Natural Language Processing

Info 159/259

Lecture 1: Introduction (Aug 23, 2018)

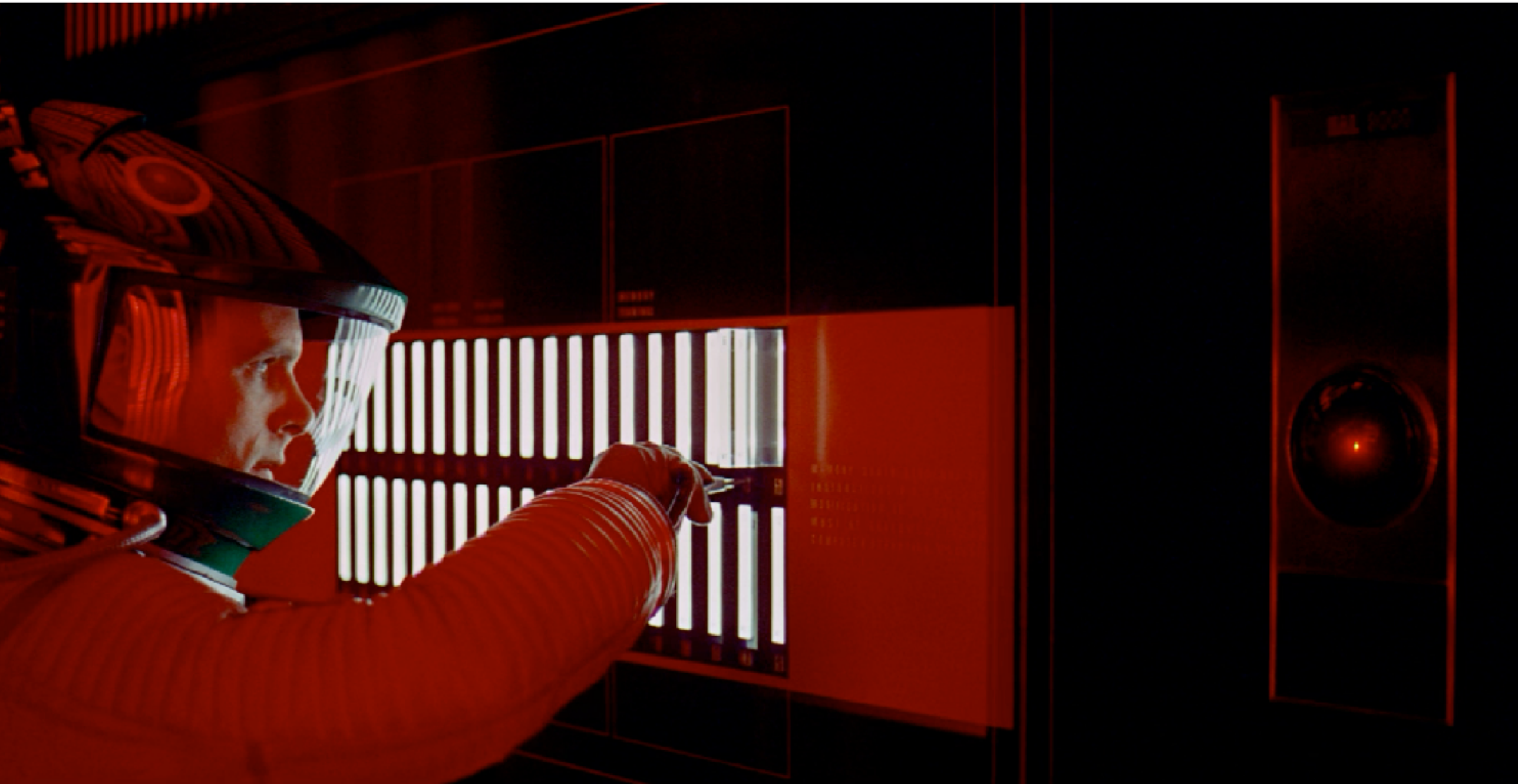
David Bamman, UC Berkeley

NLP is interdisciplinary

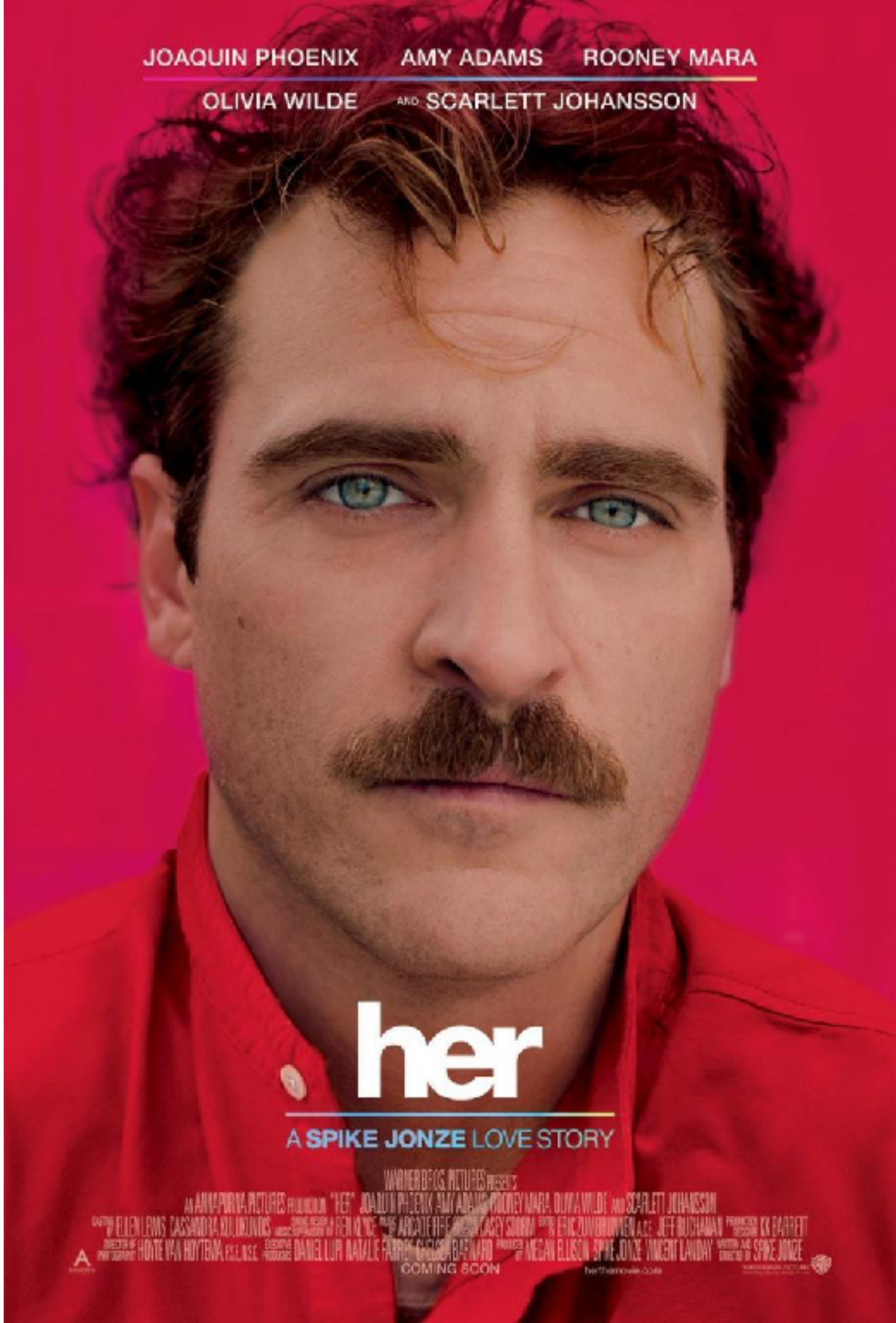
- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)

NLP = processing language with computers*

processing as “understanding”



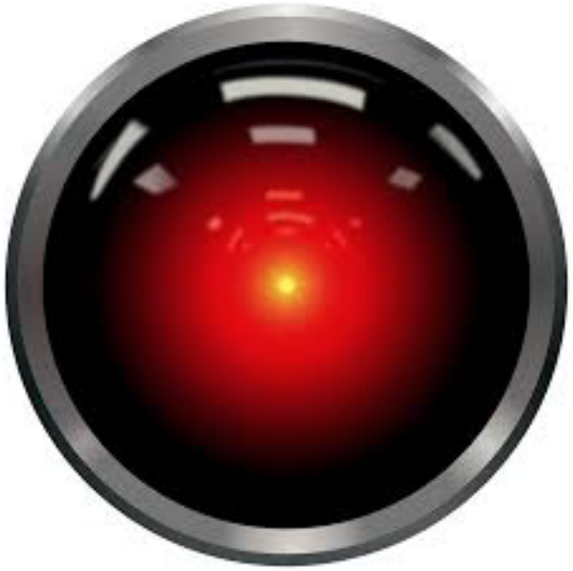
Grand Lake Theatre now!


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



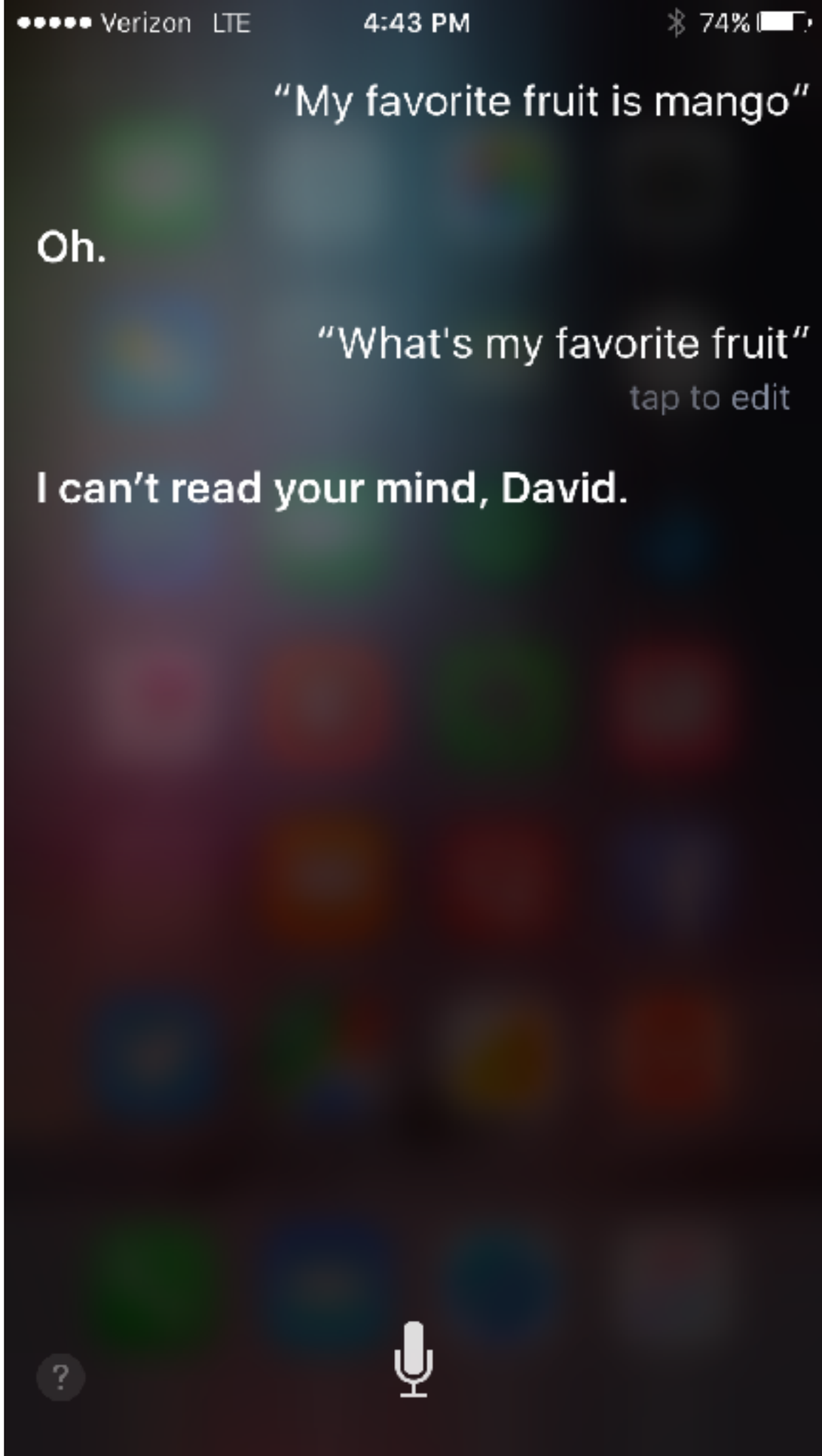
Dave Bowman: Open the pod bay doors, HAL
HAL: I'm sorry Dave. I'm afraid I can't do that

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

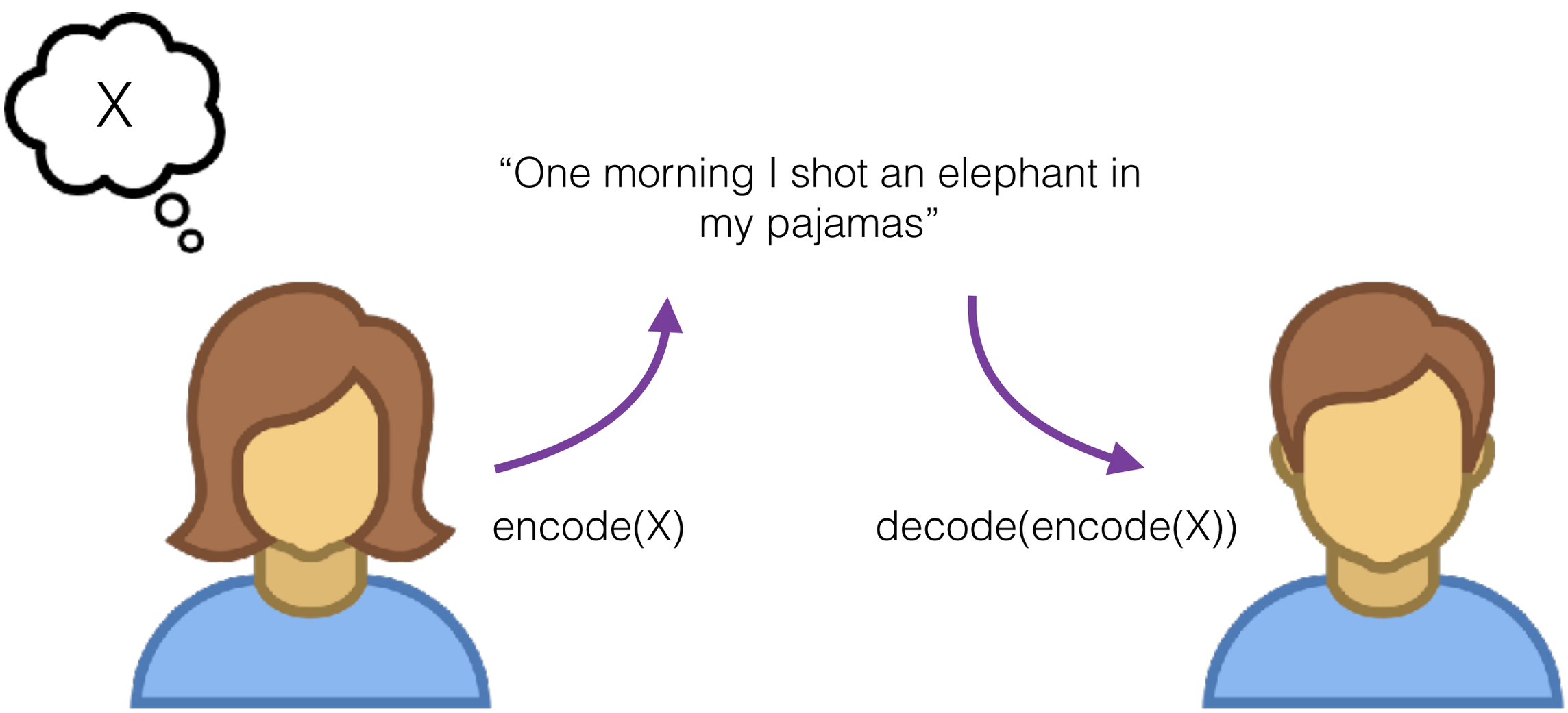
[SLP2 ch. 1]

- 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

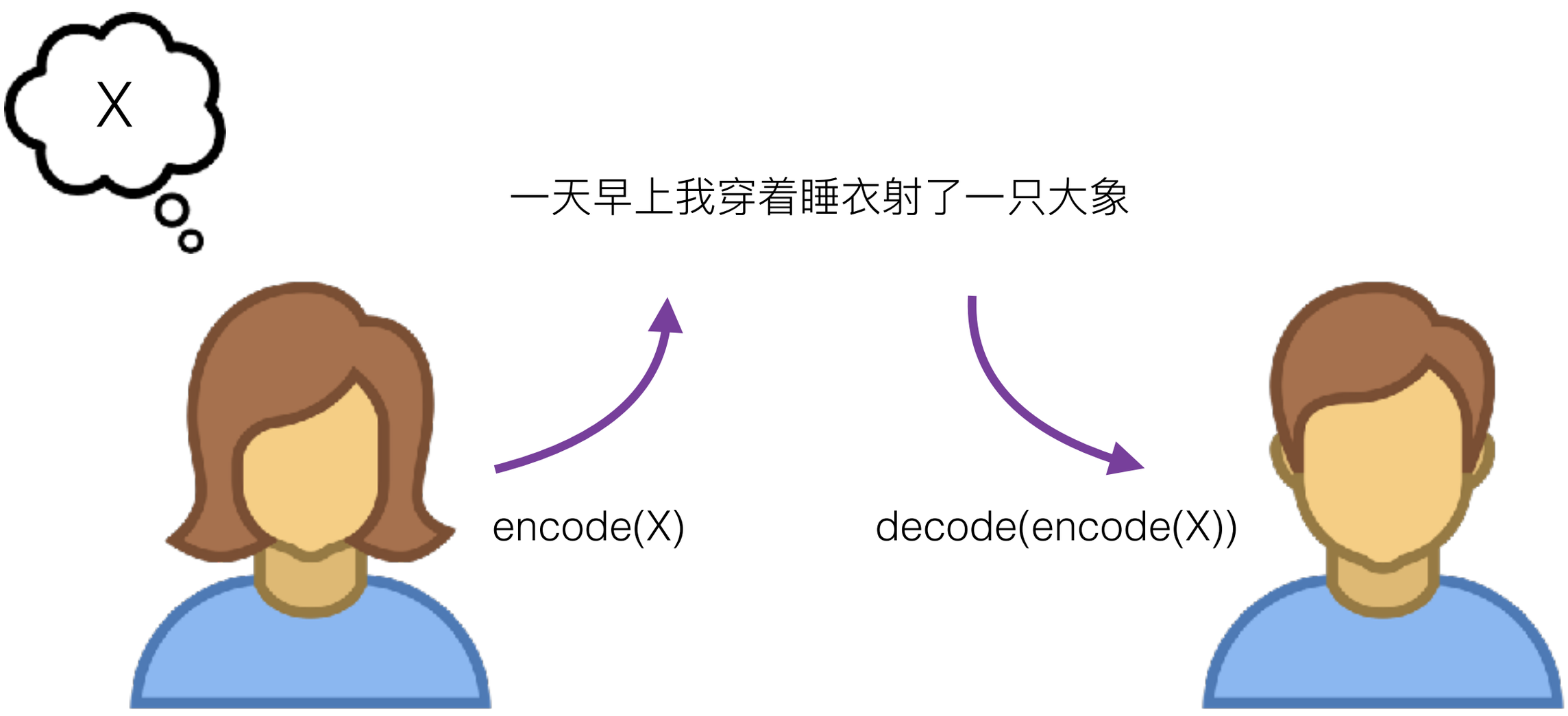
- NLP generally involves representing language for some end, e.g.:
    - dialogue
    - translation
    - speech recognition
    - text analysis

Information theoretic view



Shannon 1948

Information theoretic view

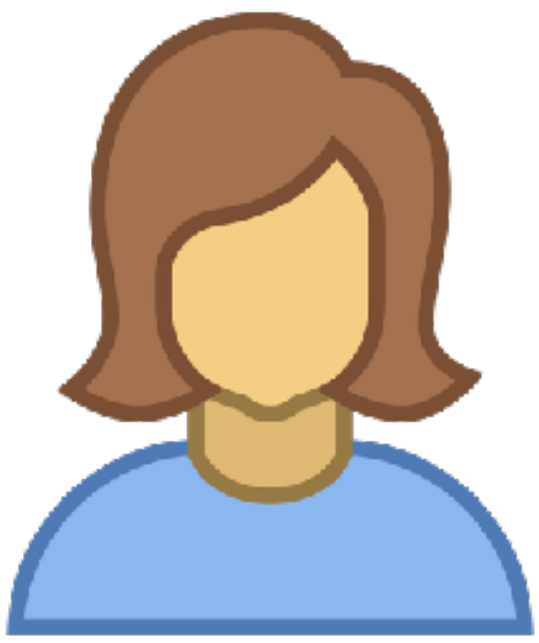


When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

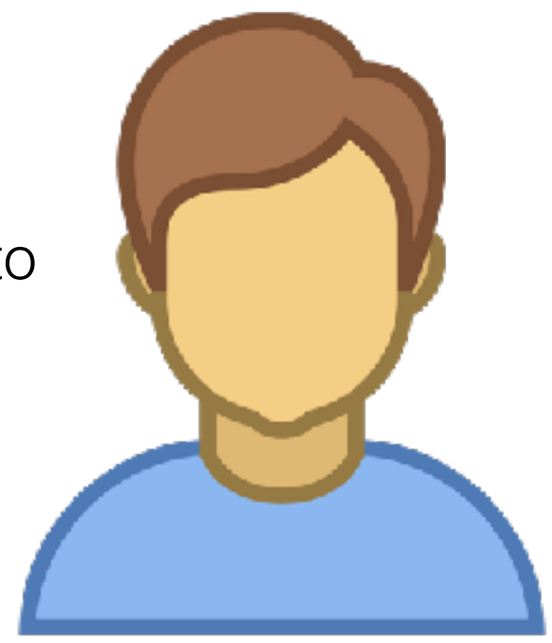
Weaver 1955

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

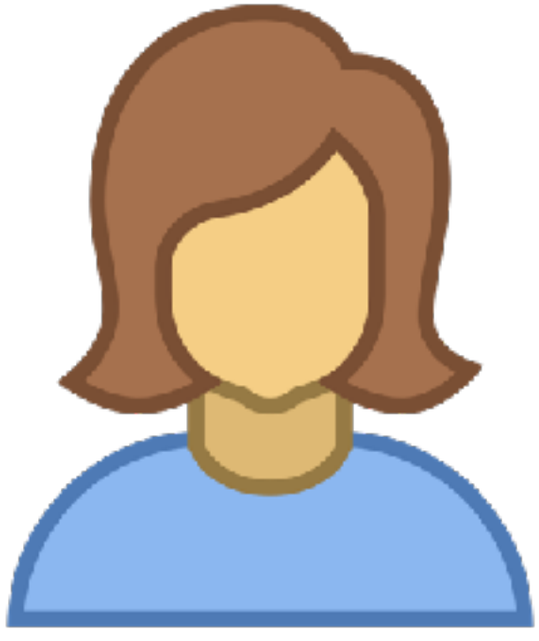
Whorfian view

“One morning I shot an elephant in my pajamas”

Weak relativism: structure of language influences thought

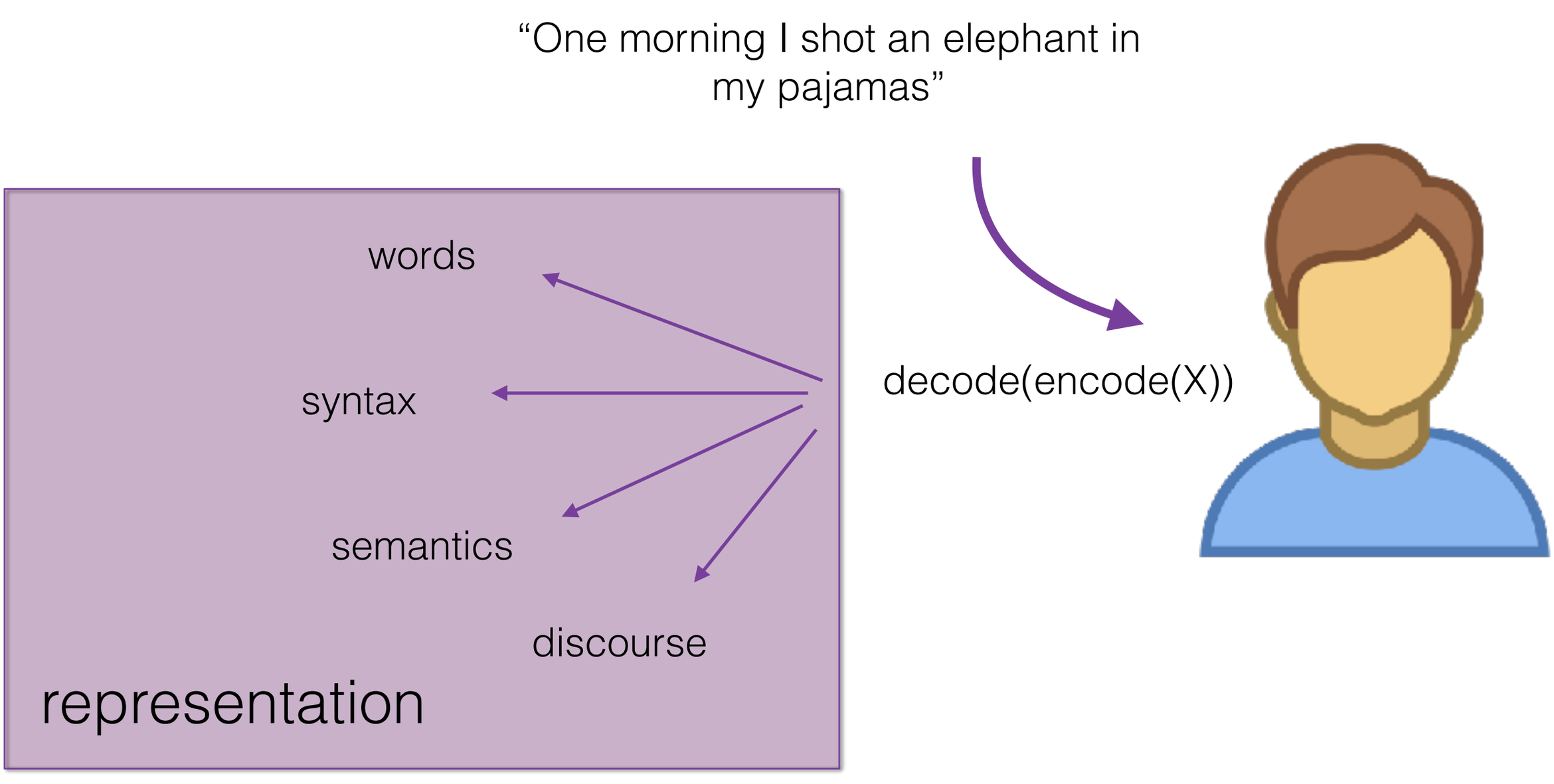
Whorfian view

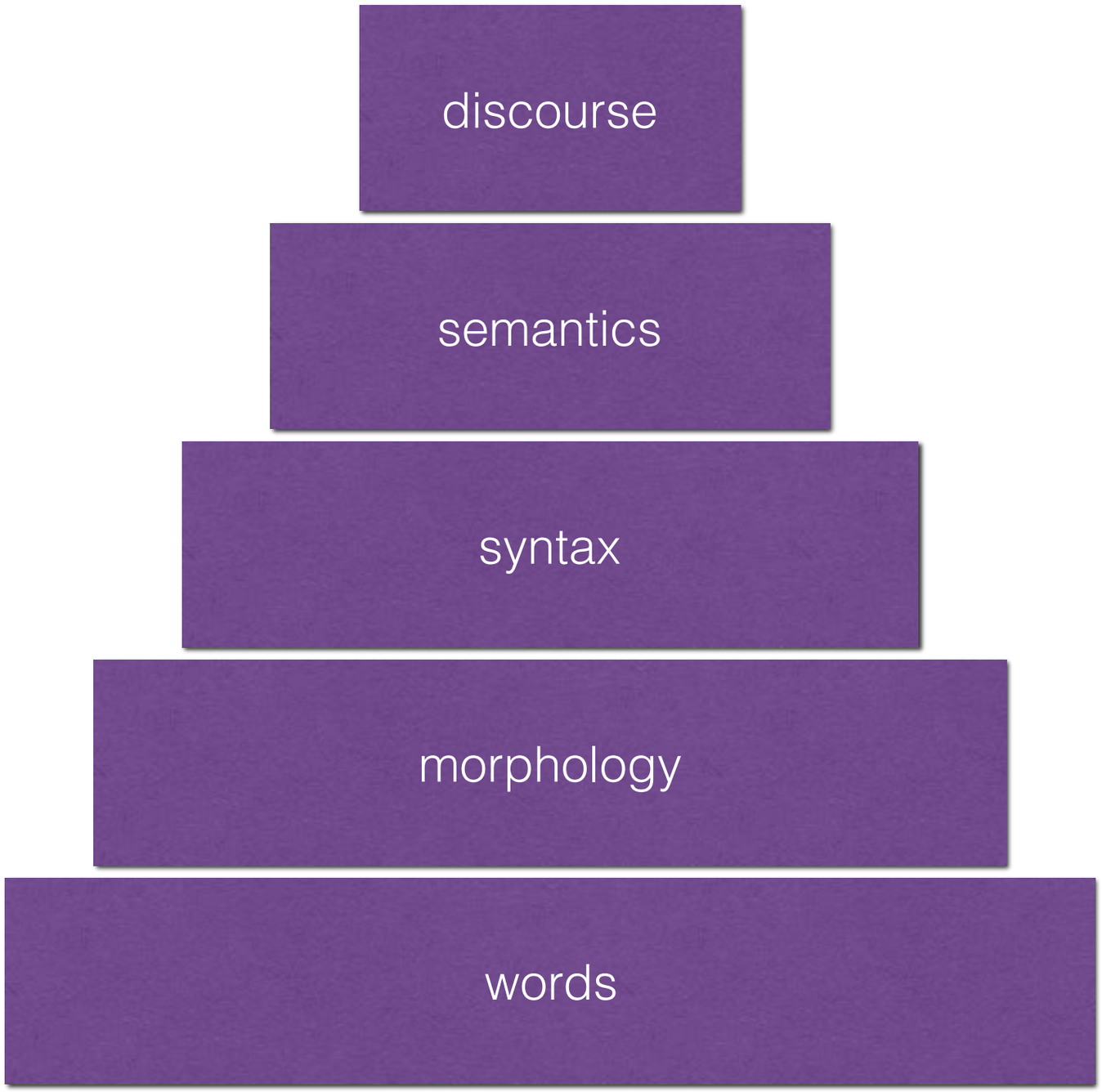
一天早上我穿着睡衣射了
一只大象



Weak relativism: structure of language influences thought

Decoding


"One morning I shot an elephant in my pajamas"
words
syntax
semantics
discourse
representation
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

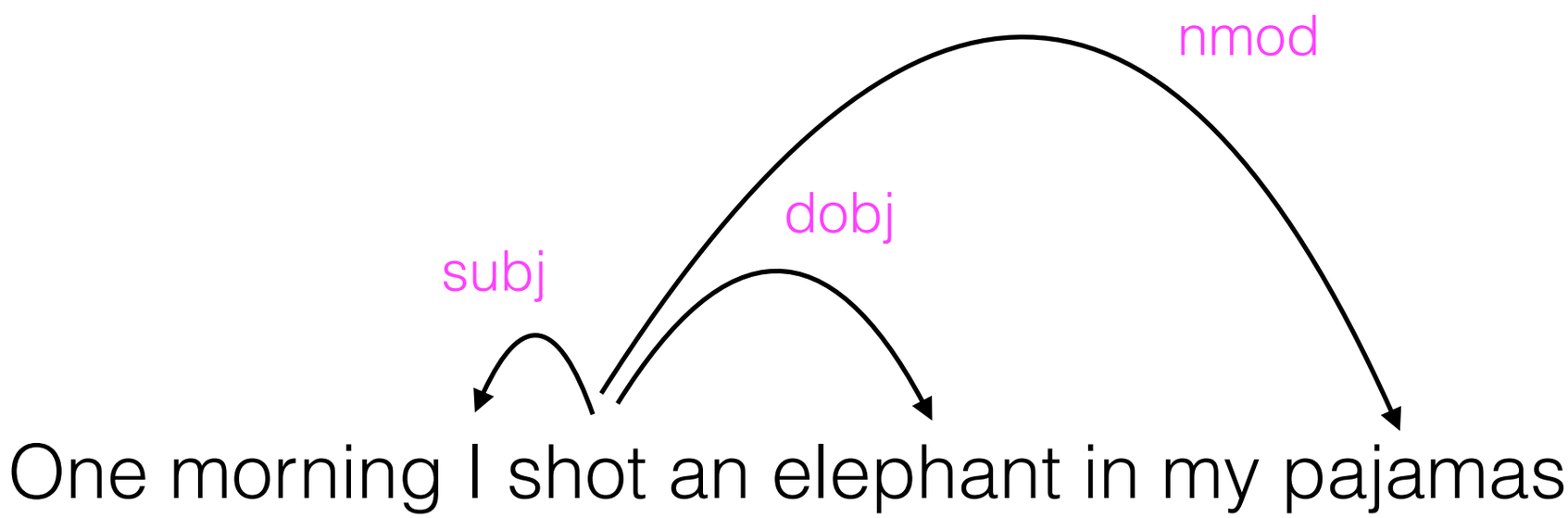
One morning I shot an elephant in my pajamas

Named entities

person

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

Syntax



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 constitutational law; at the University of Chicago Law School from 1992 to 2004.

Barack Obama

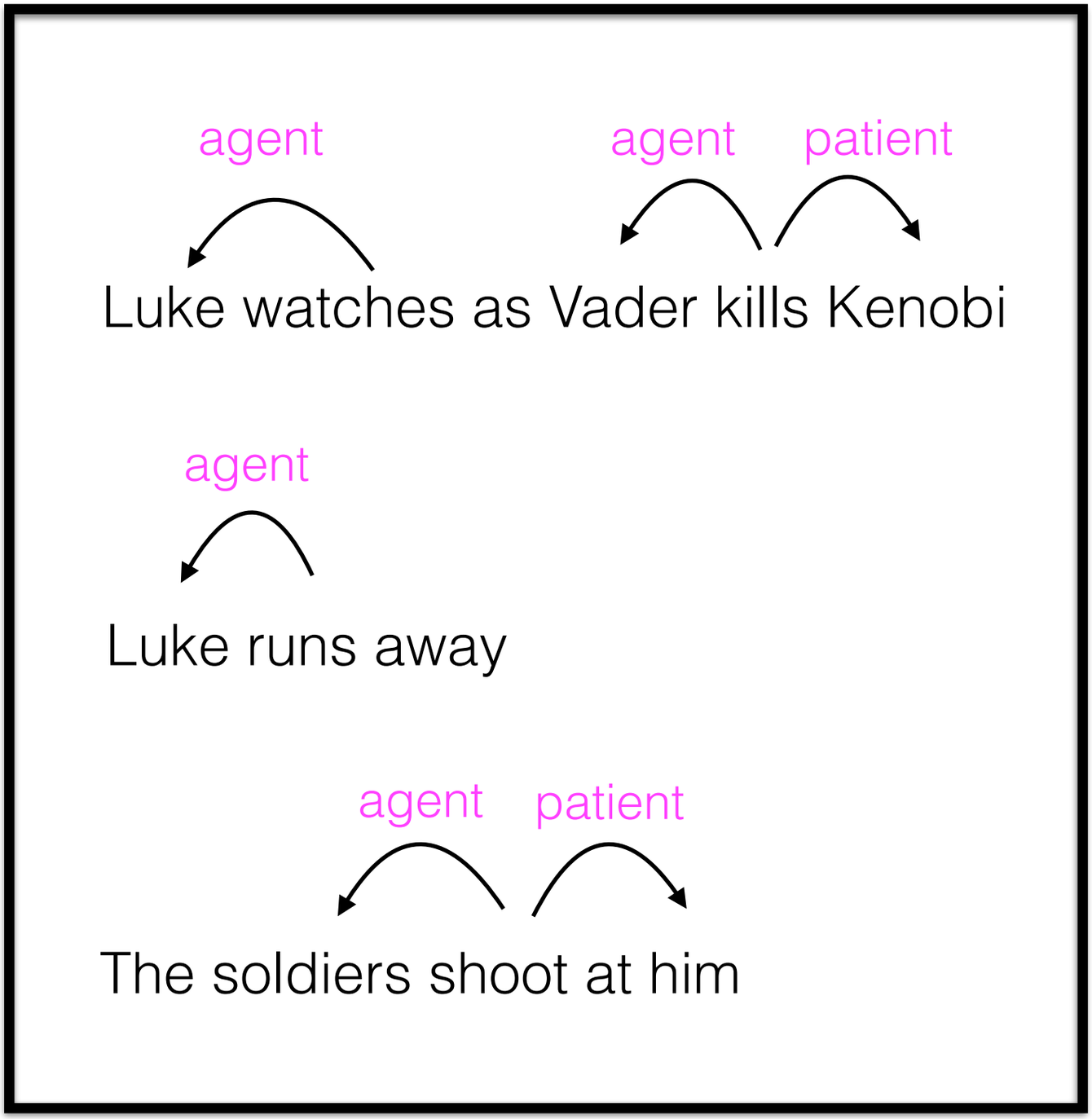
44th President of the United States

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

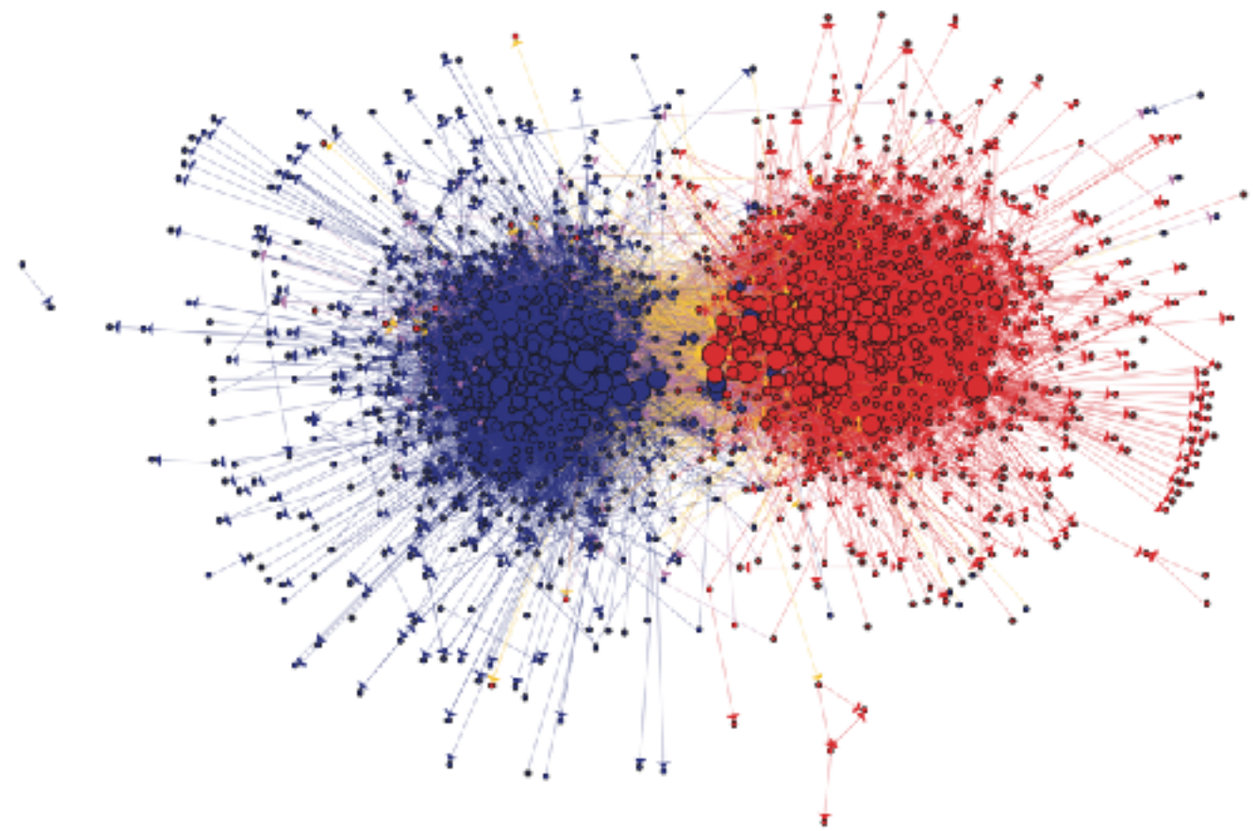
Google

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

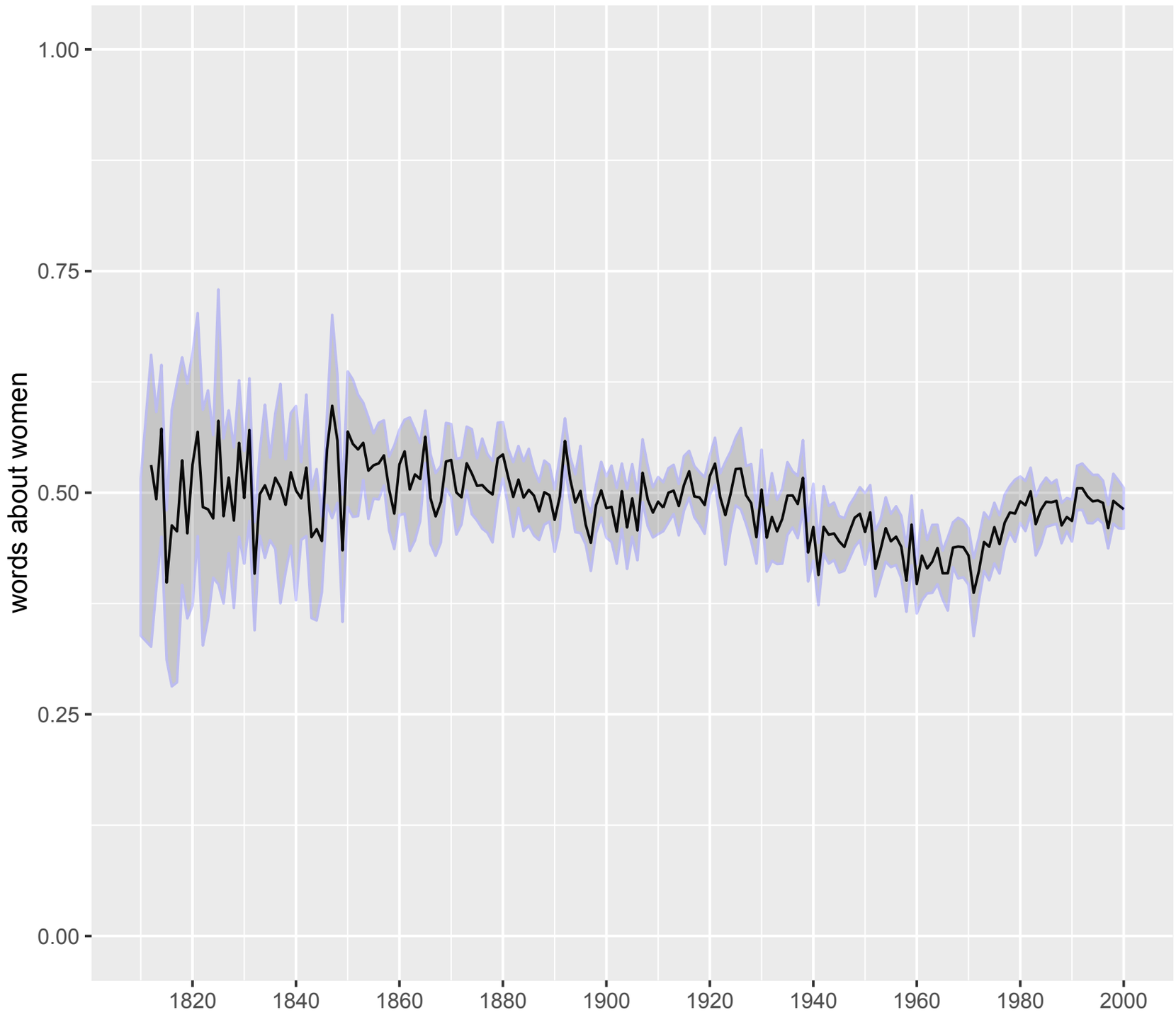
Holst Katsma (2014), Loudness in the Novel

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

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

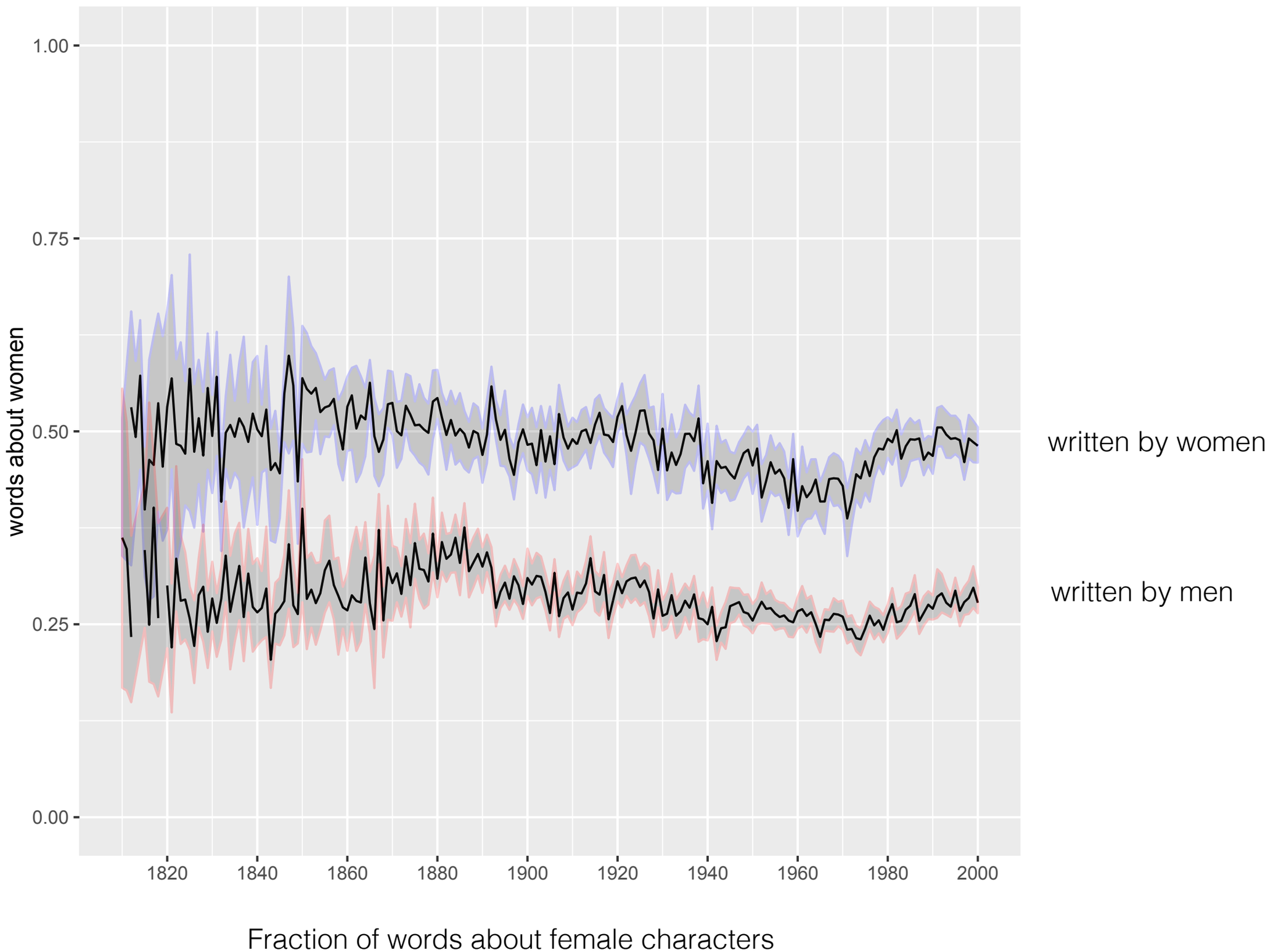
Jockers and Mimno (2013), “Significant Themes in 19th-Century Literature,”

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



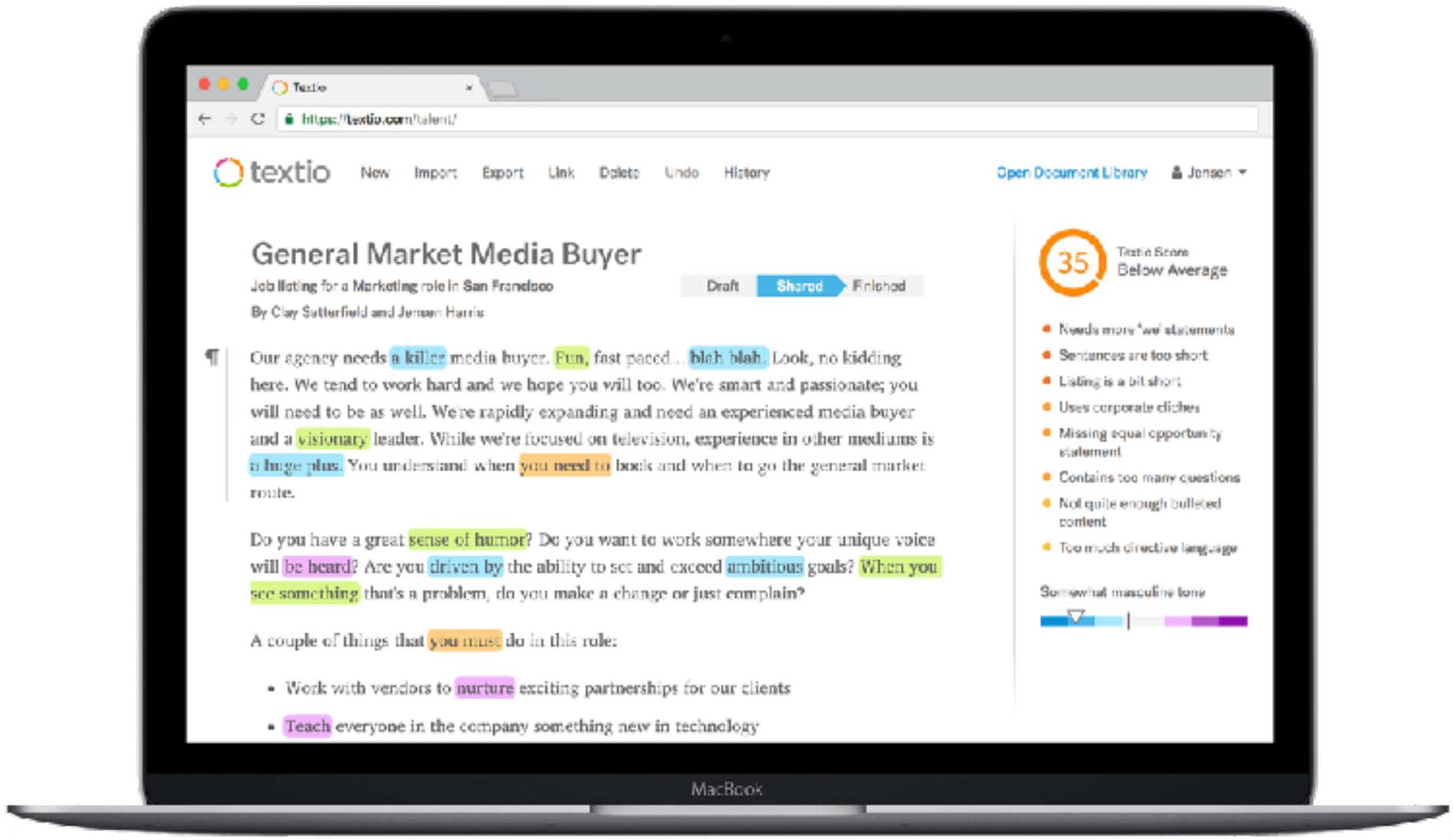
written by women

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

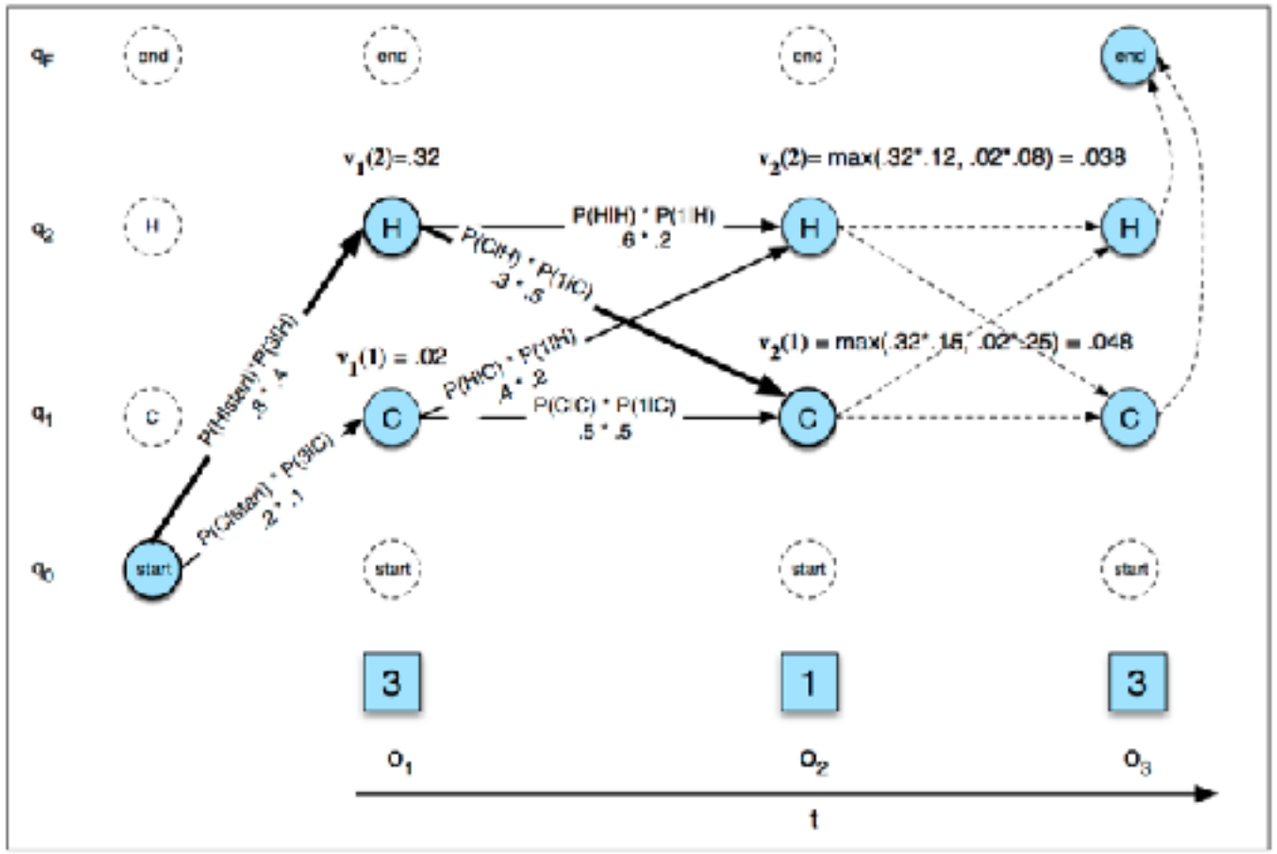
- Probabilistic models
- Naive Bayes, Logistic regression, HMM, MEMM, CRF, language models

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

Methods

- Dynamic programming (combining solutions to subproblems)

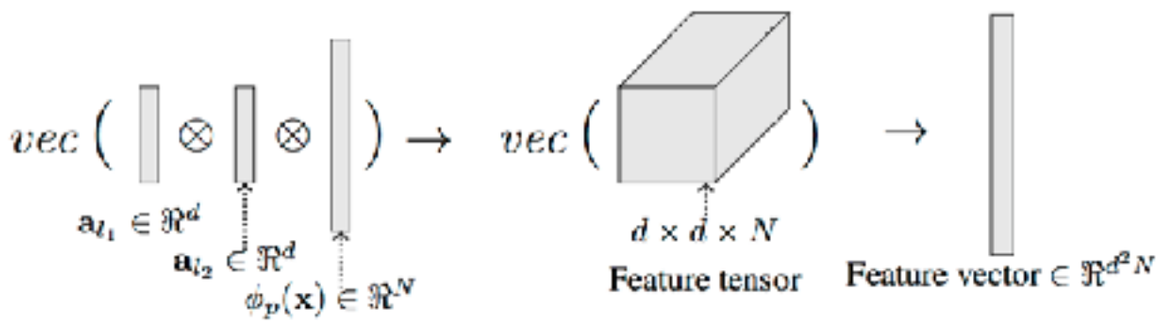
Viterbi algorithm, CKY



Viterbi lattice, SLP3 ch. 9

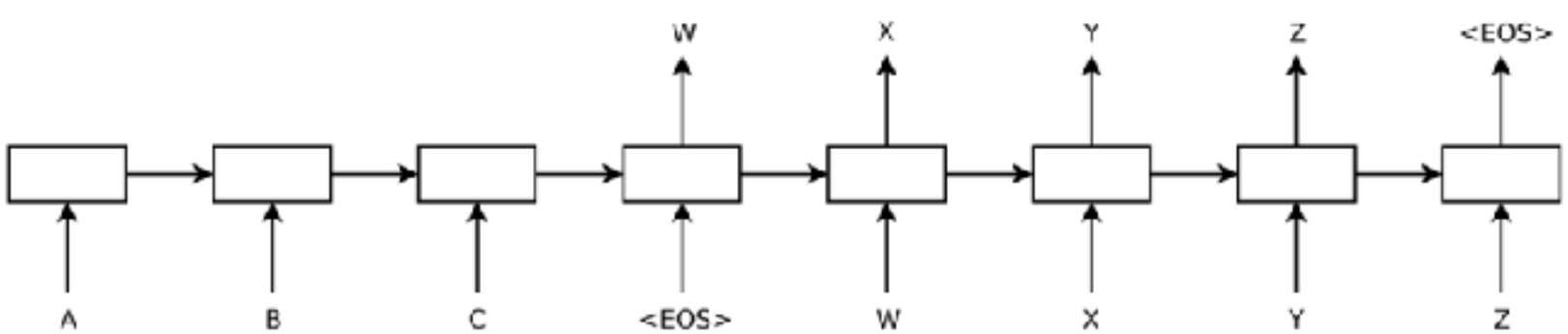
Methods

- 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”

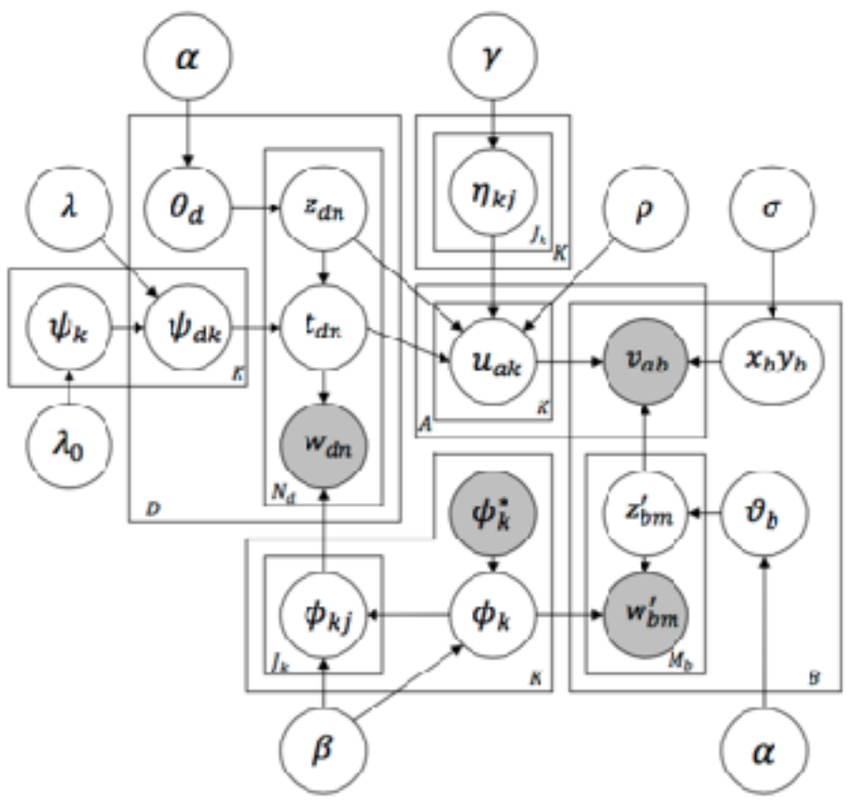


Figure 1: Plate notation diagram of HIPTM.

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] ← a_{0,s} * b_s(o_1)
    backpointer[s,1] ← 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] ← max_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)
    backpointer[s,t] ← argmax_{s'=1}^{N} viterbi[s',t-1] * a_{s',s}
viterbi[q_F,T] ← max_{s=1}^{N} viterbi[s,T] * a_{s,q_F}          ; termination step
backpointer[q_F,T] ← 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 (20%)
  - 7 short homeworks (30%)
  - 4 long homeworks (30%)

Homeworks

- Long homeworks: Modeling/algorithm exercises (derive the backprop updates for a CNN and implement it).
- Short homeworks: More frequent opportunities to get your hands dirty working with the concepts we discuss in class.

Late submissions

- All homeworks are due on the date/time specified.
- You have 2 late days total over the semester to use when turning in long/short homeworks; each day extends the deadline by 24 hours.
- You can drop 1 short homework.

Participation

- Participation can help boost your grade above a threshold (e.g., B+ →A-).
- Forms of participation:
    - Discussion in class
    - Answering questions on Piazza

Grading

- Info 259:
  - Midterm (20%) + project (30%)
  - 7 short homeworks (25%)
  - 4 long homeworks (25%)

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
    - 8-page final report, workshop quality
    - Poster presentation

ACL 2018 workshops

- Natural Language Processing Techniques for Educational Applications (NLPTEA)
- Computational Approaches to Linguistic Code-Switching (CALCS)
- Machine Reading for Question Answering (MRQA)
- Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP)
- Economics and Natural Language Processing (ECONLP)
- Representation Learning for NLP (RepL4NLP)
- Natural Language Processing for Social Media (SocialNLP)

Waitlisted

- Come to class, complete assignments

Applied NLP (Spring 2019)

- This course covers the algorithmic fundamentals of NLP to give you the core building blocks you need to innovate in NLP.

- Some graduate students may prefer my Applied NLP course in the spring; that covers the application of existing tools and methods (spacy, nltk, scikit-learn, tensorflow) for research involving text as data.

Next time

- Sentiment analysis and text classification
- Read SLP3 chapter 6 (on syllabus)
- DB office hours Wednesdays 10am-noon (314 South Hall)
- TAs:
    - Lara McConnaughey
    - Monik Pamecha
    - Brenton Chu