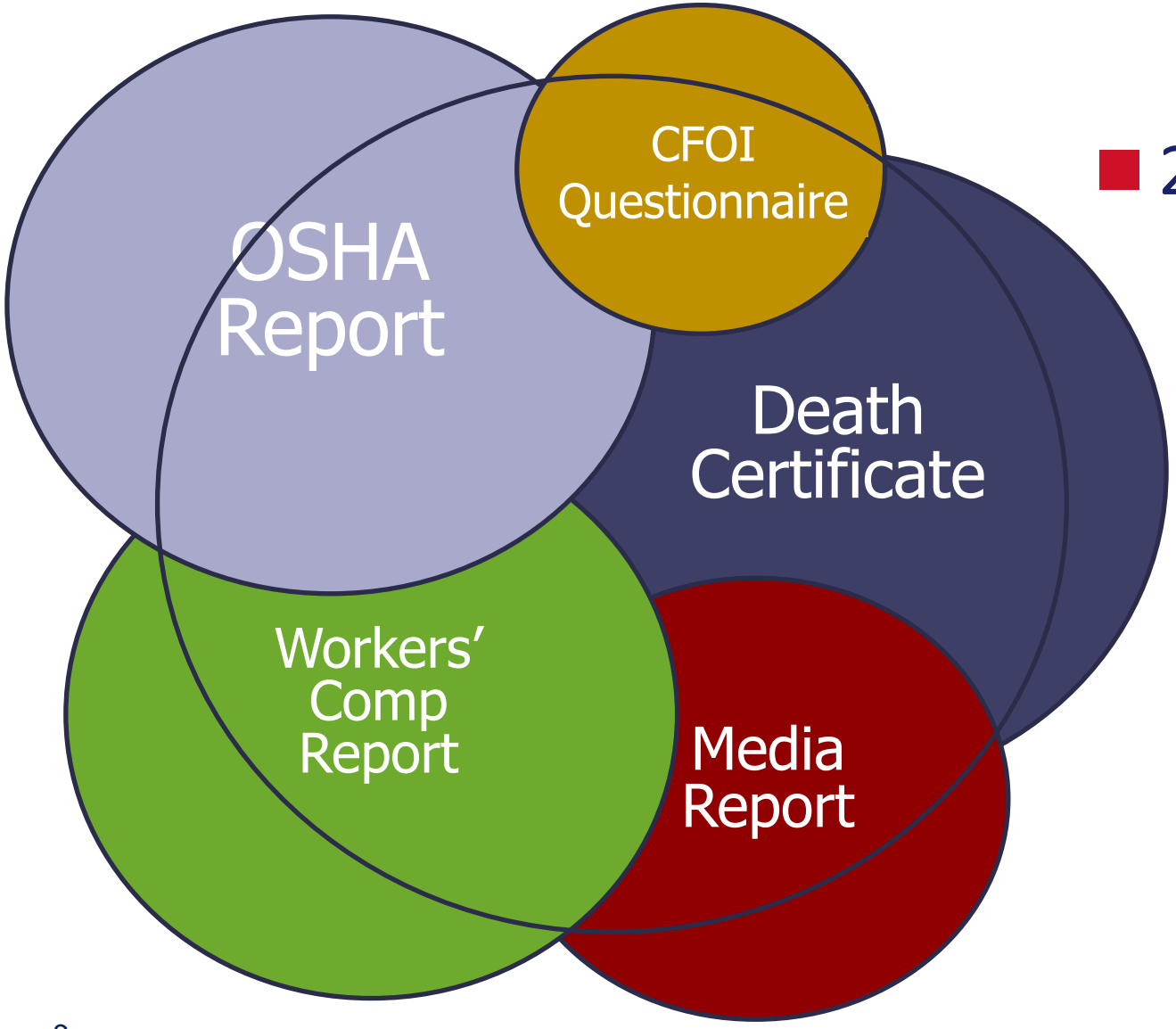
# Matching fatal injury records with machine learning

Alex Measure

# Census of Fatal Occupational Injuries

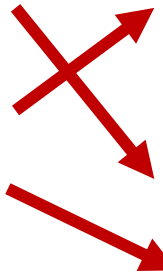



- 2019 data
  - 25,000 documents
  - 5,333 cases

# Record Matching Problem

## Census file

| Person | Company | Age | Narrative |
|---|---|---|---|
| John Smith | ACME Inc. | 25 | Car accident |
| Susan Carter | Tree Co. | 74 | Hit by tree |
| Hank Long | Big Box | 34 | Homicide |



## Source document file

| Person | Company | Union | Industry |
|---|---|---|---|
| Suzy E. Carter | Joe's Trees | Yes | 124000 |
| Frank Garcia | Cola Co. | No | 332000 |
| Jonathan Smith | A.C.M.E. | No | 429000 |
| Henry Long | BB Retail | Yes | 620000 |

# Record Matching Goal

## Census file (expanded)

| Person | Company | Age | Narrative | Union | Industry |
|---|---|---|---|---|---|
| John Smith | ACME Inc. | 25 | Car accident | No | 429000 |
| Susan Carter | Tree Co. | 74 | Hit by tree | Yes | 124000 |
| Hank Long | Big Box | 34 | Homicide | Yes | 620000 |
| Frank Garcia | Cola Co. | -- | -- | No | 332000 |

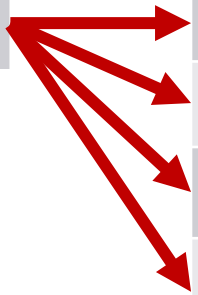
# Basic Idea: Compare Everything

## Census file

| Person | Company | Age | Narrative |
|---|---|---|---|
| John Smith | ACME Inc. | 25 | Car accident |

## Source document file

| Person | Company | Union | Industry |
|---|---|---|---|
| Suzy E. Carter | Joe's Trees | Yes | 124000 |
| Frank Garcia | Cola Co. | No | 332000 |
| Jonathan Smith | A.C.M.E. | No | 429000 |
| Henry Long | BB Retail | Yes | 620000 |



Options

- Simple rules (e.g. exact match)
  - Works poorly without unique ID's
- Probabilistic model – how do you build it?
  - We have some training data!!!
  - Supervised machine learning

# Step 1: Assemble Training Data

## Census file

| Person | Company | Age |
|---|---|---|
| John Smith | ACME Inc. | 25 |
| John Smith | ACME Inc. | 25 |
| John Smith | ACME Inc. | 25 |
| John Smith | ACME Inc. | 25 |
| Susan Carter | Tree Co. | 74 |
| Susan Carter | Tree Co. | 74 |
| … | … | … |

## Source document

| Person | Company | Age |
|---|---|---|
| Suzy E. Carter | Joe's Trees | 73 |
| Frank Garcia | Cola Co. | 32 |
| Jonathan Smith | A.C.M.E. | 26 |
| Henry Long | BB Retail | 47 |
| Suzy E. Carter | Joe's Trees | 73 |
| Frank Garcia | Cola Co. | 32 |
| … | … | … |

## Similarity metrics

| Name Sim | Age Diff | Distance | Match |
|---|---|---|---|
| 0.05 | 48 | 2 | False |
| 0.04 | 7 | 500 | False |
| 0.79 | 1 | 5 | True |
| 0.12 | 22 | 10 | False |
| 0.84 | 1 | 10 | True |
| 0.01 | 42 | 750 | False |
| … | … | … | … |

TF-IDF cosine similarity on 4 character n-grams

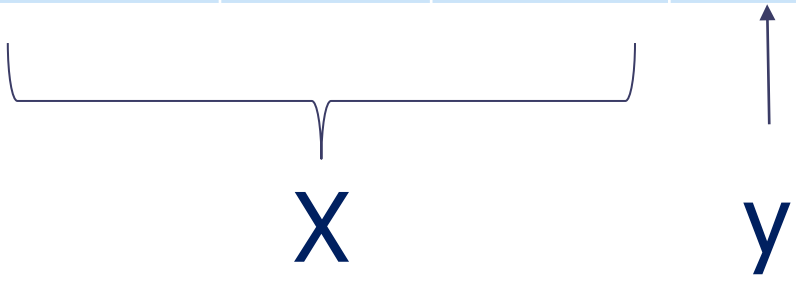
Geocoding

Manual Matches

# Step 2: Fit the Model

## Training inputs

| Name Sim | Age Diff | Distance | Match |
|---|---|---|---|
| 0.05 | 48 | 2 | False |
| 0.04 | 7 | 500 | False |
| 0.79 | 1 | 5 | True |
| 0.12 | 22 | 10 | False |
| 0.84 | 1 | 10 | True |
| 0.01 | 42 | 750 | False |
| … | … | … | … |




## Python code

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier()
clf.fit(X=X, y=y)
```

# Step 3: Use the Model

Model output

| Census file | | | Source document | | | Similarity metrics | | | |
|---|---|---|---|---|---|---|---|---|---|
| Person | Company | | Person | Company | Age | Name Sim | Age Diff | Distance | Match |
| Lisa Miller | M Auto | 43 | Kevin E. Davis | High School | 28 | 0.02 | 15 | 200 | False |
| Lisa Miller | M Auto | 43 | Andy Williams | J Food Inc. | 32 | 0.04 | 9 | 20 | False |
| Lisa Miller | M Auto | 43 | Kevin Miller | Fun Bar | 26 | 0.60 | 17 | 800 | False |
| Lisa Miller | M Auto | 43 | Lisa G. Miller | Miller Auto | 45 | 0.96 | 2 | 50 | True |
| Kevin Davis | Spring HS | 28 | Kevin E. Davis | High School | 28 | 0.92 | 0 | 50 | True |
| Kevin Davis | Spring HS | 28 | Andy Williams | J Food Inc. | 32 | 0.01 | 4 | 150 | False |
| … | … | … | … | … | … | … | … | … | … |

TF-IDF cosine similarity on 4 character n-grams

Geocoding

# Did it work?

- Scenario:
  - OSHA matching
  - No decedent or company name!
  - Previously matched by hand (but not used to train model)
- Results:
  - Model recovers 92% of manual matches
  - Where model predicts match and records show none:
    - 88% appear to be real matches that were missed!
- It works very well, now part of production

# News articles

- Challenge identifying basic info

Paul Novicki, 51, a retired first responder for Huron Township, Mich., died of covid-19.

Both Novicki and his EMT partner, Rob Nemeth, contracted the virus. Nemeth survived his bout. Novicki, 51, did not.

Most of Novicki's career was spent with the Huron Township Fire Department, from which he had recently retired. He continued working with a private ambulance company until he became ill.

Novicki brushed off his initial coronavirus symptoms as allergies. He didn't want anyone to worry about him.

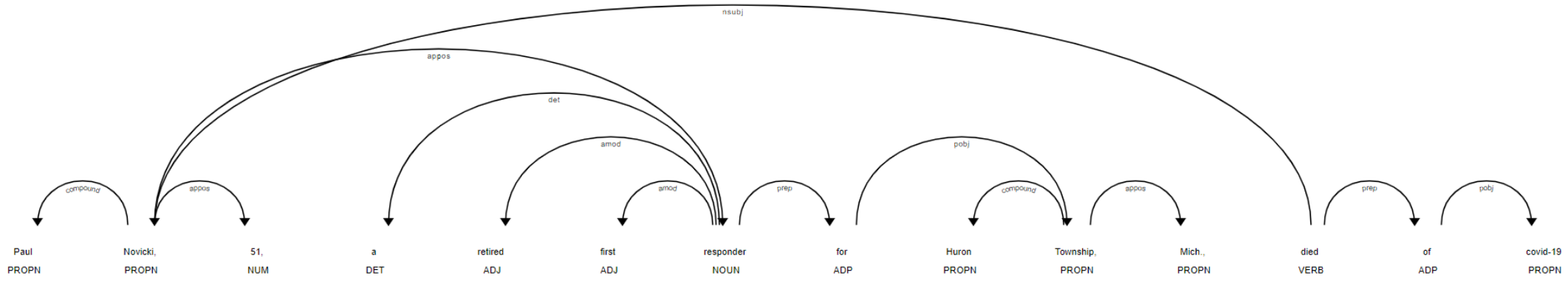
But soon, Novicki was placed on a ventilator. Otherwise healthy, he died on April 9. Local fire and police departments quickly organized a parade in his honor. "Taps" was played.

# The Traditional Way

- Named Entity Recognition

Paul Novicki PERSON , 51 DATE , a retired first ORDINAL responder for Huron Township GPE , Mich. GPE , died of covid-19

- Dependency Trees (and other parsers)




- Rules or Model to Tie Pieces Together
- Very Challenging

# A New Way: Extractive Question Answering

Q: What is the dead worker's name?

A: Paul Novicki

Q: What is the Paul Novicki's age?

A: 51

Q: What caused Paul Novicki's death?

A: covid-19

Q: On what date did Paul Novicki die?

A: April 9

Code example at https://bit.ly/2WfPbBe

Paul Novicki, 51, a retired first responder for Huron Township, Mich., died of covid-19.

Both Novicki and his EMT partner, Rob Nemeth, contracted the virus. Nemeth survived his bout. Novicki, 51, did not.

Most of Novicki's career was spent with the Huron Township Fire Department, from which he had recently retired. He continued working with a private ambulance company until he became ill.

Novicki brushed off his initial coronavirus symptoms as allergies. He didn't want anyone to worry about him.

But soon, Novicki was placed on a ventilator. Otherwise healthy, he died on April 9. Local fire and police departments quickly organized a parade in his honor. "Taps" was played.

# Contact Information

**Alex Measure**
202-691-6185
measure.alex@bls.gov
http://www.bls.gov/iif/autocoding.htm