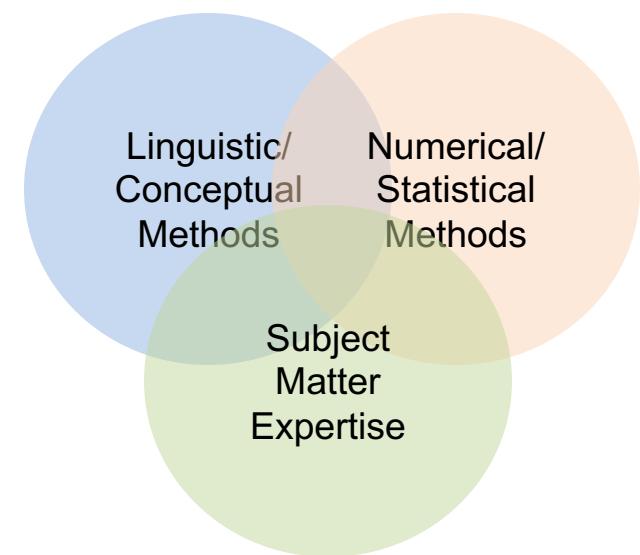


Working in NLP: NLP and Data Science

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%.

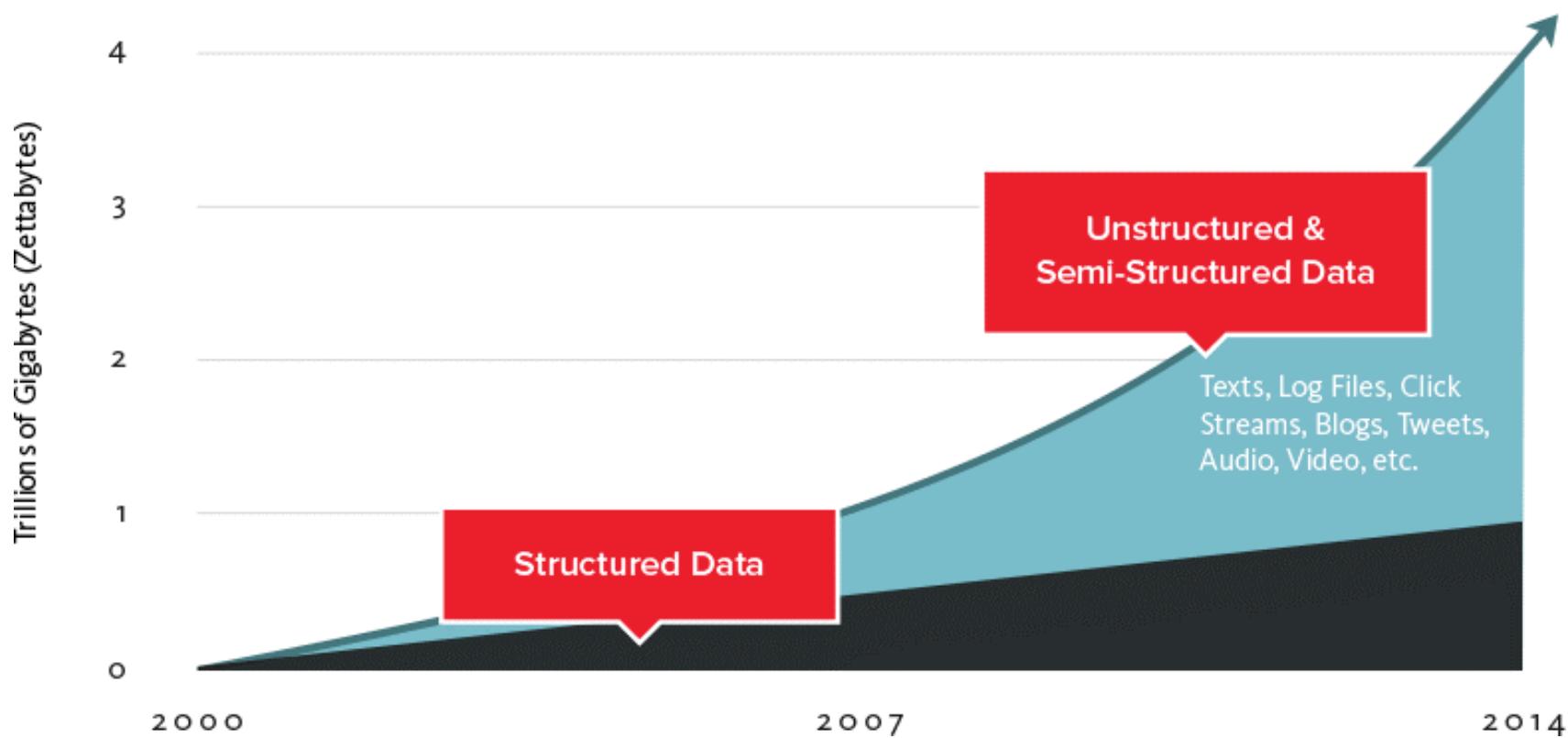


CLARABRIDGE

IBM Research

Creating Structured Data from Unstructured

And it tends to just get worse.

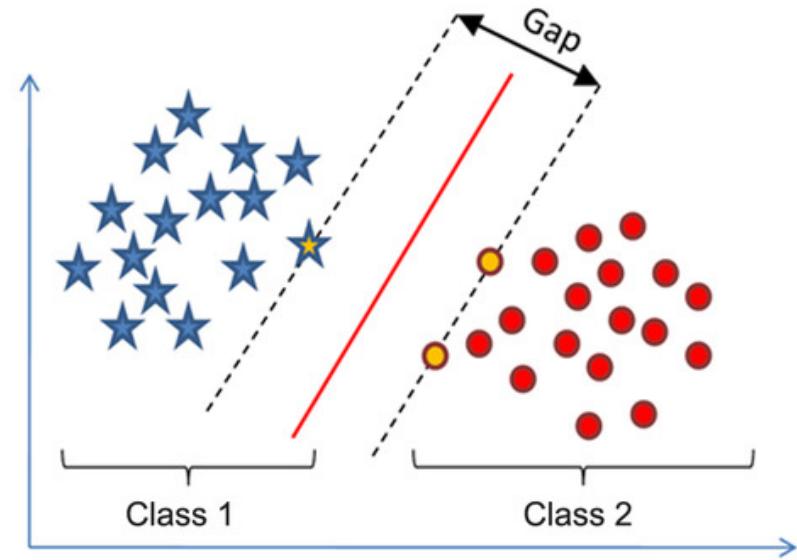


Using ML in NLP

Supervised ML is used for

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

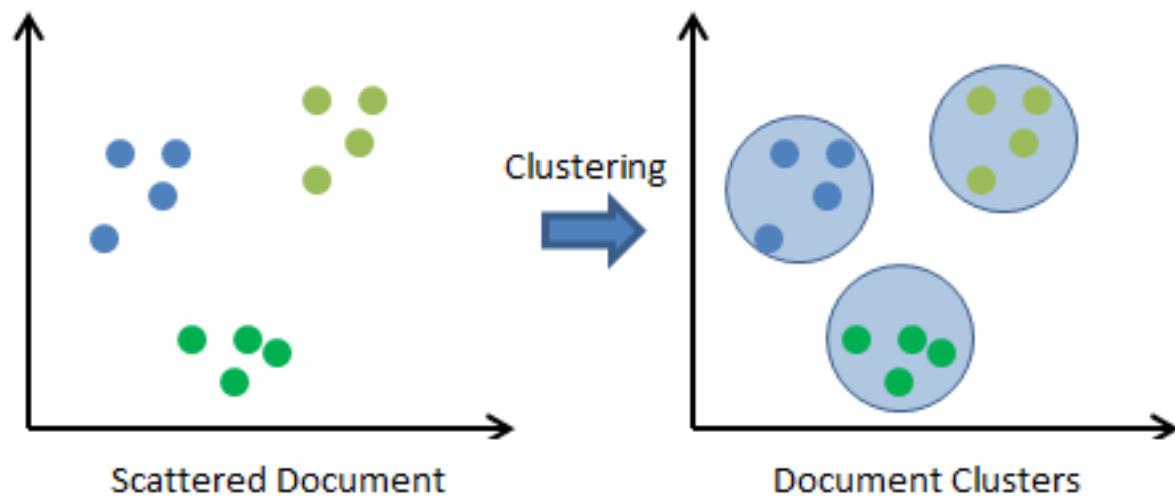
SVM



Using ML in NLP

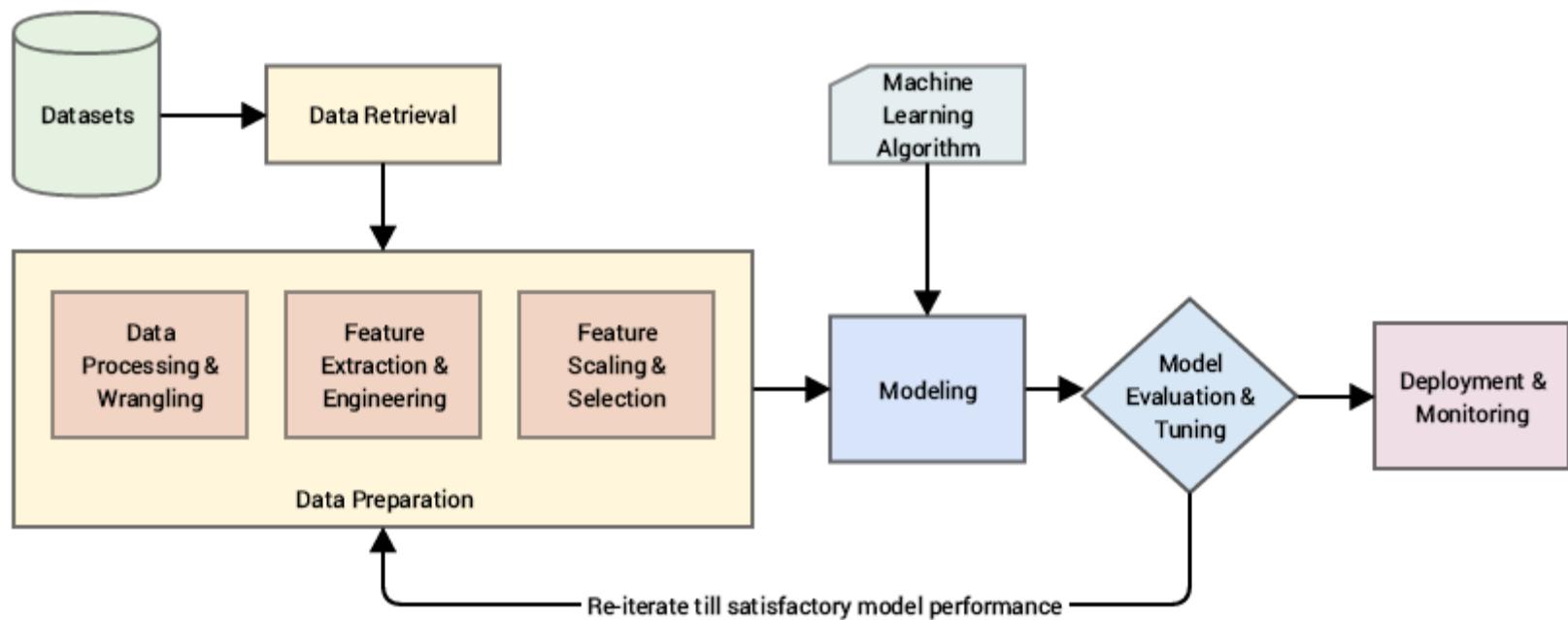
Unsupervised ML is used for document clustering:

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



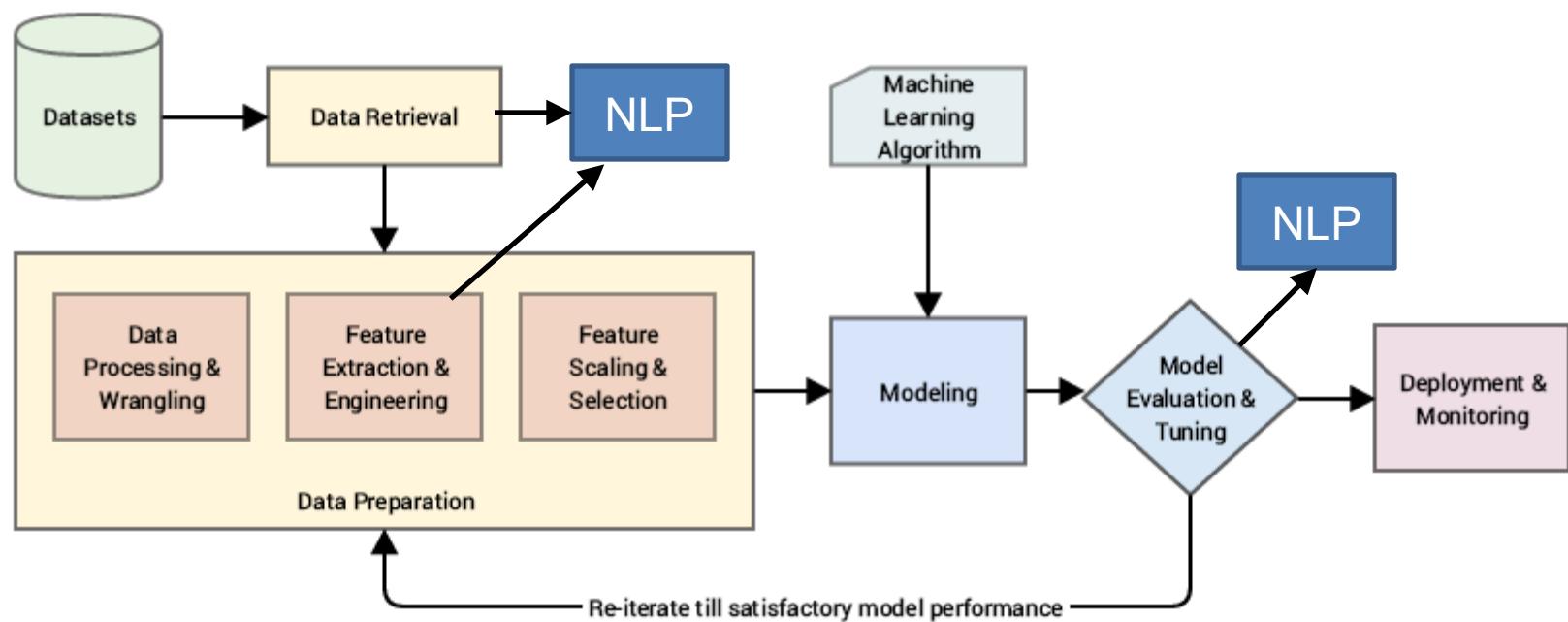
Using NLU for Features

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



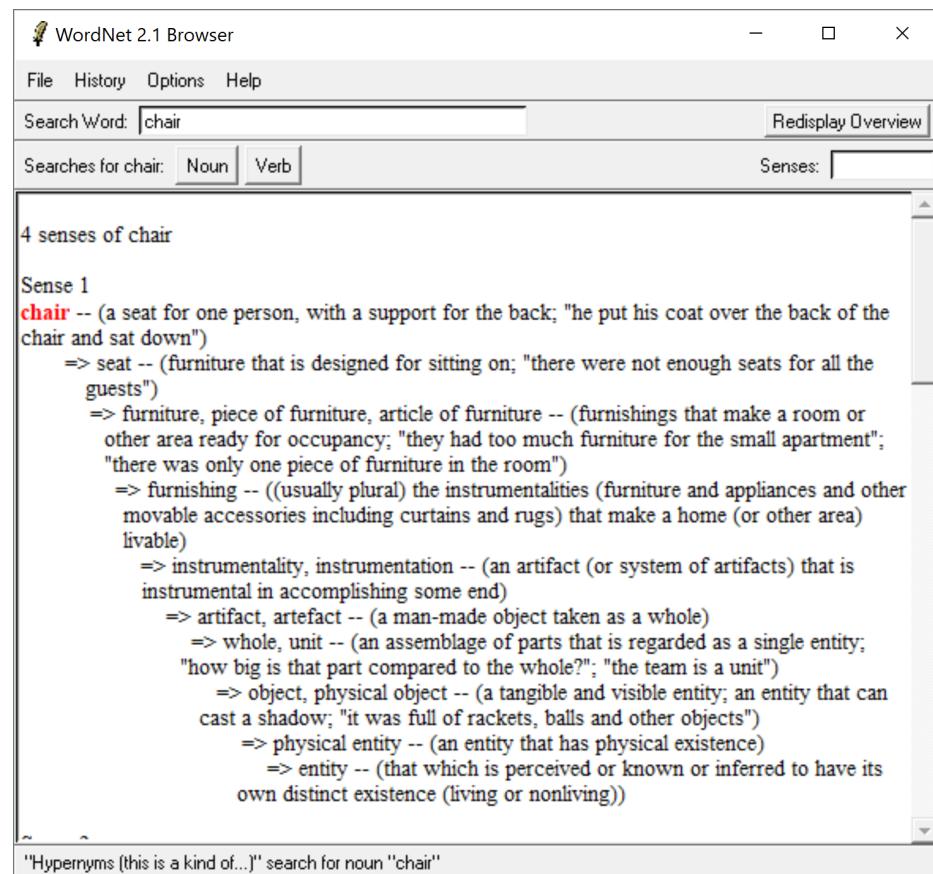
Using NLU for Features

- 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



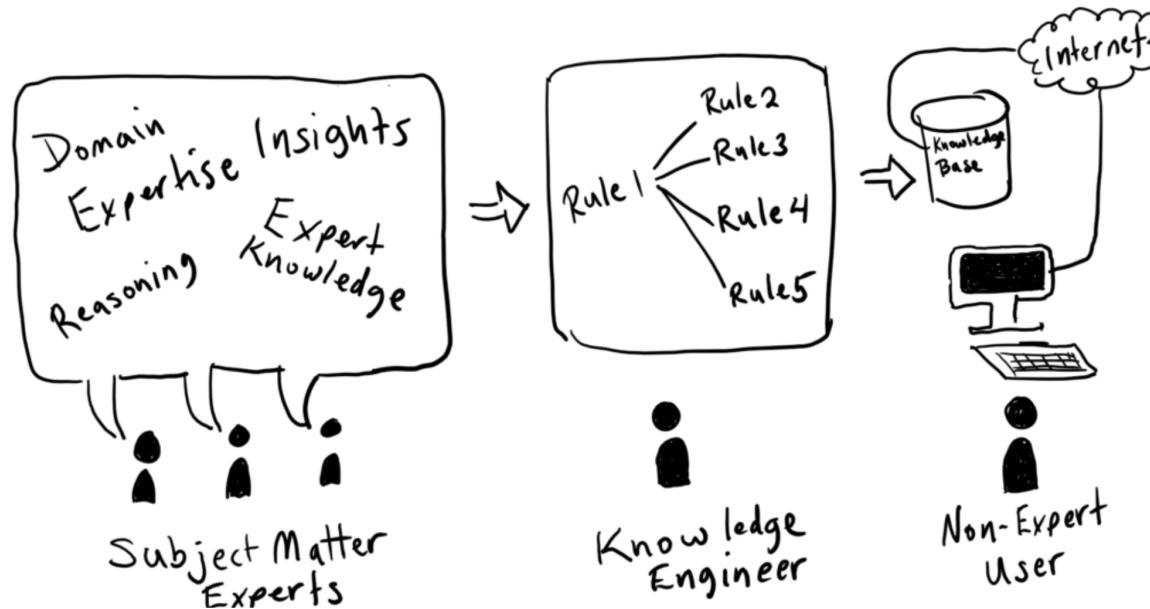
NLP Software Engineer

- 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

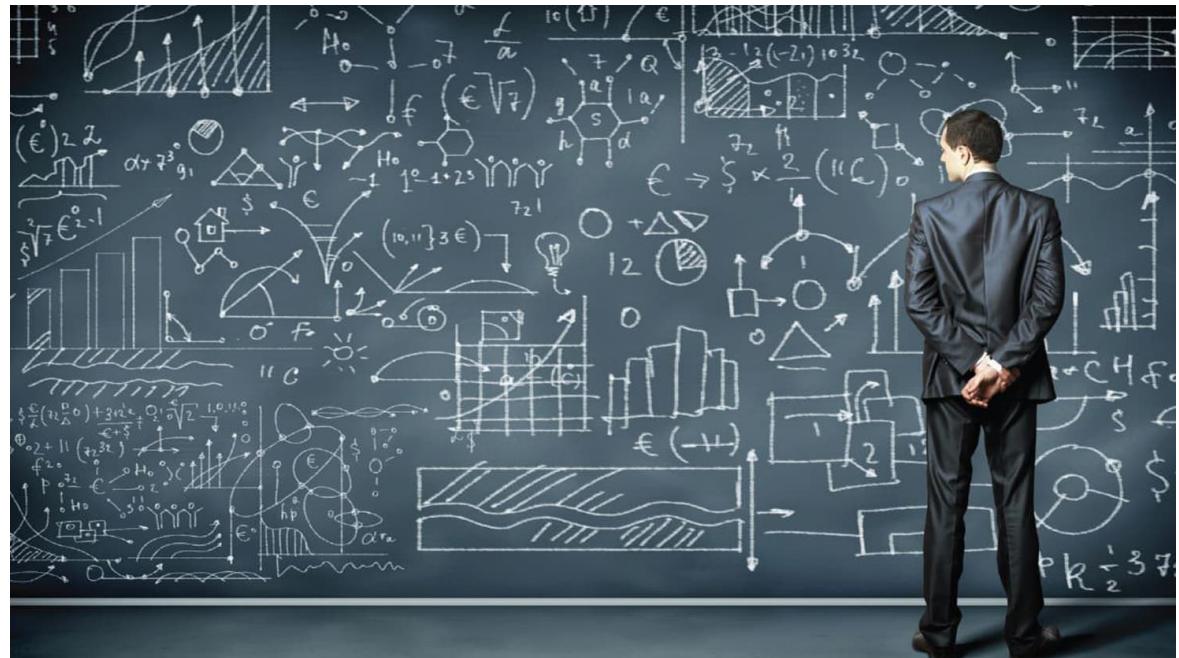
- Interfaces regularly with AI engineers and SMEs
- Demands high “people skills” (often compared to “Rogerian therapy” or “Socratic method”)
- Codifies trade secret knowledge of an organization
- Can have greater impact than the “core” AI itself



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

The screenshot shows the homepage of OnlineCorrection.com. At the top, there is a navigation bar with links for Home, About, Privacy, and Contact. Below the navigation bar, a section titled "Welcome" is displayed. It contains a brief description of the tool's purpose: "OnlineCorrection.com is a tool designed to find spelling, as well as basic grammar and stylistic mistakes, in English texts." It also encourages users to contact them if they experience any problems. A large text input area follows, containing a paragraph of text with various errors highlighted in red. A tooltip or explanatory text box is overlaid on this area, providing instructions on how to use the tool. At the bottom of the page, there are two input fields: "Autocorrect:" with a checkbox and a descriptive text, and "Dialect:" with a dropdown menu set to "American English". A blue "Submit" button is located at the bottom right.

This tool can be used to find spelling, grammar or stylistic errors in English texts. Just paste some text in the box and click 'Submit to check'. Additionally, there are many different dialects you can choose from. You can hover your mouse over a error to see its description and an useful list of possible corrections. You don't need to worry for your writing skills any more, improving your text has never been easier!

Autocorrect: Check box to correct errors automatically, where possible.
A list of all corrected errors will be shown on the results page.

Dialect: American English

Submit

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



You read an interesting article **on a free website**.

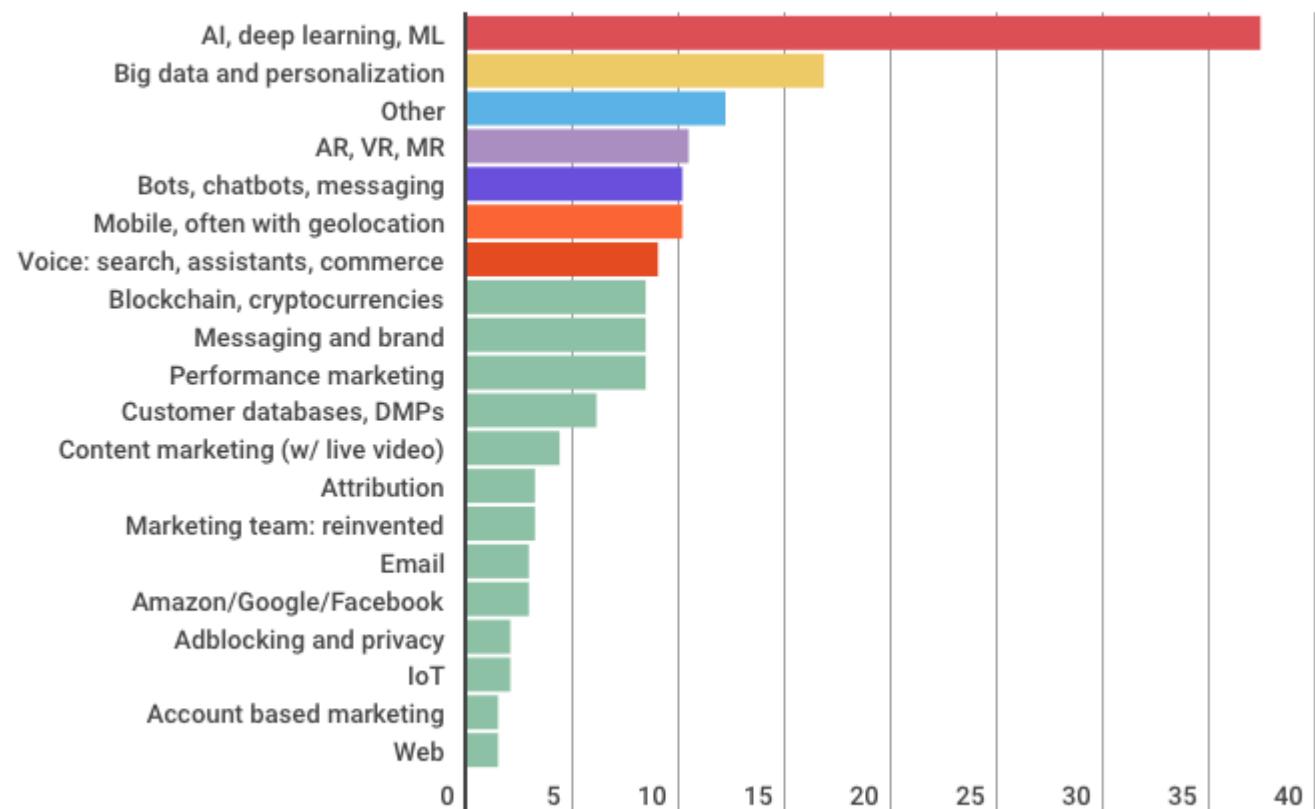
A cookie text file is saved to your browser and **lets advertisers categorize the browser** with an interest category.

Advertisers see the cookie data and **identify relevant ads** to place on websites visited by your browser.

Websites **charge advertisers** to place ads that suit your interest category. This allows websites to **offer content free to users**.

Marketing Technologist

What technology will impact marketing the most in 2018?



DataScience@SMU

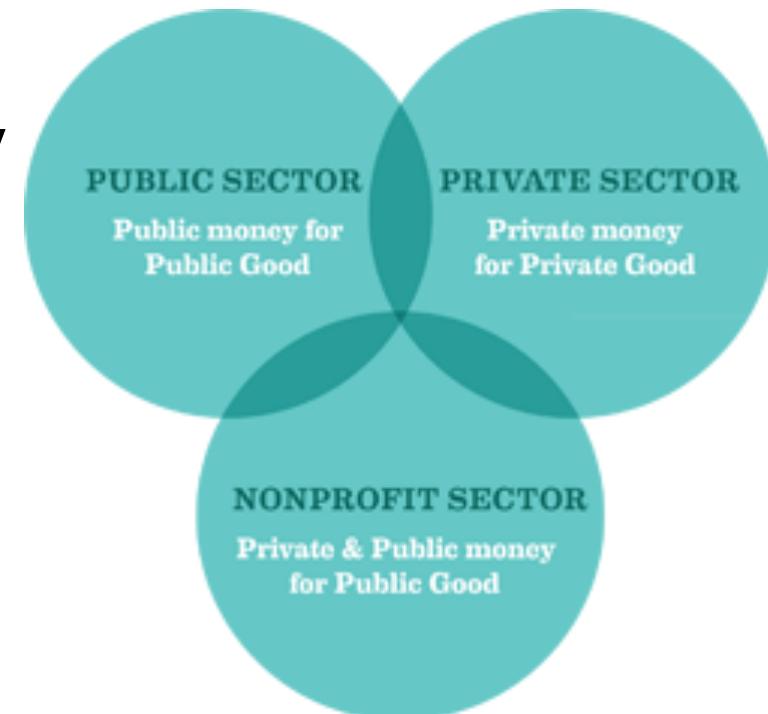
Working in NLP: Sectors Using NLP

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.



oeange count

~~~~~

**orange county**

**orange county** – County in California

**orange county** – Orange County, Florida

**orange county** – Film

**orange county airport**

**orange county airport** – John Wayne Airport, Airport in Orange County, California

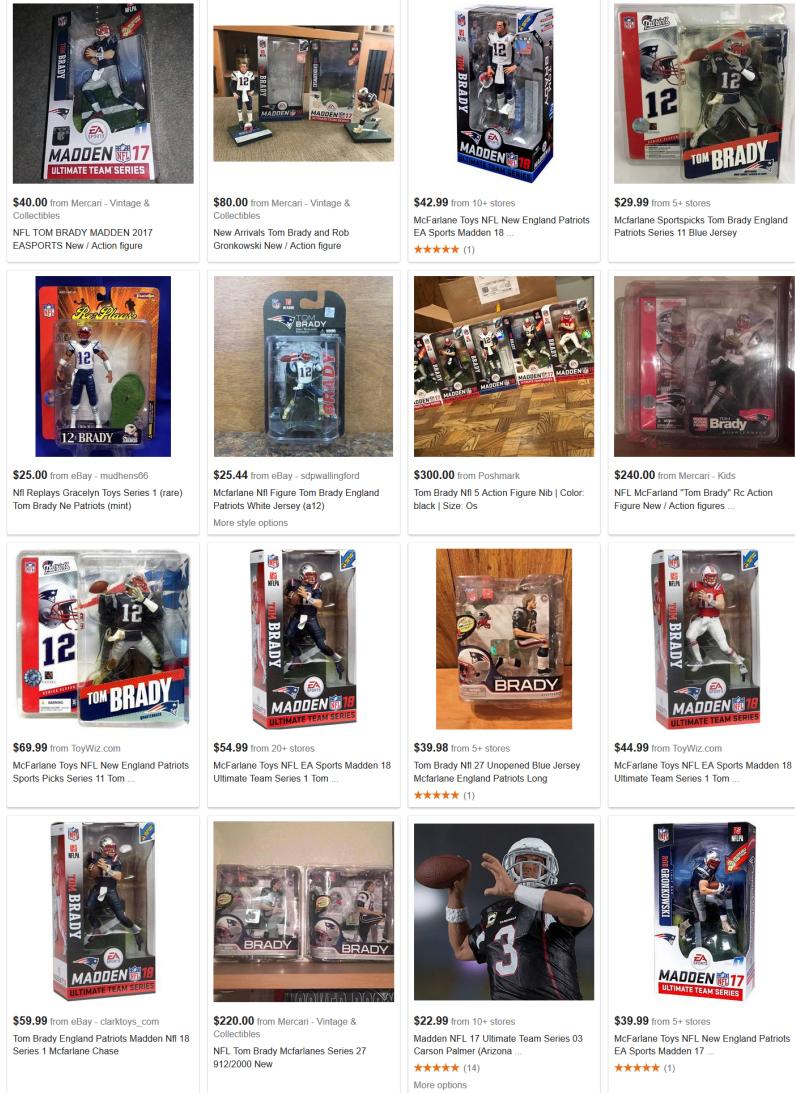
**orange county airport** – Airport in Orange County, New York

**orange county superior court**

# eCommerce

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



\$69.99 from ToyWiz.com

McFarlane Toys NFL New England Patriots Sports Picks Series 11 Tom ...



\$54.99 from 20+ stores

McFarlane Toys NFL EA Sports Madden 18 Ultimate Team Series 1 Tom ...



\$59.99 from eBay - clarktoys\_com

Tom Brady England Patriots Madden Nfl 18 Series 1 Mcfarlane Chase



\$220.00 from Mercari - Vintage & Collectibles

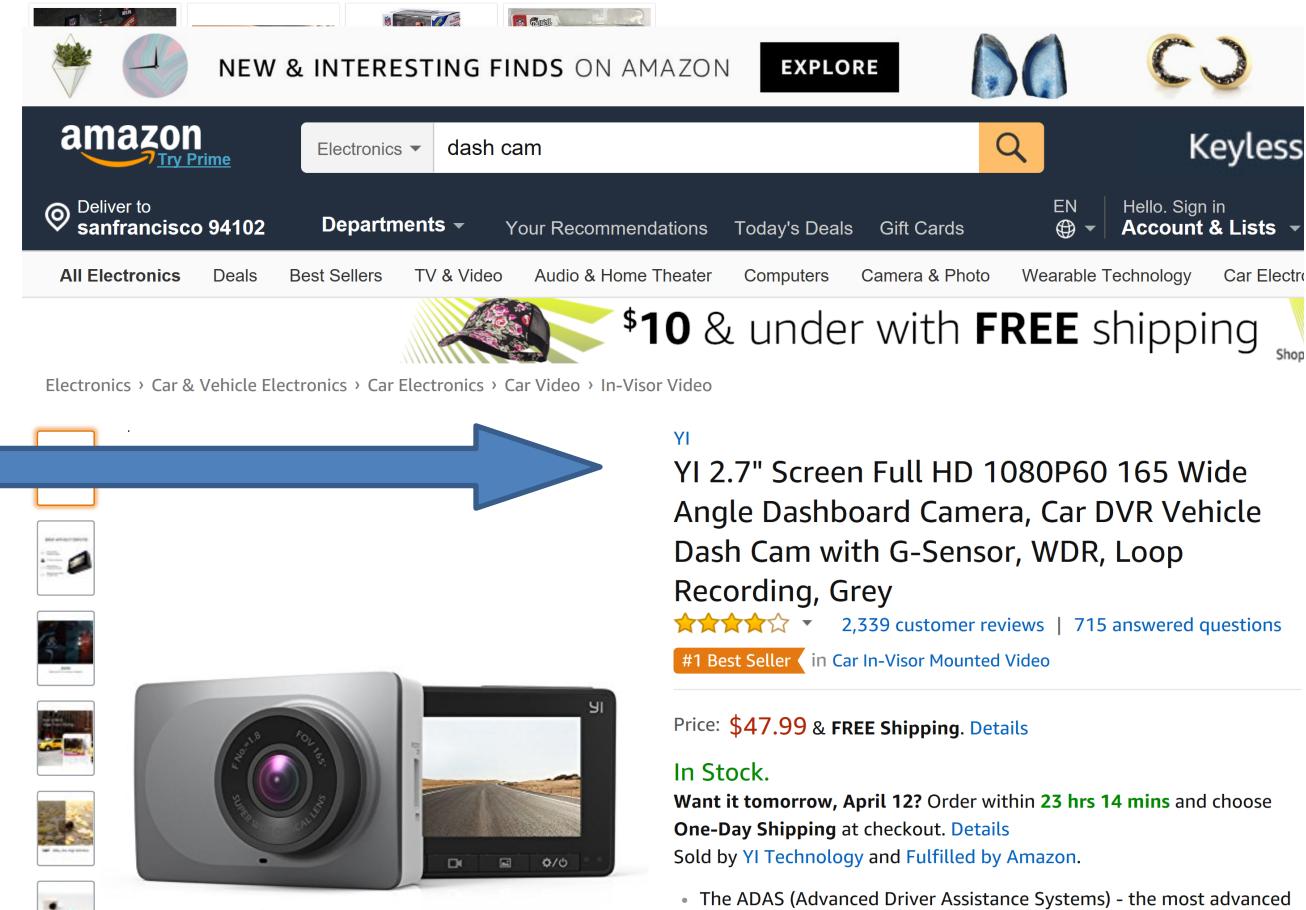
NFL Tom Brady Mcfarlanes Series 27 912/2000 New

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?!



The screenshot shows the Amazon homepage with a search bar containing "dash cam". Below the search bar, a promotional banner for "NEW & INTERESTING FINDS ON AMAZON" features a "Keyless" item. The main search results page displays a product for a "YI 2.7" Screen Full HD 1080P60 165 Wide Angle Dashboard Camera, Car DVR Vehicle Dash Cam with G-Sensor, WDR, Loop Recording, Grey". The product image shows a compact dash cam with a screen and lens. To the left of the main content, a blue callout box contains the text: "How many different data fields are crammed into the ‘product name’ field?!". A large blue arrow points from this text towards the product listing on the right.

NEW & INTERESTING FINDS ON AMAZON

EXPLORE

amazon Try Prime

Electronics dash cam

Deliver to sanfrancisco 94102 Departments Your Recommendations Today's Deals Gift Cards EN Hello, Sign in Account & Lists

All Electronics Deals Best Sellers TV & Video Audio & Home Theater Computers Camera & Photo Wearable Technology Car Electr

\$10 & under with FREE shipping

Shop

Electronics > Car & Vehicle Electronics > Car Electronics > Car Video > In-Visor Video

YI

YI 2.7" Screen Full HD 1080P60 165 Wide Angle Dashboard Camera, Car DVR Vehicle Dash Cam with G-Sensor, WDR, Loop Recording, Grey

★★★★★ 2,339 customer reviews | 715 answered questions

#1 Best Seller in Car In-Visor Mounted Video

Price: \$47.99 & FREE Shipping. Details

In Stock.

Want it tomorrow, April 12? Order within 23 hrs 14 mins and choose One-Day Shipping at checkout. Details

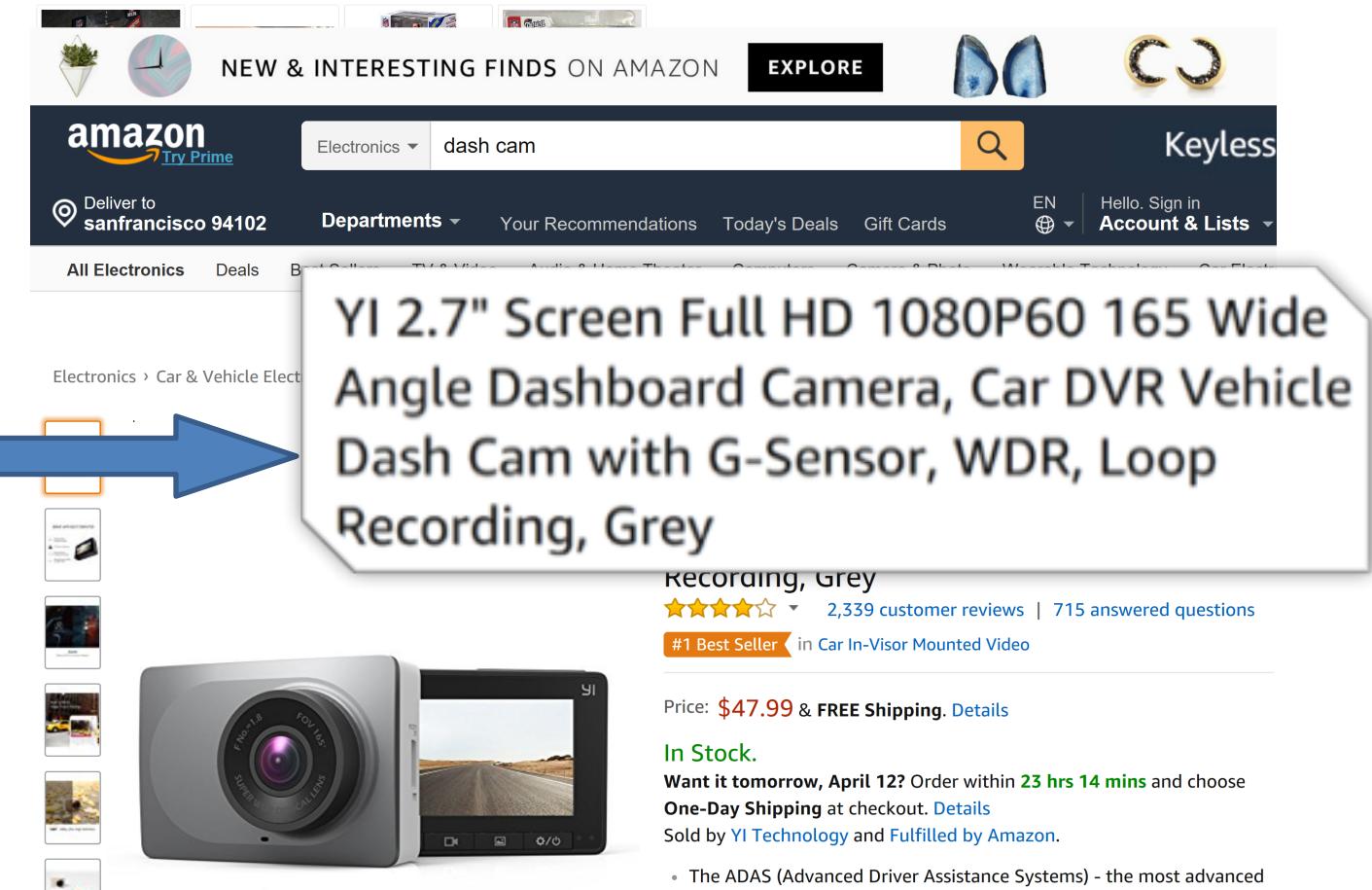
Sold by YI Technology and Fulfilled by Amazon.

- The ADAS (Advanced Driver Assistance Systems) - the most advanced

# eCommerce

Also, product descriptions are often not unitized.

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



The screenshot shows an Amazon product page for a YI 2.7" Screen Full HD 1080P60 165 Wide Angle Dashboard Camera. The product name is displayed as a single, long string of text: "YI 2.7" Screen Full HD 1080P60 165 Wide Angle Dashboard Camera, Car DVR Vehicle Dash Cam with G-Sensor, WDR, Loop Recording, Grey". A blue arrow points from the text box on the left to this product name. The page includes standard Amazon navigation elements like the search bar, categories, and account information.

YI 2.7" Screen Full HD 1080P60 165 Wide Angle Dashboard Camera, Car DVR Vehicle Dash Cam with G-Sensor, WDR, Loop Recording, Grey

Recording, Grey

★★★★★ 2,339 customer reviews | 715 answered questions

#1 Best Seller in Car In-Visor Mounted Video

Price: \$47.99 & FREE Shipping. Details

In Stock.

Want it tomorrow, April 12? Order within 23 hrs 14 mins and choose One-Day Shipping at checkout. Details

Sold by YI Technology and Fulfilled by Amazon.

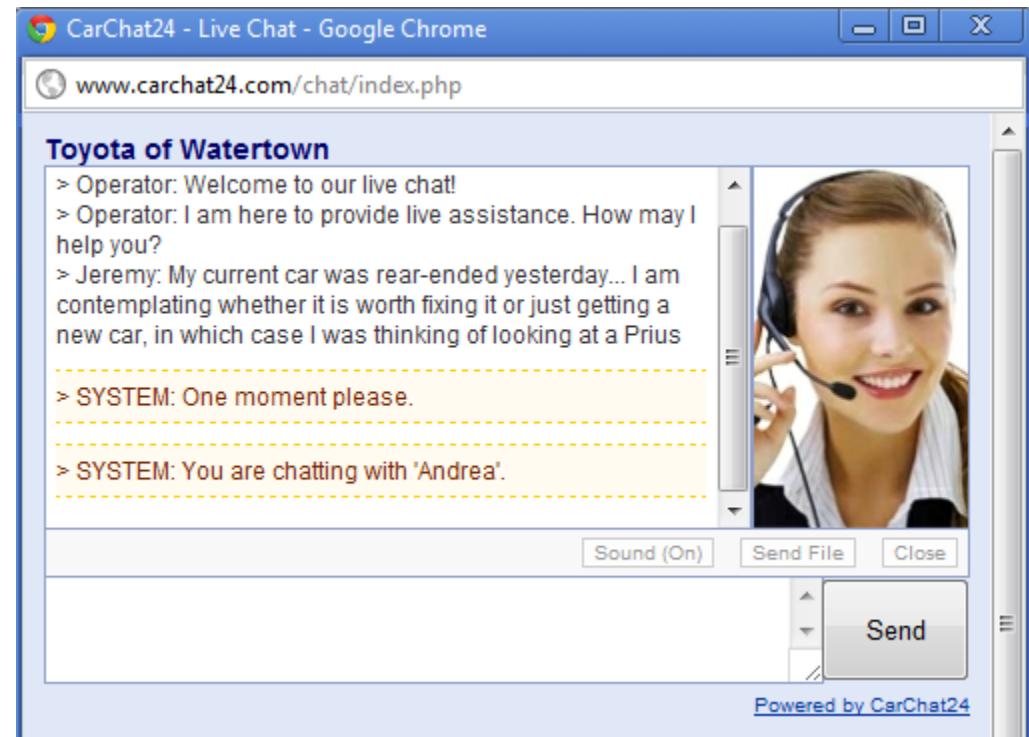
• The ADAS (Advanced Driver Assistance Systems) - the most advanced

# 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



New York blog: Through Storm Sandy, stories of courage and help  
[goo.gl/fb/bUWkb](http://goo.gl/fb/bUWkb)

[Reply](#) [Retweet](#) [Favorite](#)



New York blog: Through Storm Sandy, stories of courage and help

It looks like New York but doesn't feel like it. When I look out of my window I see a smaller city because half of it is still plunged in darkness. A tale of two cities then - one with power, one...

NDTV @ndtv · Follow

2:24 AM - 31 Oct 12 · Embed this Tweet

Flag media



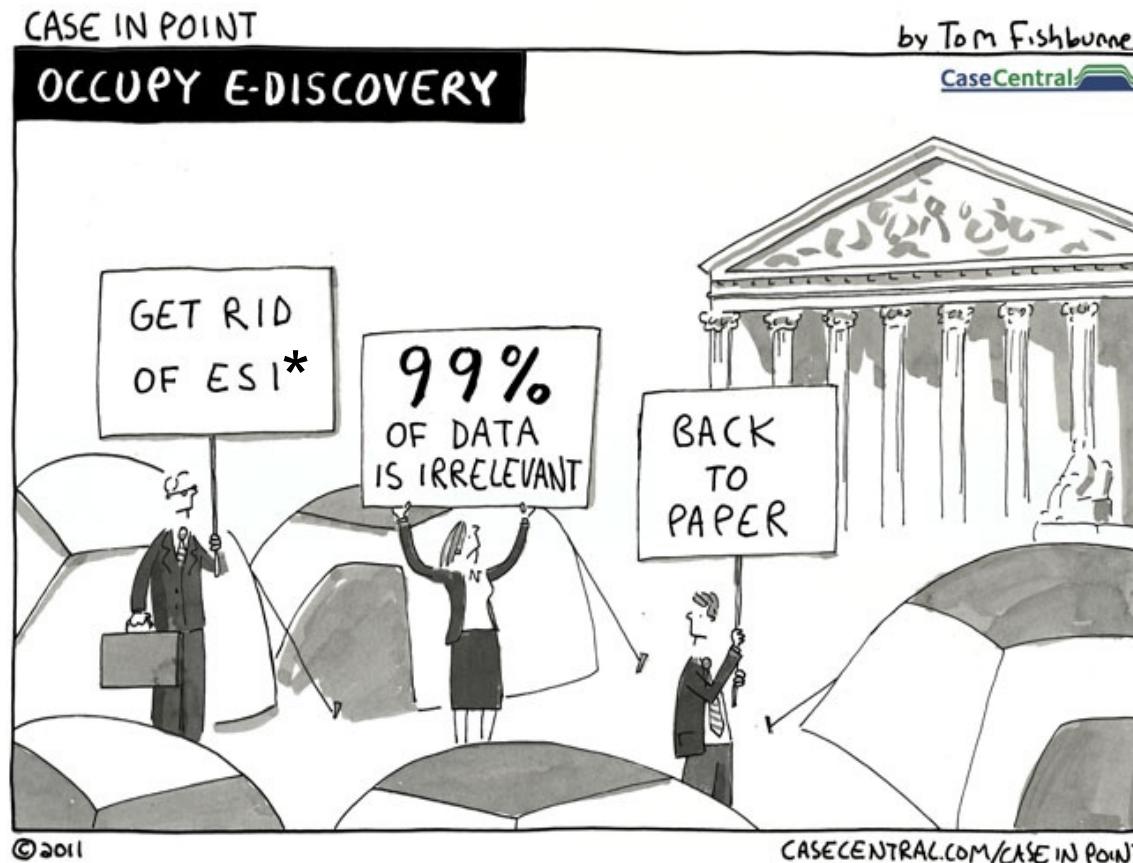
Four stages of [#redwedding](#) grief: shock, anger, depression, exhaustion. Hang in friends, you will move forward.  
[#gameofthrones](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)

7:42 PM - 2 Jun 13

# 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**

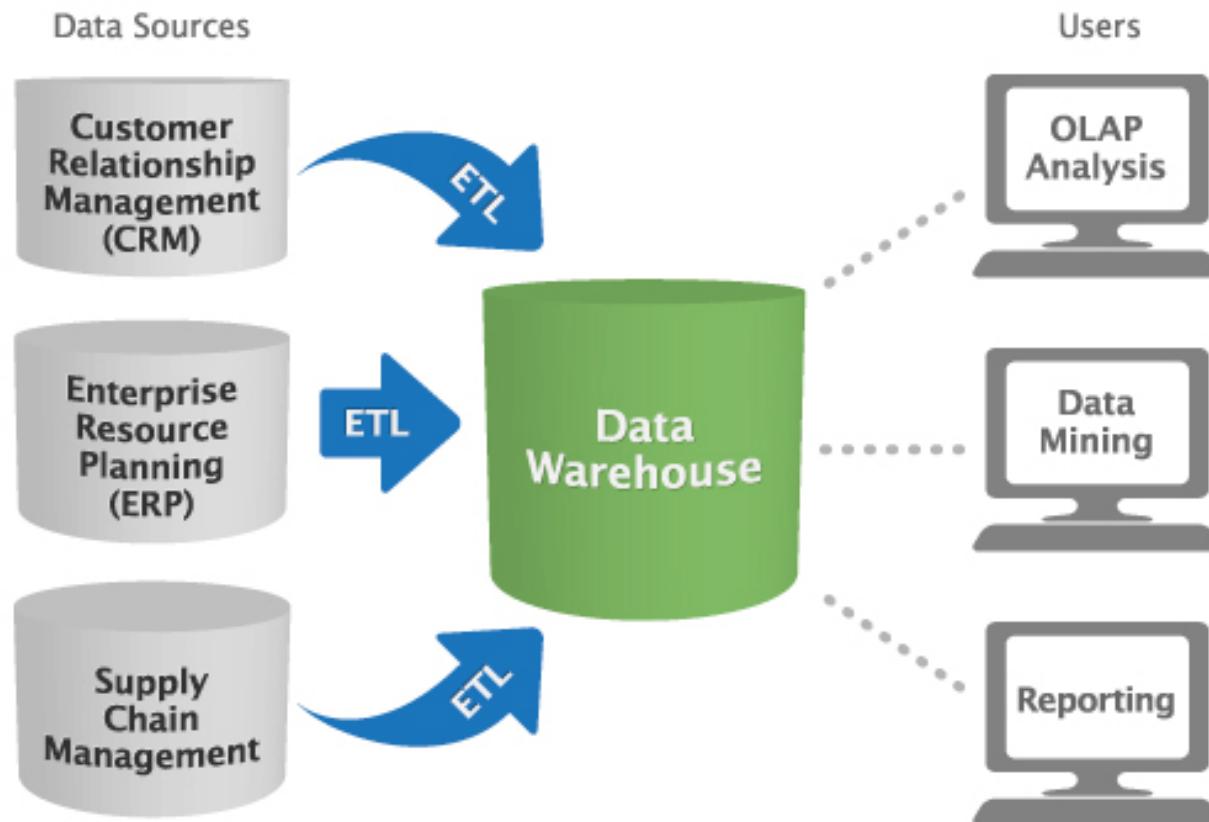
# Working in NLP: Sectors Using NLP

---

Natural Language Processing

# Business Intelligence (BI)

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



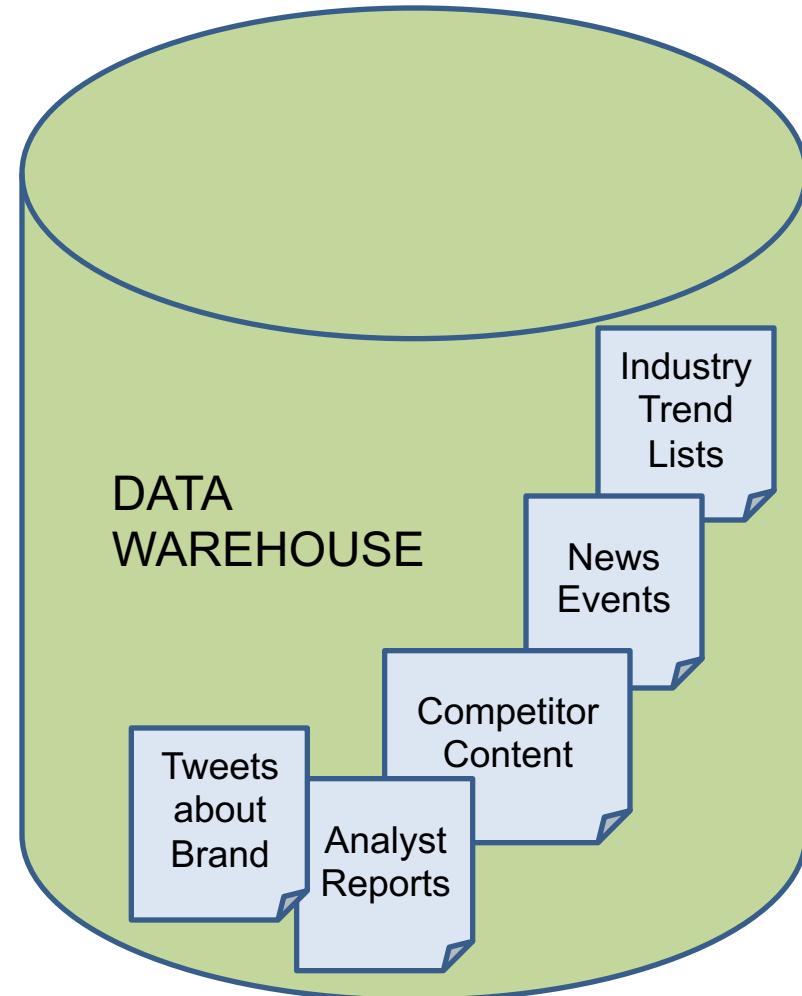
# Business Intelligence (BI)

---

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.

EDITION: ▾

ZDNet 

VIDEOS SMART CITIES WINDOWS 10 CLOUD INNOVATION SECURITY TECH PRO MORE ▾ NEWSLETTERS ALL WR

## Microsoft adds natural language search to its Power BI preview

Microsoft's Q&A natural-lang preview.  
By Mary Jo Foley for All About Microsoft

In July, Microsoft announced plans to add natural language search to its business intelligence tools, known as Power BI.

 SISENSE

Blog Home Why Sisense Platform Solutions

## Here's Why Natural Language Processing is the Future of BI

by Shelby Blitz on March 8th, 2017  
Categories: Business Perspectives

Share:  

*Natural language processing (NLP) and search-driven analytics are just a few of the new technologies companies are using to connect their most potent business minds with the right data. To learn how to harness natural language to deliver business results, [watch our webinar](#) with Aberdeen Research.*

# 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

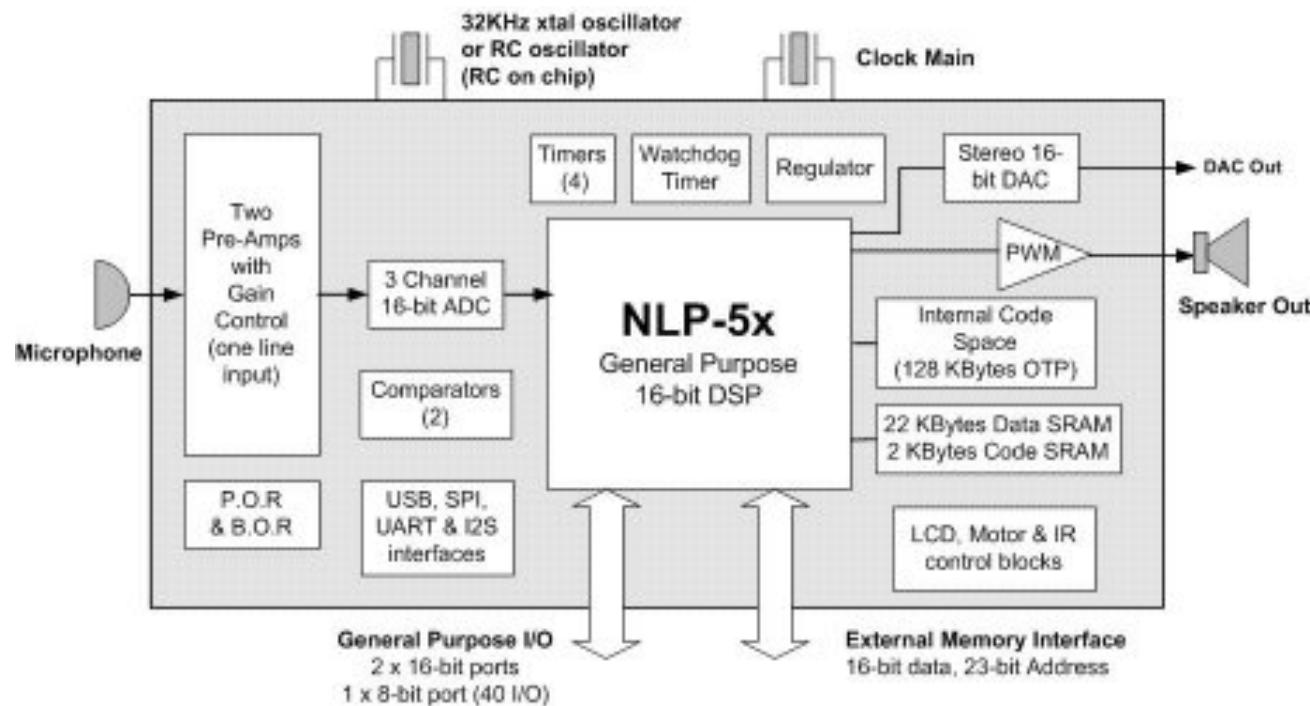
How many other device types can you think of that could use localized in-memory NLP?



# Embedded Technologies

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

- The NLP-5x from Sensory

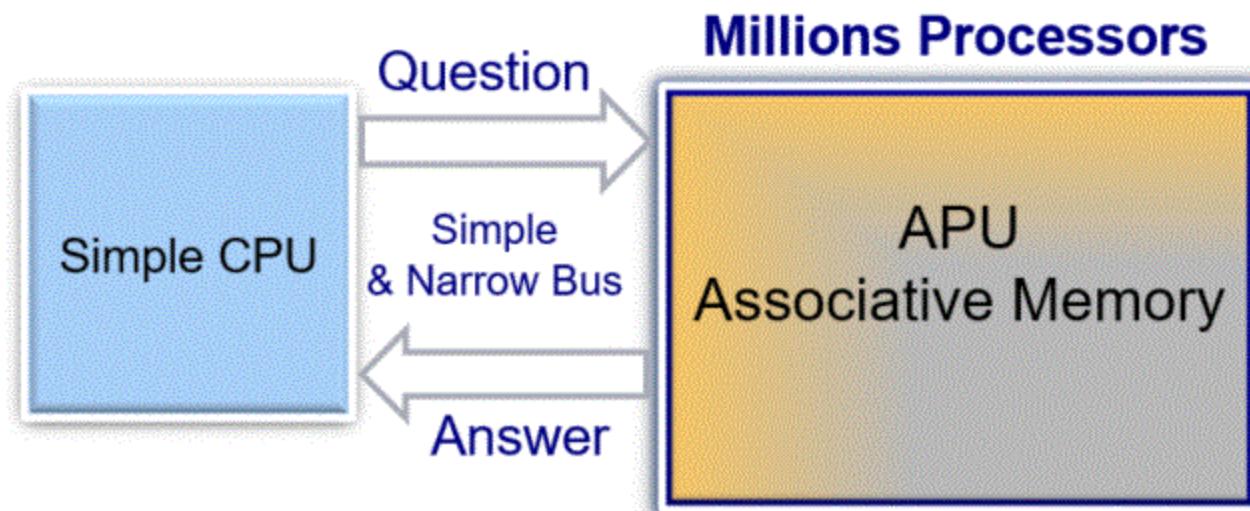


# Embedded Technologies

---

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

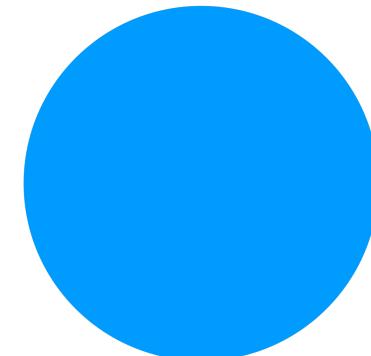
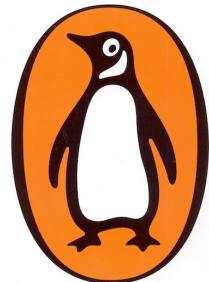


RANDOM HOUSE



The  
Economist

The  
New York  
Times

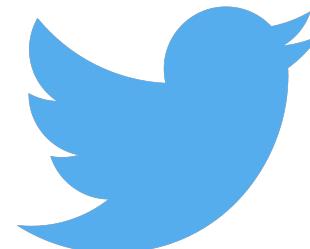


**USA  
TODAY**  
A GANNETT COMPANY

# Publishing and Media

---

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



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



**IEEE**



**IJCAI-16**



AMERICAN ASSOCIATION FOR APPLIED LINGUISTICS

**ICLA** INTERNATIONAL COGNITIVE LINGUISTICS ASSOCIATION

# 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



# IEEE

# AAAI

---

Association for the Advancement of Artificial Intelligence  
[aaai.org](http://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/](http://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>



**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



**IEEE**



**IJCAI-16**



AMERICAN ASSOCIATION FOR APPLIED LINGUISTICS

**ICLA** INTERNATIONAL COGNITIVE LINGUISTICS ASSOCIATION

# 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



# IEEE

# AAAI

---

Association for the Advancement of Artificial Intelligence  
[aaai.org](http://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/](http://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>



**DataScience@SMU**