Insights into Social Media Data:
a new formalism inspired in Thermodynamics

Mars Geldard (Me)

Programmer with a Data Science bent

Professor Paulo de Souza

Physicist

Dr James Montgomery

Evolutionary Computation Researcher

Agenda

1. **Background**
   → Literature and motivation
2. **Methodology**
   → Aims and Objectives
   → Tools and methods used
3. **Findings**
   → Analysis results
   → Discussion

Background

Why?

Social Media Usage

→ Of the world's ~4.5 billion internet users, **~3.5 billion use social media**[1]

→ From 2005-2015 **social media usage rose 10x** in the US alone[2]

→ **Hundreds of platforms**, growing by the year[3], most use two or more[4]

→ Many count users in the millions, with some reaching **> 1 billion users**[2]

→ Social media usage spans demographics[5]; includes individuals, groups and businesses of all sizes[6]

→ Critical to maintaining a competitive edge in business[7]

→ Has fundamentally **changed the way humanity communicates**[8]

Social Media Analysis

→ Understand the network structure of platforms, how people connect (**social network analysis**, **behavioural research**)

→ Understand the content being generated and shared among platforms, how people react to and propagate it (**content and sentiment analysis**, **information diffusion theories**)

→ Understand where particular interest comes from, the determinants and effects of popular topics (**topic analysis**, **trend analysis**)

Challenges

→ Social media generates content of **enormous quantity** at great speed[10]

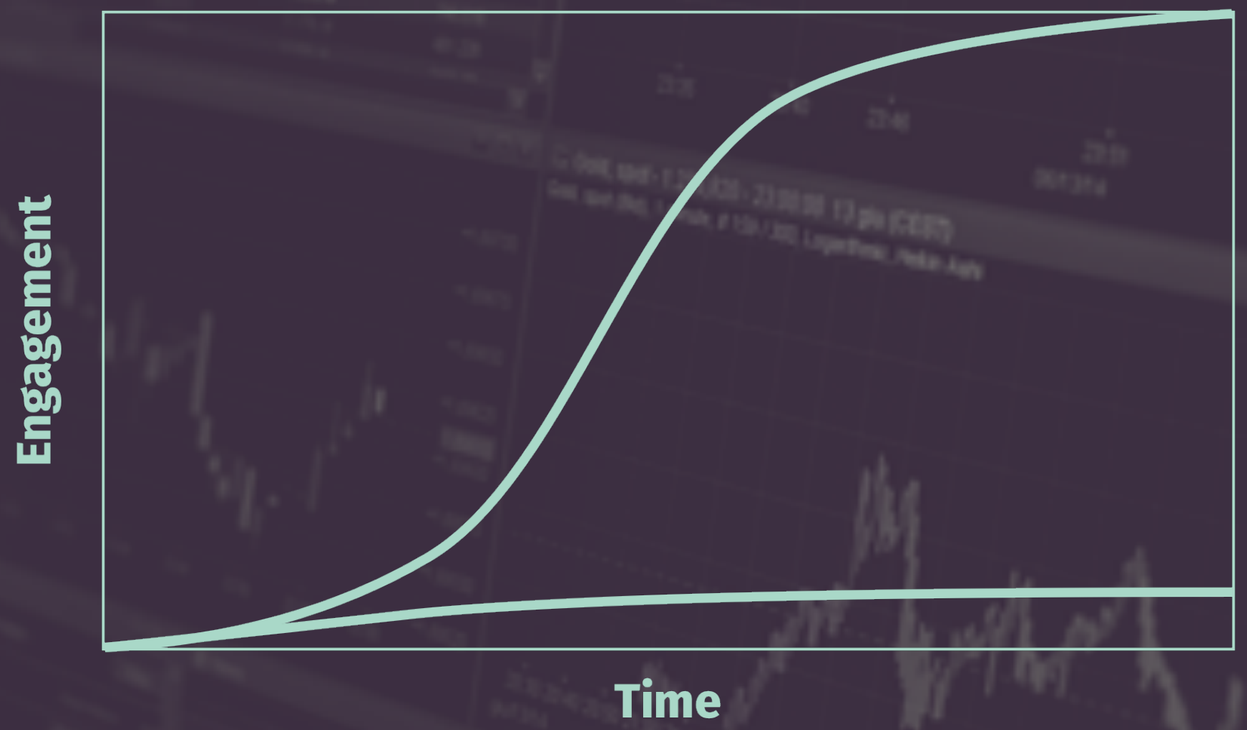
→ **Semi-structured** data that often includes **mixed media** necessary to context[11]

→ Metadata is sparse + **not always accurate**[12]

→ Hardest of **NLP**—jargon, context- + culturally-specific terms[13]

→ **Data is harder to get** over time due to platforms commoditising data[14] + increasing privacy controls[15]

Trend Analysis & "Prediction"



Trend Detection + Prediction

Detect Popular Topics	Predict how much more popular they will get
Detect Topics	Predict whether they will get popular, how popular they will get

Why?

→ Optimise marketing and outreach campaigns

→ Understand content consumption

→ Observe cultural change over time

→ Jump on bandwagons

→ Prevent impending harmful content trends

Challenges

→ Subjectivity of viewing, sharing content

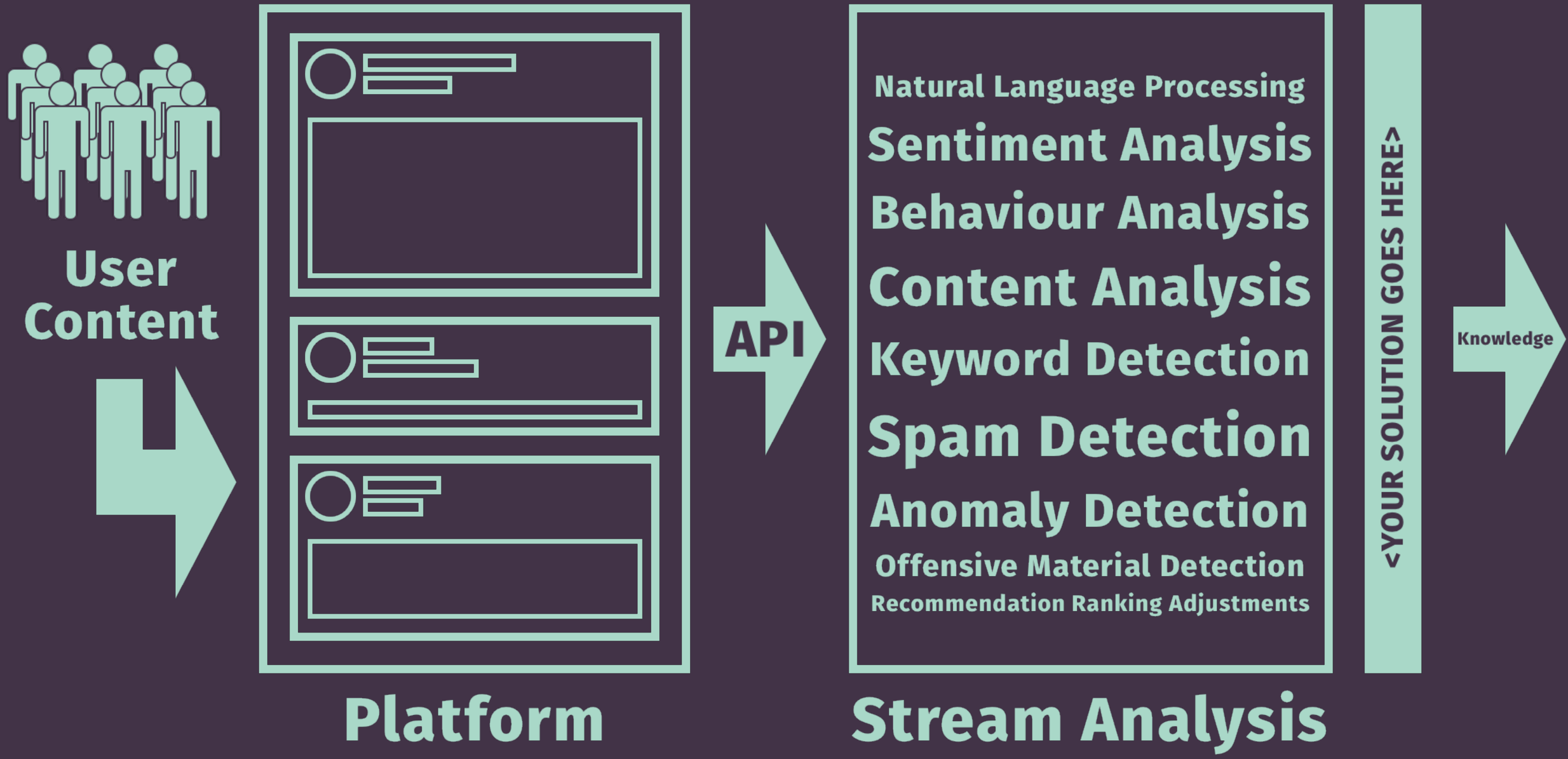
→ False positives in anomaly detection-based methods[16]

→ Reliance on other information sources in alternative methods[17]

→ Timeliness requirements for analysis and detection[18]

→ Topic evolution over time[19]

→ Predicting the future?


User Content
Platform
API
Natural Language Processing
Sentiment Analysis
Behaviour Analysis
Content Analysis
Keyword Detection
Spam Detection
Anomaly Detection
Offensive Material Detection
Recommendation Ranking Adjustments
Stream Analysis
<YOUR SOLUTION GOES HERE>
Knowledge

Multidisciplinary challenges

→ Very human data; erratic behaviours (**behavioural science**)

→ Sharing between non-fully-connected graph structure… (**graph theory**)

→ …of nodes with varying levels of transmissiveness (**social contagion**)

→ Different content and sentiments get shared in different ways (**information diffusion**)

"Social media analysis requires some new ways of thinking!"*

* see 15, 20-22

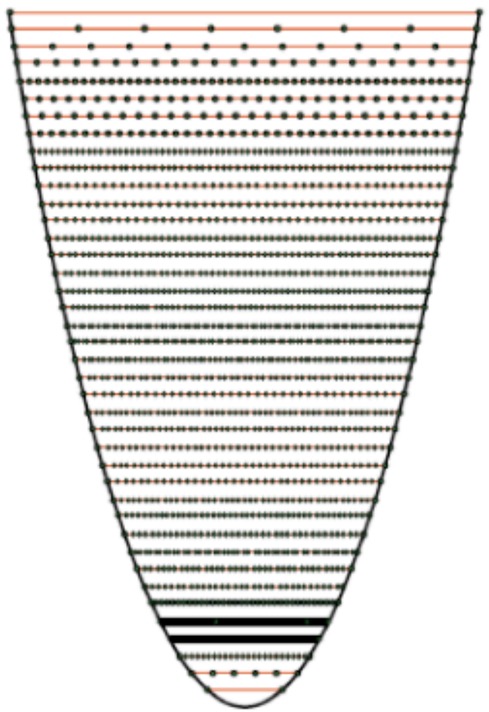
Novel cross-disciplinary applications

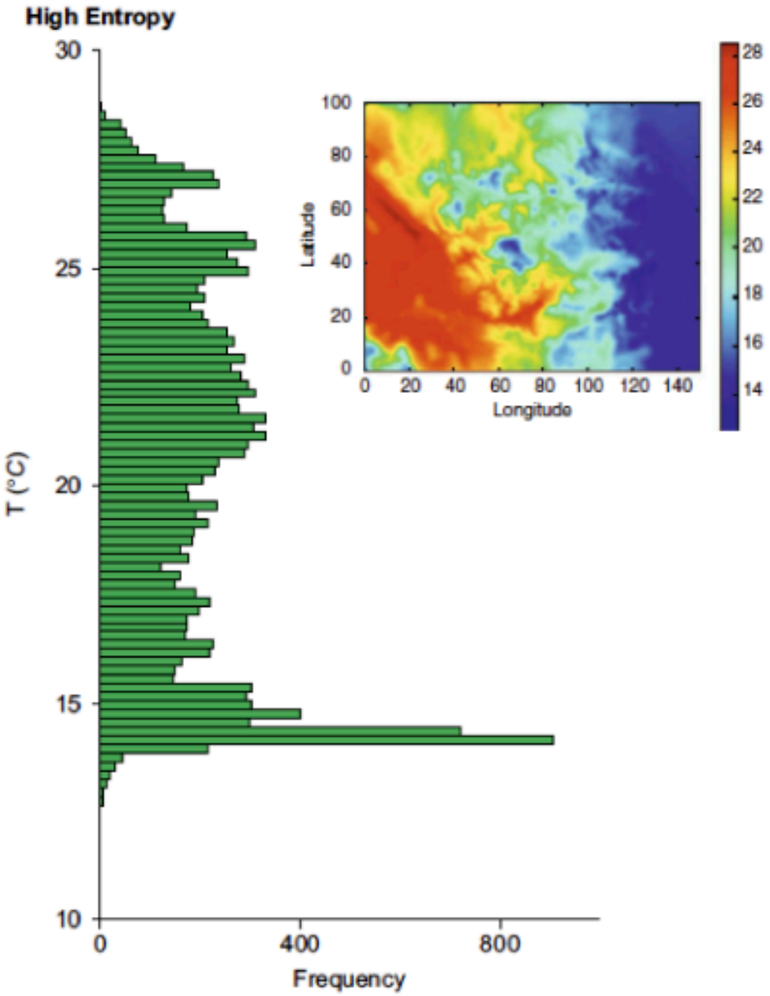
→ Social media trends as contagion theory from **Virology** in Medicine[23 - 25]

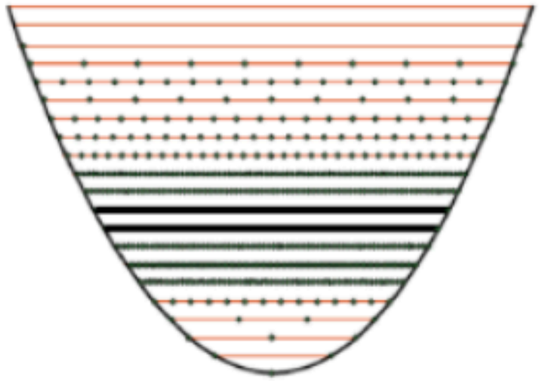
→ Social media trends as the **Technology Adoption** curve from Information Diffusion Theory[26]

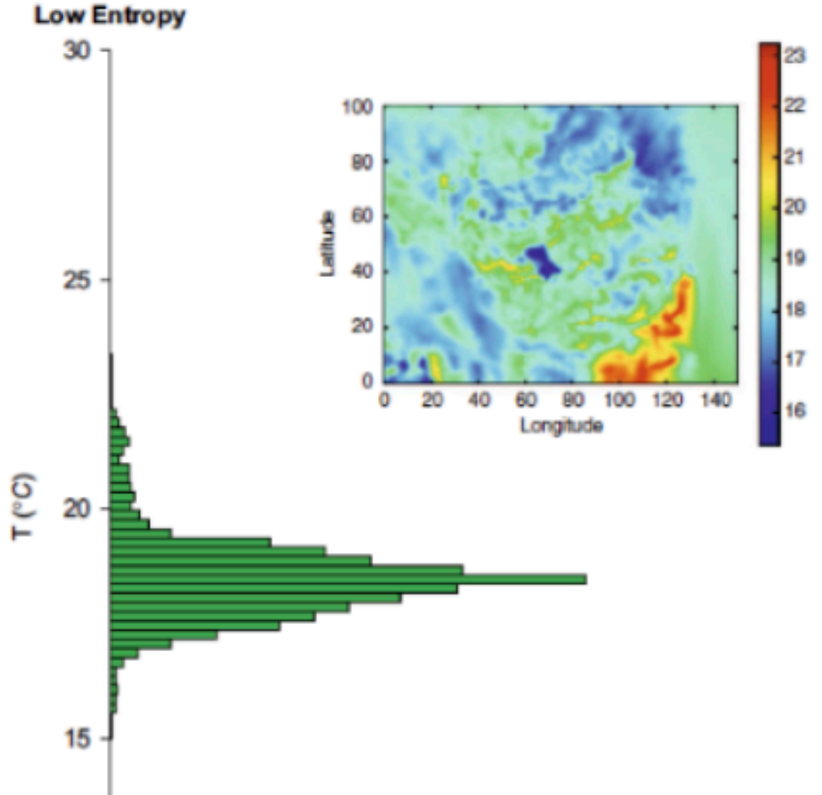
→ Social media behaviours observed as sound (**Data Sonification**) to aid in anomaly detection[27]

→ Social media network interactions as **Evolutionary Game Theory** from Evolutionary Biology[28]









Source: Mahbub, de Souza and Williams 2016

Entropy Theory

Boltzmann's $S = k_{\mathrm{B}} \ln(N)$

Upper bound of entropy in closed system at equilibrium

Gibbs' $S = -k_{\mathrm{B}} \sum_i P_i \ln P_i$

Upper bound of entropy in closed system not at equilibrium

Shannon's $S = -\sum_i P_i \ln P_i$

Average rate at which information is produced by a stochastic source of data

Entropy over time

→ Reveals descriptive statistical measures about a set of data

→ But also measures internal diversity and inconsistency

→ Takes focus away from upper and lower bounds—or even ranges—in favour of changes in internal distribution that may be more informative

→ Is in some cases sufficient to adapt analysis into Markov-model-based prediction

Aims & Objectives

Broadly

1. Can the analogous application of **Entropy Theory** present useful information about the lifetime of a trend or topic on a Social Media platform?
2. Is there indication that the method could be applied more broadly, or adapted to be **predictive**?*

*In that it could achieve earlier detection than existing early detection methods.

In this study

Experimentation aimed to confirm dual hypotheses:

1. **increasing entropy over time** may suggest impending critical interest preceding a trend developing to significance, and
2. **decreasing rate of entropy increase over time** after a high entropy or engagement level has occurred may suggest new evolution in the topic.

This required

→ data from social media platforms,

→ covering concise topics,

→ for significant periods of time, with

→ sufficiently varied trend behaviours and

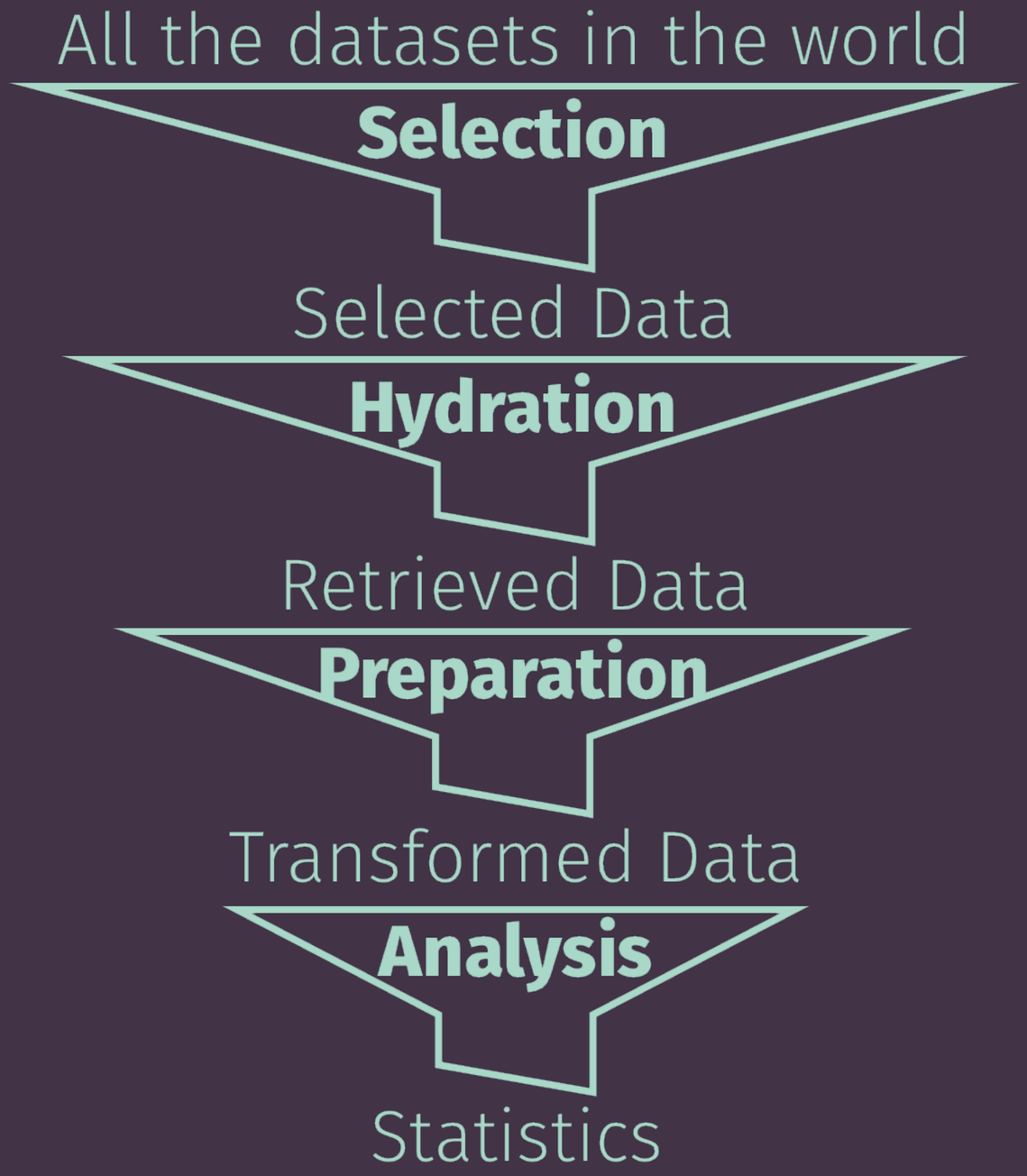
→ consistent measures of engagement.

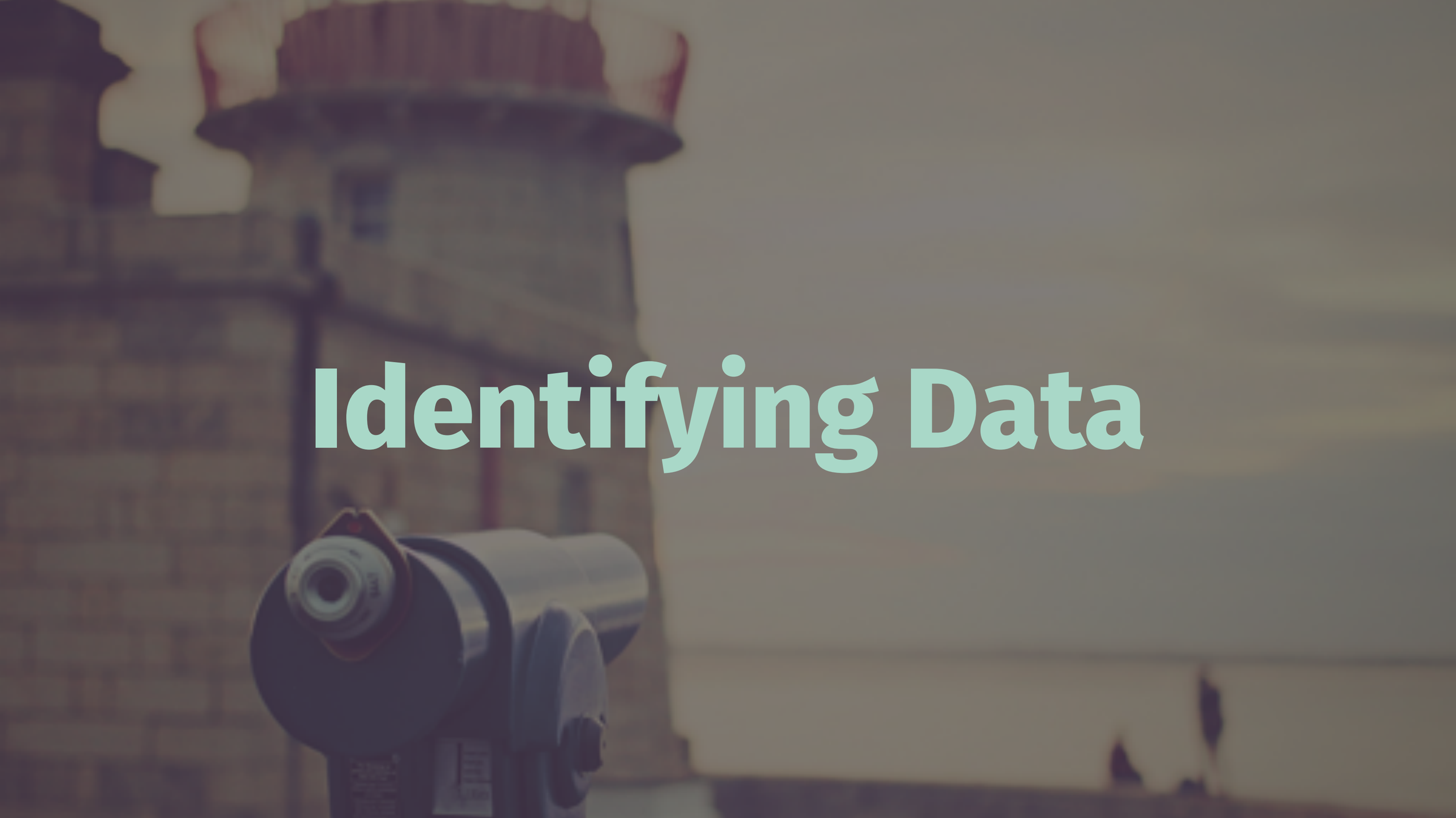
Methodology

How?

Process

1. Identifying Data
2. Retrieving Data
3. Preparing Data
4. Analysing Data

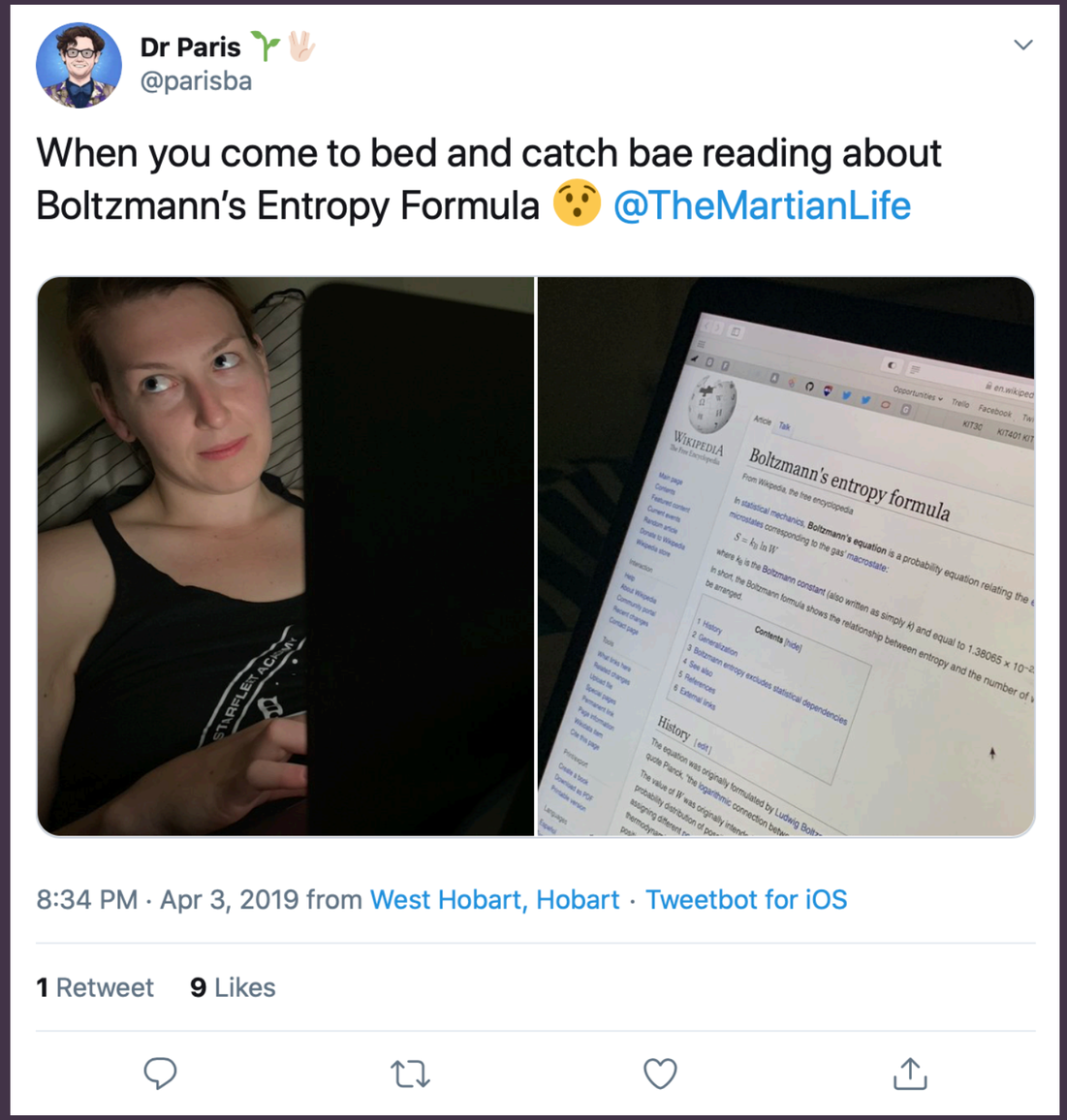




Identifying Data

Twitter!

- → Rapid generation of content, rapid decline in interest[17] => *easily observable trends*
- → Discrete content size easier to analyse[29] => *more easily verifiable datasets*
- → Breadth of functionality in API, well-established tools for research[30] => *easier to get data, perform and recreate studies*



Harvard & GW University researcher verified!

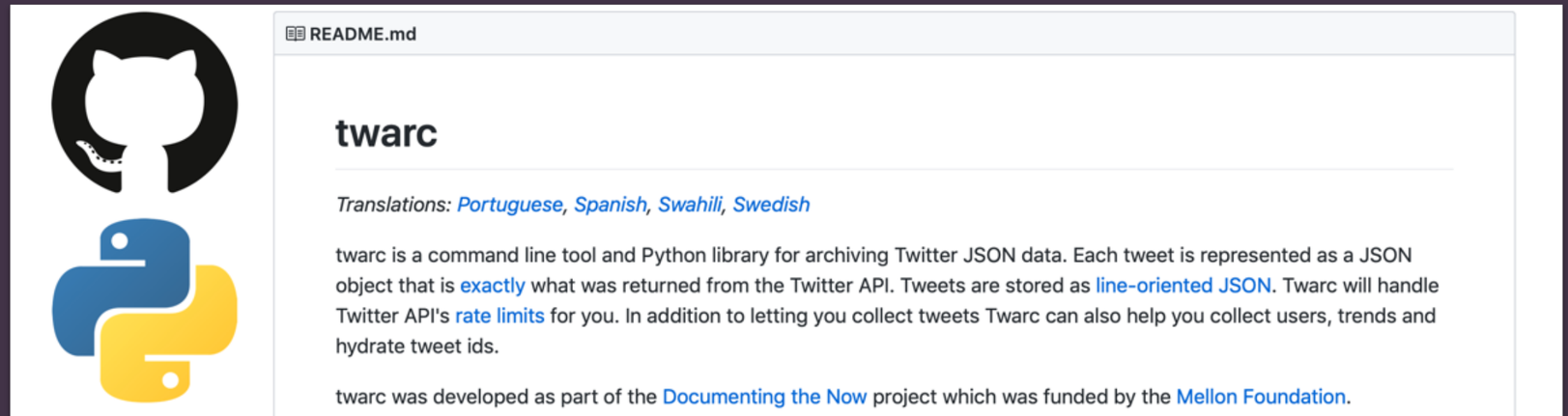
Selected topics: Hurricane Florence, Hurricane Harvey, Immigration Travel Ban, Ireland 8th Amendment Vote, Hurricane Irma, Winter Olympics, Women's March

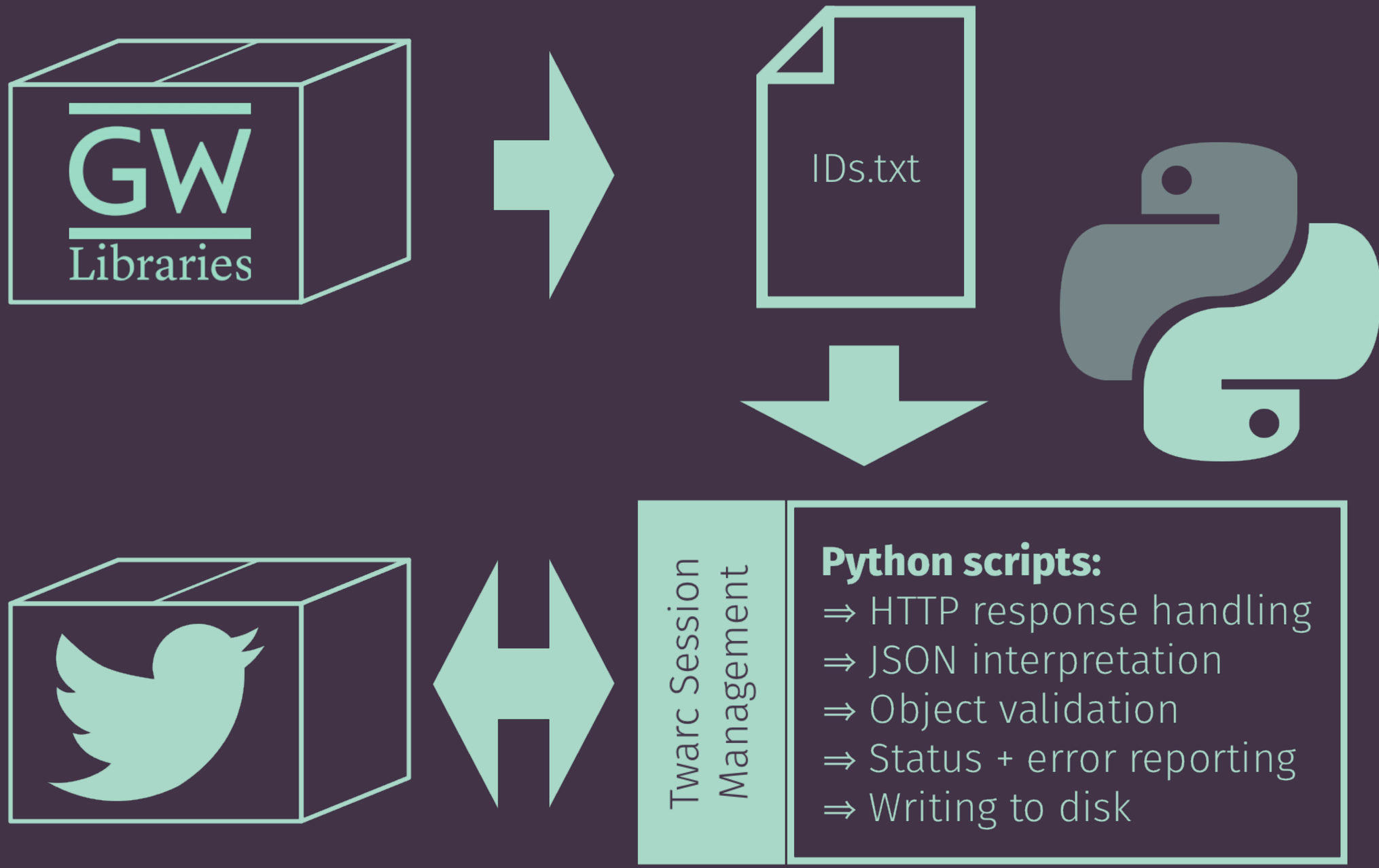
Retrieving Data

"Hydrating" Tweets: Twarc

→ made by researchers (from the *Documenting the Now* project)

→ many researchers have used it for studies before, establishing its robustness as a tool[31 - 38]

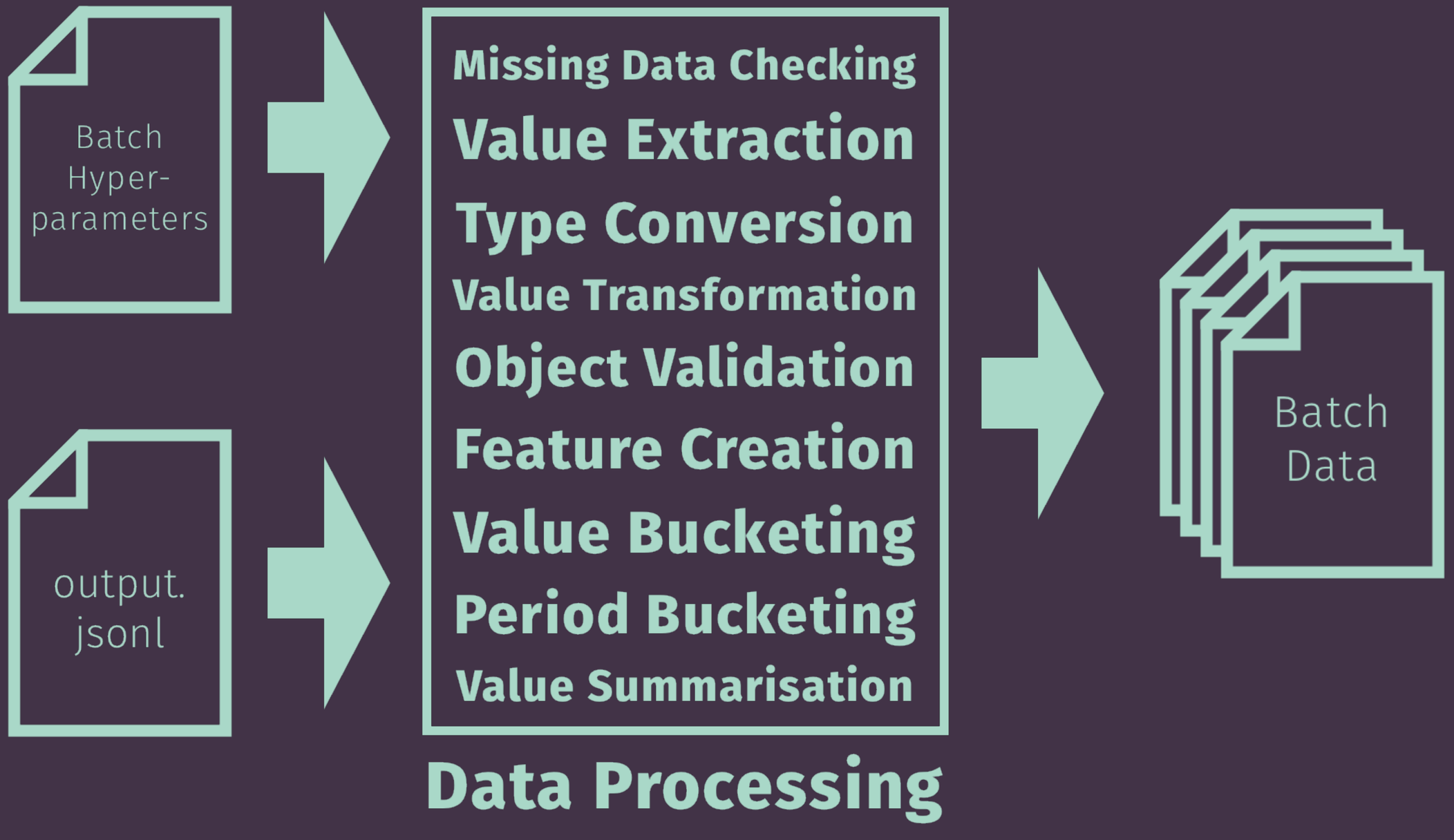



GW
Libraries
IDs.txt
Twarc Session Management
Python scripts:
⇛ HTTP response handling
⇛ JSON interpretation
⇛ Object validation
⇛ Status + error reporting
⇛ Writing to disk
Preparation

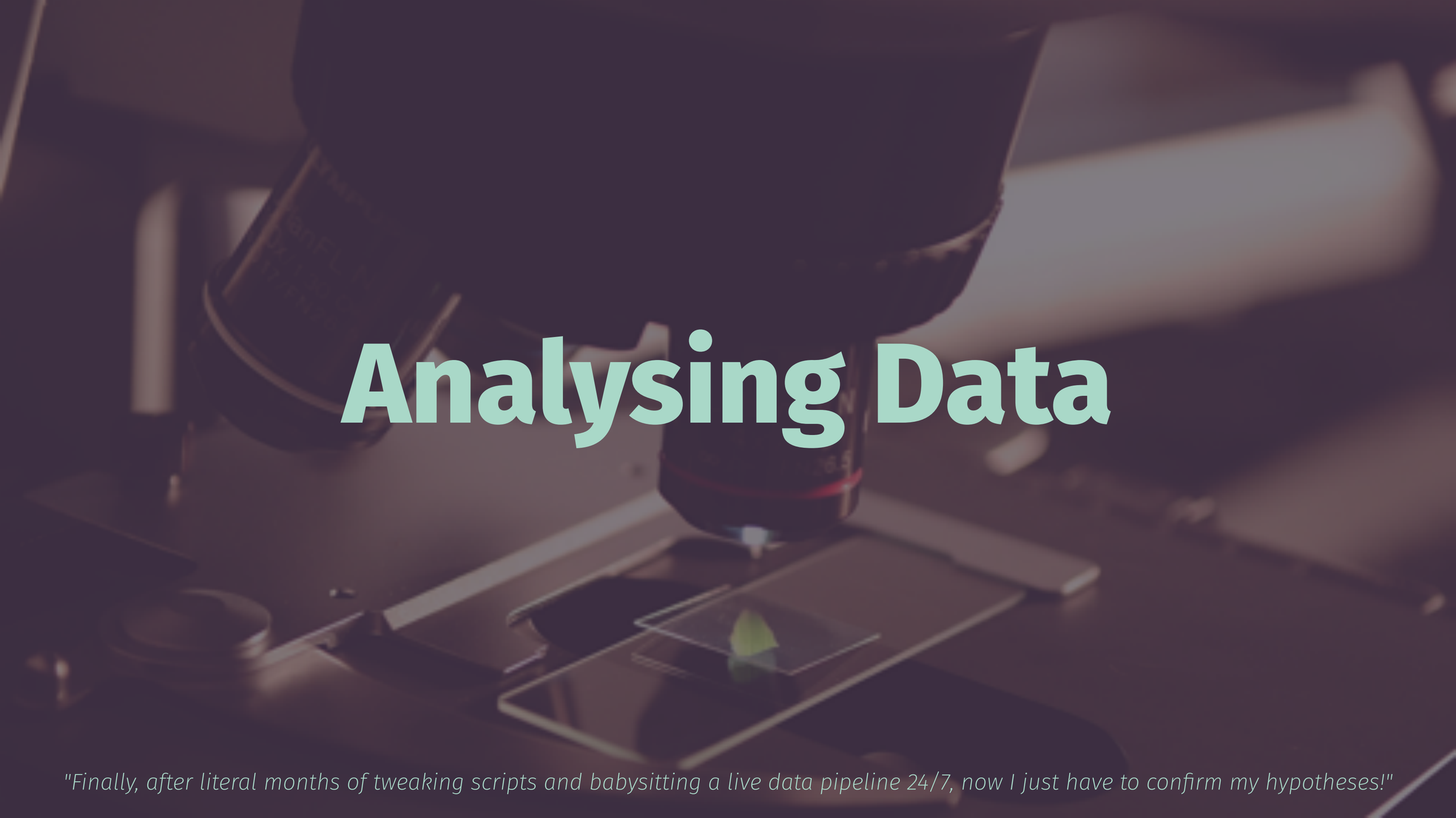
Hydration Loss

Data Group	Listed	Hydrated	Lost
Hurricane Harvey	18,336,283	13,018,265	5,318,018
Hurricane Irma	17,227,935	11,627,130	5,600,805
Immigration Travel Ban	16,842,329	11,618,463	5,223,866
Winter Olympics	13,787,299	10,567,141	3,220,158
Hurricane Florence	7,756,856	6,432,696	1,324,160
Women's March	7,263,988	4,526,921	2,737,067
Ireland Vote	2,276,808	1,784,345	492,463
Total	**83,491,498**	**59,574,961**	**23,916,537**

Preparing Data



Data Processing



Analysing Data

"Finally, after literal months of tweaking scripts and babysitting a live data pipeline 24/7, now I just have to confirm my hypotheses!"

Comparing Distributions

→ **Partitioning Data** ⇒ Segmented multiple regression, correlated with significant value thresholding, different bucketing values

→ **Assessing Data** ⇒ Conventional descriptive statistics, tests of central tendency, visualisation and exploration

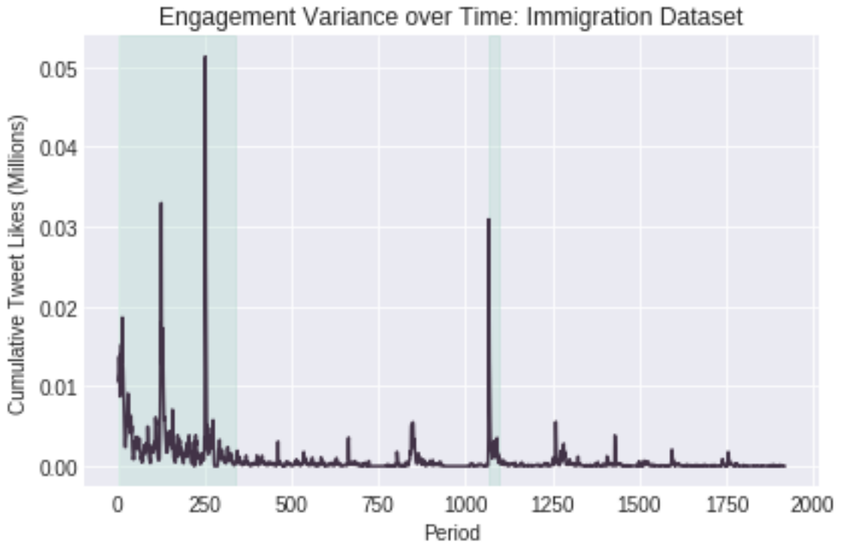
→ **Comparing Groups** ⇒ Appropriate non-parametric tests for goodness-of-fit, equal medians, expected variance

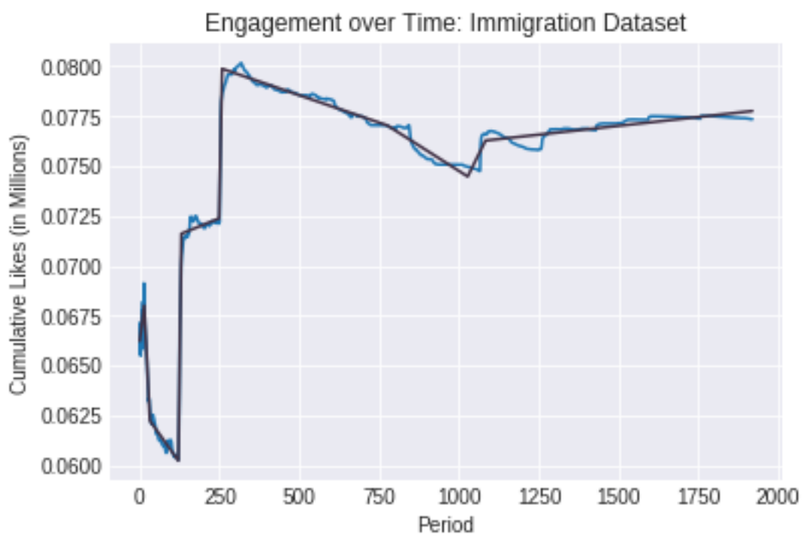
→ **Manipulating Scales** ⇒ Negating proposed baseline or previous values to assess variation, testing bucketing values

Findings

What?

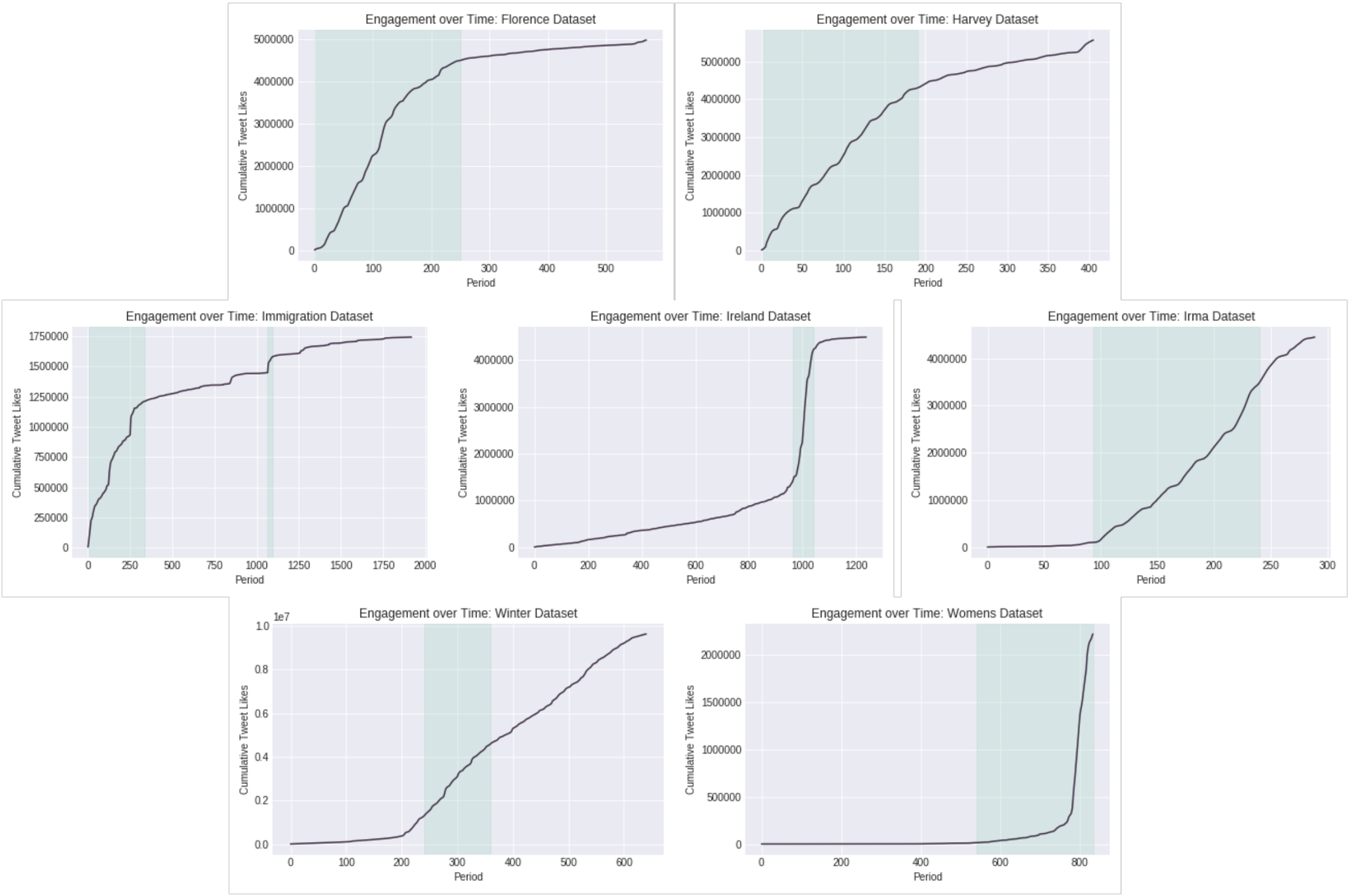
Identifying periods of significant growth



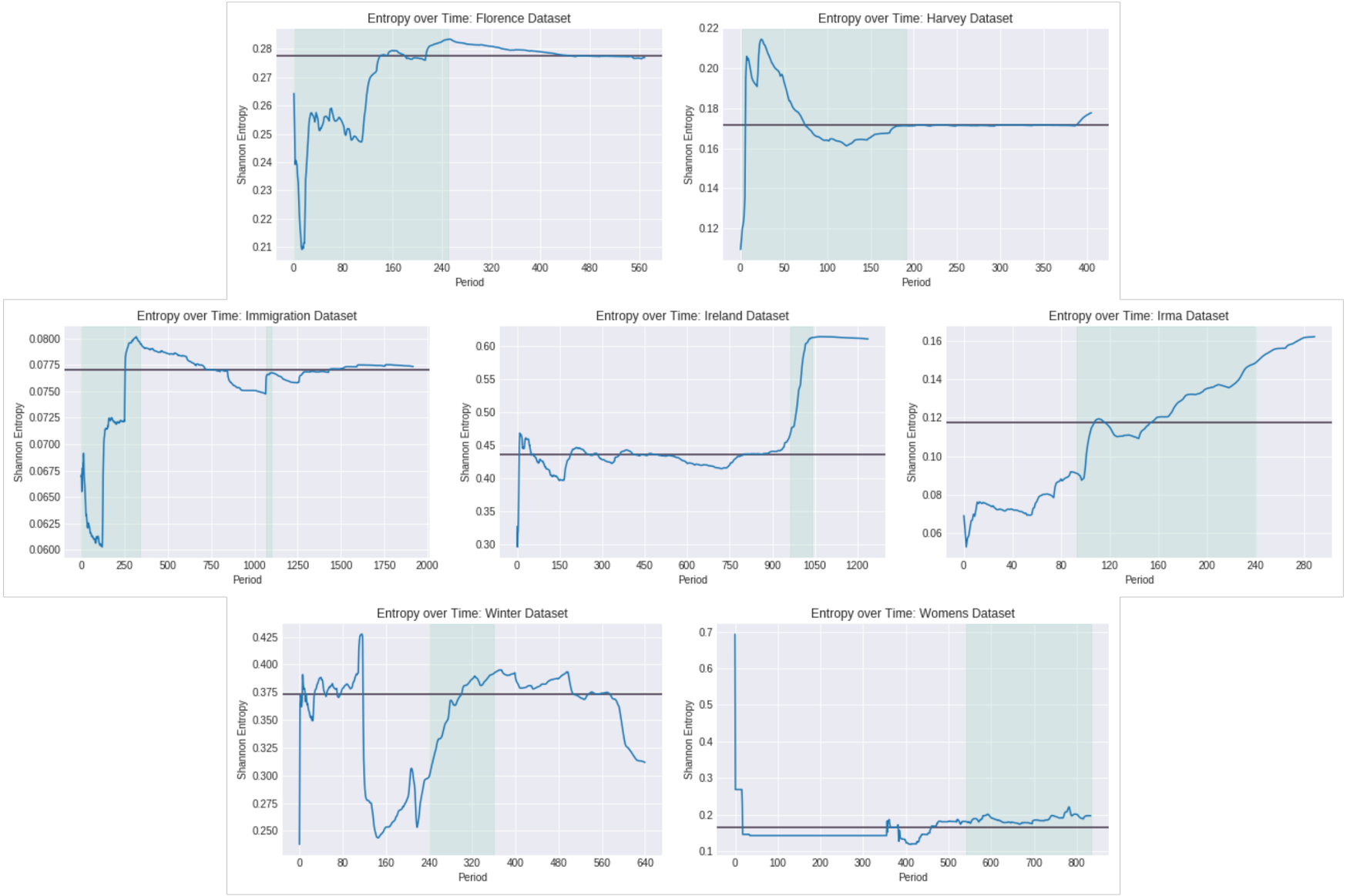


(Identify point where things change ⇒ identify periods which are significant ⇒ agree where non-significant and significant period meet at an identified point)

Engagement over time



Entropy over time

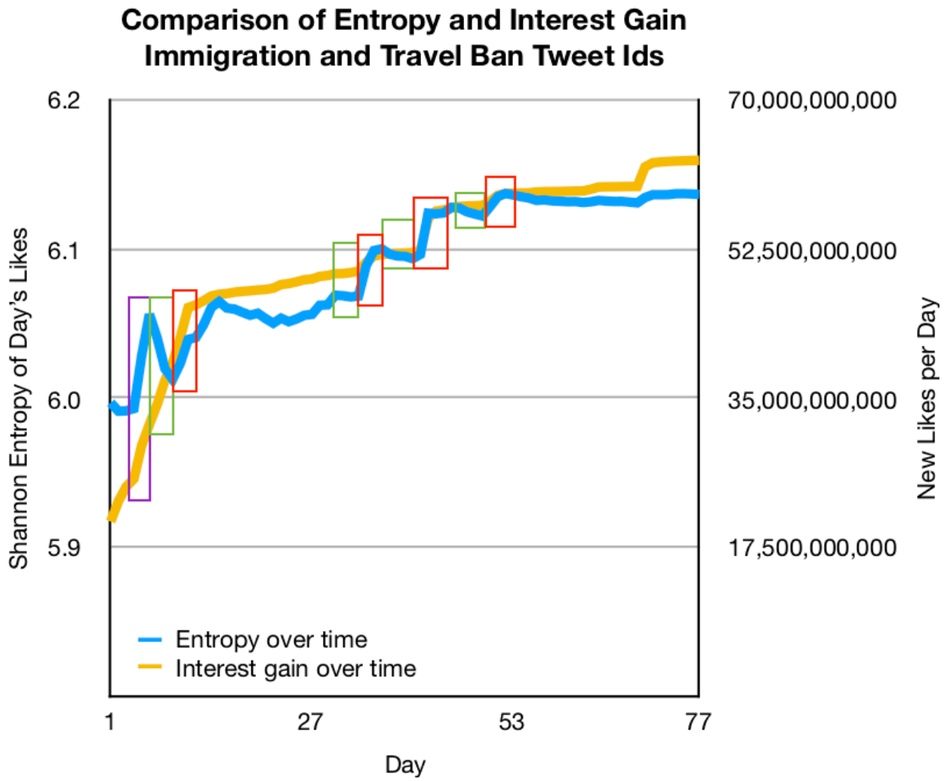


It didn't work.

It didn't work.

(please excuse the **many** plots while I rant briefly about how badly it didn’t work)

Lookback



2016 United States Presidential Election Tweet Ids

Version 3.0

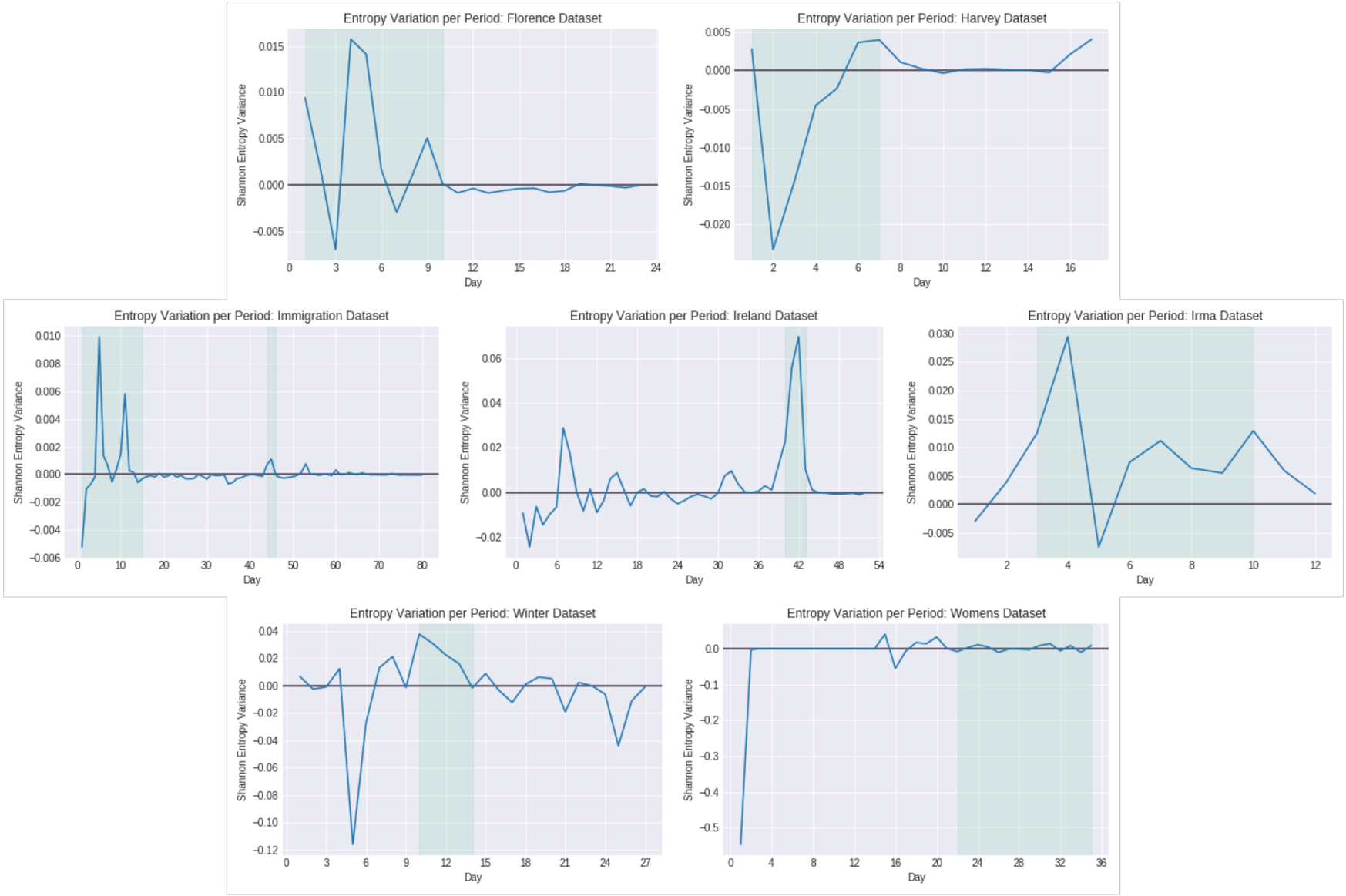
This dataset contains the tweet ids of approximately 280 million tweets related to the 2016 United States presidential election. They were collected between July 13, 2016 and November 10, 2016 from the Twitter API using Social Feed Manager.

These tweet ids are broken up into 12 collections. Each collection was collected either from the GET statuses/user_timeline method of the Twitter REST API or the POST statuses/filter method of the Twitter Stream API. The collections are:

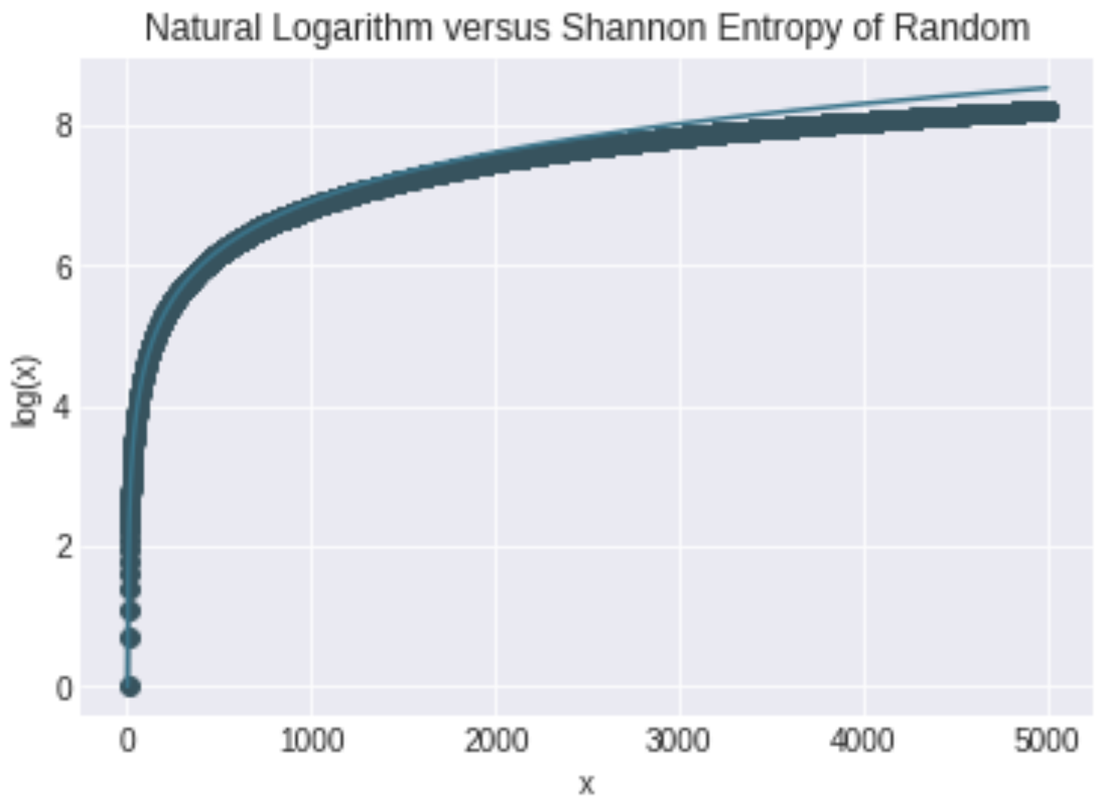
- Candidates and key election hashtags (Twitter filter): election-filter[1-6].txt
- Democratic candidates (Twitter user timeline): democratic-candidate-timelines.txt
- Democratic Convention (Twitter filter): democratic-convention-filter.txt
- Democratic Party (Twitter user timeline): democratic-party-timelines.txt
- Election Day (Twitter filter): election-day.txt
- First presidential debate (Twitter filter): first-debate.txt
- GOP Convention (Twitter filter): republican-convention-filter.txt
- Republican candidates (Twitter user timeline): republican-candidate-timelines.txt
- Republican Party (Twitter user timeline): republican-party-timelines.txt
- Second presidential debate (Twitter filter): second-debate.txt
- Third presidential debate (Twitter filter): third-debate.txt
- Vice Presidential debate (Twitter filter): vp-debate.txt

Entropy **does** spike as trends begin, but only in **half** of the observed sets, probably only because **sudden volume** spikes internal distribution inconsistency, and even then not in a **consistent** or necessarily **statistically significant** way.

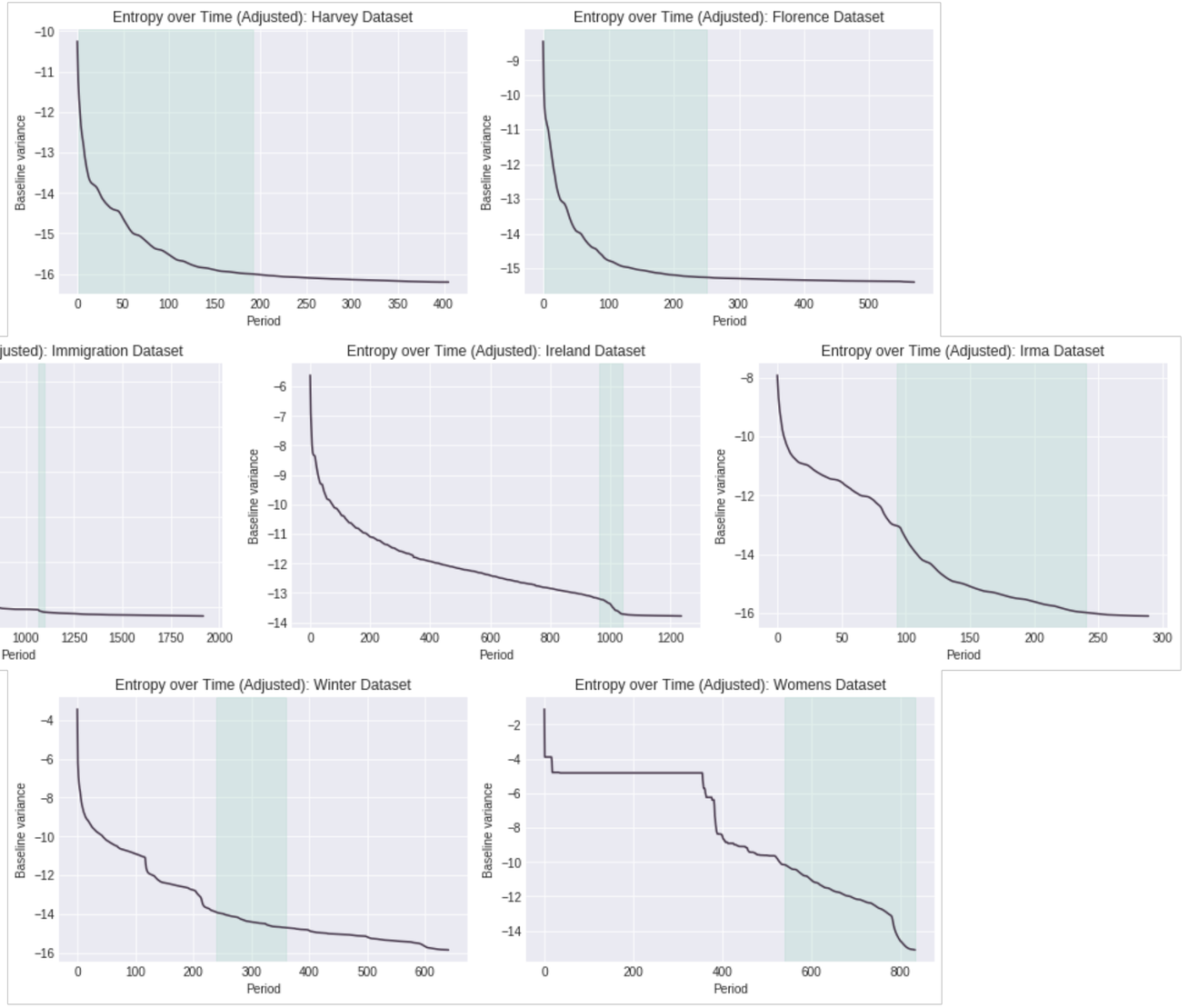
Entropy variation over time



"Hey what if I tried...?"



Entropy baseline variation over time



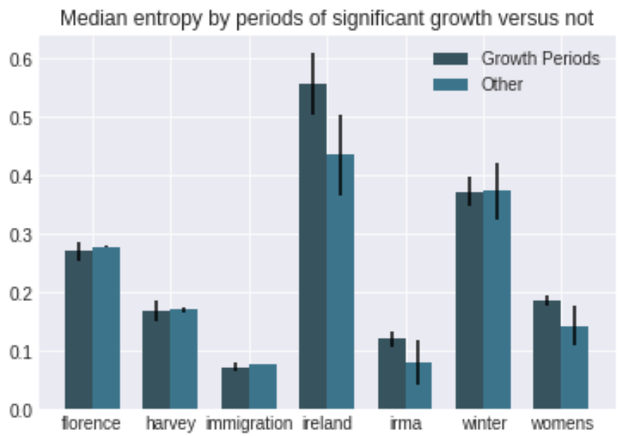
Distribution, median and variance comparisons; shape, curve and significant periods correlation; **partitioned and non-partitioned values, different bucket values, different log values**; values shifted to presumed trend, values shifted to central distribution, values shifted to inverse exponential weight by frequency, values normalised to different bounds, values transformed with all the different ANN node activation functions I knew, values shifted to weekly observed variation from some other paper I found, values masked to remove zero-value observations, values with retweets removed, time-shifted values; **regression analysis, predictive methods like LSTM network sequence prediction**; statistical tests that gave me nonsense answers for literally days before finding that one paper that says "*it only works properly for values of X type in range* $n_1 ... n_2$" that all of statistics as a discipline deigned not to publish anywhere. **Every. Tool. In. My. Statistical. Toolbelt.** Methods that required getting the original files off the server and going back to step one of data pre-processing which takes a day and a half to run, just to change one thing and find the output didn't even change sometimes...

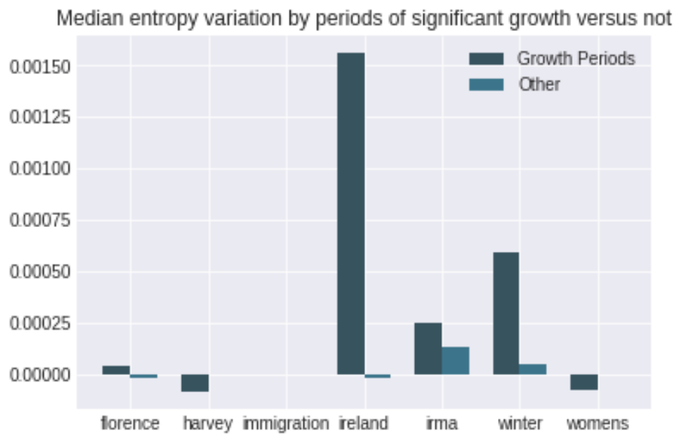
"...if the facts don't fit the theory, change the theory. But all too often it's easier to change the facts."

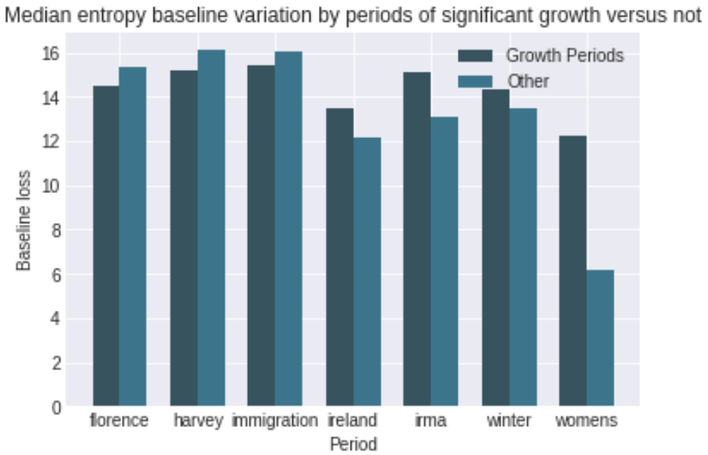
☒ Increased entropy before initial growth

☒ Decreased entropy before trend change

Observed generalisations







→ Entropy **varied significantly** during periods of growth versus not

→ Entropy **varied more** during periods of growth

→ **Which way** variation went was **inconsistent** between datasets

Useful patterns beyond the hypotheses

→ **Why** didn't it work?

→ What does it not working **imply**?

Limitations of pre-existing data

→ Start tracking once trend is detected ⇒ **first period often significant** ⇒ little opportunity to precede growth

→ Stop tracking once plateaued ⇒ **few secondary growth periods** ⇒ little opportunity to precede change

→ Scale not shared between available datasets

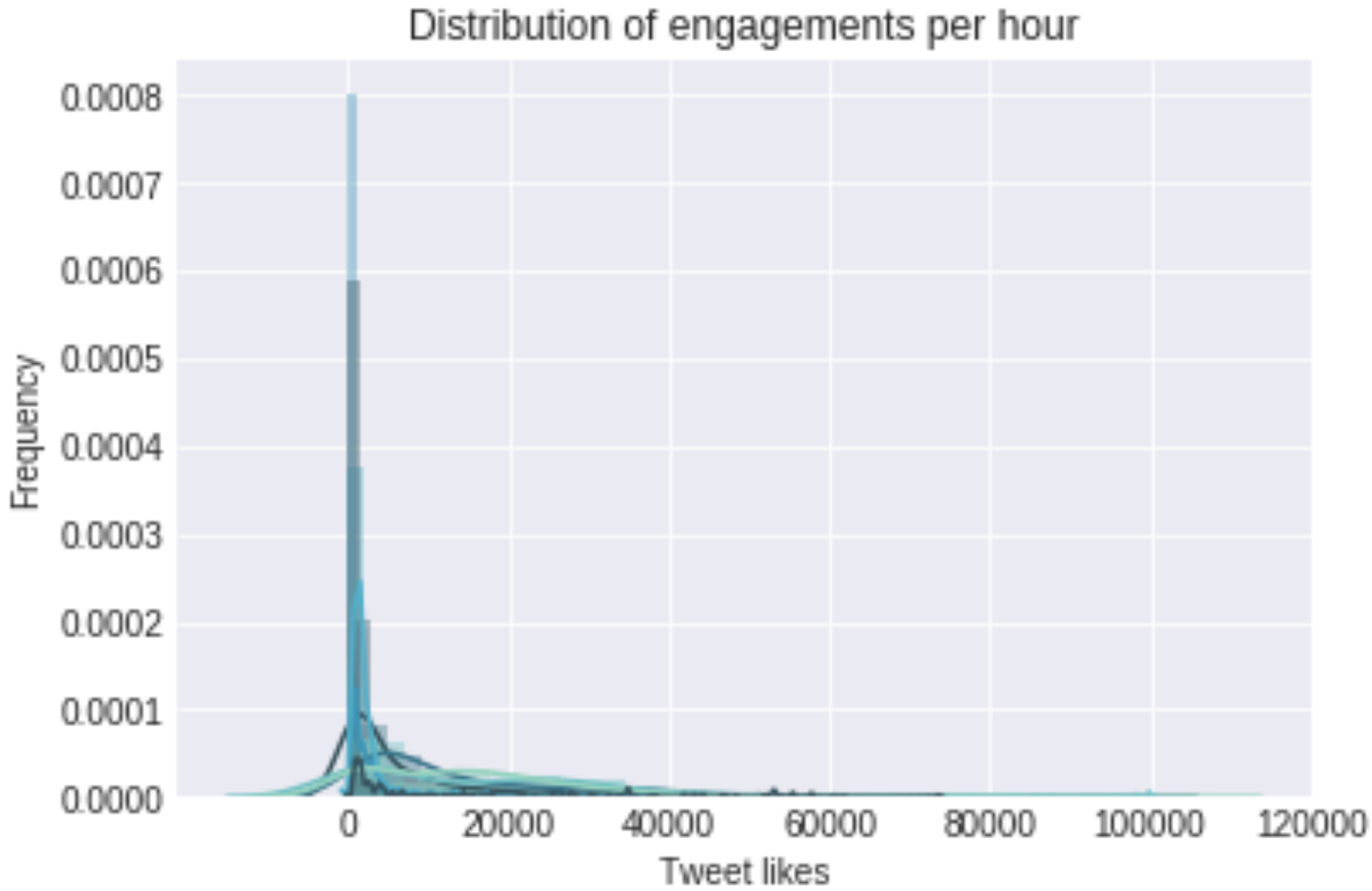
Selected 7 Datasets

→ Initial growth periods not at day zero: **4**

→ Secondary growth periods: **1**

Is this uncommon on the platforms or just in the pre-existing datasets available?

Data was sparse ⇒ dramatically left-skewed



Entropy

→ Range of datapoints

→ Volume of datapoints

→ Internal distribution

Entropy

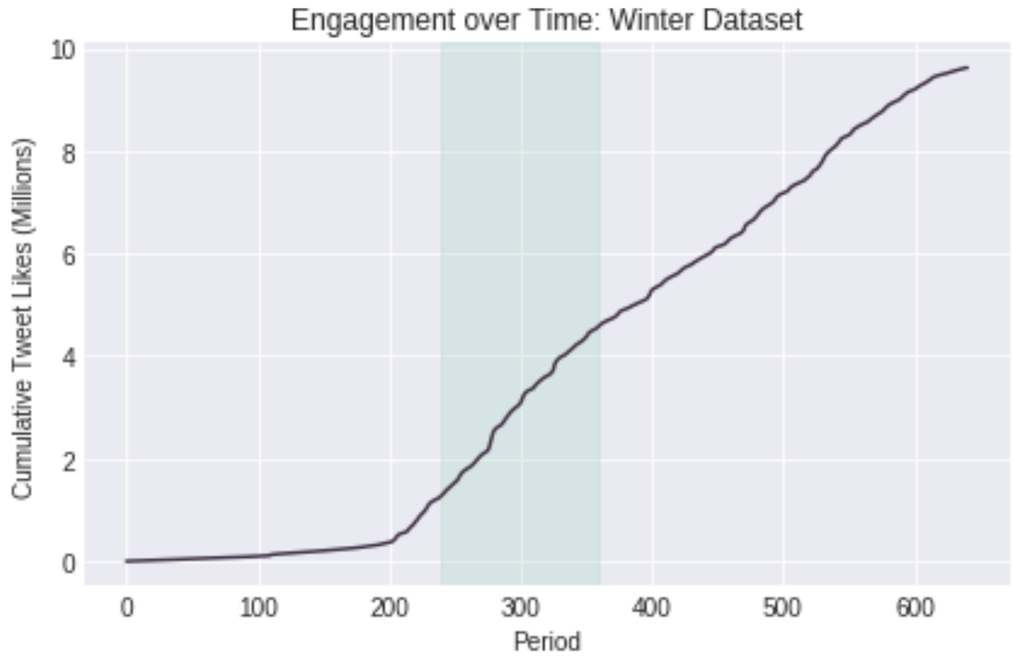
→ ~~Range of datapoints~~ (made redundant)

→ Volume of datapoints (inflated)

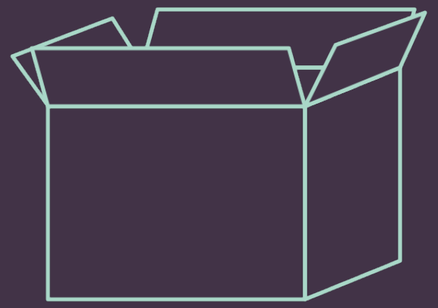
→ Internal distribution (heavily weighted towards range lower bound)

(And this couldn't be accounted for by removing zero-value observations, as it dramatically reduced data volume, skewed activity metrics and made the values being operated on too small to reach significance)

Growth could only be determined in the context of internal significance



(So it falls down when something like this happens, where a human might say **almost all of this** was significant growth)



Limited data volume and options ⇒

No ability to compare with non-trending topics, pre- and post-trend behaviours on a greater scale, etc.

What is known

→ There is still need for **computationally inexpensive methods** of gaining insight from social media data

→ Cross-disciplinary applications still show promise due to many disciplines having their own foundational methods for **complex data summarisation**

→ Entropy Theory is potentially not the solution here, unless **new procedures for data preparation** are developed that negate the skewing effect of sparse data and low medians

What this study contributed

→ A whole bunch of nicely documented Python code, bundled up in an appropriately-licensed **Open Source software** bundle released for public use

→ Knowledge of many things that do not work that now don't need to be tried in the future

→ More examples for future work to cite, justifying interest in analogous methods for social media data analysis or cross-disciplinary application of Entropy Theory

→ A conference talk at PyCon Australia that has already **sparked similar work** in the Australian data science community

What experimentation should try next

→ Performance may differ in application to **live data** analysis as opposed to retrospective

→ Entropy-based analysis may be more useful in **early-stage trend detection** as opposed to the more nebulous ongoing trend behaviour prediction

→ Incorporation of **additional information** (if possible) than that available via the Twitter platform API

→ Applicability to **other platforms** that may exhibit different content consumption and sharing behaviours


Thanks for listening!

References

[1] Hootsuite: 2018, Hootsuite's social media barometer report. https://hootsuite.com/resources/all-future*of*social-digital*in*2019-glo-en-ca-digital*in*2019-q1_2019 [Accessed 18 May 2018].
[2] Perrin, A.: 2015, Social media usage: 2005-2015, *PEW Research Center Report*.
[3] Kasemsap, K.: 2019, Professional and business applications of social media platforms, Social Entrepreneurship: Concepts, Methodologies, Tools, and Applications, *IGI Global*, pp. 824– 847.
[4] Zhao, X., Lampe, C. and Ellison, N. B.: 2016, The social media ecology: User perceptions, strategies and challenges, *Proceedings of the 2016 CHI conference on human factors in computing systems*, ACM, pp. 89–100.
[5] Kane, G. C.: 2015, Enterprise social media: Current capabilities and future possibilities., *MIS Quarterly Executive* 14(1).
[6] Kaplan, A. M. and Haenlein, M.: 2010, Users of the world, unite! the challenges and opportunities of social media, *Business horizons* 53(1), 59–68.
[7] Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G. and Westergren, M.: 2012, The social economy: Unlocking value and productivity through social technologies. https://www.mckinsey.com/industries/high-tech/our-insights/ the-social-economy [Accessed 18 May 2019].
[8] Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J. and Seymour, T.: 2011, The history of social media and its impact on business, *Journal of Applied Management and entrepreneurship* 16(3), 79–91.
[9] He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G. and Tao, R.: 2015, Gaining competitive intelligence from social media data: evidence from two largest retail chains in the world, *Industrial Management & Data Systems* 115(9), 1622–1636.
[10] Brooker, P., Barnett, J., Cribbin, T. and Sharma, S.: 2016, Have we even solved the first 'big data challenge?'practical issues concerning data collection and visual representation for social media analytics, *Digital methods for social science*, Springer, pp. 34–50.
[11] Mayeh, M., Scheepers, R. and Valos, M.: 2012, Understanding the role of social media monitor- ing in generating external intelligence, ACIS 2012: Location, location, location: *Proceedings of the 23rd Australasian Conference on Information Systems 2012*, ACIS, pp. 1–10.
[12] Halasz, C. M.: 2019, Optimizing training for sparse workloads in Tensorflow. Reinforce AI Conference. **URL**: https://reinforceconf.com/speaker/CibeleMontezHalasz
[13] Dou, W., Wang, X., Skau, D., Ribarsky, W. and Zhou, M. X.: 2012, Leadline: Interactive visual analysis of text data through event identification and exploration, *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, pp. 93–102.
[14] Hogan, B.: 2016, Social media giveth, social media taketh away: Facebook, friendships, and apis, *International Journal of Communication*, Forthcoming.
[15] Weller, K. and Kinder-Kurlanda, K. E.: 2015, Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research?, *Ninth International AAAI Conference on Web and Social Media*.

References

[16] Altshuler, Y., Pan, W. and Pentland, A. S.: 2012, Trends prediction using social diffusion models, International Conference on Social Computing, *Behavioral-Cultural Modeling and Prediction*, Springer, pp. 97–104.
[17] Sapountzi, A. and Psannis, K. E.: 2018, Social networking data analysis tools & challenges, *Future Generation Computer Systems* 86, 893–913.
[18] Figueiredo, F., Almeida, J. M., Gon¸calves, M. A. and Benevenuto, F.: 2016, Trendlearner: Early prediction of popularity trends of user generated content, *Information Sciences* 349, 172–187.
[19] Qian, S., Zhang, T., Xu, C. and Shao, J.: 2015, Multi-modal event topic model for social event analysis, *IEEE transactions on multimedia 18(2)*, 233–246.
[20] Manovich, L.: 2011, Trending: The promises and the challenges of big social data, *Debates in the digital humanities* 2, 460–475.
[21] Schroeder, R.: 2014, Big data and the brave new world of social media research, *Big Data & Society* 1(2).
[22] Sloan, L. and Quan-Haase, A.: 2017, The SAGE handbook of social media research methods, Sage.
[23] Adar, E. and Adamic, L. A.: 2005, Tracking information epidemics in blogspace, Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence, *IEEE Computer Society*, pp. 207–214.
[24] Gomez-Rodriguez, M., Leskovec, J. and Krause, A.: 2012, Inferring networks of diffusion and influence, *ACM Transactions on Knowledge Discovery from Data* (TKDD) 5(4).
[25] Cannarella, J. and Spechler, J. A.: 2014, Epidemiological modeling of online social network dynamics, *arXiv preprint arXiv:1401.4208*.
[26] Chang, H.-C.: 2010, A new perspective on twitter hashtag use: Diffusion of innovation theory, *Proceedings of the American Society for Information Science and Technology* 47(1), 1–4.
[27] Jamieson, J. and Boase, J.: 2017, Listening to social rhythms: Exploring logged interactional data through sonification, *The SAGE Handbook of Social Media Research Methods*.
[28] Liu, F., Wang, L., Johnson, H. and Zhao, H.: 2015, Analysis of network trust dynamics based on the evolutionary game, Scientia Iranica. Transaction E, *Industrial Engineering* 22(6).
[29] Schmidt, C. W.: 2012, Trending now: using social media to predict and track disease outbreaks.
[30] Zimmer, M. and Proferes, N. J.: 2014, A topology of twitter research: disciplines, methods, and ethics, *ASLIB Journal of Information Management* 66(3), 250–261.

References

[31] Milligan, I., Ruest, N. and Lin, J.: 2016, Content selection and curation for web archiving: The gatekeepers vs. the masses, *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, ACM, pp. 107–110.
[32] Ruest, N. and Milligan, I.: 2016, An open-source strategy for documenting events: The case study of the 42nd canadian federal election on twitter, *Code4Lib* 32.
[33] Juanals, B. and Minel, J.-L.: 2017, Analysing cultural events on twitter, *International Conference on Computational Collective Intelligence*, Springer, pp. 376–385.
[34] Palmer, A., Robinson, M. and Phillips, K. K.: 2017, Illegal is not a noun: Linguistic form for detection of pejorative nominalizations, *Proceedings of the First Workshop on Abusive Language Online*, pp. 91–100.
[35] Pinter, A. T., Goldman, B. and Novotny, E.: 2017, Pennsylvania perspectives of the 2016 election: A project to collect web and social media content around significant societal events, *Pennsylvania Libraries: Research & Practice* 5(2), 96–106.
[36] Aruguete, N. and Calvo, E.: 2018, Time to #protest: Selective exposure, cascading activation, and framing in social media, *Journal of Communication* 68(3), 480–502.
[37] Darwish, K.: 2018, To kavanaugh or not to kavanaugh: That is the polarizing question, *arXiv preprint arXiv:1810.06687* p. 01.
[38] Kalmar, I., Stevens, C. and Worby, N.: 2018, Twitter, gab, and racism: the case of the soros myth, *Proceedings of the 9th International Conference on Social Media and Society*, ACM, pp. 330–334.
[39] Mahbub, M.S., de Souza, P. and Williams, R., 2017. Describing environmental phenomena variation using entropy theory. *International Journal of Data Science and Analytics*, 3(1), pp.49-60.

Insights into Social Media Data:

a new formalism inspired in Thermodynamics

See **github.com/TheMartianLife/Honours-Presentation**