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

Agenda


WHY?
Background

Social Media Usage

- Of the world's ~4.5 billion internet users, almost 3.5 billion use social media[1]
- From 2005-2015 social media usage rose tenfold in the United States 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 over a 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]

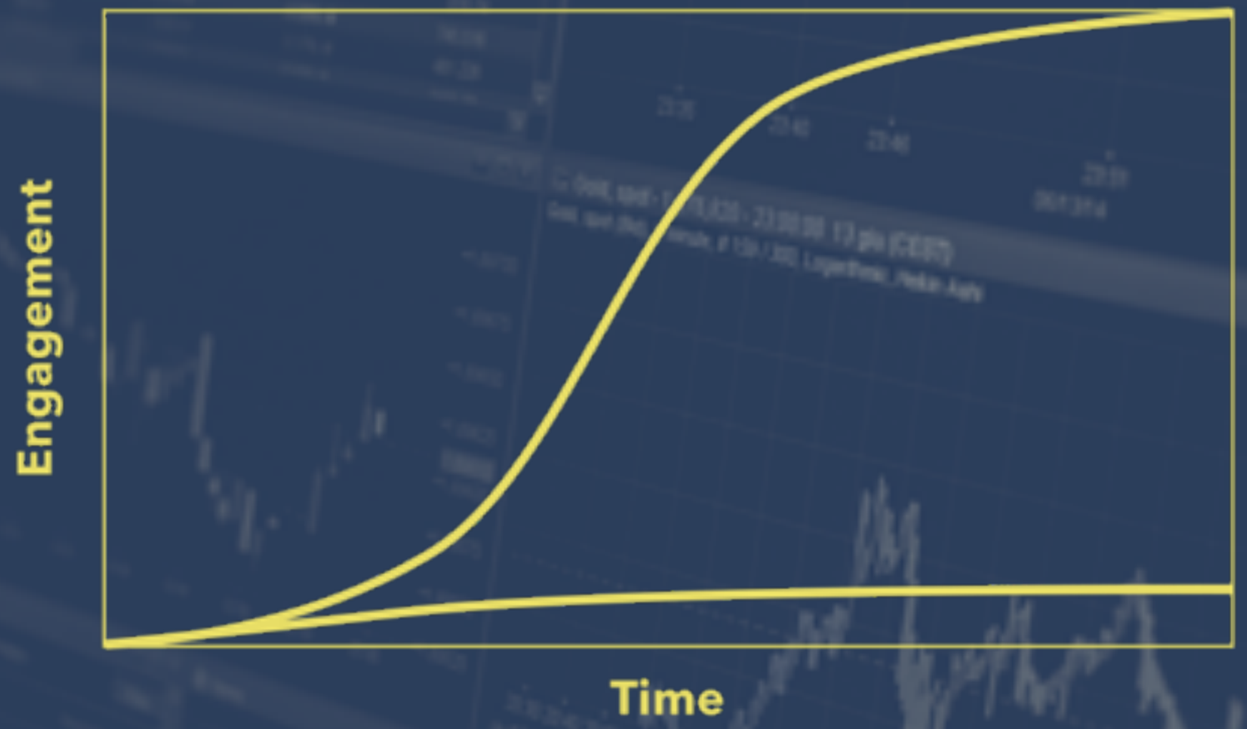
Social Media Analysis

- 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 speed[10]
- **Semi-structured** data with **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



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? 🤔

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**)

Analogous Application

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]

Entropy Theory

Boltzmann's $S = k \ln(N)$

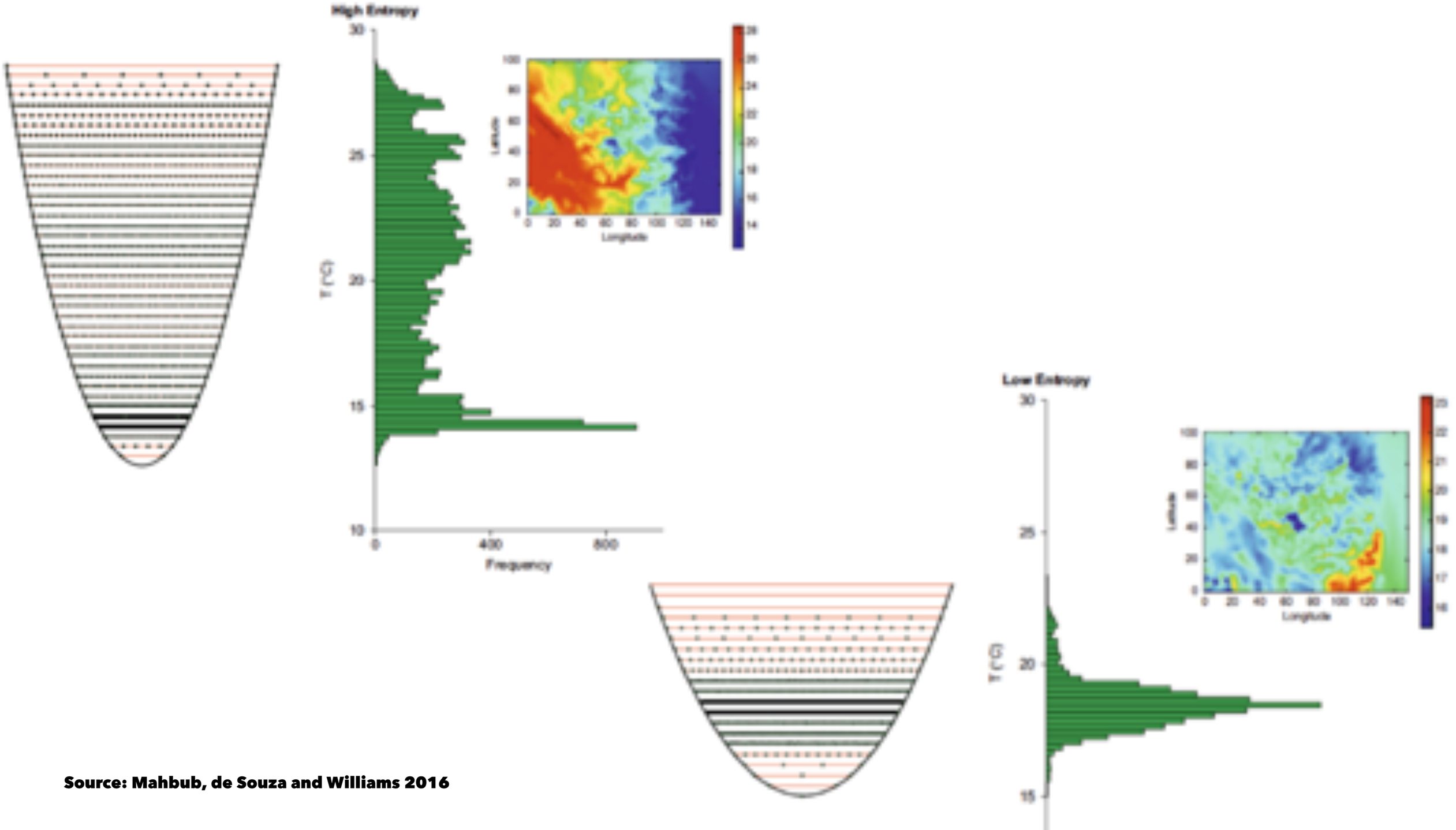
Upper bound of entropy in closed system at equilibrium

Gibb's $S = -k \sum_i P_i \ln P_i$

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



Source: Mahbub, de Souza and Williams 2016

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

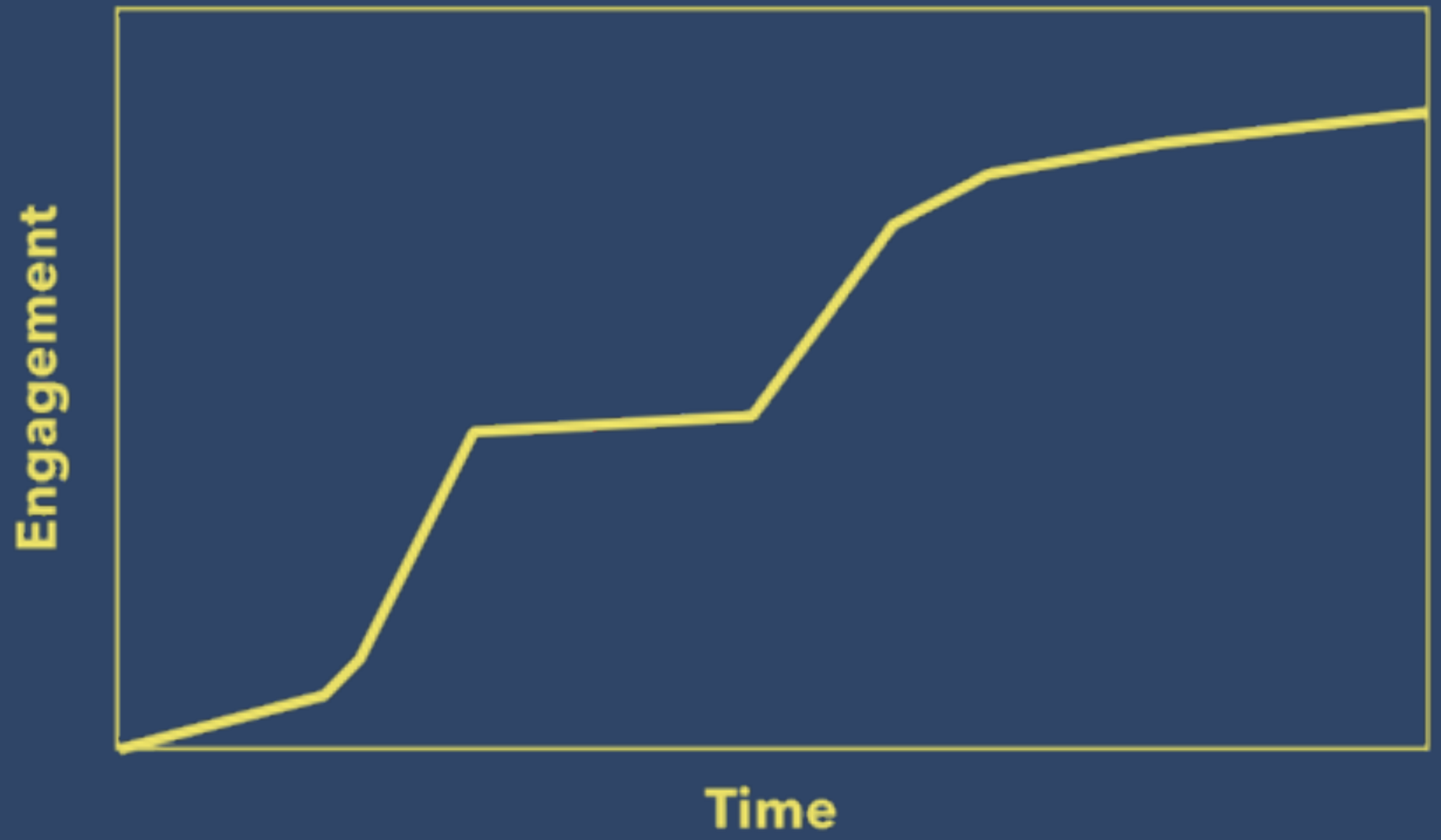
Broadly

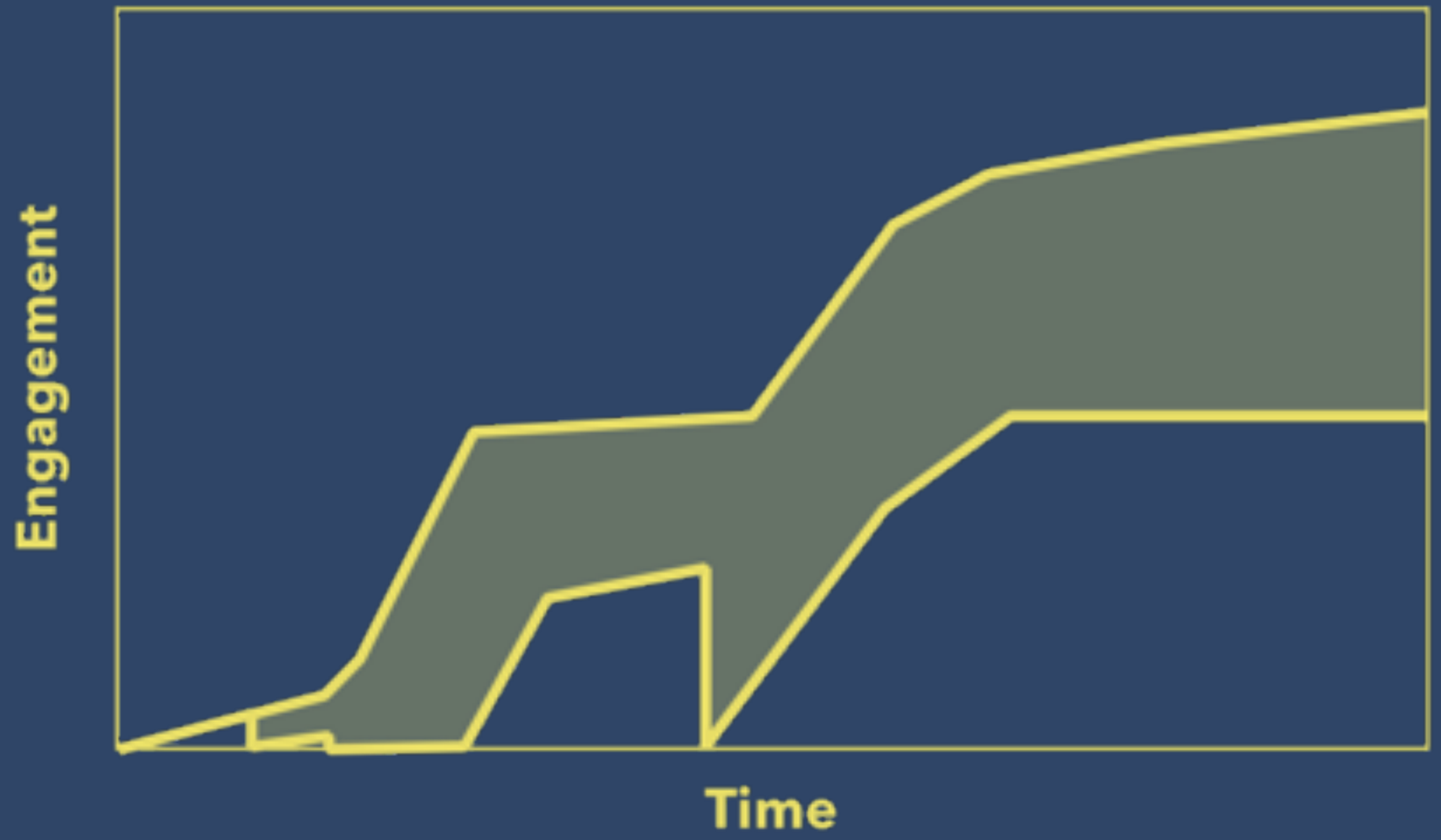
1. Can the analogous application of **Entropy Theory** present useful information about the lifetime of a trend or topic on a Social Media platform?
2. Is there indication that the method could be applied more broadly, or adapted to be **predictive**?

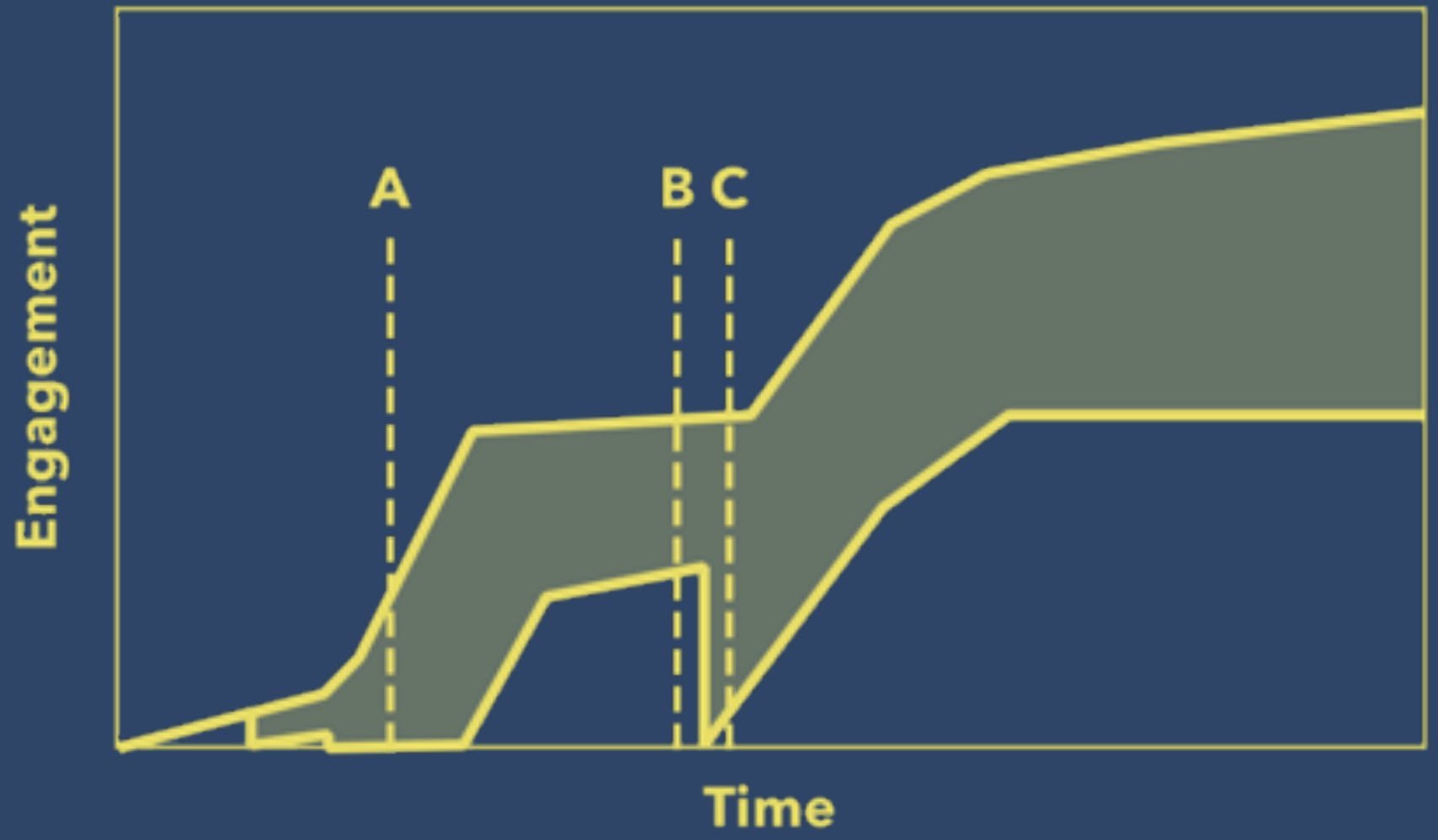
In this study

Experimentation aims to confirm the dual hypotheses as follows:

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.


Engagement
Time


Engagement
Time


A
B C
Engagement
Time

Potential commercial value

If we can predict changes in topic growth over time, this may aid in early **detection of trends**.

This may allow parties to release information at points of predicted **optimal reach** or **engagement potential**.

Potential social research value

Assuming we can predict trend decay over time if no new stimulus is received, we can identify points where new external stimulus probably was received, even where it is not evident in the platform content.

This can help identify new context or **anomalous trends**.

HOW?

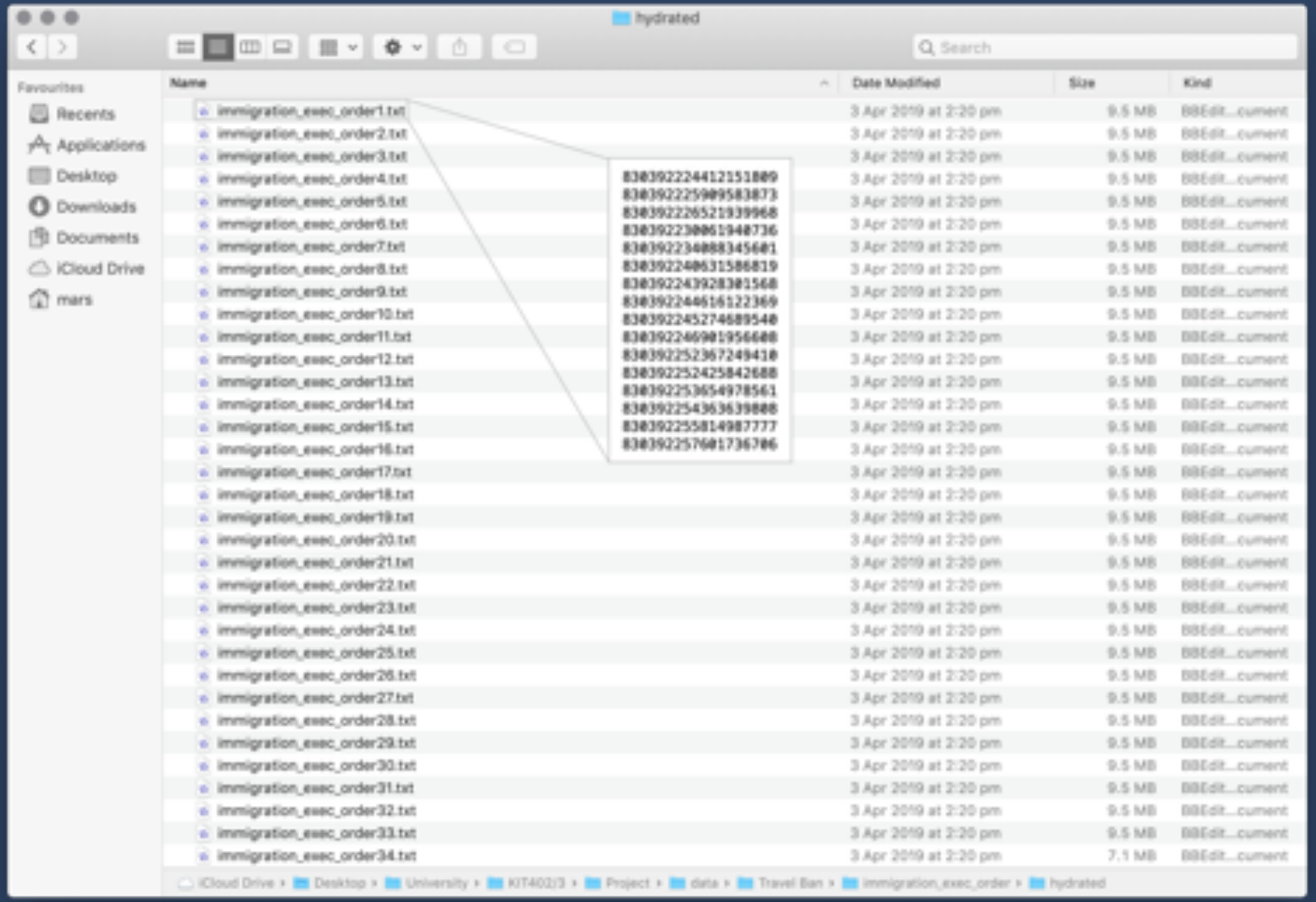
Research Design

Datasets needed for this study

1. Representing a **single** cohesive topic growing over time
2. Representing a higher-level topic with **multiple** sub-topics or evolution over time

Why 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*

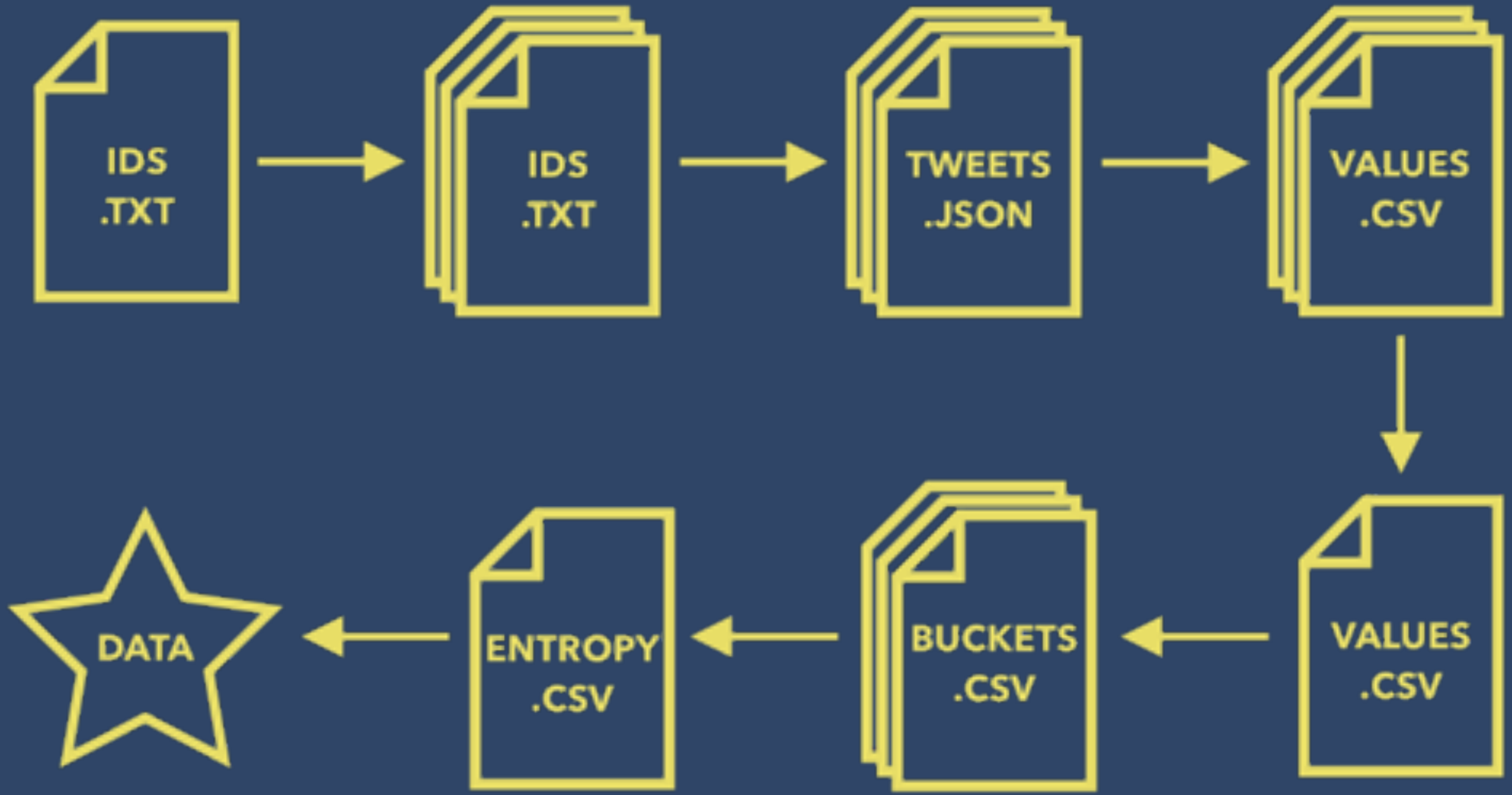

hydrated
Search
Favourites
Recents
Applications
Desktop
Downloads
Documents
iCloud Drive
mars
Name
Date Modified
Size
Kind
immigration_exec_order1.txt
immigration_exec_order2.txt
immigration_exec_order3.txt
immigration_exec_order4.txt
immigration_exec_order5.txt
immigration_exec_order6.txt
immigration_exec_order7.txt
immigration_exec_order8.txt
immigration_exec_order9.txt
immigration_exec_order10.txt
immigration_exec_order11.txt
immigration_exec_order12.txt
immigration_exec_order13.txt
immigration_exec_order14.txt
immigration_exec_order15.txt
immigration_exec_order16.txt
immigration_exec_order17.txt
immigration_exec_order18.txt
immigration_exec_order19.txt
immigration_exec_order20.txt
immigration_exec_order21.txt
immigration_exec_order22.txt
immigration_exec_order23.txt
immigration_exec_order24.txt
immigration_exec_order25.txt
immigration_exec_order26.txt
immigration_exec_order27.txt
immigration_exec_order28.txt
immigration_exec_order29.txt
immigration_exec_order30.txt
immigration_exec_order31.txt
immigration_exec_order32.txt
immigration_exec_order33.txt
immigration_exec_order34.txt
3 Apr 2019 at 2:20 pm
9.5 MB
7.1 MB
BBEdit...cument
8303922244121551809
8303922259095838737
8303922265219399688
8303922300619487367
8303922340888345601
8303922406315868197
8303922439283815688
8303922446161223697
8303922452746885407
8303922469019566088
8303922523672494107
8303922524258426887
8303922536549785617
8303922543636398887
8303922558149877777
8303922576017367067
iCloud Drive
Desktop
University
KIT402/3
Project
data
Travel Ban
immigration_exec_order
hydrated

Tweet JSON


```
{
   "created_at":"Fri Feb 10 10:05:45 2017",
   "id_str":"829994567841062 4",
   "full_text": "This is a fake Tweet.",
   "user":{
      "id_str": "972937630",
      "name": "John Doe",
      "screen_name": "johndoe",
      "protected": false,
      "verified": false,
   },
   "retweet_count": 167,
   "favorite_count": 50
}
```


"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]
- lots of visualisation and analysis tools that work with it


IDS
.TXT
IDS
.TXT
TWEETS
.JSON
VALUES
.CSV
VALUES
.CSV
BUCKETS
.CSV
ENTROPY
.CSV
DATA

File	IDs	Hydrated	Lost
immigration-exec-order1	500,000	342,551	157,449
immigration-exec-order2	500,000	341,873	158,127
immigration-exec-order3	500,000	344,425	155,575
immigration-exec-order4	500,000	375,875	124,125
immigration-exec-order5	500,000	364,173	135,827
immigration-exec-order6	500,000	364,605	135,395
immigration-exec-order7	500,000	332,024	167,976
...	...	...	...
Total	16,875,766	11,776,307	5,099,459

seemingly random yet consistent distribution of missing datapoints

=

suitably random self-selecting population sample 👍

Analysis Phase 1: Exploration

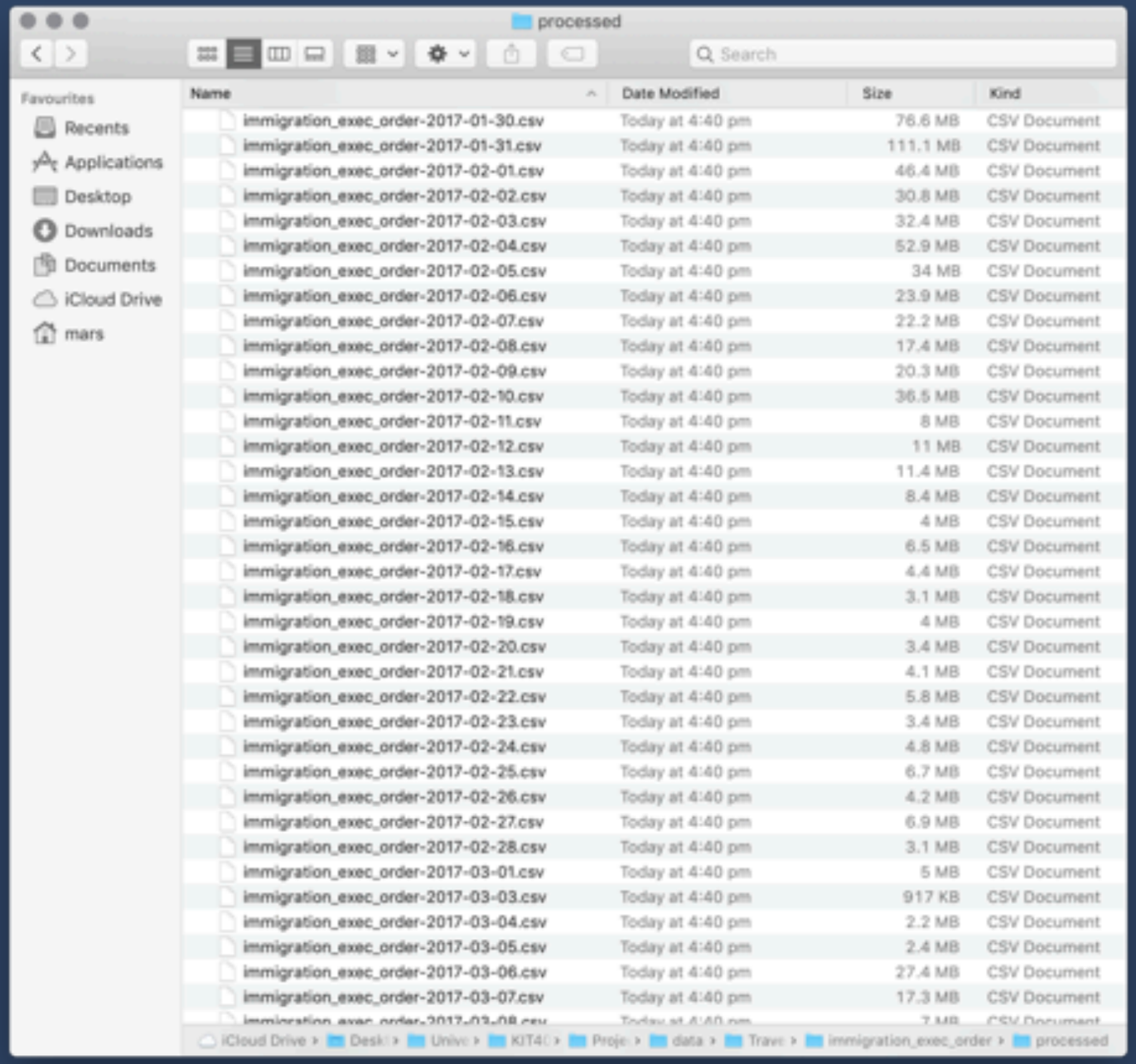
- Establish "ground truth" of trend behaviour over time portrayed in each dataset
- Decide on appropriate granularity for discretisation of data (or justify use of full range of values)
- Explore descriptive statistics to establish theoretical bounds
- Develop strategies for dealing with missing data and edge cases

⚠️ Issues with historical data

1. Shared tweets show no information about whether likes were given to original or share
2. Engagement metrics are associated with time of posting, even though they would have occurred later and over time

It is assumed that the presence of both effects in **both the test and comparison data** will nullify the effect.

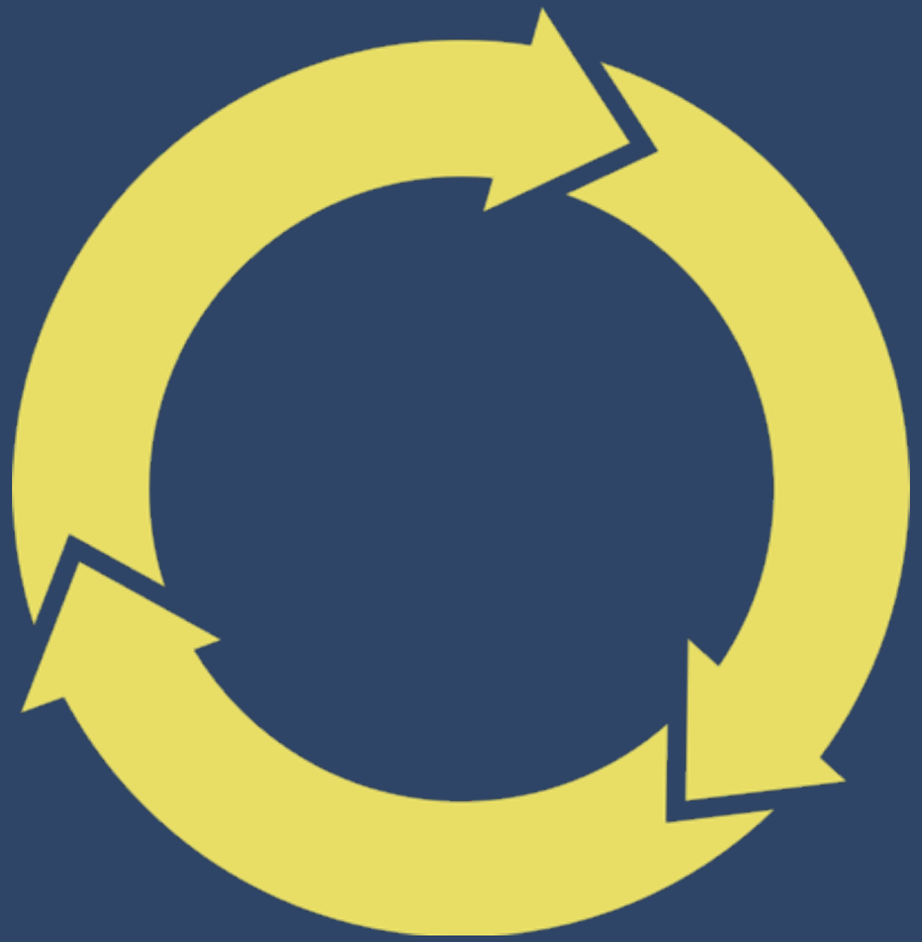
1994. II
2001 II


processed
Search
Favourites
Recents
Applications
Desktop
Downloads
Documents
iCloud Drive
mars
Name
Date Modified
Size
Kind
immigration_exec_order-2017-01-30.csv Today at 4:40 pm 76.6 MB CSV Document
immigration_exec_order-2017-01-31.csv Today at 4:40 pm 111.1 MB CSV Document
immigration_exec_order-2017-02-01.csv Today at 4:40 pm 46.4 MB CSV Document
immigration_exec_order-2017-02-02.csv Today at 4:40 pm 30.8 MB CSV Document
immigration_exec_order-2017-02-03.csv Today at 4:40 pm 32.4 MB CSV Document
immigration_exec_order-2017-02-04.csv Today at 4:40 pm 52.9 MB CSV Document
immigration_exec_order-2017-02-05.csv Today at 4:40 pm 34 MB CSV Document
immigration_exec_order-2017-02-06.csv Today at 4:40 pm 23.9 MB CSV Document
immigration_exec_order-2017-02-07.csv Today at 4:40 pm 22.2 MB CSV Document
immigration_exec_order-2017-02-08.csv Today at 4:40 pm 17.4 MB CSV Document
immigration_exec_order-2017-02-09.csv Today at 4:40 pm 20.3 MB CSV Document
immigration_exec_order-2017-02-10.csv Today at 4:40 pm 36.5 MB CSV Document
immigration_exec_order-2017-02-11.csv Today at 4:40 pm 8 MB CSV Document
immigration_exec_order-2017-02-12.csv Today at 4:40 pm 11 MB CSV Document
immigration_exec_order-2017-02-13.csv Today at 4:40 pm 11.4 MB CSV Document
immigration_exec_order-2017-02-14.csv Today at 4:40 pm 8.4 MB CSV Document
immigration_exec_order-2017-02-15.csv Today at 4:40 pm 4 MB CSV Document
immigration_exec_order-2017-02-16.csv Today at 4:40 pm 6.5 MB CSV Document
immigration_exec_order-2017-02-17.csv Today at 4:40 pm 4.4 MB CSV Document
immigration_exec_order-2017-02-18.csv Today at 4:40 pm 3.1 MB CSV Document
immigration_exec_order-2017-02-19.csv Today at 4:40 pm 4 MB CSV Document
immigration_exec_order-2017-02-20.csv Today at 4:40 pm 3.4 MB CSV Document
immigration_exec_order-2017-02-21.csv Today at 4:40 pm 4.1 MB CSV Document
immigration_exec_order-2017-02-22.csv Today at 4:40 pm 5.8 MB CSV Document
immigration_exec_order-2017-02-23.csv Today at 4:40 pm 3.4 MB CSV Document
immigration_exec_order-2017-02-24.csv Today at 4:40 pm 4.8 MB CSV Document
immigration_exec_order-2017-02-25.csv Today at 4:40 pm 6.7 MB CSV Document
immigration_exec_order-2017-02-26.csv Today at 4:40 pm 4.2 MB CSV Document
immigration_exec_order-2017-02-27.csv Today at 4:40 pm 6.9 MB CSV Document
immigration_exec_order-2017-02-28.csv Today at 4:40 pm 3.1 MB CSV Document
immigration_exec_order-2017-03-01.csv Today at 4:40 pm 5 MB CSV Document
immigration_exec_order-2017-03-03.csv Today at 4:40 pm 917 KB CSV Document
immigration_exec_order-2017-03-04.csv Today at 4:40 pm 2.2 MB CSV Document
immigration_exec_order-2017-03-05.csv Today at 4:40 pm 2.4 MB CSV Document
immigration_exec_order-2017-03-06.csv Today at 4:40 pm 27.4 MB CSV Document
immigration_exec_order-2017-03-07.csv Today at 4:40 pm 17.3 MB CSV Document
iCloud Drive › Desk › Univ › KIT4 › Proje › data › Trave › immigration_exec_order › processed

```
calculate_entropy(data)
```

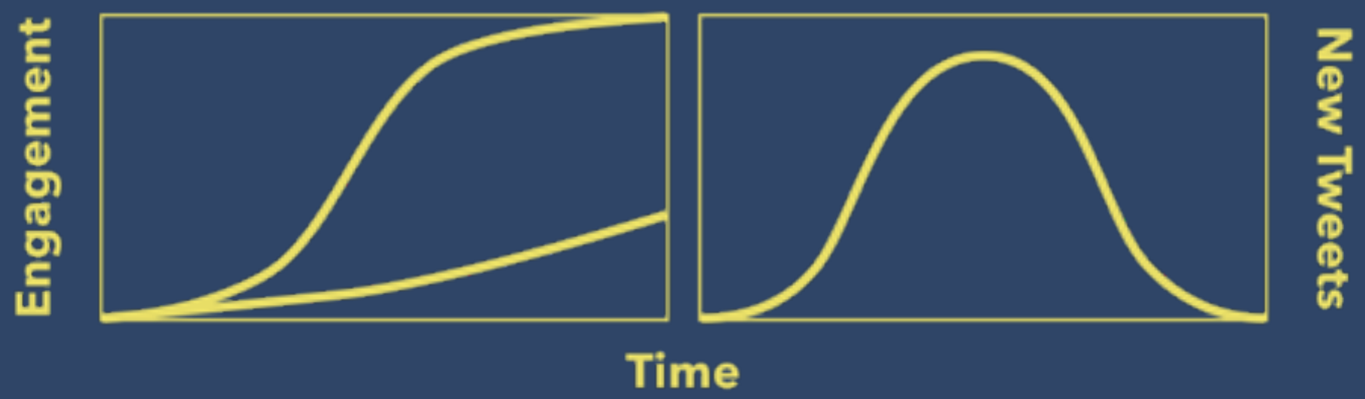
More Analysis!
Phase 2: Discrete topic
&
Phase 3: Evolving topic

Iterative experimentation



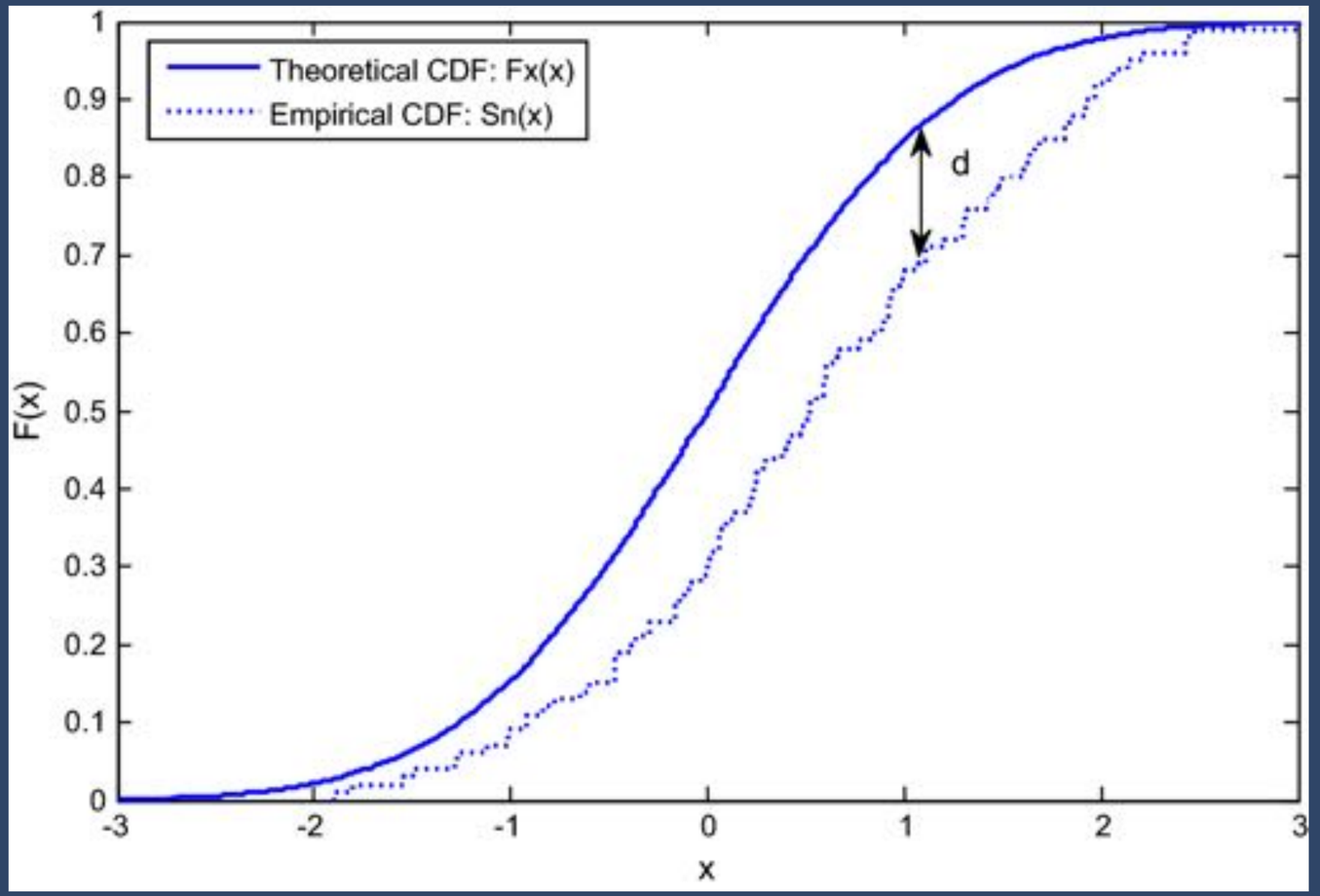

Inferential analysis =
dynamic + reponsive

Predictions

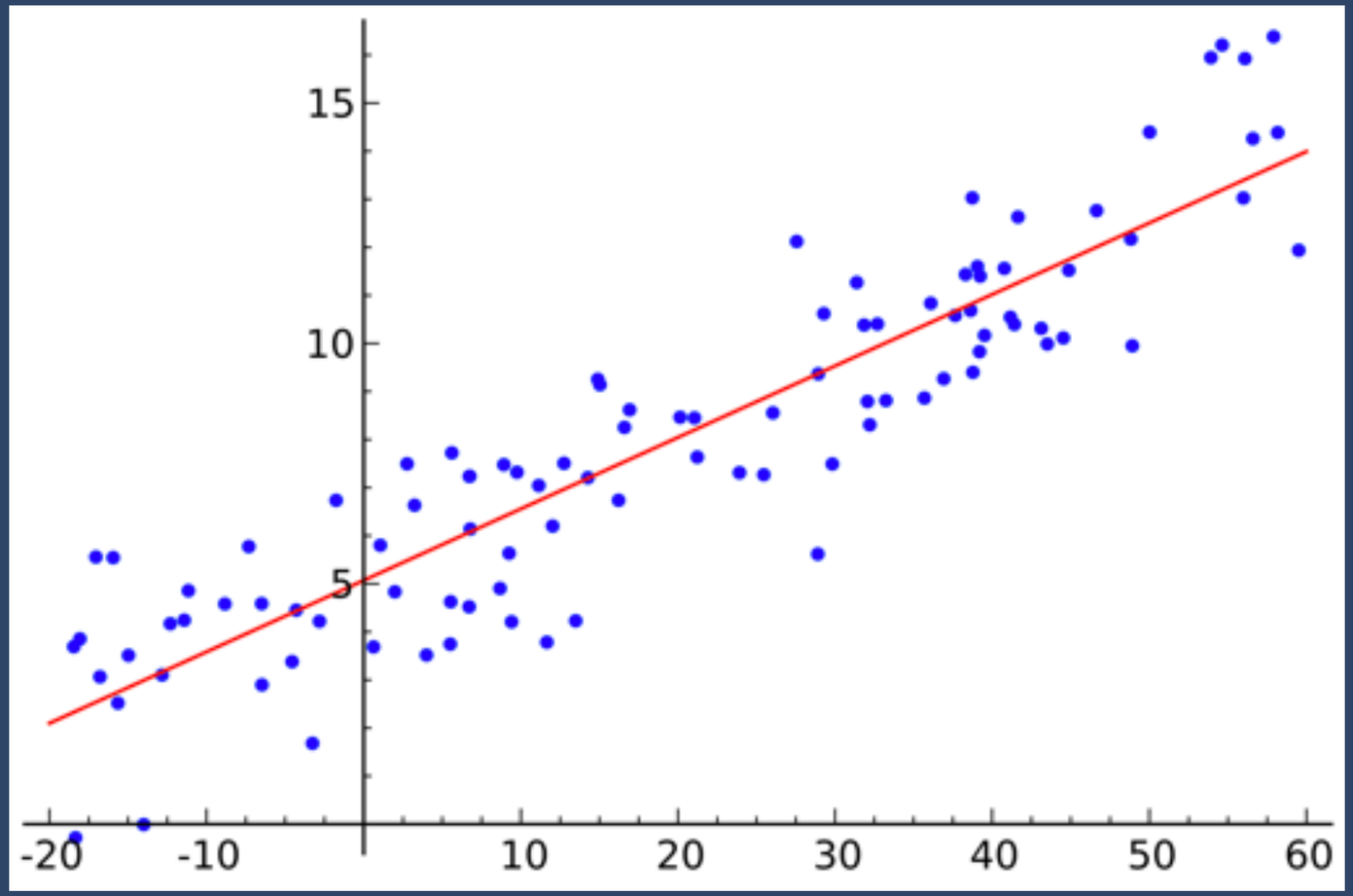


- Normal-ish distribution of new tweets
- Varying loglinear to exponential growth rates in engagement
- Engagements between 0 and ~500,000
- Clear topic plateau point towards end

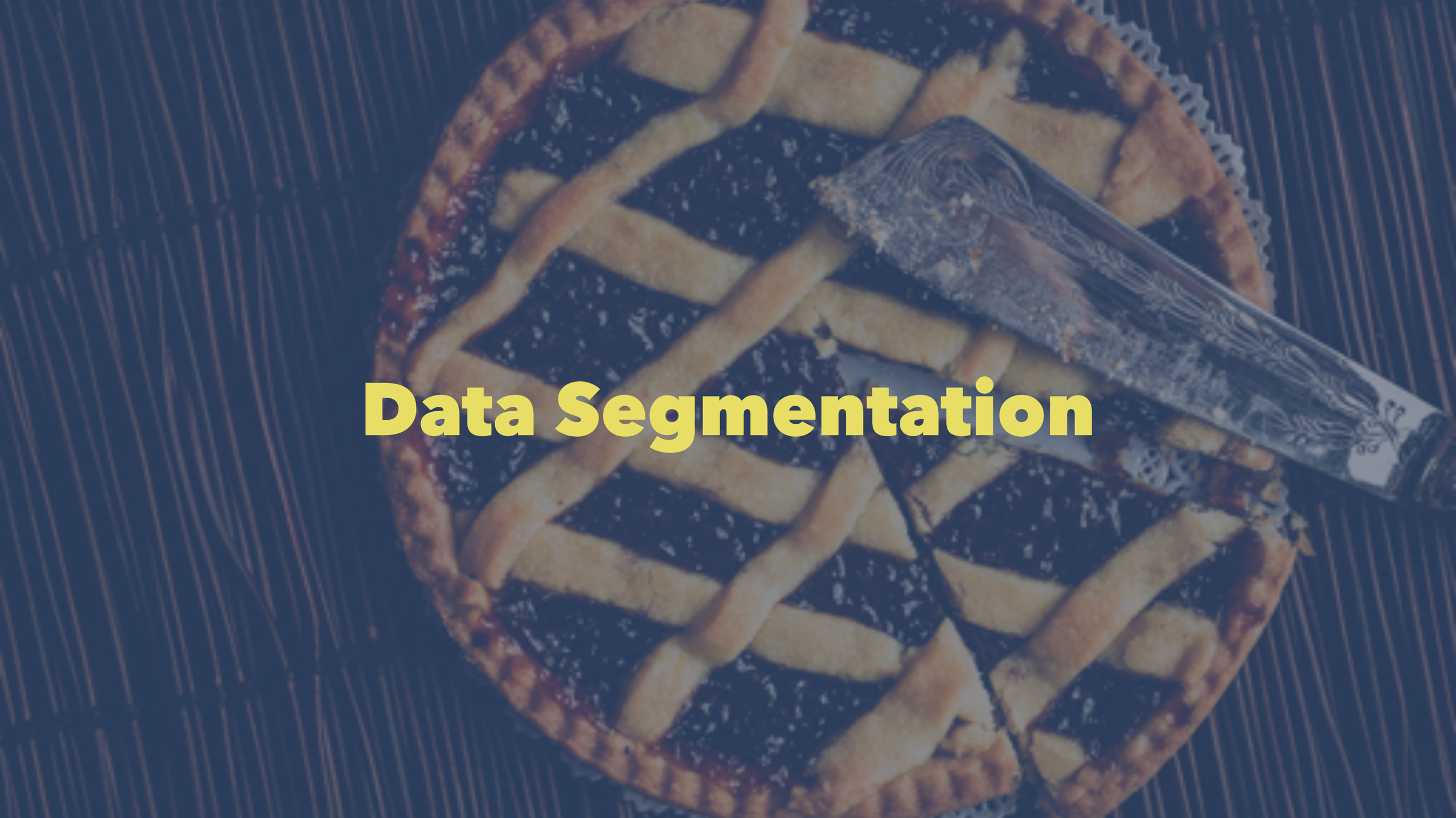
The Kolmogorov-Smirnov & Cucconi Tests



Regression analysis

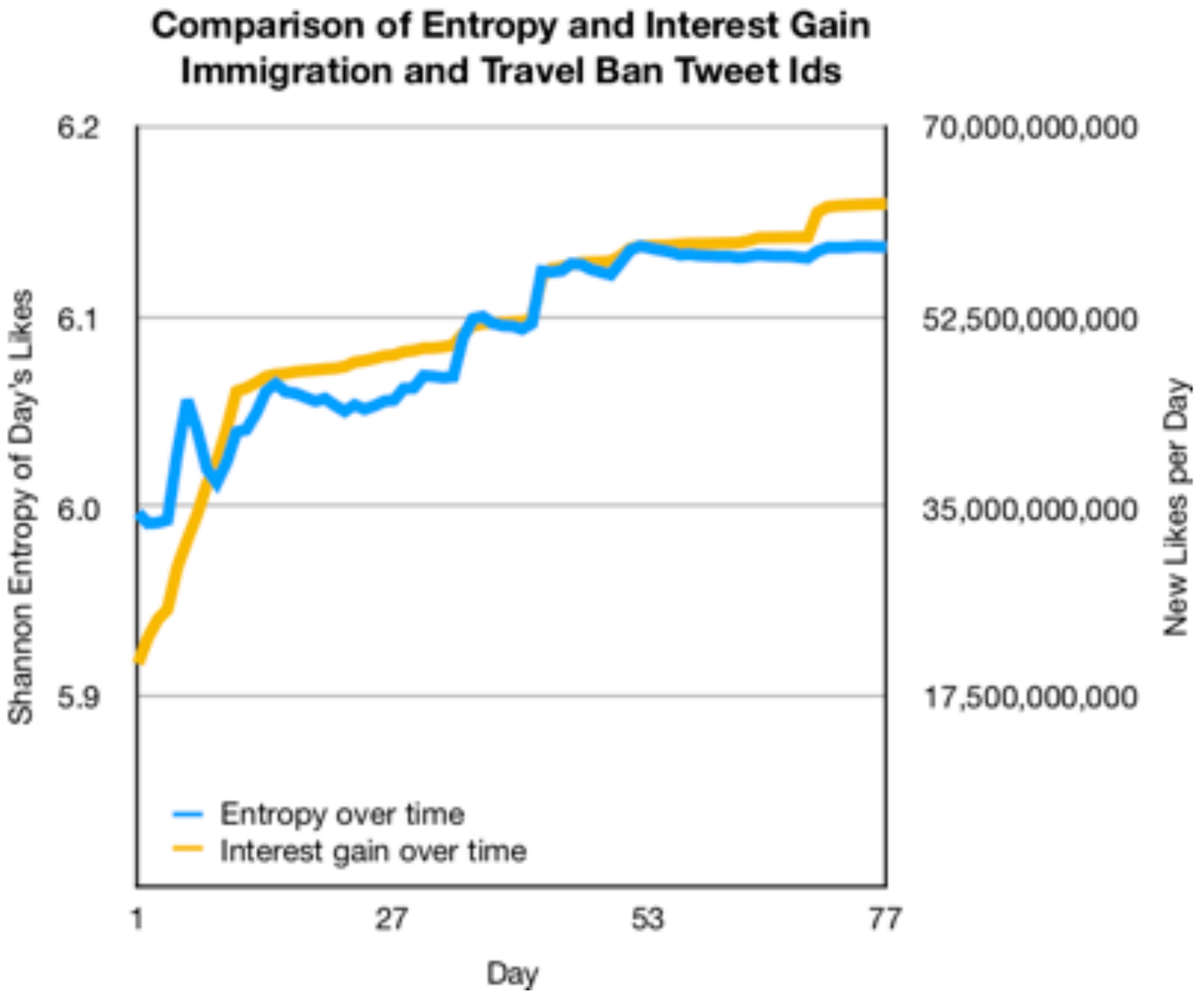


Scaling + Transforming


Data Segmentation

Preliminary Findings

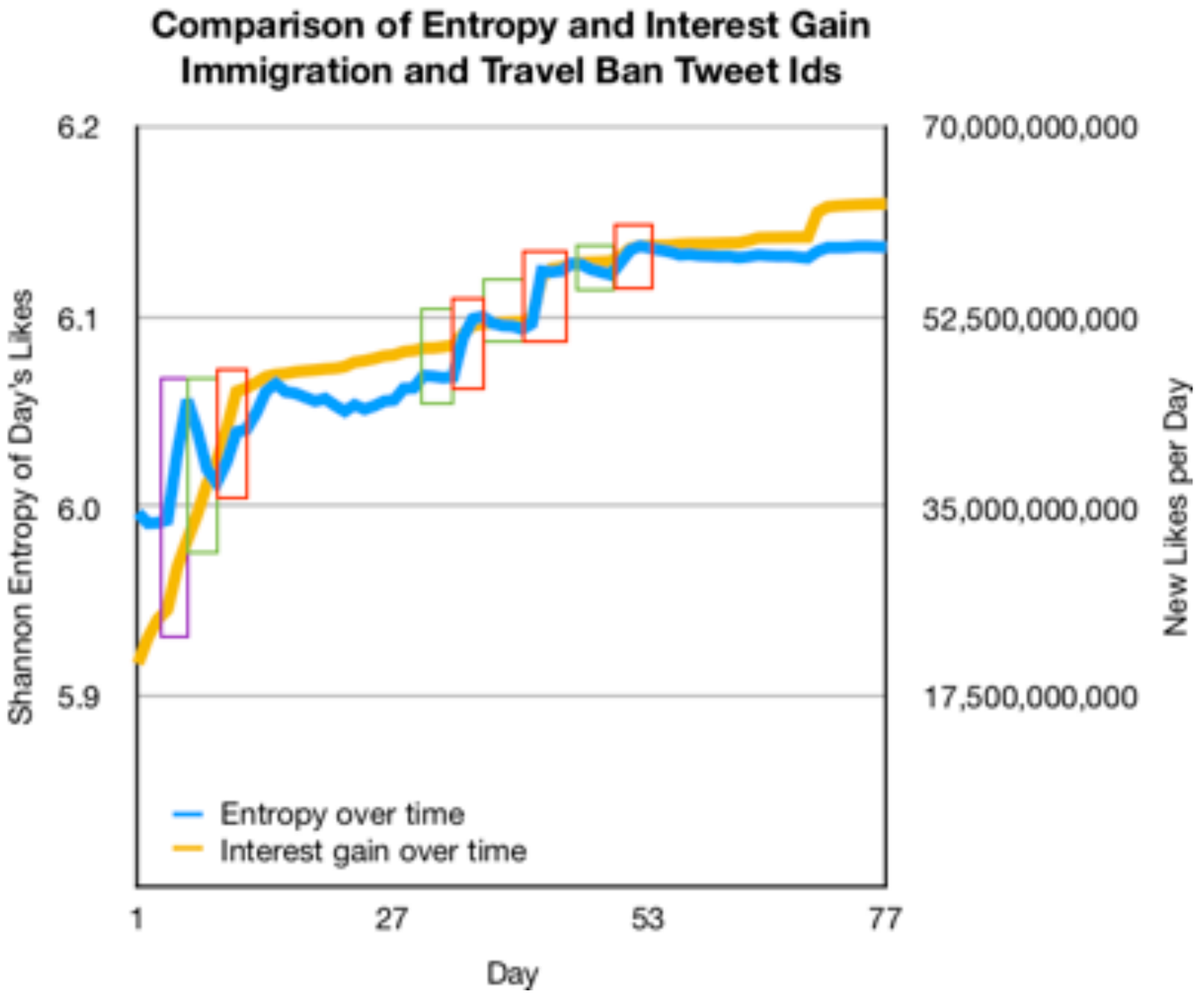
(*very* preliminary, I only just managed to hydrate a complete dataset on Sunday)


Comparison of Entropy and Interest Gain
Immigration and Travel Ban Tweet Ids
6.2
6.1
6.0
5.9
Shannon Entropy of Day's Likes
70,000,000,000
52,500,000,000
35,000,000,000
17,500,000,000
New Likes per Day
Entropy over time
Interest gain over time
1
27
53
77
Day

Revisiting hypotheses

I theorised that the patterns to emerge would be two-fold:

1. **increasing** entropy over time **preceding a trend** developing to significance, and
2. **decreasing** rate of entropy increase over time after a high entropy or engagement level **preceding changes in behaviour**.


Comparison of Entropy and Interest Gain
Immigration and Travel Ban Tweet Ids
6.2
6.1
6.0
5.9
Shannon Entropy of Day's Likes
70,000,000,000
52,500,000,000
35,000,000,000
17,500,000,000
New Likes per Day
Entropy over time
Interest gain over time
1
27
53
77
Day

...maybe that was too easy? 🤔

Where to go next?

1. **Verify and quantify** the appearance of correlation from initial data
2. **Test bounds** of correlation, extensively
3. **Experiment** with different bucketing values, differently weighted engagement metrics
4. **Observe and record** which factors increase or decrease correlation
5. **Repeat** all of the above with another dataset

References

[1] Hootsuite: 2018, Hootsuite's social media barometer report. https://hootsuite.com/resources/all-futureofsocial-digitalin2019-glo-en-ca-digitalin2019-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.

Thank you!

- All images are CC0, Pixabay-licensed or my own
- See **github.com/TheMartianLife/Honours-Presentation**