Slido Poll:

www.sli.do

(code: CBSA1)

Demo URL:

https://rCanada.shinyapps.io/demo

Canada Border Services Agency
Agence des services frontaliers du Canada

Data engineering challenges and solutions: demonstration of Shiny

Dmitry Gorodnichy
Chief Data Office
Data Science Division

PROTECTION • SERVICE • INTEGRITY

GC Data Conference 2021, Data Literacy Fest
22-23 February 2021, Ottawa

Workshop layout

- “Round table” poll (Slido)

- Common data quality problems
- Data engineering as a solution
- Common tools for common problems
- Discussion (Slido)

Common data quality problems

- Facing same challenges turning “records” into “data”

• Dates :	‘20210820’ vs. ‘dob 20 Aug 2021’
• Names:	‘Dmitry Gorodnichy’ vs. ‘Dimitri Horodnytchyyi’
• Business Names:	AC, AirCanada, Air Canada Corp.
• Geographic Names:	‘Ottawa Airport’, ‘YOW’, Ottawa International Airport

Poor data quality impedes interoperability

- Good interoperability allows various data to be linked and enriched
- Probabilistic (approximate, fuzzy) matching is used to link “noisy” data
  - All words need to be compared to each other
  - Various techniques in data linkages include: using edit metrics, look-up tables, q-grams, phonetic, heuristics, …
- However, probabilistic matching has its share of challenges as well
  - How to assign threshold?
  - How to measure quality?
  - Lost nuances?
    - E.g., Bell Canada vs. Shell Canada
- No perfect solution

Data engineering to address data quality

- Data engineers develop techniques to standardize and organize data to help address data integrity, e.g.,
  - ‘Ottawa Airport’ → ‘YOW’
  - ‘YOW’ → ‘YOW’
  - ‘Ottawa International Airport → ‘YOW’
- On average, 80% of efforts of data scientists goes to address data engineering issues

Common Tools for Common Problems

- In GoC, we are working on the same set of data engineering problems
    - Standardizing various data fields
    - Cleaning, linking and searching data

so we can carry out analysis

- Often, data scientists end up “reinventing the wheel”
- We need to build common data engineering tools for common GoC data engineering problems

Vision for Solution

- We need a set of ‘libraries’ that are built and maintained by GoC data science community that is
  - Open
  - Free
  - Available to any data scientist who needs them
  - GCcode helps us to do that
  - R already has many libraries, supported by global community

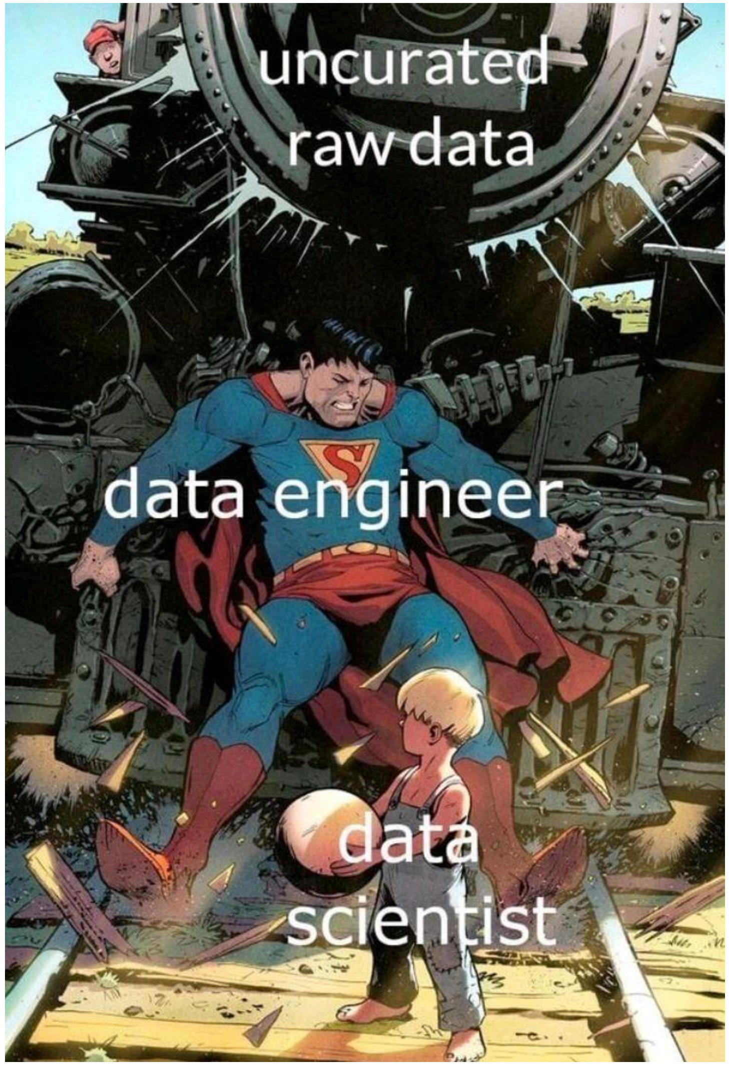
- This presentation will demo a proof of concept for feedback and discussion

Discussion

- Our works has just started. Much more ahead.
  - We build on Public Data (esp. Open Canada Data) and Public knowledge (esp. R global community)
  - We build solution (for entire GC community) and also we build Community of Practice
  - Codes and resources: https://gccode.ssc-spc.gc.ca/r4gc/
- Planned milestones:
  - rCanada Package, Testbed App, Toolkit App: 2021-2022
  - Use cases (for on-going Agency needs): Spring - Winter 2021
- We need your help!
  - curating & organizing DE challenges and public domain solutions (codes/papers)
  - curating & organizing public domain Data-sets
  - testing & benchmarking

Thank you !

Dmitry.Gorodnichy@cbsa-asfc.gc.ca



The author gratefully acknowledges the use of the following public domain material in this project:

R packages: data.table, dtplyr, lubridate, magrittr; soundex, phonetic; antiword, filehash, textreg, textreuse, stringr, stringi, textclean, syuzhet, hunspell, textshape; stringdist, recline, fuzzyjoin, RecordLinkage, fastlink; NLP, quanteda, udpipe, spacyr, tidytext; qdap, tm, lexicon, Rnewsflow, textcat; shiny, rsconnect, knitr, rmarkdown, flexdashboard, DT; cancensus; googleway, tidygeocoder; rvest, httr, xml2, jsonlite; microbenchmark; ggplot2, wordcloud.

Papers: Sariyar M / Borg A (2010). The RecordLinkage Package: Detecting Errors in Data“ The R Journal. van der Loo (2014) The stringdist package for approximate string matching. The R Journal ; L. Boytsov (2011). Indexing methods for approximate dictionary searching: comparative analyses. ACM Journal of experimental algorithmics G. Navarro (2001). A guided tour to approximate string matching. ACM Computing Surveys; Enamorado et al. (2019). ``Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records.'' American Political Science Review

Data: Statistics Canada, Post Canada, Simple Maps

Other: image by Anna Nyulund (LinkedIn post), stimulating discussions with many CBSA and GC colleagues, in particular at Friday's Data Science meet-ups.

Demo time

https://rCanada.shinyapps.io/demo