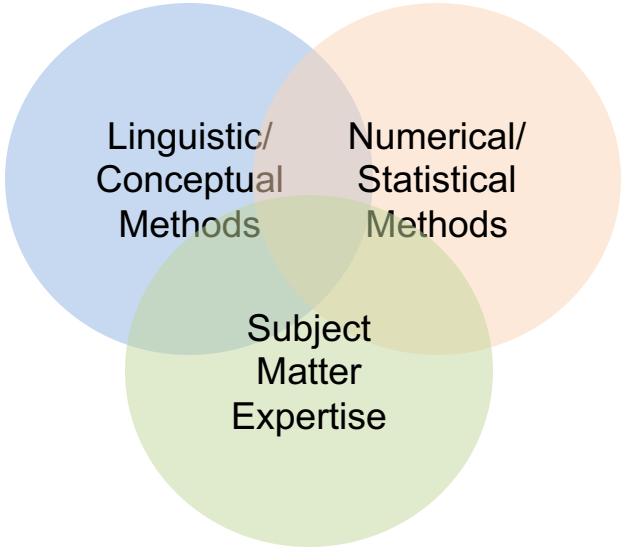
Working in NLP: NLP and Data Science

Natural Language Processing

NLP and Data Science

- NLP and data science intersect in a variety of ways:
  - NLU for creating structured data out of unstructured text
  - Machine learning (ML) in NLP
  - NLU for feature engineering and automated feature extraction to support ML
  - NLG to create “human-friendly” presentation layers for complex data analysis
  - NLU on unstructured text to validate discrete (numerical) data that is related to the text



Creating Structured Data from Unstructured

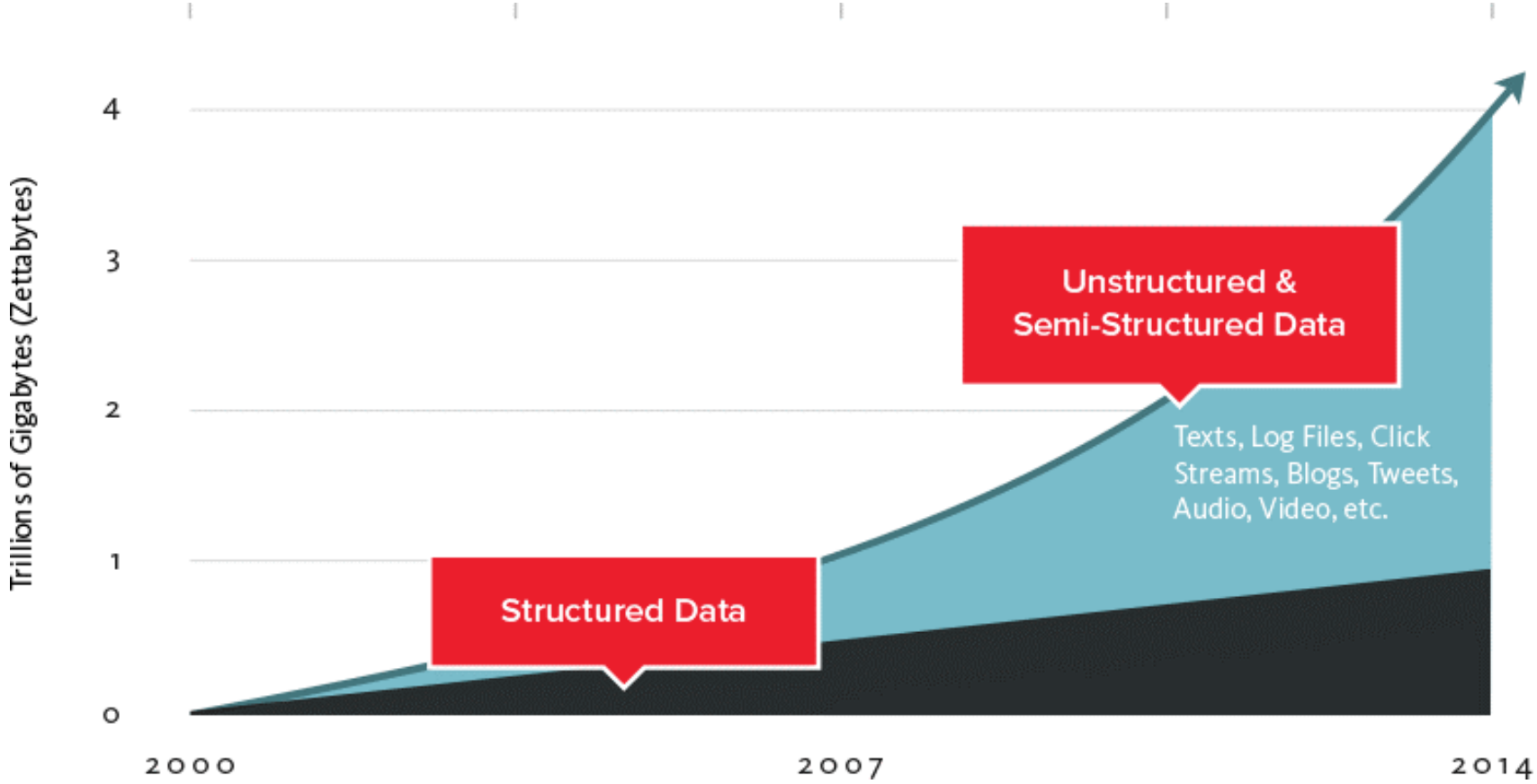
- Data science thrives on structured data:
  - Numerical
  - Scalar
  - Taxonomical
  - Relational
- The majority of the information “world” is unstructured:
  - Clarabridge has estimated that 80% of business information is unstructured.
  - IBM Research has estimated it at 85%.

CLARABRIDGE

IBM **Research**

Creating Structured Data from Unstructured

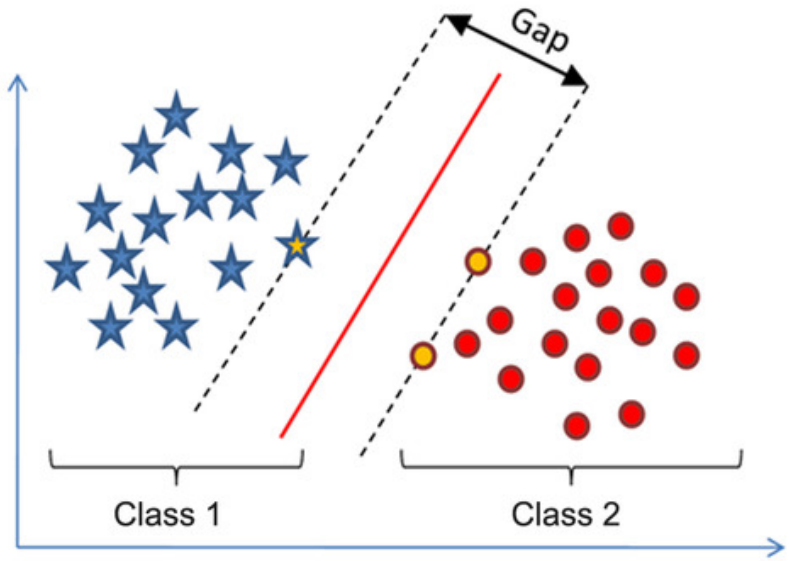
And it tends to just get worse.



Using ML in NLP

Supervised ML is used for

- Training parsers
  - Ngram trainers
  - Naïve Bayes
- Training classifiers
  - Naïve Bayes
  - Support Vector Machine (SVM)

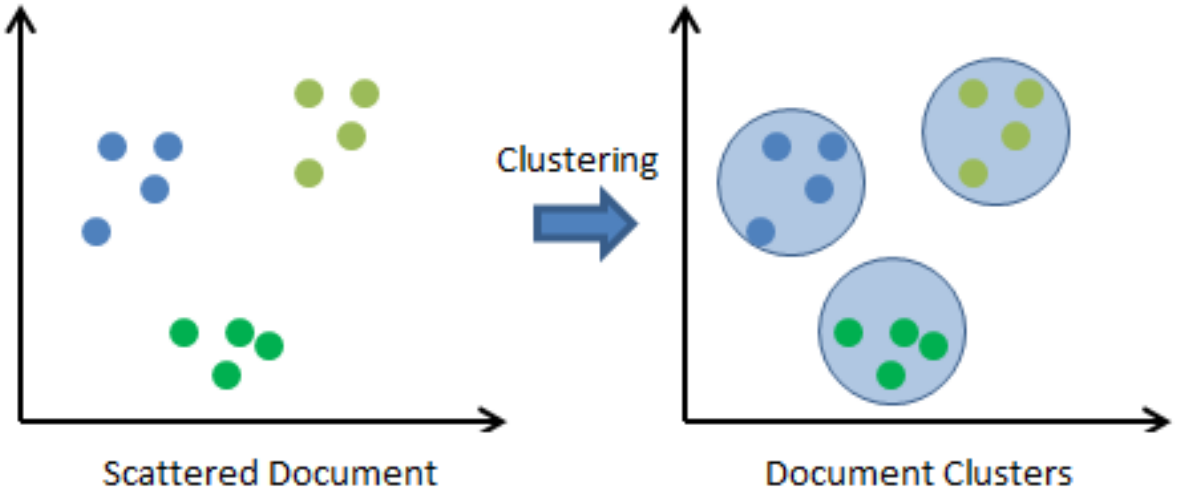


SVM

Using ML in NLP

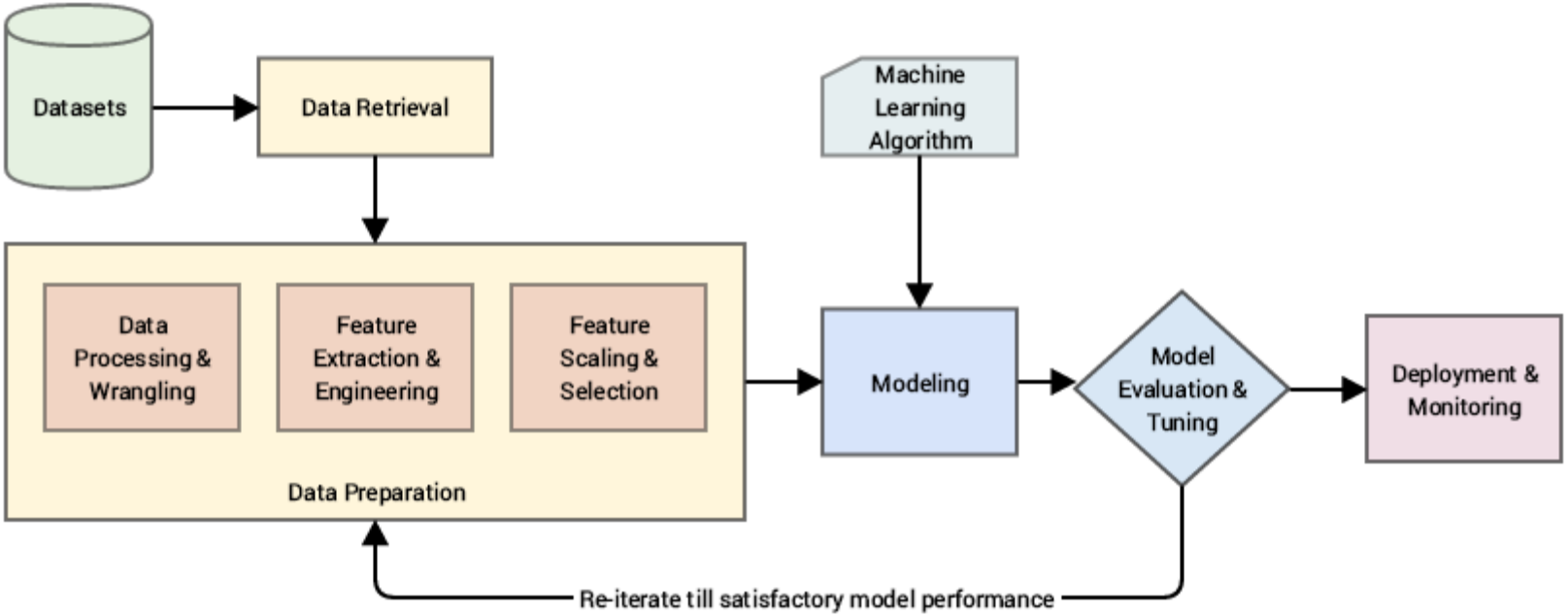
Unsupervised ML is used for document clustering:

- K-means
- Affinity propagation
- Hierarchical
  - Agglomerative
  - Divisive



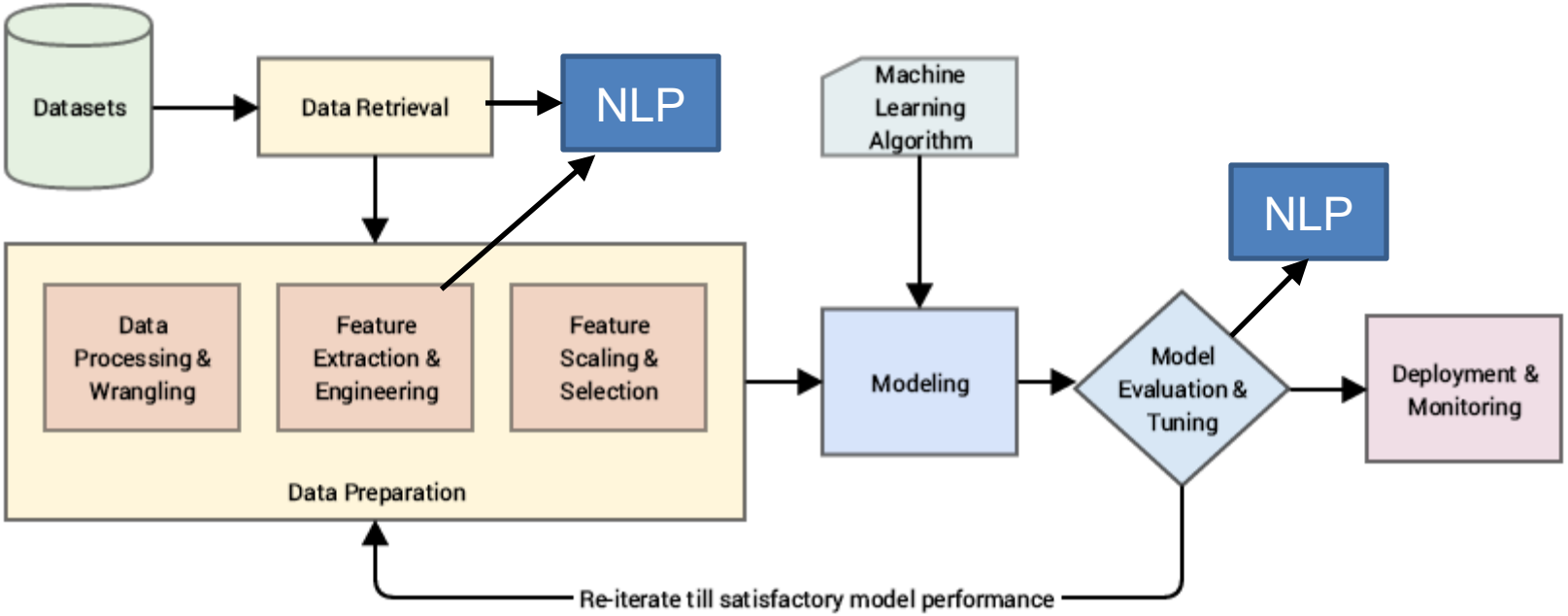
Using NLU for Features

- We can use NLU to bootstrap feature engineering.
- We cycle between human brainstorming and machine-driven suggestions/refinements.



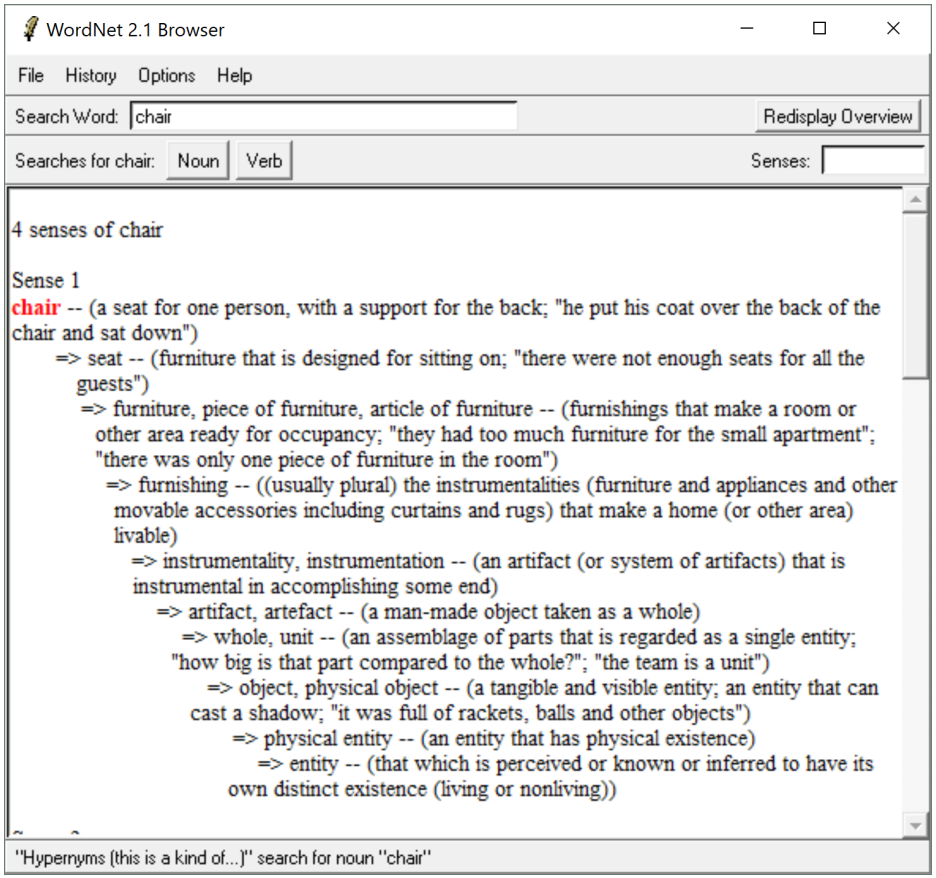
Using NLU for Features

- We can use NLU to bootstrap feature engineering.
- We cycle between human brainstorming and machine-driven suggestions/refinements.



Using NLU for Features

- We can use NLU to perform automated feature extraction:
  - Synsets
  - Hypernym trees
  - Filtering of above by pseudorelevance feedback



NLG as Presentation Layer

- We can create suggested NL (“human-friendly”) labels for clusters.
- For example we can create a micro-grammar template, such as <weighted adjective> <node name:plural>

On having clustered web pages that are all cooking recipes, this might give us cluster names such as “cheesy snacks,” “spicy snacks,” “easy desserts,” “fancy desserts,” etc.

NLG as Presentation Layer

- We can create templated synopses of automated data analyses.

BUSINESS DAY | UNBOXED

In Case You Wondered, a Real Human Wrote This Column

By STEVE LOHR SEPT. 10, 2011

"WISCONSIN appears to be in the driver's seat en route to a win, as it leads 51-10 after the third quarter. Wisconsin added to its lead when Russell Wilson found Jacob Pedersen for an eight-yard touchdown to make the score 44-3 ... . "

Those words began a news brief written within 60 seconds of the end of the third quarter of the Wisconsin-U.N.L.V. football game earlier this month. They may not seem like much — but they were written by a computer.

NLU for Validating Discrete Data

- Marketing technology distinguishes between “declared” and “inferred” data.
- We are increasingly aware of the need to use inferred data to validate declared data.
- Doing so, at scale, requires NLP.

"I'm an honest person but when I take an online survey, I'm a big liar."

DataScience@SMU

Working in NLP: Jobs Using NLP

Natural Language Processing

Jobs Using NLP

Jobs that utilize NLP include:

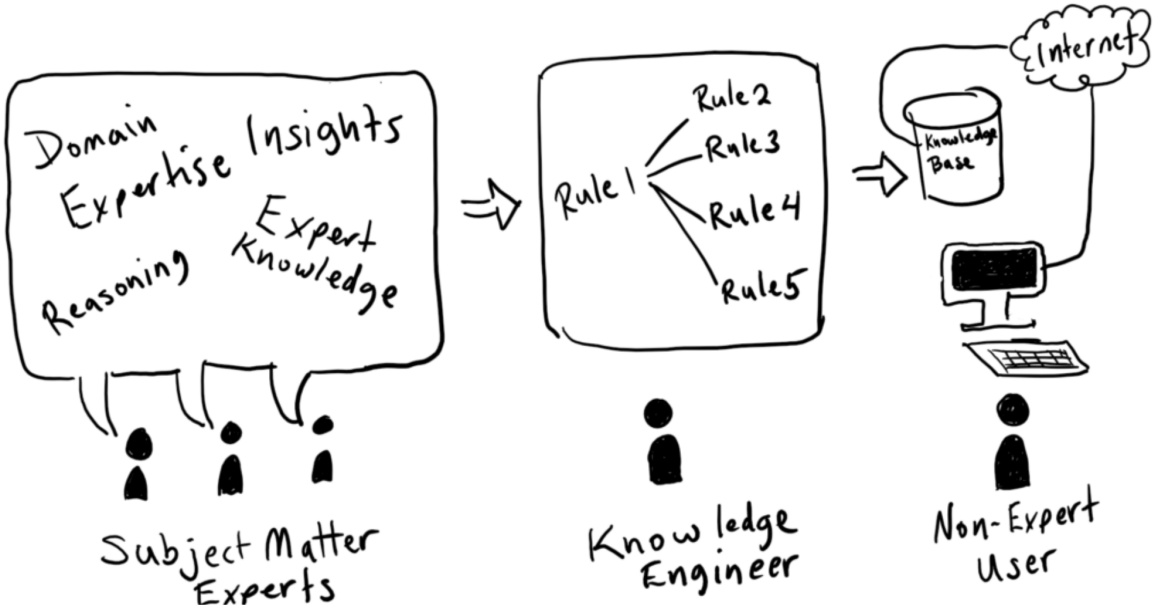
- Software engineer
- Knowledge engineer
- Data scientist
- DBA
- Applied linguistics researcher
- Cognitive scientist
- Marketing technologist

NLP Software Engineer

- Takes up the types of projects discussed in the course, and pushes them much further
- Develops any variety of applications that we have touched upon
- Becomes a master of one or a few particular methods—perhaps ontologies, or semantic parsing, or the ML-related aspects of NLP
- Is likely part of a team that is mostly non-NLP engineers

Knowledge Engineer

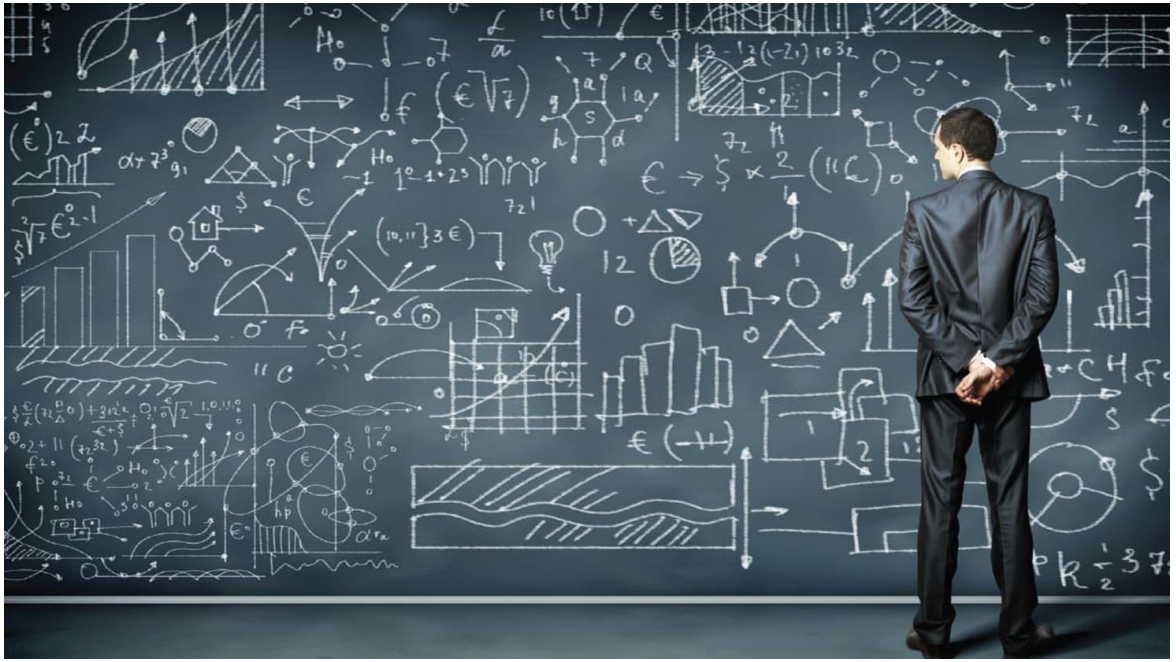
- Interfaces regularly with AI engineers and SMEs
- Demands high “people skills” (often compared to “Rogerian therapy” or “Socratic method”)
- Codifies trade secret knowledge of an organization
- Can have greater impact than the “core” AI itself



Data Scientist

A hybrid category comprising:

- Researcher
- Statistician
- Software engineer
- AI practitioner
- Visualizer
- Communicator



Database Administrator

While this role historically did not use NLP, it is beginning to do so. Activities include:

- Choosing between database paradigms to fit projects/organizations
- Designing data models
- Implementing data governance policies
- Performing ETL ("extract, transform, load")
  - Using NLU ("natural language understanding")
- Helping generate reports
  - Using NLG ("natural language generation")

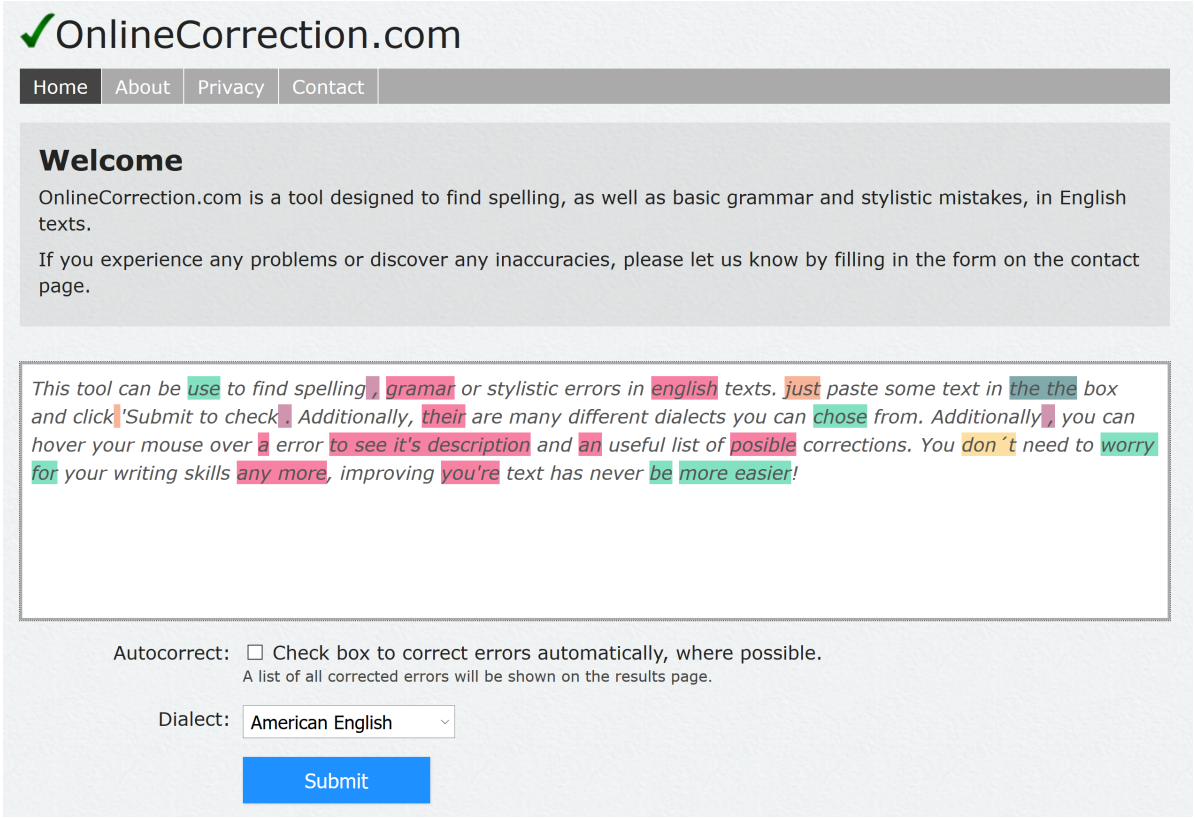
Applied Linguistics Researcher

- Applies linguistics to helping/managing real people
- Chief application space is education: help teachers better educate children in language arts
  - ESL
  - Learning impaired
  - Bilingual education

Applied Linguistics Researcher

NLP interacts with this field in numerous ways:

- Automated writing evaluation
- Online reading help
- Automated grade-level estimation
- Nonnative speaker support



Cognitive Scientist

- Joins an interdisciplinary research project
- Goes shoulder-to-shoulder with neuroscientists, psychologists, linguists, anthropologists, and philosophers
- Studies the mechanisms of human thinking, deciding, speech acts, language acquisition, and everything we do with words
- Read outside your discipline; see if your findings are confirmed or disconfirmed by others using approaches outside of AI.

Marketing Technologist

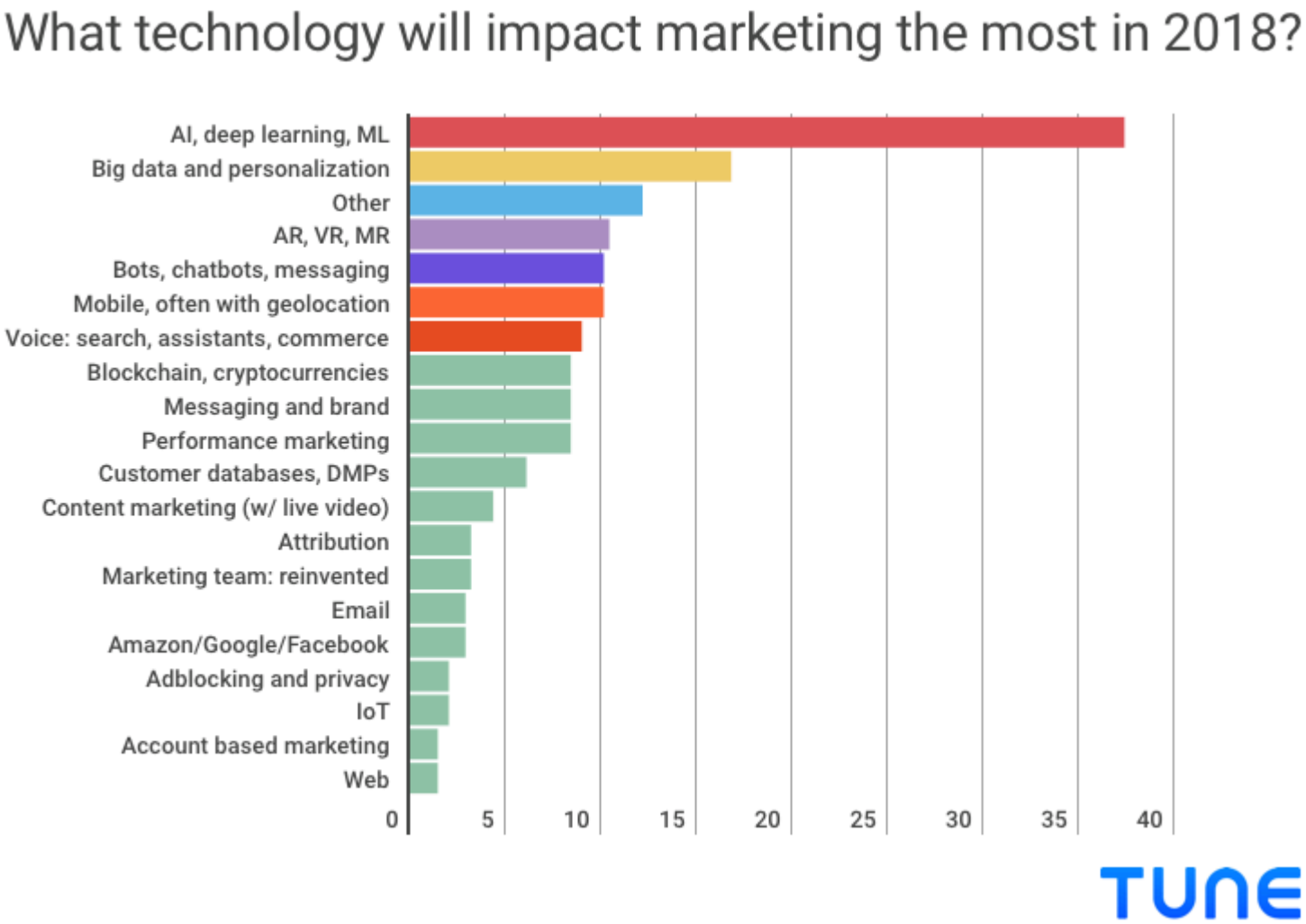
Tries to get the right product or service in front of the right person in the right place at the right time

Reading an article about travel? You'll probably see ads for travel pop on your screen shortly after. Check out a musician's website? You might get some ads for music popping up soon too.

Marketing Technologist

- Makes ads, emails, and sponsored content more relevant and more timely
- Uses NLP to build rich profiles of consumers or customers so that brands know their audience better

Marketing Technologist



TUNE

DataScience@SMU

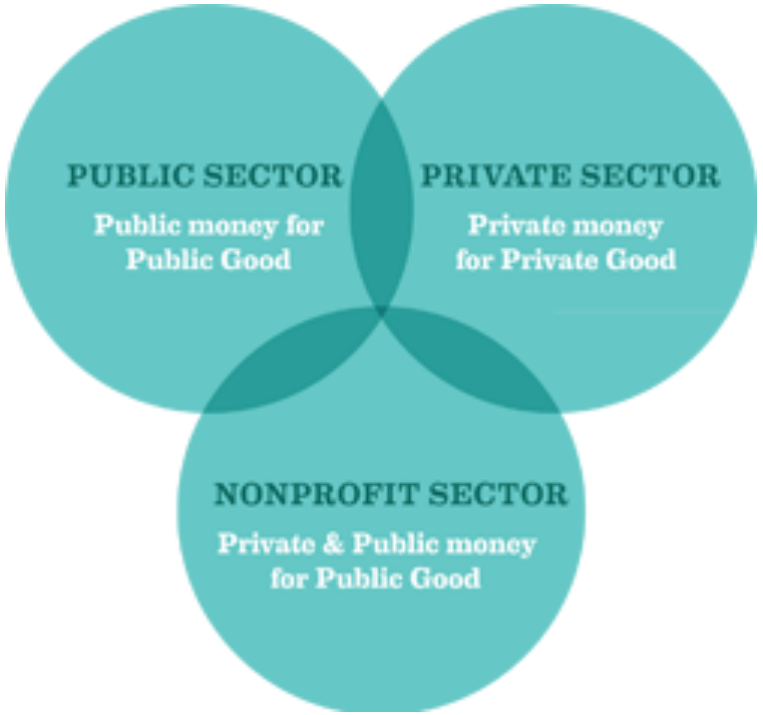
Working in NLP: Sectors Using NLP

Natural Language Processing

Sectors That Use NLP

Public and private sectors that use NLP include:

- Information services
- eCommerce
- Customer service desks
- Law enforcement or military
- Legal
- Business intelligence
- Consumer devices
- Embedded technologies
- Publishing
- Research institutes

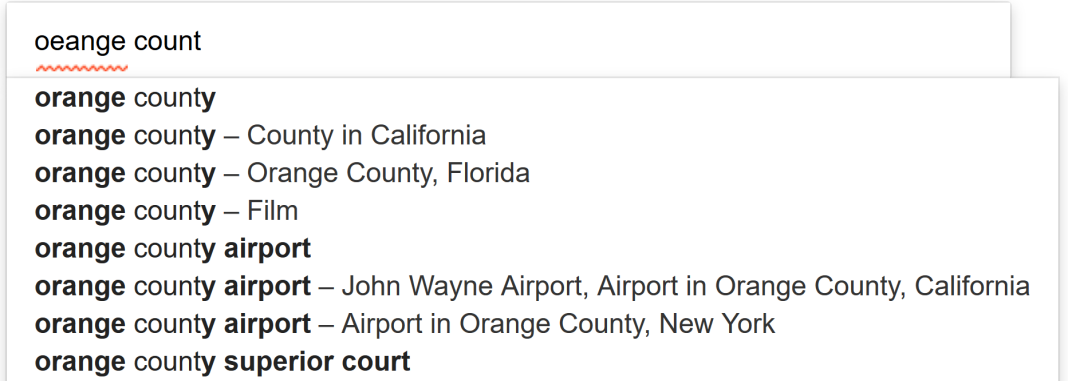


Information Services

Keyword search is rapidly being supplemented by NLP at various levels:

- Query repair
- Query refinement
- Results post-processing

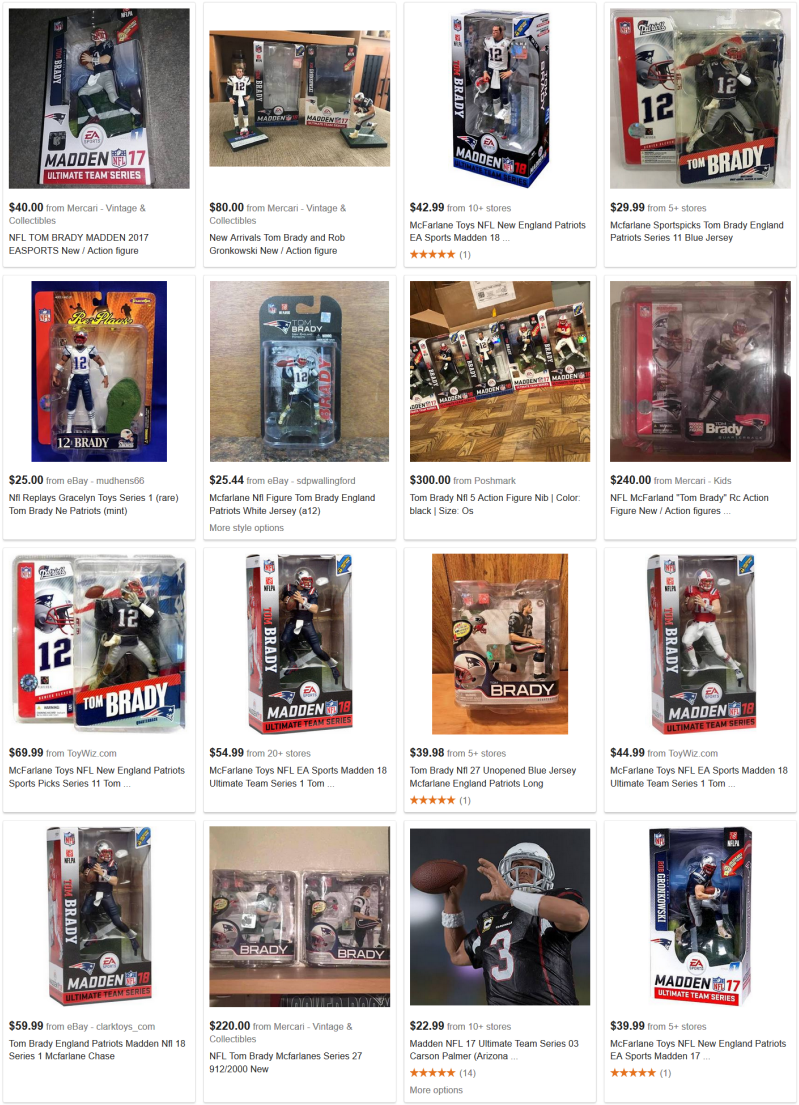
These efforts are ongoing not only at the famous search engines but also in vertical site search and special services like Lexis-Nexis.



eCommerce

Multiple manufacturers, distributers, and vendors describe products and features with variant (and errant) language.

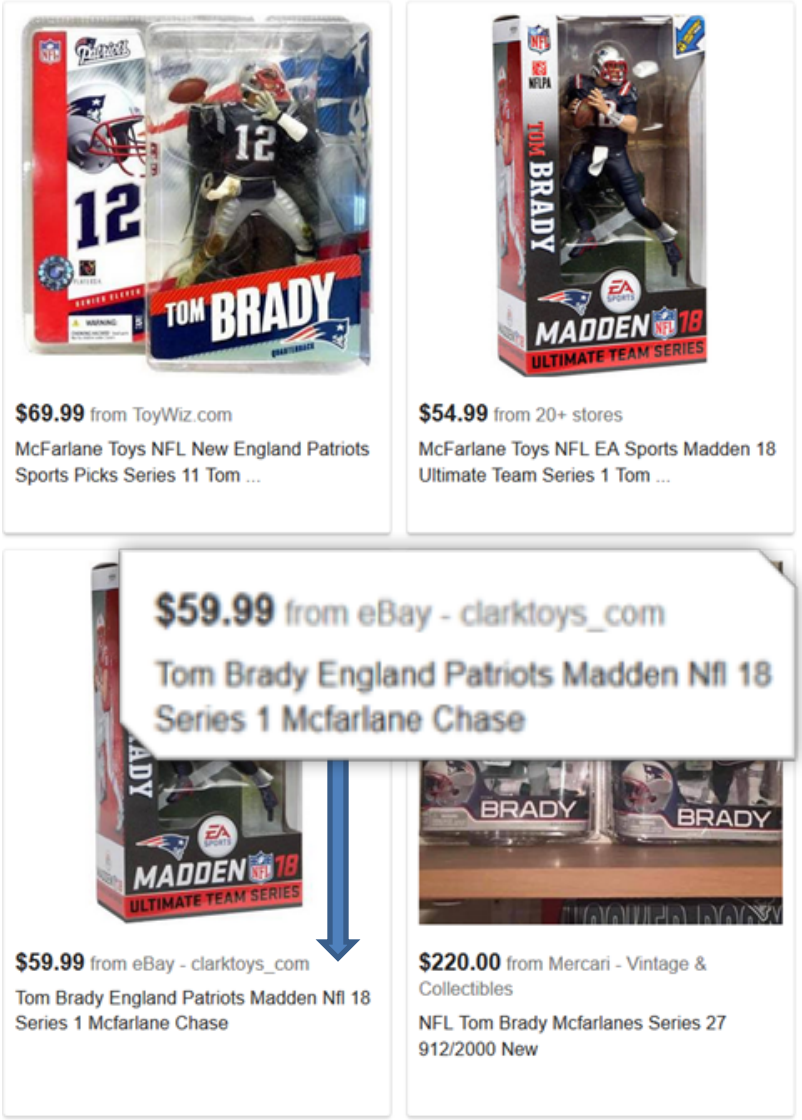
Can you tell by the text description which of these are the same?



eCommerce

Multiple manufacturers, distributers, and vendors describe products and features with variant (and errant) language.

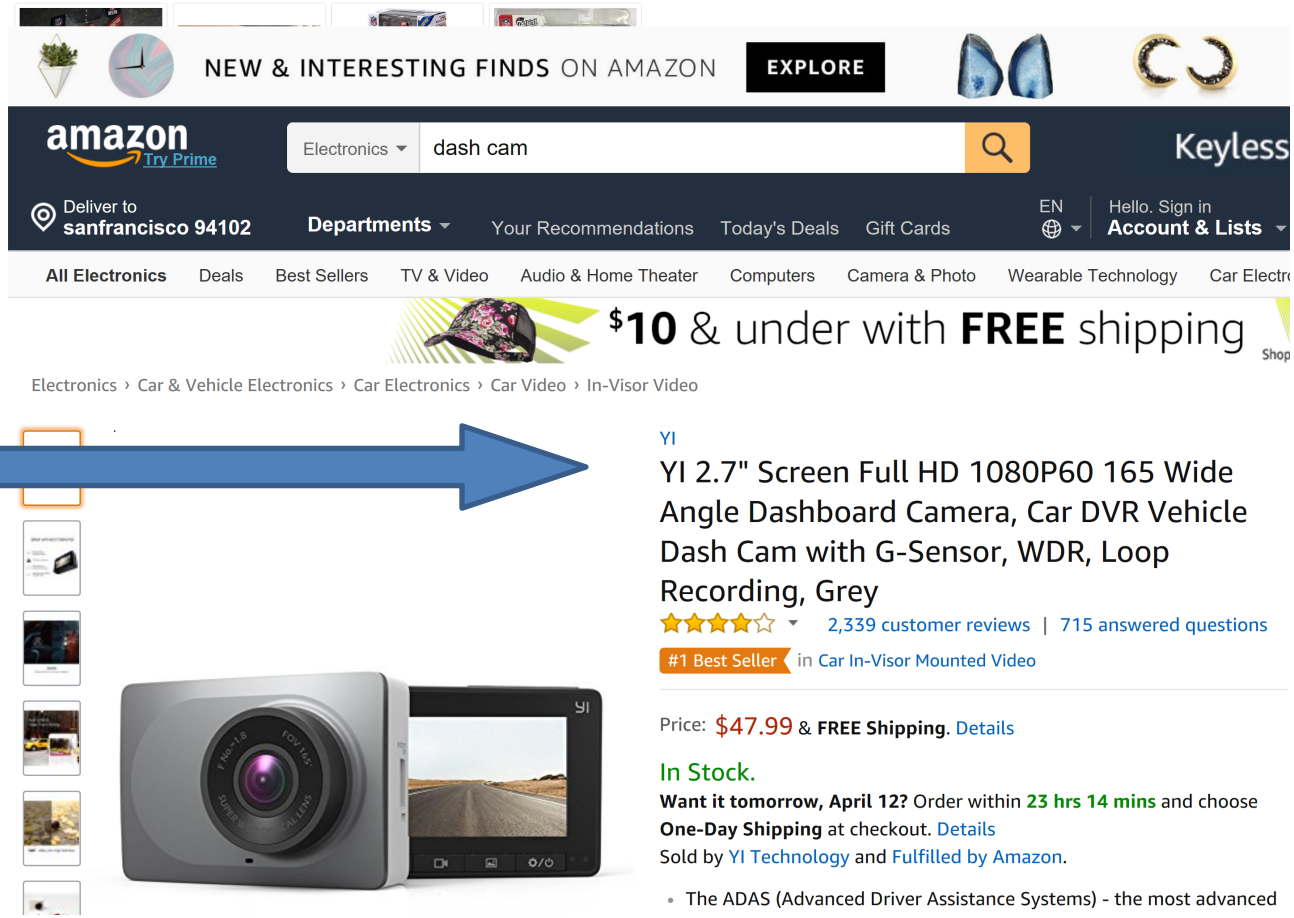
Can you tell by the text description which of these are the same?



eCommerce

Also, product descriptions are often not unitized.

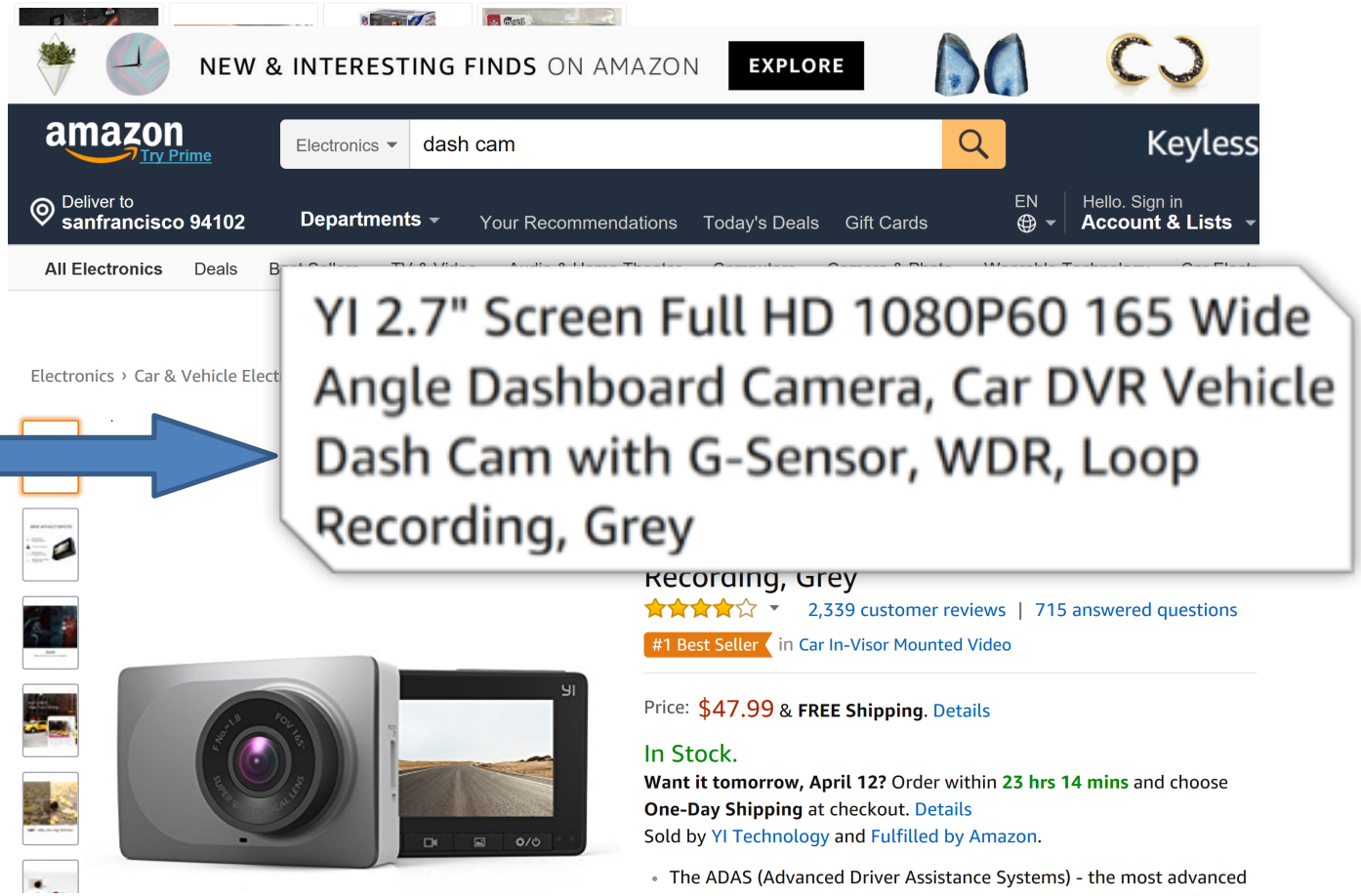
How many different data fields are crammed into the “product name” field?!



eCommerce

Also, product descriptions are often not unitized.

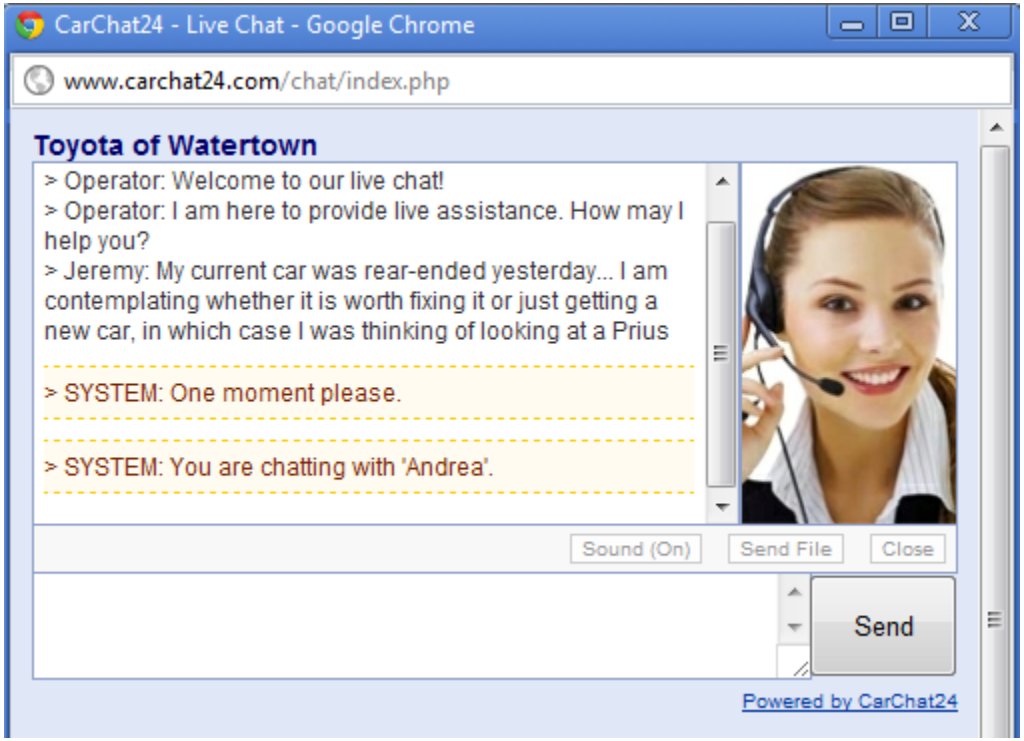
How many different data fields are crammed into the “product name” field?!



Customer Service Desks

Every customer service desk has repositories of:

- Call rep's unstructured text notes
- Text submissions from the customer
- Chat sessions between customer and rep



Customer Service Desks

These repositories are an NLP gold mine for:

- Text mining to spot trends, to inform best practices
- Case-based reasoning
- Symptom-syndrome analysis

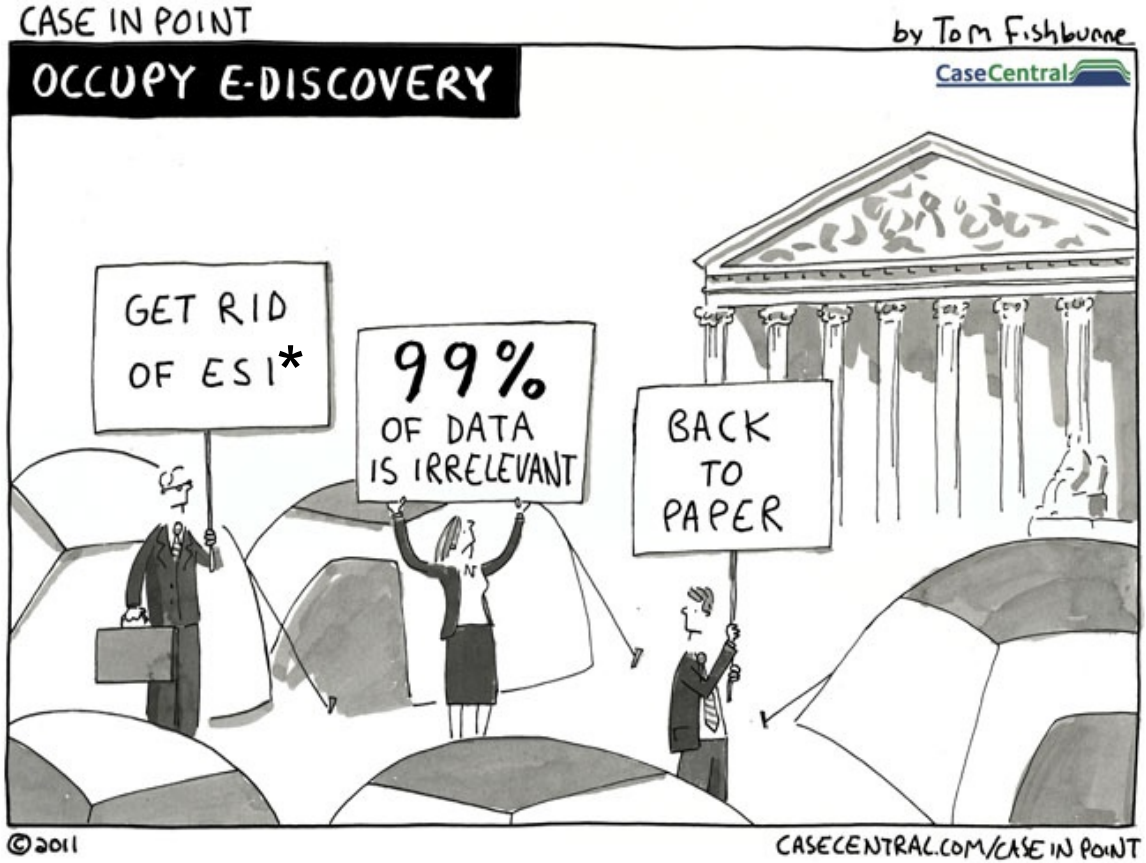
Law Enforcement or Military

NLP is needed and increasingly utilized for creating (without false positives):

- Time-sensitive fuzzy search for named entities
- Disaster alerts
- Terrorist activity alerts
- Other perpetrator detection

Legal

Legal teams have to comb mountains of documents and still do most of it manually or with merely Boolean keyword search.



*electronically stored information

Legal

Legal teams have to comb mountains of documents, and still do most of it manually or with merely Boolean keyword search.

The industry is beginning to apply NLP to:

- eDiscovery
- Scan for candidate infringers of patent/trademark IP
- Semi-automated (or “bootstrapped”) document construction

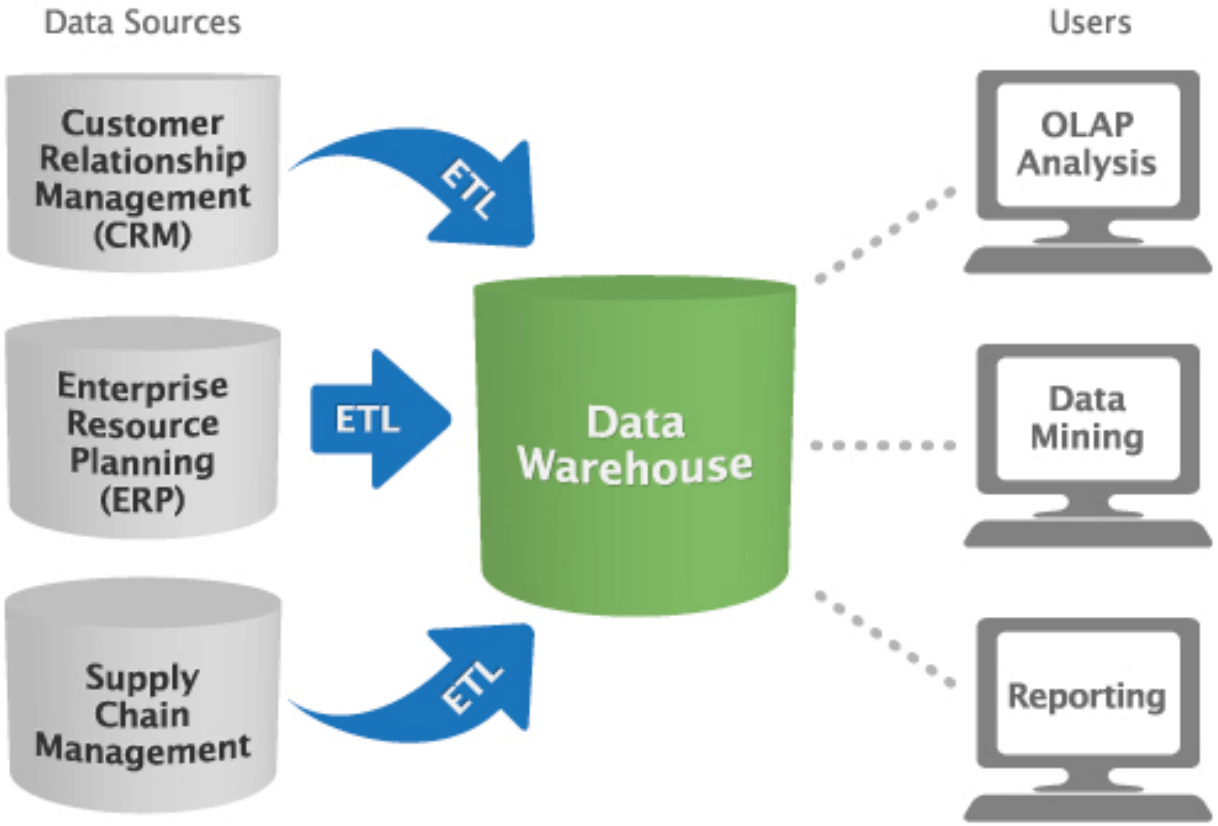
DataScience@SMU

Working in NLP: Sectors Using NLP

Natural Language Processing

Business Intelligence (BI)

Many data sources flowing into the typical BI data warehouse contain large amounts of unstructured text:

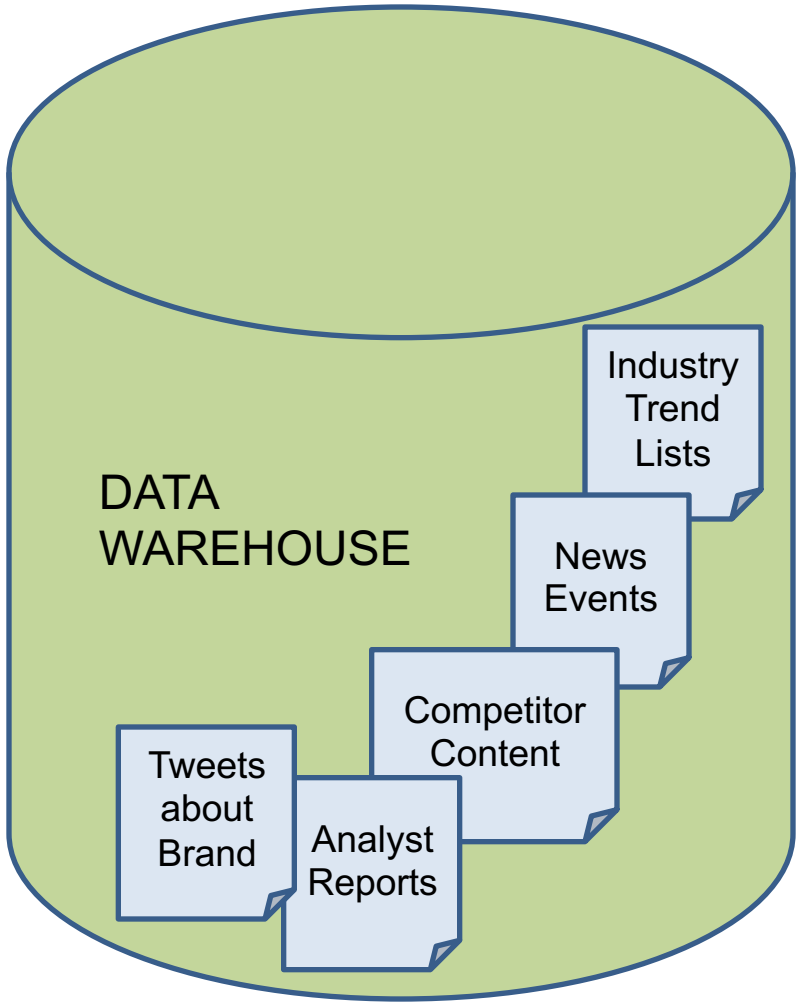


Business Intelligence (BI)

BI people are increasingly interested in automation help for analyzing:

- Competitor's marketing literature, SEM, and catalogs
- Consumer reviews and professional reviews
- Industry news
- Analyst reports
- Customer feedback records
- SM posts about your brand

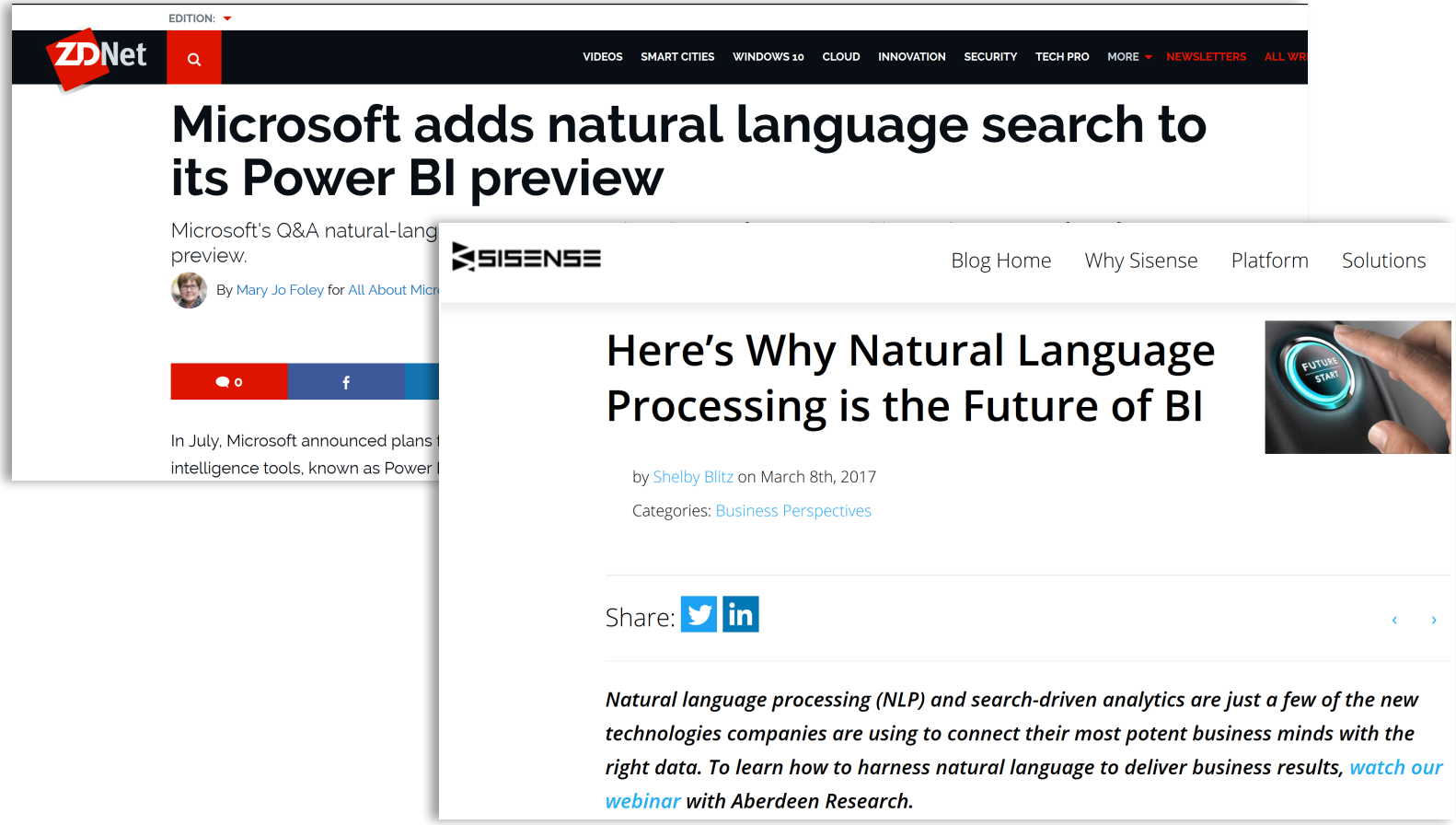
These sources need to be cross-referenced or correlated with internal KPI's* and reanalyzed all the time.



**key performance indicators*

Business Intelligence (BI)

NLP is being identified as the way forward for BI.



Consumer Devices

New consumer devices connect to a cloud-based NLI (natural language interface)

- Amazon Echo, Dot
- Google Home Smart Speaker
- Apple HomePod

Embedded Technologies

In the car:

- FORD Sync
- GM OnStar

Wearables/portables:

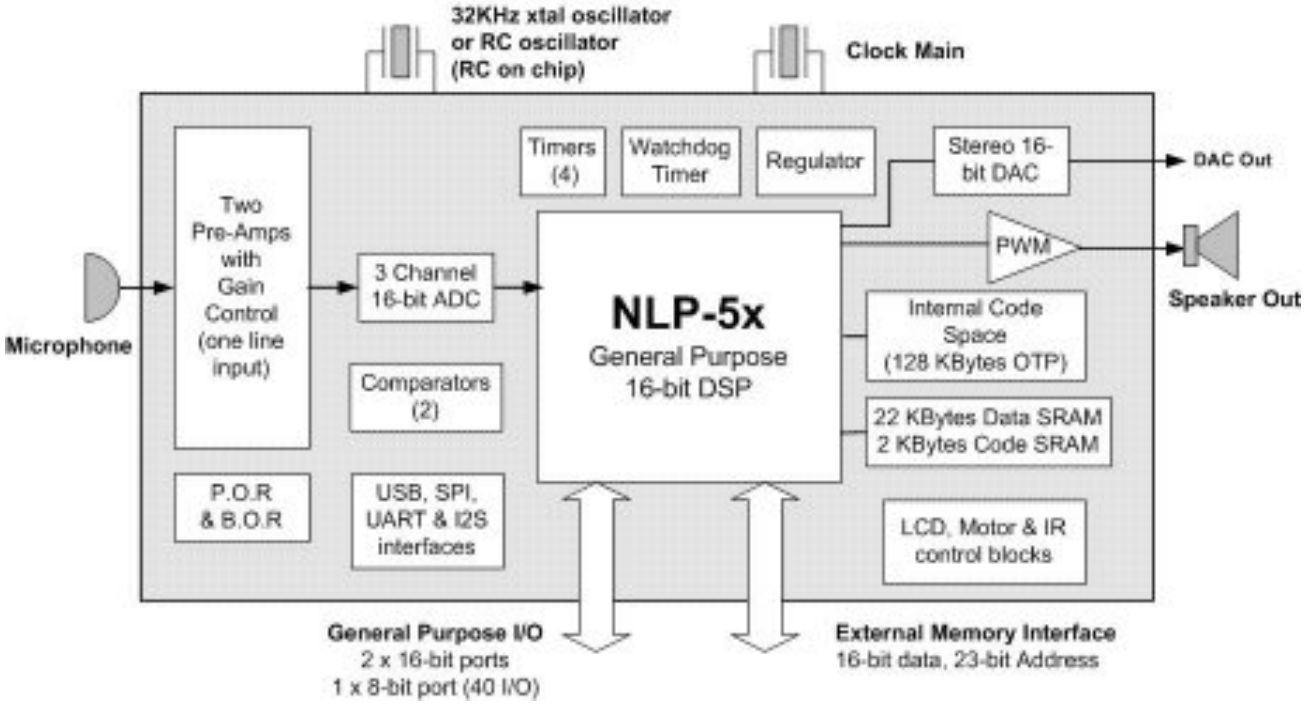
- Phones
- Watches/bands

How many other device types can you think of that could use localized in-memory NLP?

Embedded Technologies

Examples of dedicated chips for in-memory NLP (as of 2018):

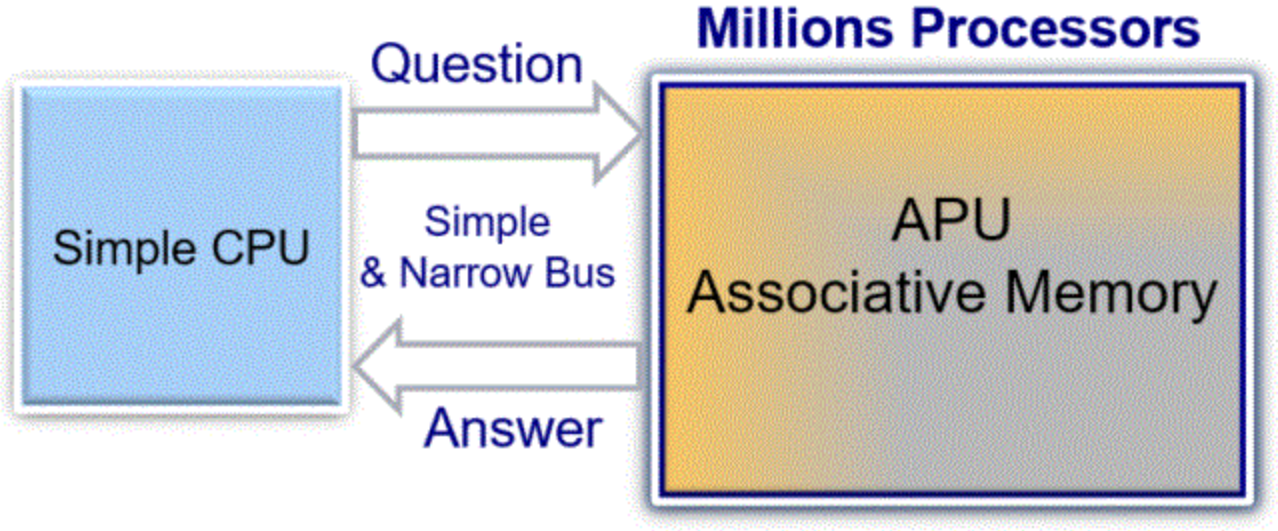
- The NLP-5x from Sensory



Embedded Technologies

Examples of dedicated chips for in-memory NLP (as of 2018):

- The NLP-5x from Sensory
- GSI Technology's APU (Associative Processing Unit)



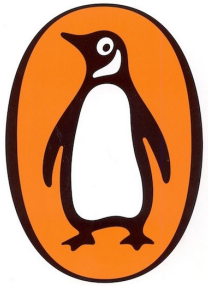
Publishing and Media

Hot button detection: News, blog, TV, podcast, magazine, and books publishers need to keep a finger on the pulse of audience interest

RANDOM HOUSE

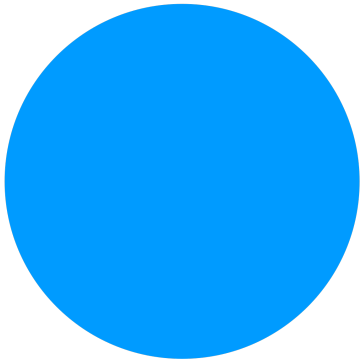
The New York Times

c|net



HEARST magazines UK

The Economist



USA TODAY

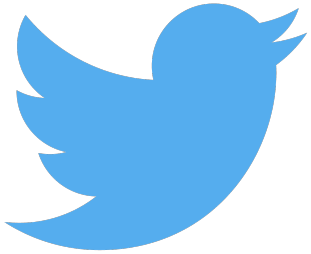
A GANNETT COMPANY

Publishing and Media

Hot button detection: NLP can do that by spotting emerging topic trends and identifying underserved content areas from these sources.

amazon

tumblr.



DataScience@SMU

Working in NLP: Organizations Supporting NLP

Natural Language Processing

Organizations That Relate to NLP

Associations supporting NLP include

- ACM
- IEEE
- AAAI
- IJCAI
- AAAL
- ICLA

ACM

Association for Computing Machinery

https://www.acm.org/special-interest-groups/

- SIGKDD (Knowledge Discovery Group)
- SIGAI (AI Group)

IEEE

International EEE

Don’t be fooled by the name; within this organization is the “IEEE Computer Society.”

Technical Committees:

- TC on Semantic Computing
  https://tc.computer.org/tcsem/
- TC on Pattern Analysis and Machine Intelligence
  https://www.computer.org/web/tcpami/
- TC on Intelligent Informatics
  http://www.wi-lab.com/tcii/index.shtml
- TC on Data Engineering
  http://tab.computer.org/tcde/

Special Technical Committees
https://www.computer.org/web/stc/

- STC on Smart Computing
- STC on Big Data
- STC on Cloud Computing

AAAI

Association for the Advancement of Artificial Intelligence

aaai.org

- If you can join only one organization for NLP, you probably want it to be this one.
- Variety of conferences and workshops always including multiple NLP topics.

IJCAI

International Joint Conference on Artificial Intelligence Organization

http://www.ijcai.org/

Why go international?

- US/UK academic culture is still more compartmentalized (less interdisciplinary) than continental academia.
- EU has placed funding with differing priorities to US, and it shows in conferences.

AAAL

American Association for Applied Linguistics

www.aaal.org/

- Addresses language in the lives of individuals
- Interdisciplinary field—draws on a wide range of approaches from the humanities to the social sciences and computer science

ICLA

International Cognitive Linguistics Association

http://www.cognitivelinguistics.org/en

Excellent background on how the field came to be: http://www.cognitivelinguistics.org/en/about-cognitive-linguistics

ICLA INTERNATIONAL COGNITIVE LINGUISTICS ASSOCIATION

DataScience@SMU

Working in NLP: Organizations Supporting NLP

Natural Language Processing

Organizations That Relate to NLP

Associations supporting NLP include

- ACM
- IEEE
- AAAI
- IJCAI
- AAAL
- ICLA

ACM

Association for Computing Machinery

https://www.acm.org/special-interest-groups/

- SIGKDD (Knowledge Discovery Group)
- SIGAI (AI Group)

IEEE

International EEE

Don’t be fooled by the name; within this organization is the “IEEE Computer Society.”

Technical Committees:

- TC on Semantic Computing
  https://tc.computer.org/tcsem/
- TC on Pattern Analysis and Machine Intelligence
  https://www.computer.org/web/tcpami/
- TC on Intelligent Informatics
  http://www.wi-lab.com/tcii/index.shtml
- TC on Data Engineering
  http://tab.computer.org/tcde/

Special Technical Committees
https://www.computer.org/web/stc/

- STC on Smart Computing
- STC on Big Data
- STC on Cloud Computing

AAAI

Association for the Advancement of Artificial Intelligence

aaai.org

- If you can join only one organization for NLP, you probably want it to be this one.
- Variety of conferences and workshops always including multiple NLP topics.

IJCAI

International Joint Conference on Artificial Intelligence Organization

http://www.ijcai.org/

Why go international?

- US/UK academic culture is still more compartmentalized (less interdisciplinary) than continental academia.
- EU has placed funding with differing priorities to US, and it shows in conferences.

AAAL

American Association for Applied Linguistics

www.aaal.org/

- Addresses language in the lives of individuals
- Interdisciplinary field—draws on a wide range of approaches from the humanities to the social sciences and computer science

ICLA

International Cognitive Linguistics Association

http://www.cognitivelinguistics.org/en

Excellent background on how the field came to be: http://www.cognitivelinguistics.org/en/about-cognitive-linguistics

ICLA INTERNATIONAL COGNITIVE LINGUISTICS ASSOCIATION

DataScience@SMU