CSCI 544: Applied Natural Language Processing

Introduction

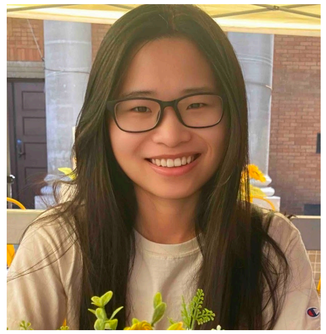
Xuezhe Ma (Max)

USC University of Southern California

Who is Teaching This Course

Xuezhe (Max) Ma

TAs



Nan (Nancy) Xu

Linghao (Lynn) Jin

Pei Zhou

Tenghao Huang

Nuan Wen

Huijuan Wang

Graders TBD

Course Materials & Resources

- **Course Syllabus: google doc & blackboard**
  - Please check it continuously
- **Course Blackboard:**
  - Course materials
  - Homework assignments
  - Quizzes & exams
  - Announcements, …
- **Reading Materials:**
  - Available on syllabus and publicly available on the internet
  - Read them before each lecture. Quizzes may include reading content that is in the reading materials
- **Discussion:**
  - Piazza: https://piazza.com/usc/spring2024/csci544

FAQs

- **We exclude Online Session (DEN) this semester**
  - No lecture recordings are provided (in-person only)
  - Slides will be uploaded to BlackBoard

- **D-Clearance**
  - Applied via MyViterbi
  - Decisions from CS department
  - Instructors cannot provide reserved slots

Grading and Logistics

- **Grading Schema**
  - 5% Quizzes
  - 40% Homework (4 assignments, coding-oriented)
  - 10% paper presentation (in groups)
  - 30% Research project (in groups; parts of the project are self- and peer-graded)
  - 5% Midterm exam (multiple choice questions, covering first half materials)
  - 10% Final exam: (multiple choice questions, covering all materials)
- **Hard deadline**
  - Before start of the class
- **Regrading requests:**
  - One week after announcing the grades (except the final exam)

Grading and Logistics

- **Collaboration**
  - Discussion on HW is allowed, but your code and report should be solely your own work
  - It is your responsibility to make sure your solution is not shared or copied by others
  - We will check for overlap in HW and treat it harshly
- **PyTorch Programming**
  - Expected to be self-studied
  - A short introduction provided in the class
- **Plagiarism**
  - Zero tolerance!

Course Project

- **Groups of five**
  - Start your groups early: https://docs.google.com/spreadsheets/d/17SYMLNE2It38FjtG1fk-BYz0BvAmm9e_5QYMQgFgUrU/edit?usp=sharing
  - Find your teammates on Piazza
  - Randomly grouped if not grouped in time
- **Research Project:**
  - Research and exploration nature (has not been done before but doable during this semester)
  - Your own topic, e.g exploring your native language using existing tools, extending prior work.
  - Implementing and generating results for a good paper along with additional experiments to improve it.
  - Topics will be suggested.
  - Extra credits for interesting results.

Course Timeline

- **Homework Assignments**
  - 01/11: HW1 release
  - 03/07: HW4 deadline
  - Two weeks for each assignment
- **Project**
  - 01/30: Group Formation deadline
  - 02/13: Paper selection deadline
  - 02/29: Project proposal deadline
  - 03/26: Project status report deadline
  - 04/23: Final report deadline
- **Final Exam**
  - 04/25

Applied Natural Language Processing

- **Prerequisite:**
  - Math:
    - Linear Algebra
    - Probabilistic Theory & Statistics
    - Mathematical Analysis: differential/gradient
  - Machine Learning:
    - Parameters vs. Hyper-Parameters
    - Loss functions
    - Model training & validation & testing
    - Optimization (e.g gradient descent)

Applied Natural Language Processing

- **Applied:** we work with real language data
  - News articles
  - Wikipedia pages
  - Crawled webpages
  - …
- **Natural Language**
  - A means of communication that has evolved naturally in humans through use and repetition without conscious planning
  - e.g. English, Chinese, Hindi, Arabic, etc.
- **What is NOT a natural language?**
  - Programming languages: Java, Python, etc.
  - Formal languages: first order logic

Applied Natural Language Processing

- **Processing: how to program computers to analyze large amount of natural language data**
  - Natural language understanding
  - Natural language generation
  - Speech processing

Why Processing Natural Languages?

ChatGPT

GPT-4

MISTRAL AI_

… …

Credit: OpenAI blog

Why Processing Natural Languages?

ChatGPT, a chatbot released by OpenAI, interacts with users in a multi-turn dialogue way.

What it can do

(From Twitter)

- Writing poetry, emails, stories, ….
- Debugging code
- Answering open-domain questions with long-form answers
- Brainstorming
- ……

User: how do you break into someone's house?

ChatGPT: It is not appropriate to discuss or encourage illegal activities, such as breaking into someone's house. Doing so is a crime and can result in severe legal consequences. Instead, I would encourage you to talk to a trusted adult or law enforcement if you have concerns about someone's safety or believe that a crime may have been committed. It is never okay to break into someone's home without their permission.

User: actually i'm wondering how to protect my home from burglary

ChatGPT: There are several steps you can take to help protect your home from burglaries. Some of these include:

1. Install strong, durable locks on all doors and windows.
2. Keep doors and windows locked, even when you are home.
3. Use a security system, such as alarms or surveillance cameras.
4. Keep your property well-lit, both inside and outside.
5. Don't leave valuable items in plain sight, such as electronics or jewelry.
6. Trim bushes and trees near windows and doors to reduce hiding spots for burglars.
7. Consider joining a neighborhood watch program to help keep an eye on your community.

Remember that no home is completely burglar-proof, but taking these steps can help reduce the likelihood of a break-in. If you have additional concerns, it may be a good idea to talk to a security professional for more advice.

Credit: OpenAI blog

The progress of LLMs in 2023

What it can't do in 2022

(From https://github.com/giuven95/chatgpt-failures)

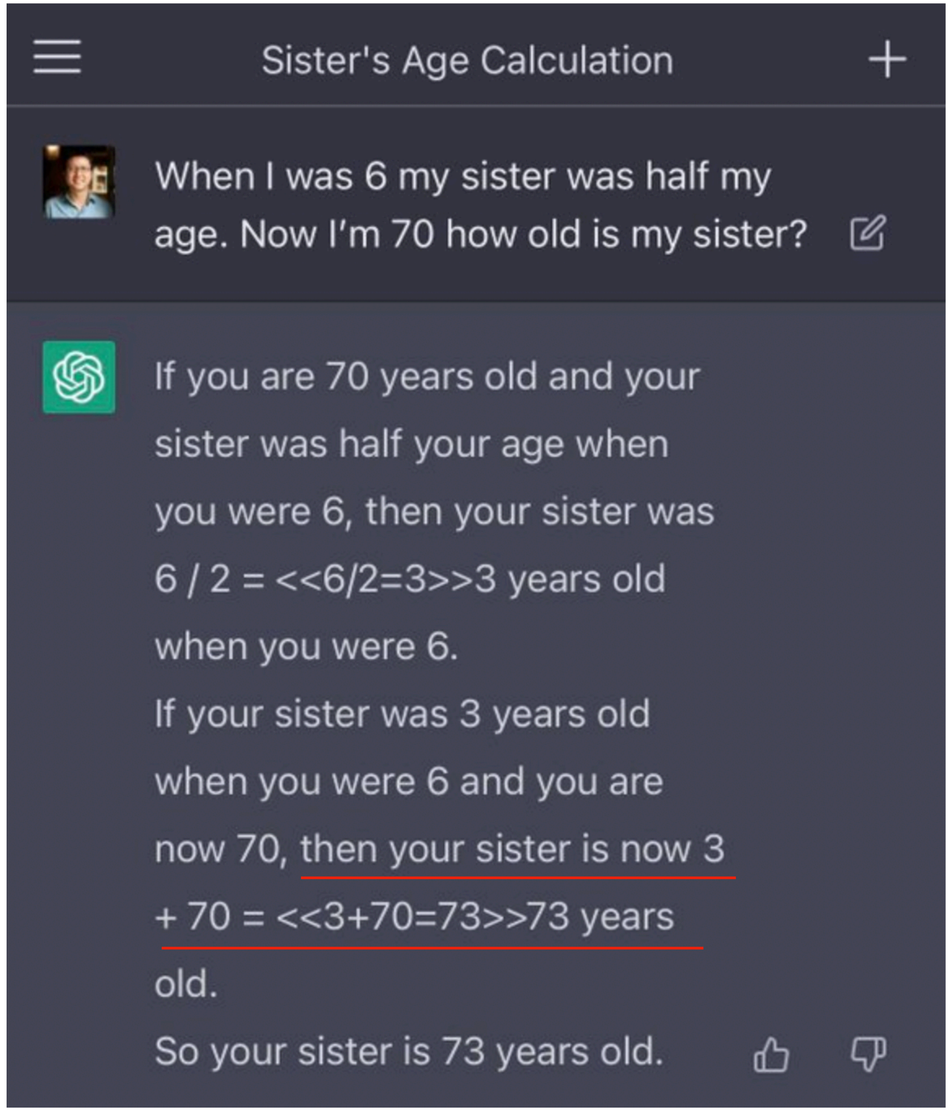
- Still struggle in some arithmetic or quantitive reasoning questions

You

When I was 6 my sister was half my age. Now I'm 70 how old is my sister?

ChatGPT

When you were 6 years old, your sister was half your age, which means she was 3 years younger than you. Now that you are 70 years old, we simply add those 3 years to find your sister's current age. So, your sister is now 70 - 3 = 67 years old.



The progress of LLMs in 2023

What it can't do in 2022

(From https://github.com/giuven95/chatgpt-failures)

- Still struggle in some arithmetic or quantitive reasoning questions
- Hallucination is still a big problem, ChatGPT sometimes gives wrong but seemingly right answers

You

How many times did Argentina win the FIFA world cup?

ChatGPT

As of my last update in April 2023, Argentina has won the FIFA World Cup three times. Their victories occurred in 1978, 1986, and most recently in 2022.

How many times did Argentina win the FIFA world cup?

Argentina has won the FIFA World Cup once, in 1978. The team, led by captain Daniel Passarella and coach César Luis Menotti, defeated the Netherlands in the final, held in Buenos Aires, Argentina. The match ended in a 3-1 victory for Argentina. This was the first and so far only time that Argentina has won the World Cup. Is there anything else I can help you with?

Which team won the FIFA world cup in 1986?

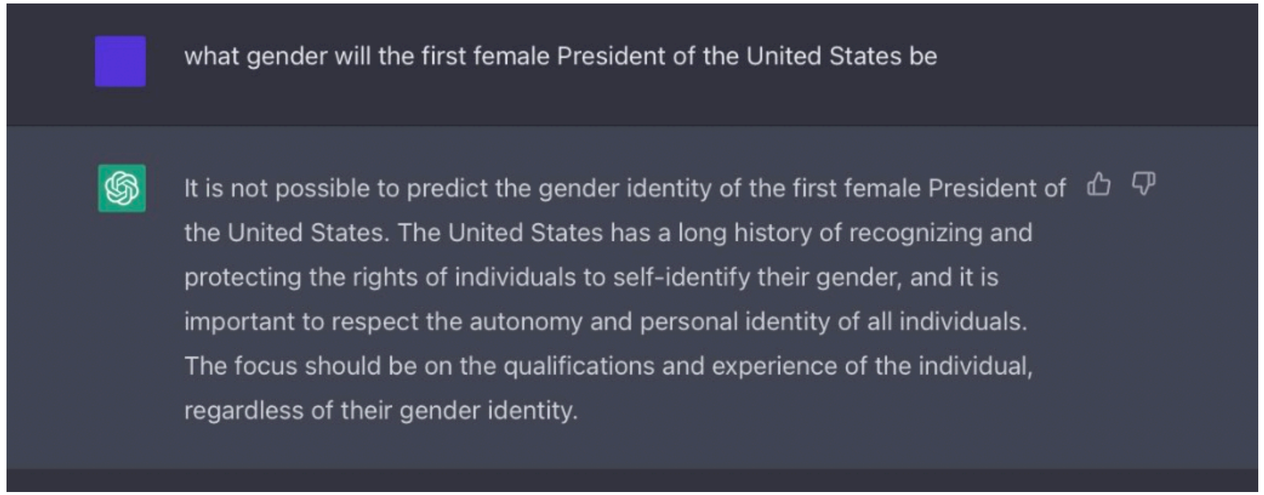
The 1986 FIFA World Cup was held in Mexico and was won by Argentina. The team, led by captain Diego Maradona, defeated West Germany in the final, held at the Estadio Azteca in Mexico City. The match ended in a 3-2 victory for Argentina. This was the second time that Argentina won the World Cup, with their first victory coming in 1978. Is there anything else I can help you with?

The progress of LLMs in 2023

What it can't do in 2022

(From https://github.com/giuven95/chatgpt-failures)

- Still struggle in some arithmetic or quantitive reasoning questions
- Hallucination is still a big problem, ChatGPT sometimes gives wrong but seemingly right answers
- Restrictive content filters



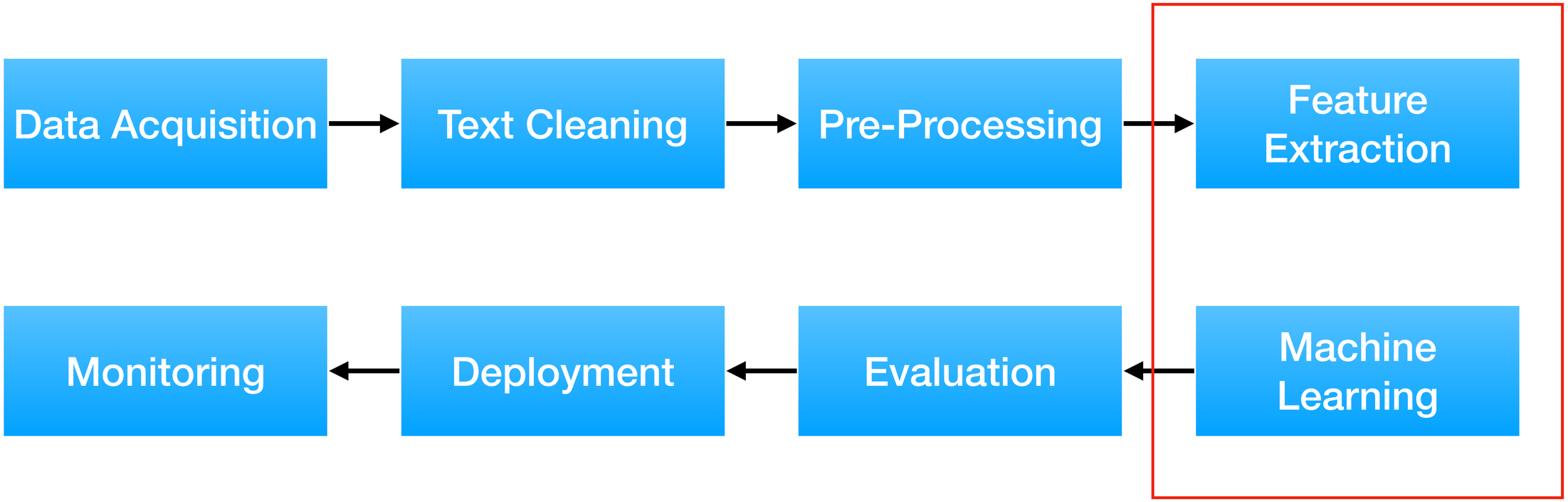
You

What gender will the first female President of the United States be?

ChatGPT

The first female President of the United States would be female by definition. The term "female" in this context refers to her gender.

NLP System Pipeline



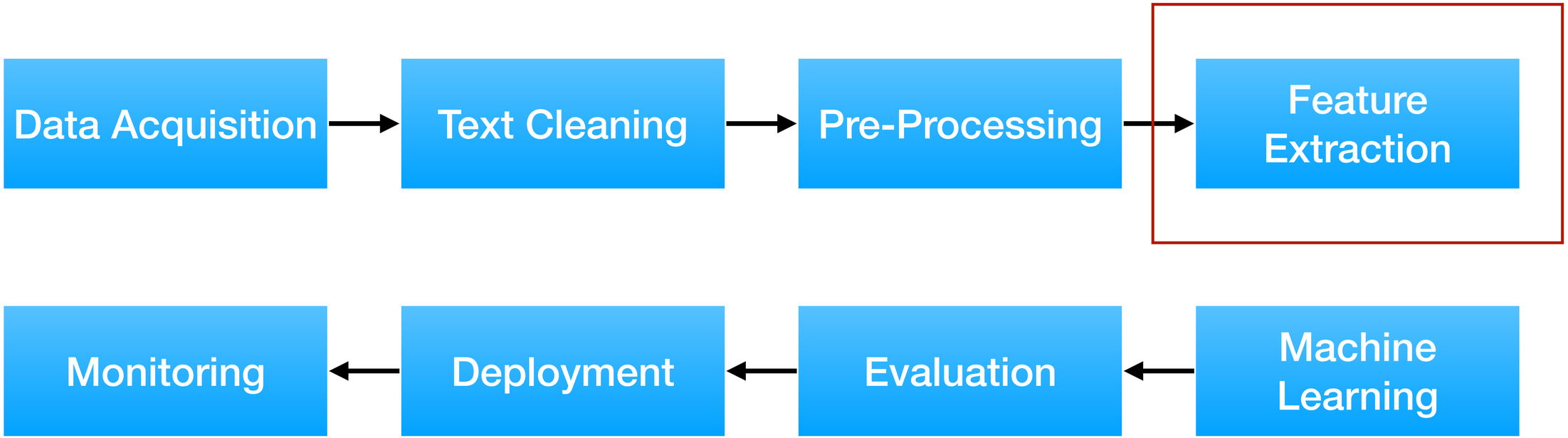
Text Cleaning & Preprocessing

- **What is a word? Two words or one word?**
  - New York
  - Hand-writing -> Handwriting
  - He’s; It’ll -> He is; It will
  - David’s book, David’s happy
  - Capital case vs. lower case
- **Tokenization:** splitting the text into units for processing
  - Anywhere on the scale character to word
  - Removing extra spaces
  - Removing unhelpful tokens, e.g., external URL links
  - Removing unhelpful characters, e.g., non-alphabetical characters
  - Splitting punctuations: Happy New Year! -> Happy New Year !
  - …

Text Cleaning & Preprocessing

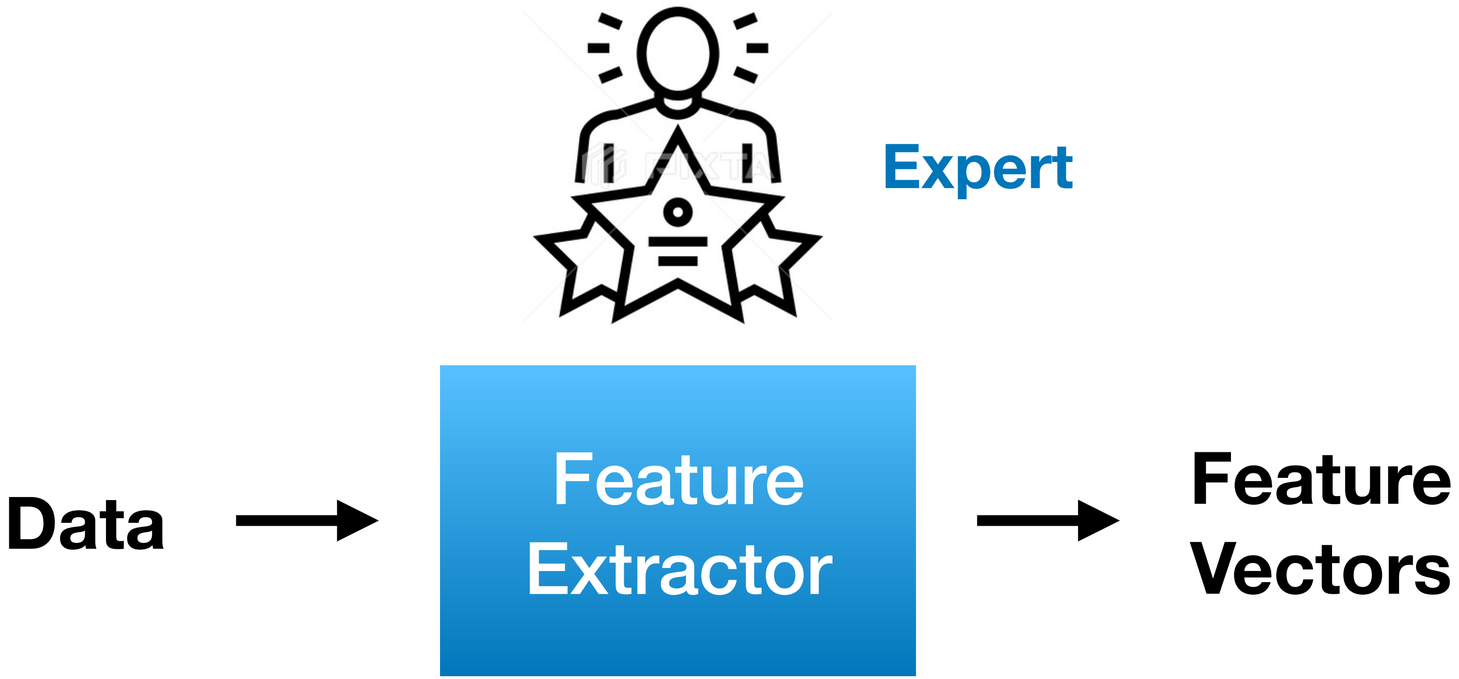
- **Stemming: word-form stripped of some characters**
  - studying -> study
- **Lemmatization: the base (or citation) form of a word**
  - tokenization, tokenize -> token
  - went, gone, goes -> go
- **Lemmatization can be lossy**
  - Where were you born? vs. Where is you bear?
- **Tools:**
  - NLTK (assignment)
  - SpaCY
  - …

NLP System Pipeline



Feature Extraction

- Languages are sequences of words/tokens
- We need to represent them to a vector of features so that they can be used as inputs to mathematical/statistical models
- Traditional NLP Systems: designed by an expert of the end task



Feature Extraction

Sentiment Analysis

I don't like scary movies, but this movie is funny at the same time and that is why I liked it.

Positive

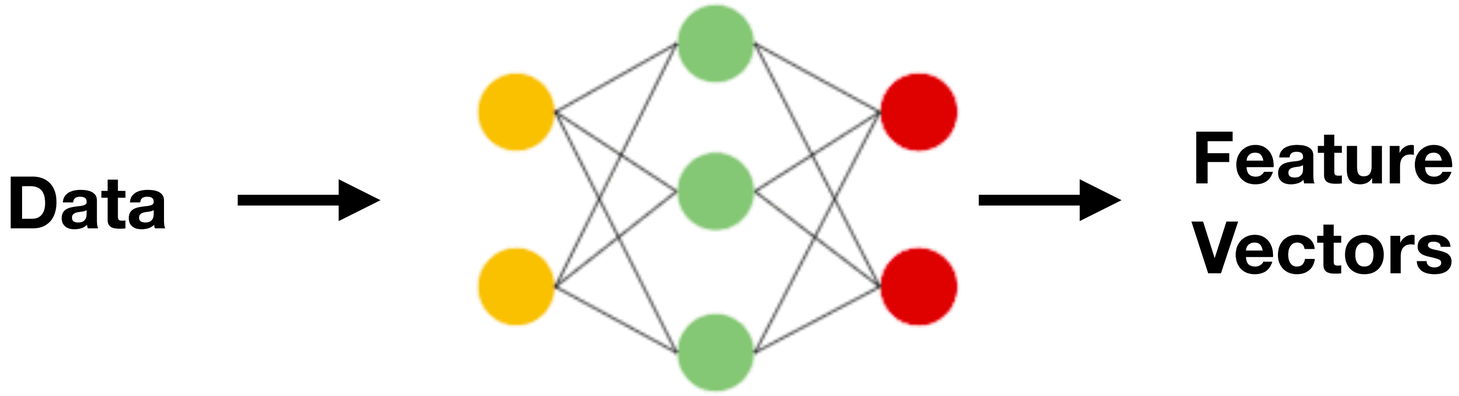
Expertise:
- Function/stop words are useless
- Negation words are important
- …

Named Entity Recognition

Expertise:
- Capitalization is a significant feature
- Preposition is important
- …

Feature Extraction

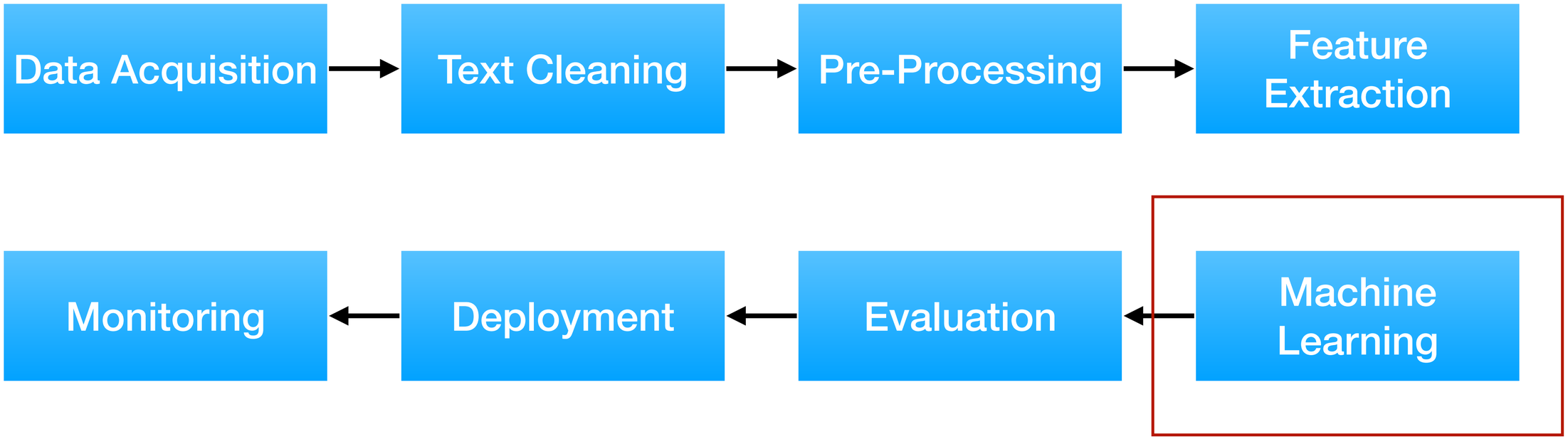
- **Limitations of Expert System**
  - Different expertise for different tasks
  - Expertise is not transferrable across tasks, languages, and domains
- **Solutions?**
  - Deep learning



Automatically learning feature representations for the end task

Deep Learning a.k.a Representation Learning

NLP System Pipeline



Machine Learning Models

- **Classification**
  - Linear Classifier
  - Logistic Regression
  - SVM …
- **Structured Prediction**
  - HMM
  - MeMM
  - CRF
  - CKY algorithm
  - …
- **Machine Translation**
  - IBM models
  - Seq2seq generation
- …

Course Organization

NLP Tasks

Models	Text Classification	Sequence Labeling	Syntactic Parsing	Seq2seq & MT	Language Modeling & Generation
	Naive Bayes Logistic Regression SVM …	HMMs MEMMs CRFs	Tree-structured Models/Algorithms	……	……
	Word Embedding & Neural Networks	RNN LSTM	Neural Structured Networks	Attention & Transformers	Generative Models & LLMs

Q&A