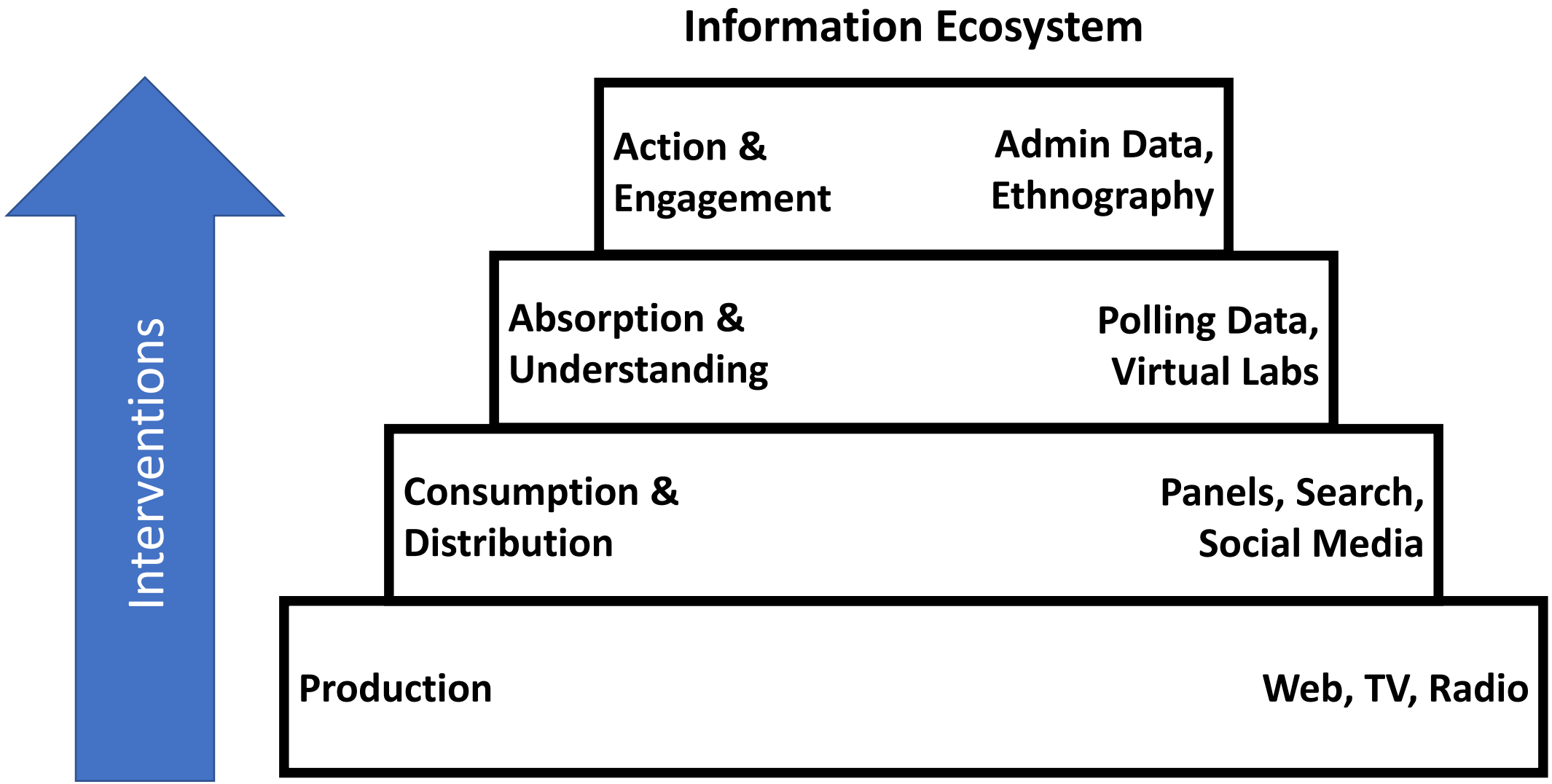
Project Ratio: problem

- News Datasets: idiosyncratic, one-off, and often small in scale
  - Comparisons difficult across different modes of media consumption, different sample populations, and different times
- Data is scattered between private companies and academic research
  - Bringing that data together is both an infrastructural and cooperative challenge
- Research scattered across several disciplines (e.g. economics, marketing, political science, communications, psychology, sociology, computer science, and network science), each with its own set of theoretical frameworks, accepted methodologies, and publishing venues
  - Collating and reconciling results across these disciplinary boundaries is difficult and often leads to contradictory or incoherent conclusions

Project Ratio: solution

- Build a large-scale, shared data infrastructure for studying the production, distribution, consumption, absorption, and impact of news over time and across the entire information ecosystem
- Develop academic-industry partnerships around data and solutions, and communicate these solutions to the public
- Build “many labs” research network to advance basic research and generate actionable insights that are relevant to business and public policy

Project Ratio



Project Ratio

- **Production:** What news is actually being produced?
  - Bias in the selection and framing of articles/events/topics
  - Provenance of articles/events
  - Networks of article/story sharing and updating
- **Consumption:** What news is actually being consumed?
  - How much news do people get
  - Where do they get their news
  - How much of it is mainstream, fake, commentary, hyper-partisan
  - Do people live in filter bubbles
- **Absorption:** What news are people actually absorbing?
- **Action:** How does absorbing news change actions or engagement?

production data

News Scraping (since 8/2014)
(3,000 US and 4,500 European pub.)

- 7,500 publishers/ 500,000 articles (top 300 publishers' home pages every 30 mins)
- Landing pages/placement of each article on the page/font-size
- Recording transcripts of all national TV channel and relevant local markets

Metadata Generation
(events, provenance)

- We partition daily newly published articles into "events"
- Within events, we analyze the provenance of articles:
  - Text from agencies such as AP and Syndicated articles
  - Original content

Research

Web Tools

- Annotated raw data and meta data becomes available for research.
- Aggregated meta data is available through a public website.

production pipeline

Input
Scraped articles
(HTML source codes of article webpages)

Text Extraction
Remove HTML code;
Remove boilerplate text

Article Selection
Exclude articles containing landing page text;
Exclude duplicate articles

TOKYO — Ever since President Trump pulled out of planned talks with Kim Jong-un, the leader of North Korea, officials from both countries have been scrambling to salvage the summit meeting and keep it on track for June 12 in Singapore.

A process that would normally take months is now being compressed into a matter of days, with meetings taking place between North Korean and American officials all over the world. Below, a rundown of who is meeting whom, and what they might be talking about.

Token Selection
Tokenize sentences, stem words, and extract entities;
Remove stop words and rare unigrams or bigrams

unigram	***tf-idf***
summit	18.808
conflict	9.838
president	13.760
denuclearization	15.735
military	4.108
bigram	***tf-idf***
demilitarized zone	17.484
human right	10.724
security concern	14.618
imminent threat	8.513
nuclear weapon	17.232
entity	***tf-idf***
north korea	8.649
kim jong-un	16.779
donald trump	13.883
united states	7.887
mike pompeo	6.994

Text Vectorization
Create daily document frequency dictionaries;
Compute tf-idf vectors

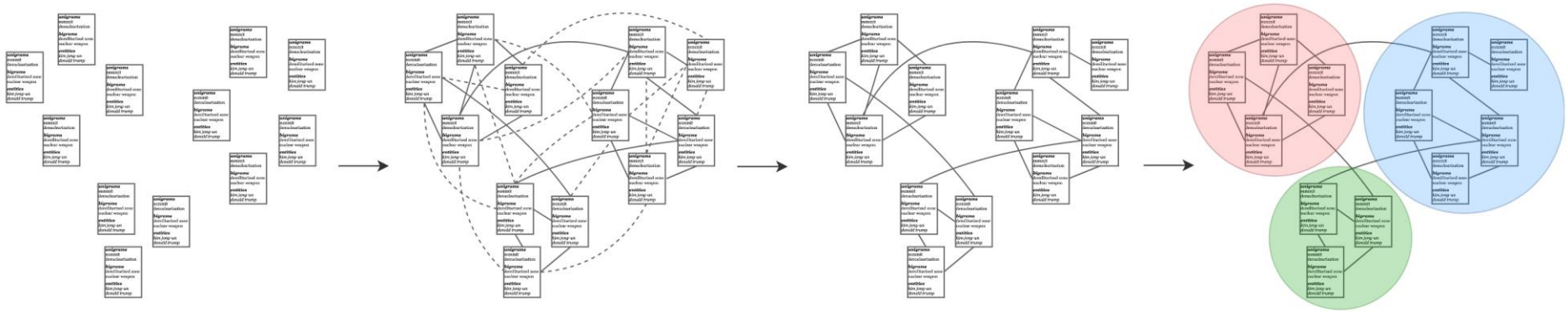
unigrams
summit
denuclearization
president
conflict
military
...

bigrams
demilitarized zone
nuclear weapon
security concern
human right
imminent threat
...

entities
kim jong-un
donald trump
north korea
united states
mike pompeo
...

Output
Article fingerprints
(Top 100 tf-idf features for unigrams, bigrams, and entities)

production pipeline



Input
Daily article fingerprints
(Top 100 tf-idf features for unigrams, bigrams, and entities)

Network Construction
Calculate and normalize similarity between article fingerprints; Average similarity scores for unigrams, bigrams, and entities; Prune edges with lowest weights

Article Clustering
Community detection using the Louvain method

Output
Daily event clusters
(Groupings of articles about the same event based on fingerprints)

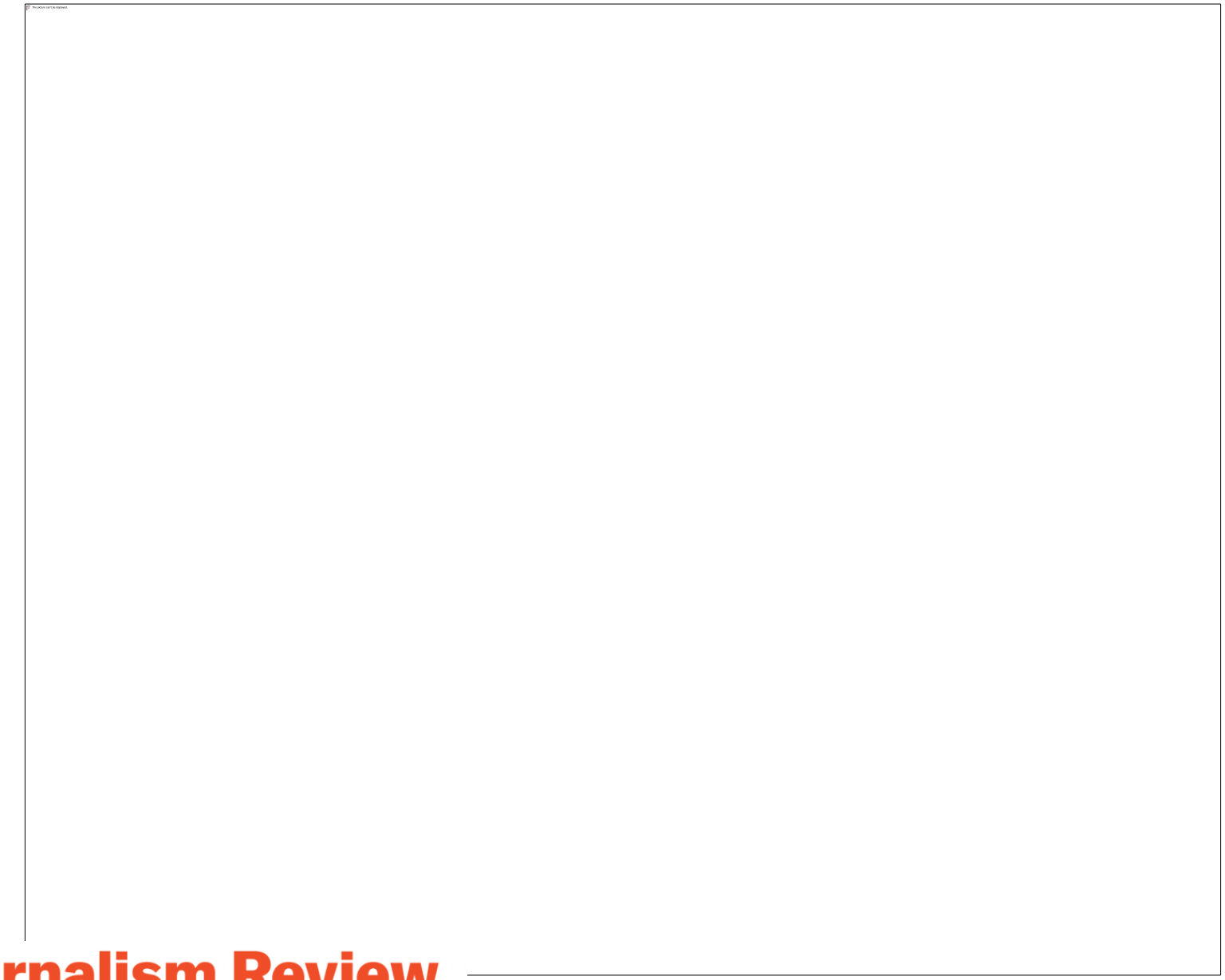
We partition articles into daily events by fingerprinting them and then clustering them.

Events: January 23, 2018

	A	B	C	D	E	F	G	H	I	J	K	L
1	eventID	label1	label2	label3	article sha	words	bigrams					
2	0	'Logan' Oscar Nominations: Film I	Here Are t	Oscars no	3.32%	nomin, os	phantom thread, ladi bird, oscar nomin, eb missouri, film acade					
3	1	Kentucky shooting: Bevin to discu	Kentucky s	Kentucky s	2.64%	school, sh	marshal counti, school shoot, benton ky, fatal school, paducah s					
4	2	Senate Votes Overwhelmingly to	Senate vot	Senate Co	2.05%	senat, den	govern shutdown, white hous, border wall, feb 8, senat democr					
5	3	National Tsunami Center cancels	Alaska ear	Alaska ear	1.79%	tsunami, a	tsunami warn, kodiak island, higher ground, british columbia, cc					
6	4	Sentencing nears end for former	USA Gymr	'Enjoy hell	1.55%	nassar, gy	usa gymnast, larri nassar, matthew dae, michigan state, victim ir					
7	5	Mueller's office spoke with Sessic	Mueller's	Mueller's	1.12%	mueller, s	russia investig, justic depart, attornei gener, session interview, n					
8	6	Tecmo Super Bowl predicts the S	Philadelph	Patriots-Ea	1.01%	bowl, supe	super bowl, eagl fan, bowl lii, tom bradi, nfl footbal, england pat					
9	7	Trump slaps new trade tariffs on	Trump sig	Trump app	0.89%	solar, tarif	wash machin, solar panel, solar cell, solar power, solar industri,					
10	8	UN readies aid as Turkey attacks	Turkey ba	Erdogan S	0.85%	syria, syria	chemic weapon, northern syria, syrian kurdish, chemic attack, le					
11	9	FBI Wray Threatens To Resign If S	Under Crit	Trump Rej	0.83%	fbi, mccab	fbi director, jame comei, andrew mccabe, white hous, act fbi, jin					
12	10	Trump on FBI Agent Strzok's Miss	Sessions L	#DeepStat	0.81%	fbi, text, st	text messag, secret societi, peter strzok, fbi agent, miss text, gut					

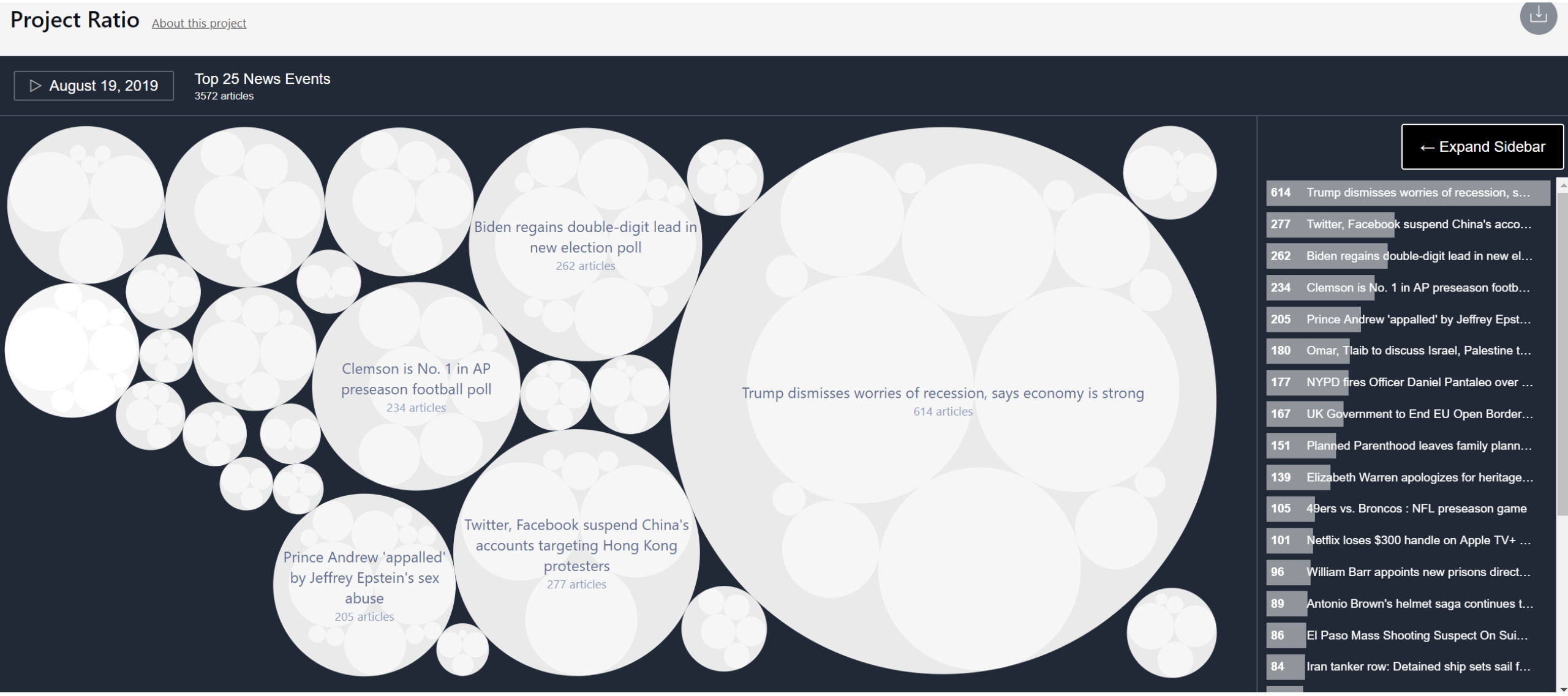
Search through these events to find school shootings.

production of real news



Columbia Journalism Review.

production dashboard



production dashboard

consumption data

Nationally broadcast TV content. Nielsen's nationally representative TV panel (N = 100,000), from which we compute the number of minutes per person per day devoted to watching specific programs.

Local TV news. Subset of the national panel (N = 50,000) sampled from the 25 largest local markets.

Online content (desktop). Nielsen's nationally representative web panel (N = 65,000), which records the complete browser history of its panelists including the specific URLs visited as well as the length of the visit.

Online content (mobile). Comscore aggregated metrics, which track total time spent and total online viewers per month for each website for mobile (including browser and app usage) and desktop respectively by age group.

consumption pipeline

Television news consumption

- Time consuming any of the roughly 400 programs that are classified by Nielsen as “news”
- Upper Bound: Magazine news (e.g. Inside Edition, Dateline), morning shows (e.g. Good Morning America, Today Show), entertainment news (e.g. TMZ, Access Hollywood), and late-night shows (e.g. The Daily Show with Trevor Noah, the Late Show with Stephen Colbert)

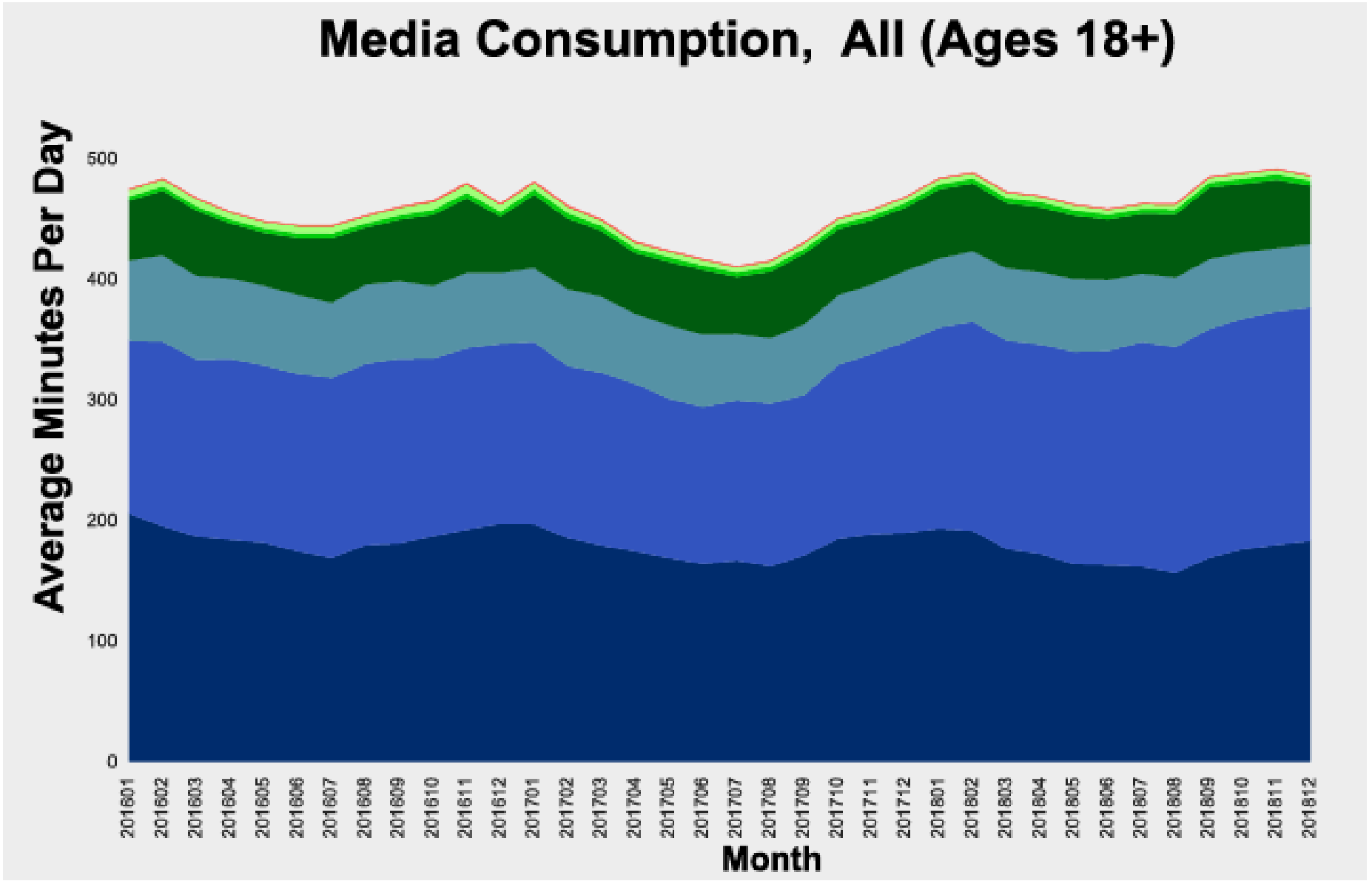
Online news consumption

- Time consuming any article published in more than 800 websites, adapted from (Athey et al 2017): primarily cover “hard” news topics like politics, business, and US and international affairs.
- Fraction of time consuming social media sites (Facebook, Twitter, Reddit), search engines (Google, Bing, and Yahoo) using referral links
- Fraction of time consuming YouTube (classified 10,000 videos internally as “news and politics”)
- 100% of time consuming portals (MSN, Yahoo, AOL)

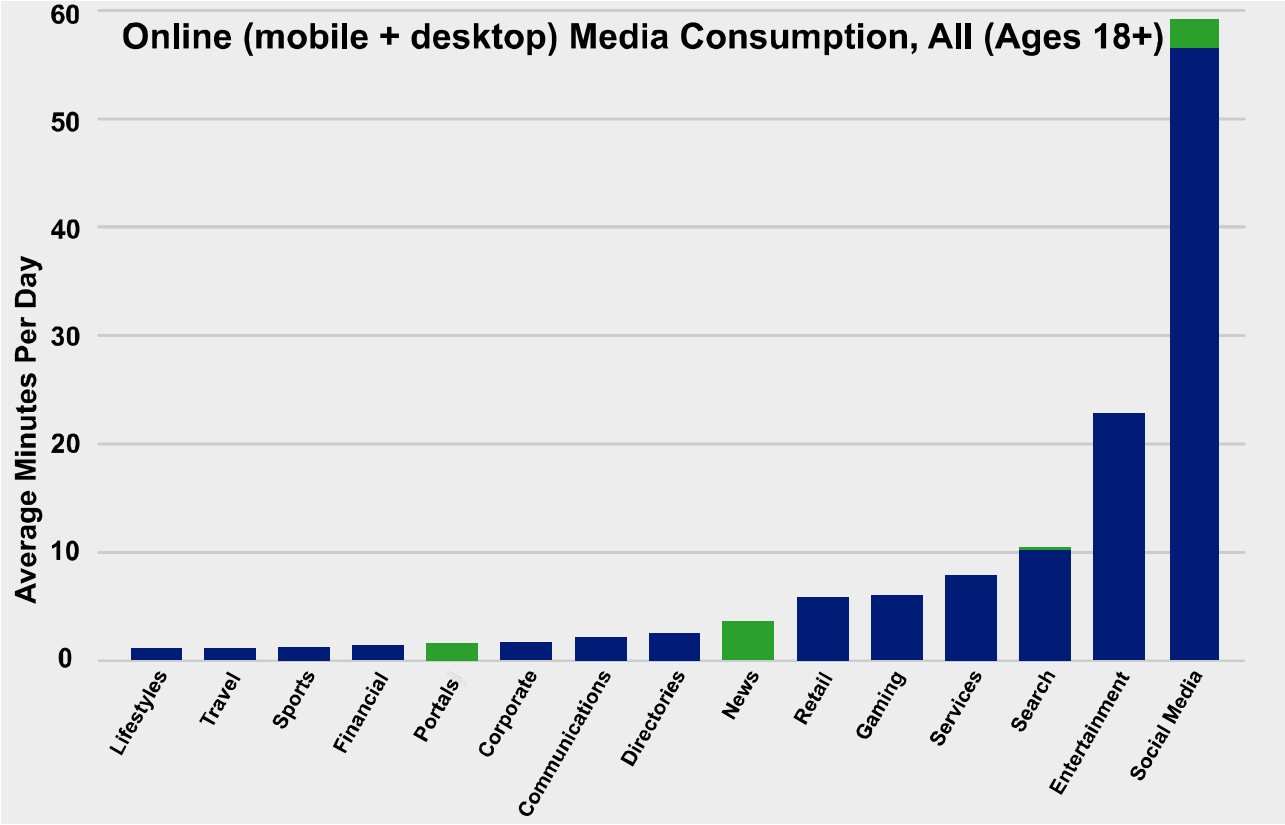
Fake news consumption

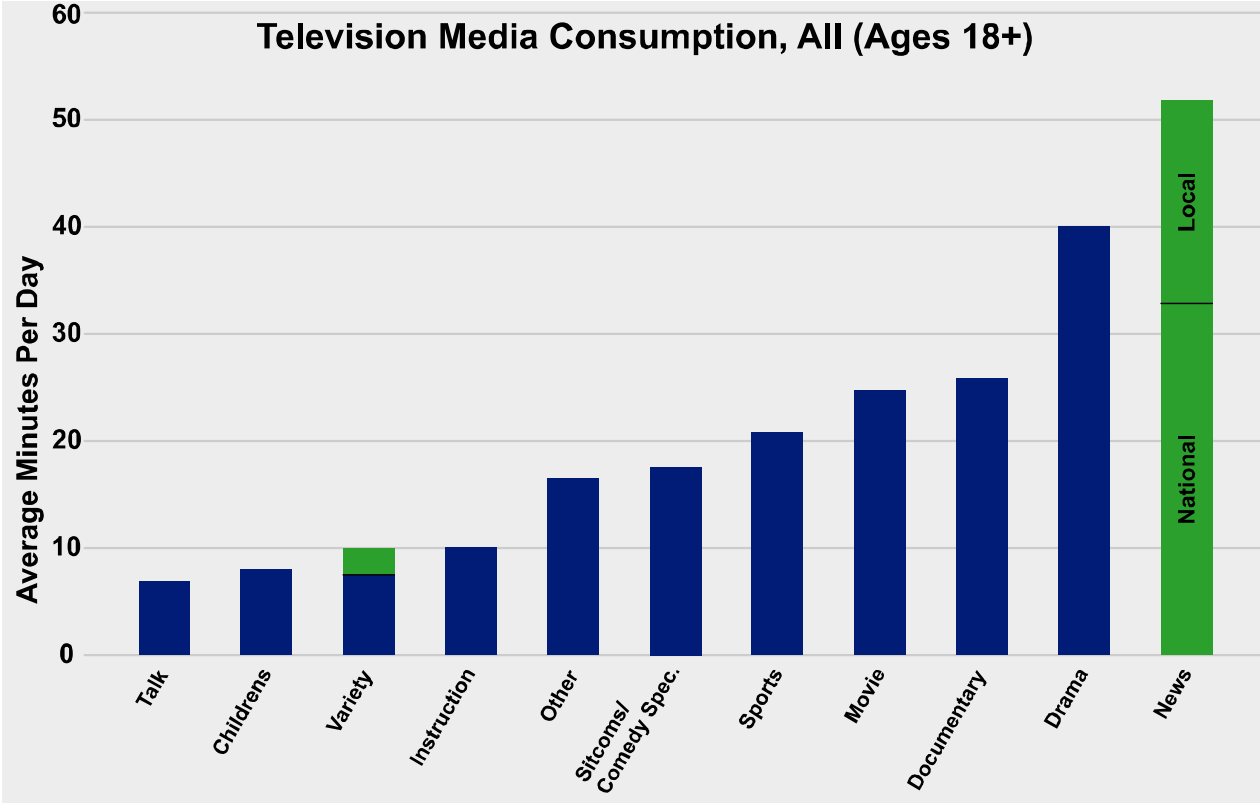
- Time consuming 1 of 98 websites previously identified by researchers (Grinberg et al 2017), professional fact checkers, journalists as: fake, deceptive, low quality, or hyperpartisan news.
- Also attribute some fraction to social media and search as above
- YouTube hand classified

Both news and fake news definitions intended to overestimate real consumption


TV: Not News
TV: News
N/A TV: Fake News
Mobile: Not News
Mobile: News
Mobile: Fake News
Desktop: Not News
Desktop: News
Desktop: Fake News
Media Consumption, All (Ages 18+)
Average Minutes Per Day
500
400
300
200
100
0
201601
201602
201603
201604
201605
201606
201607
201608
201609
201610
201611
201612
201701
201702
201703
201704
201705
201706
201707
201708
201709
201710
201711
201712
201801
201802
201803
201804
201805
201806
201807
201808
201809
201810
201811
201812
Month

News
Not News

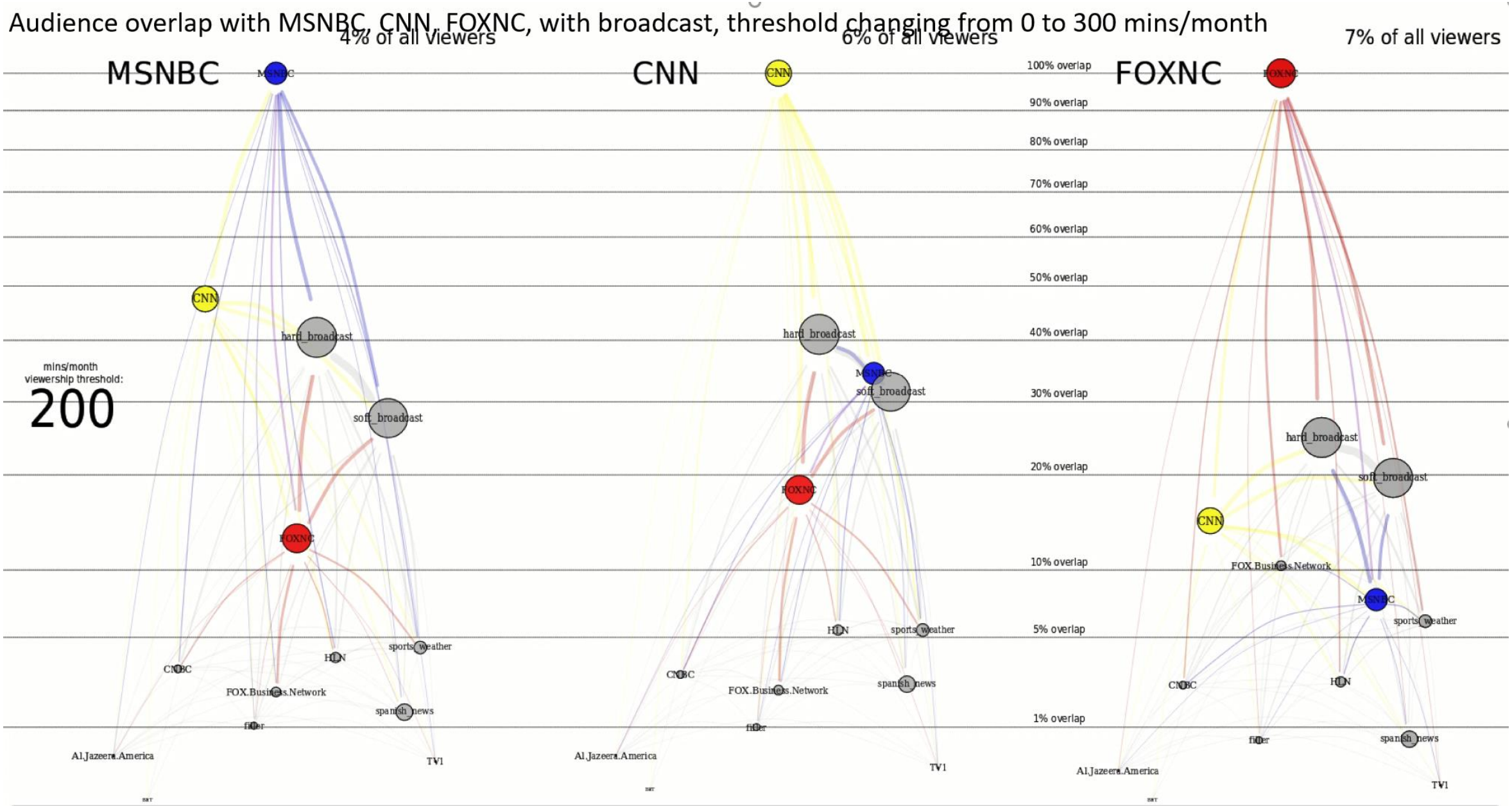




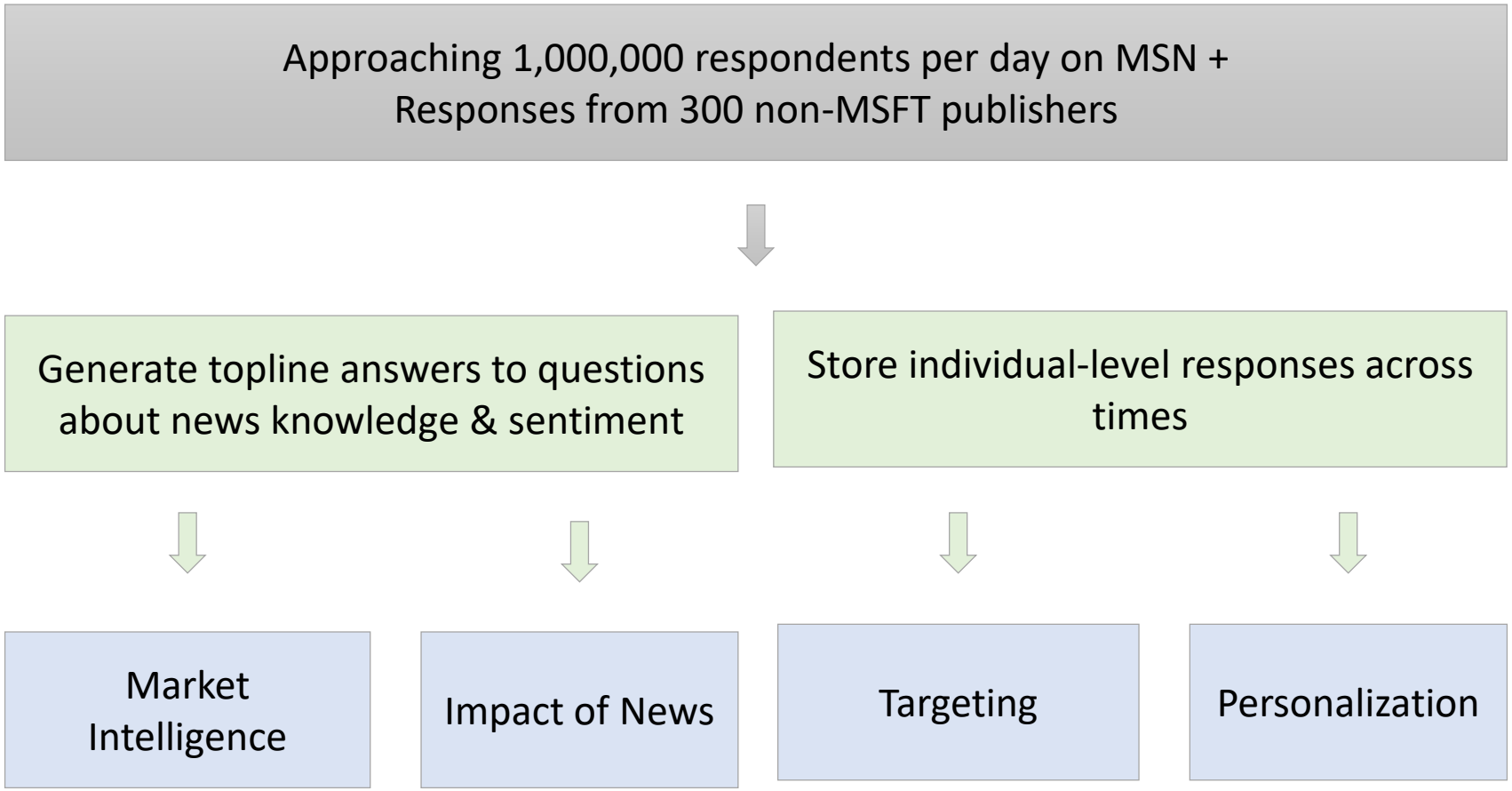
consumption: platforms, social media, search

	MSN Portal	Yahoo! Portal	AOL Portal	Bing Search	Google Search	Yahoo! Search	You-Tube	Face-book	Reddit	Twitter
% of overall online Media Consumption that is on this Platform	0.2%	0.4%	0.1%	0.2%	2.1%	0.1%	9.4%	10.6%	0.2%	0.5%
% of Consumption on Platform that is News	100%	100%	100%	4%	5%	3%	3%	8%	5%	11%
Minutes of News/Day on this Platform	0.52	0.84	0.16	0.02	0.22	0.01	0.60	1.93	0.03	0.12
% of News on Platform that is Swampy	0%	0%	0%	4%	4%	7%	38%	15%	6%	6%
Minutes of Real News/Day on this Platform	0.52	0.84	0.16	0.02	0.21	0.01	0.37	1.63	0.02	0.11

consumption: networks



absorption data



absorption data