



Big Data Challenges in the Education Industry

DABAI-EDU

Dept. of Computer Science

University of Copenhagen

DTU, Mar 29, 2017

1

Who Are We?



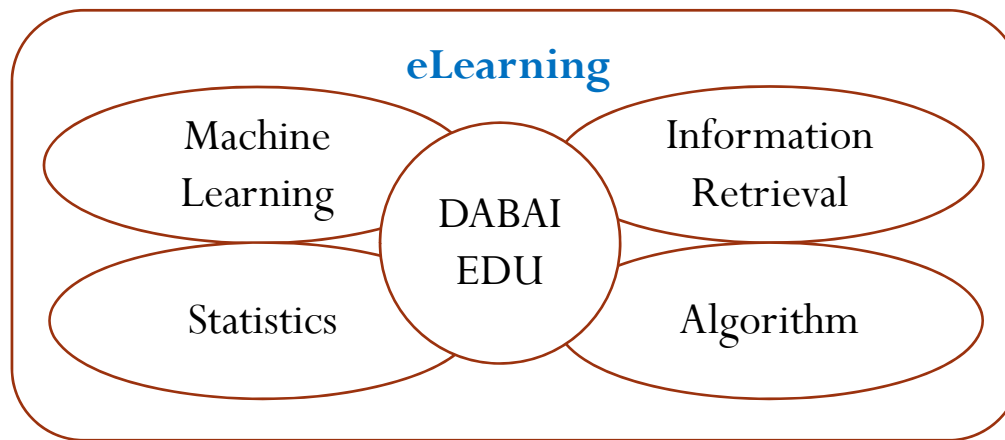
- **DABAI-EDU: Old guys**



Christian Igel



Yevgeny Seldin



Christina Lioma



Helle Rootzén



Stephen Alstrup



Mikkel Thorup

Who Are We?



- **DABAI-EDU: Young guys**



Ninh Pham

Randomized algorithms,
machine learning, big
data analytics



Niklas Hjuler

Combinatorics, big
data, machine learning



Stephan Lorenzen

Discrete algorithms, big
data, machine learning



Casper Hansen

Machine learning, big
data analytics



Christian Hansen

Machine learning, big
data, education data
mining

Some facts:



Secondary education in Denmark

- 175 gymnasiums (public and private)
- 40.718 students (started Stx and Hf in 2015)
- 10.826 teachers
- Drop out Stx: 16%
- Drop out Hf: 31%

Some facts:



Primary education in Denmark

- 2.430 schools
- 708.000 pupils
- 52.500 teachers

Some facts:



Number of full-time employed in Denmark: 2 mill. persons

Total number of students in Denmark: 1.250.000

The number of students in percentage of full-time employment in Denmark: **63%**

Education Industry Partners



Matematik
Fessor.dk

Clio Online



Lectio



Kasper Holst Hansen
Founder and CEO



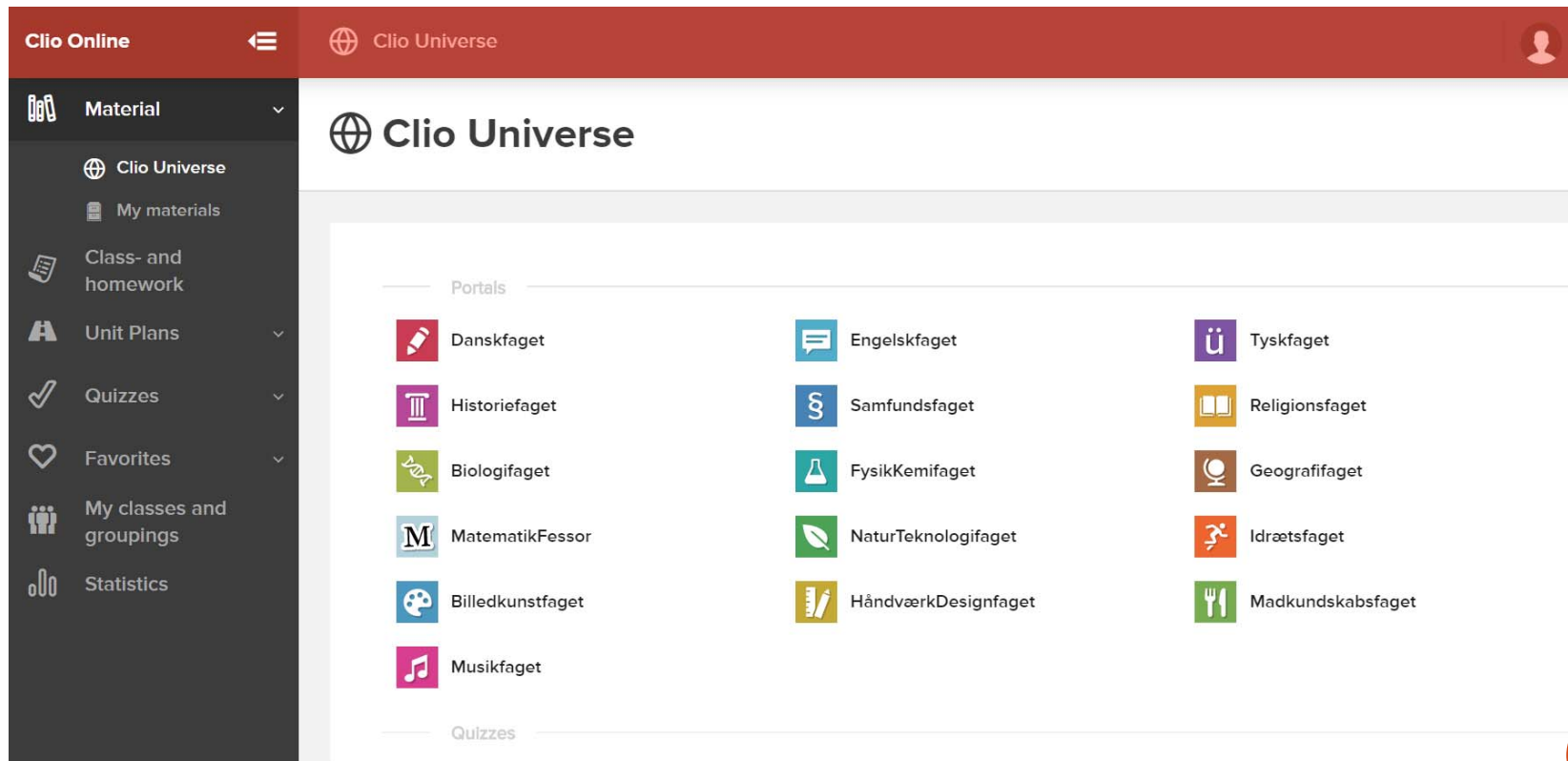
Lasse Guldsborg
Director IT & Finance



Martin Holbøll
Director & CEO

Clio Online: Case Study

- **90%** of primary schools have used
- **4 million** access per month in 2016









The screenshot displays the Clio Online user interface. At the top, there is a dark red header with the 'Clio Online' logo on the left, a hamburger menu icon, the text 'Clio Universe', and a user profile icon on the right. Below the header is a dark grey sidebar menu with the following items: 'Material' (with a dropdown arrow), 'Clio Universe', 'My materials', 'Class- and homework', 'Unit Plans' (with a dropdown arrow), 'Quizzes' (with a dropdown arrow), 'Favorites' (with a dropdown arrow), 'My classes and groupings', and 'Statistics'. The main content area is white and features the 'Clio Universe' logo at the top. Below the logo is a section titled 'Portals' containing a grid of subject-specific icons and labels: Danskfaget, Engelskfaget, Tyskfaget, Historiefaget, Samfundsfaget, Religionsfaget, Biologifaget, FysikKemifaget, Geografifaget, MatematikFessor, NaturTeknologifaget, Idrætsfaget, Billedkunstfaget, HåndværkDesignfaget, Madkundskabsfaget, and Musikfaget. At the bottom of the main content area, there is a section titled 'Quizzes'. In the bottom right corner of the slide, there is a red circular button with the number '8'.



Clio Online: Case Study

- Prediction of student performance on the quizzes system
 - Problem: estimating the score of an unseen quiz
 - Motivation: to personalize elearning, to classify quizzes...
 - Formulation:

	+	-	*	÷
	1.00	0.75	0.63	0.22
	0.75	1.00	0.91	?
	0.63	0.91	1.00	?
	0.22	?	?	1.00
	0.30	?	?	0.97
	?	0.16	0.40	0.64

- Can we predict the score ?

- Assume that there are a **small** number of latent features revealing the students and quizzes preferences



Clio Online: Case Study

- Application in Clio Online:
 - Predict the student's learning objectives on each unit plan
 - 1000 unit plans * 5 learning goals * 1M students ~ **5B** evaluations

🎯 Learning objectives

Below you can see the unit plans learning objectives, and evaluate yourself in relation to them.

Jeg kan lave en skulpturel hverdagsgenstand ud af en papkasse.

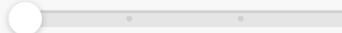
Where I am now



Cannot

Can

Where I want to be

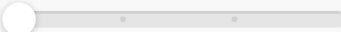


Cannot

Can

Jeg kan forklare, hvad performance er, og selv deltage i en.

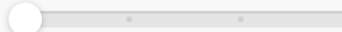
Where I am now



Cannot

Can

Where I want to be

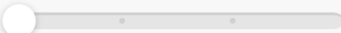


Cannot

Can

Jeg kan udvælge kraftige farver og male min skulptur med dem.

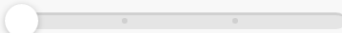
Where I am now



Cannot

Can

Where I want to be



Cannot






Can

d



Clio Online: Case Study

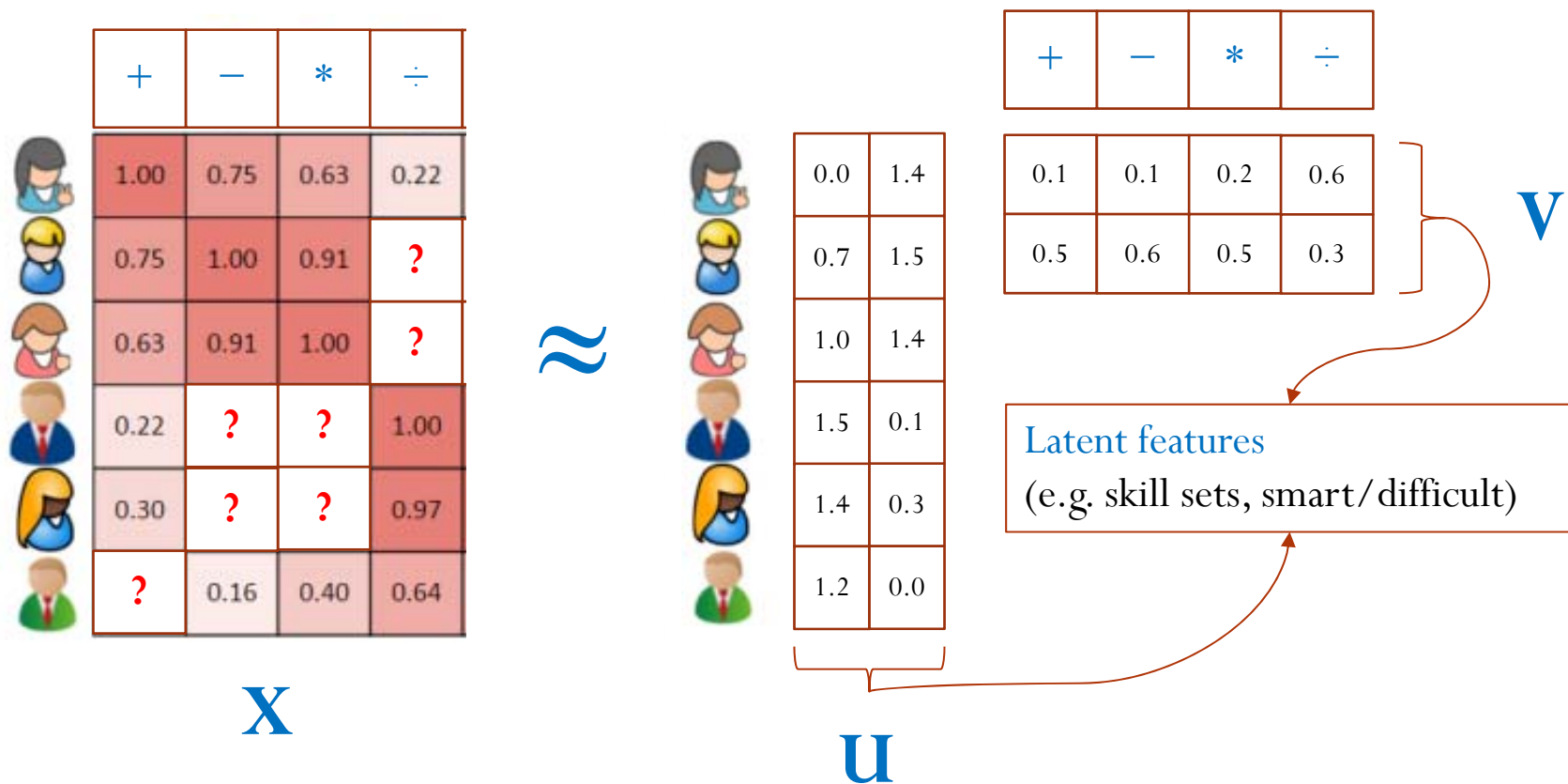
- Low-rank matrix factorization:

	+	-	*	÷
	1.00	0.75	0.63	0.22
	0.75	1.00	0.91	?
	0.63	0.91	1.00	?
	0.22	?	?	1.00
	0.30	?	?	0.97
	?	0.16	0.40	0.64

X







Clio Online: Case Study

- Low-rank matrix factorization:









Clio Online: Case Study

- Predictions:

	+	-	*	÷
	1.00	0.75	0.63	0.22
	0.75	1.00	0.91	0.9
	0.63	0.91	1.00	?
	0.22	?	?	1.00
	0.30	?	?	0.97
	0.1	0.16	0.40	0.64

X

\approx

	0.0	1.4
	0.7	1.5
	1.0	1.4
	1.5	0.1
	1.4	0.3
	1.2	0.0

U

+	-	*	÷
---	---	---	---







0.1	0.1	0.2	0.6
0.5	0.6	0.5	0.3

V

Latent features
(e.g. skill sets, smart/difficult)







Clio Online: Case Study

- Optimization problem:



	+	-	*	÷
	1.00	0.75	0.63	0.22
	0.75	1.00	0.91	0.9
	0.63	0.91	1.00	?
	0.22	?	?	1.00
	0.30	?	?	0.97
	0.1	0.16	0.40	0.64

X

≈

	0.0	1.4
	0.7	1.5
	1.0	1.4
	1.5	0.1
	1.4	0.3
	1.2	0.0

U

	+	-	*	÷
	0.1	0.1	0.2	0.6
	0.5	0.6	0.5	0.3

V

W is the weight matrix
 - 1s for observed entries
 - 0s for missing entries

$$\|\mathbf{W} \otimes (\mathbf{X} - \mathbf{UV})\|_F \text{ is minimum}$$

Clio Online: Case Study

- Techniques:

1. Randomly initialization U, V
2. Update X
3. Compute U, V such that $\|\mathbf{W} \otimes (\mathbf{X} - \mathbf{UV})\|_F$ is minimum
4. Repeat the step (2) until convergence

0.0	0.5
0.0	0.5
0.5	0.5
0.5	0.5
0.5	0.5
0.5	0.5

U

0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5

V

Optimizing



0.0	1.4
0.0	1.5
1.0	1.4
1.5	0.1
1.4	0.3
1.2	0.0

U

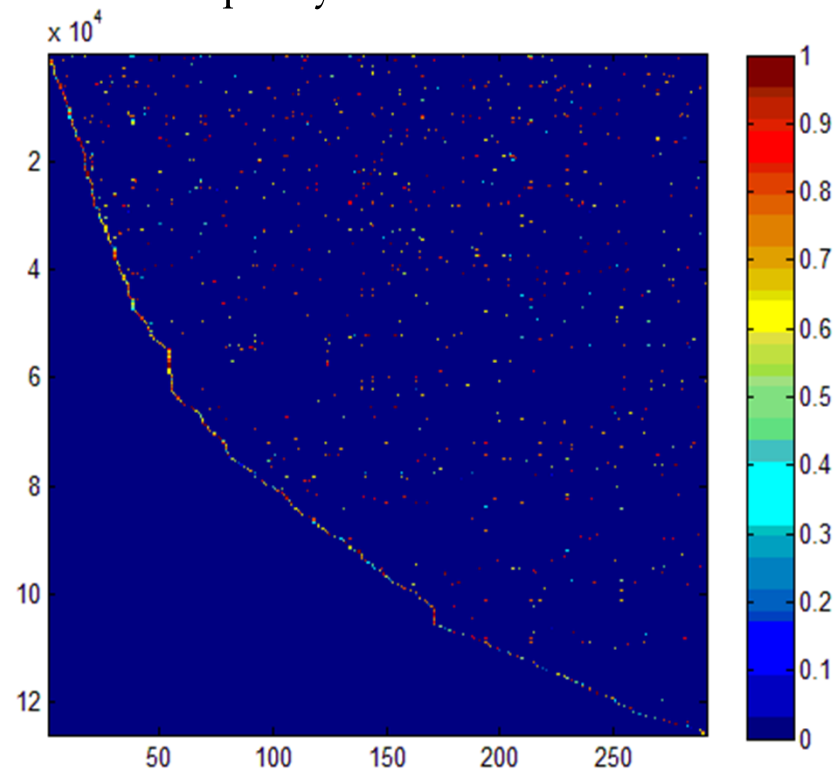
0.1	0.1	0.2	0.6
0.5	0.6	0.5	0.3

V

Clio Online: Case Study

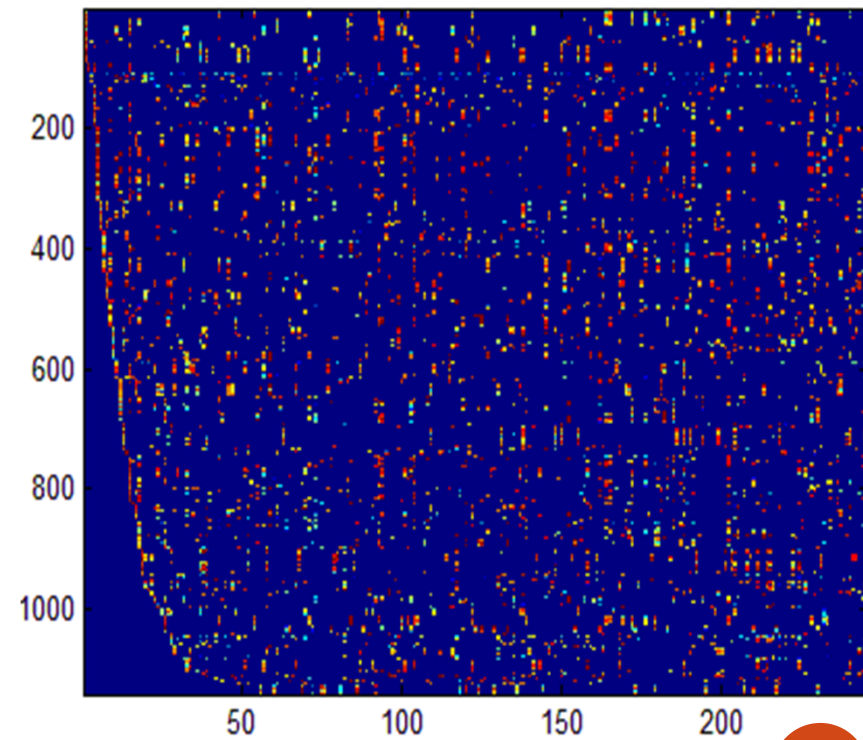
- Sample of Clio quiz:

- 126044 students, 291 quizzes
- Sparsity $\sim 1.2\%$



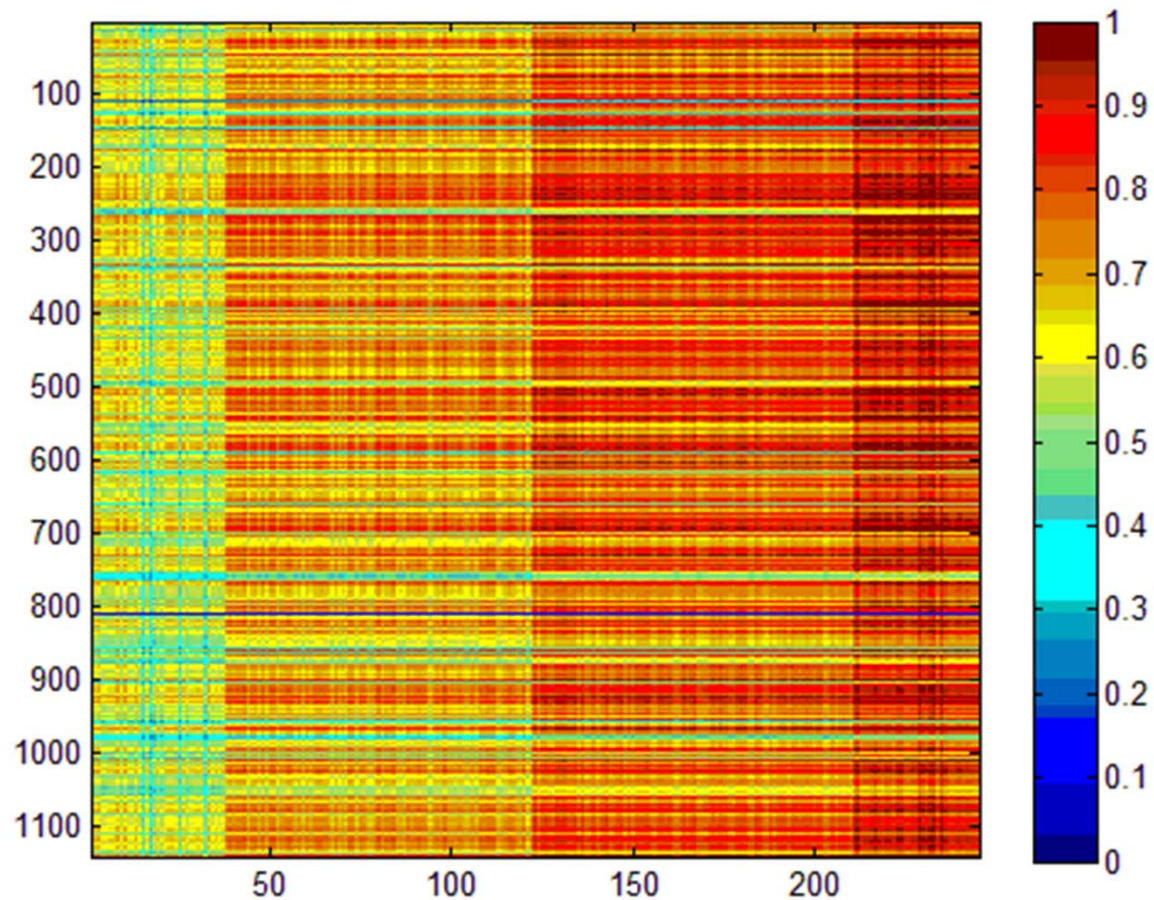
- Densify data set:

- 1141 students, 245 quizzes
- Sparsity $\sim 7.6\%$



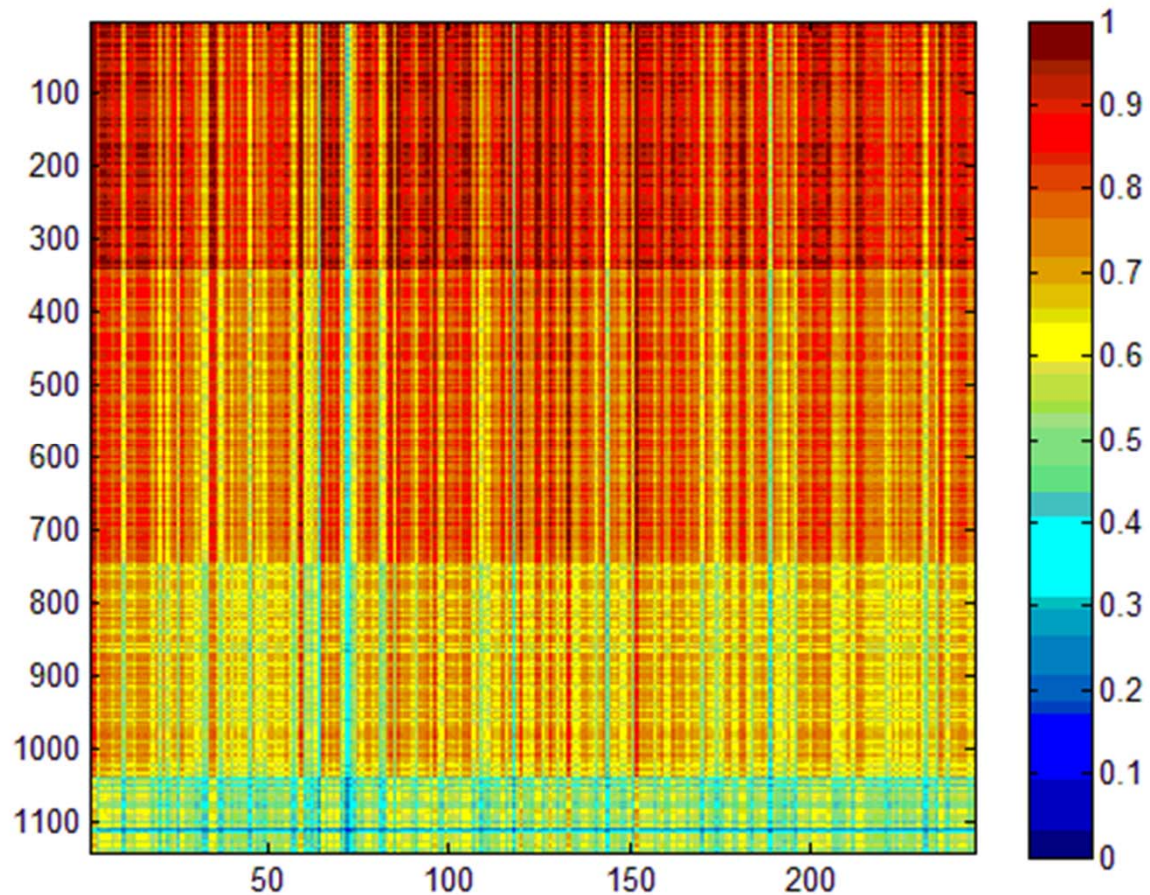
Clio Online: Case Study

- Results:
 - Quizzes classification (easy, average, difficult, very difficult)



Clio Online: Case Study

- Results:
 - Students classification (weak, average, good, very good)





Clio Online: Case Study

- **Conclusions:**

- We can predict the score in the range $x \pm 17$ (scale of 100)
- Active students tend to be good at most of quizzes
 - Average score is **73** in scale of 100
- Active students' behavior is stable in most of quizzes
 - Number of latent feature is **1**, corresponding to how good (difficult) a student (a quiz) is.
 - Most of student gets **50** scores for the very difficult quizzes, and **85** scores for easy quizzes.

- **Future work:**

- Improve the accuracy of prediction
- Efficient matrix factorization on large-scale quizzes & unit plans

EduLab: Case Study

- The largest supplier of online math in Denmark for primary school level
- 75% of all schools using the system, 1.5+ millions answered questions per day

The screenshot shows the Matematik Fessor.dk website interface. At the top, there is a navigation bar with the logo and menu items: Lektioner, Bogreolen, Træning, GeometriFessor, and Spil. The main heading is 'Lektionsoversigt'. Below this, there is a search bar labeled 'Søgeord' and a grade selector labeled 'Klassetrin' with a slider set to 7. To the right of the slider is a green button with the number 1. An illustration of school supplies (pencil, ruler, protractor, compass) is positioned to the right of the search and grade selector. On the left side, there is a vertical sidebar with a hamburger menu icon and the text 'Emneoversigt' with an arrow pointing to the right. The main content area displays a grid of lesson topics:

Afrunding af tal	Brøker og brøkretneregler
Chance og sandsynlighed	Data - Indsamling og forståelse
Decimaltal og brøker	Division
Faglig læsning - Begreber	Faglig læsning - Find tallet
Farlig læsning - Tekst opgaver	Funktioner

EduLab: Case Study

- Student profiling
 - Find common use patterns
 - Model student as histogram over behaviors
 - Detect unproductive sessions
 - Common framework usable for most log data

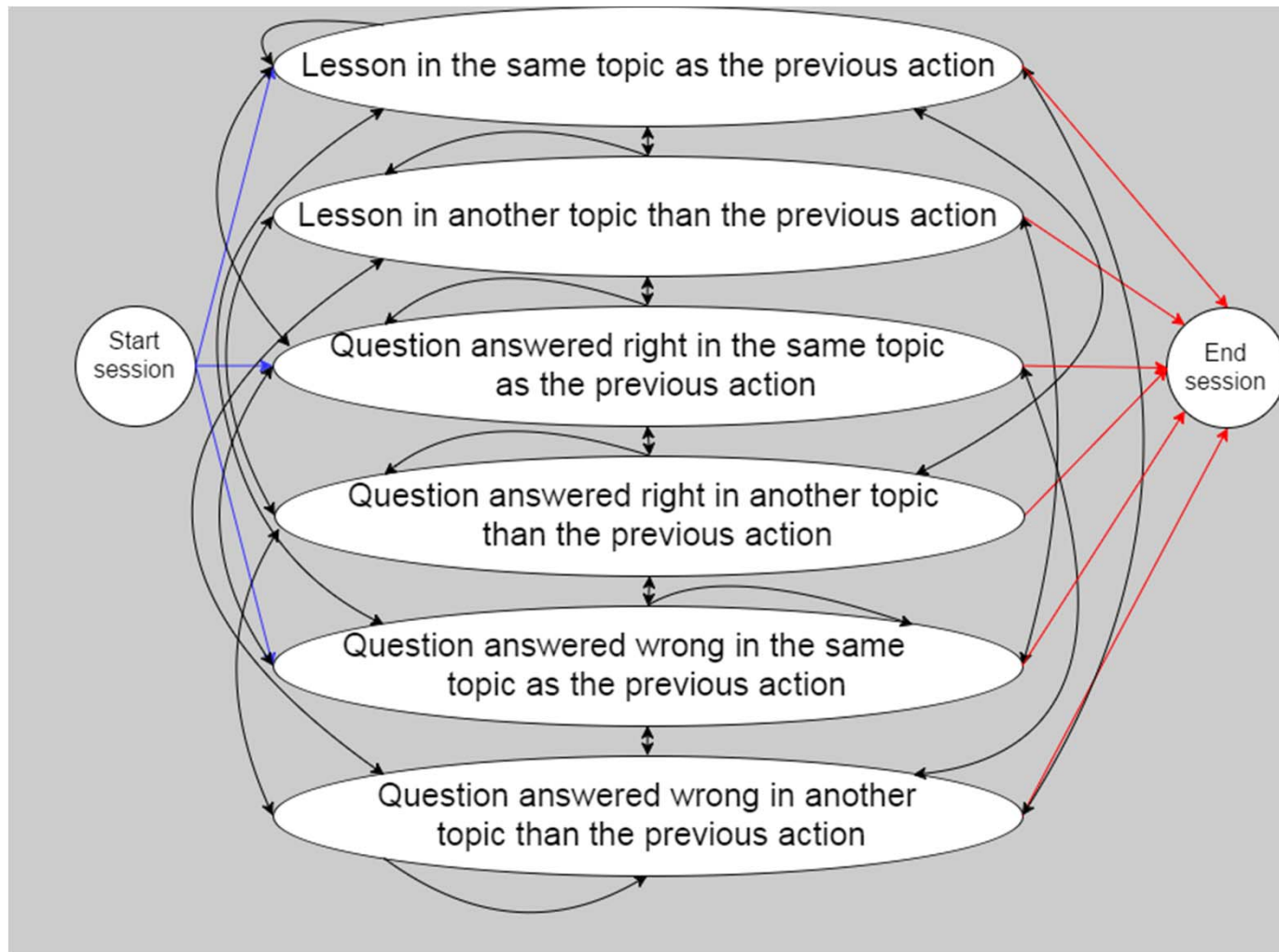


EduLab: Case Study

- **Problem definition**
 - Given a sequence of user activities, mined from logs, find common patterns in user behavior
 - Results should be interpretable by humans
- **Motivation**
 - Find unknown user behavior to grant new insight in the use of the system
- **Results**
 - Found 11% unproductive sessions leading to insight into smaller potential changes to the system.

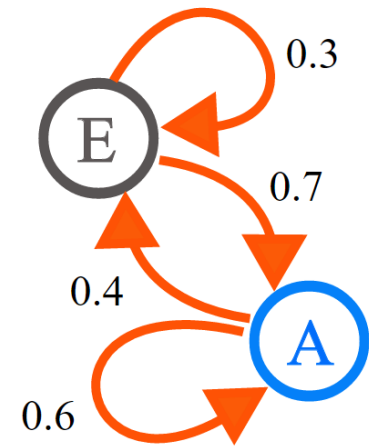
EduLab: Case Study

- Example of state space



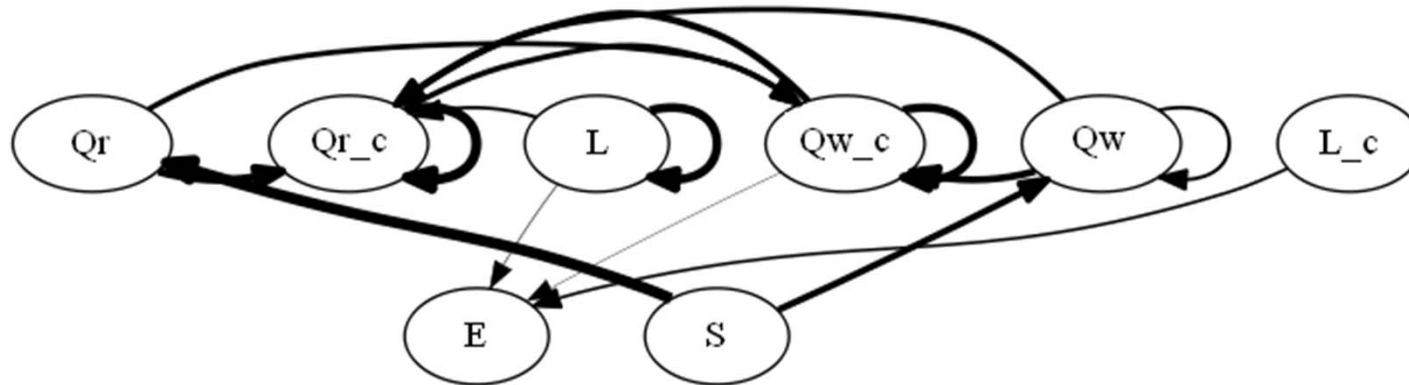
EduLab: Case Study

- Method
 - Model user behavior as first order Markov chains
 - Algorithm (Modified K-means to Markov chains)
 - Initialize by generating K random Markov chains
 - Assign each action sequence to most probable chain
 - Recompute the K chains
 - Repeat until some convergence criteria is met
 - The resulting Markov chains are analyzed for insight
- Data
 - 1.08M sessions for 7th to 8th grade students for this school year.
 - Done for $K=6$



EduLab: Case Study

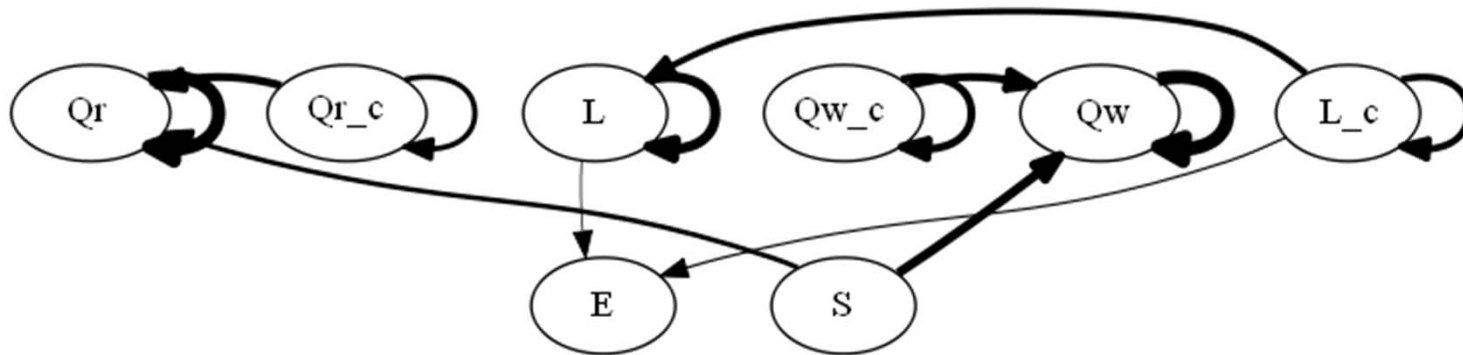
- Chain representing mixture of right and wrong answers to questions



- Qr = question correct, L=lesson, Qw= question wrong

EduLab: Case Study

- Chain representing *unproductive* sessions



- Corresponding to 11% of the user sessions.
- Model is currently being used on a variety of different state spaces.

Other Ongoing Projects



- Knowledge tracing (EduLab)
- Constrained recommender systems for learning materials (EduLab)
- Detecting ghost-writing in high school assignments (MaCom/Lectio)
- Automatic meta-tagging of learning materials (Gyldendal)
- Similarity among quizzes (Clio Online)
- Curriculum trainer (MaCom/Lectio)

Innovation Network:

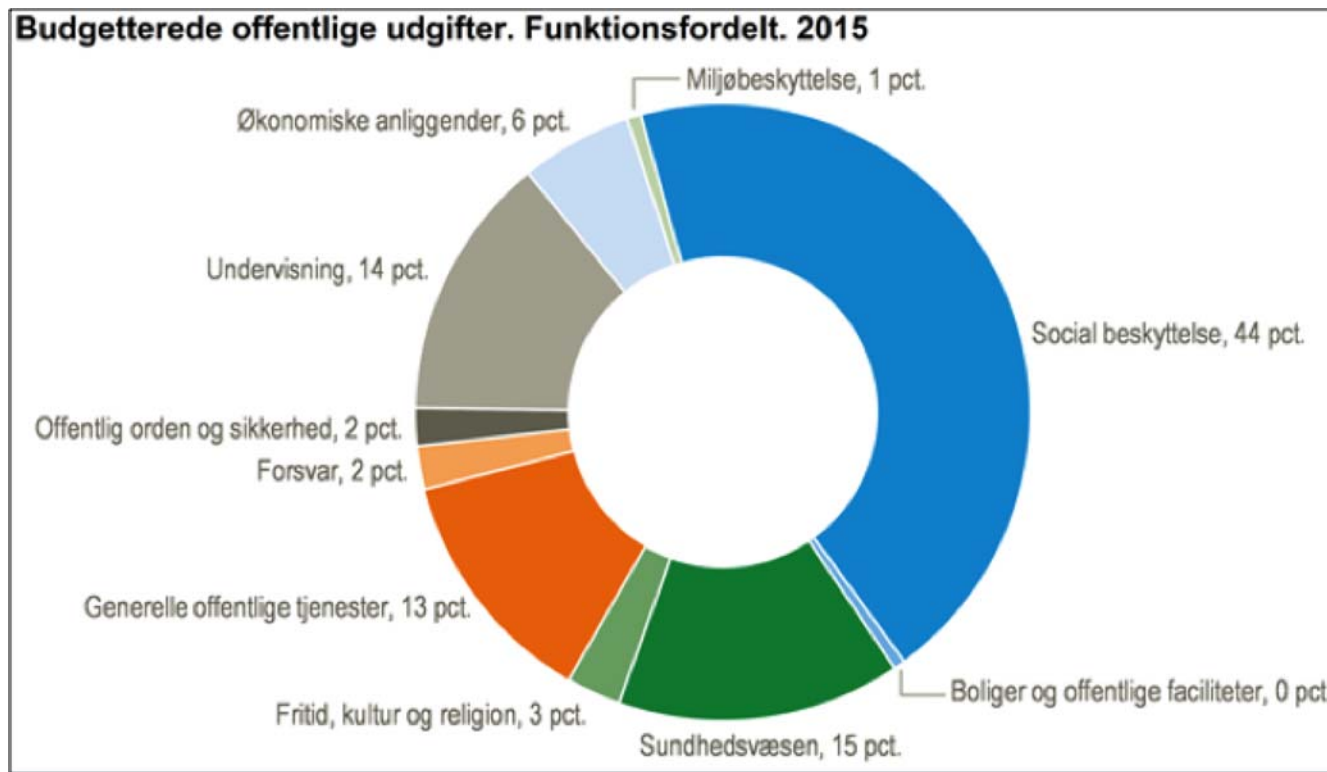


- Meetings every 6 month
- Objective: Exchange knowledge, discuss potential new project ideas
- Interested companies: Gyldendal, EasyCorrect, Writereader, MaCom, EduLab, Clio Online, Egmont Fonden
- Interested public sector: Central Region Denmark, City of Copenhagen, STIL, Capital Region Denmark, Municipality of Naestved

Some facts:



Education accounts for 14% of total public expenses



Big Data Driven Innovation
in Education
will have huge impact



- There is a huge interest
- The number of users (students, teachers, parents etc.) involved are massive
- The preliminary results from the the three cases indicates a large potential for development of improved or new products with benefits for customers and society
- Research access to data may - in some cases - be a key challenge for harvesting the full potential