Overview of Natural Language Processing

MMA 865

Moez Ali

Smith SCHOOL OF BUSINESS | Queen's University

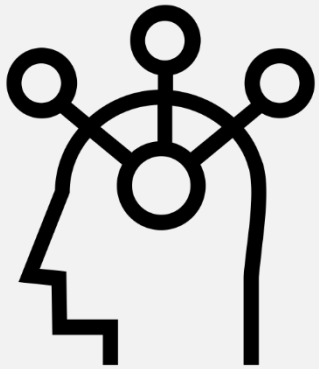
Outline

- **What is Natural Language Processing (NLP)**

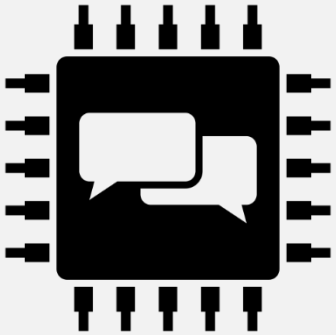
- **Major NLP practice areas**
  - Information extraction
  - Document classification
  - Sentiment analysis
  - Machine Translation
  - Chatbots
  - Language Generation (Week 3 and 4)

WHAT IS NATURAL LANGUAGE PROCESSING?

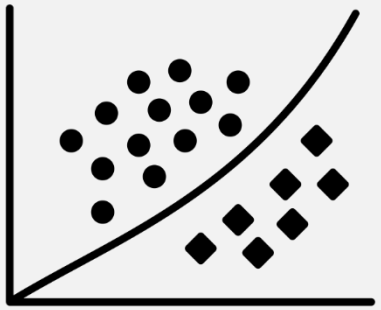
The Six AI Disciplines



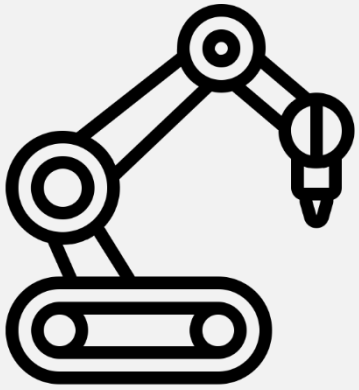
Knowledge Reasoning



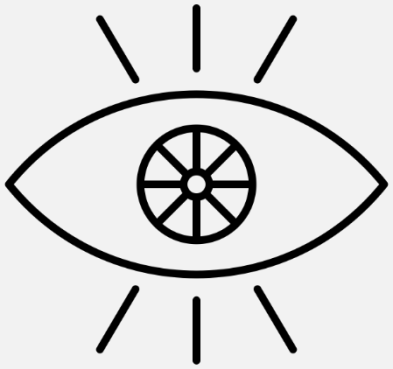
Natural Language Processing



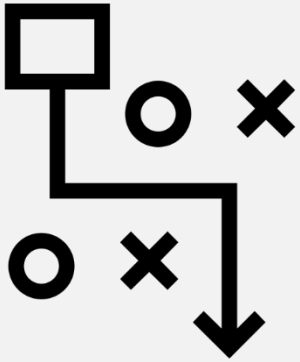
Machine Learning



Robotics



Computer Vision



Planning

What and Why?

natural language processing

noun

- Subset of AI whose goal is to understand and extract meaning from text

90%
of data is unstructured

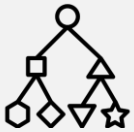
500M+
tweets per day

300B+
emails per day

90%
prefer to message


Information Extraction

Structure


Document Classification

Predict


Sentiment Analysis

Understand


Machine Translation

Translate

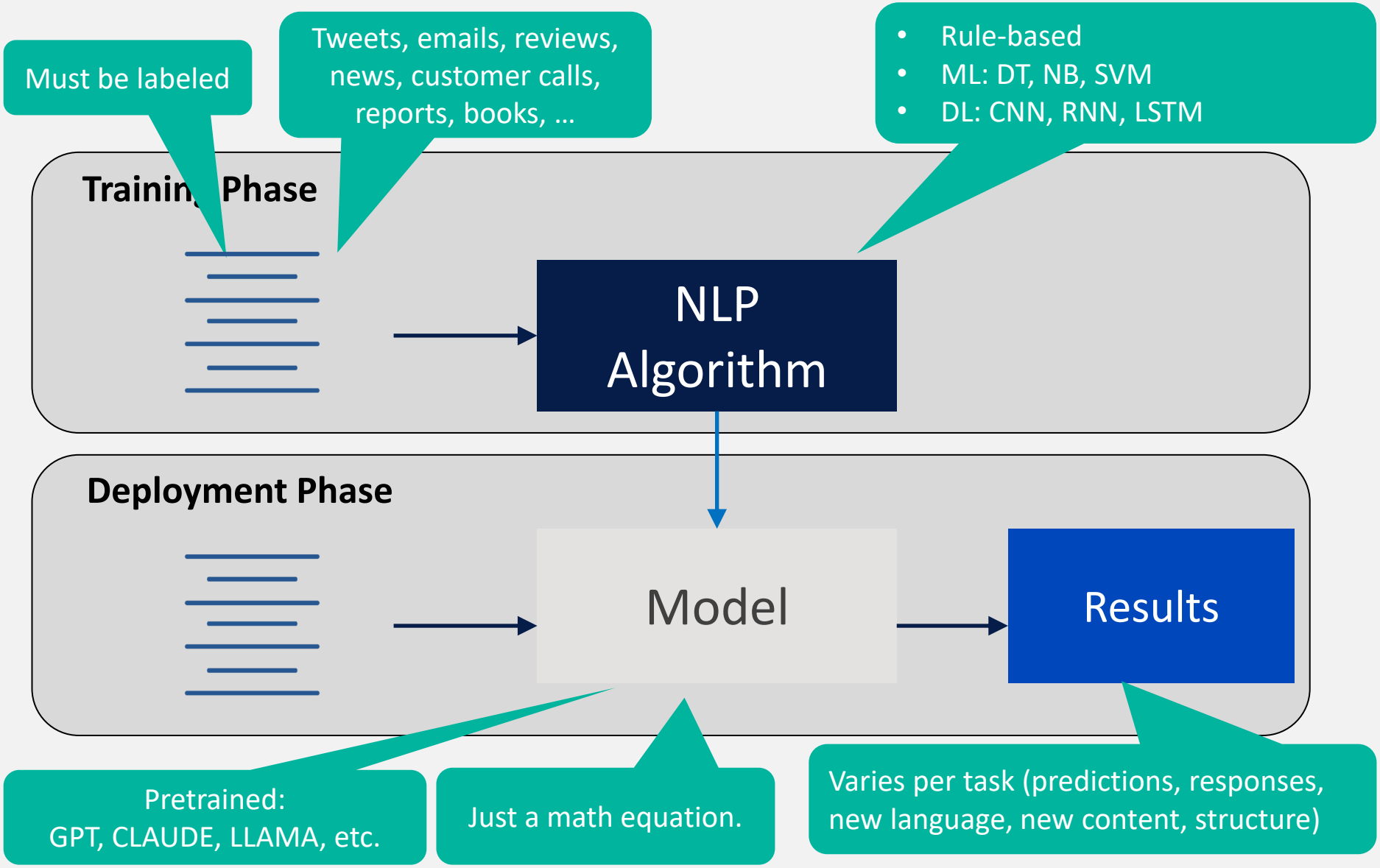

Chatbots

Respond


Language Generation

Create

Natural Language Processing



Structured *versus* Unstructured Data

Stranger Things is an American science fiction horror web television series created by the Duffer Brothers. It was written and directed by Matt and Ross Duffer and executive-produced by Shawn Levy. It stars an ensemble cast of Winona Ryder, David Harbour, Finn Wolfhard,Millie Bobby Brown, Gaten Matarazzo, Caleb McLaughlin, Natalia Dyer, Charlie Heaton, Cara Buono, and Matthew Modine. The plot follows the disappearance of a young boy, and a telekinetic girl who helps his friends in their search, while the boy's older brother and the town police chief start their own investigations. The show is set in 1980s Indiana and is a homage to '80s pop culture, channeling the works of Steven Spielberg, John Carpenter, Stephen King, and George Lucas, among others. The show was released on Netflix on July 15, 2016. It received largely positive reviews for its characterization, pacing, atmosphere, acting, soundtrack, directing, writing, and homages to 1980s genre films.

Structured

- Contains well-defined fields
- Easily parsed by a computer

Unstructured

- Produced by people, for people
- Not easily parsed by a computer
- Not random: Follows linguistic rules and structure

Common Data Sources

Social media

Books/Magazines

CRM

News Articles

Product Reviews

Research Articles

Comments/Surveys

Call Logs

Email

Web pages

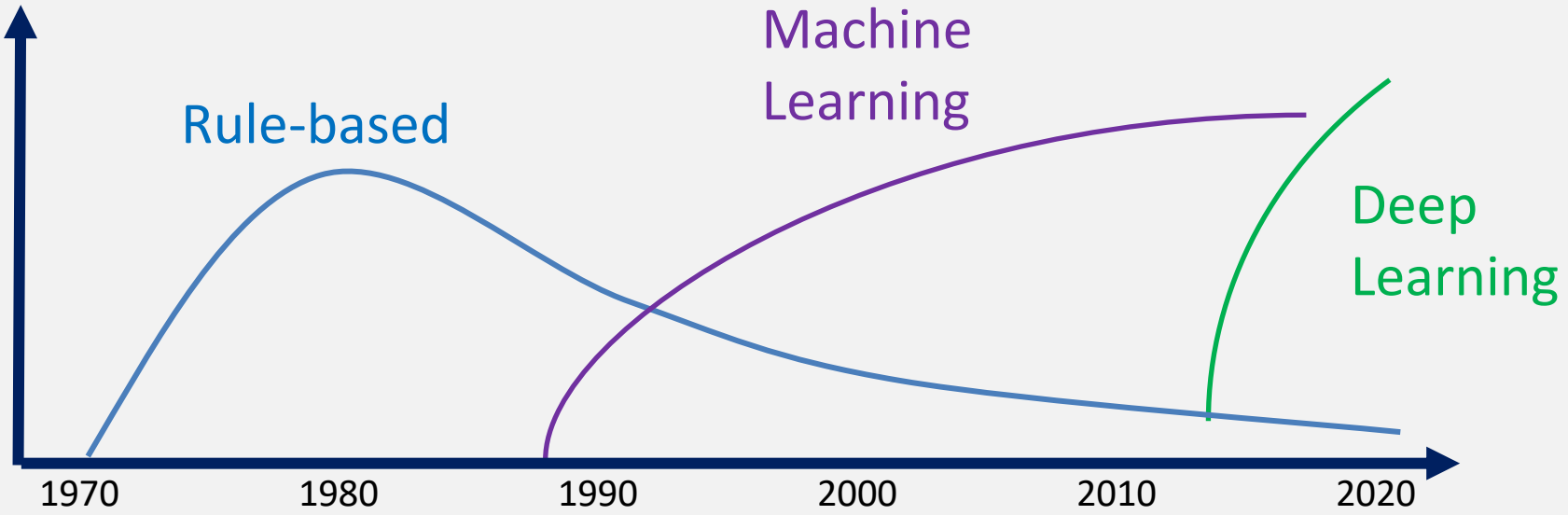
Common Crawl

Timeline of Common Crawl data [edit]

The following data have been collected from the official Common Crawl Blog.[15]

Crawl date	Size in TiB	Billions of pages	Comments
June 2023	390	3.1	Crawl conducted from May 27 to June 11, 2023
April 2023	400	3.1	Crawl conducted from March 20 to April 2, 2023
February 2023	400	3.15	Crawl conducted from January 26 to February 9, 2023
December 2022	420	3.35	Crawl conducted from November 26 to December 10, 2022
October 2022	380	3.15	Crawl conducted in September and October 2022
April 2021	320	3.1	
November 2018	220	2.6	
October 2018	240	3.0	
September 2018	220	2.8	
August 2018	220	2.65	

The Three NLP Eras



- **Rule-based**
  - Manually-created rules, dictionary lookup, regular expressions
- **Machine Learning**
  - Extract features from text and then use:
    - Decision Tree, Naïve Bayes, Random Forest, KNN, etc.
- **Deep Learning**
  - Use word embeddings (Word2vec, Glove, etc.)
  - Use Deep NN architectures (RNN, LSTM, Attention)

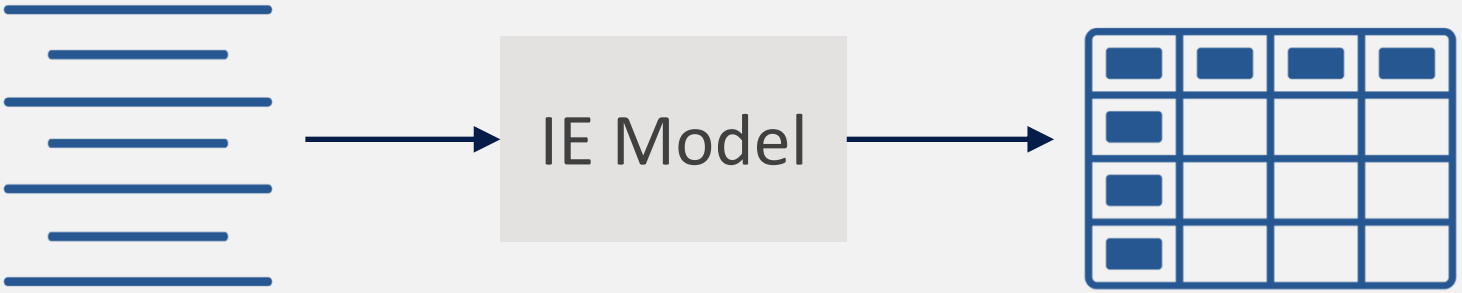
Discussion

- Can you think of data sources at your workplace that produces a lot of text data?
- Some examples:
  - Call center transcripts
  - Surveys
  - Product Reviews

INFORMATION EXTRACTION

Information Extraction

Extract entities and relationships



Example

Elon Musk co-founded Tesla.

Moez Ali works at Queen's University.

Steve Jobs, the co-founder of Apple, died of cancer in his Palo Alto home.

Entity	Relationship	Entity
Elon Musk	Works at	Tesla
Moez Ali	Works at	Queen's
Steve Jobs	Works at	Apple
Steve Jobs	Lives in	Palo Alto
Steve Jobs	Died of	Cancer

Example: Medical Notes

Pt is 40yo mother, software engineer. HPI: Sleeping trouble on present dosage of Clonidine. Severe Rash on face and leg, slightly itchy.

Meds: Vyvanse 50 mgs po at breakfast daily, Clonidine 0.2 mgs – 1 and 1 / 2 tabs po ghs

HEENT: Boggy inferior turbinates, No oropharyngeal lesion. Lungs: clear. Heart: Regular rhythm. Skin: Papular mild erythematous eruption to hairline.
Follow-up as scheduled.

→ IE Model →

Entity	Class
40yo	Age
software engineer	Profession
Sleeping trouble	Dx name
Clonidine	Medication
Rash	Dx name
face	System organ site
leg	System organ site

IE Applications

- Resume harvesting
- News and media tracking
- Patent search
- Email scanning
- Parsing medical records
- Parsing scientific literature
- Business and financial intelligence
- Entity-oriented search

Demo

First Liberal MP calls on Justin Trudeau to step down

28 June 2024

Share Save

Nadine Yousif
BBC News, Toronto


```
Entity: First Liberal MP, Label: ORG
Entity: Justin Trudeau, Label: PERSON
Entity: Justin Trudeau, Label: PERSON
Entity: 2015, Label: DATE
Entity: first, Label: ORDINAL
Entity: Canadian, Label: NORP
Entity: Justin Trudeau, Label: PERSON
Entity: Liberal Party, Label: ORG
Entity: Wayne Long, Label: PERSON
Entity: Canadian, Label: NORP
Entity: Long, Label: PERSON
```

First Liberal MP ORG calls on Justin Trudeau PERSON to step down

Prime Minister Justin Trudeau PERSON has been in office since 2015 DATE

A backbench MP has become the first ORDINAL sitting member of Canadian NORP Prime Minister Justin Trudeau PERSON 's

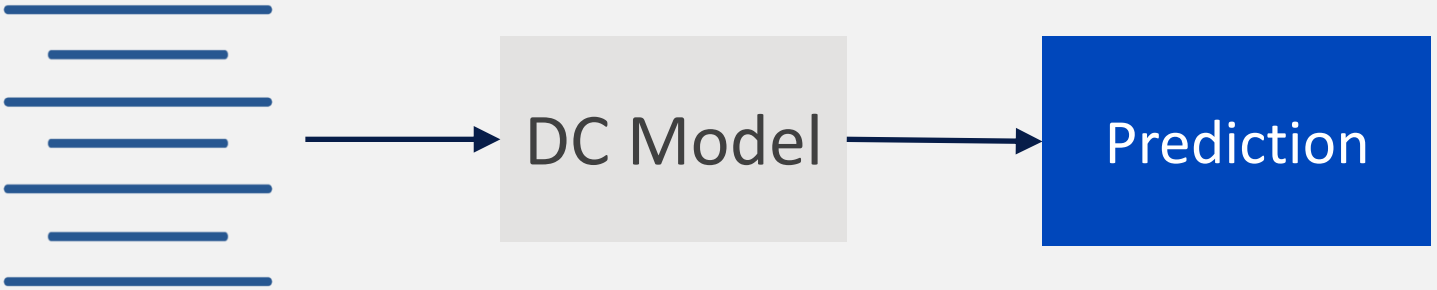
Wayne Long PERSON said new leadership was needed "for the future of our party and for the good of our country" in an email to fe

Mr Long PERSON, who represents a riding in the province of New Brunswick GPE, wrote "the voters have spoken loud and clea

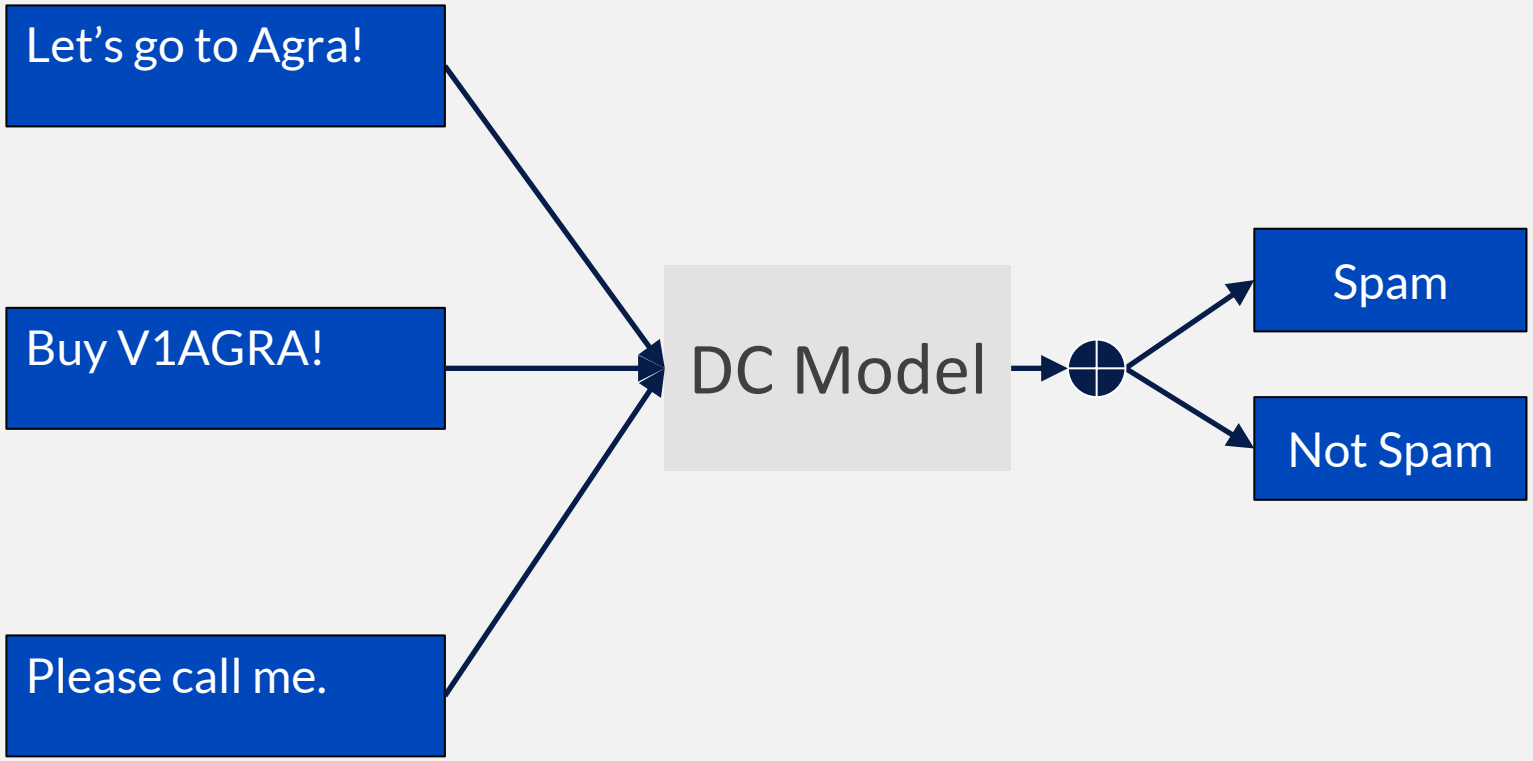
DOCUMENT CLASSIFICATION

Document Classification

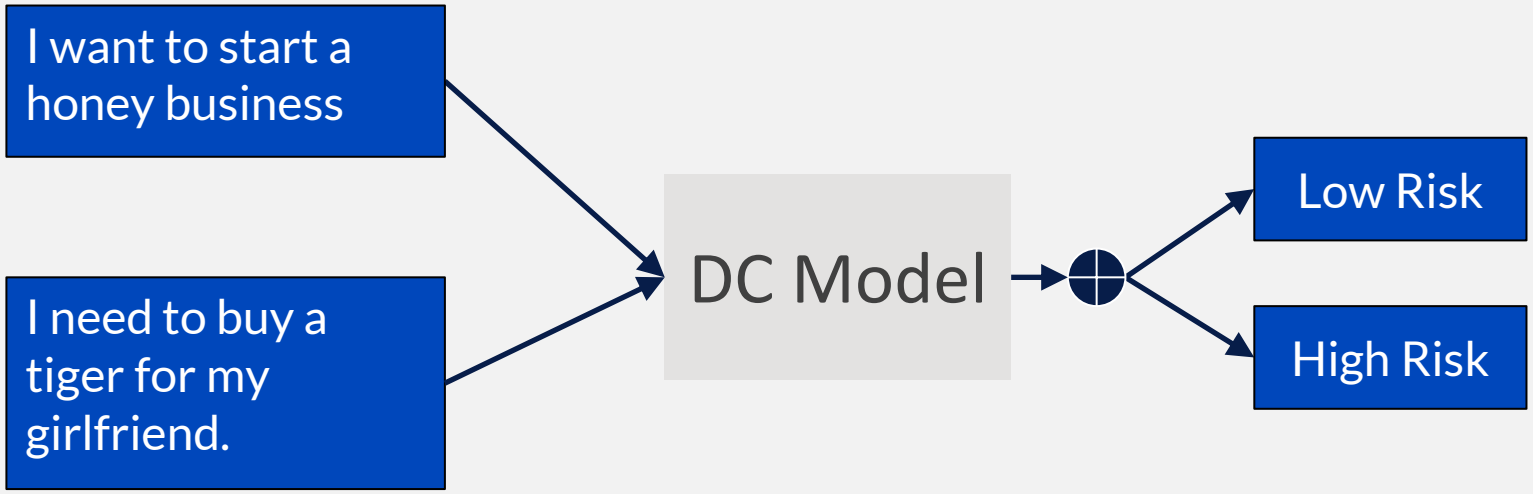
Predict category of documents



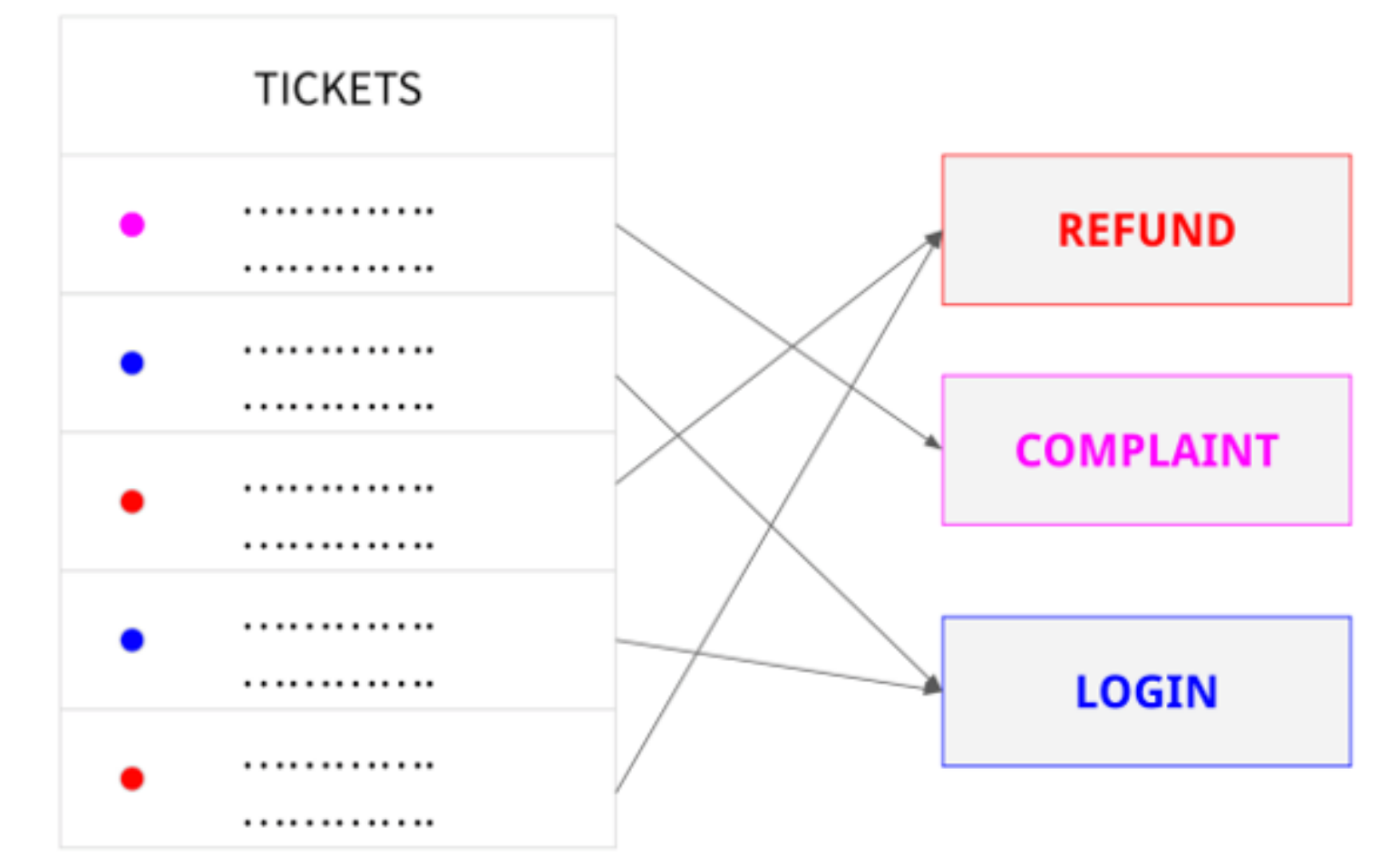
Example: Spam Detection



Example: Loan Risk Assessment

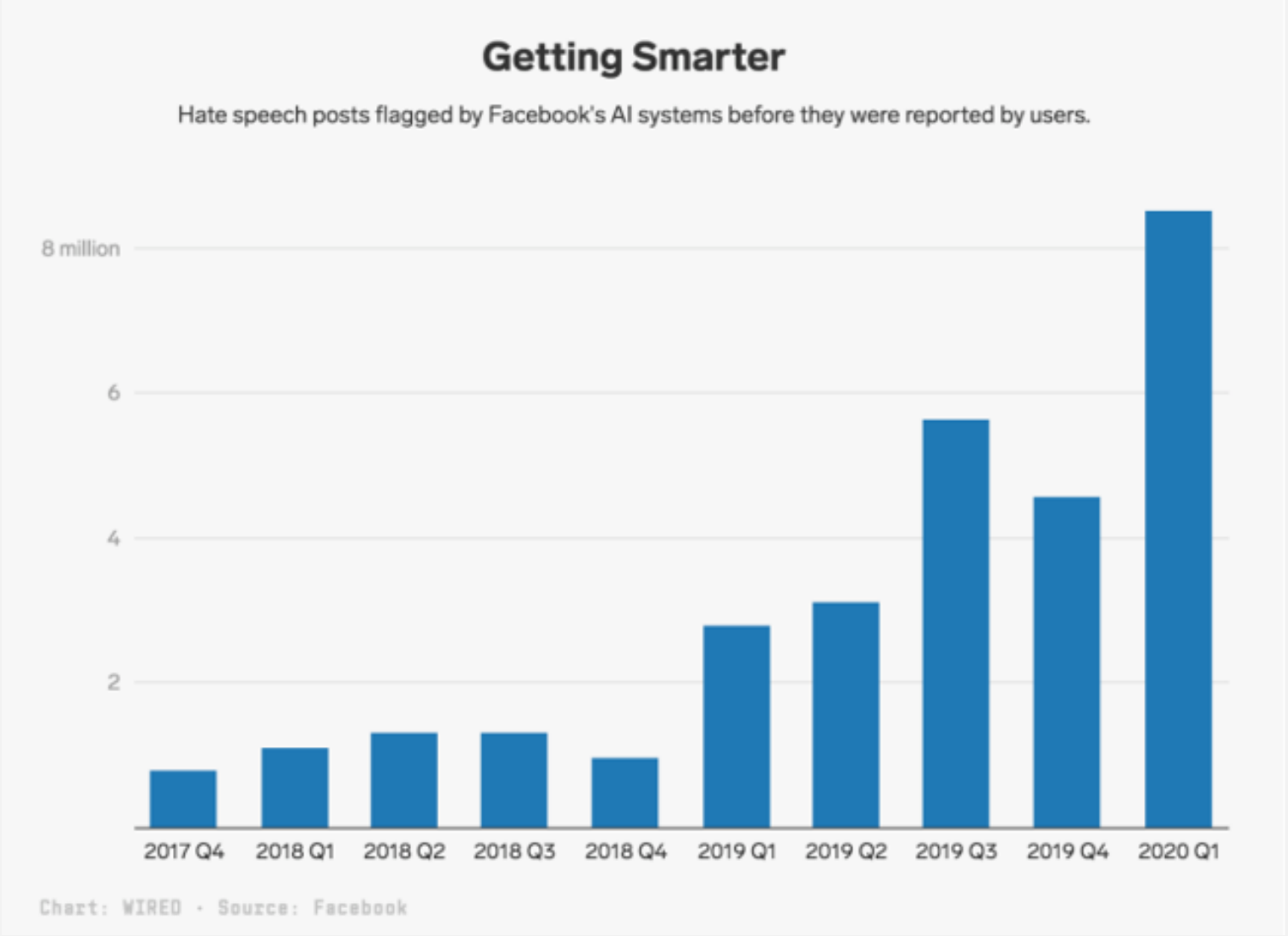


Application: Customer Support



Source: https://kavita-ganesan.com/5-real-world-examples-of-text-classification/#.YQbiYo5KhPY

Application: Facebook Hate Detection



Source: https://kavita-ganesan.com/5-real-world-examples-of-text-classification/#.YQbiYo5KhPY

Document Classification Applications

- Spam detection
- Loan risk assessment
- Churn prediction
- Insurance claim approval
- Insurance fraud detection
- Email triage

Demo

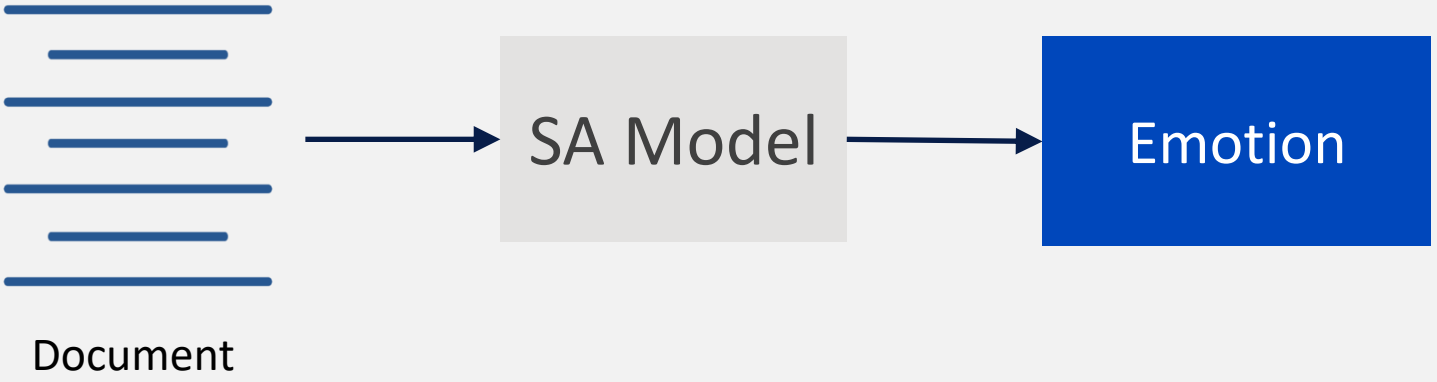
Trump	Trudeau
I will be making a major statement from the @WhiteHouse upon my return to D.C. Time and date to be set.	RT @NavdeepSBains: Innovation at work! Canadian and Indian businesses met tonight to explore opportunities for collaboration, building on oâ€¦
Just arrived at #ASEAN50 in the Philippines for my final stop with World Leaders. Will lead to FAIR TRADE DEALS, unâ€¦ https://t.co/ExRBdQnJru	RT @NavdeepSBains: L'innovation en action ! Les entreprises canadiennes et indiennes se sont rencontrÃ©es ce soir pour explorer les possibilâ€¦
After my tour of Asia, all Countries dealing with us on TRADE know that the rules have changed. The United States hâ€¦ https://t.co/MoMNa1MUI3	I'll always be grateful for George Gosbee's work on our advisory council during our time in opposition. Calgary hasâ€¦ https://t.co/ZihaF8eVoE
Great to see @RandPaul looking well and back on the Senate floor. He will help us with TAX CUTS and REFORM!	Je serai toujours reconnaissant pour le travail que George Gosbee a fait au sein de notre comitÃ© consultatif quandâ€¦ https://t.co/SIG4z2OjHL
Excited to be heading home to see the House pass a GREAT Tax Bill with the middle class getting big TAX CUTS! #MakeAmericaGreatAgainðŸ‡ºðŸ‡¸	RT @FP_Champagne: All fired up with the largest business delegation of the #TradeMissionToIndia. Let's be ambitious, let's keep building ouâ€¦

SENTIMENT ANALYSIS

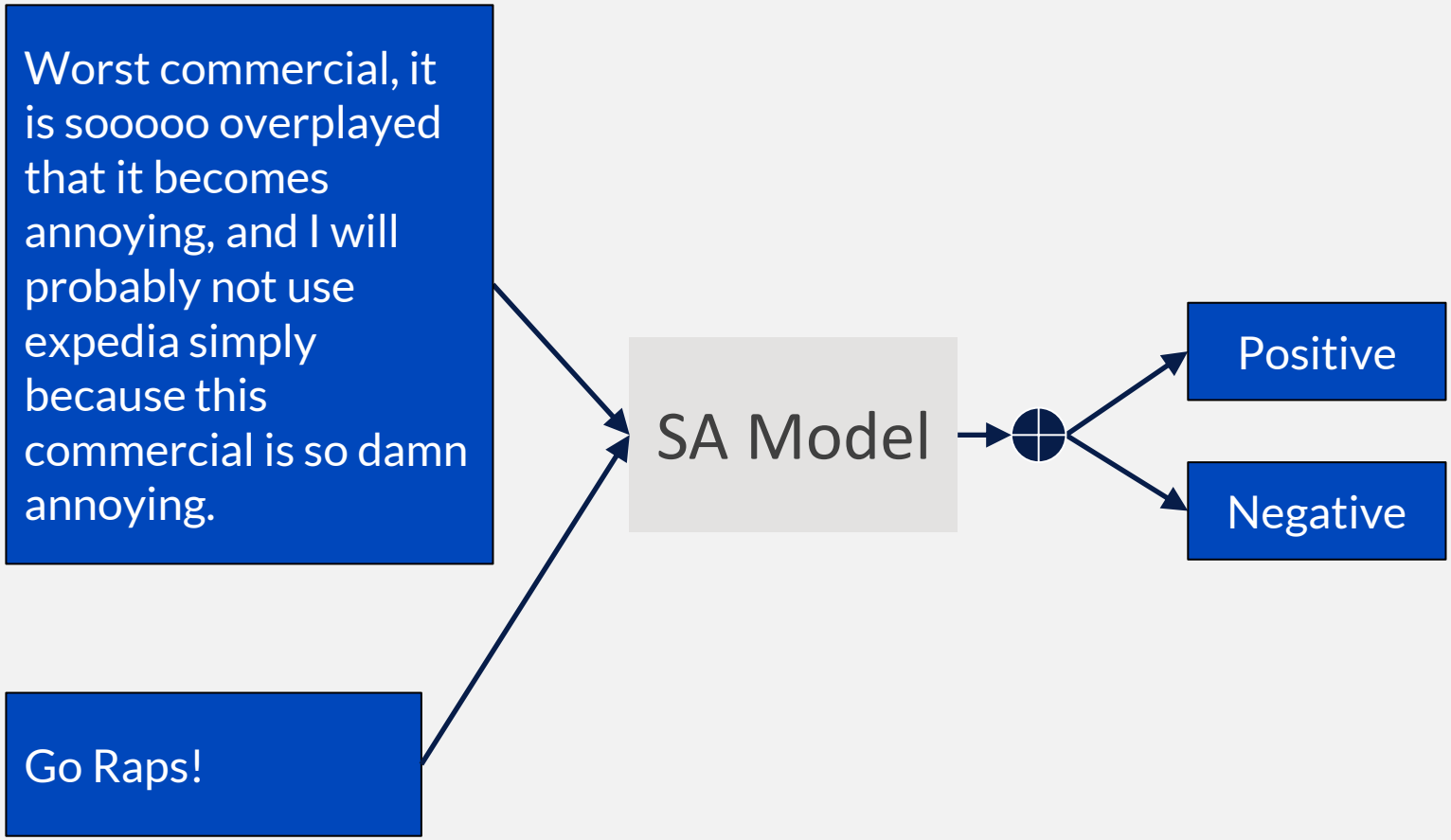
Sentiment Analysis

Automatically determine the *emotion* of a document

- Positive, negative, neutral
- Happy, sad, disgust, anger, surprise



Example: Social Media Analysis



Sentiment Analysis Approaches

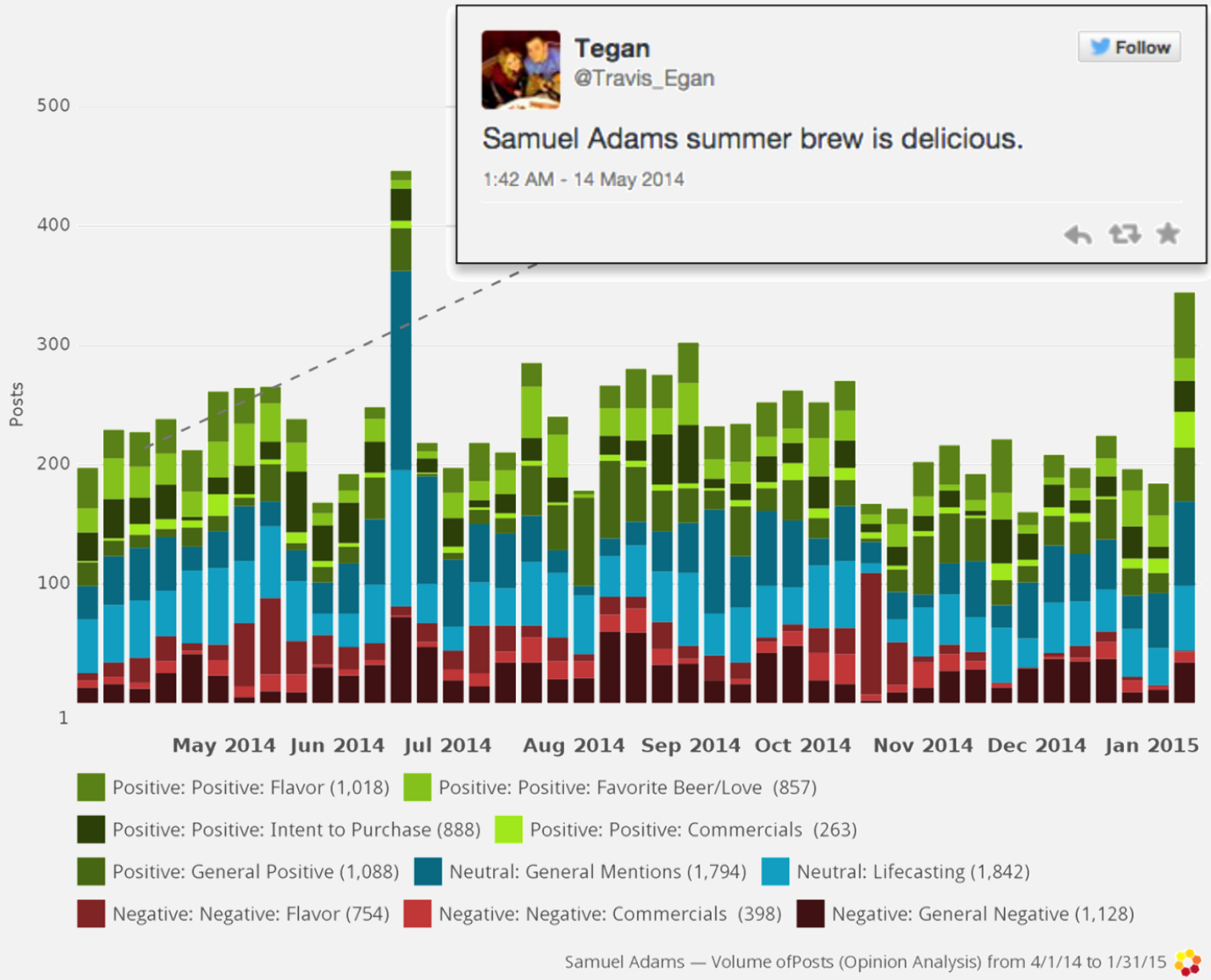
- **Lexicon-based**
  - Dictionary Based
  - Corpus Approach
    - Statistical
    - Semantic

- **Machine Learning Based**

Sentiment Analysis Frameworks in Python

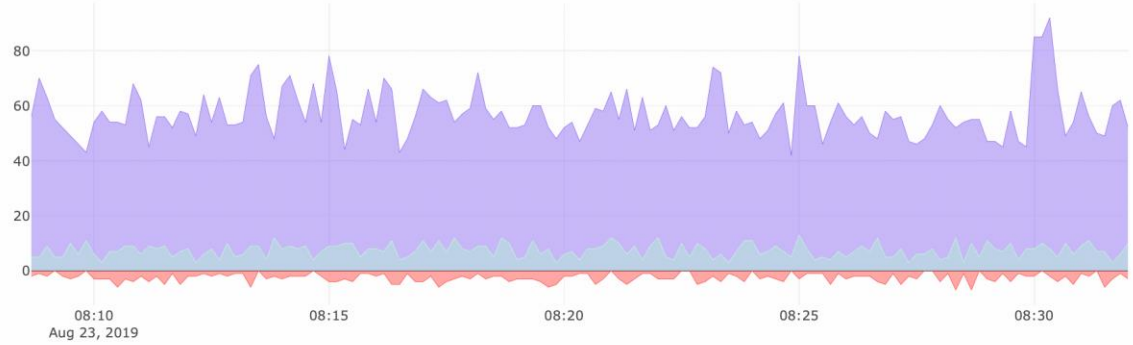
- NLTK
- TextBlob
- spaCy
- Stanford CoreNLP
- Gensim
- Hugging Face
- **ML/DL based approaches:**
  - Scikit-learn
  - TensorFlow / PyTorch

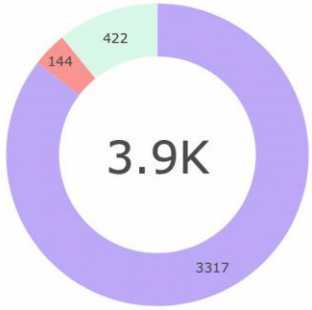
Example: Brand Monitoring



Source: Crimson Hexagon

Example: Facebook





Tweets/10 Mins Changed By

-4.77%

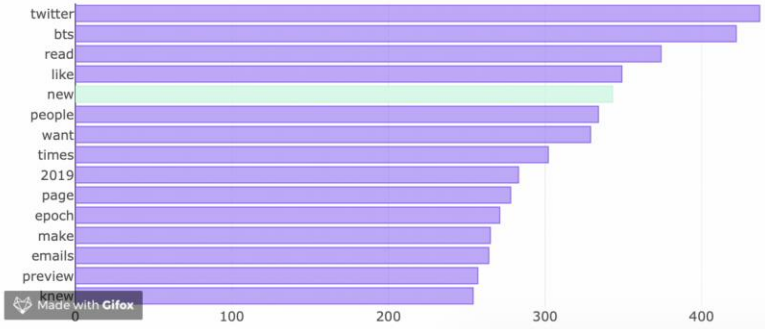
Potential Impressions Today

38.9M

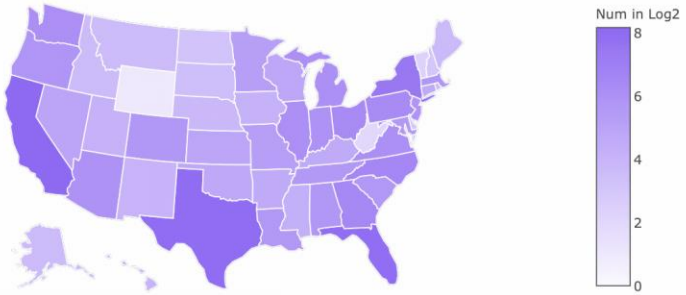
Tweets Posted Today

43.2K

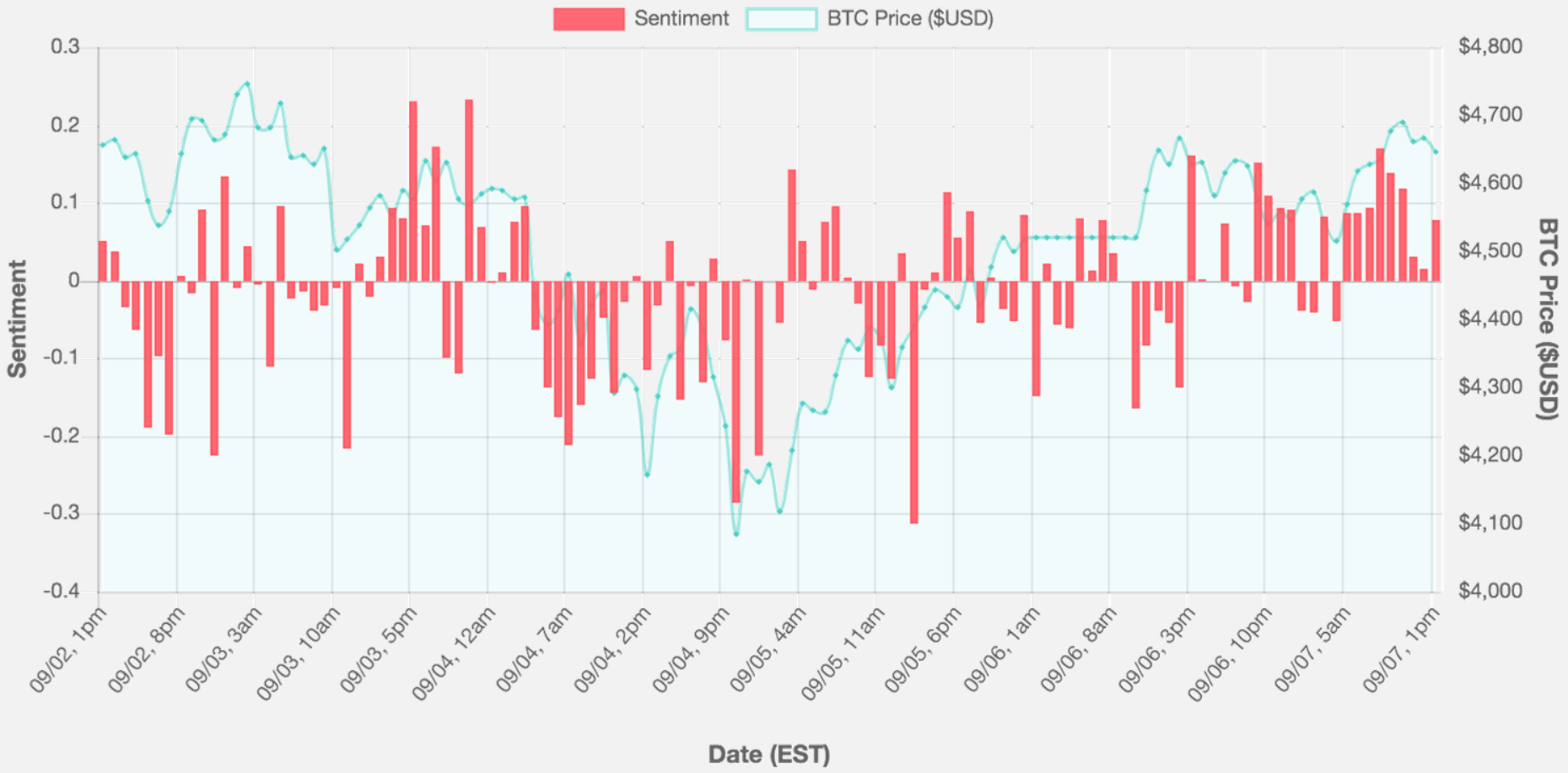
Currently tracking "Facebook" brand (NASDAQ: FB) on Twitter in Pacific Daylight Time (PDT).



Geographic Segmentation for US

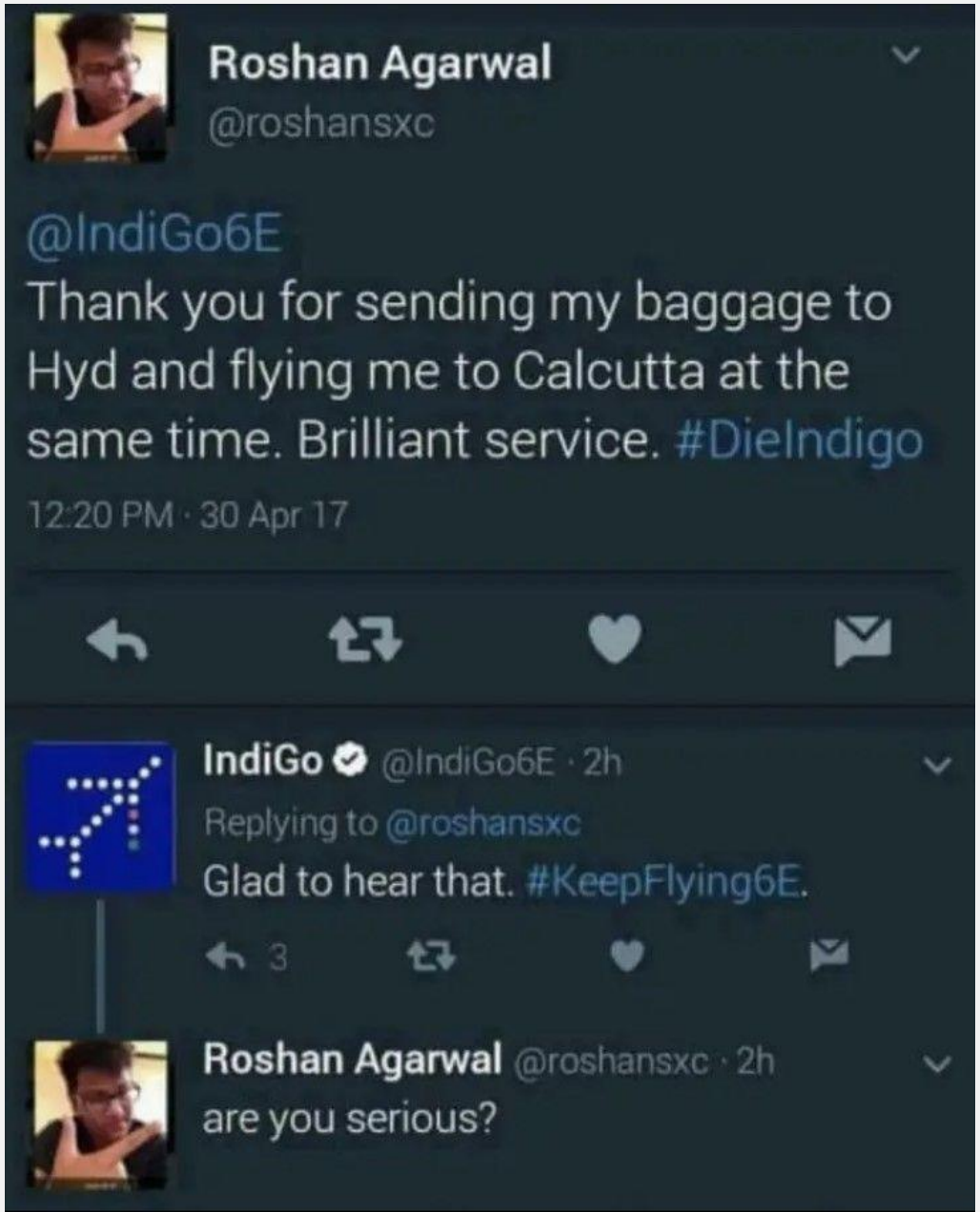


Example: Cryptocurrencies



Source: Medium, Sep 2017

Sentiment Analysis gone wrong!



Sentiment Analysis Applications

- Brand monitoring and comparison (Sam Adams vs Budweiser)
- Review ranking (Airbnb)
- Market research (iPhone, Airlines)
- Product launch (Foxcatcher)
- Literature analysis (Harry Potter)

Demo

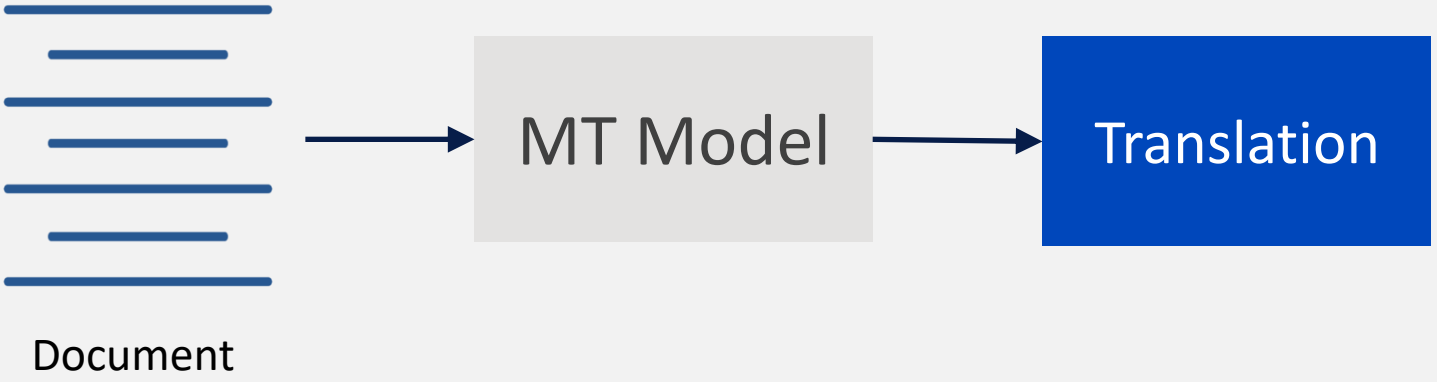
	__typename	brandType	contentDirectFeedbackPromptId	customData/doWeEndorseIt	customData/hotel	hotelId	id	impressionAnalytics	impressionAnalytics/__typename	impressionA
0	PropertyReview	Expedia	NaN	no opinion	Prague Hotel Krystal	428588	66b63b958397934320ad7f75	NaN	NaN	
1	PropertyReview	Expedia	NaN	no opinion	Prague Hotel Krystal	428588	66b503566a7cfe1353bdb91e	NaN	NaN	
2	PropertyReview	Expedia	NaN	no opinion	Prague Hotel Krystal	428588	66b4d8d6e6073643fd1b9cdc	NaN	NaN	
3	PropertyReview	Expedia	NaN	no opinion	Prague Hotel Krystal	428588	66b3c0f33677b72a4489b05f	NaN	NaN	
4	PropertyReview	Expedia	NaN	no opinion	Prague Hotel Krystal	428588	66b36f292fa7b179c05b6637	NaN	ClientSideImpressionEventAnalytics	

Smith
SCHOOL OF BUSINESS
Queen's University

MACHINE TRANSLATION

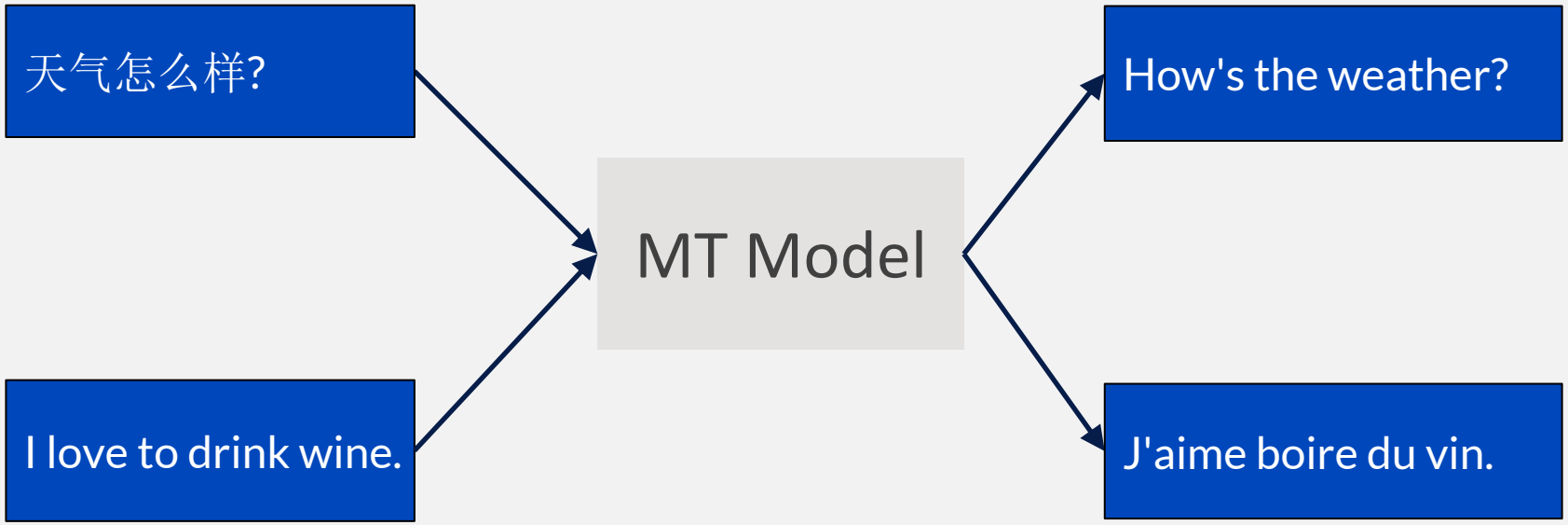
Machine Translation

Translate documents from one language to another



Example

Translate documents from one language to another



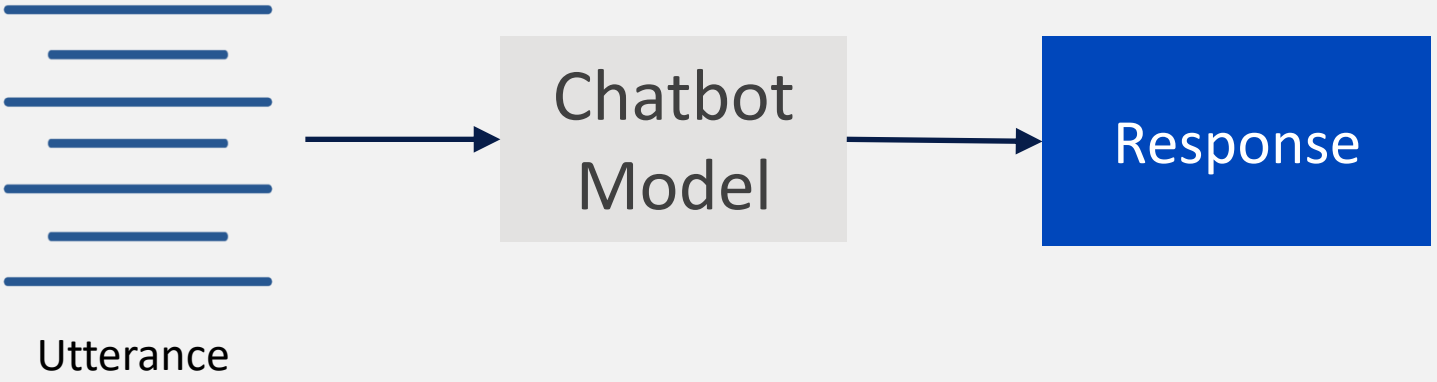
Machine Translation Applications

- Correspondence: emails, chats, etc
- Government documents
- US Military (surveillance)
- Social media
- Video subtitles
- Education
- Speech translation
- Website internationalization (product descriptions, etc)

CHATBOTS

Chatbots

Automatically answer natural language questions from humans.



Example: Virtual Assistants

What's the weather today?

What is there to do this weekend?

Chatbot Model

It will be 28 and sunny

There is a Red Hot Chili Peppers concert.

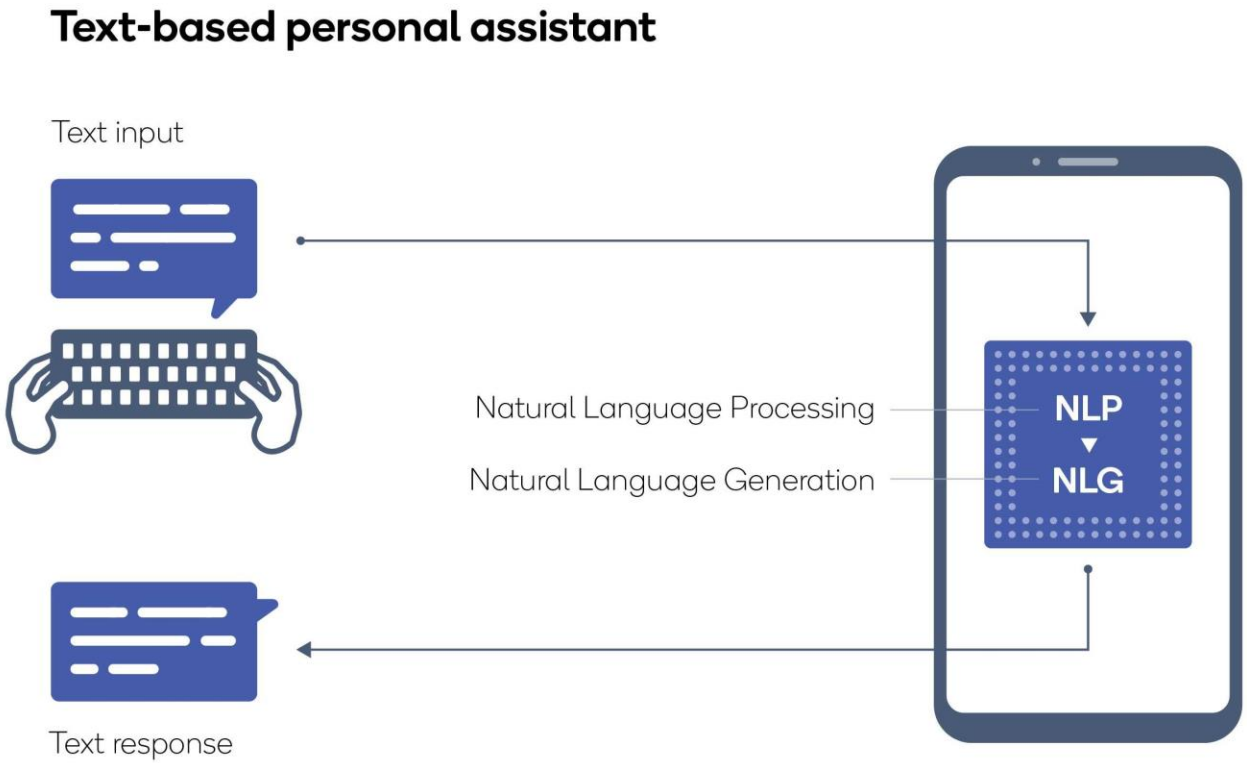
Alexa

Google Assistant

Cortana

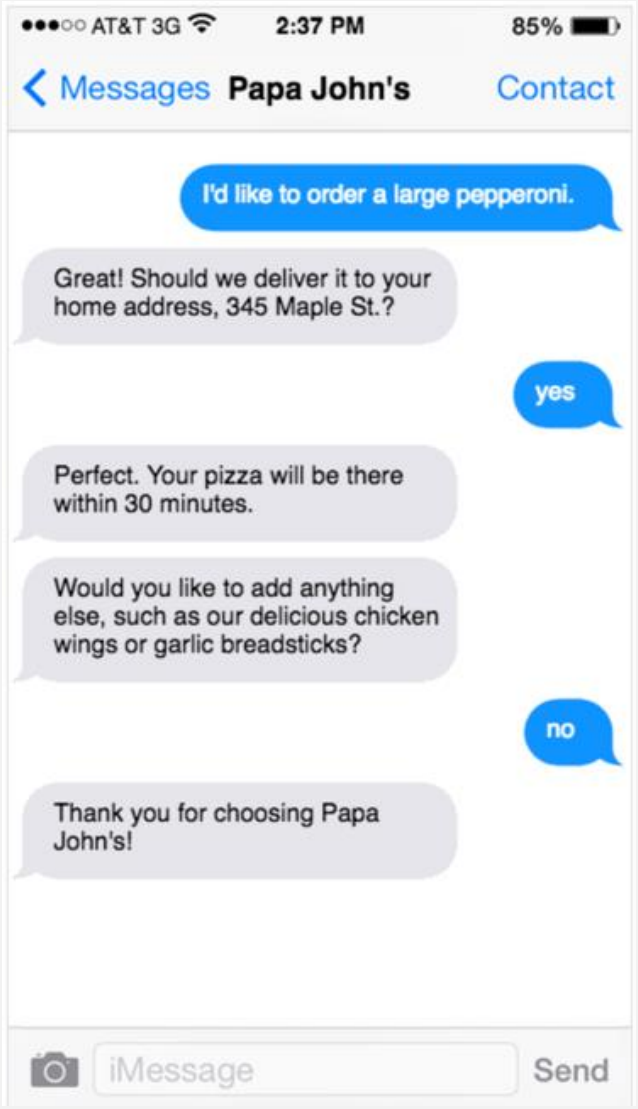
Siri

Chatbots are NLU + NLG

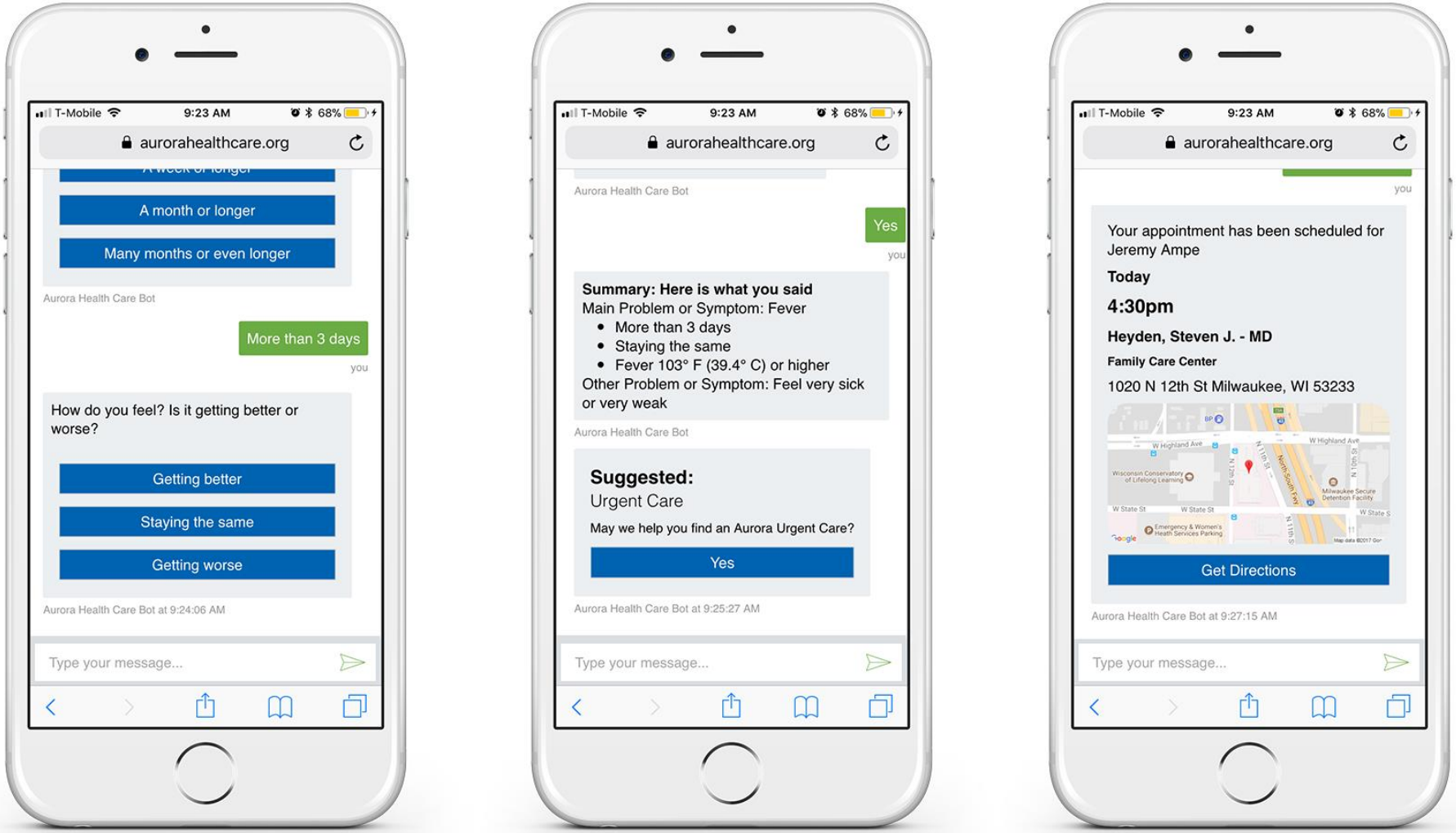


Source: https://www.qualcomm.com/news/onq/2018/01/11/integrating-compelling-chatbots-ai-development

Example: Papa John's



Example: MS Health Bot Project



https://www.microsoft.com/en-us/research/project/health-bot/

Chatbot Applications

- Customer service (Wells Fargo)
- Product orders (Papa John's, Amtrak)
- Product Q&A and product recommendation (Sephora, WholeFoods)
- Virtual assistant (Alexa, Google Home)
- Health triage (MS Healthbot)
- Health advice (Youper)
- Companionship (Endura for dementia patients)
- Games (Disney's Zootopiz)

Discussion

- What are the pros and cons of deploying chatbots
  - From both business perspective and customer perspective?

LANGUAGE GENERATION

Language Generation

Given a prompt, automatically generate content

Example

Miley Cyrus was caught shoplifting from A&F on Hollywood Boulevard today.

→ NLG Model →

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back.

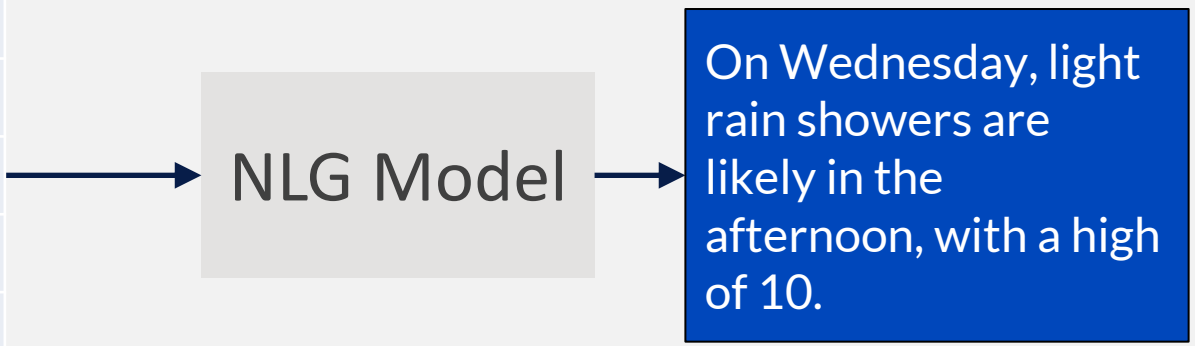
Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

Example

Date	9/9/2020
High	10
Low	7
POP	30%
Wind	7 km/h S
Humidity	90%
Pressure	102.4 kPa



Language Generation Applications

- Generating stories
- Generating weather reports
- Generating any kind of data-driven reports (E.g., financial)
- Enhanced Accessibility
- Generating product descriptions
- Generation meeting memos
- Generating song lyrics

Get started

Enter an instruction or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the model.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our models may be used to train and improve future models. Learn more.
- Our default models' training data cuts off in 2021, so they may not have knowledge of current events.

Playground

Load a preset... Save View code Share ···

Professor 'Moez Ali' needs your help. He is looking for a few ways to improve student experience and course delivery for his courses at Queens. A friend of 'Moez' told him that Natural Language Processing (NLP) could be really helpful. Please give one detailed example of a NLP use-case that can improve student experience and course delivery.

Write a long answer discussing several options for a use-case for NLP in education.

NLP can be used to help students in a number of ways, from improving their writing skills to providing personalized feedback on their assignments.

One way NLP can be used to improve student experience is by providing them with real-time feedback on their writing. This can be done by incorporating NLP into a word processing program or by using a standalone NLP tool. By giving students feedback on their grammar, spelling, and vocabulary usage, they can learn to improve their writing skills over time.

Another way NLP can be used to improve student experience is by providing personalized feedback on assignments. This can be done by using NLP to analyze a student's essay and provide feedback that is tailored to their specific needs. This type of feedback can help students identify areas where they need to improve and can also help them understand the expectations of the assignment.

Finally, NLP can also be used to create targeted content for students. This can be done by using NLP to analyze a student's interests and learning style and then creating content that is specifically tailored to their needs. This type of content can help students engage with the material and can also help them learn more effectively.

Mode

Model: text-davinci-002

Temperature: 0.7

Maximum length: 500

Stop sequences
Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

Best of: 1

Inject start text

Inject restart text

Show probabilities: Off