video demo
End-User Web Scraping: Google Scholar Edition

Sarah Chasins
data scraping tool

input
demonstration of how to collect the first row of a relational dataset

output
a script that collects the rest of the dataset
### Case Study: Google Scholar Data

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Citations</th>
<th>Authors</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>vapnik</td>
<td>The Nature of Statistical Learning Theory</td>
<td>1995</td>
<td>53976</td>
<td>V Vapnik</td>
<td>Data mining and knowledge discovery</td>
</tr>
<tr>
<td>vapnik</td>
<td>A training algorithm for optimal margin classifiers</td>
<td>1992</td>
<td>6095</td>
<td>BE Boser, IM Guyon, VN Vapnik</td>
<td>Proceedings of the fifth annual workshop on Computational learning theory ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### Case Study: Google Scholar Data

<table>
<thead>
<tr>
<th>Current Author</th>
<th>Title</th>
<th>Year</th>
<th>Citations</th>
<th>Authors</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vapnik</td>
<td>The Nature of Statistical Learning Theory</td>
<td>1995</td>
<td>53976</td>
<td>V Vapnik</td>
<td>Data mining and knowledge discovery</td>
</tr>
<tr>
<td>Vapnik</td>
<td>A training algorithm for optimal margin classifiers</td>
<td>1992</td>
<td>6095</td>
<td>BE Boser, IM Guyon, VN Vapnik</td>
<td>Proceedings of the fifth annual workshop on Computational learning theory ...</td>
</tr>
</tbody>
</table>

...
scale

authors limit
2000

papers per author limit
500

limits placed by user at demo time
two central questions

did the tool generate a good script?

at what age do researchers peak?
did the tool generate a good script?
should we trust this data at all?

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Citations</th>
<th>Journal/Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>vapnik</td>
<td>The Nature of Statistical Learning Theory</td>
<td>1995</td>
<td>53976</td>
<td>V Vapnik, Data mining and knowledge discovery</td>
</tr>
<tr>
<td>vapnik</td>
<td>A training algorithm for optimal margin classifiers</td>
<td>1992</td>
<td>6095</td>
<td>BE Boser, IM Guyon, VN Vapnik, Proceedings of the fifth annual workshop on Computational learning theory...</td>
</tr>
</tbody>
</table>

So checking up on the data afterwards is hard...
what do we expect?

2000 authors

up to 500 papers per author
what did we actually get?

rows: 157,159
what did we actually get?

rows: 157,159

unique authors: 1993
what did we actually get?

rows: 157,159

unique authors: 1993

oh no! tool messed up and I only have a week to fix it?
what did we actually get?

rows: 157,159
unique authors: 1993

possible explanations:
1. tool doesn’t work as well as I thought :( (my problem)
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)
what did we actually get?

rows: 157,159
unique authors: 1993

more thorough author analysis:
author names that appear separated by other author names:
Yves Deville : listed as author 183 and 191
Giovanni Pau : listed as author 355 and 1736
Henry Lin : listed as author 1024 and 1403
Fabrizio Messina : listed as author 1391 and 1396
authors whose citation counts jump in the middle of their runs:
Marco Ronchetti : listed as author 225 and 226
Joefon Jann : listed as author 810 and 811
Marcin Kubica : listed as author 1069 and 1070

remember papers were listed in order of decreasing citation count
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Year</th>
<th>Authors</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marco Ronchetti</td>
<td>Dynamical Properties of Classical Liquids and Liquid Mixture</td>
<td>1984</td>
<td>G Jacucci, M Ronchetti, W Schirmacher</td>
<td>Condensed Matter Research Using Neutrons, 139-161</td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Didattica per competenze: che supporto dalla tecnologia?</td>
<td></td>
<td>S Giaffredo, M Ronchetti, A Valerio</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Insegnare l’informatica a non-informatici: emergenza annunciata</td>
<td></td>
<td>S Giaffredo, L Mich, M Ronchetti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Some considerations from ontological standpoint of modeling processes in the social domain</td>
<td></td>
<td>A Ghosh, M Ronchetti, R Ferrario</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>LEZIONI SUL TELEFONINO: PORTING IN AMBIENTE SYMBIAN</td>
<td></td>
<td>M Ronchetti, J Stevovic</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Costruzione di un’interfaccia-utente per Lavagne Interattive Multimediali nel caso di simulazioni bidimensionali di fisica</td>
<td></td>
<td>M Ronchetti, N Dorigatti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>A Service-Oriented Architecture for the NEEDLE (Next gEneration sEarch engine for Digital LibrariEs) Multimodal Search Engine</td>
<td></td>
<td>M Ronchetti, MJN Krishnan, M Jarke</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Predizione contestuale di termini per fornire supporto a studenti con varie forme di disabilità</td>
<td></td>
<td>A Zanella, M Ronchetti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Spacetime: A Two Dimensions Search and Visualisation Engine Based on Linked Data</td>
<td></td>
<td>M RONCHETTI, F VALSECCHI</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Dipartimento di Informatica e Telecomunicazioni Università degli Studi di Trento, 38050 Povo (Trento) Italy</td>
<td></td>
<td>M Ronchetti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Dipartimento di Informatica e Studi Aziendali Università di Trento via F. Zeni 8, 1-38068 Rovereto (TN) ITALY</td>
<td></td>
<td>G Kovacs, G Succi, F Baruchelli, M Ronchetti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>L’uso di video su Internet nella didattica universitaria.</td>
<td></td>
<td>M Ronchetti</td>
<td></td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Bond-orientational order in liquids and glasses</td>
<td>1983</td>
<td>PJ Steinhardt, DR Nelson, M Ronchetti</td>
<td>Physical Review B 28 (2), 784</td>
</tr>
<tr>
<td>Marco Ronchetti</td>
<td>Icosahedral bond orientational order in supercooled liquids</td>
<td>1981</td>
<td>PJ Steinhardt, DR Nelson, M Ronchetti</td>
<td>Physical Review Letters 47 (18), 1297</td>
</tr>
</tbody>
</table>
what did we actually get?

rows: 157,159
unique authors: 1,993
unique author runs: 2,000
what did we actually get?
what if the runs weren’t the first 2,000?

Scholar page at end of run confirms they really were the first 2,000
what did we actually get?
what if the runs weren’t the first 2,000?
Scholar page at end of run confirms they really were the first 2,000

1. tool doesn’t work as well as I thought :-( (my problem)
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)
what did we actually get?

can we eliminate explanation 2 also?

1. tool doesn’t work as well as I thought :( (my problem)
2. data updates during scraping (problem inherent in long scraping tasks)
3. Scholar lists some authors twice (Scholar problem)
4. some authors share names (not a problem!)
what did we actually get?
what did we actually get?

label:computer_science yves deville

Yves Deville
Professor of Computer Science, Université catholique de Louvain, ICTEAM, EPL
Verified email at uclouvain.be
Cited by 3624
Computer Science, Artificial Intelligence, Constraints, Optimization

Yves Deville
Professor of Computer Science, University of Louvain
Cited by 3536
Computer Science, Artificial Intelligence, Constraints, Optimization

Dates and citation counts are estimated and are determined automatically by a computer program.
what did we actually get?

can we eliminate explanation 2 also?

1. tool doesn’t work as well as I thought :( (my problem)

2. data updates during scraping (problem inherent in long scraping tasks)

3. Scholar lists some authors twice (Scholar problem)

4. some authors share names (not a problem!)

I suspect 3 is true cause for all seven, but can’t be positive.
what did we actually get?
papers per author

what we expect to see
many authors with few papers
a few authors with many papers
spike around 500, from truncation

what we don’t want to see
spikes around multiples of 20
papers per author
papers per author

one paper authors?

turns out, yes
at what age do researchers peak?
citations by year

Number of Citations by Year of Publication

Citations of Year's Papers

Year
<table>
<thead>
<tr>
<th>Title</th>
<th>Cited by</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>A quasi-atomic model of human adenovirus type 5 capsid</td>
<td>120</td>
<td>2005</td>
</tr>
<tr>
<td>C Fabry, M Rosa-Calatrava, JF Conway, C Zubieta, S Cusack, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The EMBO Journal 24 (9), 1645-1654</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How good are humans at solving CAPTCHAs? a large scale evaluation</td>
<td>104</td>
<td>2010</td>
</tr>
<tr>
<td>E Bursztein, S Bethard, C Fabry, JC Mitchell, D Jurafsky</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security and Privacy (SP), 2010 IEEE Symposium on, 399-413</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The failure of noise-based non-continuous audio captchas</td>
<td>38</td>
<td>2011</td>
</tr>
<tr>
<td>E Bursztein, R Beauxis, H Paskov, D Perto, C Fabry, J Mitchell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security and Privacy (SP), 2011 IEEE Symposium on, 19-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure of the dodecahedral penton particle from human adenovirus</td>
<td>35</td>
<td>2006</td>
</tr>
<tr>
<td>type 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P Fuschetti, G Schoehn, P Fender, CMS Fabry, E/A Hewat, J Chroboczek, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of molecular biology 356 (2), 510-520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-dimensional structure of canine adenovirus serotype 2 capsid</td>
<td>30</td>
<td>2008</td>
</tr>
<tr>
<td>G Schoehn, M El Bakkouri, CMS Fabry, O Billot, LF Estrozli, L Le, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of virology 82 (7), 3192-3203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>An Archaeal Peptidase Assemblies into Two Different Quaternary Structures A</td>
<td>25</td>
<td>2006</td>
</tr>
<tr>
<td>TETRAHEDRON AND A GIANT OCTAHEDRON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G Schoehn, FMD Velleux, MA Dura, V Receveur-Brechot, CMS Fabry, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of Biological Chemistry 281 (47), 36327-36337</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The C-terminal domains of adenovirus serotype 5 protein IX assemble into an</td>
<td>19</td>
<td>2009</td>
</tr>
<tr>
<td>antiparallel structure on the facets of the capsid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS Fabry, M Rosa-Calatrava, C Moriscot, RWH Ruigrok, P Boulanger, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of virology 83 (2), 1135-1139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 capsid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM Fabry, M Rosa-Calatrava, JF Conway, C Zubieta, S Cusack, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMBO J 8 (24), 1645-1654</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
citations by year

papers removed for having no year information
14,115 (9.0%)

papers removed for being more than 50 years from author mean
169 (0.1%)

papers remaining
142,875 (90.9%)
citations by author-year
citations by author-year

but this allows a few authors with high citation counts to skew results
citations by author-year

Number of Citations by Author-Year of Publication

- **vapnik** *The Nature of Statistical Learning Theory* 53,976
- **vapnik** *Statistical Learning Theory* 54,228
- **David S. Johnson** *Computers and Intractability* 51,032
- **Peter E. Hart** *Pattern Classification* 46,535
citations by author-year

but this allows a few authors with high citation counts to skew results

alternatives
authors’ percent citations by year
authors’ highest cited paper years
citations by author-year

Percent of Author Citations by Author-Year of Publication

each dot is one paper
citations by author-year

Percent of Author Citations by Author-Year of Publication, Scaled by Absolute Citation Count
The average author receives about 9% of his or her total citations on papers from year 0 of his or her publishing career.
citations by author-year

but this puts extra weight on early papers because some authors have short careers for authors with 1 paper, 100% of citations in year 0...
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 10 Years or More

Average Percent of Author's Citations from Author-Year Papers

1,340 authors with 10 years or more publishing
Percent of Author Citations by Author-Year of Publication, for Authors Who Published 20 Years or More

647 authors with 20 years or more publishing
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 30 Years or More

285 authors with 30 years or more publishing
citations by author-year

Percept of Author Citations by Author-Year of Publication, for Authors Who Published 40 Years or More

110 authors with 40 years or more publishing
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 10 Years or More

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 20 Years or More

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 30 Years or More

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 40 Years or More
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 0-10 Years

751 authors with 0-10 years publishing
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 10-20 Years

732 authors with 10-20 years publishing
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 20-30 Years

Average Percent of Author's Citations from Author-Year Papers

391 authors with 20-30 years publishing
Percent of Author Citations by Author-Year of Publication, for Authors Who Published 30-40 Years

Average Percent of Author's Citations from Author-Year Papers

187 authors with 30-40 years publishing
citations by author-year

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 0-10 Years

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 10-20 Years

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 20-30 Years

Percent of Author Citations by Author-Year of Publication, for Authors Who Published 30-40 Years
citations by author-year

Number of Citations by Author-Year of Publication

- Each dot represents a paper.
- 4 papers with very high citation counts are not included.
most-cited papers

Number of Most-Cited Papers by Author-Year of Publication
most-cited papers

but still the problem with career length skewing results...
most-cited papers

Author Career Length and Author-Year of Most-Cited Paper

Each dot is one author.
most-cited papers

Author Career Length and Author-Year of Most-Cited Paper, Scaled by Absolute Citation Count

Author Publication Career Length (in Years)

Author-Year of Author's Most-Cited Paper
all papers
Range of Author-Years in Which Authors Wrote Papers Responsible for 75% of Their Citations
all papers

Range of Author-Years in Which Authors Wrote Papers Responsible for 75% of Their Citations
truncation

recent papers may not have had time to accumulate citations

authors still working may not have reached true peak yet
truncation

recent papers may not have had time to accumulate citations

authors still working may not have reached true peak yet
future work

remove the papers per author limit
good for analyzing my tool, not the author peak question
future work

not all computer science authors tagged with “computer science” label plans to search CS string and label, scrape common tags, then scrape larger set of authors

above approach -> larger data set should allow better analysis of effects of truncation
future work

collect data on conference committees (DBLP)?
aligning data with citation count data may reveal correlation

other suggestions?