Baby steps to language

2019-10-03 @ PAISS

Alejandrina (Alex) Cristia

Laboratoire de Sciences Cognitives et Psycholinguistique

Language Acquisition Across Cultures Team

Thanks to my team for feedback on the slides!

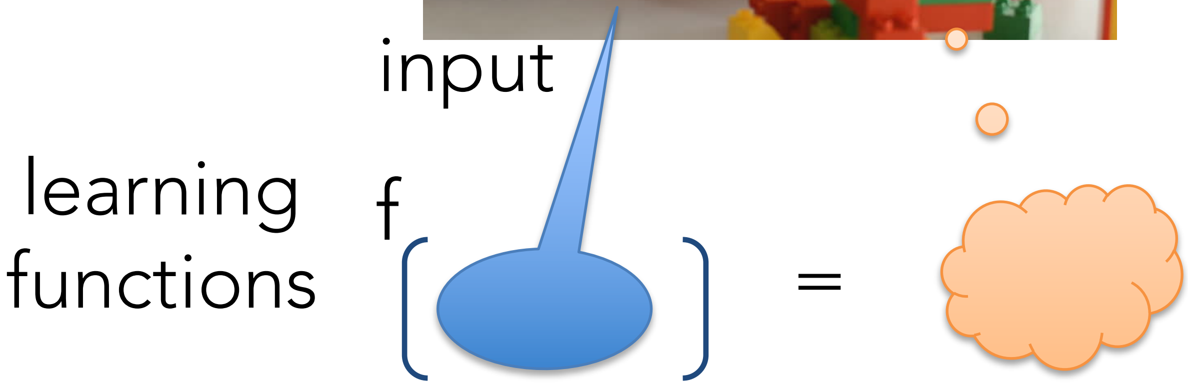
Which of the following are true?

- Newborns prefer listening to their native language than to an unfamiliar language
- Newborns know their name
- By 6 months, babies know their name
- By 6 months, babies say their first word
- By 12 months, babies say their first word

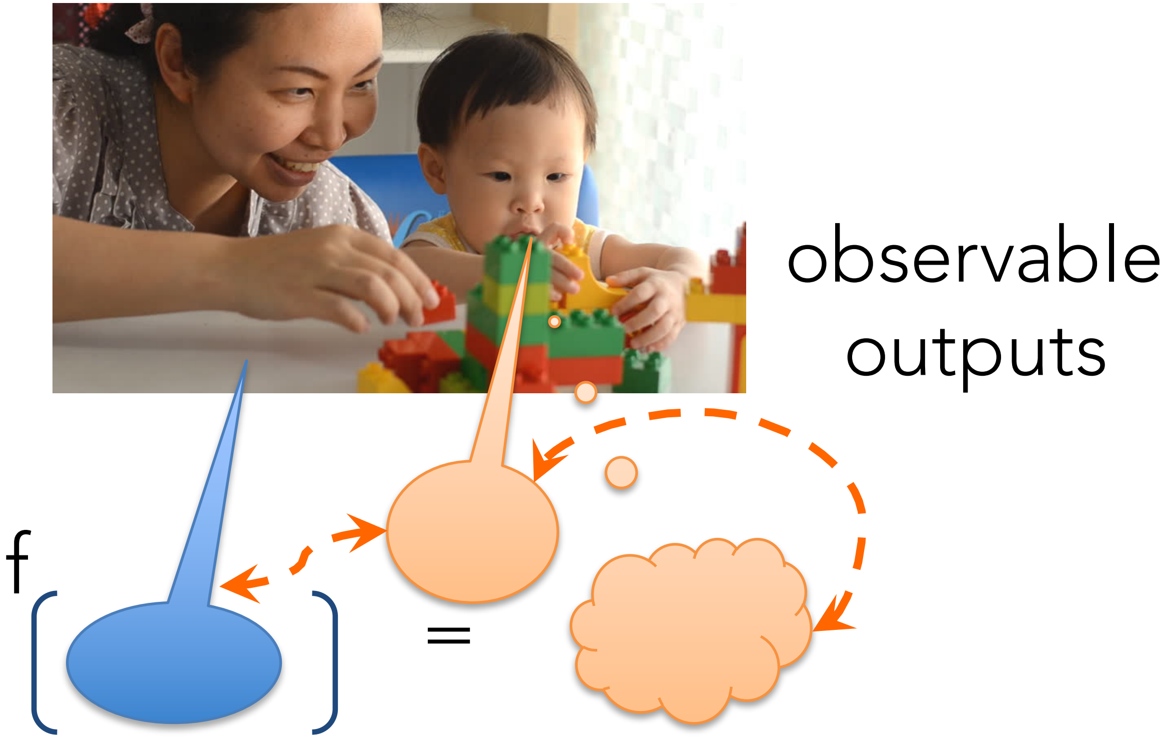
A broad language acquisition theory (v 1.0)

Mental representations appropriate to native language(s)

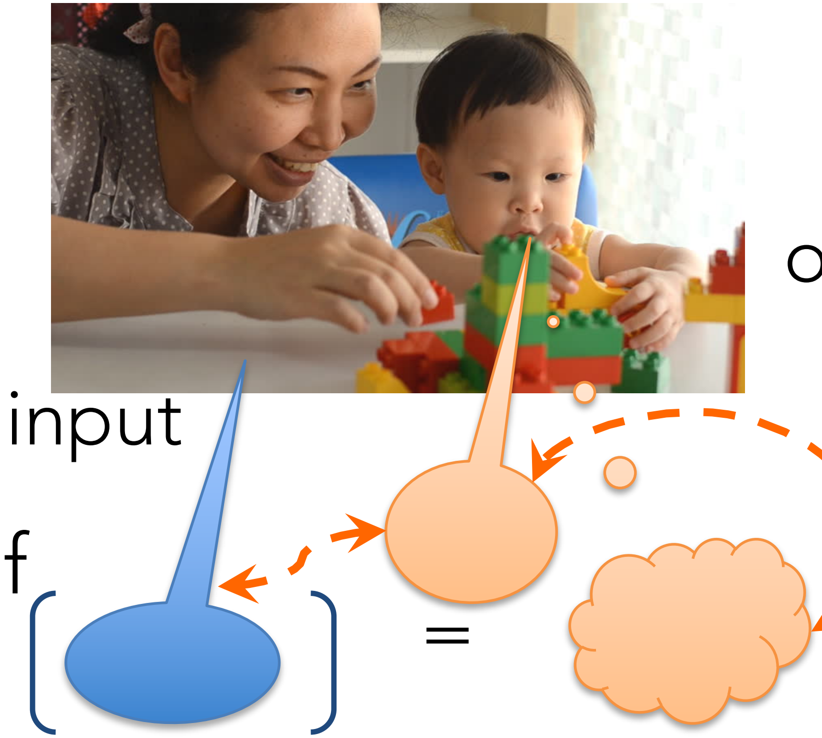
A broad language acquisition theory (v 1.0)



A broad language acquisition theory (v 1.0)

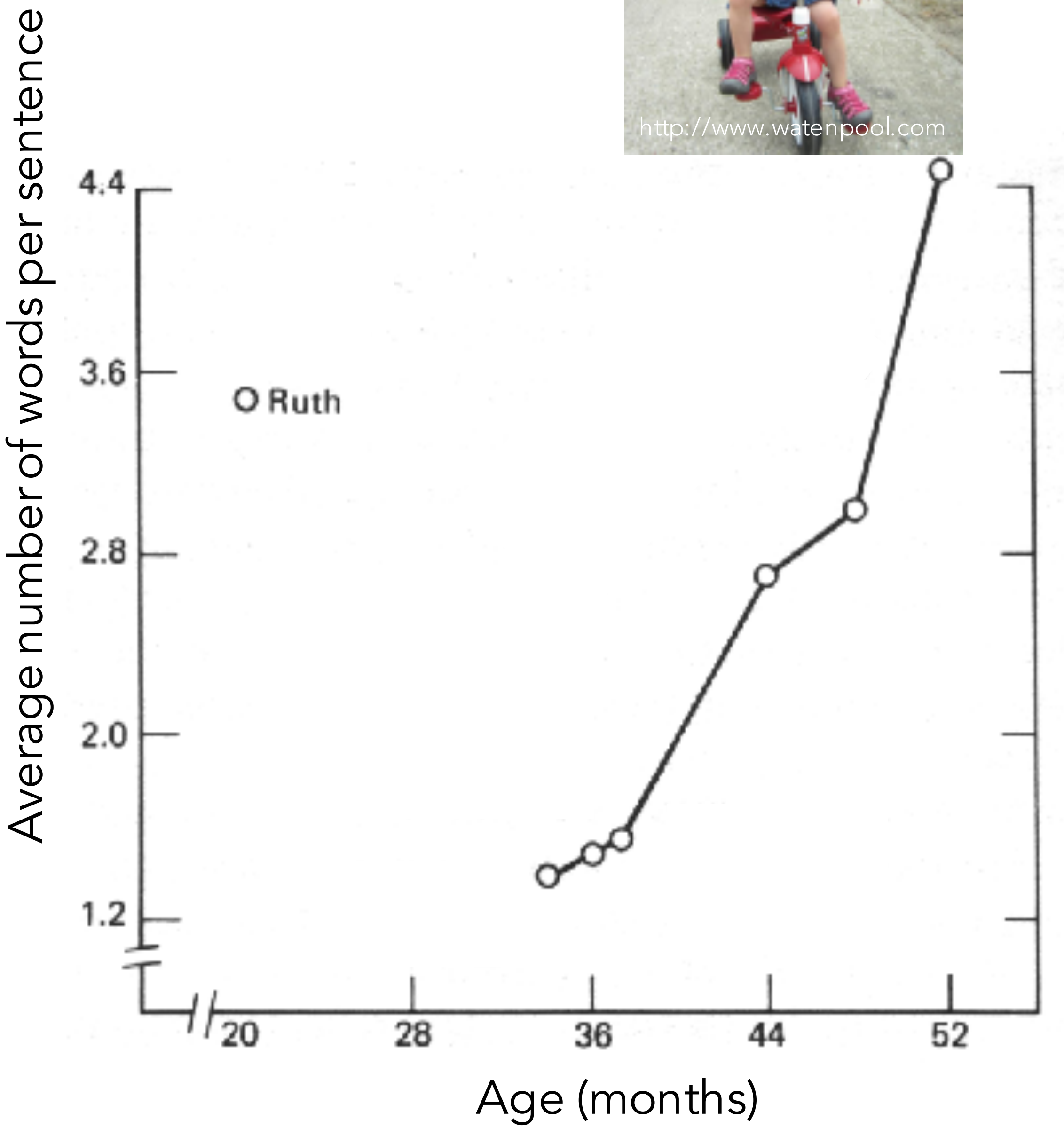


A broad language acquisition theory (v 1.0)

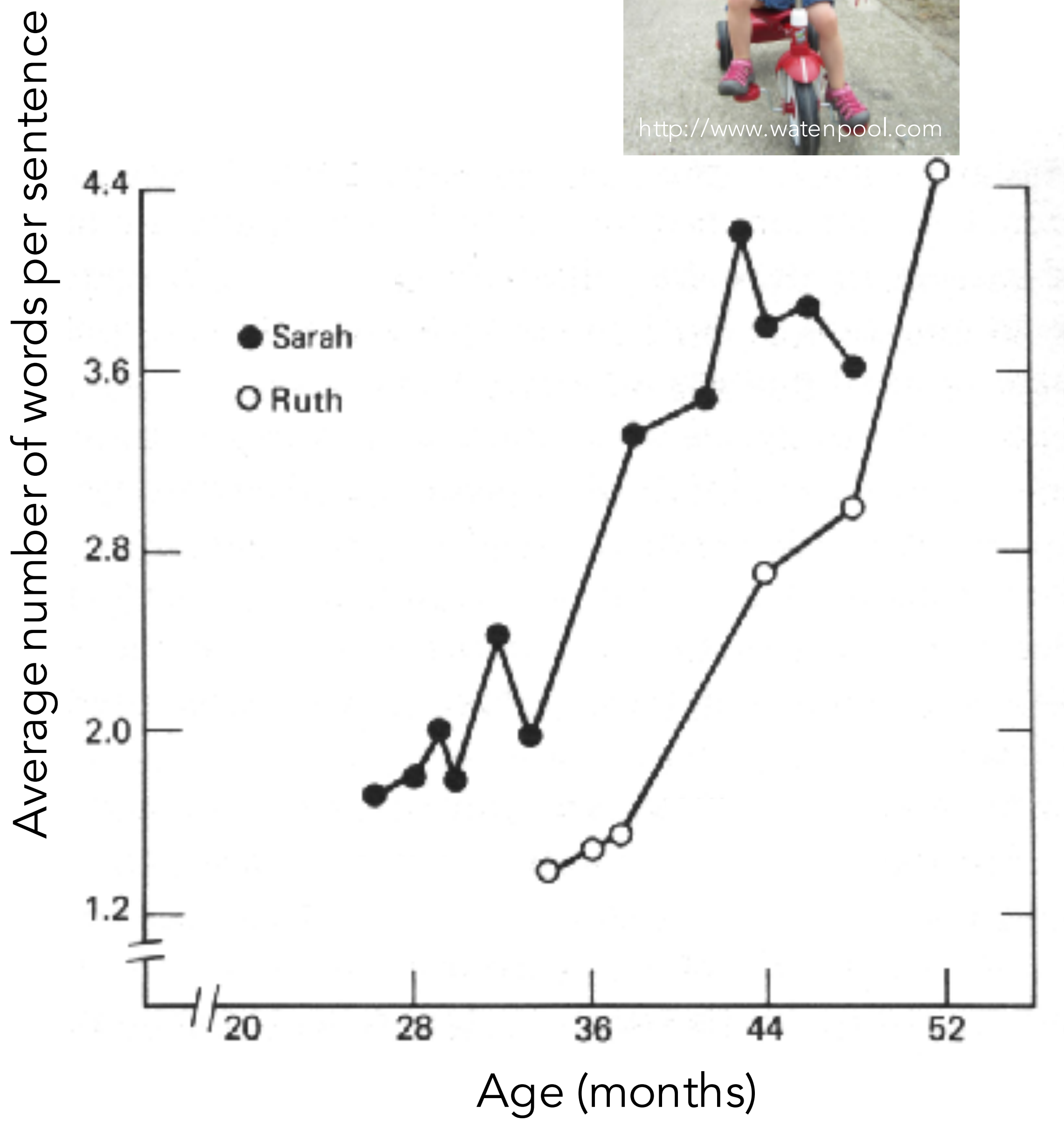


Which of the following are true?

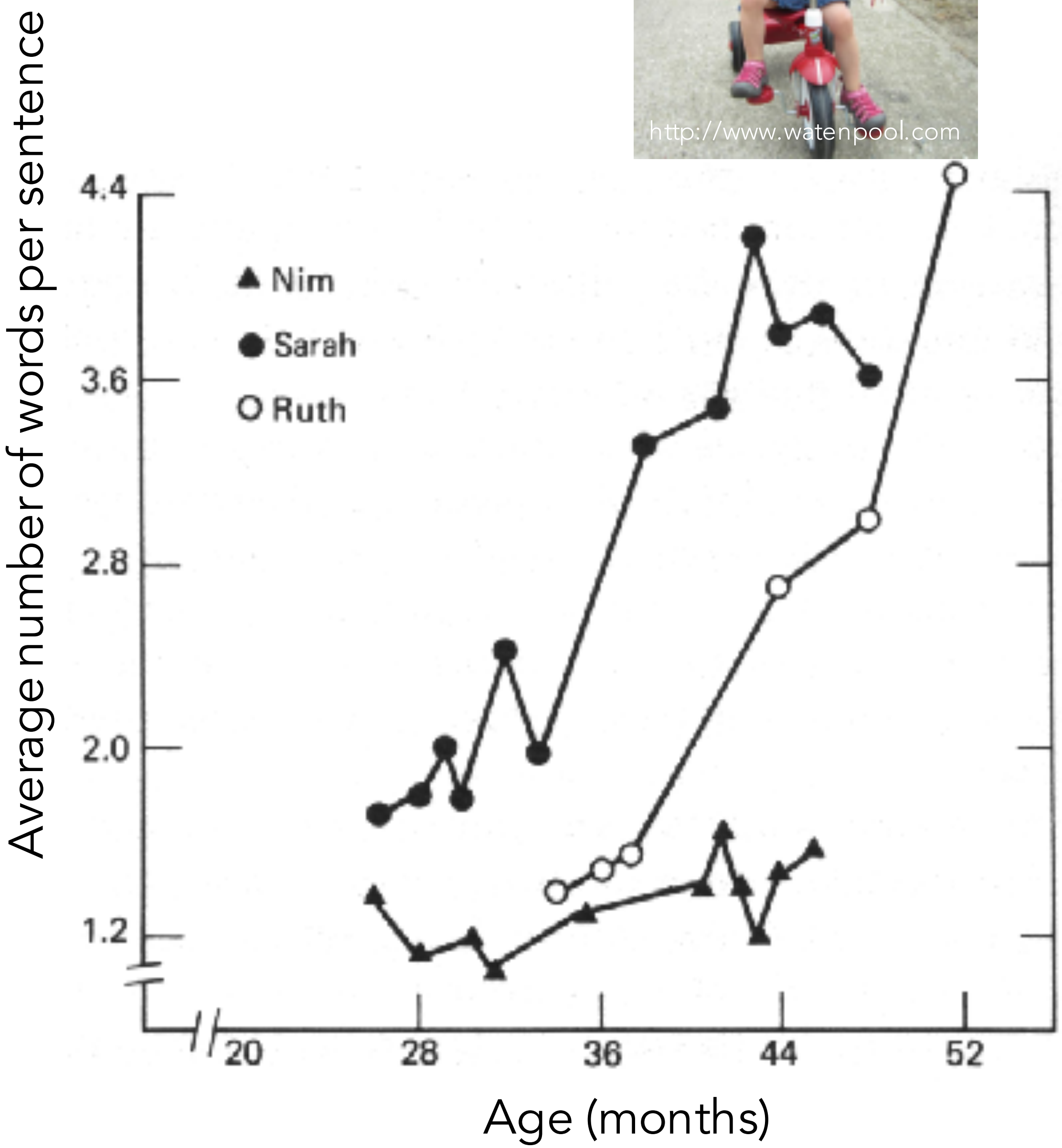
- Humans and chimpanzees share a majority of their genetic information
- In terms of their visual skills, humans and chimpanzees are more similar to each other than humans and killer whales are
- In terms of their communication system, humans and chimpanzees are more similar to each other than humans and killer whales are
- You can raise a chimpanzee to use language like human babies do



Terrace 1979 Science

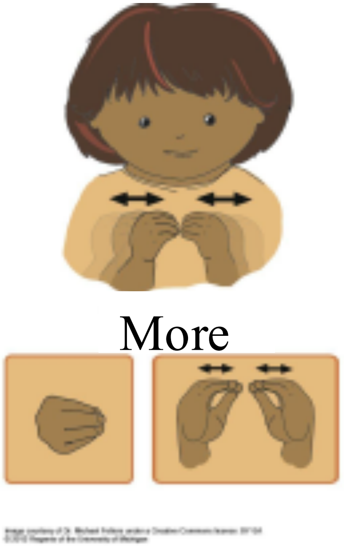


Terrace 1979 Science

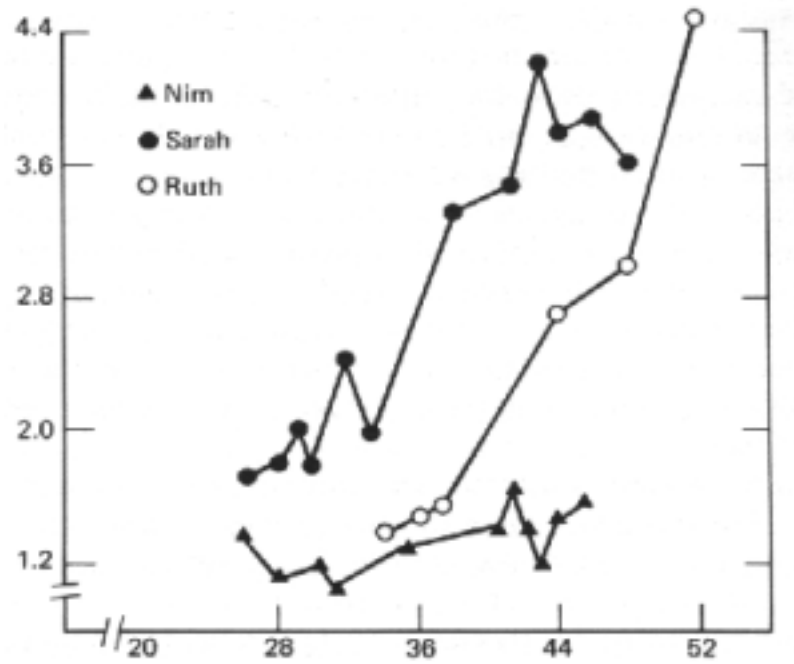


Terrace 1979 Science

Innate



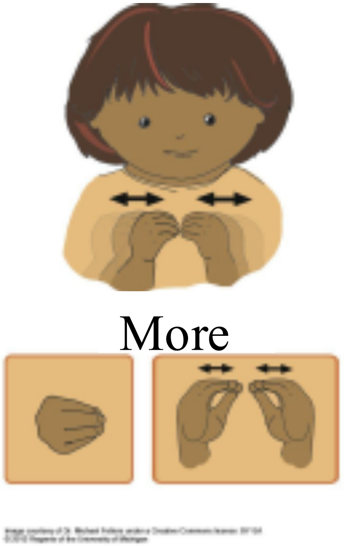
Terrace 1979 Science



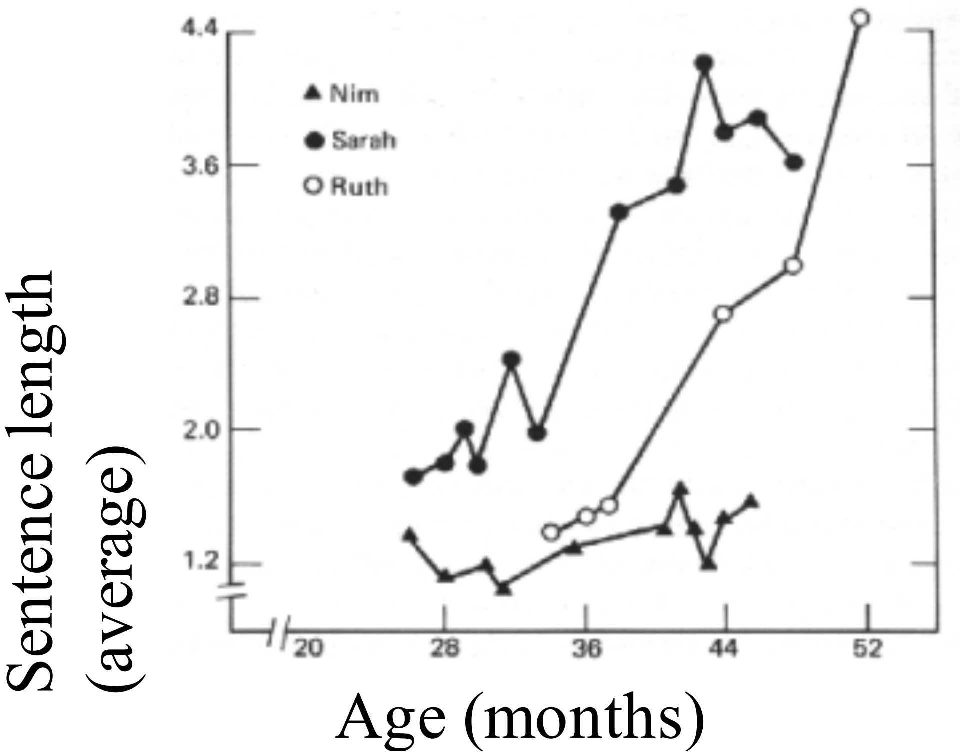
Sentence length (average)

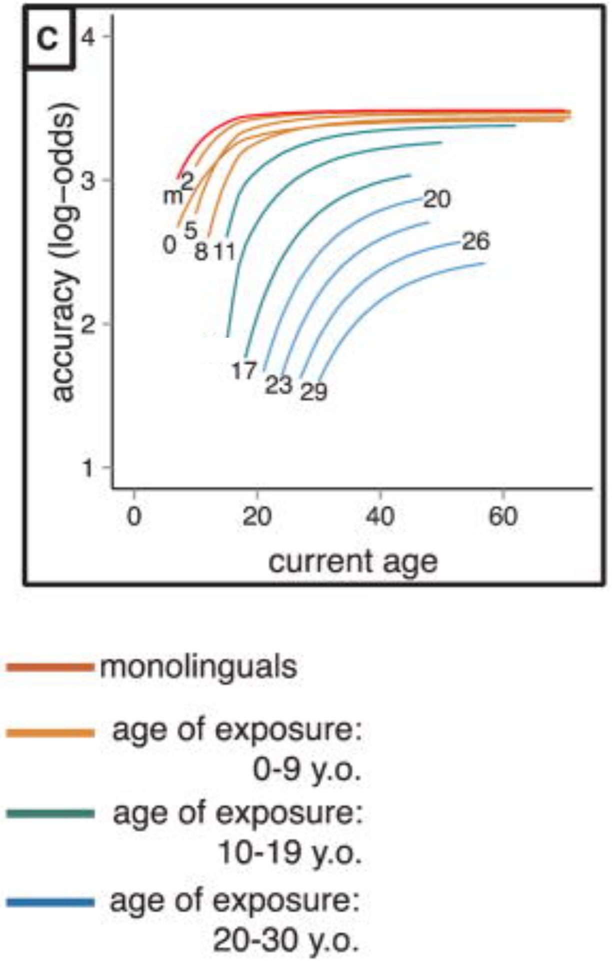
Age (months)

Innate & acquired

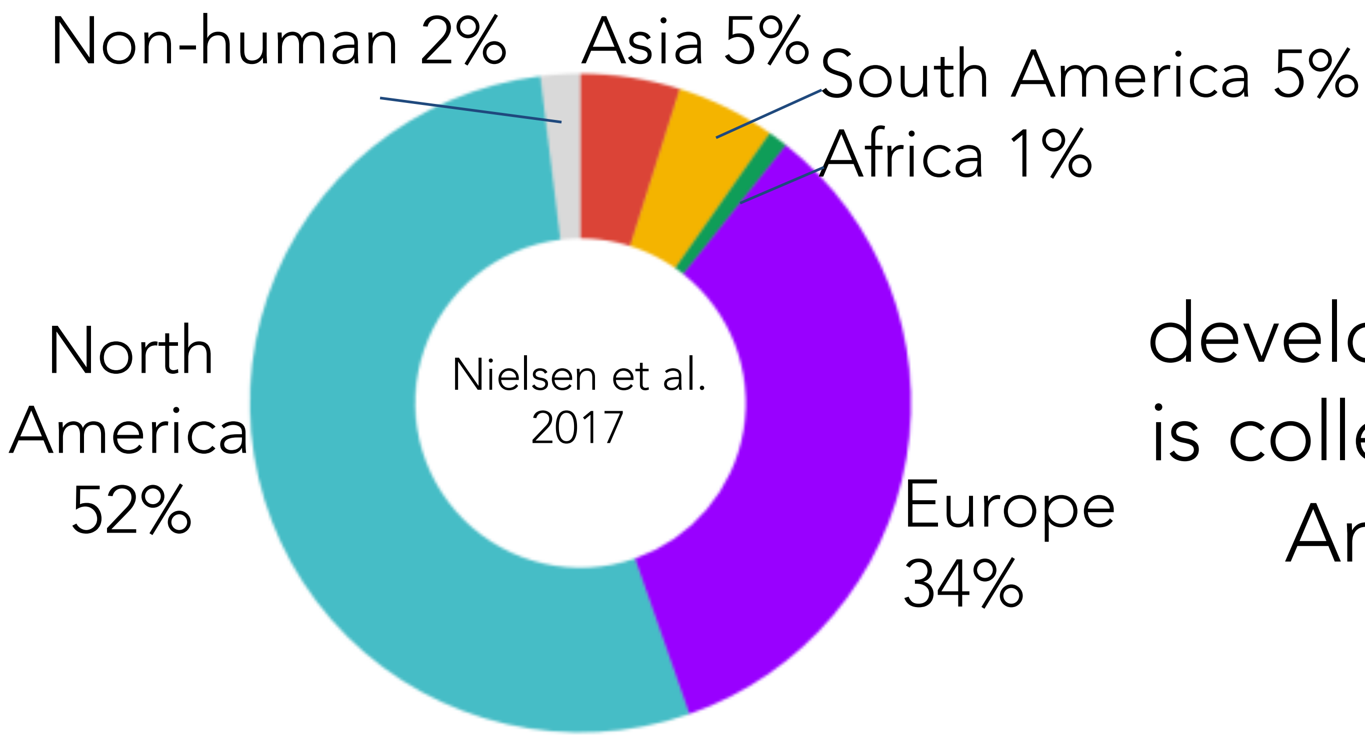


Terrace 1979 Science

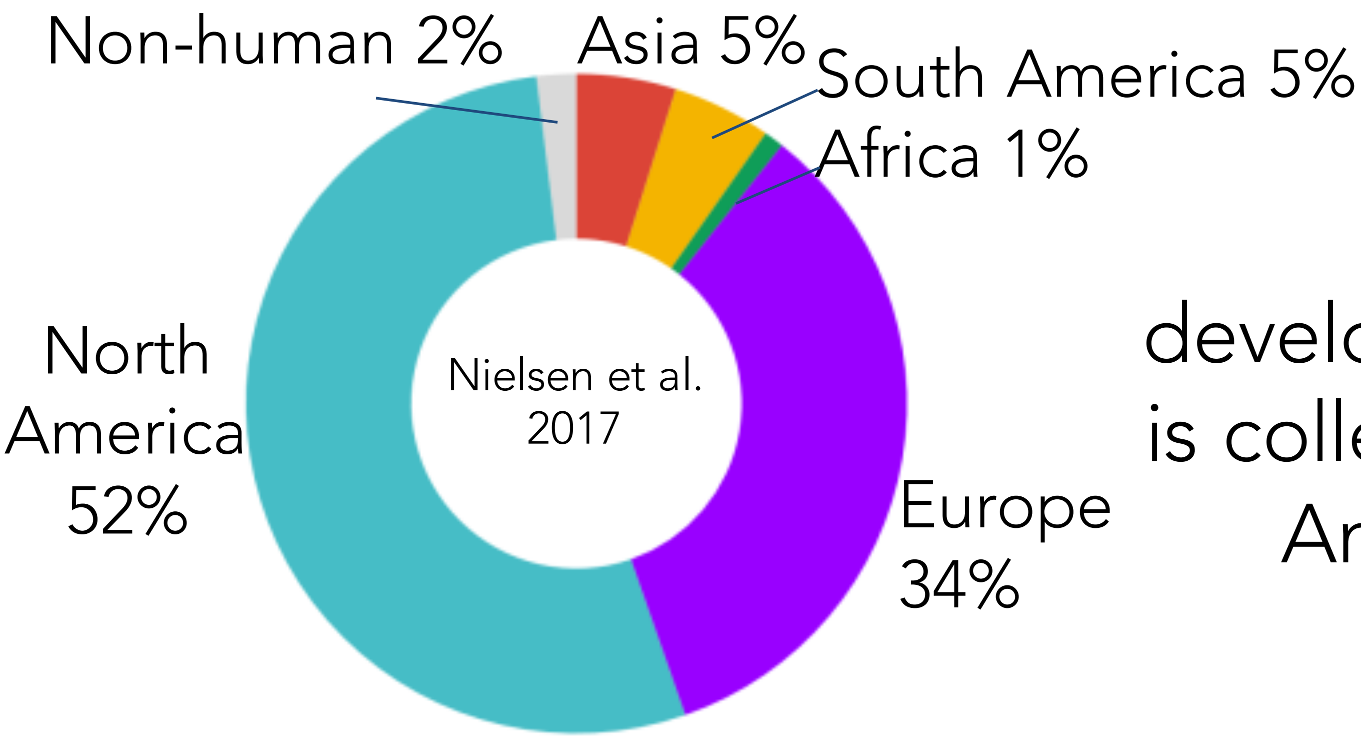




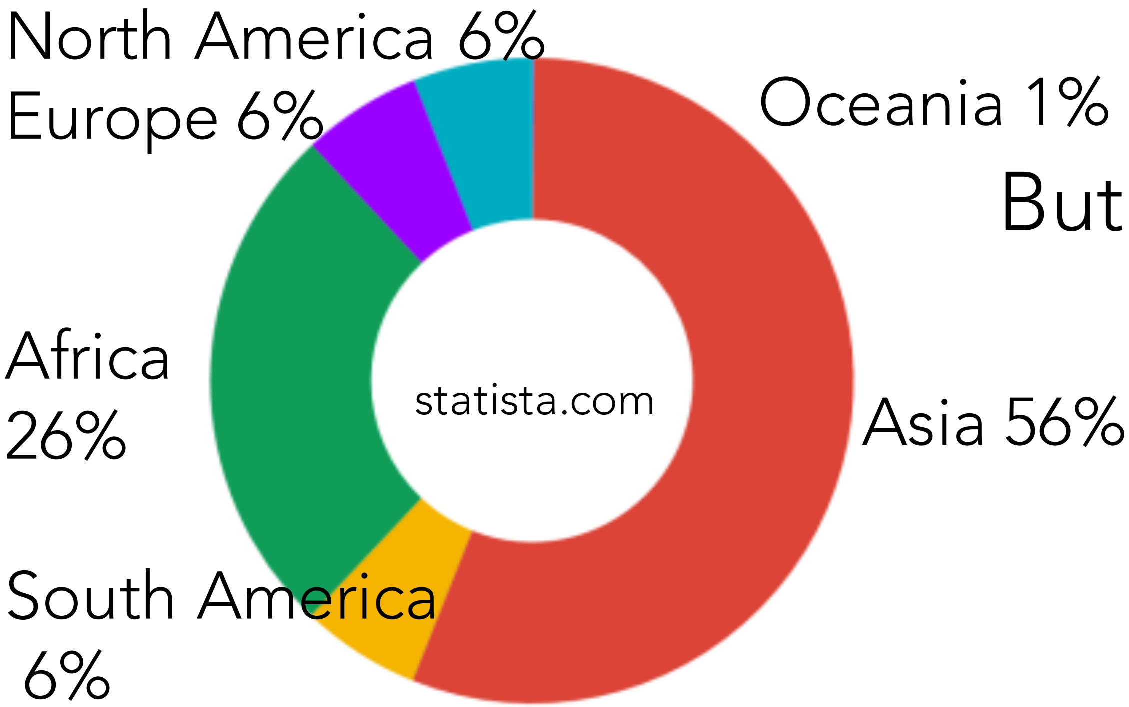
Hartshorn et al. 2018 Cognition



Most developmental data is collected in North America and Europe



Most developmental data is collected in North America and Europe

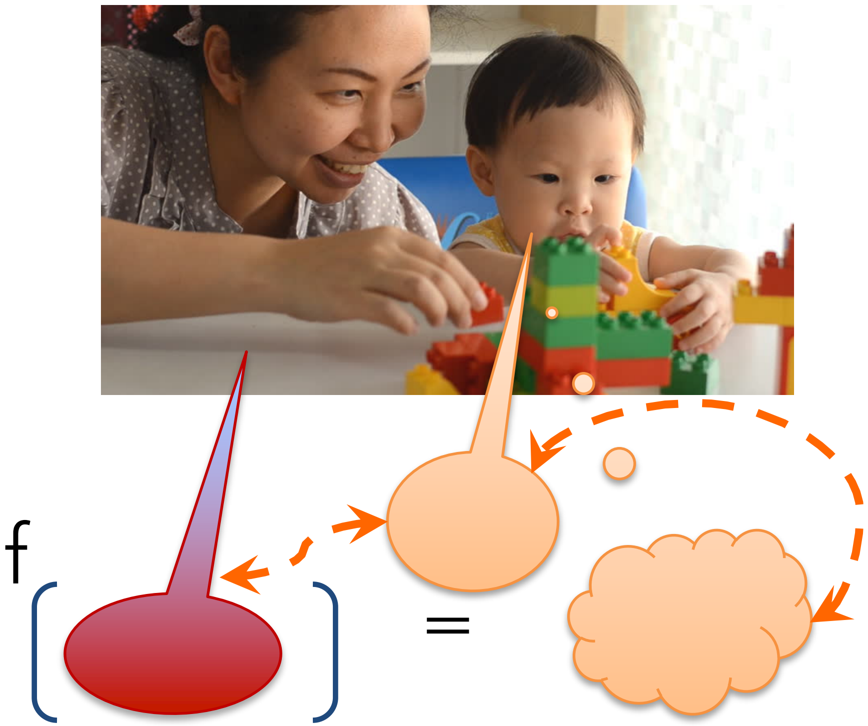


But most children live in Asia and Africa

Who grew up in…

- Europe
- North America
- South America
- Africa
- Asia
- Oceania

High quantity of high quality input

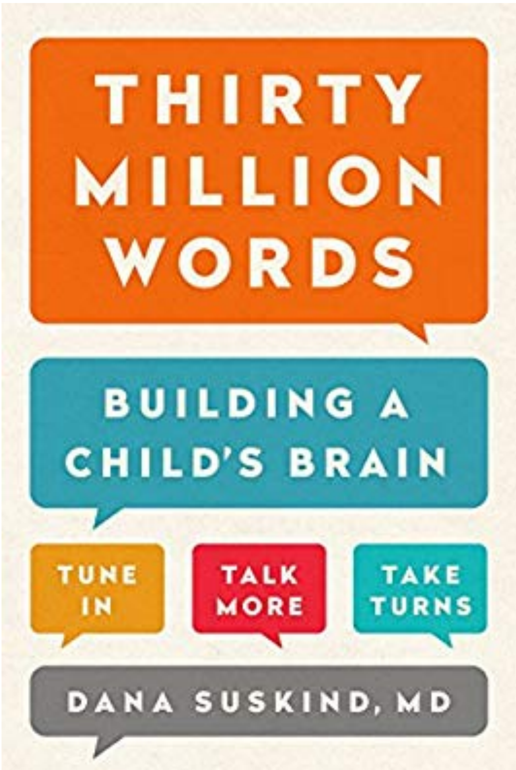


Adults' speech is **high** quality
- a stable linguistic system
- developed “theory of mind”

One on one
- topics adapted to child’s attention & abilities
- use of “Parentese”

FIRST 5 CALIFORNIA

PROVIDENCE TALKS
TALK TO TEACH



TALK WITH ME BABY

PEQUEÑOS Y VALIOSOS

UNIVISION CONTIGO

Thanks to Janet Bang for this selection!

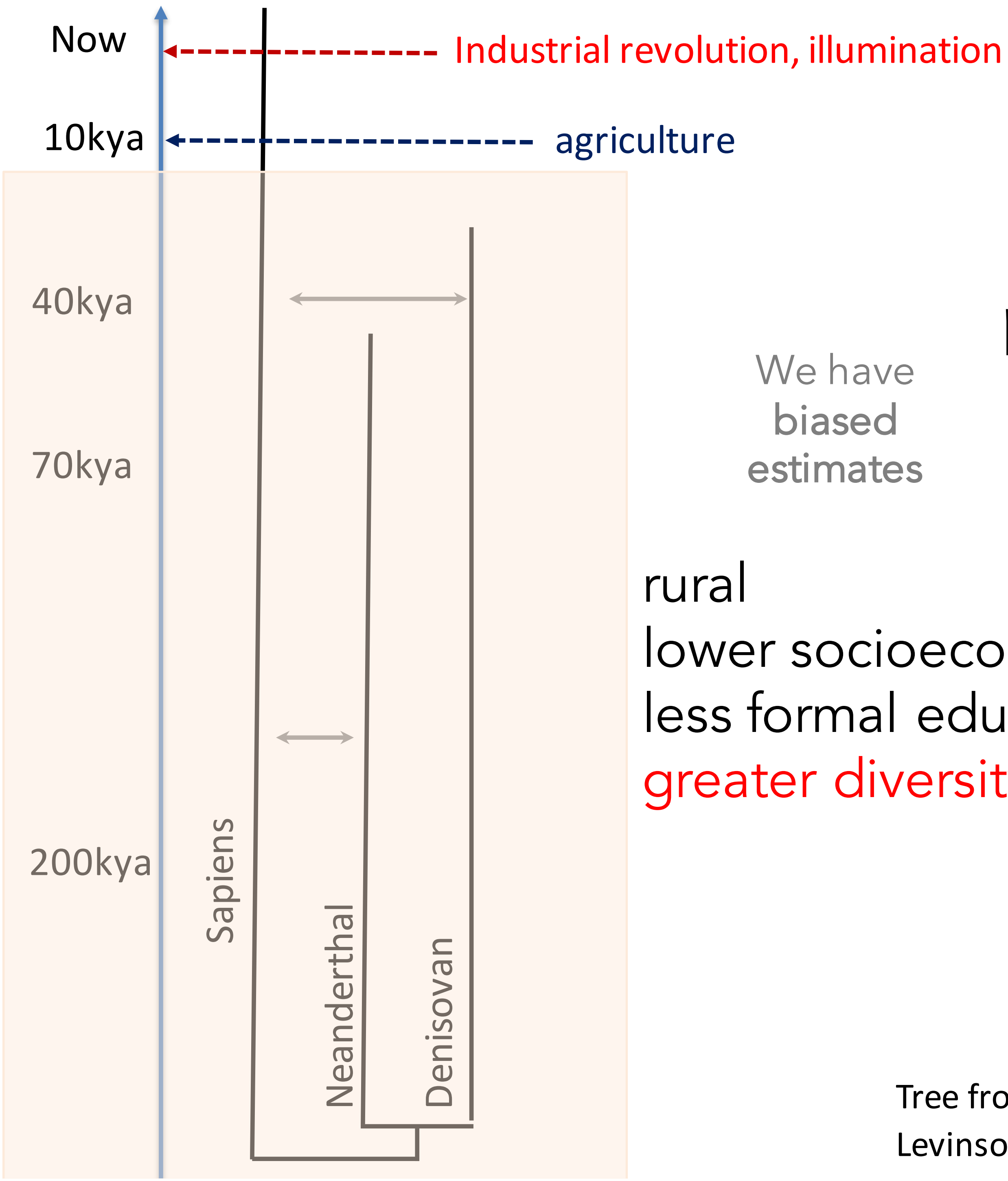
The average family across continents

WEIRD= Western, Educated, Industrialized, Rich, Democratic;

Heinrich et al. 2010

industrialized
higher socioeconomic status
more formal education
fewer children
single caregiver

rural
lower socioeconomic status
less formal education
more children
shared caregiving

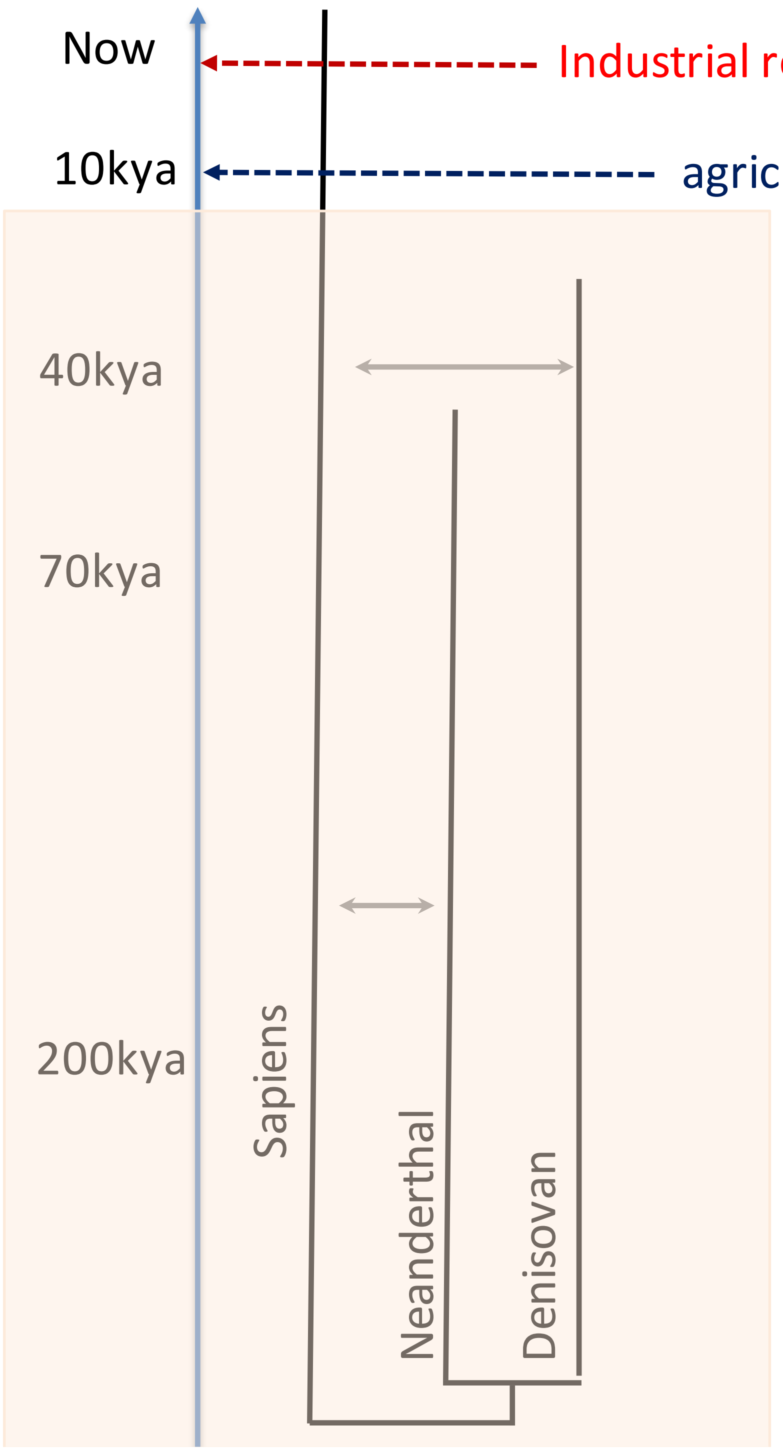


We have biased estimates

WEIRD settings do not represent human ecology

rural
lower socioeconomic status
less formal education
greater diversity in ecological settings

Tree from Dediu & Levinson 2013, *Frontiers*
Levinson & Holler, 2014 *Phil.T.R.Soc*.



We have
biased
estimates

WEIRD settings do not represent human ecology

rural
lower socioeconomic status
less formal education
greater diversity in ecological settings

Child-rearing among hunter-gatherer communities

- **Universal**
- Co-sleeping & physical contact
- Maternal primacy <1y
- Multi-age groups >1y
- Frequent breast-feeding

- **Variation**
- Non-maternal care
- Self-provisioning
- Assigned chores
- Father involvement
- Weaning age/ inter-birth interval duration

Variation in reproductive strategies

e.g. in number of children

Konner 2016

Hewlett et al. 2000

higher prevalence child-directed speech predicted

!Kung

hunter-gatherers

average # children: 4

Konner 2016

Tsimane’

hunter-farmers

average # children: 9

Stieglitz et al. 2013

lower prevalence child-directed speech predicted*

*at least due to competition

higher prevalence child-directed speech predicted

!Kung

hunter-gatherers

average # children: 4

Konner 2016

TO BE CONTINUED

Tsimane’

hunter-farmers

average # children: 9

Stieglitz et al. 2013

lower prevalence child-directed speech predicted*

*at least due to competition

homebank.talkbank.com


© Tsimane project


© Wikipedia

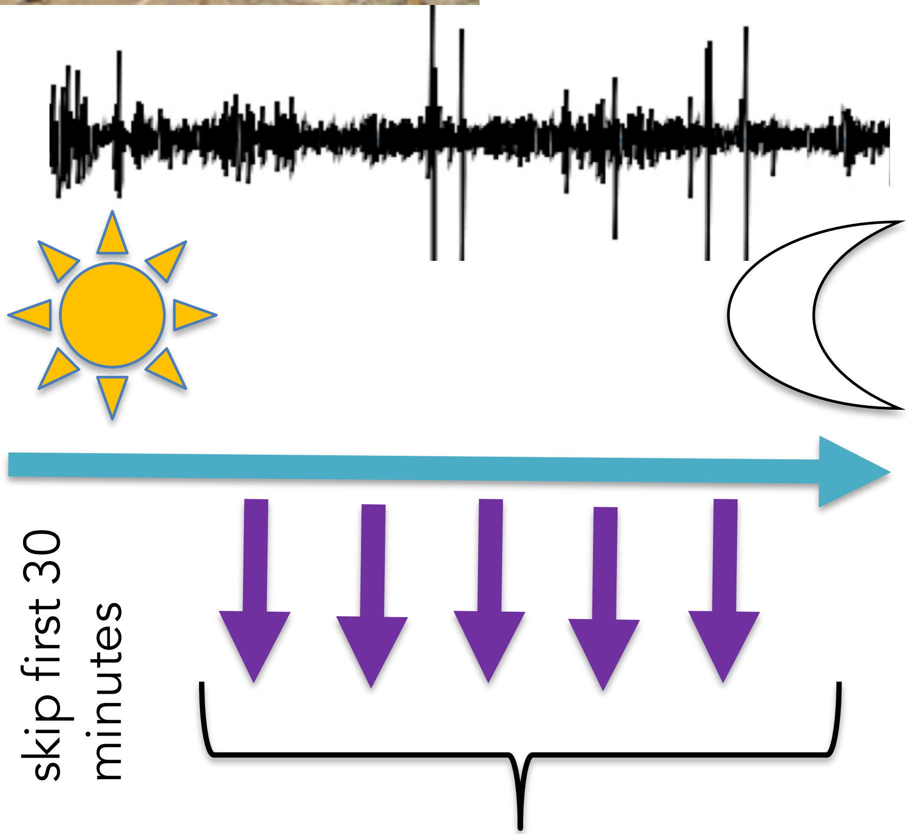

© Crumb imagecity


Photo credit:
Heidi Colleran

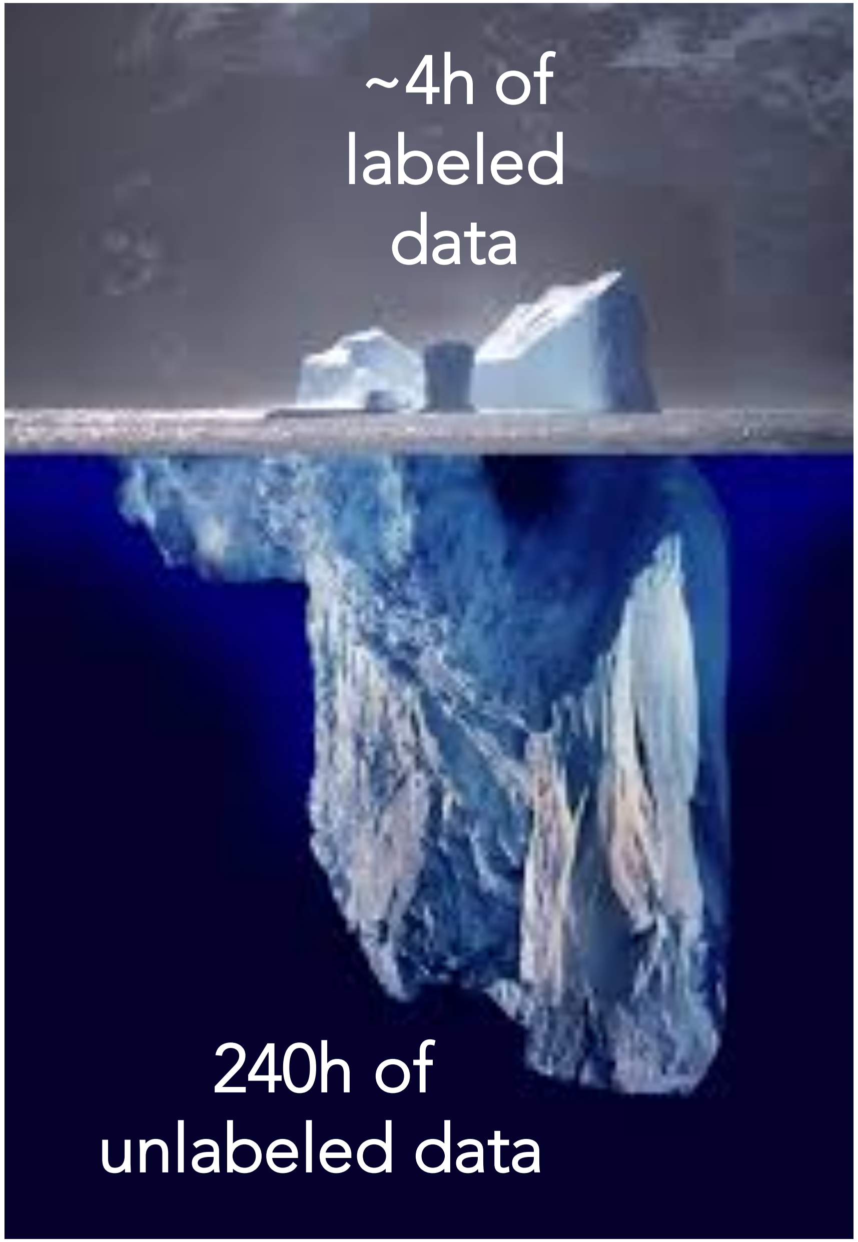
+ ecological
+ coverage

15 hours
(15$)

Casillas &
Cristia (2019)
Collabra

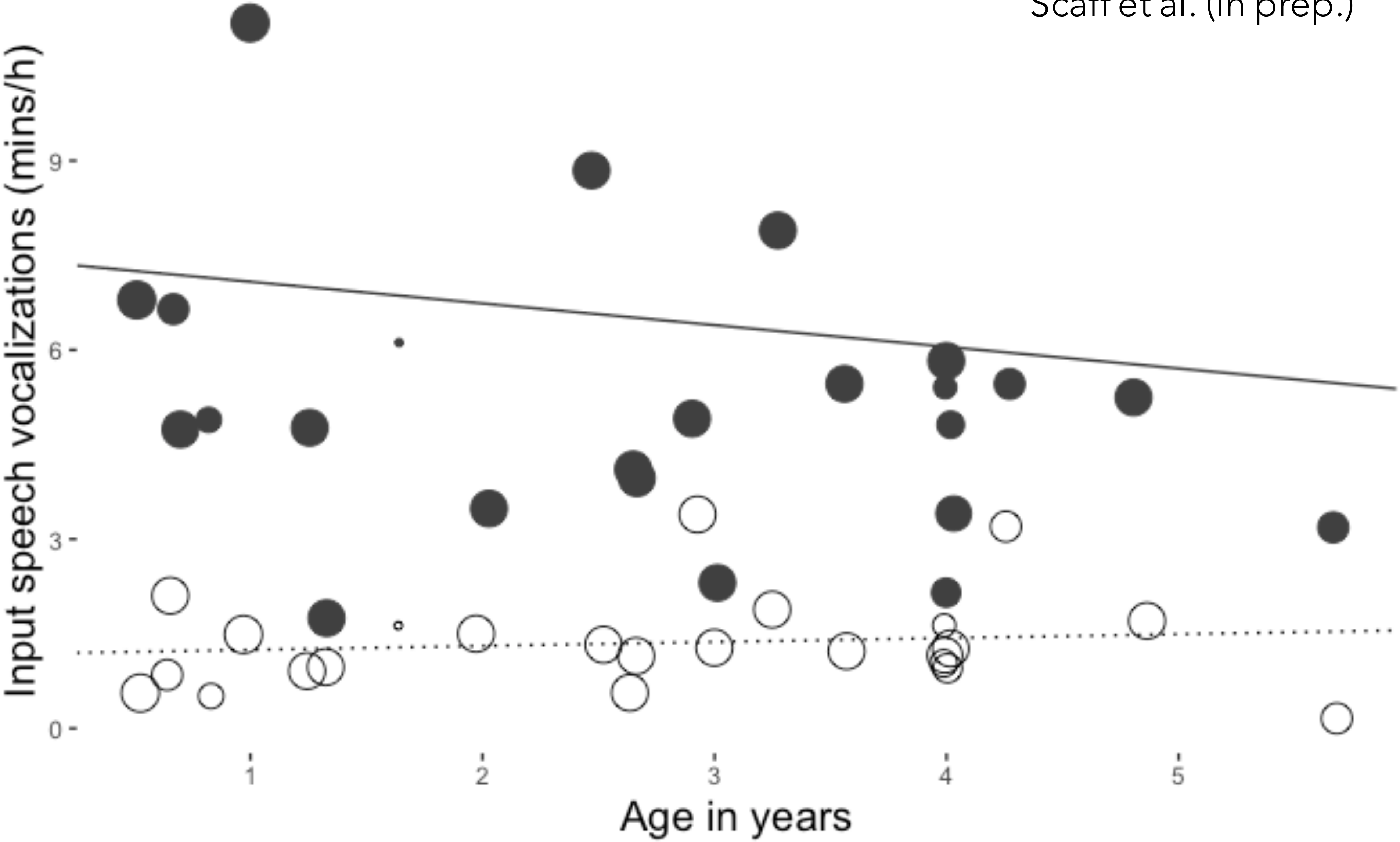


code 1 minute per hour



Input quantities among the Tsimane’

Scaff et al. (in prep.)



How much do you think American babies get talked to?

- .5 minute per hour (less than Tsimane')
- 1 minute per hour (same as Tsimane')
- 5 minutes per hour (more than Tsimane')

<u>Preliminary</u> results X-cultures

Input quantities vary a lot

e.g. Tsimane' children get 1' of child-directed speech per hour,
American kids get 11' per hour

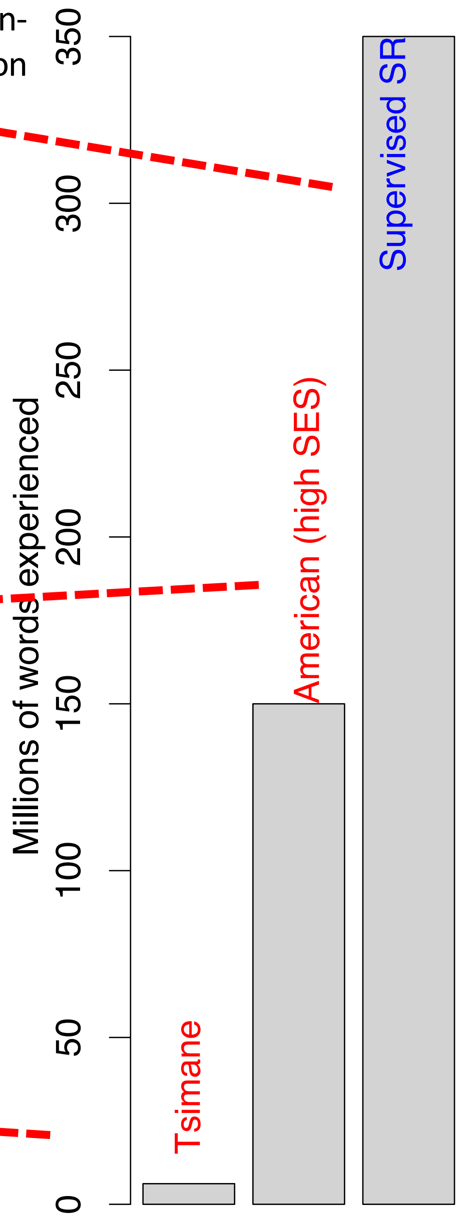
0.2h of speech/day

1.8h of speech/day

Cristia et al (2019) Child Dev
Scaff*… Cristia (in prep)

Baby-machine comparison is even more astounding:

MS's first-pass human-level ASR transcription



humans cumulated to 10 years of age

Children **everywhere** learn to **perceive (& produce) speech** with

<u>much less input</u>

& <u>supervision</u>

than machines do

Supervised SR: Xiong et al. 2016 arXiv
American: Hart & Risley (1995)
Tsimane: Cristia et al. (in press) *Child Dev*

Preliminary results X-cultures

Input quantities vary a lot

e.g. Tsimane' children get 1' of child-directed speech per hour, American kids get 11' per hour

10-fold difference

Input **sources** vary a lot

e.g. Tsimane' children get 50% speech from other children, American kids <10%

if only adult speech "counts", 20-fold difference

Cristia et al (2019) Child Dev
Scaff*… Cristia (in prep)

Maybe measures are inaccurate – they are based on very little data!

Maybe children in those cultures are “delayed” compared to Americans?

Yeah, how about the ‘output’?

Building classifiers to
generalize to unlabeled data

Building classifiers to generalize to unlabeled data

child adult

Talker diarization (who speaks when)

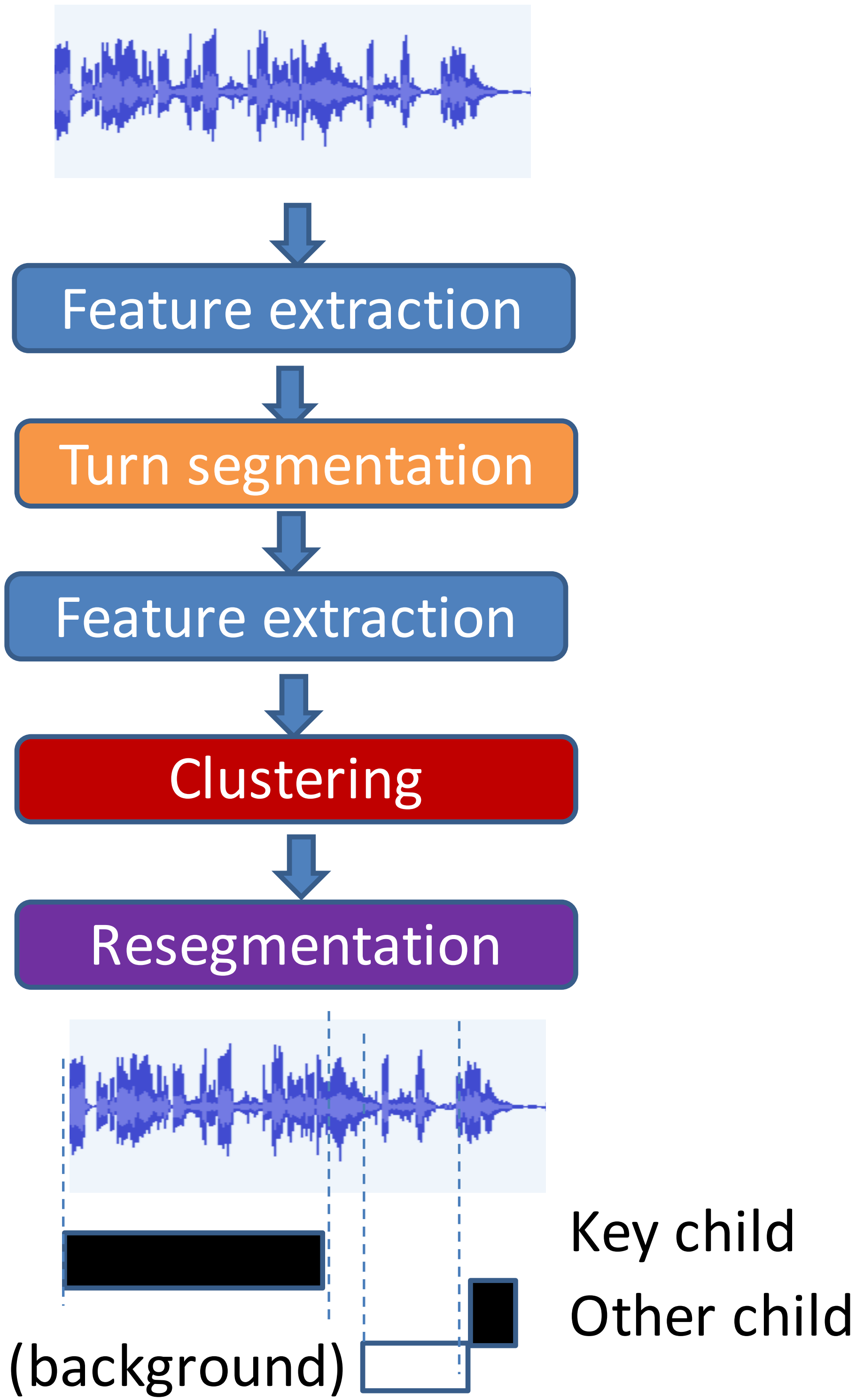
DIHARD 2018, 2019 Interspeech

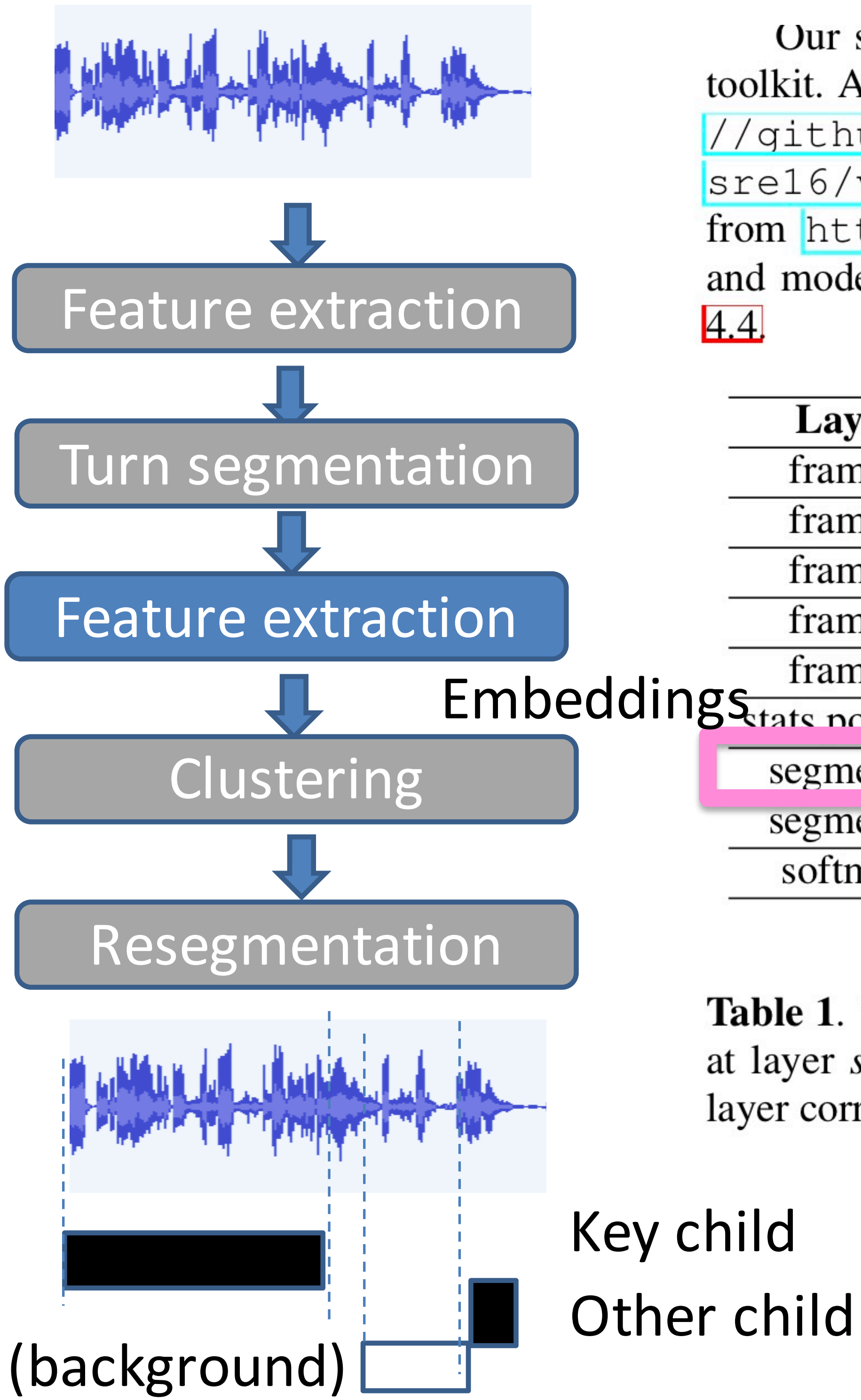
Challenge

We built a dataset

We & others compete to build the best scoring system

Ryant et al. (2018) ICASSP; (2019) Interspeech


Feature extraction
Turn segmentation
Feature extraction
Clustering
Resegmentation
Key child
Other child
(background)

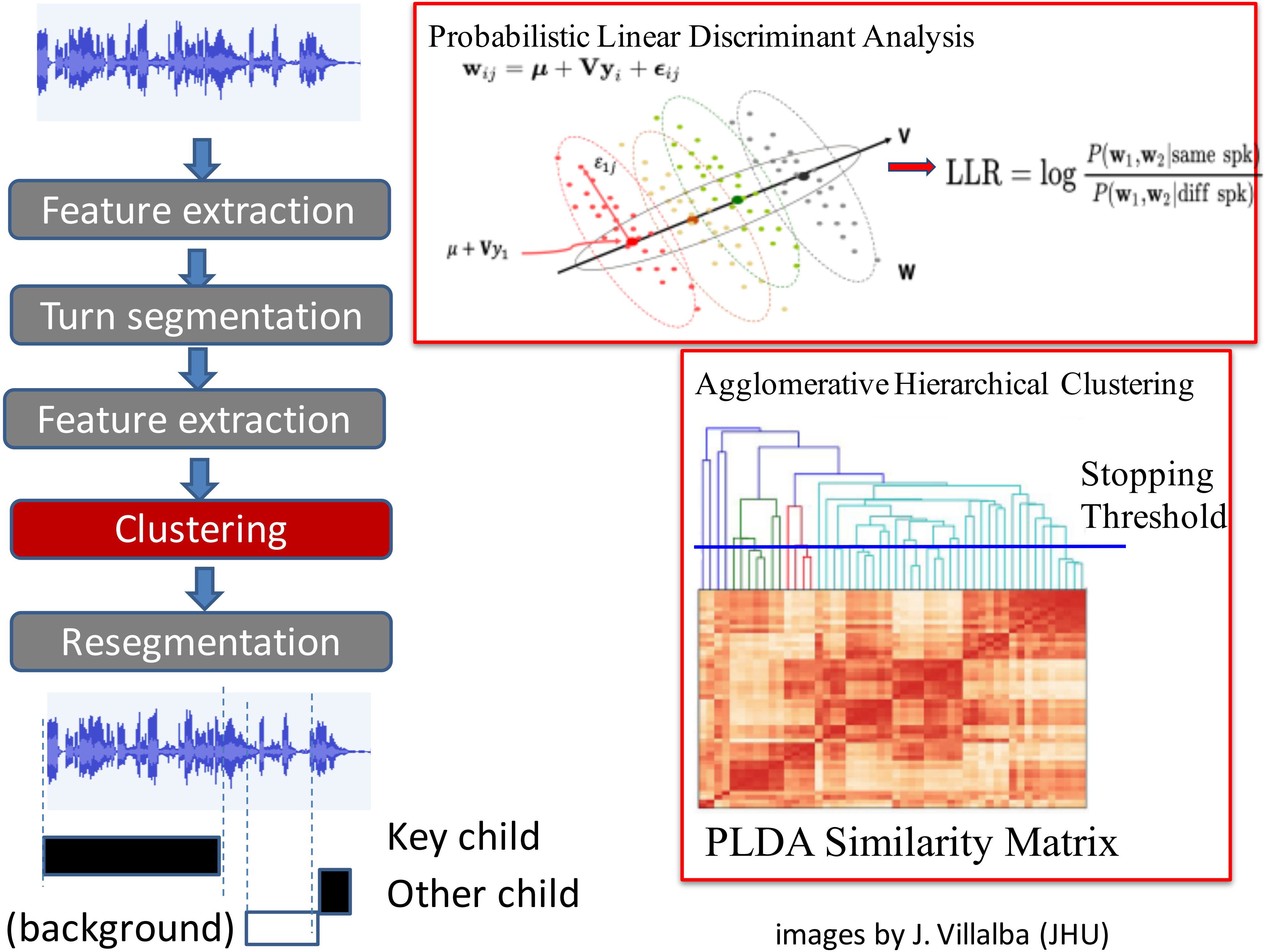


Our software framework has been made available in the Kaldi toolkit. An example recipe is in the main branch of Kaldi at `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2` and a pretrained x-vector system can be downloaded from `http://kaldi-asr.org/models.html`. The recipe and model are similar to the x-vector system described in Section 4.4.

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

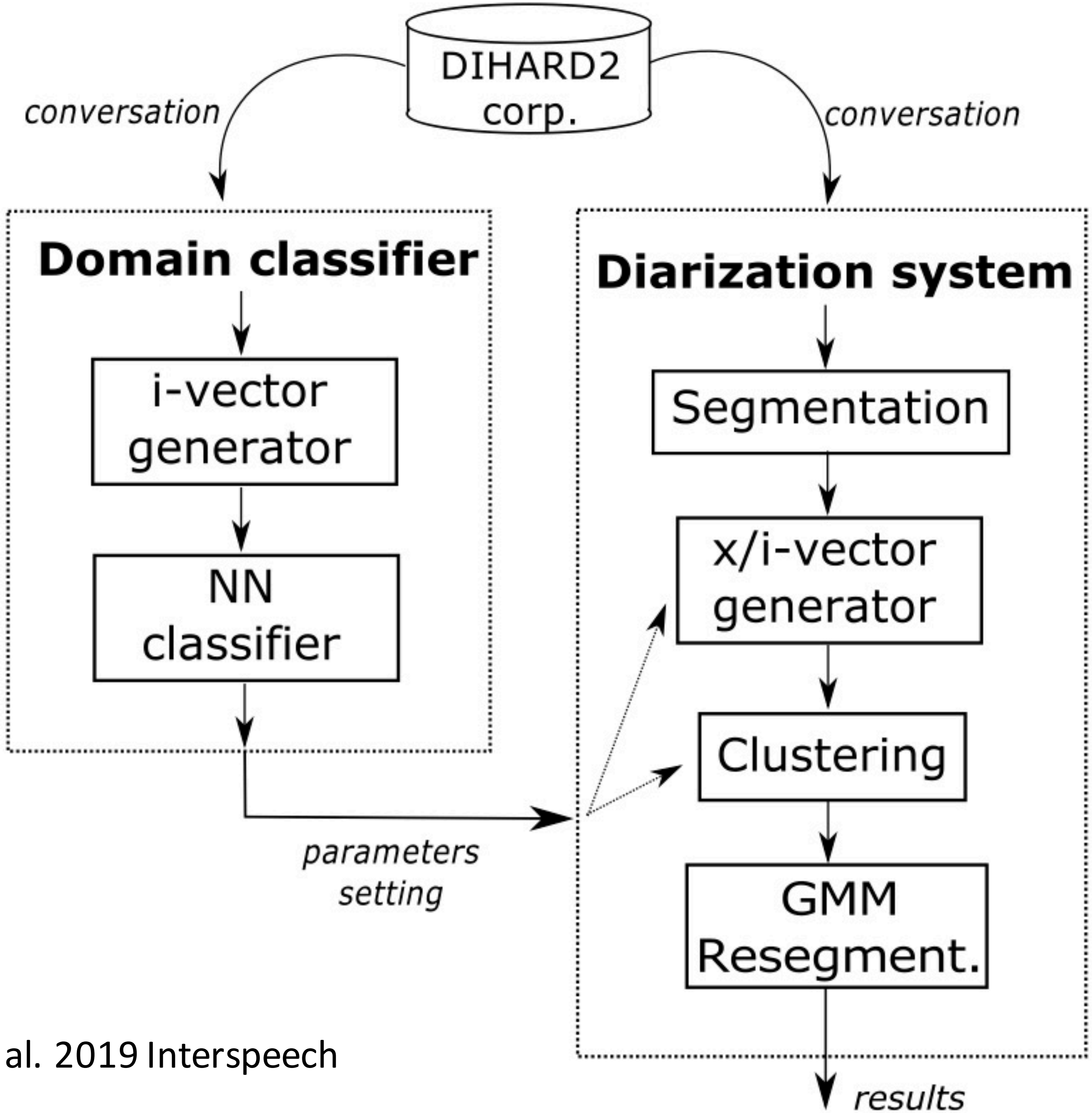
Table 1. The embedding DNN architecture. x-vectors are extracted at layer *segment6*, before the nonlinearity. The N in the softmax layer corresponds to the number of training speakers.

Snyder et al. 2018 ICASSP



Feature extraction
Turn segmentation
Feature extraction
Clustering
Resegmentation
Key child
Other child
(background)
Probabilistic Linear Discriminant Analysis
$\mathbf{w}_{ij} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i + \boldsymbol{\epsilon}_{ij}$
ϵ_{1j}
$\mu + \mathbf{V}\mathbf{y}_1$
V
W
$\mathrm{LLR} = \log \frac{P(\mathbf{w}_1, \mathbf{w}_2 | \text{same spk})}{P(\mathbf{w}_1, \mathbf{w}_2 | \text{diff spk})}$
Agglomerative Hierarchical Clustering
Stopping Threshold
PLDA Similarity Matrix
images by J. Villalba (JHU)

Heterogeneity
+
Data
sparseness



Zajic et al. 2019 Interspeech

Building classifiers to generalize to unlabeled data

Talker diarization

(who speaks when)

DIHARD 2018, 2019 Interspeech

Building classifiers to
generalize to unlabeled data

Talker diarization
(who speaks when)
DIHARD 2018, 2019 Interspeech

Addressee classification
(whom are they talking to)
ComParE 2017 Interspeech

2 classes,
no team beat the
baseline

Building classifiers to generalize to unlabeled data

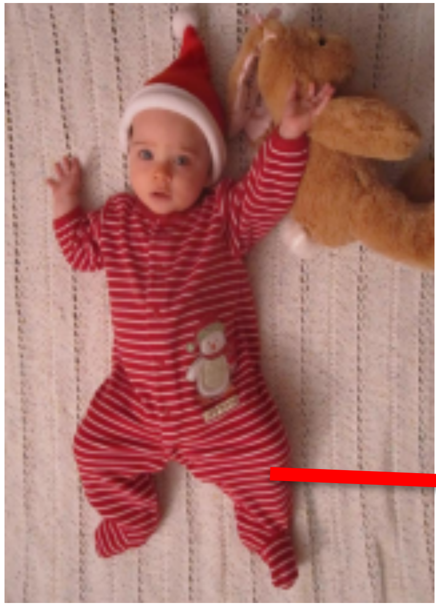
Talker diarization
(who speaks when)
DIHARD 2018, 2019 Interspeech

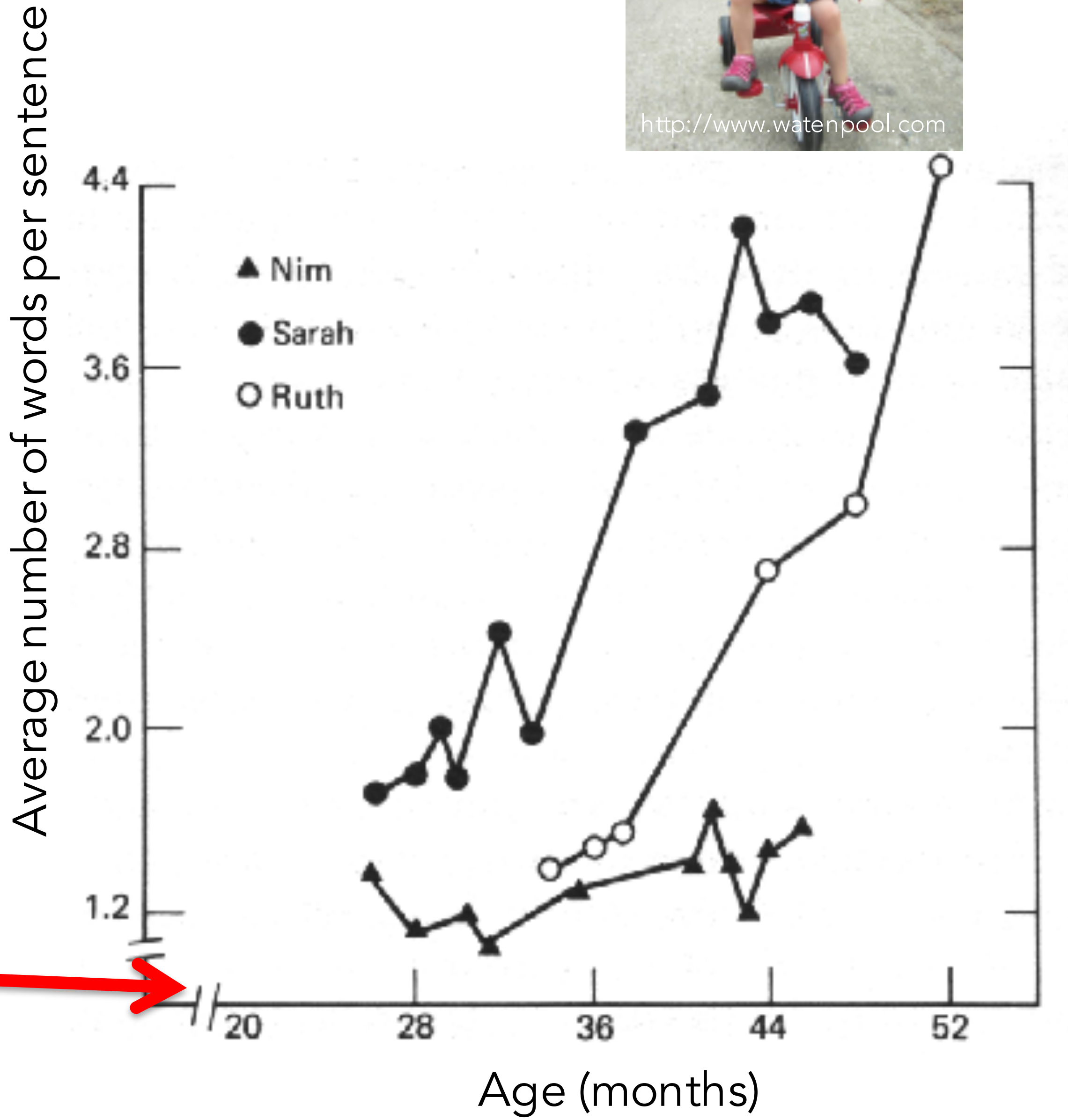
Addressee classification
(whom are they talking to)
ComParE 2017 Interspeech

Child vocalization types
(babbling, crying, …)
ComParE 2019 Interspeech

5 classes

plenty
happens
before 1 year!





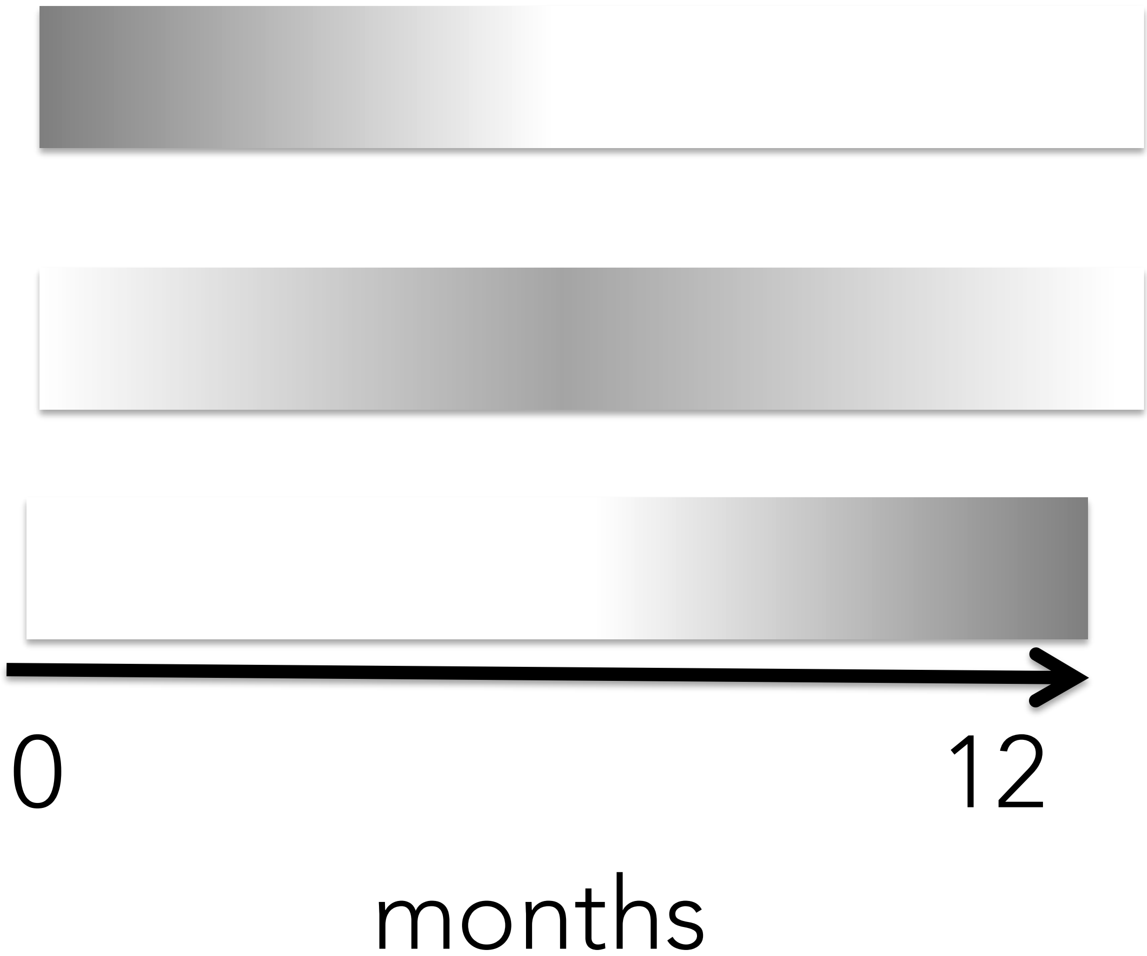
Terrace 1979 Science

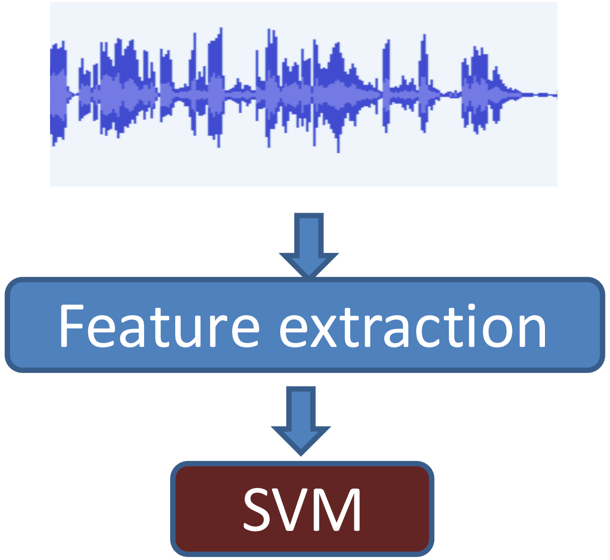
Vocalizations vary in complexity

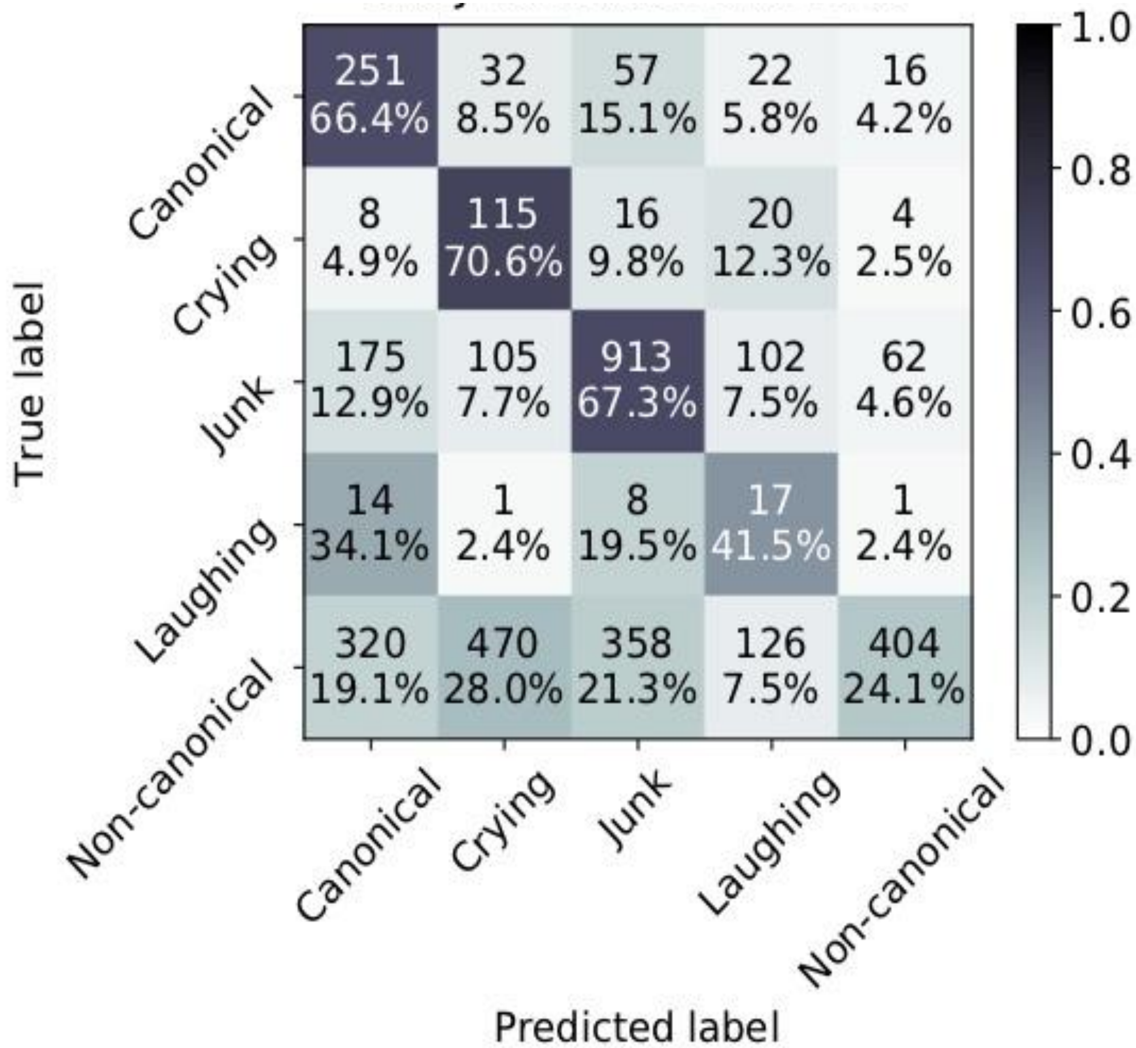
reflexive vocalizations

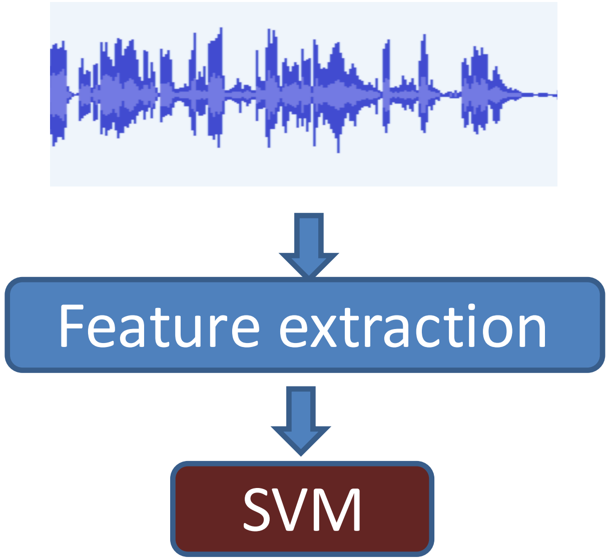
non-canonical babbling
(55″)

canonical babbling
(24″)

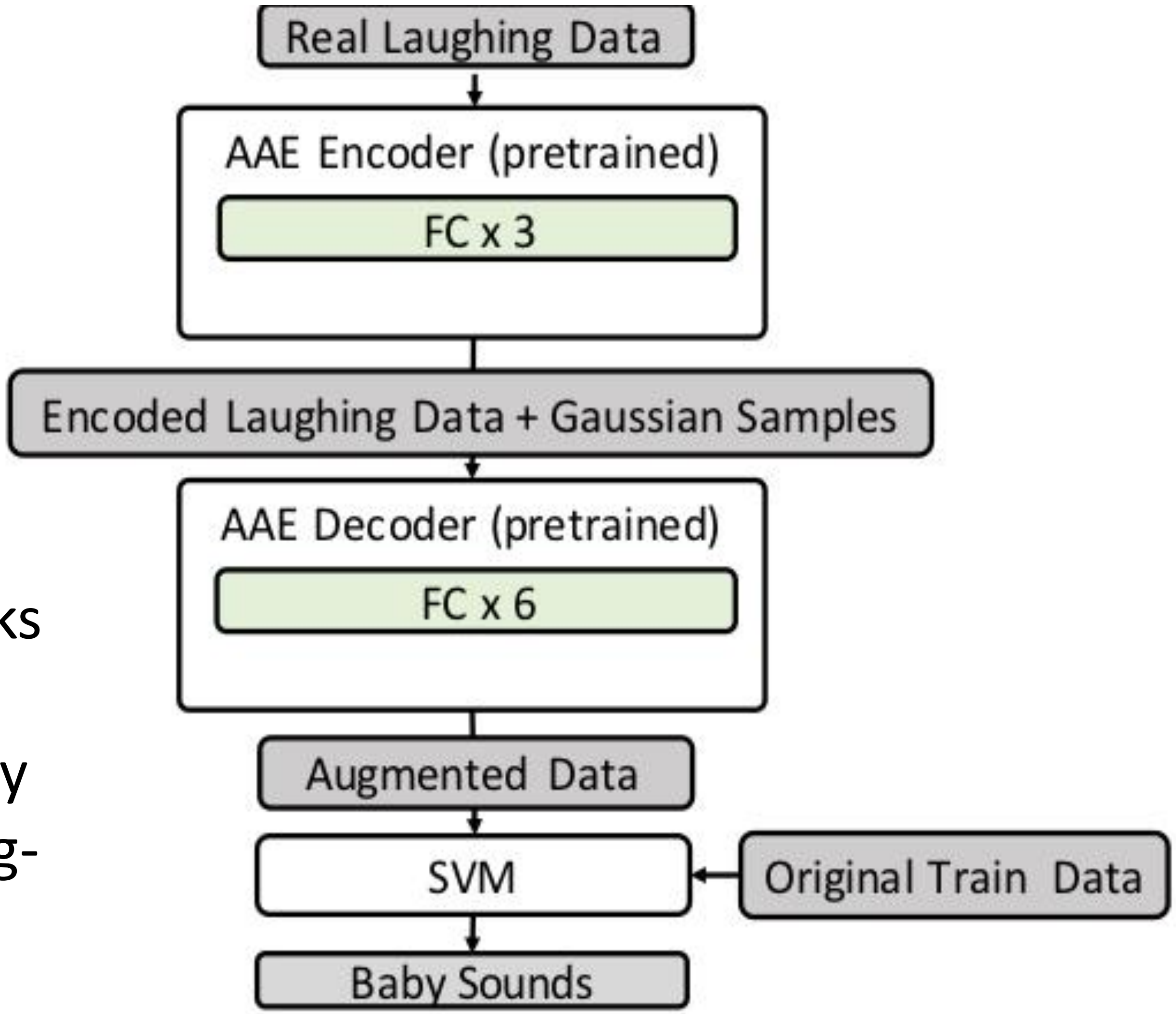



Feature extraction
SVM


True label
Canonical
Crying
Junk
Laughing
Non-canonical
251 66.4%
32 8.5%
57 15.1%
22 5.8%
16 4.2%
8 4.9%
115 70.6%
16 9.8%
20 12.3%
4 2.5%
175 12.9%
105 7.7%
913 67.3%
102 7.5%
62 4.6%
14 34.1%
1 2.4%
8 19.5%
17 41.5%
1 2.4%
320 19.1%
470 28.0%
358 21.3%
126 7.5%
404 24.1%
Canonical
Crying
Junk
Laughing
Non-canonical
Predicted label
1.0
0.8
0.6
0.4
0.2
0.0



And the winner is…



"Using Attention Networks and Adversarial Augmentation for … Baby Sound Recognition", Sung-Lin Yeh … Chi-Chun Lee

Building classifiers to generalize to unlabeled data

Talker diarization
(who speaks when)
DIHARD 2018, 2019 Interspeech

Addressee classification
(whom are they talking to)
ComParE 2017 Interspeech

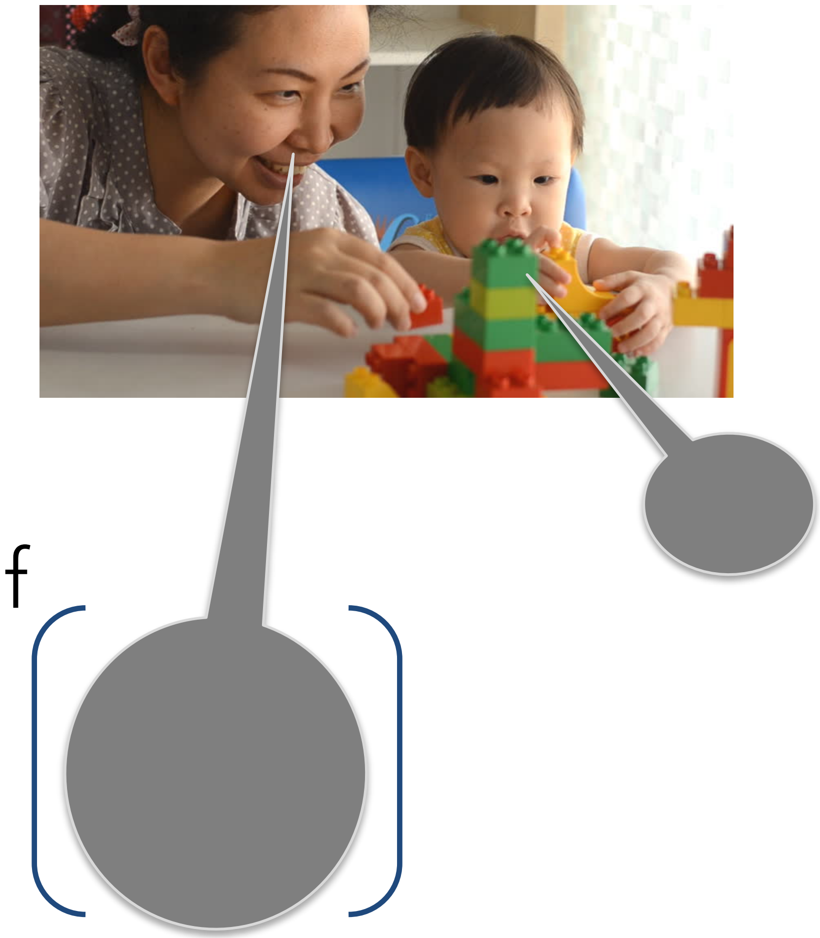
Child vocalization types
(babbling, crying, …)
ComParE 2019 Interspeech

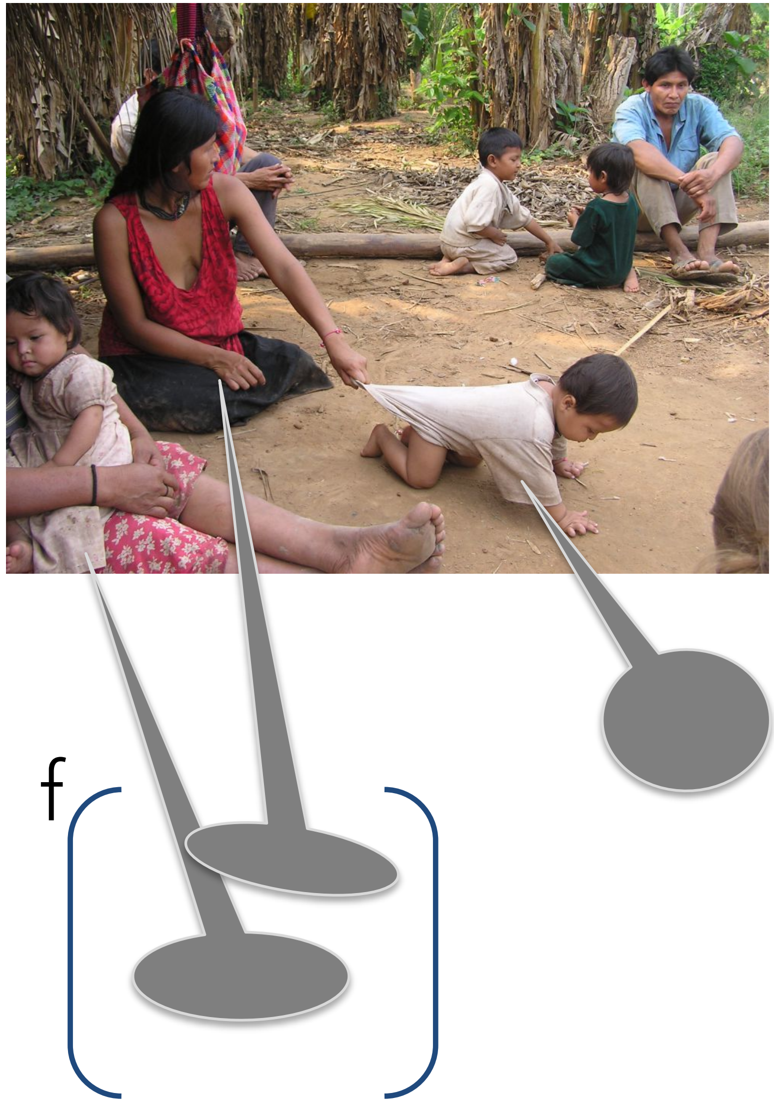
Shamelessly stolen from Y. LeCun

TO BE CONTINUED

NEEDED:
more work on unsupervised, semi-supervised, and self-supervised classification

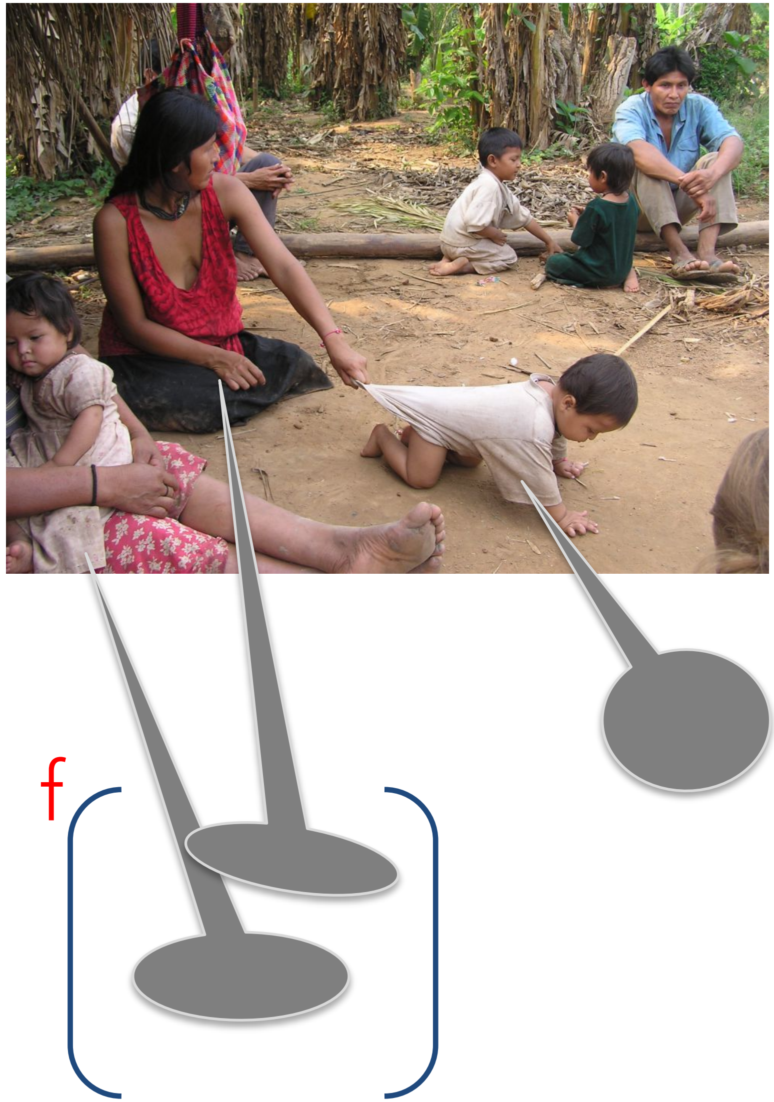
Assuming results hold, our broad language acquisition theory (v 1.1)





Assuming results hold, our broad language acquisition theory (v 1.1)

May infants learn from peers (children's speech)? from overheard speech?



Assuming results hold, our broad language acquisition theory (v 1.1)

Next step:
Learnability properties

f

May infants learn from peers (children's speech)? from overheard speech?

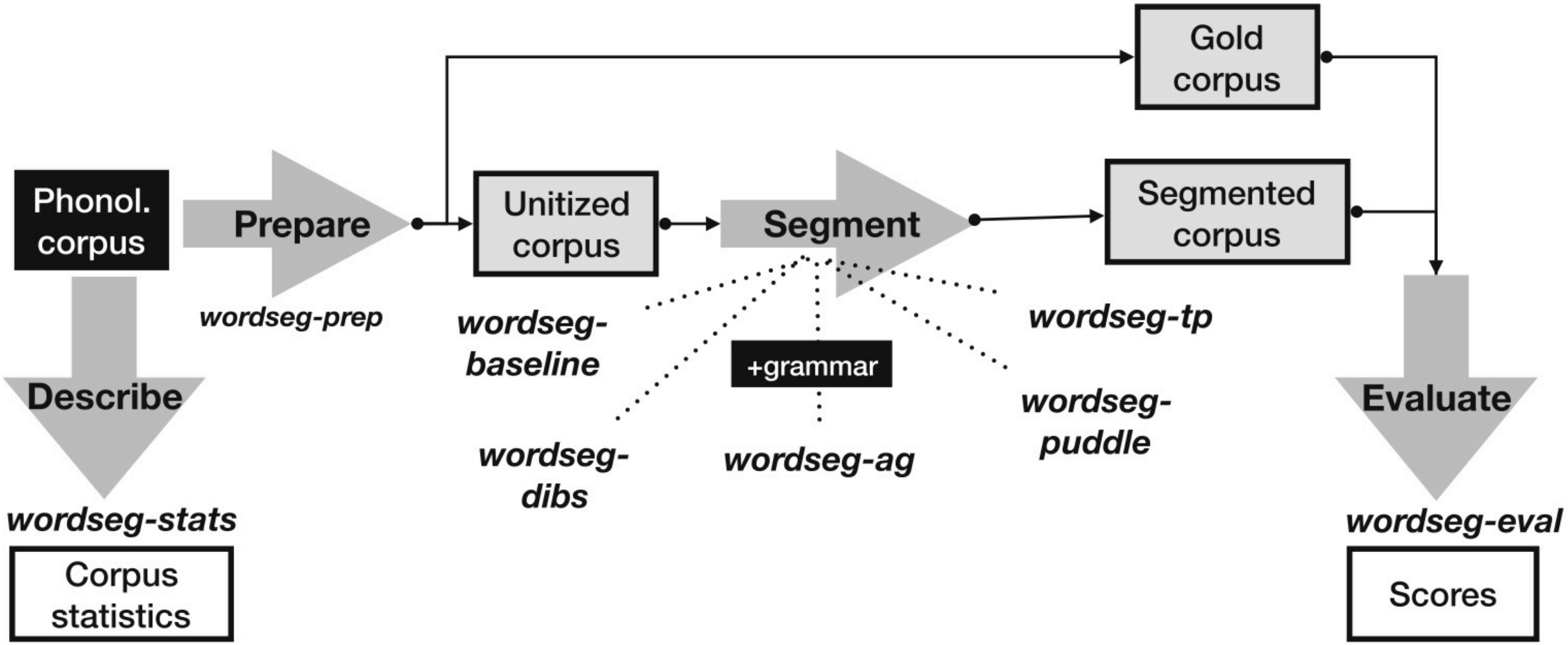
f

Studying learnability properties: Unsupervised word segmentation

f

WordSeg Package

wordseg.readthedocs.io



Example algorithms

1. Baseline

Simplest strategies

- Every sentence is a word (**SentBase**)
- Every syllable is a word (**SyllBase**)

Lignos 2012

2. Sub-lexical

Goal is to "cut" using local cues

- Transitional Probabilities (TP) x Absolute/Relative threshold
  - **TP_abs**
  - **TP_rel**
- Diphone-Based Segmentation **(DiBS)**

Daland + 2009; Saksida + 2016

3. Lexical

Goal is to learn a set of "minimal recombinable units"

- Adaptor Grammar **(AG)**
- Phonotactics from Utterances Determine Distributional Lexical Elements **(Puddle)**

Johnson + 2007; Monaghan + 2010

Package: wordseg.readthedocs.io Preprint: https://osf.io/nx49h/

Bernard et al. 2019 Beh Res Meth

Studying learnability properties: Unsupervised word segmentation

f

WordSeg Package

+

hibaby
areyouacutebaby?

Transcribed speech corpora

English may not be the best language to study learnability on…

English (and other contact/imperial languages)

Finish it, I'll be here!

He's dressed.

English may not be the best language to study learnability on…

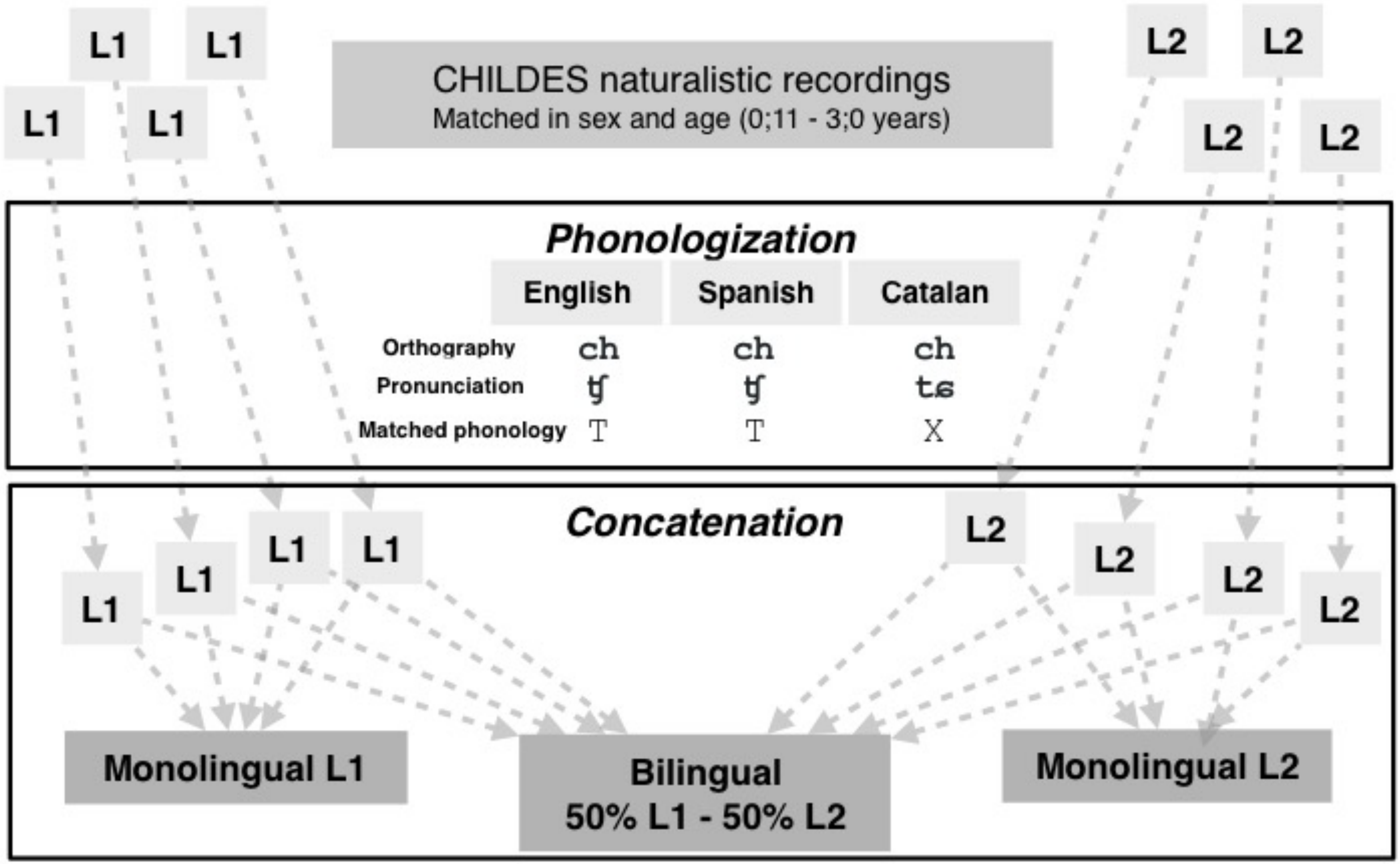
English (and other contact/imperial languages)

Inuktitut

Finish it, I'll be here! = Nungullugungai, taavanilangajualusunga!

He's dressed. = Annuraqsimajualuuman.

Creating bilingual corpora

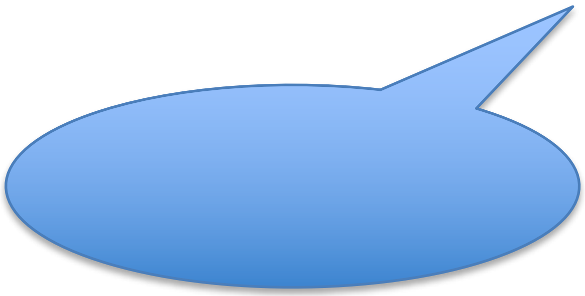


Factors we manipulated

Different processing algorithms

f

Different languages



Monolingual versus bilingual input

Which factor had the biggest impact on performance?

Different processing algorithms

$f(\qquad)$

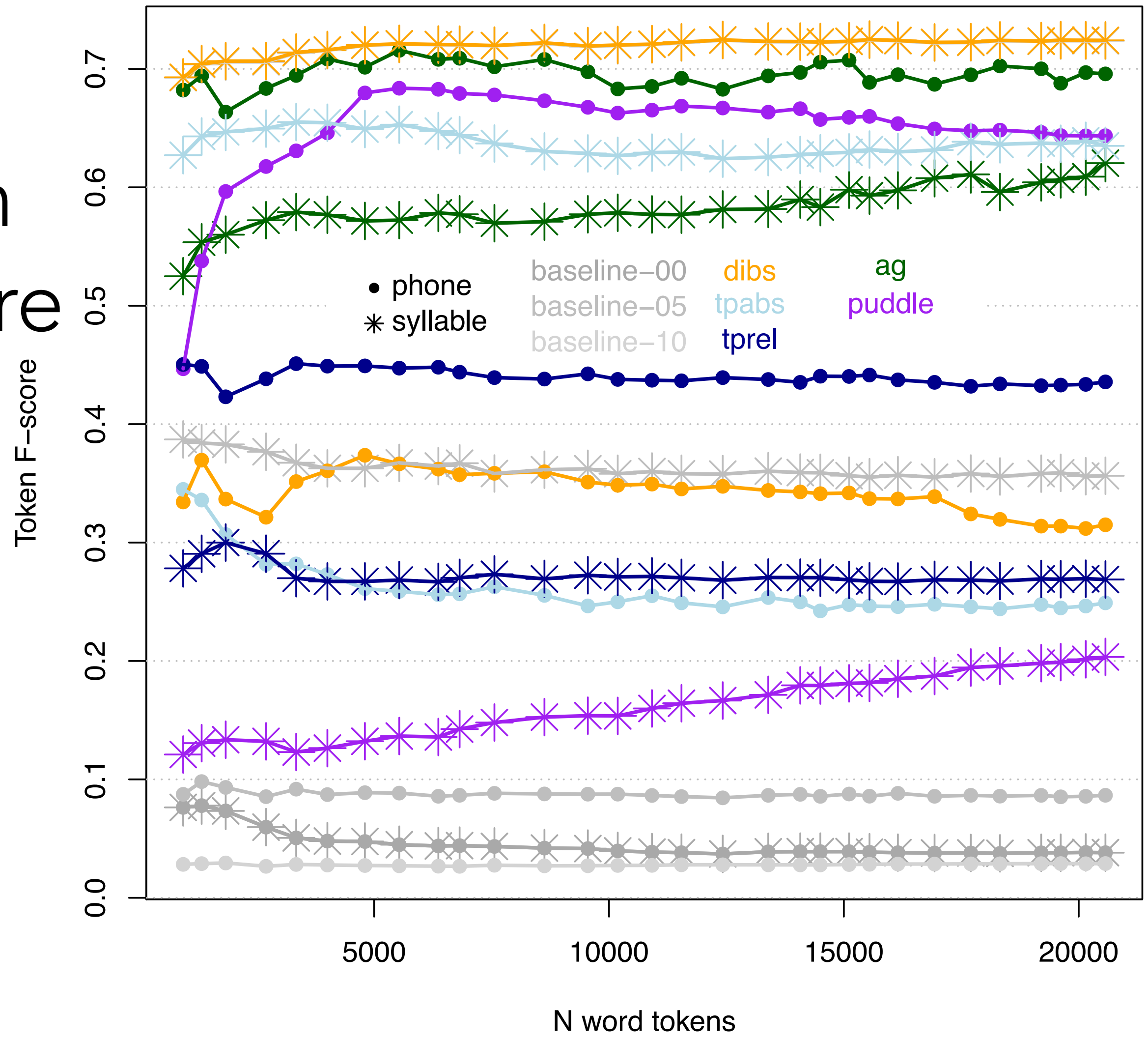
Different languages

Monolingual versus bilingual input

Results so far

f

Differences between learning algorithms are enormous (40-60%)



Mathieu … Cristia (2019) Beh Res Methods

Results so far

f

Differences between learning algorithms are enormous (40-60%)

> than that between languages as a function of morphological type (20%)

- Monolingual versus bilingual input (<5%)

Mathieu … Cristia (2019) Beh Res Methods

Loukatou … Cristia (2019) ACL
Fibla … Cristia (subm)

Results so far

f

Differences between learning algorithms are enormous (40-60%)

> than that between languages as a function of morphological type (20%)

- Monolingual versus bilingual input (<5%)

TO BE CONTINUED

Mathieu … Cristia (2019) Beh Res Methods

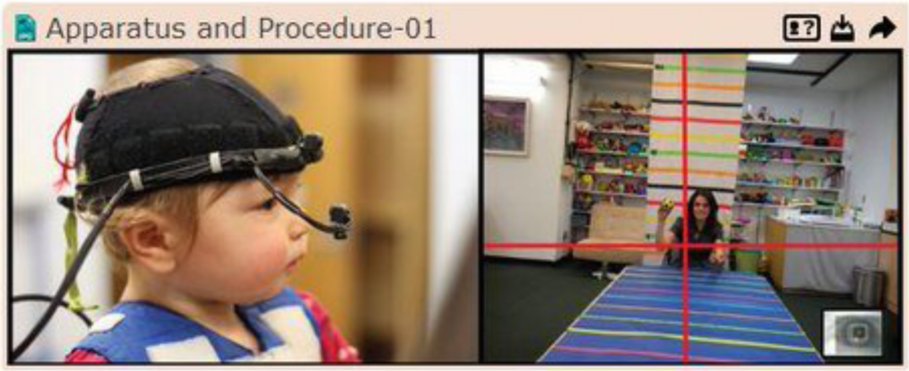
NEEDED:

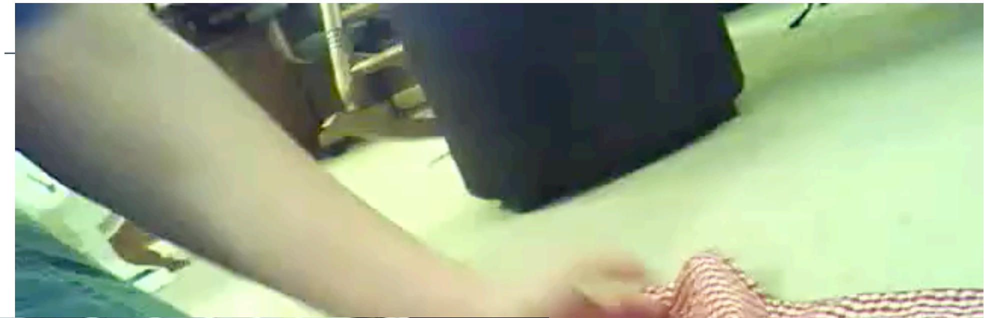
- learnability on other levels;
- *real infant evidence*

Loukatou … Cristia (2019) ACL

Fibla … Cristia (subm)

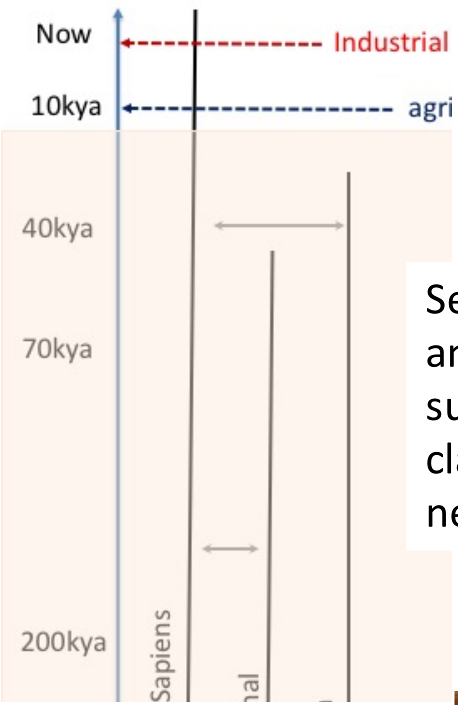
Databrary

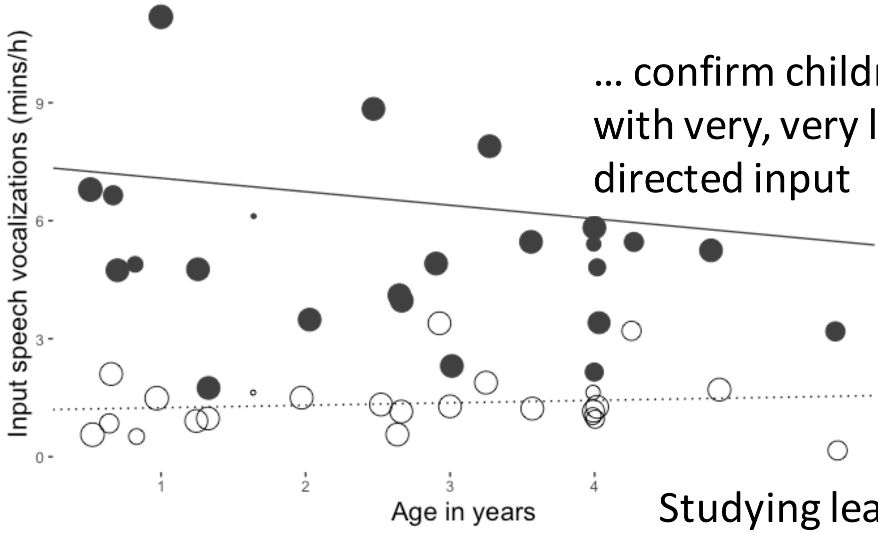




1-month-old looking
over caregiver's shoulder

All extant datasets are biased

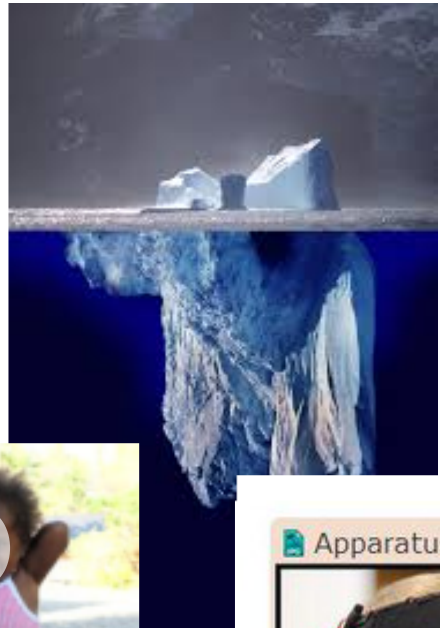


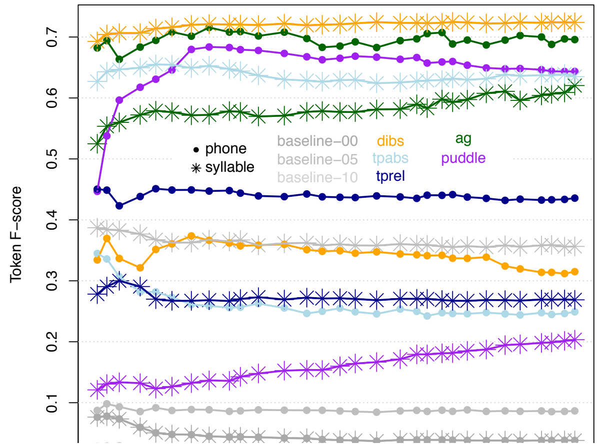


... confirm children succeed with very, very little directed input

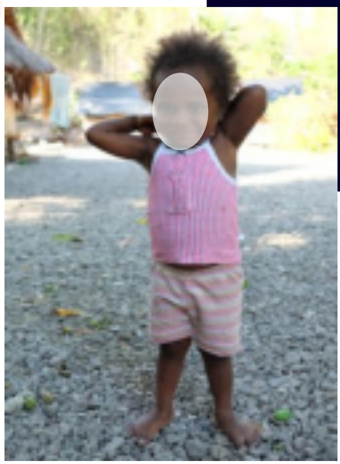
Studying learnability properties using artificial agents

Semi-, un-, and self-supervised classifiers needed!

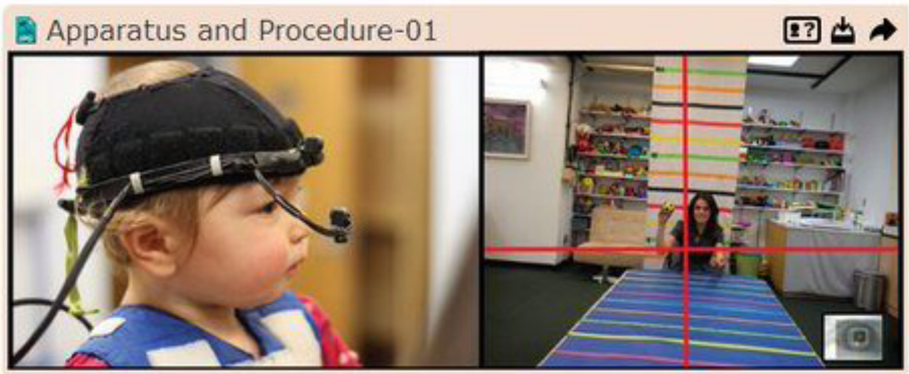




Humans evolved in a setting crucially different from that represented in those data



Naturalistic, massive datasets of child language...



Post-doctoral fellows

Naomi Havron
Effect of siblings

Christof Neumann
Parental investment

Logistics

Catherine Urban
Admin Magician

Xuan Nga Cao
Manager

Affiliated postdoc

Camila Scaff
Fieldwork

PhD student

Georgia Loukatou
Cognitive modeling

Engineer/PhD student

Marvin Lavechin
All-ologist

Interns

Ruben Bousbib (CentraleSupélec)
Chiara Semenzin (Erasmus, U Edinb)
Elisa Lannelongue (M2 Cogmaster)
Lara Oliel (M2 UPMC)
Leo Pivot (M1 Cogmaster)