DATA:SEARCH ‘18

Searching for Datasets

Brian D. Davison

Jeff Heflin

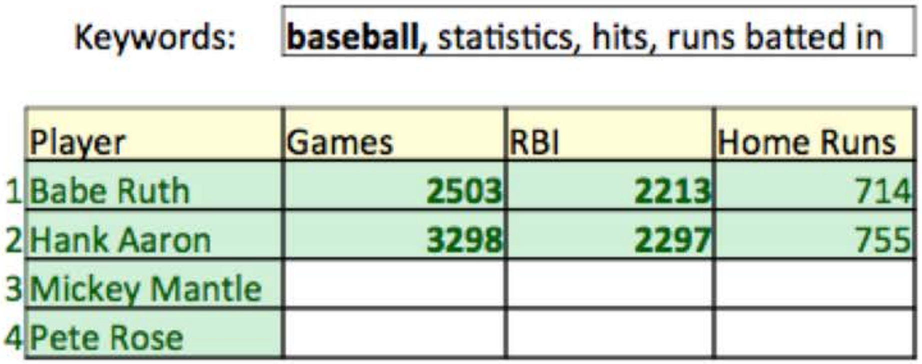
Haiyan Jia

- Y. Yi, Z. Chen, J. Heflin and B. D. Davison. **Recognizing Quantity Names for Tabular Data**. (DATA:SEARCH'18).
- Z. Chen, H. Jia, J. Heflin and B. D. Davison. **Generating Schema Labels through Dataset Content Analysis**. (Profiles & Data:Search'18), **Best paper award.**

Traditional IR

- Variety of information search modalities
  - Keyword search
  - Structured queries
  - Exemplars (find more like this)
  - Query-less search
- Search targets can vary too
  - Searching for containers/locations (documents/pages)
    - i.e., immediate information need is for a document, not the answer to a specific question
  - Searching for answers
- All of these can have analogues in DATA:SEARCH

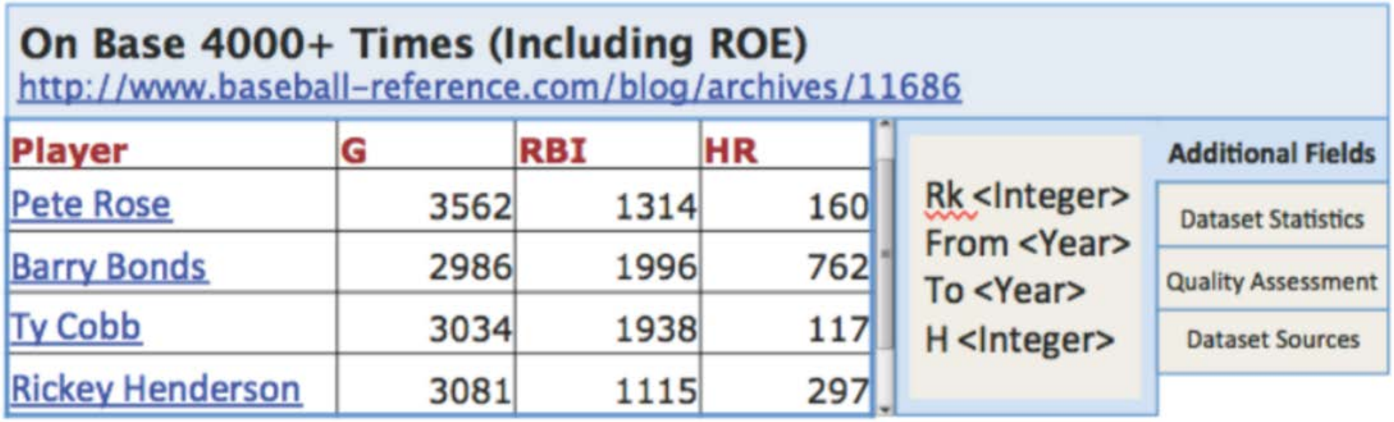
Possible Interfaces



Keywords: **baseball,** statistics, hits, runs batted in

	Player	Games	RBI	Home Runs
1	Babe Ruth	**2503**	**2213**	714
2	Hank Aaron	**3298**	**2297**	755
3	Mickey Mantle			
4	Pete Rose			

In this interface mock-up, the user searches for statistics about top baseball players. Only five items are bolded, corresponding to required matches. The attribute names are not required matches because some datasets may use different terms, and similarly the player names are not bolded because the searcher knows there are other names (“Peter Rose”, “George Herman Ruth”, “Henry Louis Aaron”) which could be used.

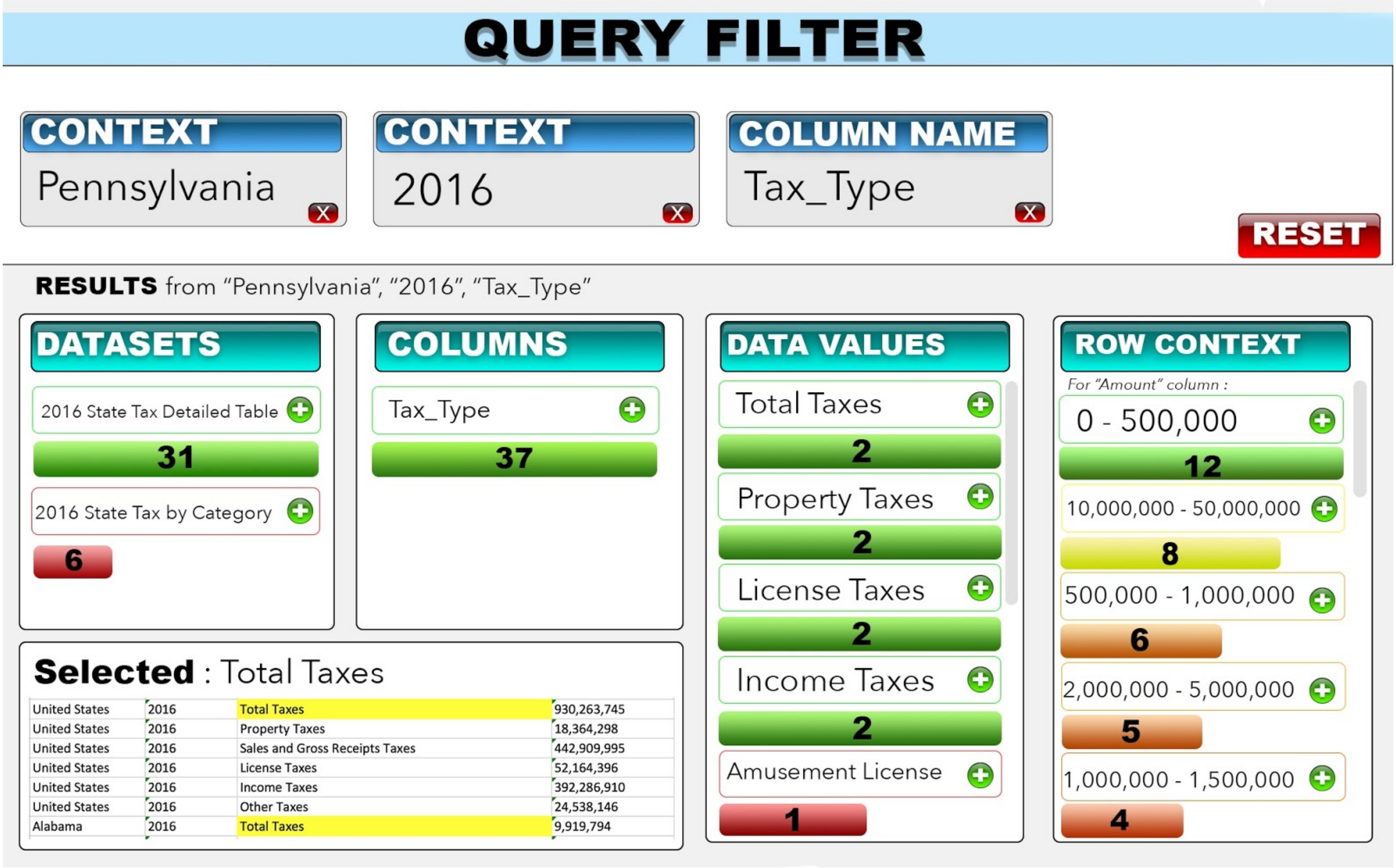


On Base 4000+ Times (Including ROE)

http://www.baseball-reference.com/blog/archives/11686

Player	**G**	**RBI**	**HR**
Pete Rose	3562	1314	160
Barry Bonds	2986	1996	762
Ty Cobb	3034	1938	117
Rickey Henderson	3081	1115	297

Rk <Integer>
From <Year>
To <Year>
H <Integer>



QUERY FILTER

CONTEXT	CONTEXT	COLUMN NAME
Pennsylvania ☒	2016 ☒	Tax_Type ☒

RESULTS from "Pennsylvania", "2016", "Tax_Type"

DATASETS

- 2016 State Tax Detailed Table ⊕ — 31
- 2016 State Tax by Category ⊕ — 6

COLUMNS

- Tax_Type ⊕ — 37

DATA VALUES

- Total Taxes ⊕ — 2
- Property Taxes ⊕ — 2
- License Taxes ⊕ — 2
- Income Taxes ⊕ — 2
- Amusement License ⊕ — 1

ROW CONTEXT

For "Amount" column :

- 0 - 500,000 ⊕ — 12
- 10,000,000 - 50,000,000 ⊕ — 8
- 500,000 - 1,000,000 ⊕ — 6
- 2,000,000 - 5,000,000 ⊕ — 5
- 1,000,000 - 1,500,000 ⊕ — 4

Selected : Total Taxes

United States	2016	Total Taxes	930,263,745
United States	2016	Property Taxes	18,364,298
United States	2016	Sales and Gross Receipts Taxes	442,909,995
United States	2016	License Taxes	52,164,396
United States	2016	Income Taxes	392,286,910
United States	2016	Other Taxes	24,538,146
Alabama	2016	Total Taxes	9,919,794

LEHIGH UNIVERSITY

DATA:SEARCH '18

LEHIGH UNIVERSITY

Thanks!

Brian D. Davison

davison@cse.lehigh.edu

Associate Prof., Lehigh University

Lehigh University's Linderman Library

We welcome collaborators – there are many directions to explore!