior in the cognitive tutor classroom: when students "game the system". In: Proceedings of SIGCHI conference on Human Factors in Computing Systems, 383-390. Vienna, Austria.
- (Baker & al. 2006) Baker, R.S.J.d., Corbett, A.T., Roll, I., and Koedinger, K.R. (2006). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287-314.
- (Baker & Yacef 2009) Baker, R.S.J.D., Yacef, K. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions", In *Journal of Educational Data Mining,* Vol. 1(1).

References

- (Barnes 2005) Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In the technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining. 8 p.
- (Cobo & al. 2012) Cobo, G., Garcia, D., Santamaria, E., Moran, J.A., Melenchon, J., Monzo, C. Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In (Dawson, S., Haythornthwaite, C. Hrsg.): Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. (Vancouver, Canada, April 29 – May 2). ACM, 248-251.
- (Ezen-Can & al. 2015) Ezen-Can, A., Grafsgaard, J.F., Lester J.C., Boyer, K. E. (2015) Classifying Student Dialogue Acts with Multimodal Learning Analytics. In LAK'15, March 16 - 20, 2015, Poughkeepsie, NY, USA. ACM, 280-289
- (Fortenbacher & al. 2013) Fortenbacher, A.; Elkina, M.; Merceron, A.: The Learning Analytics Application LeMo – Rationals and First Results. In International Journal of Computing, Volume 12, Issue 3, 2013, ISSN 1727-6209, p. 226-234.
- (Golding & Donaldson 2006) P. Golding, O. Donaldson, “Predicting Academic Performance”, Proceedings of 36th ASEE /IEEE Frontiers in Education Conference, 2006.

References

- (Huang & Fang 2013) Huang, S., Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, Computer and Education, 61, 133-145.
- (Kim & Kim 2010) Kim, J.; Li, J.; Kim. T. Towards Identifying Unresolved Discussions in Student Online Forums. In (Tetreault, J., Burstein, J., Leacock, C. Hrsg.): Proceedings of the NAACL HLT 5th Wokshop on Innovative Use of NLP for Building Educational Applications. (Los Angeles, CA, USA, June 2010). Association for Computational Linguistics, 84 -91.
- (Koedinger & al. 2010) K.R. Koedinger et al., "A Data Repository for the EDM Community: The PSLC DataShop," *Handbook of Educational Data Mining,* CRC Press, 2010, pp. 43–56.
- (Lopez & al. 2012) M. I. Lopez, R. Romero, V. Ventura, and J.M. Luna," Classification via clustering for predicting final marks starting from the student participation in Forums," In (Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. Hrsg.): Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, June15-21, pp. 148-151, 2012.
- (Luengo 2014)Luengo V.; Projet ANR HUBBLE HUman oBservatory Based on anaLysis of e-LEarning traces. https://hal.archives-ouvertes.fr/hal-01116630

References

- (Merceron & Yacef 2005) Merceron, A; Yacef, K. (2005). Educational Data Mining: a case study. In proceedings of Artificial Intelligence in Education (AIED2005) C.-K. Looi, G. McCalla, B. Bredeweg and J. Breuker Eds., 467-474, Amsterdam, The Netherlands.
- (Merceron & Yacef 2010) Merceron, A, Yacef, K. (2010). Measuring correlation of Strong Association Rules in Educational Data. Book chapter in Handbook of Educational Data Mining. 2010. edited by C. Romero, S. Ventura, M. Pechenizkiy & R.S.J.d. Baker, CRC Press, ISBN: 978-1-4398-0457-5, 2010. p. 245 -256.
- (Merceron 2014) Merceron, A. (2014). Connecting Analysis of Speech Acts nd Performance Analysis: a Initial Study. In Proceedings of the Workshop 3: Computational Approaches to Connecting Levels of Analysis in Networked Learning Communities, LAK 2014, Vol-1137
- (Pardos & al. 2007) Z. Pardos, N. Hefferman, B. Anderson, and C. Hefferman, "The effect of Model Granularity on Student Performance Prediction Using Bayesian Networks," Proceedings of the international Conference on User Modelling, Springer, Berlin, pp. 435-439, 2007

References

- (Romero & Ventura 2010) C. Romero, and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE transactions on Systems, Man and Cybernetics,* vol. 40(6), pp.601-618, 2010.
- (San Pedro & al. 2015) San Pedro, M.O., R. Baker, N. Heffernan, J. Ocumpaugh. (2015). What Happens to Students Who Game the System? . In LAK'15, March 16 - 20, 2015, Poughkeepsie, NY, USA. ACM, 36-40.
- (Suthers 2015) Suthers, D. (2015) From Contingencies to Network-level Phenomena: Multilevel Analysis of Activity and Actors in Heterogeneous Networked Learning Environments. In LAK'15, March 16 - 20, 2015, Poughkeepsie, NY, USA. ACM, 368-377.
- (Wolf & al. 2013) A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment," Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 145-149, 2013.
- (Zimmerman & al. 2011) J. Zimmermann, K. H. Brodersen, J. P. Pellet, E. August, J. M. Buhmann, "Predicting graduate-level performance from undergraduate achievements," Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.