Age Prediction in Blogs

A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations

Sara Rosenthal and Kathleen McKeown

Columbia University

ACL 2011

Motivation

- Age prediction is useful for targeted advertising
- Age prediction can be used improve results in larger tasks such as identifying opinion, persuasion, and power

Research Questions

- What are the differences in how people communicate that most directly reveal their age?
- Which age groups should be chosen to yield the best results?
  - Under and over 18
  - 10s vs. 20s vs. 30s

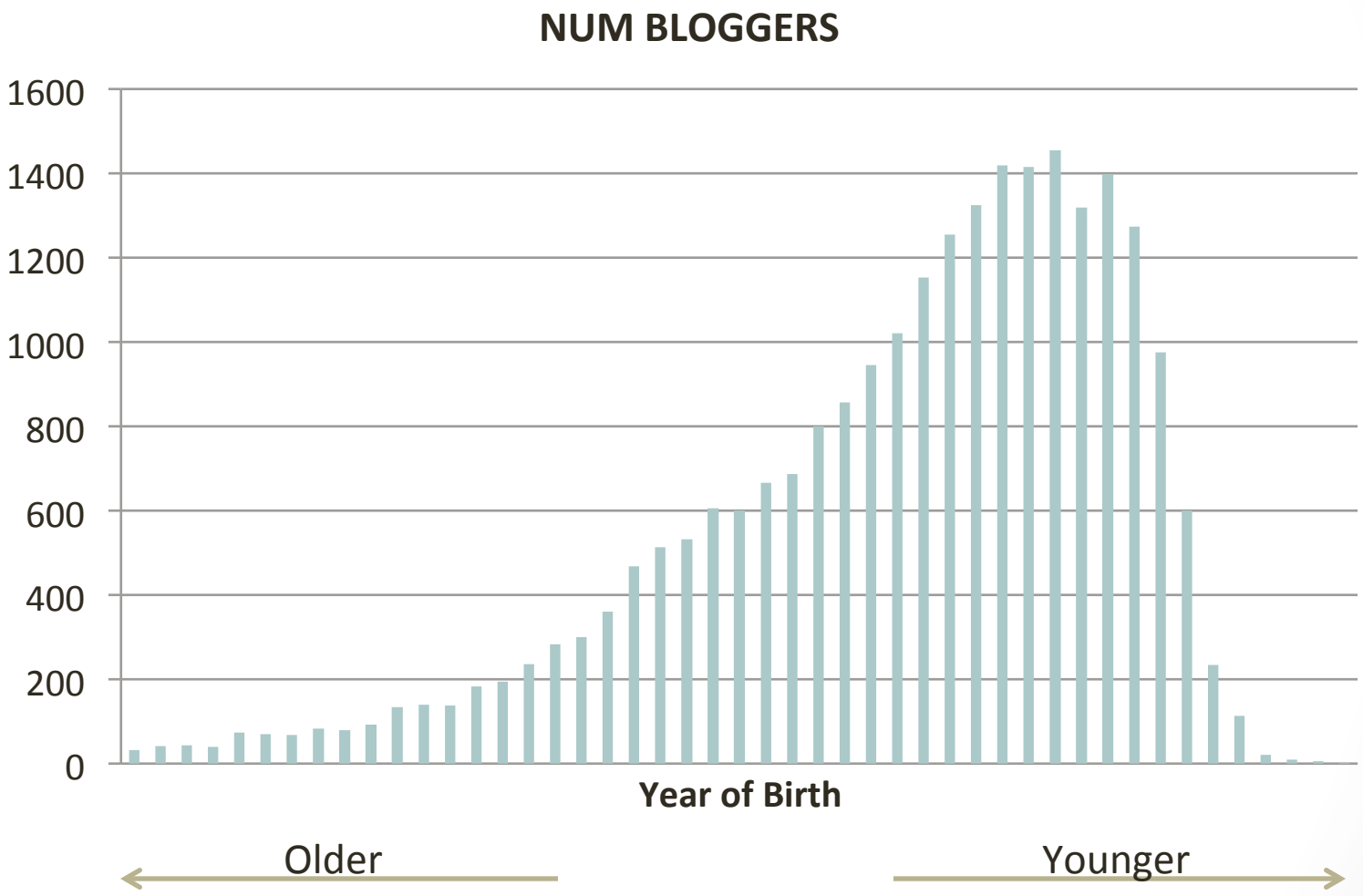
Approach

- Supervised Machine Learning
- Identify influential features
  - Online Behavior
    - e.g. Number of posts
  - Lexical-Stylistic
    - e.g. Sentence length
  - Lexical-Content
    - e.g. Bag of words
- Perform Classification
  - Using Logistic Regression

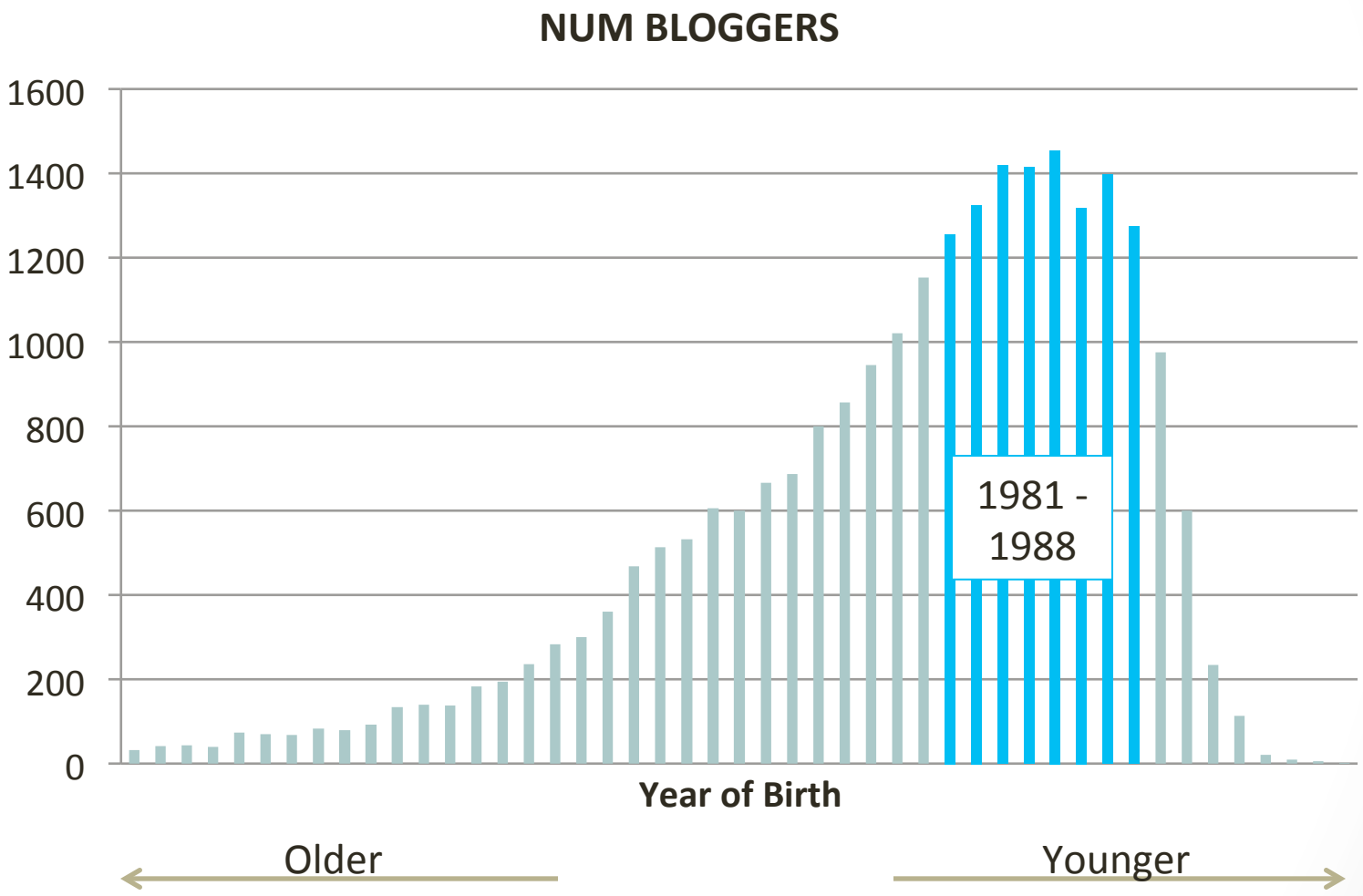
Corpus

- 24,500 LiveJournal Blogs
- Each blog is written by a person living in the US.
- Each blog includes the blogger's age in the profile
- Each blogger has written at least one entry in 2009 or later

Corpus


NUM BLOGGERS
1600
1400
1200
1000
800
600
400
200
0
Year of Birth
Older
Younger

Corpus


NUM BLOGGERS
1600
1400
1200
1000
800
600
400
200
0
1981 - 1988
Year of Birth
Older
Younger

Features: Online Behavior

Feature	Explanation	Example	Trend as Age Decreases
Interests	Top interests provided on their profile page (LiveJournal only)	disney	N/A
# of Friends	Number of friends the blogger has	45	fluctuates
# of Posts	Number of downloadable posts (0-25)	23	decrease
# of Lifetime Posts	Number of posts written in total	821	decrease
Time	The mode hour (00-23) and day the blogger posts	11, Monday	no change
Comments	Average number of comments per post	2.64	increase

Online Behavior

Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

18-22		28-32		38-42	
reading	3	reading	1	reading	1
		…			
drawing	10	love	24	sci-fi	21
fanfiction	11	drawing	25	love	34
love	15	sci-fi	37	polyamory	40
disney	39	tori amos	49	drawing	65
yaoi	40	hiking	55	sca	67
johnny depp	42	fanfiction	58	babylon 5	84
rent	44	women	61	leather	94

Online Behavior

Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

18-22		28-32		38-42	
~~**reading**~~	~~**3**~~	~~**reading**~~	~~**1**~~	~~**reading**~~	~~**1**~~
			…		
drawing	10	~~**love**~~	~~**24**~~	sci-fi	21
fanfiction	11	drawing	25	~~**love**~~	~~**34**~~
~~**love**~~	~~**15**~~	sci-fi	37	polyamory	40
disney	39	tori amos	49	drawing	65
yaoi	40	hiking	55	sca	67
johnny depp	42	fanfiction	58	babylon 5	84
rent	44	women	61	leather	94

Online Behavior

Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

18-22		28-32		38-42	
~~reading~~	~~3~~	~~reading~~	~~1~~	~~reading~~	~~1~~
			…		
drawing	**10**	~~love~~	~~24~~	**sci-fi**	**21**
fanfiction	**11**	**drawing**	**25**	~~love~~	~~34~~
~~love~~	~~15~~	**sci-fi**	**37**	polyamory	40
disney	39	tori amos	49	**drawing**	**65**
yaoi	40	hiking	55	sca	67
johnny depp	42	**fanfiction**	**58**	babylon 5	84
rent	44	women	61	leather	94

Online Behavior

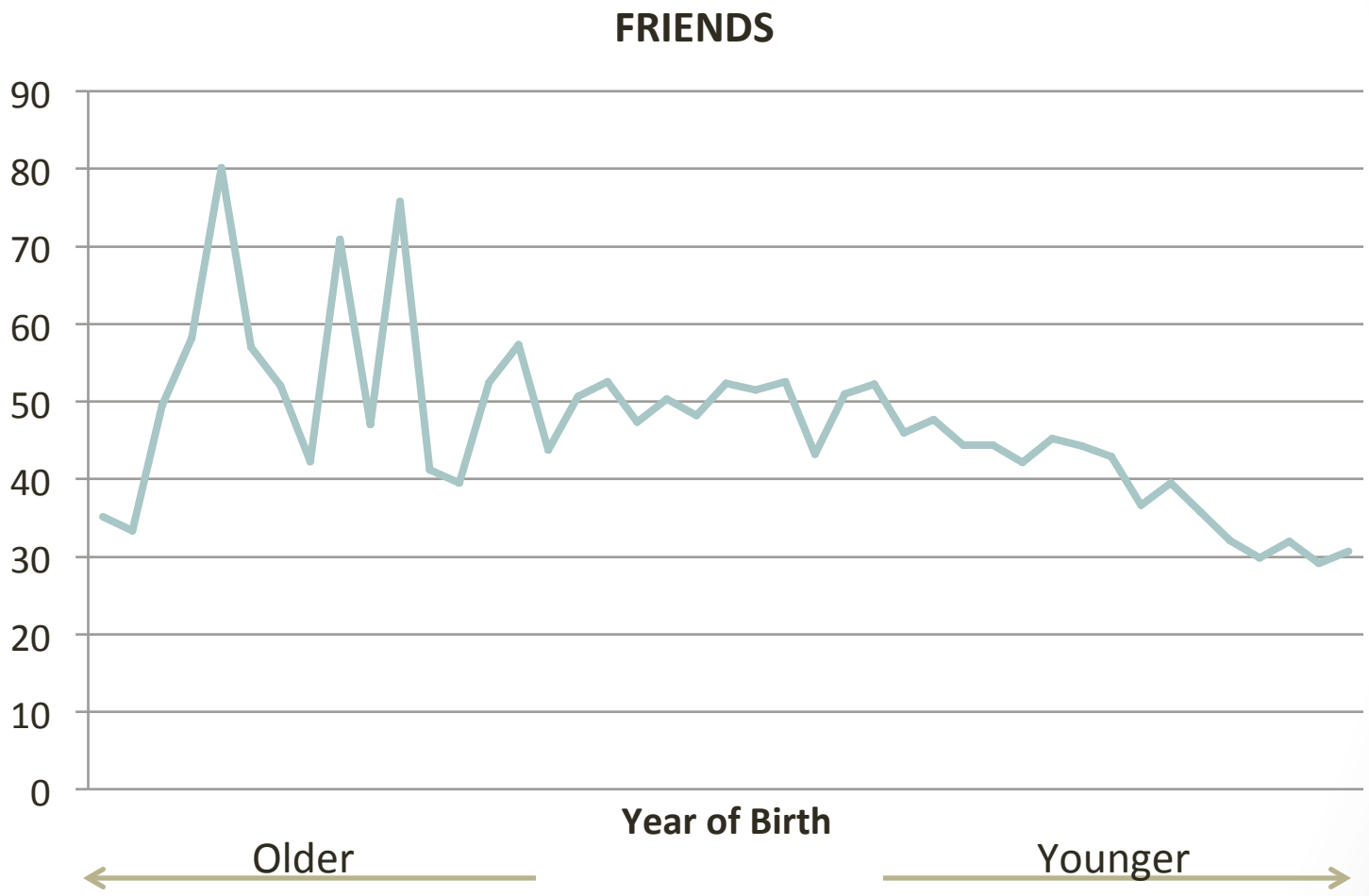
Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

18-22		28-32		38-42	
~~reading~~	~~3~~	~~reading~~	~~1~~	~~reading~~	~~1~~
			…		
drawing	10	~~love~~	~~24~~	sci-fi	21
fanfiction	11	drawing	25	~~love~~	~~34~~
~~love~~	~~15~~	sci-fi	37	polyamory	40
disney	**39**	tori amos	49	drawing	65
yaoi	40	**hiking**	**55**	sca	67
johnny depp	42	fanfiction	58	**babylon 5**	**84**
rent	44	women	61	leather	94

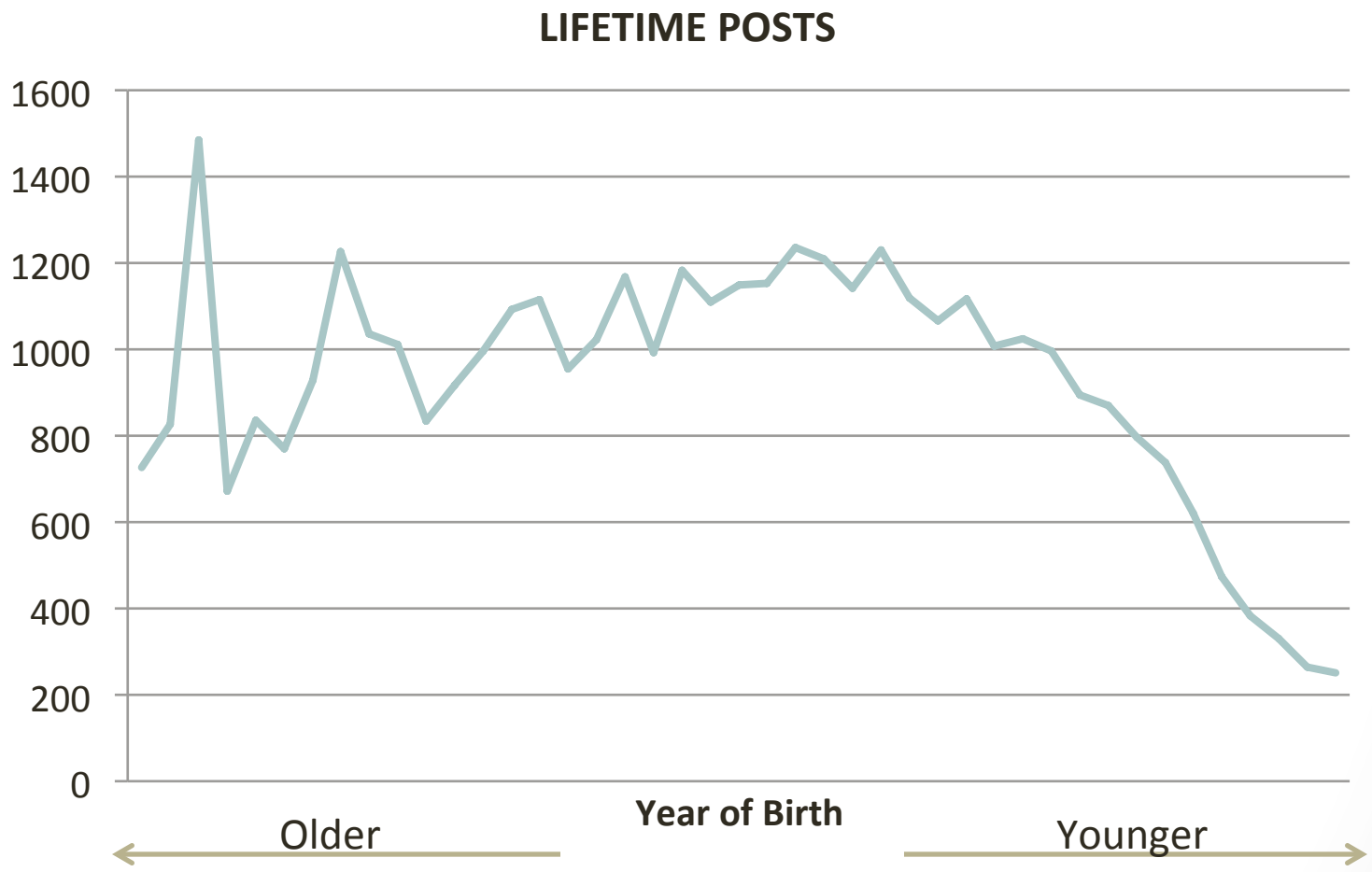
Online Behavior

Friends: *no clear trend*



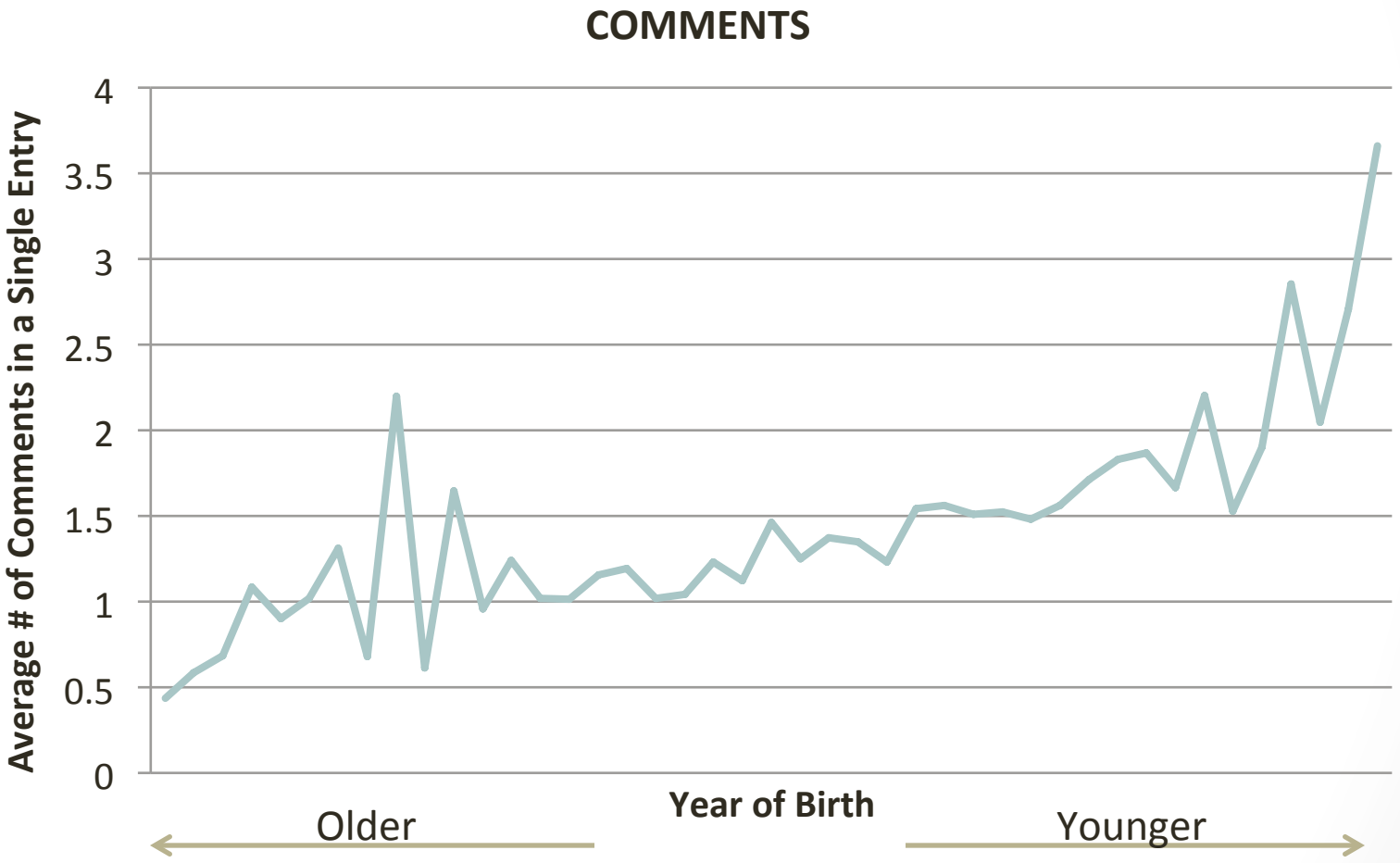
Online Behavior

Lifetime Posts: *older have more posts*



Online Behavior

Comments: *younger have more comments*



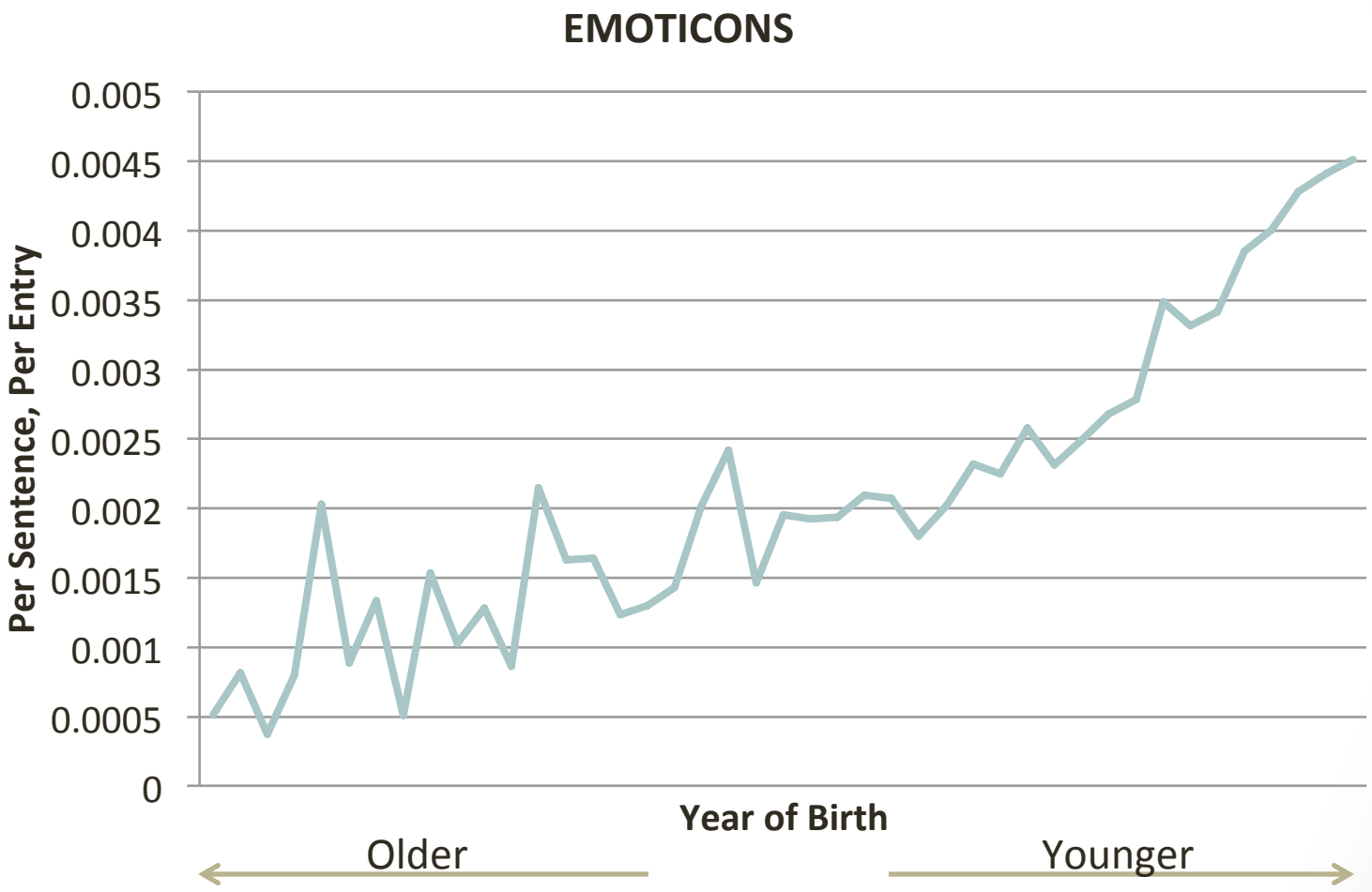
Features: Lexical - Stylistic

Feature	Explanation	Example	Trend as Age Decreases
Emoticons	Number of emoticons[1]	:)	increase
Acronyms	Number of internet acronyms[1]	lol	increase
Slang	Number of words that are not found in the dictionary[1]	wazzup	increase
Punctuation	Number of stand-alone punctuation[1]	!	increase
Capitalization	Number of words (with length of >1) that are all CAPS[1]	YOU	increase
Links/Images	Number of URL and image links[1]	www.site.com	fluctuates
Sentence Length	average sentence length	40	decrease

[1] Normalized per sentence per entry

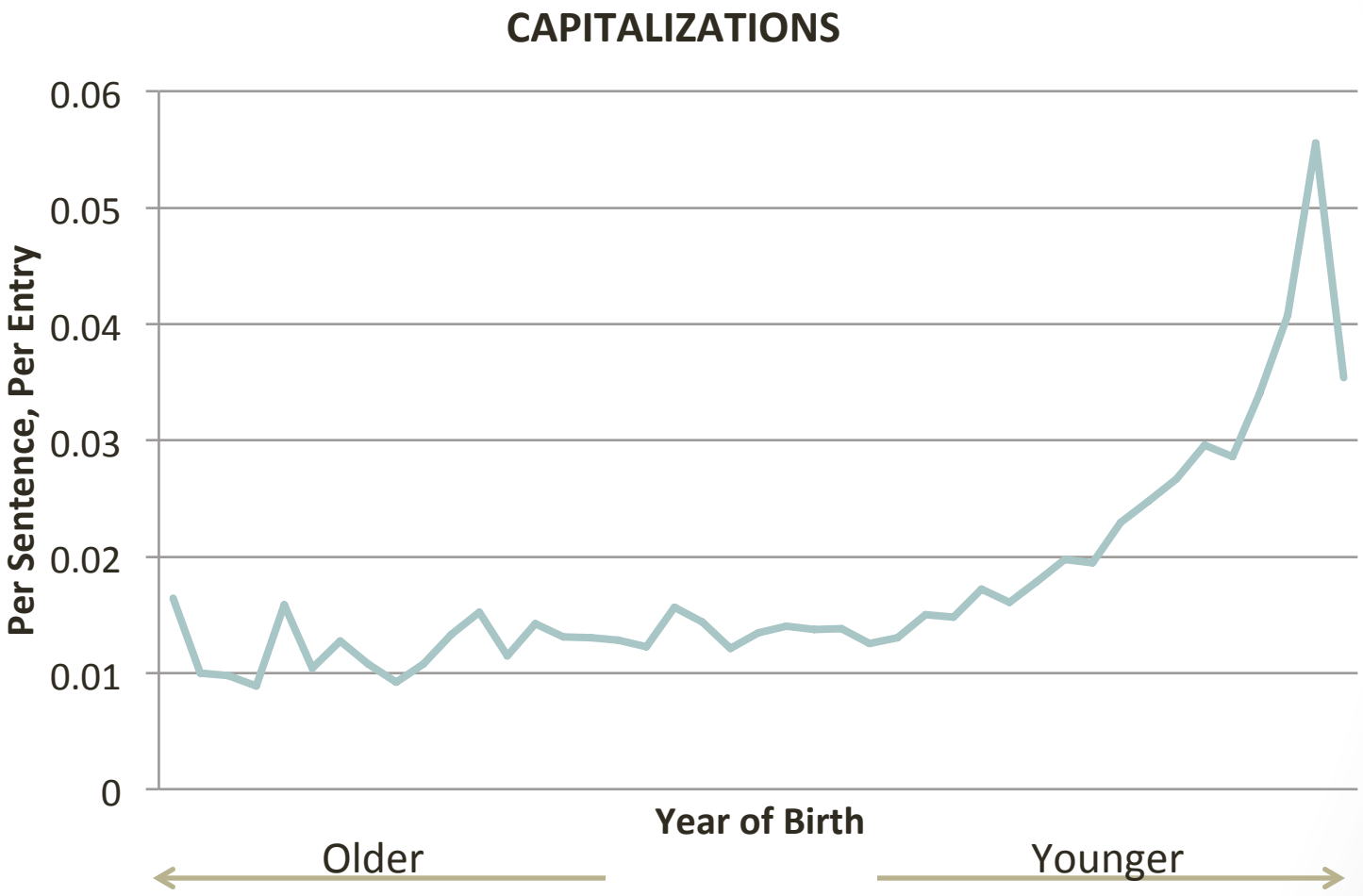
Lexical - Stylistic

Emoticons: *younger use more emoticons*



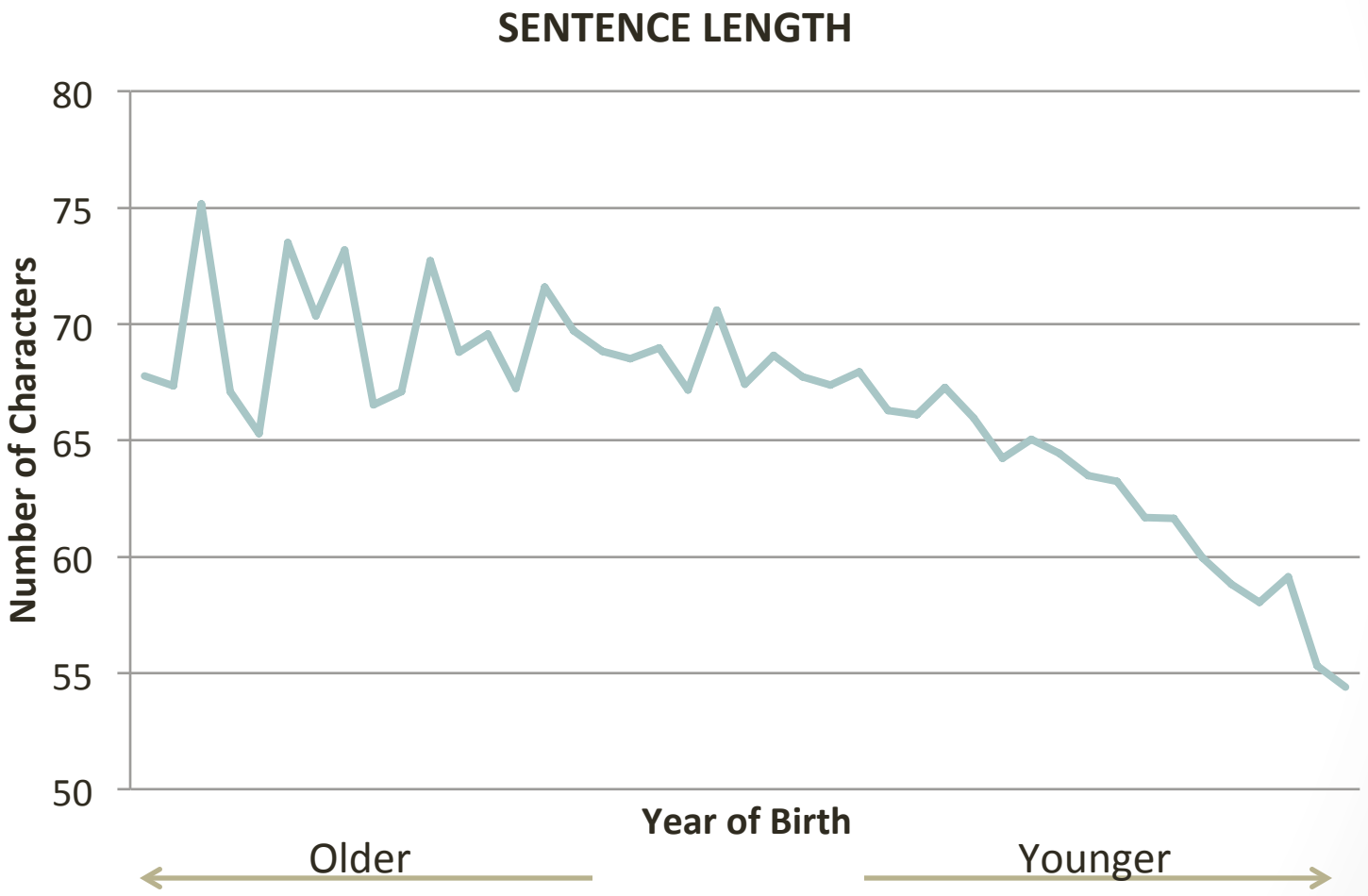
Lexical - Stylistic

Capitalizations: *younger use more capital words*



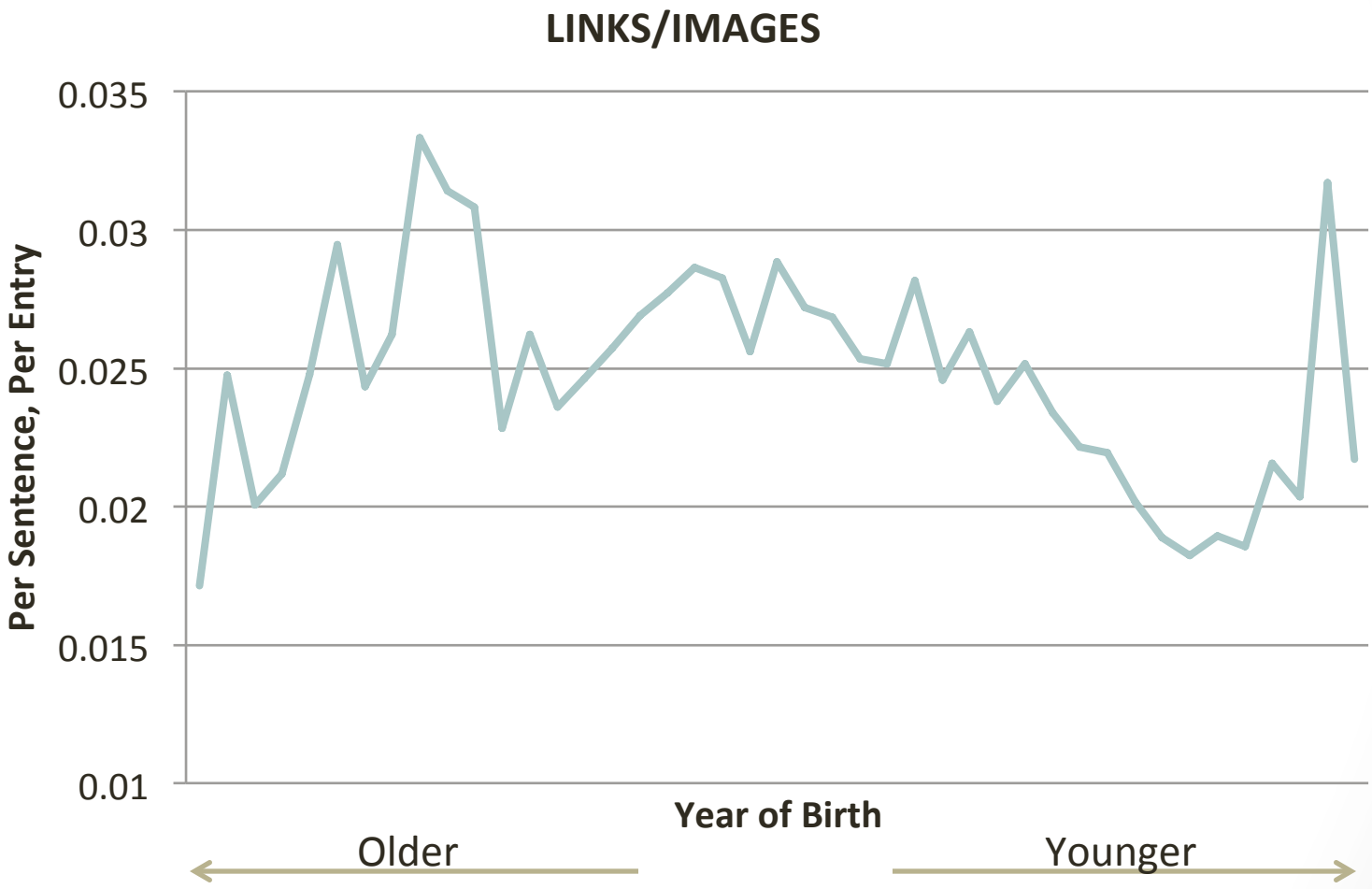
Lexical - Stylistic

Sentence Length: *older have longer sentences*



Lexical - Stylistic

Links/Images: *no clear trend*



Features: Lexical - Content

Feature	Explanation	Example
Collocations	Top collocations in the age group	in [] relationship
Syntax Collocations	Top syntax collocations in the age group[1]	have [] clue
POS Collocations	Top Part-of-Speech (POS) collocations in the age group	wouldn't VB
Words	Top words in the age group	his

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19:143– 177.

Words

- Extracted the top 200 words per age group using post frequency
- The value refers to the *position* of the word in its list.
- Keep discriminating words

	18-22		28-32		38-42	
1	ldqout (‘)	101	great	166	may	164
2	t	152	find	167	old	183
3	school	172	many	177	house	191
4	x	173	years	179	world	192
5	anything	175	week	181	please	198

Related Work

- Age and geographic inferences of the LiveJournal social network [Mackinnon and Warren 2006]
  - Use mean age of a bloggers social network
  - +/- 5 years: 98% accuracy
- An exploration of observable features related to blogger age [Burger and Henderson 2006]
  - Identify style and online behavior features
  - Under/Over 18: unsuccessful
- Effects of Age and Gender in Blogging [Schler et al 2006]
  - Use style and content features to predict whether a blogger is in their 10s, 20s, or 30s

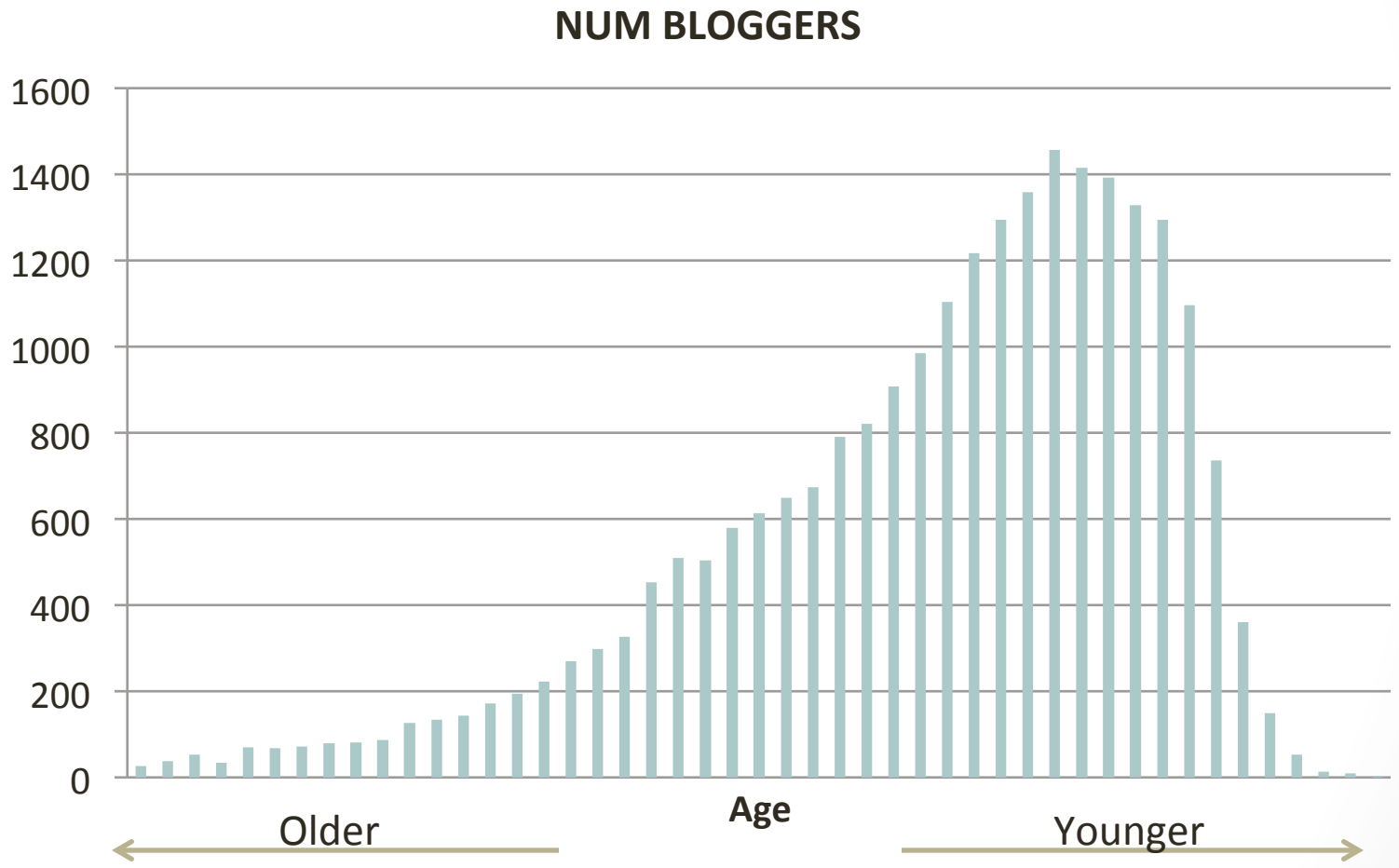
Ian Mackinnon and Robert Warren. 2006. Age and geographic inferences of the livejournal social network.
Burger, John D. and John C. Henderson. 2006. An exploration of observable features related to blogger age.
Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging.

Experiments

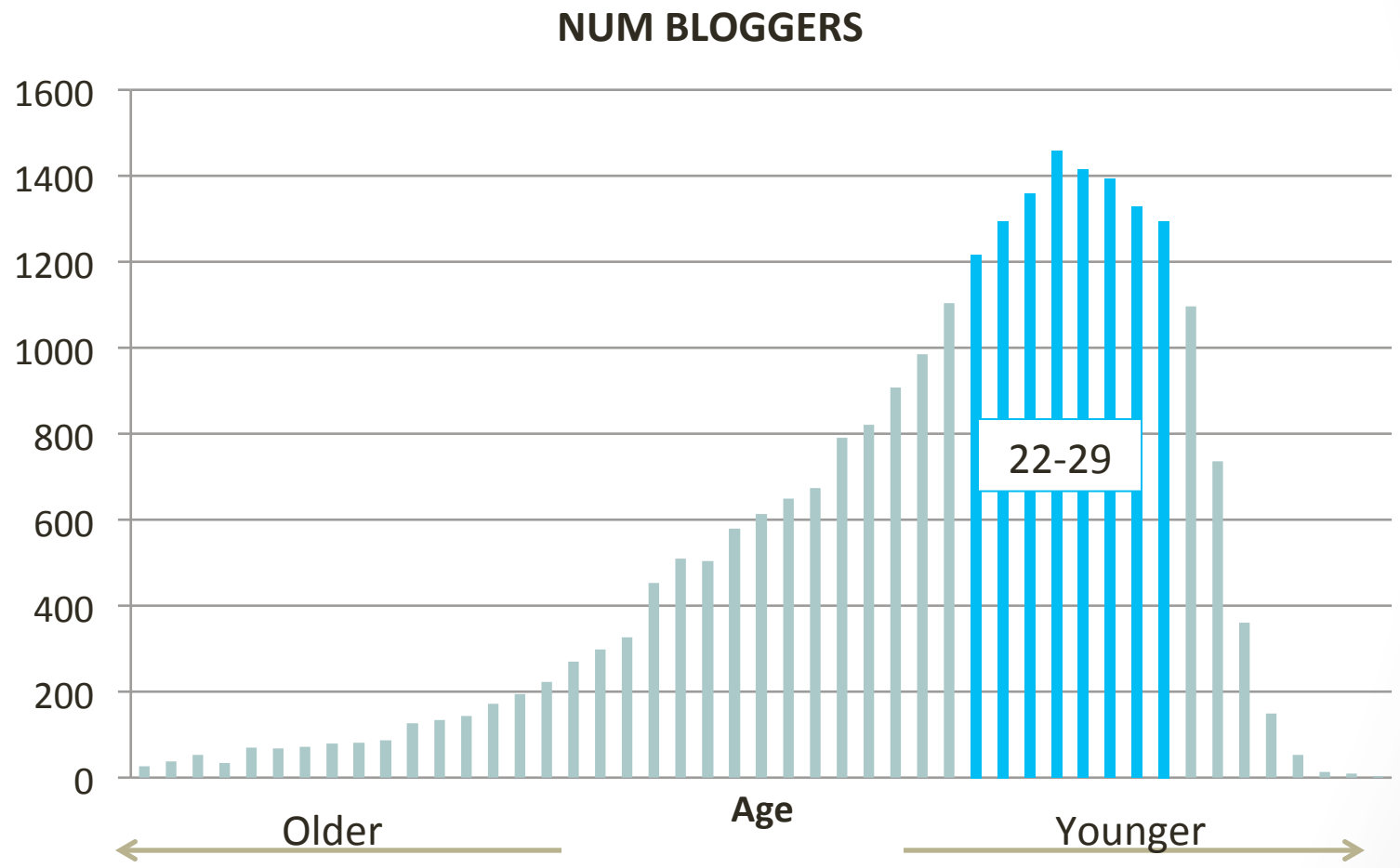
- Data: Portions of the LiveJournal corpus
- Classification: Logistic Regression and 10-fold cross-validation in Weka [Hall et al]
- Statistical significance: t-test

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update.

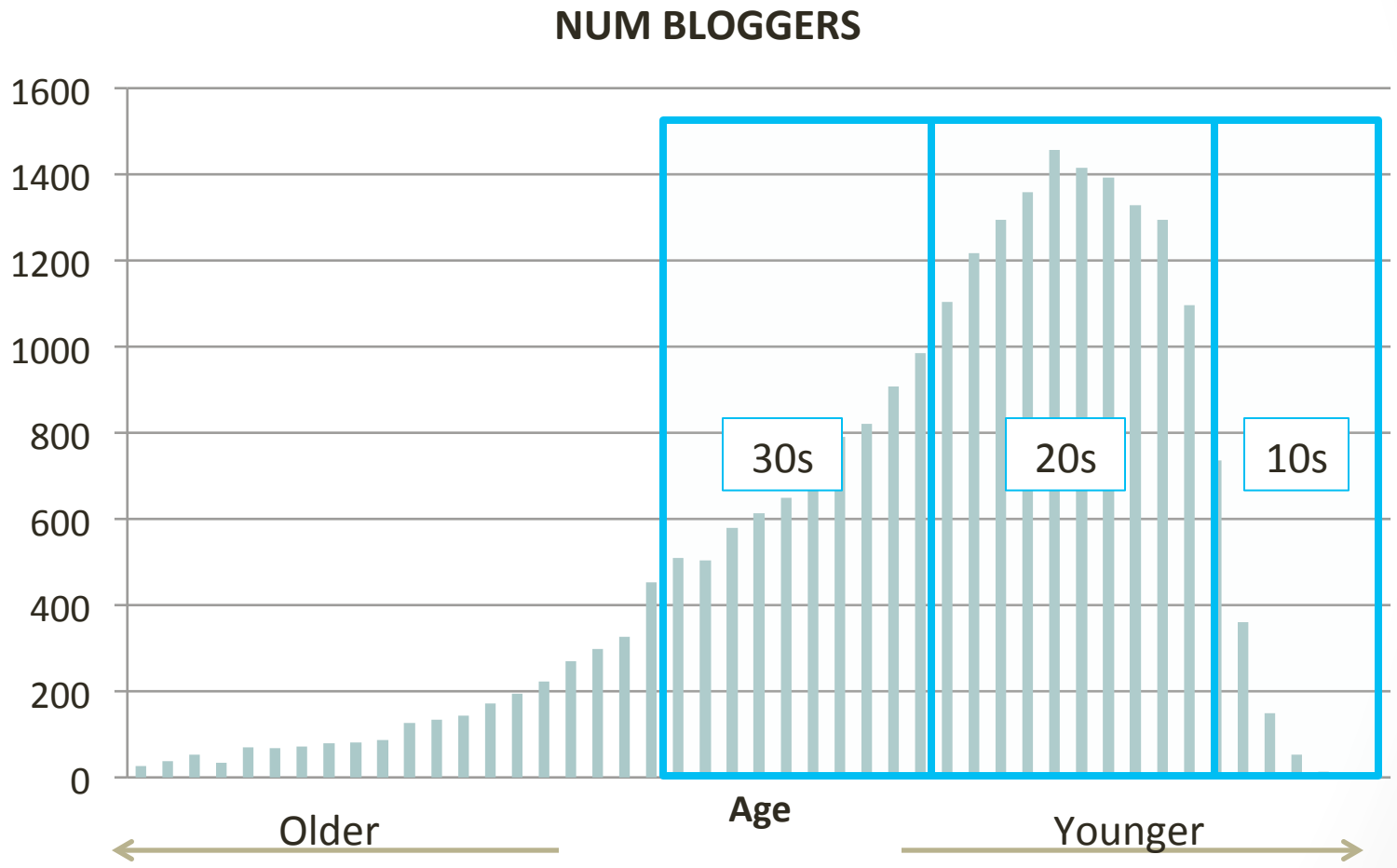
Experiment I
Age Groups



Experiment I
Age Groups

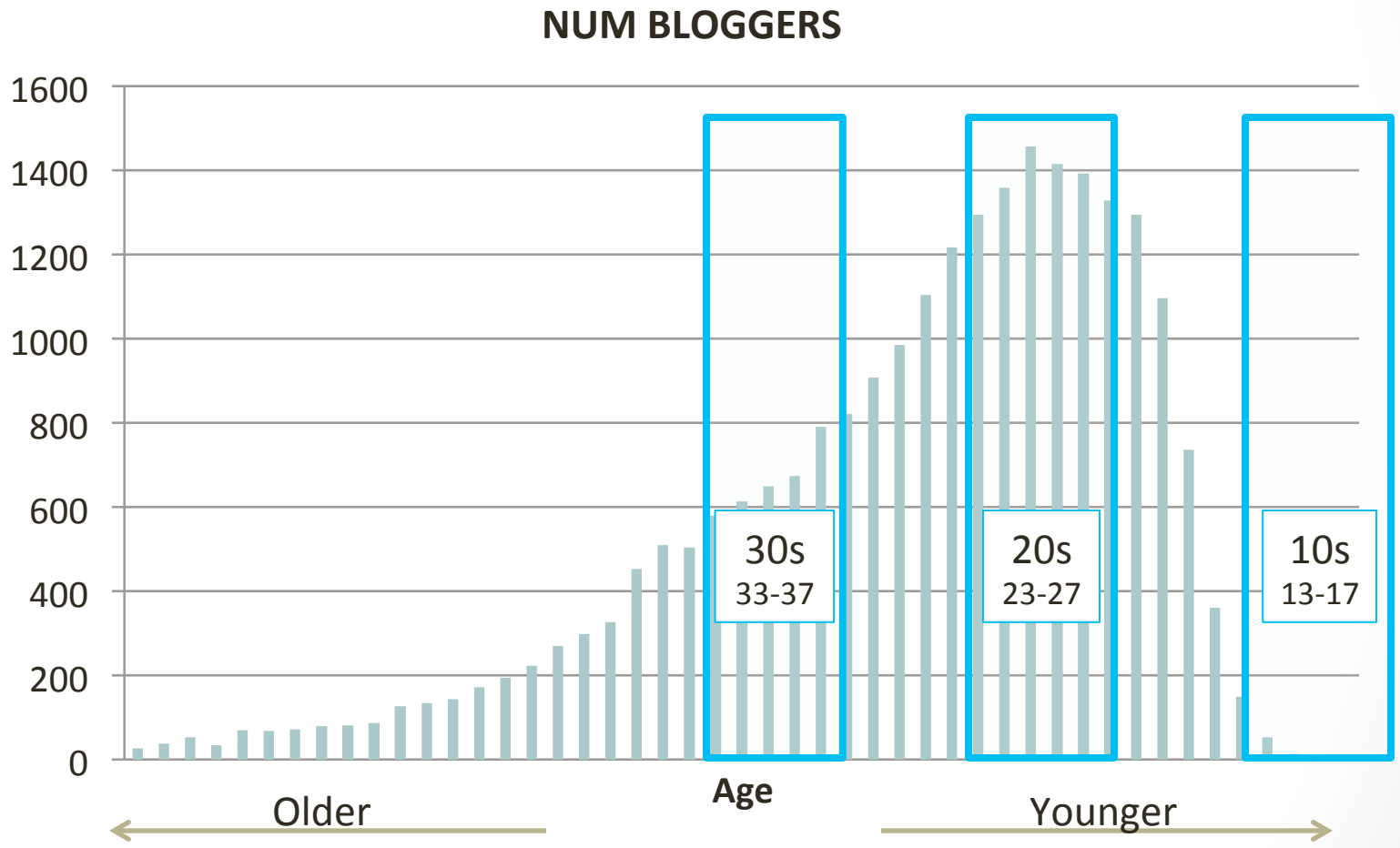


Experiment I
Age Groups

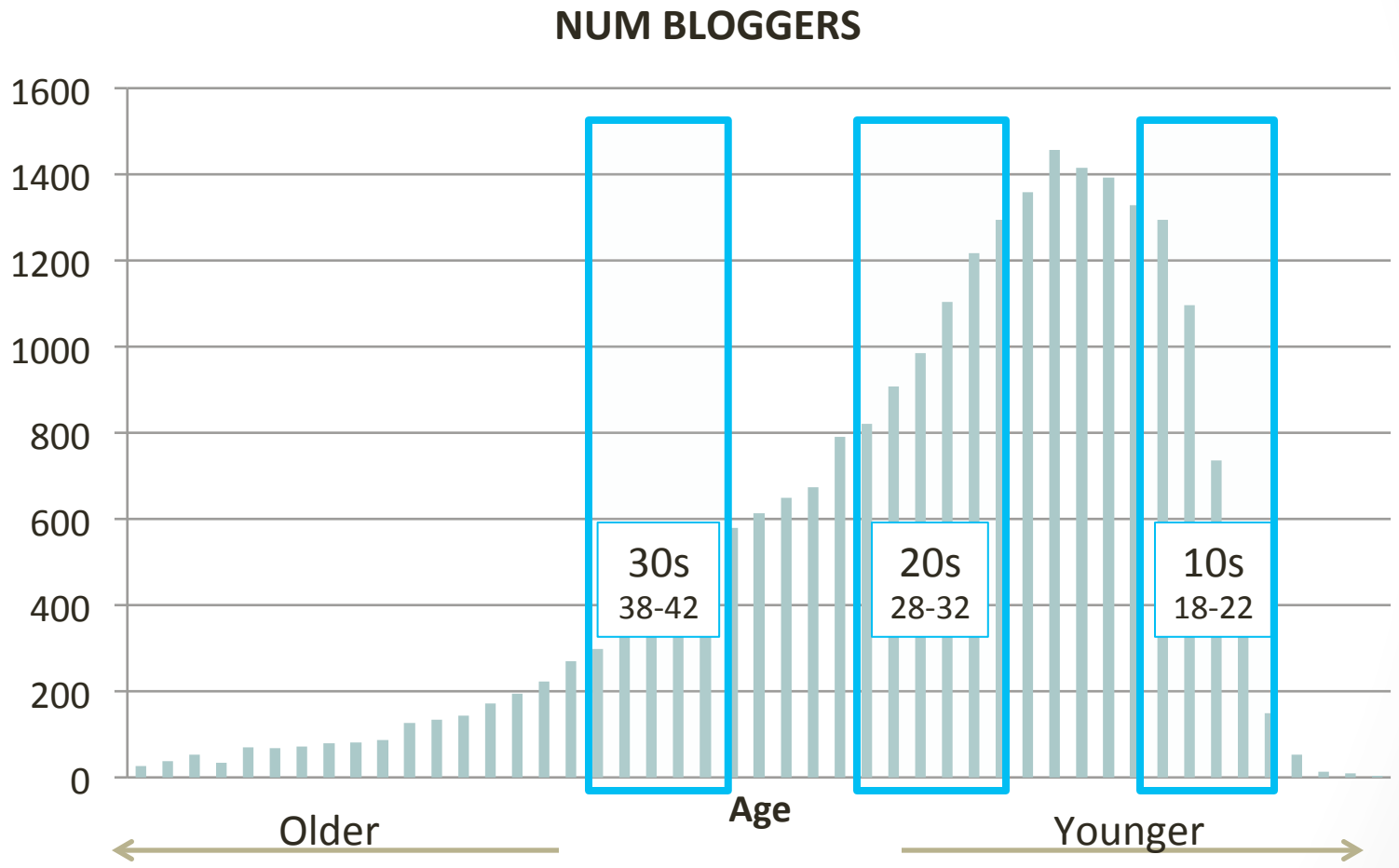


Experiment I

Age Groups – *Schler et al*



Experiment I
Age Groups



Experiment I
Age Groups

- Schler et al's corpus is much larger

	Blogger (Schler et al)	LiveJournal
# Blogs	19,320	11,521
# Posts	1.4 million	256,000
# of Words	295 million	50 million

Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I
Age Groups

- **10s vs. 30s is classified the best**
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I
Age Groups

- 10s vs. 30s is classified the best
- **10s vs. 20s are classified well**
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- **20s vs. 30s are classified the worst**
- Adding online behavior features increased accuracy by 10%

	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

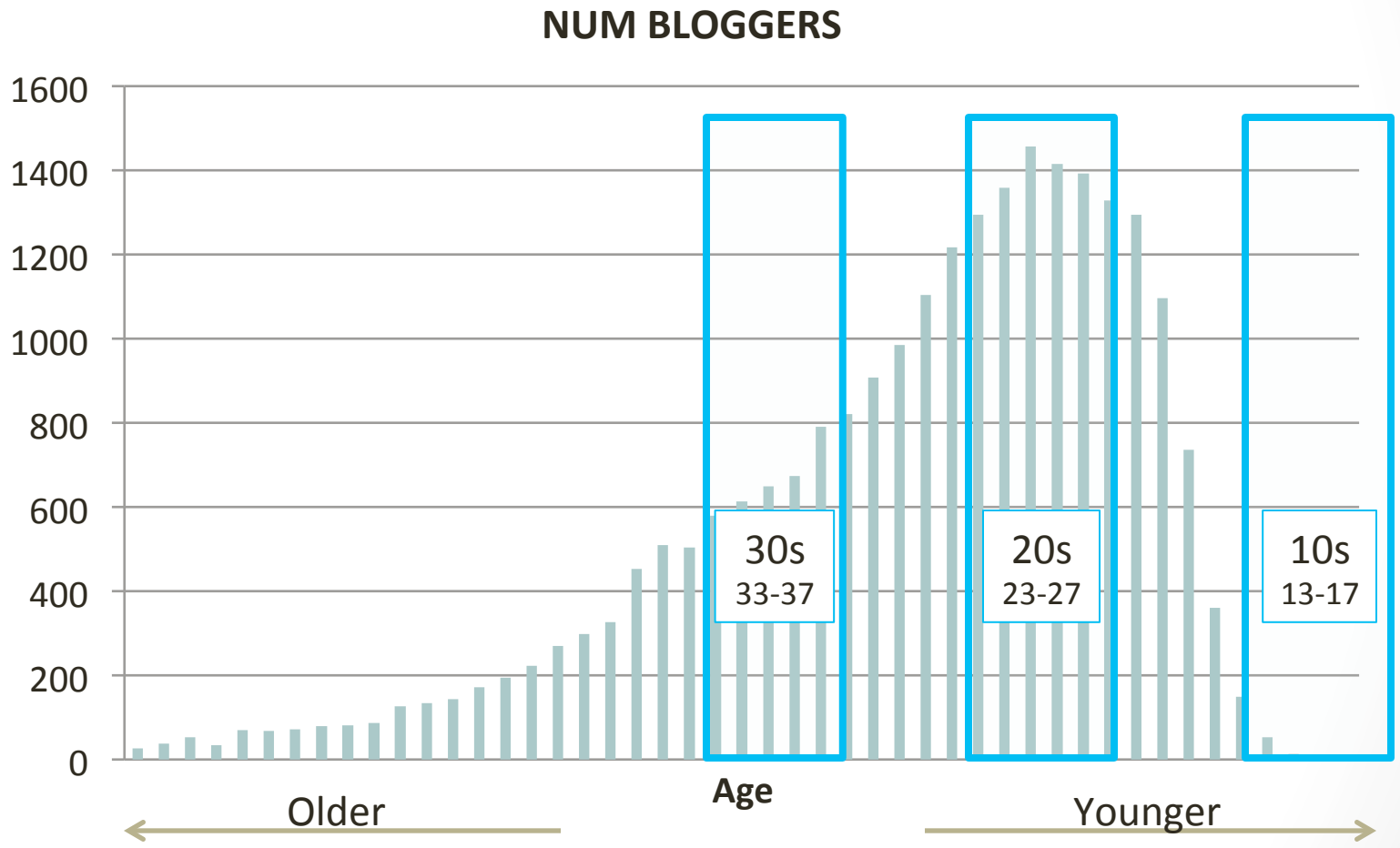
Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- **Adding online behavior features increased accuracy by 10%**

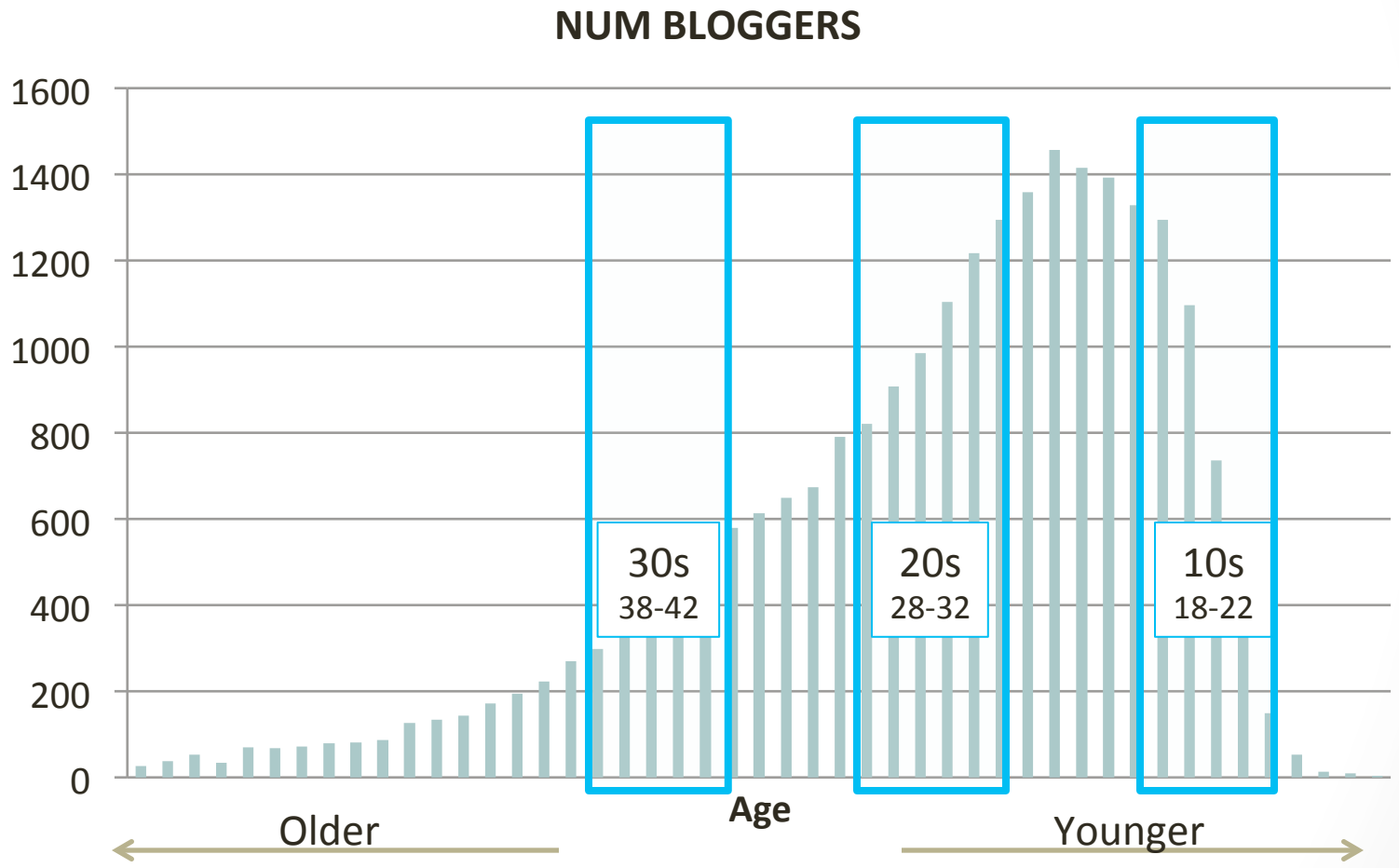
	Blogger (Schler et al)	LiveJournal	
Majority Baseline	44% (13-17)	48% (28-32)	
Style and Content	✓	✓	✓
Online-Behavior			✓
10s vs. 30s	96%	92%	95%
10s vs. 20s	87%	72%	81%
20s vs. 30s	77%	68%	72%
Overall Accuracy	76%	57%	67%

Experiment I

Age Groups – *Schler et al*



Experiment I
Age Groups

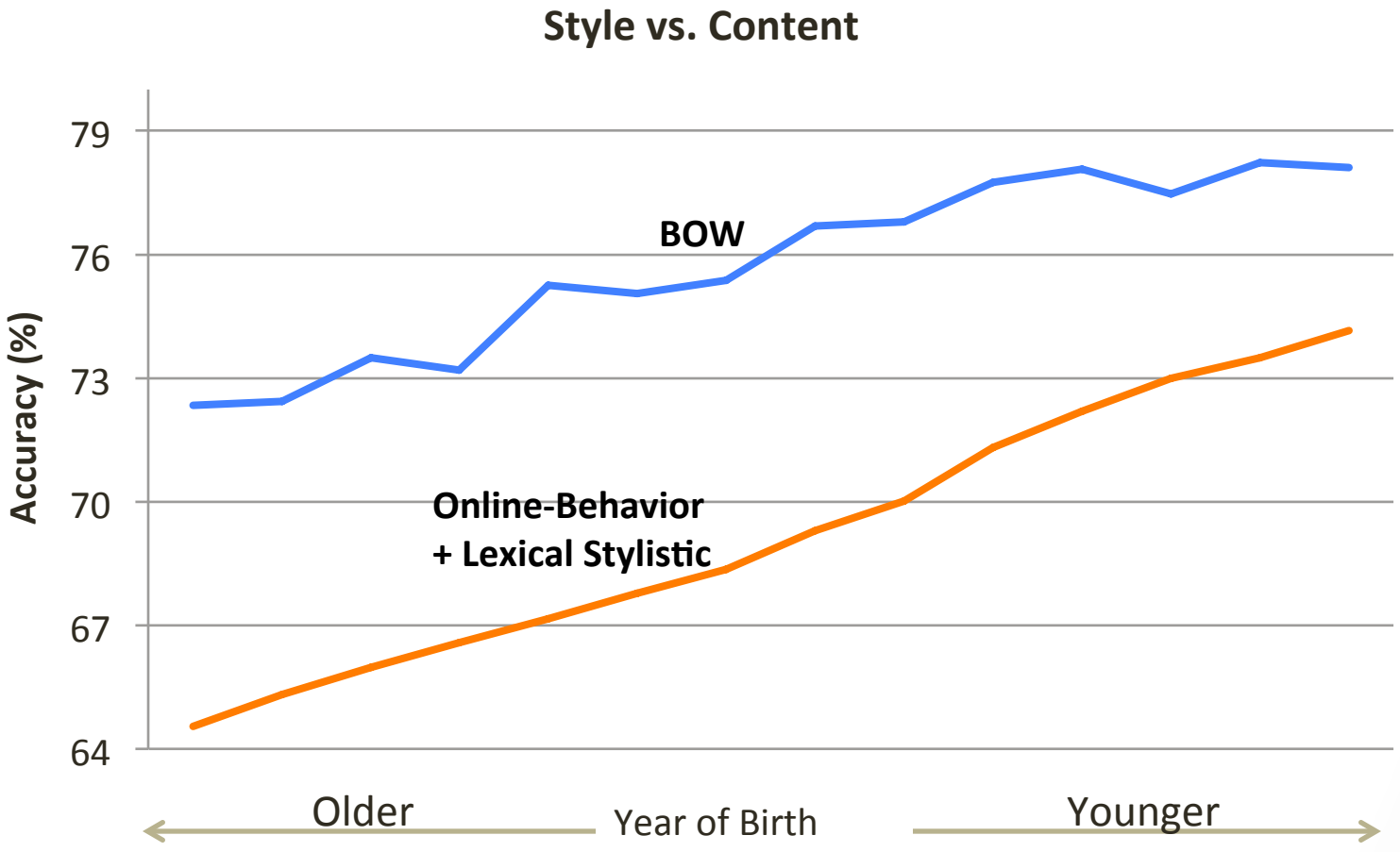


Experiment II
Social Media and Generation Y

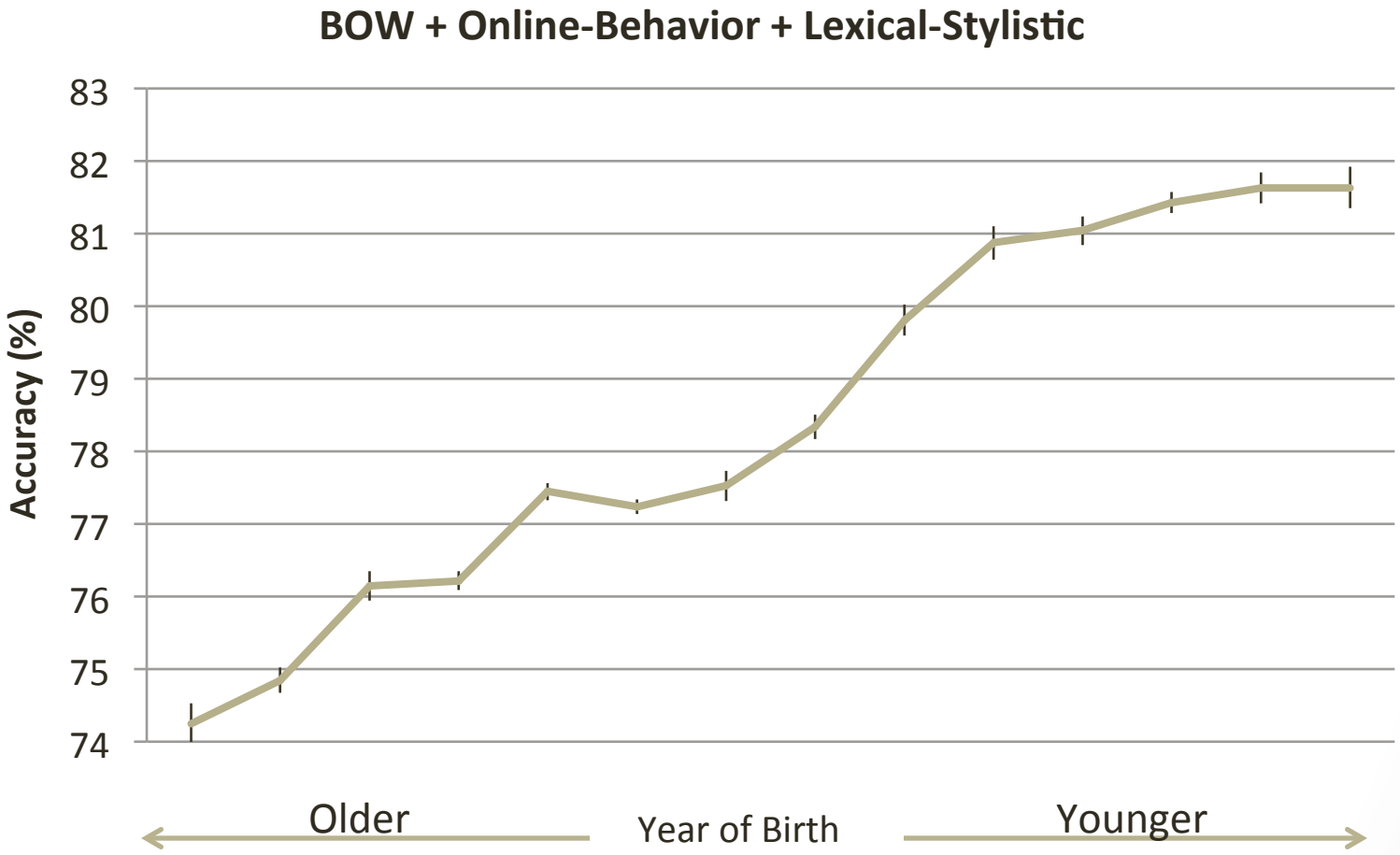
- Generation Y uses social networking, blogs, and instant messaging more than their elders [Zickuhr 2010]
- For each birth year X = 1975-1988:
  - get 1500 blogs (~33,000 posts) balanced across years BEFORE X
  - get 1500 blogs (~33,000 posts) balanced across years IN/AFTER X
  - Perform binary classification between blogs BEFORE X and IN/AFTER X

Kathryn Zickuhr. 2010. Pew Internet. Generations 2010.

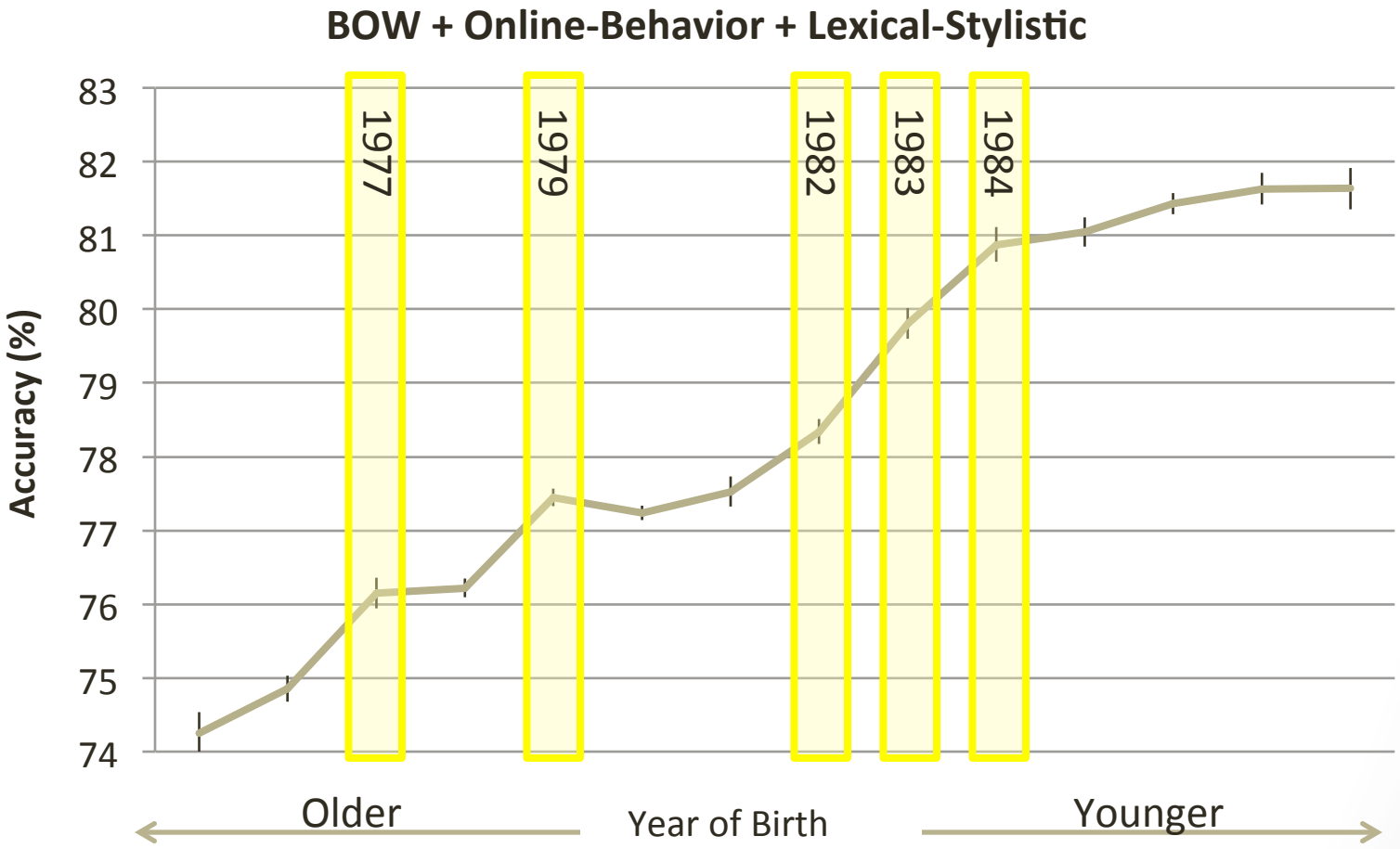
Experiment II
Social Media and Generation Y



Experiment II
Social Media and Generation Y

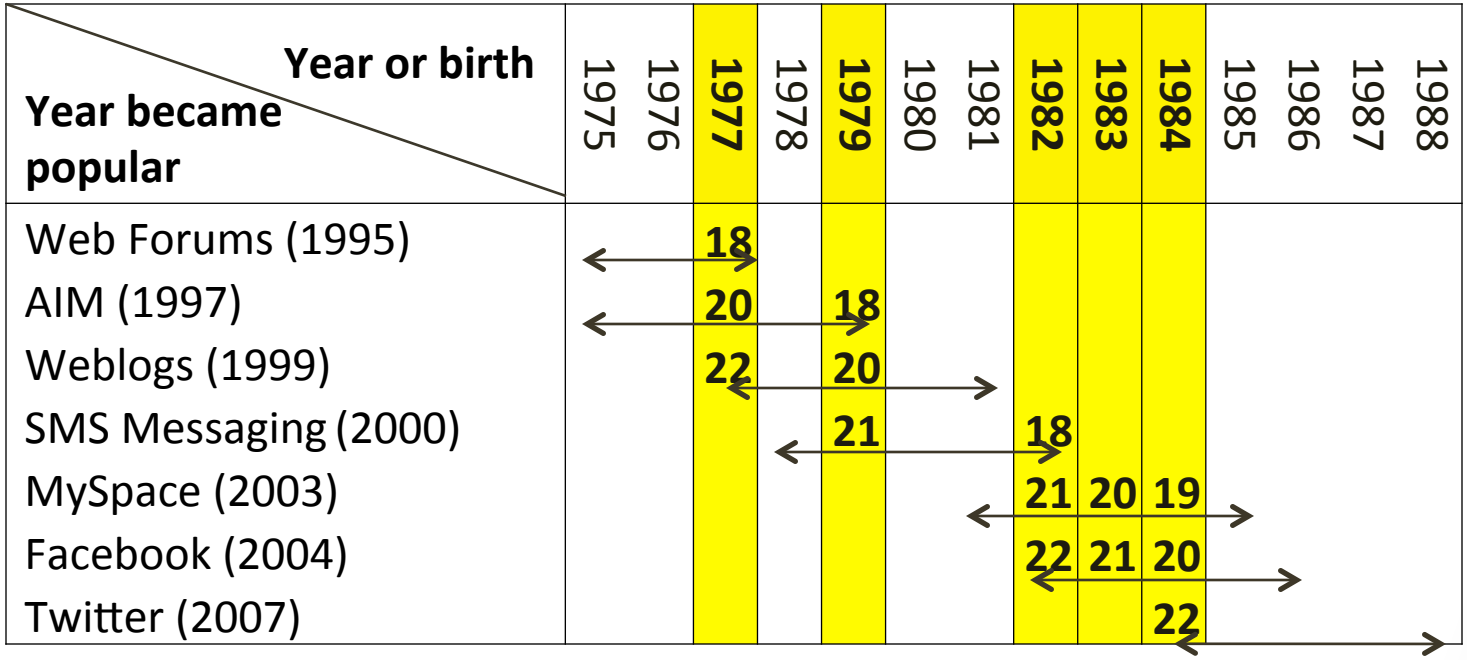


Experiment II
Year of Birth



Experiment II
Social Media and Generation Y

College Age: 22 21 20 19 18



Experiment III
A closer look

	Features	1979	1984
Baselines	Online-Behavior	59.7	61.6
	Interests	70.2	74.6
Excluding Interests	Lexical-Stylistic	65.4	67.3
	Slang + Emoticons + Acronyms	60.6	62.1
	Online-Behavior + Lexical-Stylistic	67.2	71.3
	BOW	75.3	77.8
	BOW + Online-Behavior	76.4	79.2
	BOW + Online-Behavior + Lexical-Stylistic	77.5	80.9
	BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.4
Including Interests	BOW + Online-Behavior + Interests + Lexical-Stylistic	80	81.6
	All Features	71.3	74.1

Experiment III

A closer look - *baselines*

	Features	1979	1984
Baselines	Online-Behavior	59.7	61.6
	Interests	70.2	74.6
Excluding Interests	Lexical-Stylistic	65.4	67.3
	Slang + Emoticons + Acronyms	60.6	62.1
	Online-Behavior + Lexical-Stylistic	67.2	71.3
	BOW	75.3	77.8
	BOW + Online-Behavior	76.4	79.2
	BOW + Online-Behavior + Lexical-Stylistic	77.5	80.9
	BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.4
Including Interests	BOW + Online-Behavior + Interests + Lexical-Stylistic	80	81.6
	All Features	71.3	74.1

Experiment III

A closer look - *Lexical-Stylistic*

	Features	1979	1984
Baselines	Online-Behavior	59.7	61.6
	Interests	70.2	74.6
Excluding Interests	Lexical-Stylistic	65.4	67.3
	Slang + Emoticons + Acronyms	60.6	62.1
	Online-Behavior + Lexical-Stylistic	**67.2**	**71.3**
	BOW	75.3	77.8
	BOW + Online-Behavior	76.4	79.2
	BOW + Online-Behavior + Lexical-Stylistic	77.5	80.9
	BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.4
Including Interests	BOW + Online-Behavior + Interests + Lexical-Stylistic	80	81.6
	All Features	71.3	74.1

Experiment III

A closer look - *Lexical-Content*

	Features	1979	1984
Baselines	Online-Behavior	59.7	61.6
	Interests	70.2	74.6
Excluding Interests	Lexical-Stylistic	65.4	67.3
	Slang + Emoticons + Acronyms	60.6	62.1
	Online-Behavior + Lexical-Stylistic	67.2	71.3
	BOW	75.3	77.8
	BOW + Online-Behavior	76.4	79.2
	BOW + Online-Behavior + Lexical-Stylistic	**77.5**	**80.9**
	BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations	74.8	80.4
Including Interests	BOW + Online-Behavior + Interests + Lexical-Stylistic	**80**	**81.6**
	All Features	71.3	74.1

Conclusion & Future Work

- Style, Content, and Online Behavior are useful in age prediction
- Significant changes in writing style coincide with the popularity of social media technologies
- In the future, we want to experiment with using ranking, regression, and/or clustering for age prediction

Effects of Age and Gender on Blogging

- Experiment on 3 age groups consisting of 19,320 bloggers downloaded from blogger.com in 2004
- 1,405,209 blog entries and

  295,526,889 words
- Features
  - Style-based:
    - select part of speech (pronouns, determiners)
    - function words (negation, assent)
    - blog-specific (hyperlinks, post length, blog words)
  - Content-based:
    - content words – with the highest information gain
    - LIWC categories – job, money

13-17	23-27	33-37
8240	8086	2994

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging.

Effects of Age and Gender on Blogging

- Use the Multi-Class Real Winnow learning algorithm
- Majority Baseline (13-17): 43.8%
- Results
  - Find content to be slightly more useful than style, but the combination is most useful.
  - 10s are distinguishable from 30's with accuracy above 96%
  - 10s are distinguishable from 20's with accuracy of 87.3%
  - Many 30s are misclassified as 20's, yielding overall accuracy of **76.2%**

	10s	20s	30s
10s	7036	1027	177
20s	916	6326	844
30s	178	1465	1351

- But… age changes. Will the writing style of these people change drastically in a few years?

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging.

LiveJournal

- Livejournal provides a bot policy page providing data formats to be used instead of scraping the website.
  - http://www.livejournal.com/bots/
  - XML Profile format
  - XML Entries format
  - Friends Data
  - Interests Data
- Modifications
  - Added type of friend connection to profile information
    - MutualFriend, Friend, FriendOf
  - Added interest count for entire LiveJournal network
  - Added comments to entries
- Java API
  - Can be used to download and read blogs
- Downloaded Data
  - 10000 blogs all containing age (Downloaded in 2009)
  - 30000 blogs containing age, originating from US,UK,AU,CA, and updated within the last two years (Downloaded in 2010)

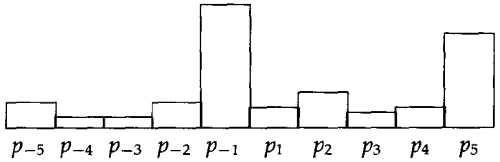
Livejournal Blog

- Profiles
  - profile name
  - full name
  - age
  - e-mail
  - bio
  - interests
  - friends list
  - Country and city/state
  - etc...

- Blog Entries
  - entries
    - date written
    - profile name
  - comments
    - profile name of person that left the comment
    - date written
    - thread structure

Xtract
Retrieving Collocations from Text

- Stage 1: Extract significant collocations from all words within a +/- 5 window. Filter out words that do not have significant ***strength*** and ***spread***. Part of speech can be used as part of the collocation as well.
  - ***strength***: w_1 occurs with w_0 more than average
  - ***spread***:



- Stage 2: Use bigrams from stage 1 to create n-grams
  - Precision: 40%
- Stage 3: Filter bigrams further to keep the ones that have good syntax relationships such as VO,SV,NN, NJ. Use those collocations to create new n-grams.
  - Precision: 80%, Recall: 90%*

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19:143– 177.
* Figure taken from paper