Applying learning analytics on predicting students’ academic performance

Stephen J.H. Yang (楊鎮華)

Distinguished Professor

National Central University, Taiwan

Outline

- Part I： Analyzing students’ learning activities
  - Dashboard, graph, charts
  - Visualization of analysis results
- Part II： Predicting at-risk students who might
  - Drop out, withdraw, fail,
  - low score, low grade
- 5 Case studies

Case study 1

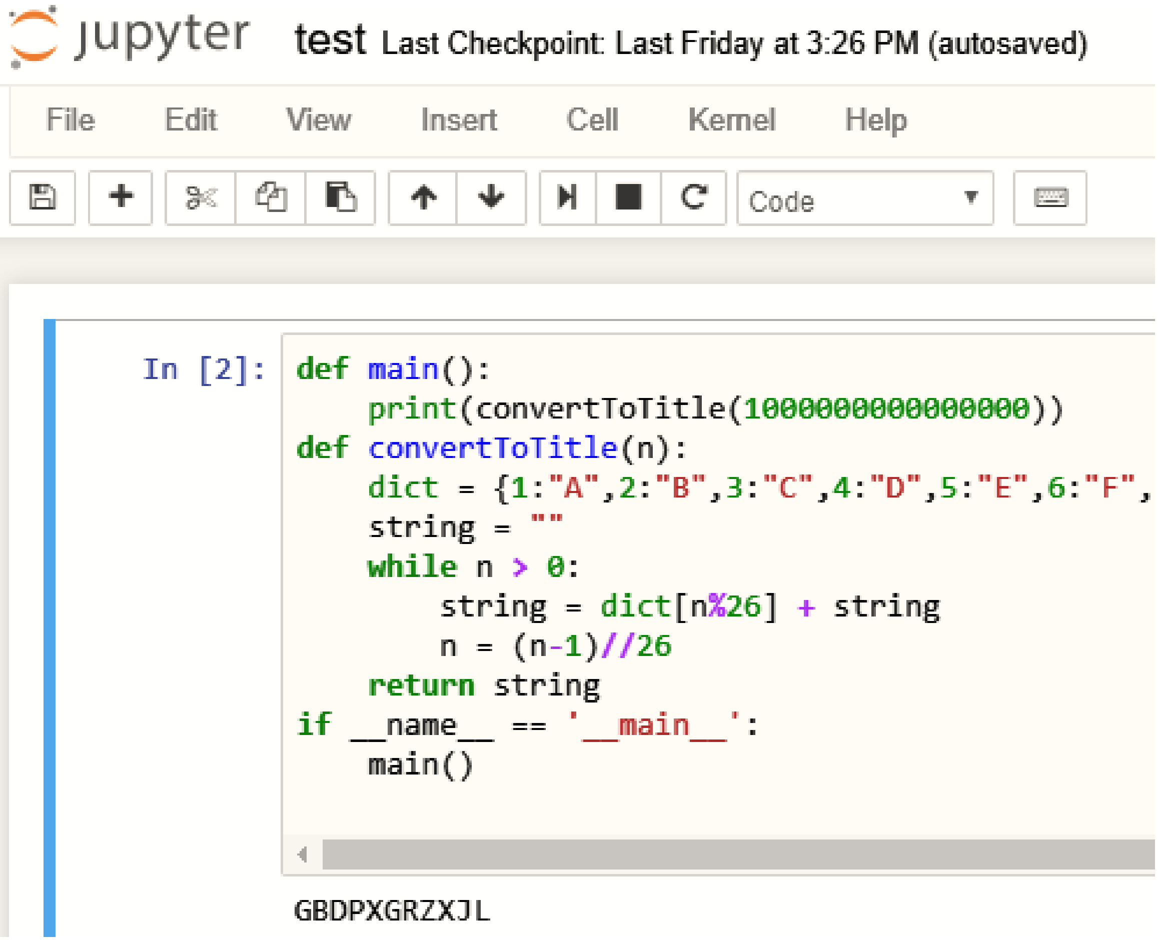
Data collection and Dashboard
for analyzing students’ coding activities
in a Python programming course

by 王紹宇, Owen

Tracking log of learning activities

- Data collection of digital footprint
  - eBook reading (lecture notes, slides)
  - Video viewing
  - Exercise practice (quiz, assessment, test)
  - Discussion forum
  - Science/engineering experiment
  - Program coding activities (hands-on exercises)

Python coding environment：Jupyter

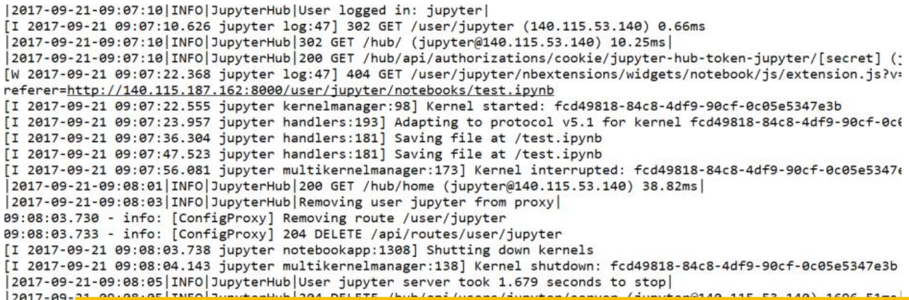


```
def main():
    print(convertToTitle(1000000000000000))
def convertToTitle(n):
    dict = {1:"A",2:"B",3:"C",4:"D",5:"E",6:"F",
    string = ""
    while n > 0:
        string = dict[n%26] + string
        n = (n-1)//26
    return string
if __name__ == '__main__':
    main()
```

```
GBDPXGRZXJL
```

Collecting students’ coding activities on Jupyter

Jupyter Log of hands-on exercise



```
|2017-09-21-09:07:10|INFO|JupyterHub|User logged in: jupyter|
[I 2017-09-21 09:07:10.626 jupyter log:47] 302 GET /user/jupyter (140.115.53.140) 0.66ms
|2017-09-21-09:07:10|INFO|JupyterHub|302 GET /hub/ (jupyter@140.115.53.140) 10.25ms|
|2017-09-21-09:07:10|INFO|JupyterHub|200 GET /hub/api/authorizations/cookie/jupyter-hub-token-jupyter/[secret] (:
[W 2017-09-21 09:07:22.368 jupyter log:47] 404 GET /user/jupyter/nbextensions/widgets/notebook/js/extension.js?v=
referer=http://140.115.187.162:8000/user/jupyter/notebooks/test.ipynb
[I 2017-09-21 09:07:22.555 jupyter kernelmanager:98] Kernel started: fcd49818-84c8-4df9-90cf-0c05e5347e3b
[I 2017-09-21 09:07:23.957 jupyter handlers:193] Adapting to protocol v5.1 for kernel fcd49818-84c8-4df9-90cf-0c
[I 2017-09-21 09:07:36.304 jupyter handlers:181] Saving file at /test.ipynb
[I 2017-09-21 09:07:47.523 jupyter handlers:181] Saving file at /test.ipynb
[I 2017-09-21 09:07:56.081 jupyter multikernelmanager:173] Kernel interrupted: fcd49818-84c8-4df9-90cf-0c05e5347
|2017-09-21-09:08:01|INFO|JupyterHub|200 GET /hub/home (jupyter@140.115.53.140) 38.82ms|
|2017-09-21-09:08:03|INFO|JupyterHub|Removing user jupyter from proxy|
09:08:03.730 - info: [ConfigProxy] Removing route /user/jupyter
09:08:03.733 - info: [ConfigProxy] 204 DELETE /api/routes/user/jupyter
[I 2017-09-21 09:08:03.738 jupyter notebookapp:1308] Shutting down kernels
[I 2017-09-21 09:08:04.143 jupyter multikernelmanager:138] Kernel shutdown: fcd49818-84c8-4df9-90cf-0c05e5347e3b
|2017-09-21-09:08:05|INFO|JupyterHub|User jupyter server took 1.679 seconds to stop|
|2017-09-
|2017-09-
|2017-09-
```

Filter & extraction of coding activities

Input

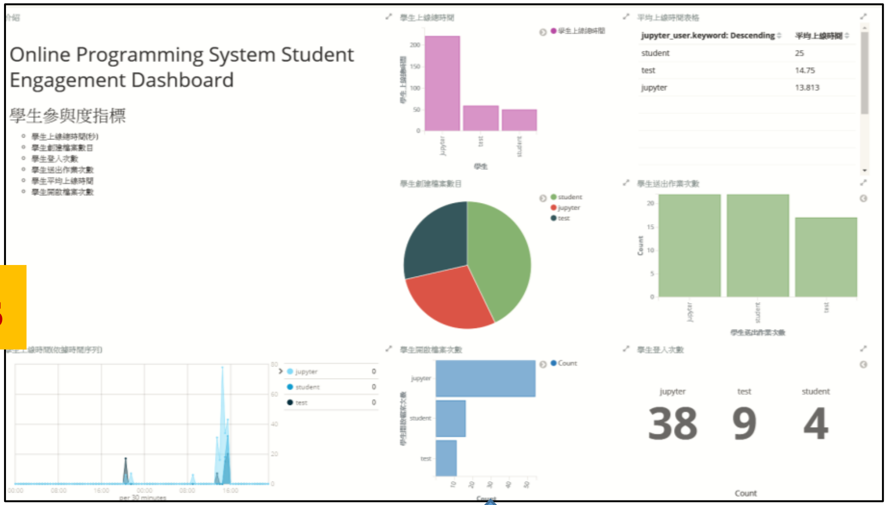
path => "/var/log/jupyter_screen.log"

Filter

Gork {match => ["timestamp", "yyyy-MM-dd HH:mm:ss.SSS"]}

Output

hosts => ”elasticsearch:9200"



Online Programming System Student Engagement Dashboard

學生參與度指標

- 學生上線總時間(秒)
- 學生創建檔案數目
- 學生登入次數
- 學生送出作業次數
- 學生平均上線時間
- 學生開啟檔案次數

jupyter_user.keyword: Descending	平均上線時間
student	25
test	14.75
jupyter	13.813

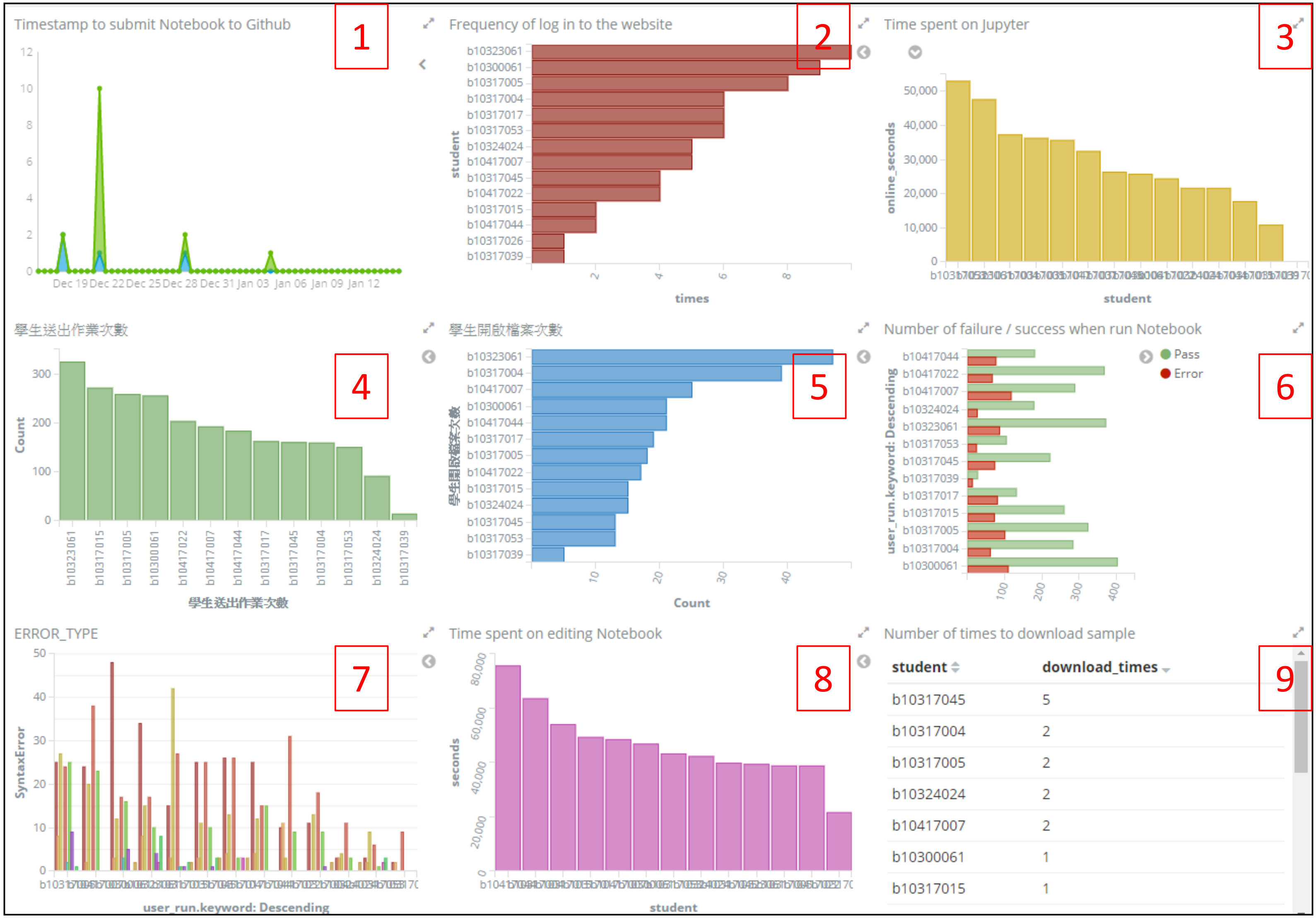
Visualization of coding activities

Elasticsearch & Kibana

Students’ coding activities of interest

NO.	Coding activities (Features)
1	Timestamp of assignments completed
2	Frequency of login to Jupyterhub
3	Time spent online (Jupyterhub)
4	Time spend on coding
5	Numbers of fail/success run
6	Types of coding error
7	Times of assignment submission
8	Times of open files
9	Number of podcasts used

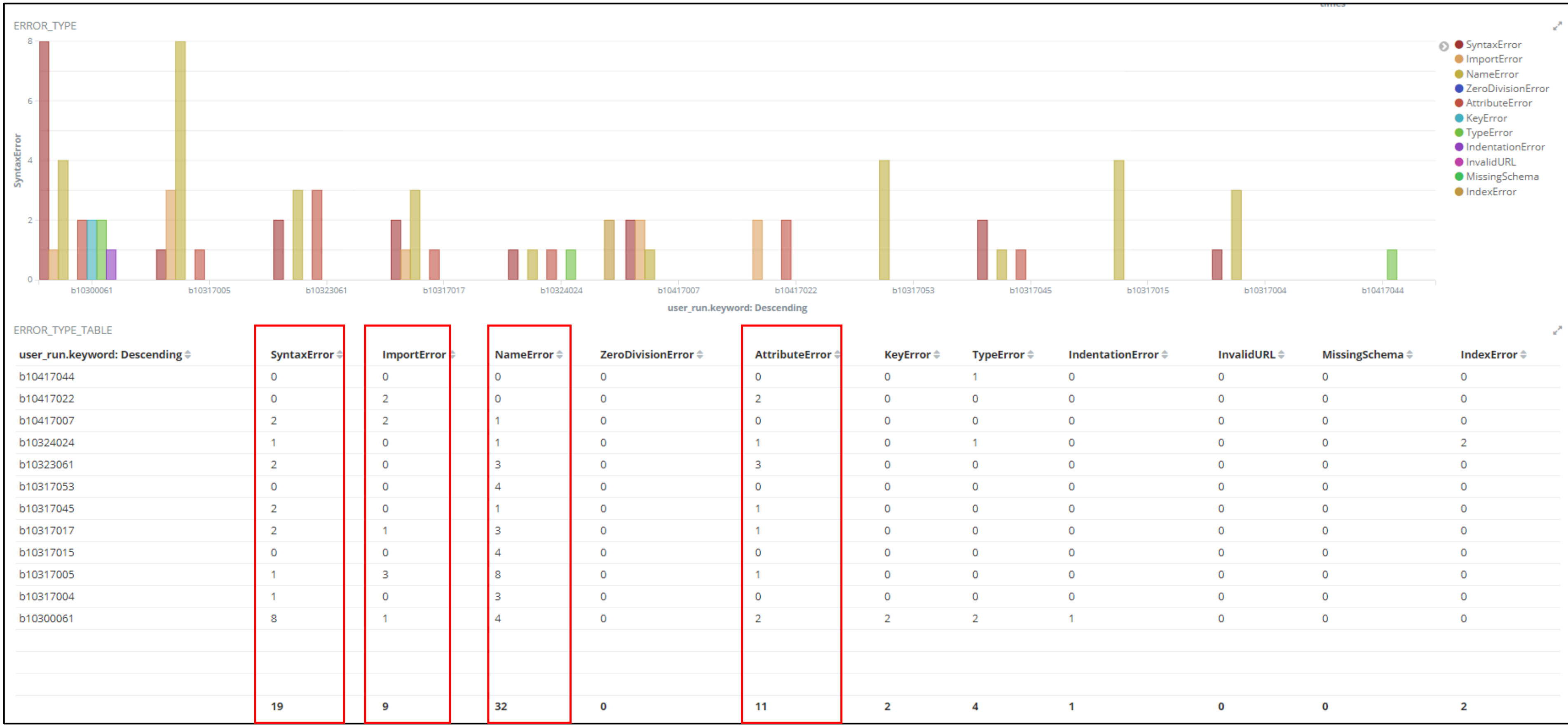
Corresponding Dashboard of 9 activities



student	download_times
b10317045	5
b10317004	2
b10317005	2
b10324024	2
b10417007	2
b10300061	1
b10317015	1

For individual student and the whole class

6 Types of coding errors and errors distribution



ERROR_TYPE_TABLE

user_run.keyword: Descending	SyntaxError	ImportError	NameError	ZeroDivisionError	AttributeError	KeyError	TypeError	IndentationError	InvalidURL	MissingSchema	IndexError
b10417044	0	0	0	0	0	0	1	0	0	0	0
b10417022	0	2	0	0	2	0	0	0	0	0	0
b10417007	2	2	1	0	0	0	0	0	0	0	0
b10324024	1	0	1	0	1	0	1	0	0	0	2
b10323061	2	0	3	0	3	0	0	0	0	0	0
b10317053	0	0	4	0	0	0	0	0	0	0	0
b10317045	2	0	1	0	1	0	0	0	0	0	0
b10317017	2	1	3	0	1	0	0	0	0	0	0
b10317015	0	0	4	0	0	0	0	0	0	0	0
b10317005	1	3	8	0	1	0	0	0	0	0	0
b10317004	1	0	3	0	0	0	0	0	0	0	0
b10300061	8	1	4	0	2	2	2	1	0	0	0
	19	**9**	**32**	**0**	**11**	**2**	**4**	**1**	**0**	**0**	**2**

Syntax Error：19　Import Error：9　Name Error：32　Attribute Error：11

Work in progress of Study 1

- Automatic detection of students'
  - Coding styles
  - Coding defects (NASA MDP datasets, transfer learning)

- Apply DANN (Domain-adversarial Neural Network) to identify students' coding defects and their relationship with coding styles

- Correlation of students' SRL capability and final academic performance
- Correlation of collaborative programming and students' final academic performance

Case study 2

What are the critical factors affecting students' academic performance in MOOCs

Part of an empirical study of Taiwan's MOOCs initiative

by Anna

Taiwan’s MOOCs initiative

- Part of Taiwan’s Digital learning initiative
- Funded by Ministry of Education in Taiwan
- National project for funding universities to develop MOOCs courseware and mechanism, as well as data analytics.

Data analytics of Taiwan’s MOOCs initiative

- Platforms (MOE official)
  - Open edX (based on MIT Open edX)
  - Insight$^+$ (tracking logs, data analytics)

- Period of data collection
  - 3 years：Sept. 2014 - Aug. 2017

- Courses: 590

- Registers: 32,120

Top 10 courses (out of 590) in terms of registers

Course name	Number of registers	Course begin	Course end
105 年課文本位閱讀理解策略教學初階課程	3,802	2016/10/ 07	2017/8/10
讓老闆不得不重用你~正在崛起的「專案管理」_201410	1,310	2014/10/20	2015/4/30
大學普通物理實驗—手作坊	1,289	2014/12/1	2015/4/12
台灣傳統糕餅文化與製作、創新	1,174	2015/6/11	2016/1/26
從車庫到金庫—看見台灣企業生命力	1,013	2014/10/26	2015/7/12
軟體設計 I- 物件導向設計	750	2016/8/1	2020/8/31
巧克力製作	727	2015/12/15	--
軟體設計 III- 設計樣式	680	2016/8/25	2020/731
2015 K-12能源科技教育種子教師培訓 (初階)	608	2015/5/14	2015/6/15
讓老闆不得不重用你~正在崛起的「專案管理」_104	610	2015/12/1	--

Top 10 courses (out of 590) in terms of activity/register

Course name	Number of Registers	Number of video viewing event	Activity/Register
1041微積分聯合教學--地科班 Calculus 101	59	33,928	575
虛擬實境	31	10,486	338
電腦攻擊與防禦	33	7,261	220
影像處理	47	9,446	200
虛擬實境	39	7257	186
台灣小說選讀	74	13,458	181
普通化學—平衡、酸與鹼、水溶液的平衡	77	10,100	131
Linux作業系統核心	48	5,300	110
書報討論	138	14,953	108

Data analytics for Calculus 101

- Calculus 101 was offered by National Central University, it is the most active course in Taiwan's MOOCs initiative

- Blended learning (59 students)
    - Open edX for online video viewing
    - Maple TA for online exercise & assessment
    - Offline assignment & exam
    - Knowing students' final academic grade

Correlation of MOOCs activities and final grade

- **Algorithms**
  - MCA (Multiple Corresponding Analysis)
  - MFA (Multiple Factor Analysis)

- **MCA** to find out the correlation of **leaning activities** and **final grade**

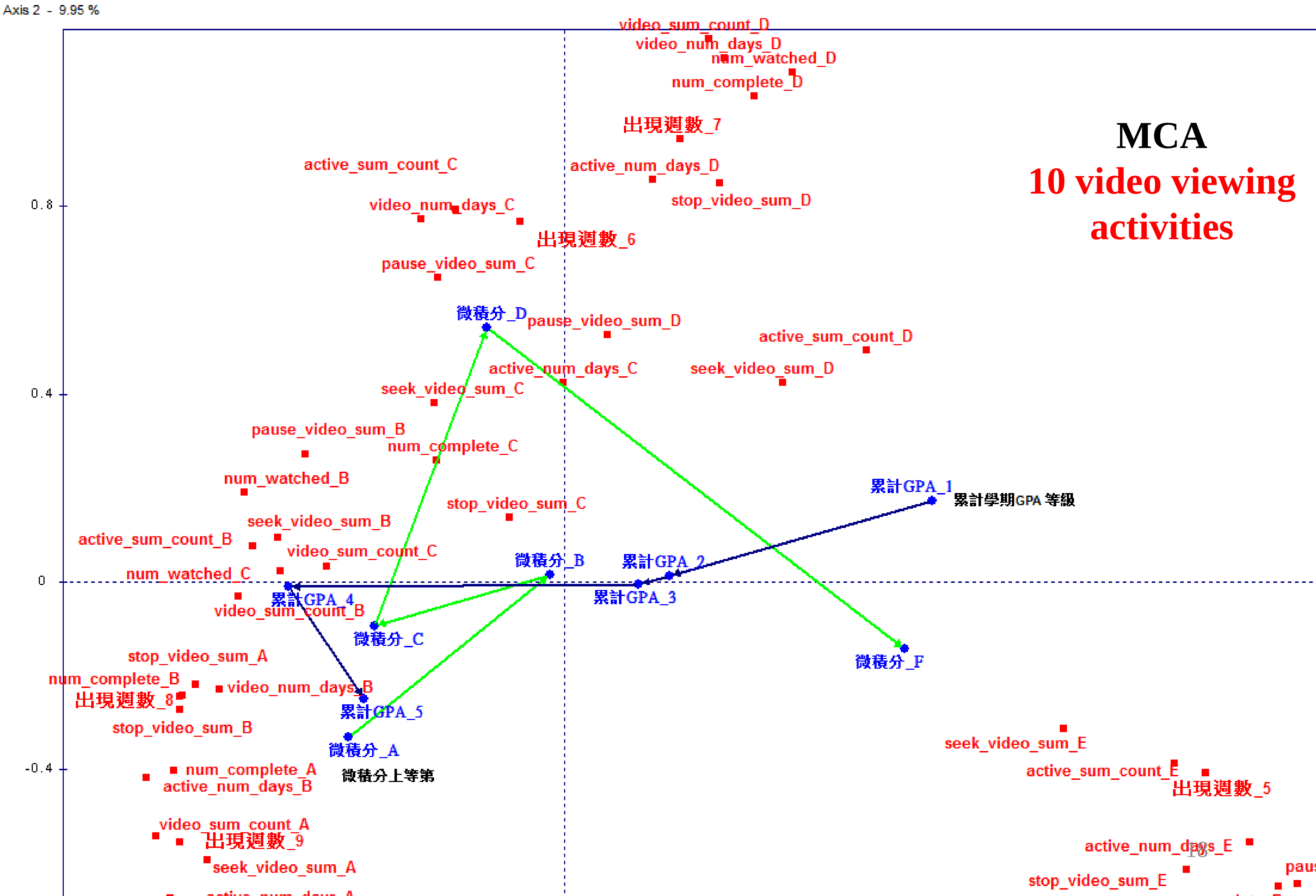
  RQ_1：How is video viewing activity correlated to grade?

  RQ_2：How is exercise practice activity correlated to grade?

- **MFA** to find out the **critical factors** affecting students **final grade**

  RQ_3：What are the factors affecting high grade and low grade?

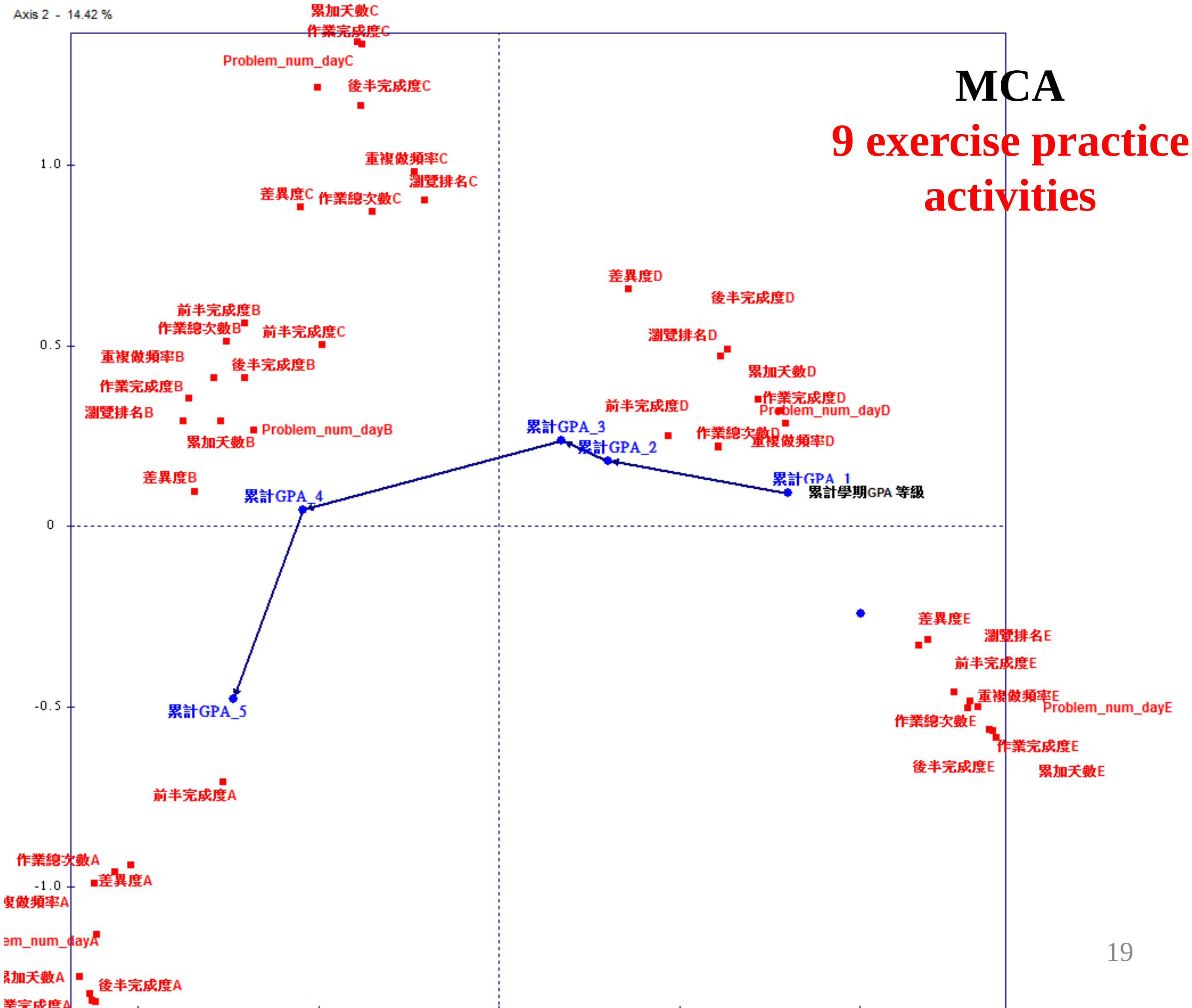
RQ_1：How is video viewing activity correlated to grade?



MCA

10 video viewing activities

RQ_2：How is exercise practice activity correlated to grade?



MCA

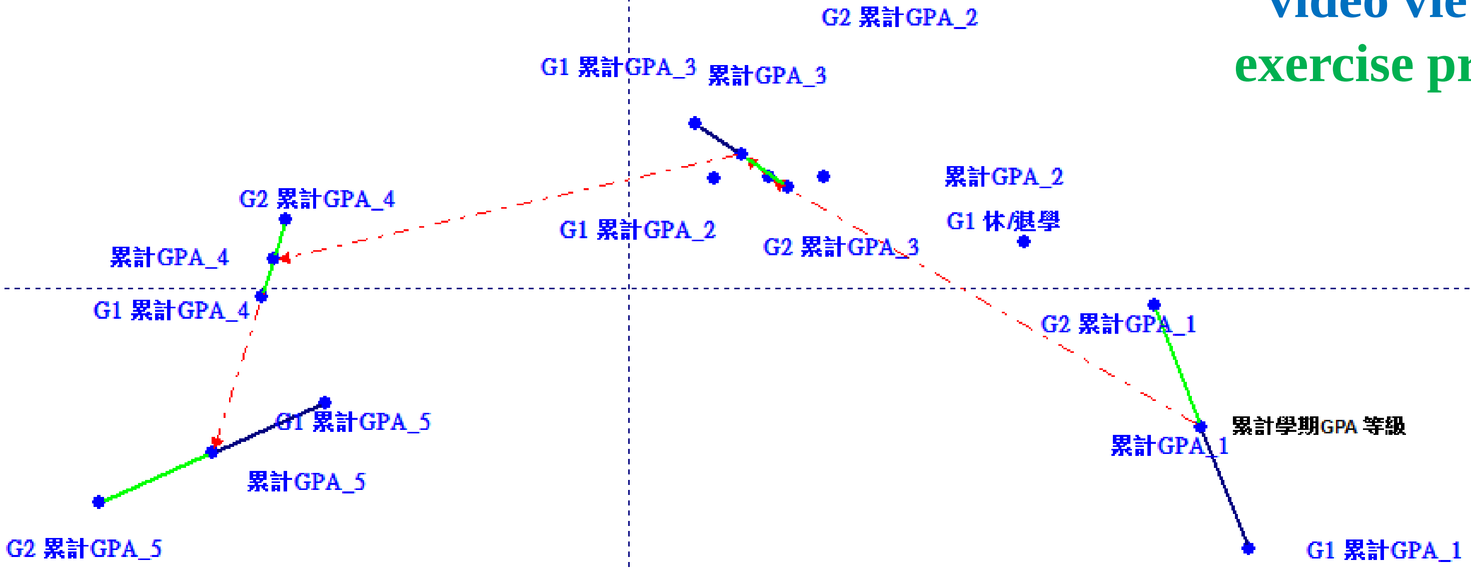
9 exercise practice activities

RQ_3：What are the factors affecting high grade and low grade?

MFA

video viewing activities

exercise practice activities



Correlation of MOOCs activities and grade

- Summary
  - Video viewing is the critical factor resulting in low grade
  - Exercise practice is the critical factor resulting in high grade

- **RQ_4**：what kinds of viewing activities resulting in low grade?

RQ_4：What kinds of viewing activities resulting in low grade?

- Group video viewing activities into 4
  - G1：video viewing events
  - G2：first time watched
  - G3：video viewing frequency/day
  - G4：video viewing completion rate

- Find out the **critical factors** for improving students' grade
  - How to improve students' grade from low to middle?
  - How to improve students' grade from middle to high?

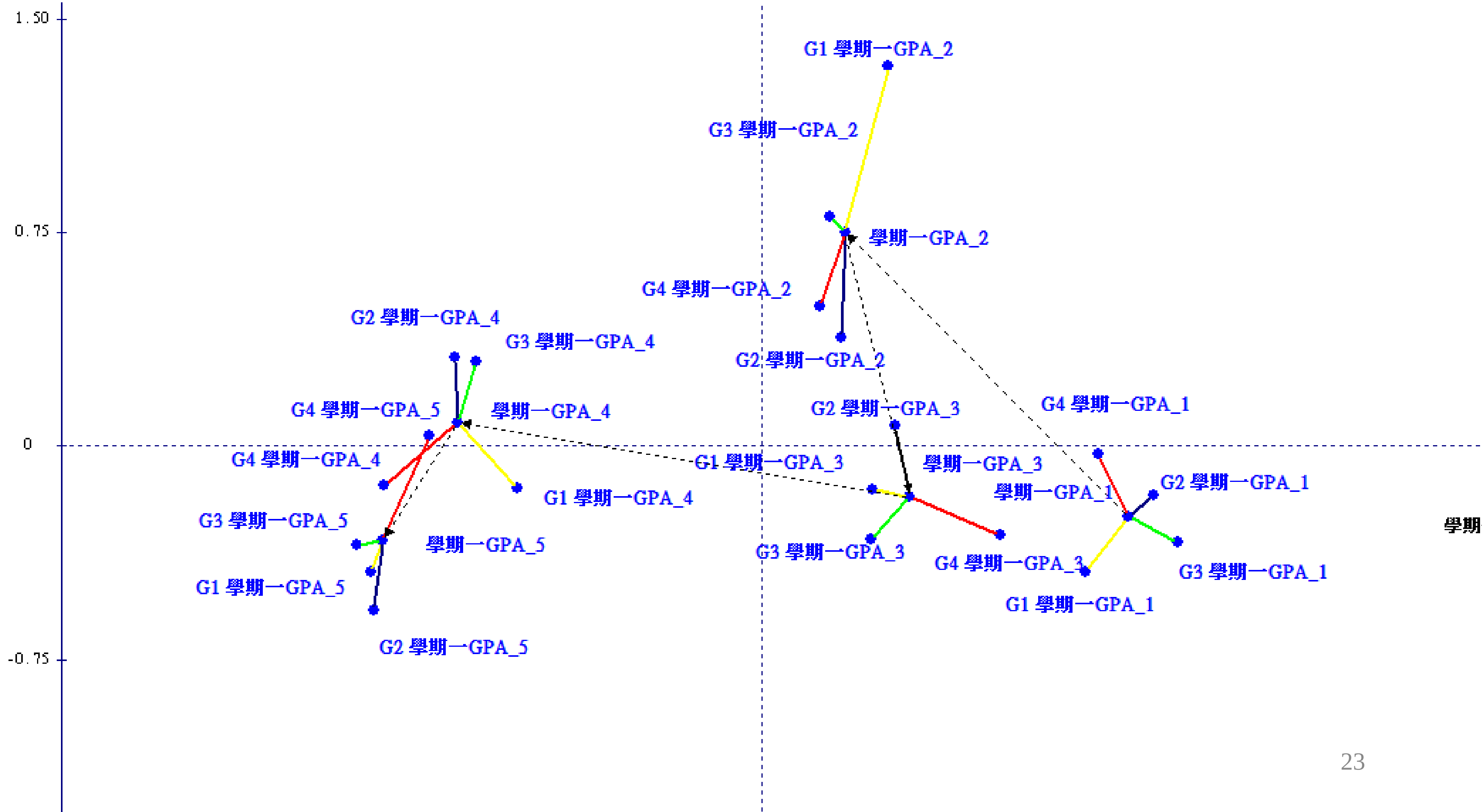
GPA 1 is highly related to G2 and G3

GPA 3 is highly related to G4

GPA 4 is highly related to G4

GPA 5 is highly related to G1、G2 and G3

MFA



Critical factors resulting in low grade

- G2, G3 are the critical factors resulting in low grade (GPA 1)
  - G2：first time watched
  - G3：video viewing frequency/day

- Suggestion：
  - Watch video early, increase viewing frequency are two critical factors for improving grade from low to middle.

Critical factors resulting in middle grade

- G4 is the critical factor resulting in middle grade (GPA 3 & GPA 4)
  - G4：video viewing completion rate

- Suggestion：
  - Complete video viewing is the critical factor for improving grade from middle to high.

Work in progress of Study 2

- Correlation of students' high school academic performance and their college performance
  - Freshmen year's vs college 4 year's performance

- Correlation of students' types of entrance channels and their the college performance
  - Freshmen year's vs college 4 year's performance

- For entrance exam, which subject is most related to students' Freshmen year's performance?

Case study 3

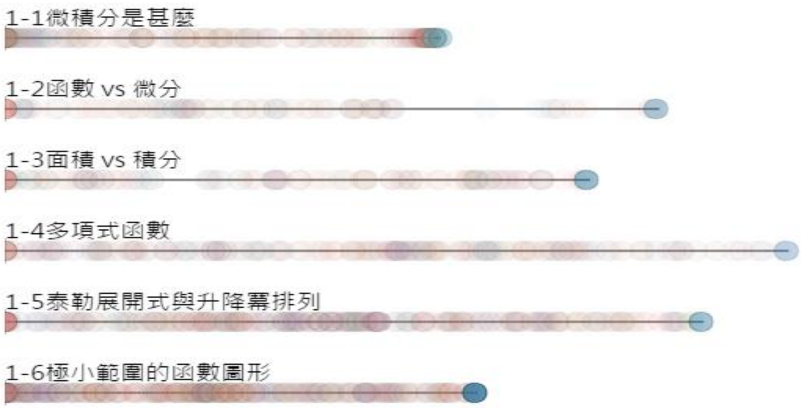
Visual analysis of students' video viewing patterns in MOOCs

Part of an empirical study of Taiwan's MOOCs initiative

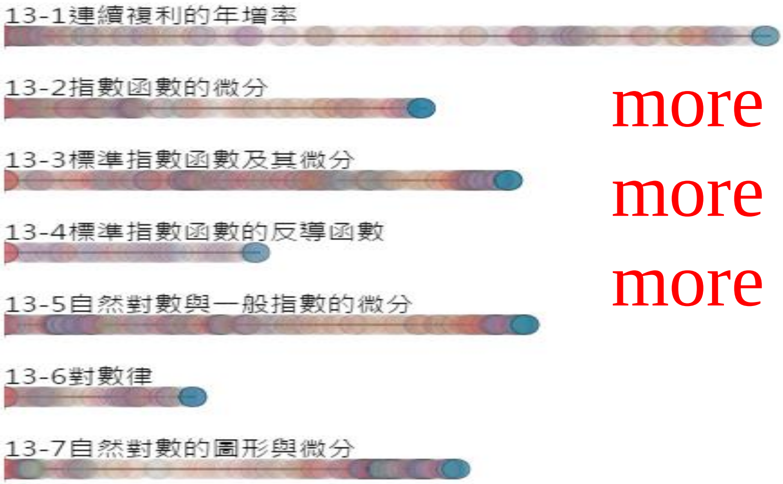
by Anna

Heat graph of viewing events (video, eBook)

Viewing events: play, pause, stop, forward, backward



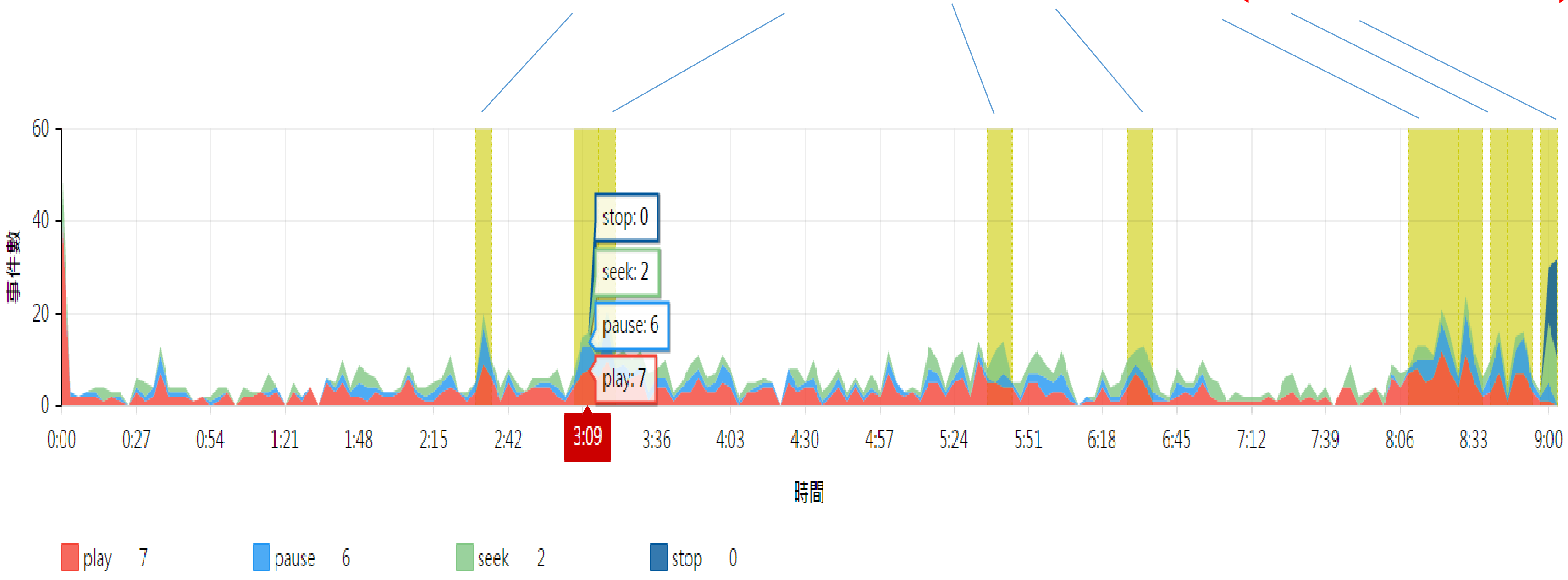
less dark,
less viewing events
less focus



more dark,
more viewing events
more focus

Peak graph of viewing events (video, eBook)

Periods of attention in a video (more focus)



Individual student’s viewing sequence in a video

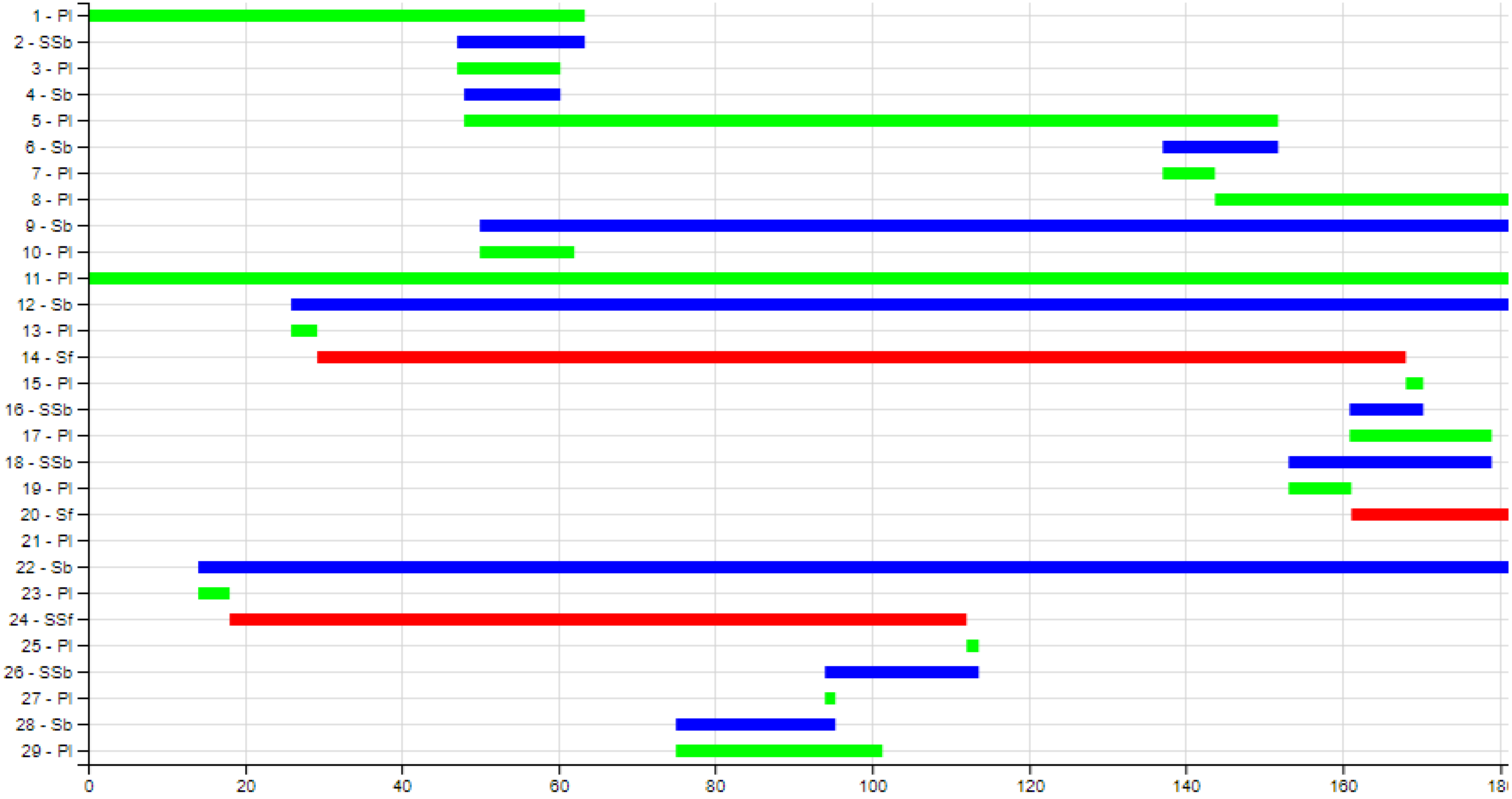
Green: Play

Blue: Backward

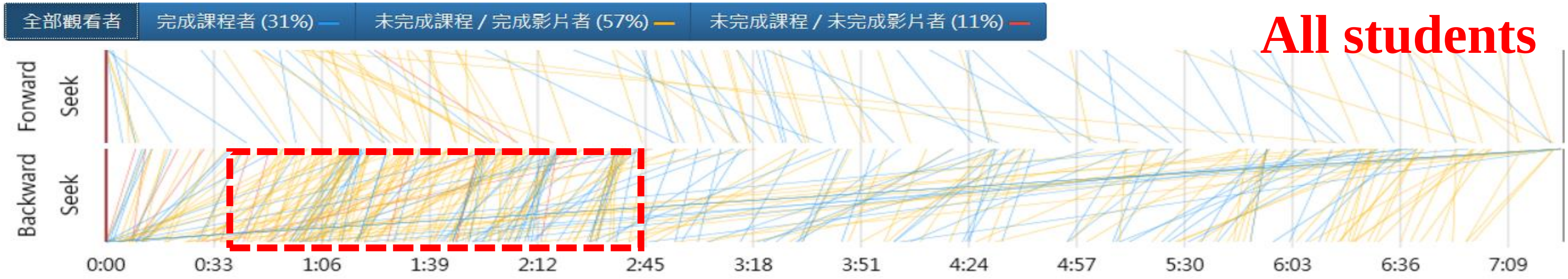
Red: Forward

X: timestamp of a video

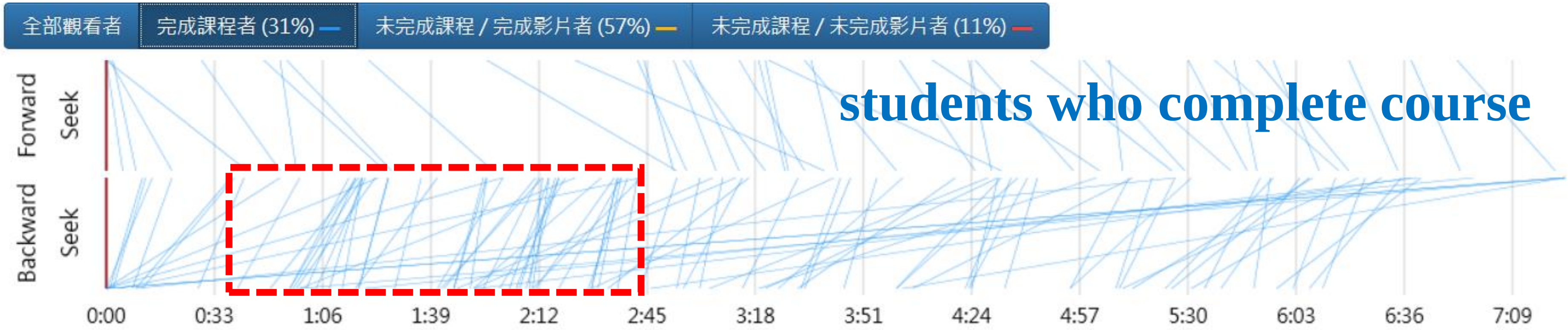
Y: Pattern



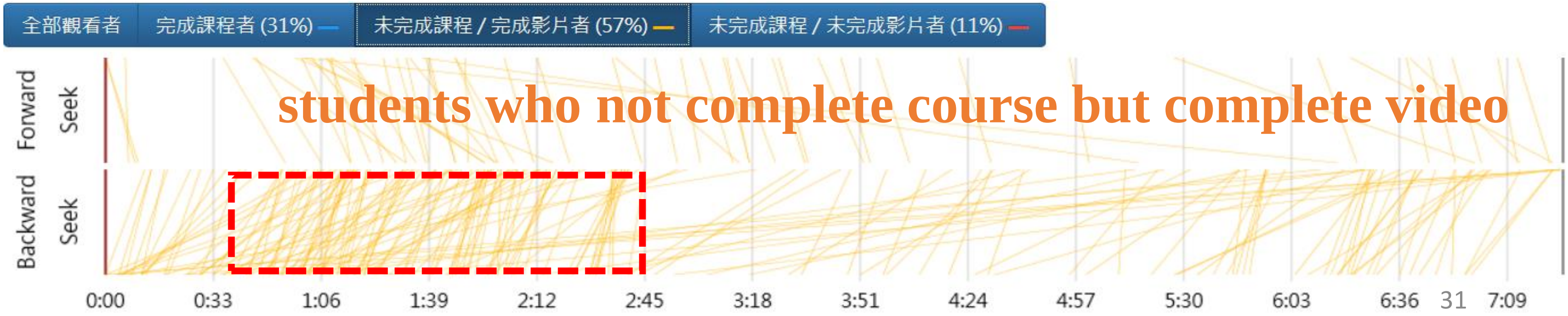
High density of **Backward** might due to material is too difficult, struggle



All students

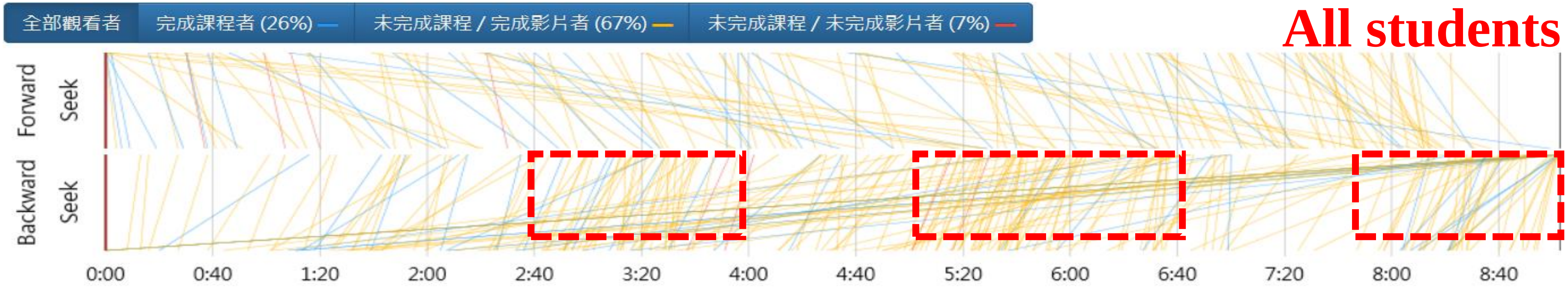


students who complete course

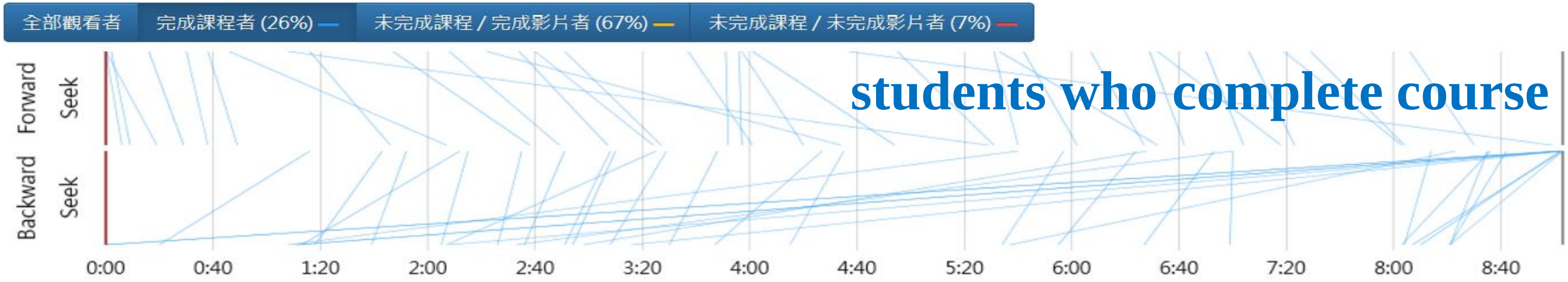


students who not complete course but complete video

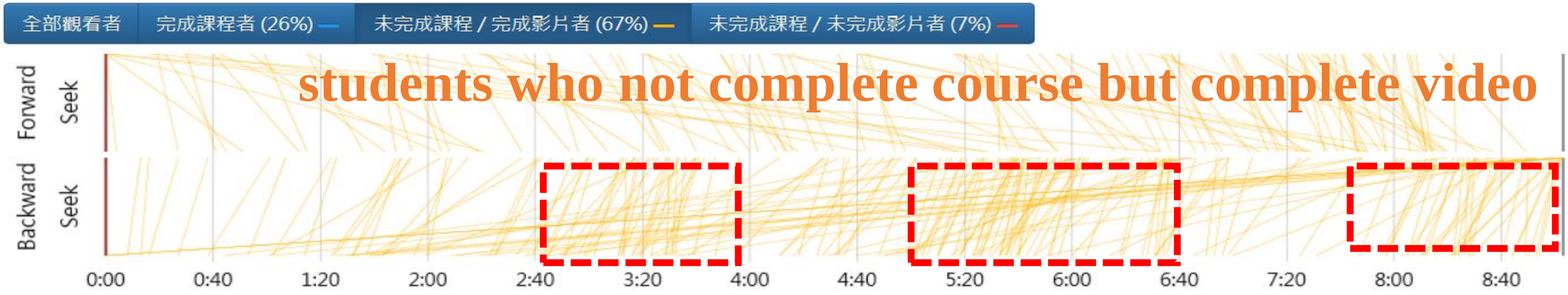
High density of **Backward** might due to lack of background knowledge



All students



students who complete course



students who not complete course but complete video

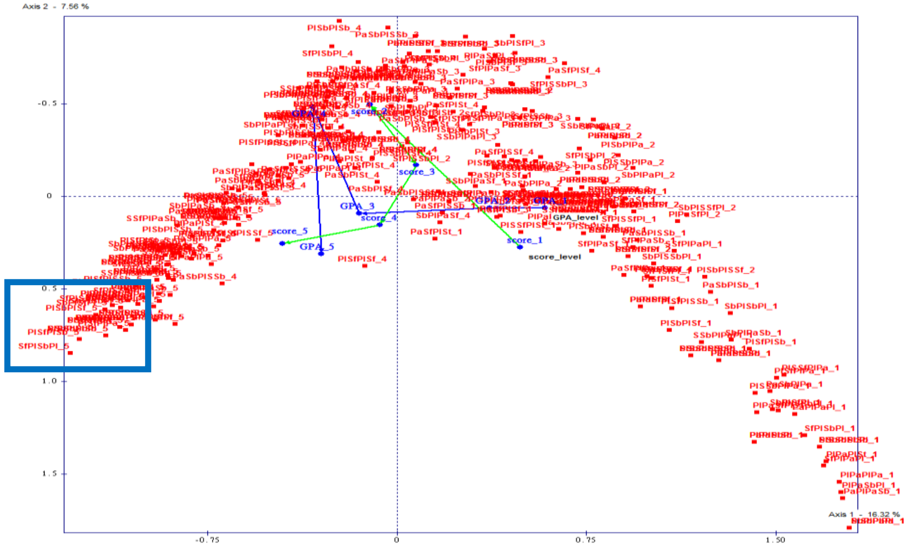
Correlation of viewing patterns and final grade

- RQ_5：
  - Can we find the viewing patterns that affect students' final grades?

- RQ_6：
  - Can we find the differences of viewing patterns between students with different grades (high and low)?

RQ_5：use MCA to find the viewing patterns that affect student’s final grades

- The patterns with more contribution to final grade are located at two sides

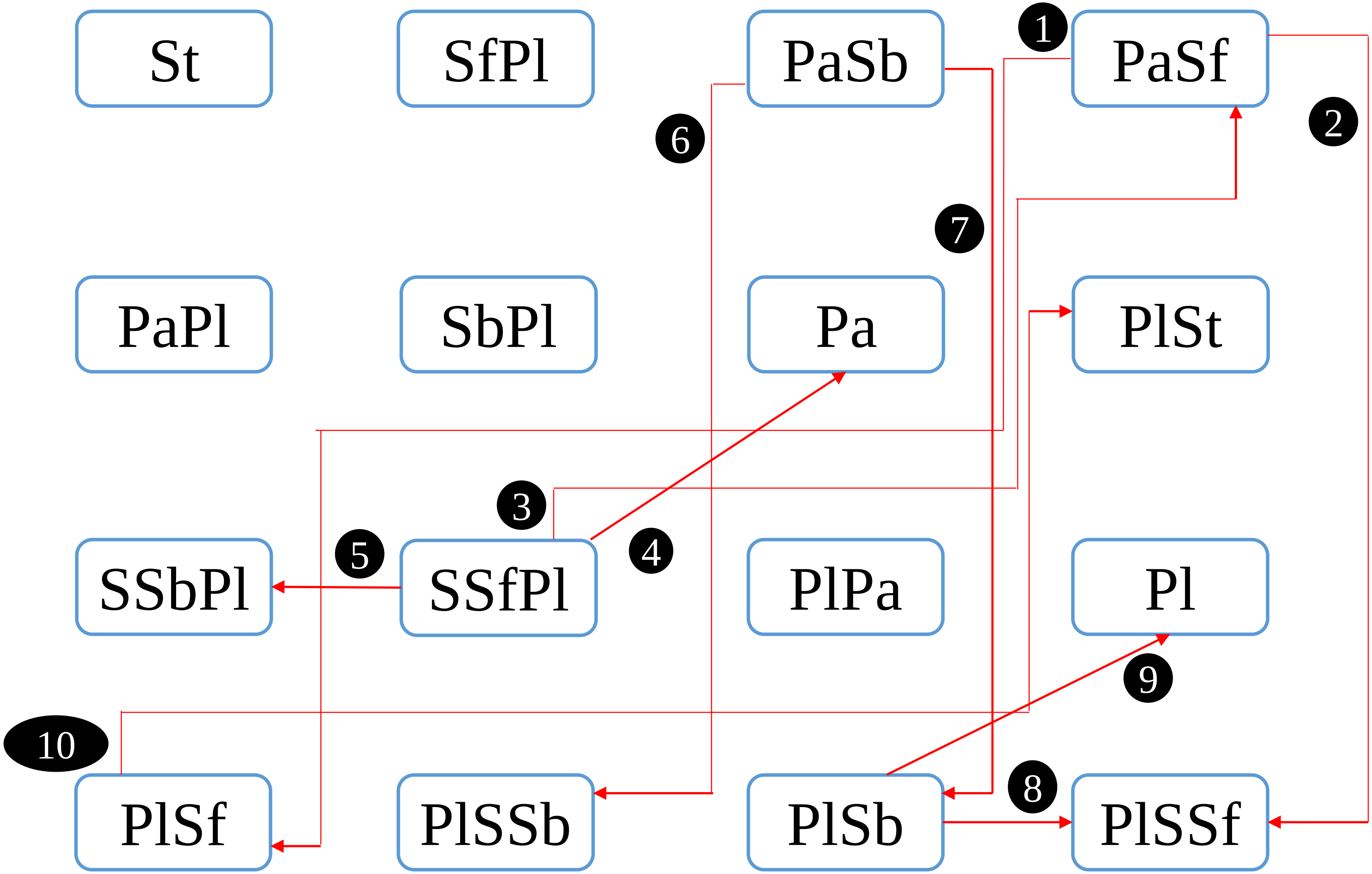


Pattern	Contribution	Viewing behavior	Behavior description	Types of learners
SfPlSbPl SfPlSSbPl	0.719 0.571	Forward, play, backward, play	Seek focus of content	Targeting learners
SbPlSfPl PlSbPlSf	0.666 0.597	Backward, play, forward, play	Confirmation of concept and seek focus	Comprehensive learners
SbPlSbPl PlSSbPlSb	0.668 0.668	Backward, play, backward, play	Repeated reconfirmation	Reflective learners

RQ_6： Use LSA to find the differences of viewing patterns between students with different grades

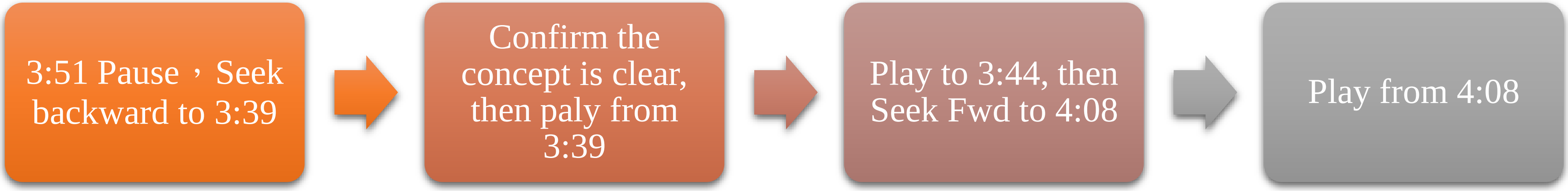
Behavioral transfer diagram of high grade

The red line indicates it is significant for high grade and insignificant for low grade



Viewing patterns of learners with high grade (Comprehensive learner)

	Sequence	z-score	Viewing pattern	Behavior description	Types of learners
1	PaSf ➔ PlSf	5.5762	Pause, Seek backward, play, and seek forward again	Confirm concept, and seek focus of content	Comprehensive learner

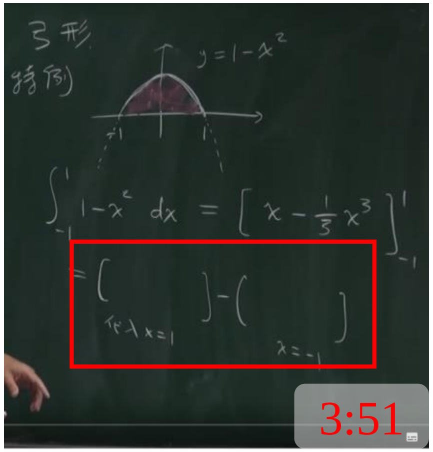


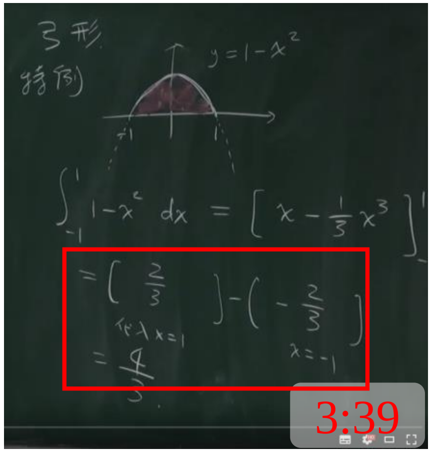
3:51 Pause，Seek backward to 3:39

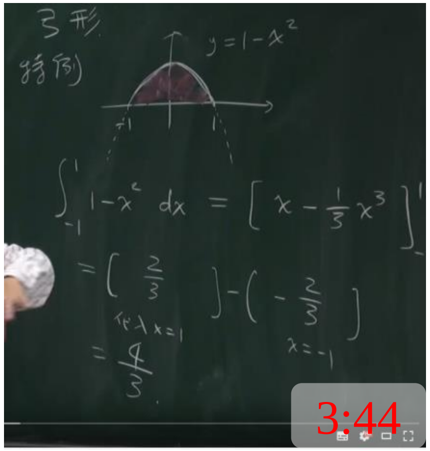
Confirm the concept is clear, then paly from 3:39

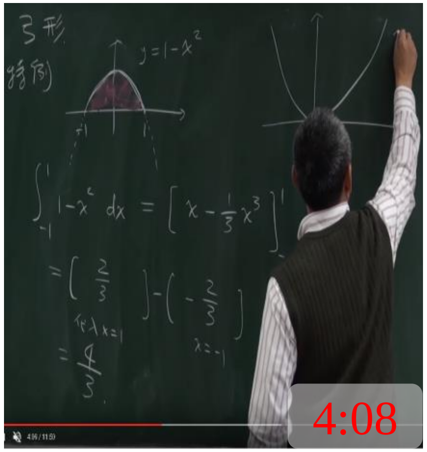
Play to 3:44, then Seek Fwd to 4:08

Play from 4:08



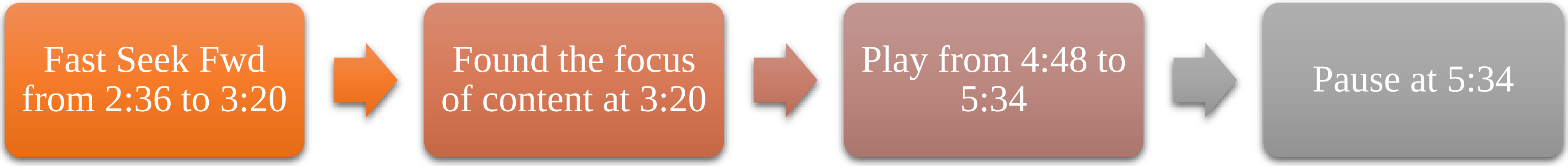




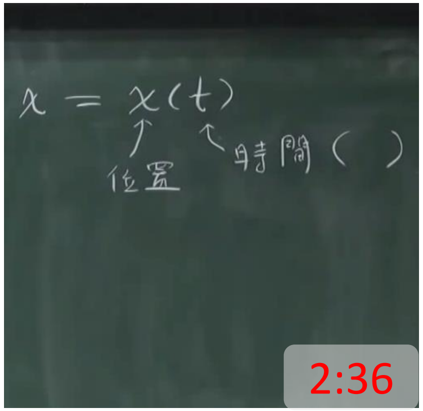


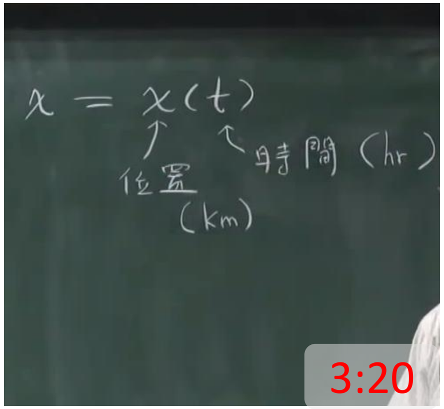
Viewing patterns of learners of high grade (Targeting learners)

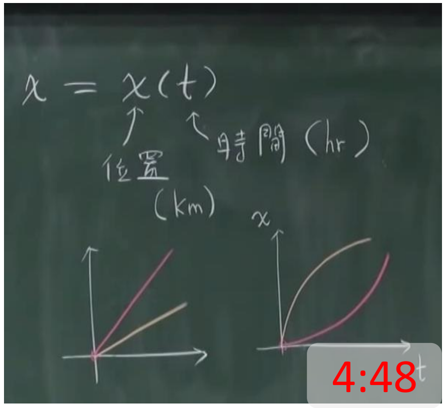
	Sequence	z- score	Viewing behavior	Behavior description	Types of learners
4	SSfPl ➔ Pa	2.6487	Fast Seek Fwd, play, then Pause	Seek focus of content	Targeting learners

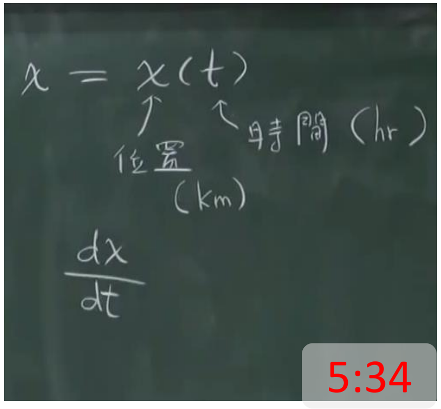


Fast Seek Fwd from 2:36 to 3:20 → Found the focus of content at 3:20 → Play from 4:48 to 5:34 → Pause at 5:34



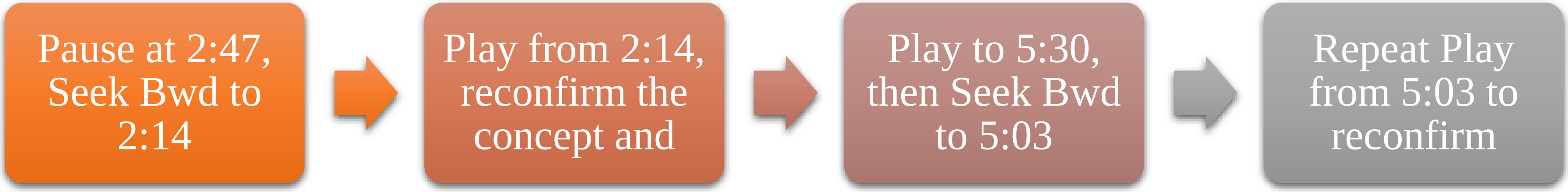






Viewing patterns of learners of high grade (Reflective learners)

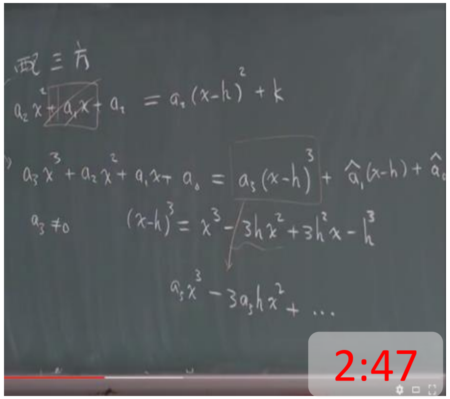
	Sequence	z- score	Viewing behavior	Behavior description	Types of learners
7	PaSb ➔PlSb	6.8895	Pause, Seek Bwd, then Play and Seek Bwd again	Repeated reconfirmation	Reflective learners

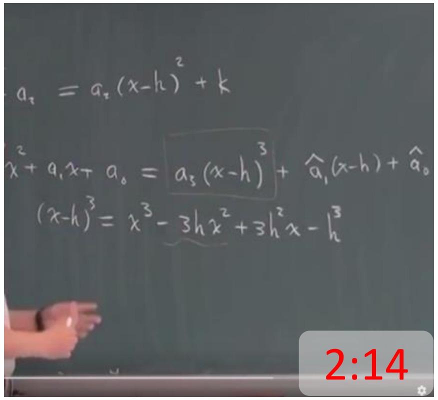

Pause at 2:47, Seek Bwd to 2:14

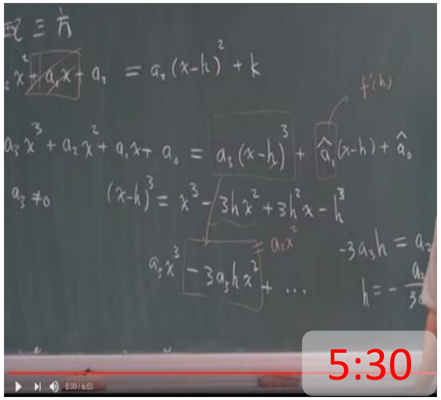
Play from 2:14, reconfirm the concept and

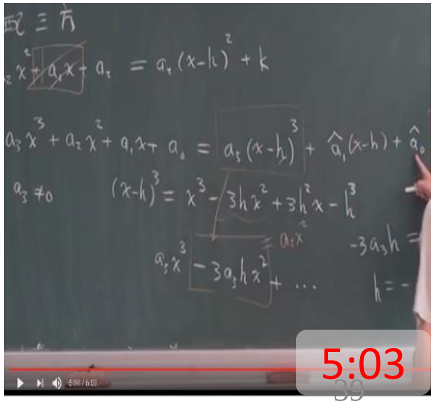
Play to 5:30, then Seek Bwd to 5:03

Repeat Play from 5:03 to reconfirm



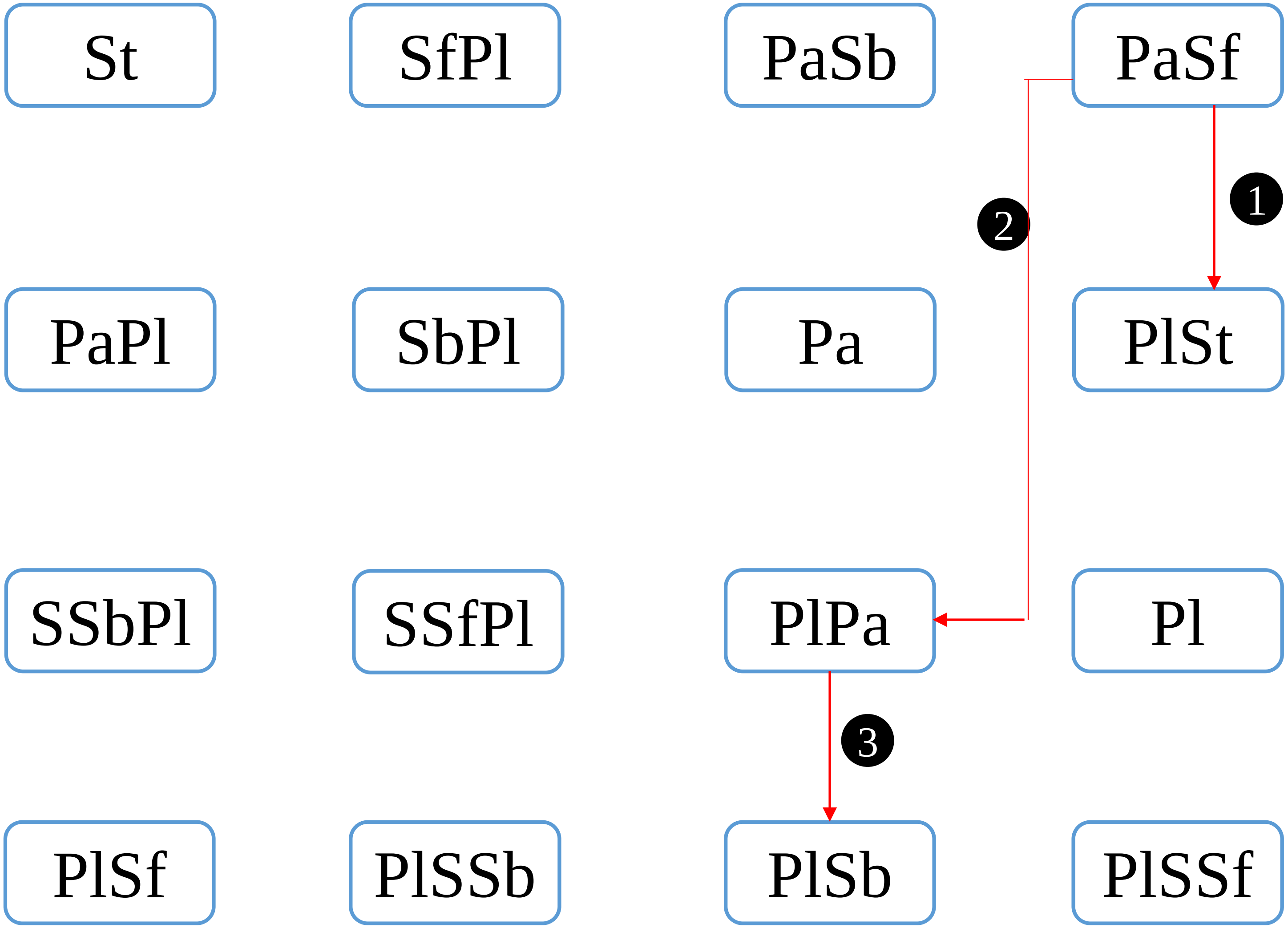






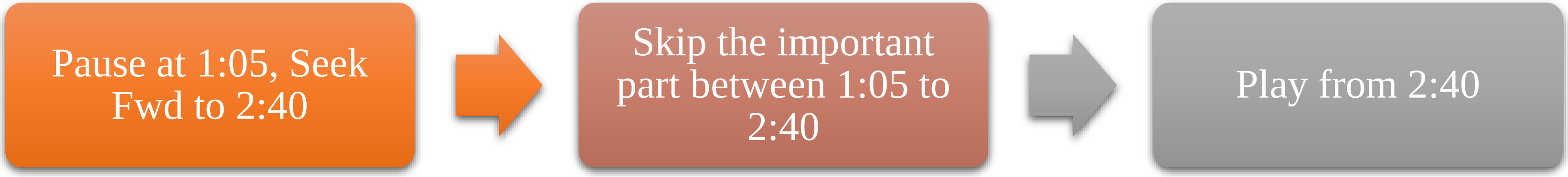
Behavioral transfer diagram of low grade

The red line indicates it is significant for low grade and insignificant for high grade



Viewing patterns of learners of low grade (Surfing learners)

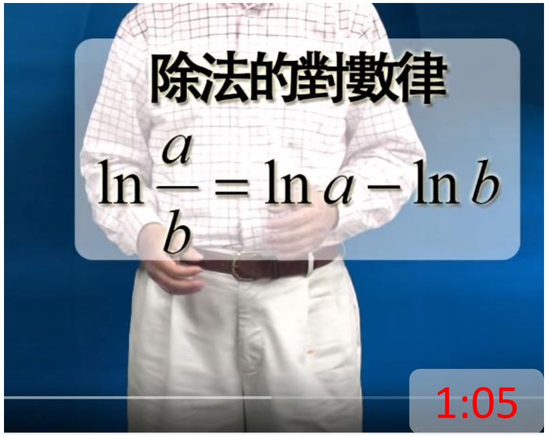
	Sequence	z- score	Viewing behavior	Behavior description	Types of learners
1	PaSf➔PlSt	4.6024	Pause, Seek Fwd, then Play and Stop	Skip most of video	Surfing learners

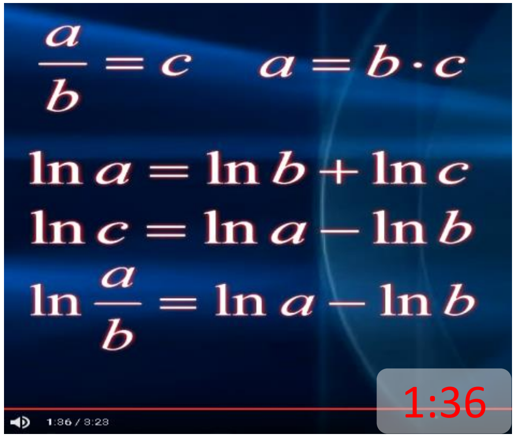


Pause at 1:05, Seek Fwd to 2:40

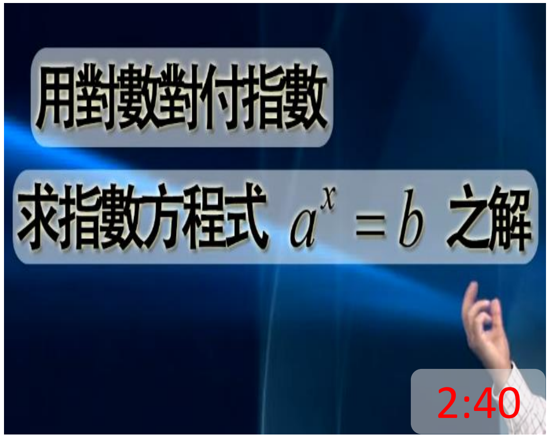
Skip the important part between 1:05 to 2:40

Play from 2:40





$\frac{a}{b}=c \quad a=b\cdot c$
$\ln a=\ln b+\ln c$
$\ln c=\ln a-\ln b$
$\ln\frac{a}{b}=\ln a-\ln b$
1:36 / 3:23
1:36



We identify 4 types of learner in Calculus 101 (knowing students' final grade)

Type name	Type description	Final Grade
Comprehensive	Confirmation of concept and seek focus Backward, play, forward, play (more activities)	high
Reflective	Repeated reconfirmation Backward, play, forward, play (more backward)	high
Targeting	Seek focus of content Forward, play, backward, play (more forward)	high
Surfing	Skip most of the video Forward, play, forward (less activities)	low

Work in progress of Study 3

- Knowing learners’ viewing patterns, how to provide timely intervention to improve their academic performance?

- Knowing learners’ struggles, how to provide timely intervention in order to improve MOOCs completion rate?

- How to provide timely intervention?
  - Recommendation systems (BookRoll + KURENAI)

Case study 4

Clustering analysis of MOOCs users' types and their learning activities

Part of an empirical study of Taiwan's MOOCs initiative

by Jeff, 健宏

Data collection of Taiwan’s MOOCs Initiative

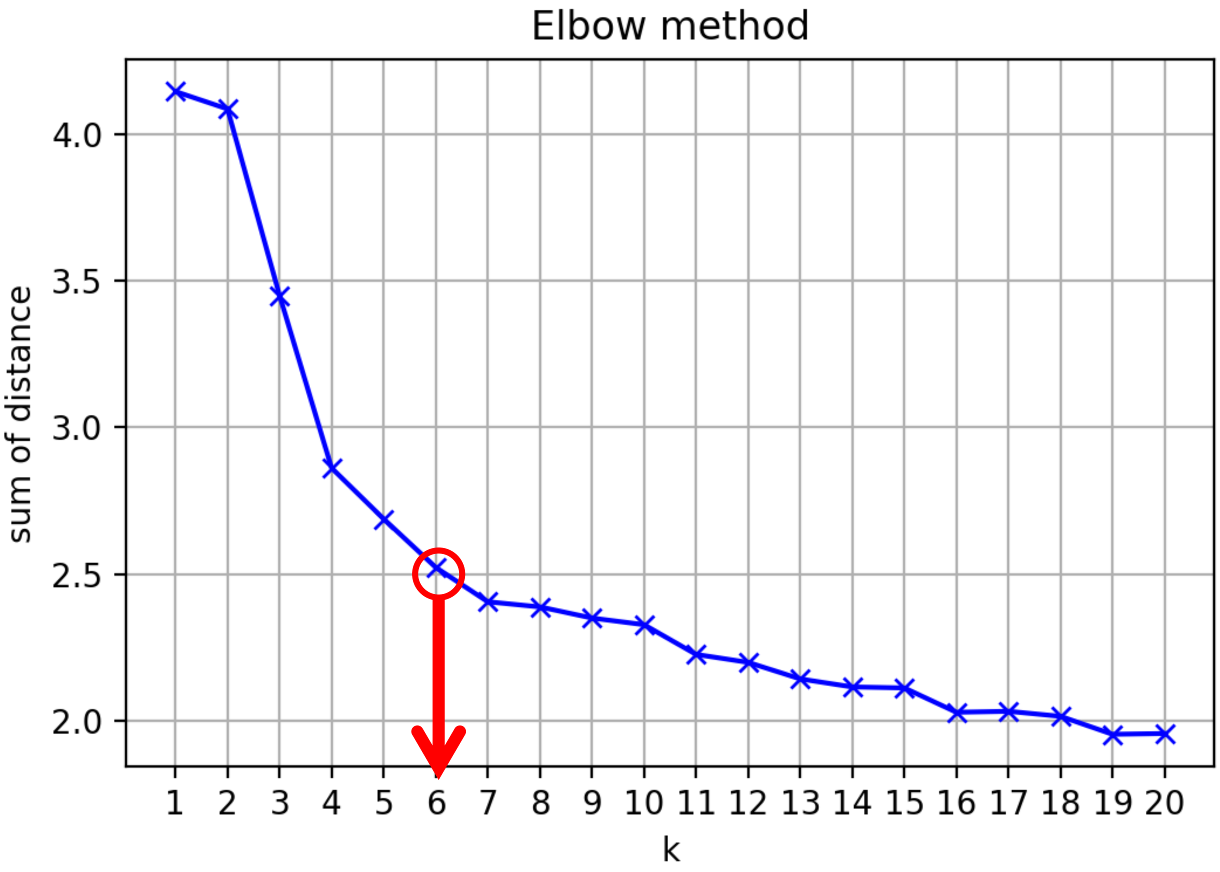
- Platforms
  - Open edX (based on MIT Open edX)
- Data collection period
  - 3 years：Sept. 2014 - Aug. 2017
  - Not knowing students’ final grade
- Courses: 590
- Registers: 32,120

Top 10 courses (out of 590) in terms of registers

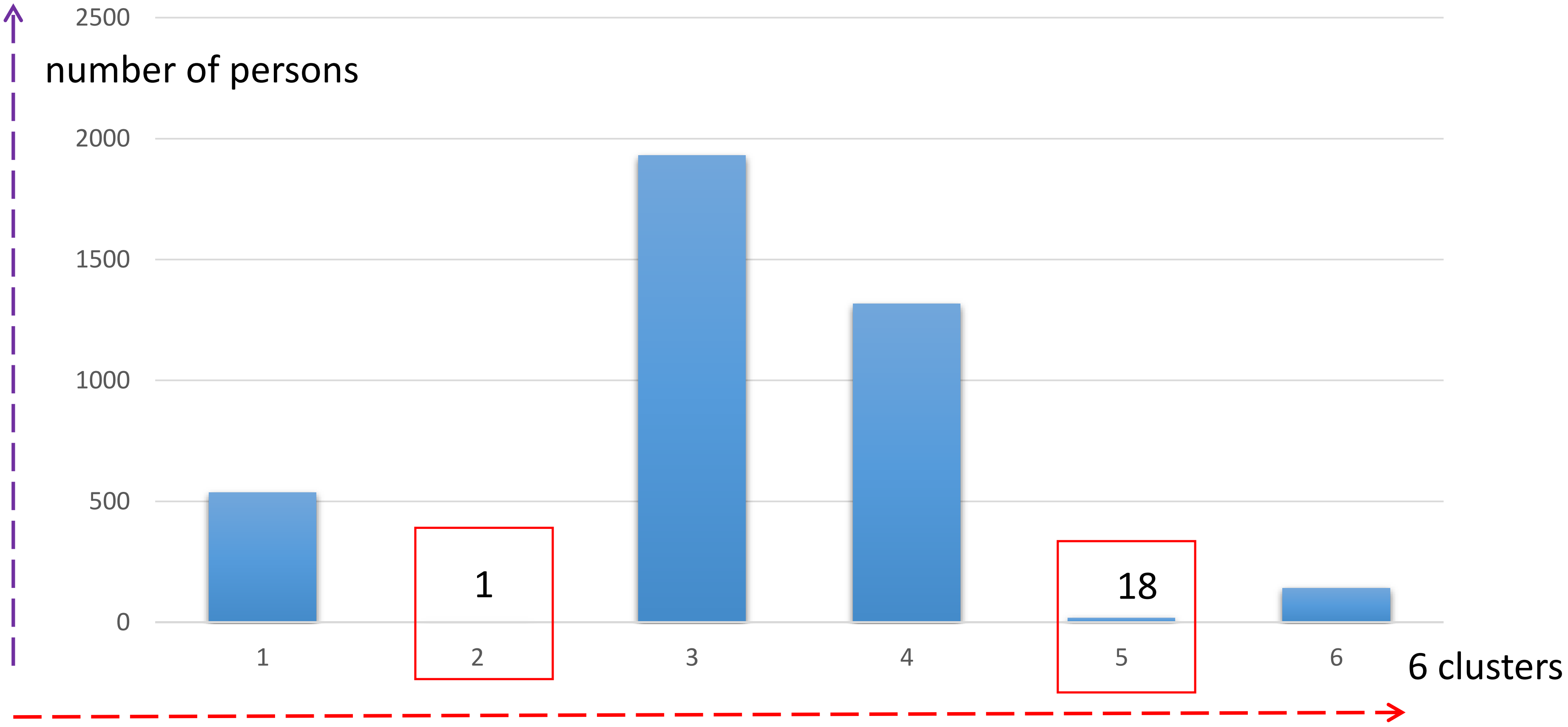
Course name	Number of registers	Course begin	Course end
105 年課文本位閱讀理解策略教學初階課程	3,802	2016/10/ 07	2017/8/10
讓老闆不得不重用你~正在崛起的「專案管理」_201410	1,310	2014/10/20	2015/4/30
大學普通物理實驗─手作坊	1,289	2014/12/1	2015/4/12
台灣傳統糕餅文化與製作、創新	1,174	2015/6/11	2016/1/26
從車庫到金庫─看見台灣企業生命力	1,013	2014/10/26	2015/7/12
軟體設計 I- 物件導向設計	750	2016/8/1	2020/8/31
巧克力製作	727	2015/12/15	--
軟體設計 III- 設計樣式	680	2016/8/25	2020/731
2015 K-12能源科技教育種子教師培訓 (初階)	608	2015/5/14	2015/6/15
讓老闆不得不重用你~正在崛起的「專案管理」_104	610	2015/12/1	--

Clustering analysis (K-means)

Feature Name
video_num_days
num_watched
num_complete
complete_rate
seek_video_sum
pause_video_sum
stop_video_sum
video_forward_seek_sum
video_backward_seek_sum
video_pause_sum
video_play_sum
video_stop_sum
video_pause_forward_seek_sum
video_pause_backward_seek_su m
video_stop_backward_seek_sum
video_events_avg

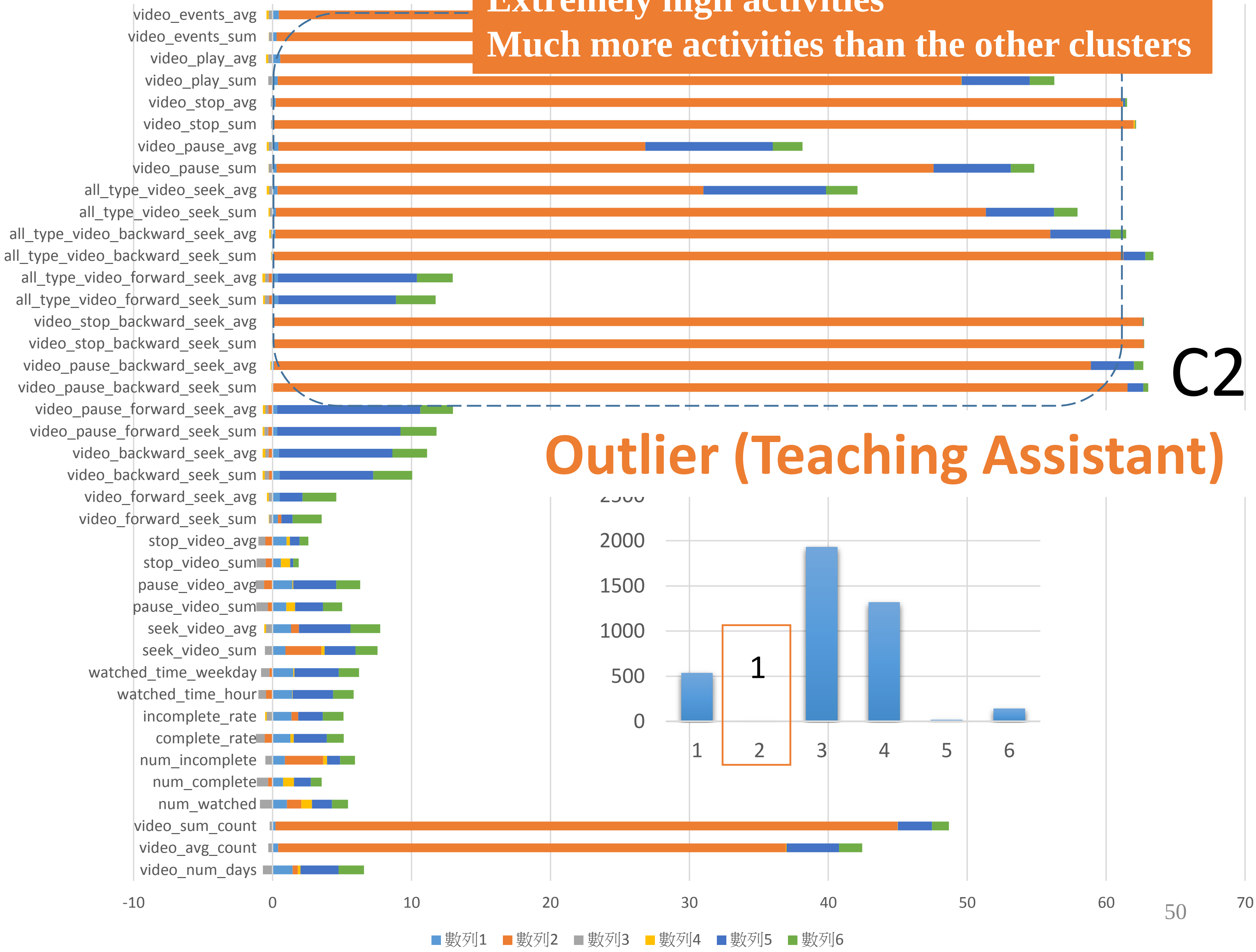


Group learners into 6 clusters.



Description of 6 types, based on # of activities per learner

Cluster	Learners	Description	Learns’ type
2	1	Extremely high activities Much more activities than the others	Outliers
5	18	Confirmation of concept and seek focus Backward, play, forward, play (more activities)	Comprehensive
6	141	Repeated reconfirmation Backward, play, forward, play (more backward)	Reflective
1	536	Seek focus of content Forward, play, backward, play (more forward)	Targeting
4	1,317	Skip most of the video Few Forward, play, forward (less activities)	Surfing
3	1,930	Little activities	Disengaged

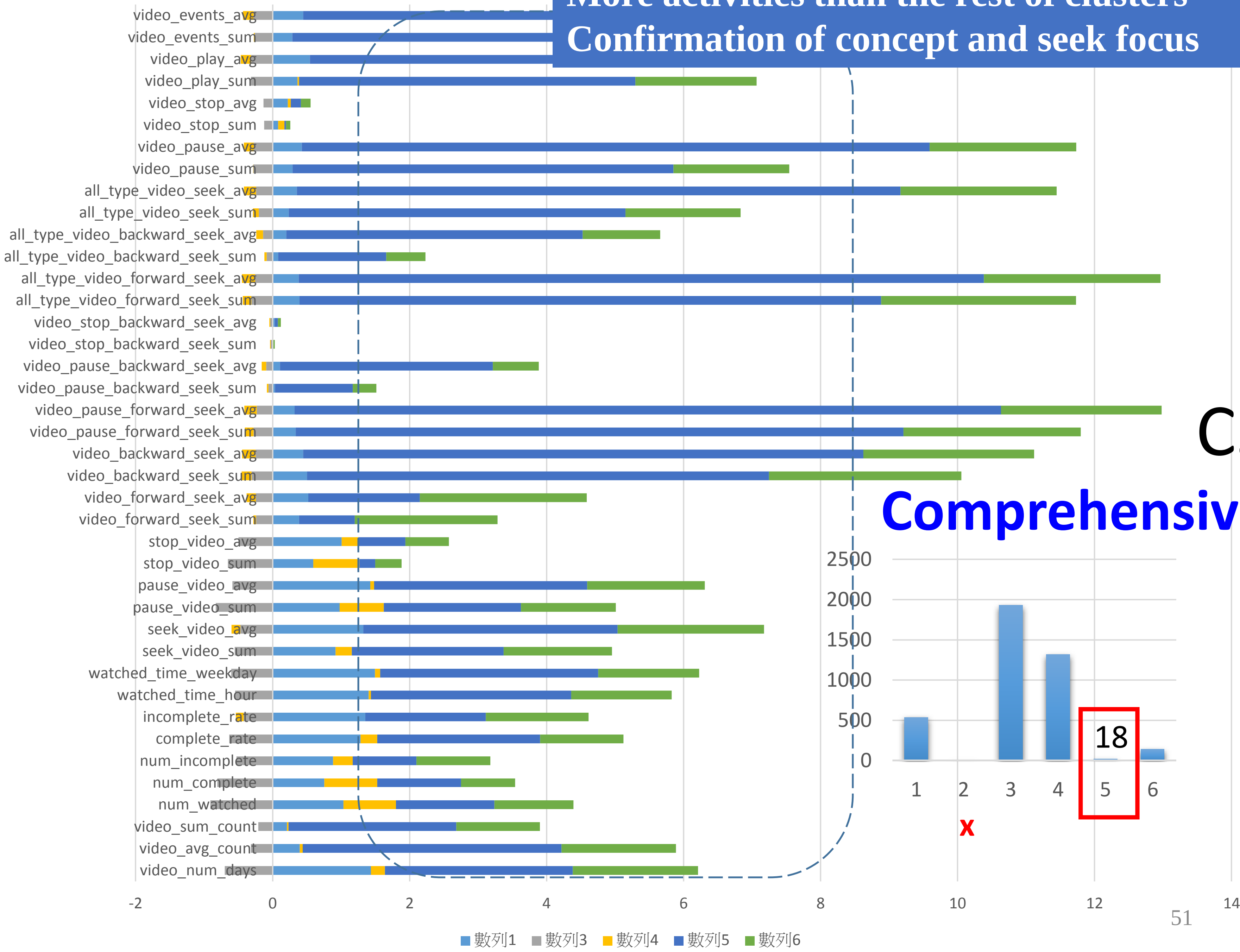


Extremely high activities
Much more activities than the other clusters

C2

Outlier (Teaching Assistant)

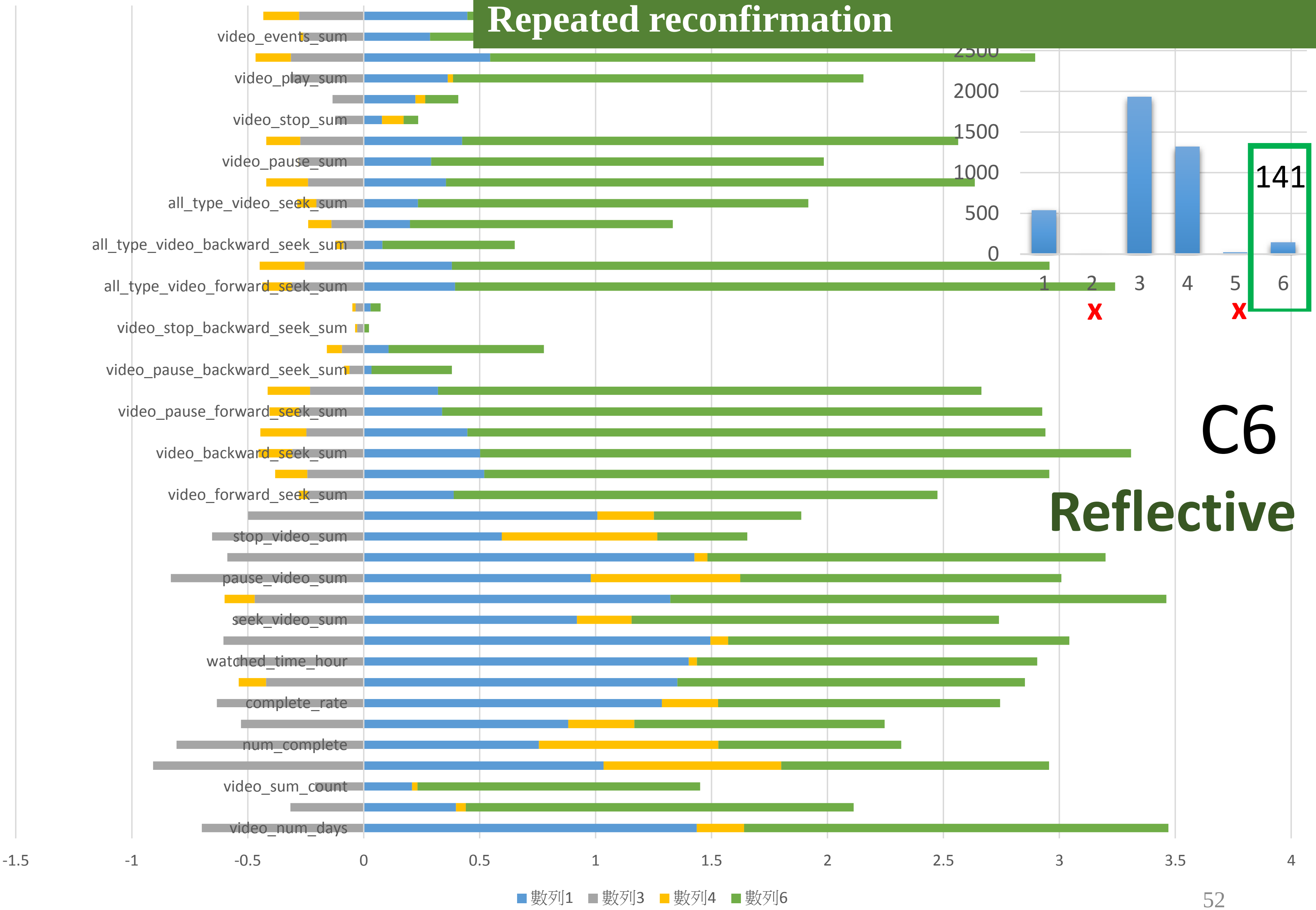
More activities than the rest of clusters
Confirmation of concept and seek focus



C5

Comprehensive

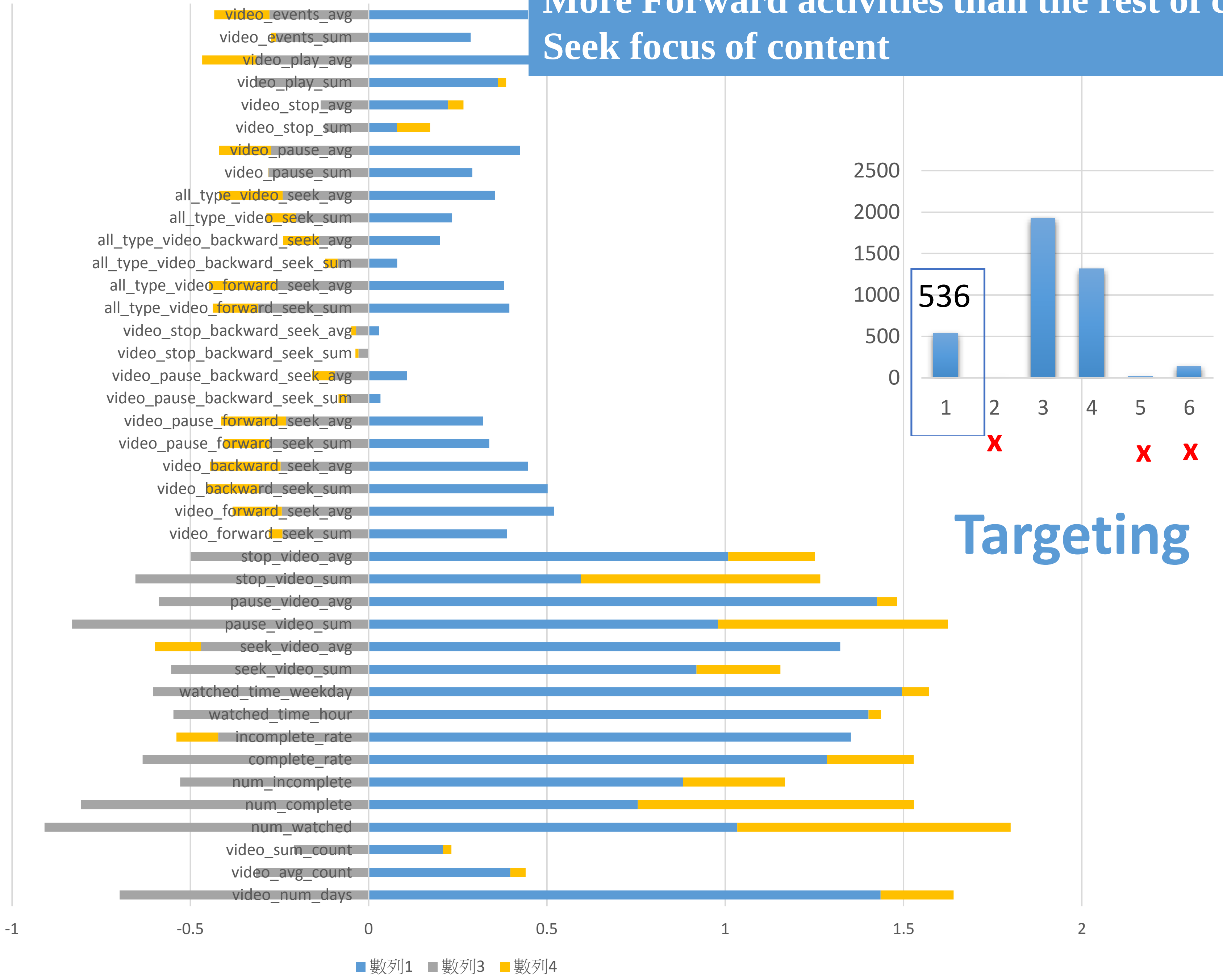
More backward activities than the rest of clusters
Repeated reconfirmation



C6

Reflective

More Forward activities than the rest of clusters
Seek focus of content



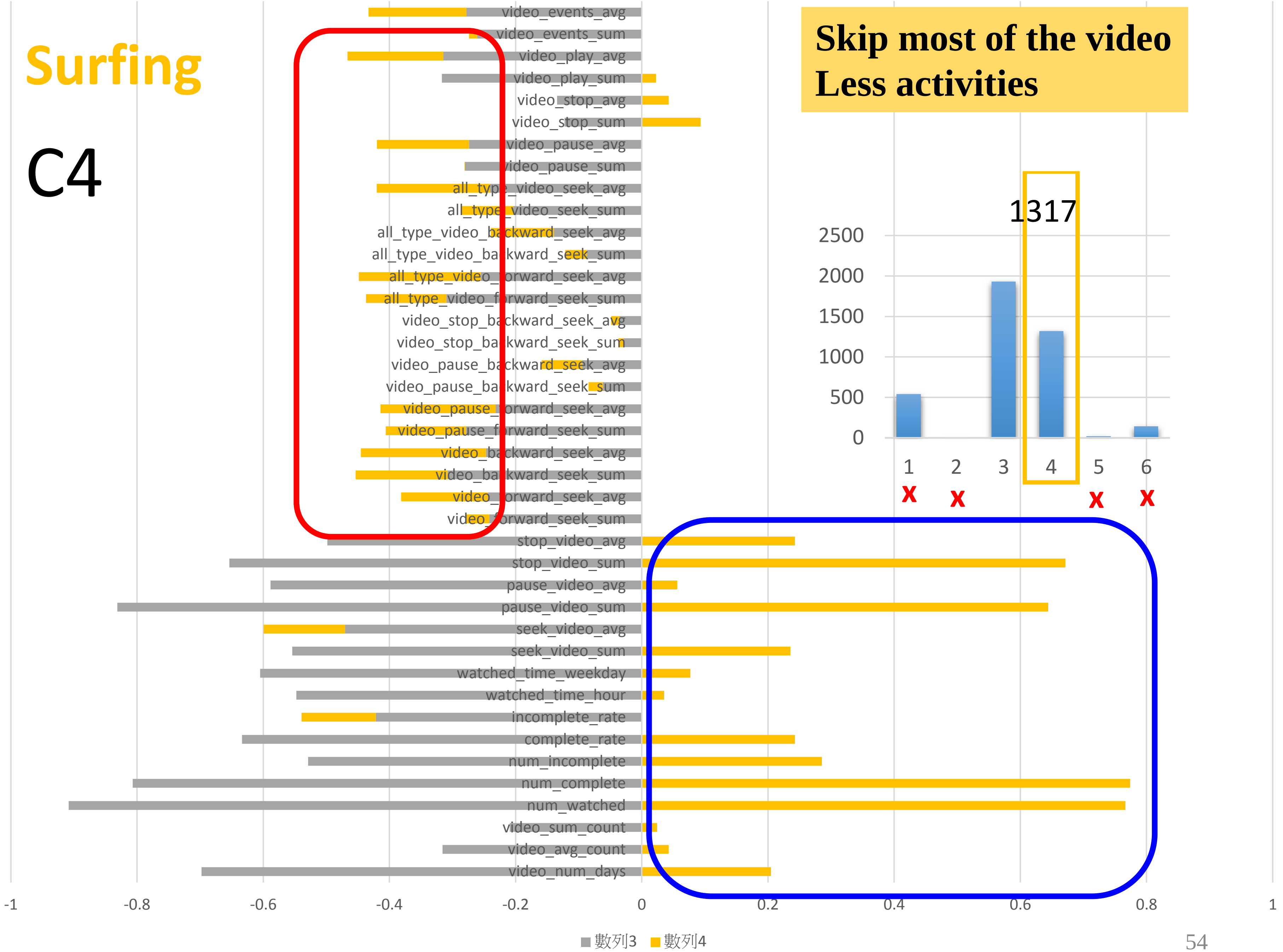
C1

Targeting

Surfing

C4

Skip most of the video
Less activities



Work in progress

- Knowing MOOCs learners' types
  - Find the correlation of learners' types and their Self-regulated learning capability.
  - Improve learning with SRL.

Outline

- Part I：Analyzing students’ learning activities
    - Dashboard, graph, charts
    - Visualization of analysis results

- Part II：Predicting at-risk students who might
    - Drop out, withdraw, fail,
    - low score, low grade

Machine learning for Prediction

- Supervised learning (with labels)
  - Classification (true/false; yes/no)
  - Regression (predict a value)
- Unsupervised learning (without labels)
  - Clustering (group formation)
- Reinforcement learning (learn from past experiences)
  - Markov chains
- Transfer learning
  - Portability of prediction model
  - From source domain to target domain

Algorithms for predicting students' academic performance

- Classification
  - pass/fail, dropout yes/no
- Regression
  - Score, grade
- Clustering
  - Group performance

Examples of classification algorithms

- Support Vector Machine
- Logistic Regression
- Decision Tree
- Random Forest
- Neural Network
- Gaussian Naive Bayes(GaNB)

Examples of regression algorithms

- Classification And Regression Tree (CART)
- Quantile Regression
- Robust Regression
- Support Vector Regression (SVR)
- Multiple Linear Regression (MLR)
- Principle Component Regression (PCR)
  - MLR plus PCA (Principal component analysis)

Case study 5

Early prediction of students' academic performance in blended learning

Part of an empirical study of Taiwan's MOOCs initiative

by Anna, 舜澤

Early perdition of students’ final scores

RQ_7：Comparing six regression algorithms, which is the best algorithm for predicting students’ academic performance (final scores)?

RQ_8：How to improve the performance of prediction?

RQ_9：Can we provide early prediction of students’ academic performance (final scores)? And how early it is?

RQ_7：Which is the best regression algorithm for predicting students' final scores ?

- Data source (Calculus 101)
  - Blended learning (59 students)
  - Knowing students' final scores

- Platforms
  - Open edX (based on MIT Open edX open source)
  - Maple TA (for Calculus exercise & assessment)
  - Insight^{+} (learning analytics)

Model evaluation (performance metrics)

- Predictive Mean squared error (pMSE)
- PCR outperform the rest of 5 algorithms

Algorithms	**pMSE**
MLR	448.754
CART	402.886
Quantile	794.252
Robust	1157.176
SVR	385.995
PCR	188.628

RQ_8：How to improve the performance of prediction?

- Training with various data sets
    - Full dataset vs. sub dataset

- Remove outliers using influence points
    - Locate influence points with Cook’s distance and DFFITS
    - Identify influence points as outliers

Training with various data sets

- Accumulated dataset
  - W1-6: week 1 ~ week 6 (1/3 of semester)
  - W1-12: week 1 ~ week 12 (2/3 of semester)
  - W1-18: week 1 ~ week 18 (full semester)

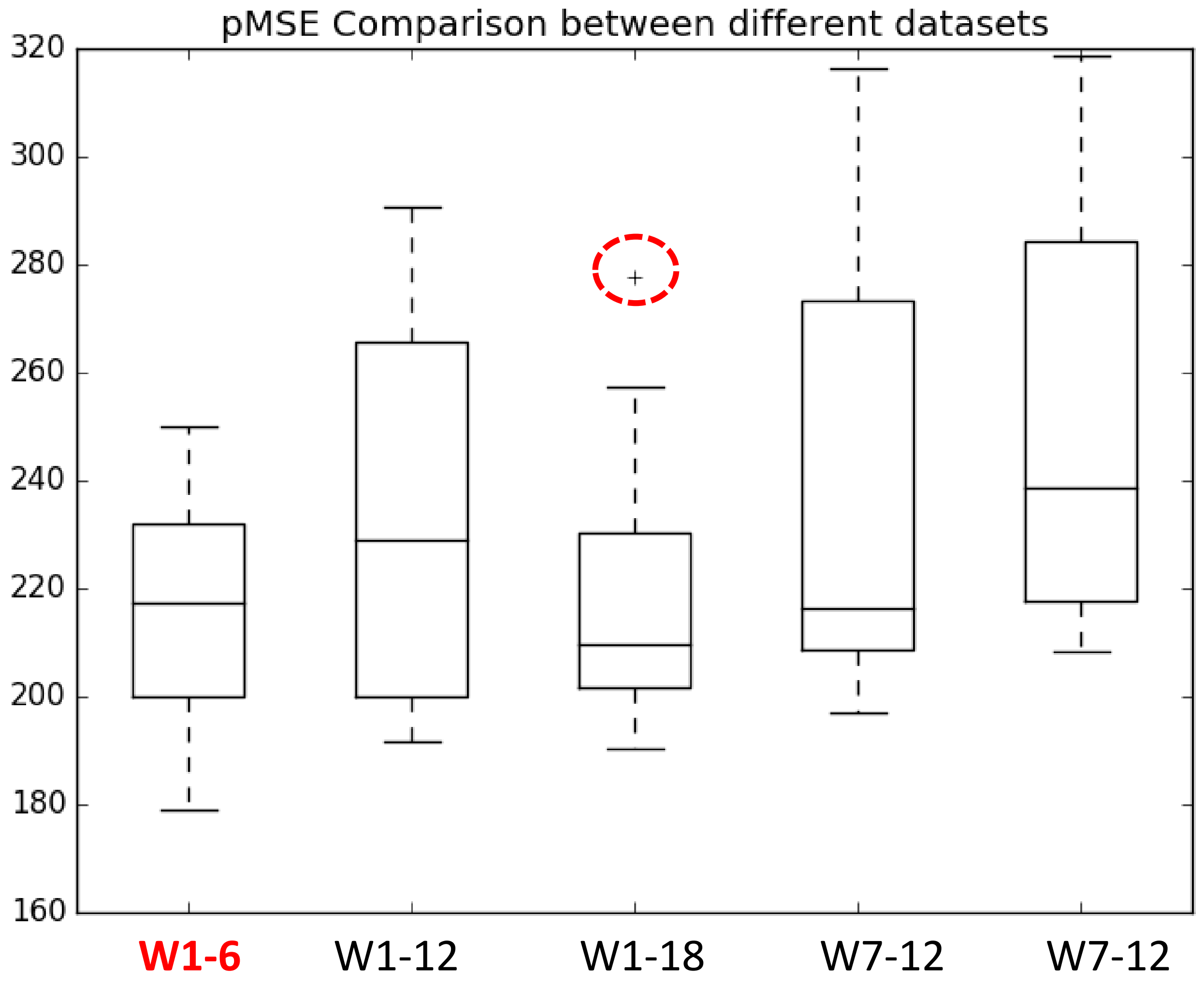
- Duration dataset
  - W7-12: week 7 ~ week 12 (middle of semester)
  - W12-18: week 13 ~ week 18 (last 1/3 of semester)

Comparing pMSE of different data sets

Dataset	pMSE
week1-week6 (W1-6)	232.1524
week1-week12 (W1-12)	242.284
week1-week18 (W1-18)	235.3709
week7-week12 (W7-12)	244.0642
week13-week18 (W13-18)	254.37

Comparing Box-plot of pMSE of different datasets

W1-6 is more stable than W1-18



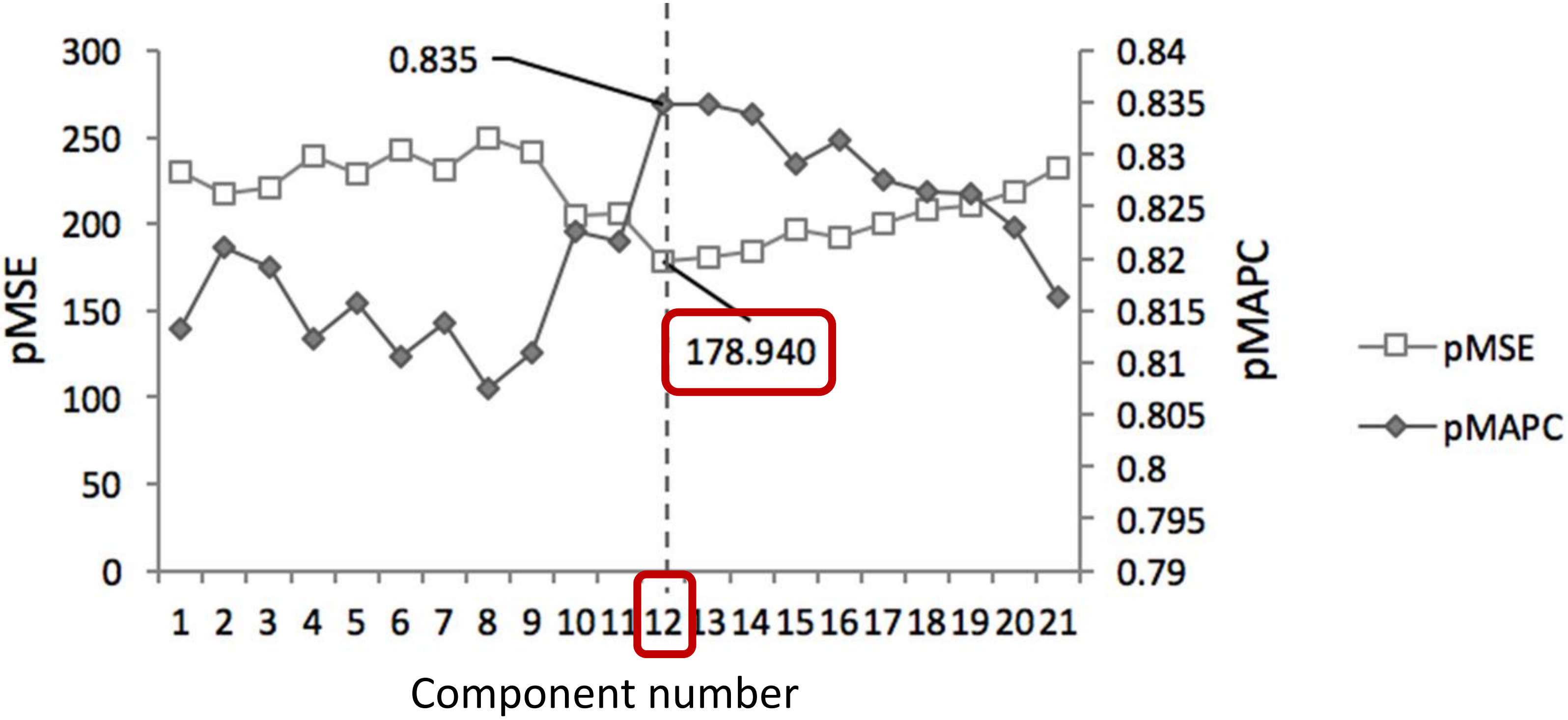
Comparing pMSE after removing outliers

Locate influence points with Cook’s distance and DFFITS

method	Data set	Influence points (outliers)	Features left	pMSE	pMSE Removing outliers
Cook’s	W1-W6	8	51	188.628	148.3545
DFFITS		9	50		184.0055

RQ_9：Can we provide early prediction of students' final scores? And how early it is?

- Dataset W1-6 with 12 components has the best pMSE
- Early prediction as early as at 6th week.



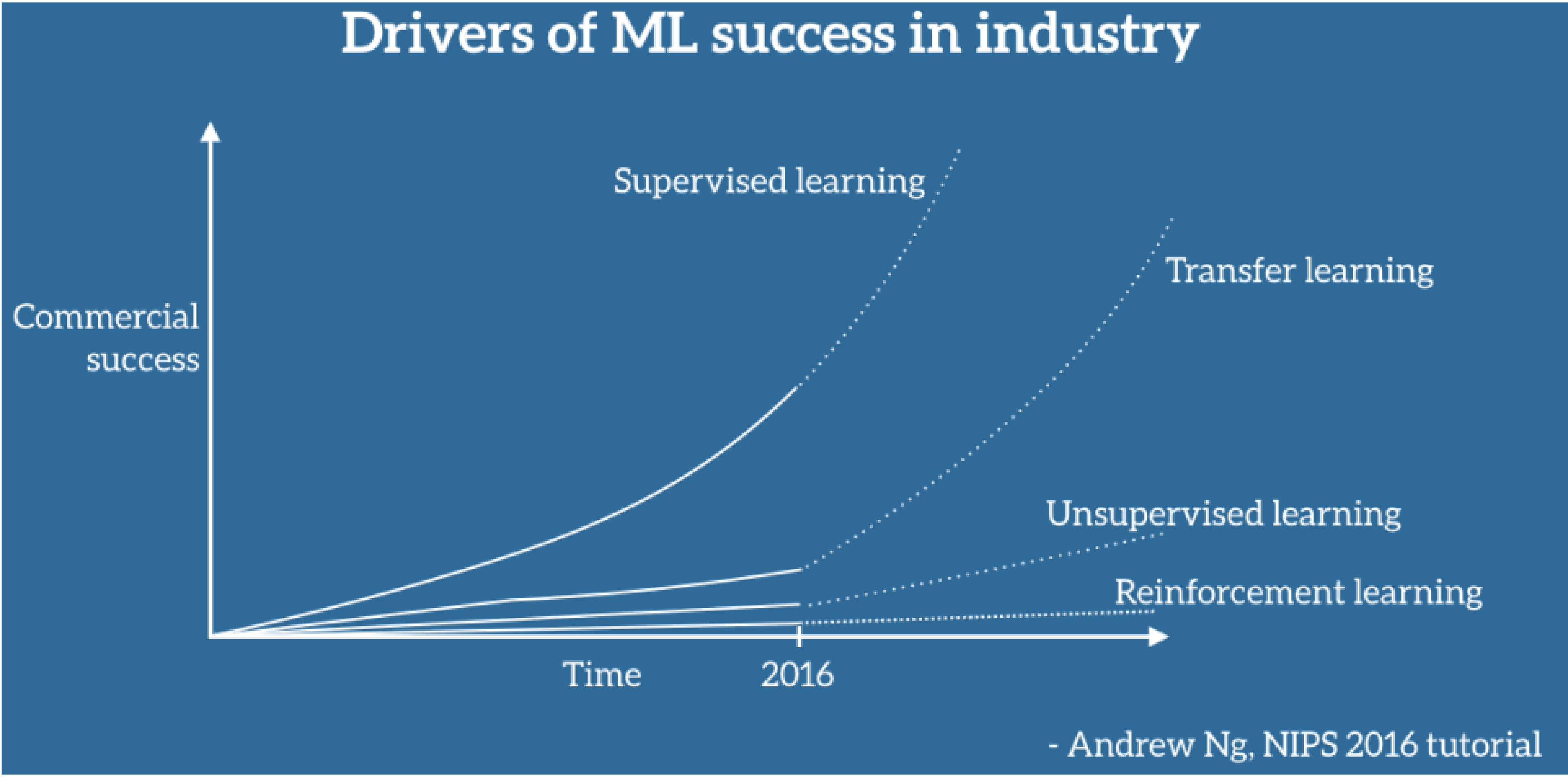
Work in progress of Study 5

- Knowing at-risk students
  - Applying CNN (Convolution Neural Network) to recommend learning resources based on concept map.

What next - Portability of prediction model

- Can we apply calculus experience (prediction model) to any courses, such as CS, Physics, Chemistry, Biology, Psychology, Philosophy,…, etc.

Our next step - Transfer learning



Drivers of ML success in industry

- Andrew Ng, NIPS 2016 tutorial

Future research

- Transfer learning between different data sets, same domain
  - Calculus domain:
    - Between College & Senior High
    - Between different colleges
    - Between different media (video, eBook, lectures)
  - Program domain: Between Python & others

- Transfer learning from source domain to target domain
  - From Calculus to CS domain
  - From programming to more CS courses

Thanks very much

Stephen J.H. Yang

楊鎮華