Automating Frame Maintenance & Probabilistic Frame Matching

U.S. Census Bureau
Keith Finlay - Research Economist
Elizabeth Willhide - Survey Statistician

CIVIC DIGITAL FELLOWSHIP

United States Census Bureau

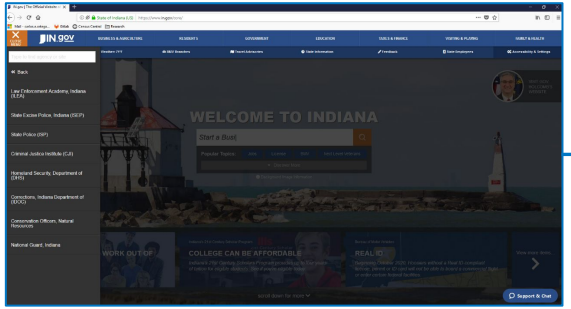
CARLOS EDUARDO ORTEGA
University of California, Berkeley
Data Science

Frame Maintenance is Costly

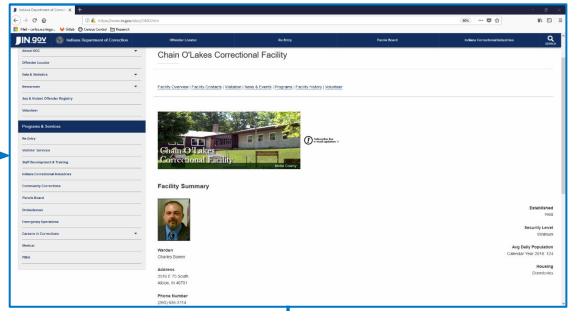
- **Public-sector collections in the Economic Reimbursable Surveys Division:**
  - Staff manually update frame information, hand match new tables to existing frames.
- **A lot of frame information is publicly available on the web.**
  - Use web scraping to bring in unstructured data.
- **How can we reduce the labor required to integrate or cross-reference new frame tables?**
  - Document parsing, fuzzy matching.

CIVIC DIGITAL FELLOWSHIP

Integrating New Data From Web Scraping



Web Crawler



Established 1968
Security Level Minimum
Warden Charles Bowen
Housing Dormitories
Address 3516 E 75 South Albion, IN 46701
Phone Number (260) 636-3114

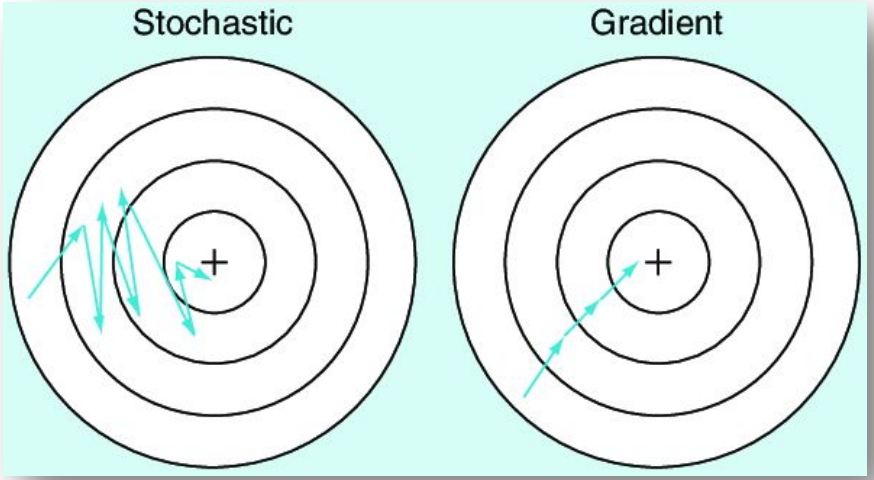
Document Parsing

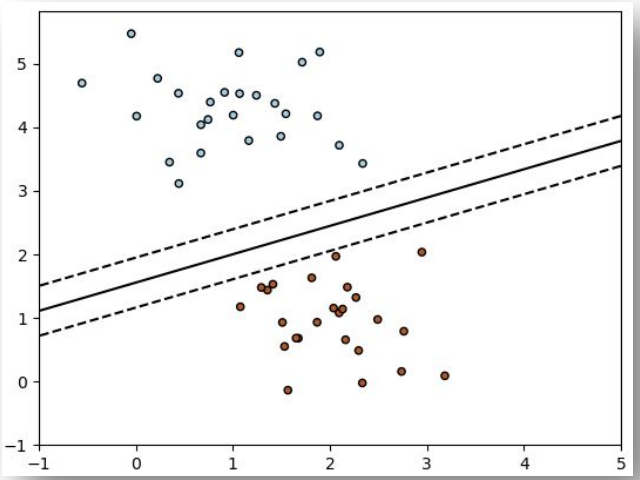
Facility	Chain O'Lakes Correctional Facility
Address	3516 E 75 South, Albion, IN 46701
Phone Number	(260) 636-3114

Fuzzy Matcher

CIVIC DIGITAL FELLOWSHIP

Identifying Relevant Pages





Stochastic Gradient Descent

Relevant vs Irrelevant Pages

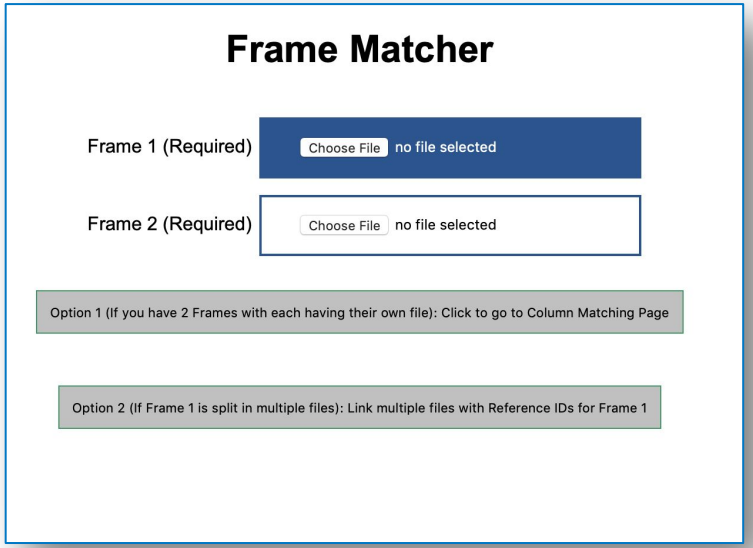
CIVIC DIGITAL FELLOWSHIP

Fuzzy Matcher to Link New Units

ID	Facility Name	City
100	Rockbourne State Prison	Silverden
101	Janville Penitentiary	Janville
102	Brightfield Jail	Brightfield
103	Fairford Holding Facility	Fairford

Facility Name	Contact Number
Rockbourne State Pris	(209) 555-0123
Janville Pen.	(510) 555-6789
Wellmill Jail	(215) 555-2154
Fairford Holding	(125) 555-8512

ID	Facility Name	City	Contact Number
100	Rockbourne State Prison	Silverden	(209) 555-0123
101	Janville Penitentiary	Janville	(510) 555-6789
102	Brightfield Jail	Brightfield	
103	Fairford Holding Facility	Fairford	(125) 555-8512
	Wellmill Jail		(215) 555-2154



CIVIC DIGITAL FELLOWSHIP

String Comparison

Q-Grams (n=2)

	CA	AR	RL	LO	OS	SO	OR	RT	TE	EG	GA	SS	SA	AN	NT	TA	NA
Carlos Ortega	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
Carlos Santana	1	1	1	1	1	0	0	0	0	0	0	1	1	2	1	1	1

Similarity

$$\frac{\boldsymbol{length}(intersection)}{\min[\boldsymbol{length}h(qgrams_1), \boldsymbol{length}(qgrams_2)]}$$

$$= \frac{\boldsymbol{length}(\text{ca}, ar, rl, lo, os)}{\min[length(ca, ar, rl, lo, os, so, or, rt, te, eg, ga),\ \boldsymbol{length}(ca, ar, rl, lo, os, ss, sa, an, nt, an, ta, na)]}$$

$$= \frac{5}{11} = 0.454545$$

CIVIC DIGITAL FELLOWSHIP

Moving Forward

Future Work

- Move frame matcher onto web server to improve speed and capacity of the program.
- Automate the search for matching columns in frame matching.
- Expand the scope of the web scraper to maintain more frames.

CIVIC DIGITAL FELLOWSHIP

Special Thanks To

- My supervisors, **Keith Finlay** & **Liz Willhide**, for supporting my work this summer and helping me navigate the Census.
- My mentor, **Mario Daniel Turse**, for providing me insight on working in Data Science within the government.
- **Rachel Dodell, Chris Kuang** and **Hillary Mclauchlin**, for organizing social and professional development events for all fellows, as well as giving us this opportunity this summer.
- My **Fellow Census Fellows**, as well as my pod, **Flora Wang, Raanan Gurewitsch**, and **Noam Kantor**, for making work feel a bit less like work.

CIVIC DIGITAL FELLOWSHIP