Language Learning: A Data-Driven Approach

Michael C. Frank
Stanford University

Intros

- Michael Frank (call me Mike)
  - Faculty at Stanford
  - Study language learning, reproducibility and replicability, pragmatics, development, metascience
  - Fun fact: play the mandolin in a bluegrass band
- You:
  - Institution, program of study
  - Research interest
  - Fun fact – or boring fact if you feel stressed

Learning goals

- Discuss the "standard model" framework for early word learning, focusing on input, processing, and uptake constructs,
- Compare different instruments and approaches for measuring child language,
- Learn a reproducible workflow for exploring language acquisition data in R, and
- Explore data from Wordbank, CHILDES, and Peekbank as a source of insights into language learning.

Course outline

- Today: Introduction to course framework and toolset
- Tuesday: Digging deeper into vocabulary data in **Wordbank**
- Wednesday: exploring transcripts of children's speech through **CHILDES** and **childes-db**
- Thursday: analyzing eye-tracking data using **Peekbank**
- Friday: Student-led mini-projects

Course prerequisites

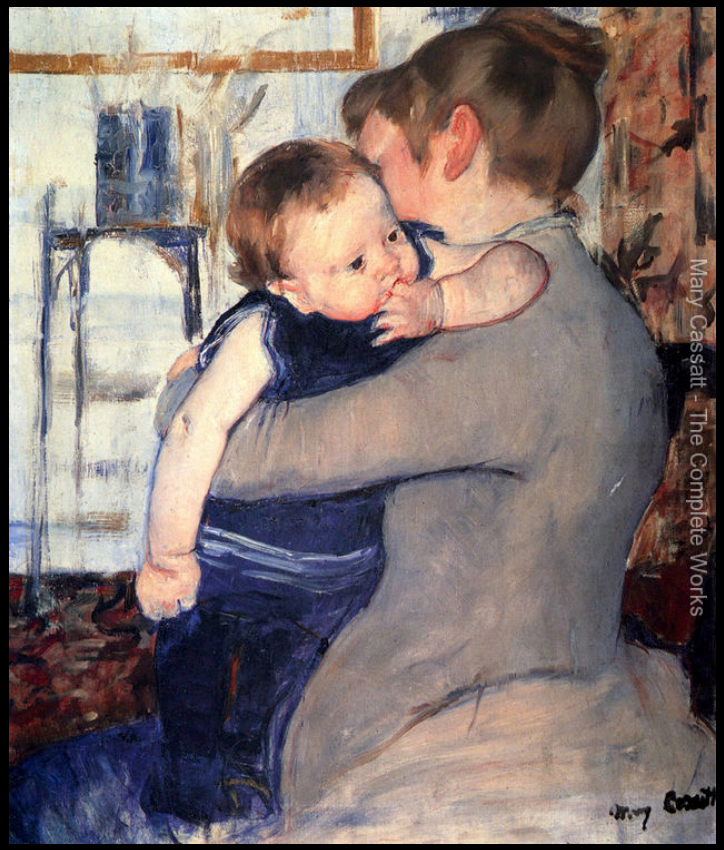
- Minimal **github** to synchronize most recent course files
- **R** and **R studio** to manipulate data
- Tidyverse including **dplyr** and **ggplot** to manipulate and visualize data
- All of these will be (briefly) introduced
- But if you do not have some of these, the second part of most classes will be frustrating

Readings and other homework

- There will be two articles assigned for each class
  - Day 1
    - Frank et al. (2021) Chapter 1, "Theoretical Foundations"
    - Kachergis, Marchman, & Frank (2022), "Towards a standard model"
  - Day 2
    - Bates & Goodman (1997), ”On the inseparability of grammar and the lexicon"
    - Frank et al. (2021), Chapter 13, "Morphology, Grammar, and the Lexicon”
  - Day 3
    - MacWhinney & Snow (1990), "CHILDES: an update"
    - Sanchez*, Meylan* et al. (2019), "childes-db: a a flexible and reproducible interface”
  - Day 4
    - Fernald et al. (1998), "Rapid gains in speed of verbal processing"
    - Zettersten et al. (2022), "Peekbank: an open, large-scale repository”
- Beyond these, please work to complete the Rmds from class
- On Day 4, you should send me an email with your group and your proposed project (1 paragraph)

Outline

1. **Basics of early word learning**
2. Toward a “standard model” of language learning
3. Developing resources (to measure quantities in the model)
   - Input – childes-db
   - Learning – wordbank
   - Processing – peekbank
4. The course toolset



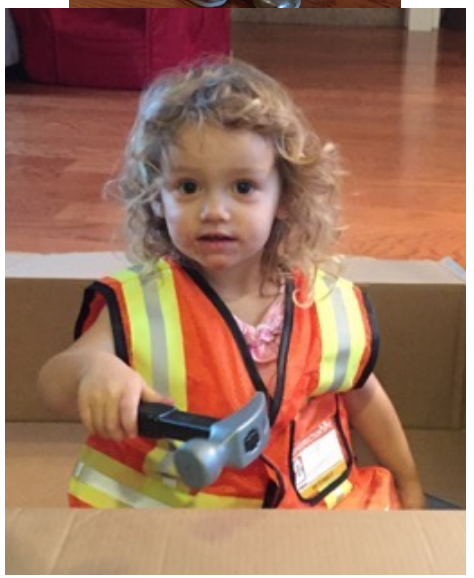
An explosion of language

18 mo: “happy-b”

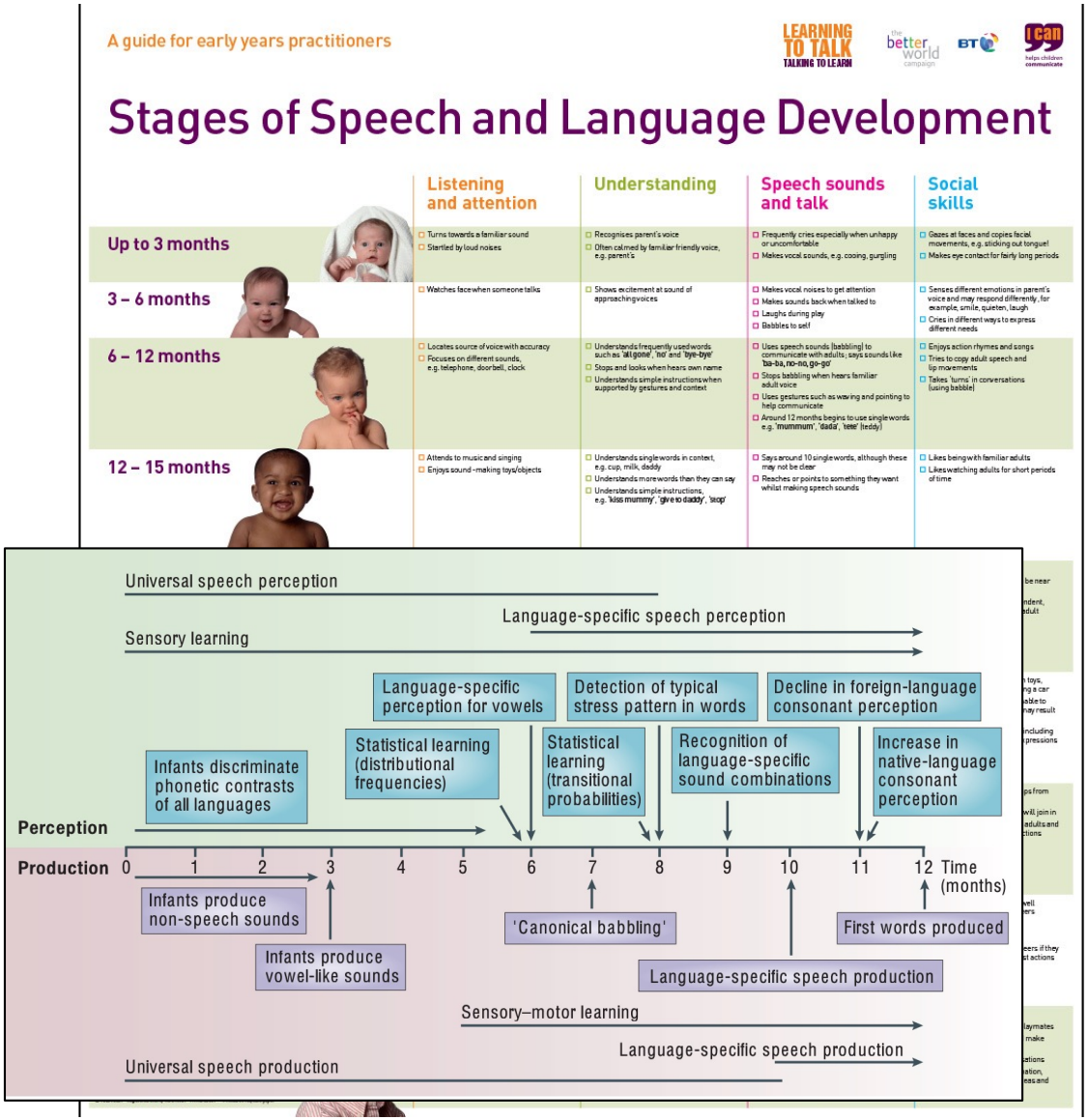
19 mo: “blue ball”

23 mo: “spike doggy no food eat dirt”

26 mo: “dada move own body, my need lilbit more space”



Ages and stages theories are not sufficient



A guide for early years practitioners

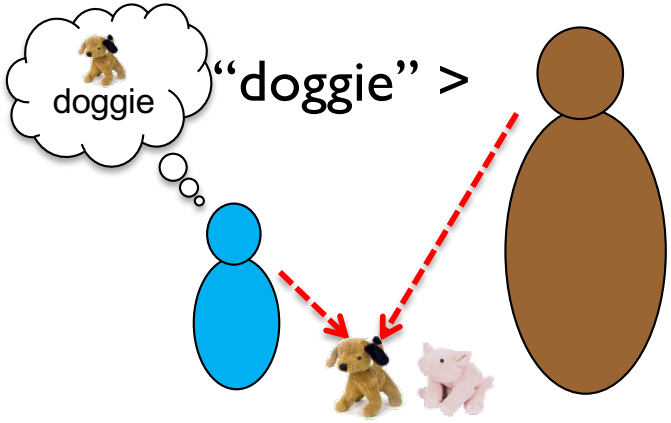
LEARNING TO TALK
TALKING TO LEARN

Stages of Speech and Language Development

	Listening and attention	Understanding	Speech sounds and talk	Social skills
Up to 3 months	Turns towards a familiar sound Startled by loud noises	Recognises parent's voice Often calmed by familiar friendly voice, e.g. parent's	Frequently cries especially when unhappy or uncomfortable Makes vocal sounds, e.g. cooing, gurgling	Gazes at faces and copies facial movements, e.g. sticking out tongue! Makes eye contact for fairly long periods
3 – 6 months	Watches face when someone talks	Shows excitement at sound of approaching voices	Makes vocal noises to get attention Makes sounds back when talked to Laughs during play Babbles to self	Senses different emotions in parent's voice and may respond differently, for example, smile, quieten, laugh Cries in different ways to express different needs
6 – 12 months	Locates source of voice with accuracy Focuses on different sounds, e.g. telephone, doorbell, clock	Understands frequently used words such as 'all gone', 'no' and 'bye-bye' Stops and looks when hears own name Understands simple instructions when supported by gestures and context	Uses speech sounds (babbling) to communicate with adults; says sounds like 'ba-ba, no-no, go-go' Stops babbling when hears familiar adult voice Uses gestures such as waving and pointing to help communicate Around 12 months begins to use single words e.g. 'mummum', 'dada', 'tete' (teddy)	Enjoys action rhymes and songs Tries to copy adult speech and lip movements Takes 'turns' in conversations (using babble)
12 – 15 months	Attends to music and singing Enjoys sound-making toys/objects	Understands single words in context, e.g. cup, milk, daddy Understands more words than they can say Understands simple instructions, e.g. 'kiss mummy', 'give to daddy', 'stop'	Says around 10 single words, although these may not be clear Reaches or points to something they want whilst making speech sounds	Likes being with familiar adults Likes watching adults for short periods of time

Early vocabulary as a case study

- Easy to observe and measure
- Units (words) are less controversial than representations underlying morphosyntax
- Nonetheless tightly linked to morphosyntax (day 2)
- Grounded in concept learning, social cognition, etc.



Language input

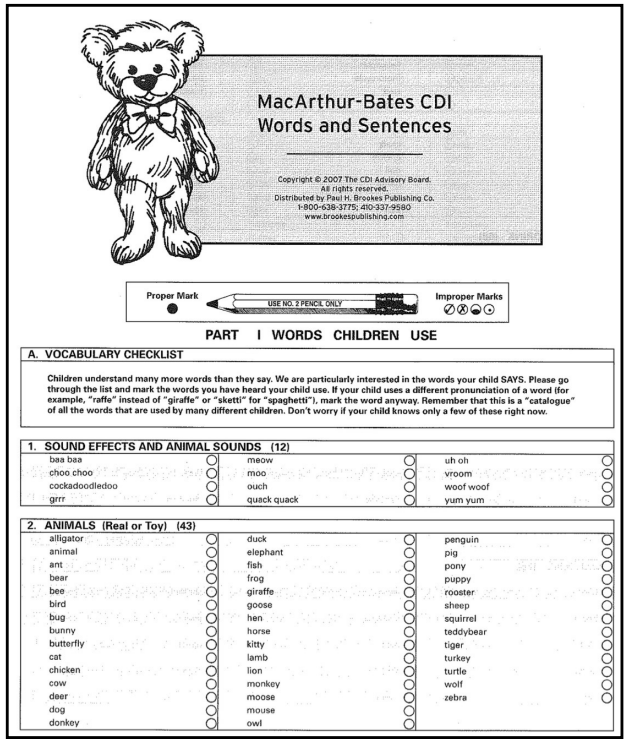
	Yearly total				Yearly CDS			
	Hours		Words (M)		Hours		Words (M)	
				Urban, high SES				
H&R (N=13)[t]	1221[w,c]	[578,1987]	11.0[c]	[5.20, 17.9]	786[w]	[372, 1279]	7.07	[3.35, 11.5]
S&G (N=6)[t]	2023[w,m]	[1243, 2858]	18.2[m]	[11.2, 25.7]	1223[w,m]	[853, 1574]	11.0[m]	[7.7, 14.2]
VdW (N=1)	931		9.28		140		1.39	
				Urban, low SES				
H&R (N=6)[t]	363[w,d]	[136, 558]	3.26[d]	[1.22, 5.02]	225[w]	[84, 346]	2.02	[0.76., 3.11]
W&F (N=29)[t]	363[w]	[52, 1049]	3.27	[0.46., 9.44]	225[w]	[32, 650]	2.03	[0.29, 5.85]
				Rural, low SES				
S&G (N=6)[t]	503[w,m]	[365, 640]	4.53[m]	[3.28, 5.76]	234[w,m]	[132, 322]	2.10[m]	[1.19, 2.90]

Surprisingly tricky to get accurate counts of words per year (need transcripts…)

All speech: 250k - 1.5M words/month
Child-directed: 150k – 1M words/month

Dupoux (2018)

Children's vocabulary growth



MacArthur-Bates CDI
Words and Sentences

Proper Mark | USE NO. 2 PENCIL ONLY | Improper Marks

PART I WORDS CHILDREN USE

A. VOCABULARY CHECKLIST

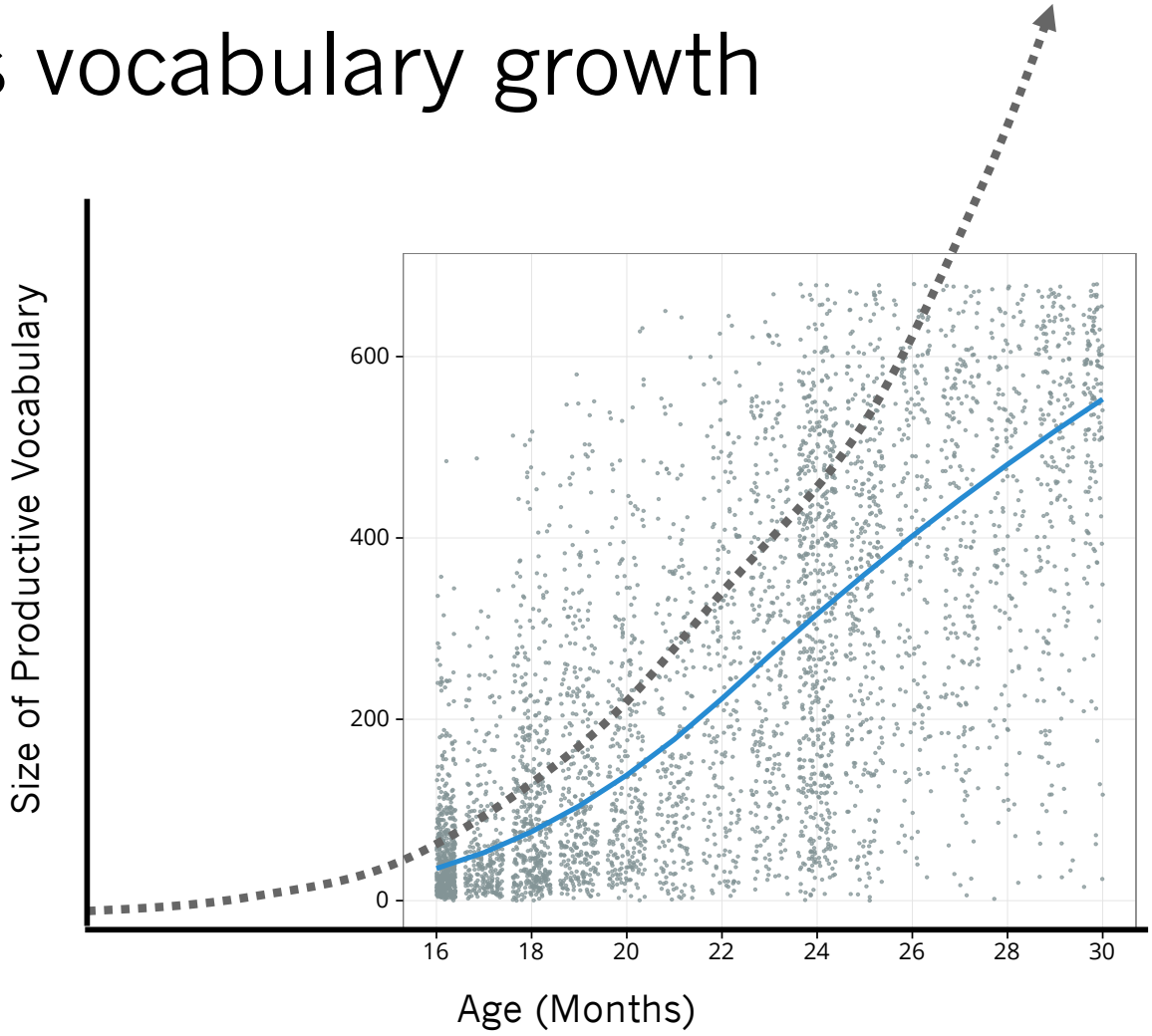
Children understand many more words than they say. We are particularly interested in the words your child SAYS. Please go through the list and mark the words you have heard your child use. If your child uses a different pronunciation of a word (for example, "raffe" instead of "giraffe" or "sketti" for "spaghetti"), mark the word anyway. Remember that this is a "catalogue" of all the words that are used by many different children. Don't worry if your child knows only a few of these right now.

1. SOUND EFFECTS AND ANIMAL SOUNDS (12)

baa baa	meow	uh oh
choo choo	moo	vroom
cockadoodledoo	ouch	woof woof
grrr	quack quack	yum yum

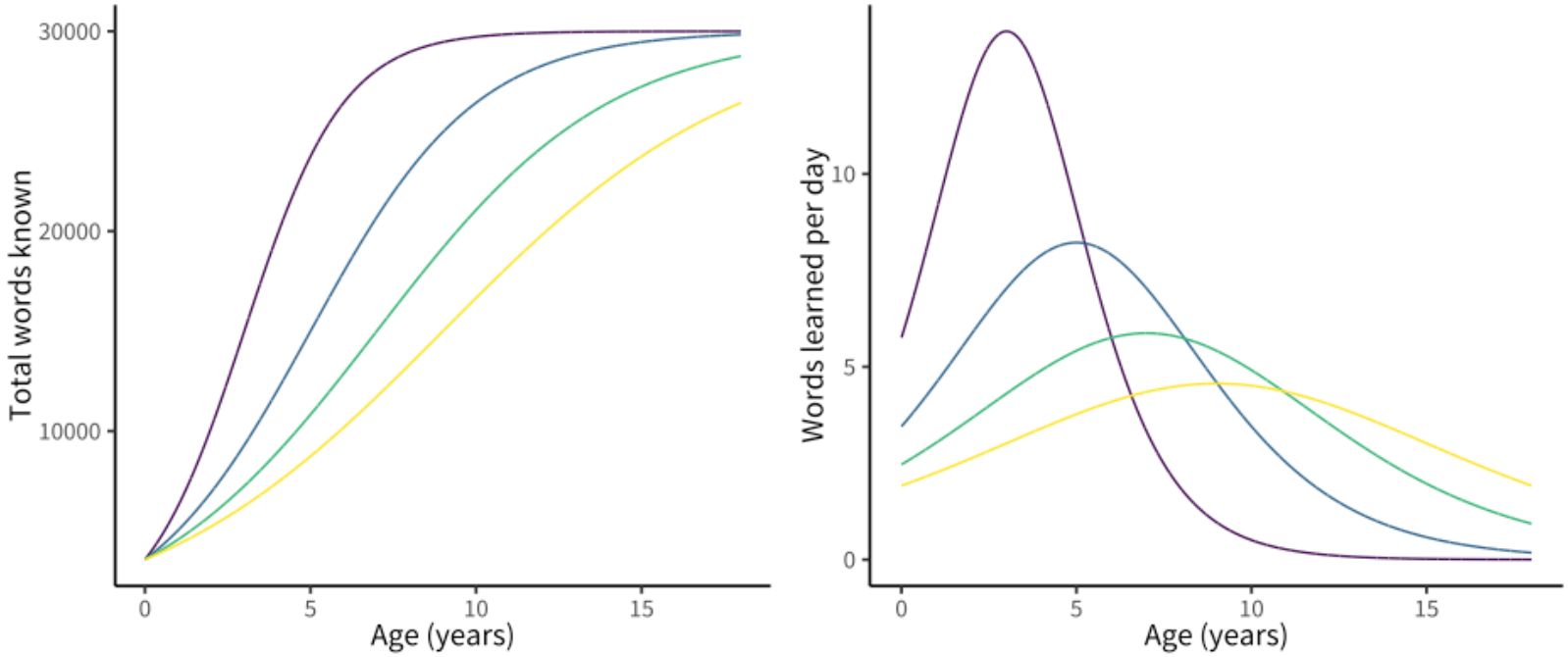
2. ANIMALS (Real or Toy) (43)

alligator	duck	penguin
animal	elephant	pig
ant	fish	pony
bear	frog	puppy
bee	giraffe	rooster
bird	goose	sheep
bug	hen	squirrel
bunny	horse	teddybear
butterfly	kitty	tiger
cat	lamb	turkey
chicken	lion	turtle
cow	monkey	wolf
deer	moose	zebra
dog	mouse	
donkey	owl	



CDI data at http://wordbank.stanford.edu

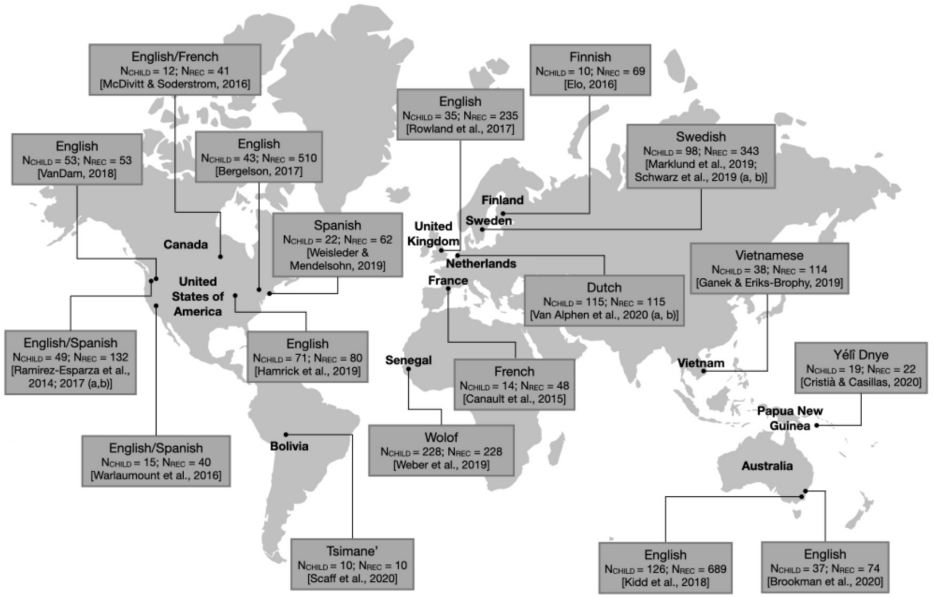
The rate of word learning

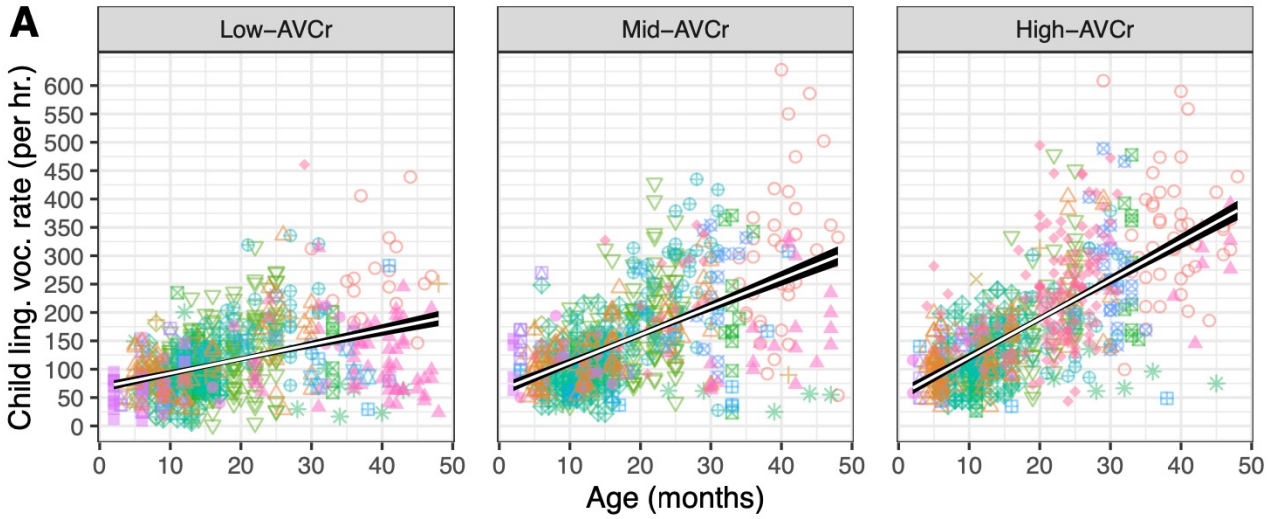


Children are “learning” several words each day!

Input to uptake

Automated analysis of vocalizations through 1000 daylong audio recordings reveals robust input-uptake correlations





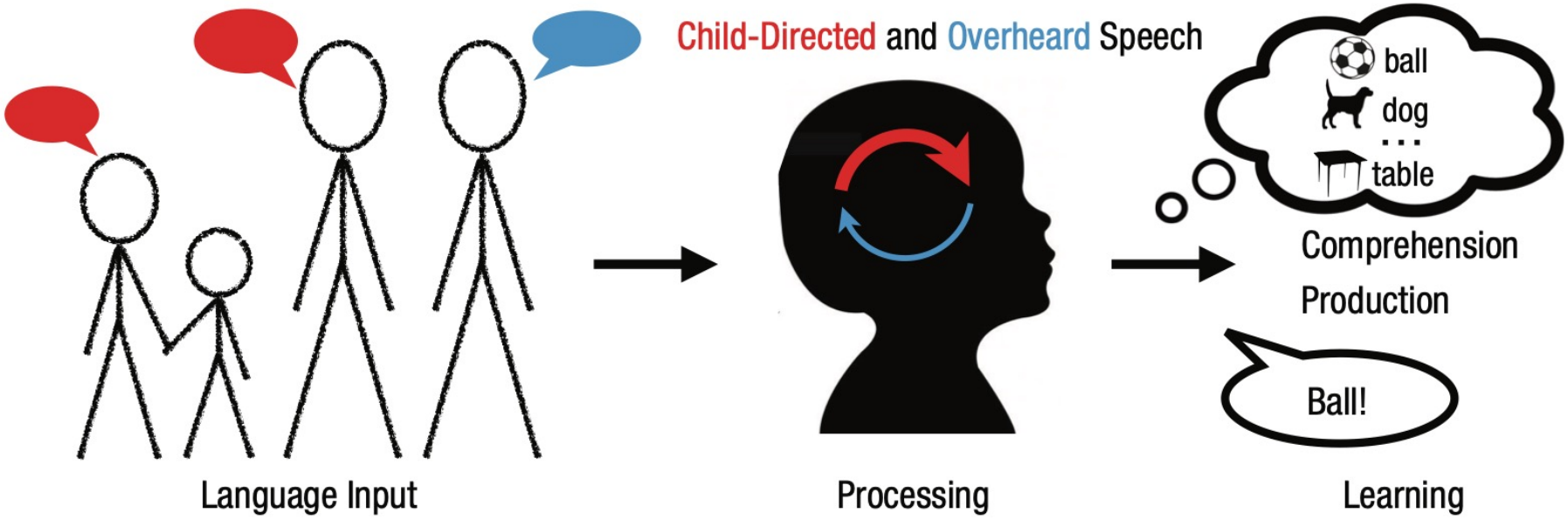
Bergelson et al. (psyarxiv)

Outline

1. Basics of early word learning
2. **Toward a "standard model" of language learning**
3. Developing resources (to measure quantities in the model)
   - Input – childes-db
   - Learning – wordbank
   - Processing – peekbank
4. The course toolset

A “standard model” for language learning

Everyone has (more or less) the same theory in mind



Can we formalize this theory and put it in contact with data? What data?

Kachergis, Marchman, & Frank (2022), *Current Directions in Psych Sci*

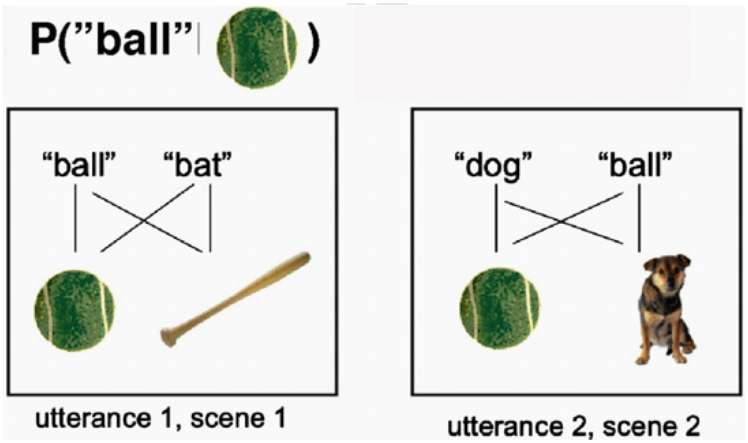
Associative learning foundations

”Statistical learning”

golabupadotitupiro
bidakugolabupadoti
golabupadotitupiro
bidakugolabupadoti

Saffran, Aslin, & Newport (1996)

“Cross-situational learning”



Yu & Smith (2008)

Defusing the Childhood Vocabulary Explosion

Bob McMurray

Between birth and adulthood, children learn about 60,000 words, on average, 8 to 10 words per day. Studies consistently reveal that, during the second postnatal year, word learning accelerates dramatically (*1*). Although the acceleration is continuous and not stagelike (*2*, *3*), this so-called vocabulary explosion is a foundational phenomenon that theories of language acquisition must address.

acquisition threshold, it is learned. This mode exhibits the characteristic pattern of slow learnin followed by acceleration (Fig. 1B, black line Deceleration is seen at the end of acquisitior something that has been hypothesized (*3*) but nc examined empirically [Supporting Onlin Material (SOM) text S2].

Further simulations examined mechanism that leverage initial words to facilitate learning

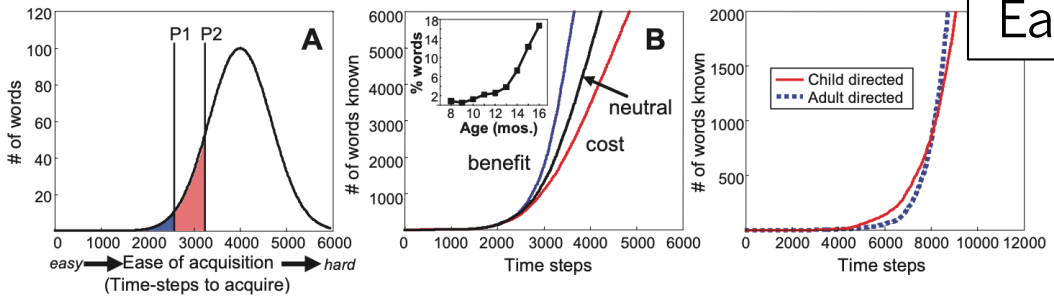
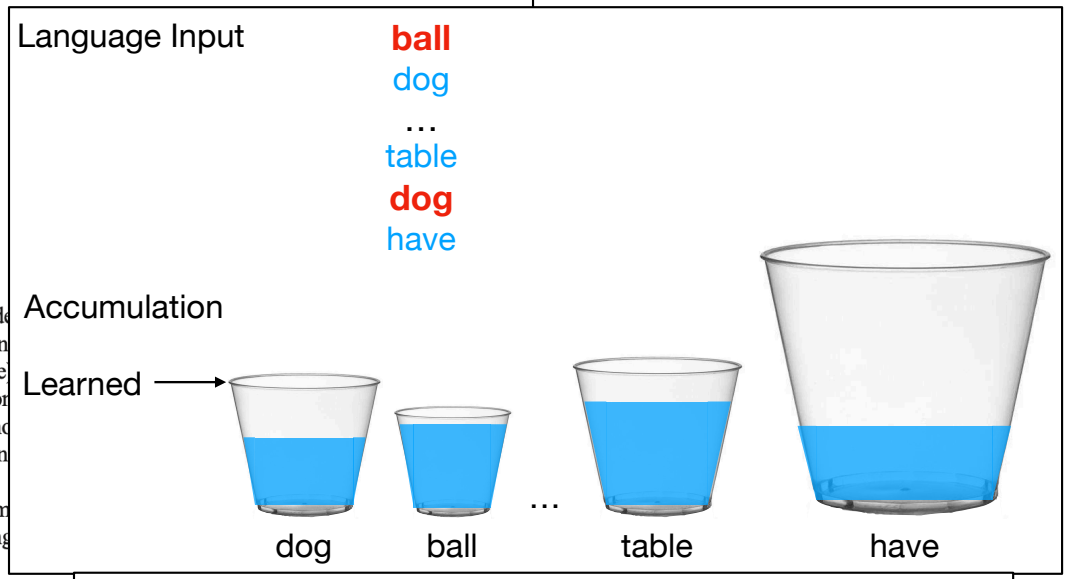


Fig. 1. (**A**) A Gaussian distribution of time to acquisition. Between 0 and 2600 time steps (P1), 1481 words are acquired. An additional 3966 are acquired in the next 600 (P2). (**B**) Vocabulary size as a function of time in initial simulations, and when learning a word offers a cost or benefit to future learning. (Inset) Percentage of words on MacArthur Communicative Development Inventory produced by children as a function of age (SOM text S1). (**C**) Acquisition in simulations based on word frequency.

processes are not causally necessary to explain acceleration; such processes may, in fact, arise in response to acceleration, or they may offset other processes that slow learning (*6*). Moreover, the model's generality suggests that any parallel learning system (e.g., motor patterns and concepts) should behave similarly. Acceleration is an unavoidable by-product of variation in difficulty. It should not be misconstrued as evidence for specialized learning.



Each word is a bucket, filled by input

Cognitive Science 33 (2009) 1503–1523

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01071.x

On Leveraged Learning in L
Relationship to

Colleen Mitchell,[a]

[a]*Department of Mathematics, and*
[b]*Department of Psychology, and*

Received 10 September 2008; received in revised

Abstract

Children at about age 18 months experience acc
sion is a robust phenomenon, although the exact sh
of explanations, which we term collectively as *le*
words helps with the learning of others. In this fra
learning is slow. As more words are acquired, new
of early words that fuels the explosion in learning
learning in the vocabulary spurt by proposing a sim
that leverage can change both the shape and timing
eration if it did not exist in the corresponding mod
to the Zipfian distribution of word frequencies, whi
ation, but t
complex.

Keywords:
ment

OPEN ACCESS Freely available online

PLOS ONE

A Computational Model Associating Learning Process, Word Attributes, and Age of Acquisition

Shohei Hidaka*

Japan Advanced Institute of Science and Techno

Abstract

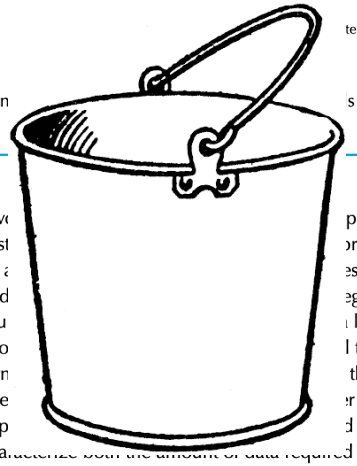
We propose a new model-based
computational tool for understand
AoAs as measures of vocabulary g
between three theoretical factors
how different learning processes,
required to acquire a word, likely
three respects. The first analysis s
standard alternative. The second
psychological attributes, such as fr
trend predicted by our estimated
on word learning in children. We

Citation: Hidaka S (2013) A Computational M
journal.pone.0076242

Editor: Johan J. Bolhuis, Utrecht University, T

How Data Drive Early Word Learning: A Cross-Linguistic Waiting Time Analysis

Francis Mollica[1] **and Steven T. Piantadosi**[1]



Keywords: word learnin

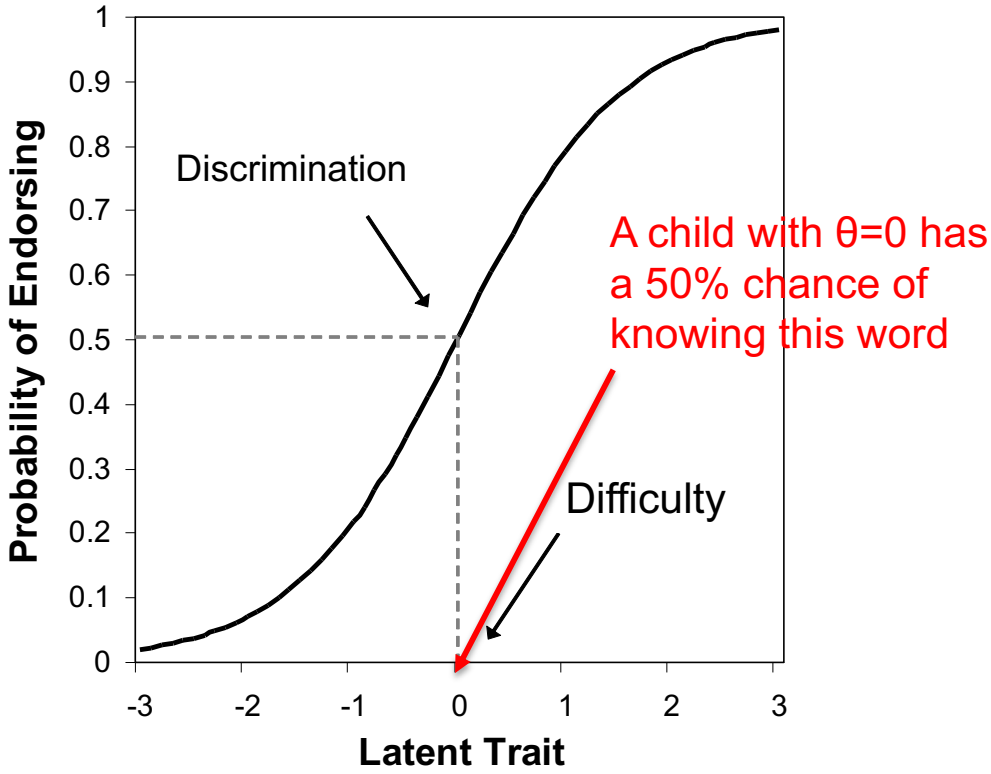
ABSTRACT

The extent to which w
is a longstanding quest
influence learning on
can rapidly learn word
information across mu
of word acquisition no
in early language learn
accurately predicts the
parameters suggest a p
the model directly cha
informative data occurs. With high statistical certainty, words require on the order of ∼ 10 learning instances, which occur on average once every two months. Our method is extremely simple, statistically principled, and broadly applicable to modeling data-driven

Can these models be used to connect variability in input to variability in uptake?

Item response theory

- A model of each child:
  - Has a latent parameter on the trait (θ)
  - Estimate of language ability
- A model of each word:
  - How difficult it is
  - Optionally: How well it discriminates low vs. high ability kids
- Allows prediction about new words and new children

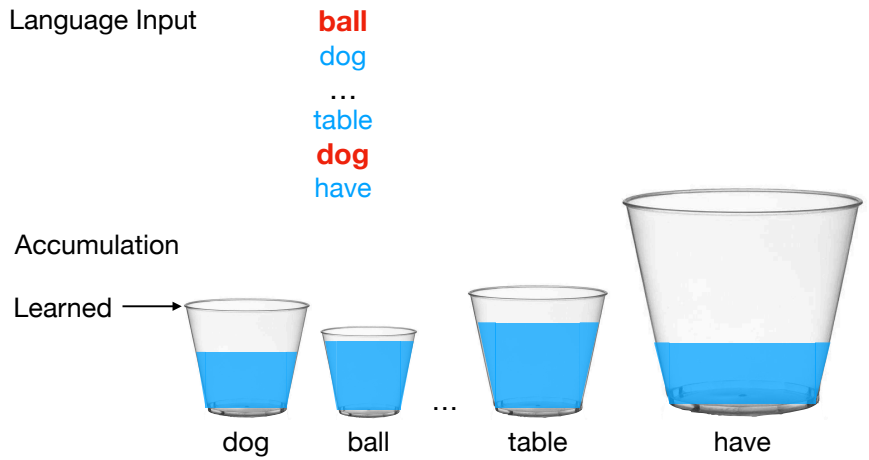


ICC = item characteristic curve

Embretson & Reise (2001)

Accumulator models are IRT models

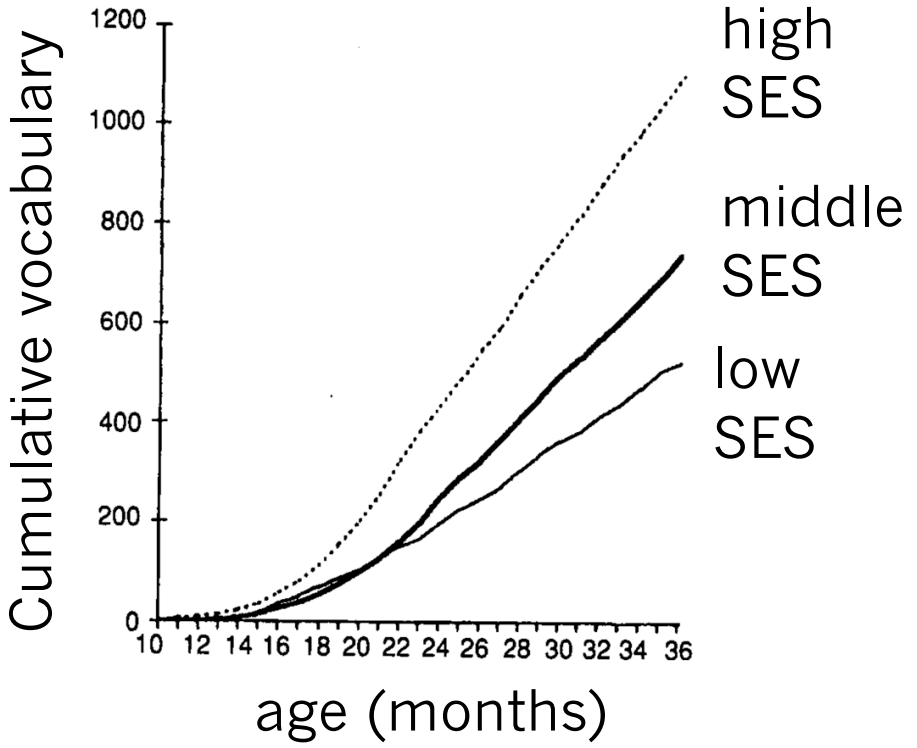
- Parameters map directly
  - Bucket size is word difficulty
  - Drip rate is child's learning rate X rate of input
- These models can be fit with mixed effects regression (de Boeck et al., 2011)
- But to be interpretable they require real measurements in <u>absolute units</u>
  - How many words per hour are heard?
  - How many total words are known?



Kachergis, Marchman, & Frank (2022)

Models are sometimes useful because they are wrong

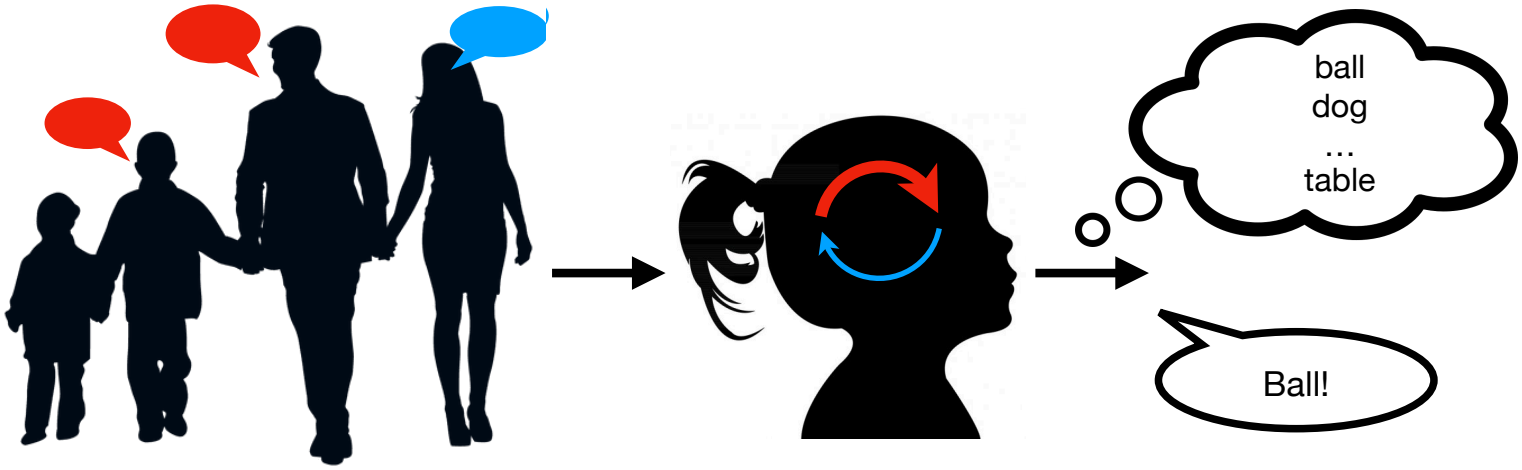
- Does input variation predict learning?
  - within culture, across populations?
  - across cultures?
- Failures to predict would constitute evidence for
  - processes beyond simple accumulation (e.g., learning from overhearing)
  - variation in learning process across populations



Hart & Risley (1995),
Cf. Raz & Beatty (2018), Sperry et al. (2018)

Outline

1. Basics of early word learning
2. Toward a “standard model” of language learning
3. **Developing resources (to measure quantities in the model)**
   - Input – childes-db
   - Learning – wordbank
   - Processing – peekbank
4. The course toolset



Language Input

Processing

Learning

CHILDES

Child Language Data Exchange System

CHILDES is the child language component of the ***TalkBank*** system.
TalkBank is a system for sharing and studying conversational interactions.

System

Ground Rules*

Contributing New Data

IRB Principles

Overviews and Introductions

Other Child Language sites

Unicode and IPA for ***Mac***

Unicode and IPA for ***Windows***

Database

Index to Corpora*

Browsable Database

LuCiD Toolkit

childes-db

CLAN

Teaching Resources

YouTube Examples

Bibliographies

Manuals

CHAT - ***CLAN*** - ***MOR***

Tutorial Screencasts

SLP's Guide to CLAN and ***中文***

Brian MacWhinney : ***homepage***

MRC lexical dictionary

ChildFREQ ***Site*** and ***Paper***

Special Procedures

CA analysis

Digitized video

Digitized audio

Versions

Derived Corpora and Counts

XML version of the database

Database Versioning

More Resources

Building a New Corpus

CCT Computerized Comprehension

LEAT Assessment Tool

The original “big data” for child language

... in a shared research environment with open tools and resources

childes-db API Tutorial Visualizations Documentation R

childes-db

A flexible and reproducible interface to CHILDES

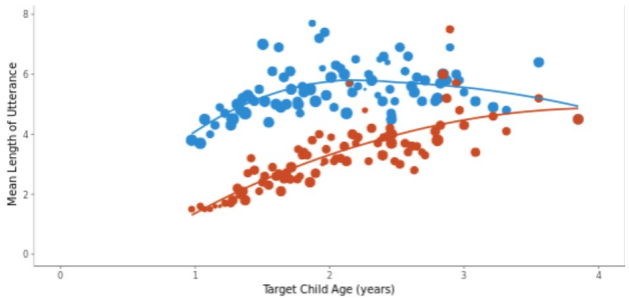
API Tutorial

Get a hands on walk-through on accessing `childes-db` through R.

```
> library(childesr)
> d_adam_prod <- get_tokens(collection = NULL,
+                           corpus = "Brown",
+                           role = "target_child",
+                           age = NULL,
+                           sex = NULL,
+                           child = "Adam",
+                           token = c("dog", "ball"))
Getting data from 1 child in 1 corpus ...
```

Visualizations

Explore the data in `childes-db` using our interactive applications.



Sanchez*, Meylan* et al. (2016), *Behavior Research Methods*

childesr 0.1.2 | Reference | Articles | Changelog

Accessing `childes-db`

2019-10-16

Source: `vignettes/access_childes_db.Rmd`

Overview

The `childesr` package allows you to access data in the childes-db from R. This removes the need to write complex SQL queries in order to get the information you want from the database. This vignette shows some examples of how to use the data loading functions and what the resulting data look like.

There are several different `get_` functions that you can use to extract different types of data from the childes-db:

- `get_transcripts()`
- `get_participants()`
- `get_tokens()`
- `get_types()`
- `get_utterances()`
- `get_speaker_statistics()`

Technical note 1: You do not have to explicitly establish a connection to the childes-db since the `childesr` functions will manage these connections. But if you would like to establish your own connection, you can do so with `connect_to_childes()` and pass it as an argument to any of the `get_` functions. If you do so, make sure to disconnect the connections you make by using `DBI::dbDisconnect()`, `childesr::clear_connections()`, or restarting your R session.

Technical note 2: We have tried to optimize the time it takes to get data from the database. But if you try to query and get all of the tokens, it will take a long time.

```
# load the library
library(childesr)
```

R API Overview

- Programmatic access to data
- Small number of calls to maintain
- Versioned database
- Caches lots of common data preprocessing
- Removes reproducibility uncertainty

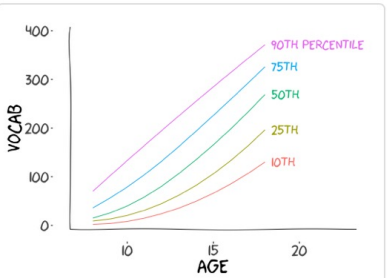
The MacArthur-Bates Communicative Development Inventory (CDI)

Ba, da, ma

Hi, bye, ball, dog, more, milk, mommy, daddy, hand, foot, nose, …

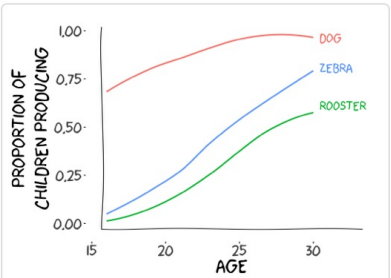
Wordbank

An open database of children's vocabulary development



Vocabulary Norms

Explore vocabulary size growth curves for various languages and demographic groups.



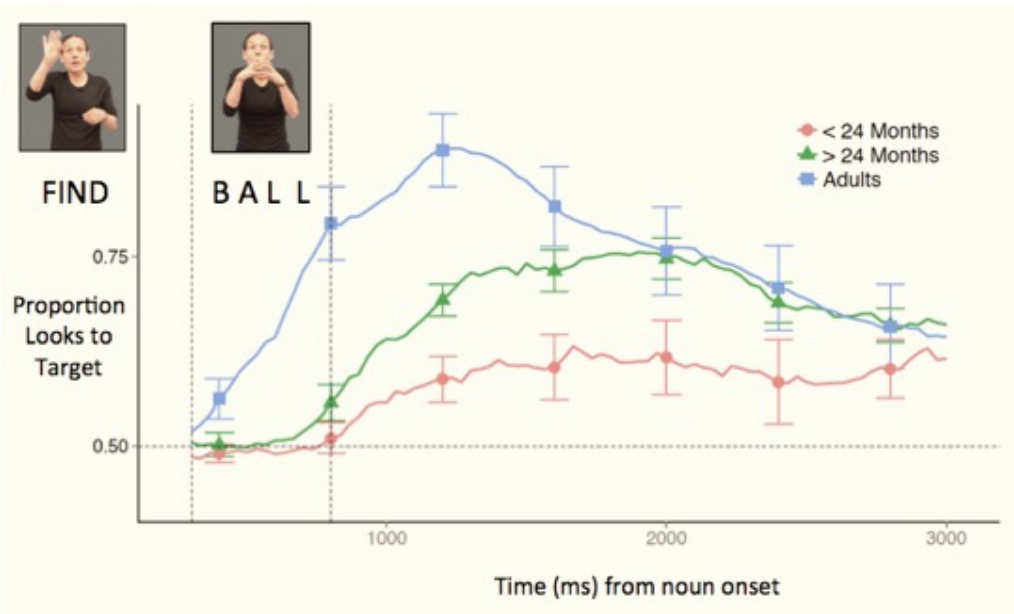
Item Trajectories

Explore trajectories of individual words, word categories, and grammar items.

Wordbank contains data from 79,027 children and 93,258 CDI administrations, across 40 languages and 76 instruments:

Frank et al. (2017), *Journal of Child Language*

Looking-While-Listening (LWL)



Parallel results in studies with diverse populations of children in the U.S.

- Spanish & English speakers in lower and higher SES families
- Late talkers
- Children born premature
- Emerging & simultaneous bilinguals
- Deaf children learning ASL

Slide courtesy Anne Fernald

MacDonald, Lamarr, Corina, Marchman, & Fernald (2018)

peekbank Data Access Tutorial Visualizations Contributors Documentation About

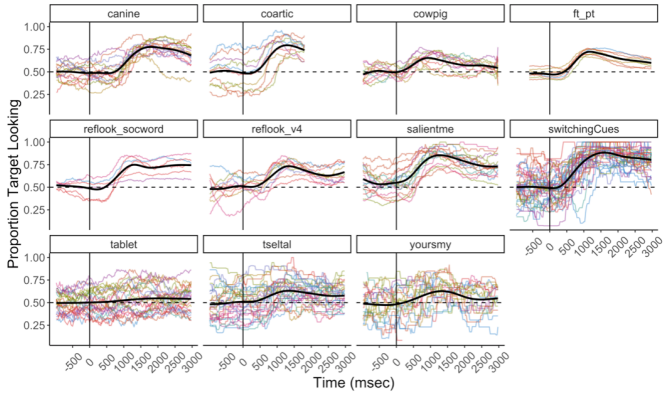
peekbank

A flexible and reproducible interface to developmental eyetracking datasets

What is peekbank?

peekbank is a flexible and reproducible interface to developmental eyetracking datasets.

The Peekbank project is an open database storing eye-tracking datasets on children's word recognition in a well-documented, easily accessible, tabular format. It also provides processing tools for standardizing eye-tracking data across data sources (`peekds` R package), interfaces for accessing the database (`peekbankr` R package), and applications for visualizing the data (Peekbank Shiny App).



Data Access Tutorial

Get started accessing the data here

Interactive Visualizations

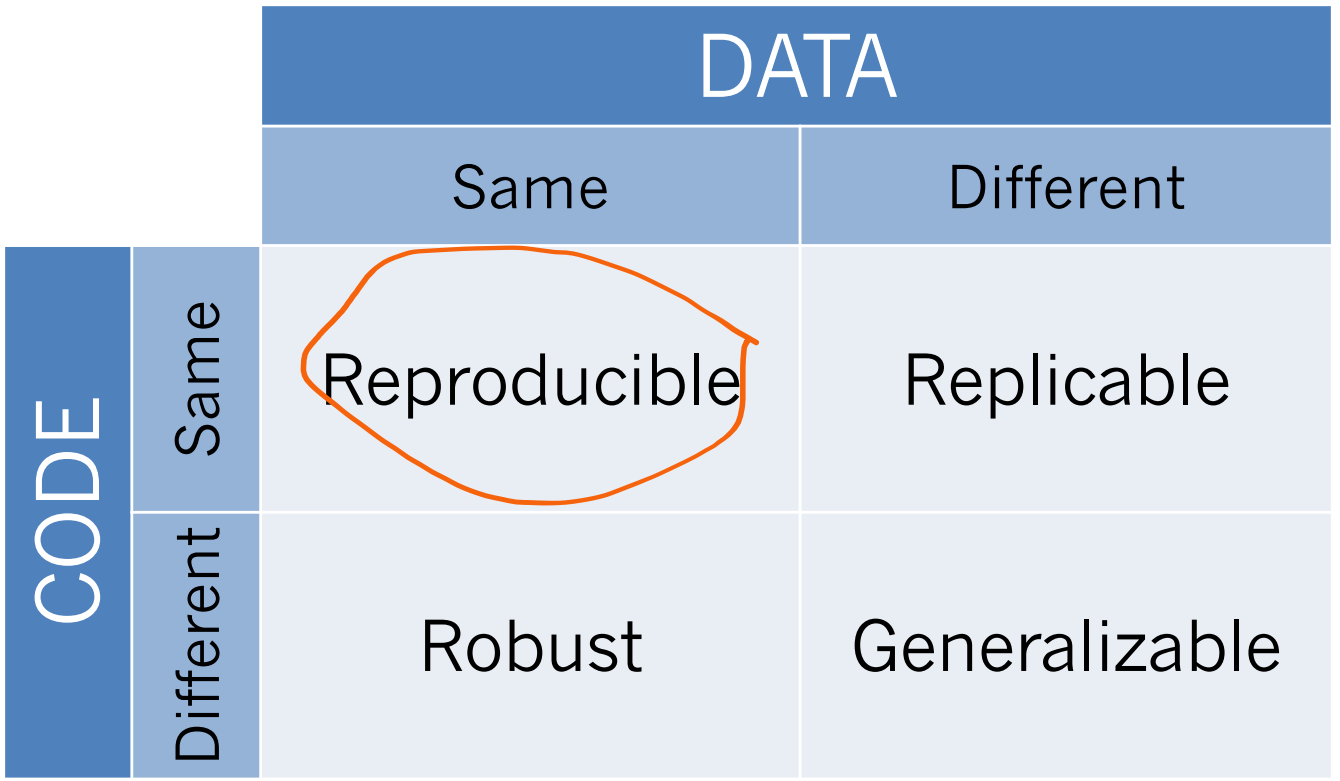
Dynamically generate visualizations of the data

Documentation

Further information about the database

Outline

1. Basics of early word learning
2. Toward a “standard model” of language learning
3. Developing resources (to measure quantities in the model)
   - Input – childes-db
   - Learning – wordbank
   - Processing – peekbank
4. **The course toolset**


DATA
Same
Different
CODE
Same
Different
Reproducible
Replicable
Robust
Generalizable

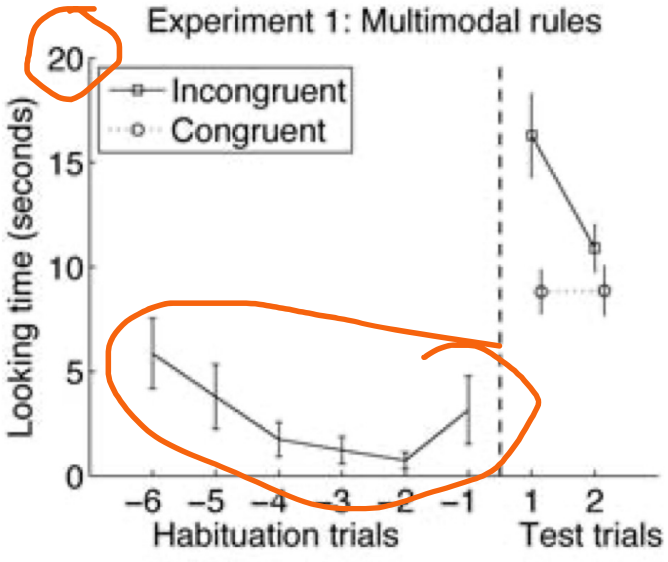
Developmental Science

Developmental Science 12:4 (2009), pp 504–509 DOI: 10.1111/j.1467-7687.2008.00794.x

PAPER

Information from multiple modalities helps 5-month-olds learn abstract rules

Michael C. Frank,[1] Jonathan A. Slemmer,[2] Gary F. Marcus[3] and Scott P. Johnson[4]



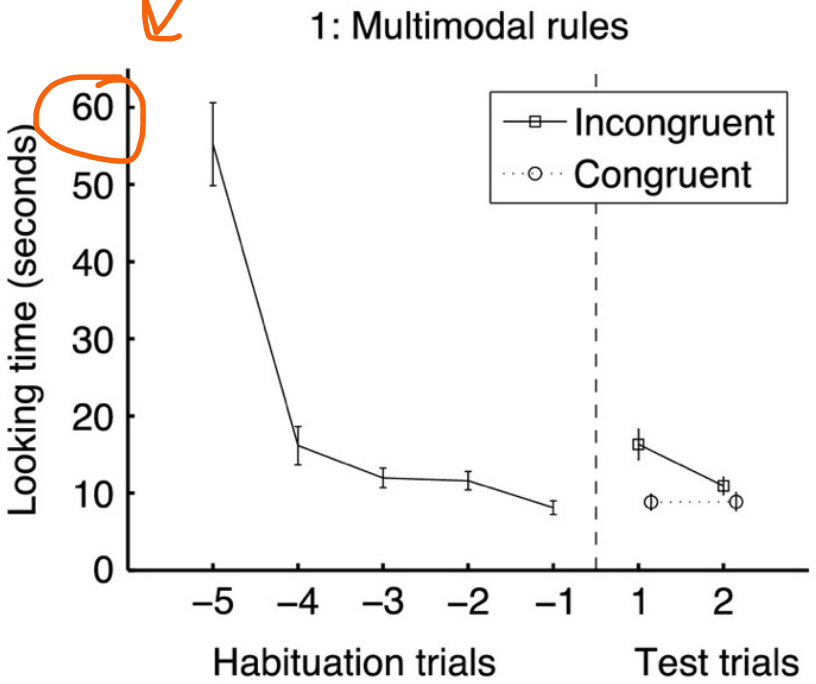
Developmental Science

Erratum | Free Access

Erratum

This article corrects the following:

First published: 23 February 2013 | https://doi.org/10.1111/desc.12060 | Citations: 1



Workflow

1. Get Excel data
2. Write Matlab code
3. Export figure as .EPS
4. Tweak in Illustrator
5. Drag into Word

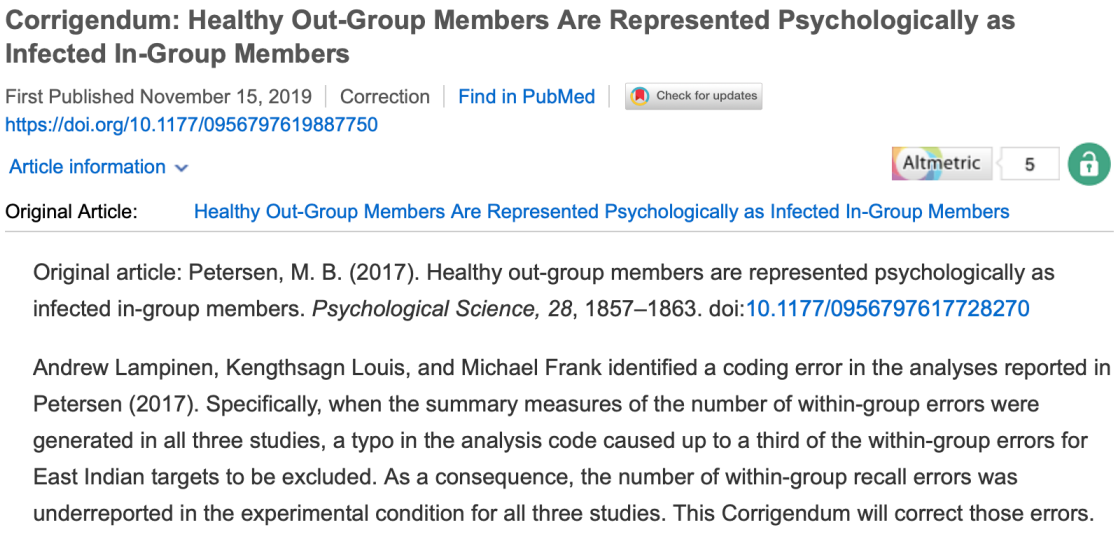
```
%% new plots
expts = {'1: Multimodal rules','2: Visual rules','3: Auditory rules
(checkerboard)','3: Auditory rules'};
expt_nums = [1 2 4];
s = {[1 2],[3 4],[6 7]};
figure(1); clf;
set(gcf,'Position',[   121    48   885   694]);

for i = 1:3
  subplot(2,4,s{i})
  set(gca,'FontSize',18);
  e = expt_nums(i);
  habit_times = [mean(HM7(EXP==e)) mean(HM8(EXP==e)) mean(HM9(EXP==e)) ...
    mean(HM10(EXP==e)) mean(HM11(EXP==e)) mean(HM12(EXP==e))]./1000;
  habit_errs = [stderr(HM7(EXP==e)) stderr(HM8(EXP==e)) stderr(HM9(EXP==e)) ...
    stderr(HM10(EXP==e)) stderr(HM11(EXP==e)) stderr(HM12(EXP==e))]./1000;
  hold on
  errorbar(1:6,habit_times(1:6),habit_errs(1:6),'-k')
  a = errorbar(7:8,[mean(N1C(EXP==e)) mean(N2C(EXP==e))]./1000,...
    [stderr(N1C(EXP==e)) stderr(N2C(EXP==e))]./1000,'-sk');
  b = errorbar([7:8]+.15,[mean(F1C(EXP==e)) mean(F2C(EXP==e))]./1000,...
    [stderr(F1C(EXP==e)) stderr(F2C(EXP==e))]./1000,':ok');
  axis([0 9 0 20]);
  if i == 1
    ylabel('Looking time (seconds)')
    xlabel('Habituation trials            Test trials','Position',[5 -2.2])
    legend([a,b],{'Incongruent','Congruent'},'Location','NorthWest')
  end
  title(expts{e});
  set(gca,'XTick',[1 2 3 4 5 6 7 8],'XTickLabel',{'-6','-5','-4','-3','-2','-
  1','1','2'})
  line([6.5 6.5],[0 20],'LineStyle','--','Color',[0 0 0],'LineWidth',1)
end
```

Wrong trial numbers!

Not just me...

```
gen recall1=.
replace recall1=0 if Q21==1 replace recall1=1 if Q21==3 | Q21==5 | Q21==6
replace recall1=2 if Q21==2 | Q21==4 | Q21==7 | Q21==8
replace recall1=0 if Q69==1
replace recall1=1 if Q69==3 | Q69==5 | Q69==6
replace recall1=2 if Q69==2 | Q69==4 | Q69==7 | Q69==8 ta recall1
```



Corrigendum: Healthy Out-Group Members Are Represented Psychologically as Infected In-Group Members

First Published November 15, 2019 | Correction | Find in PubMed | Check for updates

https://doi.org/10.1177/0956797619887750

Article information

Altmetric 5

Original Article: Healthy Out-Group Members Are Represented Psychologically as Infected In-Group Members

Original article: Petersen, M. B. (2017). Healthy out-group members are represented psychologically as infected in-group members. *Psychological Science, 28*, 1857–1863. doi:10.1177/0956797617728270

Andrew Lampinen, Kengthsagn Louis, and Michael Frank identified a coding error in the analyses reported in Petersen (2017). Specifically, when the summary measures of the number of within-group errors were generated in all three studies, a typo in the analysis code caused up to a third of the within-group errors for East Indian targets to be excluded. As a consequence, the number of within-group recall errors was underreported in the experimental condition for all three studies. This Corrigendum will correct those errors.

Low data and code availability renders most papers non-reproducible

Data & code availability (prevalence estimates)

	Data	Code
Psychology (2014-2017)[1]	2% [1-4%]*	1% [0-1%]
Social Sciences (2014-2017)[2]	7% [2-13%]	1% [0-3%]
Biomedicine (2015-2017)[3]	18.3% [11.6-27.3%]	0%

[1]Hardwicke et al. (2020a)
[2]Hardwicke et al. (2020b)
[3]Wallach et al. (2018)

*[95% confidence intervals]

Data not available on request

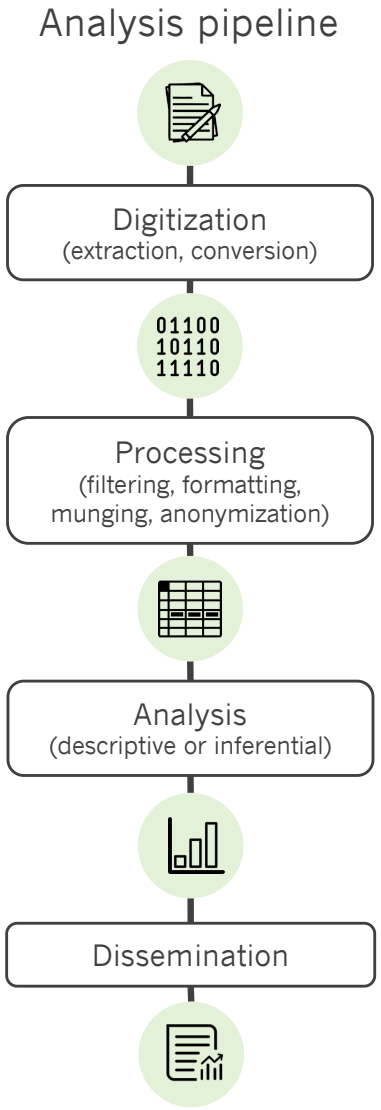
	Data shared
141 articles published in four major APA journals (2004)[4]	27%
516 ecology articles published (1991-2011)[5]	20%
111 most highly-cited articles published in psychology & psychiatry (2006-2016)[6]	14%

[4]Wicherts et al. (2006)
[5]Vines et al. (2014)
[6]Hardwicke & Ioannidis (2018)

What is a reproducible report?

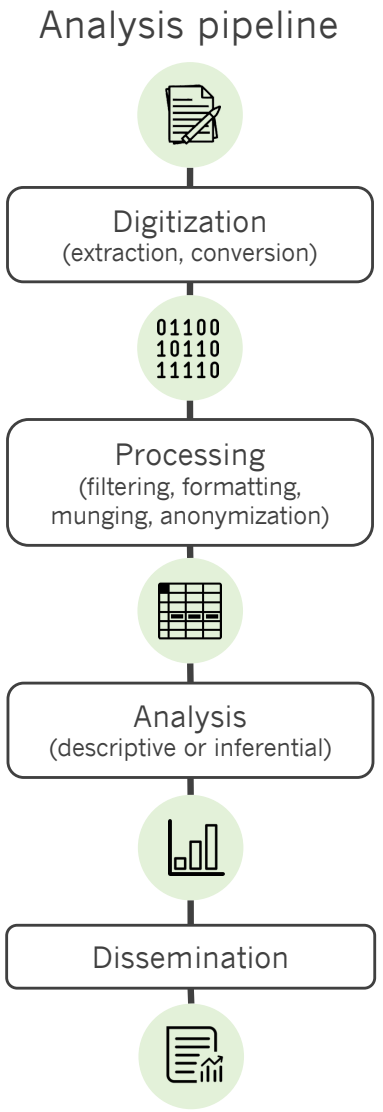
"...a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

- Buckheit & Donoho (1995)



What is a reproducible report?

A document for which the entire analysis pipeline, from primary data to reported numerical values, can be recreated, ideally by an independent scientist.



Why write reproducible reports?

1. Provenance, transparency, & verification

To know what numbers mean, we need to know where they are from.

Transparent sharing of analysis pipelines enables independent verification.

2. Error detection & mitigation

Scientists are human and humans make mistakes. Disorganized workflows with manual tasks, create opportunities for error.

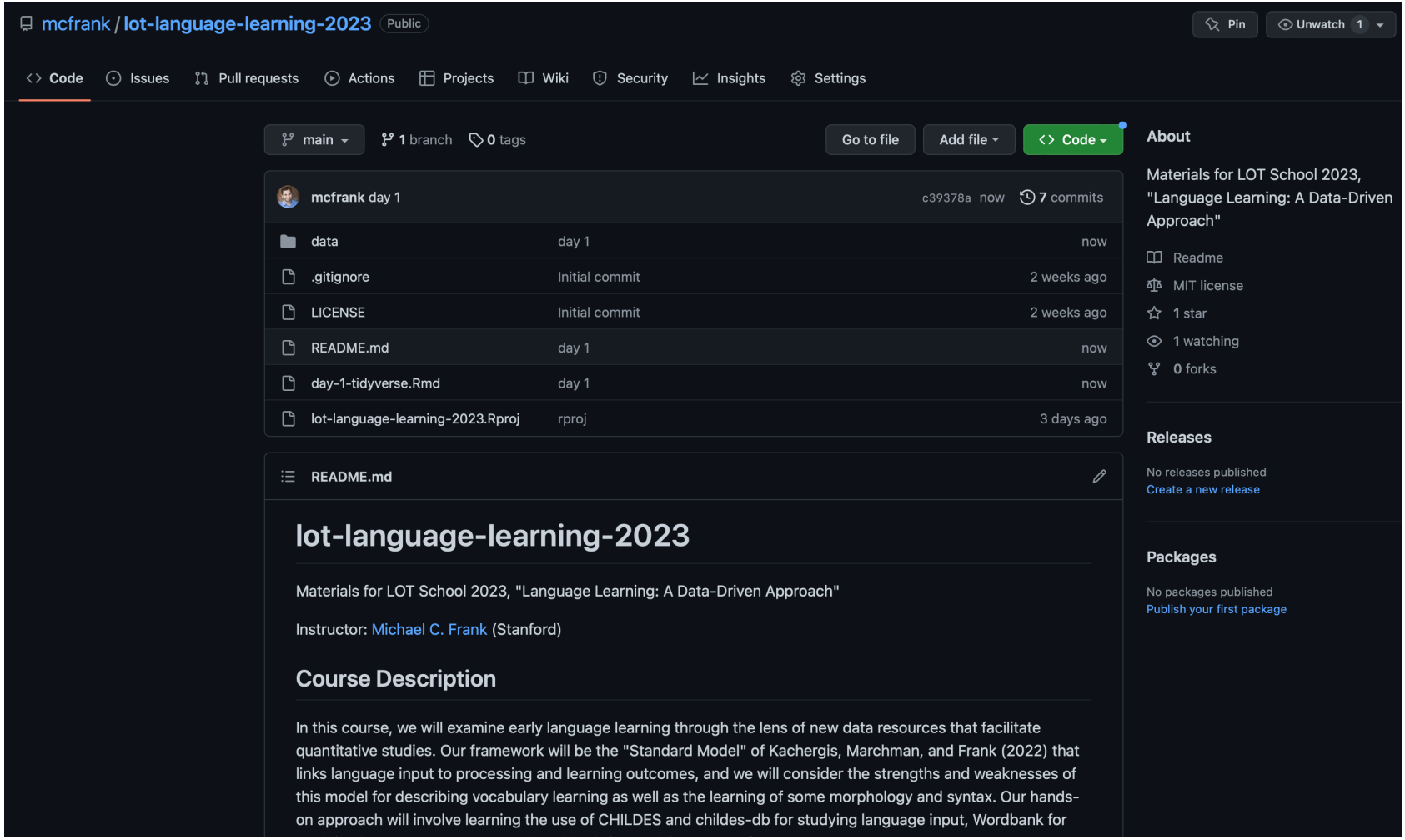
In a reproducible report, the flow of information through the analysis pipeline is directed and documented by code. This is easier to audit and manual tasks like copy & pasting are avoided, reducing opportunities for error.

3. Efficiency

Scientists often need to re-run or re-use analyses, for example in responding to reviewers, repurposing for a new project, or generating reporting outputs (posters, slides, papers etc.)

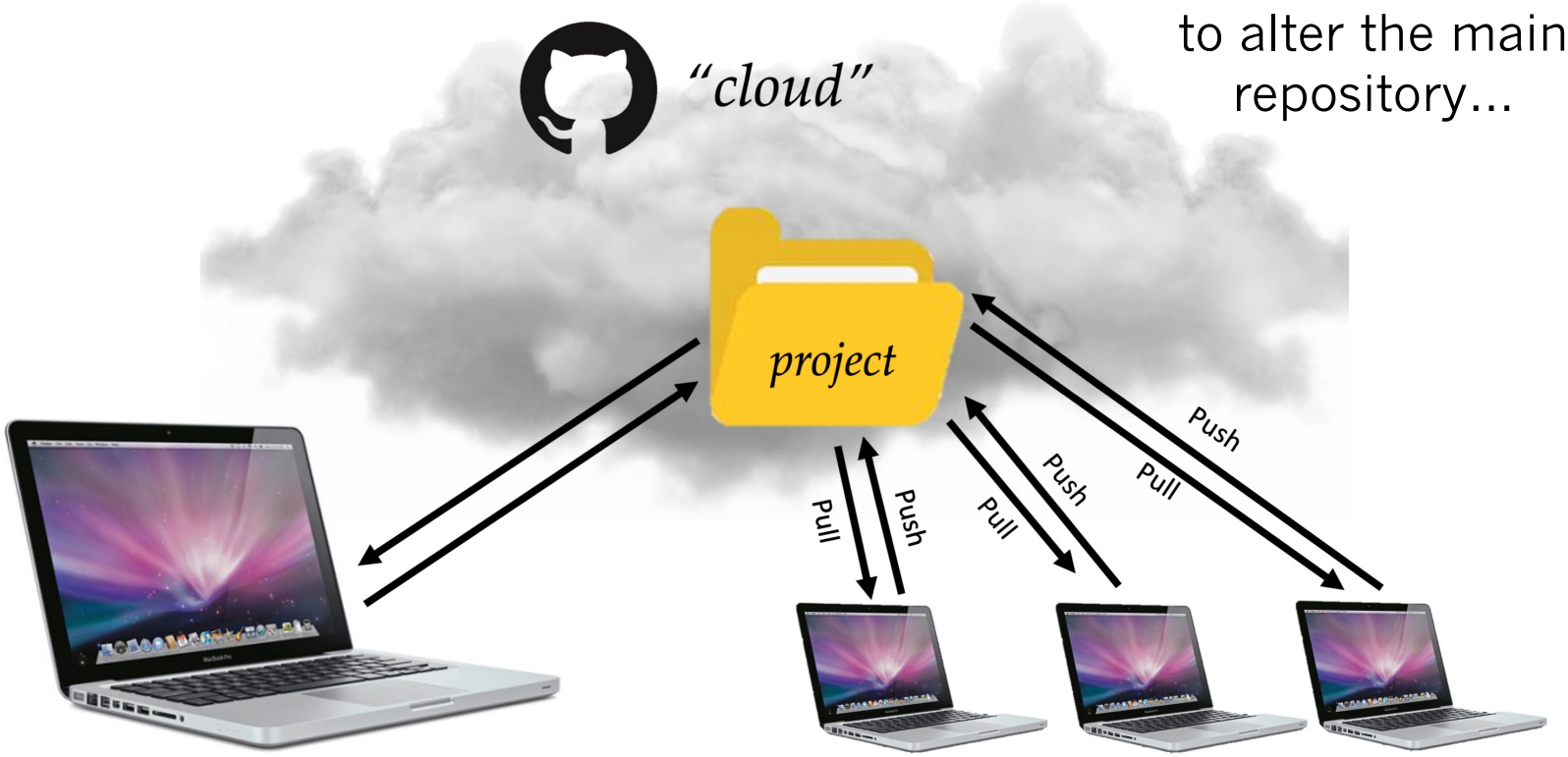
In a reproducible workflow, it is easier to build on your prior work by re-using code – there's no need to start from scratch each time or remind yourself what you did previously.

Our course repository



Git and github

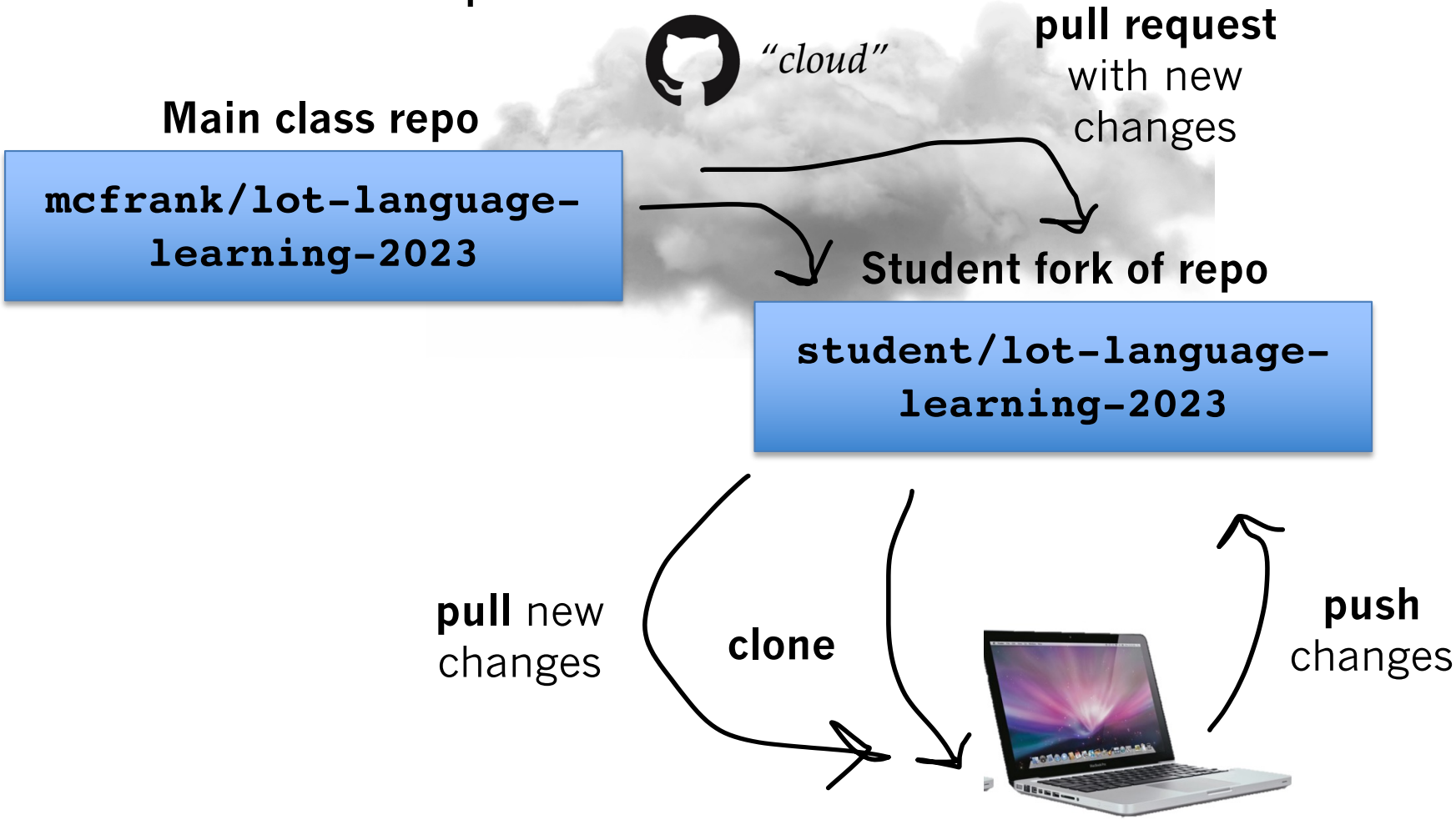
But, we don't want students to alter the main repository…



Two ways to engage with our course repo

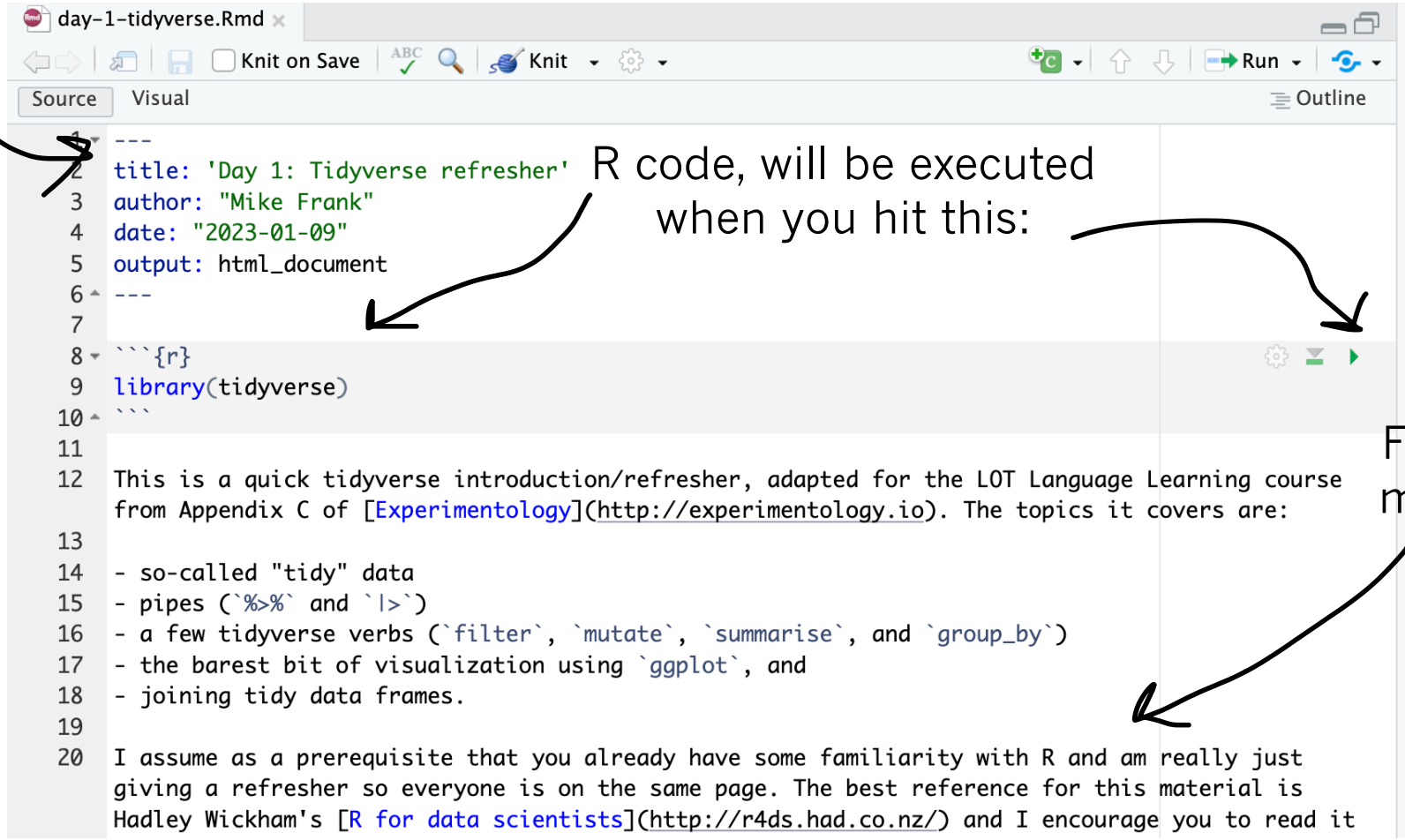
1. Download the repo directly onto your computer
   - Each day, copy the next day's code into a new Rmd file
2. "Fork" the repo on github (a bit trickier, but helps you master a good workflow)
   - Clone your fork onto your own computer
   - When you are done for the day, commit your changes to your local repo
   - The next day, pull changes from my repo to your fork
   - Then pull the changes to your local computer

The “fork and pull” model

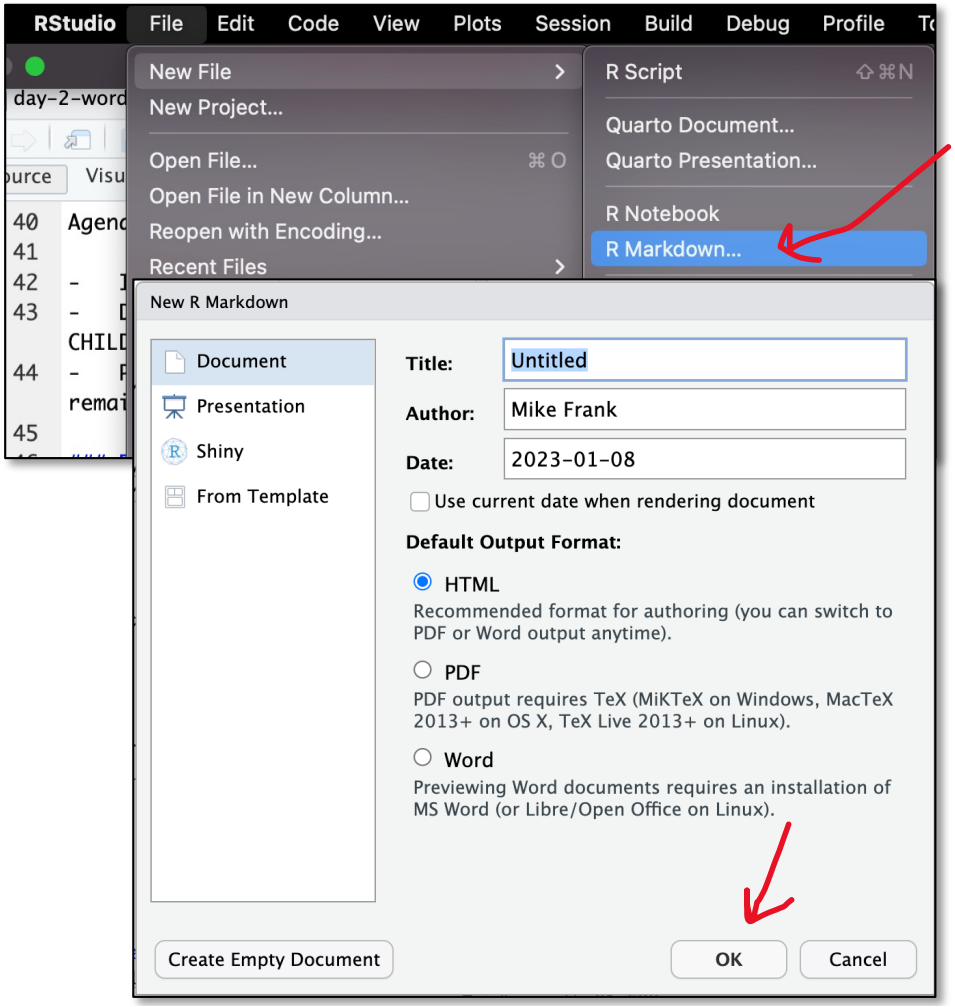


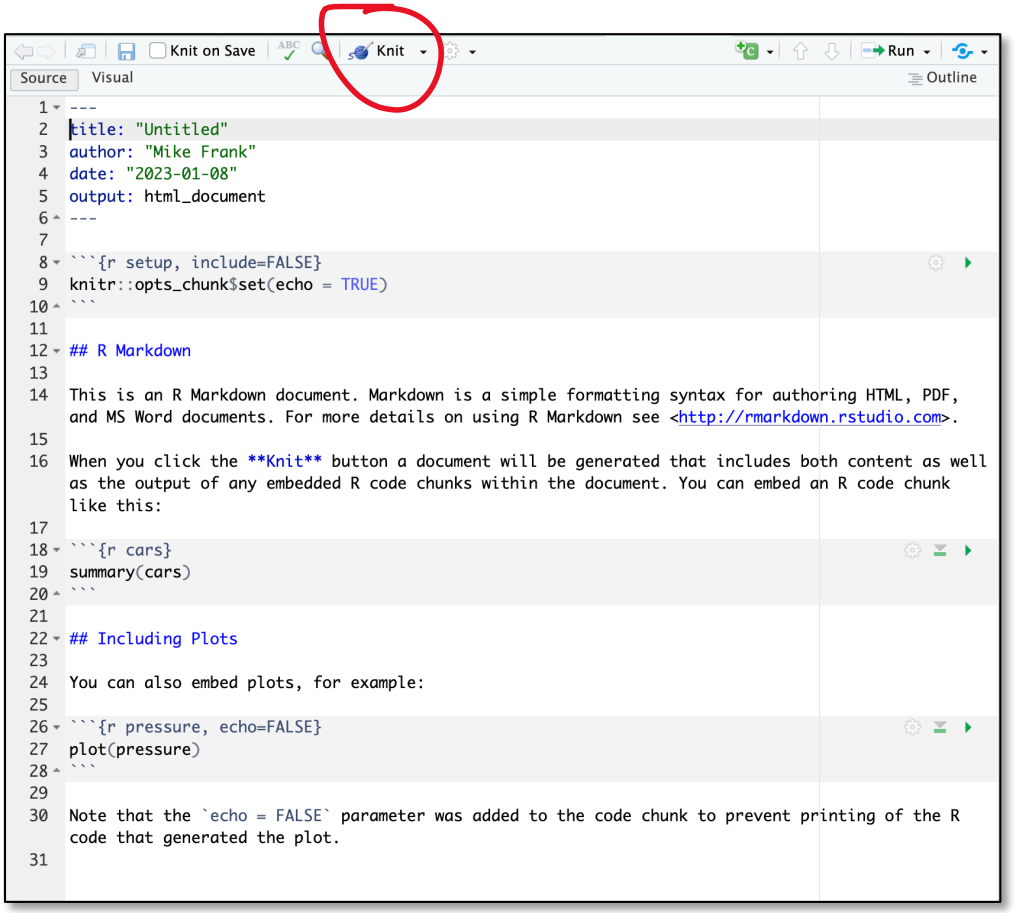
R Markdown

YAML header – don't mess with this



You try it!


RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
New File
New Project...
Open File... ⌘O
Open File in New Column...
Reopen with Encoding...
Recent Files
R Script ⇧⌘N
Quarto Document...
Quarto Presentation...
R Notebook
R Markdown...
New R Markdown
Document
Presentation
Shiny
From Template
Title: Untitled
Author: Mike Frank
Date: 2023-01-08
Use current date when rendering document
Default Output Format:
HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).
PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).
Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).
Create Empty Document
OK
Cancel


Knit on Save
Knit
Run
Source
Visual
Outline
---
title: "Untitled"
author: "Mike Frank"
date: "2023-01-08"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

The tidyverse...

- Open `lot-language-learning-2023.Rproj`
  - Project file ensures that you will be working from the correct directory
- Open `day-1-tidyverse.Rmd`
- We will gradually work through each topic, pausing for simple exercises

Exercises

I have a problem, please come help!

I'm done with this exercise and ready to move on