Can AI/DM help MOOCs?

Jie Tang
Computer Science
Tsinghua University

The slides can be downloaded at http://keg.cs.tsinghua.edu.cn/jietang

Big Data in MOOC

- 149 partners
- 2000+ courses
- 24,000,000 users

xuetangX.com 学堂在线

- 1,000+ courses
- 8,000,000 users
- **Chinese EDU association**

edX

- 110 partners
- 1,270 courses
- 10,000,000 users
- 10+ MicroMaster

MOOC

網易公开课

- host >1,000 courses
- millions of users

UDACITY

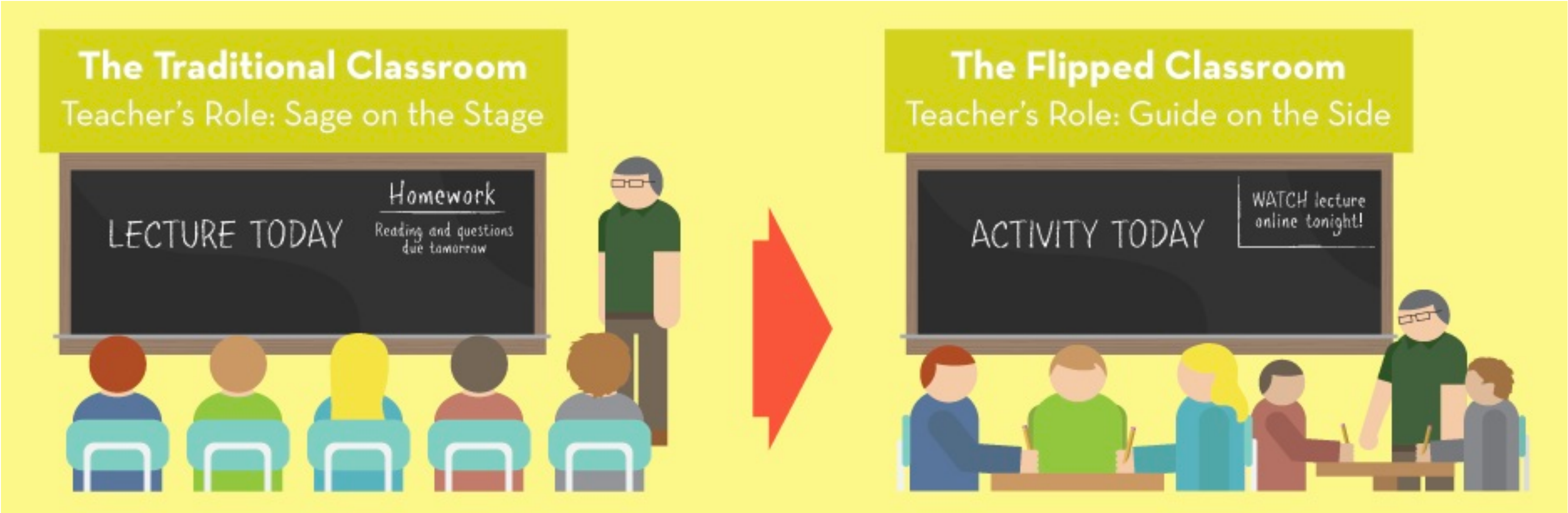
- ~10 partners
- 40+ courses
- **1.6 million** users
- **"nanodegree"**

……

launched in 2013

Some exciting data…

- Every day, there are 5,000+ new students
- An MOOC course can reach 100,000+ students
- >35% of the XuetangX users are using mobile
- traditional->flipped classroom->online degree



Some exciting data…

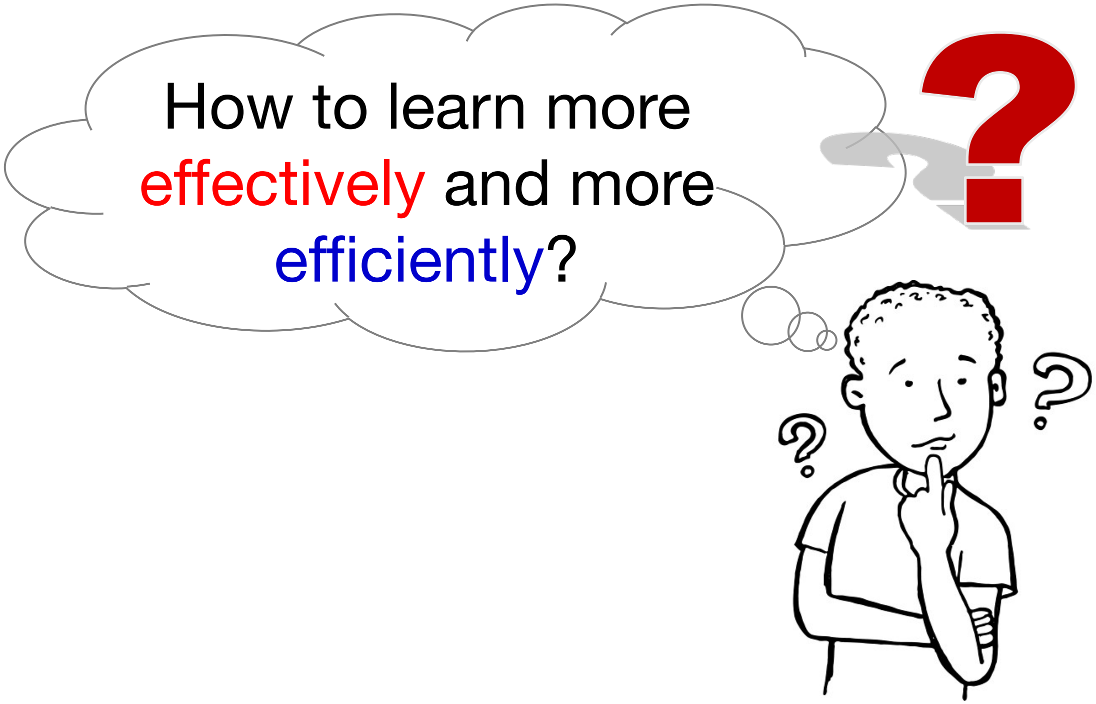
- Every day, there are 5,000+ new students
- An MOOC course can reach 100,000+ students
- >35% of the XuetangX users are using mobile
- traditional->flipped classroom->online degree
- “**Network+** EDU” (O2O)
  - edX launched 10+ MicroMaster degrees
  - Udacity launched NanoDegree program
  - GIT+Udacity launched the largest online master
  - **Tsinghua+XuetangX** will launch a MicroMaster soon

However,

- **only ~3% certificate rate**
  - The highest certificate rate is **14.95%**
  - The lowest is only **0.84%**

- Can **AI** help MOOC and how?

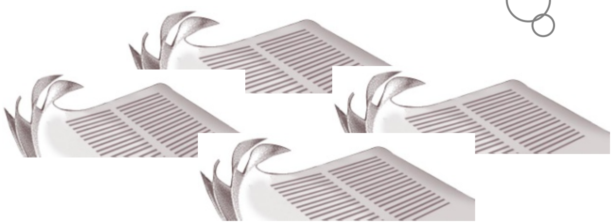
MOOC user = Student?



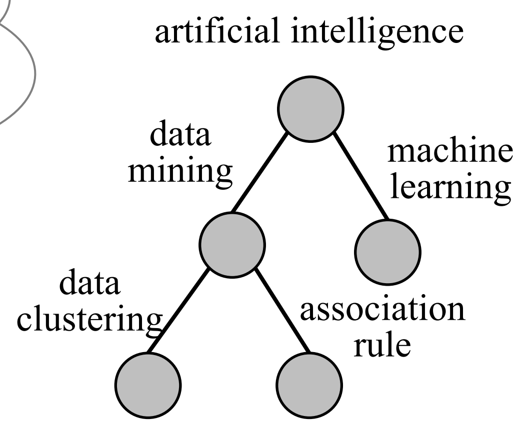
- **Who is who? background, where from?**
- **Why MOOC? motivation? degree?**
- **What is personalization? preference?**

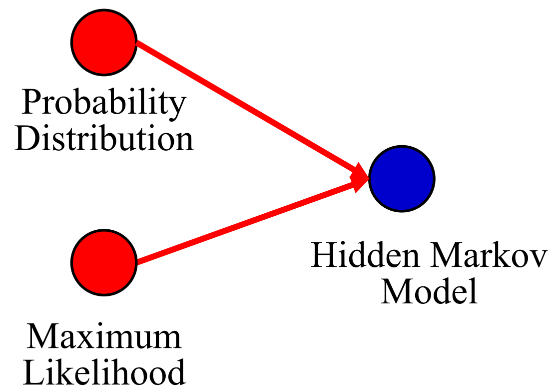
MOOC course = University course?

How to discover the prerequisite relations between concepts and generate the concept graph automatically?



Thousands of Courses





WIKIPEDIA
The Free Encyclopedia

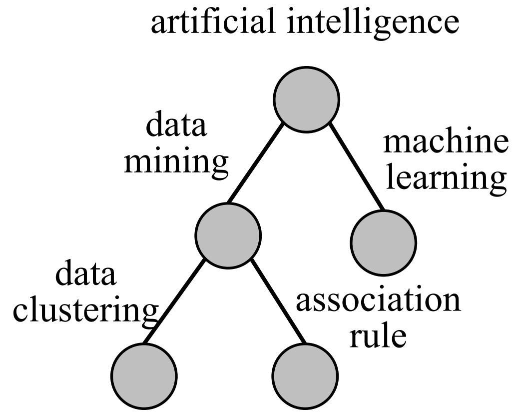
Freebase

yago
select knowledge

How to leverage the external knowledge?

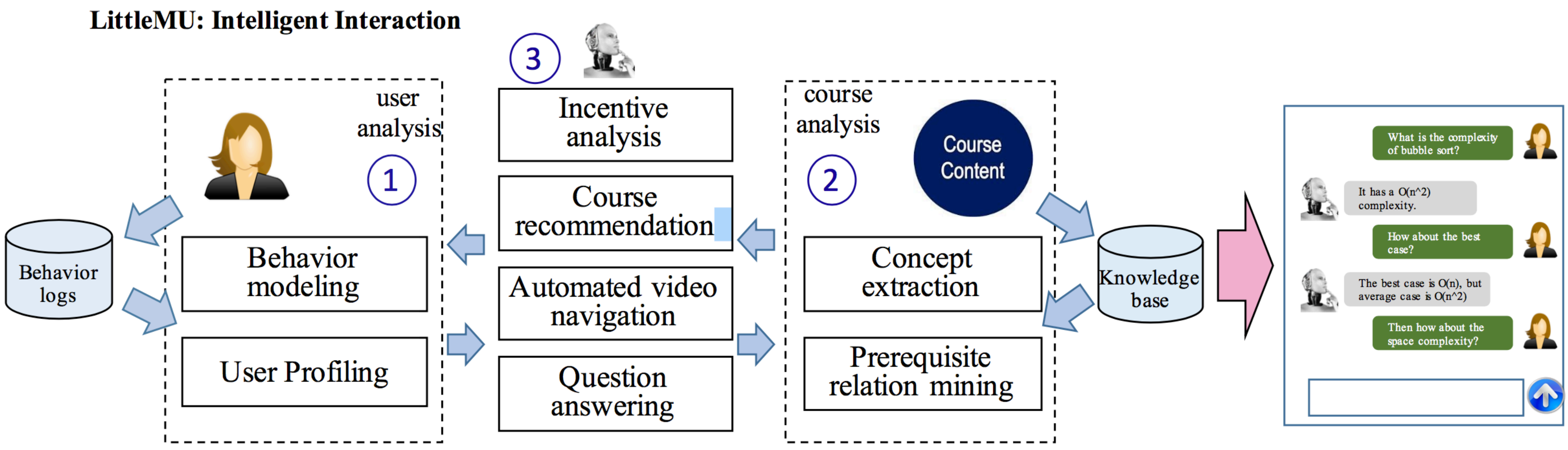
However to improve the engagement?

User



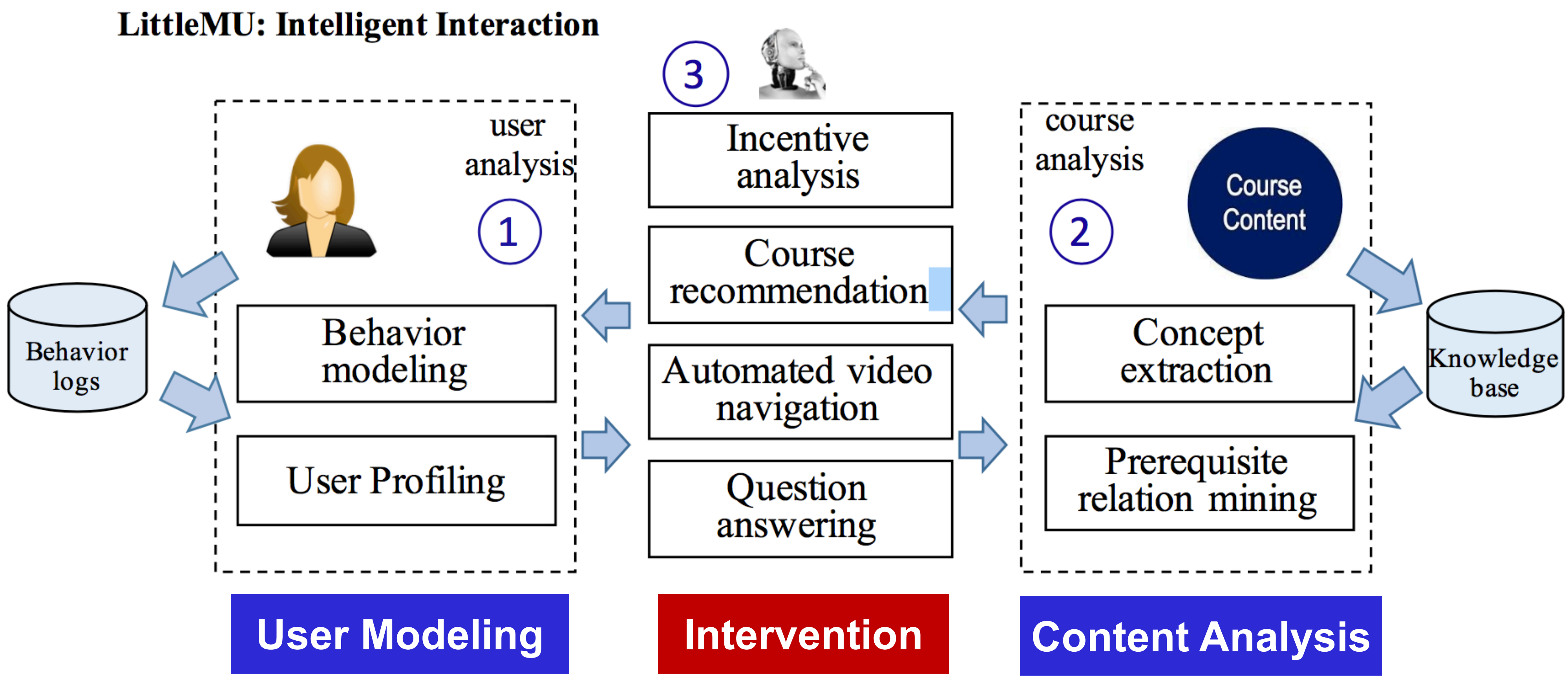
Knowledge

LittleMU (小木)



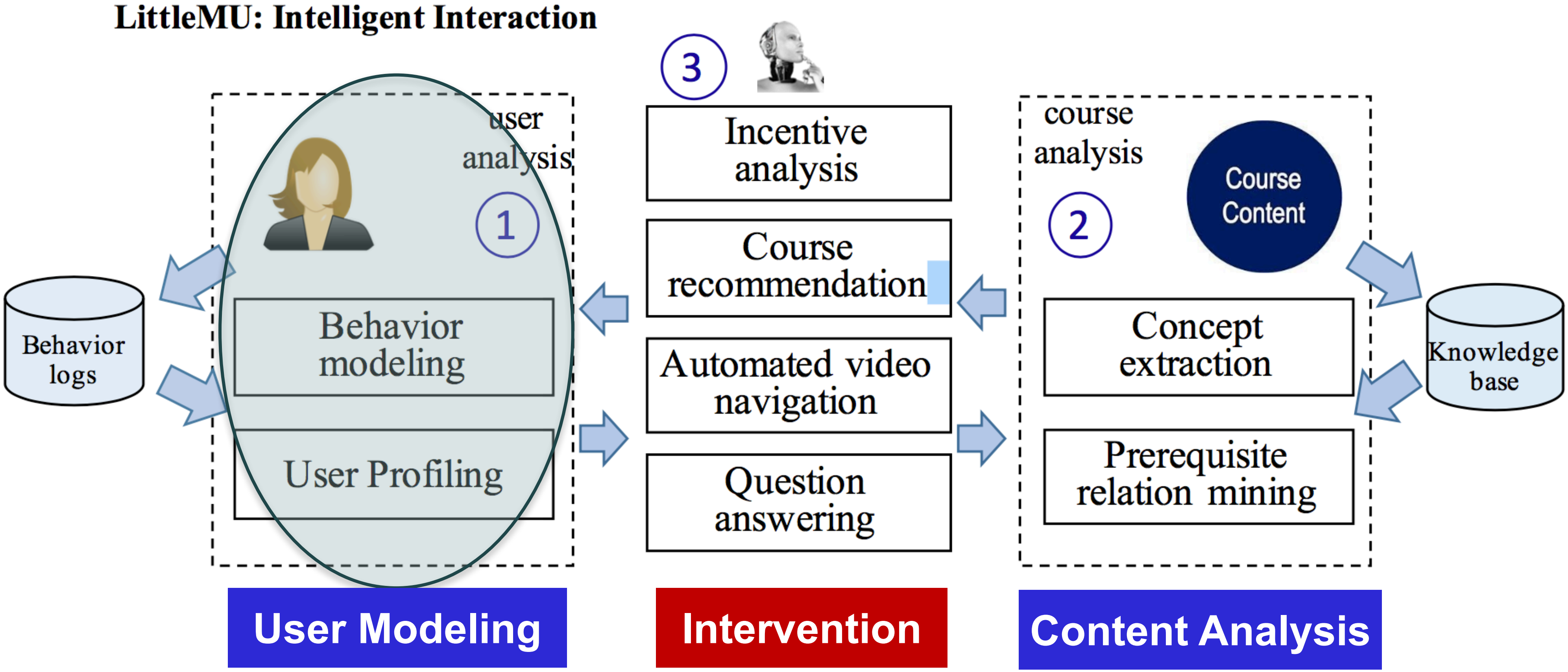
LittleMU: Intelligent Interaction
user analysis
1
Behavior modeling
User Profiling
Behavior logs
3
Incentive analysis
Course recommendation
Automated video navigation
Question answering
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
What is the complexity of bubble sort?
It has a O(n^2) complexity.
How about the best case?
The best case is O(n), but average case is O(n^2)
Then how about the space complexity?

LittleMU (小木)



LittleMU: Intelligent Interaction
3
Incentive analysis
Course recommendation
Automated video navigation
Question answering
user analysis
1
Behavior logs
Behavior modeling
User Profiling
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
User Modeling
Intervention
Content Analysis

LittleMU (小木)

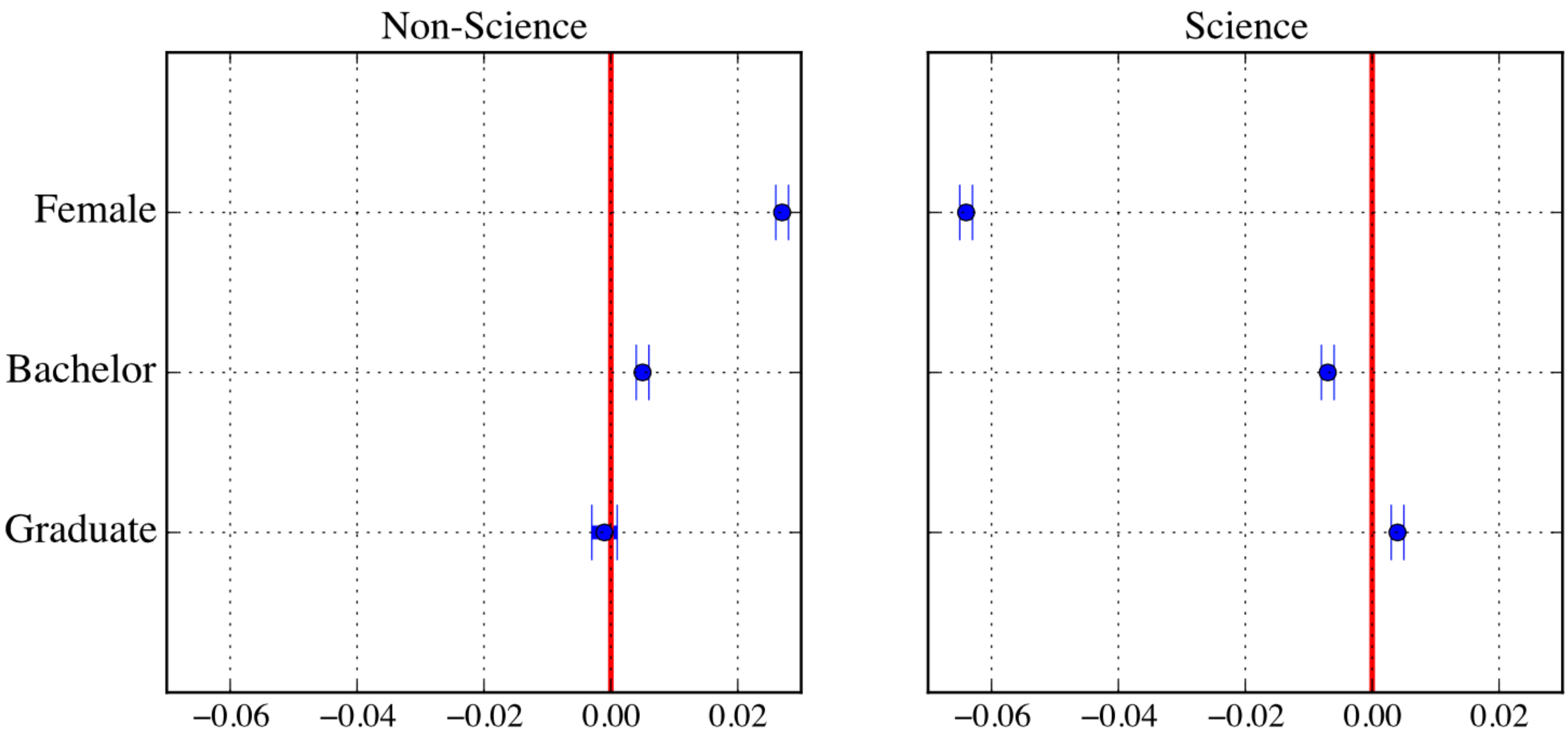


LittleMU: Intelligent Interaction
3
user analysis
1
Behavior modeling
User Profiling
Behavior logs
Incentive analysis
Course recommendation
Automated video navigation
Question answering
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
User Modeling
Intervention
Content Analysis

MOOC user

- Who is who? background, where from?
- Why MOOC? motivation? degree?
- What is personalization? preference?

Basic Analysis



Observation 1 – Gender Difference

Table 4: Regression Analysis for Certificate Rate: All Users

	Model 1		Model 2	
	Non-Science (1)	Science (2)	Non-Science (3)	Science (4)
Female	0.014***	-0.003	0.002*	0.001
	(0.002)	(0.002)	(0.001)	(0.002)
New Post	—	—	0.004***	0.038***
			(0.001)	(0.008)
Reply	—	—	0.004**	0.001*
			(0.002)	(0.001)
Video	—	—	0.000***	-0.000
			(0.000)	(0.000)
Assignment	—	—	0.003***	0.000***
			(0.000)	(0.000)
Bachelor	0.014***	0.003*	0.011***	-0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Graduate	0.007***	0.004	0.013***	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Effort	-0.072***		-0.072***	
	(0.003)		(0.003)	
Constant	0.286***	0.018***	0.280***	0.006
	(0.013)	(0.006)	(0.011)	(0.004)
Obs.	74,480	19,269	74,480	19,269
R^2	0.024	0.001	0.462	0.363

Model 1: Demographics vs Certificate

Model 2: Demographics + Learning activities vs Certificate

- Females are significantly more likely to get the certificate in non-science courses.
- The size of the gender difference decreases significantly after we control for forum learning activities.

Observation 2 – Ability v.s. Effort

Table 4: Regression Analysis for Certificate Rate: All Users

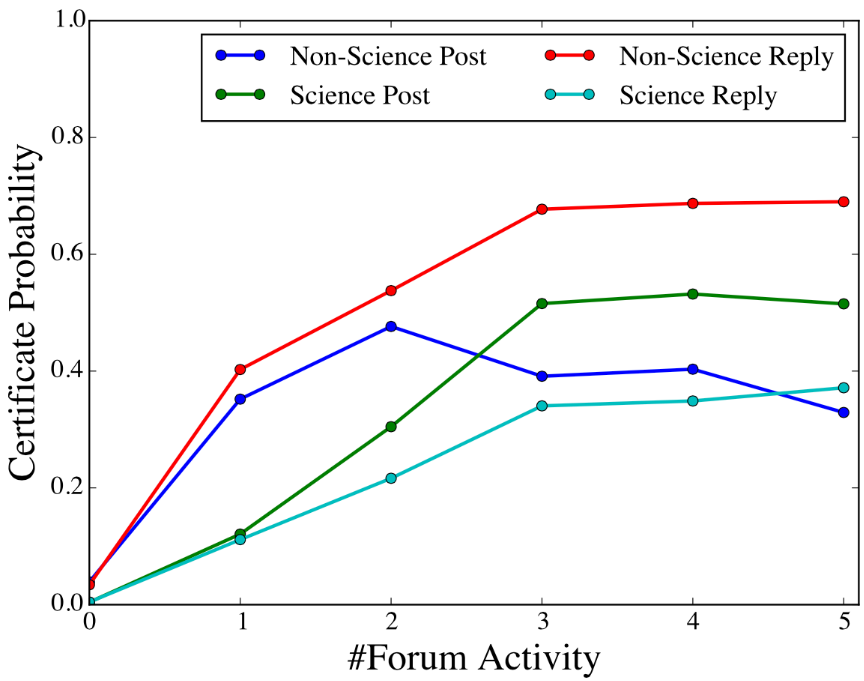
	Model 1		Model 2	
	Non-Science (1)	Science (2)	Non-Science (3)	Science (4)
Female	0.014***	-0.003	0.002*	0.001
	(0.002)	(0.002)	(0.001)	(0.002)
New Post	—	—	0.004***	0.038***
			(0.001)	(0.008)
Reply	—	—	0.004**	0.001*
			(0.002)	(0.001)
Video	—	—	0.000***	-0.000
			(0.000)	(0.000)
Assignment	—	—	0.003***	0.000***
			(0.000)	(0.000)
Bachelor	0.014***	0.003*	0.011***	-0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Graduate	0.007***	0.004	0.013***	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Effort	-0.072***		-0.072***	
	(0.003)		(0.003)	
Constant	0.286***	0.018***	0.280***	0.006
	(0.013)	(0.006)	(0.011)	(0.004)
Obs.	74,480	19,269	74,480	19,269
R^2	0.024	0.001	0.462	0.363

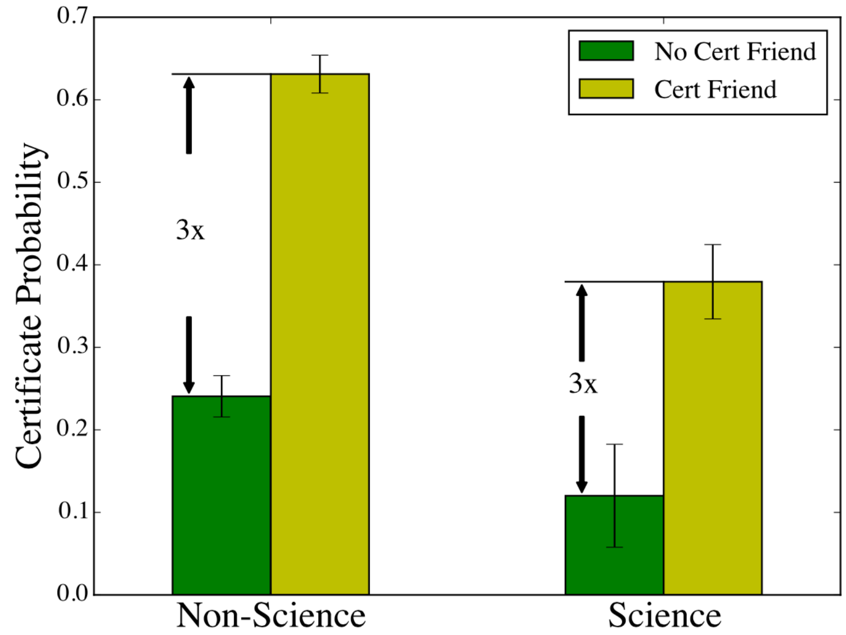
Model 1: Demographics vs Certificate

Model 2: Demographics + Learning activities vs Certificate

- Bachelors students are significantly more likely to get the certificate in non-science courses.
- Graduate students are more likely to get the certificate in science courses. After controlling for learning activities, the size of the effect is almost doubled.
- Learning activities are good predictors for getting certificates.

Forum activity vs. Certificate





Forum activity vs. Certificate

— It is more important to be present in forum, while the intensity matters less.

“近朱者赤” (Homophily)

– Certificate Probability tripled when one is aware that she has certificate friend(s)

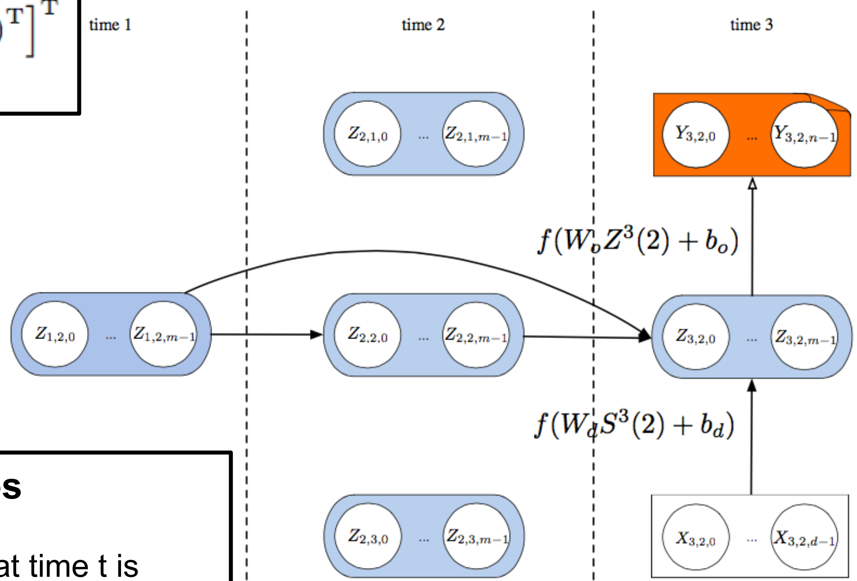
Dynamic Factor Graph Model

Model: incorporating deep learning and factor graphs

$$Y^t(i)^* = f(W_o Z^t(i) + b_o)$$

$$Z^t(i)^* = f(W_d S^t(i) + b_d)$$

$$S^t(i) = \left[\mathbf{Z}_{t-p}^{t-1}(i)^{\mathrm{T}}, X^t(i)^{\mathrm{T}}\right]^{\mathrm{T}}$$



Prediction labels:
Activities we are interested in, e.g., assignments performance and getting certificates.

$$Y^t(i) = [Y_{t,i,0}, Y_{t,i,1}, \ldots, Y_{t,i,n-1}]^{\mathrm{T}}$$

Latent learning states

Every student's status in at time t is associated with a vector representation

$$Z^t(i) = [Z_{t,i,0}, Z_{t,i,1}, \ldots, Z_{t,i,m-1}]^{\mathrm{T}}$$

All features: time-varying attributes:
1.Demographics
2.Forum Activities
3. Learning Behaviors

$$X^t(i) = [X_{t,i,0}, X_{t,i,1}, \ldots, X_{t,i,d-1}]^{\mathrm{T}}$$

[1] Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. Modeling and Predicting Learning Behavior in MOOCs. **WSDM'16**, pages 93-102.

Certificate Prediction

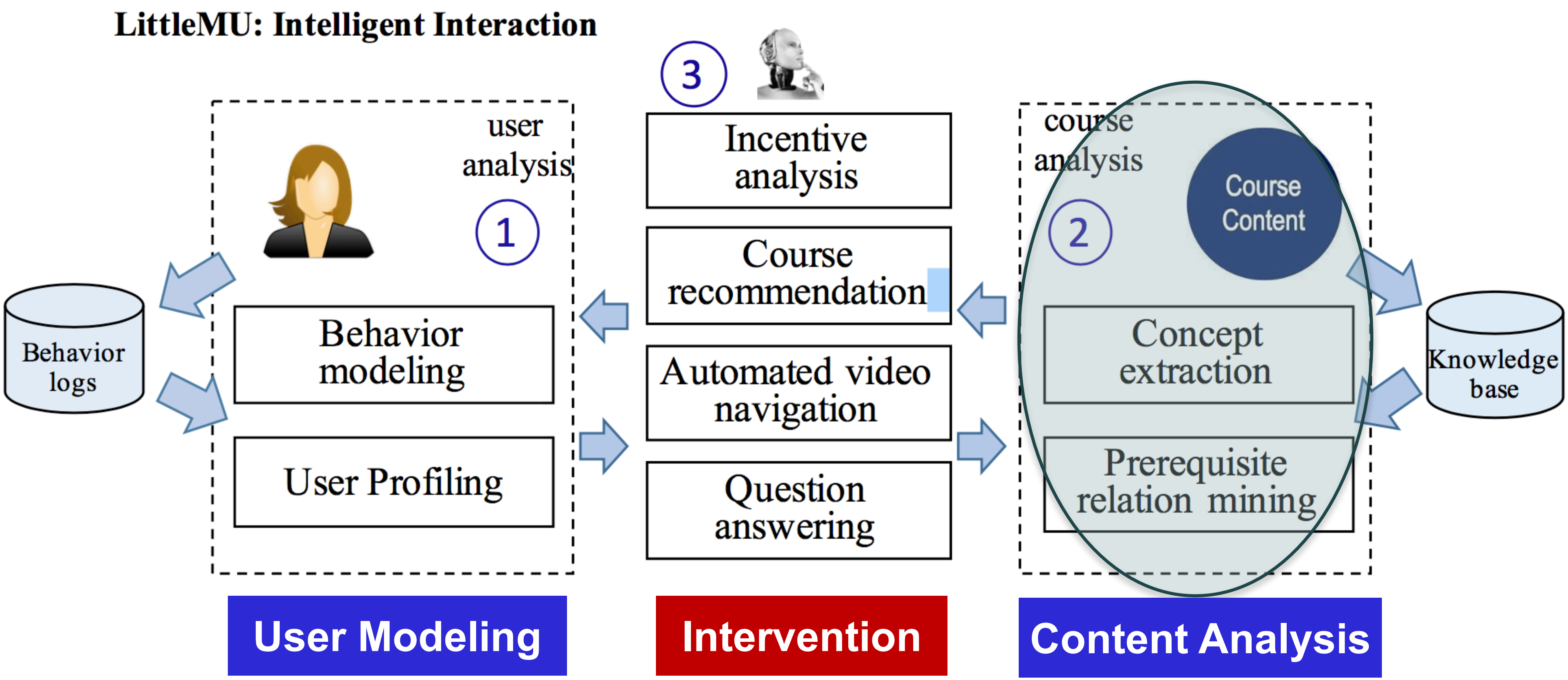
Category	Method	AUC	Precision	Recall	F1-score
Science	LRC	92.13	**83.33**	46.51	59.70
	SVM	92.67	52.17	83.72	64.29
	FM	94.48	61.54	74.42	67.37
	LadFG	**95.73**	73.91	**79.07**	**76.40**
Non-Science	LRC	94.16	76.93	89.20	82.57
	SVM	93.94	76.96	88.60	82.37
	FM	94.87	**80.22**	86.23	83.07
	LadFG	**95.54**	79.76	**89.01**	**84.10**

- LRC, SVM, and FM are different baseline models
- LadFG is our proposed model

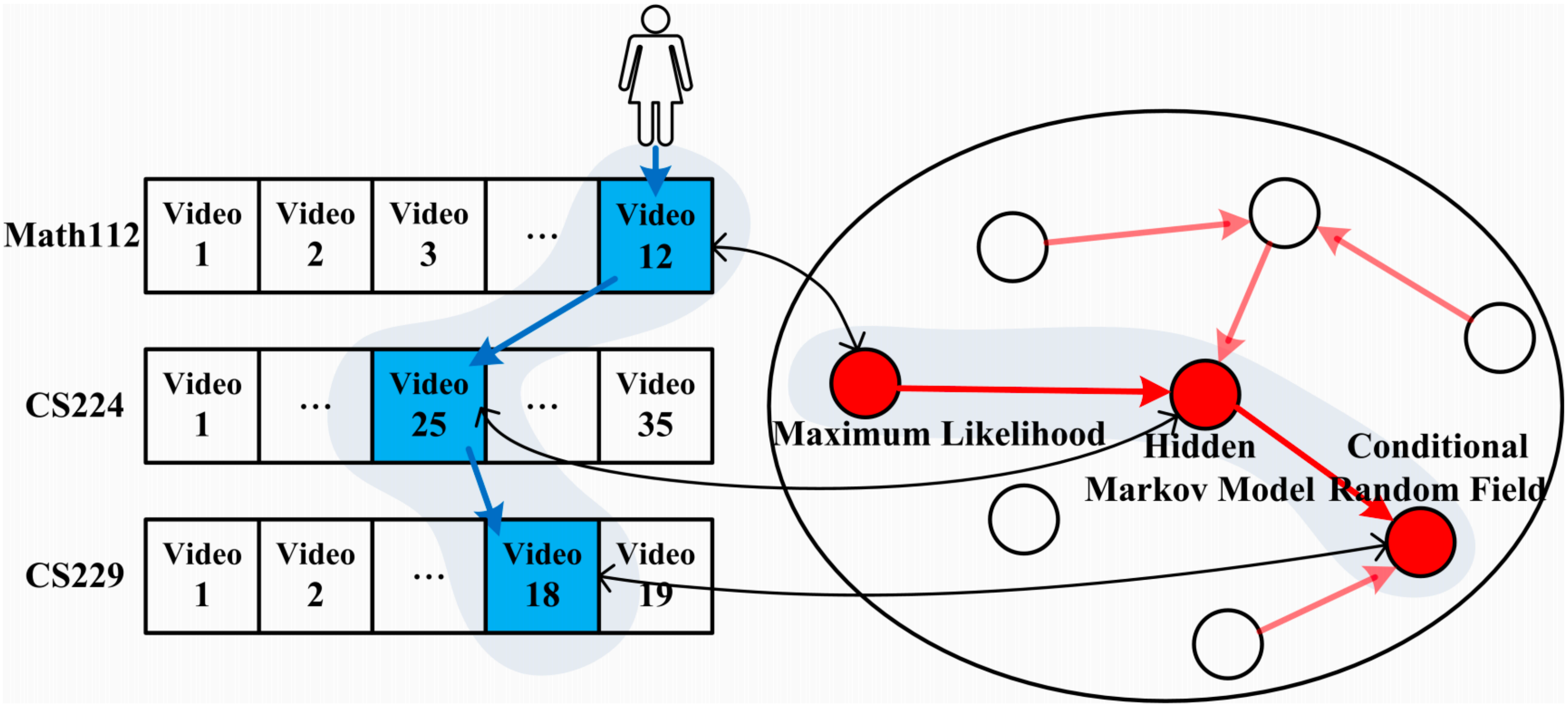
Predicting more

- Dropout
  - KDDCUP 2015, 1,000+ teams worldwide
- Demographics
  - Gender, education, etc.
- User interest
  - computer science, mathematics, psychology, etc.
- …

LittleMU (小木)


LittleMU: Intelligent Interaction
user analysis
1
Behavior modeling
User Profiling
Behavior logs
3
Incentive analysis
Course recommendation
Automated video navigation
Question answering
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
User Modeling
Intervention
Content Analysis

Knowledge Graph



- How to extract concepts from course scripts?
- How to recognize (prerequisite) relationships between concepts?

Concept Extraction

Candidate Concept Extraction → Semantic Representation Learning → Graph-based Ranking

In this course, we will teach some basic knowledge about data mining and its application in business intelligence.

Video script

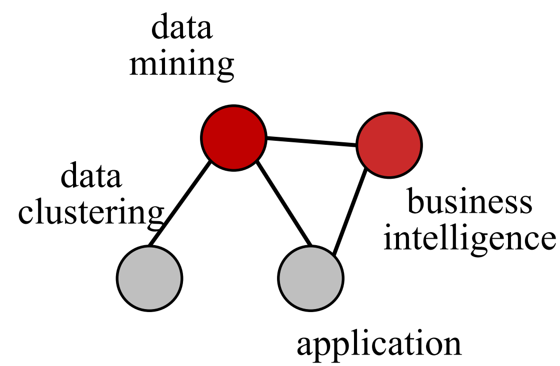
data mining

0.8	0.2	0.3	…	0.0	0.0

business intelligence

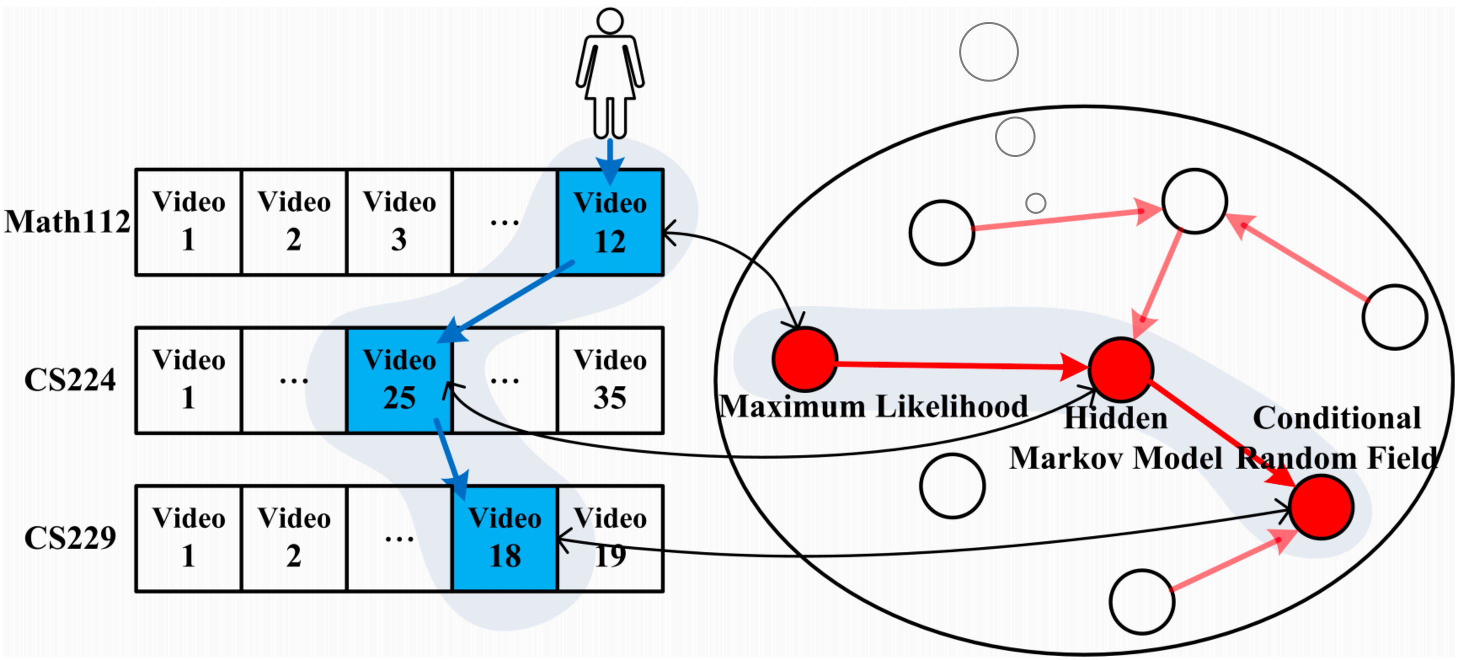
0.1	0.1	0.2	…	0.8	0.7

Vector representation
Learned via embedding or deep learning



Prerequisite Relationship

How to extract the prerequisite relationship?



[1] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. **ACL'17.**

Prerequisite Relationship Extraction

- Step 1： First extract important concepts
- Step 2： Use Word2Vec to learn representations of concepts

data mining

0.8	0.2	0.3	…	0.0	0.0

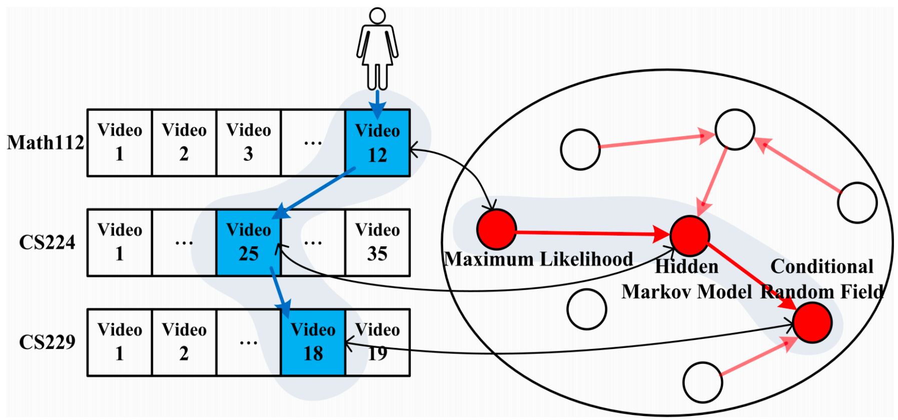
business intelligence

0.1	0.1	0.2	…	0.8	0.7

Vector representation
Learned via embedding or deep learning

Prerequisite Relationship Extraction

- Step 1： First extract important concepts
- Step 2： Use Word2Vec to learn representations of concepts
- Step 3： Distance functions
  - Semantic Relatedness
  - Video Reference Distance
  - Sentence Reference Distance
  - Wikipedia Reference Distance
  - Average Position Distance
  - Distributional Asymmetry Distance
  - Complexity Level Distance



Result of Prerequisite Relationship

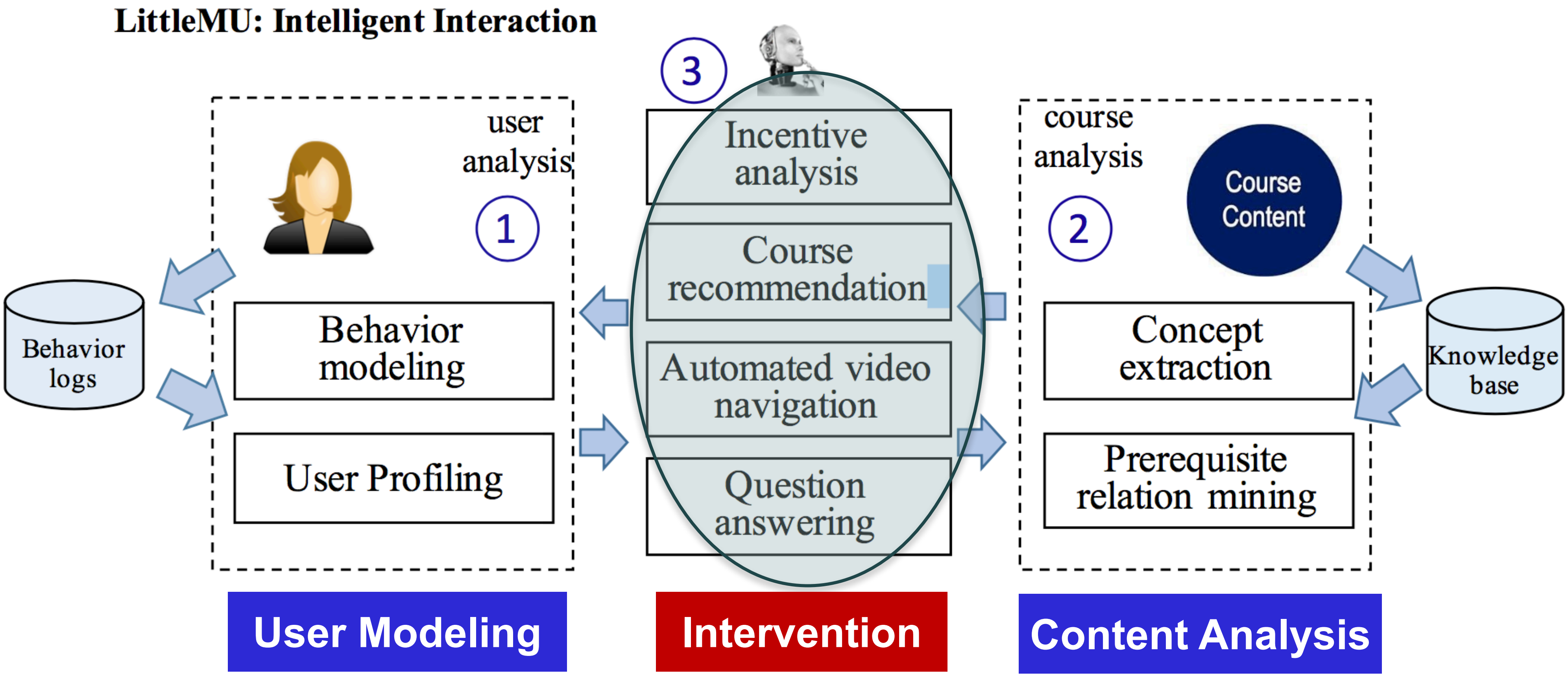
Classifier	M	ML 1	ML 10	DSA 1	DSA 10	CAL 1	CAL 10
SVM	P	63.2	60.1	60.7	62.3	61.1	61.9
	R	68.5	72.4	**69.3**	67.5	**67.9**	68.3
	F_1	65.8	65.7	64.7	64.8	64.3	64.9
NB	P	58.0	58.2	62.9	62.6	60.1	60.6
	R	58.1	60.5	62.3	61.8	61.2	62.1
	F_1	58.1	59.4	62.6	62.2	60.6	61.3
LR	P	66.8	67.6	63.1	62.0	62.7	63.3
	R	60.8	61.0	64.8	66.8	63.6	64.1
	F_1	63.7	64.2	63.9	64.3	61.6	62.9
RF	P	**68.1**	**71.4**	**69.1**	**72.7**	**67.3**	**70.3**
	R	**70.0**	**73.8**	68.4	**72.3**	67.8	**71.9**
	F_1	**69.1**	**72.6**	**68.7**	**72.5**	**67.5**	**71.1**

Table 2: Classification results of the proposed method(%).

- SVM, NB, LR, and RF are different classification models
- It seems that with the defined distance functions, RF achieves the best

[1] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. **ACL'17.**

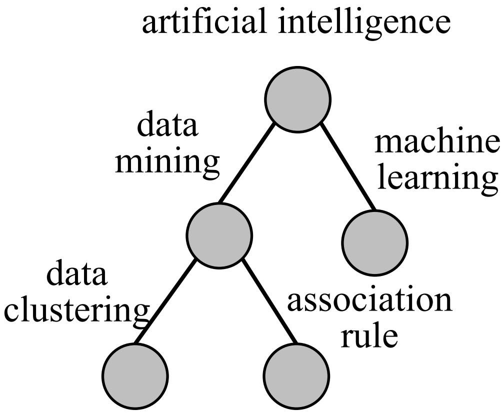
LittleMU (小木)



LittleMU: Intelligent Interaction
user analysis
1
Behavior logs
Behavior modeling
User Profiling
3
Incentive analysis
Course recommendation
Automated video navigation
Question answering
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
User Modeling
Intervention
Content Analysis

What we can do?

User modeling



Knowledge

- Let start with a simple case
  - Course recommendation based on user interest

Course Recommendation

Course topic analysis

Low frequency
- LDA training
- User clustering
- Course perquisite modeling

High frequency
- Latent interest modeling
- Collaborative filtering

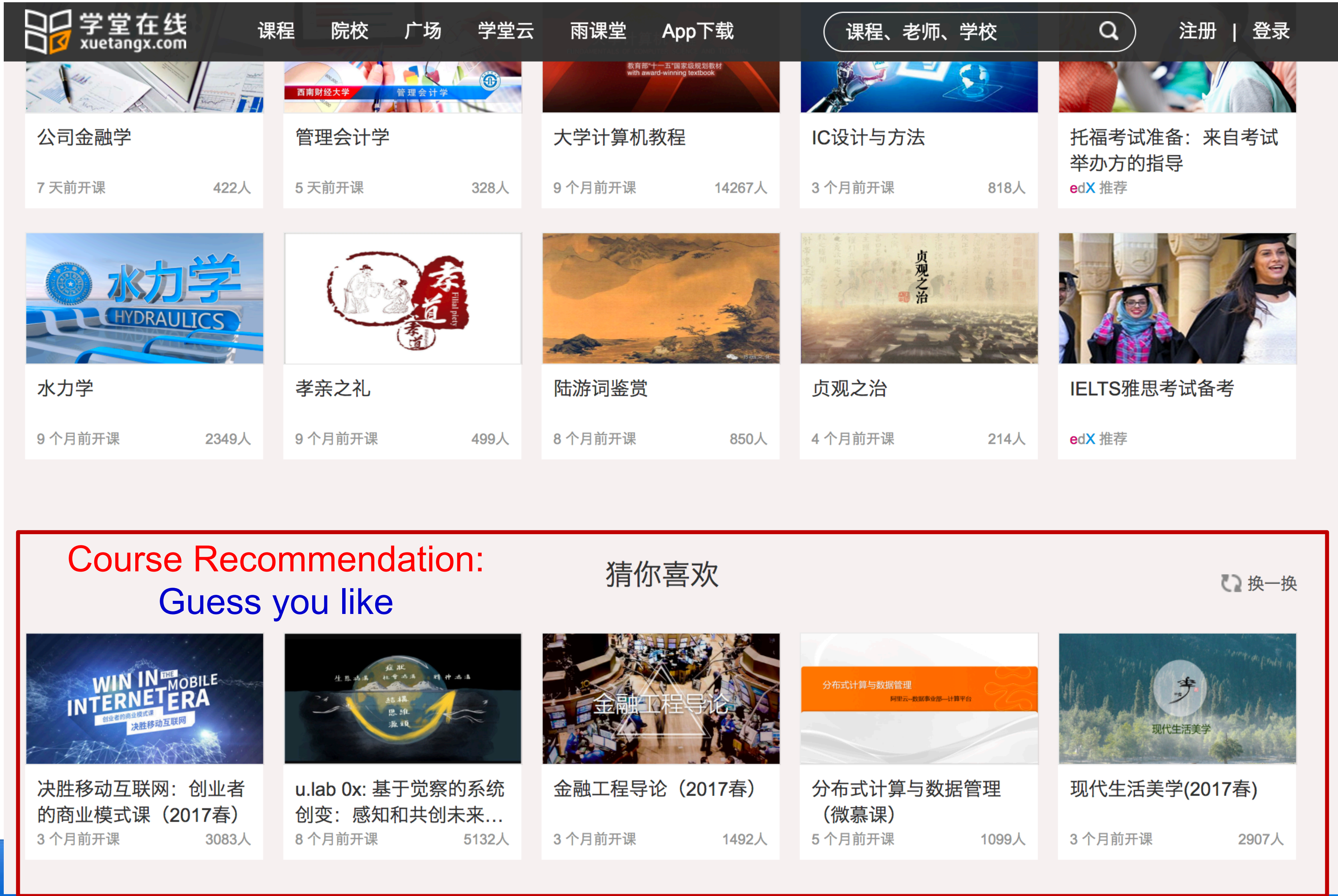
With the learned user model

Recommendation result

Rule based adjustment

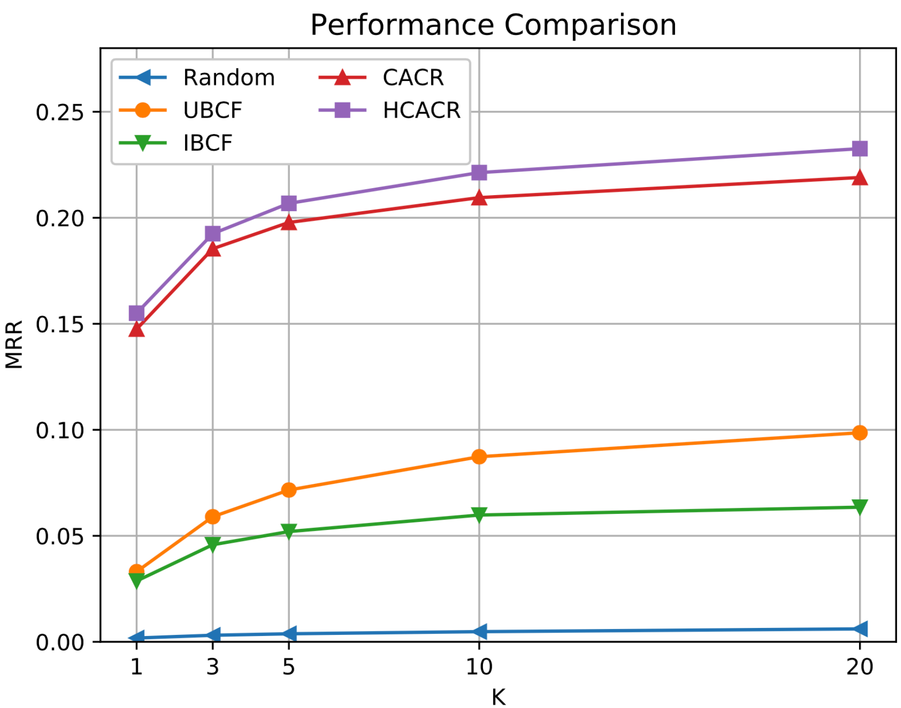
[1] Xia Jing, Jie Tang, Wenguang Chen, Maosong Sun, and Zhengyang Song. Guess You Like: Course Recommendation in MOOCs. **WI'17.**

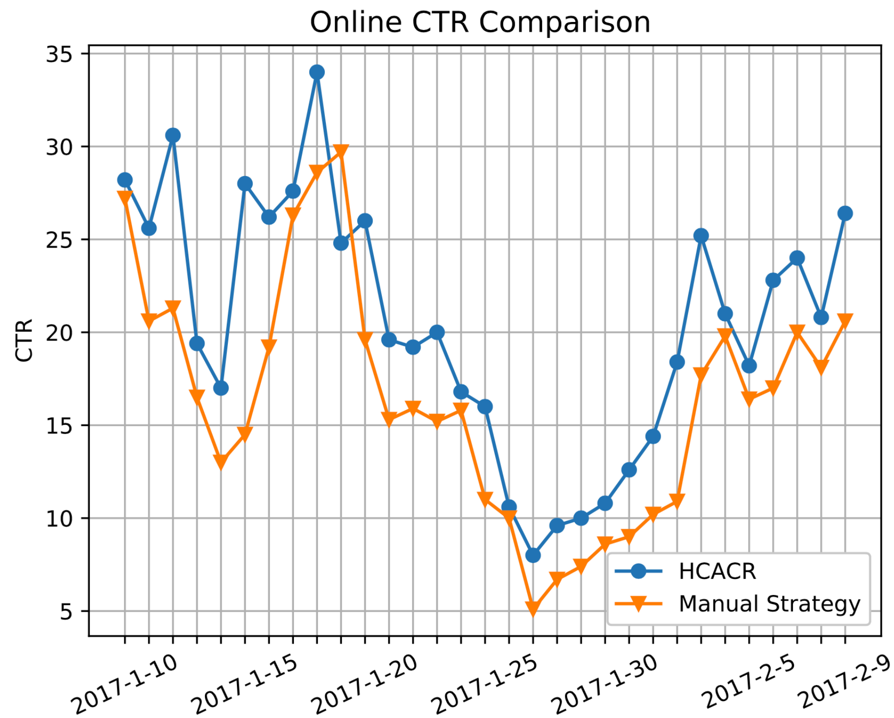
Course Recommendation



Course Recommendation:
Guess you like

Online A/B Test





Top-k recommendation accuracy (MRR)

Comparison methods:

HCACR – Hybrid Content-Aware Course Recommendation

CACR – Content-Aware Course Recommendation

IBCF – Item-Based Collaborative Filtering

UBCF – User-Based Collaborative Filtering

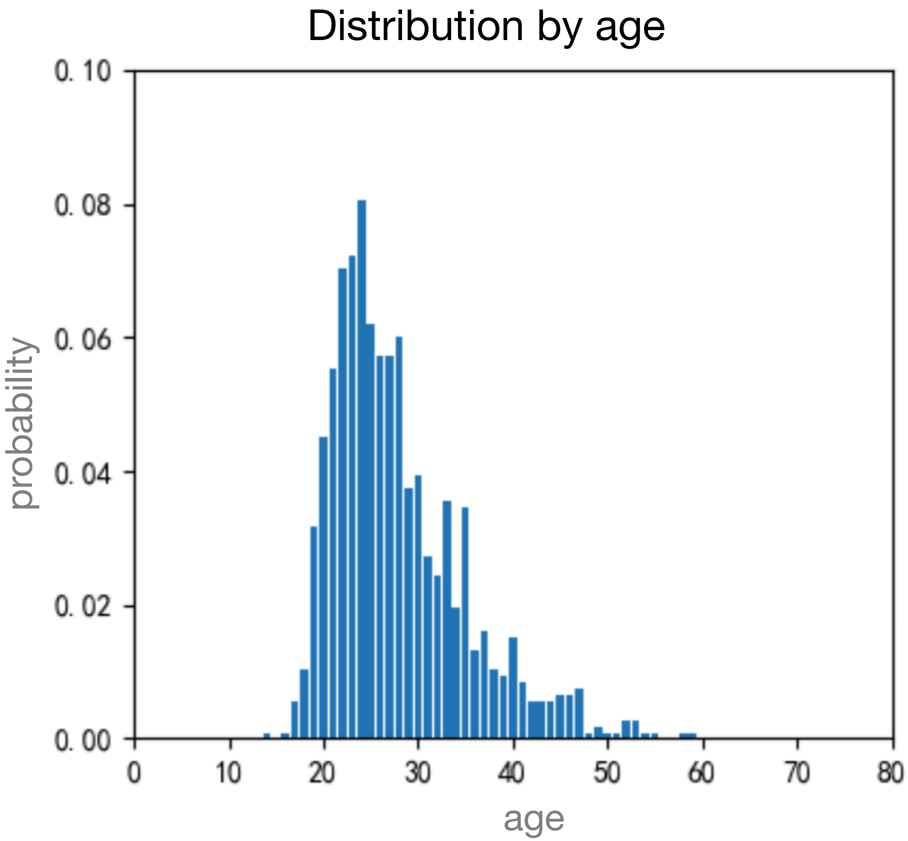
Online Click-through Rate

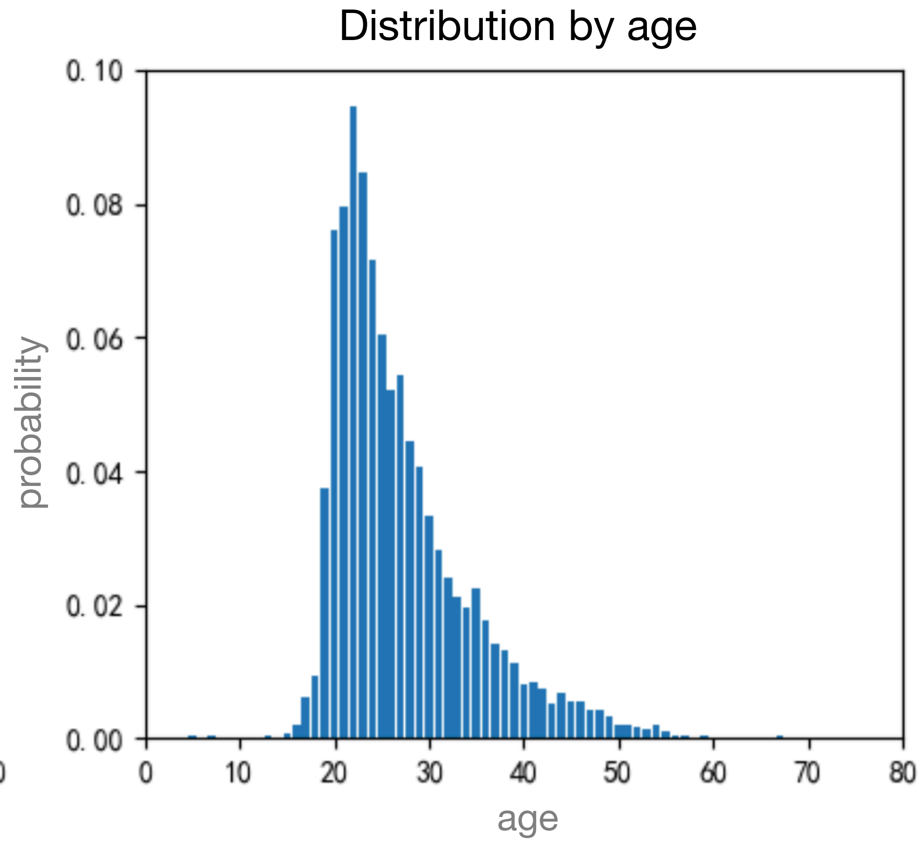
Comparison methods:

HCACR – Our method

Manual strategy

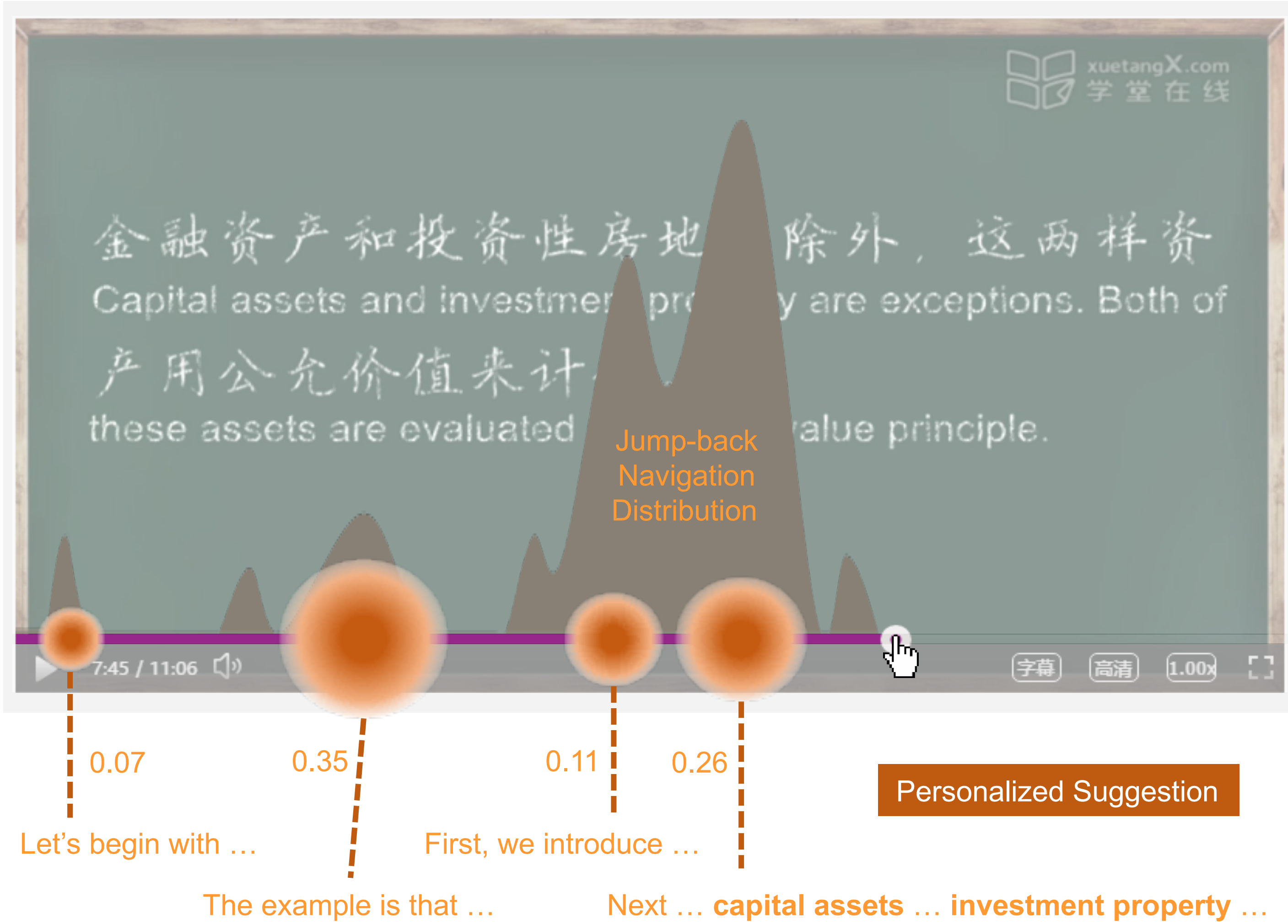
More Analysis



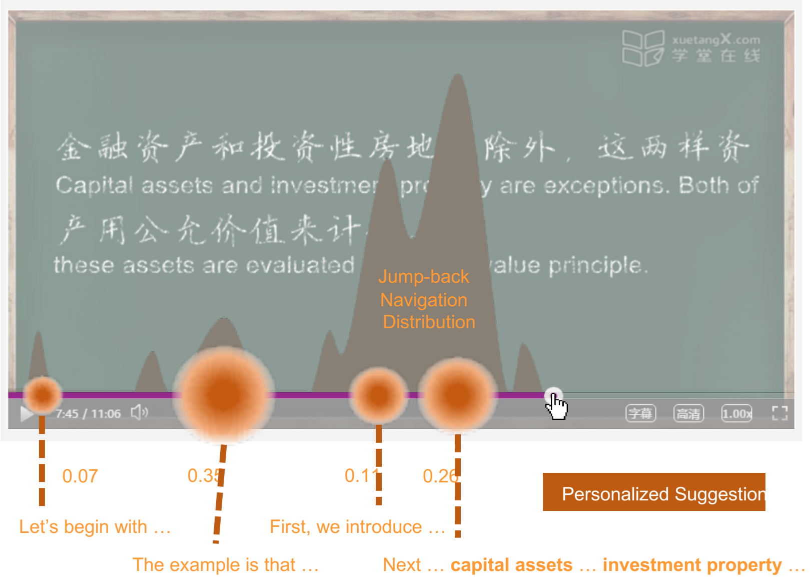


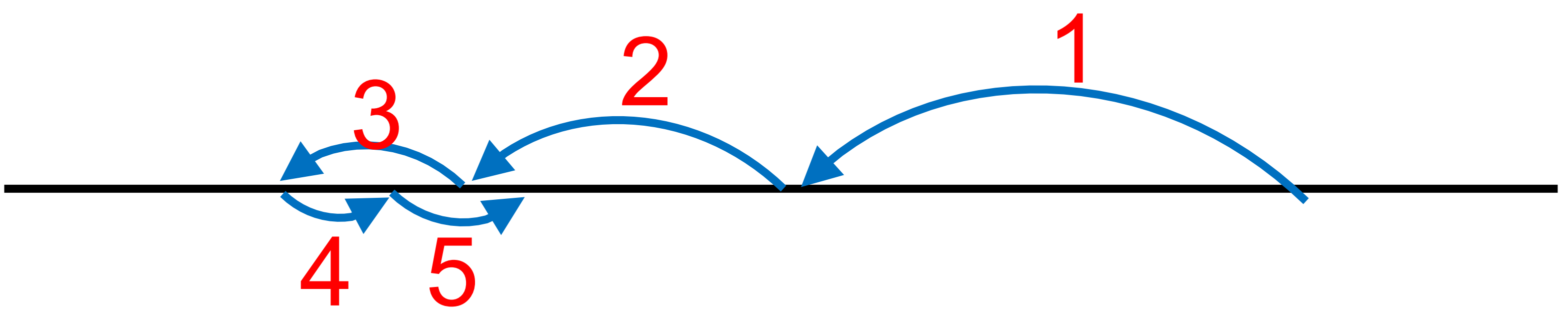
- Let start the simplest case
  - Course recommendation based on user interest
- What can we else?
  - Interaction when watching video?

Smart Jump
—Automated suggestion for video navigation



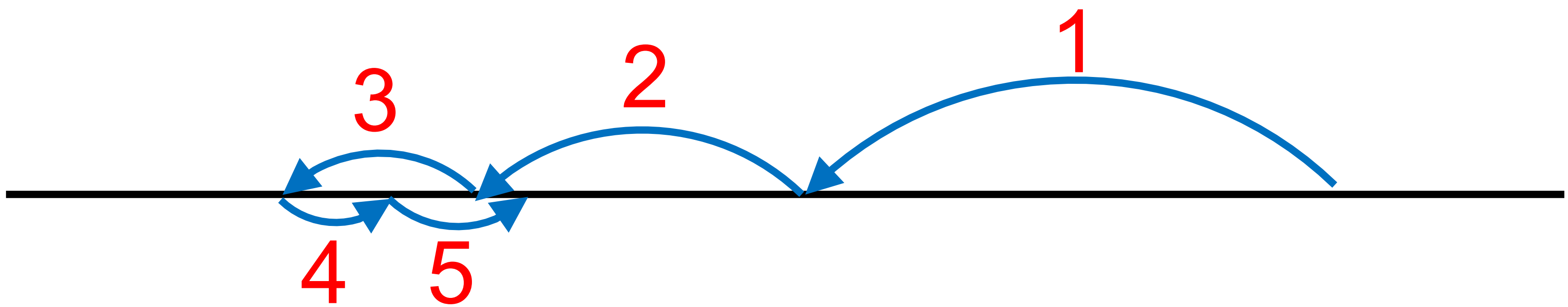
Average Jump



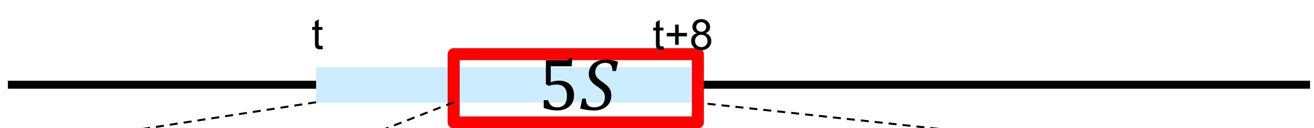


On Average: 2.6 Clicks = 5 seconds

Two Numbers



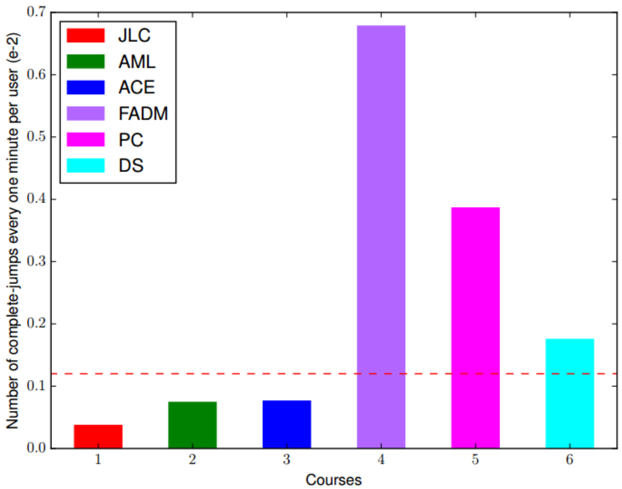
On Average: 2.6 Clicks = 5 seconds

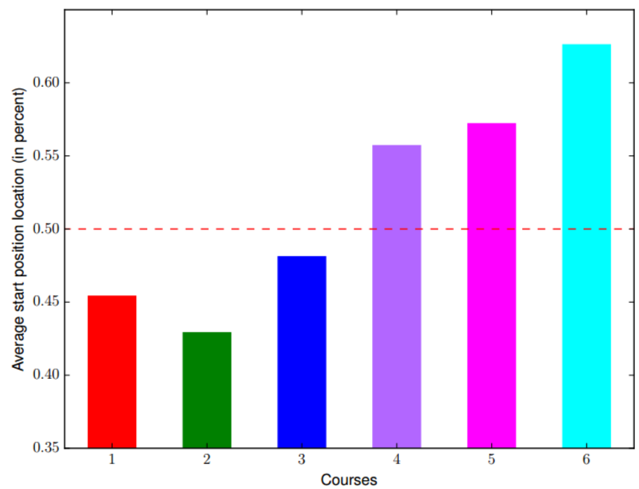


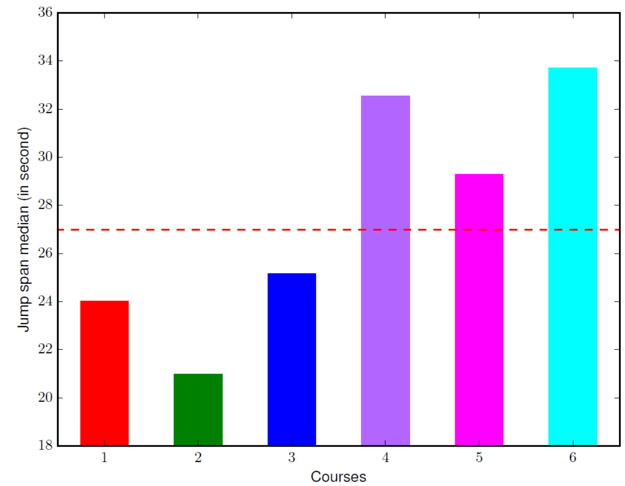
According to what we have discussed we find that the fifth activity belongs to cash outflow of a business activity.

$$5S \times 8{,}000{,}000\ users = 1.3\ years$$

Observations – Course Related





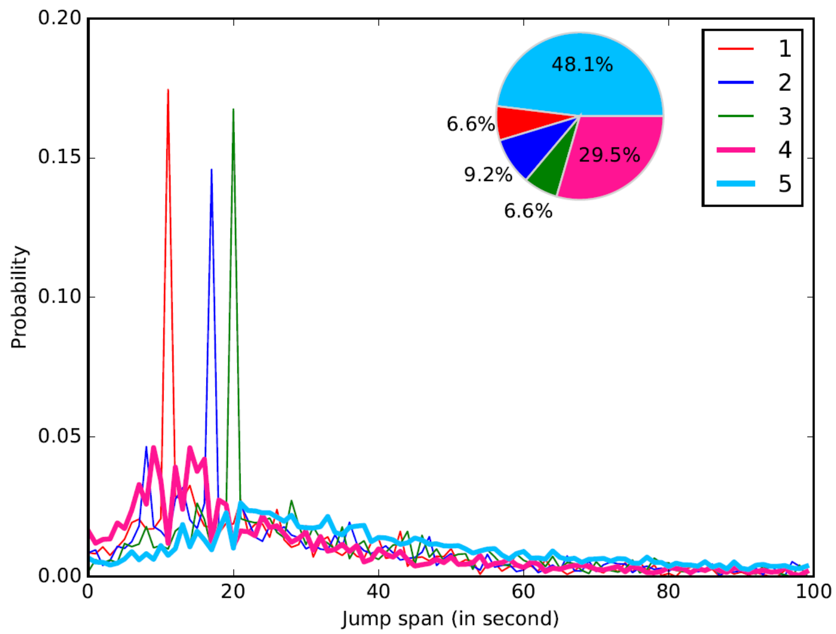


Science courses contain much more frequent jump-backs than non-science courses.

Users in non-science courses jump back earlier than users in science courses.

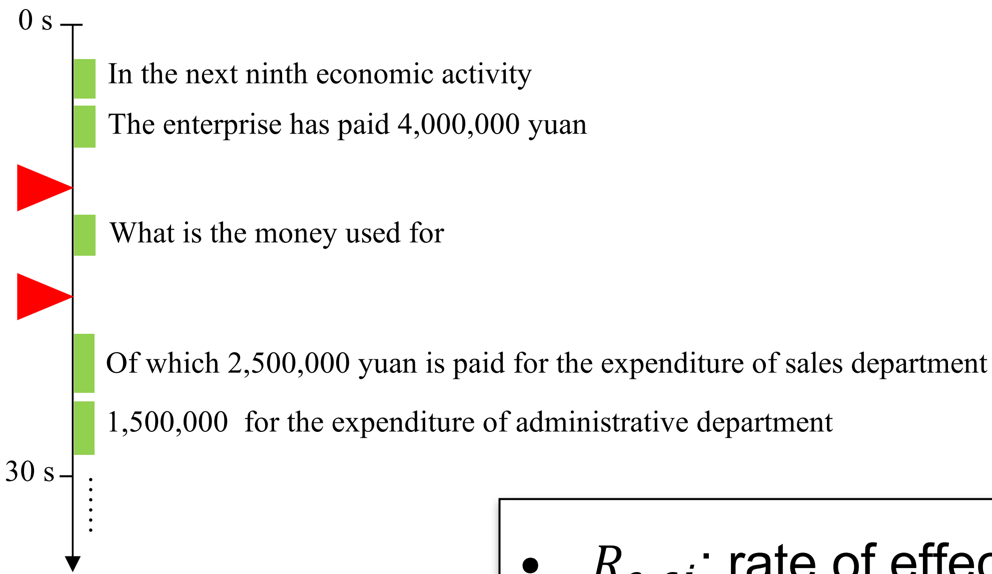
Users in science courses are likely to rewind farther than users in non-science courses.

Observations – User Related



- 6.6% users prefer 10 seconds
- 9.2% users prefer 17 seconds
- 6.6% users prefer 20 seconds

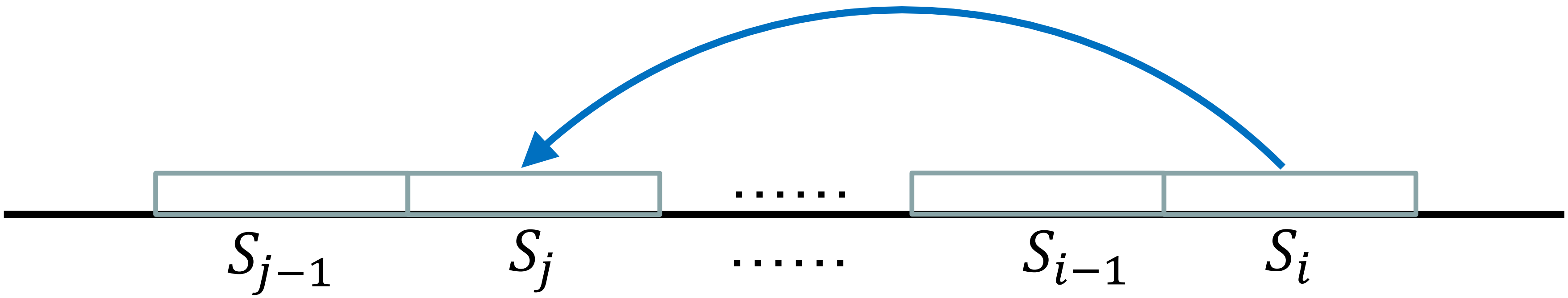
Video Segmentation



$$\underset{\Delta t}{\operatorname{argmax}}\ 2\frac{\boxed{R_{e\_cj}}\cdot\boxed{R_{n\_s}}}{R_{e\_cj}+R_{n\_s}}$$

- R_{e_cj}: rate of effective complete-jumps (start position and end position located in different segments).
- R_{n_s}: rate of non-empty segments (contains at least one start position or end position of some complete-jumps).

Problem Formulation



$$\underset{\Theta}{\operatorname{argmax}}\, P(s_j | u, v, s_i; \Theta)$$

[1] Han Zhang, Maosong Sun, Xiaochen Wang, Zhengyang Song, Jie Tang, and Jimeng Sun. Smart Jump: Automated Navigation Suggestion for Videos in MOOCs. **WWW'17**, pages 331-339.

Prediction Results

Course	Model	AUC	P@1	P@3	P@5
Science	**LRC**	72.46	35.95	65.54	80.13
	SVM	71.92	35.45	66.15	81.99
	FM	74.02	37.61	**76.04**	**89.59**
Non-science	**LRC**	72.59	69.23	73.23	89.32
	SVM	73.52	68.39	76.64	91.30
	FM	73.57	67.56	**88.43**	**96.05**

- LRC, SVM, and FM are different models
- FM is defined as follows

$$\hat{y}(\mathbf{x}_i) = w_0 + \sum_{j=1}^{d} w_j x_{i,j} + \boxed{\sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} x_{i,j} x_{i,j'} \langle \mathbf{p}_j, \mathbf{p}_{j'} \rangle}$$

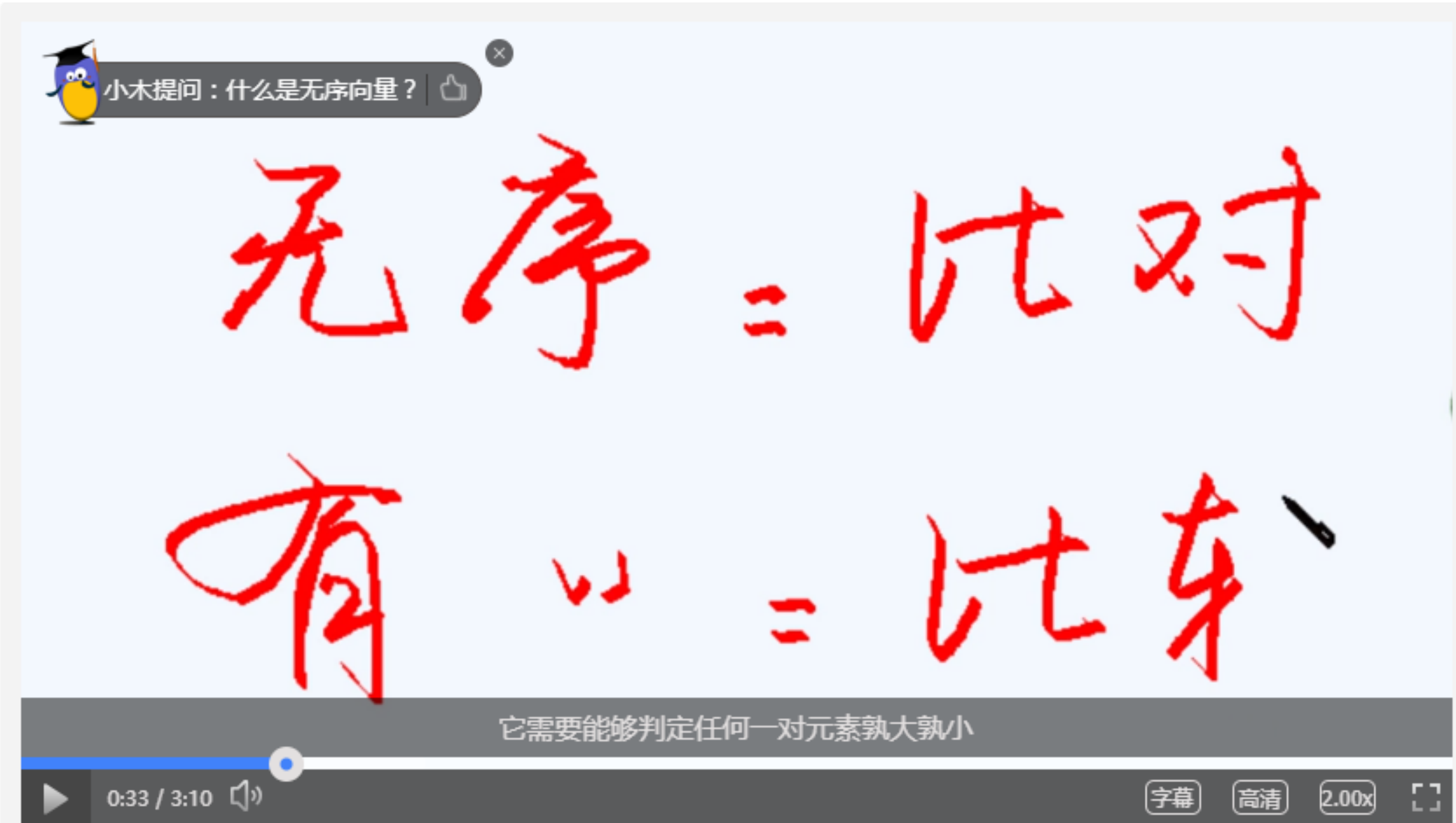
More

- Let start the simplest case
  - Course recommendation based on user interest
- What can we else?
  - Interaction when watching video?
  - Interaction->intervention

Active Question

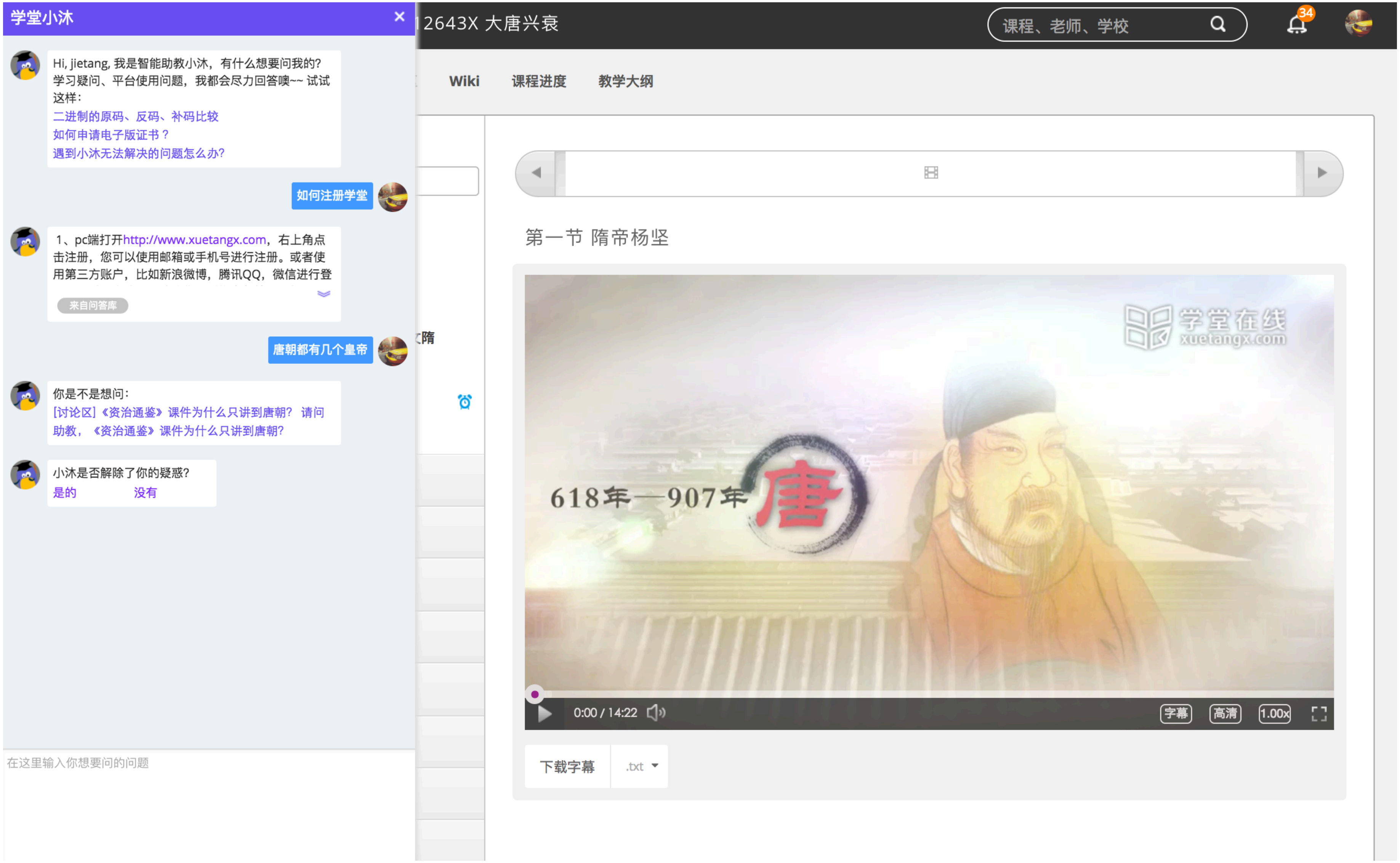
What is time complexity?

What is **Random Vector?**

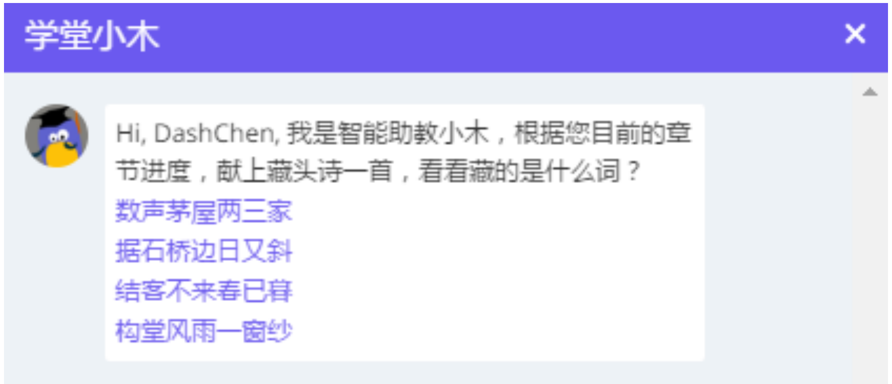


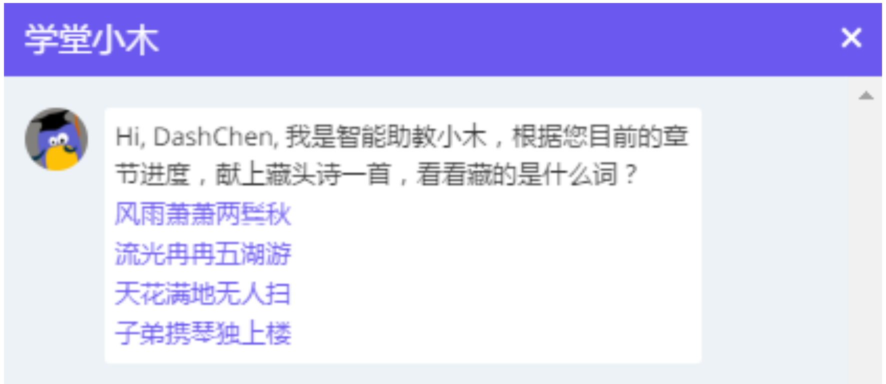
	#Questions
Total_request	30991
feedback	569
Feedback_ratio	0.0184
User-thumb_up	132
User-cancel	503
Thumb_ratio	0.24

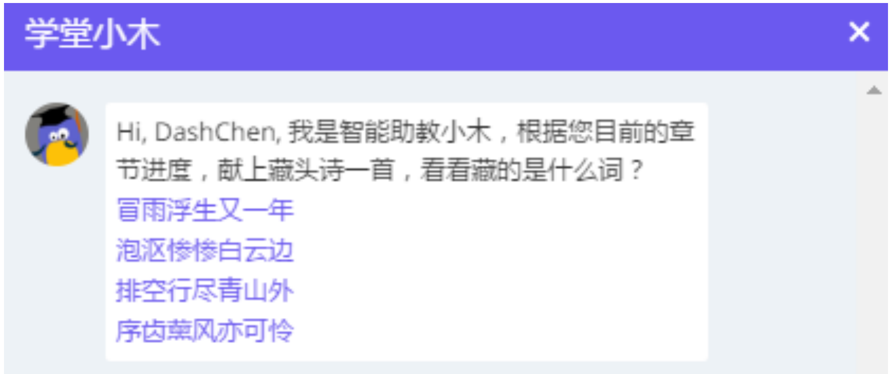
LittleMU (小木)

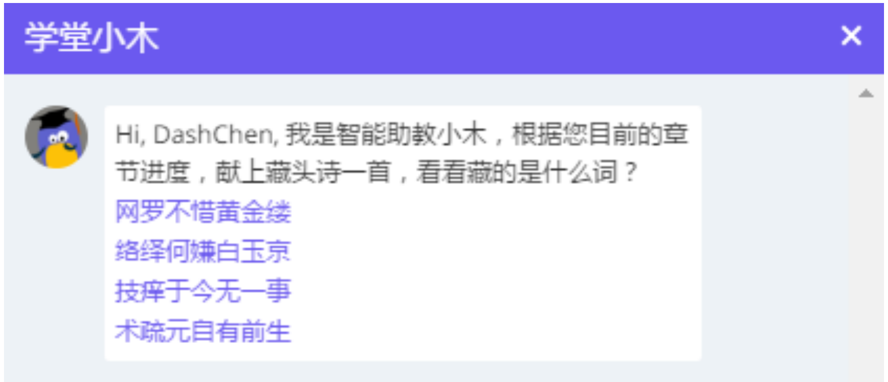

学堂小沐
Hi, jietang, 我是智能助教小沐，有什么想要问我的？学习疑问、平台使用问题，我都会尽力回答噢~~ 试试这样：
二进制的原码、反码、补码比较
如何申请电子版证书？
遇到小沐无法解决的问题怎么办?
如何注册学堂
1、pc端打开http://www.xuetangx.com，右上角点击注册，您可以使用邮箱或手机号进行注册。或者使用第三方账户，比如新浪微博，腾讯QQ，微信进行登
来自问答库
唐朝都有几个皇帝
你是不是想问：
[讨论区]《资治通鉴》课件为什么只讲到唐朝？ 请问助教，《资治通鉴》课件为什么只讲到唐朝?
小沐是否解除了你的疑惑?
是的
没有
在这里输入你想要问的问题
2643X 大唐兴衰
课程、老师、学校
Wiki
课程进度
教学大纲
第一节 隋帝杨坚
学堂在线
xuetangx.com
618年—907年
唐
0:00 / 14:22
字幕
高清
1.00x
下载字幕
.txt

Acrostic Poem: 小木作诗

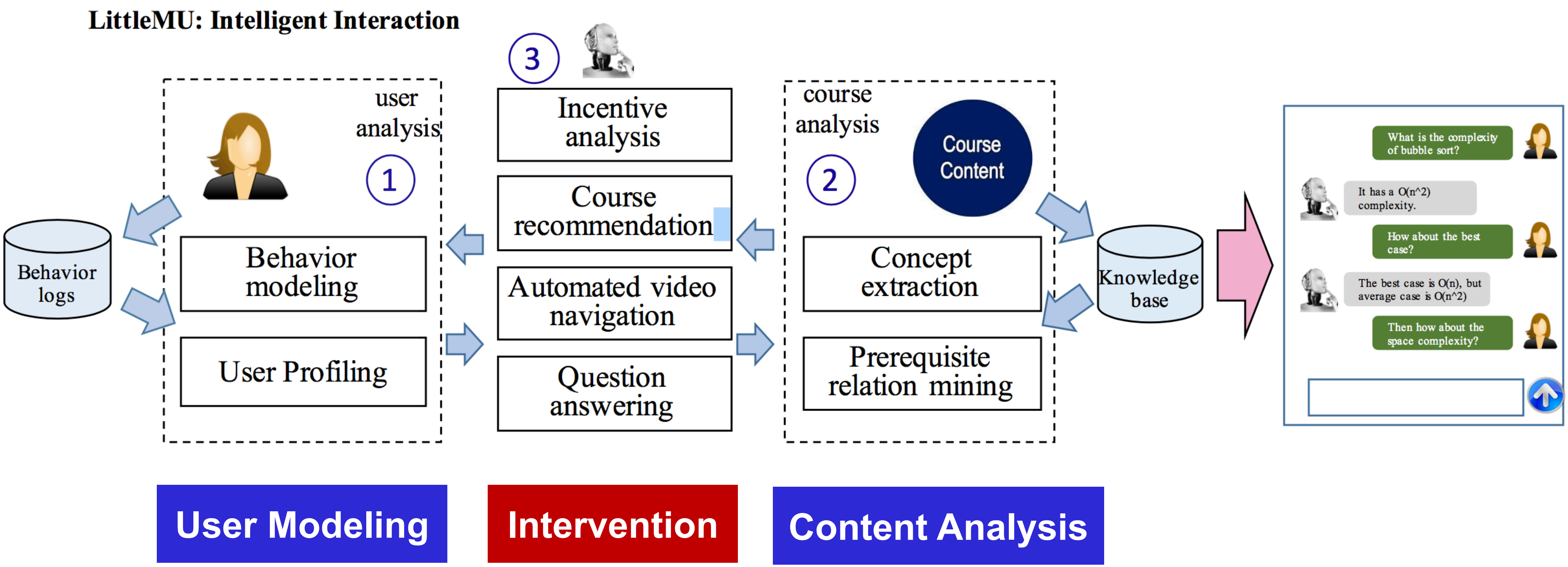








LittleMU (小木)



LittleMU: Intelligent Interaction
user analysis
1
Behavior modeling
User Profiling
Behavior logs
3
Incentive analysis
Course recommendation
Automated video navigation
Question answering
course analysis
2
Course Content
Concept extraction
Prerequisite relation mining
Knowledge base
What is the complexity of bubble sort?
It has a O(n^2) complexity.
How about the best case?
The best case is O(n), but average case is O(n^2)
Then how about the space complexity?
User Modeling
Intervention
Content Analysis

Recent Publications

- Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite Relation Learning for Concepts in MOOCs. In ACL'17.
- Xia Jing, Jie Tang, Wenguang Chen, Maosong Sun, and Zhengyang Song. Guess You Like: Course Recommendation in MOOCs. WI'17.
- Han Zhang, Maosong Sun, Xiaochen Wang, Zhengyang Song, Jie Tang, and Jimeng Sun. 2017. Smart Jump: Automated Navigation Suggestion for Videos in MOOCs. In WWW'17 Companion.
- Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. 2016. Modeling and Predicting Learning Behavior in MOOCs. In WSDM'16. 93–102.
- Jie Gong, Tracy Xiao Liu, Jie Tang, and Fang Zhang. Incentive Design on MOOC: a Field Experiment on XuetangX, Management Science (top in management). Submitted.
- Jie Tang, Tracy Xiao Liu, Zhenyang Song, Xiaochen Wang, Xia Jing, Jiezhong Qiu, Zhenhuan Chen, Chaoyang Li, Han Zhang, Liangmin Pan, Yi Qi, Xiuli Li, Jian Guan, Juanzi Li, and Maosong Sun. LittleMU: Enhancing Learning Engagement Using Intelligent Interaction on MOOCs. submitted to KDD.
- 李曼丽, 徐舜平, 孙梦嫽. MOOC 学习者课程学习行为分析——以“电路原理”课程为例[J]. 开放教育研究, 2015, 21(2): 63-69.
- 薛宇飞, 黄振中, 石菲. MOOC 学习行为的国际比较研究--以“财务分析与决策”课程为例[J]. 开放教育研究, 2015 (2015 年 06): 80-85.
- 薛宇飞，敬峡，裘捷中，唐杰，孙茂松. 一种在线课程中的作业互评方法：中国，201510531490.2.（中国专利申请号）
- 唐杰,张茜,刘德兵. 用户退课行为预测方法及装置. 201610292389.0 （中国专利申请号）

Thank you!

Collaborators: Jian Guan, Xiuli Li, Fenghua Nie (**XuetangX**)

Jie Gong (**NUS**), Jimeng Sun (**GIT**)

Maosong Sun, Tracy Liu, Juanzi Li (**THU**)

Xia Jing, Zhenhuan Chen, Liangmin Pan, Jiezhong Qiu, Han Zhang, Zhengyang Song, Xiaochen Wang, Chaoyang Li, Yi Qi (**THU**)

Jie Tang, KEG, Tsinghua U, http://keg.cs.tsinghua.edu.cn/jietang

Download all data & Codes, http://arnetminer.org/data

http://arnetminer.org/data-sna

Open Academic Graph (OAG)

https://aminer.org/open-academic-graph
https://www.openacademic.ai/

This data set is generated by linking two large academic graphs: Microsoft Academic Graph (MAG) **and** AMiner.org.
It includes **166,192,182 papers** from MAG and **154,771,162 papers** from AMiner.
We generated **64,639,608 linking (matching)** relations between the two graphs.

Data set	#Paper	#File	Total size	Date
Linking relations	64,639,608	1	1.6GB	2017-06-22
MAG papers	166,192,182	9	104GB	2017-06-09
AMiner papers	154,771,162	3	39GB	2017-03-22

Open Academic Data Challenge

https://biendata.com/competition/scholar/

OPEDAC
Open Academic Data Challenge

Microsoft, Tsinghua, CKCEST · $30,000 · 224 Teams

Open Academic Data Challenge 2017

Final Submissions

2017-07-18 2017-09-15

Information
Introduction
Rules
Data
Timeline & Prize
Evaluation
Organizers
Taskone
Submission
Make a submission
My submissions
Others
My Team

Introduction to Open Academic Data Challenge 2017

Academic data has witnessed an exponential growth in recent years as the total number of academic papers worldwide has exceeded 300 million and the number of academic researchers has reached 100 million. However, only about 3% of all the academic data contain semantic annotations. Such severe lack of semantic annotation information greatly restricts the service capacity of the academic big data' and its industrial development. Open Academic Data Challenge 2017 is hosted against such backdrop, committed to increasing the semantic annotation information in the academic database.

Hosted by Tsinghua University, Microsoft Research, the Knowledge Center of Chinese Academy of Engineering and the National Science Library of Chinese Academy of Sciences, and co-organized by Tsinghua Big Data Industries Association and IEEE Computer Society, Open Academic Data Challenge 2017 is aimed to create accurate academic profiles through mining the description of the scholars, their research interests and academic influence, and to explore the cutting-edge academic profiling techniques.

Based on the datasets provided by AMiner.org, a renowned academic data mining system and Microsoft Academic Graph,