How Best to Learn Foreign Languages, and Anything

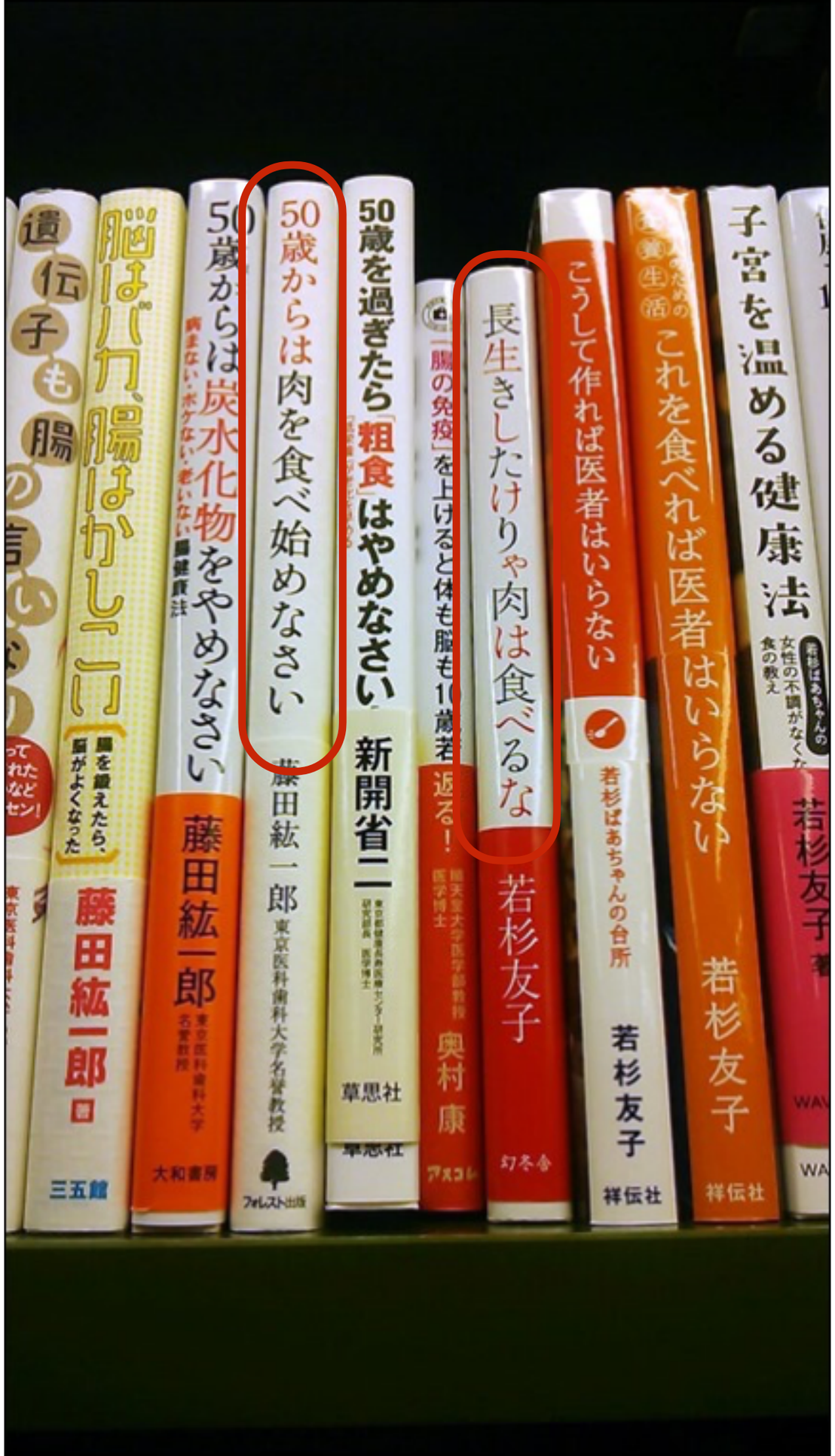
Duolingoの大規模データから導き出された効率的な言語習得法

Masato Hagiwara
萩原 正人

Software Engineer / Research Scientist


まず
白米をやめなさい!
溝口 徹
「糖質」のとりすぎが
老化と万病を引き起こす
●脳卒中 ●心臓病 ●高血圧
●認知症 ●ガン ●アトピー
●花粉症などのアレルギー ●うつ病
合格する親子は、
「塾」「家庭教師」をこう使っている
いつものパンがあなたを殺す
デイビッド・パールマター
クリスティン・ロバーグ[著]
白澤卓二[訳]
一番役に立つ「言い方」
小林弘幸
Don’t eat rice
Don’t eat bread

Eat meat after 50



Don’t eat meat to live longer

90% of English is ...

Grade 1-6

Grade 7-9

Paraphrases

Be/Do/Have

You can be fluent in ...

3 months

30 days

7 days

… or don’t study it!

Losing Weight
& Language Learning
ダイエットと外国語学習

- Requires consistent efforts + time
  継続的な努力と時間を要する

- Anyone healthy can succeed, yet most fail
  誰でも成功できるが、大部分が失敗に終わる

- Lack of knowledge of correct methodologies
  正しい方法に関する知識の不足

Second Language Acquisition (SLA)

第二言語習得

Dynamic Memory Models

Critical Period Hypothesis

Monitor Theory

Comprehensive Input Hypothesis

Interaction Hypothesis

Self-Regulation Theory

People learning English worldwide

世界中の英語学習者

1,500,000,000

(~20% of the world's population)

(世界人口の約20%)

(TESOL Report 2014)

duolingo

- More than **150 million registered users** worldwide

  登録ユーザー数：1億5000万人

- Currently offer **64 courses (teaching 21 distinct languages)**

  65コース (21言語)を学習可

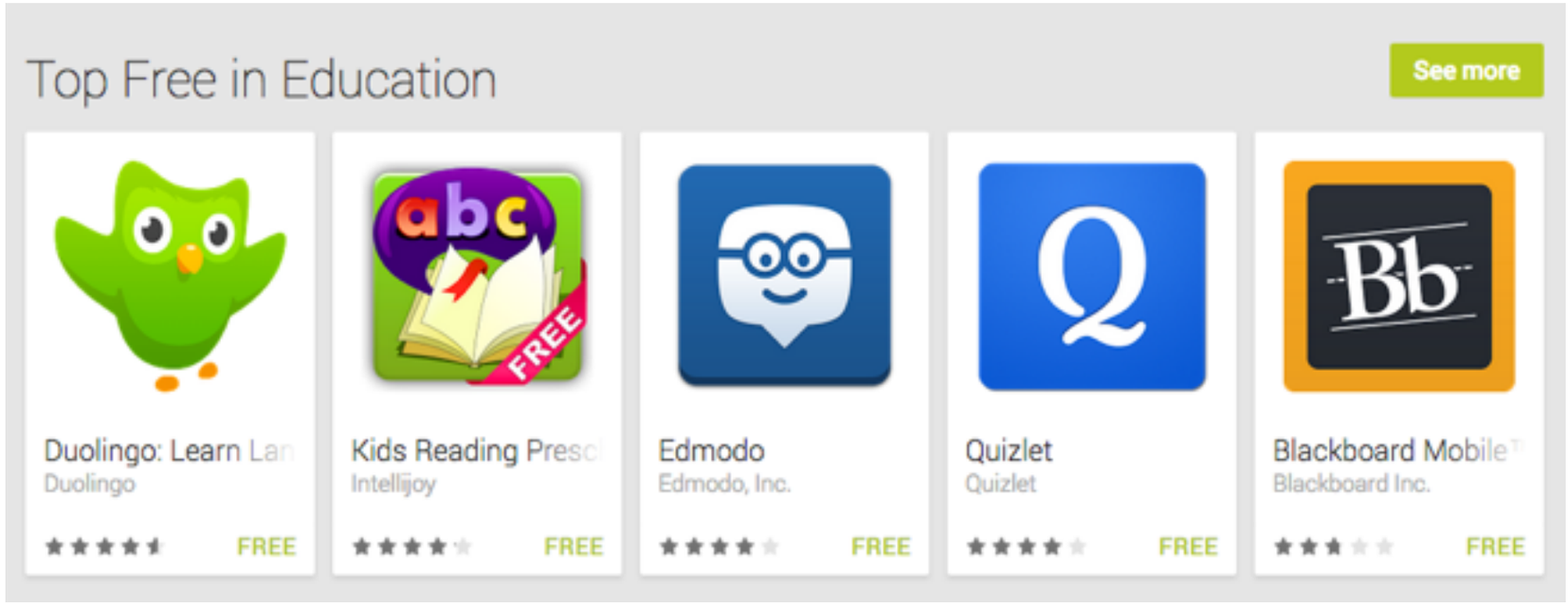
- Expanding to **87 courses soon (incl. Klingon!)**

  87コースに拡大予定（クリンゴン語含む）

- Available on **Android, iOS, Windows Phone & Web**
- 100% **FREE**　100%無料

(As of Sept. 2016)

Google Play

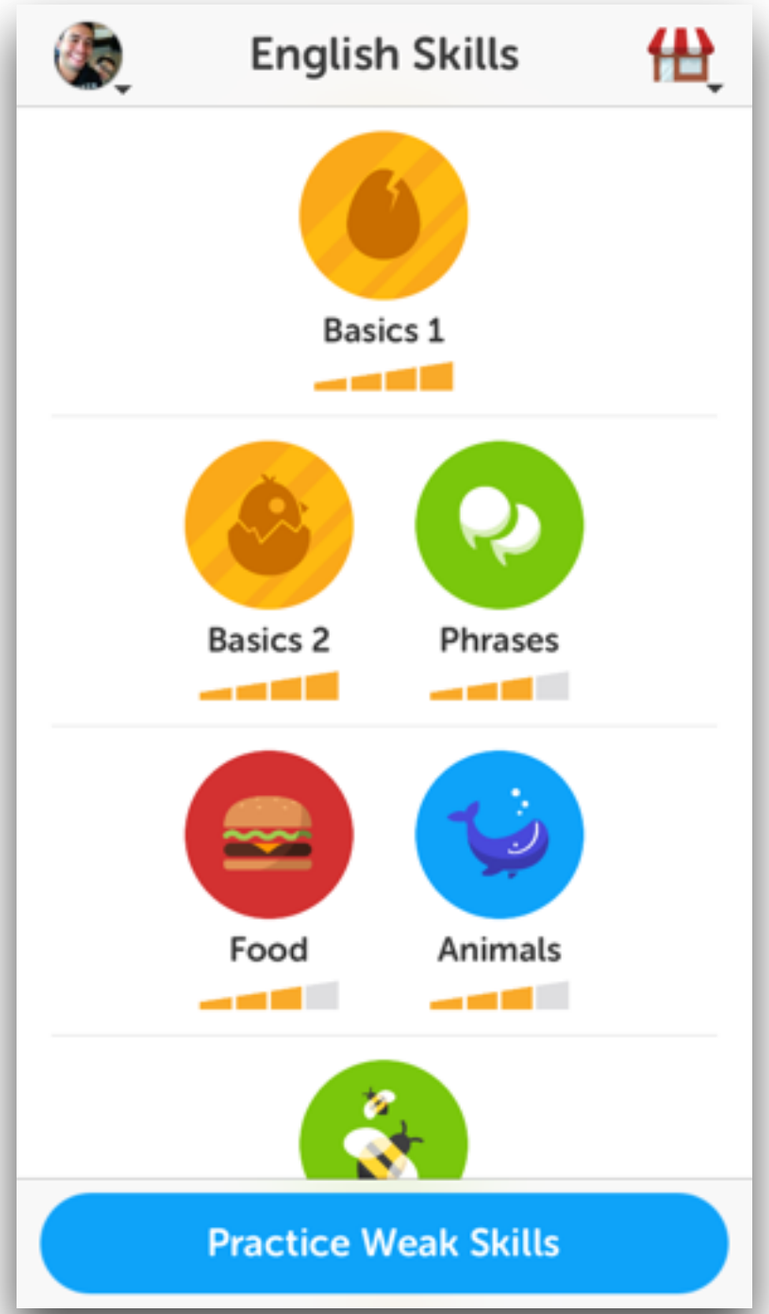


iOS App Store

GOOGLE
Best of the Best

APPLE
App of the Year

TECHCRUNCH
Education Startup of the Year



34 hours

時間

of Duolingo is
as effective as one
university semester

= 大学 1 学期相当

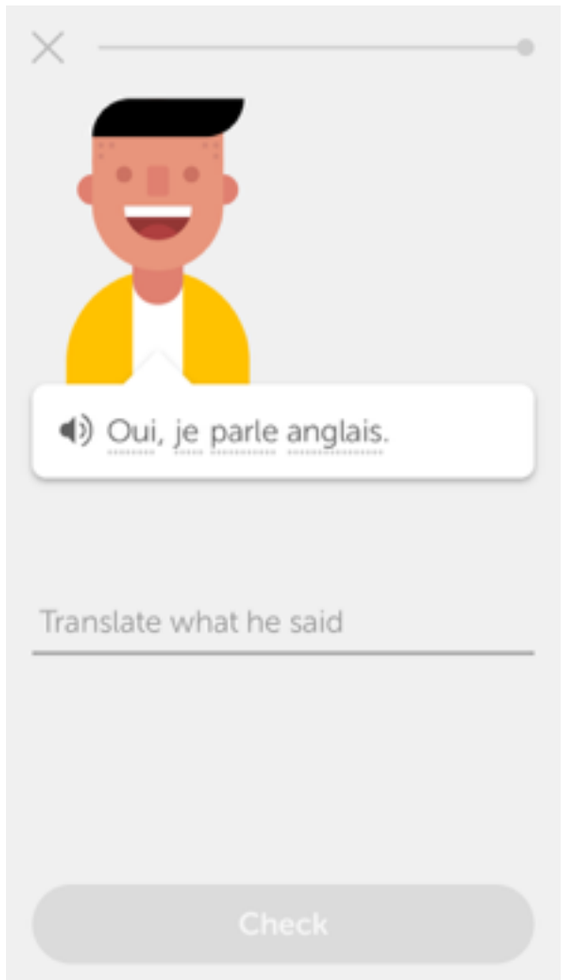
[Vesselinov & Grego, 2012]

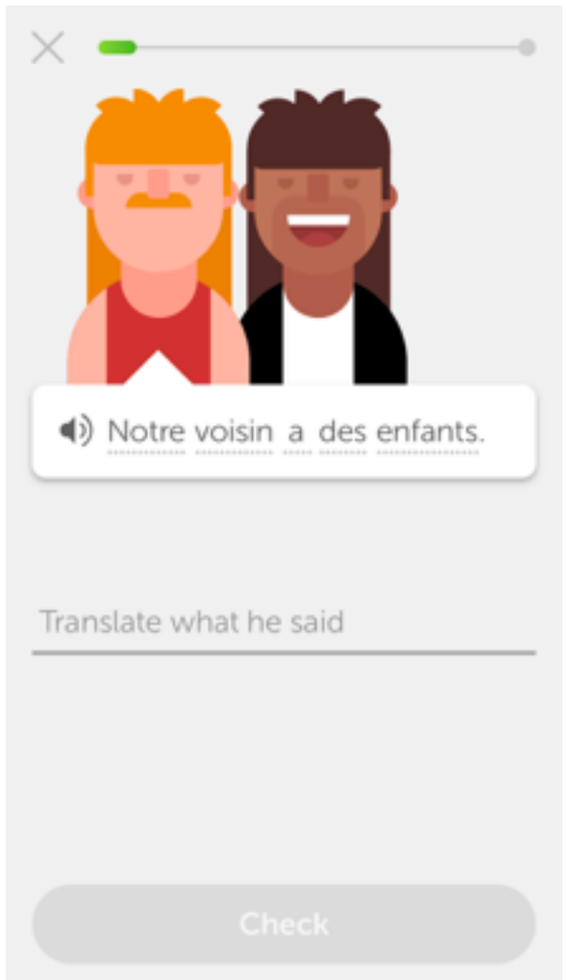
What makes learning languages difficult?

外国語習得の難しさの要因

Methodology
教授法

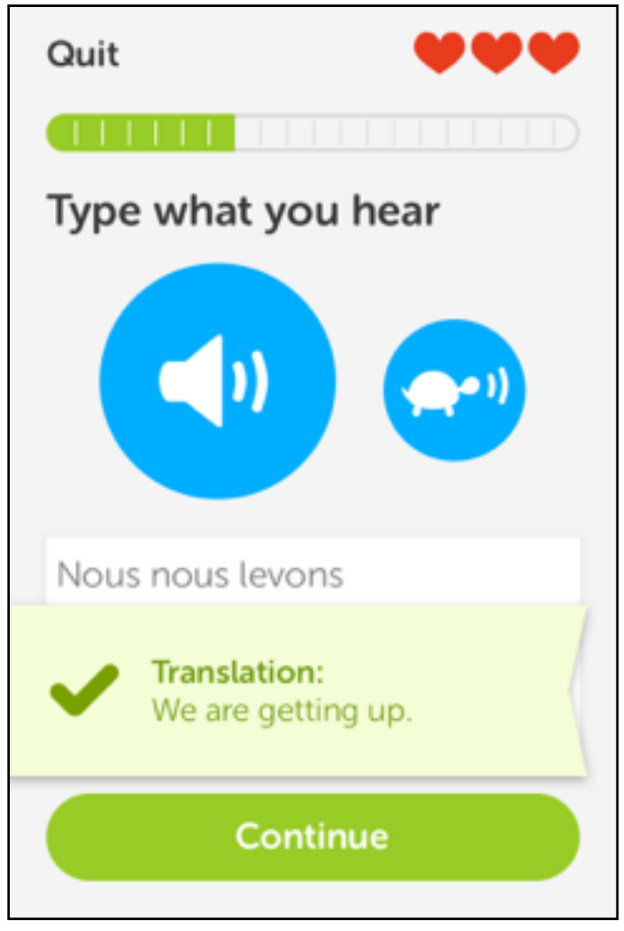
- Providing context: Teaching through whole sentences / pictures　文や写真に基づく問題

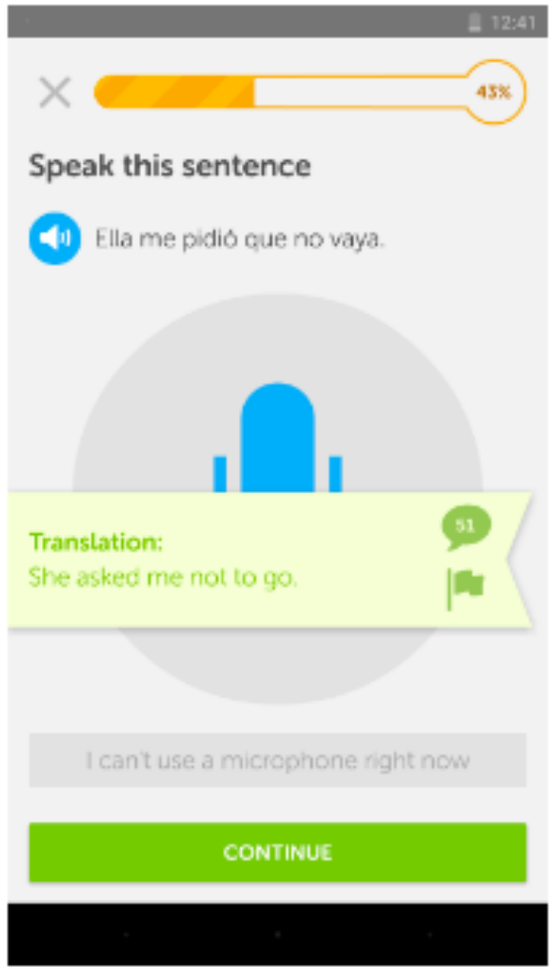




Methodology
教授法

- Listening and speaking　リスニング＆スピーキング



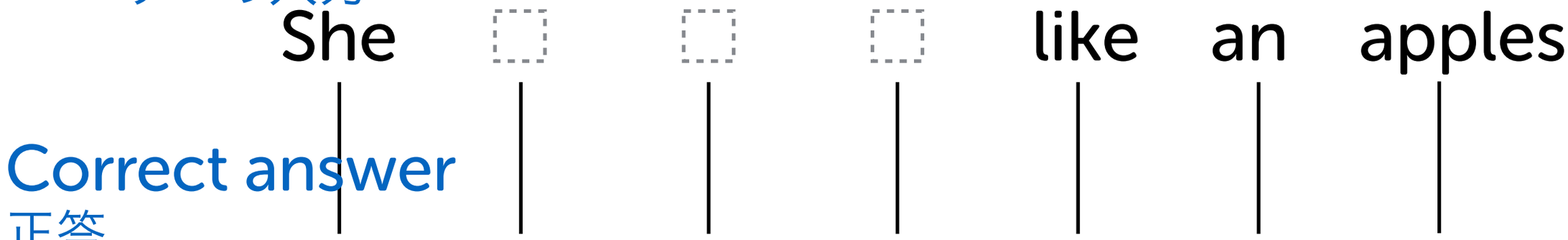


Learner Error Labeling

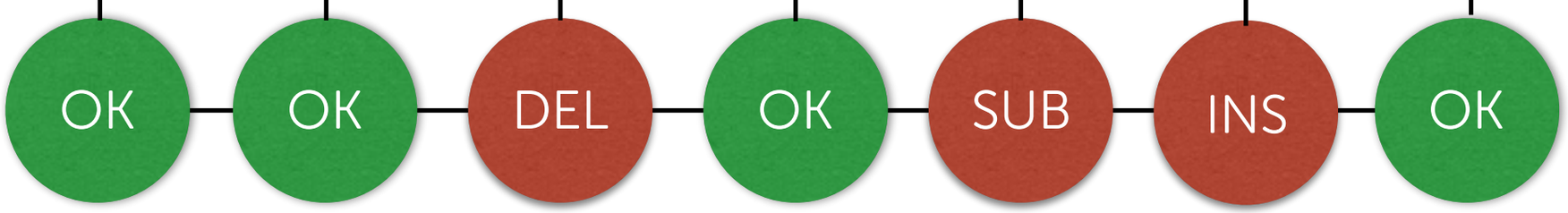
学習者誤りラベリング

User submission
ユーザーの入力

Correct answer
正答



Error labels
誤りラベル



CRF (Conditional Random Field)

$$p(\mathbf{y}|\mathbf{x}) \propto \exp(\mathbf{w}^T \phi(\mathbf{y}, \mathbf{x}))$$

Learner Error Prediction

学習者誤り推定

New sentence
正答

He ☐ often ☐ eats ☐ bananas

Label prediction
推定ラベル

? ? ? ? ? ? ?

Features
素性

{ w=he
p=PRON
m:Gender=Masc
m:Number=Sing
m:Person=3
d=nsubj->VERB }

{ w=eat
p=VERB
m:Number=Sing
m:Person=3
m:Tense=Pres
d=ROOT }

Words vs Difficulties

単語の難易度

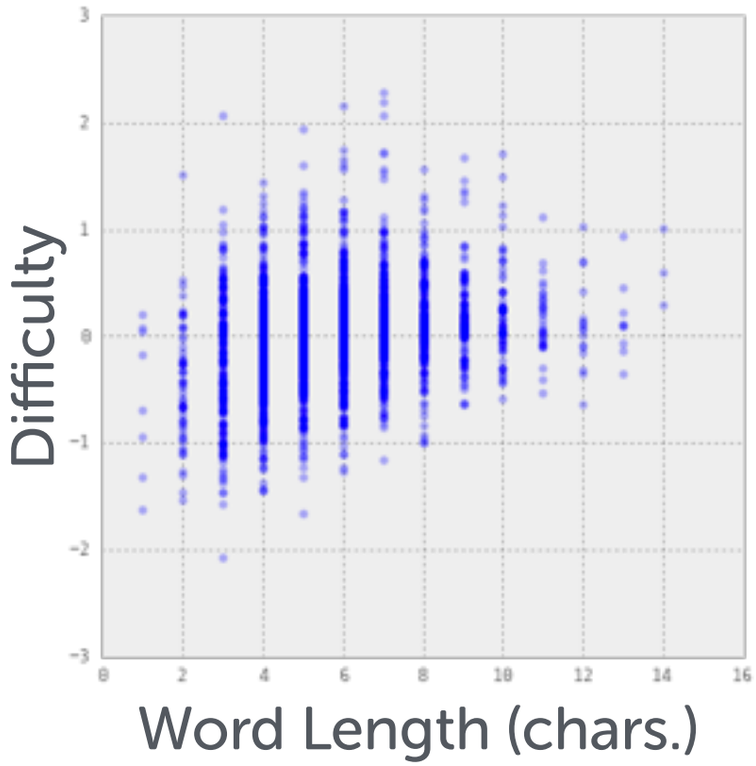
English		Spanish		French	
Easy	Hard	Easy	Hard	Easy	Hard
yes	Definitely	A	ho	simple	est-ce
table	Possibly	Y	Calcetines	combien	dix-sept
top	ceiling	casa	volvió	a	sont-ils
is	Seconds	no	Definitivamente	il	Voulez-vous
and	Almost	ayer	Oímos	nuit	soirée
‘s	Think	local	Cuándo	orange	Es-tu
old	Summer	es	Jamás	ici	Est-ce
tuna	Clearly	antes	Posiblement	animal	Dimanche
open	Portuguese	nota	Cómo	mal	Veux-tu
Pink	Private	hotel	bolígrafos	tard	Quoi

* Difficulty of word x = Likelihood of substitution on x = Weight of w[0]=x —> SUB

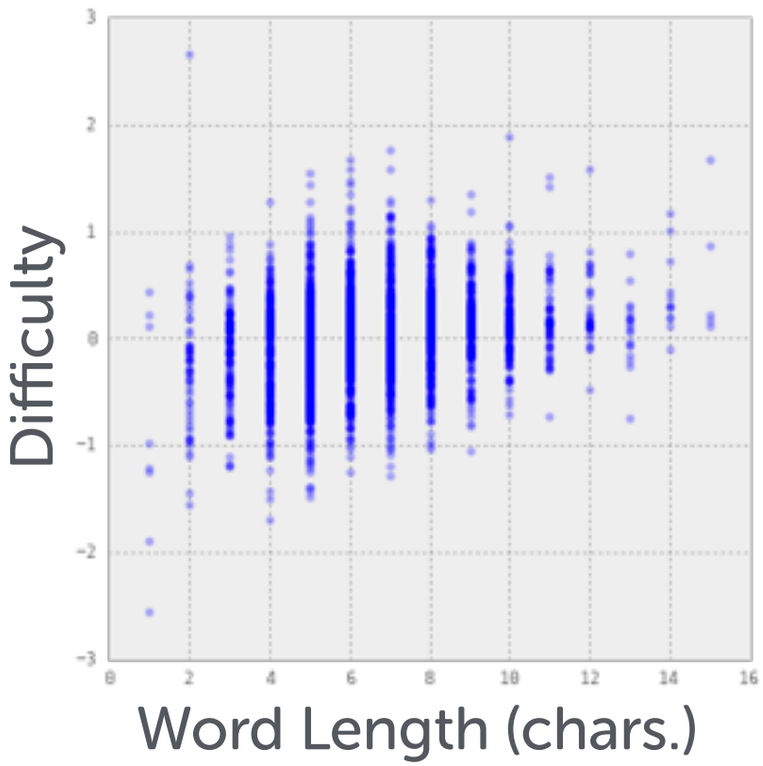
Word Lengths vs Difficulty

単語長と難易度

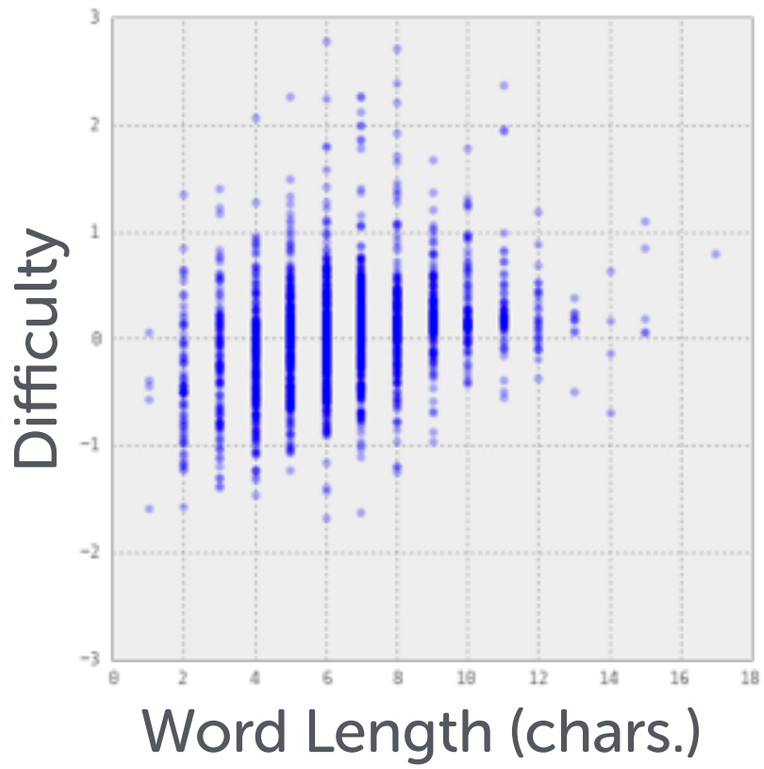
English



Spanish



French



r = 0.25 (p < 0.01)　　r = 0.24 (p < 0.01)　　**r = 0.32 (p < 0.01)**

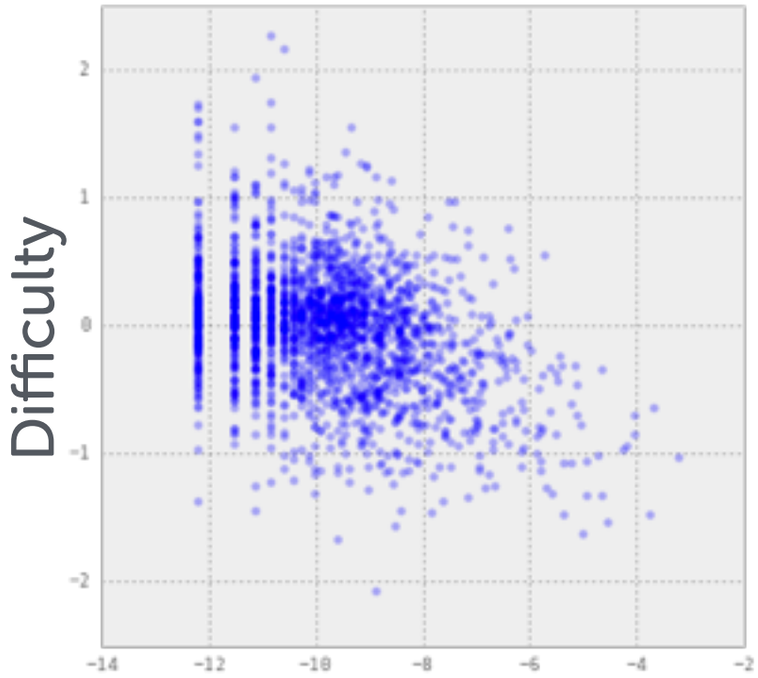
French word difficulty is most sensitive to length

フランス語の単語難易度＝単語長と相関が最も高い

Word Frequency vs Difficulty

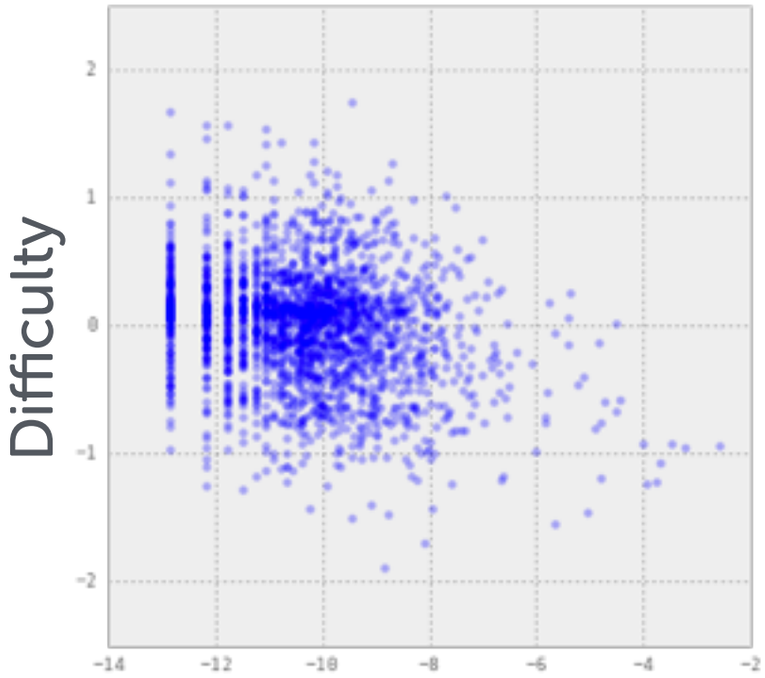
単語頻度と難易度

English



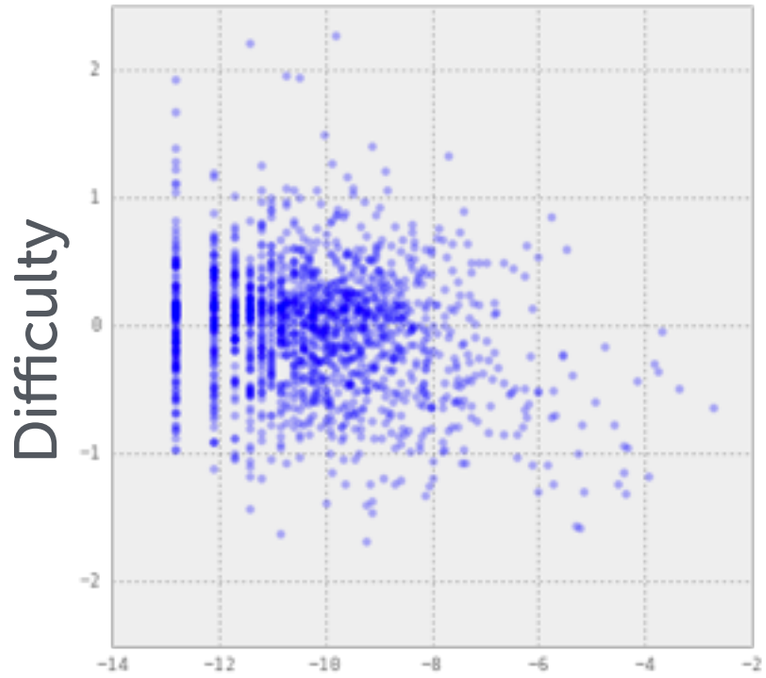
r = -0.29 (p < 0.01)

Spanish



r = -0.18 (p < 0.01)

French



r = -0.15 (p < 0.01)

English word difficulty is most sensitive to frequency

英語の単語難易度＝単語頻度と相関が最も高い

Character n-grams vs Listening Difficulties

文字nグラムとリスニングの難易度

English		Spanish		French	
Easy	Hard	Easy	Hard	Easy	Hard
pin	ma$	^a$	igu	^ea	tai
leg	gn$	ue$	^és	ge$	auv
^ye	als	on$	ést	lez	uva
six	rop	not	^oí	ix$	laq
nu$	nig	^ru	nam	ans	ère
cre	ly$	^no	^oi	^a$	eue
mal	bar	ija	oig	cas	enn
ted	ge$	se$	sig	arm	ele
^da	wom	pa$	sde	ust	vas
ta$	eke	^fr	esd	mon	^dé

Typical Errors by Native Language

第一言語別の頻出誤り

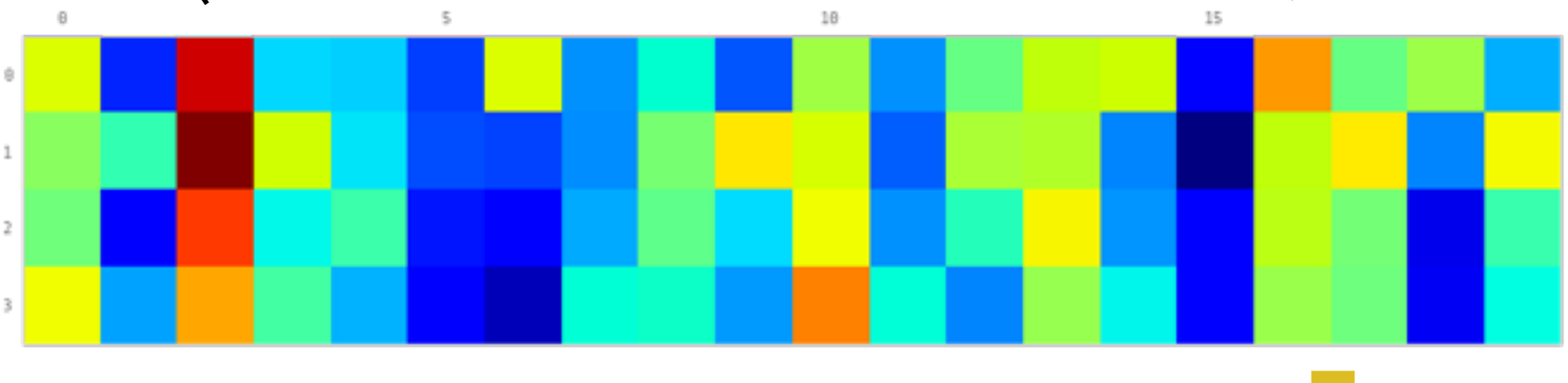
- **Portuguese**
  - *We **drinks** wine. (verb conjugation ‘drink’)
  - *I was not _ teacher. (article ‘a’)
- **Spanish**
  - *_ Need a bathroom. (subject ‘I’)
  - * She **not is** my girlfriend. (negation word order)
- **Arabic**
  - *Where **they are** from? (word order ‘are they’)
  - *The oil _ black. (be-verb ‘is’)
- **Russian**
  - *Our work week starts **in** Monday. (Preposition ‘on’)
  - *My father is _ farmer. (Article ‘a’)

Effect of Native Language

第一言語の影響

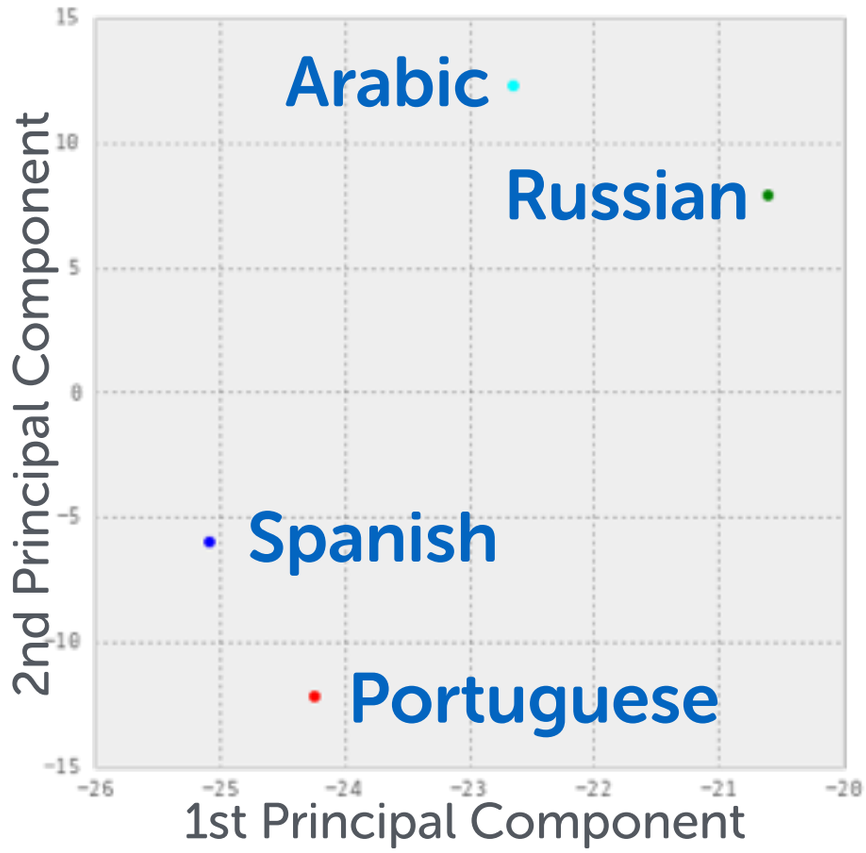
w=‘a’ —> DEL　w=‘his’ —> DEL　d=nmod —> SUB

English <- Portuguese
English <- Spanish
English <- Arabic
English <- Russian



...

SVD
(Singular Value Decomposition)
特異値分解



How about Japanese?

日本語話者の英語

Difficult Words for Japanese

日本語話者にとっての難易度（単語・文字nグラム）

Words		Character n-grams	
Easy	Hard	Easy	Hard
are	research	ges	bye
days	mouth	fiv	ye$
and	citizen	las	fly
for	On	do$	mou
One	firm	^i$	ste
cats	piece	mar	uth
ten	Impossible	epo	lde
To	unit	by$	edi
condition	ill	to$	ene
not	Walk	^it	ket

Difficult Concepts for Japanese

日本語話者にとっての難易度（文法）

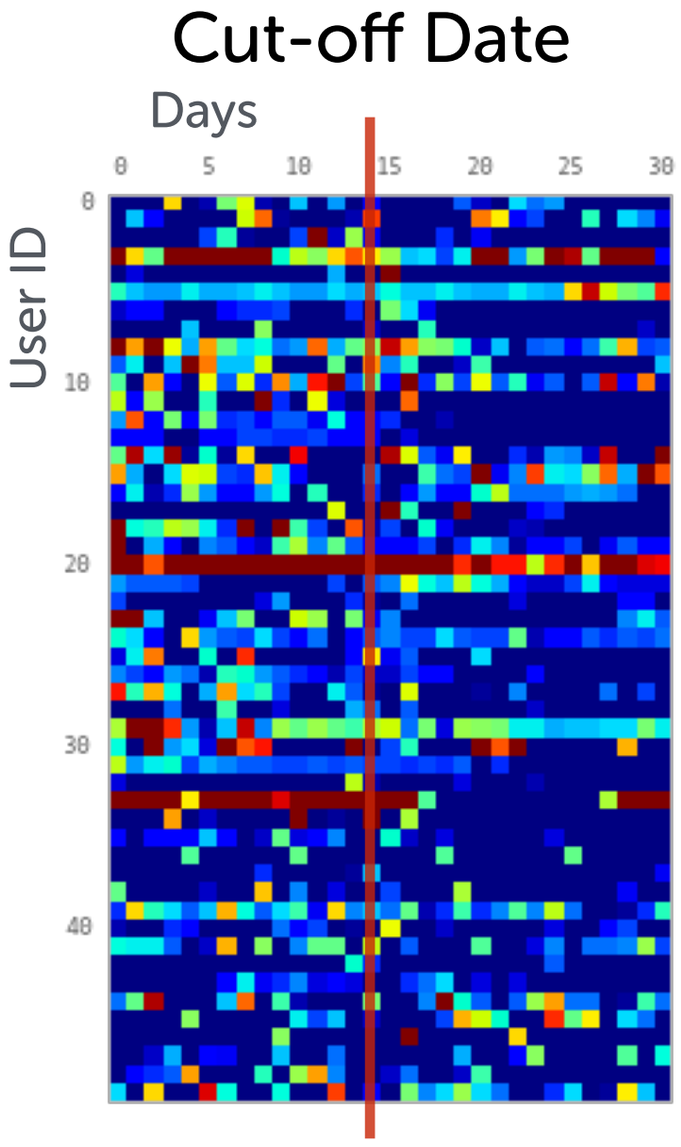
POS (Part-of-Speech)		Morphological Tags	
Easy	**Hard**	**Easy**	**Hard**
CONJ	NOUN	Degree=Cmp	Number=Sing
PUNCT	VERB	Case=Acc	Reflex=Yes
X	INTJ	Gender=Fem	VerbForm=Fin
AUX	ADV	PronType=Int	PronType=Prs
SCONJ	ADJ	Definite=Ind	Degree=Pos
NUM	ADP	Mood=Imp	PronType=Rel
PROPN	PART	NumType=Card	Mood=Ind
DET	PRON	Person=1	Tense=Pres
		VerbForm=Ger	Number=Plur
		Case=Nom	Person=3

What makes users successful in learning languages?

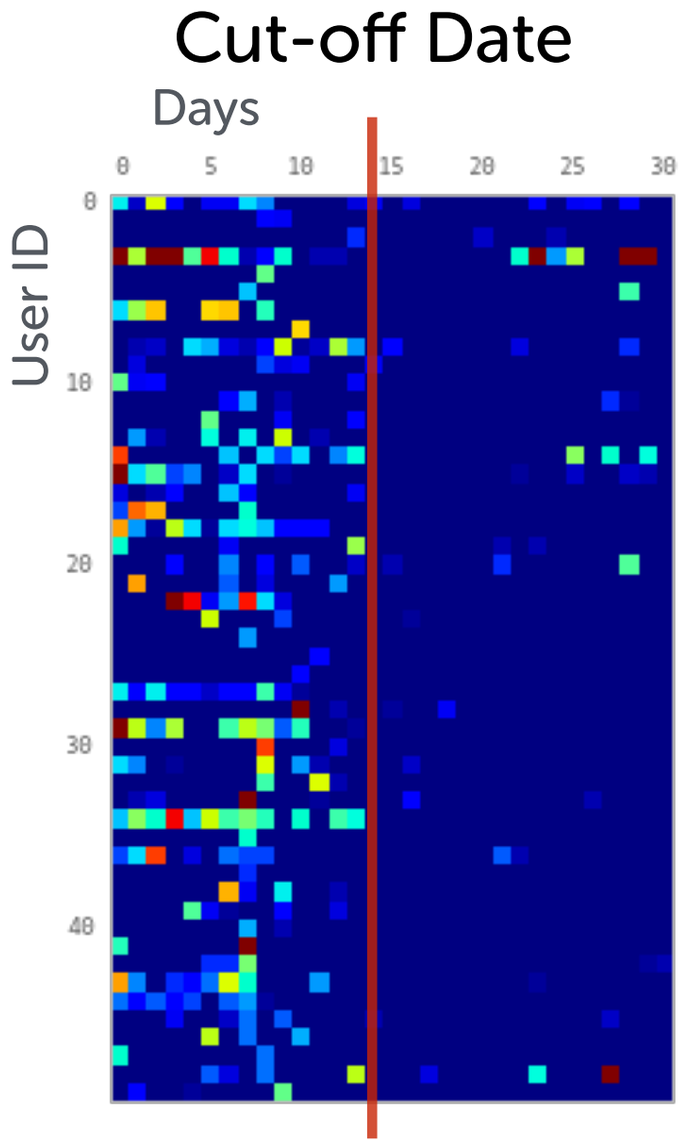
外国語学習に成功する人の特徴

Retained vs Churned Users

継続・離脱ユーザー



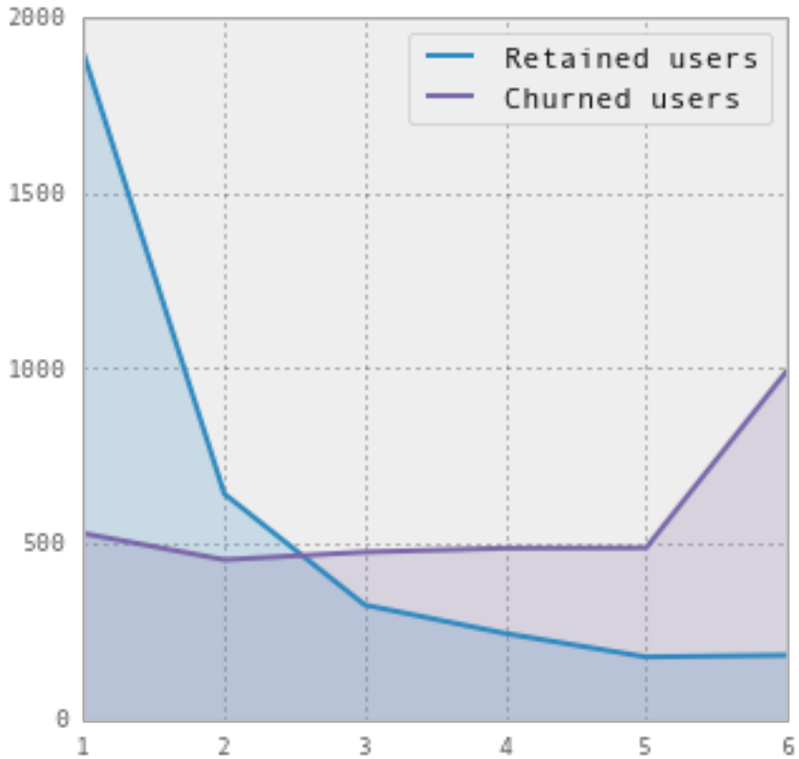
3,466 retained users



3,426 churned users

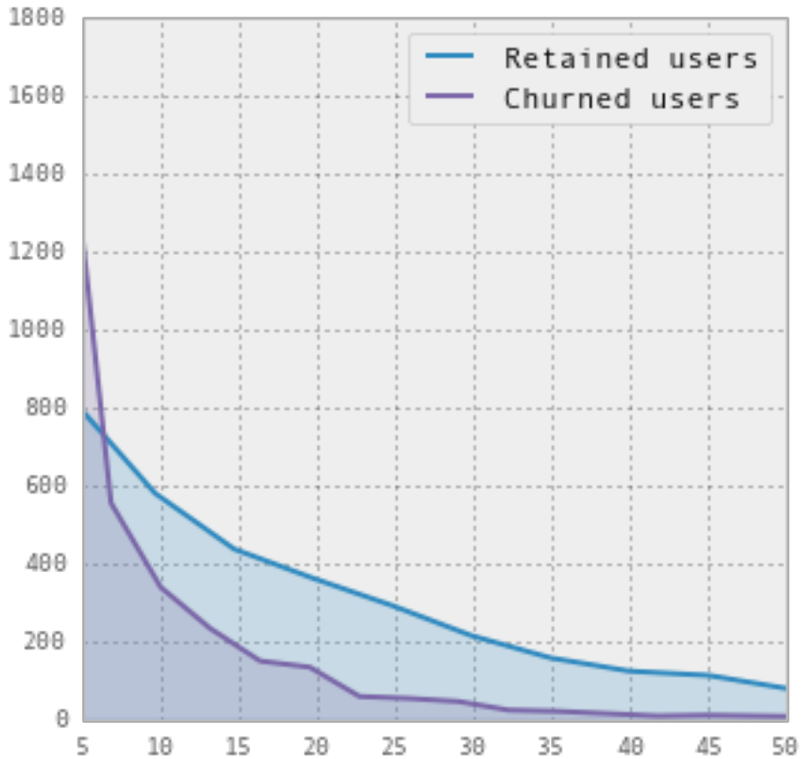
Retained vs Churned Users

継続・離脱ユーザー



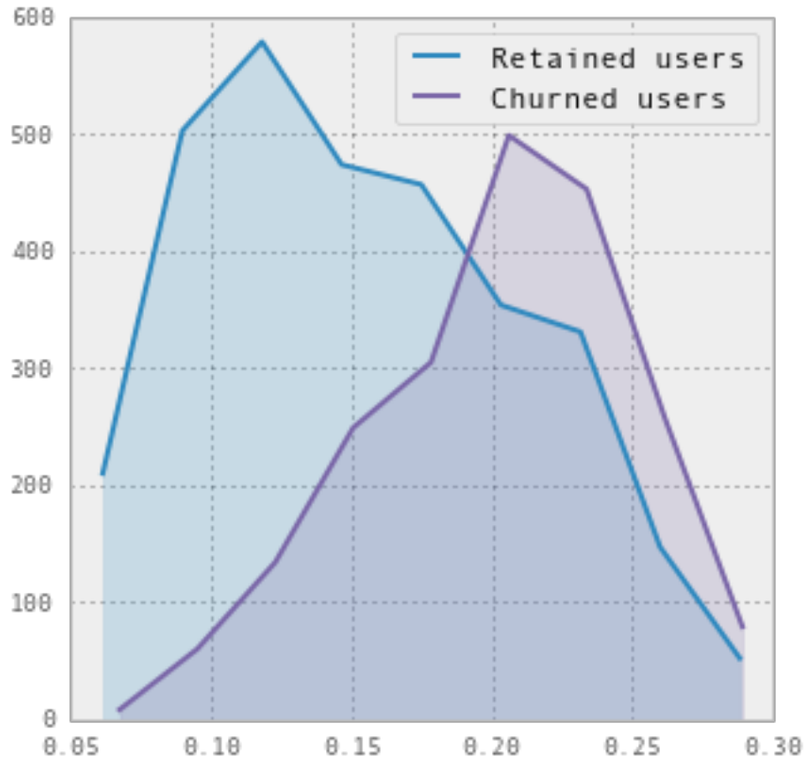
# of days since last session

最終セッションからの経過日数



# of sessions (last weeks)

セッション数/週



Relative stdev of daily # of sessions

セッション数/日の相対標準偏差

Other characteristics of successful user

成功する人の他の特徴

- Has already registered (reminder, etc.)
  ユーザー登録済み (リマインダー効果等)

- Has an observer (a teacher)
  オブザーバー（先生）による進捗管理

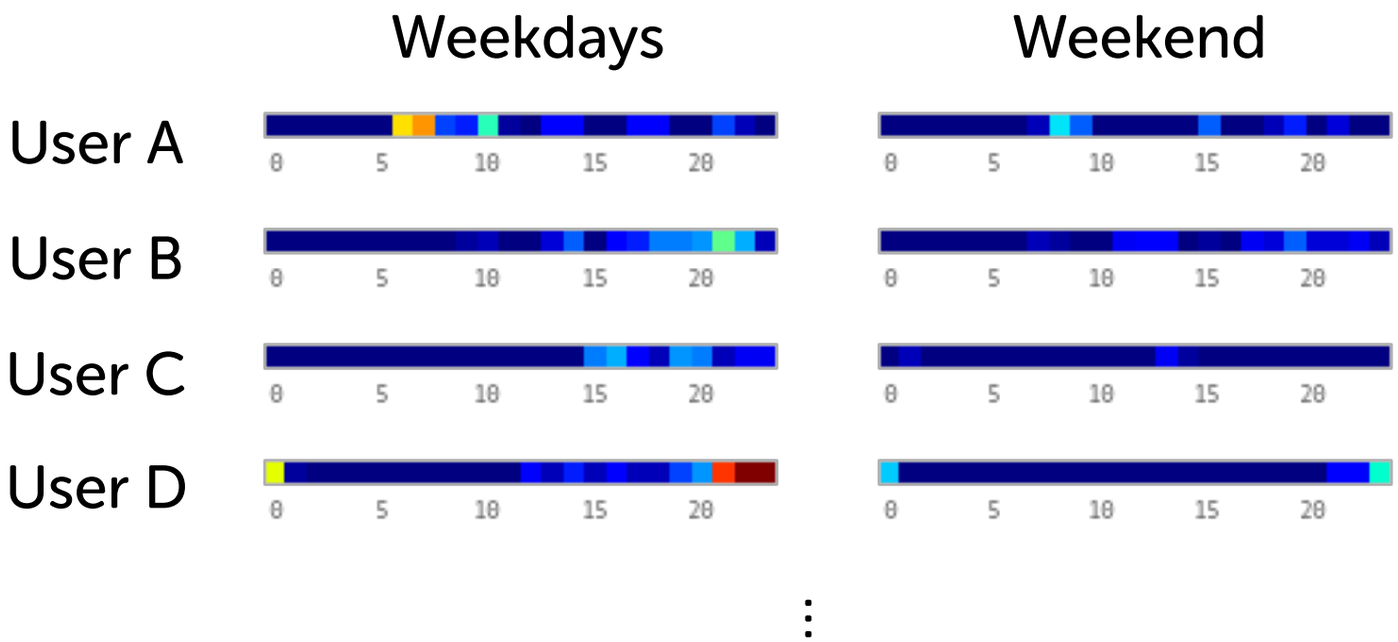
- Has longer ‘streak’
  連続日数が長い

- Does more practice
  復習回数が比較的多い

- Takes more time during a session
  平均セッション時間が長い

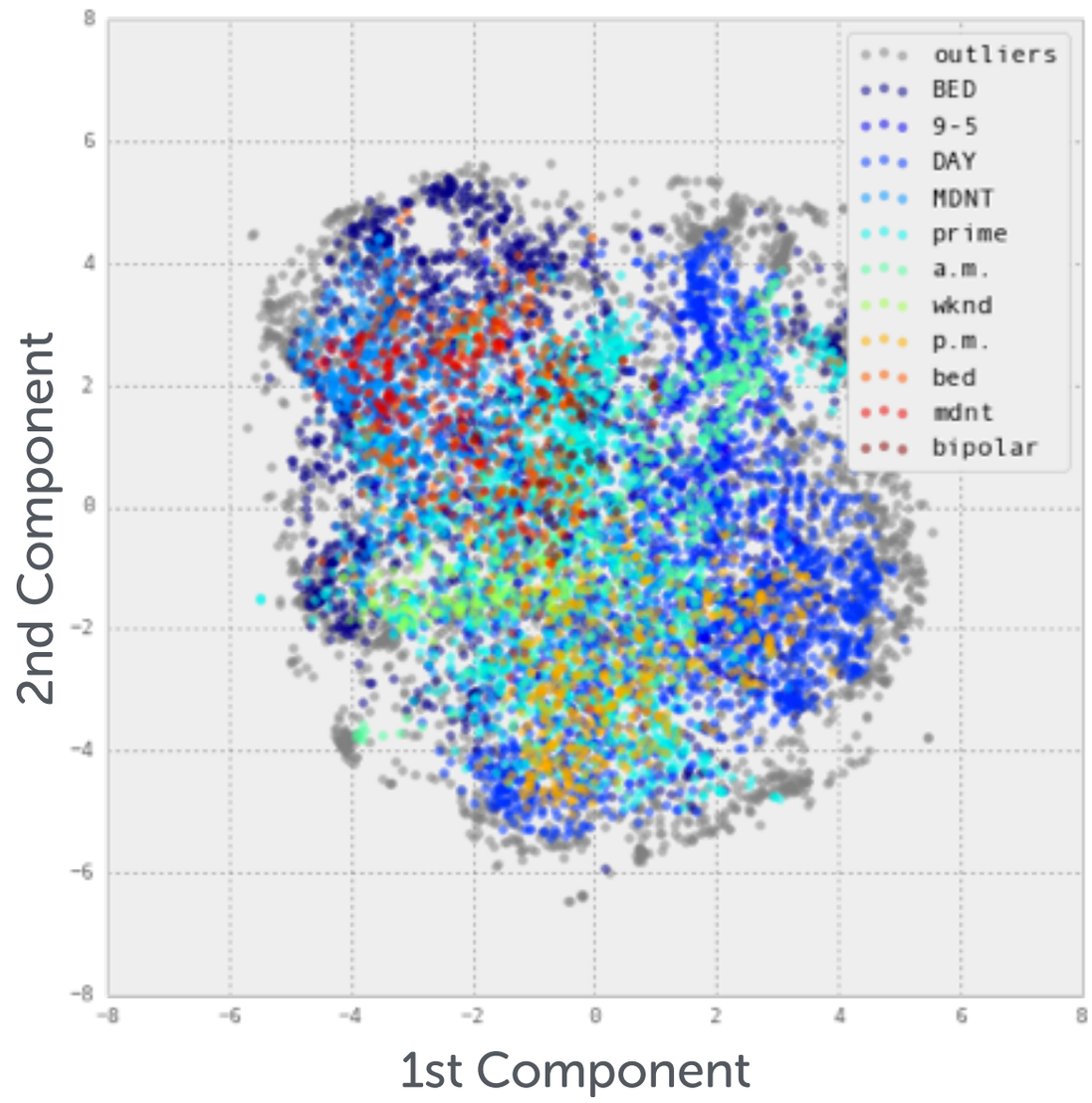
Clustering Users

ユーザーのクラスタリング



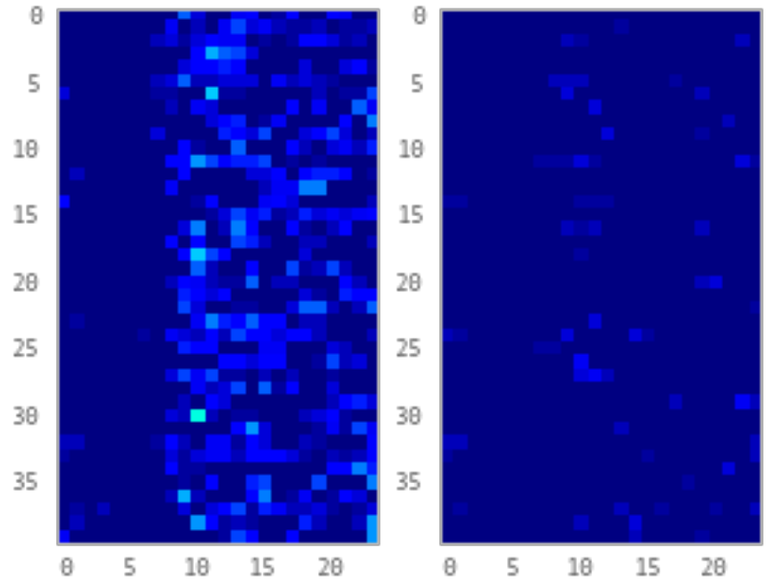
Dirichlet Process Gaussian Mixture Model (DPGMM)

Visualization by t-SNE

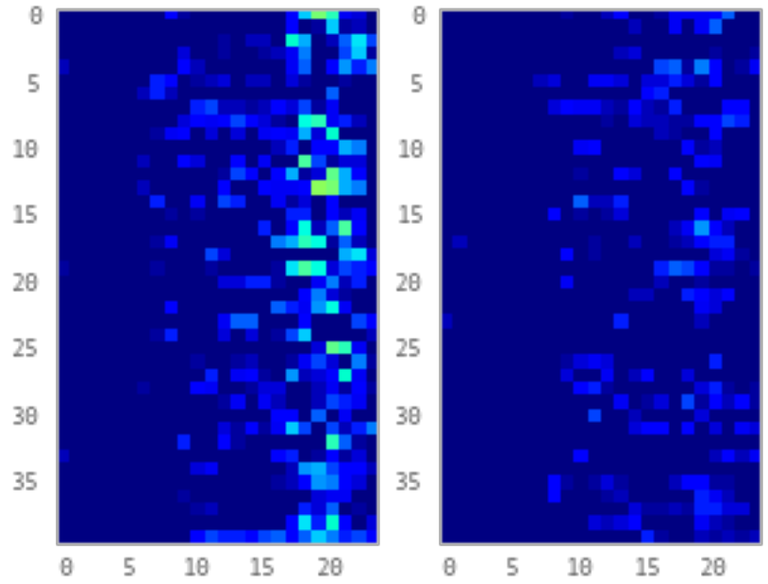


Cluster Examples

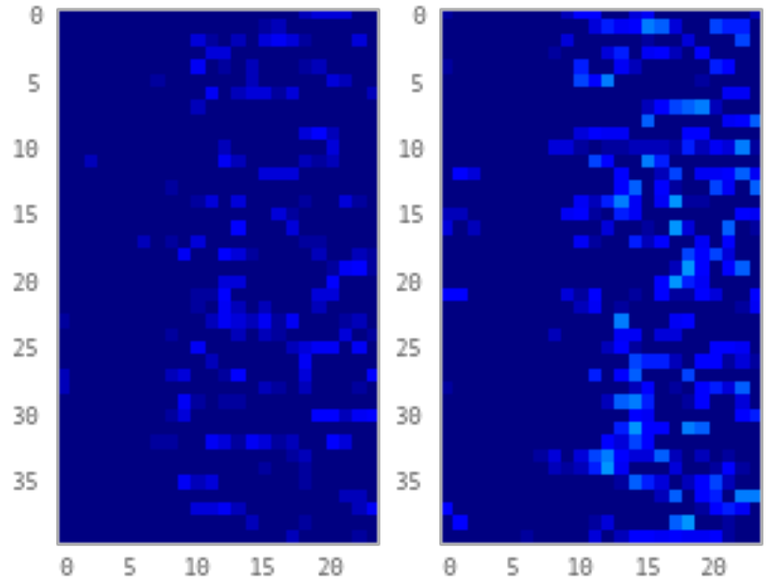
クラスターの例



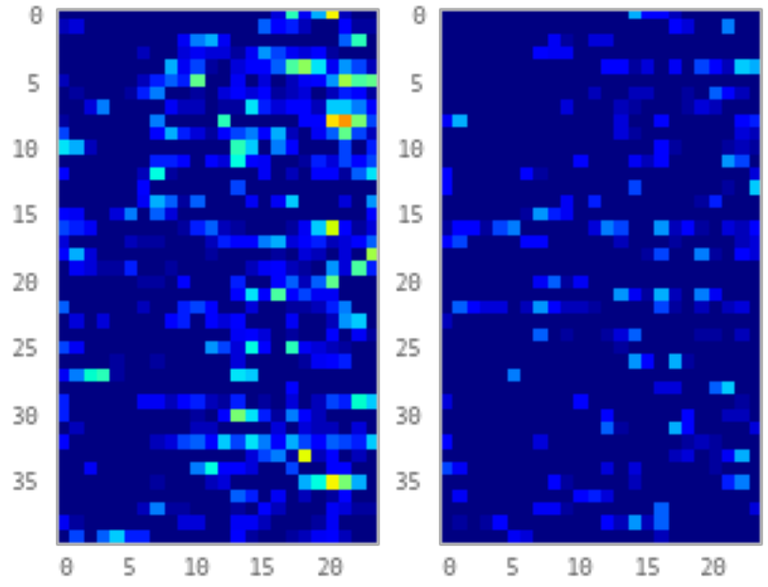
9am - 5pm



Prime time (~8pm)



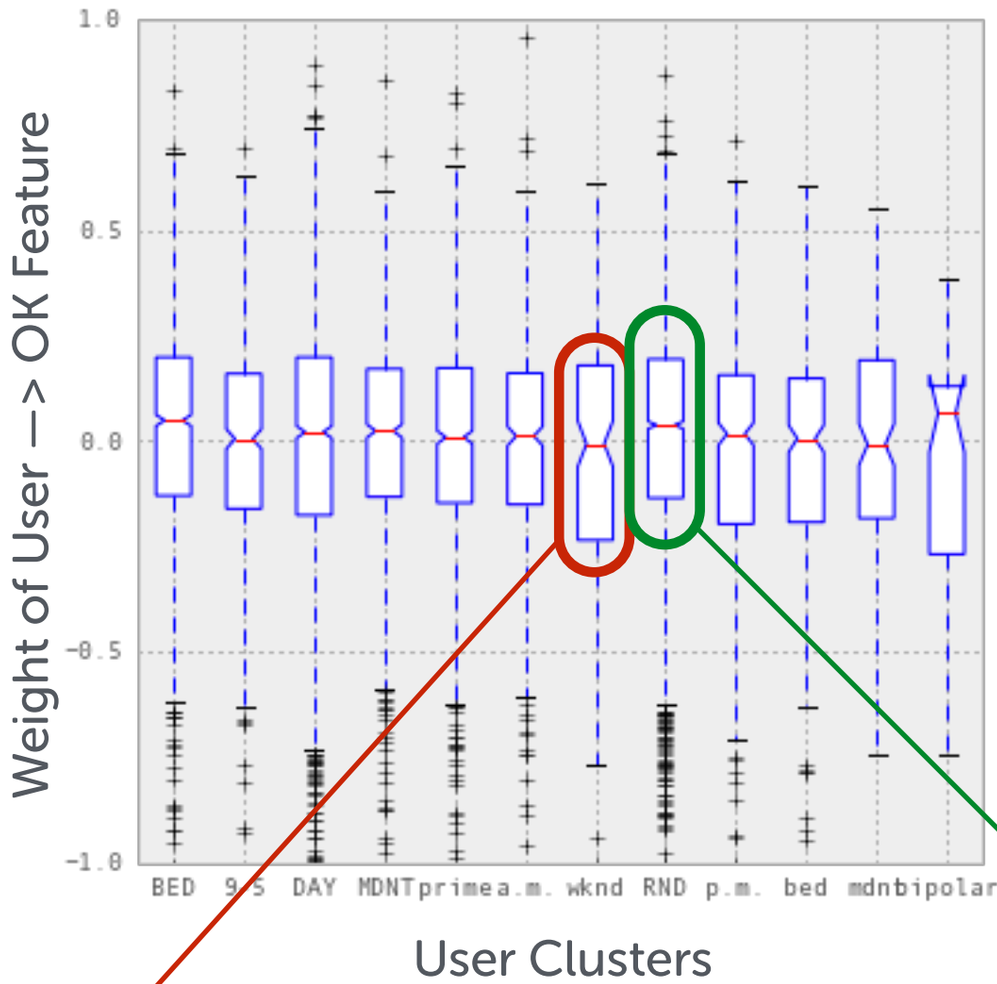
Weekend



Random

Performance by Cluster

クラスターとその成績



Weekend (worst)

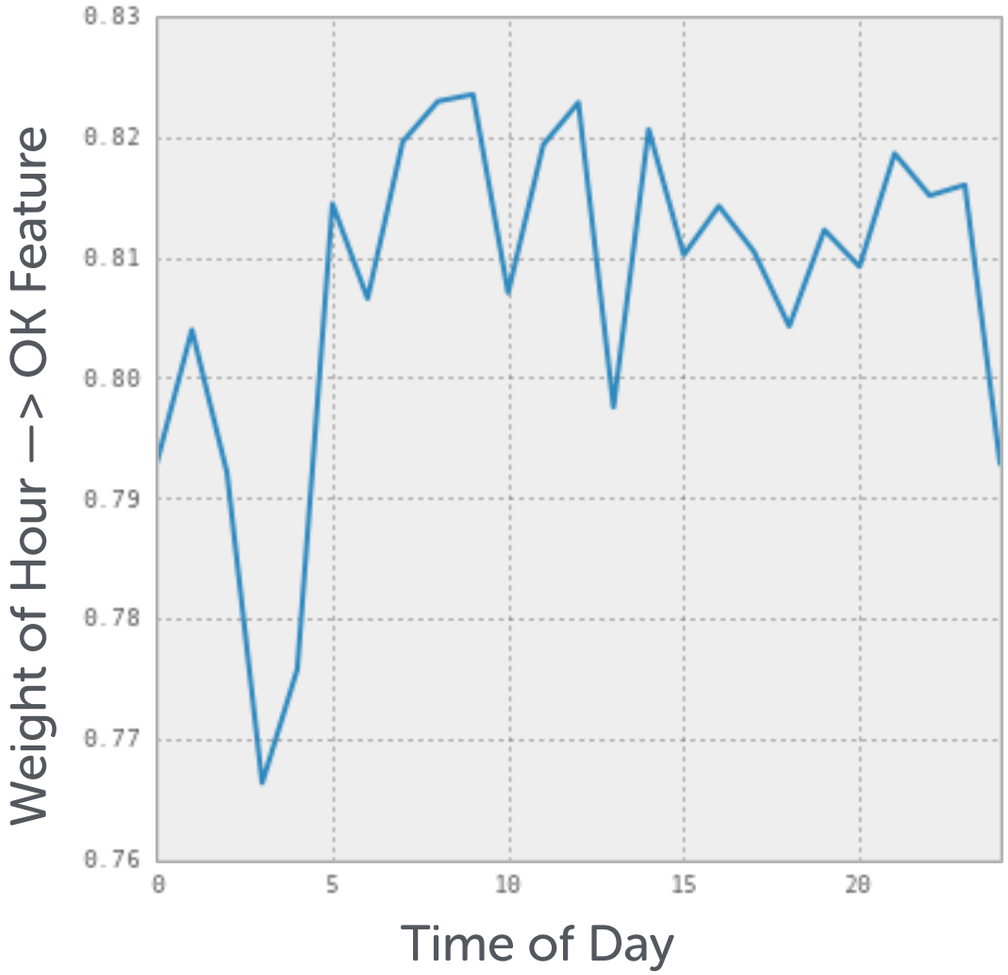
Random (best)

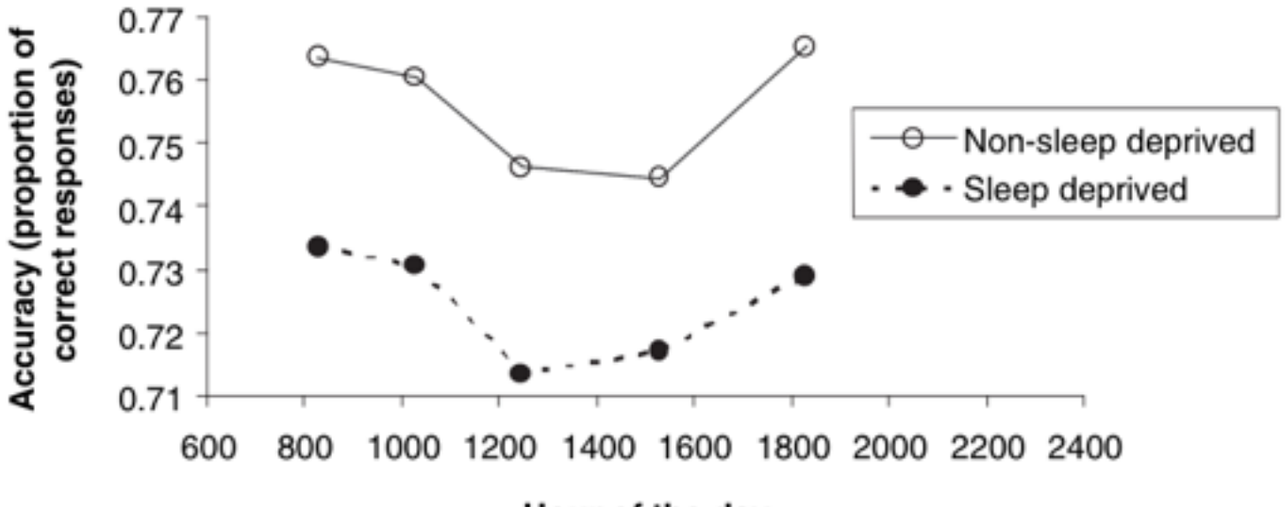
“How often you study” > “When you study”

学習頻度 > 学習時間

Time of Day vs Performance

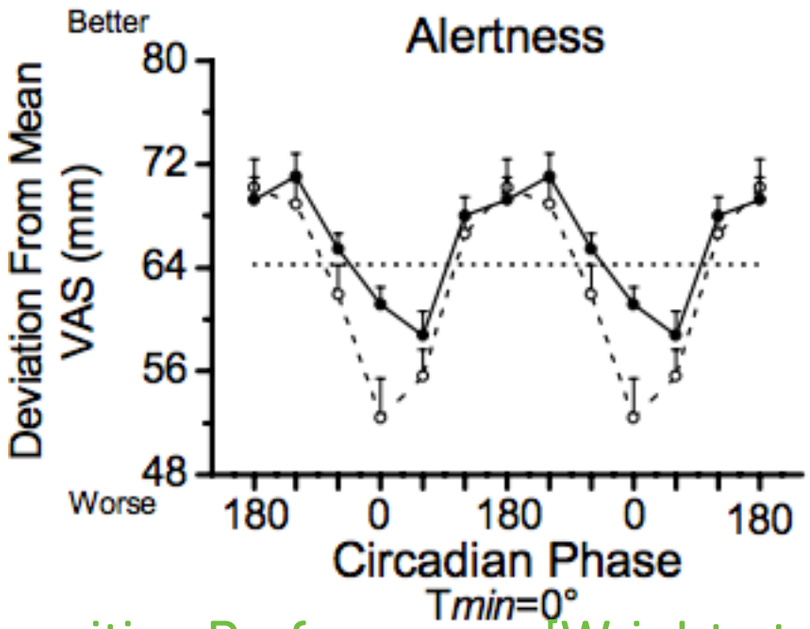
時間と成績





Auditory temporal judgement [Babkoff et al., 2004]

聴覚時間差判断



Cognitive Performance [Wright et al., 2002]

認知能力

Conclusion (so far)
これまでの結論

- Consistent practice (vs “cramming”) is the key
  詰め込みではなく、継続的な練習が重要

- Studying only on weekends may not be enough
  週末だけ勉強するのではおそらく足らない

- When you study doesn’t matter as much as how often you study (unless it’s at 3am!)
  勉強の時間より、勉強の間隔が重要

A Model for

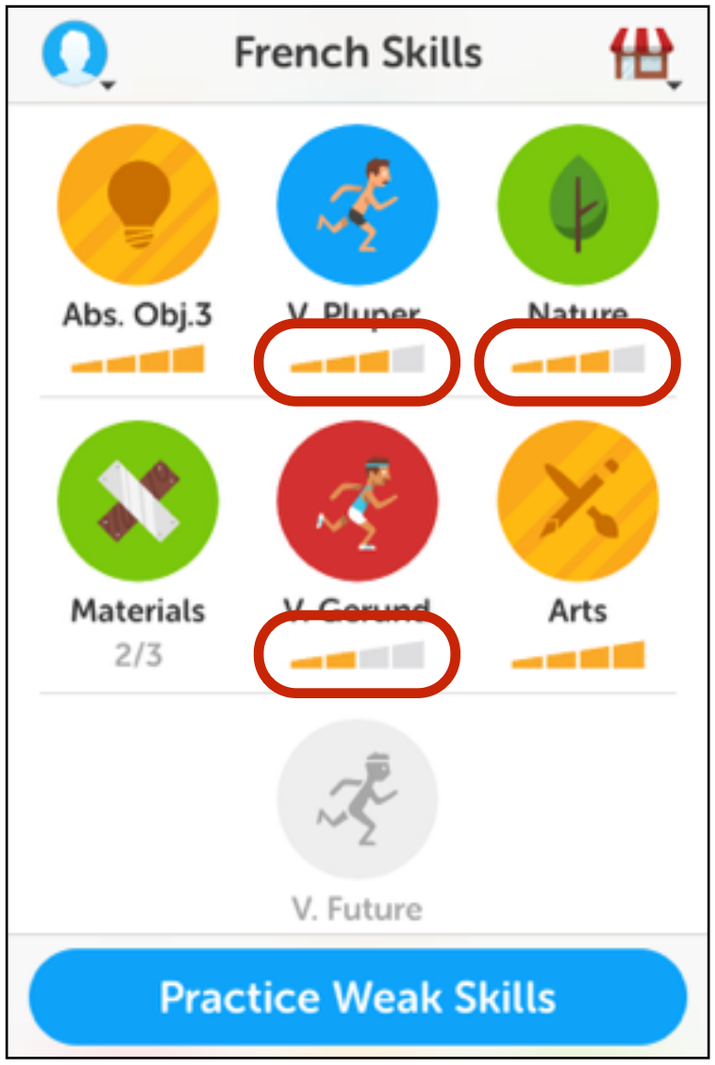
Spaced Repetition

Practice

間隔反復法モデル

(Settles and Meeder 2016)

Spaced Repetition
間隔反復

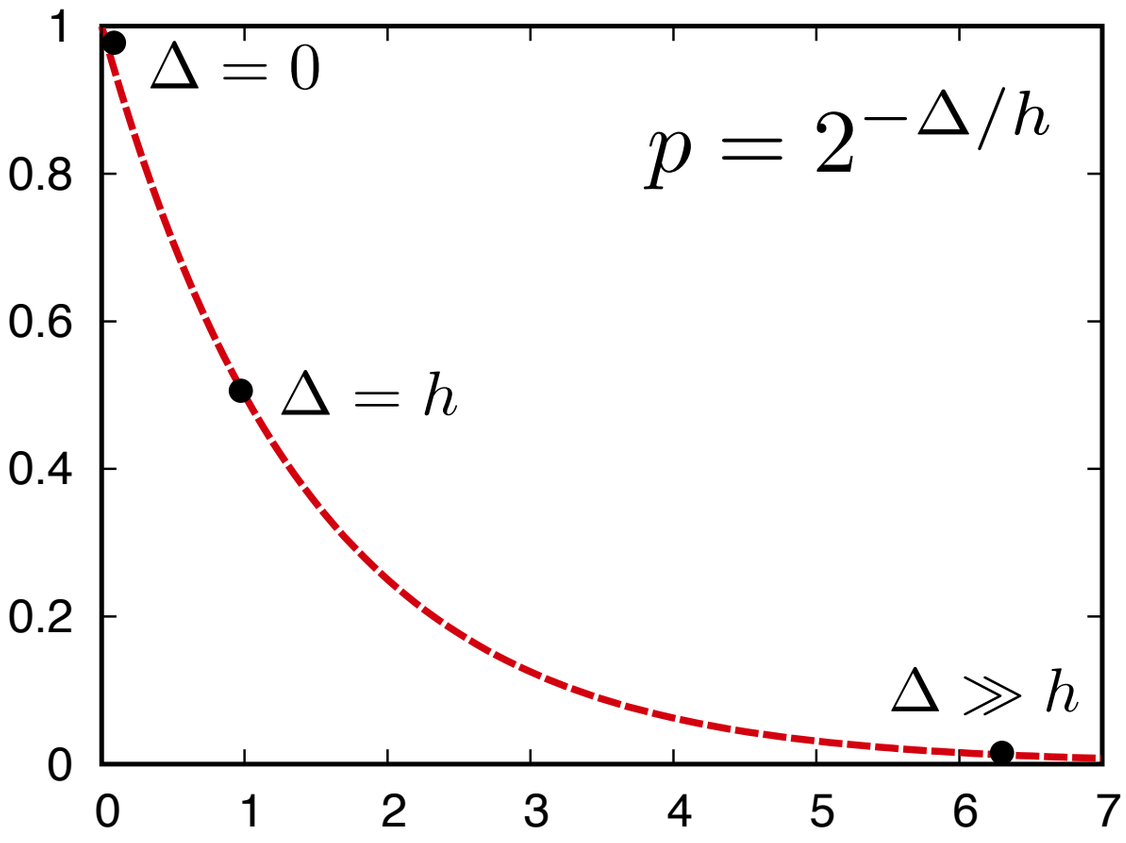


people tend to learn better when practices are **spaced over long intervals**, rather than “cramming”
詰め込みより間隔を開けた学習の方が効果的

Duolingo uses **strength bars** to indicate when lexemes/concepts in a skill need practice
強さバーを使い単語・概念の学習時期を表示

(Ebbinghaus, 1885; Pimsleur, 1967; Leitner, 1972)

Forgetting Curve
忘却曲線



the **probability** p of a correct answer as a function of:
正解率 p は以下の関数：

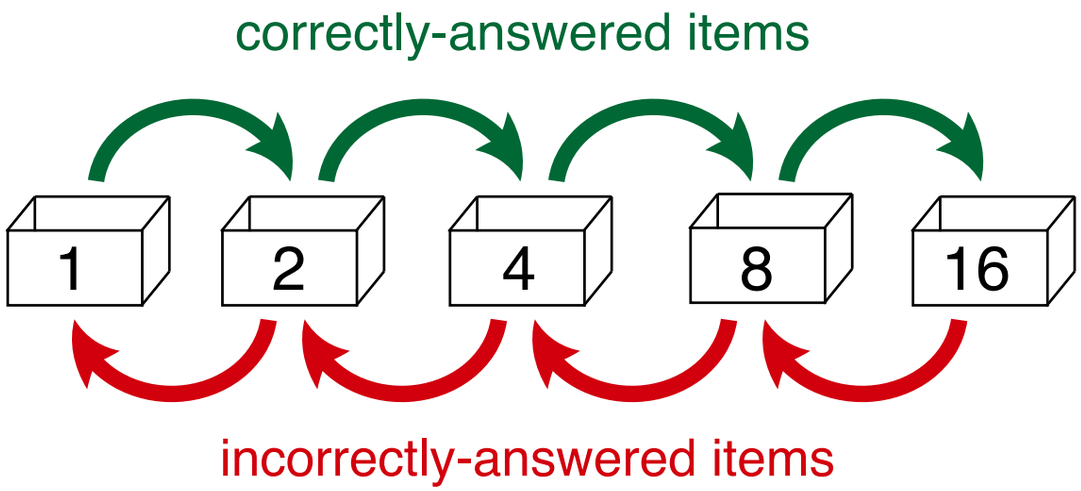
- **time** Δ since it last practice
  経過した時間 Δ
- **halflife** h in user's memory
  半減期 h

(Ebbinghaus, 1885; Pimsleur, 1967; Leitner, 1972)

Halflife?

半減期の推定

initial implementation:
Leitner flashcard algorithm
ライトナーシステム



idea:
formalize halflife into
the equation:
半減期の表現

$$h = 2^{x_{\oplus} - x_{\ominus}}$$

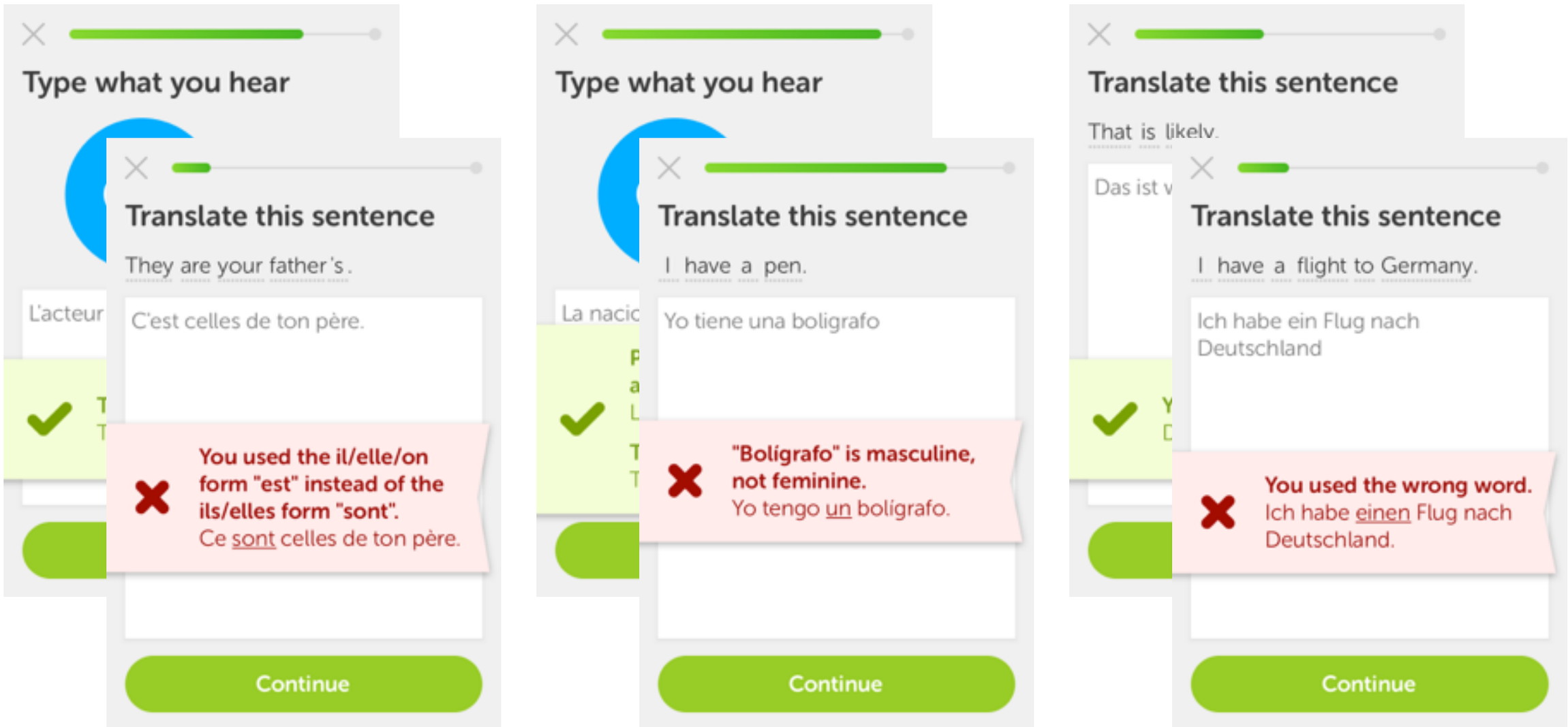
or, **more generally**:
一般化

$$h = 2^{\Theta \cdot x}$$

for model weights Θ and a
vector of predictors x

Student Performance Data

学生の成績データ



can we **learn to predict** word halflife from these data?

“Halflife Regression”

半減期回帰モデル

- learn $h = 2^{\Theta \cdot x}$ empirically from performance data:

$$\Theta^* = \arg\min \sum_i \left(y^{(i)} - 2^{(-\Delta^{(i)}/(\Theta \cdot x^{(i)}))} \right)^2 + \lambda \|\Theta\|_2^2$$

actually correct (0,1)

predicted probability correct

L2 regularization term

- fit Θ with stochastic gradient descent using 1 week of log data (13M student/word pairs)

Results: Error Rates

結果：エラー率

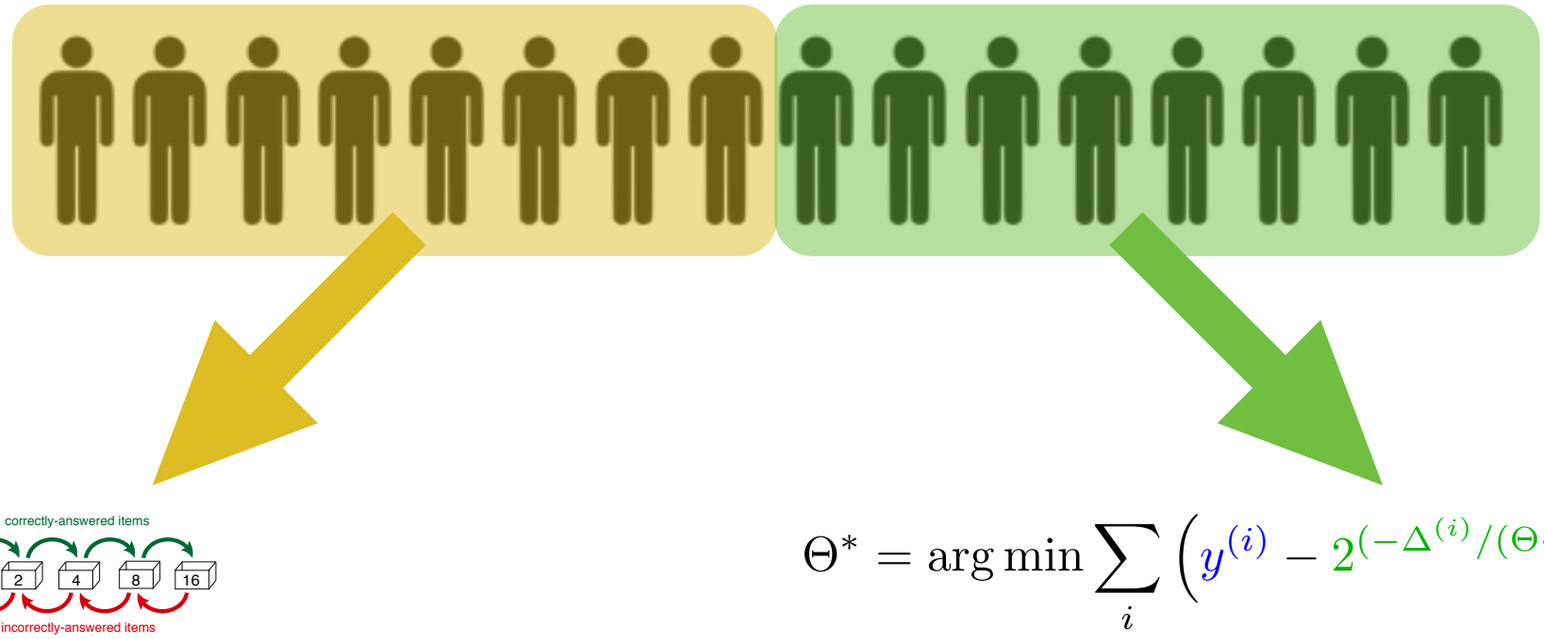
40-46% error reduction on a held-out week of future data

language course	Leitner	regression
🇺🇸 ← 🇪🇸	0.457	**0.272**
🇺🇸 ← 🇮🇪	0.465	**0.253**
🇺🇸 ← 🇧🇷	0.451	**0.269**
🇫🇷 ← 🇺🇸	0.488	**0.287**
🇩🇪 ← 🇺🇸	0.503	**0.283**
🇪🇸 ← 🇺🇸	0.503	**0.270**
🇮🇪 ← 🇺🇸	0.433	**0.262**
🇧🇷 ← 🇺🇸	0.467	**0.280**

(ca. April 2013)

Results: User Experiment

結果：ユーザー実験



$$\Theta^* = \arg\min \sum_i \left(y^{(i)} - 2^{(-\Delta^{(i)}/(\Theta \cdot x^{(i)}))} \right)^2 + \lambda \|\Theta\|_2^2$$

OLD

(6 weeks; 3.6M subjects)

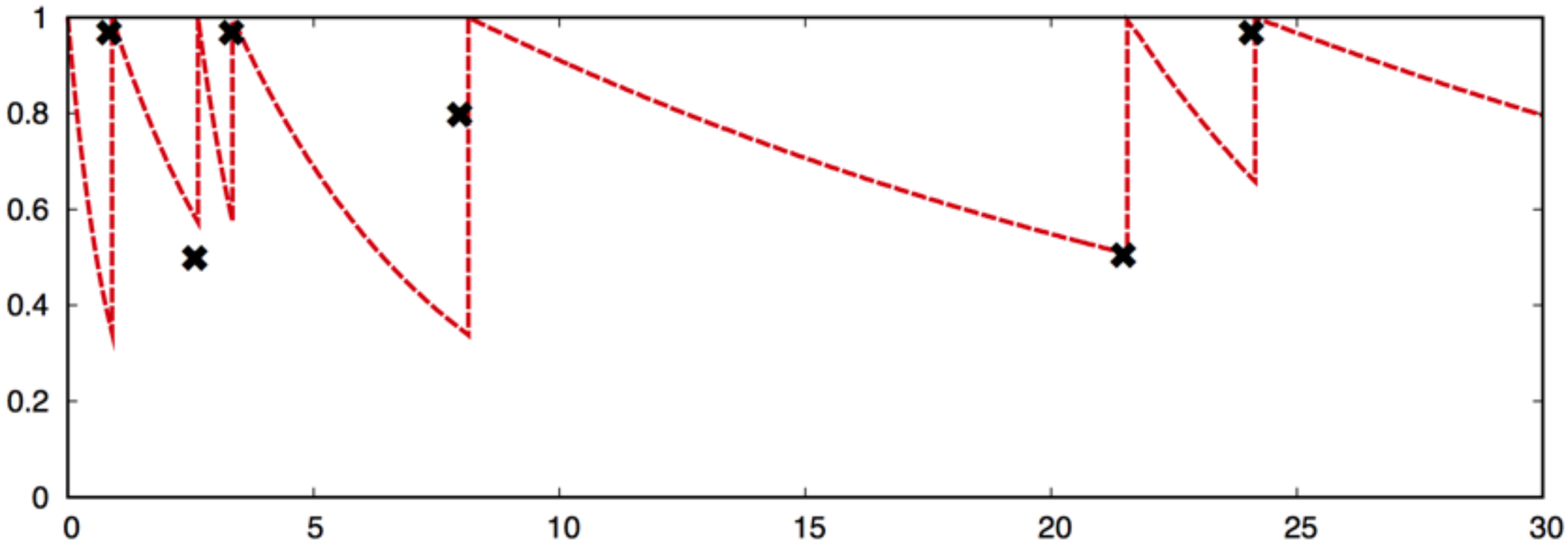
NEW

all students
who returned the next day:

+1.7% (p<0.001)

Results: Example User Trace

結果：回帰例



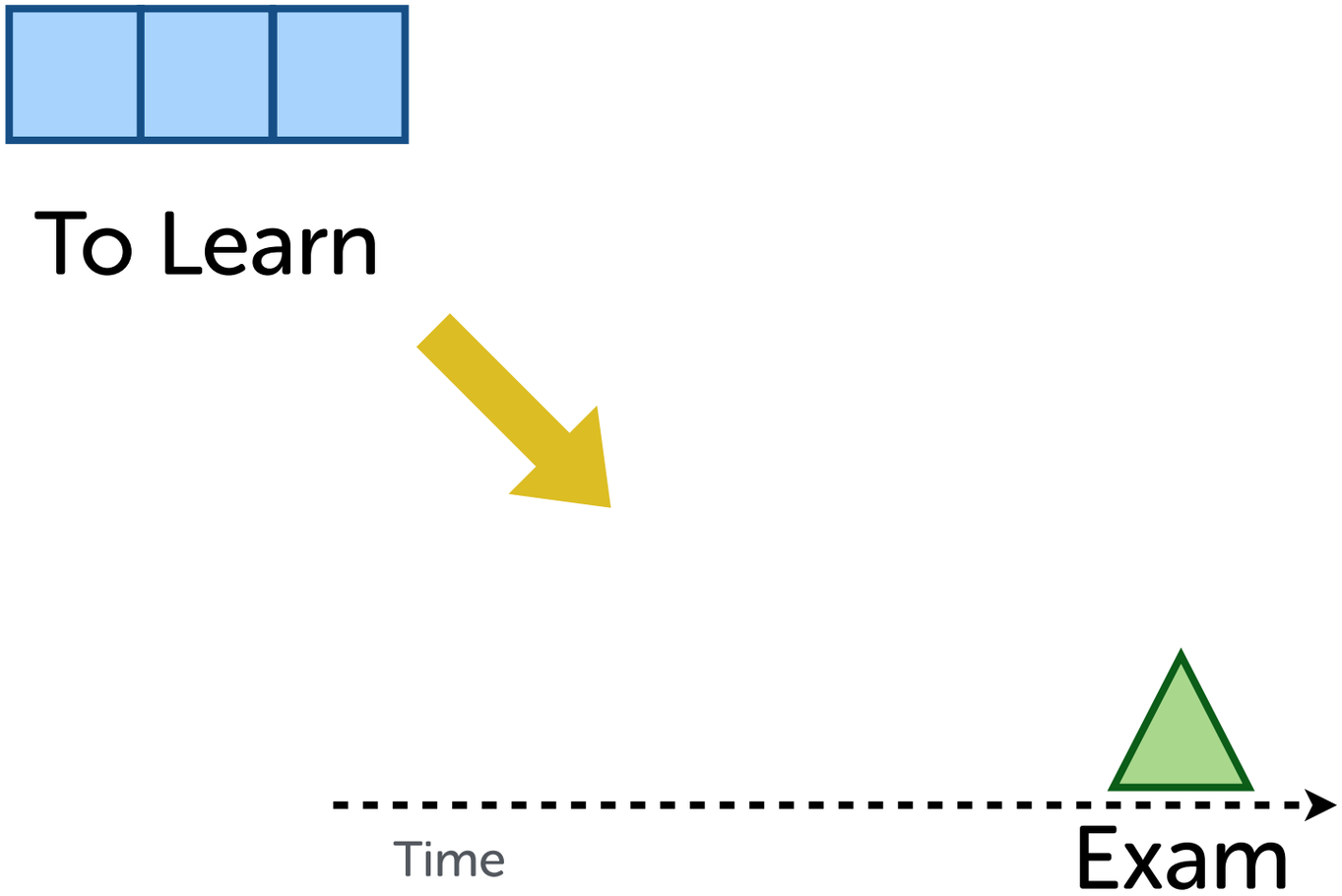
(b) 30-day student-word learning trace and predicted forgetting curve

How best to learn anything

効率的な学習法

Learning Strategy

学習法



Practice Testing

練習テスト

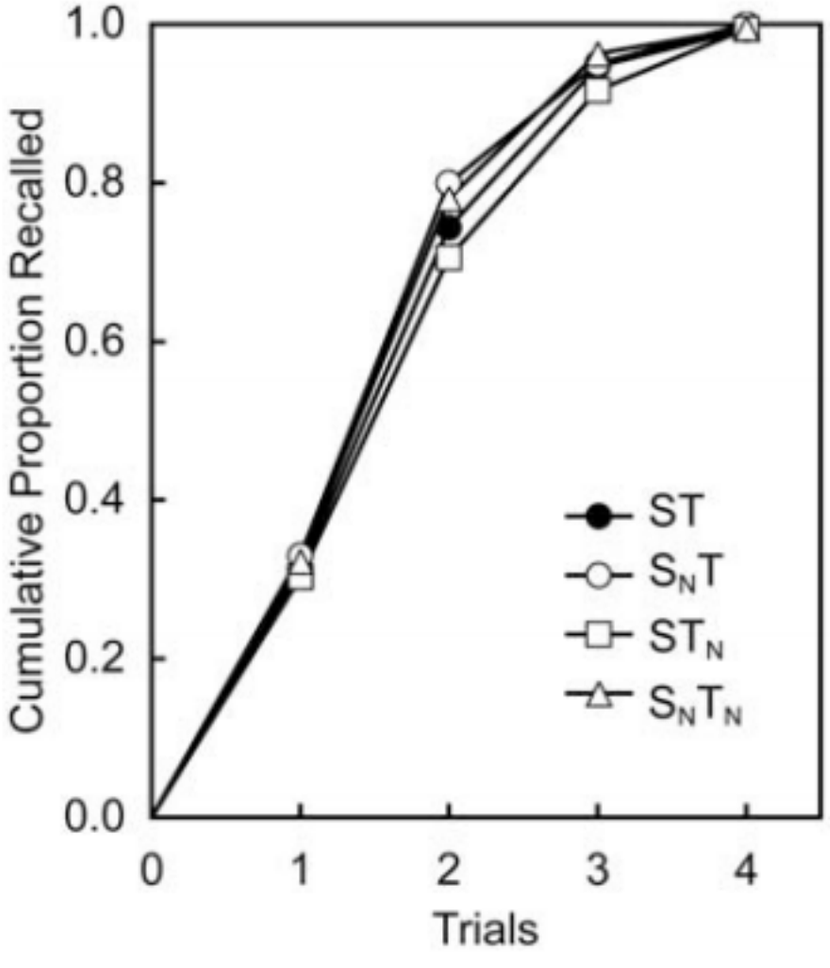
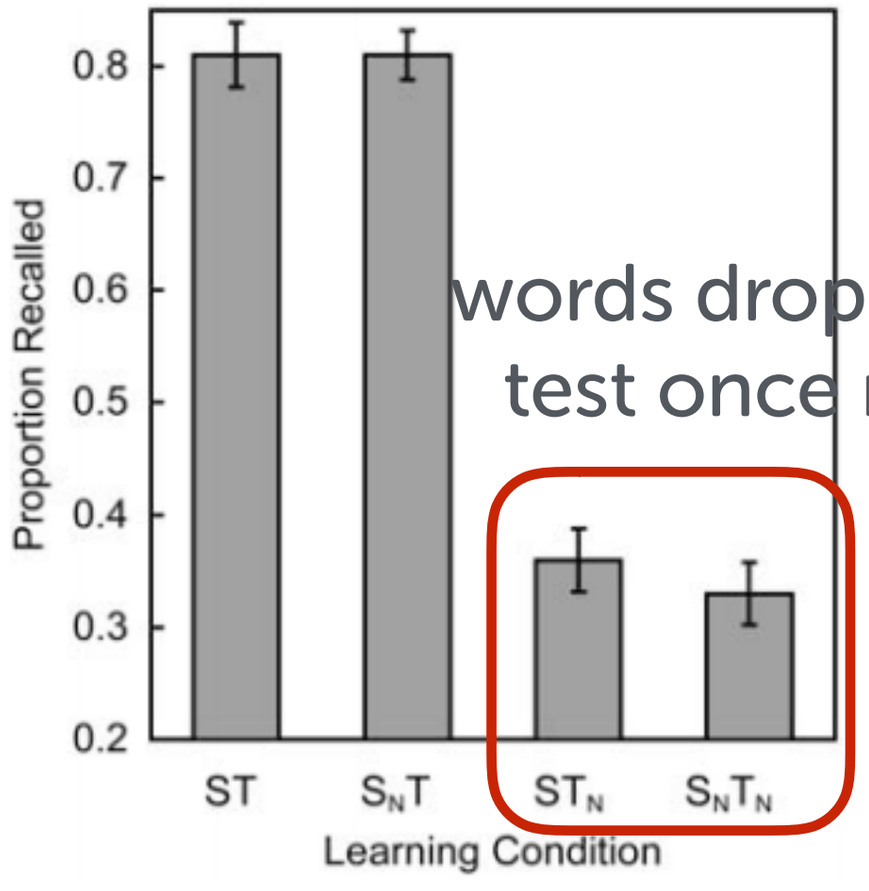


Fig. 1. Cumulative performance during the learning phase.



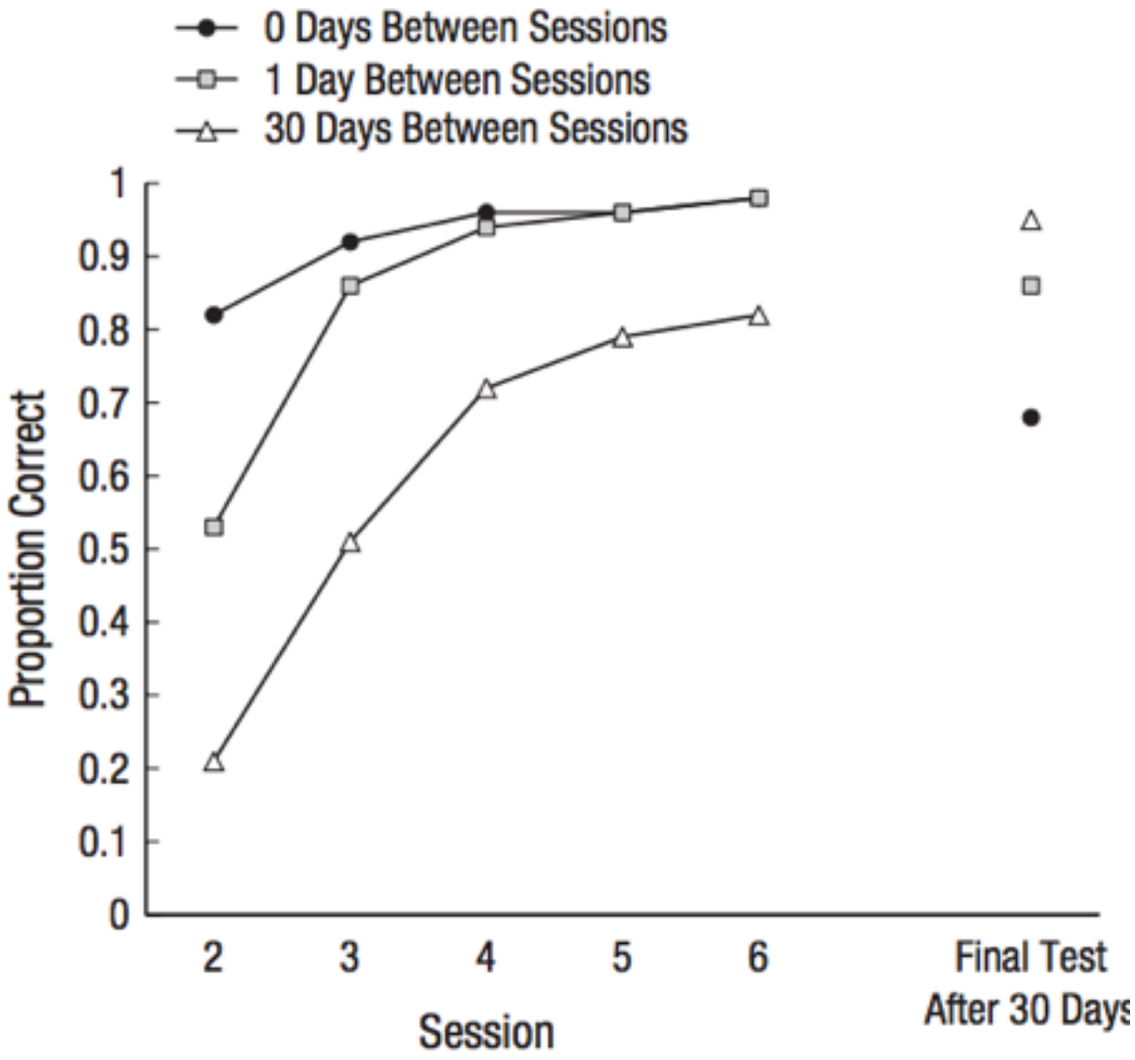
words dropped from test once recalled

Fig. 2. Proportion recalled on the final test 1 week after learning. Error bars represent standard errors of the mean.

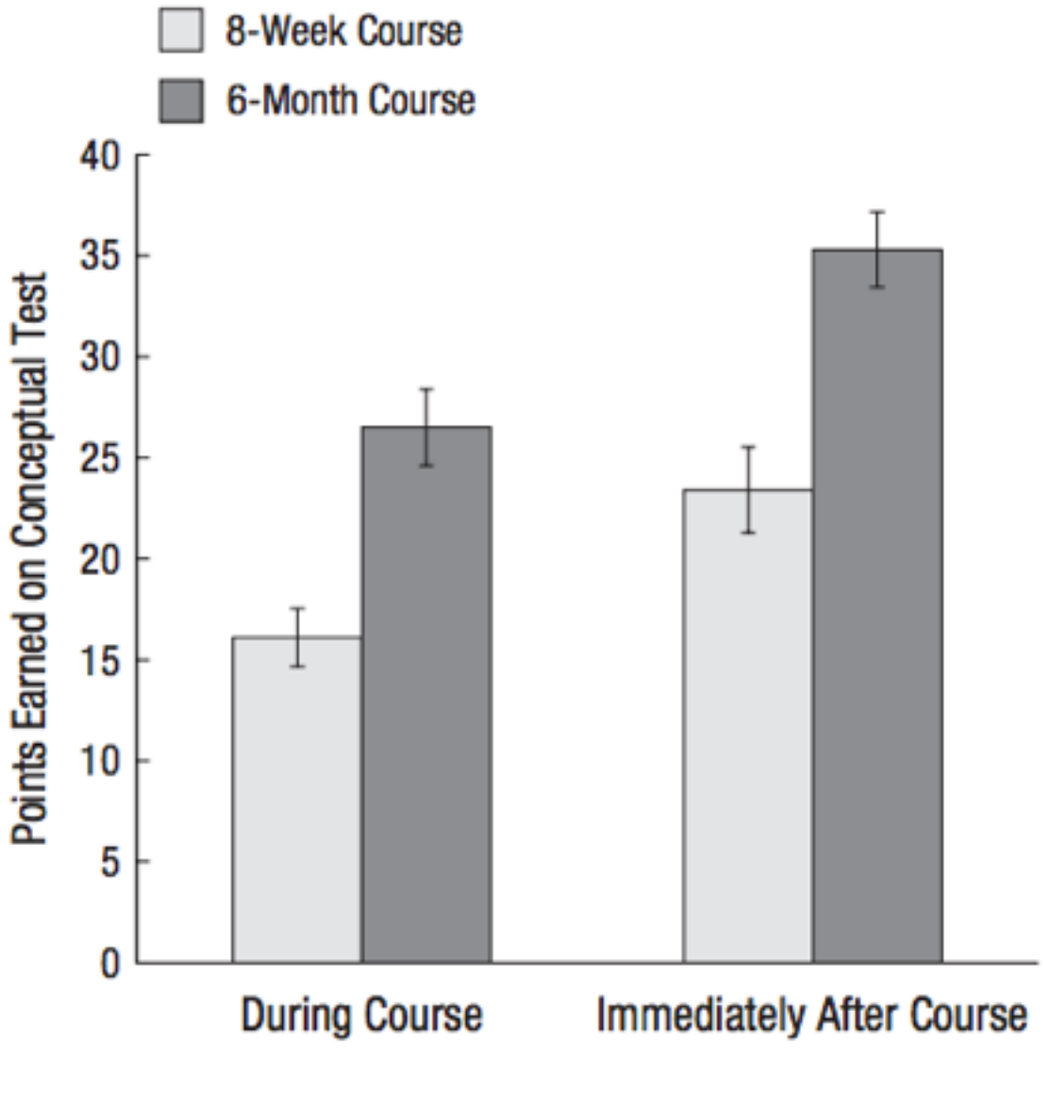
Remembering Swahili-English word pairs [Karpicke & Roediger III, 2008]

Distributed Practice

分散練習



Translation of Spanish words [Bahrick 1979]

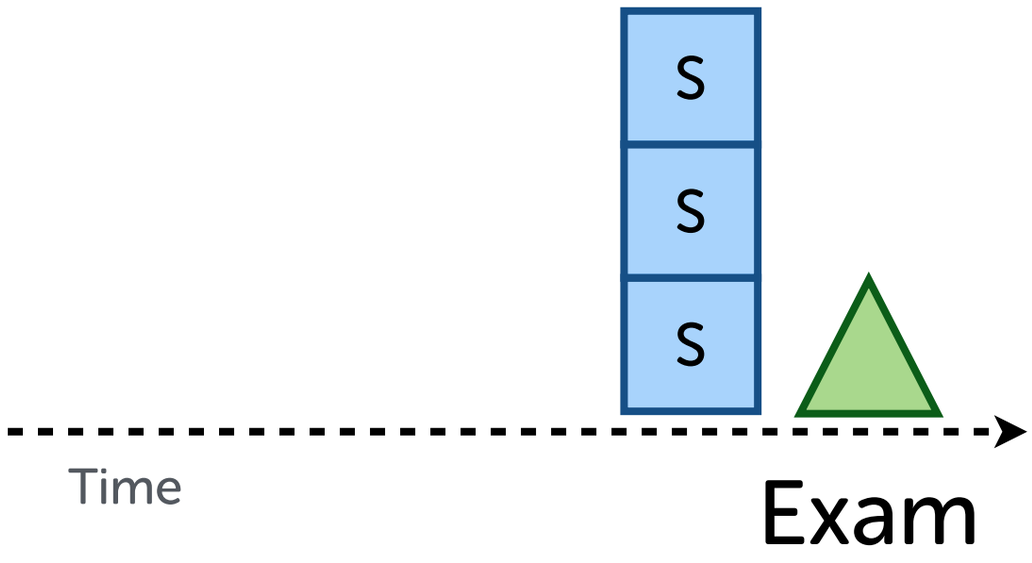


[Budé, Imbos, van de Wiel, and Berger 2011]

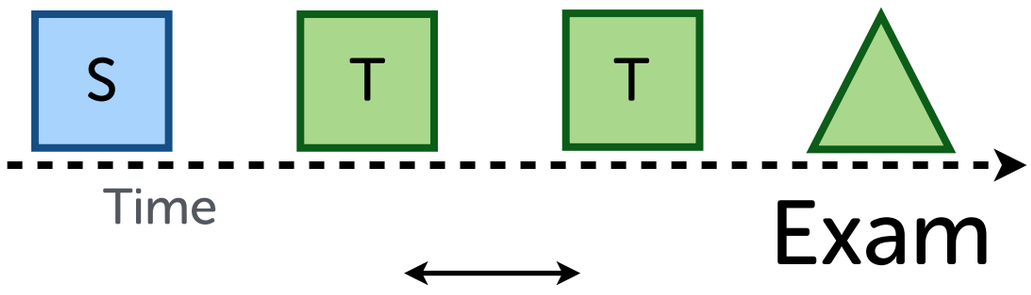
Distributed Practice

分散練習

✗



✓



10%-20% of desired retention period

[Cepeda et al., 2008]

Conclusion

結論

- _____ is the largest language learning app with 150 million registered users

  _____ は、1億5千万の登録ユーザーを有する世界最大の外国語学習アプリである

- Successful learners study language in a _____ manner (i.e., do not _____)

  成功する学習者は、_____ 的に勉強する（_____ しない）

- _____ and _____ are the two key strategies for effective learning

  効果的な学習のためには、_____ と _____ が重要である

Thanks!

Follow me on twitter @mhagiwara