AI & THE BIG DATA MOVEMENT

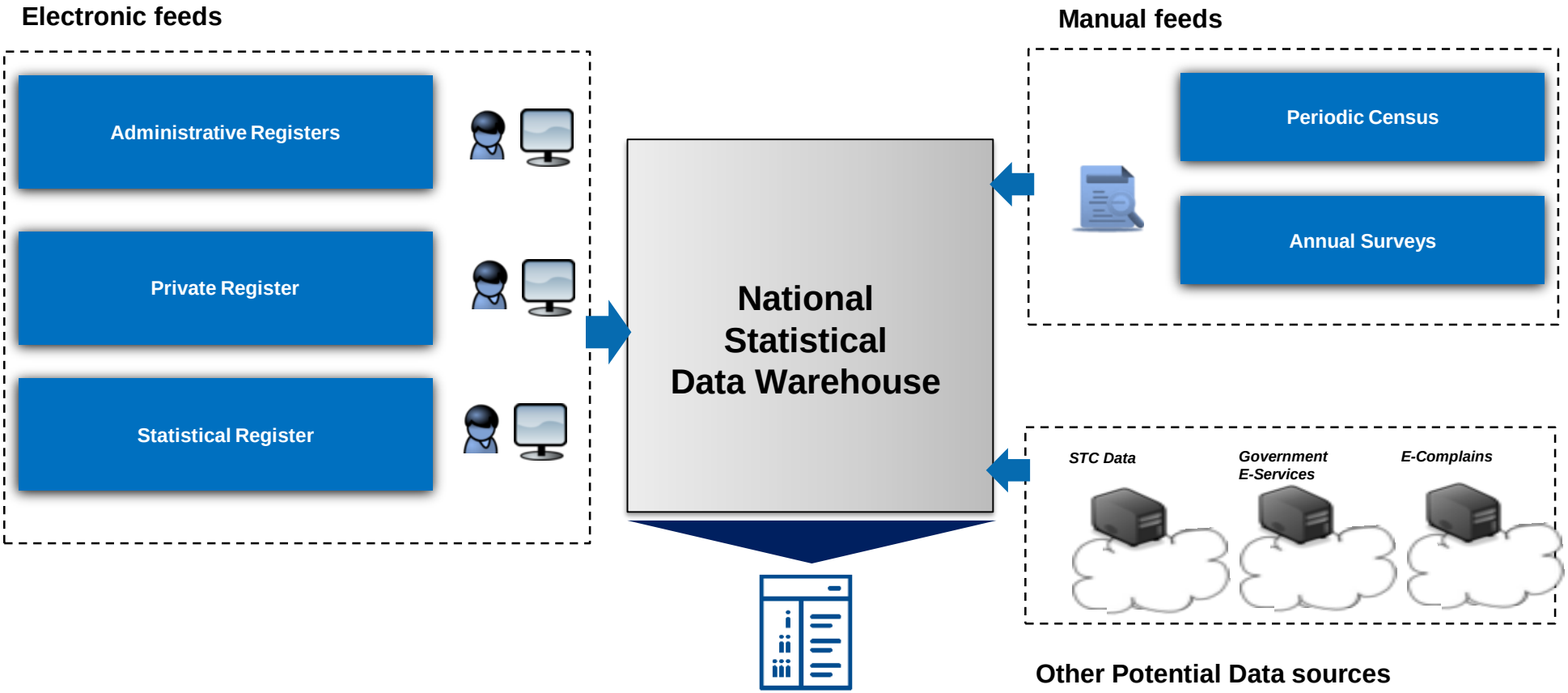
EXECUTIVE AI & BIG DATA WORKSHOP

Session 2

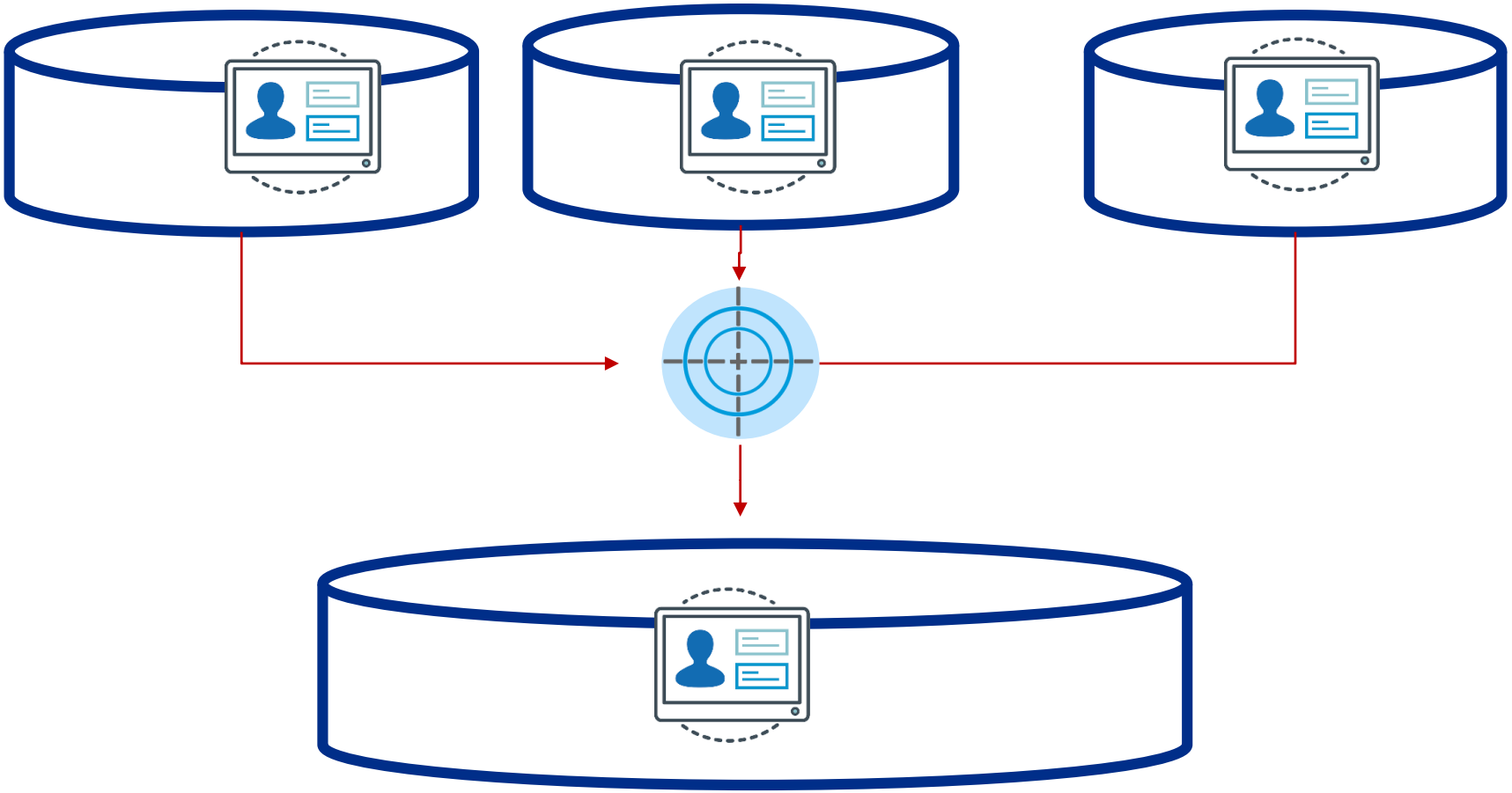
4/22/2019

COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS

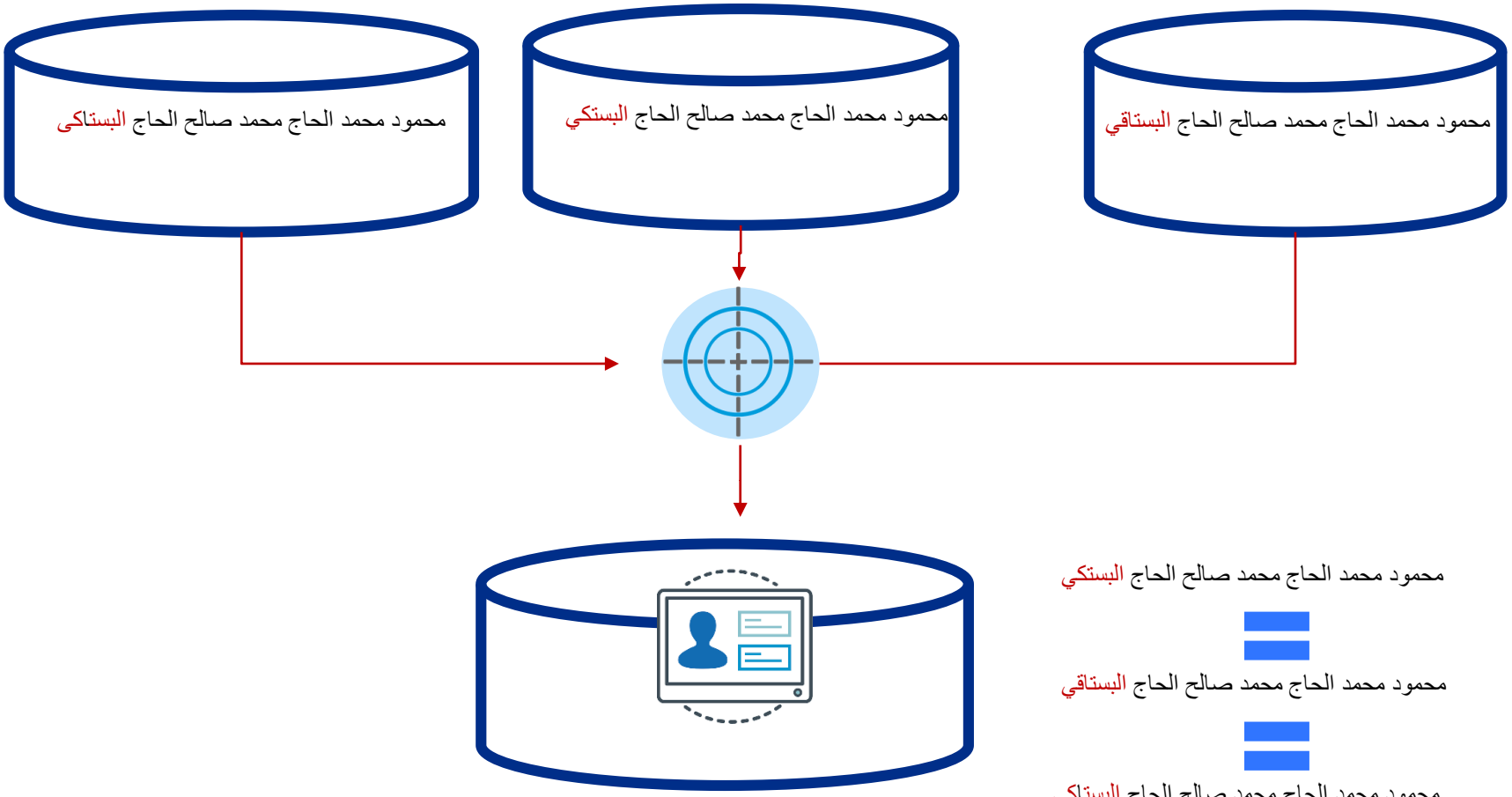
Consolidating Records from multiple Data Sources



Consolidating Records from multiple Data Databases



Consolidating Records from multiple Data Sources

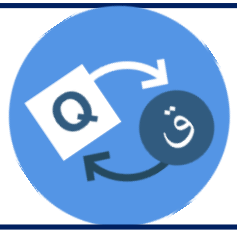


Identity Management through Name Matching (Arabic/English)

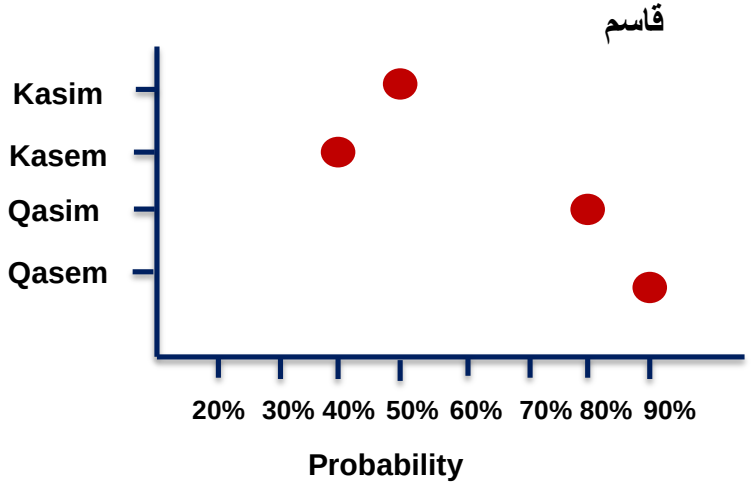
Suppose Other Phenomena Are Entered

- Nicknames
  - Johnny Smith
- Gender mistakes
  - Joan Smith
  - Joanie Smith
- Differences in relative name frequencies
  - Shaun Smith
- Initials, especially for organization names
  - IBM vs. International Business Machines

Qasem Mohamed Hisham Rawashdeh



قاسم محمد هشام رواشدة



COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS

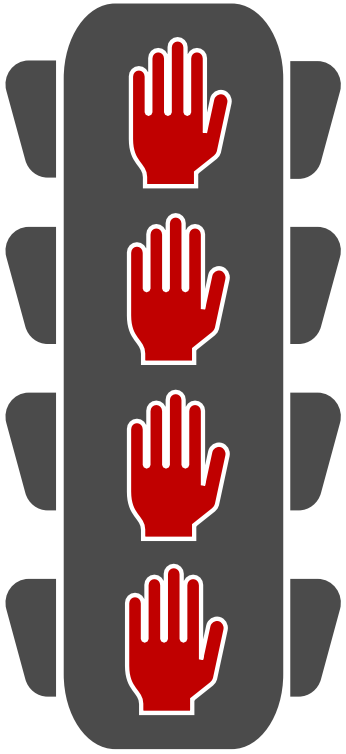
Open Source Names Data Collection

Months Dataset	UAE Brokers Dataset	Pre-Collected Research Dataset	Allah Names Dataset
Manually generated	UAE Open Data	Cognitro Research	Manually generated
https://git.io/fh80t	http://data.bayanat.ae/en_GB/dataset/brokers	https://git.io/fh80J	https://git.io/fh80s
Single Names	Multiple Names	Single Names	Single Names
12 Data Size	10K Data Size	20K Data Size	101 Data Size




COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS

Data Cleansing and Preparation

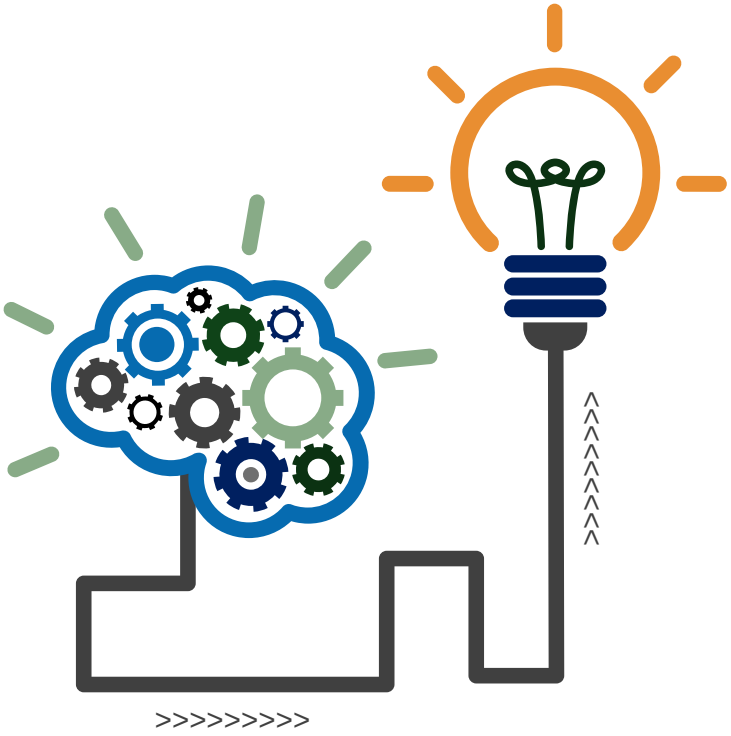


Miss-Match Records
Eng: [SAIF MOHAMMED SALEM DEAIFES]
Ar [سيف محمد سالم دعيفس الشامسى]

Names with Abbreviations
Eng: [SALEH SALEM S G AL MARRI]
Ar: [صالح سالم صالح غراب المرى]

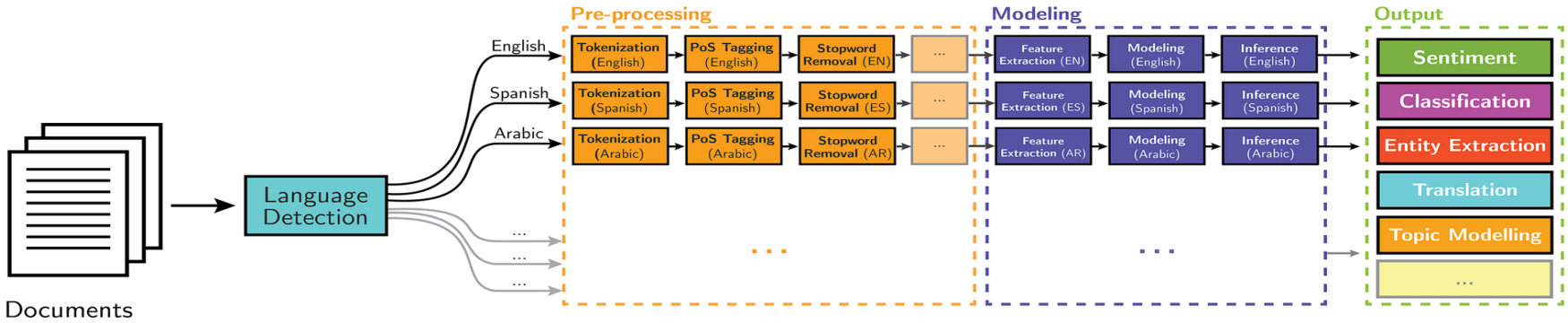
Spelling Mistakes
Eng: [ALMASA TOURTSIOM L LC]
Ar: [الماسه للسياحه ذ م م]

Translation and Transliteration Mixed
Eng: [AL SAHRAA SAFEER TRYE REPAIR AND LUBRICATION L L C]
Ar: [سفير الصحراء لتصليح الاطارات وتبديل الزيوت ذ م م]

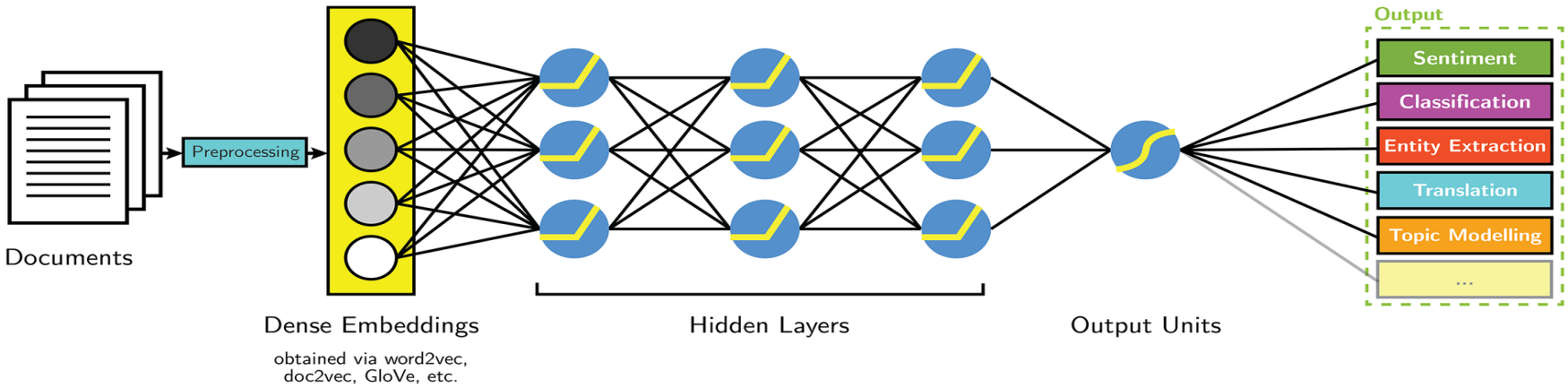


Deep Learning Vs. Classic NLP

Classical NLP



Deep Learning-based NLP



Name Matching Testing and Validation

Model accuracy and performance will be validated using Levenshtein Distance string metric as one of Word Error Rate (WER) techniques

Definition

The Levenshtein distance is a string metric for measuring difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other

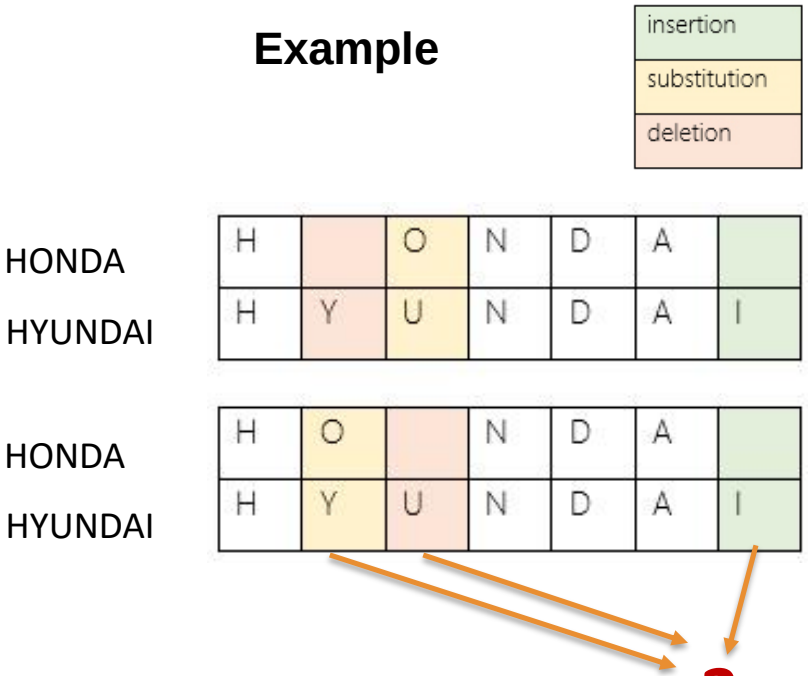
Accuracy =

1 – ((Subs+Del+Ins)/T)

Subs: substitutions
Del: deletions
Ins: insertions

T: Text Length

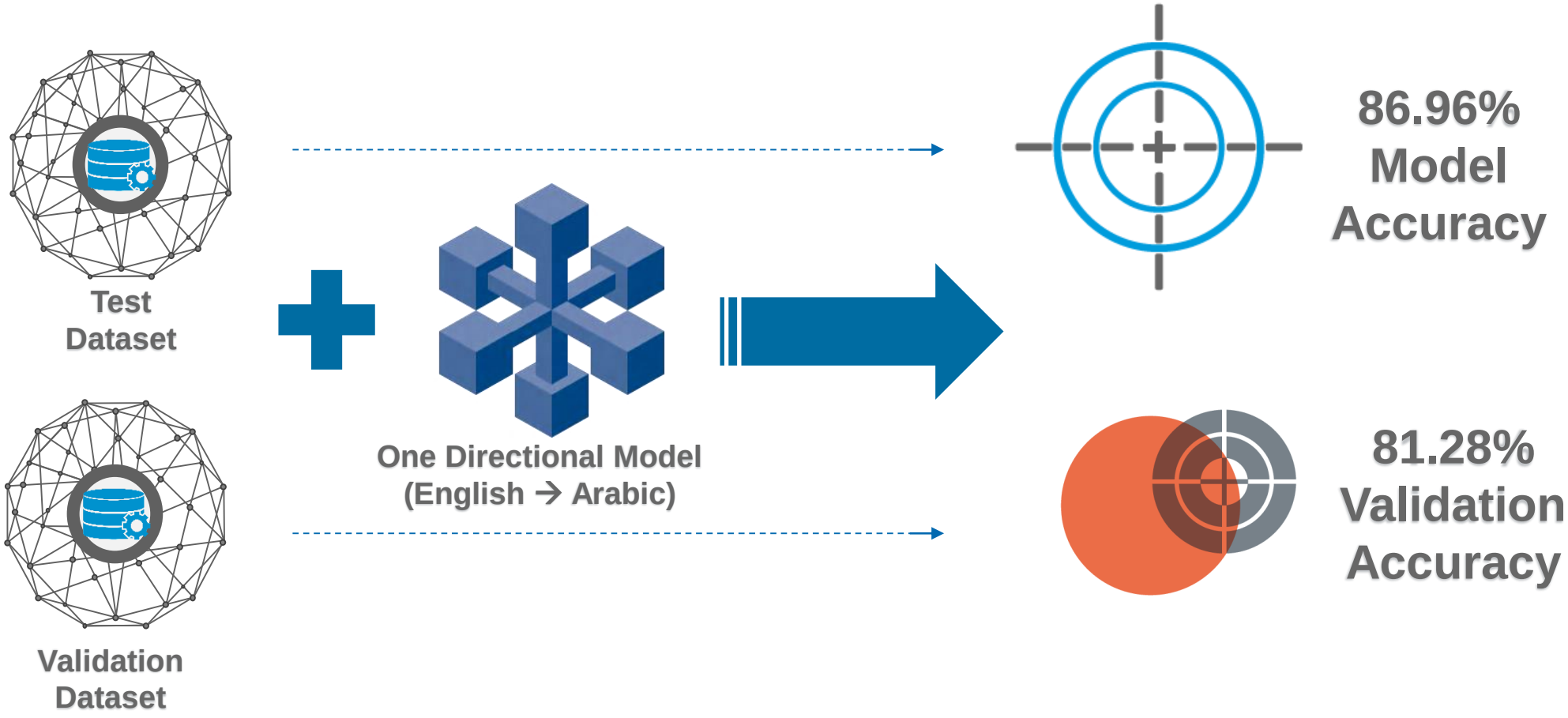
Example



Levenshtein distance between "**HONDA**" and "**HYUNDAI**" is **3**

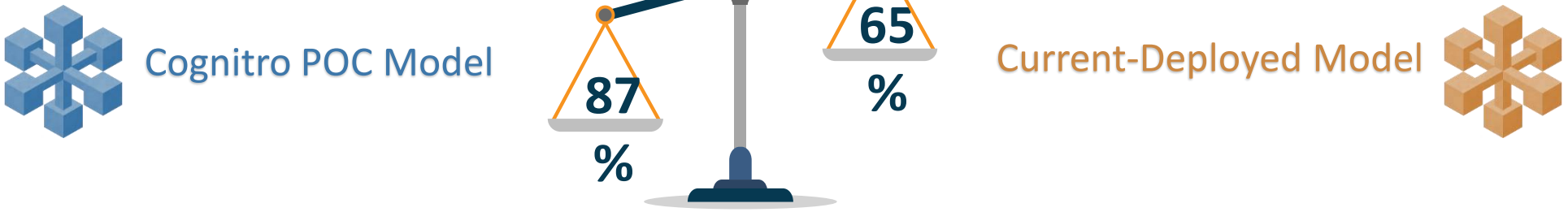
COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS

Identity Management through Name Matching (Arabic/English)



Identity Management through Name Matching (Arabic/English)

We have compared our POC Model with the current-deployed Model using 300 samples out of any training sets



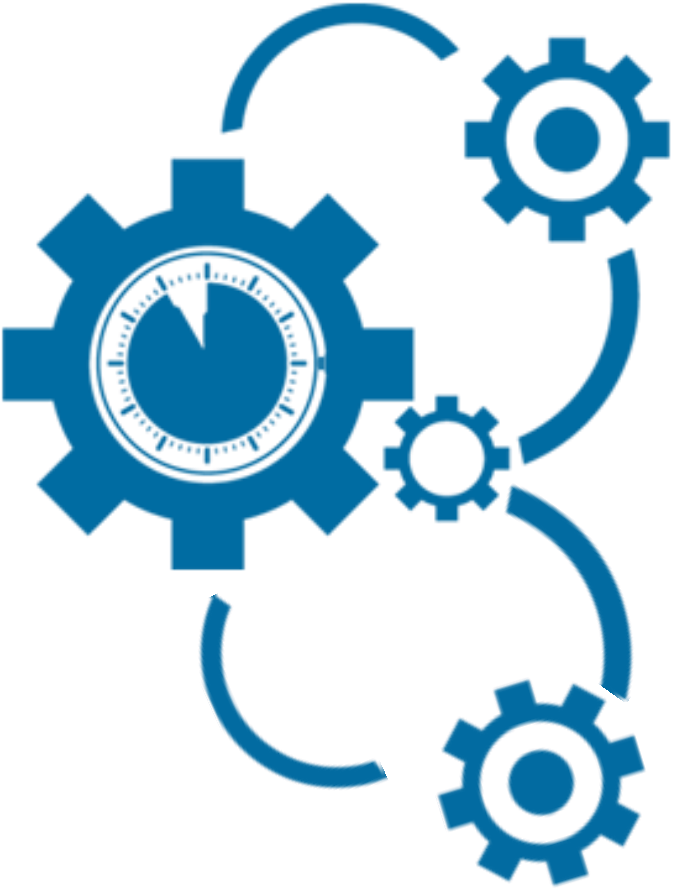
Original Text Examples

محمود محمد الحاج محمد صالح الحاج البستكى	محمود محمد الحاج محمد صالح الحاج البستكي MAHMOOD MOHD ALHAJ MOHD SALEH ALHAJ ALBASTAKI	محمود محمد الحاج محمد صالح الحاج البستاقي
عبدالناصر ابراهيم عبدالله الخياط	عبدالناصر ابراهيم عبدالله الخياط ABDULNASSER EBRAHIM ABDULLA ALKHAYYAT	عبد الناصر ابراهيم عبدالله الخيات
موزه عبيد سعيد غانم غباش المهيرى	موزه عبيد سعيد غانم غباش المهيرى MOAZA OBAID SAEED GHANEM GHUBASH ALMUHAIRI	معاذه عبيد سعيد غانم غباش المهيري

COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS

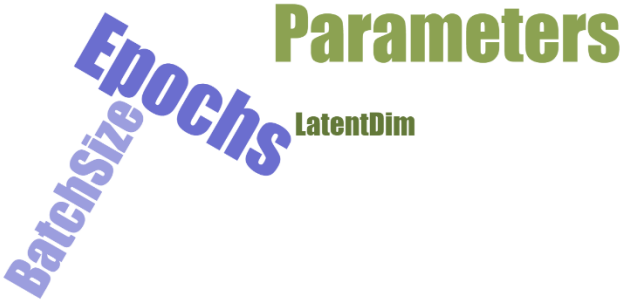
Identity Management through Name Matching (Arabic/English)

Then, we will validate and fine-tune the model by enhancing the parameters and performing further pre-processing and odd-cases procedures



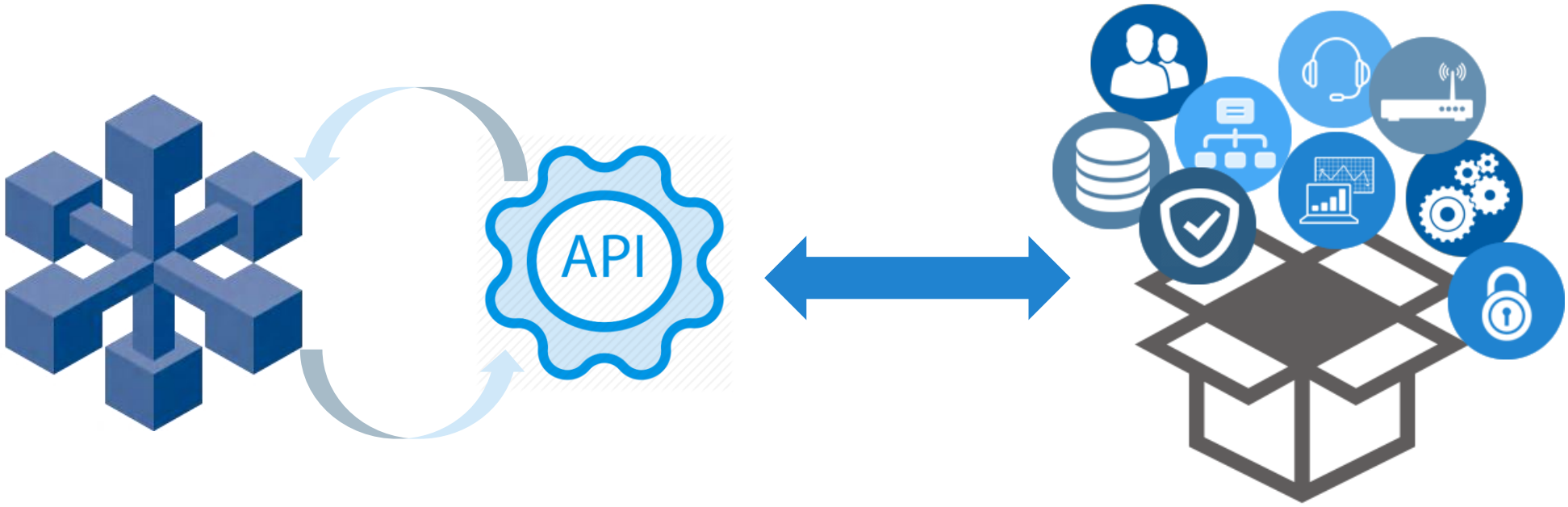
Data Preprocessing

Model Optimization



Identity Management through Name Matching (Arabic/English)

Our model will be deployed inside Tahaluf Data Center as an API that can be integrated with Tahaluf other services



Name Matching Model

Name Matching API

Enterprise- Services

COGNITRO ANALYTICS
DATA EMPOWERING DECISIONS