HOW UNIQUE ARE YOU ON TWITTER

UNDERSTANDING THE TRADEOFFS BETWEEN PRIVACY AND UTILITY

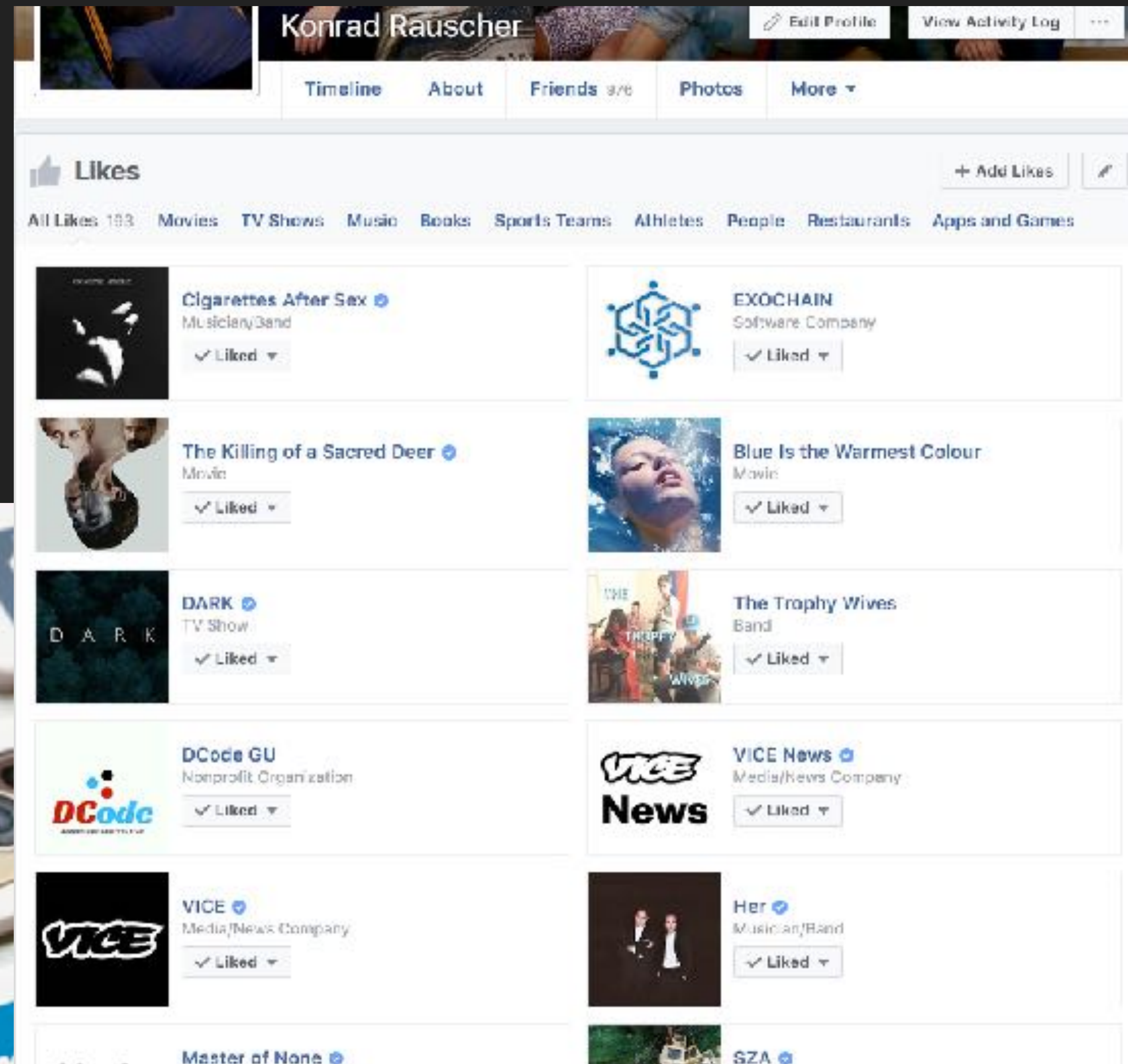
- KONRAD RAUSCHER

OUTLINE

1. Motivation & Problem
2. Related Literature
3. Methodology
4. Data Description
5. Experiments & Results
6. Contributions

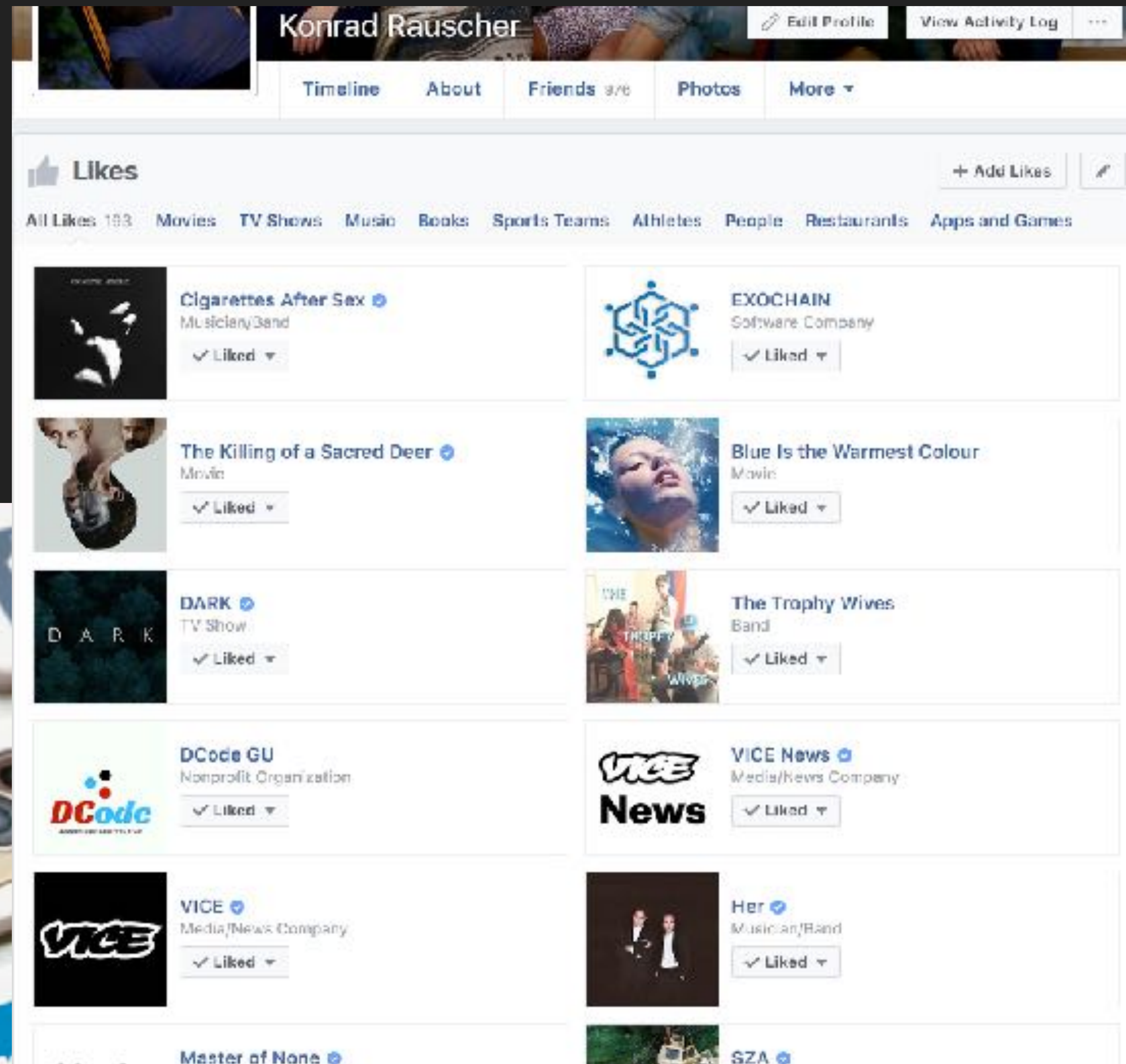
MOTIVATION

Cambridge Analytica



MOTIVATION

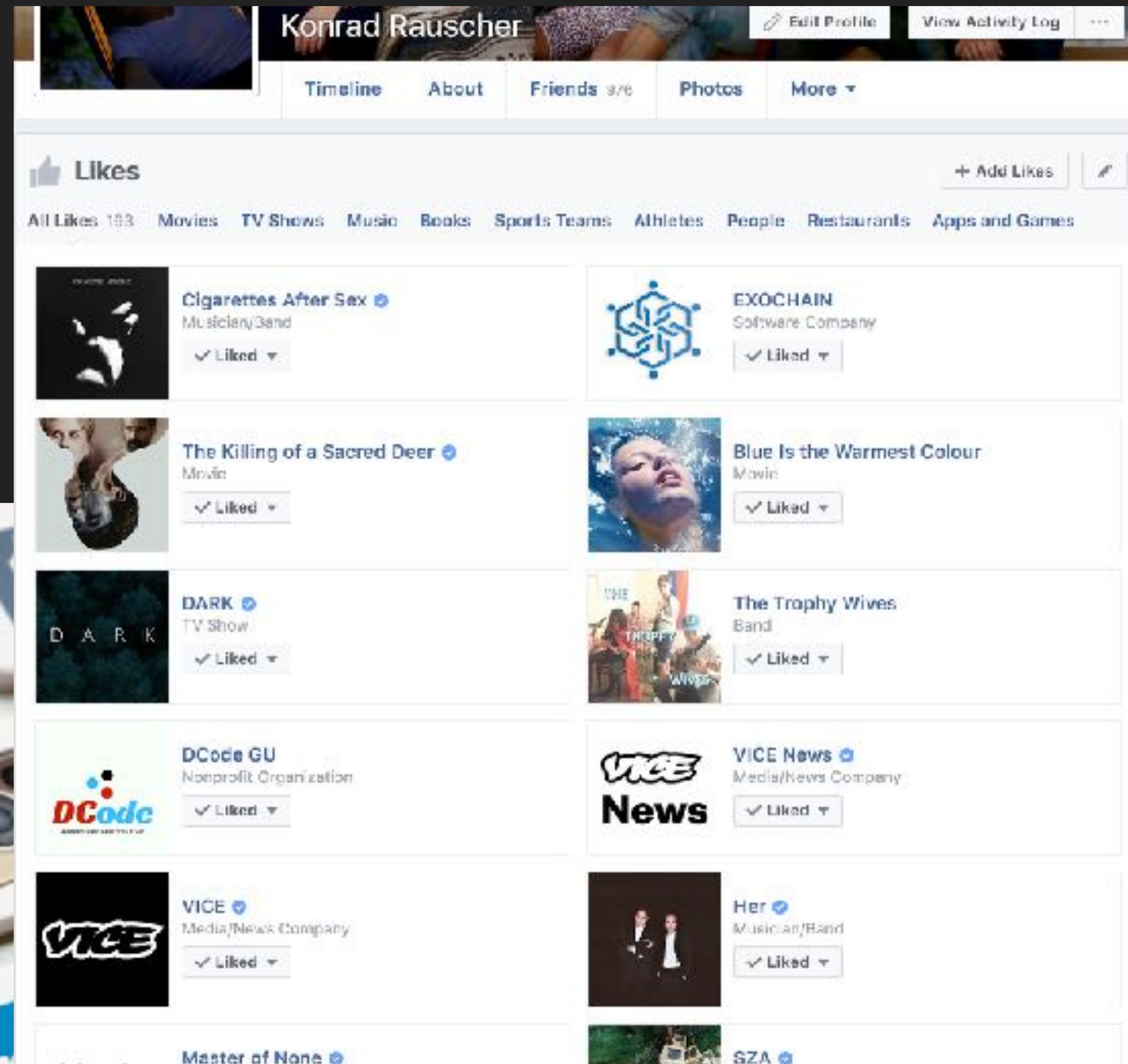
Cambridge Analytica



How unique is Twitter user content?

MOTIVATION

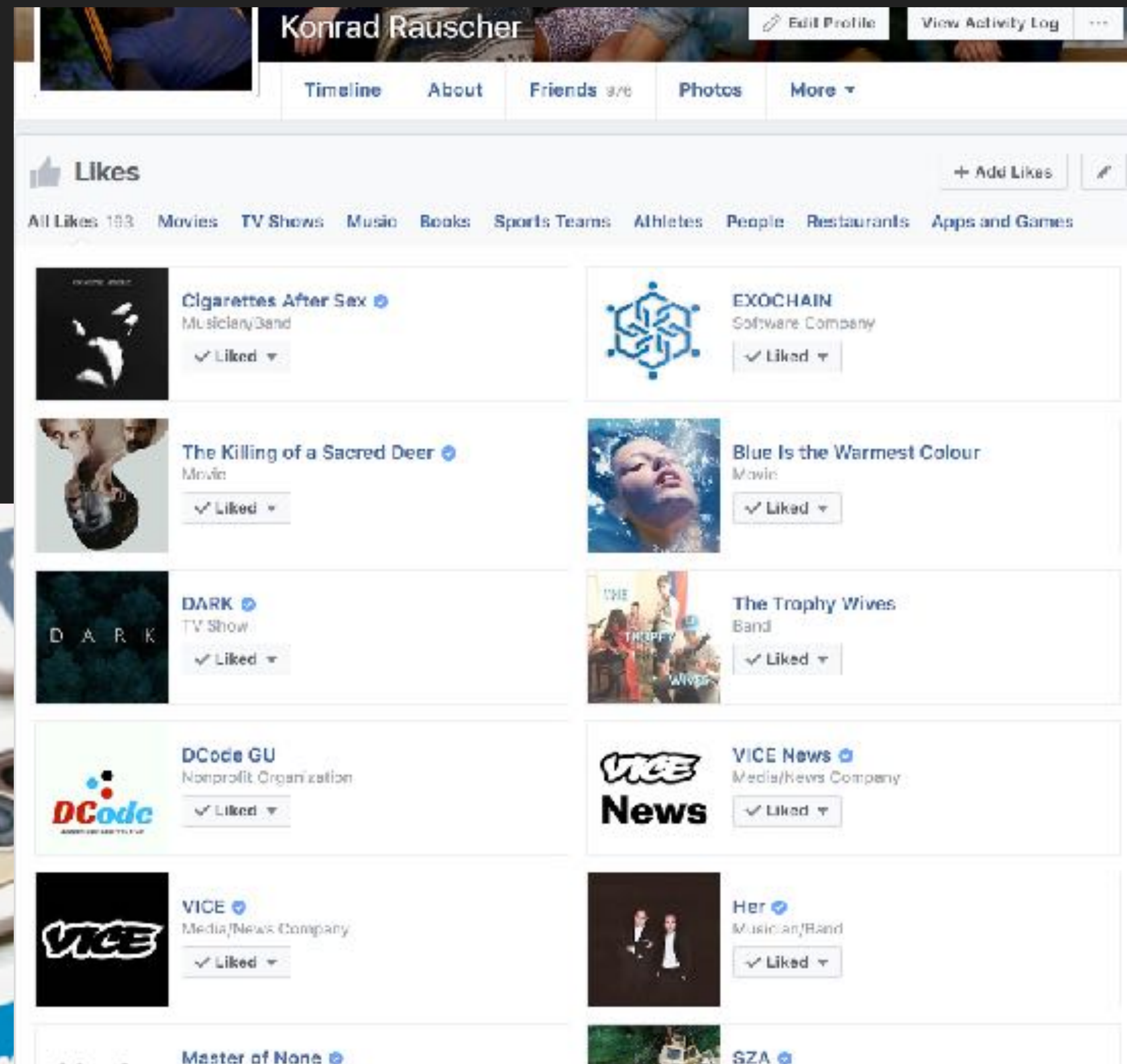
Cambridge Analytica



Can we easily make it private?

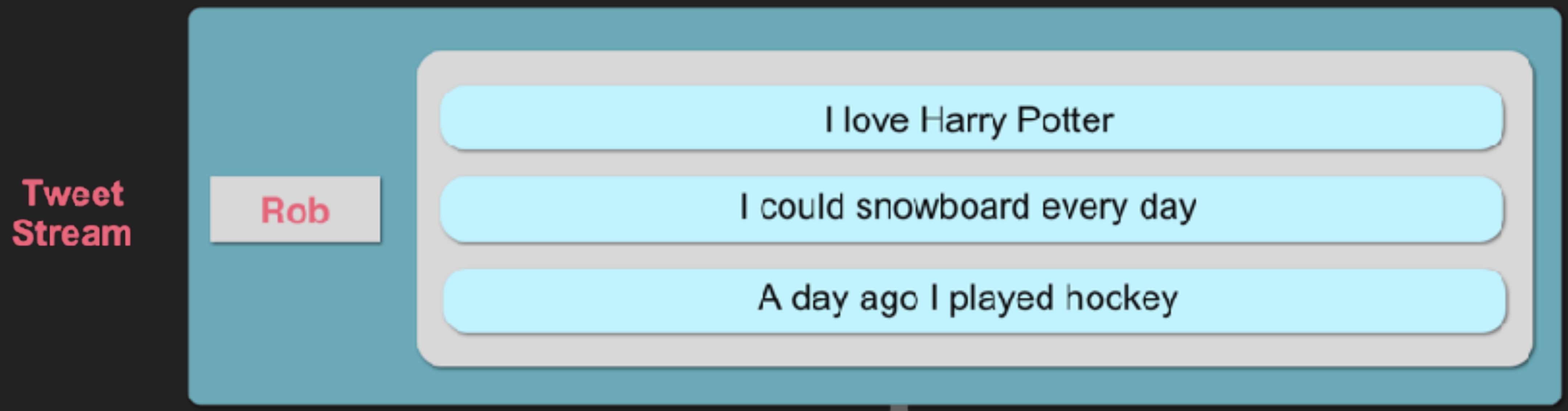
MOTIVATION

Cambridge Analytica

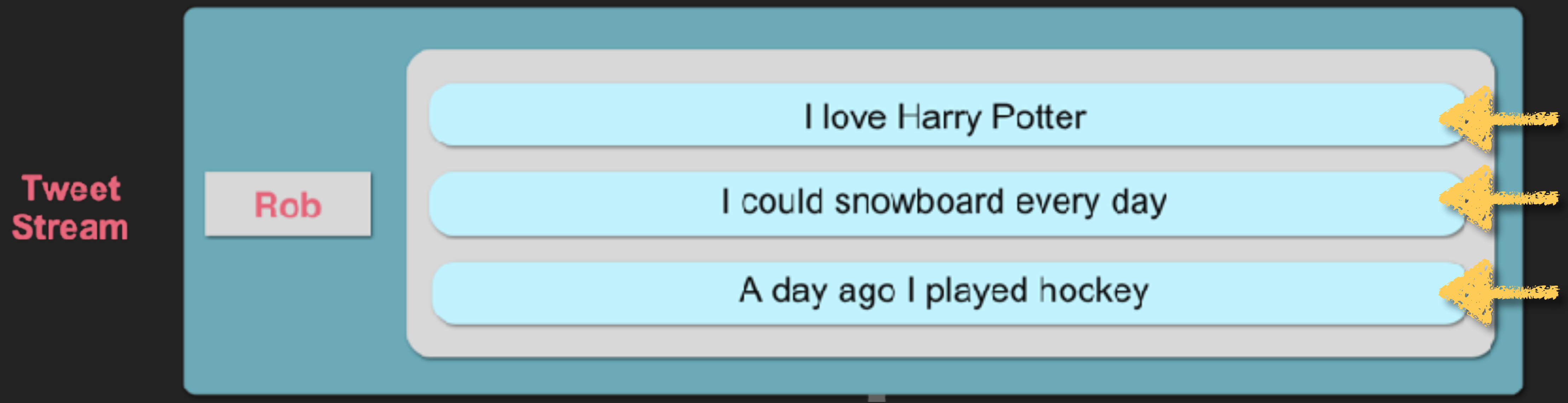


If so, still retain data utility?

EXAMPLE – UTILITY & PRIVACY OF TWEET STREAMS

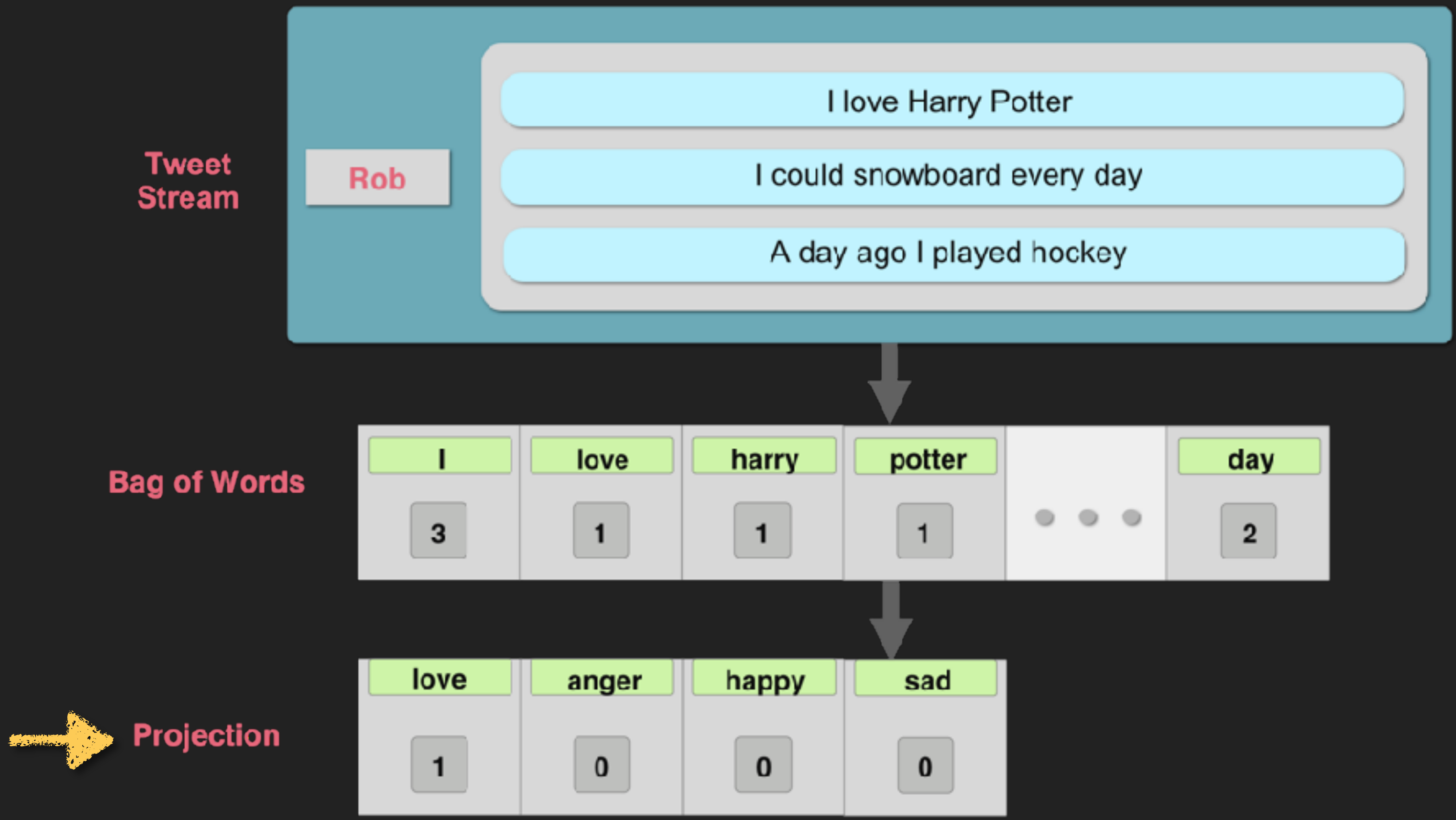


EXAMPLE – UTILITY & PRIVACY OF TWEET STREAMS



EXAMPLE – UTILITY & PRIVACY OF TWEET STREAMS

EXAMPLE – UTILITY & PRIVACY OF TWEET STREAMS



PROBLEM: TRADE-OFFS BETWEEN PRIVACY AND UTILITY

- Given a set of Twitter users, determine the tradeoff between **privacy** and **utility** by considering different projections of their tweet streams

RELATED LITERATURE

RELATED LITERATURE – PRIVACY

- Evaluate how much information is revealed by directly publishing data on the web [Singh et al., 2015]
  - Small number of social media attributes can allow for unique identification
- Measure privacy and utility loss arising from anonymization techniques utilized in microdata publishing [Li and Li, 2009]
  - Privacy is an individual concept and utility an aggregate concept
- Proposal of k-anonymity privacy measure [Sweeney, 2002]

RELATED LITERATURE - EMOTION

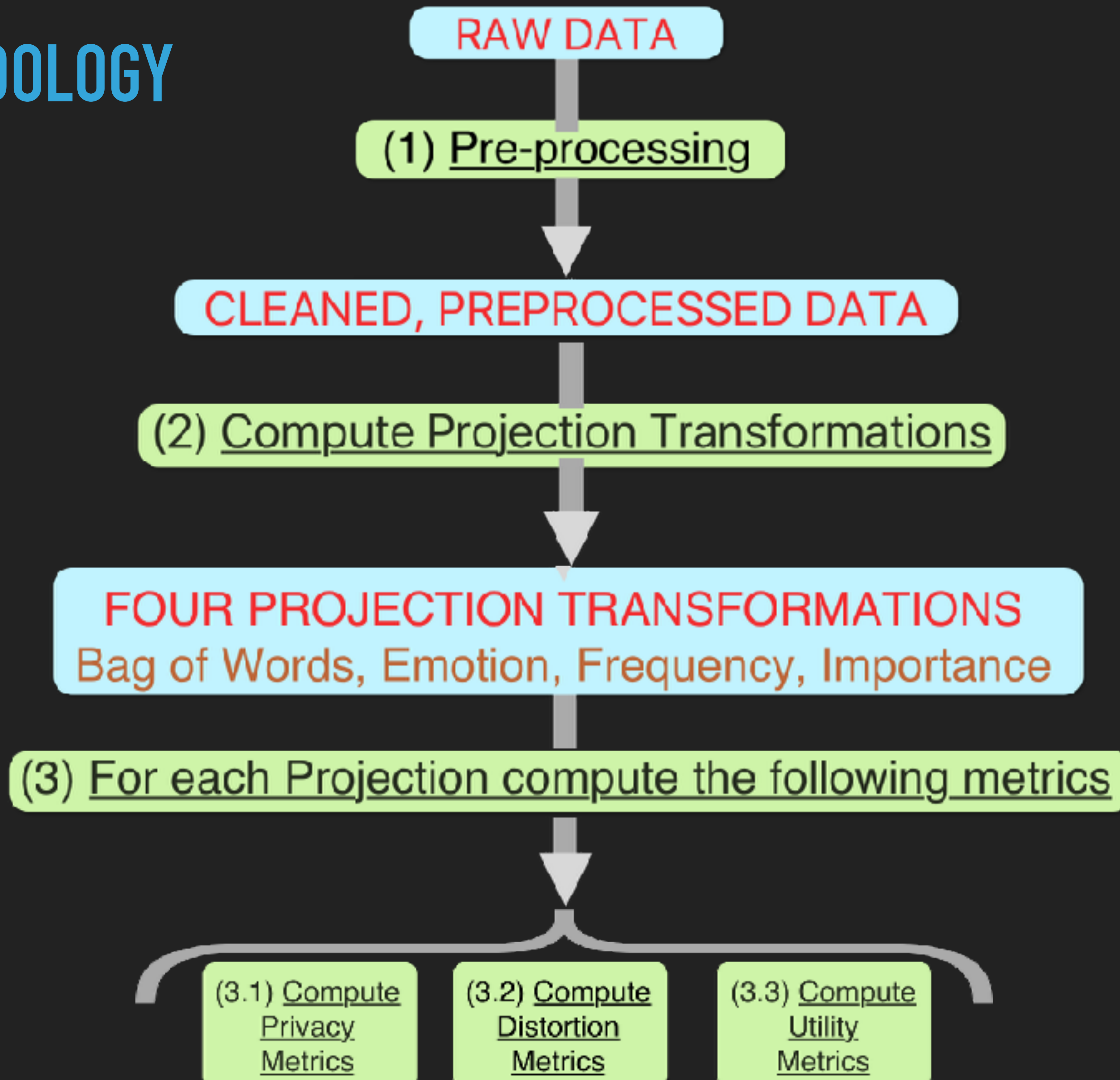
- Propose a bootstrapping algorithm to learn hashtags that convey emotion [Qadir and Riloff, 2013]
- Demonstration of positivity bias in human emotion [Doddsa et al.]
- Propose several challenges unique to the classification of individual emotions [Roberts et al., 2012]
- Propose methods for the construction and evaluation of emotional lexicons that use emoji [Yang et al., 2007]

RELATED LITERATURE – TEXT SUMMARIZATION

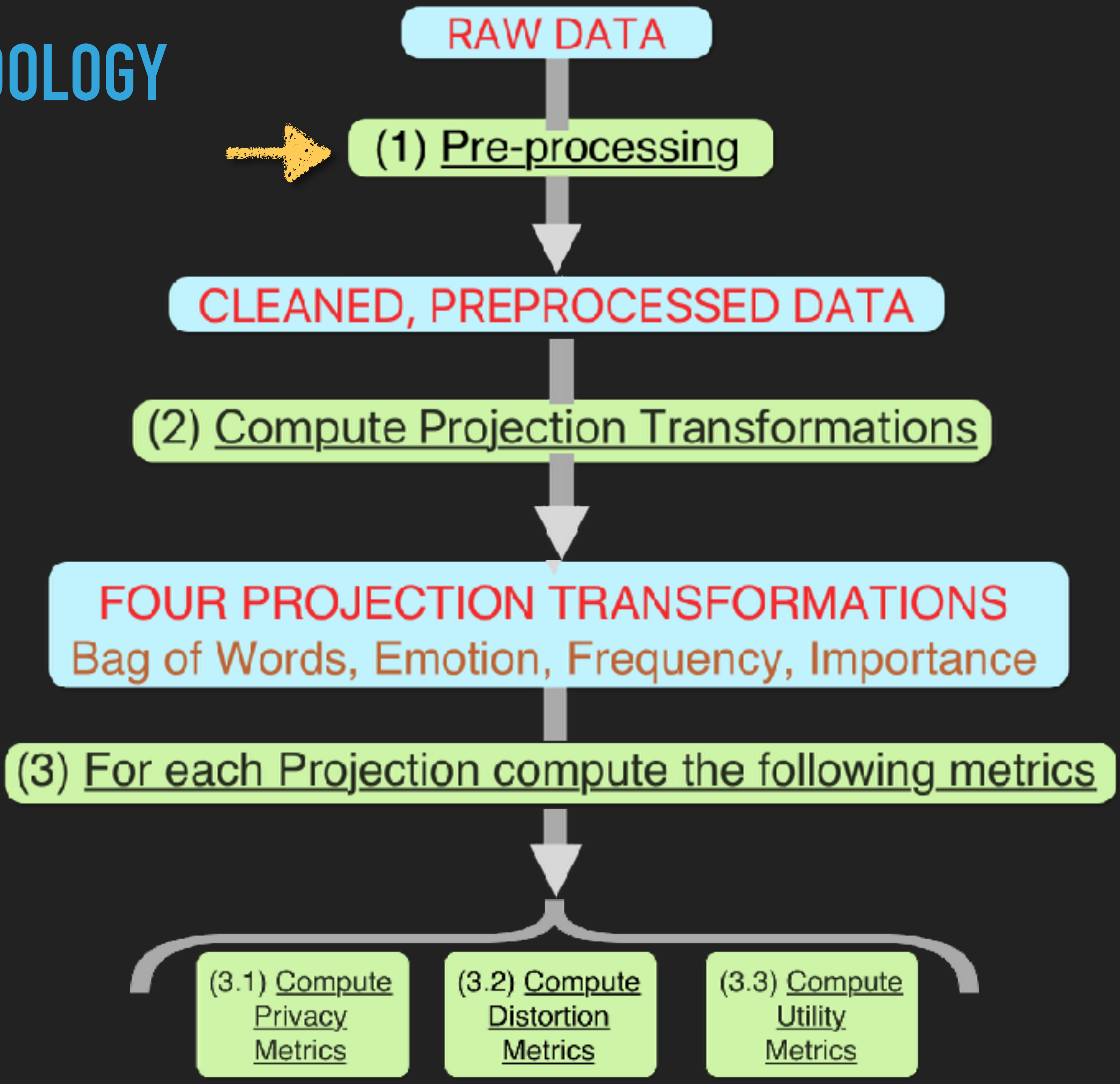
- Evaluation of novel word-based text compressions in terms of speed and achieved compression factor [Horspool and Cormack, 1992]
- Evaluation of word-based text compressions in terms of subjective quality [Witten et al., 2005]

METHODS

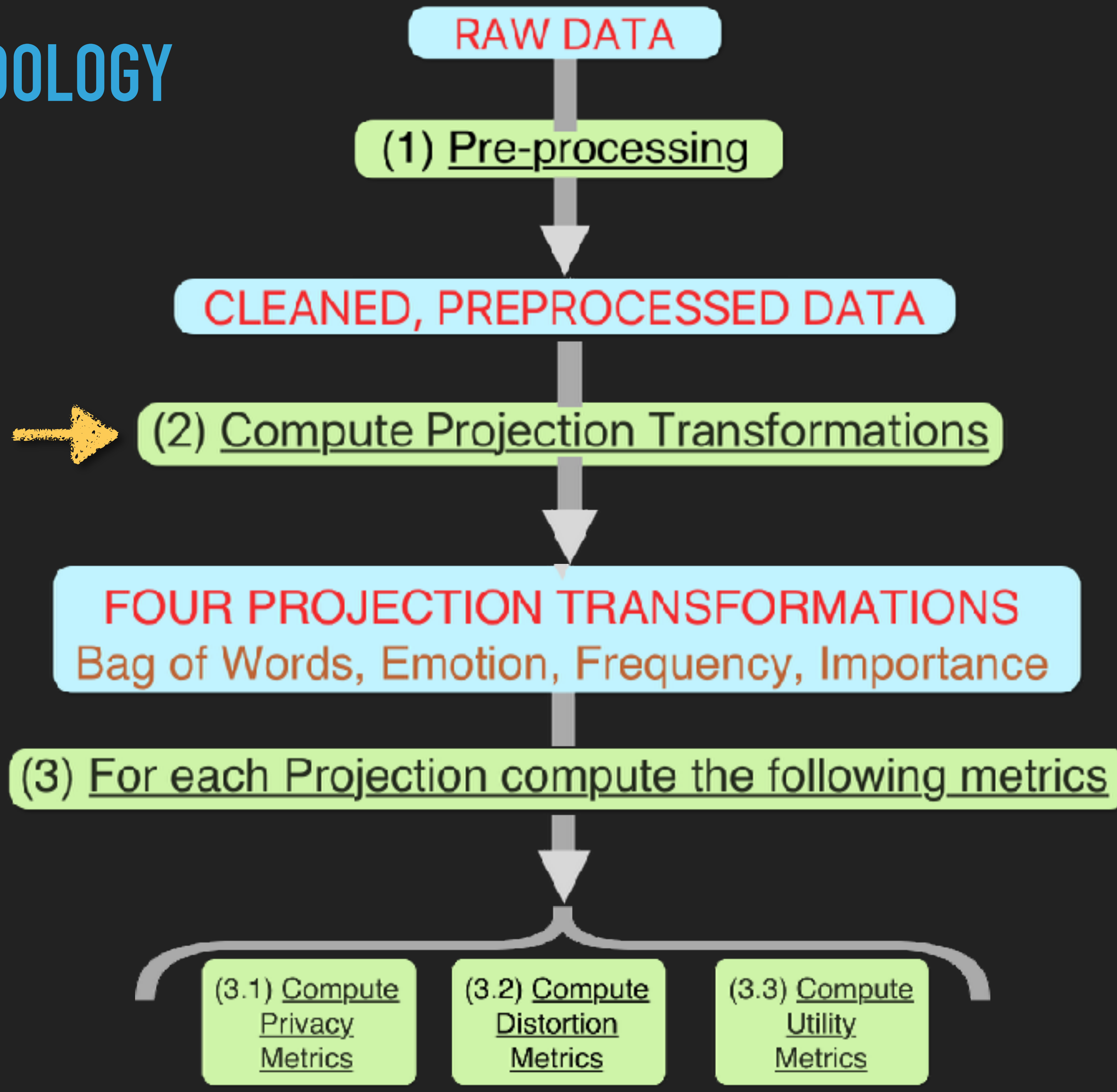
GENERAL METHODOLOGY



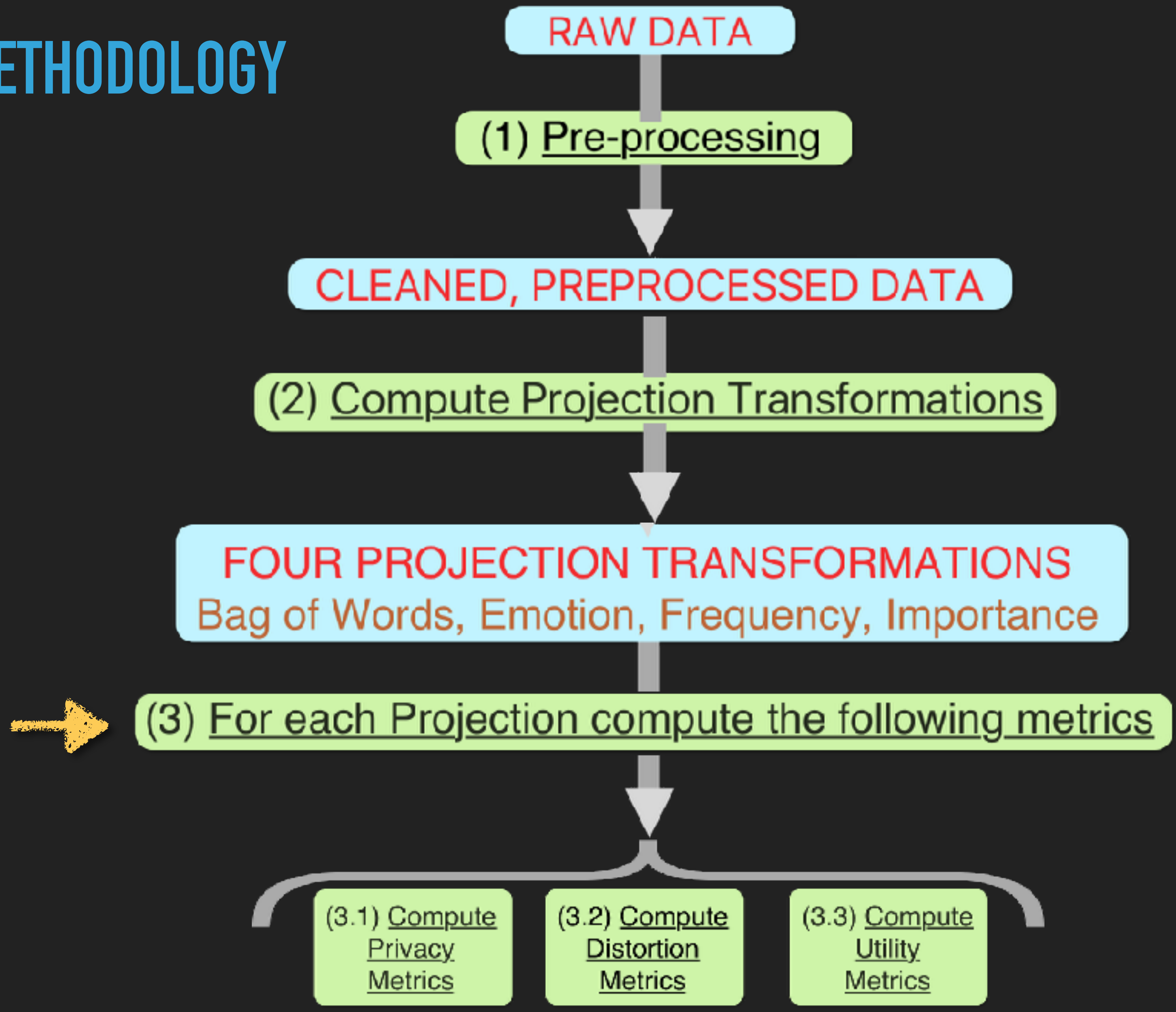
GENERAL METHODOLOGY



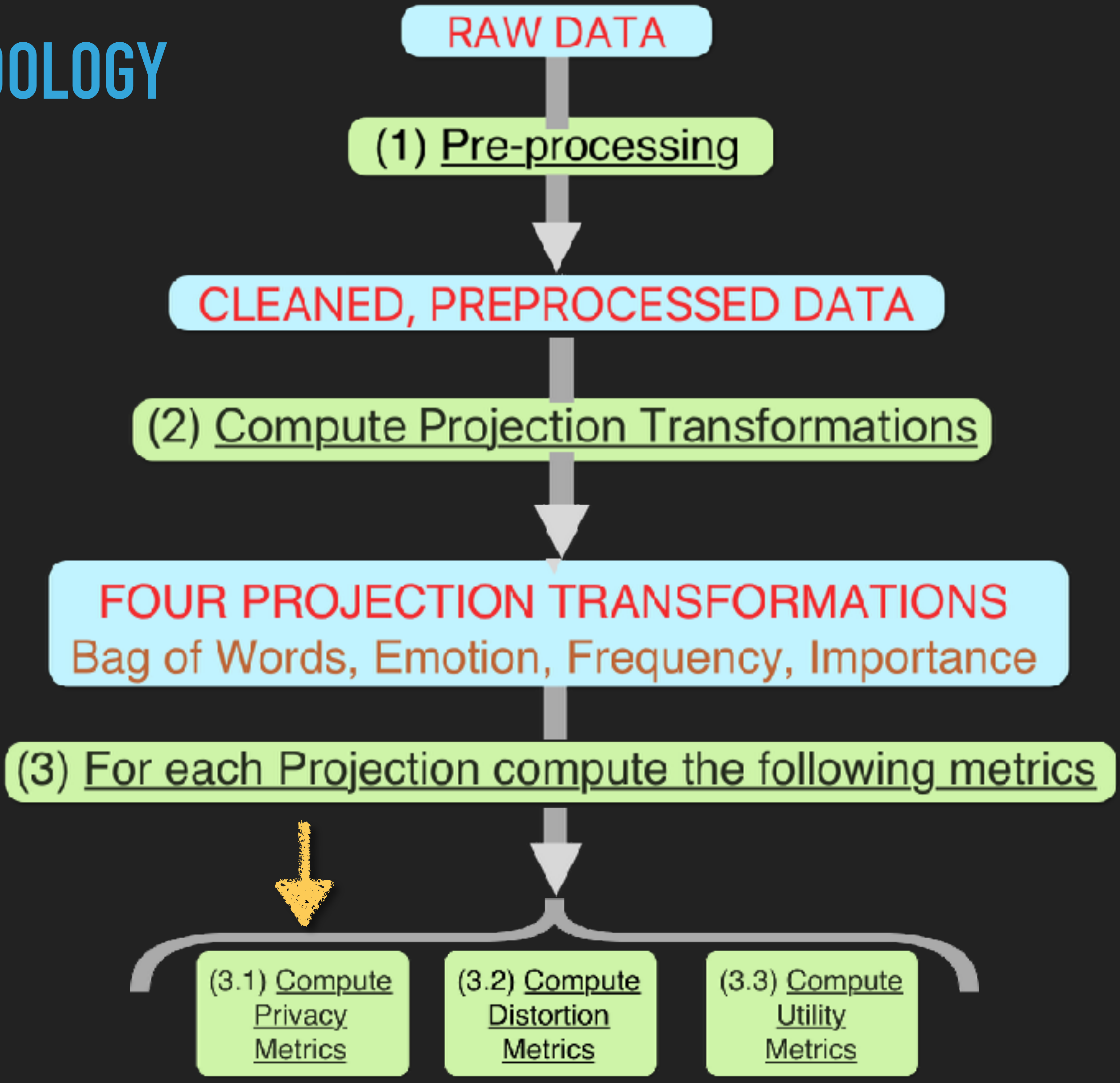
GENERAL METHODOLOGY



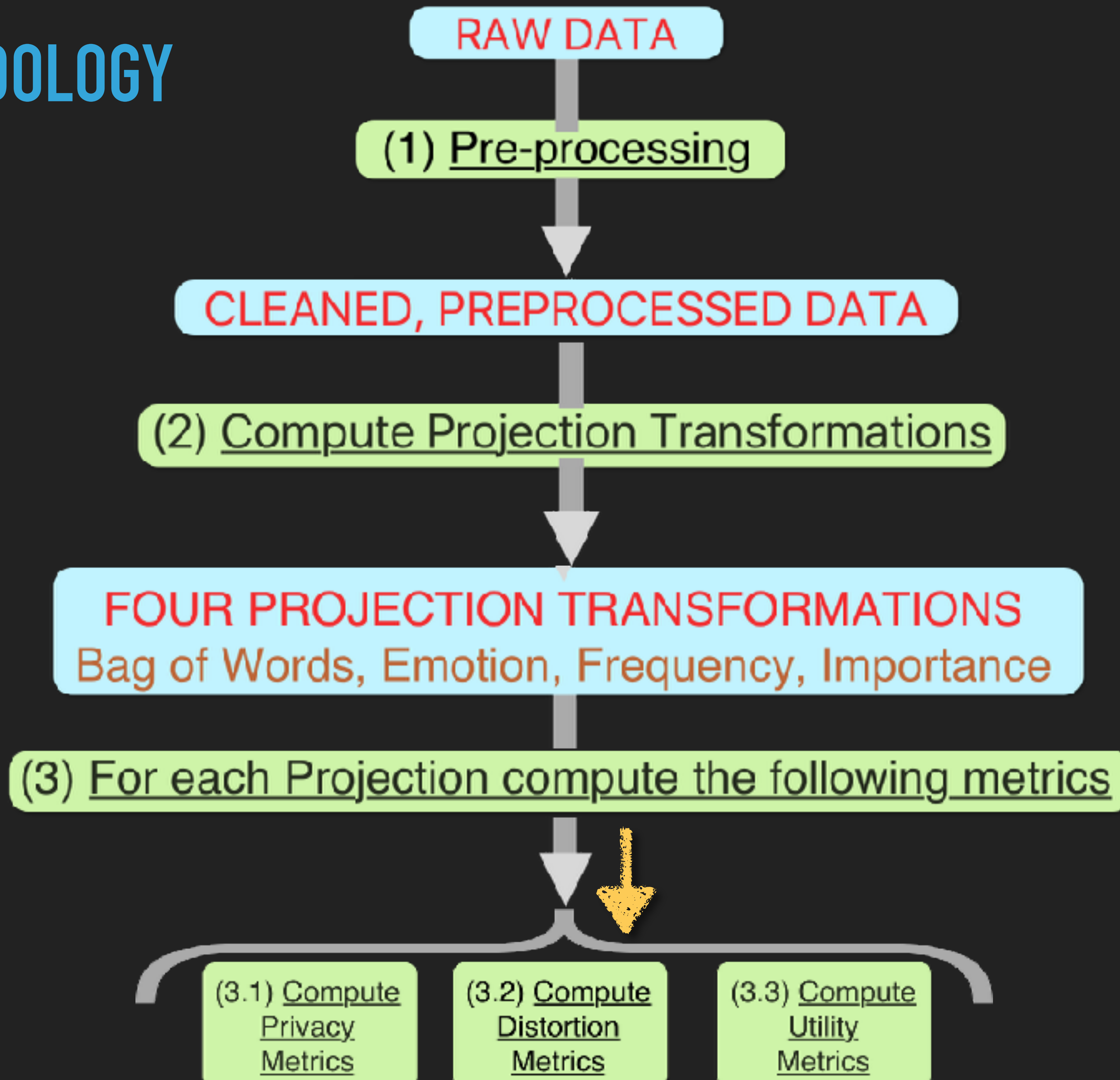
GENERAL METHODOLOGY



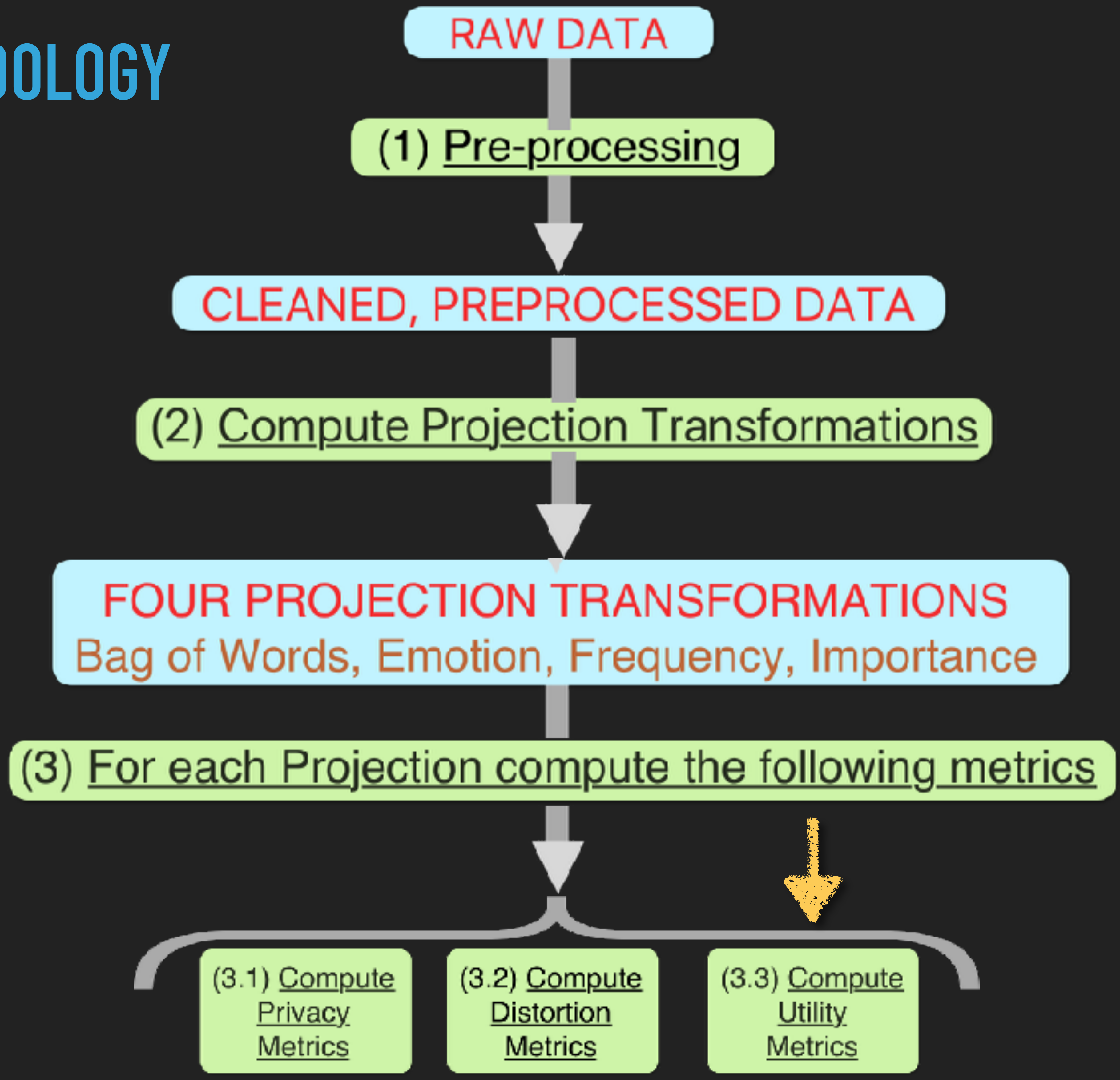
GENERAL METHODOLOGY



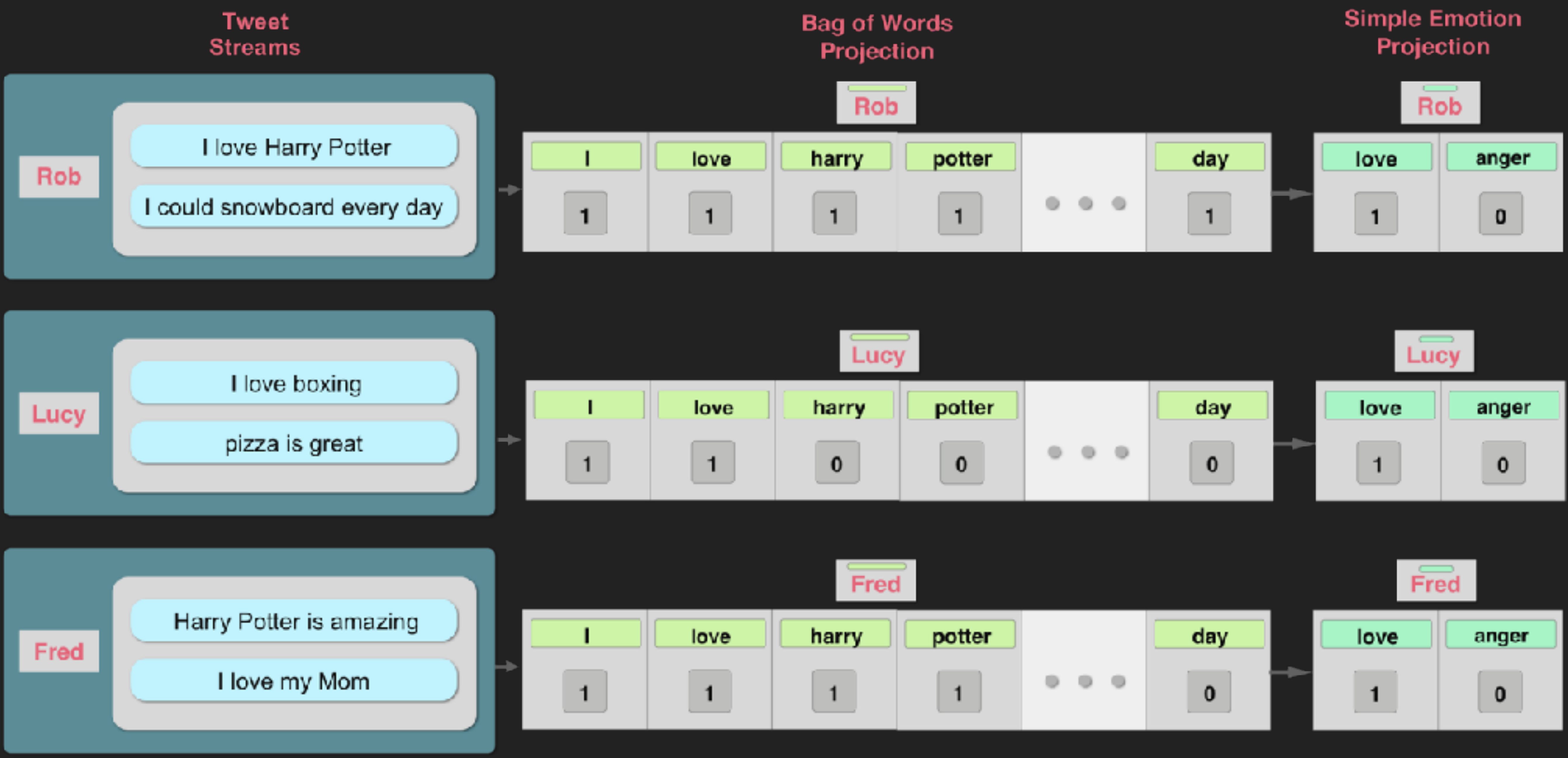
GENERAL METHODOLOGY



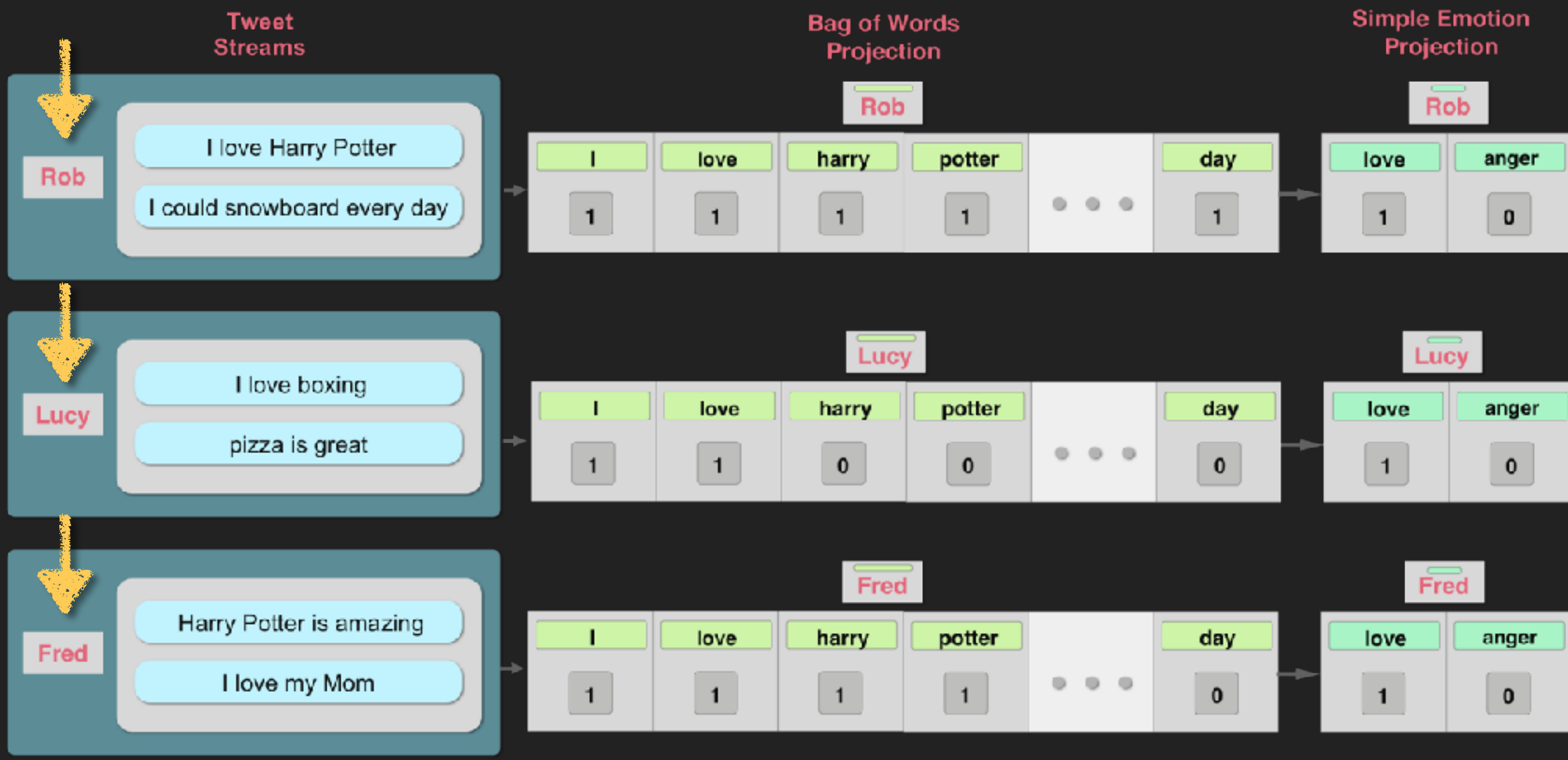
GENERAL METHODOLOGY



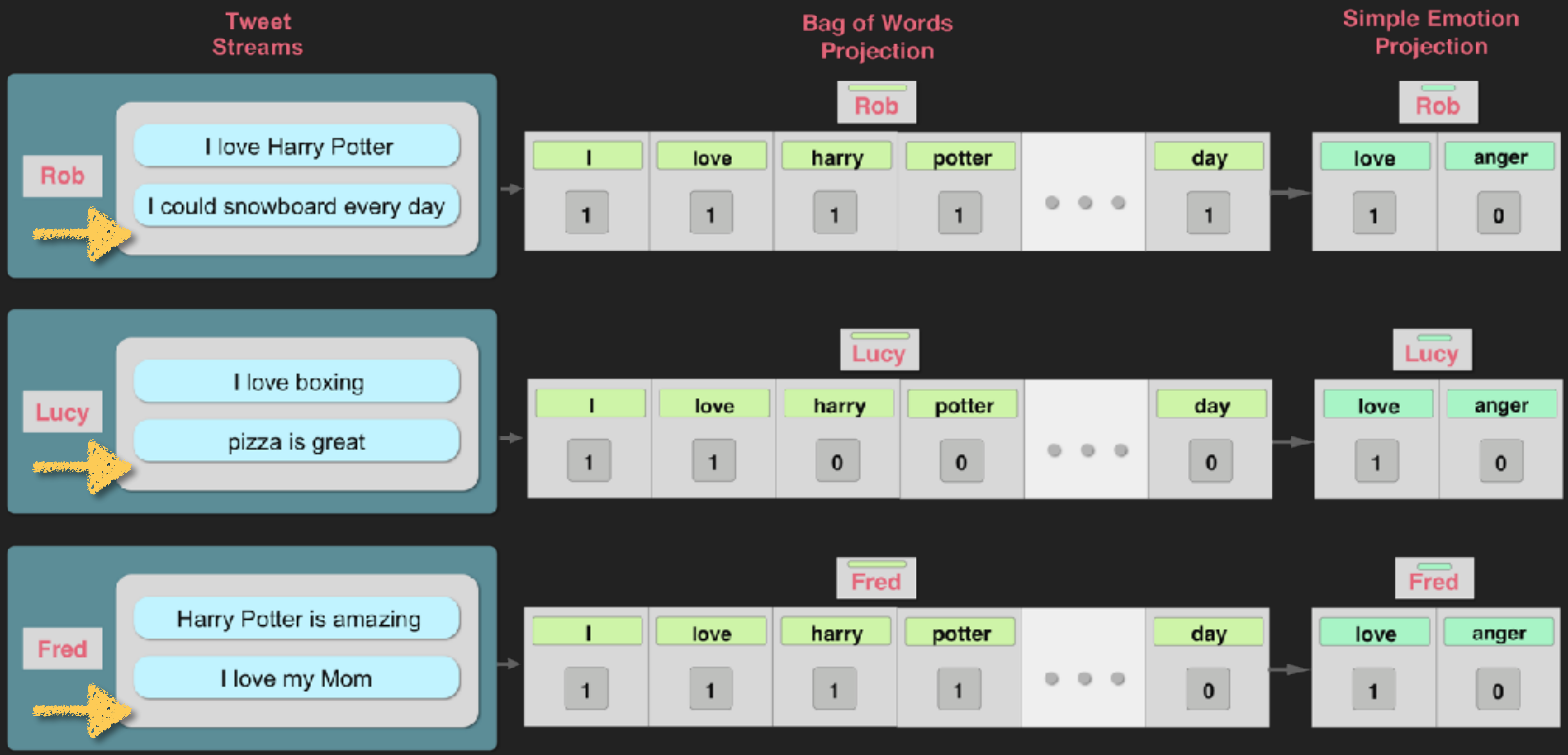
METHODOLOGY EXAMPLE



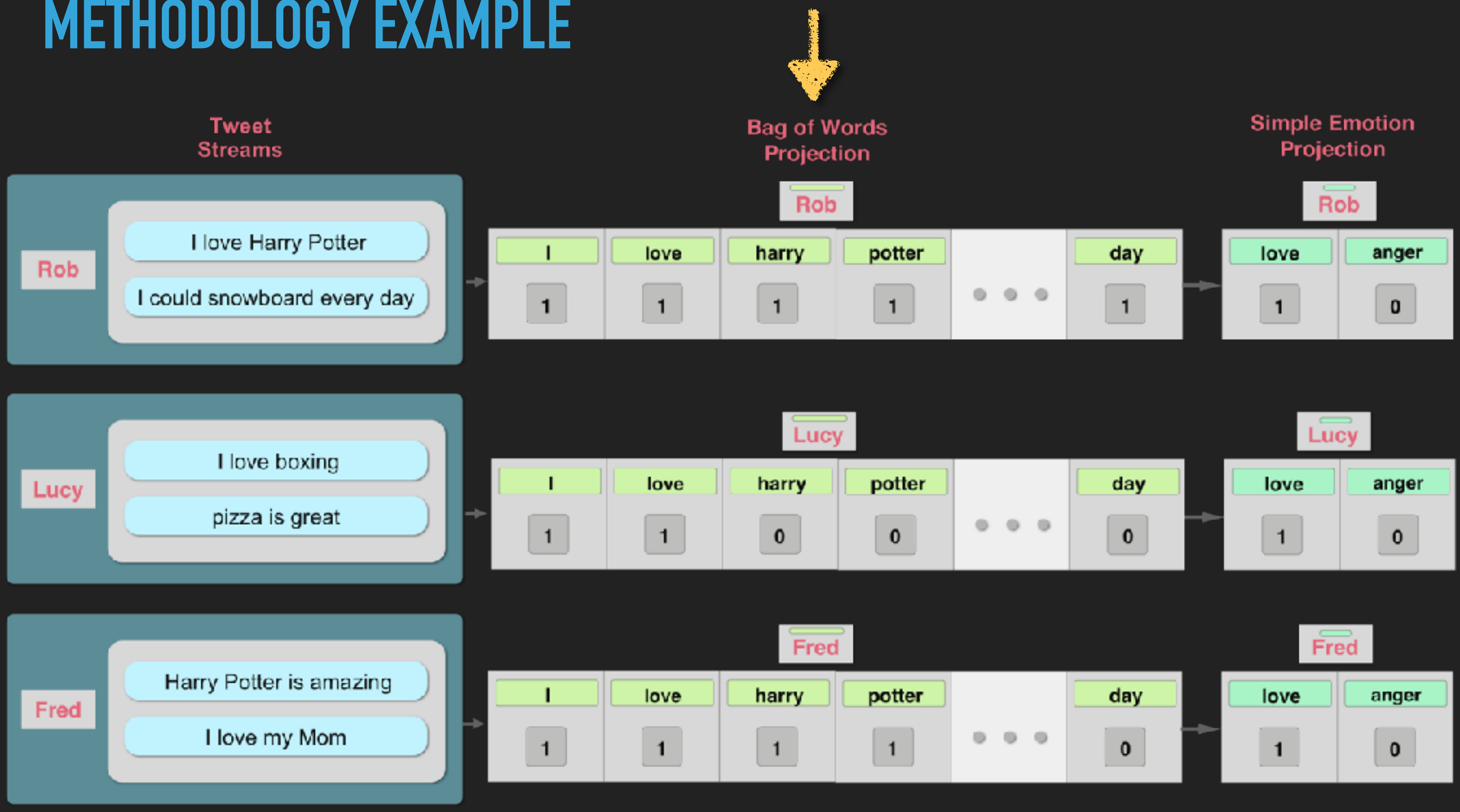
METHODOLOGY EXAMPLE



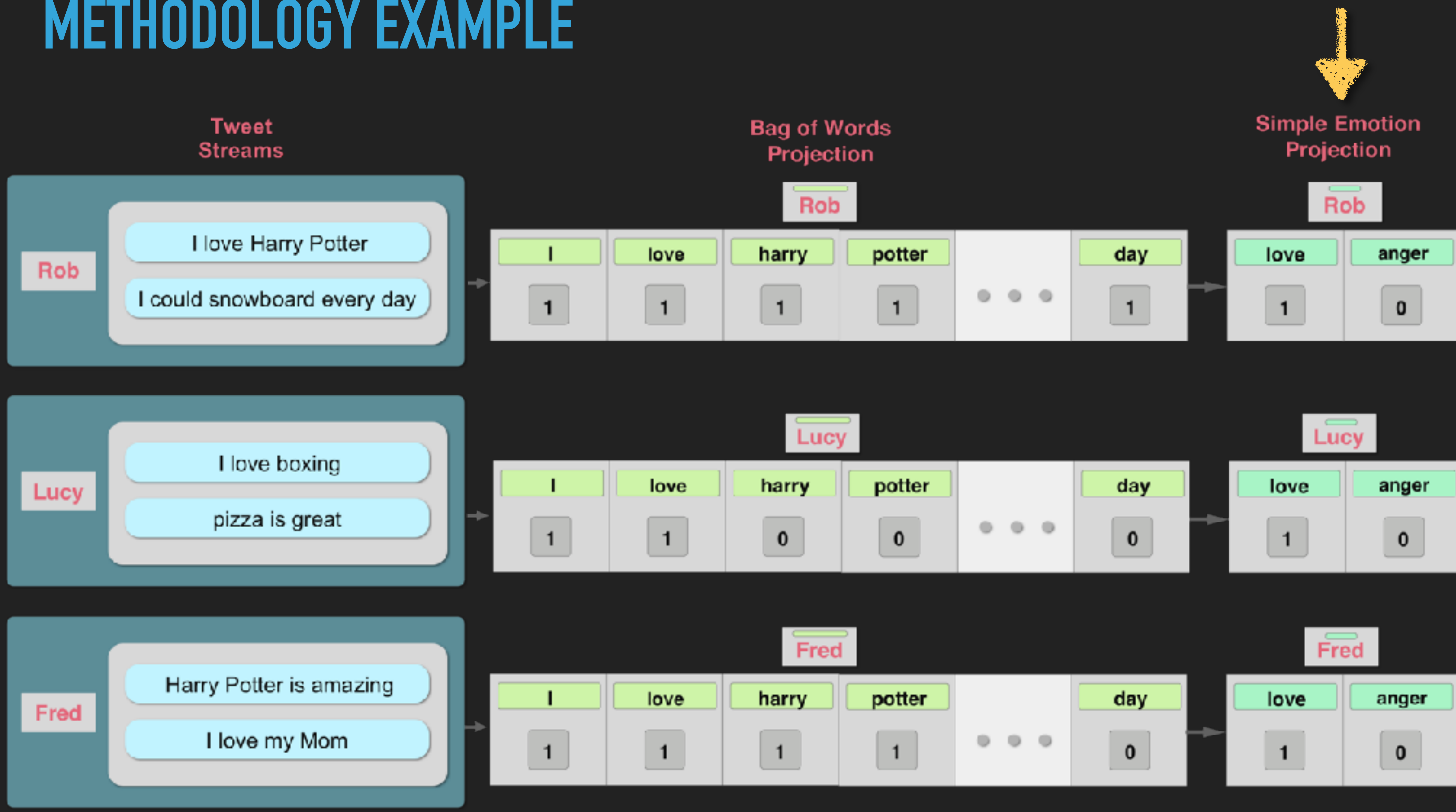
METHODOLOGY EXAMPLE



METHODOLOGY EXAMPLE

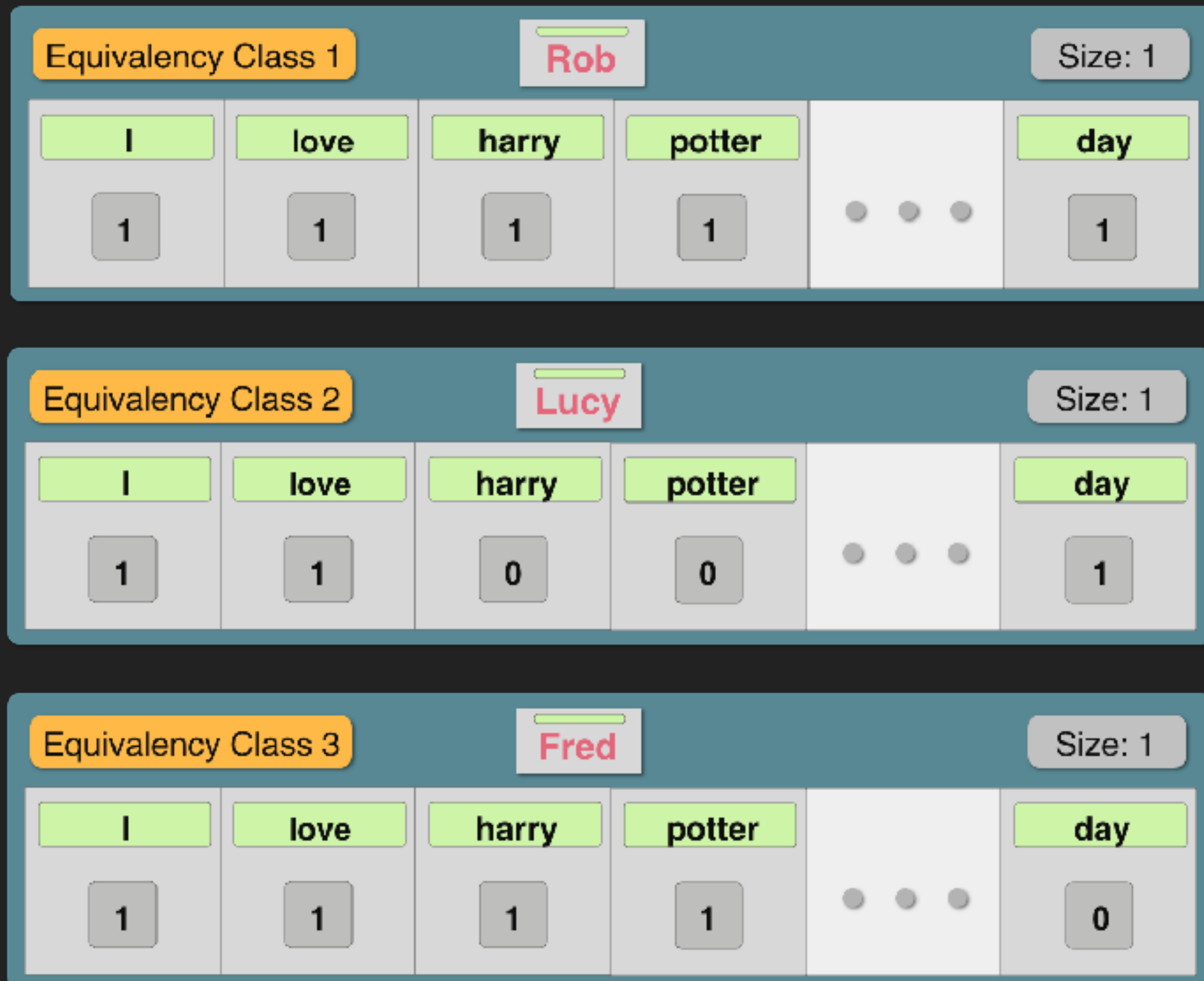


METHODOLOGY EXAMPLE



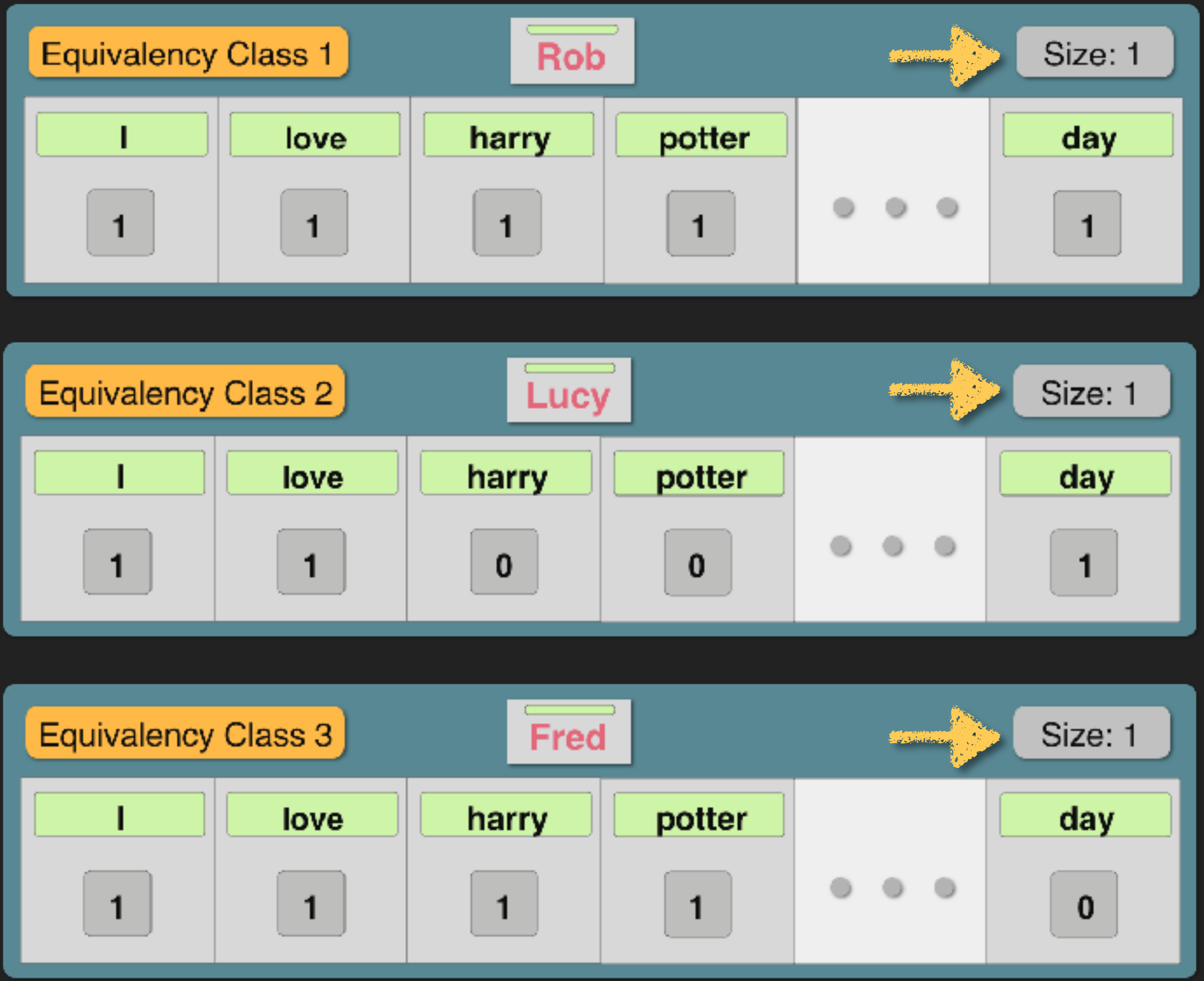
METHODOLOGY EXAMPLE – PRIVACY

- Bag of Words



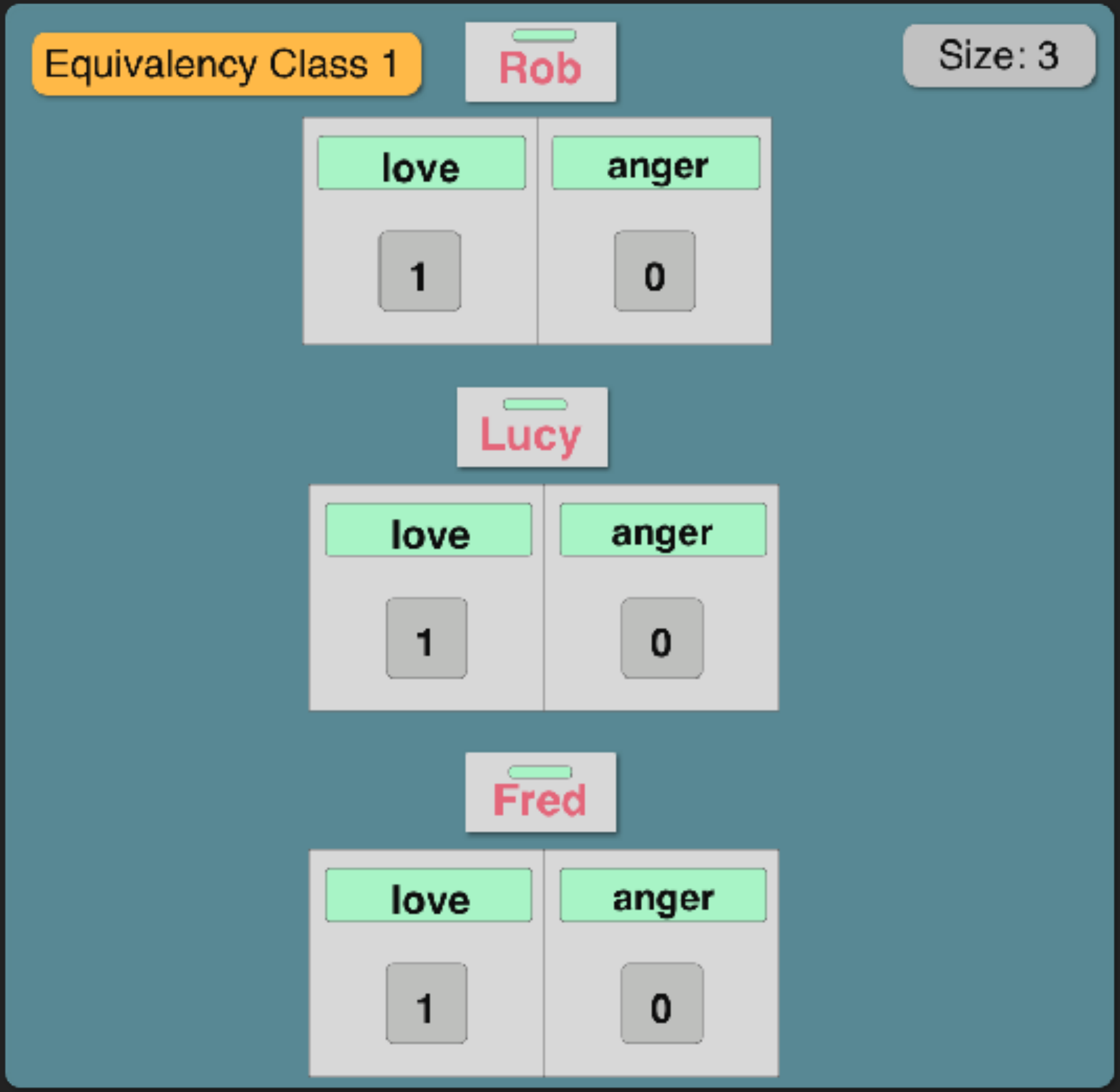
METHODOLOGY EXAMPLE – PRIVACY

- Bag of Words
- Lowest size equivalence class is 1
- k=1 for k-anonymity of the vector space



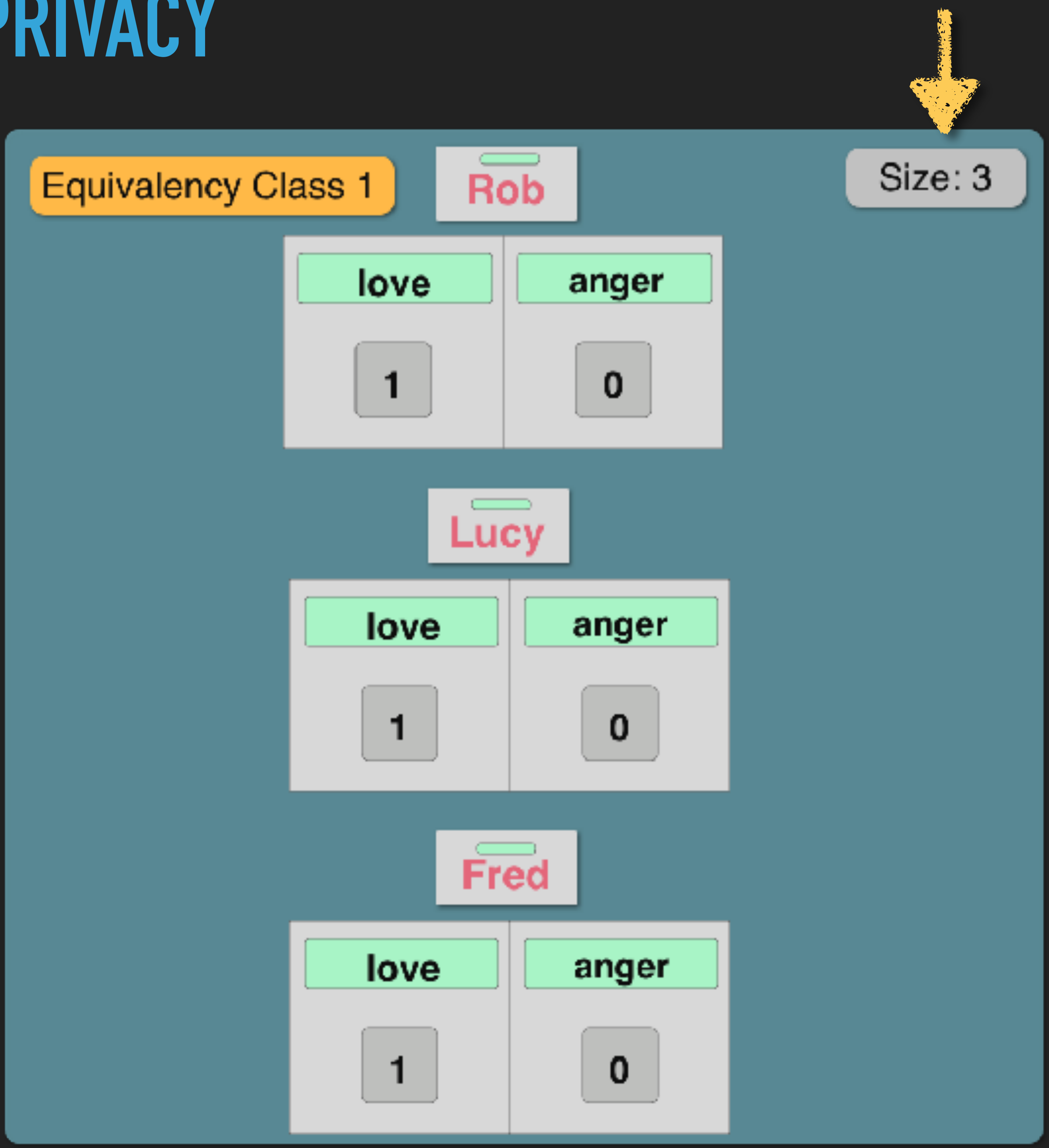
METHODOLOGY EXAMPLE – PRIVACY

- Emotion Projection



METHODOLOGY EXAMPLE – PRIVACY

- Emotion Projection
- k = 3 for k-anonymity of the vector space



DISTORTION

- Projections of tweet streams result in distortion
- Loss of potentially distinguishing features
- High distortion, fewer features to distinguish an individual

UTILITY

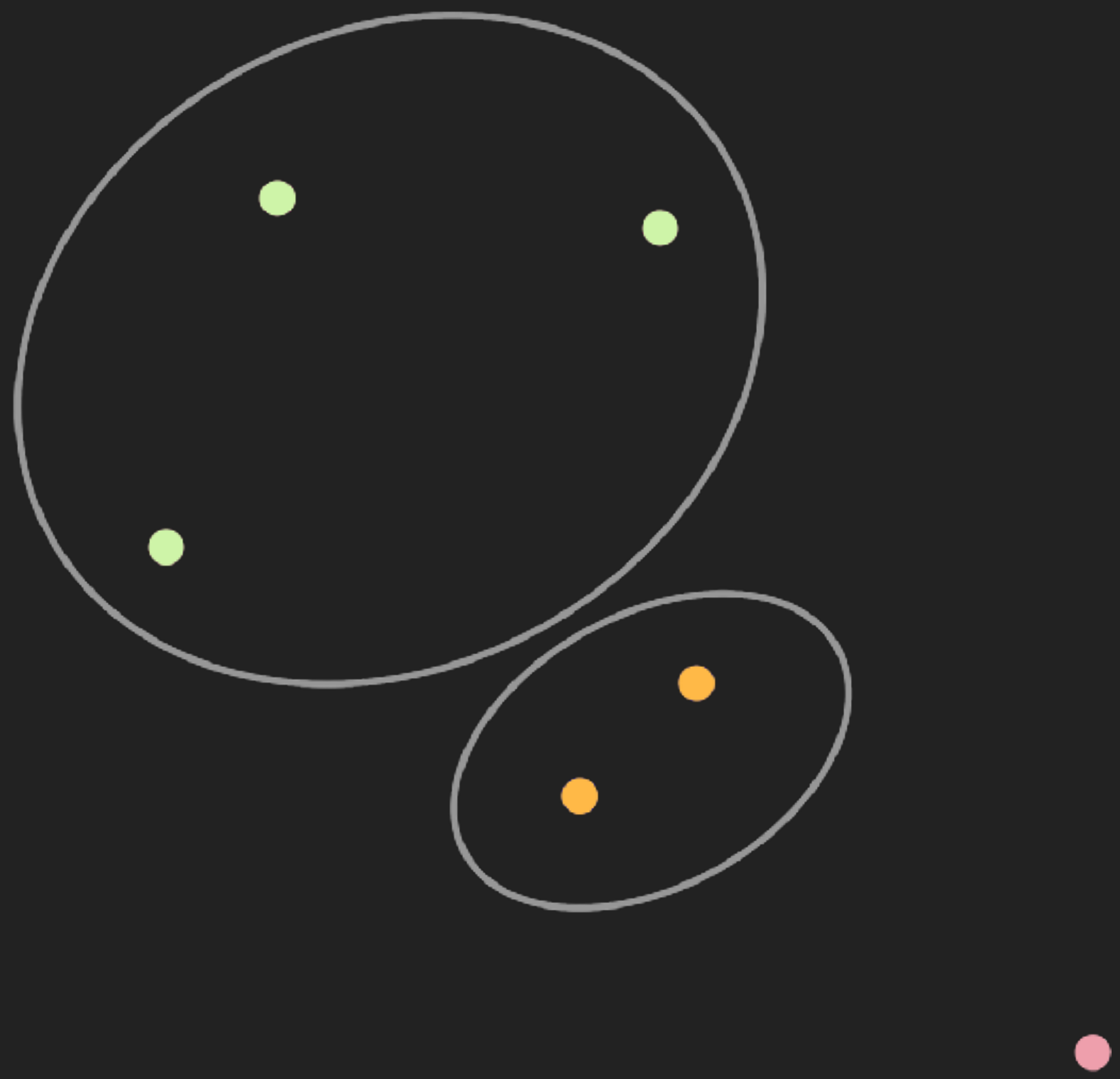
- Clustering
- Anomaly Detection
- Frequent Item Set Mining

CLUSTERING – K MEANS

- K Means [Lloyd, 1957]
  - Aims to partition *n* observations into *k* clusters
  - Each observation belongs to cluster with nearest mean

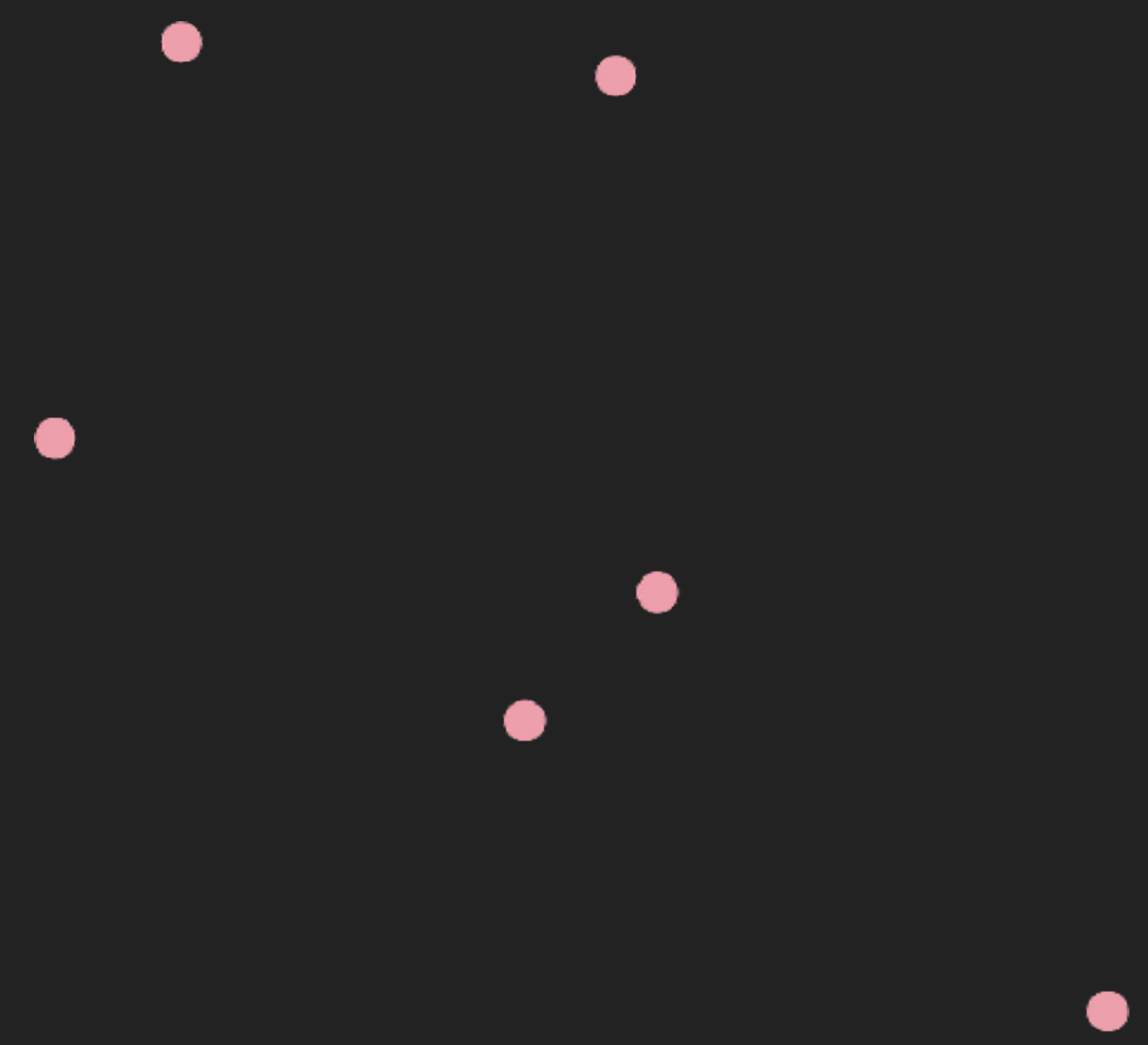
CLUSTERING – K MEANS

- k = 2



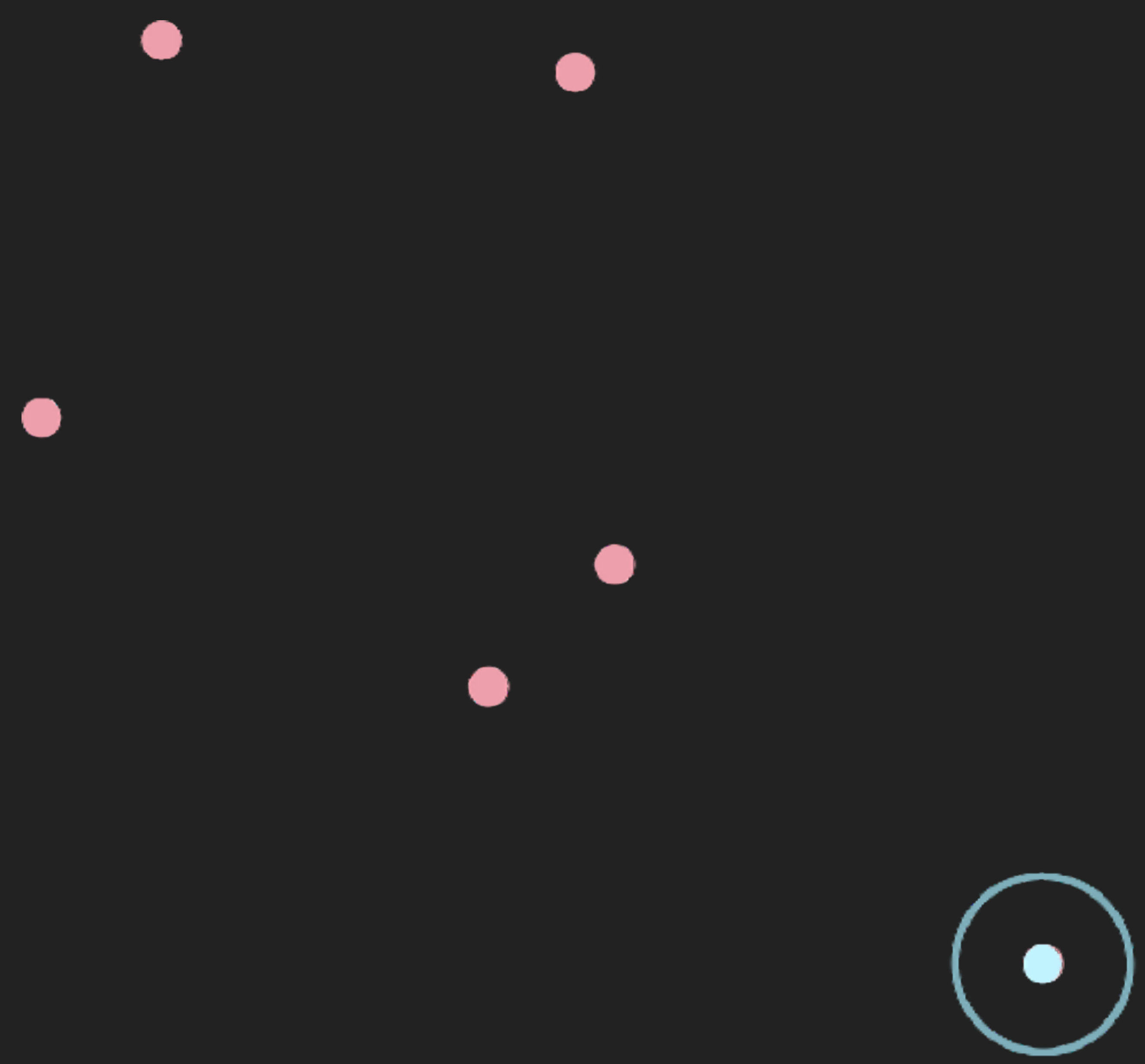
ANOMALY DETECTION – LOCAL OUTLIER FACTOR

- Local Outlier Factor [Breunig et al., 2000]
  - Observations evaluated according to density of their neighborhood
  - The lower the density of observations near an observation, the more anomalous the observation



ANOMALY DETECTION – LOCAL OUTLIER FACTOR

- Local Outlier Factor [Breunig et al., 2000]
  - Observations evaluated according to density of their neighborhood
  - The lower the density of observations near an observation, the more anomalous the observation



FREQUENT ITEM SET MINING

- Frequent Item Set Mining [Agrawal et al., 1993]
  - Compute sets of features that co-occur above a specified frequency (minimum support)

FREQUENT ITEM SET MINING

- Frequent Item Set Mining [Agrawal et al., 1993]
  - Compute sets of features that co-occur above a specified frequency (minimum support)

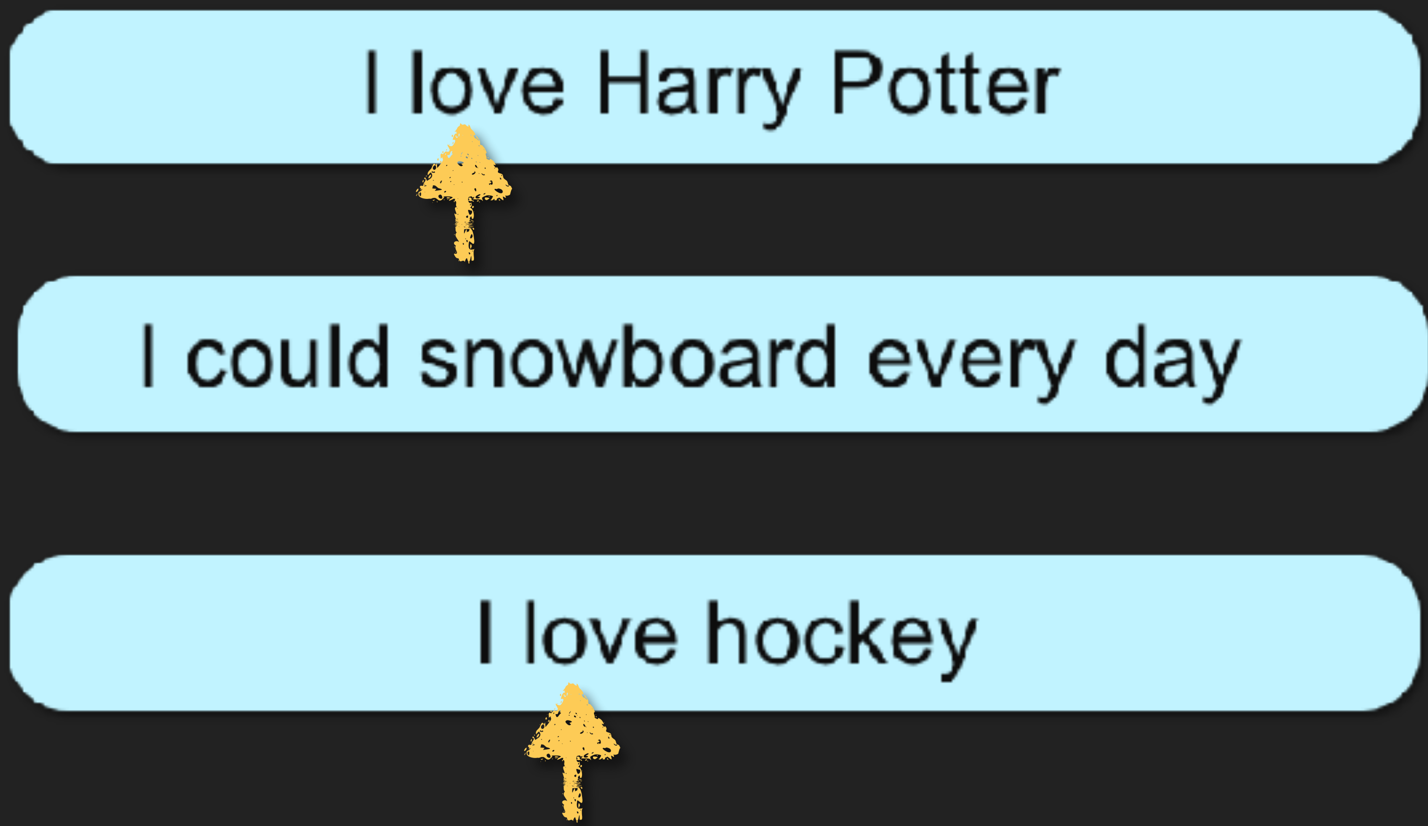
I love Harry Potter

I could snowboard every day

I love hockey

FREQUENT ITEM SET MINING

- Frequent Item Set Mining [Agrawal et al., 1993]
  - Compute sets of features that co-occur above a specified frequency (minimum support)

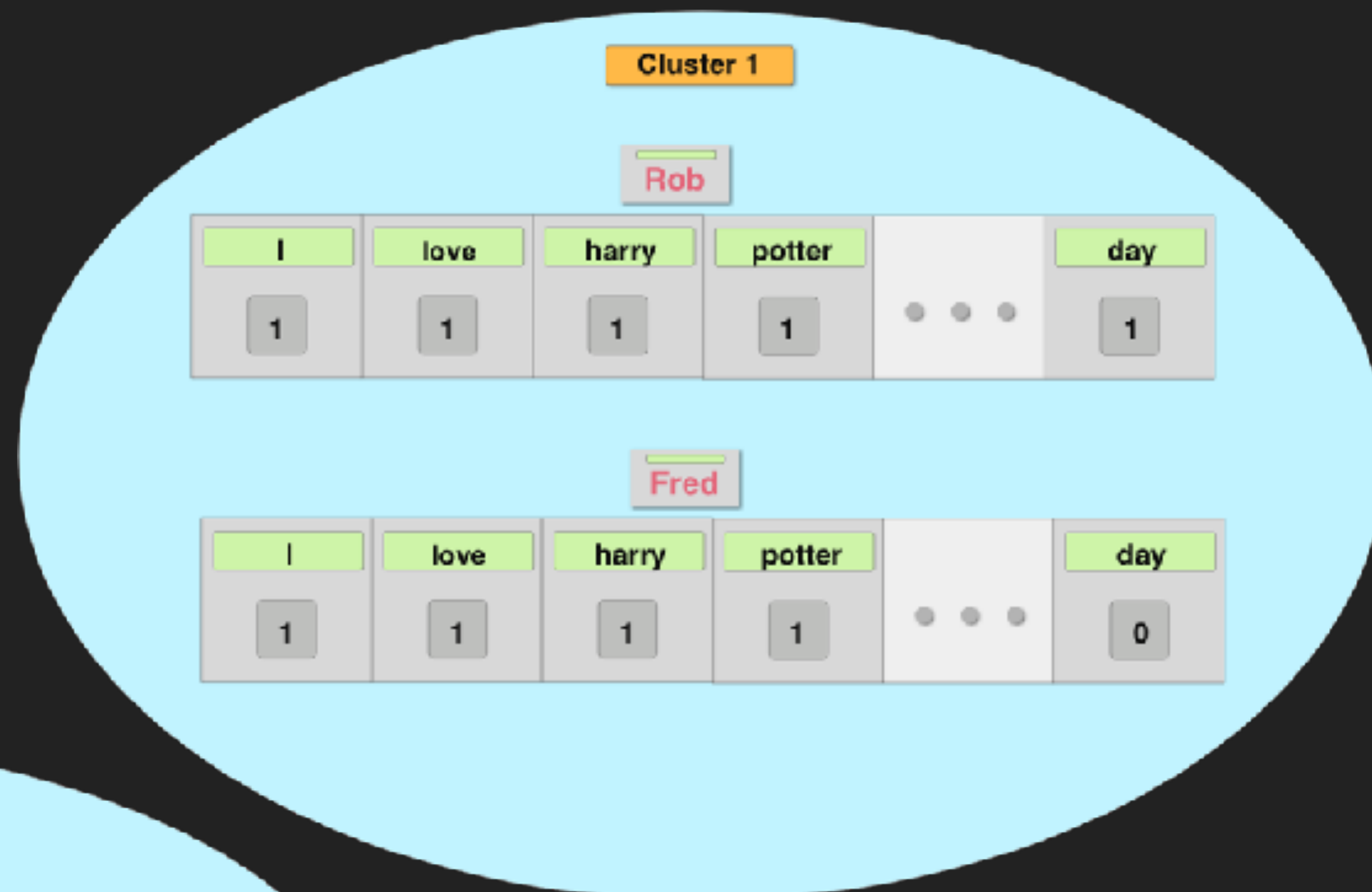


FREQUENT ITEM SET MINING

- Frequent Item Set Mining [Agrawal et al., 1993]
  - Compute sets of features that co-occur above a specified frequency (minimum support)

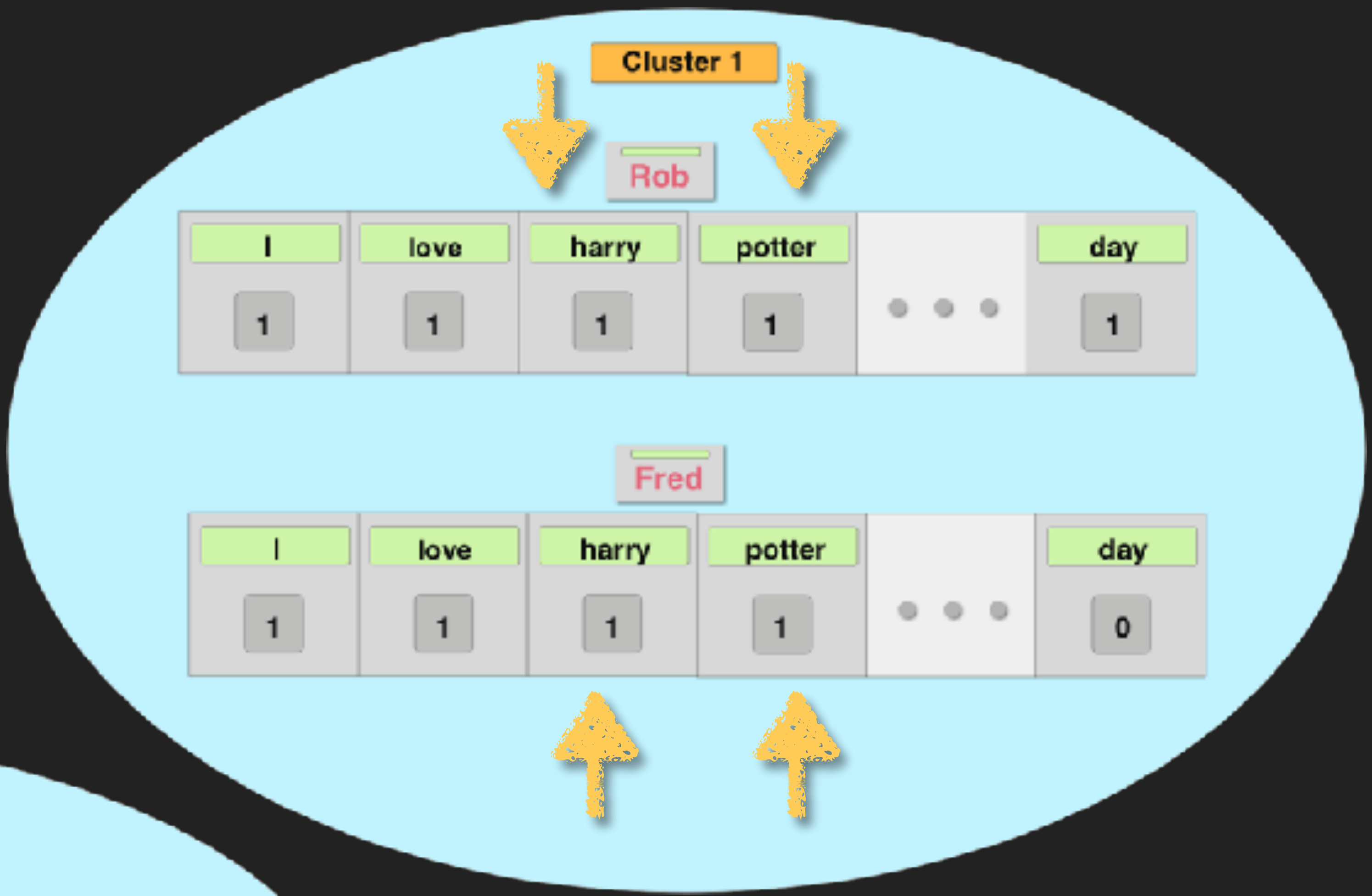
METHODOLOGY - UTILITY & DISTORTION EXAMPLE

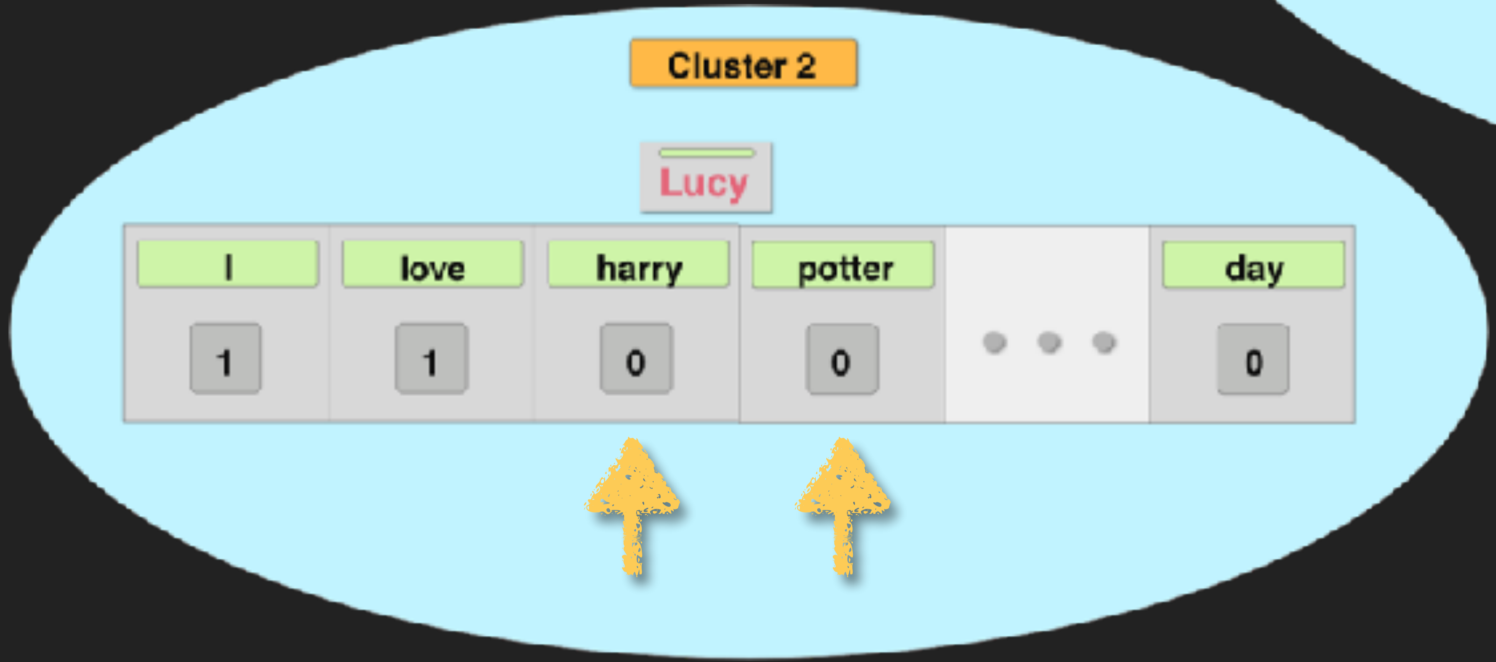
- Clustering
  - Bag of Words



METHODOLOGY - UTILITY LOSS FROM DISTORTION EXAMPLE

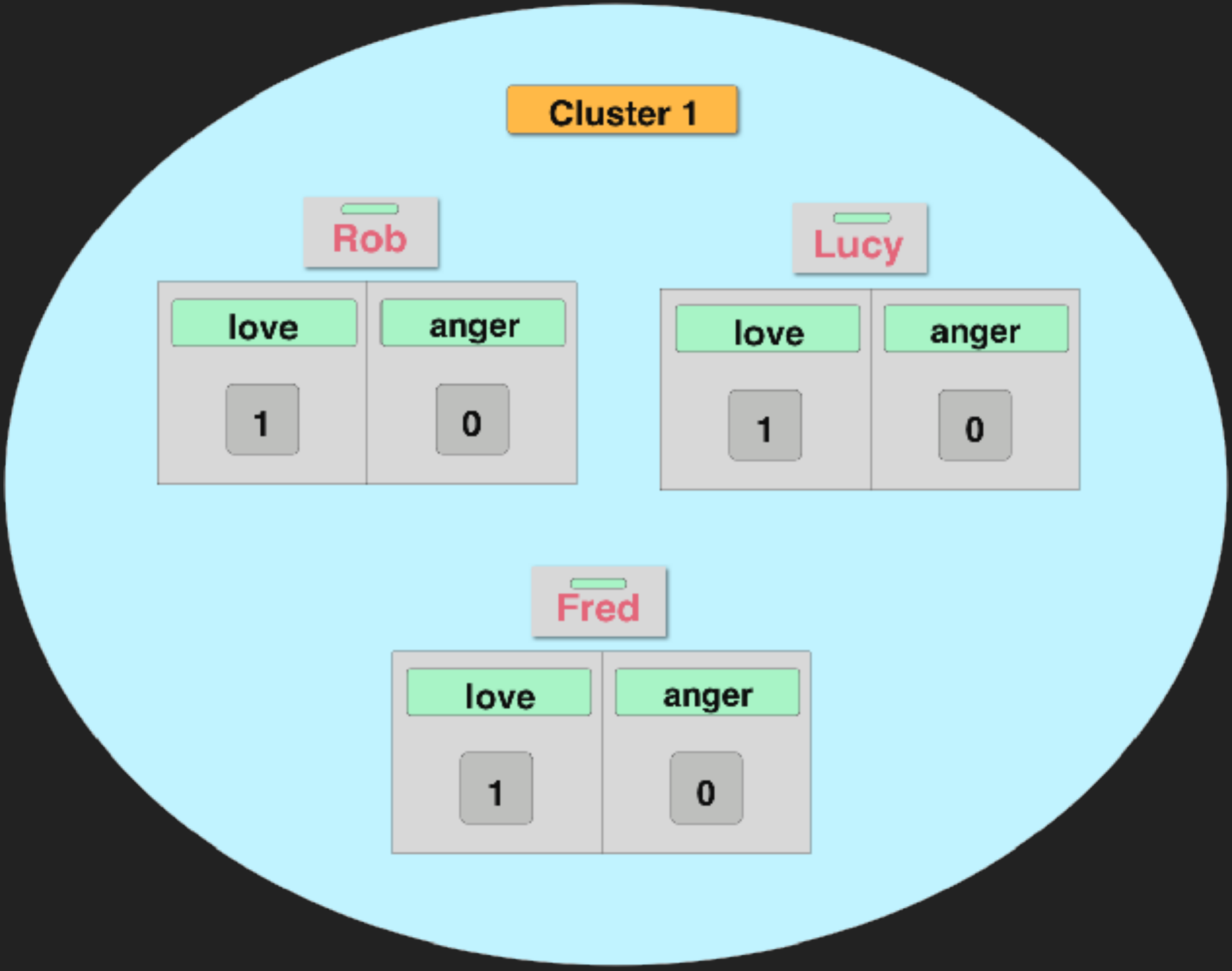
- Clustering
  - Bag of Words





METHODOLOGY - UTILITY LOSS FROM DISTORTION EXAMPLE

- Clustering
  - Emotion Projection
  - Example of high distortion impacting utility



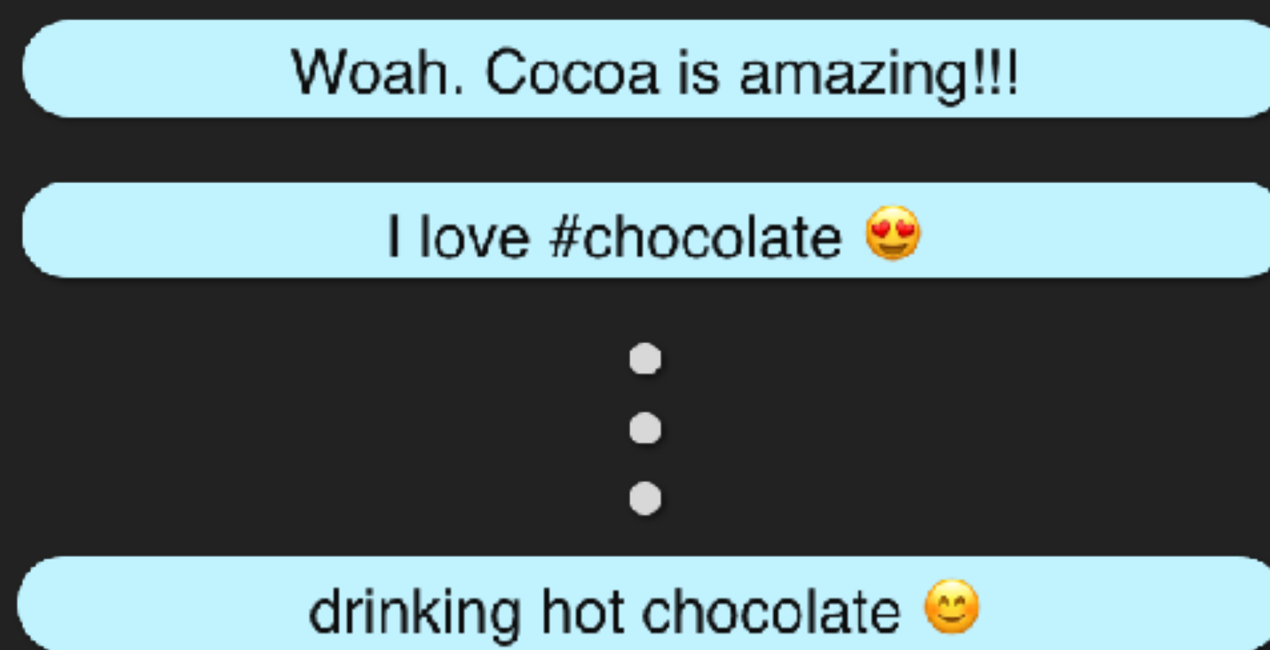
IDEAL DATA TRANSFORMATION

- High data utility
- Low individual distinguishability
- High distortion

CONTENT THAT MAY MATTER

- What makes an individual unique on Twitter?

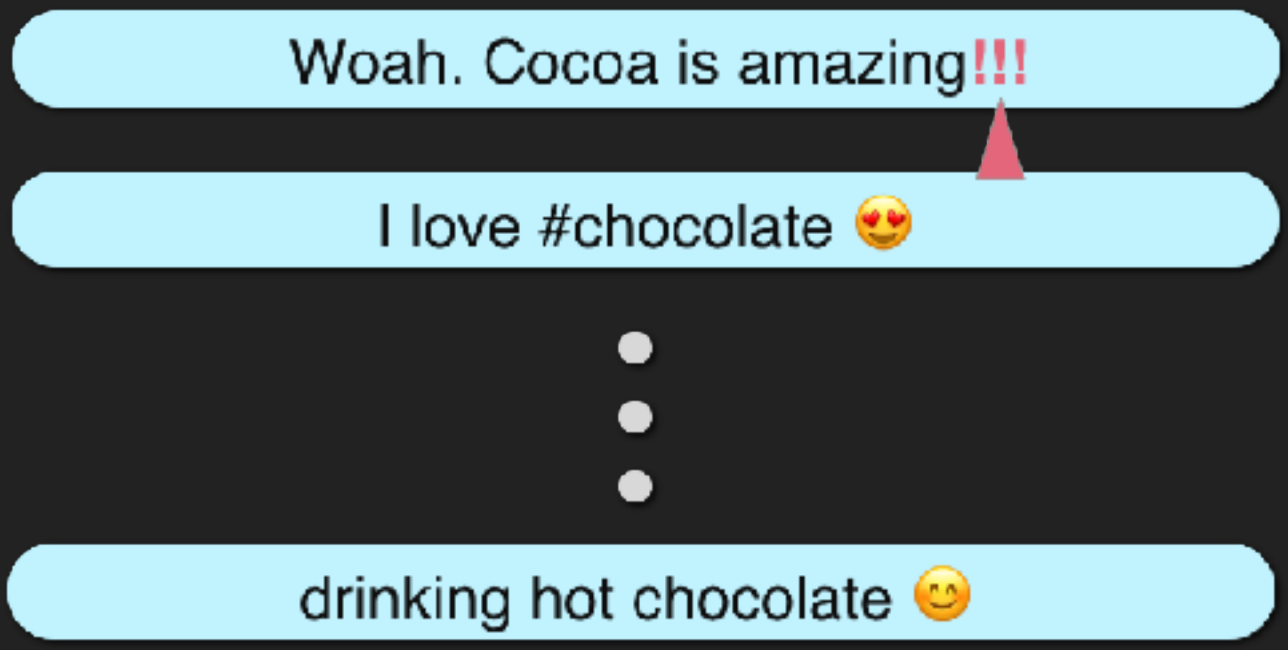
1. Words used
    - punctuation use characteristics
    - feature Frequency
2. Substantive content discussed
3. Expressed emotionality
    - words & emoji
4. Engagement level



CONTENT THAT MAY MATTER

- What makes an individual unique on Twitter?

1. **Words used**
   - **punctuation use characteristics**
   - feature Frequency
2. Substantive content discussed
3. Expressed emotionality
   - words & emoji
4. Engagement level



CONTENT THAT MAY MATTER

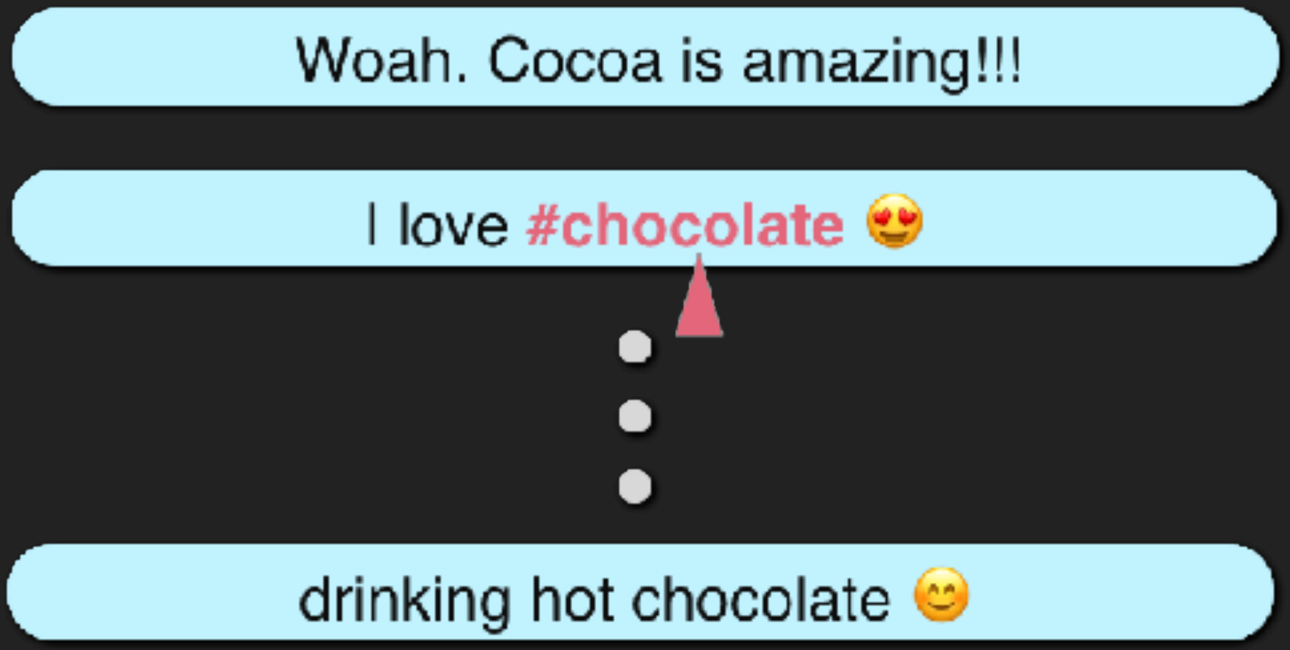
- What makes an individual unique on Twitter?

1. **Words used**
    - punctuation use characteristics
    - **feature Frequency**
2. Substantive content discussed
3. Expressed emotionality
    - words & emoji
4. Engagement level

CONTENT THAT MAY MATTER

- What makes an individual unique on Twitter?

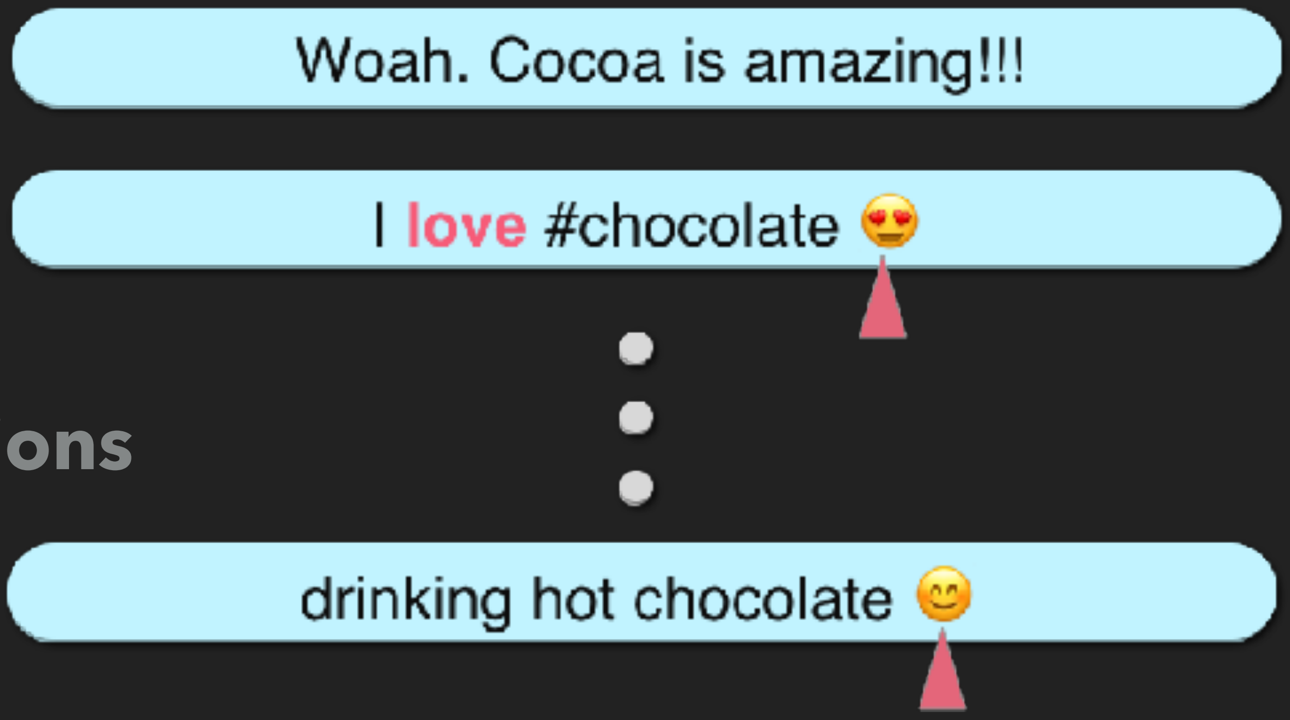
1. Words used
    - punctuation use characteristics
    - feature Frequency
2. **Substantive content discussed**
3. Expressed emotionality
    - words & emoji
4. Engagement level



CONTENT THAT MAY MATTER

- What makes an individual unique on Twitter?

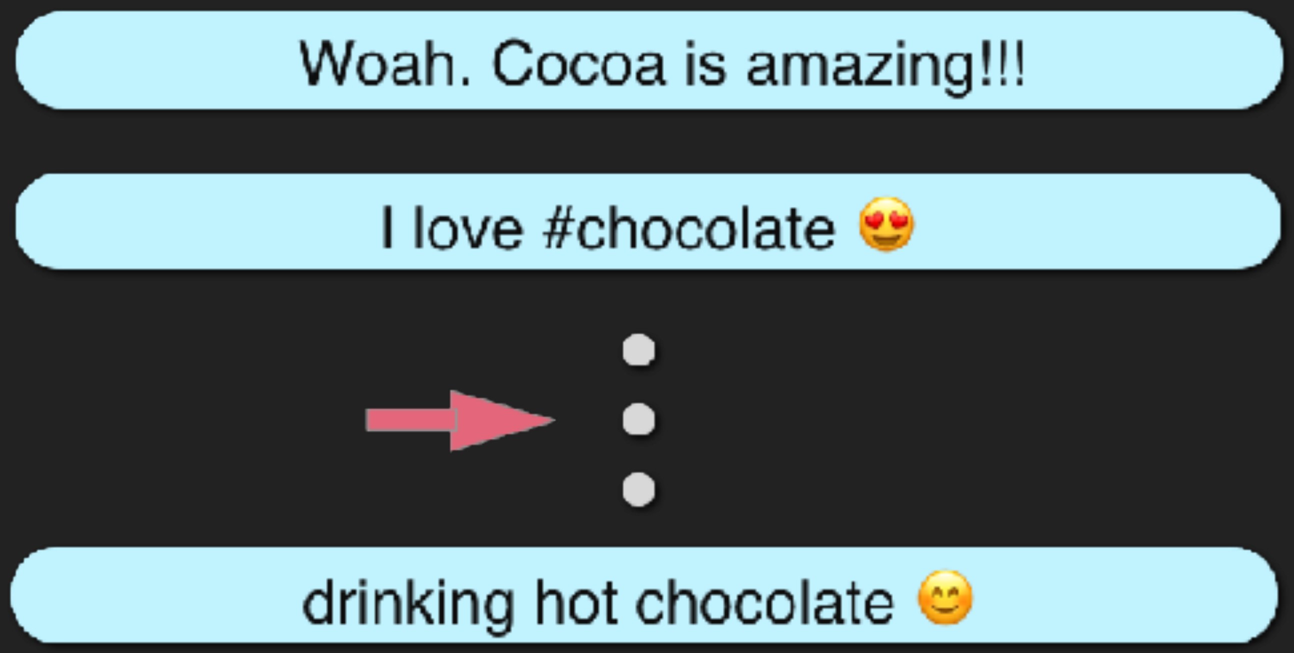
1. Words used
    - punctuation use characteristics
    - feature Frequency
2. Substantive content discussed
3. **Expressed emotionality**
    - **words & emoji & ascii expressions**
4. Engagement level



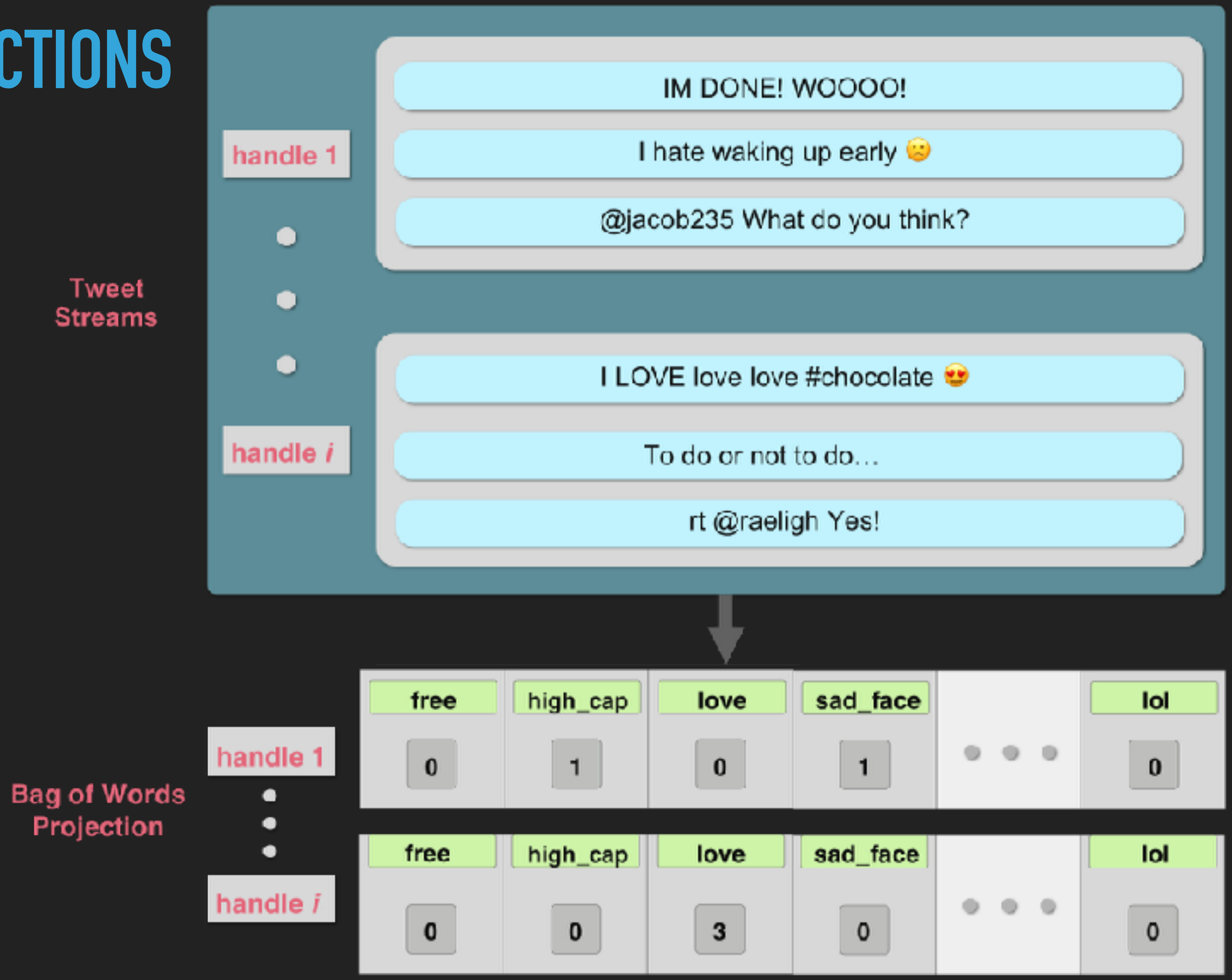
CONTENT THAT MAY MATTER

- What makes an individual unique on Twitter?

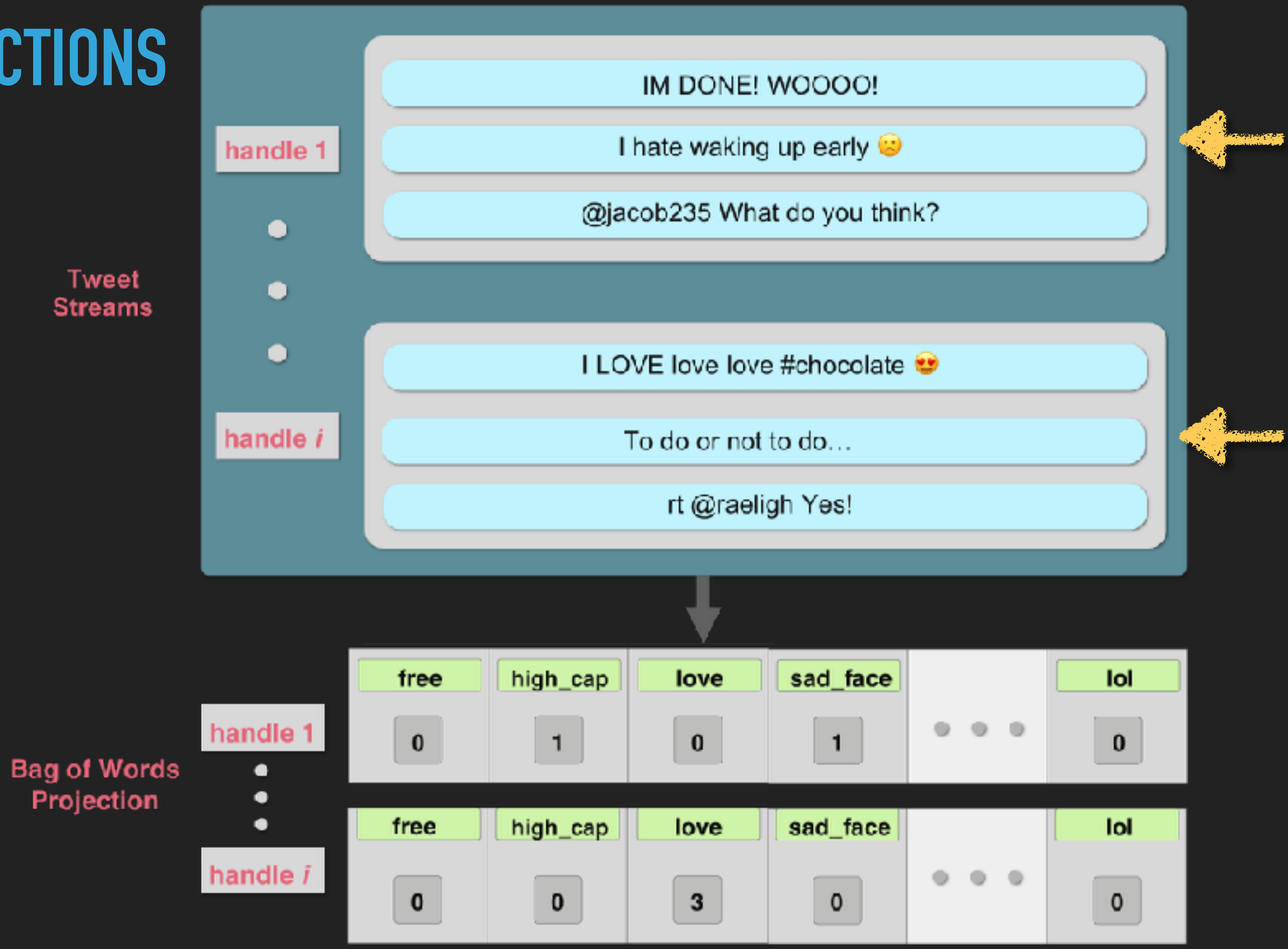
1. Words used
    - punctuation use characteristics
    - feature Frequency
2. Substantive content discussed
3. Expressed emotionality
    - words & emoji
4. **Engagement level**



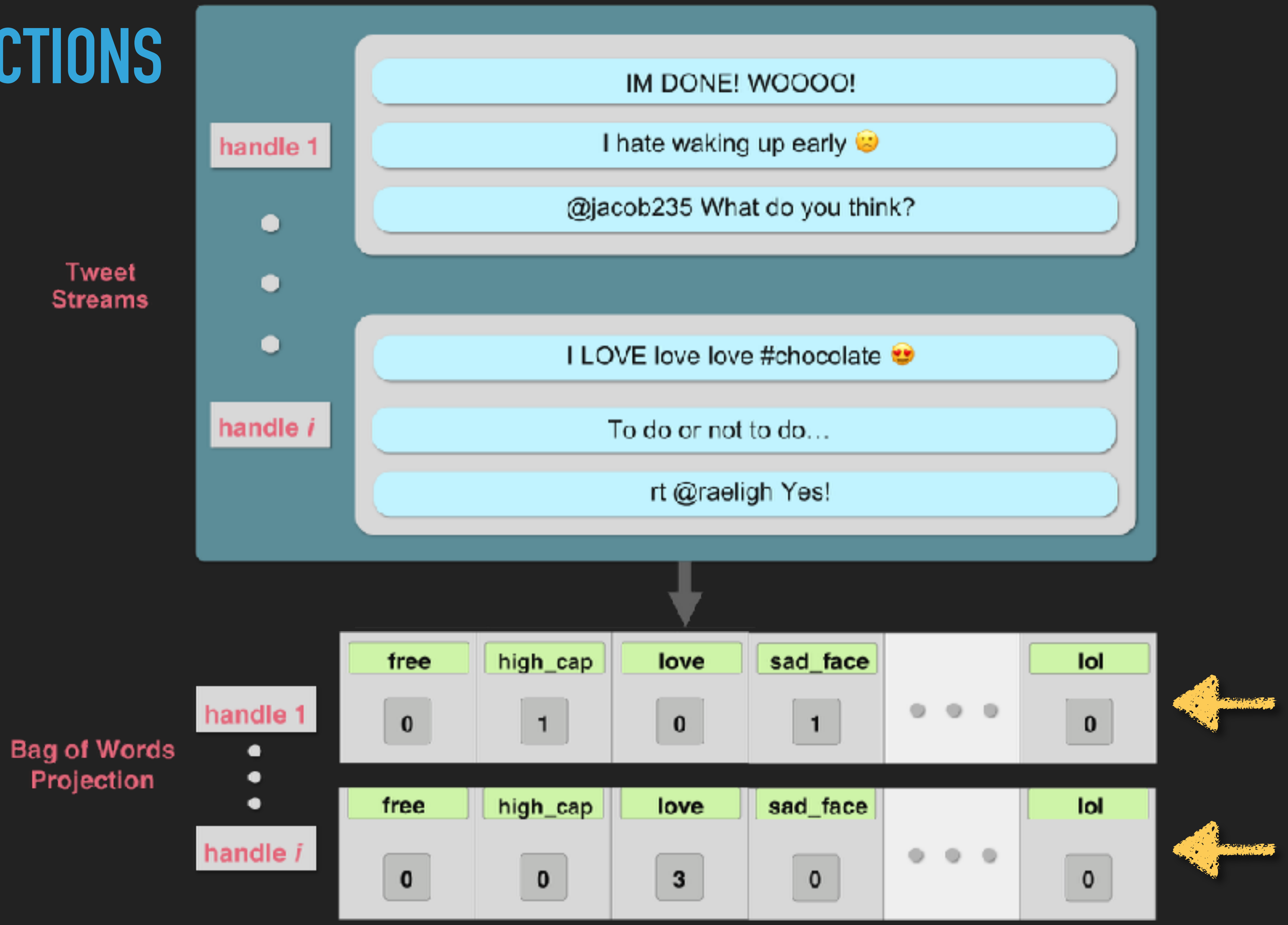
PROJECTIONS



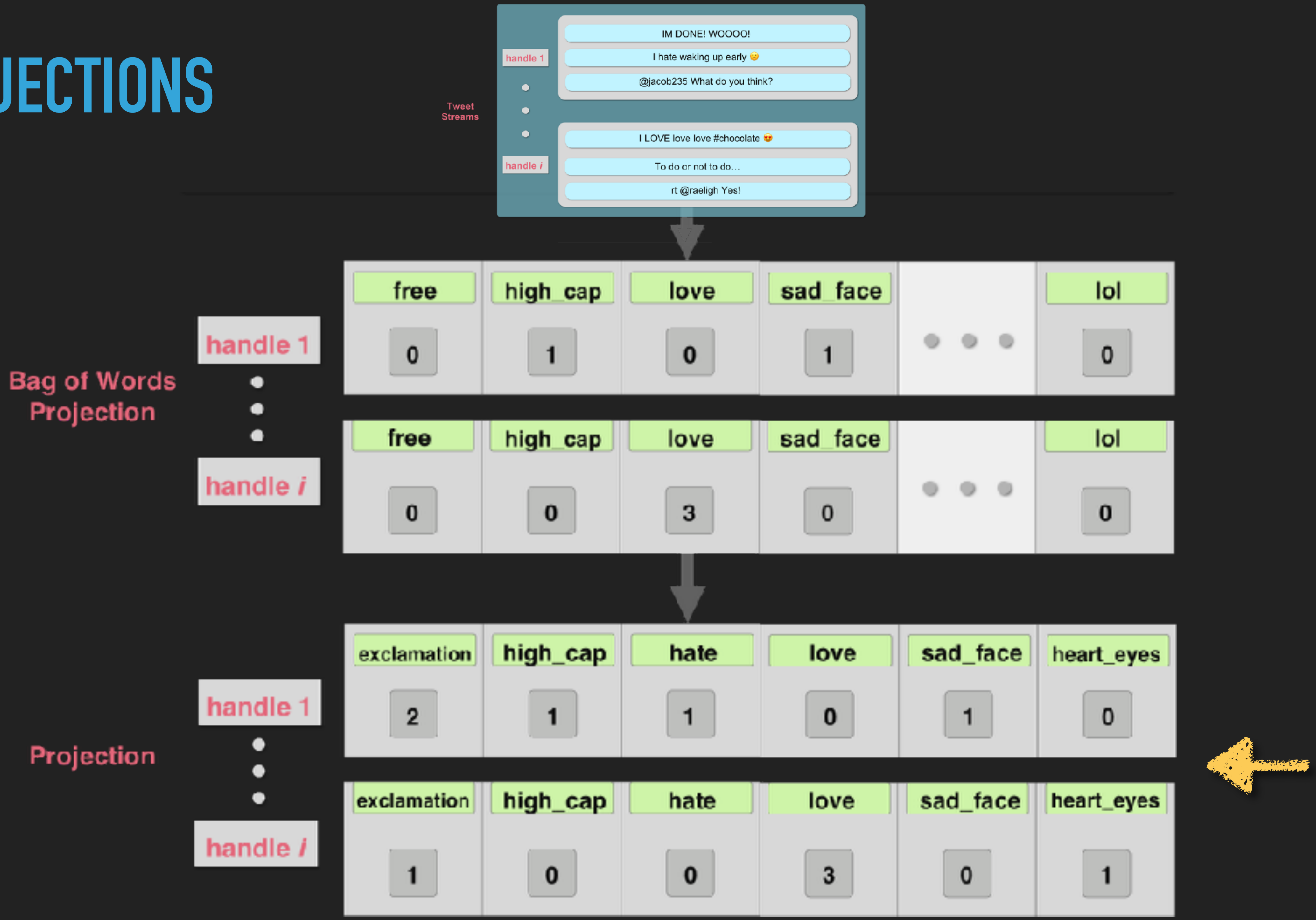
PROJECTIONS



PROJECTIONS

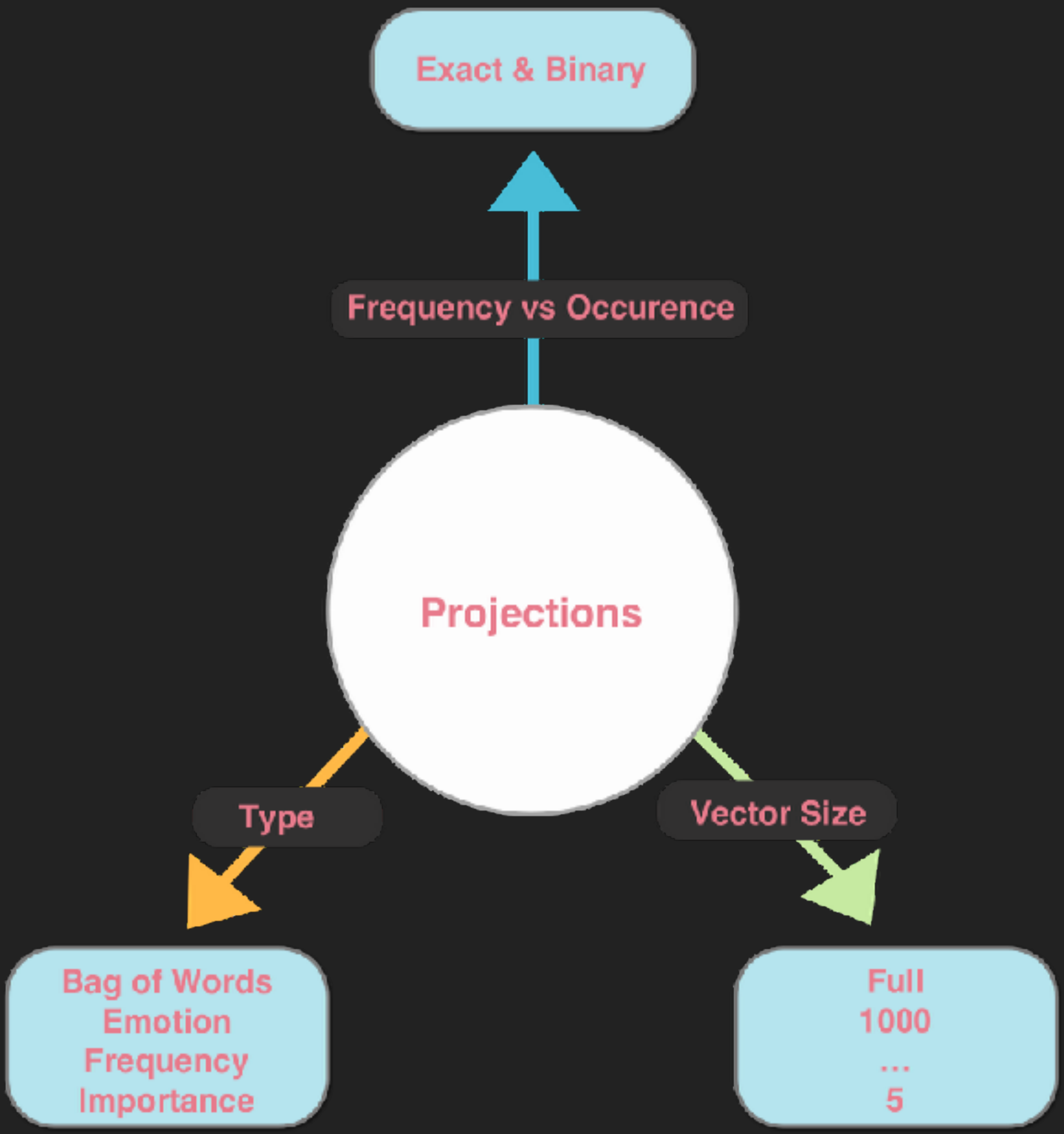


PROJECTIONS



METHODS: PROJECTIONS

- We consider projections across three dimensions:
  - Projection type
  - Binary and exact
  - Projection length



METHODS: PROJECTIONS

- Emotion
- Frequency
- Importance

METHODS: PROJECTION CONSTRUCTION

Twitter Handle
FunSpanishTchr

tweet1
[tweet_num=20]

RT @Hlenggy_H: #ItsTimeIConfess I'm afraid to sit at the front seat of a taxi! 😶😢

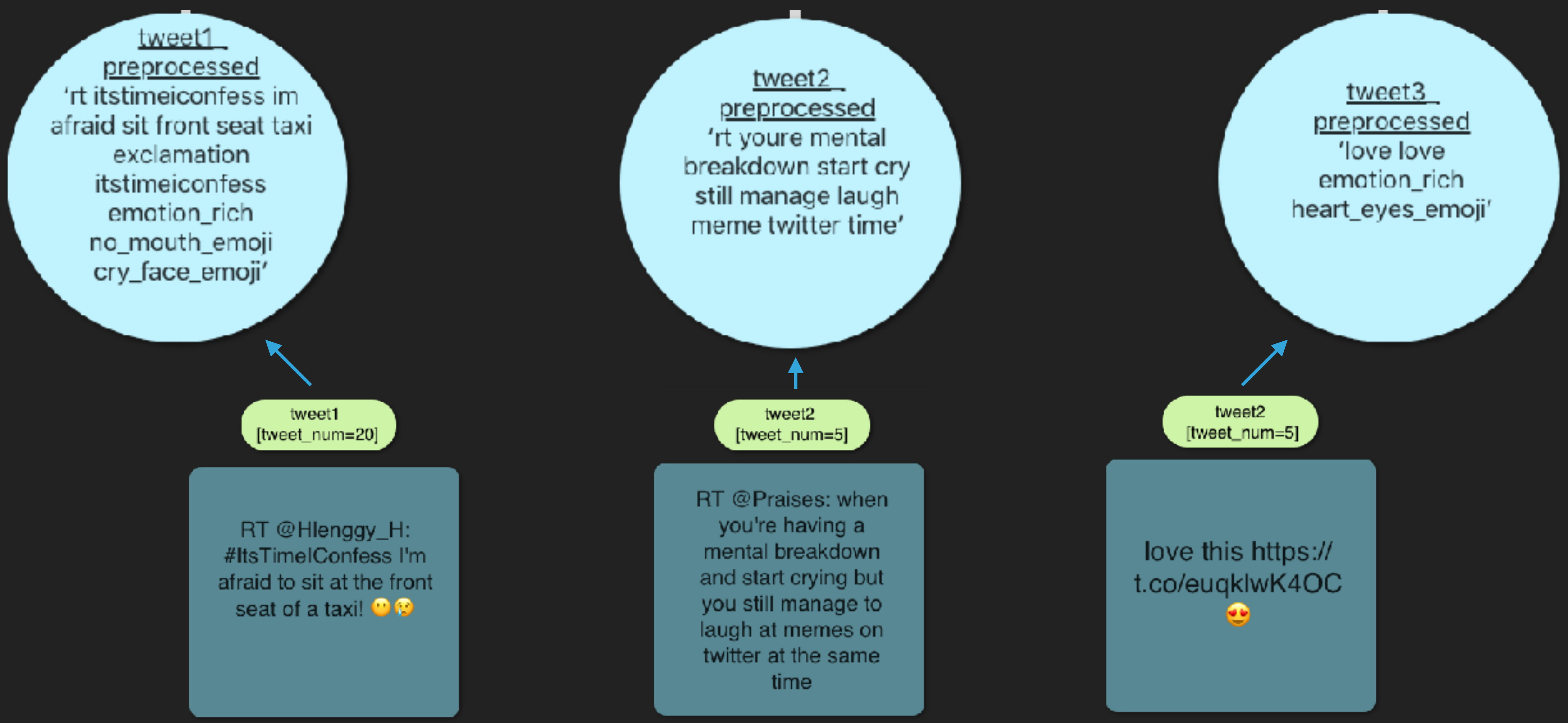
tweet2
[tweet_num=5]

RT @Praises: when you're having a mental breakdown and start crying but you still manage to laugh at memes on twitter at the same time

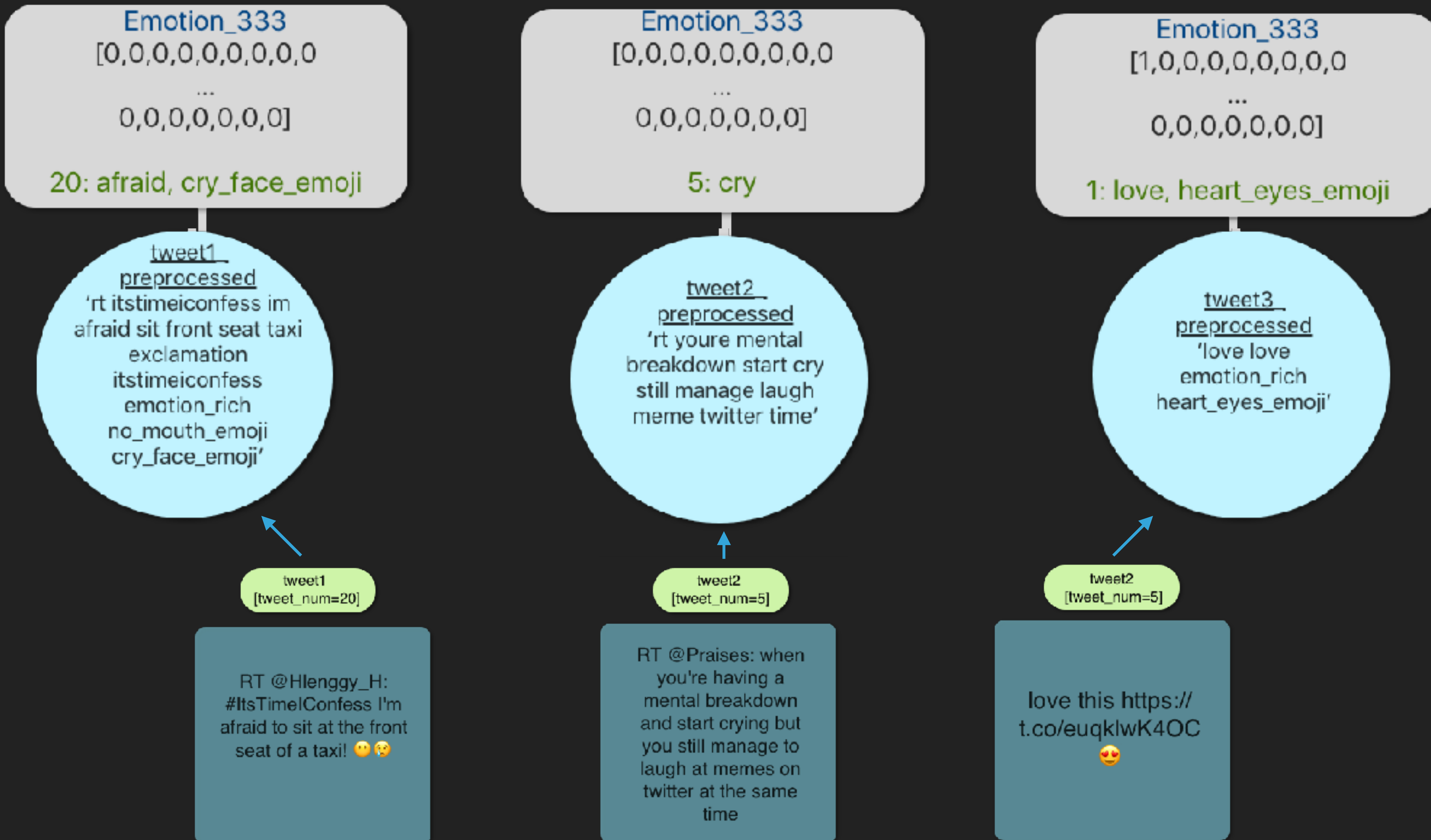
tweet2
[tweet_num=5]

love this https://t.co/euqklwK4OC 😍

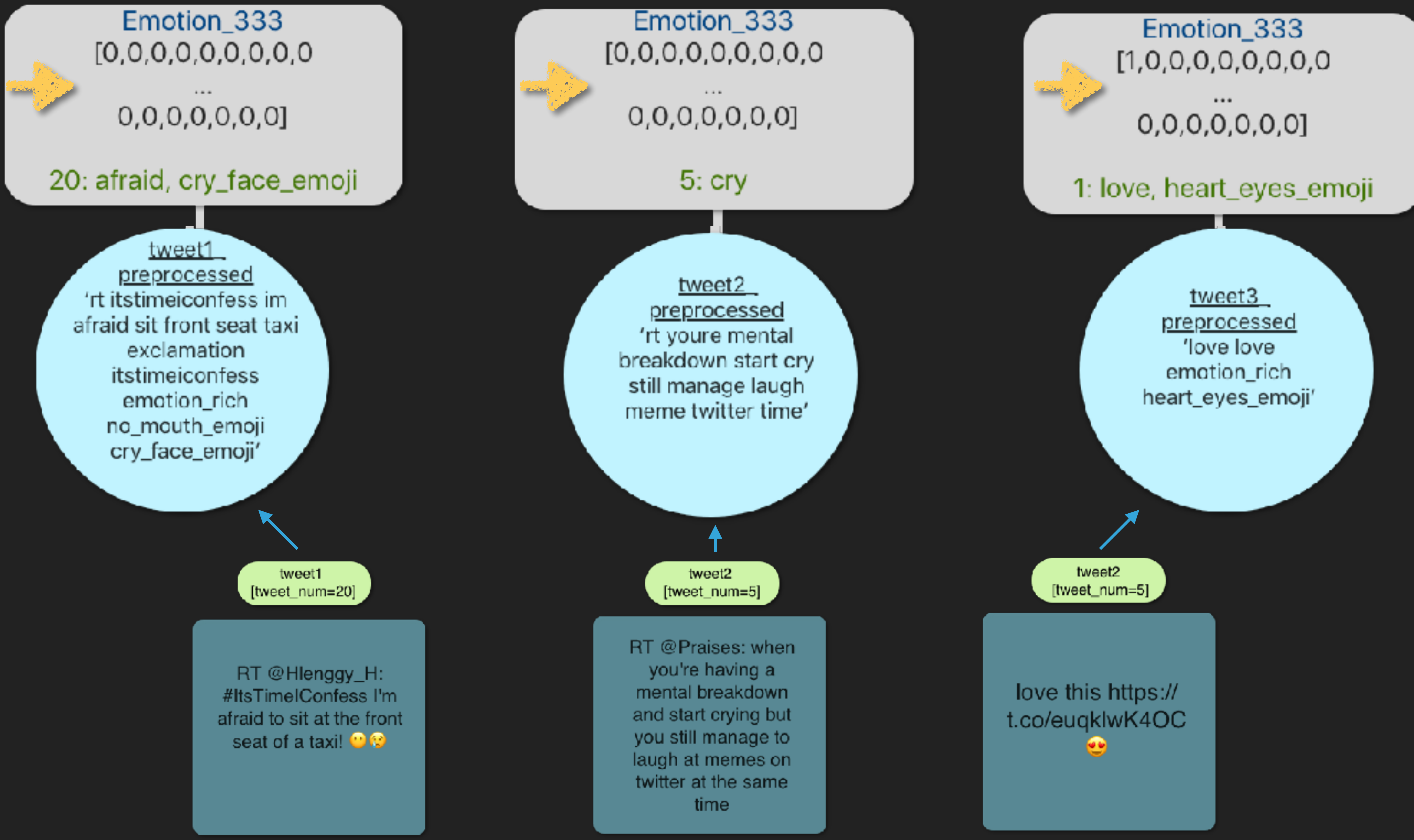
METHODS: EMOTION PROJECTION



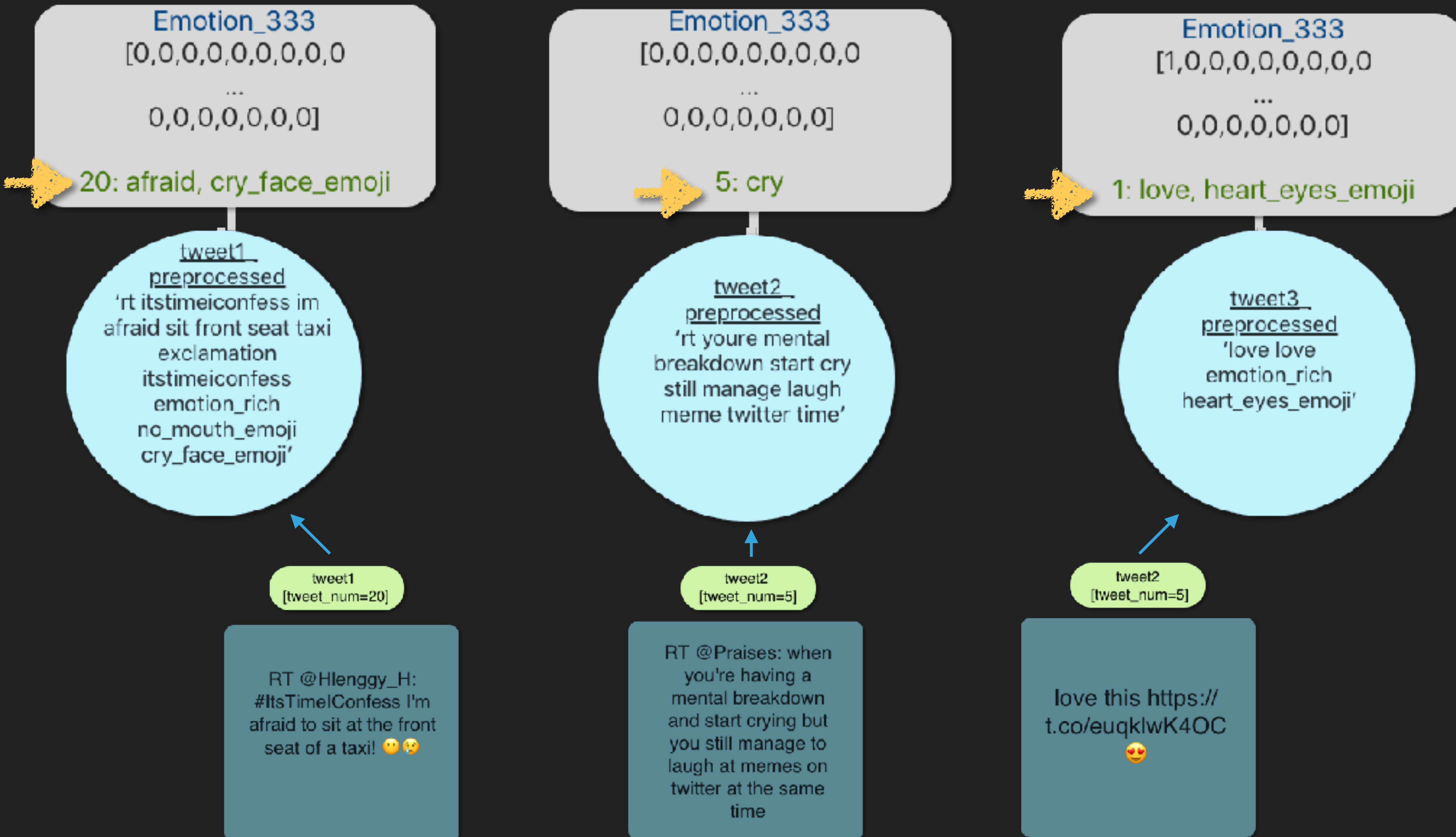
METHODS: EMOTION PROJECTION



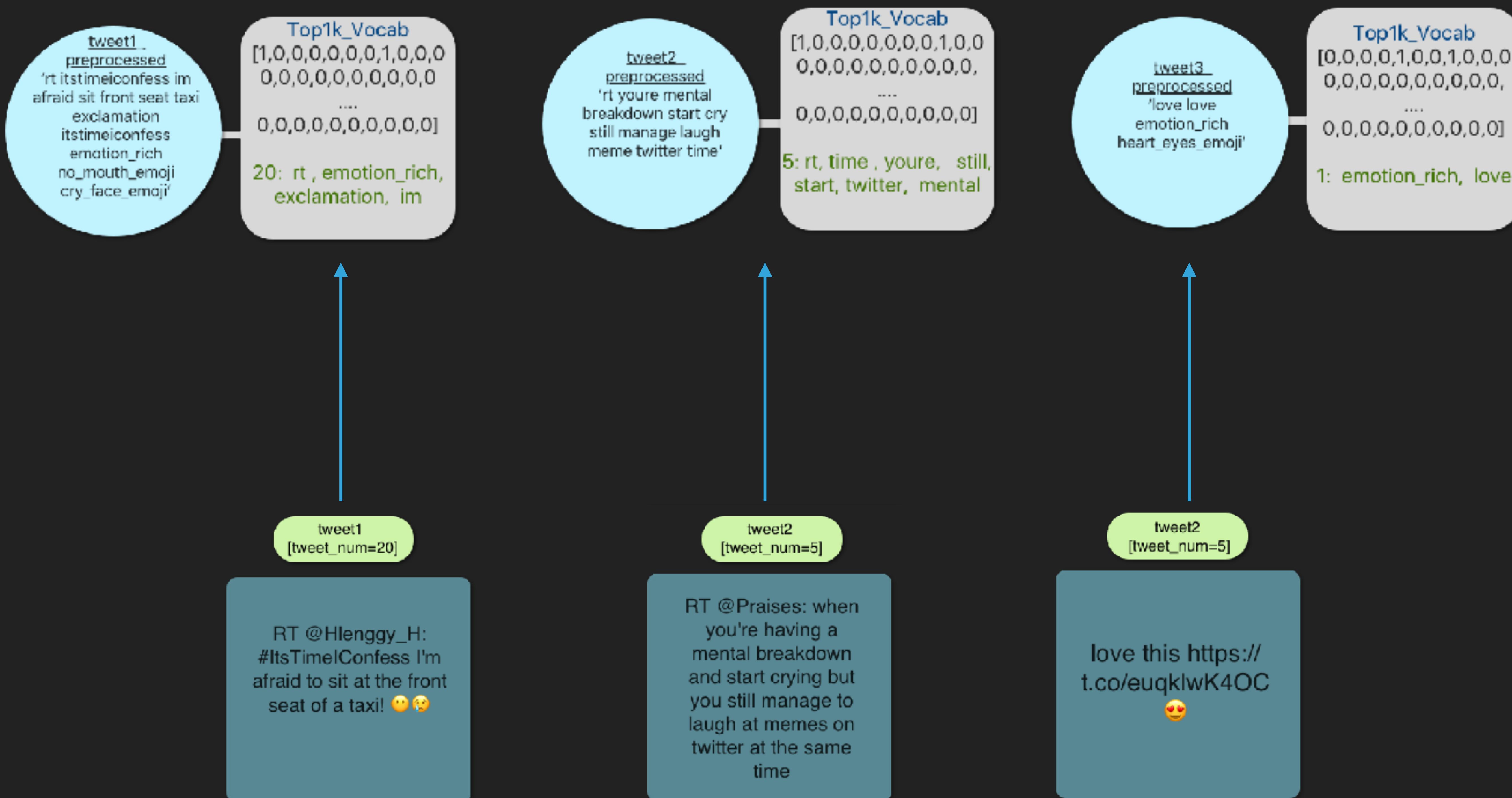
METHODS: EMOTION PROJECTION



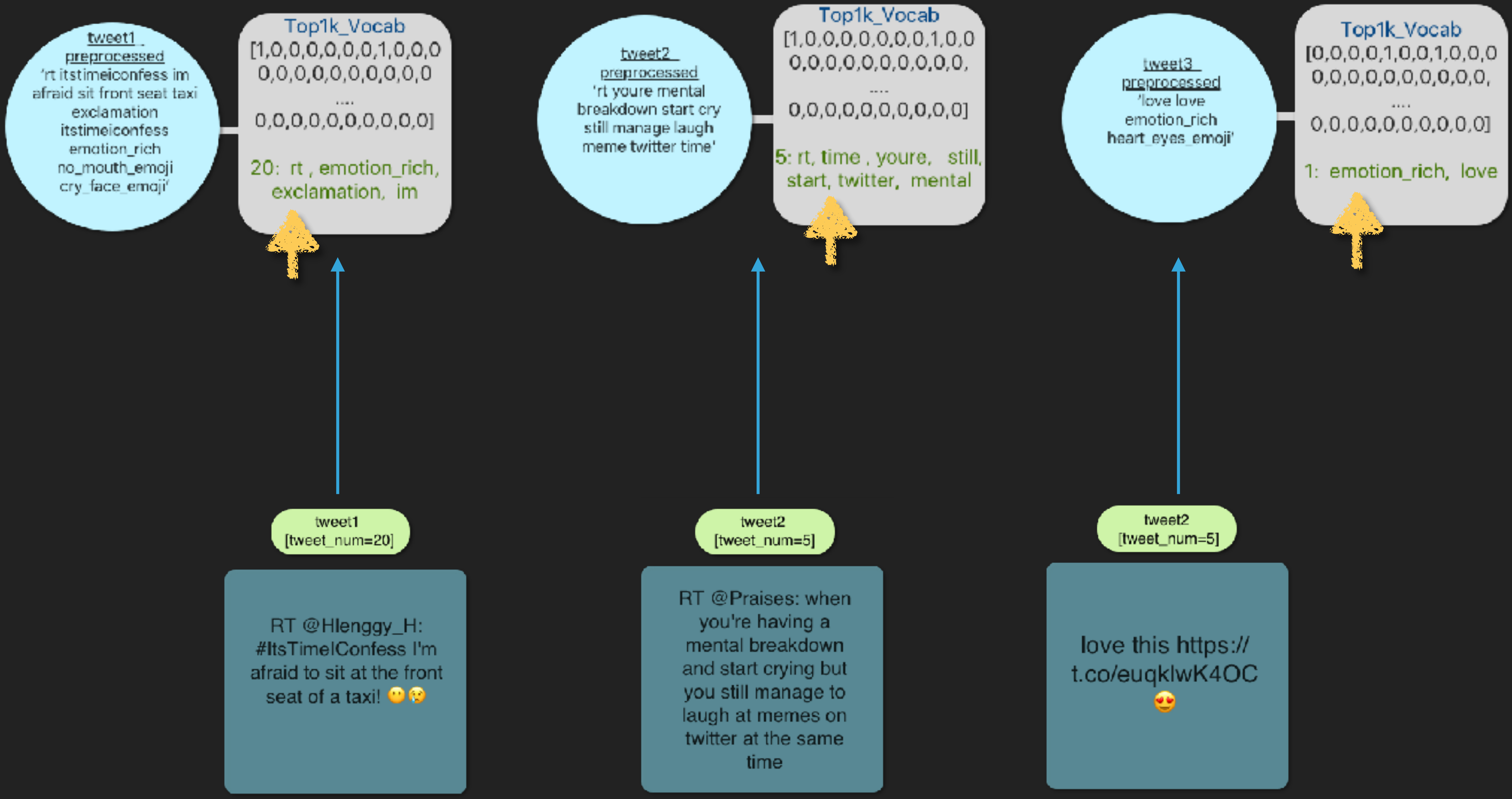
METHODS: EMOTION PROJECTION



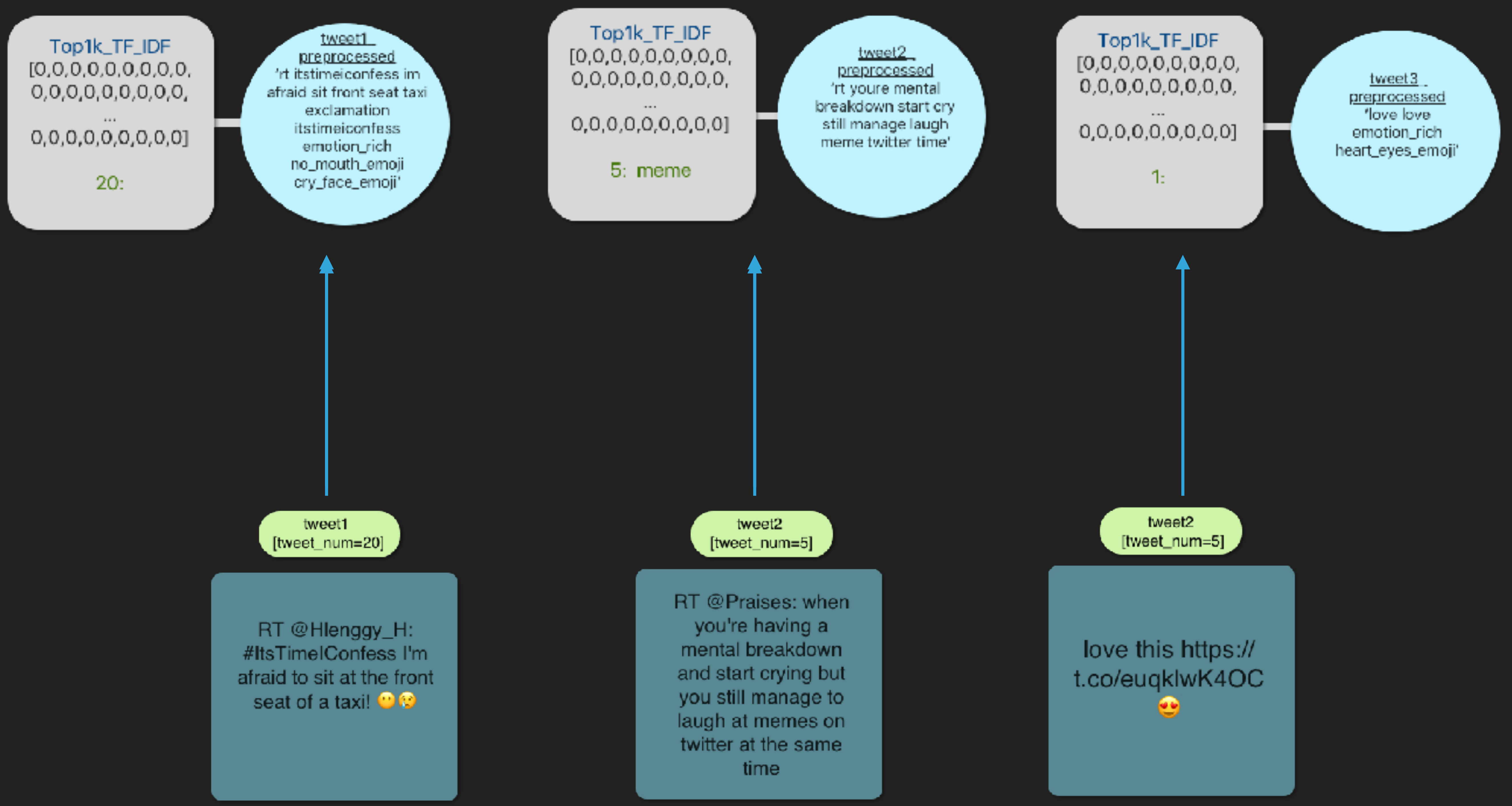
METHODS: FREQUENCY PROJECTION



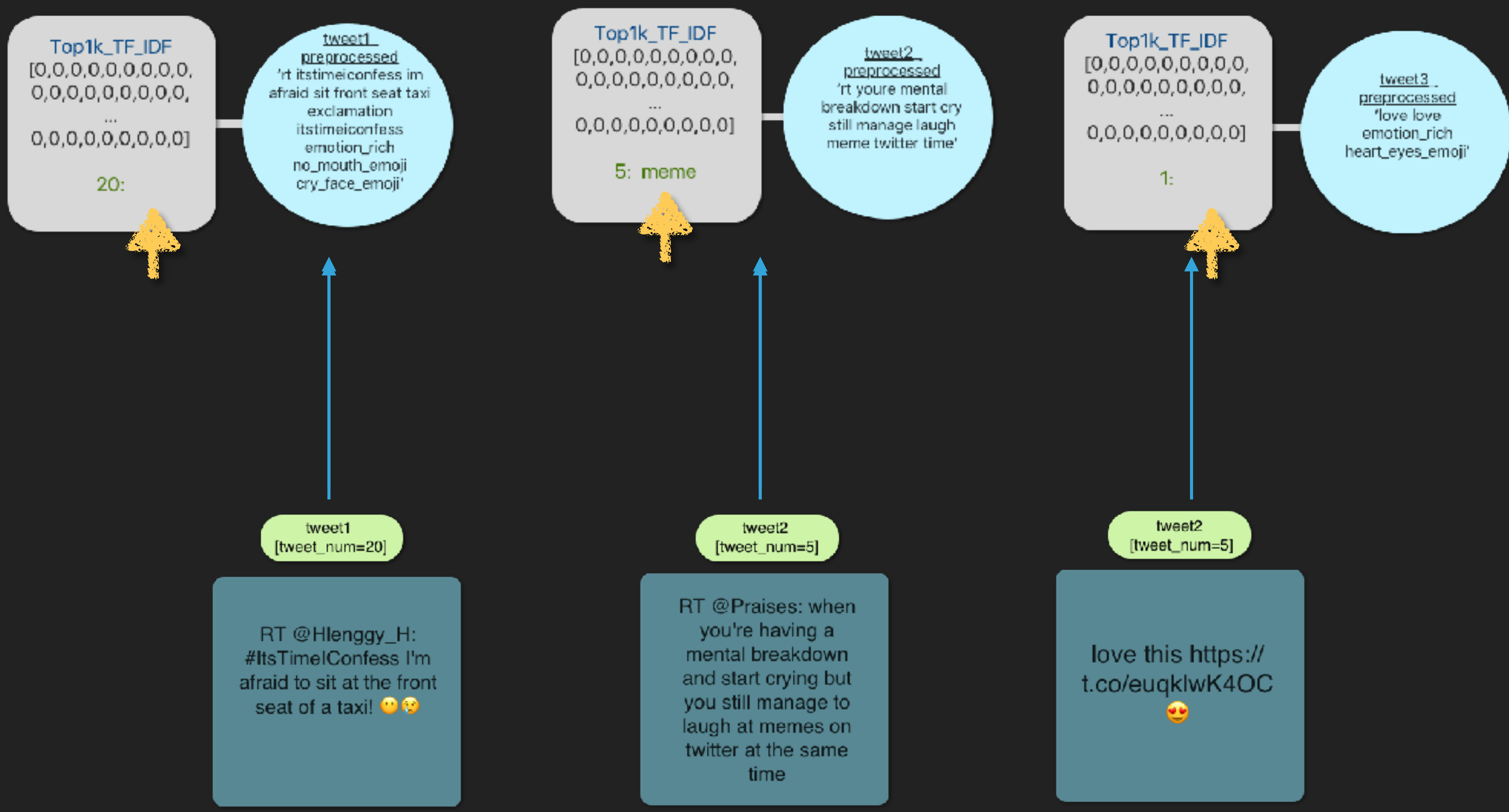
METHODS: FREQUENCY PROJECTION



METHODS: IMPORTANCE PROJECTION



METHODS: IMPORTANCE PROJECTION



METHODS: PRIVACY METRIC

- k-anonymous vector privacy measure
- Measure of privacy for least private individual for a projection

METHODS: PRIVACY METRIC - EXAMPLE

- k-anonymous vector privacy measure
- Measure of privacy for least private individual for a projection

Same

exclamation	high_cap	hate	love	sad_face	heart_eyes
2	1	1	0	1	0

exclamation	high_cap	hate	love	sad_face	heart_eyes
2	1	1	0	1	0

METHODS: PRIVACY METRIC - EXAMPLE

- k-anonymous vector privacy measure
- Measure of privacy for least private individual for a projection

Unique

exclamation	high_cap	hate	love	sad_face	heart_eyes
5	1	1	1	3	1

exclamation	high_cap	hate	love	sad_face	heart_eyes
2	1	1	0	1	0

METHODS: PRIVACY METRIC – EXAMPLE

- k-anonymous vector privacy measure
- Measure of privacy for least private individual for a projection

All-zero

- Often private, but no data utility

METHODS: UTILITY RETENTION METRICS

- Clustering
- Anomaly Detection
- Frequent Item Set Mining

METHODS: UTILITY RETENTION METRICS - CLUSTERING TASK

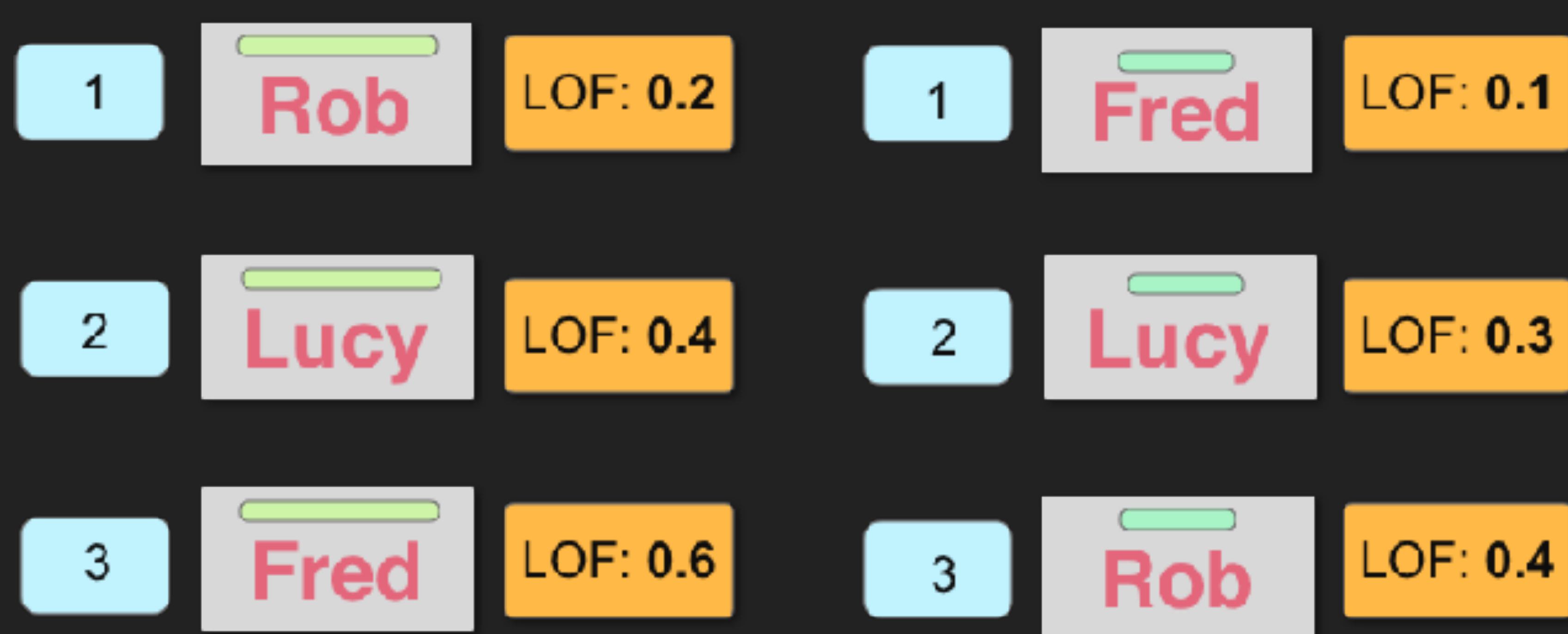
- Utility Retention Metric
  - Proportion of cluster labels shared between projection and baseline (Bag of Words)

METHODS: UTILITY RETENTION METRICS - ANOMALY DETECTION TASK

- Utility Retention Metric
  - Kendall Tau rank correlation coefficient [Kendall, 1938]
    - Measure of ordinal association between two measured quantities

METHODS: UTILITY RETENTION METRICS - ANOMALY DETECTION TASK

- Utility Retention Metric
  - Kendall Tau rank correlation coefficient [Kendall, 1938]
    - Measure of ordinal association between two measured quantities



METHODS: UTILITY RETENTION METRICS – FREQUENT ITEM SET MINING TASK

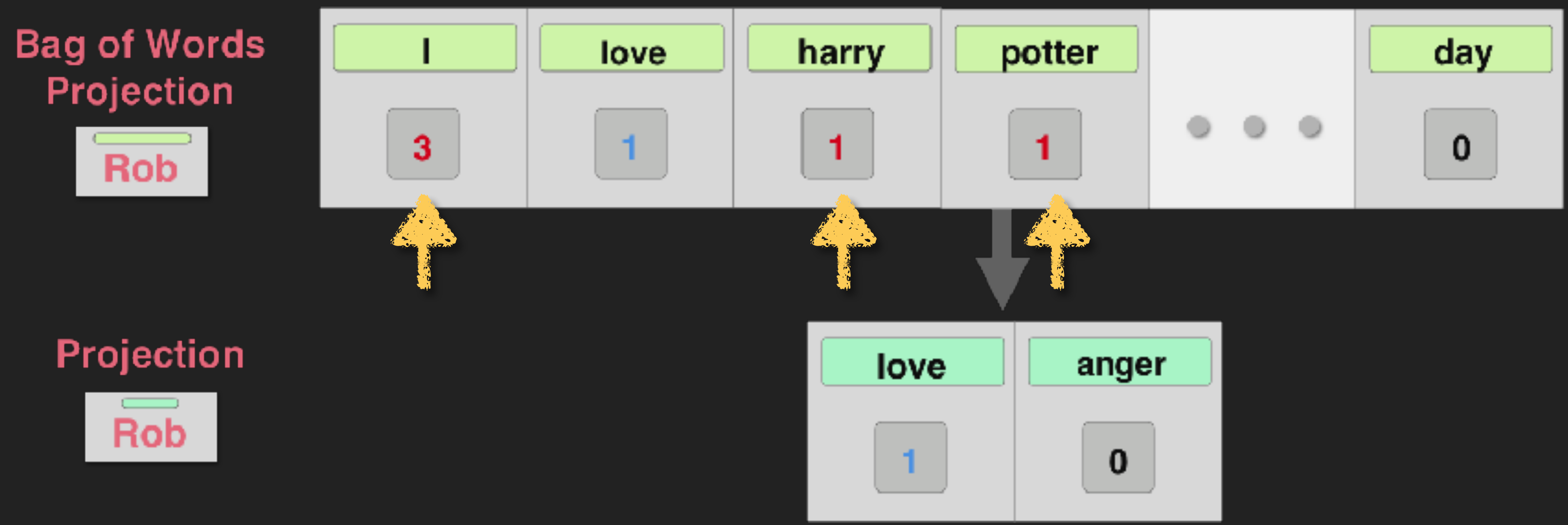
- Utility Retention Metric
  - Average of (i) proportion of items sets from projection also in Bag of Words item sets [precision] and (ii) proportion of Bag of words item sets in item sets of projection [recall]

METHODS: DISTORTION METRICS

- Proportion of features lost for each handle
- Do not count zero-valued features
  - No distinguishing features lost

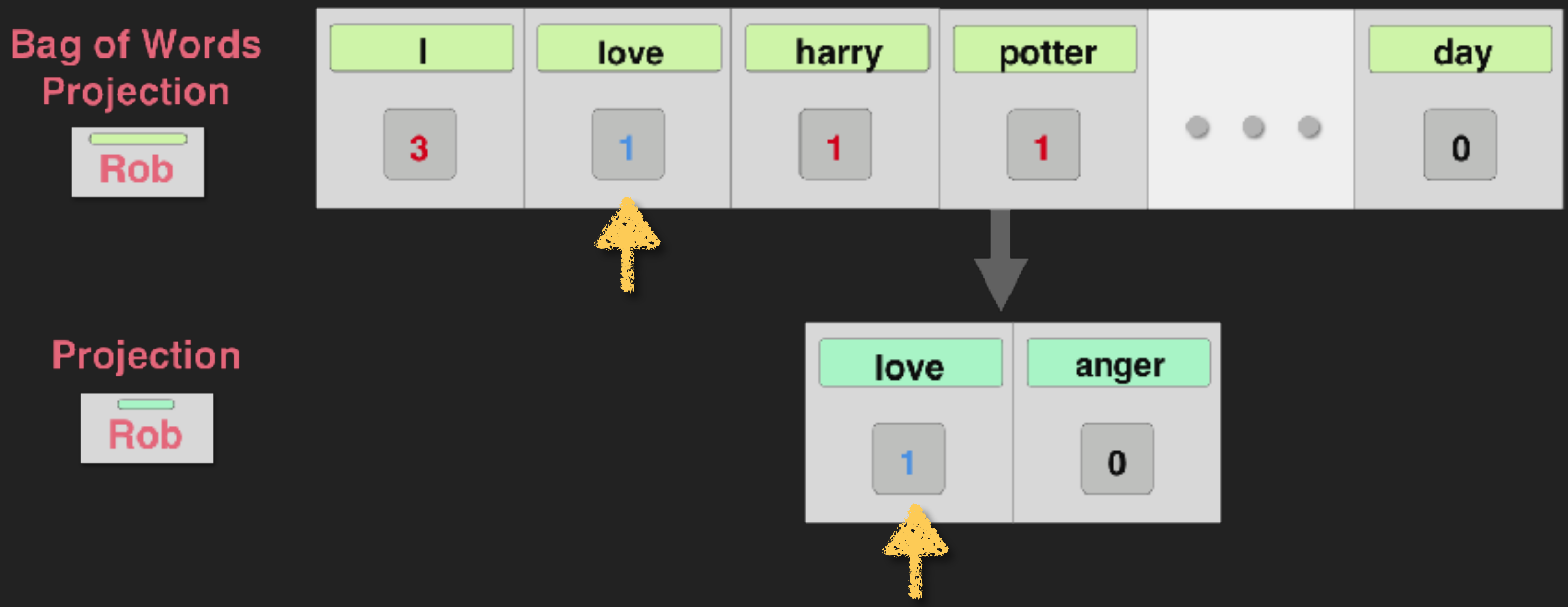
METHODS: DISTORTION METRICS

- Proportion of features lost for each handle
- Do not count zero-valued features
  - No distinguishing features lost



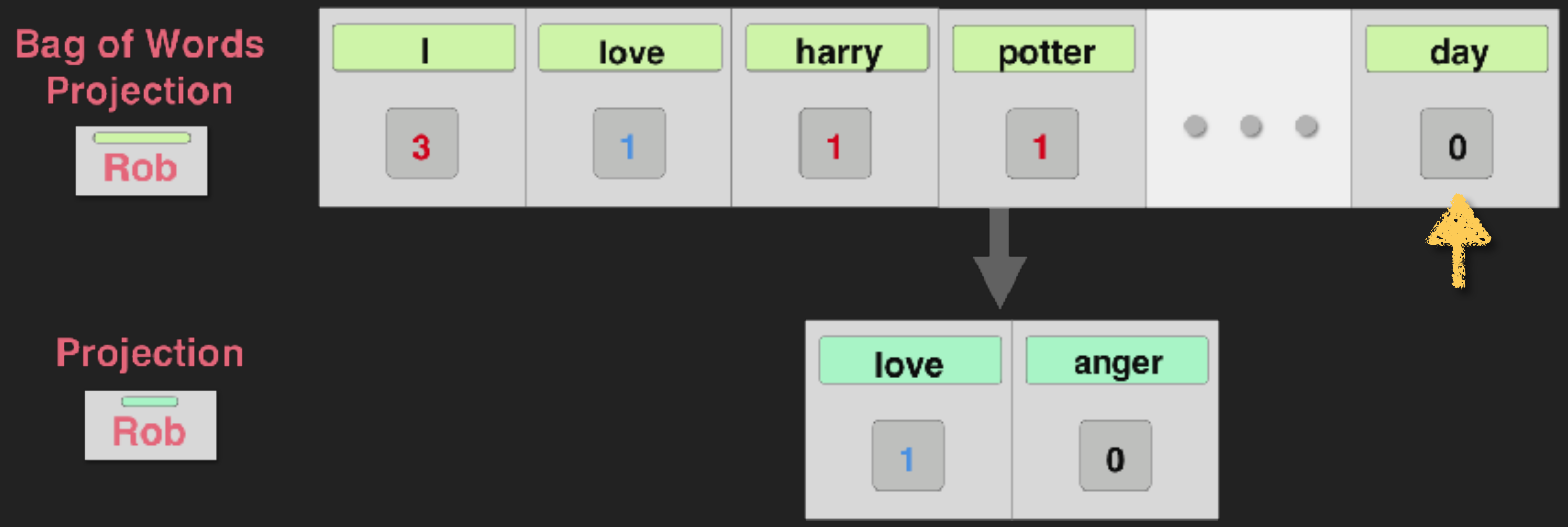
METHODS: DISTORTION METRICS

- Proportion of features lost for each handle
- Do not count zero-valued features
  - No distinguishing features lost



METHODS: DISTORTION METRICS

- Proportion of features lost for each handle
- Do not count zero-valued features
  - No distinguishing features lost



ENGAGEMENT LEVELS

- High
  - 2000 < number of tweets
- Moderate
  - 200 < number of tweets <= 2000
- Low
  - 20 <= number of tweets <= 200

DATA DESCRIPTION

TWEET COLLECTION

- 110,891 Twitter handles
  - 10 main handles, 9 corresponding to parenting and 1 to medical sites
  - ~10,000 handles following each of these main handles
- Active Collection
  - mid 2016 - end of 2017
  - Most recent 3500 tweets for each handle collected at commencement
    - Tweets may be missed in the case of high engagement

DATA SET

- Goal: ~1000 handles for each engagement level
- 5626 handles of those obtained in tweet collection
- 4 million tweets

EXPERIMENTS & RESULTS

PREPROCESSING OF DATA

- Lemmatization
  - to reduce noise
- Feature Injection
- Cleaning

PREPROCESSING OF DATA - FEATURE INJECTION

- Emoji and ascii expressions
  - Normalization
- Punctuation
  - Exclamation, extreme exclamation
- Emotional richness
- High capitalization

PREPROCESSING OF DATA - CLEANING

- Removal:
  - Hyper-links
  - Twitter handle references
  - All-number & all-punctuation words
  - Stop words
- Fix contractions
- Case lowering

COMPUTATION OF DATA PROJECTIONS - BAG OF WORDS

- 767,937 features
  - 256,135 features are hapax legomenon
- 47 all-zero vector representations of handles' tweet streams
  - Few tweets
  - No features left after removal of handle references, urls, & stop words
- Baseline projection

COMPUTATION OF DATA PROJECTIONS - EMOTION

- 333 features for the following emotional categories
  - anger_disgust
  - fear
  - joy_love
  - sadness
  - surprise
- Features include emoji & ascii expressions

COMPUTATION OF DATA PROJECTIONS - EMOTION

- Features selected by
  - Human selection of fundamental synonyms for the five emotional categories
  - High occurrence of a word in tweets classified as expressing one of the five emotional categories
    - Used Emotional Lexicon Hit Method
      - Average precision 0.551, average recall 0.662
      - Tried Naïve Bayes and SVMs as well, but Emotional Lexicon Hit Method had best performance on average
    - 600 labeled tweets

COMPUTATION OF DATA PROJECTIONS – EMOTION

- Example synonyms:

anger_disgust	fear	surprise	sadness	joy_love
anger	fear	surprise	sadness	joy
disgust	fright	amazed	sad	love
sickening	frightened	astonished	dispirited	encouraged
outraged	panic	astounded	heartbroken	smiling
ew	worry	dumbfounded	morose	laugh
angry_pout_face	fearful_face	open_mouth_face	cry_face	grin_face
stream_from_nose_face	worried_face	surprise_face	frown_face	heart_emoji
...	...	...	...	...

COMPUTATION OF DATA PROJECTIONS - EMOTION

- Example synonyms:

anger_disgust	fear	surprise	sadness	joy_love
anger	fear	surprise	sadness	joy
disgust	fright	amazed	sad	love
sickening	frightened	astonished	dispirited	encouraged
outraged	panic	astounded	heartbroken	smiling
ew	worry	dumbfounded	morose	laugh
angry_pout_face	fearful_face	open_mouth_face	cry_face	grin_face
stream_from_nose_face	worried_face	surprise_face	frown_face	heart_emoji
...	...	...	...	...

COMPUTATION OF DATA PROJECTIONS – FREQUENCY

- Feature space:
  - 1,000 most frequently occurring words across set of tweet streams
- Potentially noisy projection

COMPUTATION OF DATA PROJECTIONS - IMPORTANCE

- Feature space:
  - 1,000 most important words across set of tweet streams
  - Determined by tf-idf measure [Jones, 1972]
    - Computed for each word, for each document (handle)
  - 94% of features are hashtag words
- Projection for substantive content (using hashtags as a proxy for topics)

COMPUTATION OF DATA PROJECTIONS - FEATURE SPACE REDUCTION

- Reduce feature space of projections to evaluate impact on privacy & utility
- Frequency & Importance projections
  - Use top *n* features to represent tweet streams
- Emotion Projection
  - Represent tweet streams according to our five emotional categories
  - Value of emotional category feature corresponds to number of observed synonym occurrences for that emotional category

COMPUTATION OF UTILITY - K MEANS

- To determine value for k, conducted Sensitivity Analysis
- Evaluated for Bag of Words
  - Silhouette score [Rousseeuw, 1987]
  - Subjective similarity of vectors sharing same cluster
- 8 clusters for exact Bag of Words projection
- 6 clusters for binary Bag of Words projection

COMPUTATION OF UTILITY - LOCAL OUTLIER FACTOR

- Sensitivity Analysis: range of number of neighbors in a neighborhood values evaluated for Bag of Words
  - Silhouette score
  - Subjective similarity of vectors sharing same cluster
  - Number of tweet streams labeled as outliers
- Chose number of neighbors in a neighborhood value of 20

COMPUTATION OF UTILITY - FREQUENT ITEM SETS

- Sensitivity Analysis: conducted for a range of minimum support values, for all projections
  - number of sets of size larger than 2
  - number of sets = ~100
- Minimum support of 1% for Bag of Words
  - 113 item sets

COMPUTATION OF UTILITY – FREQUENT ITEM SETS

- Emotion
  - Minimum support of .01%
  - 44 frequent item sets
- Importance
  - Minimum support of .02%
  - 92 frequent item sets
- Frequency
  - Minimum support of 1%
  - 138 frequent item sets

COMPUTATION OF DISTORTION

- Distortion rates for projections
  - Fraction of non-zero features lost each handle vector, when represented by projection

Emotion 333	Top 1k Frequency	Top 1k Importance	Emotion 5	Frequency 5	Importance 5
0.9993	0.6054	0.9996	0.9995	0.9597	0.9997

COMPUTATION OF DISTORTION

- Distortion rates for projections
  - Fraction of non-zero features lost each handle vector, when represented by projection

Emotion 333	Top 1k Frequency	Top 1k Importance	Emotion 5	Frequency 5	Importance 5
0.9993	0.6054	0.9996	0.9995	0.9597	0.9997

- Importance and Emotion projections exhibit high distortion

COMPUTATION OF DISTORTION

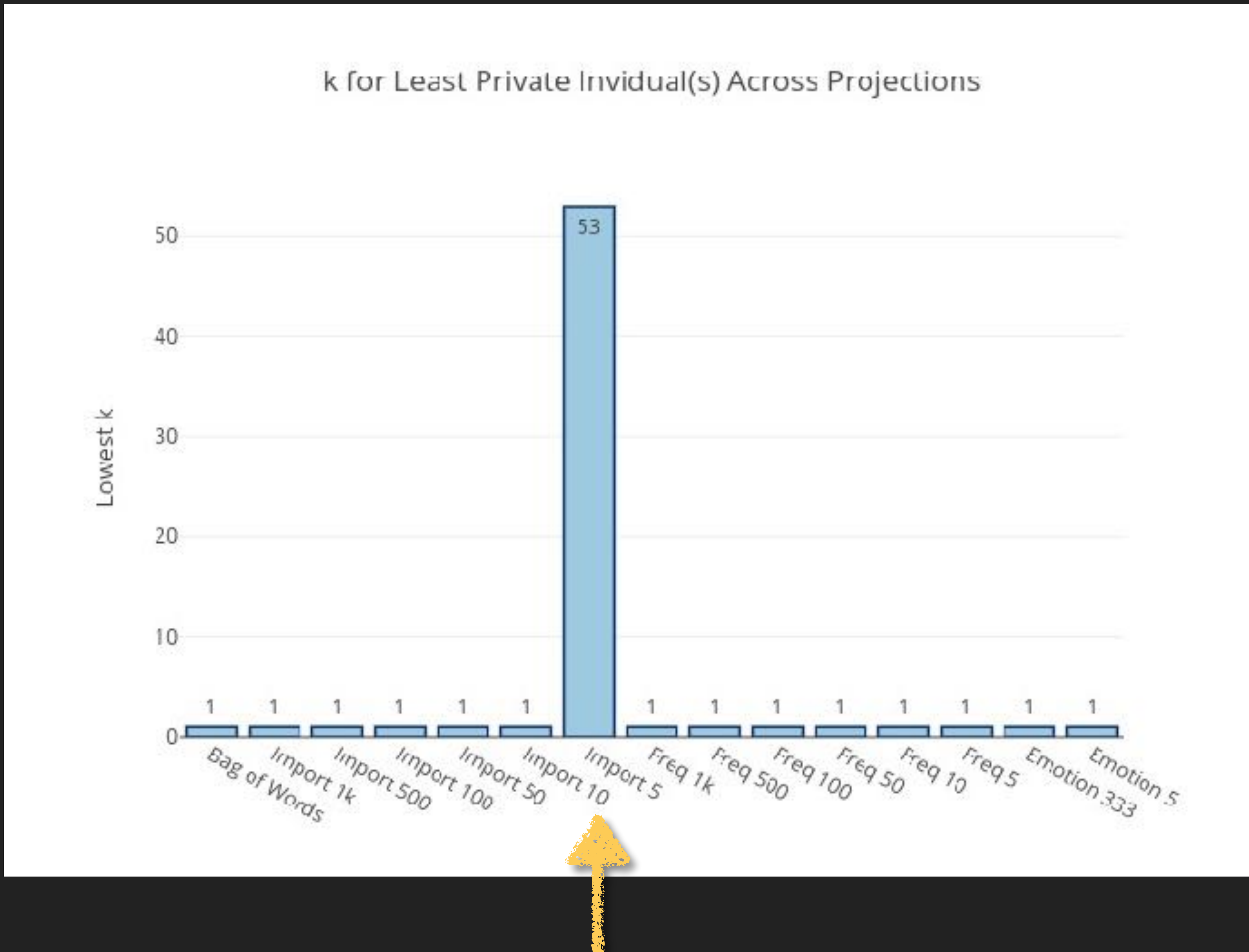
- Distortion rates for projections
  - Fraction of non-zero features lost each handle vector, when represented by projection

Emotion 333	Top 1k Frequency	Top 1k Importance	Emotion 5	Frequency 5	Importance 5
0.9993	0.6054	0.9996	0.9995	0.9597	0.9997

- Comparatively low distortion for Frequency projection

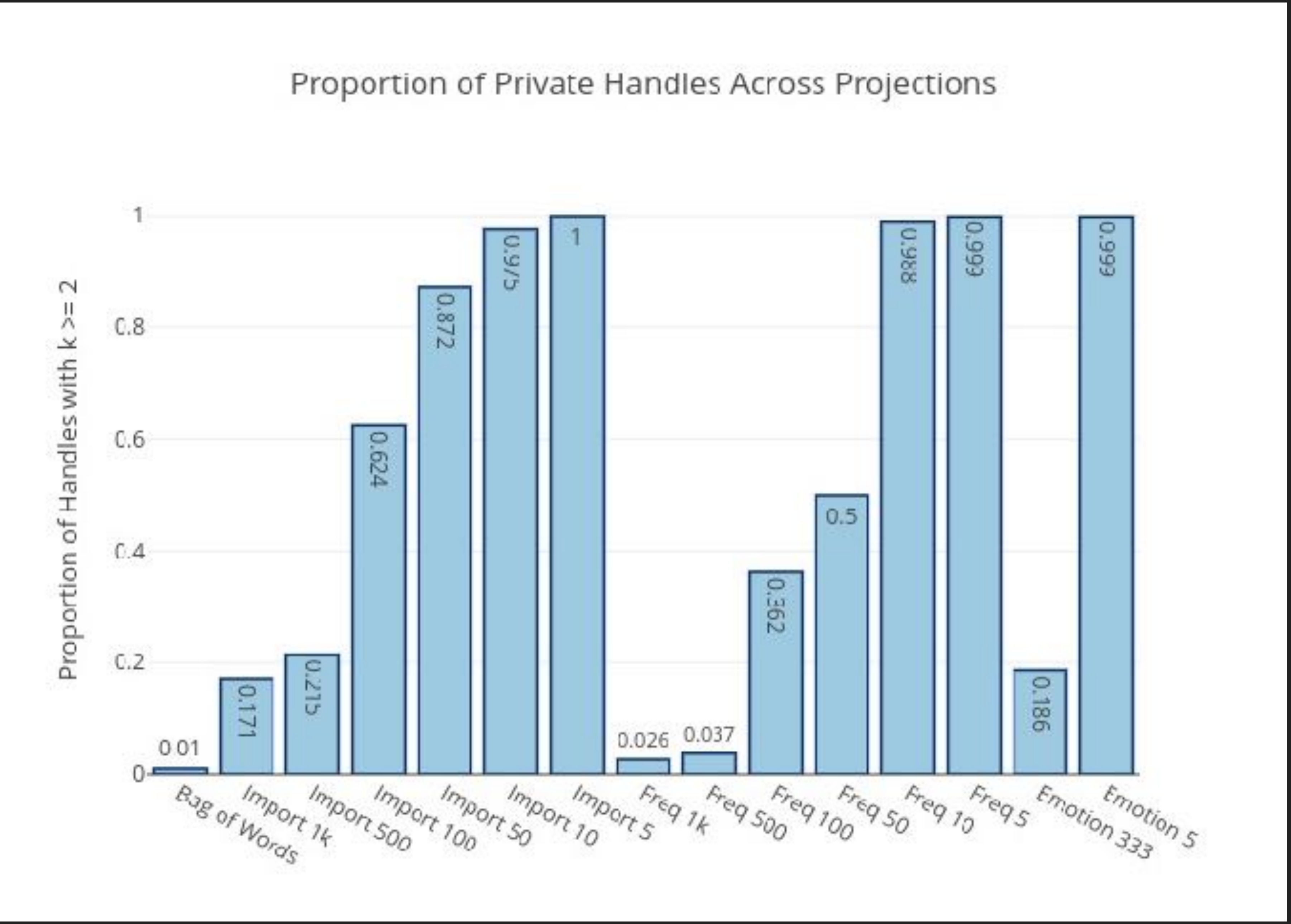
PRIVACY – K ANONYMOUS VECTOR METRIC

- Binary
- Lowest k for each projection
- Only Importance projection achieves privacy for all users



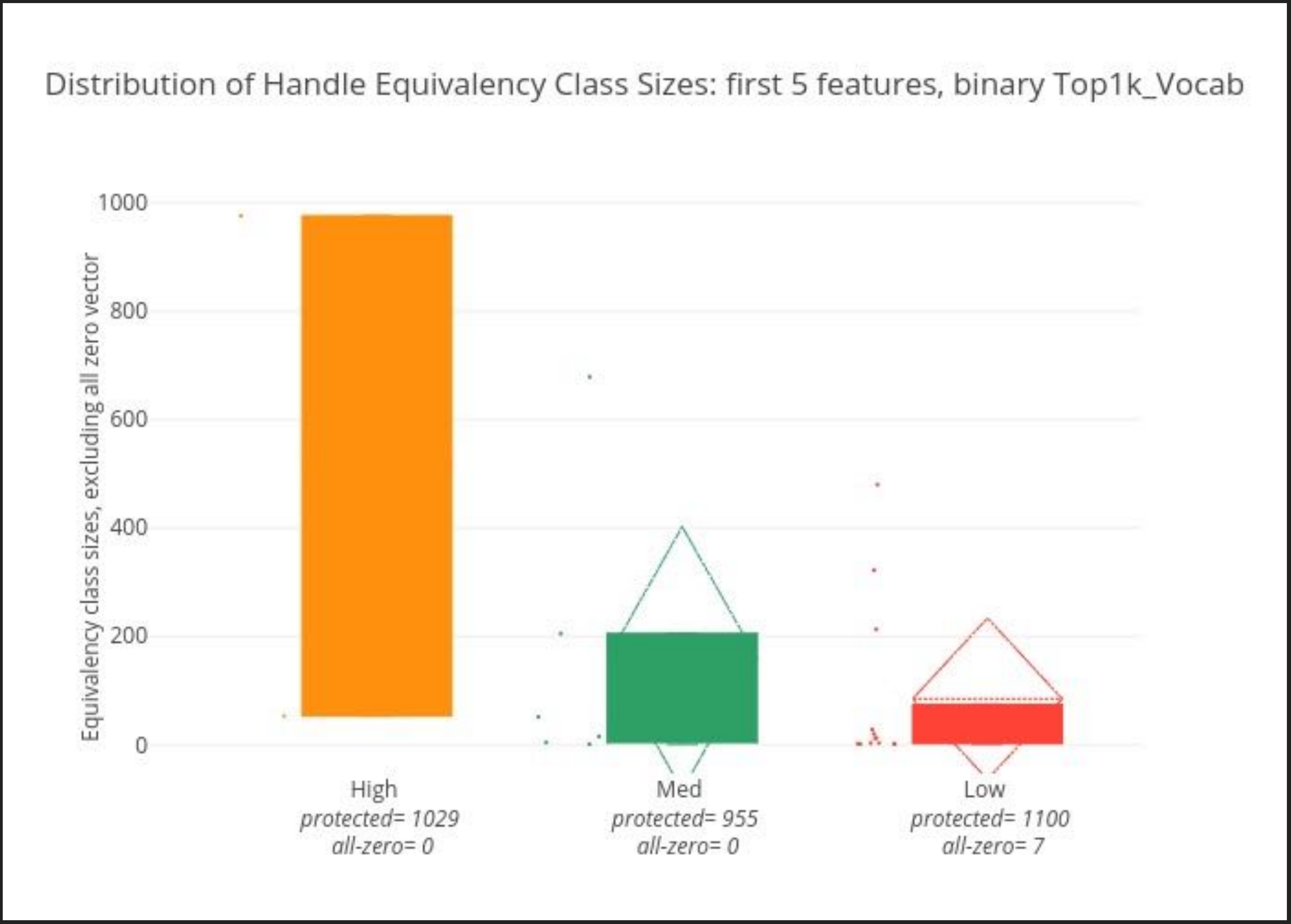
PRIVACY – K ANONYMOUS VECTOR METRIC

- Proportion of handles with privacy (k>=2) for each projection
- Low privacy unless small feature spaces used



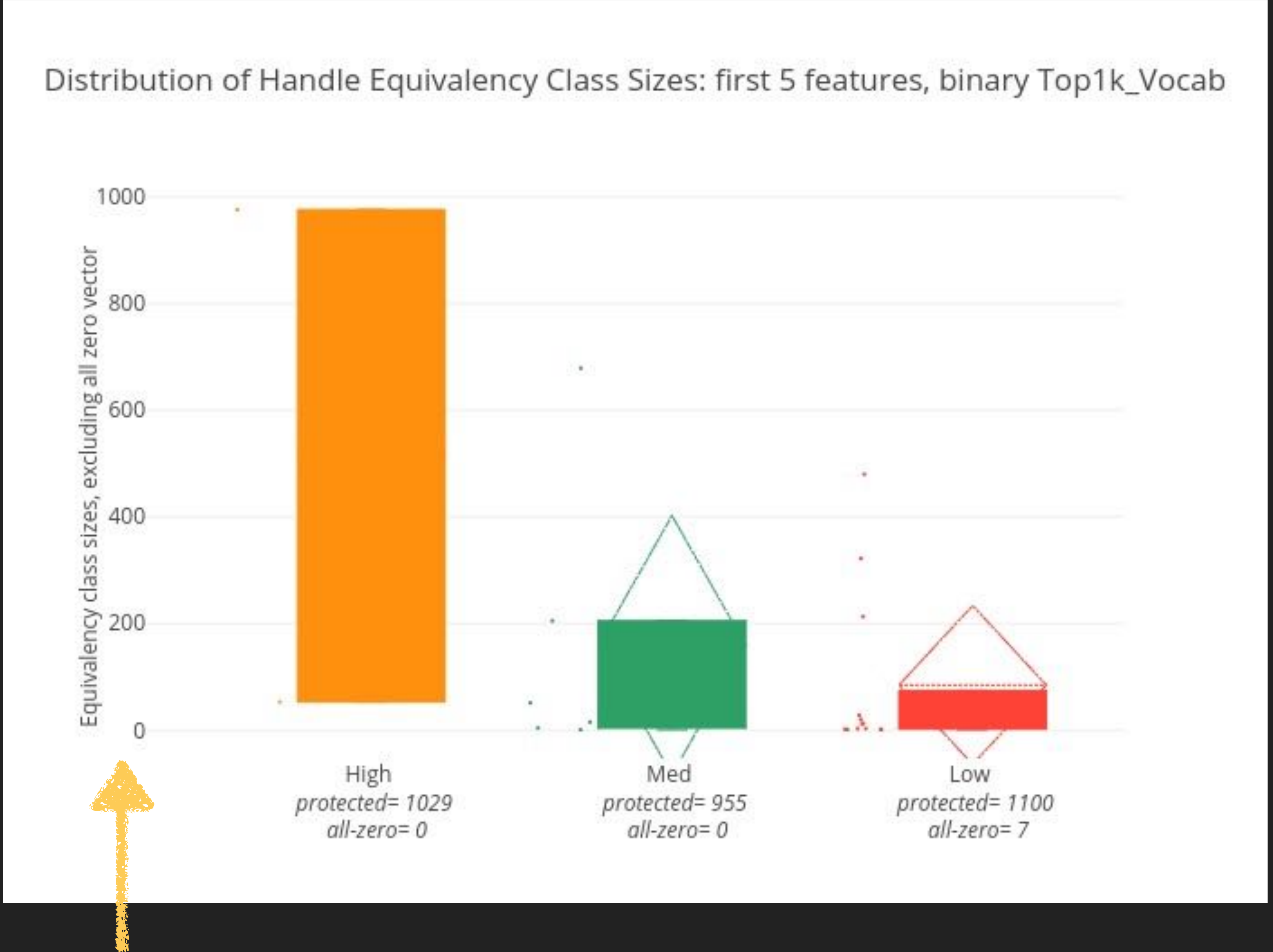
PRIVACY – ACROSS ENGAGEMENT LEVELS – FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



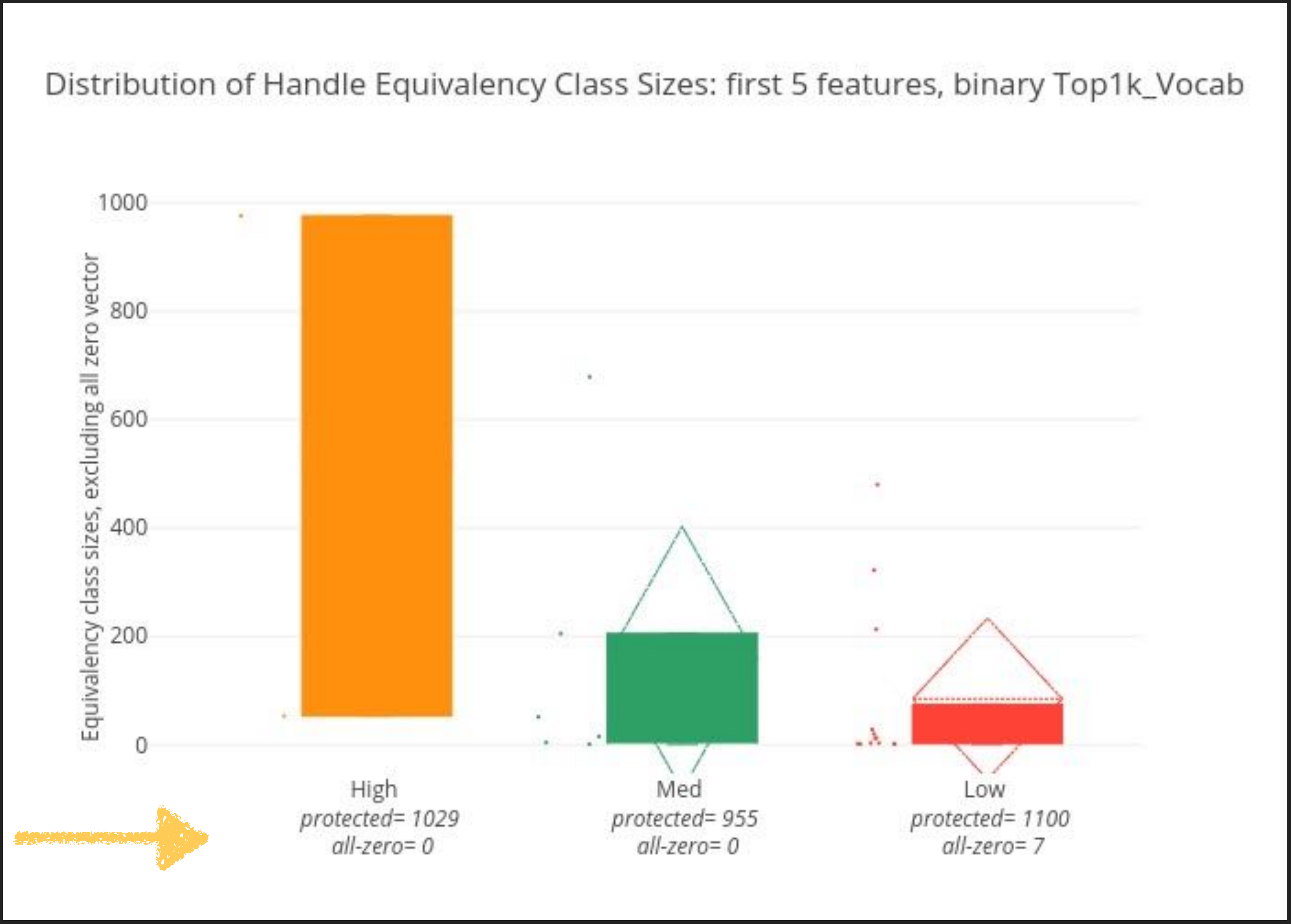
PRIVACY - ACROSS ENGAGEMENT LEVELS - FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



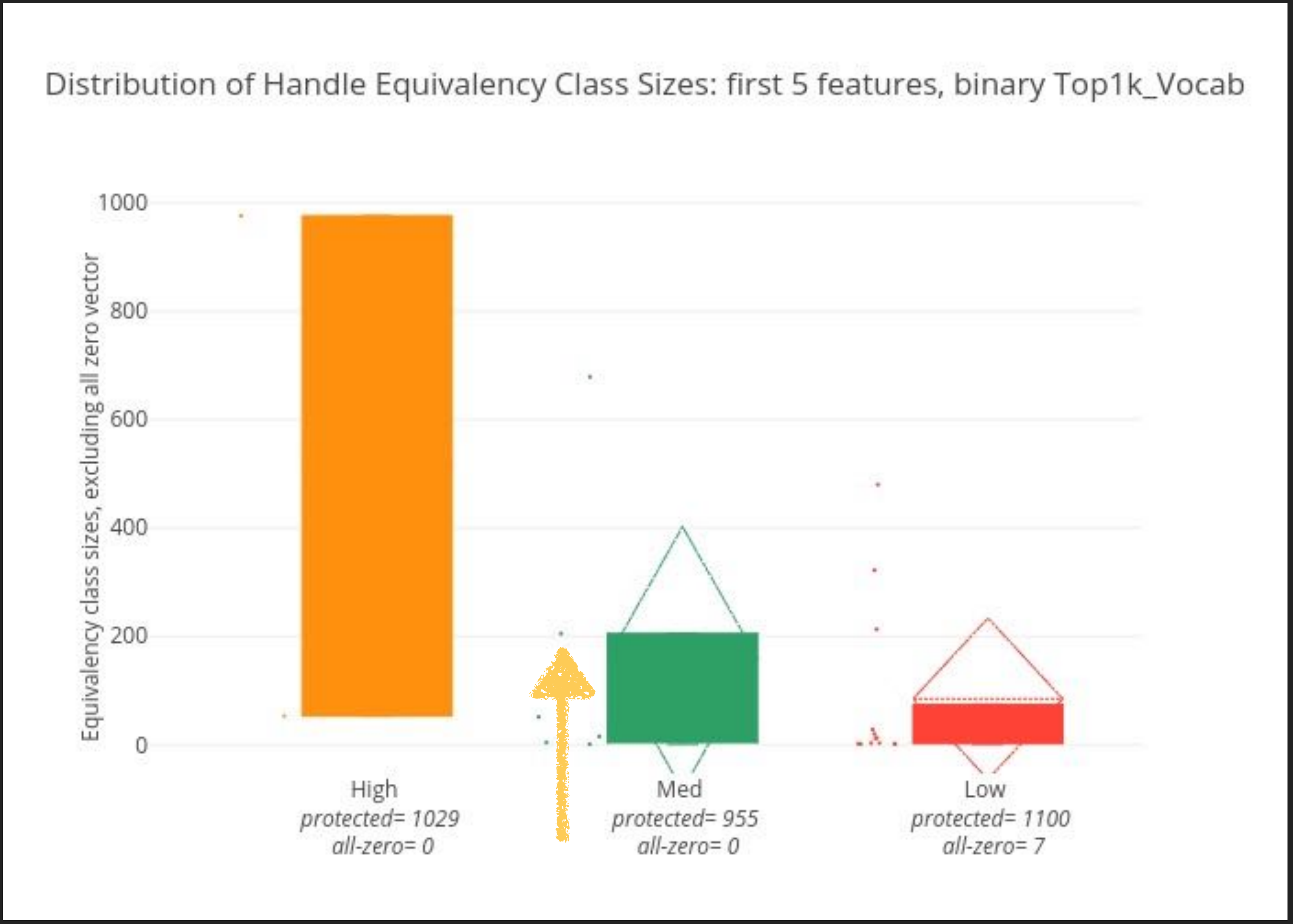
PRIVACY - ACROSS ENGAGEMENT LEVELS - FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



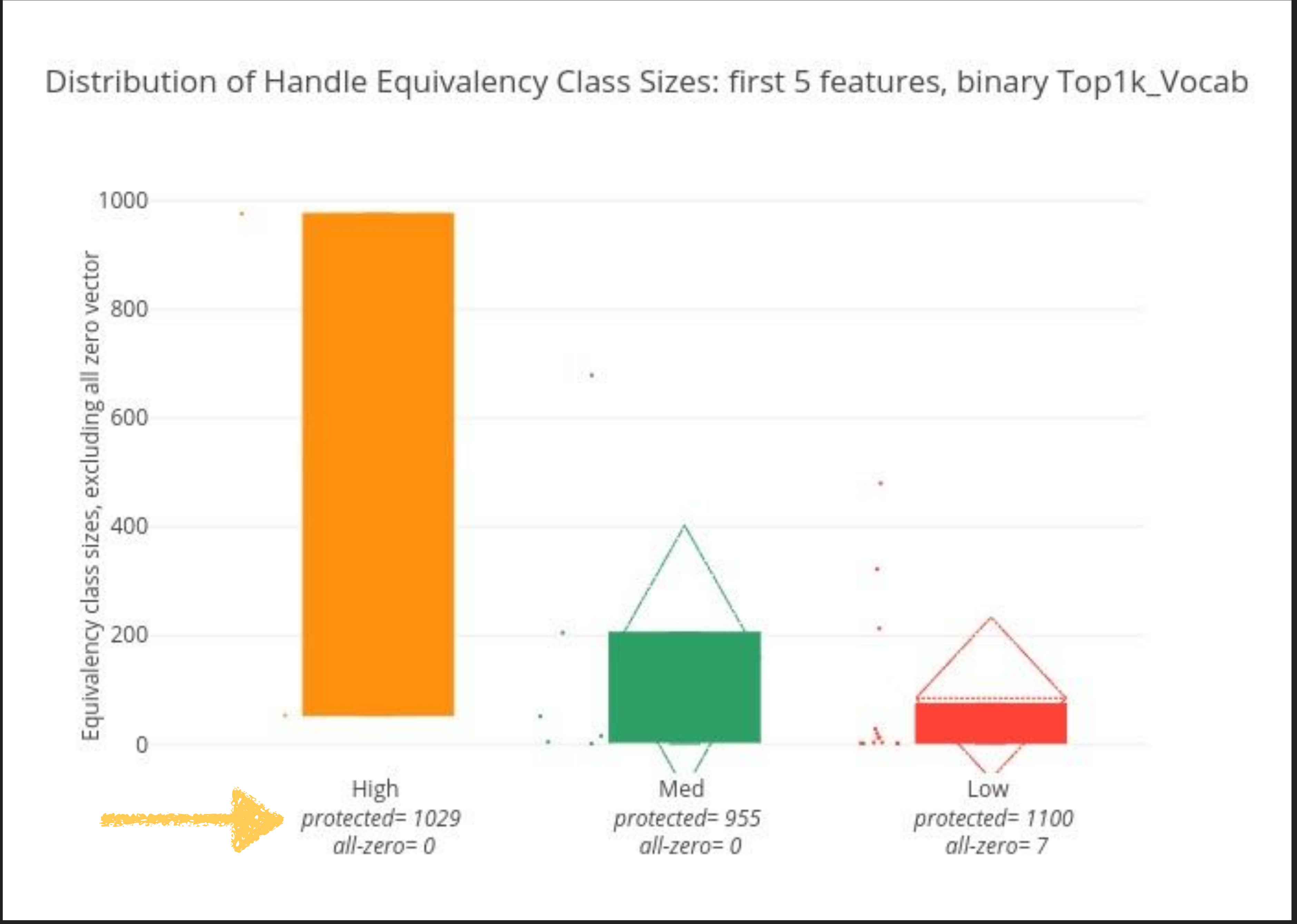
PRIVACY - ACROSS ENGAGEMENT LEVELS - FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



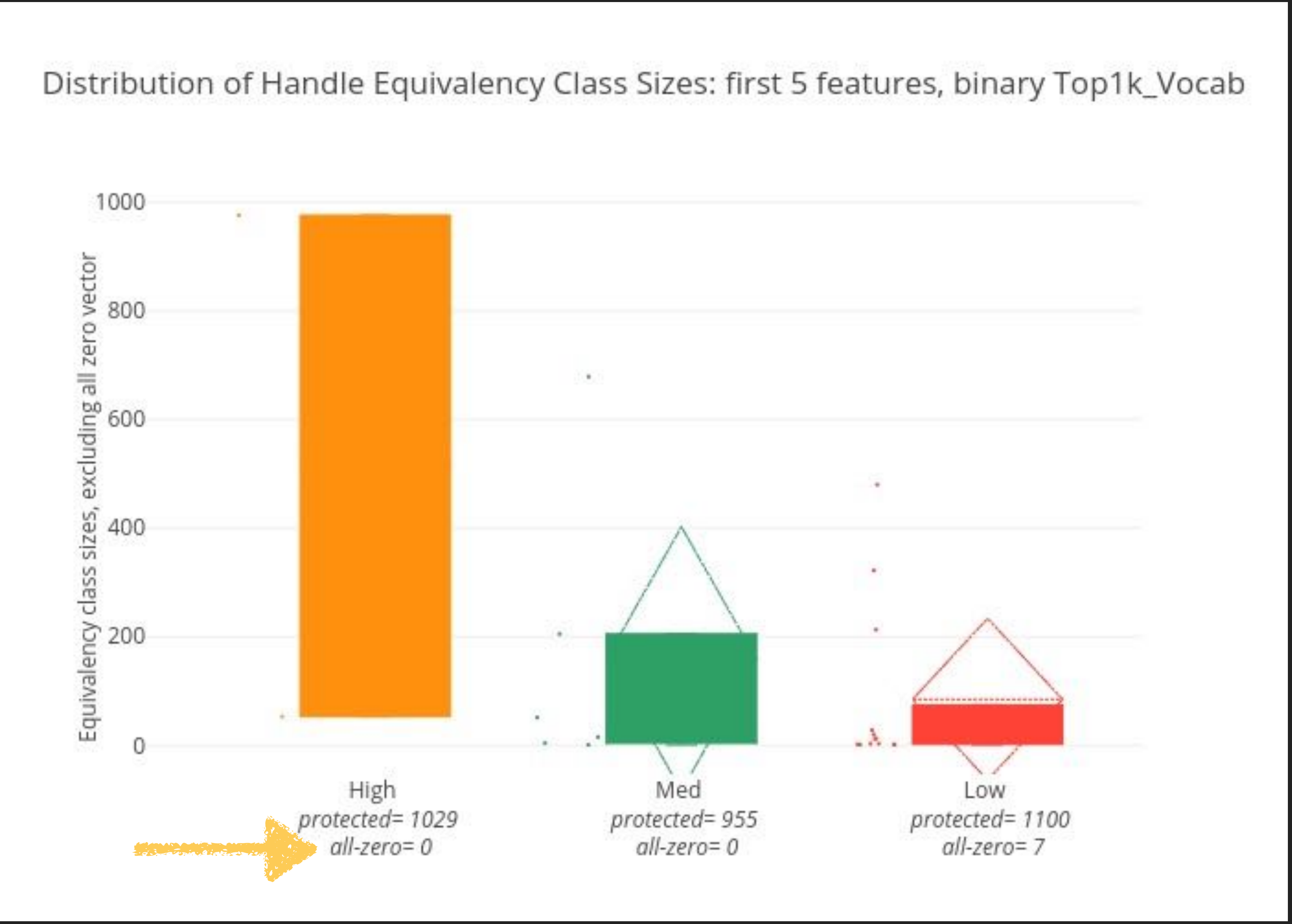
PRIVACY - ACROSS ENGAGEMENT LEVELS - FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



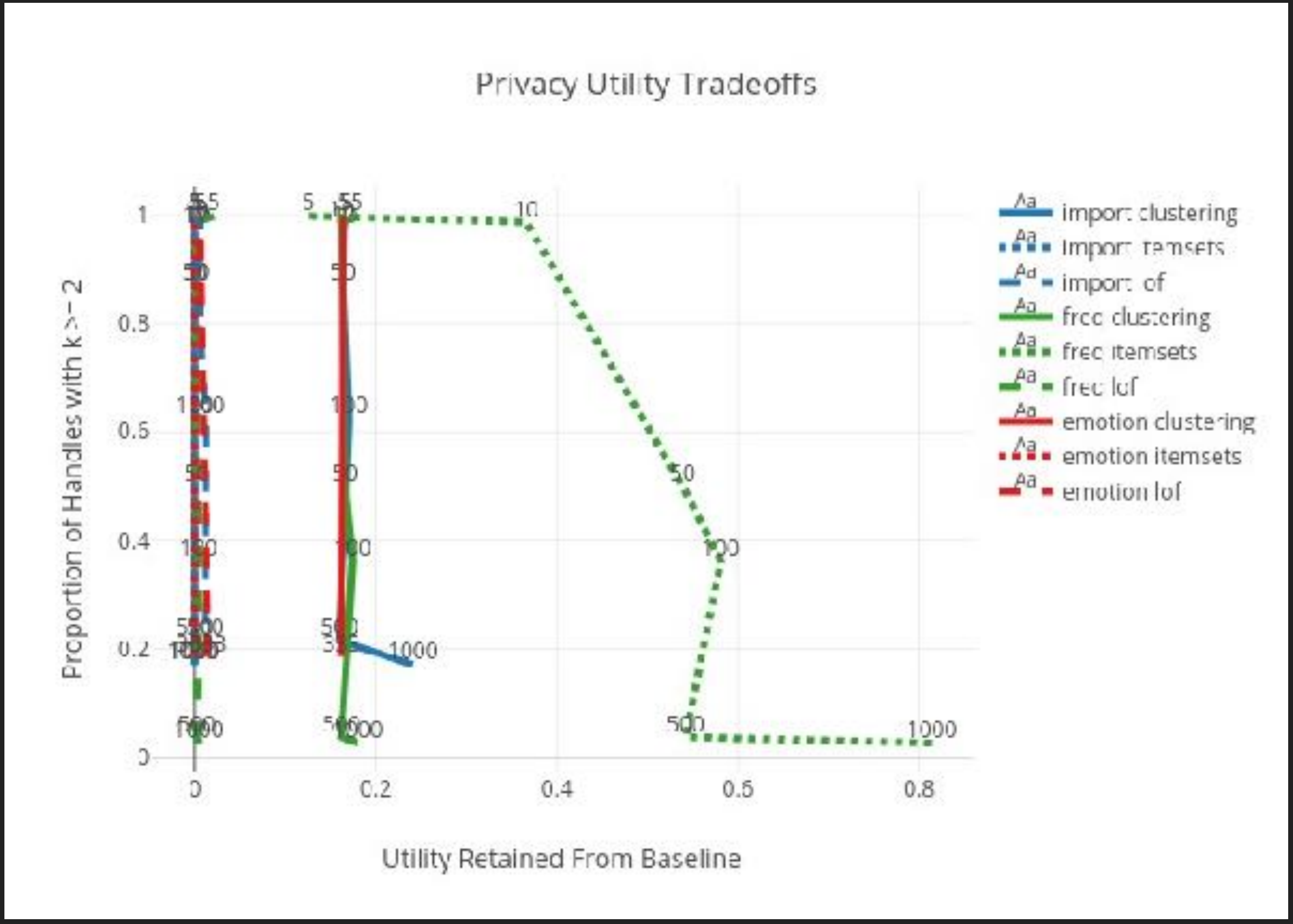
PRIVACY – ACROSS ENGAGEMENT LEVELS – FREQUENCY PROJECTION

- High average *k* for Frequency projection, small vector size
- More engagement corresponds to higher average *k*
- All individuals protected



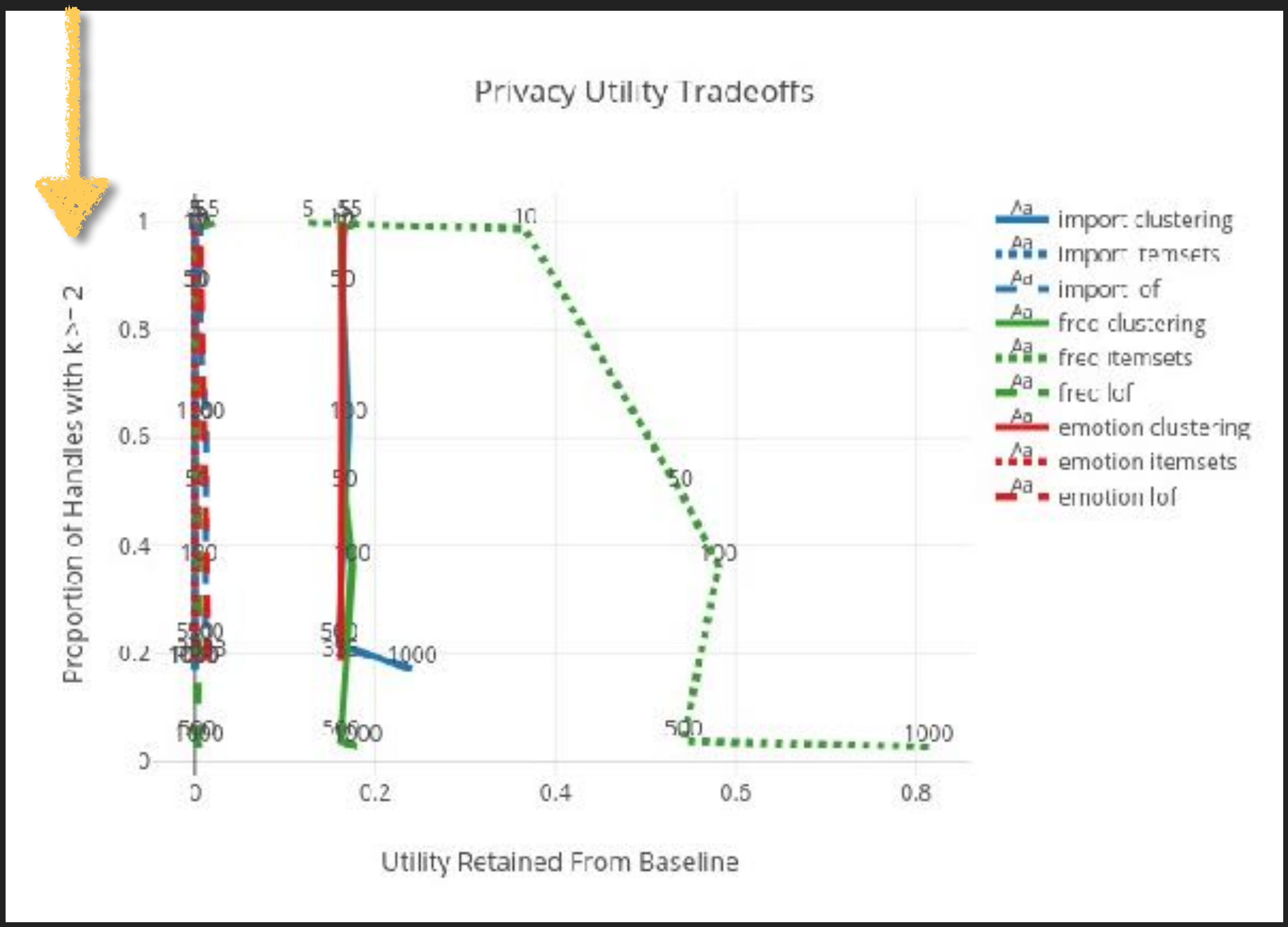
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- Only meaningful utility retention for clustering task and frequent item set mining task



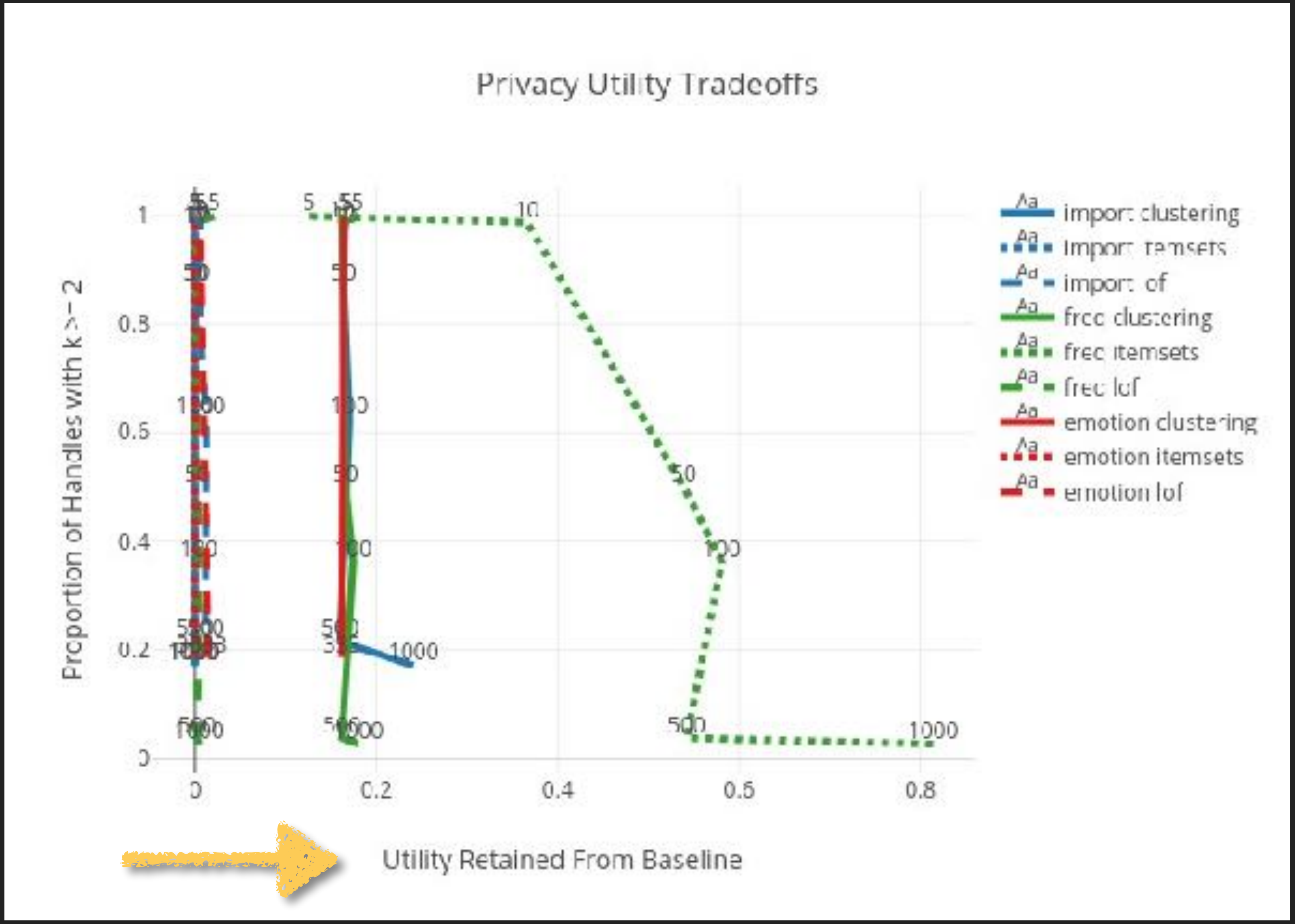
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- Only meaningful utility retention for clustering task and frequent item set mining task



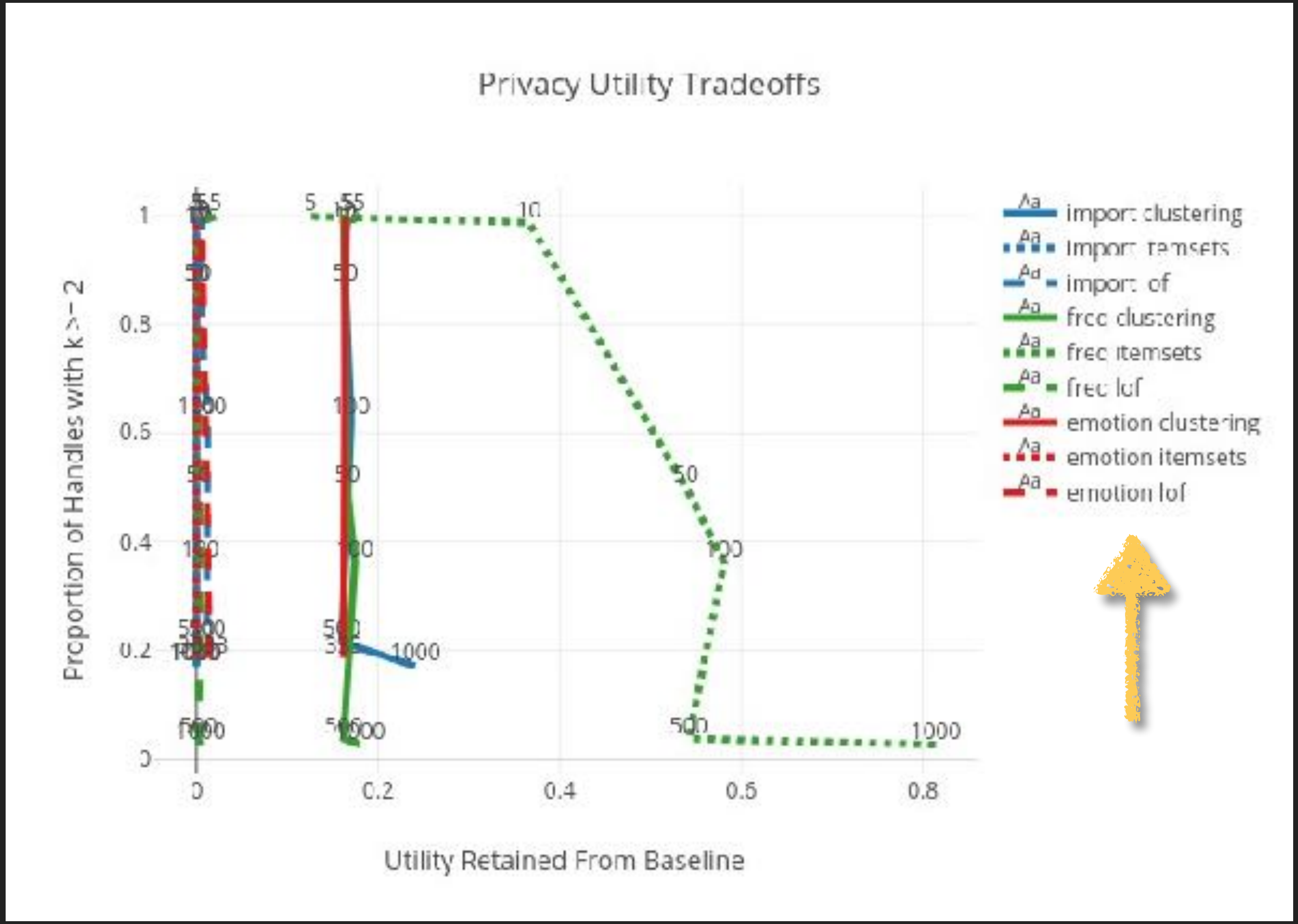
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- Only meaningful utility retention for clustering task and frequent item set mining task



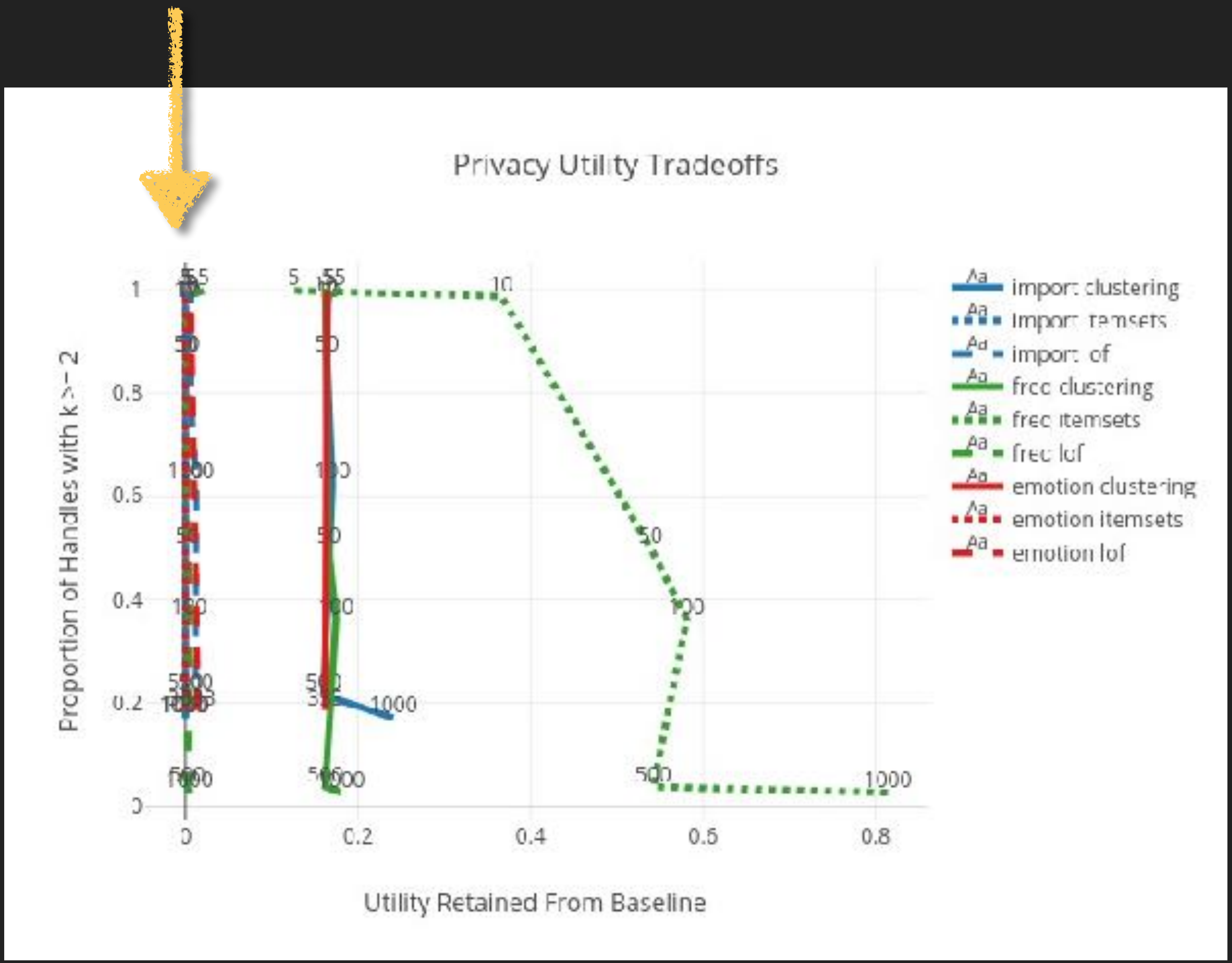
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- Only meaningful utility retention for clustering task and frequent item set mining task



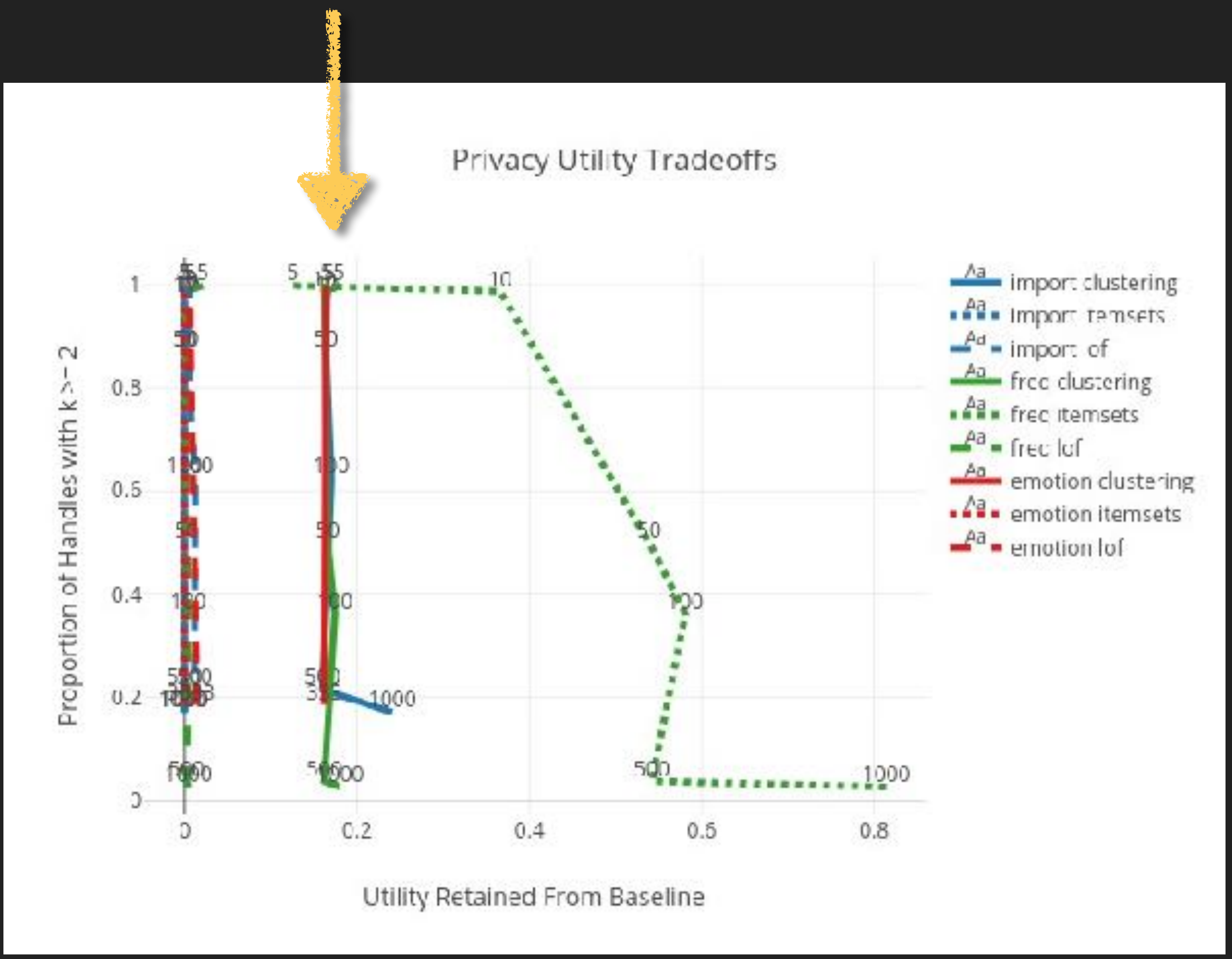
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- **Only meaningful utility retention for clustering task and frequent item set mining task**



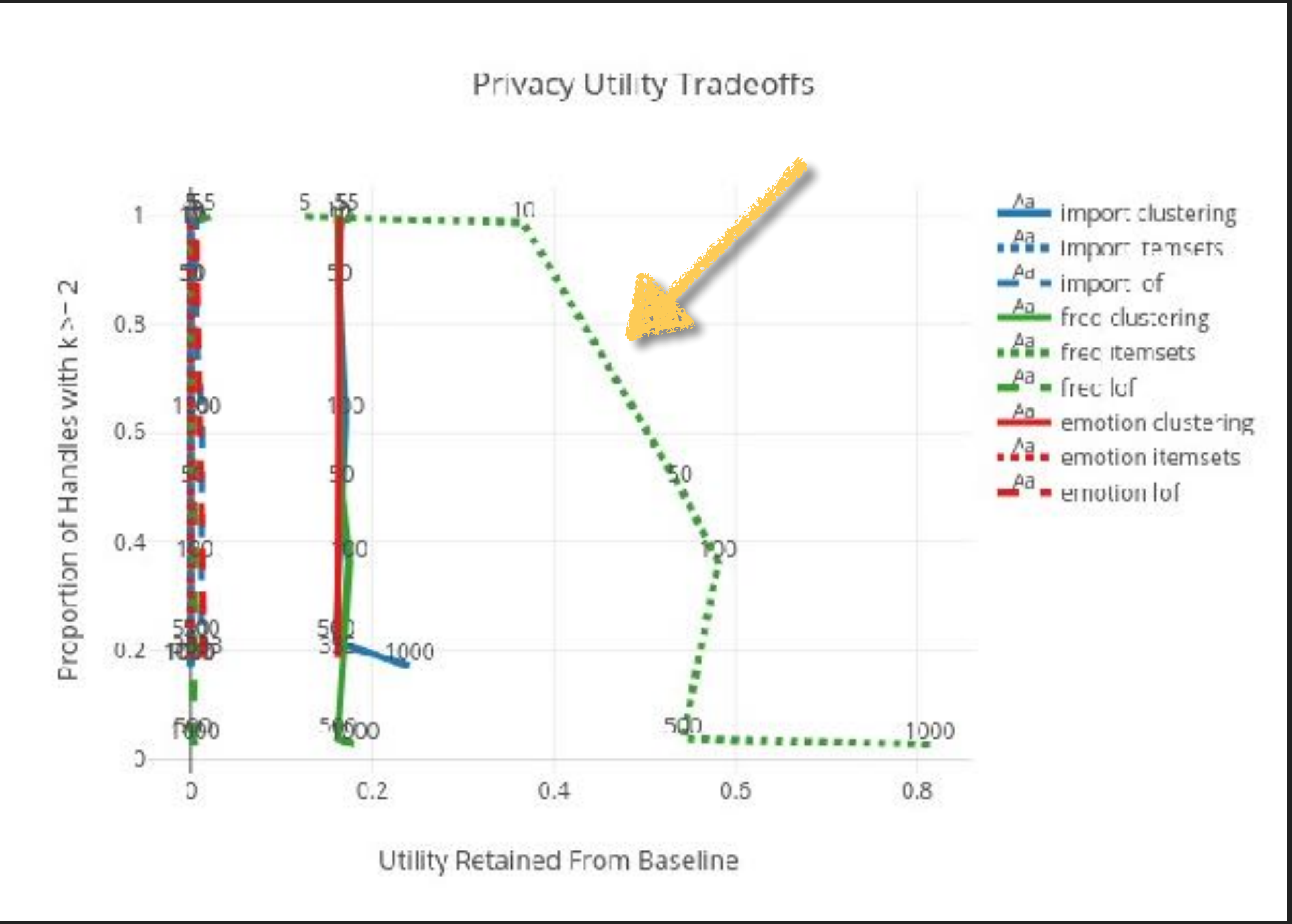
RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- **Only meaningful utility retention for clustering task and frequent item set mining task**



RESULTS - PRIVACY UTILITY TRADEOFFS

- Utility retention and proportion of individuals with privacy for a range of projection feature space sizes
- **Only meaningful utility retention for clustering task and frequent item set mining task**



RESULTS - SUMMARY

- Difficult to obtain k anonymous vector privacy measure greater than 1 for all tweet streams without a high loss of data utility
- Vector length had little or no impact on utility for all projections, besides item set mining on Frequency projection
- For Frequency projection, interesting how frequent item set mining performance changes according to vector length
- The relationship between utility and privacy is very task dependent
  - Shouldn't expect general solution - projection selection should be task dependent

CONTRIBUTIONS

- Novel analysis of social media distinguishability
- Framework for analyzing privacy-utility tradeoffs of different representations of social media texts posts
- Empirical analysis showing users are only private if represented by a small number of features, but this results in high data utility loss

FUTURE WORK

- Better understand what types of transformations (including mathematical transformations) help maintain any level of utility for specific data mining tasks
  - Because of sparseness of our projections, mathematical transformation have potential for increased privacy with low impact on data utility.
- Impact of retweet frequency on privacy

REFERENCES

- Dodds, et al. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2015.
- Horspool and Cormack. Constructing word-based text compression algorithms. In *Data Compression Conference*, 1992., pages 62–71, March 1992.
- Li and Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 517–526, New York, NY, USA, 2009. ACM.
- Qadir and Riloff. Bootstrapped learning of emotion hashtags #hashtags4you. 2013.

REFERENCES

- Roberts et al. Empatweet: Annotating and detecting emotions on twitter. 2012.
- L. Singh et al. Public information exposure detection: Helping users understand their web foot- prints. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 153–161, Aug 2015.
- Ian Witten et al. Semantic and generative models for lossy text compression. 37, 07 2000.
- Yang et al. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pages 133–136, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- Sweeney. *k*-anonymity: a model for protecting privacy. 2002.

QUESTIONS

?