NEWSWORTHY ALGORITHMS

2016–11–09_0700_US_CNN_Election_Night_in_America.txt,22913eee-a64a-11e6-bdad-089e01ba0338,election,11/9/16 7:09, "THE LEADER OF THE DEMOCRATIC PARTY, AND FOR HER NOT TO WALK –>> IS IT TOO LATE?>> IS IT TOO LATE?>> YEAH.>> I THINK AS WE’VE JUST SEEN, THIS IS AN ELECTION THAT IS DEEPLY DIVIDED THE COUNTRY. THE POLLING—WE LOOK AT—WE’LL PROBABLY GET THE POLLING OVER NIGHT, BUT THE POLLING GOING IN WAS 90% OF THE HILLARY CLINTON VOTERS ","http://www.sscnet.ucla.edu/tna/edge/video,22913eee-a64a-11e6-bdad-089e01ba0338,549000"

Source: **NewsScape Archive**[1]

ALL NEWS IS LOCAL

Saturating the ensuing pages (figure 4.1) is a list of words, each one extracted from news coverage of the 2016 US presidential election: a contest in which Donald Trump upset widespread predictions that Hillary Clinton would become the first woman commander in chief. The list begins neutrally enough, with the election characterized as simply *different*, *general*, and *important*. But the discourse soon progresses to *pivotal*, *polarizing*, and *bizarre*. At its midpoint, the election campaign is *compromised*, *looming*, and *zombie*. It ends with proclamations of *magic* and *awful* but *secure*.

All these words immediately precede the term *election* in news coverage from a period starting 418 days before the vote was held, during the first candidate debate on September 16, 2015. The list ends on election day, November 8, 2016. These words—all descriptors for the election, of one sort or another—appear in the order they were used by one of three news outlets: Cable News Network (CNN), *Wall Street Journal* (*WSJ*), and Breitbart News Network (BNN). Each word appears only once in the list, at the moment of its first use (since September 16, 2015). Color indicates the news outlet that first used it: black for CNN, blue for the *WSJ*, and red for BNN. The numbers that precede each word indicate the days left until the election. Those that follow, also colored black, blue, and red, indicate how many times the term was subsequently invoked by each news outlet.

4.1 (following pages)

The NewsSpeak algorithm returns a list of words, which precede the term *election*, in news coverage from multiple sources. Image by the author and Peter Polack.

Days until election		CNN	BNN	WSJ
418	a	18		
	an	1958	24	126
	different	30		
	general	6248	9	202
	important	161		2
	last	219	1	3
	losing	10		
	presidential	1701	17	275
	recall	12		1
	school	3		
	the	8832	62	345
	these	115	1	
	this	6280	24	85
	unusual	29		2
417	-	945	7	254
	2016	151	1	53
	buy	81		1
	come	46	1	
	competitive	14		1
	faso	1		
	first	74		2
	for	144	1	5
	huge	25		
	next	101	1	2
	six	42		2
	statewide	20		
	t	3		1
	that	260	3	9
	three	21		2
	welcome	2		
	winning	44		
416	.	567	5	68
	committing	1		
	federal	112		40
414	four	13		2
	his	114	2	22
	national	153	4	20
	nec	1		
413	early	9	2	8
	fair	87		5
	have	26		
	republican	21		2
	resounding			1
	saying	5		
412	?	12		
	every	205		5
	from	170	1	2
	greek			1
	rigorous	2		
	win	108	1	5
	your	20		
411	'	6		3
	2012	127	1	28
	to	354	2	3
410	1896			13
	2014	3		6
	entire	121		
	her	21	1	
	particularly	1		
	primary	151	10	23
409	entering	2		
	midterm	52		17
408	s	237	16	62
406	assad			1
	buying	24		1
	democratic	35		5
	landslide	12	1	3
405	full	2		
	leadership	14		4
	previous	69	3	1
404	and	207	1	13
	one	50		3
	separatist	1		
	until	418		2
403	more	16		1
402	coming	13	1	14
	local	57	5	7
	many	41		
	of	380	11	37
	own	47		2
401	2009		1	2
	hold	8		2
	legislative		3	5
400	so	6		
	way	8		
399	critical	21		1
	major	6	1	2
	robust	1		
	speaker	4		1
398	ballot	3	1	2
397	2000	43		1
	changed	24		
	close	165		3
	my	20		
	on	1081	8	44
	stolen	21		
	where	3		1
396	2008	37		6
	29			1
	big	39		2
	called	1	2	

Days until election		CNN	BNN	WSJ
	class	4		
	floor			1
	internal			2
	obama	15		2
	out	9		
	seeking	4		
	short	1		
	special	32		4
	those	39	3	2
	with	37	4	5
394	parliamentarian	1		
	parliamentary	25	4	17
	upcoming	55	2	
393	lexl	1		
392	1999			1
	contentious	8		2
	delegates	1		
	fall	52		5
	mobilization	9		
	state	87	3	9
	swung			1
	traditional	6		
	two	116	2	2
	year	14		1
391	any	54		2
	change	101		
	in	227	9	13
	larger	3		
390	breaks	1		
	few	25		
	free	22		2
	growth			1
	president	7		
	recent	45	1	7
389	actual	35		
	fascinating	5		
	flower	2		
	historic	45		2
	house	6		1
	june		6	1
	potential	2		1
388	19		1	
	into	74	1	
	lose	8		
	prior	8		3
	whether	1		2
387	another	44	2	1
	congressional	6	1	6
386	new	53		3
385	canadian	4	1	
	future	11	1	1
	whole	104	1	1
384	current	17	1	1
	director			2
	past	92		7
	sixth	1		
	staggered			1
383	be	18		
	conference	2		
	particular	37		
382	1998			1
	at	19		2
	campaign	4		
	he	5		
	positive	7		
	successful	8		2
381	--	17		
	jgeneral	3		
378	breaking	5	2	
	broke		1	
377	speakers	1		
	volatile	11		
376	1994	4		1
	is	70	1	
	wild	8		
375	crazy	19		
	easy	4		
	senate	17		
374	book	1		
	endless	22		
	hard	3		
	some	16		1
	think	14		1
372	closest	11		
	since	2		
	snap	4	1	
371	actually	3	1	
	because	16		1
	been	15		
	call	5		
	clean			2
	cliffhanger	1		
	council	4		1
	later		1	
	secretary		1	
	tendless	1		
370	country	9		
	each	22		1
	ele	22		
	happy	8		
	independent			2

Days until election		CNN	BNN	WSJ
	jrl	1		
	newsletterearly		1	
	october			3
	prnl	2		
	unexpected	1		2
	yes	2		
369	about	83		2
	hope	3		
	interesting	18		
	partisan	2		
	polarized	4		2
	sweep	2		
	taiwanese		2	1
368	addition			2
367	anymore	1		
	month	1		
365	by	46	2	1
	freest	1		
	myanmar	2		
	was		1	
364	---			1
	turvy	2		
363	heated	9		3
	impact	5		
	only	22		1
	strange	44		1
	weird	11		
362	6	6		
	having	7		
	our	367	2	2
	said	11		
	their	42		6
	track	2		
360	contested	26		3
	runaway			1
358	french	7		1
	yep	2		
357	political	29		
352	facing	3		
351	runoff	10	1	
350	regional	8	2	3
	twitter	3		
349	constitution	1		
347	british	11		1
	through	55		3
346	focused	2		
	post	21		
344	get	1		
343	clinton	8		
	official	3	1	
	private	1	3	
342	time	6		1
341	atypical	5		
	union			2
339	policy	2		
338	after	113		7
335	electione	1		
	second	4		
334	basing	2		
	e	9		
	liberty	1		
333	rag	1		
	unpredictable	13		
332	drive	1		
	electoral	1		
	serious	7		
331	november	118		38
327	before	394	2	19
	polls	2		
326	cover	3		
	security	7		
	see	2		
324	base	9		
	or	15		4
323	fun	8		
322	may			1
321	decide	4		2
	lost	16		1
320	card	1		
	couple	16		
	inconclusive			2
	mock	11		
	other	33		2
	outsider	5		
	say	8		
	thermometer	3		
	thermostat	3		
	us	4		
319	great	9		
317	annual			2
	typical	10		
316	smooth	2		
315	several	26		1
314	dramatic			1
	liberal	1		
	loses	4		
313	comes	2		
311	mean	1		
308	8	4	1	4
	february	1	1	
307	no	17		1

Days until election		CNN	BNN	WSJ
	thought	1		
306	:	2		4
	per			1
	successive	1		2
304	during	26	2	3
	failed	4		
303	disputed	4		2
	iranian	3		2
302	as	33	1	4
	wave	21		
301	mandate	1		
300	above	1		
	el	3		
	presidency	3		
299	better	2		
	covering	5		
	day	4		
	persuasion	8		
	they	1		
	what	29		
298	won	24		
297	representation			1
	resume	2		
295	expert	2		
294	2010	4		
	although		2	
	parliament	1		
293	administration			1
	te	2		
292	crucial	8		
	extreme			5
	fought	7		4
	governor	3		
	gubernatorial	5	1	1
	municipal	4	3	1
	wins	6		
291	96	1		
	consequential	31	1	
	months			1
	moving			1
288	head	2		
287	1992	3		1
	germ	9		
	representative	2		
	skwlen	1		
	talked	1		
	tumultuous	3		3
286	american	143	1	7
	but	12		
	consecutive	7		1
	overall	3		
285	alter			1
	extraordinary	6		
	jebl	2		
	large	2		
283	"	4		11
	influence	11	3	2
	like	6		
282	between	6		2
	how	14		1
	wide	3		
281	iowa	10		
	wacky	1		
280	all	17		
	changer	1		
	dunk	1		
	normal	20		1
	sanders	4		
279	march	1		
	mayoral	2		
	ordinary	10		1
	reset	1		
	term	22		
	turnout	45		
278	20	3		1
	around	14		2
	determine	2		
	overcome			1
	professional	2		
	unfair	2		
277	over	16		
	talk	1		
	top	4		
276	prescinctpreci	2		
	sanctioned	1		
	wonderful	1		
275	tough	17		
	tumble	8		
274	disturbing	2		
	open	6		1
273	hampshire	2		
272	amazing	9		
	either	1		
271	15	4		
	cnn	118		
	covered	10		
	party	5	2	1
	processing	1		
	roll	1		
	trump	13		1
270	send	1		
	single	19		2
269	1824	2		
	passing	1		
	stole			1
268	raucous	2		1
	word	2		
267	clarifying	3		
	divisive	5		1
	duck	1		
	monumental	2		
	tara	1		
265	consider	1		
	manipulating	1		
	normally			1
	protracted	1		1
264	are	30		
	f	1		
	final	19		
	make	4		1
	pivotal	6		1
	real	13		
	someone	2		
263	unique	13		
262	appn	1		
	five	21		
	polarizing	4		1
	woolly	2		
260	butter	1		
	difficult	7		
	states	14	1	2
	tight	26		2
259	[			1
	predicting	1		
258	says	2		
	today	3		
257	]	1		
	08	7		
	candidate			1
	good	11		
	gop	3		1
	held	3	1	2
	holding	5		
	nn	2		
	washington	2		
256	dirty	1		
255	ll	1		
254	carolina	6		
	numbers	1		
253	ten	1		
252	1996	3		
	jen	1		
	modern	16		1
251	1972	2		
	board	4		4
	broad	1		
	generally	3		
	had	16		
	hit	3		
	stakes	3		
	when	23		
250	1828			1
	apop	1		
	rags	1		
	supercompetitive	1		
249	1960	4		2
	colorful	3		
	inevitable	2		
	sleepy	2		
	throughout	4		
	unprecedented	15		
248	app	1		
	weekend	1		
	which	6		
247	closed	3		
246	eve	1		
	funded	1		
	half	3		
245	1984	5		
	nationwide	4		1
244	county	14		
	deadlock	1		
243	direct	6		
	handles	1		
	hypothetical			1
	macro	2		
	organizing			2
	same	5		
	surreal	4		
	suze	1		
242	15th	1		
	1980			1
241	look	5		
	ohio	1		2
	possible	1		1
	run	32		
	running	2		
	vigorous	1		
240	peaceful	2		
239	1860	1		
238	know	10		
	worst	6		1
237	2004	5		3
	certify	1		
	closer	6		
	florida	7		
	key	5		2
	mature	1		
	tuesday	20		
236	angriest	1		
	complex			1
	gentle	2		
	hillary	2		
	toxic	4		
	volcanic	1		
235	against	5		
	beyond	3		
	explosive	1		
234	nominating	3		
233	2018	1		1
	utah	2		
232	heavy	2		
	pending	6		
	power	2		
231	beg	1		
	exciting	12		
	remember	1		
230	tree	2		
229	average			1
	bipartisan	4		
	sol	2		
	surprise			1
	western	3		
228	immediate	1		
226	bizarre	5		2
	complete	7		
	remarkable	1		
	separate	2		
	september			1
	thus	1		
225	1948	2		
	emotional	2		
224	nasty	16		
223	27	1		
	54	1		
	7	1		
	seven	8		
221	2013			2
	changing	4		
220	both	11	1	1
218	century	1		
	rifting	1		
	stronger	1		
216	bush	1		
	governors	4		
	judicial	4		
	will	2		
215	contest	5		
	cruz	3		
	formal	3		
	hear	2		
	you	2		
214	towards	13		
213	delegate	18		
	disrupted	2		
212	30	2		
210	12	4		
	colorado	2		
	most	22		
	multiple	7		
	posthumous	1		
	regular	6		
	sponsored	2		
209	11	12		
	cancelled	1		
	swept	1		
208	catastrophic	1		1
	given	3		1
	risky	2		
207	decisive	3		
206	reform	1		
205	cancel	6		
	difference	2		
204	media	2		
	social	1		
203	crooked	2		
	dreamy	1		
	preference		1	
	rigged	572	3	2
	transparent	3		
202	electi	2		
	empowered	2		
	influencing	4		
	pluralities	1		
	question	3		
201	1990	1		
	does	1		
	outsiders	5		
	voterless	6		
	voting	6		
200	newy	1		
196	individual	3		
	sign	1		

Days until election		CNN	BNN	WSJ
195	confident	2		
	course	4		
	fourth	2		
	mood	8		
	predict	1		
	ultimate	6		
194	!	1		
	biggest	5		
	genderal	1		
	likely	1		
	message	2		
	repeat			1
193	eeral	1		
	referendum	2		
192	home	4		
191	california	3		
	consequences	2		
	help	1		
188	benefits	1		
	gets	1		
	presidenti	2		
187	its	4		
186	mexican	1		
185	choice	4		
	moment	2		
	turbulent			1
184	coalition	1		
	incredible	3		
183	2006	1		
	theory	2		
182	cut	1		
	understand	1		
181	50	1		
	brutal	5		
	gap	1		
	gem	2		
	nastiest	2		
	romney	1		
180	immobilization	2		
	le	4		
	overnight	2		
	spirited	1		
179	forry	1		
	usual	11		
178	looks	2		
	puzzling	1		
	research	2		
	ugly	11		
177	conventional	6		
176	faulty	2		
175	alex	1		
	controversial	2		1
174	candidates	3		
	claims			1
	lbj	3		
173	administer	1		
	chief	6		
171	putting	5		1
170	popular	2	1	
168	2015	1		1
	corporate			1
	layup	2		
	nose	2		
	uninspiring	1		
167	fantastic	1		
	swings	3		
	violated	1		
166	elec	3		
	job	1		
165	antiestablishment	1		
	break	8		
	establishment	3		
	german	1		1
	movement	2		
164	main	1		
162	financing	1		
	oning	1		
	sometimes	1		
	sway	3		
161	8th	20		
	mcgovern	1		
160	various	3		
159	legacy	1		
	tv	1		
158	64	2		
	rare	3		
156	boxing	1		
	probably	2		
155	genera	1		
154	dakota	2		
	still	2		
153	gat	3		
	superdelegate	2		
152	flagship	1		
	n	1		
	odd	4		
151	ae	1		
	hand	1		
150	2002	1		
	surprising			1
149	1964	2		

Days until election		CNN	BNN	WSJ
	equal	2		
144	shall	1		
	tougher	1		
	well	2		
141	provincial	3		
140	broader	3		
	then	3		
139	based	1	2	
	outside	2		
	unconventional	6		
138	bag	1		
	released	1		
137	20916	1		
	effective	1		
136	collection	1		
	dollar	1		
	european	2		
	overwhelming	4		
	th	3		
133	versus	1		
132	2			1
	right	8		
	shl	2		
131	04	1		
	den	3		
130	1940	1		
127	there	1		
125	face	1		
122	collects			1
120	australian	1		
119	following	2		
116	point	2		
115	near	2		1
113	indiana	1		
111	truth	2		
110	bowl	1		
	harder	2		
109	%	4		
	accepting	1		
	dprl	1		
107	beautiful	3		
	personality	4		
	unending			1
106	insurgent	1		
104	straight	2		
103	quo	1		
102	wrong	3		
101	countries	2		
	definitional	1		
	funding	2		
	goat	2		
	negative	2		
	up	6		
	using	2		
99	affects	1		
	were	8		
98	quote	1		
	quoting	1		
97	politics	5		
96	global	4		
	uncertain			1
95	costs	2		
	honest	5		
	reflective	2		
94	crossroads	5		
	questionable	2		
	umpteenth	1		
93	believe	1		1
	compromised	8		
	perhaps	2		
	respective	1		
92	historical	3		
90	13	1		
	emptive	3		
	gore	3		
	investigating	1		
	profile	1		
	steal	7		
	uplifting	2		
89	hack	2		
	lebanese	3		
	phony	6		
88	specifically	1		
86	become	2		
	bombshell	2		
	do	4		
	legit	1		
	legitimate	3		
	recruit	2		
	shell	2		
	tb	2		
	tur	2		
84	observe	1		
	rig	5		
83	despite			1
	i	3		
	mega	1		
	rae	1		
	third	4		
81	contrast	1		
	whacky	1		

Days until election		CNN	BNN	WSJ
79	finally	1		
	virginia	1		
76	rigging	9		
73	toward	3		
72	election	2		
	white	2		
71	winnable	1		
70	1988	1		
	monitor	2	1	
	viable	3		
69	arizona			2
	everything	1		
	falsify	1		
	illinois			1
	targeting	1		
	than	4		
68	characterized			1
	polish			1
67	aggressive	3		
	apples	1		
65	off	2		
	otherwise	2		
64	dispiriting	2		
	sad	11		
63	2011		2	
	foush	1		
	once	1		
	till	6		
62	trust	2		
61	ethe	2		
	life	1		
	yial	1		
58	analyze	2		
	looming	1		
	nine	2		
56	facts	1		
	ninth	2		
55	approach	6		1
	driven	2		
	even	5		
	gear	1		
54	ebb	1		
	legal	1		
	not	7		
	russian	8		
53	debates	3		
	doing	2		
	down	4		
	messy	4		
52	button	1		
	clear	40		
51	russians	1		
50	winnings	2		
48	com		1	
	performance		1	
46	making	1		2
42	certain	2		
	ddemocratic	1		
	heat	1		
	inspirational	1		
	tie	4		
41	chris	1		
	egg	1		
	straightforward	1		
40	disrupt	3		
	eat	1		
	strangest	4		
	swing	3		
	unorthodox	3		1
39	attempted	2		
	it	2		
	military			1
	tied	2		
	while	2		
38	grade	1		
37	bad	1		
	modify	1		
	slinging	1		
36	again	1		
35	true	1		
34	latest	5		
	vicious	1		
33	swan	1		
32	hoping	1		
30	31	1		
	insane	3		
	premier	1		
29	whose	6		
27	computerized	1		
	dece	1		
	fraudulent	3		
	interest	1		
	matters	1		
	officially	2		
	style	1		
	thin	1		
	though	1		
	zombie	1		
26	discouraging	2		
	downballot	2		
	why	7		

Days until election		CNN	BNN	WSJ
25	rough	1		
	sick	1		
24	apparent	1		
23	2020	1		1
	attacked	1		
	scandal	3		
22	1982	5		
	dismal	2		
	divided	18		
	example	1		
	hijacked	2		
	horrible	1		
	prosecute	2		
	prosecuted	4		
21	51	3		
	bitter	2		
	discredited	2		
	fraud	2		
	furniture	1		
	love	6		
	luck	1		
	questioning	2		
	scale	3		
	study	3		
	swayed	3		
	thinking	2		
	very	1		
	widespread	5		
20	city	3		
	conspiracy	1		
	feel	1		
	fix	2		
	fixed	3		
	govern	3		
	mind	3		
	observing	1		
	studied	4		
	traumatic	1		
	trig	2		
19	7th		2	
	if	1		
	oversee	3		
	ricked	1		
	ringed	2		
	selling	1		
	smaller	2		
	support	2		
	tabloid	1		
18	accept	6		1
	al	3		
	bogus	2		
	corrupt	2		
	decided	2		
	dozen	1		
	elect	4		
	finale	2		
	finish	2		
	orderly	2		
	rampant	1		
	thing	1		
	voted	1		
17	bruising	1		
	disputes	1		
	extended	1		
	huh	1		
	knuckles	1		
	monitoring	4		
	precise	1		
	rankerous	2		
	stealing	1		
16	anchors	2		
	defeat	2		
	nowrom	1		
	waive	2		
15	elhe	1		
	former	2		
	ged	2		
	got	2		
	magic	1		
	stressedhis	2		
14	americans	3		
	carroll	1		
	didvicive	1		
	speakership	1		
	statement	1		
	wat	1		
13	biter	2		
	celebration	2		
	chad	2		
	comparable	1		
	conduct	2		
	demographic	3		
	ending	1		
	fm	1		
	shortened	1		
	under		1	
12	club	5		
	gaffe	1		
	terrible	1		
	tremendous	1		
11	4	1		

Days until election		CNN	BNN	WSJ
	40			1
	bunker	1		
	chaotic	1		1
	draining	1		
	government			1
	international			2
	reagan			1
	skip	1		
	updated	1		
	warning	1		
10	david	1		
	dented	1		
	spending	6		
	undermine	1		
9	impending	1		
	I	1		
	predictable	1		
	sheriff	2		
7	bhfr	1		
	cannibalize	1		
	controlled		2	
	ecome	1		
	electronic		2	
	suppressed	1		
6	baked	1		
	citing	1		
	craziest	1		
	dull	1		
	electio	1		
	intense	1		
	now	1		
	princeton	1		
5	debate	2		
	defining	3		
	meanwhile	1		
	offee	1		
	plot	3		
	populous	1		
	thoo	1		
	voters	4		
	yeah	1		
4	america	4		
	bee	1		
	lekd	1		
	night	1		
	pg	1		
3	inhis	1		
	late	1		
	need	1		
	planning	1		
	requiring	2		
	watch	2		
	work	1		
2	able	2		
	bring	1		
	catch			1
	daye	2		
	foreign	2		
	healing	12		
	just	3		
	lag			2
	leak	1		
	live	2		
	shape	1		
	subsequent	2		
	tense	1		
	them	1		
	wildest	2		
1	1800	1		
	conscience	2		
	epic	1		
	similar	1		
	watershed	2		
0	1932			1
	being	1		
	criticizing	2		
	defines	1		
	destabilize	1		
	ensuring			1
	here	1		
	intermable	2		
	keeping	2		
	long	3		
	memorable	2		
	use	1		
	watching	1		1
	2e679ed	1		
	awful	1		
	especially	3		
	hopeful	1		
	justice	2		
	means	1		
	poll	1		
	preliminary	1		
	roughest			1
	secure	1		

For example, the fifth word used to describe the election is *important*. It comes from this statement made on CNN in the first few moments of a debate on September 16, 2015: “This is an *important* election with an enormous number of challenges facing the American people and the first four questions are about Donald Trump.” *Important* was eventually used directly before *election* 161 times by CNN and twice by the *WSJ*. Based on this list, BNN does not seem to have used the word *important* to characterize the election at all.

Assembled together on these pages using a simple algorithm (or computer program) that I created, called NewsSpeak, these words offer a rule-based reading of election news as it might be revealed through data.[2] NewsSpeak is not intended to offer a comprehensive view of election coverage. Nor is it meant to produce a political analysis of the news during the eventful period that it covers. The algorithm is much more mundane: it calls attention to important processes that are too often taken for granted as routine in contemporary practices of computing applied to text. NewsSpeak is designed to be an encounter with the locality of “natural language” data.

The bulk of the words listed by NewsSpeak, specifically those extracted from CNN coverage, were obtained from NewsScape, an online television news archive based at the University of California at Los Angeles that has assembled more than four hundred thousand broadcasts from around the world, some of which date back to the US Watergate scandal in the early 1970s.[3] The data excerpt at the beginning of this chapter provides an example of that data in its original form, as it was digitized from CNN’s early morning coverage on November 9, 2016, the day after the election. The date and station identification are listed first, followed by text derived from closed captioning. The final string is an internet link that leads directly to an excerpted online video of the coverage.

NewsScape is one of few such large-scale archives that seek to assemble comprehensive collections of the news as data.[4] But NewsScape doesn’t contain newspaper or online journalism. The two other sources used in NewSpeak, the *WSJ* and BNN, are scraped from ProQuest and the BNN site, respectively, using yet another algorithm,

4.2

CNN video snapshot from November 9, 2016, attained using the Internet Archive TV News Archive.

4.3

Wall Street Journal front page from November 9, 2016, archived by the *New York Times*. Victor, “Trump’s Victory, on Front Pages Worldwide.”

THE WALL STREET JOURNAL.

DOW JONES | News Corp · WEDNESDAY, NOVEMBER 9, 2016 - VOL. CCLXVIII NO. 111 · WSJ.com · ★★★★ $3.00

DJIA 18332.74 ▲ 73.14 0.4% · NASDAQ 5193.49 ▲ 0.5% · STOXX 600 334.91 ▲ 0.3% · 10-YR. TREAS. ▼ 12/32, yield 1.867% · OIL $44.98 ▲ $0.09 · GOLD $1,273.40 ▼ $4.90 · EURO $1.1025 · YEN 105.17

PRESIDENT TRUMP

POPULIST SURGE LIFTS REPUBLICAN TO UPSET

Clinton lost in key battleground states; GOP keeps Senate and House

What's News

Business & Finance

Electoral gains by Trump on Tuesday evening spurred a sharp decline in U.S. stock futures and a broad flight to safety around the globe. Government bond yields slid. A1, C1

◆ **U.S. businesses** will face uncertainties with a Trump win in the election. A3

◆ **Walgreen sued** Theranos, alleging that its former laboratory-testing partner breached a contract between the two companies. B1

◆ **A Chinese aluminum firm** is evading U.S. trade sanctions, the government said, as it tries to rein in a flood of cheap metal imports. B1

◆ **Alphabet ousted** two managers on its drone-delivery project amid infighting on its team. B1

◆ **Amazon could be** in the crosshairs of Europe's taxman, and the stakes for the online retailer are high. B7

◆ **CVS warned** that it could lose 40 million prescriptions next year as Walgreens deals shut out CVS pharmacies. B2

◆ **Valeant cut** its annual forecast again, sending shares tumbling 22%. B2

◆ **Icahn more** than doubled his Hertz stake, the same day the rental-car firm's stock sank 22.5%. B3

World-Wide

◆ **Trump won** the 2016 election, becoming the nation's 45th president by defeating Clinton in one of the biggest upsets in U.S. political history. Defying polls and political pundits, the Republican triumphed over his Democratic rival by running up extraordinary vote totals in rural areas and working-class counties across the U.S. where his message of economic populism resonated. A1

◆ **Republicans captured** a series of hard-fought Senate races Tuesday night, allowing the party to retain the chamber's majority and solidify complete GOP control of Congress. A4

◆ **Democrats registered** only modest gains in the House, falling short of their most ambitious plans to put a significant dent in the Republican majority. A4

◆ **Turkish guards** abused political prisoners in the wake of the July coup attempt, eight of the prisoners alleged in a letter. A15

◆ **Iraq began** to identify bodies in a mass grave in a village near Mosul, from which Islamic State reportedly forced out residents. A16

◆ **Hungarian lawmakers** rejected the prime minister's proposed ban on future EU resettlement of migrants in the country. A16

CONTENTS

© Copyright 2016 Dow Jones & Company. All Rights Reserved

President-elect Donald Trump addresses supporters in Manhattan early Wednesday after his stunning upset of Hillary Clinton.

Forecasts Missed Breadth of Support

By Janet Hook, Colleen McCain Nelson and Beth Reinhard

Republican Donald Trump won the 2016 election, becoming the nation's 45th president in one of the biggest upsets in U.S. political history.

Mr. Trump reached his stunning triumph over Democrat Hillary Clinton in the early morning hours Wednesday, after a cliffhanger night of vote counting that capped the bitter, monthslong presidential contest.

Defying polls and political pundits, Mr. Trump ran up extraordinary vote totals in rural areas and working-class counties across the U.S., where his message of economic populism resonated. It was an appeal the first-time candidate used in the

Please see VOTE page A8

'Deplorables' Rise Up To Reshape America

By Gerald F. Seib

ANALYSIS

The deplorables rose up and shook the world.

"Deplorables" was, of course, the disparaging term Hillary Clinton at one point applied to some supporters of Donald Trump. Many of his loyal followers proudly embraced the insult and used it as a motivating tool.

Wearing such establishment disdain as a badge of honor, the Trump army cut a deep swath through the American electoral system Tuesday, propelling the Republican nominee to a stunning victory.

In the process of exceeding virtually all expectations, Mr. Trump has remade the Republican party in his own image. He rewrote some of the GOP's

Please see RISE page A8

Markets Jittery on Early Results

By Ira Iosebashvili, Min Zeng and Chelsey Dulaney

Electoral gains by Republican presidential nominee Donald Trump on Tuesday evening spurred a sharp decline in U.S. stock futures and a broad flight to safety around the globe, underscoring trepidation among investors that the New York developer's candidacy would succeed.

S&P 500 futures slumped as much as 5%, hitting a trading limit, before slightly paring losses, as Mr. Trump pulled ahead of Democratic nominee Hillary Clinton. CME Group doesn't allow its stock-market futures contracts to drop more than 5% overnight.

Japan's Nikkei stock index was down 5.3%, crude oil slumped more than 3% and the dollar fell 3.6% against the Japanese yen, while the prices of U.S. Treasurys and other rich-country government bonds rose, sending yields down. The dollar was up as much as 13% against the Mexican peso. Gold, a haven, was up almost 5% at one point.

Tuesday evening's reversal set in stark relief the markets' fear of a possible Trump presidency. During his campaign he has advocated for sharply controlling immigration and raising tariffs on trade, decisions that many economists contend could pressure growth at trading partners such as Mexico and potentially reduce already sluggish global growth.

Many investors also are uneasy about the lack of specifics underlying some Trump proposals and the economic and geopolitical uncertainty that they contend would arise as a result of his presidency.

Please see REACTS page A12

◆ Global bond yields shrink, expectations for interest-rate hike take hit C1

ELECTION 2016

Exit polls find voters show a deep hunger for change A2

Businesses grapple with new populist landscape A3

Republicans maintain majority in Senate A4

See WSJ.com for the latest vote results and election coverage

A Campaign Unlike Any in Modern Times

By Michael C. Bender and Peter Nicholas

In his last flurry of campaign stops, the Republican nominee Donald Trump began to exercise more discipline. Finally, it seemed, months after the campaign kicked off, after his impromptu Twitter tirades, the accusations of sexual harassment, the off-the-cuff comments and the uneven debate performances, his advisers had gotten control of their unwieldy boss.

Not so, according to Mr. Trump.

"I do what I want to do," he said in an interview over the final weekend. Mr. Trump predicted an "unprecedented victory," after which he would celebrate for "about an hour. Then I'll get up Wednesday morning, and start working so hard immediately."

Few believed him. Polls showed Democrat Hillary Clinton with a slender but stabilizing lede. Her party talked of taking back the Senate. The election, an unedifying, raucous and unpredictable contest, the strangest in the modern era, defied all the predictions.

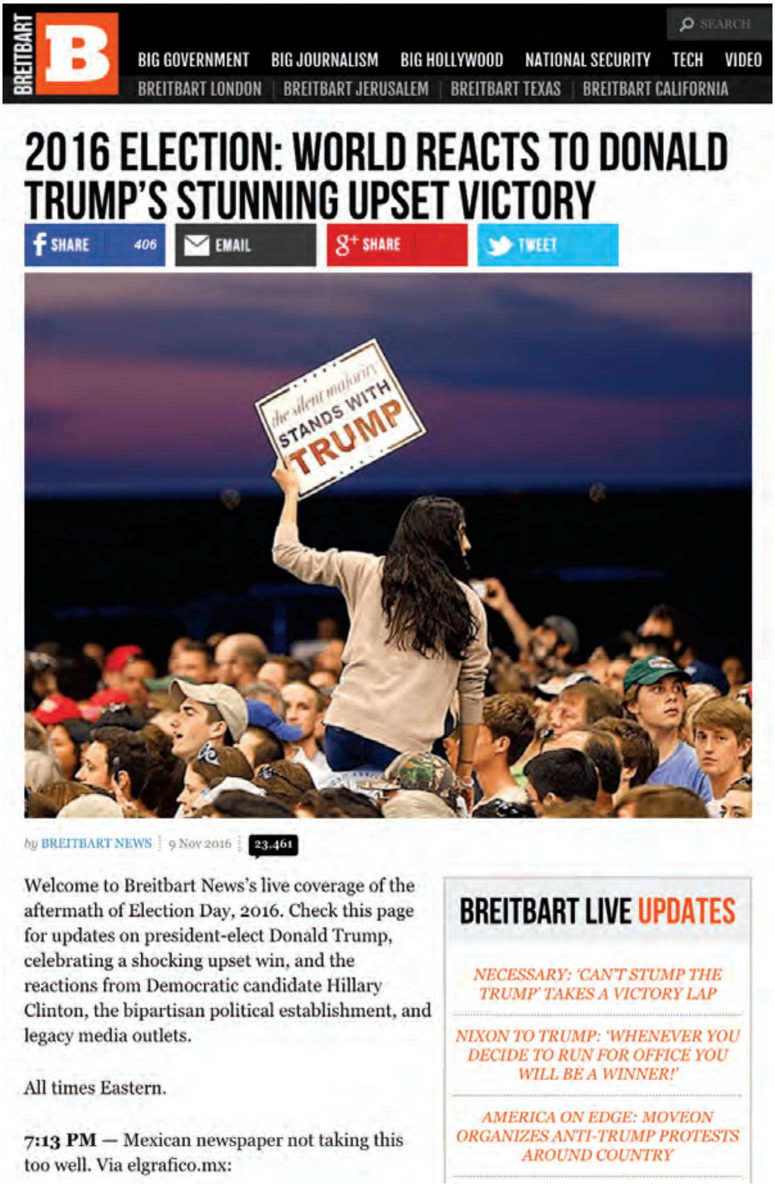
At the New York Hilton Midtown, where the Republican nominee watched election returns, cheers went up from throngs of men in suits and women in red cocktail dresses every time a state was called for Mr. Trump.

An eager volunteer handed out red "Make America Great Again" hats while supporters posed for pictures against giant campaign-branded backdrops.

"It's starting to feel electric like it does at Trump rallies," said New York City lawyer Mark Smith. As it got closer to midnight, a cake in the shape of Mr. Trump's head was wearing a campaign hat.

Mr. Trump watched from his campaign headquarters "war room." As he has been from day

Please see RACE page A10



BREITBART

BIG GOVERNMENT BIG JOURNALISM BIG HOLLYWOOD NATIONAL SECURITY TECH VIDEO

BREITBART LONDON | BREITBART JERUSALEM | BREITBART TEXAS | BREITBART CALIFORNIA

SEARCH

2016 ELECTION: WORLD REACTS TO DONALD TRUMP'S STUNNING UPSET VICTORY

SHARE 406 EMAIL SHARE TWEET

by BREITBART NEWS | 9 Nov 2016 | 23,461

Welcome to Breitbart News's live coverage of the aftermath of Election Day, 2016. Check this page for updates on president-elect Donald Trump, celebrating a shocking upset win, and the reactions from Democratic candidate Hillary Clinton, the bipartisan political establishment, and legacy media outlets.

All times Eastern.

7:13 PM — Mexican newspaper not taking this too well. Via elgrafico.mx:

BREITBART LIVE UPDATES

NECESSARY: 'CAN'T STUMP THE TRUMP' TAKES A VICTORY LAP

NIXON TO TRUMP: 'WHENEVER YOU DECIDE TO RUN FOR OFFICE YOU WILL BE A WINNER!'

AMERICA ON EDGE: MOVEON ORGANIZES ANTI-TRUMP PROTESTS AROUND COUNTRY

4.4

Breitbart web page story from November 9, 2016, captured from http://www.breitbart.com/.

the details of which I won't delve into here. Neither offers a direct means of downloading news coverage over an extended period of time.

These three sources are of interest because of their historical and material differences. The latter has to do with the "materialities of information representation" touched on in chapter 1, which Paul Dourish explains as the particular forms that information takes, and their implications for interpretation and action.[5] None of the outlets used here are what we typically think of as "local news"; they don't present a regional perspective. They nevertheless report the news of the day in varied ways: CNN is a twenty-four-hour television news channel (figure 4.2); the *WSJ* is a print and online newspaper with, as it turns out, deep historical ties to computational linguistics (figure 4.3); and BNN is a born-digital platform composed in the language of the web (figure 4.4). These differences shape the way that the news sources appear materially on the pages of this book as data. For each source must travel a distinct computational pipeline to finally arrive on the word list presented here. These pipelines are formed around the contours of each source, but also in relation to the NewsSpeak algorithm: the rules of the word list written in computer code. It is my hope that in seeing the news through the lens of NewsSpeak, readers might come to understand the fourth principle of the book: *data and algorithms are inextricably entangled*.

HOW DOES THE NEWS BECOME DATA?

Unlike purpose-built data, made in the care of scientific or historical collections such as those mentioned in previous chapters, what we might call "news data" is a product of "datafication": a process by which any incident or media might be quantified and translated into a form recognizable by computers.[6] Datafication is one of the primary mechanisms in the assembly of big data, and it has been used to digitally capture information about many of our social behaviors: our everyday movements are monitored by the handheld devices that we carry with us; our spending patterns are registered at the checkout line; and our reading habits are tracked online. In considering the results of datafication, as in other examples explored throughout this book, we must come to terms with their locality. How can the news—gathered from a variety of discordant and locally defined broadcast, print, and born-digital sources—become quantifiable and comparable as data?[7]

Understanding the locality of the news requires venturing into the places where the news is composed and viewed. But comprehending the locality of *news data* means something more: grappling with the algorithms that make the news accessible and manipulable, indeed functional, as data. Communication theorist Tarleton Gillespie writes of algorithms as "encoded procedures for transforming input data into a desired output based on specified calculations."[8] In practical terms, however, making the news into data means rendering it accessible to algorithms, like the ones that I used to produce NewsSpeak. My algorithm is intentionally simple so that readers can understand its mechanics and their relationship to different data sources.

In the *Language of New Media*, Lev Manovich explains algorithms and data as symbiotic. One cannot exist without the other. "Together, data structures and algorithms are two halves of the ontology of the world according to a computer."[9] And yet up until this chapter, I have discussed data independently of their other half. Now I expand my exploration to include the implications of locality for algorithms. In the context of new concerns about their potential biases, algorithms are rapidly becoming the subject of widespread public interest.[10] But we cannot simply think of data and algorithms as independently originating elements of computing. Specific data sets and algorithms, designed to work together, cannot be easily separated and appropriated for other ends.

Following on the 2016 election, understanding the locality of news data in the US has never been so important. First, election coverage now unfolds online at a more rapid pace than ever before, demanding our continual, twenty-four-hour attention. Second, since their failure to call the election, news sources have been under scrutiny from all sides.[11] Federal investigators claim that a foreign power, Russia, might have introduced "fake news" that influenced the vote.[12] Trump, meanwhile, routinely accuses the establishment media of being fake.[13] As a result, many Americans exclaim that they no longer trust any news media. Third, media companies such as Facebook, Twitter, and Google have been asked by the US Congress to reflect on the role that their algorithms play in curating our everyday engagements with the news.[14] That work is long overdue, for the relationship between news, data, algorithms, and reality, as currently configured by such companies, is getting more confused by the day.

My efforts to parse the election coverage as data required access to not only a wide range of news data but also the algorithmic tools for natural language processing (NLP). This is an increasingly prevalent set of algorithms created to recognize and manipulate everyday discourse.[15] NLP tools support all kinds of computer manipulations of language, from counting word frequencies to comparing writing styles. Ultimately, researchers in NLP endeavor to help computers understand complete human utterances, at least well enough to provide useful responses. The successes of NLP research can already be seen in software for language translation and digital personal assistants like Apple's Siri. In order to create user experiences that feel more intuitive, indeed more natural, developers use recent advances in machine learning: algorithmic techniques devised to train computers to recognize expected patterns of speech, using large example data sets rather than explicit rules.[16]

But let us dig beneath the placating surface of such applications in order to understand the material and historical assumptions of NLP. I built NewsSpeak as an exploratory exercise. Indeed, I found that its constitutive elements—data, algorithms, and assumptions about news media and even its connection to reality—had to be continually reworked in the process of creating it.[17] The limits of NewsSpeak, a function of the incongruent textual structures that I discovered in the news sources, then led me to an inquiry into the historical relationship between NLP algorithms and news data.

Too often, algorithms and data are discussed as related but nevertheless discrete elements of computing. I argue that being able to identify and analyze the local conditions that shape news data is crucial for understanding related algorithms as well as the limitations inherent in applying them as generic tools. Looking at the news as data reinforces a recurring theme of this book: we cannot continue to think about data sets as autonomous units to be taken up indiscriminately in practices of computing that have broad implications for society.

CREATING NEWSSPEAK

In spring 2016, I began to compose the NewsSpeak algorithm as a form of inquiry into the Natural Language Toolkit (NLTK), a widely used library for NLP, and one that is representative of a larger set of toolkits.[18] As part of that endeavor, I also had to work with Python, a computer language necessary to make use of NLTK. Along the way, I discovered that the news and NLTK articulate one another in distinct ways dependent on local formations of each. Moreover, I learned that the news has played a crucial role in the history of NLTK, which I will explain in due time. Let us start with the details of the NewsSpeak algorithm. Below is an outline of what the algorithm does, written in pseudocode: a kind of natural language summary of a computer program:[19]

```
GIVEN A CSV OF ROWS IN THE FOLLOWING FORMAT: 'date,' 'outlet,' 'sentence'

CREATE A UNIQUE_WORDS DICTIONARY
CREATE A PREVIOUS_WORDS LIST

FOR EACH ROW IN THE CSV FILE:

   SET DATE TO THE FIRST ELEMENT IN THE ROW
   SET OUTLET TO THE SECOND ELEMENT IN THE ROW
   SET SENTENCE TO THE THIRD ELEMENT IN THE ROW

   SPLIT SENTENCE INTO INDIVIDUAL WORDS IN A WORD_LIST

   FOR EACH WORD IN THE WORD_LIST:

      IF THE WORD IS 'ELECTION' AND NOT THE FIRST WORD IN THE WORD_LIST:

         SET PREVIOUS_WORD TO THE WORD BEFORE 'ELECTION' IN THE WORD_LIST

         IF PREVIOUS_WORD IS NOT ALREADY IN THE LIST OF PREVIOUS_WORDS:
            MAP PREVIOUS_WORD TO AN EMPTY INSTANCE_LIST IN UNIQUE_WORDS
            ADD PREVIOUS_WORD TO THE LIST OF PREVIOUS_WORDS
```

```
        CREATE A DATE_AND_OUTLET DICTIONARY
        MAP 'DATE' TO THE DATE
        MAP 'OUTLET' TO THE OUTLET

        ADD DATE_AND_OUTLET TO THE INSTANCE_LIST MAPPED TO PREVIOUS_WORD

FOR EACH WORD AND INSTANCE_LIST MAPPING IN UNIQUE_WORDS:
  SORT THE DATE_AND_OUTLET'S IN THE INSTANCE_LIST BY DATE IN DESCENDING ORDER
  SET FIRST_USAGE TO THE FIRST DATE_AND_OUTLET IN THE SORTED INSTANCE_LIST

  SET UNIQUE_WORD TO THE WORD
  SET USAGE_DATE TO THE VALUE OF 'DATE' IN FIRST_USAGE
  SET OUTLET TO THE VALUE OF 'OUTLET' IN FIRST_USAGE
  SET USAGE_COUNT TO THE LENGTH OF THE INSTANCE_LIST
  SET DAYS_PRIOR TO 11/8/16 - USAGE_DATE

  WRITE WORD, USAGE_DATE, OUTLET, USAGE_COUNT, DAYS_PRIOR TO A CSV FILE
```

The purpose of the NewsSpeak algorithm is threefold: to introduce readers to the basic mechanics of NLTK using data from a three-part corpus (CNN, *WSJ*, and BNN) that they are likely to be familiar with; to reveal how news data can be understood locally in relationship to the conditions of their creation and the site of their use (for instance, to create an output that fits on these printed pages); and to show how a data set, algorithm, and implicit representation of reality (that is, what the data can be used to argue) must be constructed in tandem.

My work on NewsSpeak began with sample data exported from the NewsScape archive. As an initial source, I choose to work with CNN coverage because of its accessibility via NewsScape along with its breadth and prominence within the 2016 election cycle. There are several means of accessing CNN data through NewsScape. I made use of the basic search interface that allows keyword queries filtered by organization, date, and other variables. Such search queries in NewsScape are applied to existing closed-captioned text for the news, created in the course of each broadcast.

From the outset of this work, I found the news snippets based on closed captioning to be a compelling way of excerpting the news, though with significant limitations: the original video and audio is missing, and along with them the tone and imagery that add layers of context, particularly useful in understanding the never-ending drama that characterizes twenty-four-hour television news.[20]

Yet even putting aside the missing video and audio, closed-captioned text is shaped by its own local conditions related to both the broadcast and transcription processes. Transcriptions today are created through a human-machine process, which relies on voice recognition software, but not in the way that you might think. Straight audio from a broadcast is too complex for current voice-to-text software to translate directly. Instead, human transcribers act as intermediaries, repeating (almost) everything said on-screen into a computer-connected microphone. They must enunciate in just the right way, and use command words to add punctuation as well as macros: verbal shortcuts, programmed for anticipated words and phrases that are difficult or time consuming to verbally encode.[21]

Frequently there are errors. Words are left out or mispronounced, and in turn misspelled or substituted by similar-sounding words that the system has been trained to expect. Sometimes the macros are the source of errors.[22] This work can be enhanced if transcriptionists have access to the scripts created ahead of time for newscasters. But the most urgent "breaking news" reported on the fly or from the field does not follow a script. As a result of all these contingencies, the corpus of closed-captioned text that NewsScape uses to search and share broadcasts is inherently uneven.

Early on in my examination of NLP, while using NLTK functions to explore coverage from NewsScape, I became aware of another subtle though important material way in which the natural language algorithms process news. When the techniques of NLP are applied to the news, time is flattened out. I spoke about the strange temporality of the news with Sergio Goldenberg, the director of technical product management for TV and mobile applications at CNN. He explains that we are meant to encounter the news only in the present. Every news story is broadcast, printed, or posted with the understanding that it is self-contained and ephemeral. Newsmakers cannot assume that audiences will have read or heard yesterday's news or that they will return for a follow-up story tomorrow.[23]

There is an old saying, notes Goldenberg, that "today's news is tomorrow's fish and chip papers," which both acknowledges the temporality of newspapers and puts them in perspective. Likewise, on television and radio, news happens in the moment. Archives of the news, like NewsScape, change the temporality of the news. They rely on algorithms to make news data part of the retrievable record of what has been said by journalists and their subjects across time.

Thus what originates as a sequence of statements unfolding in real time is stored by NewsScape as one contiguous list of words. The past and present now share the same frame. Only by making the news into storable, searchable, retrievable data, through the techniques of NLP, is the temporality of the news redefined as something that accretes.[24] This early lesson—that the datafication of the news challenges our temporal expectations (it's no longer just a representation of what's new)—stayed with me throughout the NewsSpeak exercise. Building on this conception of the news as something additive, my work with NewsScape, which up until this point had been broadly exploratory, started to converge on a question: Can an algorithm reveal what's new in the news?

As I worked, my data set shifted and expanded to encompass different sources. I experimented with subsets of CNN, such as individual news programs as well as different query terms and time periods. In parallel, I had to create new algorithms to augment or supplement those offered by NLTK. This was a messy process with many false starts, wrong turns, and abandoned efforts. Suffice it to say that my algorithms had to stay in sync with the changing data that I was working with. For example, before settling on the search term *election*, my earliest implementations of NewsSpeak focused on the names of presidential candidates or issues related to the election. Furthermore, I made a deliberate choice to focus on a single term (or "token" in the language of NLP). I did this for the simplicity of the algorithm, leaving aside other related terms like *race* or *vote*. I also had to decide how much data to take as input for the algorithm. As I explored various parameters for the data set, I was forced to think about my "what's new?" question in material terms. For instance, the number of days included and the page space in this book were formative variables.

Meanwhile, I was discovering just how different various news sources were in terms of their structure. Rather than seeing the limits of NLTK as a challenge, I viewed them as an opportunity to think about how close relations between data and algorithms might be examined. As mentioned previously, I added two other outlets to my investigation, from outside the confines of NewsScape: the *WSJ* and BNN.

The *WSJ* is a daily newspaper, first published in 1889 to handle business and financial news.[25] I attained digital *WSJ* coverage of the election from ProQuest, a service that requires a paid subscription. I choose the *WSJ* partially because of its historical ties to NLTK, which I will revisit later in the chapter, but also as a prototypical example of print news.

BNN has only recently gained widespread attention, in large part because of its role in the election. It is a favored news source for Trump, often inspiring his widely followed comments on Twitter. Trump's onetime chief strategist, Steve Bannon, long served as the head of the online news organization. BNN is important for another reason, though. It is an illustration of a born-digital news source—one that is structured by tags and links, as opposed to newsprint pages or broadcast times.

Taken together, these three sources certainly do not tell us everything about how discourse in the news might be made into and managed as data. But they offer a comparative understanding of how the news, created in multiple formats (each with their own constraints), becomes data. Developing an algorithm to bring together these discordant sources required a number of nested decisions about what to include, how to focus, and what to leave out for clarity. After all, NewsSpeak is not meant to be innovative, but rather instructive. If it strikes some technical readers as a routine use of NLTK, that is intentional.

Before I consider the implications of this work, let me summarize. The current version of NewsSpeak is assembled using code written in Python. It takes as input news from three sources, beginning on September 16, 2015.[26] Each of these news sources is processed through the same algorithm, which searches for the string *election* in the selected coverage (although this is done in different ways for different sources). The algorithm identifies the string directly before *election*, filtering out a variety of conditions (i.e., if the string is a "."). These exceptions are sometimes known as "stop words." If the string is new—if it has not been seen before by the algorithm—it is added to a collection of strings that are kept in the order in which they first appeared. If the string is not new, it is added to an internal tally, recording the number of times that string has been used. The collection is presented on the pages of this book in graphical form, as a list of terms that modify *election*, in order of use leading up to the day of the vote on November 8, 2016. JavaScript was used to make the resulting text list visible in graphical form, adding color for each outlet and appending the number of times that the word was used in each.

Observations from NewsSpeak

The NewsSpeak algorithm shows the news as NLP sees it: as a list of words. The most apparent observation from the list is the dominance of CNN in relation to the *WSJ* and BNN. But this should not come as a surprise to anyone who knows these outlets and the varied frequencies at which they do their reporting. CNN broadcasts news continuously, while the *WSJ* and BNN are more periodic in their output. If I were performing a legitimate discourse analysis of the news, I would probably normalize these numbers—a process introduced in chapter 3, helpful for comparing different sources on an equal basis. Yet that is not my intention here. Rather, I want to reveal what routine practices, such as normalization, conceal.

Other reflections from a high-level, quantitative view are of limited value too. The most commonly used modifiers to the term *election* are not informative: *the* (mentioned 8,832 times on CNN, 62 on the *WSJ*, and 345 on BNN), *this* (6,280 CNN, 24 *WSJ*, and 85 BNN), and *general* (6,248 CNN, 9 *WSJ*, and 202 BNN). One can make some interesting observations, however, by reading a little more closely. For instance, CNN tends to do the most editorializing. It uses the highest number of overtly negative and positive modifiers to the term *election*. The positive terms *exciting* (12 CNN), *happy* (8 CNN), and more ambivalently *clear* (40 CNN) are all used by CNN exclusively. *Healing* (12 CNN) is perhaps the most noteworthy positive modifier in that it was used twelve times in the two days leading up to the election. Meanwhile, CNN also led in the use of negative terms. *Rigged* (522 CNN, 3 *WSJ*, and 2 BNN) was the most commonly used negative modifier, by an order of magnitude. It was also one of few terms used by all outlets. But CNN was alone in its use of many other negative terms, such as *stolen* (21 CNN), *nasty* (16 CNN), *crazy* (19 CNN), *divided* (18 CNN), *ugly* (11 CNN only), and *sad* (11 CNN), just to name the most commonly used illustrations. The *WSJ* didn't seem to use *election* much at all. When it did employ the term, its preceding words, such as *municipal* (3 *WSJ*), are descriptive not editorial. Meanwhile, BNN was the first to bring up many past dates, suggesting that its narratives about the election might have a historical arc.

Such observations are of limited value, however, because of the constraints of the algorithm, which reads the election coverage selectively. Typically, this kind of word frequency counting algorithm would be the first step in a lengthier program, designed to draw quantitative conclusions about these news outlets, based on the words that they use. For example, NewsSpeak could be linked to sentiment analysis, a technique for determining whether a corpus is generally positive or negative. Such an approach could expand our understanding of the differences among these three outlets or suggest how the general sentiment of the news changed over time as the election day approached.[27] Alternatively, we could assess the partisanship of the news outlets by comparing the words in this list to known instances of partisan discourse. But I will not follow either of those models, or unpack the equally contingent algorithms that they necessitate. For as I have already indicated, my purpose here is not to directly analyze the news.

NewsSpeak is meant to create an experience of news data on these pages that we can reflect on in material terms: colored and composed for the constraints of book printing, with special relevance for audiences that followed the election as it unfolded in real time. For people who lived through the election coverage, the list acts like an index, prompting them to revisit the election in a new form, and retrace its discourse over the months leading up to the final vote. This experience might be different for each reader. After all, NewsSpeak is not meant to reveal a single pattern or conclusion. Instead, it offers a reframing of the news as data.[28] It allows us to experience the election in a novel way, enabled by a systematic computational filtering of recorded election coverage. When understood as such, the NewsSpeak algorithm reveals itself as an aesthetic encounter with the news, more akin to a poem or performance than a political analysis. It is meant to raise questions about what words do in the news, what they don't or can't do on their own, and how they might be reorganized for alternative reading experiences.[29]

Data Artifacts in NewsSpeak

What can this list of words tell us about the ways in which data and algorithms are entangled? Below I address a handful of localized incidents—referred to as *data artifacts* in previous chapters—that are not immediately apparent from the list, yet should inform our reading of it. These data artifacts are symptomatic of broader challenges in NLP that ongoing research in computer science is trying to address or circumvent.

First, the list obscures the high number of transcription errors in live news reports. Take row 25, for example. The term *welcome* does not seem out of place. But consider that it comes from this sentence:

> I learned that we have a lot of talent in the republican party and I think we're going to do very welcome election day. (CNN, September 17, 2016)

Reading the full sentence, it is easy to see that it should say “very well, come election day.” This is likely a mistake caused by rapid, on-the-fly transcription. I spoke with a professional transcriptionist, who explained that broadcast news presents special problems, illuminated in artifacts like this one.[30] As mentioned earlier, broadcasts must be transcribed in real time, commonly without a script, which is likely to lead to such errors. Often these errors are linked to the use of new words or phrases, presenting a special issue for my algorithm as it is attempting to portray the first uses of words to describe the election. New words or phrases may be mistyped the first time simply because transcriptionists are not prepared to hear them. Moreover, many standard phrases are programmed in, which makes them easier to use. Of course, these are the first uses in NewsSpeak, during a certain time period and related to the election, so they are not new words in the larger sense.

A different set of problems accompanies work with text created to be read online. Take, say, the string below from BNN, identified by NLTK as a full sentence. In this excerpt, a period (preceding and not present in the string itself) is interpreted by NLTK as an indicator of the end of the previous sentence. In actuality, it is part of a uniform resource locator, an example of nonnatural computer language or, in this case, HTML.

> com election-tracking site. (BNN, September 20, 2016)

The list of terms also doesn’t help us recognize whether a term is being used literally or as a metaphor. Consider the use of the term *school* below. In this instance, the commentators are referring to the election at hand, but only obliquely. For they are comparing it to a high school election.

> So getting small, getting—you know, making it look like it’s just a high-school election and name calling and all of those things, that can actually take you out of the campaign tonight. Stay big. I think that’s one of the things we want to see. (CNN, September 15, 2016)

The hyphen used above is not picked up by NLTK (a common error), so we only see *school* on the list. Another problem involves identifying complex nested quotes, and who ultimately is being quoted. In an entry to the list from a CNN broadcast, *weird* refers to a term that Trump used, not commentary from a journalist:

> Stuff with jeb it gets covered because trump is the candidate that the media is more obsessed with than anyone else. >> i love that he says this is a weird election as if he’s removed has nothing to do with the fact that this has been a weird election. >> yeah thanks to trump it’s weird right? >> it is a very weird. (CNN, November 10, 2015)

As in the case above, journalists are frequently analyzing the speech of others, complicating efforts to assign attribution. Finally, there is the problem of identifying scope. Are journalists talking about the 2016 US presidential election or another one? The terms *endless* and *cliff-hanger* are about the 2000 election, when George W. Bush defeated Al Gore, and *supercompetitive* refers to the 2008 elections. Also, I enjoyed discovering the following humorous example:

> SIGN UP FOR OUR NEWSLETTERA man disguised as the Star Wars character Chewbacca was arrested in Ukraine for breaking election laws after he drove â€œDarth Vaderâ€ the Internet Party candidate for mayor of Odessa to vote in local elections. (BNN, October 26, 2015)

Because of these artifacts and others, many forays into the analysis of news data collections use methods other than NLP. The Internet Archive, a nonprofit library of internet sites, has used face recognition in order to track reoccurring news coverage of individuals.[31] Media Cloud, an "online platform for media analysis," has made effective use of network visualization of the news using the identifiers of news sources alone.[32] Another project, titled "Page X," adapts a traditional form of news analysis—how much "real estate" on a page does a story get from different outlets—for born-digital news.[33]

Using NLP to effectively analyze the news at a large scale is not that practical today. In fact, it is an unresolved research problem, but an active one. For instance, the NLP community is currently using open competitions to synchronize efforts to improve automatic news comprehension. One recent example targets the identification of fake news—a problem identified by the 2016 election.

Were US flags "banned from display at the 2016 Democratic National Convention"?[34] Did Trump once exclaim that Republicans are the "Dumbest Group of Voters"? Was a pedophile ring "operating out of a Clinton-linked pizzeria called Comet Ping Pong"? None of these headlines are true. But such illustrations of fake news, or simply propaganda, might have influenced the 2016 election by reaching millions of users on Facebook and other social media sites.[35]

The NLP Fake News Challenge proposes that we use algorithms to identify and dethorn fake news stories before they can do more harm. Although this challenge represents the highest ambitions of the NLP community, it is articulated as a problem of aiding human analysts rather than replacing them:

> The goal of the Fake News Challenge is to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem. We believe that these AI technologies hold promise for significantly automating parts of the procedure human fact checkers use today to determine if a story is real or a hoax.[36]

Indeed, today's algorithms are not actually autonomous; they must work in tandem with human users and preexisting data sources. Developing NewsSpeak helped me come to terms with some of the resistances involved in bringing together unrelated data and algorithms. I had to wrangle the data and algorithm in tandem. As I modified the data set—by changing the sources, scope, or search terms—the associated algorithms had to be rewritten or repaired accordingly. Data, by virtue of their representation in linguistic form, cannot simply be plugged into existing NLP toolkits.

This kind of necessary tinkering, with both data and the algorithms, is a mundane yet necessary part of data work. By many accounts, practitioners spend the majority of their time "cleaning" the data or "debugging" algorithms. But such language, which suggests that any problems in either data or algorithms come from the outside—from errant data dirt or computer bugs, which have infiltrated an otherwise-pristine system—obscures the complex local conditions in which both data and, by extension, algorithms are made.

Rather, local sources of the news and NLP algorithms vary in their assumptions about what constitutes "natural language." Those assumptions are best understood by examining the history of engagements between algorithms for NLP and various data sources. In their early development, the algorithms that make up the the NLTK library and in fact all NLP toolkits were trained on particular formulations of natural language.

HISTORICALLY NATURAL LANGUAGE

The history of NLP offers a striking example of how algorithms develop in conversation with research goals as well as changing ideas about what counts as "good" data. Algorithms and data shape one another, not only in projects like NewsSpeak, but over the course of long-term efforts in research and development. Up until 2001, the evolution of NLP could be explained in four distinct phases, according to Karen Spärck Jones, one of the most prominent researchers in the area and a former president of the Association for Computational Linguistics.[37] Each phase was grounded in the local conditions of its time.

The first phase ran from the 1940s to the 1960s and focused on machine translation. Although work on translation was international from the start, funding in the United States concentrated on Russian-English translation and was contingent on an encompassing Cold War political setting, which provided the motivation and funding from the military. Due to their practical defense-related needs, translation algorithms were simple at first, revolving around on word- or sentence-level discourse.

Work during the second phase shifted to pick up momentum from early robotics work in the late 1960s and 1970s. During this time, NLP was framed around the problem of manipulating knowledge representations. In a project called SHRDLU, a simulated robot designed to manipulate blocks in a computer model could carry out basic commands, such as "pick up the red pyramid." The project, led by artificial intelligence pioneer Terry Winograd, demonstrated that a computer could understand natural language, albeit in a highly restricted and quantifiable domain: inside the curated virtual space of a computer. Any attempt to change the setting or scale of the activity resulted in dramatically reduced effectiveness.[38]

The third phase, from the late 1970s to late 1980s, was a time of significant progress for grammar theory. A focus on semantics led researchers to investigate how meaning inevitably referred to the context of discourse and an encompassing "world model." Most notably, this era produced some of the first practical and more widely accessible toolkits, such as the Alvey Natural Language Processing Tools.[39]

Only in the fourth phase, beginning in the early 1990s, were the tools developed in the previous eras turned toward the analysis of large corpora of text. Finally, the news became a prominent source for the development of NLP algorithms. Early on, processing the news centered on *information extraction*, which can be defined as identifying predefined classes of terms within existing texts, ranging from syntactic entities, such as parts of speech, to semantic entities, such as subjects and objects of action.[40] As in previous eras, the most effective systems were locally constrained: they could only function in narrow, predefined settings. A veteran researcher in computational linguistics, Charles Fillmore, explains an early system designed to read the daily news:

> I witnessed work on information retrieval in the form of a system that automatically collected information from newspaper accounts of traffic accidents. My impression was that the system was given texts that were known to be about traffic accidents and it was already provided with a checklist of information to look for, based ultimately on the style sheets used by reporters working on traffic accident assignments, or, really ultimately, on the reporting traditions of the local police departments.[41]

These limitations were widely acknowledged within the NLP community. Jones notes that the most effective systems at the time of her writing were those with the simplest tasks and most tightly constrained domains. Jones's story ends in 2001, the same year that NLTK was created for a computational linguistics course in the University of Pennsylvania's Department of Computer and Information Science. Today NLTK is freely available online. It has become one of the most popular computational toolkits for working with language in a range of applications including "predictive text and handwriting recognition," commonly found in messaging applications, and "web search engines that give access to information locked up in unstructured text."[42]

Like other programming toolkits, NLTK is not a single algorithm but instead a library: a collection of prewritten code that can be invoked by a programmer in order to create their own software. Libraries contain multiple algorithms. NLTK is based on statistical models developed in order to address the limitations of brute force algorithms used in earlier eras. This probabilistic approach, informed by grammatical theory, has given rise to systems that can perform work that even native speakers sometimes find tricky, like identifying parts of speech.[43] NLTK, however, is limited in its handling of language. It is a good illustration of how all algorithms are trained or at the very least tested on local, sample data sets, which fundamentally shape their normative assumptions: what the algorithms should and can be used for.

As such, libraries for NLP such as NLTK are not neutral or general. They do not offer a complete and unbiased comprehension of language use. Computer scientist Joseph Weizenbaum once wrote, and I believe it still holds, that there is no such thing as a general language processor, not even a human one. He invented an early chat bot, named ELIZA, programmed to parody the role of a psychotherapist. Weizenbaum was deeply disturbed by what he saw as unearned reactions to the algorithms that drive ELIZA: "A widespread, and to me surprising, reaction to the ELIZA program was the spread of a belief that it demonstrated a general solution to the problem of computer understanding of natural language. ... [E]ven people are not embodiments of any such general solution."[44]

But NLP tools are also not "inhuman" or "alien," as they are sometimes characterized.[45] Rather, such algorithms are themselves local, for they are created and applied within particular historical and material conditions. For example, data based on the *WSJ* provides a basis for NLTK.[46] As such, the toolkit relies on the insufficiently acknowledged labor of journalists and editors whose English language knowledge has now been subsumed by the algorithms. Furthermore, countless person-hours were contributed by highly trained linguists who annotated the original corpus, converting it into a form that existing machines can process. We might say that NLP is actually a type of human reading—one that is mediated by an increasingly sophisticated system, which masks the work of its many individual contributors.

Let me explain how this particular set of algorithms came to be entangled with an oddly specific data setting, defined by the *WSJ* and its human creators. NLTK relies on a kind of machine learning, trained using the Penn Treebank, a corpus of over 4.5 million words in American English first created in 1989, and painstakingly annotated for both syntactic and semantic structures by expert linguists.[47] A treebank attempts to represent the hierarchical treelike structure of language. It is a "bank of linguistic trees."[48] In the Penn Treebank's most recent manifestation, the primary component of this corpus is a million words from editions of the *WSJ* published in the late 1980s.[49] Tagging this corpus consists of identifying distinct grammatical behaviors within it and ideally coding each word with a single behavior (that is, *VB* for infinitive or imperative verbs, and *NN* for singular common nouns). Authors of the NLTK book, a widely used resource for learning NLP, explain that "the training process involves inspecting the tag of each word and storing the most likely tag for any word in a dictionary."[50] NLTK has a variety of automatic tagging algorithms, each of which accounts for the linguistic context of a word in different ways. Only by determining the proper context, an ongoing research project in computational linguistics, can algorithms determine how a word should be categorized.

As a means of further understanding NLTK, it might be helpful to revisit some of the problems introduced in the example algorithm, NewsSpeak: mistakes in transcription, misinterpretation of punctuation, and a general lack of context. But NLTK researchers are perhaps most challenged by ambiguities or double meanings, which are an inherent part of language. They often trip up human readers, so how can we expect computers to do any better? Jokes frequently play on such ambiguities. The authors of the Penn Treebank demonstrate this using a humorous, if somewhat dated, quip from the actress Katherine Hepburn, presumably about Cary Grant, with whom she made several films: "Grant can be outspoken—but not by anyone I know."[51] Here the term *outspoken* will likely be interpreted by an NLP "part of speech" tagger as an adjective, even though, in the context of the joke, it is meant to be recognized as the past participle of the verb *outspeak*. Hopefully some readers will get this and be amused. Such jokes, however, can easily miss their mark with an unprepared audience. Many younger readers might wonder, who is Cary Grant?

The Penn Treebank contains thirty-six parts of speech tags, and twelve tags for punctuation and other symbols—a dramatic reduction and streamlining of previous models. The annotation was originally performed through a combination of machine tagging and human correction, which was found to be considerably faster than human annotation alone. Nevertheless, the cost of annotation, which requires a high level of expertise, is prohibitive, and one of the major reasons that the Penn Treebank has not be updated often or replicated.[52]

When they wrote "rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials," NLTK's creators did not necessarily intend *naturally occurring* to mean a set of completely generalizable examples of American English.[53] Researchers are much too nuanced for that.

Early on the Treebank processed a number of other specific corpora, including WBUR transcripts (an NPR-affiliate news radio station based in Boston), IBM computer manuals, and Library of America texts, the last of which consist of small passages from US-based authors like Mark Twain, Herman Melville, W. E. B. DuBois, and Ralph Waldo Emerson. All these choices reveal much about the expectations among NLTK's designers as to what are model forms of language use: examples considered canonical or commonplace (all by men, most of them white). It was only later that they turned to the *WSJ* as an exemplar of everyday language use.[54]

Readers who are not part of the computational linguistics community might need to be attuned to some of the compromises implicit in the efforts to apply computers to the understanding of natural language. I do not suggest that NLP researchers are unaware of these concessions. Only that they shape what the rest of us must come to understand as a localized model of natural language, informed by circumstances that should not be overlooked. First, NLP accepts more or less that language can be disassembled into a list of words to be analyzed and manipulated as discrete, self-contained units of expression.[55] As such, researchers must accept a view of language as principally informational as opposed to poetic or even perhaps intentionally opaque, as in the case of lingo, jargon, and slang. Second, researchers have confined themselves to a certain type of grammatically precise language at a level only possible in print or scripted speech. NLP works on the supposition that language is textbook correct, and has been professionally edited for spelling and grammar. Third, language is expected to create meaning independently of other media, such as images, which are routinely part of the news (even print news like the *WSJ*). Finally, there is a regrettable assumption in the continued use of historical NLP resources like the Penn Treebank that the rules of language use do not change. Yet tags created in 1989 to describe articles from the *WSJ* cannot be effective indefinitely. Linguists know that even punctuation has not always been part of natural language use and continues to evolve, as the emergence of emoji use reveals.[56]

Many of these assumptions, necessary for NLP to produce results, do not hold in the majority of language use. Try running with these assumptions on Twitter, where rapidly changing forms of language use, such as the introduction of hashtags or the intentional appropriation of humorous errors, make NLP use difficult.[57] Nevertheless, they have been adopted, often consciously, as a part of the ongoing struggle to get computers to recognize our communications.

Current approaches to NLP are not wrong, of course. They are simply based on limited models of language use as well as situated human labor, much of which unfortunately goes unacknowledged. It is important to concede that NLP algorithms are shaped by the locality of data, for they hold significant implications for the future trajectory of public discourses in the news. Indeed, some news makers are already bending their products into forms that are more easily recognizable by current algorithms.

DATA, ALGORITHMS, AND THE REALITIES THEY SUPPORT

Only recently have algorithms been appreciated as subjects of broad relevance to everyday life. They introduce new and opaque procedures to important domains of public understanding and decision making, such as the news, finance, criminal justice, and even love.[58] Algorithms have become the matchmakers of our time: they illuminate connections across data from diverse sources. But algorithms are not just technical procedures. Social studies of algorithms have revealed them to be complex sociotechnical artifacts: fragile, multiple, and situated in ad hoc practices of computational work.[59] Algorithms are local, not in small part because they rely on data for their development and testing. I would take that argument one step further: collections of data and algorithms should not be considered as entirely independent components of computation. Indeed, they are entangled with each one another, materially and historically.

Let us return to a question from the beginning of this chapter: How does the news become data? We can now see that processes of datafication are enmeshed in complex local conditions. Searching for a way to characterize how data are made during the all-consuming rush to digitize media—images, videos, music, text, and more—starting in the 1990s, Manovich posited that a new "cultural algorithm" had emerged:[60]

reality -> media -> data -> database

Reality, explains Manovich, is represented through media, which are then digitized as data, and ultimately structured and stored in databases for future use. His prescient concern at the time was that data might eventually overshadow their original sources. For viral images, video, and texts had already spawned innumerable copies online. Invoking Jorge Luis Borges, a favored storyteller of media theorists, Manovich observes that data are a new means by which the map can exceed the territory.[61] Data, he argues, have become a dominant means of sense making, eclipsing direct engagement with reality.

This concern, however, presupposes a stable and singular reality that can be represented by media, and ultimately shoehorned into a database. We might call this a "realist" explanation of how data come to be in the world. When applied to our thinking about the news, the realist explanation suggests that the world is authentically captured by news media first, then later made to conform to the constraints of contemporary data and database structures, such as those that govern how news is shared on contemporary social media platforms. As for the role of algorithms in this explanation, that is left more abstract. Manovich presents algorithms as processes that govern the relationship between reality, media, data, and the database, but that nonetheless exist independently from them.

This may appear to be the case if we look narrowly at the way in which the NewsScape archive ingests the news into its databases. Yet I believe Manovich's realist characterization of data needs some rethinking. For example, it's not difficult to imagine how the arrows of Manovich's algorithm might go in the other direction:

database -> data -> media -> reality

Depending on whether the database is relational, hierarchical, or networked, its structure will shape the kind of data that can be created.[62] Data in turn can be used to create media in the form of visualizations, graphs, charts, and other forms of analysis around which conceptions of reality are ultimately built. As Jean Baudrillard asserts, the map can precede the territory.[63] Conceptions of reality must conform to predefined categories of data in order to be legitimized. Consider how NewsScape prioritizes closed-captioning text from broadcasts; it is easier to translate those elements of reporting into a searchable form. Meanwhile, the video and audio are given less attention. As another illustration, the reader may recall from the last chapter how a library must catalog its new arrivals into a predetermined system of organization, such as those established by the Library of Congress, Dewey decimal classification, or Dublin Core.

This reverse ordering of Manovich's cultural algorithm might be called a "constructivist" explanation.[64] In contrast with Manovich's realist version, it suggests that conceptions of reality are, as the name implies, constructed from explanatory media and based on data that are dependent on an underlying database infrastructure. In this formulation, algorithms are still independent, directing the progress from database to reality from some external position. Yet this explanation does not hold water either. From what is data derived if not reality? The constructivist explanation has no clear ground.

Whether it proceeds forward or backward, by realist or constructivist steps, Manovich's cultural algorithm is meant to explain how data, algorithms, and reality are connected. By calling this algorithm "cultural," Manovich gestures to the fact that a range of processes (not all technical) are at work, helping to create and manipulate data. Rather, I would like to suggest that unpacking the algorithm—understanding it in specific rather than general terms, as I have sought to do here—can help us see how various elements of computation coevolve in a parallel instead of sequentially. The relationships between data and algorithms are not generic. They are bound up in specific conditions, such as the choice made for the Penn Treebank to use the *WSJ* to shape contemporary NLP toolkits.

There is another kind of explanation, commonly employed in social studies of science and technology, exemplified by "actor-network theory" and "ontological politics," that can help us understand how reality gets done locally.[65] Reality is not universally fixed, these explanations assert. Rather, to adapt an assertion made by Bruno Latour, reality simply consists of the set of statements too costly to call into question.[66] Such "network" perspectives would hold that reality is not at the beginning of the algorithm (the input), nor its end (the output). In fact, statements about reality are another element at play in the continual search for an effective arrangement of media, data, databases, and algorithms. A network perspective on datafication assumes that all these elements coevolve:

reality + media + data + databases + algorithms

Significantly, network explanations imply that there are multiple realities, each resonant with an organization of algorithms, data, databases, and media. If we accept this view, we can move beyond the limits of the realist and constructivist explanations. The elements of a network are not discrete, linear steps to be connected by an algorithm.[67] Rather, these elements and the relationships between them are perpetually under construction. What's more, algorithms and even reality are themselves part of the arrangement.

Bringing a network perspective to news data can help us understand datafication with tools that are useful for thinking about the past, present, and even future of the news. This perspective helps us describe a condition that the material and historical inquiries earlier in the chapter support: news media, data, databases, algorithms, and accompanying conceptions of reality are in continual flux, while being in tension with one another too.

We are always encountering localized arrangements of these elements. The contents of news media vary in response to a broader cultural climate, but also in relationship to the network of technologies within which they are composed. Data structures for accessing news online vary as well: some news is delivered as text-based HTML; other news is broadcast as streaming video; and many sites now offer affordances for interaction and commentary. Meanwhile, today's social media, enabled by large scale databases, are increasingly the settings in which news is read, watched, annotated, and shared. Finally, algorithms for NLP and an awareness of their limitations have affected all these other elements: social media platforms use such algorithms to curate news content and target advertising toward readers; data structures include tags and other forms of annotation meant to ease the burden on NLP; and even the way that the news is made is poised to change, to be more structured from investigation through delivery. Someday soon, each news story may be conceived of as an algorithmically manipulable composite of claims and evidence, organized for not only reference but also recomposition and reuse.[68]

There are other, darker implications to these new configurations of media, data, databases, and algorithms. First, without low-cost algorithmically generated advertising that can zero in on particular demographics, respond to those audiences dynamically, and direct them to an immediate point of sale, most small local news outlets have closed. This has left local governments and communities virtually unsupervised; traditionally, local news has played an important watchdog role.[69] Second, the large national and international news outlets that have survived are considering how they might further conform to new expectations for everything to be searchable as data. Third, news-viewing habits are becoming more siloed. In 2019, we have come to accept that social media filter our news into personal feeds, effectively cutting off segments of the population from news that they might find disagreeable.[70]

Lastly, new forms of fake news have emerged. Although news-like propaganda and overzealous editorials have been around for a long time, networks of technologies that support the emergence of news data open up the possibility for more insidious ways to fake the news. In social media feeds, fake news propagates, indistinguishable from less biased, more conventional alternatives. The limitations of current instantiations of NLP that prevent its algorithms from accurately detecting fake news might be overcome someday by incremental technical improvements. But it is more likely that all the elements of news datafication will converge in new ways: news media will retool themselves, data formats will be restructured, databases and platforms for hosting the news will be reinvented, and the boundaries between what is real and what is fake will shift.

As in any arms race, fake news will continue to defy the algorithms devised to detect it. All these changes are happening in parallel but informed by one another. Understanding the locality of datafication means acknowledging all these shifting elements: reality, media, data, databases, and algorithms along with their interrelationships.

CONCLUSION

Data and algorithms are not stand-alone elements of computing, independently applicable anywhere. They are created in close coordination with one another, and with prevailing conceptions of the reality that they seek to represent, analyze, or predict. Moreover, they only function symbiotically, in contingent local conditions that are both materially and historically grounded. The news is a crucial site in which those contingencies might be better understood.

The nascent algorithmic turn in the analysis of news heralds several profound implications. First, old news is no longer trash; it becomes part of the consultable archive of what has been said. News archives and the algorithms that bring them to life can help us hold public figures and processes accountable, and protect against biases or even fake news. Yet the form of the news is not static. Producers of the news must choose whether to lean into or challenge the operation of normative algorithms for NLP. Already some news organizations are actively rethinking the way that news is generated, from robot reporters on the beat in areas like sports and weather, to more structured approaches to journalistic investigation that promise to make news data more accountable and accessible to remixing. In the coming years, all these elements of the news will continue to change and reciprocally inform one another: media, data, databases, algorithms, and even conceptions of what counts as news.

Beyond the roles that they play in the news, algorithms for NLP are pervasive: on social media, in internet searches, and even under the hood of the word processor I am using to write this book. I have barely begun to explore the multiple localities of NLP. Further research is needed to understand how such algorithms and the human labor that they subsume work together with data to shape important domains of public life.

The next chapter will continue to investigate algorithms as part of another computational structure: the interface. Already an element of everyday life, the interface also faces challenges posed by the locality of data. Rather than preserving the originating context of data, many interfaces today seek to recontextualize data, with critical implications for how data are made actionable and for whom. The illustrative case for that chapter is housing data, in many ways the highest-stake case in the book as well as the last one.

5