

# Issues and Achievements of Computer Science Students by Historical Data Analyses - Are We Ready for Education Big Data?

Ivan Luković,  
University of Novi Sad,  
Faculty of Technical Sciences



---

*15th Workshop DAAD*

# Agenda

---

- Motivation
- Data Analyses - Some references
- EDM in CS Programs – Current Activities
- EDM in CS Programs – Future Activities

# Motivation

---

- Initiating research activities in **Business Intelligence (BI)**
  - by deploying interdisciplinary knowledge and often complex technologies
- Our experience shows
  - **a lack of well educated experts**, capable of covering all necessary disciplines
    - **Computer Science**, including System Architecture Design, Languages, Databases, Artificial Intelligence, Data Mining, Formal Methods in Software Engineering
    - **Mathematics**, including Logic, Statistics and Optimization Methods
    - **Management, Organization and Psychology**
    - **Problem domain knowledge**



# Motivation

- Initiating new study programs at **B.Sc. and M.Sc. levels** in
  - **Information Engineering** and
  - **Information and Analytics Engineering**
  - three programs already approved at National Accreditation Committee of Republic of Serbia
    - (talks from the last two year workshops)
- To support these initiatives
  - **new research projects** in the area of **BI** and **Historical Data Analyses (HDA)** are necessary



# Motivation

---

- New BI and HDA projects – nice idea, but...
- Evident practical problems
  - our industry **is not often ready** for such projects
    - despite many declarations that such kind of projects are important and necessary for a success in business
  - even, if there is a clear declaration to initiate such industry projects, **practical difficulties arise**
    - preventing real development of such projects
- Typical examples
  - inability to create a good development team
  - lack of expert knowledge
  - inability to provide and collect historical data



# Motivation

---

- **Inability to provide and collect historical data**
  - always complains about lack of data for HDA
  - how to obtain data - **a hardest issue in performing data analyses**
- Some aspects of the problem
  - **psychological**
    - not a real willingness of stakeholders to provide data
  - **technical**
    - poorly organized and stored data, heterogeneity of data sources, poor data trustfulness
  - **law and information confidence**
    - a fear of law or user rights violation



# Agenda

---

- Motivation
- Data Analyses - Some references
- EDM in CS Programs – Current Activities
- EDM in CS Programs – Future Activities

# Data Analyses - Some references

---

- Some our research activities **in the area of Health Care**
  - great possibilities **for lowering expenses** by relatively cheap methods based on BI projects
  - by our experience, **very poorly exploited** in Serbia
  - an attempt with the Institute for Healthcare of Republic of Serbia in Belgrade
    - quite unsuccessful
    - not a real willingness to cooperate
  - the new attempt with the Institute for Healthcare of Autonomous Province of Vojvodina (APV) in Novi Sad
    - with partial success
    - we have obtained data about patient absentisms





# Data Analyses - Some references

- References with the Inst. for Healthcare of APV
  - Ivančević V., Knežević M., Simić M., Mandić D., Luković I: ***Dr Warehouse - An Intelligent Software System for Epidemiological Monitoring, Prediction, and Research***; 5th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), Seville, Spain, 2013. (Best paper award)
  - Ivančević V., Knežević M., Simić M., Luković I., Mandić D: ***Public Healthcare and Epidemiology with Dr Warehouse***; International Journal on Advances in Software, International Academy, Research, and Industry Association (IARIA), ISSN: 1942-2628, Vol. 6, No. 3 & 4, 2013, pp. 329-342.
- nice, but further attempts to continue collaboration were unsuccessful
  - not a real willingness to cooperate
  - they feel themselves better in the usage of Excel tables



# Data Analyses - Some references

---

- The next example in the child dental healthcare
  - relatively small amount of provided data
    - with very big efforts invested by individual dentists to collect all data about patients
- Reference (in press)
  - Ivančević V., Tušek I., Tušek J., Knežević M., Elheshk S., Luković I: ***Using Association Rule Mining to Identify Risk Factors for Early Childhood Caries***; Computer Methods and Programs in Biomedicine, Elsevier, ISSN: 0169-2607, DOI: 10.1016/j.cmpb.2015.07.008, 2015.
  - finding the most important factors influencing ECC
    - Cost: 100 ~ 1000 EUR / 1 patient
    - Over 30% of child population with ECC



# Agenda

---

- Motivation
- Data Analyses - Some references
- EDM in CS Programs – Current Activities
- EDM in CS Programs – Future Activities

# EDM in CS Programs – Current Activities

---

- **Educational Data Mining (EDM)**
  - “a perfect” area for initiating BI and HDA projects
  - as we generate, store and collect our own data
    - by this, become independent of other stakeholders
- However, the important issues are
  - Do we have a systematic way of collecting, storing and analyzing "fine" data about students' performance
  - How often we analyze our historical data about students and their results
    - quantitative analyses vs. speculative approach



# EDM in CS Programs – Current Activities

---

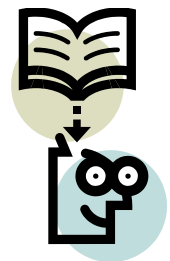
- **Current situation at the Faculty of Technical Sciences (FTS)**
  - **operational data about student performance**
    - systematically collected at the level of final exam, by Student Service
    - finer granularity of data, e.g. at the level of each week, is not systematically collected
      - it is a subject of each lecturer
  - **aggregated data about student performance**
    - not systematically derived and stored
    - rarely and not systematically used for some data analyses



# EDM in CS Programs – Current Activities

---

- **Our research initiatives in the data collection and analysis**
  - **(A)** from our “massive courses”, as in **Programming Languages**, at the 1<sup>st</sup> year of study
    - comprising about 200 students per year
    - about students’ behavior and performances in a computer laboratory
    - by analyzing the appropriate log files and obtained results
  - **(B)** from our courses **Databases 1** and **Databases 2**
    - after dividing the former unified **Databases** course (4<sup>th</sup> year) into the two new ones (3<sup>rd</sup> and 4<sup>th</sup> year)
    - comprising more than 100 students per year
    - by analyzing logs with students’ performance during the semester and also higher years of study



# EDM in CS Programs – Current Activities

- Research initiative **(A)** - References
  - Ivančević V., Čeliković M., Luković I.: ***Analysing Student Spatial Deployment in a Computer Laboratory***, 4th International Conference on Educational Data Mining (EDM), Eindhoven, the Netherlands, 2011.
  - Ivančević V., Čeliković M., Luković I.: ***The Individual Stability of Student Spatial Deployment and its Implications***, XIV International Symposium on Computers in Education (SIIE), Andorra la Vella, Andorra, 2012.
  - Ivančević V., Knežević M., Luković I.: ***Academic Achievement and Choices of Computing and Control Engineering Students in relation to Gender***, 41th SEFI Conference, Leuven, Belgium, 2013.



# EDM in CS Programs – Current Activities

---

- Research initiative **(A)** – Some findings
  - Application of DM techniques for the investigation of student seating arrangements in the classroom
  - With the increased location distance from the instructor, scores tend to drop in test score
  - Border locations are outliers in terms of associated test scores and occupancy
  - Students who do not change the seating location have, on average, a one grade level higher score than the others
  - Students with higher levels of spatial consistency have around 10% higher assessment scores when compared to those of low spatial consistency





# EDM in CS Programs – Current Activities

---

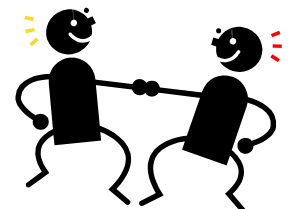
- Research initiative **(A)** – Some findings
  - Nearly one third of the students never changed their seat during the semester
  - Female students have better spatial consistency when compared to male students
  - Female students complete slightly more assignments
  - Female students outperform male students in mathematics and advanced scientific/academic courses
  - Male students have a slight advantage in programming courses
  - Many students gravitate towards such behaviour



# EDM in CS Programs – Current Activities

---

- Research initiative **(B)** - References
  - Ivančević V., Knežević M., Čeliković M., Aleksić S., Luković I.: **Database Courses: Curriculum Changes and Student Results**, 6th PSU-UNS International Conference on Engineering and Technology (PSU-UNS ICET), Novi Sad, Serbia, 2013.



# EDM in CS Programs – Current Activities

---

- Research initiative **(B)** – Some findings
  - Over the years, the distribution of scores in theoretical assessments has become right-tailed and the mean value has decreased, i.e., more students have been achieving lower scores
  - Very high theory scores imply comparably good practical scores, oral exam performance, and attendance
    - students who master theoretical foundations with a high or very high score have no problem mastering the practical aspects of databases, as well as the complete course



# EDM in CS Programs – Current Activities

---

- Research initiative **(B)** – Some findings
  - Decision to split the original course and move its modules to a lower year of study has not negatively impacted student performance
  - The new advanced course **Databases 2** has led to largely improved theory scores



# Agenda

---

- Motivation
- Data Analyses - Some references
- EDM in CS Programs – Current Activities
- EDM in CS Programs – Future Activities

# EDM in CS Programs – Future Activities

---

- For more than 10 years, at FTS **various questionnaires** are collected
  - Students’ satisfiability questionnaires
    - at the end of each semester
    - at the end of studies
  - Employees’ satisfiability questionnaires
    - each year
  - Data Analyses are performed “on the fly”
    - relatively primitive statistic analyses
    - after publishing reports, data are archived and never used again
  - **HDA have been never performed!**
    - even, there is no awareness about its importance



# EDM in CS Programs – Future Activities

---

- Our initiative – establishing an **EDM BI research project** through at least one Ph.D. thesis
  - HDA of students' questionnaires, spanning 10 years
  - Explore the importance of many factors, influencing students' satisfiability with the education process
  - Establish the appropriate IT infrastructure for HDA
  - Perform Data Mining, as well as Text Mining (sentiment) analyses
    - as we have both well structured data and poorly structured textual information, in the form of students' comments
  - The main problem, again
    - obtaining **a permission to get historical data**
  - After months of attempts, we succeeded!



# Final words

---

- It seems that the application of BI and HDA may bring a great value to the business
  - also in the area of high education
- Promising technologies and approaches already exist
- Huge amounts of operational data are already collected, in a less or more better way
- However, we are still far from real practical benefits





**End of Presentation**

# Issues and Achievements of Computer Science Students by Historical Data Analyses - Are We Ready for Education Big Data?

Ivan Luković,  
University of Novi Sad,  
Faculty of Technical Sciences



---

*15th Workshop DAAD*