



Applied Natural Language Processing

Info 256

Lecture 1: Introduction (Jan 22, 2019)

David Bamman, UC Berkeley

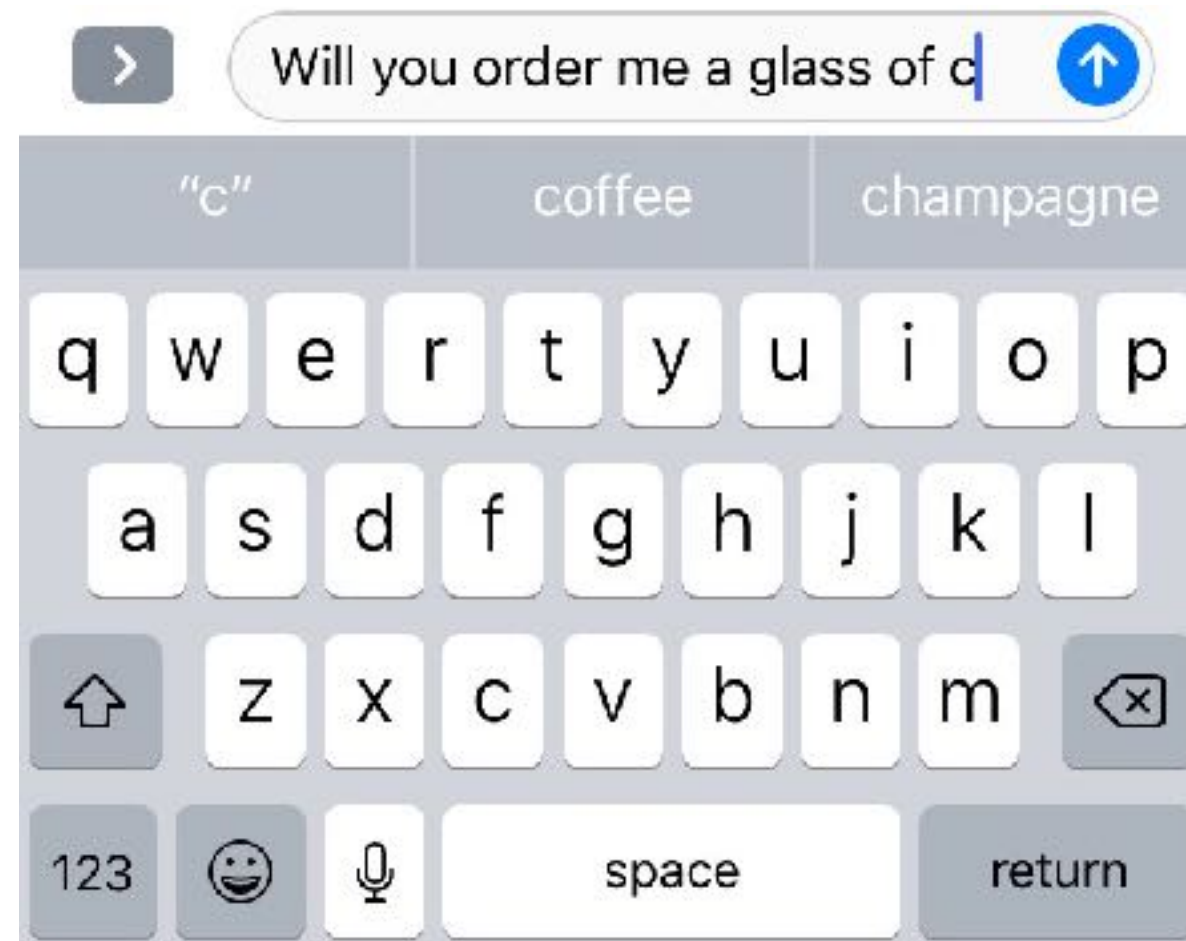
NLP



Google

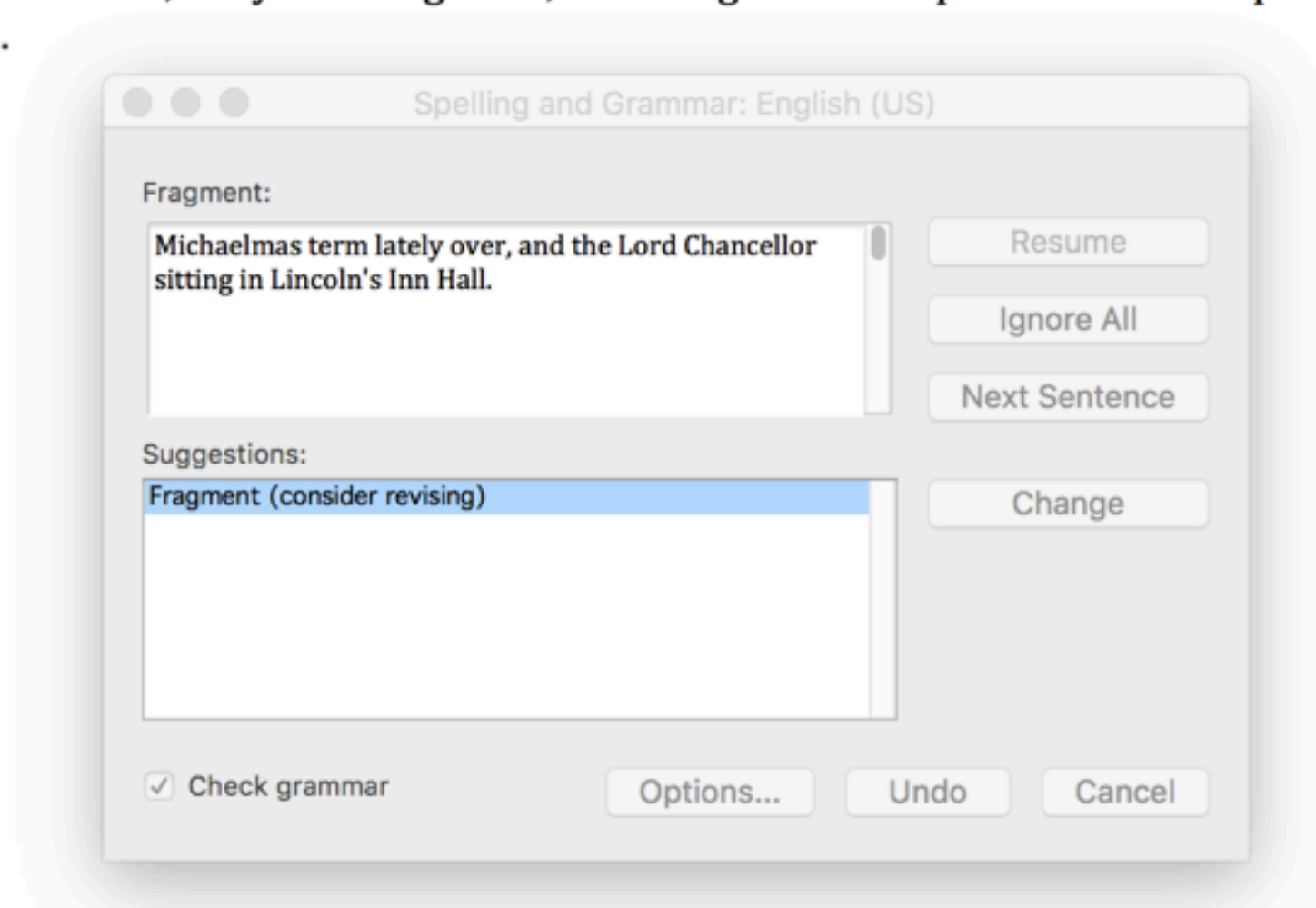


Predictive text messaging

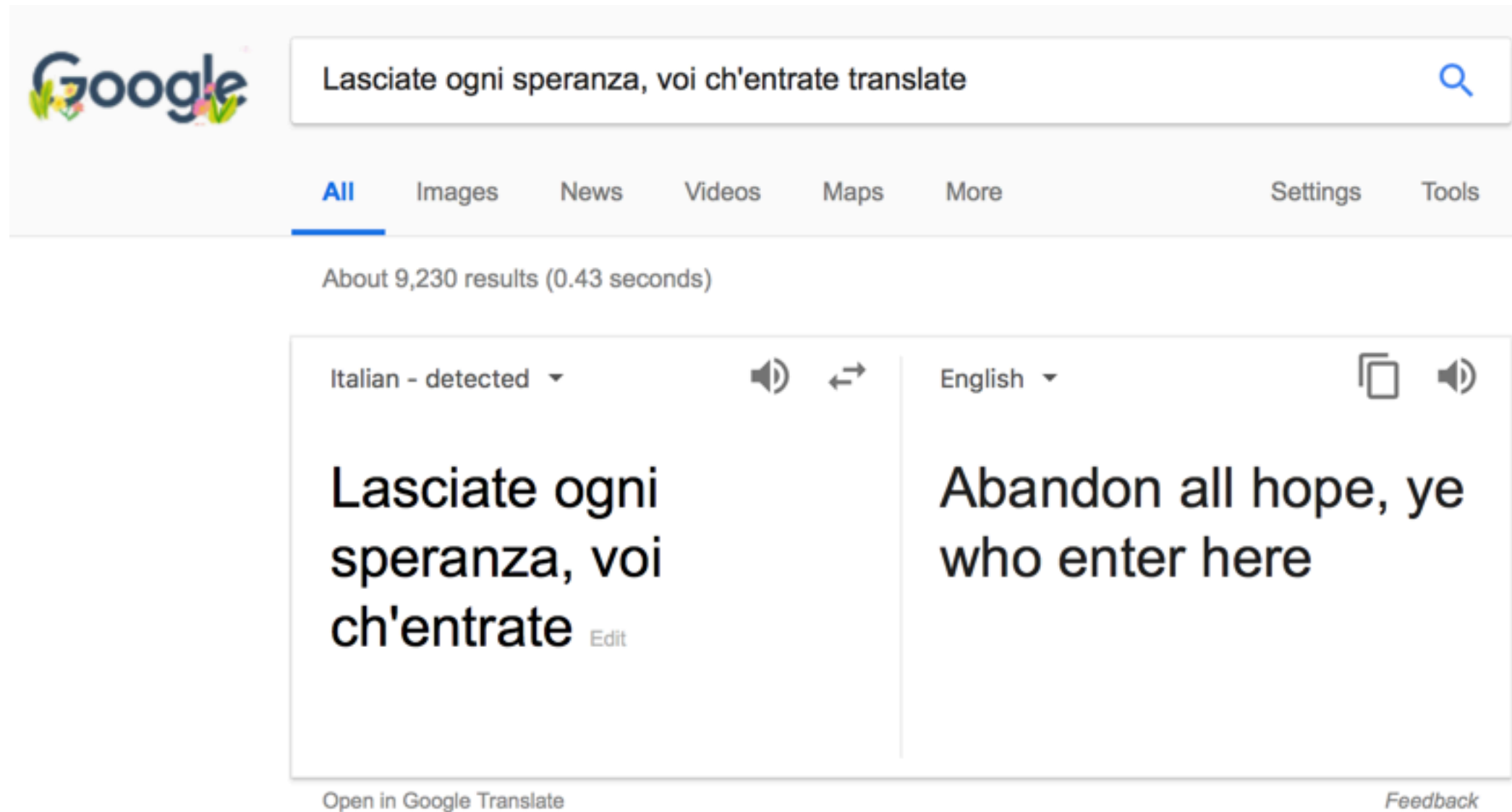


Grammar checking

London. Michaelmas term lately over, and the Lord Chancellor sitting in Lincoln's Inn Hall. Implacable November weather. As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.



Machine translation



The screenshot shows a Google search interface. The search bar contains the text "Lasciate ogni speranza, voi ch'entrate translate". Below the search bar, the "All" tab is selected. The search results show "About 9,230 results (0.43 seconds)". The main content area displays a machine translation from Italian to English. The Italian text is "Lasciate ogni speranza, voi ch'entrate" and the English translation is "Abandon all hope, ye who enter here".

Google

Lasciate ogni speranza, voi ch'entrate translate

All Images News Videos Maps More Settings Tools

About 9,230 results (0.43 seconds)

Italian - detected

English

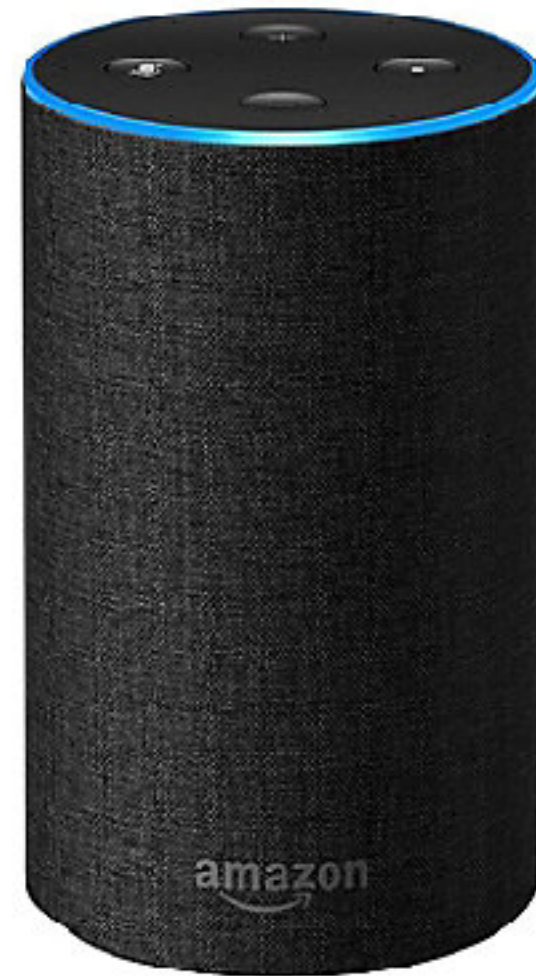
Lasciate ogni
speranza, voi
ch'entrate [Edit](#)

Abandon all hope, ye
who enter here

[Open in Google Translate](#) [Feedback](#)

Speech Recognition

“Alexa, how many cups are
in a quart?”



Question Answering



when was the last total eclipse in the united states



All

News

Images

Videos

Shopping

More

Settings

Tools

About 63,600,000 results (0.72 seconds)

August 21, 2017

See more photos of the **August 21** eclipse. Bottom line: After the **August 21, 2017**, eclipse, the next total solar eclipse visible from North America will be **April 8, 2024**. Jul 5, 2018



[When's the next total solar eclipse for North America? | Astronomy ...](https://earthsky.org/astronomy-essentials/whens-the-next-total-solar-eclipse-in-the-us)

<https://earthsky.org/astronomy-essentials/whens-the-next-total-solar-eclipse-in-the-us>

NLP

If you're interested in the core methods and algorithms, take [Info 159/259 \(NLP\)](#) instead.

- language modeling
- sequence labeling
- phrase-structure parsing
- dependency parsing
- dynamic programming
- MT

Applied NLP

How do we use the methodologies in NLP
toward some end?

Software/Libraries



spaCy

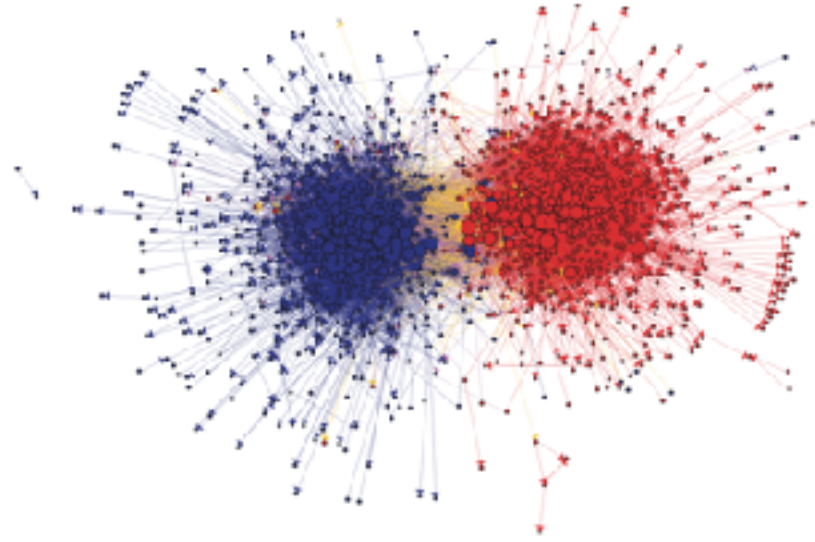


NLTK

NLP is interdisciplinary

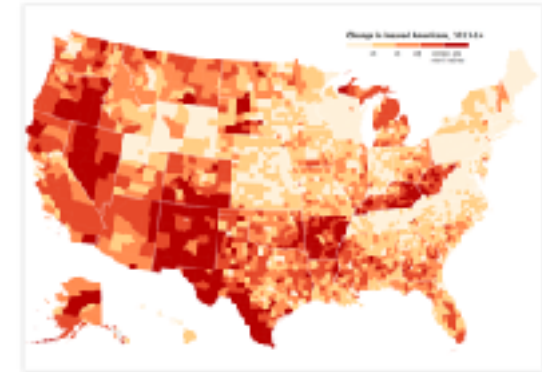
- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)

Computational Social Science



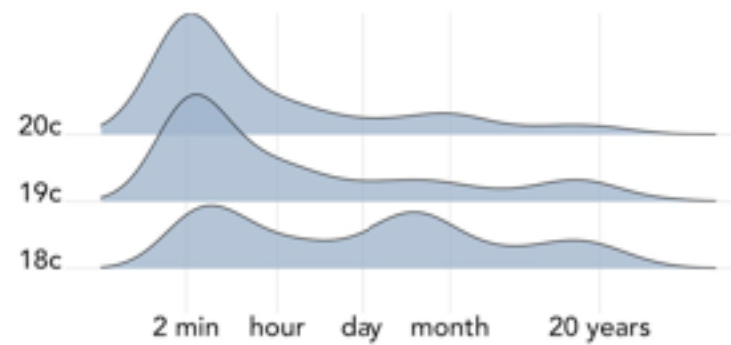
Adamic and Glance 2005

Computational Journalism



Change in insured Americans under the ACA,
NY Times (Oct 29, 2014)

Computational Humanities



Underwood 2018

Movie revenues

Input: text of movie review

Output: box office revenue

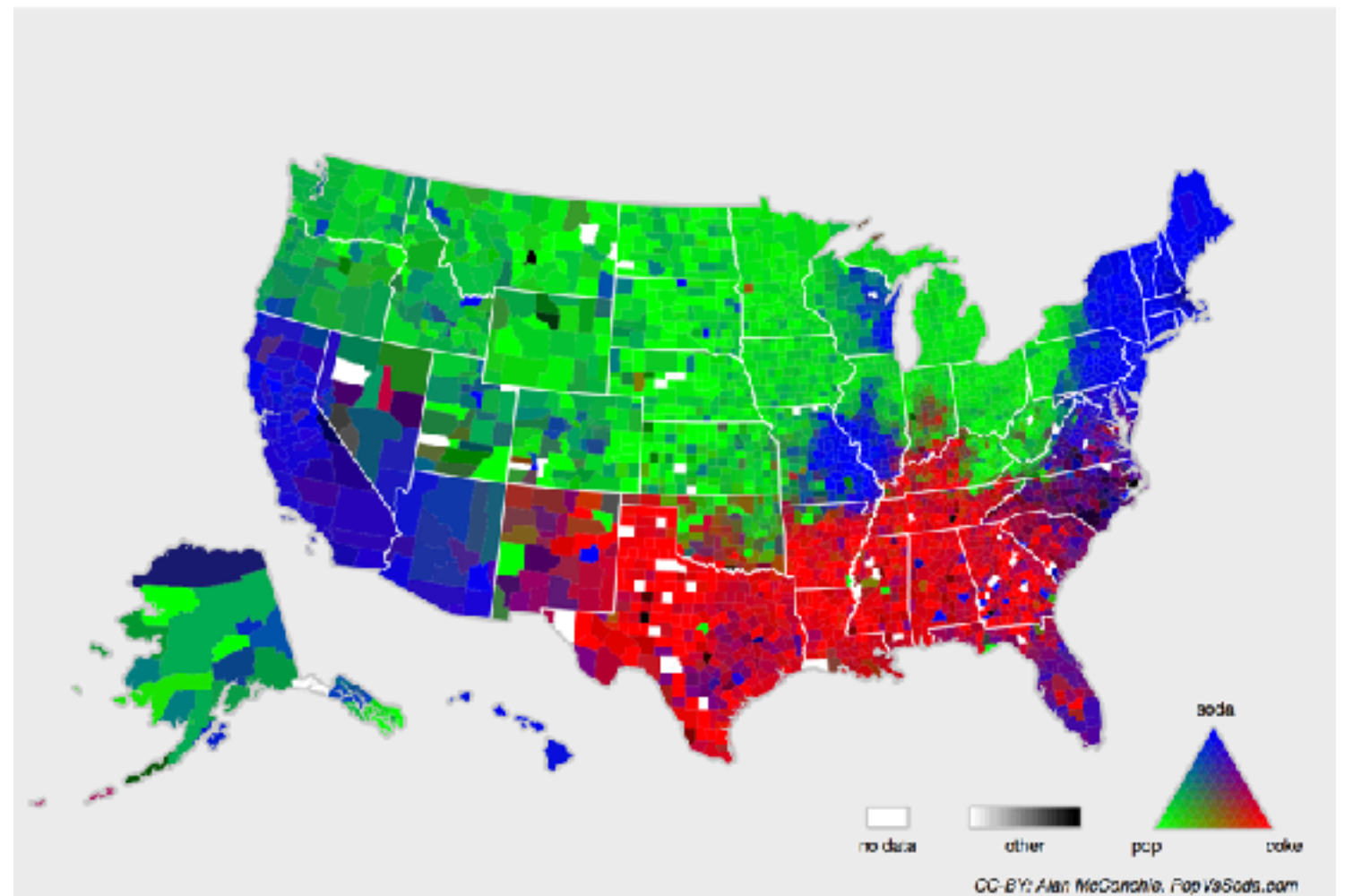


Geographical location

POP vs SODA

Input: tweet

Output: latitude,
longitude



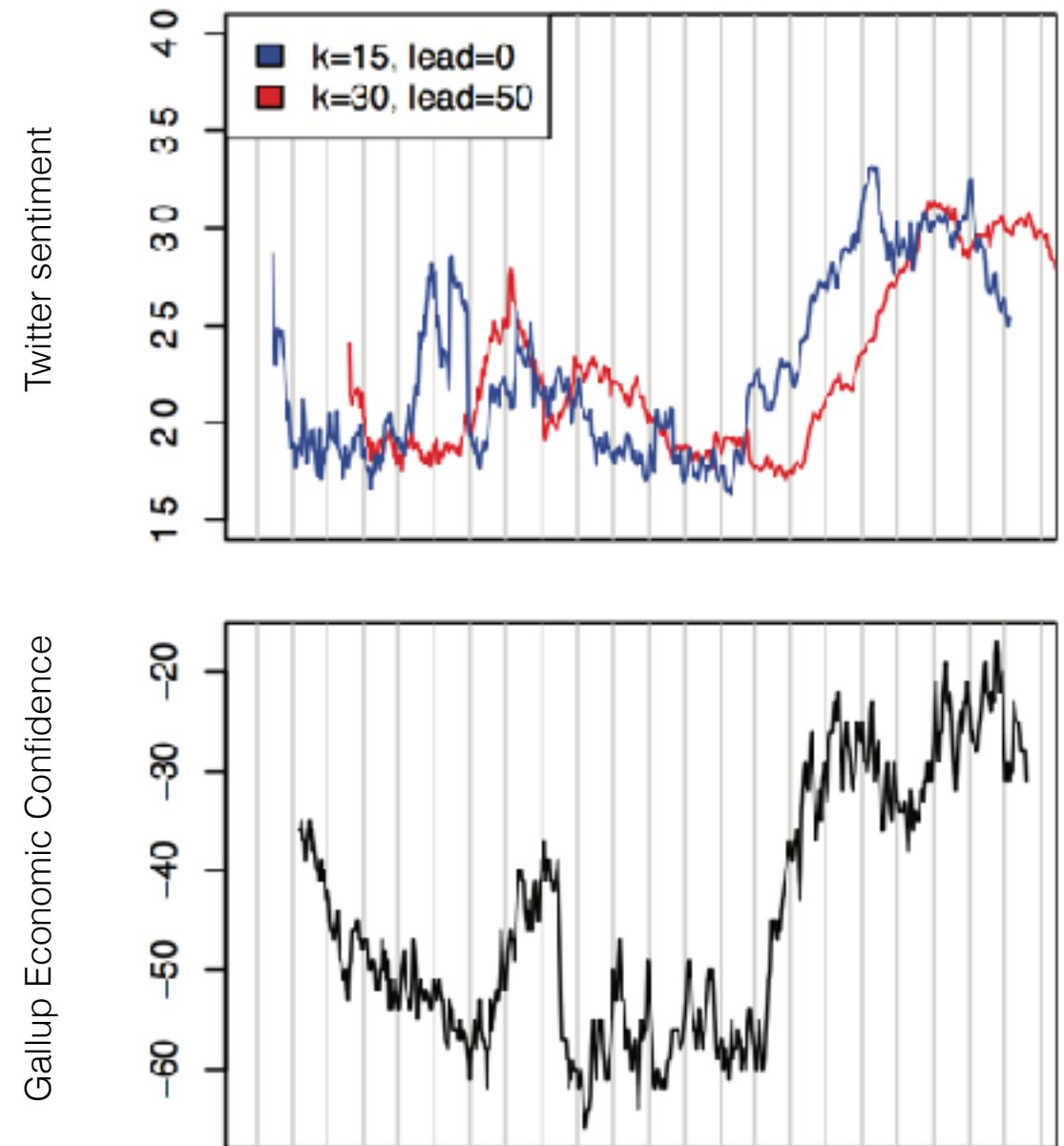
<http://popvssoda.com>

Wing and Baldrige (2011), "Simple supervised document geolocation with geodesic grids" (ACL)

Consumer sentiment

Input: tweets

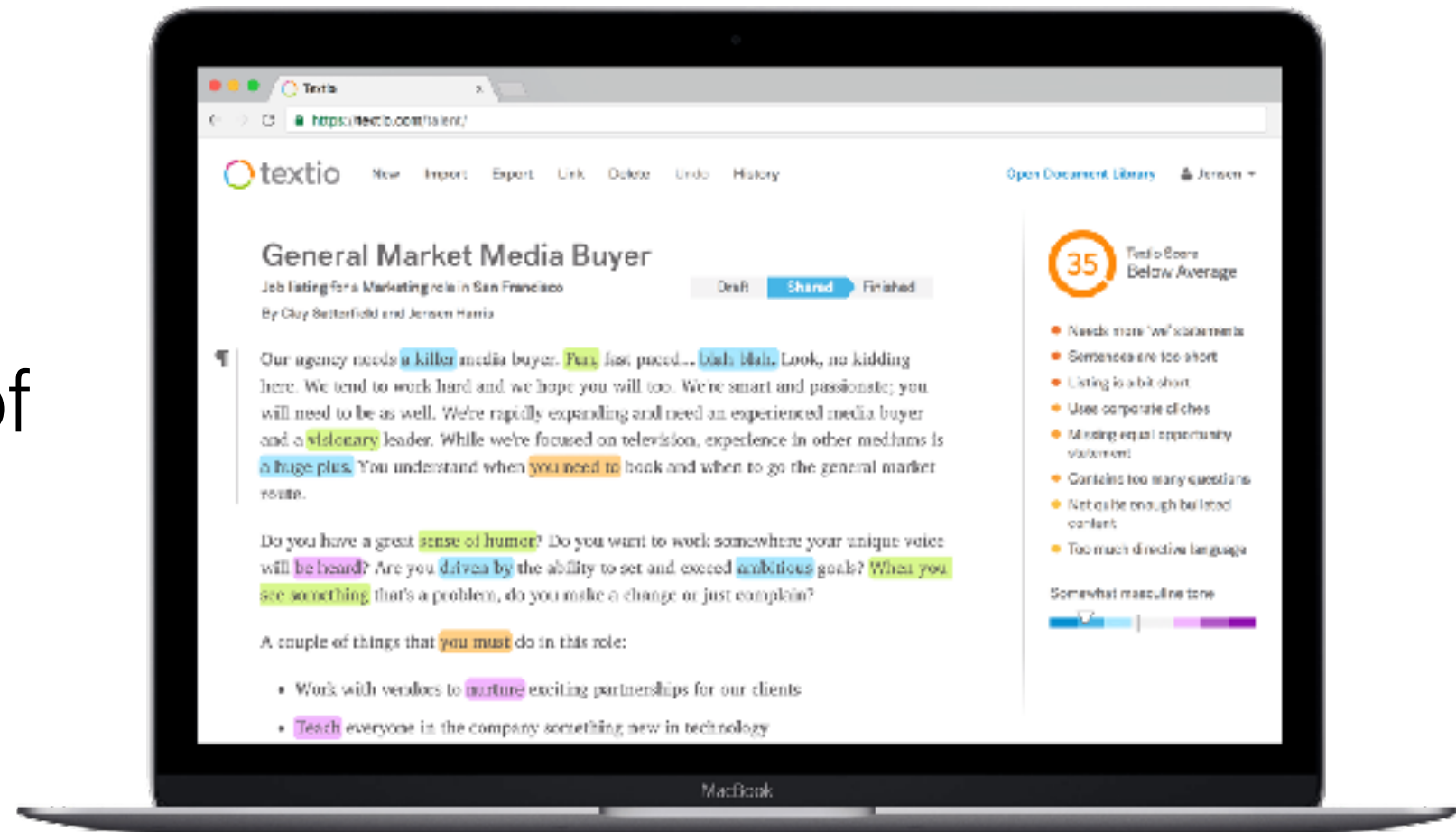
Output: Gallup
economic confidence
score



Hiring practices

Input: job ads

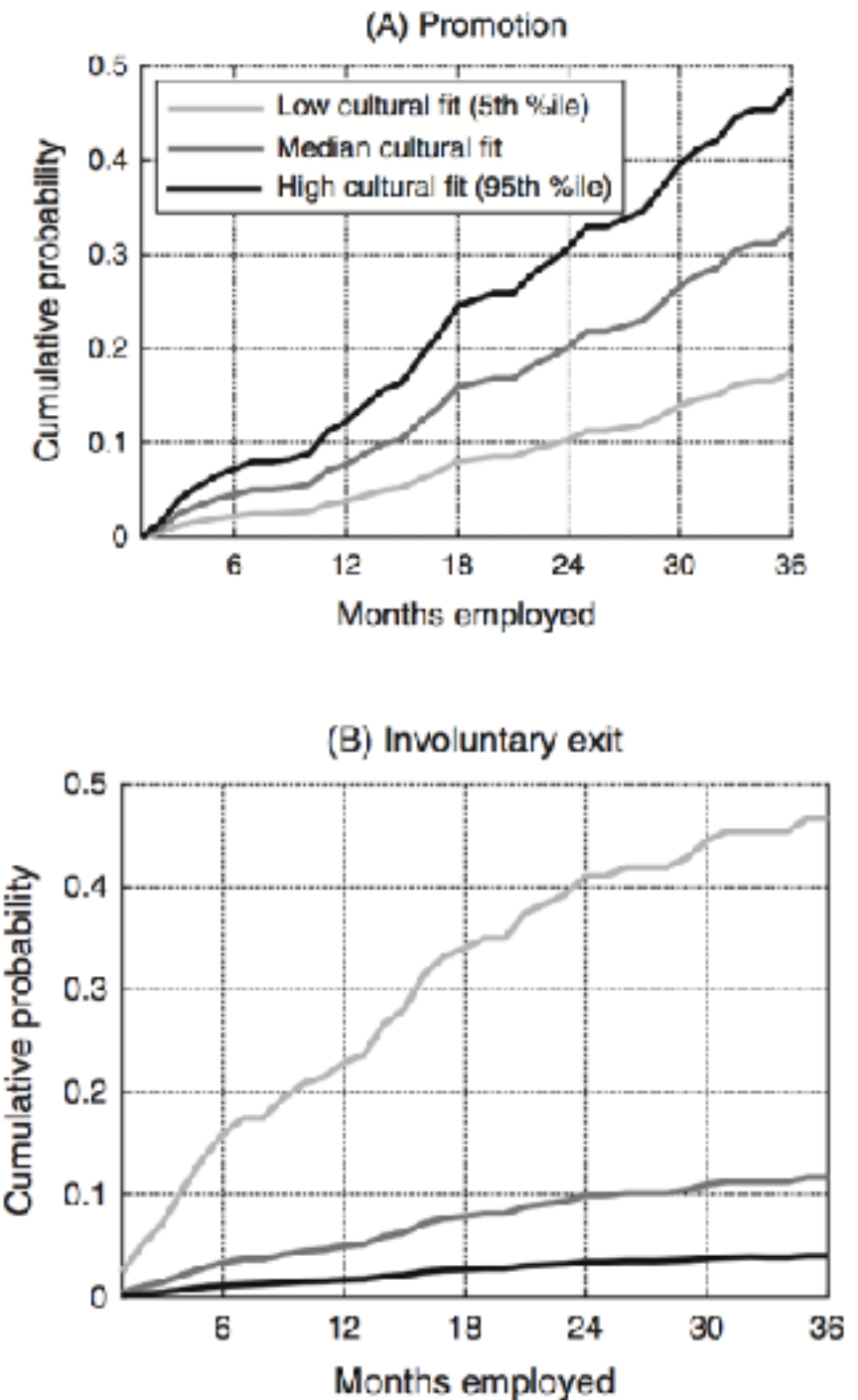
Output: gender ratio of applicants



Enculturation

Input: employee emails

Output: promotion to manager, time to separation



Srivastava et al. (2017), "Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations" (*Management Science*)



RANDOM ACTS OF PIZZA | READ THE RULES BEFORE POSTING.

HOT

NEW

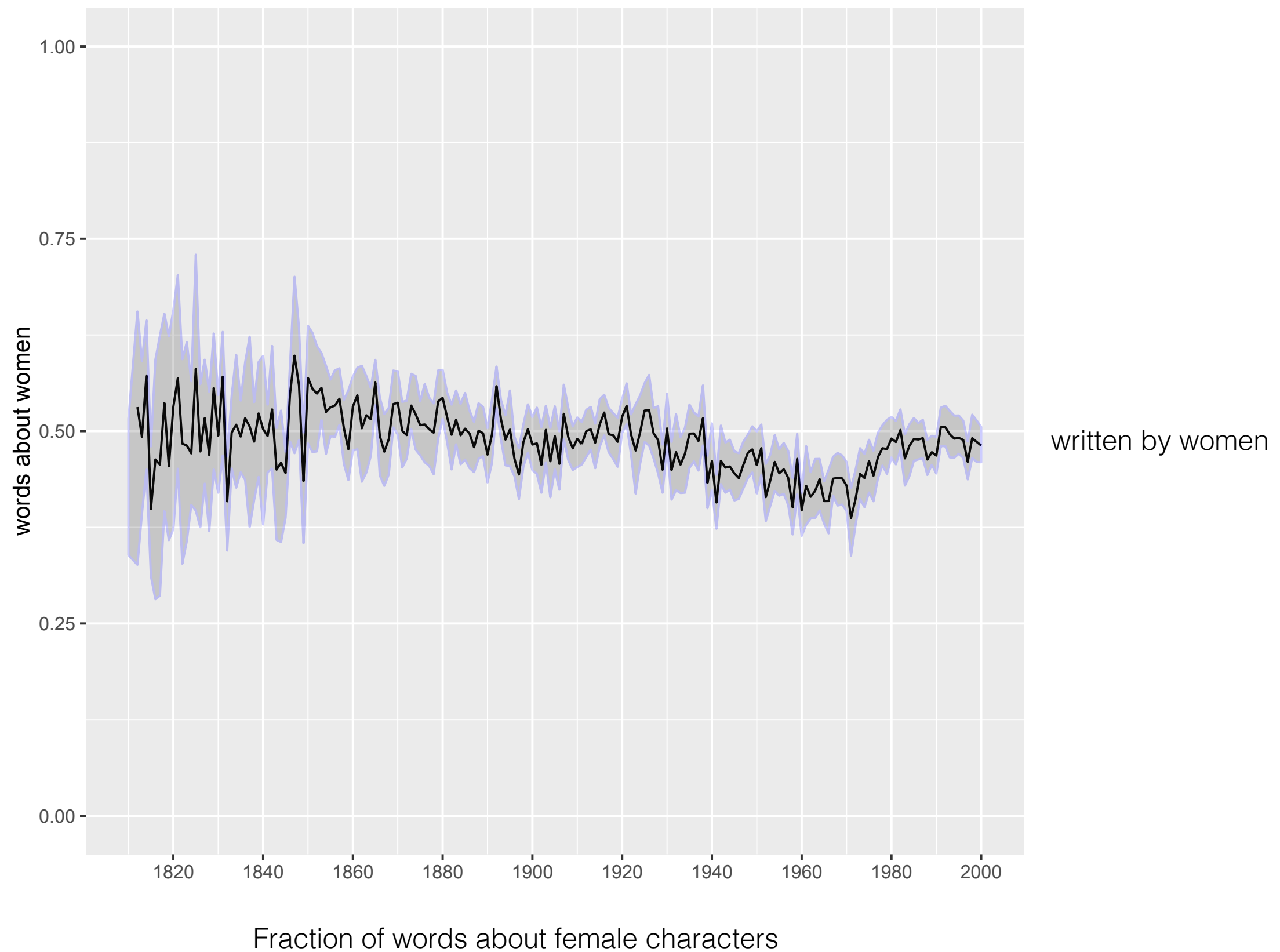
RISING

CONTROVERSIAL

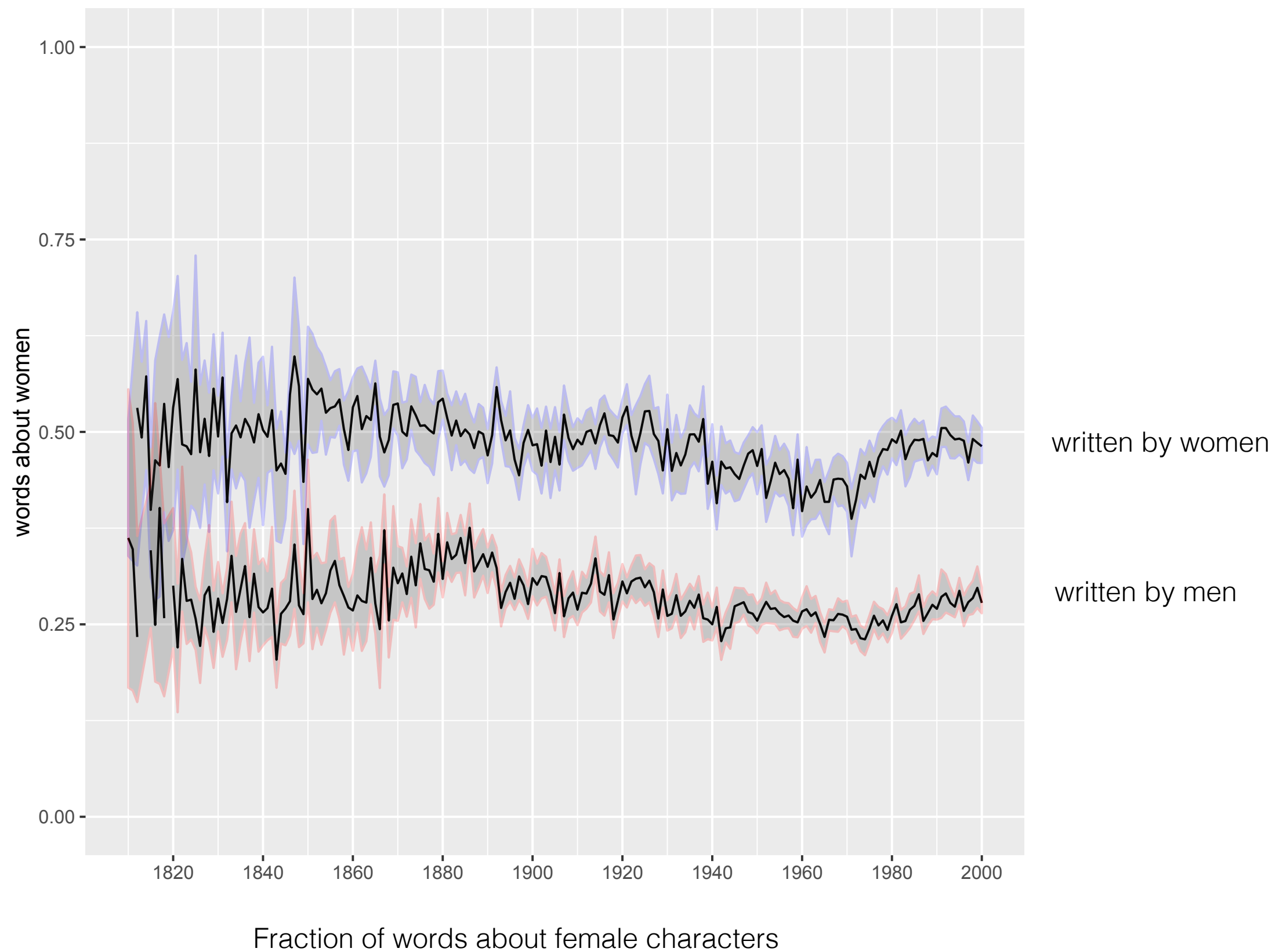
TOP

WIKI

- Data: Random acts of pizza (subreddit)
- Response: Is a request successful in getting a pizza?



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*)

Measurement

- This is fundamentally a problem of measurement: how do we design an algorithmic instrument that can transform a text into a quantity?

“TOM!” No answer. “TOM!” No answer. “What's gone with that boy, I wonder? You TOM!” No answer. The old lady pulled her spectacles down and looked over them about the room; then she put them up and looked out under them. She seldom or never looked *through* them for so small a thing as a boy; they were her state pair, the pride of her heart, and were built for “style,” not service--she could have seen through a pair of stove-lids just as well. She looked perplexed for a moment, and then said, not fiercely, but still loud enough for the furniture to hear: “Well, I lay if I get hold of you I'll--” She did not finish, for by this time she was bending down and punching under the bed with the broom, and so she needed breath to punctuate the protest. She corrected nothing but the cat. “I never did see the beat of that boy!” She went to the open door and stood in it and looked out among the tomato vines and “jimpson” weeds that constituted the garden. No Tom. So she lifted up her voice at an angle calculated for distance and shouted: “Y-o-u-u TOM!” There was a slight noise behind her and she turned just in time to seize a small boy by the slack of his roundabout and arrest his flight. “There! I might 'a' thought of that closet. What you been doing in there?” “Nothing.” “Nothing! Look at your hands. And look at your mouth. What *is* that truck?” “I don't know, aunt.”

0.53



The Adventures of Tom Sawyer

MARK TWAIN

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Bag of Words

tom no answer tom no
answer what's gone with
that boy , I wonder ? you
tom ! no answer the old lady
pulled her spectacles down
and looked over them about
the room .



The Adventures of Tom Sawyer

MARK TWAIN



The Adventures of Tom Sawyer

MARK TWAIN

"TOM!"

No answer.

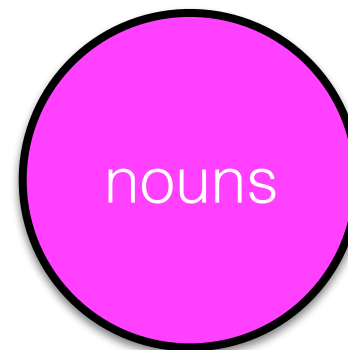
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



"TOM!"

No answer.

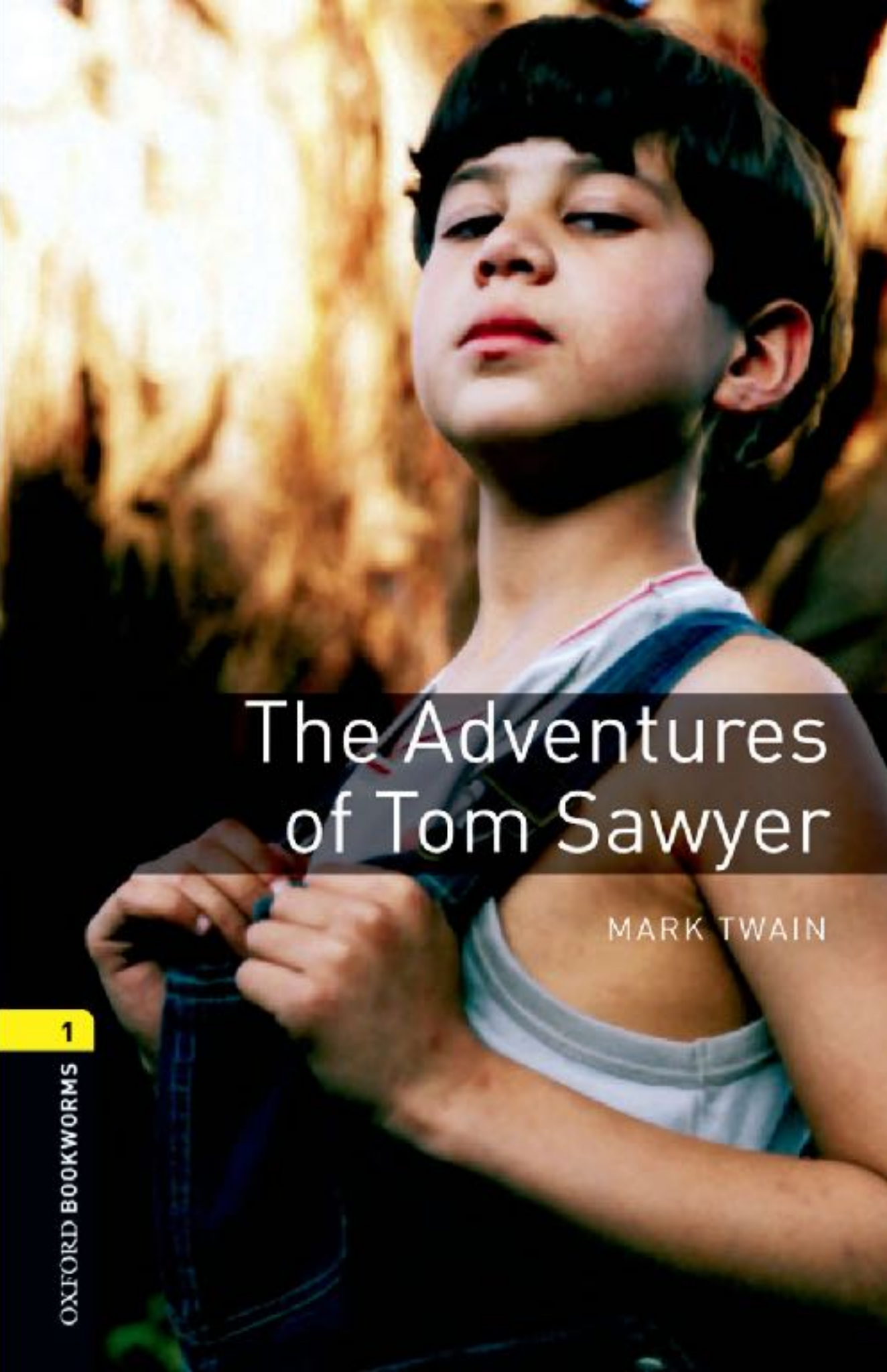
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

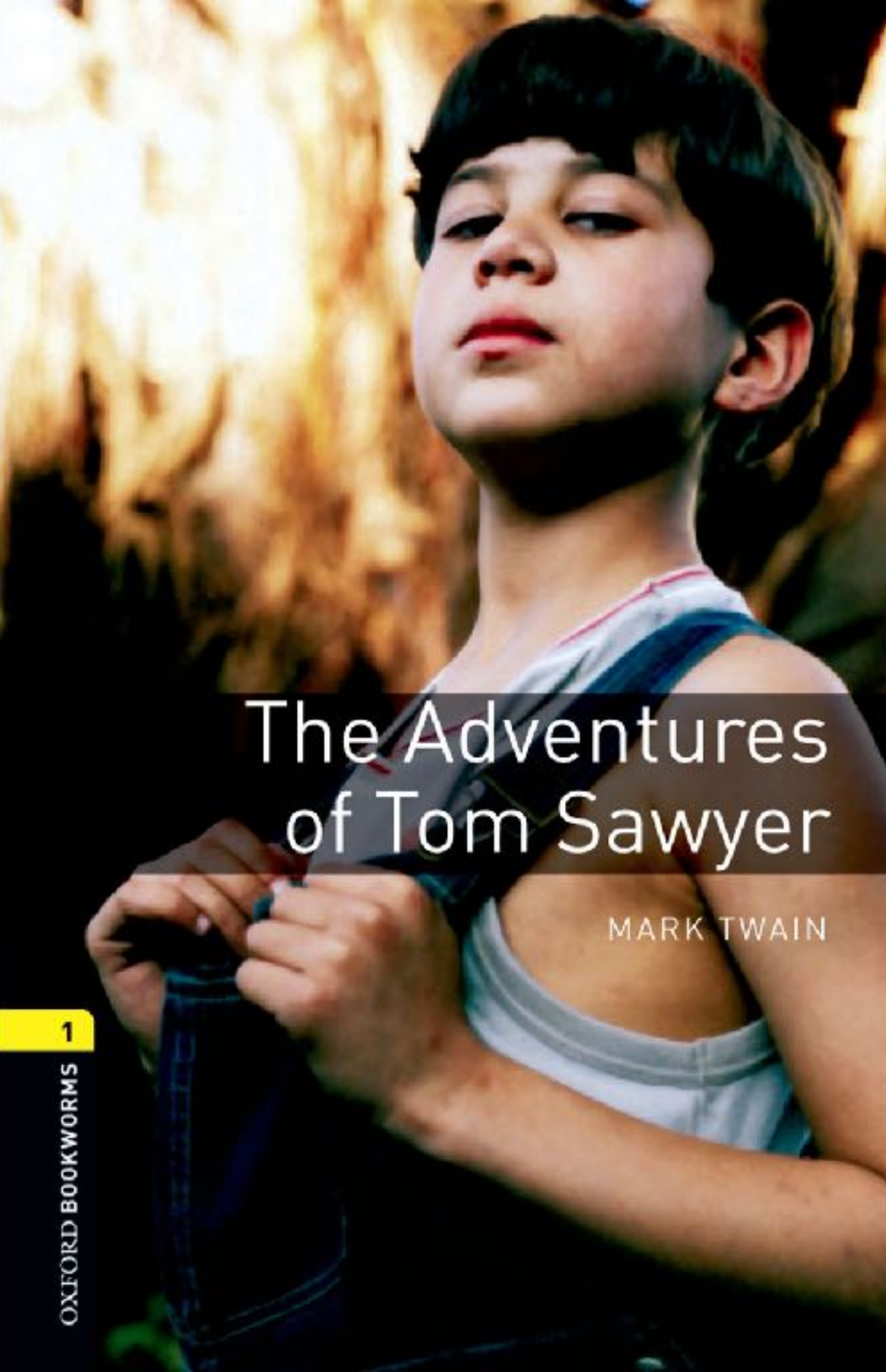
No answer.

The old lady pulled her spectacles down and looked over them about the room.



The Adventures of Tom Sawyer

MARK TWAIN



The Adventures of Tom Sawyer

MARK TWAIN



"TOM!"

No answer.

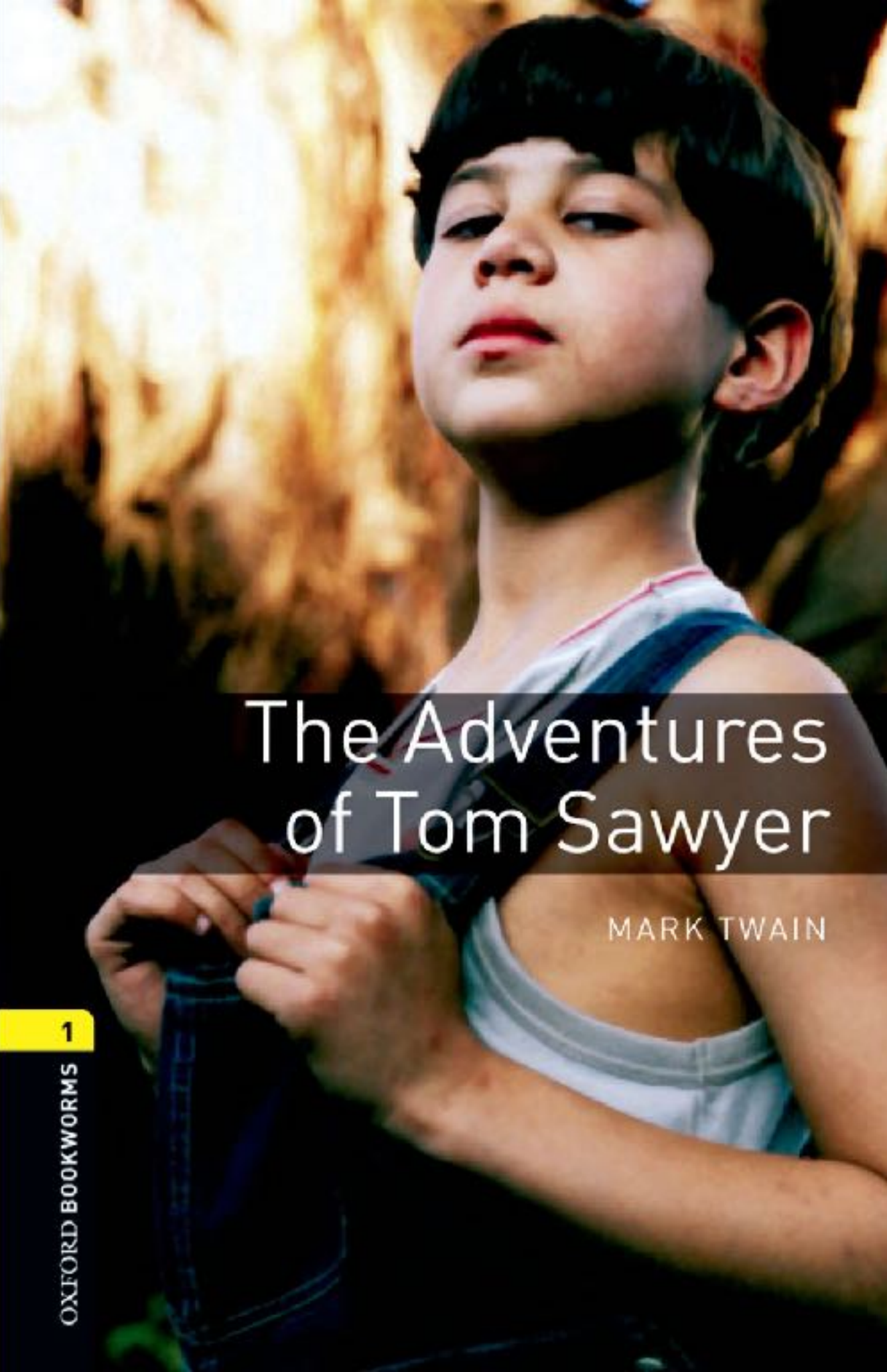
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



The Adventures of Tom Sawyer

MARK TWAIN

1

OXFORD BOOKWORMS

"TOM!"

No answer.

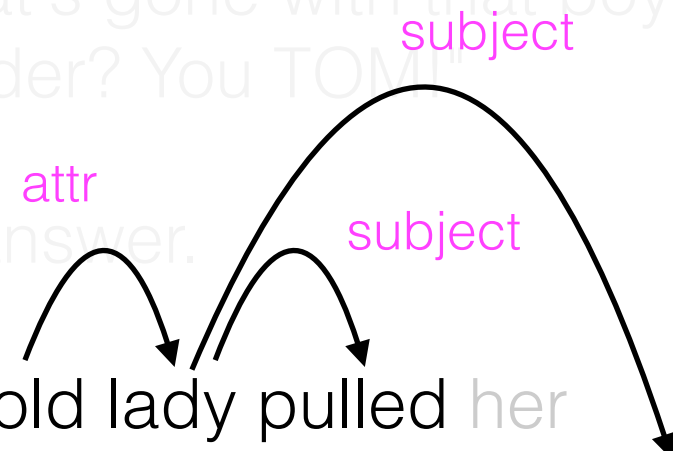
"TOM!"

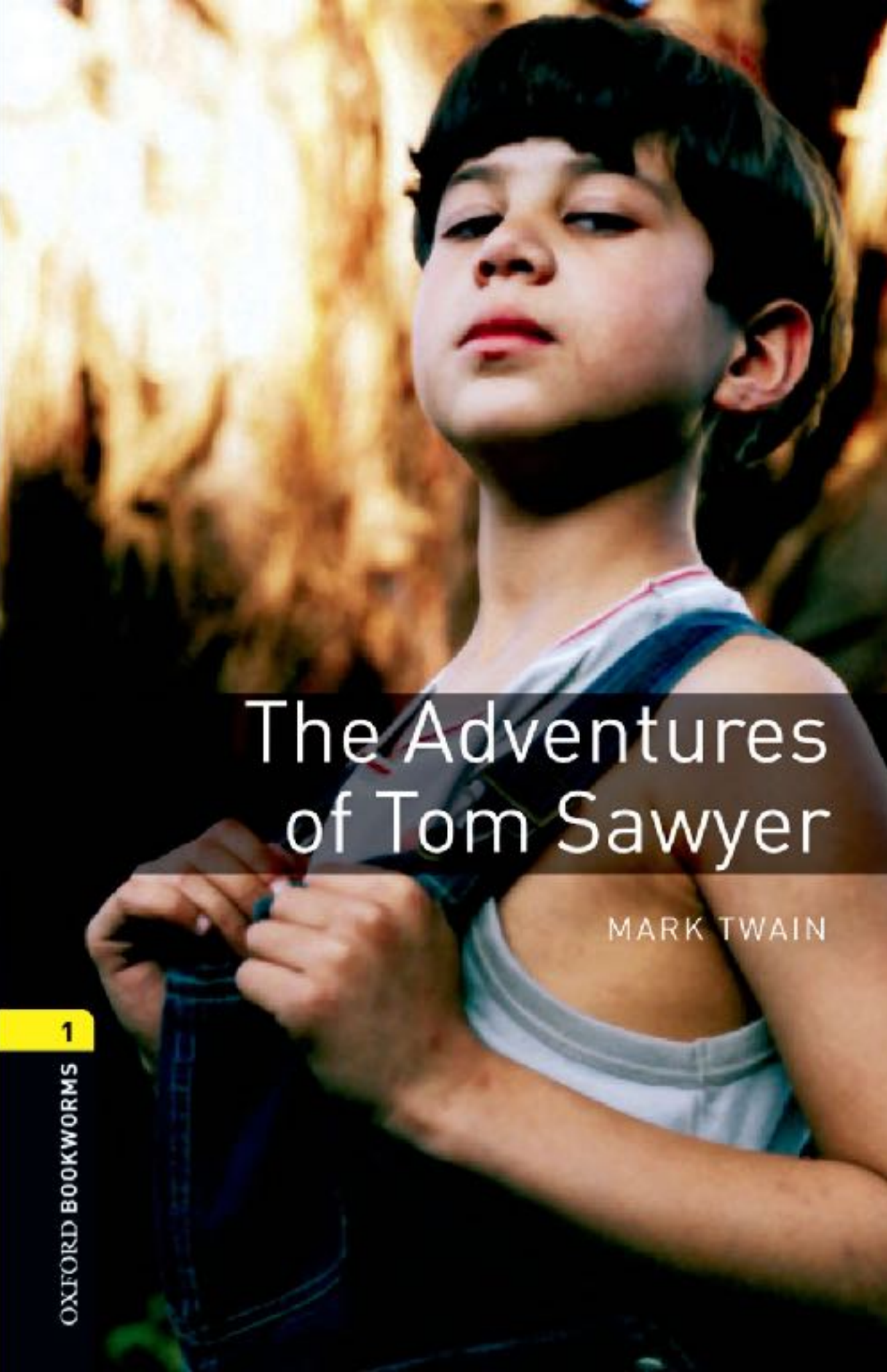
No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.





"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

```
graph LR; S1[The old lady] -- attr --> V1[pulled]; S1 -- agent --> V2[looked];
```


Temporal sequence

pulled her spectacles down



looked over them

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

agent

agent

Speaker identification



"TOM!"

No answer.

"TOM!"

No answer.



"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Coreference

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with *that boy*, I wonder? *You TOM!*"

No answer.

The old lady pulled *her* spectacles down and looked over them about the room.



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

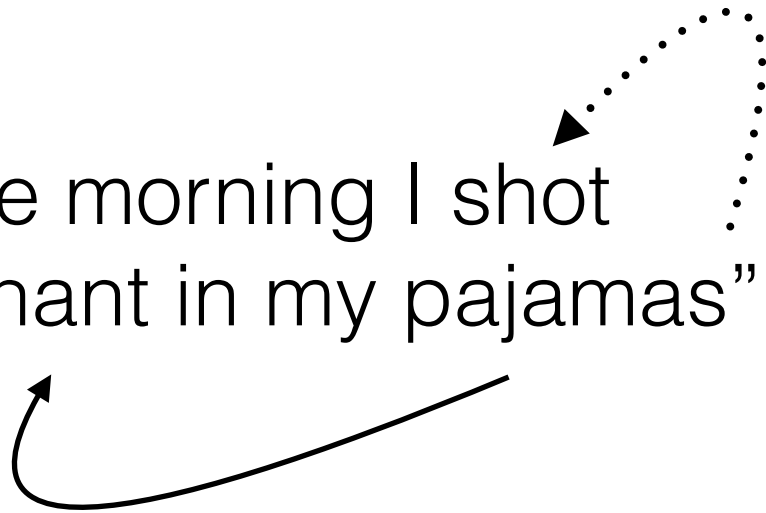
“One morning I shot
an elephant in my pajamas”



Animal Crackers

Ambiguity

“One morning I shot
an elephant in my pajamas”



Animal Crackers

Ambiguity

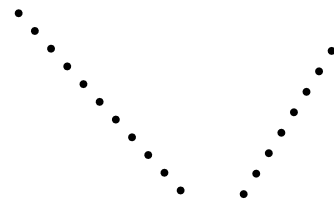


“One morning I shot
an elephant in my pajamas”



Ambiguity

verb noun



“One morning I shot
an elephant in my pajamas”



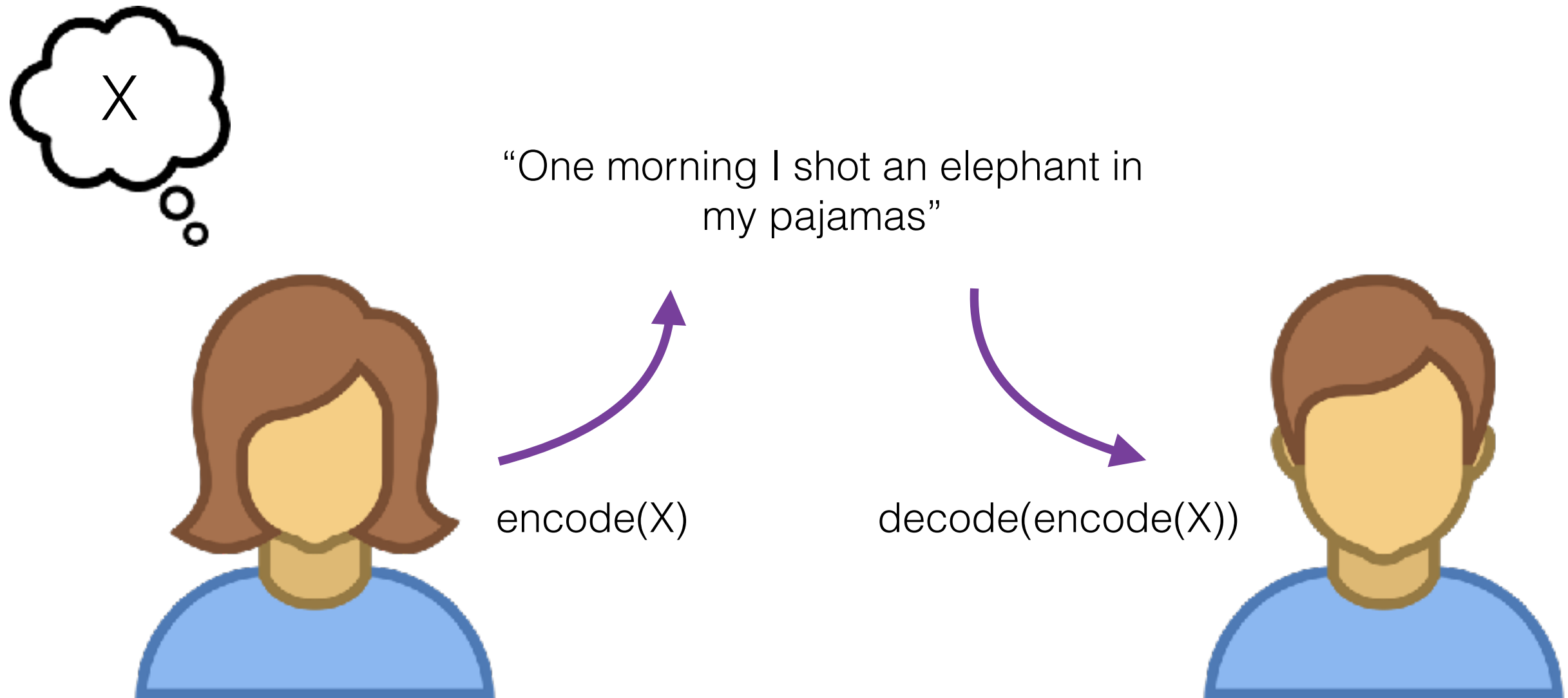
Animal Crackers

I made her duck

[SLP2 ch. 1]

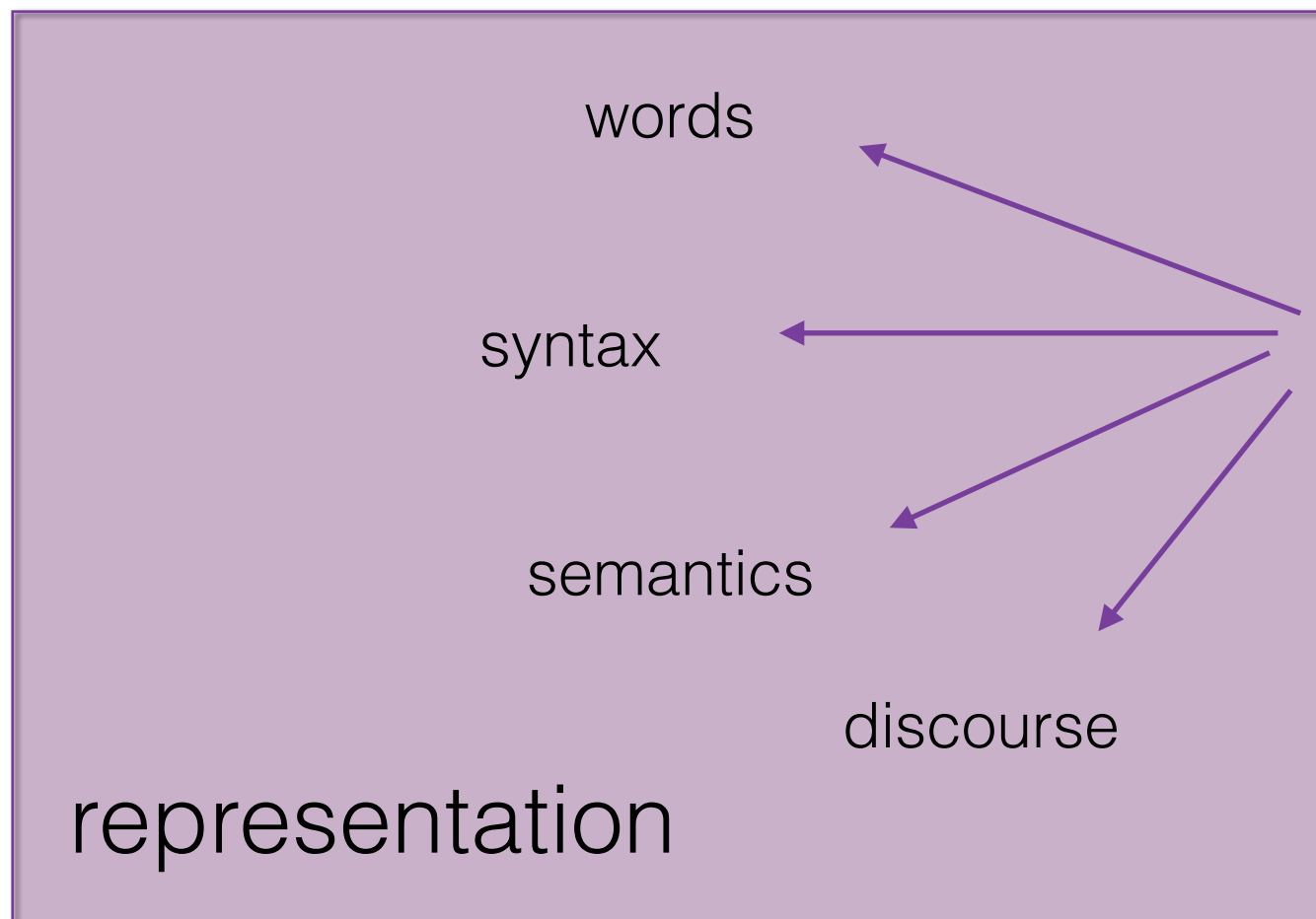
- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- ...

Information theoretic view

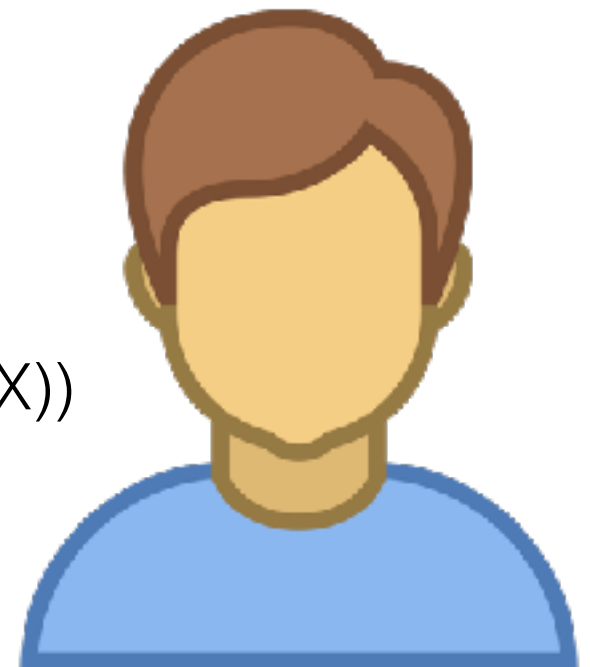


Decoding

“One morning I shot an elephant in
my pajamas”



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



“Raw” data

- We often want to make **claims** about the world using textual data.
- Data is not self-evident, neutral or objective
- Data is collected, stored, processed, mined, interpreted; each stage requires our **participation**.
- What is the **process** by which the data you have got to you?

Administrivia

- David Bamman
dbamman@berkeley.edu

Office hours: **Wednesdays 10am-noon**, 314 SH
— or by appointment

- Masha Belyi, TA
mashabelyi@berkeley.edu

Info 256

- Each class period will be divided between:
 - a short lecture; and
 - in-class lab work using Jupyter notebooks
- Students must prepare for each class and submit homeworks **before** class; attendance in class is required.

Grading

- Homeworks (40%)
- Participation (10%)
- Group project (50%)

Late submissions

- All homeworks are due on the date/time specified, before each class. We'll go over the homework in class, so **no late homeworks**.
- You can drop 2 homeworks.

Homeworks

- Homeworks will be frequent; you are free to discuss them at a high level with your classmates, but **all coding must be done individually.**
- If you use or build on others' code (e.g., from StackOverflow), you must cite its source.
- UC Berkeley code of conduct:
<http://sa.berkeley.edu/code-of-conduct>

Participation

- Participation includes:
 - Coming to class and working in groups (attendance is required!)
 - Peer assessment of homework and project deliverables.
 - Answering Piazza questions from your classmates

Course project

- Semester-long project (involving 1-3 students) , involving natural language processing in support of an empirical research question.
 - Project proposal/literature review
 - Midterm report
 - 8-page final report, workshop quality
 - Project presentation

ACL 2019 workshops

- BioNLP 2019
- BlackboxNLP 2019: Analyzing and interpreting neural networks for NLP
- The Thirteenth Linguistic Annotation Workshop (LAW XIII)
- The Third Workshop on Abusive Language Online
- Second Workshop on Storytelling (StoryNLP)
- Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)
- 1st International Workshop on Computational Approaches to Historical Language Change
- The 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)
- 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)
- Gender Bias in Natural Language Processing

Github

- Course homework will be on Github:
<https://github.com/dbamman/anlp19>
- Sign up for an account right now if you don't have one!

In class

- `anlp19/0.setup`
- Install anaconda environment + libraries we'll use frequently.