Developing a generalizable detector of when students game the system

Type	Journal Article
Author	Ryan S. J. d Baker
Author	Albert T. Corbett
Author	Ido Roll
Author	Kenneth R. Koedinger
URL	http://link.springer.com/article/10.1007/s11257-007-9045-6
Volume	18
Issue	3
Pages	287-314
Publication	User Modeling and User-Adapted Interaction
ISSN	0924-1868, 1573-1391
Date	2008/08/01
Journal Abbr	User Model User-Adap Inter
DOI	10.1007/s11257-007-9045-6
Accessed	12/19/2016, 3:46:10 PM
Library Catalog	link.springer.com
Language	en
Abstract	Some students, when working in interactive learning environments, attempt to "game the system", attempting to succeed in the environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly. In this paper, we present a system that can accurately detect whether a student is gaming the system, within a Cognitive Tutor mathematics curricula. Our detector also distinguishes between two distinct types of gaming which are associated with different learning outcomes. We explore this detector's generalizability, and find that it transfers successfully to both new students and new tutor lessons.
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Notes:

ABSTRACT

Some students, when working in interactive learning environments, attempt to "game the system", attempting to succeed in the environment by exploiting properties of the system rather than by learning the material and trying to use that
knowledge to answer correctly. In this paper, we present a system that can accurately detect whether a student is gaming the system, within a Cognitive Tutor mathematics curricula. Our detector also distinguishes between two distinct types of gaming which are associated with different learning outcomes. We explore this detector's generalizability, and find that it transfers successfully to both new students and new tutor lessons.

An action is assessed to be gaming or not, by a function on parameters composed of the features drawn from each action's characteristics.

Generalization: A detector that transfers to new populations of students, within a single tutor lesson, can be used as the basis for a system that responds to gaming the system, within that lesson.

Attachments

- Full Text PDF
- Snapshot

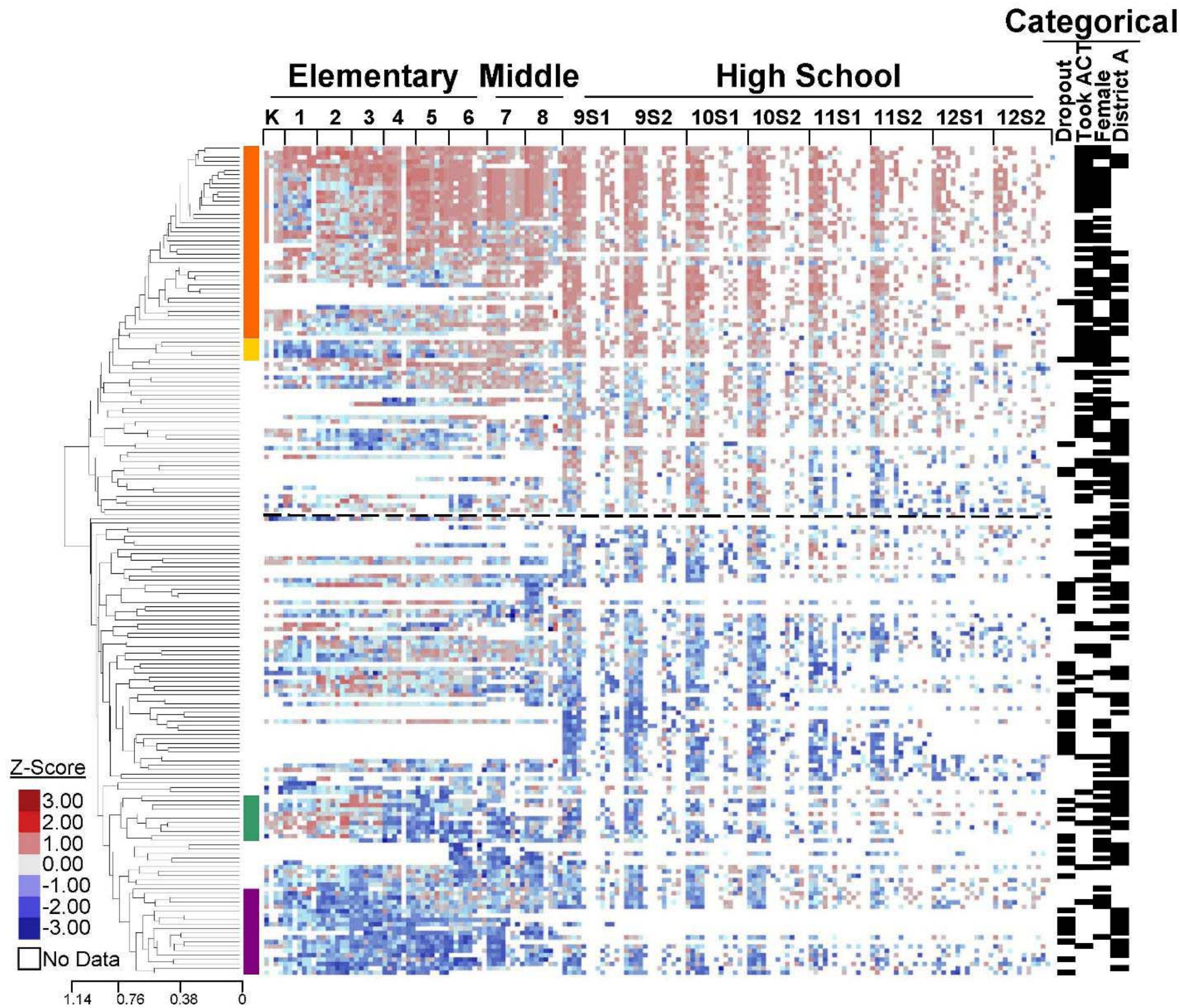
Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis

Type	Journal Article
Author	Alex J. Bowers
URL	http://www.pareonline.net/pdf/v15n7.pdf
Volume	15
Issue	7
Pages	1–18
Publication	Practical Assessment Research and Evaluation
Date	2010
Accessed	12/21/2016, 2:20:43 PM
Library Catalog	Google Scholar
Short Title	Analyzing the longitudinal K-12 grading histories of entire cohorts of students
Date Added	12/21/2016, 2:20:43 PM
Modified	12/21/2016, 2:20:43 PM

Notes:

The central purpose of this study is to introduce hierarchical cluster analysis and pattern visualization methods from the data mining literature and demonstrate the method's utility through one example, identification of student dropout from student K-12 longitudinal grades.

- For educational data, the method provides a useful and interesting means to visualize and assess an entire disaggregated data history pattern for a student in comparison with every other student's data pattern in a sample.
- The clustergram allows for the visualization and interpretation of every data point. Each student's data pattern is proximal in the clustergram to students with similar patterns, facilitating system-wide analysis and identification of specific clusters in the dataset.



Attachments

- v15n7.pdf

  The central purpose of this study is to introduce
  hierarchical cluster analysis and pattern visualization
  methods from the data mining literature and demonstrate the method's utility through one example, identification of student dropout from student K-12 longitudinal grades. For educational data, the method provides a useful and interesting means to visualize and assess an entire disaggregated data history pattern for a student in comparison with every other student's data
  pattern in a sample. The clustergram allows for the visualization and interpretation of every data point. Each student's data pattern is proximal in the clustergram to students with similar patterns, facilitating system-wide analysis and identification of specific clusters in the dataset.

Data wranglers: human interpreters to help close the feedback loop

Type	Conference Paper
Author	Doug Clow
Publisher	ACM
Pages	49–53
Date	2014
Proceedings Title	Proceedings of the Fourth International Conference on Learning Analytics And Knowledge
Date Added	12/21/2016, 2:10:18 PM
Modified	12/21/2016, 2:10:18 PM

Notes:

NOTES

ABSTRACT

Closing the feedback loop to improve learning is at the heart of good learning analytics practice. However, the quantity of data, and the range of different data sources, can make it difficult to take systematic action on that data. Previous work in the literature has emphasised the need for and value of human meaning-making in the process of interpretation of data to transform it in to actionable intelligence.

This paper describes a programme of human Data Wranglers
deployed at the Open University, UK, charged with making sense of a range of data sources related to learning, analysing that data in the light of their understanding of practice in individual faculties/departments, and producing reports that summarise the key points and make actionable recommendations.

The evaluation of and experience in this programme of work
strongly supports the value of human meaning-makers in the
learning analytics process, and suggests that barriers to
organisational change in this area can be mitigated by embedding learning analytics work within strategic contexts, and working at an appropriate level and granularity of analysis.

Using data for direct observations may use human interventions for interpretation.

- Internet based learning has a lot of data, this technique is useful in that domain.
- Implementation requires more than skill. Buy-in from stakeholders and policy makers is important.
- Data wranglers should work alongside domain experts to ensure accurate interpretation.
- Making scalable environments is a concern.
- Bad data management is a concern.
- Interpretation can be a concern (student insights vs performance of the system).

Attachments

- Clow-DataWranglers-final.pdf

ABSTRACT

Closing the feedback loop to improve learning is at the heart of
good learning analytics practice. However, the quantity of data,
and the range of different data sources, can make it difficult to
take systematic action on that data. Previous work in the literature has emphasised the need for and value of human meaning-making in the process of interpretation of data to transform it in to actionable intelligence.

This paper describes a programme of human Data Wranglers
deployed at the Open University, UK, charged with making sense
of a range of data sources related to learning, analysing that data in the light of their understanding of practice in individual
faculties/departments, and producing reports that summarise the
key points and make actionable recommendations.

The evaluation of and experience in this programme of work
strongly supports the value of human meaning-makers in the
learning analytics process, and suggests that barriers to
organisational change in this area can be mitigated by embedding
learning analytics work within strategic contexts, and working at
an appropriate level and granularity of analysis.

Key: the value of human "meaning-making" in this data driven automated space.

Take Aways:

1. **we're seeing great applications for this in regards to internet based learning.**
2. **Needs a bottom up approach for sense making at the same time the has to buy in and push the agenda.**
   Scalable efforts are important
3. Educational systems, leaders and academics need to be supported and encouraged to interpret and design LA through a contextual framework.
4. Communities of practice
5. Data wranglers must work in tandumn with academics exchanging expertise for all to improve
6. This is time consuming
7. the process can be hampered by poor data management

Human "data wranglers" from the UK charged with the task of summarizing a range of data sources and making sense of what they learned and analyzed - in light of their understanding of practice in diff. departments of study.

Then by summarizing key points make actionable suggestions.

Learning is complex the tools we uses to measure and analyze learning and learners are complex thus we need teams of people across disciplines to make sense of the info and find actions that close the feedback loop.

How:

1. Scalable efforts are important
2. Educational systems, leaders and academics need to be supported and encouraged to interpret and design LA through a contextual framework.
3. Communities of practie

** there will be pushback --> this could be come a metrics agenda. **

Big Data in Education 2.4 Diagnostic Metrics: Correlation

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=7r3hfJW1gz0&feature=youtu.be
Accessed	12/21/2016, 1:09:17 PM
Library Catalog	YouTube
Running Time	519 seconds
Short Title	Big Data in Education 2.4 Diagnostic Metrics
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Validation and Correlation/Regressions

0.3 is a decent correlation for education data (relative to data/field).

R^2 is used for model goodness/Prediction of linear models.

Mean-Absolute Deviation Error - How far the model predicts compared to the expected absolute value.

Low Root Mean Squared Error with High Mean-Absolute Deviation is a good model.

Look at scatter plot or heat map for identifying outliers.

Validation metrics can be biased, use multiple methods.

Bayesian (BiC) - good for interpretability. Similar to K-Fold cross Validation.

AiC - Akaikes Information: trade off of goodness of fit and flexibility of fit. Prone to over-fitting.

Big Data in Education 1.3 Classifiers, Part 1

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=k9Z4ibzH-1s&feature=youtu.be
Accessed	12/21/2016, 12:23:25 PM
Library Catalog	YouTube
Running Time	636 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Prediction Modes can be used to predict the future and make inferences about the present.

Classifier models are a type of prediction model.

- The "Label" is the thing you want to predict - the answer is a set of categories, not a number.
- Labes can come from
- Software performance
- school records
- Test data
- Survey data
- Field Observations
- Raw text

"Features" (variables) are used to predict the label.

There are hundreds of classification algorithms

Which combination of features build the best classification model?

No one knows why some algorithms are better than others between domains.

- These algorithm are useful in specific; education:
  - step regression
    - binary classifications (0,1)
    - conservative
    - **not good when interaction common**
  - logistical regression
    - predictor variables
    - conservative
    - good for where changes in predictor var. have an impact on probability of predicted variable class
    - **not good when interaction common**
  - J48/ C4.5 Decision Trees
    - **Good for interactions**
    - **open source**
    - **uses numerical and categorical**
    - **relatively conservative**
  - JRIP Decision Rues
  - K*Instance-Based Classifiers

Big Data in Education 2.3 Diagnostic Metrics Part 2

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=9PDwRdyb6Sw&feature=youtu.be
Accessed	12/21/2016, 12:50:16 PM
Library Catalog	YouTube
Running Time	723 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

ROC

Something with two outcomes and you're hoping to know how good your model is.

Receiver-Operator Characteristic curve (ROC).
Used to determine correct/incorrect, gaming the system, dropout/retainment. Only predicts 2 values (0 and 1, Yes and No).

Four possibilities

1. True Positive - both model and data say it's 1.
2. False Positive - data says it's 0 but the model says it's 1.
3. True Negative - model and data say it's 0.
4. False Negative - data says it's 1 but the model says it's 0.

Big Data in Education 1.1 Introduction

Type Video Recording
Director ColumbiaLearn
URL https://www.youtube.com/watch?v=dc5Nx3tyR8g&feature=youtu.be
Accessed 12/20/2016, 12:04:21 PM
Library Catalog YouTube
Running Time 417 seconds
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Notes:

Goals are to promote:

- New science/discoveries
- Better assessment of learners
- Create/improve real-time support.

Methods come from:

- Data Mining/Machine Learning Domains
- Psychometrics/Traditional Statistics Domain

Trends in Education Data

- Ed Data trends do not match industry trends (not "Google Big".
- Neural Networks aren't used frequently
- They tend to "overfit"
- key types of EDM is
  - Predictive Modeling
  - Structure Discovery/ clustering
  - knowledge engineering

Big enough that r^2 is frequently found as statistically significant at 0.0019

EDM is just emerging because the data is now available and we have the technology. Data used to be summative, but now lots of students use ed software at a fine grain scale.

Big Data in Education 6.1 Learning Curves

Type Video Recording
Director ColumbiaLearn
URL https://www.youtube.com/watch?v=Mr17Z0nZUQc&feature=youtu.be
Accessed 10/25/2016, 9:34:17 PM
Library Catalog YouTube
Running Time 504 seconds
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Notes:

Learning Curves create insights into the structure of data (e.g. reveal levels or easy comparisons).

Visualizations in Education Data:

Tags: Big Data, Big Data in Education, data analysis, Learning, Learning analytics, visualization

Big Data in Education 7.6 Knowledge Inference: Q-Matrix Knowledge Structure

Type Video Recording
Director ColumbiaLearn
URL https://www.youtube.com/watch?v=oFSV6-opnws&feature=youtu.be
Accessed 12/21/2016, 12:13:43 PM
Library Catalog YouTube
Running Time 527 seconds
Short Title Big Data in Education 7.6 Knowledge Inference
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Notes:

Q-Matrix are setup as a row for each item and a column for each skill. Model is looking at assessing performance of each observed item by student on each skill.

Barnes model: are skills conjunctive - need all the skills or are they compensatory (only need of the skills)

John Stamper worked on smoothing the learning curve for skills discovery. If a random spike is found, two skills are most likely being treated as one in the model.

PSLC has tools for this.

Big Data in Education 2.2 Diagnostic Metrics Part 1

Type Video Recording
Director ColumbiaLearn
URL https://www.youtube.com/watch?v=fGMFYTHhcHg&feature=youtu.be
Accessed 12/21/2016, 12:42:45 PM
Library Catalog YouTube
Running Time 517 seconds
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Notes:

Accuracy is a poor metric for assessment b/c the metrics are weighted different.

We use **Kappa**

- Kappa = 1 - perfect agreement
- Kappa = 0 - agreement is a chance
- Kappa = -1 -agreement is the inverse (model is exact opposite)
- Kappa > 1 = you messed up
- Kappa < 0 = model is worse than chance
- 0<Kappa<1 = For data mined model, 0.3 - 0.5 is considered good enough to call the model better than chance and publishable.
- There is no standard for Kappa thresholds of success. Sometimes you can use 0.8. But it's influenced by the data set.

Kappa is the actual agreements minus the expected agreements / divided by the total possibility of agreement minus the expected agreements.

Big Data in Education 7.2 Validation and Selection

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=B9dvJYwBfmk&feature=youtu.be
Accessed	12/21/2016, 11:41:53 AM
Library Catalog	YouTube
Running Time	322 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Euclidean distance: distance from A to B in two dimensions
Cross validations doesn't work with determining centroids.

Best value of K with BiC or AIC is a good combination.

Tags: Cluster analysis, k-means clustering

Big Data in Education 6.4 State Space Diagrams

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=WI1AVcpCYgk&feature=youtu.be
Accessed	10/25/2016, 9:53:01 PM
Library Catalog	YouTube
Running Time	241 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

State Space Diagrams.

- State = a complete characterization of the situations
- Also known = student learning pathways

State Space Diagram "Refractions" (example) -> Game

- Can show how a student or groups go through a process or fail to go through a process.
- Groups: you can see a "logic" path
  - How common a people take particular pathways.
- Use:
  - Study specific student trajectories
  - study productive paths
  - study rare but productive paths
  - Make recommendations to students based on their path (Hints)

Big Data in Education 6.3 Scatter Plots

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=oTiixxmh9-Q&feature=youtu.be
Accessed	10/25/2016, 9:44:07 PM
Library Catalog	YouTube
Running Time	406 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Visualization - Outliers and Data Structure

Scatter plots don't scale to big data sets. Preference for smaller/traditional data only.

Heat Map are better for large scale data. They indicate intensity and density

Parameter Space Maps are a type of heat map. They can indicate how often a proportion of skills will be the best.

Tags: Big Data, Big Data in Education, data analysis, heat maps, scatter plots, visualization

---

Big Data in Education 7.1 Clustering

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=mgXm3AwLxP8&feature=youtu.be
Accessed	12/21/2016, 11:28:15 AM
Library Catalog	YouTube
Running Time	820 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Structure Discovery Algorithms can reveal the structure of the data when it is unknown through finding previously unknown groupings/characteristics between data.

Process for K-Means Clustering

1. Select the number of Clusters
2. Choose random number of centroids
3. Define initial clusters w/ a voyanod diagram
4. Re-fit the centroids
5. Repeat until convergence - where the centroid is in the center.

Tags: Cluster analysis, k-means clustering

Big Data in Education 2.5 Cross-Validation and Over-Fitting

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=1P34cxpEdKA&feature=youtu.be
Accessed	12/21/2016, 1:34:08 PM
Library Catalog	YouTube
Running Time	419 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Over-Fitting is when you fit the model to the data so it accounts for the extra noise. Cross-validation can test to see if the model is good or if it fits to the noise.

Models that are over fit are bad for new data.

Ways to avoid:

- Fewer variables (BiC, AIC, Occam's Razor).
- Less Complex Functions (MDL)

Over-fitting cannot be eliminated, only reduced.

Check to see how generalizable the model is. build on some data then test on the rest. Or use cross-validation to train the model by testing groups of data against a control.

K-Fold - pick a specific split of numbers. You select the folds. But this is faster.

Leave-out one - Every data is a fold.

Cross Validation Variant

Flat Cross - each point has equal chance of being placed into fold.

Stratified cross - biases fold elections to create equality among variables in fold.

If you plan to use data from NEW STUDENTS - Student-Level Cross Validation is important.

Big Data in Education 1.4 Classifiers Part 2

Type	Video Recording
Director	ColumbiaLearn
URL	https://www.youtube.com/watch?v=8X0UlMShss4&feature=youtu.be
Accessed	12/21/2016, 12:33:35 PM
Library Catalog	YouTube
Running Time	496 seconds
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Notes:

Algorithms for decision rules can function as if/then rules.

JRip and Part

Create many if/then statements then take the best functioning statements out of the trial/error.

K* ("K Star")

K* is good when the data is very divergent. Data that are similar tend to be in the same classification. Looks where data falls and the characteristics of the nearest data points.

Sometimes works when nothing else does.

Bagged Stumps

Related to decision tree. Creates lots of trees with only the first feature. A close variant is Random Forest.

Random Forest

Create lots of trees then aggregates them.

ALL GOOD BECAUSE THEY ARE SIMPLE.

Other, more complex, methods

Support Vector Machines

Genetic Algorithms

Neural Networks - hard to interpret

These methods have not produced the best solutions from working with educational data.

Knowledge tracing: Modeling the acquisition of procedural knowledge

Type	Journal Article
Author	Albert T. Corbett
Author	John R. Anderson
URL	http://link.springer.com/article/10.1007/BF01099821
Volume	4
Issue	4
Pages	253-278
Publication	User Modeling and User-Adapted Interaction
ISSN	0924-1868, 1573-1391
Date	1994/12/01
Journal Abbr	User Model User-Adap Inter
DOI	10.1007/BF01099821
Accessed	12/19/2016, 3:46:26 PM
Library Catalog	link.springer.com
Language	en
Abstract	This paper describes an effort to model students' changing knowledge state during skill acquisition. Students in this research are learning to write short programs with the ACT Programming Tutor (APT). APT is constructed around a production rule cognitive model of programming knowledge, called theideal student model. This model allows the tutor to solve exercises along with the student and provide assistance as necessary. As the student works, the tutor also maintains an estimate of the probability that the student has learned each of the rules in the ideal model, in a process calledknowledge tracing. The tutor presents an individualized sequence of exercises to the student based on these probability estimates until the student has 'mastered' each rule. The programming tutor, cognitive model and learning and performance assumptions are described. A series of studies is reviewed that examine the empirical validity of knowledge tracing and has led to modifications in the process. Currently the model is quite successful in predicting test performance. Further modifications in the modeling process are discussed that may improve performance levels.
Short Title	Knowledge tracing
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Notes:

This paper describes an effort to model students' changing knowledge state during skill acquisition. Students in this research are learning to write short programs with the ACT Programming
Tutor (APT). APT is constructed around a production rule cognitive model of programming knowledge, called the ideal student model. This model allows the tutor to solve exercises along with the student and provide assistance as necessary. As the student works, the tutor also maintains an estimate of the probability that the student has learned each of the rules in the ideal model, in a process called knowledge tracing. The tutor presents an individualized sequence of exercises to the student
based on these probability estimates until the student has 'mastered' each rule. The programming tutor, cognitive model and learning and performance assumptions are described. A series of studies is reviewed that examine the empirical validity of knowledge tracing and has led to modifications in the
process. Currently the model is quite successful in predicting test performance. Further modifications in the modeling process are discussed that may improve performance levels.

Conclusion

> "We started this research effort with an executable cognitive model of Lisp programming skill. This model consisted of a set of ideal production rules and formed the core of a successful intelligent programming tutor. The goal of the research was to implement a simple student modeling process that would allow the tutor to monitor the student's knowledge state and tailor the sequence of practice exercises to the student's needs. The resulting knowledge tracing process models the student
> as an overlay of the production rules and the mastery-based curriculum structure allows us to associate each programming action with a single production rule. A simple two-state learning model enables us to estimate the student's knowledge
> state from performance and predict performance from that knowledge state. Successive evaluations led us to (1) abandon an initial ideal student model and to model a sufficient set of rules, (2) to model differences in rule difficulty and (3) to model individual differences among students in learning and performance. The resulting model predicts student performance quite well and enables most students
> to reach a high level of task performance. It may be possible to improve on this level of performance and enable more students to reach mastery by manipulating incentive in testing, by providing additional procedural practice or by monitoring and remediating students' knowledge of key declarative concepts."

Tags: Knowledge Tracing

Attachments

- Full Text PDF
- Snapshot

Using data mining to predict secondary school student performance

Type	Journal Article
Author	Paulo Cortez
Author	Alice Maria Gonçalves Silva

URL http://repositorium.sdum.uminho.pt/handle/1822/8024
Date 2008
Accessed 12/21/2016, 2:41:49 PM
Library Catalog Google Scholar
Date Added 12/21/2016, 2:41:49 PM
Modified 12/21/2016, 2:41:49 PM

Attachments

- USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

Tags:

business analytics, data analysis, Data mining

Rob's Notes:

Using BI/ Data Mining to help students in Portugal. Using real world data,, student grades, demographic info, SES. Collected via questionnaire.

Models used (Decision tree, random forest, neutral networks, support vector machines).

The results:

- a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available.
  - Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption).
- As a direct outcome of this research, more efficient student prediction tools can be be developed, improving the quality of education and enhancing school resource management.

i) binary classification (pass/fail);

ii) classification with five levels (from I very good or excellent to V - insufficient); and

iii) regression, with a numeric output that ranges between zero (0%) and twenty (100%).

BI can capture information that is overlooked by human being due to lack of expertise or human error.

Its important to capture demographic information along with school performance over time (years over years).

When the systems in play don't have the infrastructure to capture the things you need it is necessary to generate your own data through questionnaires.

decision tree = IF-THEN rules

Final Thoughts: I can not remember the journal but there was an entire section discussing the various vulnerabilities associated with EDM. Specifically the situational vulnerability. Which refer to systems tracking student performances and assigning classifications to based on an event in the students life that is temporary (death in in fam, etc.), this leads to permanent limitations of that student.

Zuckerberg is ploughing billions into 'personalised learning' – why?

Type Web Page
Author Elizabeth FitzGerald
Author Natalia Kucirkova
URL http://theconversation.com/zuckerberg-is-ploughing-billions-into-personalised-learning-why-51940
Accessed 12/19/2016, 3:45:46 PM
Abstract Zuckerburg wants to plough billions into personalised learning, but his way may not be the right way.
Website Title The Conversation
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Notes:

Three flaws were proclaimed:

1. Education has always been about acquiring knowledge and skills relevant to a profession, but also about acquiring general knowledge. By feeding children only the content they're interested in, we may end up with many specialists and few generalists.
2. While learners may cope poorly with trying to learn in a way that's not suited to them, in the real world life will not always be so accommodating. Their lack of ability to compensate may mean they suffer as a result.
3. Children's preferences are not fixed. To predict content relevant for children there needs to be sensitive, human-directed input – not automation. Otherwise we end up with what might be called de-personalised learning, and classrooms with little conversation between student and teacher.

In subcontracting out teaching to technology, the risk is that the valuable social contact between students, teachers and parents that's inherent to effective learning will be reduced.

I do not agree with the authors opinion.

Attachments

- Snapshot

Translating Learning into Numbers: A Generic Framework for Learning Analytics

Type Journal Article
Author Wolfgang Greller
Author Hendrik Drachsler
URL http://search.proquest.com.eduproxy.tc-library.org:8080/docview/1287024919/abstract/863A340DAE0C4C20PQ/1
Rights Copyright International Forum of Educational Technology & Society 2012
Volume 15
Issue 3
Pages n/a
Publication Journal of Educational Technology & Society
ISSN 11763647
Date 2012
Accessed 9/20/2016, 10:19:17 AM
Library Catalog ProQuest
Language English
Abstract With the increase in available educational data, it is expected that Learning Analytics will become a powerful means to inform and support learners, teachers and their institutions in better understanding and predicting personal learning needs and performance. However, the processes and requirements behind the beneficial application of Learning and Knowledge Analytics as well as the consequences for learning and teaching are still far from being understood. In this paper, we explore the key dimensions of Learning Analytics (LA), the critical problem zones, and some potential dangers to the beneficial exploitation of educational data. We propose and discuss a generic design framework that can act as a useful guide for setting up Learning Analytics services in support of educational practice and learner guidance, in quality assurance, curriculum development, and in improving teacher effectiveness and efficiency. Furthermore,

the presented article intends to inform about soft barriers and limitations of Learning Analytics. We identify the required skills and competences that make meaningful use of Learning Analytics data possible to overcome gaps in interpretation literacy among educational stakeholders. We also discuss privacy and ethical issues and suggest ways in which these issues can be addressed through policy guidelines and best practice examples. [PUBLICATION ABSTRACT]

Short Title Translating Learning into Numbers
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Tags:

Data for learning, Domain design, educational data mining, Ethics, Framework, Learning analytics

Notes:

The processes and requirements behind the beneficial application of Learning and Knowledge Analytics as well as the consequences for learning and teaching are still far from being understood.

They can be used for personalization. The driving factors include quality, effectiveness, effiiciency, reduced costs, and accessibility.

Rob's Class notes:

It tells you how to do learning Analytics - the 6 dimensions you need to consider - if you don't you're doing it wrong.

Dimensions:

1. stakeholders
2. Objectives
3. Data
4. Instruments -
5. Internal Limitations
6. External Limitations

What could be missing: The argument is that these are the six dimensions that provide the ethics you're looking for... but that is not guaranteed, thus, ethics should be it's own.

Why be afraid of Ed being out there?

- Education tracking that "tracks" poor performing students into tracks
- Companies try to manufacture superficial education and thus we lose track of how to truly education because of "marketing designed" to look like education.
- Metrics in education are very high stakes (hints: SAT); we need to move to a low stake world.

My Thoughts: Data and policy are separate. We look for insights, we don't dictate how they are used. Taking technology out of education will never happen, so a balance needs to be struck.

Attachments

- Full Text PDF

Translating Learning into Numbers: A Generic Framework for Learning Analytics.

Type Journal Article
Author Wolfgang Greller
Author Hendrik Drachsler
URL http://www.jstor.org/stable/jeductechsoci.15.3.42
Volume 15
Issue 3
Pages 42–57
Publication Educational technology & society
Date 2012
Accessed 12/21/2016, 2:20:36 PM
Library Catalog Google Scholar
Short Title Translating Learning into Numbers
Date Added 12/21/2016, 2:20:36 PM
Modified 12/21/2016, 2:20:36 PM

Notes:

With the increase in available educational data, it is expected that Learning Analytics will become a powerful means to inform and support learners, teachers and their institutions in better understanding and predicting personal learning needs and performance. However, the processes and requirements behind the beneficial application of Learning and Knowledge Analytics as well as the consequences for learning and teaching are still
far from being understood. In this paper, we explore the key dimensions of Learning Analytics (LA), the critical problem zones, and some potential dangers to the beneficial exploitation of educational data. We propose and discuss a generic design framework that can act as a useful guide for setting up Learning Analytics services in support of educational practice and learner guidance, in quality assurance, curriculum development, and in improving teacher effectiveness and efficiency.

Furthermore, the presented article intends to inform about soft barriers and limitations of Learning Analytics. We identify the required skills and competences that make meaningful use of Learning Analytics data possible to overcome gaps in interpretation literacy among educational stakeholders. We also discuss privacy and ethical issues and suggest ways in which these issues can be addressed through policy guidelines and best practice examples.

Attachments

- jeductechsoci.15.3.42.pdf

ABSTRACT

With the increase in available educational data, it is expected that Learning Analytics will become a powerful means to inform and support learners, teachers and their institutions in better understanding and predicting personal learning needs and performance. However, the processes and requirements behind the beneficial application of Learning and Knowledge Analytics as well as the consequences for learning and teaching are still
far from being understood. In this paper, we explore the key dimensions of Learning Analytics (LA), the critical problem zones, and some potential dangers to the beneficial exploitation of educational data. We propose and discuss a generic design framework that can act as a useful guide for setting up Learning Analytics services in support of educational practice and learner guidance, in quality assurance, curriculum development, and in
improving teacher effectiveness and efficiency.

Furthermore, the presented article intends to inform about soft barriers and limitations of Learning Analytics. We identify the required skills and competences that make meaningful use of Learning Analytics data possible to overcome gaps in interpretation literacy among educational stakeholders. We also discuss privacy and ethical issues and suggest ways in which these issues can be addressed through policy guidelines and best practice examples.

Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research

Type Journal Article
Author D. Z. Grunspan
Author B. L. Wiggins
Author S. M. Goodreau
URL http://www.lifescied.org/cgi/doi/10.1187/cbe.13-08-0162

Volume 13
Issue 2
Pages 167-178
Publication Cell Biology Education
ISSN 1931-7913
Date 2014-06-01
DOI 10.1187/cbe.13-08-0162
Accessed 12/21/2016, 2:20:27 PM
Library Catalog CrossRef
Language en
Short Title Understanding Classrooms through Social Network Analysis
Date Added 12/21/2016, 2:20:27 PM
Modified 12/21/2016, 2:20:27 PM

Attachments

- 167.full.pdf

ABSTRACT

Social interactions between students are a major and underexplored part of undergraduate education.Understanding how learning relationships form in undergraduate classrooms, as well as the impacts these relationships have on learning outcomes, can inform educators in unique ways and improve educational reform. Social network analysis (SNA) provides the necessary tool kit for investigating questions involving relational data. We introduce basic concepts in SNA, along with methods for data collection, data processing, and data analysis, using a previously collected example study on an undergraduate biology classroom as a tutorial. We conduct descriptive analyses of the structure of the network of costudying relationships.We explore generative processes that create observed study networks between students and also test for an association between network position and success on exams. We also cover practical issues, such as the unique aspects of human subjects review for network studies. Our aims are to convince readers that using SNA in classroom environments allows rich and informative analyses to take place and to provide some initial tools for doing so, in the process inspiring future educational studies incorporating relational data.

Rob's Thoughts:

I found this study interesting and informative of the strengths of SNA, but also understanding its limits as a predictor of performance.

The Big Five and Visualisations of Team Work Activity

Type Conference Paper
Author Judy Kay
Author Nicolas Maisonneuve
Author Kalina Yacef
Author Peter Reimann
URL http://link.springer.com/chapter/10.1007/11774303_20
Publisher Springer, Berlin, Heidelberg
Pages 197-206
Date 2006/6/26
Extra DOI: 10.1007/11774303_20
Accessed 12/19/2016, 3:44:12 PM
Library Catalog link.springer.com
Conference Name International Conference on Intelligent Tutoring Systems
Language en
Abstract We have created a set of novel visualisations of group activity: they mirror activity of individuals and their interactions, based upon readily available authentic data. We evaluated these visualisations in the context of a semester long software development project course. We give a theoretical analysis of the design of our visualizations using the framework from the “Big 5” theory of team work as well as a qualitative study of the visualisations and the students’ reflective reports. We conclude that these visualisations provide a powerful and valuable mirroring role with potential, when well used, to help groups learn to improve their effectiveness.
Proceedings Title Intelligent Tutoring Systems
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Attachments

- Full Text PDF
- Snapshot

Teaching Recommender Systems at Large Scale: Evaluation and Lessons Learned from a Hybrid MOOC

Type Journal Article
Author Joseph A. Konstan
Author J. D. Walker
Author D. Christopher Brooks
Author Keith Brown
Author Michael D. Ekstrand
URL http://doi.acm.org/10.1145/2728171
Volume 22
Issue 2
Pages 10:1–10:23
Publication ACM Trans. Comput.-Hum. Interact.
ISSN 1073-0516
Date April 2015
DOI 10.1145/2728171
Accessed 12/19/2016, 3:45:16 PM
Library Catalog ACM Digital Library
Abstract In the fall of 2013, we offered an open online Introduction to Recommender Systems through Coursera, while simultaneously offering a for-credit version of the course on-campus using the Coursera platform and a flipped classroom instruction model. As the goal of offering this course was to experiment with this type of instruction, we performed extensive evaluation including surveys of demographics, self-assessed skills, and learning intent; we also designed a knowledge-assessment tool specifically for the subject matter in this course, administering it before and after the course to measure learning, and again 5 months later to measure retention. We also tracked students through the course, including separating out students enrolled for credit from those enrolled only for the free, open course. Students had significant knowledge gains across all levels of prior knowledge and across all demographic categories. The main predictor of knowledge gain was effort expended in the course. Students also had significant knowledge retention after the course. Both of these results are limited to the sample of students who chose to complete our knowledge tests. Student completion of the course was hard to predict, with few factors contributing predictive power; the main predictor of completion was intent to complete. Students who chose a concepts-only track with hand exercises achieved the same level of knowledge of recommender systems concepts as those who chose a programming track and its added assignments, though the programming students gained additional programming knowledge. Based on the limited data we were able to gather, face-to-face students performed as well as the online-only students or better; they preferred this format to traditional lecture for reasons ranging from pure convenience to the desire to watch videos at a different pace (slower for English language learners; faster for some native English speakers). This article also includes our qualitative observations, lessons learned, and future directions.
Short Title Teaching Recommender Systems at Large Scale
Date Added 12/21/2016, 1:18:55 PM

Modified 12/21/2016, 1:18:55 PM

Tags:

learning assessment, Massively Online Open Course (MOOC)

Attachments

- ACM Full Text PDF

Machine Beats Experts: Automatic Discovery of Skill Models for Data-Driven Online Course Refinement

Type	Book
Author	Noboru|Furukawa Matsuda
URL	http://eric.ed.gov/?id=ED560513
Publisher	International Educational Data Mining Society
Date	2015/06/00
Accessed	12/19/2016, 3:45:42 PM
Library Catalog	ERIC
Language	en
Abstract	How can we automatically determine which skills must be mastered for the successful completion of an online course? Large-scale online courses (e.g., MOOCs) often contain a broad range of contents frequently intended to be a semester's worth of materials; this breadth often makes it difficult to articulate an accurate set of skills and knowledge (i.e., a skill model, or the QMatrix). We have developed an innovative method to discover skill models from the data of online courses. Our method assumes that online courses have a pre-defined skill map for which skills are associated with formative assessment items embedded throughout the online course. Our method carefully exploits correlations between various parts of student performance, as well as in the text of assessment items, to build a superior statistical model that even outperforms human experts. To evaluate our method, we compare our method with existing methods (LFA) and human engineered skill models on three Open Learning Initiative (OLI) courses at Carnegie Mellon University. The results show that (1) our method outperforms human-engineered skill models, (2) skill models discovered by our method are interpretable, and (3) our method is remarkably faster than existing methods. These results suggest that our method provides a significant contribution to the evidence-based, iterative refinement of online courses with a promising scalability. [For complete proceedings, see ED560503.]
Short Title	Machine Beats Experts
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Tags:

Automation, Comparative Analysis, correlation, data, Formative Evaluation, Models, Online Courses, Skills

Notes:

Summary

How can we automatically determine which skills must be mastered for the successful completion of an online course? Large-scale online courses (e.g., MOOCs) often contain a broad range of contents frequently intended to be a semester's worth of materials; this breadth often makes it difficult to articulate an accurate set of skills and knowledge (i.e., a skill model, or the QMatrix). We have developed an innovative method to discover skill models from the data of online courses. Our method assumes that online courses have a pre-defined skill map for which skills are associated with formative assessment items embedded throughout the online course. Our method carefully exploits correlations between various parts of student performance, as well as in the text of assessment items, to build a superior statistical model that even outperforms human experts. To evaluate our method, we compare our method with existing methods (LFA) and human engineered skill models on three Open Learning Initiative (OLI) courses at Carnegie Mellon University. The results show that (1) our method outperforms human-engineered skill models, (2) skill models discovered by our method are interpretable, and (3) our method is remarkably faster than existing methods. These results suggest that our method provides a significant contribution to the evidence-based, iterative refinement of online courses with a promising scalability.

The method outperforms human-engineered skill models,

The skill models discovered by the method are interpretable, and

The method is remarkably faster than existing methods.

We can use this technique to evaluate instructional design/rubrics against student performance within a course. This can be used to identify gaps in entry-level skills needed for performing well in the educational system.

Attachments

- Full Text PDF
- Snapshot

Predicting college enrollment from student interaction with an intelligent tutoring system in middle school

Type	Conference Paper
Author	Maria Ofelia Pedro
Author	Ryan Baker
Author	Alex Bowers
Author	Neil Heffernan
URL	http://www.educationaldatamining.org/conferences/index.php/EDM/2013/paper/download/1035/1001
Date	2013
Accessed	12/21/2016, 2:21:21 PM
Library Catalog	Google Scholar
Proceedings Title	Educational Data Mining 2013
Date Added	12/21/2016, 2:21:21 PM
Modified	12/21/2016, 2:21:21 PM

Attachments

- EDM2013_SBBH.pdf

Summary

Research shows that middle school is an important juncture for a student where he or she starts to be conscious about academic achievement and thinks about college attendance. It is already known that access to financial resources, family background, career aspirations and academic ability are indicative of a student's choice to attend college; though these variables are interesting, they do not necessarily give sufficient actionable information to instructors or guidance counselors to intervene for individual students. However, increasing numbers of students are using educational software at this phase of their education, and detectors of specific aspects of student learning and engagement have been developed for these types of learning environments. If these types of models can be used to predict college attendance, it may provide more actionable information than the previous generation of predictive models. In this paper, we predict college attendance from these types of detectors, in the context of 3,747 students using the ASSISTment system in New England, producing detection that is both successful and potentially more actionable than previous approaches; we can distinguish between a student who will attend college and a student who will not attend college 68.6% of the time.

Conclusion:

- Many factors influence a students decision to enroll in college
- mostly external and social:
  - financial
  - parental support
  - school support
- One's ability to engage
- **The study of the ASSITment data + EDM techniques can accurately predict students college enrollment by 68.6%'**
  - Sig Predictors:
    - Boredom
    - confusion
    - slip/carelessness
  - Success in middle school math is key (indicated by correct answers and high probability of knowledge of subject
  - Potenially.... software that tackle the areas of disenagement in different ways could prevent disengagement.

------***************

- Middle school is where students start to engage and disengage from learning.
  - this can follow them throughout their schooling
- Studies support the idea of predicting drop out rates based on late elementary/ middle school
- Conversely students who begin to plan for College in middle school are more likely to achieve
- Major issue is that intervention at the levels of HS are too late as the measurements of failure in those grades are too late.

This Paper:

- Study how learning and engagement in middle school – as assessed by this type of automated detector – can be used to predict college enrollment using:
  - EDM techniques
  - Log files from edu. software
  - constructs such as:
    - gaming the system
    - off-task behavior
    - carelessness
    - boredom
    - frustration
    - engaged concentration
  - using automated detectors to track these factors over time

Methodology:

- The ASSISTment System (tutoring system):
  - assesses middle school math skills
  - uses scaffolding to help bridge knowledge gaps
- ASSISTments Data:
  - N = 3,747 students
  - New England area
  - 2 years long use
  - Demo: mix of urban, low income, english second language & suburb middle class
  - Used college enrollment data from the Clearinghouse to track if these students were in school or not
- Feature Creation:
  - knowledge estimate
  - student affect (boredom, confusion, engagement)
  - student disengagement (off-task, gaming system,)
    - How:
      - observation of students + log files of students at time of observation
      - Used **Cohen's Kappa to legitimize approach?**
      - Used REP-Tree Algorithm
      - K algorithm
      - Contextual Slip model was created using Bayesian + linear regression
        - Linear regression is good for "noisy" education data
  - Usage data (right answers, attempts)
  - Bayesian Knowledge Tracing = Latent knoweldge
- Logical Regression:
  - They created an average score for factors (ave. boredom, confusion, etc.)
  - Choosing logistic regression allows for relatively good interpretability of the resultant models, while matching the statistical approach used in much of the other work on predicting college attendance
- Results:
  - Conducting an independent samples t-test (equal variances assumed) indicates that, with the exception of off-task, the difference of means of each feature between the two groups are statistically significant

Learning analytics and educational data mining: towards communication and collaboration

Type	Conference Paper
Author	George Siemens
Author	Ryan SJ d Baker
URL	http://dl.acm.org/citation.cfm?id=2330661
Publisher	ACM
Pages	252–254
Date	2012
Accessed	12/21/2016, 2:20:01 PM
Library Catalog	Google Scholar
Proceedings Title	Proceedings of the 2nd international conference on learning analytics and knowledge
Short Title	Learning analytics and educational data mining
Date Added	12/21/2016, 2:20:01 PM
Modified	12/21/2016, 2:20:01 PM

Notes:

Summary

Growing interest in data and analytics in education, teaching, and learning raises the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics. Two research communities – Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have developed separately to address this need. This paper argues for increased and formal communication and collaboration between these communities in order to share research, methods, and tools for data mining and analysis in the service of developing both LAK and EDM fields.

Attachments

- LAKs reformatting v2.pdf

  ABSTRACT

  Growing interest in data and analytics in education, teaching, and learning raises the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics. Two research communities – Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have developed separately to address this need. This paper argues for increased and formal communication and collaboration between these communities in order to share research, methods, and tools for data mining and analysis in the service of developing both LAK and EDM fields.

Feature Selection - Georgia Tech - Machine Learning

Type	Video Recording
Director	Udacity
URL	https://www.youtube.com/watch?v=8CpRLplmdqE
Accessed	12/21/2016, 11:49:53 AM
Library Catalog	YouTube
Running Time	194 seconds
Abstract	Watch on Udacity: https://www.udacity.com/course/viewer... Check out the full Advanced Operating Systems course for free at: https://www.udacity.com/course/ud262 Georgia Tech online Master's program: https://www.udacity.com/georgia-tech
Date Added	12/21/2016, 2:13:52 PM
Modified	12/21/2016, 2:13:52 PM

Tags:

computer science, Georgia Tech, Machine learning, Udacity, unsupervised learning

Notes:

Which sub-set of features actually matter.

- Good for **interpretability** and insight
- Curse of dimensionality
  - as you add more features you'll need more data to really gain any insight.

The design and development of the dragoon intelligent tutoring system for model construction: lessons learned

Type	Journal Article
Author	Jon Wetzel
Author	Kurt VanLehn
Author	Dillan Butler
Author	Pradeep Chaudhari
Author	Avaneesh Desai
Author	Jingxian Feng
Author	Sachin Grover
Author	Reid Joiner
Author	Mackenzie Kong-Sivert
Author	Vallabh Patade
Author	Ritesh Samala
Author	Megha Tiwari
Author	Brett van de Sande
URL	http://dx.doi.org/10.1080/10494820.2015.1131167
Volume	0

Issue 0
Pages 1-21
Publication Interactive Learning Environments
ISSN 1049-4820
Date March 11, 2016
DOI 10.1080/10494820.2015.1131167
Accessed 12/21/2016, 3:10:05 PM
Library Catalog Taylor and Francis+NEJM
Abstract This paper describes Dragoon, a simple intelligent tutoring system which teaches the construction of models of dynamic systems. Modelling is one of seven practices dictated in two new sets of educational standards in the U.S.A., and Dragoon is one of the first systems for teaching model construction for dynamic systems. Dragoon can be classified as a step-based tutoring system that uses example-tracing, an explicit pedagogical policy and an open learner model. Dragoon can also be used for computer-supported collaborative learning, and provides tools for classroom orchestration. This paper describes the features, user interfaces, and architecture of Dragoon; compares and contrasts Dragoon with other intelligent tutoring systems; and presents a brief overview of formative and summative evaluations of Dragoon in both high school and college classes. Of four summative evaluations, three found that students who used Dragoon learned more about the target system than students who did equivalent work without Dragoon.
Short Title The design and development of the dragoon intelligent tutoring system for model construction
Date Added 12/21/2016, 3:10:05 PM
Modified 12/21/2016, 3:10:05 PM

Tags:

classroom orchestration tools, dynamic systems, Intelligent tutoring system, learning by authoring intelligent tutoring systems, model construction, open learner model

Notes:

- § A moocRP includes the following features.
  - It integrates data request/authorization and distribution workflow features as well as provides a simple analytics module upload format to enable reuse and replication of analytics results among instructors and researchers.
  - Responsive visualizations based on different data models can be generated and distributed in the learning community.
  - The tool is tailored for application in Massive Open Online Communities (MOOCs) of learning.
  - A simple, aggregated interface to all of its basic functions enables access to data and collaboration/sharing of analytical modules.

Takeaway Message

- § Tools like moocRP are needed to realize the impact of each researcher's work in the learning analytics community and foster collaborations; these tools can help researchers build upon each other's work, and facilitating this should prove a productive and valuable area for future work.

Research Questions and Findings

Research Question:

- § To what extent does the novel tool, moocRP, function in various practical scenarios?

Key Findings:

- § The educational data mining and learning analytics communities have been a driving force behind academic and industry efforts towards collecting, curating, and operationalizing education data in service to the learner.
- § Centralized repositories, such as the LearnLab's DataShop, are not supported by an institution and rely on government funding to keep its services running.
- § moocRP proposed features to be implemented, with community support and continued engineering work, include:
  - automated analytic module security screening,
  - alternative authentication protocols and modularization,
  - integration of data pre-processing scripts,
  - support for scripts written in alternative languages (e.g. Matlab, Perl, and Python that could be used for machine learning analytics),
  - additional statistics on data model and analytic usage information, and
  - various server and implementation optimizations.
- § A visualized tree structure using an expandable layout is created from parsing of the course file data and. While this visualization shows only an interactive course tree structure, it can be used as the basis for navigating to analytics about different elements of the course, such as selecting the problem to receive Bayesian problem analytics
- § The visualization is accurately presented along with age and level of education information so that an instructor may inspect for which demographic students lacked the requisite prior knowledge.
- § Additional attributes can be added to this visualization, including country, survey information, and outcomes for previous parts of the course.
- § Tools like moocRP are needed to realize the impact of each researcher's work in the learning analytics community and foster collaborations; these tools can help researchers build upon each other's work, and facilitating this should prove a productive and valuable area for future work.

Attachments

- Snapshot

Evaluating Machine Learning Models - O'Reilly Media

Type Web Page
Author Alice Zheng
URL http://www.oreilly.com/data/free/evaluating-machine-learning-models.csp?intcmp=il-data-free-lp-lgen_free_reports_page
Accessed 12/19/2016, 3:46:17 PM
Abstract Data science today is a lot like the Wild West: there's endless opportunity and excitement, but also a lot of chaos and confusion. If you're new to data science and applied machine learning, evaluating a machine-learning model can seem pretty overwhelming...
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Attachments

- Snapshot

  Zheng first introduces the machine-learning workflow, and then dives into evaluation metrics and model selection. The latter half of the report focuses on hyperparameter tuning and A/B testing, which may benefit more seasoned machine-learning practitioners.

  - Learn the stages involved when developing a machine-learning model for use in a software application
  - Understand the metrics used for supervised learning models, including classification, regression, and ranking
  - Walk through evaluation mechanisms, such as hold?out validation, cross-validation, and bootstrapping
  - Explore hyperparameter tuning in detail, and discover why it's so difficult
  - Learn the pitfalls of A/B testing, and examine a promising alternative: multi-armed bandits
  - Get suggestions for further reading, as well as useful software packages

---

Chicken or the Egg? Granger-Causality for the masses

Type Blog Post
URL https://www.r-bloggers.com/chicken-or-the-egg-granger-causality-for-the-masses/
Date 2013-06-02T20:46:00+00:00
Accessed 12/21/2016, 1:07:57 PM
Abstract


```
<div><br></div> <div><span>When I first learned about Granger-causality this past February, I was bemused and quite skeptical of the whole procedure.  I felt it belonged on the
scrapheap of impractical academic endeavors, preferring to possibly use an ARIMA transfer function model for the same task.  However, several contemporaries threw the red
challenge flag and upon further review, my initial impressions have been overturned.   Not only am I fascinated by the technique, in my attempt to discover its value I have
became a raving R fan.  As such, my first blog entry is to provide some simple code to allow anyone to utilize this obscure econometric technique, but first some background.<p></p>
</span></div> <div><span>Given two sets of time series data, x and y, granger-causality is a method which attempts to determine whether one series is likely to influence change in the
other. This is accomplished by taking different lags of one series and using that to model the change in the second series. We create two models which predict y, one with only past values of y
(&Omega;), and the other with past values of y and x (&pi;). The models are given below where k is the number of lags in the time series:<p></p></span></div> <div><br></div> <div>
<span>Let &Omega; = y<sub>t</sub> = &beta;<sub>0</sub> + &beta;<sub>1</sub>y<sub>t-1</sub> +&hellip;+ &beta;<sub>k</sub>y<sub>t-k </sub>+ </span><span>e</span>
<span><p></p></span> </div> <div><br></div> <div> <span>And &pi; = y<sub>t</sub> = &beta;<sub>0</sub> + &beta;<sub>1</sub>y<sub>t-1</sub> +&hellip;+ &beta;
<sub>k</sub>y<sub>t-k </sub>+ &alpha;<sub>1</sub>x<sub>t-1</sub> +&hellip;+ &alpha;<sub>k</sub>x<sub>t-k</sub> + </span><span>e</span><span><p></p></span> </div>
<div><br></div> <div><span>The residual sums of squared errors are then compared and a test is used to determine whether the nested model (&Omega;) is adequate to explain the future
values of y or if the full model (&pi;) is better. The F-test, t-test or Wald test (used in R) are calculated to test the following null and alternate hypotheses:<p></p></span></div> <div><br>
</div> <div><span>H<sub>0</sub>:  &alpha;<sub>i</sub> = 0 for each i of the element [1,k]<p></p></span></div> <div><span>H<sub>1</sub>: &alpha;<sub>i</sub> &ne; 0
for at least 1 i of the element [1,k]<p></p></span></div> <div><br></div> <div><span>Essentially, we are trying to determine whether we can say that statistically x provides more
information about future values of y than past values of y alone. Under this definition it is clear that we are not trying to prove actual causation, only that the two values are related by some
phenomenon.  Along those lines, we must also run this model in reverse to verify that that y does not provide information about future values of x. If we find that this is the case, it is
likely that there is some exogenous variable, z, which needs to be controlled or could be a better candidate for granger causation.  <p></p></span></div> <div><span>For a detailed
explanation, one can read the original paper on the subject:<p></p></span></div> <div><span>Granger, Clive W., &ldquo;Investigating Causal Relations by Econometric Models and
Cross-Spectral Methods&rdquo;, <i>Econometrica, </i>37(1969): 424-38<p></p></span></div> <div><span>The R package &ldquo;lmtest&rdquo; incorporates the granger causality
procedure, including a data set to answer the age old question of what came first, &ldquo;the chicken or the egg&rdquo;.  The data was presented by Walter Thurman and Mark Fisher
in the American Journal of Agricultural Economics, May 1988, titled &ldquo;Chickens, Eggs, and Causality, or Which Came First?&rdquo;  It consists of two time series from 1930 to
1983, one of U.S. egg production and the other the estimated U.S. chicken population.<a href="http://www.blogger.com/blogger.g?blogID=1476380772205204590" name="_GoBack">
</a><p></p></span></div> <div><span>Let's get some code going, first loading the data from a saved .csv file.<p></p></span></div> <div><span>&gt; chickegg &lt;-
read.csv(file.choose())<p></p></span></div> <div><span>&gt; head(chickegg)<p></p></span></div> <div><span>  Year chicken  egg<p></p></span></div>
<div><span>1 1930  468491 3581<p></p></span></div> <div><span>2 1931  449743 3532<p></p></span></div> <div><span>3 1932  436815 3327<p>
</p></span></div> <div><span>4 1933  444523 3255<p></p></span></div> <div><span>5 1934  433937 3156<p></p></span></div> <div><span>6 1935  389958
3081<p></p></span></div> <div><span>&gt; attach(chickegg)<p></p></span></div> <div><span>&gt; # plot the time series<p></p></span></div> <div><span>&gt;
par(mfrow=c(2,1))<p></p></span></div> <div><span>&gt; plot.ts(chicken)<p></p></span></div> <div><span>&gt; plot.ts(egg)<p></p></span></div> <div><br></div> <div></div>
<br><div><a href="http://3.bp.blogspot.com/-1Gtzau_3n1c/UawDMbtg02I/AAAAAAAAAAo/ZZzZ1H4edng/s1600/Picture1.png"><img border="0" height="319"
src="http://3.bp.blogspot.com/-1Gtzau_3n1c/UawDMbtg02I/AAAAAAAAAAo/ZZzZ1H4edng/s320/Picture1.png" width="320"></a></div> <div></div> <div><br></div>
<div><span>The plots provide little information other than the data is likely not stationary.  I've just started using the forecast package, so let's load it and test for what will achieve
stationarity.<p></p></span></div> <div><br></div> <div><span>&gt; library(forecast)<p></p></span></div> <div><span>&gt; # test for unit root and number of differences required,
you can also test for seasonality with nsdiffs<p></p></span></div> <div><span>&gt; ndiffs(chicken, alpha=0.05, test=c("kpss"))  <p></p></span></div> <div><span>[1]
1<p></p></span></div> <div><span>&gt; ndiffs(egg, alpha=0.05, test=c("kpss"))<p></p></span></div> <div><span>[1] 1<p></p></span></div> <div><br></div> <div><span>&gt; #
differenced time series<p></p></span></div> <div><span>&gt; dchick &lt;- diff(chicken)<p></p></span></div> <div><span>&gt; degg &lt;- diff(egg)<p></p></span></div>
<div><br></div> <div><span>&gt; plot.ts(dchick)<p></p></span></div> <div><span>&gt; plot.ts(degg)<p></p></span></div> <div><br></div> <div><span>Much better!<p></p>
</span></div> <div><a href="http://1.bp.blogspot.com/-684DEXy_UWQ/UawDOktQQfI/AAAAAAAAAAw/vC2x8euFW2E/s1600/Picture2.png"><img border="0" height="319"
src="http://1.bp.blogspot.com/-684DEXy_UWQ/UawDOktQQfI/AAAAAAAAAAw/vC2x8euFW2E/s320/Picture2.png" width="320"></a></div> <div><br></div> <div><span>
<br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span>
</div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div>
<div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span><br></span></div> <div><span>
<br></span></div> <div><span>That's pretty standard stuff, but this is where the magic happens!  There are several ways to find the optimal lag, which I will skip in the interest of
time, but let's say four is the magic number.<p></p></span></div> <div><br></div> <div><span>&gt; # do eggs granger cause chickens? <p></p></span></div> <div><span>&gt;
grangertest(dchick ~ degg, order=4)<p></p></span></div> <div><span>Granger causality test<p></p></span></div> <div><br></div> <div><span>Model 1: dchick ~ Lags(dchick, 1:4)
+ Lags(degg, 1:4)<p></p></span></div> <div><span>Model 2: dchick ~ Lags(dchick, 1:4)<p></p></span></div> <div><span>  Res.Df Df     
F      Pr(&gt;F)   <p></p></span></div> <div><span>1    
40                        <p></p></span></div>
<div><span>2     44   -4   4.1762  <b>0.006414</b>**<p></p></span></div> <div><span>---<p></p></span></div>
<div><span>Signif. codes:  0 &lsquo;***&rsquo; 0.001 &lsquo;**&rsquo; 0.01 &lsquo;*&rsquo; 0.05 &lsquo;.&rsquo; 0.1 &lsquo; &rsquo; 1<p></p></span></div> <div><br>
</div> <div><span>Highly significant p-value, but what about the other direction?<p></p></span></div> <div><br></div> <div><span>&gt; # do chickens granger cause eggs, at lag
4?<p></p></span></div> <div><span>&gt; grangertest(degg ~ dchick, order=4)<p></p></span></div> <div><span>Granger causality test<p></p></span></div> <div><br></div>
<div><span>Model 1: degg ~ Lags(degg, 1:4) + Lags(dchick, 1:4)<p></p></span></div> <div><span>Model 2: degg ~ Lags(degg, 1:4)<p></p></span></div> <div><span>  Res.Df
Df      F Pr(&gt;F)<p></p></span></div> <div><span>1    
40                    <p></p></span></div>
<div><span>2     44 -4 0.2817 0.8881<p></p></span></div> <div><br></div> <div><span>It is not significant, so we can say the eggs Granger-Cause chickens!
<p></p></span></div> <div><span>This is just the tip of the iceberg, but should be enough to strike up your curiosity and to make you dangerous.  I'm working on commodity prices,
bond prices and the U.S. stock market, but that is better left for another day.<p></p></span></div> <div><br></div> <div><span>Reference:<p></p></span></div> <div>
<span>Thurman W.N. &amp; Fisher M.E. (1988), Chickens, Eggs, and Causality, or Which Came First?,<span> </span><em>American Journal of Agricultural Economics</em>,
237-238.</span><span><p></p></span> </div> <br><div><br></div>
```


Blog Title R-bloggers
Short Title Chicken or the Egg?
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Attachments

- Snapshot

---

data-wrangling-cheatsheet.pdf

Type Attachment
URL http://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
Accessed 12/19/2016, 3:43:47 PM
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:40:26 PM

Quick reference sheet for manipulating Data in R (Studio) using *dplyr* and *tidyr* packages.

Tidy data complements R's vectorized operations. R will automatically preserve observations as you manipulate variables.
No other format works as intuitively with R.

Includes directions to understand syntax of the libraries, reshape, subset observations (rows) & variables (columns), summarise data, create new variables, combine data sets, and group data.

How to analyze smartphone sensor data with R and the BreakoutDetection package

Type Blog Post
URL https://www.r-bloggers.com/how-to-analyze-smartphone-sensor-data-with-r-and-the-breakoutdetection-package/
Date 2015-01-10T08:56:46+00:00
Accessed 12/21/2016, 1:08:03 PM
Abstract


```
<p>Yesterday, J&ouml;rg has written a <a href="http://beautifuldata.net/2015/01/data-storytelling-stepwise-abstraction-from-raw-data/">blog post</a> on Data Storytelling with
Smartphone sensor data. Here&rsquo;s a practical approach on how to analyze <a>smartphone sensor data</a> with R. In this example I will be using the accelerometer smartphone data
that Datarella provided in its <a href="http://datarella.com/call-for-data-fiction/">Data Fiction competition</a>. The dataset shows the acceleration along [...]</p>
```


Blog Title R-bloggers

Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Notes:

Examine smartphone data from the accelerometer along three axes:

- x – sideways acceleration of the device
- y – forward and backward acceleration of the device
- z – acceleration up and down

Conduct breakout detection with the library(BreakoutDetection) in R.

Example:

> "install.packages("devtools") devtools::install_github("twitter/BreakoutDetection") library(BreakoutDetection) bo <- breakout(df$x[df$timestamp >= '2014-05-12 12:32:00' & df$timestamp < '2014-05-12 12:35:00'], min.size=10, method='multi', beta=.001, degree=1, plot=TRUE) bo$plot"

Attachments

- Snapshot

Introduction to Social Network Methods: Chapter 1: Social Network Data

Type Web Page
URL http://faculty.ucr.edu/~hanneman/nettext/C1_Social_Network_Data.html
Accessed 12/19/2016, 3:44:57 PM
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Notes:

Social Network Analysis (SNA) are present in other traditional methods, like cross-sectional survey research.

"Conventional" social science data consist of a rectangular array of measurements.

"Network" data (in their purest form) consist of a square array of measurements.

Network data are defined by actors and by relations (or "nodes" and "edges").

Social network analysts rarely draw samples in their work. Most commonly, network analysts will identify some population and conduct a census (i.e. include all elements of the population as units of observation).

Network analyst can also find individual people nested within networks of face-to-face relations with other persons... networks of interpersonal relations become "social facts" (e.g. people within a family share that common trait).

"Full network" methods yield the maximum of information, but can also be costly and difficult to execute.

"Snowbal"l methods begin with a focal actor or set of actors. Each of these actors is asked to name some or all of their ties to other actors. Then, all the actors named (who were not part of the original list) are tracked down and asked for some or all of their ties. The process continues until no new actors are identified, or until we decide to stop (usually for reasons of time and resources, or because the new actors being named are very marginal to the group we are trying to study).

ROB Notes

SN data lead us to think differently about how to apply statistics.

Major Emphasis:

1. The first major emphasis is seeing how actors are embedded in the overall network
2. Seeing how the whole pattern of individual choices give more holistic patterns.

SN analysis looks at actors and relations vs. conventitional looking at actors and attributes. (NOT SURE OF THE DIFFERENCE)

SN have two main parts Nodes (actors) and edges (relations)

- the subject or nodes isn't about an individual but samples elements.
  - not john and who are his friends
  - but instead john, his friends, and their friends connection
  - thus SN typically doesn't use "samples' but a large portion of a population.
    - all the classrooms in a school
    - all the words in a text
    - An entire social class

Modality: Understanding that every population is needed in multiple networks a student in a class, a class in a school, etc.

Two mode network: A data set that contains information about 2 types of social entities (say persons and organ.)

Measuring Sampling ties (Edges)

Methods:

- full network - max info, but can also be costly and difficult ot execute/ generalize.
  - taking a census of ties.
  - its difficult to collect every information from all individuals, which is needed.
  - most people have limited number of ties anyway.
- Snowball - begins with a focal actor(s)
  - actors are asked to name others, then the named are asked to name, until no new actors are listed.
  - good for special populations (business contacts, community elites)
  - limitation:
    - isolated folk are not collected
    - overstating connectiveness of a population
    - Thus it is best to start with the most powerful and well connected and work down. (although you still miss those not connected to the elite)
- ego-centric -
  - alter connections:
    - asking one actor to list folk and list how those folk may know eachother. (localize network)
    - you can also find trends and cliques within that mini network
  - ego only:

Scales of measurement

- Understanding how you want to rank relationships.
  - the presence of a tie or the strength of a tie
  - in doing this "dummy" coding you lose a ton of info
- Thus you might need to engage in full-rank (but uncommon)

Tags: network analysis

Attachments

- Introduction to Social Network Methods: Chapter 1: Social Network Data

---

mori-mori-mendiburu-etal.pdf

Type	Attachment
URL	https://journal.r-project.org/archive/accepted/mori-mori-mendiburu-etal.pdf
Accessed	12/21/2016, 1:08:06 PM
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Passing the Privacy Test as Student Data Laws Take Effect (EdSurge News)

Type	Web Page
URL	https://www.edsurge.com/news/2016-01-12-passing-the-privacy-test-as-student-data-laws-take-effect
Accessed	12/19/2016, 3:38:44 PM
Abstract	On January 1, 2016, " SOPIPA"—the recently passed California student data privacy law that defines how edtech companies can use student data became effective. About 25 other states have passed similar laws that are already in effect, or will become effective. At the same time, more than 200 sc
Website Title	EdSurge
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Attachments

- Snapshot

---

r - How to statistically compare two time series? - Cross Validated

Type	Web Page
URL	http://stats.stackexchange.com/questions/19103/how-to-statistically-compare-two-time-series
Accessed	12/21/2016, 1:07:50 PM
Short Title	r - How to statistically compare two time series?
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Attachments

- Snapshot

Saturday Morning Breakfast Cereal - Clock

Type	Web Page
URL	http://www.smbc-comics.com/index.php?id=3978
Accessed	12/19/2016, 3:43:26 PM
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Attachments

- Saturday Morning Breakfast Cereal - Clock

  lol

Shiny - The R Markdown Cheat sheet

Type	Web Page
URL	http://shiny.rstudio.com/articles/rm-cheatsheet.html
Accessed	12/19/2016, 3:38:41 PM
Date Added	12/21/2016, 1:18:55 PM
Modified	12/21/2016, 1:18:55 PM

Notes:

Quick Reference Sheet for R Markdown and Shiny Apps

Process:

1. Workflow R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more.

2. Open File Start by saving a text file with the extension .Rmd, or open an RStudio Rmd template

3. Markdown Next, write your report in plain text. Use markdown syntax to describe how to format text in the final report.

4. Choose Output Write a YAML header that explains what type of document to build from your R Markdown file.

5. Embed Code Use knitr syntax to embed R code into your report. R will run the code and include the results when you render your report.

6. Render Use your .Rmd file as a blueprint to build a finished report.

7. Interactive Docs Turn your report into an interactive Shiny document in 3 steps

8. Publish Share your report where users can visit it online

9. Learn More:
Documentation and examples - rmarkdown.rstudio.com
Further Articles - shiny.rstudio.com/articles

Attachments

- Shiny - The R Markdown Cheat sheet
- rmarkdown-cheatsheet.pdf

  Quick Reference Sheet for R Markdown and Shiny Apps

  Process:

  1. Workflow R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more.

  2. Open File Start by saving a text file with the extension .Rmd, or open an RStudio Rmd template

  3. Markdown Next, write your report in plain text. Use markdown syntax to describe how to format text in the final report.

  4. Choose Output Write a YAML header that explains what type of document to build from your R Markdown file.

  5. Embed Code Use knitr syntax to embed R code into your report. R will run the code and include the results when you render your report.

  6. Render Use your .Rmd file as a blueprint to build a finished report.

  7. Interactive Docs Turn your report into an interactive Shiny document in 3 steps

  8. Publish Share your report where users can visit it online

  9. Learn More:
  Documentation and examples - rmarkdown.rstudio.com
  Further Articles - shiny.rstudio.com/articles

Why Students Should Own Their Educational Data – Wired Campus - Blogs - The Chronicle of Higher Education

Type Blog Post
URL http://www.chronicle.com/blogs/wiredcampus/why-students-should-own-their-educational-data/54329
Accessed 12/19/2016, 3:44:05 PM
Date Added 12/21/2016, 1:18:55 PM
Modified 12/21/2016, 1:18:55 PM

Notes:

Mr. Rose (Harvard Professor; Graduate School of Education): non-profit for for supporting the development of techniques that teach at an individual level. -> Center of Individual Opportunity

Learning Style theory:

We need to be able to model education to understand personal patterns of how humans learn across dimensions.

Starts with personality research and how that is contextual not stable.

Hot Take: private R&D around education is hurting education b/c these companies own student data and they don't have access to info to know how they themselves learn.

Tags: custom learning, data, learning style theory, Massively Online Open Course (MOOC), MOOCs

Attachments

- Snapshot

Passing the Privacy Test as Student Data Laws Take Effect (EdSurge News)

Type Web Page
URL https://www.edsurge.com/news/2016-01-12-passing-the-privacy-test-as-student-data-laws-take-effect
Date 2016-01-12
Accessed 9/20/2016, 9:57:07 AM
Abstract On January 1, 2016, “ SOPIPA”—the recently passed California student data privacy law that defines how edtech companies can use student data became effective. About 25 other states have passed similar laws that are already in effect, or will become effective. At the same time, more than 200 sc
Website Title EdSurge
Date Added 12/21/2016, 2:13:52 PM
Modified 12/21/2016, 2:13:52 PM

Notes:

The SOPIPA laws don't require tools available to the general public that are being used by student to redesign their products even if the school uses them.

Students are typically protected by other federal and state laws.

Why of data that can be used:

Access,must delete, do not require students to share data unless it is for educational purposes.

This varies leaning towards the safe end - only allowing anonymized data at times.

Private v. Public institutions:

Public institutions have to comply w/ legislative moves around education data. Private do not.

Attachments

- Snapshot