

# Optimizing Human Language Learning

Masato Hagiwara

Software Engineer / Researcher



duolingo



279. (a) Trans la korto staras nia domo. Si logas trans la rivero.  
Ili restas trans la maro. Li vagas trans montoj kaj maro.

(b) Across the field is a rose-garden. Stand and wait on-the-other-side-of the street. Scotland is there, across the valley.

#### KE

280. KE=that, used as a link-word joining two sentences (a "conjunction"). Mi kredas, ke vi estas prava, *I believe (that) you are right*. Here ke joins the sentences Mi kredas and vi estas prava. Mi scias, ke li logas tie, *I know (that) he lives there*. Mi bedaŭras, ke vi malsanas; *I am-sorry (regret) (that) you are ill*.

In English *that* is often omitted, but you must not omit the ke. Place a comma before ke, to divide the sentences.

281. Mi ĉial ĝojas, ke mi estas Esperantisto. Cu vi sentas, ke mi eraras? Nia opinio estas (ni opinias) (*If it-our-opinion, we think*), ke tio estas ideo idiota. Kiel vi supozas, ke tio ne estas ebla? Si pretendas (claims), ke ti estas reĝidino. Mi divenas (guess), ke vi estas okjara.

Mi respektas lin tial, ke (because, for the reason that) li estas tre modesta. (Tial ke=a strong form of car).

282. Tri gradoj de mar-malsano. (1) Mi kredas, ke mi mortas. (2) Mi scias, ke mi mortas. (3) Mi timas, ke mi ne mortas.

"Cu vi opinias, ke mi estas idioto?" "Tute ne! Tamen eble mi eraras."

283. Make sentences using ke after some of the words bedaŭr- (diri- frustr- gej- instru- juĝ- kompren- konfes- konsent- leg- lern- memor- opinii- plend- predik- pruv- rakont- respond- rev- ripet- sci- sent- skrib- sonig- telefon- telegraf- jas. Do not forget the comma.

(b) Complete the following sentences. Ne forgesu, ke... Memoru, ke... Ne respondu, ke... Kredu al mi, ke... Li ofte assertas, ke... Kiel vi supozas, ke... Cu vi do timas, ke...? Cu vi vere suspektas, ke...?

#### THAT

284. This English word has many meanings, and must be translated accordingly. He says about that, that that "that" "that is in your essay is wrong. Li diras pri tio, ke tiu "that", kiu estas en via eseo, estas malgusta.

285. She says (that) he is rich, but for-some-reason I do not suppose (that) that is true. Tell him (that) I am not a man of-that-kind. I feel (that) you are wrong on that matter. She asserts (that) she is right about that question. Do not believe that all that is in that book is true. That proves the earth is round.

#### KE JES, KE NE

286. The translation of ke jes, ke ne, depends on the context. Cu li estas tie? (Mi kredas, ke jes, *I believe he is, I think so*). Mi supozas, ke ne, *I suppose he isn't, I suppose not*). Cu vi komprenas? (Mi timas, ke ne, *I'm afraid not, I fear I don't*). Mi scias, ke jes, *I know I do*).

287. Write similar questions and answers, using ke jes, ke ne, after some of the verbs assert-(bedaŭr- dir- esper- ŝoj- insist- konfes- konsent- kred- opinii- ripet- sci- skrib- supoz- suspekt- telefon- tim- vid- jas (Compare 104-5).

#### NOTE ON PREPOSITIONS

288. Al, apud, de, dum, en, inter, kun, por, pri, sen, sub, super, sur, tra, trans, are prepositions (96).

289. In English conversation we often end a sentence with a preposition, thus: "What are you sitting on?" "Whom are you thinking of?"\*—in spite of the famous rule "A preposition is a word you must not end a sentence with"! But in Esperanto you may NOT do this. Write, therefore: "Sur kio vi sidas?" "Pri kiu vi pensas?"

290. Translate these questions and answer them. What are you complaining of? Which town do you come from? What are we reading through? What are we fighting for? What is your money in? Whom are you sitting near? What are you doubtful about? Whom are you waiting for? What does an apple grow on? Whom are you writing to? What are you talking about? Whom are you dreaming of?

#### "SUBJECT" AND "OBJECT"

291. Most sentences contain (1) a SUBJECT (the thing or person we are talking about) and (2) a VERB (which says what the subject is or does). Thus, in *la suno brilas*, and *li staras tie*, *la suno* is the subject of the verb *brilas*, and *li* is the subject of the verb *staras*. The answer to *What?* *Who?* asked before a verb, is the subject of the sentence. *What shines?* (*la suno*). *Who stands?* (*li*).

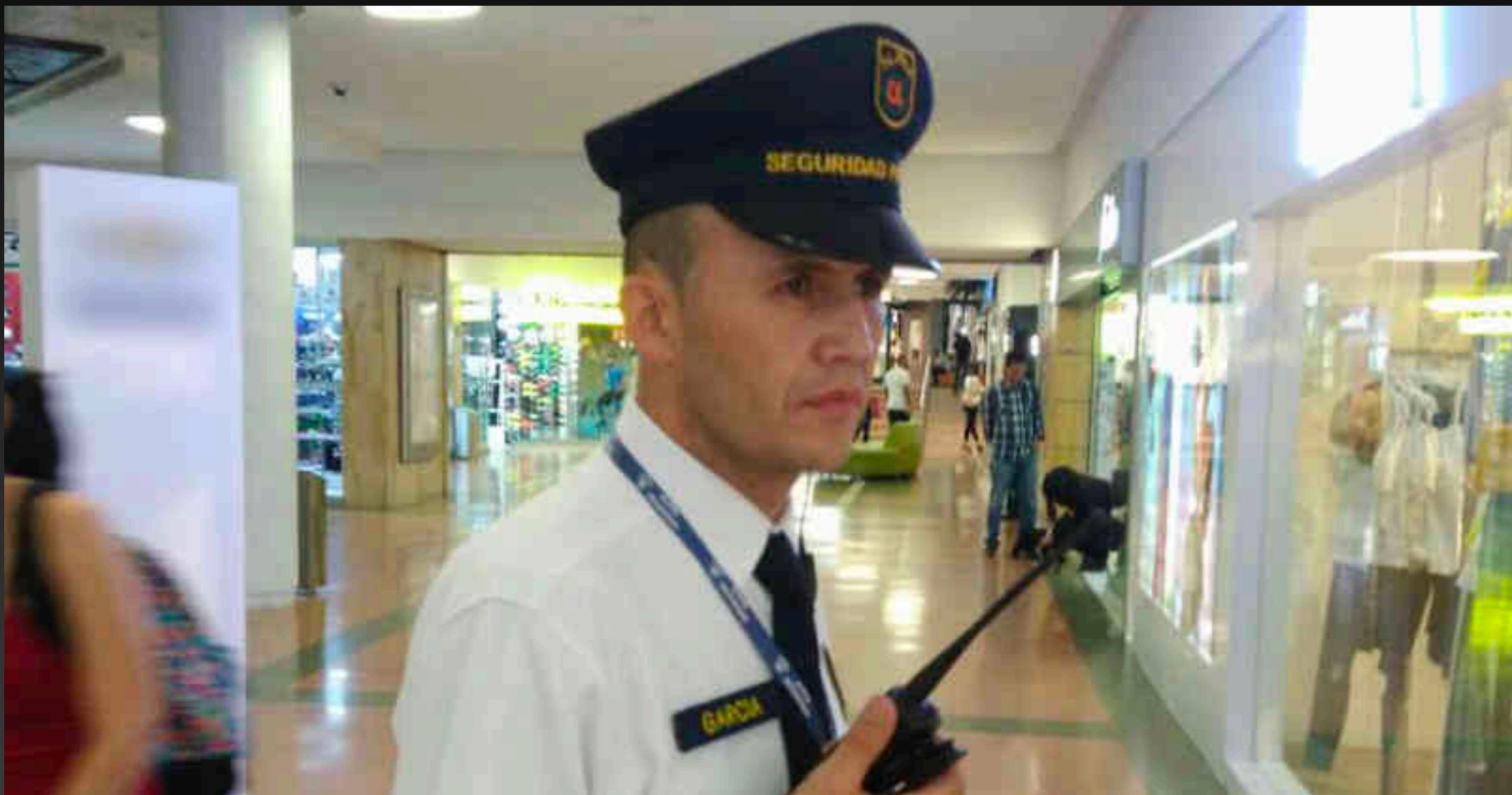
292. Usually the subject comes before the verb. But it need not do so (20, 23). One may say, for example, *Brilas la suno. Tie staras li*.

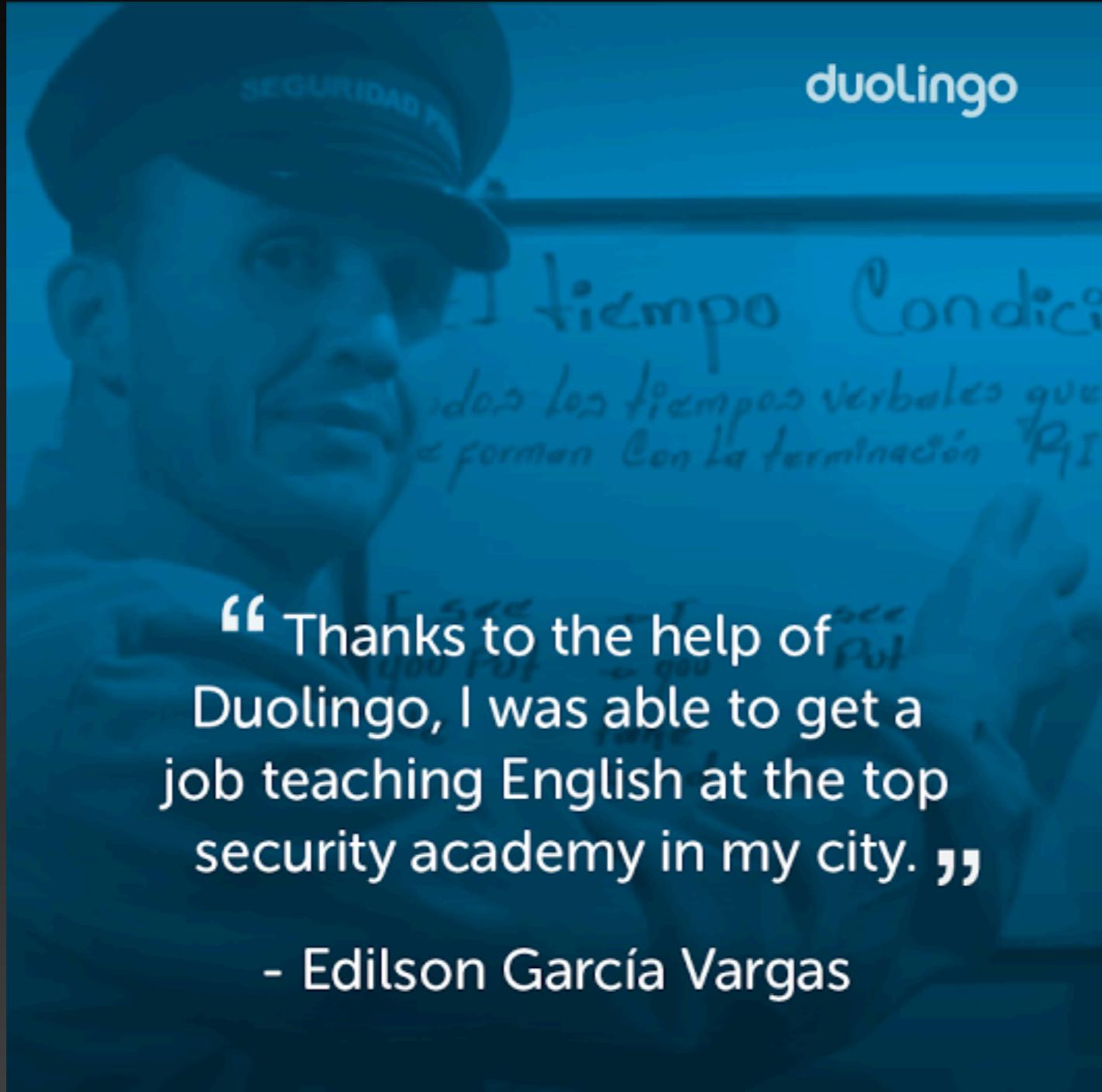
\*Even "What was the man who asked what you chose that book to be read to out of for up to?"!

“Learning Esperanto used to be a solitary quest. You could practise it by sitting for weeks with a book and a dictionary, figuring out the rules and memorising the words. But there was usually no professor to correct your mistakes or polish your pronunciation.”

“Newer generations are not as patient, and they don’t have to be. Unlike most of their elders, who rarely had the chance to speak Esperanto, today’s speakers can use the language every day online.”

— Jose Luis Penarredonda, BBC, 2018





duolingo

“ Thanks to the help of Duolingo, I was able to get a job teaching English at the top security academy in my city. ”

- Edilson García Vargas

people learning a second language

1,200,000,000

(~16% of the world's population)

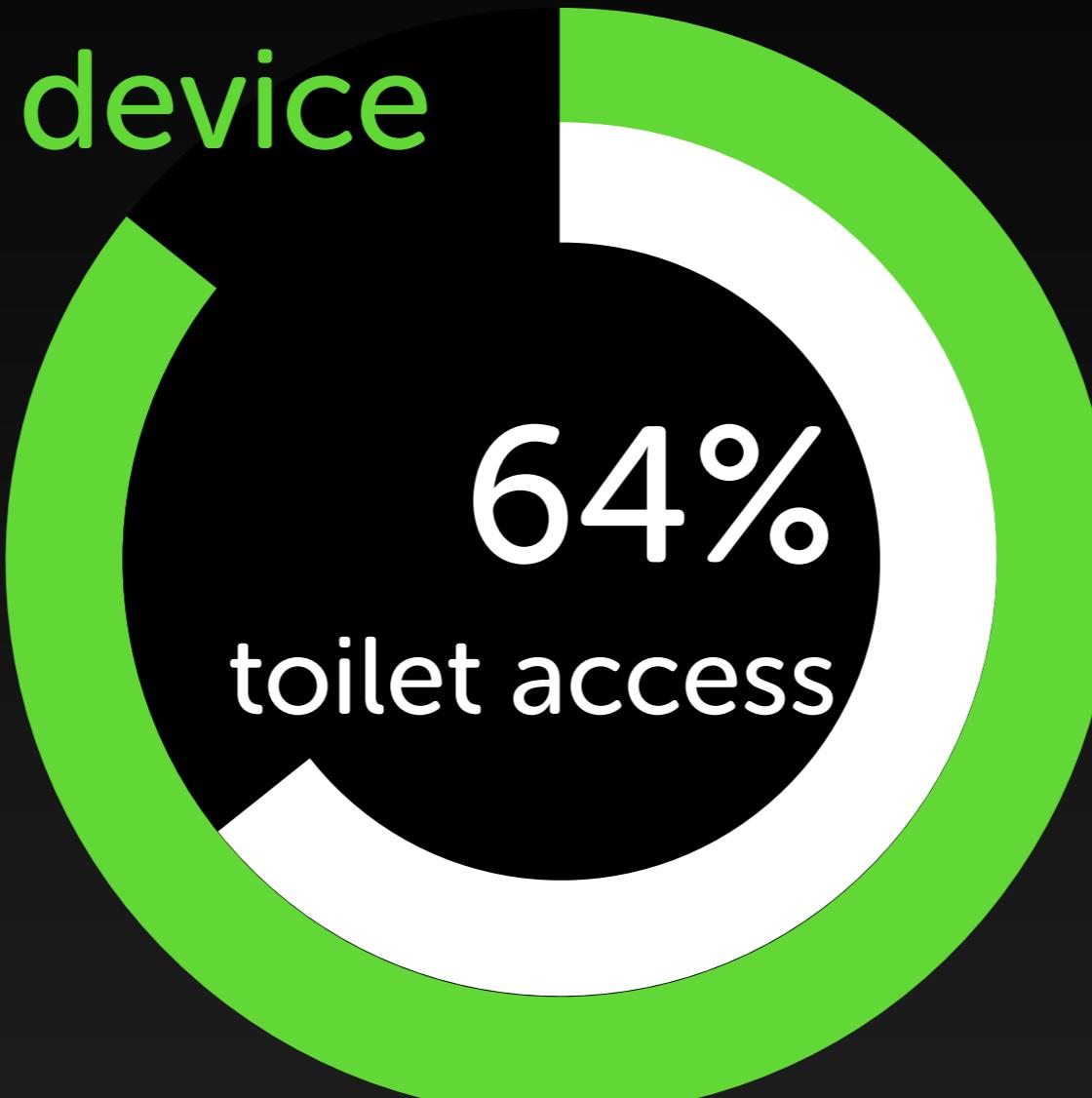
~800M satisfy three properties:

- learning English
- in a developing country
- to gain more opportunity

(Source: British Council)

86%

mobile device  
access



64%

toilet access

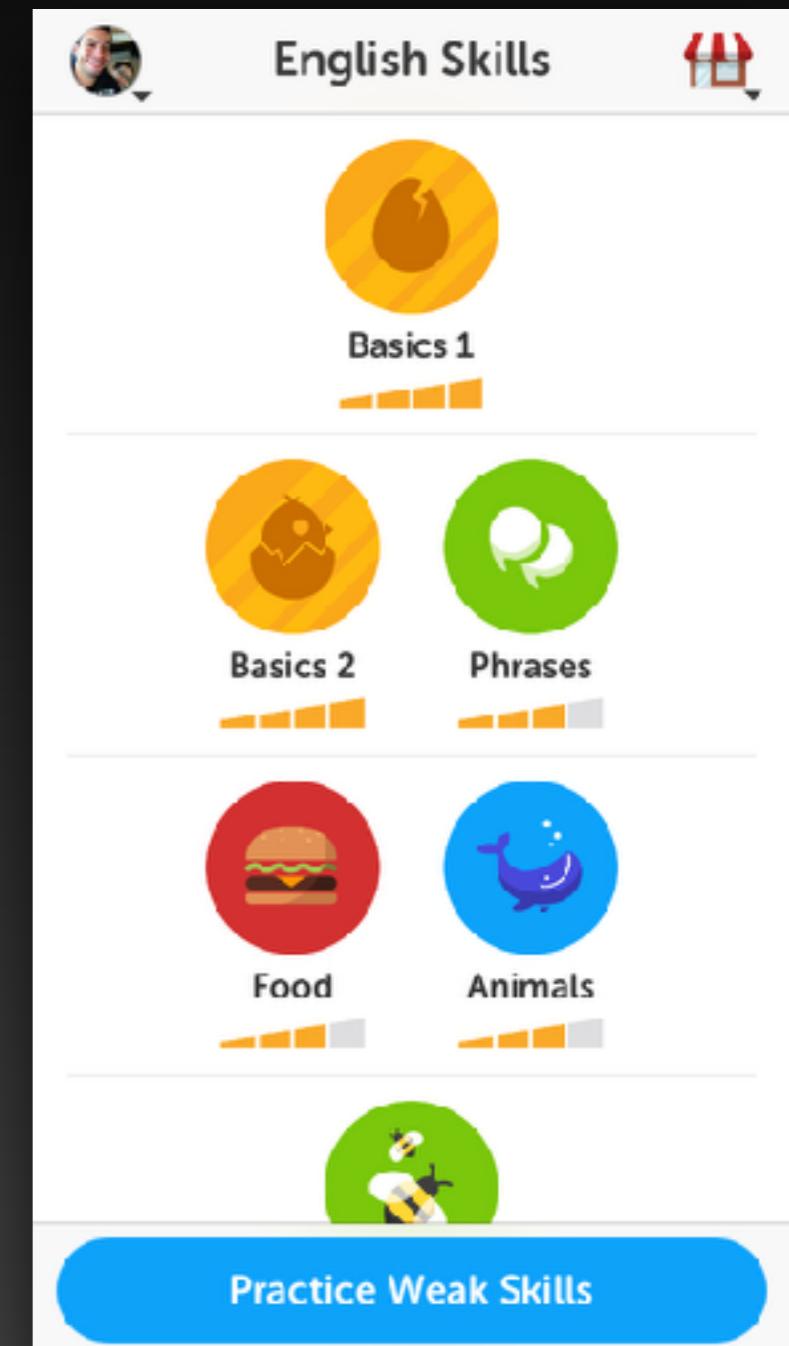
(Source: U.N. Report, 2015)

# duolingo

200M+  
registered users

79  
courses

31  
languages



(As of May. 2018)

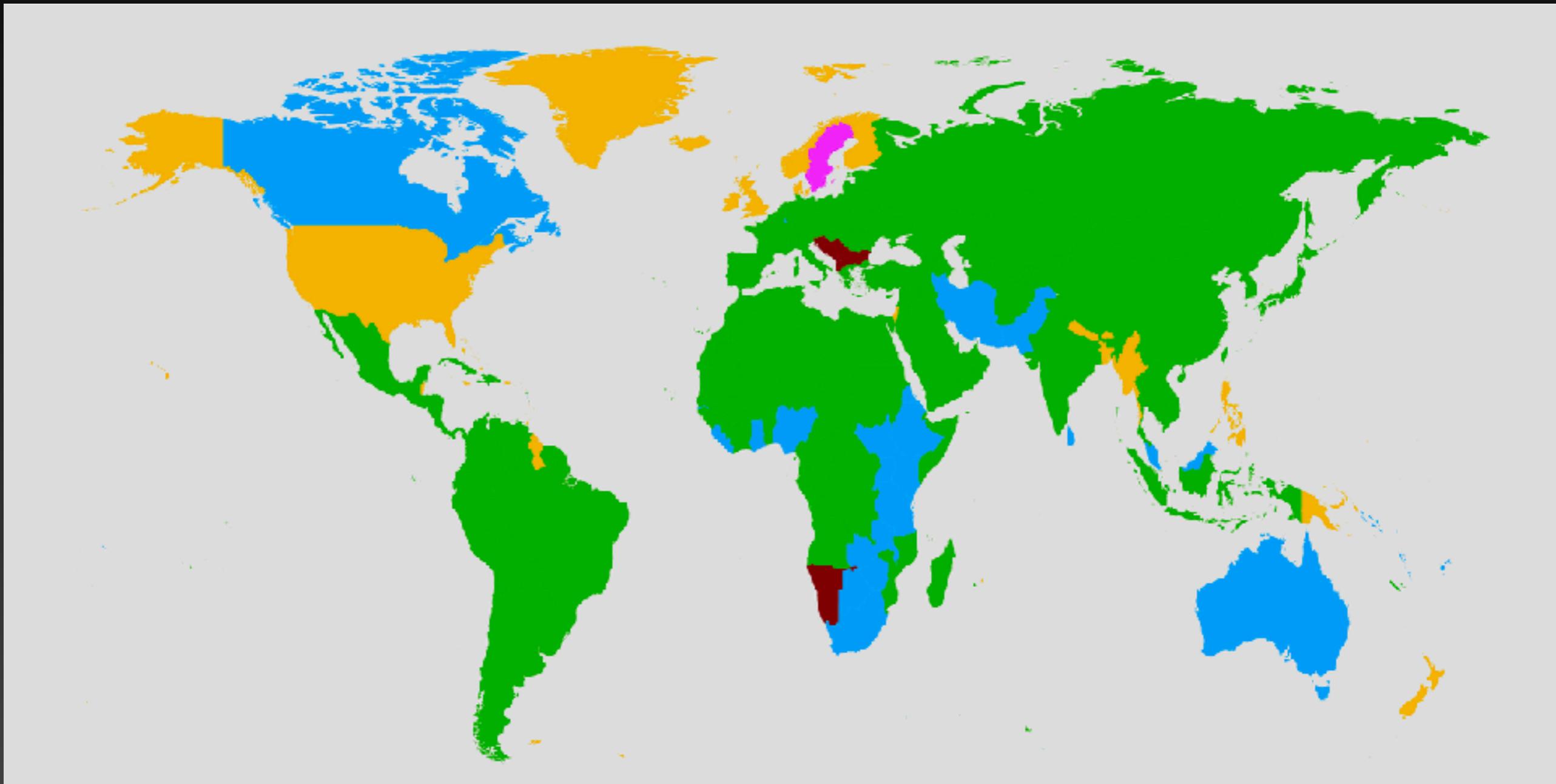
Spanish

French

English

German

Swedish



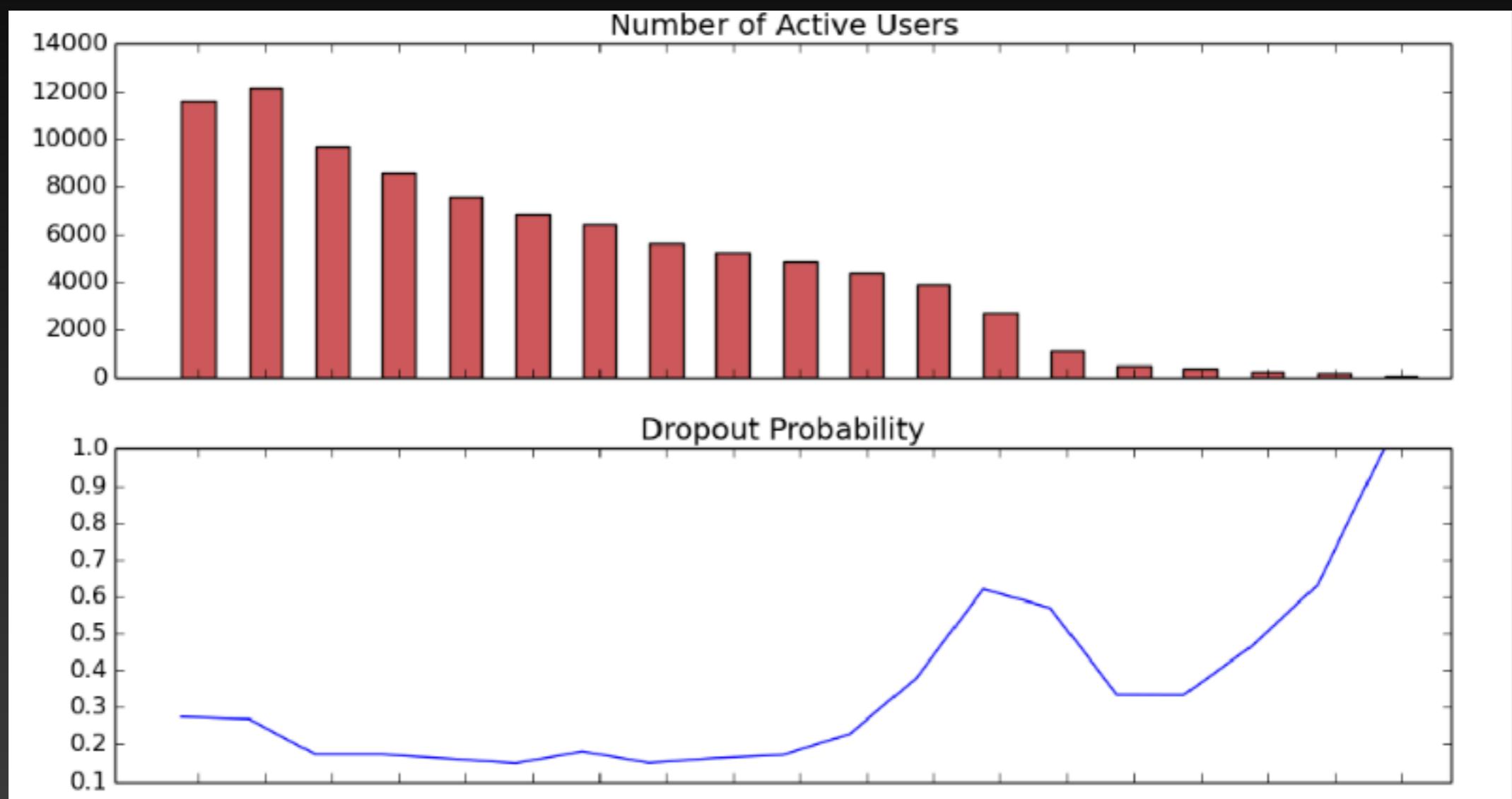
Most popular language studied on Duolingo in each country

# Human

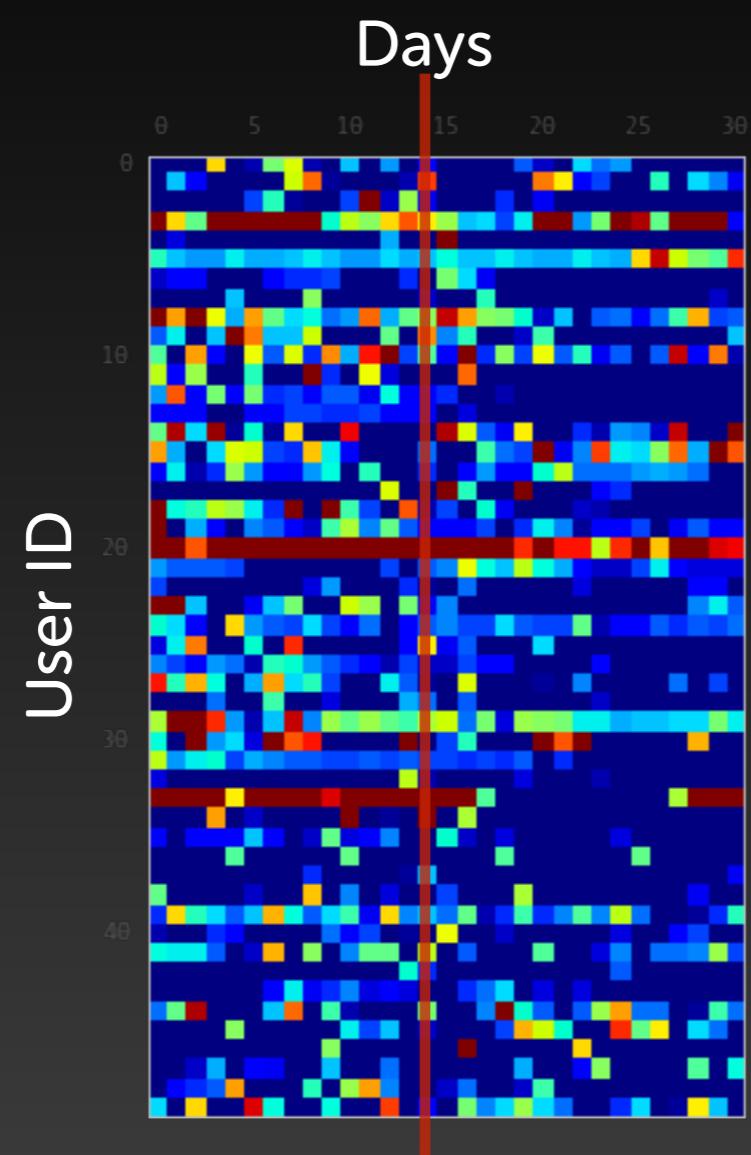
# “User attention is expensive”



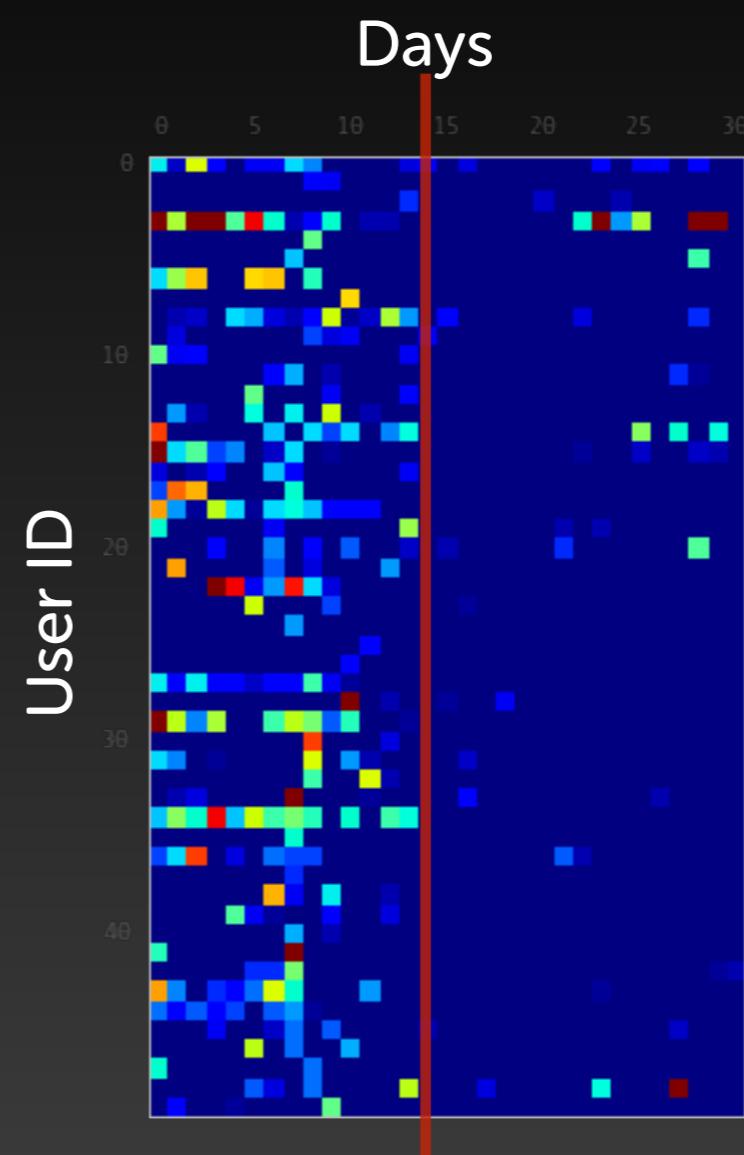
Source: Ultimate App Icons Set Sketch Resource



week

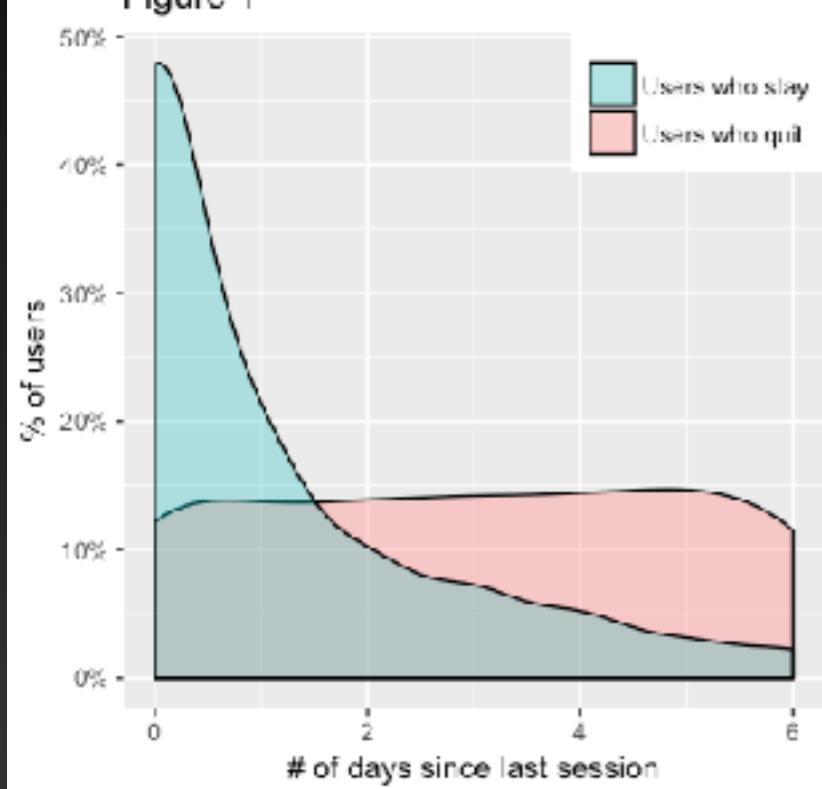


3,466 users who stayed



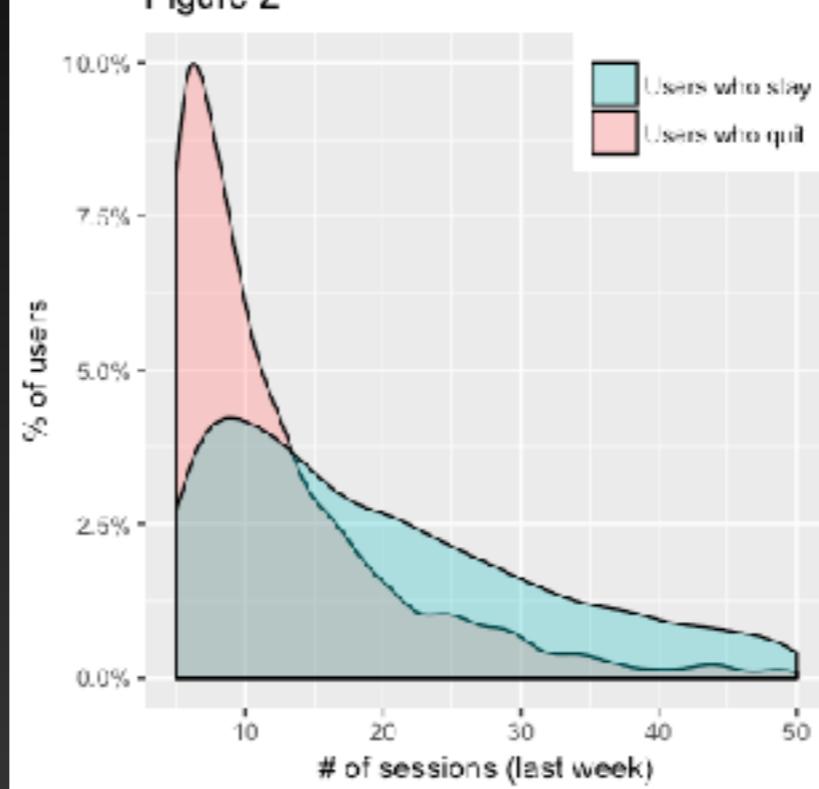
3,426 users who quit

**Figure 1**



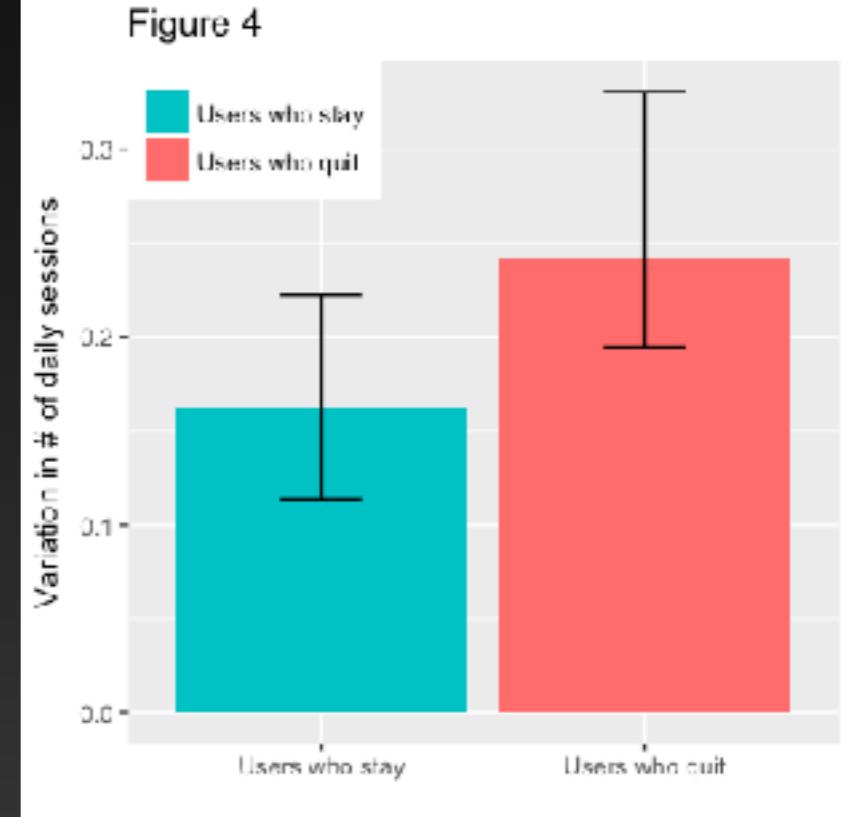
**# of days since  
last session**

**Figure 2**

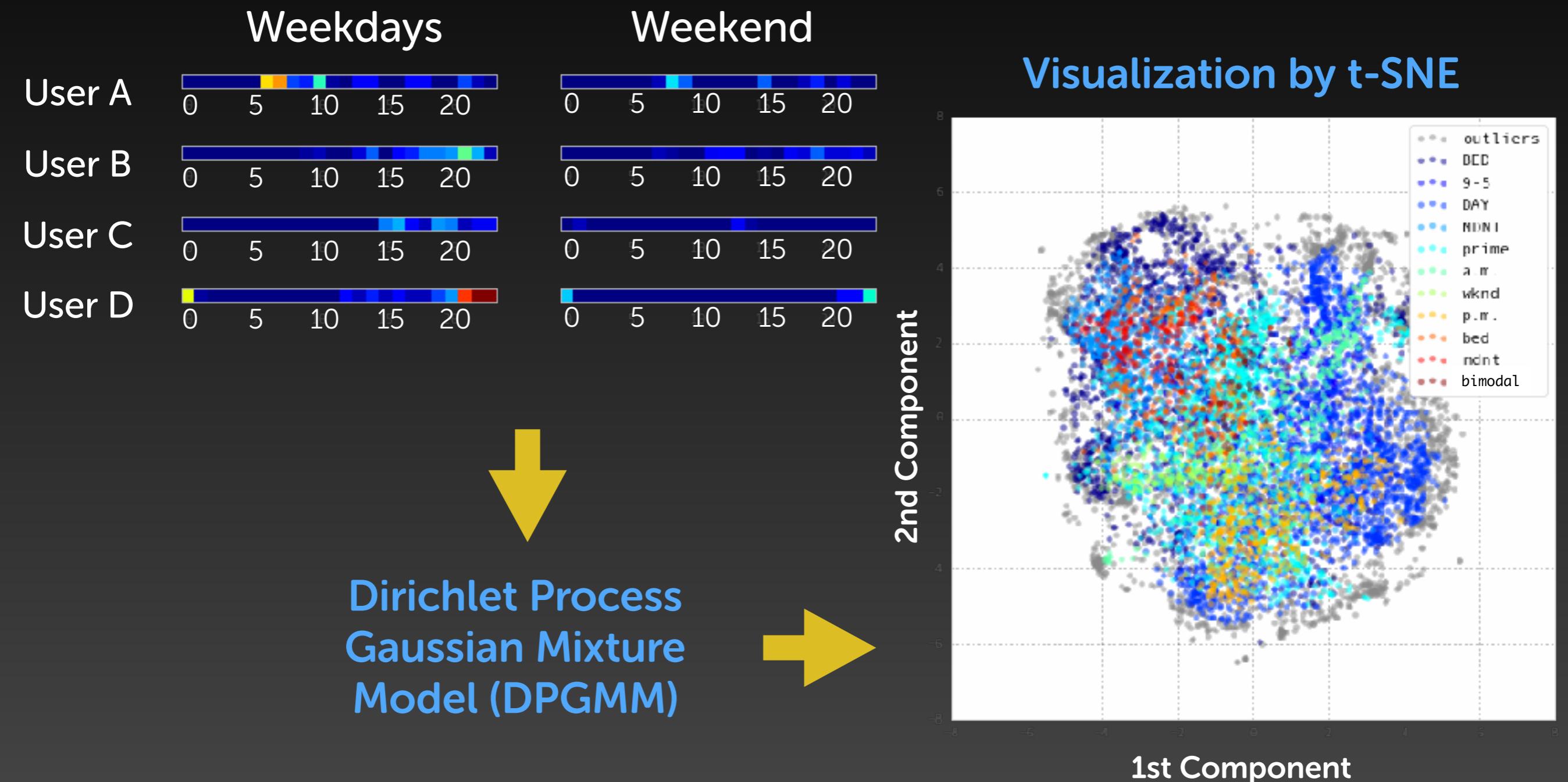


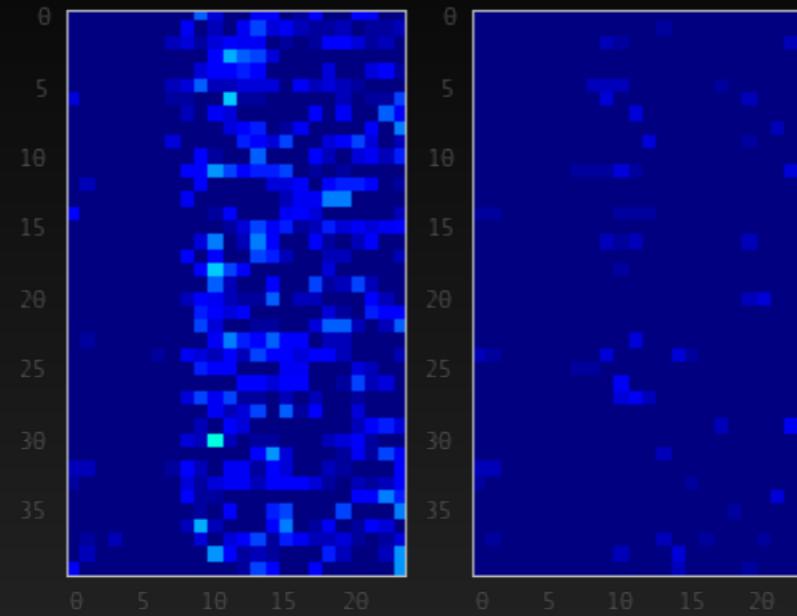
**# of sessions  
(last weeks)**

**Figure 4**

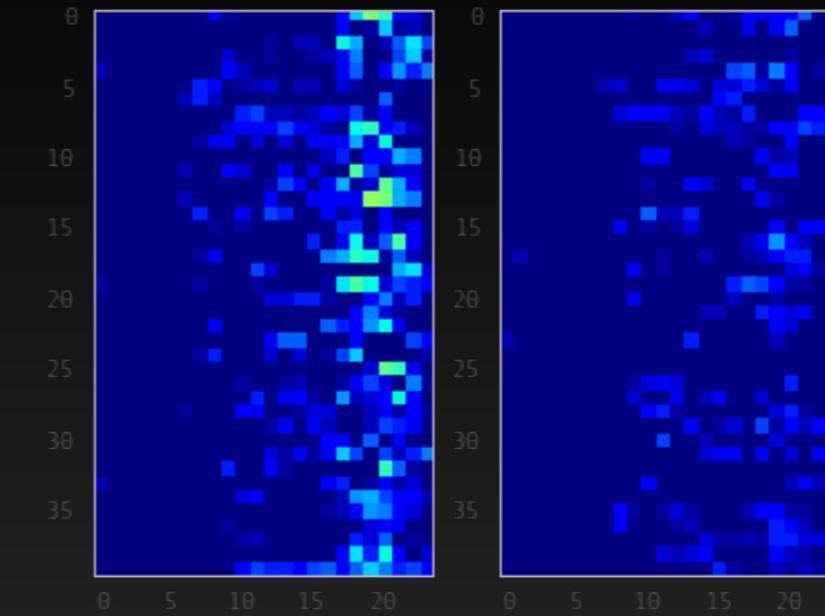


**Relative stdev of  
daily # of sessions**

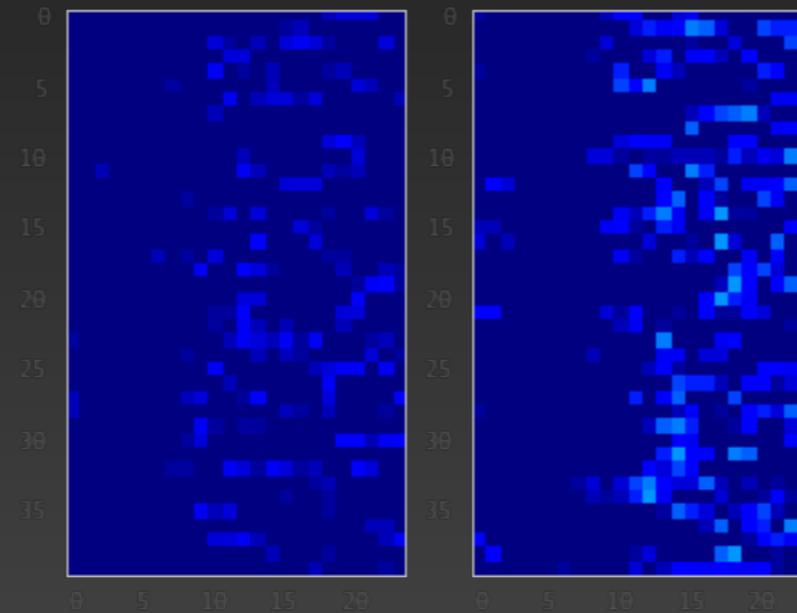




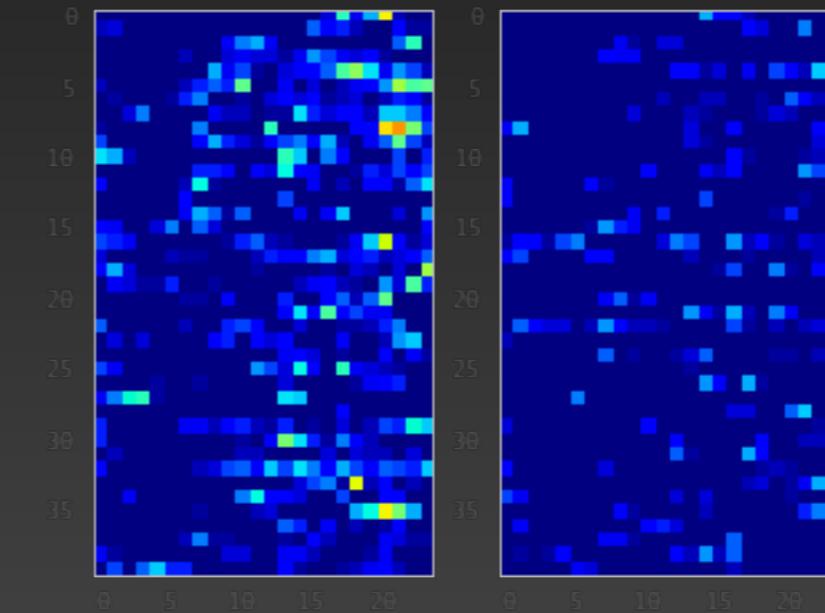
9-to-5



Prime time (~8pm)

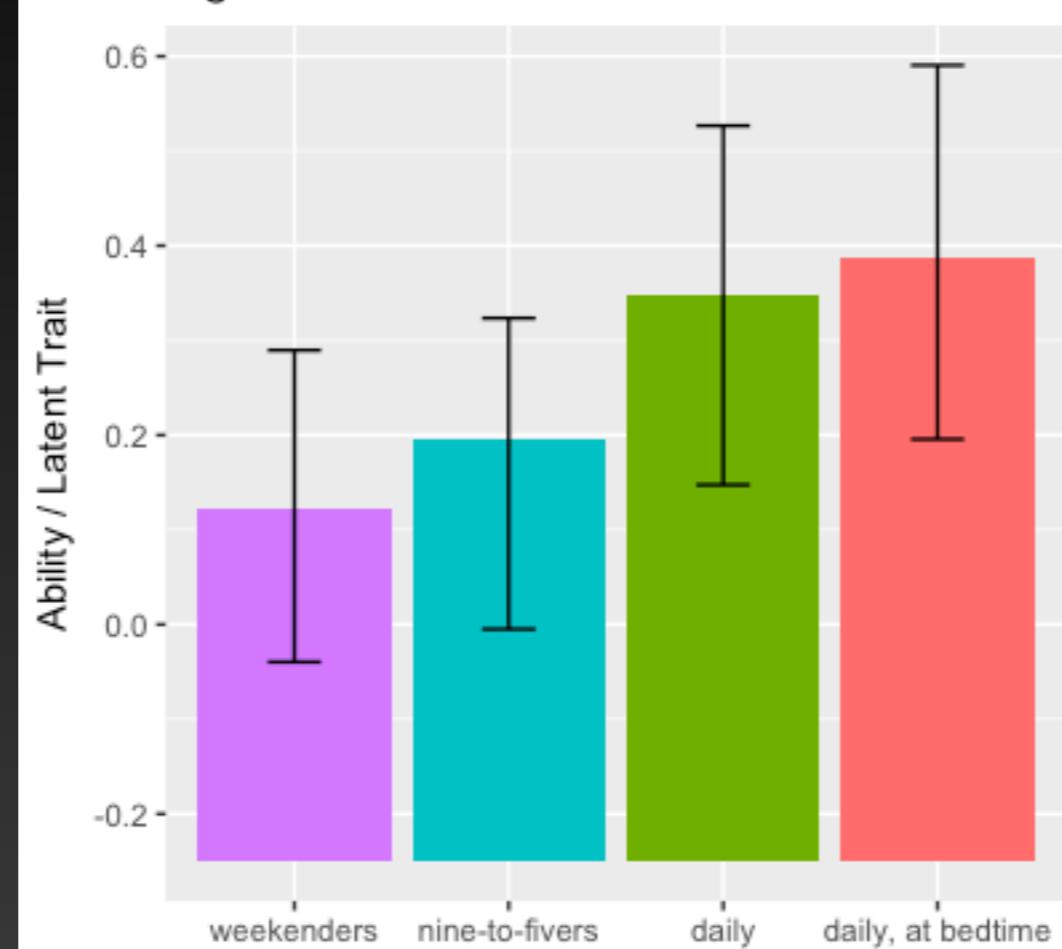


Weekenders



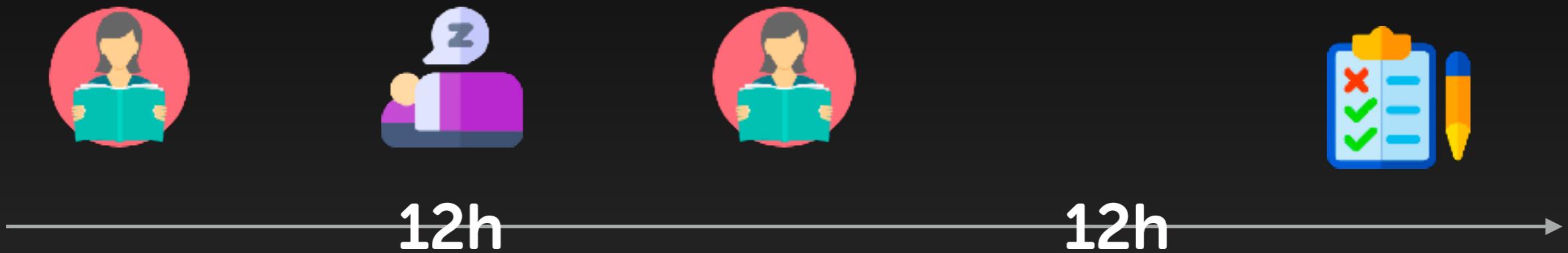
Arbitrary

Figure 3

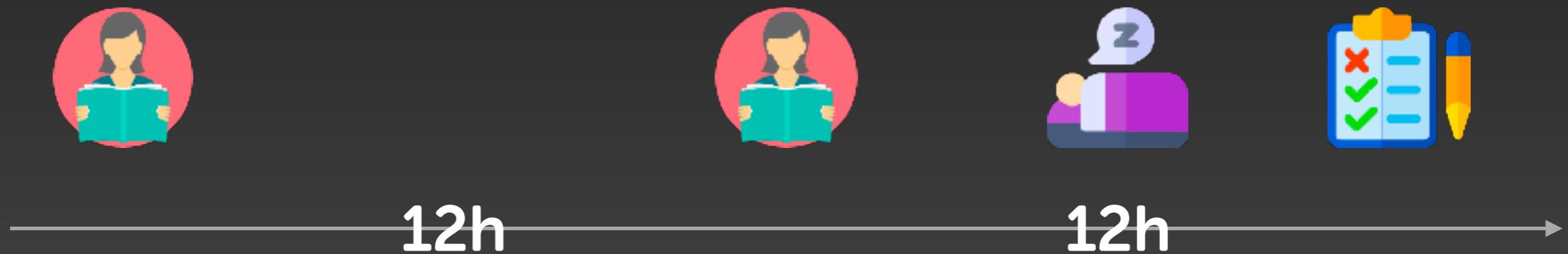


~30%+ retention  
2x relearning

A



B



Source: (Mezza et al. 2016)

"Learning" by Icon Pond (Flaticon)  
"Healthy Lifestyle" and "Employment" by Freepik (Flaticon)

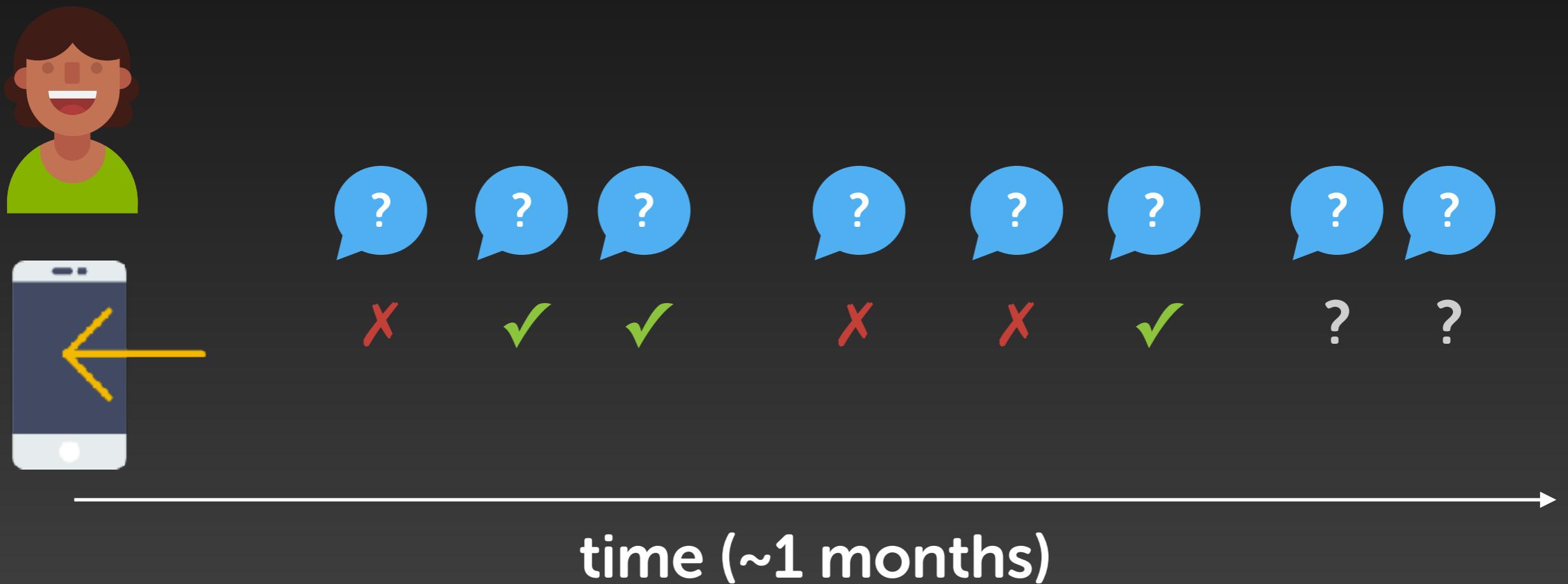
# Language

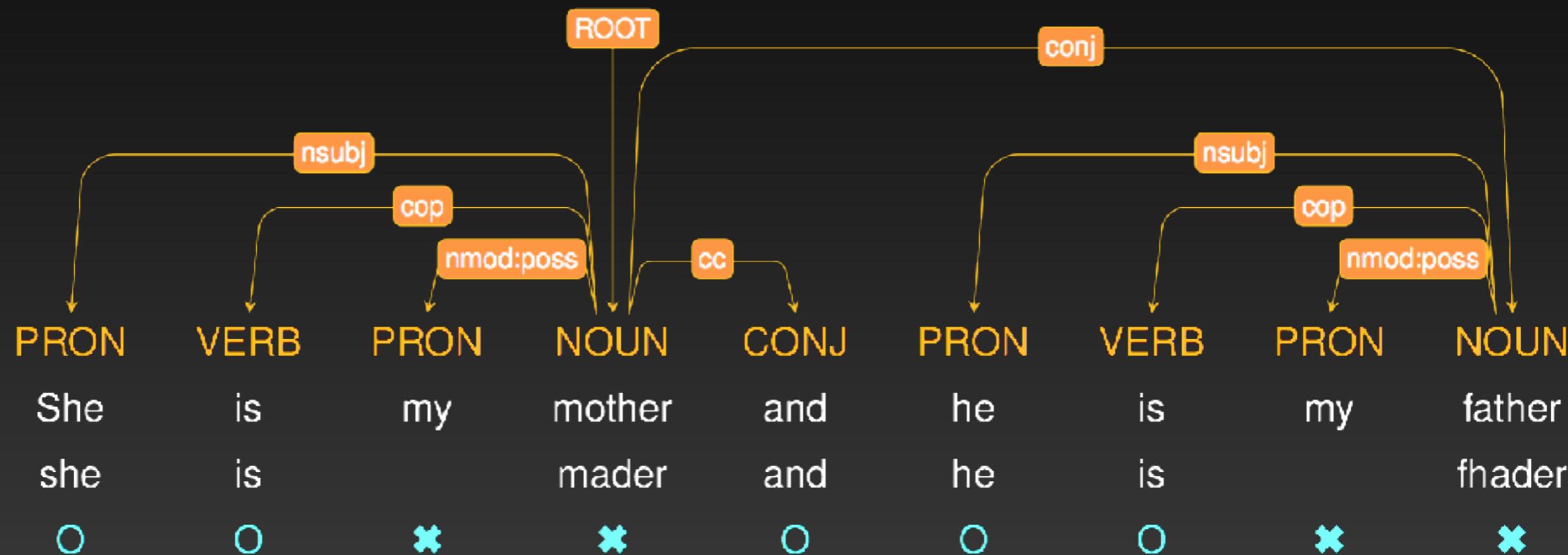
Language  
learning

Multilingual

Longitudinal

# 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)





# user:D2inSf5+ countries:MX days:1.793 client:web session:lesson format:reverse\_translate time:16

8rgJEAPw1001	She	PRON	Case=Nom Gender=Fem Number=Sing Person=3 PronType=Prs fPOS=PRON++PRP	nsubj	4	0
8rgJEAPw1002	is	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin fPOS=VERB++VBZ	cop	4	0
8rgJEAPw1003	my	PRON	Number=Sing Person=1 Poss=Yes PronType=Prs fPOS=PRON++PRP\$	nmod:poss	4	1
8rgJEAPw1004	mother	NOUN	Degree=Pos fPOS=ADJ++JJ	ROOT	0	1
8rgJEAPw1005	and	CONJ	fPOS=CONJ++CC	cc	4	0
8rgJEAPw1006	he	PRON	Case=Nom Gender=Masc Number=Sing Person=3 PronType=Prs fPOS=PRON++PRP	nsubj	9	0
8rgJEAPw1007	is	VERB	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin fPOS=VERB++VBZ	cop	9	0
8rgJEAPw1008	my	PRON	Number=Sing Person=1 Poss=Yes PronType=Prs fPOS=PRON++PRP\$	nmod:poss	9	1
8rgJEAPw1009	father	NOUN	Number=Sing fPOS=NOUN++NN	conj	4	1

# user:D2inSf5+ countries:MX days:2.689 client:web session:practice format:reverse\_translate time:6

oMGsnnH/0101	When	ADV	PronType=Int fPOS=ADV++WRB	advmod	4	1
oMGsnnH/0102	can	AUX	VerbForm=Fin fPOS=AUX++MD	aux	4	0
oMGsnnH/0103	I	PRON	Case=Nom Number=Sing Person=1 PronType=Prs fPOS=PRON++PRP	nsubj	4	1
oMGsnnH/0104	help	VERB	VerbForm=Infl fPOS=VERB++VB	ROOT	0	0

	Users	Tokens		
		TRAIN	DEV	TEST
English	2.6k	2.6M	387k	387k
Spanish	2.6k	2.0M	289k	289k
French	1.2k	927k	138k	136k
Overall	6.4k	5.5M	814k	804k

# 2018 Duolingo SLAM Shared Task

- Phase 1 (8 weeks): TRAIN and DEV sets released
- Phase 2 (10 days): TEST set released
- Metrics: AUROC (Area Under ROC curve) + F1
- 15 teams participated (11 submitted paper)

				English	Spanish	French	Average
Team	RNN	Ensemble	Multitask	AUC	AUC	AUC	Rank
<u>SanaLabs</u>	✓	✓		0.861	0.838	0.857	1.0
<u>singsound</u>	✓			0.861	0.835	0.854	1.7
<u>NYU</u>		✓	✓	0.859	0.835	0.854	2.3
<u>TMU</u>	✓		✓	0.848	0.824	0.839	4.3
<u>CECL</u>			✓	0.846	0.818	0.843	4.7
<u>Cambridge</u>	✓			0.841	0.807	0.835	6.0
<u>UCSD</u>		✓		0.829	0.803	0.823	7.0
<u>LambdaLab</u>		✓		0.821	0.801	0.815	7.6
<u>Grotoco</u>				0.817	0.791	0.813	9.0
<u>nihalnayak</u>				0.821	0.790	0.811	9.0
<u>jilljenn</u>				0.815	0.788	0.809	10.7
<u>SLAM_baseline</u>				0.774	0.746	0.771	14.7

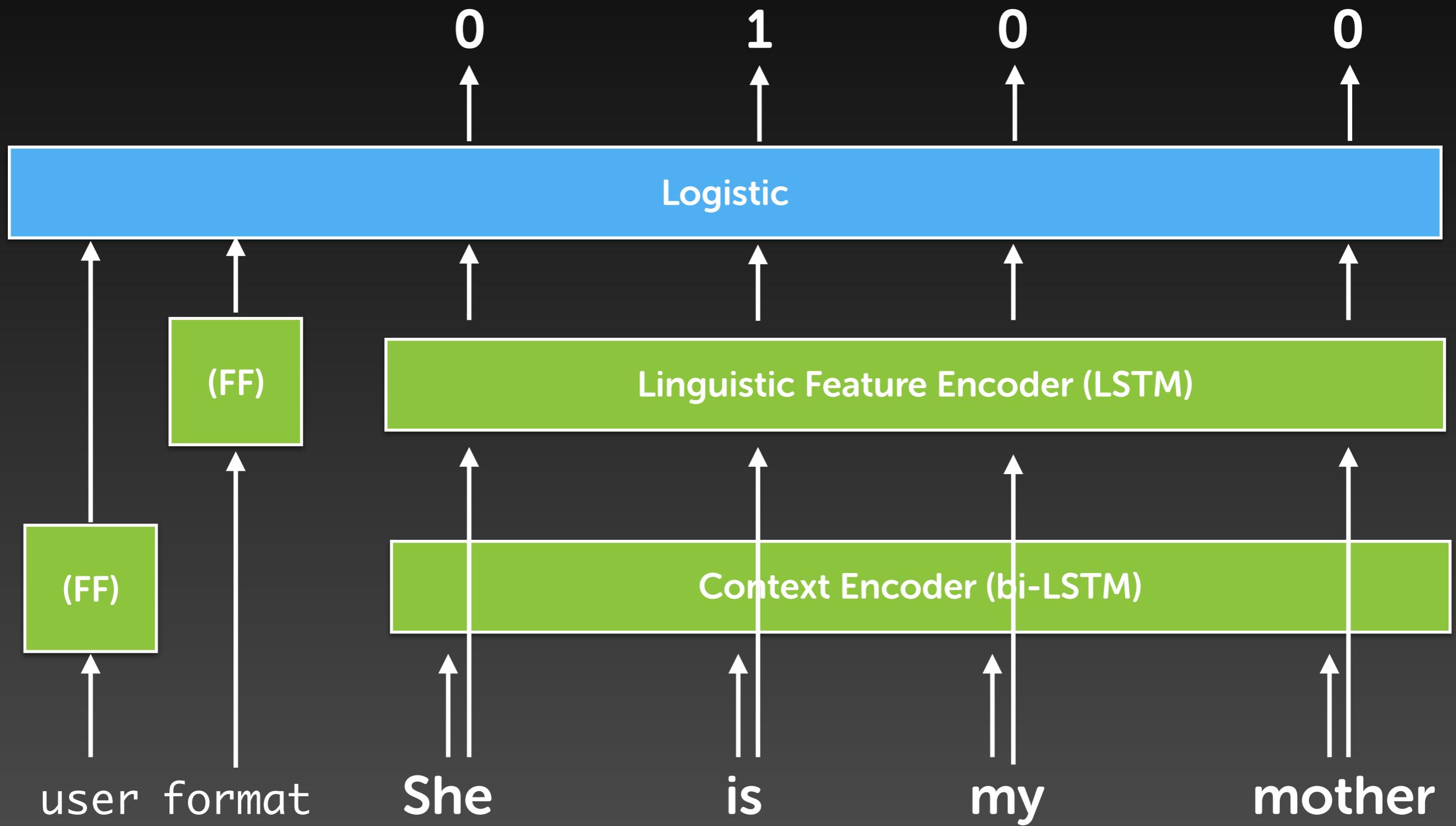
# #1. SanaLabs

(Osika et al., 2018)

- Combination of GBDT (Gradient Boosted Decision Tree) and RNN (LSTM)
  - GBDT – good for tabular data
  - RNN – good for sequential data
- Additional features (e.g., context, times seen)

# #2. singsound

(Xu et al., 2018)

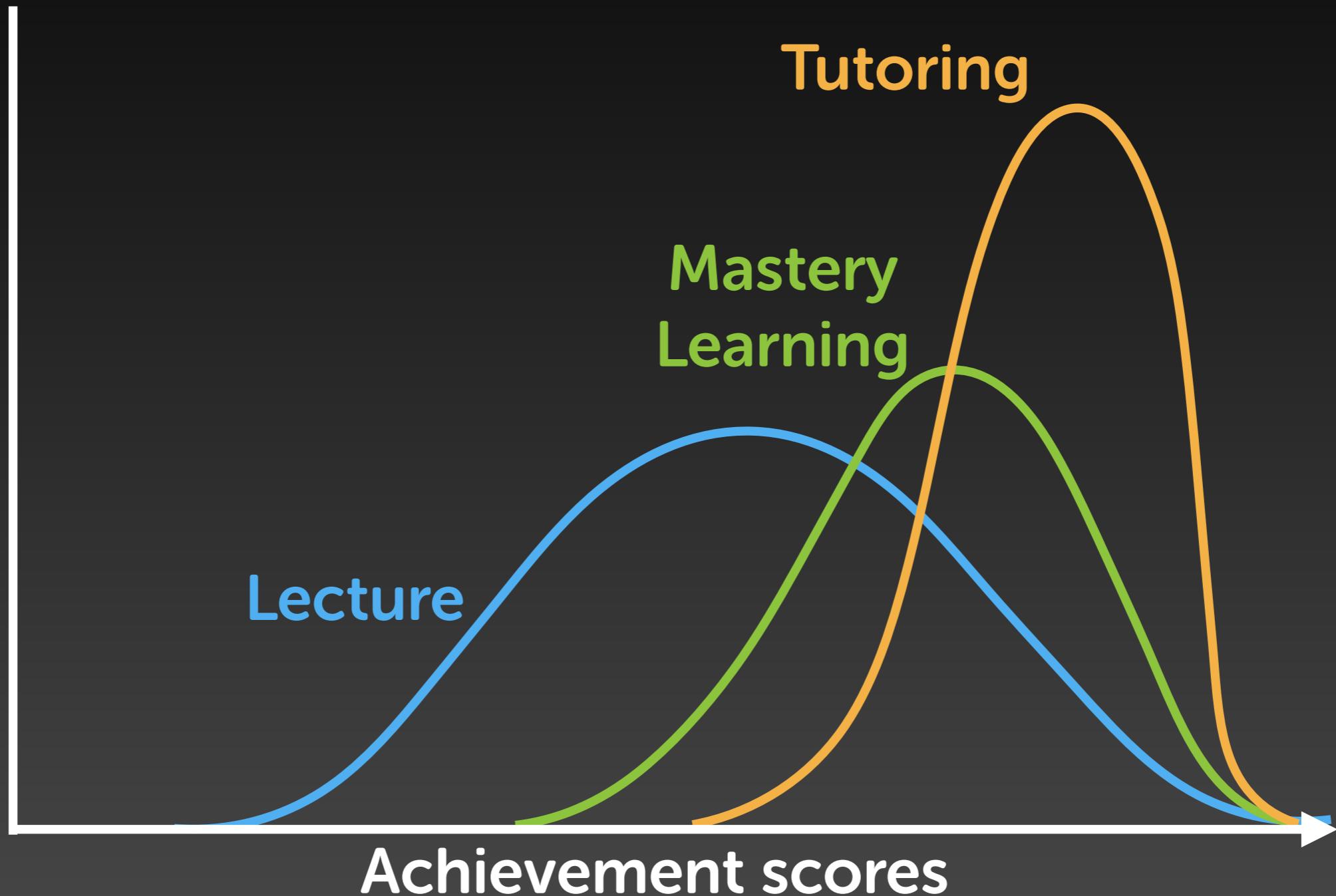


# Shared Task Takeaways

- Choice of learning algorithm > feature engineering
  - Sequence models (RNN)
  - Tree ensembles (GBDT)
  - Multitask learning
- Mostly beginners
  - vs more longitudinal, diverse user base

# Learning

# The 2 Sigma Problem



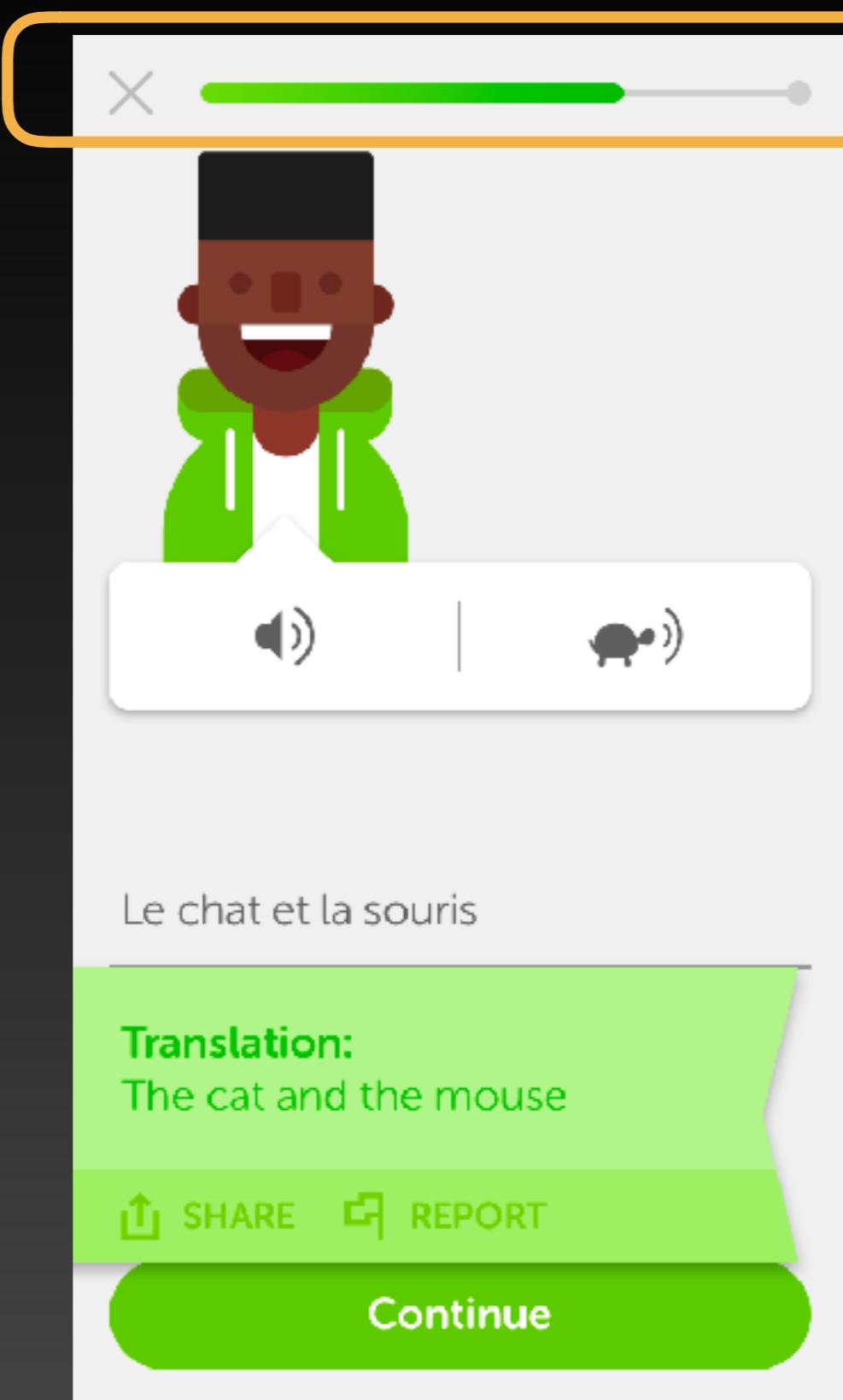
Source: (Bloom 1984)

# A Tale of Two Bars

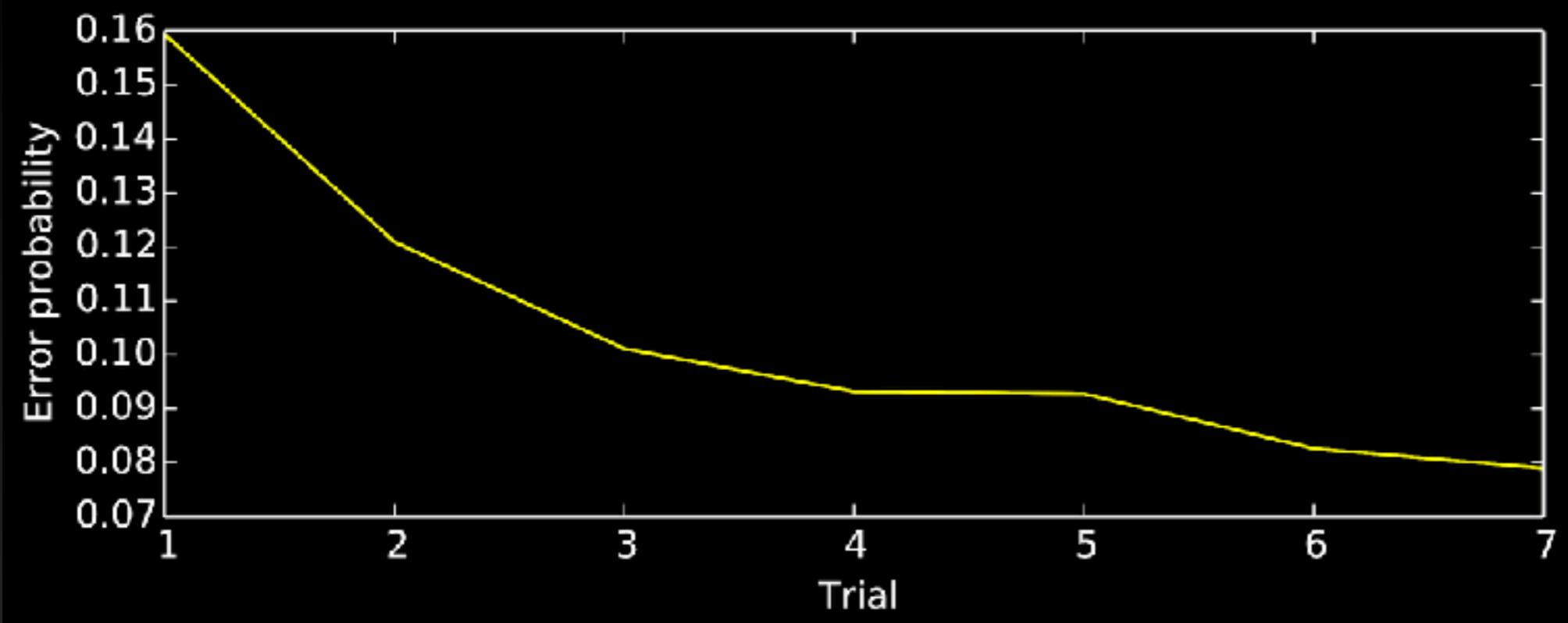
(or, Duolingo's answers to  
the two sigma problem)

## Progress Bar

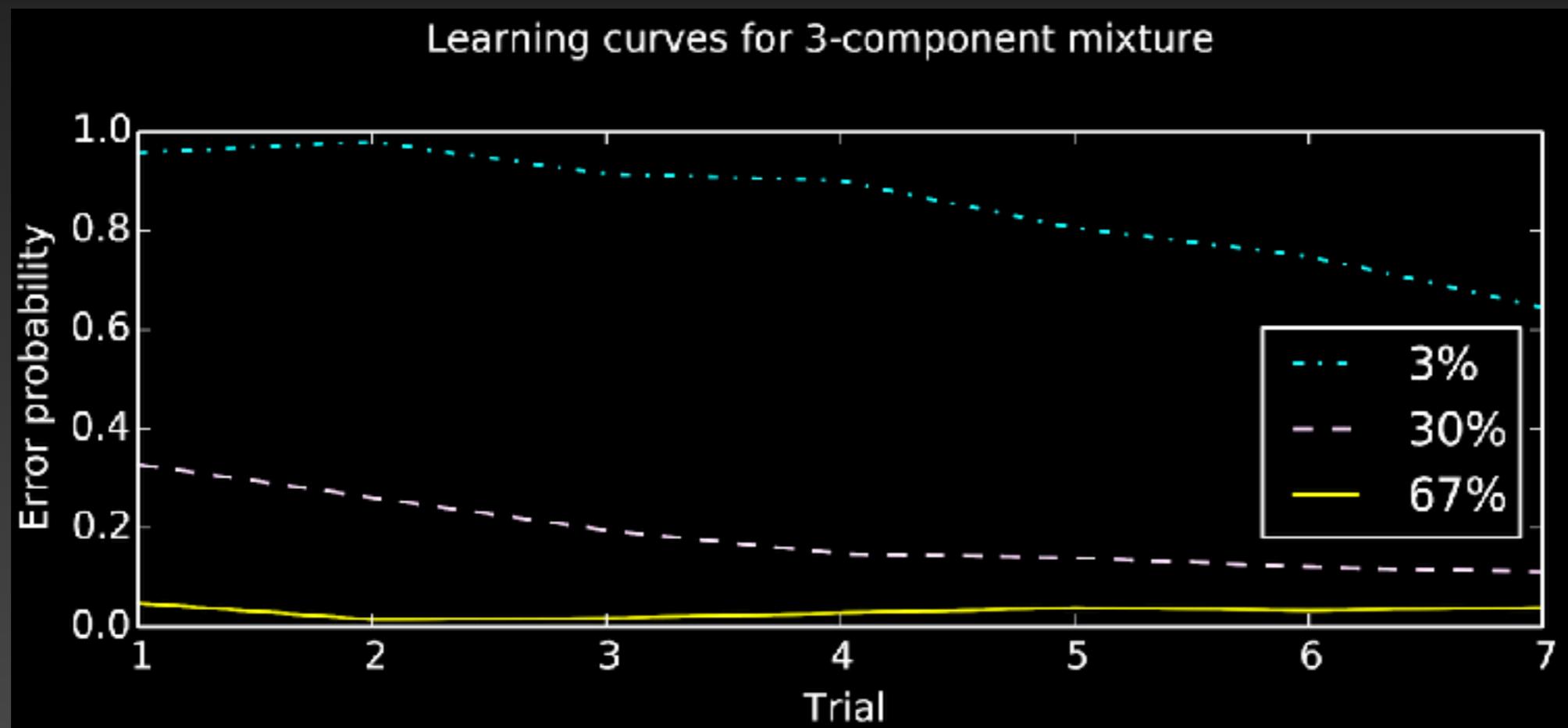
$$\prod_{j=1}^J p(c_j|v)$$



Population learning curve

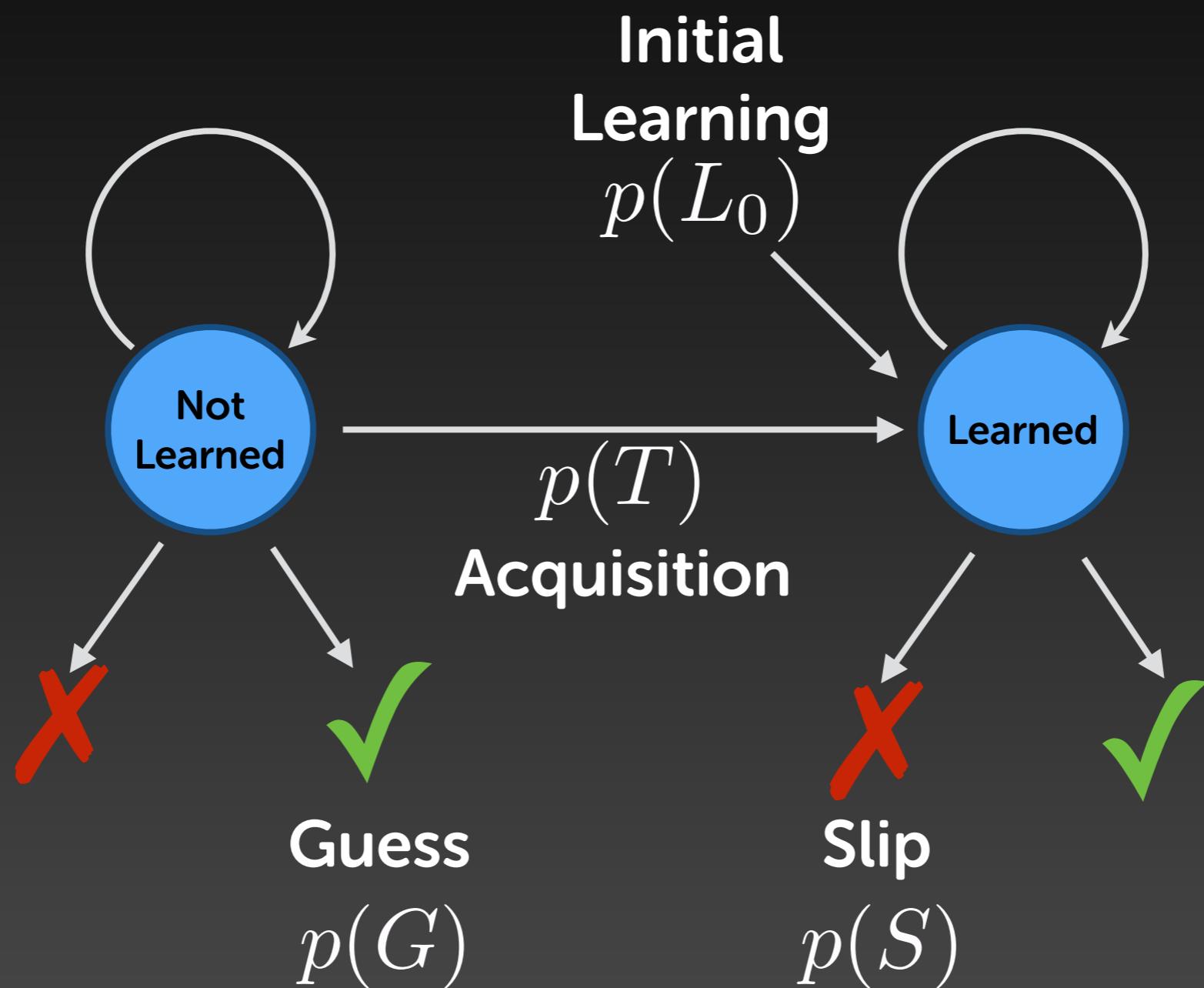


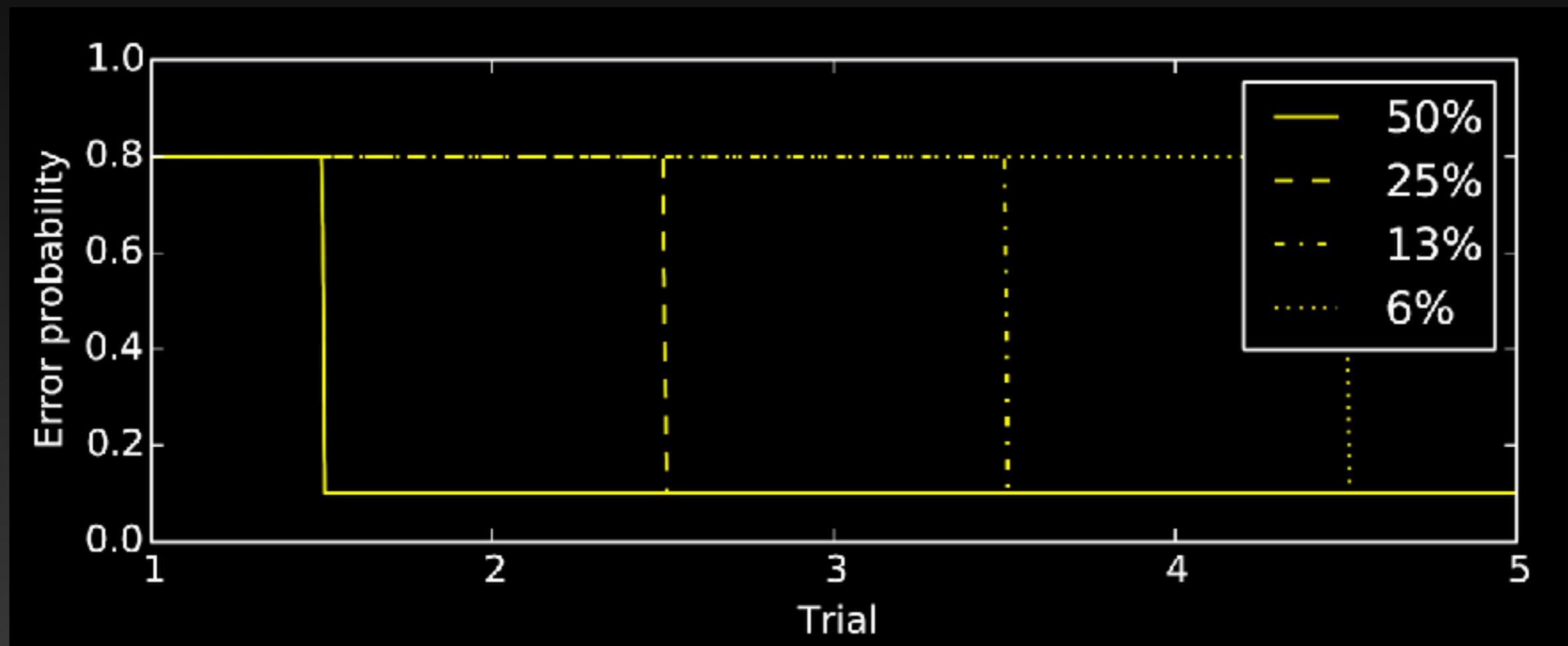
Learning curves for 3-component mixture



Source: (Streeter 2015) - Users learning es<-en on Duolingo, "una" on listen items

# Bayesian Knowledge Tracing





## Item Response Theory (1PL)

$$\theta_i + \beta_j$$

## Additive Factor Model [Cen et al. 2006]

$$\theta_i + \beta_j + \gamma_j T_{ijt}$$

## Performance Factor Model [Pavlik et al. 2009]

$$\theta_i + \beta_j + \alpha_j S_{ijt} + \rho_j F_{ijt}$$

$i$  student

$j$  problem/KC

$T_{ijt}$  number of practice opportunities

$S_{ijt}$  number of successes

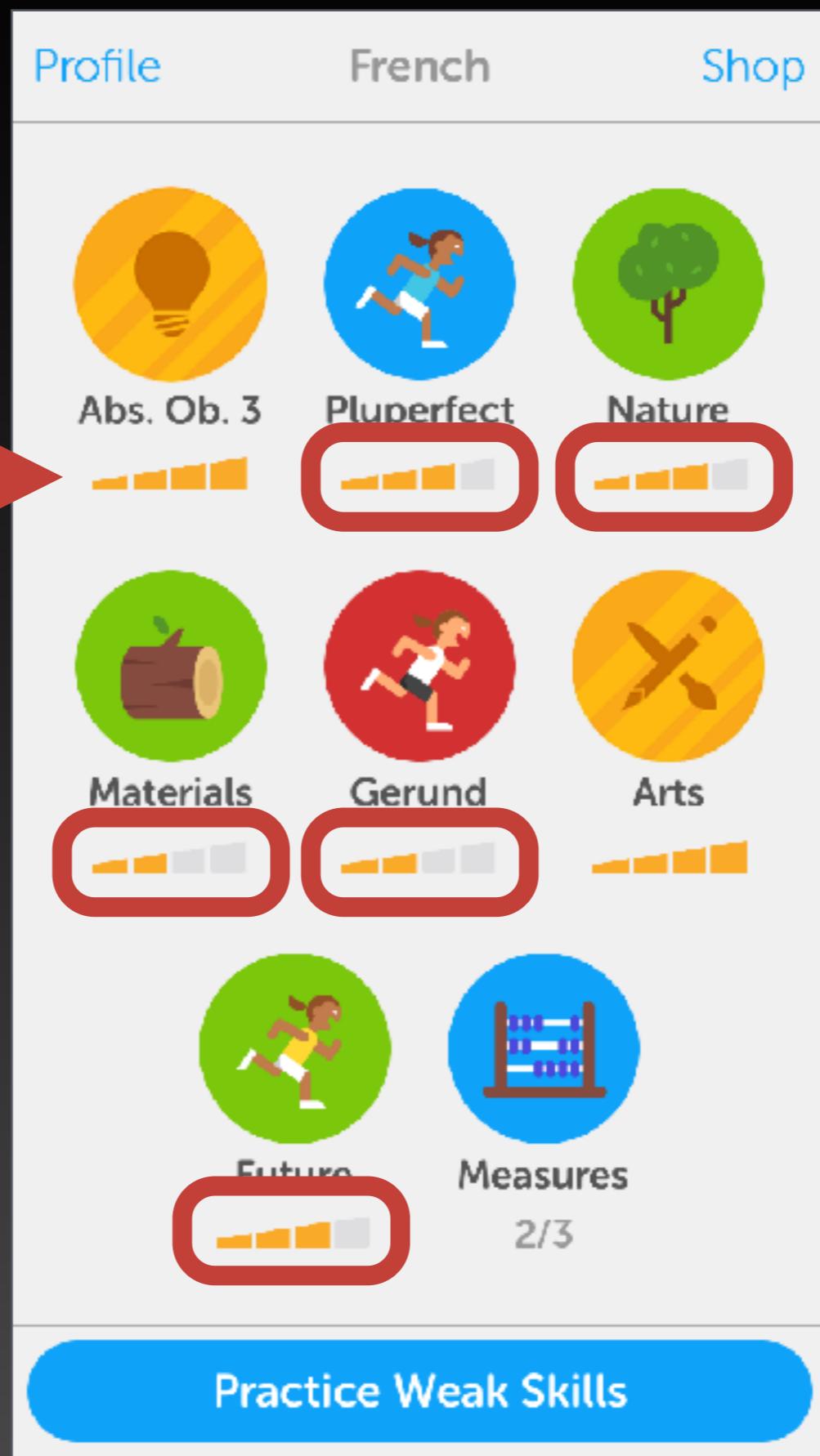
$F_{ijt}$  number of failures



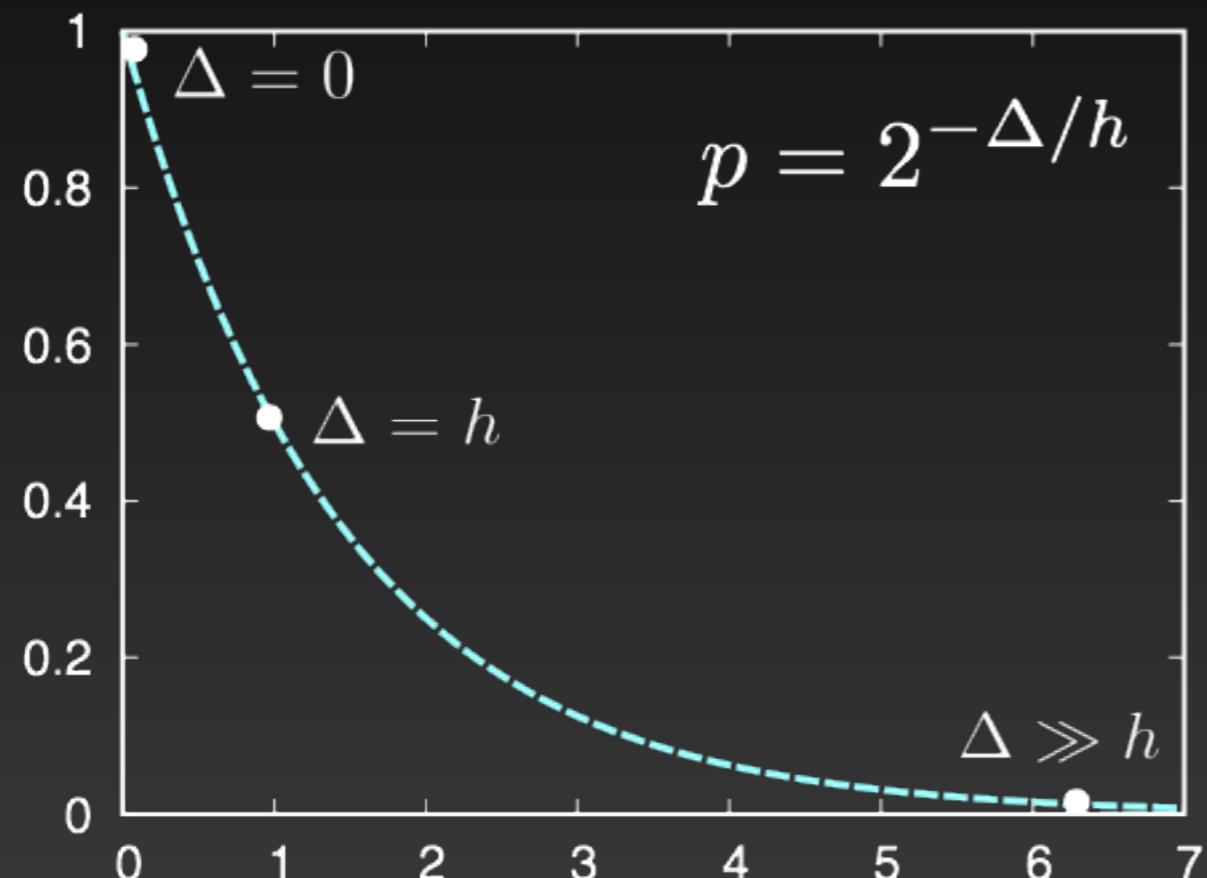
$$\sum_{k=1}^K p^k \prod_{t=1}^T \mathcal{B}(\sigma(\phi_{i,j,t} \cdot \beta^k), v_t)$$

Method	Training log loss	Test log loss	Training AUC loss	Test AUC loss
Knowledge Tracing	0.3429	0.3441	0.3406	0.3460
Performance Factors Analysis	0.3248	0.3285	0.2774	0.2865
Additive Factor Model	0.2869	0.3250	0.1629	0.2789
AFM Mixture (3 components)	<b>0.2818</b>	<b>0.3220</b>	<b>0.1598</b>	<b>0.2760</b>

## Strength Bar



# Forgetting Curve

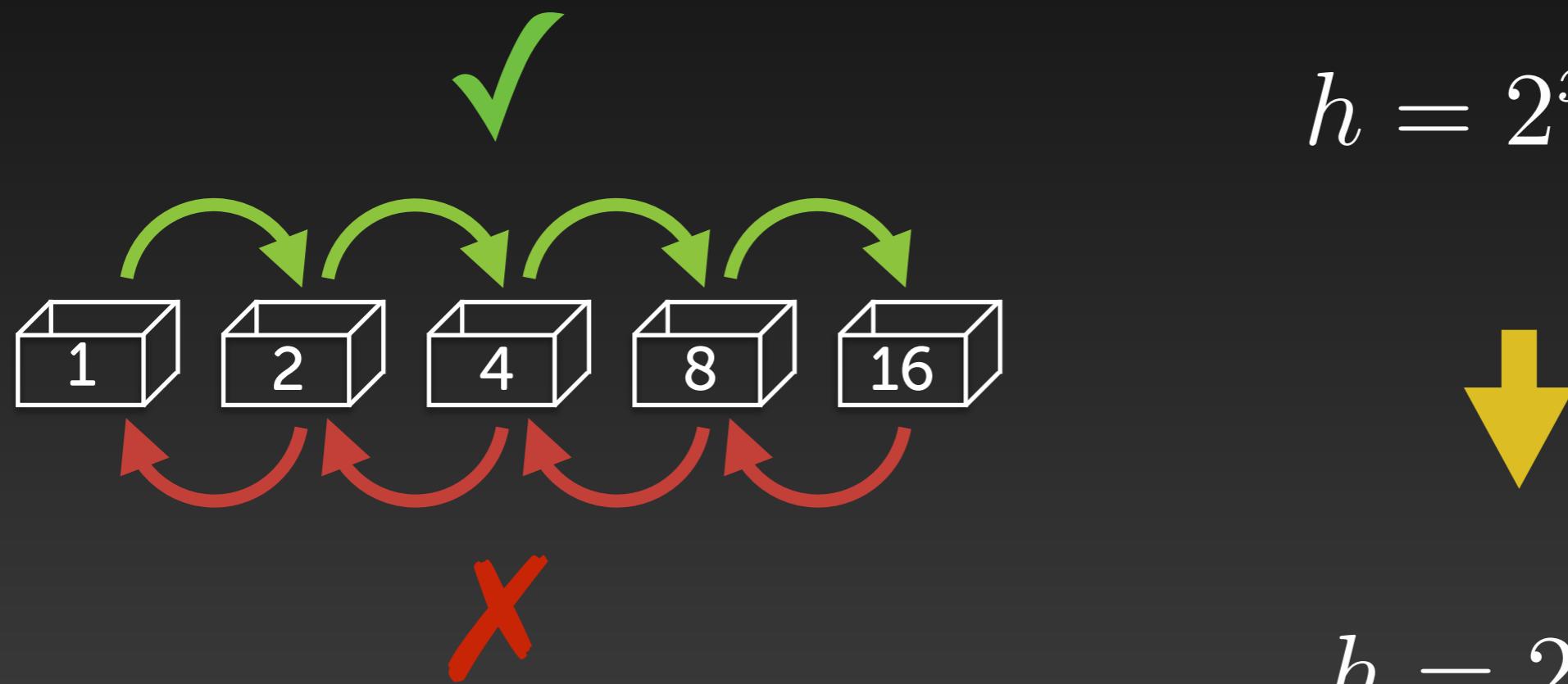


the **probability**  $p$  of a correct answer as a function of:

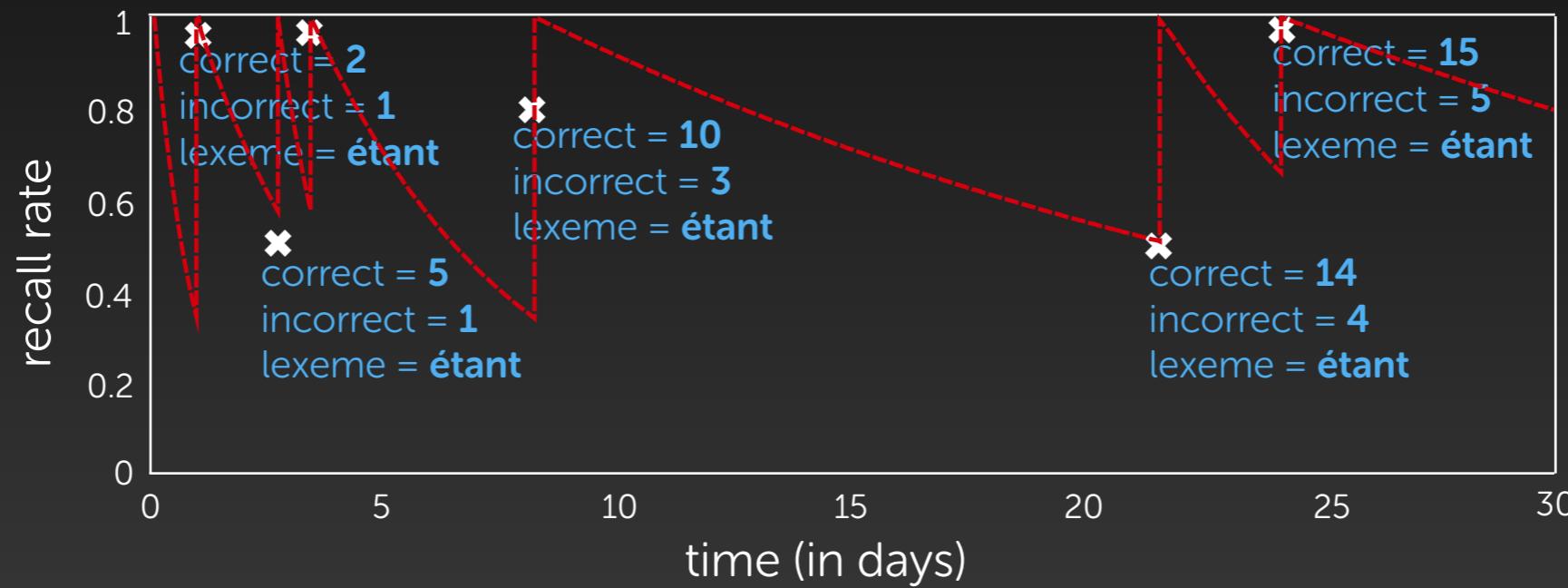
- **time**  $\Delta$  since it last practice
- **halflife**  $h$  in user's memory

# The Leitner System

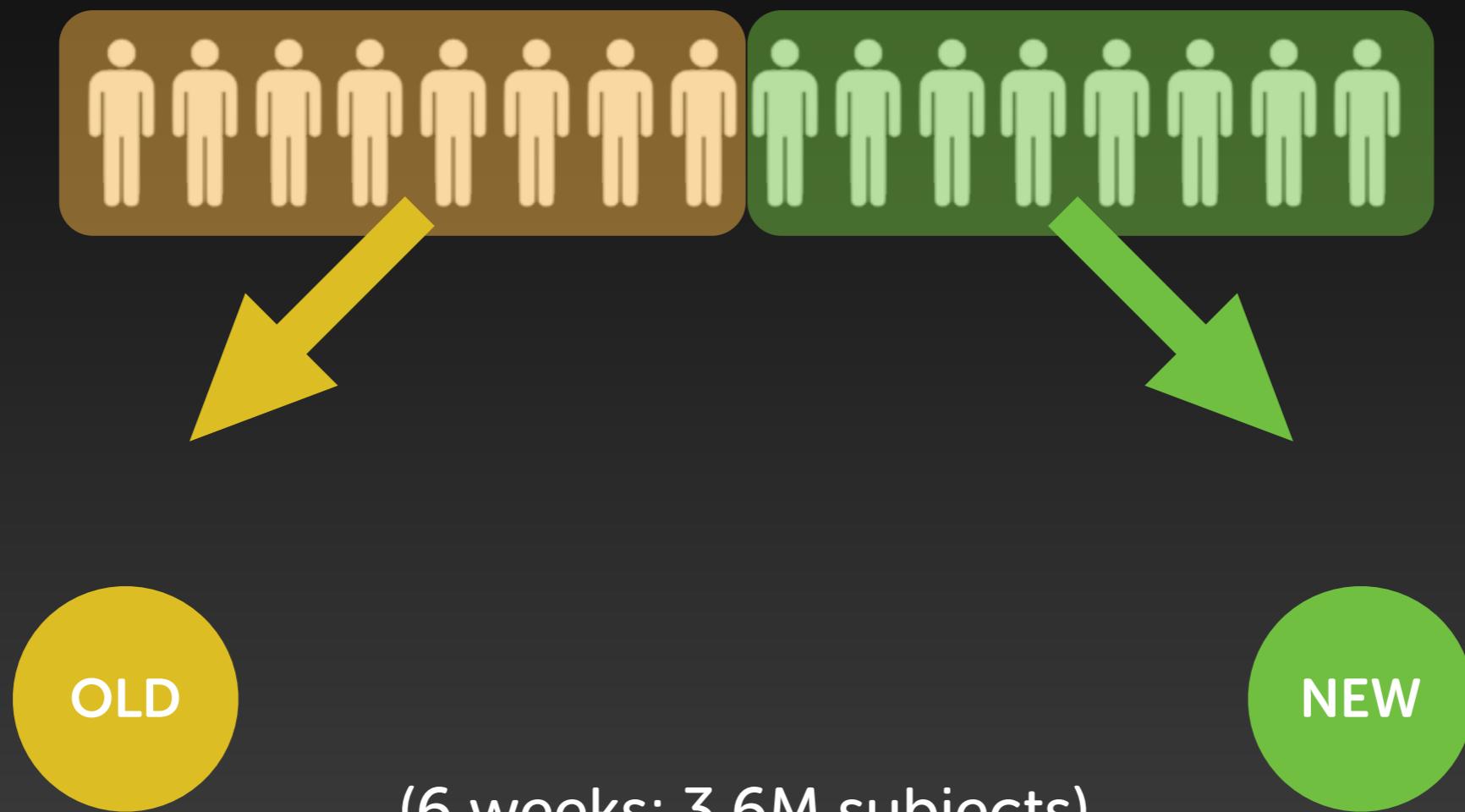
(Leitner, 1972)



# Half-life Regression (HLR)



# Results: User Experiment



all students  
who returned the next day:  
**+1.7% (p<0.001)**



"By working on something every day, recognising words and remembering grammar, I was improving my short-term memory. It gave me something to centre myself on so I didn't feel so lost in my day-to-day life."

— Shannon O'Neill

# Thanks!

<http://masatohagiwara.net/its2018.html>

... and we're hiring!