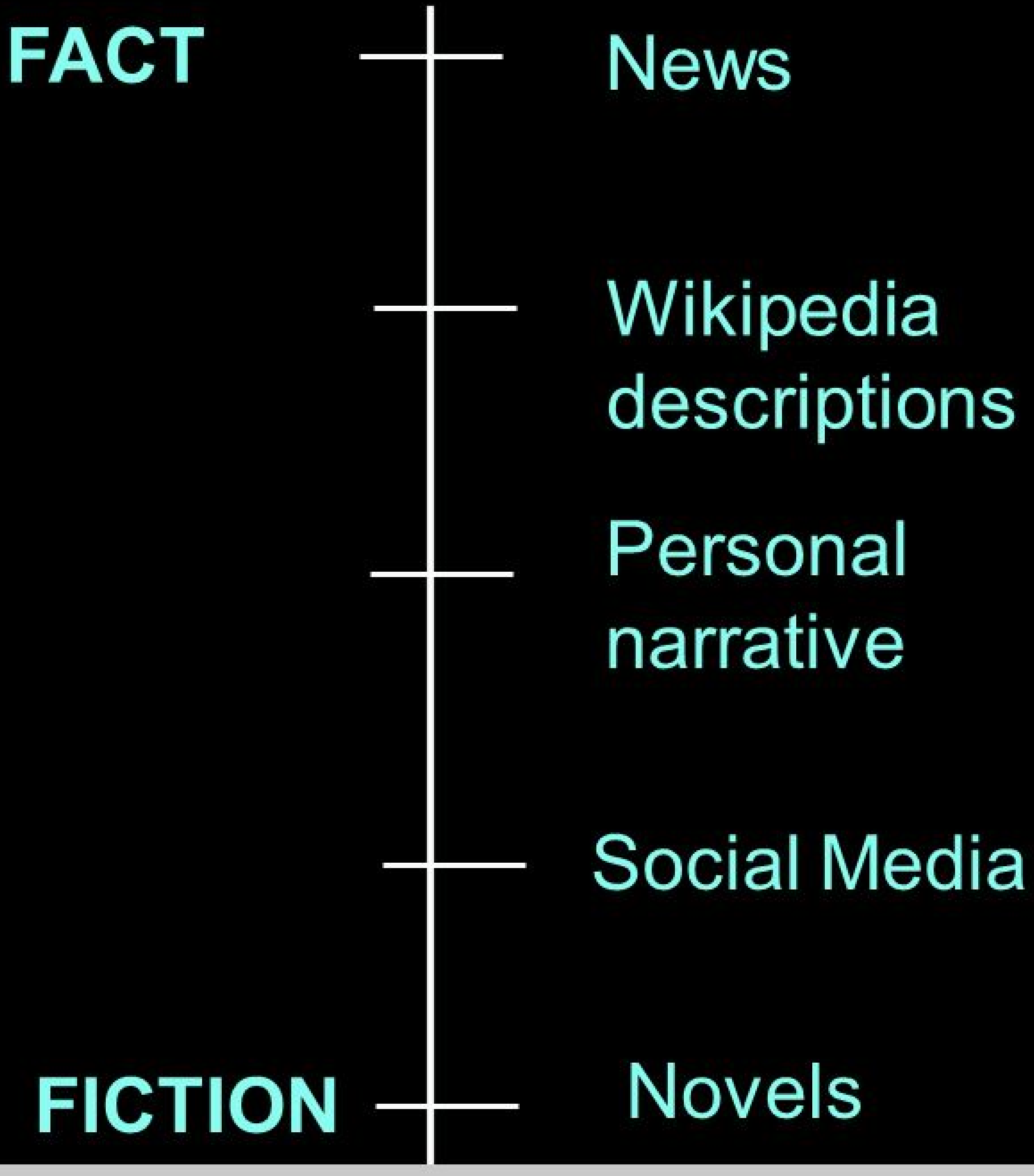
At the Intersection of Language and Data Science

Kathleen McKeown
Department of Computer Science
Columbia University

DATA SCIENCE INSTITUTE
NEW MEDIA FOR DATA SCIENCE

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



FACT

News

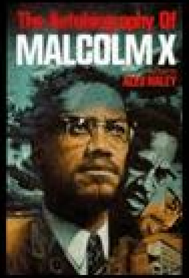
Wikipedia descriptions

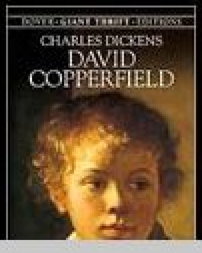
Personal narrative

Social Media

FICTION

Novels





Why is Summarization Hard?

- Seems to require both interpretation and generation of text
- Handle input documents from unrestricted domains robustly
- Operate without full semantic interpretation

Leads many summarization researchers to use sentence selection

Our Approach

- Edit selected sentences
  - Correct infelicitous references

    Nenkova et al,, HLT 2005
  - Remove extraneous material through **compression**

    Jing, ANLP 2000, Galley & McKeown, NAACL 2007
  - Make fluent sentences from disfluent **translated** sentences

    Siddharthan & McKeown, HLT 2005, Parton et al, EAMT 2012
- Generate new sentences from selected phrases through **fusion**

  Barziilay&McKeown 2005

Walking a fine line: it's easy to make a good sentence bad

Text to Text Generation

Model text transformation as a *structured prediction* problem

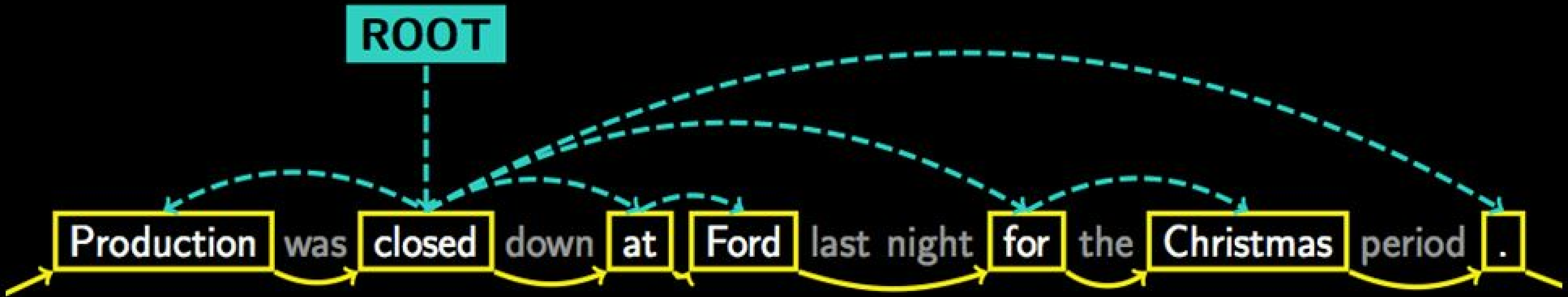
- Input: One or more sentences with parses
- Output: Single sentence + parse

Joint inference over

- **word choice**,
- **n-gram ordering**
- **dependency structure**

Sentence Compression

- Input: single sentence
- Output: sentence with **salient** information
- Dataset + baseline from *Clarke & Lapata (2008)*

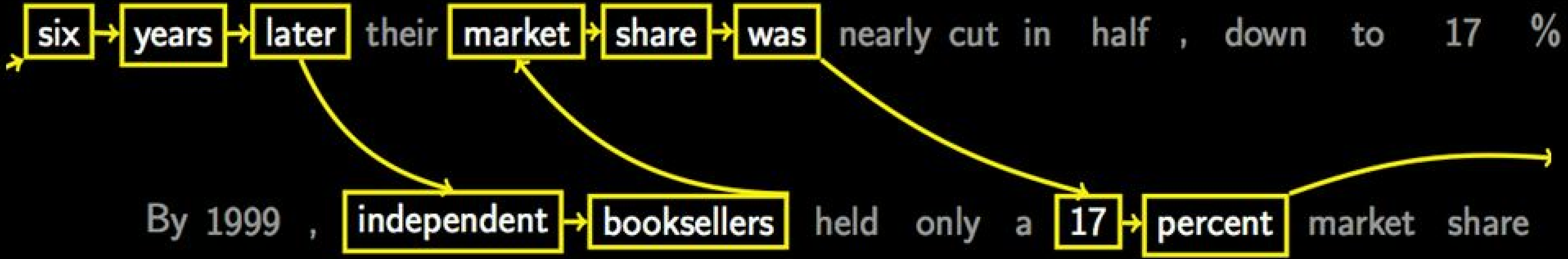


ROOT

Production was closed down at Ford last night for the Christmas period .

Sentence Fusion

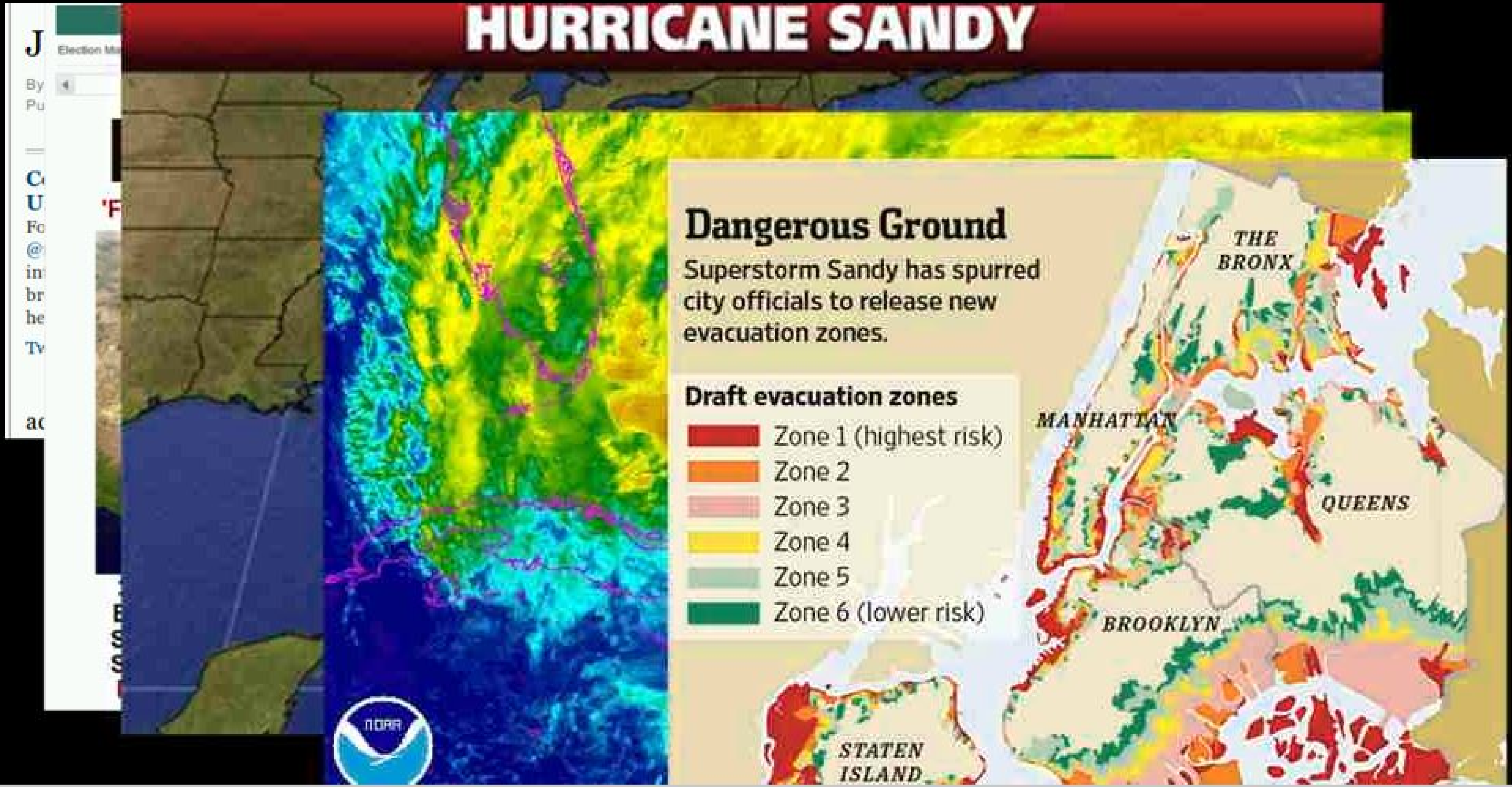
- Input: multiple sentences
- Output: sentence with ***common*** information
- Dataset created from summarization evaluations
- Fusion-specific features, e.g., repetition



six → years → later their market → share → was nearly cut in half , down to 17 %

By 1999 , independent → booksellers held only a 17 → percent market share

Problem: Identifying needs during disaster



Dangerous Ground

Superstorm Sandy has spurred city officials to release new evacuation zones.

Draft evacuation zones

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 language model scores sentences by how typical they are of the language – higher scores mean more fluent

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