Social Media Data Analysis and Monitoring in Financial Services

Dr. Dani Madrid-Morales | dmmorales2@uh.edu | @DMadrid_M

24 February 2022

UNIVERSITY of HOUSTON

Outline

1. Why use **social media data** to monitor financial services' consumers?
2. Sample **research project** in three countries.
3. Overview of a **recommended workflow**/research approach.
4. Approaches for **computational analysis** of social media text data.
5. Q&A

Why use social media data?

UNIVERSITY of HOUSTON

What can social media data provide?

- Monitoring of social media offers opportunity to collect **observational data** about consumers' opinions, attitudes and behaviors.
- Insights gathered from the **analysis of social media data** can
  - Help monitoring in real-time of issues/events
  - Be incorporated in policy interventions, A/B testing…
  - Be used in predictive modeling

Sample project

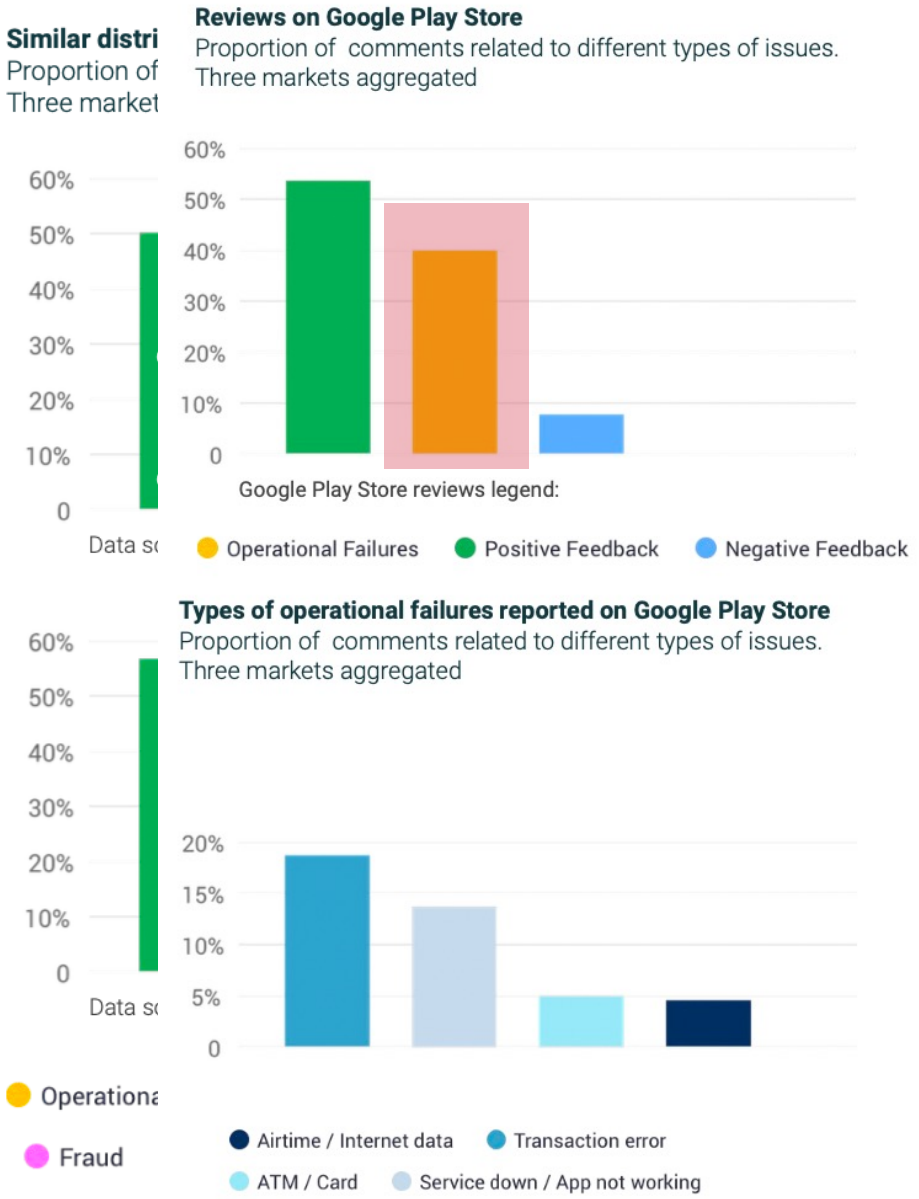
Example: Research Design

- IPA worked with Citibeats to conduct a social monitoring project
- Data collection driven by study goal:
  - Understand types of problems faced by **digital finance consumers**
- Social media data collected in **Nigeria**, **Kenya** and **Uganda**
  - Data in multiple languages
- Data comes from Twitter, Facebook Public Pages, and Google Play Store
- Longitudinal study: from **July 1, 2019**, to **July 1, 2020**
  - We cover roughly 6 months pre and post COVID-19

Example: Results (1)

1. Twitter and Facebook are mainly used to report **consumer protection issues, particularly customer care**.

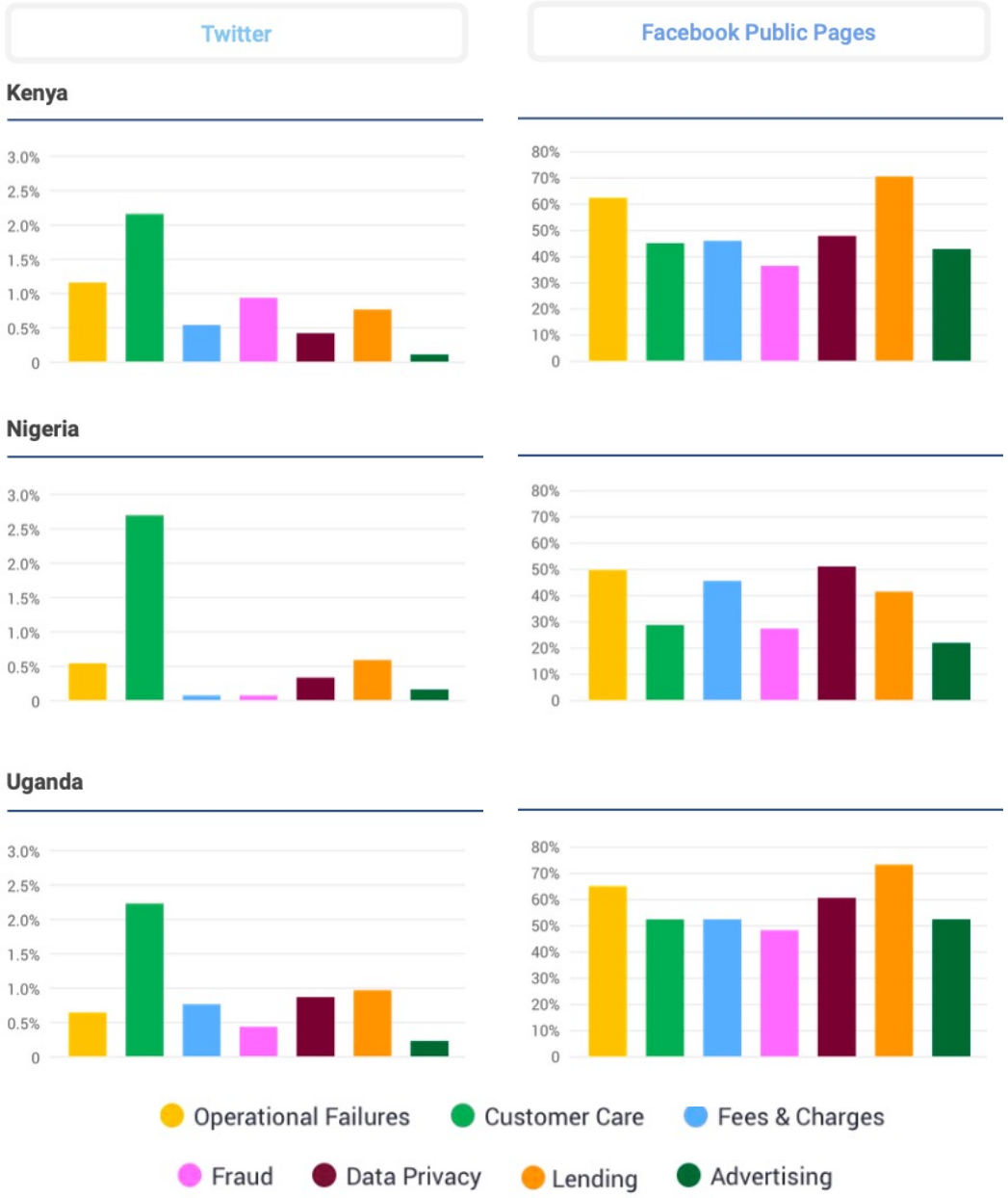
2. Google Play Store reviews focus on **app performance (positive reviews) and operational failures (negative)**
   - The most common operational failure reported was **Transaction errors**



Example: Results (2)

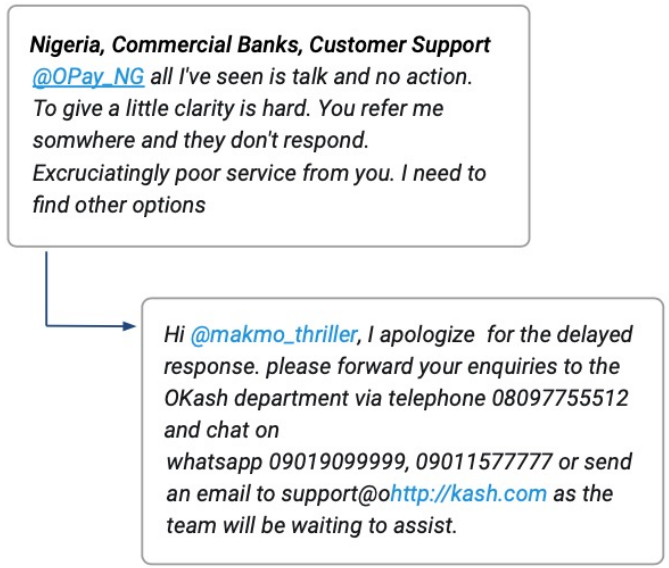
3. Financial providers' response rates **vary considerably** across Twitter, Facebook and Google Play Store.

4. Replies on Twitter are more **concentrated on customer care** issues; Facebook and Google Play responses are **more distributed among different issues**.



Example: Results (3)

5. Nigerian Commercial Banks & Fintech tend to **move to DM more often**, suggesting a better and more structured customer care policy.



Type of provider responses on Twitter
Proportion of type of responses by type of bank and country

	DM	Public response
UG | Microfinance	0%	100%
KE | Commercial Banks	29,9%	76,1%
NI | Commercial Banks	**80,7%**	19,2%
UG | Commercial Banks	20,2%	79,6%
KE | Microfinance	0%	100%
UG | Telecomms	19,3%	80,7%
KE | Telecomms	23,8%	76,2%
UG | Fintech	12,5%	87,5%
NI | Fintech	**55,5%**	44,6%
NI | Microfinance	35,7%	64,3%
NI | Telecomms	48%	51,9%
KE | Fintech	0%	100%

Data source: Twitter

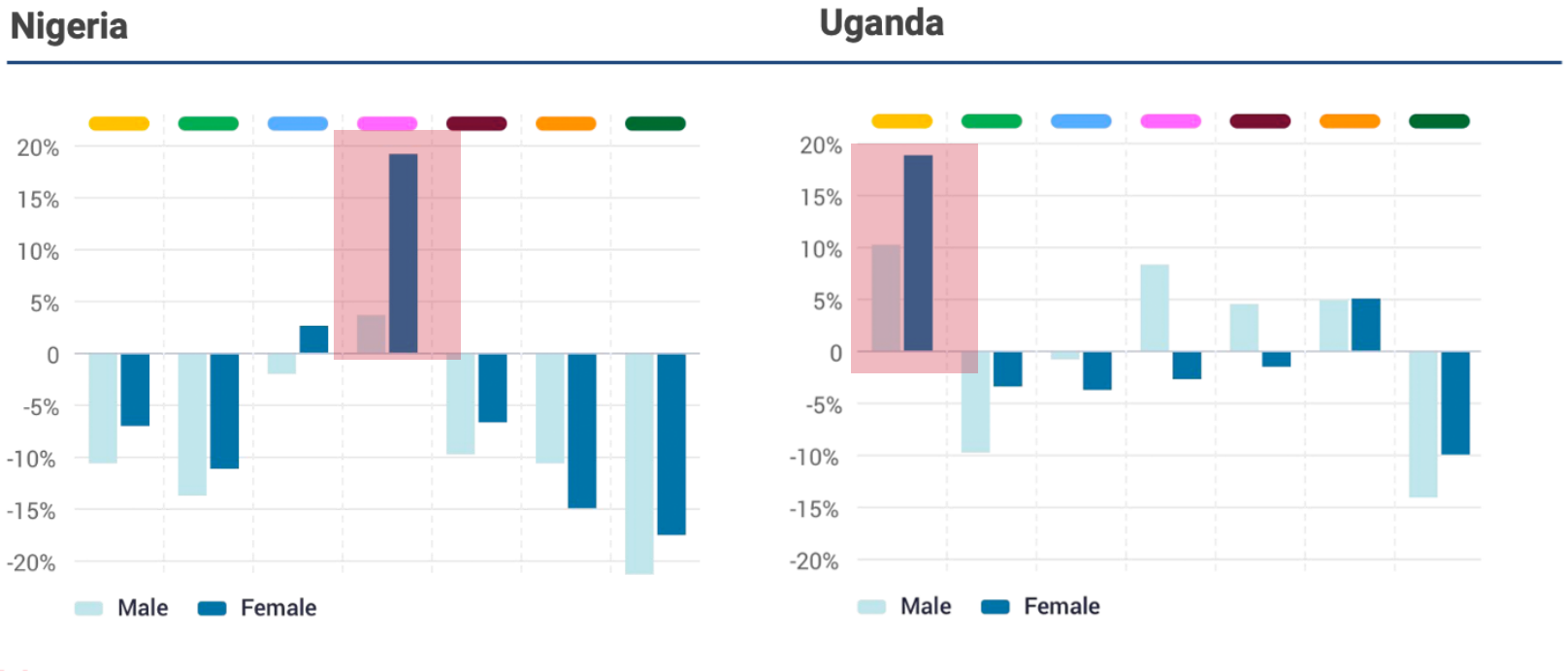
Example: Results (4)

6. The use of social media channels to communicate issues and interact with financial providers **increased** across the three markets after the Covid-19 pandemic.

7. The distribution of issues **did not change** post-Covid-19.

Example: Results (5)

8. On Twitter, women in Nigeria have **significantly increased** their rate of **complaints about fraud** compared to men after the outbreak of COVID-19, while in Uganda **customer care reports** have also **risen** for women.



Example: Approach/Workflow

Step 1
- Collecting social media data at scale

Step 2
- Defining categories/topics of interest

Step 3
- Using word frequencies and probabilities to locate topics in data

Step 4
- Deeper analysis by incorporating user information

Example: Approach/Workflow

Step 1

- Collecting social media data at scale

Step 2

Step 3

Step 4

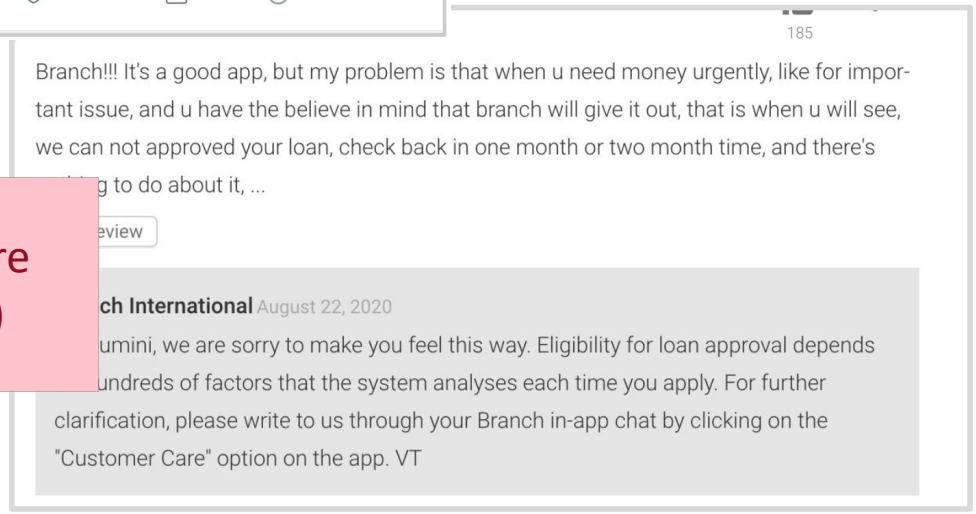
Example: Data sources

- For this project, we collected **4.5 million social media messages** from
  - Commercial Banks
    - (e.g. Equity, Polaris Bank, Stanbic Bank…)
  - Telecommunication Companies offering mobile money services
    - (e.g. Telekom T-kash, Airtel Money, UTL…)
  - Fintech start-ups offering lending/payment
    - (e.g. Okolea, Sokoloan, Tala)
  - Microfinance institutions
    - (e.g. Uwezo Kash, Fortis Mobile Money, Tugende)

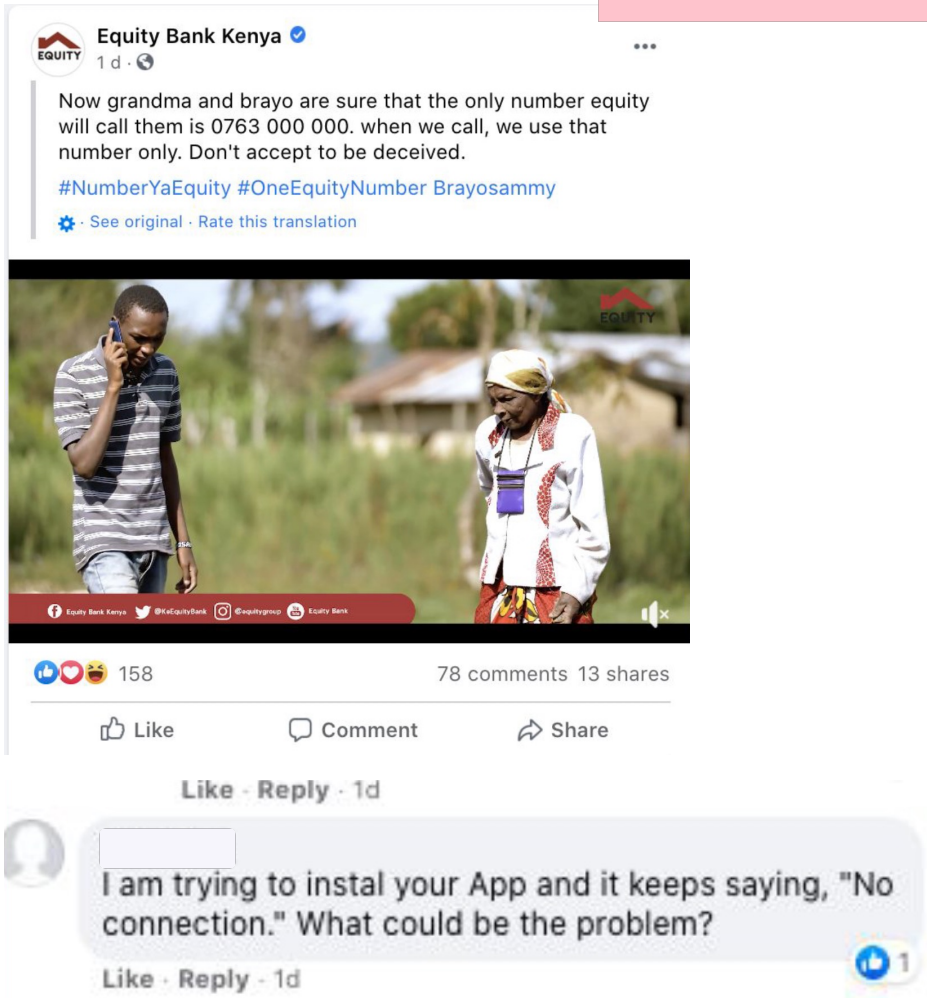
Example: Data

Facebook
830,939 (42 %)

Twitter
1,651,659 (85 %)



Google Play Store
294,950 (40 %)



Example: Approach/Workflow

Step 1
- Collecting social media data at scale

Step 2
- Defining categories/topics of interest

Step 3

Step 4

Example: Defining categories/topics of interest

- We combine **top-down** and **bottom-up** approaches to identifying salient categories in the data
- Through interviews and expert advice, **seven areas of interest** were identified before the analysis:
  - Operational failures, consumer care, fees & charges, fraud, data privacy, lending, advertising
- Using text analysis tools (cluster analysis), and human input (individual analysis of sample messages), sub-topics were identified.

Example: Approach/Workflow

Step 1
- Collecting social media data at scale

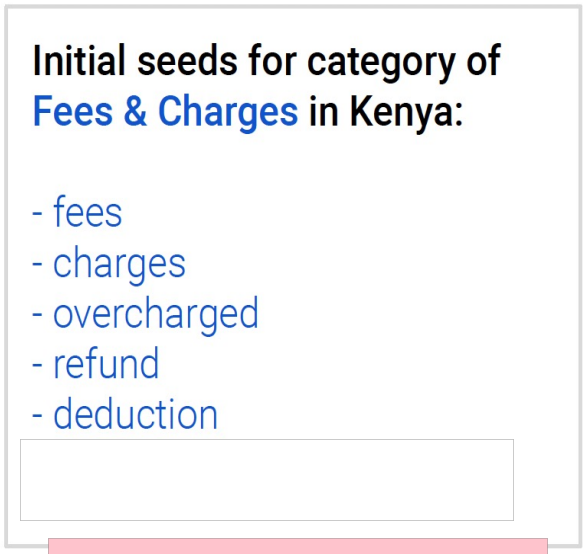
Step 2
- Defining categories/topics of interest

Step 3
- Using word frequencies and probabilities to locate topics in data

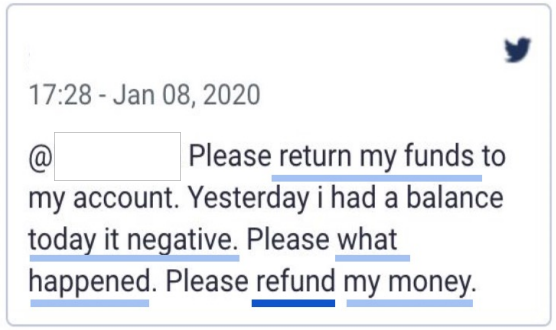
Step 4

Example: Semi-supervised machine learning

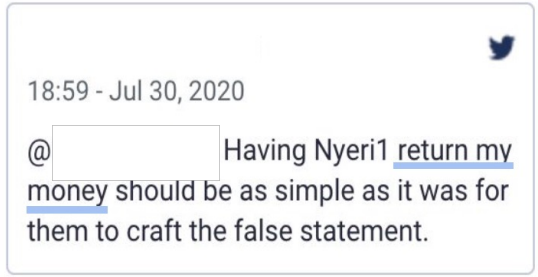
- During the analysis, we used (semi-supervised) **machine-learning** to go from unstructured text data to structured

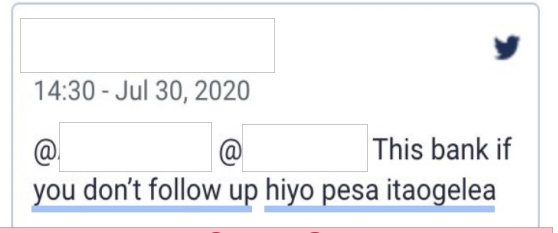


Step 1
User defined dictionary of keywords



Step 2
Computing topic probabilities from keywords & context





Step 3
Newly learned words help determine topics for items with no keywords

Example: Approach/Workflow

Step 1
- Collecting social media data at scale

Step 2
- Defining categories/topics of interest

Step 3
- Using word frequencies and probabilities to locate topics in data

Step 4
- Deeper analysis by incorporating user information

Example: Incorporating user metadata


```
  "place":
{
  "attributes":{},
   "bounding_box":
  {
     "coordinates":
     [[
           [-77.119759,38.791645],
           [-76.909393,38.791645],
           [-76.909393,38.995548],
           [-77.119759,38.995548]
     ]],
     "type":"Polygon"
  },
   "country":"United States",
   "country_code":"US",
   "full_name":"Washington, DC",
   "id":"01fbe706f872cb32",
   "name":"Washington",
   "place_type":"city",
   "url":"http://api.twitter.com/1/geo/id/0172cb32.j
son"
}
```


Location

User device

Gender

Adapting the workflow

A Proposed Workflow

Step 1
- Determining research goals and scope

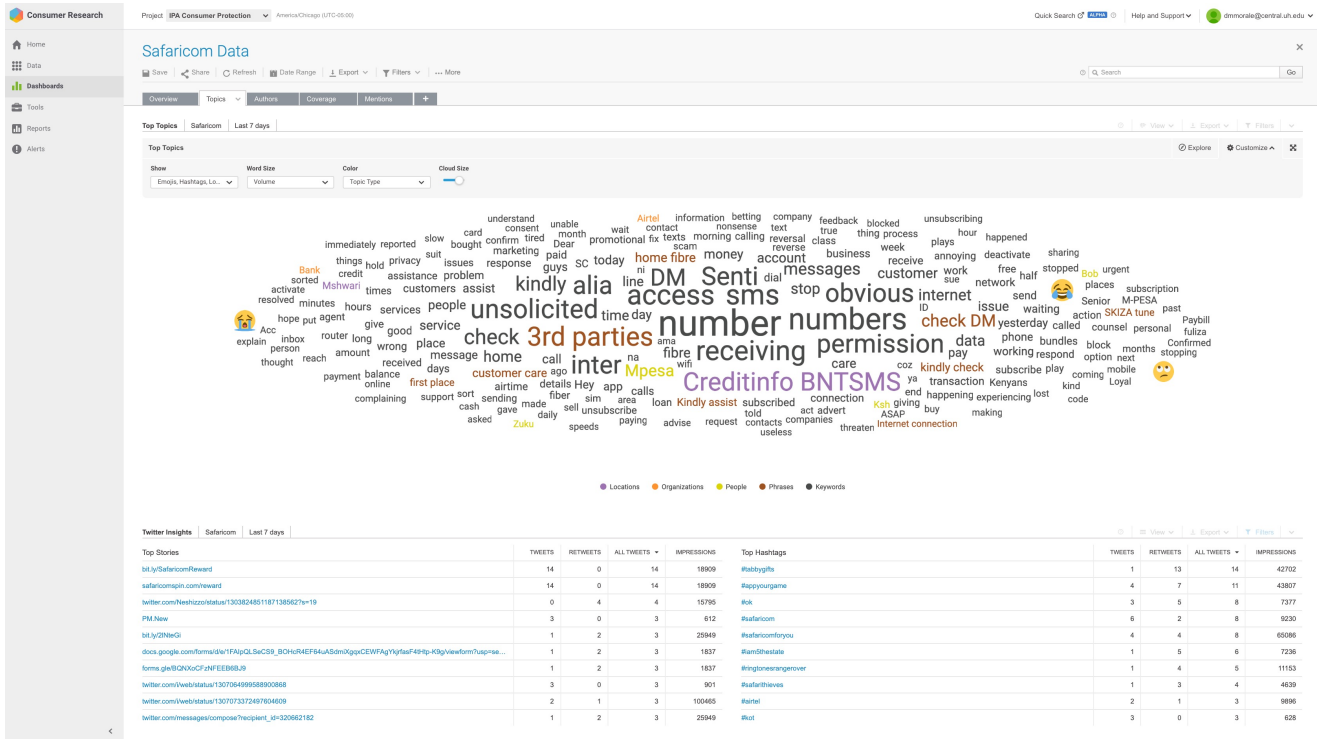
Step 2
- Automated collection of social media data at scale

Step 3
- Identifying the most suitable approach for automated analysis of data

Step 4
- Validation of computational findings, and interpretation

Social media data collection

1. There are **commercial solutions** to access social media data, such as Brandwatch, Crimson Hexagon and other similar products.



Social media data collection

2. APIs (**application programming interfaces**): a set of structured http requests that (usually) return JSON or XML data
   - It is possible to collect data via direct queries through http calls
   - Or, more commonly, queries can be sent through API Clients (e.g. `rtweet`, `tuber`, `RedditExtractoR`).
   - Not all social media platforms have an API (most notably, Facebook).

```
https://www.googleapis.com/youtube/v3/playlists?part=snippet
&channelId=UC_x5XG1OV2P6uZZ5FSM9Ttw
&key={YOUR_API_KEY}
```

Social media data collection


```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
        Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 54873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 756411,
  "favorite_count": 288867,
  "lang": "en"
}
```


API data is often returned in JSON formats, which can easily be turned into tabular data formats

Social media data collection

3. Web scraping: extract data from source code of website, with html parser and/or regular expressions

   - Scales well for large projects, it is reproducible.
   - Some websites explicitly prohibit the scraping of data (e.g., Facebook)
   - Packages in R for that purpose include `rvest, httr, XML2`

A Proposed Workflow

Step 1
- Determining research goals and scope

Step 2
- Automated collection of social media data at scale

Step 3
- Identifying the most suitable approach for automated analysis of data

Step 4
- Validation of computational findings, and interpretation

Computational approaches

UNIVERSITY of HOUSTON

What Can Computational Text Methods Do?

Haystack metaphor ~ **Improve Reading**

✗ Interpreting meaning of a phrase [**Analyzing a straw of hay**]

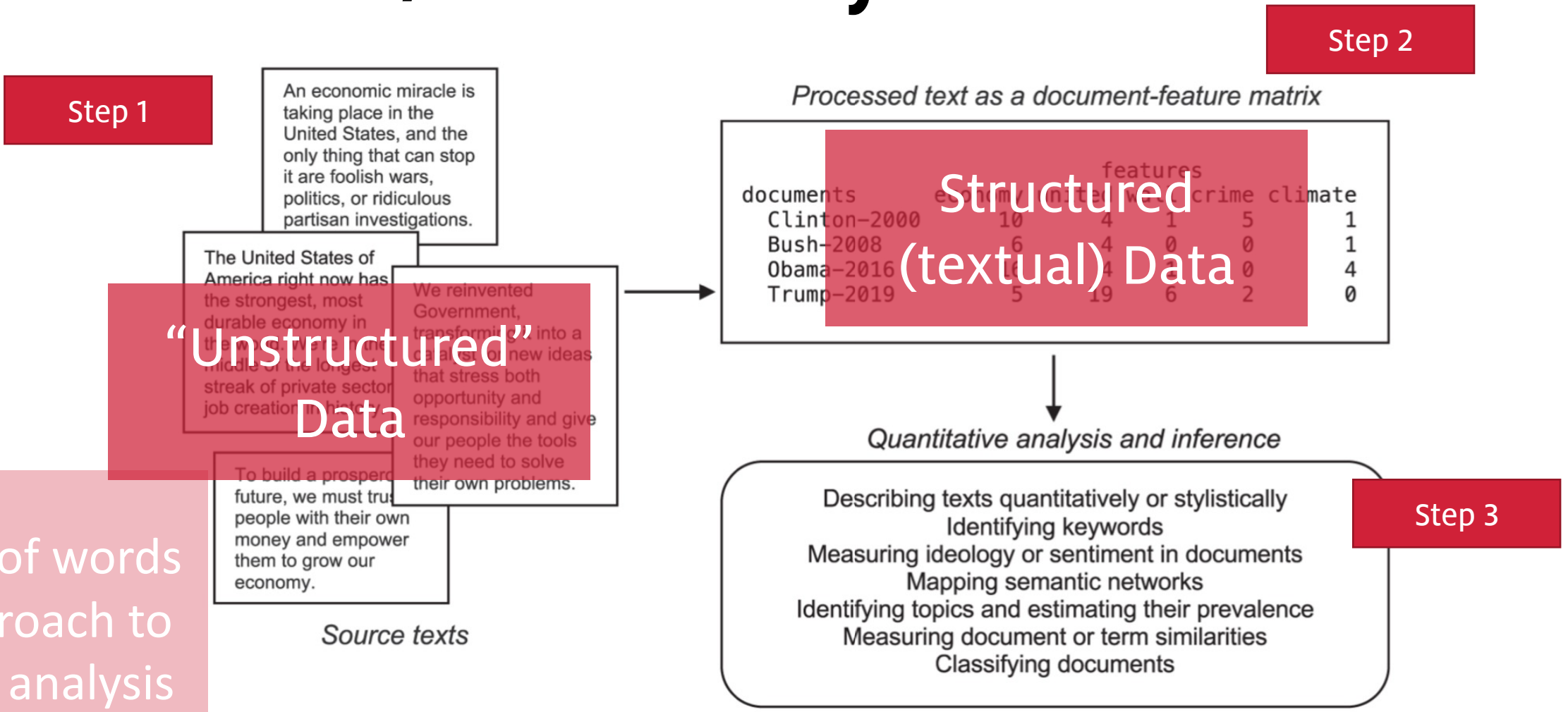
- Humans: amazing! (Straussian political theory, analysis of English poetry…)
- Computers: struggle ☹

Comparing, Organizing, & Classifying Texts [**Organizing haystack**]

- Humans: terrible. Tiny active memories 😐
- Computers: amazing!

Grimmer (2018a)

Text → DTM/DFM → Analysis



Bag of words approach to text analysis

Benoit (2020)

Document-term matrix (or DTM)

	Word 1	*Word 2*	*Word 3*	*Word 4*	*Word 5*	...	*M Words*
Document 1	1	3	2	0	0	...	
Document 2	0	0	1	1	0	...	
Document 3	1	1	0	2	3	...	
Document 4	3	1	0	0	0	...	
Document 5	0	1	0	3	1	...	
...							
Document n	0	1	1	0	1	...	

$$\times = \begin{pmatrix} 2 & 1 & 0 & \cdots & 2 \\ 1 & 0 & 1 & \cdots & 3 \\ 3 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & \cdots & 3 \end{pmatrix}$$

From words to numbers

1. Preprocess text (raw data)

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

Tweet 2 ***"@MEPcandidate You're an idiot, I would never vote for you"***

From words to numbers

1. Preprocess text: lowercase

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@mepcandidate thank you and congratulations, you're the best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@mepcandidate you're an inept, i would never vote for you"

From words to numbers

1. Preprocess text: lowercase, remove stop words, remove punctuation

Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@mepcandidate thank congratulations you're best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@mepcandidate you're inept never vote"

From words to numbers

1. Preprocess text: lowercase, remove stop words, remove punctuation, stem, tokenize

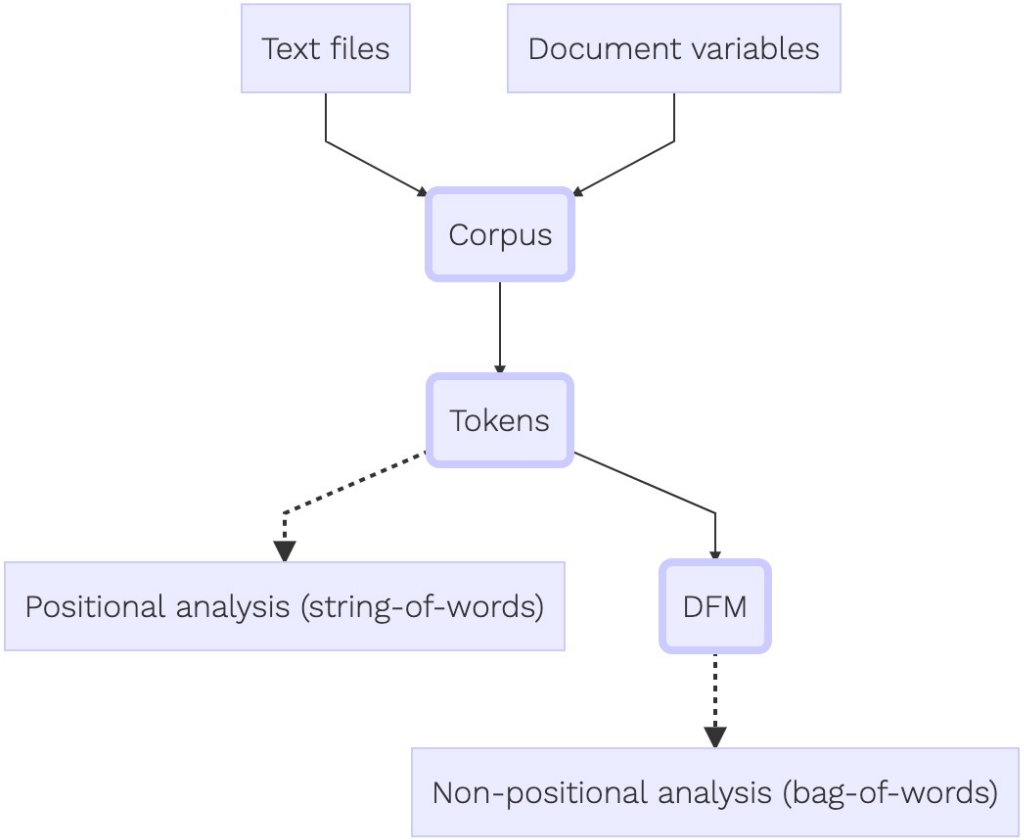
Tweet 1 ***"@MEPcandidate thank you and congratulations, you're the best #EP2014"***

"@ thank congratul you'r best #ep2014"

Tweet 2 ***"@MEPcandidate You're an inept, I would never vote for you"***

"@ you'r inept never vote"

Standard QTA procedure in quanteda



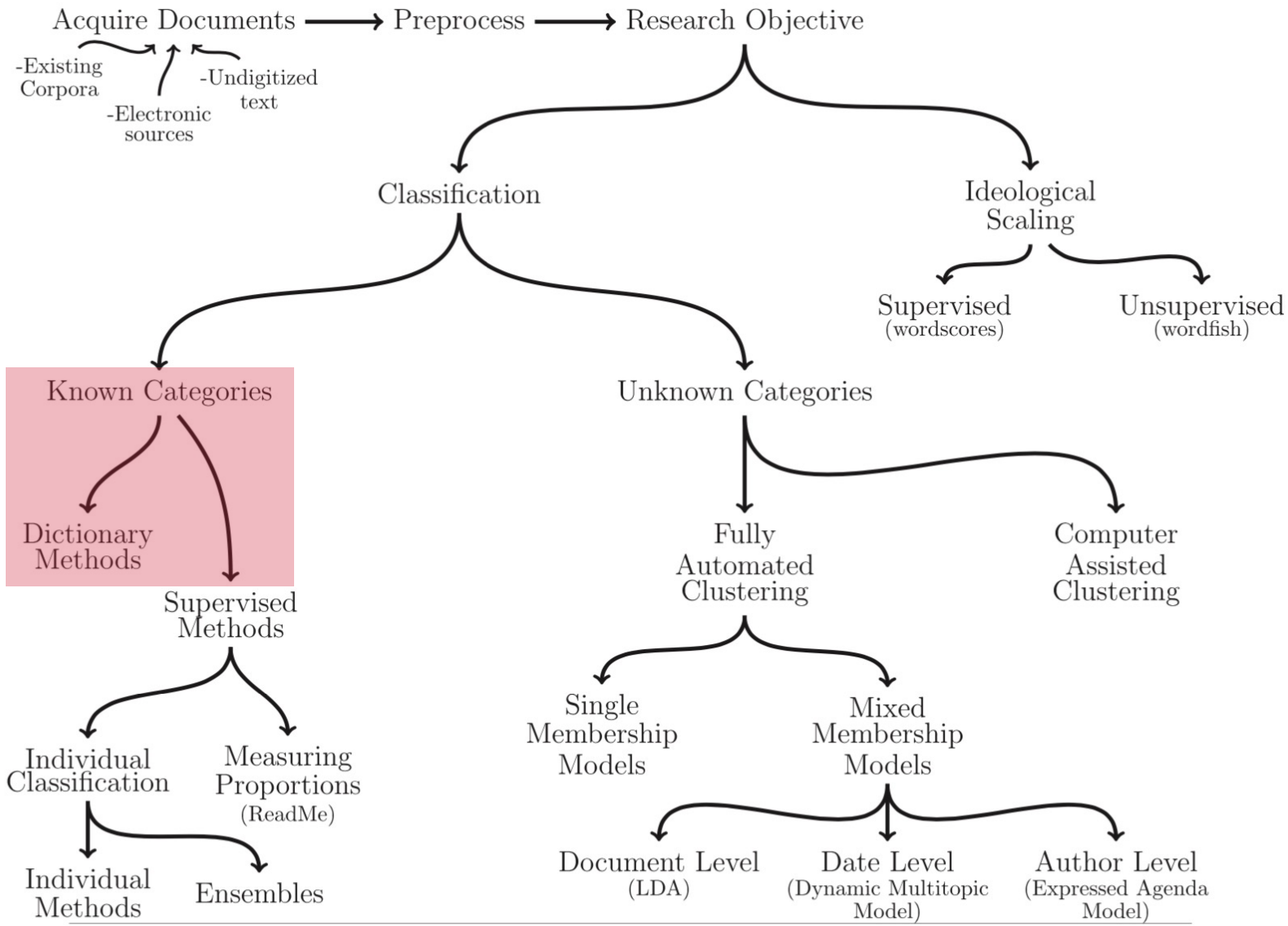


Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Dictionary methods

Classifying documents when **categories are known** using dictionaries:

1. Lists of words that correspond to each category:
   - Positive or negative (for sentiment)
   - Sad, happy, angry, anxious (for emotions)
   - Insight, causation, discrepancy, tentative (for cognitive processes)
   - Sexism, homophobia, xenophobia, racism (for hate speech)

Adapted from Barberá (2016)

Dictionary methods

2. Count **number of times** they appear in each document
3. Normalize by document length (optional)
4. Validate, **validate**, validate.
   - Check sensitivity of results to exclusion of specific words
   - Code a few documents manually and see if dictionary prediction aligns with human coding of document

Adapted from Barberá (2016)

Dictionaries

Bing Liu Sentiment Lexicon

word	label
zombie	negative
zippy	positive
zest	positive
zenith	positive
zealously	negative
zealot	negative
zeal	positive
zaps	negative
zapped	negative
zap	negative

AFINN-111 Dictionary

word	value
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
adduction	-2
abhor	-3
abhorred	-3
abhorrent	-3
abhors	-3
abilities	2

Loughran-McDonald Lexicon

word	value
compelling	constraining
compensatory	litigious
complain	negative
compliment	positive
confuses	uncertainty
extant	superfluous
Failed	negative
forego	negative
honors	positive
hurt	negative

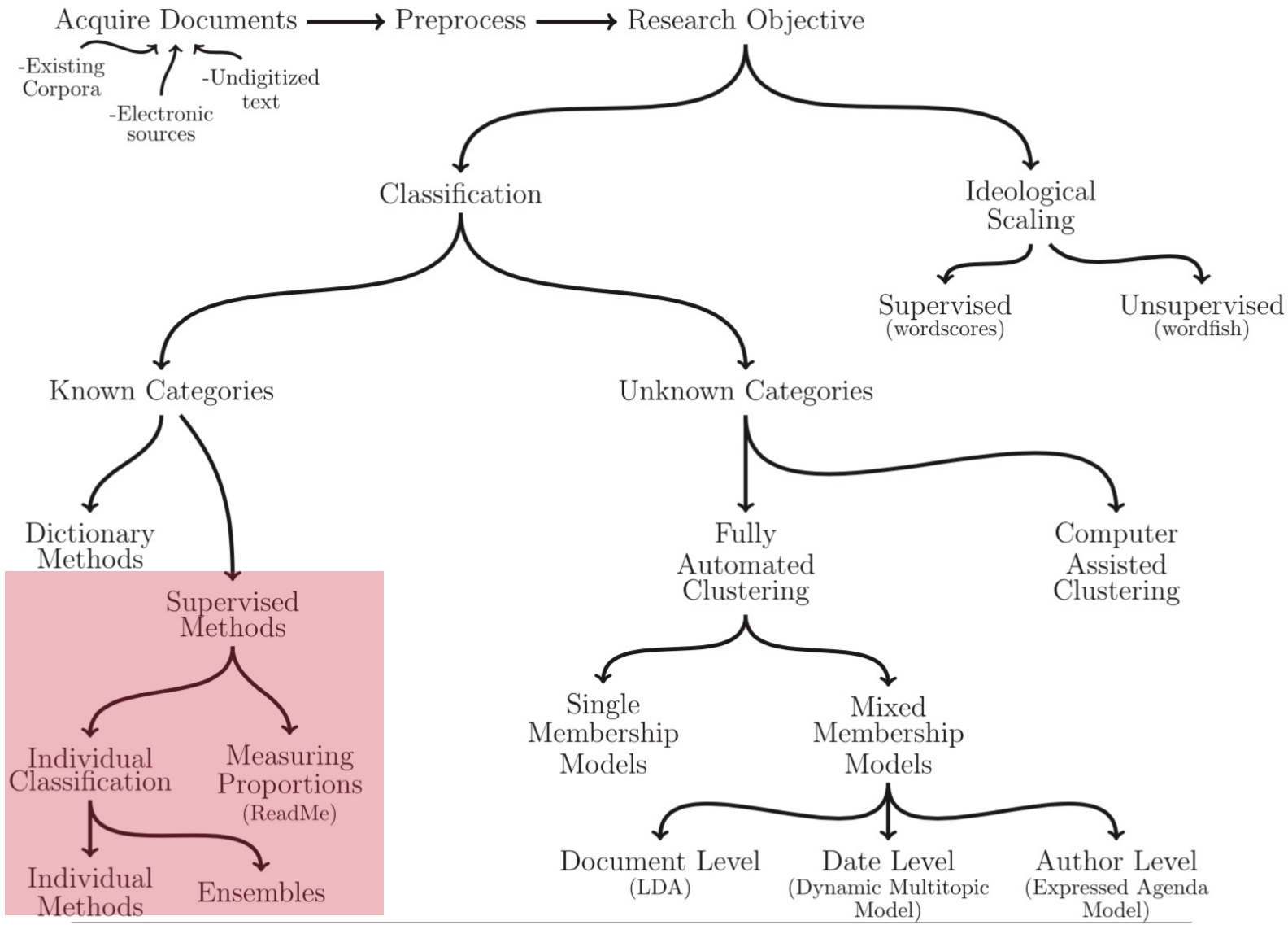


Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Supervised Machine Learning

- Machine Learning in QTA refers to training statistical models on a set of **annotated texts**, that are used to predict the category of **unseen texts**.
- Machine learning uses a number of variables, called features (IV), to predict a target category of class (DV).
- In text mining, IVs are generally **term frequencies**.

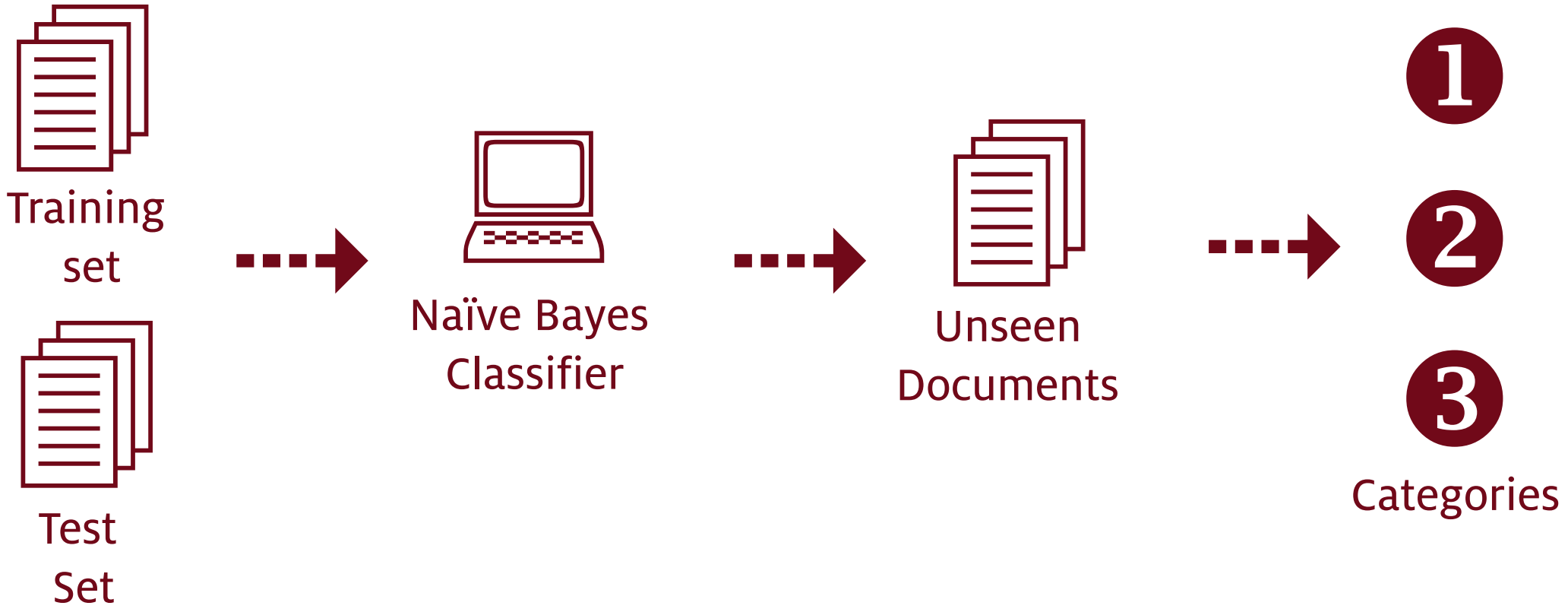
Van Attenveldt (2016)

Supervised Machine Learning

- Our goal is to **classify documents** into pre-existing categories, such as
  - sentiment of tweets
  - types of user complaints
  - types of services being discussed
- Some of these tasks could be done using a dictionary-based approach, but ML algorithms can **help avoid some of the pitfalls of dictionaries**

Adapted from Barberá (2019)

Supervised ML (Naïve Bayes Classifier)

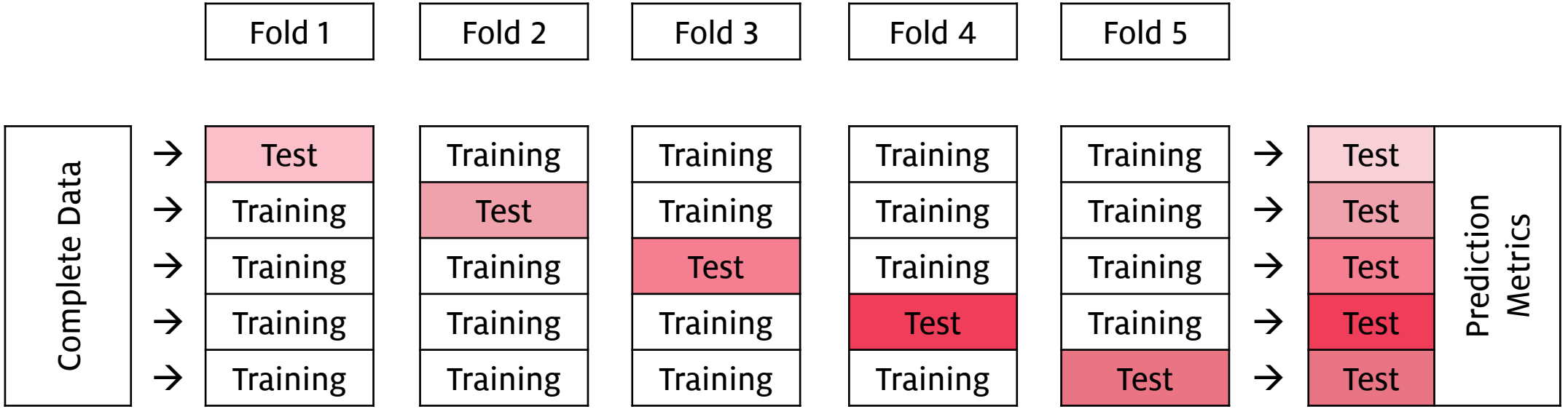


Steps to Supervised ML Text Classification

1. Construct a **corpus** and a **labeled dataset** (sometimes, the same)
2. Construct **feature vectors**
3. Split the **labeled data** into a **training** and **test** set
4. Select **one (or several) algorithms**
5. Tune **parameters** (if necessary)
6. Run **algorithm on training set**
7. Evaluate **accuracy** on test set
8. Repeat **steps 4-7** until you have a satisfactory model

Lukito & Sun (2019)

Measuring performance



Adapted from Barberá (2019)

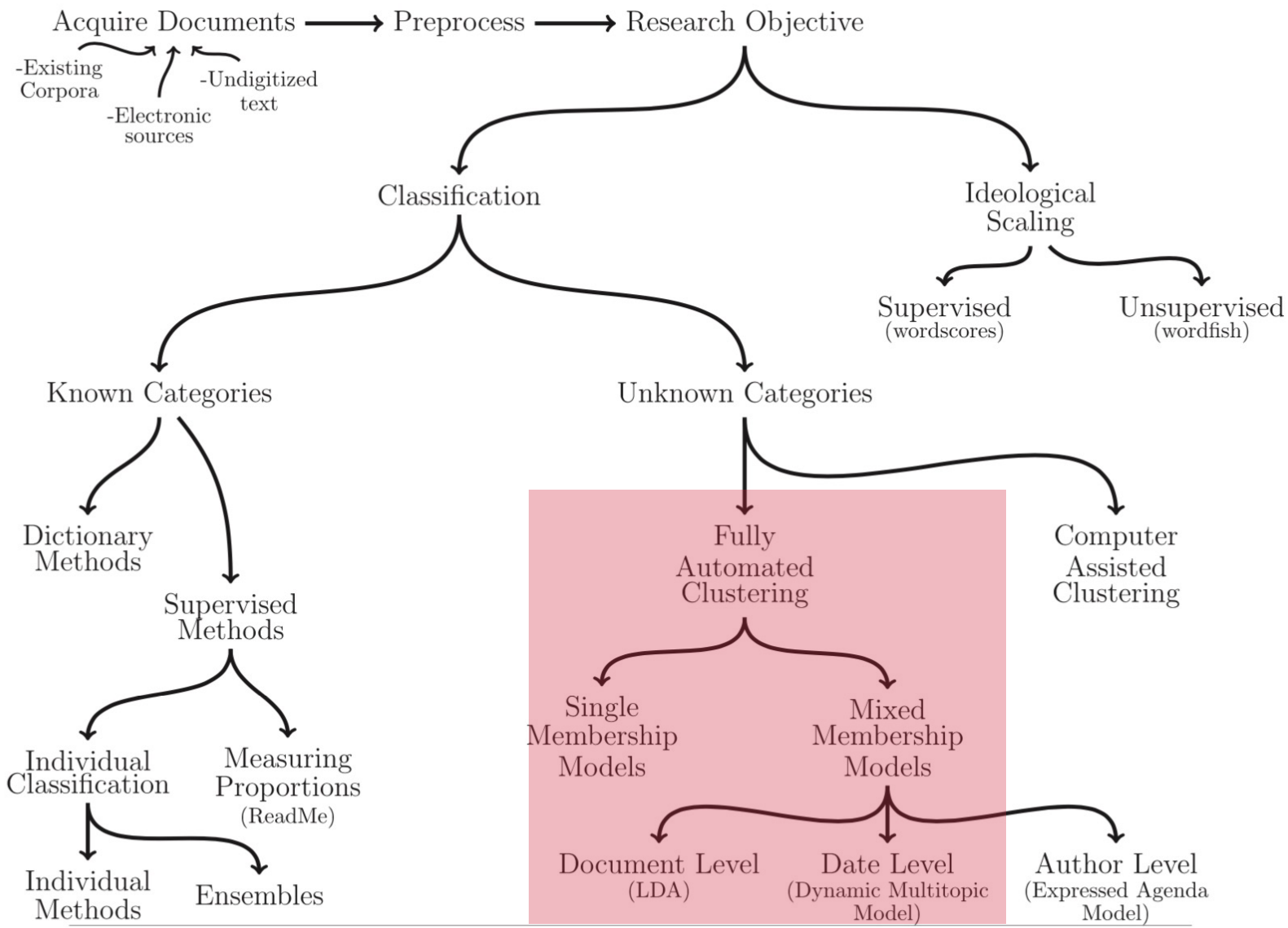


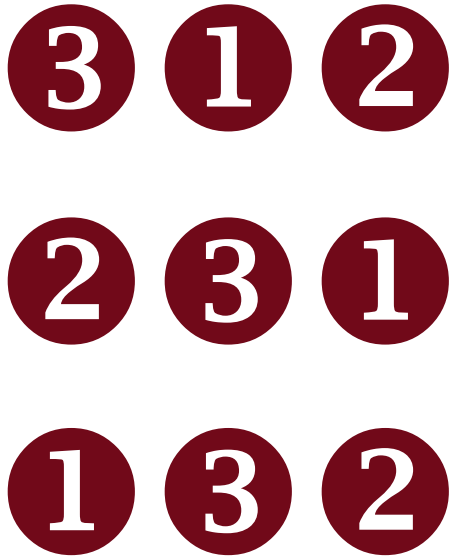
Fig. 1 An overview of text as data methods.

Grimmer and Stewart (2013)

Unsupervised ML (Topic Modeling)

Unseen Documents

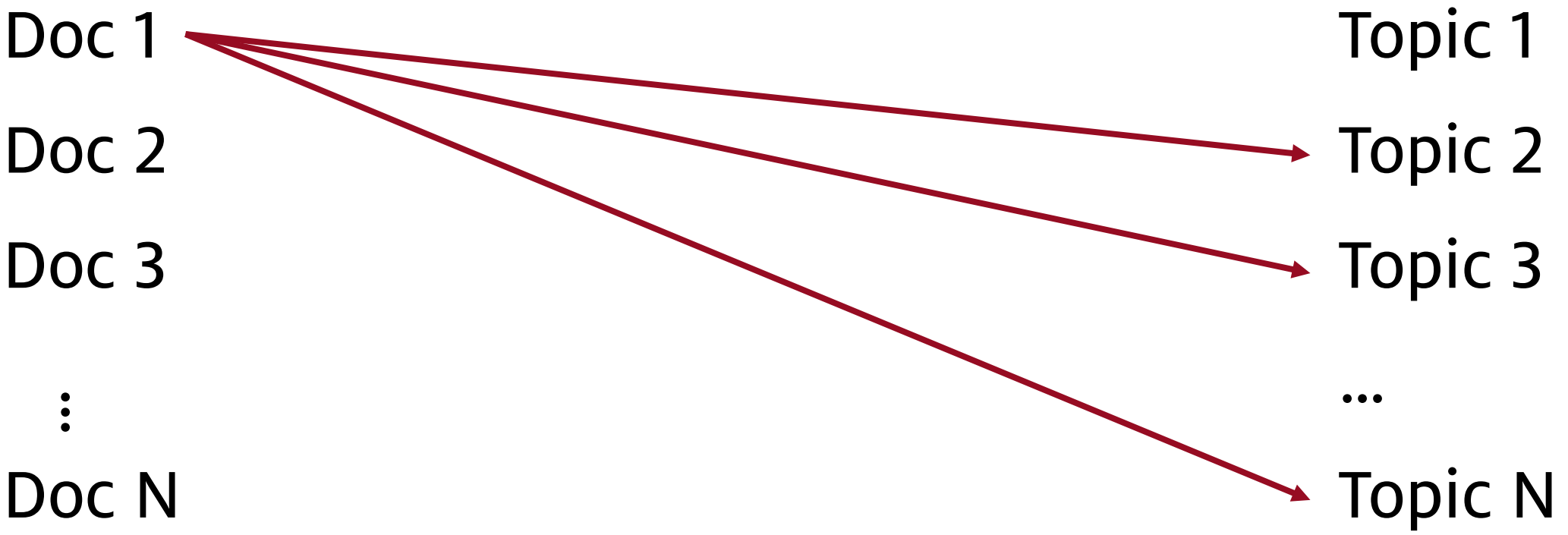
Topic Modeling



Multiple Categories

Single vs. Mixed Membership models

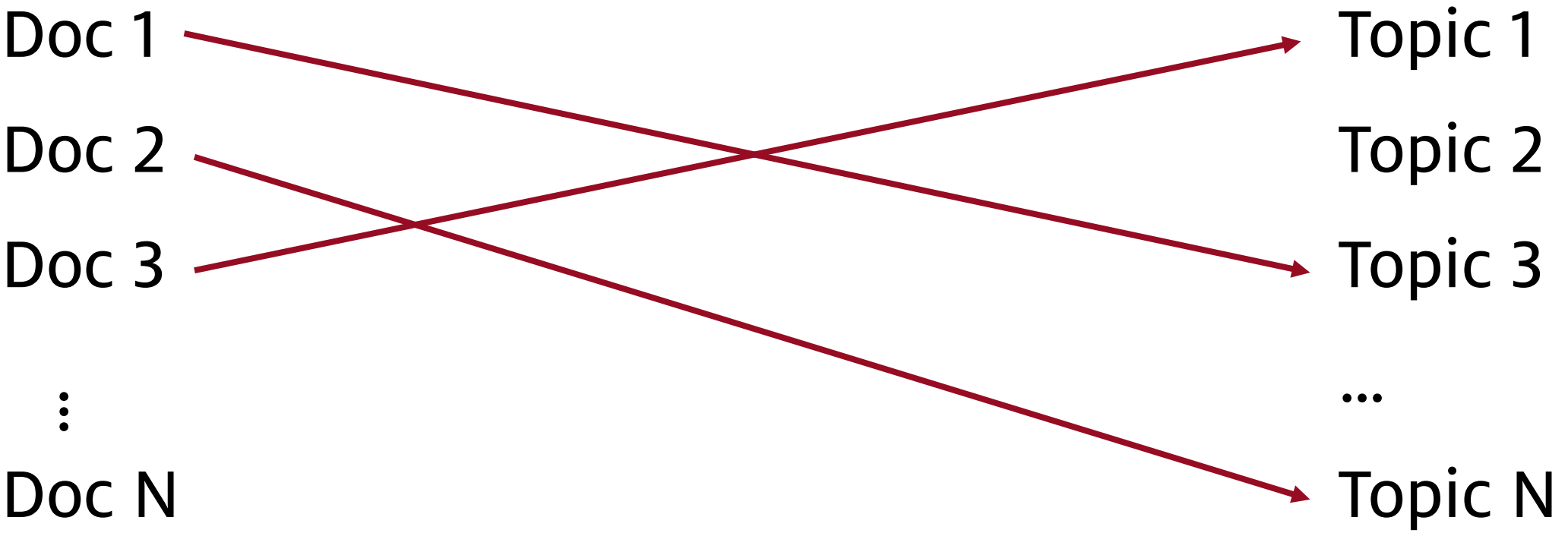
Topic modeling (e.g. LDA)



Adapted from Terman (2018)

Single vs. Mixed Membership models

Supervised ML (e.g. Naïve Bayes Classifier)



Adapted from Terman (2018)

Latent Dirichlet Allocation (LDA)

- Inputs
  - A document feature matrix (or any multidimensional dataset)
  - K: the desired number of topics.
- Outputs
  - π_{κ} : Topic distribution over words.
  - θ_{i} : Document distribution over topics.

Adapted from Terman (2018)

Latent Dirichlet Allocation (LDA)

- Goal: Topic model the following documents:
  1. Russia **threatens Georgia** with **sanctions** again.
  2. New **threat** against **Ukraine** and **Georgia** over **election** interfering.
  3. The rising **price** of **oil** and **gold**.
  4. UAE reduces **production** of **oil** significantly.
  5. With new **sanctions** coming, an increase in **oil prices** looms over **Georgia**.

Topic A (interpreted as '**politics**')

Topic B (interpreted as '**economy**')

Adapted from Terman (2018)

LDA Output (π_k)

Topic distribution over words (π_k)

Topic	threat	Ukraine	sanctions	oil	gold	price	election	loom	Total
A	.30	.25	.20	.01	.01	.01	.12	.10	1
B	.01	.01	.01	.35	.24	.25	.08	.05	1

Adapted from Terman (2018)

LDA Output (θ_i)

Document distribution over topics (θ_i)

Docs	Topic A Weight	Topic B Weight	Total
1	.99	.01	1
2	.99	.01	1
3	.01	.99	1
4	.01	.99	1
5	.60	.40	1

Adapted from Terman (2018)

Researcher Involvement in Different Types of QTA

Human Coding	A lot
Supervised QTA	Some
Unsupervised QTA	A bit less

A Proposed Workflow

Step 1
- Determining research goals and scope

Step 2
- Automated collection of social media data at scale

Step 3
- Identifying the most suitable approach for automated analysis of data

Step 4
- Validation of computational findings, and interpretation

Validate, Validate, Validate

- It is common to use four metrics to evaluate the accuracy of machine learning algorithms:
  - Accuracy: percent of overall correct classified items
  - Precision: ability of a classification model to identify **all relevant instances**
  - Recall: ability of a classification model to return **only relevant instances**
  - F1 score: single metric that **combines recall and precision** using the harmonic mean

Confusion matrix

	Actual label	
Classification (algorithm)	*Negative*	*Positive*
Negative	*True negative*	*False negative*
Positive	*False positive*	*True positive*

Performance metrics

$$Accuracy = \frac{TrueNeg + True\ Pos}{TrueNeg + True\ Pos + FalseNeg + FalsePos}$$

$$Precision_{positive} = \frac{True\ Pos}{True\ Pos + FalsePos}$$

$$Recall_{positive} = \frac{True\ Pos}{True\ Pos + FalseNeg}$$

$$Fscore\ or\ F1 = 2\frac{(Precision \times Recall)}{Precision + Recall}$$

Q&A

UNIVERSITY of HOUSTON

Social Media Data Analysis and Monitoring in Financial Services

Dr. Dani Madrid-Morales | dmmorales2@uh.edu | @DMadrid_M

24 February 2022

UNIVERSITY of HOUSTON