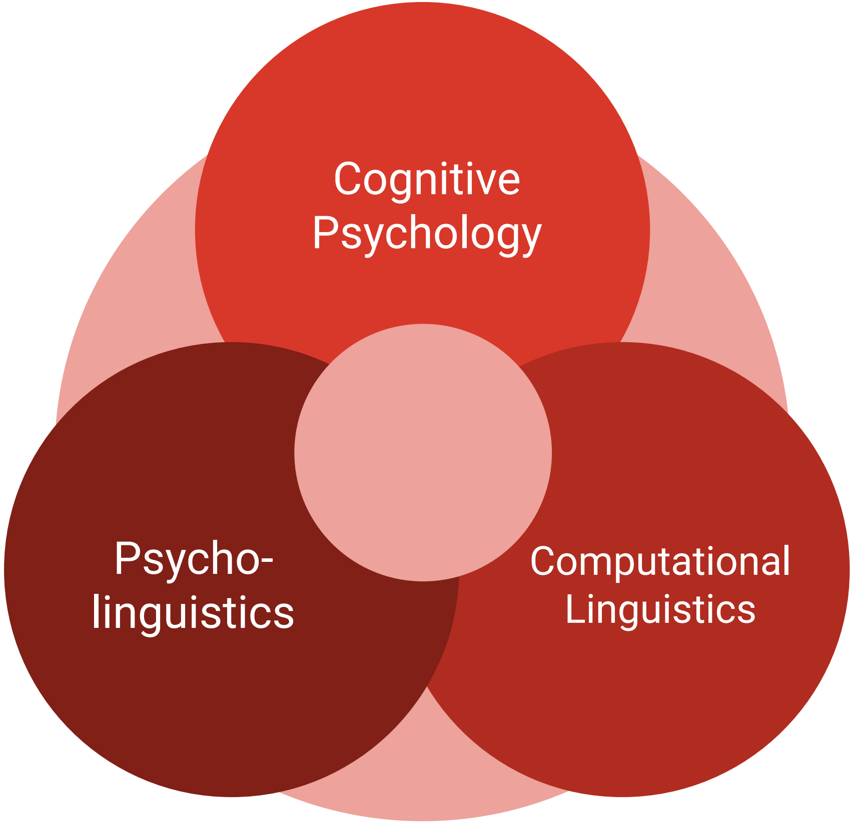
SPAM-L

PSACON2020 Study 007 Announcement

Semantic Priming Across Many Languages

- Erin M. Buchanan*, Harrisburg University of Science and Technology
- Maria Montefinese, University of Padova and University College London
- Felix Henninger, University of Mannheim
- Jack Taylor, University of Glasgow
- K. D. Valentine, Massachusetts General Hospital

Overview


Cognitive Psychology
Psycho-linguistics
Computational Linguistics

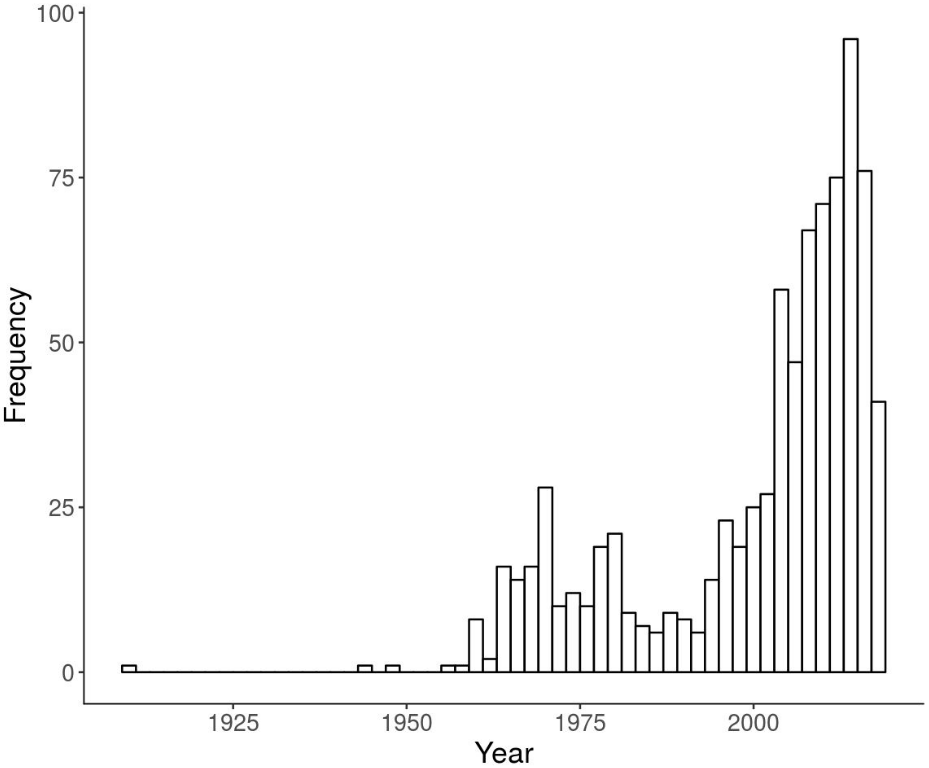
Mix + Match = A Mess

- We understand the importance of experimental control
- Many early studies used in-lab normed stimuli
  - Both Lucas (2000) and Hutchison (2003) have discussed how stimuli often were not “semantic”
- The definitions of similarity varies across studies

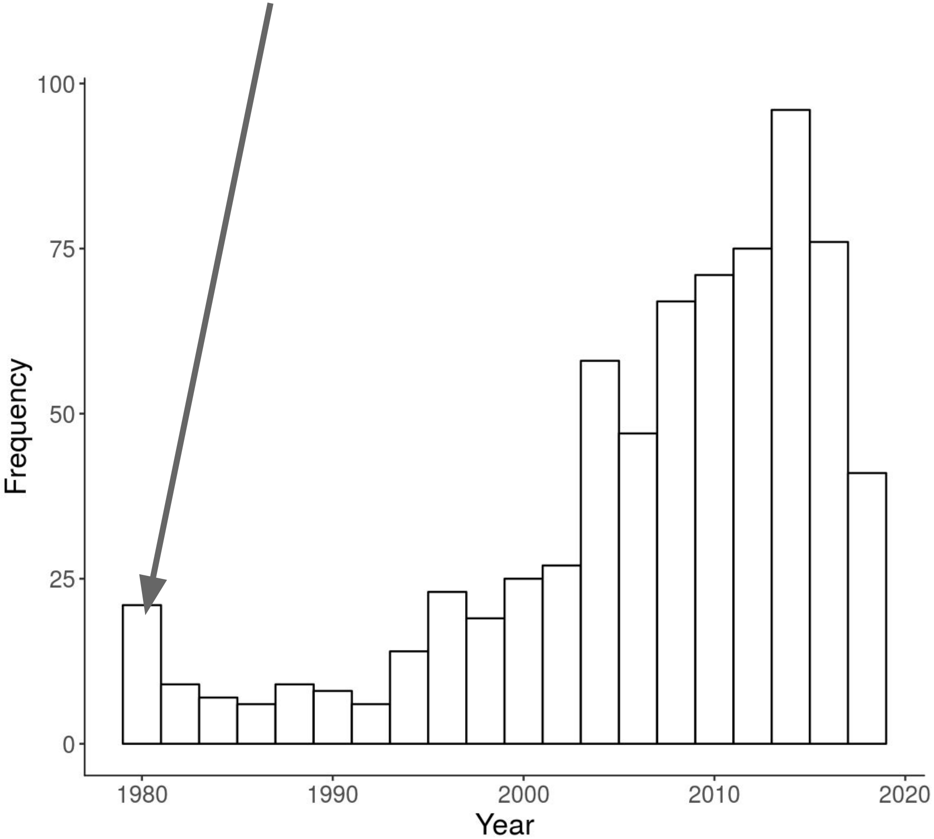
Normed Stimuli to the Rescue

- Buchanan, Valentine, & Maxwell (2019)
- Linguistic Annotated Bibliography
- https://wordnorms.com/

Normed Stimuli to the Rescue



Snodgrass & Vanderwart



Normed Stimuli to the Rescue

- Important!
- Controlled stimuli for new studies!
  - Reproducibility!
  - Replication!
- New and interesting research hypotheses!

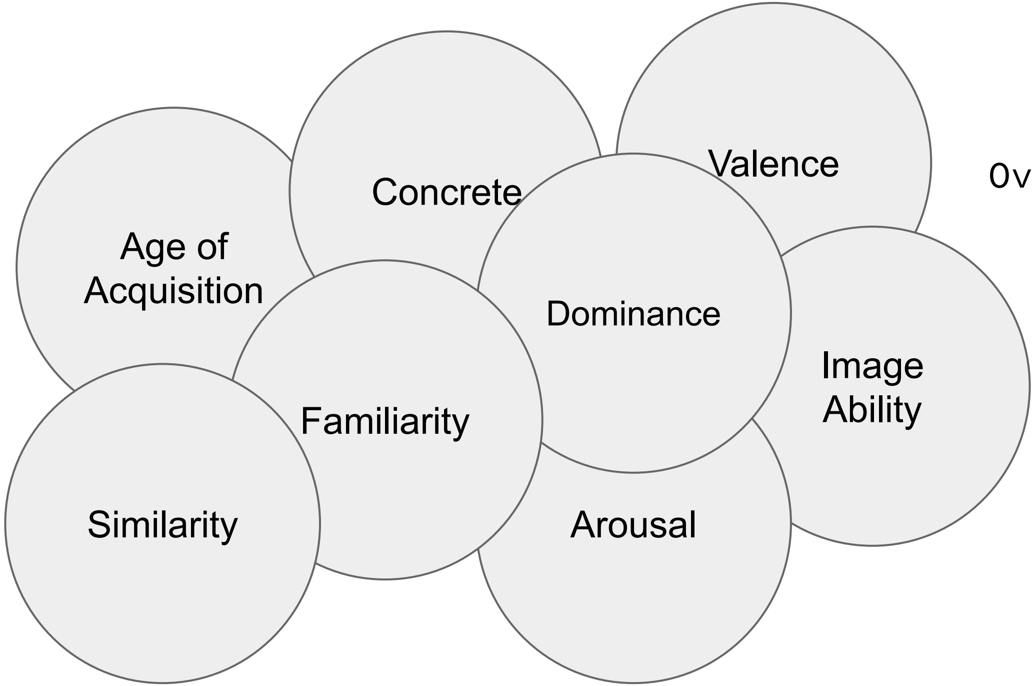
However, The work Sucks ...

- Buchanan, Valentine, & Maxwell (2019)
  - And previously, Buchanan et al. (2013)
- De Deyne, Navarro, Perfors, Brysbaert, & Storms (2019)
- Montefinese, Vinson, Vigliocco, & Ambrosini (2019)
  - And more from Montefinese et al. (2013)^2

Where's the Data?

- Corpus style norms
  - Subtitles
  - Twitter
  - Books
- Subjective norms
  - Ratings
  - Judgments

What's in the Data



Overlap = poor

Multiple languages?

What do we want to do?

- Online platform for data collection
- Semantic priming data + many languages + many variables
- R/Python/Shiny packages to connect to the data
- Secondary data challenge

Semantic priming

- Let’s do a demo of a lexical decision task!
- Words are linked in pairs:
  - Cue: *doctor*
  - Unrelated target: *tree*
  - Related target: *nurse*
  - Nonsense target: *tren*
- Semantic priming occurs when related words are responded to *faster* than other trial types.

Semantic Priming

- The Semantic Priming Project: Hutchison et al. (2013)
  - 1661 English words in lexical decision and naming tasks
  - These were paired with unrelated, related (two types), and nonsense words

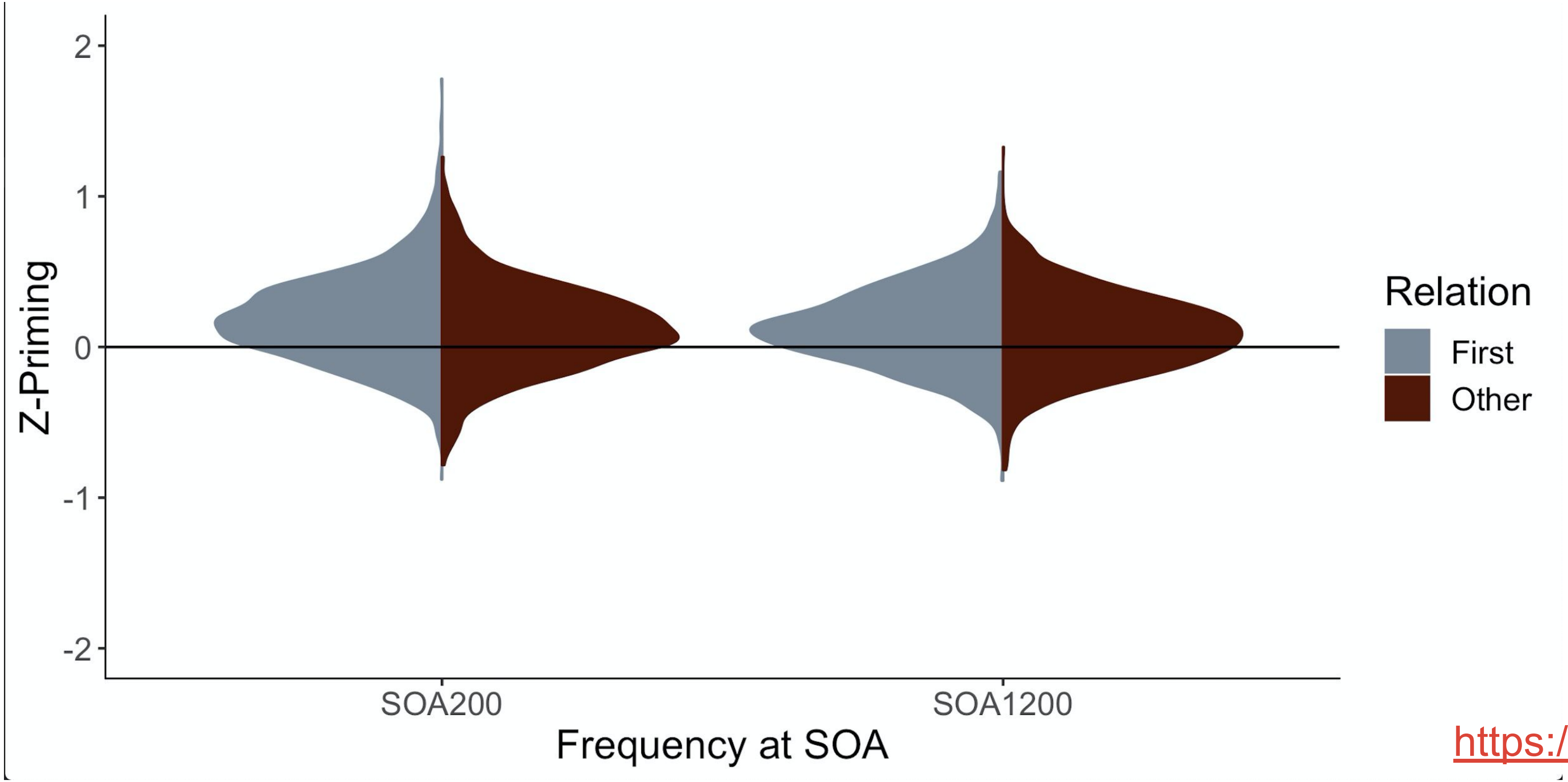
Key Differences

- Why do we need another study?
  - English only
  - Focused on target only lexical decision with two different stimulus onset asynchronies
  - Similarity defined by free association norms: Nelson et al. (2004)
  - Sample size *n* ~ 32 per pair by condition

Key Issues

- Sample size is probably too small for coverage/power
- Overlap with other stimuli still poor
- Is priming even reliable?
  - Heyman et al. (2016, 2018)
- Is priming even predictable?
  - Hutchison et al. (2008), see next slide

Key Issues



https://osf.io/74esw/

Outcome 1: Online Portal

- We will create an online portal to collect, store, and share the data
- https://smallworldofwords.org/en
  - Lowers the burden on research labs
  - Allows for data collection to occur in waves
  - Publication updates for data versus one-shot paper

Outcome 1: Online Portal

- The experiment will be programmed with labjs (what you saw in the demo!)
- Labjs has extensively worked on millisecond timing in browser (it's good stuff)
  - Some precident for collecting this data online (SPALEX: Aguasvivas et al., 2018)

Outcome 1: Online Portal

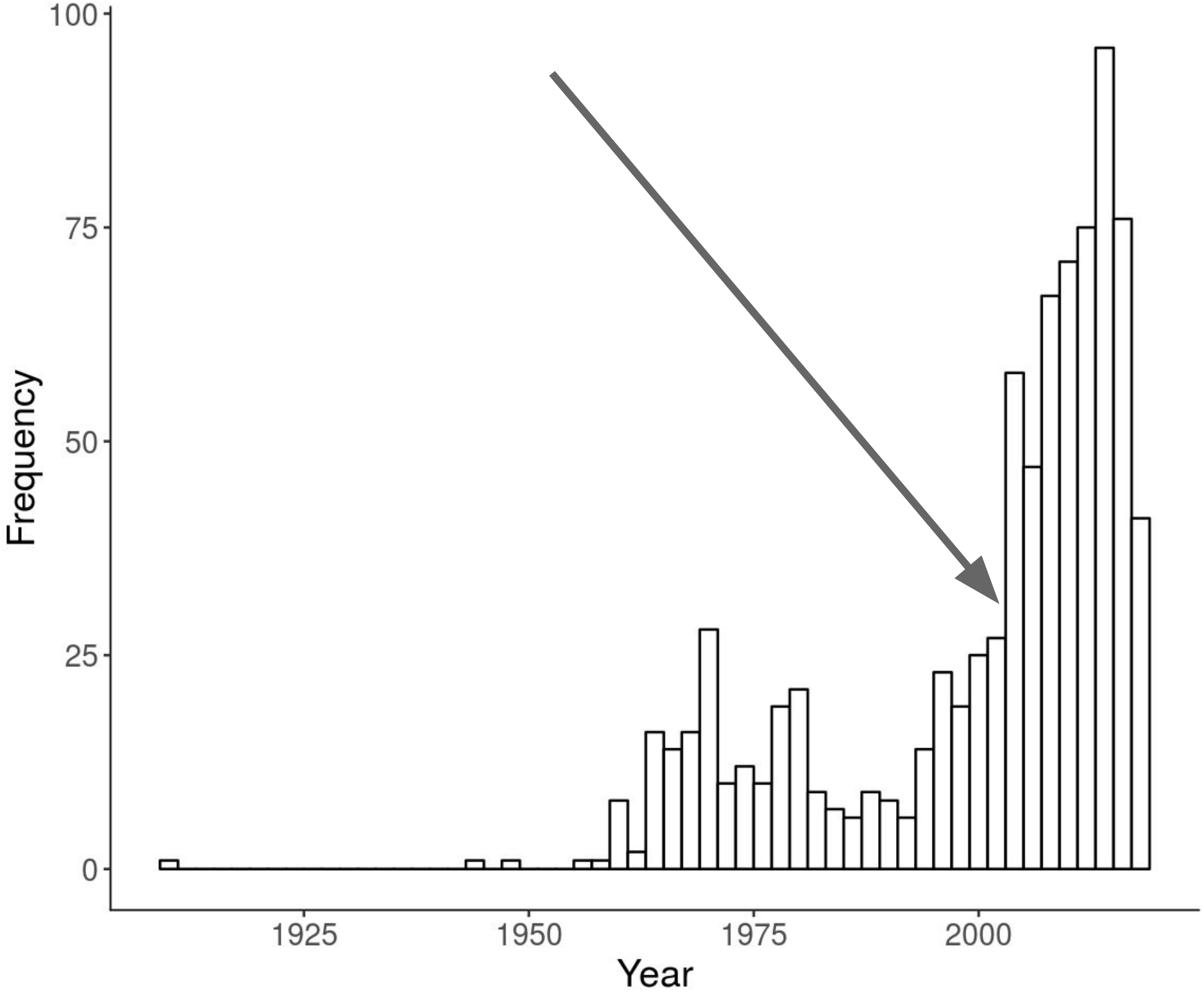
- Data is stored in a sqlite file, which can be accessed for the online display of data or through the packages (outcome 3)
- Labs can used specialized links
- Many languages can be provided for participants

Outcome 2: Loads O' Data

- Corpus Text Data
  - Subtitle Projects Analyzed (2 projects)
- Semantic Priming Data
  - Based on subtitle work above
- Subjective Rating data
  - Filling in the gaps from what is currently avaliable

Outcome 2: Loads O' Data

- Corpus Text Data: Open Subtitles Project
  - Freely available subtitles in ~60 languages for computational analysis
  - Approximately 51 languages contain enough data to be useable for these projects
  - BONUS: Translation pairs are included (translators rejoice!)
- *The Subtitle Projects have had a serious impact on our field.*


100
75
50
25
0
Frequency
1925
1950
1975
2000
Year

Outcome 2: Loads O' Data

- Corpus Text Data: Ongoing projects*
- Subs2strudel
  - Convert the subtitle data into concept-feature pairs
  - Example: zebra (concept) has stripes (feature)
  - STRUDEL: structured dimension extraction and labeling (Baroni et al., 2010)
  - Concept-feature pairs can be used to calculate similarity!

* Happy to have help! Let me go on vacation first, see you in October.

Outcome 2: Loads O' Data

- Corpus Text Data: Ongoing projects*
- Words2manylanguages
  - A recent publication of subs2vec, which converts the subtitle projects to FastText computational models
  - *Concerns are had*
  - Provide word2vec models of each subtitle language, which allows for similarity calculation

* Happy to have help! Let me go on vacation first, see you in October.

Outcome 2: Loads O' Data

- Semantic Priming Data
  - Related stimuli will be selected using similarity values from the first two analyses described
  - Unrelated stimuli are re-paired words with no similarity (close to zero as possible)
  - Nonsense words are created by changing one letter of the other stimuli, while mantaining valid phonetic pronunciation

Outcome 2: Loads O' Data

- Semantic Priming Data
    - The translations provided in the Open Subtitle Projects will be used to cross reference across languages
    - We hope to have approximately 1000 of the *same pairs* in languages with roughly *the same similarity.*

Outcome 2: Loads O' Data

- Semantic Priming Data
  - A single stream lexical decision task will be used
- Trials are formatted as:
  - A fixation cross (+) for 500 ms
  - CUE or TARGET in uppercase Serif font
  - Lexical decision response (word, nonsense word)
- Practice timing will determine number of trials (~400-600)

Outcome 2: Loads O' Data

- Semantic Priming Data
  - This procedure creates data at many levels
  - Item level: for each individual item, rather than just cue or just concept
  - Subject level: for every participant
  - Priming level: for each related pair compared to the unrelated pair
    - Nonsense words have a purpose!

Outcome 2: Loads O' Data

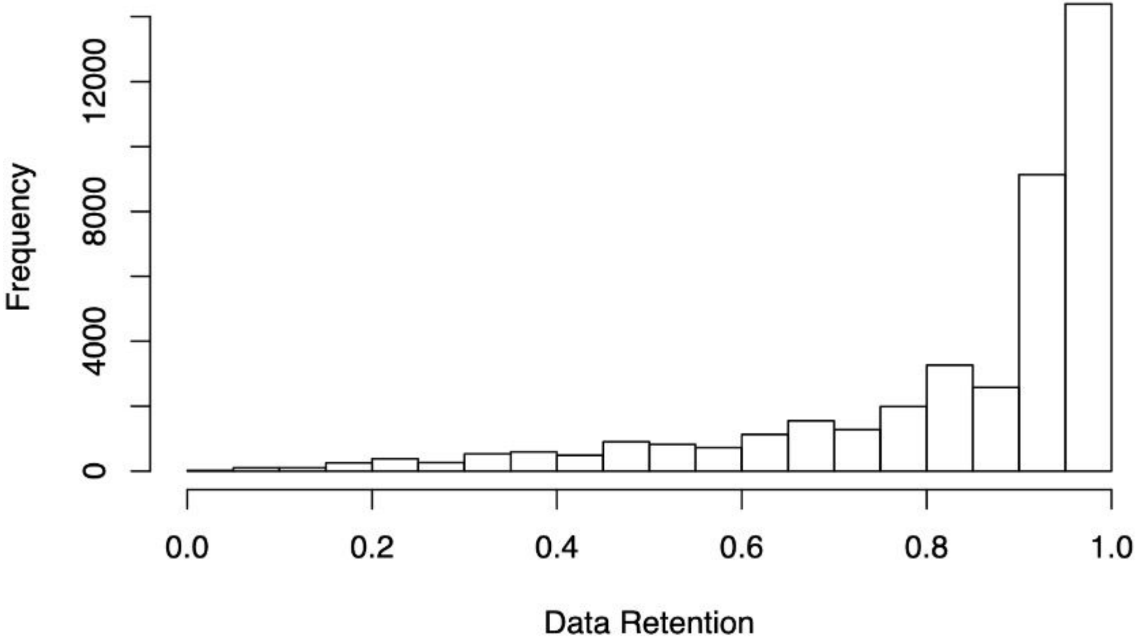
- Subjective Rating data
    - Filling in the gaps from what is currently avaliable
    - Ask participants to randomly complete one of these tasks based on what is needed.
    - Target variables: age of acquisition, imageability, concreteness, valence, arousal, dominance, familarity
    - These are the most studied and popular measures!

Outcome 2: Loads O' Participants

- Power for non-hypothesis tests is tricky
- AIPE: Accuracy in parameter estimation approach may be best (see anything by Ken Kelley)
  - Power to create a “sufficiently narrow” confidence interval
- So, we simulated using the English Lexicon Project (Balota et al., 2007) and the previous priming data

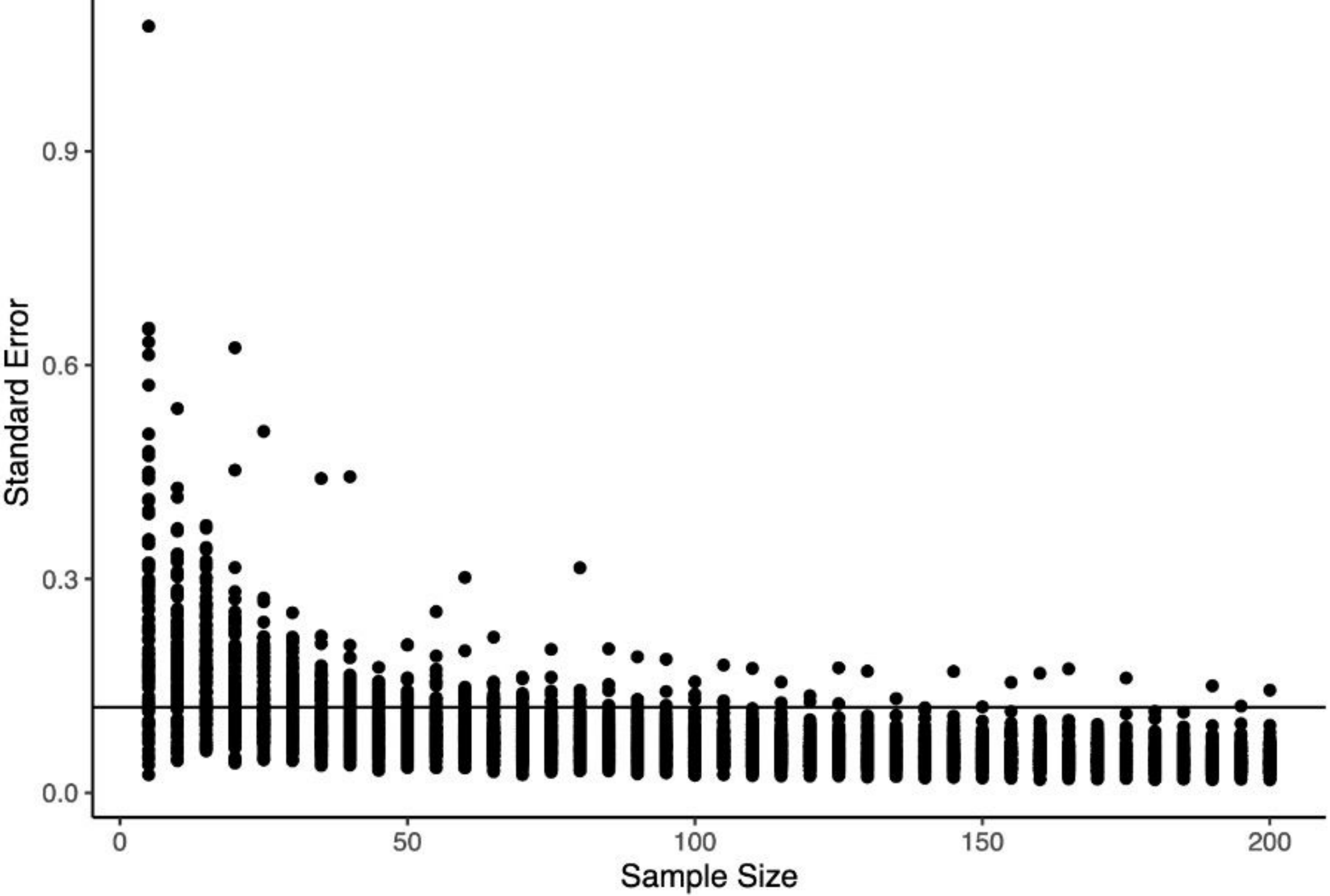
Outcome 2: Loads O' Participants

- Expect about 84% data retention (people get things wrong, which you can't use)



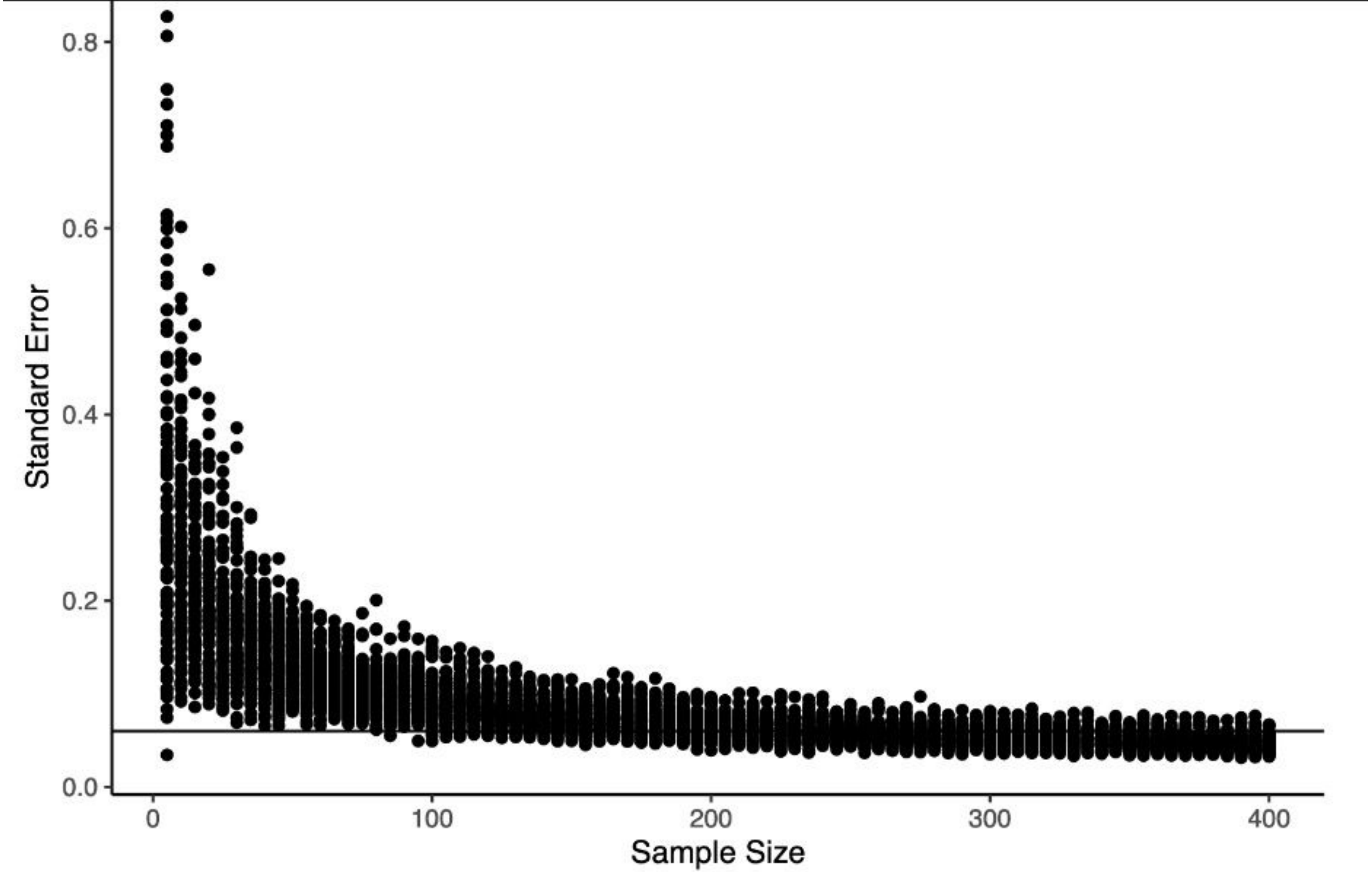
Outcome 2: Loads O' Participants

- Calculated the standard error for response latencies
- Randomally sampled from the data simulating *n* = 5, 10, … 200
- At what point is the standard error of 80% of the samples < our target standard error?


Standard Error
0.9
0.6
0.3
0.0
0
50
100
150
200
Sample Size

Outcome 2: Loads O' Participants

- *N* = 50 per word! Not so bad!
- Until you look at priming data …
  - Same procedure, this time with priming data
- Likely to pick some compromise of the two approaches


Standard Error
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Sample Size

Outcome 2: Loads O' Participants

- Therefore, we will use a minimum, stopping rule, and maximum sample (pre-registered)
  - Minimum number of participants per word = 50
  - Stopping rule = after 50, examine the SE until it reaches the desired “sufficiently narrow window”
  - Maximum number of participants = 320
  - *also a paper we are working on, if interested

Outcome 3: Data Access + Packages

- LexOPS is amazing!
  - Allows for stimuli selection and comparison
- We would try to convert to Python and supplement LexOPS with functions for acquiring/importing the data from this project.
- All the other data collected as well

Outcome 4: Secondary Data Challenge

- We will support ($) a secondary data challenge timed with the release of the first round of data.
  - Computational linguistics rejoice!

Check it Out

- I have learned a lot of new code tricks (and Python) since I wrote this proposal but you can check out all the background code, math, and ideas at:
- https://github.com/SemanticPriming/SPAML/

Questions

- All thoughts welcome!

- Twitter: @aggieerin
- Email: buchananlab@gmail.com
- GitHub: doomlab
- Find me on the PSA Slack