Applied Data Analysis (CS401)

Lecture 14
ADA in action
21 Dec 2022

EPFL

Robert West

dlab

Announcements

- Today: last lecture
- No lab session this Friday
- Final project milestone P3 due on Fri 23 Dec 2022, 23:59
- Final exam: Tue 17 Jan 2023, 15:15-18:15
  - Announcement re: exam protocol, room assignment, etc., to be made in early January (on Ed with email notification)

Give us feedback on this lecture here:
https://go.epfl.ch/ada2022-lec14-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Today I will

- present a research paper of mine,
- highlight how everything you've learned in lectures 1–13 comes together in one project.

Paper available at https://doi.org/10.1073/pnas.2106152118

Check for updates

Postmortem memory of public figures in news and social media

Robert West[a,1], Jure Leskovec[b], and Christopher Potts[c]

[a]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [b]Department of Computer Science, Stanford University, Stanford, CA 94305; and [c]Department of Linguistics, Stanford University, Stanford, CA 94305

Edited by Henry L. Roediger III, Washington University in St. Louis, St. Louis, MO, and approved June 24, 2021 (received for review April 1, 2021)

Deceased public figures are often said to live on in collective memory. We quantify this phenomenon by tracking mentions of 2,362 public figures in English-language online news and social media (Twitter) 1 y before and after death. We measure the sharp spike and rapid decay of attention following death and model collective memory as a composition of communicative and cultural memory. Clustering reveals four patterns of postmortem memory, and regression analysis shows that boosts in media attention are largest for premortem popular anglophones who died a young, unnatural death; that long-term boosts are smallest for leaders and largest for artists; and that, while both the news and Twitter are triggered by young and unnatural deaths, the news additionally curates collective memory when old persons or leaders die. Overall, we illuminate the age-old question of who is remembered by society, and the distinct roles of news and social media in collective memory formation.

computational social science | collective memory | news and social media analysis | forgetting

Being remembered after death has been an important concern for humans throughout history (1), and conversely, many cultures have considered *damnatio memoriae*—being purposefully erased from the public's memory—one of the most severe punishments conceivable (2). To reason about the processes by which groups and societies remember and forget, the French philosopher and sociologist Maurice Halbwachs introduced the concept of collective memory in 1925 (3), which has since been a subject of study in numerous disciplines, including anthropology, ethnography, philosophy, history, psychology, and sociology, and which gave rise to the new discipline of memory studies (4). Over the decades, collective memory has moved from being a purely theoretical construct to becoming a practical phenomenon that can be studied empirically (5), e.g., in order to quantify to what extent US presidents are remembered across generations (6) or how World War II is remembered across countries (7).

Whereas oral tradition formed the basis for collective memory in early human history, today the media play a key role in determining what and who is remembered, and how (8–11). Researchers have studied the role of numerous media in constructing the postmortem memory of deceased public figures. A large body of work has investigated the journalistic format of the obituary (12–16), which captures how persons are remembered around the time of their death (14). Taking a more long-term perspective, other work has considered how deceased public figures are remembered in the media over the course of years and decades (17–21). As ever more aspects of life are shifting to the online sphere, the Web is also gaining importance as a global memory place (22), which has led researchers to study, e.g., how social media users (23–27) and Wikipedia editors (28) react to the death of public figures. In the context of social media, the detailed analysis of highly visible individual cases, such as Princess Diana (24), pop star Michael Jackson (25, 26), or race car driver Dale Earnhardt (27), has revealed how people experience and overcome the collective trauma that can ensue following the death of celebrities.

Although such studies of individuals have led to deep insights at a fine level of temporal granularity, they lack breadth by excluding all but some of the very most prominent public figures. What is largely absent from the literature is a general understanding of patterns of postmortem memory in the media that goes beyond single public figures.

To bridge this gap, we draw inspiration from a body of related work that has studied the temporal evolution of collective memory using large-scale datasets—although, unlike our work, not with a focus on the immediate postmortem period of public figures. For instance, van de Rijt et al. (20) tracked thousands of person names in news articles, finding that famous people tend to be covered by the news persistently over decades. In a similar analysis, Cook et al. (19) further showed that the duration of fame had not decreased over the course of the last century. Beyond news corpora, the online encyclopedia Wikipedia has become a prime resource for the data-driven study of collective memory. Researchers have leveraged the textual content of Wikipedia articles (29), as well as logs of both editing (30) and viewing (31, 32), as proxies for the collective memory of traumatic events such as terrorist attacks or airplane crashes. Jatowt et al. (33) characterized the coverage and popularity of historical figures in Wikipedia, observing vastly increased page-view counts for people from the 15th and 16th centuries, a fact that Jara-Figueroa et al. (34) later attributed to the invention of the printing press. In addition to news and

Significance

Who is remembered by society after they die? Although scholars as well as the broader public have speculated about this question since ancient times, we still lack a detailed understanding of the processes at work when a public figure dies and their media image solidifies and is committed to the collective memory. To close this gap, we leverage a comprehensive 5-y dataset of online news and social media posts with millions of documents per day. By tracking mentions of thousands of public figures during the year following their death, we reveal and model the prototypical patterns and biographic correlates of postmortem media attention, as well as systematic differences in how the news vs. social media remember deceased public figures.

Author contributions: R.W., J.L., and C.P. designed research; R.W. performed research; R.W. analyzed data; and R.W., J.L., and C.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: robert.west@epfl.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2106152118/-/DCSupplemental.

Published September 15, 2021.

Philip Seymour Hoffman † 2014

Amy Winehouse † 2011

Questions

- Who is remembered by society after death?
  - “Postmortem collective memory”
- Are there prototypical patterns of postmortem collective memory?
- Are certain kinds of people remembered in certain ways?
- Are dead people remembered differently in news vs. social media?

Why should we care about postmortem collective memory?

Fact: Humans care a lot about being remembered after death

Ara Pacis, Rome
[Wikipedia]

Damnatio memoriae
[Wikipedia]

An ADA approach

Stuffed Count von Count counting stuff

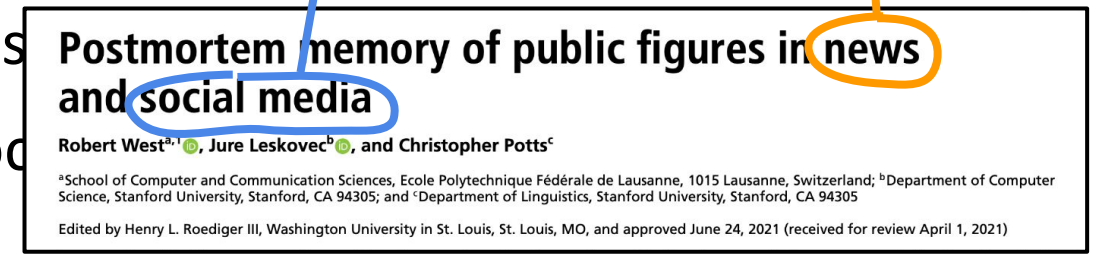
Let’s use lots of data and count stuff!

- **Detect names** of dead people in big corpus of news and social media
- Build time series of name **counts**
- Analyze the **shape of time series**
- **Correlate** shapes with biographic info about dead people from knowledge base

Part 1:
Getting the data

The raw data: spinn3r

- “Spinn3r provides APIs for social media, weblogs, news, video, and live web content to our customers in any language and in large volumes.” (Source: spinn3r.com)
- Firehose stored to dis
- Several billions of doc
- 40 terabytes



Postmortem memory of public figures in news and social media

Robert West[a,1], Jure Leskovec[b], and Christopher Potts[c]

[a]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [b]Department of Computer Science, Stanford University, Stanford, CA 94305; and [c]Department of Linguistics, Stanford University, Stanford, CA 94305

Edited by Henry L. Roediger III, Washington University in St. Louis, St. Louis, MO, and approved June 24, 2021 (received for review April 1, 2021)

ada #13 Scaling to massive data

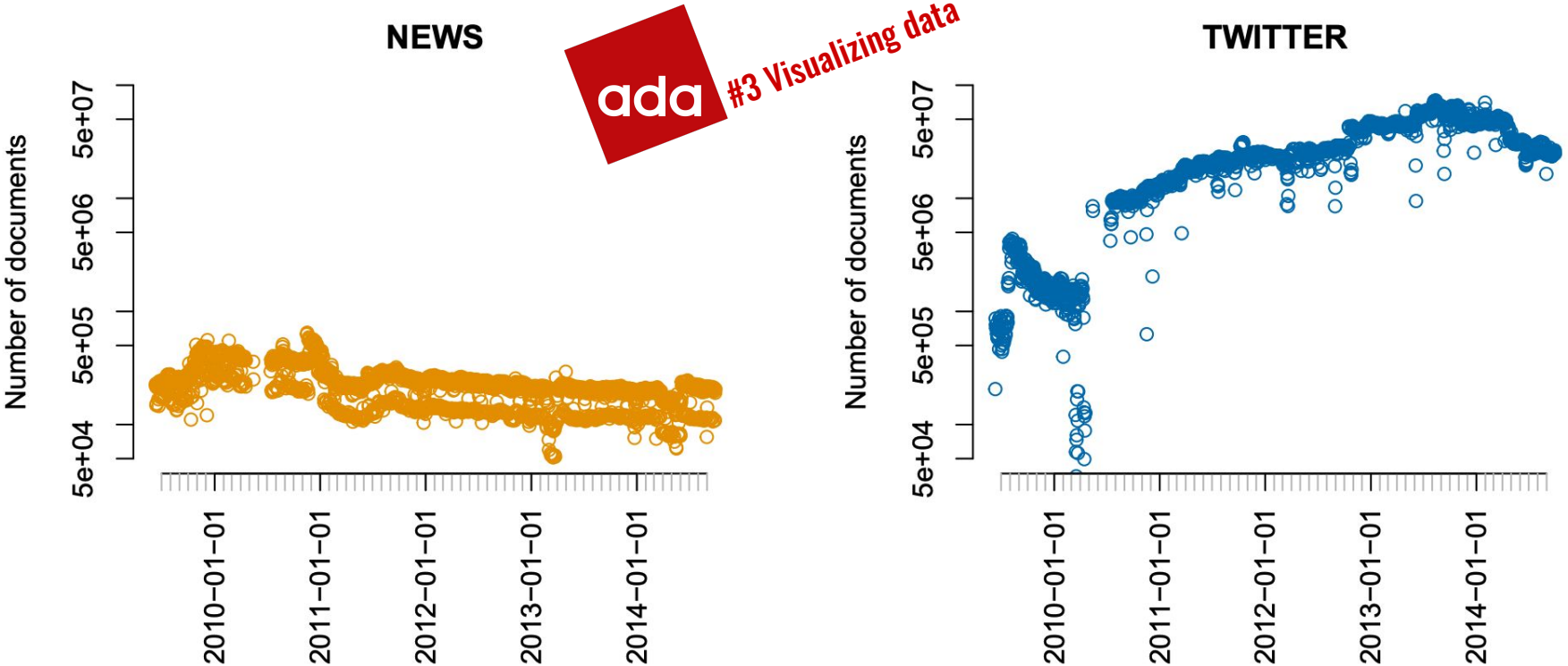
Detecting news and social media in Spinn3r

- Found a list of all 151K online news articles about Osama bin Laden's killing (2 May 2011) indexed by Google News
- Assume that every relevant news outlet had reported on bin Laden's death
- News defined as documents from the 6,608 Web domains appearing in the "bin Laden list"

- Detect via URL pattern: e.g., https://twitter.com/UserName/status/1605097142552403968

ada #2 Handling data

Data volume: Number of docs per day



Detecting mentions of people names

ada

#2 Handling data
#12 Handling network data
#10+11 Handling text data

- Easy: unambiguous names
  - e.g., “Amy Winehouse”
- Hard: ambiguous names
  - e.g., “Michael Jackson”

Wikipedia-based solution:

- Given: a name *X* (e.g., “Michael Jackson”)
- Consider all *N* links in English Wikipedia with *X* as anchor text (cf. examples on the right)
- Build distribution over the *N* link targets
- If there is a target *T* to which ≥90% of all links point: consider *X* “sufficiently unambiguous” and assume that all mentions of *X* in Spinn3r refer to *T*
- Else: consider *X* “too ambiguous” and ignore it in the analysis

The city of Gary is known as the birthplace of singer Michael Jackson.

Beer and whisky expert Michael Jackson was born in Leeds.

…

WIKIPEDIA

Michael Jackson (disambiguation)

Article Talk

Michael Jackson (1958–2009) was an American singer, songwriter and dancer known as the "King of Pop".

Michael Jackson, **Mike Jackson**, or **Mick Jackson** may also refer to:

People

Entertainment industry

- Michael Jackson (radio commentator) (1934–2022), American radio talk show host, KABC and KGIL, Los Angeles
- Michael Jackson (writer) (1942–2007), *Beer Hunter* show host, beer and whisky expert
- Mick Jackson (director) (born 1943), British film and TV director, known for *The Bodyguard*
- Michael J. Jackson (born 1948), English actor from Liverpool, best known for his role in *Brookside*
- Michael Jackson (television executive) (born 1958), British television executive
- Mick Jackson (author) (born 1960), British writer, known for *The Underground Man*
- Mike Jackson (photographer) (born 1966), British abstract and landscape photographer, known for *Poppit Sands* images
- Michael Jackson (actor) (born 1970), Canadian actor
- Mike Jackson (film producer) (born 1972), American film producer and talent manager
- Michael R. Jackson (born 1981), American playwright, composer, and lyricist

Musicians

- Mike Jackson (musician) (1888–1945), American jazz pianist and composer
- Mike Jackson (Australian entertainer) (born 1946), Australian multi-instrumentalist, songwriter and children's entertainer
- Mick Jackson (singer) (born 1947), English singer-songwriter
- Michael Gregory (jazz guitarist) (born 1953), American jazz guitarist, born Michael Gregory Jackson
- Mike and Michelle Jackson, Australian multi-instrumental duo
- Michael Jackson (English singer) (born 1964), British singer with the heavy metal band Satan/Pariah
- Oh No (musician), birth name Michael Woodrow Jackson (born 1978), American rapper
- Michael Lee Jackson, guitarist
- Mick Jackson, bassist with British band Love Affair (1950-)

Military and militants

- Michael Jackson (American soldier) (1734–1801), soldier from Massachusetts, wounded at Bunker Hill
- Mike Jackson (British Army officer) (born 1944), former head of the British

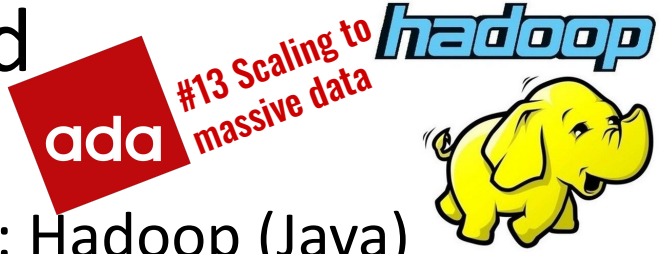
Recruiting the army of the dead: Freebase

	All	Included
Age		
N/A	10%	6%
1st quartile	68	64
Mean	76	74
Median	80	77
3rd quartile	88	87
Gender		
N/A	27%	7%
Female	16%	17%
Male	84%	83%
Manner of death		
N/A	76%	60%
Natural	85%	86%
Unnatural	15%	14%
Language		
N/A	45%	27%
Anglophone	60%	82%
Non-anglophone	40%	18%
Notability type		
N/A	1%	0%
Arts	40%	50%
Sports	14%	14%
Leadership	11%	14%
Known for death	26%	16%
General fame	7%	4%
Academia/engineering	2%	2%
Count	33 340	2 362

- Knowledge graph
  - Now defunct (a.k.a. dead…), but still available
- Contains information about >3M people
  - e.g., Philip Seymour Hoffman == /m/02qgqt
  - More info about some, less about others
- Start from 33K people with death date 2009–2014
  - Further filtering → 2,362 people
- Extract biographic information from Freebase

ada #2 Handling data #12 Handling network data

Tools used



- Counting names in Spinn3r dump: Hadoop (Java)
- Extracting info from Freebase: Python, Perl
- Once data was small enough: R (script, repo)
  - Statistical analyses
  - Plotting
  - Read data as CSV once (minutes), then serialized to binary format and deserialized in later runs (seconds)

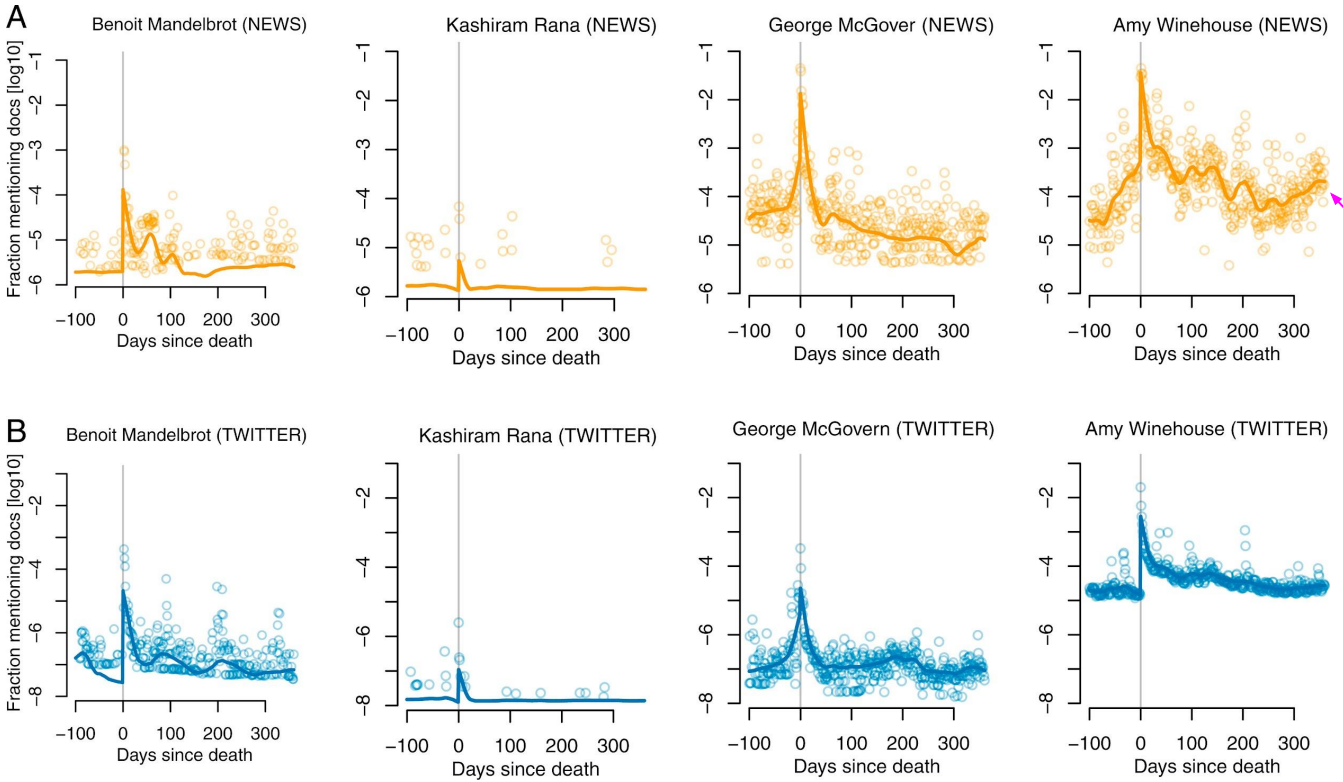
Intrat: Mention time series (our protagonist)

- t: relative time t, counting days since death
  - $t = 0$: day of death
- $S_i(t)$: fraction of documents in which person i was mentioned, out of all documents published on day t
  - For mention time series, consider logarithms: $\log_{10} S_i(t)$

Mention time series: examples

News

Twitter

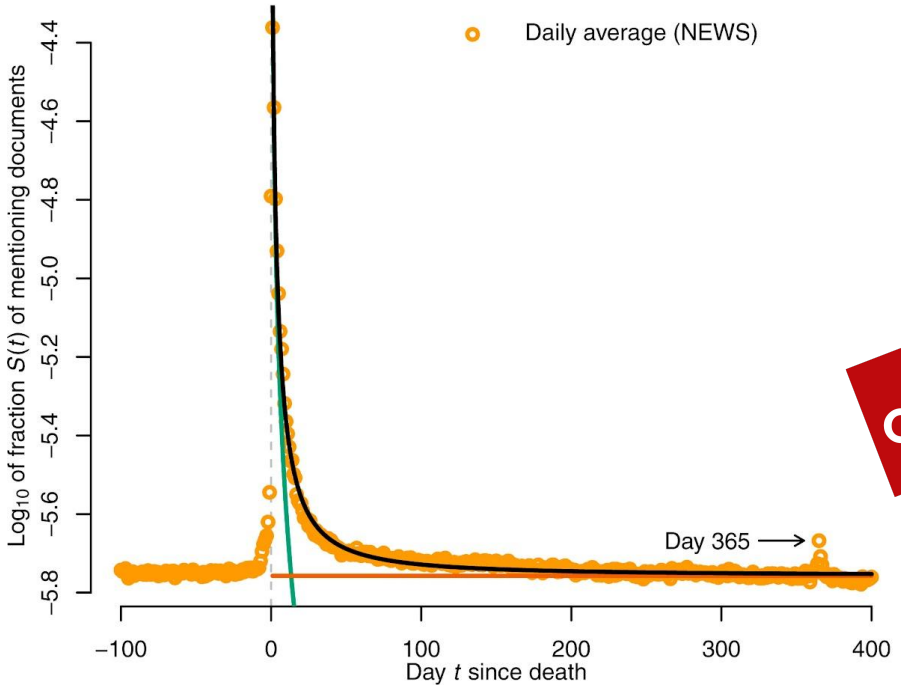


Smoothed via "Friedman's super smoother"

Part 2:
The shape of postmortem memory

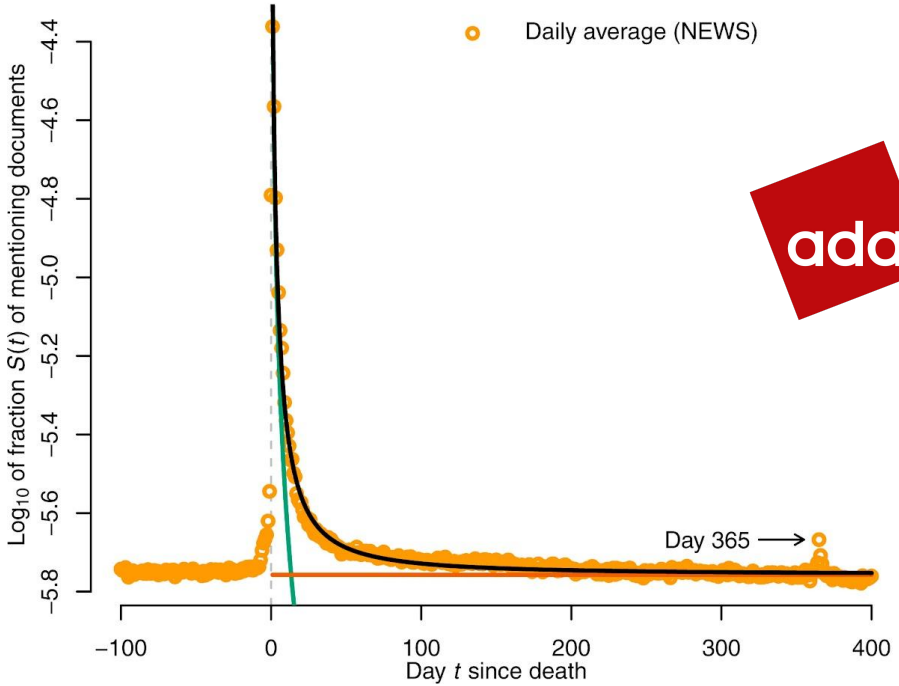
Average mention time series

News



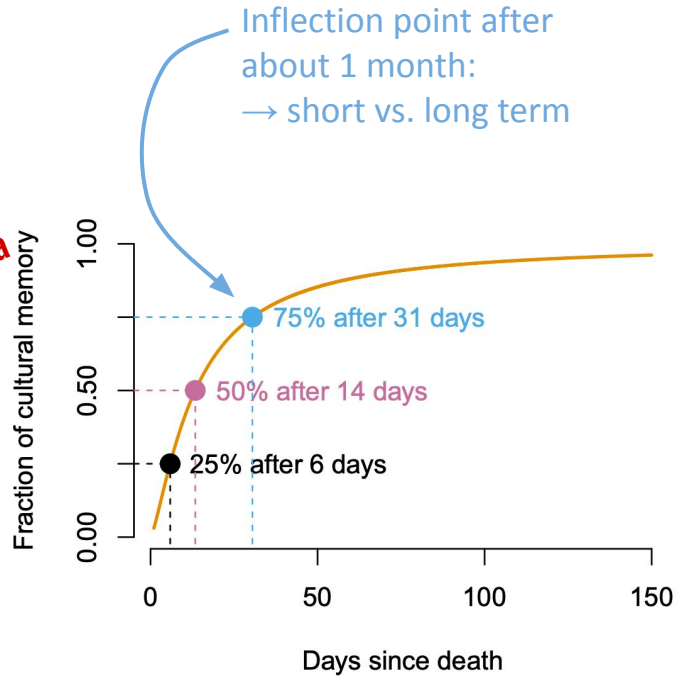
ada #4 Describing data

Modeling the data

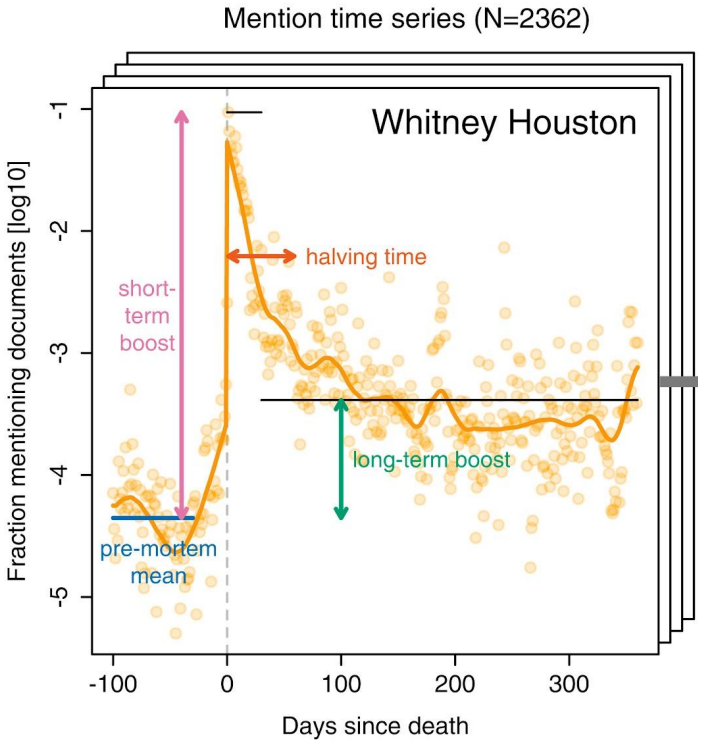


ada #3 Visualizing data

$y(t) = v(t)/(u(t) + v(t))$



Curve characteristics



Pre-mortem mean: arithmetic mean of days 360 through 30 before death

Short-term boost: maximum of days 0 through 29 after death, minus the premortem mean

Long-term boost: arithmetic mean of days 30 through 360 after death, minus the premortem mean

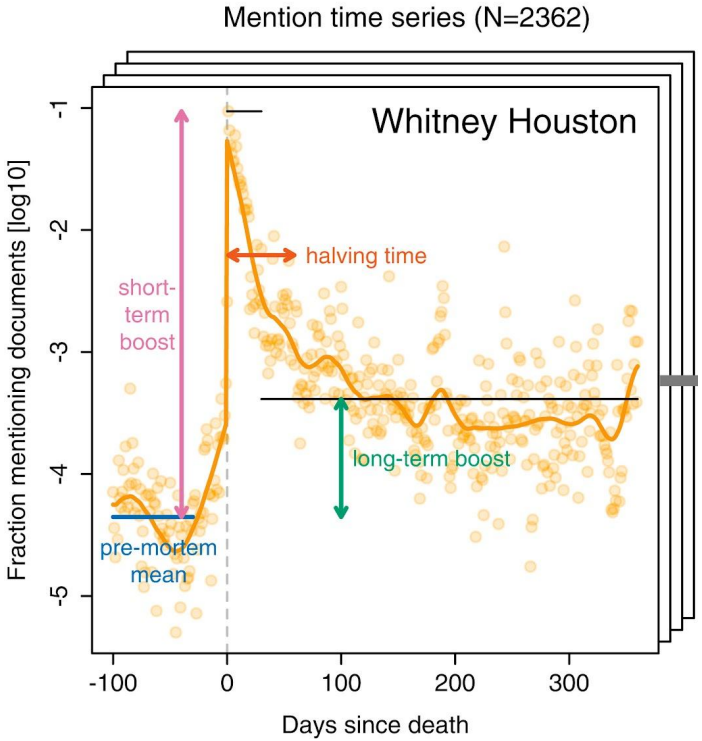
Halving time: number of days required to accumulate half of the total area between the postmortem curve (including the day of death) and the minimum postmortem value

Median over people:
1.98
95% CI [1.90, 2.03]

Median over people:
0.00055
95% CI [-0.00091, 0.0017]

ada #4 Describing data

Are there prototypical curve shapes?



- Each curve one data point
- Represented via its 4 curve characteristics
  - ≈ manual dimensionality reduction
  - Values standardized via *z*-scores
- Cluster via *k*-means

ada

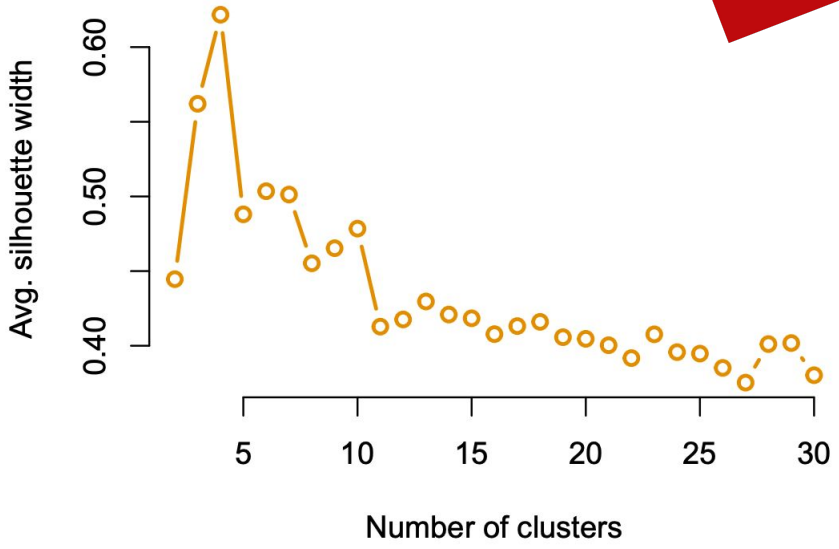
#8 Applied ML
#9 Unsupervised learning

Q: How to find the best number *k* of clusters?

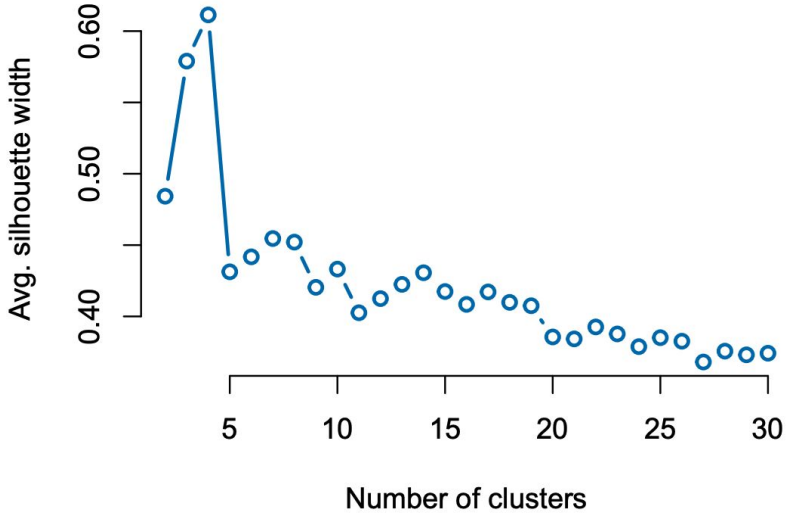
Average silhouette width!

ada #9 Unsupervised learning

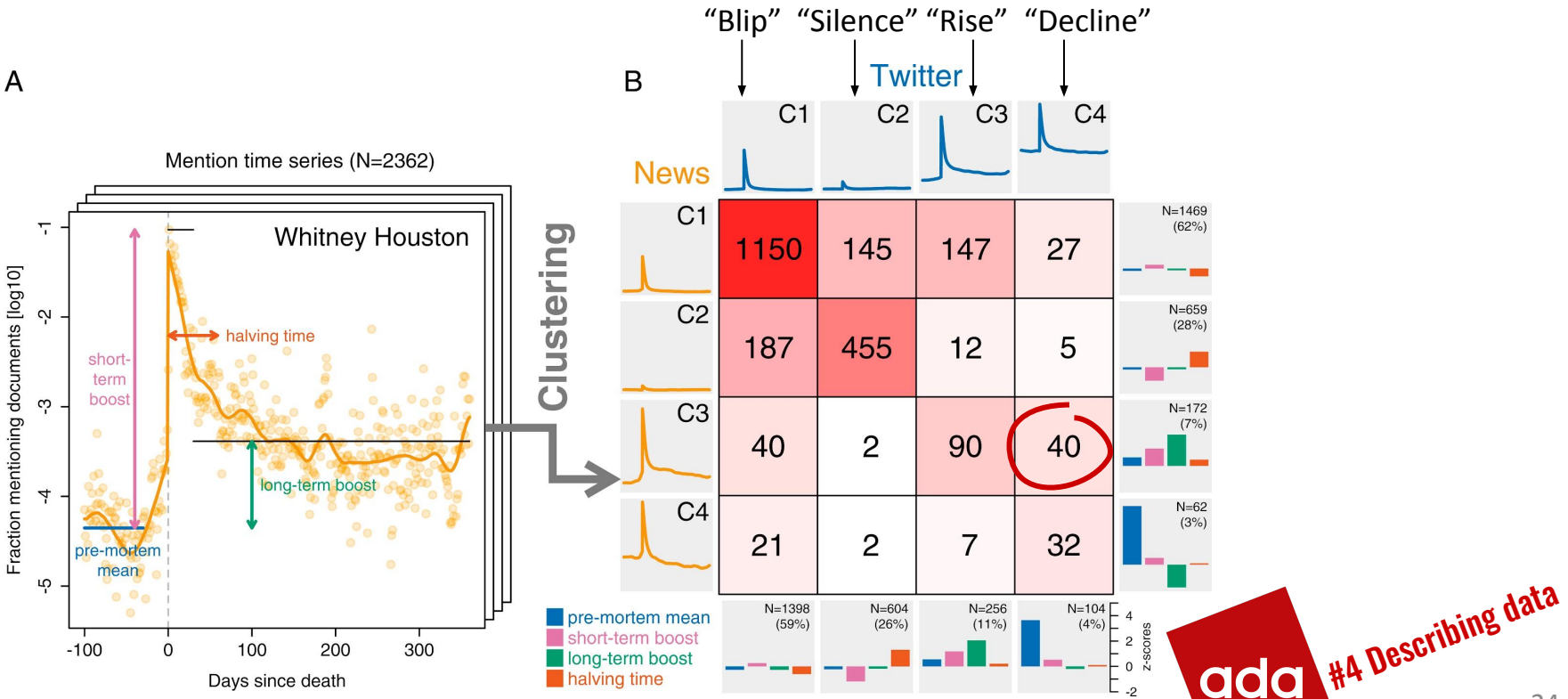
News



Twitter



Cluster analysis



ada #4 Describing data

Part 3:
Biographic correlates of postmortem memory

A first stab

- Measure correlation coefficient between outcome (e.g., short-term memory boost) and each biographic properties (e.g., gender)
- Higher correlation ⇒ higher outcome for the respective group of people (e.g., women)

Age
N/A
1st quartile
Mean
Median
3rd quartile
Gender
N/A
Female
Male
Manner of death
N/A
Natural
Unnatural
Language
N/A
Anglophone
Non-anglophone
Notability type
N/A
Arts
Sports
Leadership
Known for death
General fame
Academia/engineering
Count

Problem: Biographic properties are correlated

E.g., leaders (politicians, CEOs, etc.) are

- more likely to have died old,
- more likely to have died of a natural death,
- more likely to be men,

compared to artists

Regression analysis allows us to compare averages across subgroups of the data while accounting for correlations among averaged values!

Linear regression

Avg. outcome for "baseline persona": male anglophone artist of median premortem popularity who died a natural death at age 70–79

$$y_i = \beta_0 + \beta_1 \text{premortem\_mention\_freq}_i + \beta_2 \text{age\_at\_death}_i + \beta_3 \text{manner\_of\_death}_i + \beta_4 \text{notability\_type}_i + \beta_5 \text{language}_i + \beta_6 \text{gender}_i$$

Rank-transformed, then linearly scaled/shifted to [-0.5, 0.5]; i.e., median has value 0

Outcome for person *i*:

- short-term boost or
- long-term boost

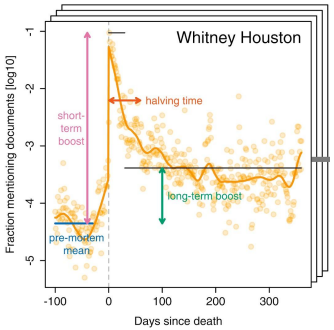
8 discrete levels (dummy-coded): 20–29, 30–39, …, 70–79, 80–89, 90–99

2 levels: natural, unnatural

6 levels: arts, sports, leadership, known for death, general fame, academia/engineering

3 levels: anglophone, non-anglophone, unknown

2 levels: male, female

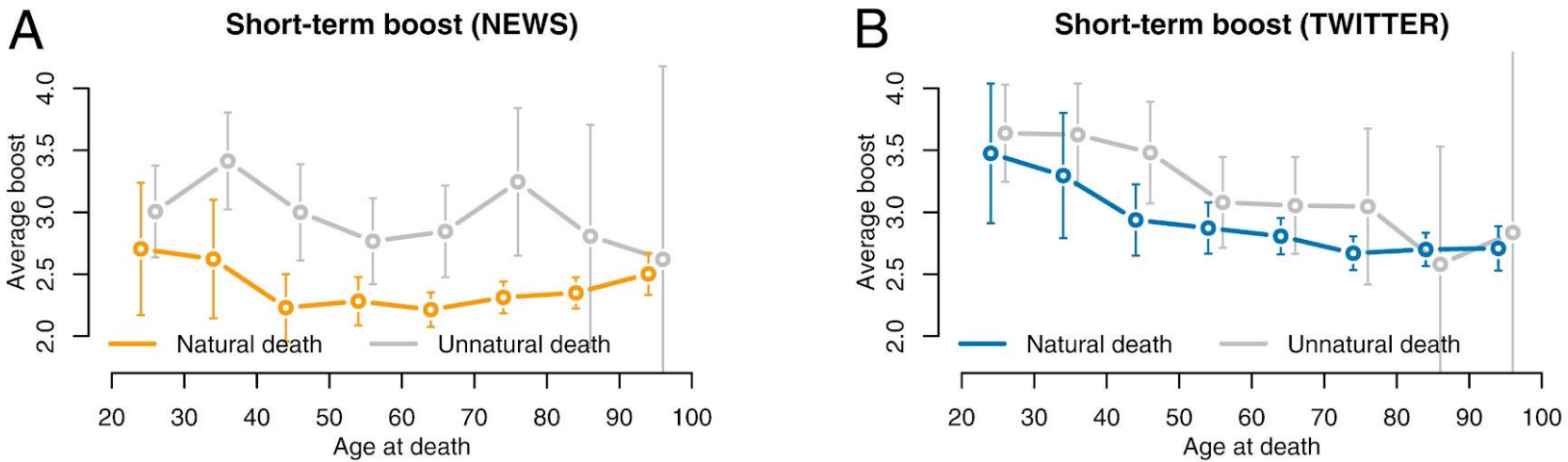


Linear regression results

	Short-term boost (news)	Short-term boost (Twitter)	Long-term boost (news)	Long-term boost (Twitter)
(Intercept)	2.322 (0.063)***	2.670 (0.067)***	0.088 (0.014)***	0.095 (0.015)***
Premortem mean (relative rank)	0.804 (0.093)***	0.948 (0.100)***	0.031 (0.020)	0.086 (0.022)***
Manner of death: unnatural	0.618 (0.095)***	0.282 (0.100)**	0.097 (0.021)***	0.106 (0.022)***
Language: non-anglophone	−0.316 (0.074)***	−0.116 (0.078)	−0.061 (0.016)***	−0.037 (0.017)*
Language: unknown	−0.446 (0.086)***	−0.325 (0.091)***	−0.079 (0.019)***	−0.081 (0.020)***
Gender: female	0.083 (0.072)	−0.034 (0.076)	0.034 (0.016)*	0.006 (0.017)
Notability type: academia/engineering	0.181 (0.197)	0.340 (0.208)	−0.032 (0.043)	0.023 (0.046)
Notability type: general fame	0.070 (0.124)	0.132 (0.131)	−0.010 (0.027)	−0.008 (0.029)
Notability type: known for death	−0.107 (0.099)	−0.088 (0.106)	−0.021 (0.022)	0.008 (0.023)
Notability type: leadership	0.152 (0.083)	0.113 (0.087)	−0.058 (0.018)**	−0.040 (0.019)*
Notability type: sports	0.049 (0.083)	0.072 (0.088)	−0.034 (0.018)	−0.034 (0.020)
Age: 20–29	0.162 (0.170)	0.718 (0.180)***	0.060 (0.037)	0.192 (0.040)***
Age: 30–39	0.400 (0.167)*	0.649 (0.177)***	0.028 (0.037)	0.118 (0.039)**
Age: 40–49	−0.046 (0.126)	0.351 (0.133)**	−0.001 (0.028)	0.100 (0.030)***
Age: 50–59	−0.075 (0.099)	0.181 (0.104)	−0.058 (0.022)**	−0.024 (0.023)
Age: 60–69	−0.109 (0.082)	0.130 (0.086)	−0.050 (0.018)**	−0.025 (0.019)
Age: 80–89	0.022 (0.078)	0.021 (0.082)	−0.018 (0.017)	−0.013 (0.018)
Age: 90–99	0.174 (0.098)	0.034 (0.103)	−0.011 (0.021)	−0.024 (0.023)
R^2	0.213	0.192	0.123	0.178
Adj. R^2	0.197	0.176	0.106	0.161
No. obs.	870	870	870	870
RMSE	0.772	0.815	0.169	0.181

SEs of coefficients are in parentheses. ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$.

Age at death vs. postmortem memory

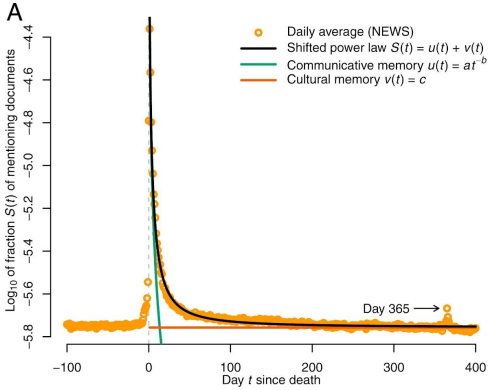


News plays two simultaneous roles (more so than Twitter):

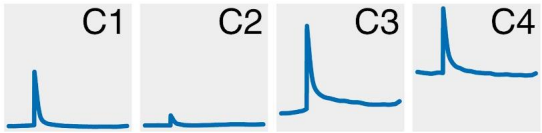
- Catering to public curiosity stirred by a young or unnatural death
- But also when old person or accomplished leader dies

Part 4:
Discussion

Summary: The shape of postmortem memory



- Sharp **pulse** of media attention with death:
  - Median: +9,400% in news, +28,000% on Twitter
- Then sharp **drop** (around 1 month long) toward premortem level
- **Two components** of collective memory:
  - Baseline level of **cultural** memory built up during life (constant)
  - Added layer of **communicative** memory sparked by death (power law)
- **Cluster analysis** revealed a set of four prototypical memory patterns: “Blip”, “silence”, “rise”, “decline”
- Same patterns in news and Twitter; **same person** tends to fall into the **same cluster** across the two media



Summary: Biographic correlates

- Notability types: all regression coefficients for long-term boosts negative
  ⇒ All types have lower average long-term boost than default type (artists)
  ⇒ **Artists more present in collective memory**
- **Low R^2** (10–20%): human lives/legacies rich, hard to model
  - But all **model fits highly significant** (*F*-statistic, *p*-value)
  - **Effects** not only significant, but also **large:** e.g., short-term boost (on linear scale) for unnatural vs. natural death: 4x in news, 2x on Twitter
- **Largest boost:**
  - **Premortem popular anglophones who died a young, unnatural death**
  - Long-term boosts **largest for artists, smallest for leaders**

Merry Christmas ADA happy New Year!