Natural Language Processing

Info 159/259

Lecture 1: Introduction (Aug 24, 2017)

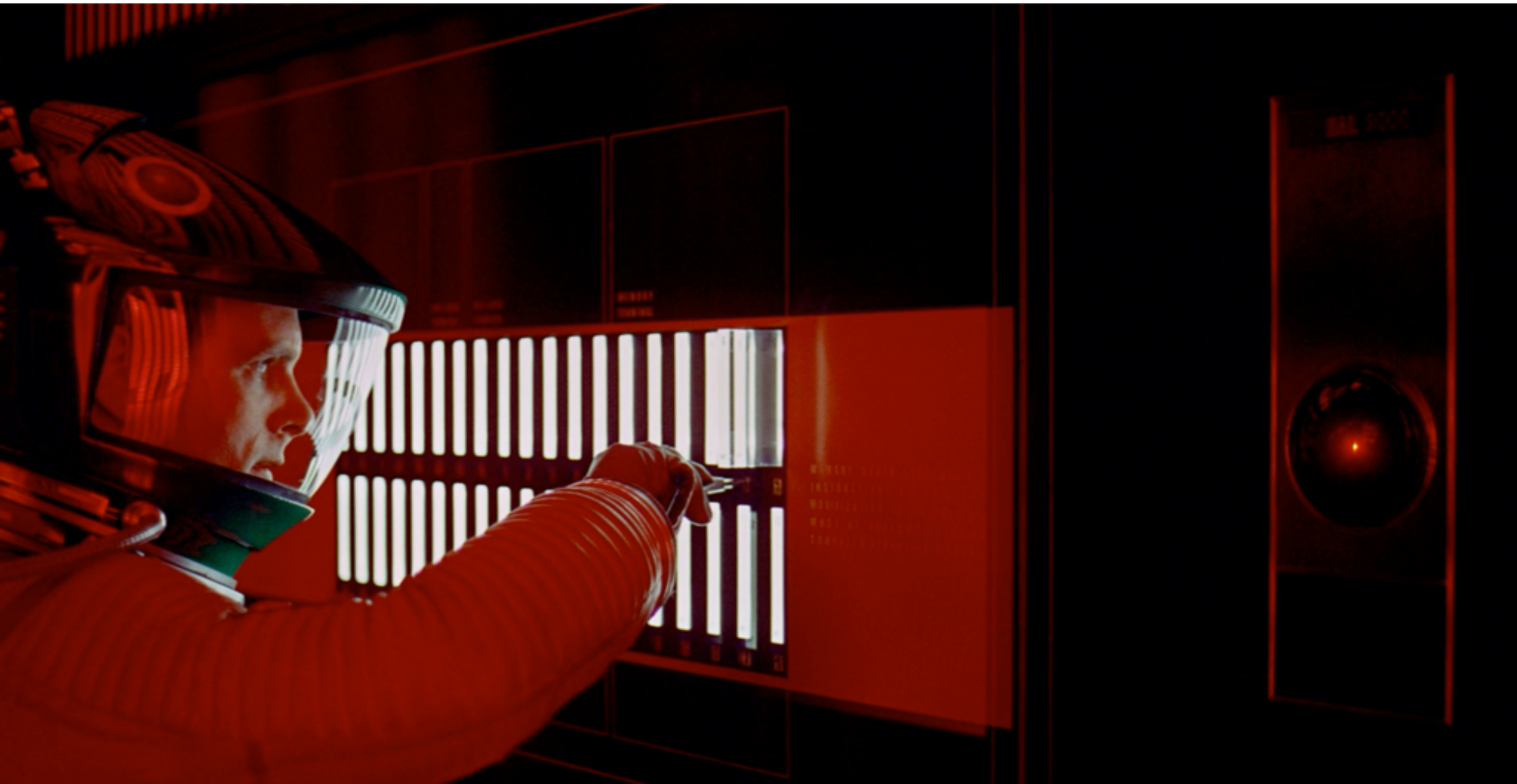
David Bamman, UC Berkeley

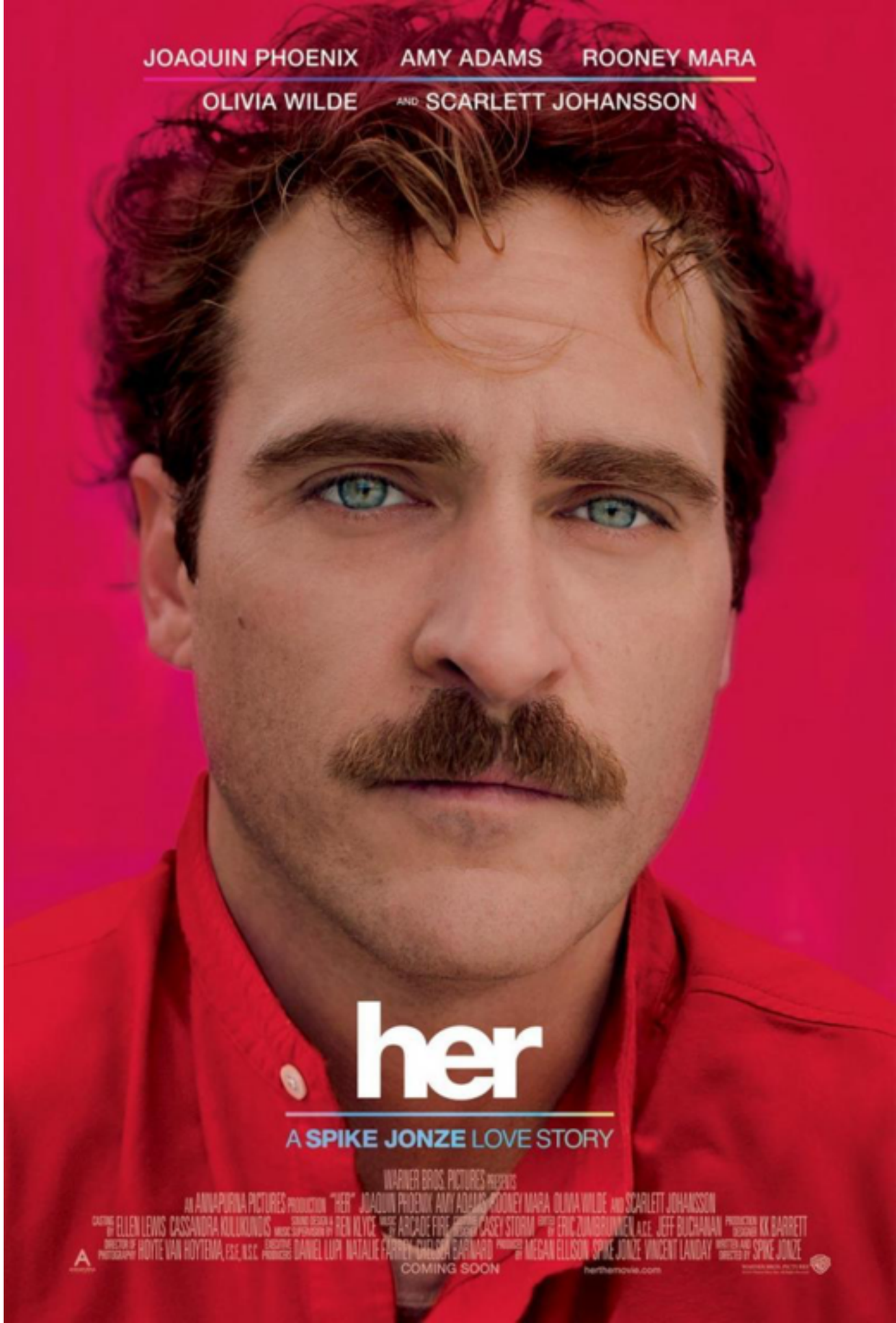
NLP is interdisciplinary

- 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”

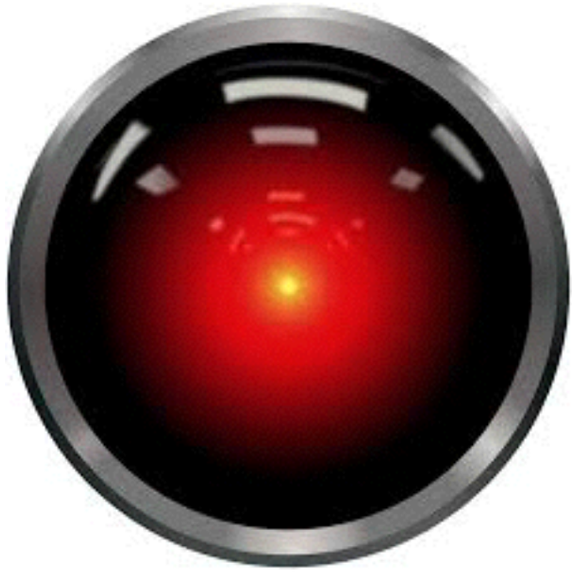



JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
COMING SOON

Turing test

Distinguishing human vs. computer only through written language

Turing 1950



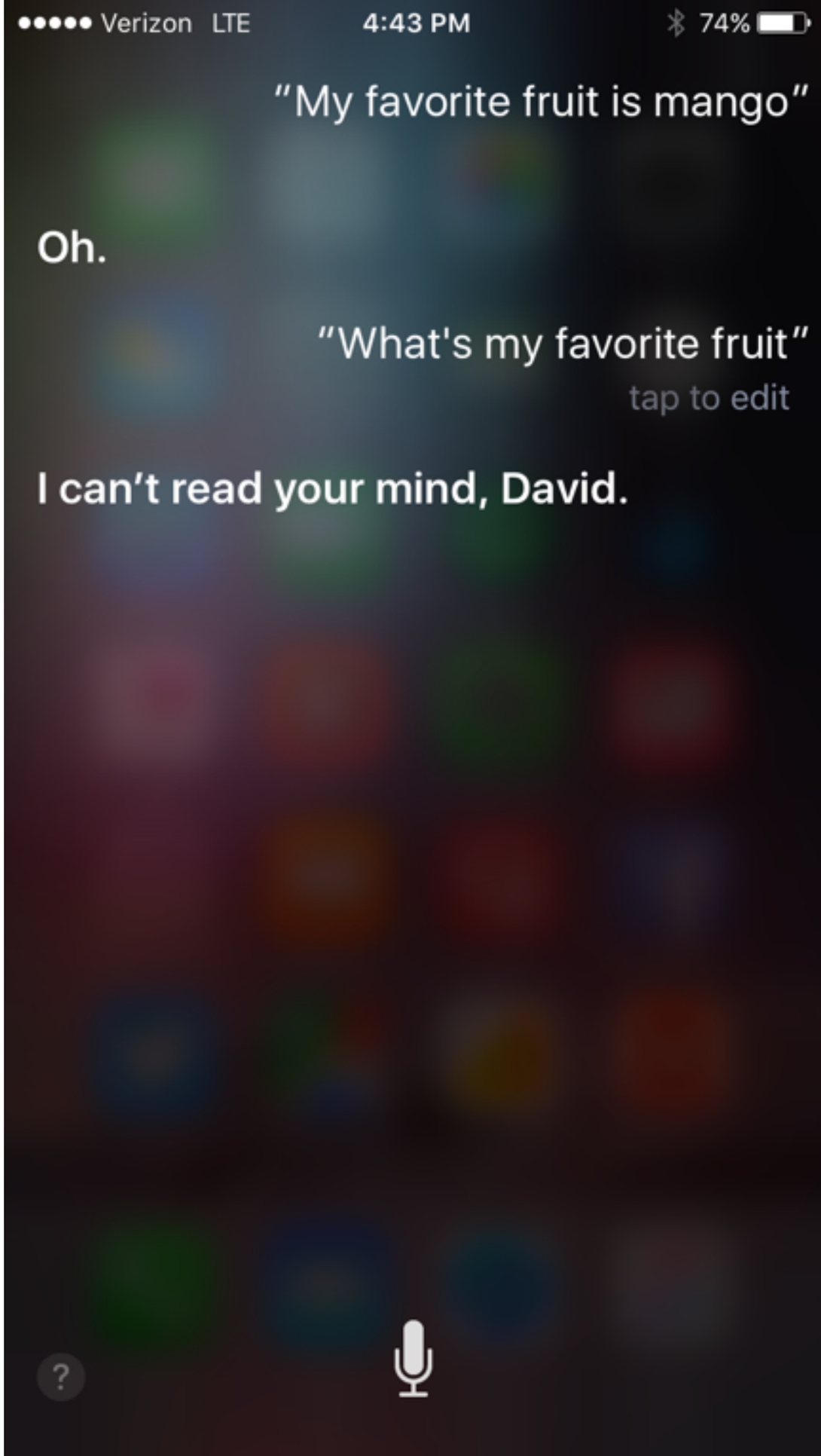
Dave Bowman: Open the pod bay doors, HAL
HAL: I’m sorry Dave. I’m afraid I can’t do that

Agent	Movie	Complex human emotion mediated through language
Hal	2001	Mission execution
Samantha	Her	Love
David	Prometheus	Creativity

Where we are now

Where we are now

Where we are now



Baseline mutual information model (Li et al. 2015)
A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...

A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about . (6)
A: You don't know what you are saying. (7)
...

Li et al. (2016), "Deep Reinforcement Learning for Dialogue Generation" (EMNLP)

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 (“Clinton is running for election”)
- 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]

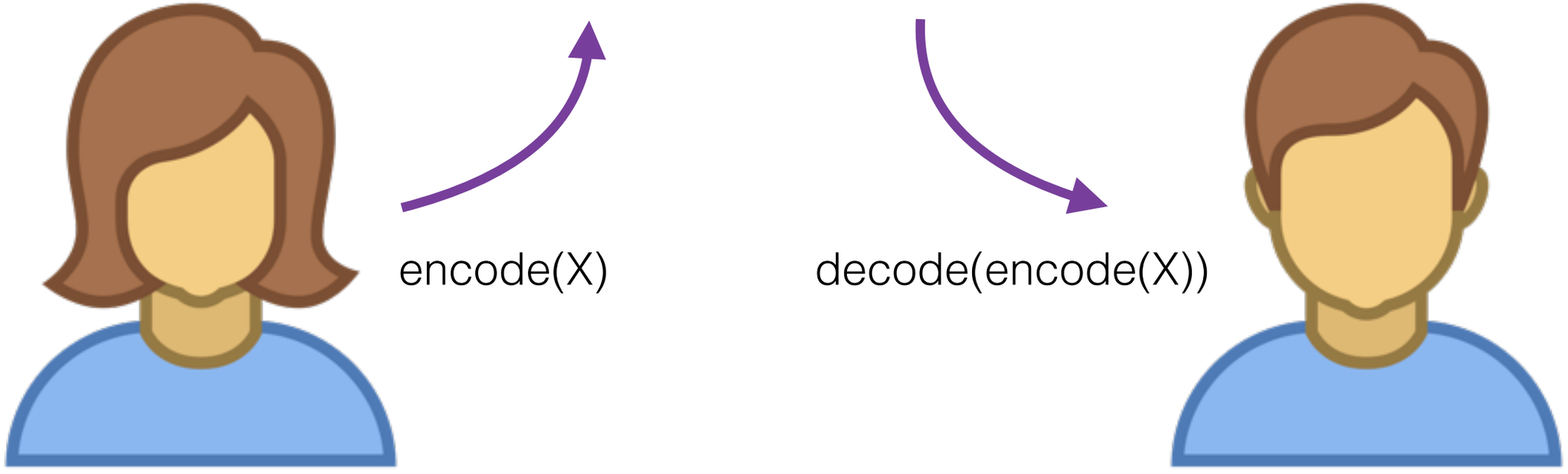
- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- …

processing as representation

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

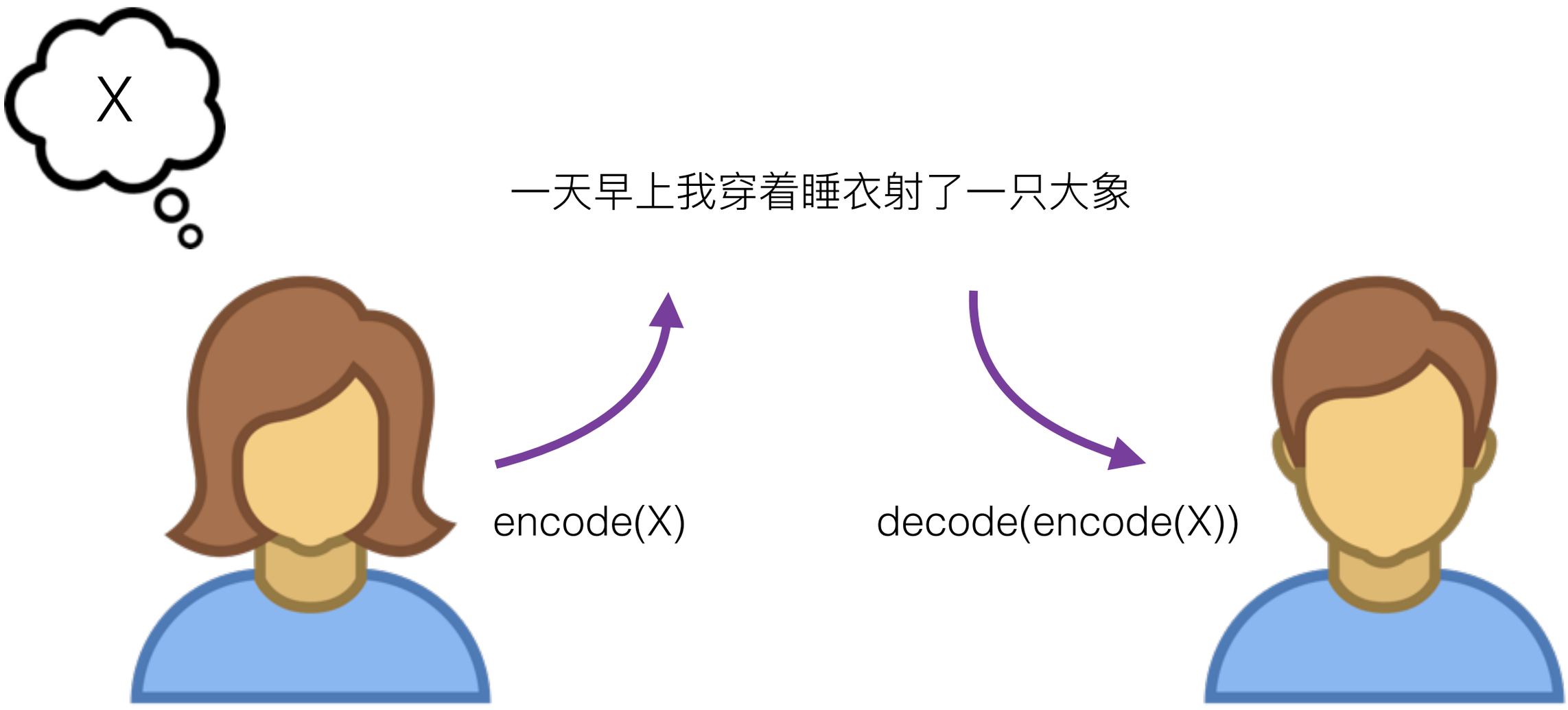
Information theoretic view

"One morning I shot an elephant in my pajamas"



Shannon 1948

Information theoretic view

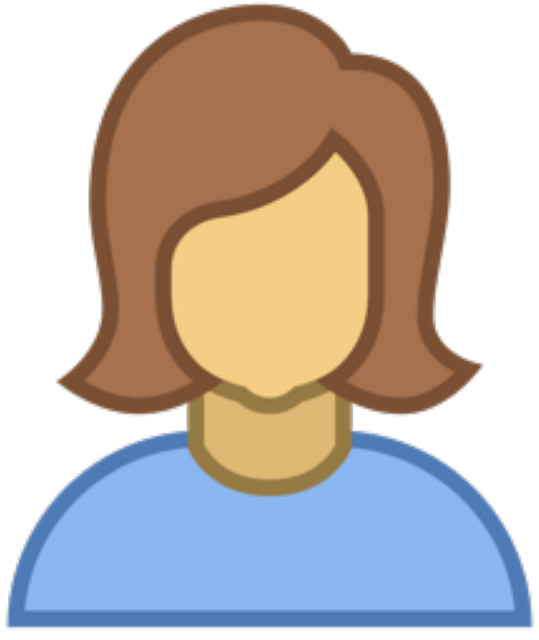


When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Weaver 1955

Rational speech act view

"One morning I shot an elephant in my pajamas"

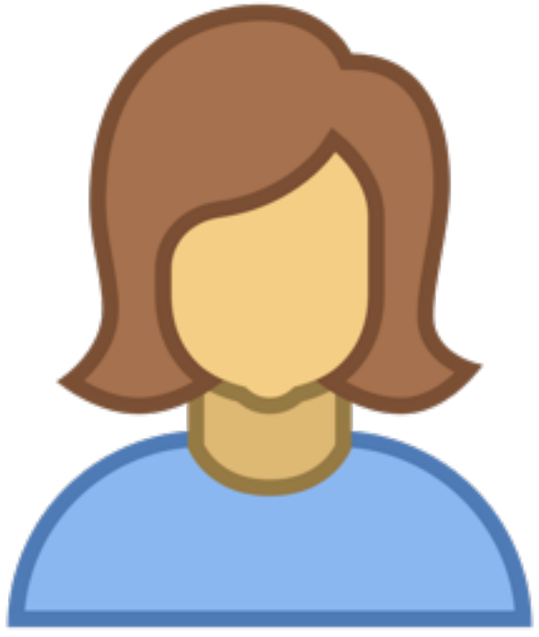


Communication involves recursive reasoning: how can X choose words to maximize understanding by Y?

Frank and Goodman 2012

Pragmatic view

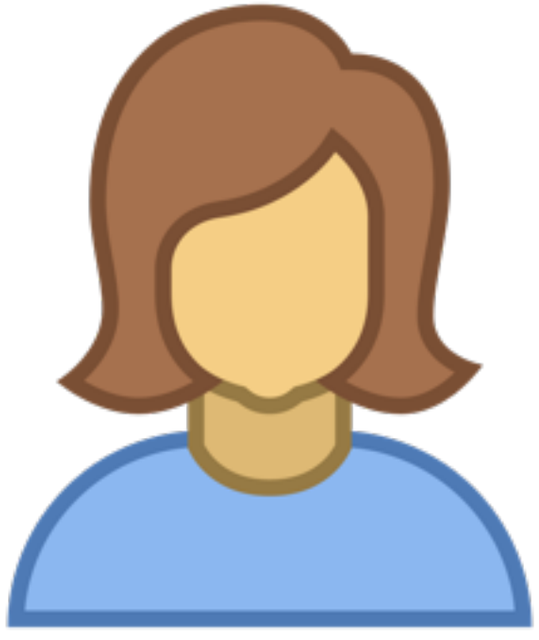
"One morning I shot an elephant in my pajamas"



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

Whorfian view

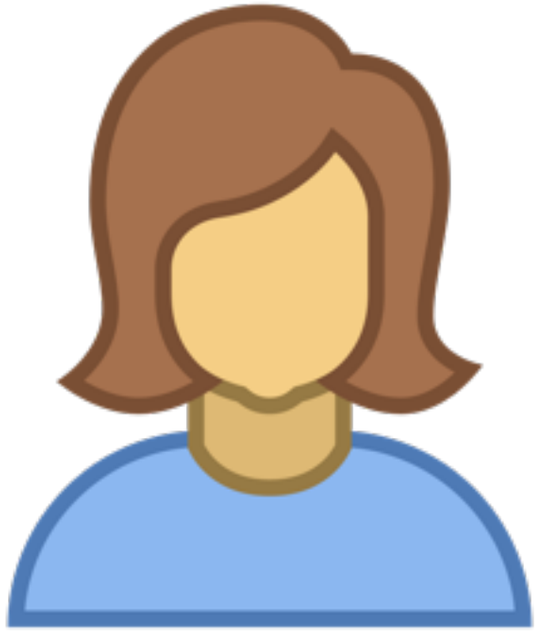
"One morning I shot an elephant in my pajamas"



Weak relativism: structure of language influences thought

Whorfian view

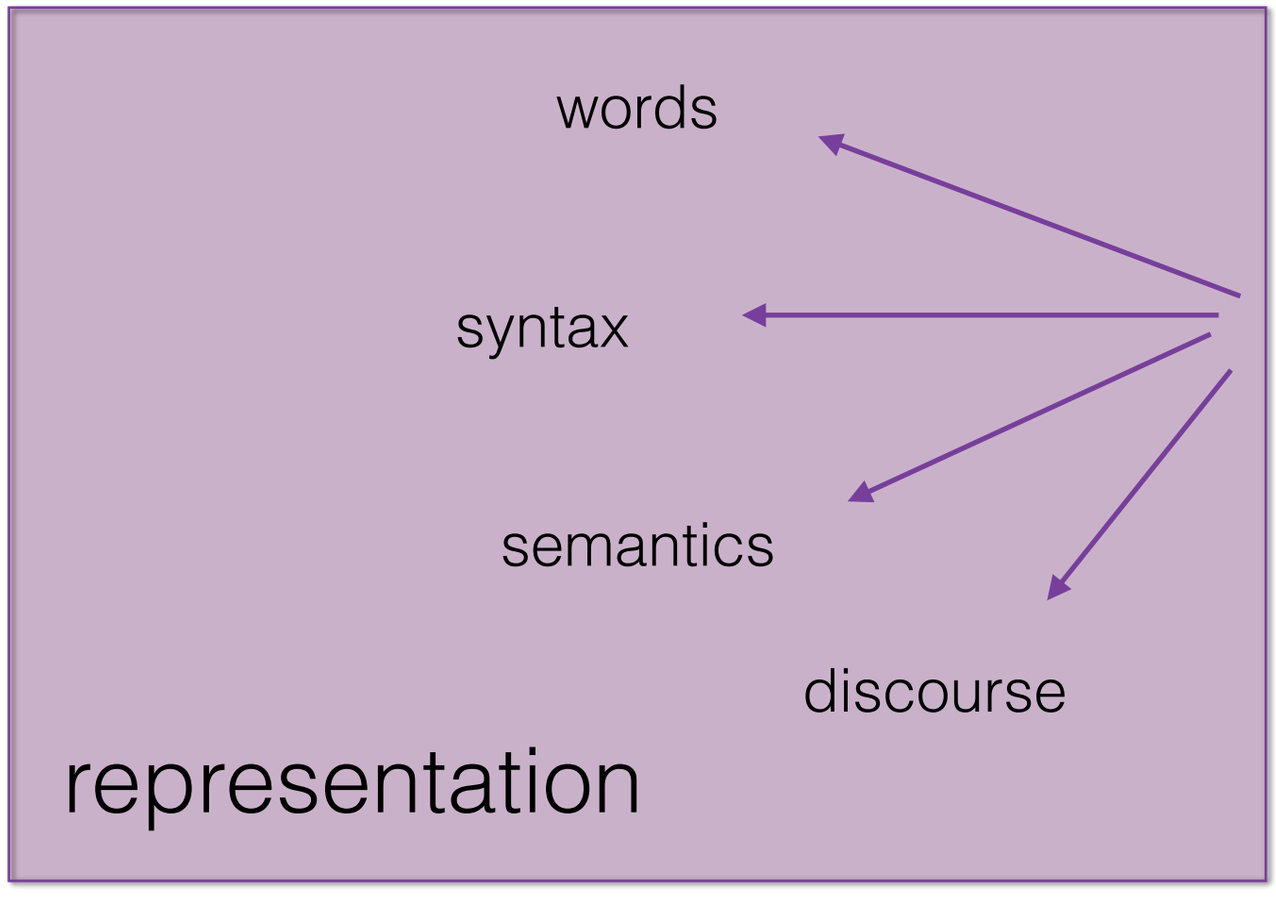
一天早上我穿着睡衣射了
一只大象



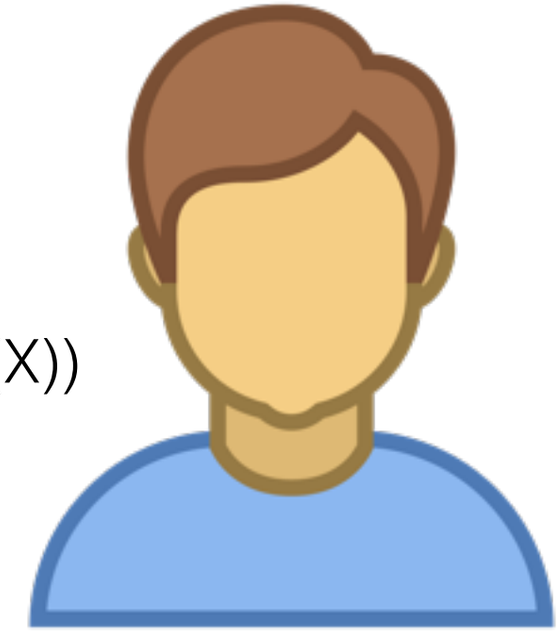
Weak relativism: structure of language influences thought

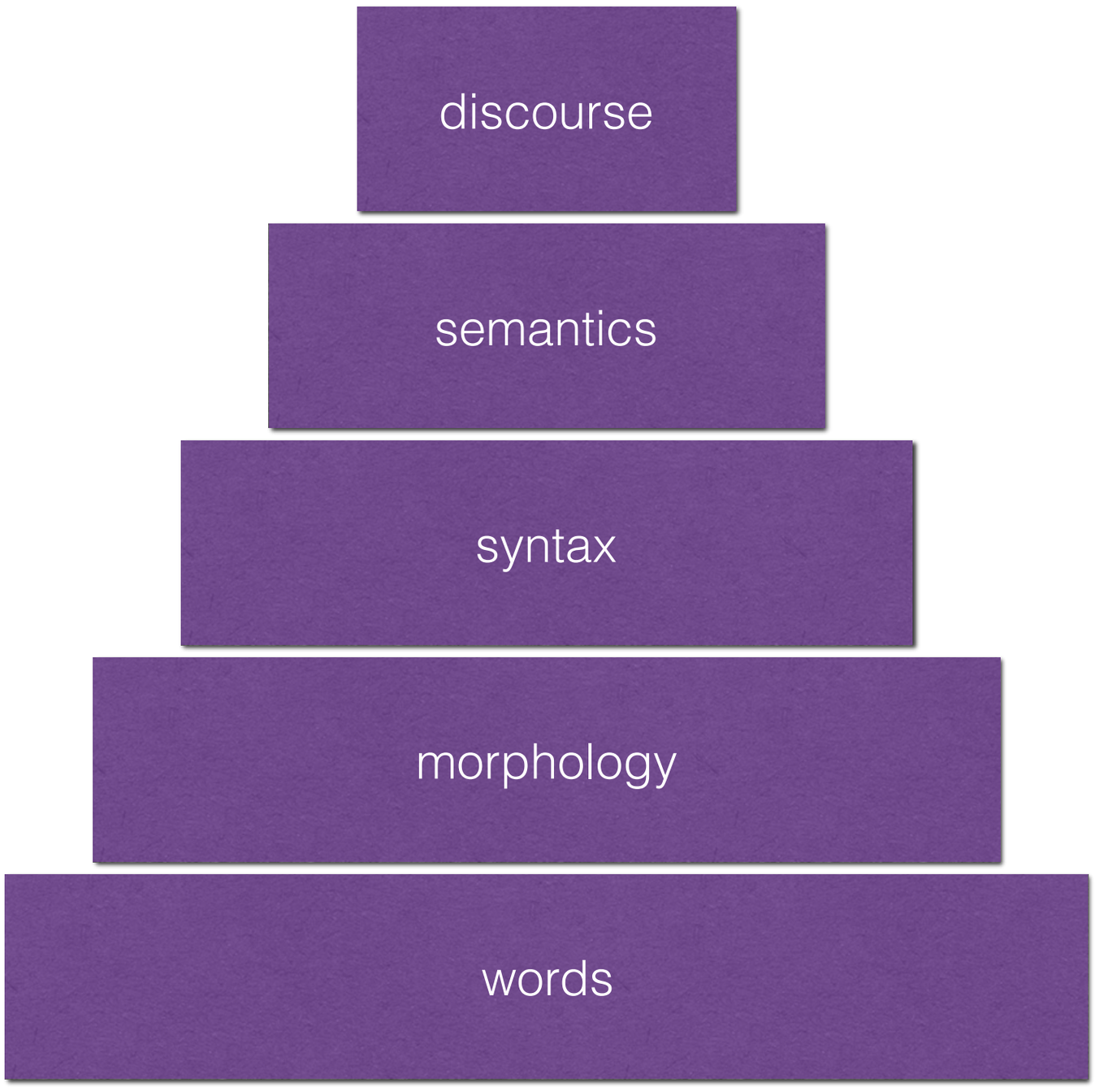
Decoding

"One morning I shot an elephant in my pajamas"



decode(encode(X))




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

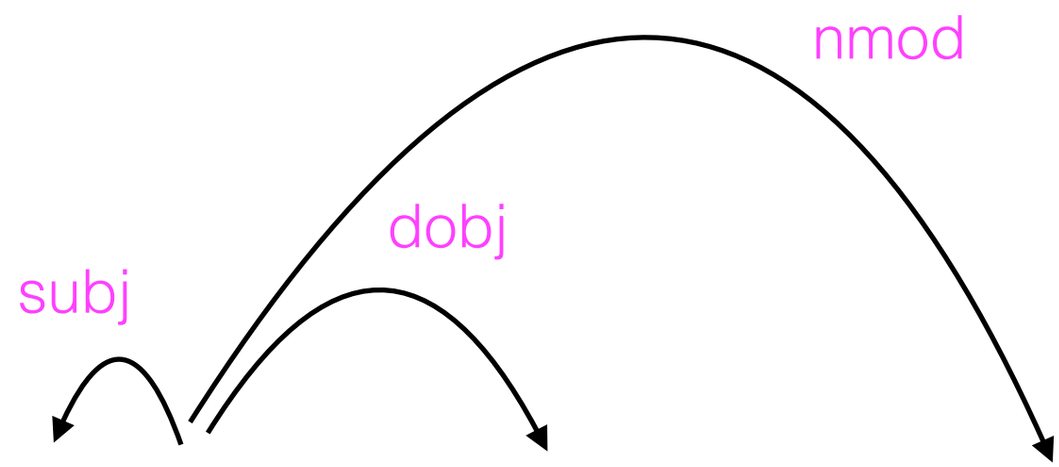
One morning I shot an elephant in my pajamas

Named entities

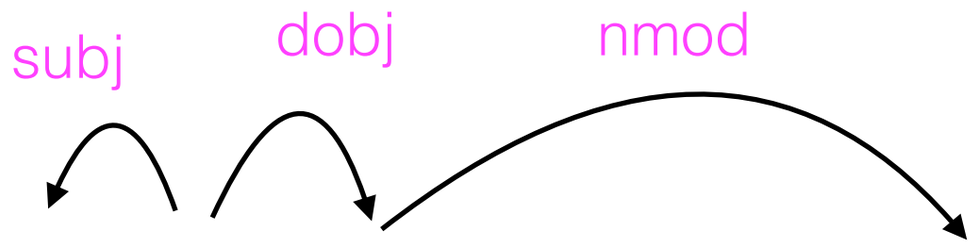
person

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

Syntax



One morning I shot an elephant in my pajamas



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

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

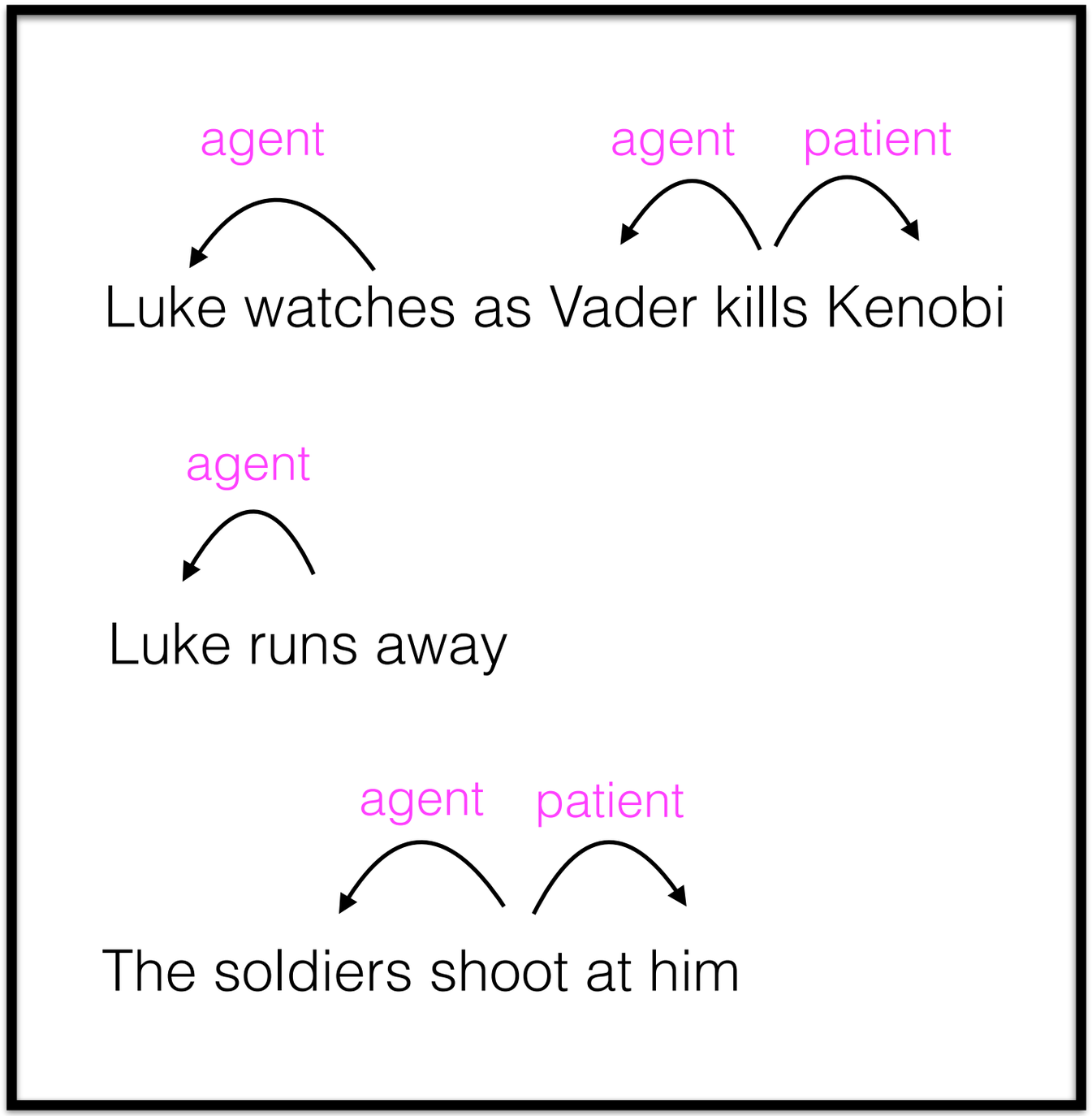
What did Barack Obama teach?

Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School between 1992 and 2004.

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

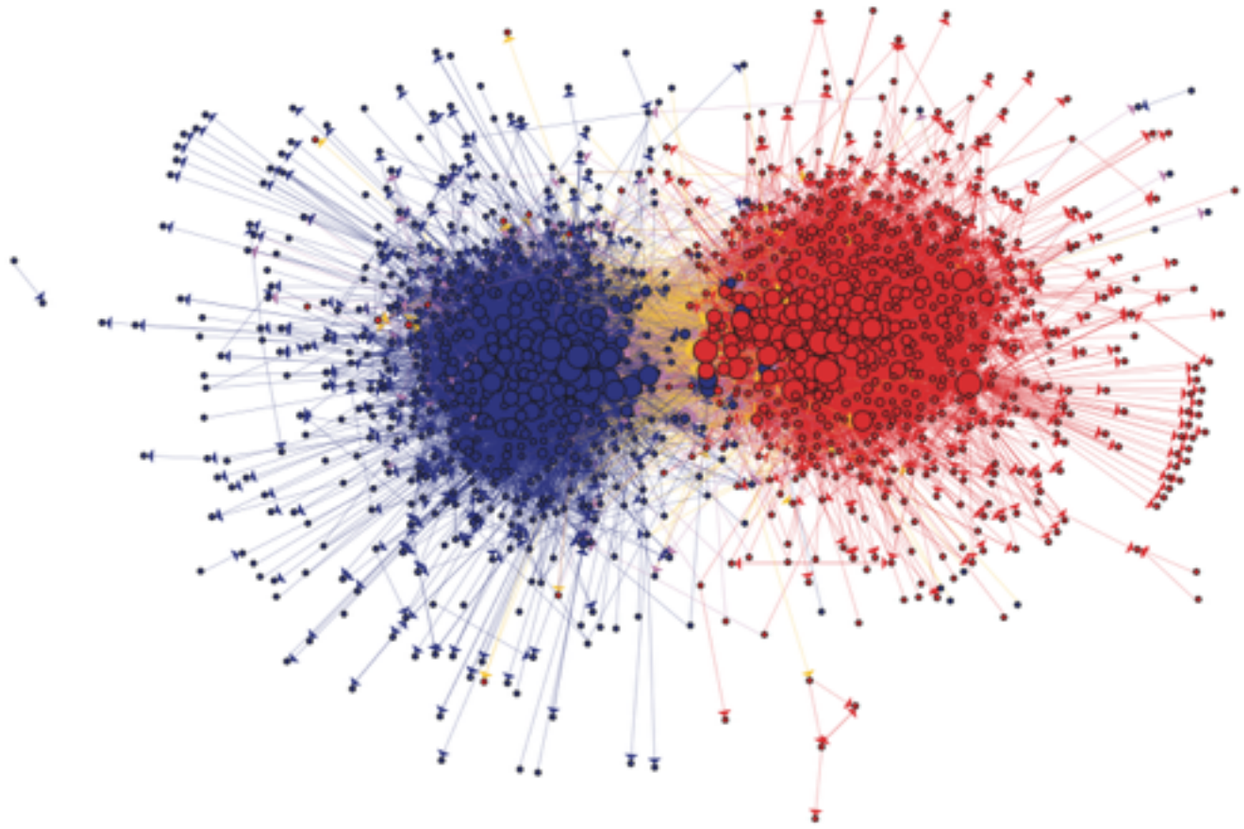
Google

IBM WATSON

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 (2016), “The Life Cycles of Genres,” Cultural Analytics

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

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

Andrew Goldstone and Ted Underwood (2014), “The Quiet Transformations of Literary Studies,” New Literary History

Franco Moretti (2005), Graphs, Maps, Trees

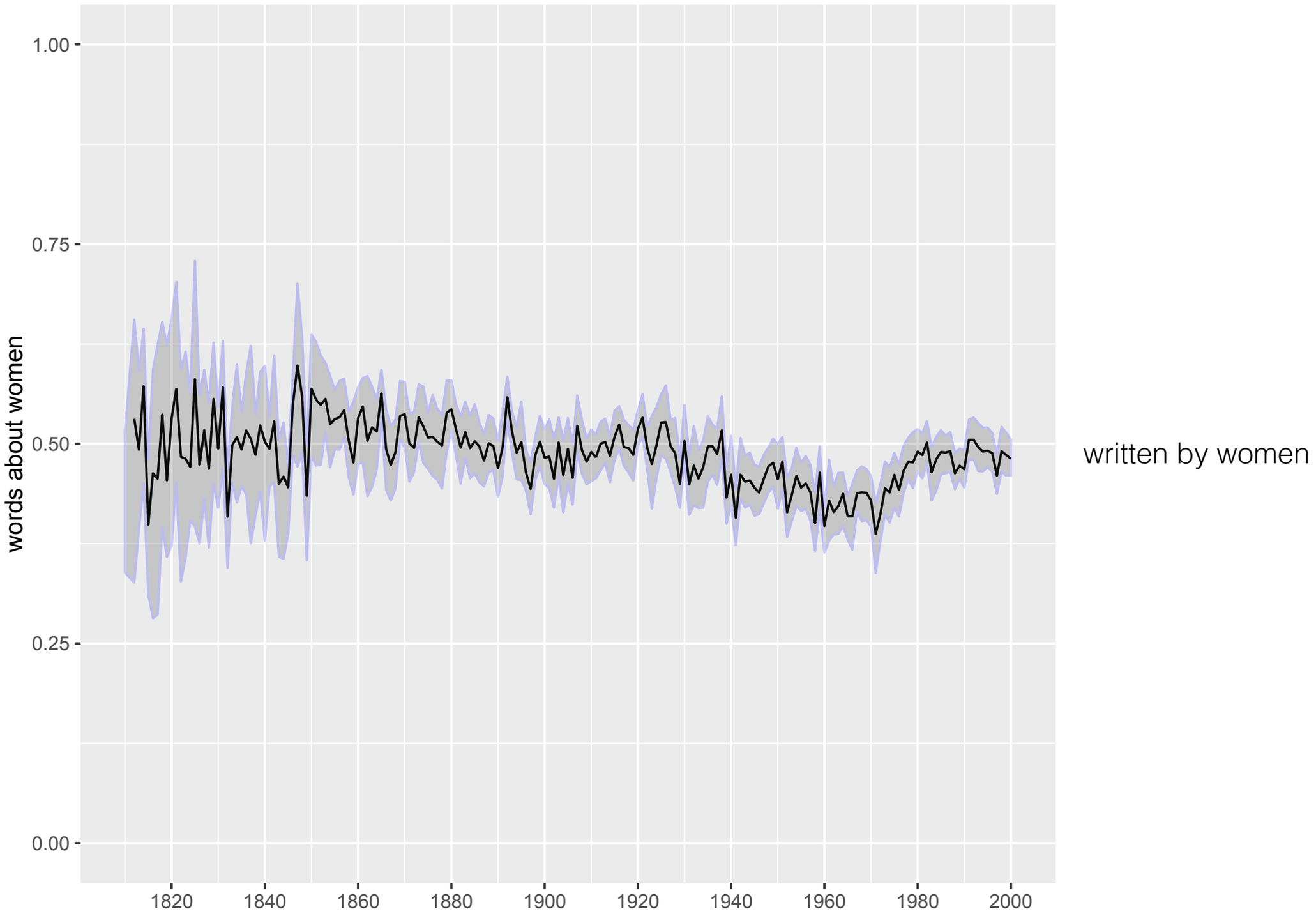
Holst Katsma (2014), Loudness in the Novel

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

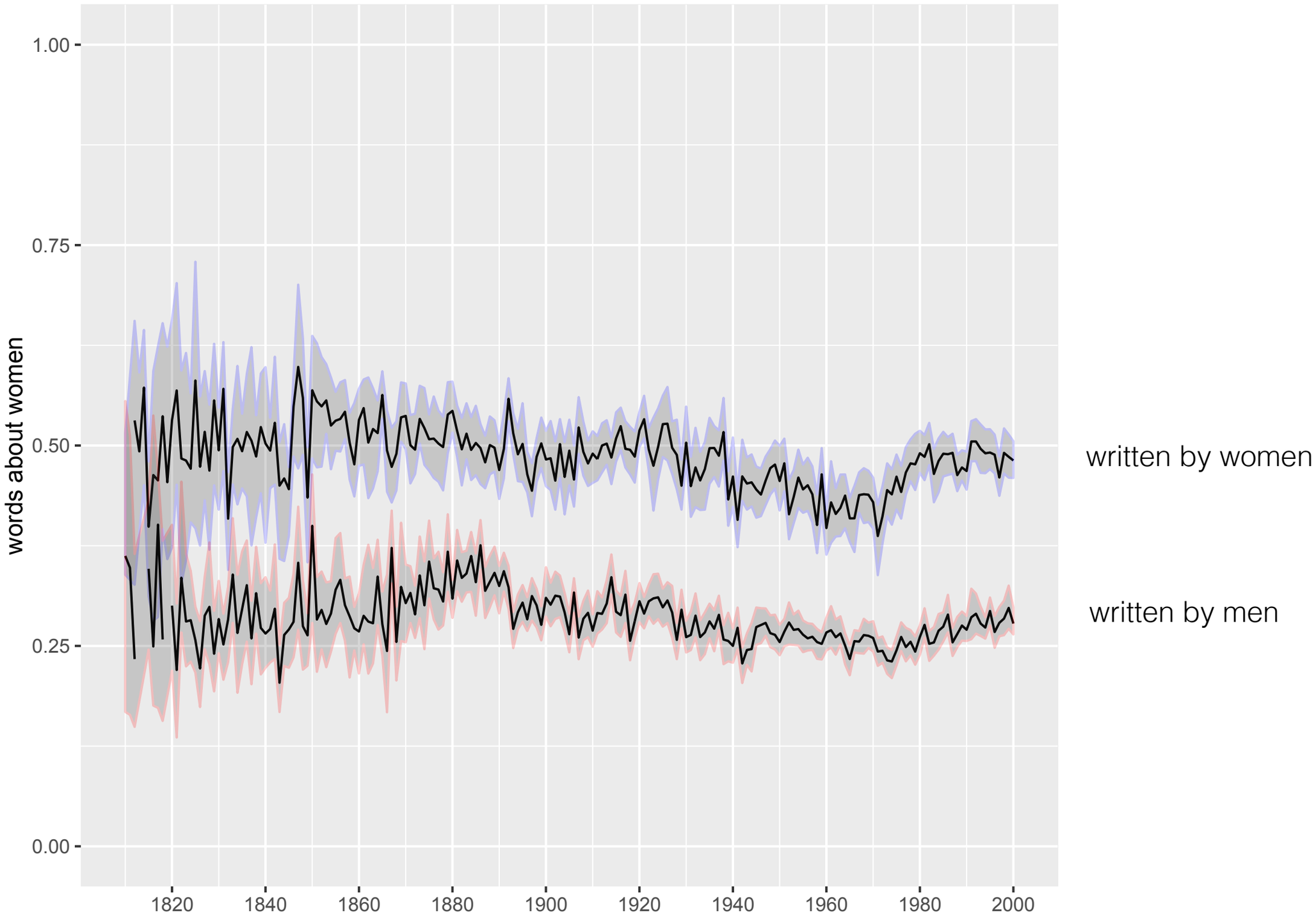
Matt Wilkens (2013), “The Geographic Imagination of Civil War Era American Fiction”

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

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

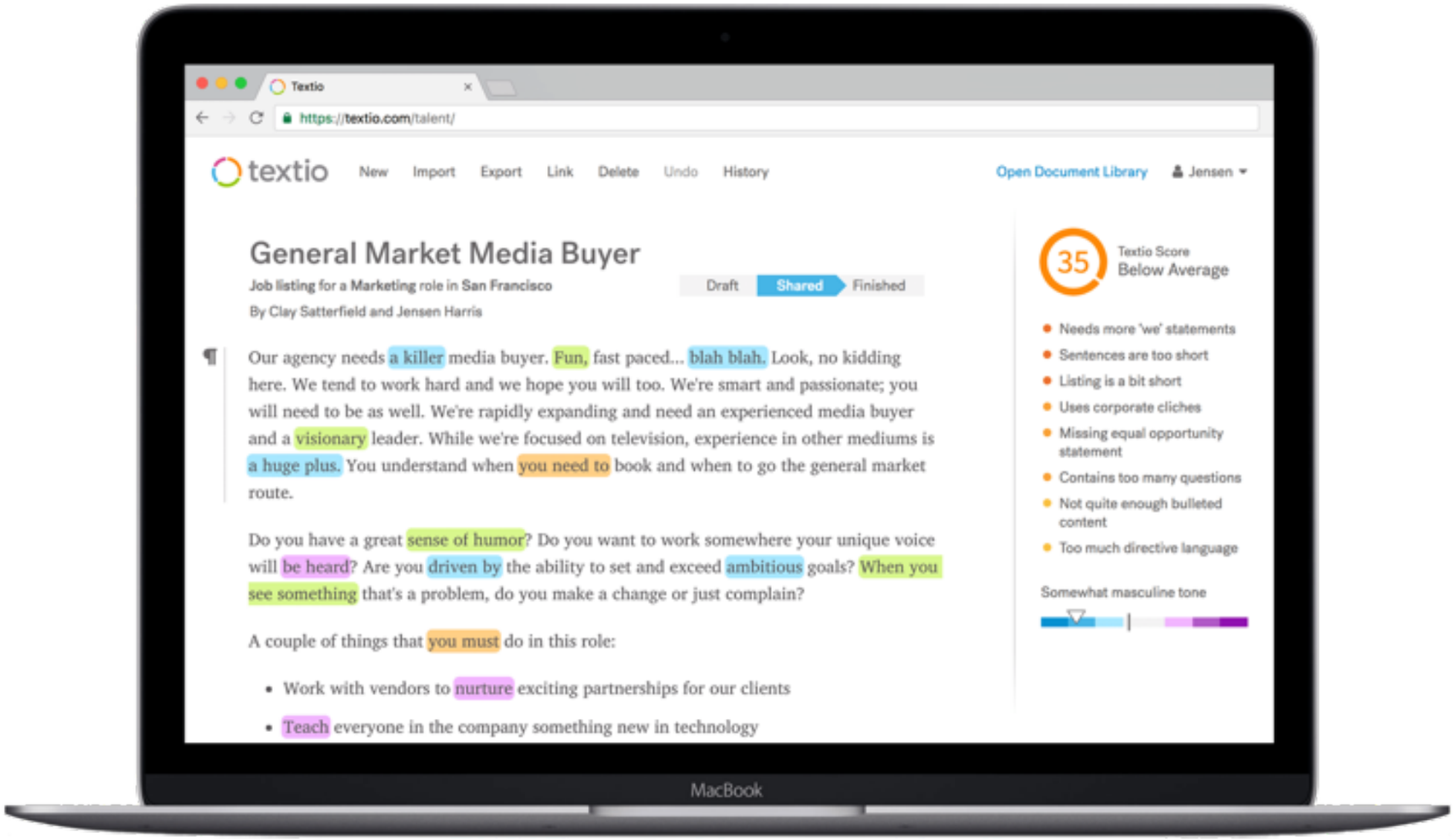


Ted Underwood and David Bamman (2016), “The Instability of Gender” (MLA);
“The Gender Balance of Fiction” (2017).



Ted Underwood and David Bamman (2016), “The Instability of Gender” (MLA);
“The Gender Balance of Fiction” (2017).

Text-driven forecasting



Methods

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

- Rule-based systems

Methods

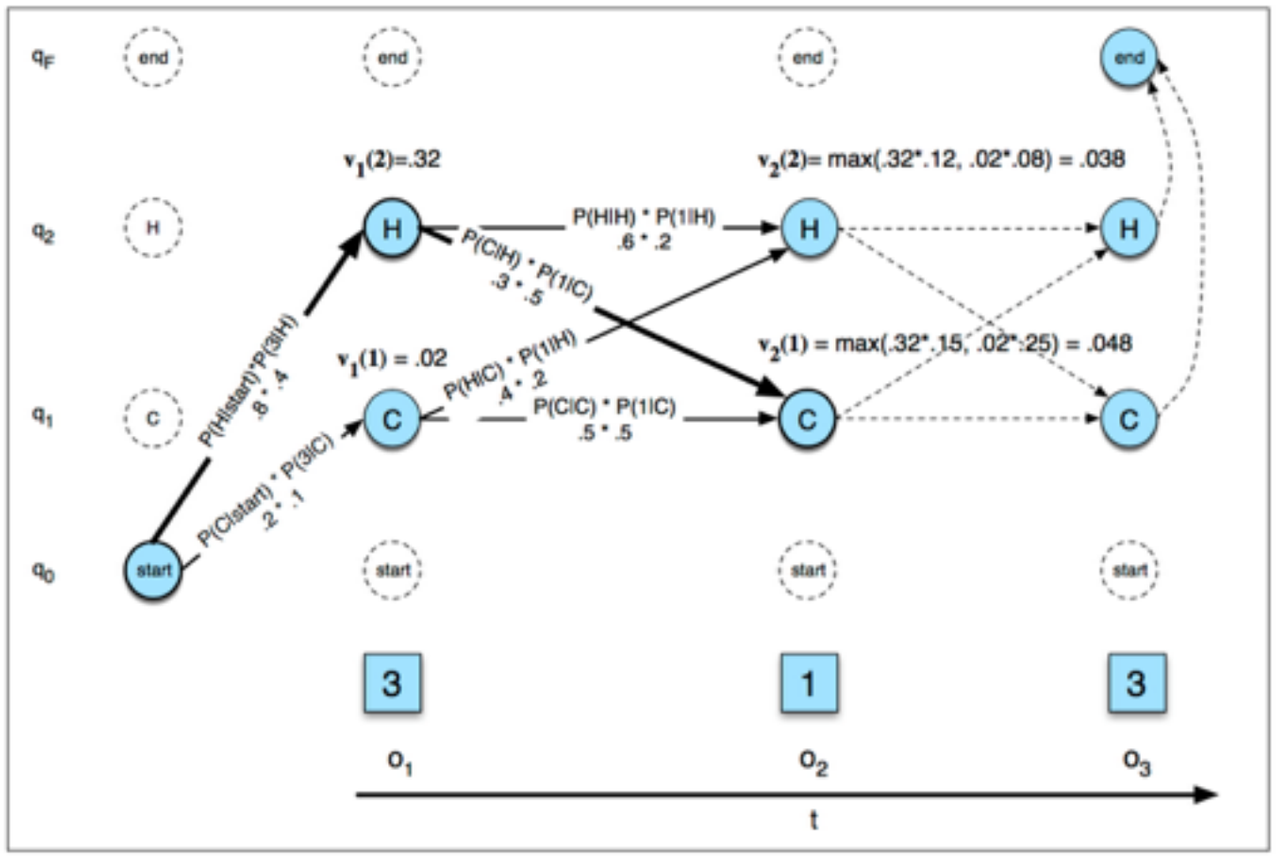
- Probabilistic models
- Naive Bayes, Logistic regression, HMM, MEMM, CRF, language models

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

Methods

- Dynamic programming (combining solutions to subproblems)

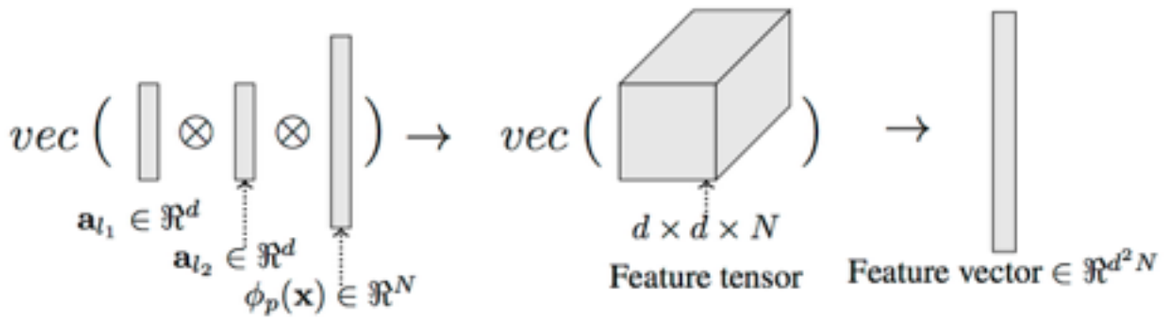
Viterbi algorithm, CKY



Viterbi lattice, SLP3 ch. 9

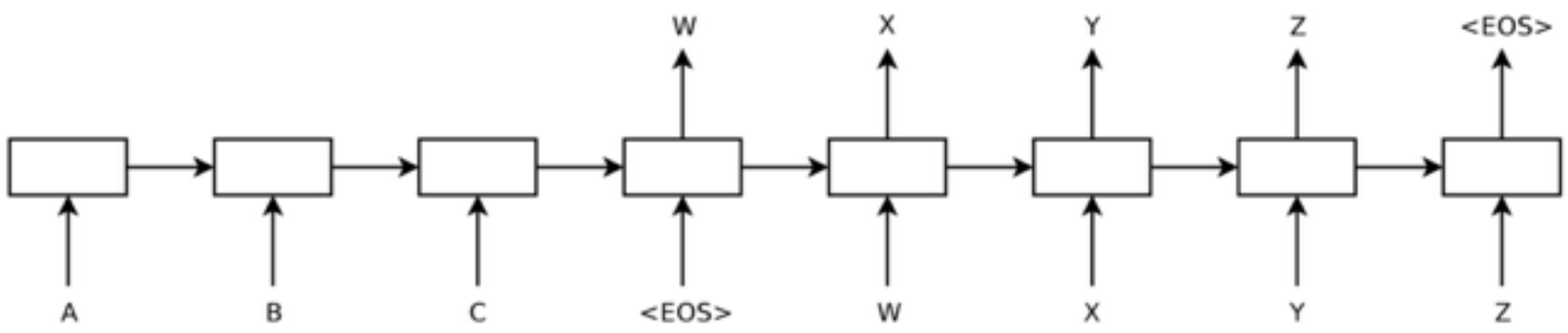
Methods

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



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

- Multiple, highly parameterized layers of (usually non-linear) interactions mediating the input/output ("deep neural networks")



Sutskever et al (2014), "Sequence to Sequence Learning with Neural Networks"

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”

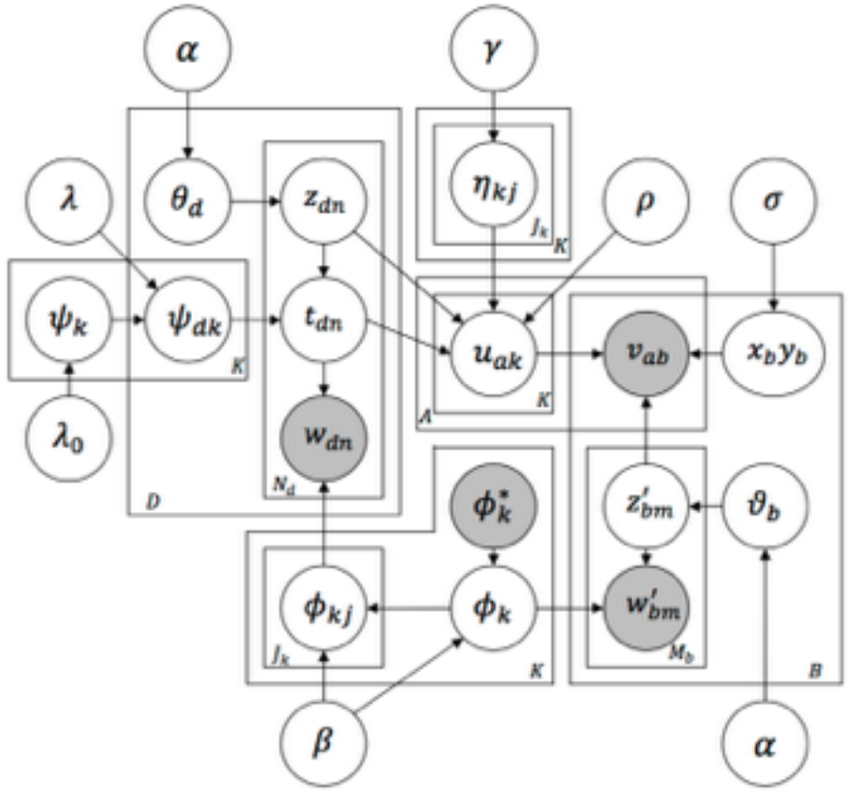


Figure 1: Plate notation diagram of HIPTM.

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 $backpointer[q_F,T]$

Viterbi algorithm, SLP3 ch. 9

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

Grading

- Info 159:
  - Midterm (20%) + Final exam (30%)
  - Take-home homeworks and in-class short quizzes (drop 3 lowest scores).

Quizzes

- Cover any material in current reading for that day or any material in previous lectures.

Homeworks

- ~ Half annotation exercises (learn the universal dependency representation of syntax and annotate some text)

- ~ Half modeling/algorithm exercises (derive the backprop updates for a CNN and implement it).

Late submissions

- All homeworks are due on the date/time specified; late homeworks won't be accepted after the deadline

- Note you can drop the lowest 3 scores on homeworks/quizzes; be judicious in how you manage that.

Grading

- Info 259:
  - Midterm (20%) + project (30%)
  - Take-home homeworks and in-class short quizzes (drop 3 lowest scores).

259 Project

- Semester-long project (involving 1 or 2 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
    - 8-page final report, workshop quality
    - Poster presentation

ACL 2017 workshops

- CLPsych: Computational Linguistics and Clinical Psychology
- Workshop on NLP and Computational Social Science
- Repl4NLP: 2nd Workshop on Representation Learning for NLP
- LaTeCH-CLfL: Workshop on Computational Linguistics for Literature
- TextGraphs-11: Graph-based Methods for NLP
- ALW1: 1st Workshop on Abusive Language Online
- EventStory: Events and Stories in the News

Waitlisted

- Come to class, complete assignments

Next time

- Sentiment analysis and text classification
- Read SLP3 chapter 6 (on syllabus)
- DB office hours tomorrow 10am-noon (314 South Hall)
- TAs (office hours Friday 9/1 2:30-3:30pm):
    - Yiyi Chen
    - Sayan Sanyal