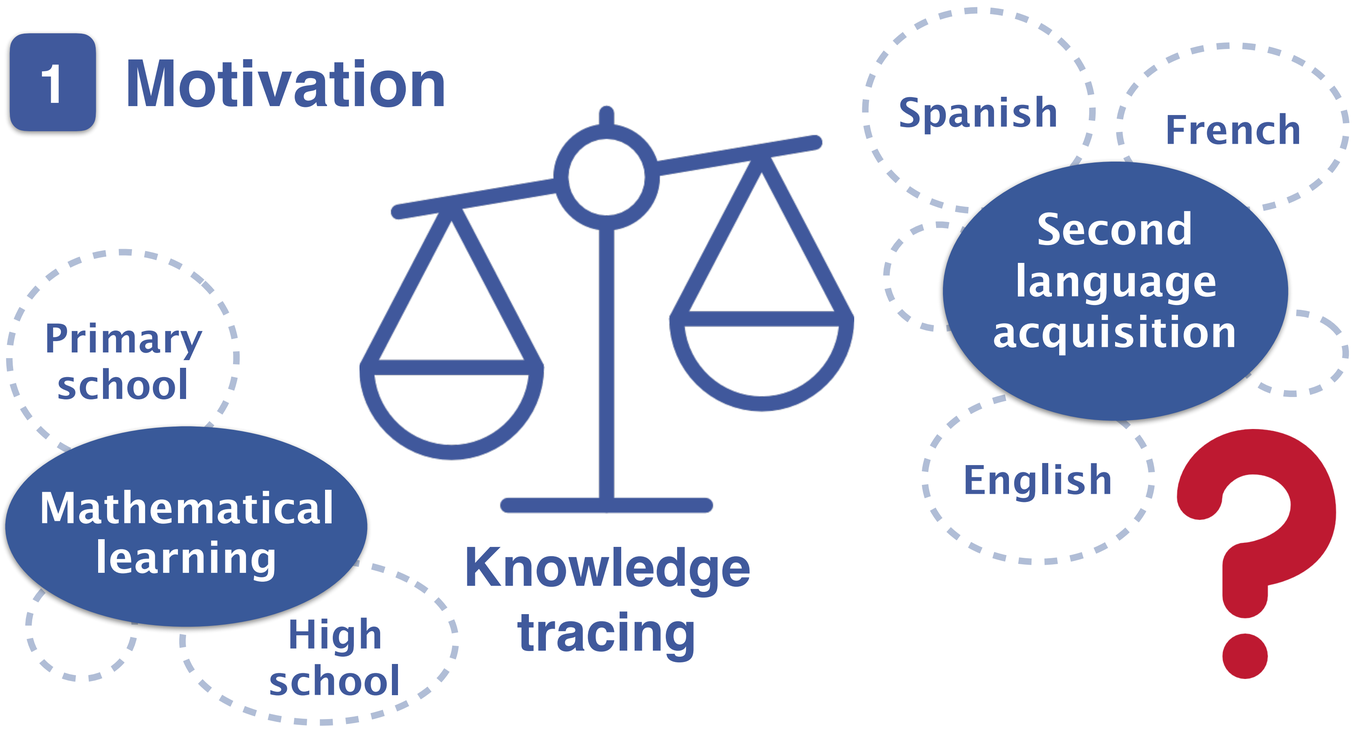
Feature Engineering for Second Language Acquisition Modeling

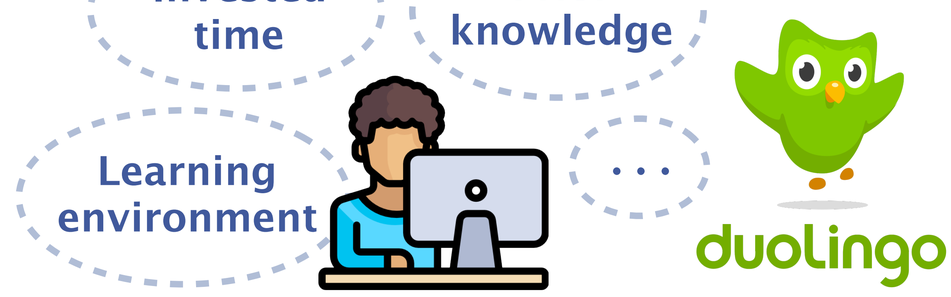
Guanliang Chen, Claudia Hauff, Geert-Jan Houben
Web Information Systems, TU Delft

1 Motivation



2 Research Question

What <u>factors</u> impact students' language learning performance?



3 Research Hypotheses

H1 A student's <u>living community</u> affects her learning performance.

Analyzing by grouping students living in different <u>countries</u>.

H2 The more <u>engaged</u> a student is, the more words she masters.

	Stud-Acc			Mast-Word		
	FR-EN	ES-EN	EN-ES	FR-EN	ES-EN	EN-ES
# Exercises Attempted	-0.05 *	-0.09 *	-0.08 *	0.85 *	0.87 *	0.79 *
# Words Attempted	-0.06 *	-0.08 *	-0.08 *	0.85 *	0.86 *	0.80 *
Time Spent	-0.13 *	-0.14 *	-0.22 *	0.73 *	0.79 *	0.61 *

H3 The <u>more time</u> a student spends on <u>solving an exercise</u>, the more likely she will get it wrong.

	FR-EN	ES-EN	EN-ES
Correlation	-0.16 *	-0.18 *	-0.18 *

H4 <u>Contextual factors</u> such as the device being used, learning type and exercise format impact a student's learning.

	FR-EN	ES-EN	EN-ES
Avg.	84.29	86.31	87.96
Client			
Web	80.64 *	85.44 *	85.68 *
iOS	86.45 *	87.90 *	88.10 *
Android	83.92 *	84.88 *	88.92 *
Session			
Lesson	85.43 *	87.23 *	88.76 *
Practice	80.94 *	83.92 *	84.19 *
Test	82.19 *	84.34 *	84.66 *
Format			
Reverse Translate	77.92 *	85.88 *	85.42 *
Listen	78.30 *	77.01	82.78 *
Reverse Tap	92.51 *	94.84 *	95.48 *

H5 <u>Repetition</u> is useful and necessary for a student to master a word.

	FR-EN	ES-EN	EN-ES
# Previous attempts	-0.05 *	-0.04 *	-0.07 *
Time elapsed	0.05 *	0.06 *	0.07 *

H6 A student with a <u>high-spacing learning routine</u> is more likely to learn more words than one with a low-spacing learning routine.

Analyzing by grouping students according to their spent time and learning spacing routine.

4 Knowledge Tracing

Step 1: Feature Engineering

Features	Granularity Level: User	Word	Exercise
Student ID	√		
Word		√	
Countries	√		
Format			√
Type			√
Device			√
Time spent (exercise)			√
# Exercises attempted	√		
# Words attempted	√		
# Unique words attempted	√		
# sessions	√		
Time spent (learning)	√		
# Previous attempts	√	√	
# Correct times	√	√	
# Incorrect times	√	√	
Time elapsed	√	√	
Word-Acc	√	√	
Std. timestamps (exercise)	√		√
Std. timestamps (word)	√	√	
Std. timestamps (session)	√		
Std. timestamps (word-session)	√	√	
Std. timestamps (word-correct)	√	√	
Std. timestamps (word-incorrect)	√	√	

23 features

Step 2: Gradient Tree Boosting

	Methods	AUC	F1
FR-EN	Baseline	0.77	0.28
	GTB	0.82 *	0.41 *
ES-EN	Baseline	0.75	0.18
	GTB	0.80 *	0.34 *
EN-ES	Baseline	0.77	0.19
	GTB	0.82 *	0.39 *

Gradient Tree Boosting is most effective with 9 of the designed features, e.g., learning type and exercise format, for second language acquisition modeling.

More efforts on feature engineering are needed.

TU Delft Extension School