Social Media Text Analysis

Stony Brook University
CSE545, Fall 2016

Basics of Natural Language Processing

- Tokenization
  - Sentence
  - Word

- Part of Speech Tagging

- Syntactic Parsing

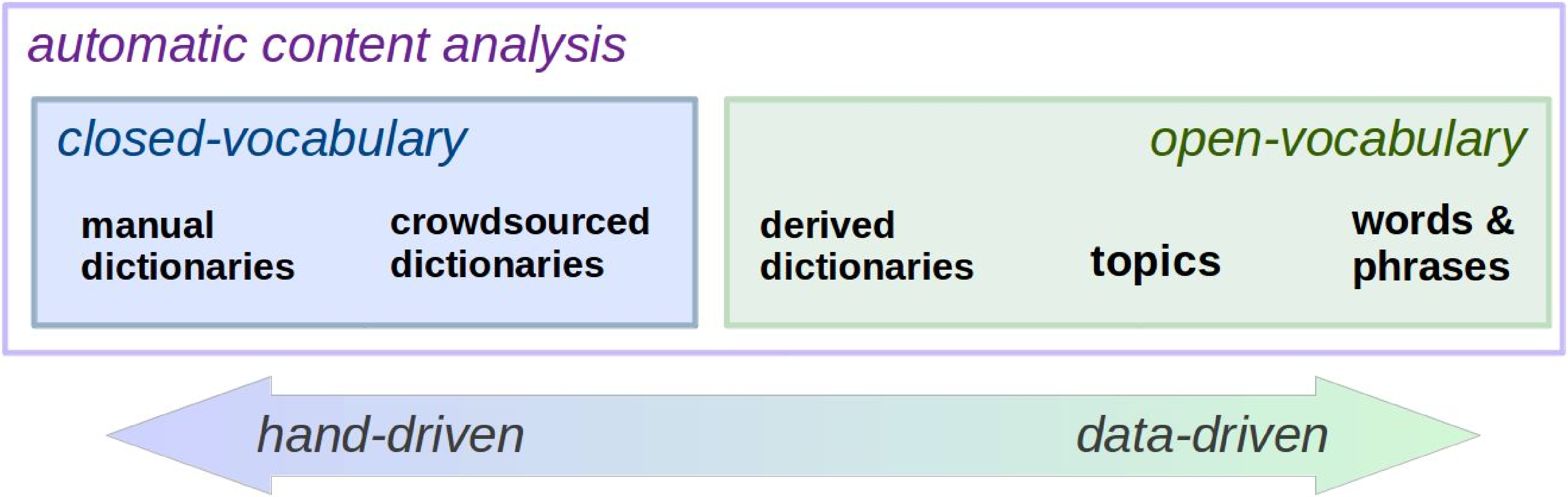
From language to features

Feature encodings

- Count
- Relative Frequency
- TF-IDF

- Dimensionally Reduced

Features: Closed-to-Open Vocabulary



automatic content analysis

closed-vocabulary

manual dictionaries

crowdsourced dictionaries

open-vocabulary

derived dictionaries

topics

words & phrases

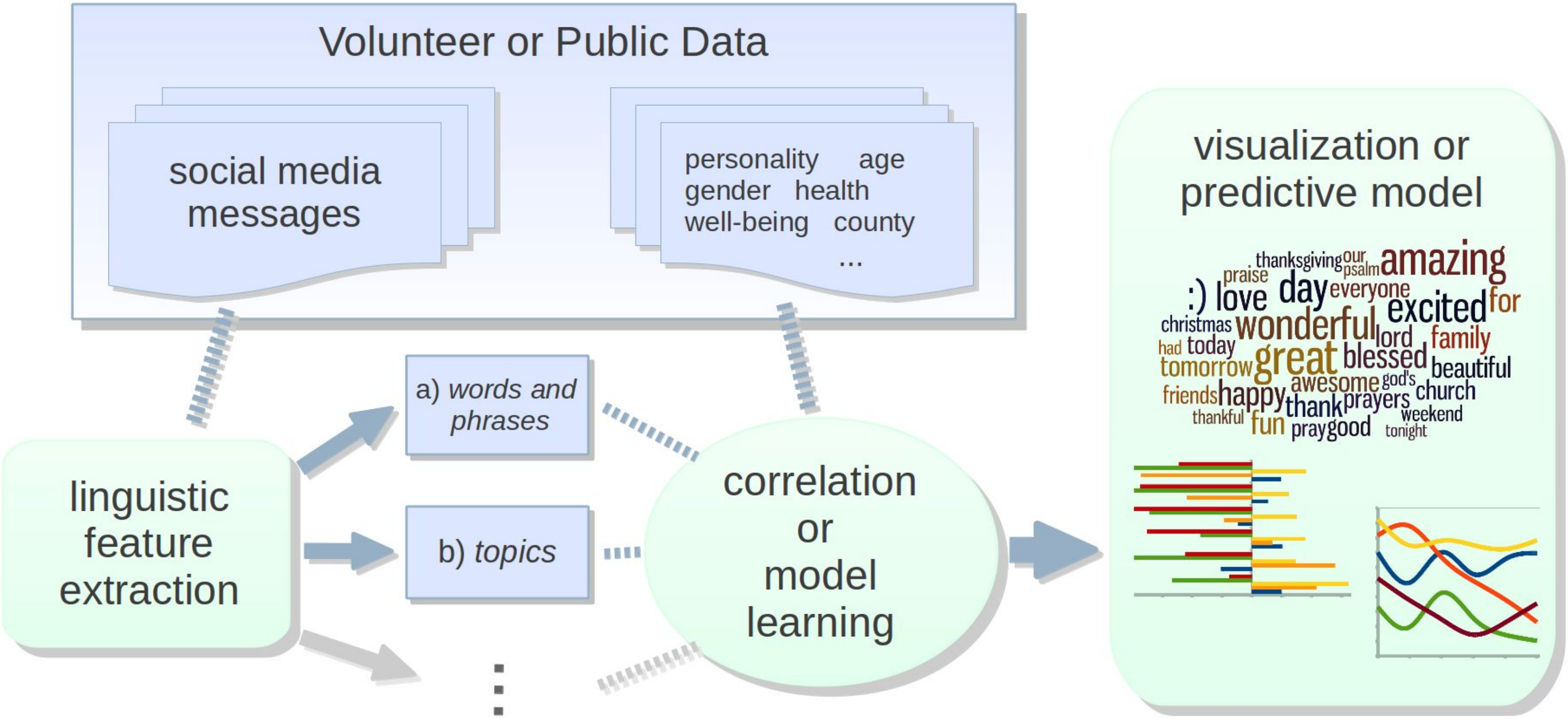
hand-driven

data-driven

Standard Tasks

- Insight
- Prediction

General "Insight" Framework



Volunteer or Public Data

social media messages

personality age gender health well-being county ...

linguistic feature extraction

a) *words and phrases*

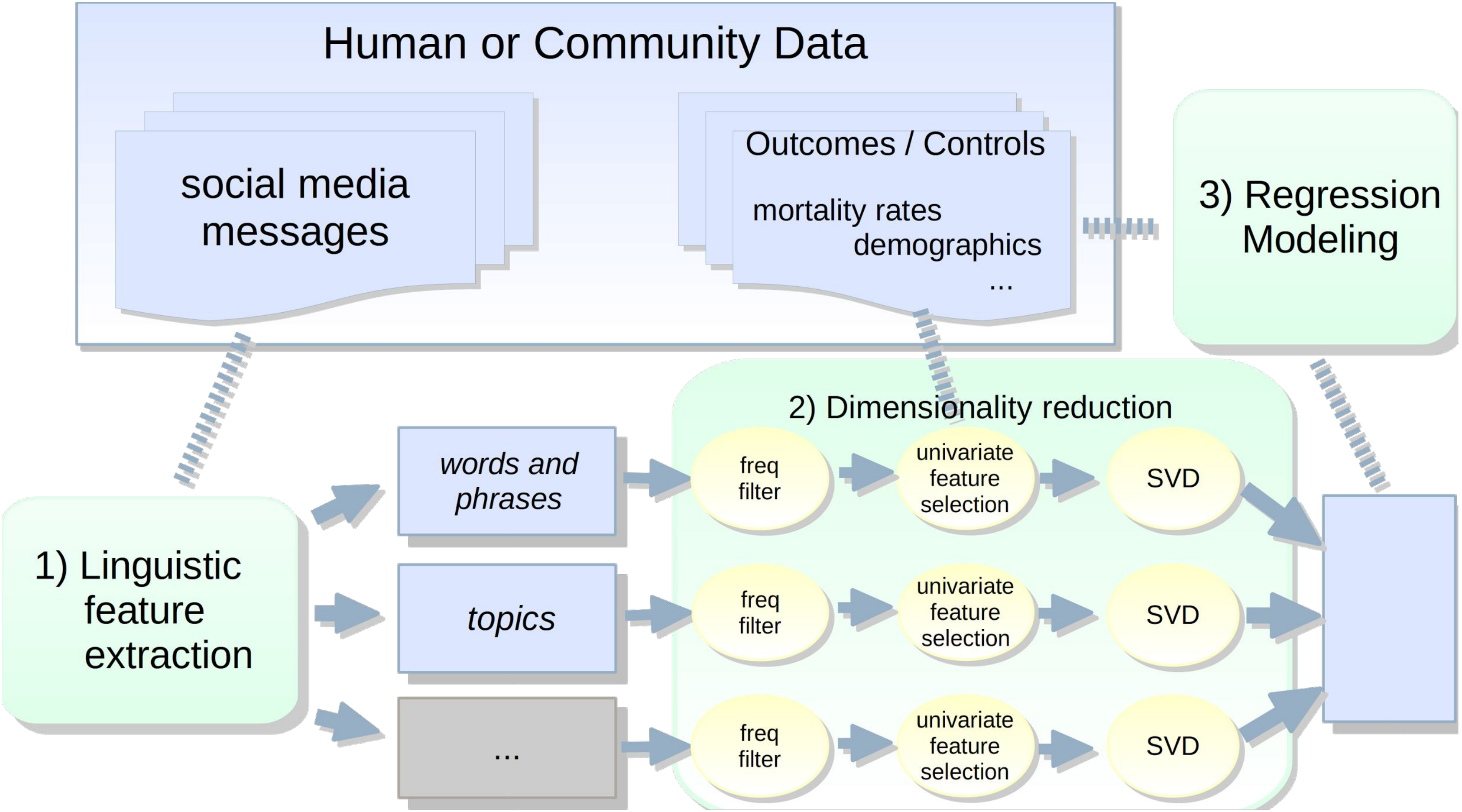
b) *topics*

...

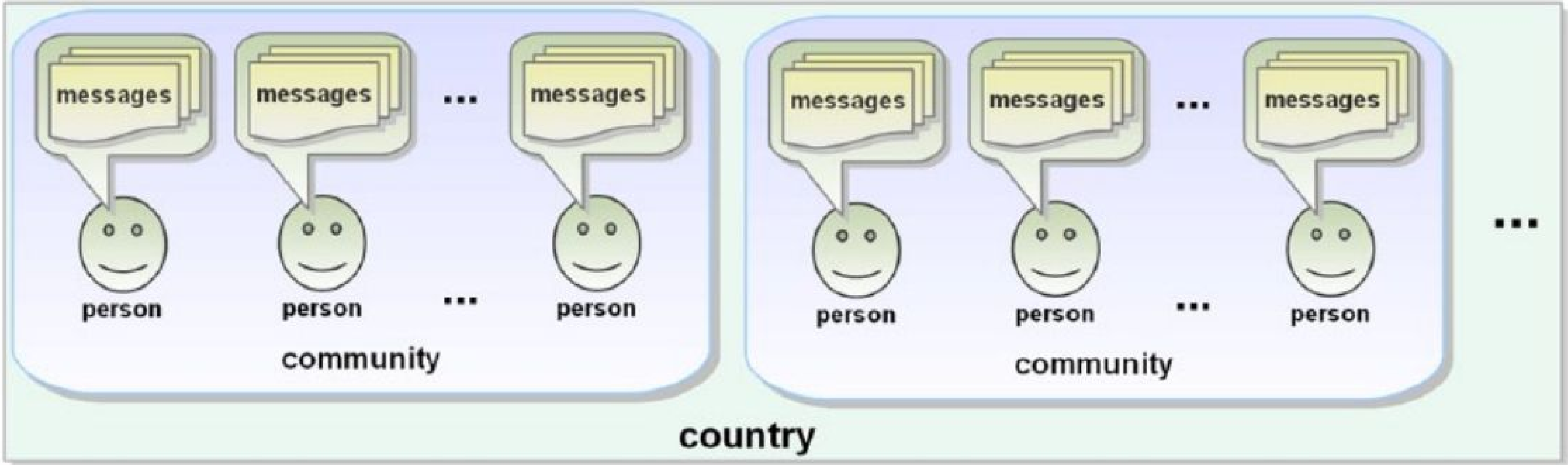
correlation or model learning

visualization or predictive model

Prediction Framework



Levels of Analysis



Example Tasks

1. Text-based Geolocation

2. Community Health Prediction
   (Handling many features, few observations)

3. Human Temporal Orientation
   (Sophisticated Features)

1. **Text-based Geolocation**

GOAL: Determine where a given user lives.

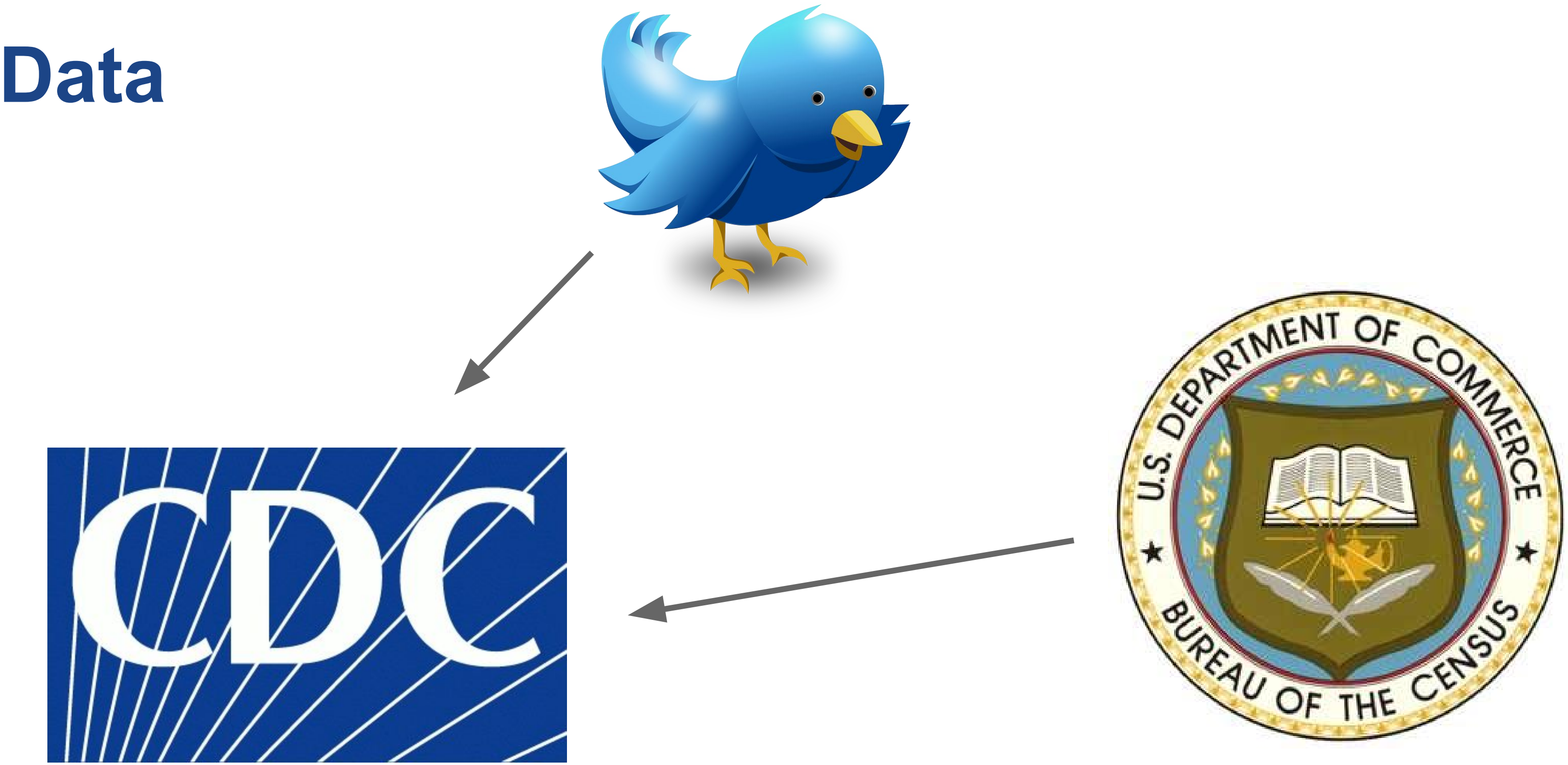
Versions

1. Based on posts (e.g. status updates, tweets)
2. Based on profile information

Gold-Standard: Geo-coordinates (lat+lon)

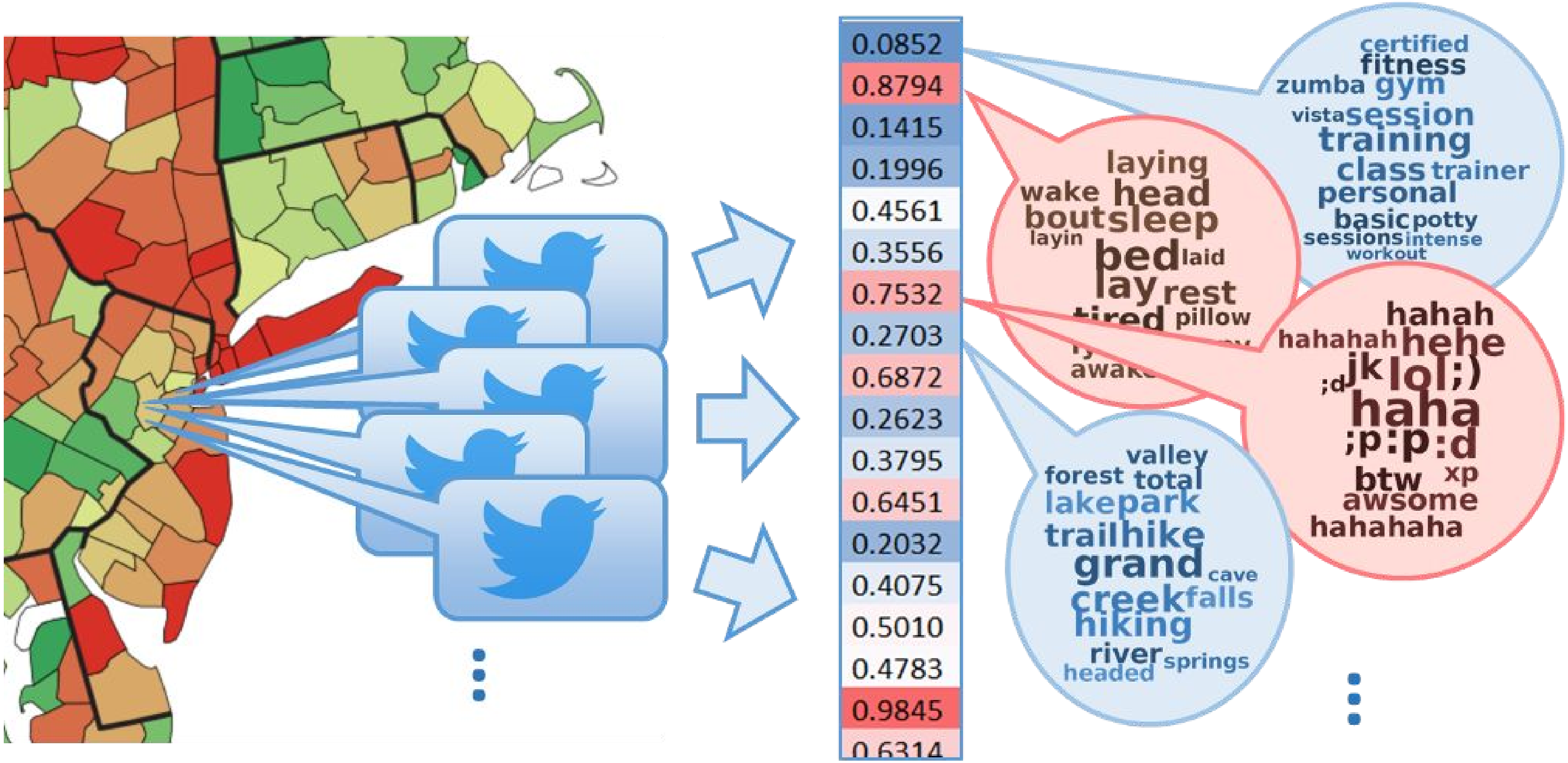
2. Community Health Prediction

Data

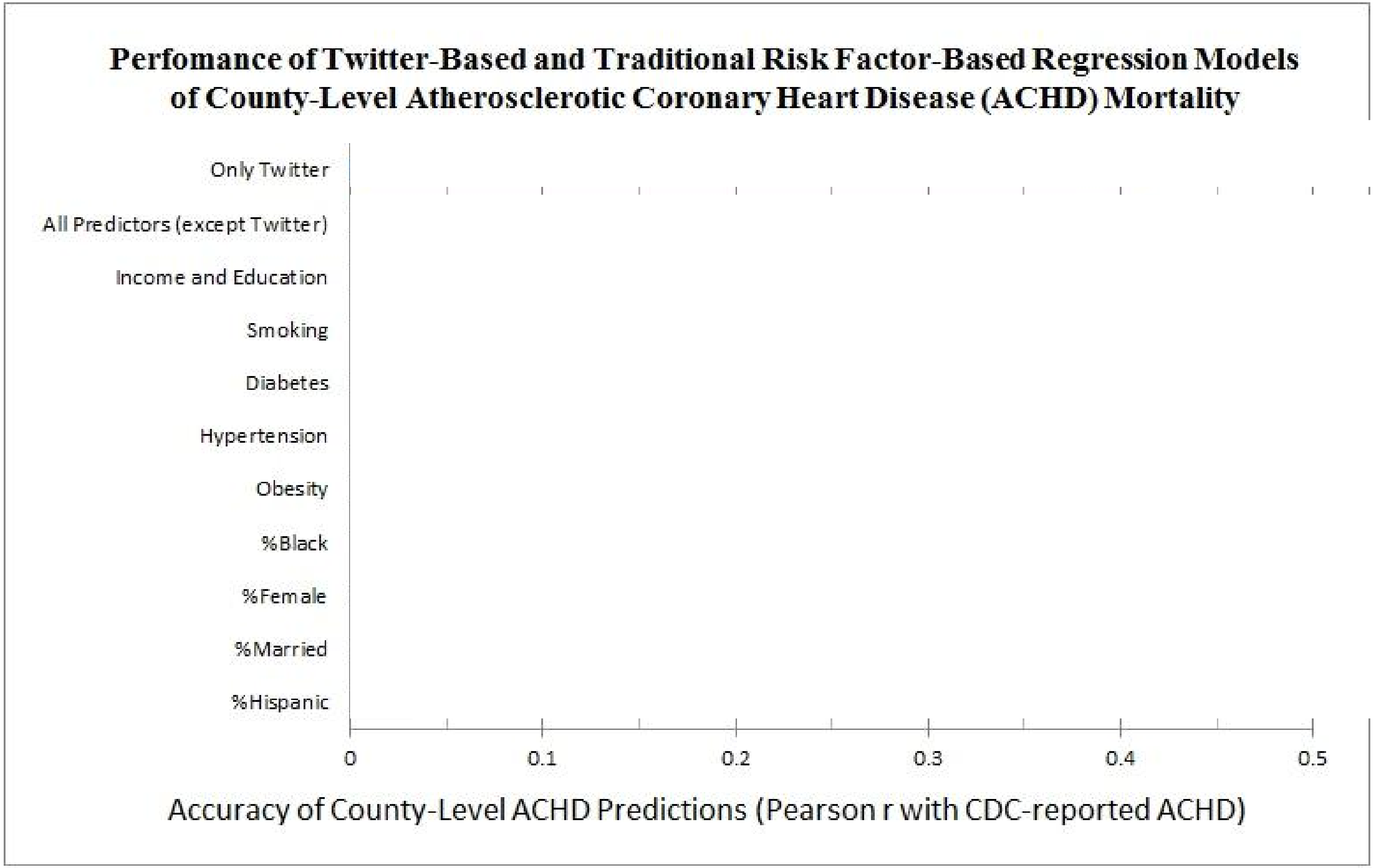


Atherosclerotic heart disease mortality

Encoding a community

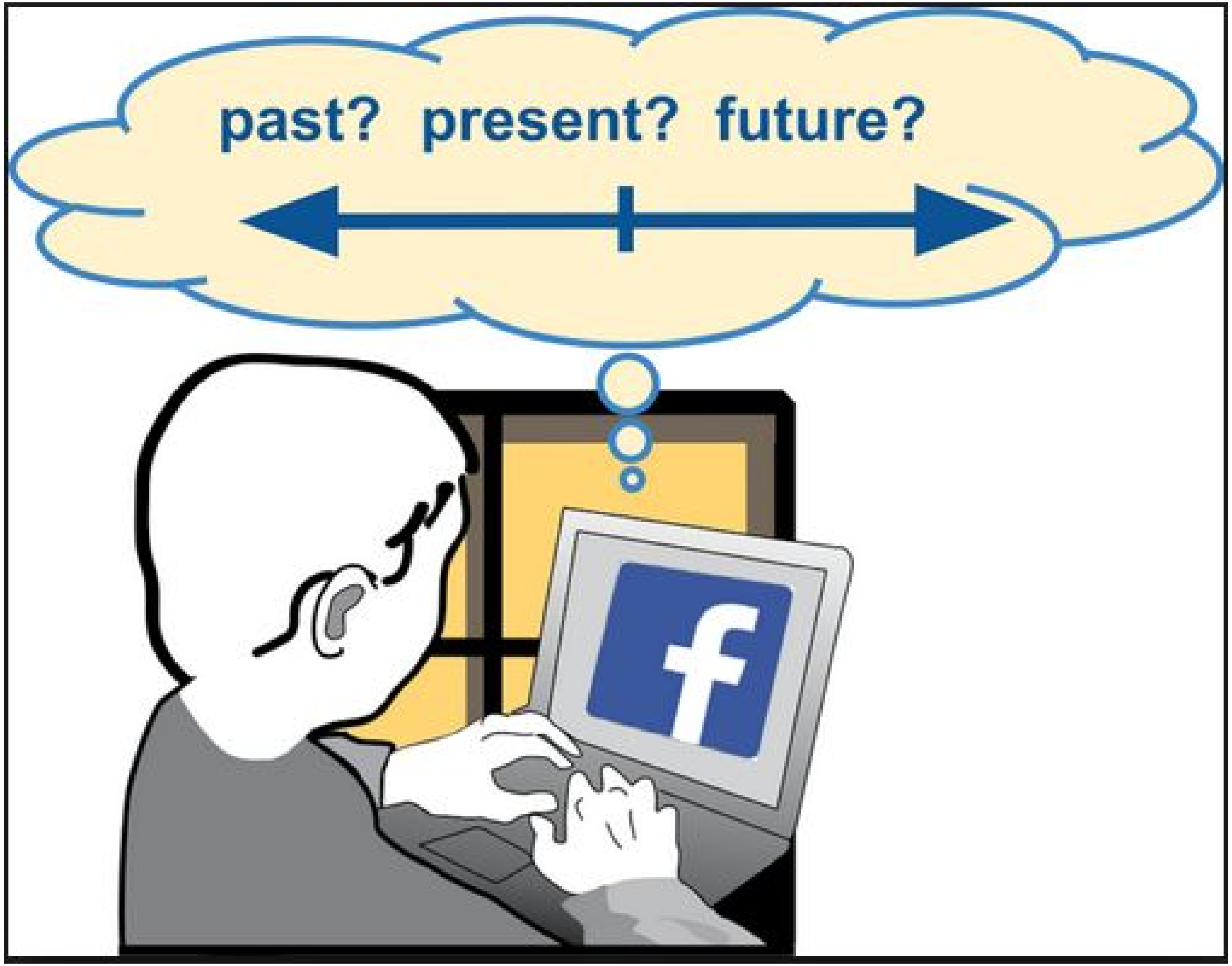

0.0852
0.8794
0.1415
0.1996
0.4561
0.3556
0.7532
0.2703
0.6872
0.2623
0.3795
0.6451
0.2032
0.4075
0.5010
0.4783
0.9845
0.6314
certified fitness zumba gym vista session training class trainer personal basic potty sessions intense workout
laying wake head bout sleep layin bed laid lay rest tired pillow awake
hahah hahahah hehe ;d jk lol ;) haha ;p :p :d btw xp awsome hahahaha
valley forest total lake park trail hike grand cave creek falls hiking river springs headed

Twitter Predicts Heart Disease



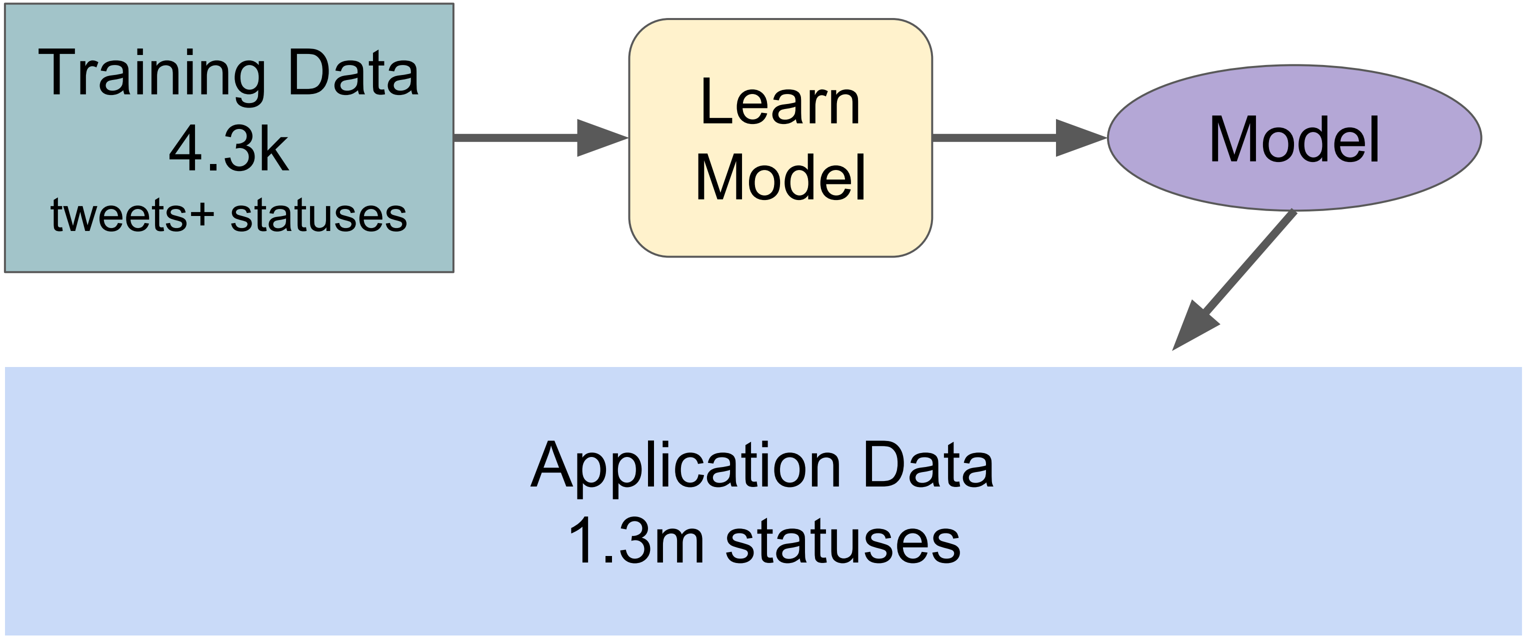
Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G.,..., Ungar, L. H., & Seligman, M. E. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science 26*(2), 159-169

3. Human Temporal Orientation


past? present? future?
f

Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future



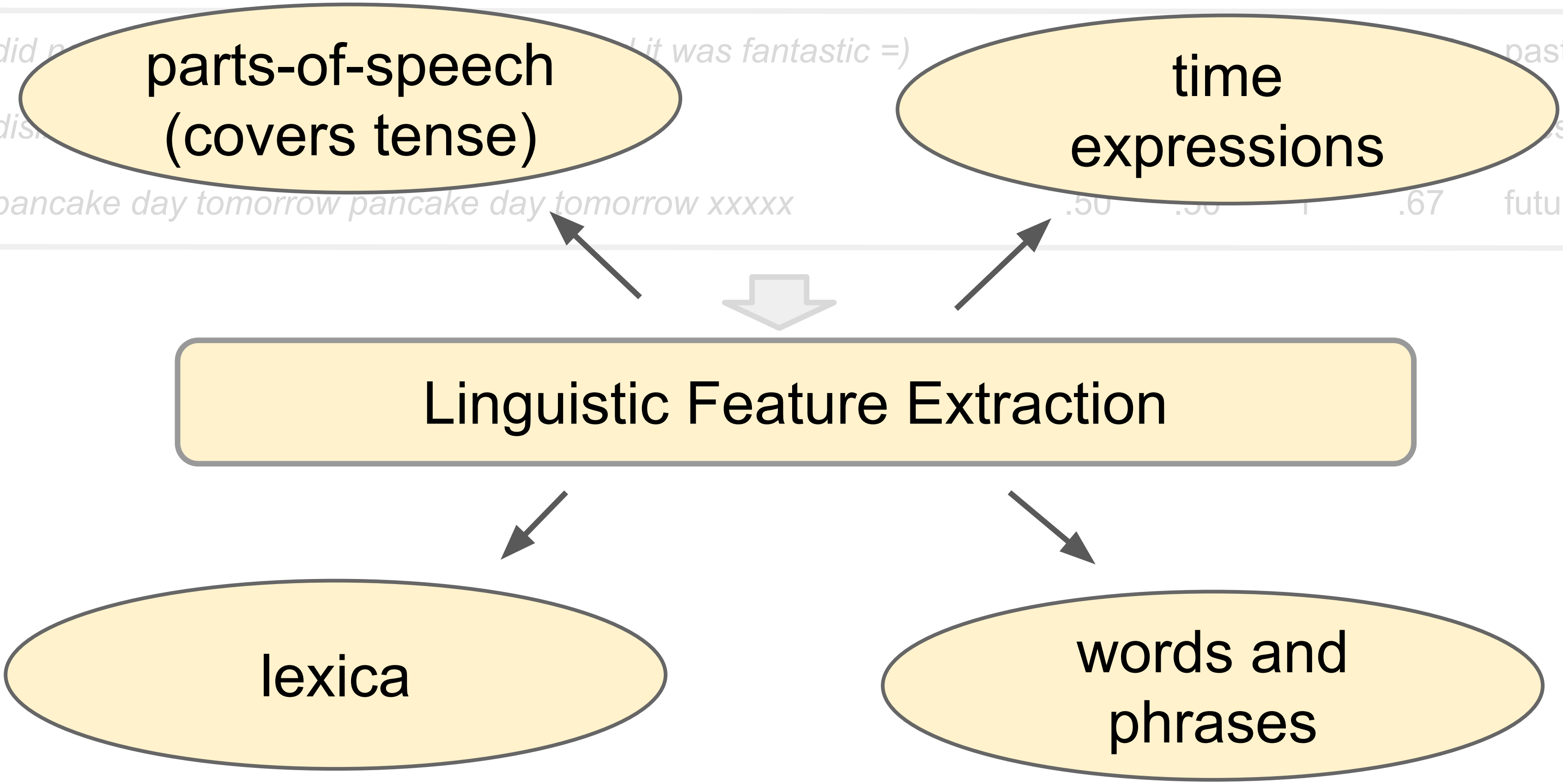
Building a model

message	**R1**	**R2**	**R3**	**m**	**class**
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future

Linguistic Feature Extraction

Building a model

message	R1	R2	R3	m	class
did n … it was fantastic =)					past
dis…					…sent
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future



parts-of-speech (covers tense)

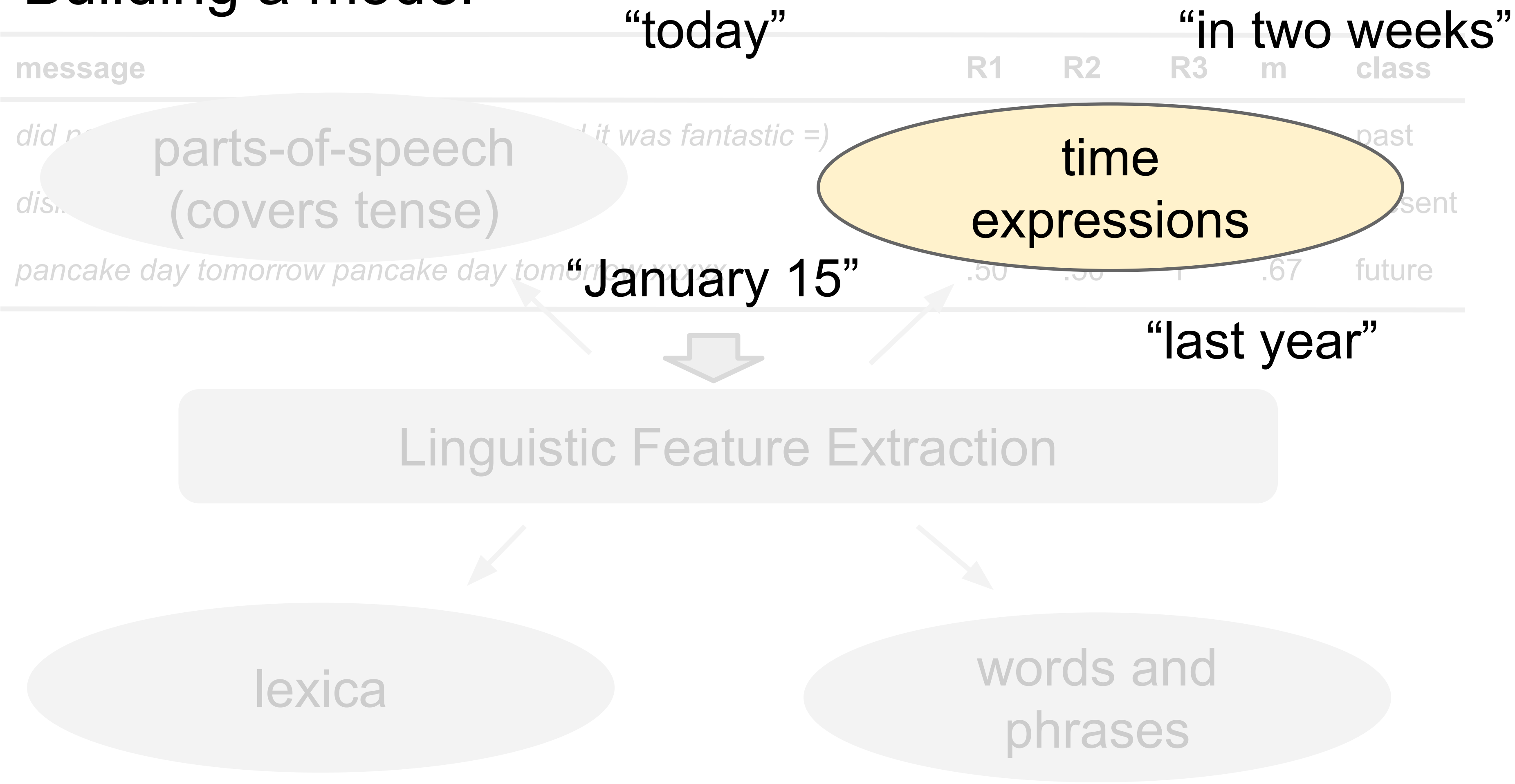
time expressions

Linguistic Feature Extraction

lexica

words and phrases

Building a model



“today”

“in two weeks”

message	R1	R2	R3	m	class
did n… it was fantastic =)					past
dis…					…sent
pancake day tomorrow pancake day tom…	.50	.50	1	.67	future

parts-of-speech (covers tense)

time expressions

“January 15”

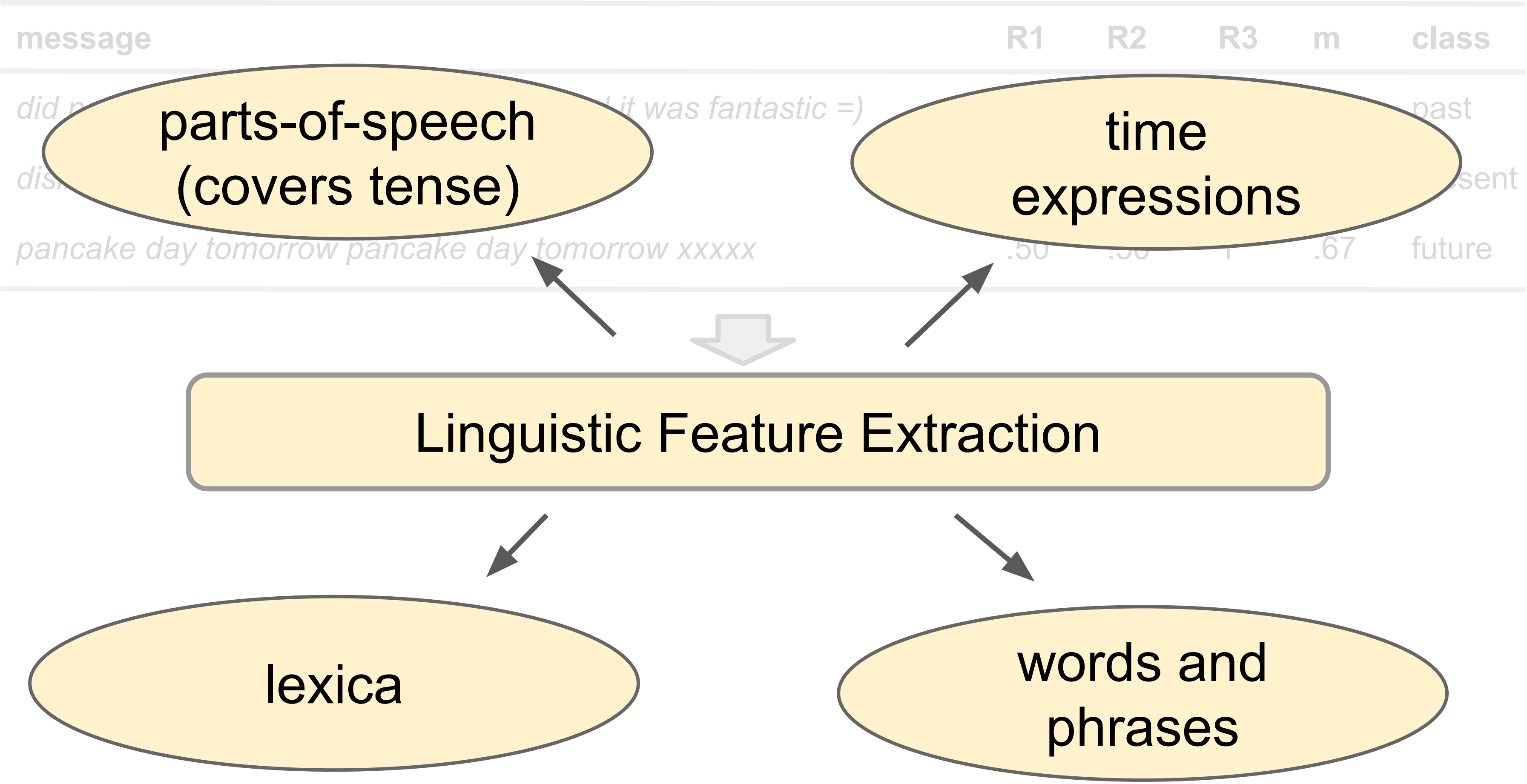
“last year”

Linguistic Feature Extraction

lexica

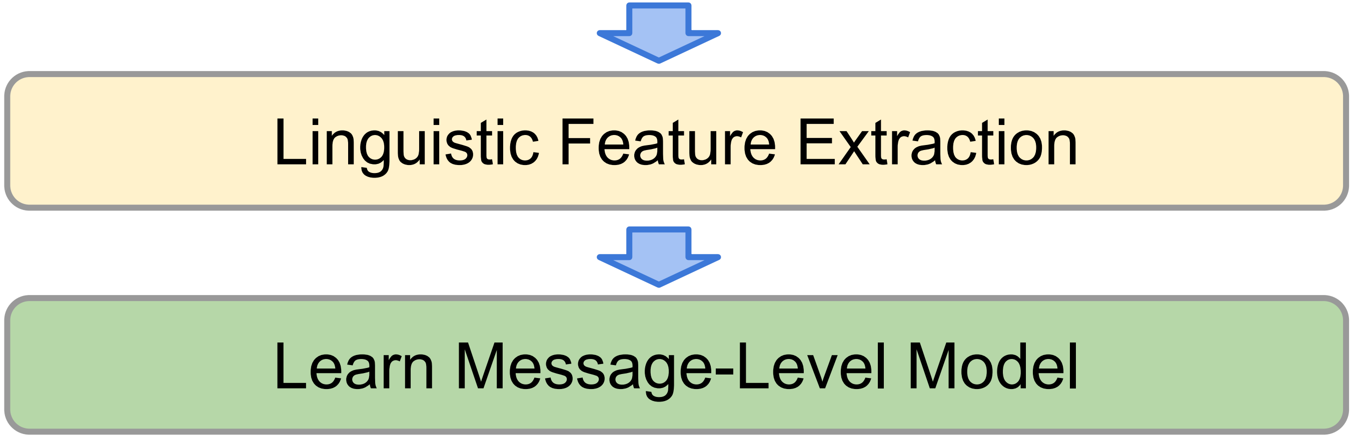
words and phrases

Building a model


message
R1
R2
R3
m
class
did n
it was fantastic =)
past
dis
sent
pancake day tomorrow pancake day tomorrow xxxxx
.50
1
.67
future
parts-of-speech (covers tense)
time expressions
Linguistic Feature Extraction
lexica
words and phrases

Building a model

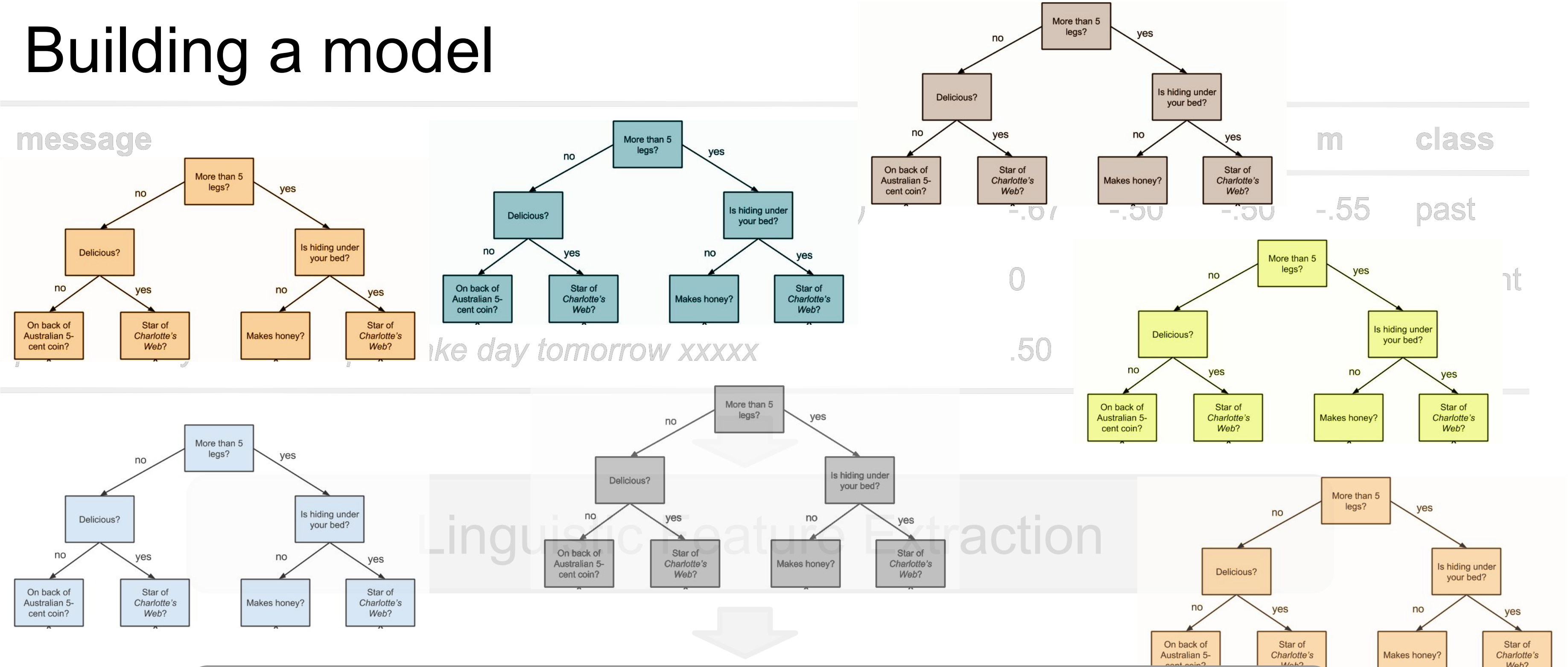
message	**R1**	**R2**	**R3**	**m**	**class**
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future



Linguistic Feature Extraction

Learn Message-Level Model

Building a model



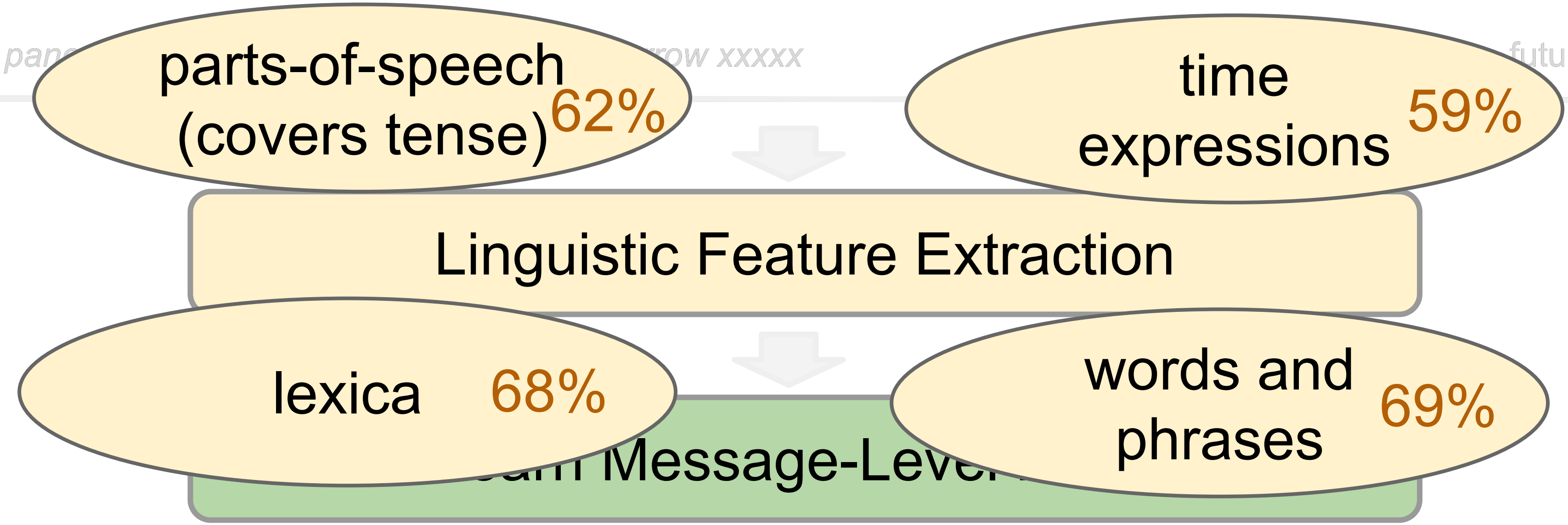
Learn Message-Level Model

Accuracy over a held-out set: 72%; baseline: 53%

Schwartz, H. A., Park, G., Sap, M., ..., & Ungar, L. (2015). Extracting Human Temporal Orientation from Facebook Language. *NAACL-2015: Conference of the North American Chapter of the Association for Computational Linguistics*

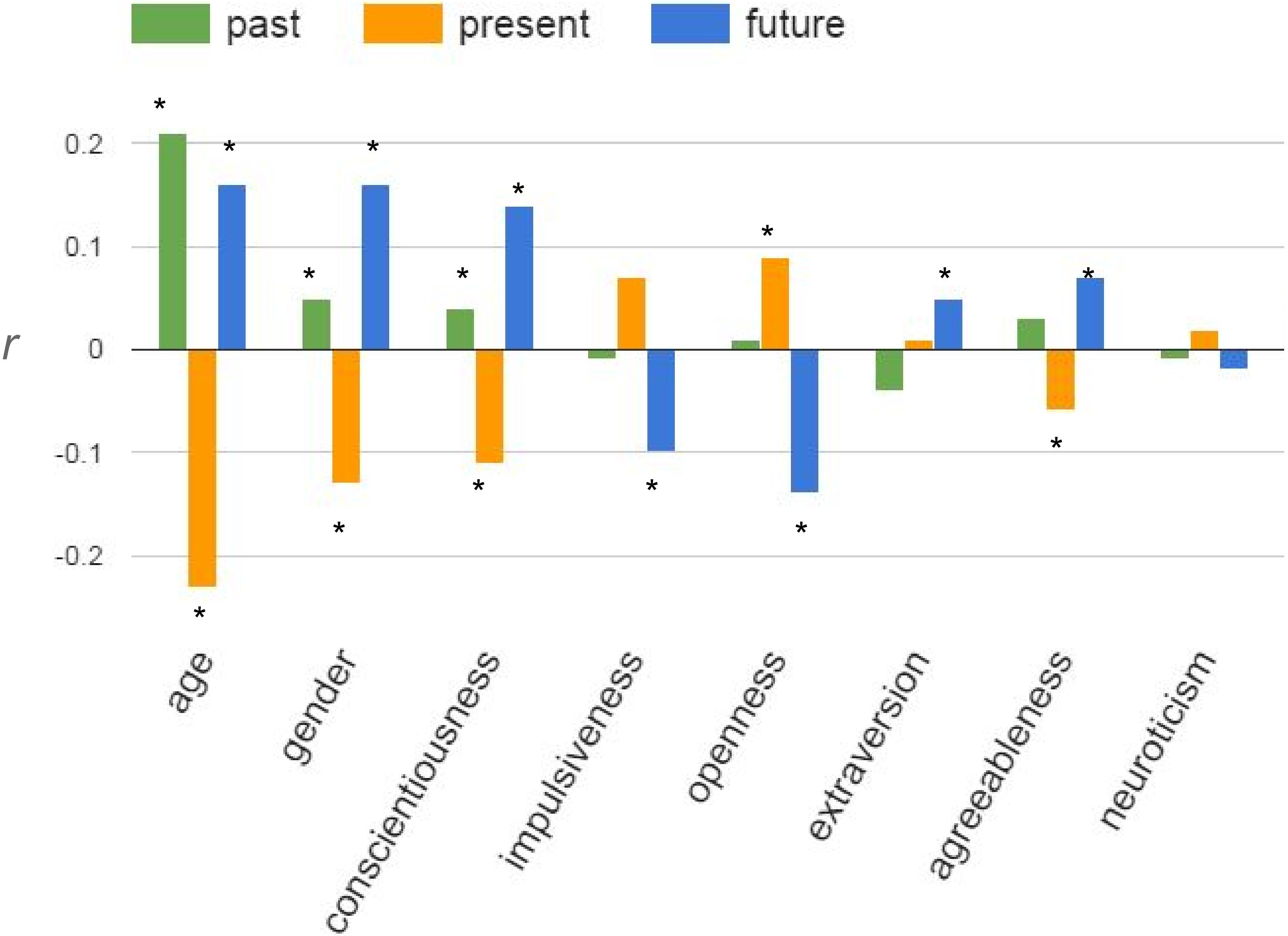
Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pan rrow xxxxx					future



Accuracy over a held-out set: 72%;   baseline: 53%

Schwartz, H. A., Park, G., Sap, M., ..., & Ungar, L. (2015). Extracting Human Temporal Orientation from Facebook Language. *NAACL-2015: Conference of the North American Chapter of the Association for Computational Linguistics*



Apply to Participant Messages