Data-Driven Education: Technologies and Directions

•••

Rakesh Agrawal
Data Insights Laboratories, San Jose, CA
Visiting Professor, DIAS Lab, EPFL, Switzerland

November 2, 2017

Datafying Education: A Research Expedition

How to enhance the quality of the electronic textbooks?

How to form teams of students in a class?

How to create study plans for courses?

Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research

Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research

Data Mining for Enhancing Electronic Textbooks

Algorithmic enhancement of textbooks for enriching reading experience

References to selective web content

Links to authoritative articles [AGK+10], images [AGK+11b] and videos [ACG+14], based on the focus of the section

References to prerequisites

Links to concepts necessary for understanding the present section, derived using a model of a how students read textbooks [AGK+13]

Diagnostic tools for identifying weaknesses in textbooks

Within section deficiencies

Complexity of writing and dispersion of concepts in the section [AGK+11a]

Across sections deficiencies

Comprehension burden due to non-sequential presentation of concepts [ACG+12]

Validation on textbooks from U.S.A and India, on different subjects, across grades

Joint work with Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, *et al.*

Data Mining for Enhancing Electronic Textbooks

Algorithmic enhancement of textbooks for enriching reading experience

References to selective web content

Links to authoritative articles [AGK+10], images [AGK+11b] and videos [ACG+14], based on the focus of the section

References to prerequisites

Links to concepts necessary for understanding the present section, derived using a model of a how students read textbooks [AGK+13]

Diagnostic tools for identifying weaknesses in textbooks

Within section deficiencies

Complexity of writing and dispersion of concepts in the section [AGK+11a]

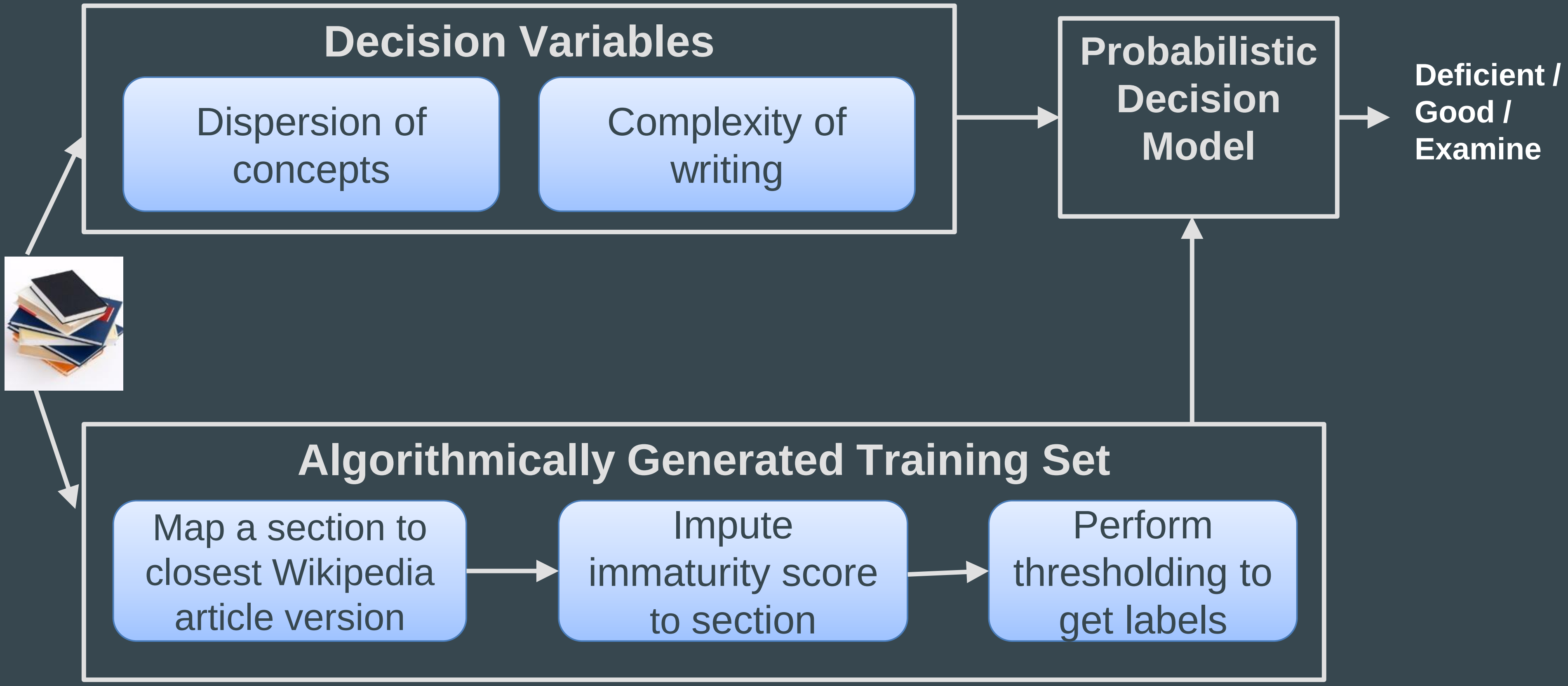
Across sections deficiencies

Comprehension burden due to non-sequential presentation of concepts [ACG+12]

Validation on textbooks from U.S.A and India, on different subjects, across grades

Joint work with Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, *et al.*

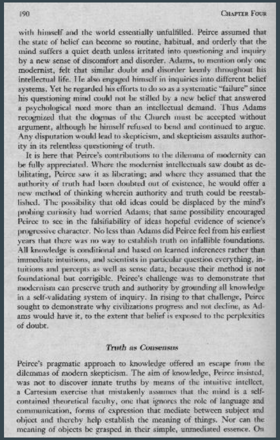
Identification of Deficient Sections



See our WWW-2011 paper

Concept Dispersion

Many unrelated concepts ➔ Hard to understand section


190 CHAPTER FOUR

with himself and the world essentially unfulfilled. Peirce assumed that the state of belief can become so routine, habitual, and orderly that the mind suffers a quiet death unless irritated into questioning and inquiry by a new sense of discomfort and disorder. Adams, to mention only one modernist, felt that similar doubt and disorder keenly throughout his intellectual life. He also engaged himself in inquiries into different belief systems. Yet he regarded his efforts to do so as a systematic "failure" since his questioning mind could not be stilled by a new belief that answered a psychological need more than an intellectual demand. Thus Adams recognized that the dogmas of the Church must be accepted without argument, although he himself refused to bend and continued to argue. Any disputation would lead to skepticism, and skepticism assaults authority in its relentless questioning of truth.

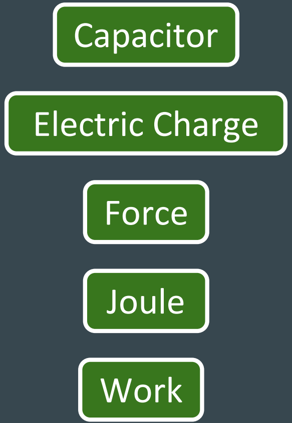
It is here that Peirce's contributions to the dilemma of modernity can be fully appreciated. Where the modernist intellectuals saw doubt as debilitating, Peirce saw it as liberating; and where they assumed that the authority of truth had been doubted out of existence, he would offer a new method of thinking wherein authority and truth could be reestablished. The possibility that old ideas could be displaced by the mind's probing curiosity had worried Adams; that same possibility encouraged Peirce to see in the falsifiability of ideas hopeful evidence of science's progressive character. No less than Adams did Peirce feel from his earliest years that there was no way to establish truth on infallible foundations. All knowledge is conditional and based on learned inferences rather than immediate intuitions, and scientists in particular question everything, intuitions and percepts as well as sense data, because their method is not foundational but corrigible. Peirce's challenge was to demonstrate that modernism can preserve truth and authority by grounding all knowledge in a self-validating system of inquiry. In rising to that challenge, Peirce sought to demonstrate why civilizations progress and not decline, as Adams would have it, to the extent that belief is exposed to the perplexities of doubt.

Truth as Consensus

Peirce's pragmatic approach to knowledge offered an escape from the dilemmas of modern skepticism. The aim of knowledge, Peirce insisted, was not to discover innate truths by means of the intuitive intellect, a Cartesian exercise that mistakenly assumes that the mind is a self-contained theoretical faculty, one that ignores the role of language and communication, forms of expression that mediate between subject and object and thereby help establish the meaning of things. Nor can the meaning of objects be grasped in their simple, unmediated essence. On

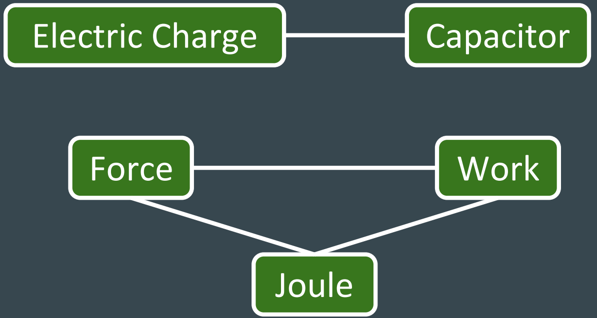
Text Analysis ➔

Terminological Noun Phrases A*N+



Concept Graph ➔

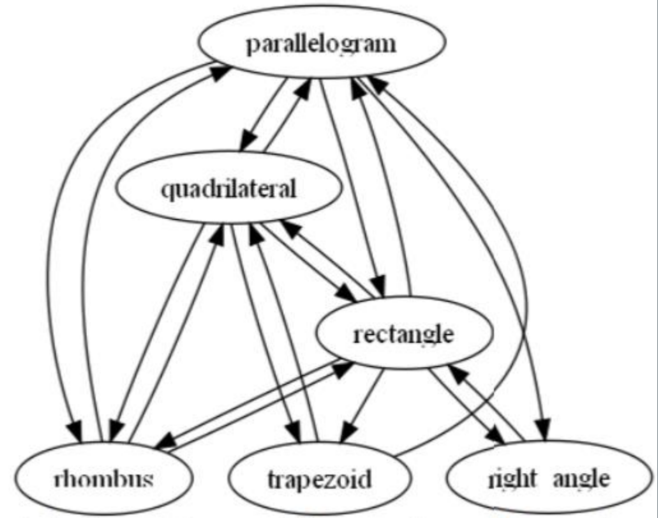
Edge (c1, c2) if c1 and c2 are related



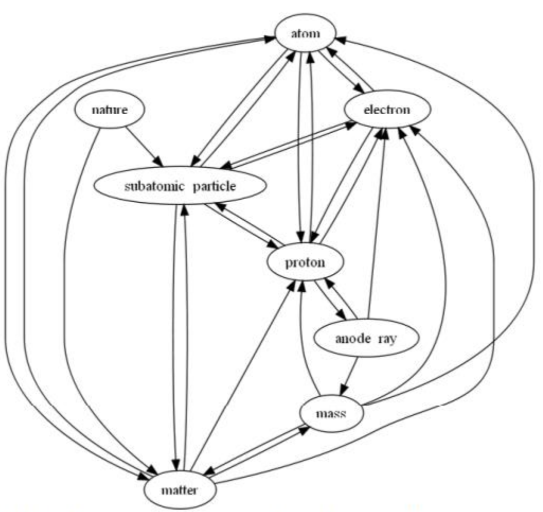
Dispersion(*s*) := Fraction of unrelated concept pairs

- (1 – Edge Density) of the concept graph

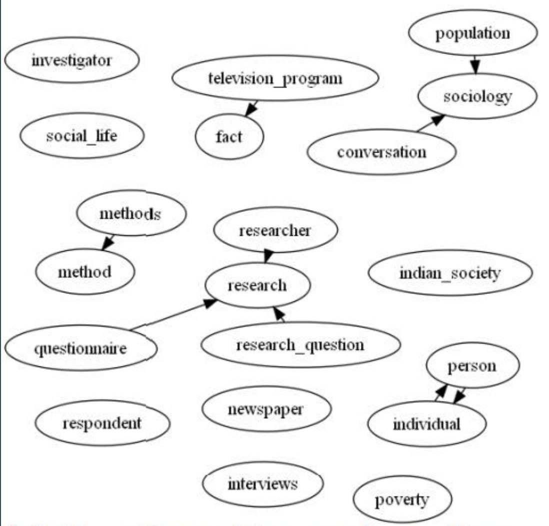
Concept Dispersion



(a) A section with very low dispersion (Grade IX Mathematics book)



(b) A section with low dispersion (Grade IX Science book)



(c) A section with very large dispersion (Grade XII Sociology book)

Writing Complexity

- ➢ Readability Formulas (~100 years of research)
- ➢ More than 200 formulas in widespread use

Flesch Reading Ease Score [17]	206.835	−	84.6	×	S/W	−	1.015	×	W/T
Flesch-Kincaid Grade Level [31]	−15.59	+	11.8	×	S/W	+	0.39	×	W/T
Dale-Chall Grade Level [14]	14.862	−	11.42	×	D/W	+	0.0512	×	W/T
Gunning Fog Index [23]			40	×	C/W	+	0.4	×	W/T
SMOG Index [37]	3.0	+	$\sqrt{30}$	×	$\sqrt{C/T}$				
Coleman-Liau Index [10]	−15.8	+	5.88	×	L/W	−	29.59	×	T/W
Automated Readability Index [46]	−21.43	+	4.71	×	L/W	+	0.50	×	W/T

C	=	Number of words with three syllables or more
D	=	Number of words on the Dale Long List
L	=	Number of letters
S	=	Number of syllables
T	=	Number of sentences
W	=	Number of words

- ➢ Sentence Length:
  - ○ Avg. number of words per sentence
- ➢ Word Length:

Illustrative Result: Deficient Section

CHAPTER 2

FORMS OF BUSINESS ORGANISATION

2.7 CHOICE OF FORM OF BUSINESS ORGANISATION

After studying various forms of business organisations, it is evident that each form has certain advantages as well as disadvantages. It, therefore, becomes vital that certain basic considerations are kept in mind while choosing an appropriate form of

operations. Cooperative societies and companies have to be compulsorily registered. Formation of a company involves a lengthy and expensive legal procedure. From the point of view of initial cost, therefore, sole proprietorship is the preferred form as it involves least expenditure. Company form of organisation, on the other hand, is more complex and involves greater costs.

(ii) Liability: In case of sole proprietorship and partnership firms, the liability of the owners/partners is unlimited. This may call for paying the debt from personal assets of the owners. In joint Hindu family business, only the *karta* has unlimited liability. In cooperative societies and companies, however, liability is limited and creditors can force payment of their claims only to the extent of the company's assets.

in nature and require professionalised management, company form of organisation is a better alternative. Proprietorship or partnership may be suitable, where simplicity of operations allow even people with limited skills to run the business. Thus, the nature of operations and the need for professionalised management affect the choice of the form of organisation.

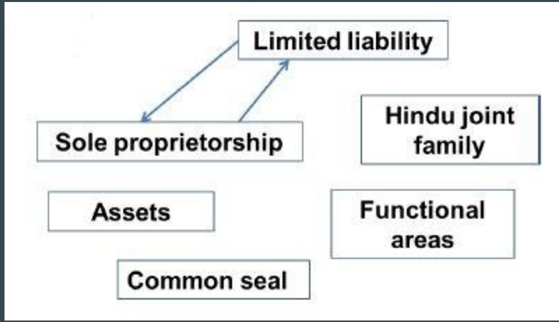
(v) Capital considerations: Companies

above are inter-related. Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be

organisations one by one. In Table 2.5, we analysed characteristics of different forms of organisations taken together so as to enable you to understand on a comparative basis as to where a form of organisation stands in comparison to others in respect of select features.

Tested on every grade 9-12 NCERT textbook in India

❖ High dispersion:



❖ Long sentences:

Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be appropriate when the same business is carried on a large scale.

Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research

Optimal Grouping of Students in a Large Class

Given:

- a class of *N* students
- each exhibiting a different ability level, $\theta_i \in \mathbb{R}_{\geq 0}$

How to partition them into *k* groups, each of size *n*, so that the overall gain from peer learning is maximized ($N = k \times n$)

Ability Score θ_i :

- Measured via a test (e.g. using Item Response Theory)

Work with Sharad Nandanwar & M.N. Murty (Under Submission)
Extension of work with Behzad Golshan & Evimaria Terzi (KDD 2014)

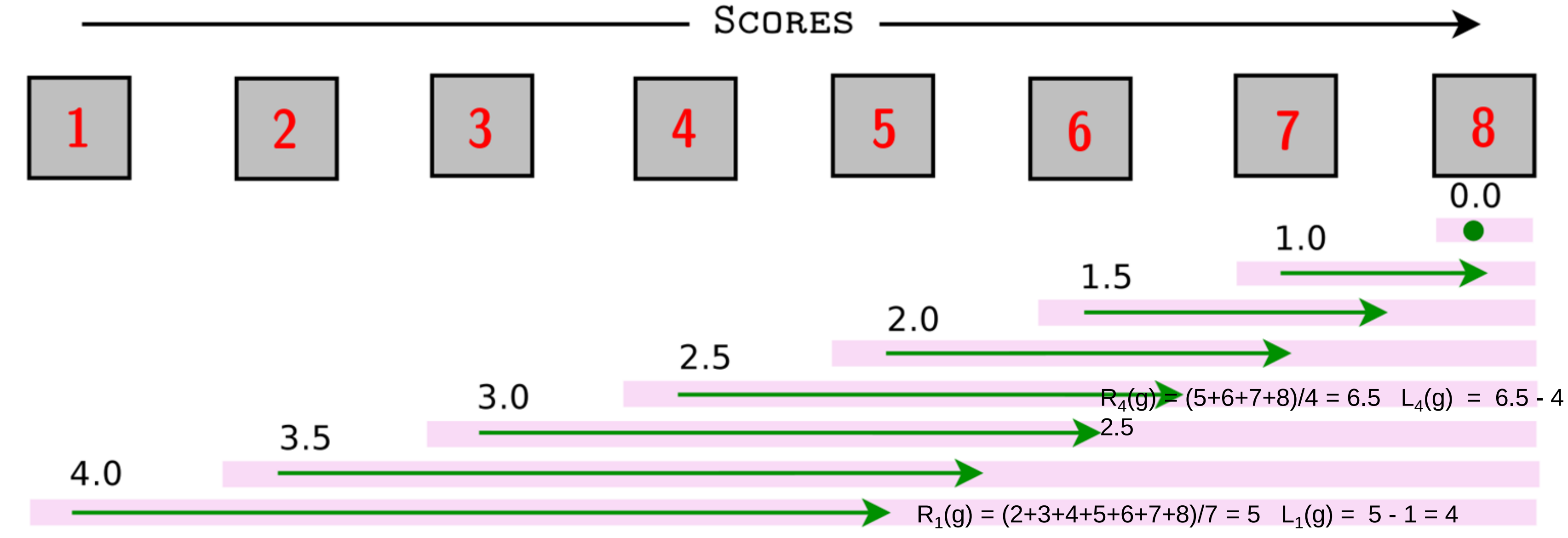
Prevalent Approaches

- Stratified (Ability-based grouping)
  - Put best with the best
- Pseudo-random (Diversity-based grouping)
  - Group students of all abilities together
- Inconclusive verdict from the empirical studies on the effectiveness [Richer76, Kulik92, Grossen96]
- Any computational alternative?

Model

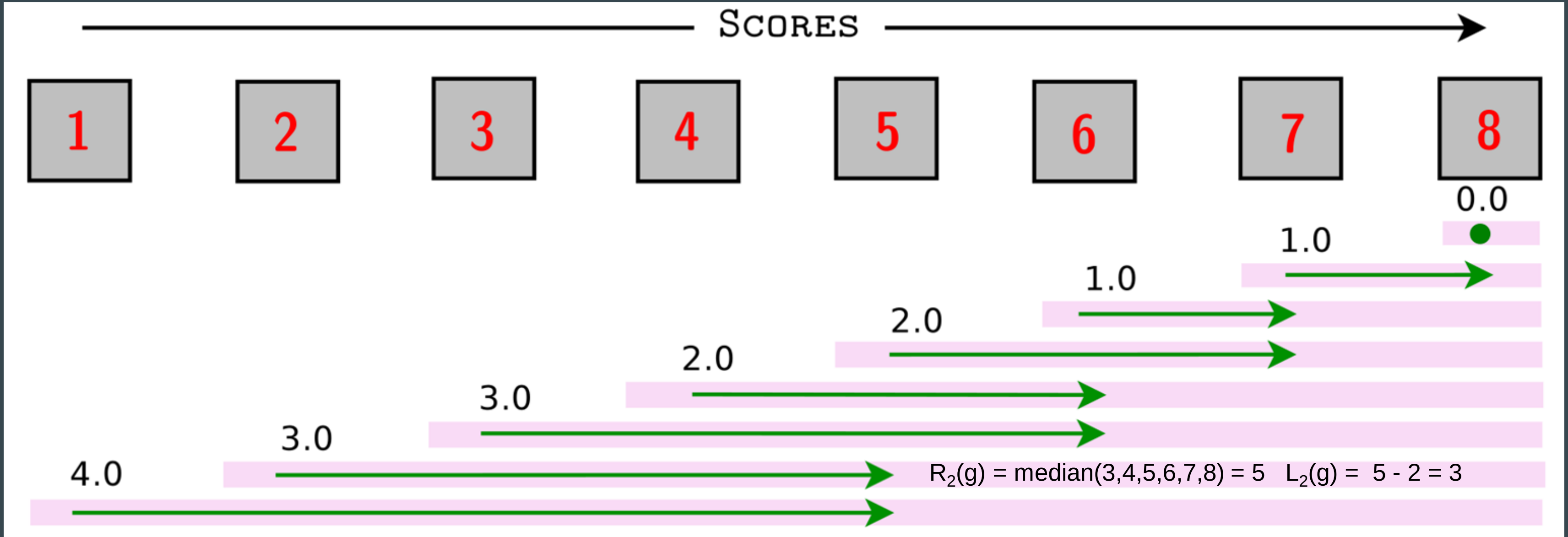
- Every student gains from higher-ability peers [Vygotski]
- Learning gain for student *i* in group *g*: $L_i(g) = R_i(g) - \theta_i$
  - ➔ $R_i(g)$ is a function of *i*'s superior peers in group *g*
  - ➔ $R_i(.)$ is different for different students in the same group
  - ➔ $R_i(.)$ is different for the same student in different groups
  - ➔ Examples of $R_i(g)$: Mean, Median, p percentile

Illustration: $R_i(g)$ = mean of the scores of the superior peers



Box *i* has the score of student *i*

Illustration: $R_i(g)$ = median of the scores of the superior peers



Box *i* has the score of student *i*

Grouping Algorithm: Magic Partitions

Optimal for:

- $R_i(g)$ = mean of the scores of the superior peers
- $R_i(g)$ = median of the scores of the superior peers
- $R_i(g)$ = *p*-percentile scores of the superior peers where *100-p* divides 100

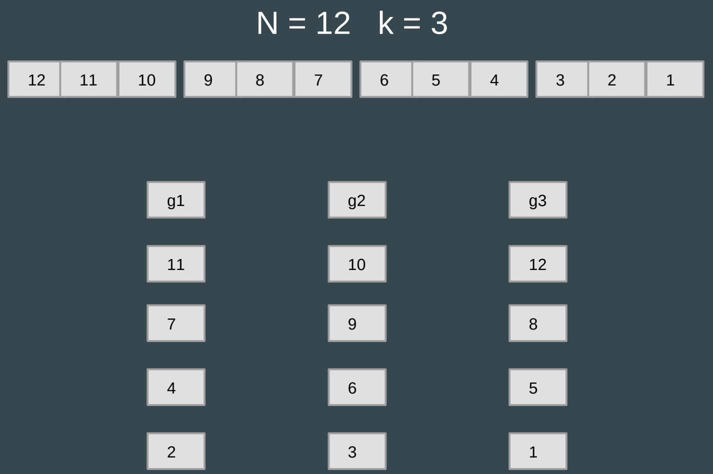
Magic Partitions in Action

Assume scores are unique and *k* divides *N*

1. Sort the scores in descending order

2. Partition the sorted scores into *N*/*k* blocks of *k* scores (students) each

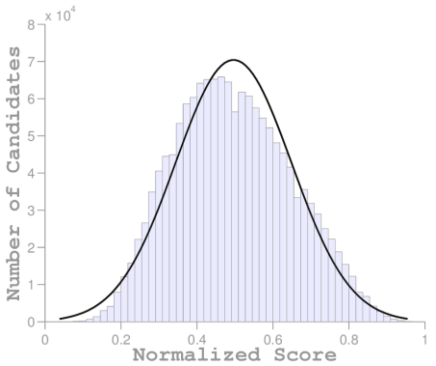
3. Assign randomly from each block exactly one score (student) to each group



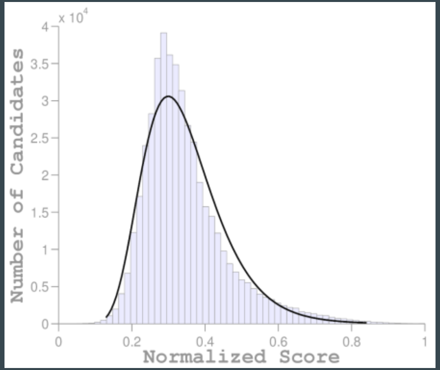
Experiment: Learning Gain

Data Distribution

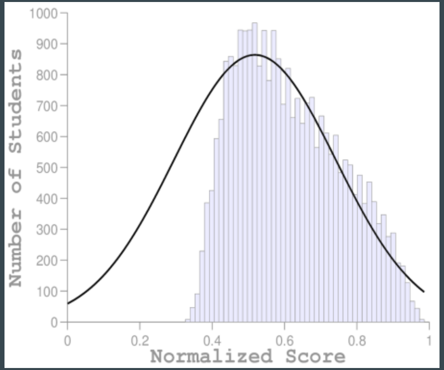
SSC (Normal)



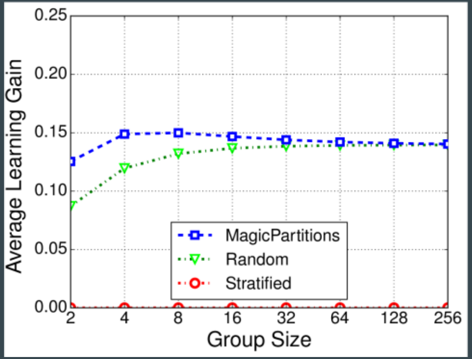
GATE (Log Normal)

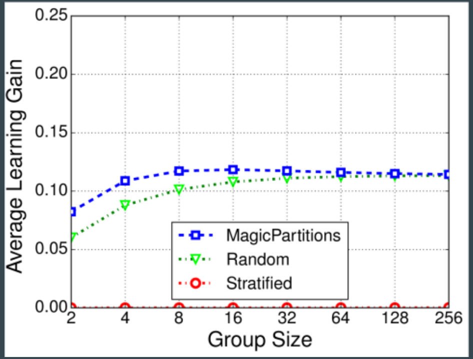


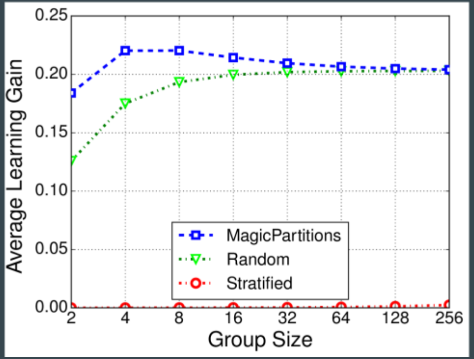
J&K Board (Truncated Normal)



Average Learning Gain







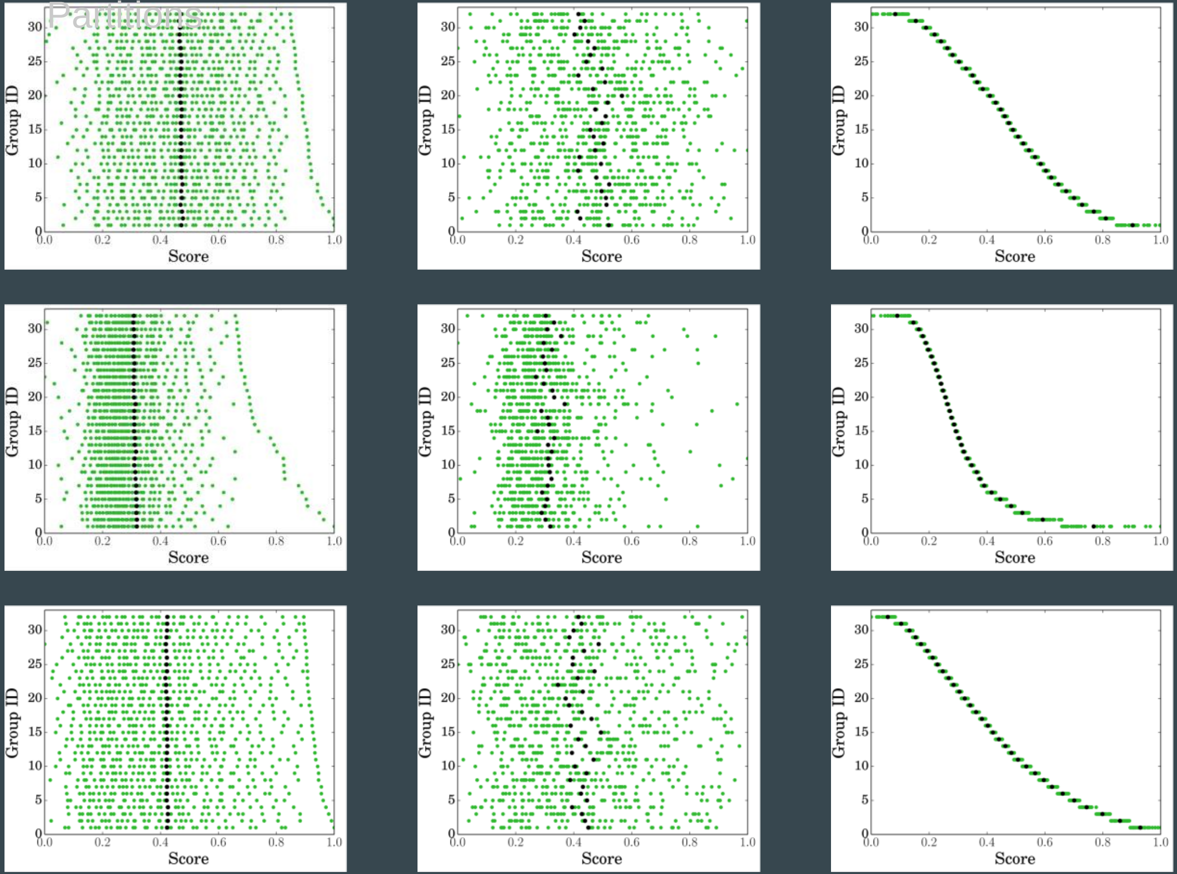
Experiment: Group Structure

Magic-Partitions | Random | Stratified

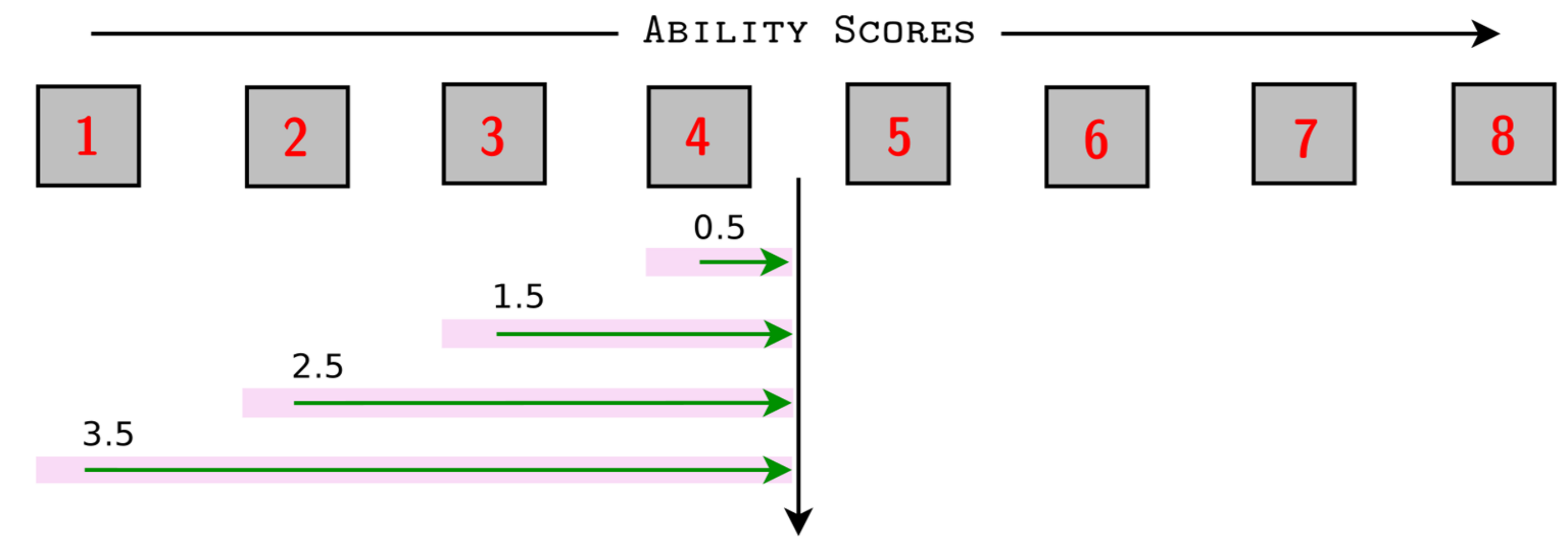
SSC

GATE

J & K Board



Prior Work: [AGT KDD-2014]



- Learning gain only for below average students gained
- Time complexity of the proposed algorithm left open for future work

Grouping Students (Recap)

- Computational approach points to a grouping strategy better from the conventional strategies
- For the objective of maximizing overall gain, the proposed partitions are optimal for a variety of reasonable learning models
- Low complexity of the algorithm
- Ongoing work: Incorporate learning gain from teaching, social constraints

Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research

Synthesizing Study Plans

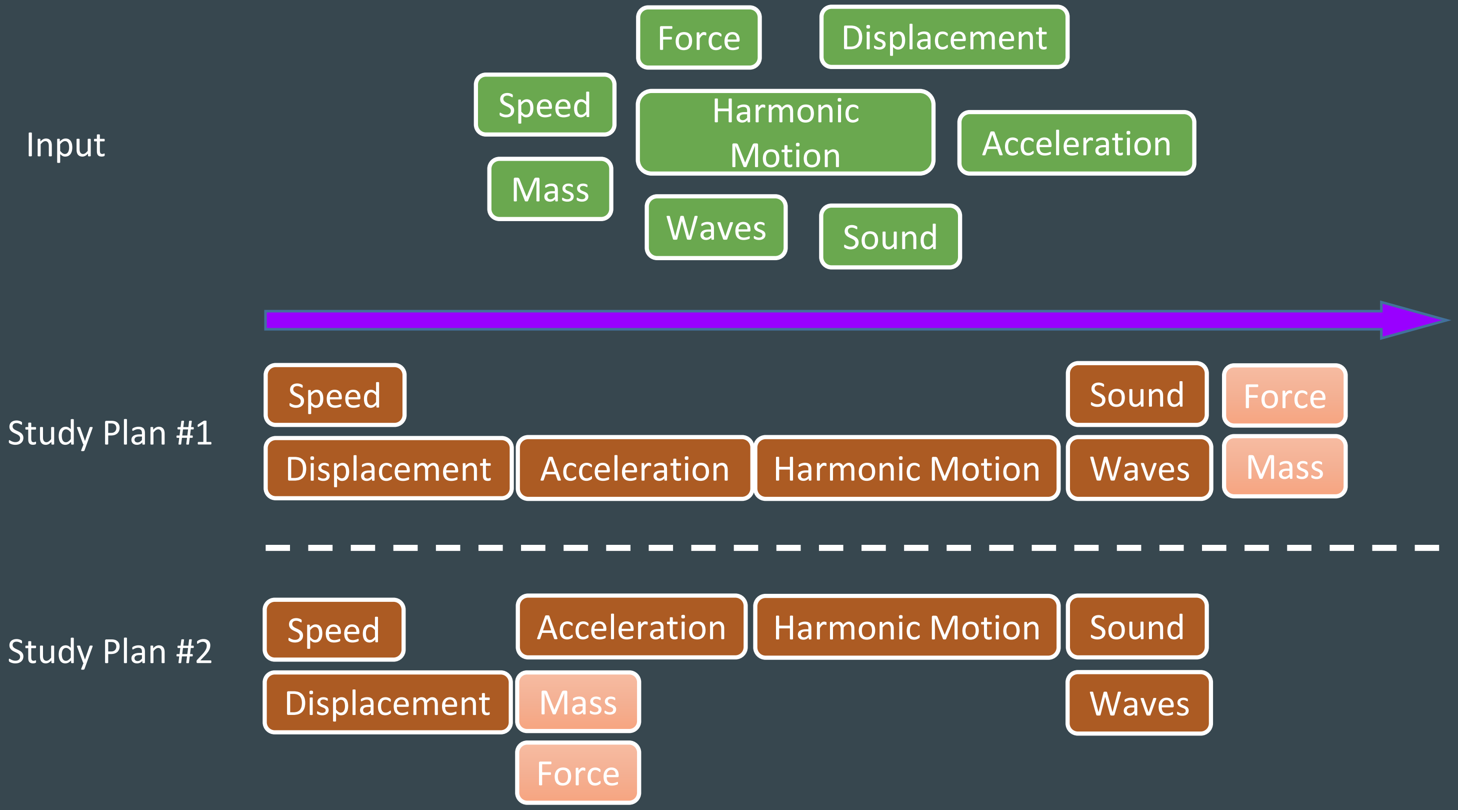
Imagine you are an instructor who wants to offer a new course

You know the concepts you want to teach in the course, but need help with formulating the study plan:

a. What concepts should you cover in one session
b. The sequencing of sessions

Work with Behzad Golshan and Evangelos Papalexakis (EDM 2106)

Study Plans


Input
Force
Displacement
Speed
Harmonic Motion
Acceleration
Mass
Waves
Sound
Study Plan #1
Speed
Sound
Force
Displacement
Acceleration
Harmonic Motion
Waves
Mass
Study Plan #2
Speed
Acceleration
Harmonic Motion
Sound
Displacement
Mass
Waves
Force

Axioms

- Learning Unit
  - *A group of coherent concepts suitable to be covered together*
  - Cohesion: Concepts within a learning unit must be closely related
  - Isolation: Concepts in different learning units must be independent
  - Unity: A concept should be covered in one unit
- Study plan
  - *An ordering of some number of learning units*
  - Prerequisite compliance: L1 < L2 => concepts in L2 not needed for L1
  - Locality of references: L2 builds upon L1 => L2 should come soon after L1

Problem Statement

Given a set of concepts,

- Partition them into a given number of learning units, and
- Provide a sequencing of learning units

such that an objective function *f* is minimized

PROBLEM 1 (STUDY PLAN DESIGN PROBLEM). *Given a concept graph* $G = \langle V, E \rangle$ *with* $n > 0$ *nodes, and the number of desired learning units* m ($m \leq n$), *output an ordered vector of learning units* $\mathcal{L} = \langle L_1, L_2, \cdots, L_m \rangle$ *to*

$$\textit{Minimize: } f(\mathcal{L})$$
$$\textit{s.t. } \forall i : L_i \subseteq V, L_i \neq \phi, \textit{and}$$
$$\cup_i L_i = V.$$

Objective Function

$$f(L) = \sum_{\pi(u) > \pi(v)\ \&\ (u,v) \in E} (\pi(u) - \pi(v))\, C_p$$

Penalize if u is taught after v but v is a prerequisite for understanding v

$$+ \sum_{\pi(u) < \pi(v)\ \&\ (u,v) \in E} (\pi(v) - \pi(u))\, C_r$$

Penalize if u is a prerequisite for v but u is taught much earlier

$$+ \sum_{\pi(u) = \pi(v)\ \&\ (u,v) \notin E} C_c$$

Penalize if u and v placed in the same learning unit but are unrelated

π(u): gives the position of the learning unit in which the concept u is covered

Prerequisite Compliance Violation

Locality of Reference Violation

Cohesion Violation (Also Isolation)

Unity Violation penalized by first two terms

Problem Complexity

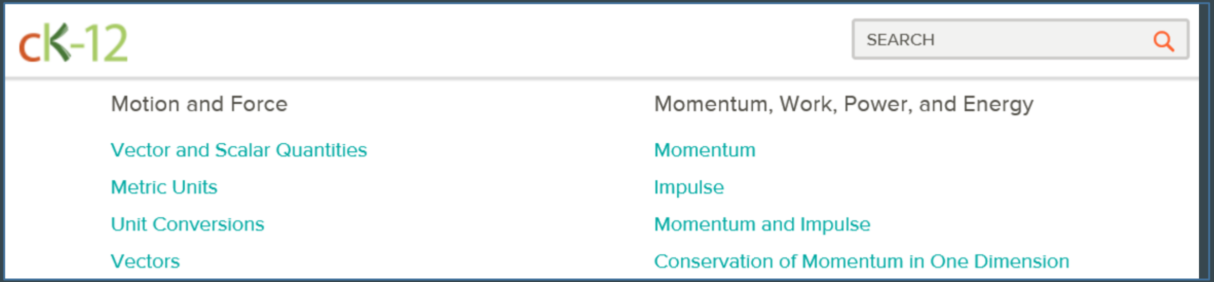
NP-Complete

- Minimum Linear Arrangement (minLA) problem reduces to our problem

See [AGP EDM2016] for our solution

Experiment

Input: 139 high school physics concepts from CK12.org



Synthesized Study Plan

Unit 1(20 concepts)	**Unit 2** (21 concepts)	**Unit 3** (14 concepts)	**Unit 4** (18 concepts)
buoyancy	acceleration	atom	calorimetry
euclidean vector	angular momentum	bohr model	change of state
force	angular velocity	conservation of energy	combined gas law
free body diagram	centripetal force	elastic collision	conversion of units
friction	circular motion	inelastic collision	double-slit experiment
impulse	displacement	kinetic energy	energy
inclined plane	keplers laws of ...	mass versus weight	energy conversion

Unit 5 (27 concepts)	**Unit 6** (28 concepts)	**Unit 7** (11 concepts)
ammeter	beat	doppler effect
capacitor	color	general relativity
capacitors in circuits	concave lens	half-life
electric charge	conduction	length contraction
electric current	curved mirror	mathematical physics
electric field	diffraction	newtons law of univer...
electric power	diffraction grating	rc time constant

User Study

- Recruited 9 domain experts (Physics teachers, Graduate students)
- They were given the following tasks:

1) *Count the number of odd concepts in each learning unit that you believe do not belong to the unit*
2) *Without changing any of the learning units proposed, what order do you suggest?*

Results of the User Study

- Number of concepts that do not belong in the respective unit:

	Min	Max	Median	Mean	# Concepts
Unit 1	1	6	3.0	3.4	20
Unit 2	0	3	1.0	1.1	20
Unit 3	1	7	3.5	3.7	14
Unit 4	0	5	2.0	1.8	18
Unit 5	0	4	1.0	1.0	26
Unit 6	0	3	0.5	0.9	28
Unit 7	0	5	1.0	1.4	11

- Only two participants ordered the units somewhat differently
- The high school Physics teacher: our study plan was very clever

Synthesizing Study Plans (Recap)

- Formalized the problem of synthesizing study plans automatically
- Provided a novel and pragmatic solution
- The proposed method did not use domain specific knowledge
  - Generalizing to other areas seems promising
- Experimental results as well as the user study show that the problem of creating study plans is amenable to computational approaches

Further Work

- Incorporate user modeling into the system
  - Creating study plans that suit students background/interests/abilities
- Investigate how human input (implicit or explicit) can improve the quality of generated study plans

Outline

1. Enhancing the quality of the electronic textbooks
2. Grouping students in a class
3. Synthesizing study plans
4. Opportunities for Future Research

Future Research Opportunities

- ❖ Validation of experimental results through deployment
- ❖ Synergies with crowd-sourcing approaches
- ❖ Use of logs of interactions data and personalization
- ❖ Performance evaluation methodologies and benchmarks
- ❖ Issues related to privacy, security, confidentiality, copyright, royalty

…

Magic happens when what is desperately needed
meets what is technically feasible

Selected References

[AGK+10] Rakesh Agrawal, Sreenivas Gollapudi, Krishnaram Kenthapadi, Nitish Srivastava, Raja Velu. "Enriching Textbooks Through Data Mining". DEV 2010.

[AGK+11a] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Identifying Enrichment Candidates in Textbooks". WWW 2011.

[AGK+11b] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Enriching Textbooks With Images". CIKM 2011.

[ACG+12] Rakesh Agrawal, Sunandan Chakraborty, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Empowering Authors to Diagnose Comprehension Burden in Textbooks". KDD 2012.

[AGK+13] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Studying from Electronic Textbooks". CIKM 2013.

[AJK14] Rakesh Agrawal, M. Hanif Jhaveri, and Krishnaram Kenthapadi. "Evaluating Educational Interventions at Scale". LAS 2014.

[ACG+14] Rakesh Agrawal, Maria Christoforaki, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, Adith Swaminathan. "Augmenting Textbooks with Videos". ICFCA 2014.

[AGT14] Rakesh Agrawal, Behzad Golshan, Evimaria Terzi. "Grouping Students in Educational Settings". KDD 2014.

[AGP16] Rakesh Agrawal, Behzad Golshan, Evangelos Papalexakis. "Toward Data-Driven Design of Educational Courses''. EDM 2016.

[ANM17] Rakesh Agrawal, Sharad Nandanwar, M.N. Murty. "Grouping Students for Maximizing Learning from Peers''. EDM 2017.


Questions?