CS 1671 / CS 2071 / ISSP 2071
Human Language Technologies

Session 5: Machine learning intro, NLP tasks and applications

Michael Miller Yoder

January 28, 2026

University of Pittsburgh | School of Computing and Information

Course logistics: quiz

- First in-class quiz is next class, **Mon Feb 2**
  - Covers readings from all the sessions up to that point
  - Looking over the reading is a great way to prepare
  - Session 4: J+M 2-2.6, 2.8, 2.10
  - Can cover content assigned in reading that is not discussed in class
  - Content from other sessions will not be included
- 3-4 questions
- Conceptual, not programming
- Lowest quiz score in the course will be dropped
- Quizzes are 15% of your course grade total

Course logistics: quiz

- In class on Canvas, 10 minutes to complete it (1-1:10pm)
- Allowed resources
    - Textbook
    - Your notes (on a computer or physical)
    - Course slides and website
- Resources not allowed
    - Generative AI
    - Internet searches
- If you won't be in class, let me know and I can accommodate

Course logistics

- Homework 1 has been released. Is **due Feb 12 at 11:59pm**
- Homework assignments are programming-based
- Homework 1 covers text processing and regular expressions in Python
- Please remind me of your name before asking or answering a question

Course logistics

- A form to submit project ideas you may want to work on will be released this Fri Jan 29
  - Project idea submission form will be due next Thu Feb 5
- Take a look at the example projects on the project website. You can submit one or more of those for the form, or submit your own idea!
- Have a potential project idea that involves deriving insight from a dataset of text, or building an NLP system that can do something with text. You can submit it!
  - Ideas do not need to be well-formed
  - Ideas that have data already available are more realistic
- You will later choose from an anonymized list of project ideas on Project Match Day, Feb 11

Hacking4Humanity 2026: Challenging AI Injustice, Building Ethical Futures

- Tech and policy hackathon
- Feb 6-20
- Teams from SCI have won in the past and were invited to Harrisburg to present their projects to members of Governor Shapiro’s staff
- More information at https://www.duq.edu/research/centers-and-institutes/grefenstette-center/hacking4humanity.php

Overview: Machine learning intro, NLP tasks and applications

- Intro to machine learning
  - Definitions
  - Models and algorithms
  - Data: training, development, test
- NLP applications
- NLP “core tasks”
- Coding activity: clickbait classification

Review activity:
Define a term from last session about text preprocessing

Intro to (supervised) machine learning

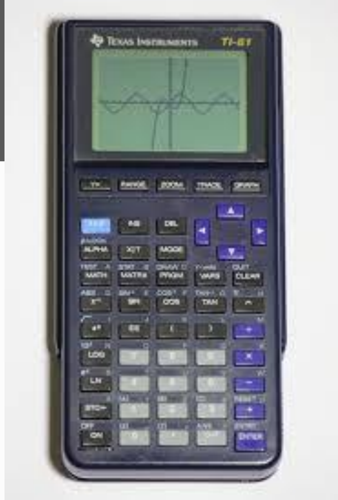
What is machine learning?

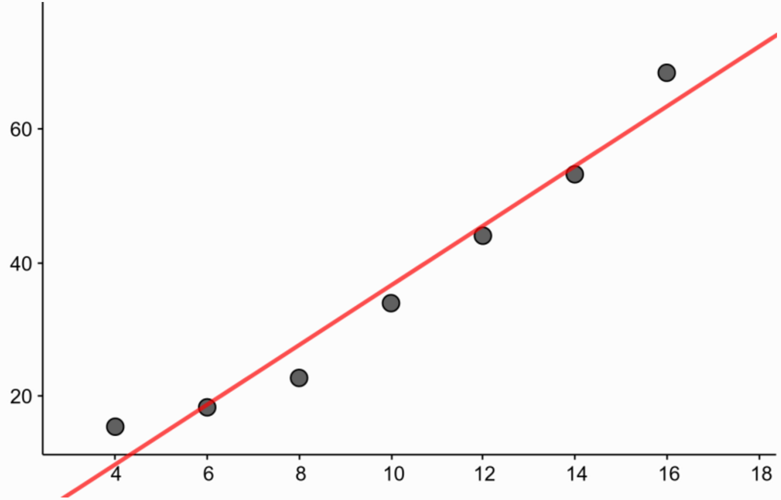
- A system that learns a function (maps from an input to an output) from examples/data
- Can predict things and perform tasks **without** explicit instructions
- Learns patterns from data with statistical algorithms
- Examples
  - Predict the weather tomorrow
  - Predict the best next driving action to take in an autonomous vehicle

What can you do with machine learning?

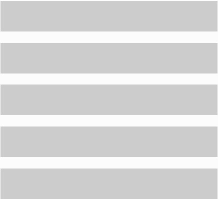
- Prediction: predict an output from an unseen input
  - That fits the pattern learned by looking at input it has seen before
- Interpretation
  - Examine the learned model weights to characterize the relationship between variables

y = 4*x* - 10

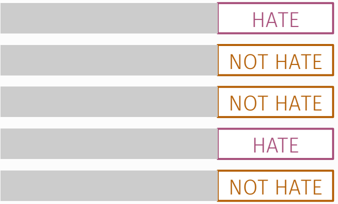




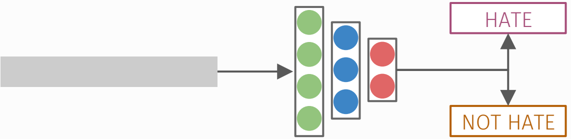
Supervised machine learning process



Data
(input text, *X*)



Annotate
labels (*Y*)



Train a model to
predict labels (*Y*)
from input text (*X*)

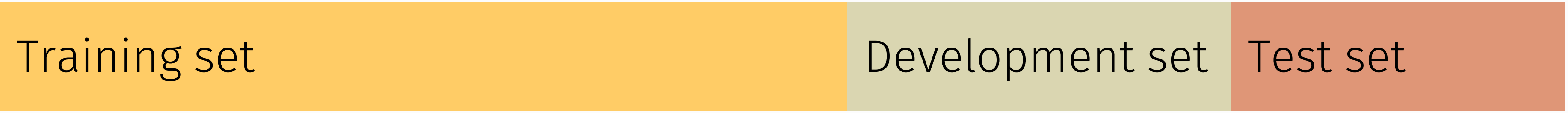
Machine learning models

- Transform an input to an output with a “model”: a simplified mathematical/statistical version of reality
- Models have parameters **learned from patterns in data**
  - Usually encode how variables relate to each other

Machine learning algorithms

- Algorithms are systematic ways of doing things
- In machine learning, “algorithms” refers to systematic ways of estimating model parameters from data.
- How does the model learn the patterns that enable it to make predictions? That’s the machine learning algorithm
- We’ll go over many in this class, including:
    - Logistic regression
    - Neural networks
    - Transformers

Training and test sets (and phases)



Training set

Development set

Test set

- Train parameters of the model on training set (training phase)
  - Sees examples of input and (assumed correct) output that it will mimic
- Development set to run tests of the model and choose hyperparameters
- Test time
  - Freeze parameters of the model
  - Predict input from an unseen set
  - Evaluate on correct answers and see how well the model performs
- **Don’t look at the test set too much when developing/choosing models**

NLP applications

Core tasks and applications of NLP

APPLICATIONS

machine translation

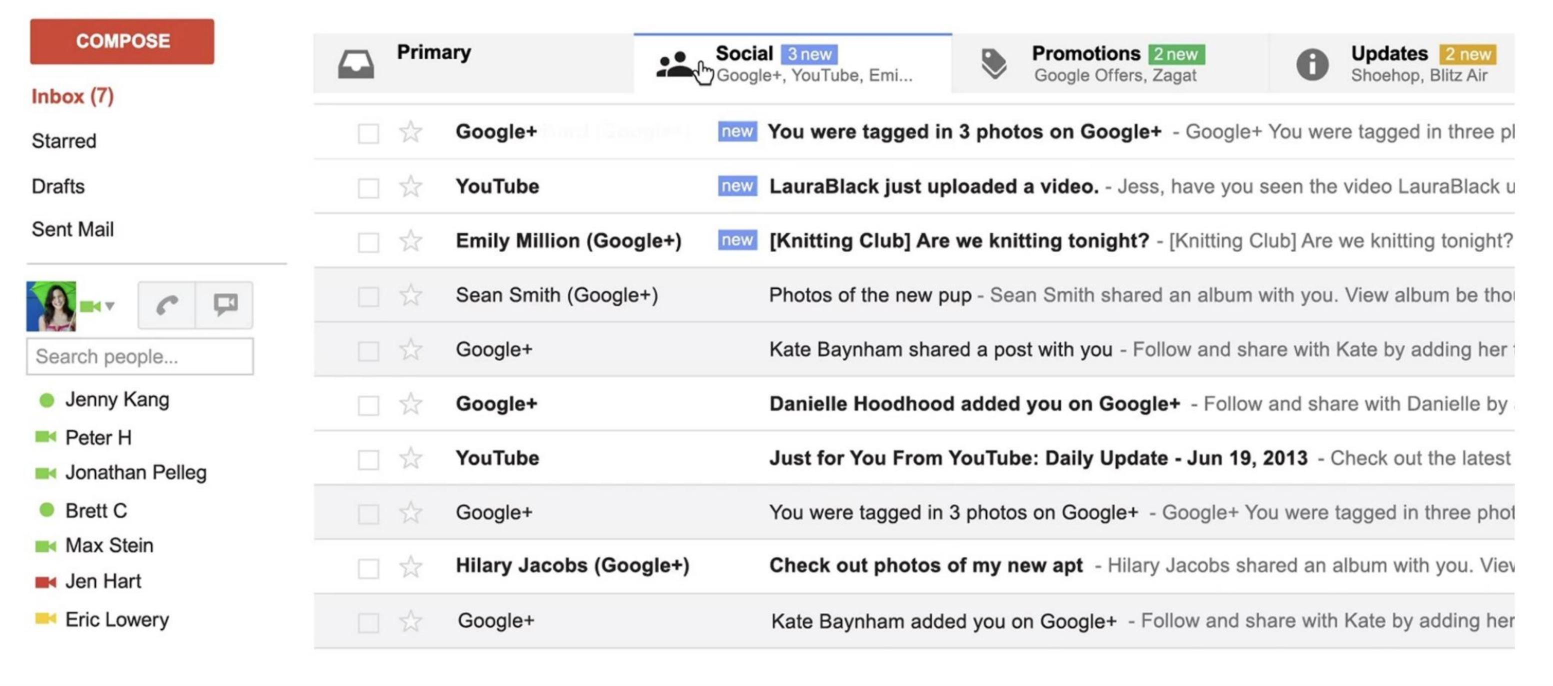
chatbots

information retrieval

summarization

question answering

NLP applications: email classification



Google+ new **You were tagged in 3 photos on Google+** - Google+ You were tagged in three pl

YouTube new **LauraBlack just uploaded a video.** - Jess, have you seen the video LauraBlack u

Emily Million (Google+) new **[Knitting Club] Are we knitting tonight?** - [Knitting Club] Are we knitting tonight?

Sean Smith (Google+) Photos of the new pup - Sean Smith shared an album with you. View album be tho

Google+ Kate Baynham shared a post with you - Follow and share with Kate by adding her

Google+ **Danielle Hoodhood added you on Google+** - Follow and share with Danielle by

YouTube **Just for You From YouTube: Daily Update - Jun 19, 2013** - Check out the latest

Google+ You were tagged in 3 photos on Google+ - Google+ You were tagged in three phot

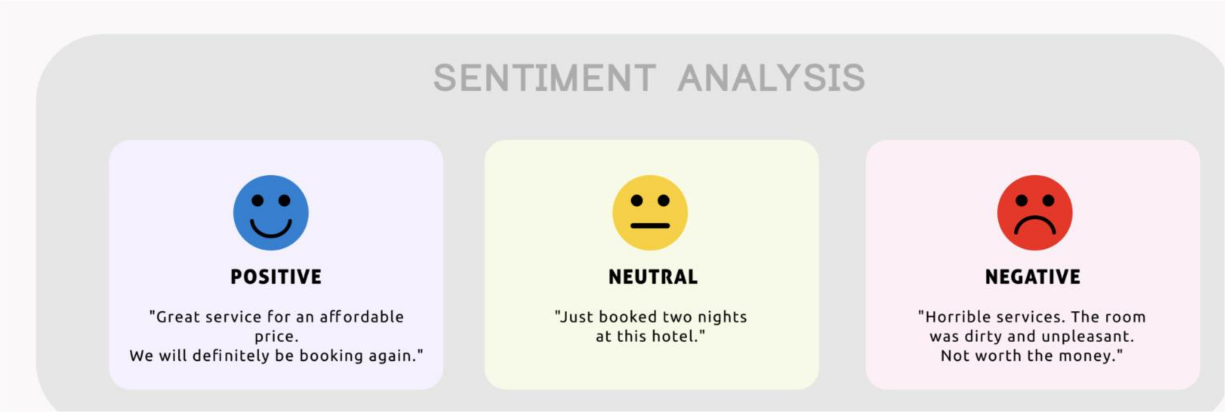
Hilary Jacobs (Google+) **Check out photos of my new apt** - Hilary Jacobs shared an album with you. View

Google+ Kate Baynham added you on Google+ - Follow and share with Kate by adding her

- Spam / Not spam
- Priority Level
- Category (primary / social / promotions / updates)

Slides Credit: Kevin Gimpel

NLP applications: sentiment analysis



SENTIMENT ANALYSIS

POSITIVE

"Great service for an affordable price.
We will definitely be booking again."

NEUTRAL

"Just booked two nights at this hotel."

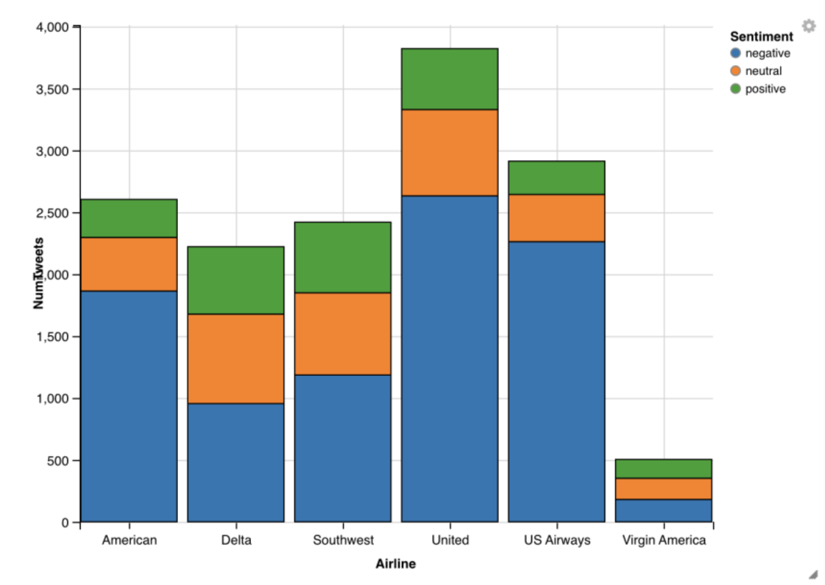
NEGATIVE

"Horrible services. The room was dirty and unpleasant.
Not worth the money."

Hotel review sentiment

Slides Credit: Kevin Gimpel

NLP applications: sentiment analysis

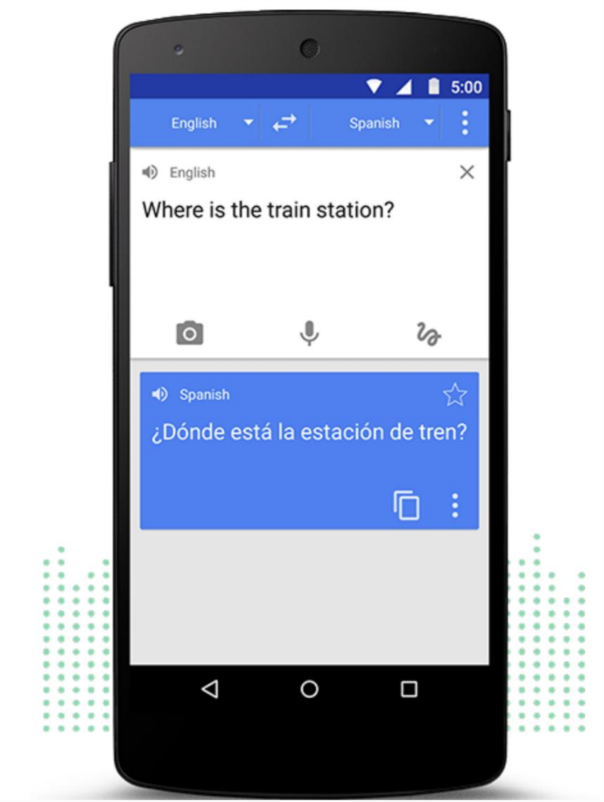


US Airline review sentiment

Slides Credit: Kevin Gimpel

https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

NLP applications: machine translation



English

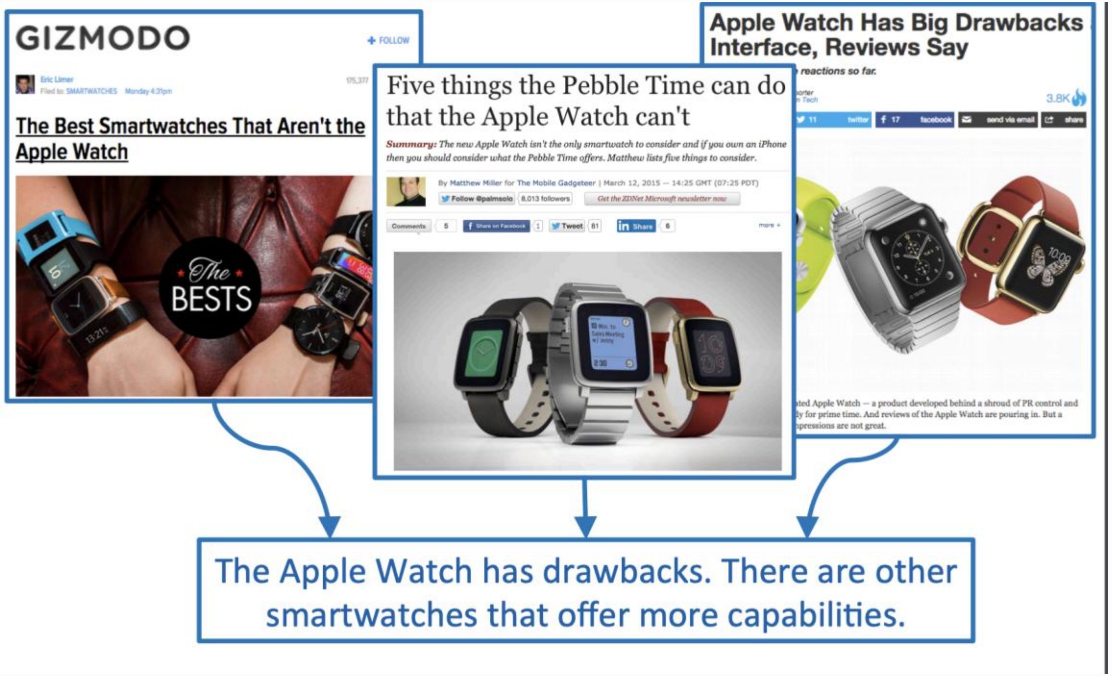
Where is the train station?

Spanish

¿Dónde está la estación de tren?

Slides Credit: Kevin Gimpel

NLP applications: summarization



GIZMODO

+ FOLLOW

Eric Limer

Filed to: SMARTWATCHES Monday 4:31pm

175,377

The Best Smartwatches That Aren't the Apple Watch

Five things the Pebble Time can do that the Apple Watch can't

Summary: The new Apple Watch isn't the only smartwatch to consider and if you own an iPhone then you should consider what the Pebble Time offers. Matthew lists five things to consider.

By Matthew Miller for The Mobile Gadgeteer | March 12, 2015 -- 14:25 GMT (07:25 PDT)

Follow @palmsolo 8,013 followers *Get the ZDNet Microsoft newsletter now*

Comments 5 Share on Facebook 1 Tweet 81 in Share 6 more +

Apple Watch Has Big Drawbacks Interface, Reviews Say

reactions so far.

porter
n Tech

3.8K

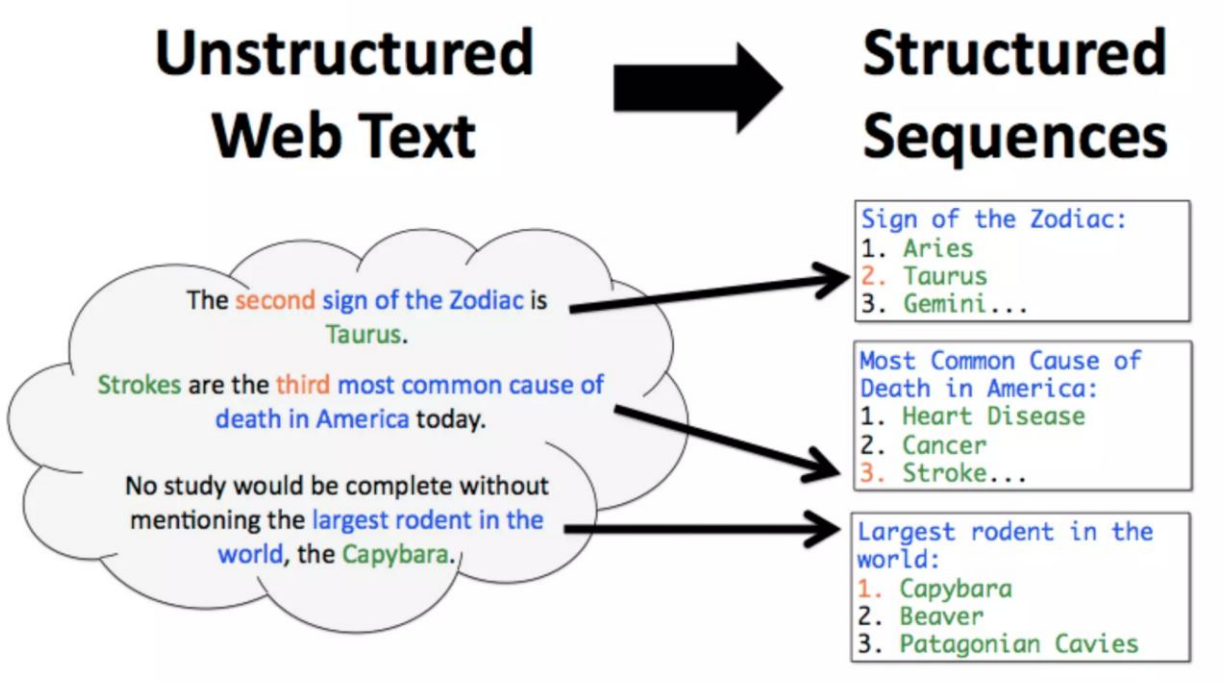
11 twitter 17 facebook send via email share

ated Apple Watch -- a product developed behind a shroud of PR control and
ly for prime time. And reviews of the Apple Watch are pouring in. But a
pressions are not great.

The Apple Watch has drawbacks. There are other smartwatches that offer more capabilities.

Slides Credit: Kevin Gimpel

NLP applications: information extraction



Unstructured Web Text

Structured Sequences

The second sign of the Zodiac is Taurus.

Strokes are the third most common cause of death in America today.

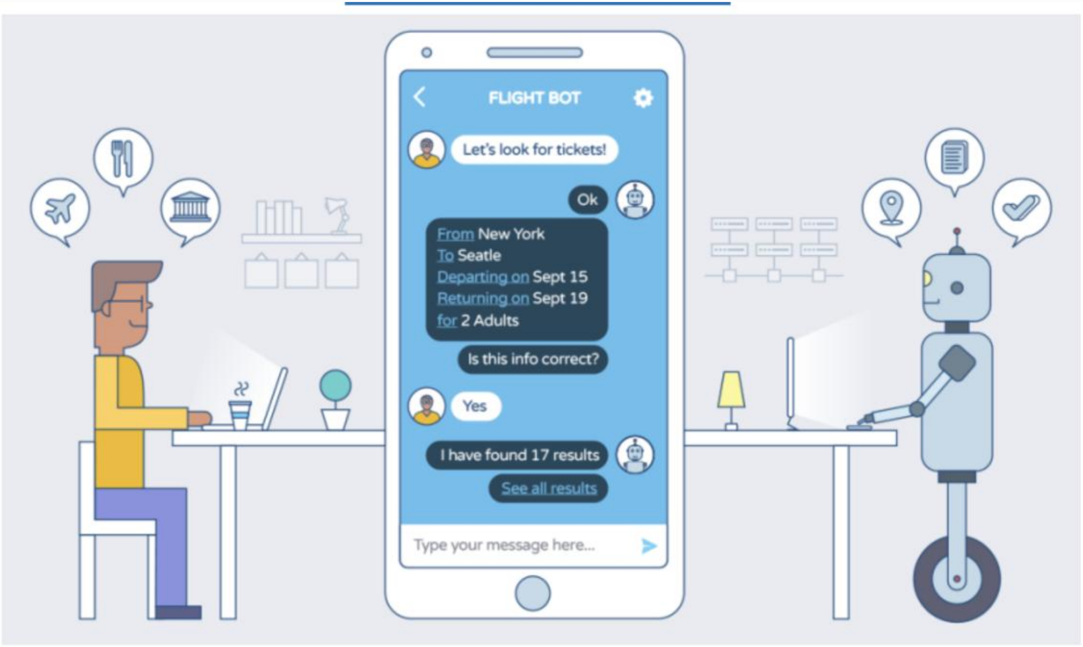
No study would be complete without mentioning the largest rodent in the world, the Capybara.

Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...

Most Common Cause of Death in America:
1. Heart Disease
2. Cancer
3. Stroke...

Largest rodent in the world:
1. Capybara
2. Beaver
3. Patagonian Cavies

NLP applications: dialogue systems/chatbots



Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

Is this info correct?

Yes

I have found 17 results

See all results

Slides Credit: Kevin Gimpel

NLP applications: question answering

“Alexa, who was President when Barack Obama was nine?”

“Alexa, how’s my commute?

“Alexa, what’s the weather?”

“Alexa, did the 49ers win?”

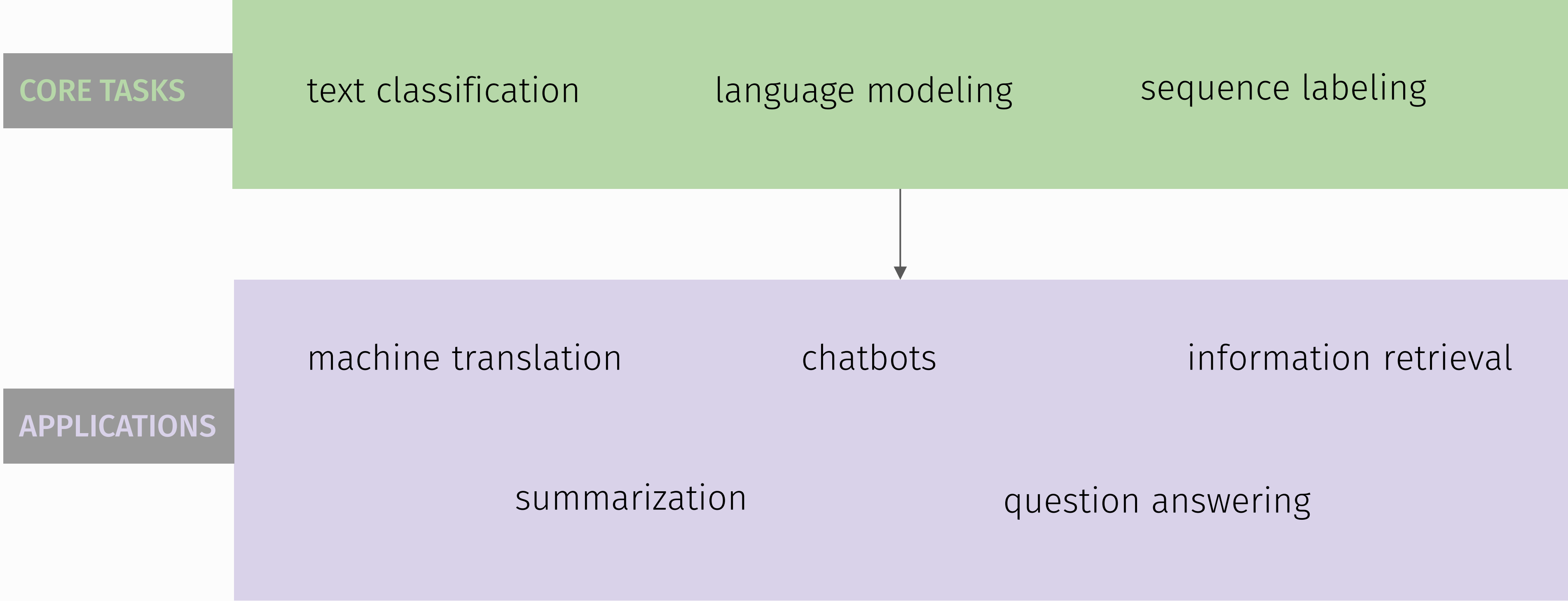
Slides Credit: Kevin Gimpel

Discuss with a neighbor:

What NLP applications, if any, do you use?

NLP core tasks

Core tasks and applications of NLP



CORE TASKS

text classification

language modeling

sequence labeling

APPLICATIONS

machine translation

chatbots

information retrieval

summarization

question answering

Text classification

- Input: a span of text
- Output: a label from a set of discrete options
- *Example:* sentiment analysis
    - *Text* -> {positive, neutral, negative}

Language modeling

- Input: a span of text, or no text at all
- Output: the next word
- *Example:* text generation for chatbots (ChatGPT)
  - *context text -> next word*

Sequence labeling

- Input: a span of text
- Output: a sequence of labels, one for each word (token)
- *Example:* part-of-speech tagging
    - *The book was brilliant -> DET NOUN VERB ADJ*

Coding activity: clickbait classification

Load in-class notebook

1. Go to this nbgitpuller link (also available on course website)
2. Log in with your Pitt username if necessary
3. Start a server with **TEACH – 6 CPUs, 48 GB**
4. Load custom environment at **/ix1/cs1671-2026s/class_env**
5. This should pull the cs1671_spring2026_jupyterhub folder into your JupyterLab
6. Open **session5_clickbait_classification.ipynb**