

# Text mining with ngram variables

Matthias Schonlau, Ph.D.

# The most common approach to dealing with text data

- The most common approach to dealing with text data is as follows:
- Step 1: encode text data into numeric variables
  - Ngram variables
- Step 2: analysis
  - E.g. Supervised learning on ngram variables
  - E.g. Topic modeling (clustering)

(\*) Another common approach is to run neural network models. This gives higher accuracy in the presence of large amount of data.

# Overview

- Ngram variables approach to Text mining
- Example 1: Immigrant Data (German)
- Example 2: Patient Joe (Dutch)

# Text mining: “bag of words”

- Consider each distinct word to be a feature (variable)
- Consider the text “The cat chased the mouse”
  - 4 distinct features (words)
  - Each word occurs once except “the” which occurs twice

# Unigram variables

```
. input strL text
      text
1. "The cat chased the mouse"
2. "The dog chases the bone"
3. end;

. set locale_functions en

. ngram text threshold(1) stopwords(.)

. list t_* n_token
```

	t_bone	t_cat	t_chased	t_chases	t_dog	t_mouse	t_the	n_token
1.	0	1	1	0	0	1	2	5
2.	1	0	0	1	1	0	2	5

- Single-word variables are called unigrams
- Can use frequency or indicators (0/1)

# Unigram variables

- Threshold is the minimum number of observations in which the word has to occur before a variable is created.
- Threshold(2) means that all unigrams occurring only in one observation are dropped
- This is useful to limit the number of variables being created

```
. ngram text, threshold(2)  
stopwords(.)
```

```
. list t_* n_token
```

```
+-----+  
| t_the  n_token |  
+-----+  
1. |          2          5 |  
2. |          2          5 |  
+-----+
```

# Removing stopwords

- Remove common words “stopwords” unlikely to add meaning e.g. “the”
- There is a default list of stopwords
- The stopword list can be customized

```
. set locale_functions en
```

```
. ngram text threshold(1)
```

```
Removing stopwords specified in stopwords_en.txt
```

```
. list t_* n_token
```

```
+-----+
| t_bone  t_cat  t_chased  t_chases  t_dog  t_mouse  n_token |
+-----+
1. |      0      1      1      0      0      1      5 |
2. |      1      0      0      1      1      0      5 |
+-----+
```

# Stemming

- “chased” and “chases” have the same meaning but are coded as different variables.
- Stemming is an attempt to reduce a word to its root by cutting off the end
- E.g. “chased” and “chases” turns to “chase”
- This often works well but not always
- E.g. “went” does not turn into “go”
- The most popular stemming algorithm, the Porter stemmer, is implemented



# Stemming

```
. set locale_functions en
. ngram text threshold(1) stemmer
Removing stopwords specified in stopwords_en.txt
stemming in 'en'
```

```
. list t_* n_token
```

```
+-----+
| t_bone  t_cat  t_chase  t_dog  t_mous  n_token |
|-----|
1. |      0      1      1      0      1      5 |
2. |      1      0      1      1      0      5 |
+-----+
```

# “Bag of words” ignores word order

```
. input strL text
```

```
text
```

1. "The cat chased the mouse"
2. "The mouse chases the cat"
3. end;

- Both sentences have the same encoding!

```
. set locale_functions en
```

```
. ngram text threshold(1) stemmer degree(1)
```

```
Removing stopwords specified in  
stopwords_en.txt  
stemming in 'en'
```

```
. list t_* n_token
```

```
+-----+  
| t_cat   t_chase  t_mous  n_token |  
+-----+  
1. |      1      1      1      5 |  
2. |      1      1      1      5 |  
+-----+
```

# Add Bigrams

- Bigrams are two-word sequences
- Bigrams partially recover word order
- But ...

```
. ngram text threshold(1) stemmer degree(2)
Removing stopwords specified in
stopwords_en.txt
stemming in 'en'
```

```
. list t_chase_mous t_mous_chase
```

	t_chas~s	t_mous~e
1.	1	0
2.	0	1

# Add Bigrams

- ... But the number of variables grows rapidly

```
. describe simple
text          t_mous          t_cat_ETX          t_chase_mous      n_token
t_cat         t_STX_cat         t_cat_chase        t_mous_ETX
t_chase       t_STX_mous        t_chase_cat        t_mous_chase
```

Special bigrams:

STX\_cat : “cat” at the start of the text

cat\_ETX: “cat at the end of the text

# Ngram variables works

- While easy to make fun of the ngram variable approach works quite well on moderate size texts
- Does not work as well on long texts (e.g. essays, books) because there is too much overlap in words.

# Spanish

- Don Quijote de la Mancha
- “Give credit to the actions and not to the words “

```
. input strL text

      text
1. "Dad crédito a las obras y no a las palabras."
2. end;
```

```
.
. set locale_functions es

. ngram text, threshold(1) stemmer
Removing stopwords specified in stopwords_es.txt
stemming in 'es'
```

```
. list t_* n_token
```

```
+-----+
| t_crédit  t_dad  t_obras  t_palabr  n_token |
+-----+
1. |          1      1          1          1          10 |
+-----+
```

## Default Spanish Stopwords

de	le	les	nada	mi	estoy	estabais	he	habíais	soy	erais	tengo	teníais
la	ya	ni	muchos	mis	estás	estaban	has	habían	eres	eran	tienes	tenían
que	o	contra	cual	tú	está	estuve	ha	hube	es	fui	tiene	tuve
el	este	otros	poco	te	estamos	estuviste	hemos	hubiste	somos	fuiste	tenemos	tuviste
en	sí	ese	ella	ti	estáis	estuvo	habéis	hubo	sois	fue	tenéis	tuvo
y	porque	eso	estar	tu	están	estuvimos	han	hubimos	son	fuimos	tienen	tuvimos
a	esta	ante	haber	tus	esté	estuvisteis	haya	hubisteis	sea	fuisteis	tenga	tuvisteis
los	entre	ellos	estas	ellas	estés	estuvieron	hayas	hubieron	seas	fueron	tengas	tuvieron
del	cuando	e	algunas	nosotras	estemos	estuviera	hayamos	hubiera	seamos	fuera	tengamos	tuviera
se	muy	esto	algo	vosotros	estéis	estuvieras	hayáis	hubieras	seáis	fueras	tengáis	tuvieras
las	sin	mí	nosotros	vosotras	estén	estuviéramos	hayan	hubiéramos	sean	fuéramos	tengan	tuviéramos
por	sobre	antes	nuestra	os	estaré	estuvierais	habré	hubierais	seré	fuerais	tendré	tuvierais
un	ser	algunos	nuestros	mío	estarás	estuvieran	habrás	hubieran	serás	fueran	tendrás	tuvieran
para	también	qué	nuestras	mía	estará	estuviese	habrá	hubiese	será	fuese	tendrá	tuviese
con	me	unos	vuestro	míos	estaremos	estuvieses	habremos	hubieses	seremos	fueses	tendremos	tuvieses
no	hasta	yo	vuestra	mías	estaréis	estuviésemos	habréis	hubiésemos	seréis	fuésemos	tendréis	tuviésemos
una	hay	otro	vuestros	tuyo	estarán	estuviéseis	habrán	hubiéseis	serán	fueseis	tendrán	tuviéseis
su	donde	otras	vuestras	tuya	estaría	estuviesen	habría	hubiesen	sería	fuesen	tendría	tuviesen
al	quien	otra	esos	tuyos	estarías	estando	habrías	habiendo	serías	siendo	tendrías	teniendo
es	desde	él	esas	tuyas	estaríamos	estado	habríamos	habido	seríamos	sido	tendríamos	tenido
lo	todo	tanto		suyo	estaríais	estada	habríais	habida	seríais		tendríais	tenida
como	nos	esa		suya	estarían	estados	habrían	habidos	serían		tendrían	tenidos
más	durante	estos		suyos	estaba	estadas	había	habidas	era		tenía	tenidas
pero	todos	mucho		suyas	estabas	estad	habías		eras		tenías	tened
sus	uno	quienes		nuestro	estábamos		habíamos		éramos		teníamos	

# French

- Le Petit Prince
- “Please ... draw me a sheep...”

```
. input strL text

      text
1. "S'il vous plaît...dessine-moi un mouton..."
2. end;

. set locale_functions fr

. ngram text, threshold(1) stemmer
Removing stopwords specified in stopwords_fr.txt
stemming in 'fr'

. list t_* n_token
```

```
+-----+
| t_dessin  t_mouton  t_plaît  n_token |
+-----+
1. |          1          1          1          8 |
+-----+
```



# Swedish

“I have never tried that before, so I can definitely do that”

Pippi Longstocking  
(Astrid Lindgren)

```
. input strL text
```

```
text
```

1. "Det har jag aldrig provat tidigare så det klarar jag helt säkert."
2. end;

```
. set locale_functions sv
```

```
. ngram text, threshold(1) stemmer
```

```
Removing stopwords specified in stopwords_sv.txt  
stemming in 'sv'
```

```
. list t_* n_token
```

```
+-----+  
| t_aldr  t_helt  t_klar  t_prov  t_säkert  t_så  t_tid  n_token |  
+-----+  
1. |      1      1      1      1      1      1      1      12 |  
+-----+
```

# Internationalization

- The language affects ngram in 2 ways:
  - List of stopwords
  - Stemming
- Supported Languages are shown on the right along with their locale  
set locale\_functions <locale>
- These are European languages. Ngram does not work well for logographic languages where characters represent words (e.g. mandarin)
- Users can add stopword lists for additional languages, but not stemmers

da (Danish)  
de (German)  
en (English)  
es (Spanish)  
fr (French)  
it (Italian)  
nl (Dutch)  
no (Norwegian)  
pt (Portuguese)  
ro (Romanian)  
ru (Russian)  
sv (Swedish)

# Statistical learning algorithms in Stata

## **Traditional learning Algorithms:**

- boost: Gradient boosting (\*)
- svmachines: Support Vector Machines (\*)
- randomforest: Random Forests (\*)
- discrim knn: k Nearest Neighbor classification (no regression)

## **Regularized regressions:**

- plogit: penalized logistic regression
- lars: least angle regression
- krls: kernel-based regularized least squares
- lassopack: various Lasso algorithms

(\*) My programs are currently in transition from being downloadable from [schonlau.net](http://schonlau.net) to [ssc](http://ssc).

See User's corner on machine learning for some others:  
<https://www.stata.com/stata-news/news33-4/users-corner/>

# Immigrant Data

- As part of their research on cross-national equivalence of measures of xenophobia, Braun et al. (2013) categorized answers to open-ended questions on beliefs about immigrants.
- German language

Braun, M., D. Behr, and L. Kaczmirek. 2013. Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research* 25(3): 383{395.

# Open-ended question asked

- (one of several) statement in the questionnaire:
  - “Immigrants take jobs from people who were born in Germany”.
- Rate statement on a Likert scale 1-5
- Follow up with a probe:
  - “Which type of immigrants were you thinking of when you answered the question? The previous statement was: [text of the respective item repeated].”

# Immigrant Data

This question is then categorized by (human) raters into the following outcome categories:

- **General reference to immigrants**
- Reference to specific countries of origin/ethnicities (Islamic countries, eastern Europe, Asia, Latin America, sub-Saharan countries, Europe, and Gypsies)
- Positive reference of immigrant groups ("people who contribute to our society")
- Negative reference of immigrant groups ("any immigrants that[. . .] cannot speak our language")
- Neutral reference of immigrant groups ("immigrants who come to the United States primarily to work")
- Reference to legal/illegal immigrant distinction ("illegal immigrants not paying taxes")
- Other answers ("no German wants these jobs")
- **Nonproductive [Nonresponse or incomprehensible / unclear answer ( "its a choice")]**

# Key Stata code

```
set locale_functions de
ngram probe_all, degree(2) threshold(5) stemmer binarize

boost y t_* n_token if train, dist(multinomial) influence pred(pred) ///
      seed(12) interaction(3) shrink(.1)
```

- 242 ngram Variables created based on training 500 observations
  - Total data set had N=1006
- This is not a lot of variables; you can easily exceed 1000 variables

# Which ngram options do well?

- Use the options that perform best on a test data set

German Stemming	Remove German Stopwords	binarize	Accuracy
yes	remove	yes	61.9 %
yes	remove	no	62.5 %
no	remove	yes	61.9 %
no	keep	yes	68.2%
yes	keep	yes	71.2%

- The key message: keep German stopwords
  - This is not always true



# Default German Stopword List

Stopword lists are computed as the most common words in the language

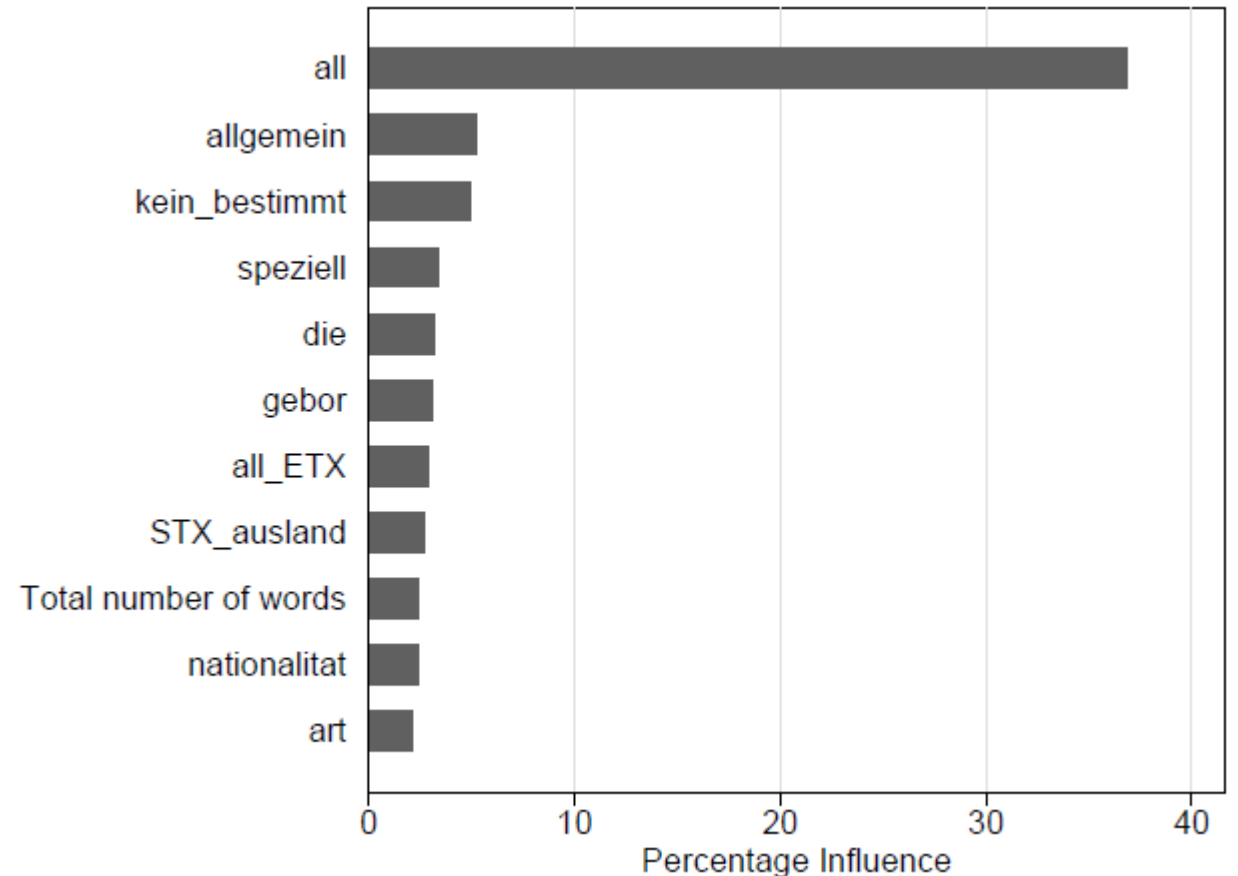
aber	deiner	hier	meines	war	bis	einigem	jenen	so	würden	
alle	deines	hin	mit	waren	bist	einigen	jener	solche	zu	
allem	denn	hinter	muss	warst	da	einiger	jenes	solchem	zum	
allen	derer	ich	musste	was	damit	einiges	jetzt	solchen	zur	
aller	dessen	mich	nach	weg	dann	einmal	kann	solcher	zwar	
alles	dich	mir	nicht	weil	der	er	kein	solches	zwischen	
als	dir	ihr	nichts	weiter	den	ihn	keine	soll		
also	du	ihre	noch	welche	des	ihm	keinem	sollte		
am	dies	ihrem	nun	welchem	dem	es	keinen	sondern		
an	diese	ihren	nur	welchen	die	etwas	keiner	sonst		
ander	diesem	ihrer	ob	welcher	das	euer	keines	über		
andere	diesen	ihres	oder	welches	daß	eure	können	um		
anderem	dieser	euch	ohne	wenn	derselbe	eurem	könnte	und		
anderen	dieses	im	sehr	werde	derselben	euren	machen	uns		
anderer	doch	in	sein	werden	denselben	eurer	man	unse		
anderes	dort	indem	seine	wie	desselben	eures	manche	unsem		
anderem	durch	ins	seinem	wieder	demselben	für	manchem	unsen		
andern	ein	ist	seinen	will	dieselbe	gegen	manchen	unser		
anderr	eine	jede	seiner	wir	dieselben	gewesen	mancher	unses		
anders	einem	jedem	seines	wird	dasselbe	hab	manches	unter		
auch	einen	jeden	selbst	wirst	dazu	habe	mein	viel		
auf	einer	jeder	sich	wo	dein	haben	meine	vom		
aus	eines	jedes	sie	wollen	deine	hat	meinem	von		
bei	einig	jene	ihnen	wollte	deinem	hatte	meinen	vor		
bin	einige	jenem	sind	würde	deinen	hatten	meiner	während		

# Interpretable black-boxes

- In linear regression we can interpret every coefficient
- Statistical learning models are black-box models and generally difficult to interpret
  - with potentially thousands of coefficients
- One of the great joys is to look at influential variables

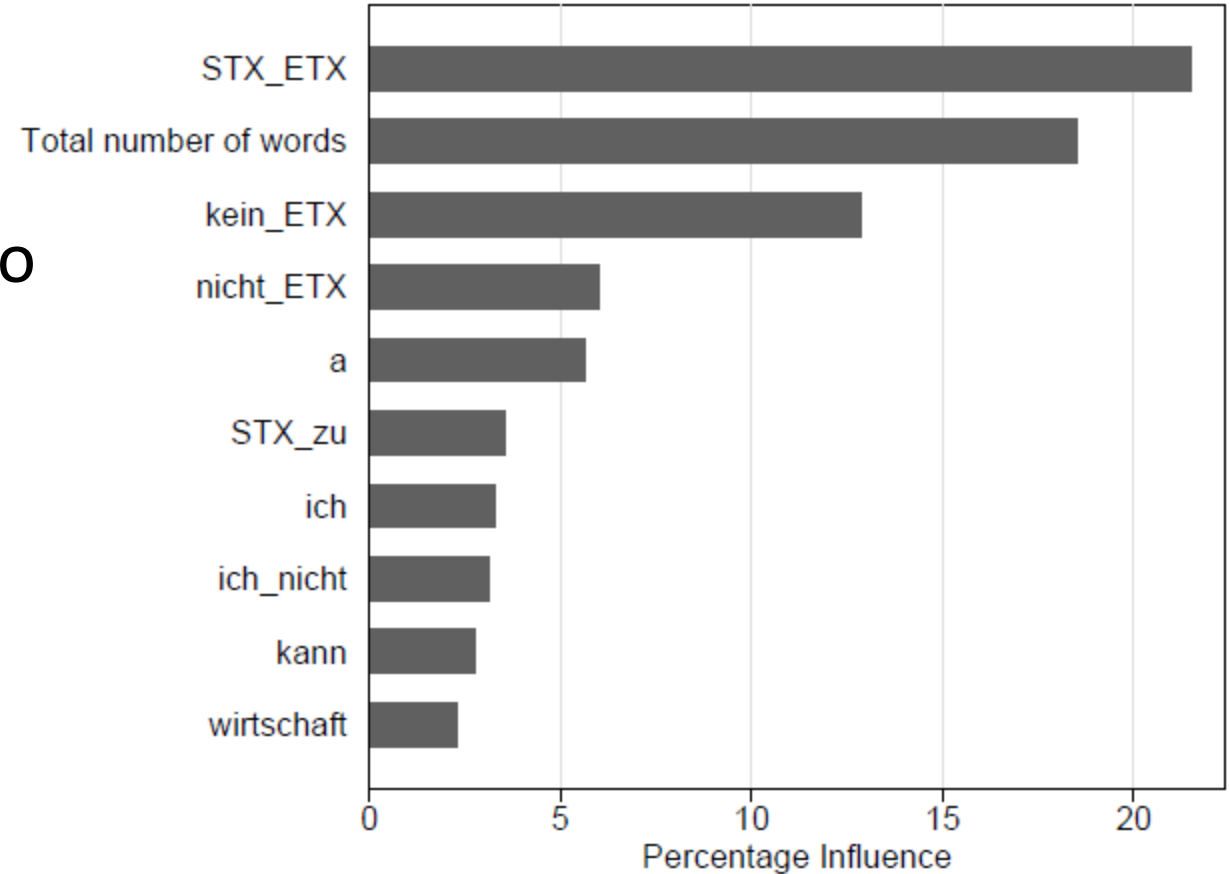
# Influential variables for the outcome “general”

- Influential words for outcome “general”
- “all” (same meaning in English).
- “allgemein”, means “general”
- “kein bestimmt” translates to “no particular” as in “no particular type of foreigner”.
- Several other influential variables refer to general groups of foreigners such as stemmed words of nationality, and foreigners



# Influential variables for the outcome “non-productive”

- STX\_ETX is a line with zero words  
May contain “-”, “.” and “???”
- “kein ETX” and “nicht ETX” refer to the words “kein” (no, none) and “nicht” (not) appearing as the lastword in the text.



## Default German Stopword List

Stopword lists are computed as the most common words in the language

aber	deiner	hier	meines	war	bis	einigem	jenen	so	würden	
alle	deines	hin	mit	waren	bist	einigen	jener	solche	zu	
allem	denn	hinter	muss	warst	da	einiger	jenes	solchem	zum	
allen	derer	ich	musste	was	damit	einiges	jetzt	solchen	zur	
aller	dessen	mich	nach	weg	dann	einmal	kann	solcher	zwar	
alles	dich	mir	nicht	weil	der	er	kein	solches	zwischen	
als	dir	ihr	nichts	weiter	den	ihn	keine	soll		
also	du	ihre	noch	welche	des	ihm	keinem	sollte		
am	dies	ihrem	nun	welchem	dem	es	keinen	sondern		
an	diese	ihren	nur	welchen	die	etwas	keiner	sonst		
ander	diesem	ihrer	ob	welcher	das	euer	keines	über		
andere	diesen	ihres	oder	welches	daß	eure	können	um		
anderem	dieser	euch	ohne	wenn	derselbe	eurem	könnte	und		
anderen	dieses	im	sehr	werde	derselben	euren	machen	uns		
anderer	doch	in	sein	werden	denselben	eurer	man	unse		
anderes	dort	indem	seine	wie	desselben	eures	manche	unsem		
anderem	durch	ins	seinem	wieder	demselben	für	manchem	unsen		
andern	ein	ist	seinen	will	dieselbe	gegen	manchen	unser		
anderr	eine	jede	seiner	wir	dieselben	gewesen	mancher	unses		
anders	einem	jedem	seines	wird	dasselbe	hab	manches	unter		
auch	einen	jeden	selbst	wirst	dazu	habe	mein	viel		
auf	einer	jeder	sich	wo	dein	haben	meine	vom		
aus	eines	jedes	sie	wollen	deine	hat	meinem	von		
bei	einig	jene	ihnen	wollte	deinem	hatte	meinen	vor		
bin	einige	jenem	sind	würde	deinen	hatten	meiner	während		

# Why stopwords were needed

- The reason why removing the stopwords was a bad idea, is that words like “kein” and “keine” were very influential in this data set.

# Example: Patient Joe

- The following open-ended question was asked in a web survey in a subset of the Dutch LISS panel.
- “Joe’s doctor told him that he would need to return in two weeks to find out whether or not his condition had improved. But when Joe asked the receptionist for an appointment he was told that it would be over a month before the next available appointment. What should Joe do?”

# Most influential variables

Variable	Translation	Counterproductive	Passive	Somewhat	Proactive
Number of words	Number of words	13.9	10.7	2.6	14.6
andere	other	13.8	0.2	0.0	0.1
staan	stand	0.0	1.8	0.0	8.2
telefonisch	telephonic	0.0	0.0	7.5	0.0
ergens	somewhere	0.0	0.0	6.7	0.0
bol_geen	bol_no	6.3	0.0	0.0	0.2
arts	doctor	0.0	4.5	0.4	6.2
met_de	with_the	0.0	4.3	0.1	6.0
bol_een	bol_a	4.8	0.0	0.1	0.0
naar_de	to_the	0.0	0.0	0.0	4.6
bol_niet	bol_not	4.2	0.2	0.0	0.2
ander_ziekenhuis	other_hospital	4.2	0.0	0.0	0.0
iemand	somebody	0.0	0.0	3.9	0.0
tussen	between	0.0	0.0	3.2	0.0
schuiven	Push/ fit in	0.0	0.0	3.1	0.0
conditie	(medical) condition	0.0	3.1	0.0	0.3
arts_zoeken	seek_doctor	3.0	0.0	0.0	0.0



# Counterproductive

- Counterproductive patients leave established care to go to another doctor/hospital or patient leaves without any appointment.
- words: “other” “other hospital”
- first word “no” “not”
  - e.g. “Geen afspraak maken.[...]” (Make no appointment)

# Passive

- Passive patients take no action that has a reasonable chance attaining patient's goal.
- absence of “doctor” “with the”
- presence of (medical) “condition”
  - e.g. “[...] wachten tot de arts klaar is met de volgende afspraak [...]” (wait until the doctor agrees with the next appointment)

# Somewhat proactive

- Somewhat proactive patients accepts the appointment but ask to be called
- “telephonic” and “somewhere” “between” “fit in”
  - e.g. [...] binnen de gestelde termijn er **tussen** door moeten **schuiven**”  
(Loosely: They have to **fit** him **in** in **between**.)

# Proactive

- Proactive patients take active steps towards getting an appointment in two weeks before leaving the doctor's office.
- “stand” and “doctor” “with the” and “near the”
- “er op **staan** dat er met de dokter wordt overlegd voor een afspraak over 14 dagen” (to **insist** there is a consultation . with the doctor within 14 days)

# References

## **Methodology open-ended questions**

- Schonlau, M., Couper M. Semi-automated categorization of open-ended questions. *Survey Research Methods*. August 2016, 10(2), 143-152.
- McLauchlan, C, Schonlau, M. Are Final Comments in Web Survey Panels Associated with Next-Wave Attrition? *Survey Research Methods*, Dec 2016, 10(3), 211-224.
- Gweon, H., Schonlau, M., Kaczmirek L., Blohm, M., Steiner, S. Three Methods for Occupation Coding Based on Statistical Learning. *Journal of Official Statistics* 2017, 33 (1), 101-122.

## **Stata Software**

- Schonlau, M., Guenther, N. Sucholutsky, I. Text mining using n-gram variables. *The Stata Journal*. Dec 2017, 17(4), 866-881.
- Guenther, N., Schonlau. M. Support vector machines. *The Stata Journal*. Dec 2016, 16(4), 917-937.
- Schonlau M. Boosted Regression (Boosting): An introductory tutorial and a Stata plugin. *The Stata Journal*, 2005; 5(3):330-354
- Randomforest (paper submitted)

# THE END

Contact info:

[schonlau at uwaterloo dot ca](mailto:schonlau@uwaterloo.ca)

[www.schonlau.net](http://www.schonlau.net)

I gratefully acknowledge funding from the Social Sciences and Humanities Research Council (SSHRC) of Canada