NATIONAL CHENGCHI UNIVERSITY

Tamkang University
淡江大學

Social Media and Opinion Mining
(社群媒體與意見探勘)

時間：2016/10/25 (二) (2:10-5:00pm)
地點：政治大學綜合院館270407，北棟407教室
主持人：陳恭 主任

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/

Outline

- Social Media
  - Social Media Marketing Analytics (社群媒體行銷分析)
- Opinion Mining
  - Text Mining and Analytics Technology (文字探勘分析技術)

Tamkang University
淡江大學

Social Media Marketing Analytics
(社群媒體行銷分析)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail. tku.edu.tw/myday/

2016-07

Outline

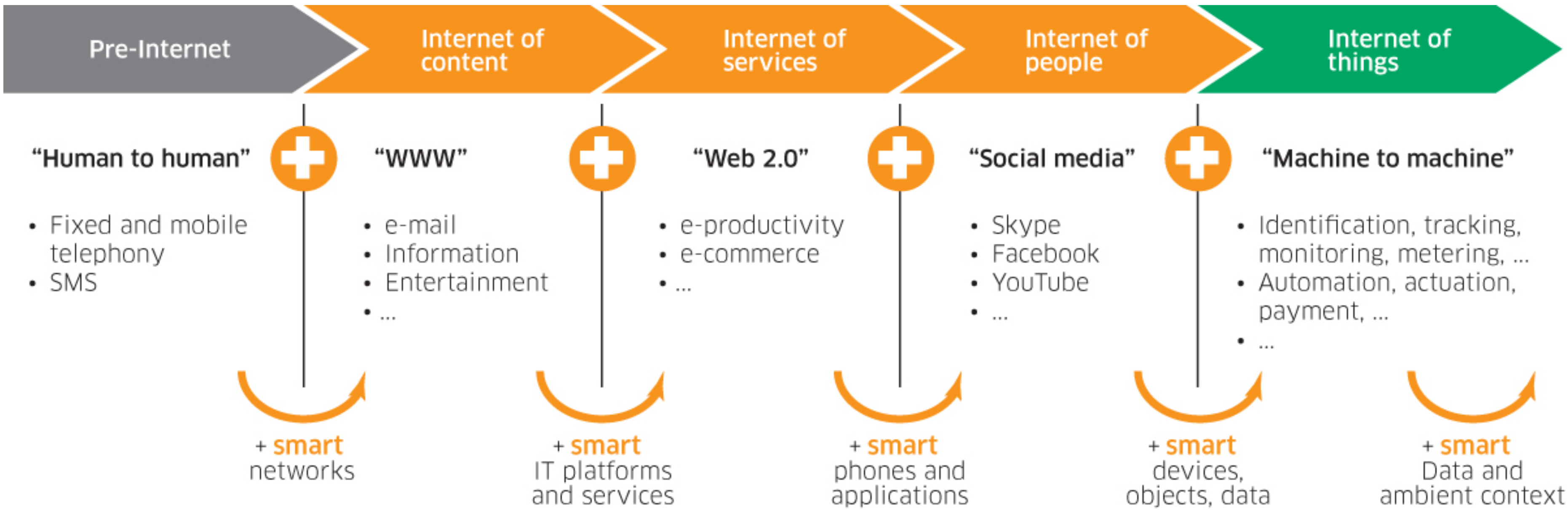
- Consumer Psychology and Behavior on Social Media
- Social Media Marketing Analytics
  - Social Media Listening
  - Search Analytics
  - Content Analytics
  - Engagement Analytics
- Social Analytics Lifecycle

Social Media

Source: http://hungrywolfmarketing.com/2013/09/09/what-are-your-social-marketing-goals/

Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine



Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015
http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/

Emotions

Love

Joy

Surprise

Anger

Sadness

Fear

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. … "

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too **expensive**, and wanted me to return it to the shop. … "

+Positive Opinion

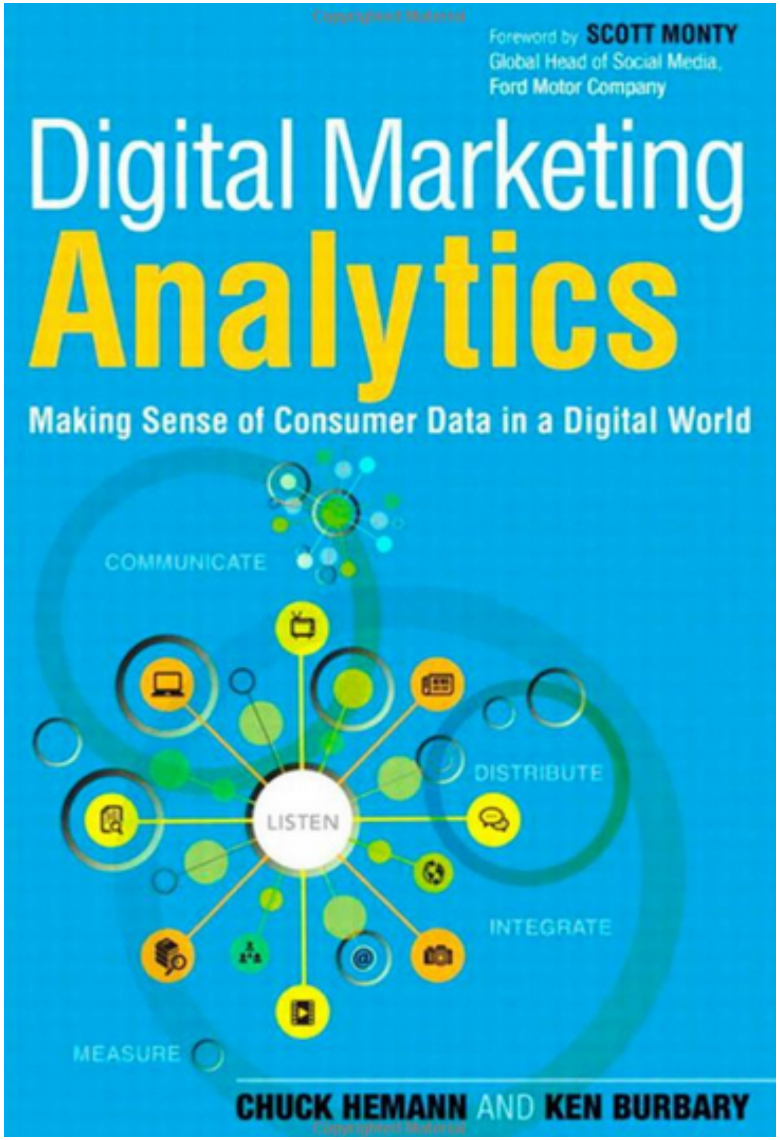
-Negative Opinion

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Social Media Marketing Analytics

Digital Marketing Analytics:

Making Sense of Consumer Data in a Digital World, Chuck Hemann and Ken Burbary, Que. 2013



Source: http://www.amazon.com/Digital-Marketing-Analytics-Consumer-Biz-Tech/dp/0789750309

Consumer Psychology and Behavior on Social Media

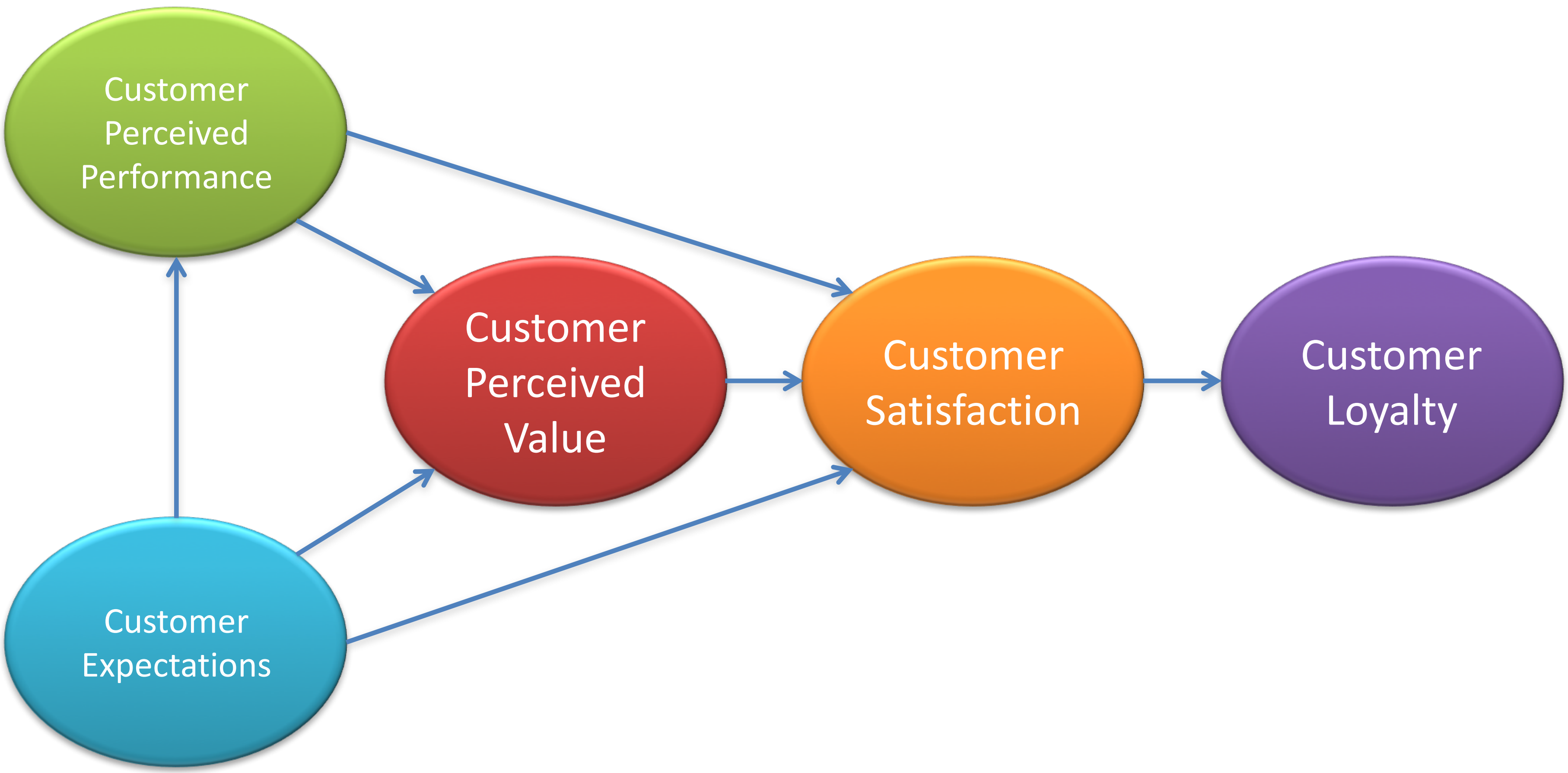
How consumers think, feel, and act

Analyzing Consumer Markets

- The aim of marketing is to meet and satisfy target customers' needs and wants better than competitors.
- Marketers must have a thorough understanding of
  how consumers think, feel, and act
  and offer clear value to each and every target consumer.

Source: Philip Kotler & Kevin Lane Keller, Marketing Management, 14th ed., Pearson, 2012

Customer Perceived Value, Customer Satisfaction, and Loyalty



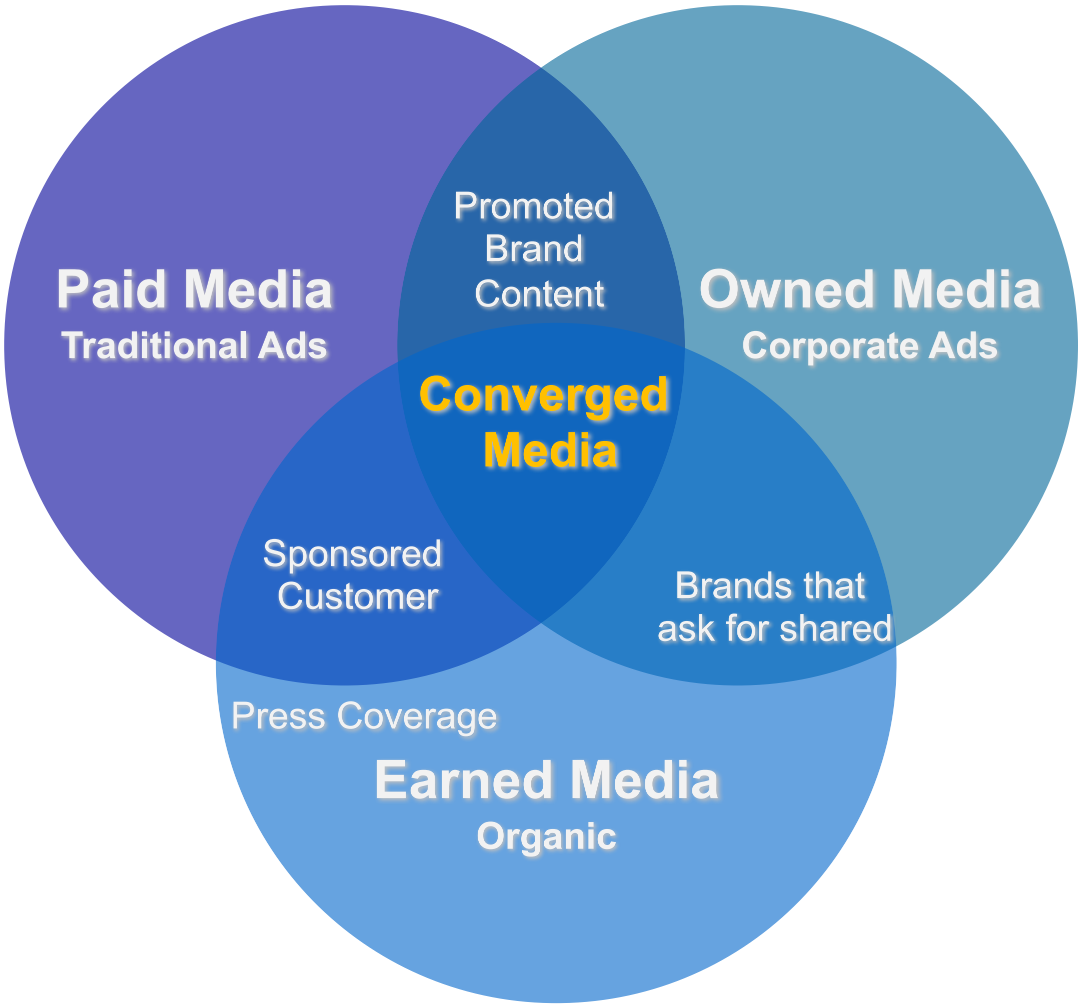
Source: Philip Kotler & Kevin Lane Keller, Marketing Management, 14th ed., Pearson, 2012

Social Media Marketing Analytics

- Social Media Listening
- Search Analytics
- Content Analytics
- Engagement Analytics

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

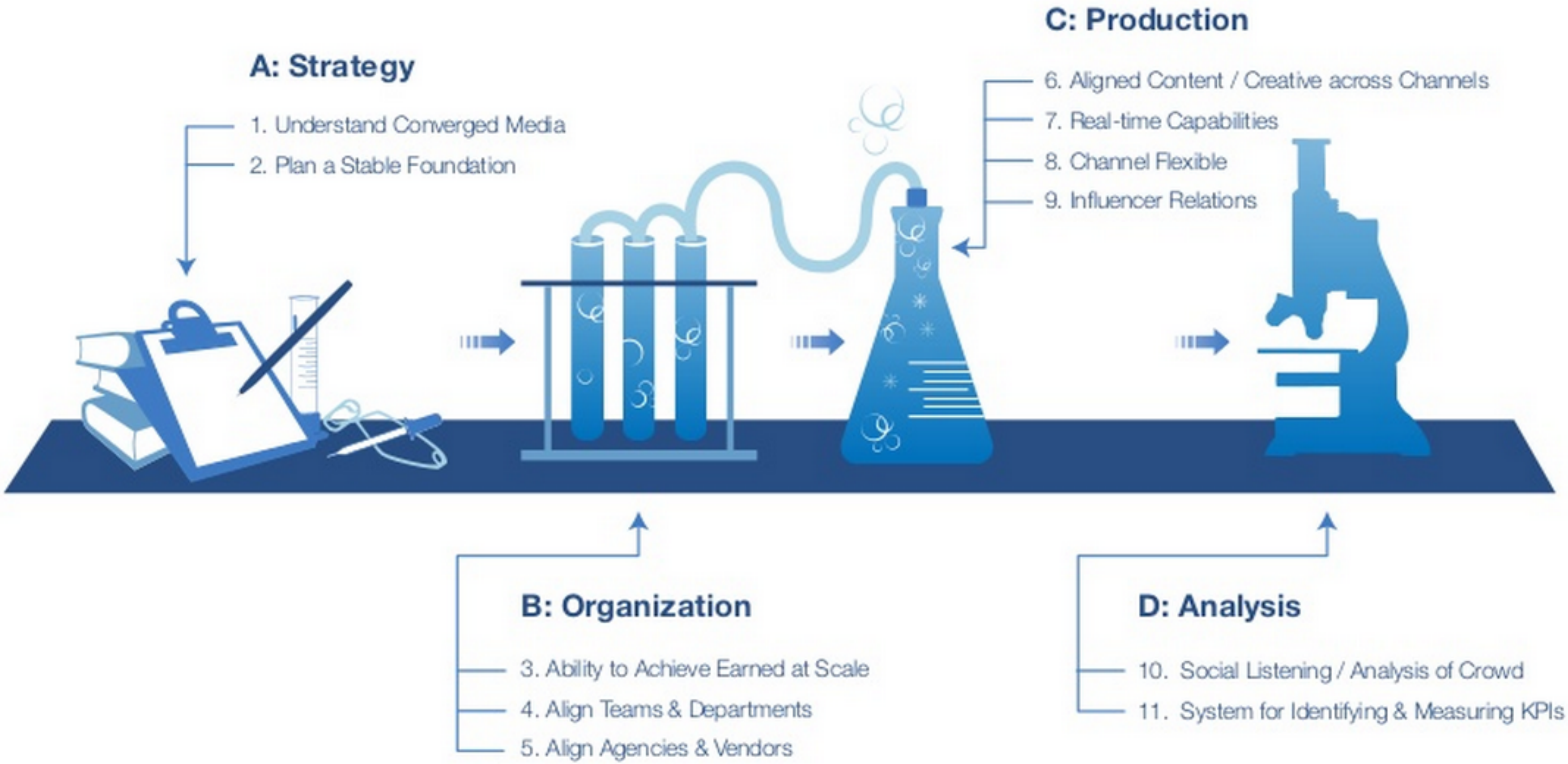
The Convergence of Paid, Owned & Earned Media



Source: "The Converged Media Imperative: How Brands Will Combine Paid, Owned and Earned Media", Altimeter Group, July 19, 2012)
http://www.altimetergroup.com/2012/07/the-converged-media-imperative/

Converged Media
Top 11 Success Criteria

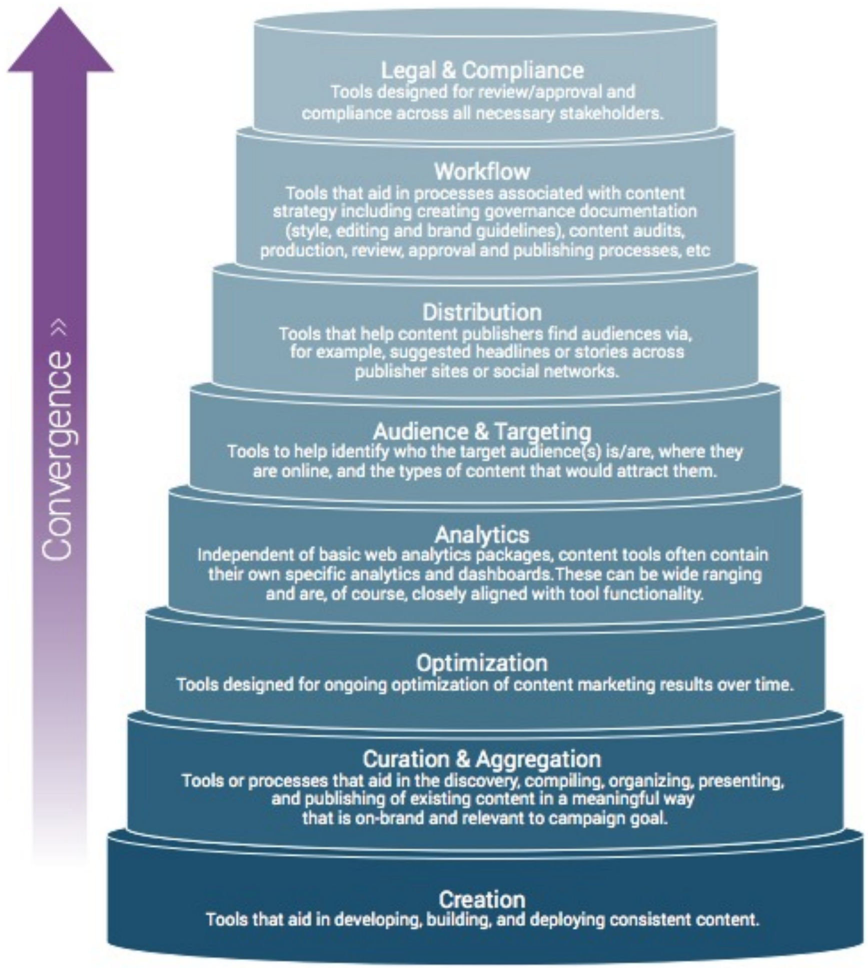
Social Listening / Analysis of Crowd



Source: "The Converged Media Imperative: How Brands Will Combine Paid, Owned and Earned Media", Altimeter Group, July 19, 2012)
http://www.altimetergroup.com/2012/07/the-converged-media-imperative/

Content Tool Stack Hierarchy

Figure 3 Content Tool Stack Hierarchy



Source: Altimeter Group

Source: Rebecca Lieb, "Content marketing in 2015 -- research, not predictions", December 16, 2014
http://www.imediaconnection.com/content/37909.asp

Competitive Intelligence

- Gather competitive intelligence data

Google Alexa Compete

- Which audience segments are competitors reaching that you are not?
- What keywords are successful for your competitors?
- What sources are driving traffic to your competitors' websites?

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Competitive Intelligence

- Facebook competitive analysis
- Facebook content analysis
- YouTube competitive analysis
- YouTube channel analysis
- Twitter profile analysis

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Web Analytics (Clickstream)

- Content Analytics
- Mobile Analytics

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Mobile Analytics

- Where is my mobile traffic coming from?
- What content are mobile users most interested in?
- How is my mobile app being used?
  What’s working?
  What isn’t?
- Which mobile platforms work best with my site?
- How does mobile user’s engagement with my site compare to traditional web users’ engagement?

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Identifying a Social Media Listening Tool

- Data Capture
- Spam Prevention
- Integration with Other Data Sources
- Cost
- Mobile Capability
- API Access
- Consistent User Interface
- Workflow Functionality
- Historical Data

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Search Analytics

- Free Tools for Collecting Insights Through
  – Search Data
  – Google Trends
  – YouTube Trends
  – The Google AdWords Keyword Tool
  – Yahoo! Clues
- Paid Tools for Collecting Insights Through Search Data
- The BrightEdge SEO Platform

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Owned Social Metrics

- Facebook page
- Twitter account
- YouTube channel

Own Social Media Metrics: Facebook

- Total likes
- Reach
  - Organic
  - Paid reach
  - Viral reach
- Engaged users
- People taking about this (PTAT)
- Likes, comments, and shares by post

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Own Social Media Metrics: Twitter

- Followers
- Retweets
- Replies
- Clicks and click-through rate (CTR)
- Impressions

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Own Social Media Metrics: YouTube

- Views
- Subscribers
- Likes/dislikes
- Comments
- Favorites
- Sharing

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Own Social Media Metrics: SlideShare

- Followers
- Views
- Comments
- Shares

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Own Social Media Metrics: Pinterest

- Followers
- Number of boards
- Number of pins
- Likes
- Repins
- Comments

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Own Social Media Metrics: Google+

- Number of people who have an account circled
- +1s
- Comments

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

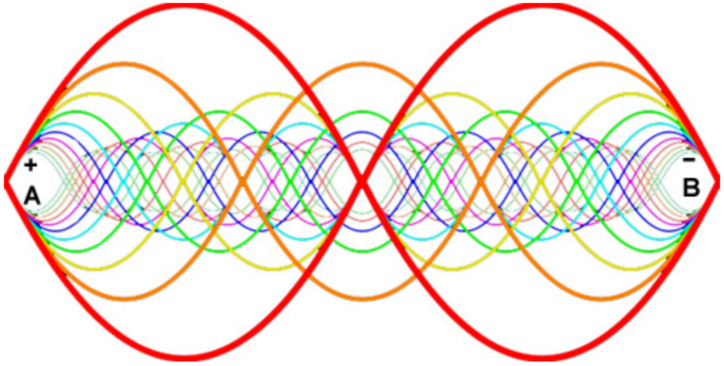
Earned Social Media Metrics

- Earned conversations
- In-network conversations

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Earned Social Media Metrics: Earned conversations

- Share of voice
- Share of conversation
- Sentiment
- Message resonance
- Overall conversation volume



Source: http://www.elvtd.com/elevation/p/beings-of-resonance

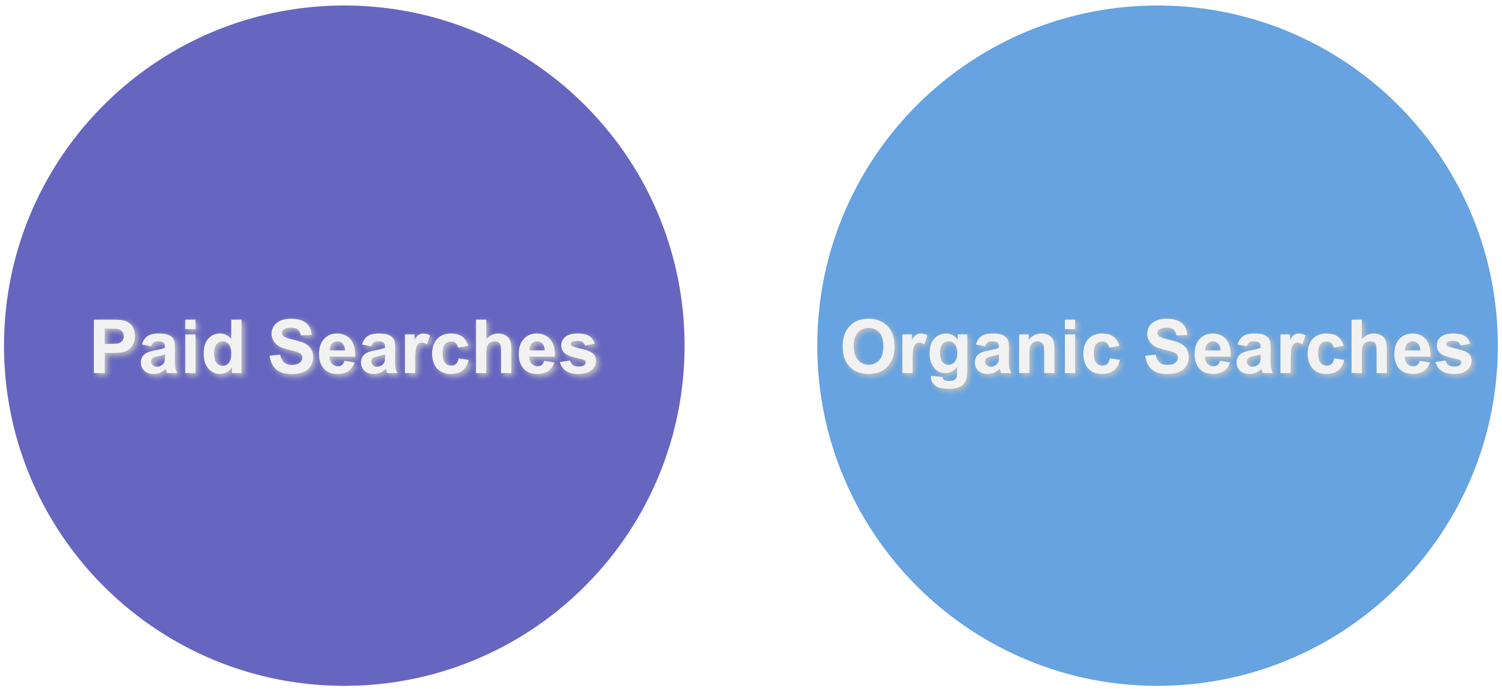
Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Demystifying Web Data

- Visits
- Unique page views
- Bounce rate
- Pages per visit
- Traffic sources
- Conversion

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Searching for the Right Metrics



Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Paid Searches

- Impressions
- Clicks
- Click-through rate (CTR)
- Cost per click (CPC)
- Impression share
- Sales or revenue per click
- Average position

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Organic Searches

- Known and unknown keywords
- Known and unknown branded keywords
- Total visits
- Total conversions from known keywords
- Average search position

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Aligning Digital and Traditional Analytics

- Primary Research
  - Brand reputation
  - Message resonance
  - Executive reputation
  - Advertising performance
- Traditional Media Monitoring
- Traditional CRM Data

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Media Listening Evolution

- Location of conversations
- Sentiment
- Key message penetration
- Key influencers

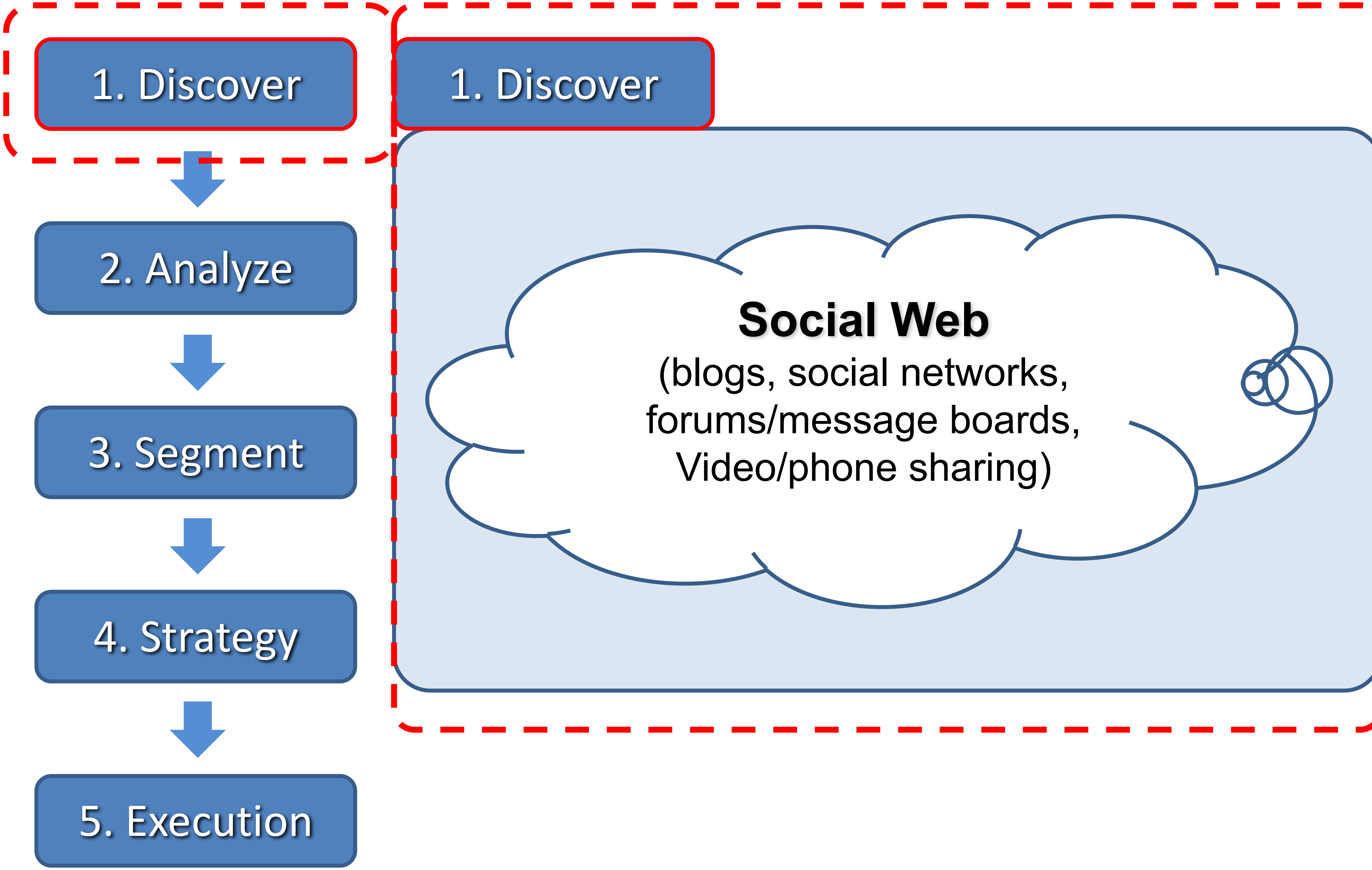
Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)



Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)

1. Discover

2. Analyze

3. Segment

4. Strategy

5. Execution

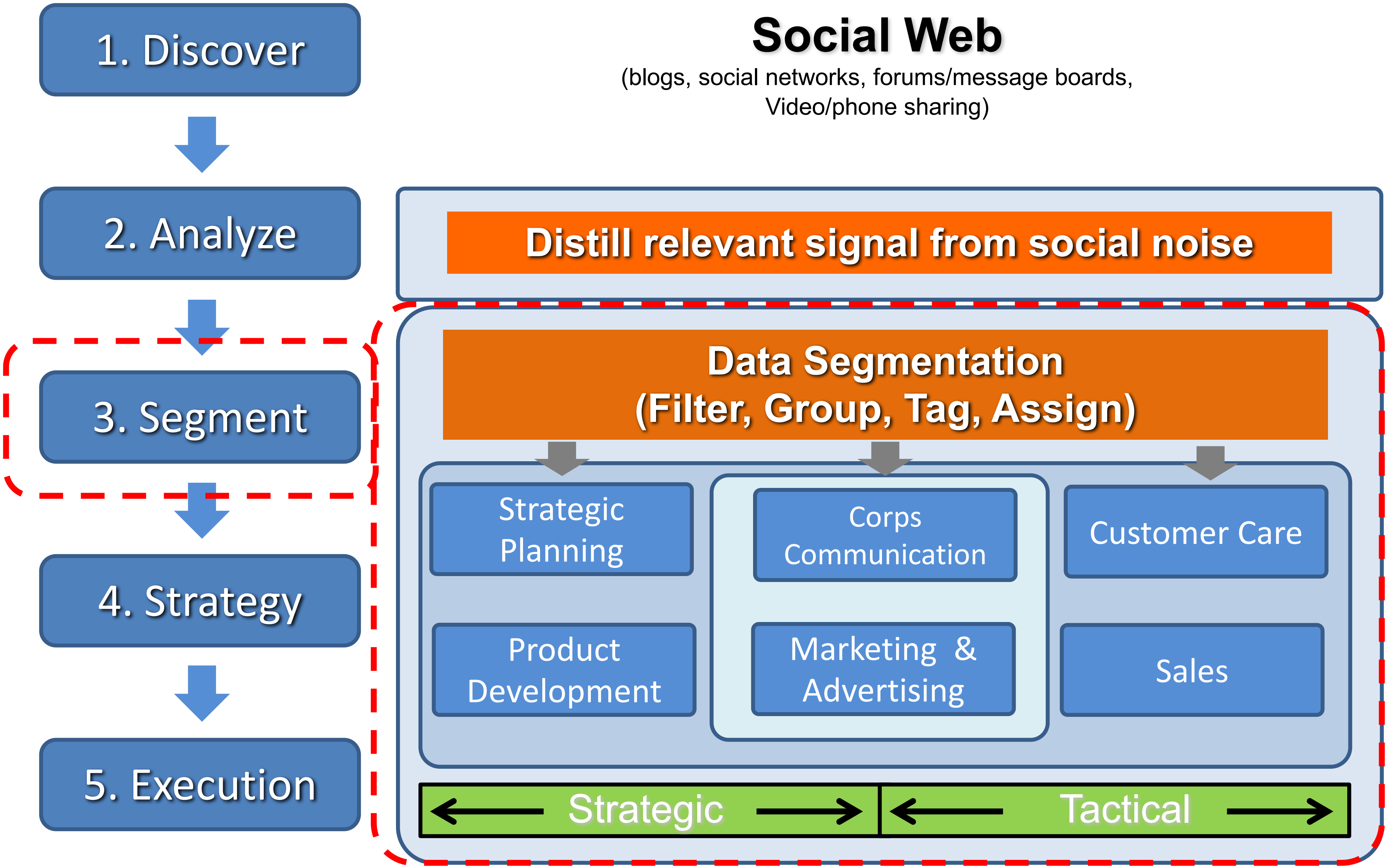
Social Web

(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)



Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web

(blogs, social networks, forums/message boards,
Video/phone sharing)

Distill relevant signal from social noise

Data Segmentation
(Filter, Group, Tag, Assign)

Insights drive focused business strategies

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web
(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Data Segmentation
(Filter, Group, Tag, Assign)

Insights drive focused business strategies

Innovation | Future Direction | Reputation Management | Campaigns | CRM | Customer Satisfaction Improvements

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

Social Analytics Lifecycle (5 Stages)

1. Discover
2. Analyze
3. Segment
4. Strategy
5. Execution

Social Web

(blogs, social networks, forums/message boards, Video/phone sharing)

Distill relevant signal from social noise

Data Segmentation (Filter, Group, Tag, Assign)

Insights drive focused business strategies

- Innovation
- Future Direction
- Reputation Management
- Campaigns
- CRM
- Customer Satisfaction Improvements

Source: Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013

How consumers think, feel, and act

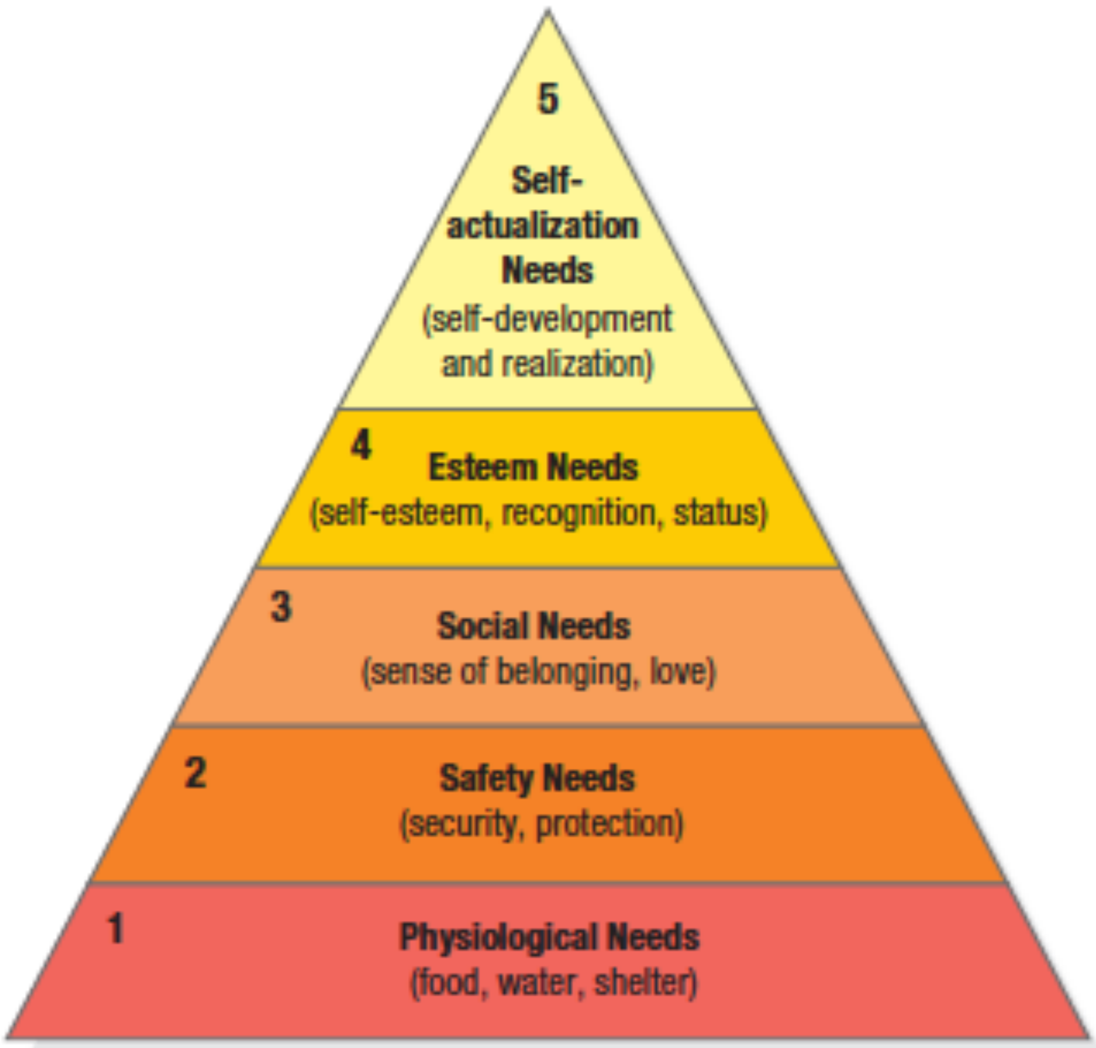
Emotions

Love	Anger
Joy	Sadness
Surprise	Fear

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Maslow's Hierarchy of Needs



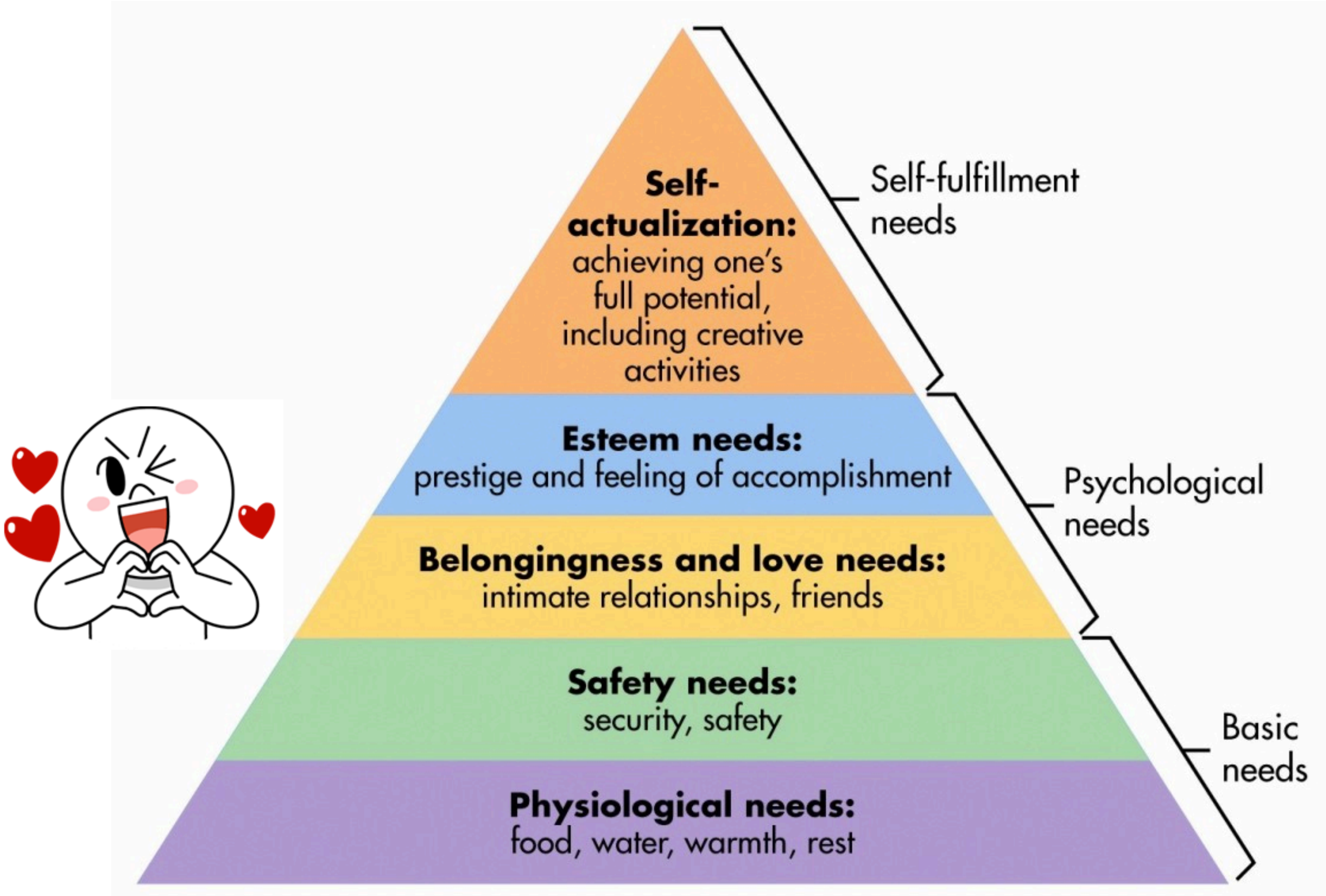
Source: Philip Kotler & Kevin Lane Keller, Marketing Management, 14th ed., Pearson, 2012

Maslow's hierarchy of human needs

(Maslow, 1943)

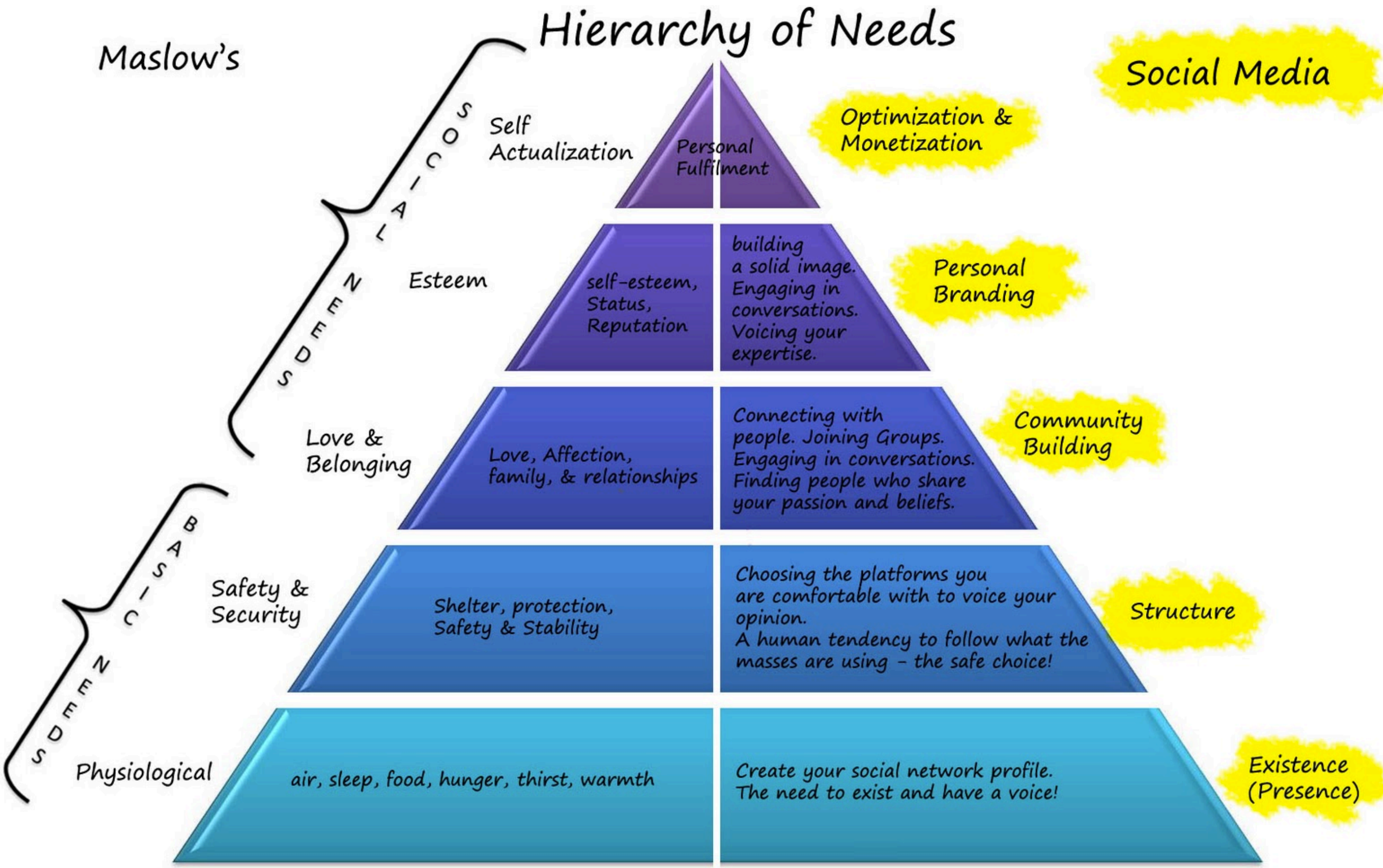
Source: Backer & Saren (2009), Marketing Theory: A Student Text, 2nd Edition, Sage

Maslow’s Hierarchy of Needs



Source: http://sixstoriesup.com/social-psyche-what-makes-us-go-social/

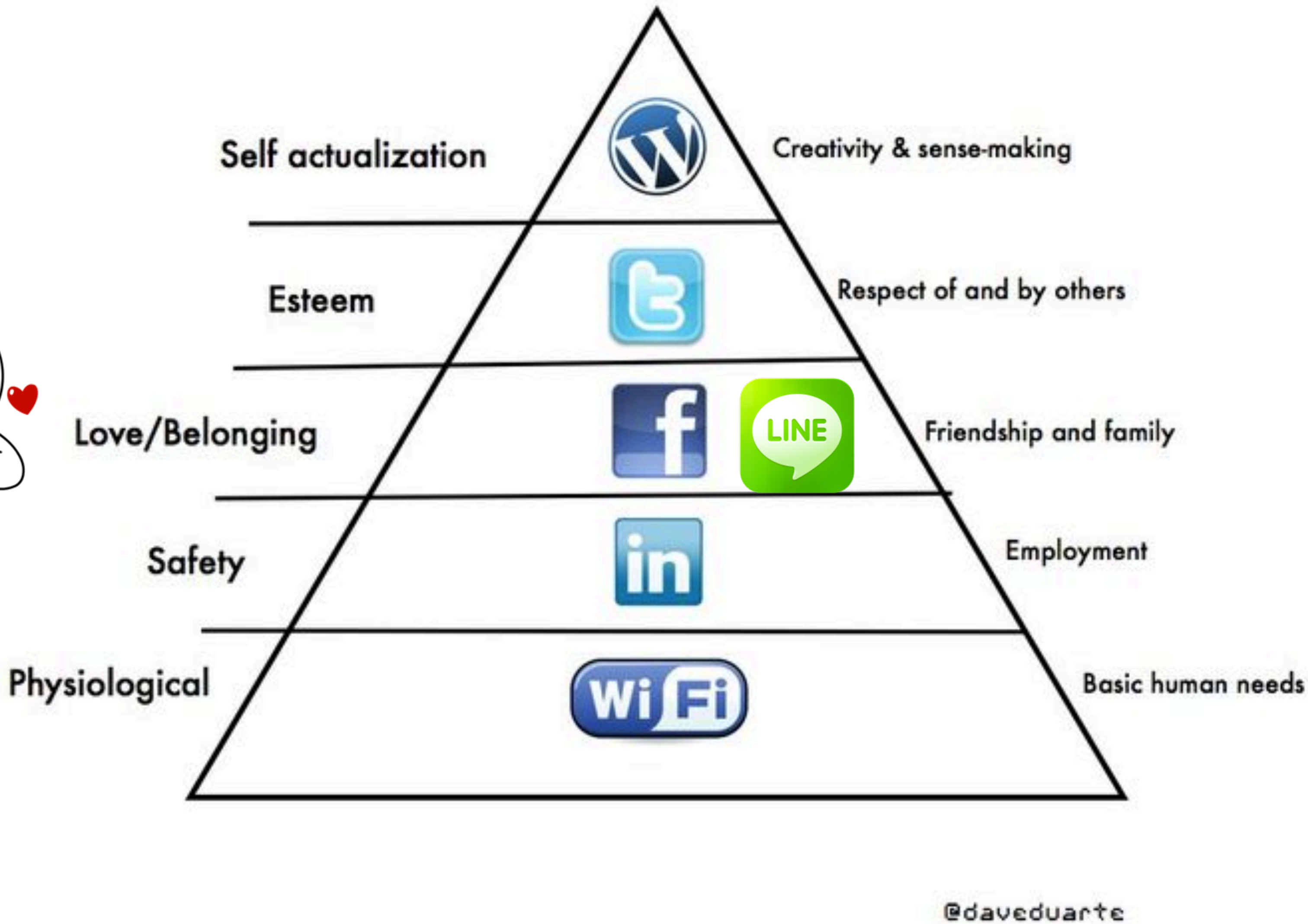
Social Media Hierarchy of Needs



Social Media Hierarchy of Needs - by John Antonios

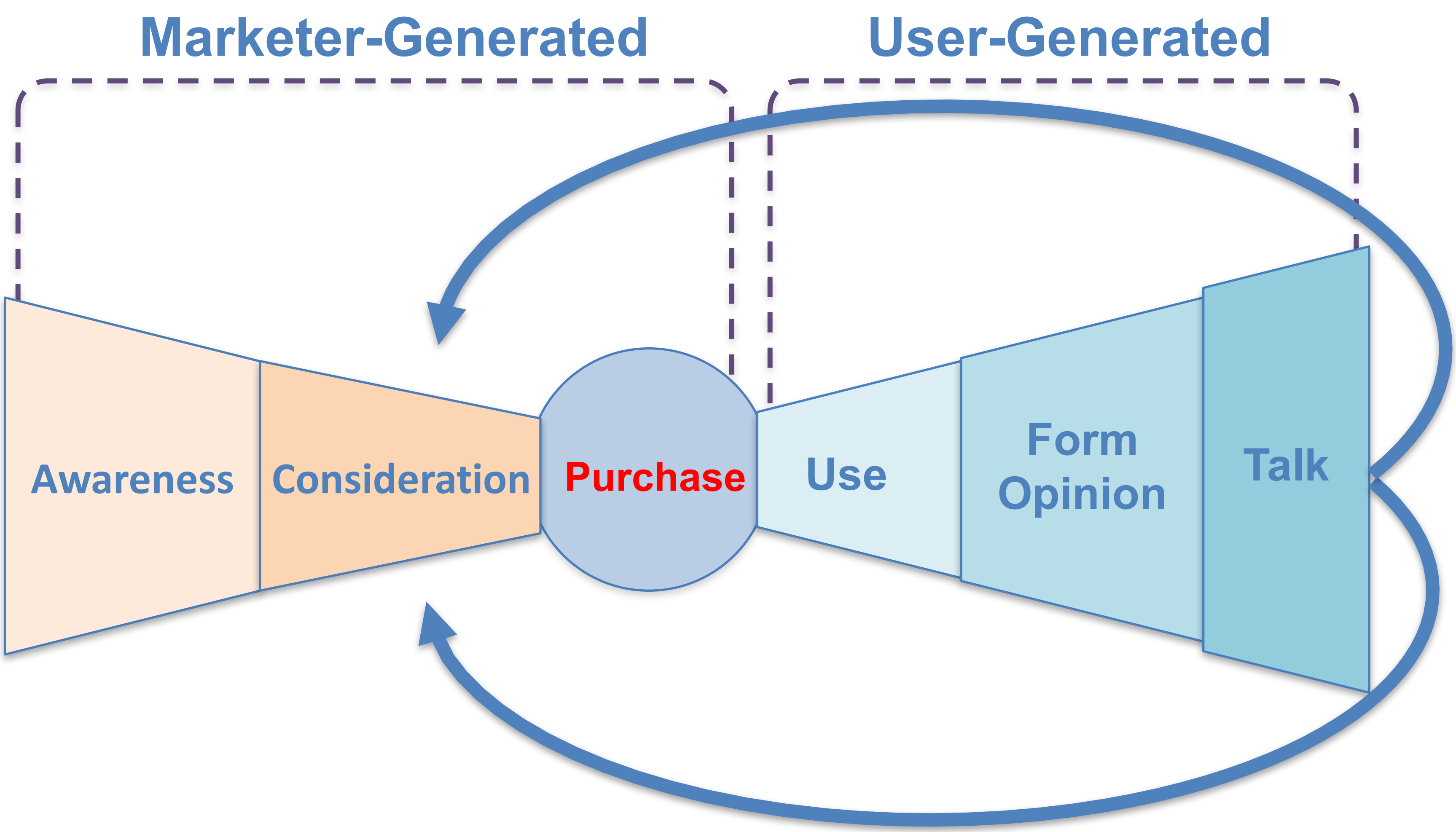
Source: http://2.bp.blogspot.com/_Rta1VZltiMk/TPavcanFtfI/AAAAAAAAACo/OBGnRL5arSU/s1600/social-media-heirarchy-of-needs1.jpg

Social Media Hierarchy of Needs



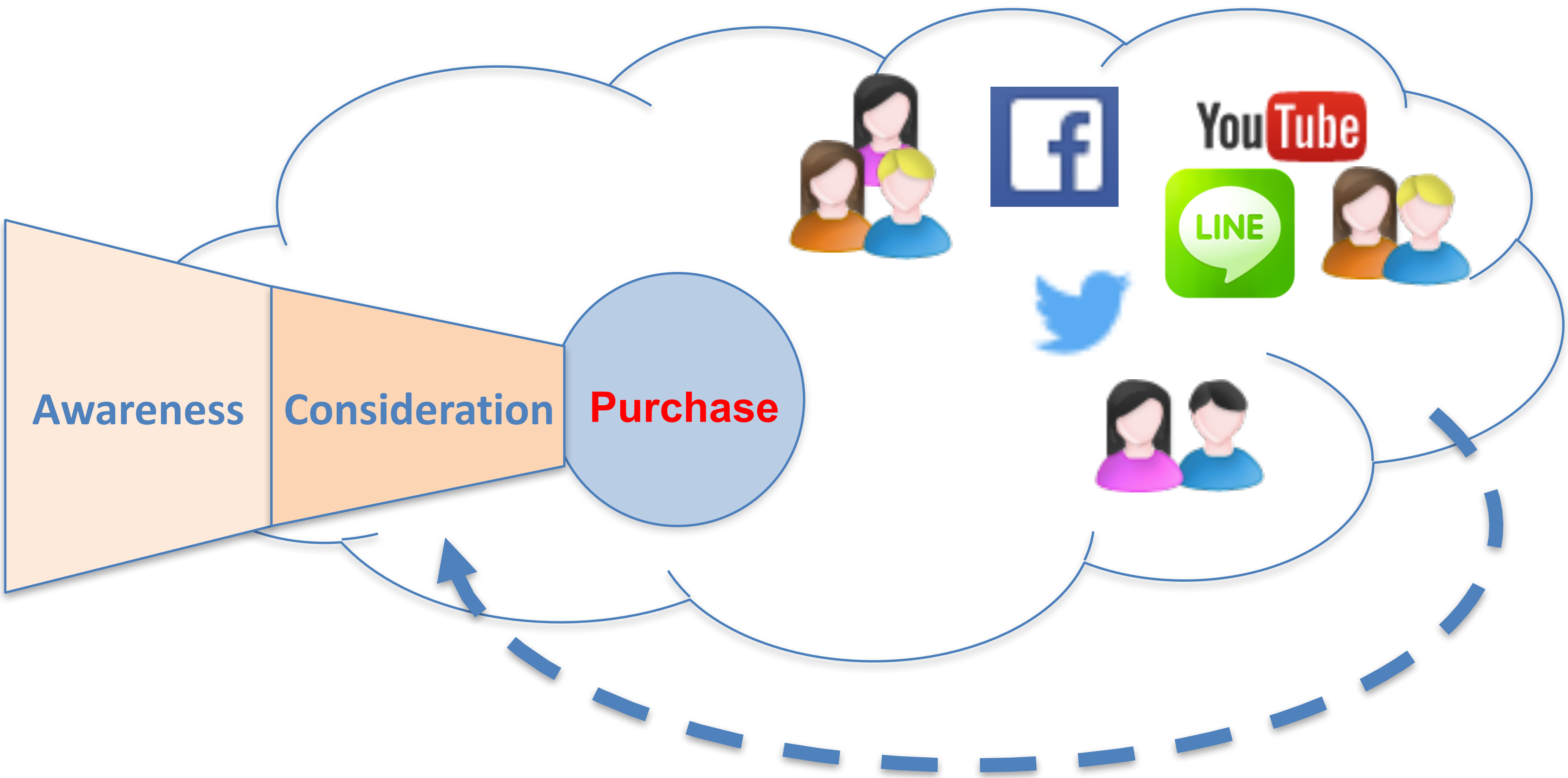
Source: http://www.pinterest.com/pin/186477859309035585/

The Social Feedback Cycle
Consumer Behavior on Social Media



Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement

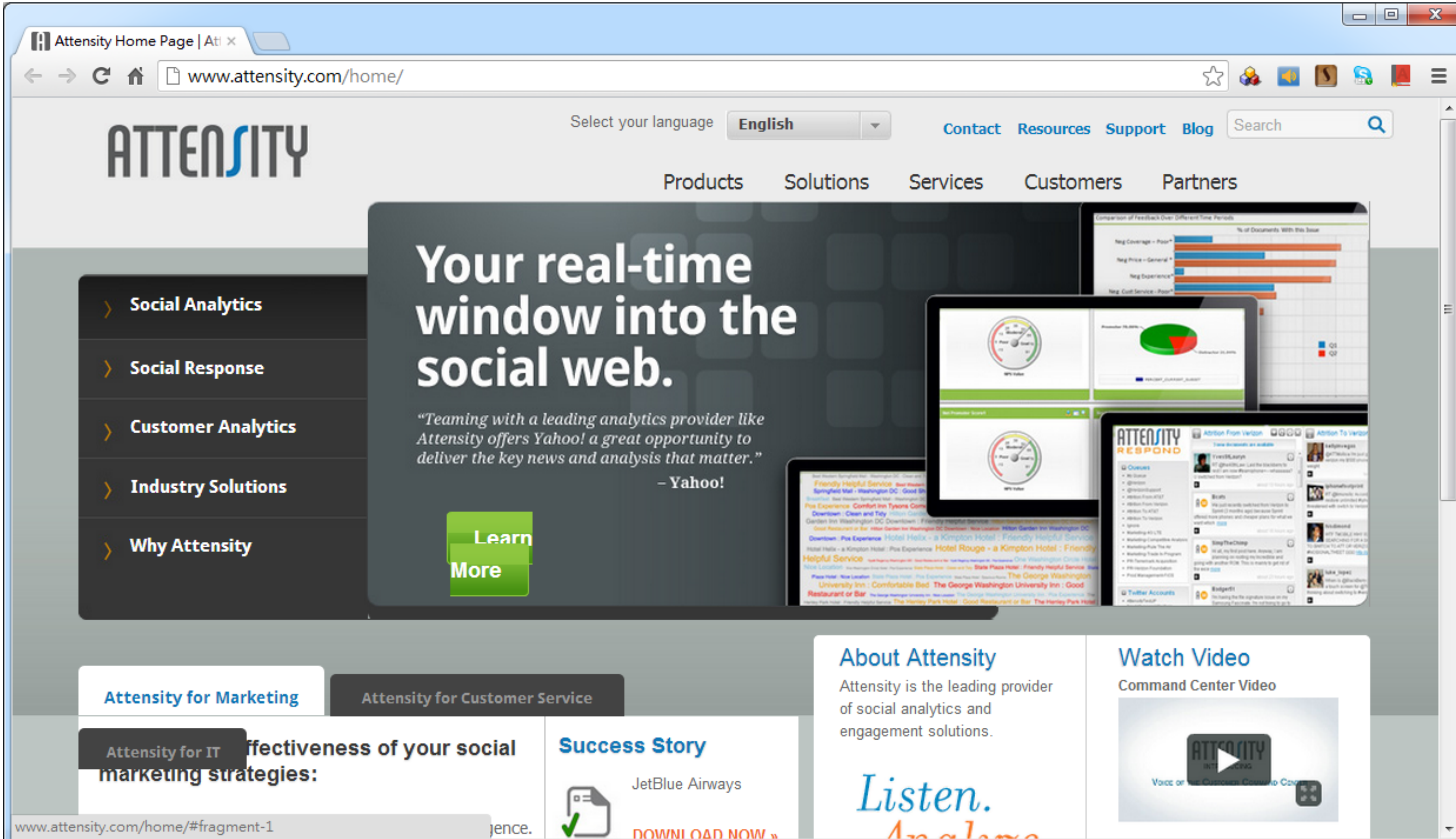
The New Customer Influence Path



Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement

Attensity: Track social sentiment across brands and competitors

http://www.attensity.com/



http://www.youtube.com/watch?v=4goxmBEg2Iw#!

Sentiment Analysis vs. Subjectivity Analysis

Sentiment Analysis	Subjectivity Analysis
Positive Negative	Subjective
Neutral	Objective

Example of SentiWordNet

POS ID PosScore NegScore SynsetTerms Gloss

a 00217728 0.75 0 beautiful#1 delighting the senses or exciting intellectual or emotional admiration; "a beautiful child"; "beautiful country"; "a beautiful painting"; "a beautiful theory"; "a beautiful party“

a 00227507 0.75 0 best#1 (superlative of `good') having the most positive qualities; "the best film of the year"; "the best solution"; "the best time for planting"; "wore his best suit“

r 00042614 0 0.625 unhappily#2 sadly#1 in an unfortunate way; "sadly he died before he could see his grandchild“

r 00093270 0 0.875 woefully#1 sadly#3 lamentably#1 deplorably#1 in an unfortunate or deplorable manner; "he was sadly neglected"; "it was woefully inadequate“

r 00404501 0 0.25 sadly#2 with sadness; in a sad manner; "`She died last night,' he said sadly"

Summary

- Consumer Psychology and Behavior on Social Media
- Social Media Marketing Analytics
  - Social Media Listening
  - Search Analytics
  - Content Analytics
  - Engagement Analytics
- Social Analytics Lifecycle

References

- Chuck Hemann and Ken Burbary, Digital Marketing Analytics: Making Sense of Consumer Data in a Digital World, Que. 2013
- Dave Evans, Susan Bratton, and Jake McKee, Social Media Marketing: The Next Generation of Business Engagement, , Sybex, 2010
- Liana Evans, Social Media Marketing: Strategies for Engaging in Facebook, Twitter & Other Social Media, Que, 2010.
- Hiroshi Ishikawa, Social Big Data Mining Hardcover, CRC Press, 2015
- Data Science for Business: What you need to know about data mining and data-analytic thinking, Foster Provost and Tom Fawcett, O'Reilly, 2013

Tamkang University
淡江大學

Text Mining and Analytics Technology
(文字探勘分析技術)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/

2016-07

Outline

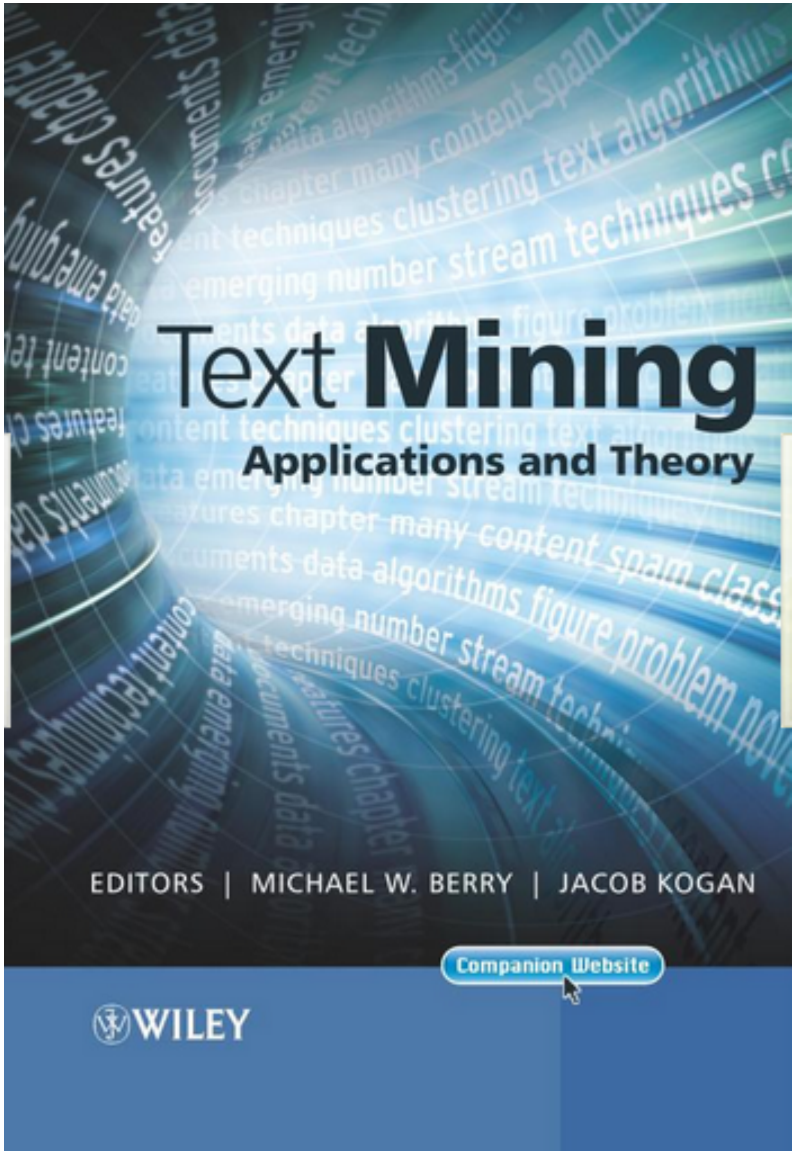
- Text Mining
  - Differentiate between
    text mining, Web mining and data mining
- Natural Language Processing (NLP)
- Text Mining Tools and Applications

Text Mining and Analytics Technology

Text Mining Techniques

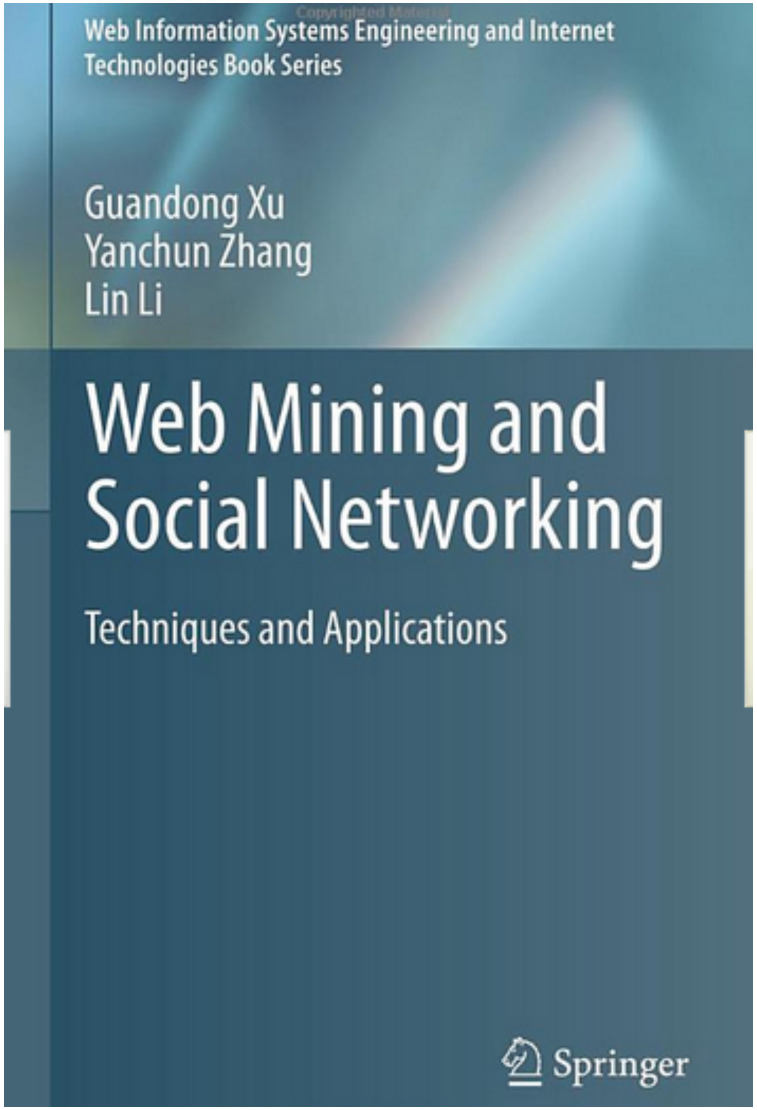
Natural Language Processing (NLP)

Text Mining



http://www.amazon.com/Text-Mining-Applications-Michael-Berry/dp/0470749822/

Web Mining and Social Networking

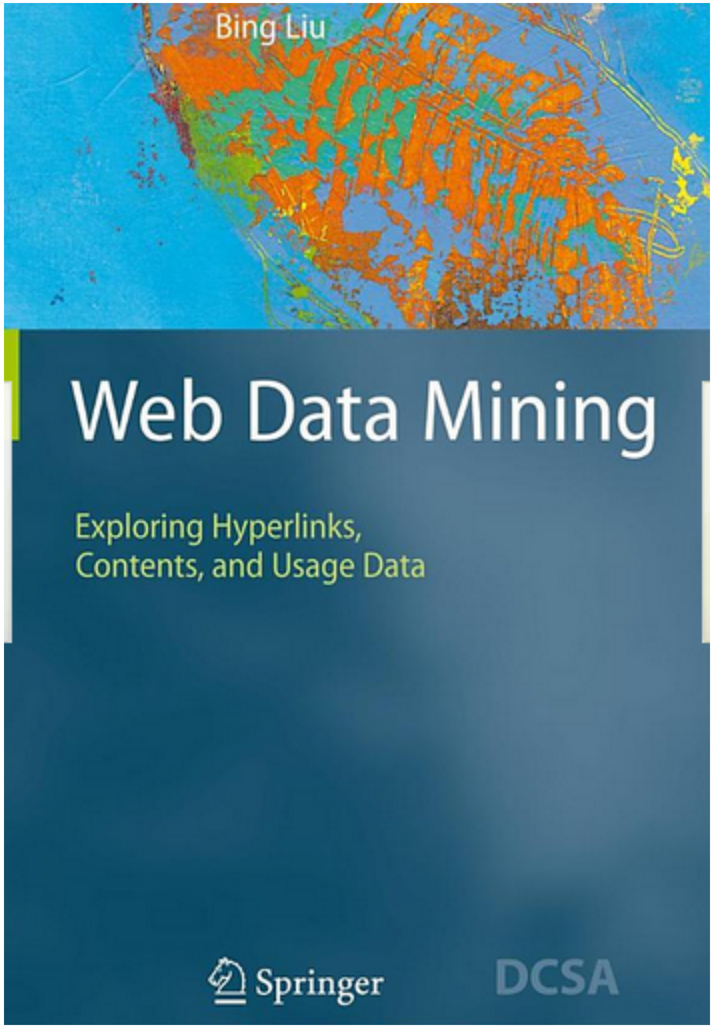


http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

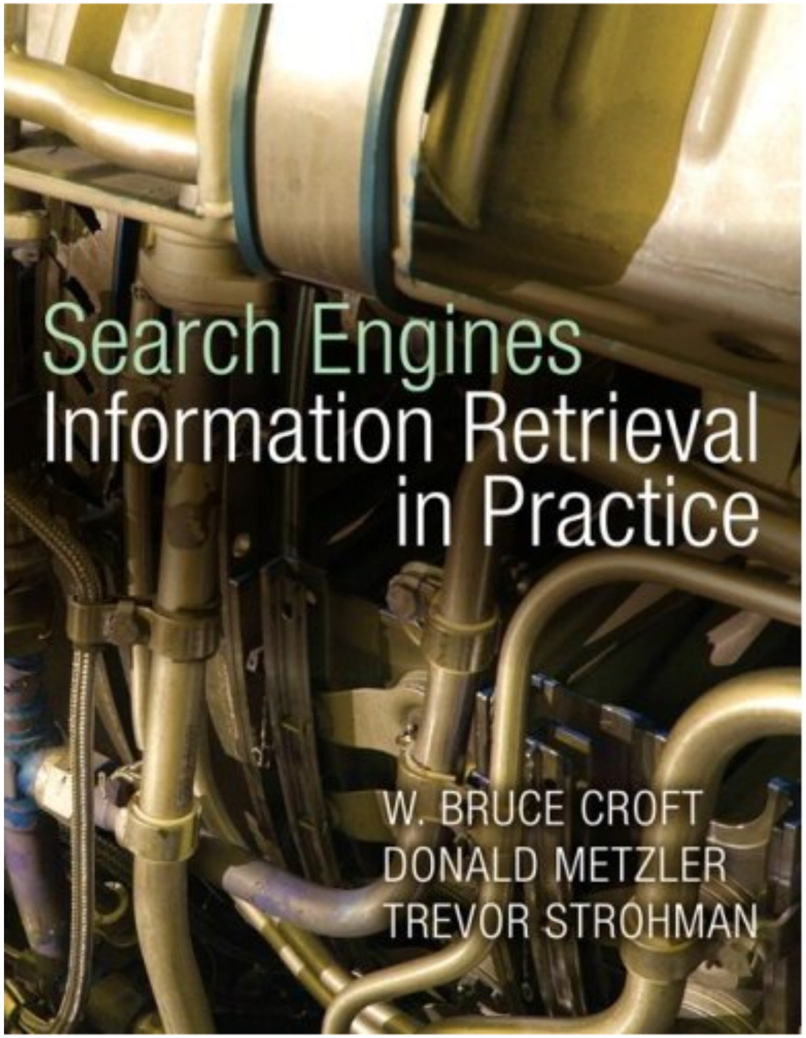
http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345

Web Data Mining:
Exploring Hyperlinks, Contents, and Usage Data



http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3540378812

Search Engines: Information Retrieval in Practice

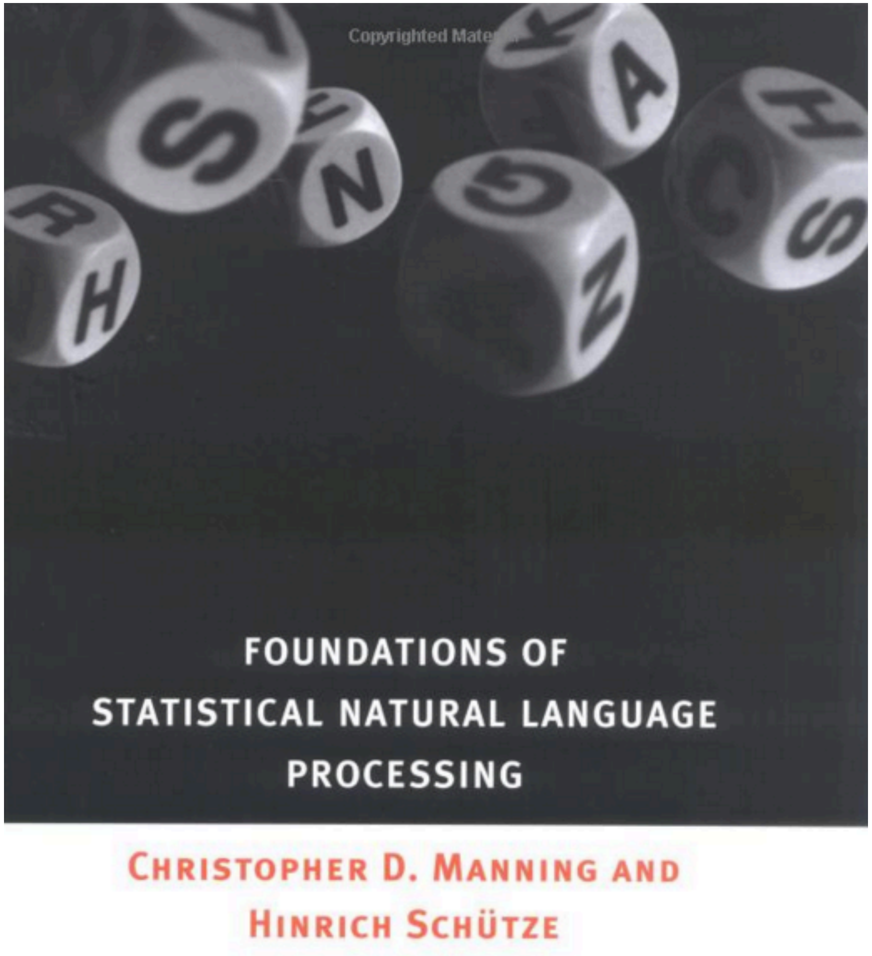


http://www.amazon.com/Search-Engines-Information-Retrieval-Practice/dp/0136072240

Christopher D. Manning and Hinrich Schütze (1999),

Foundations of Statistical Natural Language Processing,

The MIT Press

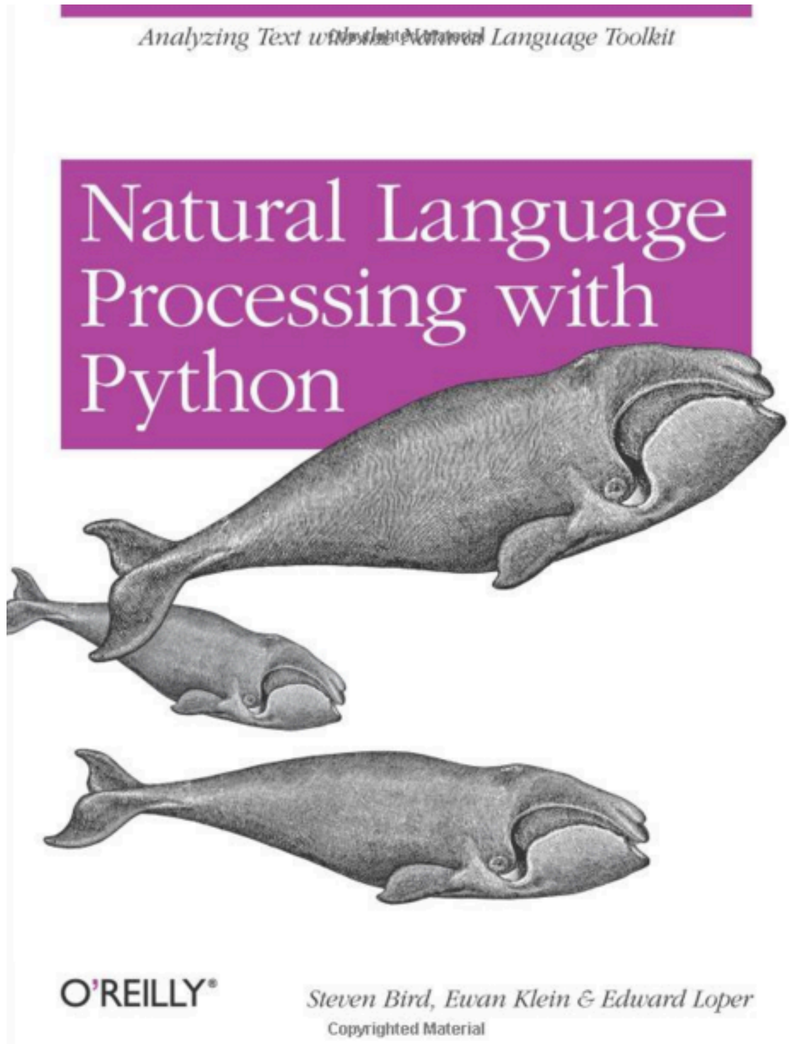


http://www.amazon.com/Foundations-Statistical-Natural-Language-Processing/dp/0262133601

Steven Bird, Ewan Klein and Edward Loper (2009),

Natural Language Processing with Python,

O'Reilly Media



http://www.amazon.com/Natural-Language-Processing-Python-Steven/dp/0596516495

Natural Language Processing with Python
– Analyzing Text with the Natural Language Toolkit

www.nltk.org/book/

Natural Language Processing with Python

– Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

The NLTK book is currently being updated for Python 3 and NLTK 3. This is work in progress; chapters that still need to be updated are indicated. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/. A second edition of the book is anticipated in early 2016.

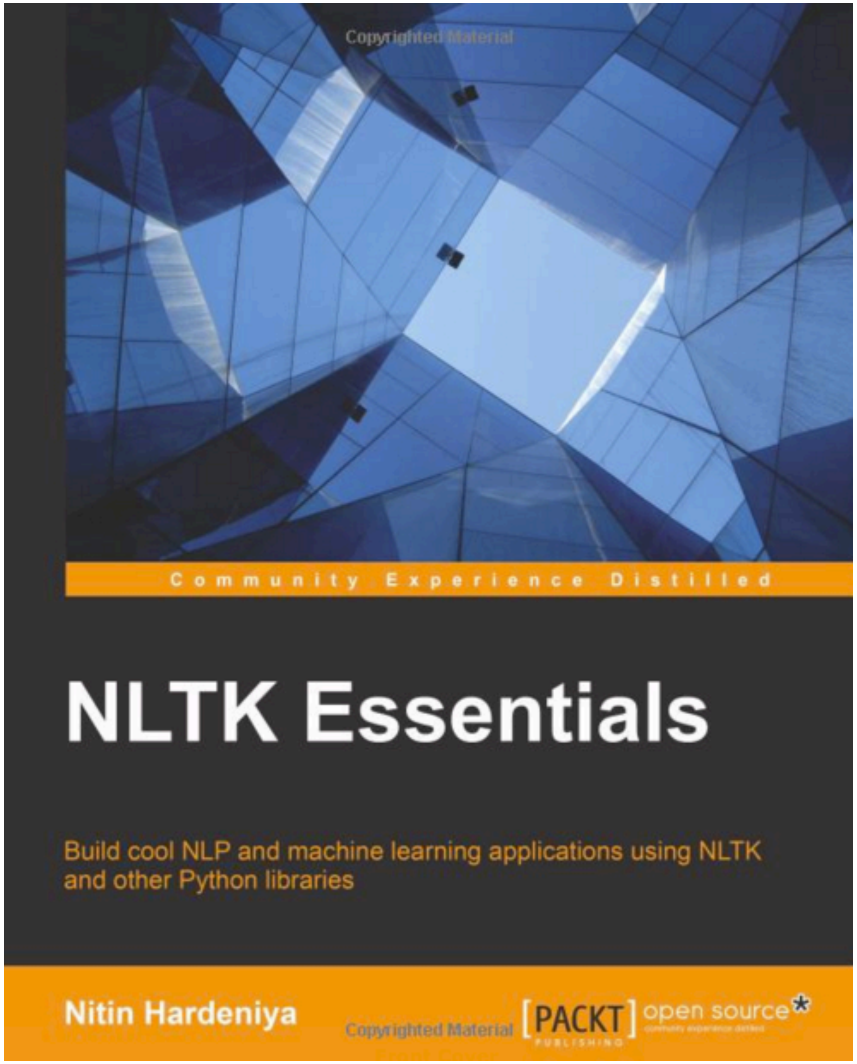
0. Preface
1. Language Processing and Python
2. Accessing Text Corpora and Lexical Resources
3. Processing Raw Text
4. Writing Structured Programs
5. Categorizing and Tagging Words (minor fixes still required)
6. Learning to Classify Text
7. Extracting Information from Text
8. Analyzing Sentence Structure
9. Building Feature Based Grammars
10. Analyzing the Meaning of Sentences (minor fixes still required)
11. Managing Linguistic Data (minor fixes still required)
12. Afterword: Facing the Language Challenge

Bibliography
Term Index

Please post any questions about the materials to the nltk-users mailing list. Please report any errors on the issue tracker.

http://www.nltk.org/book/

Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing



http://www.amazon.com/NLTK-Essentials-Nitin-Hardeniya/dp/1784396907

Text Mining (text data mining)

the process of deriving high-quality information from text

http://en.wikipedia.org/wiki/Text_mining

Typical Text Mining Tasks

- Text categorization
- Text clustering
- Concept/entity extraction
- Production of granular taxonomies
- Sentiment analysis
- Document summarization
- Entity relation modeling
  - i.e., learning relations between named entities.

http://en.wikipedia.org/wiki/Text_mining

Web Mining

- Web mining
  - discover useful information or knowledge from the **Web hyperlink structure, page content, and usage data.**
- Three types of web mining tasks
  - Web structure mining
  - Web content mining
  - Web usage mining

Source: Bing Liu (2009) Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

Text Mining Concepts

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining versus Text Mining

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data:
  - Structured versus unstructured data
  - Structured data: in databases
  - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on
- Text mining – first, impose structure to the data, then mine the structured data

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Concepts

- Benefits of text mining are obvious especially in text-rich data environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communization records (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

Text Mining Application Area

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Text Mining Terminology

- Unstructured or semistructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Terminology

- Term dictionary
- Word frequency
- Part-of-speech tagging (POS)
- Morphology
- Term-by-document matrix (TDM)
  - Occurrence matrix
- Singular Value Decomposition (SVD)
  - Latent Semantic Indexing (LSI)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- Structuring a collection of text
  - Old approach: bag-of-words
  - New approach: natural language processing
- NLP is …
  - a very important concept in text mining
  - a subfield of artificial intelligence and computational linguistics
  - the studies of "understanding" the natural human language
- Syntax versus semantics based text mining

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- What is "Understanding" ?
  - Human understands, what about computers?
  - Natural language is vague, context driven
  - True understanding requires extensive knowledge of a topic

  - Can/will computers ever understand natural language the same/accurate way we do?

Natural Language Processing (NLP)

- Challenges in NLP
  - Part-of-speech tagging
  - Text segmentation
  - Word sense disambiguation
  - Syntax ambiguity
  - Imperfect or irregular input
  - Speech acts
- Dream of AI community
  - to have algorithms that are capable of automatically reading and obtaining knowledge from text

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- WordNet
  - A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  - A major resource for NLP
  - Need automation to be completed
- Sentiment Analysis
  - A technique used to detect favorable and unfavorable opinions toward specific products and services
  - CRM application

NLP Task Categories

- Information retrieval (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition (OCR)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Marketing applications
  - Enables better CRM
- Security applications
  - ECHELON, OASIS
  - Deception detection (…)
- Medicine and biology
  - Literature-based gene identification (…)
- Academic applications
  - Research stream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

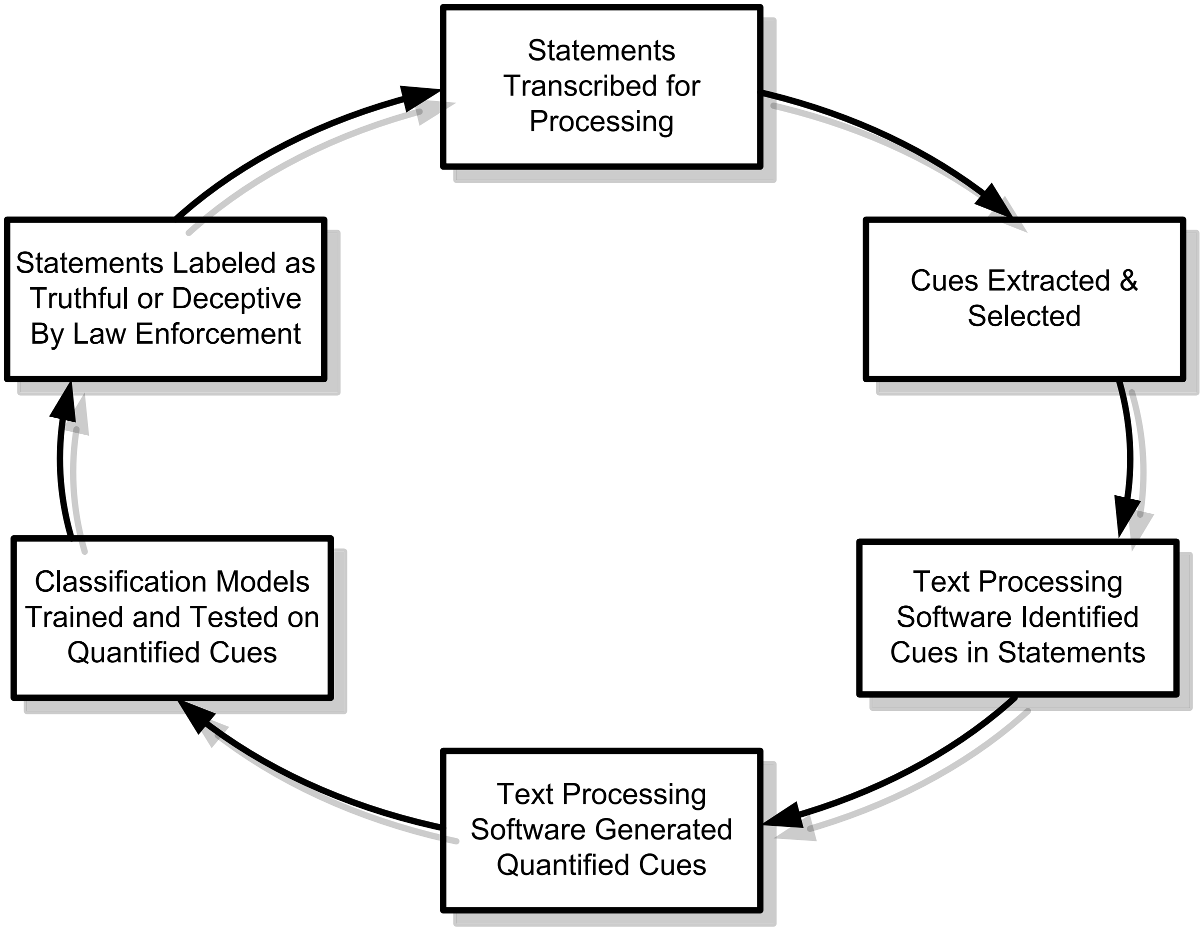
Text Mining Applications

- Application Case: Mining for Lies
- Deception detection
  - A difficult problem
  - If detection is limited to only text, then the problem is even more difficult
- The study
  - analyzed text based testimonies of person of interests at military bases
  - used only text-based features (cues)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Application Case: Mining for Lies



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Application Case: Mining for Lies

Category	Example Cues
Quantity	Verb count, noun-phrase count, ...
Complexity	Avg. no of clauses, sentence length, …
Uncertainty	Modifiers, modal verbs, ...
Nonimmediacy	Passive voice, objectification, ...
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, ...
Informality	Typographical error ratio
Specificity	Spatiotemporal, perceptual information …
Affect	Positive affect, negative affect, etc.

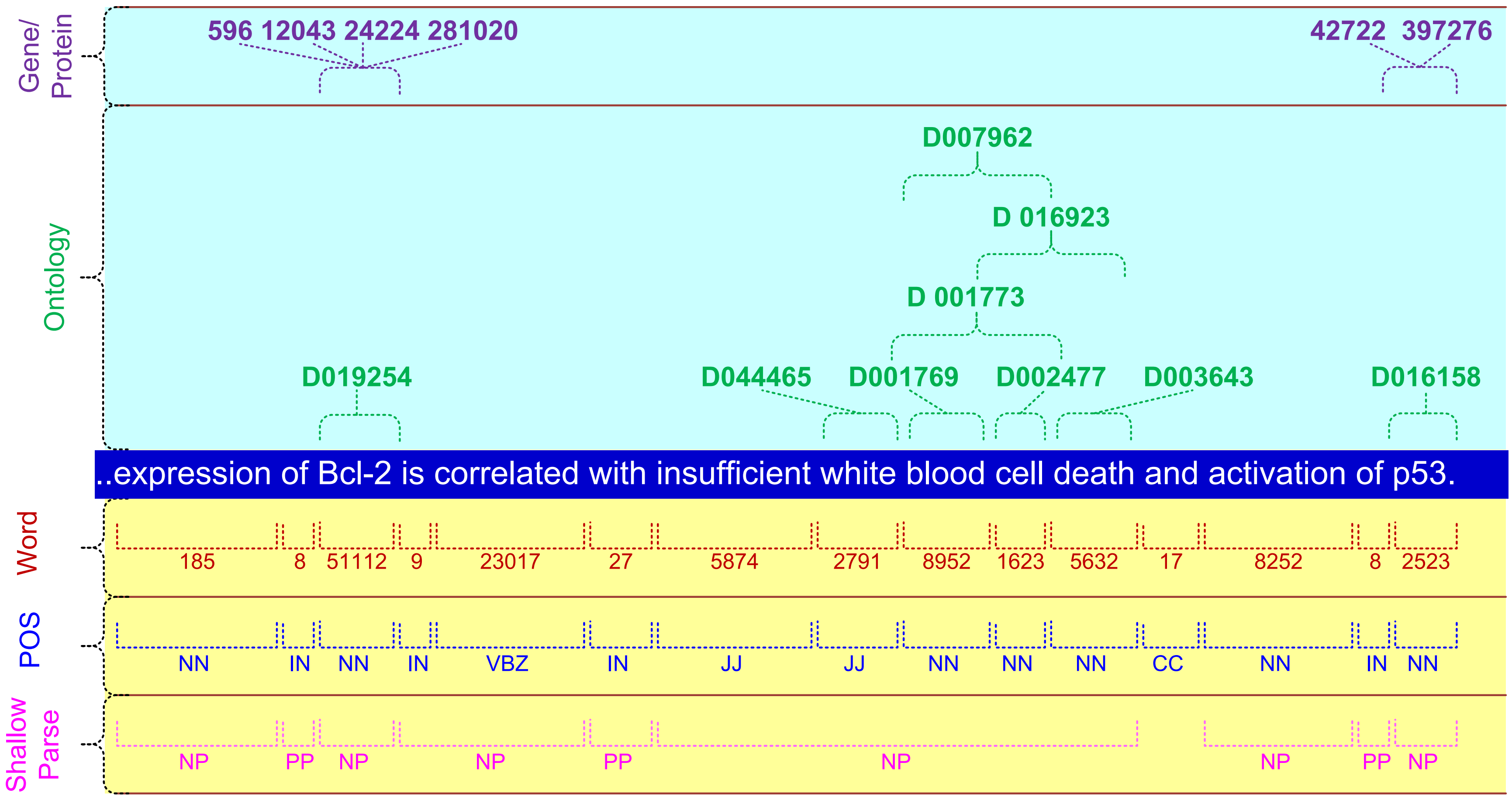
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Application Case: Mining for Lies
  - 371 usable statements are generated
  - 31 features are used
  - Different feature selection methods used
  - 10-fold cross validation is used
  - Results (overall % accuracy)
    - Logistic regression 67.28
    - Decision trees 71.60
    - Neural networks 73.46

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

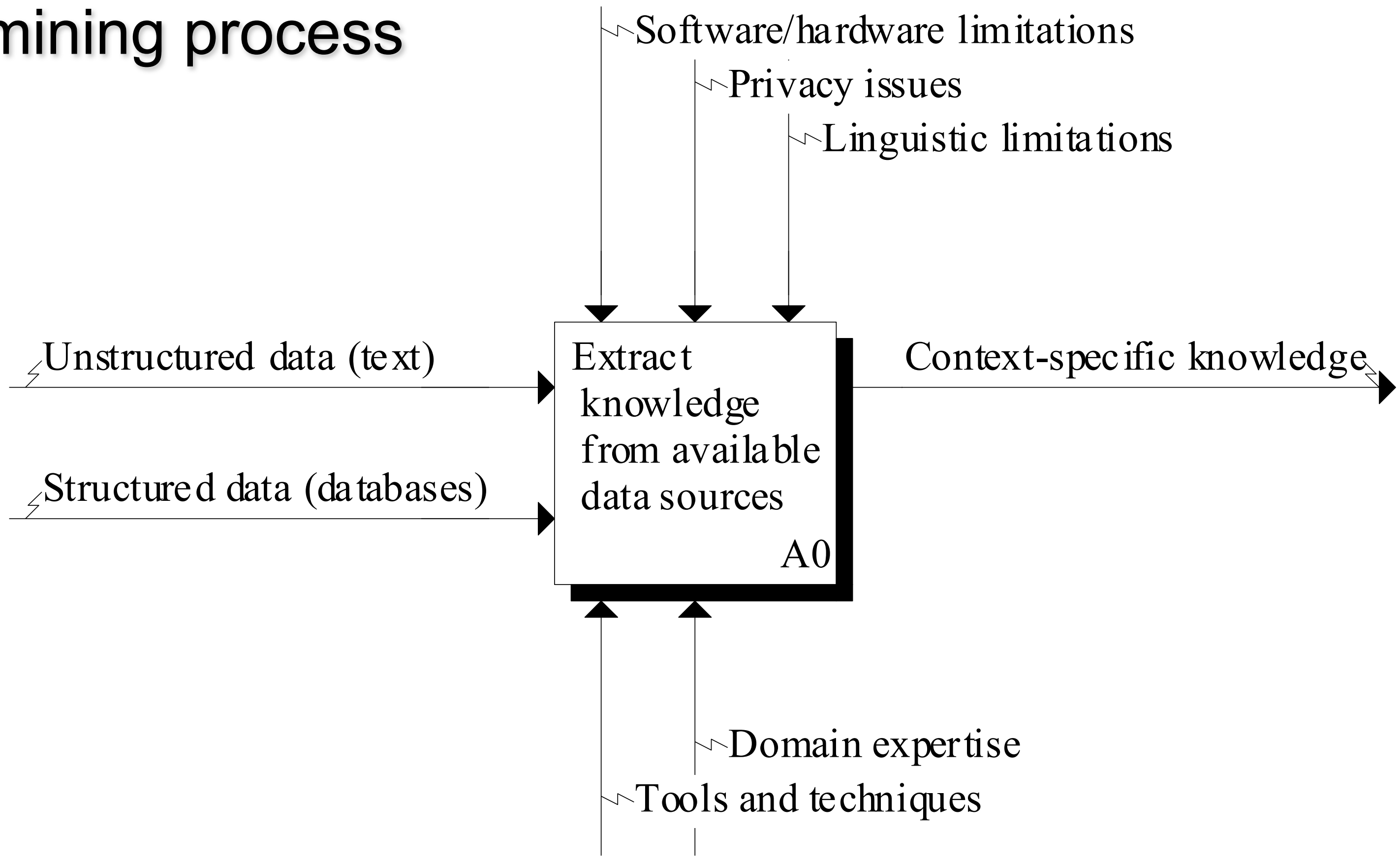
Text Mining Applications
(gene/protein interaction identification)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

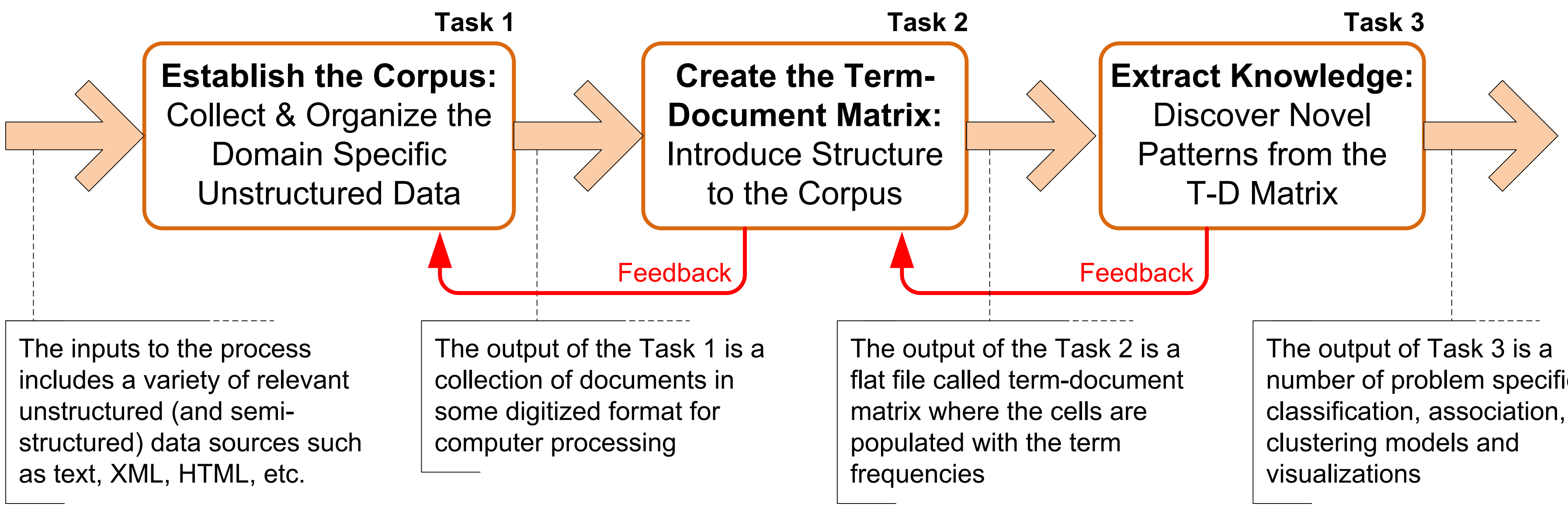
Text Mining Process

Context diagram for the text mining process



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process



The three-step text mining process

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process

- Step 1: Establish the corpus
  - Collect all relevant unstructured data (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings…)
  - Digitize, standardize the collection (e.g., all in ASCII text files)
  - Place the collection in a common place (e.g., in a flat file, or in a directory as separate files)

Text Mining Process

- Step 2: Create the Term–by–Document Matrix

Documents \ Terms	investment risk	project management	software engineering	development	SAP	...
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
...						

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process

- Step 2: Create the Term–by–Document Matrix (TDM), cont.
  - Should all terms be included?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming
  - What is the best representation of the indices (values in cells)?
    - Row counts; binary frequencies; log frequencies;
    - Inverse document frequency

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process

- Step 2: Create the Term–by–Document Matrix (TDM), cont.
  - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    - Manual - a domain expert goes through it
    - Eliminate terms with very few occurrences in very few documents (?)
    - Transform the matrix using singular value decomposition (SVD)
    - SVD is similar to principle component analysis

Text Mining Process

- Step 3: Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (…)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application
(research trend identification in literature)

- Mining the published IS literature
  - MIS Quarterly (MISQ)
  - Journal of MIS (JMIS)
  - Information Systems Research (ISR)

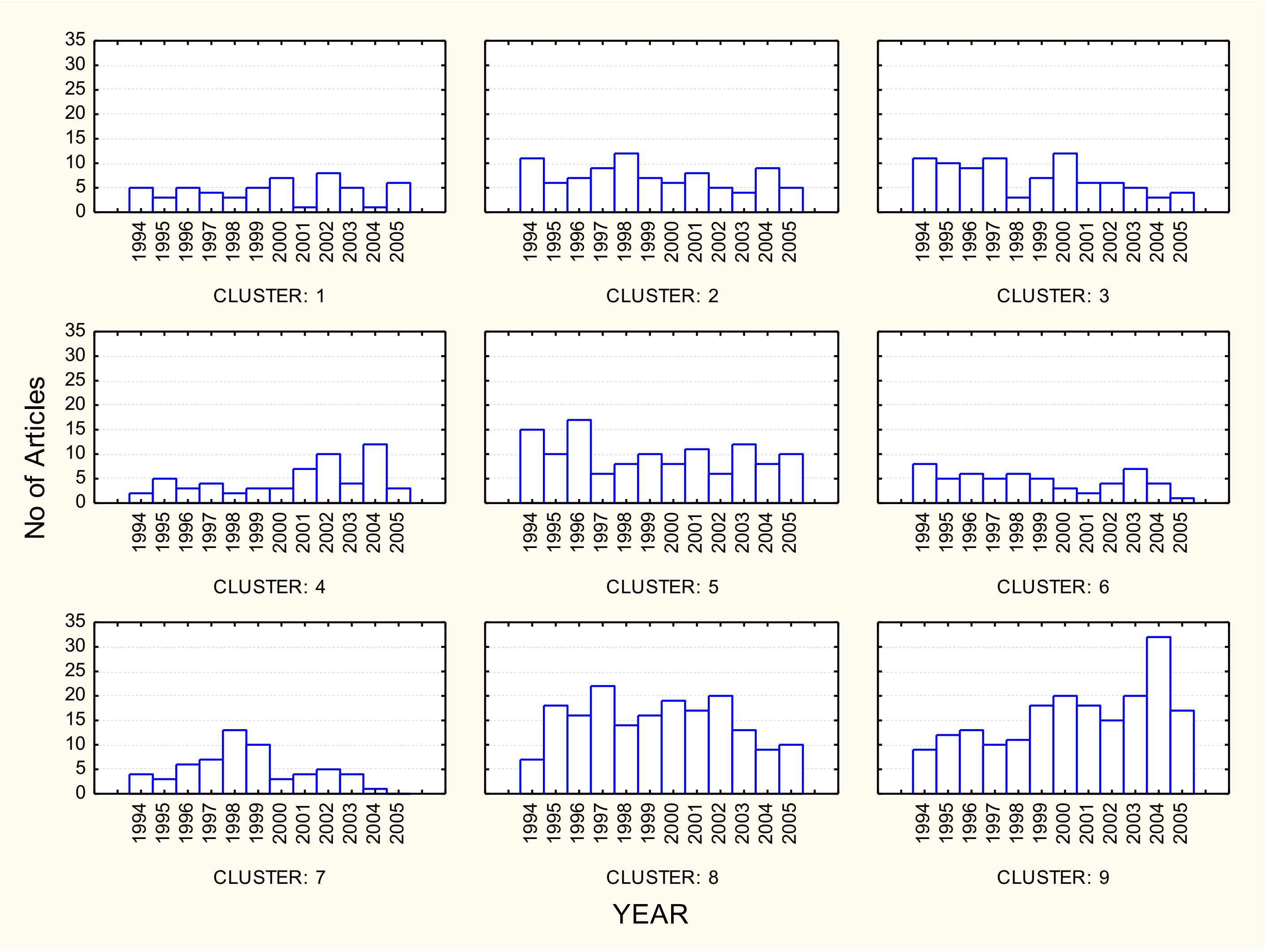
  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

Text Mining Application
(research trend identification in literature)

Journal	Year	Author(s)	Title	Vol/No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-187	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M. C. Boudreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications	2-Oct	167-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products	18/2	65-88	information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of
…	…	…	…	…	…	…	…

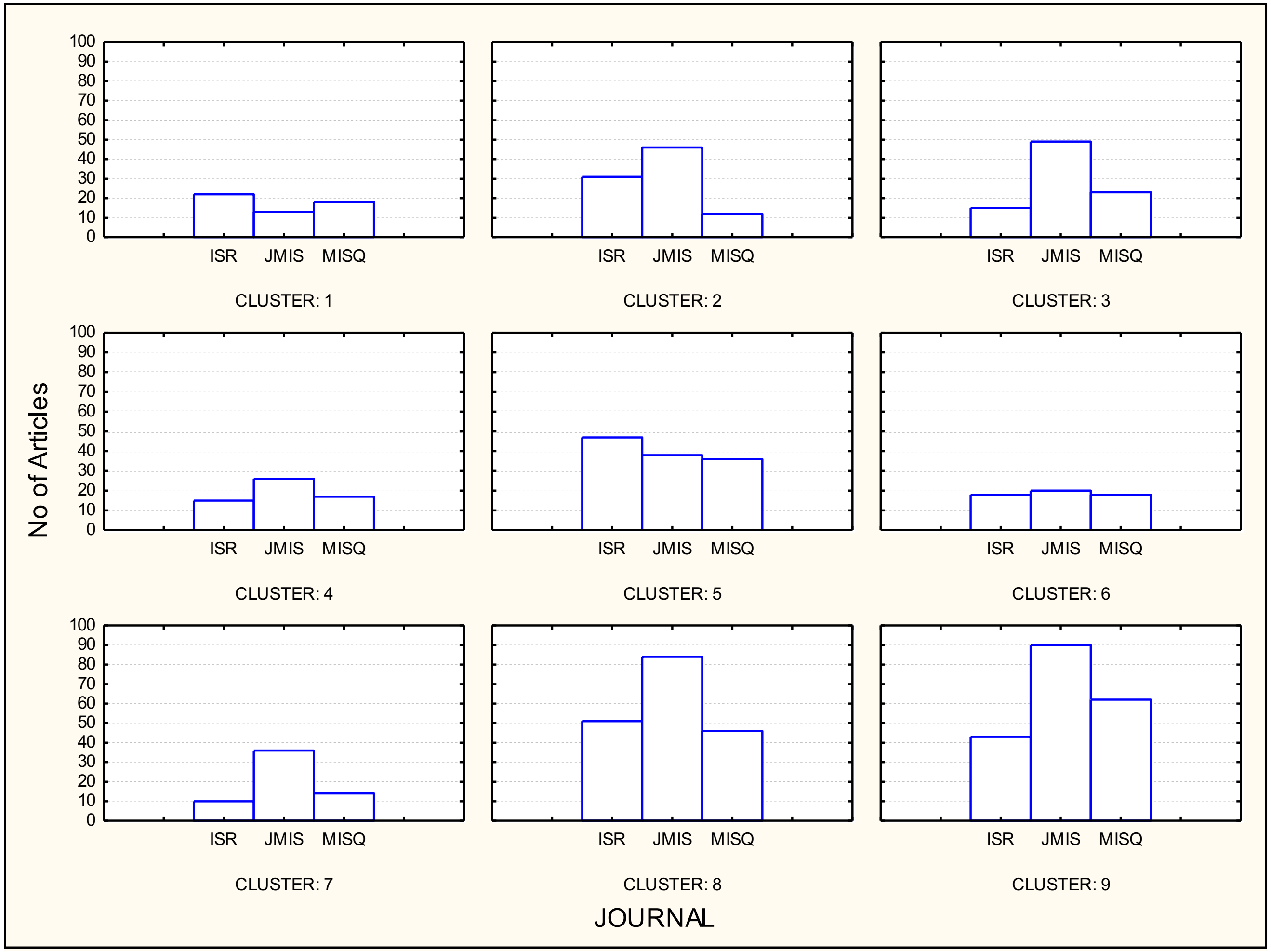
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application
(research trend identification in literature)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application
(research trend identification in literature)



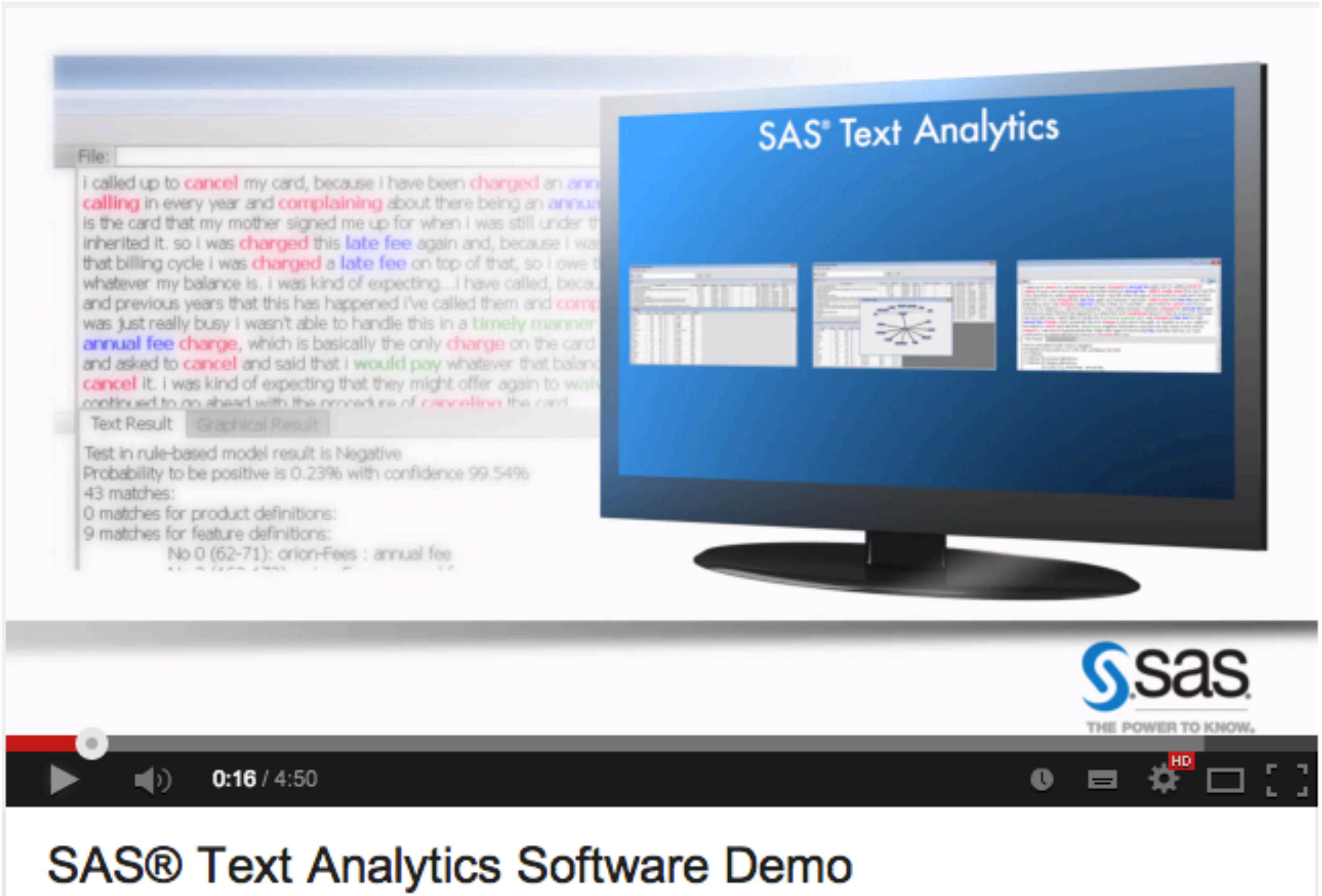
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Tools

- Commercial Software Tools
  - SPSS PASW Text Miner
  - SAS Enterprise Miner
  - Statistica Data Miner
  - ClearForest, …
- Free Software Tools
  - RapidMiner
  - GATE
  - Spy-EM, …

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

SAS Text Analytics



https://www.youtube.com/watch?v=l1rYdrRCZJ4

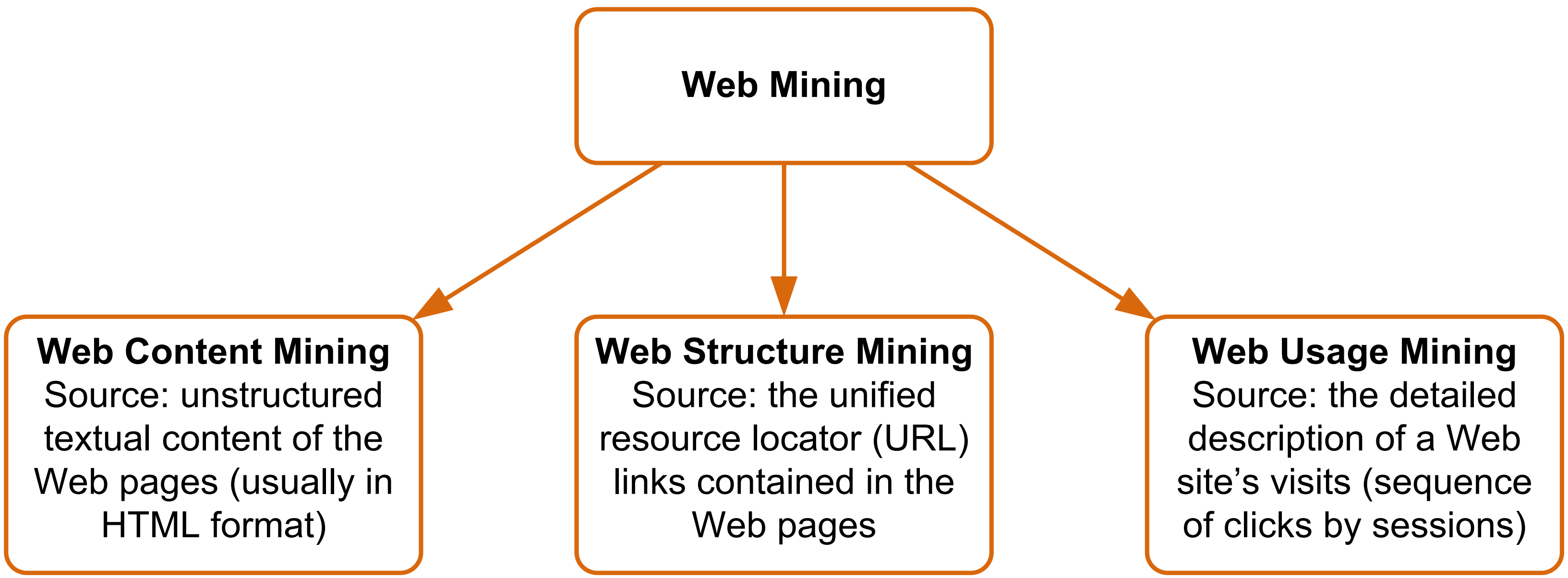
Web Mining Overview

- Web is the largest repository of data
- Data is in HTML, XML, text format
- Challenges (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- Opportunities and challenges are great!

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining

- Web mining (or Web data mining) is the <u>process</u> of discovering intrinsic relationships from Web data (textual, linkage, or usage)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Content/Structure Mining

- Mining of the textual content on the Web
- Data collection via Web crawlers

- Web pages include hyperlinks
  - Authoritative pages
  - Hubs
  - hyperlink-induced topic search (HITS) alg

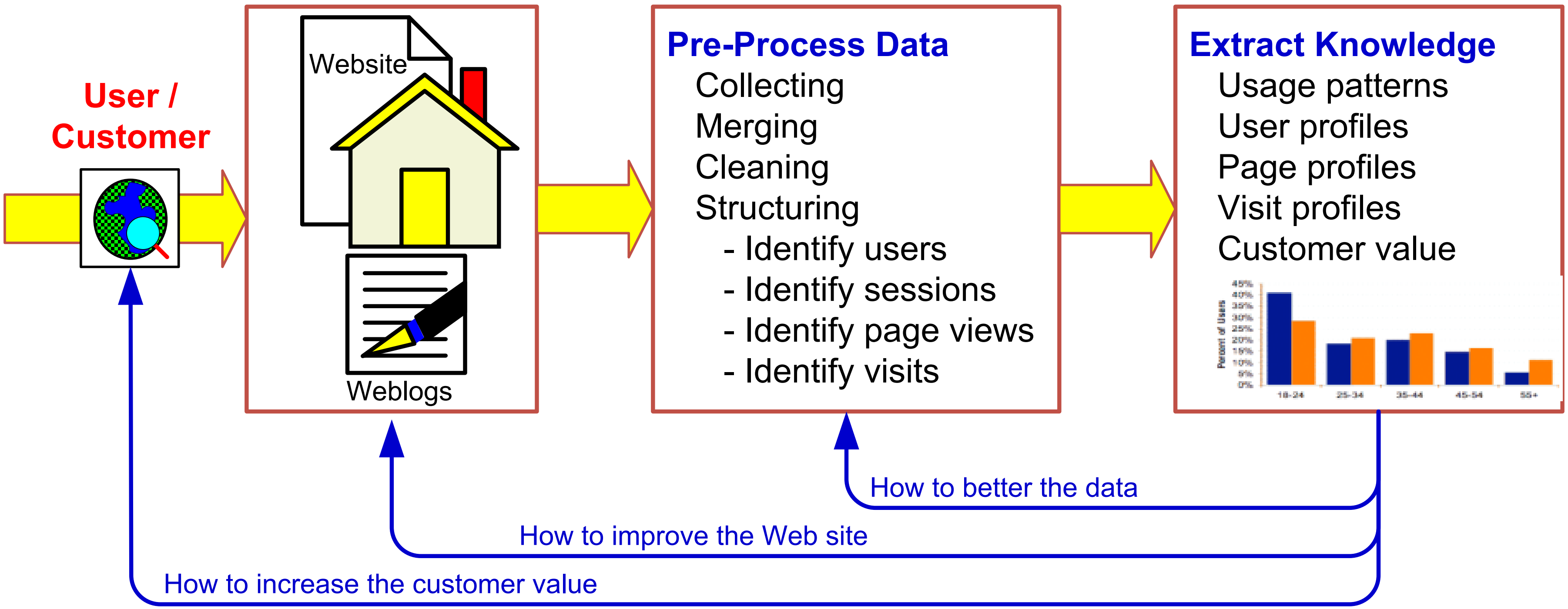
Web Usage Mining

- Extraction of information from data generated through Web page visits and transactions…
  - data stored in server access logs, referrer logs, agent logs, and client-side cookies
  - user characteristics and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream data
- Clickstream analysis

Web Usage Mining

- Web usage mining applications
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products.
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups based on user access patterns
  - Predict user behavior based on previously learned rules and users' profiles
  - Present dynamic information to users based on their interests and profiles…

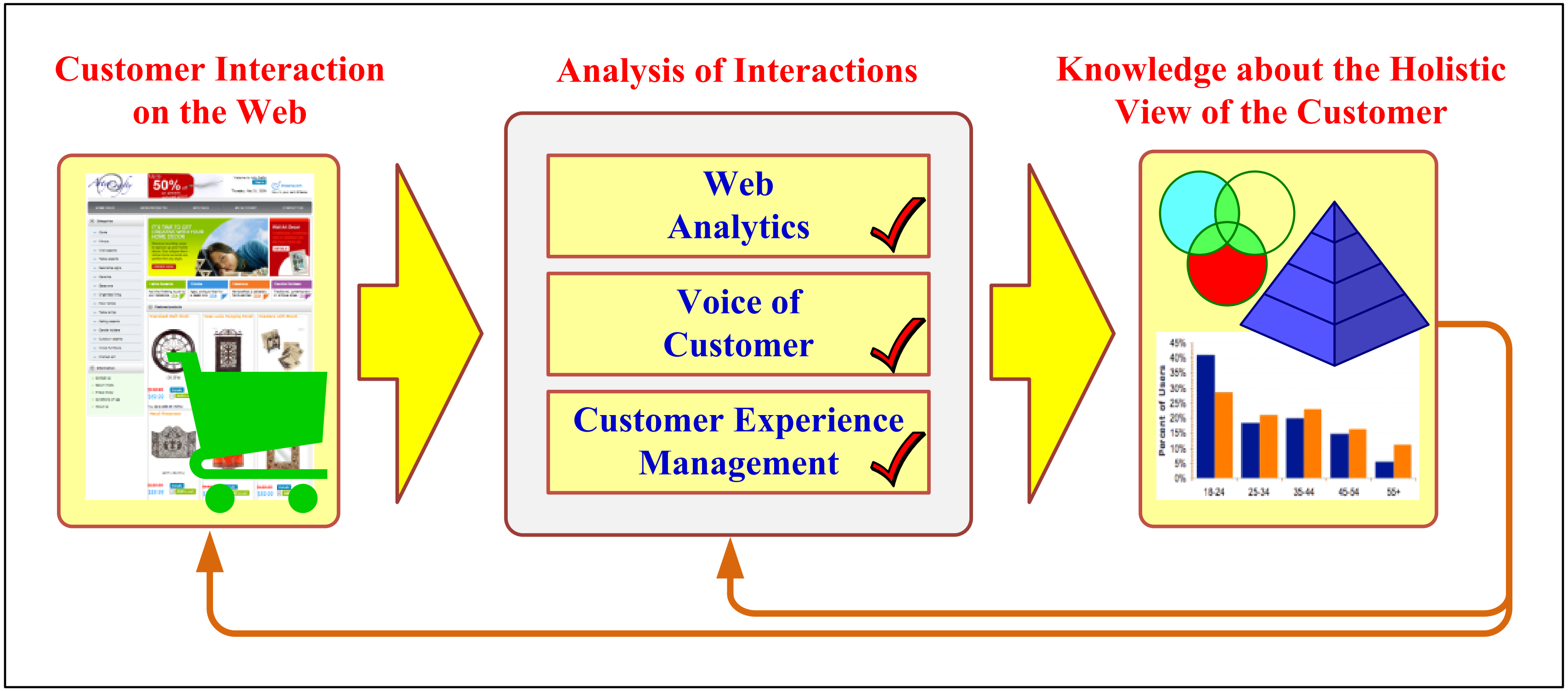
Web Usage Mining
(clickstream analysis)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, …
- Website Optimization Ecosystem



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

CKIP 中研院中文斷詞系統

http://ckipsvr.iis.sinica.edu.tw/

中文斷詞系統

相關系統： 斷詞系統 | 剖析系統 | 詞首詞尾 | 平衡語料庫 | 廣義知網 | 句結構樹庫 | 錯字偵測

- 簡介
- 未知詞擷取做法
- 詞類標記列表
- 線上展示
- 線上服務申請
- 線上資源
- 公告
- 聯絡我們

隱私權聲明 | 版權聲明

數位典藏國家型科技計畫 National Digital Archives Program

線上展示使用簡化詞類進行斷詞標記，僅供參考並且系統不再進行更新。線上服務斷詞和授權mirror site僅提供精簡詞類，結果也與舊版的展示系統不同。

自 2014/01/06 起，本斷詞系統已經處理過 929135 篇文章

送出 清除

歐巴馬是美國的一位總統

歐巴馬是美國的一位總統

文章的文字檔

擷取未知詞過程

包含未知詞的斷詞標記結果

未知詞列表

歐巴馬(Nb) 是(SHI) 美國(Nc) 的(DE) 一(Neu) 位(Nf) 總統(Na)

中文文字處理：中文斷詞

抗氣候變遷 白宮籲採緊急行動

中央通訊社 THE CENTRAL NEWS AGENCY 中央社 – 2014年5月6日 下午10:58

（中央社華盛頓6日綜合外電報導）白宮今天公布全球暖化對全美及美國經濟關鍵產業造成何種衝擊的新報告，呼籲採取緊急行動對抗氣候變遷。

這份為期4年的調查警告，極端氣候事件將對住家、基礎設施及產業帶來嚴重威脅。

美國總統歐巴馬2008年當選總統時曾在競選造勢時誓言，要讓美國成為對抗氣候變遷與相關「安全威脅」的領頭羊。

但歐巴馬在任上一直未能說服美國國會採取重大行動。

在本週對這項議題採取的新作為中，歐巴馬今天將與數名氣象學家接受電視訪問，討論美國全國氣候評估第3版調查結果。

美國數百名來自政府與民間的頂尖氣候科學家及技術專家，共同投入這項研究，檢視氣候變遷對當今帶來的衝擊並預測將對下個世紀帶來何種影響。

研究人員警告，加州可能發生旱災、奧克拉荷馬州發生草原大火，東岸則可能遭遇海平面上升，尤其佛羅里達，而這些事件多為人類造成。

海平面上升也將吞噬密西西比等低窪地區。

至於超過8000萬人居住且擁有全美部分成長最快都會區的東南部與加勒比海區，「海平面上升加上其他與氣候變遷有關的衝擊，以及地層下陷等既有問題，將對經濟和生態帶來重大影響」。

抗氣候變遷 白宮籲採緊急行動
中央社中央社 – 2014年5月6日 下午10:58
（中央社華盛頓6日綜合外電報導）白宮今天公布全球暖化對全美及美國經濟關鍵產業造成何種衝擊的新報告，呼籲採取緊急行動對抗氣候變遷。
這份為期4年的調查警告，極端氣候事件將對住家、基礎設施及產業帶來嚴重威脅。
美國總統歐巴馬2008年當選總統時曾在競選造勢時誓言，要讓美國成為對抗氣候變遷與相關「安全威脅」的領頭羊。
但歐巴馬在任上一直未能說服美國國會採取重大行動。
在本週對這項議題採取的新作為中，歐巴馬今天將與數名氣象學家接受電視訪問，討論美國全國氣候評估第3版調查結果。
美國數百名來自政府與民間的頂尖氣候科學家及技術專家，共同投入這項研究，檢視氣候變遷對當今帶來的衝擊並預測將對下個世紀帶來何種影響。
研究人員警告，加州可能發生旱災、奧克拉荷馬州發生草原大火，東岸則可能遭遇海平面上升，尤其佛羅里達，而這些事件多為人類造成。
海平面上升也將吞噬密西西比等低窪地區。
至於超過8000萬人居住且擁有全美部分成長最快都會區的東南部與加勒比海區，「海平面上升加上其他與氣候變遷有關的衝擊，以及地層下陷等既有問題，將對經濟和生態帶來重大影響」。
報告並說：「過去被認為是遙遠未來議題的氣候變遷，已著實成為當前議題。」（譯者：中央社蔡佳伶）1030506

https://tw.news.yahoo.com/%E6%8A%97%E6%B0%A3%E5%80%99%E8%AE%8A%E9%81%B7-%E7%99%BD%E5%AE%AE%E7%B1%B2%E6%8E%A1%E7%B7%8A%E6%80%A5%E8%A1%8C%E5%8B%95-145804493.html

CKIP 中研院中文斷詞系統

http://ckipsvr.iis.sinica.edu.tw/

CKIP 中研院中文斷詞系統

http://ckipsvr.iis.sinica.edu.tw/

中文斷詞系統

相關系統： 斷詞系統 | 剖析系統 | 詞首詞尾 | 平衡語料庫 | 廣義知網 | 句結構樹庫 | 錯字偵測

- 簡介
- 未知詞擷取做法
- 詞類標記列表
- 線上展示
- 線上服務申請
- 線上資源
- 公告
- 聯絡我們

隱私權聲明 | 版權聲明

數位典藏國家型科技計畫 National Digital Archives Program

抗(VJ) 氣候(Na) 變遷(VH) 白宮(Nc) 籲(VE) 採(VC) 緊急(VH) 行動(Na) 中央社(Nc) 中央社(Nc) ２０１４年(Nd) ５月(Nd) ６日(Nd) 下午(Nd) １

５８(Neu) （(PARENTHESISCATEGORY) 中央社(Nc) 華盛頓(Nc) ６日(Nd) 綜合(A) 外電(Na) 報導(VE) ）(PARENTHESISCATEGORY) 白宮(Nc) 今天(Nd

呼籲(VE) 採取(VC) 緊急(VH) 行動(Na) 對抗(VC) 氣候(Na) 變遷(VH) 。(PERIODCATEGORY)

這(Nep) 份(Nf) 為期(VH) ４年(Nd) 的(DE) 調查(VE) 警告(VE) ，(COMMACATEGORY)

極端(VH) 氣候(Na) 事件(Na) 將(D) 對(P) 住家(Na) 、(PAUSECATEGORY) 基礎(VH) 設施(Na) 及(Caa) 產業(Na) 帶來(VC) 嚴重(VH) 威脅(Na) 。

美國(Nc) 總統(Na) 歐巴馬(Nb) ２００８年(Nd) 當選(VG) 總統(Na) 時(Ng) 曾(D) 在(P) 競選(VC) 造勢(VB) 時(Ng) 誓言(VE) ，(COMMACATEGORY

要(D) 讓(VL) 美國(Nc) 成為(VG) 對抗(VC) 氣候(Na) 變遷(VH) 與(Caa) 相關(VH) 「(PARENTHESISCATEGORY) 安全(VH) 威脅(Na) 」(PARENTHES

但(Cbb) 歐巴馬(Nb) 在任(VH) 上(Ng) 一直(D) 未(D) 能(D) 說服(VF) 美國(Nc) 國會(Nc) 採取(VC) 重大(VH) 行動(Na) 。(PERIODCATEGORY)

在(P) 本(Nes) 週(Nf) 對(P) 這(Nep) 項(Nf) 議題(Na) 採取(VC) 的(DE) 新作(Na) 為(P) 中(Ncd) ，(COMMACATEGORY)

歐巴馬(Nb) 今天(Nd) 將(D) 與(P) 數(Neu) 名(Nf) 氣象學家(Na) 接受(VC) 電視(Na) 訪問(VC) ，(COMMACATEGORY)

討論(VE) 美國(Nc) 全國(Nc) 氣候(Na) 評估(VE) 第３(Neu) 版(Na) 調查(VE) 結果(Dk) 。(PERIODCATEGORY)

美國(Nc) 數百(Neu) 名(Nf) 來自(VJ) 政府(Na) 與(Caa) 民間(Nc) 的(DE) 頂尖(VH) 氣候(Na) 科學家(Na) 及(Caa) 技術(Na) 專家(Na) ，(COMM

共同(A) 投入(VC) 這(Nep) 項(Nf) 研究(Na) ，(COMMACATEGORY)

檢視(VC) 氣候(Na) 變遷(VH) 對(P) 當今(Nd) 帶來(VC) 的(DE) 衝擊(Na) 並(D) 預測(VE) 將(D) 對(P) 下(Nes) 個(Nf) 世紀(Na) 帶來(VC) 何

研究(Na) 人員(Na) 警告(VE) ，(COMMACATEGORY)

加州(Nc) 可能(D) 發生(VJ) 旱災(Na) 、(PAUSECATEGORY) 奧克拉荷馬州(Nc) 發生(VJ) 草原(Na) 大火(Na) ，(COMMACATEGORY)

東岸(Nc) 則(D) 可能(D) 遭遇(VJ) 海平面(Na) 上升(VA) ，(COMMACATEGORY)

尤其(D) 佛羅里達(Nc) ，(COMMACATEGORY)

而(Cbb) 這些(Neqa) 事件(Na) 多(D) 為(VG) 人類(Na) 造成(VK) 。(PERIODCATEGORY)

海平面(Na) 上升(VA) 也(D) 將(D) 吞噬(VC) 密西西比(Nb) 等(Cab) 低窪(VH) 地區(Nc) 。(PERIODCATEGORY)

至於(P) 超過(VJ) ８０００萬(Neu) 人(Na) 居住(VA) 且(Cbb) 擁有(VJ) 全美(Nb) 部分(Neqa) 成長(VH) 最(Dfa) 快(VH) 都會區(Nc) 的(DE) 東南

「(PARENTHESISCATEGORY) 海平面(Na) 上升(VA) 加上(VC) 其他(Neqa) 與(Caa) 氣候(Na) 變遷(VH) 有關(VJ) 的(DE) 衝擊(Na) ，(COMMACATEGOR http://nlp.stanford.edu/software/index.shtml

The Stanford Natural Language Processing Group

home · people · teaching · research · publications · software · events · local

The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Stanford NLP Software

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

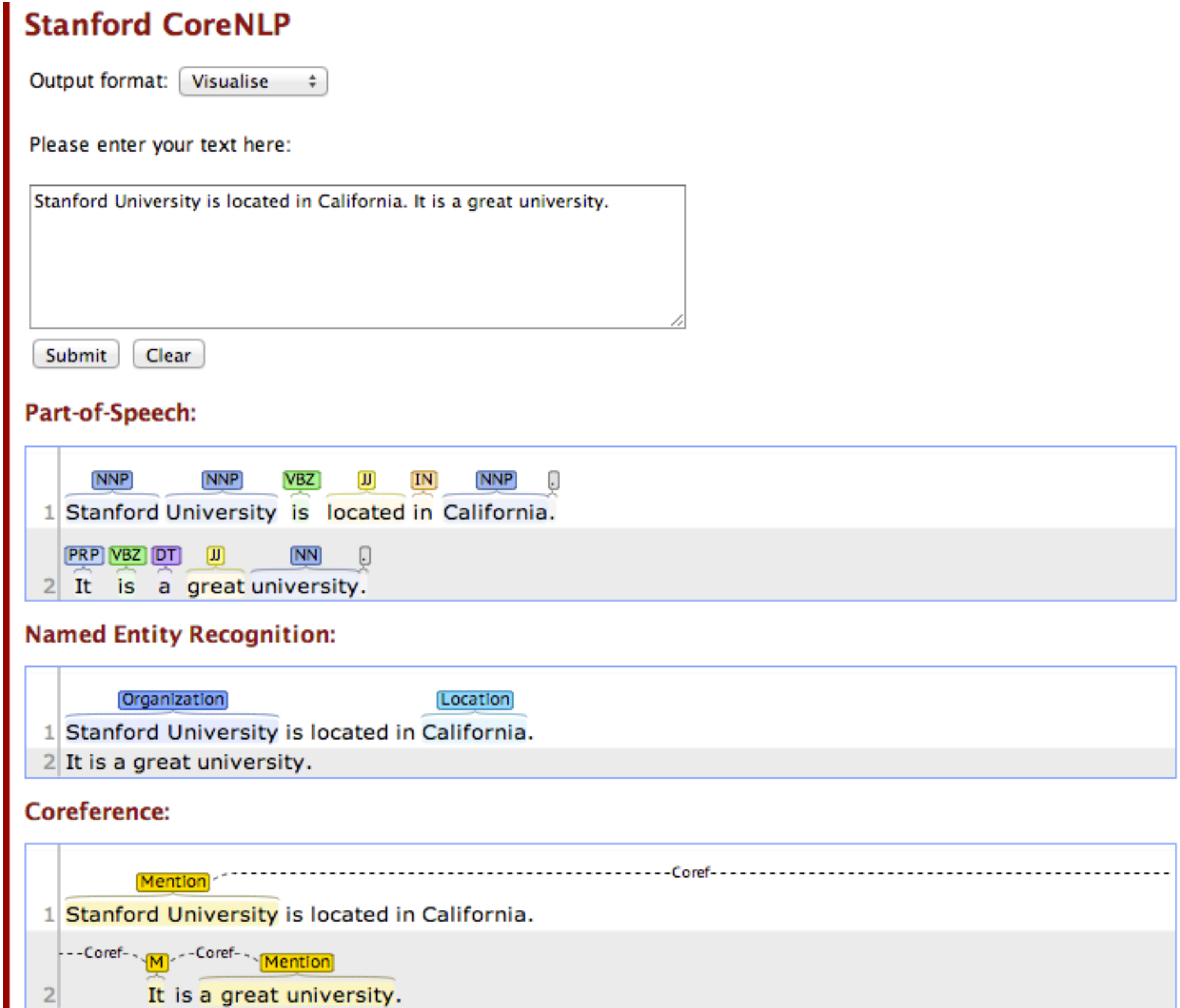
All these software distributions are open source, **licensed under the GNU General Public License** (v2 or later). Note that this is the *full* GPL, which allows many free uses, but *does not allow* its incorporation into any type of distributed proprietary software, even in part or in translation. **Commercial licensing** is also available; please contact us if you are interested.

Stanford CoreNLP
: An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

Stanford Parser
: Implementations of probabilistic natural language parsers in Java: highly optimized PCFG and dependency parsers, a lexicalized PCFG parser, and a deep learning reranker. See also: Online parser demo, the Stanford Dependencies page, and Parser FAQ.

Stanford POS Tagger
: A maximum-entropy (CMM) part-of-speech (POS) tagger for English,

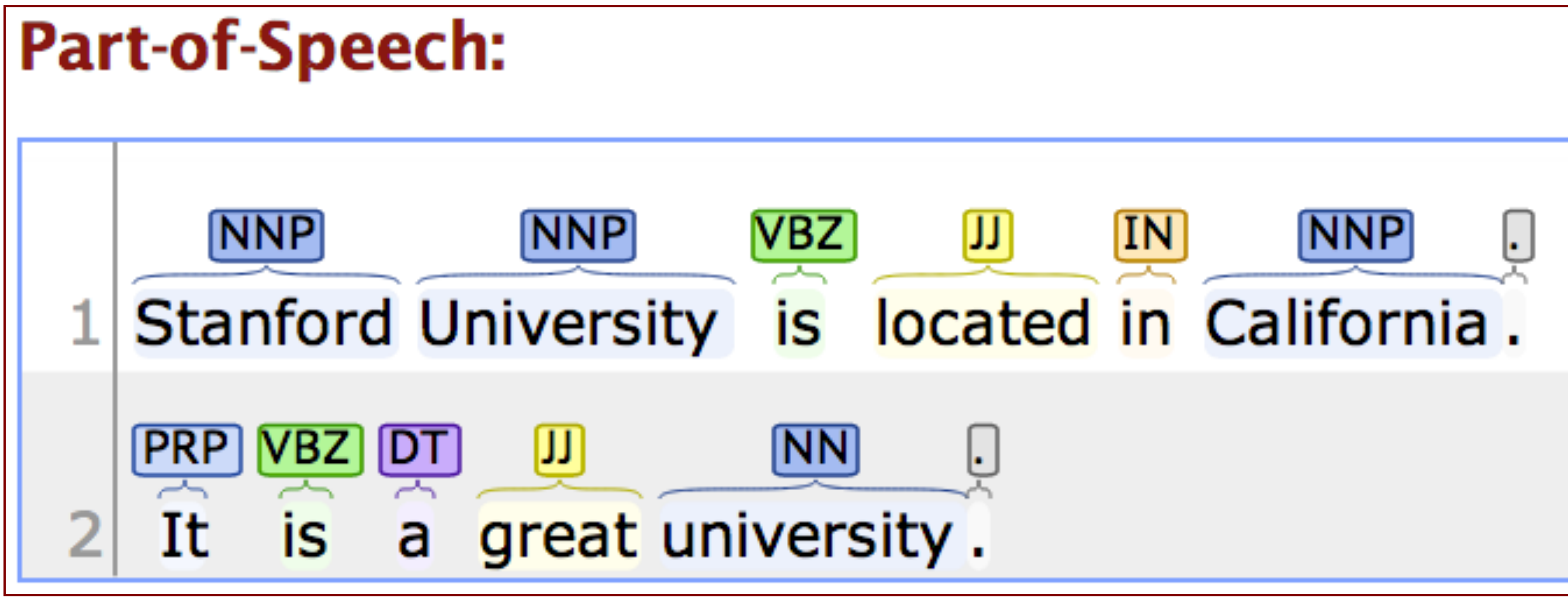
Stanford CoreNLP http://nlp.stanford.edu:8080/corenlp/process



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

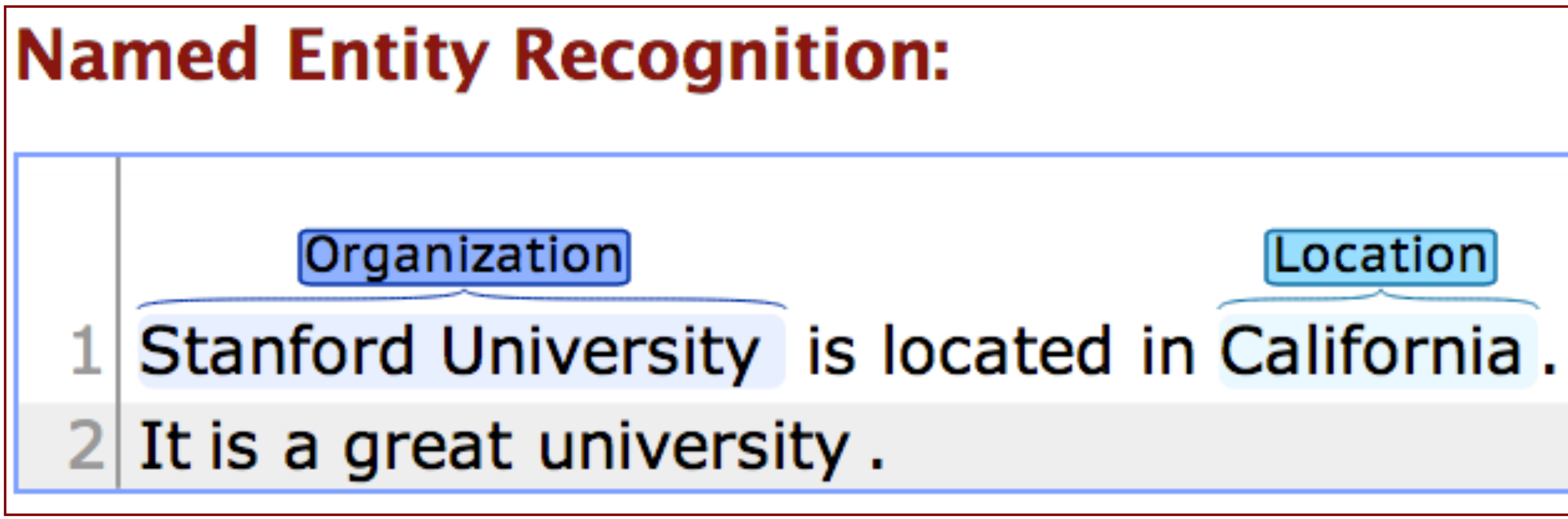
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

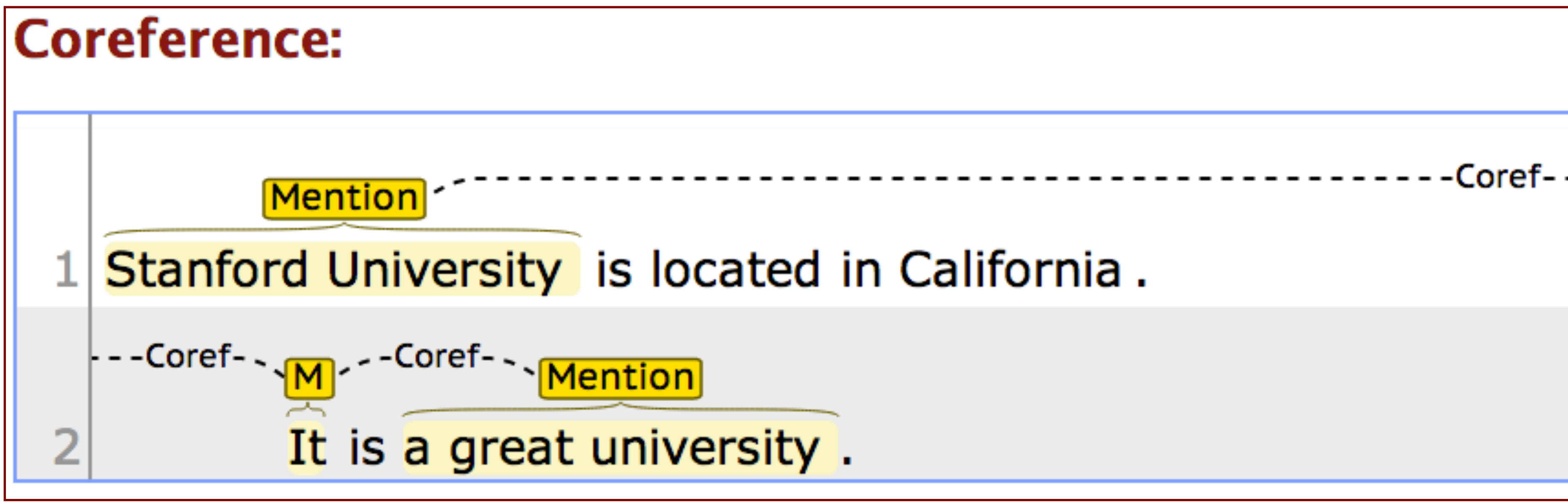
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

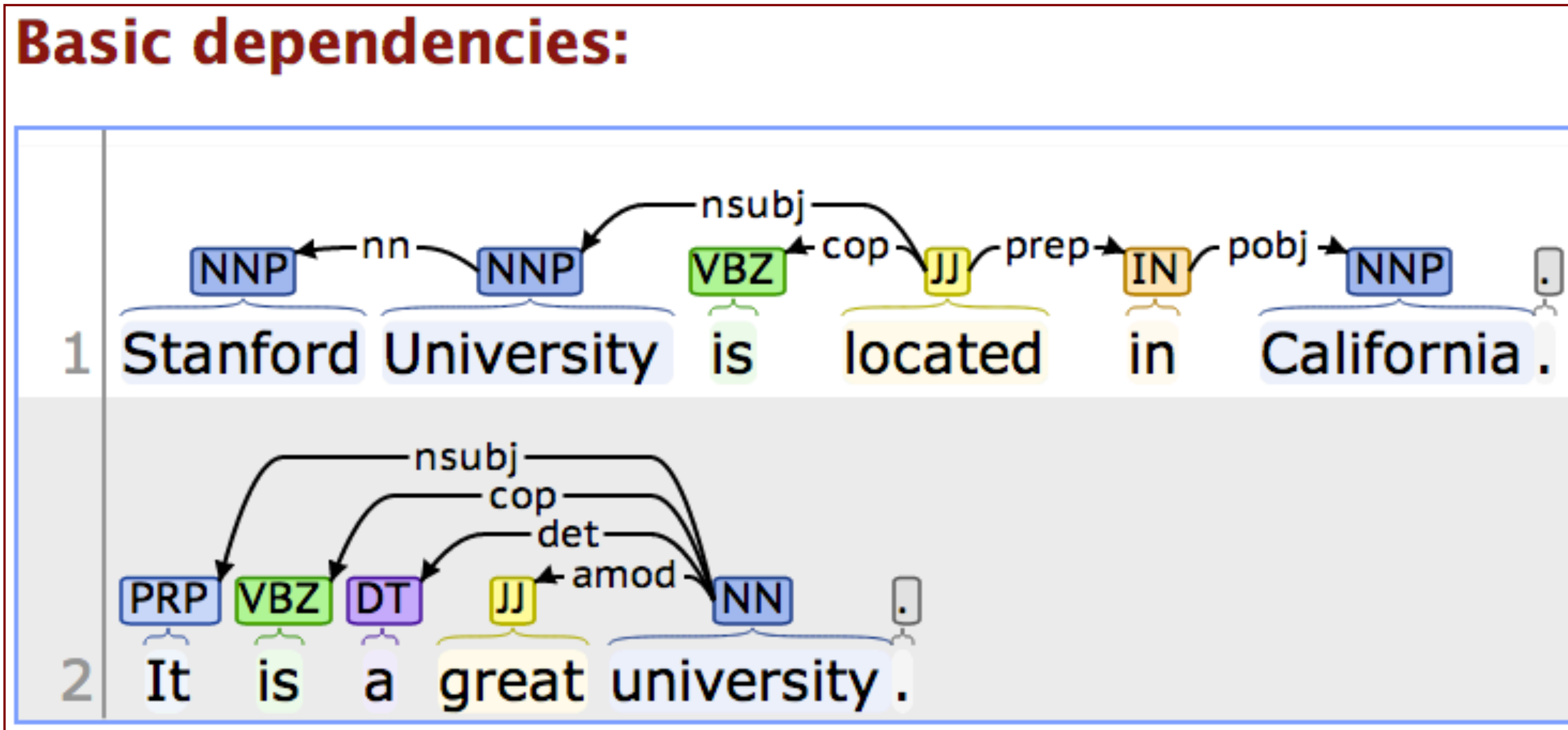
Stanford University is located in California.
It is a great university.



Stanford CoreNLP

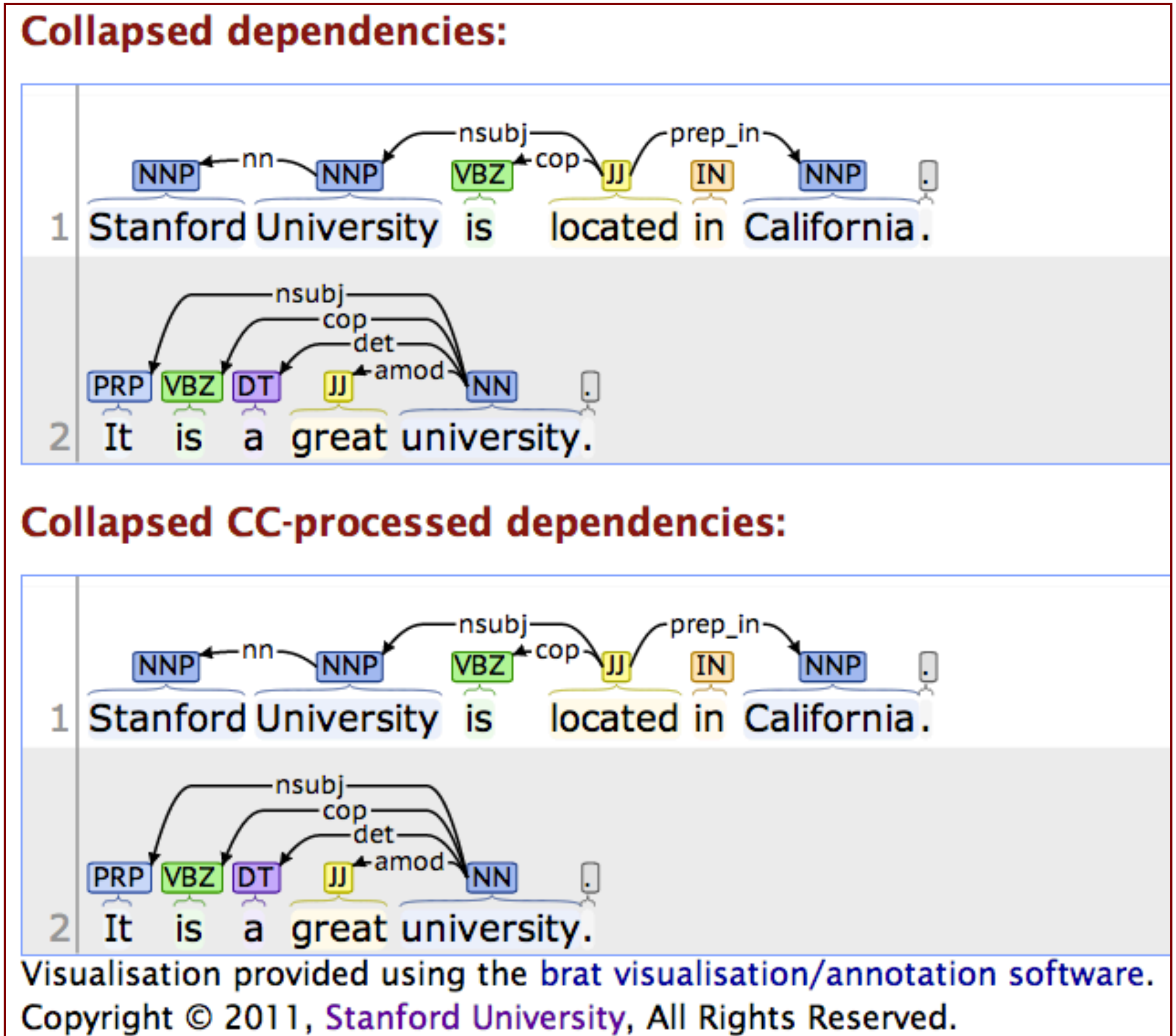
http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process



Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Output format: Pretty print

Please enter your text here:

Stanford University is located in California. It is a great university.

Submit Clear

Stanford CoreNLP XML Output

Document

Document Info

Sentences

Sentence #1

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O		PER0
4	located	located	23	30	JJ	O		PER0
5	in	in	31	33	IN	O		PER0
6	California	California	34	44	NNP	LOCATION		PER0
7	.	.	44	45	.	O		PER0

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Sentence #1

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O		PER0
4	located	located	23	30	JJ	O		PER0
5	in	in	31	33	IN	O		PER0
6	California	California	34	44	NNP	LOCATION		PER0
7	.	.	44	45	.	O		PER0

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Sentence #2

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	It	it	46	48	PRP	O		PER0
2	is	be	49	51	VBZ	O		PER0
3	a	a	52	53	DT	O		PER0
4	great	great	54	59	JJ	O		PER0
5	university	university	60	70	NN	O		PER0
6	.	.	70	71	.	O		PER0

Parse tree

(ROOT (S (NP (PRP It)) (VP (VBZ is) (NP (DT a) (JJ great) (NN university))) (. .)))

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Coreference resolution graph

1.

Sentence	Head	Text	Context
1	2 (gov)	Stanford University	
2	1	It	
2	5	a great university	

Tokens

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker
1	Stanford	Stanford	0	8	NNP	ORGANIZATION		PER0
2	University	University	9	19	NNP	ORGANIZATION		PER0
3	is	be	20	22	VBZ	O		PER0
4	located	located	23	30	JJ	O		PER0
5	in	in	31	33	IN	O		PER0
6	California	California	34	44	NNP	LOCATION		PER0
7	.	.	44	45	.	O		PER0

Parse tree

(ROOT (S (NP (NNP Stanford) (NNP University)) (VP (VBZ is) (ADJP (JJ located) (PP (IN in) (NP (NNP California))))) (. .)))

Uncollapsed dependencies

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep (located-4 , in-5)
pobj (in-5 , California-6)

Collapsed dependencies

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep_in (located-4 , California-6)

Collapsed dependencies with CC processed

root (ROOT-0 , located-4)
nn (University-2 , Stanford-1)
nsubj (located-4 , University-2)
cop (located-4 , is-3)
prep_in (located-4 , California-6)

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Stanford University is located in California.
It is a great university.

Stanford CoreNLP

http://nlp.stanford.edu:8080/corenlp/process

Output format: XML

Please enter your text here:

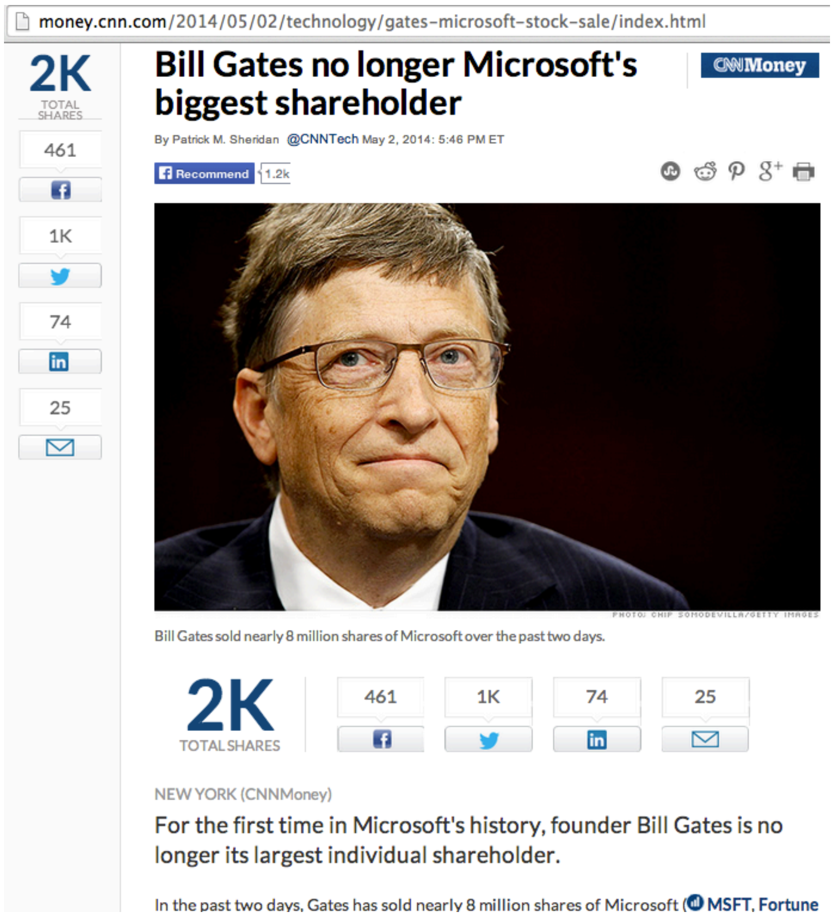
Stanford University is located in California. It is a great university.

Submit Clear


```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="CoreNLP-to-HTML.xsl" type="text/xsl"?>
<root>
  <document>
    <sentences>
      <sentence id="1">
        <tokens>
          <token id="1">
            <word>Stanford</word>
            <lemma>Stanford</lemma>
            <CharacterOffsetBegin>0</CharacterOffsetBegin>
            <CharacterOffsetEnd>8</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
          <token id="2">
            <word>University</word>
            <lemma>University</lemma>
            <CharacterOffsetBegin>9</CharacterOffsetBegin>
            <CharacterOffsetEnd>19</CharacterOffsetEnd>
            <POS>NNP</POS>
            <NER>ORGANIZATION</NER>
            <Speaker>PER0</Speaker>
          </token>
```

NER for News Article

http://money.cnn.com/2014/05/02/technology/gates-microsoft-stock-sale/index.html



Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

NEW YORK (CNNMoney)

For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder.
In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million.

That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares.
Related: Gates reclaims title of world's richest billionaire
Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires.
It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation.
The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:

- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: inlineXML

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: xml

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

<wi num="0" entity="O">Bill</wi> <wi num="1" entity="O">Gates</wi> <wi num="2" entity="O">no</wi> <wi num="3" entity="O">longer</wi> <wi num="4" entity="ORGANIZATION">Microsoft</wi><wi num="5" entity="O">'s</wi> <wi num="6" entity="O">biggest</wi> <wi num="7" entity="O">shareholder</wi> <wi num="8" entity="O">By</wi> <wi num="9" entity="PERSON">Patrick</wi> <wi num="10" entity="PERSON">M.</wi> <wi num="11" entity="PERSON">Sheridan</wi> <wi num="12" entity="O">@CNNTech</wi> <wi num="13" entity="DATE">May</wi> <wi num="14" entity="DATE">2</wi><wi num="15" entity="DATE">,</wi> <wi num="16" entity="DATE">2014</wi><wi num="17" entity="O">:</wi> <wi num="18" entity="O">5:46</wi> <wi num="19" entity="O">PM</wi> <wi num="20" entity="O">ET</wi> <wi num="21" entity="O">Bill</wi> <wi num="22" entity="O">Gates</wi> <wi num="23" entity="O">sold</wi> <wi num="24" entity="O">nearly</wi> <wi num="25" entity="O">8</wi> <wi num="26" entity="O">million</wi> <wi num="27" entity="O">shares</wi> <wi num="28" entity="O">of</wi> <wi num="29" entity="ORGANIZATION">Microsoft</wi> <wi num="30" entity="O">over</wi> <wi num="31" entity="O">the</wi> <wi num="32" entity="O">past</wi> <wi num="33" entity="O">two</wi> <wi num="34" entity="O">days</wi><wi num="35" entity="O">.</wi> <wi num="0" entity="LOCATION">NEW</wi> <wi num="1" entity="LOCATION">YORK</wi> <wi num="2" entity="O">-LRB-</wi><wi num="3" entity="O">CNNMoney</wi><wi num="4" entity="O">-RRB-</wi> <wi num="5" entity="O">For</wi> <wi num="6" entity="O">the</wi> <wi num="7" entity="O">first</wi> <wi num="8" entity="O">time</wi> <wi num="9" entity="O">in</wi> <wi num="10" entity="ORGANIZATION">Microsoft</wi><wi num="11" entity="O">'s</wi> <wi num="12" entity="O">history</wi><wi num="13" entity="O">,</wi> <wi num="14" entity="O">founder</wi> <wi num="15" entity="PERSON">Bill</wi> <wi num="16" entity="PERSON">Gates</wi> <wi num="17" entity="O">is</wi> <wi num="18" entity="O">no</wi> <wi num="19" entity="O">longer</wi> <wi num="20" entity="O">its</wi> <wi num="21" entity="O">largest</wi> <wi num="22" entity="O">individual</wi> <wi num="23" entity="O">shareholder</wi><wi num="24" entity="O">.</wi> <wi num="0" entity="O">In</wi> <wi num="1" entity="O">the</wi> <wi num="2" entity="DATE">past</wi> <wi num="3" entity="DATE">two</wi> <wi num="4"

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: slashTags

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE,/DATE 2014/DATE:/O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION over/O the/O past/O two/O days/O./O NEW/LOCATION YORK/LOCATION -LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O history/O,/O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O shareholder/O./O In/O the/O past/DATE two/DATE days/DATE,/O Gates/O has/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION,/O Fortune/O 500/O-RRB-/O,/O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O./O That/O puts/O him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON who/O owns/O 333/O million/O shares/O./O Related/O:/O Gates/O reclaims/O title/O of/O world/O's/O richest/O billionaire/O Ballmer/PERSON,/O who/O was/O Microsoft/ORGANIZATION's/O CEO/O until/O earlier/DATE this/DATE year/DATE,/O was/O one/O of/O Gates/O'/O first/O hires/O./O It/O's/O a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O./O Gates/O now/O spends/O his/O time/O and/O personal/O fortune/O helping/O run/O the/O Bill/ORGANIZATION &/ORGANIZATION Melinda/ORGANIZATION Gates/ORGANIZATION foundation/O./O The/O foundation/O has/O spent/O $/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/DATE./O

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.conll.4class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON
- MISC

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford Named Entity Tagger

Classifier: english.all.3class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Bill Gates no longer Microsoft's biggest shareholder
By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET

Bill Gates sold nearly 8 million shares of Microsoft over the past two days.

Submit Clear

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON

Classifier: english.muc.**7class**.distsim.crf.ser.gz

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Classifier: english.all.**3class**.distsim.crf.ser.gz

Bill Gates no longer Microsoft's biggest shareholder By Patrick M. Sheridan @CNNTech May 2, 2014: 5:46 PM ET Bill Gates sold nearly 8 million shares of Microsoft over the past two days. NEW YORK (CNNMoney) For the first time in Microsoft's history, founder Bill Gates is no longer its largest individual shareholder. In the past two days, Gates has sold nearly 8 million shares of Microsoft (MSFT, Fortune 500), bringing down his total to roughly 330 million. That puts him behind Microsoft's former CEO Steve Ballmer who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire Ballmer, who was Microsoft's CEO until earlier this year, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the Bill & Melinda Gates foundation. The foundation has spent $28.3 billion fighting hunger and poverty since its inception back in 1997.

Potential tags:
- LOCATION
- ORGANIZATION
- PERSON

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford NER Output Format: inlineXML

Bill Gates no longer <ORGANIZATION>Microsoft</ORGANIZATION>'s biggest shareholder By <PERSON>Patrick M. Sheridan</PERSON> @CNNTech <DATE>May 2, 2014</DATE>: 5:46 PM ET Bill Gates sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> over the past two days. <LOCATION>NEW YORK</LOCATION> (CNNMoney) For the first time in <ORGANIZATION>Microsoft</ORGANIZATION>'s history, founder <PERSON>Bill Gates</PERSON> is no longer its largest individual shareholder. In the <DATE>past two days</DATE>, Gates has sold nearly 8 million shares of <ORGANIZATION>Microsoft</ORGANIZATION> (<ORGANIZATION>MSFT</ORGANIZATION>, Fortune 500), bringing down his total to roughly 330 million. That puts him behind <ORGANIZATION>Microsoft</ORGANIZATION>'s former CEO <PERSON>Steve Ballmer</PERSON> who owns 333 million shares. Related: Gates reclaims title of world's richest billionaire <PERSON>Ballmer</PERSON>, who was <ORGANIZATION>Microsoft</ORGANIZATION>'s CEO until <DATE>earlier this year</DATE>, was one of Gates' first hires. It's a passing of the torch for Gates who has always been the largest single owner of his company's stock. Gates now spends his time and personal fortune helping run the <ORGANIZATION>Bill & Melinda Gates</ORGANIZATION> foundation. The foundation has spent <MONEY>$28.3 billion</MONEY> fighting hunger and poverty since its inception back in <DATE>1997</DATE>.

Stanford Named Entity Tagger (NER)

http://nlp.stanford.edu:8080/ner/process

Stanford NER Output Format: slashTags

Bill/O Gates/O no/O longer/O Microsoft/ORGANIZATION's/O biggest/O shareholder/O By/O Patrick/PERSON M./PERSON Sheridan/PERSON @CNNTech/O May/DATE 2/DATE,/DATE 2014/DATE:/O 5:46/O PM/O ET/O Bill/O Gates/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION over/O the/O past/O two/O days/O./O NEW/LOCATION YORK/LOCATION -LRB-/OCNNMoney/O-RRB-/O For/O the/O first/O time/O in/O Microsoft/ORGANIZATION's/O history/O,/O founder/O Bill/PERSON Gates/PERSON is/O no/O longer/O its/O largest/O individual/O shareholder/O./O In/O the/O past/DATE two/DATE days/DATE,/O Gates/O has/O sold/O nearly/O 8/O million/O shares/O of/O Microsoft/ORGANIZATION -LRB-/OMSFT/ORGANIZATION,/O Fortune/O 500/O-RRB-/O,/O bringing/O down/O his/O total/O to/O roughly/O 330/O million/O./O That/O puts/O him/O behind/O Microsoft/ORGANIZATION's/O former/O CEO/O Steve/PERSON Ballmer/PERSON who/O owns/O 333/O million/O shares/O./O Related/O:/O Gates/O reclaims/O title/O of/O world/O's/O richest/O billionaire/O Ballmer/PERSON,/O who/O was/O Microsoft/ORGANIZATION's/O CEO/O until/O earlier/DATE this/DATE year/DATE,/O was/O one/O of/O Gates/O'/O first/O hires/O./O It/O's/O a/O passing/O of/O the/O torch/O for/O Gates/O who/O has/O always/O been/O the/O largest/O single/O owner/O of/O his/O company/O's/O stock/O./O Gates/O now/O spends/O his/O time/O and/O personal/O fortune/O helping/O run/O the/O Bill/ORGANIZATION &/ORGANIZATION Melinda/ORGANIZATION Gates/ORGANIZATION foundation/O./O The/O foundation/O has/O spent/O $/MONEY28.3/MONEY billion/MONEY fighting/O hunger/O and/O poverty/O since/O its/O inception/O back/O in/O 1997/DATE./O

Summary

- Text Mining
  - Differentiate between text mining, Web mining and data mining
- Natural Language Processing (NLP)
- Text Mining Tools and Applications

References

- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- Steven Bird, Ewan Klein and Edward Loper, Natural Language Processing with Python, 2009, O'Reilly Media, http://www.nltk.org/book/ , http://www.nltk.org/book_1ed/
- Nitin Hardeniya, NLTK Essentials, 2015, Packt Publishing
- Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
- Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
- Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
- Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2009, Springer
- Bruce Croft, Donald Metzler, and Trevor Strohman, Search Engines: Information Retrieval in Practice, 2008, Addison Wesley, http://www.search-engines-book.com/
- Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, 1999, The MIT Press
- Text Mining, http://en.wikipedia.org/wiki/Text_mining