At the Intersection of Language and Data Science

Kathleen McKeown
Department of Computer Science
Columbia University

NEW MEDIA FOR DATA SCIENCE

Develop the tools and talent to enhance communication and interactions within communities

Vision

- Generating presentations that connect
  - Events
  - Opinions
  - Personal accounts
  - Their impact on the world

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What data is available for learning?

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What features yield good predictions?

FACT

News

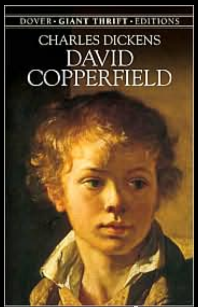
Wikipedia
descriptions

Personal
narrative

Social Media

FICTION

Novels



FACT

News

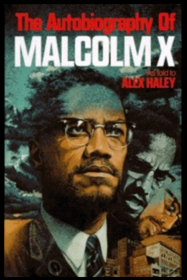
Wikipedia descriptions

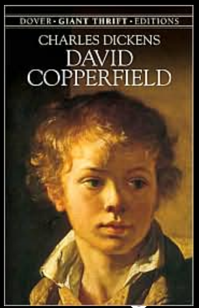
Personal narrative

Social Media

FICTION

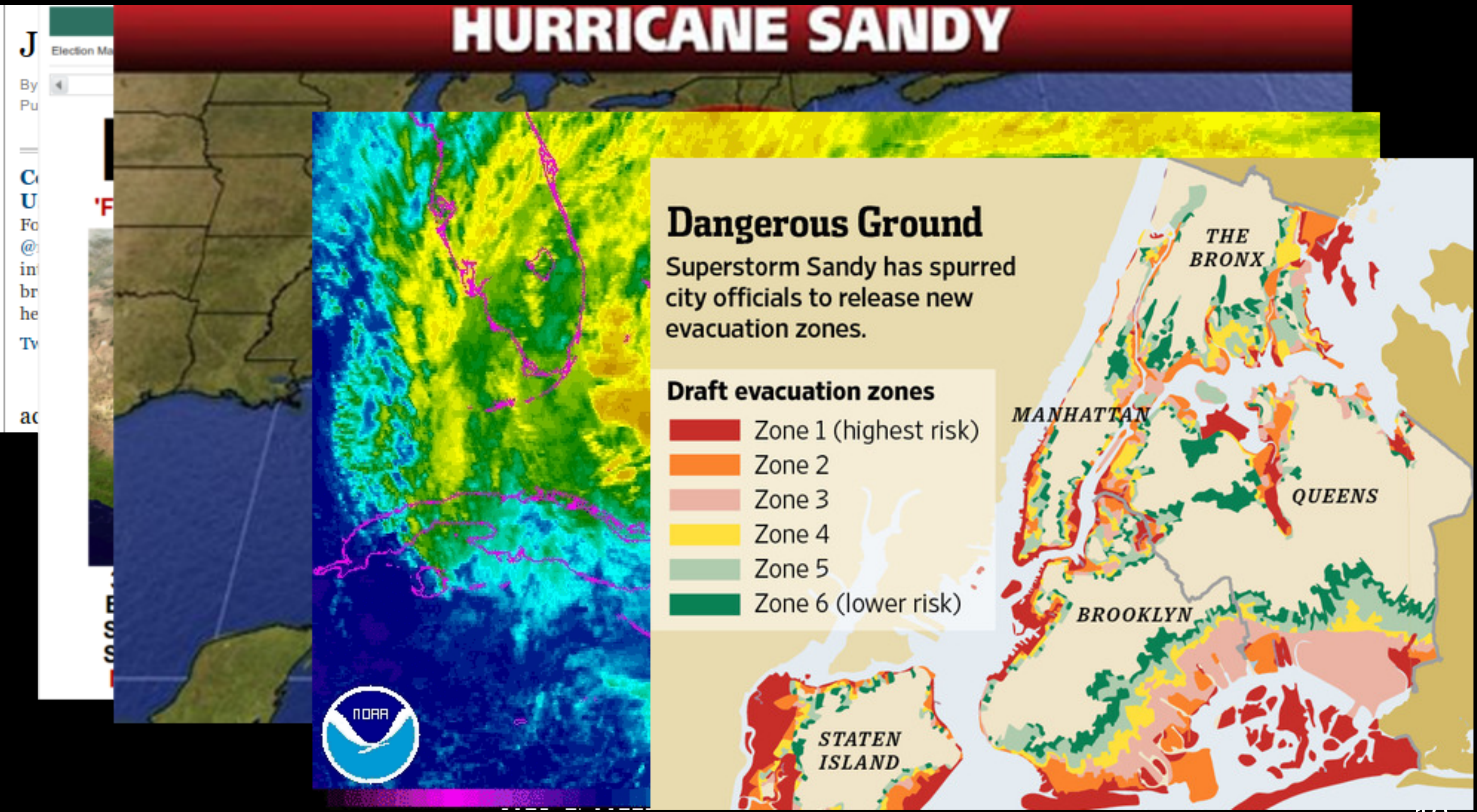
Novels






ČÍTAJTE DNES PRÍLOHU
strana 2
Slovensko si pripomína 63. výročie SNP

Problem: Identifying needs during disaster



Monitor events over time

- Input: streaming data
- News, web pages
- At every hour, what’s new

Track events and SubEvents

Hurricane
Sandy

Manhattan Blackout

Breezy Point fire

Public Transit Outage

Data from NIST: 2011 – 2013
Web Crawl, 11 categories

:15

nbcdfw.com

local • news • classifieds

headlines: Rothko at the Modern ... city insists brown water safe to drink

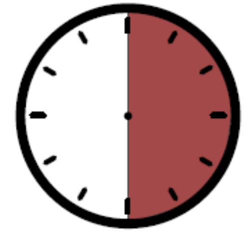
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.

The magnitude-7.5 quake, about 20 miles deep, was centered off the town of Champerico.

People fled buildings in Guatemala City, in Mexico City and in the capital ofthe Mexican state of Chiapas, across the border from Guatemala.

Would you like to contribute to this story? Start a discussion.

:30



nbcdfw.com
local • news • classifieds
headlines: Rothko at

The U.S
said t
Pacifi
there
withir
but t
warni

The
miles
Cham

People
in Me
Mexican state of Chiapas, across the border from Guatemala.

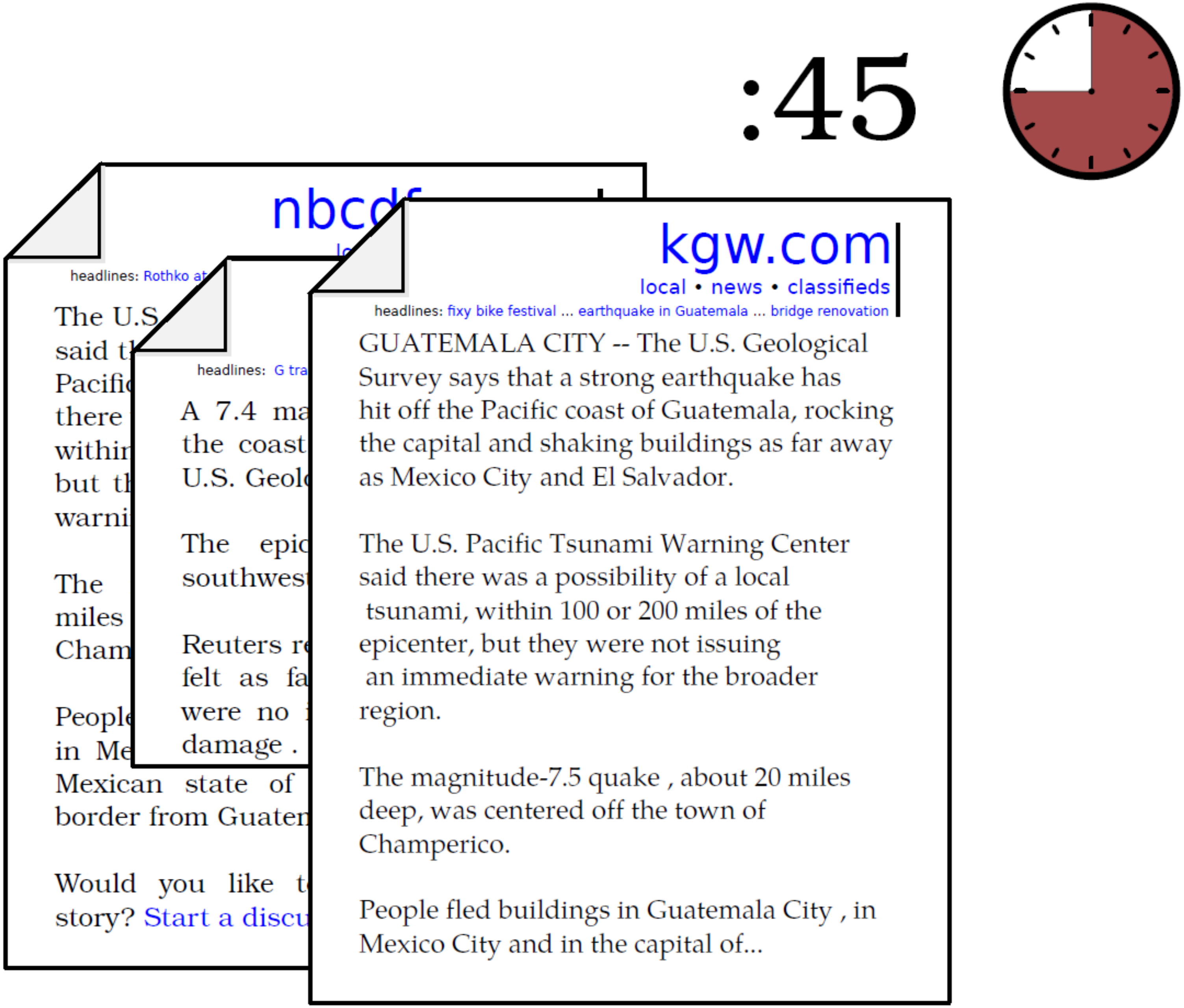
Would you like to contribute to this story? Start a discussion.

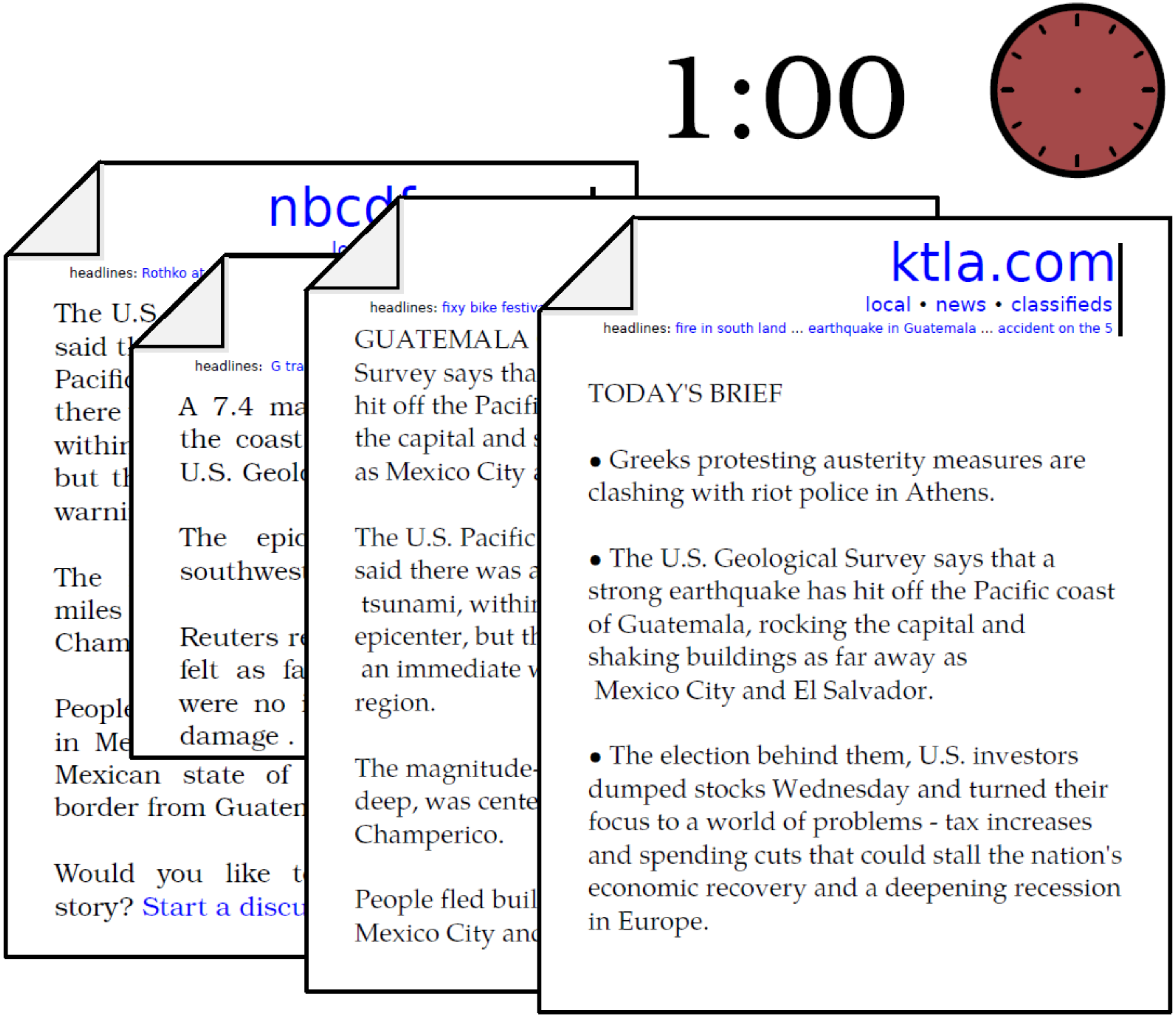
ny1.com
local • news • classifieds
headlines: G train stuck forever ... weather on the 1's ... rats eat tourists

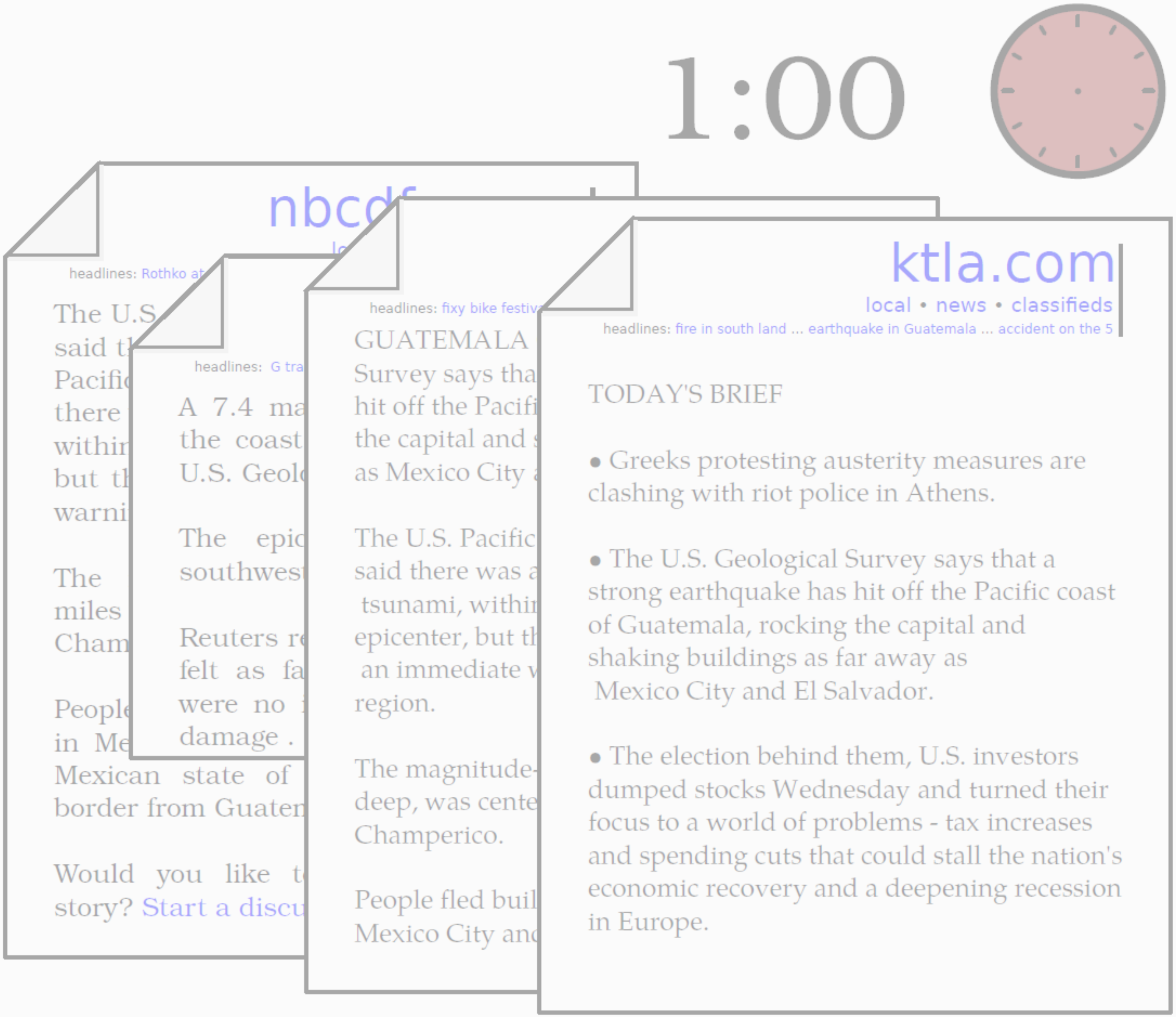
A 7.4 magnitude earthquake struck off the coast of Guatemala Wednesday, the U.S. Geological Survey reported.

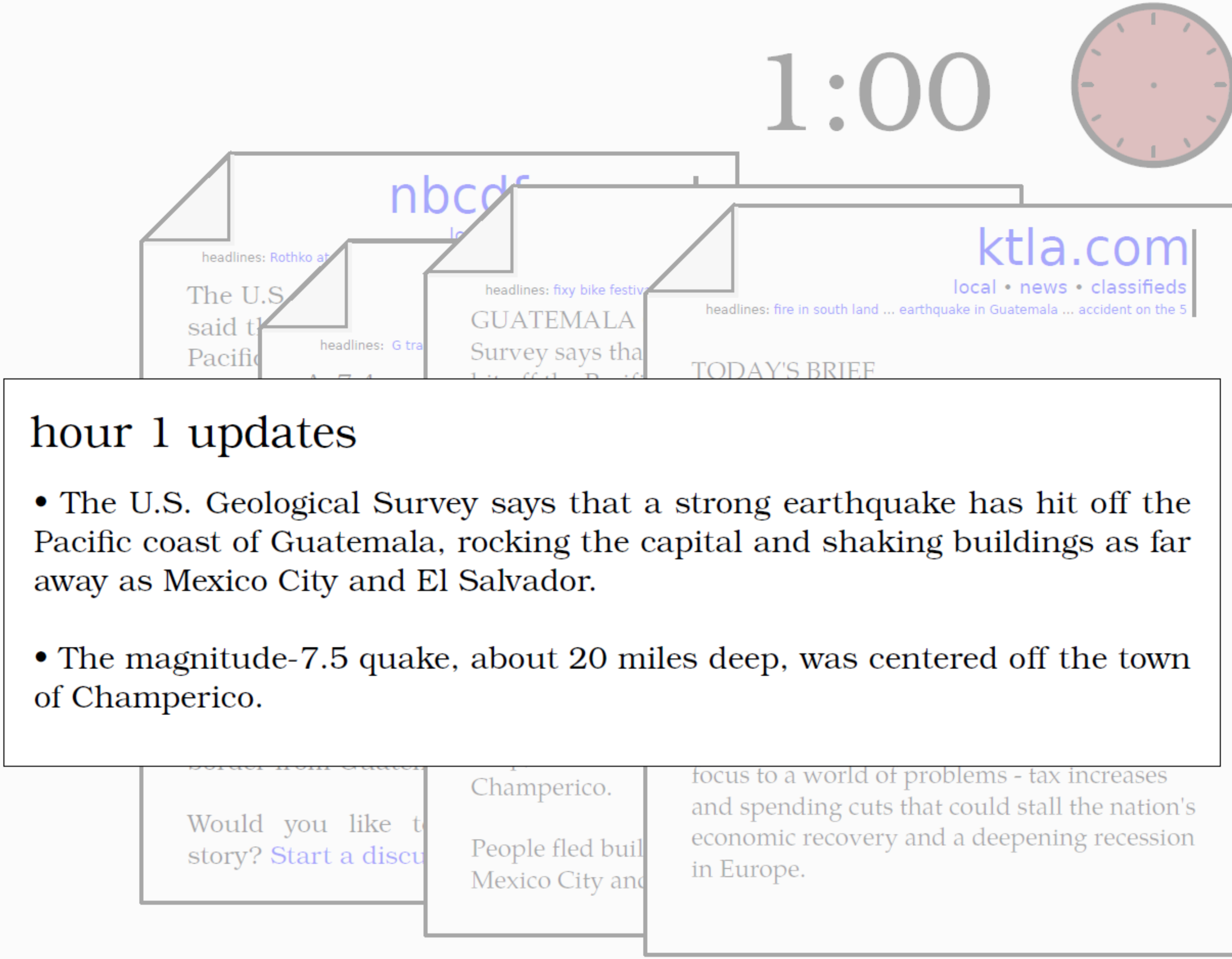
The epicenter was 124 miles west southwest of Guatemala City.

Reuters reported that the quake could be felt as far away as Mexico City. There were no immediate reports of injury or damage .


:45
kgw.com
local • news • classifieds
headlines: fixy bike festival ... earthquake in Guatemala ... bridge renovation
GUATEMALA CITY -- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
The U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter, but they were not issuing an immediate warning for the broader region.
The magnitude-7.5 quake , about 20 miles deep, was centered off the town of Champerico.
People fled buildings in Guatemala City , in Mexico City and in the capital of...


1:00
ktla.com
local • news • classifieds
headlines: fire in south land ... earthquake in Guatemala ... accident on the 5
TODAY'S BRIEF
• Greeks protesting austerity measures are clashing with riot police in Athens.
• The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
• The election behind them, U.S. investors dumped stocks Wednesday and turned their focus to a world of problems - tax increases and spending cuts that could stall the nation's economic recovery and a deepening recession in Europe.


1:00
ktla.com
local • news • classifieds
headlines: fire in south land ... earthquake in Guatemala ... accident on the 5
TODAY'S BRIEF
• Greeks protesting austerity measures are clashing with riot police in Athens.
• The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
• The election behind them, U.S. investors dumped stocks Wednesday and turned their focus to a world of problems - tax increases and spending cuts that could stall the nation's economic recovery and a deepening recession in Europe.



hour 1 updates

- The U.S. Geological Survey says that a strong earthquake has hit off the Pacific coast of Guatemala, rocking the capital and shaking buildings as far away as Mexico City and El Salvador.
- The magnitude-7.5 quake, about 20 miles deep, was centered off the town of Champerico.

Temporal Summarization Approach

At time **t**:

1. Predict salience for input sentences
   - Disaster-specific features for predicting salience
2. Remove redundant sentences
3. Cluster and select exemplar sentences for **t**
   - Incorporate salience prediction as a prior

Kedzie & al, Bloomberg Social Good Workshop, KDD 2014
Kedzie & al, ACL 2015

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

- generic news corpus (10 years AP and NY Times articles)
- domain specific corpus (disaster related Wikipedia articles)

A domain specific language model scores sentences by how typical they are of the disaster type

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

- generic news corpus (10 years AP and NY Times articles)
- domain specific corpus (disaster related Wikipedia articles)

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and oce buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

- tag input with Named-Entity tagger
- get coordinates for locations and mean distance to event

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and oce buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

- tag input with Named-Entity tagger
- get coordinates for locations and mean distance to event

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

Medium Salience

People streamed out of homes, schools and oce buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

Predicting Salience: Model Features

Language Models (5-gram Kneser-Ney model)

Geographic Features

Semantics

- number of event type synonyms, hypernyms, and hyponyms

High Salience

Nicaragua's disaster management said it had issued a local tsunami alert.

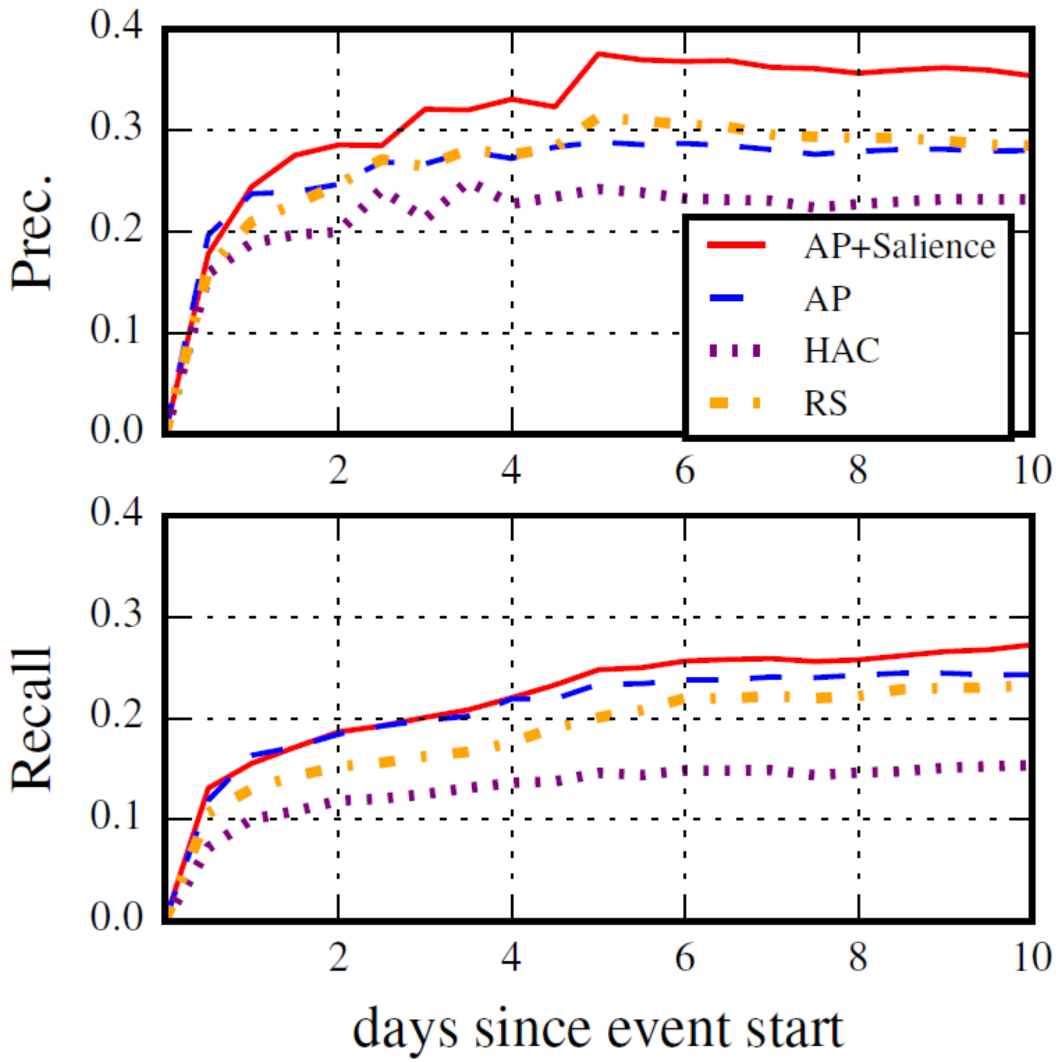
Medium Salience

People streamed out of homes, schools and oce buildings as far north as Mexico City.

Low Salience

Add to Digg Add to del.icio.us Add to Facebook Add to Myspace

What Have We Learned?



- Salience predictions lead to high precision quickly
- Salience predictions allow us to more quickly recover more information

What's next?

- Experimenting with neural nets for summarization updates

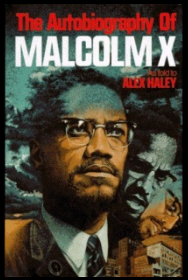
- Aiming for abstractive summarization
  - New sentences generated from phrases

FACT

News

Wikipedia descriptions

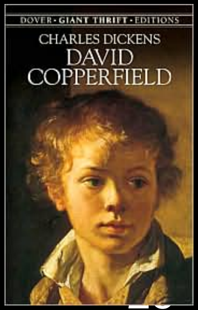
Personal narrative

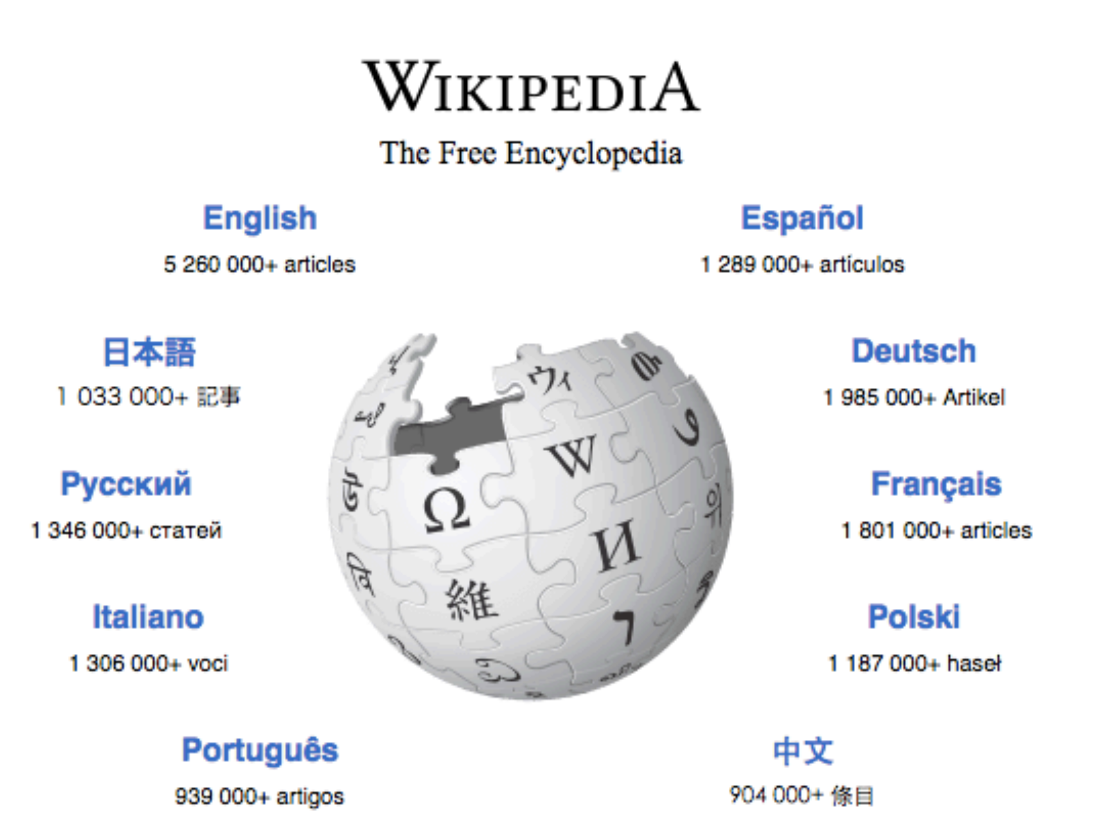


Social Media

FICTION

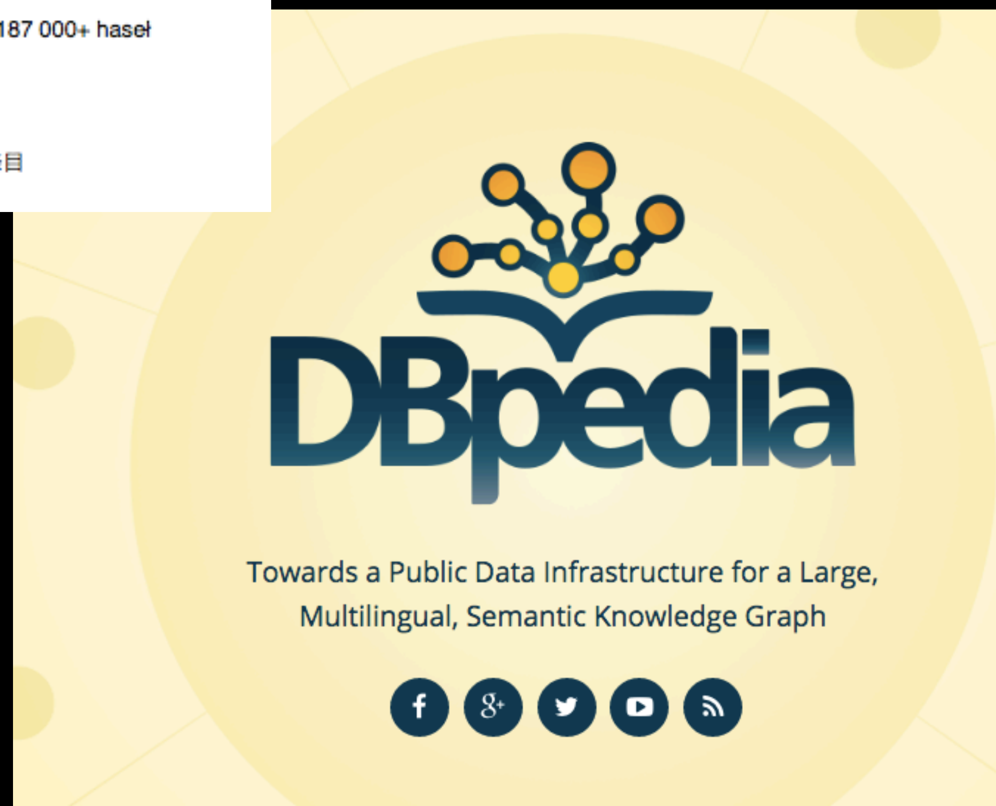
Novels





Output

Input



How do Wikipedia descriptions differ?

- Formal language
- Stylized format
- Less varied content (for certain kinds of entries)

RDF Applications

Applications that generate descriptions of semantic web entities

- Biographies
- Company descriptions

RDF Applications

Biographies

<typeOf> <Person>

<profession>

- Politicians
- Models

Company descriptions

<typeOf> <Company>

<industry>

- Automotive
- Video games

RDF Applications

Core messages: messages built from rdf triples with instance entity as subject

- The [predicate] of [subject] is [object]
- [subject]'s [predicate] is [object]

- The location of Tessla is Palo Alto.

Paraphrasal Template Extraction

Mining paraphrasal templates from a domain corpus (Wikipedia)

Use taxonomy to find richly-typed paraphrasal templates

Find paraphrasal templates from non-paraphrase sentences

Sentences from the corpus

- GM Taiwan was founded in August 1989
- In 1904, the company was established as Oscar Lear Automobile Company

Entities/dates identified

- [GM Taiwan] was founded in [August 1989]
- In [1904], the company was established as [Oscar Lear Automobile Company]

Types replace entities/dates

Paraphrases:

- [company] was founded in [date]
- In [date], the company was established as [company]

Core Message:

- The [start-date] of [company] was [date]

- Tesla Motors was founded by JB Straubel, Martin Eberhard and Elon Musk…Tesla Motors is a privately held company with approximately 6,000 employees. The product of Tesla Motors is Luxury vehicle. Tesla Motors' location is Palo Alto, California. In 2003, the company reorganized, adapting its current name, Tesla Motors. In May 2010, Toyota launched a collaboration with Tesla Motors to create electric vehicles. The "Tesla Factory" is an automobile manufacturing plant in Fremont, California, US, owned and operated by Tesla Motors.

- Tesla Motors was founded by JB Straubel, Martin Eberhard and Elon Musk…Tesla Motors is a privately held company with approximately 6,000 employees. The product of Tesla Motors is Luxury vehicle. Tesla Motors' location is Palo Alto, California. In 2003, the company reorganized, adapting its current name, Tesla Motors. In May 2010, Toyota launched a collaboration with Tesla Motors to create electric vehicles. The "Tesla Factory" is an automobile manufacturing plant in Fremont, California, US, owned and operated by Tesla Motors.

Hybrid Approach

- Can we augment RDF triples with information drawn from the web?
- Summarization approach: Find all relevant sentences that match the entity and/or contain RDF triples on the web.

- Tesla Motors was founded by JB Straubel, Martin Eberhard and Elon Musk…Tesla Motors is a privately held company with approximately 6,000 employees. The product of Tesla Motors is Luxury vehicle. Tesla Motors' location is Palo Alto, California. In 2003, the company reorganized, adapting its current name, Tesla Motors. *In May 2010, Toyota launched a collaboration with Tesla Motors to create electric vehicles. The "Tesla Factory" is an automobile manufacturing plant in Fremont, California, US, owned and operated by Tesla Motors.*

Evaluation

100 texts generated from each application-domain combination, in 5 versions

- Full system
- No paraphrasal templates
- No hybrid and no components (baseline)

Annotators shown two texts, asked which is better (or equal)

On criteria:

- Content
- Ordering
- Style
- Overall

What have we learned?

	Preference	Content	Ordering	Style	Overall
Baseline VS Full System	Baseline	20%	27%	24%	22%
	Equal	14%	11%	20%	14%
	Full System	66%	62%	56%	64%
	Winning difference	**46% †**	**35% †**	**32% †**	**42% †**

What have we learned?

	Preference	Content	Ordering	Style	Overall
Baseline VS Full System	Baseline	20%	27%	24%	22%
	Equal	14%	11%	20%	14%
	Full System	66%	62%	56%	64%
	Winning difference	**46%** †	**35%** †	**32%**†	**42%** †
No Paraphrases VS Full System	No Paraphrases	29%	33%	29%	30%
	Equal	31%	26%	28%	27%
	Full System	40%	41%	43%	43%
	Winning difference	**11%** †	**8%** †	**14%** †	**13%** †

What's next?

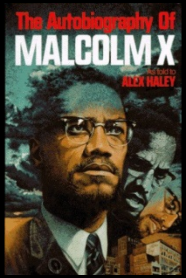
- Handling pronouns and other style questions
- Extending to other types of entities
- Developed similar approaches for explanation for machine learning

FACT

News

Wikipedia descriptions

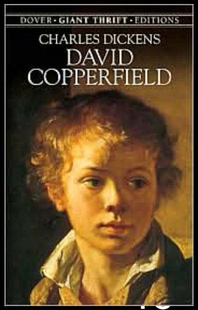
Personal narrative

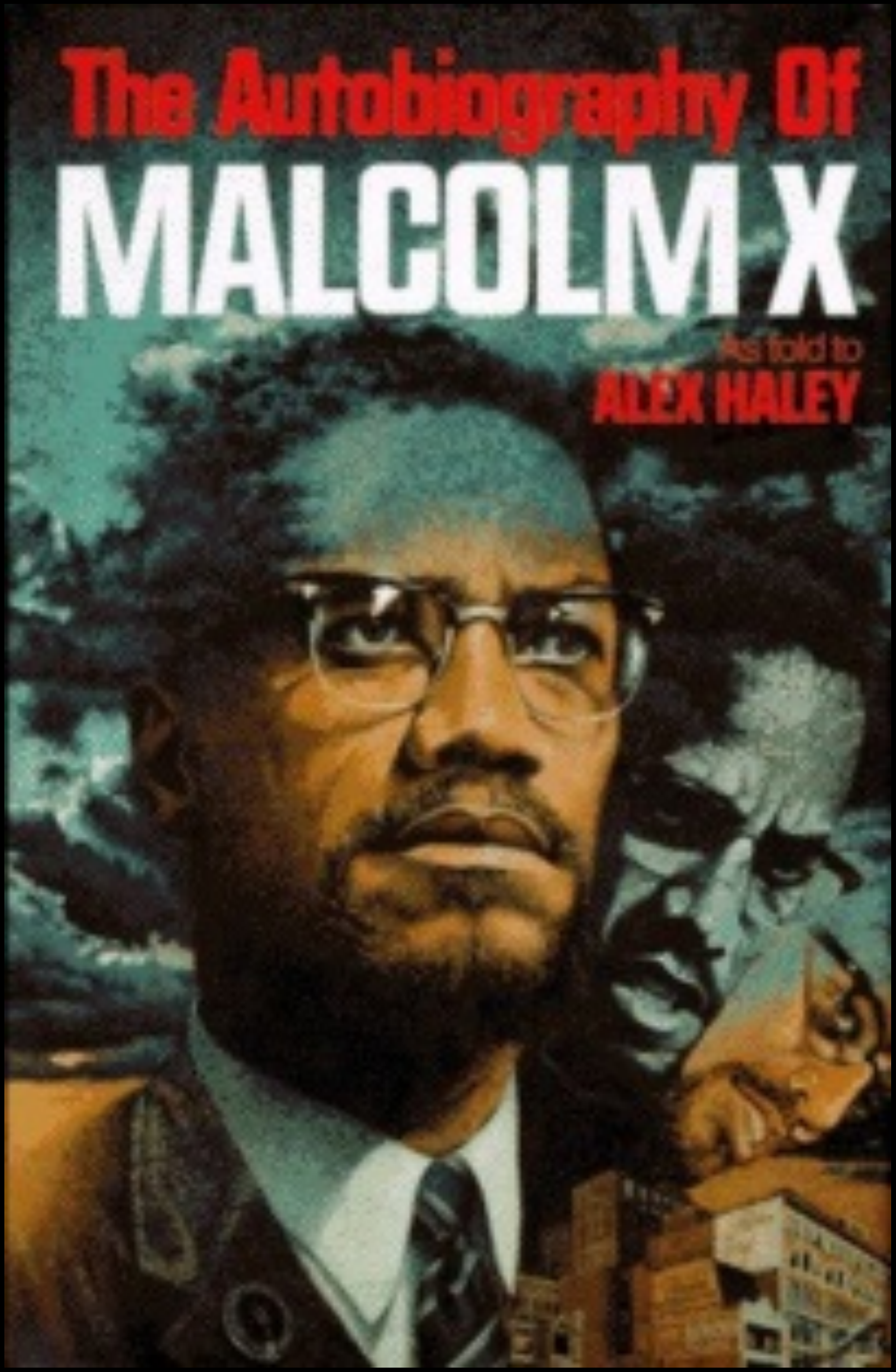


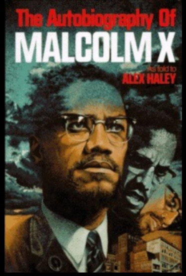
Social Media

Novels

FICTION



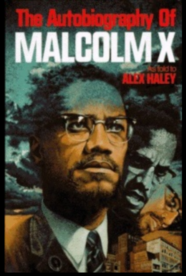

The Autobiography Of
MALCOLM X
As told to
ALEX HALEY



How is Personal Narrative Different?

- Coherent telling of a story
- Compelling component
- Monologue
- Informal language

PERSONAL VIEWS



We were sitting down to a late dinner on Monday night when the storm was supposed to hit. It was incredibly windy but the rain really hadn't been that bad.

…

"By 10 p.m., the skies lit up in a purple and blue brilliance and the power started to go out here and there….That's when I noticed neighbors across the street running out of their homes and fire trucks racing down the block. I saw a trickle of steady water coming down the street on both sides and then water began pouring in through the creaks in the basement door, so my husband went to grab the pump. He went upstairs to get a tool and in those few seconds, ocean waves broke the steel door lock and flooded the basement 6 feet high in minutes."

We were sitting down to a late dinner on Monday night when the storm was supposed to hit. It was incredibly windy but the rain really had

Background

…

"By 10 p.m., the skies lit up in a purple and blue brilliance and the power started to go out here and there….That's when I noticed neighbors across the street running out of their homes and fire trucks racing down the block. I saw a trickle of steady water coming down the street on both sides and then water began pouring in through the creaks in the basement door, so my husband went to grab the pump. He went upstairs to get a tool and in those few seconds, ocean waves broke the steel door lock and flooded the basement 6 feet high in minutes."

We were sitting down to a late dinner on Monday night when the storm was supposed to hit. It was incredibly windy b... n't been that bad.

Complicating action

"By 10 p.m., the skies lit up in a purple and blue brilliance and the power started to go out here and there….That's when I noticed neighbors across the street running out of their homes and fire trucks racing down the block. I saw a trickle of steady water coming down the street on both sides and then water began pouring in through the creaks in the basement door, so my husband went to grab the pump. He went upstairs to get a tool and in those few seconds, ocean waves broke the steel door lock and flooded the basement 6 feet high in minutes."

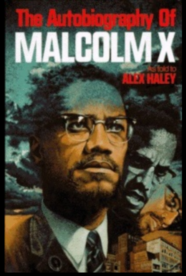
We were sitting down to a late dinner on Monday night when the storm was supposed to hit. It was incredibly windy but the rain really hadn't been that bad.

…

"By 10 p.m., the skies lit up in a purple and blue brilliance and the power started to go out here and there….That's when I noticed neighbors across the street running out of their homes [illegible]g down the block. I saw a trickle of steady [illegible]the street on both sides and then water [illegible]ough the creaks in the basement door, so my husband went to grab the pump. He went upstairs to get a tool and in those few seconds, ocean waves broke the steel door lock and flooded the basement 6 feet high in minutes."

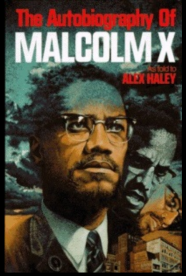
Reportable event

Identify the Reportable Event



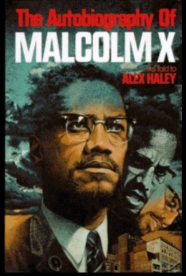
- Which sentence(s) convey the compelling event?
- The reportable event could serve as a summary for “what is this story about?”

Data

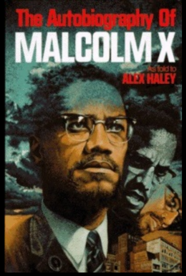


- AskReddit subreddit: e.g., ``What's your creepiest real life story?''
  - 3000 stories
- Small amount manually labeled (seed)
- Large amount automatically labeled using distant supervision

Linguistic Theory



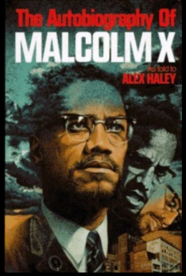
- Prince: stories about change
- Polanyi: turning point marked by change in formality, style, emphasis
- Labov: a change in verb tense often accompanies the MRE

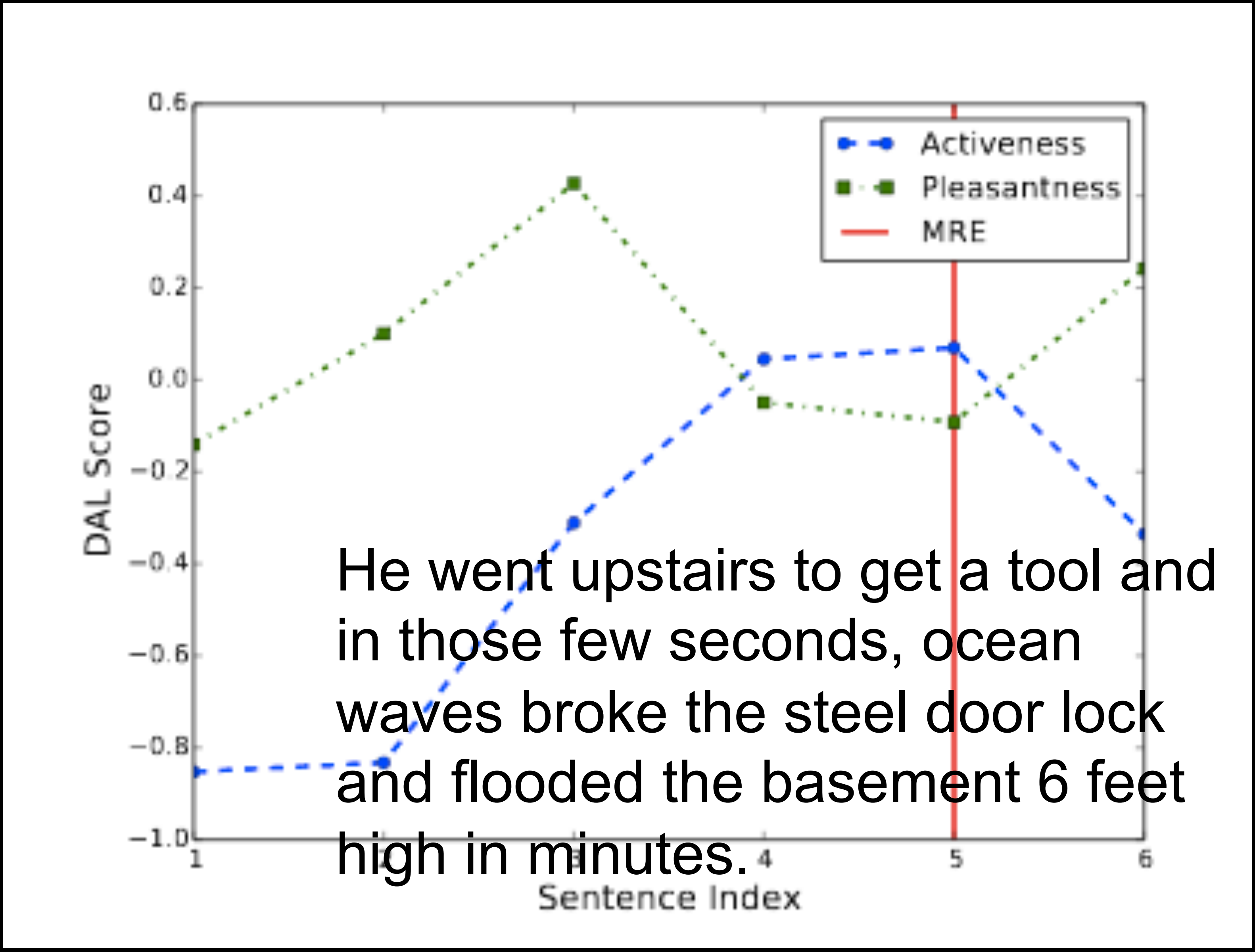


Sentence scores

- Syntactic: e.g., sentence length
- Semantic: similarity to surrounding sentences
- Affect: pleasantness, activeness, imagery

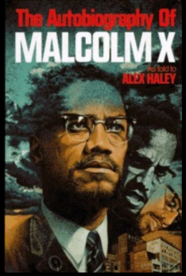
Features: Change in Affect





He went upstairs to get a tool and in those few seconds, ocean waves broke the steel door lock and flooded the basement 6 feet high in minutes.

What have we learned?



- Change features are most effective

- How to use the data
  - Experimented with seed only (small), distant supervision (large but noisy) and self-training

	Precision	Recall	F-measure
Seed only*	0.374	0.617	0.466
Dist. supervision*	0.398	0.745	0.519
Self-training*	**0.478**	**0.946**	**0.635**

Ouyang &McKeown, EMNLP 2015

What's next?

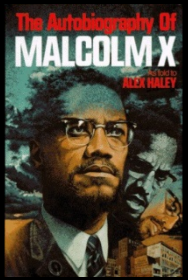
- Rewriting the extracted sentences for better summaries
- Browsing interface with summaries of different experiences

FACT

News

Wikipedia descriptions

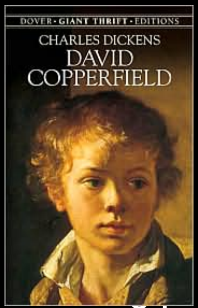
Personal narrative



Social Media

FICTION

Novels



How is social media different

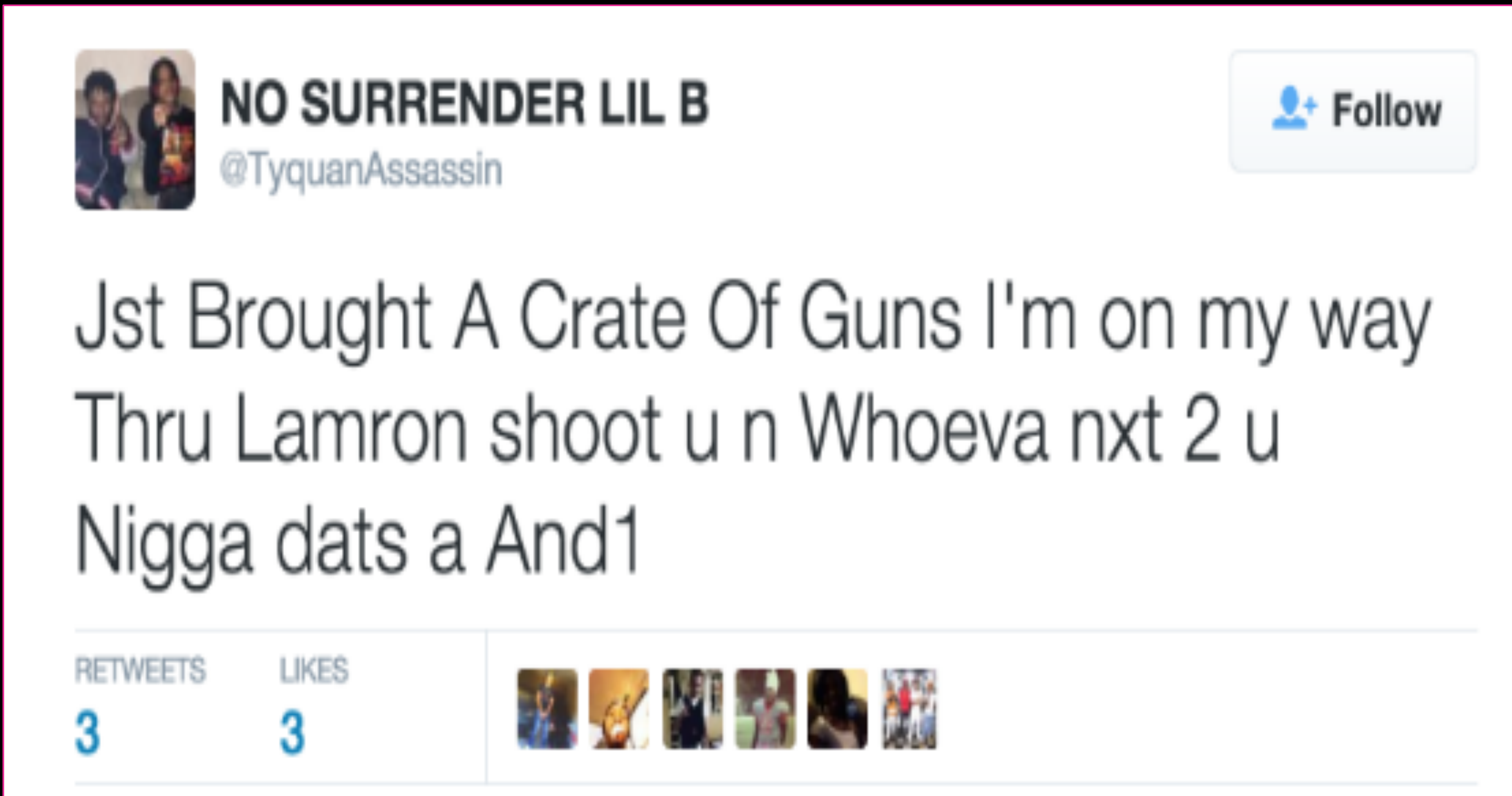
- Informal
- Slang…. Not the language of news!
- Dialog
- Each contribution short

Problem: The U.S. has the highest rate of firearm-related deaths when compared to other industrialized countries.

Violence particularly affects low-income, cities like Chicago, which had **more than 3,000 shooting victims in 2015**.

Gang Violence & Social Media

Can we automatically detect aggressive posts?



Approach

- Collaboration between social workers and natural language processing

- Annotation of tweets using “deep read” by social workers working with Chicago youth
    - Trigger events, triggered events, tone, meaning

- Used as labels for automatic prediction

- Alert community outreach

Case Study

Gakirah Barnes @TyquanAssassin

- **Recently deceased gang member in Chicago**
- **9 killings to her name until she was killed at the age of 17**
- **27,000 tweets from December 2011 to April 11, 2014**
- **~ 4,200 followers on Twitter**

Themes That Emerge From Coding

Aggression

- Insults, threats, bragging, hypervigilance, and challenges with authority.

Grief

- Distress, sadness, loneliness, and death.

Other

- General conversations between users, discussions about women, and tweets that represented happiness.

Qualitative Analysis

If We see a opp Fuck it We Gne smoke em 😈	Aggression (Threat)
Dnt get caught on Dat 800 block lame ass Lil niggas Betta take Dat Shyt on stony spot	Aggression (Insult)
Young niggas still getting shot babies still dying 🙏	Loss

Natural Language Tools

- Part of speech tagger
  - Supervised machine learning plus domain adaptation
  - Semantic clusters plus character ngrams
- Bilingual dictionary
  - Glossed tweets into standard American English and automatically aligned
  - “smoke” -> “kill”

Aggression/Loss Classifier

- Emotion lexicon
  - Scores words according to their affect
- Part-of-speech tags
- Bilingual lexicon
- F-measure: 63%

What have we learned?

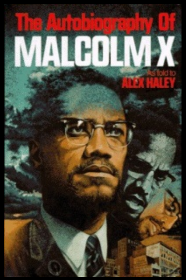
- Need collaboration to make progress
- Context essential to annotation
- Next steps: generating explanation of prediction

FACT

News

Wikipedia descriptions

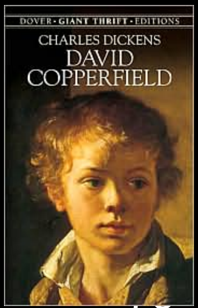
Personal narrative

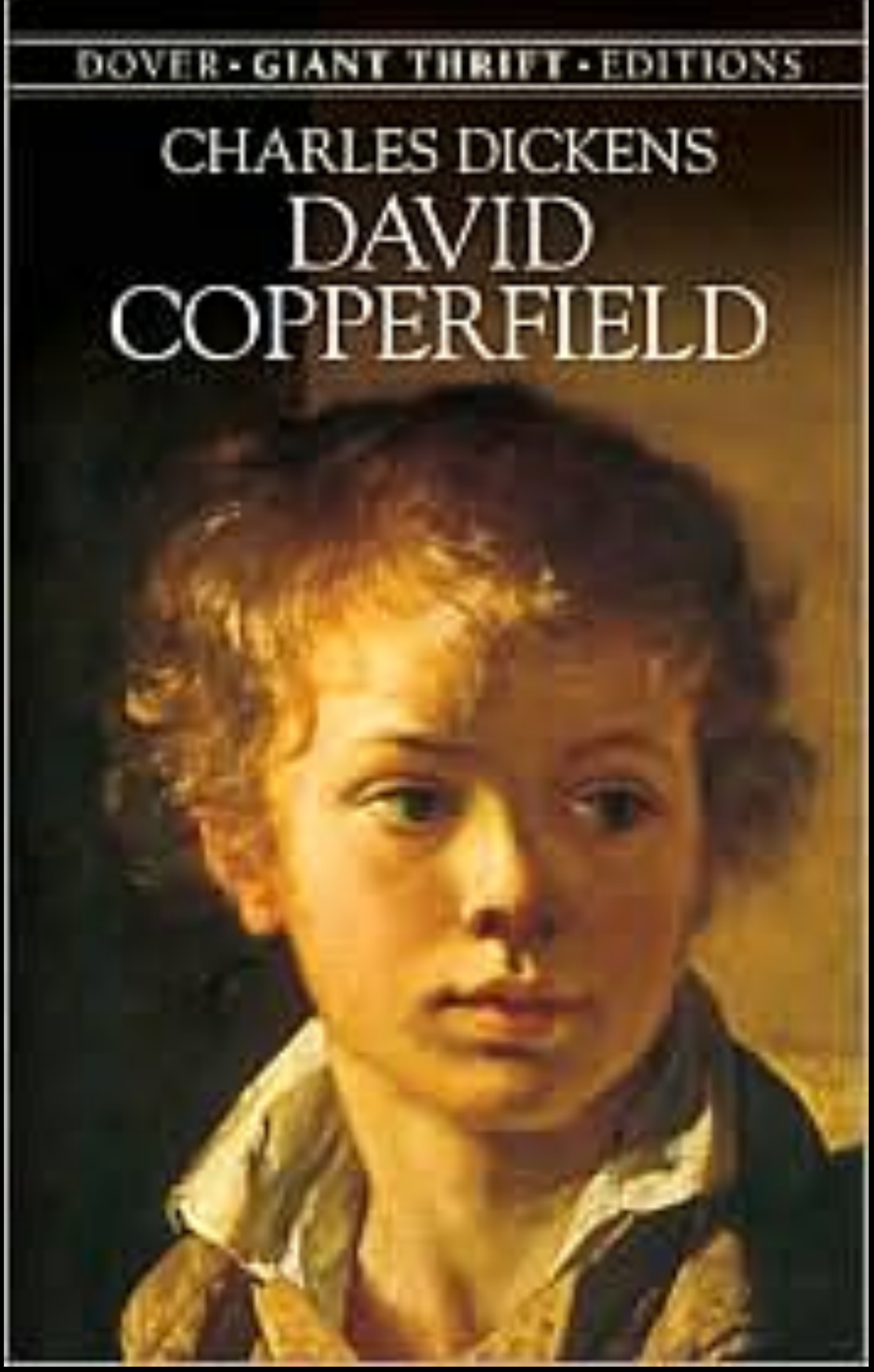


Social Media

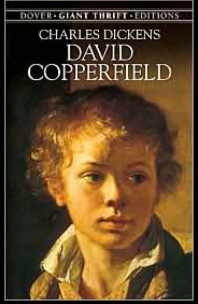
FICTION

Novels




DOVER • GIANT THRIFT • EDITIONS
CHARLES DICKENS
DAVID
COPPERFIELD

Novels



- *It was broad day—eight or nine o'clock; the storm raging, in lieu of the batteries; and someone knocking and calling at my door.*

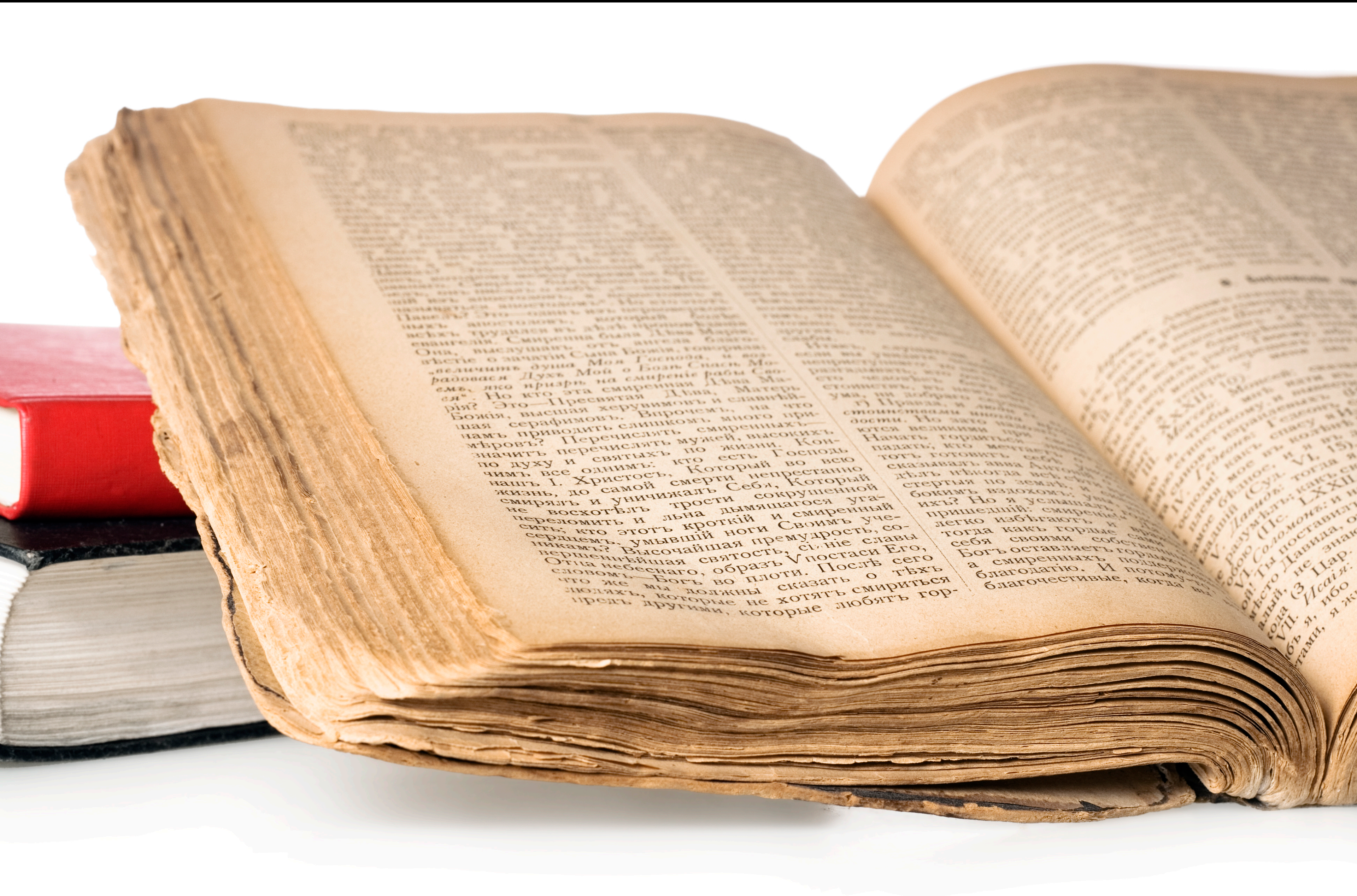
  'What is the matter?' I cried.

  'A wreck! Close by!'

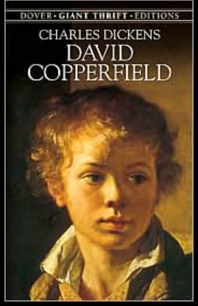
  I sprung out of bed, and asked, what wreck?

  'A schooner, from Spain or Portugal, laden with fruit and wine. Make haste, sir, if you want to see her! …'

Dickens, *David Copperfield*



Computer Science and Comparative Literature



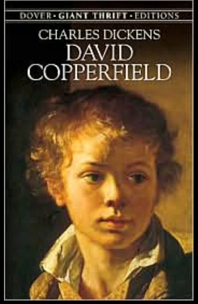
- Provide evidence for or against literary theory

- Social network extraction from literature
  - Corpus of 19^{th} Century British literature
  - Network based on conversation

- Method based on quoted speech
  - Identify who talks to whom
  - Extract graph features that evaluate hypotheses

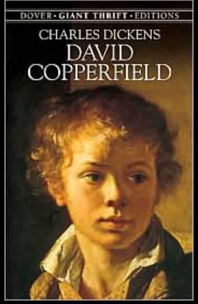
Elson, Dames, & McKeown: ACL 2010

Hypothesis #1



- Larger conversational networks (with more people) tend to be less connected
  - *Franco Moretti*: : at 10 or 20 characters, possible to include "distant and openly hostile groups"
  - *Terry Eagleton:* in a large community, "most of our encounters consist of seeing rather than speaking, glimpsing each other as objects rather than conversing as fellow subjects"
- **Can we show empirically that conversational networks with fewer people are more closely connected?**

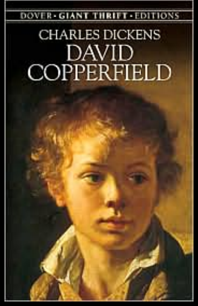
Constructing the Social Network



- Nodes = people who said something
  - Quote attribution with 83% accuracy
- Edges = people who are talking to each other
  - Quote adjacency used as heuristic for detecting conversations
  - Edge weight set to share of detected conversation
  - 95% precision, 51% recall

(Elson and McKeown AAAI 2010)

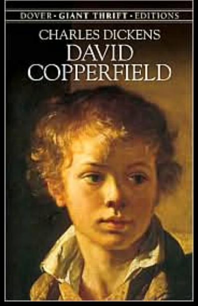
Impact of Network Size



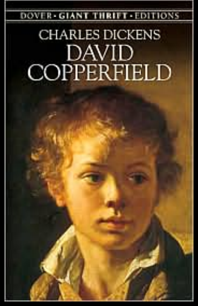
- As the number of named characters increases, we expect:
  - Same or less total speech
    - Weak yes: Normalized number of quotes flat at r=.16
  - Less lopsided distribution of quotes among speakers
    - Yes: Share of quotes by top 3 speakers decreases at r=-.61

Using Pearson's product-moment correlation coefficient

Impact of Network Size



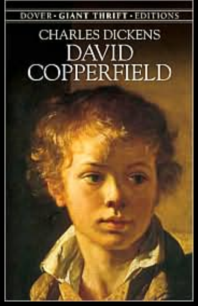
- As the number of named characters increases, we expect:
  - Lower density (fewer conversational partners as percentage of population)
    - **No**: Increases at r=.30. Larger networks are more connected
  - Same or fewer cliques
    - **No**: 3-clique rate increases at r=.38. Larger networks form cliques more often



Impact of Network Size

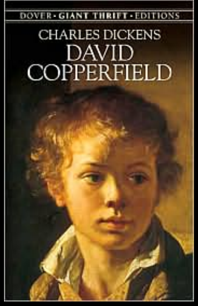
- As the number of speakers increases, we expect:
  - Less overall dialogue ("glimpsing rather than speaking")
    - **No**: Increases at .50. Larger networks are more talkative
  - Lower density
    - **No**: Increases at .49. In larger networks, people know more of their neighbors

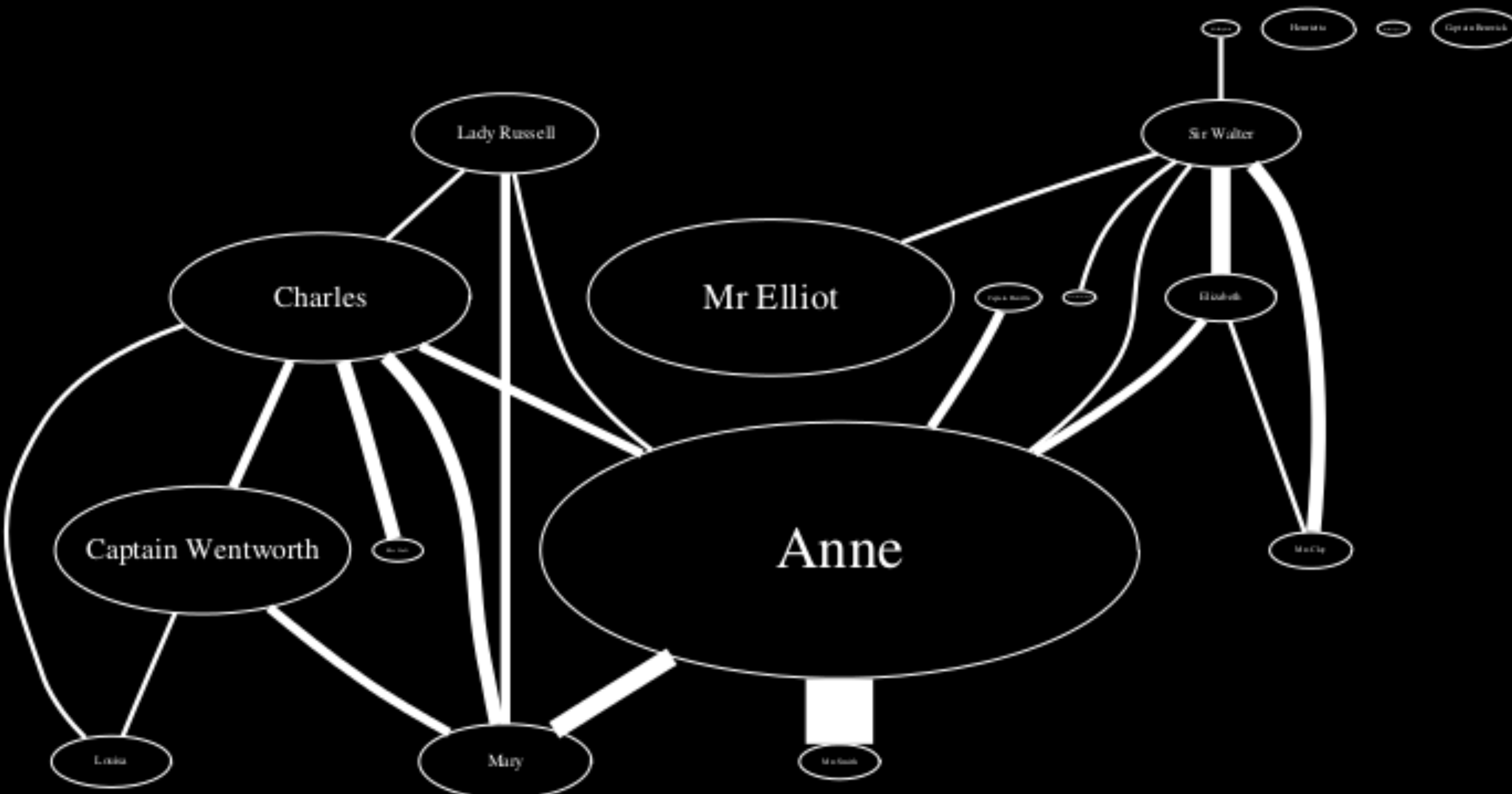
Alternate Explanation



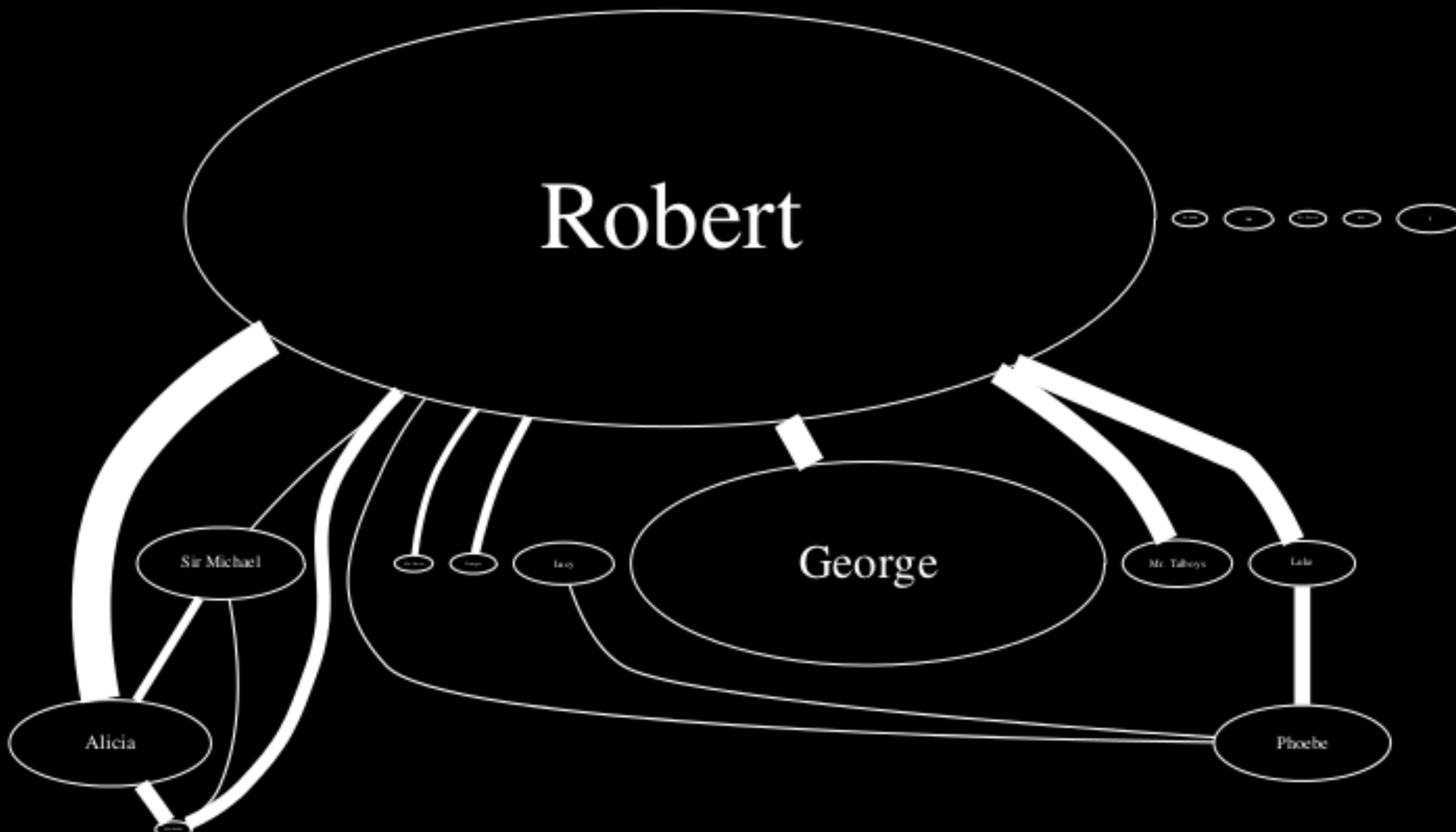
- Text **perspective** *dominates* network shape
  - 3rd person tellings: Significant increases in
    - Normalized number of quotes ($p<.05$)
    - Average degree ($p<.005$)
    - Graph density ($p<.05$)
    - Rate of 3-cliques ($p<.005$)
      - …With no significant difference in number of characters or speakers
  - Hypothesis: First-person narrators not privy to other characters' conversations with each other

3rd Person Narrative
(Austen, Persuasion)



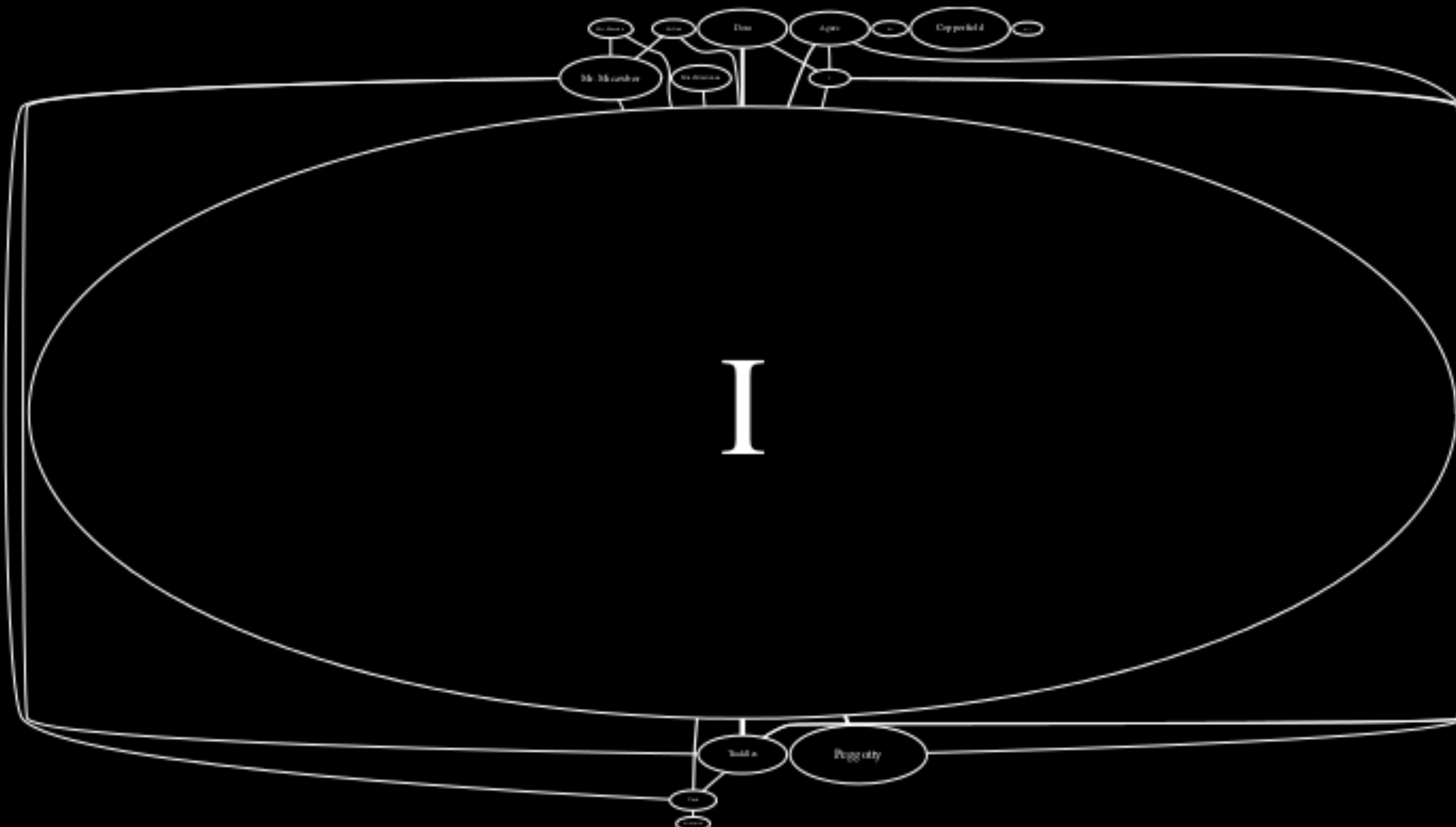


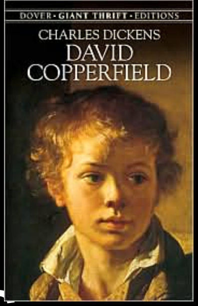
“Close 3rd” Narrative
(Braddon, Lady Audley’s Secret)



1^{st} *Person Narrative*

(Dickens, David Copperfield)





What have We Learned?

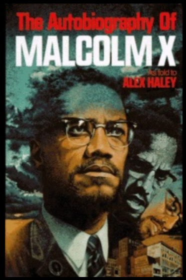
- High-precision conversational networks can be extracted from literature

- Time is right for exploring analysis of fiction

  What might we have overlooked?
  Working now on detection of scene changes

News

Wikipedia
Texts



Personal
narrative

Online discussion forums

News

Semantics

Web

Wikipedia Texts

Personal narrative

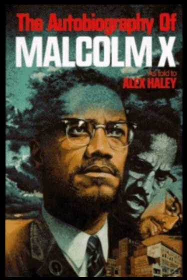
Online discussion forums

News

Semantics

Web

Wikipedia
Texts



Personal
narrative

Pragmatics

Real-world
context

Online discussion forums

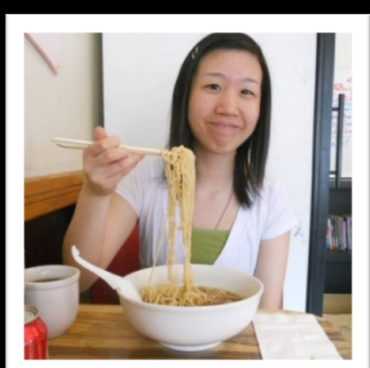
Current PhD Students

Or Biran

Noura Farra

Chris Hidey

Chris Kedzie



Jessica Ouyang

Melody Ju

Fei-Tzin Lee

Elsbeth Turcan

Past Students


Regina Barzilay


Sasha Blair-Goldensohn


Andrea Danyluk


Galina Datskovsky Moerdler


Pablo Duboue


Michael Elhadad


Noemie Elhadad


David Elson


David Evans


Elena Filatova


Pascale Fung


Michael Galley


Vasileios Hatzivassiloglou


Hongyan Jing


Min Yen Kan


Ani Nenkova


Shimei Pan


Cecile Paris


Kristen Parton


Dragomir Radev


Jacques Robin


Carl Sable


Barry Schiffman

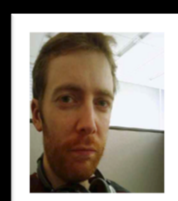

James Shaw


Eric Siegel


Frank Smadja


Ursula Wolz


Weiyun Ma


Yves Petinot


Sara Rosenthal


Kapil Thadani

Thank You!

- The research presented here has been supported in part by DARPA and NSF.