CSCI-548:
Data Cleaning
Spring 2015

Bo Wu

University of Southern California/ISI

Outline

- Introduction
- OpenRefine
- Data Wrangler
- FlashFill

Problem

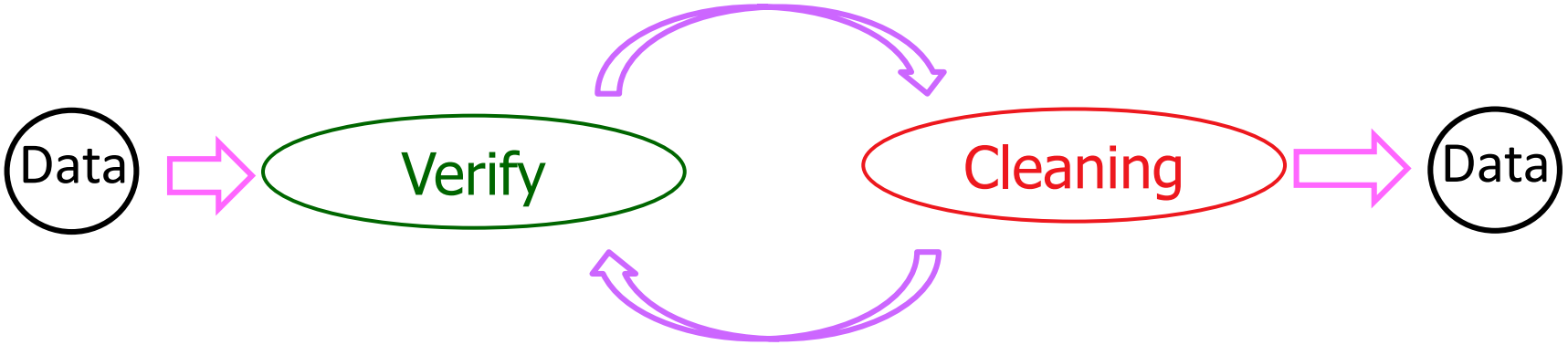
- Data is not in the expected format

artist	artistyear	artyear	credittext	dimensions	photo
FRANK STELLA	born 1936	1970	Gift of Alice Pratt Brown	120 x 600 inches	Oil on canvas
JASPER JOHNS	born 1930		Museum purchase with funds provided by; the Agnes Cullen Arnold Endowment Fund	75 x 50 inches	http://www.mfah.org/site_media/cache/e5/37/e537510de3215396c6c0336b17eab82b.jpg
GEORGE BELLOWS	1882 - 1925	1914	Gift of Mr. and Mrs. Meredith Long in memory of Agnes Cullen Arnold	38 x 30 inches	http://www.mfah.org/site_media/cache/72/7c/727c9b61b85b16d432a79031137e3d6b.jpg

Solutions

- Prevent dirty data from getting into the system
  - Enforce source integrity constraints
    - Not allow “null” for a field
    - Only allow numbers
    - …
- Clean data
  - Manually clean the data
  - Create clean scripts

Data Cleaning Workflow



Data Transformation Operations

- One to one mapping
- One to many/ many to one mapping
- Look up and join
- Positional
- Filter
- Common functions: sum, min, max, avg …

One Row to One Row Mappings

- Drop Column
- Copy Column
- Add Column
  - Constant value, random number, serial number
- Merge Columns with Glue
- Value update

Name
George
Paul
Ann

(.*) to <\C>\1</\C>

Name
<Name>George</Name>
<Name>Paul</Name>
<Name>Ann</Name>

- Split Column
  - by position
  - by regular expression (first match)

slides from Pedro Szekely

Example

		Stewart,Bob
Anna	Davis	
		Dole,Jerry
Joan	Marsh	

Format '(.*), (.*)' to '\2 \1'

		Bob Stewart
Anna	Davis	
		Jerry Dole
Joan	Marsh	

Split at ' '

		Bob	Stewart
Anna	Davis		
		Jerry	Dole
Joan	Marsh		

2 Merges

Bob	Stewart
Anna	Davis
Jerry	Dole
Joan	Marsh

slides from Pedro Szekely

One Row To Many Rows: Fold

A	B	C	D	E	F
a1	b1	c1	d1	e1	f1

fold(D, E, F)

A	B	C	
a1	b1	c1	d1
a1	b1	c1	e1
a1	b1	c1	f1

slides from Pedro Szekely

One Row To Many Rows: Fold

Name	Math	Bio
Ann	43	78
Bob	96	54

2 Formats (demotes)

Name		
Ann	Math:43	Bio:78
Bob	Math:96	Bio:54

Fold

Name	
Ann	Math:43
Ann	Bio:78
Bob	Math:96
Bob	Bio:54

Split

Name		
Ann	Math	43
Ann	Bio	78
Bob	Math	96
Bob	Bio	54

slides from Pedro Szekely

Vijayshankar Raman, Joseph M. Hellerstein

Many Rows to One Row: Unfold

unfold(col_1, col_2)

Name		
George	Math	65
George	French	42
Anna	Math	43
Anna	French	78
Bob	English	96
Bob	French	54
Joan	English	79

unfold(2,3)

Name	Math	French	English
George	65	42	
Anna	43	78	
Bob		54	96
Joan			79

slides from Pedro Szekely

OpenRefine

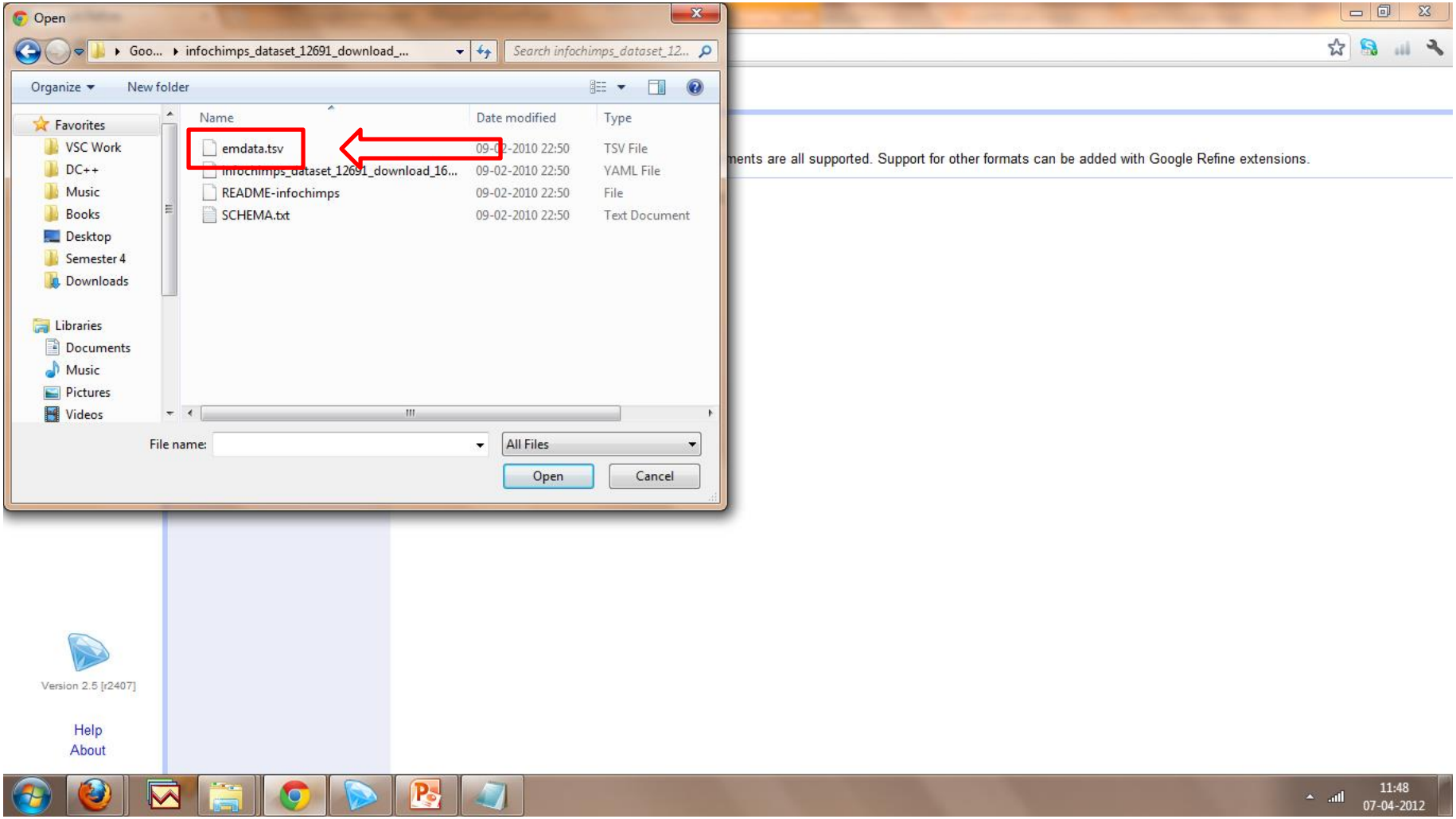
- A powerful tool that can be effectively used for data cleansing.
- It helps in working with raw data, cleaning it up, transforming from one format to other, encompassing it with web services and linking it to databases.
- It is very easy to use and has a web interface.
- It is freely available and works well with any browser.
- a desktop application and it runs a small web server on your system and we need to point our browser to the server to use refine.

slides from Sathishwaran.R and Vijaya Prabhu

Importing Data

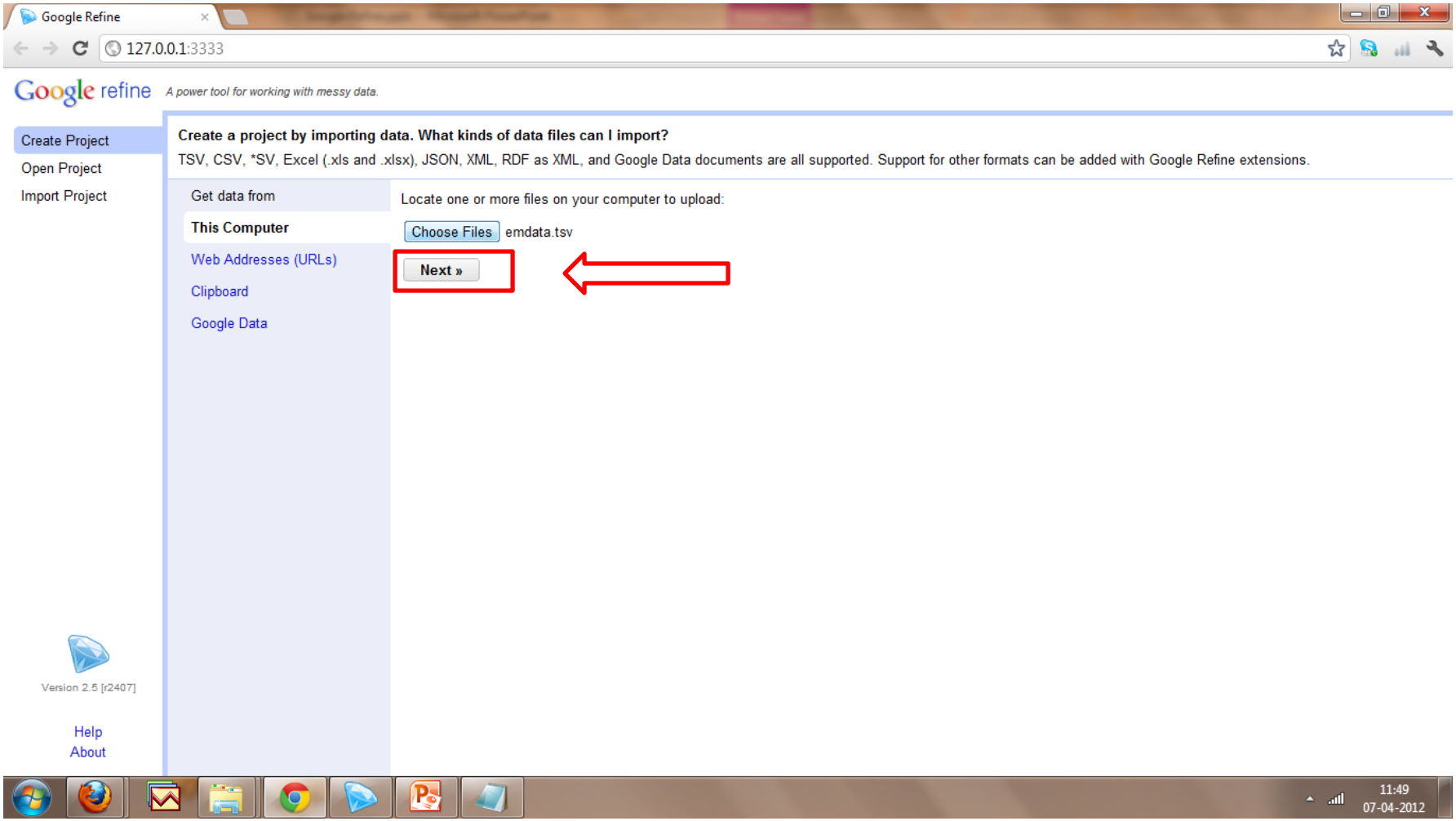
- Google Refine supports TSV, CSV, Excel (.xls and .xlsx), JSON, XML, and Google data document formats.
- Once imported the data is in Google Refine's own data format.
- We have used TSV data on Disasters worldwide from 1900-2008 available from http://www.infochimps.com/datasets/disasters-worldwide-from-1900-2008 for the tutorial.

Importing Data



slides from Sathishwaran.R and Vijaya Prabhu

Importing Data



slides from Sathishwaran.R and Vijaya Prabhu

Creating Project

Data Uploaded

Google Refine

127.0.0.1:3333

Google refine — A power tool for working with messy data.

Create Project
Open Project
Import Project

« Start Over | Configure Parsing Options | Project name: emdata tsv | Create Project »

	Start	End	Country	Location	Type	Sub_Type	Name	Killed	Cost	Affected	Id	Column 12
1.	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	Drought			280000		2008-9475	
2.	72006	2006	Afghanistan		Drought	Drought			1900000		2006-9570	
3.	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	Drought		37	2580000	0.05	2000-9186	
4.	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	Drought					1971-9085	
5.	11969	1969	Afghanistan	Paktia province	Drought	Drought			48000	0.2	1969-9007	
6.	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	935		2006-0405	
7.	13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	Earthquake (ground shaking)		5	501		2005-0686	
8.	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1		0.05	2005-0575	

Parse data as

- CSV / TSV / separator-based files
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)
- RDF/XML files

Character encoding

Update Preview

Columns are separated by
- commas (CSV)
- tabs (TSV)
- custom \t

Escape special characters with \

- [ ] Ignore first 0 line(s) at beginning of file
- [x] Parse next 1 line(s) as column headers
- [ ] Discard initial 0 row(s) of data
- [ ] Load at most 0 row(s) of data

- [x] Parse cell text into numbers, dates, ...
- [x] Quotation marks are used to enclose cells containing column separators
- [x] Store blank rows
- [x] Store blank cells as nulls
- [ ] Store file source (file names, URLs) in each row

Version 2.5 [r2407]

Help
About

11:52
07-04-2012

Slides from Sathishwaran.R and Vijaya Prabhu

Creating Project

Project Created

17828 rows

Extensions: Freebase

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 10** next › las

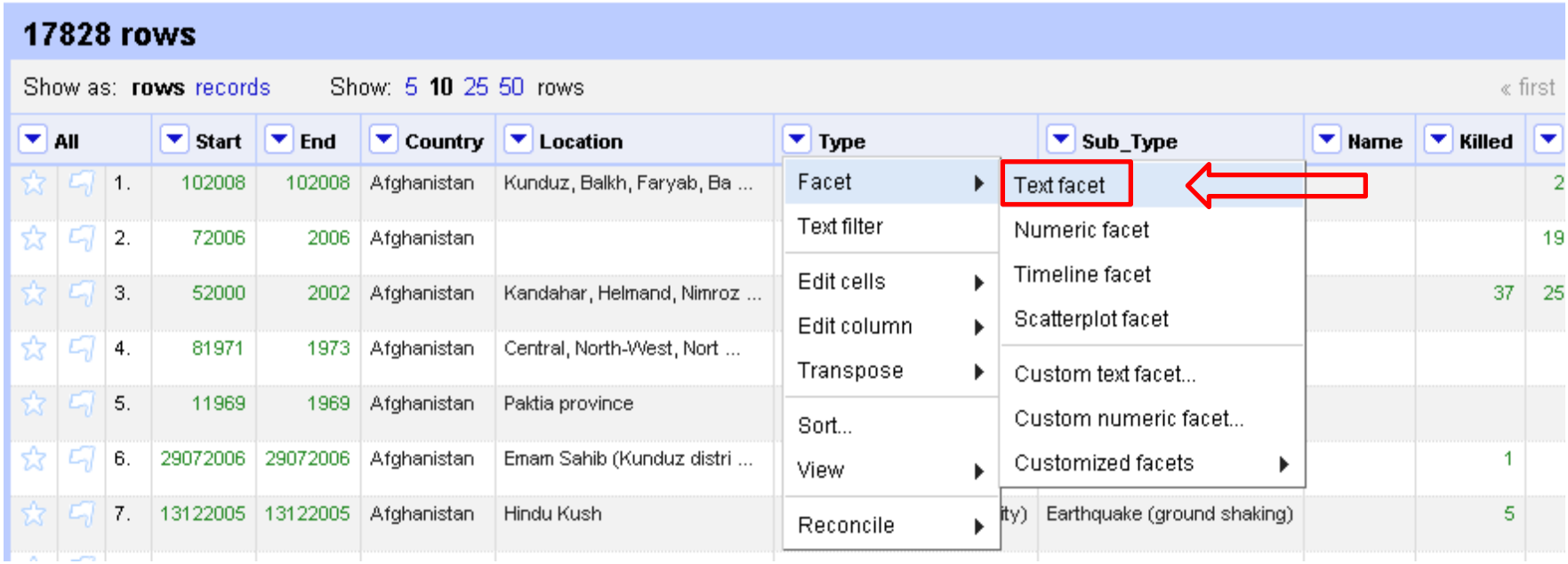
All		Start	End	Country	Location	Type	Sub_Type	Name	Killed	Cost	Affected	Id
1.		102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	Drought			280000		2008-9475
2.		72006	2006	Afghanistan		Drought	Drought			1900000		2006-9570
3.		52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	Drought		37	2580000	0.05	2000-9186
4.		81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	Drought					1971-9085
5.		11969	1969	Afghanistan	Paktia province	Drought	Drought			48000	0.2	1969-9007
6.		29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	935		2006-0405
7.		13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	Earthquake (ground shaking)		5	501		2005-0686
8.		8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1		0.05	2005-0575
9.		18072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	Earthquake (ground shaking)		2	1040		2004-0436
10.		10042003	10042003	Afghanistan	Yakabagh (Takhar province ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	1001		2003-0236

slides from Sathishwaran.R and Vijaya Prabhu

Faceting

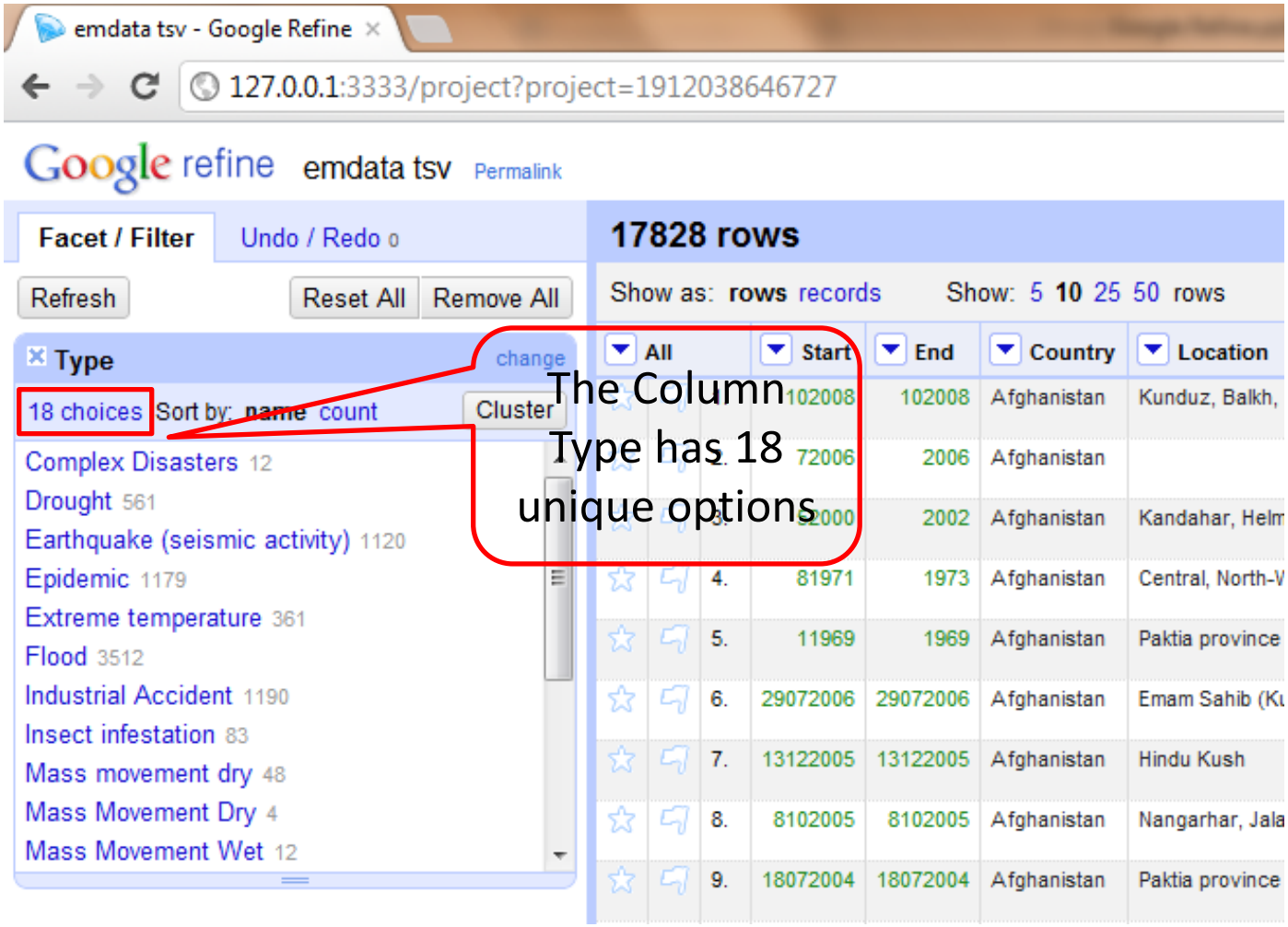
- Faceting is about seeing the big picture and filtering based on rows to work on data you want to change in bulk.
- We can create a facet for a column to get the details about that column and then we can filter to a subset of rows with a constraint.
- We can perform text facet, Numeric facet, timeline facet and scatterplot facet. Also various customized facets can be designed.

Faceting



slides from Sathishwaran.R and Vijaya Prabhu

Faceting



The Column Type has 18 unique options

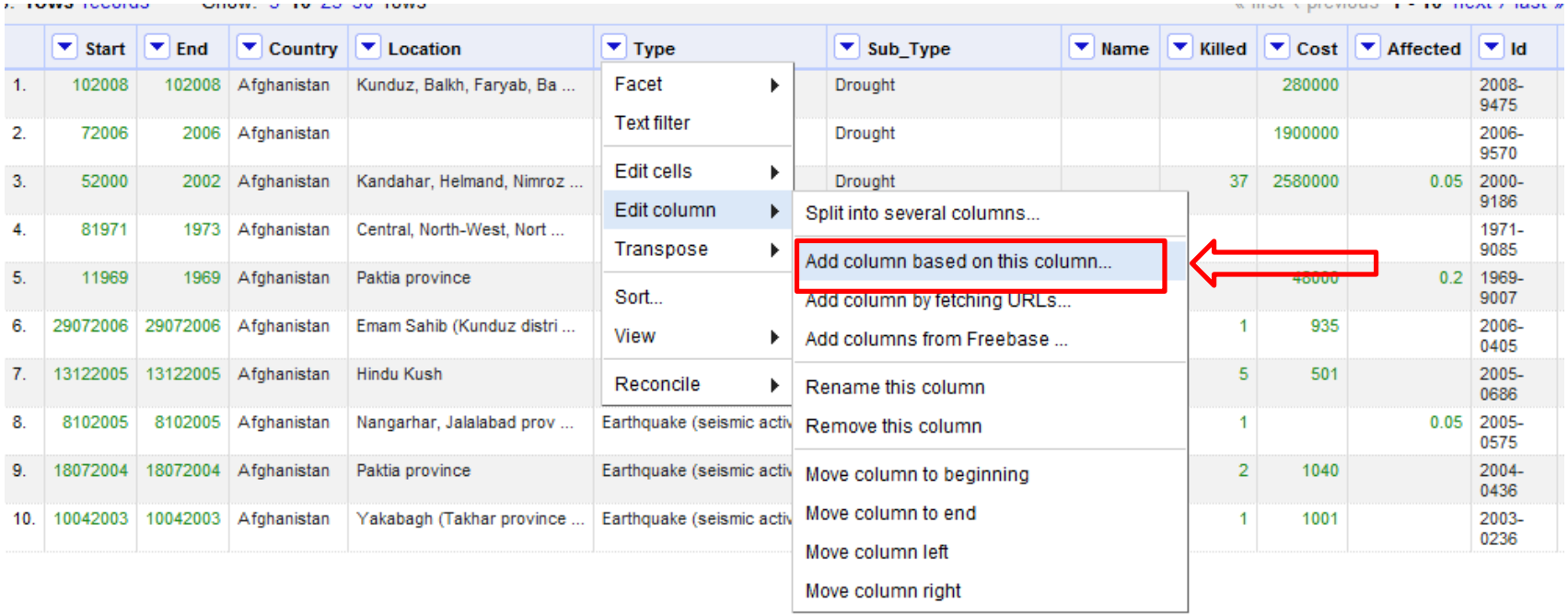
slides from Sathishwaran.R and Vijaya Prabhu

Removing Redundancy

Even though they are of same type, shows as different options due to case

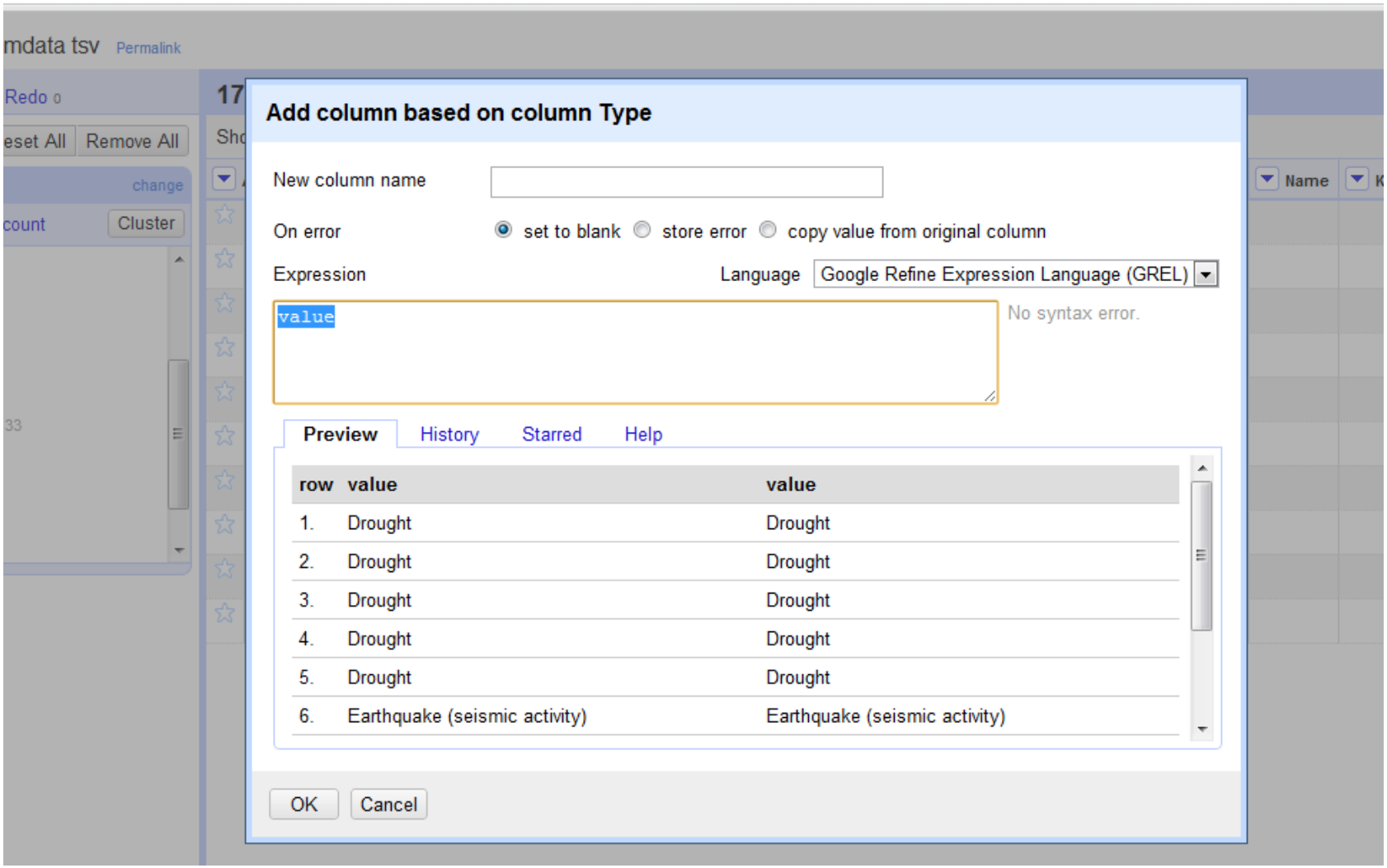
slides from Sathishwaran.R and Vijaya Prabhu

Removing Redundancy



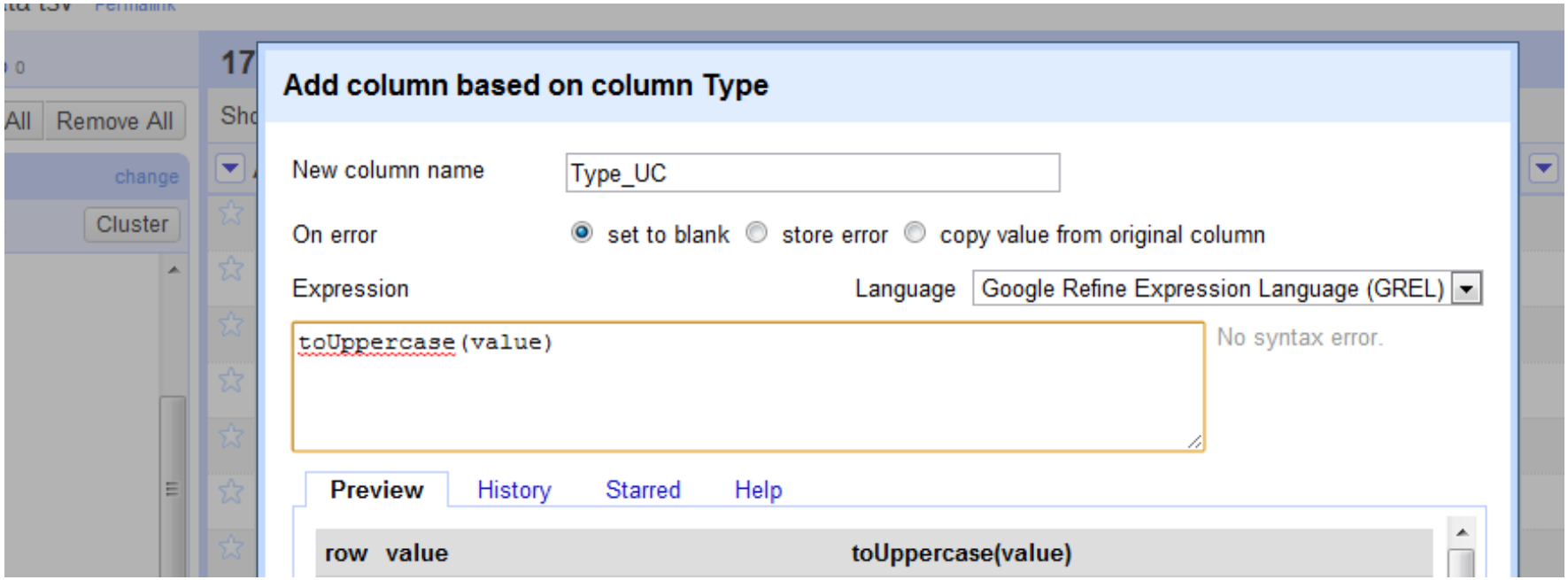
slides from Sathishwaran.R and Vijaya Prabhu

Removing Redundancy



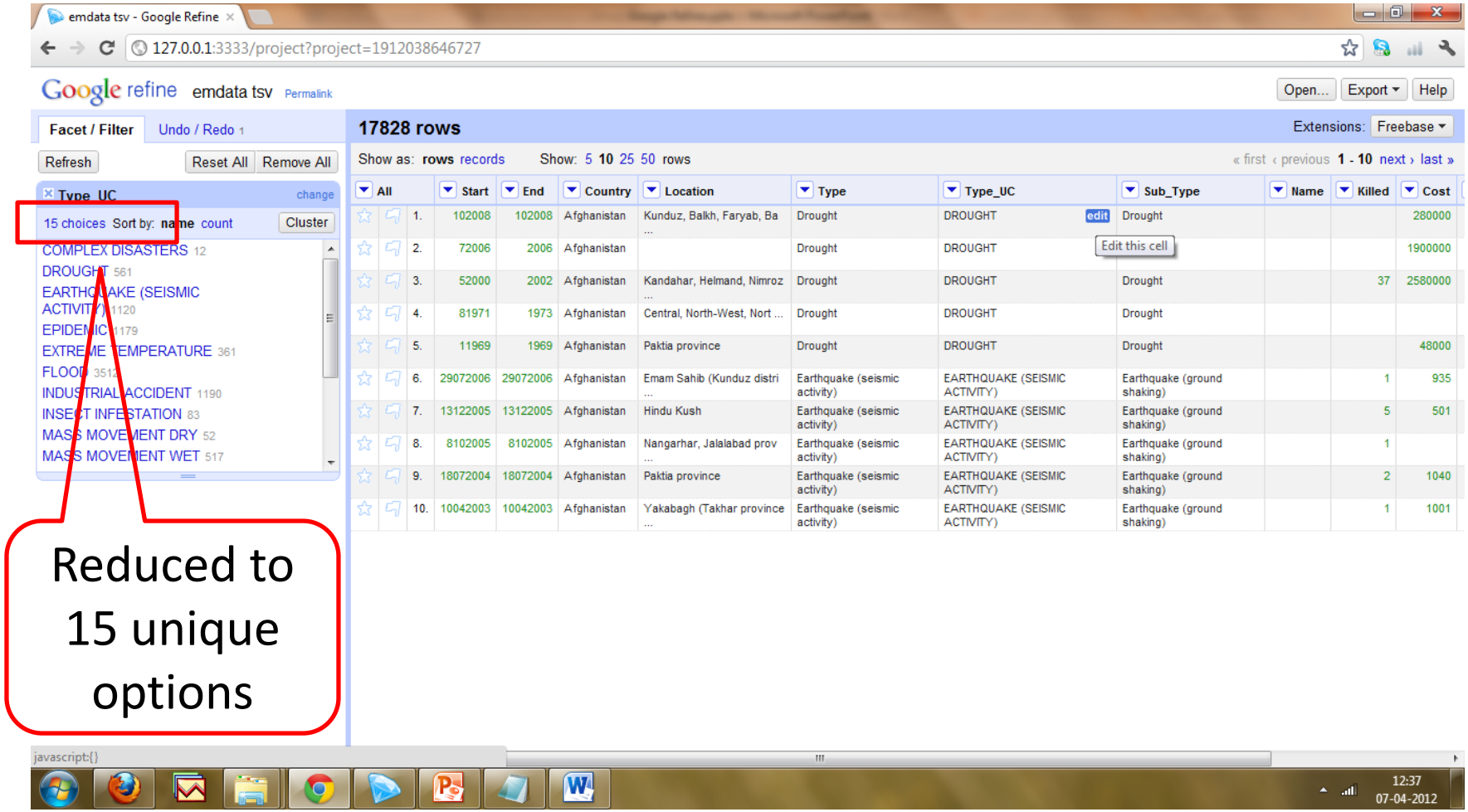
slides from Sathishwaran.R and Vijaya Prabhu

Removing Redundancy



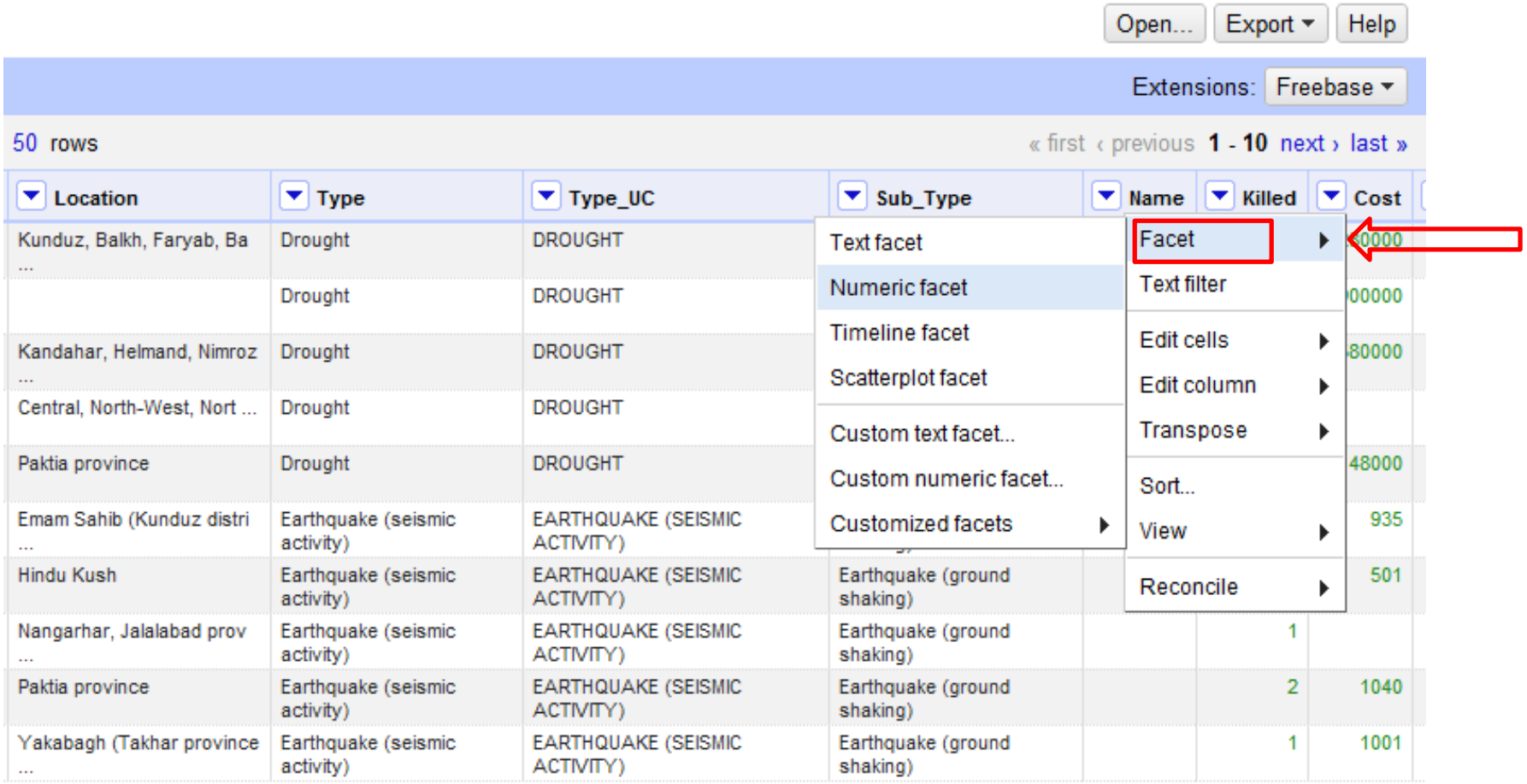
slides from Sathishwaran.R and Vijaya Prabhu

Removing Redundancy



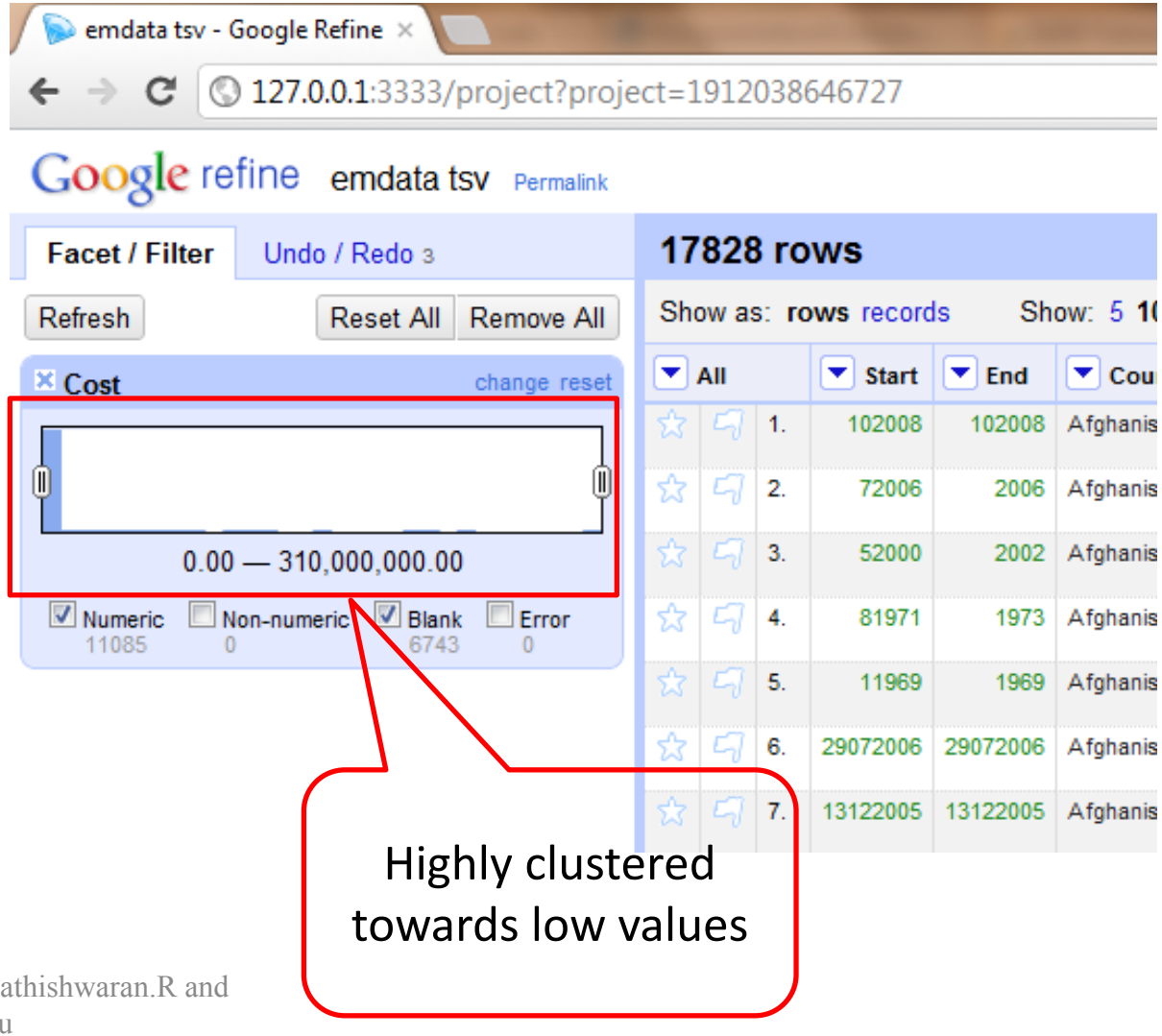
slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting



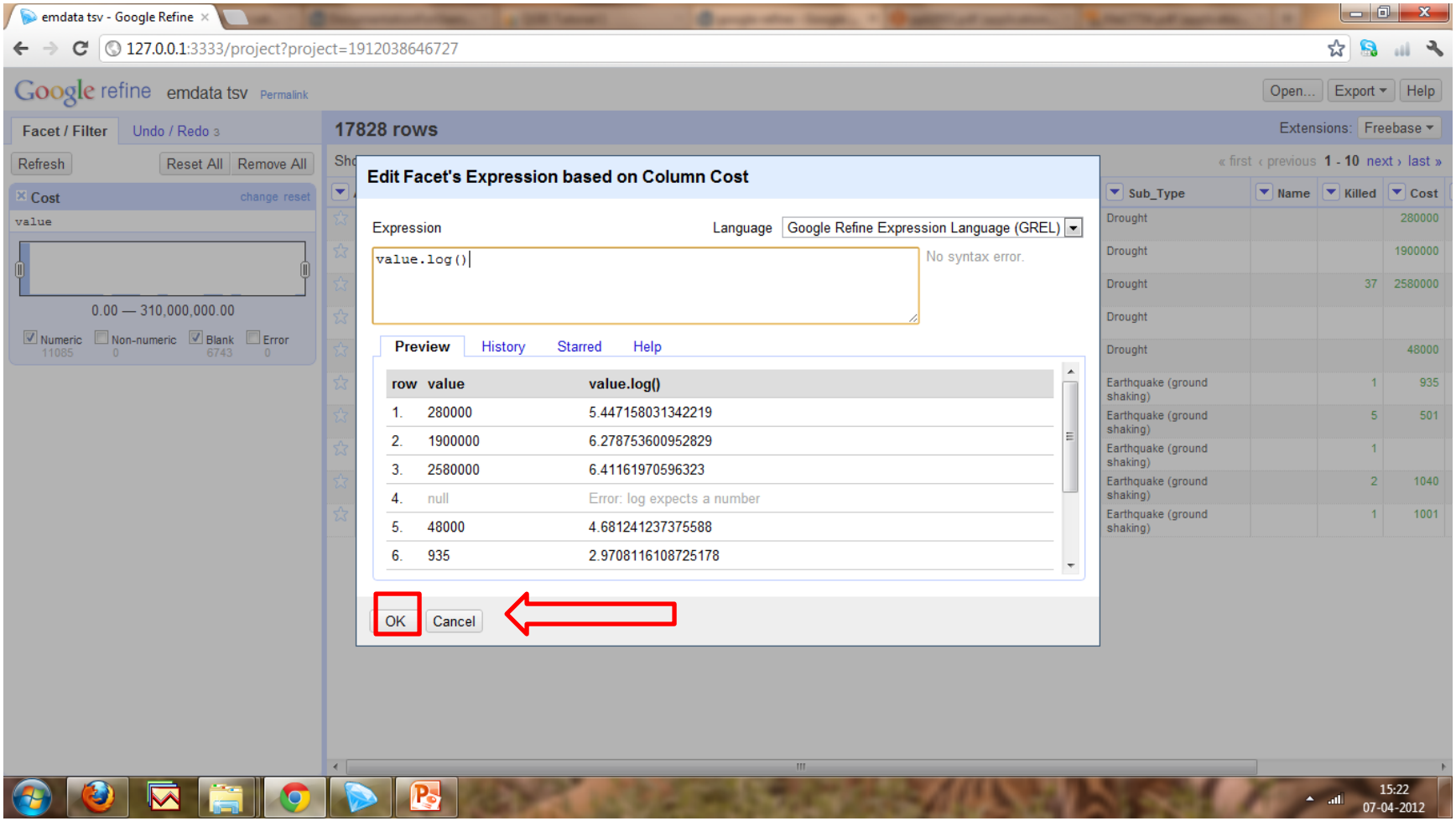
slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting



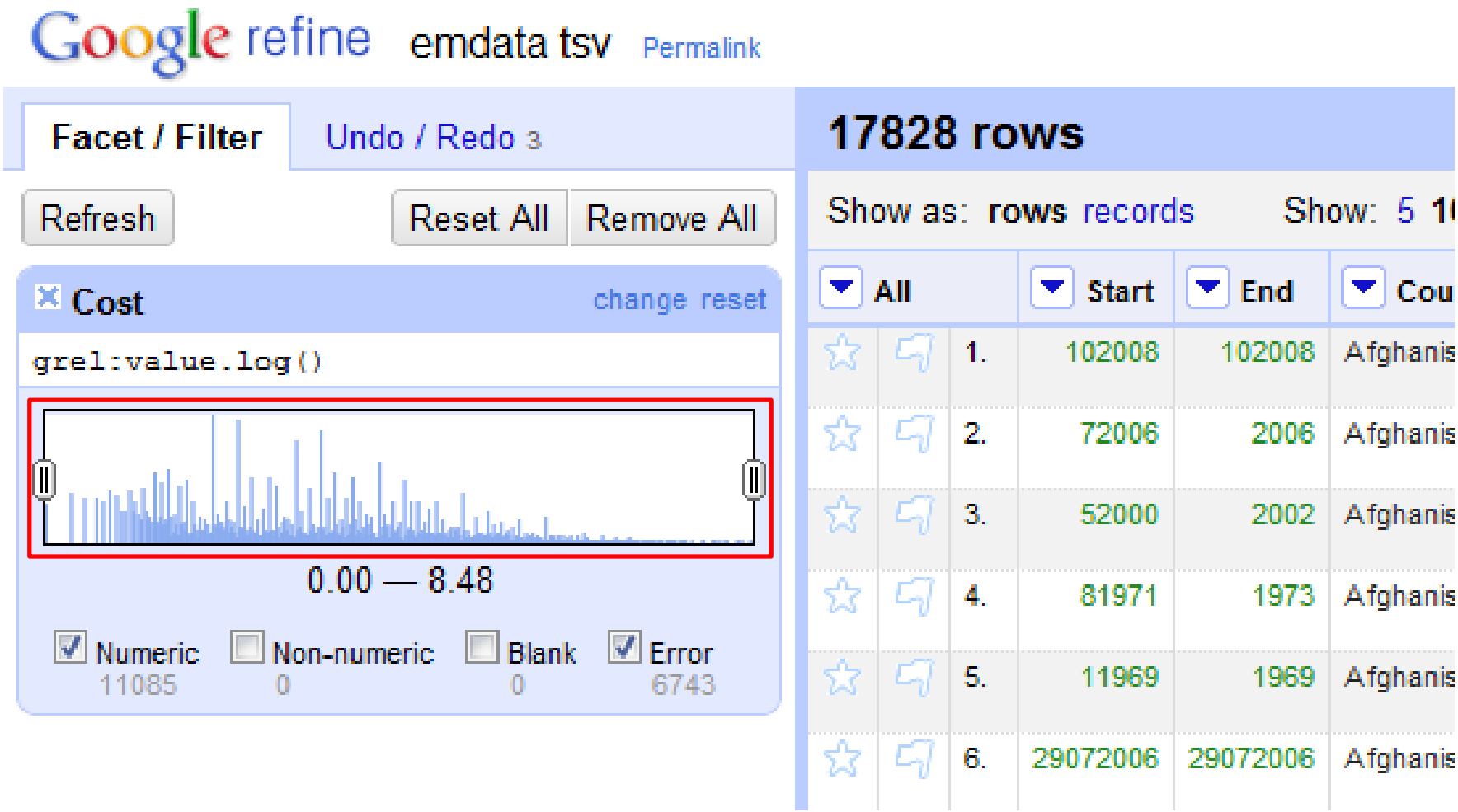
slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting



slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting



slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting

emdata tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata tsv Permalink

Open... Export Help

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

Cost change reset

grel:value.log()

0.00 — 8.48

Numeric 11085 Non-numeric 0 Blank 0 Error 6743

6743 matching rows (17828 total)

Extensions: Freebase

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

« first ‹ previous 1 - 10 next › last »

All		Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost
4.	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	DROUGHT	Drought				
8.	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1		
15.	25022001	edit 2001	Afghanistan	Faizabad region (Badakhsh ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)				
16.	19012000	19012000	Afghanistan	Hindu Kush	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)				
18.	20021998	20021998	Afghanistan	Rostaq area	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)				
21.	10051997	10051997	Afghanistan	Herat area	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			5	
24.	20051992	20051992	Afghanistan		Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			14	
25.	18041991	18041991	Afghanistan	Keshim (Badakhshan provin ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			24	
32.	10061954	10061954	Afghanistan	North Region	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			2000	
36.	24102002	24102002	Afghanistan	Badakhshan	Epidemic	EPIDEMIC		Unknown		100	

Cost column is blank and has no value

15:26 07-04-2012

slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting

emdata tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata tsv Permalink

Open... Export Help

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

Cost change reset

grel:value.log()

0.00 — 0.94

Numeric 11085 Non-numeric 0 Blank 0 Error 6743

699 matching rows (17828 total)

Extensions: Freebase

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

« first ‹ previous 1 - 10 next › last »

All		Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost
52.	25031998	8051998	Afghanistan	Rustaq district (province ...	Epidemic	EPIDEMIC	Viral Infectious Diseases	Acute haemorrhagic fever syndrome	12	7
85.	29072003	30072003	Afghanistan	Panjshir valley (Kapisa p ...	Flood	FLOOD	General flood		24	3
136.	20072002	20072002	Afghanistan	Bamiyan region	Miscellaneous accident	MISCELLANEOUS ACCIDENT	Explosion		13	6
153.	8062002	8062002	Afghanistan	Tangi Abrishim	Transport Accident	TRANSPORT ACCIDENT	Road		45	2
230.	9102002	12102002	Algeria	Batna, Biskra, Hamma and ...	Flood	FLOOD	General flood		13	4
233.	13102000	13102000	Algeria	Ain TÃ©mouchent	Flood	FLOOD	Flash flood		9	5
268.	22005	22005	Algeria	Annaba, ChetaÃ¯bi, Kanylie ...	Storm	STORM			10	5
274.	20082006	20082006	Algeria	Near El Ancer (Jijel regi ...	Transport Accident	TRANSPORT ACCIDENT	Road		11	7
275.	1012006	1012006	Algeria	Near Messaad (Djelfa regi ...	Transport Accident	TRANSPORT ACCIDENT	Road		[illegible]	5
282.	29082003	29082003	Algeria	BÃ©ni Saf (AÃ¯n TÃ©mouchent ...	Transport Accident	TRANSPORT ACCIDENT	Road		10	1

Calamities with low cost

15:33 07-04-2012

slides from Sathishwaran.R and Vijaya Prabhu

Numeric Faceting

emdata tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata tsv Permalink

Open... Export Help

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

Cost change reset

grel:value.log()

7.52 — 8.46

Numeric 11085 Non-numeric 0 Blank 0 Error 6743

30 matching rows (17828 total)

Extensions: Freebase

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

« first ‹ previous 1 - 10 next › last »

All	Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost	Affe
940.	20062004	3082004	Bangladesh	Sylhet, Sunamganj, Moulvi ...	Flood	FLOOD	General flood		730	36000000	
979.	61988	91988	Bangladesh		Flood	FLOOD			2379	45000000	
997.	71974	71974	Bangladesh		Flood	FLOOD			28700	38000000	5
2663.	12003	12003	China P Rep	Inner Mongolia Autonomous ...	Drought	DROUGHT	Drought			48000000	
2666.	42002	2002	China P Rep	Guangdong, Fujian, Guangx ...	Drought	DROUGHT	Drought			60000000	
2676.	11994	121994	China P Rep	Xuwen, Leizhou, Yangjiang ...	Drought	DROUGHT	Drought			82000000	137
2680.	61988	1988	China P Rep	Hubei,Jiangsu, Henan, Anh ...	Drought	DROUGHT	Drought		1400	49000000	942
2694.	12052008	12052008	China P Rep	Wenchuan country, Wencgua ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		87476	45976596	8
2815.	10012008	5022008	China P Rep	Zhejiang, Sichuan, Anhui, ...	Extreme temperature	EXTREME TEMPERATURE	Extreme winter conditions		129	77000000	2
2840.	15062007	72007	China P Rep	Sichuan, Anhui, Hubei , S ...	Flood	FLOOD	General flood		535	105004000	4425

Calamities with high cost

slides from Sathishwaran.R and Vijaya Prabhu

Clustering

- Clustering is used to merge choices which look similar.

emdata tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata tsv Permalink

Open... Export Help

Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

Country change

227 choices Sort by: name count Cluster

Afghanistan 163
Albania 26
Algeria 113
American Samoa 6
Angola 82
Anguilla 14
Antigua and Barbuda 11
Argentina 111
Armenia 11
Australia 221
Austria 49

17828 rows

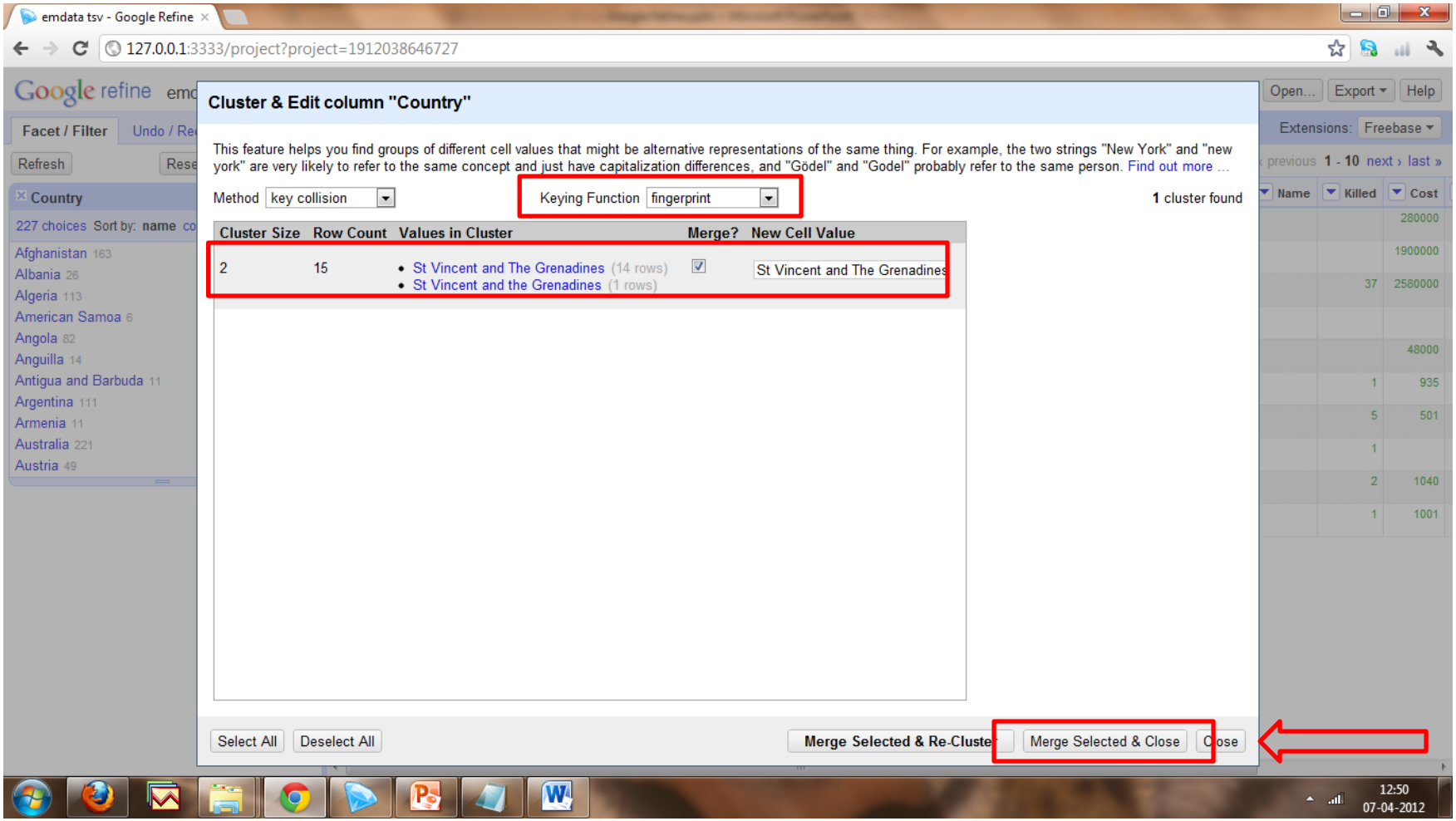
Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All		Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost
	1.	[illegible]	[illegible]008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	DROUGHT	Drought			280000
	2.	72006	2006	Afghanistan		Drought	DROUGHT	Drought			1900000
	3.	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	DROUGHT	Drought		37	2580000
	4.	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	DROUGHT	Drought			
	5.	11969	1969	Afghanistan	Paktia province	Drought	DROUGHT	Drought			48000
	6.	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	935
	7.	13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		5	501
	8.	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	
	9.	18072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		2	1040
	10.	10042003	edit 2003	Afghanistan	Yakabagh (Takhar province ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	1001

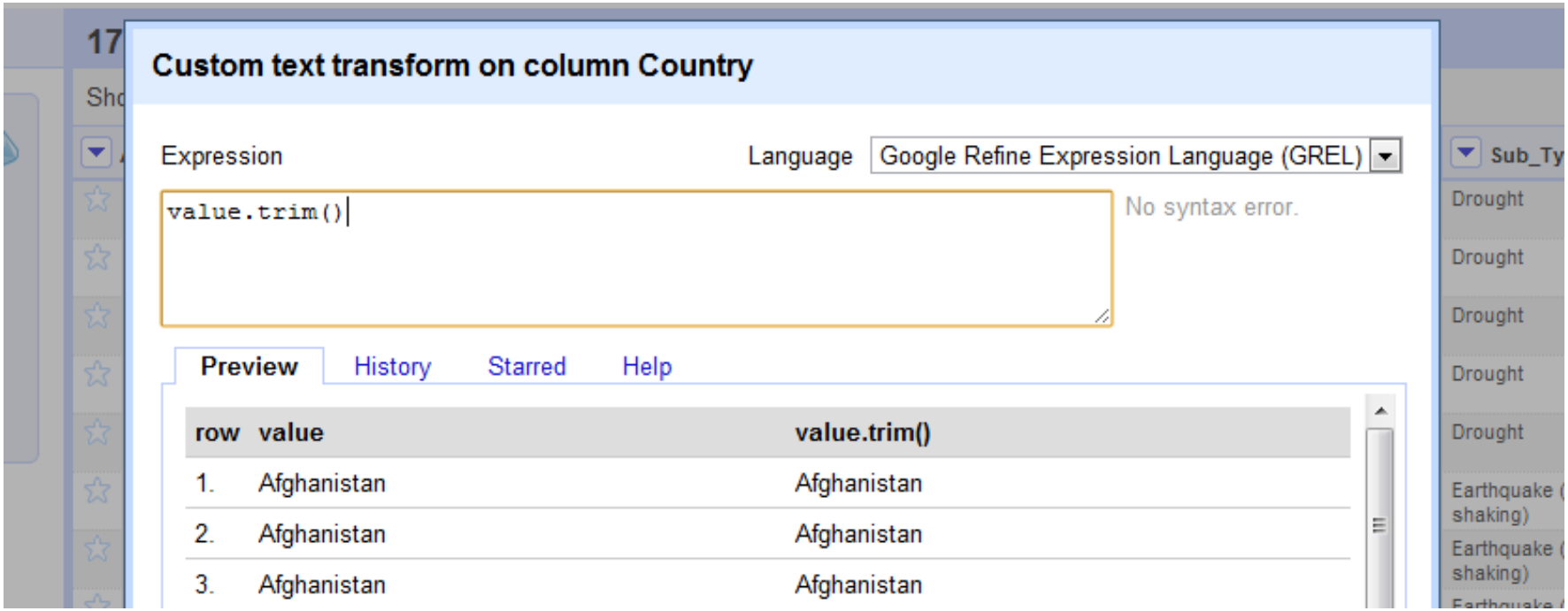
12:45
07-04-2012

Clustering



slides from Sathishwaran.R and Vijaya Prabhu

Using Expressions



slides from Sathishwaran.R and Vijaya Prabhu

Using Expressions

emdata tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata tsv Permalink

Text transform on 1 cells in column Country:
grel:value.trim() Undo

Open... Export Help

Facet / Filter Undo / Redo 3

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

17828 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All		Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost
1.	102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	DROUGHT	Drought				280000
2.	72006	2006	Afghanistan		Drought	DROUGHT	Drought				1900000
3.	52000	2002	Afghanistan	Kandahar, Helmand, Nimroz ...	Drought	DROUGHT	Drought			37	2580000
4.	81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	DROUGHT	Drought				
5.	11969	1969	Afghanistan	Paktia province	Drought	DROUGHT	Drought				48000
6.	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			1	935
7.	13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			5	501
8.	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			1	
9.	18072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			2	1040
10.	10042003	10042003	Afghanistan	Yakabagh (Takhar province ...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)			1	1001

14:37
07-04-2012

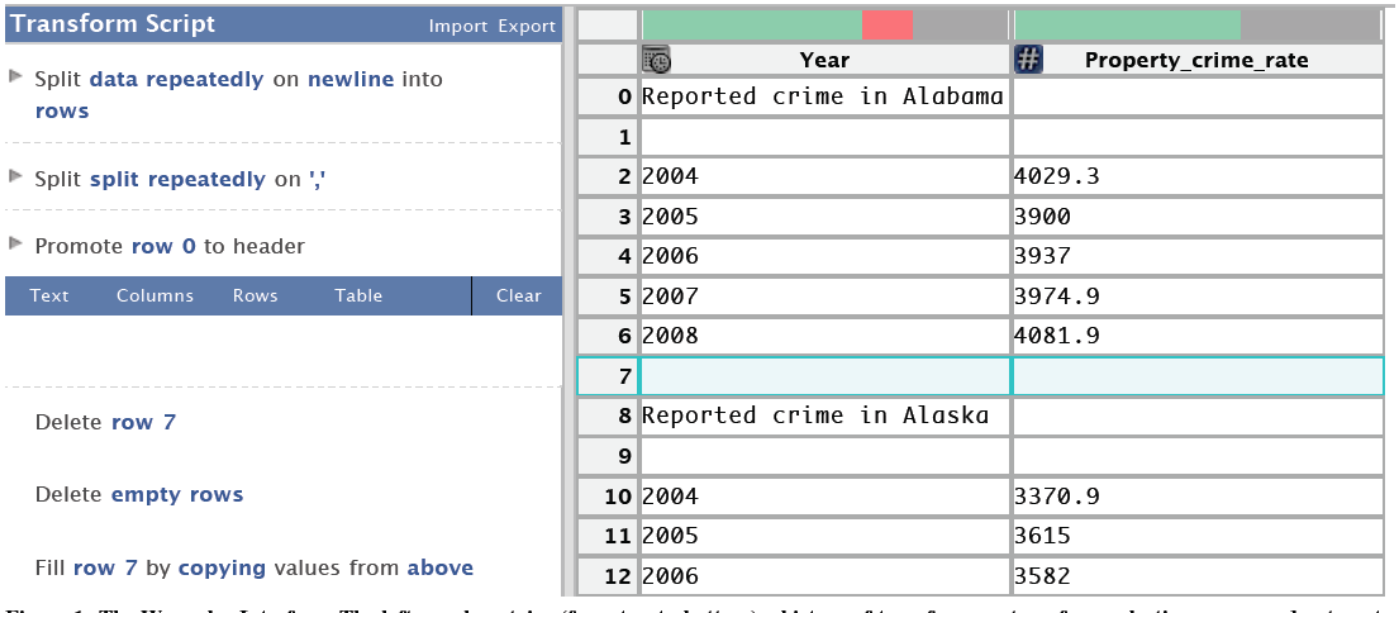
slides from Sathishwaran.R and Vijaya Prabhu

OpenRefine

- Nice tutorial
  - http://www.intersect.org.au/docs/GoogleRefine_Slides.pdf
  - http://www.intersect.org.au/docs/GoogleRefine_Exercises.pdf

Data Wrangler

- An interactive system for creating data transformations
  - Suggest transforms
  - Iteratively explore the space of applicable operations



[(Video) http://vimeo.com/19185801](http://vimeo.com/19185801)

Inference Engine

- Inference parameter sets based on users' input
  - Row, type and text selection
- Identify all compatible operations
- Rank the operations

An Example

Transform Script Import Export

- Split **data repeatedly** on **newline** into **rows**
- Split **split repeatedly** on **','**
- Promote **row 0** to header
- Delete **empty rows**
- Extract from **Year** after **'in '**
- Set **extract's** name to **State**
- Fill **State** by **copying** values from **above**

Text Columns Rows Table Clear

Delete **rows where Year starts with 'Reported'**

Delete **rows where Year contains 'Reported'**

Extract from **Year** between positions **0, 8**

	Year	State	Property
0	Reported crime in Alabama	Alabama	
1	2004	Alabama	4029.3
2	2005	Alabama	3900
3	2006	Alabama	3937
4	2007	Alabama	3974.9
5	2008	Alabama	4081.9
6	Reported crime in Alaska	Alaska	
7	2004	Alaska	3370.9
8	2005	Alaska	3615
9	2006	Alaska	3582
10	2007	Alaska	3373.9
11	2008	Alaska	2928.3
12	Reported crime in Arizona	Arizona	
13	2004	Arizona	5073.3
14	2005	Arizona	4827
15	2006	Arizona	4741.6
16	2007	Arizona	4502.6
17	2008	Arizona	4087.3
18	Reported crime in Arkansas	Arkansas	
19	2004	Arkansas	4033.1

FlashFill in Excel

- Transforming by entering input-output examples

  (https://www.youtube.com/watch?v=aMdnbMQImVg)

A Data Table

Accession	Credit	Dimensions	Medium	Name
01.2	Gift of the artist	5.25 in HIGH x 9.375 in WIDE	Oil on canvas	John Mix Stanley
05.411	Gift of James L. Edison	20 in HIGH x 24 in WIDE	Oil on canvas	Mortimer L. Smith
06.1	Gift of the artist	Image: 20.5 in. HIGH x 17.5 in. WIDE	Oil on canvas	Theodore Scott Dabo
06.2	Gift of the artist	9.75 in|16 in HIGH x 13.75 in|19.5 in WIDE	Oil on canvas	Leon Dabo
…				
09.8	Gift of the artist	12 in|14 in HIGH x 16 in|18 in WIDE	Oil on canvas	Gari Melchers

Programming by Example

	Raw Value	Target Value
R1	5.25 in HIGH x 9.375 in WIDE	**9.375**
R2	20 in HIGH x 24 in WIDE	**24**
R3	Image: 20.5 in. HIGH x 17.5 in. WIDE	**17.5**
R4	9.75 in|16 in HIGH x 13.75 in|19.5 in WIDE	13.75
	. . .	
R5	12 in|14 in HIGH x 16 in|18 in WIDE	null

Transformation Program

BNK: blankspace
NUM[0-9]+: 98
UWRD[A-Z]: I
LWRD[a-z]: mage
WORD[a-zA-Z]
START:
END:

Transform(*value*)

Conditional Statement

```
label = classify(value)
switch label:
    case "partition1":
```

Partition Transformation Program

pos_1 = value.indexOf('BNK', 'NUM', -1)
pos_2 = value.indexOf('NUM','BNK',2)
output = value.substring(pos_1,pos_2)

```
    case "partition2":
```

Partition Transformation Program

pos_1 = value.indexOf('BNK', 'NUM', 1)
pos_2 = value.indexOf('NUM','BNK',3)
output = value.substring(pos_1,pos_2)

return output

Creating Hypothesis Spaces

- Segmenting the Outputs (n! traces)

1st

I: 5.25 in HIGH x 9.375 in WIDE

O: 9.375

Constant

2nd 9.375

3nd 9.375

…

O(N^2) unique substrings

- Create Hypothesis Space

Substring

1.subString(starPosition, endPosition)
2. Constant String

StartPostion
1. 15
2. (BNK, NUM, 1)
3. (BNK,9,-1)
4. (LWRD BNK, NUM '.',1)
…

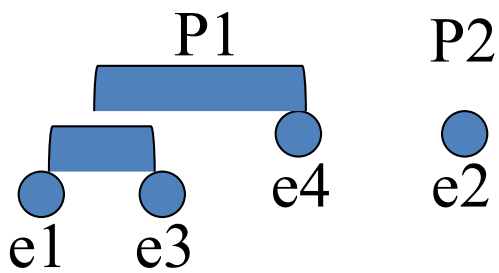
Generating Programs

- Generate programs from Hypothesis Space
  - Generate and then verify
  - Select values for non-terminals based on a partial order
    - Number of the Segments
    - Length of the context
    - Token class or actual text

Learn Conditional Statement

- Clustering:

1. Agglomerative Clustering



2. Compatibility Score

- Thank you