The SPADE project:

large-scale analysis of a spoken language across space and time

Morgan Sonderegger,
The SPADE Consortium

SPADE
SPeech Across Dialects of English

LSCP Language Group, Paris
26 Mar, 2019

SPADE

SPeech Across Dialects of English

August 2017 – July 2020

http://spade.glasgow.ac.uk/

Trans-Atlantic Platform
Social Sciences and Humanities

DIGGING INTO DATA CHALLENGE

THE UNIVERSITY of EDINBURGH

University of Glasgow

McGill

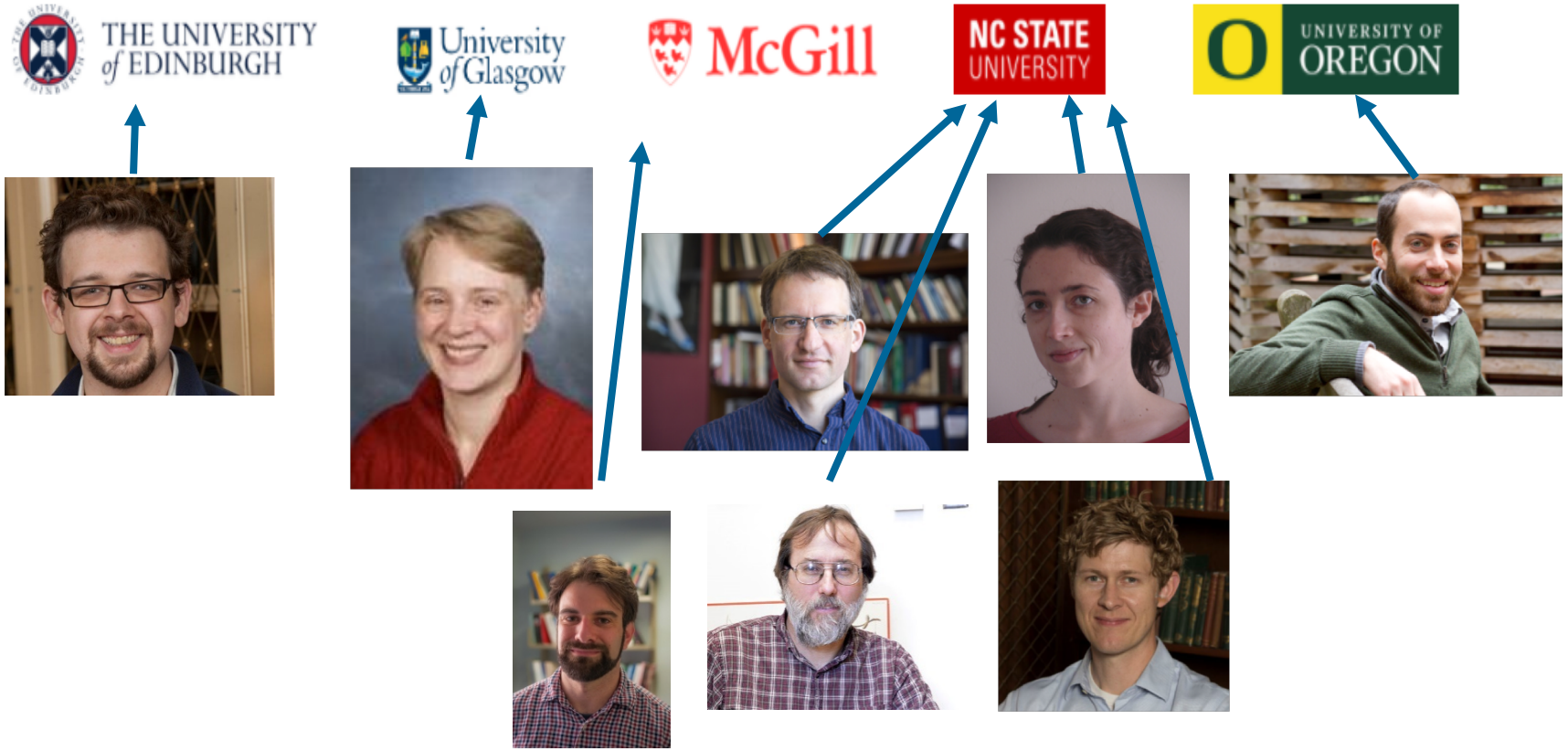
NC STATE UNIVERSITY

UNIVERSITY OF OREGON

SPADE

SPeech Across Dialects of English

investigators



http://spade.glasgow.ac.uk/

SPADE

SPeech Across Dialects of English

Postdocs

THE UNIVERSITY of EDINBURGH

University of Glasgow

McGill

NC STATE UNIVERSITY

UNIVERSITY OF OREGON

Project manager + data

Software development

Graduates + undergrads: many

http://spade.glasgow.ac.uk/

SPADE

SPeech Across Dialects of English

Project goals

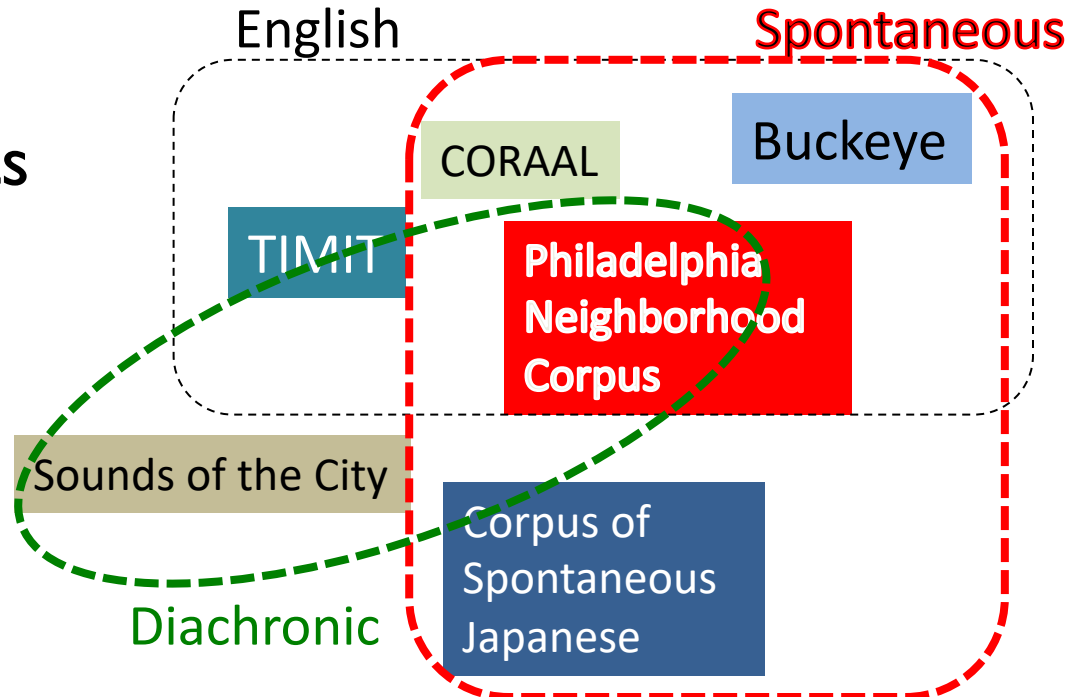
- Software for large-scale automated analysis of speech datasets
- Data from ~40 dialects
  - public & private
  - Focus: sociolinguistic data
- Case studies: investigate how ”English” varies in time and space

Motivation

- Huge amount of annotated speech data exists
  - Corpora, academic labs…

At least orthography + audio

- Across
  - Languages/dialects
  - Speech styles
  - Time



Motivation

- Huge amount of annotated speech data exists
  - Corpora,
    academic labs,
    fieldwork…
- Across
  - Languages/dialects
  - Speech styles
  - Time
- + ever-better (semi)-automatic speech measurement tools

Motivation

- Great potential for speech analysis for different purposes
  - Bigger haystacks, same-sized needle…
  - ... need a bigger magnet
- Requires software for unified corpus analysis
  - Integrating speech datasets
  - Querying across them
- SPADE focus: sociolinguistic, phonetic datasets

Barriers

- Speech datasets:
  - Large
  - Complex
  - Diverse formats

- Access to many speech datasets
  - Costly or ethically restricted

Most sociolinguistic, laboratory data

Switchboard: $3000+

- Result: requires lots of specialized code, €€, effort, computational power

Software goals

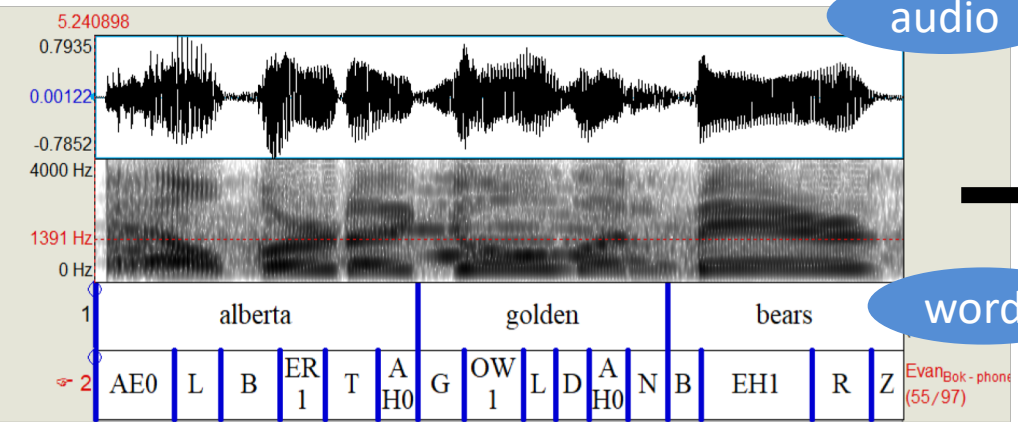
- Scalable & fast
- Require minimal technical skill from user
- Abstraction away from dataset format
- Querying dataset without access to raw data

- ⇒ Easier large-scale studies using speech corpora

- To motivate structure of software:
  - think about steps researcher goes through to do a (speech) corpus study
  - Running example: vowel formants

- Setting: sociolinguistic study, or laboratory phonology, phonetics, etc.

Raw data



audio

words

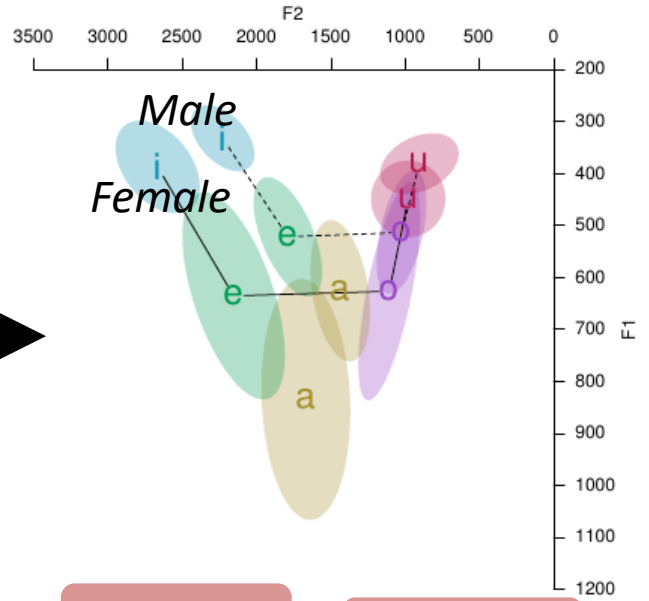
phones

utterances

Speaker M01
Gender: M
Age: 35
…

speaker questionnaire

Analysis



Gender

formants

Age

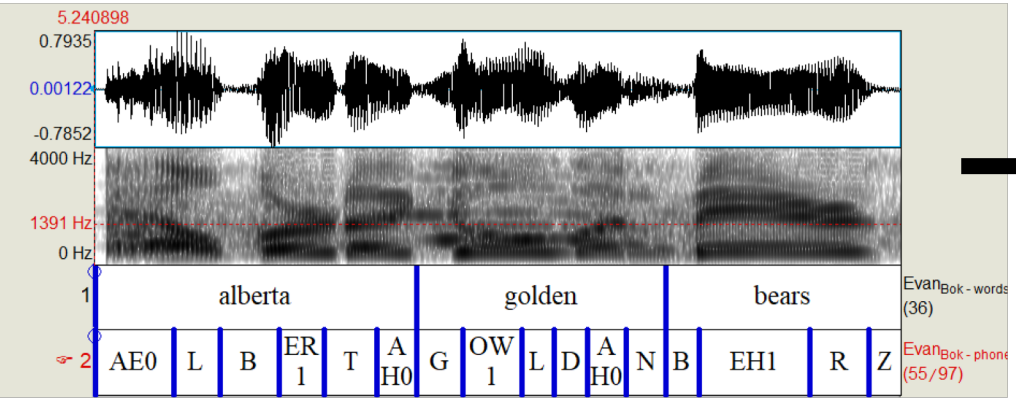
adjacent segments

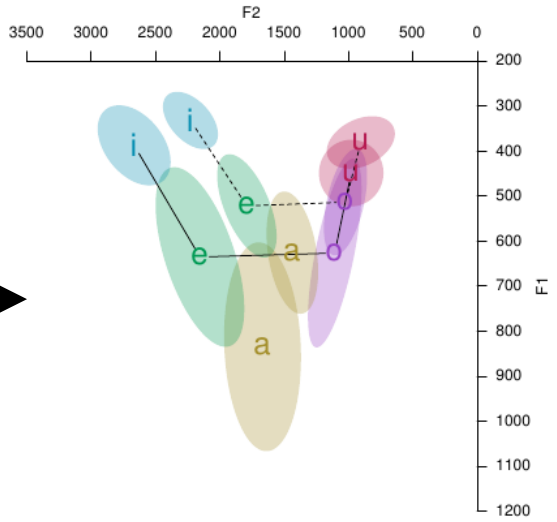
Vowel duration

Source: http://drammock.github.io/phonR/

Raw data

Analysis





(R, Goldvarb…)

How?

Data file (CSV)

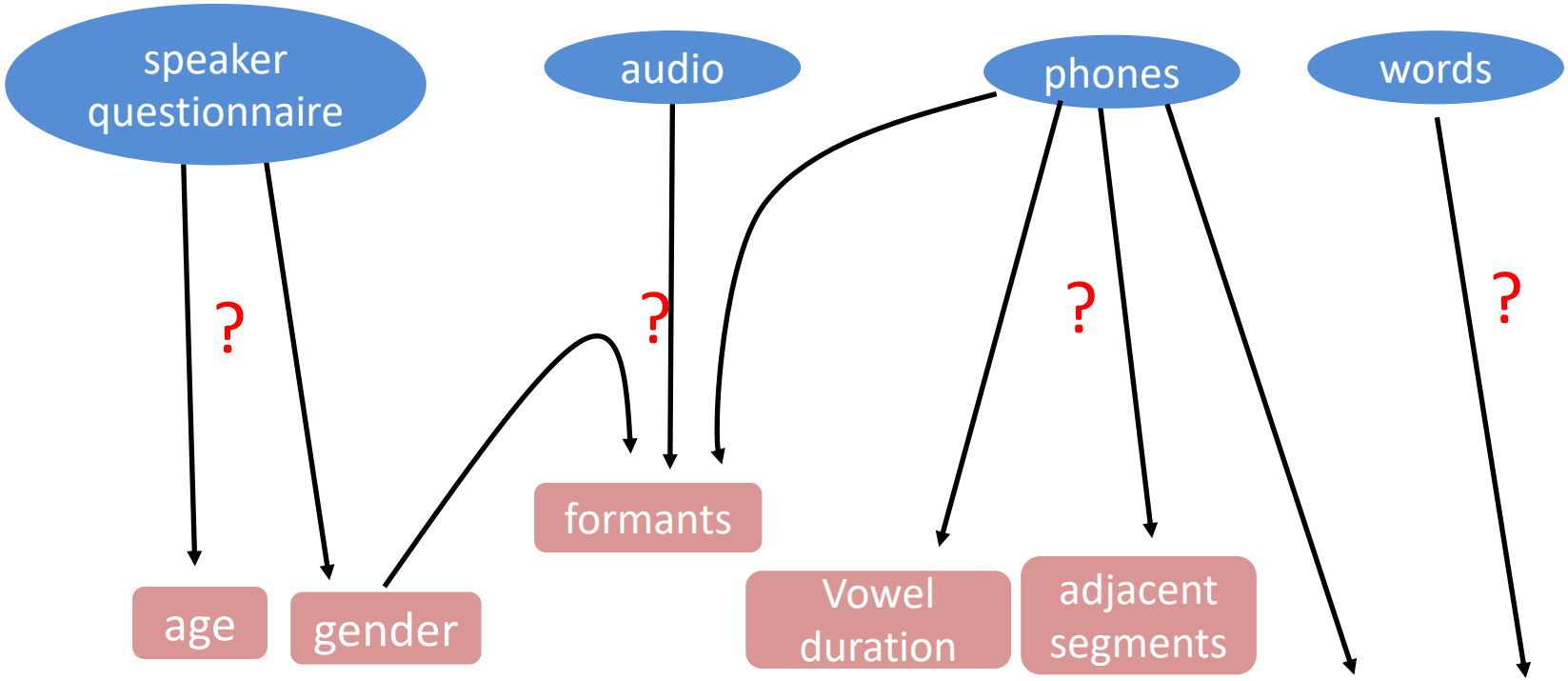
- Gender
- formants
- Vowel duration
- Age
- adjacent segments

1. Process raw data
2. Make measures
3. Find relevant tokens
4. End up with usable spreadsheet

Raw data

SpkrID	*Age*	*Gender*	*F1*	*F2*	*V_dur*	*C_left*	*C_right*	*Phone*	*Word*
M01	35	M	340	1410	0.11 s	[g]	[s]	UW	goose
F01	22	F	480	1050	0.15 s	[r]	[r]	UW	truer
…	…							…	…

Data file(CSV)

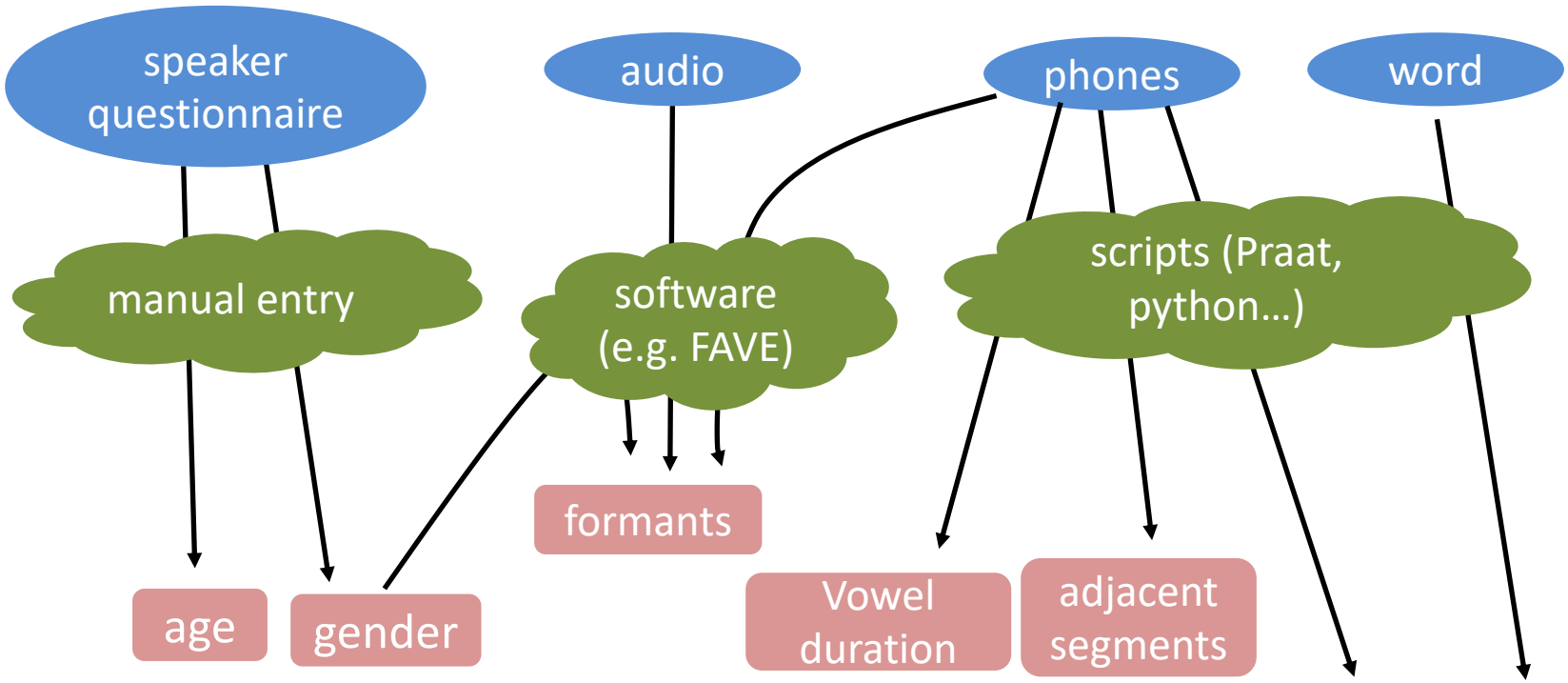


SpkrID	*Age*	*Gender*	*F1*	*F2*	*V_dur*	*C_left*	*C_right*	*Phone*	*Word*
M01	35	M	340	1410	0.11 s	[g]	[s]	UW	goose
F01	22	F	480	1050	0.15 s	[r]	[r]	UW	truer
…	…							…	…

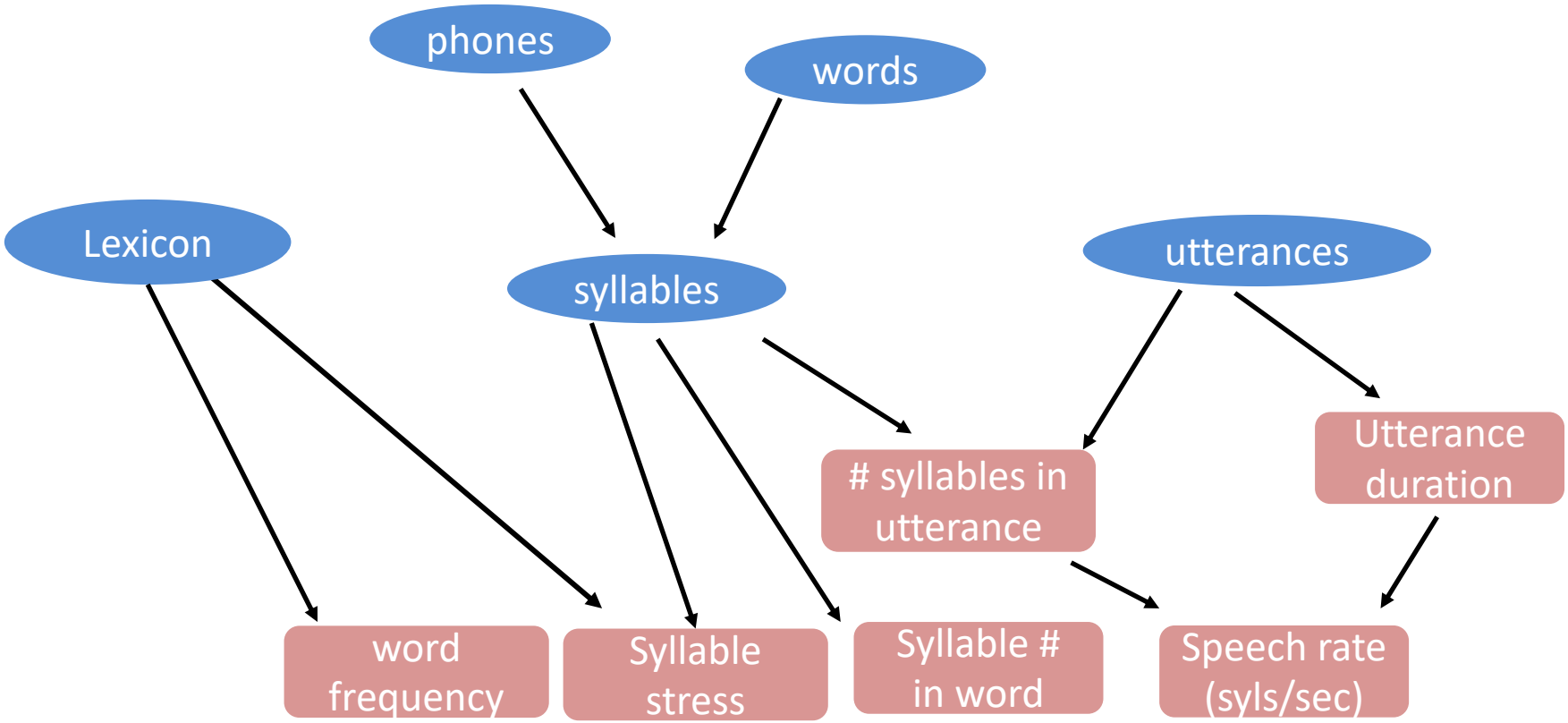
social factors

acoustic measures

phone info



ID	*Age*	*Gender*	*F1*	*F2*	*V_dur*	*C_left*	*C_right*	*Phone*	*Word*
M01	35	M	340	1410	0.11 s	[g]	[s]	UW	goose
F01	22	F	480	1050	0.15 s	[r]	[r]	UW	truer
…	…							…	…



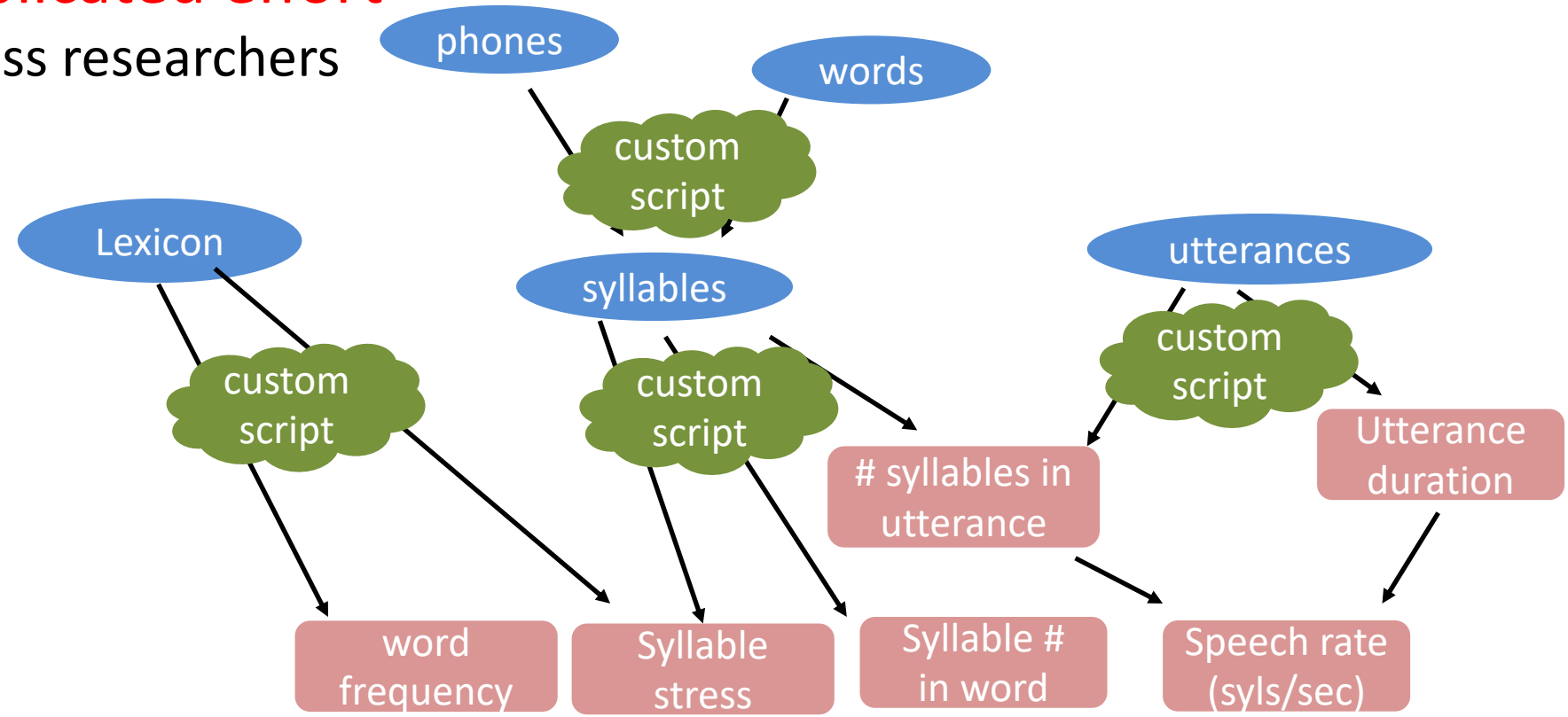
Phone	*Word*	*Frequency*	*Stress*	*Syll_#*	*Speech rate (sylls/sec)*
UW	goose	15	Y	1	0.11 s
UW	truer	25	Y	1	0.15 s
…	…				…

Lexical info

Prosodic/suprasegmental

Duplicated effort

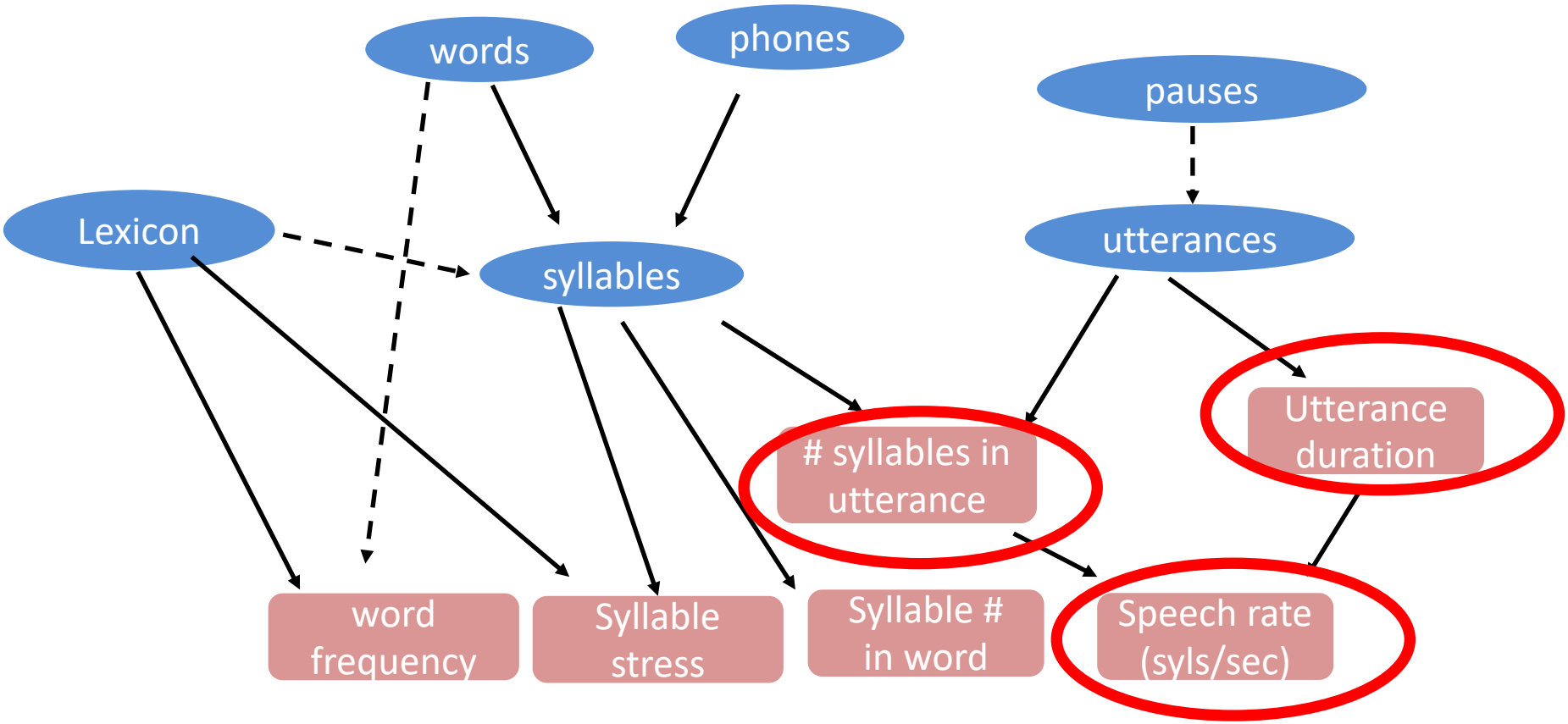
across researchers



Phone	*Word*	*Frequency*	*Stress*	*Syll_#*	*Speech rate (sylls/sec)*
UW	goose	15	Y	1	0.11 s
UW	truer	25	Y	1	0.15 s
…	…				…

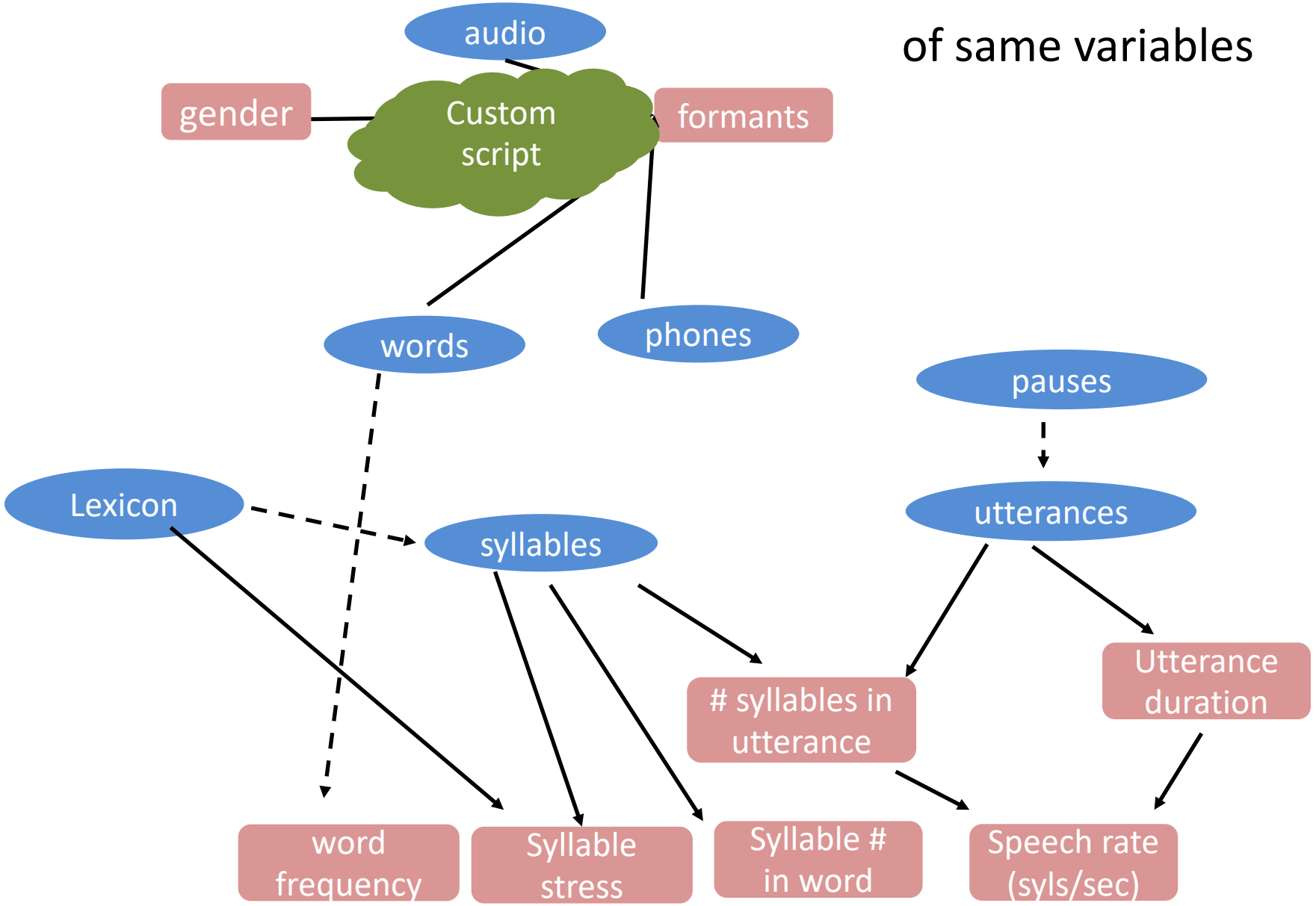
Lexical info

Prosodic/suprasegmental


words
phones
pauses
Lexicon
syllables
utterances
syllables in utterance
Utterance duration
word frequency
Syllable stress
Syllable # in word
Speech rate (syls/sec)

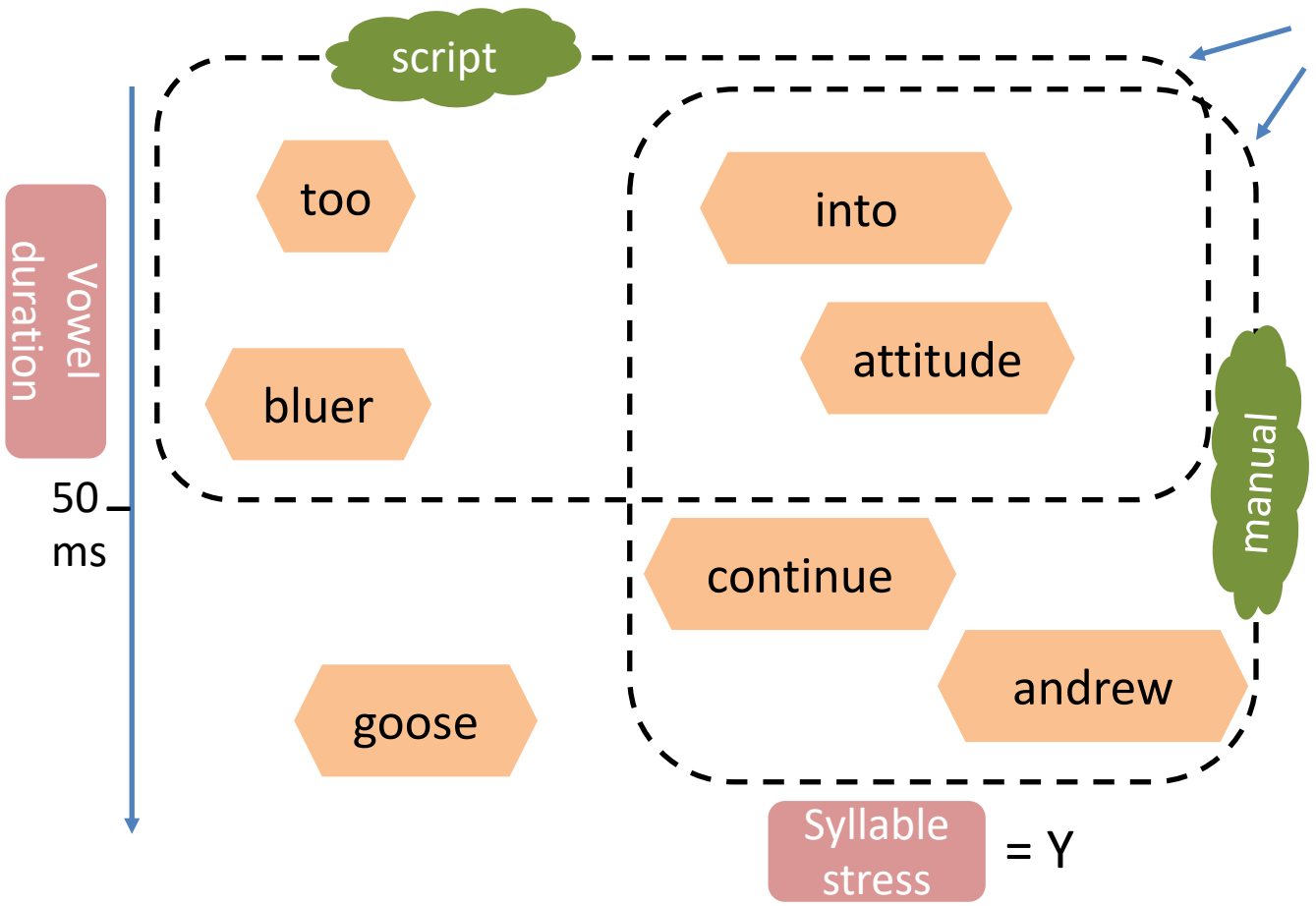
Different definitions

of same variables



1. Process raw data
2. Make measures
3. Find relevant tokens
4. End up with usable spreadsheet

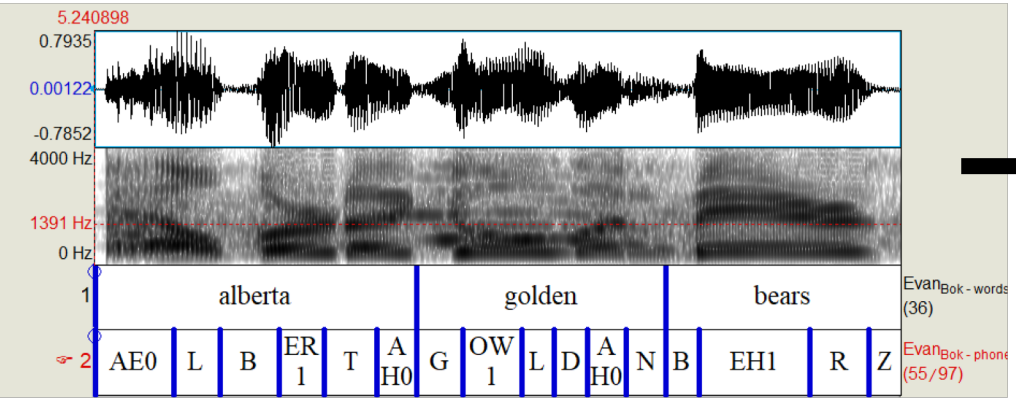
e.g. UW (GOOSE) tokens, stressed syllables > 50ms

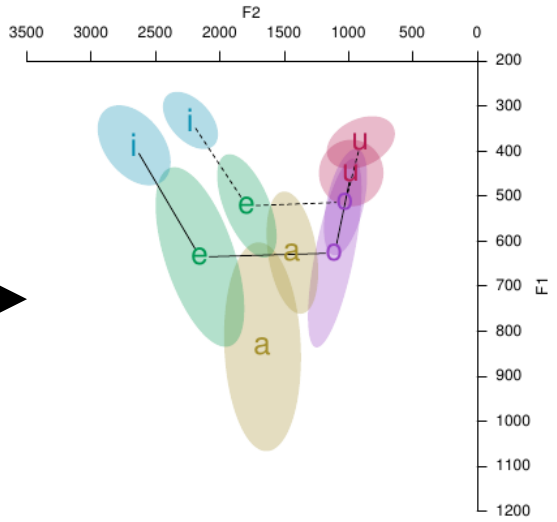


Same primitives as other steps

Raw data

Analysis





(R, Goldvarb…)

- **scripts**
- **manual entry**
- **software**

Data file (CSV)

Gender

formants

Vowel duration

Age

adjacent segments

Why 'Integrated' Speech Corpus ANalysis?

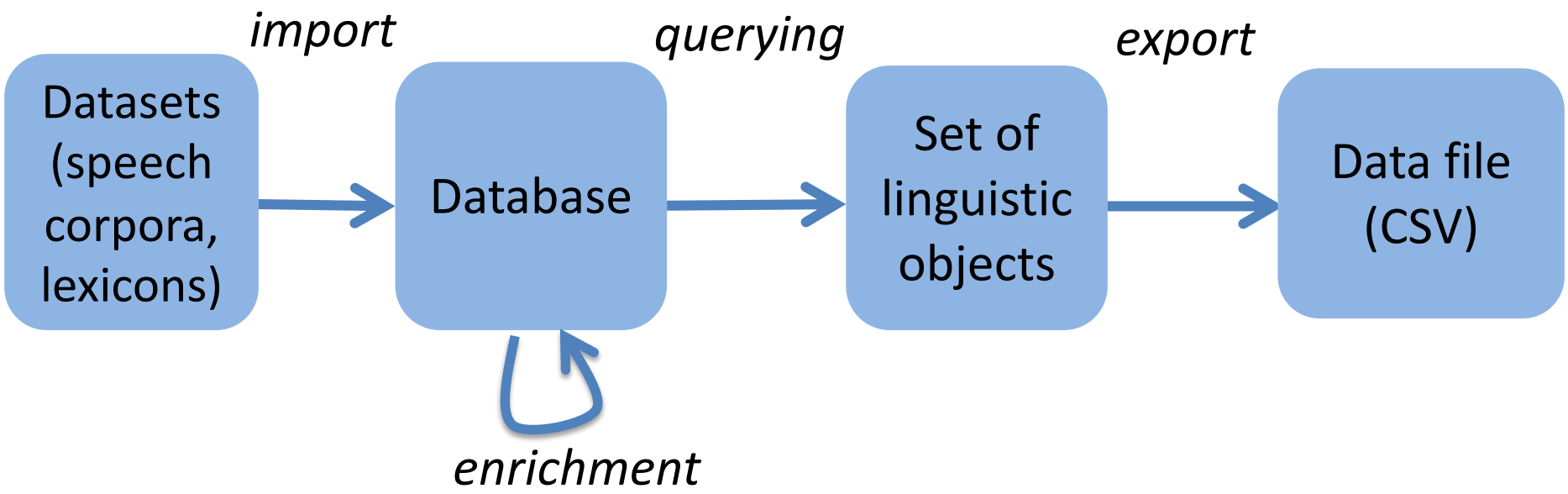
ISCAN

- Practical reasons
  - Technical skill
  - Time/duplication of effort
  - Availability
- Methodological/theoretical reasons
  - Standardized, customizable linguistic measures
- much more difficult with 1+ corpora…

ISCAN: A SYSTEM FOR INTEGRATED PHONETIC ANALYSES ACROSS SPEECH CORPORA

Michael McAuliffe[a], Arlie Coles[a], Michael Goodale[a], Sarah Mihuc[a], Michael Wagner[a], Jane Stuart-Smith[b], Morgan Sonderegger[a]

Proc. ICPhS 2019



- Implementation
  - Python API
  - Graphical User Interface

https://github.com/MontrealCorpusTools/iscan-spade-server

https://iscan.readthedocs.io/

- (show GUI here)

- Note:
  - Server-client architecture enables analysis without access to raw data
  - Permissions system controls who can see/hear tokens
  - Can be installed on web server (default) or personal computer

SPADE: datasets

- 43+ datasets, 4 countries, 115 years
- heterogeneous corpus formats
- public and private

SPADE: datasets

To date:

- Acquired: 20
- Measurements generated: 10
- ~ 10 dialect regions
  - ~500 hours (?)

SPADE: ethics and credit

- For private datasets (data guardians):
  ethics complex: GDPR + US laws
- Data transfer agreement
  - data use in keeping with original permissions, as far as is possible
- We welcome new datasets!

- "SPADE consortium" author on everything
- Datasets of measures -> data guardians at end of project

Case studies

Done	Sibilants Stuart-Smith et al. *Proc. ICPhS 2019*	Vowels Mielke et al. *Proc. ICPhS 2019*
In progress	Vowels 2 Stops	Vowels 3
Planned	Vowels (dynamic)	r, l

/s/-retraction in English

- /s/ → [ʃ]-like sound in /str/
  - *string, street*

- Sound change, varies by dialect:
  - Ex: London, Philadelphia, NZ English
  - but not others, e.g. RP, Australian English

    (e.g. Baker et al, 2011; Stevens and Harrington, 2016)

- varies by individual speaker within dialect

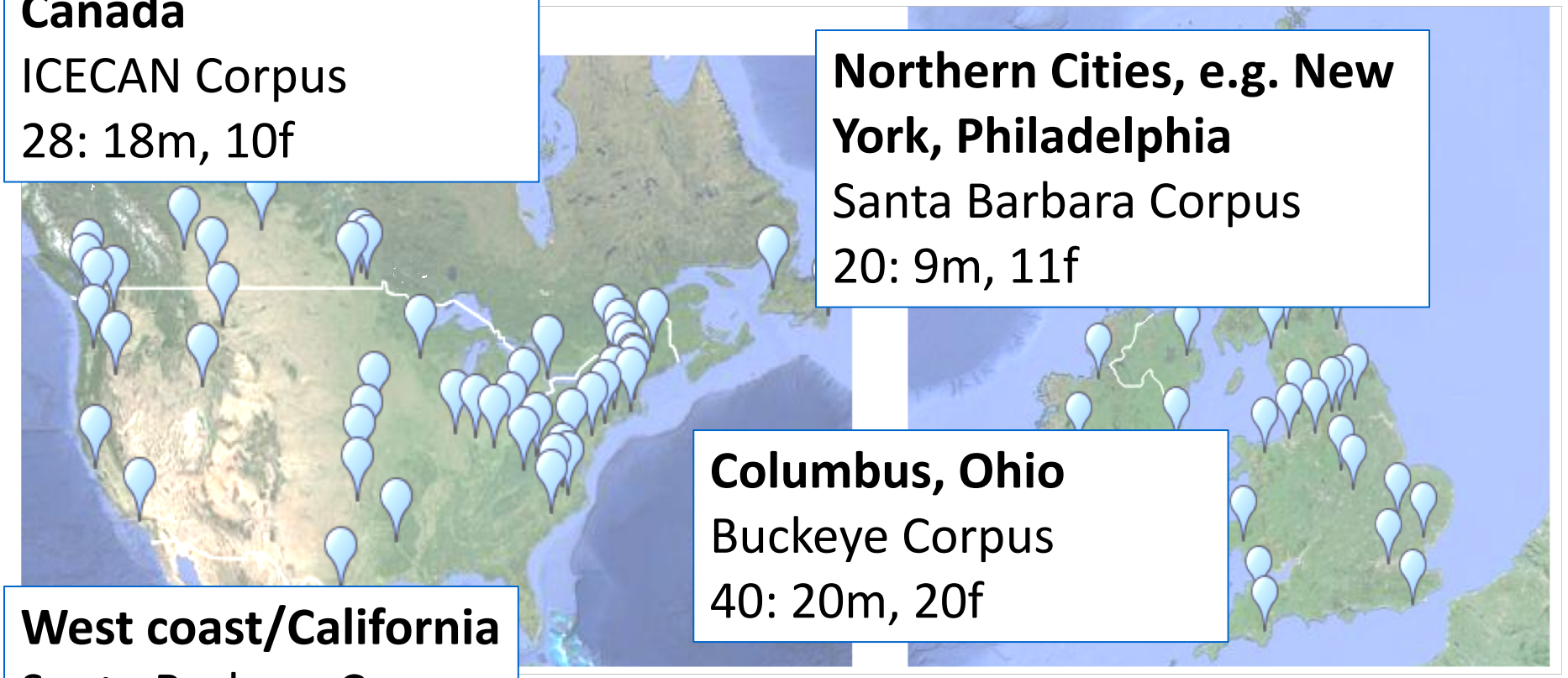
Research questions

- Q1: what is the evidence for /s/-retraction across English?
- Q2: Do English dialects show a dichotomous pattern of /s/-retraction
  - or a continuum?
- Received wisdom: there are "retracting"/non-retracting dialects and speakers

SPADE

Sample for this study: New World

Canada
ICECAN Corpus
28: 18m, 10f



Northern Cities, e.g. New York, Philadelphia
Santa Barbara Corpus
20: 9m, 11f

Columbus, Ohio
Buckeye Corpus
40: 20m, 20f

West coast/California
Santa Barbara Corpus
46: 20m, 26f

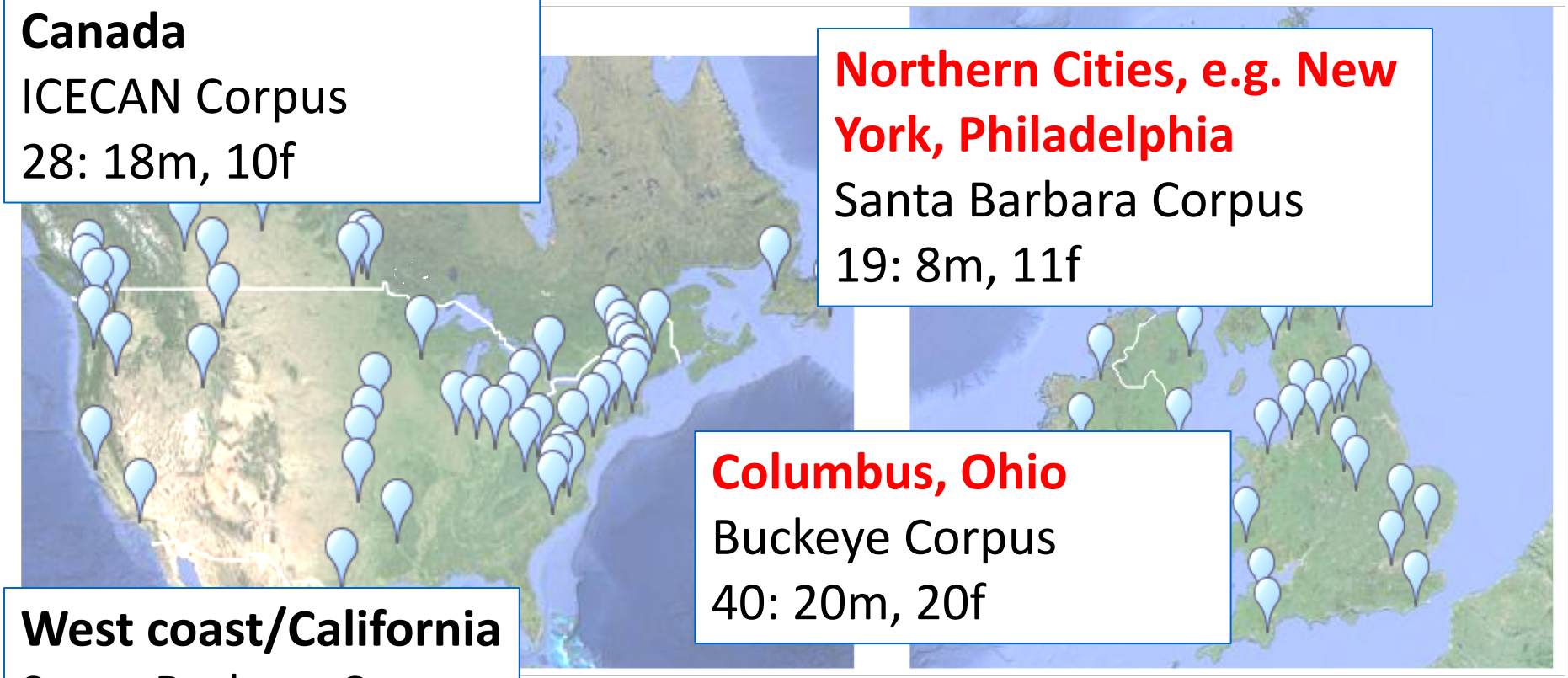
www.google.com/maps/

Raleigh, North Carolina
Raleigh Corpus
101: 50m, 51f

235 speakers

SPADE Sample for this study: New World

Canada
ICECAN Corpus
28: 18m, 10f



Northern Cities, e.g. New York, Philadelphia
Santa Barbara Corpus
19: 8m, 11f

Columbus, Ohio
Buckeye Corpus
40: 20m, 20f

West coast/California
Santa Barbara Corpus
43: 20m, 23f

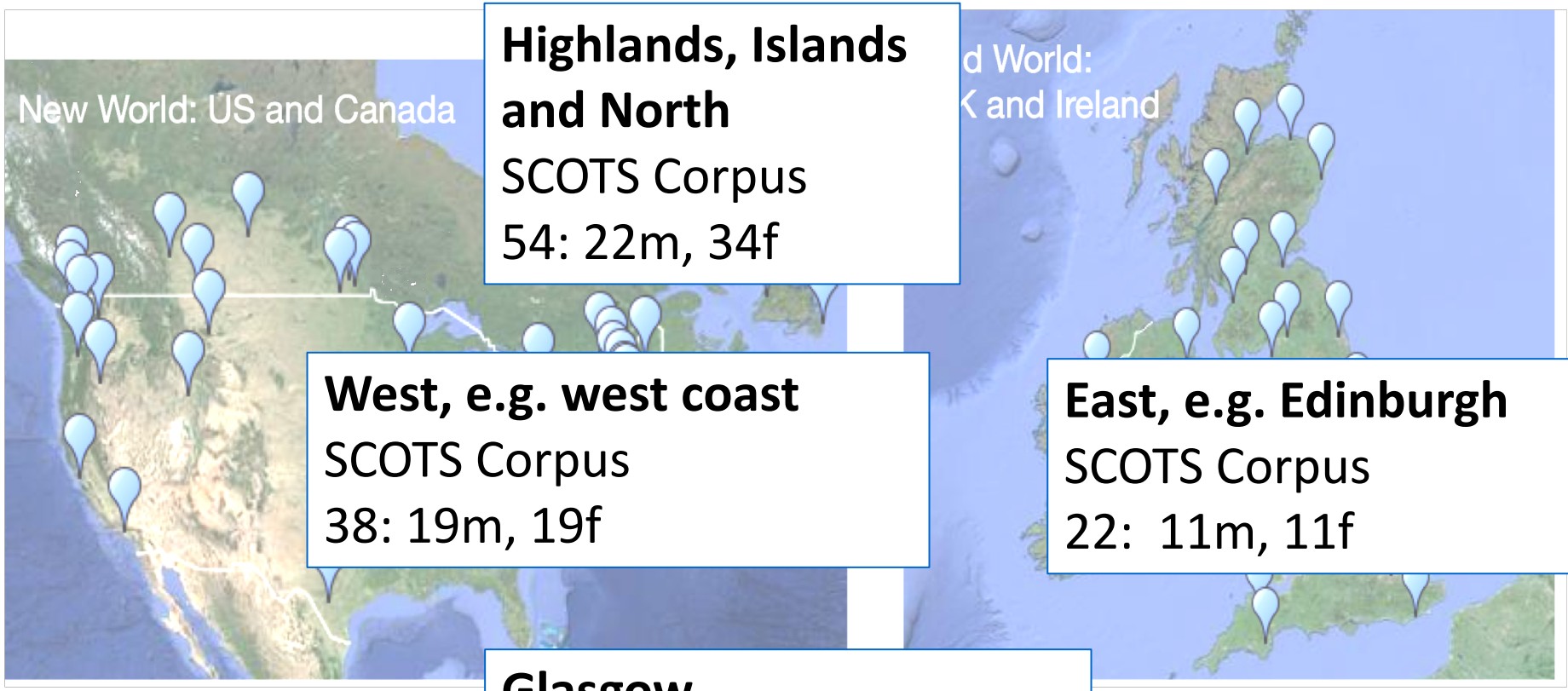
www.google.com/maps/

Raleigh, North Carolina
Raleigh Corpus
101: 50m, 51f

235 speakers

reported to show /s/-retraction

SPADE Sample for this study: Old World



Highlands, Islands and North
SCOTS Corpus
54: 22m, 34f

West, e.g. west coast
SCOTS Corpus
38: 19m, 19f

East, e.g. Edinburgh
SCOTS Corpus
22: 11m, 11f

Glasgow
Sounds of the City
70: 35m, 36f

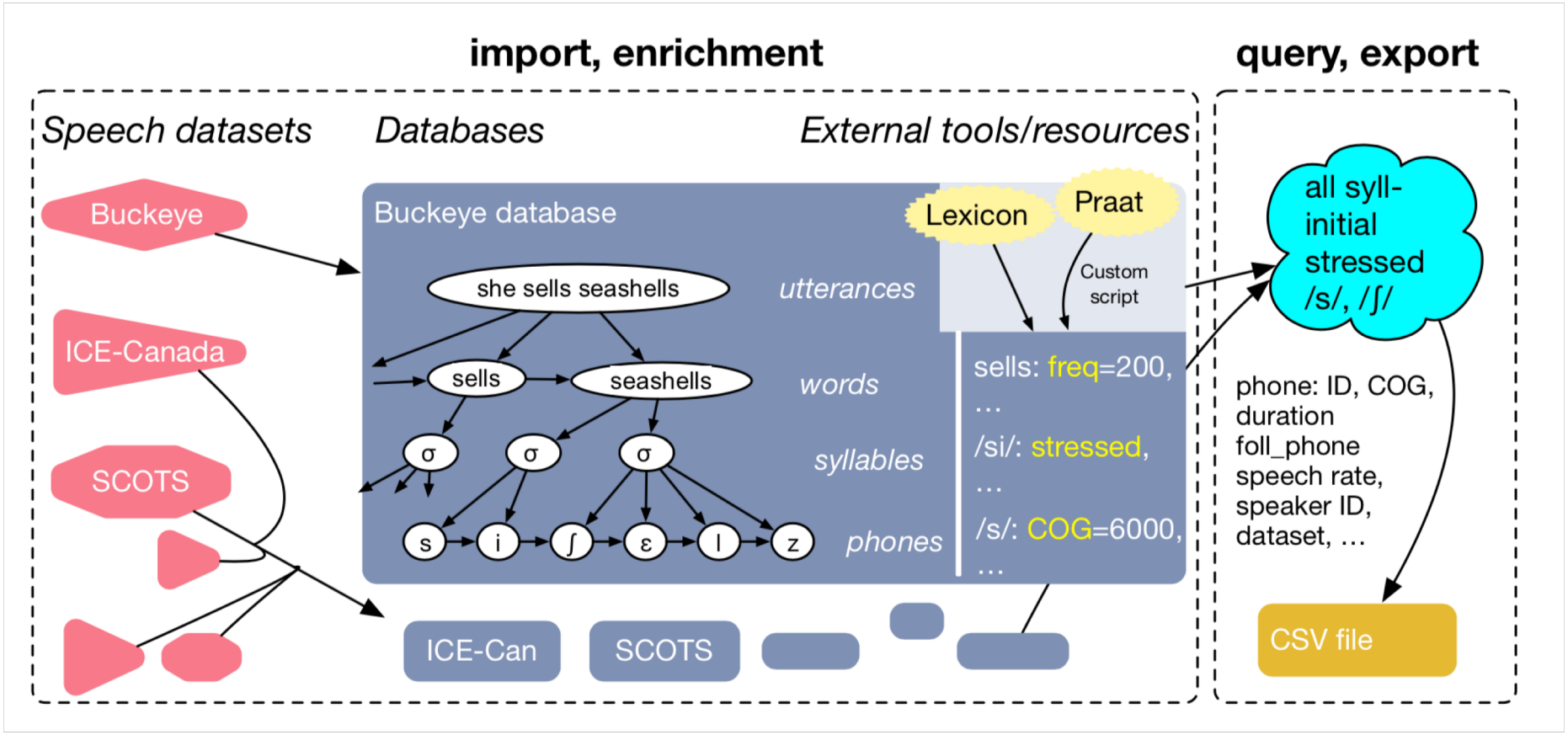
www.google.com/maps/

185 speakers

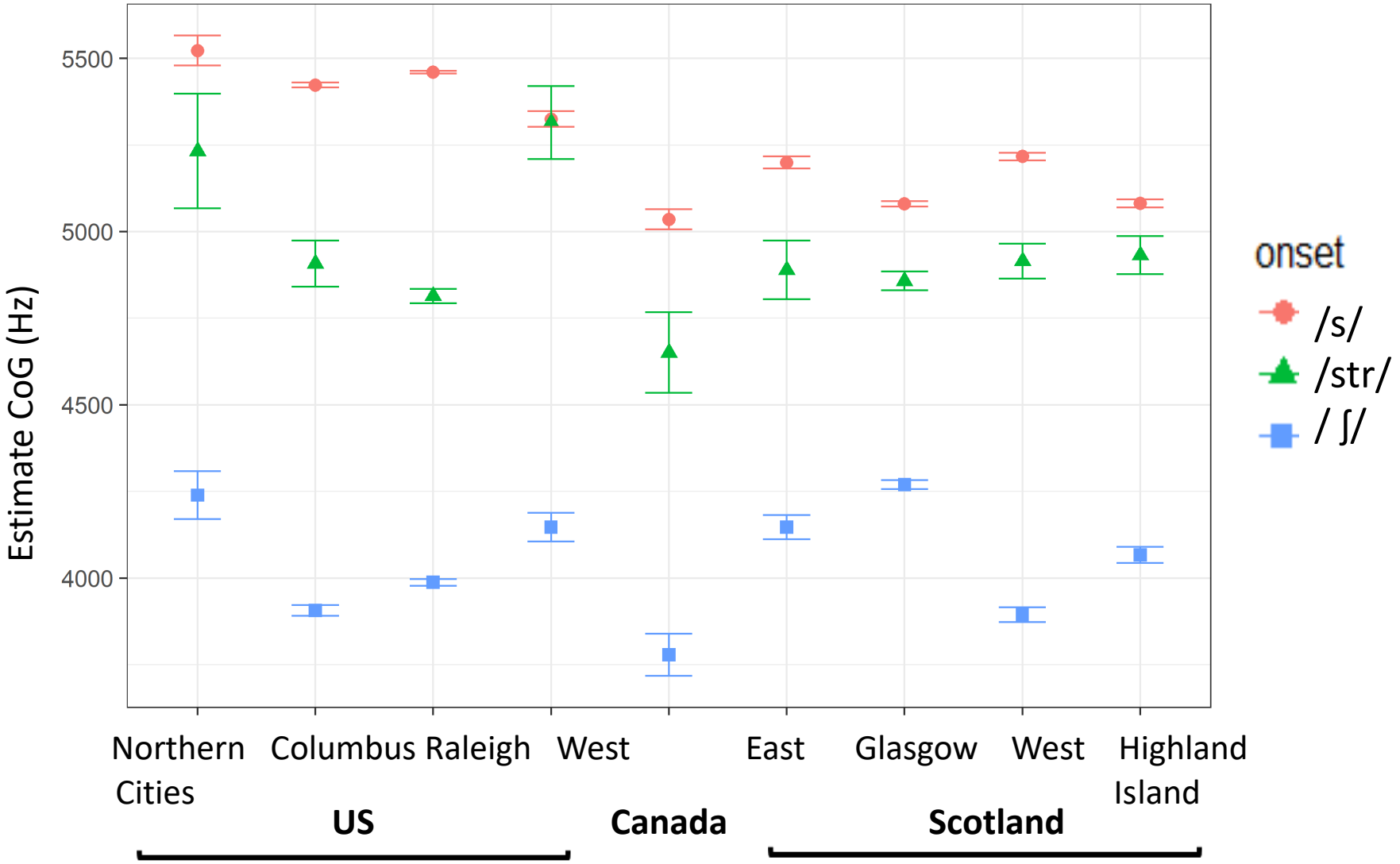
Data

- All instances of stressed, word-initial /s/
- Acoustic measures: peak, spectral Centre of Gravity (CoG)
  - 1-16 kHz
  - Middle 50%
- Data cleaning
- N = 76,440
- Prediction: **/s/ > /str/ > /ʃ/**

ISCAN usage

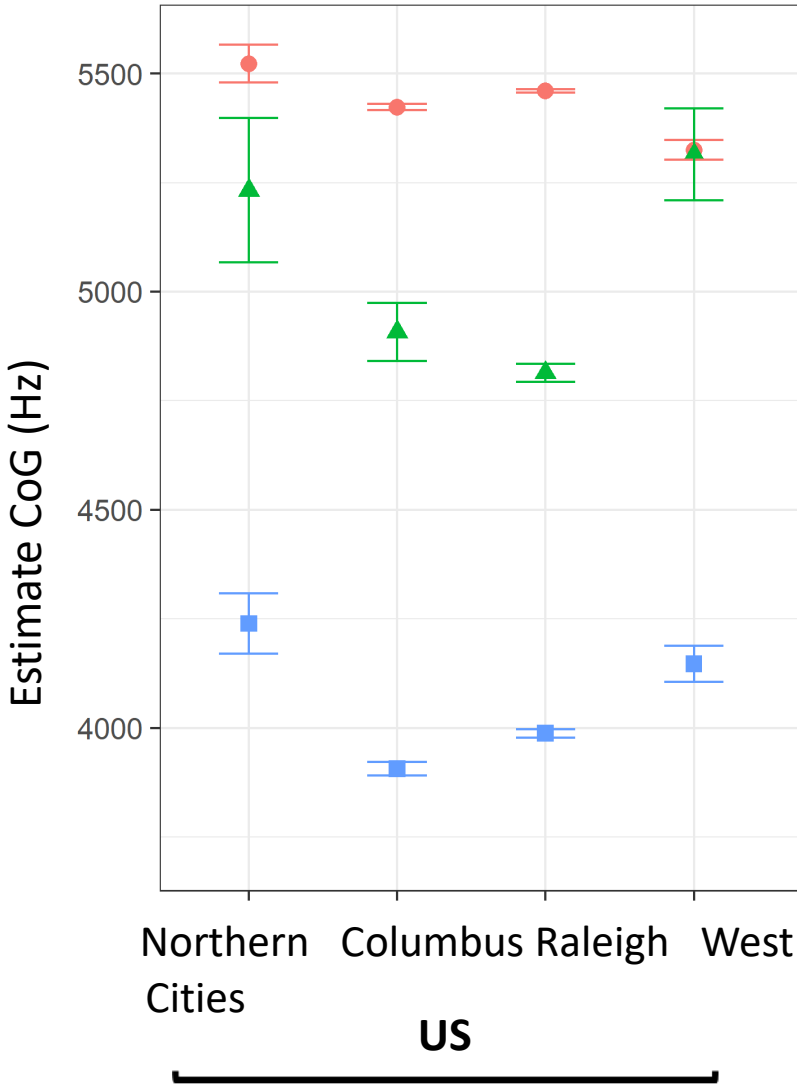

import, enrichment
query, export
Speech datasets
Databases
External tools/resources
Buckeye
ICE-Canada
SCOTS
Buckeye database
she sells seashells
utterances
sells
seashells
words
σ
σ
σ
syllables
s
i
ʃ
ɛ
l
z
phones
Lexicon
Praat
Custom script
sells: freq=200,
…
/si/: stressed,
…
/s/: COG=6000,
…
ICE-Can
SCOTS
all syll-initial stressed /s/, /ʃ/
phone: ID, COG, duration foll_phone speech rate, speaker ID, dataset, …
CSV file

Results



N = 76,440

- /str/ shows substantial variation across dialects

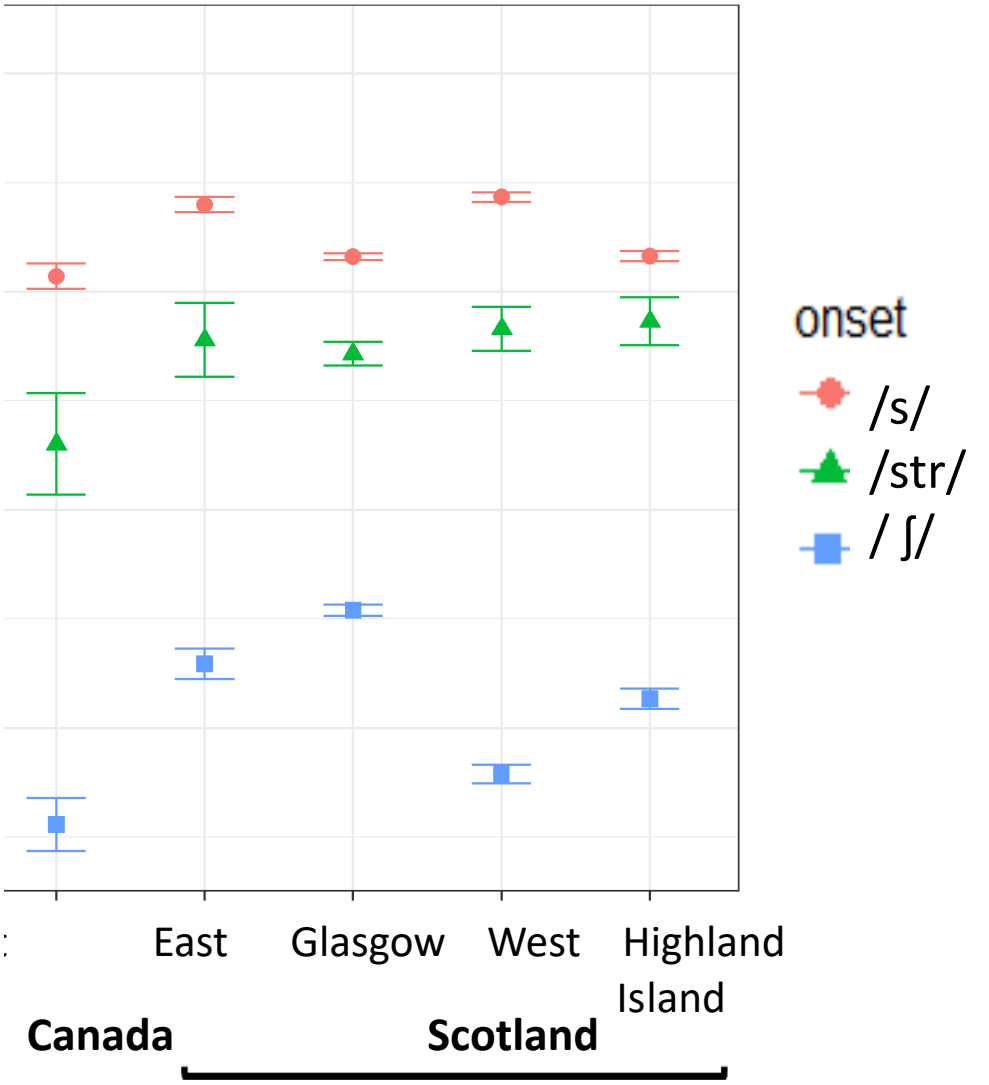


In **US** dialects, large differences in lowering of /str/ with respect to /s/

N = 76,440

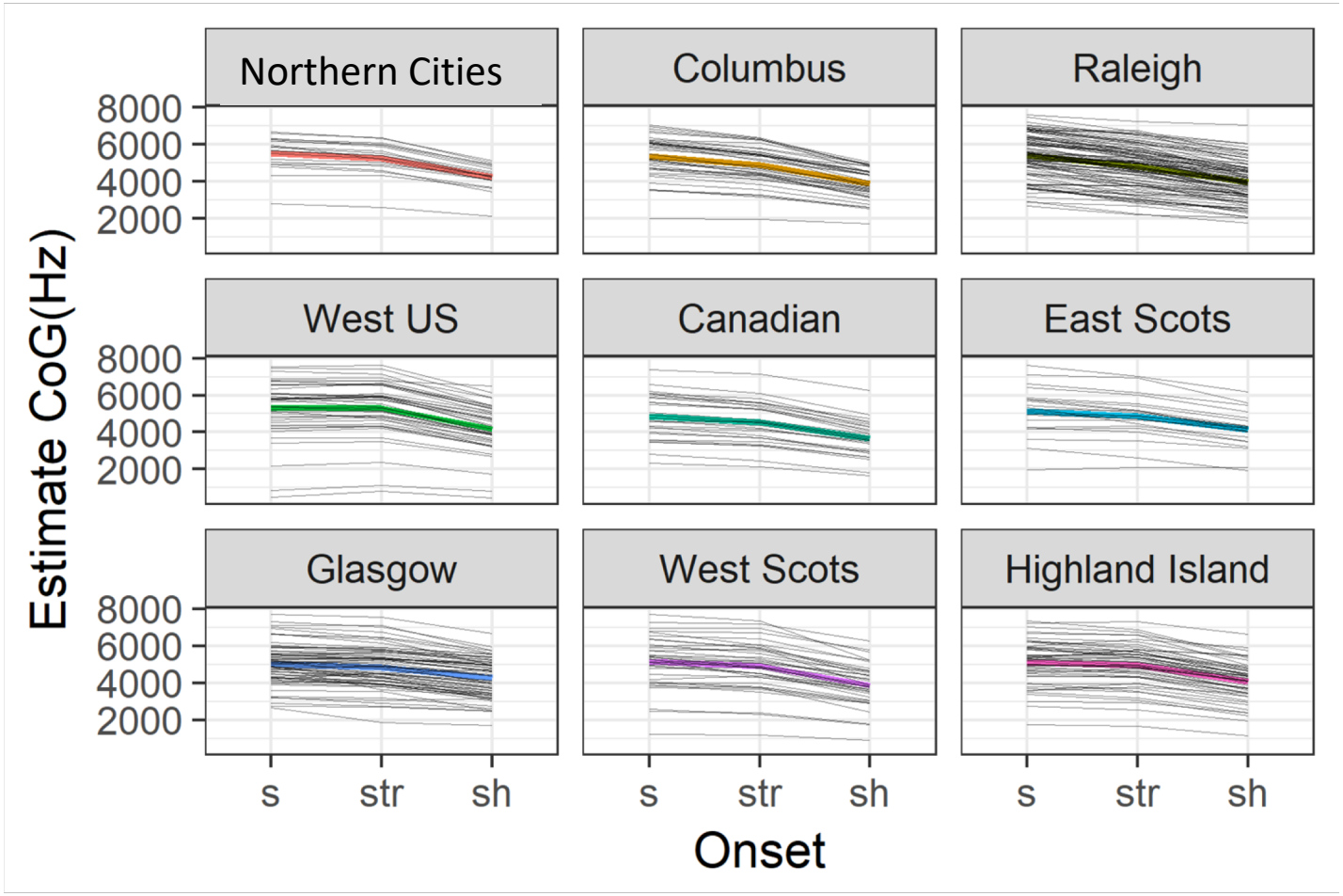
In **Scottish** and **Canadian** dialects, smaller differences between /str/ and /s/

/s/ is lower in frequency overall.



N = 76,440

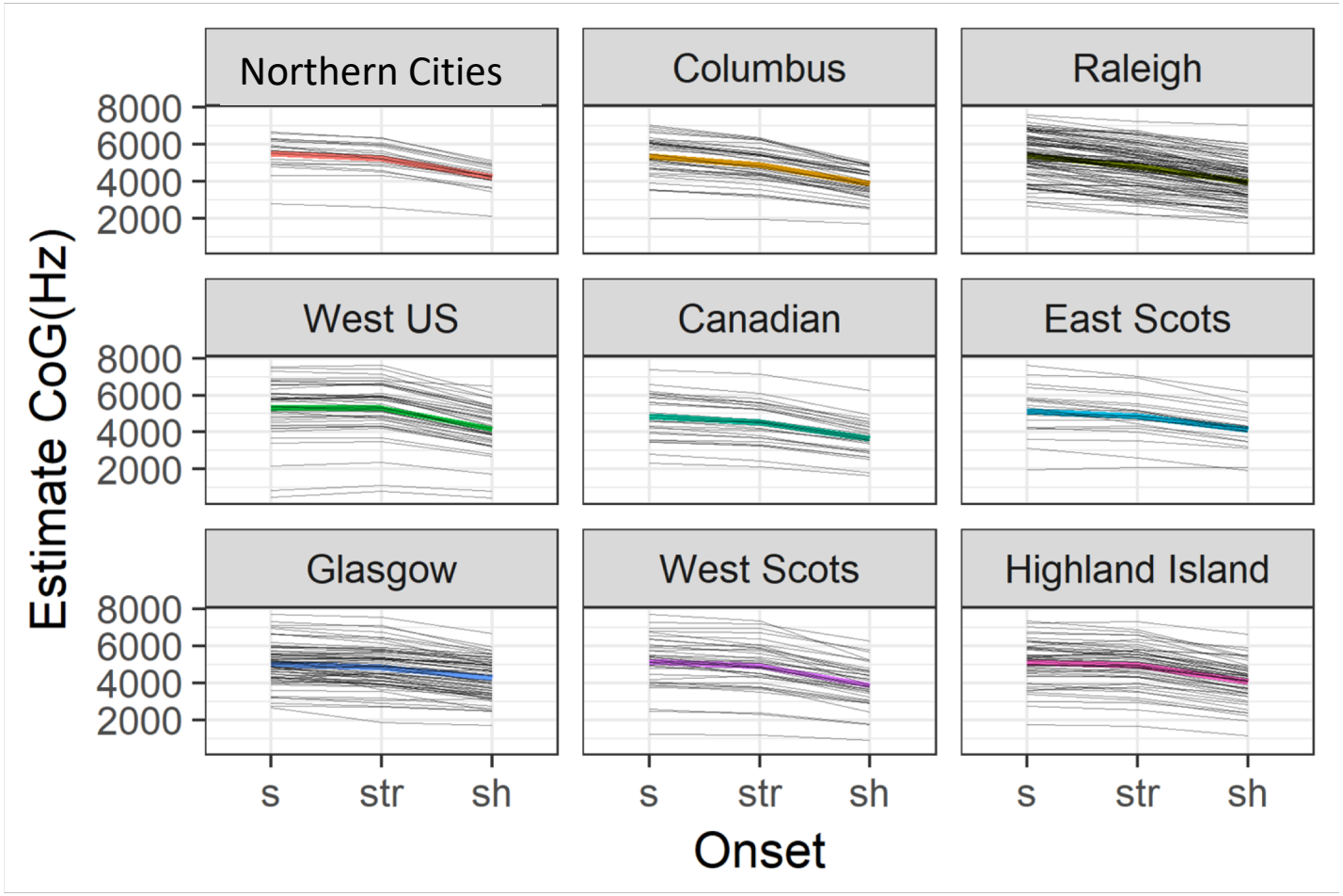
/s/-retraction in individuals within dialects



N = 76,440

retracting pattern shared by many Raleigh and Glasgow speakers.

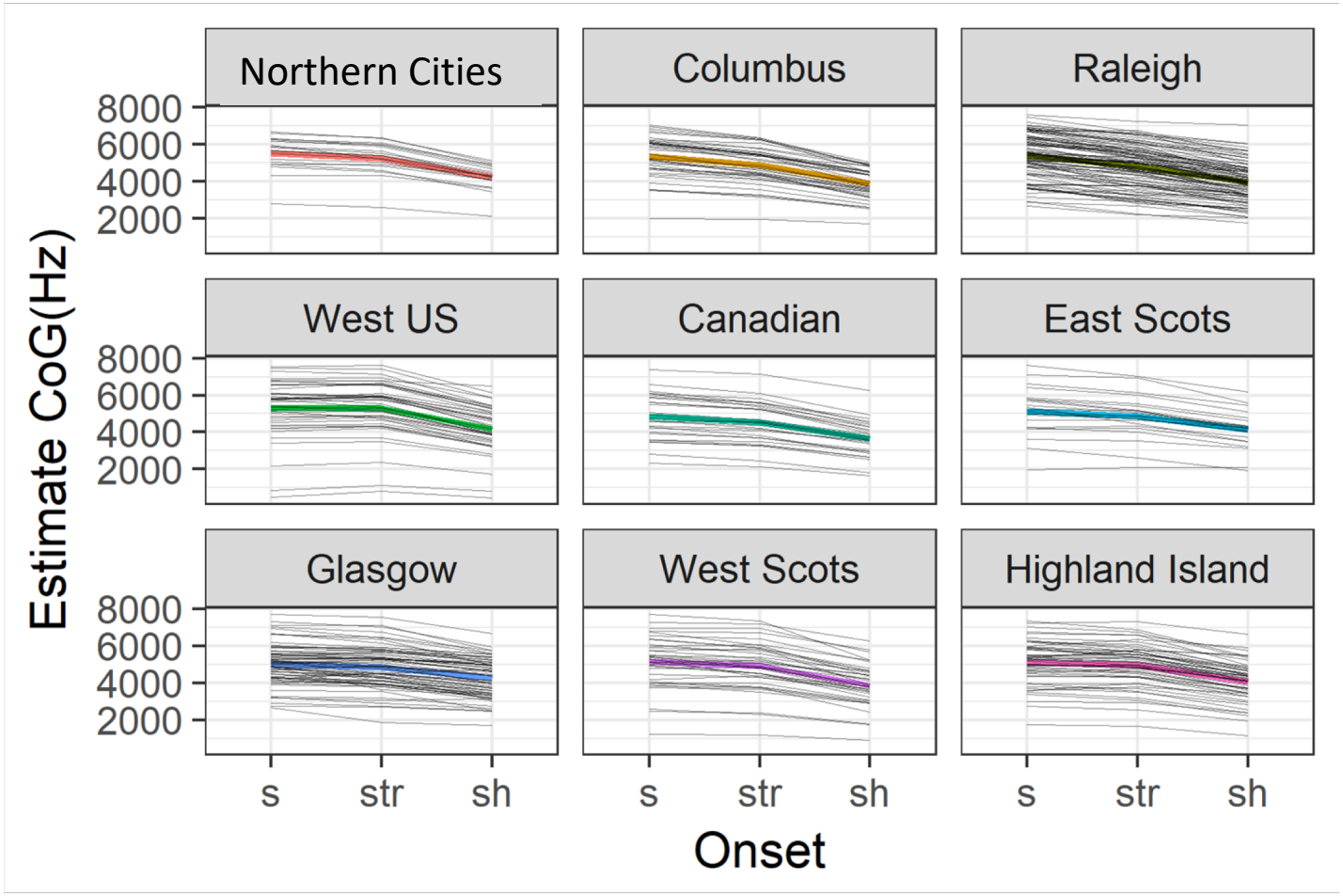
/s/-retraction in individuals within dialects



N = 76,440

non-retracting pattern in individual speakers in West US and Northern Cities dialects

/s/-retraction in individuals within dialects



N = 76,440

Both retracting and non-retracting patterns seen in some individuals in all dialects.

Discussion

- Q1: what is evidence for /s/ retraction across English?
  - Some dialects show retraction of /str/
  - Large differences by dialect and by country
  - Impression of "/s/-retraction" depends on which dialects are considered

- Q2: continuum vs. dichotomy in /s/-retraction
  - ?

- Scaling up analysis across dialects, with consistent measures, allows identification of new patterns

Study 2: vowel formants

- Influential hypothesis from sociolinguistics:
  (Labov, 1994)
  - Intraspeaker variation in vowel production ~ same axis as diachronic change in community

Mielke et al. *Proc. ICPhS 2019*

- Intuitively plausible, but unchecked

- (go to poster..)

Thanks!

- SPADE Team, especially
  - Jane Stuart-Smith, Michael McAuliffe, James Tanner, Vanna Willerton, Jeff Mielke

- MCQLL lab RAs
  - Michael Goodale, Arlie Coles, Elias Stengel-Eskin

- Funding
  - Digging Into Data, SSHRC, NSERC

Questions