Demonstrations of linguistic data matching, consolidation, and cleansing

Jeff Fried
VP Advanced Solutions
Jeff.Fried@fastsearch.com

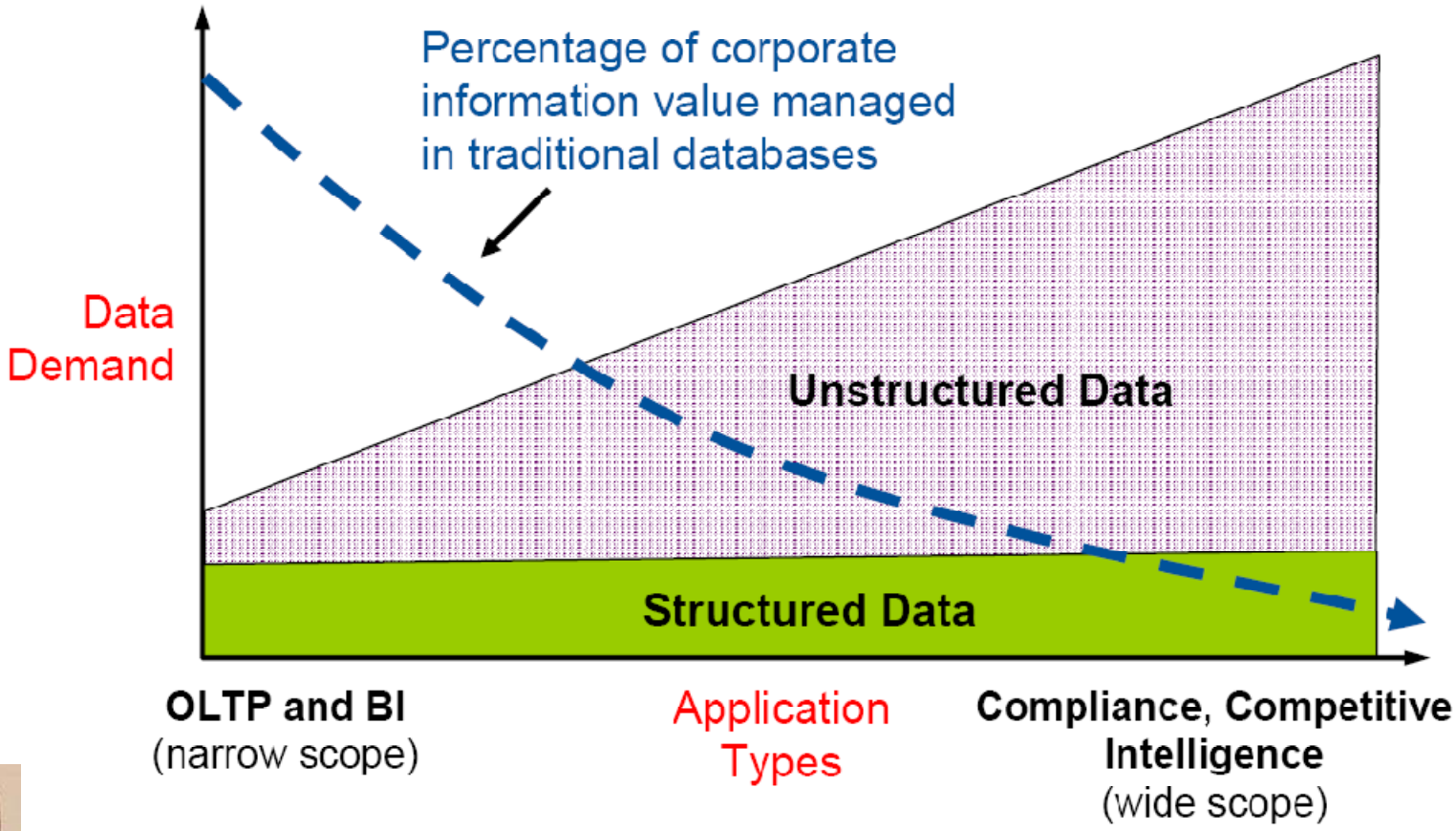
fast

Data Quality

- Still people entering the data
  - We do typos
  - We duplicate
  - We mess up

fast

Unstructured Information in the Warehouse



Mark Beyer

Gartner

Data Consolidation and ETL Studio

Direct access to RDBMs for info from some Telco's

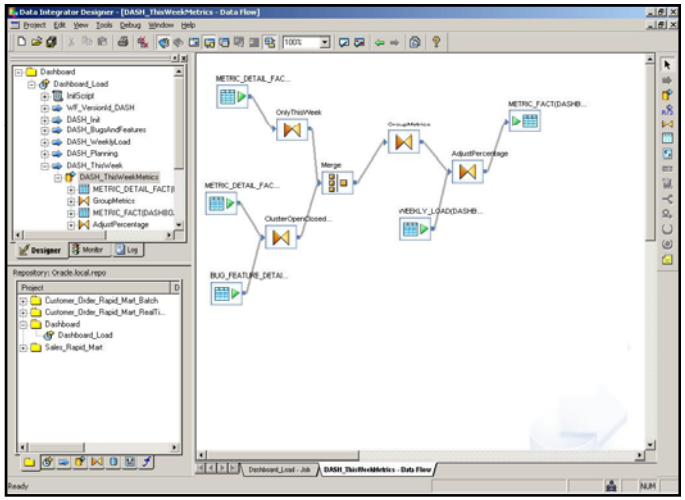
XML feed from other Telco's

XML

Flat files (CSV or fixed) from the 'laggards'

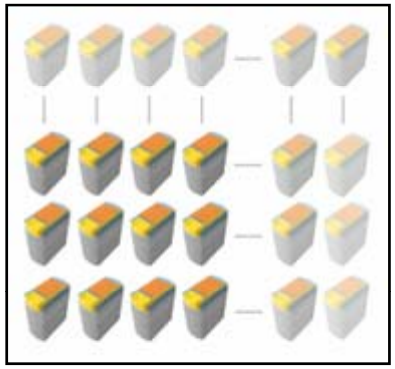
Logic for ETL

Logic for matching and cleansing



Lookup to ESP for Matching

Ordered hits (by quality)



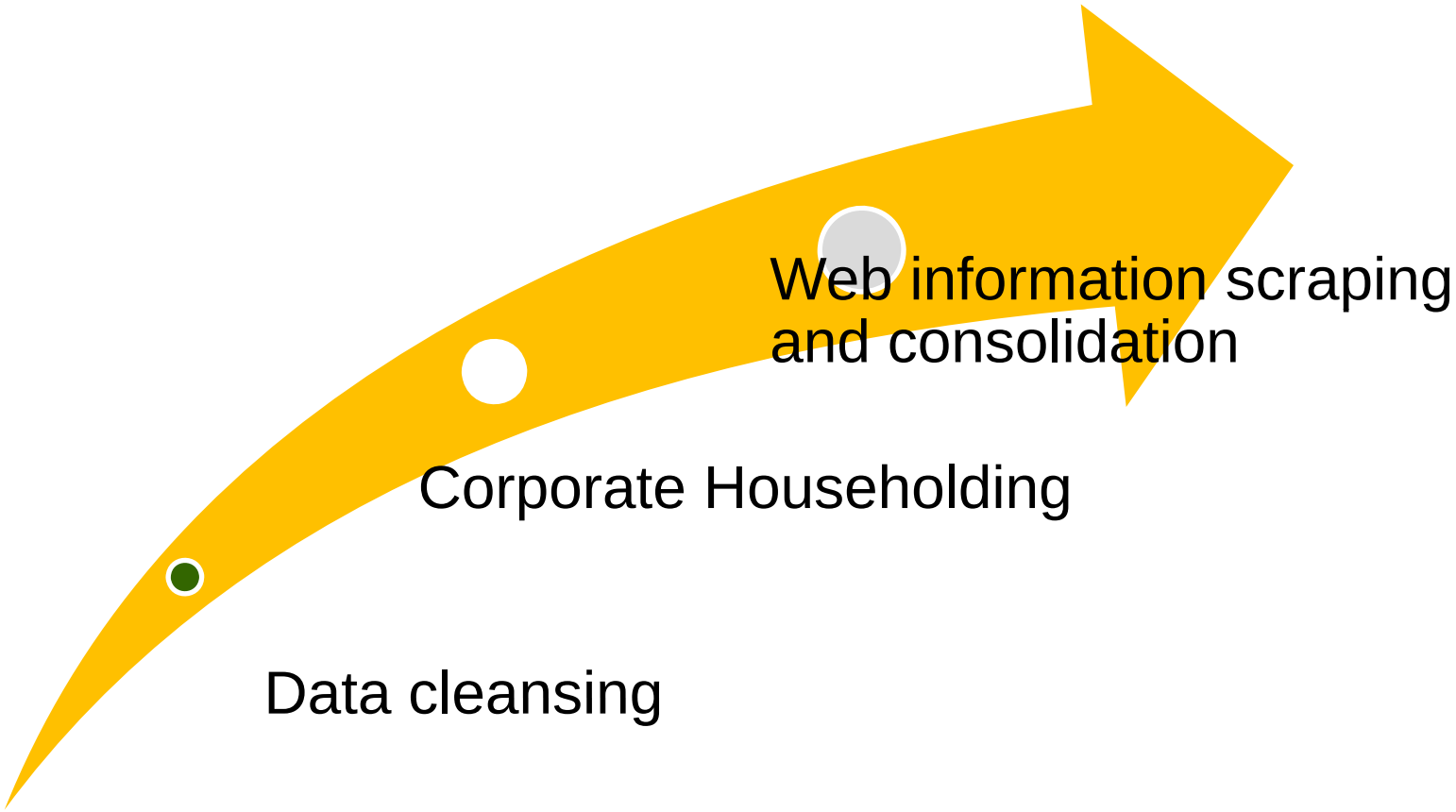
Cleansed data Back to ESP for next round Cleansing or for 'search'

Ambigous data (close hits or unidentified)

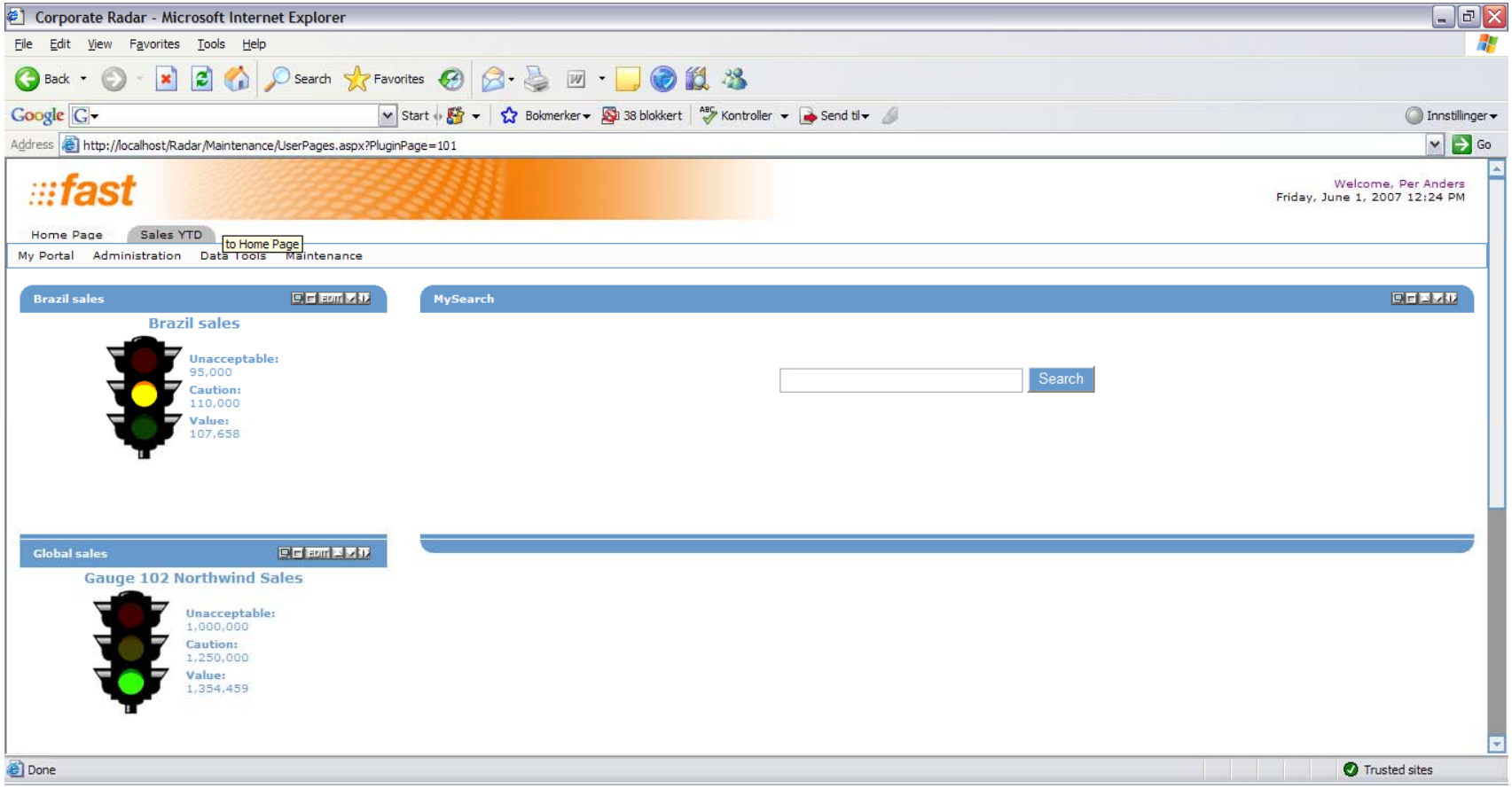
'Error' database for manual inspection, correction, storage/learning

clean data

Master database for persistant storage

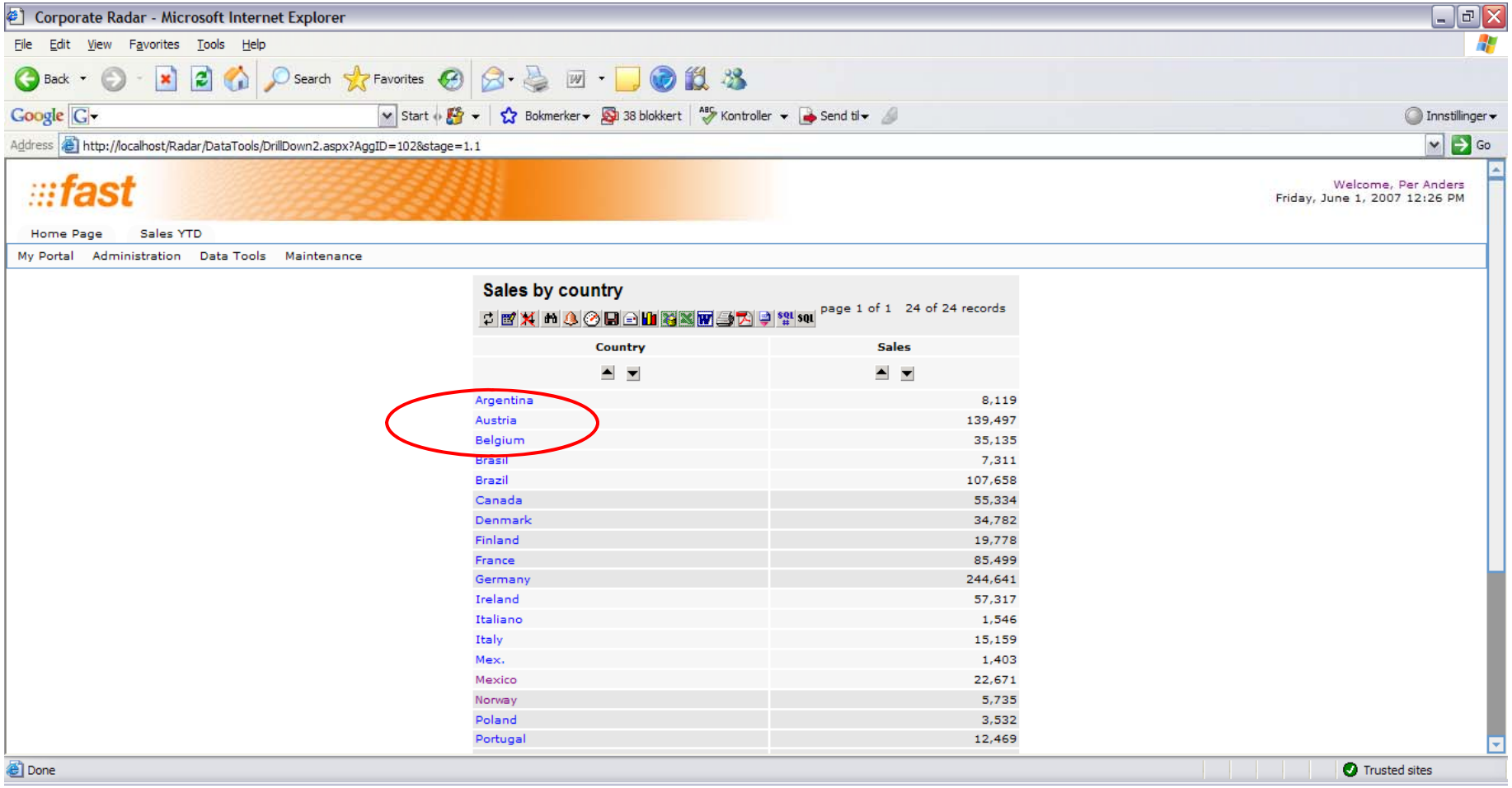

Web information scraping and consolidation
Corporate Householding
Data cleansing

Viewing data, we notice issues



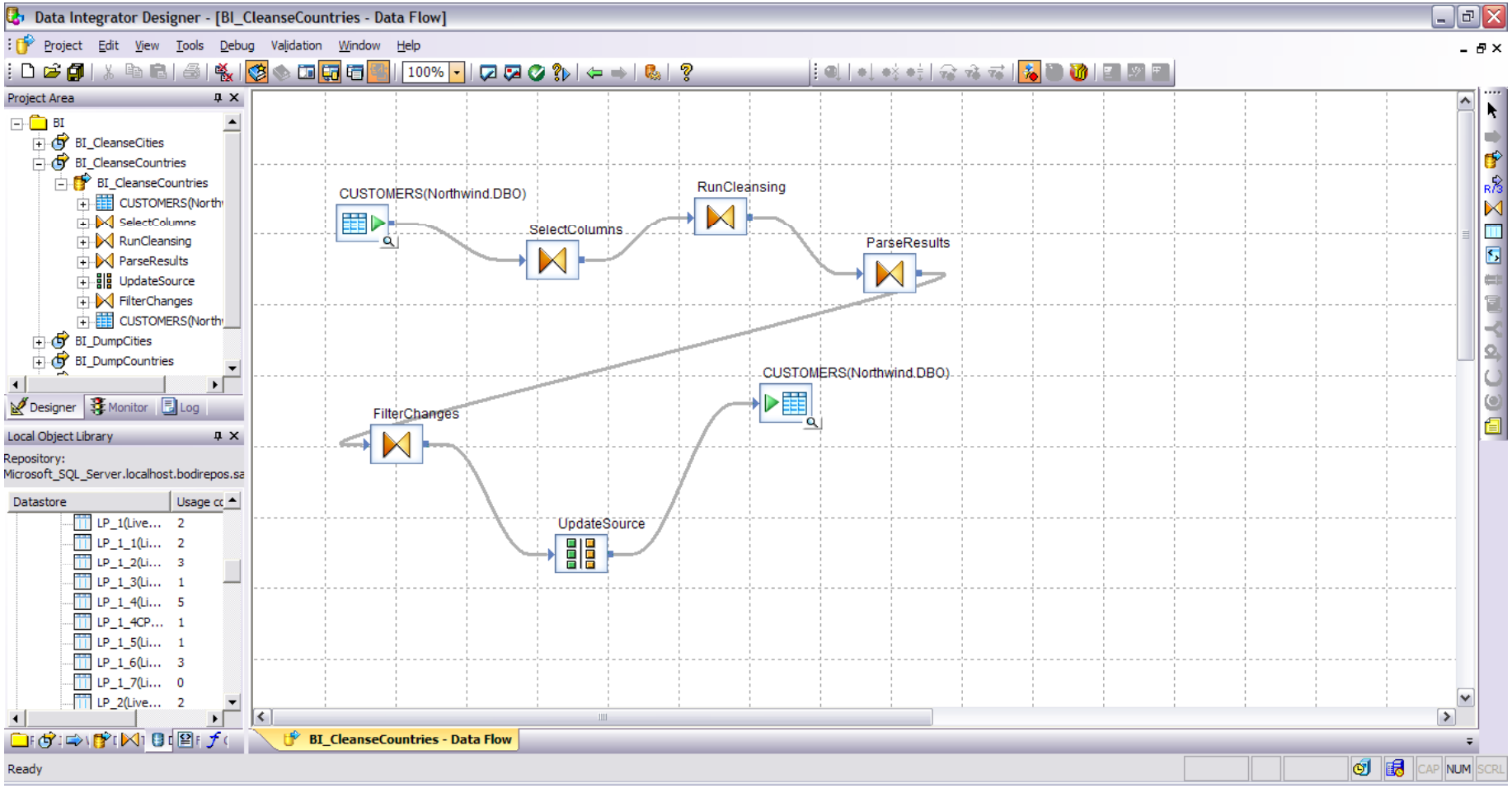
fast

We click on the gauge

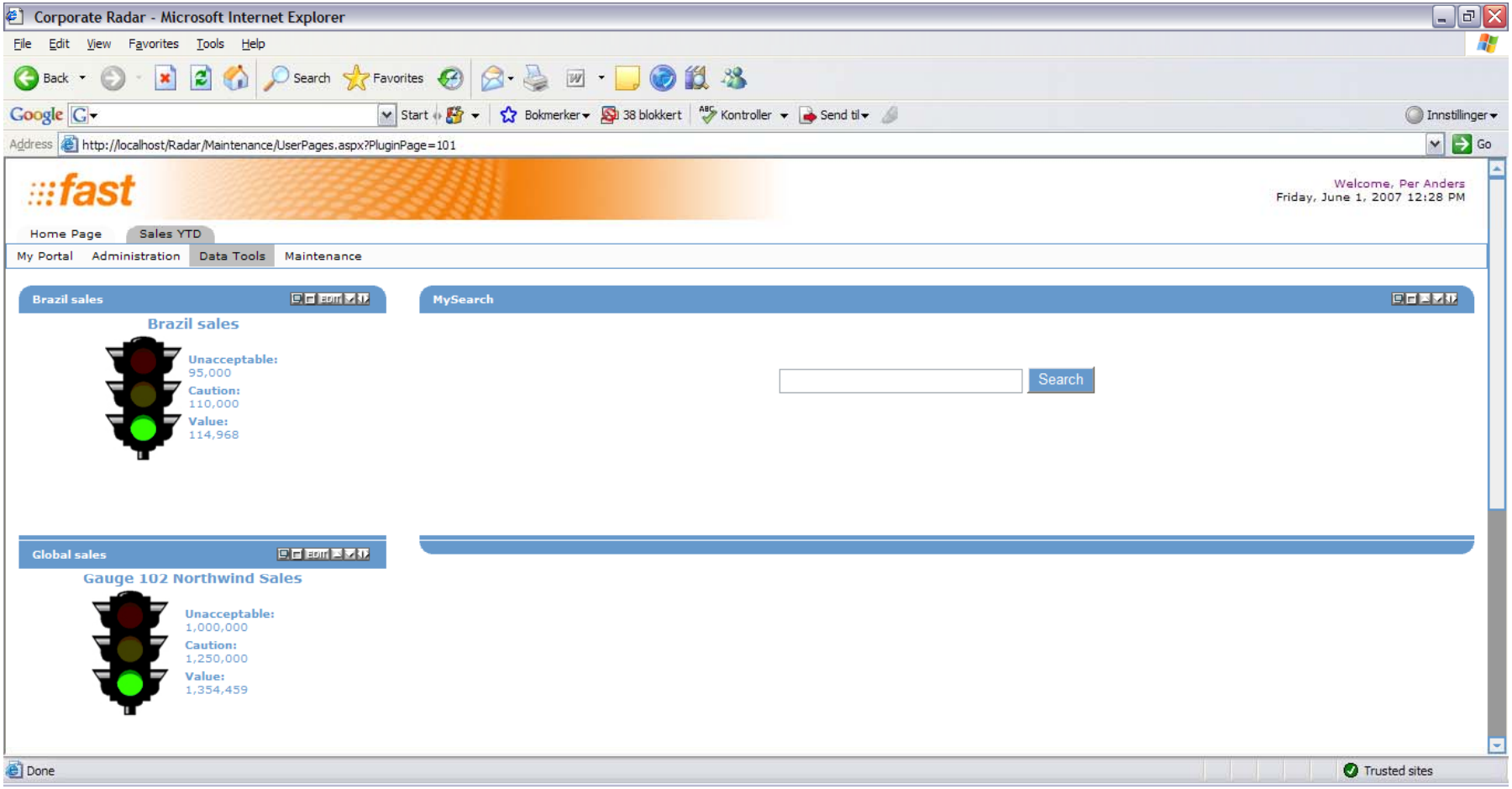


fast

We cleanse the data

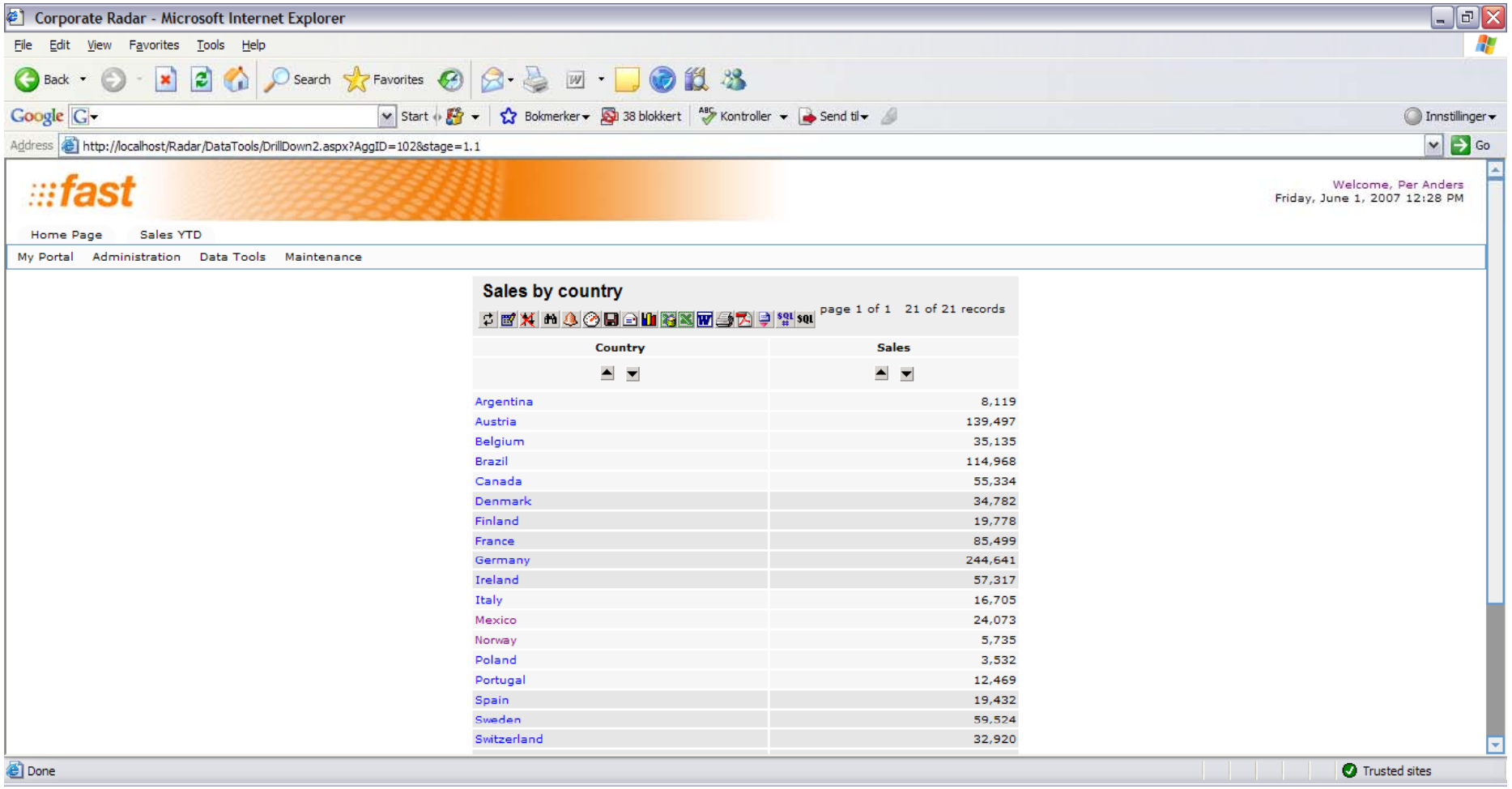


Now look at the result



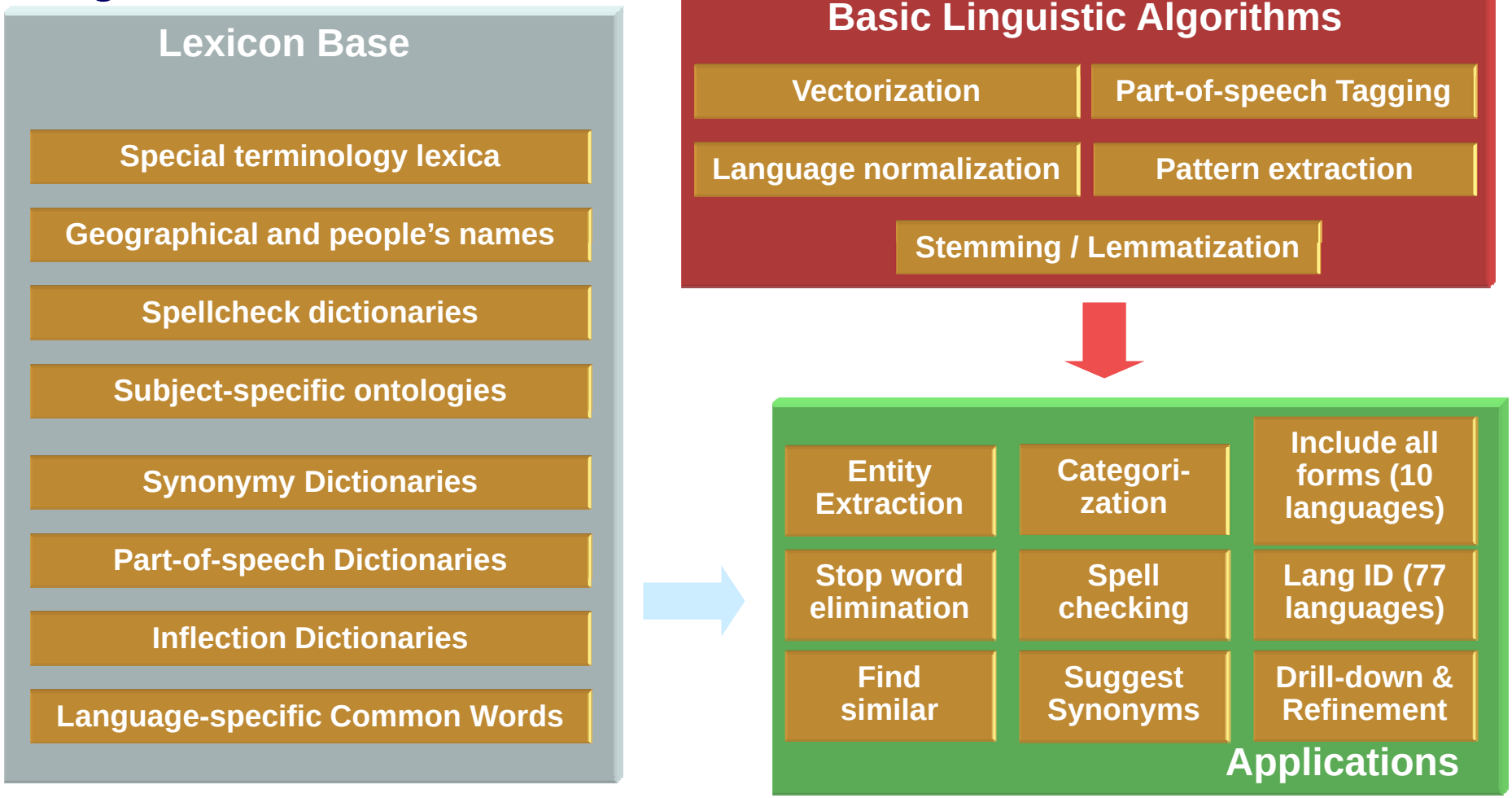
fast

Details



fast

Linguistic Fundamentals



Q&A

Thank you!

find the real value of search

Jeff Fried
VP Advanced Solutions
Jeff.Fried@fastsearch.com