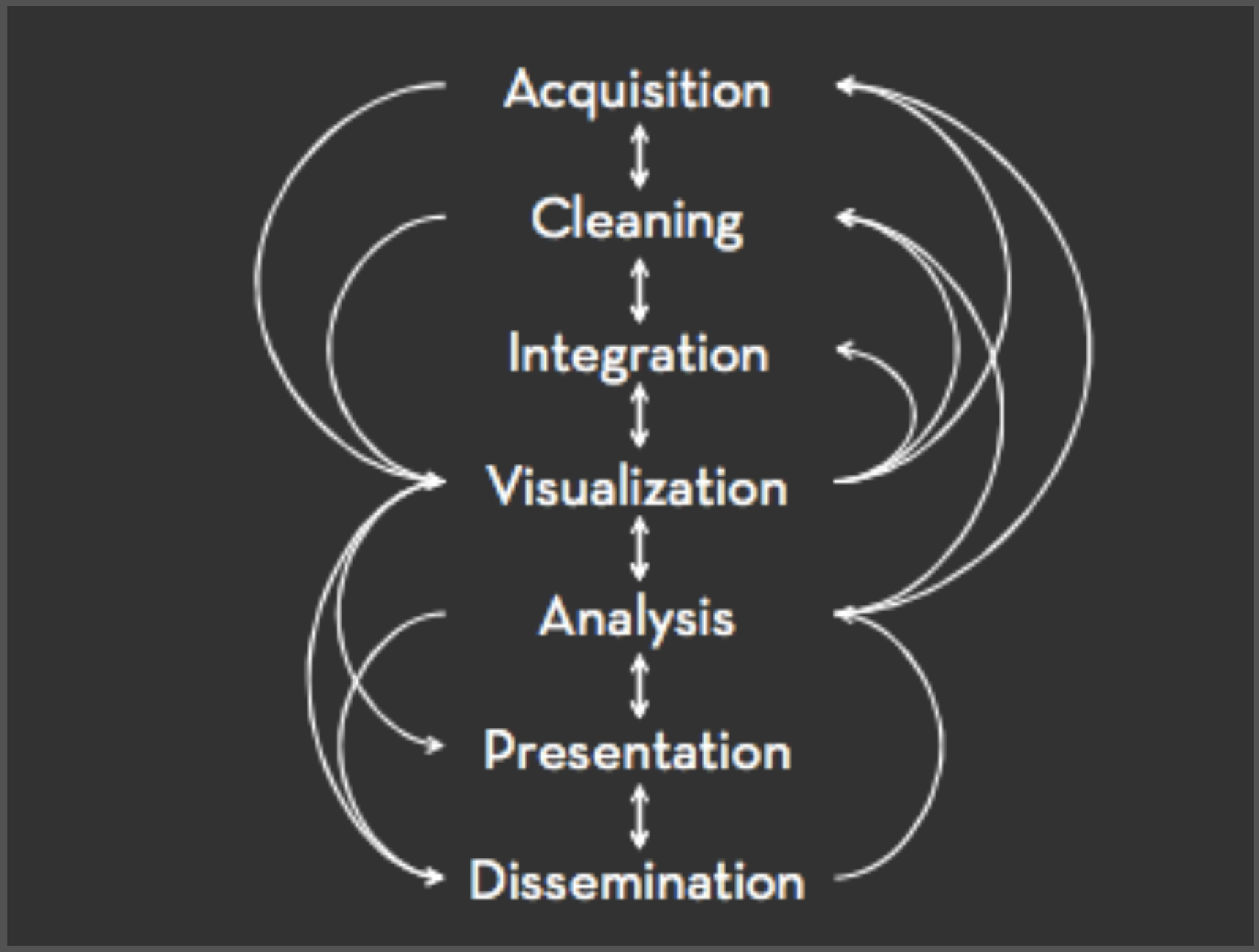
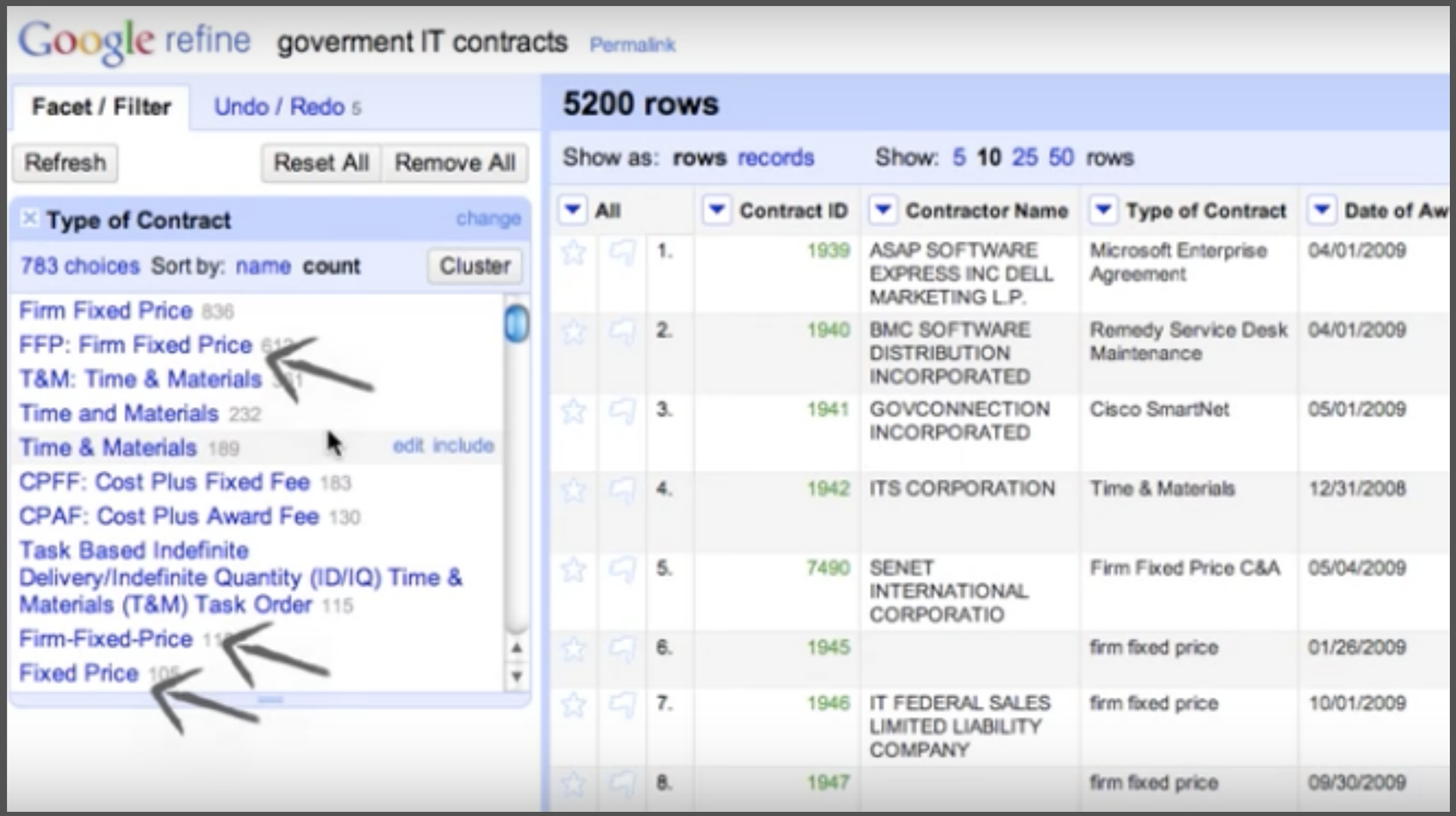
# Data cleaning

# Data lifecycle



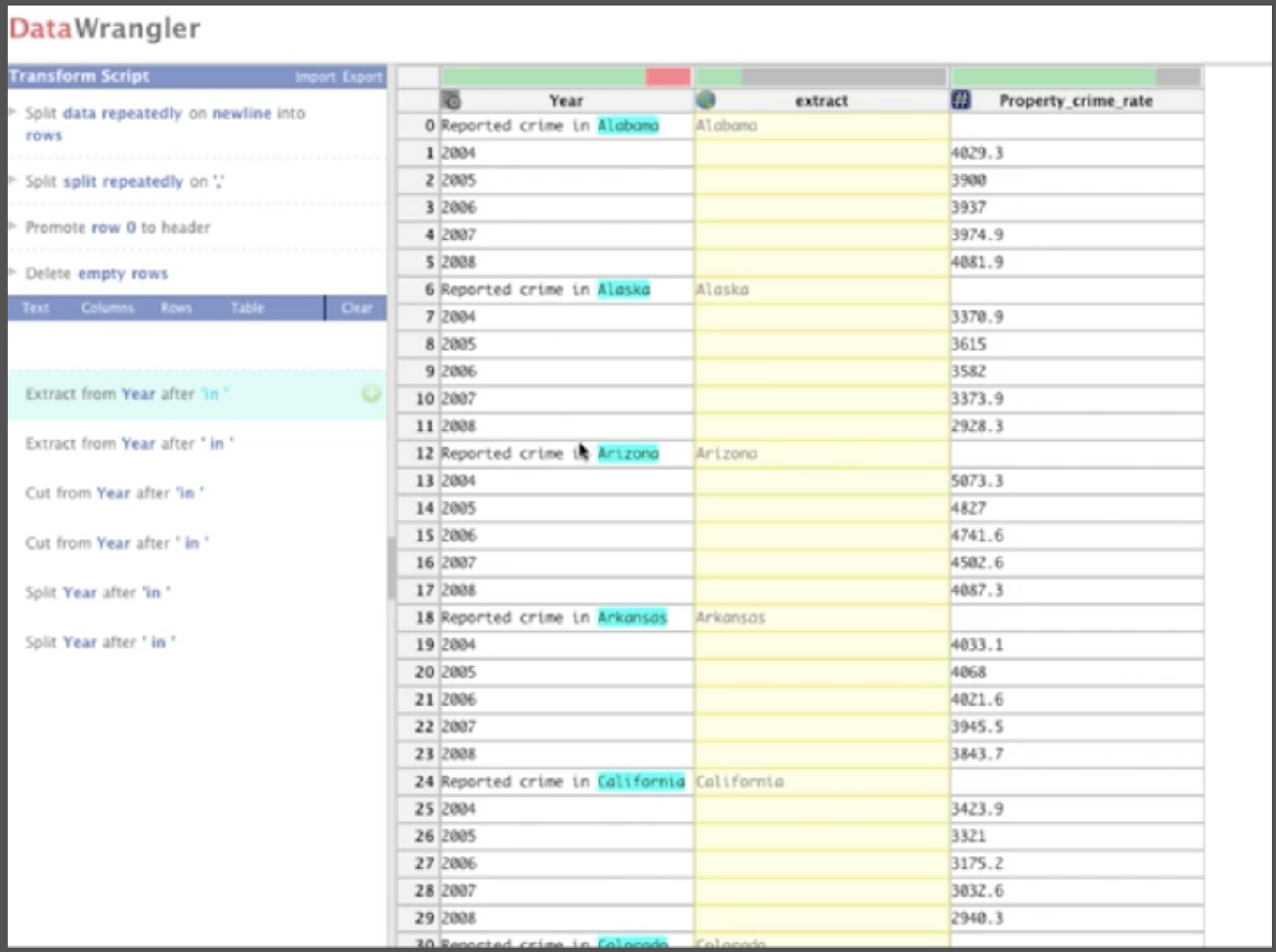

Graphic: Jeff Heer

# Open Refine




http://openrefine.org

# Data Wrangler (Stanford)




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?

…

http://eci.nic.in/eci_main1/ElectionStatistics.aspx

# Surf

Entity resolution tool to map similar values in databases

Approximate “Indian name” matching

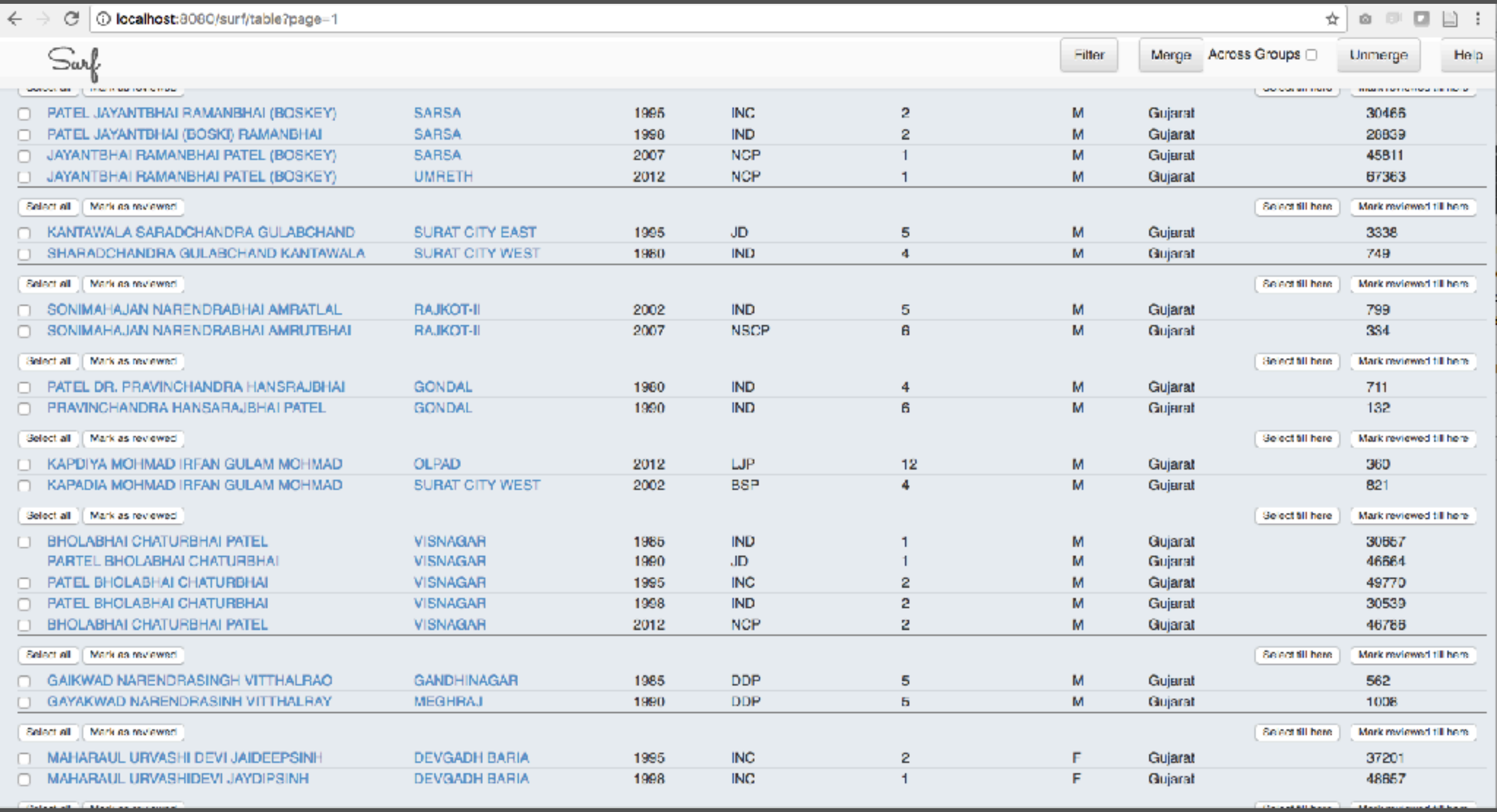
e.g. V -> W, EE -> I, TH->T, KSH -> X, SH->S

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


localhost:8000/surf/table?page=1
Surf
Filter
Merge
Across Groups
Unmerge
Help
PATEL JAYANTBHAI RAMANBHAI (BOSKEY) SARSA 1995 INC 2 M Gujarat 30466
PATEL JAYANTBHAI (BOSKI) RAMANBHAI SARSA 1990 IND 2 M Gujarat 26839
JAYANTBHAI RAMANBHAI PATEL (BOSKEY) SARSA 2007 NCP 1 M Gujarat 45811
JAYANTBHAI RAMANBHAI PATEL (BOSKEY) UMRETH 2012 NCP 1 M Gujarat 57363
KANTAWALA SARADCHANDRA GULABCHAND SURAT CITY EAST 1995 JD 5 M Gujarat 3338
SHARADCHANDRA GULABCHAND KANTAWALA SURAT CITY WEST 1980 IND 4 M Gujarat 749
SONIMAHAJAN NARENDRABHAI AMRATLAL RAJKOT-II 2002 IND 5 M Gujarat 799
SONIMAHAJAN NARENDRABHAI AMRUTBHAI RAJKOT-II 2007 NSCP 6 M Gujarat 334
PATEL DR. PRAVINCHANDRA HANSRAJBHAI GONDAL 1980 IND 4 M Gujarat 711
PRAVINCHANDRA HANSARAJBHAI PATEL GONDAL 1990 IND 6 M Gujarat 132
KAPDIYA MOHMAD IRFAN GULAM MOHMAD OLPAD 2012 LJP 12 M Gujarat 360
KAPADIA MOHMAD IRFAN GULAM MOHMAD SURAT CITY WEST 2002 BSP 4 M Gujarat 821
BHOLABHAI CHATURBHAI PATEL VISNAGAR 1985 IND 1 M Gujarat 30557
PARTEL BHOLABHAI CHATURBHAI VISNAGAR 1990 JD 1 M Gujarat 46564
PATEL BHOLABHAI CHATURBHAI VISNAGAR 1995 INC 2 M Gujarat 49770
PATEL BHOLABHAI CHATURBHAI VISNAGAR 1998 IND 2 M Gujarat 30539
BHOLABHAI CHATURBHAI PATEL VISNAGAR 2012 NCP 2 M Gujarat 46765
GAIKWAD NARENDRASINGH VITTHALRAO GANDHINAGAR 1985 DDP 5 M Gujarat 562
GAYAKWAD NARENDRASINH VITTHALRAY MEGHRAJ 1990 DDP 5 M Gujarat 1006
MAHARAUL URVASHI DEVI JAIDEEPSINH DEVGADH BARIA 1995 INC 2 F Gujarat 37201
MAHARAUL URVASHIDEVI JAYDIPSINH DEVGADH BARIA 1998 INC 1 F Gujarat 48857