Data Matching Methods and Their Uses
Intro/Overview

Profs Steve Fienberg, Rebecca Nugent
TA Sam Ventura

36-491, Mini 1
Department of Statistics
Carnegie Mellon University

August 28, 2013

Who is Sam Ventura?

Sam Ventura

Samuel Ventura

Samuel L. Ventura

Samuel Lukas Ventura

Who is Sam Ventura?

Who is NOT Samuel L. Ventura?

Will the Real Steve Fienberg Please Stand Up?

- Misspellings: Steve, Steven and Stephen; Fienberg, Feinberg, Fineberg, Fienburg, Feinburg, Steinberg, etc.

United States Census Bureau Master Address File

Which addresses correspond to the same unique location?

Address	City	State
...	...	...
123 East Main Street Unit 1	Pittsburgh	Pennsylvania
123 E. Main St. Apt. A	Pgh.	Pa.
123 Main St.	East Pittsburgh	PA
123 East Main Street	Portland	MN
123 East Main Street	Portland	N/A
...	...	...

United States Patent and Trademark Office

Which inventor records from the US Patent & Trademark Office (USPTO) database correspond to the same unique individuals?

Last	First	Middle	City	St	Assignee
...	...	...	...	...	...
Millar	David	A.	Stanford	CA	Stanford University
Miller	David	A.	Fair Haven	NJ	UNC
Miller	David	A.B.	Stanford	CA	Stanfrod University
Miller	David	Andrew	Stanford	CA	Lucent Technologies
Miller	David	Andrew	Fair Haven	NJ	Lucent Technologes
Miller	David	B.	Los Angeles	CA	Agilent Technologies
Miller	David	D.	Billerca	MA	Lucent Technologies
...	...	...	...	...	...

USPTO: 8 million patents, multiple inventors per patent

More USPTO Inventors

Last	First	Mid	City	St	Assignee
Zarian	James	R.	Corona Del Mar	CA	Lumenyte
Zarian	James	R.	Newport Beach	CA	Lumenyte
Zarian	Jashmid	J.	Woodland Hills	CA	Lumenyte
Zarian	Jashmid	NA	Woodland Hills	CA	Lumenyte
Zara	Michael	NA	Vienna	VA	Duke Univ.
Zara	Michael	NA	Vienna	VA	GW Univ.

Identifying Victims/Casualties

Homicide Victims in Columbia

- three different homicide record-systems; don't agree on number of deaths
- Colombian Census Bureau; Columbian National Police; Colombian Forensics Institute
- issues with conceptual, methodological differences; geographical coverage

Casualties in the Syrian conflict

- seven different lists
- data collection on the ground
- lots of missing data: gender, age, civilian vs military

Linking Records Together

- What features/variables would be useful?
- How would we work with text?
- Exact Matching?
- Partial Matching?
- What kinds of errors might we expect to have?

Labeled vs Unlabeled Data

Labeled Data:

- Pros: can build supervised models; ideas?
- Cons: Expensive; how would we get labeled data?

Unlabeled Data:

- Pros: cheap, “easy” to scrape, parse, collect public data
- Cons: Don’t know how we’re doing; what kind of statistical analysis can we do?

What else could we link with? What about faces?