Preparations

- Experience Poll via Whova
  - Familiarity with Excel
  - Familiarity with R
  - Familiarity with Language Modeling
- Set up collaborative environment
  - Access to QR Code and Jamboards
- Set out sticky notes and permanent markers on tables

Accessibility

We are committed to providing a session that is accessible to the widest possible audience, regardless of technology or ability.

We are actively working to increase the accessibility and usability of our work and in doing so adhere to many of the available standards and guidelines.

Please do not hesitate to let one of our presenters know if you are in need of accommodations to view and engage with the content.

Natural Language Processing (NLP) Data Techniques with Symbolic & Nonsymbolic Mathematical Language

Discovery Research PreK-12 PI Meeting in Arlington, VA

Workshop facilitation by Cassondra Griger, M.S.Ed.

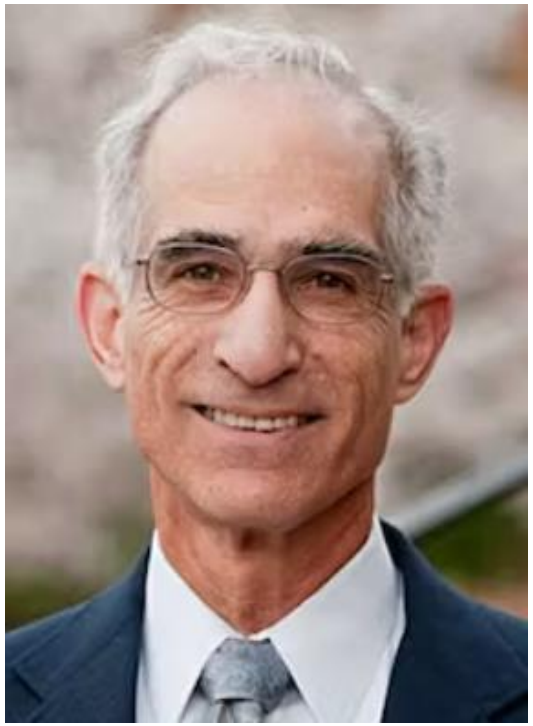
Supported by DRL-1813760 (PI: Yasemin Copur-Gencturk, Ph.D.)

June 29, 2023

USC Rossier School of Education

Science of Teaching Math

Yasemin Copur-Gencturk, Ph.D.
opurgen@usc.edu



Allan S. Cohen, Ph.D.
acohen@uga.edu

Chandra Orrill, Ph.D.
chandra.orrill@rethinklearning.com

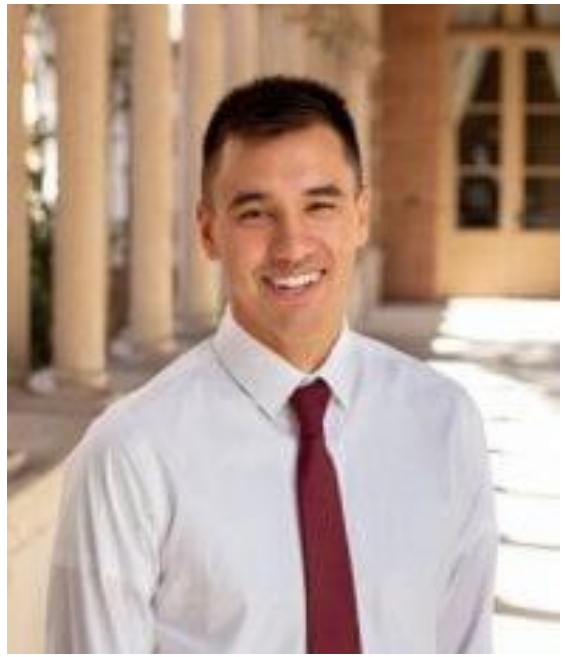
Jonathan Templin, Ph.D.
jonathan-templin@uiowa.edu

Science of Teaching Math

- ✓ teachers' acquisition of knowledge and pedagogical skills
- ✓ transfer of these skills into their practice, with special attention to issues of equity

@scienceofteach

Cassondra Griger, M.S.Ed.
Ph.D. Student
cgriger@uiowa.edu



John Ezaki
Ph.D. Candidate
jezaki@usc.edu

Cigdem Toptas, Ph.D.
Successful Defense
c.toptas@uga.edu

Sebnem Atabas, Ph.D.
Assistant Professor
sebnem.atabas@gmail.com

INFORMATION LIMITATION

Selected responses afford a limited observation of valuable, contextual information

MIXED METHODS

Constructed responses allow us to capture greater detail

NLP can qualitatively and quantitatively capture patterns of knowledge

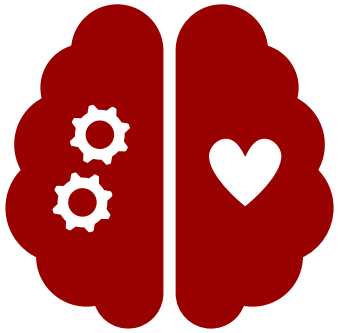
Objectives

… to **equip participants with data preprocessing techniques** specifically tailored to handle both symbolic and nonsymbolic mathematical language, from constructed response data.

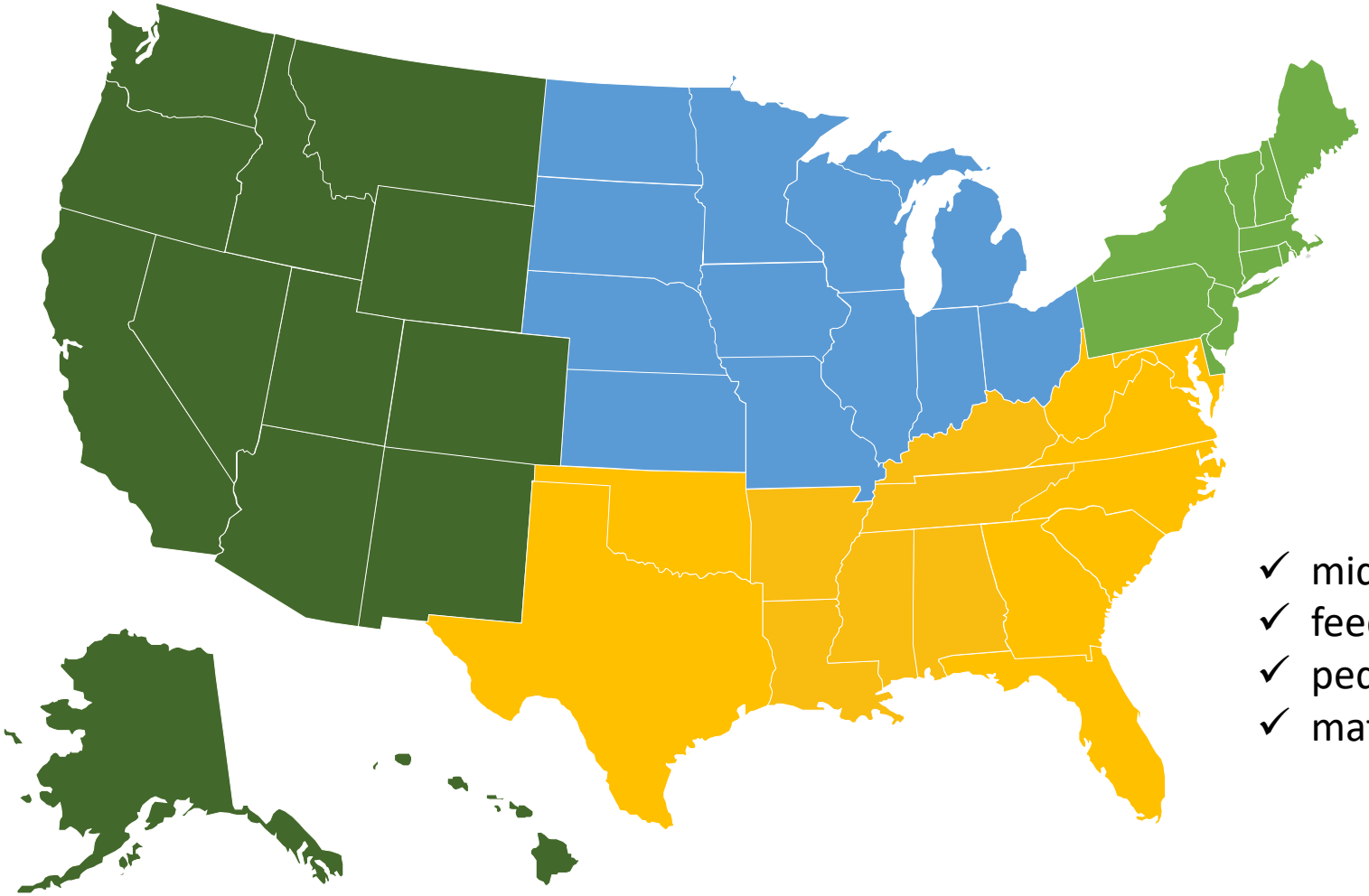
… to **gain knowledge and skills necessary to effectively process and analyze mathematical text data**, leveraging NLP methods to uncover insights and extract meaningful information from these specialized domains.

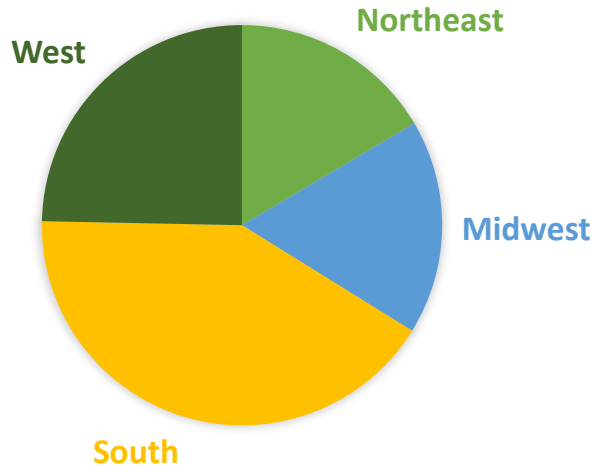
… to **explore practical applications of NLP in educational settings**, enabling them to make informed decisions and derive valuable insights from mathematical content.

Agenda



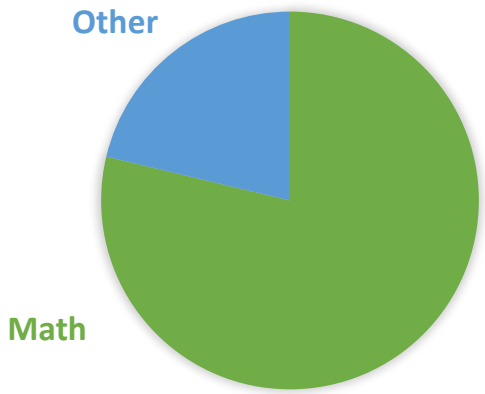
1. Experience Poll
2. Describing PCK Construct the Data
3. Experiential Processing Exercise
4. Preprocessing Data Techniques for Topic Modeling
5. Math Vocabulary Brainstorm Exercise
6. Mathematical Semantics: Symbolic vs Nonsymbolic Language
7. Questions & Discussion



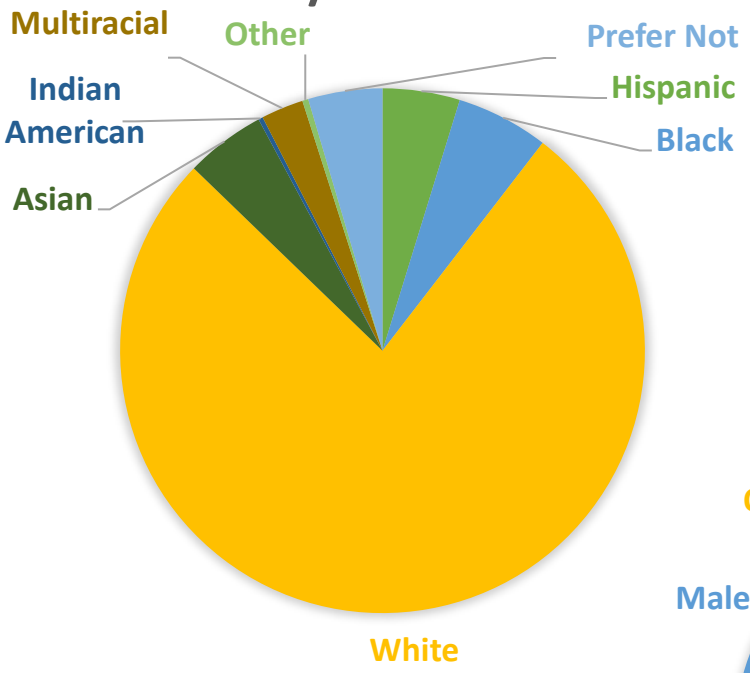


- ✓ middle school in-service teachers
- ✓ feedback to student written responses
- ✓ pedagogical content knowledge (PCK) schemas
- ✓ mathematical contexts of proportional reasoning

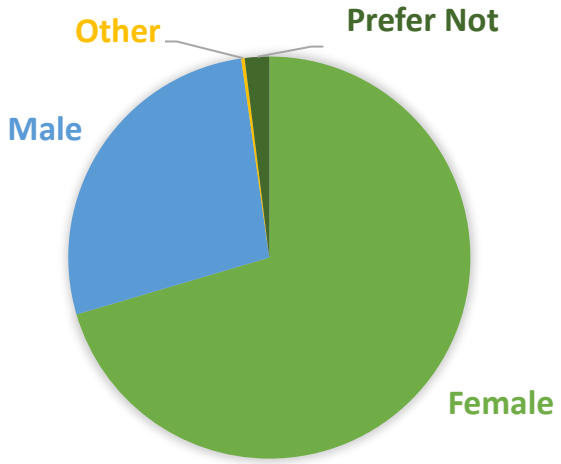
MATH CREDENTIALS



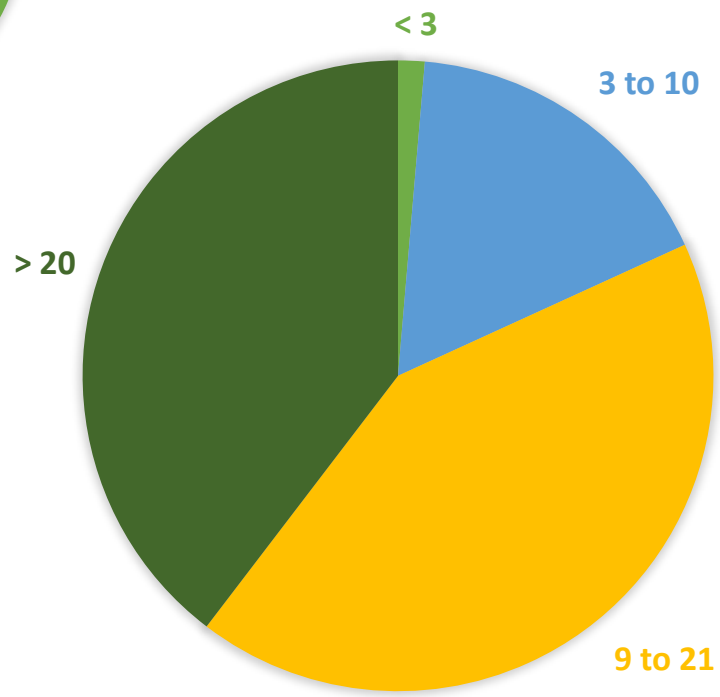
RACE/ETHNICITY



GENDER



YEARS OF EXPERIENCE



USC Rossier School of Education

Science of Teaching Math

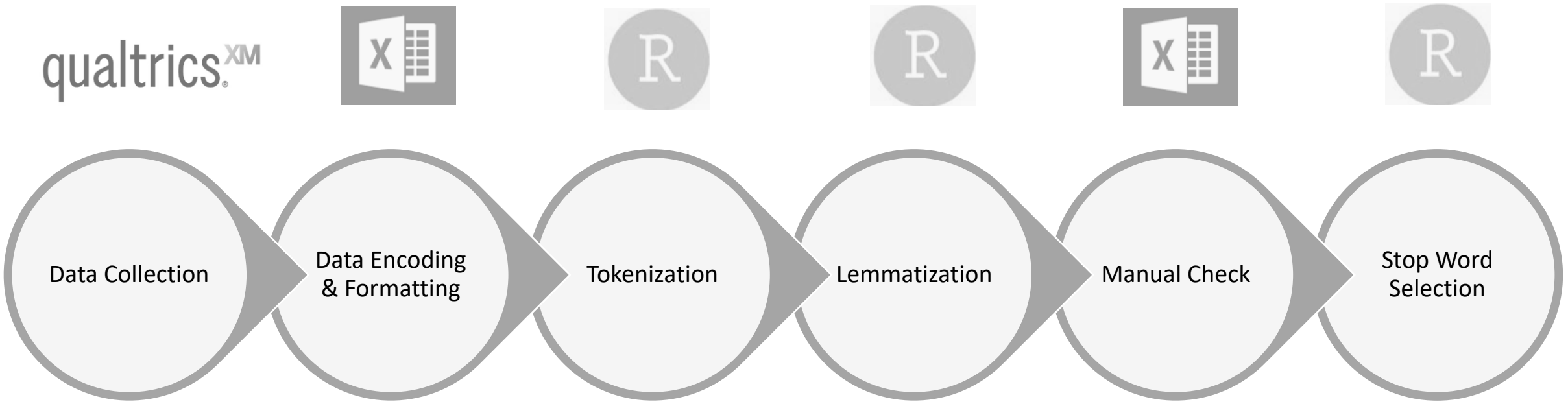
Pedagogical Content Knowledge (PCK)

PCK is an external and internal construct, as it is constituted by:

- ✓what a teacher knows
- ✓what a teacher does
- ✓the reasons for the teacher's actions

Baxer ve Lederman (1999, p. 158)
Gess-Newsome (1999)
Magnusson, Krajacik, & Borko (1999)
Van Driel, Bijaard & Verloop (2001)

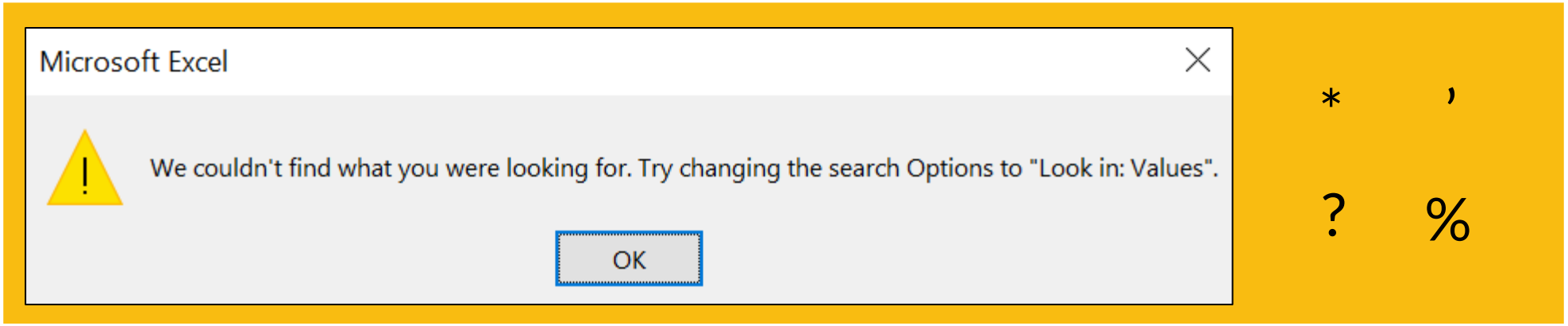
Stages of Processing Response Data



Data Encoding & Formatting

Exportation of raw data sometimes results in funky characters since Qualtrics can only export in CSV with UTF-8 encoding and TSV.

Without reformatting, special characters like 是 or ñ may not show correctly upon import. Or conversely, symbols are encoded as 是



```r
## locate "/" and change to "_" (specify number/number, versus word/word)
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                             pattern = "([\\d|\\w])\\/([\\d|\\w])",
                                             replacement = "\\1_\\2")
## locate ":" and change to "to"
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                             pattern = "([\\d|\\w])\\:([\\d|\\w])",
                                             replacement = "\\1to\\2")
## locate "-" and change to "" (specify that letters need to be on either side of the "-")
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                             pattern = "([\\d]) to ([\\d])",
                                             replacement = "\\1to\\2")
## locate "ft" and change to "foot" (do for all units of measures: ht, in, oz, ...)
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                             pattern = "[:space:]ft[:space:]?",
                                             replacement = " foot")
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                             pattern = "[:space:]ht[:space:]?",
                                             replacement = " height ")
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                            pattern = "[:space:]oz[:space:]?",
                                            replacement = " ounce ")
textdata$feedbackresponses = str_replace_all(textdata$feedbackresponses,
                                            pattern = "[:space:]cm[:space:]?",
                                            replacement = " centimeter ")
```

Tokenization

```
125  unigram_tokens = unnest_tokens(tbl = textdata,
126                                 output = word,
127                                 input = feedbackresponses,
128                                 drop = FALSE,
129                                 to_lower = TRUE)
```

Lemmatization

```
specialcases = function(term) {
  stems = lemmatize_words(x = tolower(term))

  for (i in 1:length(stems)) { # the last word presented in "" is the replacement word

    stems[i] = ifelse(stems[i] == "concentrate", "orangeconcentrate", stems[i])
    stems[i] = ifelse(stems[i] == "measurement", "measure", stems[i])
    stems[i] = ifelse(stems[i] == "numb", "number", stems[i])
    stems[i] = ifelse(stems[i] == "characteristics", "characteristic", stems[i])
    stems[i] = ifelse(stems[i] == "stair", "staircase", stems[i]) # case has two different meanings
    stems[i] = ifelse(stems[i] == "emphasize", "emphasis", stems[i])
    stems[i] = ifelse(stems[i] == "growth", "grow", stems[i])
    stems[i] = ifelse(stems[i] == "percentage", "percent", stems[i])
    stems[i] = ifelse(stems[i] == "ration", "ratio", stems[i])
    stems[i] = ifelse(stems[i] == "simply", "simple", stems[i])
    stems[i] = ifelse(stems[i] == "vs", "versus", stems[i])
    stems[i] = ifelse(stems[i] == "c.o.p", "constantofproportionality", stems[i])

    stems[i] = ifelse(stems[i] %in% c("juice", "oj"), "orangejuice", stems[i])
    stems[i] = ifelse(stems[i] %in% c("relation", "relationship"), "relate", stems[i])
    stems[i] = ifelse(stems[i] %in% c("flag", "pole"), "flagpole", stems[i])
    stems[i] = ifelse(stems[i] %in% c("light", "house"), "lighthouse", stems[i])
    stems[i] = ifelse(stems[i] %in% c("bargain", "hut"), "bargainhut", stems[i])

    stems[i] = ifelse(stems[i] %in% c("addend", "additive", "addend"), "add", stems[i])
    stems[i] = ifelse(stems[i] %in% c("subtraction", "subtractive"), "subtract", stems[i])
    stems[i] = ifelse(stems[i] %in% c("addend", "additive", "addend"), "add", stems[i])
```

Lemmatization

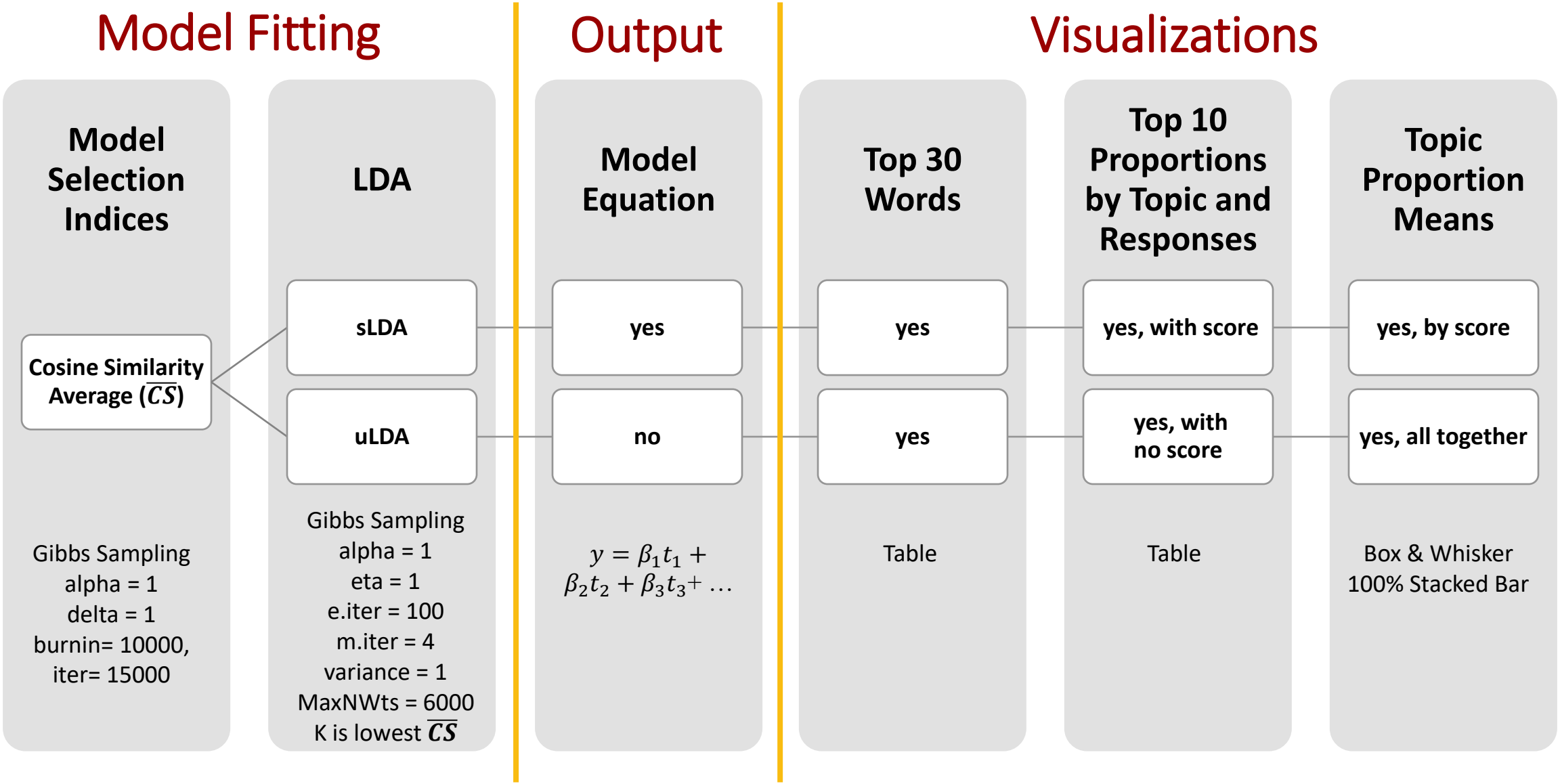
```
161     stems[i] = ifelse(stems[i] %in% c("multiplicative", "multiplicatively", "multiplication",
162                                       "multiplicity", "multiplier"), "multiply", stems[i])
163
164     stems[i] = ifelse(stems[i] %in% c("generalizable", "generalization", "generalize", "generally"), "general", stems[i])
165     stems[i] = ifelse(stems[i] %in% c("proportionality", "proportionally"), "proportional", stems[i])
166
167     stems[i] = ifelse(stems[i] %in% c("dont", "donts", "doesnt", "didnt",
168                                       "arent", "isnt", "havent",
169                                       "cant", "cannot", "couldnt", "shouldnt",
170                                       "wont", "werent", "wouldnt", "wasnt"), "not", stems[i])
171     }
172   stems # prints stems
173 }
174
175 ## combining columns from unigram dataframe and function
176 stemlist = cbind(unigram_tokens,
177                  stem = specialcases(unigram_tokens$word),
178                  finalstem = specialcases(unigram_tokens$word))
179
180 stemlist$finalstem = ifelse(stemlist$stem == "much", stemlist$word, stemlist$finalstem) # more/most - much - more/most
181 stemlist$finalstem = ifelse(stemlist$stem == "less", stemlist$word, stemlist$finalstem) # least - less - least
182 stemlist$finalstem = ifelse(stemlist$stem == "little", stemlist$word, stemlist$finalstem) # less - little - less
183 stemlist$finalstem = ifelse(stemlist$stem == "good", stemlist$word, stemlist$finalstem) # better/best - good - better/best
184 stemlist$finalstem = ifelse(stemlist$stem == "great", stemlist$word, stemlist$finalstem) # greater - great - greater
185 stemlist$finalstem = ifelse(stemlist$stem == "far", stemlist$word, stemlist$finalstem) # further - far - further
```

Stop Words

Type	Examples / Notes
generally high frequency words	i, we, he, she, they, you, your
articles, prepositions, conjunctions	the, a, and, but, or, so, however, since, because, any (-times, -thing, -one, -way), some (-times, -thing, -one), other, therefore, thus, back, up, through, upon, whether, yet
question starters	who, what, where, when, why, how, if, which
past and future verb tenses	should, could, would
extremely low frequency words (frequency <= 5)	made little difference in model fit (cite)

Data Structure

	A	B	C	D	E	F	G	H	I	J	K	
1	id	item	word	stem	finalstem	context_exclusion	stoplist_inclusion	CTn	CGn	JEnc	YCG	feedbackresponses
484218	R_2XhblWQR8qsv0A4	T21T	proporiton	proporiton	proportion	0	0					They are not unde
484219	R_1n0GqaTyY8gWLjB	T21T	proporiton	proporiton	proportion	0	0					They dont underst
484220	R_3O66AkteFRoIFSa	T78T	proporiton	proporiton	proportion	0	0					Daniel is looking at
484221	R_300JFOUIzibIX3J	V2T	proportio	proportio	proportion	0	0					I would explain tha
484222	R_3ESZuimLOI6hwyz	T79T	proportion	proportion	proportion	0	0					With a proportion
484223	R_tSUUQ7MrTvyE9BD	T22T	proportion	proportion	proportion	0	0					probing questions
484224	R_tSUUQ7MrTvyE9BD	T23T	proportion	proportion	proportion	0	0					Guiding the kids to
484225	R_tSUUQ7MrTvyE9BD	T23T	proportion	proportion	proportion	0	0					Guiding the kids to
484226	R_tSUUQ7MrTvyE9BD	T79T	proportion	proportion	proportion	0	0					You could put both


Model Fitting
Output
Visualizations
Model Selection Indices
LDA
Model Equation
Top 30 Words
Top 10 Proportions by Topic and Responses
Topic Proportion Means
Cosine Similarity Average ($\overline{CS}$)
sLDA
uLDA
yes
no
yes
yes
yes, with score
yes, with no score
yes, by score
yes, all together
Gibbs Sampling
alpha = 1
delta = 1
burnin= 10000,
iter= 15000
Gibbs Sampling
alpha = 1
eta = 1
e.iter = 100
m.iter = 4
variance = 1
MaxNWts = 6000
K is lowest $\overline{CS}$
$y = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \ldots$
Table
Table
Box & Whisker
100% Stacked Bar

Descriptives

Corpus: Responses & Words

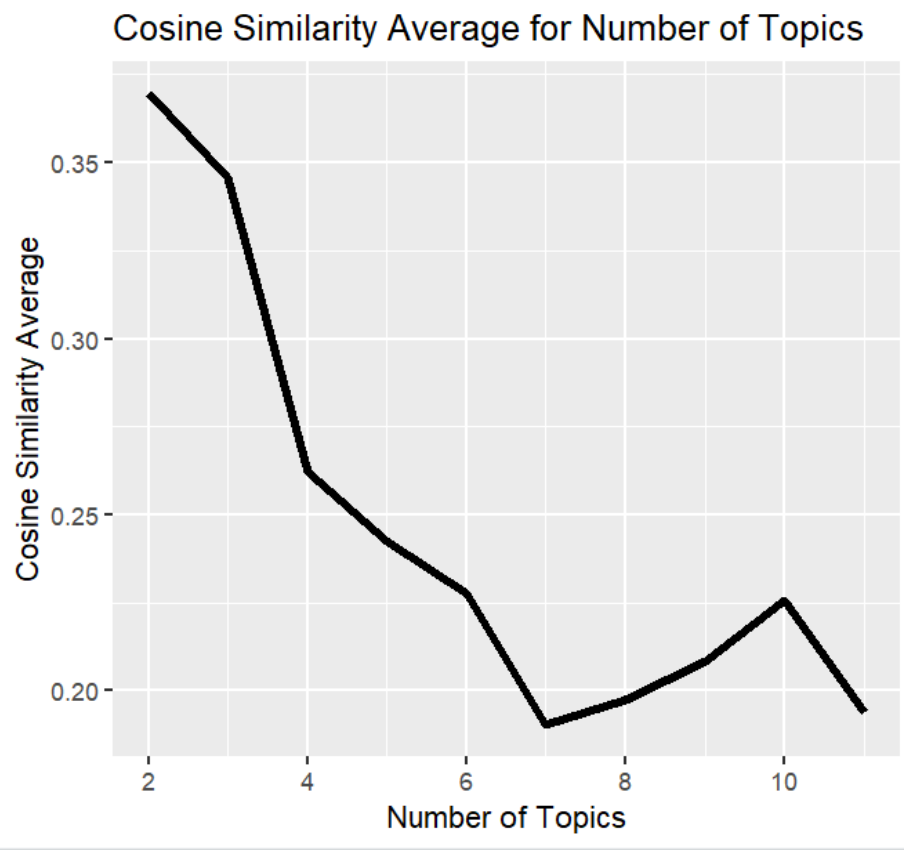
Cleaning	Number of Total Words	Number of Words	Number of Responses	Average Length
Pre	31599	1430	1072	29.47668
Post	15405	1239	1072	14.37034

MC Score Distribution

Score	Frequency
0	360
1	712

Model Selection Indices

candidate_k	model_sel_cs
7	0.190203
11	0.193853
8	0.197539
9	0.208324
10	0.22552
6	0.227434
5	0.242556
4	0.262265
3	0.345814
2	0.369588



Latent Dirichlet Allocation (LDA)

```
220 uni_slda_mod = slda.em(documents = ldaData$documents,
221                          K = Kvalue,
222                          vocab = ldaData$vocab,
223                          num.e.iterations = 100,
224                          num.m.iterations = 4,
225                          alpha = 1,
226                          eta = 1,
227                          params = params,
228                          variance =1,
229                          MaxNWts=6000,
230                          logistic=F,
231                          annotations = as.integer(unigram_documents_slda$score),
232                          method = "sLDA")
```

Model Equation

$$y = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \ldots$$

Topics	Estimate	S.E.	t	p
Beta_1	1.022	0.039	26.255	0
Beta_2	-0.095	0.044	-2.175	0.03
Beta_3	0.282	0.051	5.495	0
Beta_4	1.131	0.04	28.115	0
Beta_5	1.168	0.048	24.393	0
Beta_6	0.768	0.047	16.454	0

Topic 1	Proportion	Topic 2	Proportion	Topic 3	Proportion	Topic 4	Proportion	Topic 5	Proportion	Topic 6	Proportion
difference	0.13	dimension	0.076	side	0.171	ratio	0.173	close	0.138	square	0.111
three	0.088	each	0.064	rectangle	0.12	close	0.116	one	0.113	fifty	0.108
small	0.066	three	0.063	square	0.115	length	0.098	ratio	0.075	47	0.103
larger	0.055	same	0.062	equal	0.061	square	0.084	47_50	0.063	xdimension	0.055
inch	0.052	width	0.059	length	0.061	one	0.081	square	0.036	twenty	0.046
less	0.051	rectangle	0.044	look	0.042	width	0.073	divide	0.027	seventeen	0.037
between	0.047	length	0.043	four	0.029	1to1	0.053	17_20	0.024	most	0.027
more	0.04	square	0.037	longer	0.02	rectangle	0.028	37_40	0.018	rectangle	0.021
dimension	0.035	shape	0.028	three	0.017	47to50	0.025	0.94	0.018	94	0.021
appear	0.024	unit	0.025	each	0.016	look	0.023	compare	0.018	look	0.018
look	0.024	increase	0.023	short	0.016	side	0.017	27_30	0.017	area	0.018
make	0.021	inch	0.022	more	0.016	47	0.014	dimension	0.017	thirty	0.017
length	0.02	height	0.02	close	0.013	most	0.013	fifty	0.016	unit	0.017
rectangle	0.018	equal	0.018	two	0.013	fifty	0.013	equal	0.015	perfect	0.015
big	0.017	not	0.018	same	0.012	number	0.006	94	0.015	forty	0.015
notice	0.015	proportional	0.018	most	0.011	dimension	0.006	side	0.014	each	0.015
percent	0.015	similar	0.018	increase	0.01	choose	0.005	0.85	0.013	percent	0.013
unit	0.015	different	0.014	dimension	0.01	1.06	0.005	fraction	0.011	37	0.012
number	0.01	only	0.012	not	0.007	more	0.005	85	0.011	number	0.012
measure	0.01	think	0.011	example	0.007	0.94	0.004	whole	0.009	27	0.011
largest	0.008	apart	0.01	47	0.006	equivalent	0.004	0.925	0.009	find	0.011
only	0.008	make	0.009	choice	0.006	give	0.003	number	0.009	compare	0.01
size	0.007	size	0.009	option	0.005	greatest	0.003	0.9	0.008	85	0.009
two	0.007	ten	0.009	less	0.005	whole	0.003	decimal	0.007	dimension	0.008
3_50	0.006	question	0.008	angle	0.005	equal	0.003	last	0.007	multiply	0.008
least	0.006	number	0.008	right	0.004	between	0.003	1_1	0.007	ninety	0.007
significant	0.006	away	0.008	answer	0.004	50_47	0.003	second	0.007	close	0.006
even	0.005	oval	0.006	congruent	0.004	perfect	0.003	value	0.006	340	0.006
relative	0.005	look	0.006	none	0.004	approach	0.002	most	0.006	high	0.006
miss	0.005	amount	0.006	small	0.003	compare	0.002	give	0.006	92.5	0.005

Top Proportions by Topic and Responses

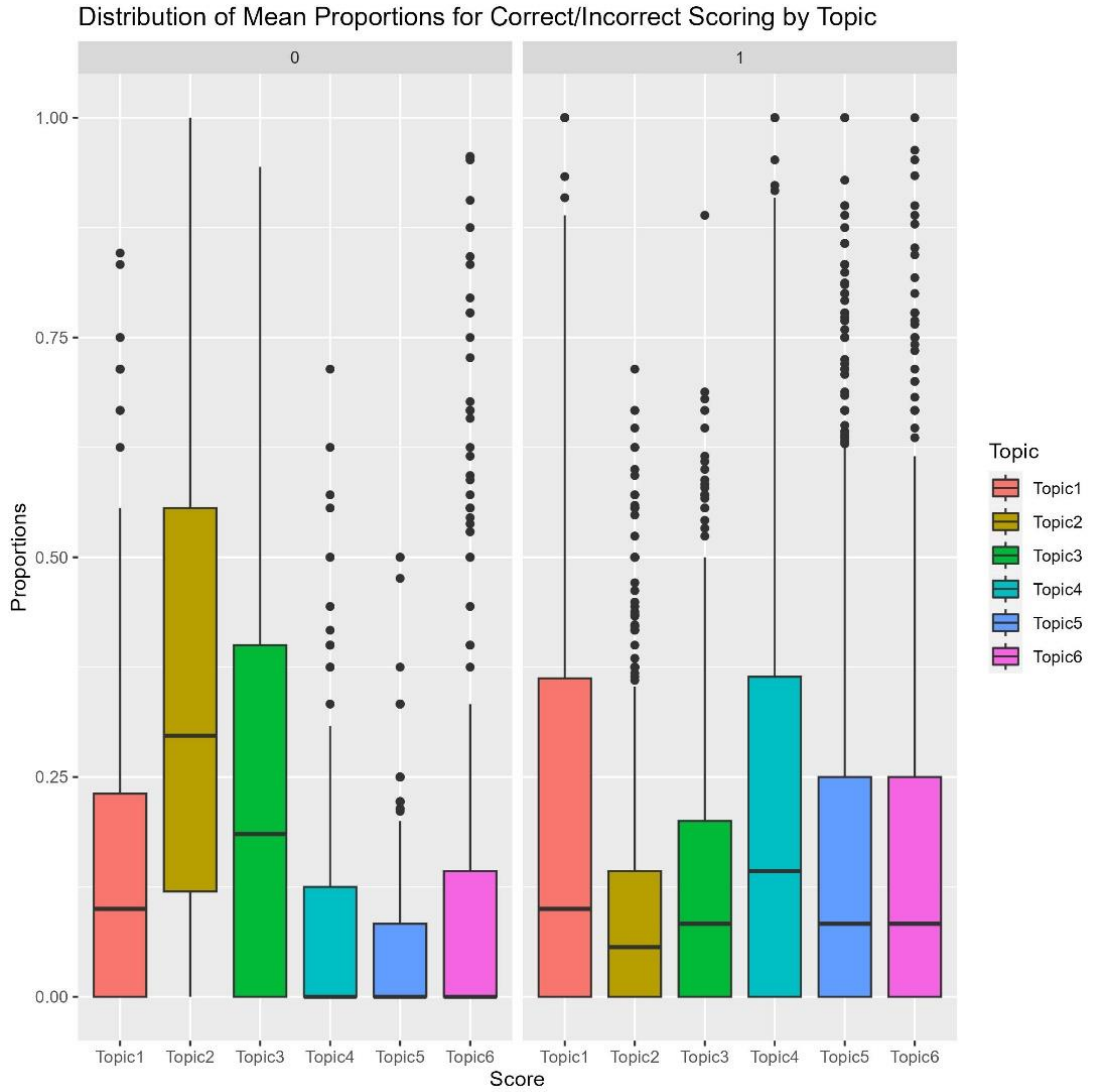
Topic 1

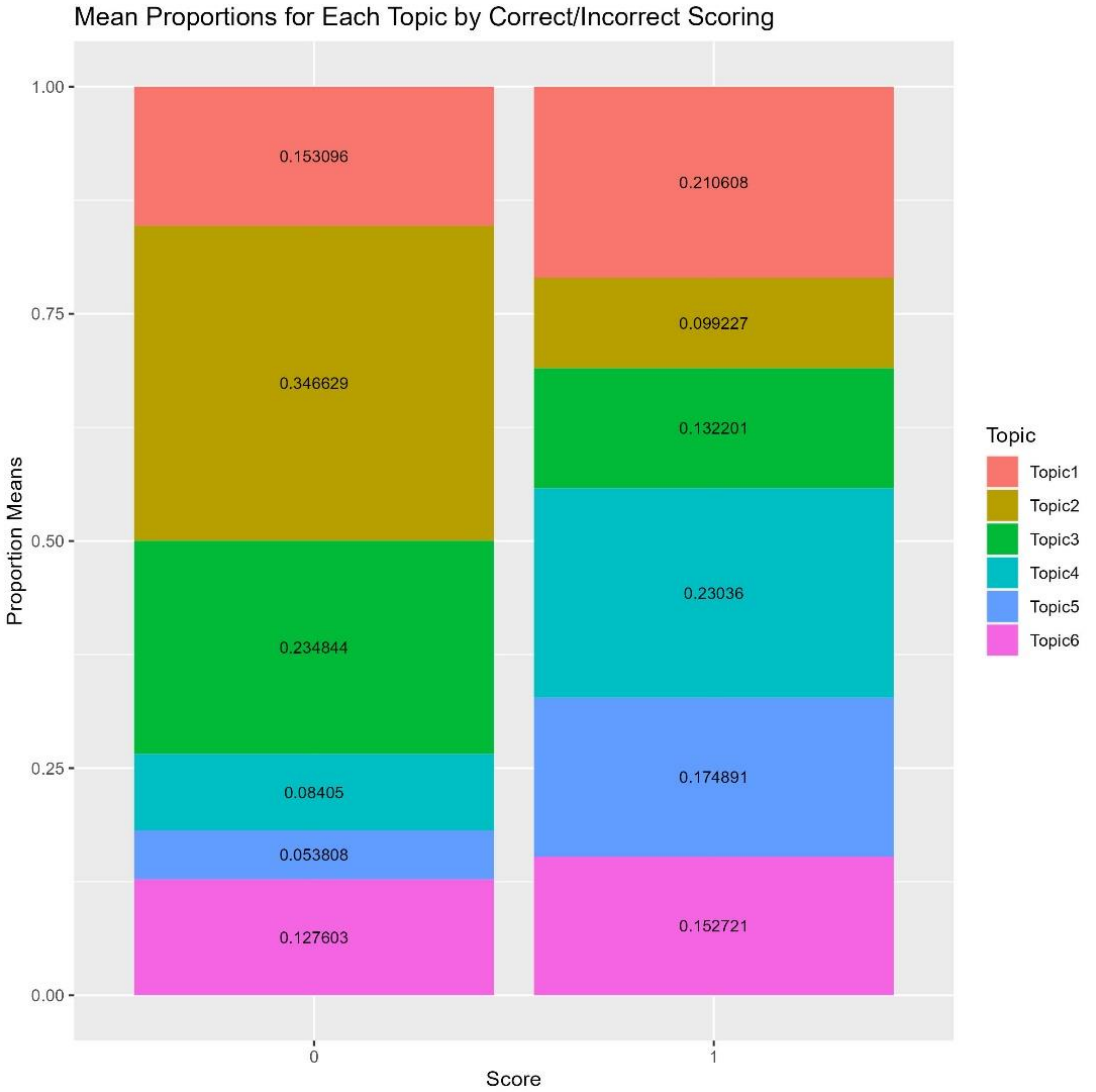
id	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	response	score
R_1GEcIHSwdm0fsjK	1	0	0	0	0	0	As the dimensions get larger, the relative difference of 3 units will make less of a difference compared to the smaller dimensions	1
R_1PeeXP08shYRUzD	1	0	0	0	0	0	Although they all have a difference of 3 units, the larger the size, the less noticeable the change.	1
R_2Y3Tp7lddDk9re3	1	0	0	0	0	0	They are all 3 inches different so the larger the numbers the smaller the difference in percentage.	1
R_2amDfbLAOHntHxn	1	0	0	0	0	0	The bigger the dimension gets, the less noticeable the difference of 3 appears.	1
R_xsaR0JNgTagZkn7	1	0	0	0	0	0	The difference of three inches would appear less significant as the numbers get larger.	1
R_xztavohLZVG61Xz	1	0	0	0	0	0	Much like the oval problem earlier, the larger the dimensions get the 3 unit difference will appear less and less.	1
R_30qnBSaI0pzM8Aq	0.933	0	0	0	0.067	0	As with the previous circle problem the bigger you get the less the the 3 unit difference makes. It becomes more and more difficult to see the difference as they get bigger	1
R_p5eErWswmp6u26R	0.909	0	0	0	0	0.091	A difference of 3 inches is going to be much more noticeable on a smaller object than it would a larger object.	1
R_2dvkeSHRRCRz0Z8	0.889	0.111	0	0	0	0	The bigger the rectangle the less you notice those 3 inches of difference between dimensions.	1
R_3FRIvdL7b9VHOud	0.867	0.067	0	0	0	0.067	"When you compare the dimensions, The difference in size is the least	1

Top Proportions by Topic and Responses

Topic 2

id	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	response	score
R_1Kx3yCfAGgImq21	0	1	0	0	0	0	Each width is exactly 3 units shorter than the length so they will be similar	0
R_26hhQzUbUZTqg4s	0	1	0	0	0	0	All the dimensions are 3 away from each other so they will be similar in shape	0
R_2B4XxLemv3S6e4R	0	1	0	0	0	0	the width for each rectangle is 3 inches shorter than the length for each rectangle so the dimensions are proportional for each.	0
R_3F40VNxJW8w4EOn	0	1	0	0	0	0	THAY ALL WOULD LOOK SIMIIAR AS THEY DIFFER BY THE SAME AMOUNT THUS THEY ARE IN A PROPORTIONAL RRELATIONSHIP	0
R_3I6PwrO2gN9D28I	0	1	0	0	0	0	The length and the width both increase at the same intervals (increase dimensions).	0
R_3M5zi6ZrFOUHqN7	0	1	0	0	0	0	I think they would be the same as theyre all 3 different, but im not really sure.	0
R_3lS0rC1mo0QTJtT	0	1	0	0	0	0	All the rectangles are similar due to the measurements being used are all 3 apart	0
R_3rTmVxEJ24GFxct	0	1	0	0	0	0	Proportionally they are all the same amount away from having equal dimensions	0
R_PNvqQWD5S5dFUrL	0	1	0	0	0	0	All of the rectangles are increasing size at the same rate and their dimensions are similar so they would all hold the same shape.	0
R_2V9cYg2tG5CkprJ	0	0.9	0	0	0	0.1	The increase in each length and each width are consistent. Each is an increase by 10 units.	0





USC Rossier School of Education

Science of Teaching Math

Math Vocabulary Brainstorm Exercise

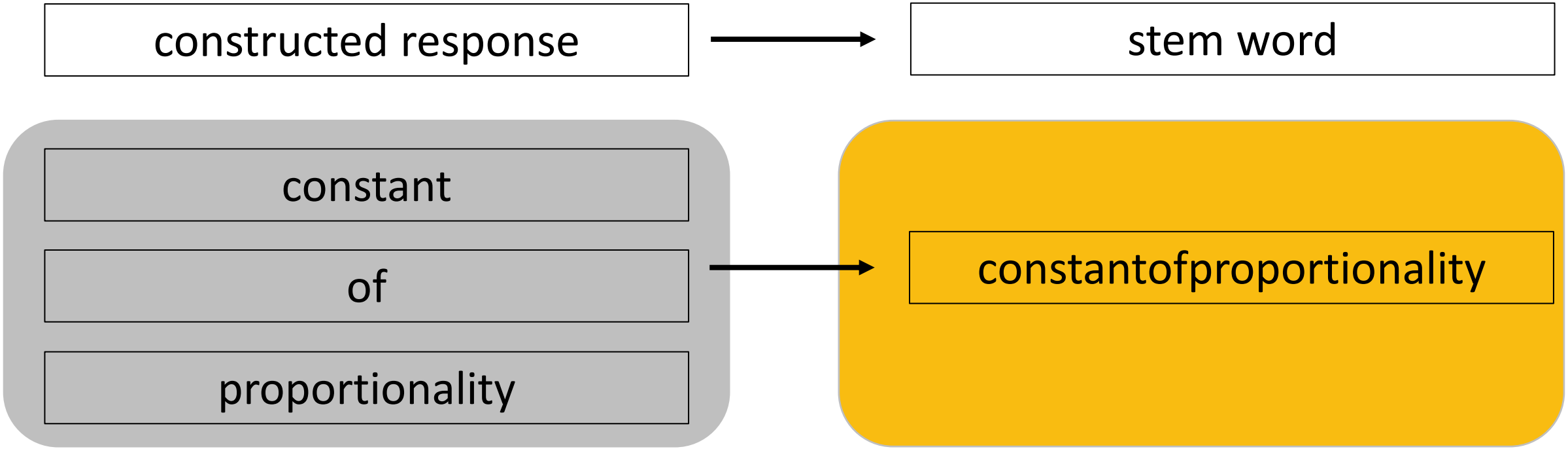
USC Rossier School of Education

Science of Teaching Math

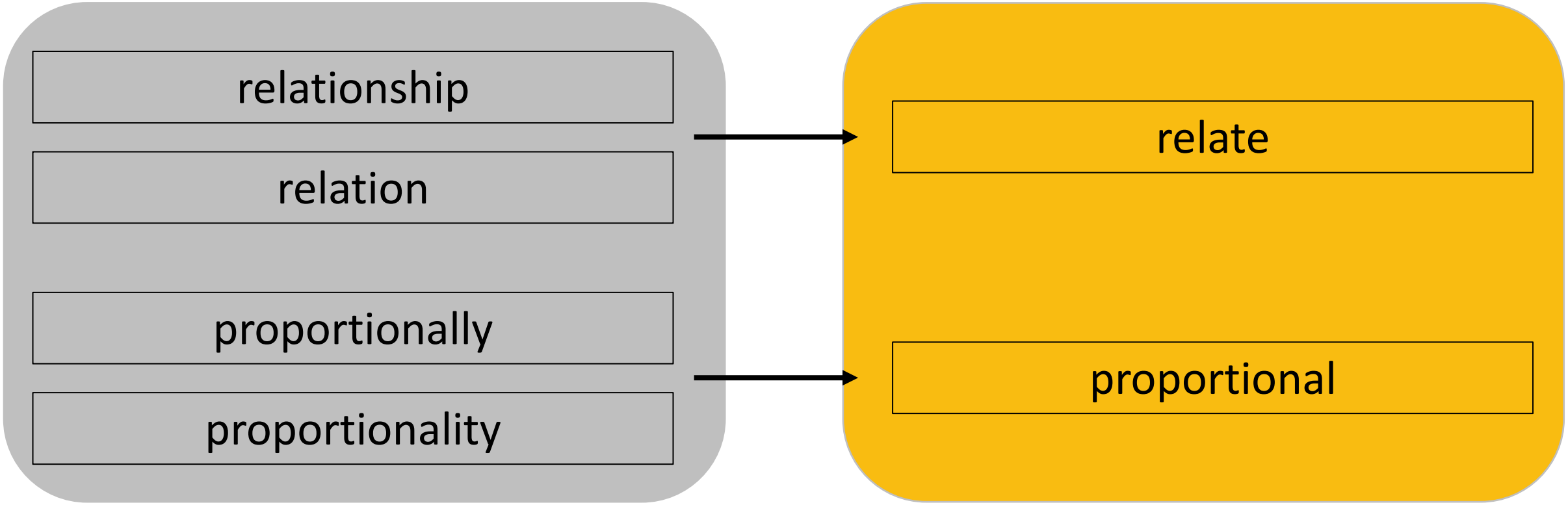
Math Vocabulary

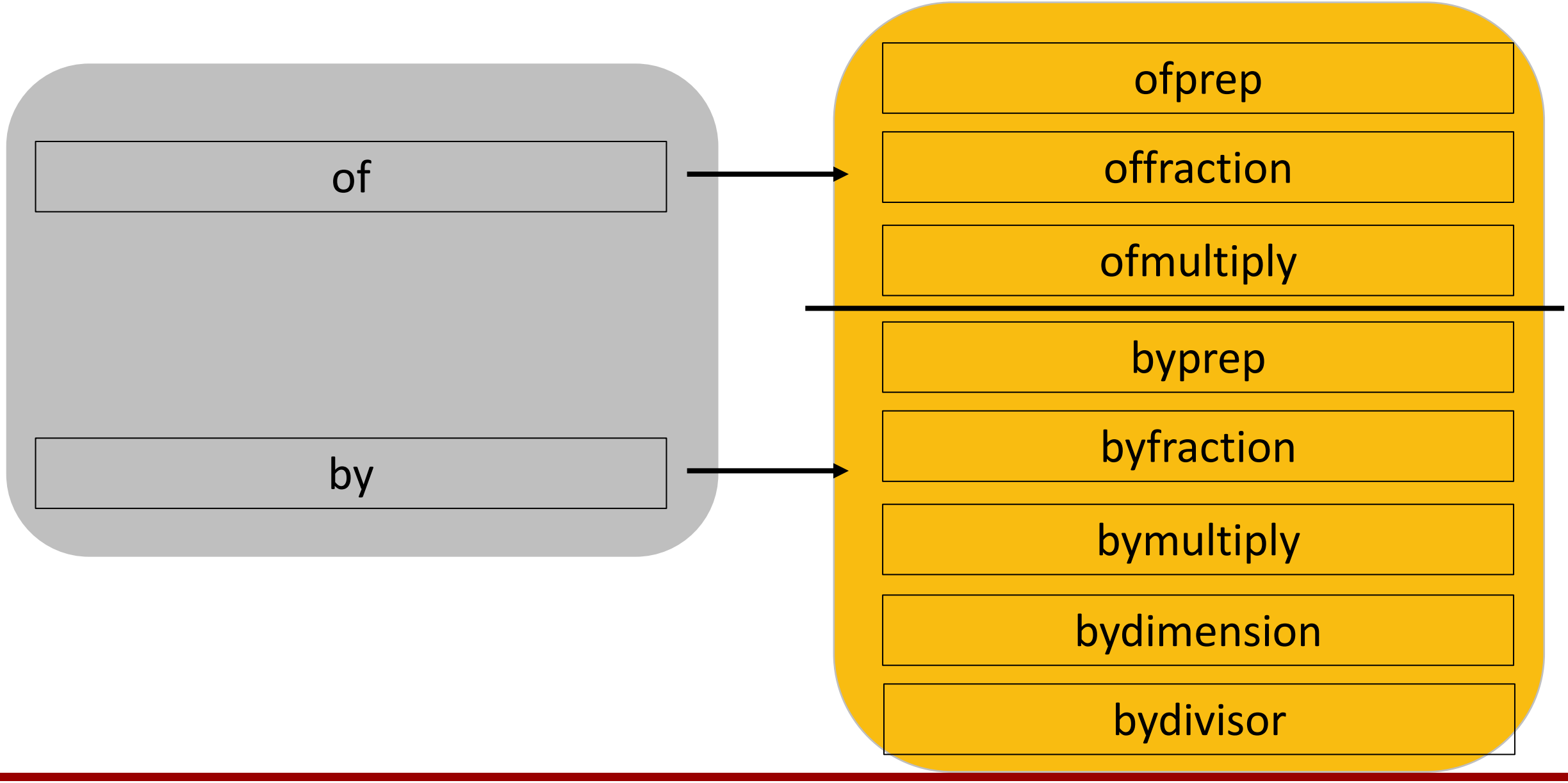
	Symbolic	Nonsymbolic
Addition	+	and, add x and y, by
Subtraction	-	from, subtract x by y, take away
Multiplication	* x	of, by
Division	/	by, over
Order / Sequence	1^{st} 2^{nd} 3^{rd}	first, second, third
Parentheses / Coordinates	() [] {}	
Exponents	^0 ^{n}	raised to the power, squared, n-th
Fraction	/	of, over, n-th
Decimal	. ,	tens, ones, hundredths, thousandths
Percent	% 0.00 - 000.00	percent, out of 100
Ratio	: /	to, the ratio of x to y
Proportion	x/y = z/a	x over y is to z over a
Equation	=	equals, is
Variables	a, b, c, x, y, z	variable a

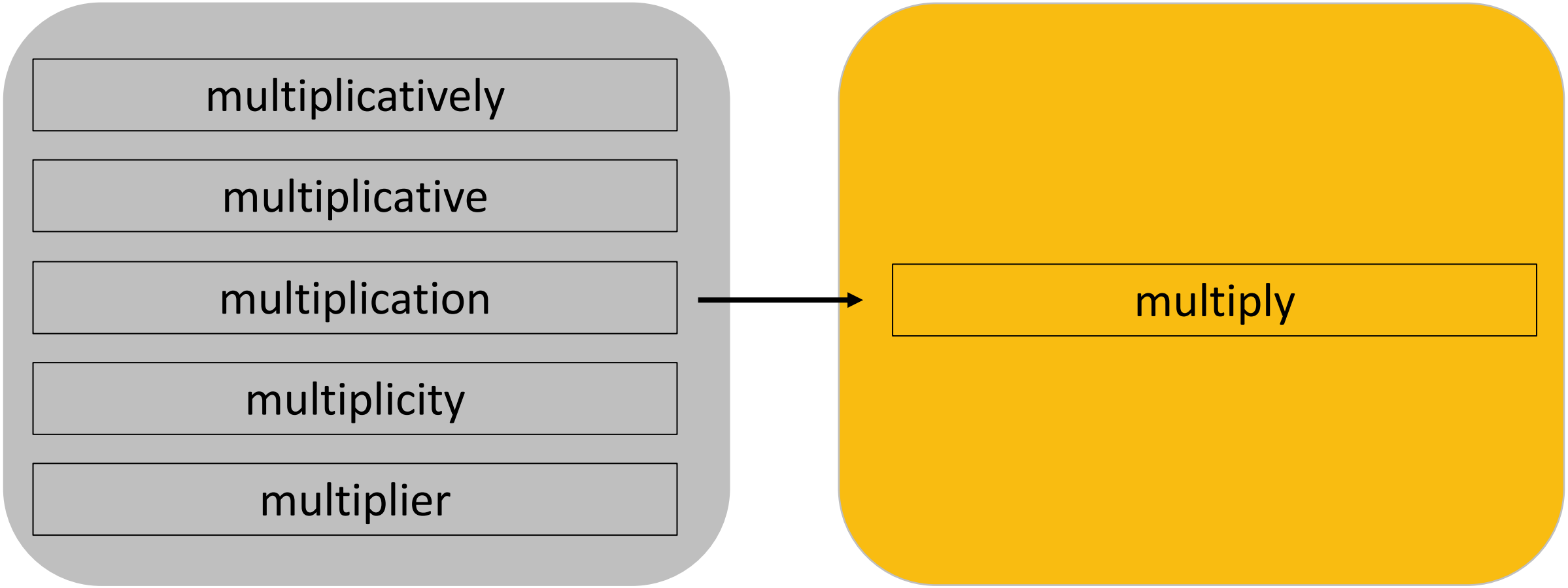
Many concepts can be understood with unigrams, such as ratio, fraction, proportion, etc. However, sometimes multigrams are required. That is, more than 1 word is needed to understand the context.



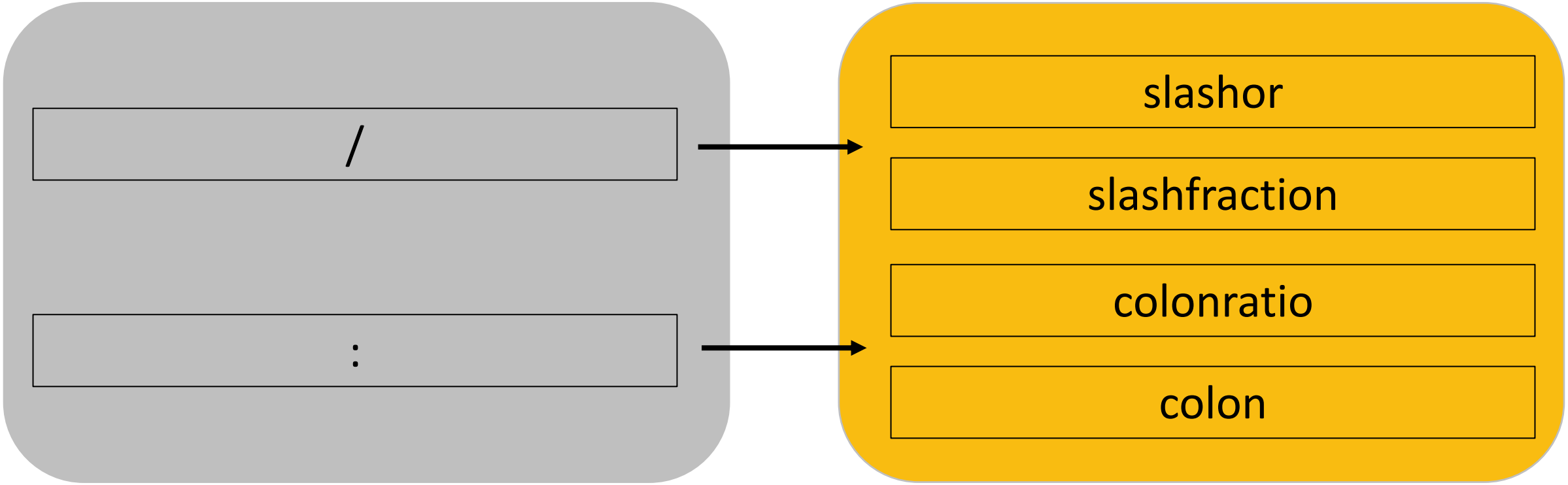
Nonsymbolic Context

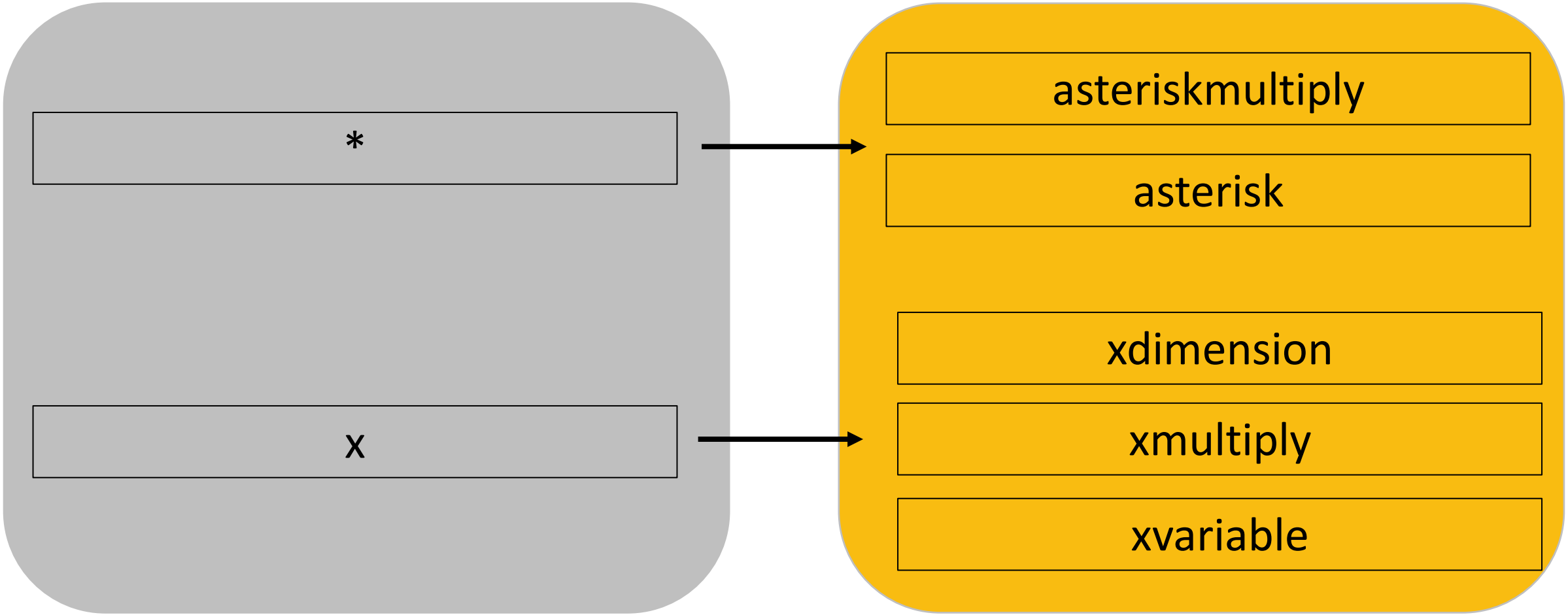



of
by
ofprep
offraction
ofmultiply
byprep
byfraction
bymultiply
bydimension
bydivisor

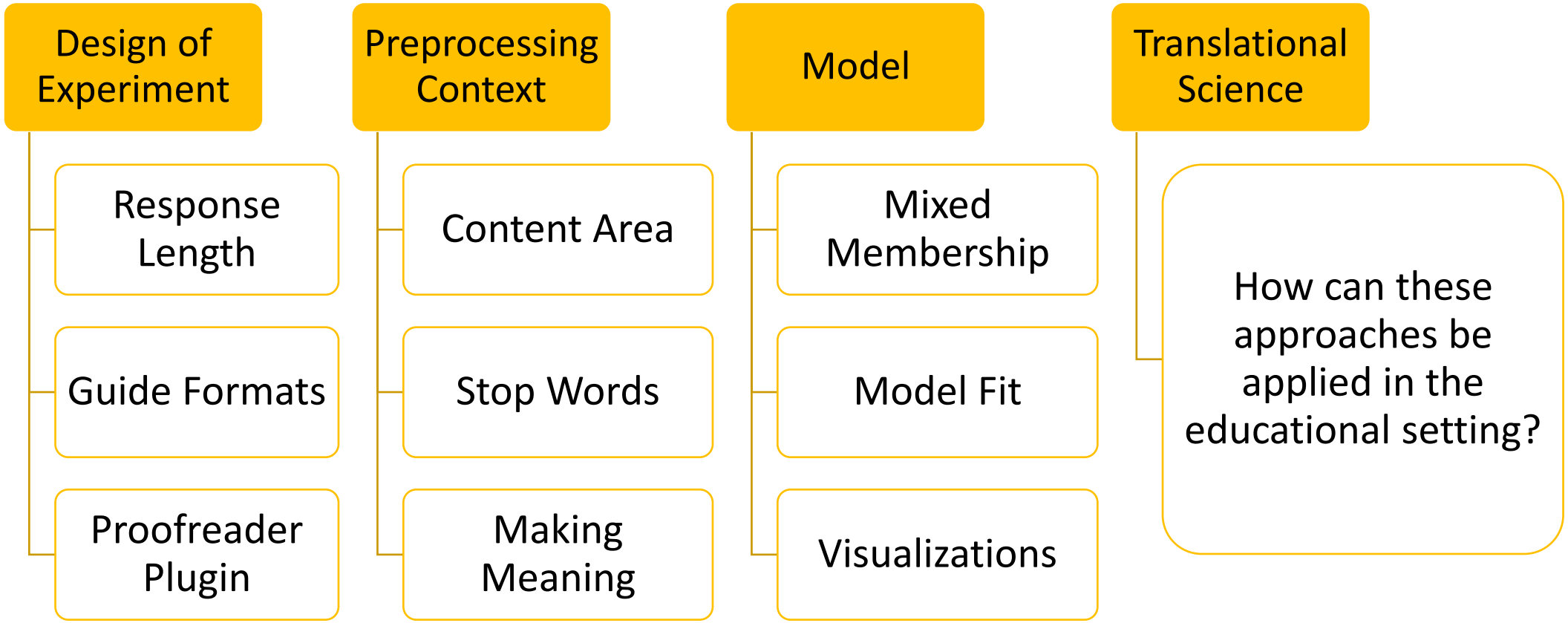

multiplicatively
multiplicative
multiplication
multiplicity
multiplier
multiply

Symbolic Context



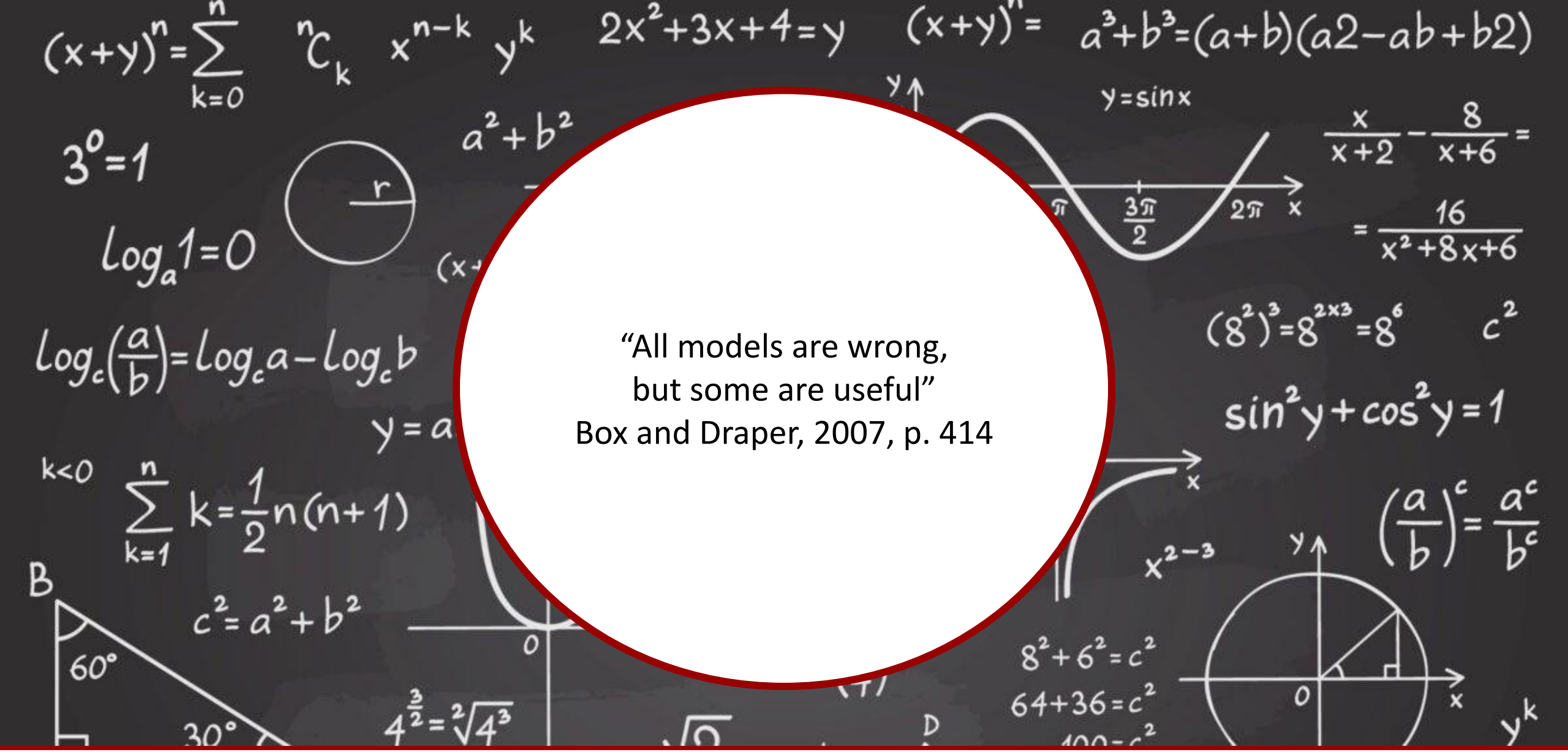

*
x
asteriskmultiply
asterisk
xdimension
xmultiply
xvariable

Research Considerations



Classroom Considerations

Gouet, Carvajal, Halberda, & Peña (2020)
Newcombe, Levine, & Mix, (2015)
Price & Fuchs (2016)
Price & Wilkey (2017)
Schneider, Beeres, Coban, Merz, Schmidt, Stricker, & De Smedt (2016)
Tikhomirova, Kuzmina, Lysenkova, & Malykh (2019)



"All models are wrong,
but some are useful"
Box and Draper, 2007, p. 414

Thank-you

USC Rossier School of Education

Science of Teaching Math

Machine Learning (ML) & Deep Learning (DL)

AI is the **concept of mimicking the performance of human intelligence.**

ML gives **machines the ability to learn from and make decisions or predictions to identify patterns and make decisions based on large data.**

DL relies on **neural networks with many layers.**