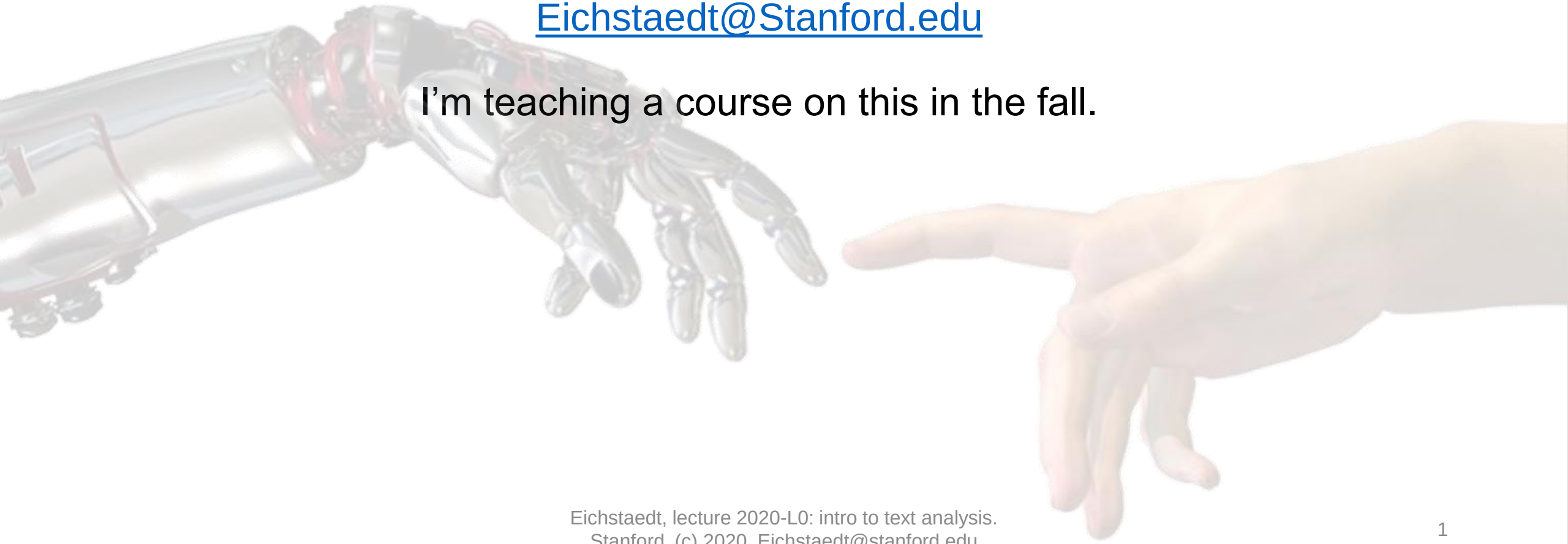
Introduction to Modern Language Analysis
(and some Machine Learning)

Johannes Eichstaedt
Eichstaedt@Stanford.edu

I'm teaching a course on this in the fall.



Goals

- Give you a quick overview of the dominant approaches, as an orientation.

- Build a first set of intuitions about the strengths and weaknesses of these approaches.

- Give you enough of an introduction to see if you want to take my applied course on this in the fall.

Outline

- Intro to language Variables
- General language analysis pipeline
- TOP DOWN: LIWC dictionaries as features
- BOTTOM UP: words as features
- BOTTOM UP: topics as features

- Intro to Machine learning
- regularized regression

- Preview of Fall Course

Why care about text?


2.27 billion


0.32 billion


1 billion


1 billion


0.46 billion

(4.5 billion+ monthly-active users)

Social media is the largest longitudinal and cross-sectional dataset in human history.

Most of it is text.

Text encodes behaviors, thoughts, emotions -- and health

You can understand psychological constructs with text

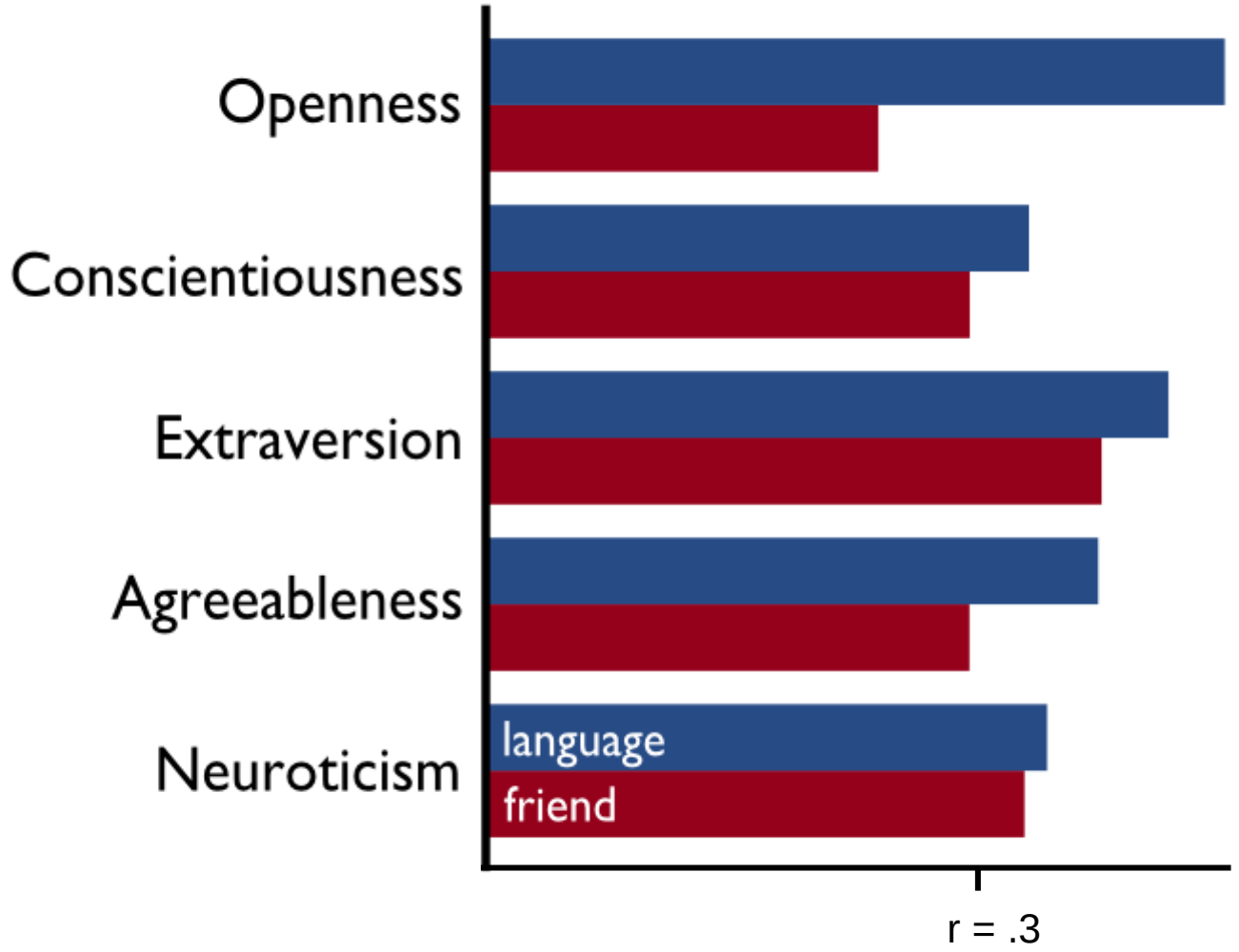
Language of **Agreeableness**

Schwartz et al., 2013, PLoSone

You can measure psychological constructs with text

Predicting Personality: **Accuracy**

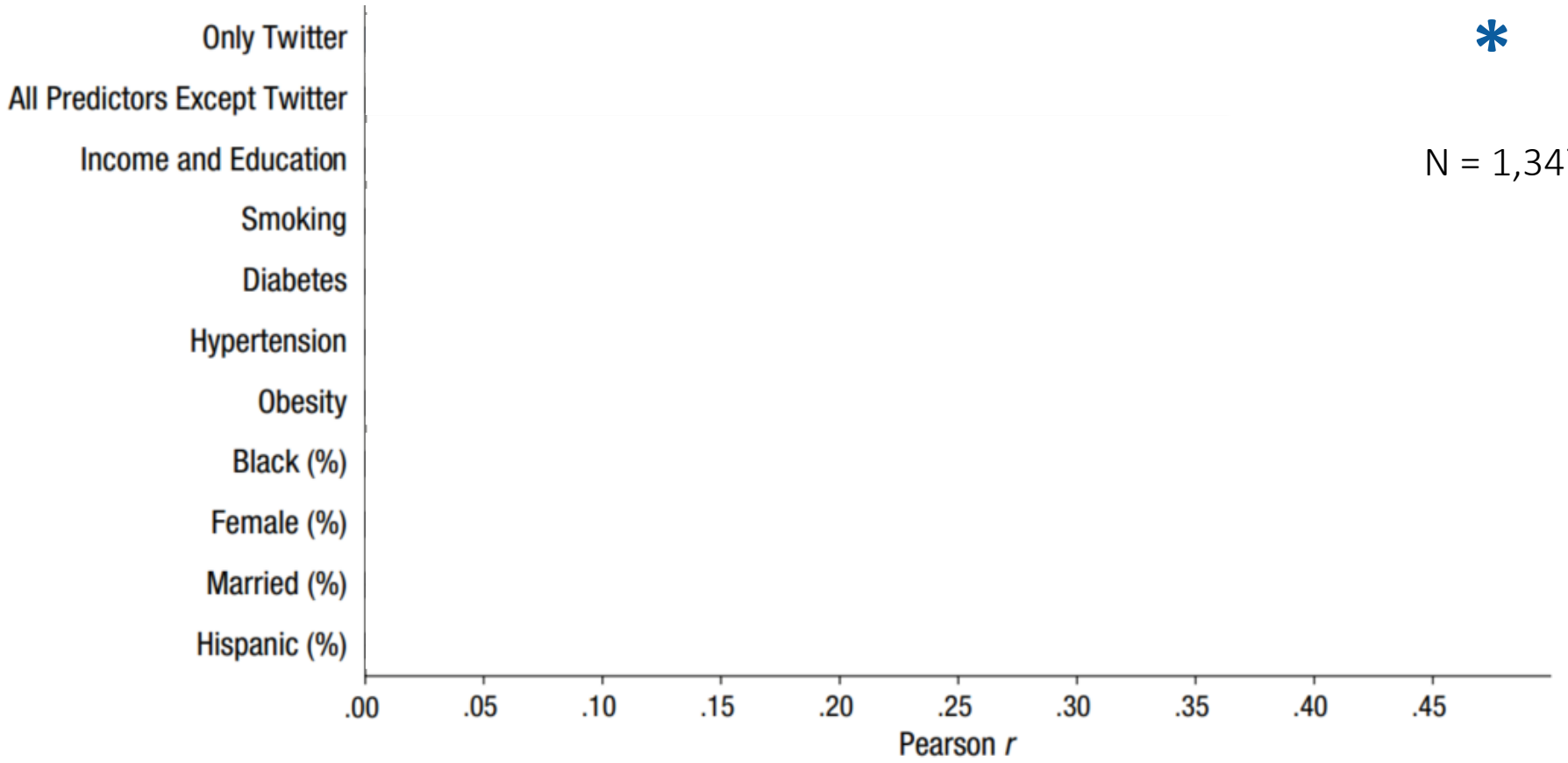
N = ~70,000 people



Park, Schwartz, Eichstaedt et al., 2014, *JPSP*

You can also measure health through text

Predicting county heart disease mortality: **Accuracy**



* **2009-2010**

N = 1,347 counties with Twitter

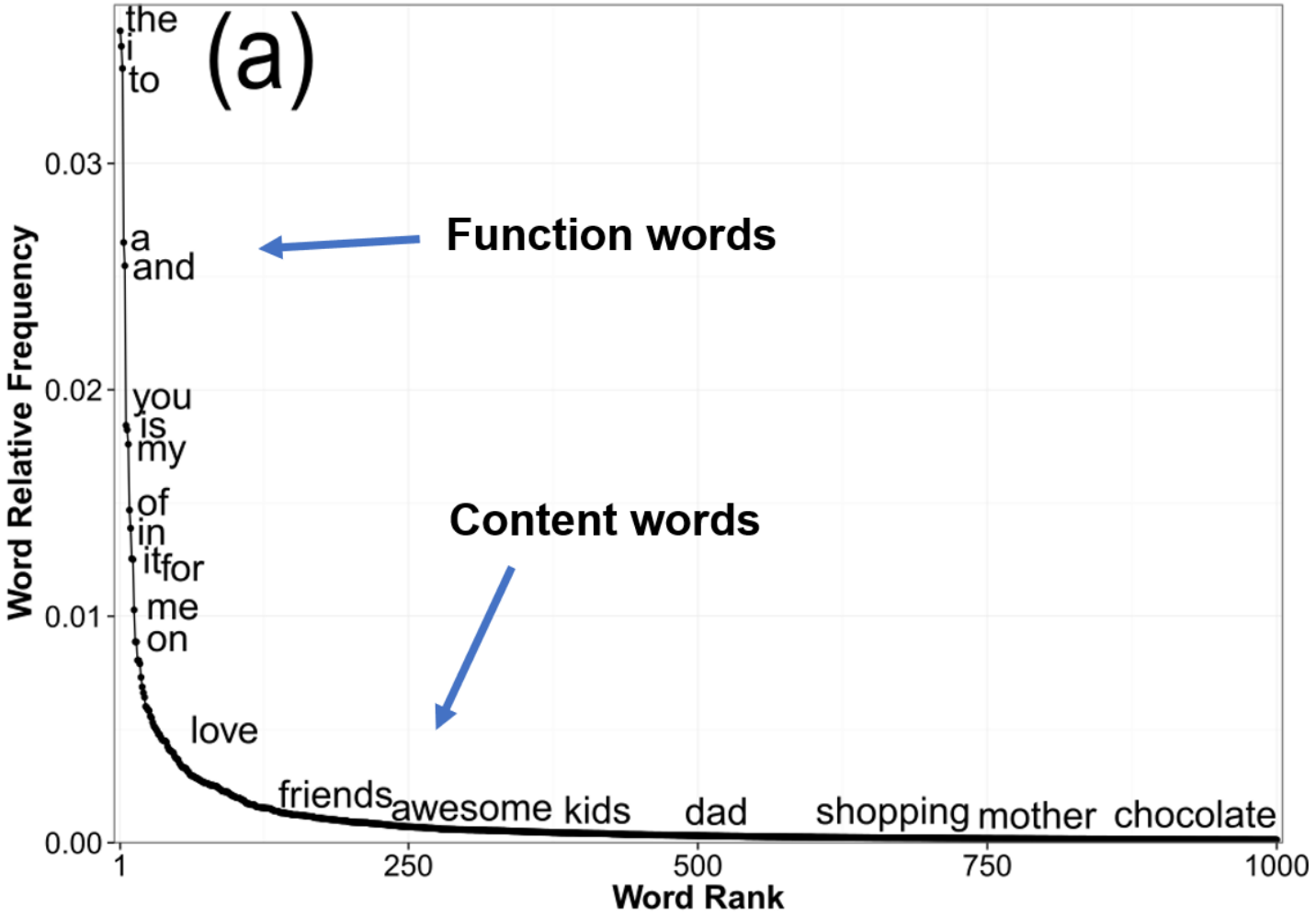
Eichstaedt et al. (2015) Psych Science

A first warning:
language is a weird variable

Language is Weird

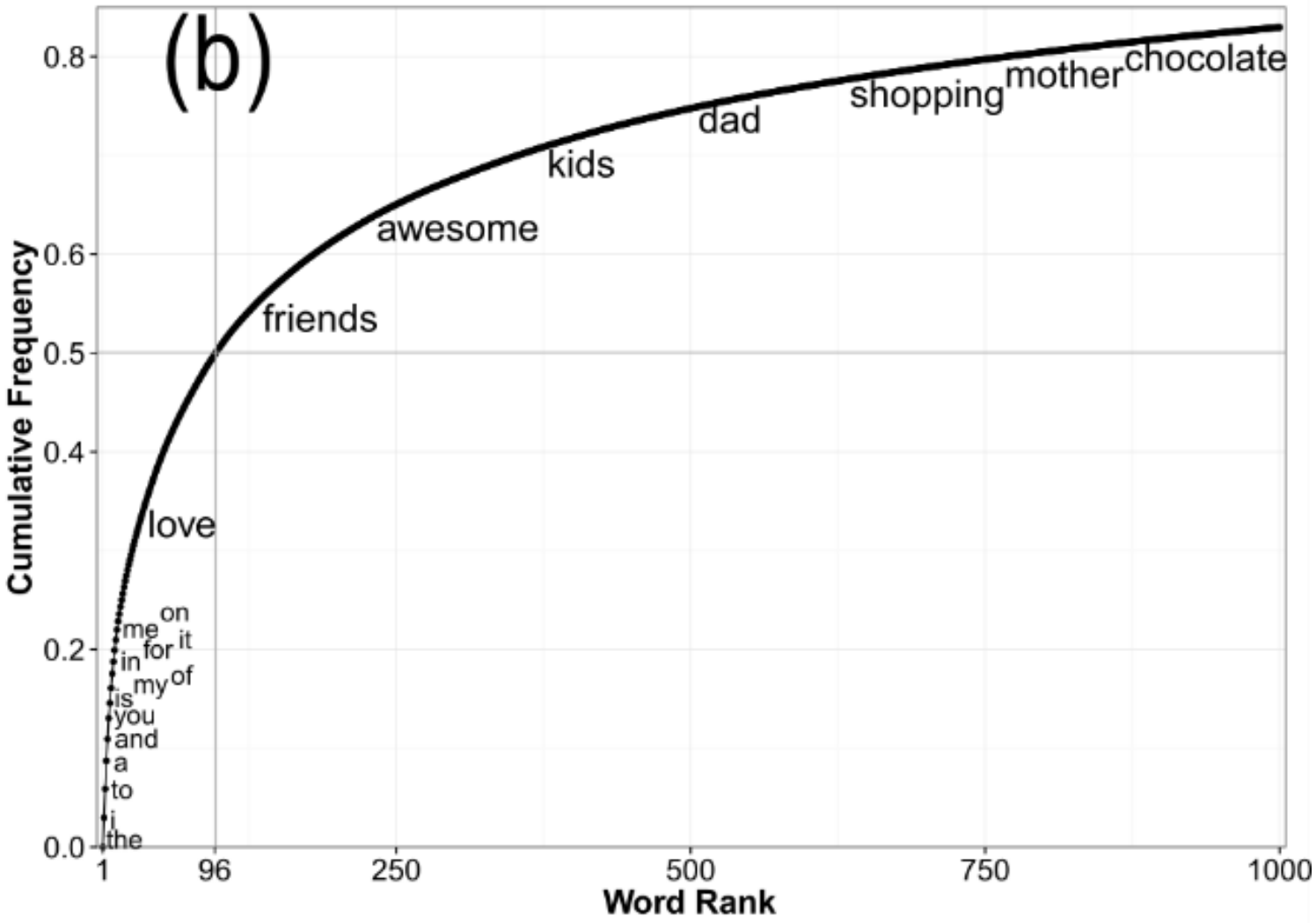
$$P(w_r) = \frac{0.1}{r}$$

w_r = word with rank r



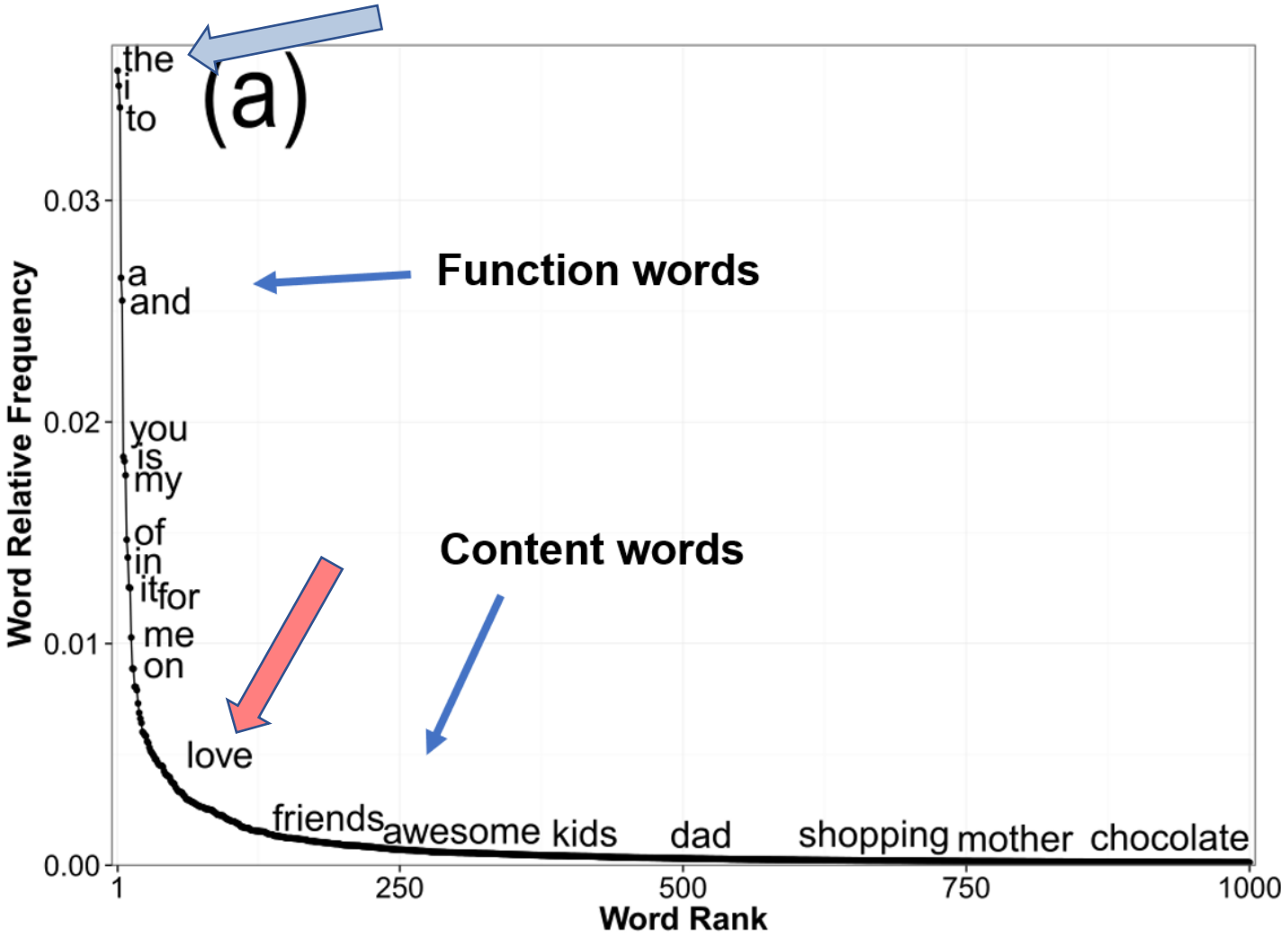
Source: [1]

Language is Weird



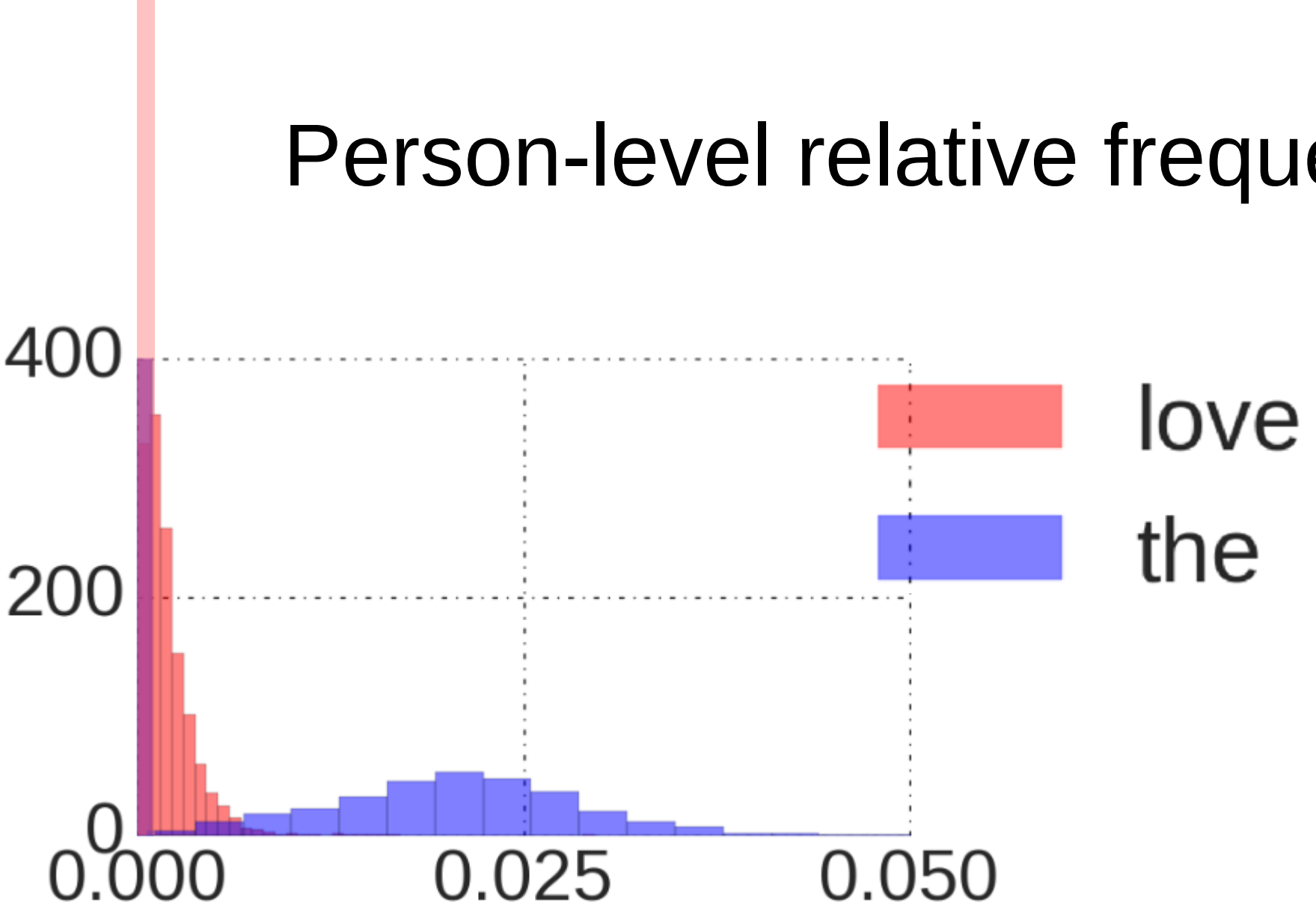
Source: [1]

Language is Weird



Source: [1]

Person-level relative frequencies



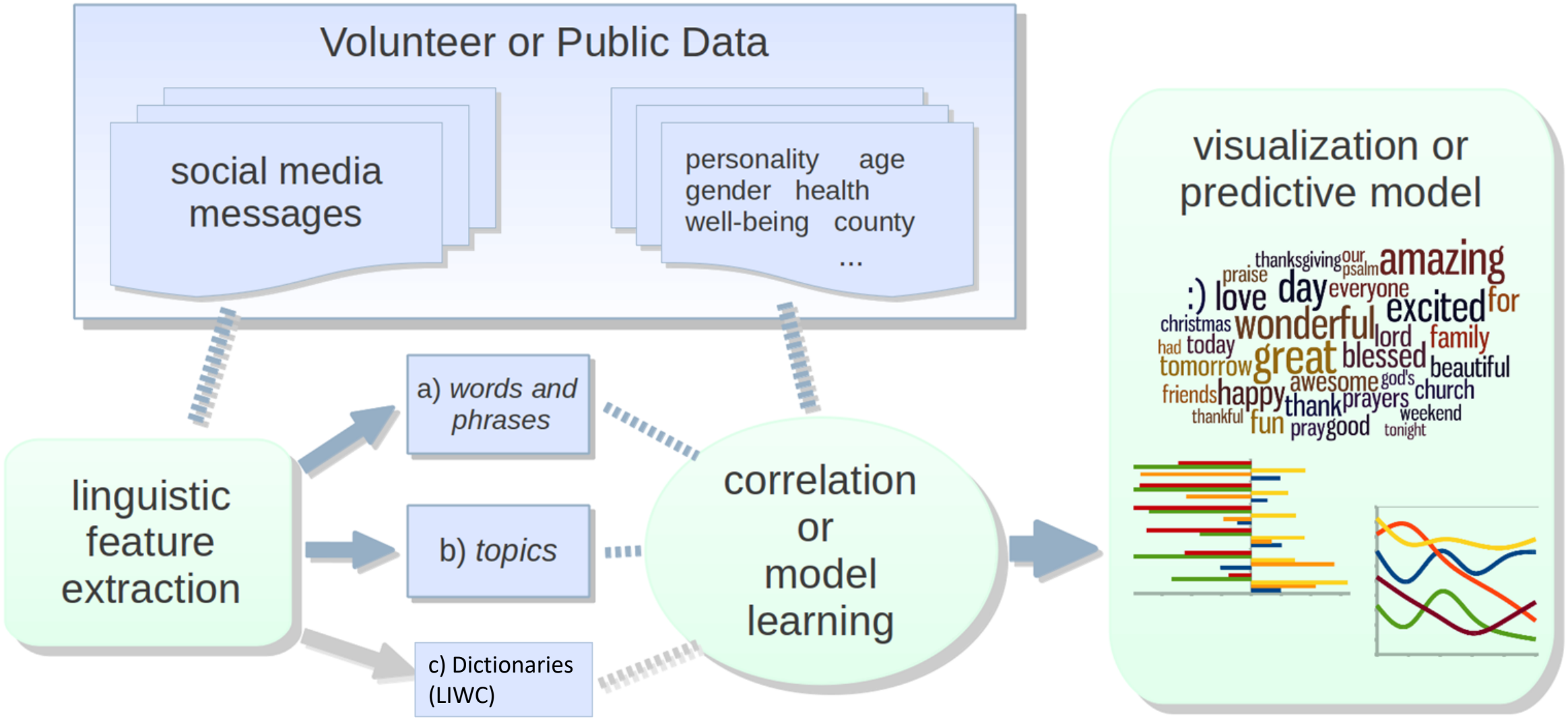
Source: [7, ACL]

Summary: Language is Weird

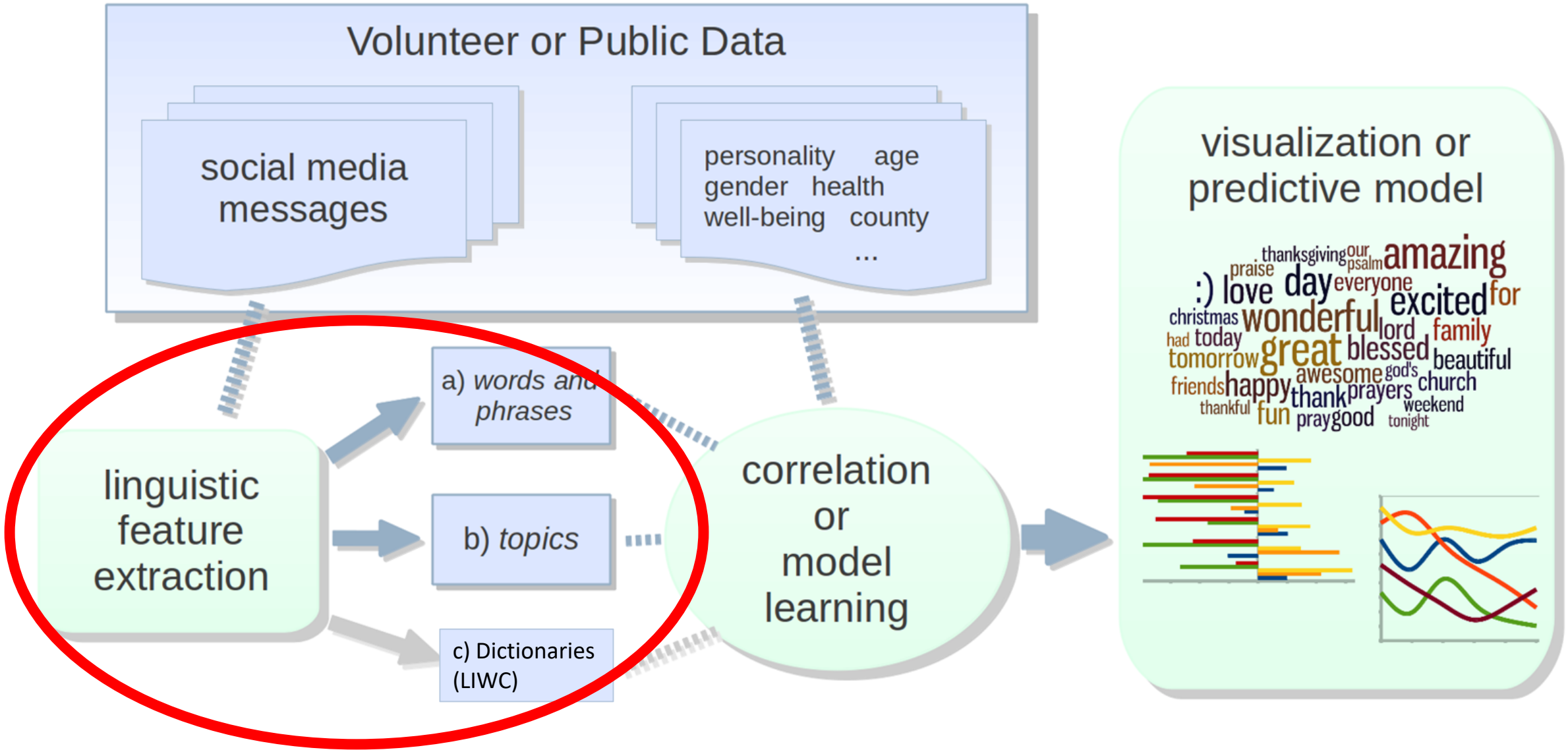
- Total real world vocabularies: 10,000 to 30,000 words
- Content words make for really difficult units of analysis – **most words are never used by most people!!**
- Very rarely are language frequencies remotely normal
- You have to carefully attend to the combination of sample size, minimal word count per observation, and choice of method.

How do we analyze language?

Full Basic Framework



Full Basic Framework



The Steps for Language Analysis

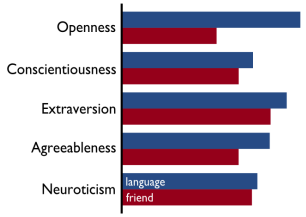
1) Collect/get data

2) Tokenize the text

3) Select language features (words? dictionaries?)

Correlate features with outcome

And/or



Build predictive model

1) Collect data

Pick unit of analysis

- Tweets, people, counties, …

Language and “labels” (outcome)

- Language: Scrape Twitter? Get facebook statuses from people? Writing task?
- Labels: personality survey? Experimental condition?

Need ~ 500 words/person, or ~30,000+ words/county

Controls?

- age & gender
- SES

2) Tokenize the text

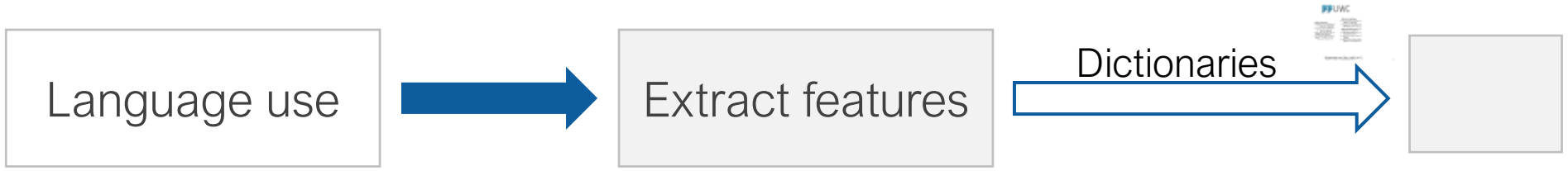
- Break text into “tokens”
  - yesterday, I got sick of my freind :-(#badnews
- Use “emoticon-aware” tokenization:)
- Don’t stem or otherwise normalize

Worked example:

Text1: “The cat cat died unexpectedly:(”

Text ID	token	count
1	The	1
1	cat	2
1	died	1
1	unexpectedly	1
1	:(	1
2	(…)	(…)

3) Extract language Features

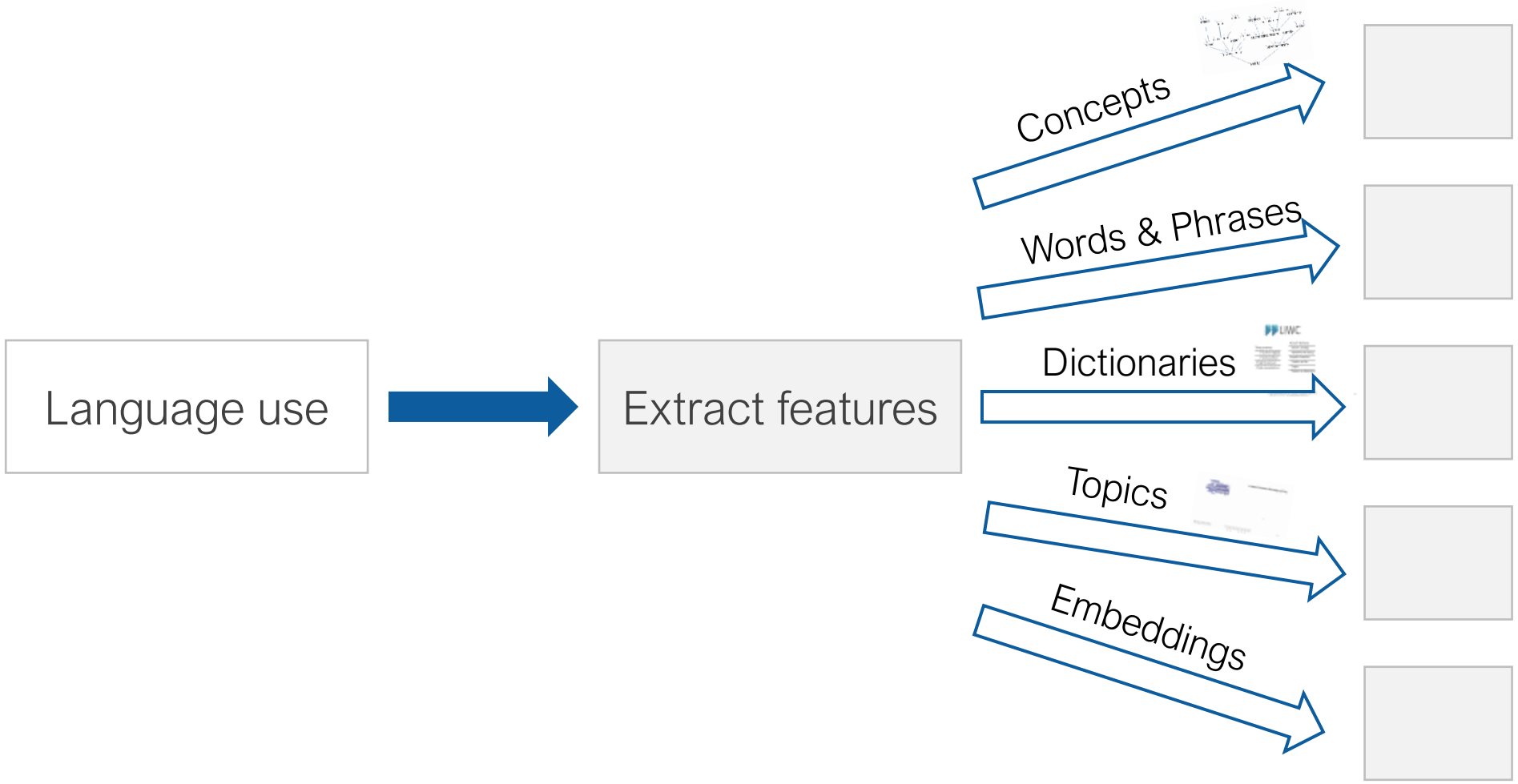


Total pronouns
1st person singular
1st person plural
Total first person
Total second person

Positive Emotions
Positive feelings
Optimism and energy
Negative Emotions
Anxiety or fear
Anger
Sadness or depression

Pennebaker et al. (2001, 2007, 2015)

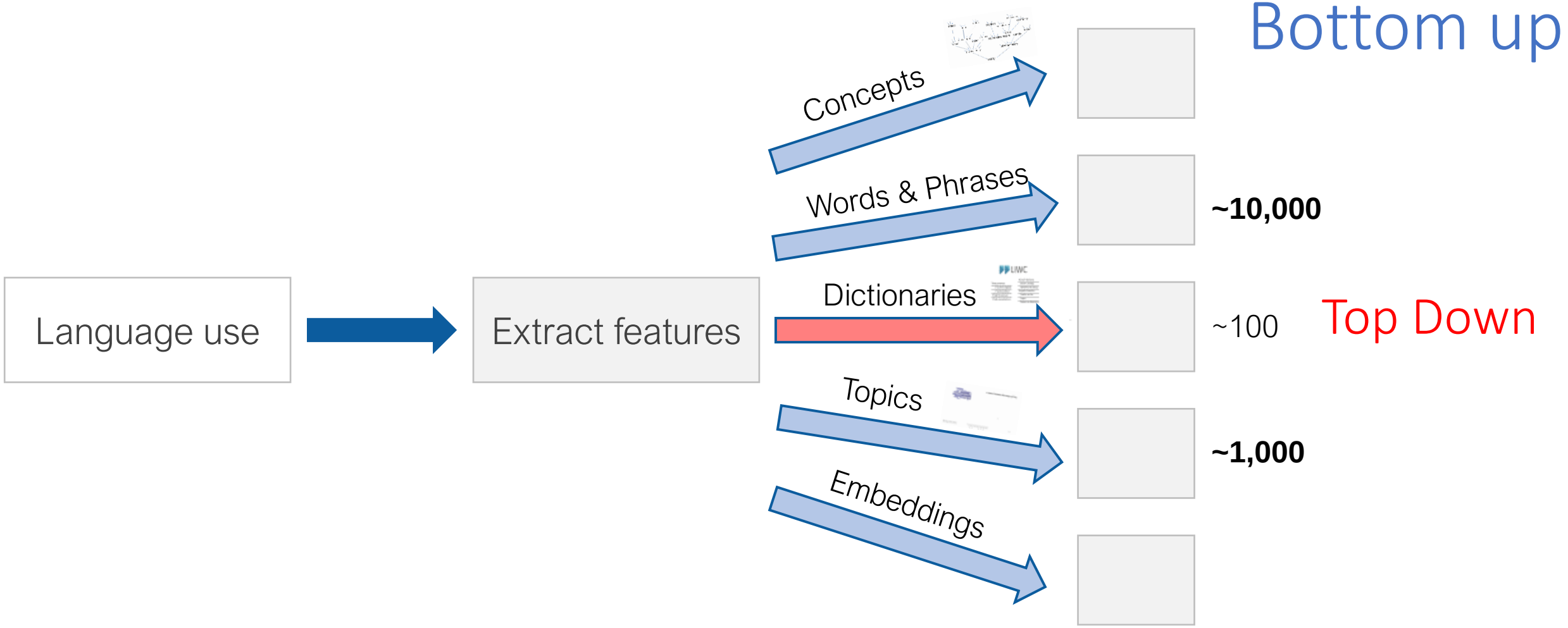
3) Extract language Features



Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan, 2003

3) Extract language Features



Feature extraction –
Language Analysis Methods (psychology view)

Top Down

Dictionaries

Also:
"theory-based"
"predetermined"
"hand-driven"
"closed vocabulary"

Examples:
Linguistic Inquiry and Word Count (LIWC),
DICTION, General Inquirer

Dictionary counts

Bottom up

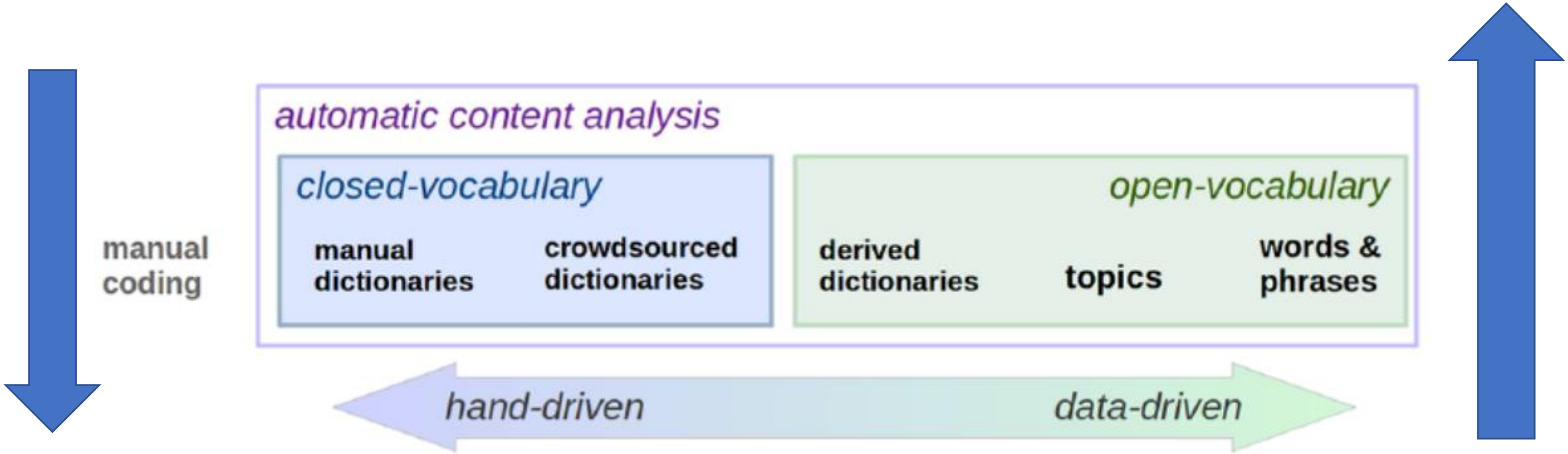
Language variables

Also:
"data-driven"
"open vocabulary"

Examples:
Topic modelling,
Latent Semantic Analysis,
Embeddings,
Differential Language Analysis

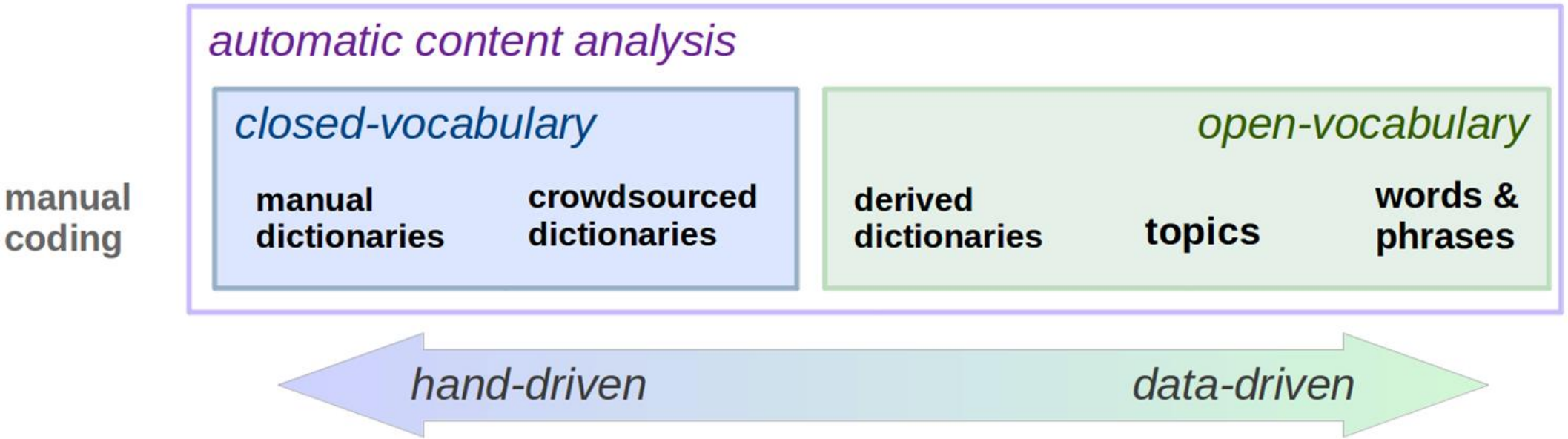
Language data

Feature extraction –
Language Analysis Methods (computer science view)

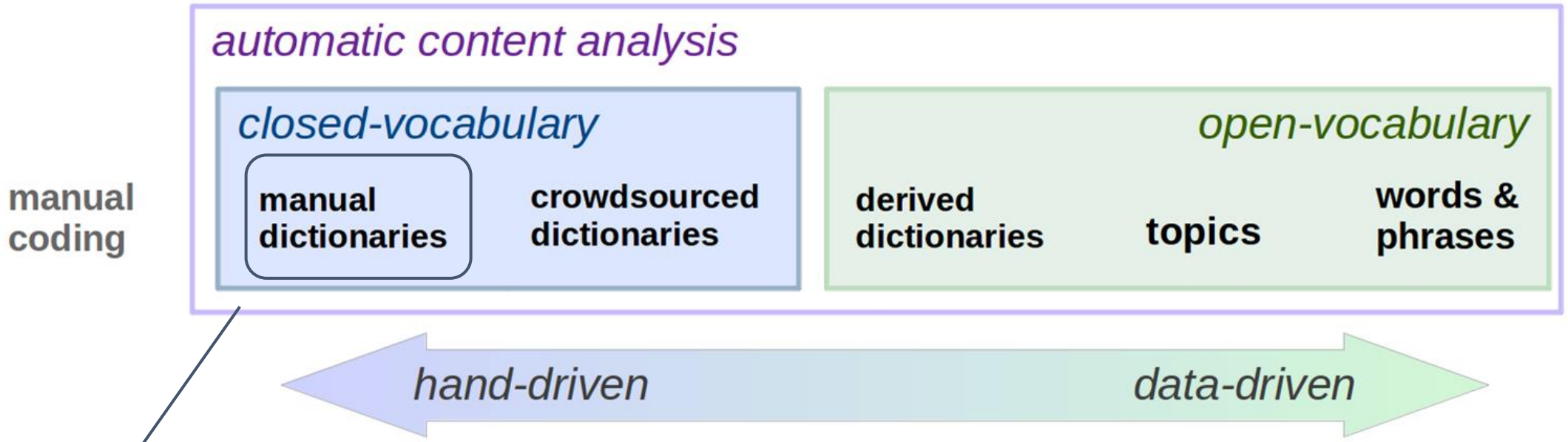


Source: [4]

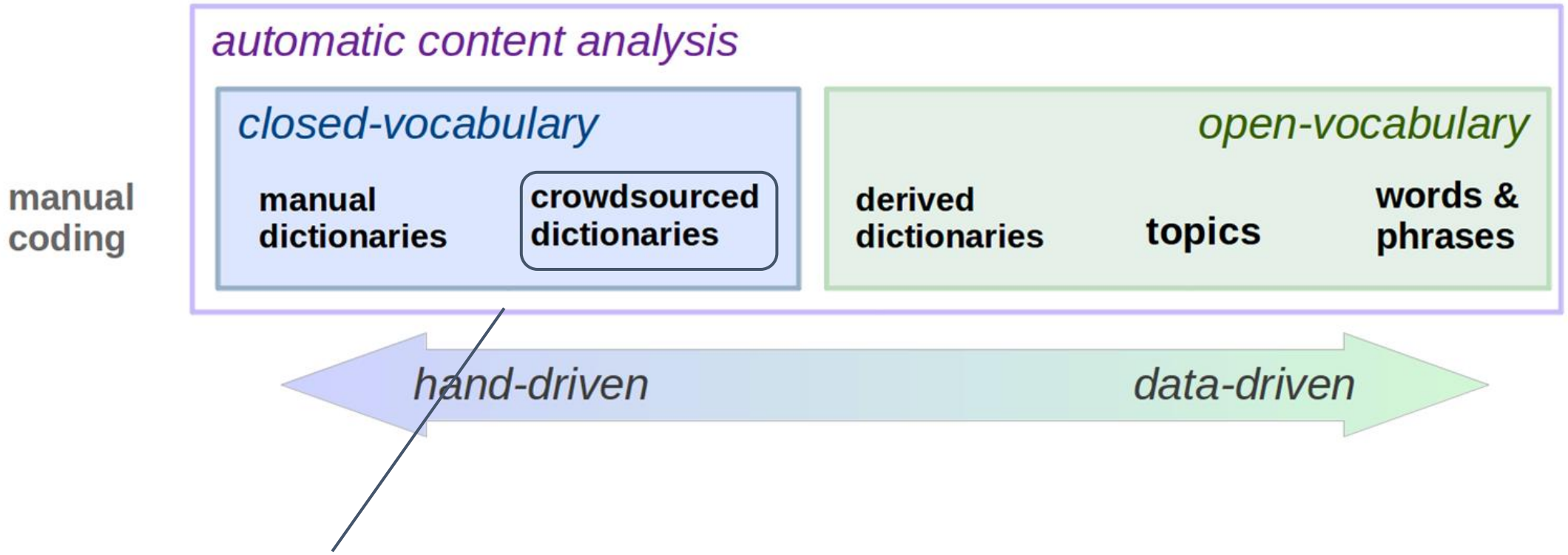
Feature extraction –
Language Analysis Methods (computer science view)



Source: [4]



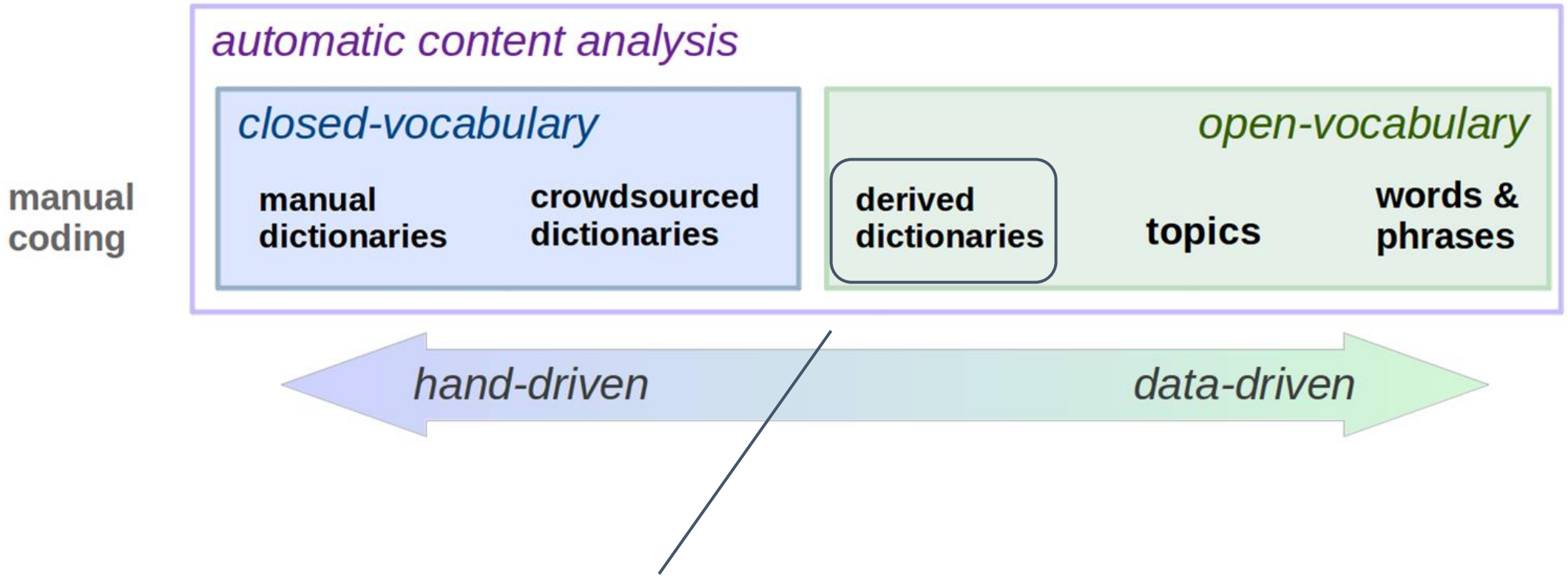
Example:
Linguistic Inquiry and Word Count
(LIWC; Pennebaker et al., 2007)



Examples:

- ANEW
  (Bradley & Lang, 1996)
- Hedonometer
  (Dodds et al., 2011)

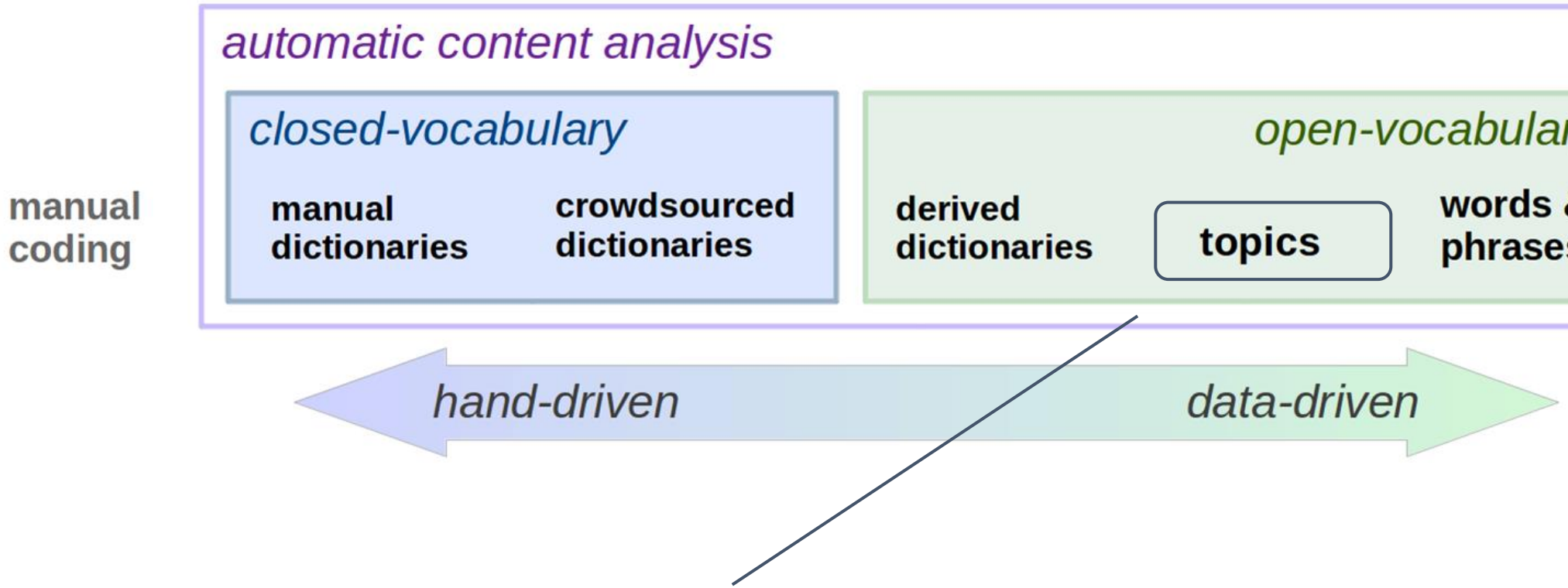
+ often weighted
+ fuller coverage



Examples:

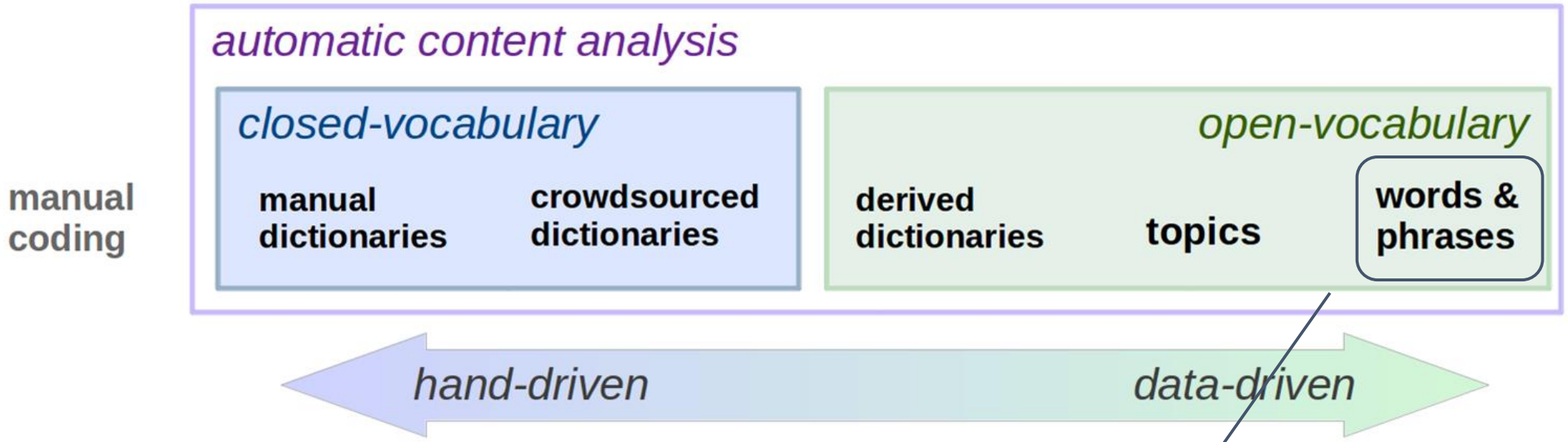
- Sentiment
  (Pang & Lee, 2002)
- Affect & Intensity
  (Preotiuc et al., 2016)

+ real world distributions
(still hypothesis driven)



Topic Modeling:

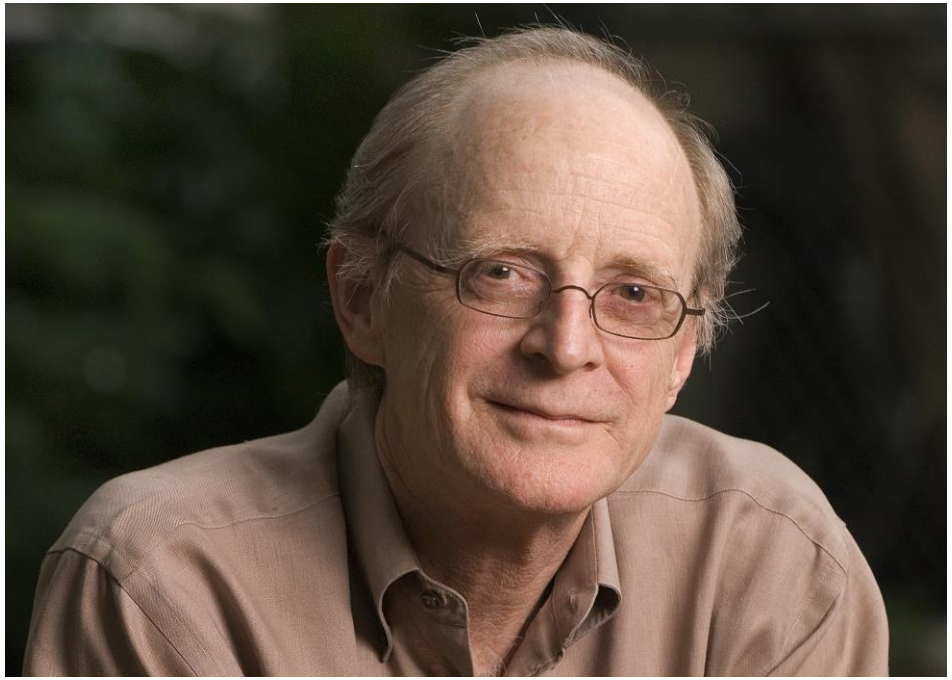
+ completely data-driven
+ “digestable”
(still losing some information)



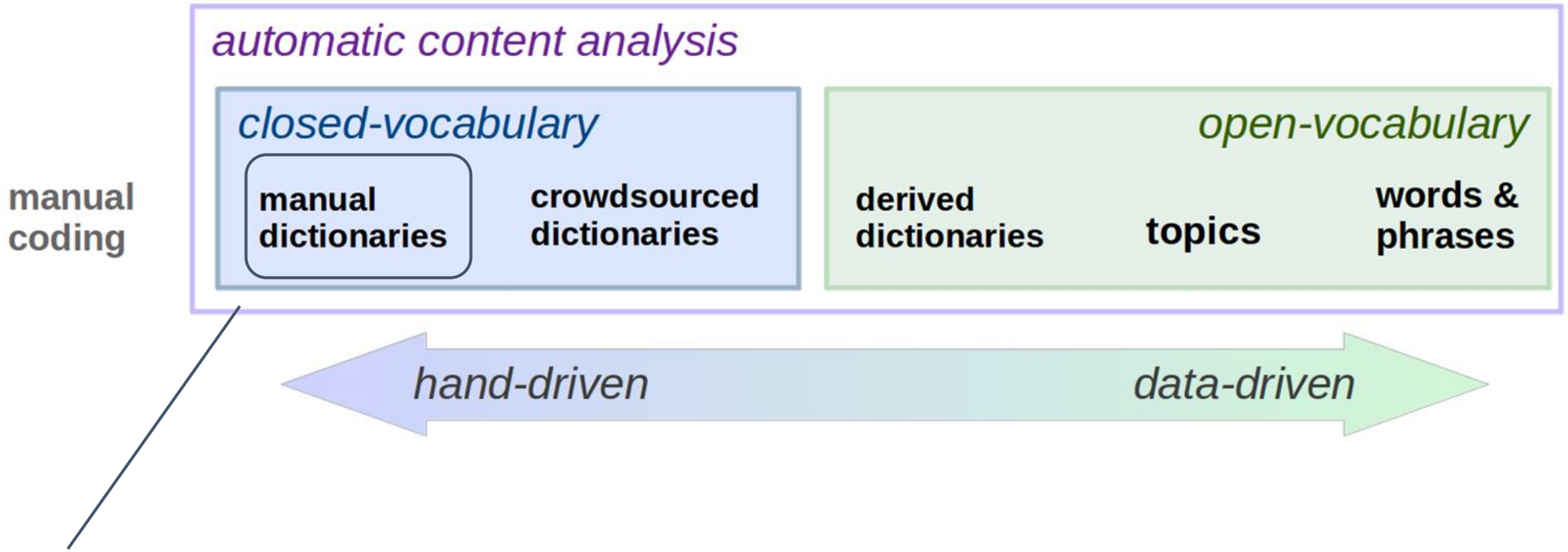
+ wide coverage
+ fine-grained information
phrases: capture some context

(not as “digestable”)

The top down classic: Linguistic Inquiry and Word Count (LIWC)



Jamie Pennebaker



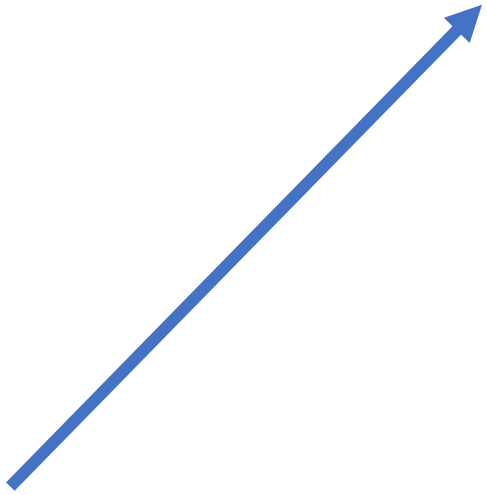
Example:
Linguistic Inquiry and Word Count
(LIWC; Pennebaker et al., 2015)

Structures within LIWC 2015

(LIWC 2007: most widely used in the literature, worse than 2015)

(LIWC 2001: ???)

73 LIWC dictionaries ("categories")



FUNCTION WORDS

Linguistic Dimensions
Total function words
Total pronouns
Personal pronouns
1st pers singular
1st pers plural
2nd person
3rd pers singular
3rd pers plural
Impersonal pronouns
Articles
Prepositions
Auxiliary verbs
Common Adverbs
Conjunctions
Negations

Source: [6]

CONTENT WORDS

Psychological Processes
Affective processes
Positive emotion
Negative emotion
Anxiety
Anger
Sadness
Social processes
Family
Friends
Female references
Male references
Cognitive processes
Insight
Causation
Discrepancy
Tentative
Certainty
Differentiation

Drives
Affiliation
Achievement
Power
Reward
Risk
Time orientations
Past focus
Present focus
Future focus
Relativity
Motion
Space
Time

Personal concerns
Work
Leisure
Home
Money
Religion
Death
Informal language
Swear words
Netspeak
Assent
Nonfluencies
Fillers

Source: [1]

LIWC 2015 Dictionary example words

LIWC category	Words
Cognitive processes	
Insight	think, know, consider
Tentativeness	maybe, perhaps, guess
Differentiation	hasn't, but, else
Cause	because, effect, hence
Discrepancies	should, would, could
Affective processes	
Anger	hate, kill, pissed
Negative emotion	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Sadness	grief, cry, sad

Summary Table Linking LIWC Word Categories to Published R

Category	Examples	Words in Category
Linguistic processes		
Word count		
Words/sentence		
Dictionary words	(Percentage of all words captured by the program)	
Words >6 letters	(Percentage of all words longer than 6 letters)	
Total function words		464
Total pronouns	I, them, itself	116
Personal pronouns	I, them, her	70
First-person singular	I, me, mine	12
First-person plural	We, us, our	12
Second person	You, your, thou	20
Third-person singular	She, her, him	17
Third-person plural	They, their, they'd	10

: 10,000 to 30,000 words

Source: [3: table]

Summary Table Linking LIWC Word Categories to Published Research Studies

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Linguistic processes				
Word count			Talkativeness, verbal fluency	2, 9, 18, 19, 20, 24, 32, 35, 36, 39, 40, 48, 53, 54, 57, 60, 66, 70, 72, 73, 74, 86, 89, 103, 115
Words/sentence			Verbal fluency, cognitive complexity	3, 7, 39, 43
Dictionary words	(Percentage of all words captured by the program)		Informal, nontechnical language	19, 42, 43, 65, 66, 85, 89
Words >6 letters	(Percentage of all words longer than 6 letters)		Education, social class	3, 19, 20, 27, 35, 36, 42, 43, 73, 74, 79, 89, 90, 93, 103, 115
Total function words		464		
Total pronouns	I, them, itself	116	Informal, personal	1, 19, 36, 43, 55, 89, 90, 119
Personal pronouns	I, them, her	70	Personal, social	58, 79
First-person singular	I, me, mine	12	Honest, depressed, low status, personal, emotional, informal	1, 3, 4, 5, 11, 13, 18, 27, 35, 36, 46, 55, 56, 64, 65, 66, 68, 69, 72, 73, 74, 78, 80, 81, 87, 89, 90, 92, 93, 94, 100, 101, 105, 108, 109, 112, 113, 115
First-person plural	We, us, our	12	Detached, high status, socially connected to group (sometimes)	1, 4, 13, 18, 35, 46, 55, 64, 65, 74, 78, 81, 87, 90, 93, 94, 97, 100, 103, 104, 105, 106, 113
Second person	You, your, thou	20	Social, elevated status	1, 18, 27, 41, 55, 90, 100, 105, 106
Third-person singular	She, her, him	17	Social interests, social support	1, 3, 14, 36, 39, 55, 64, 66, 80, 87, 88, 90, 95
Third-person plural	They, their, they'd	10	Social interests, out-group awareness (sometimes)	1, 3, 14, 39, 55, 64, 80, 87, 88, 95

Source: [3: table]

Category	Examples	Words in Category	Psychological Correlates	Published Articles
Positive emotion	Love, nice, sweet	406		2, 3, 4, 5, 6, 8, 10, 12, 15, 17, 21, 22, 23, 25, 28, 30, 31, 33, 36, 37, 38, 41, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 57, 59, 60, 61, 62, 64, 66, 67, 68, 69, 70, 71, 73, 74, 75, 76, 77, 81, 82, 85, 89, 91, 93, 94, 96, 99, 107, 108, 109, 110, 113, 115, 117, 118
Negative emotion	Hurt, ugly, nasty	499		2, 3, 4, 6, 10, 12, 13, 16, 17, 20, 21, 22, 25, 28, 29, 30, 31, 33, 35, 37, 40, 44, 45, 46, 47, 48, 50, 51, 52, 53, 55, 57, 59, 61, 62, 63, 64, 66, 67, 70, 71, 72, 73, 74, 76, 79, 80, 81, 82, 84, 85, 89, 91, 92, 93, 94, 96, 99, 102, 107, 113, 115, 117, 119, 121
Anxiety	Worried, nervous	91		6, 28, 50, 66, 68, 77, 84, 85, 92
Anger	Hate, kill, annoyed	184		6, 28, 33, 50, 58, 66, 72, 74, 92
Sadness	Crying, grief, sad	101		6, 28, 33, 38, 50, 63, 66, 77, 84, 90
Cognitive processes	Cause, know, ought	730		2, 3, 5, 8, 13, 18, 21, 23, 31, 32, 34, 46, 47, 49, 55, 58, 61, 68, 69, 71, 75, 83, 84, 85, 86, 89, 92, 93, 102, 104, 119, 120
Insight	Think, know, consider	195		1, 4, 18, 19, 25, 35, 37, 45, 53, 59, 68, 73, 76, 89, 90, 91, 92, 93, 97, 99, 111, 113, 115, 118, 119, 121
Causation	Because, effect, hence	108		10, 13, 16, 20, 35, 37, 39, 45, 53, 72, 76, 89, 90, 91, 93, 97, 99, 115, 121, 122
Discrepancy	Should, would, could	76		10, 16, 18, 19, 49, 63, 74, 89, 115
Tentative	Maybe, perhaps, guess	155		18, 19, 24, 37, 38, 49, 73, 87, 89, 98, 115
Certainty	Always, never	83	Social/verbal skills, emotional stability	38
Inhibition	Block, constrain, stop	111		1, 16, 18, 19, 49, 90, 111
Inclusive	And, with, include	18		41, 60, 73, 74, 89, 115
Exclusive	But, without, exclude	17	Cognitive complexity, honesty	24, 49, 73, 80, 89, 92, 93, 115

Source: [3: table]

LIWC is great for context within the larger literature.

LIWC 2015 results for personality

The following slides:

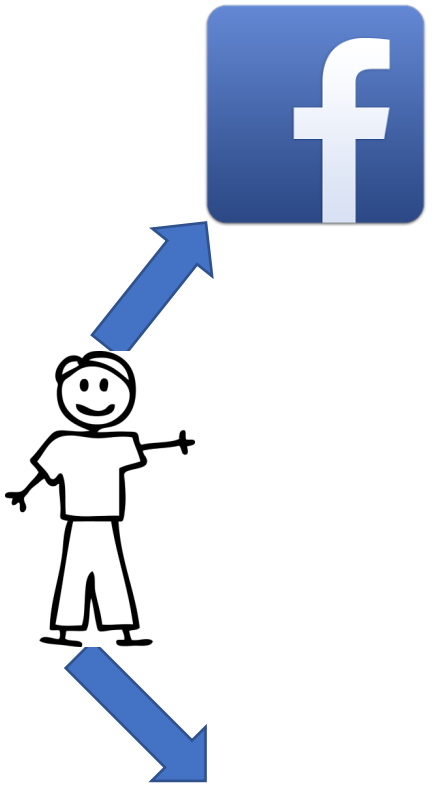
Source: [1]

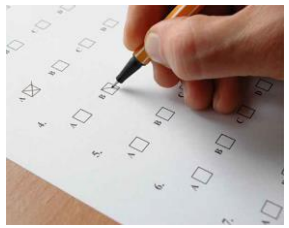
N = 65,000 Facebook users,

status histories +
IPIP Big Five personality survey data

Controlling for age and gender

LIWC personality correlation

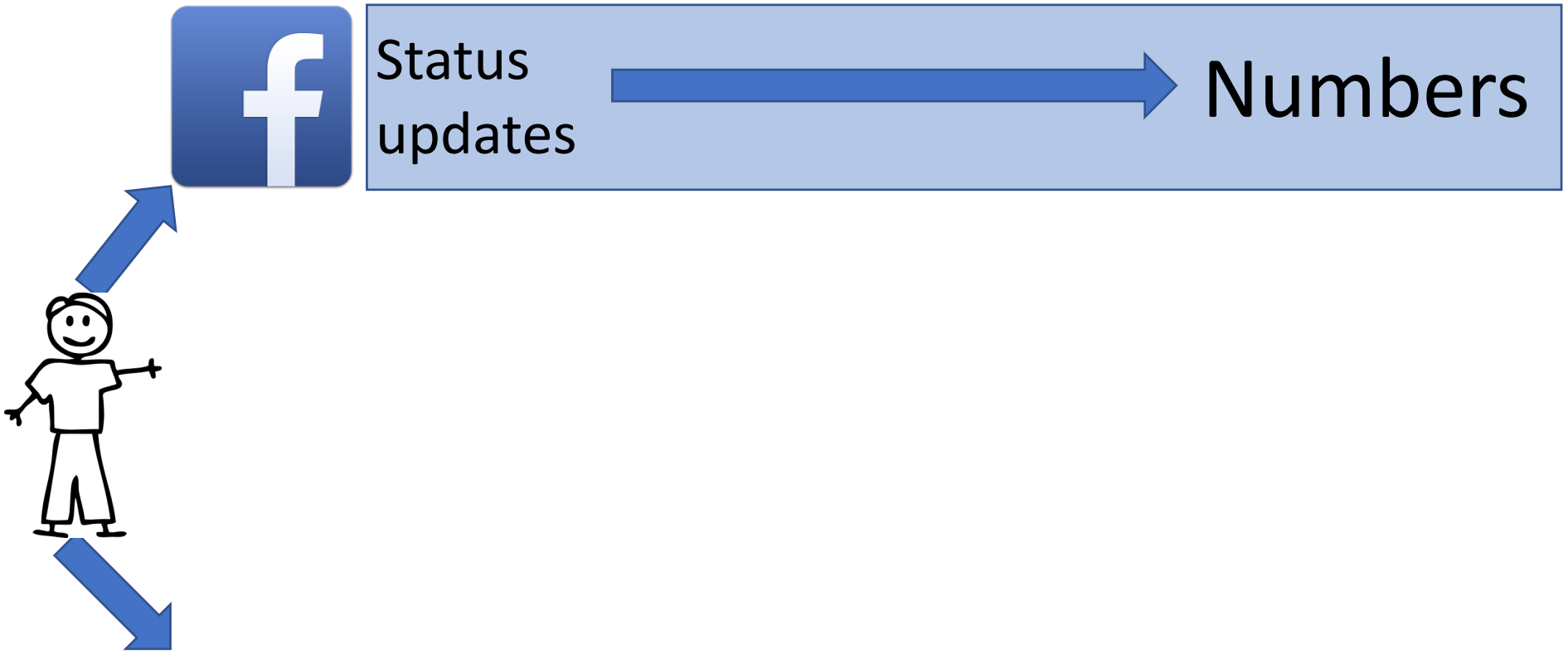




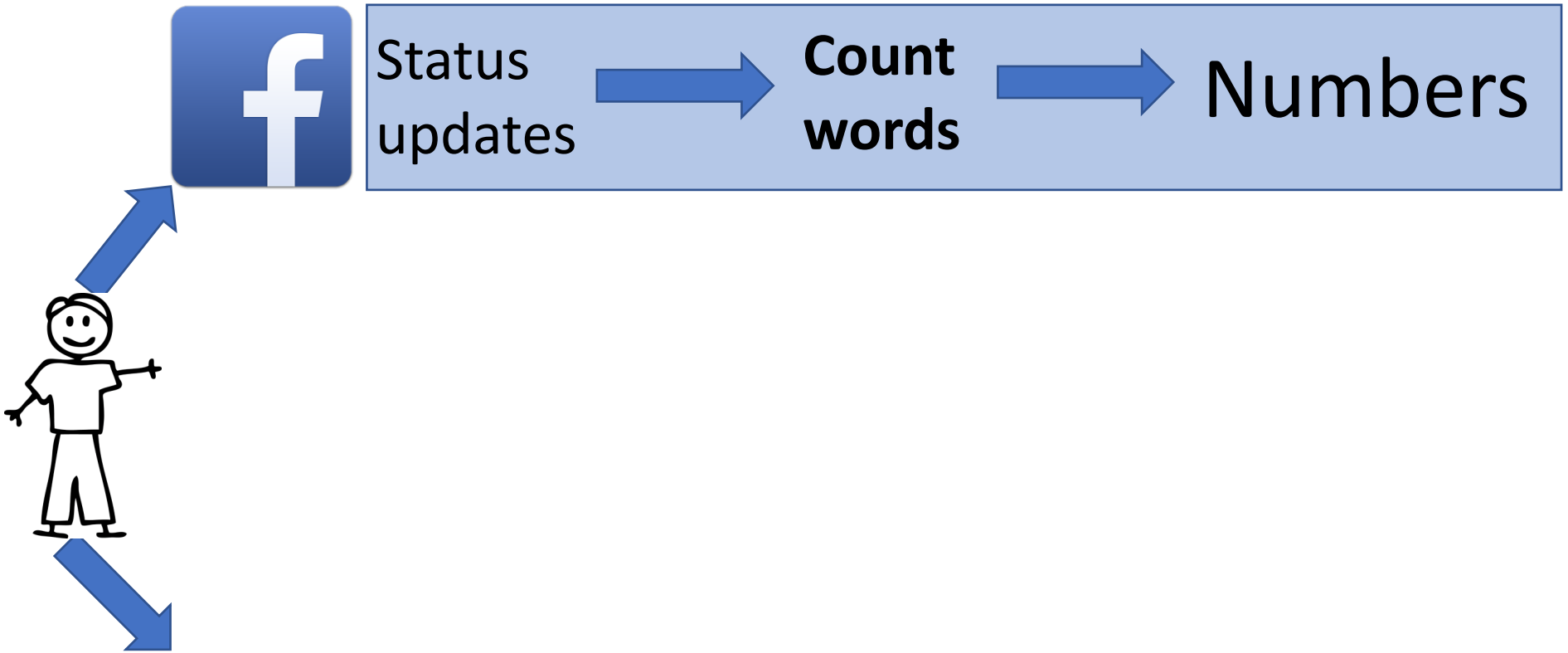
Normal Variables (gender, Big5…)

LIWC personality correlation

LIWC personality correlation

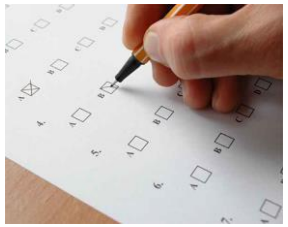


LIWC personality correlation



LIWC personality correlation

Status updates → **Count words** → Relative Frequencies of LIWC dictionaries



Normal Variables (gender, Big5…)

Worked Example

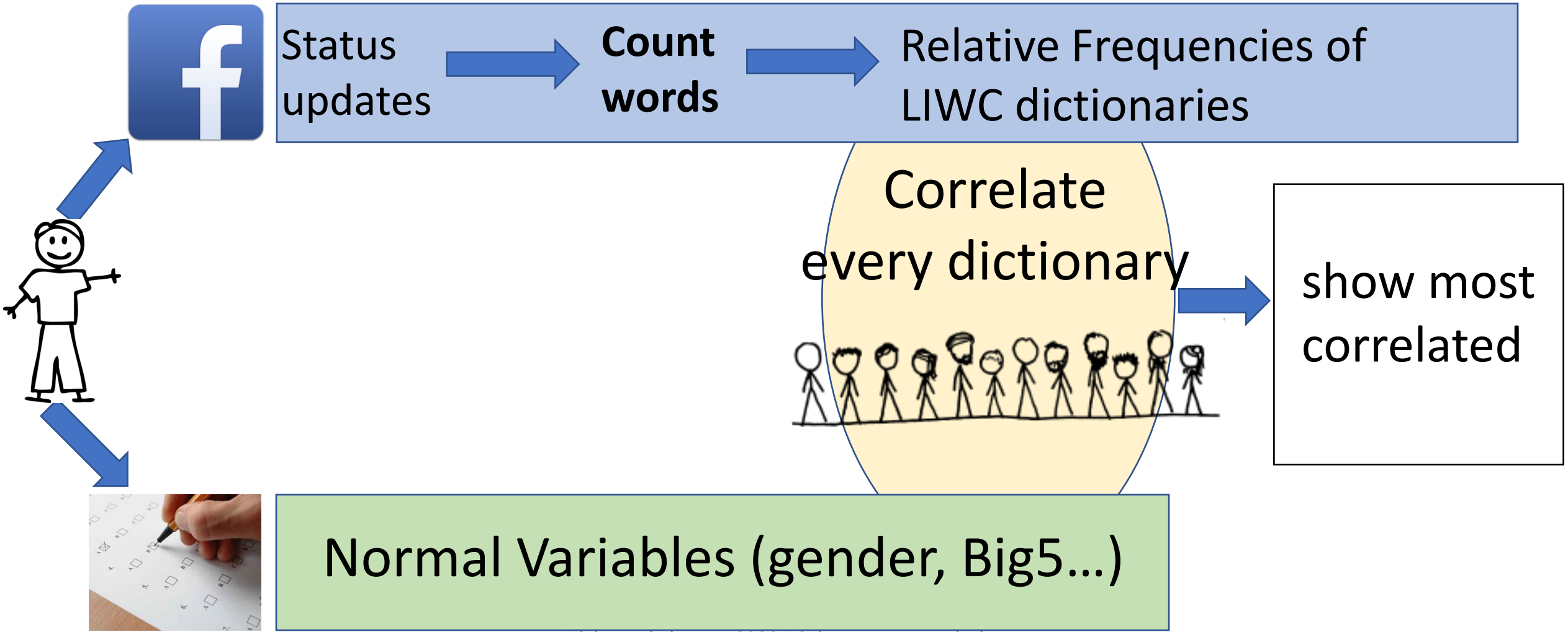
"The cat cat died unexpectedly:("

Extraction of LIWC language features.

Text ID	token	count
1	The	1
1	cat	2
1	died	1
1	unexpectedly	1
1	:(	1
2	(…)	(…)

Text ID	LIWC DICTIONARY	relativeFrequency
1	DEATH	0.166
1	NETSPEAK	0.166
1	ARTICLES	0.166

LIWC personality correlation



Gender

	Linguistic Inquiry and Word Count (LIWC 2015)			
	LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β
Female	Emotional tone (m)	.27	Social processes	.12
	Personal pronoun	.17	Female reference	.30
	1st pers singular	.16	Family	.28
	3rd pers singular	.11	Affective process	.25
	2nd person	.07	Positive emotion	.29
	Total pronouns	.11	Home	.21
	Common adverbs	.09	Netspeak	.18
	Common verbs	.07	Affiliation	.17
	Conjunctions	.07	Future focus	.10
	Common adjectives	.06	Nonfluencies	.10

Source: [1]

Gender

	Linguistic Inquiry and Word Count (LIWC 2015)			
	LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β
Female	Emotional tone (m)	.27	Social processes	.12
	Personal pronoun	.17	Female reference	.30
	1st pers singular	.16	Family	.28
	3rd pers singular	.11	Affective process	.25
	2nd person	.07	Positive emotion	.29
	Total pronouns	.11	Home	.21
	Common adverbs	.09	Netspeak	.18
	Common verbs	.07	Affiliation	.17
	Conjunctions	.07	Future focus	.10
	Common adjectives	.06	Nonfluencies	.10
Male	Articles	.24	Death	.22
	Analytical thinking (m)	.19	Anger	.21
	Comparisons	.12	Drives	
	Prepositions	.12	Power	.20
	Impersonal pronouns	.08	Achievement	.13
	Quantifiers	.06	Risk	.09
	Interrogatives	.06	Swear words	.19
	3rd pers plural	.05	Sexual	.19
	Numbers	.04	Space	.16
			Money	.11
			Tentative	.09

Source: [1]

Agreeableness

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Emotional tone (m)	.21	Positive emotion	.14
Clout (m)	.07	Drives	.09
Personal pronouns		Affiliation	.09
1st pers plural	.06	Reward	.06
Common adjectives	.05	Achievement	.05
Prepositions	.04	Relativity	.07
Quantifiers	.04	Time	.08
Authentic (m)	.03	Motion	.05
		Future focus	.07
		Religion	.07

Source: [1]

Extraversion

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Emotional tone (m)	.18	Positive emotion	.16
Clout (m)	.06	Drives	
Personal pronoun	.03	Affiliation	.12
2nd person	.04	Reward	.09
1st pers plural	.03	Netspeak	.11
1st pers singular	.02	Social processes	.05
		Friends	.09
		Family	.05
		Leisure	.07
		Future focus	.05
		Biological processes	.04

Source: [1]

Extraversion

Intraversion

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Emotional tone (m)	.18	Positive emotion	.16
Clout (m)	.06	Drives	
Personal pronoun	.03	Affiliation	.12
2nd person	.04	Reward	.09
1st pers plural	.03	Netspeak	.11
1st pers singular	.02	Social processes	.05
		Friends	.09
		Family	.05
		Leisure	.07
		Future focus	.05
		Biological processes	.04
Negations	.06	Personal concern	
Auxiliary verbs	.06	Death	.10
Personal pronouns		Work	.05
3rd pers plural	.06	Cognitive process	.09
Impersonal pronouns	.05	Tentative	.09
Common verbs	.05	Insight	.09
Common adverbs	.05	Differentiation	.08
Articles	.04	Causation	.07
Comparisons	.04	Risk	.08
Interrogatives	.04	Negative emotion	.07
		Anxiety	.07

Source: [1]

Conscientiousness

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Emotional tone (m)	.17	Drives	.12
Prepositions	.07	Achievement	.12
Clout (m)	.06	Reward	.09
Quantifiers	.06	Affiliation	.07
Personal pronouns		Relativity	.10
1st pers plural	.05	Time	.11
Analytical thinking(m)	.05	Motion	.06
Common adjectives	.05	Positive emotion	.11
Authentic (m)	.04	Work	.11
Articles	.03	Future focus	.07

Source: [1]

Neuroticism

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Negations	.07	Negative emotion	.15
Common adverbs	.05	Anger	.11
Common verbs	.05	Sadness	.09
Personal pronouns	.03	Anxiety	.08
1st pers singular	.05	Death	.08
3rd pers singular	.02	Cognitive process	.06
Auxiliary verbs	.04	Discrepancy	.07
Conjunctions	.03	Tentative	.06
		Biological processes	
		Body	.06
		Sexual	.06

Source: [1]

Openness to Experience

Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Articles	.15	Cognitive process	.09
Total function words	.08	Insight	.12
Auxiliary verbs	.07	Causation	.07
Comparisons	.06	Tentative	.07
Impersonal pronouns	.06	Death	.12
Conjunctions	.06	Perceptual process	.12
Prepositions	.05	Hear	.08
1st pers singular	.04	See	.07
Interrogatives	.04	Anxiety	.08
Quantifiers	.04	Space	.05

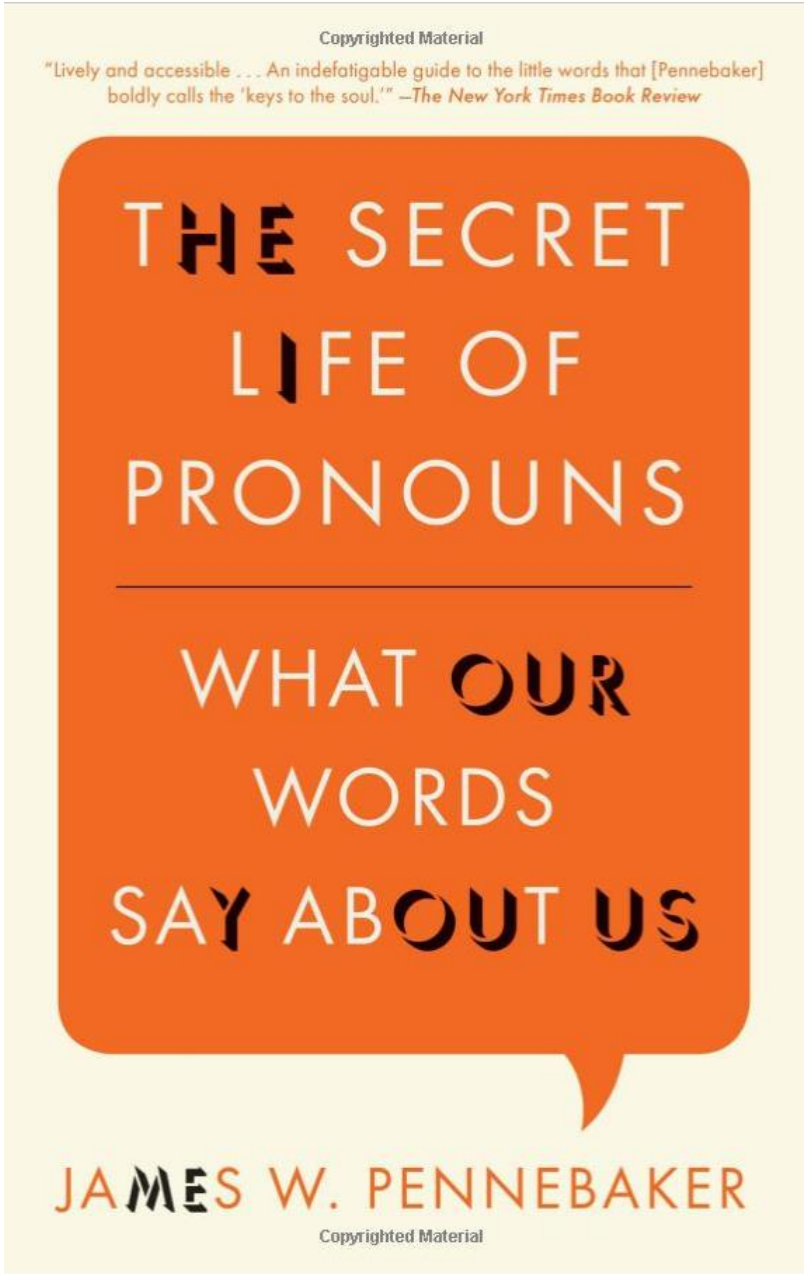
Mehl assessment

From an evolutionary perspective, it is unlikely that language has evolved as a vehicle to express emotion. Instead, humans use intonation, facial expression, or other nonverbal cues to convey feelings. Emotional tone is also expressed through metaphor and other means not related to emotion words. Taken together, embarking on emotion words to study human emotions has not emerged as a particularly promising strategy (Pennebaker et al., 2003).

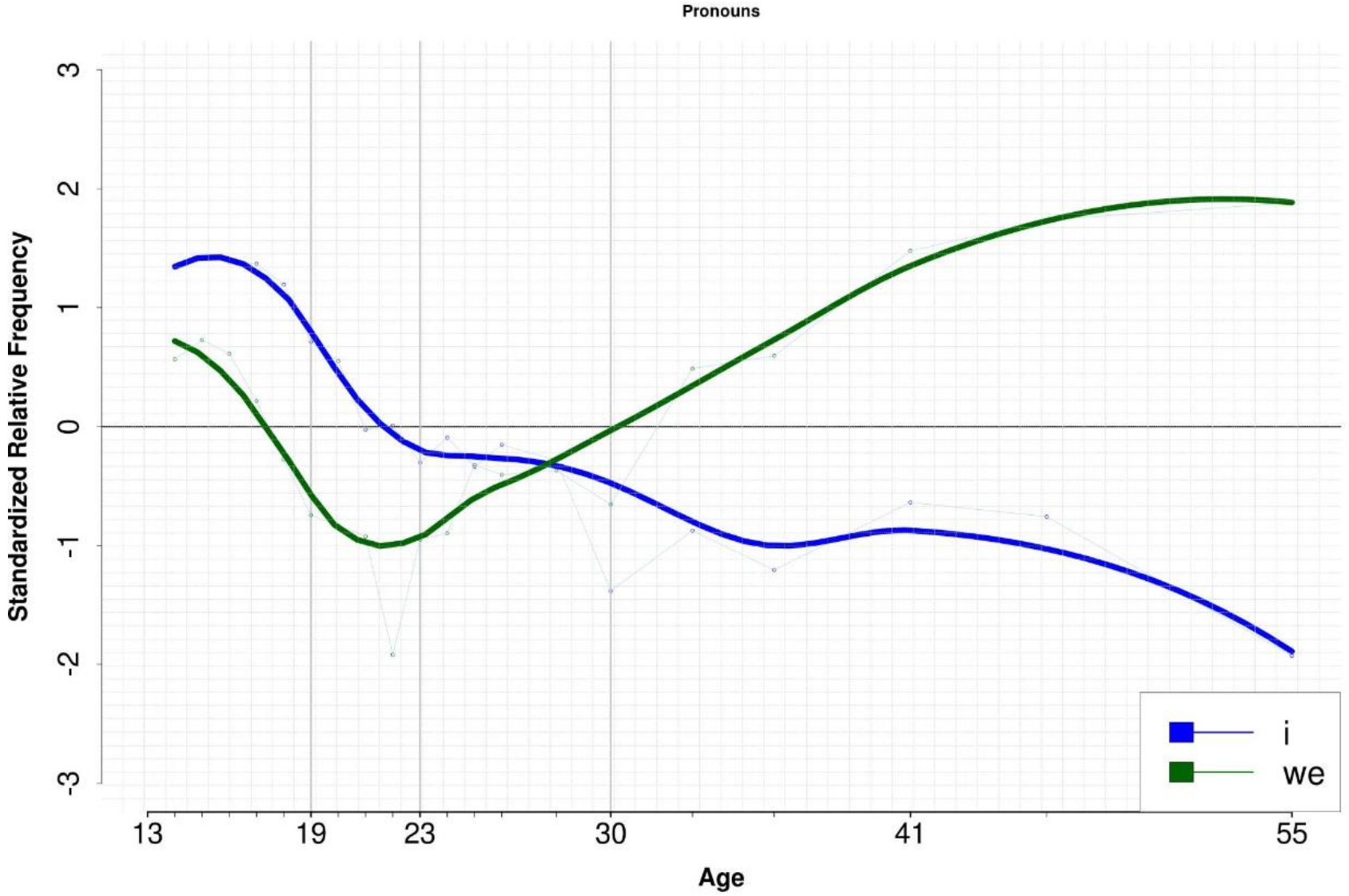
Source: [9, Mehl handbook]

Cool results have mostly been about pronouns.

The LIWC literature is mostly about “function” words.



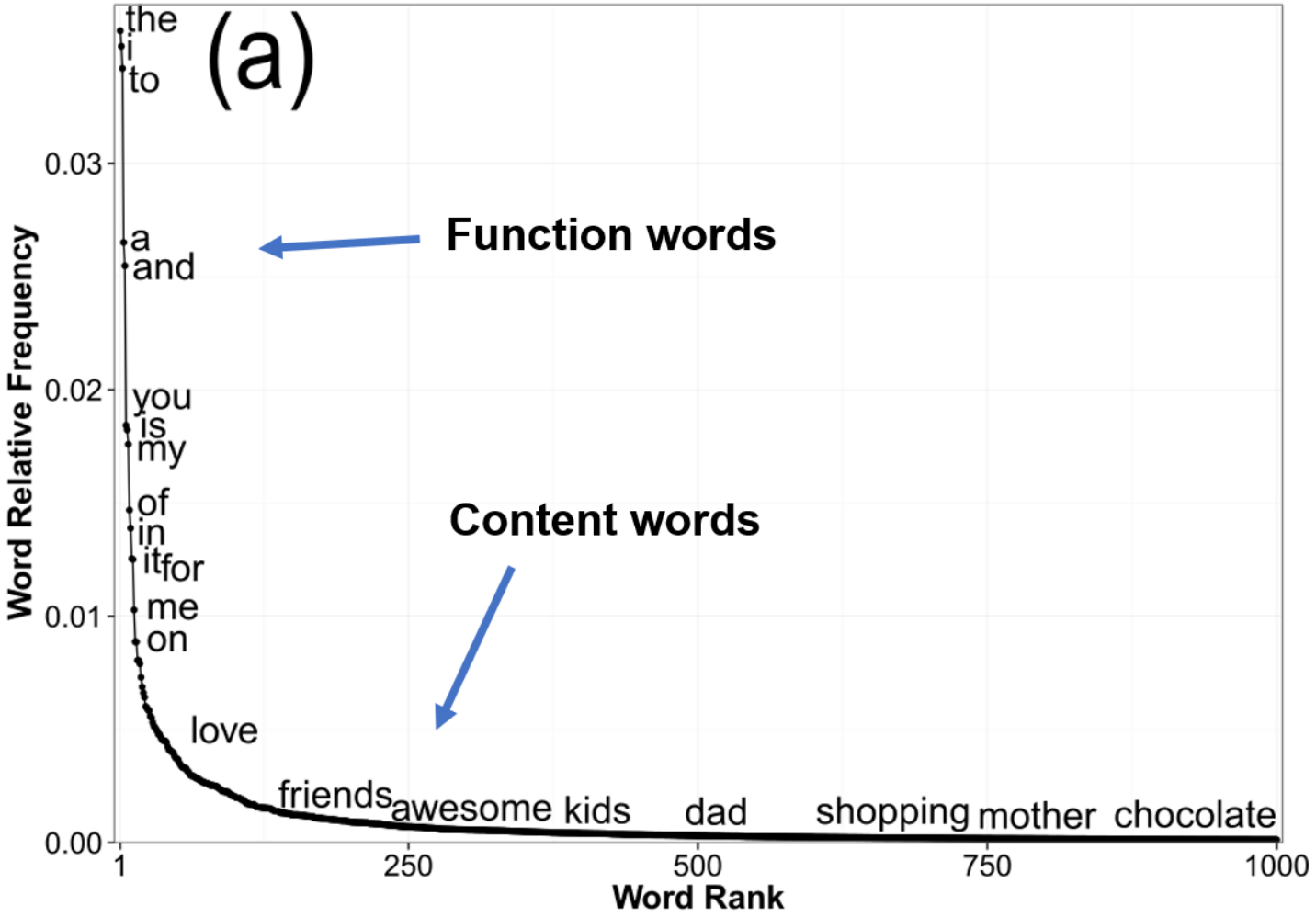
An example: I/we over age



Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., … & Seligman, M. E. P. (2014). From “sooo excited!!!” to “so proud”: Using language to study development. Developmental Psychology, 50, 178-188

A second warning:
Dictionaries aren’t always what you think

Language is Weird



Source: [1]

Dictionaries contain many words, right?

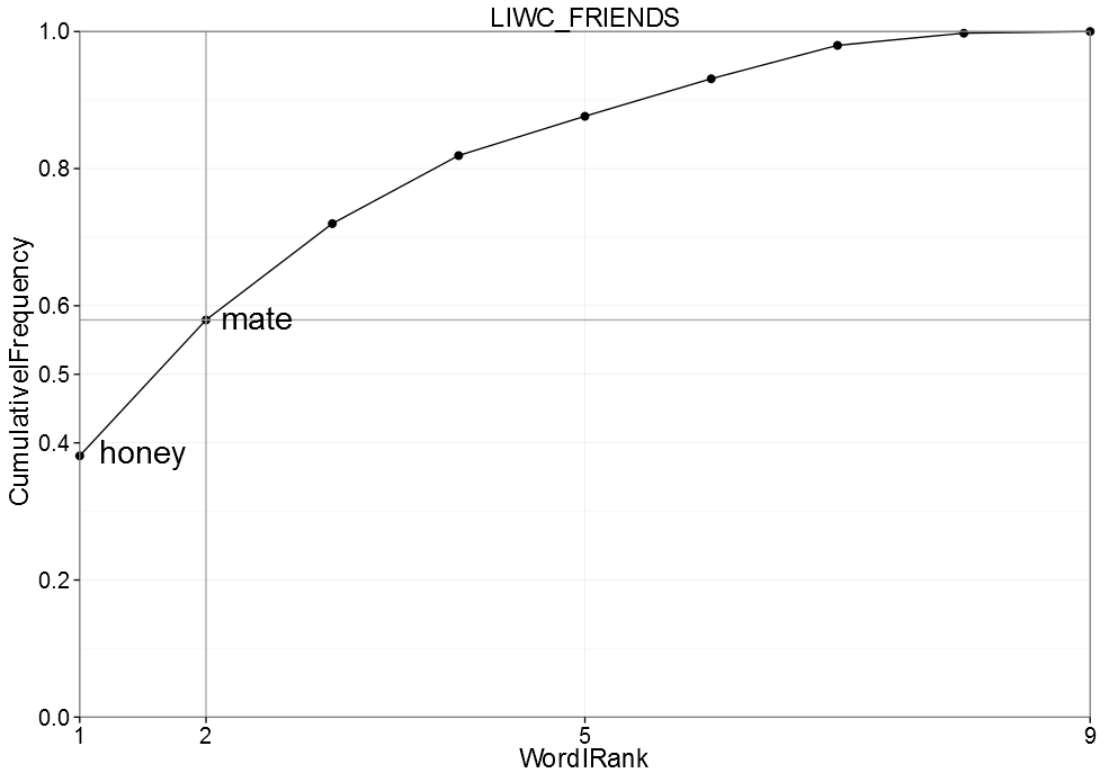
Category	Examples	Words in Category
Positive emotion	Love, nice, sweet	406
Negative emotion	Hurt, ugly, nasty	499
Anxiety	Worried, nervous	91
Anger	Hate, kill, annoyed	184
Sadness	Crying, grief, sad	101
Cognitive processes	Cause, know, ought	730
Insight	Think, know, consider	195
Causation	Because, effect, hence	108
Discrepancy	Should, would, could	76
Tentative	Maybe, perhaps, guess	155
Certainty	Always, never	83
Inhibition	Block, constrain, stop	111
Inclusive	And, with, include	18
Exclusive	But, without, exclude	17

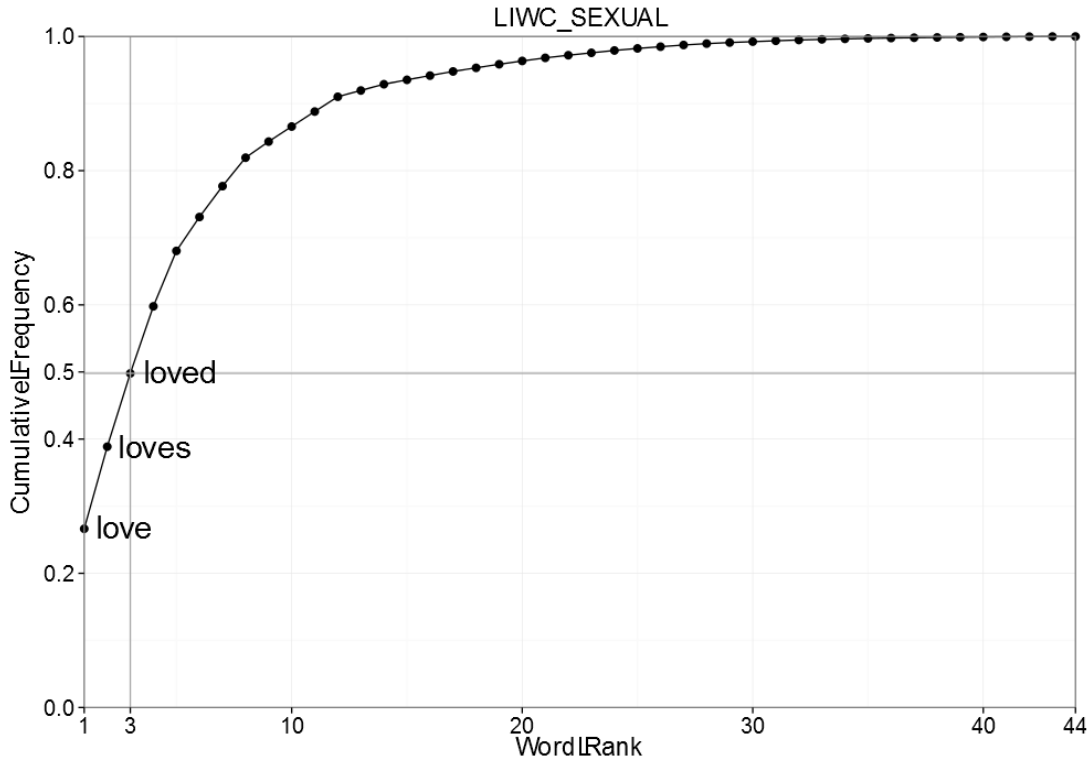
Words within LIWC 2015

FEMALE	Female references	her	32.2%	she	22.1%	girl	8.7%	mom	8.0%	she's	3.2%
MALE	Male references	he	24.0%	his	16.6%	man	12.5%	him	11.8%	boy	4.0%
COGPROC	Cognitive processes	all	5.8%	but	5.5%	not	5.0%	if	4.3%	know	3.0%
INSIGHT	Insight	know	16.9%	think	10.8%	feel	8.1%	find	4.3%	feeling	3.8%
CAUSE	Causation	how	20.3%	make	14.9%	why	13.3%	because	9.4%	made	7.6%
DISCREP	Discrepancy	if	23.4%	want	11.1%	need	10.1%	would	8.2%	should	5.8%
TENTAT	Tentative	if	19.5%	or	12.0%	some	10.7%	hope	4.4%	any	3.8%
CERTAIN	Certainty	all	39.8%	never	10.7%	ever	8.2%	always	7.1%	every	5.3%
DIFFER	Differentiation	but	18.1%	not	16.5%	if	14.4%	or	8.8%	really	6.7%
PERCEPT	Perceptual processes	see	9.7%	feel	6.0%	say	5.9%	watching	3.3%	look	3.2%
SEE	See	see	20.8%	watching	7.0%	look	6.8%	looking	5.9%	watch	5.2%
HEAR	Hear	say	29.8%	said	13.3%	says	8.0%	hear	6.6%	listening	4.8%
FEEL	Feel	feel	20.3%	feeling	9.5%	hard	8.8%	cold	6.1%	hot	6.0%
BIO	Biological processes	love	19.3%	life	11.0%	sleep	4.7%	tired	3.5%	heart	3.5%
BODY	Body	sleep	15.6%	heart	11.4%	head	8.2%	face	7.2%	ass	5.1%
HEALTH	Health	life	36.9%	tired	11.7%	sick	9.1%	live	8.8%	pain	4.8%
SEXUAL	Sexual	fuck	46.9%	gay	11.0%	sex	8.9%	sexy	8.5%	dick	3.7%

Source: [1]

Unequal distribution within dictionaries (LIWC 2007)





Source: [1]

Most words in dictionaries don’t matter!!

For all LIWC 2015 frequency distributions, see [1] and https://osf.io/h4y56/

Summary – Top Down language analyses

- LIWC 2015 provides the cleanest/best set of dictionaries
- It is split into “function” word and “content” word dictionaries.
- Using LIWC allows for easy connections to literature
- Function word (mostly pronoun) dictionaries have yielded theoretically interesting findings.

However:

- Content word (work, social…) dictionaries rarely lead to insights
- Some dictionaries are driven by very few words.

Can we do this differently?
Bottom up methods

Feature extraction –
Language Analysis Methods (psychology view)

Bottom up

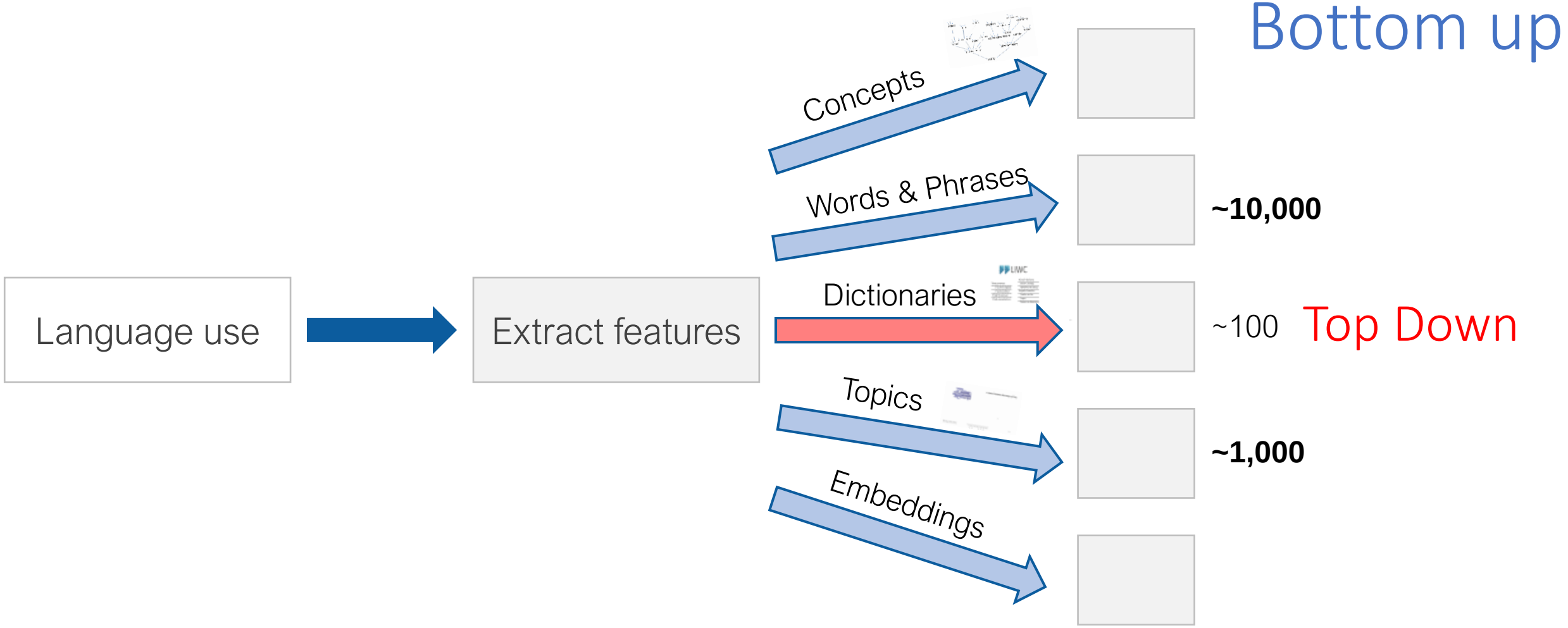
Language variables

Also:
“data-driven”
“open vocabulary”

Examples:
Topic modelling,
Latent Semantic Analysis,
Embeddings,
Differential Language Analysis

Language data

Basic Framework: Feature extraction



The Steps for Language Analysis

1) Collect/get data

2) Tokenize the text

3) Select language features (words? dictionaries?)

3) Feature Extraction, Revisited: words and phrases

Pick a larger language feature set!

Up to 20,000 words and phrases:
1-to-3 word sequences more likely to occur together than chance.

Up to 2,000 LDA topics:
Clusters of semantically-related words found via *latent Dirichlet allocation*

But do you have the sample size?

Power

Typical LIWC studies

Typical open-vocabulary study

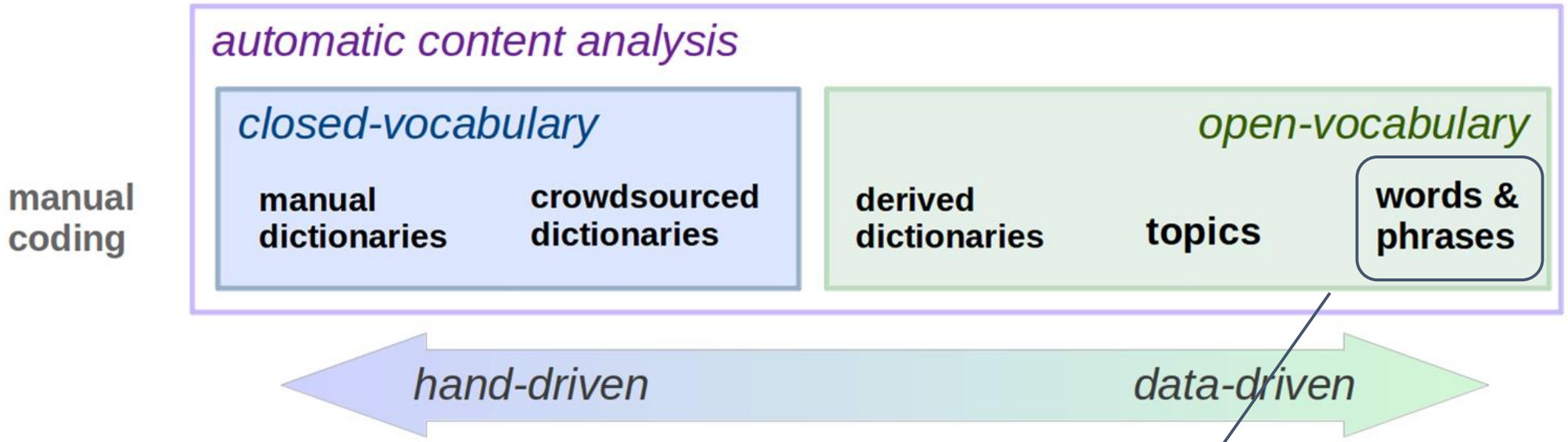
Minimal Samples needed for Exploratory Language Analyses

Thresholds of significant correlates:	Demographics		Big Five Personality					
	Gender	Age	Agr.	Con	Ext.	Neur.	Ope.	**(avg.)**
10 (out of 73) LIWC dictionaries	200	150	800	400	800	1,100	550	**750**
100 (out of 2,000) LDA topics	250	150	1,100	550	800	1,800	550	**1,000**
200 (out of 11,894) 1-to-3 grams	650	200	3,650	1,850	2,600	4,750	2,100	**3,000**

Note. Sample sizes (N) needed to observe 10 significantly associated LIWC dictionaries (out of 73), 100 LDA topics (out of 2,000) or 200 1-to-3 grams (out of 11,894) for gender, age, and personality (using all the users' Facebook posts). The significance threshold of $p = .05$ was Benjamini-Hochberg corrected for multiple comparisons.

Source: [1]

Words and phrases



+ wide coverage
+ fine-grained information
phrases: capture some context

(not as “digestable”)

Worked Example

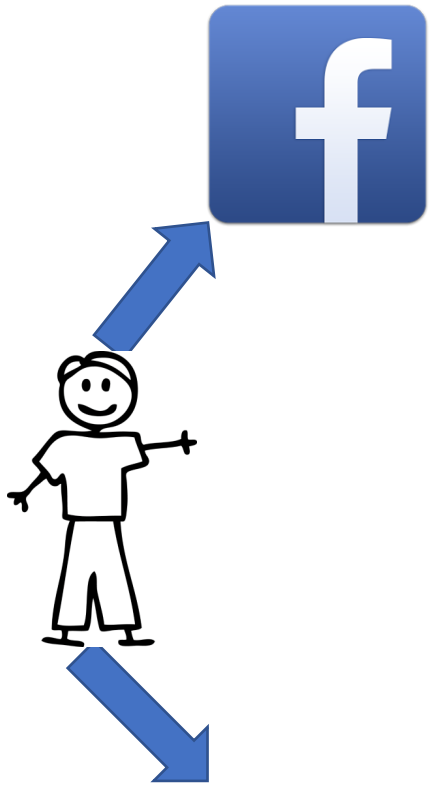
"The cat cat died unexpectedly:("

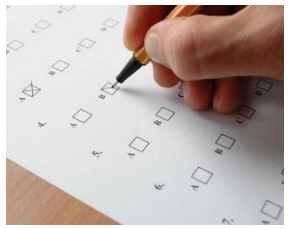
Extraction of word language features.

Text ID	token	count
1	The	1
1	cat	2
1	died	1
1	unexpectedly	1
1	:(	1
2	(…)	(…)

Text ID	token	relativeFrequency
1	The	0.166
1	cat	0.333
1	died	0.166
1	unexpectedly	0.166
1	:(	0.166
2	(…)	(…)

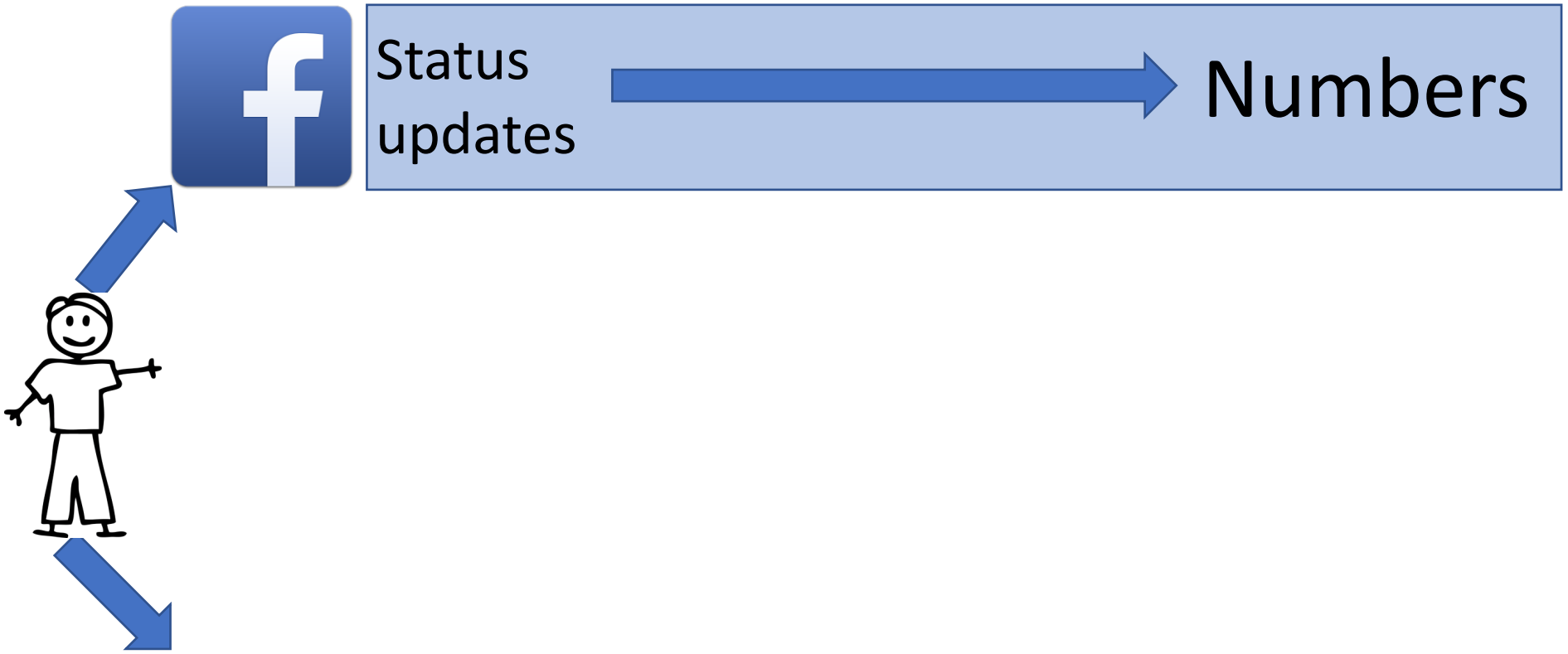
Differential Language Analysis



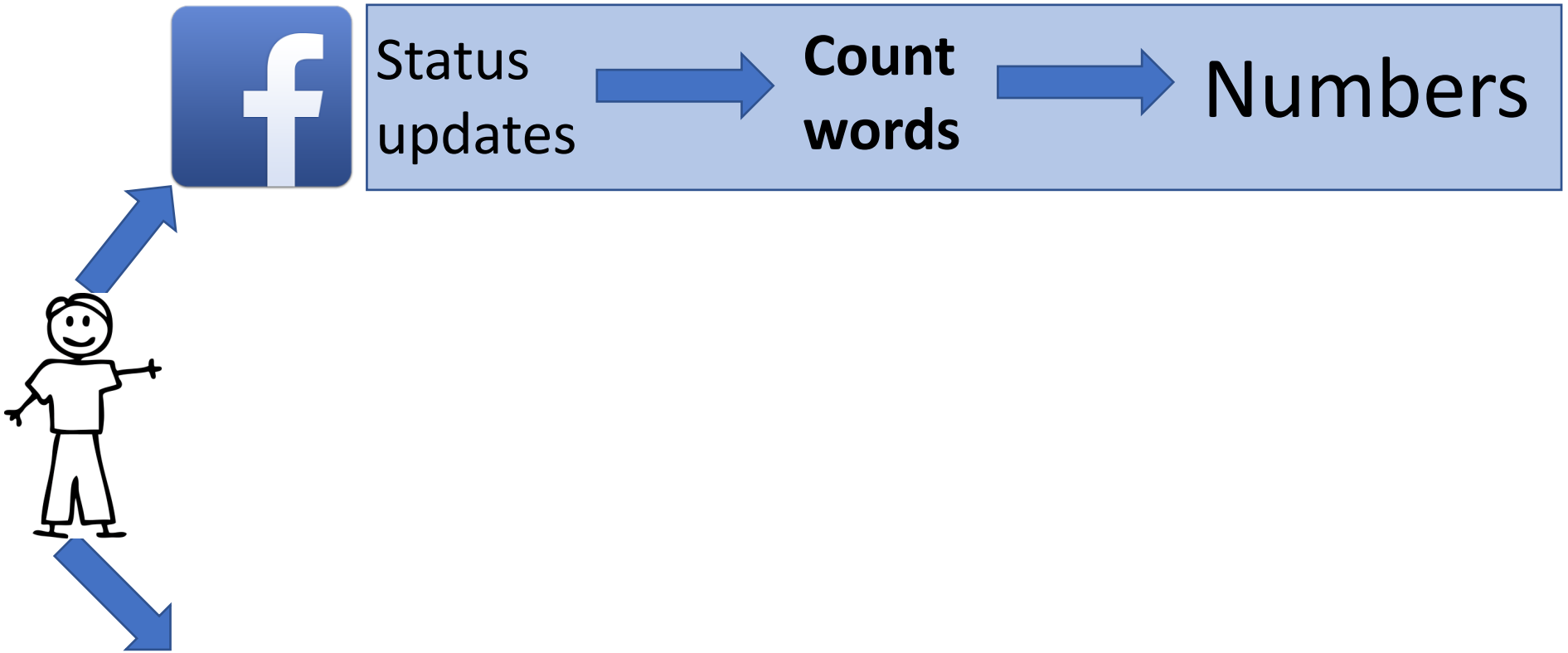


Normal Variables (gender, Big5…)

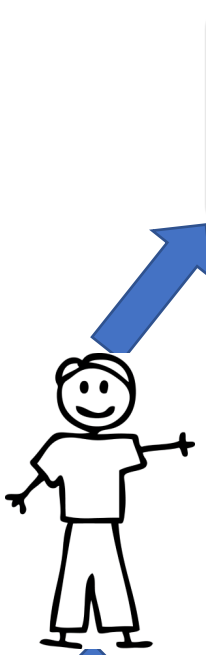
Differential Language Analysis

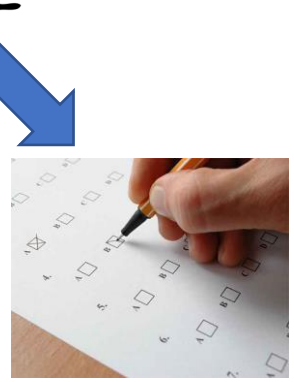


Differential Language Analysis



Differential Language Analysis





Status updates → **Count words** → Relative Frequencies

Correlate every word

show most correlated

Normal Variables (gender, Big5…)

Language of **Women**

Schwartz, Eichstaedt, Kern et al., 2013, *PLoSone*

Linguistic Inquiry and Word Count (LIWC 2015)

LIWC (other)		LIWC (psych. processes)	
Dictionary	β	Dictionary	β
Emotional tone (m)	.27	Social processes	.12
Personal pronoun	.17	Female reference	.30
1st pers singular	.16	Family	.28
3rd pers singular	.11	Affective process	.25
2nd person	.07	Positive emotion	.29
Total pronouns	.11	Home	.21
Common adverbs	.09	Netspeak	.18
Common verbs	.07	Affiliation	.17
Conjunctions	.07	Future focus	.10
Common adjectives	.06	Nonfluencies	.10

Language of **Men**

Schwartz, Eichstaedt, Kern et al., 2013, *PLoSone*

Articles	.24
Analytical thinking (m)	.19
Comparisons	.12
Prepositions	.12
Impersonal pronouns	.08
Quantifiers	.06
Interrogatives	.06
3rd pers plural	.05
Numbers	.04

Death	.22
Anger	.21
Drives	
Power	.20
Achievement	.13
Risk	.09
Swear words	.19
Sexual	.19
Space	.16
Money	.11
Tentative	.09

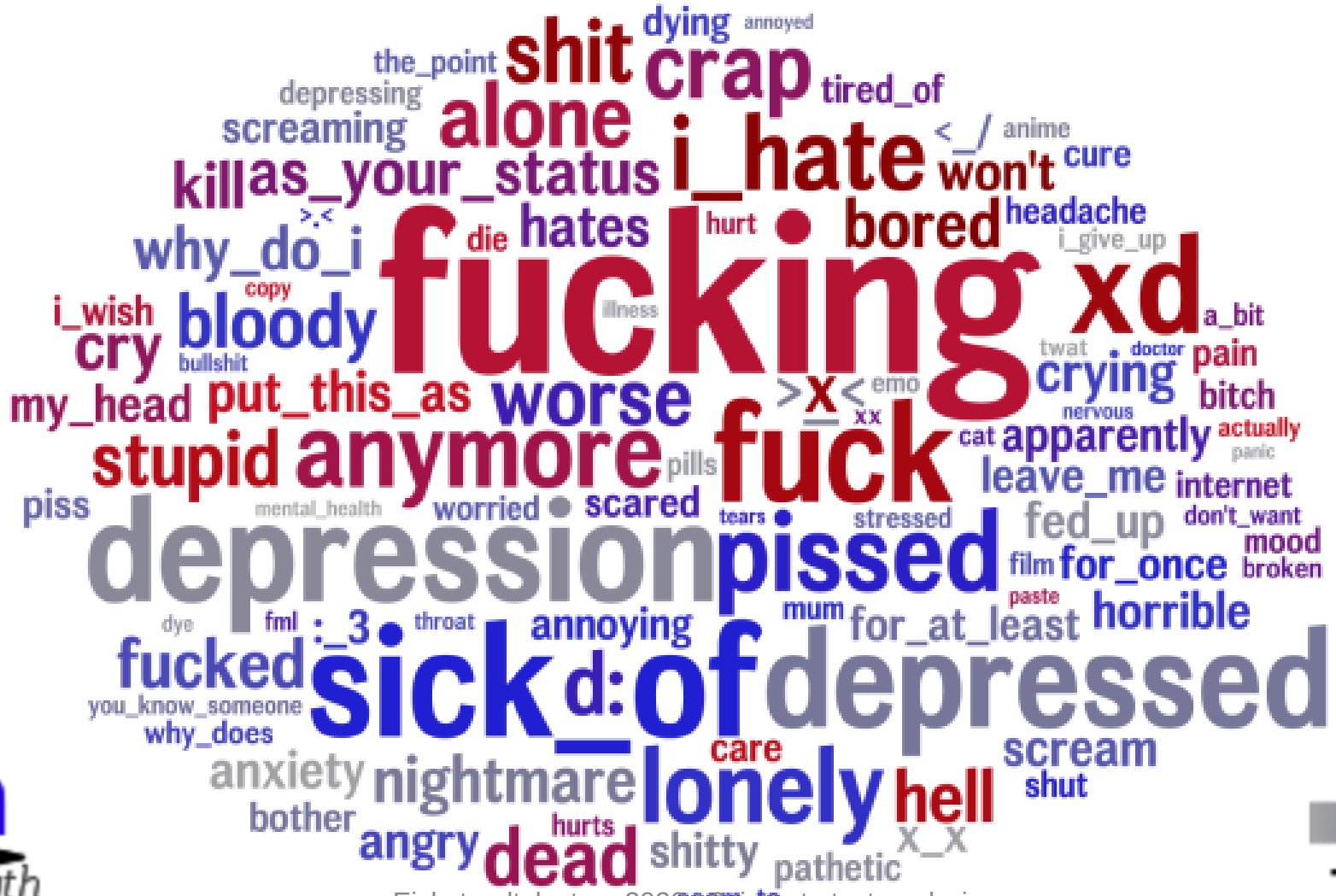
Language of **Conscientiousness**

a a a

correlation strength

relative frequency

Language of **Neuroticism**



a a a

correlation strength

relative frequency

Language of **Emotional Stability**

a a a

correlation strength

relative frequency

Language of **Agreeableness**

a a a

correlation strength

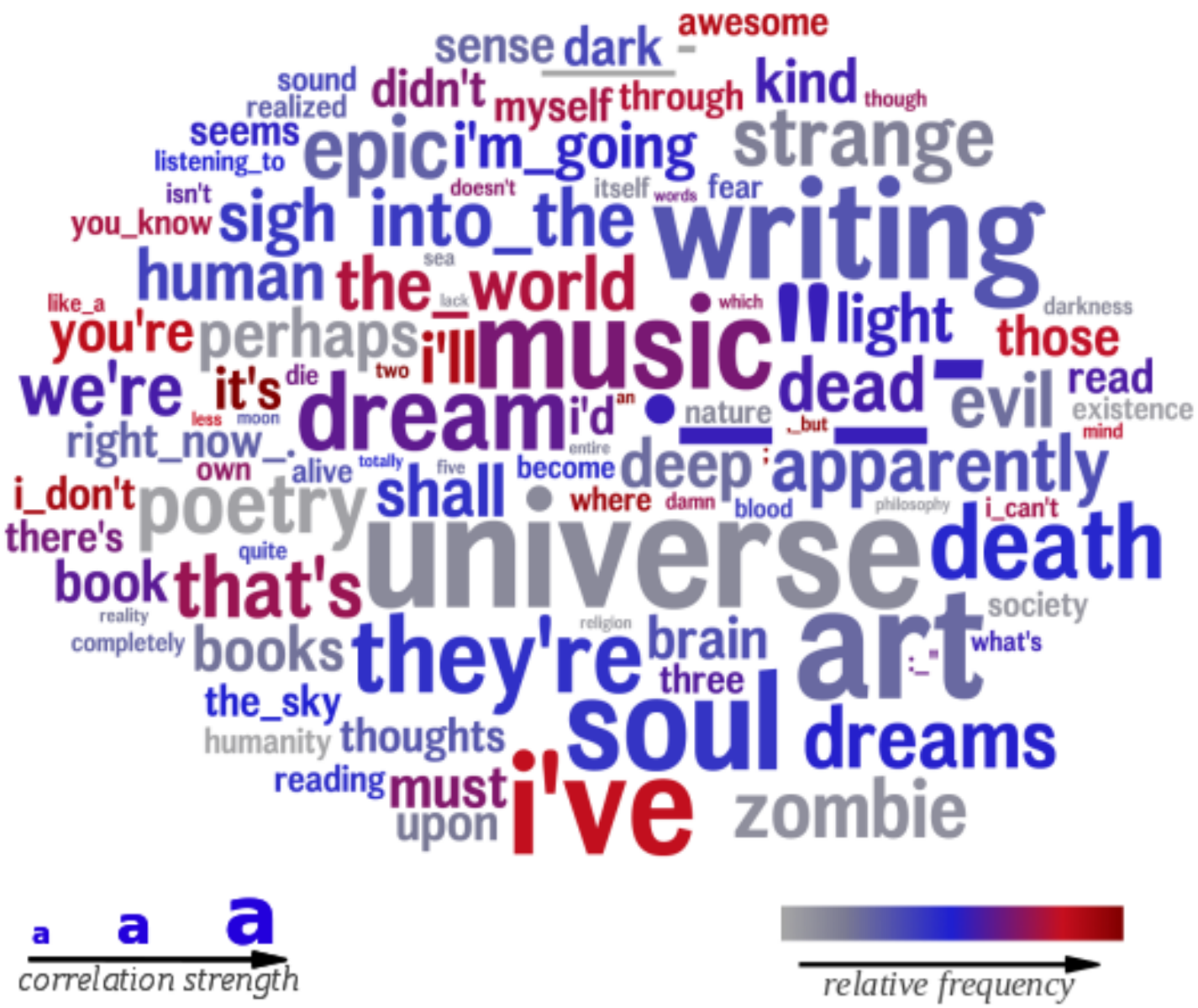
relative frequency

Language of **Extraversion**

correlation strength

relative frequency

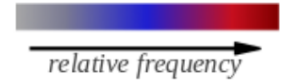
Openness

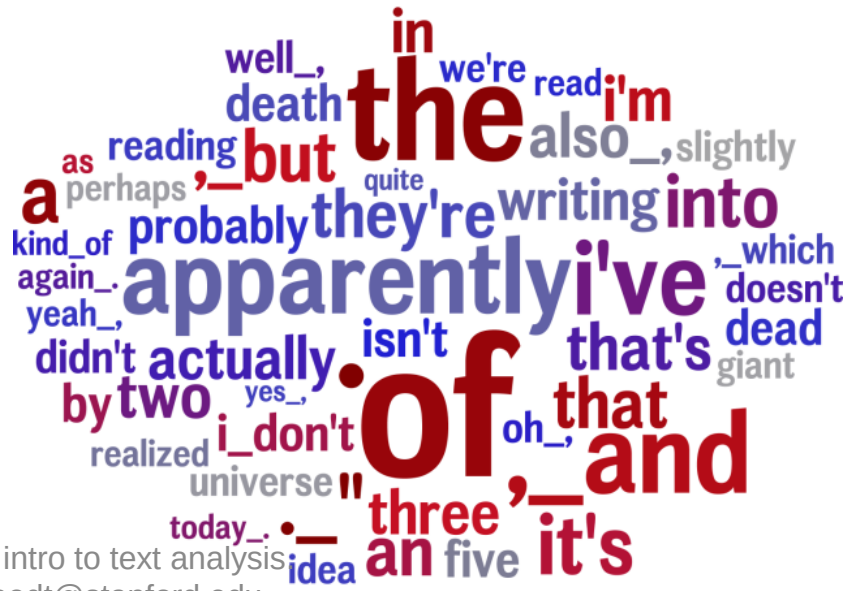


Linguistic Inquiry and Word Count (LIWC 2015)			
LIWC (other)		**LIWC (psych. processes)**	
Dictionary	β	Dictionary	β
Articles	.15	Cognitive process	.09
Total function words	.08	Insight	.12
Auxiliary verbs	.07	Causation	.07
Comparisons	.06	Tentative	.07
Impersonal pronouns	.06	Death	.12
Conjunctions	.06	Perceptual process	.12
Prepositions	.05	Hear	.08
1st pers singular	.04	See	.07
Interrogatives	.04	Anxiety	.08
Quantifiers	.04	Space	.05

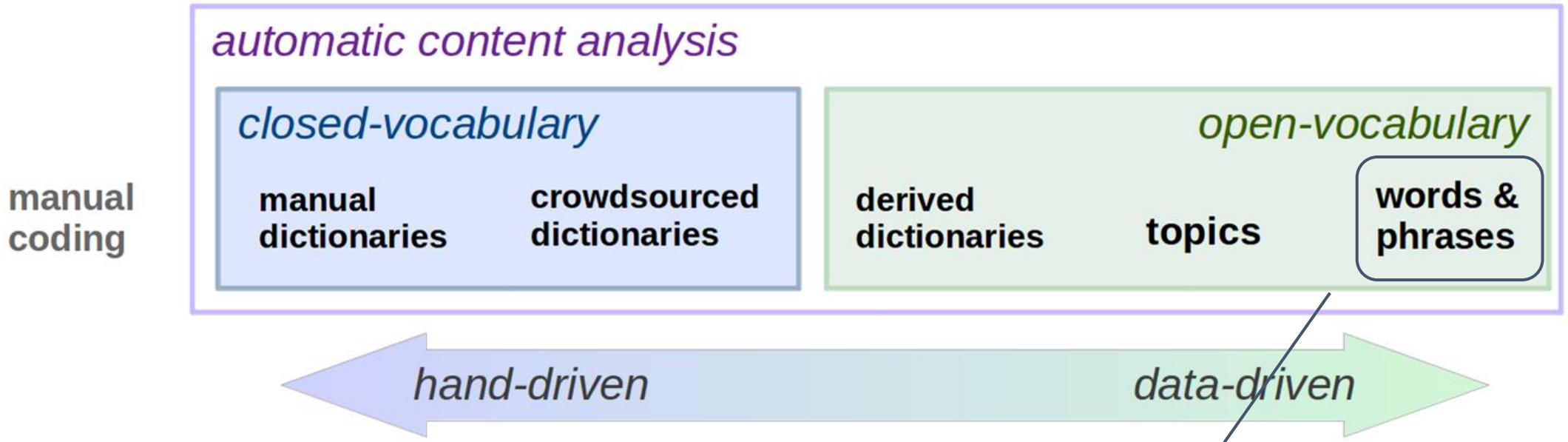
Source: [1]

Openness Facets



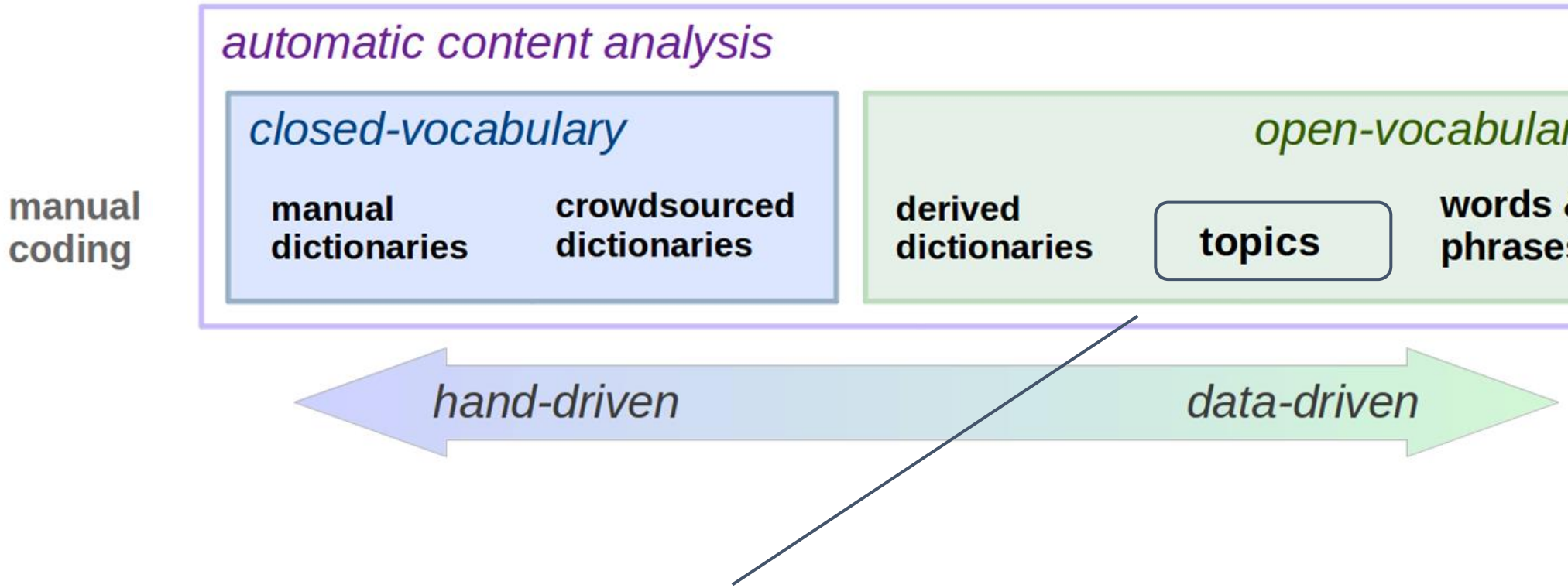


Topic modeling



+ wide coverage
+ fine-grained information
phrases: capture some context

(not as “digestable”)



Topic Modeling:

+ completely data-driven
+ “digestable”
(still losing some information)

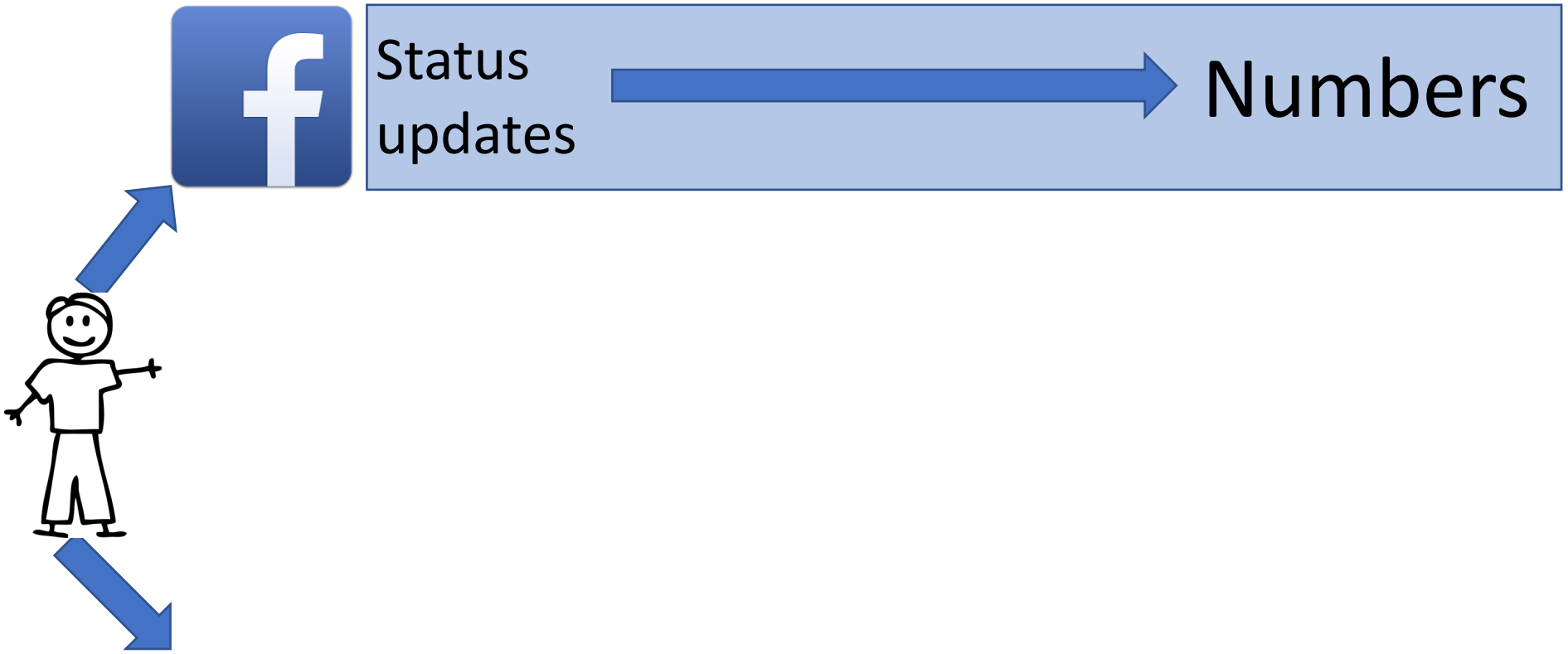
Latent Dirichlet topic models -- examples

(from: ALL topics from a 2,000 topics run over 14 m statuses at wwbp.org/data)

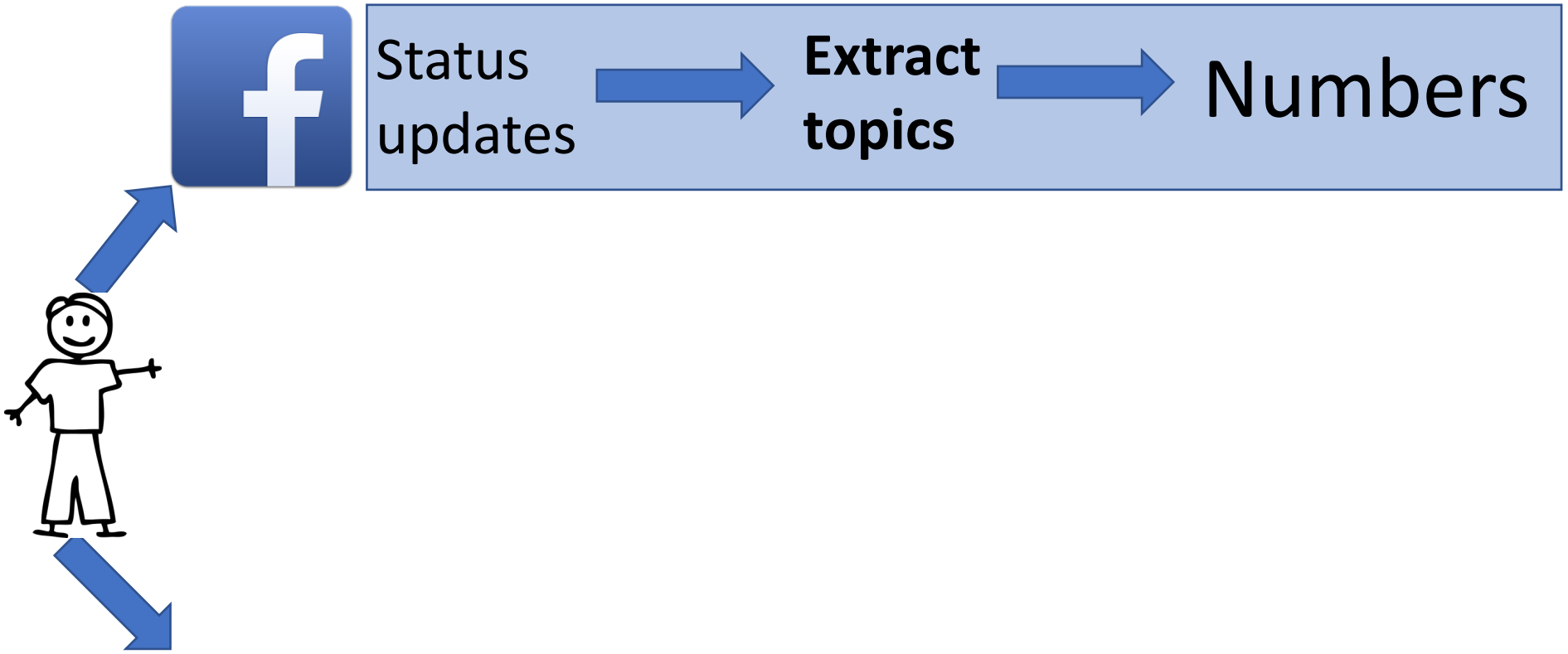
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

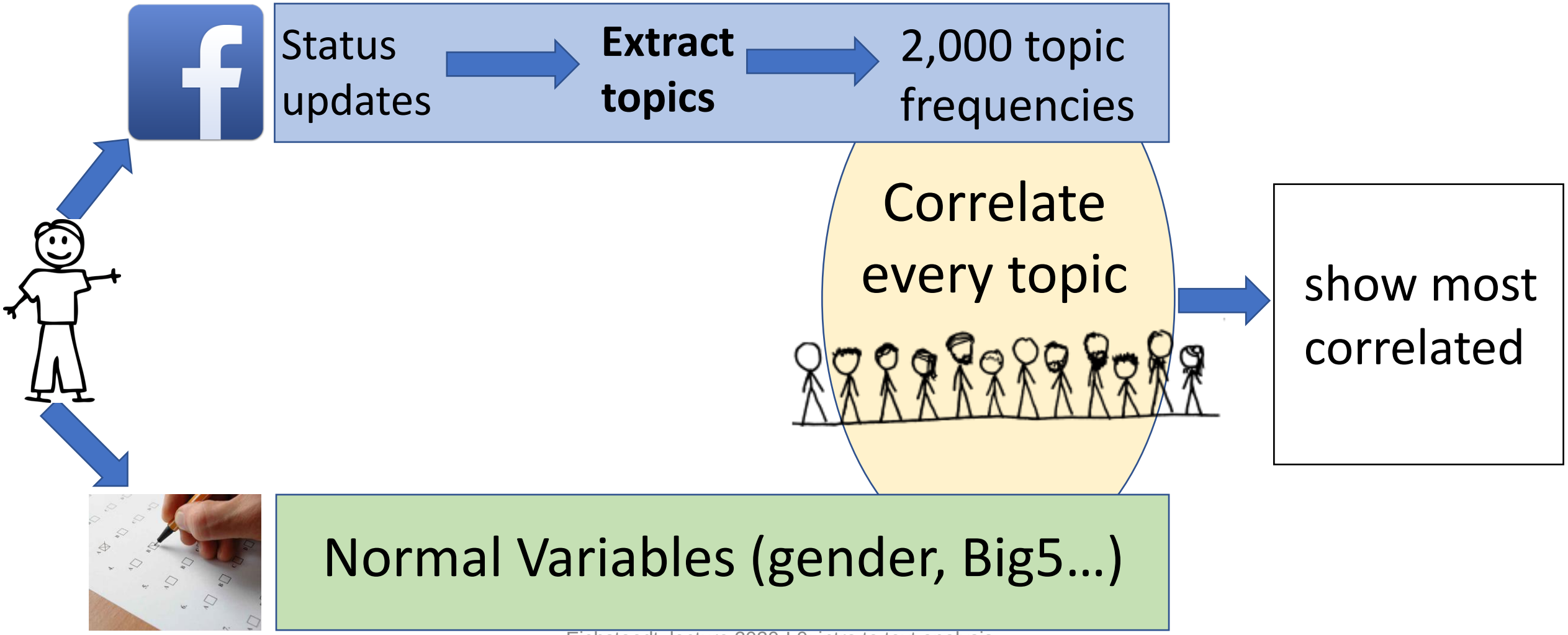
Differential Language Analysis



Differential Language Analysis



Differential Language Analysis

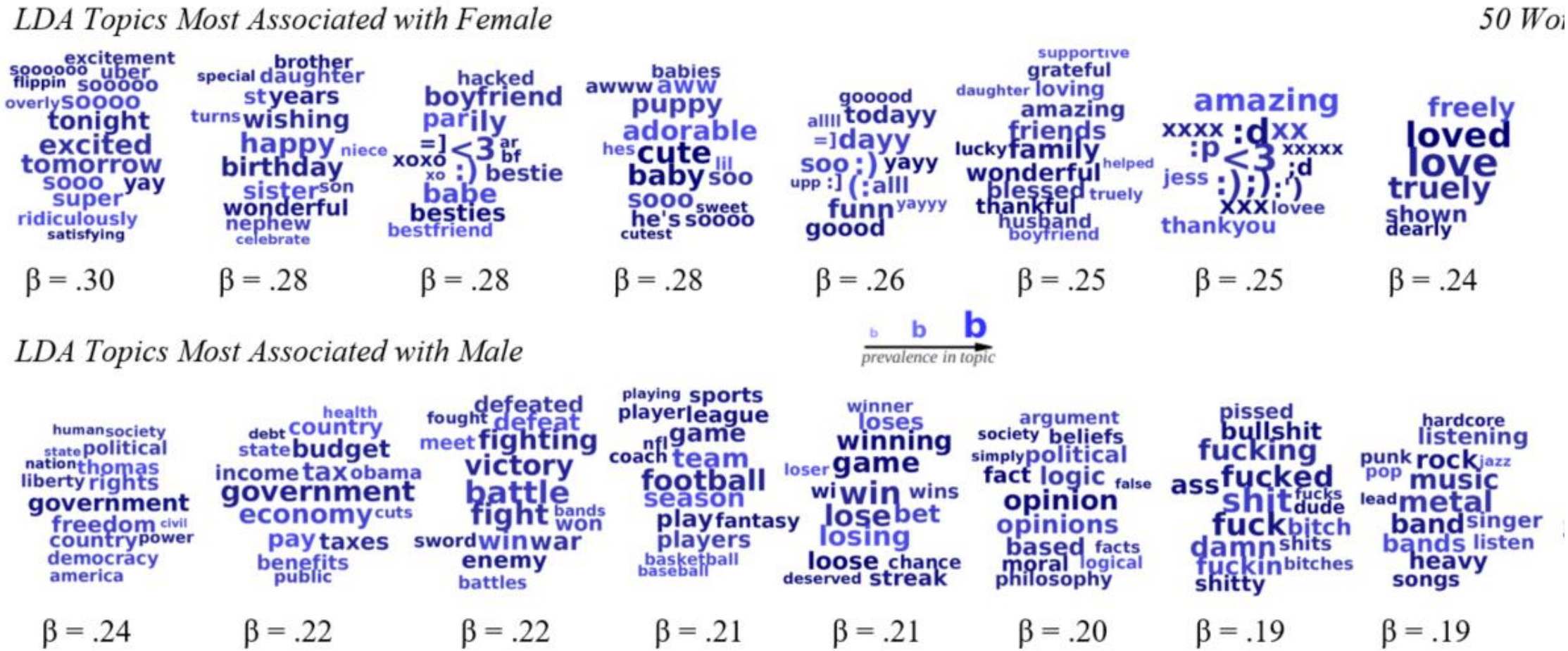


Most correlated gender topics

LDA Topics Most Associated with Female *50 Wo*

β = .30 β = .28 β = .28 β = .28 β = .26 β = .25 β = .25 β = .24

Most correlated gender topics



Full comparison

	General Inquirer						DICTION		Linguistic Inquiry and Word Count (LIWC 2015)			
	Lasswell		Harvard IV		Stanford				LIWC (other)		LIWC (psych. processes)	
	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β	Dictionary	β
Female	Affect		Pleasure	.29	Affiliation	.12	Optimism (m)	.14	Emotional tone (m)	.27	Social processes	.12
	Affect-Other	.28	Females	.28	Passive	.09	+Satisfaction	.22	Personal pronoun	.17	Female reference	.30
	Affect-Domain	.21	Emotion	.25	Positive	.09	+Praise	.08	1st pers singular	.16	Family	.28
	Affect-Gain	.16	Kinship	.20	Weak	.06	+Inspiration	.05	3rd pers singular	.11	Affective process	.25
	Affect-Participants	.05	Self	.15	Submit	.05	-Blame	.04	2nd person	.07	Positive emotion	.29
	Wellbeing-Total	.15	Children	.15			Certainty (m)		Total pronouns	.11	Home	.21
	Wellbeing-Psych.	.24	Independent Adj.	.12			+Insistence	.07	Common adverbs	.09	Netspeak	.18
	Wellbeing-Participants	.16	State Verb	.12			-Self-reference	.15	Common verbs	.07	Affiliation	.17
	Positive-Affect	.11	Need	.11			+Tenacity	.06	Conjunctions	.07	Future focus	.10
	Transaction-Gain	.10	Evaluation 2	.10			Human Interest	.12	Common adjectives	.06	Nonfluencies	.10
	Respect-Lose	.07					Temporal	.05				
Male	Wealth-Total	.19	Military	.21	Strength	.09	Realism		Articles	.24	Death	.22
	Wealth-Other	.19	Movement-Exert	.21	Hostile	.08	+Familiarity	.09	Analytical thinking (m)	.19	Anger	.21
	Power-Total	.18	Political	.19	Negative	.07	+Spatial	.09	Comparisons	.12	Drives	
	Power-Arenas	.15	Economic	.16	Understated	.06	-Complexity	.08	Prepositions	.12	Power	.20
	Power-Conflict	.14	Region	.15	Active	.06	Activity		Impersonal pronouns	.08	Achievement	.13
	Power-Participants (Ordinary)	.14	Space	.15	Power	.06	+Aggression	.10	Quantifiers	.06	Risk	.09
			Doctrine	.15			+Accomplishment	.07	Interrogatives	.06	Swear words	.19
	Power-Authority	.13	Abstract vocab.	.14			+Communication	.07	3rd pers plural	.05	Sexual	.19
	Power-Loss	.12	Collectives	.14			Commonality		Numbers	.04	Space	.16
	Arenas	.17	Expressive	.13			+Centrality	.08			Money	.11
	Religion	.14					-Diversity	.06			Tentative	.09
							-Exclusion	.05				
							Collectives	.06				

Note. All coefficients are significant at $p < .001$, corrected for multiple comparisons. (m) designates "master" categories that combine frequencies of multiple dictionaries.

LDA Topics Most Associated with Female

β = .30 β = .28 β = .28 β = .28 β = .26 β = .25 β = .25 β = .24

50 Words and Phrases Most Associated with Female

LDA Topics Most Associated with Male

β = .24 β = .22 β = .22 β = .21 β = .21 β = .20 β = .19 β = .19

Male

Source: [1]

The three-legged stool of language analysis

Summary – Bottom up language analyses

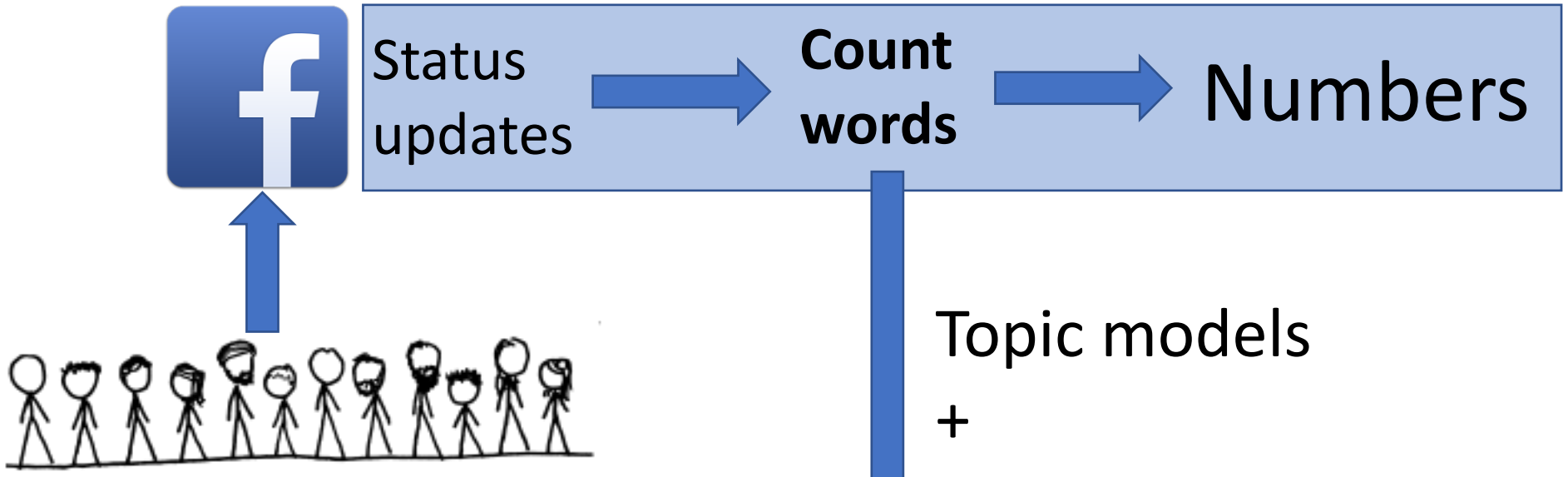
- Open vocabulary results are great for construct exploration.
- You need to have sufficient sample sizes
- N > 1,000 x 500 words each for topics
- N > 3,000 x 500 words each for words and phrases
- You can pick the number of topics to model – too many will create duplicates. You can filter for duplicates.

However:

- Results can become unwieldy.
- Always combine with LIWC 2015 analyses

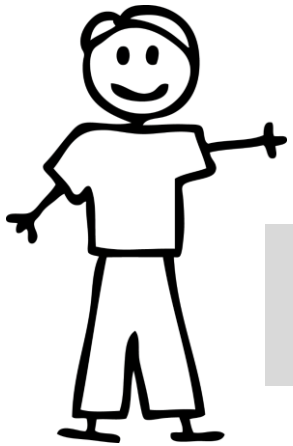
Can we use language to predict/measure?

Language **Prediction**



Schwartz et al., 2013, *PLoSone*

Predicting **Gender**



Gender

?

Gender

Accuracy? **92%**

Sap, Park, Eichstaedt et al., 2014, *EMNLP*

Predicting Personality

Personality

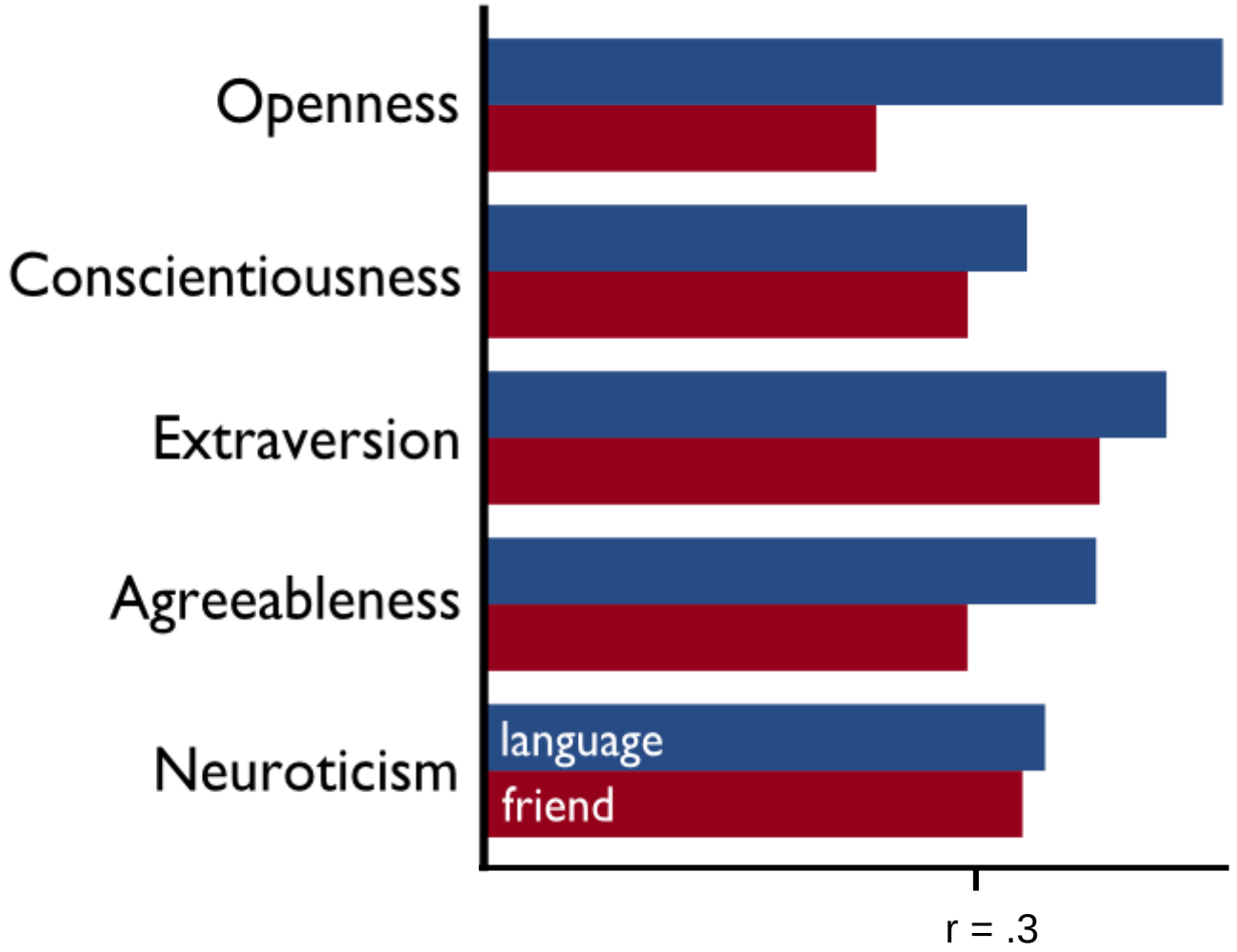
?

Personality

?

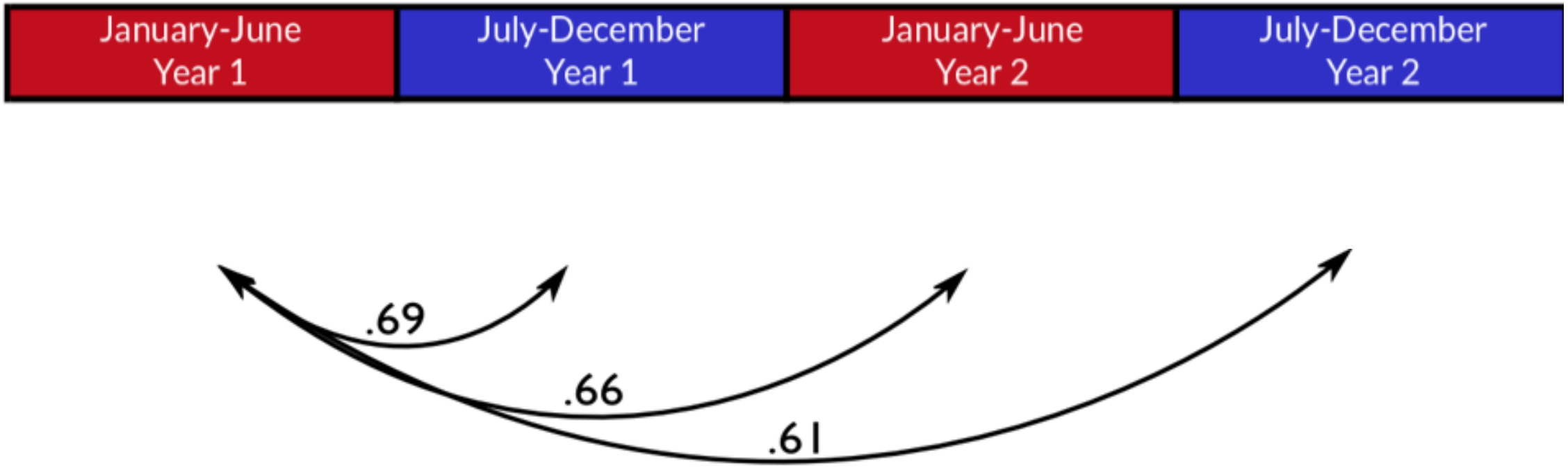
Park, Schwartz, Eichstaedt et al., 2014, *JPSP*

Predicting Personality: **Accuracy**



Park, Schwartz, Eichstaedt et al., 2014, *JPSP*

Predicting Personality: Test-Retest Reliability



Park, Schwartz, Eichstaedt et al., 2014, *JPSP*

Super quick intro to machine learning

Build predictive model

Step 1: Pick **feature set** (topics? Words and phrases?)

- Match number of features and observations
  (you have 3,000 people? Use less than 3,000 variables per person)

Step 2:

Build model << Insert machine learning here >>

- Usually linear or logistic regression (with regularization)

Test out-of-sample (this is cross-validation)

- Usually 10-fold

"Supervised learning"

Ok, step 2: Machine Learning

Basic recipe: Two conceptual ingredients

1) Build model with Regularization
- With Hyperparameter tuning

2) Cross-validation (test on new data)

The problem: high dimensional data

$$\begin{cases} x - 2y + z + 3t = 1 \\ 2x - 2y - 2z - 2t = 5 \\ x - 0.25y + 4z + 7t = -7 \\ x + y + z + t = 3 \end{cases}$$

More equations (observations = N) than variables/parameters (p)? **Great!**

Same number of equations (observations = N) as variables/parameters (p)? **NP!**

More variables/parameters (p) than equations (observations = N)? **Machine learning!**
Dimensionality reduction!

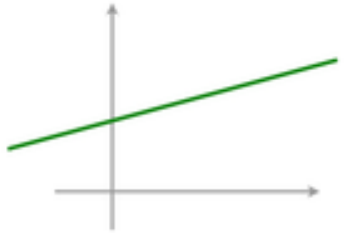
Intuition: higher order polynomials

1st degree polynomial

$y = a + bx^1$

straight line with no peaks and no valleys

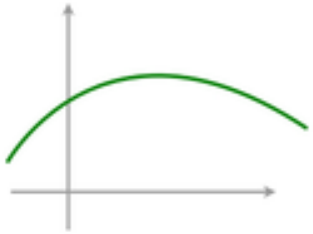
often written as $y = a + bx$



2nd degree polynomial

$y = a + bx^1 + cx^2$

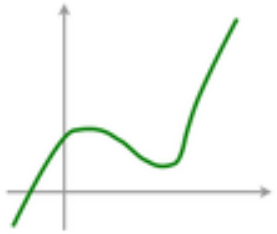
curved line with only one peak or one valley.

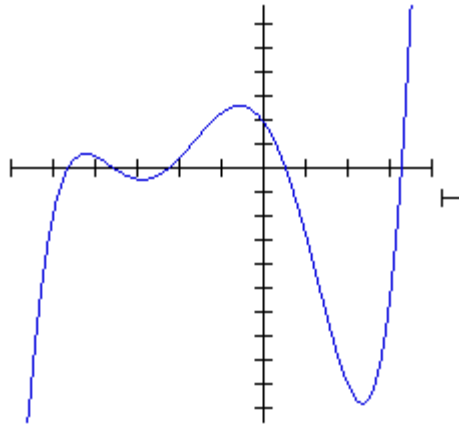


3rd degree polynomial

$y = a + bx^1 + cx^2 + dx^3$

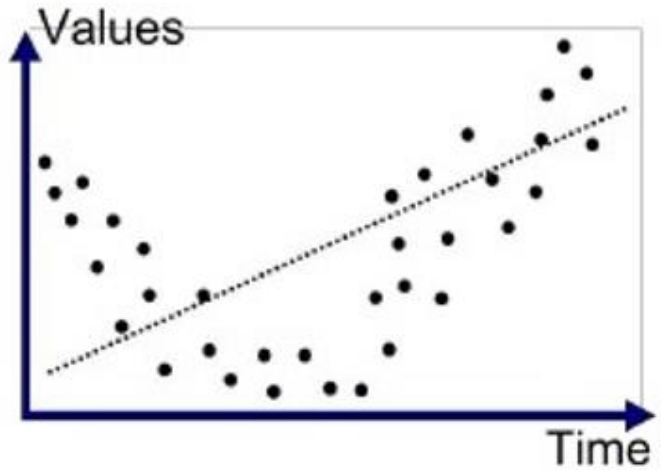
curved line with multiple peaks & valley.



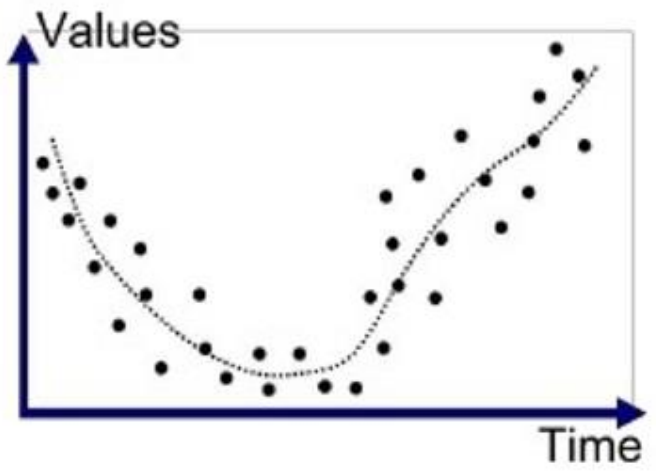


Fifth order polynomial

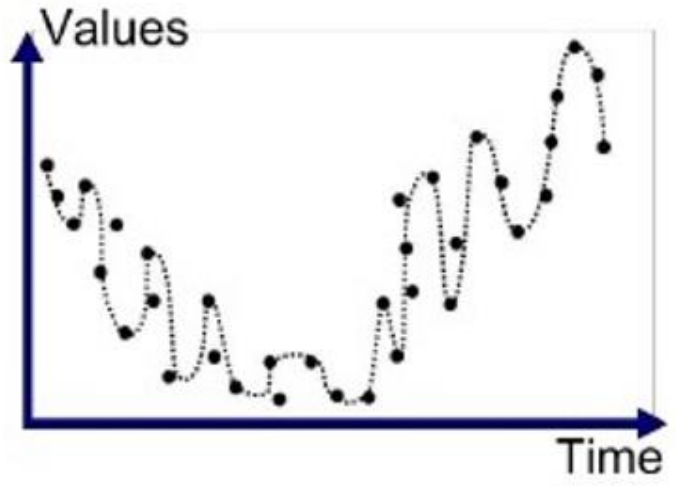
1) Regularization



Underfitted

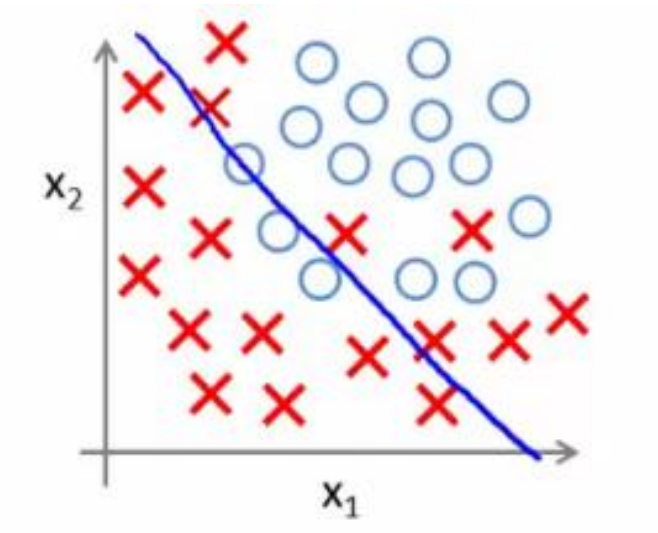


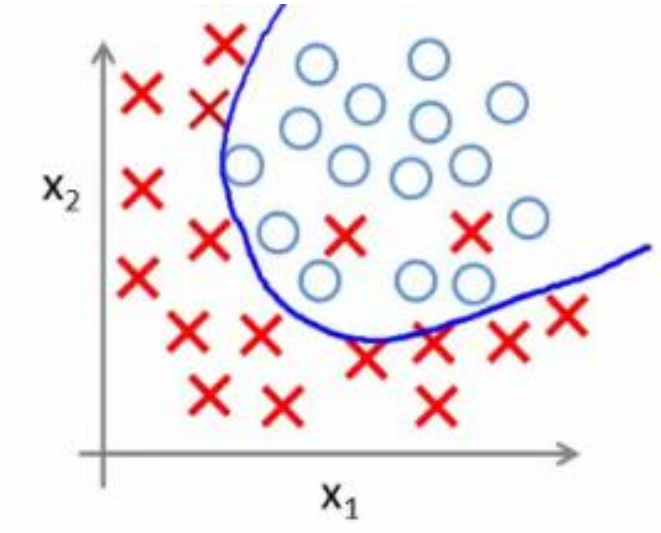
Good Fit/Robust

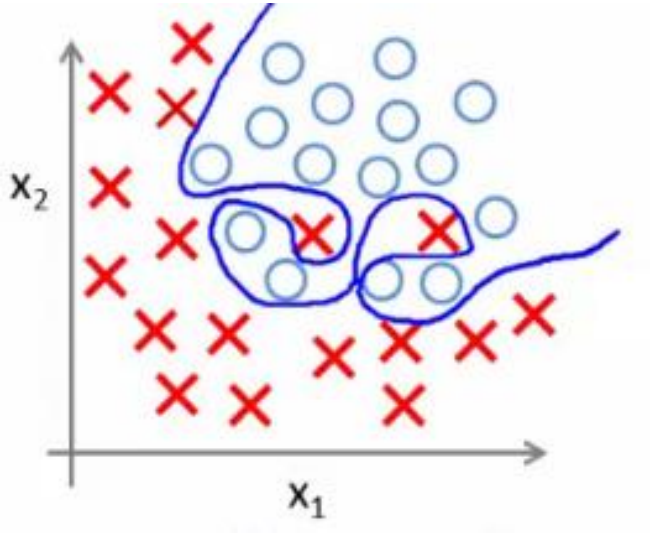


Overfitted

1) Regularization







Ordinary Least-Squares Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}}$$

(from: http://www.stat.cmu.edu/~ryantibs/datamining/lectures/16-modr1.pdf)

Ridge Regression

1) Regularization

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$$

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Type of penality

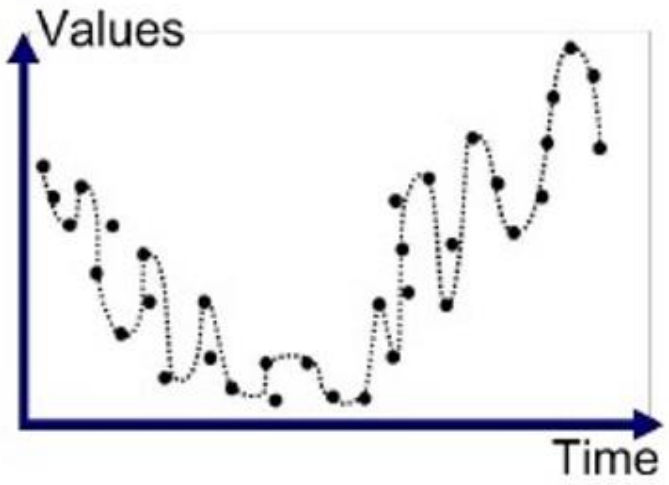
$$= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

Hyperparameter

"Penalty" = "Regularization"

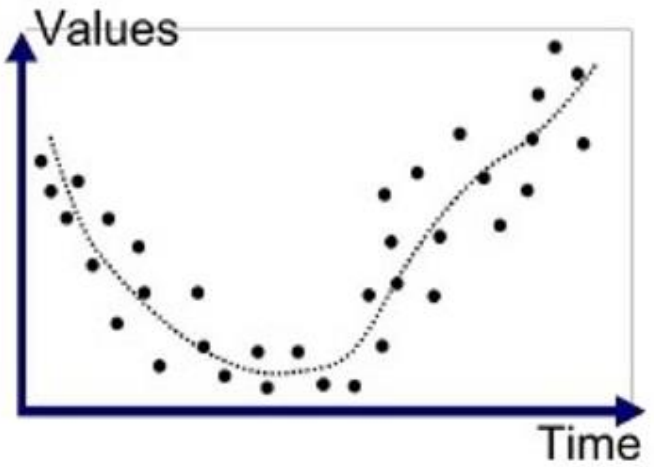
(from: http://www.stat.cmu.edu/~ryantibs/datamining/lectures/16-modr1.pdf)

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$



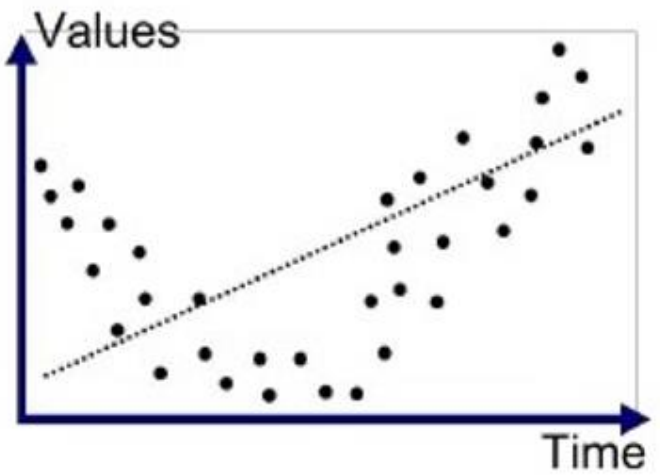
Overfitted

λ too low



Good Fit/Robust

λ just right
(when tested on new data)



Underfitted

λ too high

λ is your hyperparameter

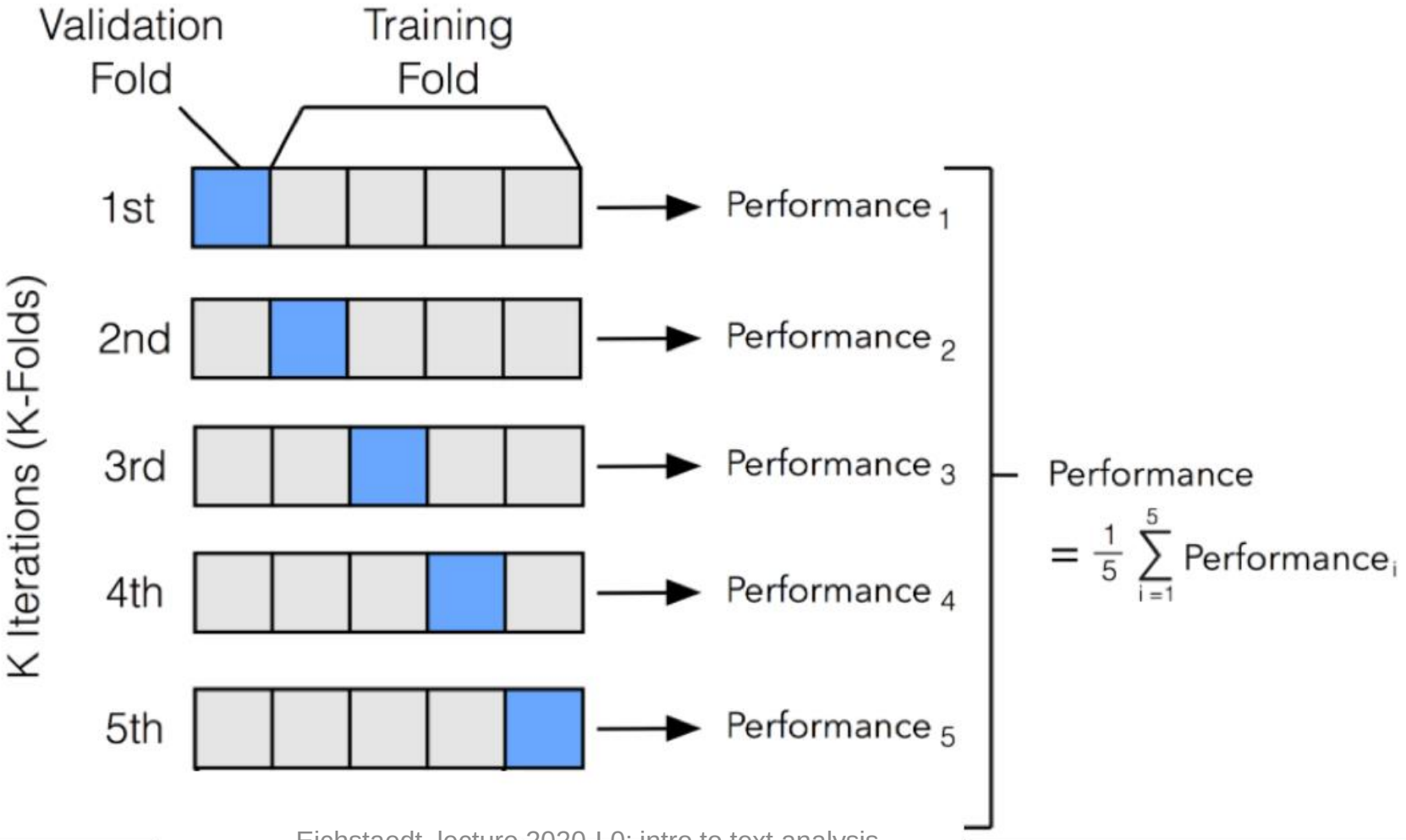
How do you find a good λ?

You try all of them…

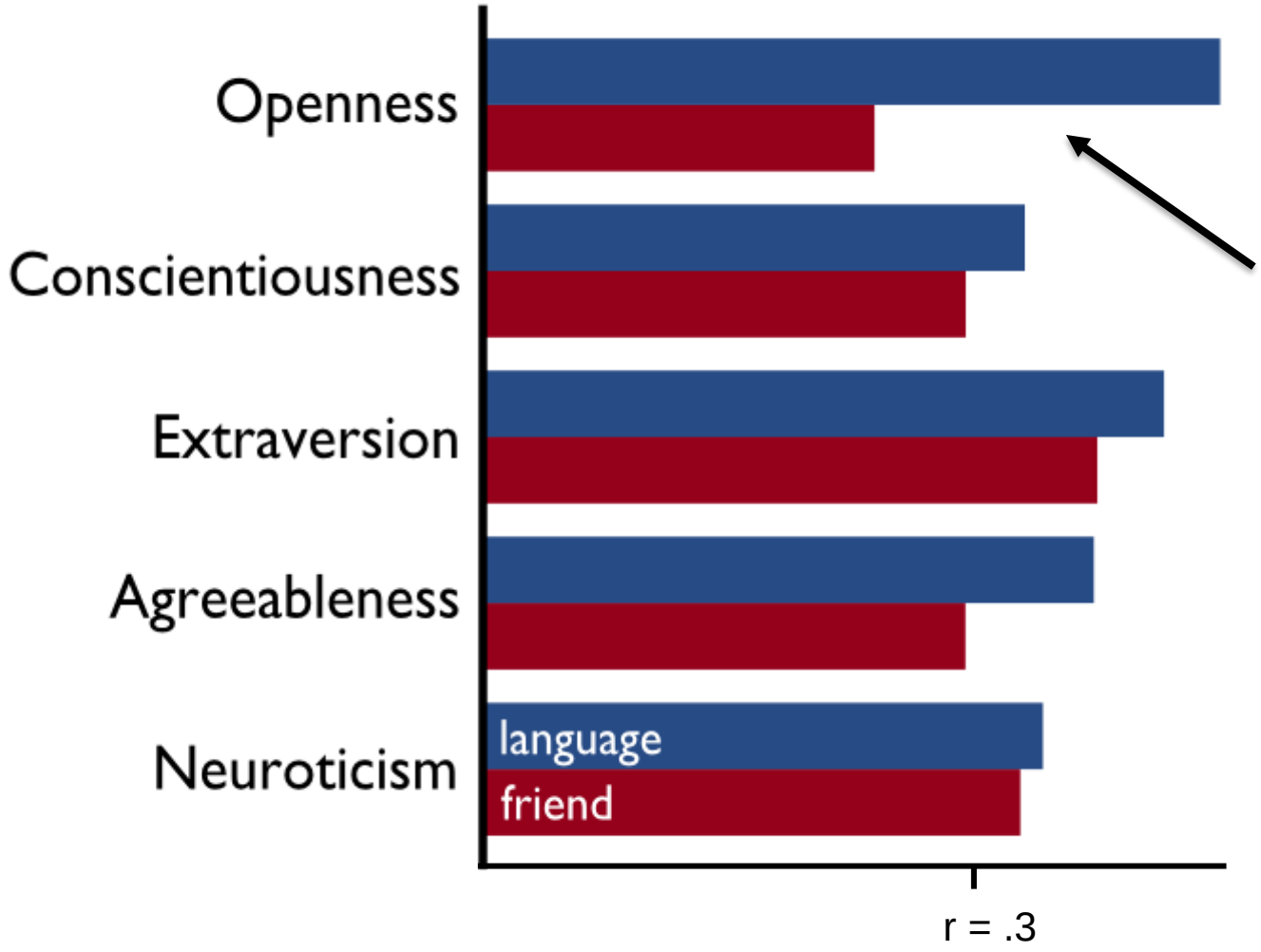
…BUT always test on new data.

You keep the one that best generalizes to new data.

Try different λ's in cross-validation.



Predicting Personality: Accuracy with language (cross-validated)

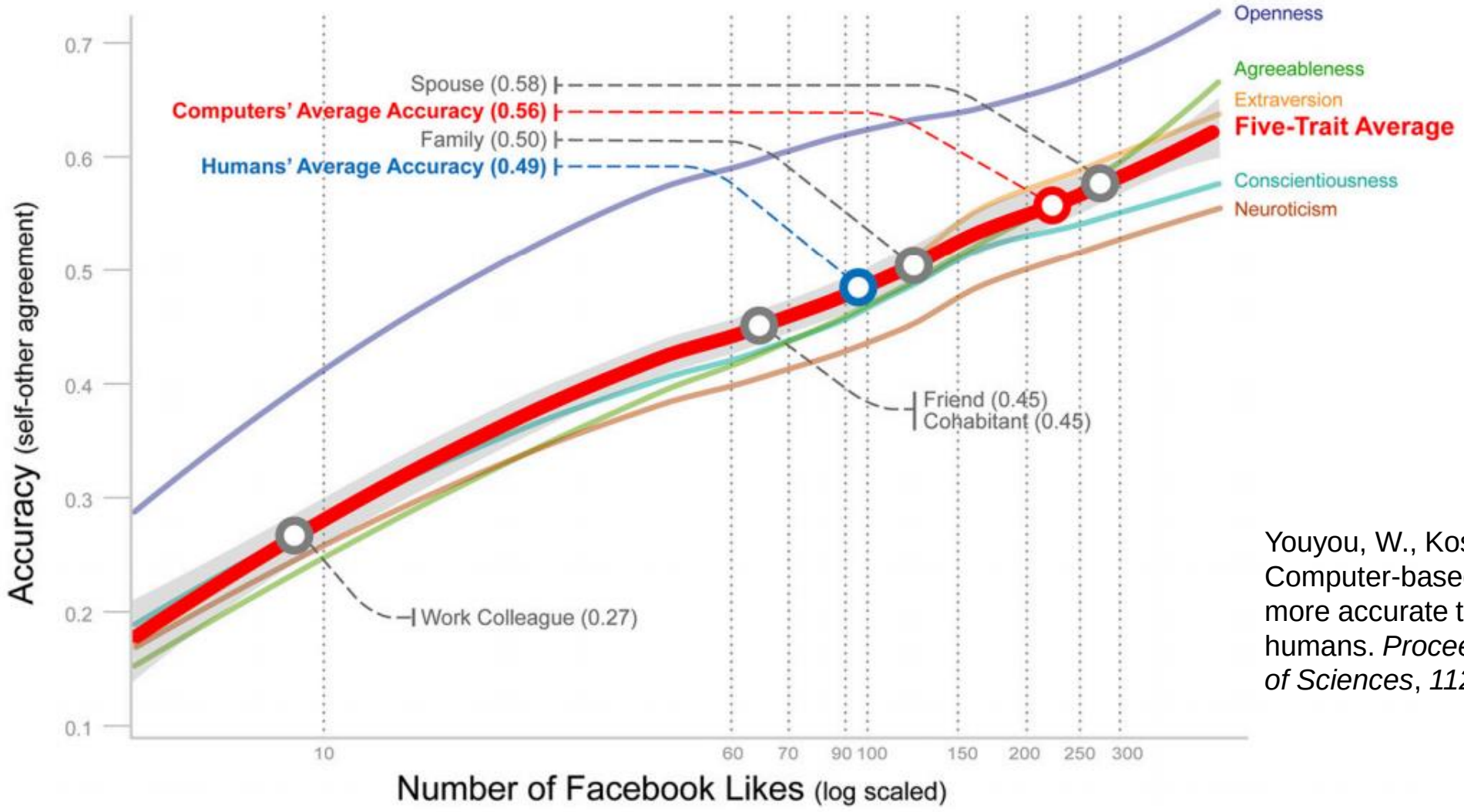


Note Openness!

Park, Schwartz, Eichstaedt et al., 2014, *JPSP*

Predicting Personality: Accuracy with Facebook likes (cross-validated)

Note Openness!



Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036-1040.

There are many algorithms…

- Some are regression-y with a regularizer
- But also different kinds of algorithms, like decision trees

Fantastic Intro to decision trees:

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Summary: Machine Learning

- Machine Learning is useful when you have **as many variables** (or more) **as observations.** That happens a lot with language.

- Basic machine learning is a **straightforward extension of regression**. Models get much fancier, but it rarely makes more than 20% difference in accuracy.

- "**Data is King**:" having more data (say twice as much) will make a much bigger difference for your prediction accuracies than having fancier prediction algorithms (say, deep learning).

Data is Always King

Summary-Summary

If you liked the content of this lecture…

… consider joining my course in the Fall of 2020
("Modern Language Analysis in the Social Sciences")

Preview of Fall 2020 Course

- Learn a modern python-based environment to do all of this directly from and to SQL databases (see DLATK.wwbp.org)
- No Python or SQL knowledge required!
- Do everything we talked about in this introduction
- Master both bottom up and top down language analyses, understand which method is appropriate when
- Learn how to write up for publication for your projects
- Basic machine learning with language variables

Johannes `Eichstaedt@Stanford.edu`

Sources/References

[1] Eichstaedt, J. C., Kern, M. L., Tobolsky, V., Yaden, D. B., Schwartz, H. A., Park, G., Hagan, C. A., Smith, L. K., Buffone, A., Iwry, J., Ungar, L. H., & Seligman, M. E. P. (2020) From Hypothesis-Testing to Hypothesis-Generation with Text Analysis: A Review and Quantitative Comparison of Open and Closed-Vocabulary Approaches on a Large Facebook Dataset. Manuscript in revision. - **EMAIL ME**

[2] [**pdf**] Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining Insights From Social Media Language: Methodologies and Challenges. *Psychological Methods*.

[3: table] [pdf] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

[4] [pdf] Schwartz, H. A., & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. The ANNALS of the American Academy of Political and Social Science, 659, 78-94. bibtex

[5] Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., ... & Schwartz, H. A. (2019). Evaluating the predictability of medical conditions from social media posts. PloS one, 14(6), e0215476.

[6: manual] https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

[7, ACL] Almodaresi, F., Ungar, L., Kulkarni, V., Zakeri M., Giorgi, S. & Schwartz, H. A. (2017). On the Distribution of Lexical Features in Social Media. Annual Meeting of the Association for Computational Linguistics. Bibtex

[8, Mehl]; Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. Annual review of psychology, 54(1), 547-577.

[9, handbook] Mehl, M. R. (2006). Quantitative Text Analysis.

[**pdf**] Schwartz, H. A., **Eichstaedt, J. C.**, Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinki, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8, e73791.

Highlighted ones are particularly recommended.