Module 1: Introduction to Machine Learning and Entity Resolution

Rebecca C. Steorts

Reading: Binette and Steorts (2022)

August 20, 2024

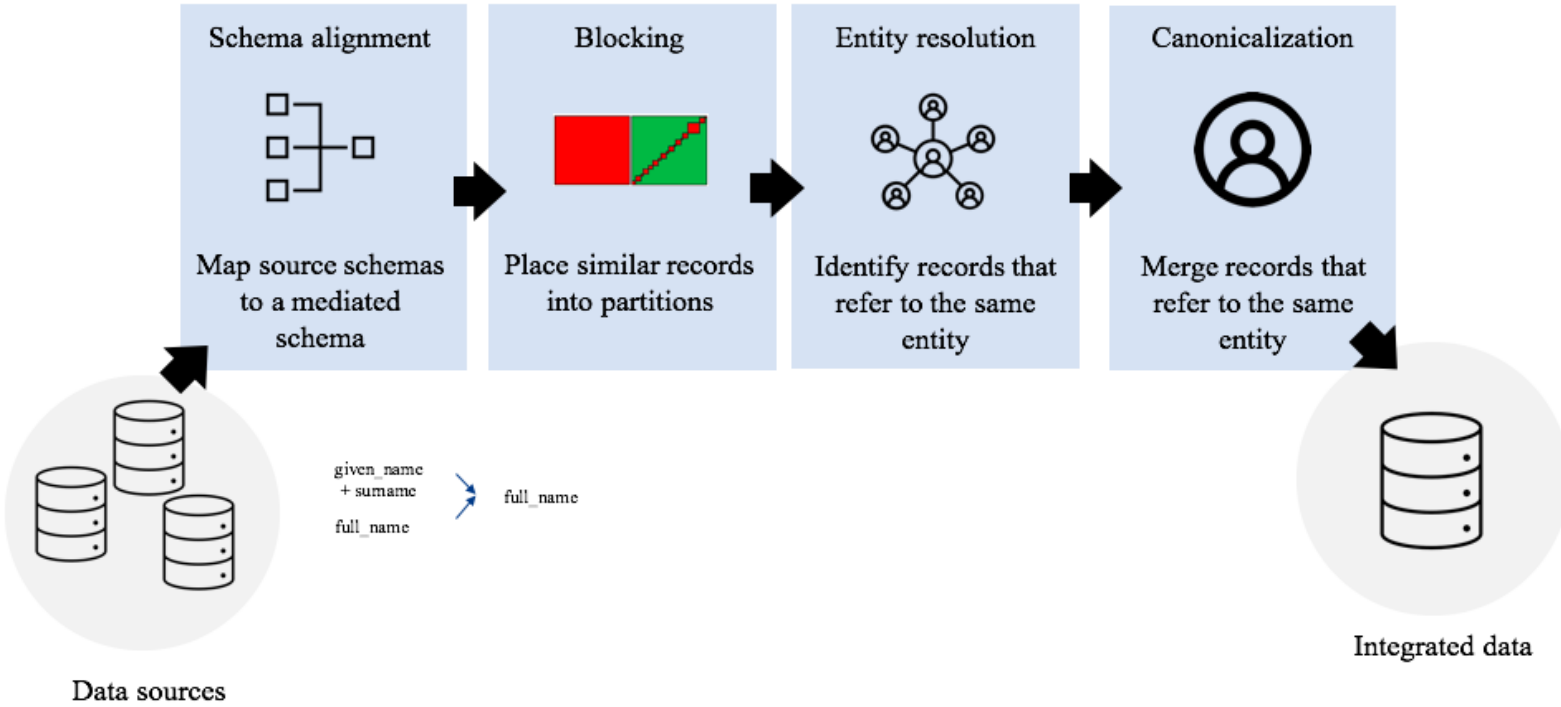
"Statistics is the science of learning from data. Machine Learning (ML) is the science of learning from data. These fields are identical in intent although they differ in their history, conventions, emphasis and culture."

- Larry Wasserman, Rise of the Machines

What are some examples that you have learned of machine learning in prior classes, internships, or elsewhere?

- Machine learning and statistics have much to learn from each other.
- In this course, we will focus on machine learning, whereas in other courses, you will learn the fundamentals of statistics.
- To be successful, you need both insights and perspectives. (As a follow up class, take Cynthia Rudin's machine learning course).

In this course, we're going to focus on one type of practical, applied machine learning that appears widely in both industry and academia, known as "data cleaning."


Schema alignment
Map source schemas to a mediated schema
Blocking
Place similar records into partitions
Entity resolution
Identify records that refer to the same entity
Canonicalization
Merge records that refer to the same entity
given_name + surname
full_name
full_name
Data sources
Integrated data

Let's consider a simple data cleaning application.

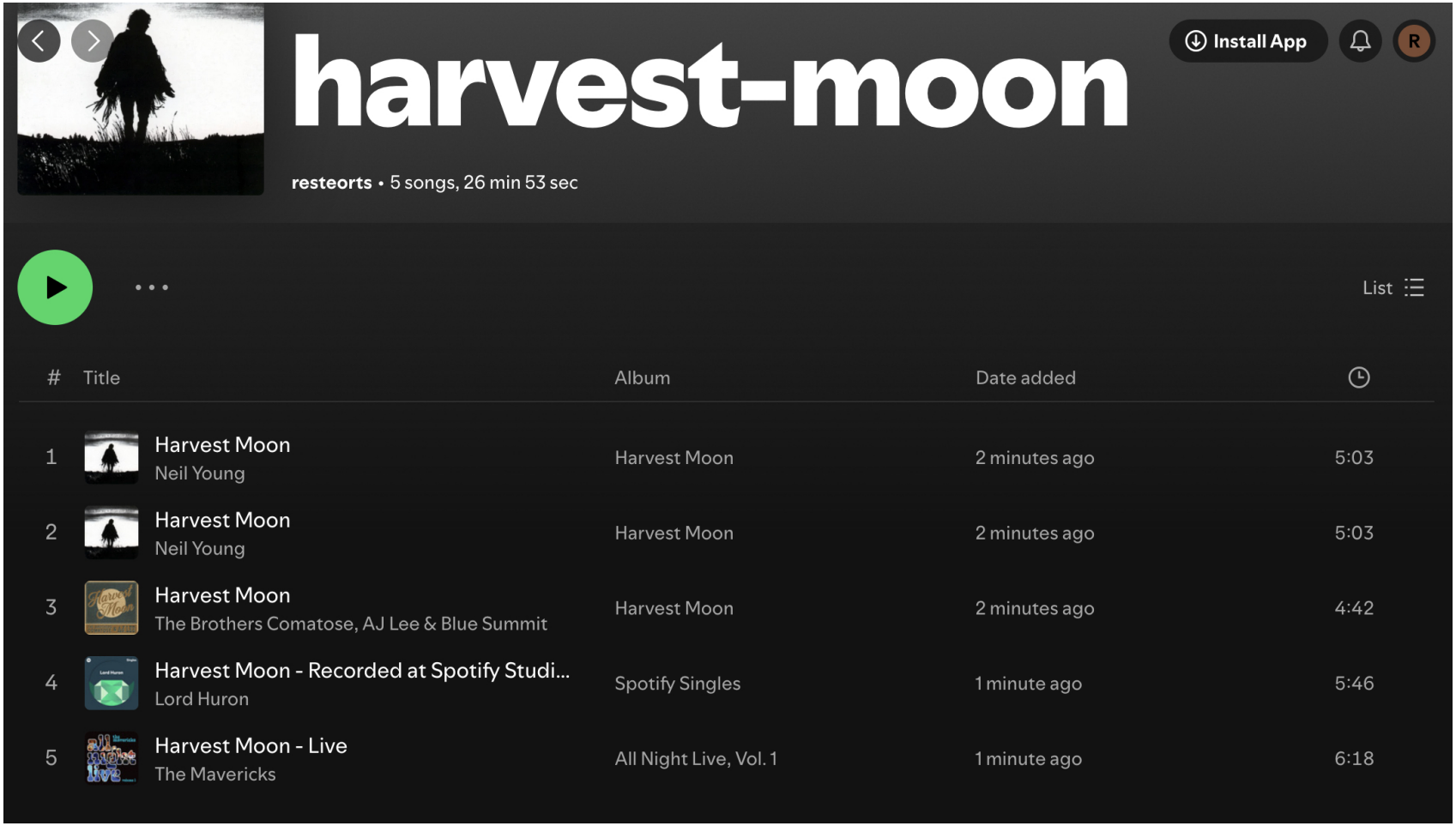


Figure: How many unique songs can you identify visually?

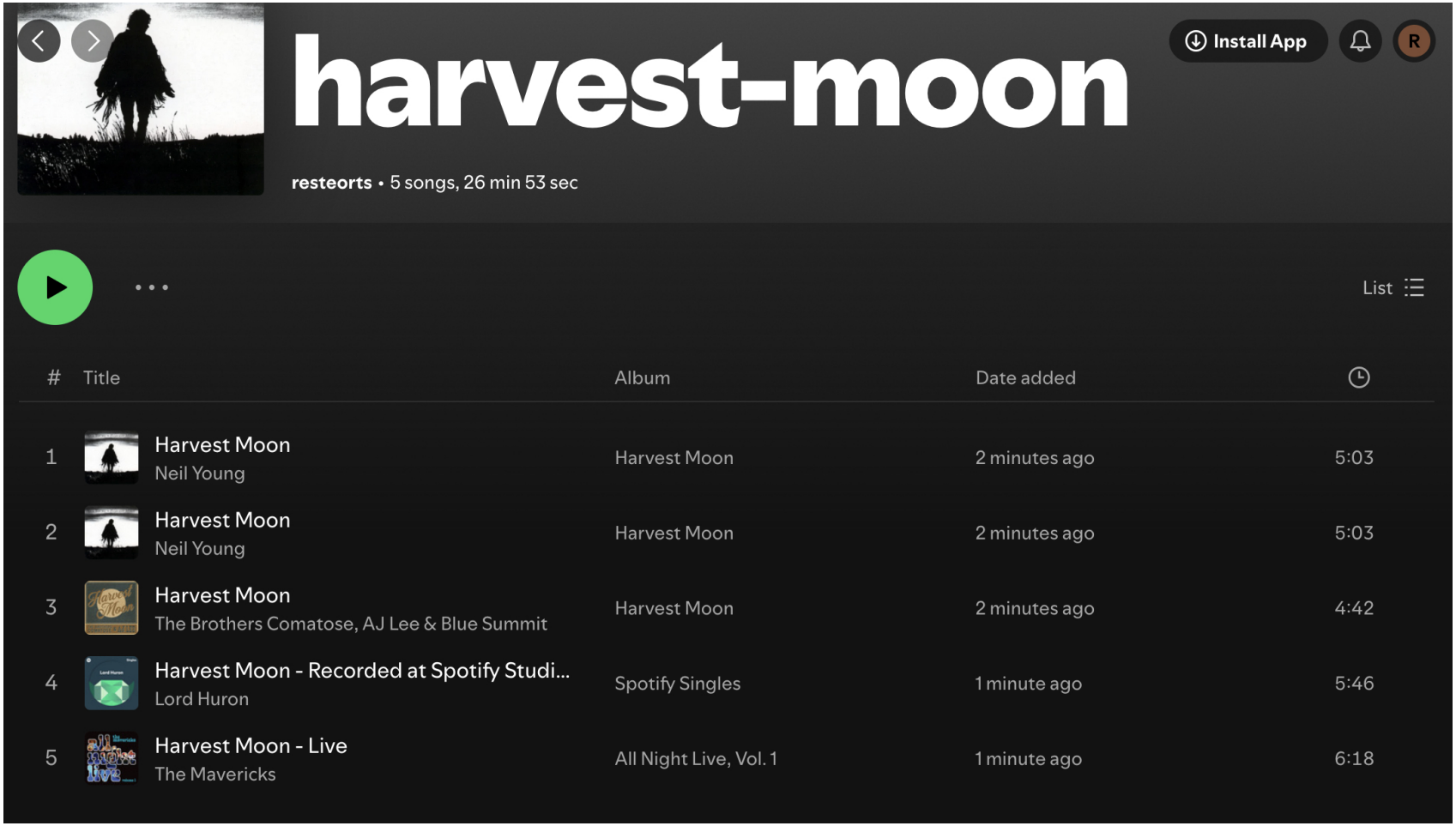
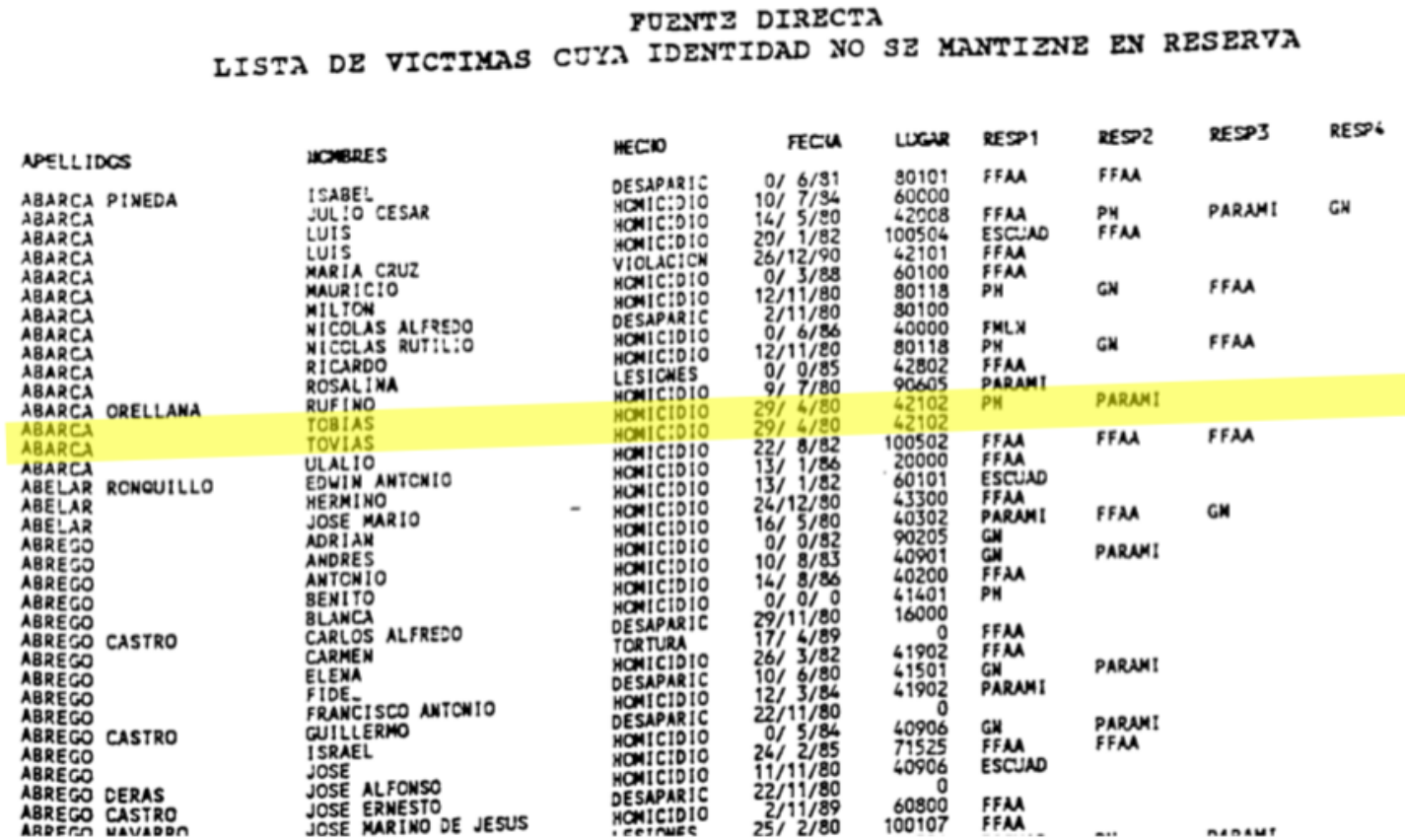


Figure: How could you automate this with millions of songs to have a clean, unique database of songs?

Let's consider a more challenging one!

Human Rights Applications



FUENTE DIRECTA

LISTA DE VICTIMAS CUYA IDENTIDAD NO SE MANTIENE EN RESERVA

APELLIDOS	NOMBRES	HECHO	FECHA	LUGAR	RESP1	RESP2	RESP3	RESP4
ABARCA PINEDA	ISABEL	DESAPARIC	0/ 6/81	80101	FFAA	FFAA		
ABARCA	JULIO CESAR	HOMICIDIO	10/ 7/34	60000				
ABARCA	LUIS	HOMICIDIO	14/ 5/80	42008	FFAA	PM	PARAMI	GN
ABARCA	LUIS	HOMICIDIO	20/ 1/82	100504	ESCUAD	FFAA		
ABARCA	MARIA CRUZ	VIOLACION	26/12/90	42101	FFAA			
ABARCA	MAURICIO	HOMICIDIO	0/ 3/88	60100	FFAA			
ABARCA	MILTON	HOMICIDIO	12/11/80	80118	PM	GN	FFAA	
ABARCA	NICOLAS ALFREDO	DESAPARIC	2/11/80	80100				
ABARCA	NICOLAS RUTILIO	HOMICIDIO	0/ 6/86	40000	FMLN			
ABARCA	RICARDO	HOMICIDIO	12/11/80	80118	PM	GN	FFAA	
ABARCA	ROSALINA	LESIONES	0/ 0/85	42802	FFAA			
ABARCA	RUFINO	HOMICIDIO	9/ 7/80	90605	PARAMI			
ABARCA	TOBIAS	HOMICIDIO	29/ 4/80	42102	PM	PARAMI		
ABARCA	TOVIAS	HOMICIDIO	29/ 4/80	42102				
ABARCA	ULALIO	HOMICIDIO	22/ 8/82	100502	FFAA	FFAA	FFAA	
ABELAR RONQUILLO	EDWIN ANTONIO	HOMICIDIO	13/ 1/86	20000	FFAA			
ABELAR	HERMINO	HOMICIDIO	13/ 1/82	60101	ESCUAD			
ABELAR	JOSE MARIO	HOMICIDIO	24/12/80	43300	FFAA			
ABREGO	ADRIAN	HOMICIDIO	16/ 5/80	40302	PARAMI	FFAA	GN	
ABREGO	ANDRES	HOMICIDIO	0/ 0/82	90205	GN			
ABREGO	ANTONIO	HOMICIDIO	10/ 8/83	40901	GN	PARAMI		
ABREGO	BENITO	HOMICIDIO	14/ 8/86	40200	FFAA			
ABREGO	BLANCA	HOMICIDIO	0/ 0/ 0	41401	PM			
ABREGO CASTRO	CARLOS ALFREDO	DESAPARIC	29/11/80	16000				
ABREGO	CARMEN	TORTURA	17/ 4/89	0	FFAA			
ABREGO	ELENA	HOMICIDIO	26/ 3/82	41902	FFAA			
ABREGO	FIDEL	DESAPARIC	10/ 6/80	41501	GN	PARAMI		
ABREGO	FRANCISCO ANTONIO	HOMICIDIO	12/ 3/84	41902	PARAMI			
ABREGO CASTRO	GUILLERMO	DESAPARIC	22/11/80	0				
ABREGO	ISRAEL	HOMICIDIO	0/ 5/84	40906	GN	PARAMI		
ABREGO	JOSE	HOMICIDIO	24/ 2/85	71525	FFAA	FFAA		
ABREGO DERAS	JOSE ALFONSO	HOMICIDIO	11/11/80	40906	ESCUAD			
ABREGO CASTRO	JOSE ERNESTO	DESAPARIC	22/11/80	0				
ABREGO NAVARRO	JOSE MARINO DE JESUS	HOMICIDIO	2/11/89	60800	FFAA			
		LESIONES	25/ 2/80	100107	FFAA			

Extract from Report of the UN Truth Commission of El Salvador (1993)

Figure: Original Information from the El Salvadoran Conflict before data is cleaned.

Human Rights Applications

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Figure: Snapshot from El Salvadoran Conflict. What seems difficult regarding this type of information?

Voter Registration Applications

Name	Street Address	Age	Sex	Race	Birth	Party
Domineck Q. AAshad Jr	914 Monmouth Ave #3	26	M	B	–	LIB
Domineck Q. AAshad Sr	1408 Auburndale Dr	55	M	B	NY	DEM
Xiomara A. Martinez	1715 Cole Mill Rd	31	F	O	HL	REP
Xiomara A. Martinez	2923 Forrestal Dr	31	F	O	HL	–
Virginia, L. Mullinix	749 Ninth St #480	101	F	W	PA	REP

Figure: Snapshot from the North Carolina Voter Registration Data Set

- Can you think of other applications where entity resolution would be needed or you have seen this before?
- Can you think of why the problem would be important?
- Can you think of some challenges of this problem?
- Why does this problem impact general machine learning, such as prediction and inference?

Questions?

beka@stat.duke.edu

Webpage: `resteorts.github.io`

Software: `https://github.com/orgs/cleanzr/`

Paper: `https://arxiv.org/abs/2008.04443`