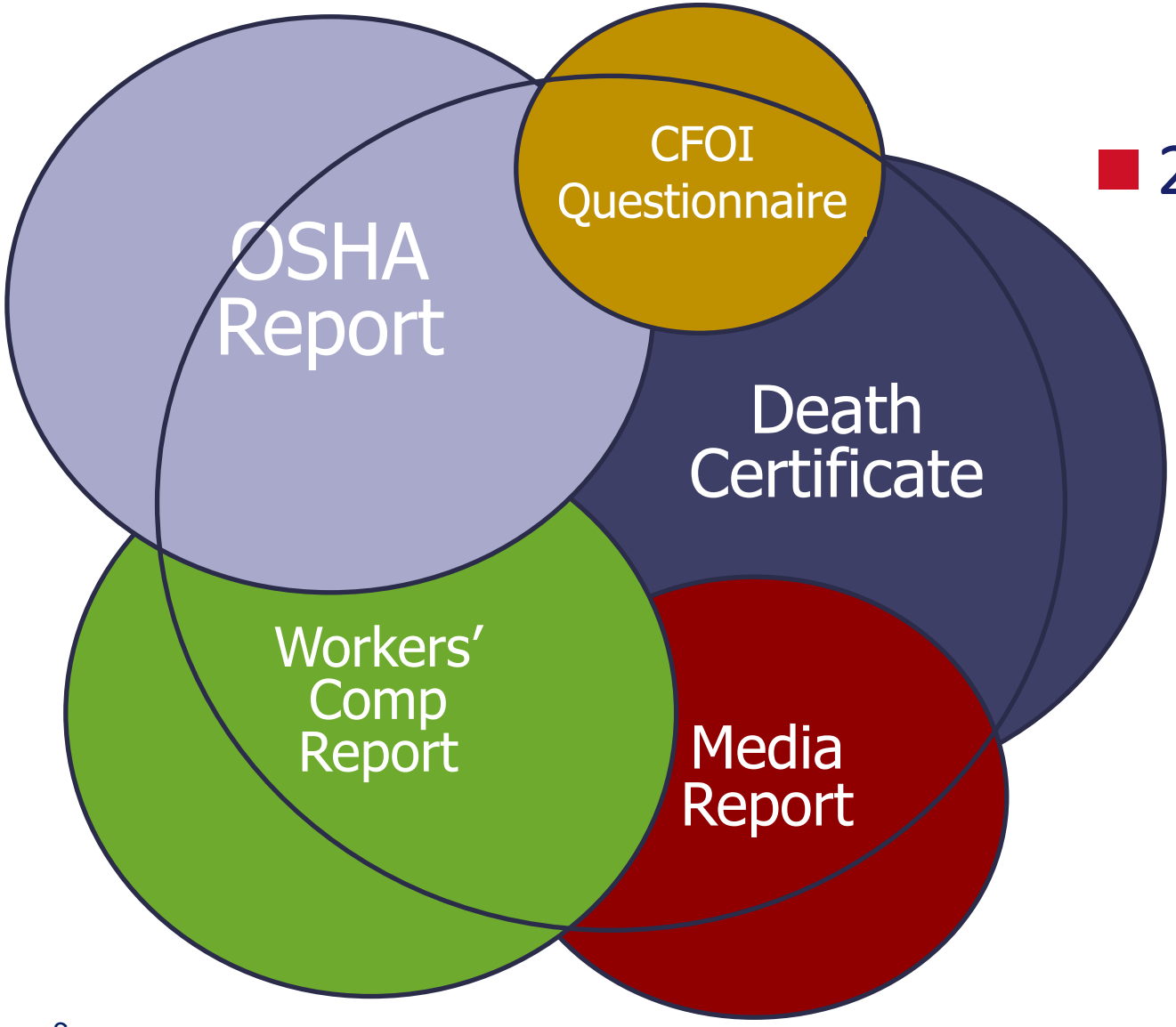
Matching fatal injury records with supervised machine learning

Alex Measure

Census of Fatal Occupational Injuries



- 2018 data
  - Almost 25,000 SDs
  - 5,250 cases
  - OSHA reports supported 2,026 cases in 2018

Record Matching Problem

CFOI case 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

OSHA inspection 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

Added Difficulty, Missing Names!

CFOI case file

Person	Company	Age	Narrative
XXXXXXX	XXXXXX	25	Car accident
Susan Carter	Tree Co.	74	Hit by tree
XXXXXXX	XXXXXX	34	Homicide

OSHA inspection 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

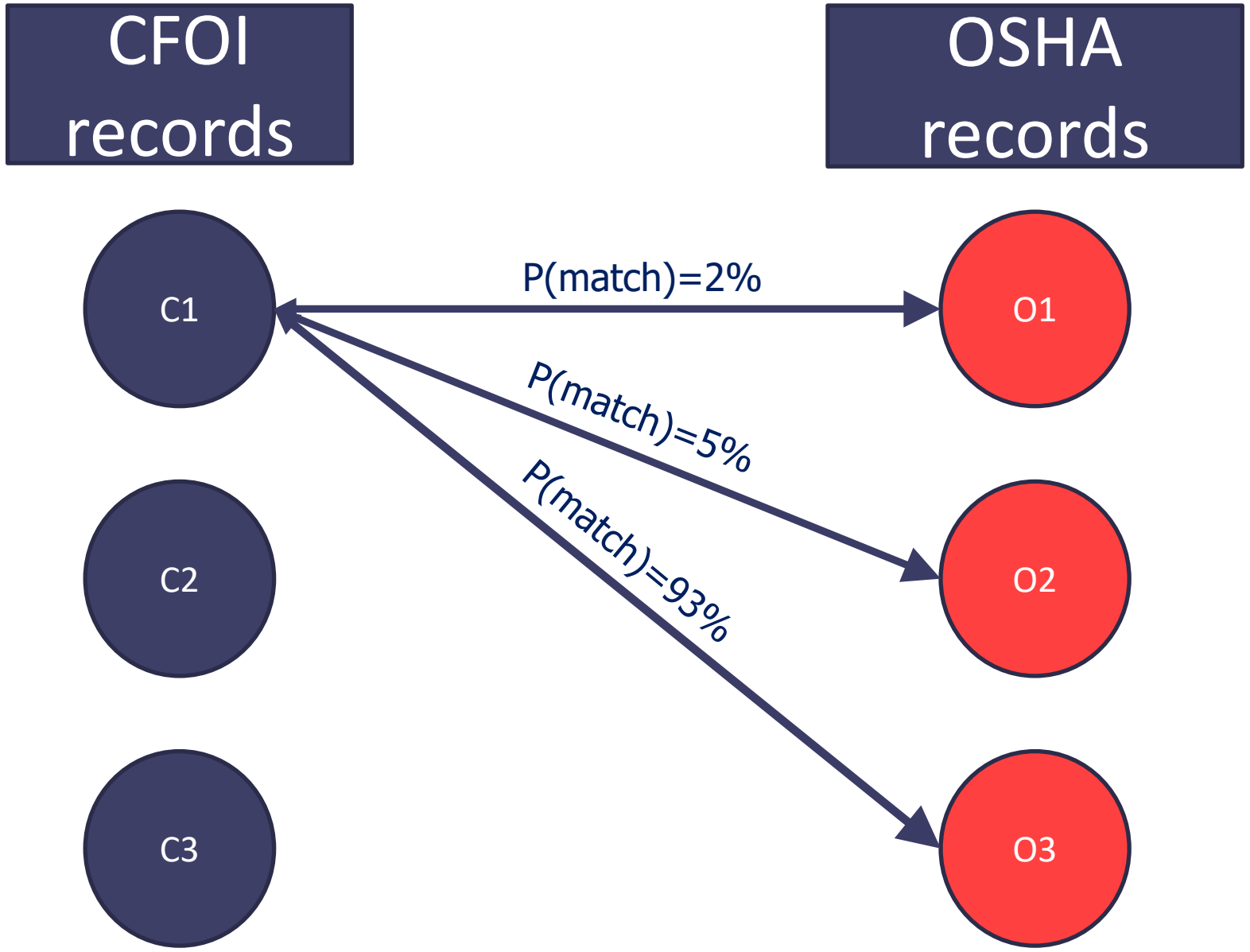
What else is there?

CFOI case file

Date of Death	Zip Code	Age	Narrative
12-02	53241	25	Car accident
4-15	20043	74	Hit by tree
9-24	92150	34	Homicide

OSHA inspection file

Time of Incident	City	Age	Narrative
4-10	Green Bay	70	Tree
7-29	NYC	No	Heart
12-01	Chicago	27	Hit by car
9-24	Miami	34	Gunshot


CFOI records
OSHA records
C1
C2
C3
O1
O2
O3
P(match)=2%
P(match)=5%
P(match)=93%

How do you estimate P(match)?

How similar are the records?

- Name: "**John Smith**" more likely to match "**J Smith**" than "**Henry Jay**"
- Age: "**57**" more likely to match with "**56**" than "**27**"
- Geography: "**NYC**" more likely to match "**Manhattan**" than "**Dallas**"
- Incident: "**Traffic accident**" more likely to match "**Car accident**" than "**Fire**"
- Date: "**Jan 25th, 2019**" more similar to "**1-24-2019**" than to "**2-25-2018**"

How important are the similarities?

- "Name" similarity might be more important than "Age" similarity, except when "Name" is not available

How to combine info?

- Fellegi-Sunter
  - For each pair of similar fields in datasets
  - Calculate similarity
  - Calculate P(match | similarity)
  - Overall P = product of individual P (assumption of independence)
- Many alternatives available today
  - Random Forest

Does it work?

- When CFOI records indicate a match:
  - 92%
- When CFOI records show nothing, but program says “match”?
  - ~88% appear to be real matches that were missed
- This is without even using decedent / company name info!

Contact Information

Alex Measure
202-691-6185
measure.alex@bls.gov