



Natural Language Processing

Info 159/259

Lecture 1: Introduction (Jan 19, 2021)

In-class questions:

<http://bit.ly/nlpqs>

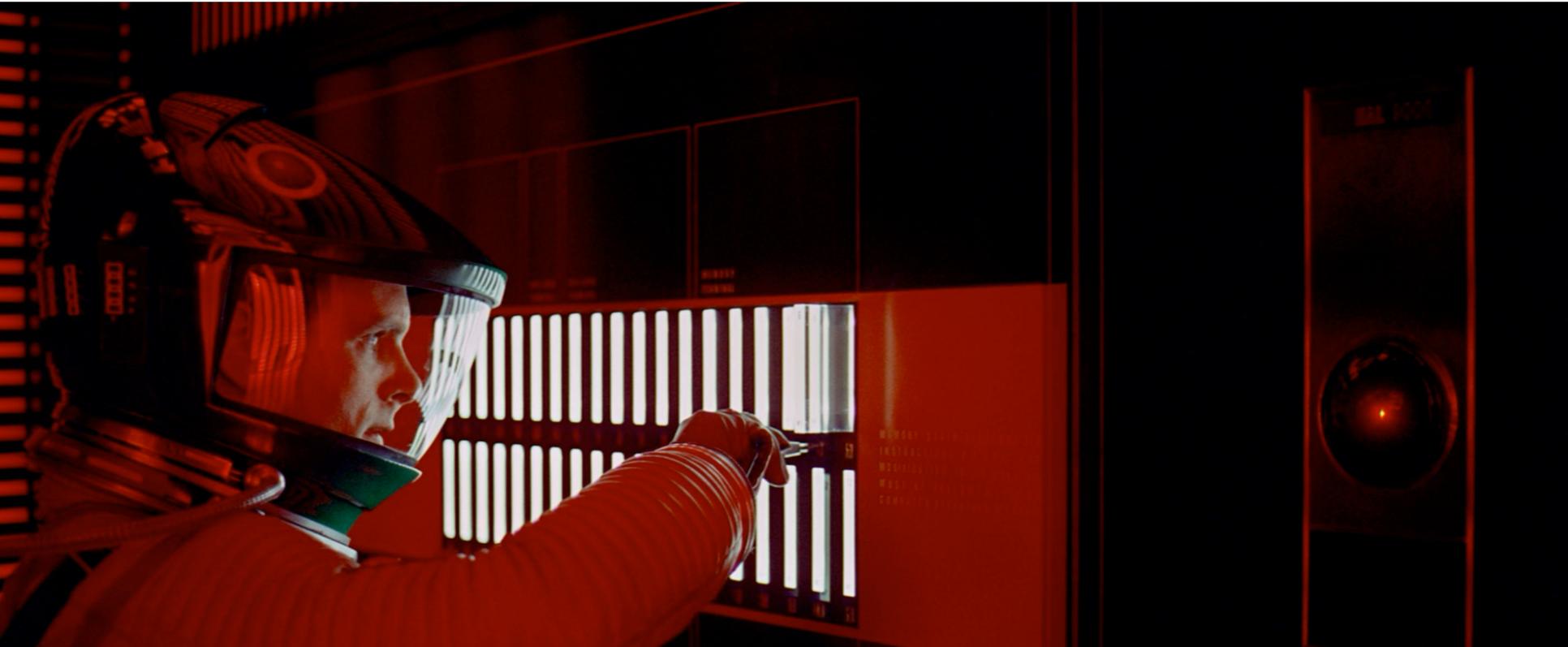
David Bamman, UC Berkeley

NLP is interdisciplinary

- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)

NLP = processing^{*} language with computers

processing as “understanding”



JOAQUIN PHOENIX AMY ADAMS ROONEY MARA

OLIVIA WILDE AND SCARLETT JOHANSSON

her

A SPIKE JONZE LOVE STORY

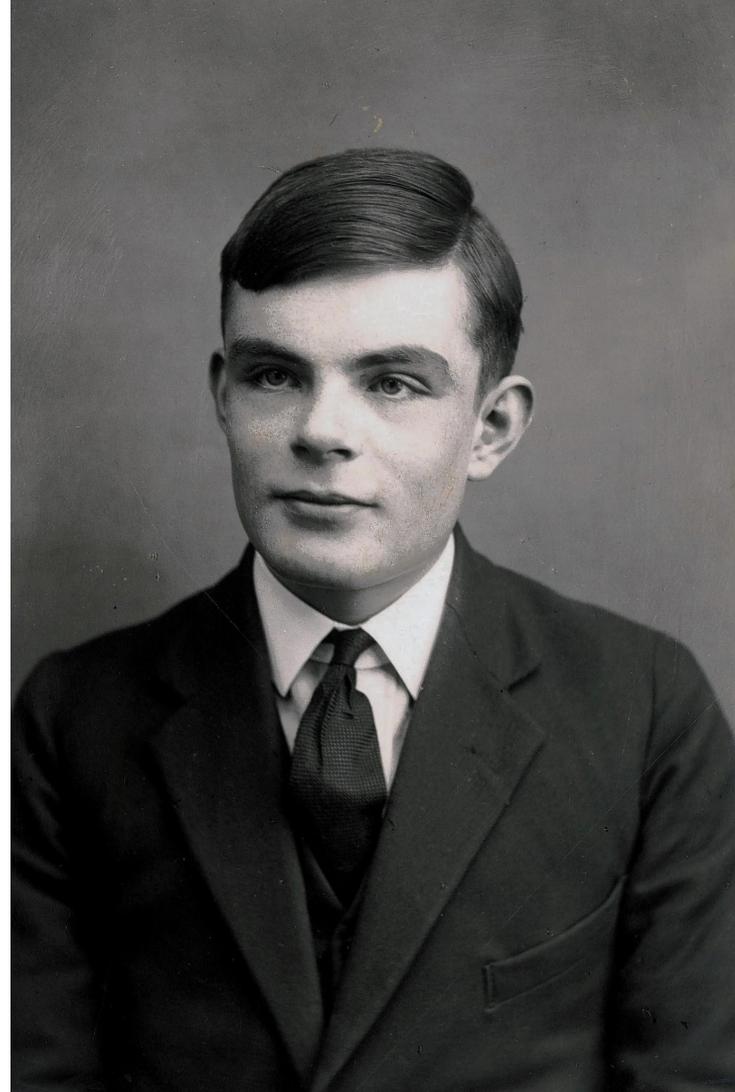
WARNER BROS. PICTURES PRESENTS
AN ANIMA PICTURES PRODUCTION "HER" JOAQUIN PHOENIX AMY ADAMS ROONEY MARA OLIVIA WILDE AND SCARLETT JOHANSSON
DIRECTED BY SPIKE JONZE
CASTING BY ELLEN LEVINS CASSANDRA KALLIKONDIS COSTUME DESIGNER JENNIFER FRENKEL EXECUTIVE PRODUCERS ARCADE FIRE PRODUCED BY CASEY STORINA PRODUCED BY ERIC ZWARGROWEN ALICE JEFF BUCHANAN PRODUCED BY KX BARRETT
PHOTOGRAPHER HOYTE VAN HOYTERA, C.S.C. MUSIC BY DANIEL LUPU NATALIE PATRICK CHARLIE HARRARD EDITOR MEGAN ELLISON SPIKE JONZE VINCENT LANDAY WRITTEN BY SPIKE JONZE
COMING SOON herthemovie.com

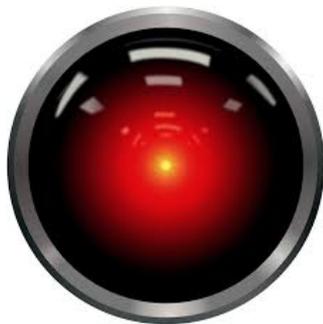


Turing test

Distinguishing human vs.
computer only through
written language

Turing 1950





Dave Bowman: Open the pod bay doors, HAL
HAL: I'm sorry Dave. I'm afraid I can't do that

Agent	Movie	Complex human emotion mediated through language
Hal	2001	Mission execution
Samantha	Her	Love
IG-11	The Mandalorian	"to nurse and protect"

Where we are now

Verizon LTE 4:28 PM 84%

"Open the pod bay doors HAL"
tap to edit

Wait, I think I know that one...

2001: A Space Odyssey

MGM (1968)

Director
Stanley Kubrick

Starring
Keir Dullea
Gary Lockwood
William Sylvester
Daniel Richter
Leonard Rossiter

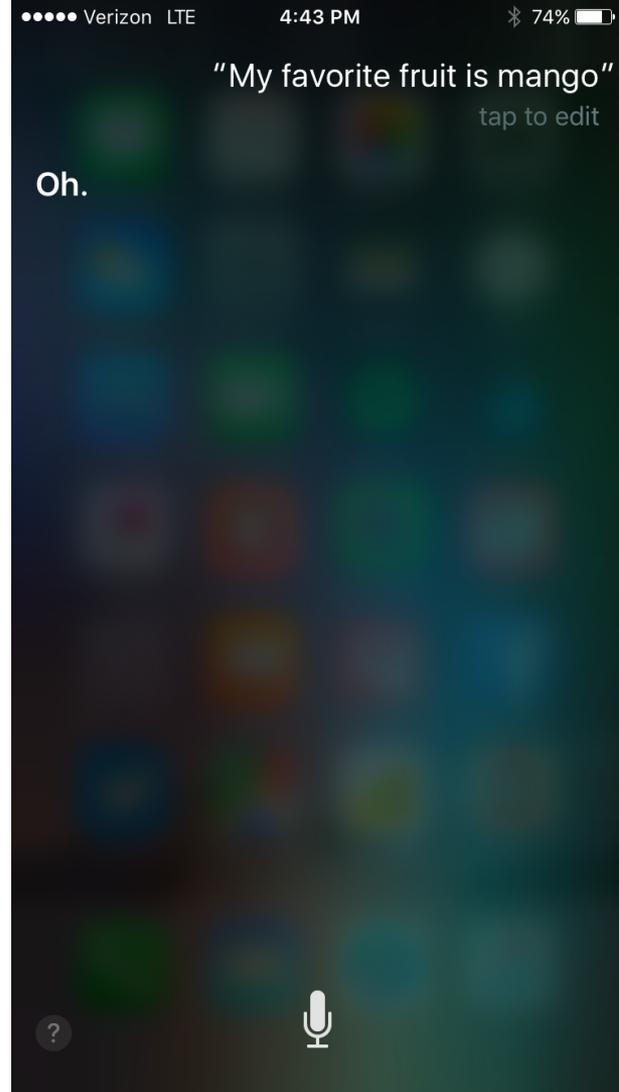
Runtime:
2h 19m

G

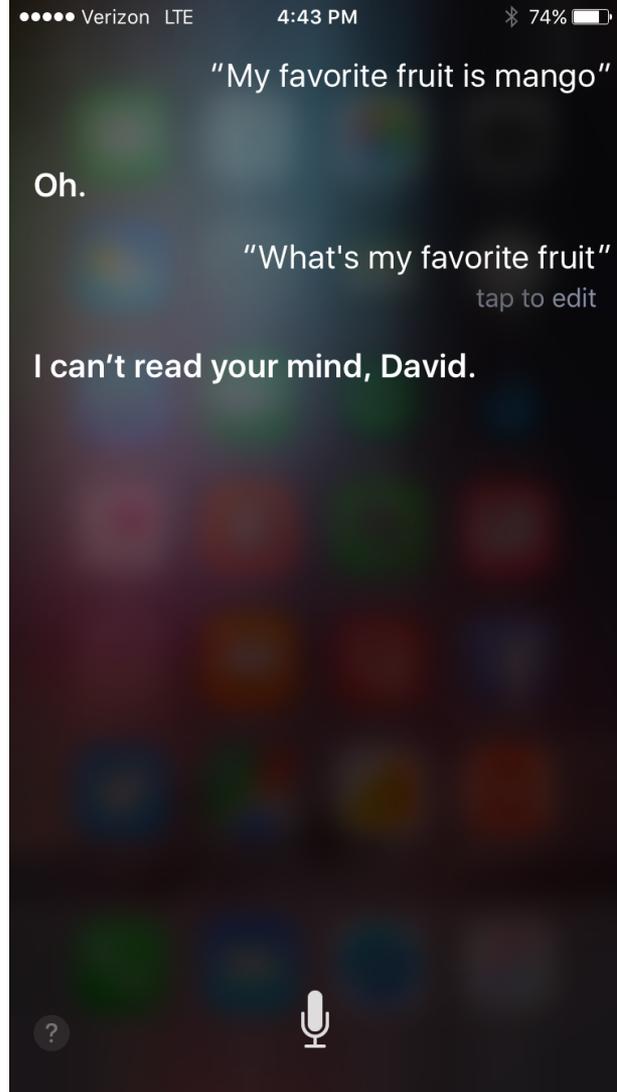


The movie poster for '2001: A Space Odyssey' features a close-up profile of a human face in a blue, ethereal light, set against a dark background with a starry space scene. The title '2001' is prominently displayed in a bold, white font, with 'A SPACE ODYSSEY' written in smaller text below it. The name 'STANLEY KUBRICK'S' is visible at the top of the poster.

Where we are now



Where we are now



Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many bonks are in a quoit?

A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

What makes language hard?

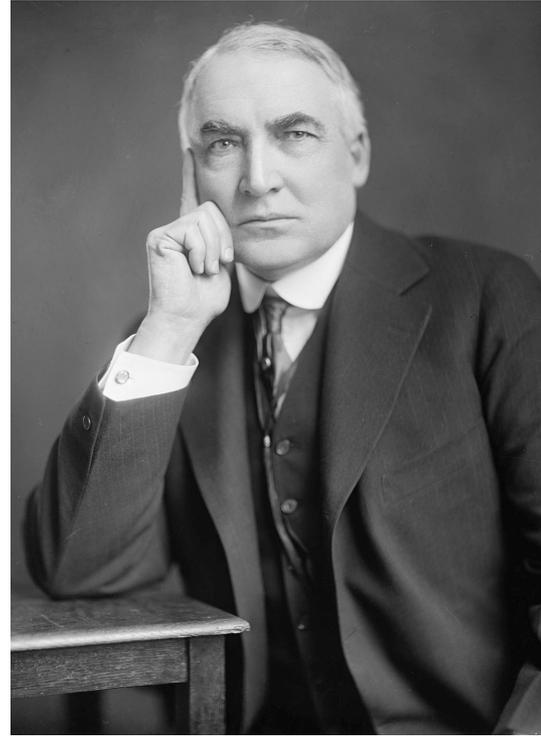
- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Modeling it is **AI-complete** (requires first solving general AI)

What makes language hard?

- Speech acts (“can you pass the salt?”)
[Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
[Grice 1975]
- Shared knowledge (“Warren ran for president”)



Elizabeth Warren
2020



Warren G. Harding
1920

What makes language hard?

- Speech acts (“can you pass the salt?”)
[Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
[Grice 1975]
- Shared knowledge (“Warren ran for president”)
- Variation/Indexicality (“This homework is wicked hard”)
[Labov 1966, Eckert 2008]

Ambiguity

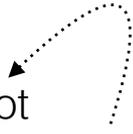
“One morning I shot
an elephant in my pajamas”



Animal Crackers

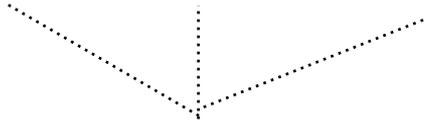
Ambiguity

“One morning I shot
an elephant in my pajamas”



Animal Crackers

Ambiguity



“One morning I shot
an elephant in my pajamas”



Ambiguity

verb noun



“One morning I shot
an elephant in my pajamas”



Animal Crackers

I made her duck

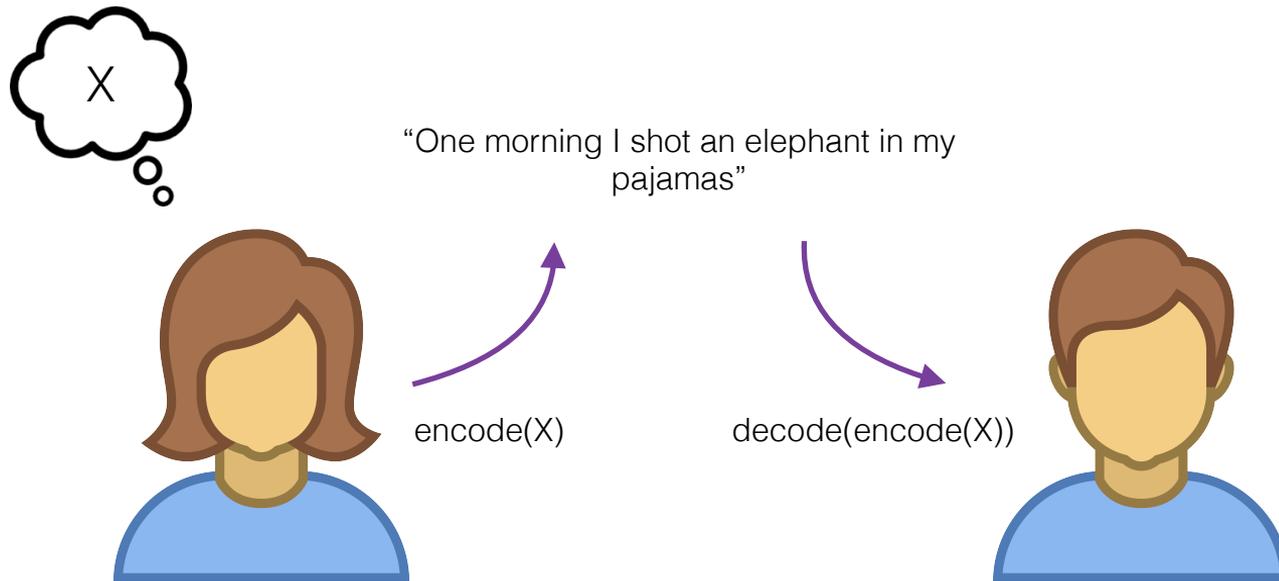
[SLP2 ch. 1]

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- ...

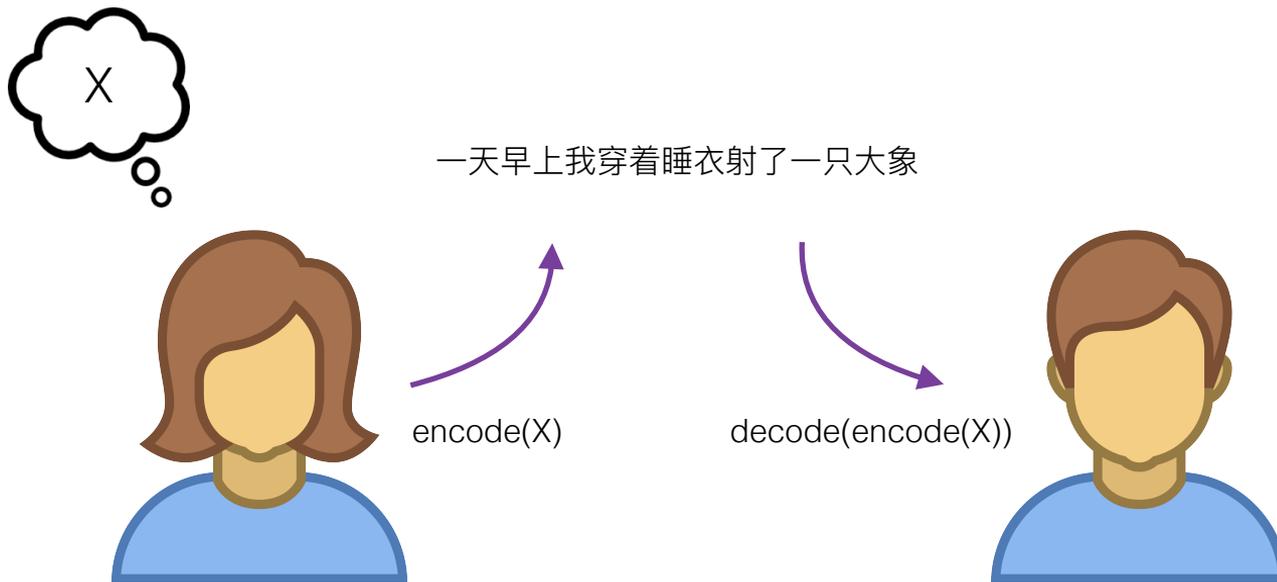
processing as representation

- NLP generally involves **representing language** for some end, e.g.:
 - dialogue
 - translation
 - speech recognition
 - text analysis

Information theoretic view



Information theoretic view



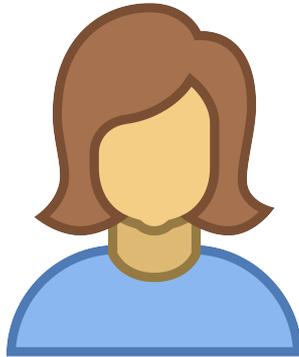
When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Weaver 1955

Rational speech act view



“One morning I shot an elephant in my pajamas”



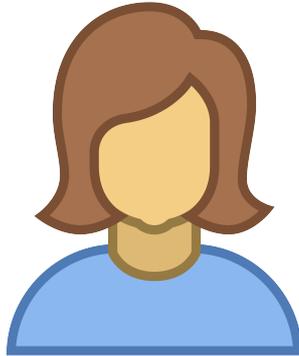
Communication involves **recursive reasoning**: how can X choose words to maximize understanding by Y?



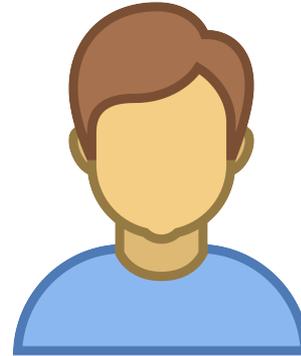
Pragmatic view



“One morning I shot an elephant in my pajamas”



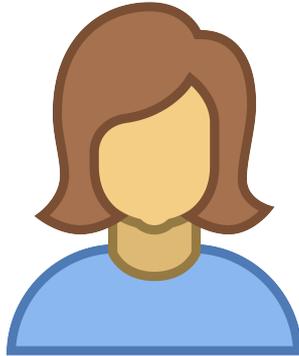
Meaning is co-constructed by the interlocutors and the **context** of the utterance



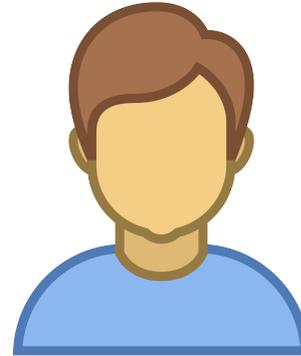
Whorfian view



“One morning I shot an elephant in my pajamas”



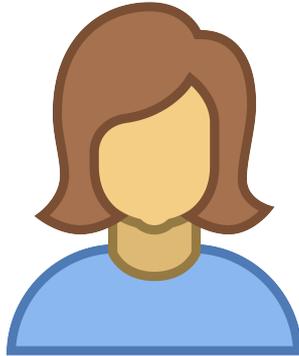
Weak relativism: structure of language influences thought



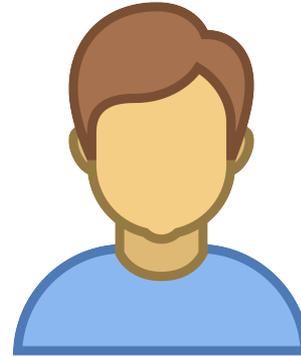
Whorfian view



一天早上我穿着睡衣射了一
只大象

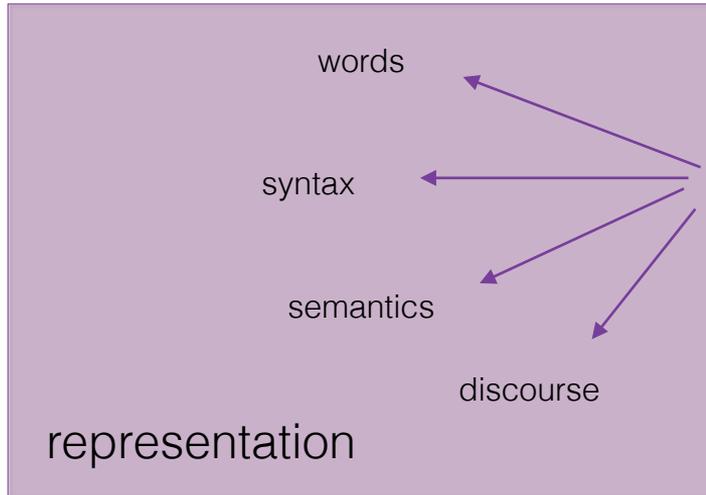


Weak relativism: structure of
language influences thought

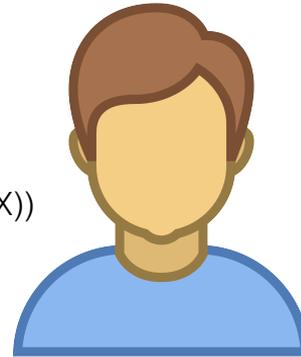


Decoding

“One morning I shot an elephant in my pajamas”



$\text{decode}(\text{encode}(X))$



A pyramid diagram with five levels, each represented by a purple rectangular block. The blocks are stacked from top to bottom, with the top block being the smallest and the bottom block being the largest. The text on each block is centered and white.

discourse

semantics

syntax

morphology

words

Words

- One morning I shot an elephant in my pajamas
- I didn't shoot an elephant
- **Imma** let you finish but Beyonce had one of the best videos of all time
- 一天早上我穿着睡衣射了一只大象

Parts of speech

noun

verb

noun

noun

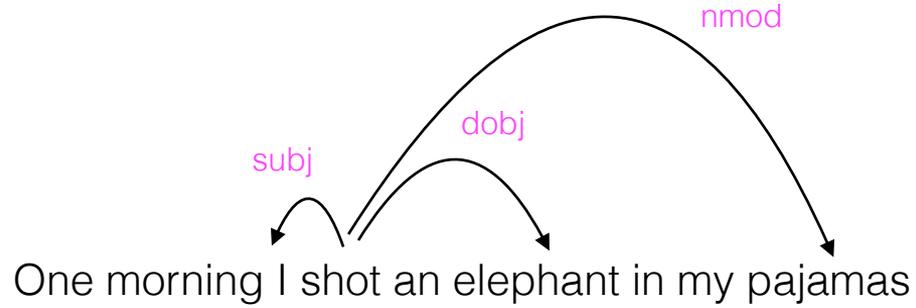
One morning I shot an elephant in my pajamas

Named entities

person

Imma let you finish but Beyonce had one of the best videos of all time

Syntax



Sentiment analysis



"Unfortunately I already had this exact picture tattooed on my chest, but **this shirt** is very useful in colder weather."

[overlook1977]

Question answering

What did Barack Obama teach?

Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th [President of the United States](#) from January 20, 2009, to January 20, 2017. A member of the [Democratic Party](#), he was the first [African American](#) to serve as president. He was previously a [United States Senator](#) from [Illinois](#) and a member of the [Illinois State Senate](#).

Obama was born in 1961 in [Honolulu, Hawaii](#), two years after the territory was [admitted to the Union](#) as the [50th state](#). Raised largely in Hawaii, he also spent one year of his childhood in [Washington state](#) and four years in [Indonesia](#). After graduating from [Columbia University](#) in 1983, he worked as a [community organizer](#) in [Chicago](#). In 1988, he enrolled in [Harvard Law School](#), where he was the first black president of the [Harvard Law Review](#). After graduating, he became a [civil rights attorney](#) and a professor, teaching [constitutional law](#); at the [University of Chicago Law School](#) from 1992 to 2004.

Barack Obama



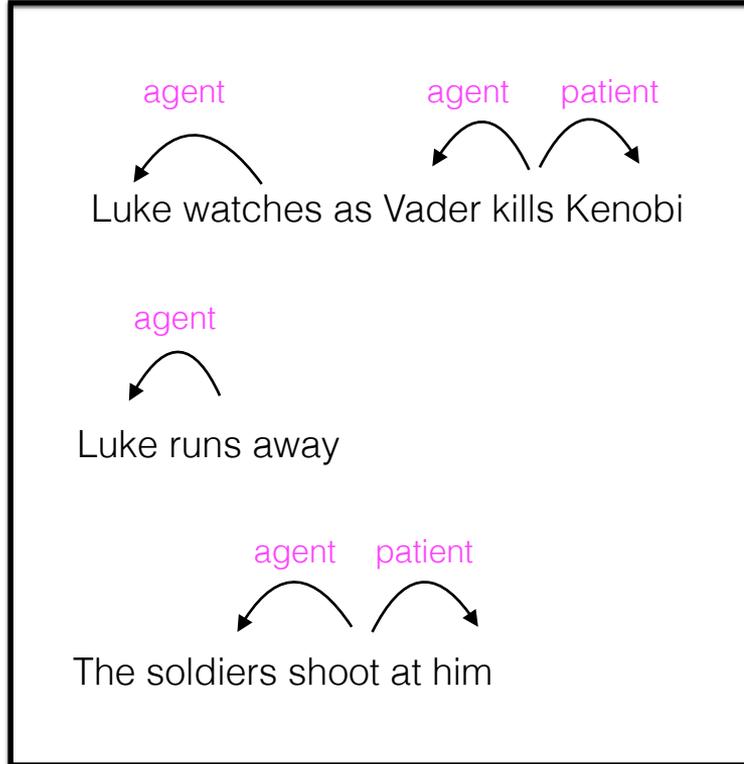
44th [President of the United States](#)

In office

Inferring Character Types

Input: text
describing plot of a
movie or book.

Structure: NER,
syntactic parsing +
coreference

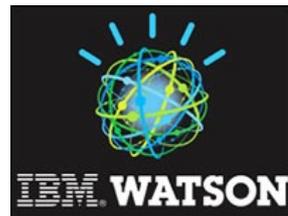


NLP

- Machine translation
- Question answering
- Information extraction
- Conversational agents
- Summarization



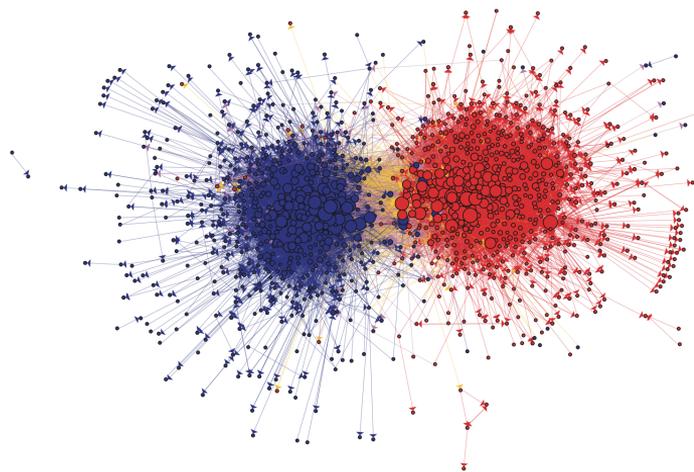
Google



NLP + X

Computational Social Science

- Inferring ideal points of politicians based on voting behavior, speeches
- Detecting the triggers of censorship in blogs/social media
- Inferring power differentials in language use



Link structure in political blogs
Adamic and Glance 2005

Computational Journalism

What do Journalists do with Documents? Field Notes for Natural Language Processing Researchers

Jonathan Stray
Columbia Journalism School
jms2361@columbia.edu

- Robust import
- Robust analysis
- Search, not exploration
- Quantitative summaries
- Interactive methods
- Clarity and Accuracy

Computational Humanities

Ted Underwood (2018), “Why Literary **Time** is Measured in Minutes”

Ryan Heuser, Franco Moretti, Erik Steiner (2016), The **Emotions** of London

Richard Jean So and **Hoyt Long** (2015), “Literary Pattern Recognition”

Ted Underwood, David Bamman and Sabrina Lee, The Transformation of **Gender** in English-Language Fiction (2018)

Franco Moretti (2005), Graphs, Maps, Trees

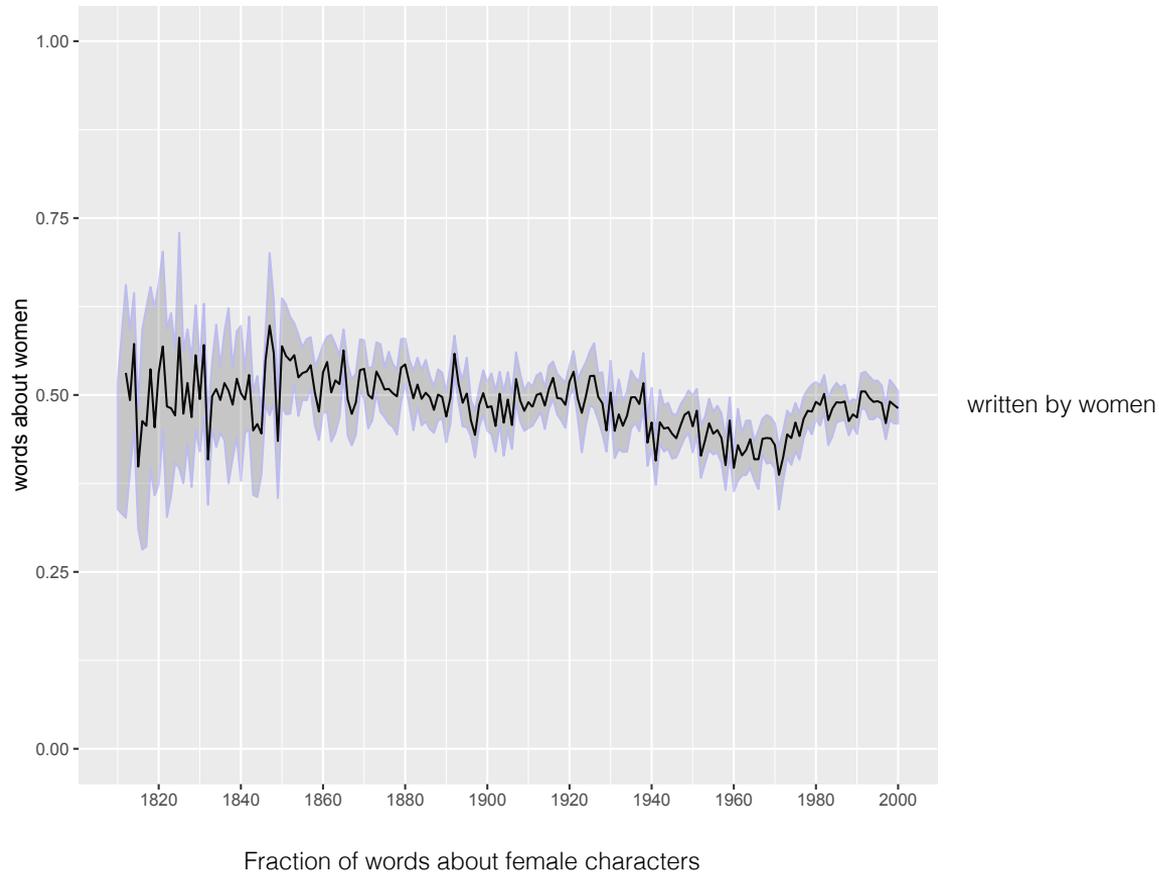
Holst Katsma (2014), **Loudness** in the Novel

So et al (2014), “**Cents** and Sensibility”

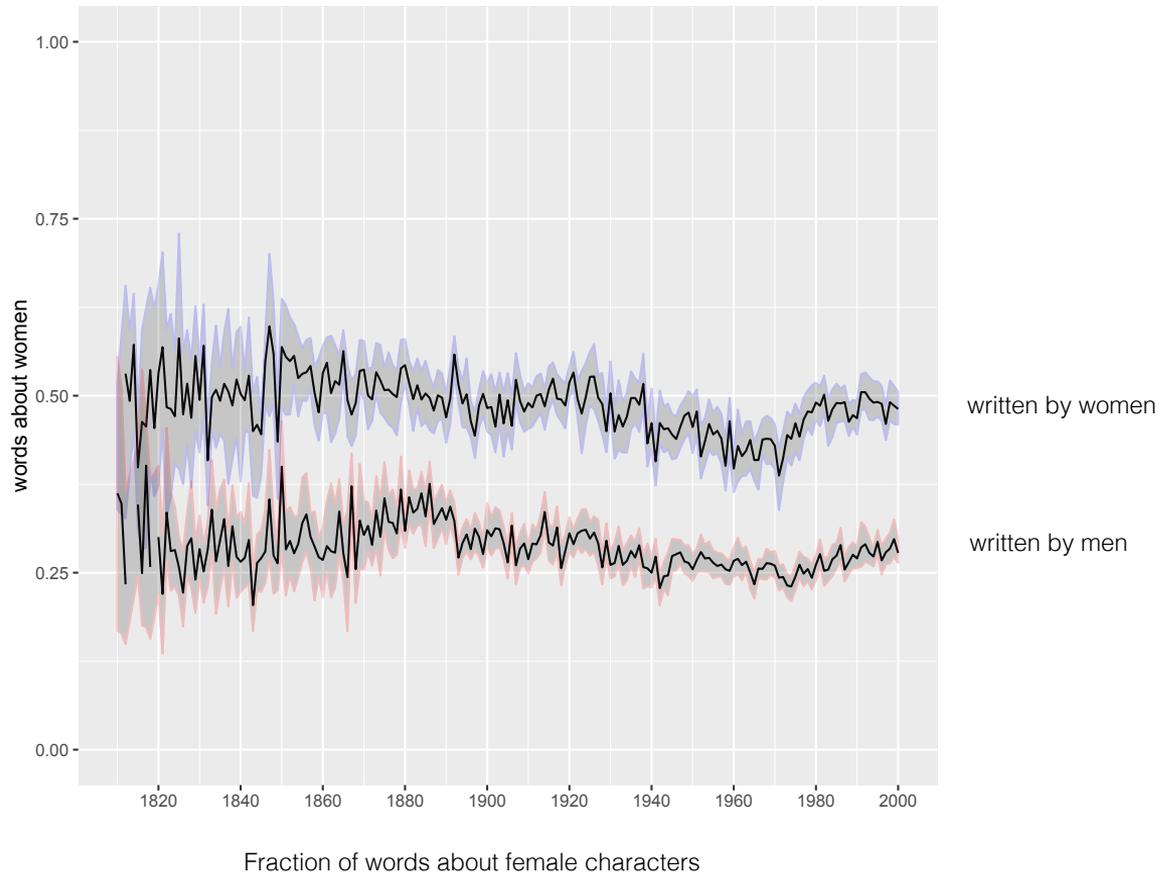
Matt Wilkens (2013), “The **Geographic** Imagination of Civil War Era American Fiction”

Jockers and Mimno (2013), “Significant **Themes** in 19th-Century Literature,”

Ted Underwood and Jordan Sellers (2012). “The Emergence of **Literary Diction**.” JDH

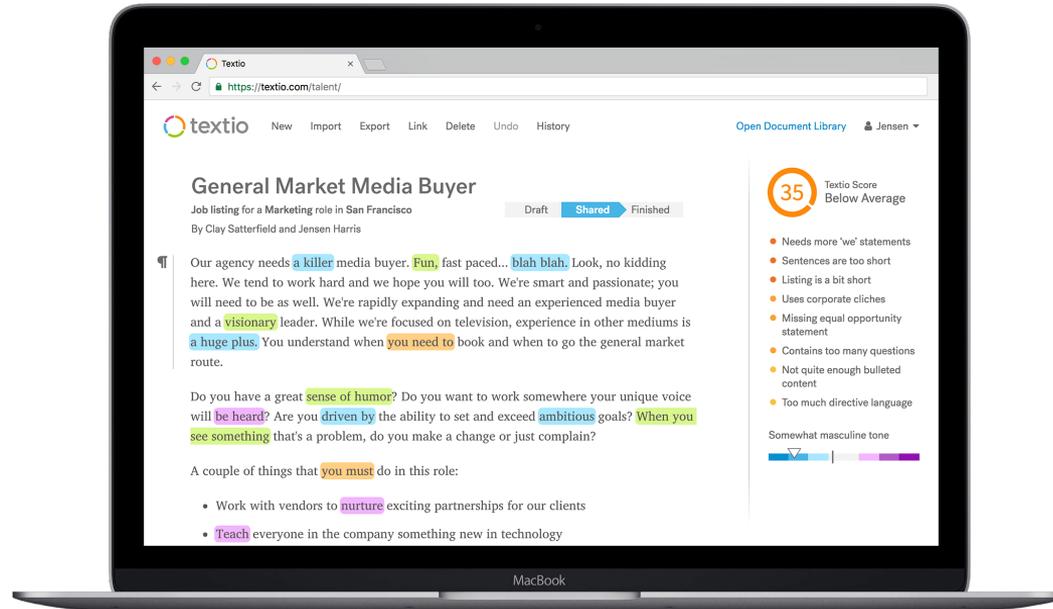


Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," *Cultural Analytics*



Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," *Cultural Analytics*

Text-driven forecasting



Methods

- Finite state automata/transducers (tokenization, morphological analysis)
- Rule-based systems

Methods

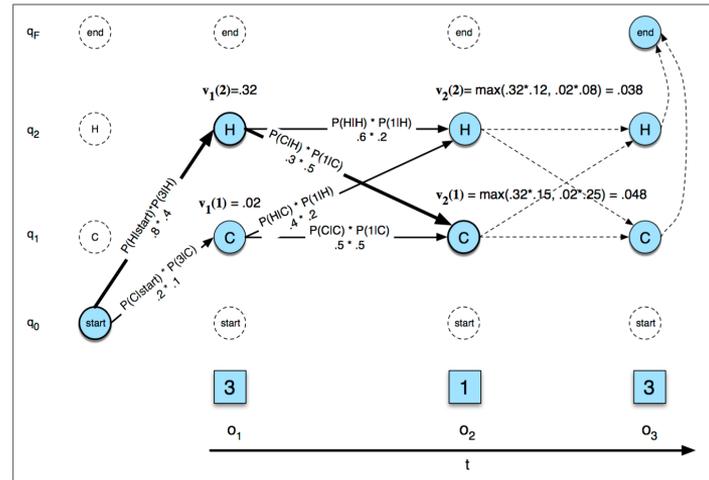
- Probabilistic models
- Naive Bayes, Logistic regression, HMM, MEMM, CRF, language models

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{\sum_y P(Y = y)P(X = x|Y = y)}$$

Methods

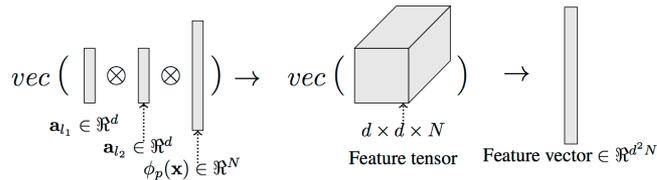
- Dynamic programming (combining solutions to subproblems)

Viterbi algorithm,
CKY



Methods

- Dense representations for features/labels (generally: inputs and outputs)



Srikumar and Manning (2014), "Learning Distributed Representations for Structured Output Prediction" (NIPS)

- Neural networks: multiple, highly parameterized layers of (usually non-linear) interactions mediating the input/output

Vaswani et al. (2017), "Attention is All You Need" (NeurIPS)

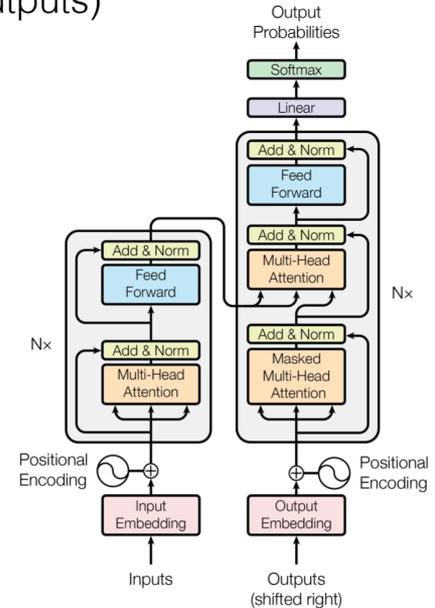


Figure 1: The Transformer - model architecture.

Methods

- Latent variable models (specifying probabilistic structure between variables and inferring likely latent values)

Nguyen et al. 2015, "Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress"

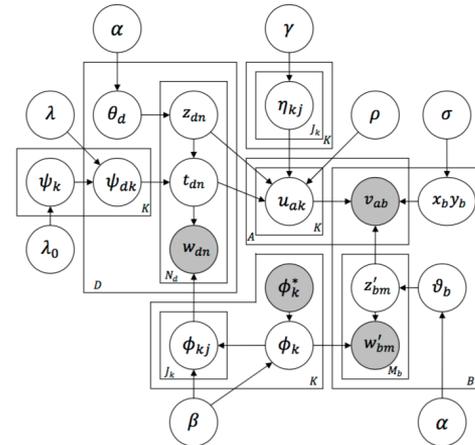


Figure 1: Plate notation diagram of HIPTM.

Info 159/259

- This is a class about **models**.
 - You'll learn and implement algorithms to solve NLP tasks efficiently and understand the fundamentals to innovate new methods.
- This is a class about the **linguistic representation** of text.
 - You'll annotate texts for a variety of representations so you'll understand the phenomena you'll be modeling

Prerequisites

- Strong programming skills
 - Translate pseudocode into code (Python)
 - Analysis of algorithms (big-O notation)
- Basic probability/statistics
- Calculus

function VITERBI(*observations* of len T , *state-graph* of len N) **returns** *best-path*

create a path probability matrix $viterbi[N+2, T]$

for each state s **from** 1 **to** N **do** ; initialization step

$viterbi[s, 1] \leftarrow a_{0,s} * b_s(o_1)$

$backpointer[s, 1] \leftarrow 0$

for each time step t **from** 2 **to** T **do** ; recursion step

for each state s **from** 1 **to** N **do**

$viterbi[s, t] \leftarrow \max_{s'=1}^N viterbi[s', t-1] * a_{s',s} * b_s(o_t)$

$backpointer[s, t] \leftarrow \operatorname{argmax}_{s'=1}^N viterbi[s', t-1] * a_{s',s}$

$viterbi[q_F, T] \leftarrow \max_{s=1}^N viterbi[s, T] * a_{s,q_F}$; termination step

$backpointer[q_F, T] \leftarrow \operatorname{argmax}_{s=1}^N viterbi[s, T] * a_{s,q_F}$; termination step

return the backtrace path by following backpointers to states back in time from $backpointer[q_F, T]$

$$\frac{dx^2}{dx} = 2x$$

Grading

- Info 159:
 - Homeworks (50%)
 - Weekly quizzes (10%)
 - Midterm (20%)
 - NLP subfield survey (20%)

NLP subfield survey

- 4-page survey for a specific NLP subfield of your choice (e.g., coreference resolution, question answering, interpretability, narrative generation, etc.), synthesizing at least 25 papers published at ACL, EMNLP, NAACL, EACL, AACL, *Transactions of the ACL* or *Computational Linguistics*.
- This survey should be able to provide a newcomer (such as yourself at the start of the semester) a sense of the current state of the art in that subfield in 2021, the major historical papers that have defined that area, and the different schools of thought within it.

Grading

- Info 259:
 - Homeworks (40%)
 - Weekly quizzes (10%)
 - Midterm (20%)
 - Project (30%)

259 Project

- Semester-long project (involving 1-3 students) involving natural language processing -- either focusing on core NLP methods or using NLP in support of an empirical research question
 - Project proposal/literature review
 - Midterm report
 - 6-page final report, *workshop quality*
 - Poster presentation

ACL 2021 workshops

- *SEM 2021: The 10th Joint Conference on Lexical and Computational Semantics
- 2nd International Workshop on Computational Approaches to Historical Language Change (LChange'21)
- Workshop on Natural Language Processing for Programming
- Third Workshop on Gender Bias for Natural Language Processing
- Workshop on Online Abuse and Harms
- 17th Workshop on Multiword Expressions (MWE 2021)
- 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)
- Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)

Exams

- We'll have one exam:
 - Midterm (3/11, tentatively in-class)
- We will **not** be offering alternative exam dates, so if you anticipate a conflict, don't take this class!

Late submissions

- All homeworks and quizzes are due on the date/time specified.
- You have 3 late days total over the semester to use when turning in homeworks/quizzes; each day extends the deadline by 24 hours. If all late days have been used up, homeworks/quizzes can be turned in up to 48 hours late for 50% credit; anything submitted after 48 hours late = 0 credit.
- Late days are assessed immediately once homeworks or quizzes are submitted late and can't be retroactively changed (if you submit 2 homeworks and 2 quizzes late, for example, you can't decide after the fact which ones to apply your 3 slip days to -- they apply to whichever homeworks or quizzes use them up first).

Academic integrity

- We'll follow the UC Berkeley code of conduct
<http://sa.berkeley.edu/code-of-conduct>
- You may discuss homeworks at a high level with your classmates (if you do, include their names on the submission), but each homework deliverable must be completed independently -- all writing and code must be your own; and all quizzes and exams must be completed independently.

Academic integrity

- If you mention the work of others, you must be clear in citing the appropriate source:
<http://gsi.berkeley.edu/gsi-guide-contents/academic-misconduct-intro/plagiarism/>
- This holds for source code as well: if you use others' code (e.g., from StackOverflow), you must cite its source.
- We have zero tolerance policy for cheating and plagiarism; violations will be referred to the Center for Student Conduct and will likely result in failing the class.

Curve

Grades in this class will **not** be curved.

Lectures

- Recordings of lectures will be available on bCourses.
- Attendance is not required for lectures, but **strongly recommended**.

Waitlisted

- We'll be adding students from the waitlist for 159 according to their order (respecting priorities for juniors/seniors, data science majors)

Piazza

- We'll use Piazza as a platform for asking and answering questions about the course material, including homeworks.
- Students are encouraged to actively participate on this forum and help others by answering questions that arise (helpful students can see a grade bump across a threshold (e.g., B+ to A-) for this participation).
- When helping with homework questions, keep the discussion to the high-level concepts; don't post answers to homeworks or quiz/exam questions.

TAs

- Katie Stasaski (OH **Tues** 10-11:30am)
- Jon Gillick (OH **Wed** 3:30-5pm)
- Chloe Lee (OH **Thurs** 3:30-5pm)
- Janaki Vivrekar (OH **Fri** 12-1:30pm)
- Visit TA office hours for help with homeworks/quizzes/exams or just to chat about NLP.
- TA OH will be held through Discord; each OH will have channels organized around topics that you can chat with other students about.

TAs

- Keep academic integrity in mind during TA office hours: you may discuss homework questions at a high level with others present, but don't discuss specific answers or share screens with code solutions. Neither the TA office hours or Piazza should be used for pre-grading (asking if a specific answer to a homework or quiz question is correct before the assignment is due).

DB office hours

- DB office hours **Wed** 10am-noon (Zoom link on bCourses)
- Come talk to me to discuss concepts from class and NLP more generally — I'm happy to chat!



Next time:

Construction of truth; ethics