

Towards Reading Session-based Indicators in Educational Reading Analytics

Madjid SADALLAH

CERIST, Algeria

Benoît ENCELLE

University of Lyon, France

Azze-Eddine MAREDJ

CERIST, Algeria

Yannick PRIÉ

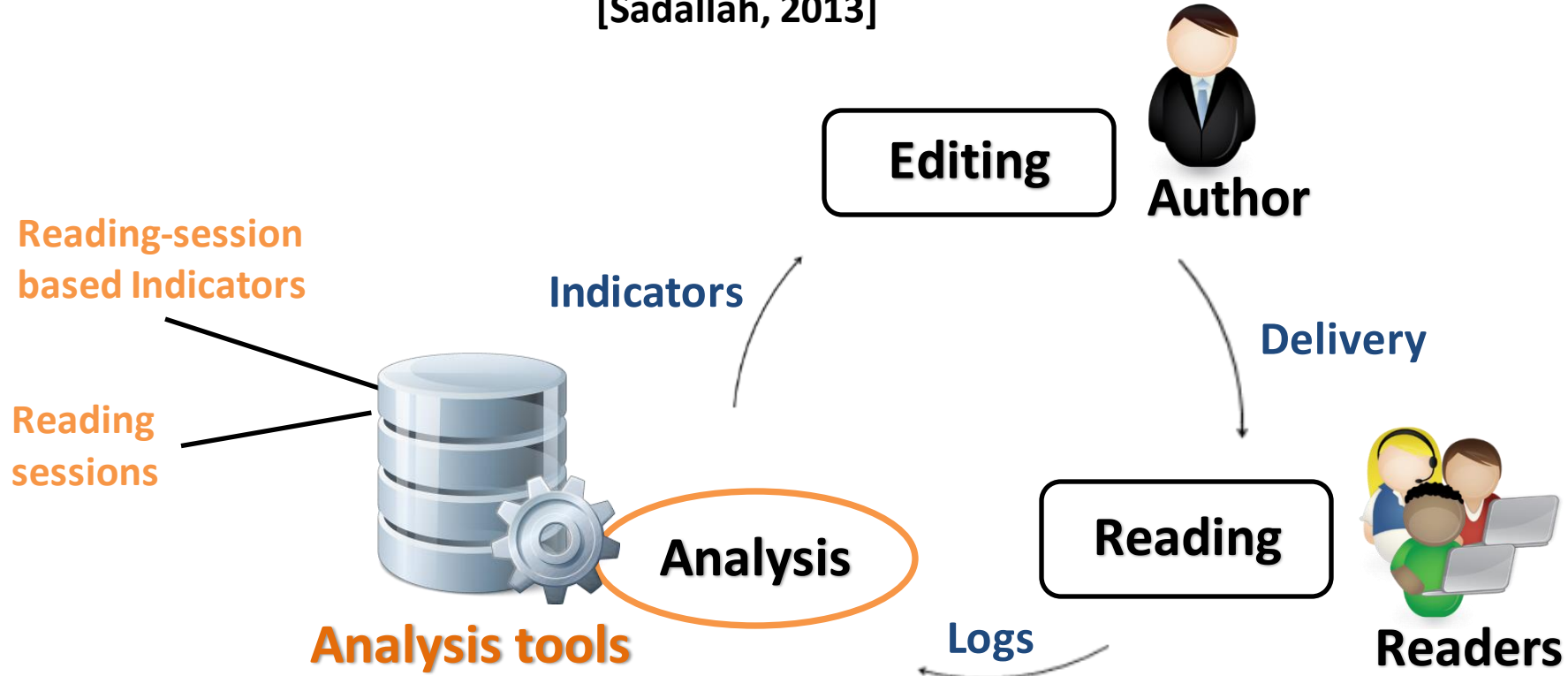
University of Nantes, France



EC-TEL 2015 - Toledo, Spain - 2015/09/17

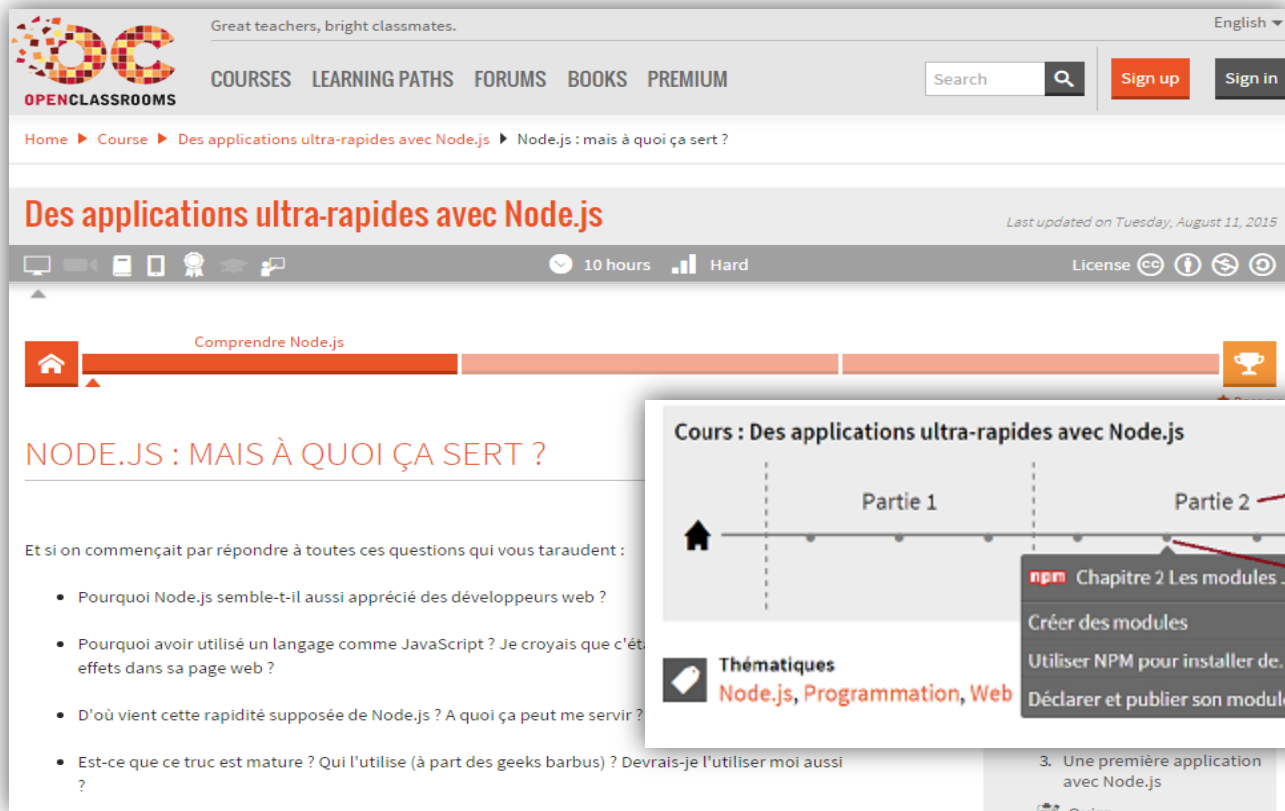
Usage-based document reengineering

[Sadallah, 2013]



Target: User, Course, page,...
Based on the events sequence

Platform & data

A screenshot of the OpenClassrooms website. The top navigation bar includes the OC logo, the tagline 'Great teachers, bright classmates.', a language selector set to 'English', and links for 'COURSES', 'LEARNING PATHS', 'FORUMS', 'BOOKS', and 'PREMIUM'. A search bar and 'Sign up'/'Sign in' buttons are also present. The main content area shows a course titled 'Des applications ultra-rapides avec Node.js' with a progress bar and a list of chapters. The first chapter, 'NODE.JS : MAIS À QUOI ÇA SERT ?', is expanded, showing a list of questions and a progress bar. A tooltip is visible over the second chapter, 'Chapitre 2 Les modules...', showing its subchapters: 'Créer des modules', 'Utiliser NPM pour installer de...', and 'Déclarer et publier son module'.

ors/month)

Course_part

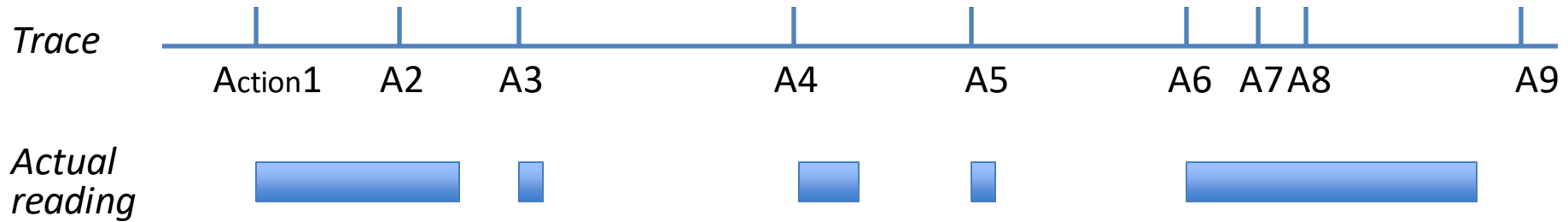
Chapter

Subchapter

Outline

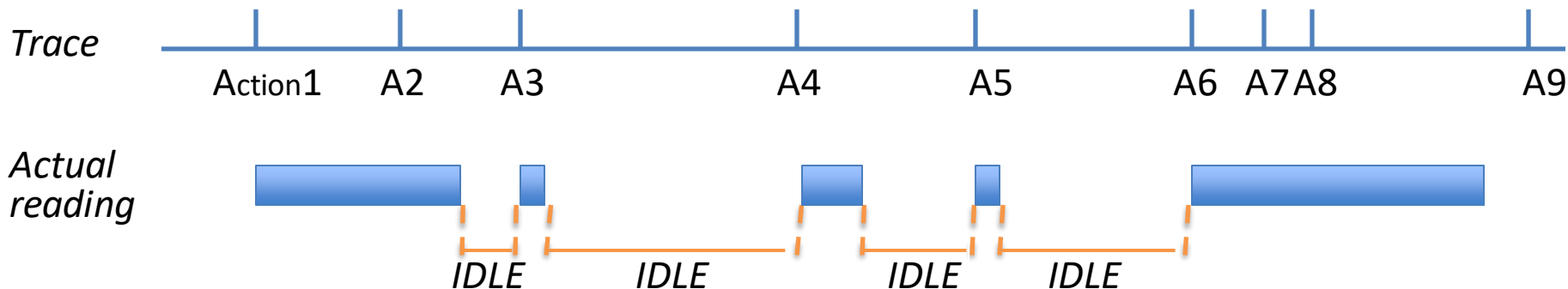
- 1. Time-based analysis of reading in eLearning**
2. A proposal for detecting reading sessions
3. Towards Reading Session-based Indicators

Why time-based indicators?



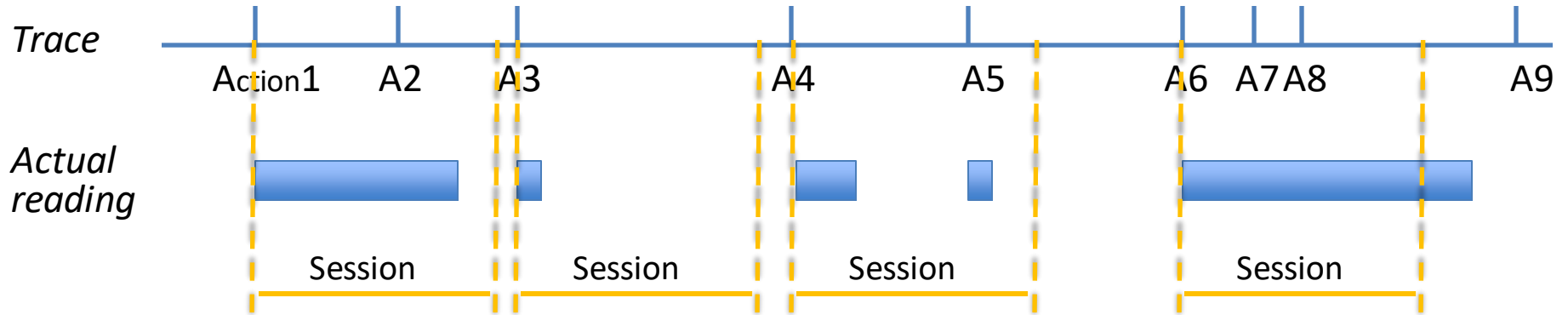
- More "accurate" estimate of student learning [Kovanovic, 2015]
- Best reflect and predict user behavior and environment over time [Hofmann, 2006]

Issue in time-based analysis



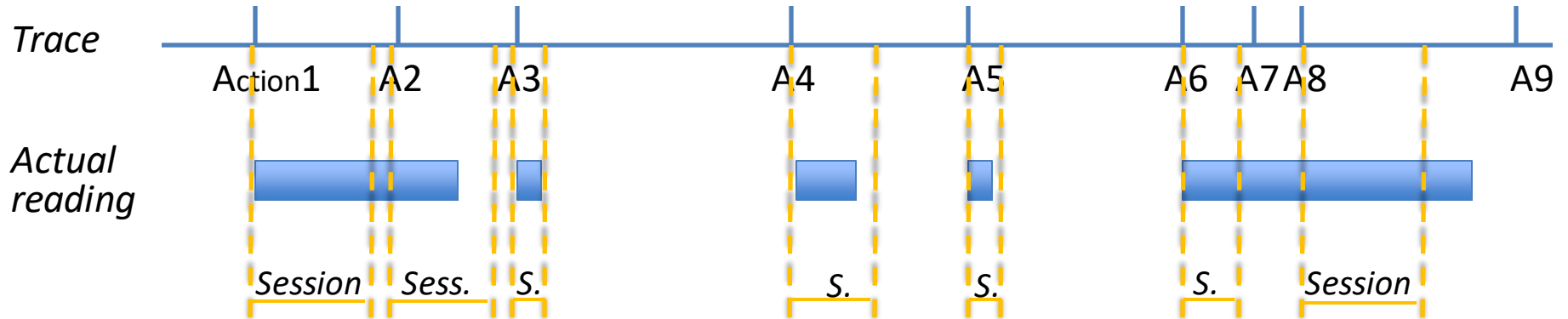
- Activity in time: active and inactive periods
- Session = active periods
 - Not logged
 - No time-on-task data to reconstruct the real sessions
 - Web Usage Mining methods for approaching them

Session duration threshold method



- Limit on the total duration of the session: 30 min
 - But: cuts continuous activities and merges separate short ones including potential in-between inactive periods

Page-stay threshold method



- Limit on the page-stay time with a predefined threshold: 10 min
 - But: some pages may be read faster or slower

Outline

1. Time-based analysis of reading in eLearning
- 2. A proposal for detecting reading sessions**
3. Towards Reading Session-based Indicators

Requirements & proposals

- Each course/course part has its inner-complexity
 - Limit sessions by a threshold on page-stay
 - Each page has its own threshold value
- Reading depends on its context (courses, pages, readers) which changes over time
 - Thresholds grounded from users logs and updated with new incoming data

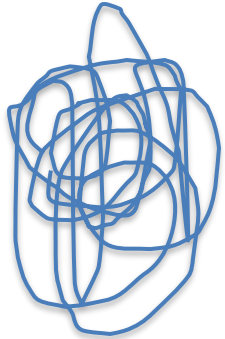
An algorithm for extracting reading sessions

**Pre-process
Data**






*Calculate
Tresholds*

*Delimit
Reading Sessions*

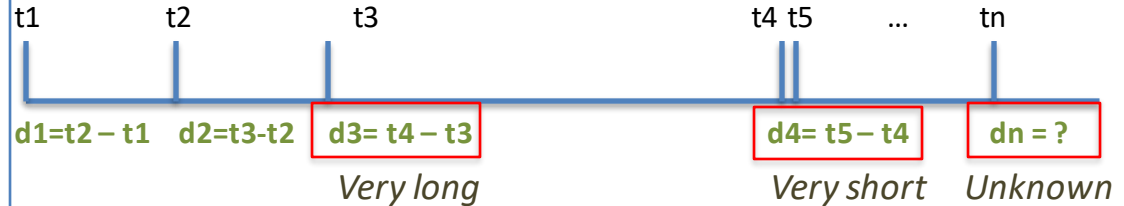
Raw data



Identify users

User 1 
User 2 
User 3 
User 4 
...
User n 

Compute durations



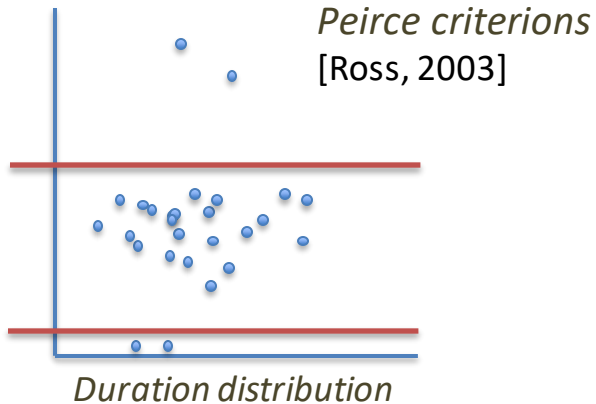
An algorithm for extracting reading sessions

Pre-process
Data

Calculate
Thresholds

*Delimit
Reading Sessions*

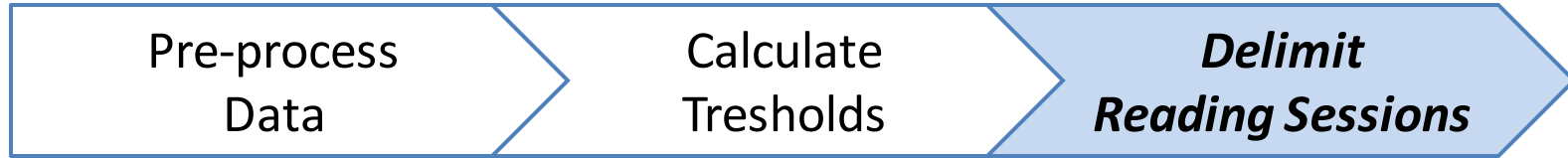
Eliminate outliers



Compute thresholds values

Parts	Durations	Threshold
Part 1	d11, d12, d13,.....	=MAX(d11,...)
Part 2	d21, d22, d23,.....	=MAX(d21,...)
...
Part n	dn1, dn2, dn3,.....	=MAX(dn1,...)

An algorithm for extracting reading sessions



Deal with unknown durations



Delimit sessions



Session ends on an action if:

Action duration > time threshold of the read part

Overview of the constructed reading sessions

	# reading sessions	#RS per user (median)	RS duration (median)	RS Size (median, in parts count)
Nodejs	13 431	3	8 min	3
Screensaver	977	2	2 min	2
XML	1 223	3	10 min	4
Java	14 042	2	27 min	2

- First indicators of reading orchestration
 - But: hard to find immediately reading patterns / issues
 - Higher level indicators aim at that

How to evaluate the quality of reading sessions?

- Build a groundtruth
 - not always feasible
- Use quality metrics
 - compliance with parts complexity
 - quality of the reconstruction

1. Compliance with parts complexity

- Complexity in terms of size
 - Important part size \rightarrow Important threshold value ?
- Pearson correlation coefficient between part size \Leftrightarrow part threshold :
 $r = 0,82$

2. Quality of the reconstruction

- Power Law distribution [Berendt, 2001]
 - most visits to a website are concentrated on a small number of pages

	Reading session		30 mn- Session-duration Threshold		10 mn Page-stay Threshold	
	R^2	<i>Err.</i>	R^2	<i>Err.</i>	R^2	<i>Err.</i>
Nodejs	0.94	0.40	0.92	0.42	0.87	0.31
Screensaver	0.86	0.33	0.76	0.48	0.27	0.20
XML	0.89	0.47	0.82	0.45	0.79	0.51
Java	0.95	0.24	0.94	0.23	0.94	0.25

Outline

1. Time-based analysis of reading in eLearning
2. A proposal for detecting reading sessions
- 3. Towards Reading Session-based Indicators**

Reading session-based Indicators

- 27 basic indicators
 - leading to 21 reading indications (higher level)
 - within 4 categories

Global Facts

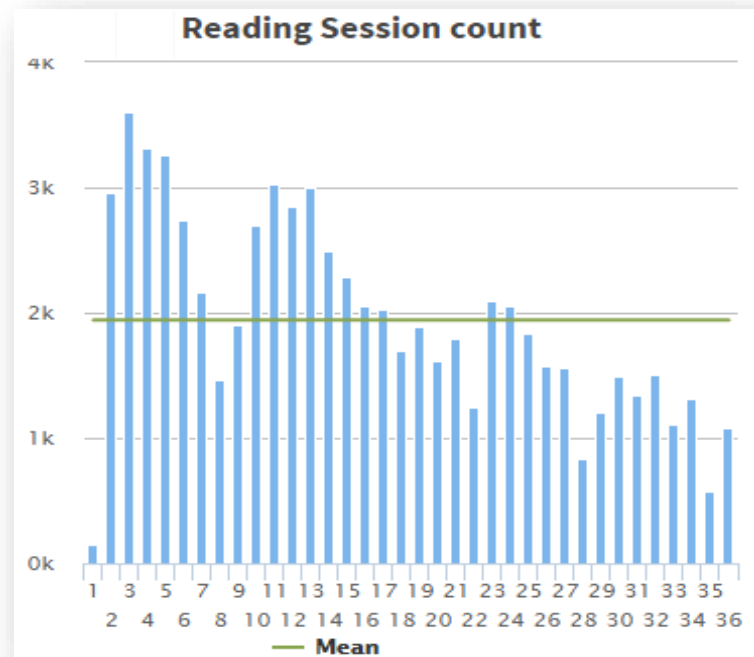
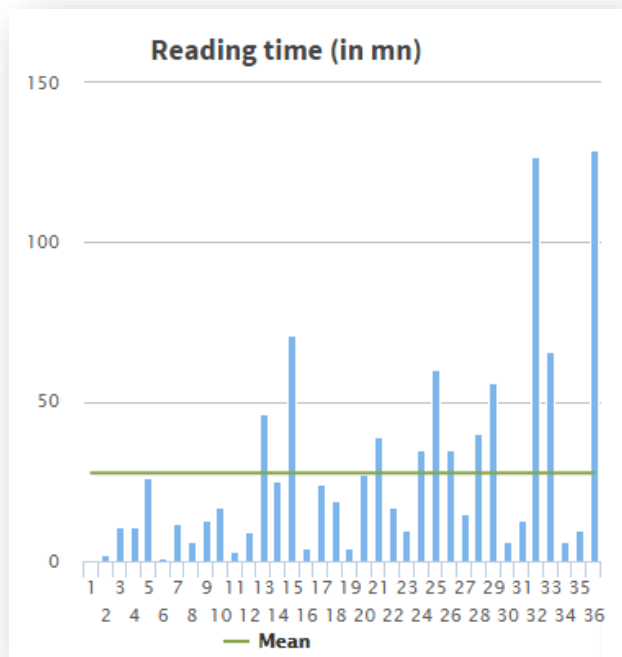
Reading paths and transitions

Rereading indicators

Reading session interruptions




1 : Global Facts

- Basic statistics and hints to characterize reading

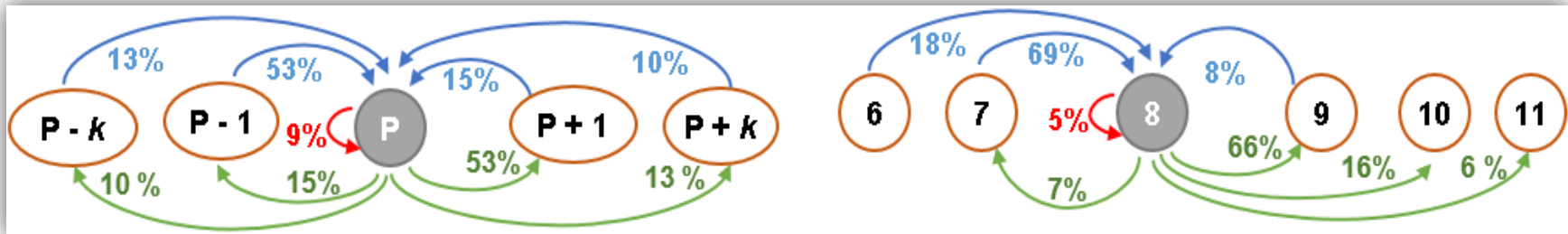


2 : Reading paths and transitions

- Path : sequence of parts read in a reading session

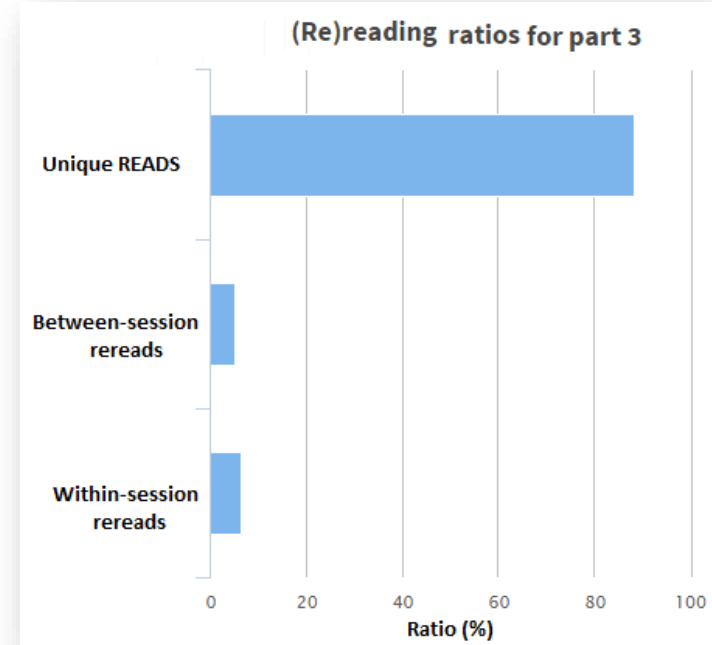
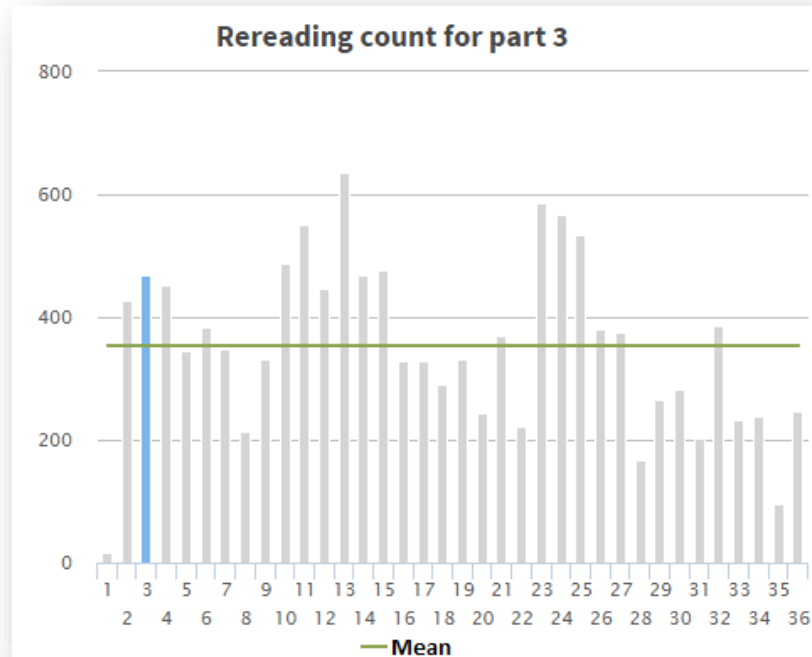
User	Reading session	parts count	Total read parts	Start part	Path	Path Graph	Duration
175	1	7	07/36	3	3;4;5;6;10;34;36		2m
	2	8	10/36	2	2;3;4;5;6;7;8;10		15m 2s
	3	4	13/36	10	10;11;12;13		6m 56s

- Transition → a relation between parts



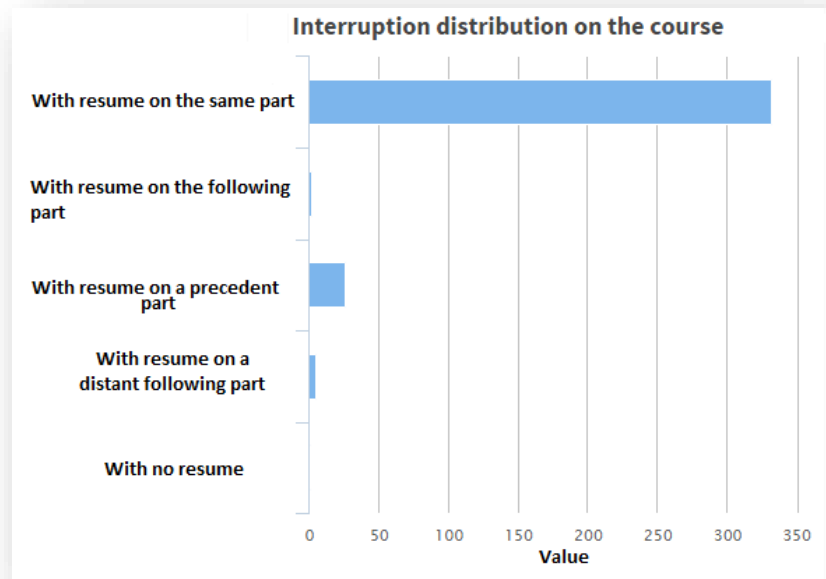
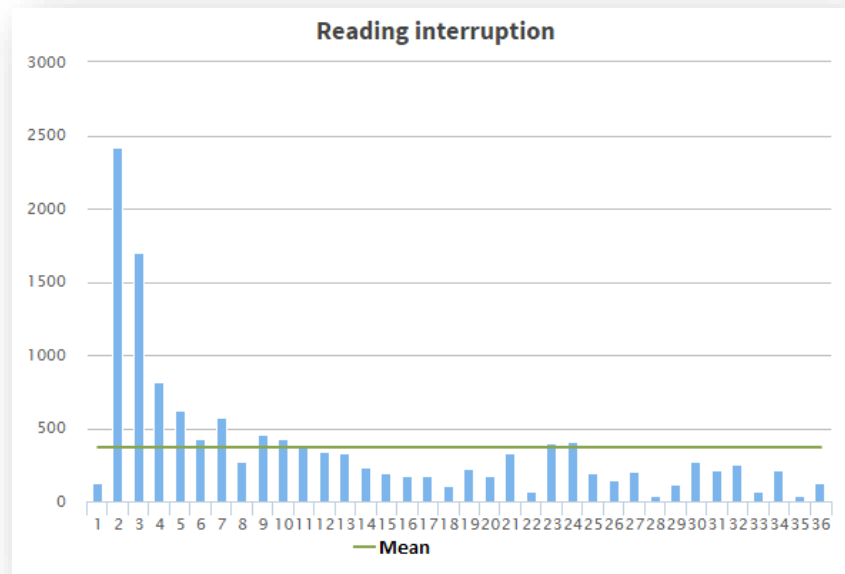
3 : Rereading indicators

- Within-session rereads. Readers struggling with the part?
- Between-session rereads. Reminder needed?



4 : Reading session interruption

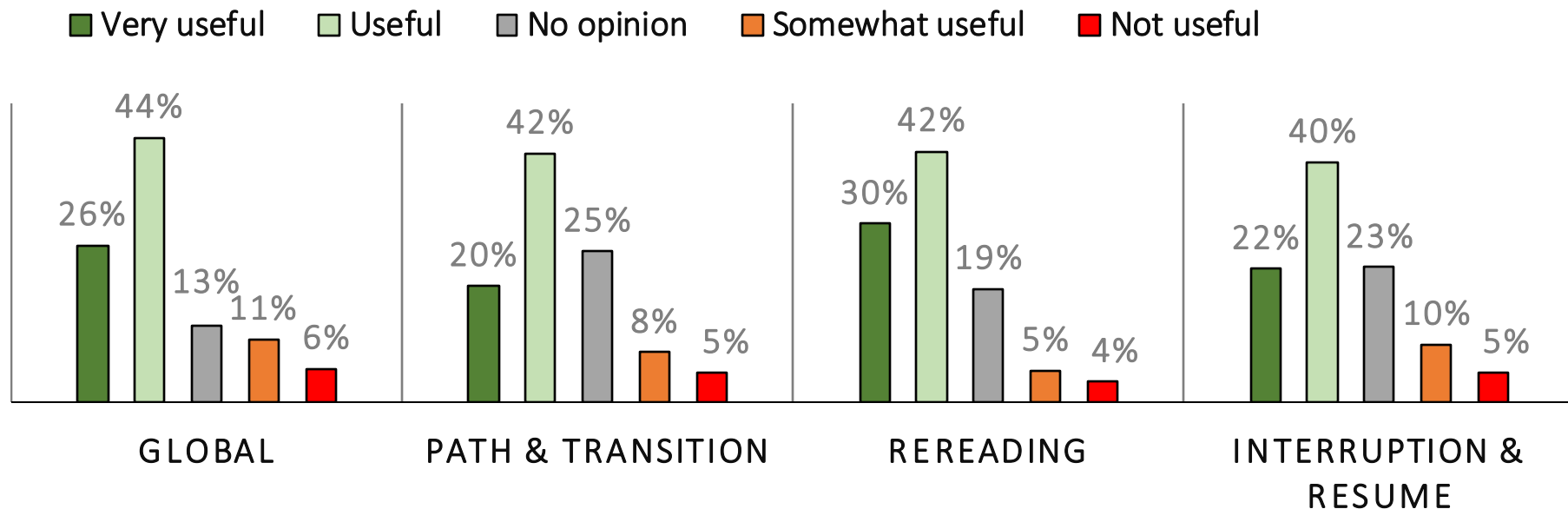
- Interruptions : *Final* or *With Resume*
- Resume: *Same part*, *Next* or *Previous part*, *Distant part*



Evaluation of the suggested indicators & classes usefulness for reconception

- Method
 - online survey
 - likert scales + free comments
- Participants
 - 105 OpenClassrooms course authors

Authors survey: indicators rating



Top indicators: #parts per session, rereaders count and definitive stops

Flop indicators: reading speed per part, session count

Authors survey: comments & opinions

- Exchanges between authors and readers are important

“Making possible for authors and readers to communicate is essential if we want to produce more interesting and productive documents”

“Why not to include direct exchanges between authors and readers through comments and forums ?”

Authors survey: comments & opinions

- Exchanges between authors and readers are important
- Proposed indicators: relevant for reconception. Too many?

“These are important metrics and yes, they would help me understand how to rethink my course document”

“While they seem interesting, I think you would have to select the more interesting to present to authors. The other ones can serve for deeper analysis”

Authors survey: comments & opinions

- Exchanges between authors and readers are important
- Proposed indicators: relevant for reconception. Too many?
- Privacy

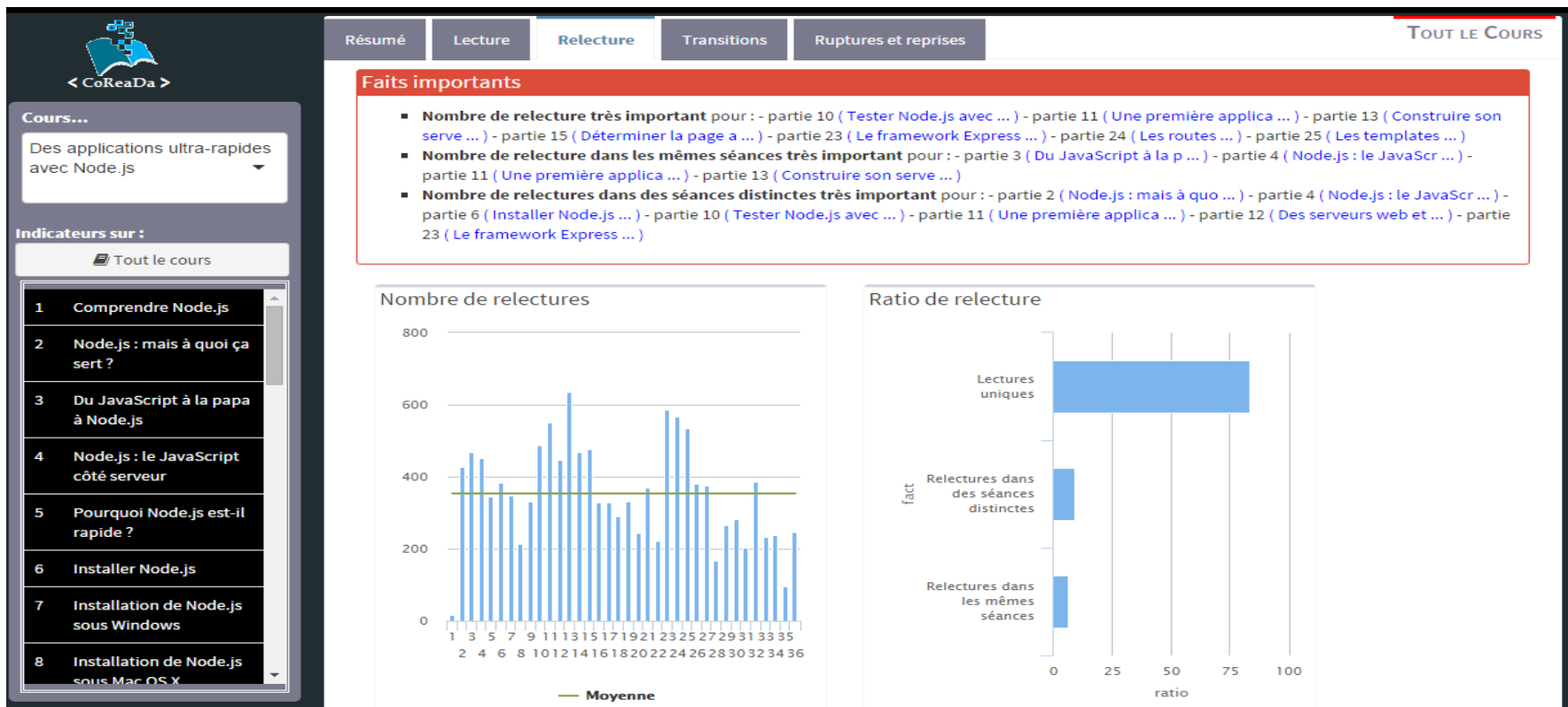
“Be careful not to abuse the personal data of users. The reader should actually be informed that his reading is logged and analyzed”

Main Contributions & Future Work

Contribution 1: Reading Sessions

- grounded on data: real interaction + part characteristics
- threshold values dynamic: become more precise with course, reading and learners' evolution
- ongoing
 - further verification. Which metrics?

Main Contributions & Future Work



Thank you!

[Berendt, 2001] B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. Workshop on Web Mining, 2001.

[Hofmann, 2006] K. Hofmann, C. Reed, and H. Holz. Unobtrusive data collection for web-based social navigation. In Workshop on the Social Navigation and Community based Adaptation Technologies. 2006.

[Kovanovic, 2015] V. Kovanović, D. Gašević, S. Dawson, S. Joksimović, R. S. Baker, and M. Hatala. Penetrating the black box of time-on-task estimation. In Proceedings of the 5th International Conference on Learning Analytics And Knowledge, pp. 184–193. ACM, 2015.

[Marquardt, 2004] C. G. Marquardt, K. Becker, and D. D. Ruiz. A pre-processing tool for web usage mining in the distance education domain. In Database Engineering and Applications Symposium. pp 78–87. IEEE, 2004.

[Ross, 2003] S. M. Ross. Peirce's criterion for the elimination of suspect experimental data. Journal of Engineering Technology, 20(2):38–41, 2003

[Sadallah, 2013] M. Sadallah, B. Encelle, A.-E. Maredj, and Y. Prié. A framework for usage-based document reengineering. In Proceedings of the 2013 ACM Symposium on Document Engineering, DocEng '13, pp. 99–102, New York, NY, USA, 2013. ACM.