

ML for Education NIPS Workshop
12/10/2016

ML Approaches for Learning Analytics: Collaborative Filtering Or Regression With Experts?

Kangwook Lee

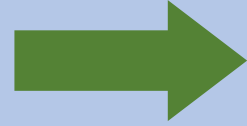


Joint work w/ Jichan Chung, Yeongmin Cha, and Changho Suh

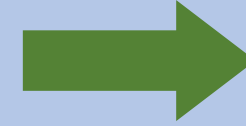


Learning Analytics

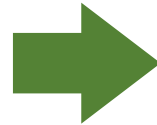
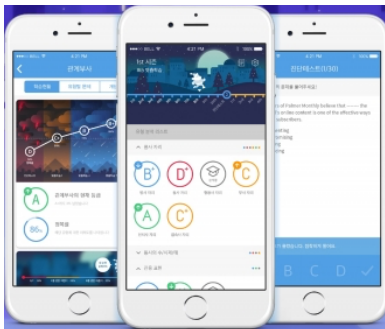
Data Collection



Data Analysis



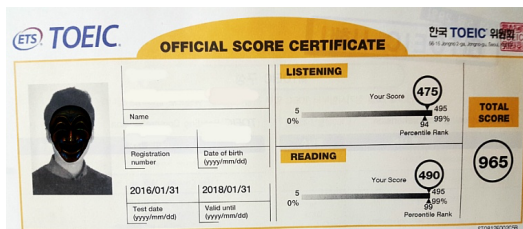
Optimize Learning



- Prediction
- Recommendation
- Personalization
- Content Design
- ...

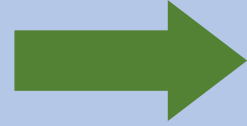


- Rule-based
- Machine Learning

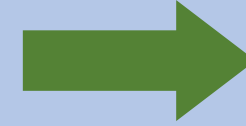


Learning Analytics

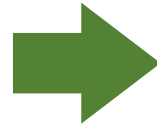
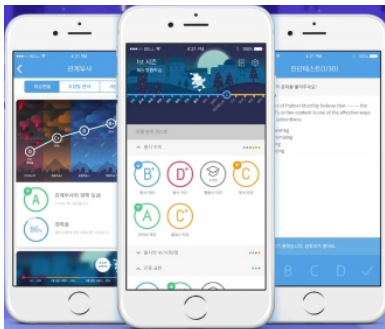
Data Collection



Data Analysis



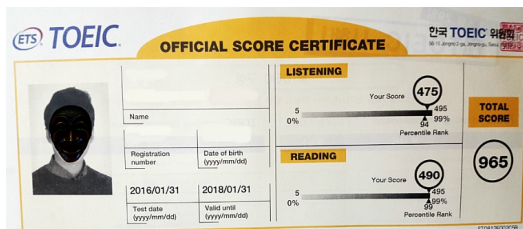
Optimize Learning



- **Prediction**
- Recommendation
- Personalization
- Content Design
- ...

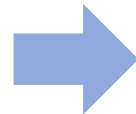
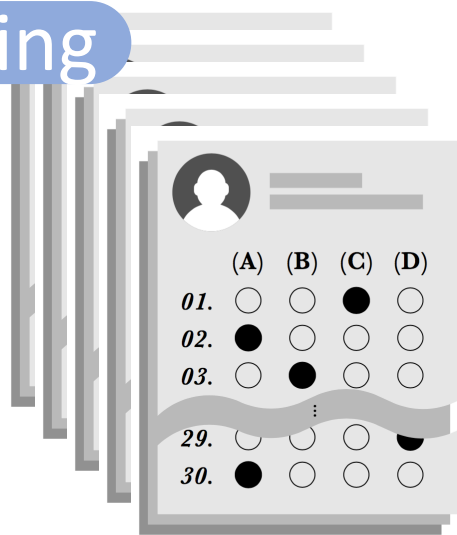


- Rule-based
- **Machine Learning**

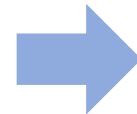


Prediction

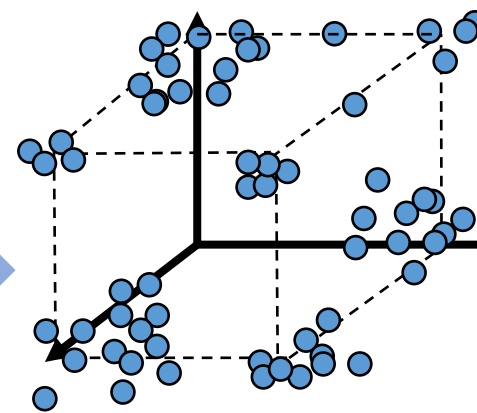
Training



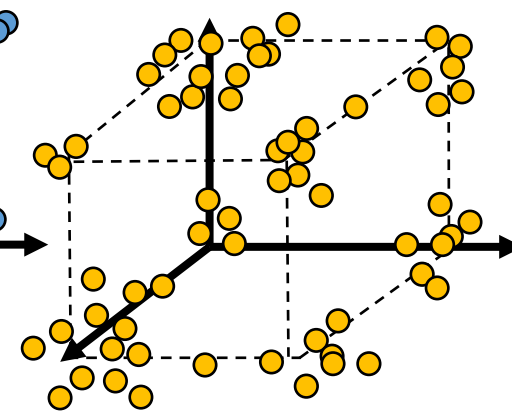
Learning
Algorithm



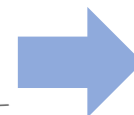
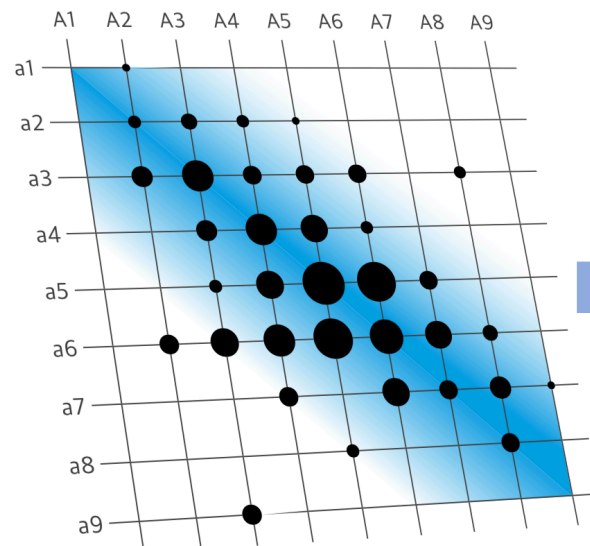
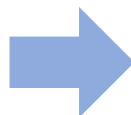
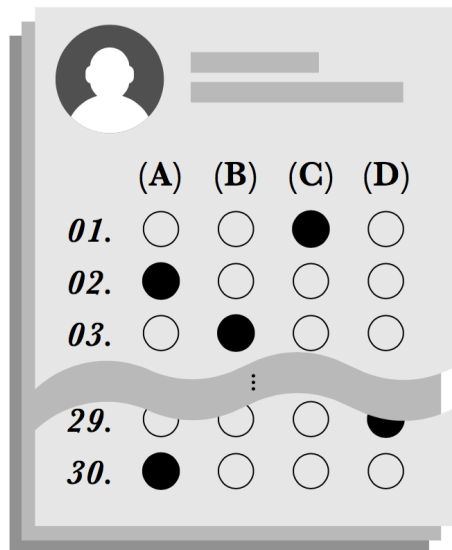
Student Features



Question Features



Prediction



Actual

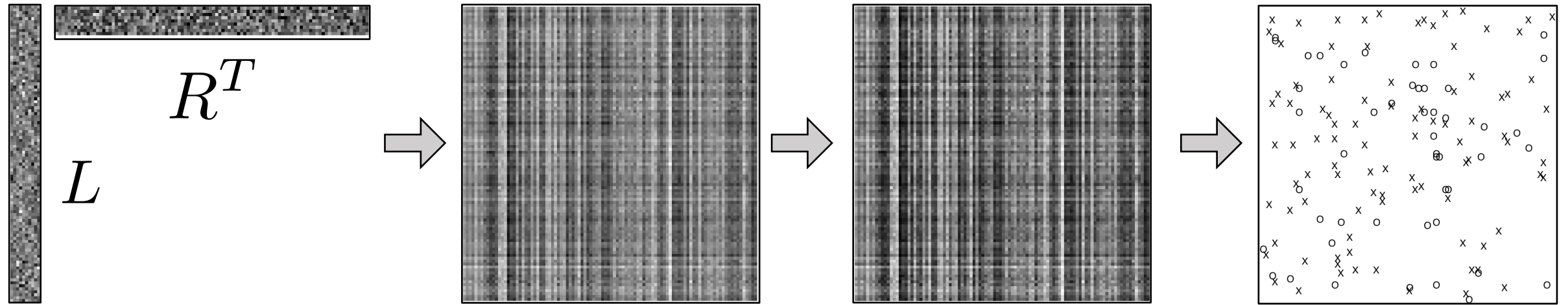
	(A)	(B)	(C)	(D)
31	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
32	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
33	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
34	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
35	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
36	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
37	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
38.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
30	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Predicted

	(A)	(B)	(C)	(D)
31.	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
32.	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
33.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
34.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
35.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
36.	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
37.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
38.	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
30	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

Response Model & Learning

[Bergner et al., 2012], [Lan, Studer, Baraniuk, 2014]



$$L \in [0, 1]^{n \times r}$$
$$R \in [0, 1]^{m \times r}$$

$$X = L \times R^T$$

$$P = \phi(X)$$

Y

Student & Question
features

Level of understanding

Probability of
correct guess

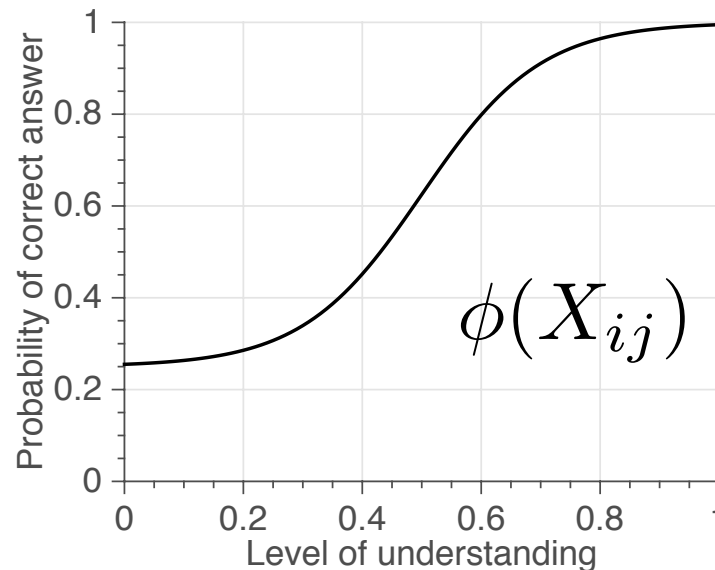
Responses



Learning Algorithm

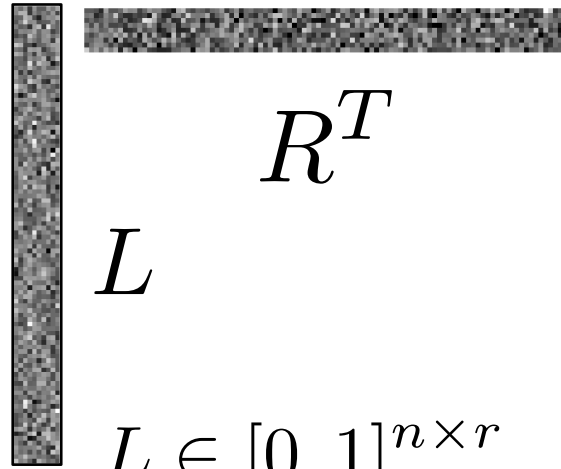
Response Model

- A variation of M2PL (Multidimensional Two-Parameter Logistic) model
- $L_{i,j}$: the level of student i's understanding of the j^{th} hidden concept.
- $R_{i,j}$: the contribution of the j^{th} hidden concept to question i
- R is normalized to sum up 1 so that $X_{i,j} = L_i R_j^T$ is in $[0,1]$
- Two additional concepts for difficulty & outliers:
 - $(r+1)^{\text{th}}$ concept for what is known to everyone
 - $(r+2)^{\text{th}}$ concept for what is not known to everyone (e.g., difficult vocab)
- $P = \text{Logistic}(X)$
- $Y = \text{Bernoulli}(P)$

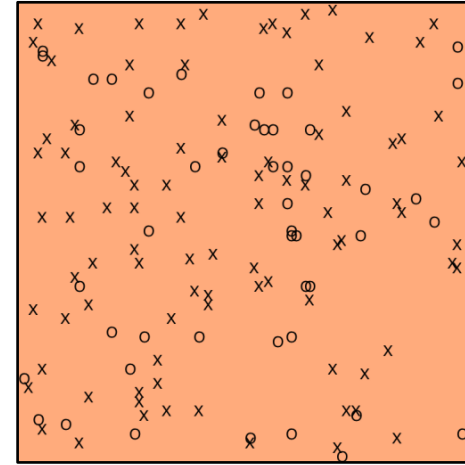
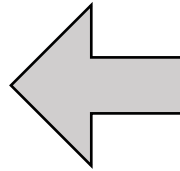


$$\phi(X_{ij}) = \phi_a + \frac{1 - \phi_a}{1 + e^{-\phi_c(X_{ij} - \phi_b)}}$$

Logistic Regression w/ Experts



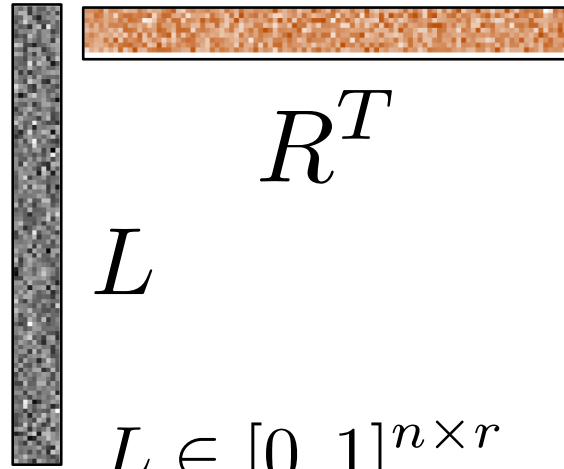
$$L \in [0, 1]^{n \times r}$$
$$R \in [0, 1]^{m \times r}$$



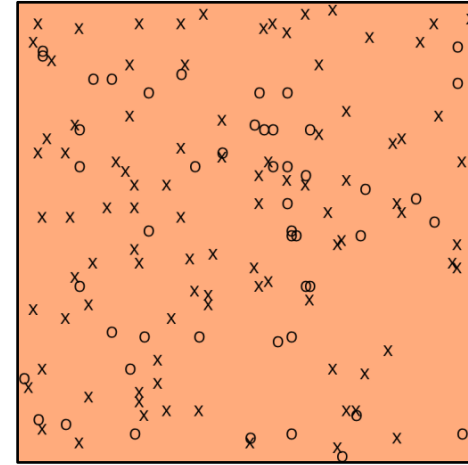
$$Y = \text{Bern}(\phi(LR^T))$$

Logistic Regression w/ Experts

If experts can provide us w/ R ,



$$L \in [0, 1]^{n \times r}$$
$$R \in [0, 1]^{m \times r}$$



$$Y = \text{Bern}(\phi(LR^T))$$

The MLE of L is

$$\min_{L_i} \sum_{j \in \Omega_{i*}} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})]$$

$$\text{s.t. } 0 \leq L_{ij} \leq 1, \quad \sum_j L_{ij} = 1, \quad P_{ij} = L_i R_j^T.$$

Logistic Regression w/ Experts

Question j

— who want to apply for this position are requested to submit their performance.

- (A) You
- (B) Those
- (C) Another
- (D) Some



- Experts
- Crowdsourcing



"Pronoun"
"personal pronoun"

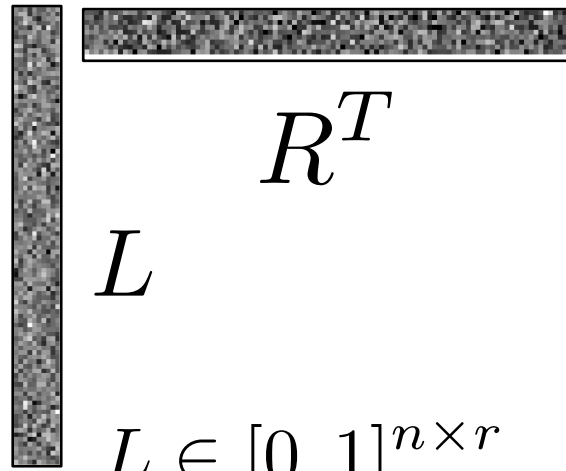
"Long
subjects (?)"
"prepositional
phrase"

R_j

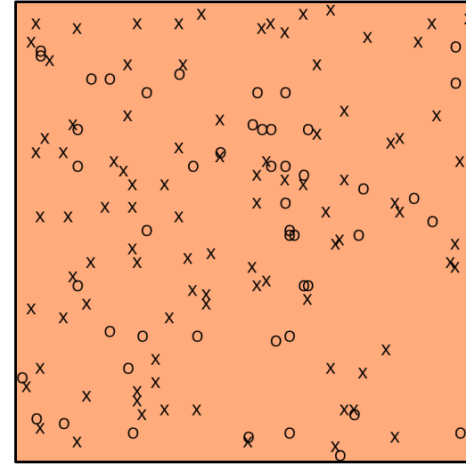
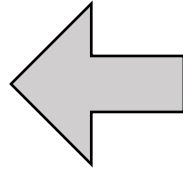
0	.5	.3	0	0	...	0	.1	.1	0
---	----	----	---	---	-----	---	----	----	---

- Noisy, subjective, ...
- (Observation) # of concepts is usually very large (prone to overfitting)
- Depends on human knowledge

Binary Matrix Completion (BMC)



$$L \in [0, 1]^{n \times r}$$
$$R \in [0, 1]^{m \times r}$$



$$Y = \text{Bern}(\phi(LR^T))$$

Estimate L and R by solving

$$\min_{L, R} \sum_{(i, j) \in \Omega} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] + \mu \|LR^T\|_*$$

$$\text{s.t. } 0 \leq L_{ij} \leq 1, \ 0 \leq R_{ij} \leq 1, \ P = LR^T, \ \sum_j L_{ij} = 1, \ \forall i.$$

Algorithm for BMC

[Recht, Ré, 2013]

$$\begin{aligned} \min_{L,R} \quad & \sum_{(i,j) \in \Omega} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] + \mu \|LR^T\|_* \\ \text{s.t.} \quad & 0 \leq L_{ij} \leq 1, \ 0 \leq R_{ij} \leq 1, \ P = LR^T, \ \sum_j L_{ij} = 1, \ \forall i. \end{aligned}$$

$$\begin{aligned} \min_{L,R} \quad & \sum_{(i,j) \in \Omega} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] + \frac{\mu}{2} (\|L\|_F^2 + \|R\|_F^2) \\ \text{s.t.} \quad & 0 \leq L_{ij} \leq 1, \ 0 \leq R_{ij} \leq 1, \ P = LR^T, \ \sum_j L_{ij} = 1, \ \forall i. \end{aligned}$$

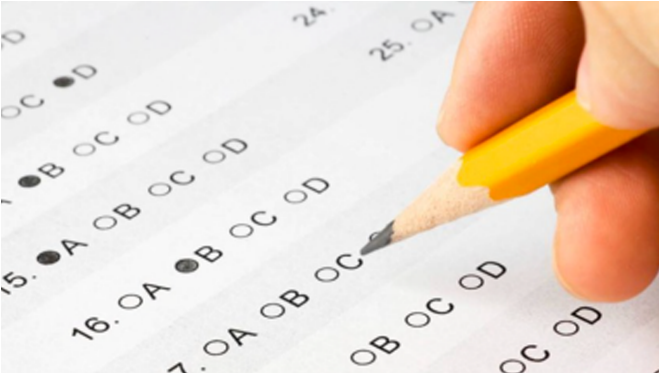
Approximation

$$\|X\|_* = \min_{X=LR^T} \frac{1}{2} (\|L\|_F^2 + \|R\|_F^2)$$

$$\begin{aligned} L_{i_k}^{(k+1)} &= \Pi_{P_L} \left(\left(1 - \frac{\mu_1 \alpha_k}{|\Omega_{i_k \star}|} \right) L_{i_k}^{(t)} - \alpha_k \frac{\phi_c(Y_{i_k j_k} - \phi(L_{i_k} R_{j_k}^T))}{\phi(L_{i_k} R_{j_k}^T)(1 + e^{-\phi_c(L_{i_k} R_{j_k}^T - \phi_b)})} R_{j_k}^{(t)} \right), \\ R_{j_k}^{(k+1)} &= \Pi_{P_R} \left(\left(1 - \frac{\mu_1 \alpha_k}{|\Omega_{\star j_k}|} \right) R_{j_k}^{(t)} - \alpha_k \frac{\phi_c(Y_{i_k j_k} - \phi(L_{i_k} R_{j_k}^T))}{\phi(L_{i_k} R_{j_k}^T)(1 + e^{-\phi_c(L_{i_k} R_{j_k}^T - \phi_b)})} L_{i_k}^{(t)} \right) \end{aligned}$$

Projected SGD

Experiments: Data Set



TOEIC (Test Of English for International Communication)

-A test with 150 multiple-choice questions

-7 parts

- Part 5, Part 6

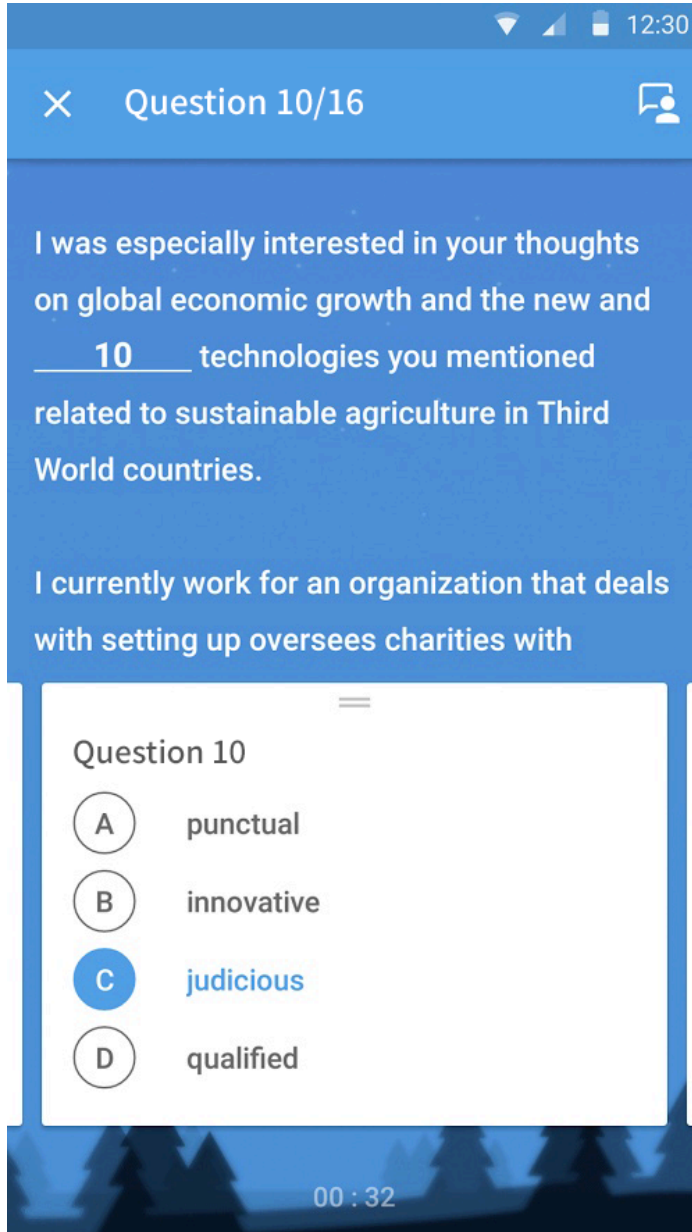
Our office security door is scheduled to _____ this week so all staff members are required to return their security cards to the front desk.

- (A) replace
- (B) replaced
- (C) being replaced
- (D) be replaced

Seasons Greetings. As a _____ customer, we wanted you to be among the first to know about our upcoming holiday sale. All craft paper, specialty printer paper, and decorative envelopes will be reduced by 50% for the month of December.

- (A) value
- (B) valued
- (C) valid
- (D) validate

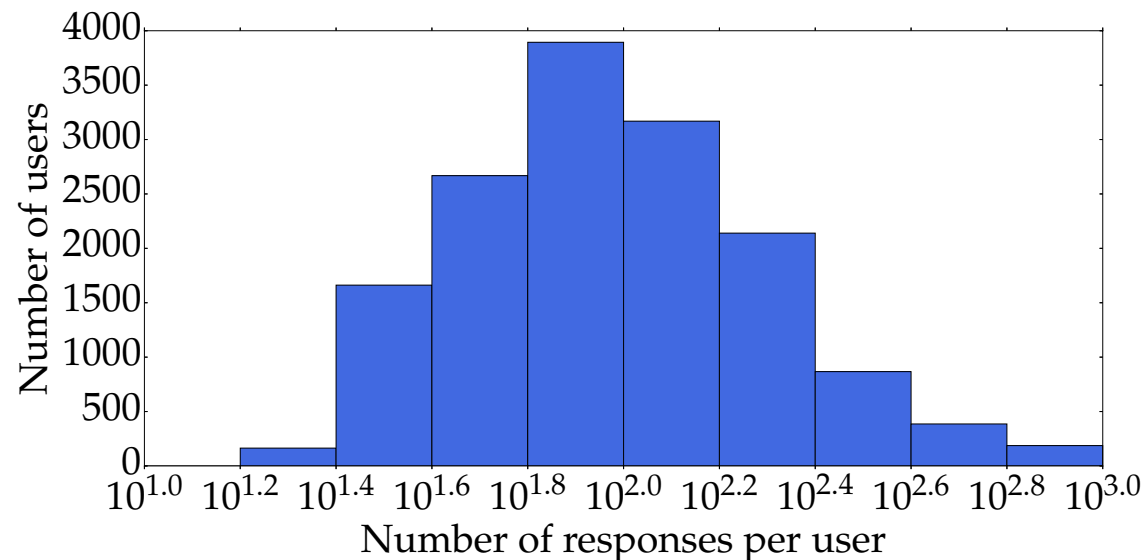
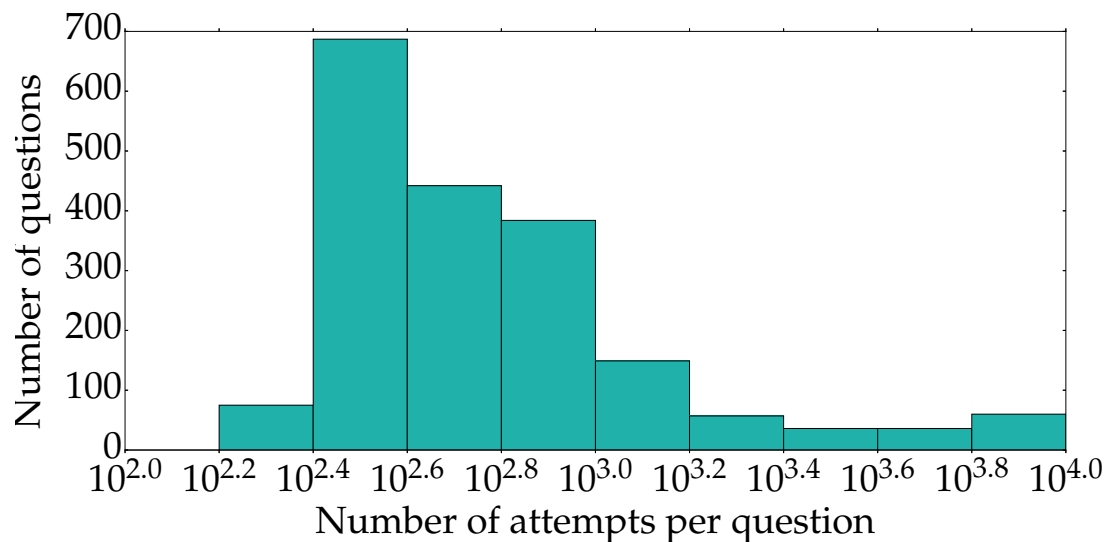
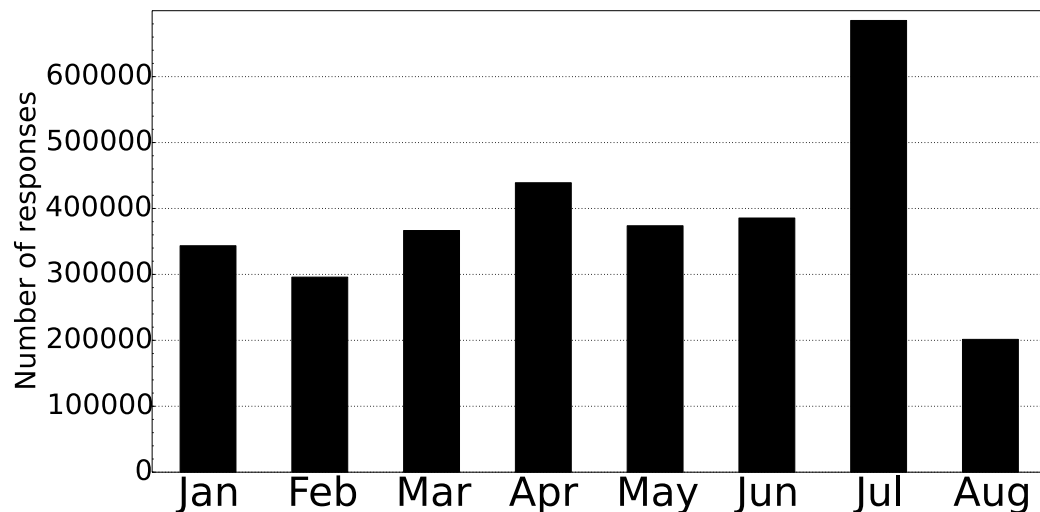
Experiments: Data Set



- Mobile applications (iOS/Android) launched in Korea
- Equipped w/ 4,202 TOEIC questions
- Data was collected from 1/1/2016 to 8/10/2016
- As a result,
 - 106k students signed up, 13m responses collected
 - => On average 130 questions per student
 - Many many outliers
 - Our app became so popular that a lot of people signed up just for checking out
 - Needed to preprocess the data

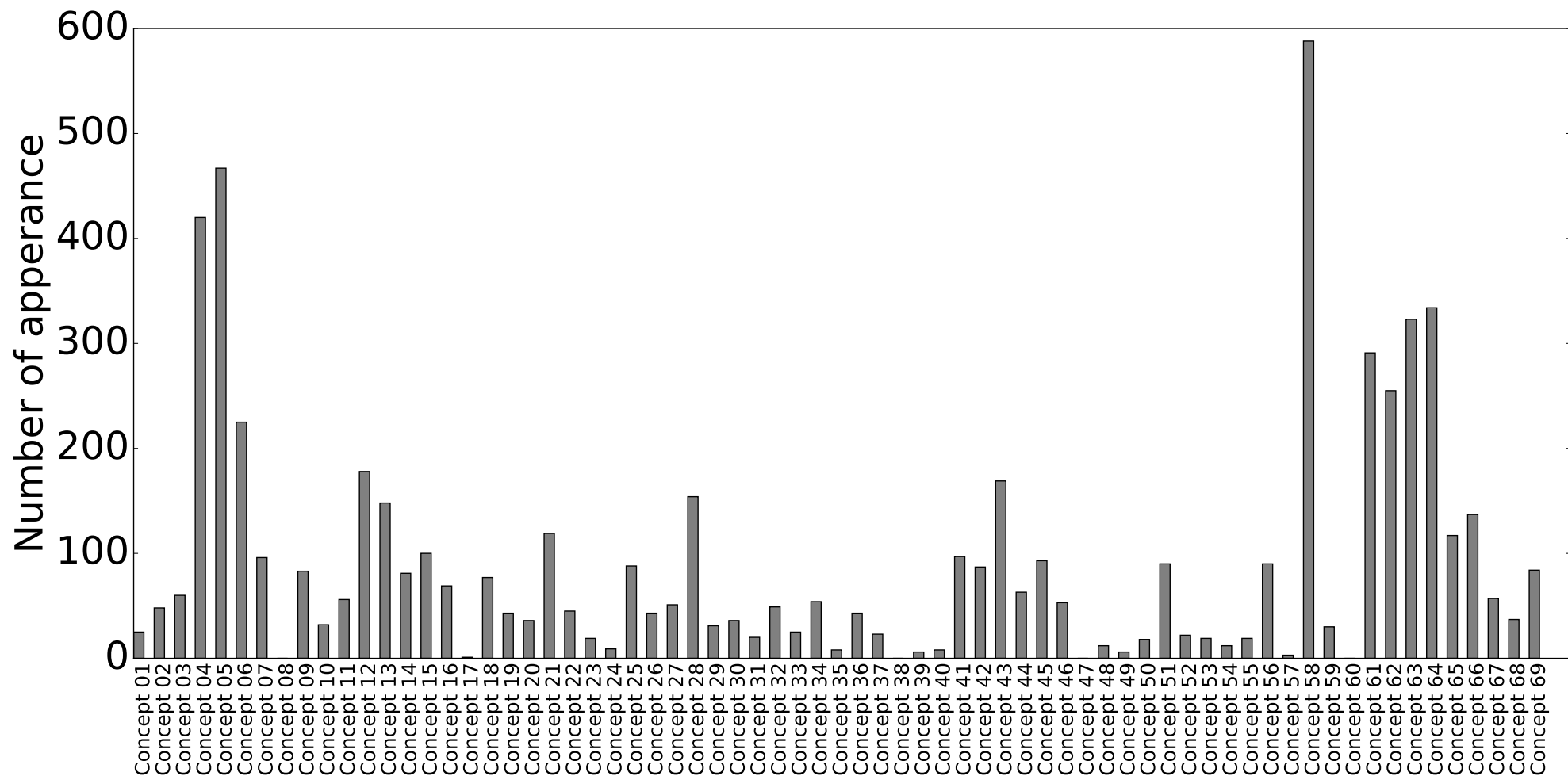
Experiments: Data Set

- Data Filtering
 - > 30 questions per student
 - > 3 seconds per question (on average)
 - > 400 students per question
- After filtering
 - $n \approx 15k$ students
 - $m \approx 2k$ questions
 - # of observed entries
 $\approx 1.9m$ questions (6.5%)

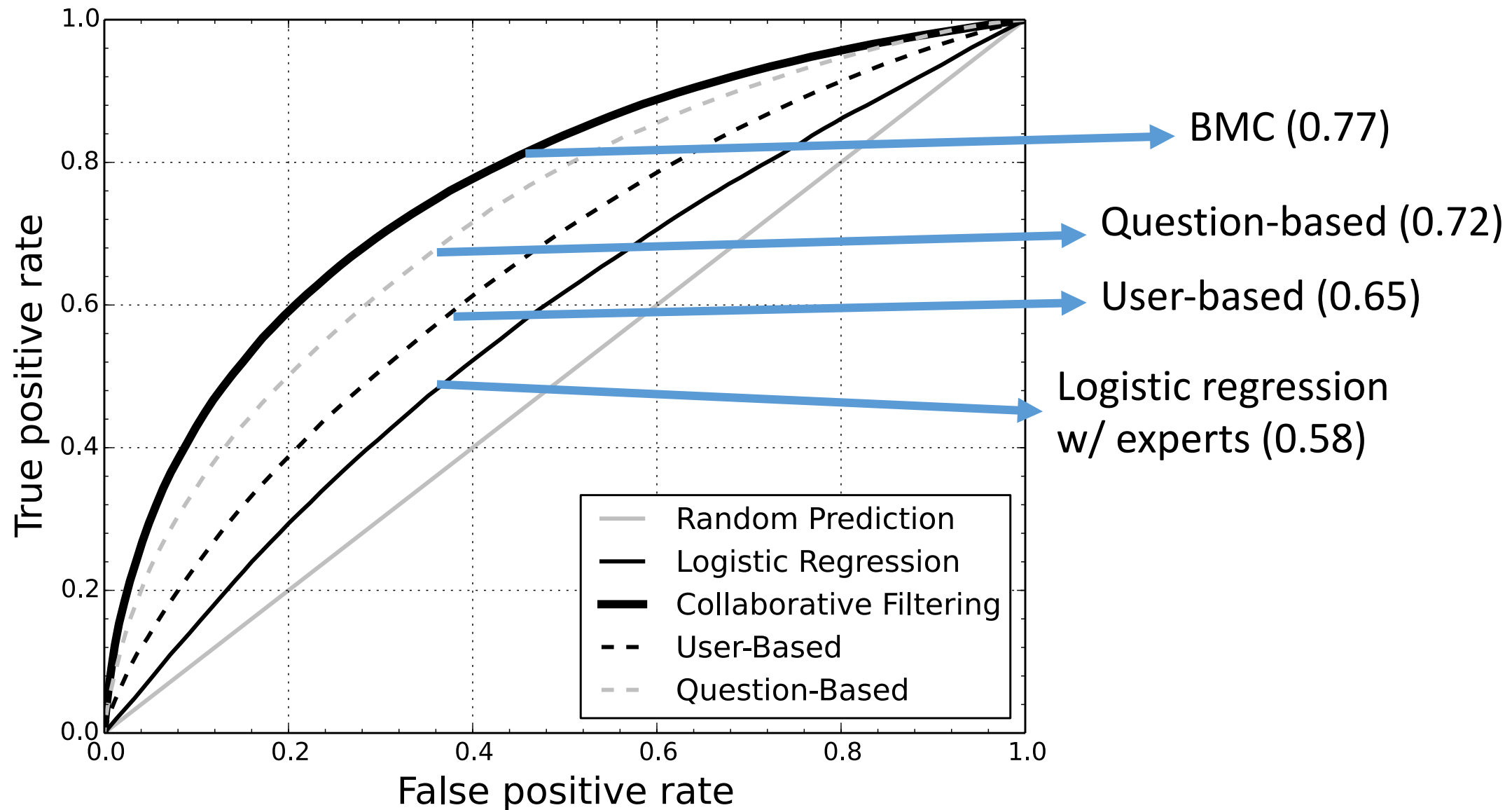


Experiments: Data Set

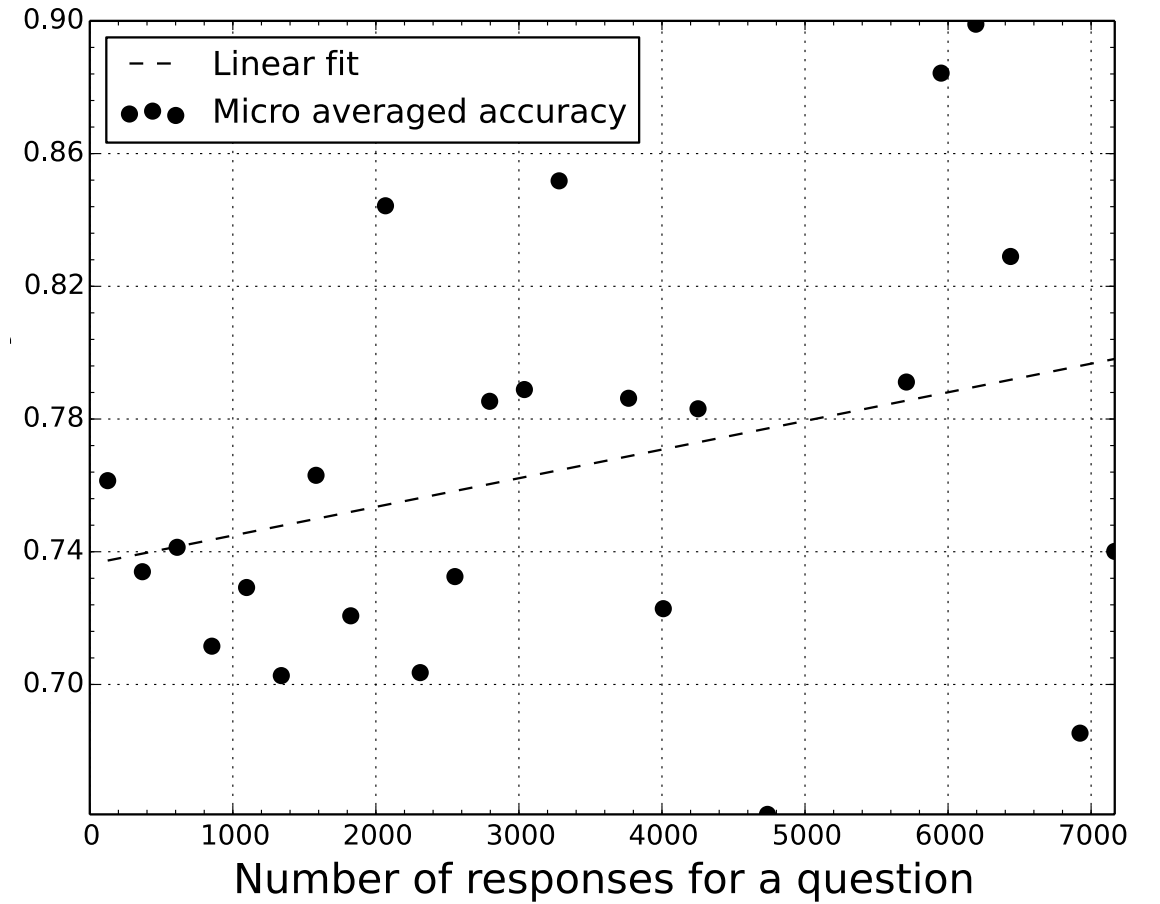
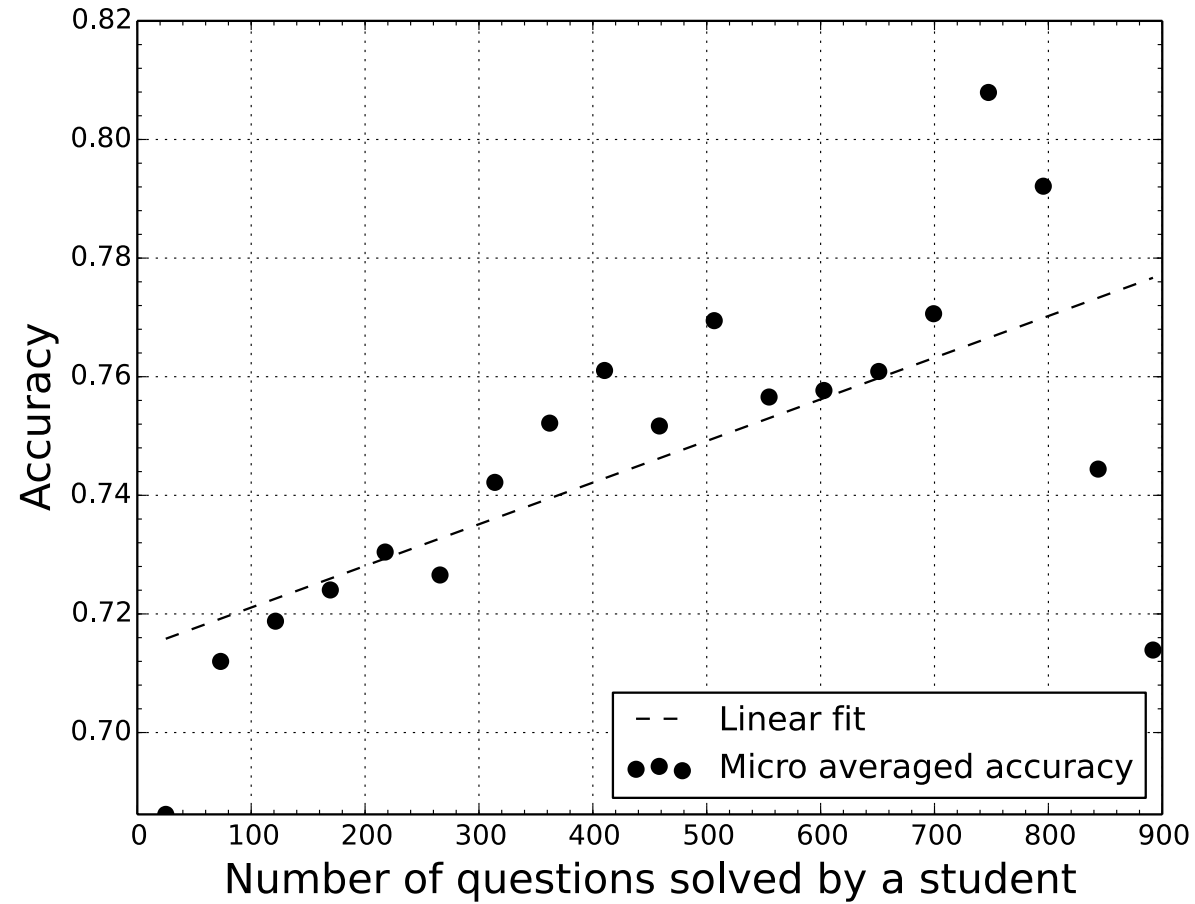
- $m \approx 2k$ questions are manually tagged by experts
- 15 experts first come up with 69 concepts for describing part 4/5 questions
- Each question is randomly assigned to 2 experts among 15 experts



Experiments: Results (AUC)



Experiments: Results (Accuracy)



Prediction API in Products

진단테스트 시작하기

아래의 버튼을 누르면 테스트가 시작됩니다.

테스트는 총 30 문제로 구성되어 있습니다.

신중하게 풀어주세요!



1

진단 테스트 문제 풀이



2

예측 결과 발송



3

예측 테스트 문제 풀이



4

예측 결과 확인

Diagnostic tests

● 진단 테스트 문제 풀이 ● 해독 결과 발송 ● 예측 테스트 문제 풀이 ● 해독 결과 발송

1 / 30

문제 1

Since advertising on the radio is normally not as _____ as advertising on television, many businesses prefer to use local radio stations for their marketing needs.

(A) expense
(B) expensively
(C) expensive
(D) expenses

☐ A
☐ B
☐ C
☐ D

결과보기

kw1jjang@kaist.ac.kr

나의 테스트 결과 **정답 7 / 30 개**머신러닝 정답/오답 예측 성공률 **77%**

퀴즈의 머신러닝 기술은 정답 / 오답 뿐만 아니라,
당신이 선택할 보기까지 예측할 수 있습니다.

머신러닝 보기 예측 성공률 20%



문제 1

문제확인

Comparison

보기 예측 실패

문제정답

내가 선택한 보기

A

내가 선택할 것으로 예측된 보기

D

The screenshot shows a digital interface for a listening test. At the top, there are navigation buttons: '전단 테스트 문제 풀이' (Previous Test Question Solution), '계속 경과 방송' (Continue Broadcast), '계속 테스트 문제 풀이' (Continue Test Question Solution), and '계속 경과 방송' (Continue Broadcast). Below these is a progress indicator '1 / 30'. The main content area displays '문제 1' (Question 1) and the text: 'Since advertising on the radio is normally not as _____ as advertising on television, many businesses prefer to use local radio stations for their marketing needs.' Below the text are four multiple-choice options: (A) expense, (B) expensively, (C) expensive, and (D) expenses. At the bottom left, there are four radio button options: A, B, C, and D. At the bottom right, there is a large '2nd test' watermark and a button labeled '시험 완료 >' (Test Completed >).

2nd test

본 메일은 위이드의 머신러닝 분석기술 제형의 예측 결과를 담고있습니다.
예측 테스트 물이 이후, 예측 결과에 피구상이 든다면 아래 예측결과와 비교해보세요!

물체 id	1st 예측값	2nd 예측값
325884	A	B
324080	D	A
313684	B	D
298106	A	C
301926	D	C
321012	D	B
298106	D	B
314192	D	A
311345	D	A
322940	D	A
301926	B	A
298880	D	C
315206	A	C
323684	A	B
328068	A	B
328068	D	D
298824	B	D
324454	B	C
314760	C	D
322558	D	C
301762	A	C
299226	B	C
301644	B	A
301398	C	B
301908	C	C
314172	C	A
325978	A	D
313696	A	D
327820	A	D
301808	C	A
314802	C	C
	D	D

Email w/ prediction

Conclusion & Discussion

- ML framework for response prediction
 - Based on a variation of M2PL
 - Two algorithms:
 - Logistic regression with manually tagged questions
 - Binary matrix completion
- A large-scale experiment
 - Collected 13m responses from 106k students
 - A filtered data set is used for this work
 - Experimental results show that BMC works the best
- Deployed in products (email me if you want to try it yourself 😊)
- Many open problems & new directions
 - Interpretation of hidden concepts for an efficient design of edu. resources
 - Prediction of choices
 - Time-varying L, Sparse R
 - Convergence, Sample complexity, Biased sampling