At the Intersection of Language and Data Science

Kathleen McKeown
Department of Computer Science
Columbia University

Vision

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

COLUMBIA'S DATA SCIENCE INSTITUTE CENTER FOR NEW MEDIA

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

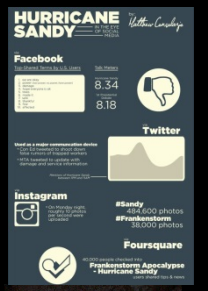
What features yield good predictions?

FACT

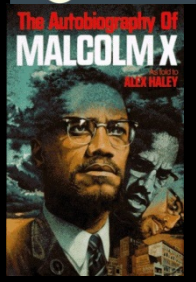
Scientific Journal Articles

News

Online discussion forums

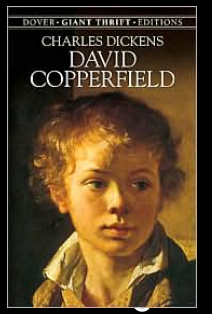


Personal narrative



FICTION

Novels

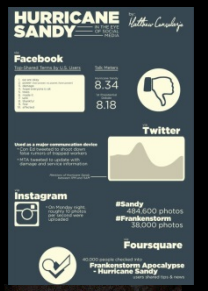


FACT

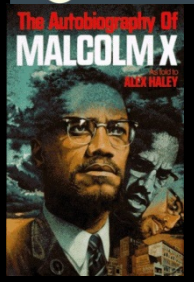
Scientific Journal Articles

News

Online discussion forums

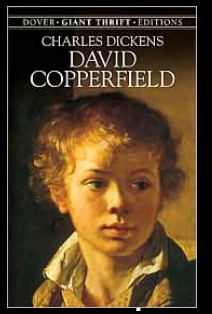


Personal narrative



FICTION

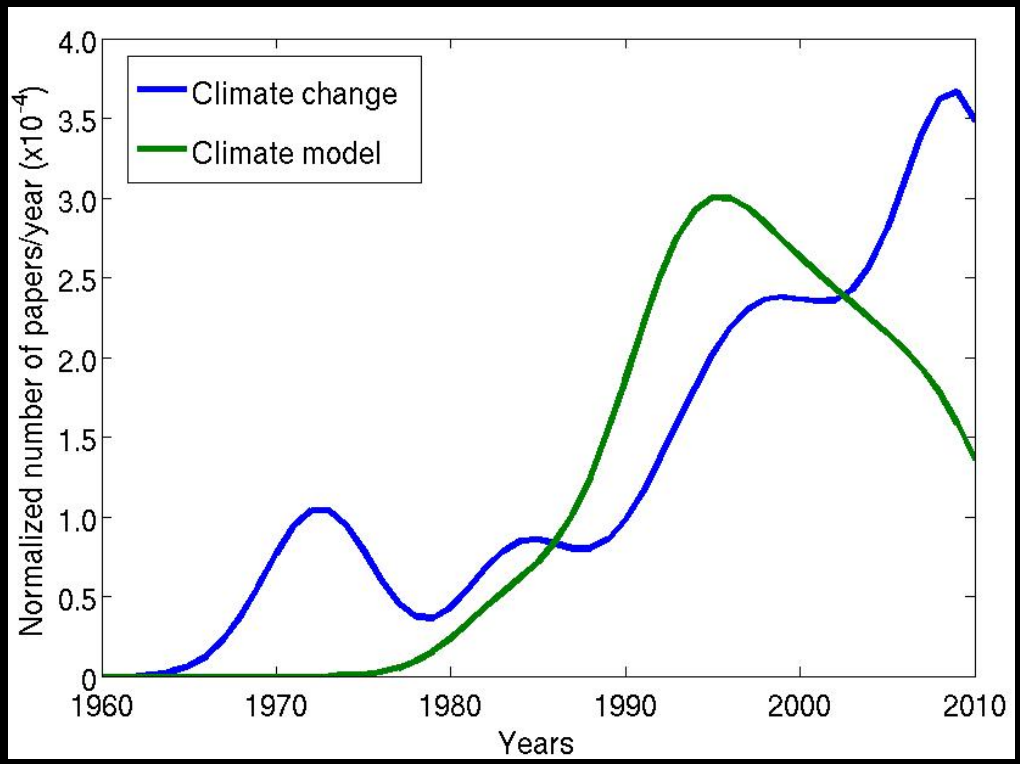
Novels




nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
WOMEN'S WORK
Why is science still institutionally sexist? PAGE 21
HISTORY
RENAISSANCE THINKING
Science and culture in Medici Florence
PAGE 46
GEOPHYSICS
ICE, ICE EVERYWHERE
How the oceans would fare in a Snowball Earth
PAGE 90
CAREERS
COMMUNITY SPIRIT
The rewards of teaching in US community colleges
PAGE 129
NATURE.COM/NATURE
7 March 2013 £10
Vol 495, No 7439
TIFF Preview

Predicting Future Scientific Impact

- Input: term, document
- Extract features from full text
- Predict prominence of term, document

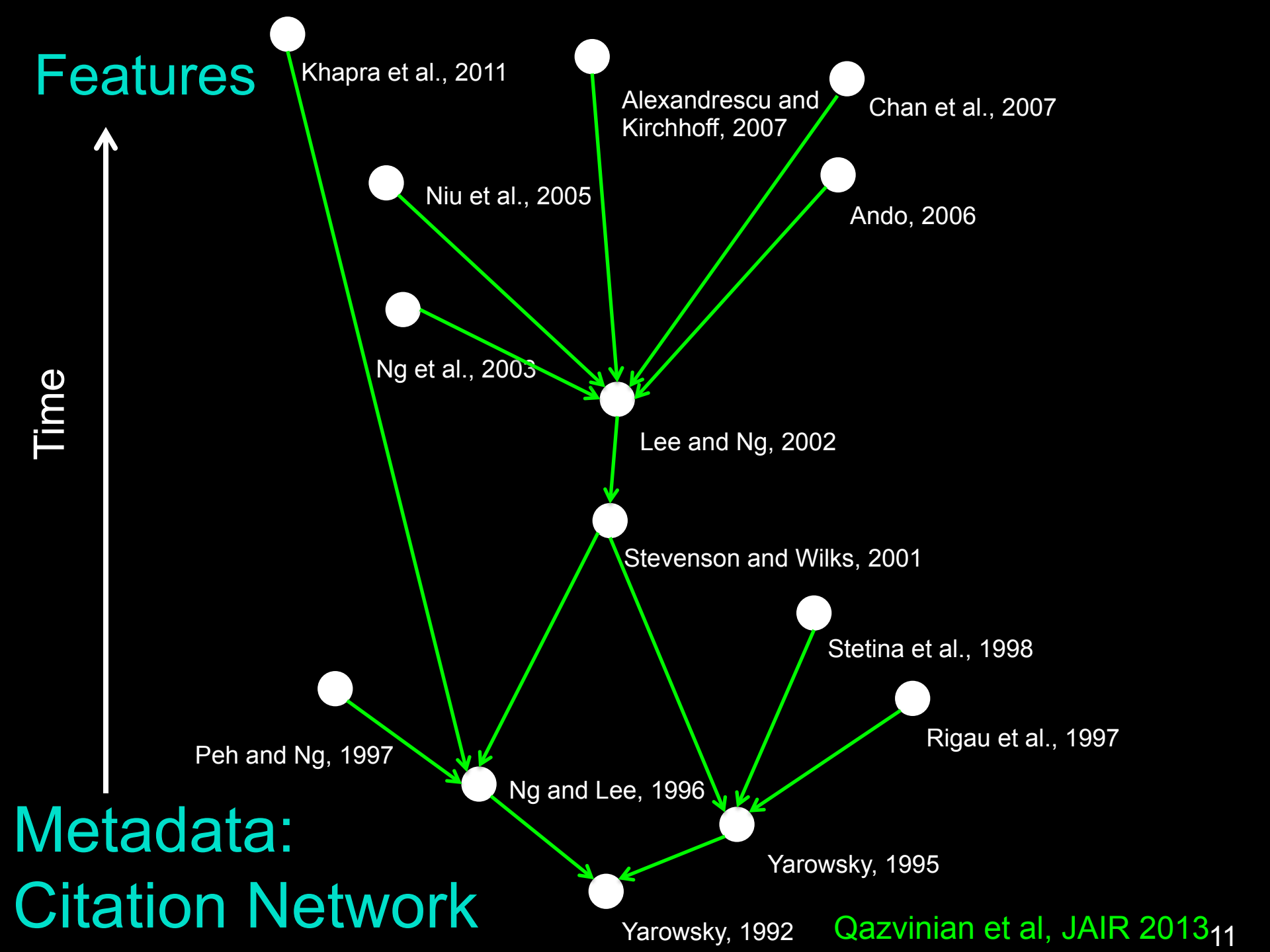


Climate change, Climate model

McKeown et al, JASIST forthcoming

Data

- 4 million full-text Elsevier journal articles
- 48 million Web of Science metadata records
- Fields: medical, chemistry, biology, computer science, …

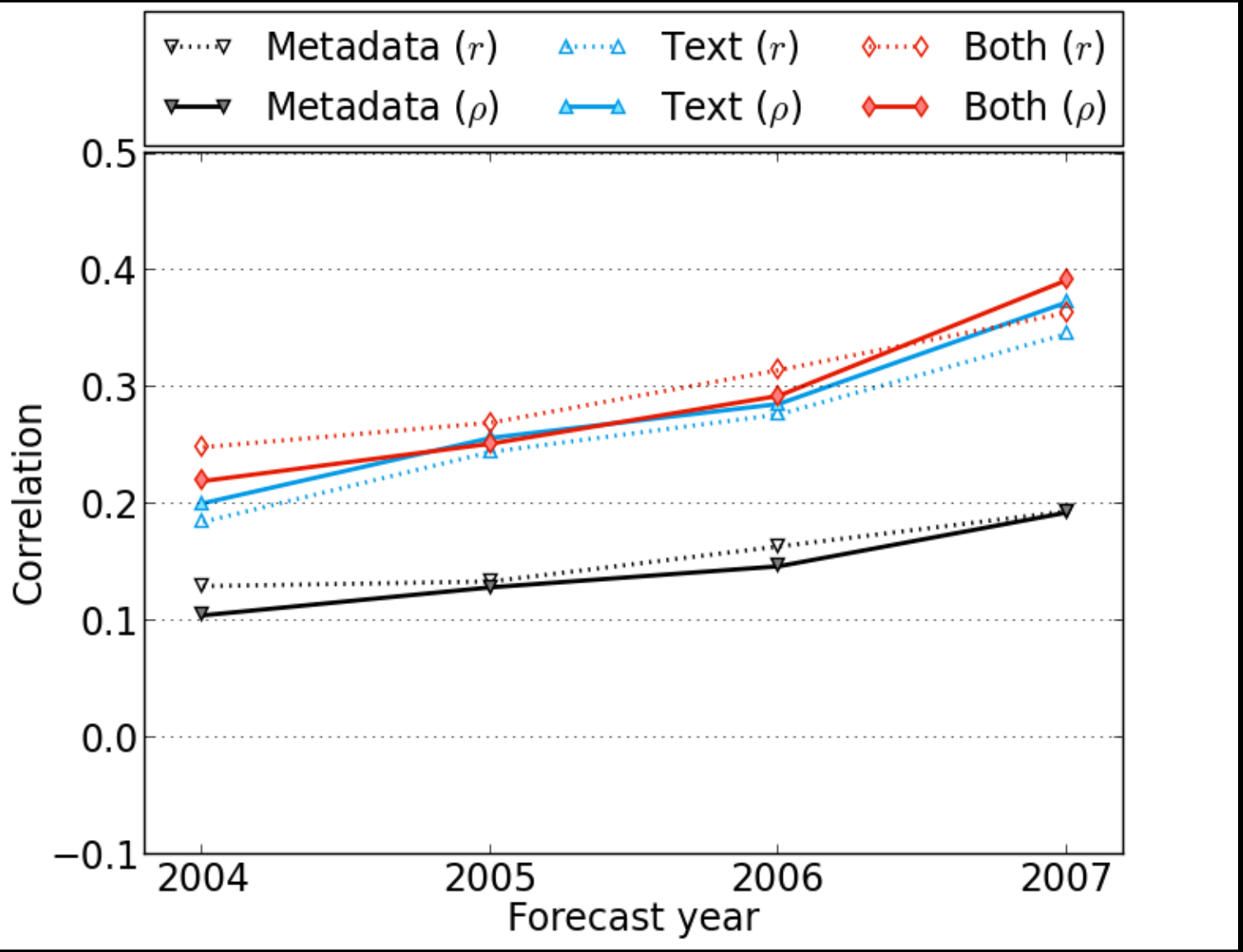


Qazvinian et al, JAIR 2013

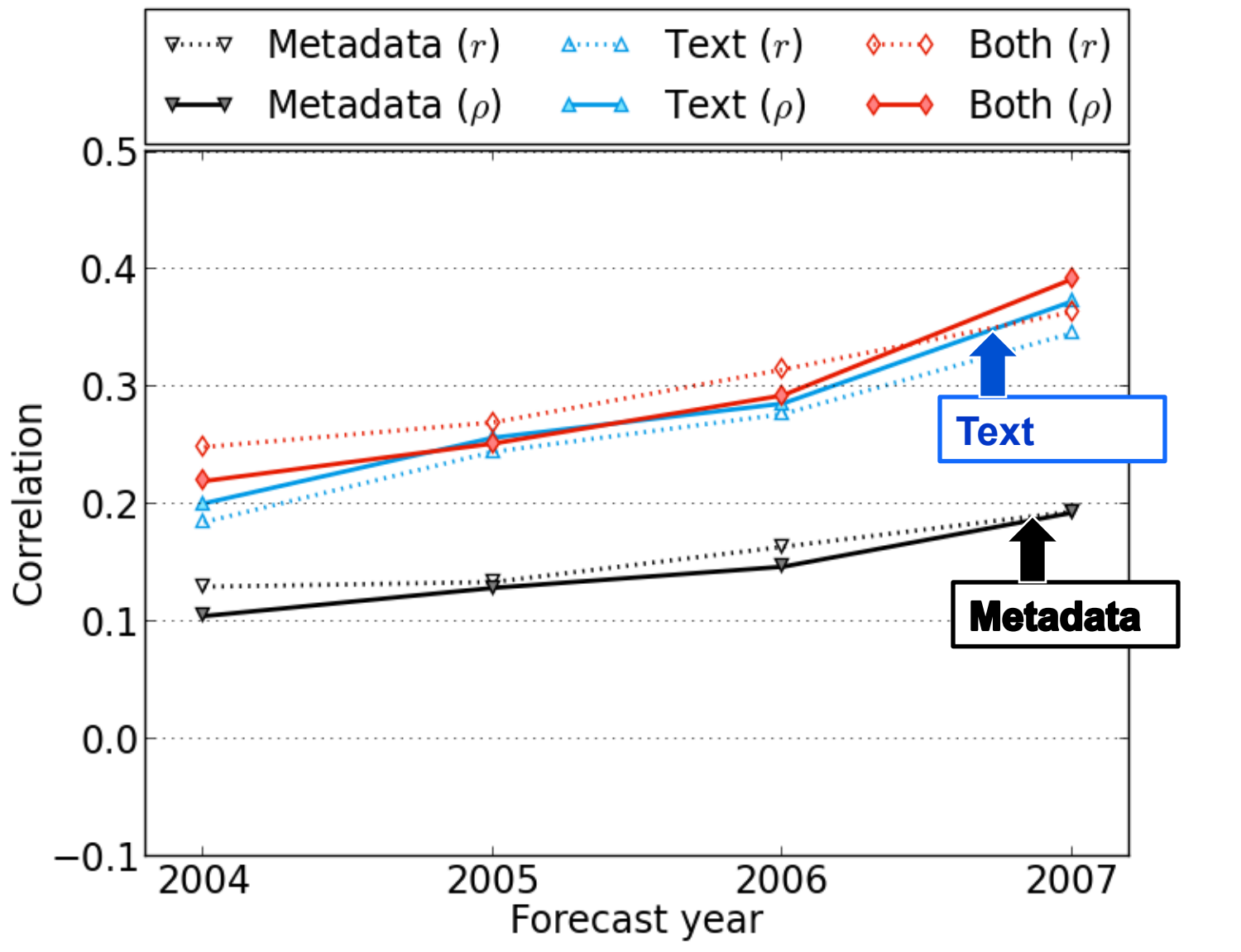
Features drawn from article text

- Rhetorical function (Teufel, 1996)
  - "Here, we present quantitative estimates of the global biological impacts of climate changes." [AIM]
- Citation sentiment and scope
  - "The approach of economists takes a broader view. … We argue that this approach misses biologically important phenomena."

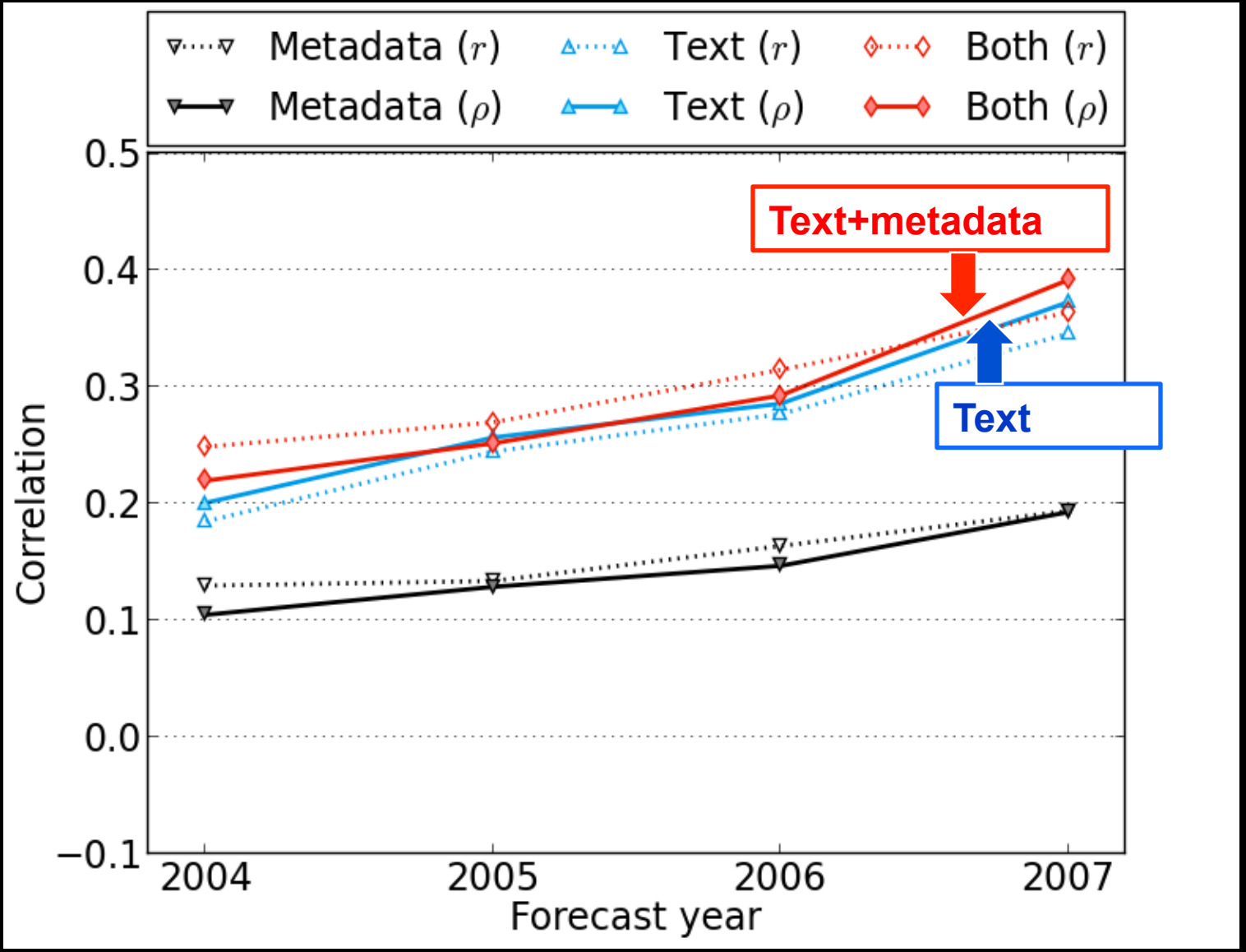
What have we learned?



Text features alone outperform metadata Elsevier data



Metadata adds value when combined with text
Elsevier data

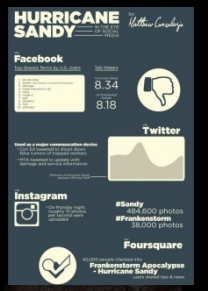


FACT

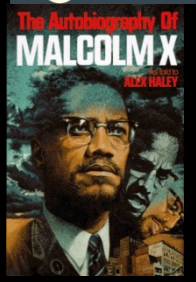
Scientific Journal Articles

News

Online discussion forums

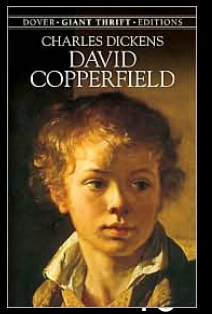


Personal narrative



FICTION

Novels



MONITOR EVENTS OVER TIME

Data

- NIST evaluation
  - hourly web crawl
  - October 2011 - February 2013
  - 16.1TB
- Training Data drawn from Wikipedia

2012 Guatemalan Earthquake – HOUR 1

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 Ch amperico .
People fled buildings in Guatemala City , in Mexico City and in the capital of the Mexican state of Chiapas , across the border from Guatemala .
Would you like to contribute to this story ?
Start a discussion .
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 .
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 buildi ngs 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 the Mexican state of Chiapas , across the border from Guatemala .
Get the biggest news in your email or cellphone as it 's happening .
We welcome your comments on this story , but please be civil .
Do not use profanity , hate speech , threats , personal abuse , images , internet links or any device to draw undue attention .
Read our full comment policy .
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 Pacifi c coast of Guatemala , rocking the capital and shaking buildings as far away as Mexico City and El Salvador .
Investors are dumping stocks as they turn their focus to a world of problems now that the election is over - tax increases and spending cuts that could stall the nation 's recovery and a deepening recession in Europe .
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 .

2012 Guatemalan Earthquake – HOUR 1

The U.S. Pacific Tsunami Warning Center said there was a possibility of a local tsunami, within 100 or 200 miles of the epicenter,

There were no immediate reports of injury or damage .
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 buildi ngs 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 the Mexican state of Chiapas , across the border from Guatemala .
Get the biggest news in your email or cellphone as it 's happening .
We welcome your comments on this story , but please be civil .
Do not use profanity , hate speech , threats , personal abuse , images , internet links or any device to draw undue attention .
Read our full comment policy .
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 Pacifi c coast of Guatemala , rocking the capital and shaking buildings as far away as Mexico City and El Salvador .
Investors are dumping stocks as they turn their focus to a world of problems now that the election is over - tax increases and spending cuts that could stall the nation 's recovery and a deepening recession in Europe .
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 .

2012 Guatemalan Earthquake – HOUR 1

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 Ch amperico .
People fled buildings in Guatemala City , in Mexico City and in the capital of the Mexican state of Chiapas

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

Reuters reported that the quake could be felt as far away as Mexico City .
There were no immediate reports of injury or damage .
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 buildi ngs 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 the Mexican state of Chiapas , across the border from Guatemala .
Get the biggest news in your email or cellphone as it 's happening .
We welcome your comments on this story , but please be civil .
Do not use profanity , hate speech , threats , personal abuse , images , internet links or any device to draw undue attention .
Read our full comment policy .
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 Pacifi c coast of Guatemala , rocking the capital and shaking buildings as far away as Mexico City and El Salvador .
Investors are dumping stocks as they turn their focus to a world of problems now that the election is over - tax increases and spending cuts that could stall the nation 's recovery and a deepening recession in Europe .
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 .

System Update

- 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.

Features to Predict Salience

Language Models (5-gram Kneser-Ney model)

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

Features to Predict Salience

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

Features to Predict Salience

Geographic Features

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

Features to Predict Salience

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.

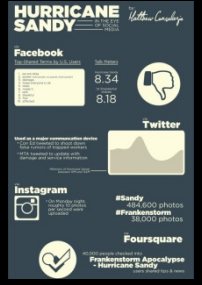
Medium Salience

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

What have We Learned?



Salience features improve results

We are not yet able to model redundancy well

ROUGE-2	
Features	F-measure
All	.049
No Language Model	.038
No Geographical	.042
No Redundancy	.056

Kedzieet al, ACL 2015

SubEvent Identification

Decompose articles on a main event into related sub-events:

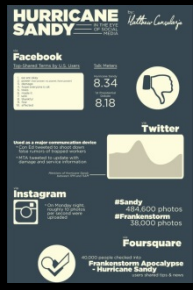
Hurricane Sandy

Manhattan Blackout

Breezy Point fire

Public Transit Outage

Going Forward: Social Media



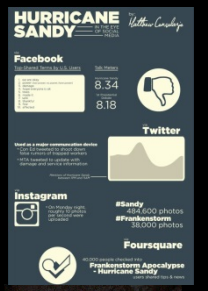
- Drawing what happened from social media
  - It's dark.There is minor price gouging. There are restaurants selling hot food through their bay windows.
- How do people feel about impending storms?
  - Excited, scared, nervous, blasé
- In collaboration with social scientists
  - How does this impact preparedness?
  - Correlate news reports with reactions across events
  - What language in news engenders a reaction that helps people prepare?

FACT

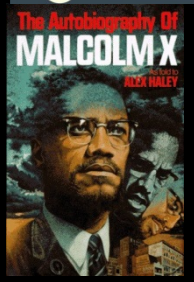
Scientific Journal Articles

News

Online discussion forums

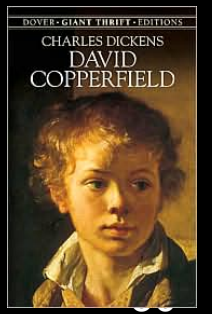


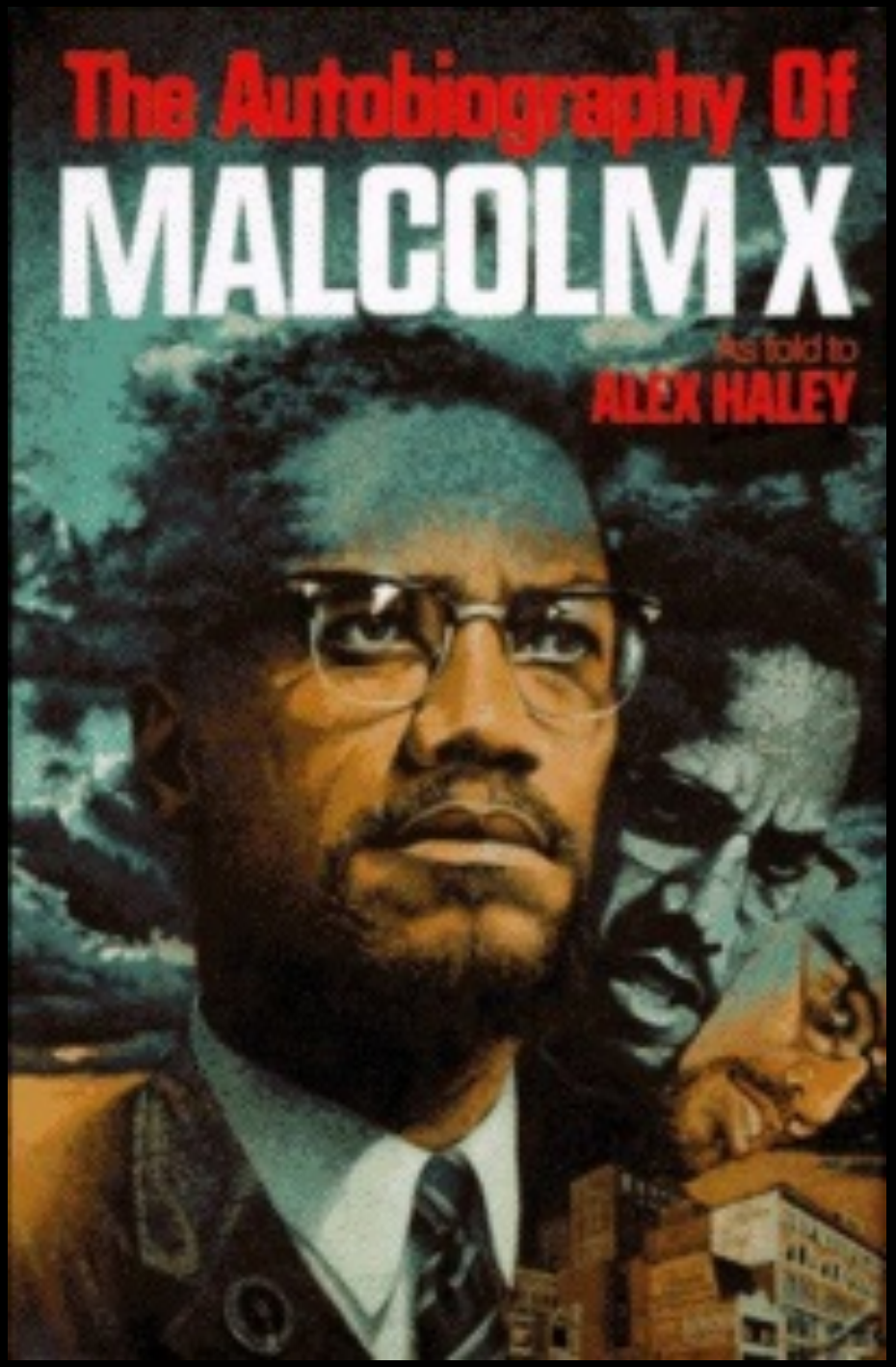
Personal narrative



FICTION

Novels




The Autobiography Of
MALCOLM X
As told to
ALEX HALEY

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 [illegible] 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 down the block. I saw a trickle of steady he street on both sides and then water 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?”

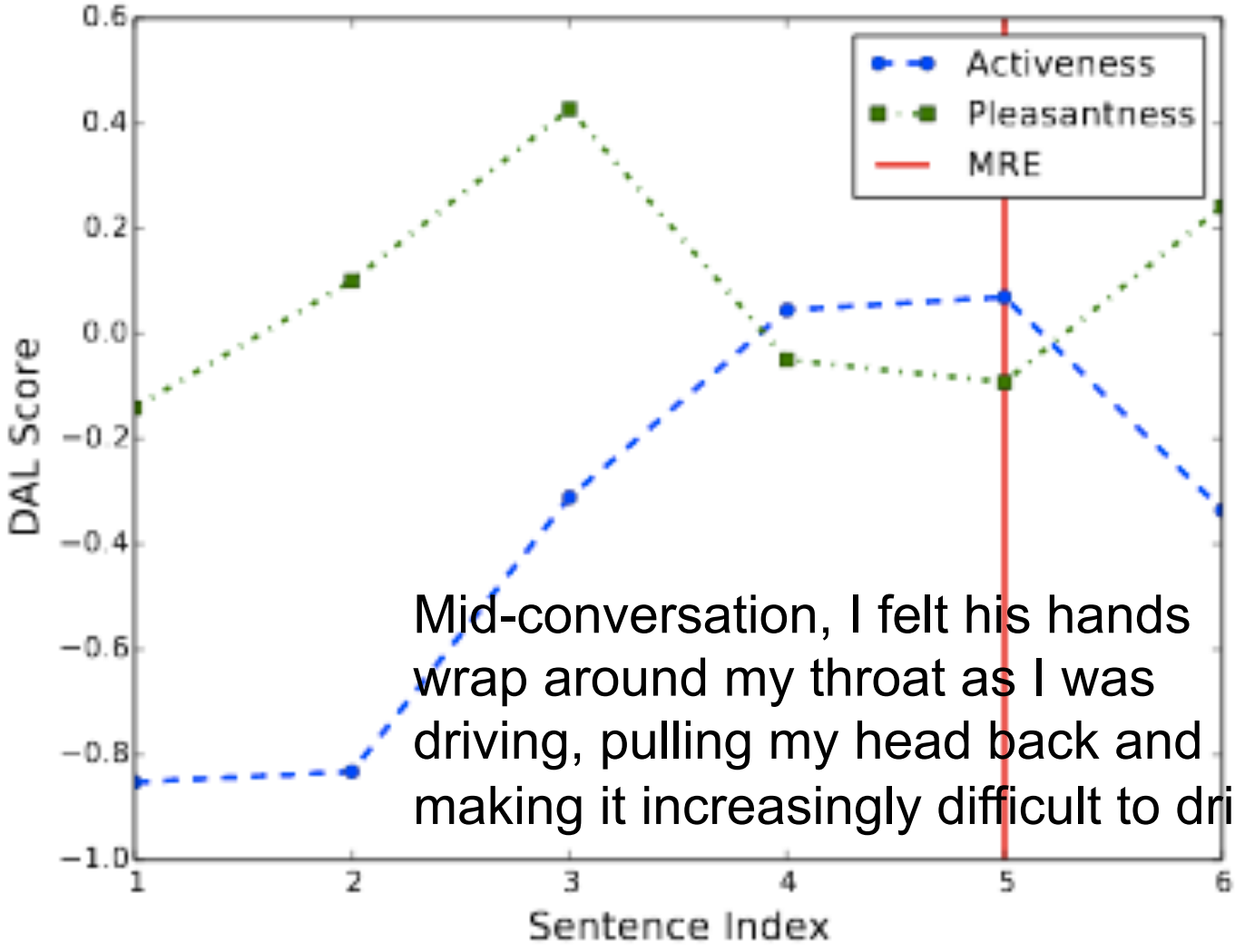
Data

- 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

Features: Change in Affect



Mid-conversation, I felt his hands wrap around my throat as I was driving, pulling my head back and making it increasingly difficult to drive.

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**

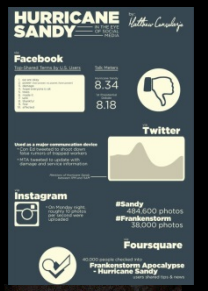
Ouyang &McKeown, EMNLP 2015

FACT

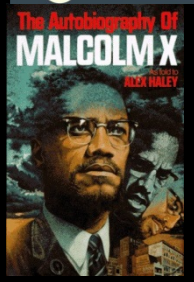
Scientific Journal Articles

News

Online discussion forums

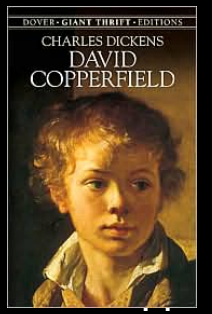


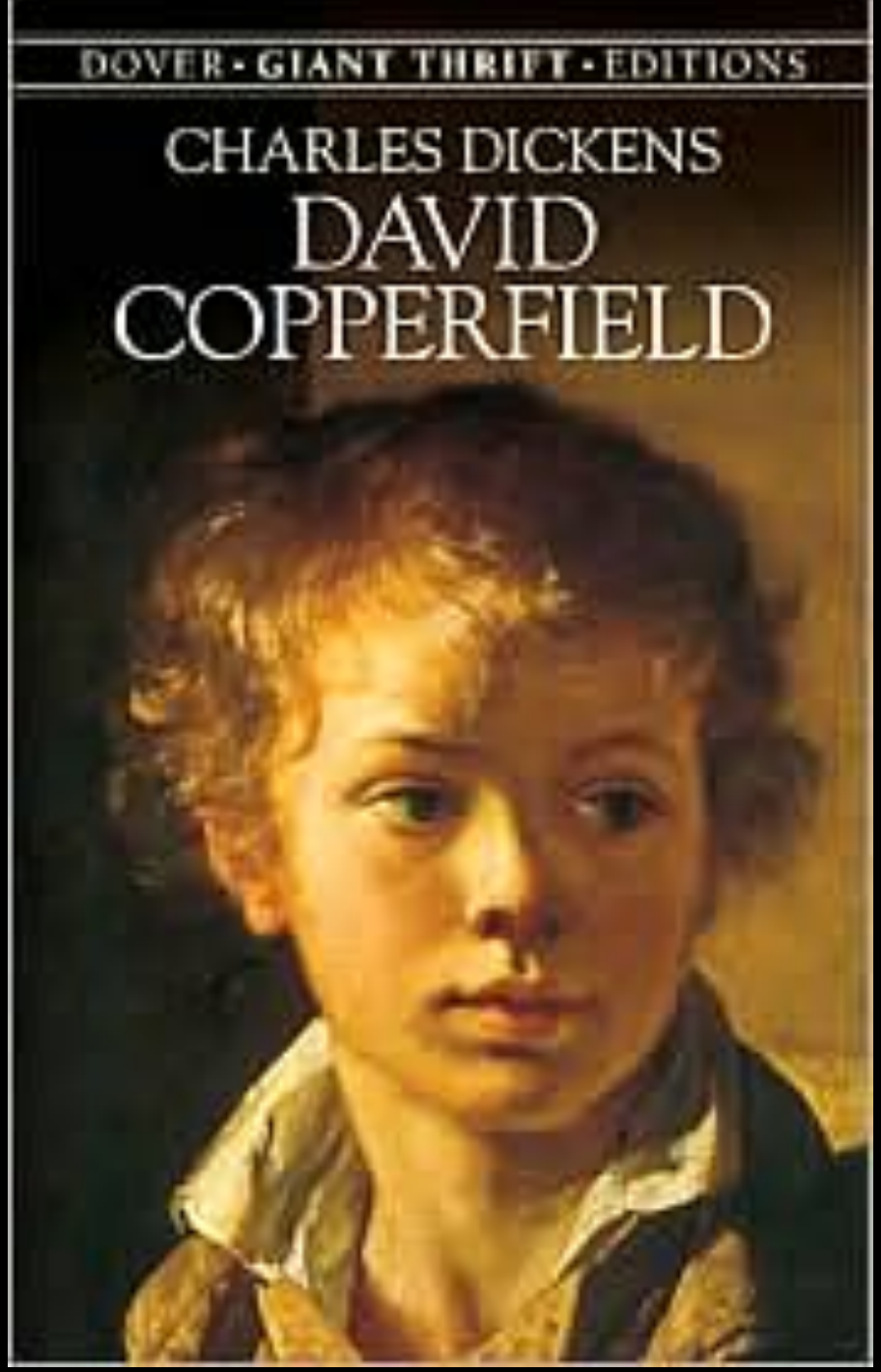
Personal narrative



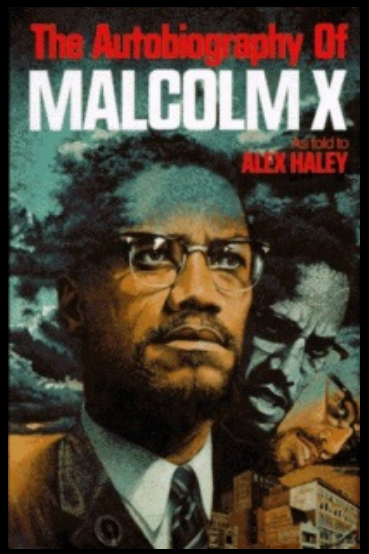
FICTION

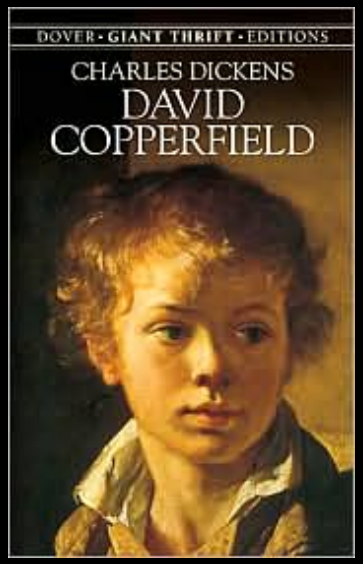
Novels




DOVER · GIANT THRIFT · EDITIONS
CHARLES DICKENS
DAVID
COPPERFIELD

Data Science Yields Solutions





Across disciplines

Current PhD Students

Or Biran

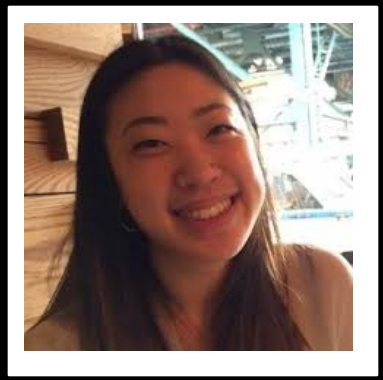
Noura Farra

Chris Hidey

Chris Kedzie

Jessica Ouyang

Yves Petinot



Melody Ju

Past Students


Regina Barzilay


Sasha Blair-Goldensohn


Andrea Danyluk


Galina Datskovsky Moerdler


Pablo Duboue


Michael Elhadad


Noemie Elhadad


David Elson


David Evans

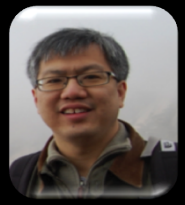

Elena Filatova


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


Sara Rosenthal


Kapil Thadani

Thank You!

- The research presented here has been supported in part by DARPA BOLT, DARPA DEFT, IARPA FUSE, IARPA SCIL and NSF.