SIIM Data Mining Lab Draft

SIIM 2016 v 20160621

Friday, July 1 | 12:45 pm – 2:45 pm Meeting Room E145-146

Authors:

- Chris Meenan, CIIP, University of Maryland School of Medicine
- Luciano Prevedello, MD, Ohio State Health System
- Ashish Sharma, PhD, Emory University

Idea: Give Students real world experience using data mining tools with messy data

Materials:

- Laptop
- Internet
- Software: Open Refine version 2.5
- Data file named: SIIM2016_Messy_Fake_EMRdata.csv

Downloading Open Refine - Prerequisite steps:

1. Go to http://www.openrefine.org
2. Scroll down to Download Open Refine
3. Select the appropriate operating system for your computer and download and install the latest version (Last stable release is Refine v 2.5).
4. Open “Refine” and verify installation was successful:

Note about information security:

Please observe and follow all institutional, state and national security precautions and legal requirements when working with patient data.

All data from Open Refine is stored on your own computer in the workspace directory. To open it, click "Browse workspace directory" on your OpenRefine application home page, by default at http://127.0.0.1:3333/

MacOSX:
~/Library/Application Support/Google/Refine/
Logging is to /var/log/daemon.log - grep for com.google.refine.Refine

Windows: Depending on OS version, stored in one of these directories:
C:\Documents and Settings\(user id)\Local Settings\Application Data\Google\Refine
C:\Users\(user id)\AppData\Roaming\Google\Refine
C:\Users\(user id)\Google\Refine

Linux:
~/.local/share/google/refine/
Open Refine Documentation: http://openrefine.org/OpenRefine/documentation
Lab #1: Introduction to open refine (formerly Google Refine)

Scenario: You are a data analyst at the Southern Immediate Innovation Medical System (SIIM). You have been asked to provide volume data across several sites including an academic medical center, a private practice site and a community hospital that are a part of the SIIM system. The hospital system shares the same registration system (so patients have a unique MRN) and the same RIS (so individually unique accession numbers), however they have a mixture of resource names and modality type has been excluded from your report.

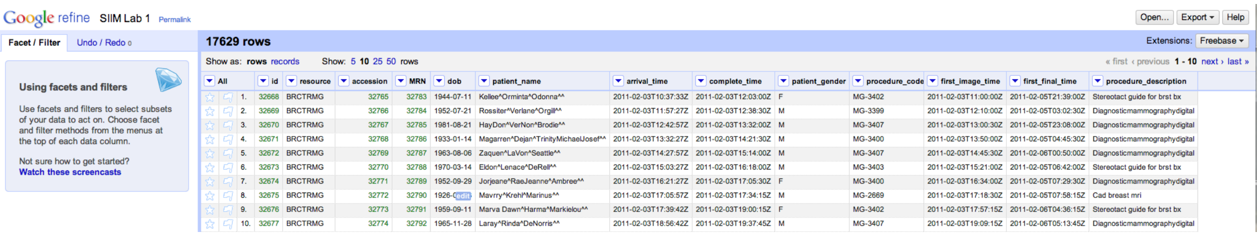
As an analyst, you are presented with multi-institution data dump from ACME RIS.

Goal #1: Normalize data dump with Google refine tools.

Goal #2: Figure out what the annual volumes are per modality per site to send back to your administrative team.

Lab Steps:

1. Before you start, make sure you have downloaded Refine tools as described in the lab prerequisites. If you have Refine installed, move to the next step.
2. Download a file named SIIM2016_Messy_Fake_EMRdata.csv and save to your desktop.
3. Load the data into Refine by selecting: "Create a project", and "get data from This Computer" and Choose the location of the file (your desktop).
4. Hit Next (the system will upload the data and show you a preview).
5. In the "Project Name" field, rename the project to SIIM Lab 1 and hit Create Project.
6. You should now see a view of your project sorted by various column headers. (see sample image below).



Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

17629 rows

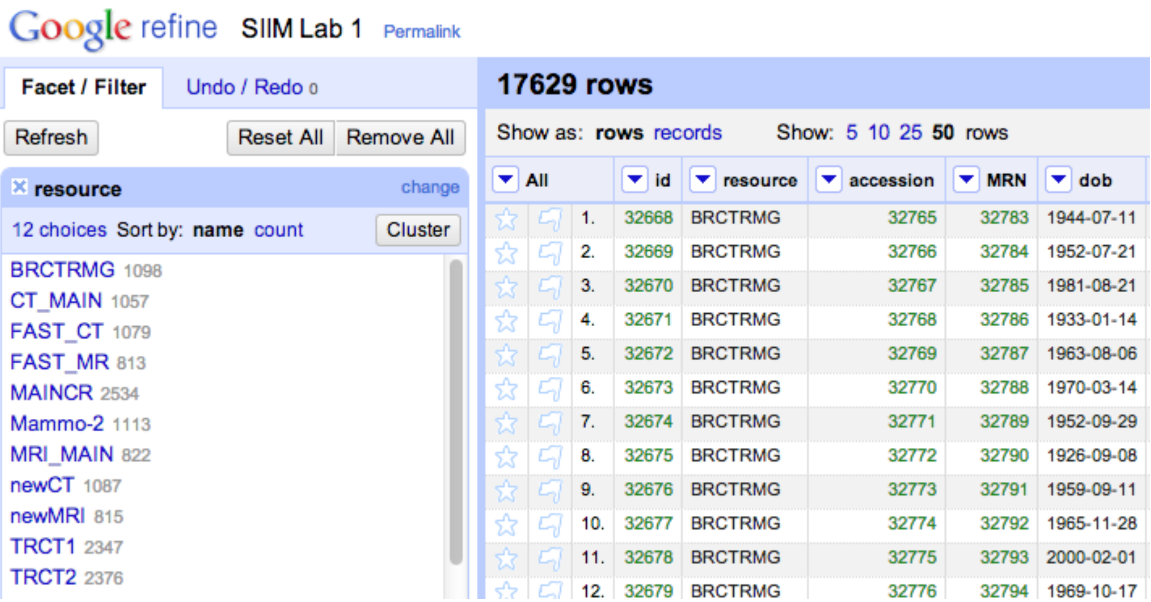
Show as: rows records Show: 5 10 25 50 rows

All	id	resource	accession	MRN	dob	patient_name	arrival_time	complete_time	patient_gender	procedure_code	first_image_time	first_final_time	procedure_description
1.	32668	BRCTRMG	32765	32783	1944-07-11	Kellee^Orminta^Odonna^^	2011-02-03T10:37:33Z	2011-02-03T12:03:00Z	F	MG-3402	2011-02-03T11:00:00Z	2011-02-05T21:39:00Z	Stereotact guide for brst bx
2.	32669	BRCTRMG	32766	32784	1952-07-21	Rossiter^Verlane^Orgill^^	2011-02-03T11:57:27Z	2011-02-03T12:38:30Z	M	MG-3399	2011-02-03T12:10:00Z	2011-02-05T03:02:30Z	Diagnosticmammographydigital
3.	32670	BRCTRMG	32767	32785	1981-08-21	HayDon^VerNon^Brodie^^	2011-02-03T12:42:57Z	2011-02-03T13:32:00Z	M	MG-3407	2011-02-03T13:00:30Z	2011-02-05T23:08:00Z	Diagnosticmammographydigital
4.	32671	BRCTRMG	32768	32786	1933-01-14	Magarren^Dejan^TrinityMichaelJosef^^	2011-02-03T13:32:27Z	2011-02-03T14:21:30Z	M	MG-3400	2011-02-03T13:50:00Z	2011-02-05T04:45:30Z	Diagnosticmammographydigital
5.	32672	BRCTRMG	32769	32787	1963-08-06	Zaquen^LaVon^Seattle^^	2011-02-03T14:27:57Z	2011-02-03T15:14:00Z	M	MG-3407	2011-02-03T14:45:30Z	2011-02-06T00:50:00Z	Diagnosticmammographydigital
6.	32673	BRCTRMG	32770	32788	1970-03-14	Eldon^Lenace^DeRell^^	2011-02-03T15:03:27Z	2011-02-03T16:18:00Z	M	MG-3403	2011-02-03T15:21:00Z	2011-02-05T06:42:00Z	Stereotact guide for brst bx
7.	32674	BRCTRMG	32771	32789	1952-09-29	Jorjeane^RaeJeanne^Ambree^^	2011-02-03T16:21:27Z	2011-02-03T17:05:30Z	F	MG-3400	2011-02-03T16:34:00Z	2011-02-05T07:29:30Z	Diagnosticmammographydigital
8.	32675	BRCTRMG	32772	32790	1926-0edit	Mavrry^Krehl^Marinus^^	2011-02-03T17:05:57Z	2011-02-03T17:34:15Z	M	MG-2669	2011-02-03T17:18:30Z	2011-02-05T07:58:15Z	Cad breast mri
9.	32676	BRCTRMG	32773	32791	1959-09-11	Marva Dawn^Harma^Markielou^^	2011-02-03T17:39:42Z	2011-02-03T19:00:15Z	F	MG-3402	2011-02-03T17:57:15Z	2011-02-06T04:36:15Z	Stereotact guide for brst bx
10.	32677	BRCTRMG	32774	32792	1965-11-28	Laray^Rinda^DeNorris^^	2011-02-03T18:56:42Z	2011-02-03T19:37:45Z	M	MG-3407	2011-02-03T19:09:15Z	2011-02-06T05:13:45Z	Diagnosticmammographydigital

7. Remember, your goal is to determine modality volumes. Start by looking at the Resource column. You can look through many different views of your data by a) selecting number of rows to display and b) looking through each sheet by selecting "next" on the upper right hand side.

8. Reviewing your data shows a variety of resource names

9. Select the drop down arrow next to "resource" and choose "facet" then "text facet".
10. You should now see the resources broken down by resource name and number sorted by frequency. Note now you also see the number of resource names (Qty. 12). (see image below)



17629 rows

12 choices Sort by: name count

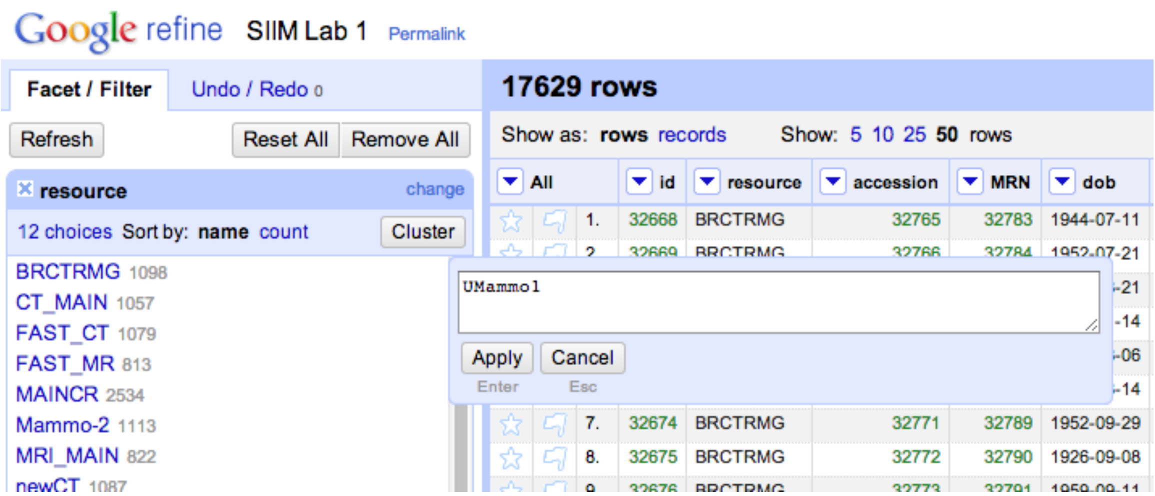
resource	count
BRCTRMG	1098
CT_MAIN	1057
FAST_CT	1079
FAST_MR	813
MAINCR	2534
Mammo-2	1113
MRI_MAIN	822
newCT	1087
newMRI	815
TRCT1	2347
TRCT2	2376

All	id	resource	accession	MRN	dob
1.	32668	BRCTRMG	32765	32783	1944-07-11
2.	32669	BRCTRMG	32766	32784	1952-07-21
3.	32670	BRCTRMG	32767	32785	1981-08-21
4.	32671	BRCTRMG	32768	32786	1933-01-14
5.	32672	BRCTRMG	32769	32787	1963-08-06
6.	32673	BRCTRMG	32770	32788	1970-03-14
7.	32674	BRCTRMG	32771	32789	1952-09-29
8.	32675	BRCTRMG	32772	32790	1926-09-08
9.	32676	BRCTRMG	32773	32791	1959-09-11
10.	32677	BRCTRMG	32774	32792	1965-11-28
11.	32678	BRCTRMG	32775	32793	2000-02-01
12.	32679	BRCTRMG	32776	32794	1969-10-17

11. Clicking on each facet on the left hand side will take you to the related data for each facet. Click on TRCT1. Look at the procedure description on the right hand side. This is clearly a CT scanner!

12. What kind of resource is BRCTRMG? Click on the facet name BRCTRMG and the related data for the facet will appear.

13. The resource names are messy. Let's try something more readable. You call your friendly neighborhood PACS Administrator who knows the resource names and where they are located. Highlight the resource name over BRCTRMG and click "edit". Rename this to UMG1. (see figure below)



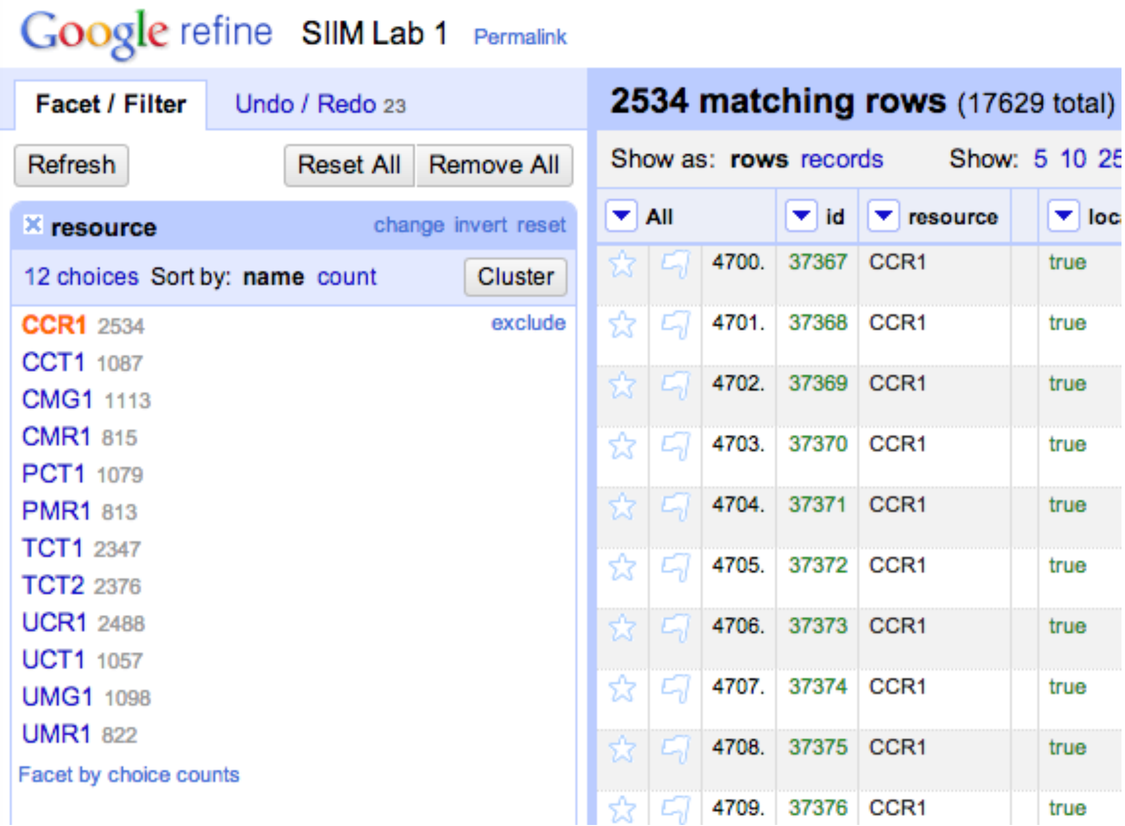
BRCTRMG 1098
CT_MAIN 1057
FAST_CT 1079
FAST_MR 813
MAINCR 2534
Mammo-2 1113
MRI_MAIN 822
newCT 1087

UMammo1

All	id	resource	accession	MRN	dob
1.	32668	BRCTRMG	32765	32783	1944-07-11
2.	32669	BRCTRMG	32766	32784	1952-07-21
					-21
					-14
					-06
					-14
7.	32674	BRCTRMG	32771	32789	1952-09-29
8.	32675	BRCTRMG	32772	32790	1926-09-08
9.	32676	BRCTRMG	32773	32791	1959-09-11

14. Let's use a convention to name resources:
    Location: One or two letter location U,C,P,TR
    Modality: Two letter modality description (CR, MR)
    Number: In order of progression. Rename the rest of your resources to friendly names (and buy the PACS administrator lunch) as follows:
    a. Mammo-2 -> CMG1
    b. XRAY -> UCR1
    c. MAINCR -> CCR1
    d. MRI_MAIN -> UMR1
    e. newMRI -> CMR1
    f. FAST_MR -> PMR1
    g. TRCT1 -> TCT1
    h. TRCT2 -> TCT2
    i. CT_MAIN -> UCT1
    j. newCT -> CCT1
    k. FAST_CT ->PCT1

    Now that we have the resource data in a more reasonable format, it's easy to see volumes per site per modality. When you are done, your resource facet should look like this:



2534 matching rows (17629 total)

resource

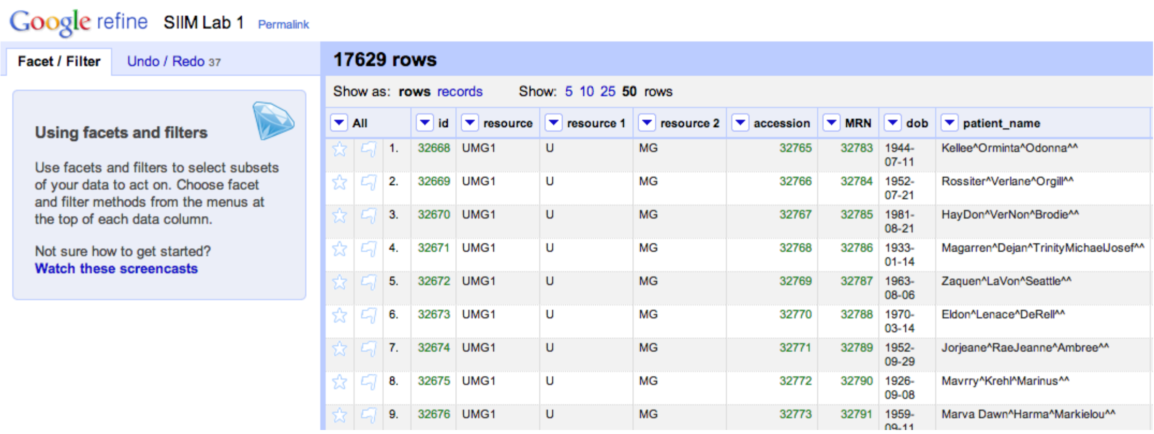
12 choices Sort by: name count

resource	count
CCR1	2534
CCT1	1087
CMG1	1113
CMR1	815
PCT1	1079
PMR1	813
TCT1	2347
TCT2	2376
UCR1	2488
UCT1	1057
UMG1	1098
UMR1	822

All	id	resource	loc
4700.	37367	CCR1	true
4701.	37368	CCR1	true
4702.	37369	CCR1	true
4703.	37370	CCR1	true
4704.	37371	CCR1	true
4705.	37372	CCR1	true
4706.	37373	CCR1	true
4707.	37374	CCR1	true
4708.	37375	CCR1	true
4709.	37376	CCR1	true

15. Let’s clean up the data a bit more. Click the “x” on the resource facet. The left hand side of your refine lab should be blank now (no facets or filters showing).

16. Click on the drop down arrow next to the resource column, choose “edit column”, then “Split into several columns”. This will give us a number of ways to split up the data. We are going to select “by field lengths” and enter 1, 2. Before you press OK, be sure to unselect the “remove this column” button in the “After Splitting” section of the window. You should now see two new columns (resource 1, and resource 2).



Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

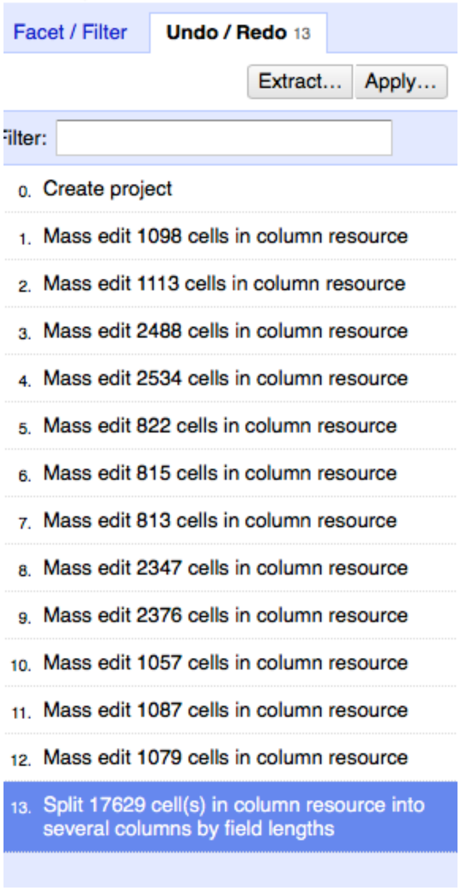
Not sure how to get started?
Watch these screencasts

17629 rows

Show as: **rows** records   Show: 5 10 25 **50** rows

All	id	resource	resource 1	resource 2	accession	MRN	dob	patient_name
1.	32668	UMG1	U	MG	32765	32783	1944-07-11	Kellee^Orminta^Odonna^^
2.	32669	UMG1	U	MG	32766	32784	1952-07-21	Rossiter^Verlane^Orgill^^
3.	32670	UMG1	U	MG	32767	32785	1981-08-21	HayDon^VerNon^Brodie^^
4.	32671	UMG1	U	MG	32768	32786	1933-01-14	Magarren^Dejan^TrinityMichaelJosef^^
5.	32672	UMG1	U	MG	32769	32787	1963-08-06	Zaquen^LaVon^Seattle^^
6.	32673	UMG1	U	MG	32770	32788	1970-03-14	Eldon^Lenace^DeRell^^
7.	32674	UMG1	U	MG	32771	32789	1952-09-29	Jorjeane^RaeJeanne^Ambree^^
8.	32675	UMG1	U	MG	32772	32790	1926-09-08	Mavrry^Krehl^Marinus^^
9.	32676	UMG1	U	MG	32773	32791	1959-09-11	Marva Dawn^Harma^Markielou^^

HELP! - What if I forgot to select "remove this column ?!? – It's OK, simply select the "undo" tab and select the last step that you performed. Refine keeps a running log of changes so you can easily go backwards!

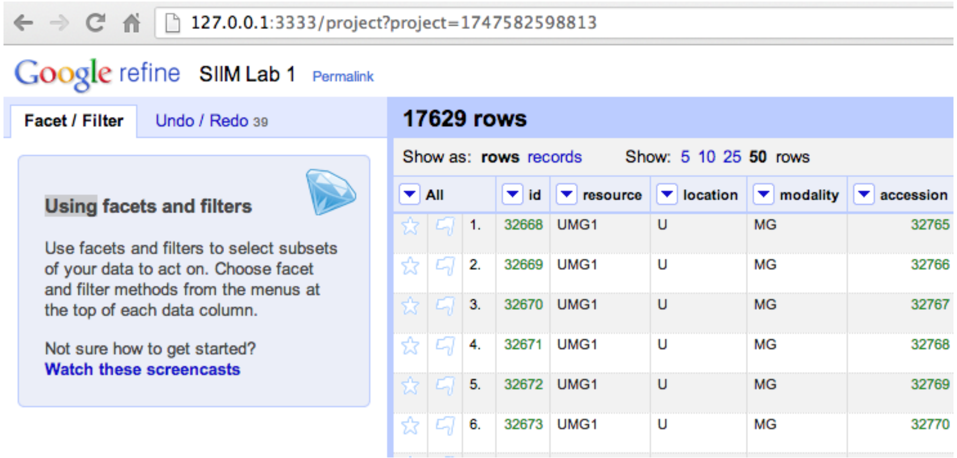


0. Create project
1. Mass edit 1098 cells in column resource
2. Mass edit 1113 cells in column resource
3. Mass edit 2488 cells in column resource
4. Mass edit 2534 cells in column resource
5. Mass edit 822 cells in column resource
6. Mass edit 815 cells in column resource
7. Mass edit 813 cells in column resource
8. Mass edit 2347 cells in column resource
9. Mass edit 2376 cells in column resource
10. Mass edit 1057 cells in column resource
11. Mass edit 1087 cells in column resource
12. Mass edit 1079 cells in column resource
13. Split 17629 cell(s) in column resource into several columns by field lengths

The Undo-feature!

17. Let's rename these columns. Click on the drop down box next to the "Resource 1" Column, select Edit Column, Rename this Column and rename the column to "Location". Click on the drop down box next to the "Resource 2" Column, select Edit Column, Rename this Column and rename the column to "Modality".

    When you are done, your refine window should look like this:



127.0.0.1:3333/project?project=1747582598813

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

17629 rows

Show as: rows records   Show: 5 10 25 50 rows

All	id	resource	location	modality	accession
1.	32668	UMG1	U	MG	32765
2.	32669	UMG1	U	MG	32766
3.	32670	UMG1	U	MG	32767
4.	32671	UMG1	U	MG	32768
5.	32672	UMG1	U	MG	32769
6.	32673	UMG1	U	MG	32770

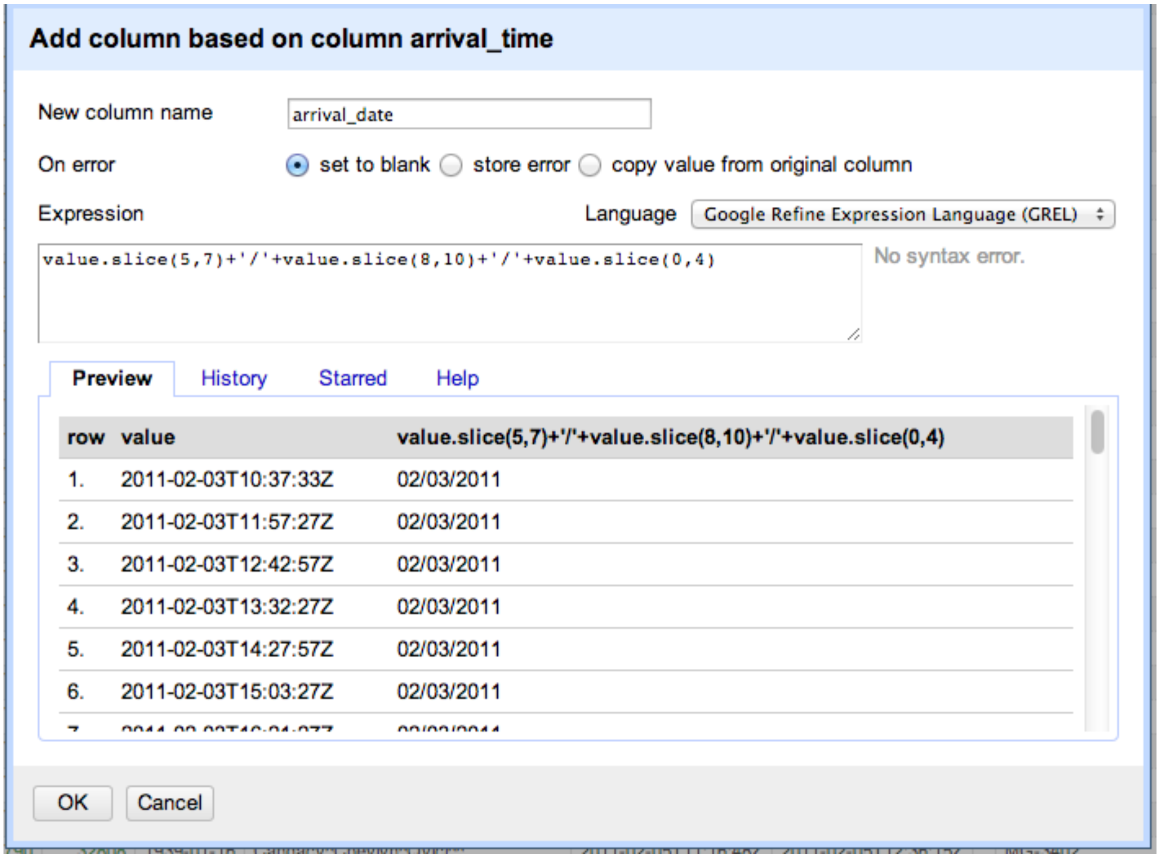
18. Now we can create a facet to easily look at global modality volumes, choose the modality column, select facet, text facet and you can now see the total modality volumes across all sites for the period selected.

19. Let's try to look at dates now. The date format is not quite what we need as it's currently a combination of date and time jammed together (e.g 2011-02-03-T10:37:33Z). Let's change the format to something more readable. Let's pretend that we don't care about the time for this report, and we just want to look at arrival date. Click on the arrow next to arrival time, and choose "column, add column based on this column". Let's name this new column arrival_date. In the expression field, lets enter the following expression:

```
value.slice(5, 7) + '/' + value.slice(8, 10) + '/' + value.slice(0, 4)
```

    It should look like this (notice if you have it right, refine reports NO SYNTAX ERROR:

    Press OK!



Add column based on column arrival_time

New column name: arrival_date

On error: ● set to blank ○ store error ○ copy value from original column

Expression — Language: Google Refine Expression Language (GREL)

```
value.slice(5,7)+'/'+value.slice(8,10)+'/'+value.slice(0,4)
```

No syntax error.

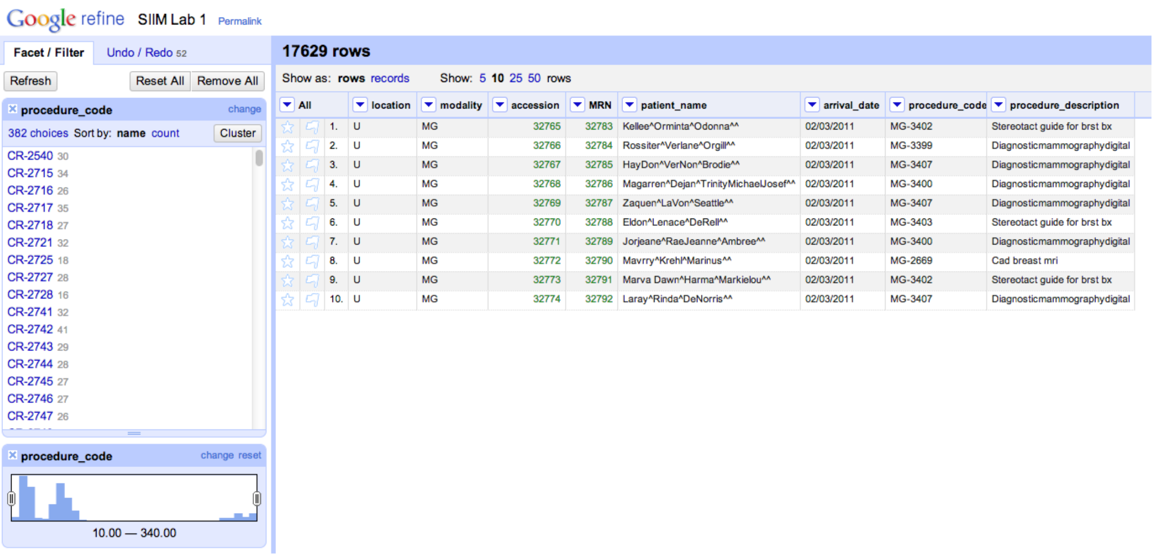
row	value	value.slice(5,7)+'/'+value.slice(8,10)+'/'+value.slice(0,4)
1.	2011-02-03T10:37:33Z	02/03/2011
2.	2011-02-03T11:57:27Z	02/03/2011
3.	2011-02-03T12:42:57Z	02/03/2011
4.	2011-02-03T13:32:27Z	02/03/2011
5.	2011-02-03T14:27:57Z	02/03/2011
6.	2011-02-03T15:03:27Z	02/03/2011

20. Let’s clean up our data display now by getting rid of a few columns. Lets remove a few columns by clicking on Edit Column and remove this column. Remove the following columns:

    remove -> first final time
    remove -> first image time
    remove -> complete_time
    remove -> patient gender
    remove -> arrival_time
    remove -> DOB
    remove -> id
    remove -> resource

    The list looks much cleaner now, and is broken down by location, modality type, acc#, MRN, name, date, procedure code and procedure description.

21. For fun, lets facet on the procedure code column, select facet and text facet. Scroll to the very bottom of the facet and select “facet by choice counts”. You should see a very nice histogram of the frequency of your procedure codes volumes.



procedure_code	count
CR-2540	30
CR-2715	34
CR-2716	26
CR-2717	35
CR-2718	27
CR-2721	32
CR-2725	18
CR-2727	28
CR-2728	16
CR-2741	32
CR-2742	41
CR-2743	29
CR-2744	28
CR-2745	27
CR-2746	27
CR-2747	26

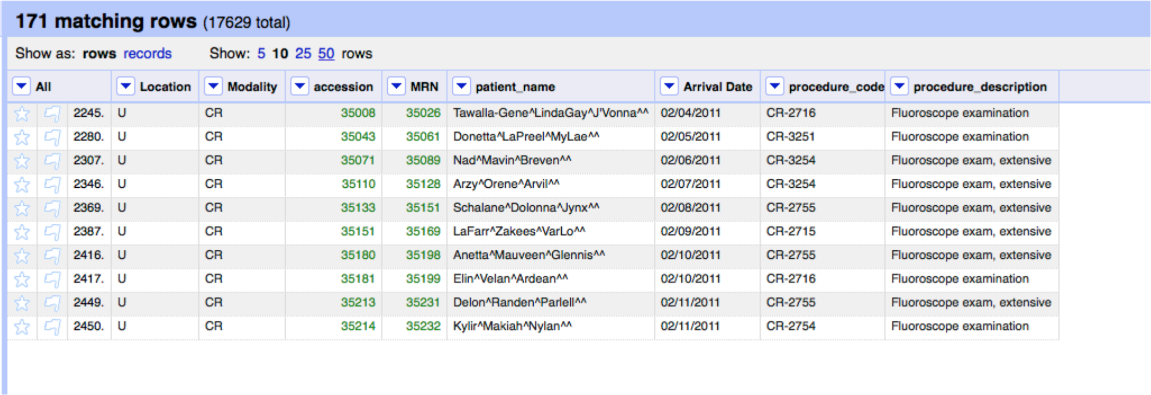
All	location	modality	accession	MRN	patient_name	arrival_date	procedure_code	procedure_description
1.	U	MG	32765	32783	Kellee^Orminta^Odonna^^	02/03/2011	MG-3402	Stereotact guide for brst bx
2.	U	MG	32766	32784	Rossiter^Verlane^Orgill^^	02/03/2011	MG-3399	Diagnosticmammographydigital
3.	U	MG	32767	32785	HayDon^VerNon^Brodie^^	02/03/2011	MG-3407	Diagnosticmammographydigital
4.	U	MG	32768	32786	Magarren^Dejan^TrinityMichaelJosef^^	02/03/2011	MG-3400	Diagnosticmammographydigital
5.	U	MG	32769	32787	Zaquen^LaVon^Seattle^^	02/03/2011	MG-3407	Diagnosticmammographydigital
6.	U	MG	32770	32788	Eldon^Lenace^DeRell^^	02/03/2011	MG-3403	Stereotact guide for brst bx
7.	U	MG	32771	32789	Jorjeane^RaeJeanne^Ambree^^	02/03/2011	MG-3400	Diagnosticmammographydigital
8.	U	MG	32772	32790	Mavrry^Krehl^Marinus^^	02/03/2011	MG-2669	Cad breast mri
9.	U	MG	32773	32791	Marva Dawn^Harma^Markielou^^	02/03/2011	MG-3402	Stereotact guide for brst bx
10.	U	MG	32774	32792	Laray^Rinda^DeNorris^^	02/03/2011	MG-3407	Diagnosticmammographydigital

What patterns do you see in your volumes? Clearly there are many procedures that you do a small number of, and a few procedures that you do a high volume of. Drag the slider bars around the histogram to explore the data volumes. What happens if you set the procedure code slider to 110-290? What's happening?

22. Remove all existing facets by clicking on the x. Let's look at the most frequently occurring procedure description. Click on the procedure_description column, then facet, text facet. Click sort by count on the facet to see the most frequently occurring procedure description (what is it?).

23. Let's see how clean our procedure description names are as they relate to the procedure codes and make sure they match and are unique (e.g. sometimes if you are looking at data from multiple RIS in one report, you may see different descriptions for the same procedure codes). Looking at the procedure description facet you created in step 21, click the cluster button. Then select the metaphone3 method under "Keying Function".

24. This is essentially refine's function looking for procedure descriptions that may be different words for the same thing. Scroll down to the very bottom, and you will see two fluoroscopy exams. Hover over the two exams and select "browse this cluster". (What do you see? Are these really the same thing?). Close the facet to exit.



171 matching rows (17629 total)

Show as: **rows** records Show: 5 **10** 25 50 rows

All	Location	Modality	accession	MRN	patient_name	Arrival Date	procedure_code	procedure_description
2245.	U	CR	35008	35026	Tawalla-Gene^LindaGay^J'Vonna^^	02/04/2011	CR-2716	Fluoroscope examination
2280.	U	CR	35043	35061	Donetta^LaPreel^MyLae^^	02/05/2011	CR-3251	Fluoroscope examination
2307.	U	CR	35071	35089	Nad^Mavin^Breven^^	02/06/2011	CR-3254	Fluoroscope exam, extensive
2346.	U	CR	35110	35128	Arzy^Orene^Arvil^^	02/07/2011	CR-3254	Fluoroscope exam, extensive
2369.	U	CR	35133	35151	Schalane^Dolonna^Jynx^^	02/08/2011	CR-2755	Fluoroscope exam, extensive
2387.	U	CR	35151	35169	LaFarr^Zakees^VarLo^^	02/09/2011	CR-2715	Fluoroscope exam, extensive
2416.	U	CR	35180	35198	Anetta^Mauveen^Glennis^^	02/10/2011	CR-2755	Fluoroscope exam, extensive
2417.	U	CR	35181	35199	Elin^Velan^Ardean^^	02/10/2011	CR-2716	Fluoroscope examination
2449.	U	CR	35213	35231	Delon^Randen^Parlell^^	02/11/2011	CR-2755	Fluoroscope exam, extensive
2450.	U	CR	35214	35232	Kylir^Makiah^Nylan^^	02/11/2011	CR-2754	Fluoroscope examination

25. Let's remove the remaining columns with PHI, so remove columns marked "accession #, MRN and patient name". (edit column, remove this column)

26. So you've done it, you have cleaned up your data and are ready to send a spreadsheet on to your chief operating officer of volumes per modality. Select the export button, choose excel and refine should save a spreadsheet file directly to your desktop.

References:

http://openrefine.org/ (great videos)
https://code.google.com/p/google-refine/wiki/Recipes
https://trip-workflow.siimweb.org