



# NEW MEDIA AND SENTIMENT MINING

## PRACTICAL ASPECTS OF SENTIMENT MINING

OVER  
**6,250** GRADUATE  
ALUMNI

OFFERING OVER  
**150** ENTERPRISE IT, INNOVATION  
& LEADERSHIP PROGRAMMES

TRAINING OVER  
**135,000** DIGITAL LEADERS  
& PROFESSIONALS

# PRACTICAL ASPECTS OF SENTIMENT MINING

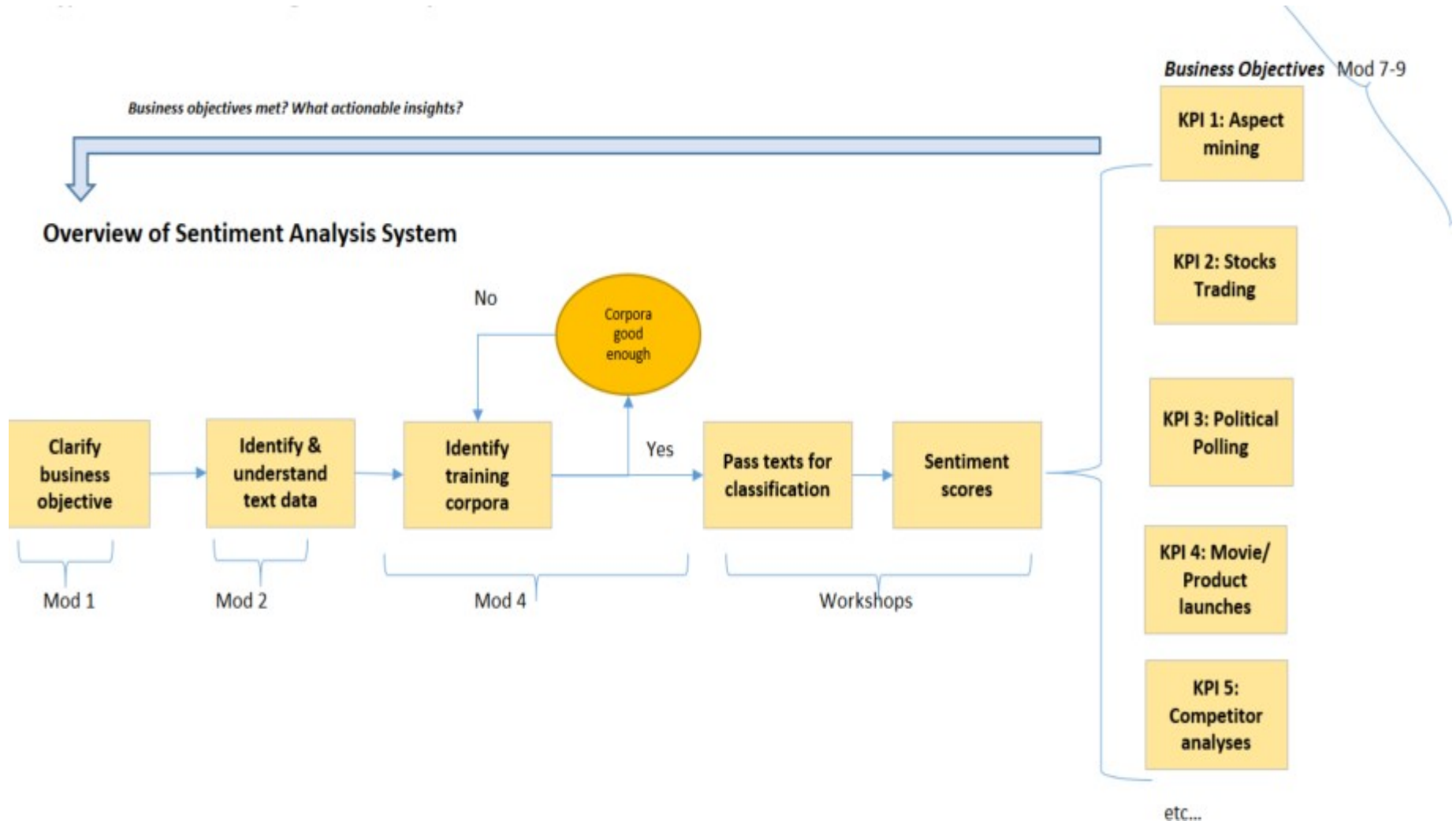


Eric Tham  
**Senior Lecturer**  
[isstyc@nus.edu.sg](mailto:isstyc@nus.edu.sg)

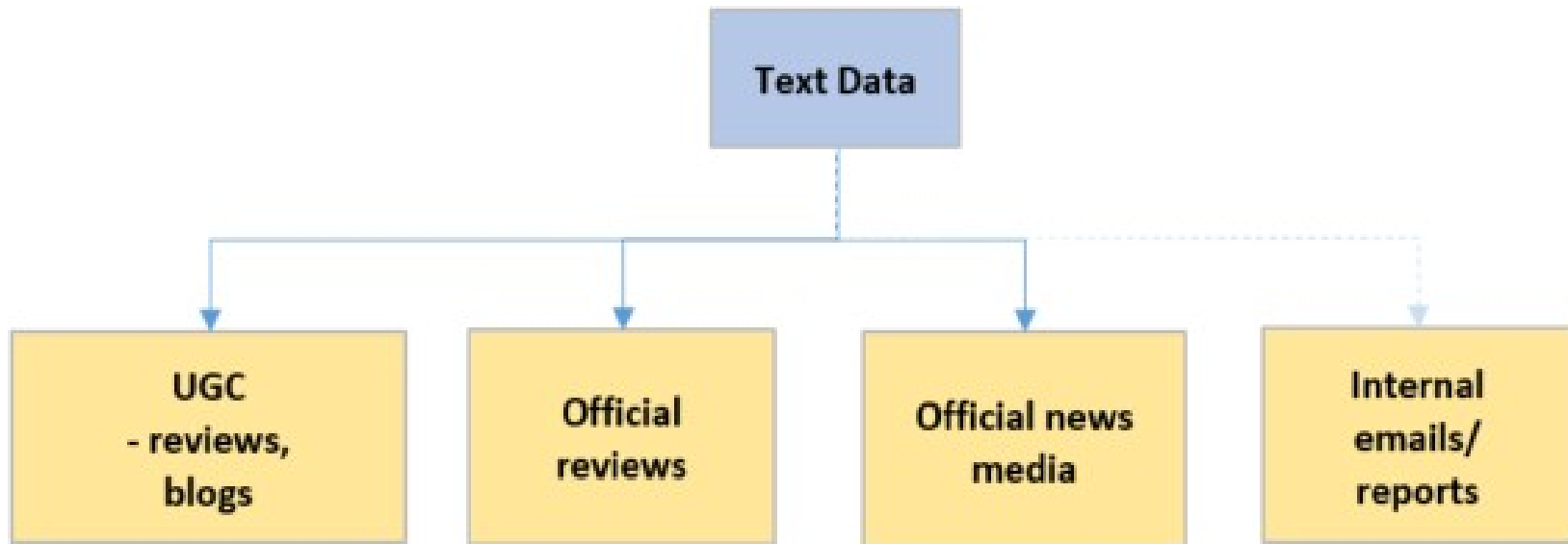
# Agenda

- Sentiment summarisation and visualization
- Psychological aspects of sentiment
- Other sentiment metrics/ indices
- Sentiment mining libraries:
  - Building a sentiment analysis engine

# Overview of Sentiment Analysis System



# Types of Text Data



## Types of Text Data in Digital Economy

# Why build a sentiment engine?

- There must be *actionable insights* from the generation of sentiment scores
  - These insights must be well communicated to stakeholders via sentiment visualization (eg dashboards).
  - Egs of actionable insights
    - Identify under-valued stocks to buy?
    - Identify customers feeling unhappy about services/ products?
- 
-

- Sentiment summarisation:
  - Aggregates sentiment from different texts over time to communicate insights
- Recall the quadruple (o, t, t, s) from sentiment definition and quintuple (o, a, t, t, s) for aspect-based sentiment analysis?
- Summarisation is done on three main levels:
  - the overall sentiment/ opinion across the writers (o)
  - the sentiment scores across time (t)
  - The aspects (a) or entity targets talked about (t)

# Sentiment aggregation (opinion holders)

- Aggregating over opinion holders, differing weights:
  - Which opinion holders is most important in influencing opinions?
  - What are the factors to consider?
- Consumer marketing
  - Influencers (likely to cause virality)
  - No of tweets/ re-tweets (from meta-data)
- Financial news
  - Types of news: regular or extraordinary reporting
  - News source: more reputed sources WSJ or social media?



# Sentiment aggregation (over time)

- Aggregating over time
  - This needs to be considered with the trading strategies and sentiment decay.
  - Trading horizon:
    - Often sentiment is cited for intra-day trading. In this case, should aggregate over shorter time-frame?
    - Also, need to tie in with the advertising campaign. Over days or weeks?

An example will be demonstrated in the workshop.

# Sentiment visualization

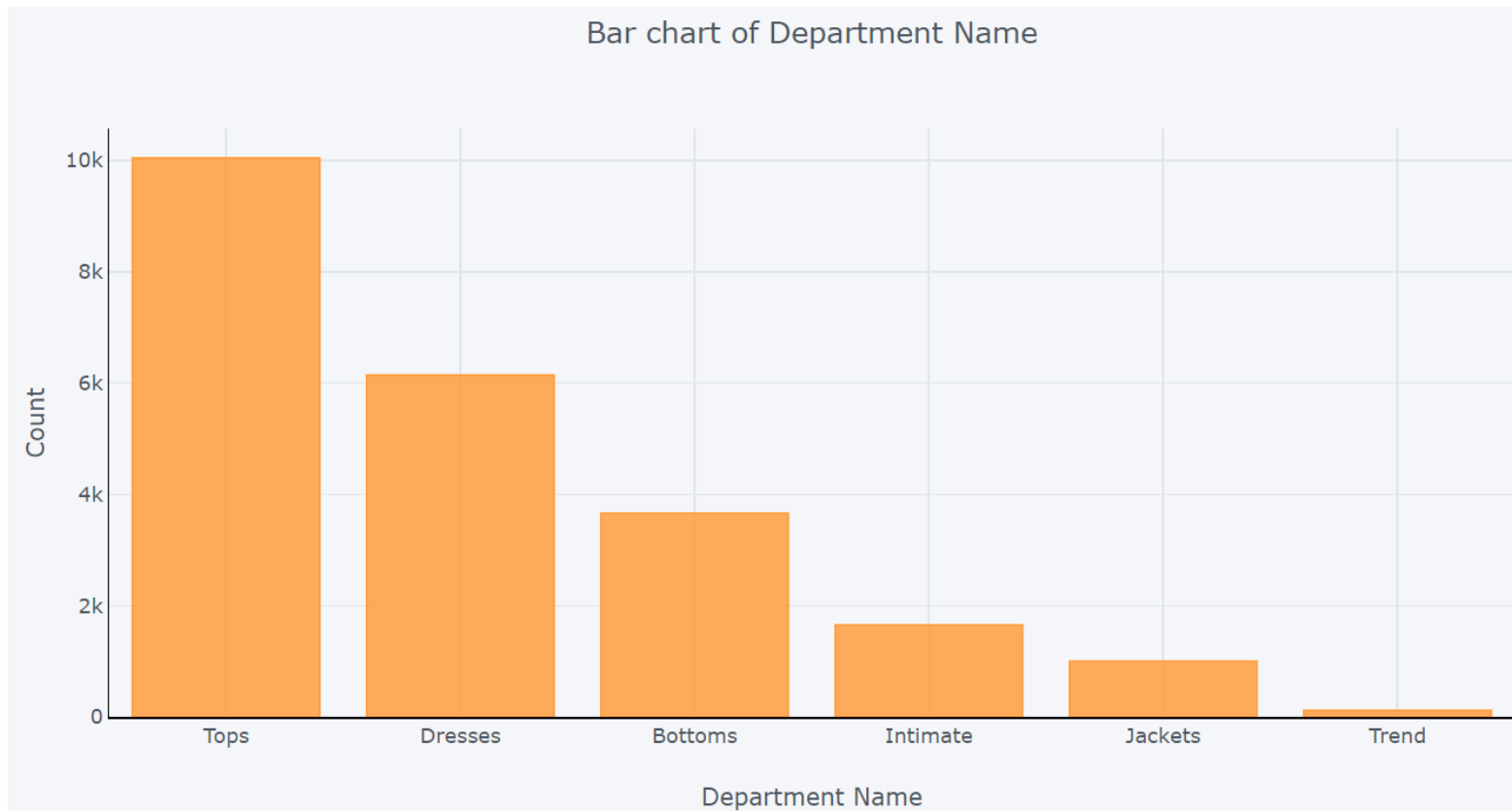
- Objectives of sentiment visualisation:
  - Communicate key results for stakeholders and enable decision-making
- First, some rules of thumbs for visualization:
  - The visualization needs to communicate only **2-3 points**, which needs to jump out of your visuals. This is a right number of 'points' to capture the audience's attention and interest. Too many, it becomes too complicated.
  - It should encourage **data exploration**. The visuals should invoke questions from the audience.
  - Optionally, it should be **visually aesthetic**. Everyone remembers a 'beautiful' diagram. Of course, aesthetics needs to go hand in hand with substance.

# Visualisation Libraries

- Python libraries to use:
  - bokeh, seaborn (and the basic matplotlib)
  - bokeh:
    - <https://bokeh.pydata.org/en/latest/index.html>
    - Allows some interactivity
  - seaborn
    - <https://seaborn.pydata.org/>
    - Useful for heatmap and treemap
  - dash plotly (demonstration)
- d3.js, crossfilter.js
  - Requires javascript knowledge but best in class for visualisation for sentiment (and other visualisation)

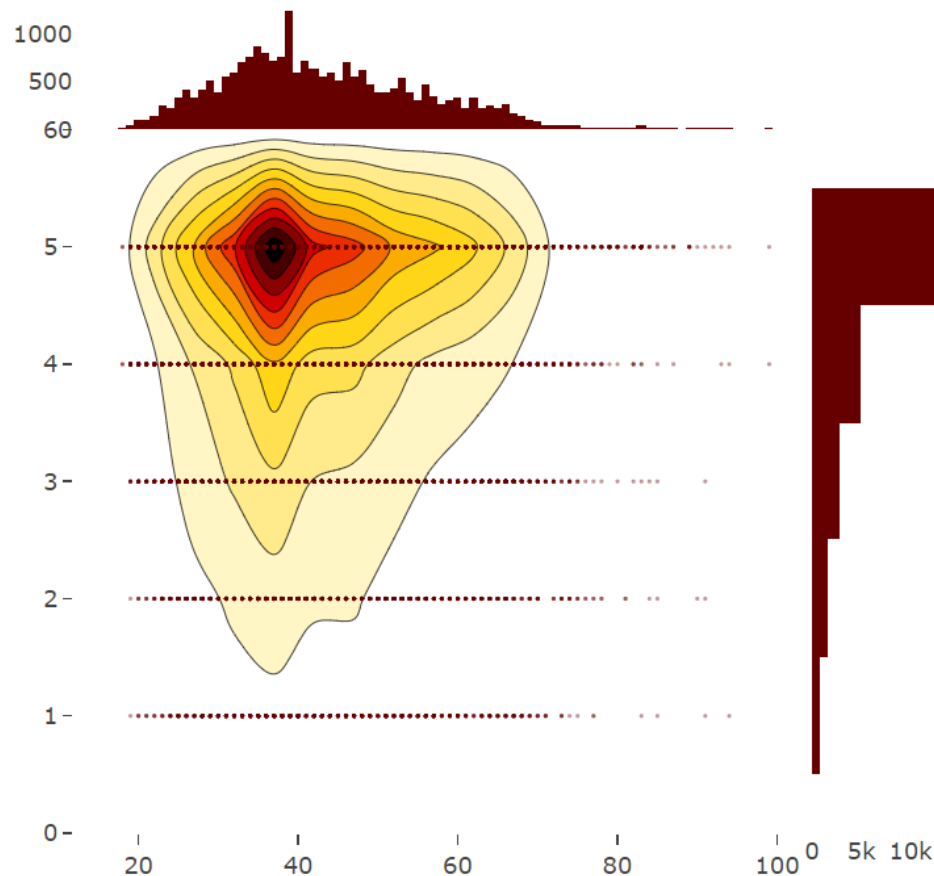
# Sentiment Aggregation

- Dimensions of how you want the sentiment to be aggregated by:
  - Time, organisational units, aspects, products etc.
  - Metrics can be both sentiment and no. of comments



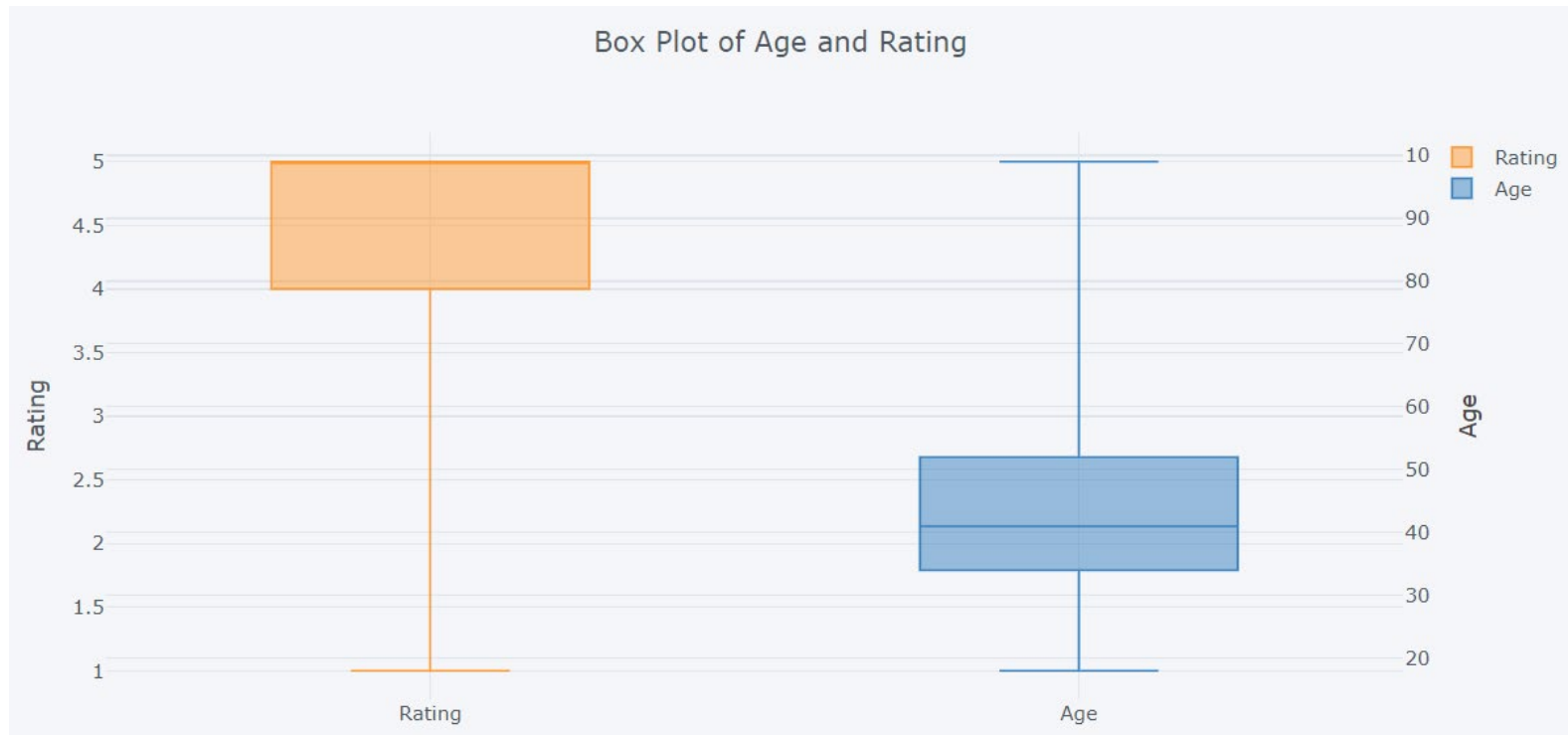
# Sentiment Aggregation (2d)

- 2-D segmentation: in this case by age and no of ratings to the rating score (contour)



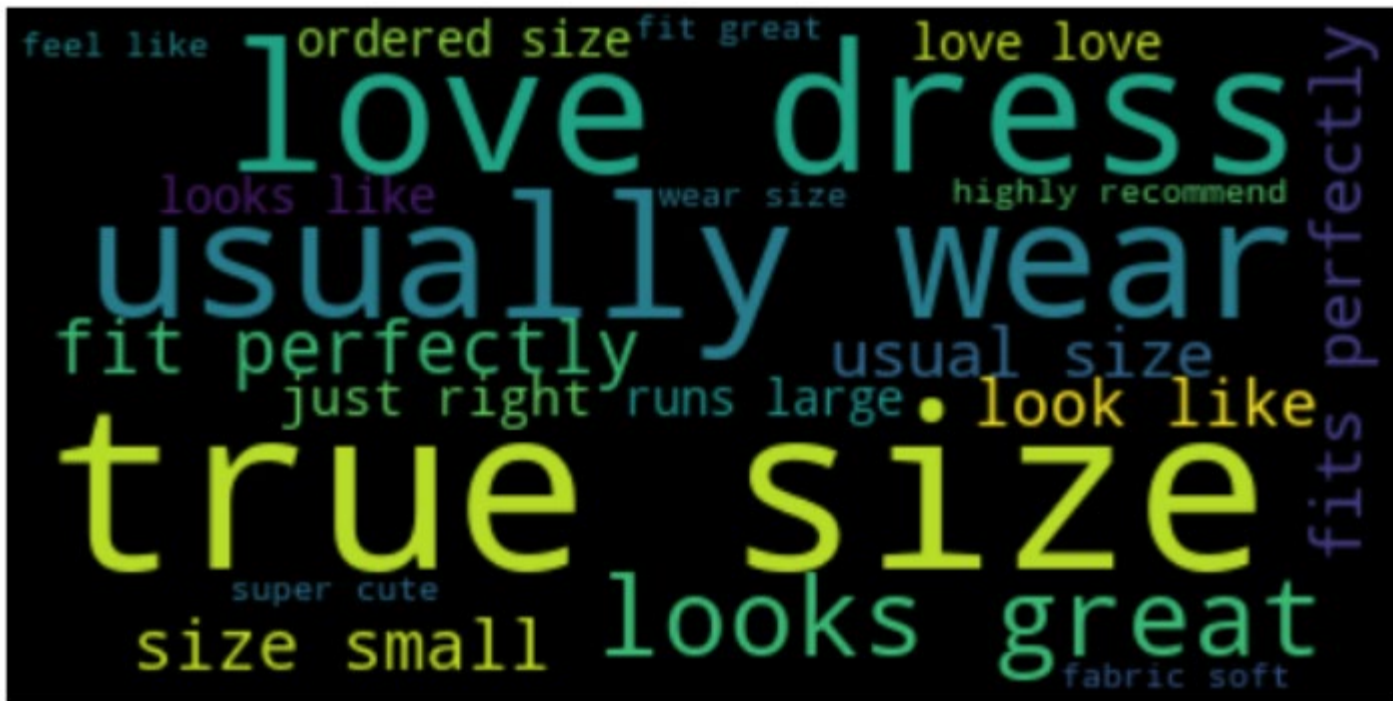
# Boxplot (range of ratings)

- Rating scores by age



# Word Clouds

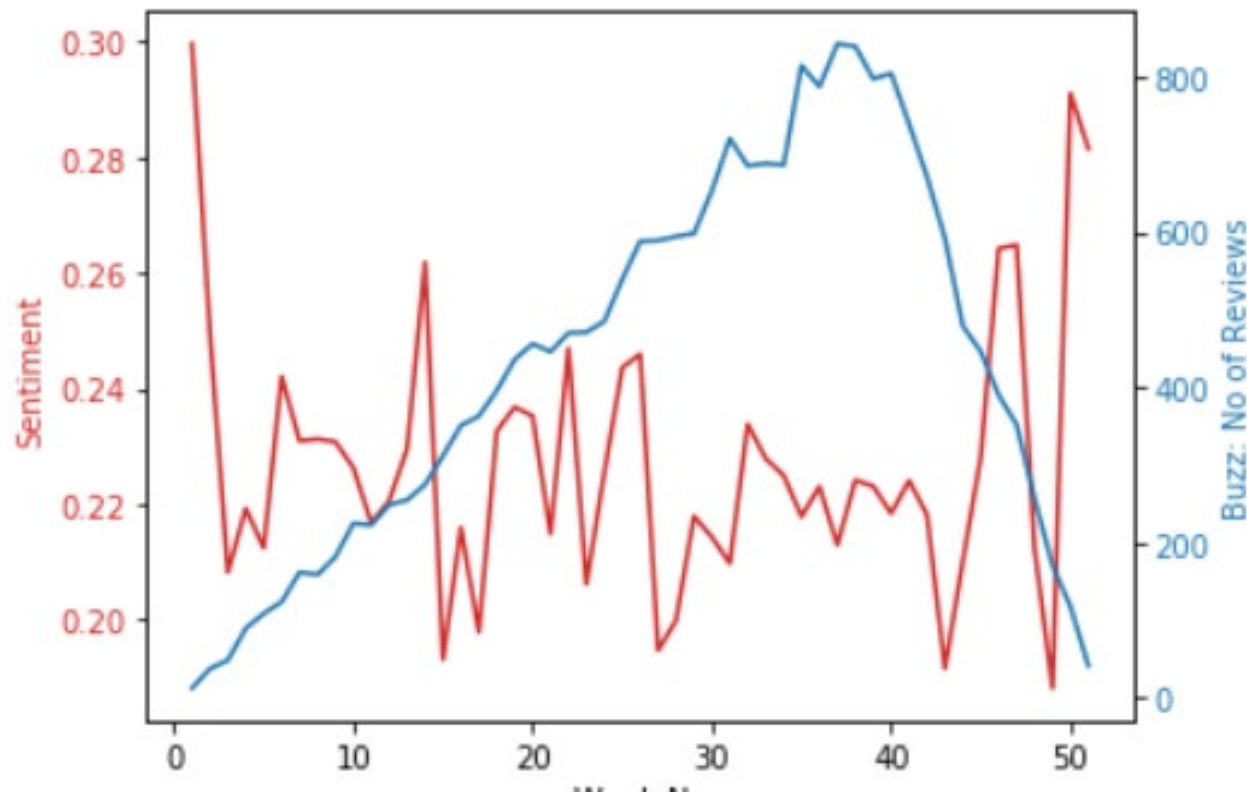
- Visualises the most common words in comments. Can be a prelude to using PMI or most associated aspect words to entities



Pointwise mutual information (PMI), or point mutual information, is a measure of association used in information theory and statistics. In contrast to mutual information (MI) which builds upon PMI, it refers to single events, whereas MI refers to the average of all possible events.

# Time series of sentiment

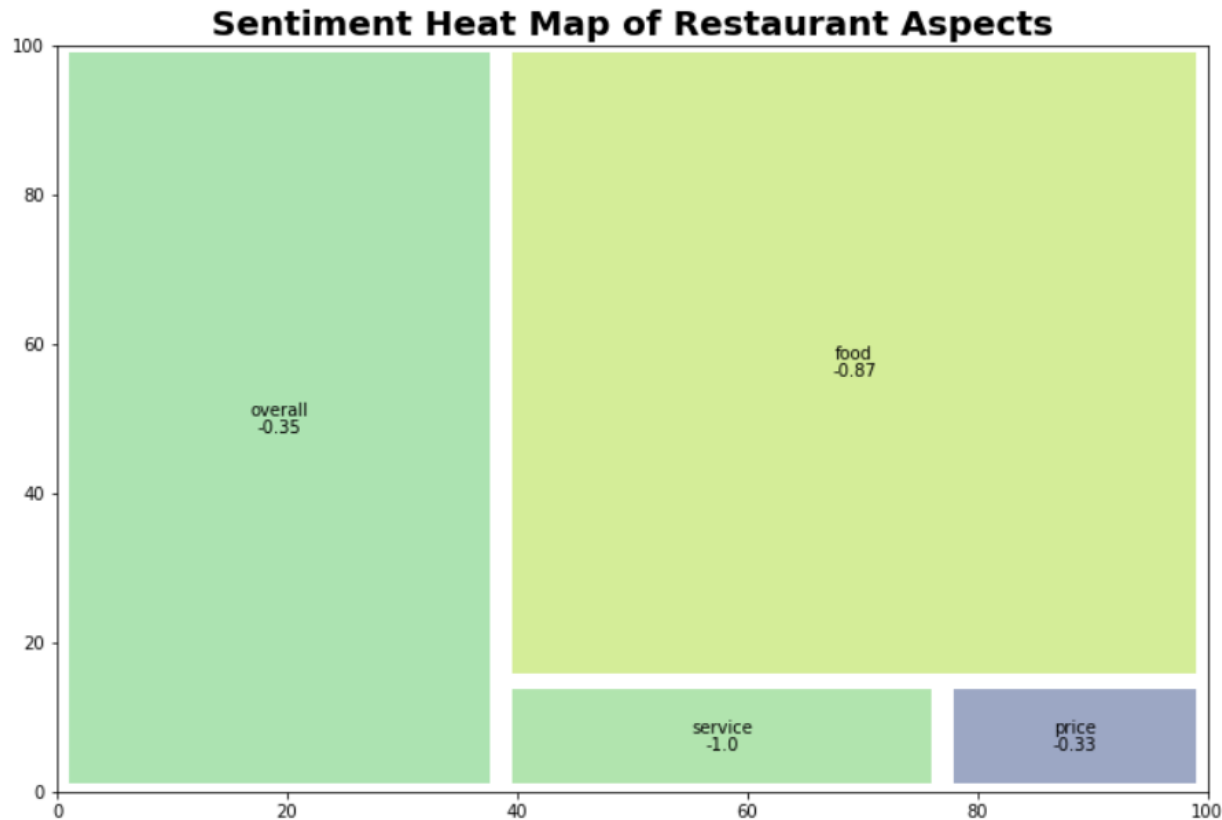
- Plotting sentiment and buzz. (note in this case, the buzz is simulated data)





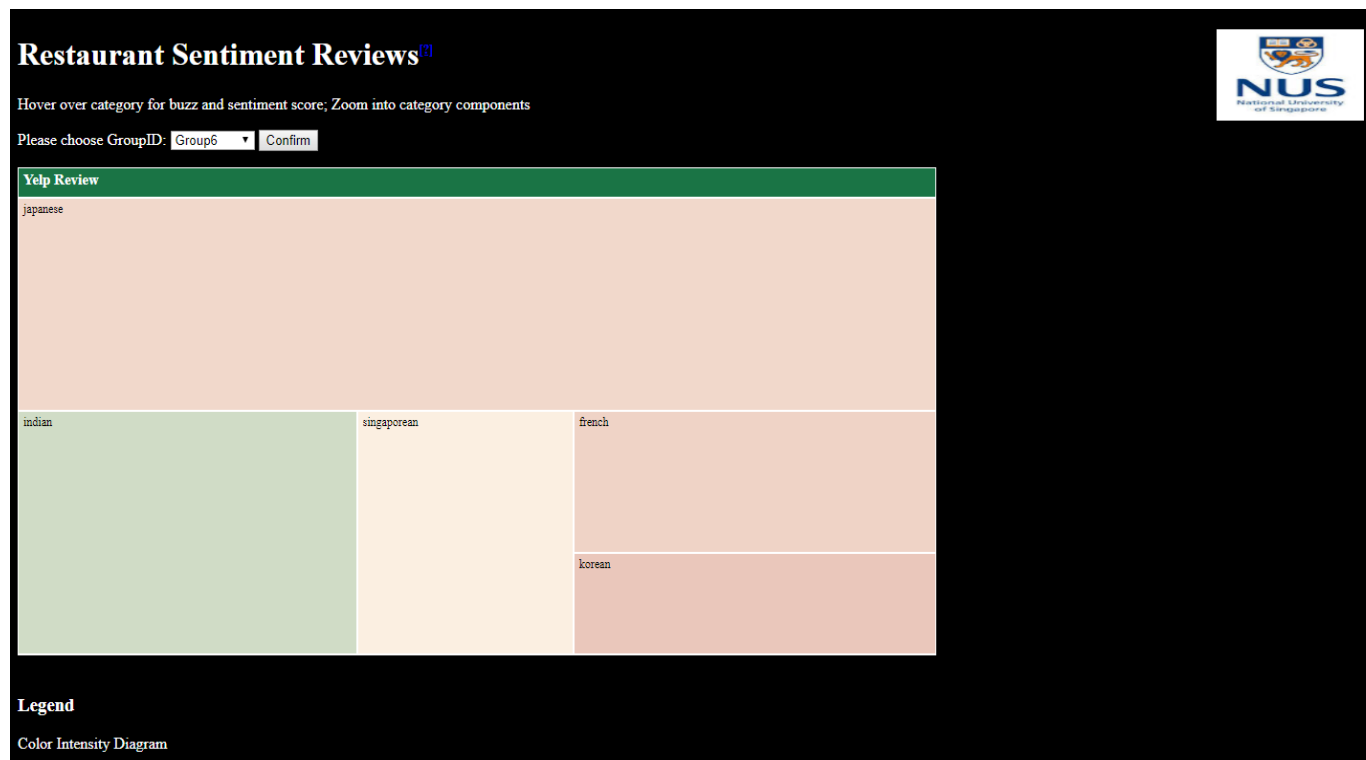
# Sentiment Heat-Map

- Allows view of most commonly talked about topic (aspects) and the sentiment score. See example workbook 3.1



# Sentiment Heat-Map

- (info only) D3.js allows the zooming-in of the sentiment categories. D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of Scalable Vector Graphics, HTML5, and Cascading Style Sheets standards. It is the successor to the earlier Protovis framework.



# Take-aways

- The sentiment aggregation can be done based on the sentiment definition of (o, a, t, t, s).
- How it is aggregated and visualised depends very much on business needs.
- Some examples of visuals that are useful for sentiment analysis include:
  - Histogram, boxplots, 2d density graphs, time series, word clouds and heatmaps

# Sentiment as an emotion

- Sentiment is an emotion
- Ekman : 6 basic emotions
- Valence: positive or negative affectivity
- Arousal:
  - intensity of the emotion, normally generates an action

These are often used in industry products.

# Questions I

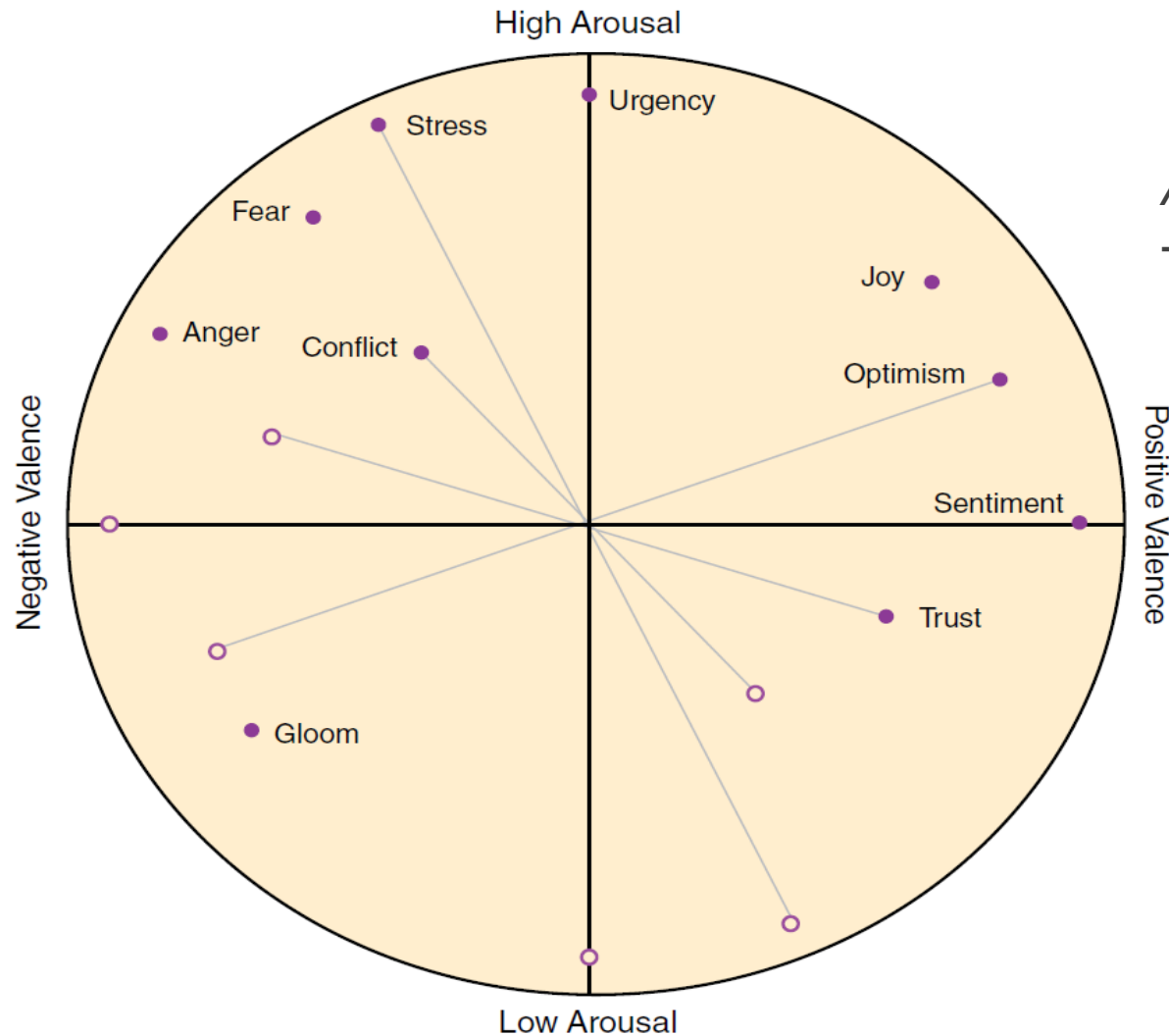
- Which emotions are positive and negative (valence)?

---

---

---

# The effects of emotion



*Affective circumplex*  
- how emotions vary

# Other sentiment ‘metrics’

- Buzz:
  - no of comments on the product/ event
  - Or no of reach of the posts (Google trends/ Google analytics)
- Proxy for ‘attention’:
  - Also a behavioural phenomenon – attention in both consumer marketing and in finance






*What information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.*

*“Designing Organizations for an Information-Rich World”*

Herbert Simon, Nobel Laureate in Economics

- Website that shows the number of users querying particular word or topic.
- <https://trends.google.com>

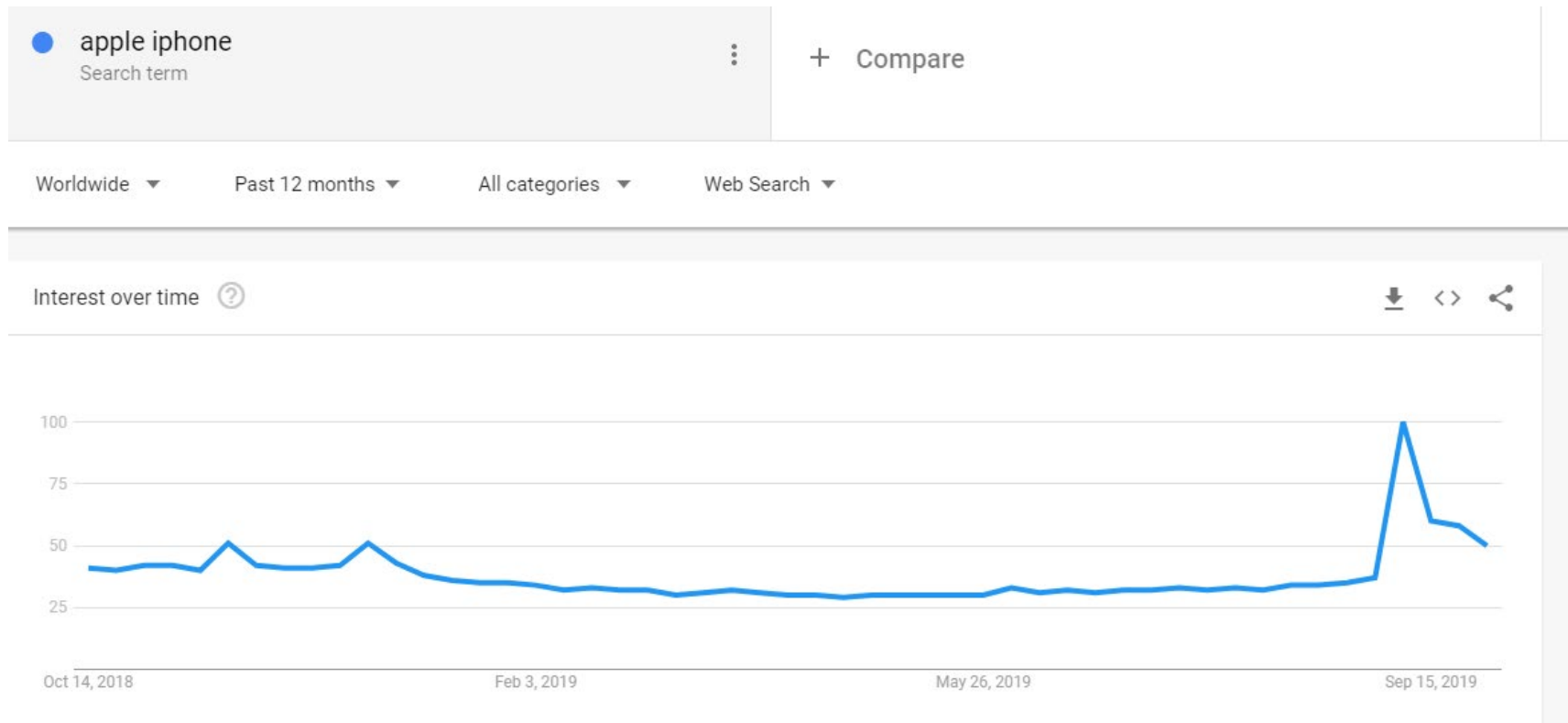
Allows segmentation by country. For eg. what is the most searched stock in Singapore? Related to investors' attention.

Related queries 		Rising    
1	sheng siong share price	Breakout
2	adsk	Breakout
3	nintendo stocks	+500%
4	singapore fixed deposit promotion	+500%
5	intc stock	+400%



# Google trends

- Apple iPhone search term. Notice the attention given to it during the launch of the iPhone 11 Pro



- Measured in 2 metrics
  - Duration
  - Intensity
- Consumers/ markets can only be ‘attentive’ to 1 thing at a time. Use attention to your benefit.
- Thence? Create a positive sentiment – ‘valence’ and ‘arousal’ in marketing materials to attract attention.
  - Some words generally have a higher arousal or valence. Give egs?

- ***Prospect Theory***
  - Nobel prize subject and cornerstone of marketing
  - Decision making is a two-staged process.
    - The first is the **scanning of decision outcomes and variables.**
    - The second is how do these variables and outcomes affect **one's relative welfare?**

Prospect theory is a psychology theory that describes how people make decisions when presented with alternatives that involve risk, probability, and uncertainty. It holds that people make decisions based on perceived losses or gains.

# Question III

- In the context of Prospect Theory, how does sentiment analysis fit in? Which stage does sentiment shed light on?
- 
-

# Decision-Making Process II

- *Affects 1<sup>st</sup> stage.* Sentiment analysis of text sheds light on each individual's assessment of products and services.
- However, the 2<sup>nd</sup> stage of the Prospect Theory remains 'unknown'. How do investors or consumers actually weigh their personal welfare?
- In layman terms, talk is cheap. Do you dare to put your money where your mouth is?

# Second stage of Prospect Theory

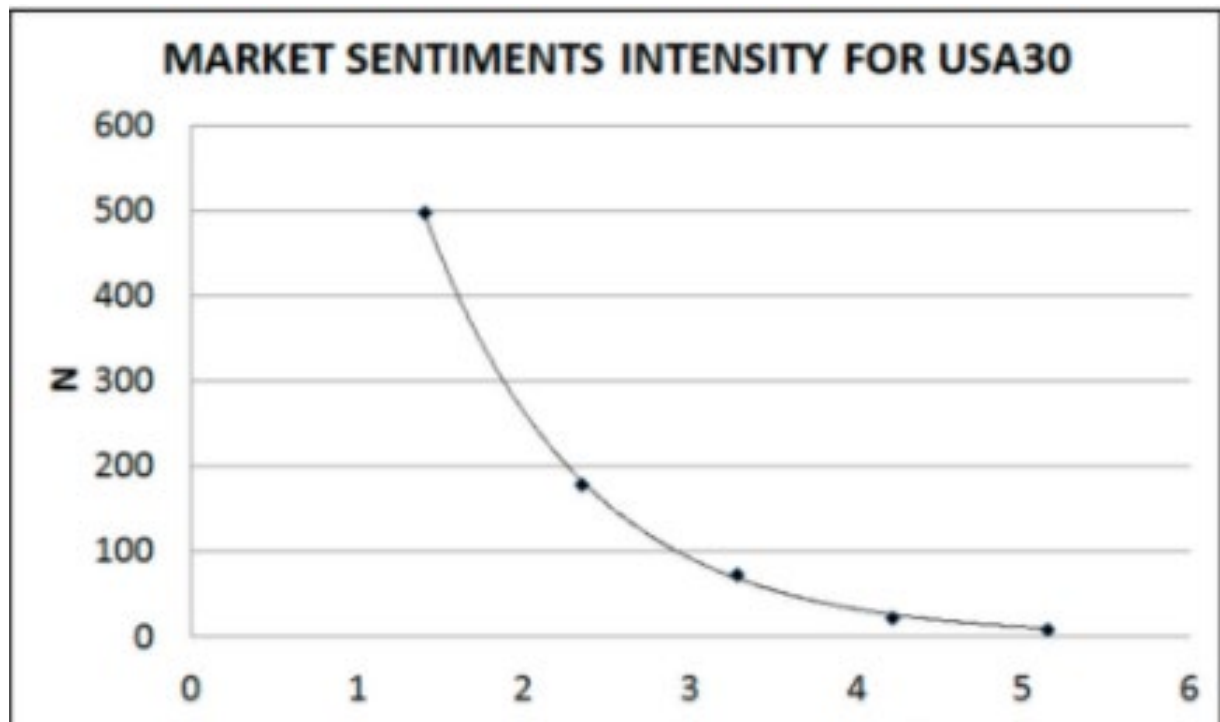
- This is also one particular reason why sentiment works or doesn't work in different scenarios.
- The influx of Big Data social media allows unprecedented 'textual' analysis on the 1st stage process of the public perception.
- What about the second stage?
- Thence sentiment analysis is used often with other data – for eg. customer data/ demographics (customer 360)

# Second stage of Prospect Theory

- Donald Trump's election poll shows him to be lagging, yet he won the election eventually.
- It is 'easy' to write reviews and do an official poll but in actual voting, people's actions may change. How about the cross-section of the polling public?
- The same for stocks trading. Even if a stock is popular and people talk about it doesn't mean the investors buy the stocks.

# Other 'Sentiment Behaviour'

- **Sentiment decay:**
  - in an exponential manner (has a half life, if no further 'shocks')
  - emotions almost never remain high





# Pre-built open source libraries

- TextBlob & Vader
  - No need for pre-training; fast prototyping;
  - But not customised training sets

- TextBlob:

<https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>

- Vader:

<http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>

In the lab session, we will also build a custom sentiment engine with Spacy.

- The Valence Aware Dictionary for sentiment Reasoning (VADER) is based on a list of lexical features and its associated sentiment intensity measures.
- These features are then combined with 5 general rules for expressing sentiment:
  - Certain punctuation, “!”, increases the magnitude of intensity without modifying the semantic orientation
  - Capitalization also increases magnitude
  - Certain intensifiers or degree adverbs can impact the sentiment intensity negatively or positively
  - Contrastive word, “but”, signals a shift in sentiment polarity, with the sentiment of the following words being more dominant

Good for social media eg twitter

- Returns polarity and subjectivity
- Sentiment polarity:  $[-1, 1]$
- List of words to determine subjectivity and polarity

<https://github.com/sloria/TextBlob/blob/dev/textblob/en/en-sentiment.xml>

- Considers also a lexicon approach by weighing words to calculate both polarity and subjectivity.

# Bootstrapping of datasets for sentiment training

Bootstrapping is any test or metric that uses random sampling with replacement, and falls under the broader class of resampling methods.

- Other training sets:
  - use labelled results with high probability ( $>0.7$ ) of being positive or negative.
- Assign as training sets
- Pros:
  - Enlarged training set with more features
- Cons:
  - Transferability? Training in one domain may not be applicable to other domains
  - What issues can you think of?

# Why do you want to build your own sentiment engine?

- Before training the sentiment algorithm classifiers, the
- training corpora is KEY.
- Experience shows it is more important than the algorithms itself.  
Garbage in garbage out is very true.
- **Training corpora** (body) needs to be as similar as possible to the test set. Needs to continually update this training corpora.
- **Training algorithms** covered on another day.

- Familiarise with visualisation packages useful for sentiment analysis
- Understand which sentiment emotion is more important for different cases.
- Familiarise with prototyping libraries for sentiment libraries

# References:

- Tham, Eric, Prospect Theory. The Unbearable Lightness of Expectations of the Chinese Investor (November 5, 2015). Handbook of Sentiment Analysis in Finance (2015) . Available at SSRN: <https://ssrn.com/abstract=3168041>
- 5 Ways To Grab Your Customer's Attention In A Distracted World  
<https://www.braze.com/blog/5-ways-grab-customers-attention-distracted-world/>
- The Rising Cost of Consumer Attention: Why You Should Care, and What You Can Do about It  
[http://www.hbs.edu/faculty/publication%20files/14-055\\_2ef21e7e-7529-4864-b0f0-c64e4169e17f.pdf](http://www.hbs.edu/faculty/publication%20files/14-055_2ef21e7e-7529-4864-b0f0-c64e4169e17f.pdf)
- In Search of Attention  
<https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2011.01679.x>