Second Language Acquisition Modeling

2018 NAACL / BEA Shared Task Report

Burr Settles
Duolingo

Chris Brust
Duolingo

Erin Gustafson
Duolingo

Masato Hagiwara
Duolingo

Nitin Madnani
Educational Testing Service

why should we care about modeling second language acquisition?

people learning a second language

1,200,000,000

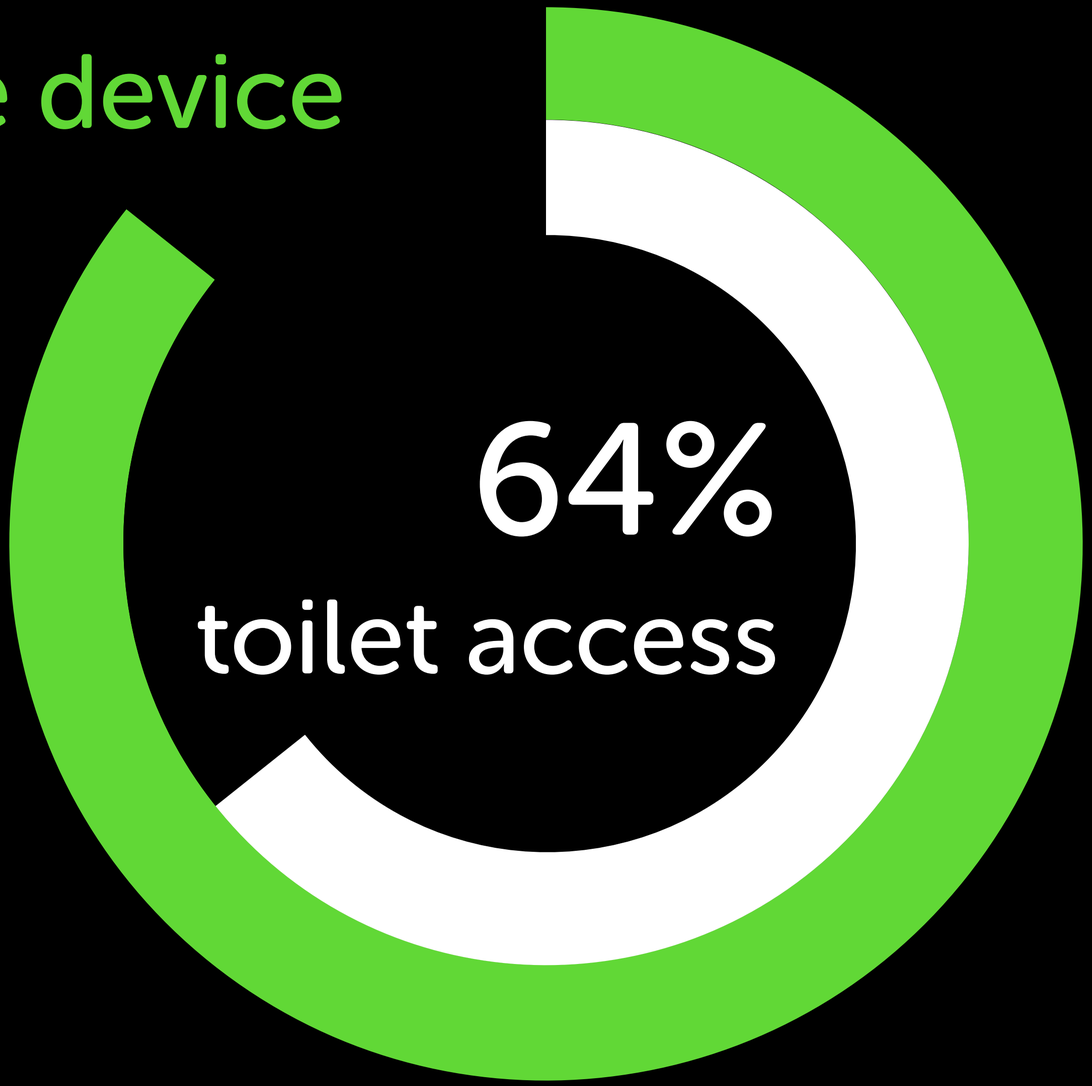
(~16% of the world's population)

~800M satisfy three properties:

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

(Source: British Council)

86%
mobile device
access

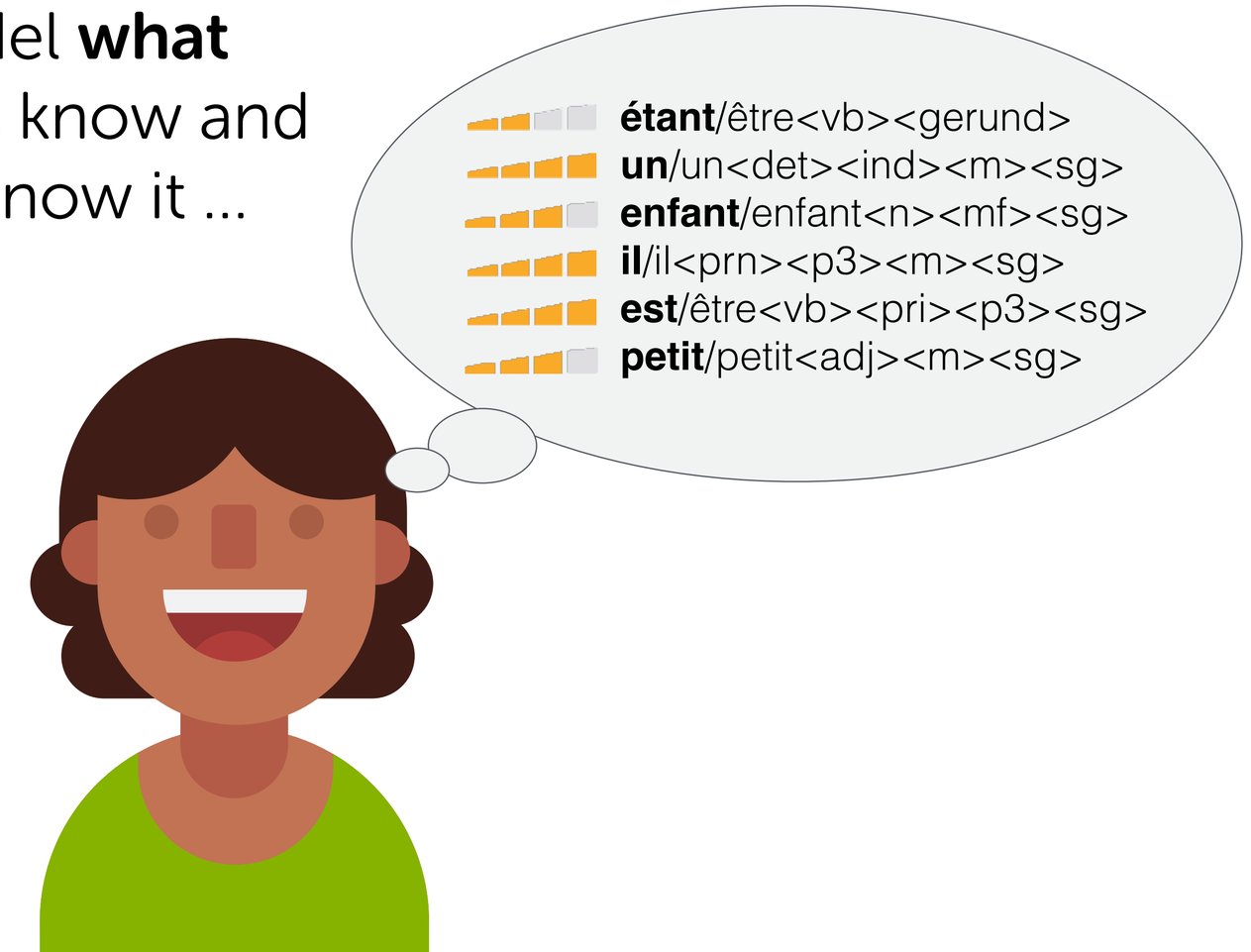


(Source: U.N. Report, 2015)

enormous potential for computer-based, adaptive language-learning!

SLA Modeling

accurately model **what** language-learners know and **how well** they know it …



SLA Modeling

... and do so in a
personalized way
(that adapts + learns over time)

Learner Modeling in Other Domains

DataShop (Koedinger et al., 2010)

- **150 public** research data sets
- mostly **math + physics** domains, largely multiple-choice items
- still relatively **small**:
  - 71 avg students (5k max)
  - 880 avg instances (1.5M max)

Our Goals for the SLAM Task

- facilitate dialog among ML/NLP/CogSci fields through a common **large-scale** empirical task
- accessible, **familiar data format** + task definition (e.g., classification similar to other shared tasks)
- include languages **other than English**
- start with **beginners** who are learning **over time**

duolingo

DEMO!

launched in **2012** (CMU research spinoff)

more than **200 million students** globally

currently **79 courses** (incl. Irish, Esperanto, + Klingon!)

expanding to **93 courses** (incl. Arabic + Hindi!)

content is **FREE**

GOOGLE
Best of the Best

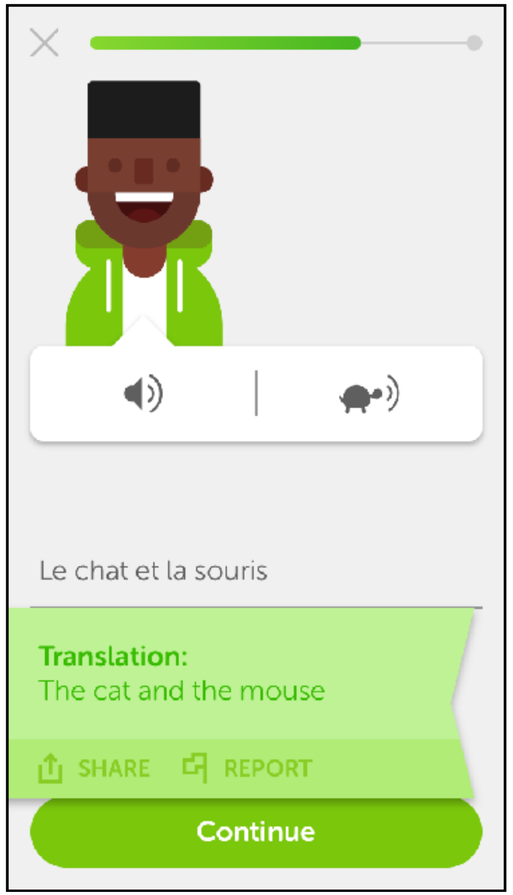
APPLE
App of the Year

TECHCRUNCH
Education Startup of the Year

The Data

`reverse_translate`

`reverse_tap`

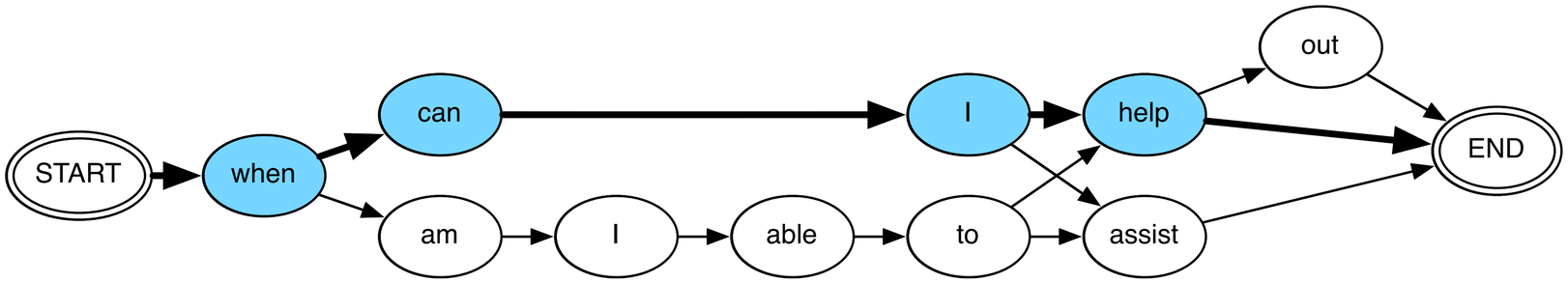


`listen`

The Data

prompt: cuándo puedo ayudar

when (can i|am i able to) (help (out|)|assist)



student: wen can help

reference: when can I help

label: 1 0 1 0

The Data

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=Inf|fPOS=VERB++VB	ROOT	0	0

IDs | reference answer tokens | morpho-syntactic features | labels

The Data

user + session-level metadata

```
# user:XEinXf5+  countries:CO  days:2.678  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=Inf|fPOS=VERB++VB                                  ROOT       0  0
```

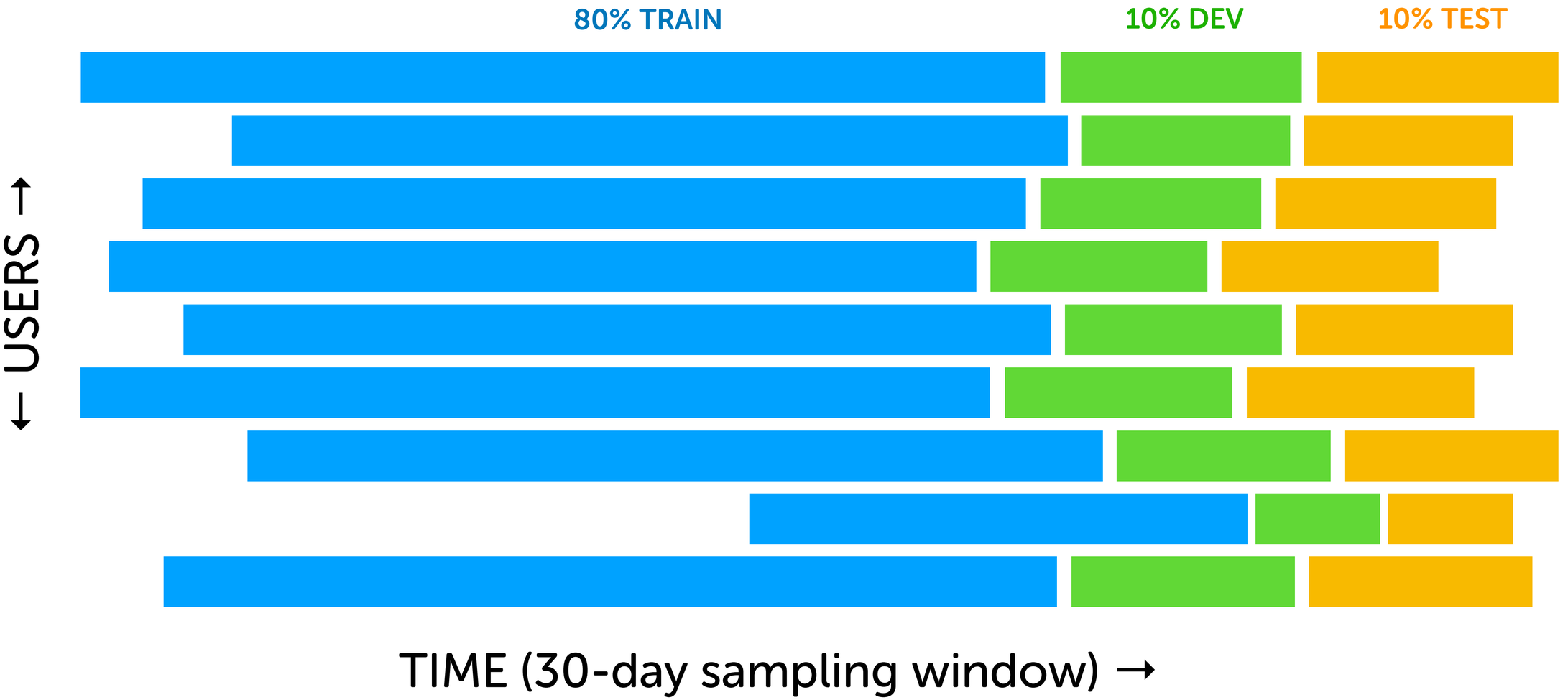
The Data

```
# user:XEinXf5+  countries:CO  days:2.678  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=Inf|fPOS=VERB++VB                                          ROOT       0  0

# user:XEinXf5+  countries:CO  days:5.707  client:android  session:practice  format:reverse_translate  time:22
W+QU2fm70301   He        PRON   Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP  nsubj      3  0
W+QU2fm70302   's        AUX    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=AUX++VBZ    aux        3  1
W+QU2fm70303   wearing   VERB   Tense=Pres|VerbForm=Part|fPOS=VERB++VBG                             ROOT       0  0
W+QU2fm70304   two       NUM    NumType=Card|fPOS=NUM++CD                                           nummod     5  0
W+QU2fm70305   shirts    NOUN   Number=Plur|fPOS=NOUN++NNS                                          dobj       3  0

# user:XEinXf5+  countries:CO  days:10.302  client:web  session:lesson  format:reverse_translate  time:28
vOeGrMgP0101   We        PRON   Case=Nom|Number=Plur|Person=1|PronType=Prs|fPOS=PRON++PRP           nsubj      2  0
vOeGrMgP0102   eat       VERB   Mood=Ind|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP                     ROOT       0  1
vOeGrMgP0103   cheese    NOUN   Degree=Pos|fPOS=ADJ++JJ                                             dobj       2  1
vOeGrMgP0104   and       CONJ   fPOS=CONJ++CC                                                       cc         2  0
vOeGrMgP0105   they      PRON   Case=Nom|Number=Plur|Person=3|PronType=Prs|fPOS=PRON++PRP           nsubj      6  0
vOeGrMgP0106   eat       VERB   Mood=Ind|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP                     conj       2  1
vOeGrMgP0107   fish      NOUN   fPOS=X++FW                                                          dobj       6  0
```

Data Partitions (Sequential)



Three Language Tracks



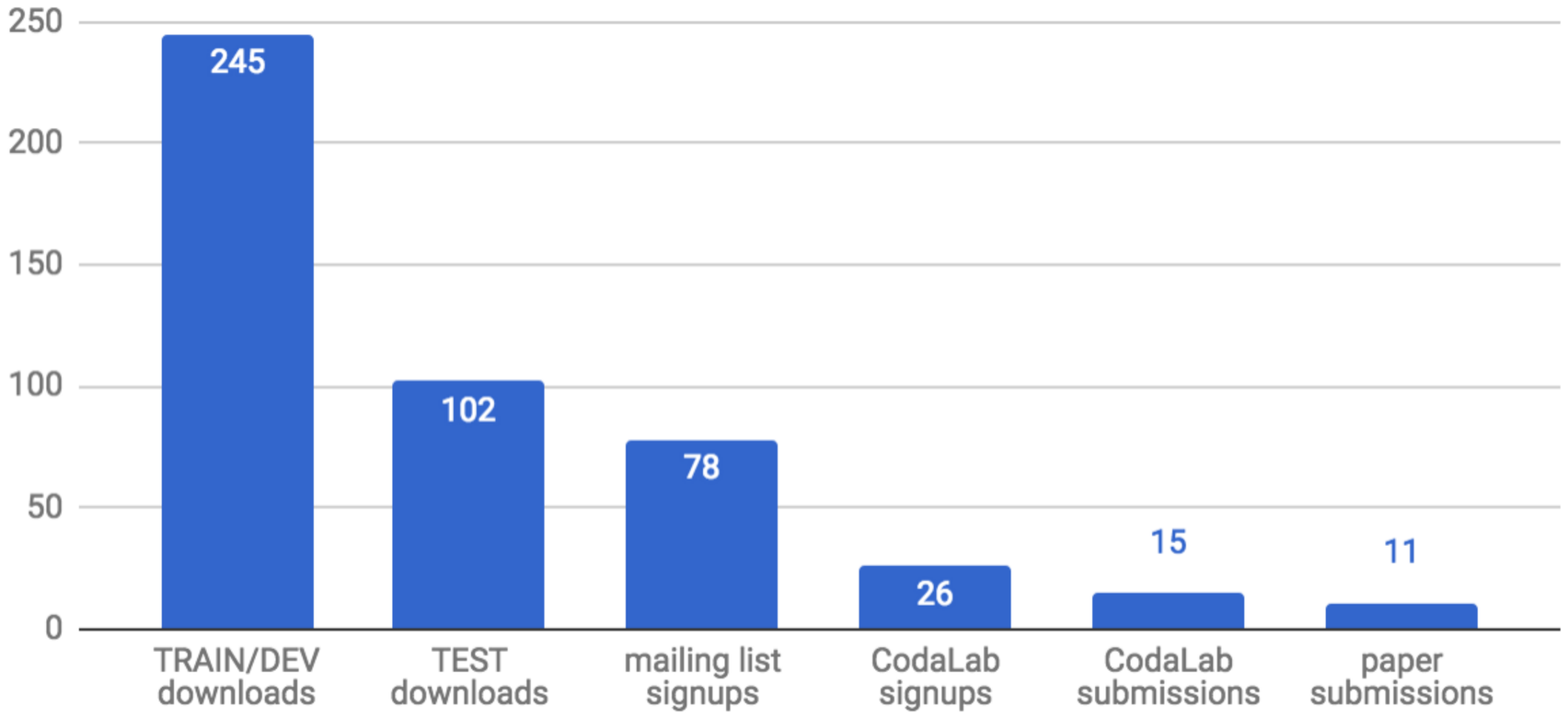
	English (EN ← ES)	Spanish (ES ← EN)	French (FR ← EN)	TOTAL (All 3 Tracks)
USERS	2,593	2,643	1,213	**6,449**
TRAIN (tokens)	2,622,958	1,973,558	926,657	**5,523,173**
DEV (tokens)	387,374	288,864	137,571	**813,809**
TEST (tokens)	386,604	282,181	135,525	**804,310**
TOTAL (tokens)	3,396,936	2,544,603	1,199,753	**7,141,292**

Duolingo's three **largest courses** (~1/3 of users)

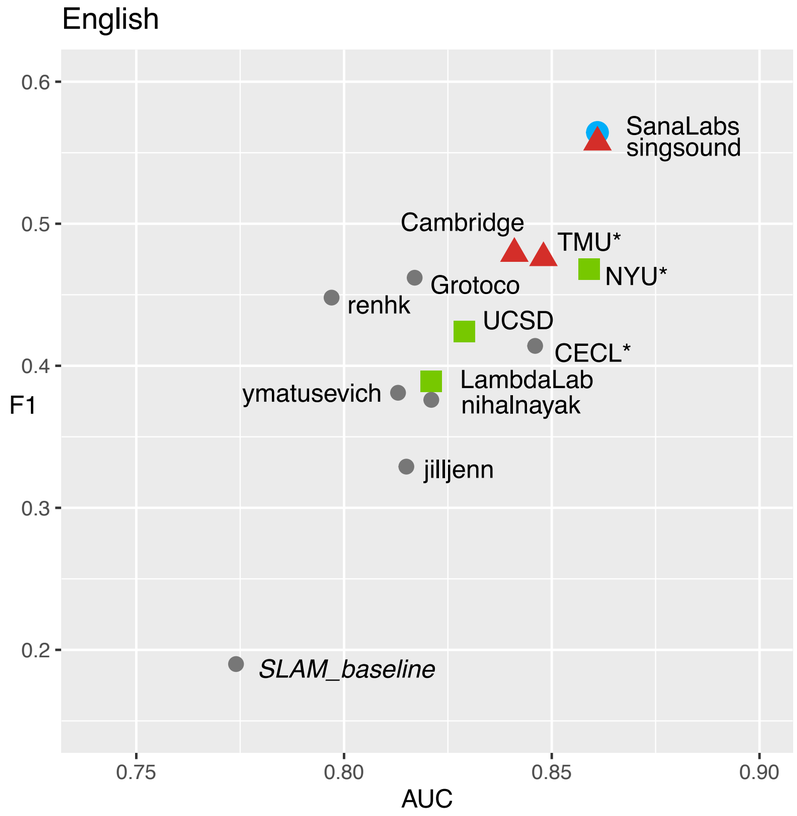
Other Details

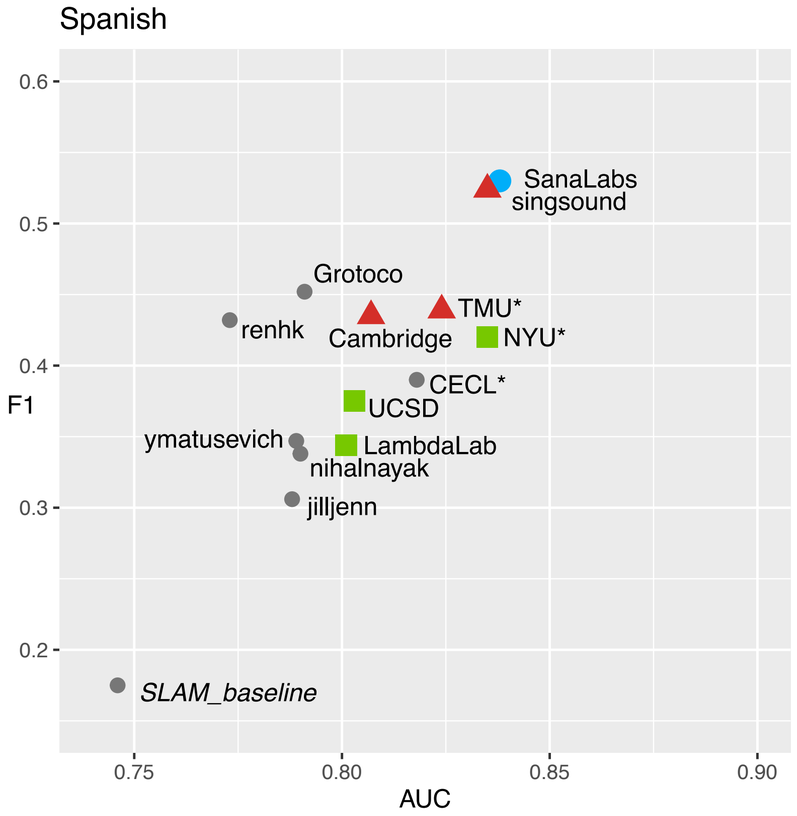
- **evaluation:** AUC (official metric) + F1
- **development phase (TRAIN + DEV):** 8 weeks
- **test phase (TEST):** 10 days
  - blind TEST set submissions via CodaLab
  - teams allowed to use both TRAIN+DEV to train

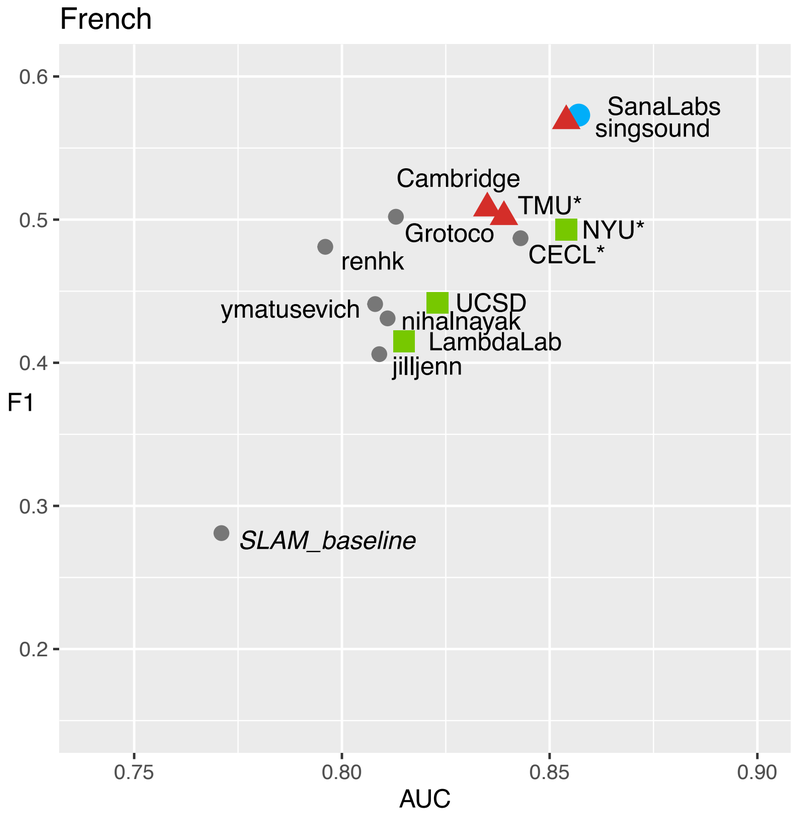
Participation Pipeline



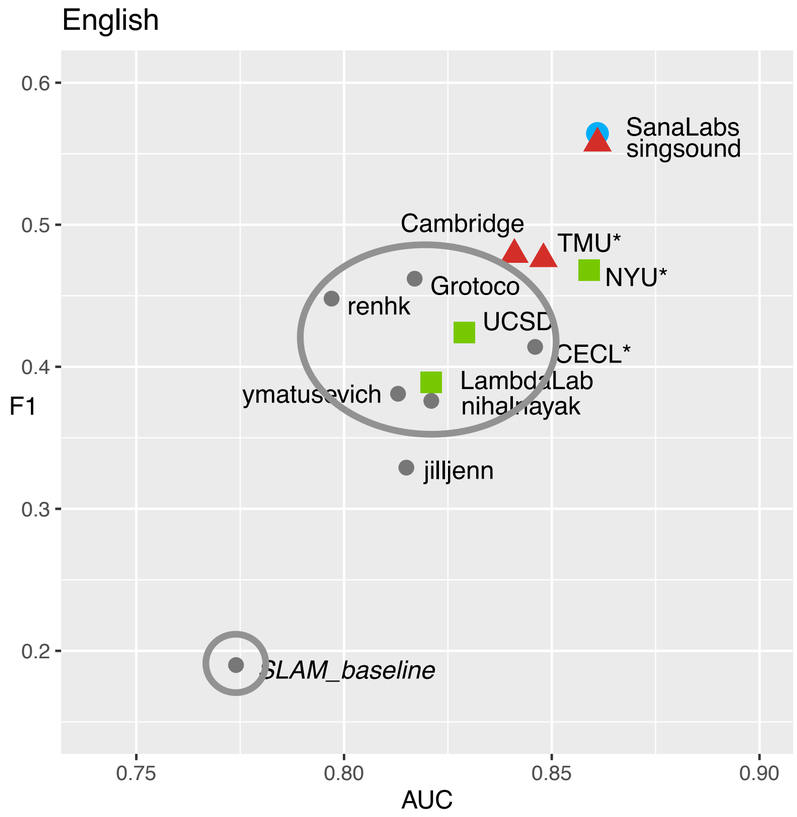
Official Results

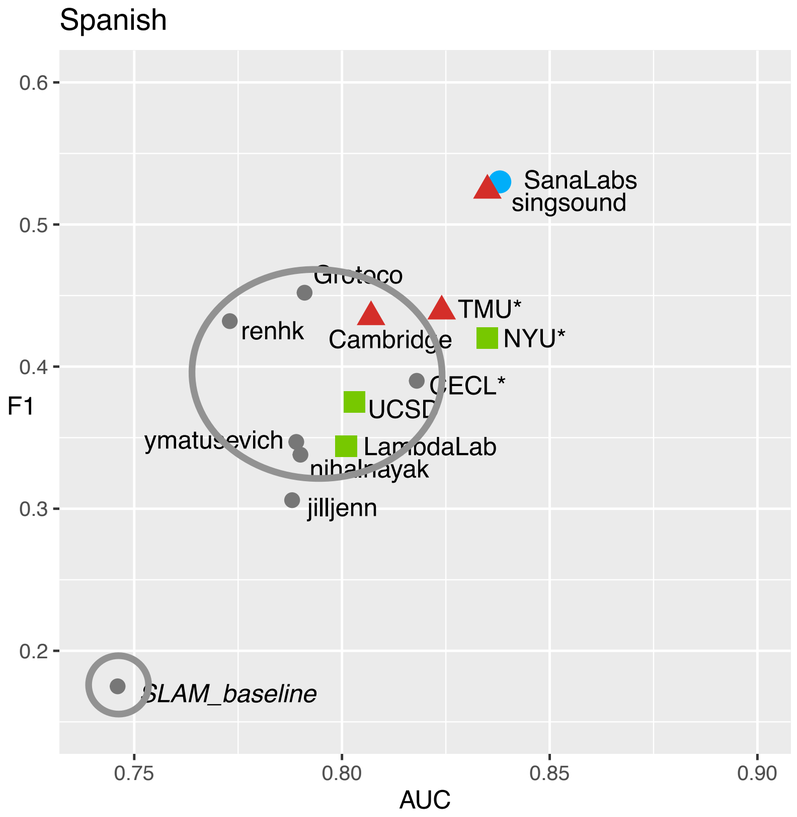


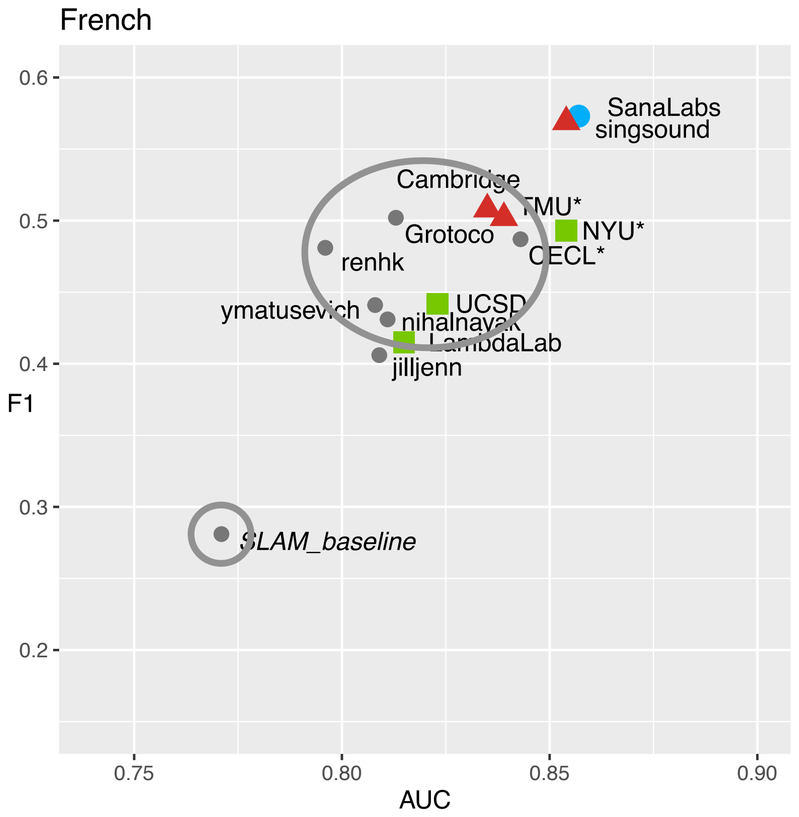




Official Results

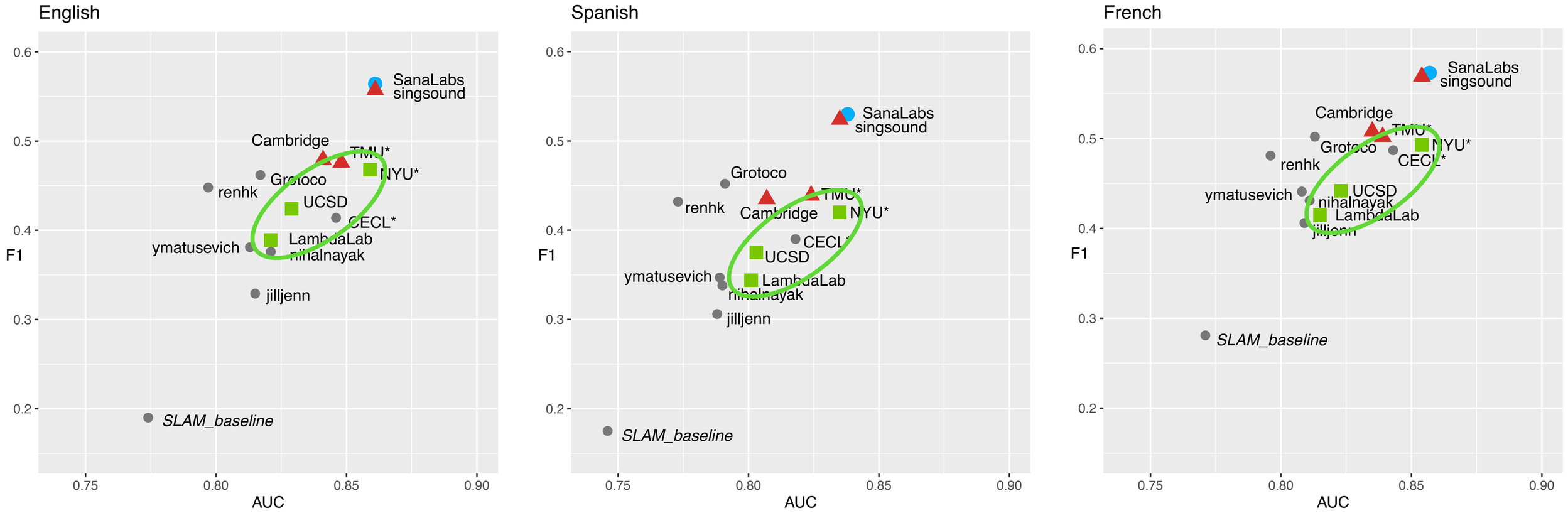




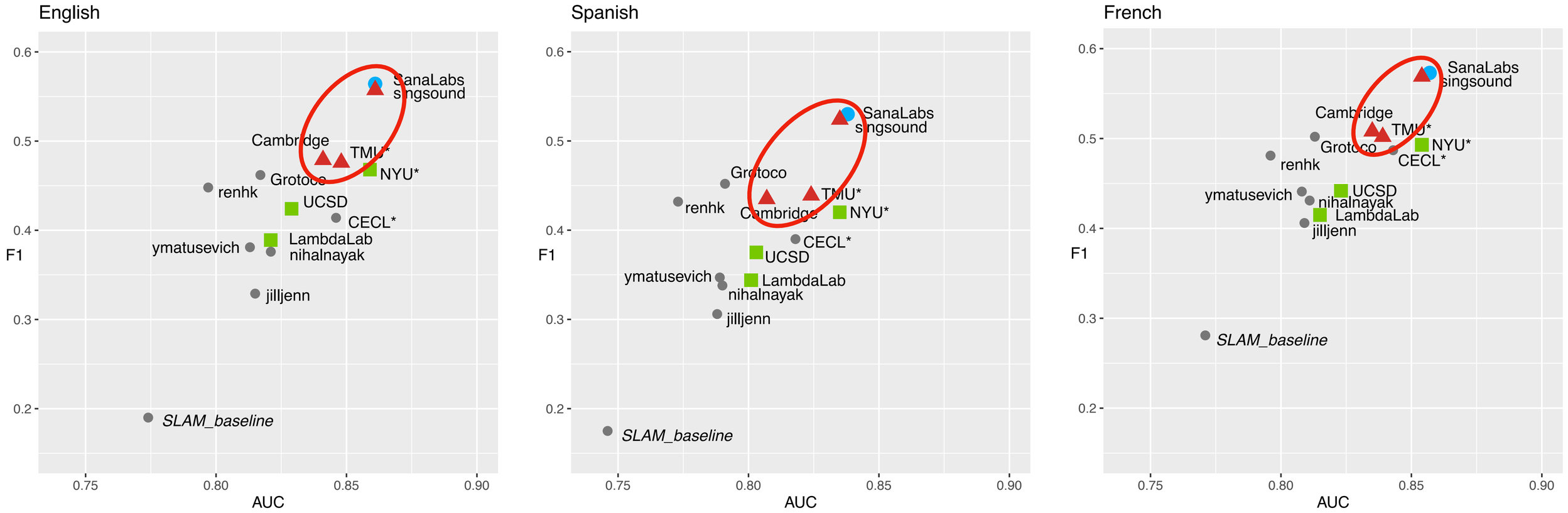


Linear models

Official Results



Official Results

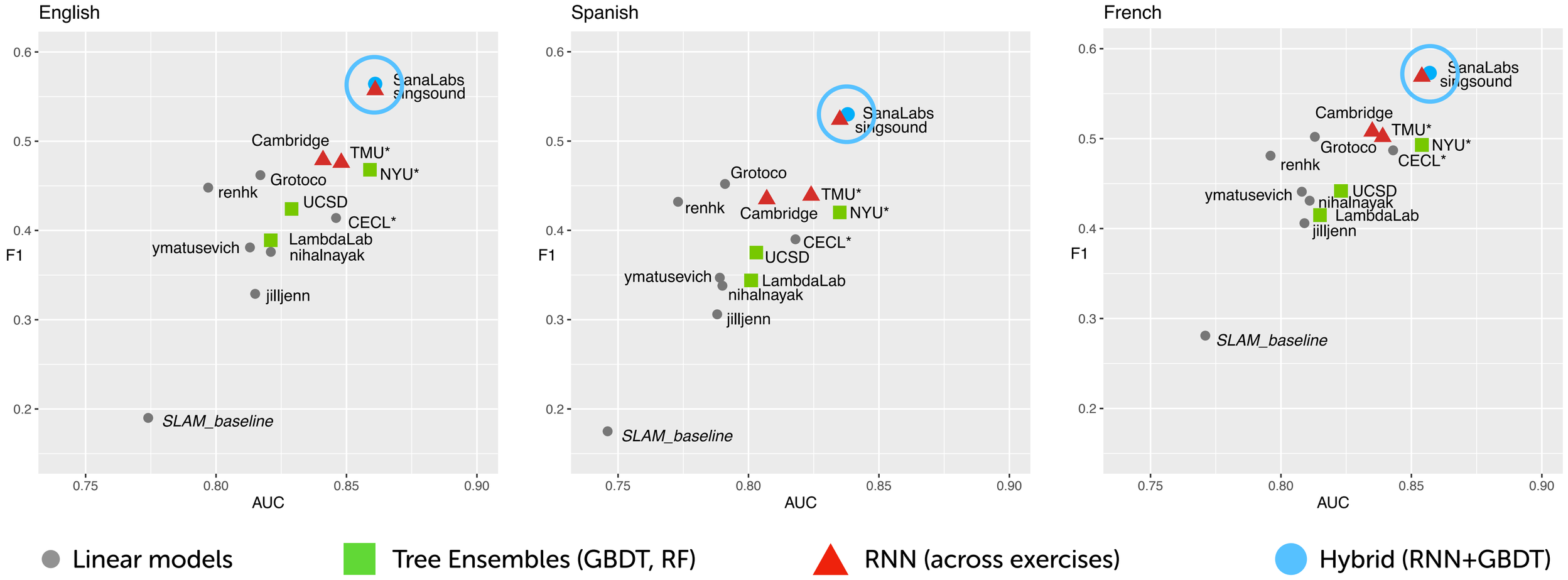


Linear models

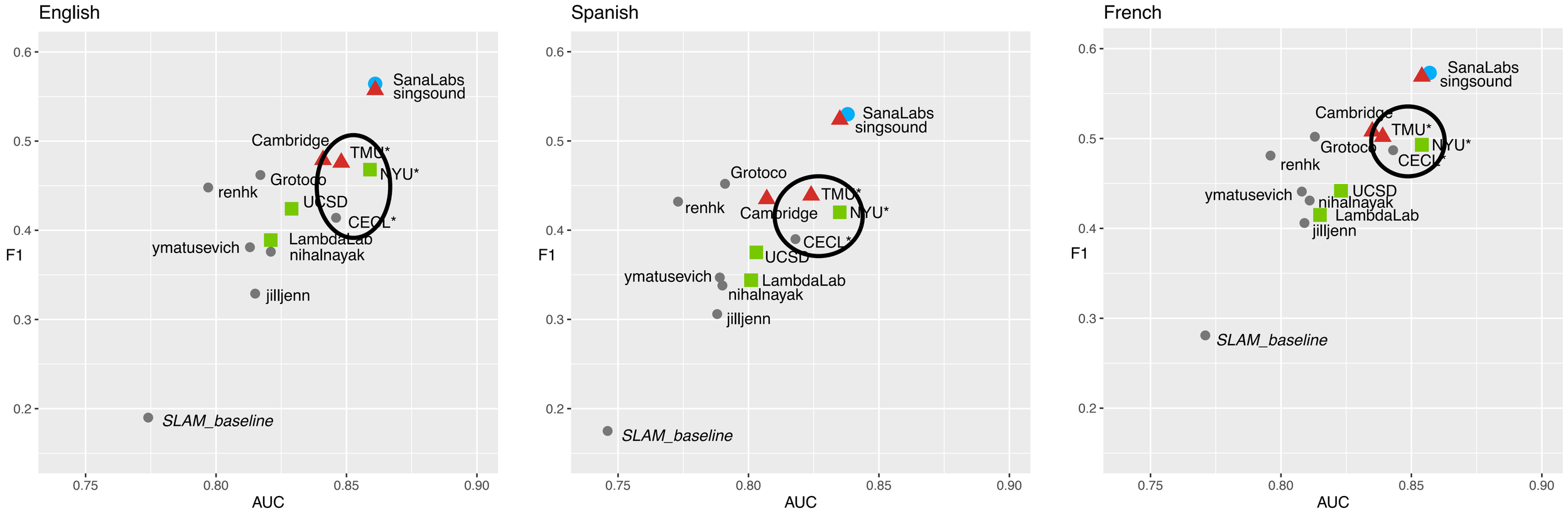
Tree Ensembles (GBDT, RF)

RNN (across exercises)

Official Results



Official Results



Linear models | Tree Ensembles (GBDT, RF) | RNN (across exercises) | Hybrid (RNN+GBDT)

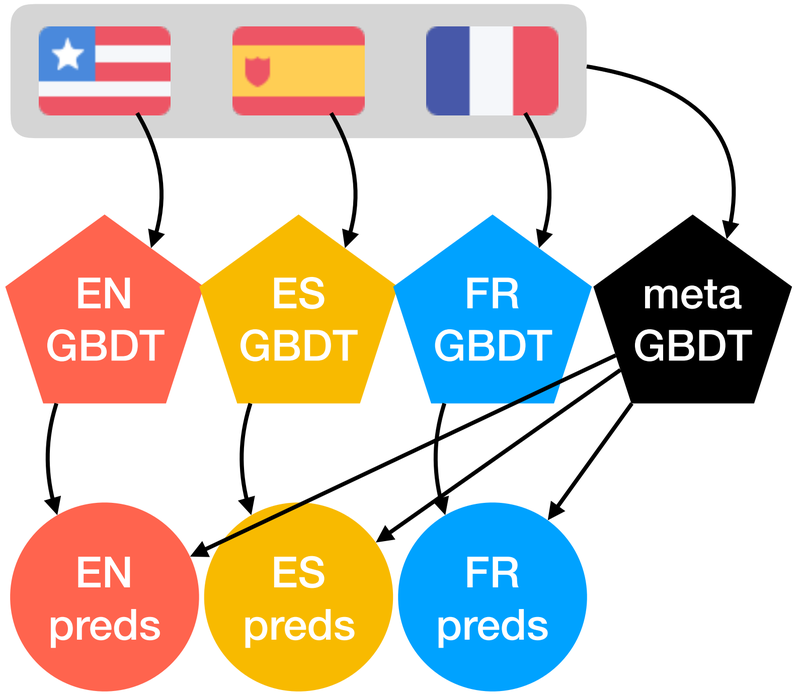
* Multitask learning (i.e., unified model across all three tracks)

Does the Algorithm Matter?

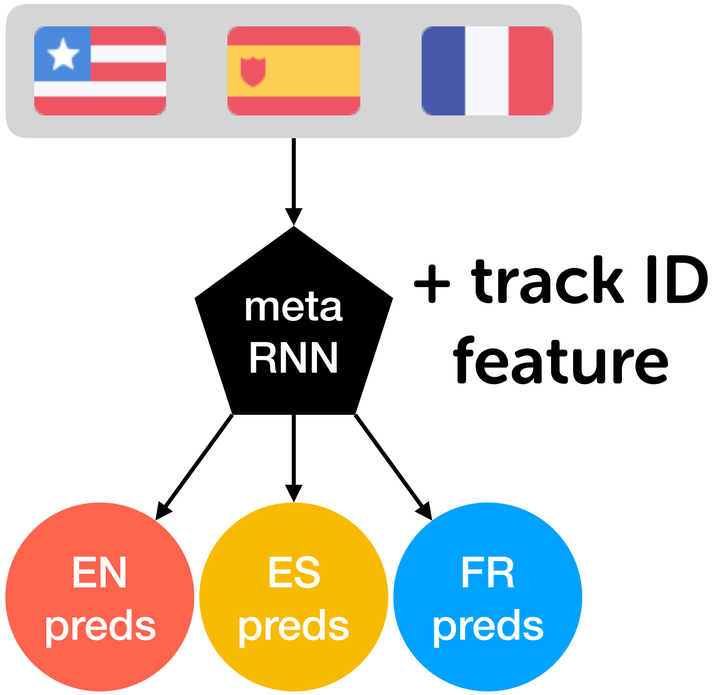
Fixed effects (algorithm choices)	**Effect**	***p*-value**	
Intercept	.786	$<.001$	***
Recurrent neural network	$+.028$	.012	*
Decision tree ensemble	$+.018$	.055	.
Linear model (e.g., IRT)	$-.006$	.541	
Multitask model	$+.023$	.017	*
Random effects	**St. Dev.**		
User ID	$\pm.086$		
Team ID	$\pm.013$		
Track ID	$\pm.011$		

linear mixed-effects analysis of **learning algorithms**

Example Multitask Approaches



NYU (Rich et al., 2018) — 3rd



TMU (Kaneko et al., 2018) — 4th

Other Algorithm Notes

- **linear classifiers** are effectively **item response theory** models, specifically AFMs (Cen et al., 2008)
- the **RNN systems** are examples of **deep knowledge tracing** (Piech et al., 2015), an extension of BKT
- the only linear model to rank in the top 5 was CECL, which used logistic regression with **feature conjunctions**
    - effectively modifies the decision surface to be **nonlinear**
    - RNN **hidden nodes** + GBDT **constituent trees** may be representing these same conjunctions

Does the Feature Set Matter?

Features used	Popularity	Effect	
Word (surface form)	▮▮▮▮▮▮▮▮▮▮▮	+.005	
User ID	▮▮▮▮▮▮▮▮▮▮	+.014	
Part of speech	▮▮▮▮▮▮▮▮▮▮	−.008	
Dependency labels	▮▮▮▮▮▮▮▮	−.011	
Morphology features	▮▮▮▮▮▮▮▮	−.021	
Response time	▮▮▮▮▮▮▮▮	+.028	*
Days in course	▮▮▮▮▮▮▮	+.023	.
Client	▮▮▮▮▮▮▮	+.005	
Countries	▮▮▮▮▮▮	+.012	
Dependency edges	▮▮▮▮	−.000	
Session	▮▮▮	+.014	

time-related features appear to help somewhat

morpho-syntactic features seem to hurt slightly?

linear mixed-effects analysis of **provided features**

Parsing (+ Alignment) Errors

Token	POS	DEP	Label
A	DET	det	0
man	NOUN	ROOT	0
a	PUNCT	punct	0
woman	DET	det	0

Token	POS	DEP	Label
The	DET	det	1
judge	ADJ	amod	1
returns	NOUN	ROOT	1

Cambridge (Yuan, 2018)

Along with the tokens themselves we encoded each instance word's part of speech, morphological features, and dependency edge label. We noticed that some words in the original dataset were paired with the wrong morphological features, particularly near where punctuation had been removed from the sentence. To fix this, we reprocessed the data using Google SyntaxNet[3].

NYU (Rich et al., 2018)

Does the Feature Set Matter?

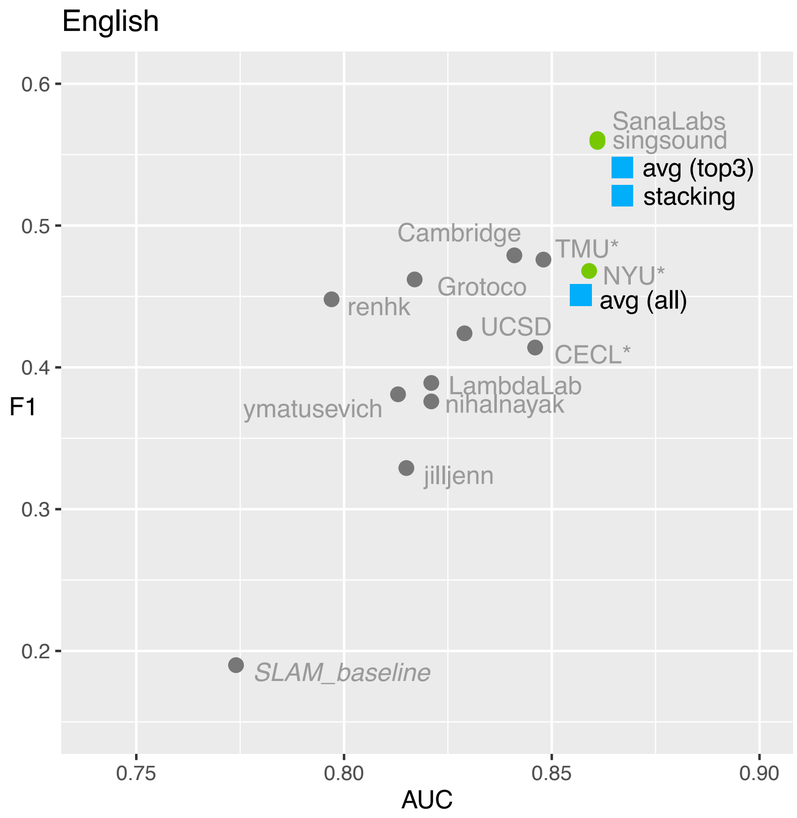
> 30 days might make these more useful

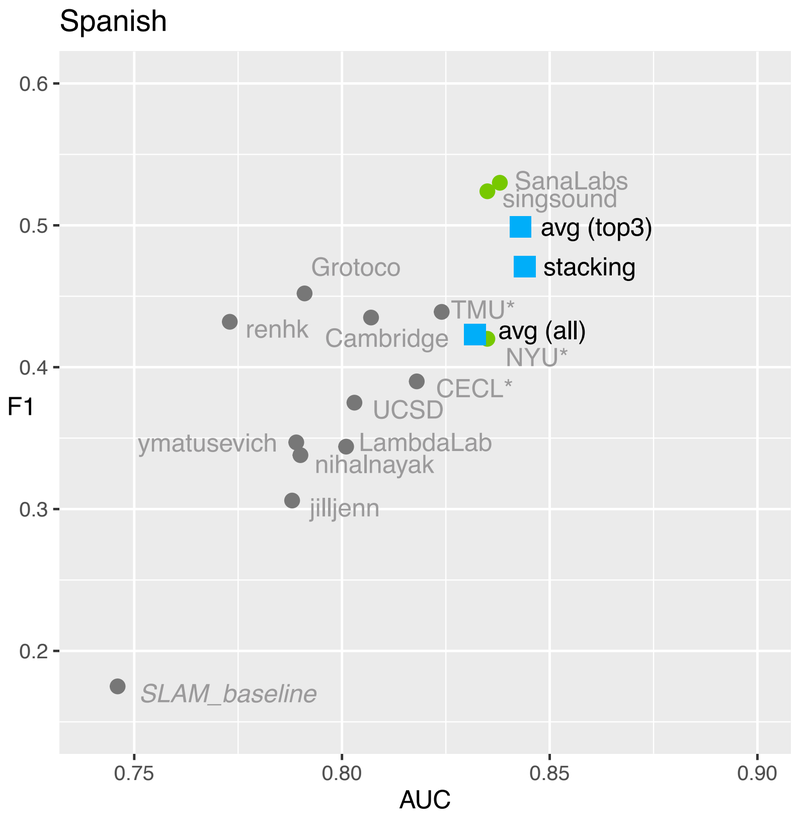
Features used	Popularity	Effect
Word corpus frequency	▮▮▮▮▮▮	+.008
Spaced repetition features	▮▮▮▮▮	+.013
L1-L2 cognates	▮▮▮▮	+.001
Word embeddings	▮▮▮	+.020
Word stem/root/lemma	▮▮▮	+.007

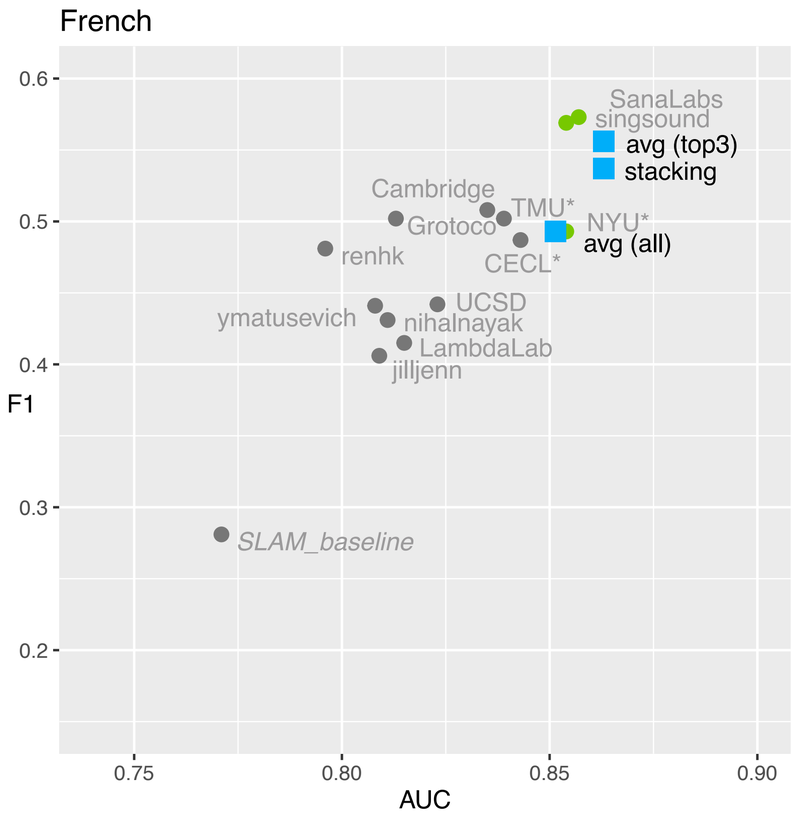
more linguistically diverse data might make these more useful

linear mixed-effects analysis of **novel features**

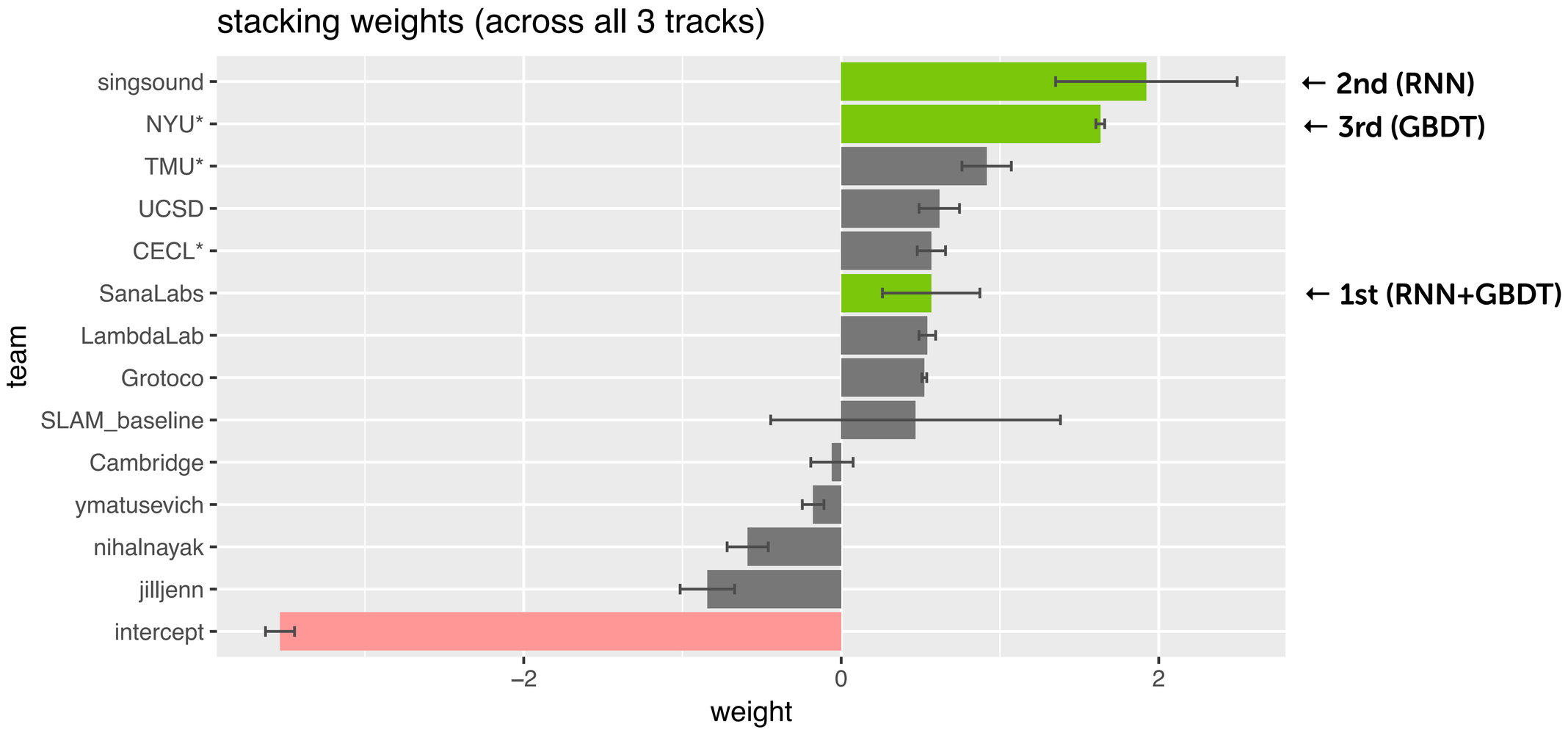
Can An Ensemble Do Better?







Can An Ensemble Do Better?



Summary

- **first SLA modeling task:** attracted 15 teams from diverse fields
- **learning algorithm choices** (RNNs, GBDTs, multitask) appear to be more impactful than **clever feature engineering**
- morpho-syntactic features **did not seem to help**, possibly due to systematic parsing (+ alignment) errors
- a more **longitudinal** SLA modeling task (> 30 days) + more linguistic diversity (multiple L1s; intermediate-advanced) might let **psychologically-inspired features be more useful**

Questions?

corpus, papers, starter code, etc. available at:
http://sharedtask.duolingo.com

special thanks to: Bozena Pajak, Joseph Rollinson, Hideki Shima, Eleanor Avrunin, Natalie Glance,
Anastassia Loukina, Kristen K. Reyher, the BEA workshop organizers,
+ of course the participating teams!