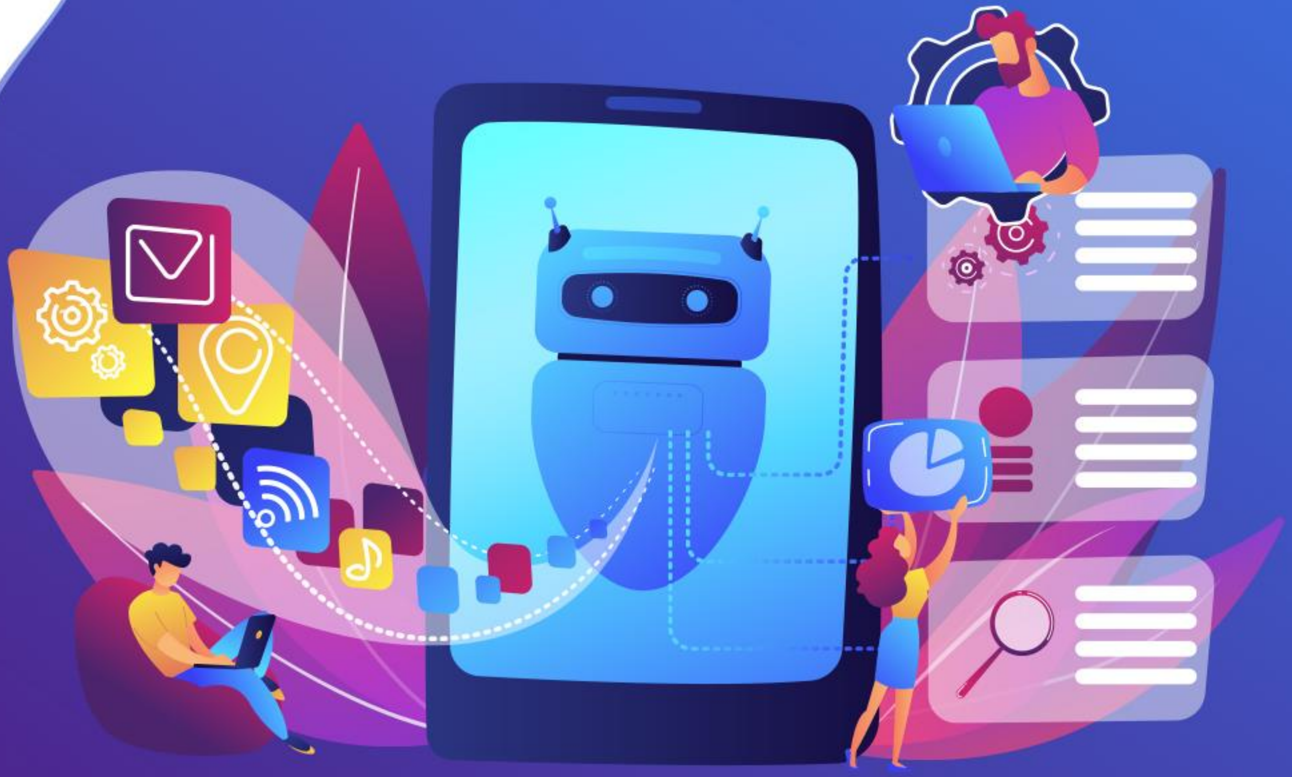


Natural language processing

Welcome on board

Lecture 1

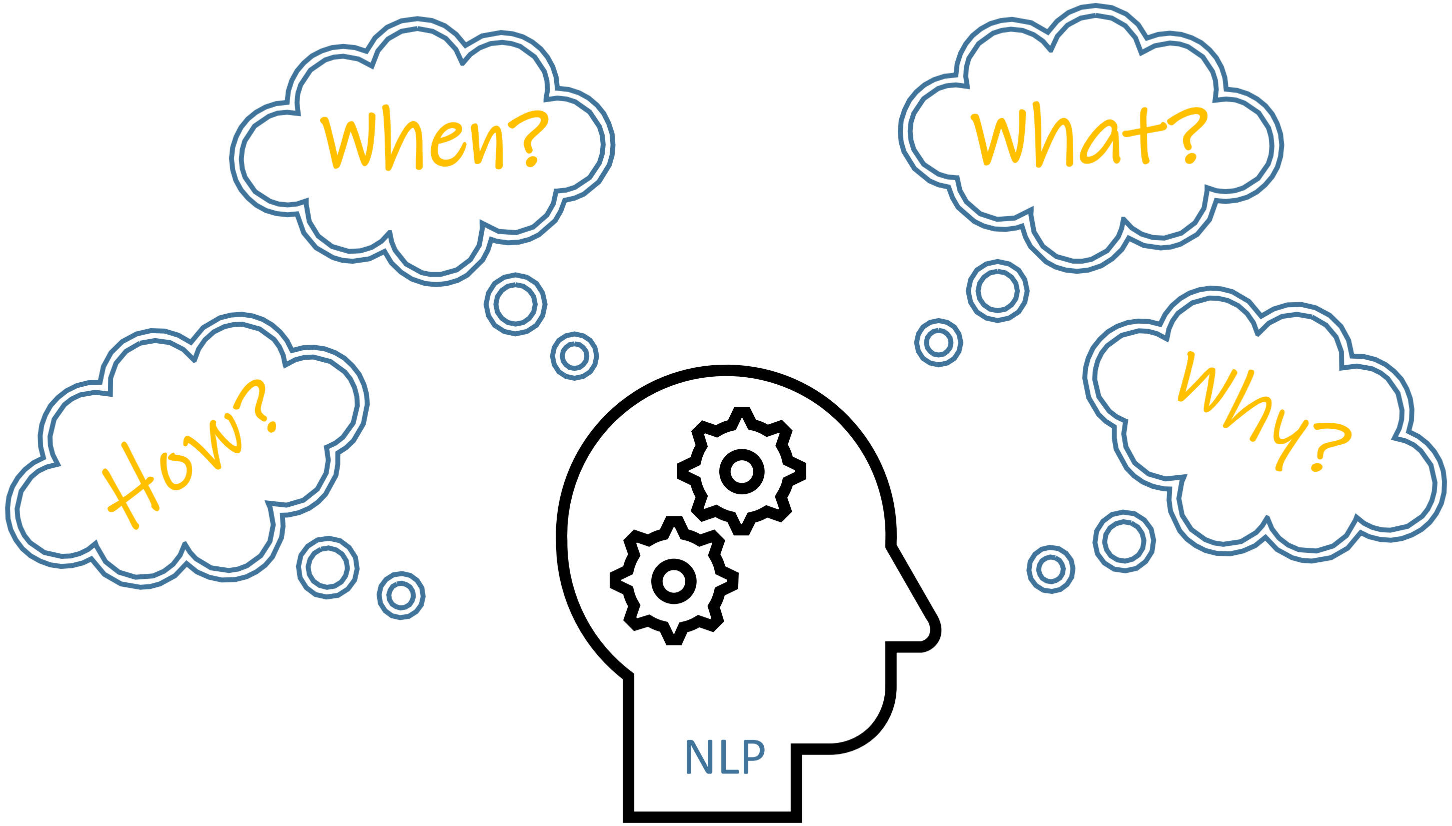
Rana Salama

INTRODUCTION TO STATISTICAL NLP

FALL 2020

Outline

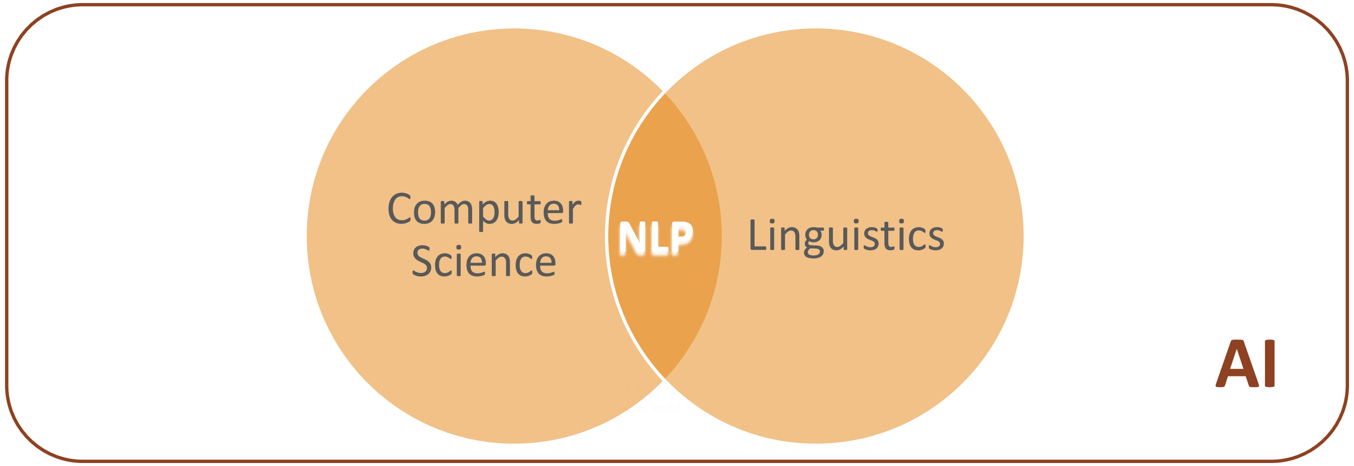
- **Course Introduction**
- Course Information
- Deadlines



Course Introduction

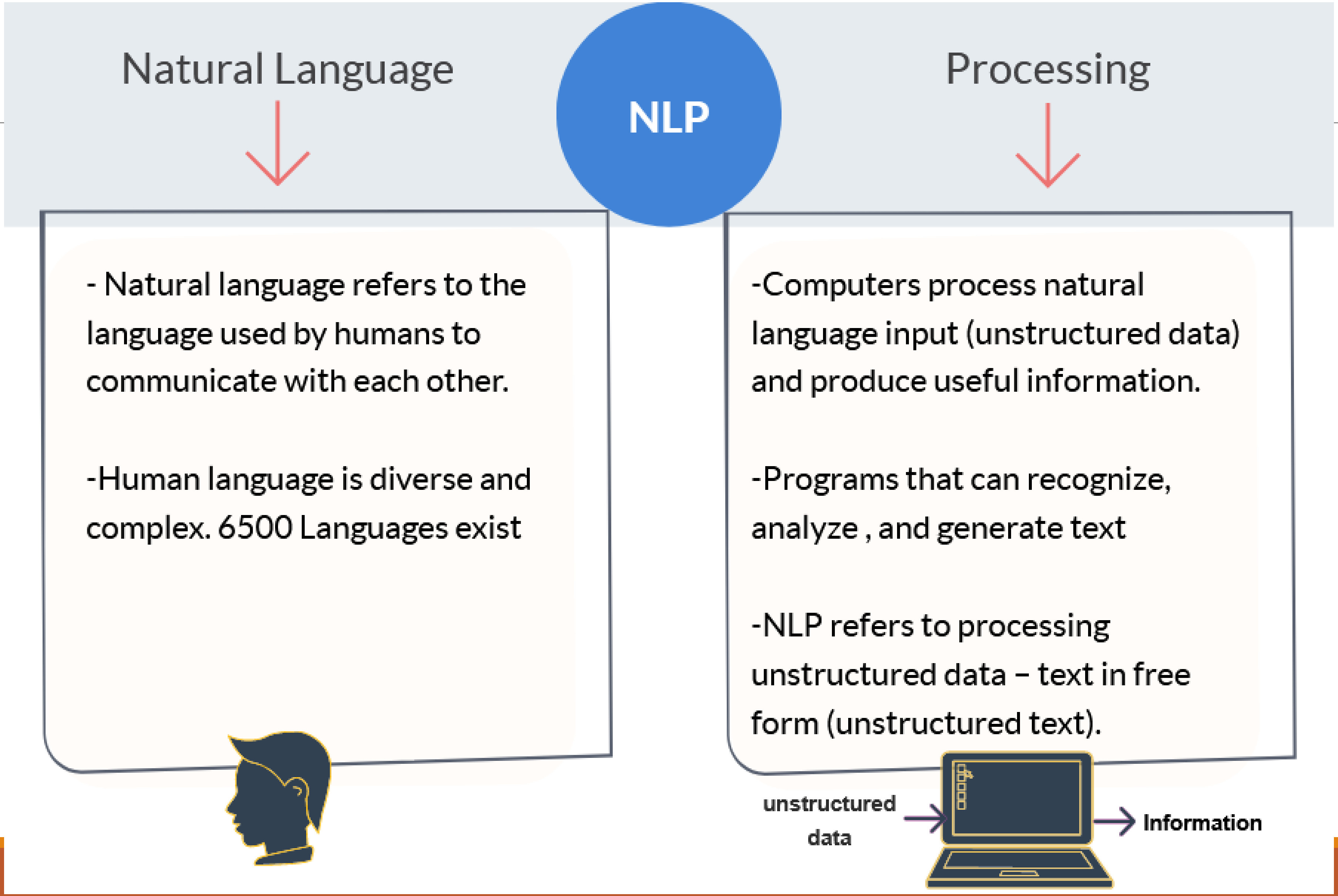
What is NLP?

- Natural language processing (NLP) is the interdisciplinary field of computer science and linguistics



- **Goal:** Have computers *understand* natural language in order to perform useful tasks (human-machine communication or improving human-human communication)

What is NLP?



Natural Language

- Natural language refers to the language used by humans to communicate with each other.

-Human language is diverse and complex. 6500 Languages exist

Processing

-Computers process natural language input (unstructured data) and produce useful information.

-Programs that can recognize, analyze , and generate text

-NLP refers to processing unstructured data – text in free form (unstructured text).

Natural Language

Language = Words (Dictionary) and Rules (Grammar)

Dictionary: set of words defined in the language; open (dynamic)

Grammar: set of rules which describe what is allowable in a language

-**Classical Grammars**: meant for humans; mainly supported by examples; no (or almost no) formal description tools; cannot be programmed

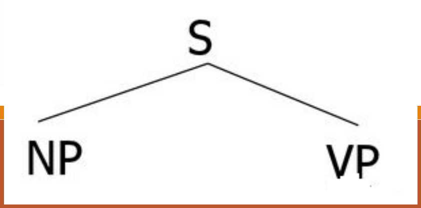
A complete sentence must include a noun and a verb

"the bird flew"

- **Explicit Grammar**: (CFG, Dependency Grammars, Link Grammars,...) formal description; can be programmed & tested on data (texts)

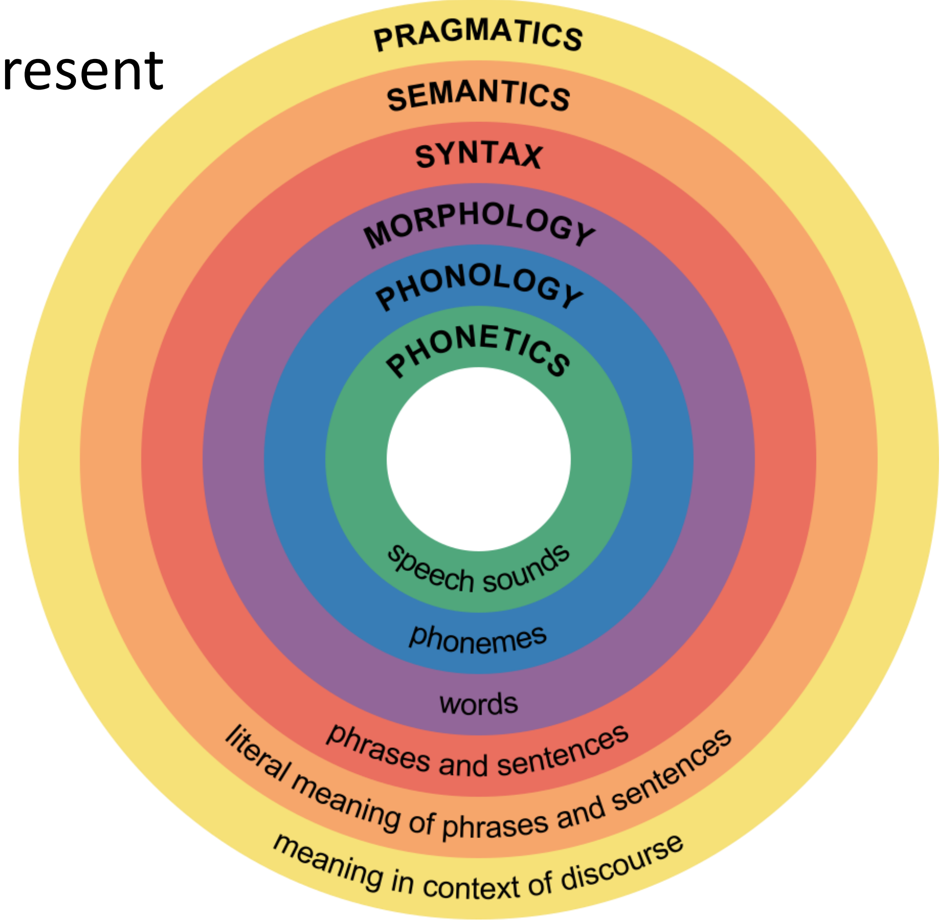
S-> NP VP

"I prefer a morning flight"



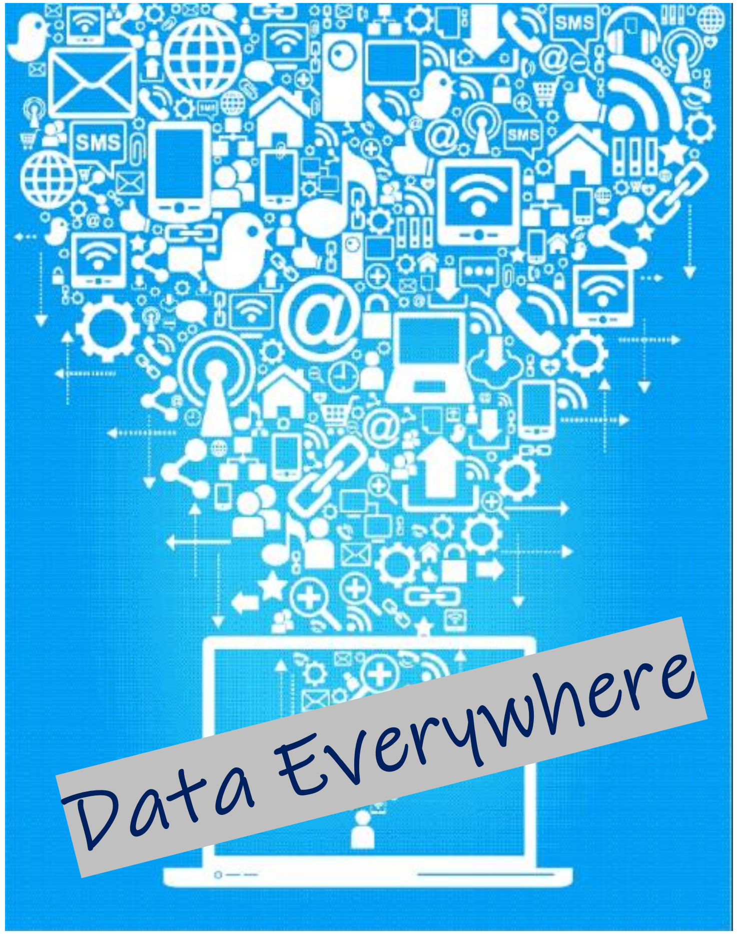
Layers of linguistic analysis

- Linguistic is the science of language.
- Its study include 6 basic levels (more or less explicitly present in most theories):
  - ❑Phonetics and Phonology: sounds
  - ❑Morphology : word formation
  - ❑Syntax : structural relationships between words, sentence formation
  - ❑Semantics : knowledge of meaning
  - ❑Pragmatic: connected sentences
- Each level has an input and output representation
- Output from one level is the input to the next (upper) level
- Sometimes levels might be skipped (merged) or split

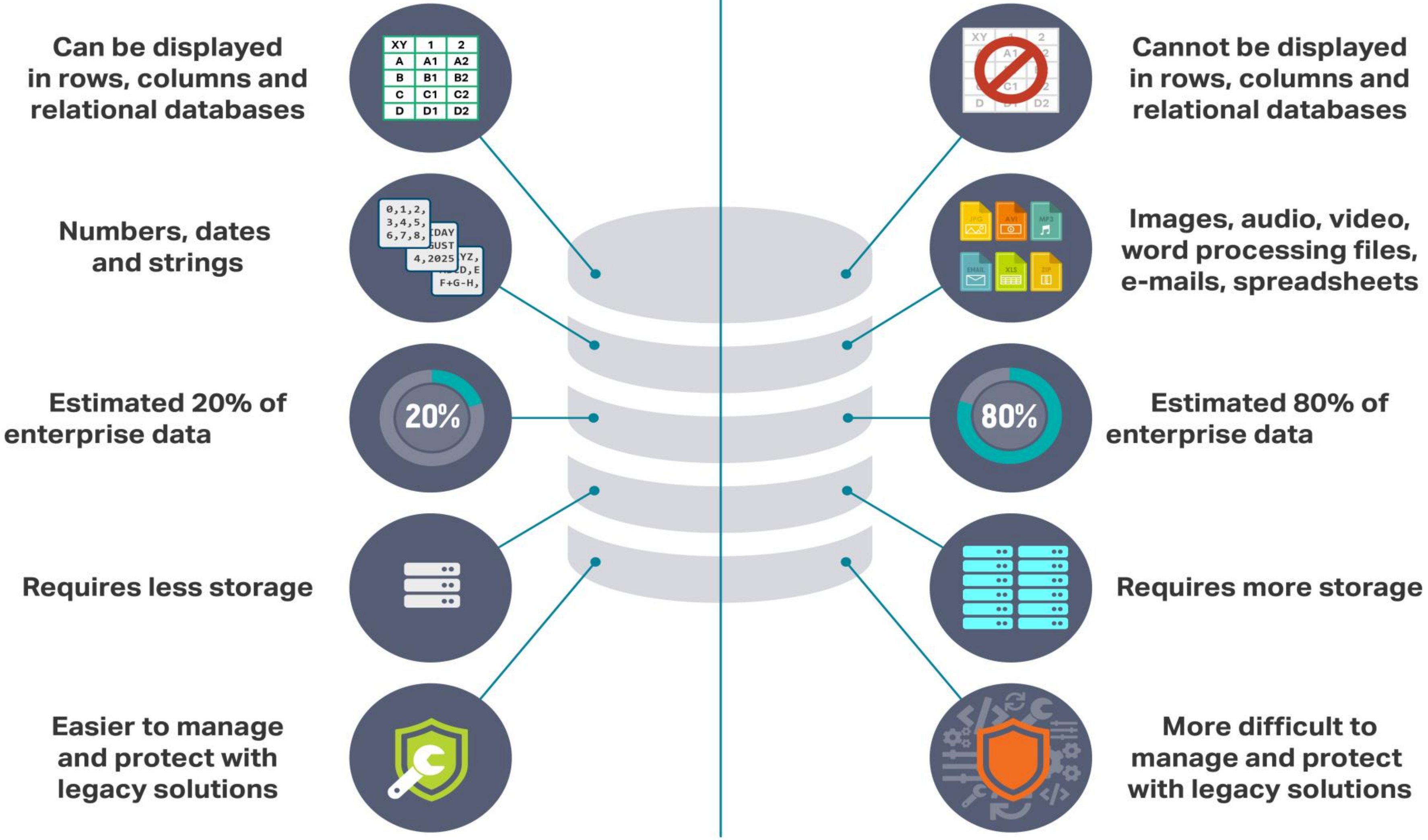


Why NLP?

- An enormous amount of knowledge is now available in machine readable form as unstructured natural language text from different resources (heterogeneous data)
- Going from the largely unstructured languages of the web to useful information
- Conversational agents are becoming an important form of human-computer communication
- Much of human-human communication is now mediated by computers.
- Very cool stuff! And with lots of commercial interest.



Structured Data vs Unstructured Data



XY	1	2
A	A1	A2
B	B1	B2
C	C1	C2
D	D1	D2

Structured Data

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data

Requires less storage

Easier to manage and protect with legacy solutions

Unstructured Data

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data

Requires more storage

More difficult to manage and protect with legacy solutions

https://www.igneous.io/blog/structured-data-vs-unstructured-data

Why NLP ?

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoatnt tihng is taht the frist and lsat ltteers be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm.

Why NLP ?

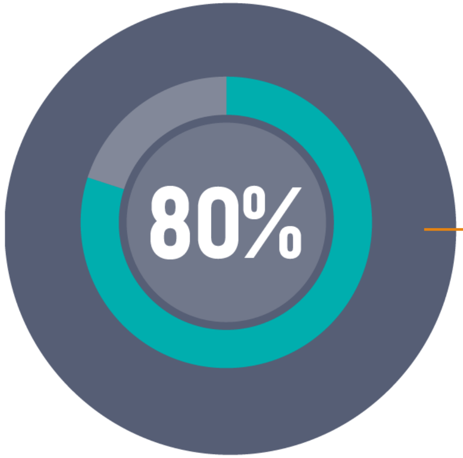
- People have no trouble understanding language
  - Commonsense knowledge
  - Reasoning capacity
  - Experience
- However, Computers have
  - No commonsense knowledge
  - No reasoning capacity

Unless we teach them!

Why NLP ?

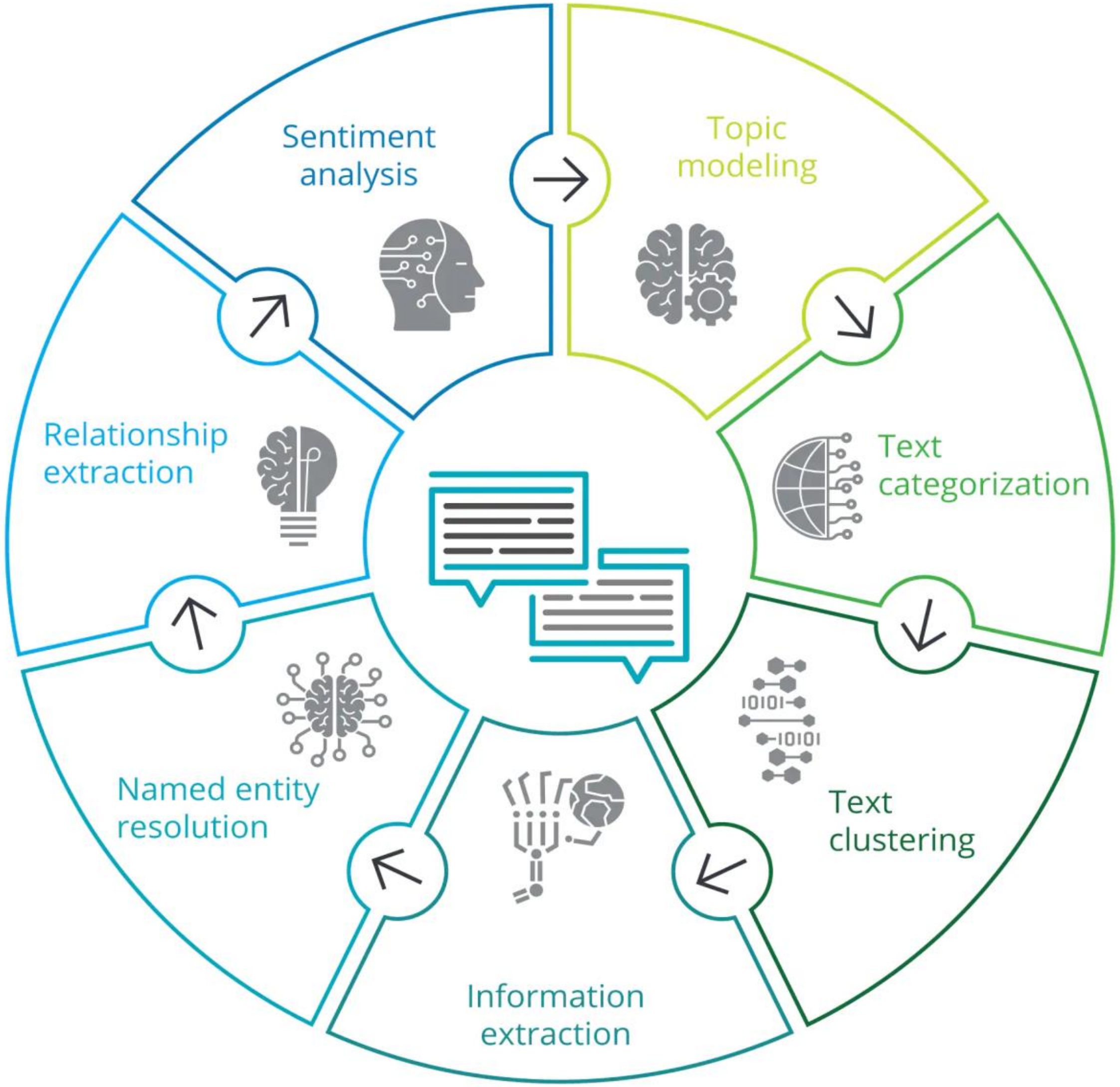
We need computers to :

- Classify text into categories
- Index and search large texts
- Automatic machine translation
- Speech understanding – Understand phone conversations
- Information extraction – Extract useful information from resumes
- Automatic summarization – Condense 1 book into 1 page •
- Question answering
- Knowledge acquisition
- Text generation / dialogs



Unstructured Data

NLP



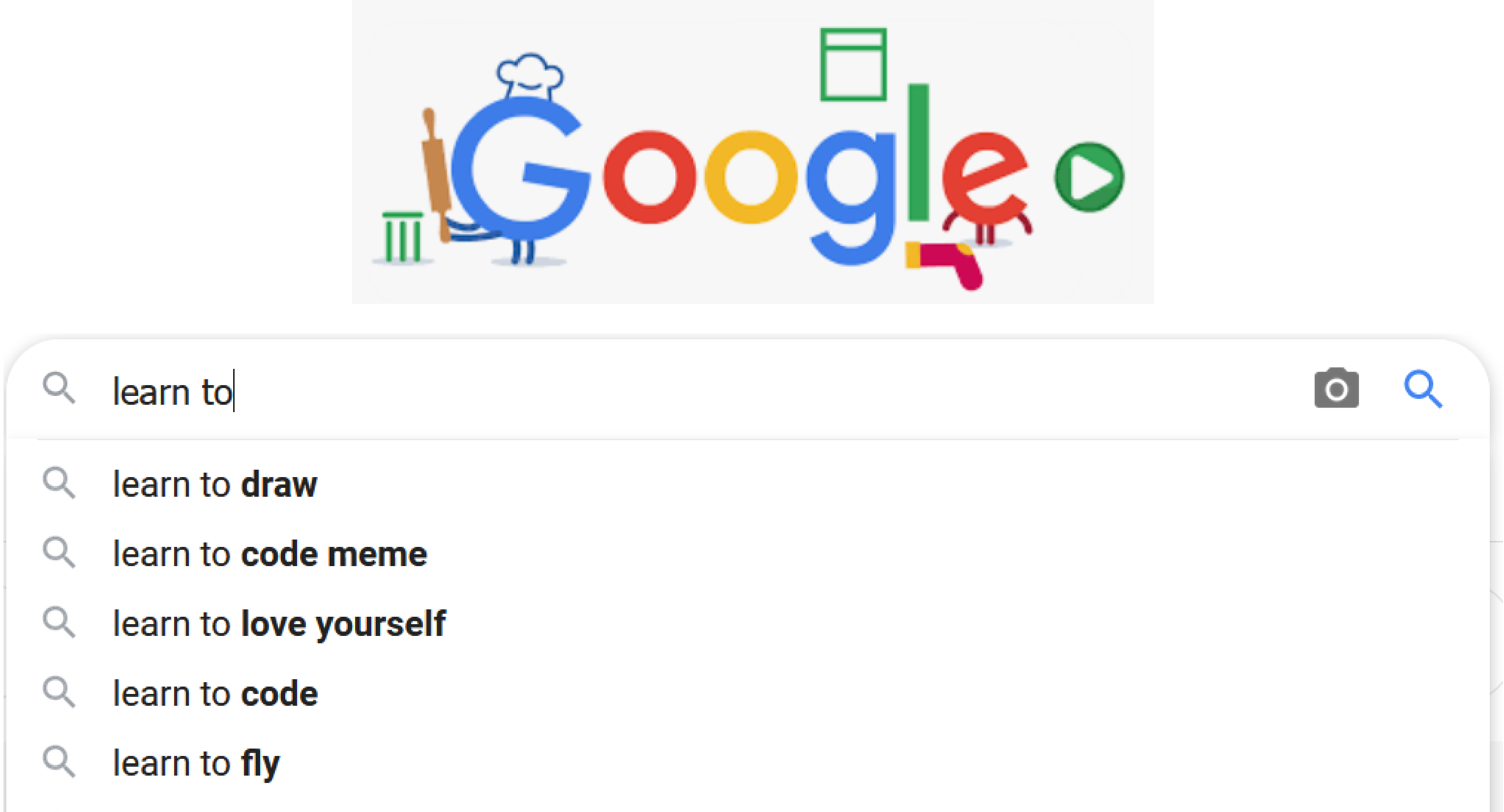
Source: Deloitte analysis.

NLP Applications

NLP Everywhere

USER PERSPECTIVE

Auto-complete



learn to

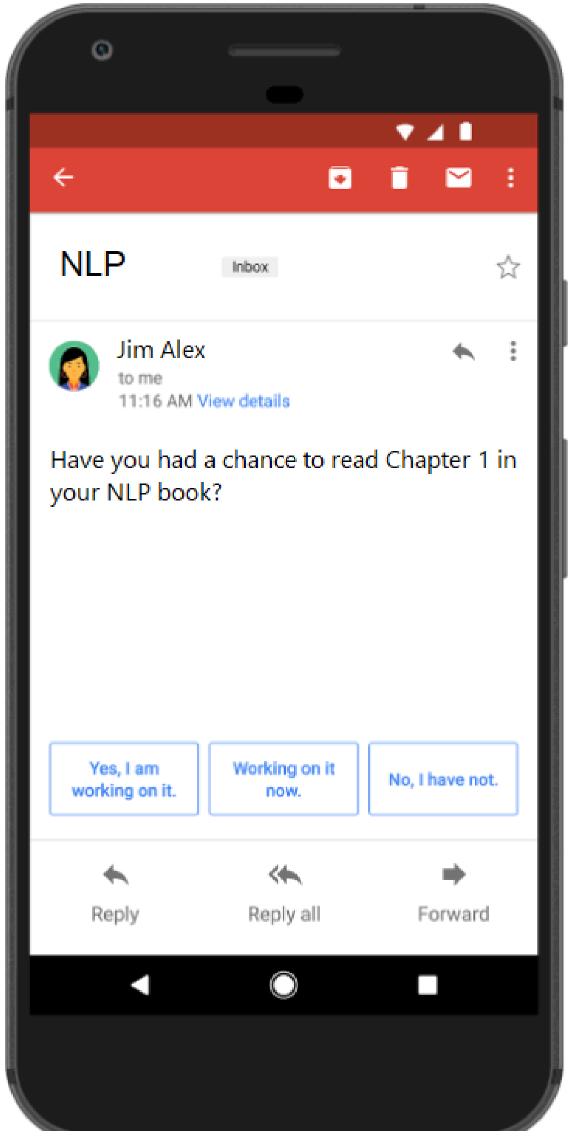
learn to **draw**

learn to **code meme**

learn to **love yourself**

learn to **code**

learn to **fly**


NLP

Inbox

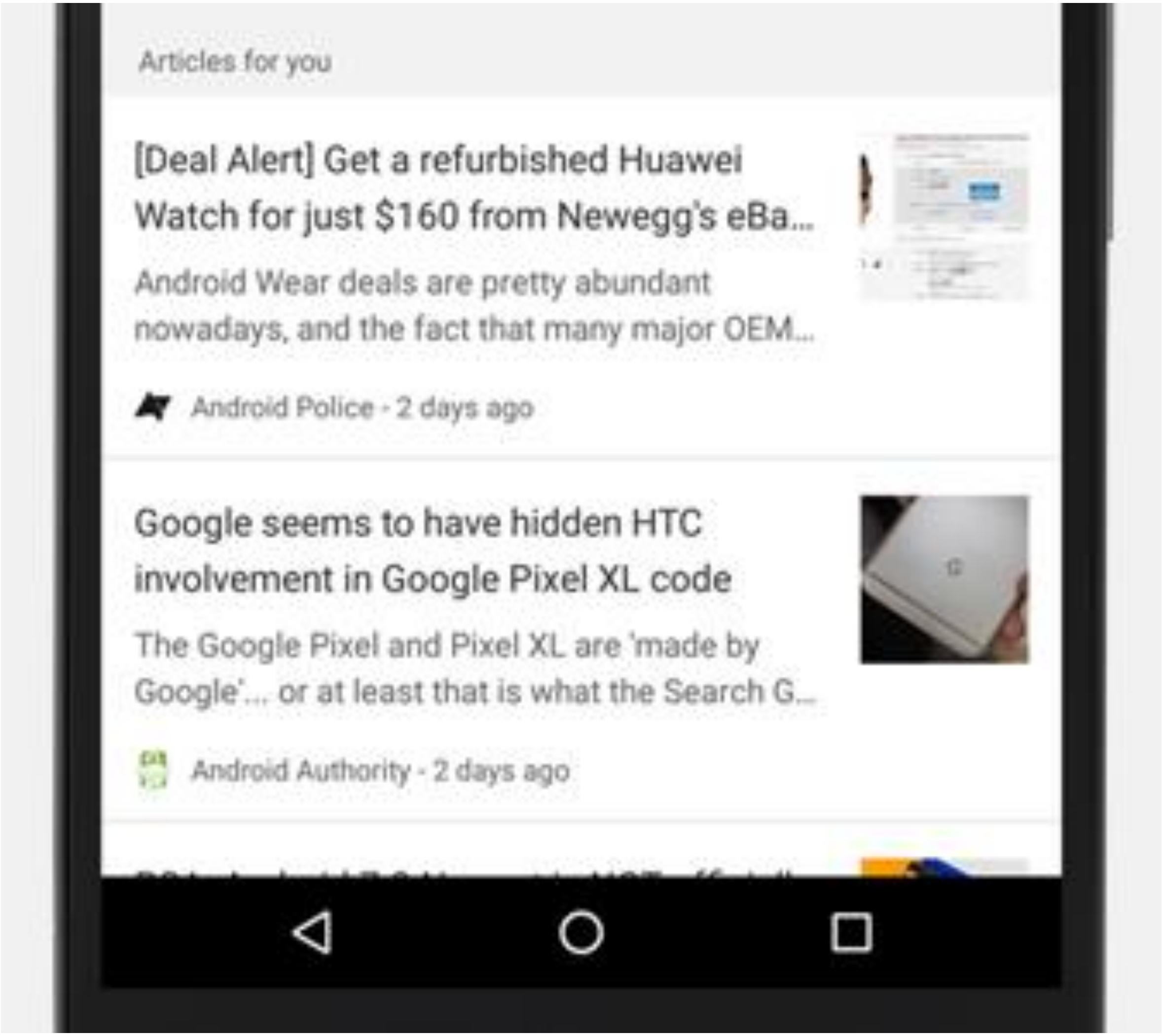
Jim Alex

to me

11:16 AM View details

Have you had a chance to read Chapter 1 in your NLP book?

Automatic Replies


Articles for you

[Deal Alert] Get a refurbished Huawei Watch for just $160 from Newegg's eBa...

Android Wear deals are pretty abundant nowadays, and the fact that many major OEM...

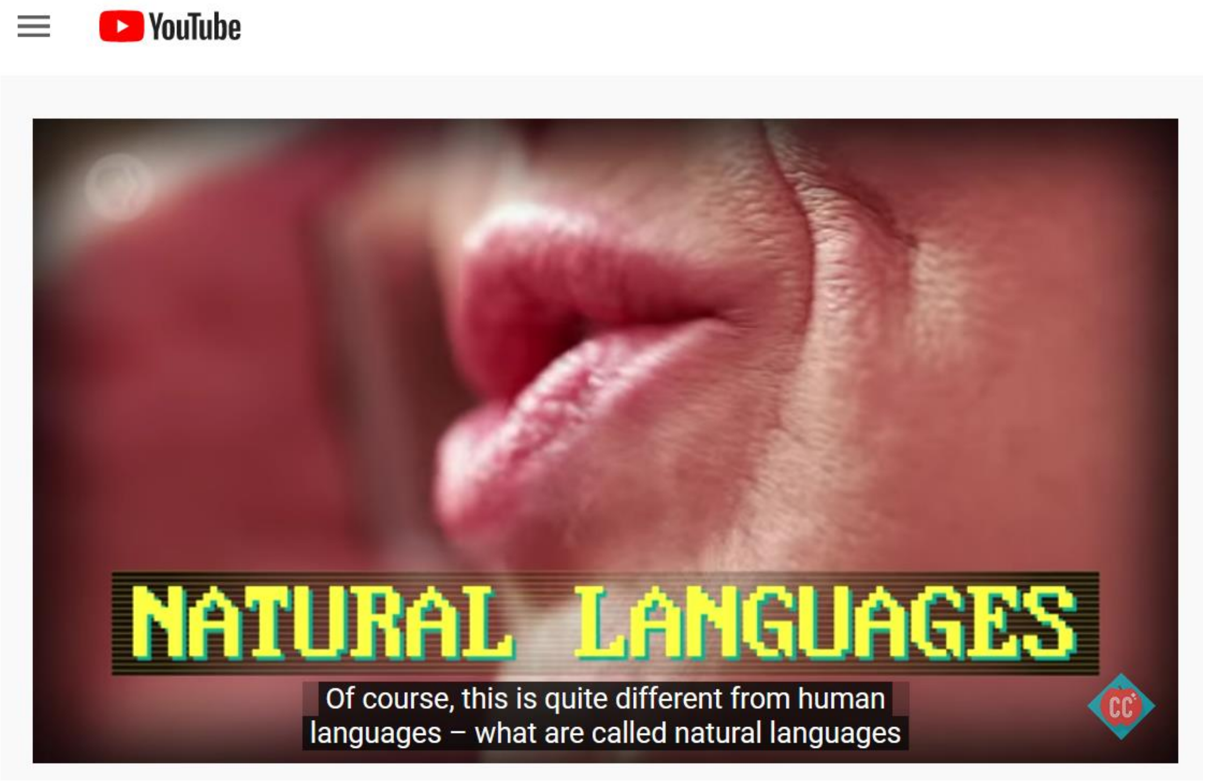
Android Police - 2 days ago

Google seems to have hidden HTC involvement in Google Pixel XL code

The Google Pixel and Pixel XL are 'made by Google'... or at least that is what the Search G...

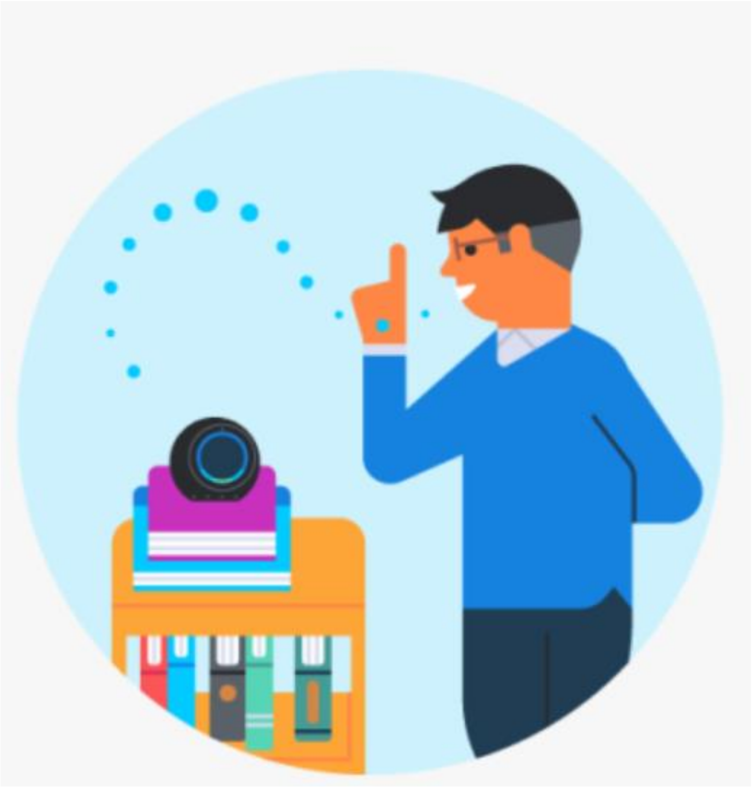
Android Authority - 2 days ago

News site's suggested articles



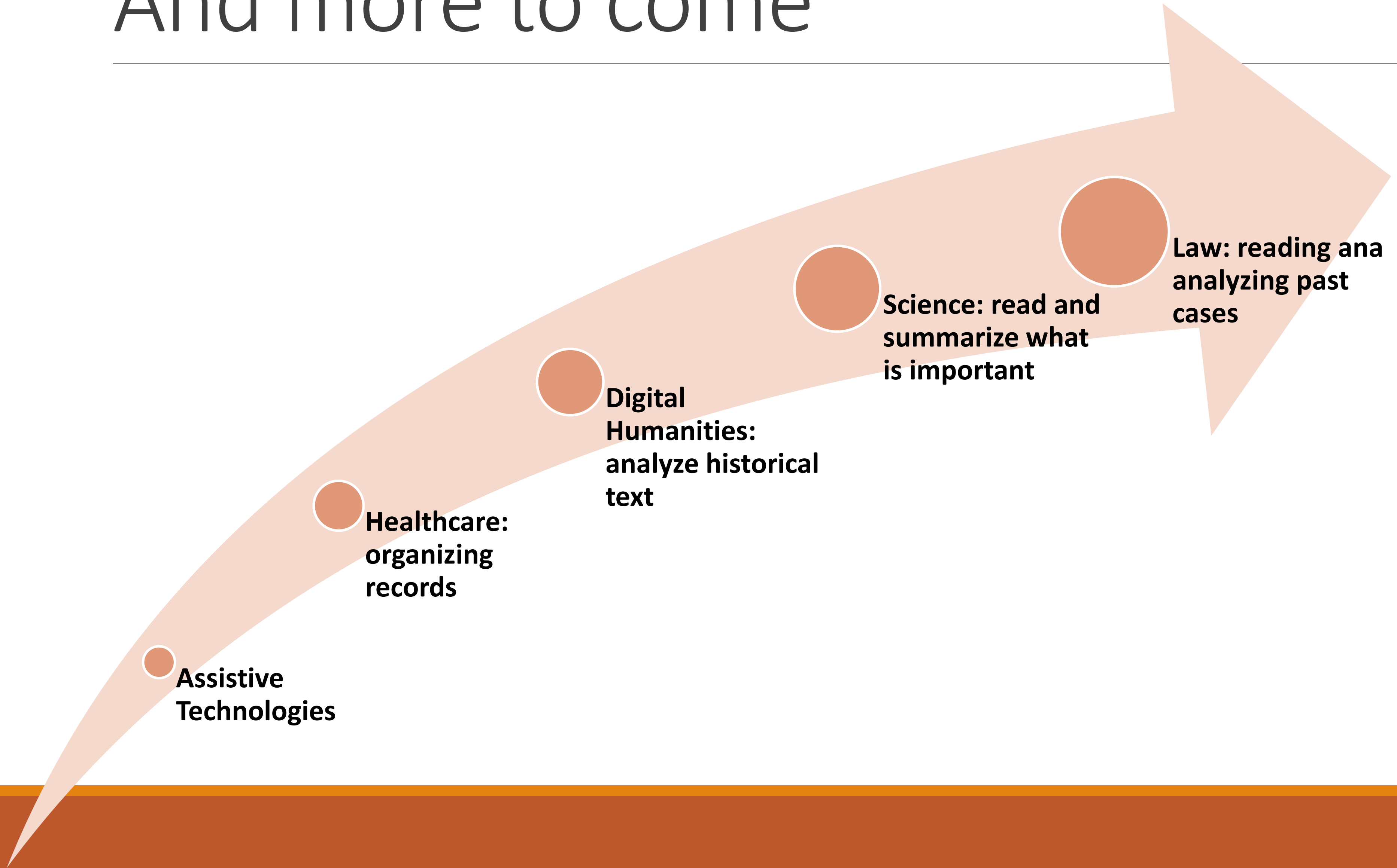
Of course, this is quite different from human languages – what are called natural languages

Auto-generated video captions



Hey Alexa,..

And more to come



Assistive Technologies

Healthcare: organizing records

Digital Humanities: analyze historical text

Science: read and summarize what is important

Law: reading ana analyzing past cases

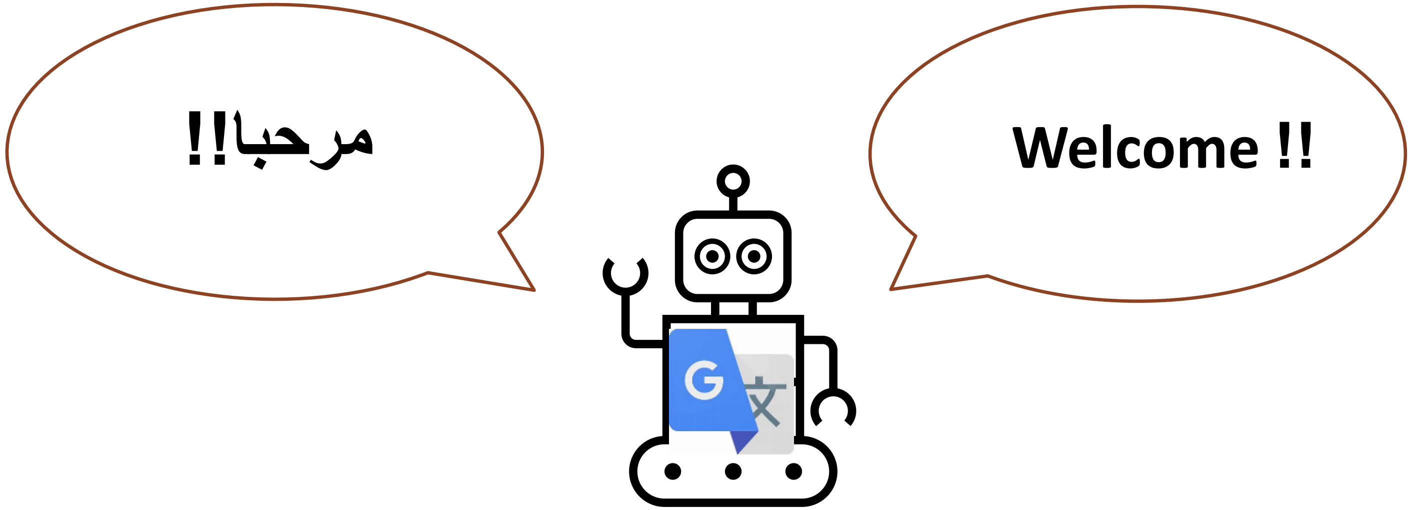
and more…..

NLP Applications

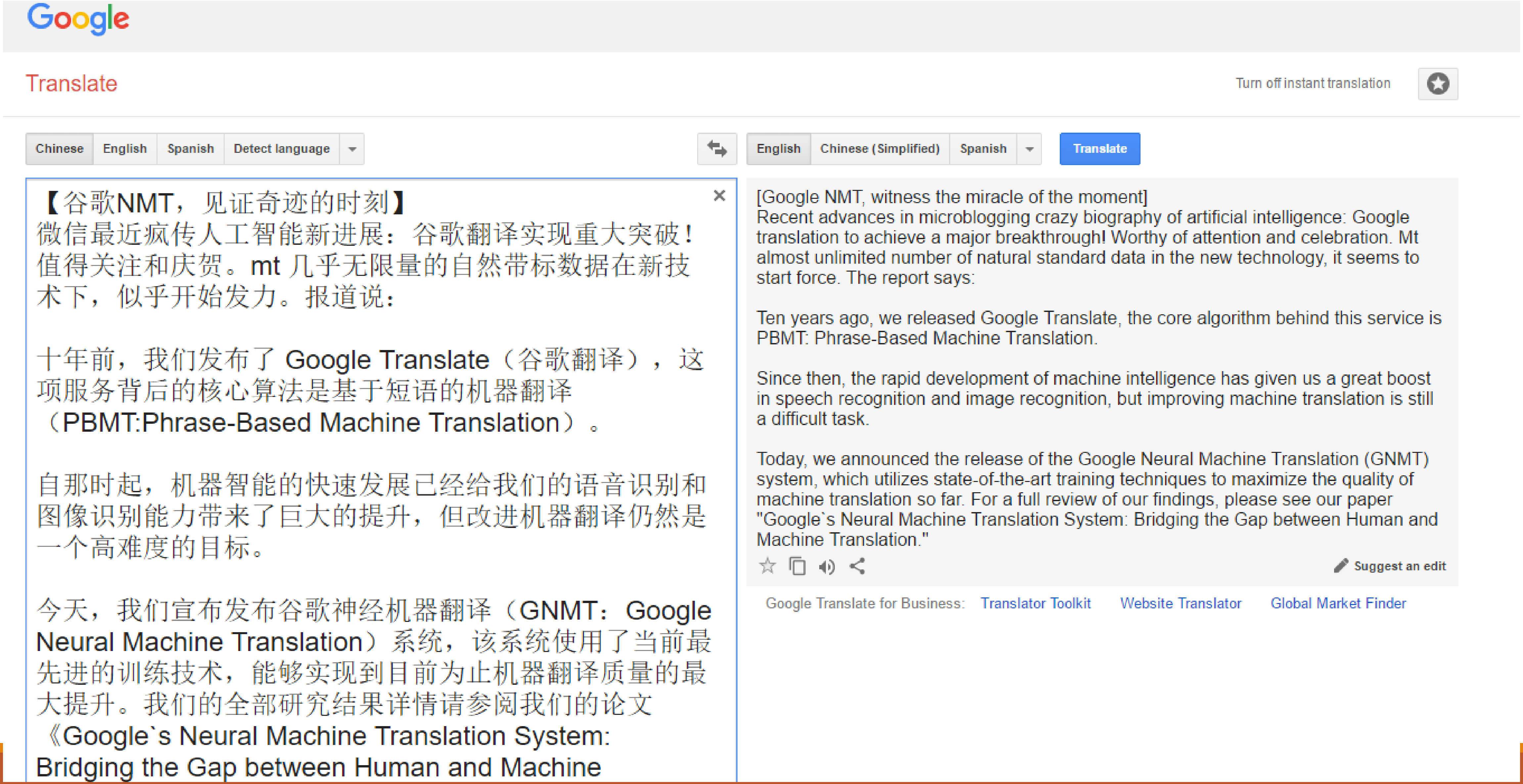
A CLOSER LOOK

Machine Translation

- **Goal**: Computers understand and translate between one language and another: e.g. Google Translate.



Machine Translation



【谷歌NMT，见证奇迹的时刻】
微信最近疯传人工智能新进展：谷歌翻译实现重大突破！值得关注和庆贺。mt 几乎无限量的自然带标数据在新技术下，似乎开始发力。报道说：

十年前，我们发布了 Google Translate（谷歌翻译），这项服务背后的核心算法是基于短语的机器翻译（PBMT:Phrase-Based Machine Translation）。

自那时起，机器智能的快速发展已经给我们的语音识别和图像识别能力带来了巨大的提升，但改进机器翻译仍然是一个高难度的目标。

今天，我们宣布发布谷歌神经机器翻译（GNMT：Google Neural Machine Translation）系统，该系统使用了当前最先进的训练技术，能够实现到目前为止机器翻译质量的最大提升。我们的全部研究结果详情请参阅我们的论文《Google`s Neural Machine Translation System: Bridging the Gap between Human and Machine

[Google NMT, witness the miracle of the moment]
Recent advances in microblogging crazy biography of artificial intelligence: Google translation to achieve a major breakthrough! Worthy of attention and celebration. Mt almost unlimited number of natural standard data in the new technology, it seems to start force. The report says:

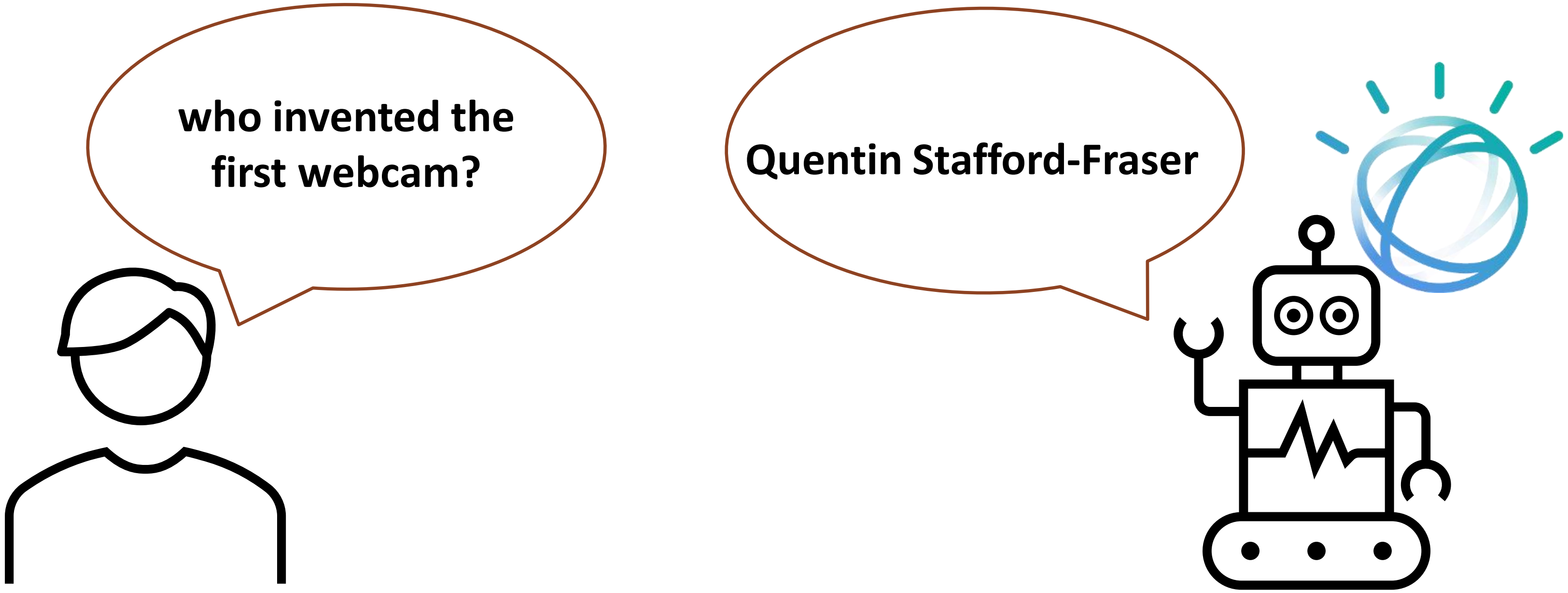
Ten years ago, we released Google Translate, the core algorithm behind this service is PBMT: Phrase-Based Machine Translation.

Since then, the rapid development of machine intelligence has given us a great boost in speech recognition and image recognition, but improving machine translation is still a difficult task.

Today, we announced the release of the Google Neural Machine Translation (GNMT) system, which utilizes state-of-the-art training techniques to maximize the quality of machine translation so far. For a full review of our findings, please see our paper "Google`s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."

Questioning Answering

- **Goal**: building systems that automatically answer questions posed by humans in a natural language.
- IBM's Watson Jeopardy! (2011), DARPA who/what/where…, Ask Jeeves



Sentiment Analysis

Discovering people opinions, emotions and feelings about a product or service

http://paxcom.net/blog/why-is-customer-sentiment-analysis-important-for-your-brand/

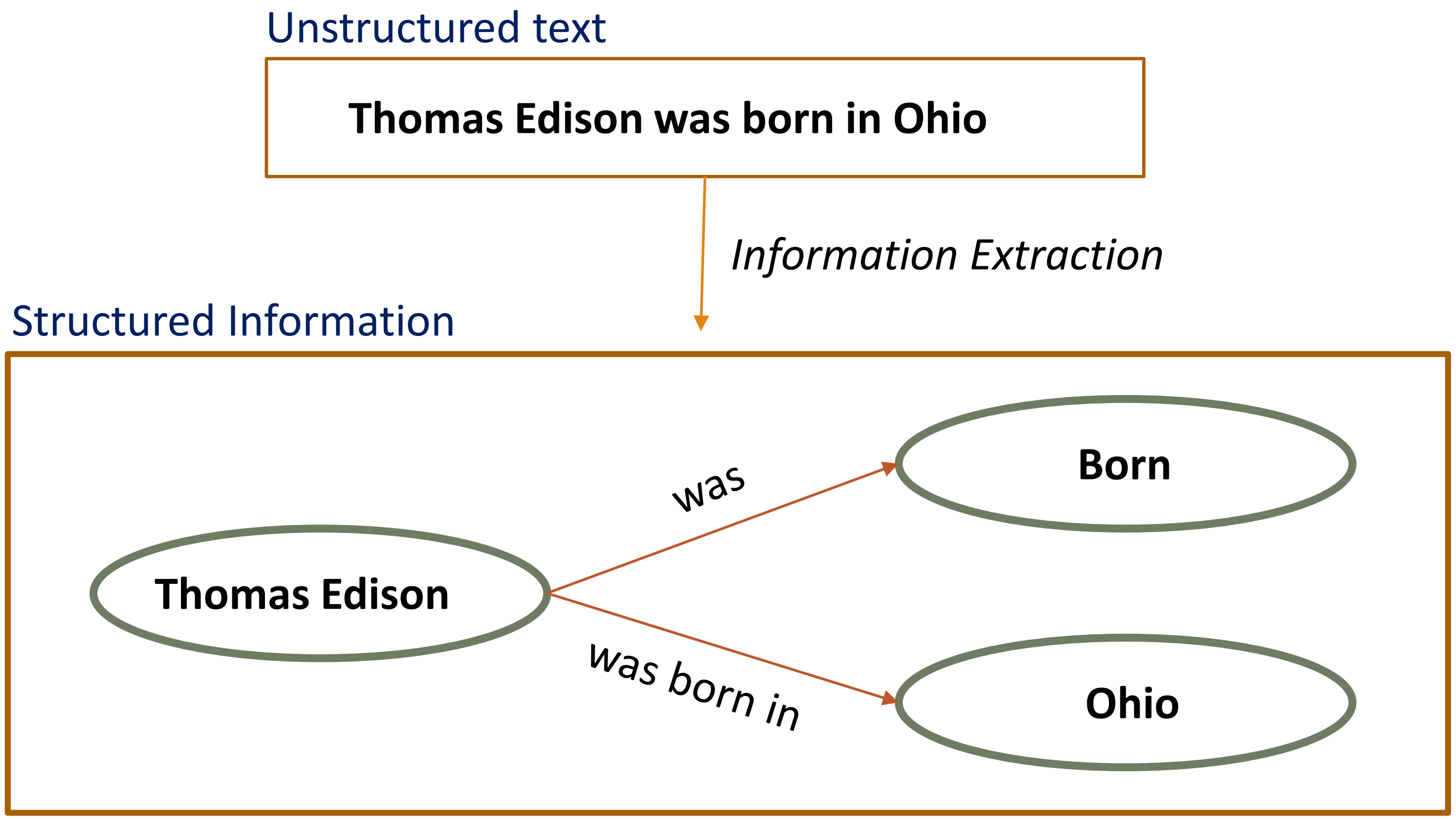
Sentiment Analysis

- Wow, great place!
- Wow, 35 minutes to get a cup of coffee? Great job.
- Not great but works as expected.
- At first I hated it, but once the story hooked me, I found it difficult to put the book down

http://paxcom.net/blog/why-is-customer-sentiment-analysis-important-for-your-brand/

Information Extraction

- **Goal**: Extracting specific information from textual sources.



Unstructured text

Thomas Edison was born in Ohio

Information Extraction

Structured Information

Thomas Edison

was

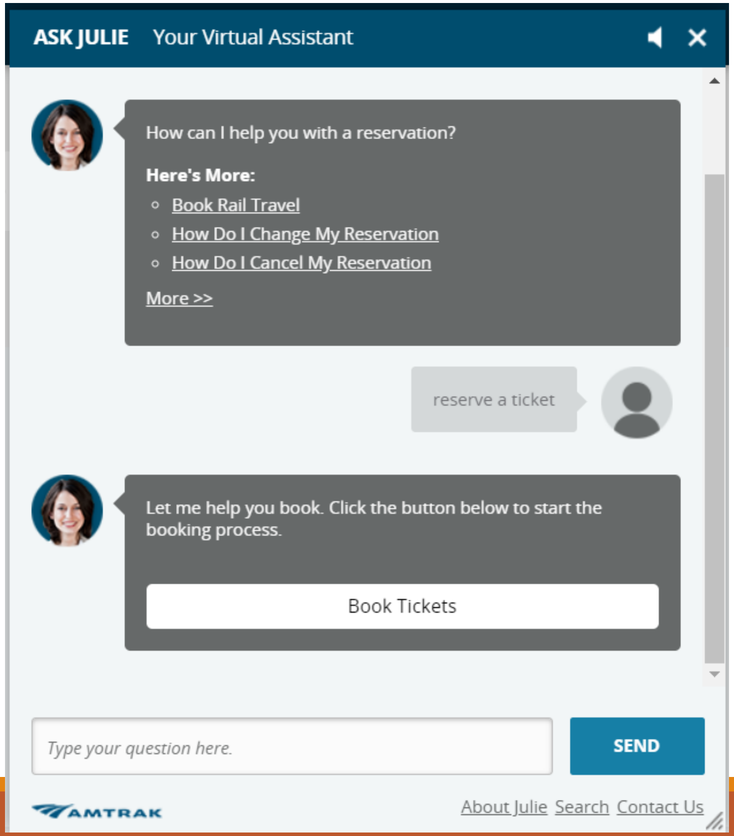
Born

was born in

Ohio

Dialogue Systems

- Goal: A computer system intended to converse with a human.
- Example: Amtrak's 'Julie' and Google Assistant



ASK JULIE Your Virtual Assistant

How can I help you with a reservation?

Here's More:

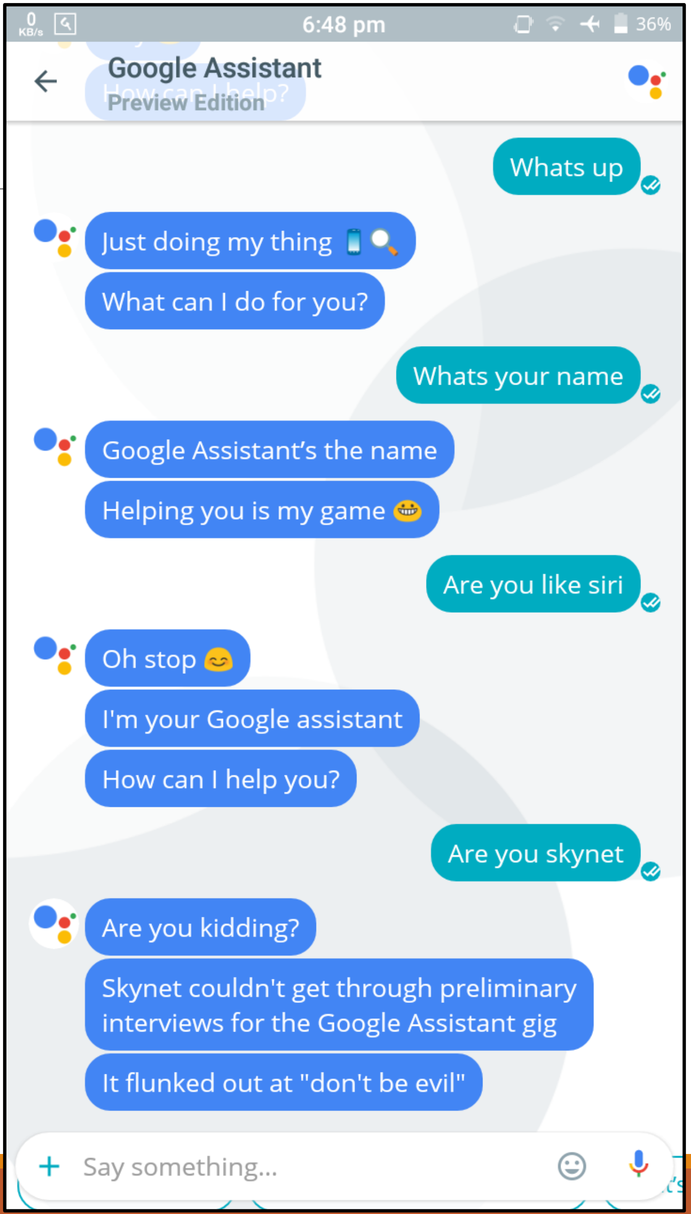
- Book Rail Travel
- How Do I Change My Reservation
- How Do I Cancel My Reservation

More >>

reserve a ticket

Let me help you book. Click the button below to start the booking process.

Book Tickets
Type your question here.
SEND
AMTRAK
About Julie Search Contact Us



Whats up

Just doing my thing 📱🔍

What can I do for you?

Whats your name

Google Assistant's the name

Helping you is my game 😁

Are you like siri

Oh stop 😊

I'm your Google assistant

How can I help you?

Are you skynet

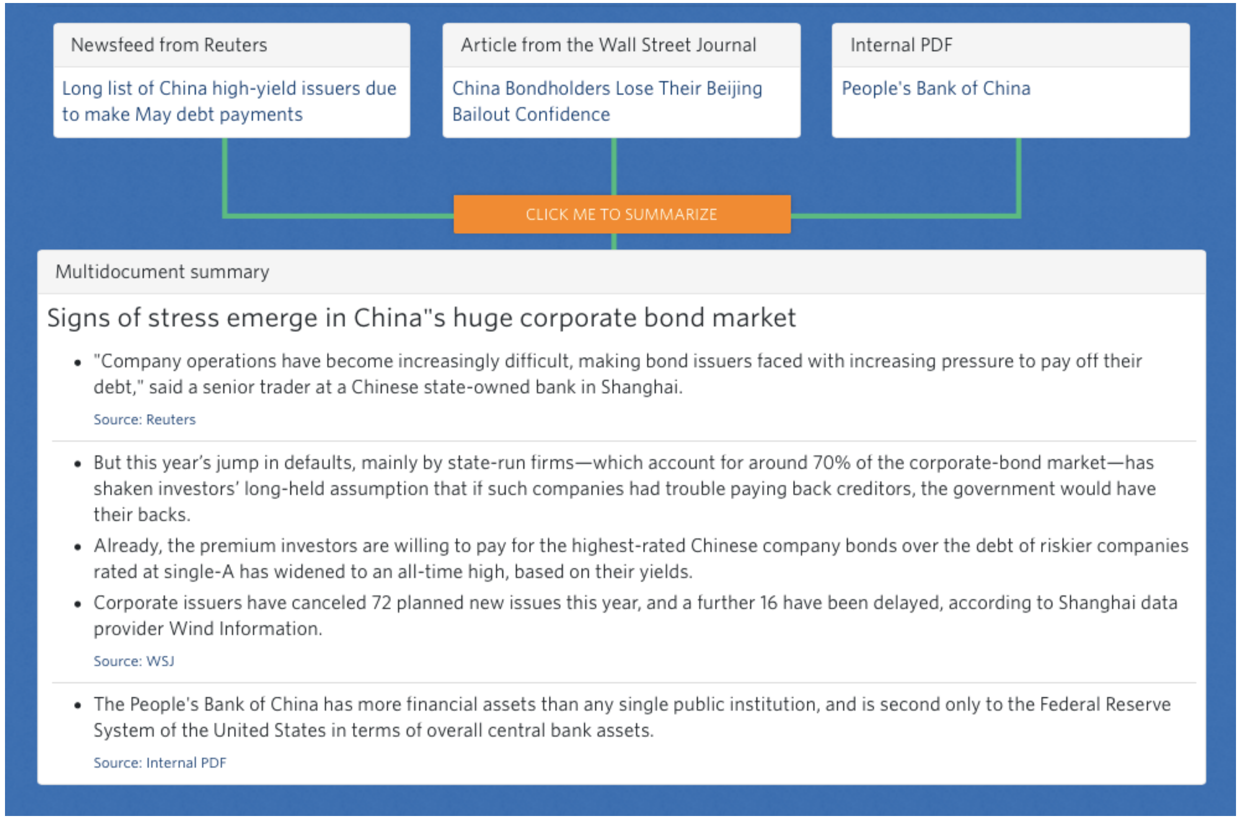
Are you kidding?

Skynet couldn't get through preliminary interviews for the Google Assistant gig

It flunked out at "don't be evil"

Summarization

- **Goal**: Summarizing very large amounts of text or speech: e.g. your email, the news, voicemail
- Example: https://www.agolo.com



Newsfeed from Reuters
Long list of China high-yield issuers due to make May debt payments
Article from the Wall Street Journal
China Bondholders Lose Their Beijing Bailout Confidence
Internal PDF
People's Bank of China
CLICK ME TO SUMMARIZE
Multidocument summary

Signs of stress emerge in China"s huge corporate bond market

- "Company operations have become increasingly difficult, making bond issuers faced with increasing pressure to pay off their debt," said a senior trader at a Chinese state-owned bank in Shanghai.

Source: Reuters

- But this year's jump in defaults, mainly by state-run firms—which account for around 70% of the corporate-bond market—has shaken investors' long-held assumption that if such companies had trouble paying back creditors, the government would have their backs.
- Already, the premium investors are willing to pay for the highest-rated Chinese company bonds over the debt of riskier companies rated at single-A has widened to an all-time high, based on their yields.
- Corporate issuers have canceled 72 planned new issues this year, and a further 16 have been delayed, according to Shanghai data provider Wind Information.

Source: WSJ

- The People's Bank of China has more financial assets than any single public institution, and is second only to the Federal Reserve System of the United States in terms of overall central bank assets.

Source: Internal PDF

What other NLP applications that you can think of?

Is NLP difficult?

What Makes NLP Difficult?

- **The nature of languages** → Language encode meaning. Language is learned intuitively

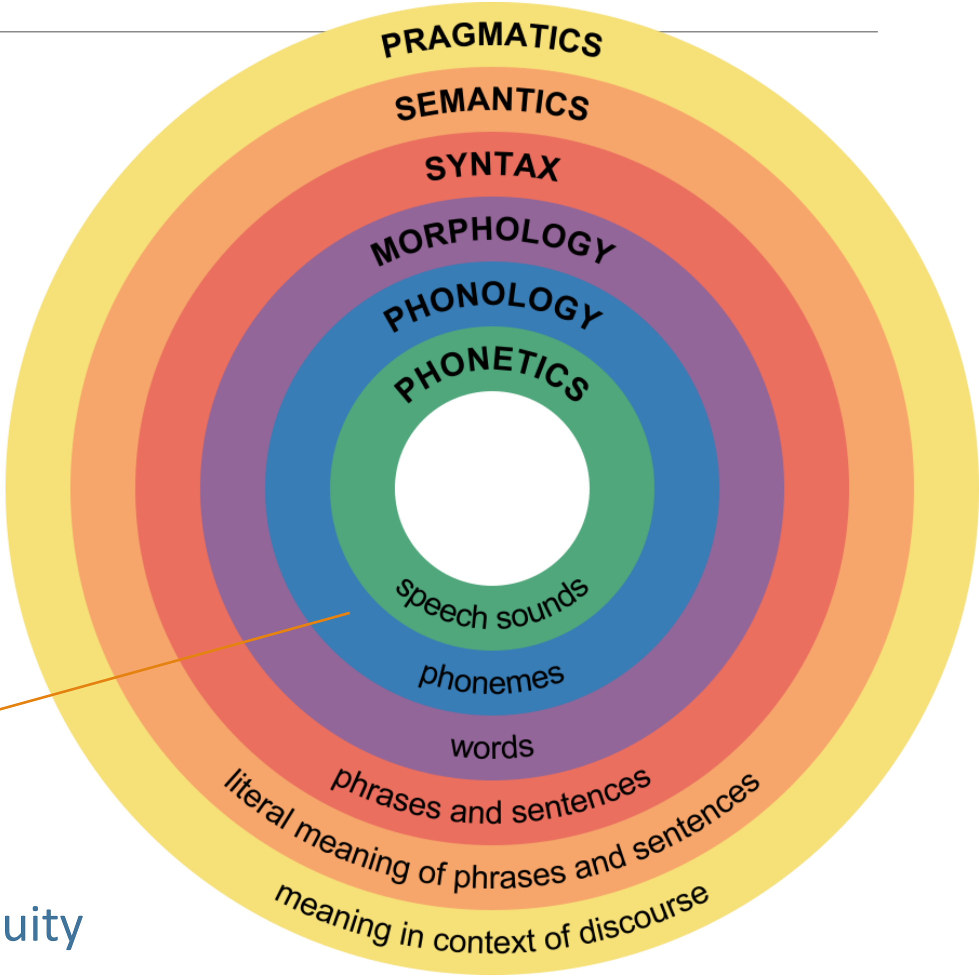
- Non-standard text
  - “ we’re soooo proud of u!”
- Idioms and metaphors
  - “dark horse” “cold feet” “lose face”
- Segmentation
  - “The New York-New Haven railroad”
- Named entities
  - “*Let It Be* sold millions”

→ **Language Variation**

- **Ambiguity**

Ambiguity

- Language is ambiguous
- Ambiguity involves multiple or alternative linguistic structures
- Ambiguity results from the existence of multiple possibilities for linguistic levels
- All 6 levels of linguistic knowledge require resolving ambiguity



-two, too, to
-ice cream, I scream

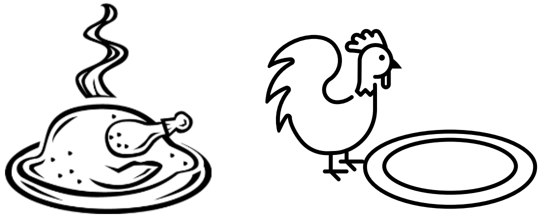
➢ Phonological Ambiguity

Ambiguity

Morphological Ambiguity

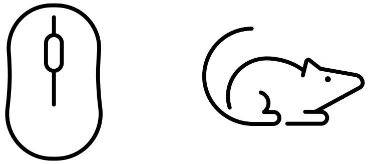
Unlockable: [[un-lock]-able]
[un-[lock-able]]

Syntactic Ambiguity

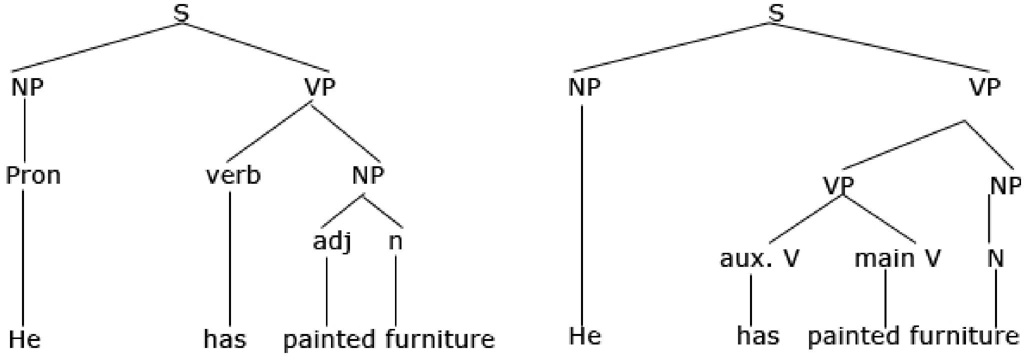


The chicken is ready to eat

Lexical Ambiguity



Pass me the mouse



Ambiguity

I made her duck

- I cooked waterfowl for her
- I cooked the waterfowl that belongs to her
- I created the ceramic duck she owns
- I caused her to quickly lower her head
- And more….

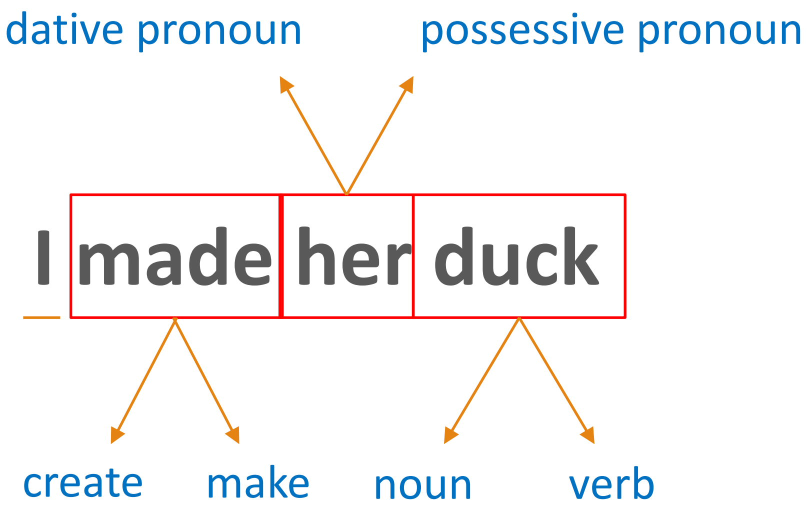
I made her duck for lunch

I made her duck with clay

I made her duck as the ball was about to hit her

Context really matters

Ambiguity



dative pronoun

possessive pronoun

I made her duck

create

make

noun

verb

Dealing with Ambiguity

- Tightly coupled interaction among processing levels; knowledge from other levels can help decide at ambiguous levels.
- Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.
- Probabilistic approaches based on making the most likely choices.
- Don’t do anything, maybe it won’t matter.
  - We’ll leave when the duck is ready to eat.
  - The duck is ready to eat now.
    - Does the “duck” ambiguity matter with respect to whether we can leave?

Making Progress …

The task is difficult! What tools do we need?

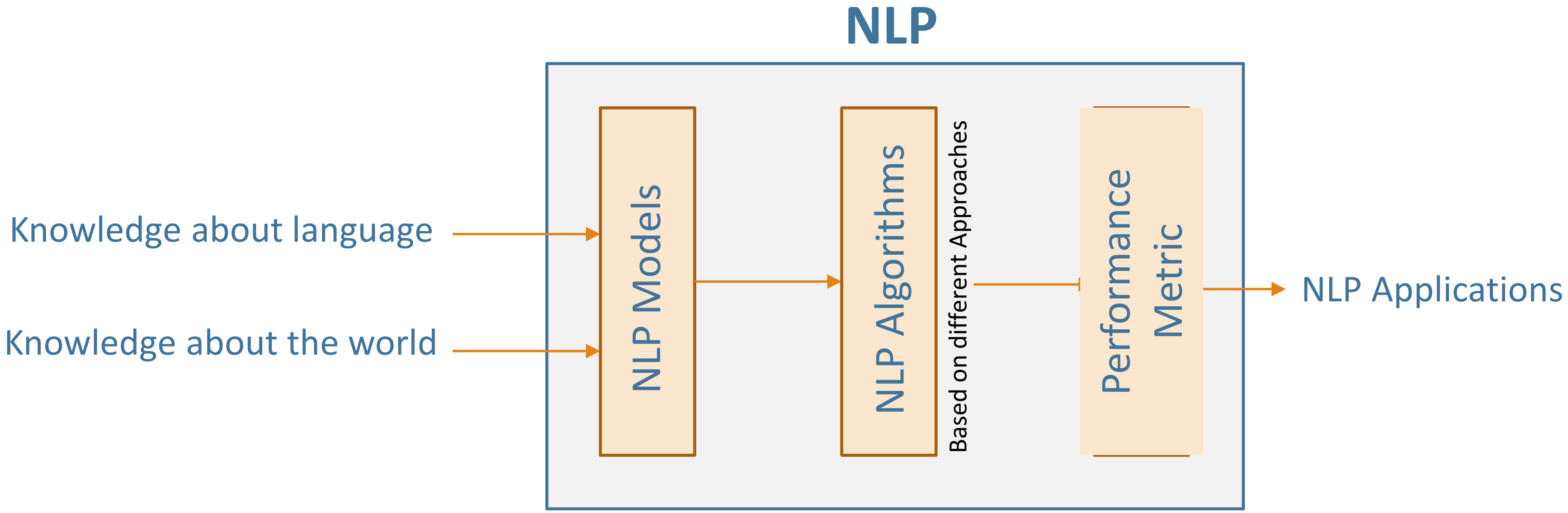
- Knowledge about language
- Knowledge about the world
- A way to combine knowledge sources

How we generally do this:

- probabilistic models built from language data
  - P(“maison” → “house”) high
  - P(“noir” → “moon”) low

Luckily, rough text features can often do half the job.

Making Progress …



NLP Models

State Machines

- Finite state automata, transducers

Formal Rule Systems

- Regular Grammars, Context Free Grammars

Logic

- First order logic, predicate calculus

Probability Theory

- Associating probabilities with the previous machinery
- Crucial for capturing every kind of linguistic knowledge

Vector-space Models

- An algebraic model for representing text documents (and any objects, in general) as vectors

NLP Algorithms

- State space search algorithms, such as dynamic programming
- Expectation-Maximization (EM)
- Machine learning algorithms, such as classifiers and sequence models, which play a significant role in many language processing tasks

Machine Learning

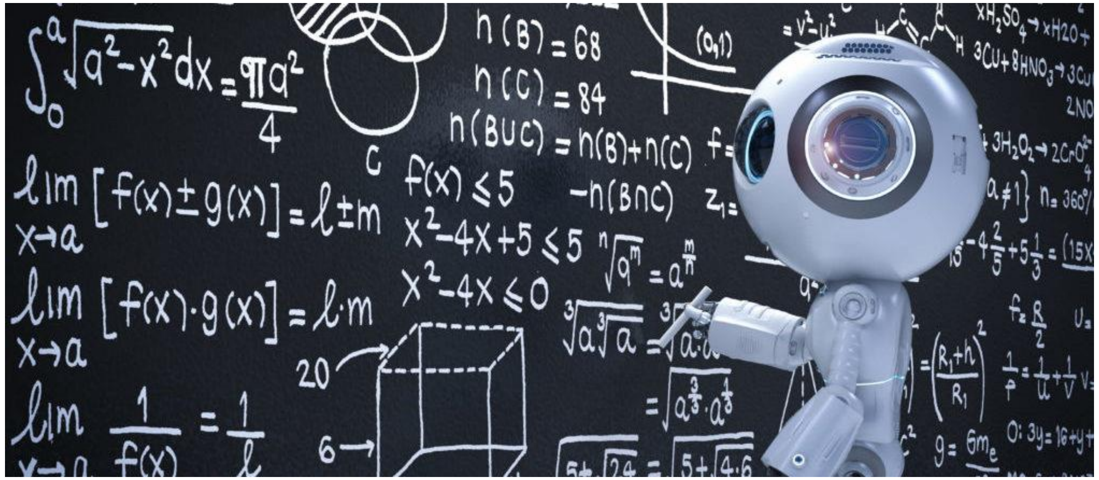
Machine learning based classifiers that are trained to make decisions based on (implicitly or explicitly modeled) features from context

Simple Classifiers:

- Naïve Bayes
- Logistic Regression
- Decision Trees
- Neural Networks

Sequence Models:

- Hidden Markov Models
- Maximum Entropy Markov Models
- Conditional Random Fields
- Recursive Neural Networks (RNNs, LSTMs)



Machines learn from data

NLP Approaches

Rule-based/Symbolic Approaches

- Linguists write rules that are applied by the machines

Corpus-based/Statistical Approaches

- Machines learn the “rules” from training data
  - Annotated data – supervised methods
    - Parallel Corpora: translated text collections
    - Treebanks: manually syntactically analyzed texts
    - Speech Corpora with transcripts
  - Unannotated data – unsupervised methods
  - Semi-supervised methods

Performance Metrics

Methods to evaluate the performance of an NLP system

- Extrinsic Evaluation:
    - Incorporate NLP system into action
- Intrinsic Evaluation:
    - Automatic Evaluation
        - ✓ Does system agree with pre-judged examples?
    - Human Post-hoc Evaluation

Historical Notes

1940s - 1950s

- Automata, regular expressions, and Formal Language theory
- Information Theory and foundational research in speech recognition (digits).
- "MT is rather easy" (MIT, Georgetown)

1960s

- "MT is too hard." (ALPAC report)
- Cancelled all work on Machine Translation in the US.
- CL/NLP research starts (e.g. ACL)
- Transformational paradigm in linguistics (Chomsky)

Historical Notes

1970s

- "MT Winter" in US
- Parsing comes of age: CFGs, ATNs,…
- Speech understanding starts

1980s

- Use of probabilistic models in MT (IBM)
- New focus on model evaluation.
- Natural language generation

1990s

- The first search engine using indexing and Google’s success
- More progress

Historical Notes

2000s

- Use of Neural models
- Multitask learning
- Sequence to sequence models
- Attention and pretrained language models

In summary

- Language to Knowledge
  - A lot more to do
- NLP is difficult
  - Ambiguous and various
  - Context really matters
- Machine and deep Learning
  - With enough data and some math computers can do it
  - The future looks exiting for NLP

Outline

- Course Introduction
- **Course Information**
- Deadlines

This Course

Key theory and methods for statistical NLP:

- Finite State Machines and Transducers
- N-gram language modeling
- Hidden Markov Models
- Discriminative classifiers
- Syntactic and Statistical Parsing
- Vector models of meaning
- Word Embedding
- Deep Learning Models

Practical, real-world applications

- Information extraction
- Spelling correction
- Sentiment analysis
- Machine Translation
- Summarization
- Dialogue Systems

Skills you will need

- Basic probability theory
- Simple linear algebra (vectors, matrices)
- Machine Learning Techniques
- Proficiency in Python programming

Logistics

Lectures: Tuesday 06:10PM - 08:40PM

Office Hours: Thursday 11:00AM-1:00PM or by appoitment

Contact: **raref@gwu.edu**

Syllabus: available on BB

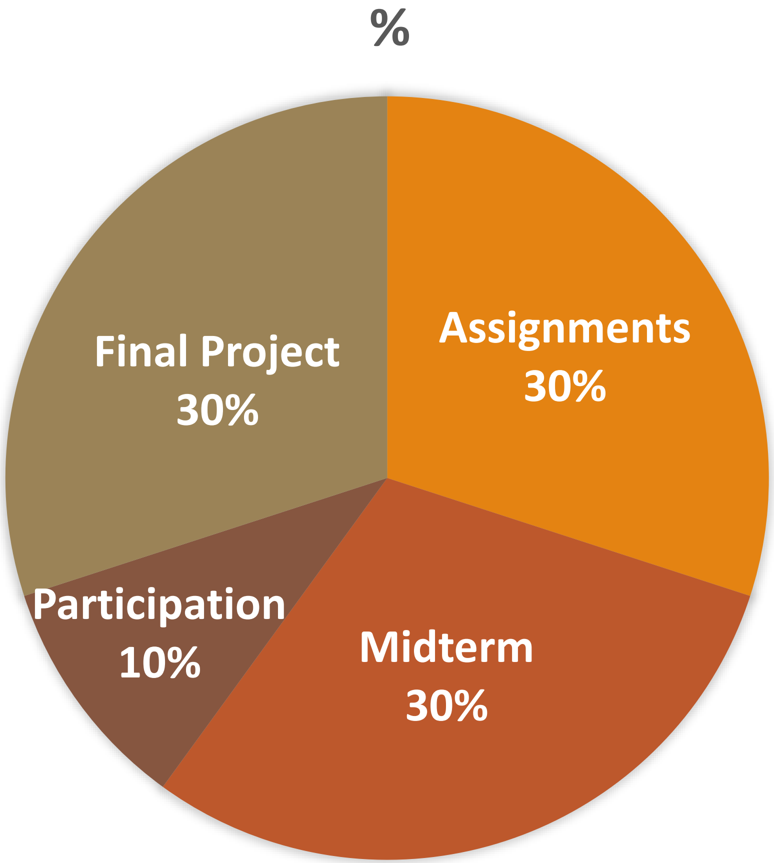
Discussion: piazza.com/gwu/fall2020/csci39076907nlp

Text: Mainly from the Jurafsky and Martin 2nd edition and 3rd edition(available online)

Coursework and Grading

- Course Material
  - Course slides and recorded class videos will be available on BB shortly after a class ends
- Assignments
  - Will be posted and announced on Blackboard
  - 4 Programming Assignments and Reading Assignments (Book Chapters and Research Papers)
  - The first assignment will be released next week
  - You have 7 free grace days for the whole semester
- Midterm
  - 10/13
- No Final Exam but a Final Project

Grading


%
Assignments
30%
Midterm
30%
Participation
10%
Final Project
30%

Final Project

- Groups
  - ✓Start forming groups of maximum 4
- Project Topic
  - ✓ Try to think of a novel problem
- Project proposal (20%)
- Class presentation (20%)
- Final report or short paper (30%)
- System Implementation (30%)

Outline

- Course Introduction
- Course Information
- **Deadlines**

Due Dates

9

- Assignment 1, 9/25
- Topic of interest, 9/21
- Project groups 9/29

10

- Assignment 2, 10/9
- Midterm, 10/13
- Project Proposal, 10/30

11

- Assignment 3, 11/13
- Assignment 4, 11/27

12

- Final Project Presentations and Submissions, 12/1 and 12/8

Questions??

Contributions to the course material & slides

Slides are sometimes adapted (with permission) from other great slide sets, namely from:

- Mona Diab, Chris Manning, Dan Jurafsky, Jason Eisner, Rada Mihalcea, Michael Collins, Alessandro Moschitti, Julia Hirschberg, Kathleen McKeown, Dragomir Radev.