The Meaning Extraction Method and its Friendly Helper: The What, Why, and How of the MEM

Downloads for today's tutorial:
Software: http://meh.ryanb.cc/download
Sample Text: http://meh.ryanb.cc/AustinSubScrape.zip
Lecture: http://meh.ryanb.cc/MEM_Lecture.pdf

Ryan L. Boyd
The University of Texas at Austin
2014-02-17

Our Goals for Today

- 1. The “What” of MEM
  - What are the basic concepts and use of meaning extraction?
- 2. The “Why” of MEM
  - Why should I care?
- 3. The “How” of MEM
  - Show me the magical ways of meaning extraction.
- 4. Let’s Try MEM
  - Demonstration using Meaning Extraction Helper

The "What" of MEM

- The Meaning Extraction Method
  - Introduced in 2008 by Chung & Pennebaker
  - Described as an automated method for extracting themes from a body of text
  - Demonstrably useful from a psychological research standpoint

The "What" of MEM

- The Meaning Extraction Method
  - The main idea: find meaningful word clusters
    - Function vs. Content words
  - "A factor analysis of [common] words"
    - Simply put, a statistical way of finding correlation clusters
    - Requires a guiding hand (yours!)

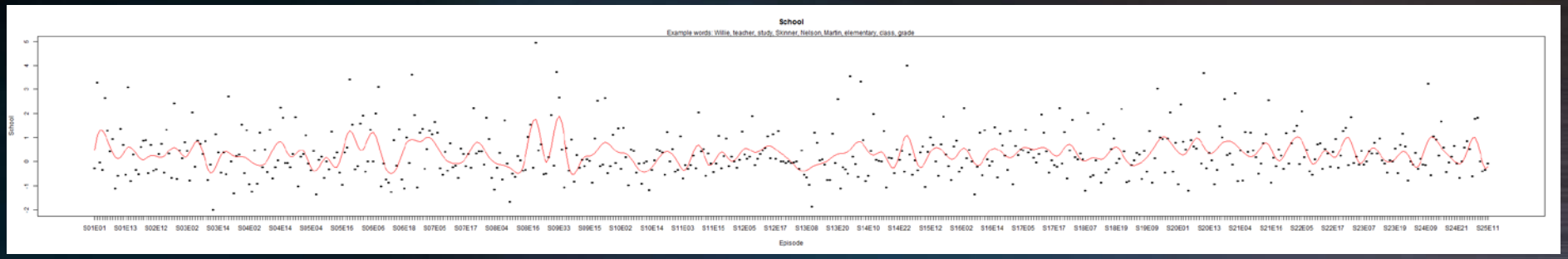
The "What" of MEM

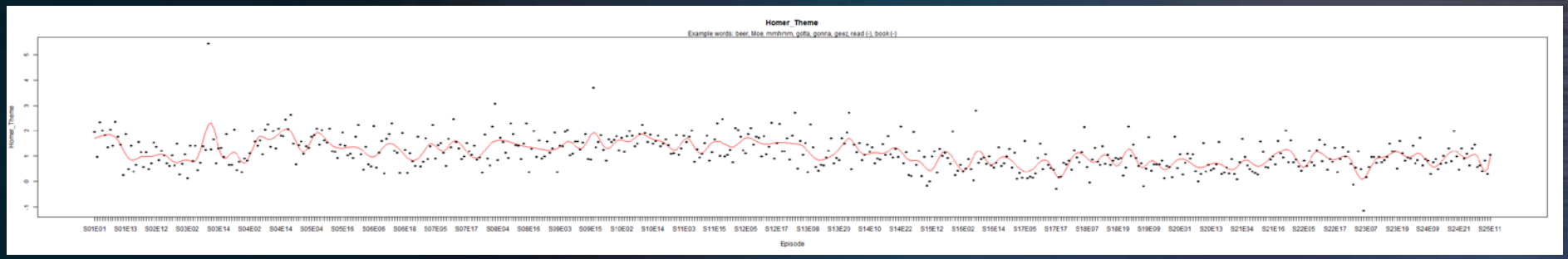
Sample: The Simpsons (all 541 episodes)	
School	teacher, test, study, student, Willie, Skinner, principal
Money	store, sold, sell, pay, money, dollar, cost, buy, buck
Homer	uh-oh, mmmm, d'oh, Homer, hmm, beer
Family	son, Simpson, mom, Lisa, honey, family, dad
Food	taste, stomach, restaurant, hungry, food, eat, delicious
Recreation	Saturday, radio, party, fun, city, car, park
Religion	Sunday, Reverend, miracle, Lord, Jesus, holy, heaven

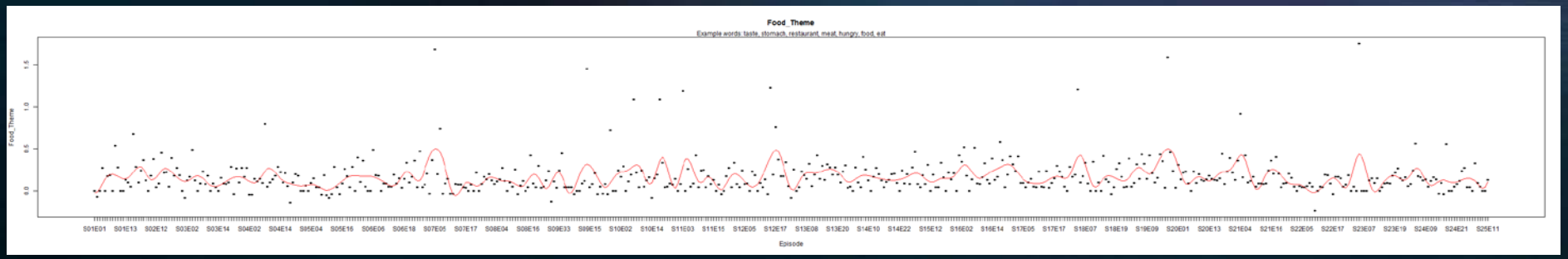
The "What" of MEM

- Once you have established themes using MEM, it is possible to determine how the themes are used in your sample.
- Use theme quantification like any other language measure (e.g., LIWC categories)
  - The powerful utility of MEM is the quantification of themes without a need for raters / subjective judgment calls on a "per subject" basis

The “What” of MEM

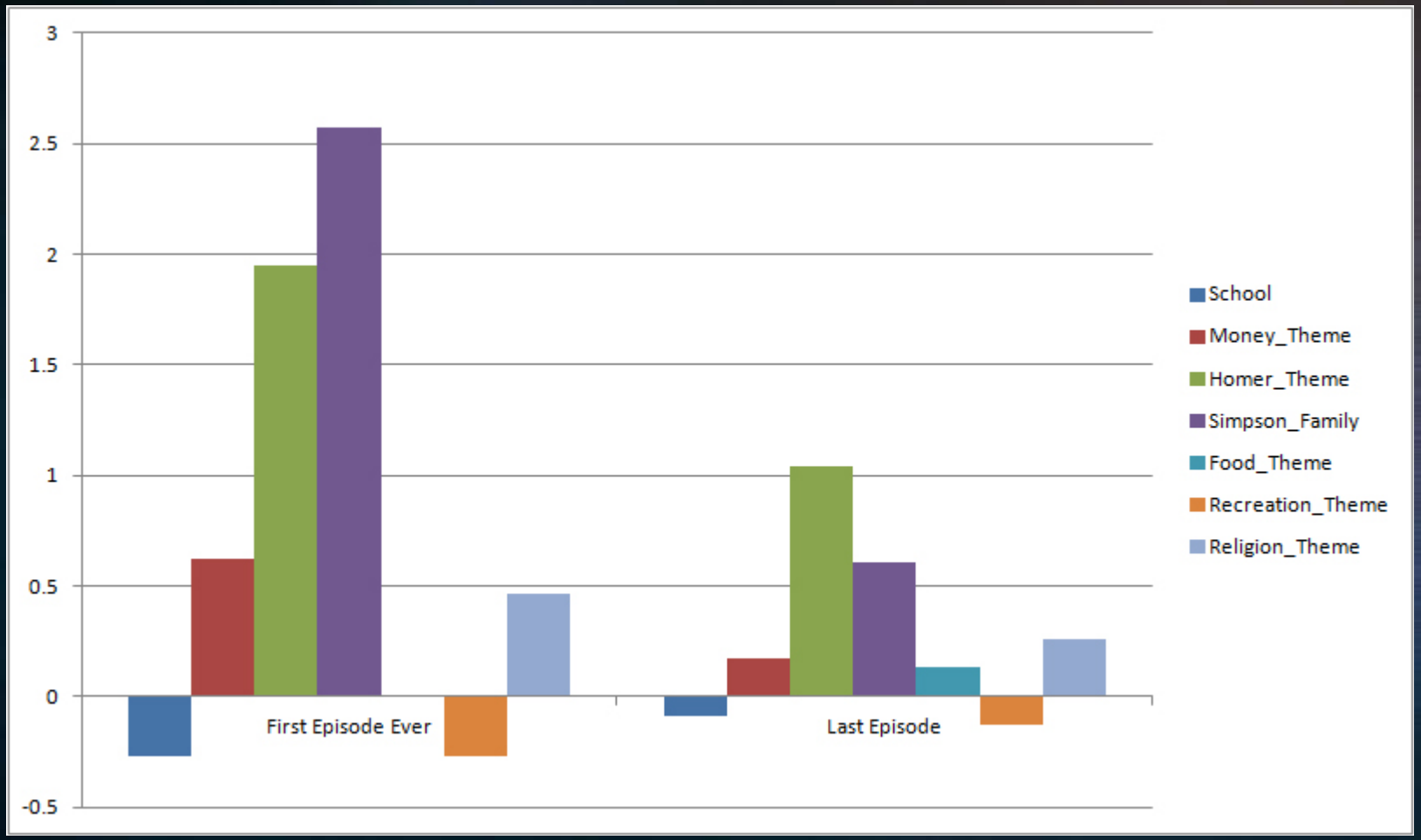






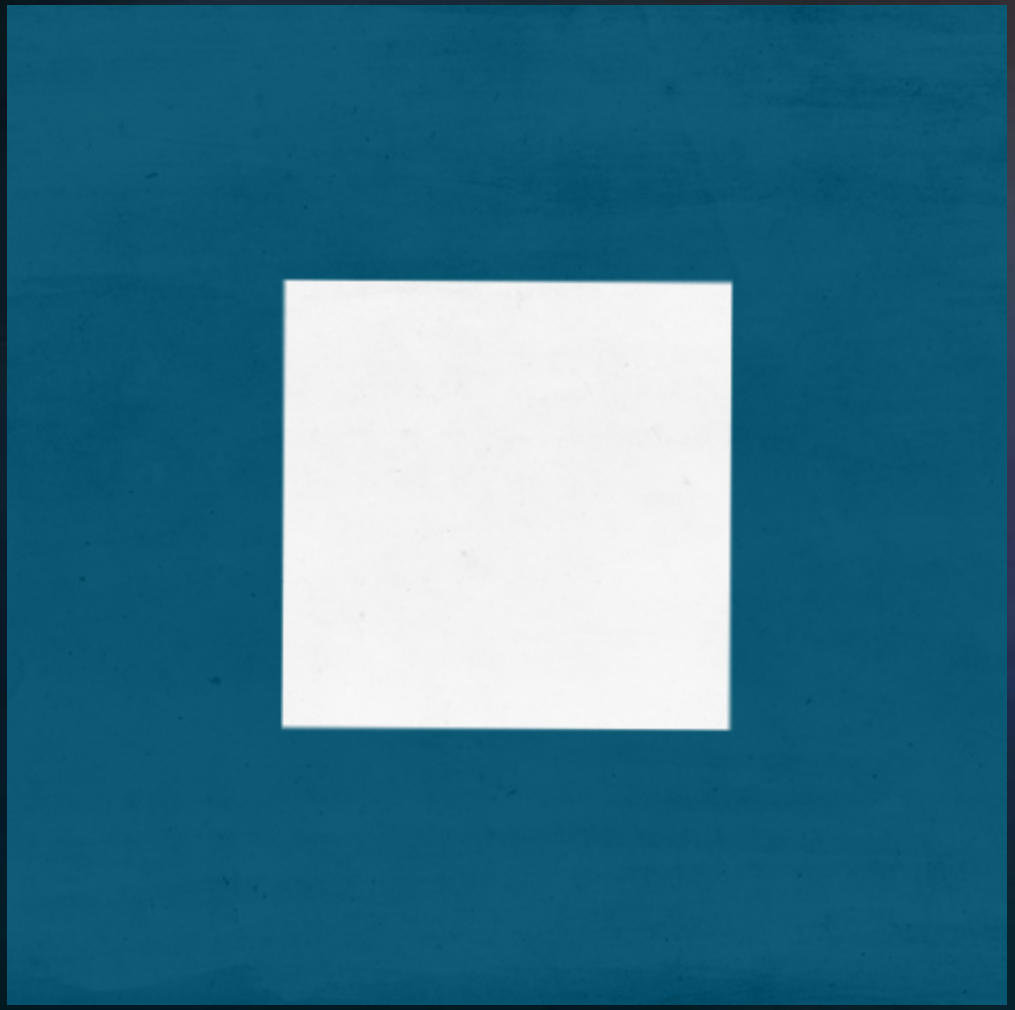
For example, look for changes across time…

The "What" of MEM



…or look at composites.

Questions?

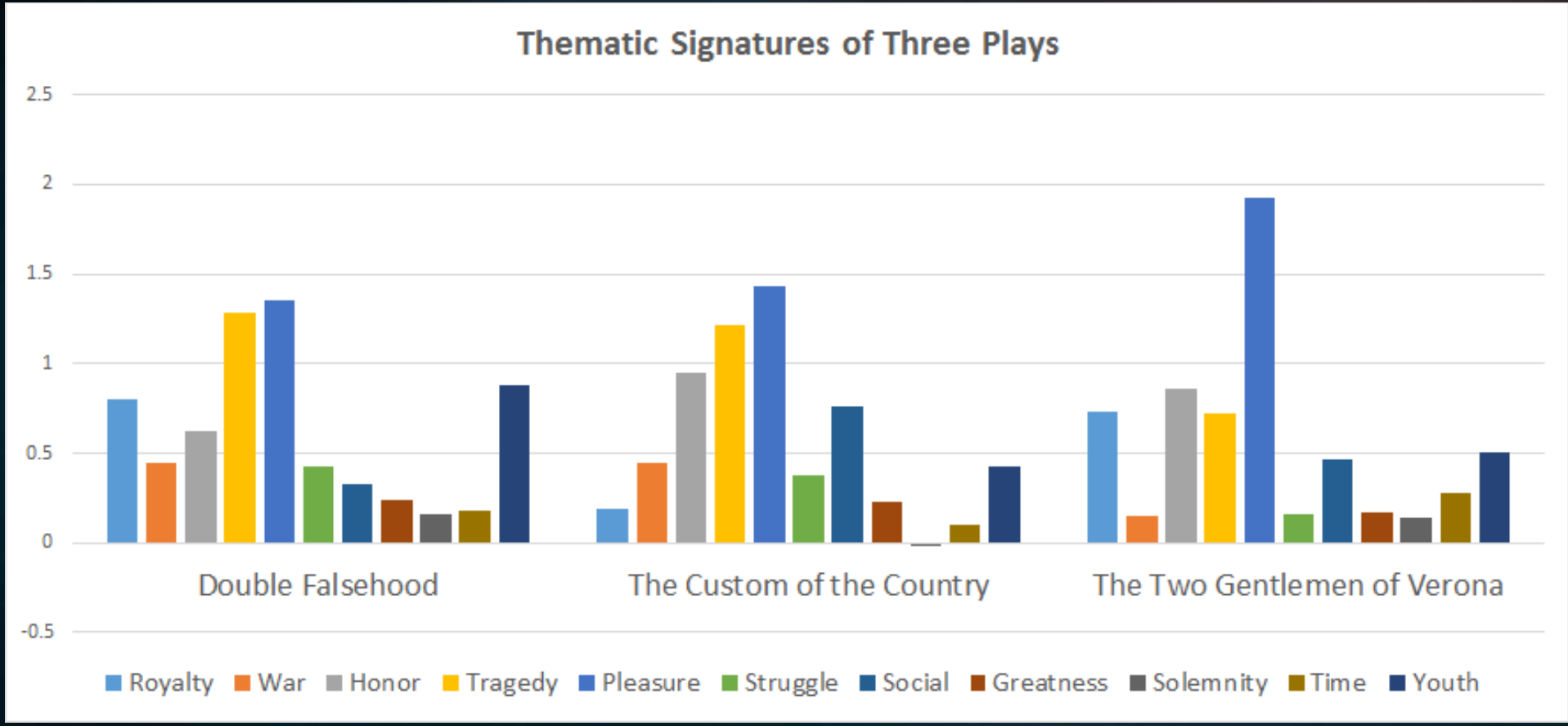


- **The "Why" of MEM**

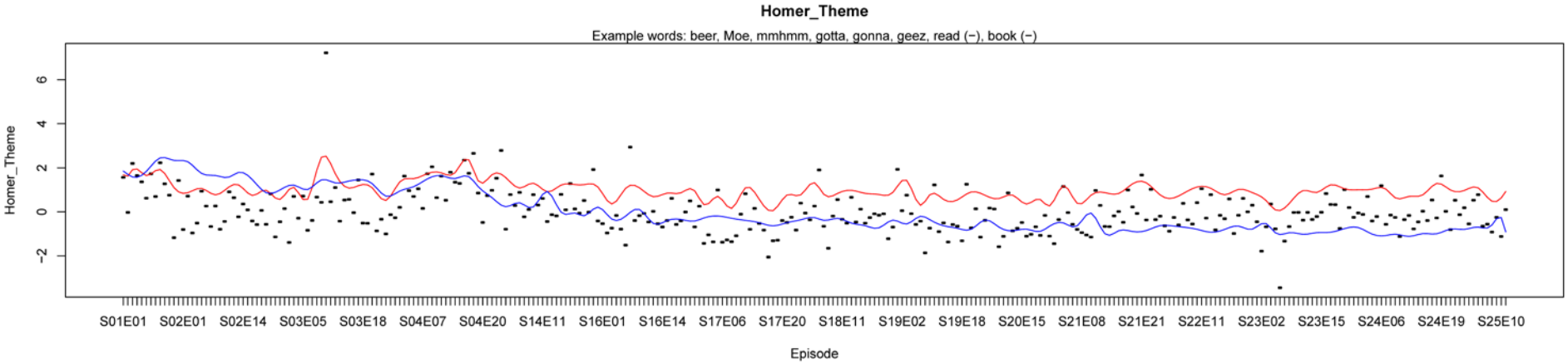
  - Why should I care?

    - Chung & Pennebaker, 2008: Personality correlates
    - Pennebaker, 2011: An insight into focus / thinking style
    - Dream research: experience → dream thematics
    - Pasca research: Changes over time
    - Double Falsehood
    - The Simpsons

The "Why" of MEM



The "Why" of MEM

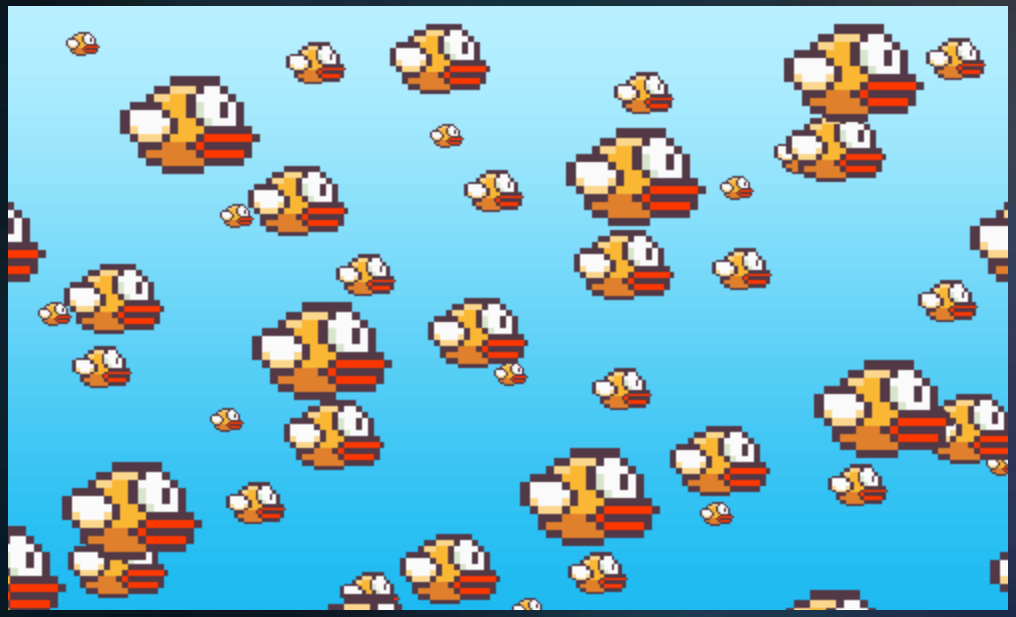


Red = Theme trend (smoothed)

Blue = Viewership trend (smoothed)

*Out of more than 600 language variables (including LIWC variables), the "Homer" theme was the 2nd strongest predictor of episode viewership (r = .37, p < .0001).

Questions?



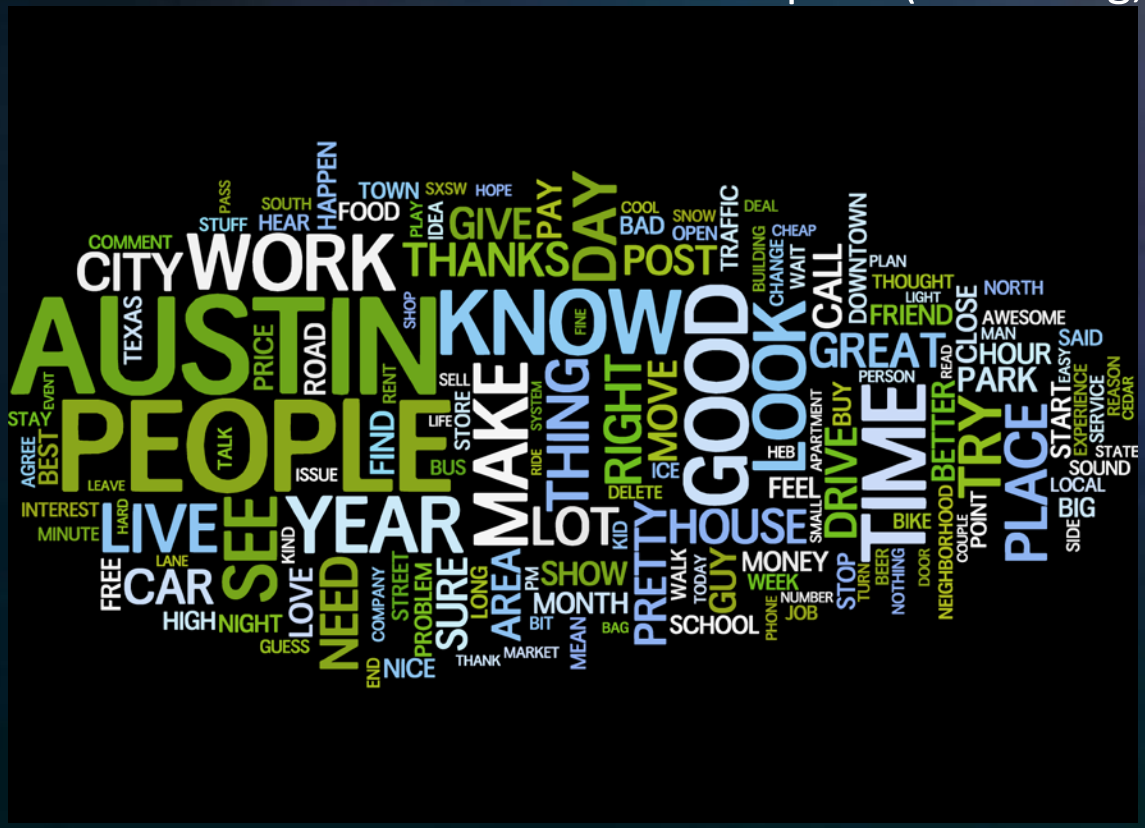
The "How" of MEM

- Overview
  - 1. Discover High-Frequency Content Words
    - Ignore function words
    - Ignore words of non-interest
    - Make sure inflections are collapsed
      - "talk" = "talking" = "talks" = "talked"
  - 2. Scan for these Words
    - Figure out when Stage 1 words are being used
    - Approximately equal size samples (min, max)
  - 3. Find Out How They Relate to Each Other
    - Principal Components Analysis

The "How" of MEM

Stage 1

- 1. Discover High-Frequency Content Words
  - Ignore function words
  - Ignore words of non-interest
  - Make sure inflections are collapsed (stemming, lemmatization)



The “How” of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

The "How" of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

Word	Frequency
OVER	1
THERE	1
YOU	2
CAN	2
SEE	1
THE	3
VEGETABLE	1
VEGETABLES	2
STORE	3
BUY	1
ESPECIALLY	1
AT	2
THEY	1
SELL	1
MANY	1
TYPES	1
OF	1

The "How" of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

Many words do not carry any useful meaning for us here…

Word	Frequency
OVER	1
THERE	1
YOU	2
CAN	2
SEE	1
THE	3
VEGETABLE	1
VEGETABLES	2
STORE	3
BUY	1
ESPECIALLY	1
AT	2
THEY	1
SELL	1
MANY	1
TYPES	1
OF	1

The “How” of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

…so we ignore them – these are called “stop words”.

Word	Frequency
SEE	1
VEGETABLE	1
VEGETABLES	2
STORE	3
BUY	1
ESPECIALLY	1
SELL	1
TYPES	1

The "How" of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

Now, we want to collapse words into their common root…

Word	Frequency
SEE	1
VEGETABLE	1
VEGETABLES	2
STORE	3
BUY	1
ESPECIALLY	1
SELL	1
TYPES	1

The "How" of MEM

Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

…lemmatization does this with high accuracy.

Word	Frequency
SEE	1
VEGETABLE	3
STORE	3
BUY	1
ESPECIALLY	1
SELL	1
TYPE	1

The "How" of MEM

Stage 2

- Figure out when Stage 1 words are being used, and where
- Approximately equal size samples (min, max)
  - If our body of text is largely comprised of observations that all have approximately the same number of words, this is not a problem.
    - An exception is if all of the samples are very large
  - If we have observations with lots of different word counts, or we have a sample of really large observations, we need to "segment" them into samples of about the same size

The “How” of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

The “How” of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

[1] Over there, you can see the vegetable store.
You can buy vegetables, especially at the store.
At the store, they sell many types of vegetables.

The "How" of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

> [1] Over there, you can see the vegetable store.
> You can buy vegetables, especially at the store.
> At the store, they sell many types of vegetables.

- Vegetable(s) = 100% of observations
- Store = 100% of observations
- Co-occurrence = 100%
- But, we only have 1 observation. Let's split it up!

The “How” of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

[1] Over there, you can see the vegetable store.
[2] You can buy vegetables, especially at the store.
[3] At the store, they sell many types of vegetables.

The "How" of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

[1] Over there, you can see the vegetable store.
[2] You can buy vegetables, especially at the store.
[3] At the store, they sell many types of vegetables.

- Vegetable(s) = 100% of observations
- Store = 100% of observations
- Co-occurrence = 100%
- Same outcome, but better power. What if we continue?

The “How” of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

[1] Over there, you can	[2] see the vegetable store.
[3] You can buy vegetables,	[4] especially at the store.
[5] At the store, they sell	[6] many types of vegetables.

The "How" of MEM

- Finding the right segmentation numbers is a bit like tuning a radio.

[1] Over there, you can	[2] see the vegetable store.
[3] You can buy vegetables,	[4] especially at the store.
[5] At the store, they sell	[6] many types of vegetables.

- Vegetable(s) = 50% of observations
- Store = 50% of observations
- Co-occurrence = 16.6%
- Too much splitting hides the interesting word patterns.

The "How" of MEM

- As a rule of thumb…
  - Try maximum word counts between 100-500
  - Use minimum word counts that are approximately half of the maximum
    - For example:
      - Min = 50, Max = 100
      - Min = 100, Max = 200
      - Min = 125, Max = 250
      - Min = 250, Max = 500

The “How” of MEM

Stage 3

- Find out how words relate to each other

The "How" of MEM

[1] Over there, you can see the vegetable store.
[2] You can buy vegetables, especially at the store.
[3] At the store, they sell many types of vegetables.

	VEGETABLE(S)	STORE
Segment 1	1	1
Segment 2	1	1
Segment 3	1	1

Good segmentation example – Variables form a good factor

Correlation = 1.00

The "How" of MEM

[1] Over there, you can	[2] see the vegetable store.
[3] You can buy vegetables,	[4] especially at the store.
[5] At the store, they sell	[6] many types of vegetables.

	VEGETABLE(S)	STORE
Segment 1	0	0
Segment 2	1	1
Segment 3	1	0
Segment 4	0	1
Segment 5	0	1
Segment 6	1	0

Bad segmentation example – Variables might not factor at all

Correlation = -0.33

Questions?

- **The "How" of MEM**

- Finally, you can find out how each observation (text sample) scores on each theme.
  - For example, rescan files for theme words in a LIWC-style manner, then combine word variables appropriately
- You can then use these scores in your statistical analyses.

Let's try MEM!

- …with a little bit of help, of course.
- Meaning Extraction Helper (MEH) is able to do all of Stage 1 and Stage 2 for us in an automated fashion.
- This makes the Meaning Extraction Method fast, easy, and highly accurate.

Website contains a lot of useful information about using MEH.

Simpsons Example:
http://imgur.com/a/jRd2Y

/r/Austin Example:
http://imgur.com/a/YtMmk

Contact: ryanboyd@utexas.edu
Feel free to ask for help!