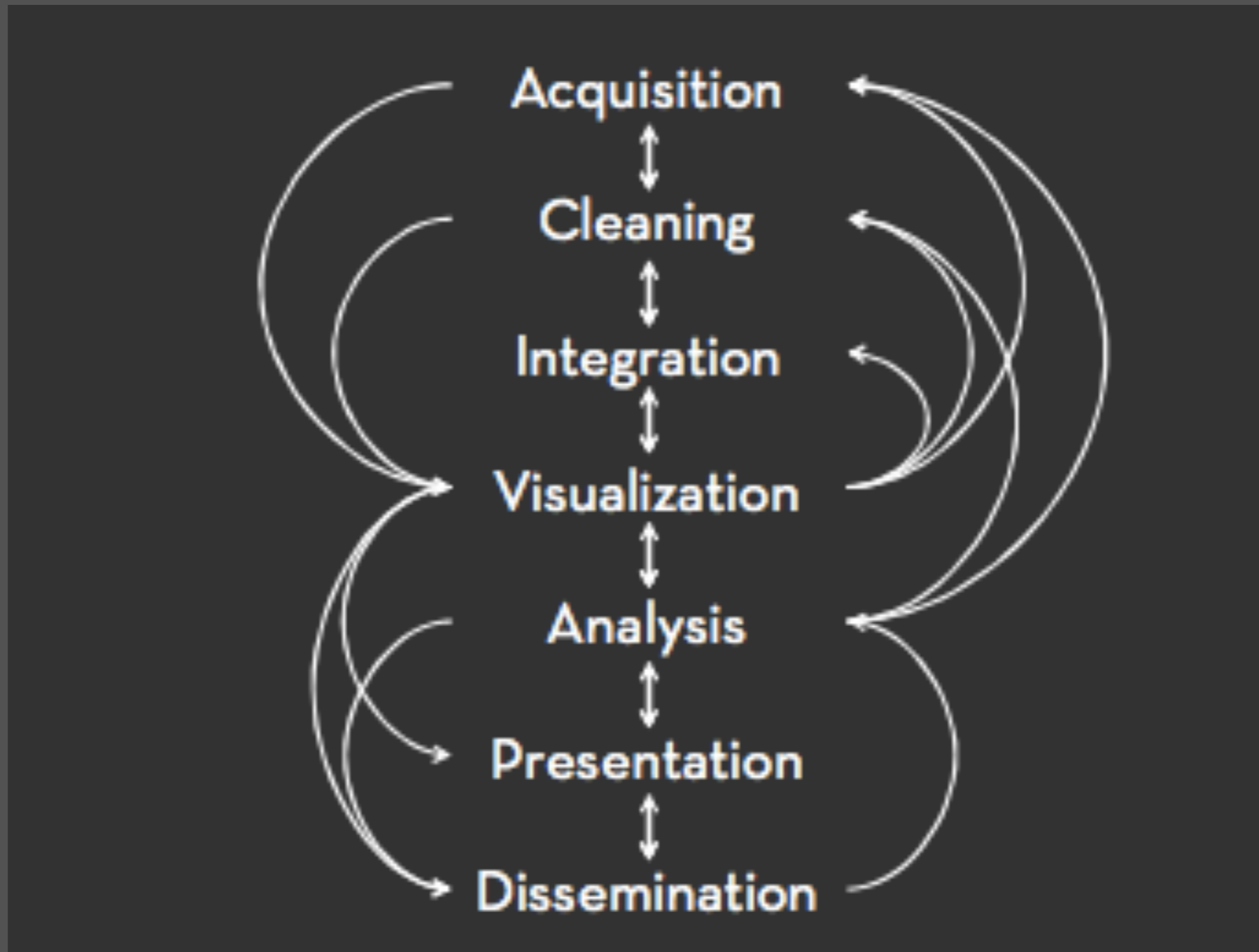


Data cleaning

Data lifecycle



Graphic: Jeff Heer

Open Refine

Google refine government IT contracts [Permalink](#)

Facet / Filter Undo / Redo 5

Refresh Reset All Remove All

Type of Contract change
783 choices Sort by: name count Cluster

- Firm Fixed Price 836
- FFP: Firm Fixed Price 612
- T&M: Time & Materials 351
- Time and Materials 232
- Time & Materials 189 edit include
- CPFF: Cost Plus Fixed Fee 183
- CPAF: Cost Plus Award Fee 130
- Task Based Indefinite Delivery/Indefinite Quantity (ID/IQ) Time & Materials (T&M) Task Order 115
- Firm-Fixed-Price 115
- Fixed Price 105

5200 rows

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

All	Contract ID	Contractor Name	Type of Contract	Date of Aw
1.	1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2009
2.	1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/2009
3.	1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/2009
4.	1942	ITS CORPORATION	Time & Materials	12/31/2008
5.	7490	SENET INTERNATIONAL CORPORATIO	Firm Fixed Price C&A	05/04/2009
6.	1945		firm fixed price	01/26/2009
7.	1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY	firm fixed price	10/01/2009
8.	1947		firm fixed price	09/30/2009

Data Wrangler (Stanford)

Data Wrangler

Transform Script Import Export

- Split data repeatedly on newline into rows
- Split split repeatedly on ','
- Promote row 0 to header
- Delete empty rows

Text Columns Rows Table Clear

Extract from Year after 'in'

Extract from Year after 'in'

Cut from Year after 'in'

Cut from Year after 'in'

Split Year after 'in'

Split Year after 'in'

	Year	extract	Property_crime_rate
0	Reported crime in Alabama	Alabama	
1	2004		4029.3
2	2005		3900
3	2006		3937
4	2007		3974.9
5	2008		4081.9
6	Reported crime in Alaska	Alaska	
7	2004		3370.9
8	2005		3615
9	2006		3582
10	2007		3373.9
11	2008		2928.3
12	Reported crime in Arizona	Arizona	
13	2004		5073.3
14	2005		4827
15	2006		4741.6
16	2007		4502.6
17	2008		4087.3
18	Reported crime in Arkansas	Arkansas	
19	2004		4033.1
20	2005		4068
21	2006		4021.6
22	2007		3945.5
23	2008		3843.7
24	Reported crime in California	California	
25	2004		3423.9
26	2005		3321
27	2006		3175.2
28	2007		3032.6
29	2008		2940.3
30	Reported crime in Colorado	Colorado	

<https://vimeo.com/19185801>

Problem

No single identifier for a person in the political process

Cannot ask questions like:

What are the incumbency metrics?

What is the political trajectory of a given person?

More or less people participating? Are they staying longer?

Age/Caste/Gender trends?

Regional differences?

...

Surf

Entity resolution tool to map similar values in databases

Approximate “Indian name” matching

e.g. V -> W, EE -> I, TH->T, KSH -> X, SH->S

Drop titles: Mr/Mrs./Dr./Adv./General/Retd./Thiru, etc.

Edit distance metric (LALKRUSHNA ADVANI = LAL KRISHNA ADVANI)

Compatible names (e.g. L.K. ADVANI = ADVANI, LAL KRISHNA)

User interface for user to make difficult calls

<input type="checkbox"/>	PATEL JAYANTBHAI RAMANBHAI (BOSKEY)	SARSA	1965	INC	2	M	Gujarat	30465
<input type="checkbox"/>	PATEL JAYANTBHAI (BOSKI) RAMANBHAI	SARSA	1990	IND	2	M	Gujarat	20839
<input type="checkbox"/>	JAYANTBHAI RAMANBHAI PATEL (BOSKEY)	SARSA	2007	NCP	1	M	Gujarat	45811
<input type="checkbox"/>	JAYANTBHAI RAMANBHAI PATEL (BOSKEY)	UMRETH	2012	NCP	1	M	Gujarat	67363
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	KANTAWALA SARADCHANDRA GULABCHAND	SURAT CITY EAST	1965	JD	5	M	Gujarat	3338
<input type="checkbox"/>	SHARADCHANDRA GULABCHAND KANTAWALA	SURAT CITY WEST	1980	IND	4	M	Gujarat	749
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	SONIMAHAJAN NARENDRABHAI AMRATLAL	RAJKOT-II	2002	IND	5	M	Gujarat	799
<input type="checkbox"/>	SONIMAHAJAN NARENDRABHAI AMRUTBHAI	RAJKOT-II	2007	NSCP	6	M	Gujarat	334
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	PATEL DR. PRAVINCHANDRA HANSRAJBHAI	GONDAL	1960	IND	4	M	Gujarat	711
<input type="checkbox"/>	PRAVINCHANDRA HANSARAJBHAI PATEL	GONDAL	1990	IND	6	M	Gujarat	132
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	KAPDIYA MOHMAD IRFAN GULAM MOHMAD	OLPAD	2012	LJP	12	M	Gujarat	360
<input type="checkbox"/>	KAPADIA MOHMAD IRFAN GULAM MOHMAD	SURAT CITY WEST	2002	BSP	4	M	Gujarat	821
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	BHOLABHAI CHATURBHAI PATEL	VISNAGAR	1965	IND	1	M	Gujarat	30557
<input type="checkbox"/>	PARTEL BHOLABHAI CHATURBHAI	VISNAGAR	1990	JD	1	M	Gujarat	46564
<input type="checkbox"/>	PATEL BHOLABHAI CHATURBHAI	VISNAGAR	1995	INC	2	M	Gujarat	49770
<input type="checkbox"/>	PATEL BHOLABHAI CHATURBHAI	VISNAGAR	1998	IND	2	M	Gujarat	30539
<input type="checkbox"/>	BHOLABHAI CHATURBHAI PATEL	VISNAGAR	2012	NCP	2	M	Gujarat	46785
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	GAIKWAD NARENDRASINGH VITTHALRAO	GANDHINAGAR	1965	DDP	5	M	Gujarat	562
<input type="checkbox"/>	GAYAKWAD NARENDRASINH VITTHALRAY	MEGHRAJ	1990	DDP	5	M	Gujarat	1008
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								
<input type="checkbox"/>	MAHARAUL URVASHI DEVI JAIDEEPSINH	DEVGADH BARIA	1995	INC	2	F	Gujarat	37201
<input type="checkbox"/>	MAHARAUL URVASHIDEVI JAYDIPSINH	DEVGADH BARIA	1998	INC	1	F	Gujarat	48657
<input type="button" value="Select all"/> <input type="button" value="Mark as reviewed"/> <input type="button" value="Select till here"/> <input type="button" value="Mark reviewed till here"/>								