Computational methods + sociolinguistic perspectives
Making the case for computational sociolinguistics

Dan Villarreal

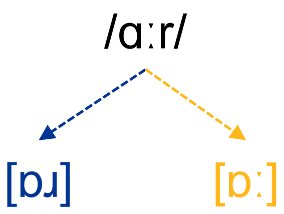
February 11, 2020

FYI: *This presentation is my 2020 job talk for a computational sociolinguistics position at the University of Pittsburgh. Rather than the usual job-talk format, I was asked to spend about half of the talk "outlining your vision for the new subfield of computational sociolinguistics more generally (e.g., what is computational sociolinguistics to you, and how do your current and future projects fit in to it?)"*

Sociolinguistics: Finding patterns in complexity

- As a field, sociolinguistics has always been about finding patterns in complexity

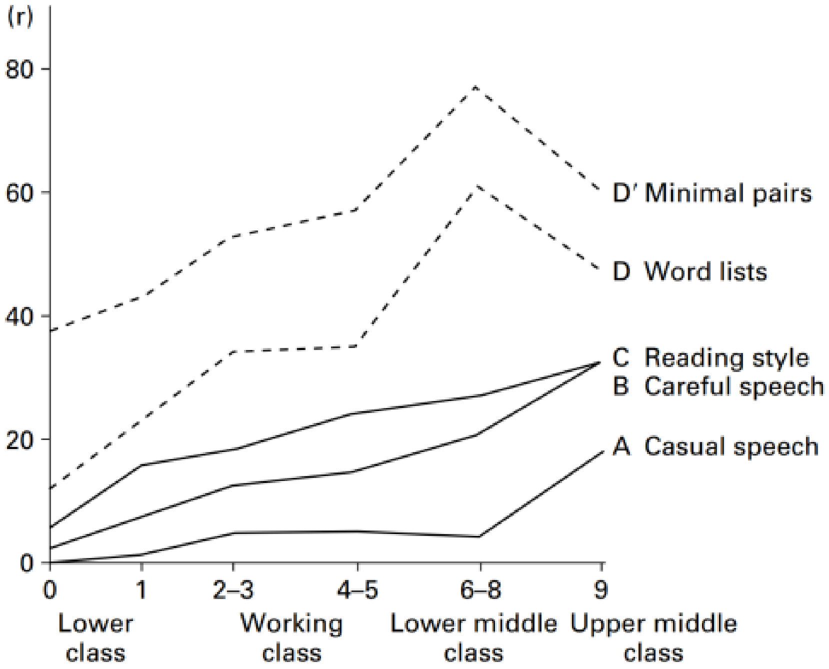
free variation



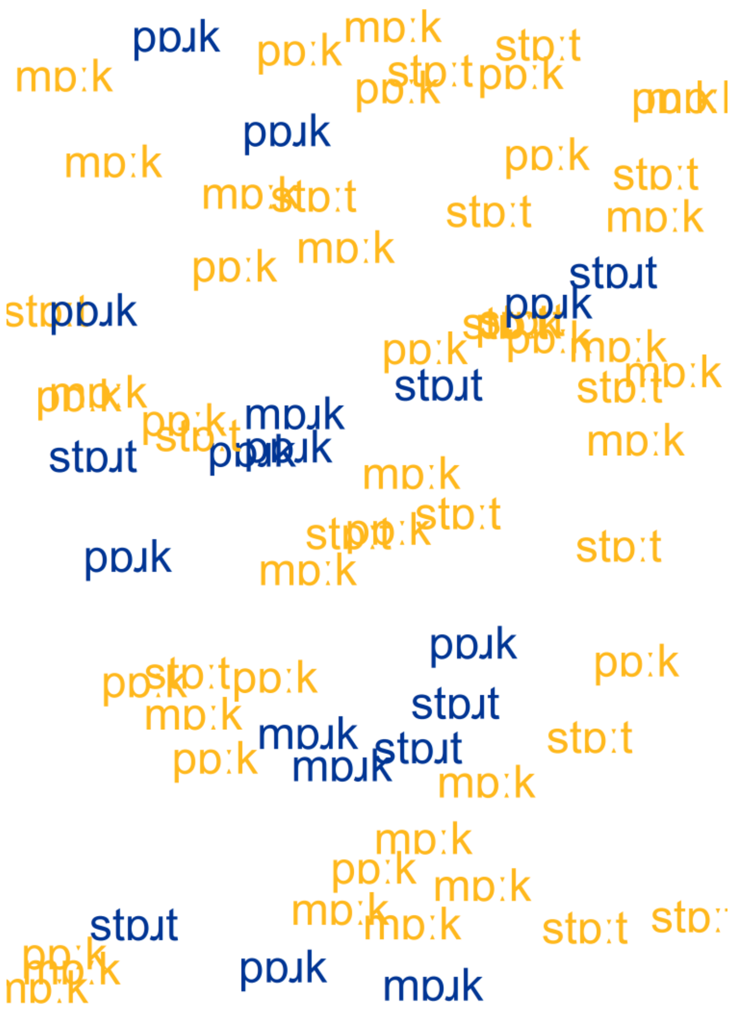
/ɑːr/

[ɒɹ]

[ɒː]

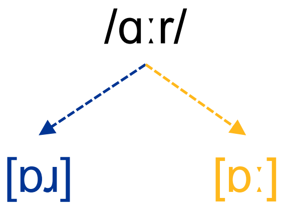


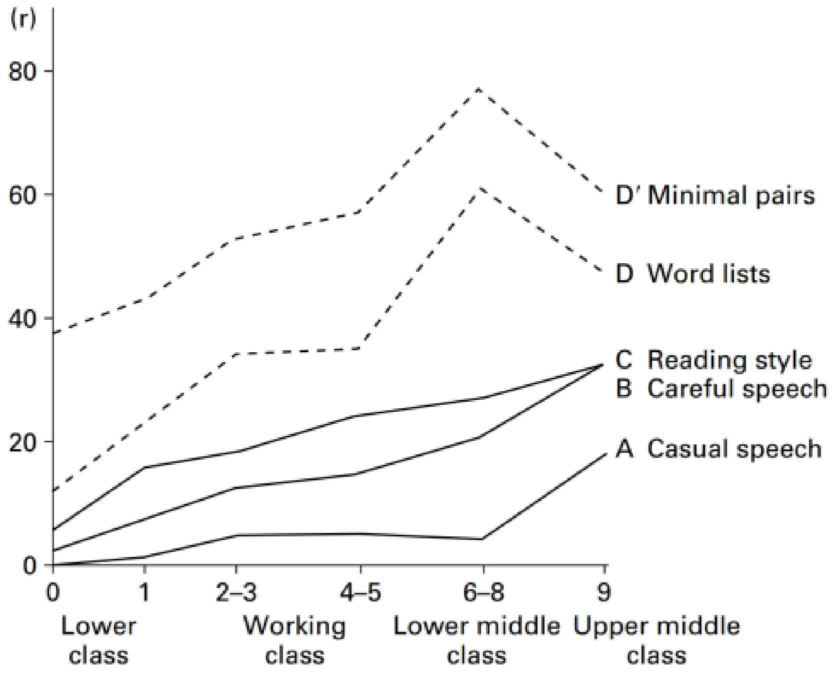
(Labov 2006/1966:151)



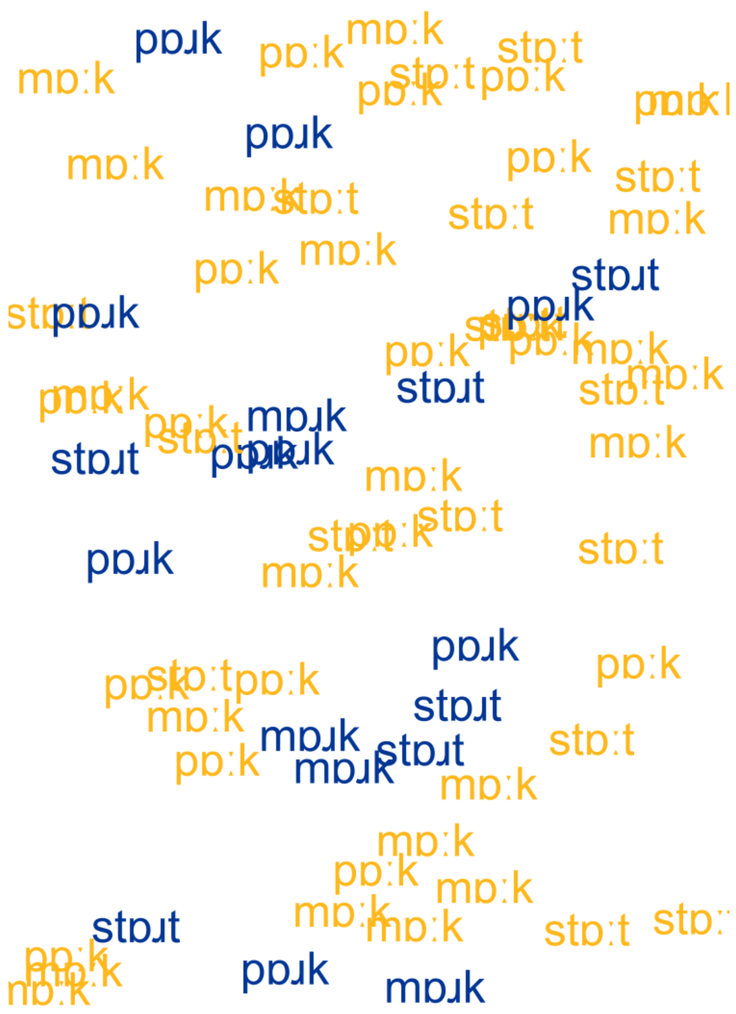
Sociolinguistics: Finding patterns in complexity

- As a field, sociolinguistics has always been about finding patterns in complexity
  - Ex: From *free variation* to *orderly heterogeneity*



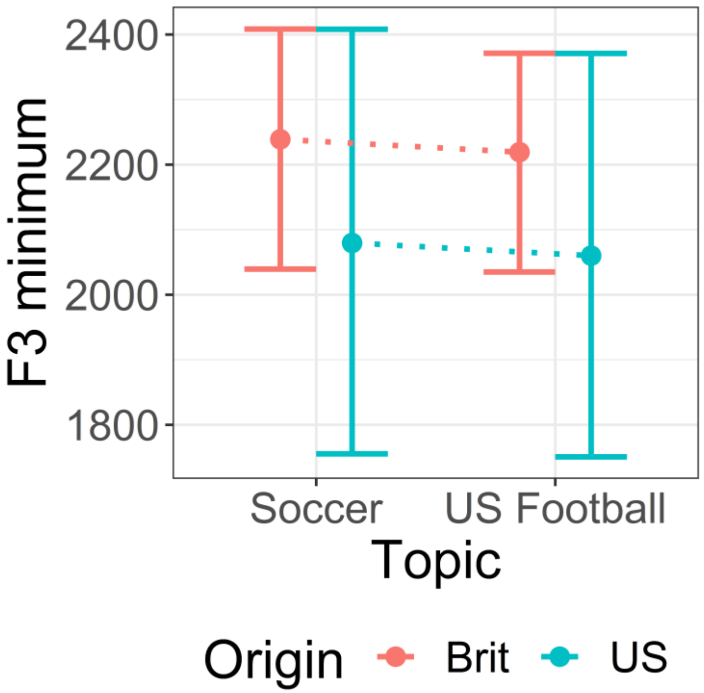


(Labov 2006/1966:151)



Sociolinguistics: Finding patterns in complexity

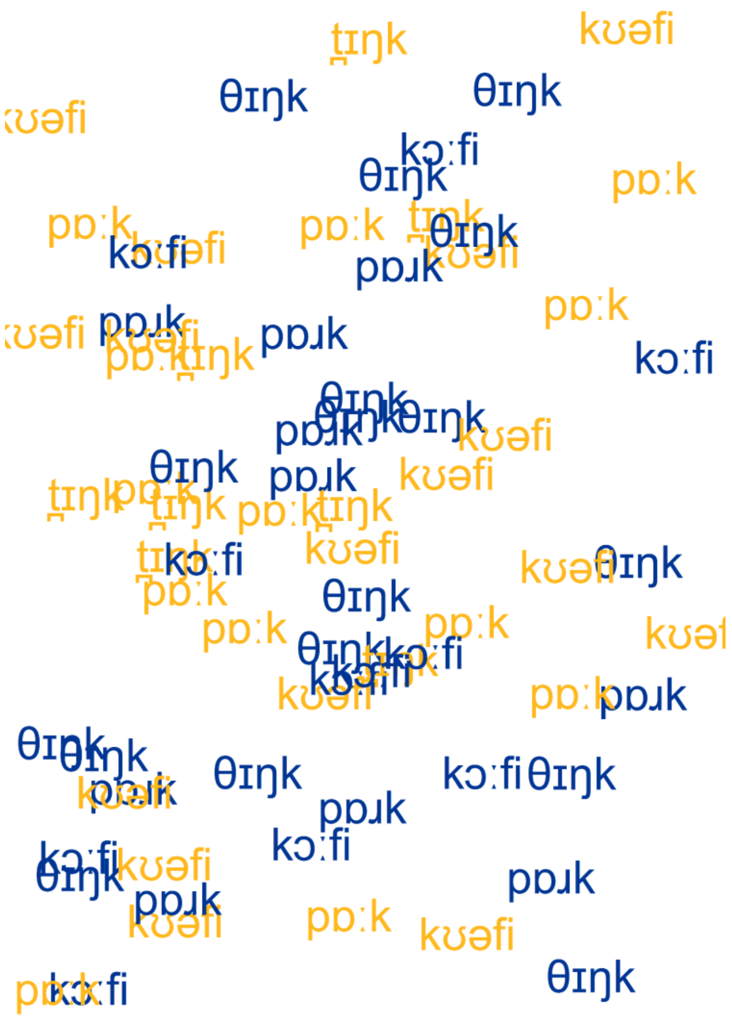
- As a field, sociolinguistics has always been about finding patterns in complexity
  - Ex: From *free variation* to *orderly heterogeneity*
- But as the field continues to develop, we're increasingly asking research questions that deal with a new degree of complexity
  - Ex: Given that many socially-meaningful sociophonetic variables exist along a phonetic gradient, how does fine-grained phonetic variation convey social meaning?



(Love & Walker 2013)

Sociolinguistics: Finding patterns in complexity

- As a field, sociolinguistics has always been about finding patterns in complexity
  - Ex: From *free variation* to *orderly heterogeneity*
- But as the field continues to develop, we're increasingly asking research questions that deal with a new degree of complexity
  - Ex: Given that many socially-meaningful sociophonetic variables exist along a phonetic gradient, how does fine-grained phonetic variation convey social meaning?
  - Ex: Given the multiplicity of sociolinguistic variables that occur in even short stretches of running speech, how do listeners pull out social meaning?



Making the case for computational sociolinguistics

- Among the benefits of computational methods is their ability to sift through complex data
- So a computational approach to sociolinguistics brings together the best of both worlds
- Today, I want to share my vision for a **computational sociolinguistics** that brings together computational methods and sociolinguistic perspectives

Outline

- What is computational sociolinguistics?
- A case study: “From categories to gradience: Auto-coding sociophonetic variation with random forests”
  - Background
  - Classifier training
  - Listening experiment
  - Discussion
- My broader agenda within computational sociolinguistics

What is computational sociolinguistics?

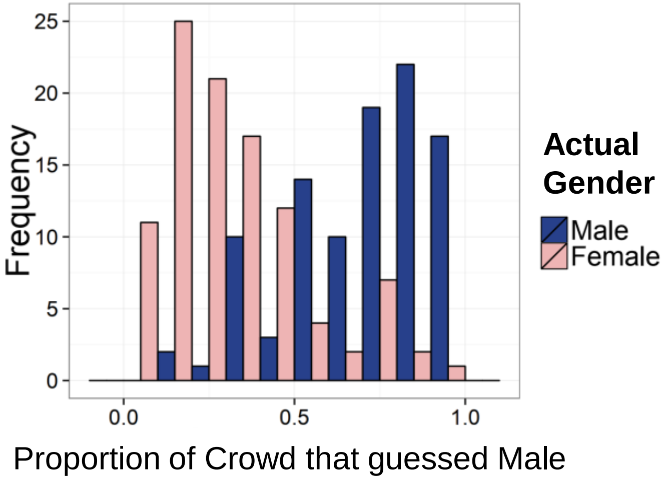
Computational methods + sociolinguistic perspectives

- Simply put, computational sociolinguistics is the marrying of computational methods with sociolinguistic perspectives
- Computational sociolinguistic research exists on a spectrum between more computational-sited research & more sociolinguistics-sited research
  - But the computational and the sociolinguistic goals are closely intertwined

Computational questions, sociolinguistic questions

- Nguyen et al. (2014a, b): TweetGenie predicted gender & age of Twitter users based on words in tweets
- Comp ling Q: Does this work? How well?
  - Trained regression models on unigrams & bigrams
  - Participants guessed gender of anonymized tweets
- Socioling Q: Which words are used more by women?
  - Women: *my man*, *bye*, *omg*, *mom*, *sweet*
  - Men: *man*, *bro*, [name of soccer team], *fifa*, *beer*
- Despite the use of gender and age as static variables in comp ling, Twitter users emphasize these aspects of identity in varying ways

		Crowd guesses	
		Male	Female
Auto-coding	Male	68	22
	Female	30	80



(Nguyen et al. 2014a:1955)

Sociolinguistics, computational-style

- The use of computational methods to answer sociolinguistic research questions manifests in three ways:
- Answering pre-existing sociolinguistic RQs **faster**
  - Forced phonetic aligners (Rosenfelder et al. 2011)
- Answering pre-existing sociolinguistic RQs **better**
  - Isogloss-drawing algorithms (Nerbonne 2009)
- Answering sociolinguistic RQs **that weren’t previously possible**
  - Lexical diffusion on Twitter (Grieve et al. 2018)

Computational linguistics, sociolinguistics-style

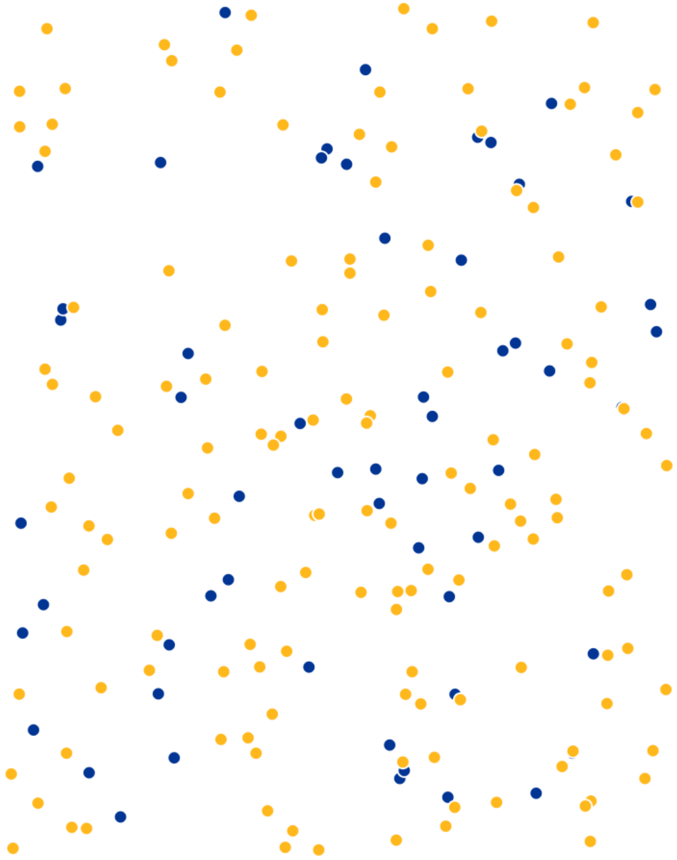
- Sociolinguistic perspectives can enrich computational approaches
- NORMs (non-mobile, older, rural, males) in dialectology analogous to WEIRDs (Western, educated, industrialized, rich, democratic) today
  - NLP APIs (e.g., Microsoft Azure, IBM Watson) classify AAL tweets as non-English to a higher degree than GenAm tweets (Blodgett et al. 2017)
  - YouTube ASR captions most accurate for white GenAm speakers (Tatman & Kasten 2017)
  - Barth et al. (2020) developed an iterative technique for building an acoustic model for Matukar Panau from scratch
- Moving beyond a conception of identity as fixed & immutable (Bamman et al. 2014; Nguyen et al. 2014b)

Meeting in the middle

- Making resources available (open data, extensive how-tos, open access)
  - Increase access to communities beyond WEIRDs
- Cyclical relationship between theory and methods
- Adjusting expectations
  - Theory-motivated predictors vs. data dredging
  - Skepticism and the validation of algorithms
- Converging on standard practices that fulfill the needs of both computational & sociolinguistic approaches

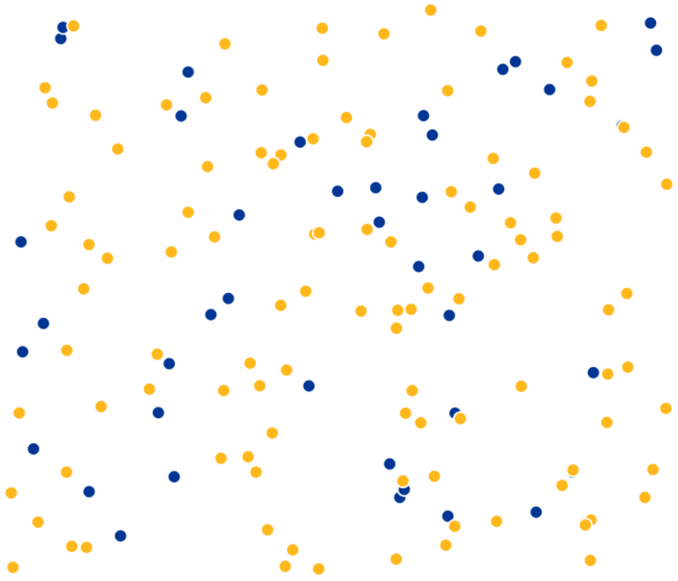
Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag



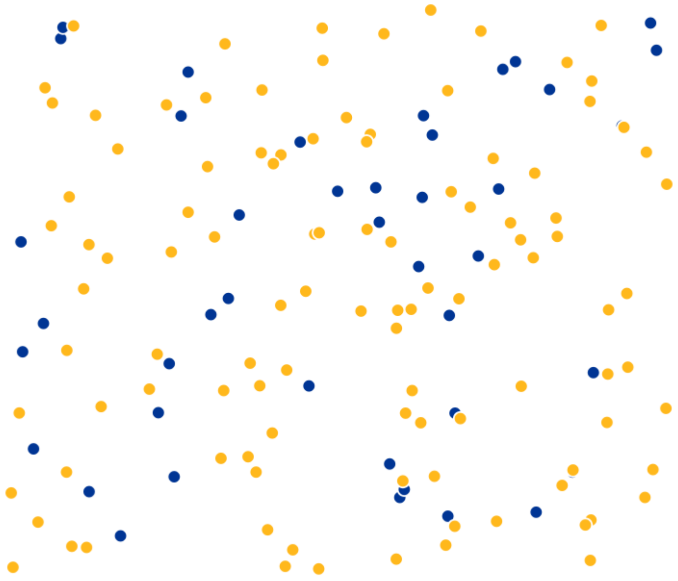
Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag



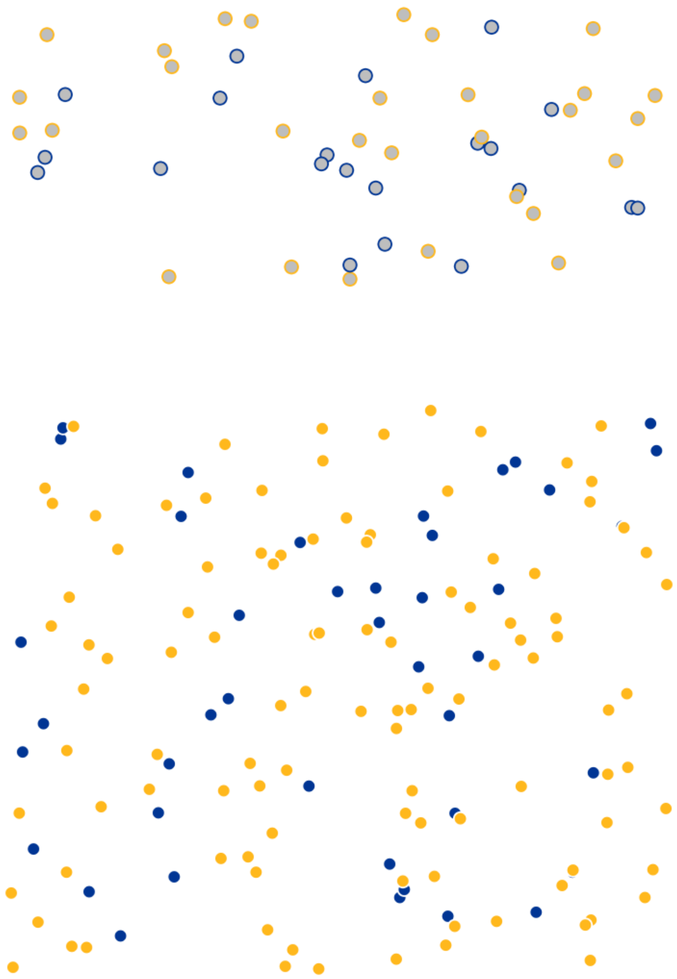
Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag



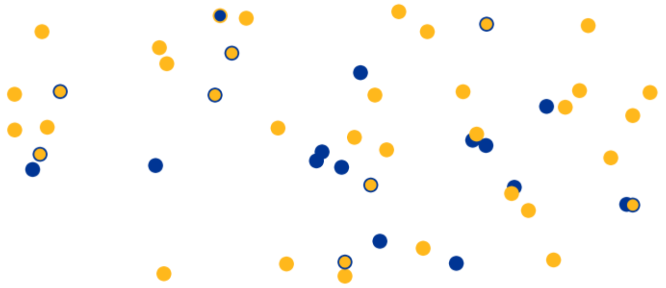
Converging on standard practices

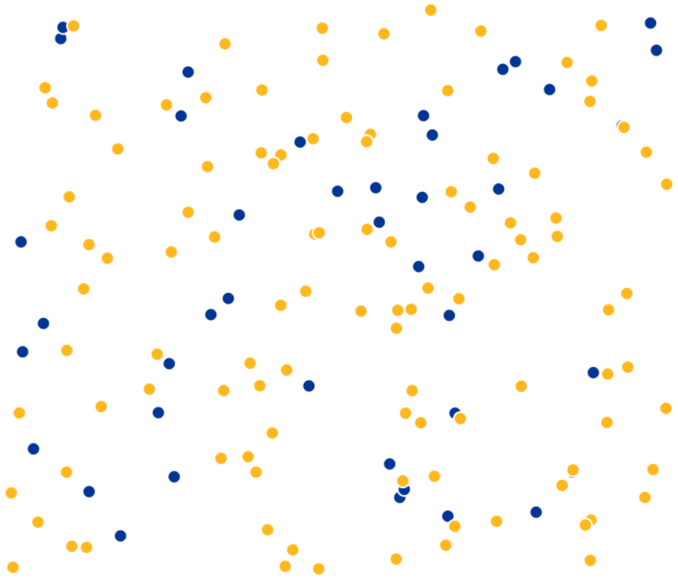
- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag



Converging on standard practices

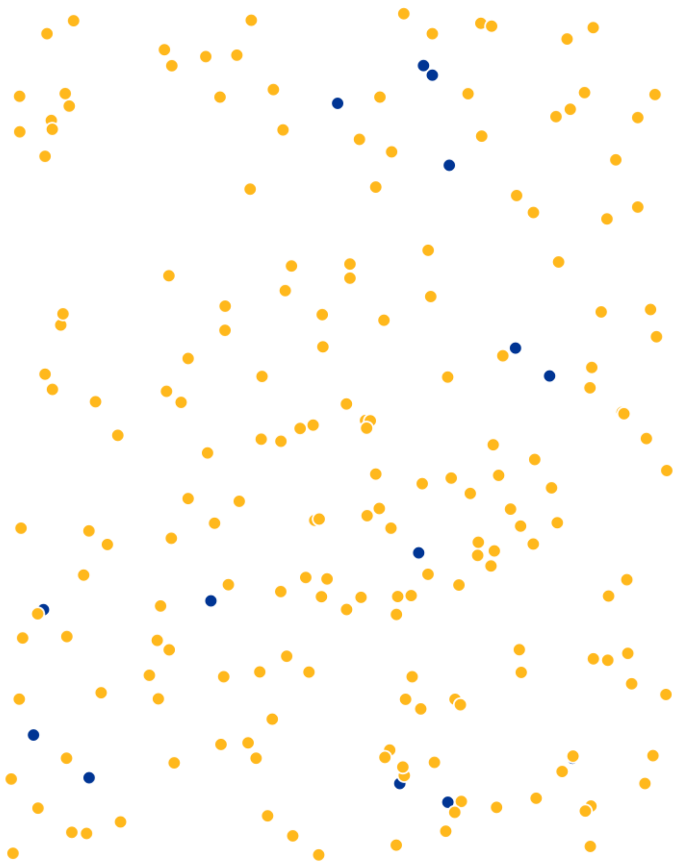
- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag





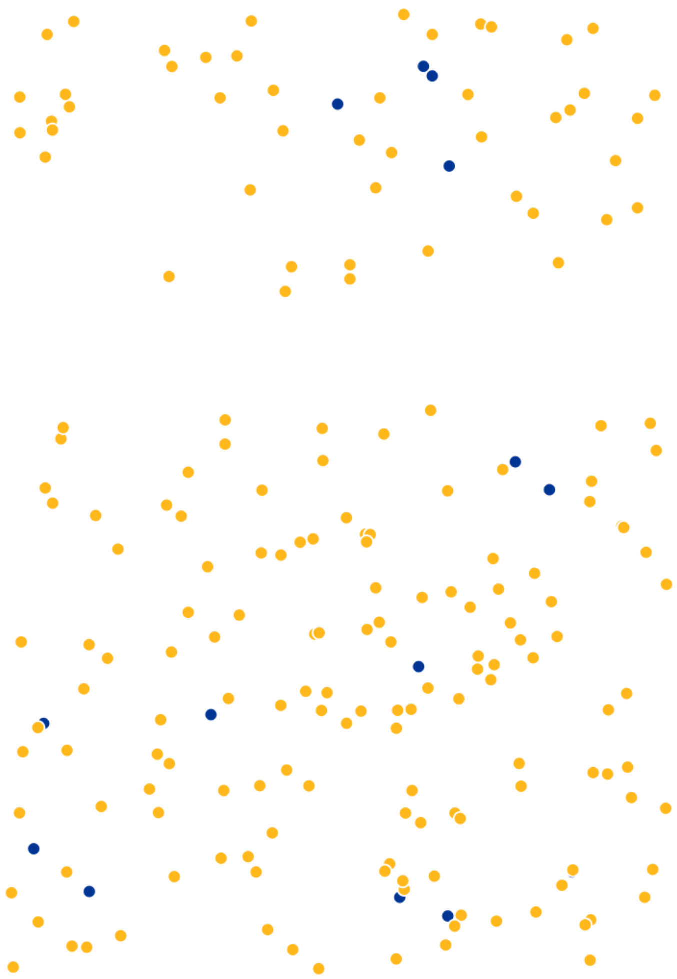
Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes



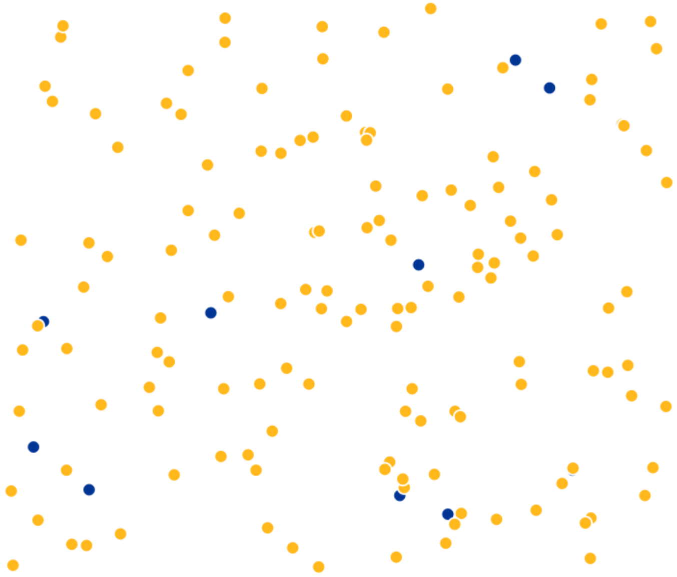
Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes



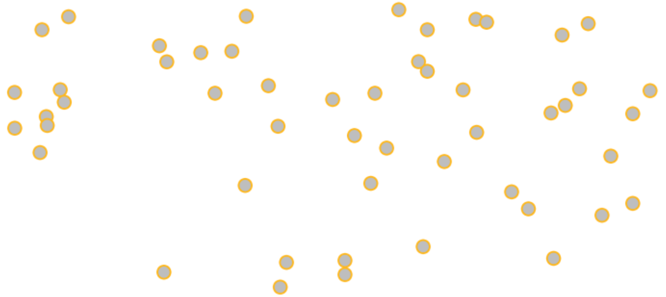
Converging on standard practices

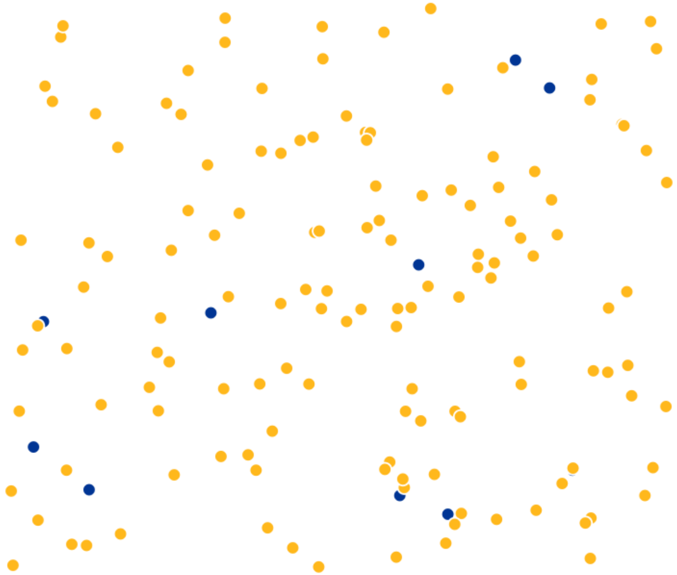
- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes



Converging on standard practices

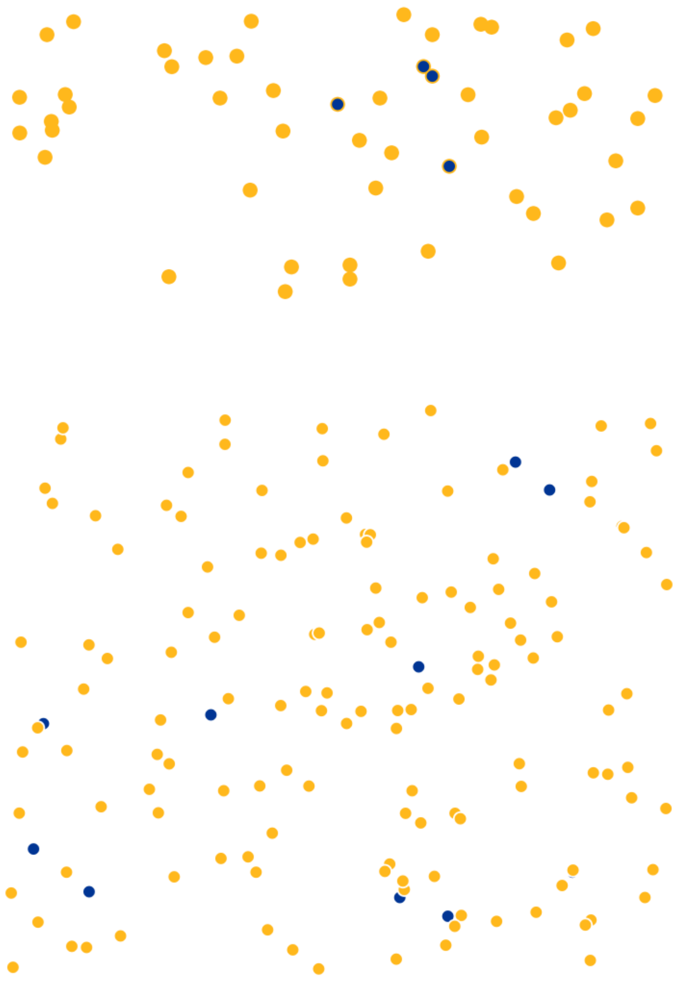
- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes





Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes



Converging on standard practices

- Ex: Performance metrics for evaluating classifiers
- In supervised ML applications (where classes are known), performance is typically assessed on observations that are held out-of-bag
  - But simply maximizing accuracy can lead to poor results if the classifier fails to balance the two classes
- F1 score sensitive to the choice of ‘positive’ class
- In the case study: Accuracy × AUC balances accuracy and representativeness
  - AUC quantifies a classifier’s ability to balance true positives and true negatives

A case study: “From categories to gradience: Auto-coding sociophonetic variation with random forests”

Villarreal, Dan, Lynn Clark, Jennifer Hay and Kevin Watson. revisions under review. From categories to gradience: Auto-coding sociophonetic variation with random forests. Revisions re-submitted to *Laboratory Phonology*

Background

- RQ: Can machine learning methods effectively automate the coding of sociophonetic variables?
- Complicating properties of sociophonetic variables:
  - Not clear what characterizes /r/ acoustically (Heselwood 2009; Lawson et al. 2014, 2018; Stuart-Smith 2007; Zhou et al. 2008)
  - Acoustic correlates exist on a gradient (e.g., Love & Walker 2013; Temple 2014)
  - /r/ (and liquids generally) notorious for low inter-coder reliability (Fosler-Lussier et al. 2007; Lawson et al. 2014; Pitt et al. 2005; Yaeger-Dror et al. 2009)
- If the method works, we get extra benefits beyond auto-coding: cue weighting, gradient predictions
- Classifiers on 2 NZ English variables: /r/ & intervocalic /t/

Training data

- /r/ data came from Southland New Zealand English corpus of sociolinguistic interviews (Bartlett 2002)
  - 4,689 tokens (28 speakers) hand-coded into 2 classes:
    Present 27.8%; Absent 72.2%
  - 180 acoustic measures: formants & formant differences, formant bandwidths, pitch, amplitude, duration
- /t/ data came from QuakeBox corpus of New Zealand English monologues (Clark et al. 2016)
  - 4,218 tokens (168 speakers) hand-coded into 6 classes:
    [ɾ] 39.2%; [s] 39.0%; [d] 9.5%; [t] 7.2%; [ʔ] 3.3%; [∅] 1.8%
  - Later collapsed to 2 classes: Voiced 50.5%; Voiceless 49.5%
  - 113 acoustic measures: amplitude, center of gravity, dispersion, kurtosis, periodicity, spectral tilt, formant bandwidths, formant transitions, duration

Running the classifiers

- Classifiers run as random forests in R with `ranger` and `caret` (Kuhn 2017; R Core Team 2018; Wright 2018)
  - Random forests (ensembles of CART trees) reduce variance while avoiding overfitting (Breiman 2001)
  - Why RF? RF don’t overfit when predictors are collinear (Matsuki et al. 2016; Strobl & Zeileis 2008)
  - Try it at home! is.gd/ClassifieR
- Trained and tuned 3 different classifiers
  - /r/: Absent vs. Present
  - 6-way /t/: [ɾ s d t ʔ ∅]
  - 2-way /t/: Voiceless ([s t ʔ]) vs. Voiced ([ɾ d ∅])
- Assessed performance via cross-validation using accuracy × AUC

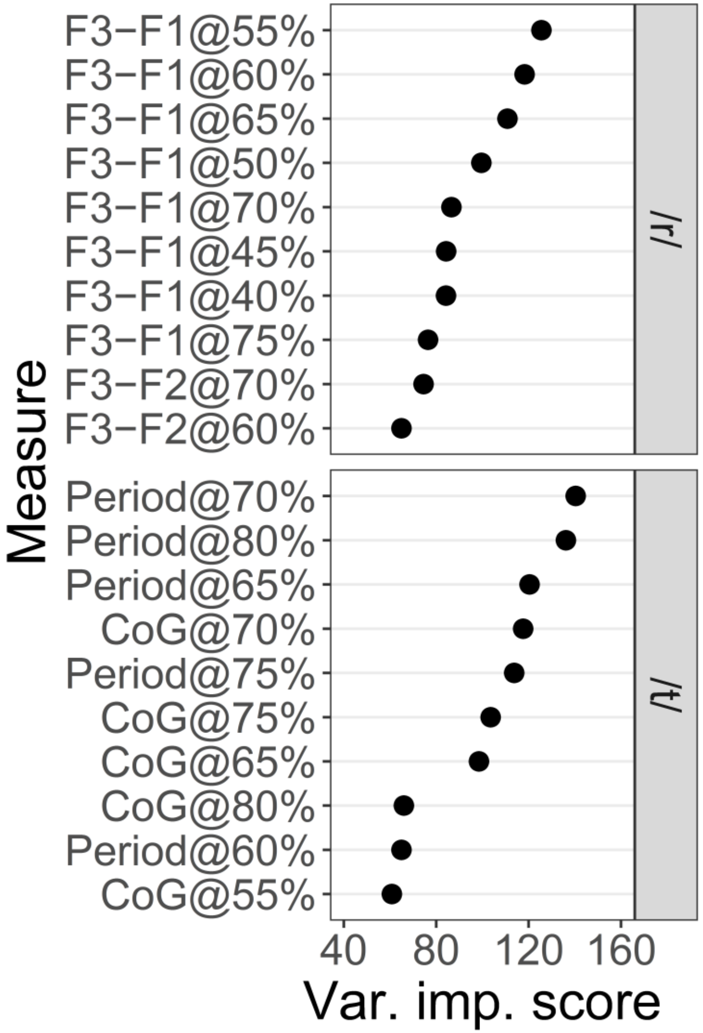
Evaluating the classifiers

Performance measure	/r/
Overall accuracy × AUC	.7423
Overall accuracy	.8447
AUC	.8786
Class accuracies	Absent 93.1% Present 62.2%

- Overall accuracy for /r/ classifier compares favorably with human inter-coder reliability for rhoticity; but more accuracy for Absent (majority)
- 6-way /t/ classifier had high accuracy for majority classes but very poor accuracy for minority classes
- 2-way /t/ classifier had excellent, balanced performance; but failed to preserve sociolinguistically salient distinctions between [ɾ s t]

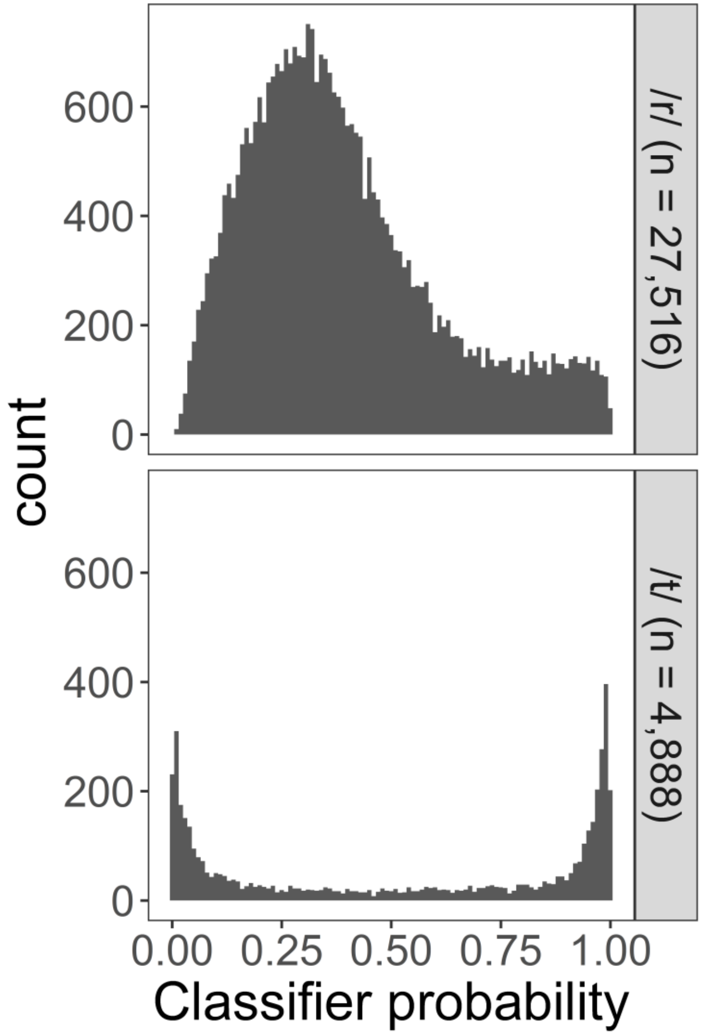
Acoustic cue weighting

- What characterizes /r/ acoustically?
  - Sociophonetic studies have used F3 minimum (e.g., Hay & Maclagan 2010, 2012); acoustic studies have suggested other cues (e.g., Heselwood 2009)
- /r/ classifier revealed most important measures were F3−F1 toward the end of the token
  - F3 minimum 14th out of 180 measures
- Not surprisingly, periodicity and center of gravity highly important in 2-way /t/ classifier



Gradient predictions

- Gradient predictions (*classifier probabilities*): how Present-like is each /r/ token? how Voiced-like is each /t/ token?
    - Categorical variables’ classifier probabilities should distribute bimodally
- Auto-coded tokens: bimodal distribution for /t/ and unimodal distribution for /r/
- Suggests that variation in /r/ isn’t so categorical after all, with a large gray area between Absent/Present

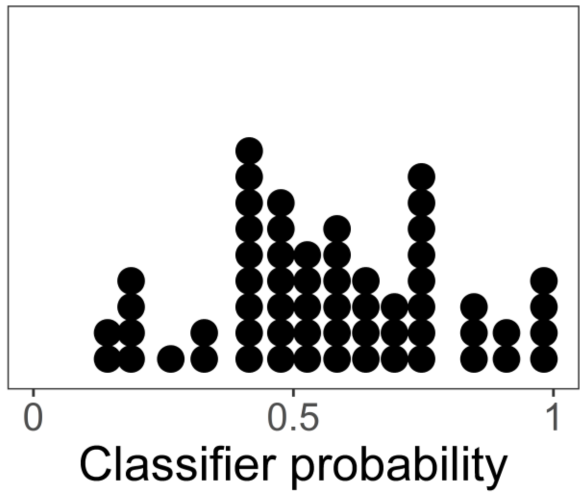


Listening experiment: Overview

- Typically, cross-validated performance results are considered sufficient to evaluate a classifier
- But to be extra sure, I decided to compare the classifier's test-set codes to human judgments
- In particular, I was curious how humans would deal with the tokens the classifier identified as being in the gray area between Absent and Present: Is the classifier merely uncertain about these tokens or do they reflect phonetic gradience?

Listening experiment: Methods

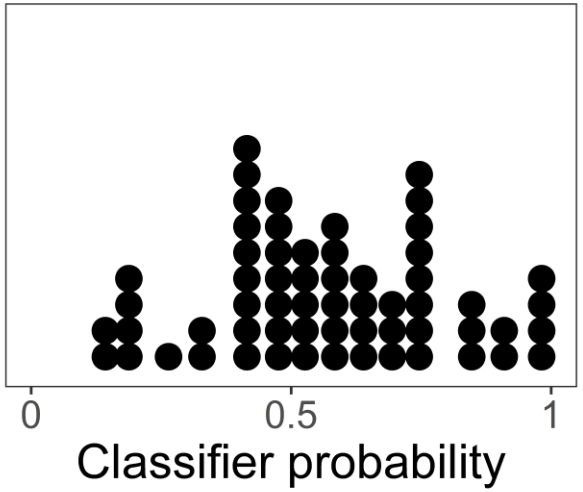
- 60 stimuli from auto-coded data
  - Male speakers, preceding NURSE vowel, following /l m n/, stressed monosyllables
  - Middle classifier probabilities oversampled
  - Isolated words, repeated w/ 750ms buffer



Stimulus	*turn*	*turned*	*girls*
Classifier probability	0.1755185	0.5010963	0.9833926

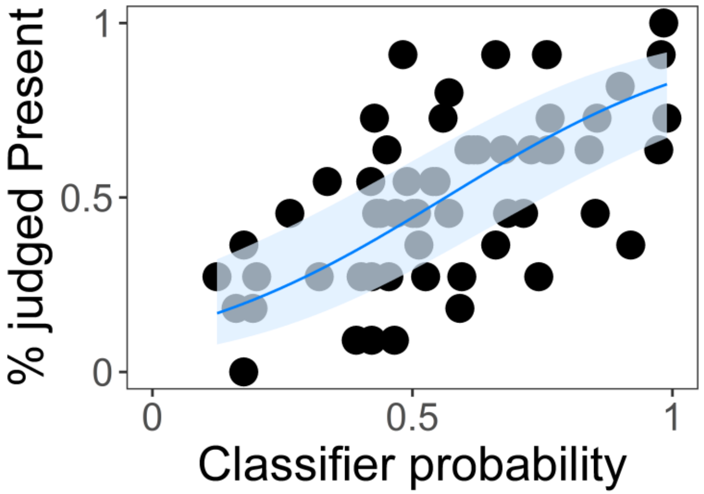
Listening experiment: Methods

- 60 stimuli from auto-coded data
  - Male speakers, preceding NURSE vowel, following /l m n/, stressed monosyllables
  - Middle classifier probabilities oversampled
  - Isolated words, repeated w/ 750ms buffer
- 11 phonetically trained linguists judged each stimulus as Present or Absent
  - Highly proficient English users residing in NZ with exposure to both rhotic and non-rhotic accents
  - 7/11 were L1 English speakers; 5/11 rhotic speakers (neither factor significantly affected judgments)
  - Listened with headphones on individual computers



Listening experiment: Results

- Substantial variability in judgments
  - Only 2 stimuli had total agreement
  - Typically, we'd look at these results and say there's insufficient inter-coder agreement (a bug in the system)
- Positive effect of classifier probability on listener judgments of rhoticity (*p* < .0001)
- Classifier probability successfully captured gradience in /r/ beyond the Absent/Present binary
  - Suggests inter-coder disagreement is feature, not bug
  - Lends weight to interpretation of classifier probability as representing phonetic gradience



Listening experiment: Results

- Since F3 minimum has been used as a gradient metric of rhoticity (Hashimoto 2019; Hay & Clendon 2012; Hay & Maclagan 2010, 2012; Love & Walker 2013), why bother to use classifier probability?
- Just like classifier probability, F3 minimum also significantly predicts listener judgments (p < .001)
  - Model with F3 minimum *and* classifier probability significantly improves on model with just F3 minimum (p < .001)
- Classifier probability, as a composite of numerous individual cues, captured something about listener judgments above and beyond what individual measures (like F3 minimum) can capture

Discussion

- This method promises to expand sociolinguists’ capacity to answer sociolinguistic questions
  - Benefits aren’t limited to auto-coding—something that humans can do but is tedious/time-consuming—but extend to cue-weighting and the generation of gradient predictions
  - Future work is necessary to determine broader applicability
- Suggestion of gradience in /r/ (but not /t/) realizations dovetails with growing evidence questioning the categoricity of supposedly categorical variables:
  - /r/ studies using F3 minimum (Hashimoto 2019; Hay & Clendon 2012; Hay & Maclagan 2010, 2012; Love & Walker 2013)
  - English coronal stop deletion, based on both acoustic (Temple 2014) and articulatory (Purse 2019) approaches

My broader agenda within computational sociolinguistics

Computational methods + sociolinguistic research questions

- My research agenda covers all of the ways that comp methods can aid the pursuit of sociolinguistic RQs:
- Answering pre-existing sociolinguistic RQs **faster**
  - Auto-coding yielded a finding that challenges sociolinguistic theory: Southland women & men had different internal constraints on /r/ variation
- Answering pre-existing sociolinguistic RQs **better**
  - Classifier probability captured gradience in listener judgments of rhoticity above and beyond F3 minimum
- Answering sociolinguistic RQs **that weren't previously possible**
  - Future direction: Applying cue weighting to address the 'multiplicity of cues to social meaning' problem

Bringing computational sociolinguistics into being

- I work toward the strategies that will help computational sociolinguistics emerge as a field
- Making resources available
  - Open data: github.com/nzilbb/Sld-R-Data
  - Auto-coder training how-to: is.gd/ClassifieR
- Cyclical relationship between theory and methods
  - RQs about Southland /r/ yielded classifier method, which opened up questions about categoricity vs. gradience
- Moving toward standard practices
  - Accuracy × AUC as a performance metric that balances accuracy and representativeness (rather than F1 or others)

Making the case for computational sociolinguistics

- The pursuit of sociolinguistic research questions stands to benefit from computational methods
- Computational methods stand to benefit from a sociolinguistic approach to variation
- By marrying computational methods with sociolinguistic perspectives (and by developing shared practices), computational sociolinguistics stands to become a valuable field in its own right

Thanks! Questions?

Thanks to Bob Bayley, Vai Ramanathan, Lynn Clark, Jen Hay, Kevin Watson, Vica Papp, Mary Kohn, James Brand, Jacq Jones, Marie Fournier, Simon Todd, Donald Derrick, Jonathan Dunn, Christina Calvillo

Dan Villarreal

daniel.j.villarreal@gmail.com

References

Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135-160.

Barth, Danielle, James Grama, Simon Gonzalez and Catherine E. Travis. 2020. Using forced alignment for sociophonetic research on a minority language. *Penn Working Papers in Linguistics* 25(2). 2.

Bartlett, Christopher. 2002. The Southland Variety of New Zealand English: Postvocalic /r/ and the BATH vowel: University of Otago. Unpublished PhD thesis.

Blodgett, Su Lin and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. Paper presented to 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2017.

Breiman, Leo. 2001. Random forests. *Machine learning* 45(1). 5-32.

Clark, Lynn, Helen MacGougan, Jennifer Hay and Liam Walsh. 2016. “Kia ora. This is my earthquake story”. Multiple applications of a sociolinguistic corpus. *Ampersand* 3. 13-20.

Fosler-Lussier, Eric, Laura Dilley, Na'im R. Tyson and Mark A. Pitt. 2007. The Buckeye Corpus of Speech: Updates and enhancements. Paper presented to Interspeech 8, Antwerp, 2007.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4). 293-319.

Hashimoto, Daiki. 2019. Loanword phonology in New Zealand English: Exemplar activation and message predictability. Christchurch, New Zealand: University of Canterbury. PhD thesis.

Hay, Jennifer and Alhana Clendon. 2012. (Non)rhoticity: Lessons from New Zealand English. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the history of English*, 761-772. Oxford: Oxford University Press.

Hay, Jennifer and Margaret Maclagan. 2010. Social and phonetic conditioners on the frequency and degree of 'intrusive /r/' in New Zealand English. In Dennis R. Preston & Nancy A. Niedzielski (eds.), *A reader in sociophonetics*, 41-69. New York: De Gruyter Mouton.

Hay, Jennifer and Margaret Maclagan. 2012. /r/-sandhi in early 20th century New Zealand English. *Linguistics* 50(4). 745-763.

Heselwood, Barry. 2009. Rhoticity without F3: Lowpass filtering and the perception of rhoticity in 'NORTH/FORCE,' 'START,' and 'NURSE' words. *Leeds Working Papers in Linguistics and Phonetics* 14. 49-64.

Kuhn, Max. 2018. caret [R package], vers. 6.0-81. https://CRAN.R-project.org/package=caret

Labov, William. 2006/1966. *The social stratification of English in New York City*, 2nd edition. Cambridge: Cambridge University Press.

Lawson, Eleanor, James Scobbie and Jane Stuart-Smith. 2014. A socio-articulatory study of Scottish rhoticity. In Robert Lawson (ed.), *Sociolinguistics in Scotland*, 53-78. London: Palgrave Macmillan.

Lawson, Eleanor, Jane Stuart-Smith and James Scobbie. 2018. The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *Journal of the Acoustical Society of America* 143(3). 1646-1657.

Love, Jessica and Abby Walker. 2013. Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech* 56(4). 443-460.

Matsuki, Kazunaga, Victor Kuperman and Julie A. Van Dyke. 2016. The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading* 20(1). 20-33.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175-198.

Nguyen, Dong, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder and Franciska De Jong. 2014a. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. Paper presented to COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, 2014.

Nguyen, Dong, Dolf Trieschnigg and Theo Meder. 2014b. TweetGenie: Development, evaluation, and lessons learned. Paper presented to Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, Dublin, 2014.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling and William Raymond. 2005. The Buckeye Corpus of Conversational Speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89-95.

Purse, Ruaridh. 2019. The articulatory reality of coronal stop "deletion". Paper presented to 19th International Congress of Phonetic Sciences, Melbourne, 2019.

R Core Team. 2018. R: A language and environment for statistical computing [vers. 3.5.2]. https://www.R-project.org/

Rosenfelder, Ingrid, Joe Fruehwald, Keelan Evanini and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) program suite [forced phonetic aligner]. http://fave.ling.upenn.edu/

Strobl, Carolin and Achim Zeileis. 2008. Danger: High power! Exploring the statistical properties of a test for random forest variable importance. Paper presented to 18th International Conference on Computational Statistics, Porto, Portugal, 2008.

Stuart-Smith, Jane. 2007. A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. Paper presented to 16th ICPhS, Saarbrücken, Germany, 2007.

Tatman, Rachael and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions. Paper presented to Interspeech 2017.

Temple, Rosalind A. M. 2014. *Where and what is (t, d)? A case study in taking a step back in order to advance sociophonetics*. Amsterdam: John Benjamins.

Wright, Marvin N. and Andreas Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1). 1-17.

Yaeger-Dror, Malcah, Tyler Kendall, Paul Foulkes, Dominic Watt, Jillian Oddie, Daniel Ezra Johnson and Philip Harrison. 2009. Perception of 'r': A cross-dialect comparison. Paper presented at the Linguistic Society of America, San Francisco.

Zhou, Xinhui, Carol Y. Espy-Wilson, Suzanne Boyce, Mark Tiede, Christy Holland and Ann Choe. 2008. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *Journal of the Acoustical Society of America* 123(6). 4466-4481.