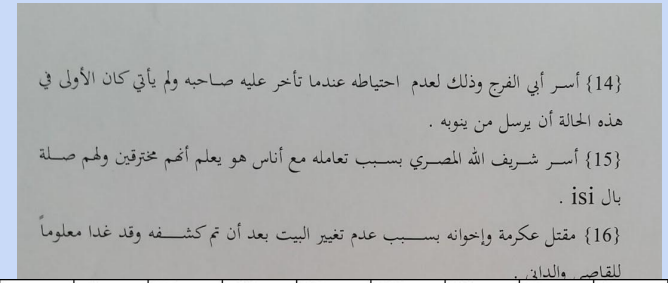
Upshot from Dayton: Found in Translation

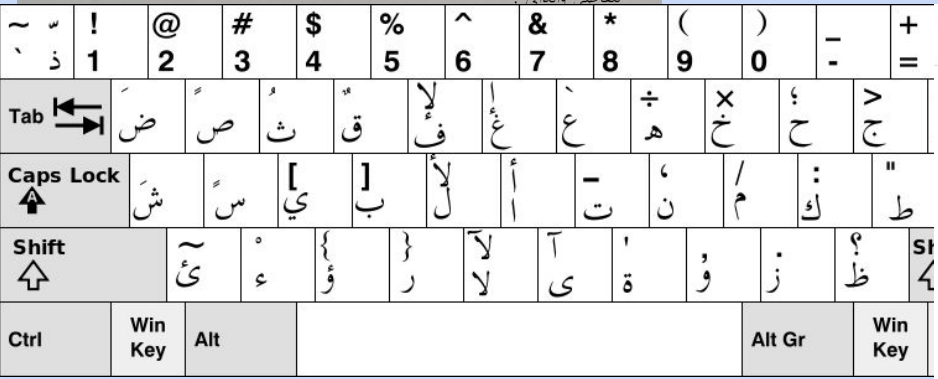
Ranking Arabic Documents with Machine Learning

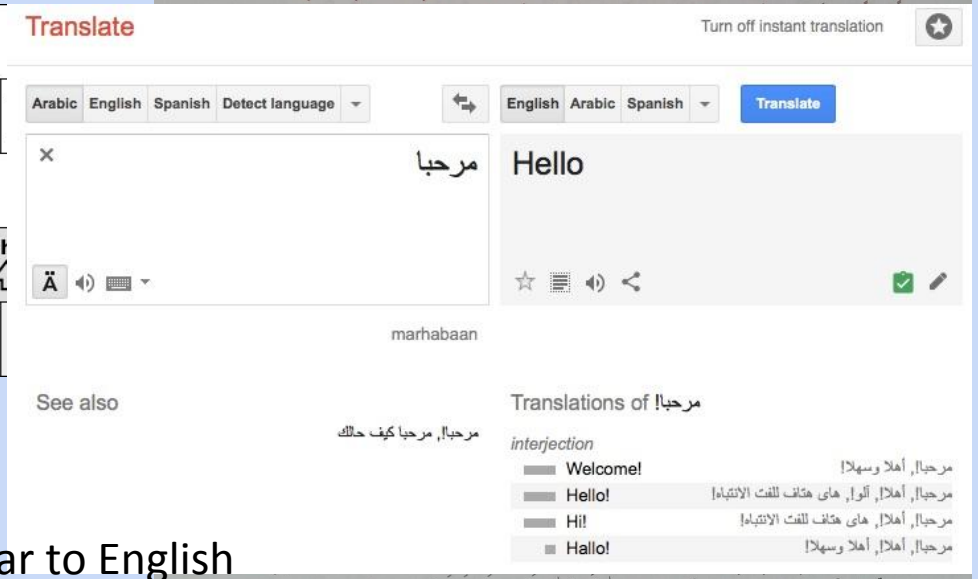
Team Members: Joann Luu, Jun Huang, Craig Boman, Carla Davis, Amy Magnus, Ron McChesney, Samuel Rivera


{14} أسر أبي الفرج وذلك لعدم احتياطه عندما تأخر عليه صاحبه ولم يأتي كان الأولى في هذه الحالة أن يرسل من ينوبه .

{15} أسر شريف الله المصري بسبب تعامله مع أناس هو يعلم أنهم مخترقين ولهم صلة بال isi .

{16} مقتل عكرمة وإخوانه بسبب عدم تغيير البيت بعد أن تم كشفه وقد غدا معلوماً للقاصي والداني .





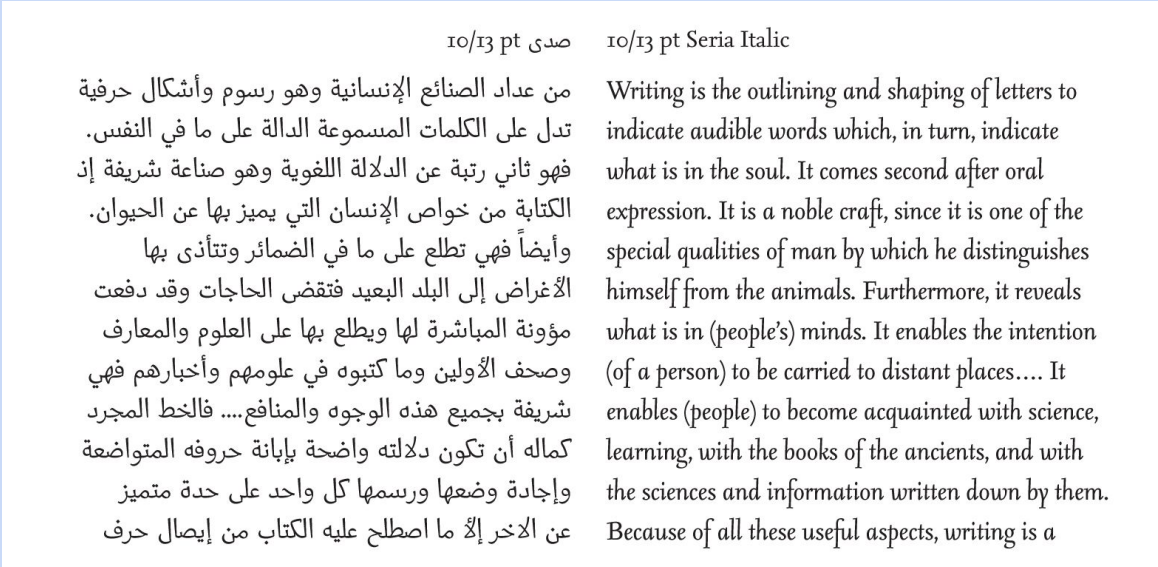
- 28 Letter Alphabet
- Logical structure (nouns and verbs) similar to English
- Written and read from right to left.
- Written in cursive, making text recognition challenging.

Concept of Operation

- Understand the value of the document
- Limited Interpreters - upshoot documents that have urgent relevance
- Mobile app translation is nice to have

Problem Statement

- Given a document photo, what are some other translated documents like?
- Should someone translate it?



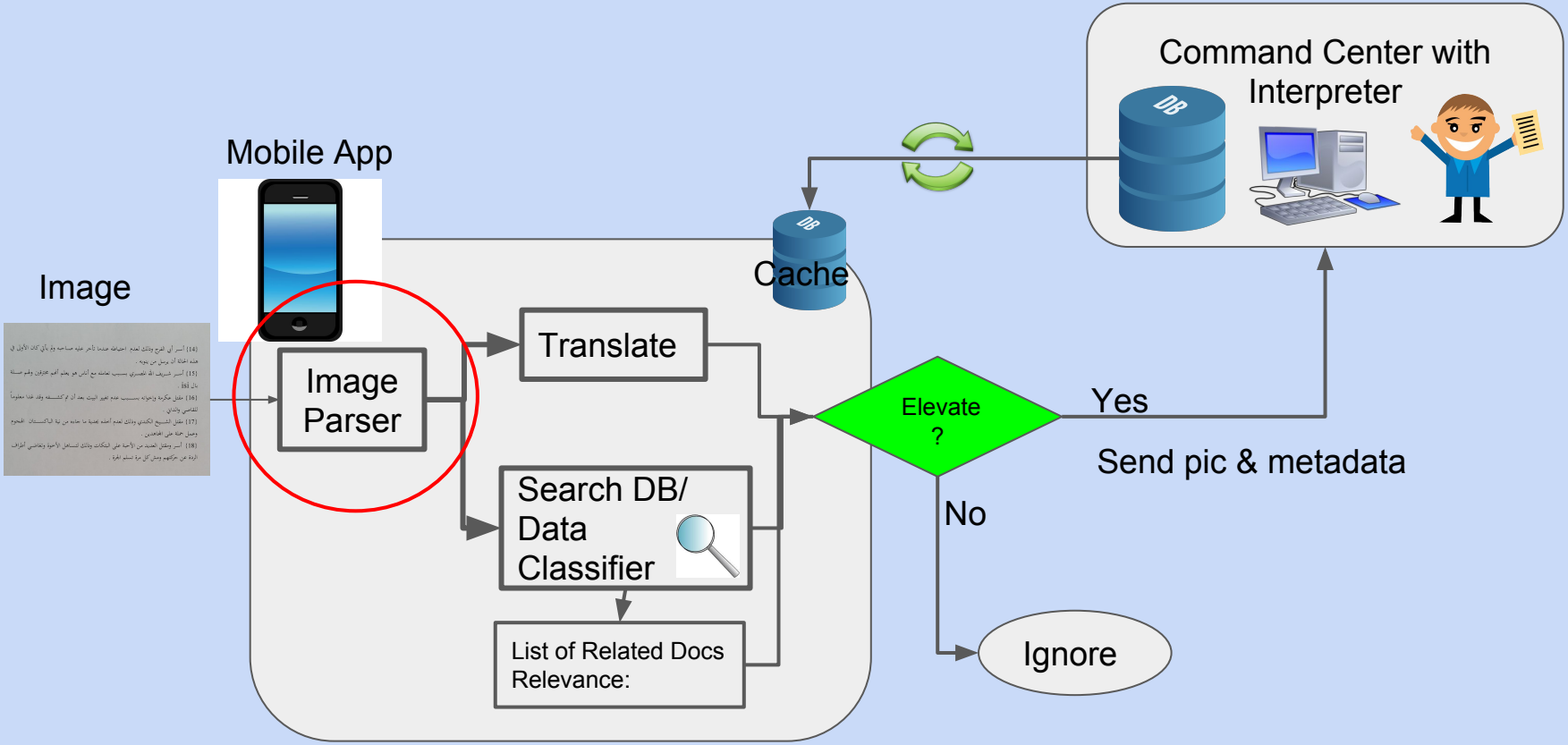
صدى 10/13 pt

من عداد الصنائع الإنسانية وهو رسوم وأشكال حرفية
تدل على الكلمات المسموعة الدالة على ما في النفس.
فهو ثاني رتبة عن الدلالة اللغوية وهو صناعة شريفة إذ
الكتابة من خواص الإنسان التي يميز بها عن الحيوان.
وأيضاً فهي تطلع على ما في الضمائر وتتأدى بها
الأغراض إلى البلد البعيد فتقضى الحاجات وقد دفعت
مؤونة المباشرة لها ويطلع بها على العلوم والمعارف
وصحف الأولين وما كتبوه في علومهم وأخبارهم فهي
شريفة بجميع هذه الوجوه والمنافع.... فالخط المجرد
كماله أن تكون دلالته واضحة بإبانة حروفه المتواضعة
وإجادة وضعها ورسمها كل واحد على حدة متميز
عن الاخر إلاّ ما اصطلح عليه الكتاب من إيصال حرف

10/13 pt Seria Italic

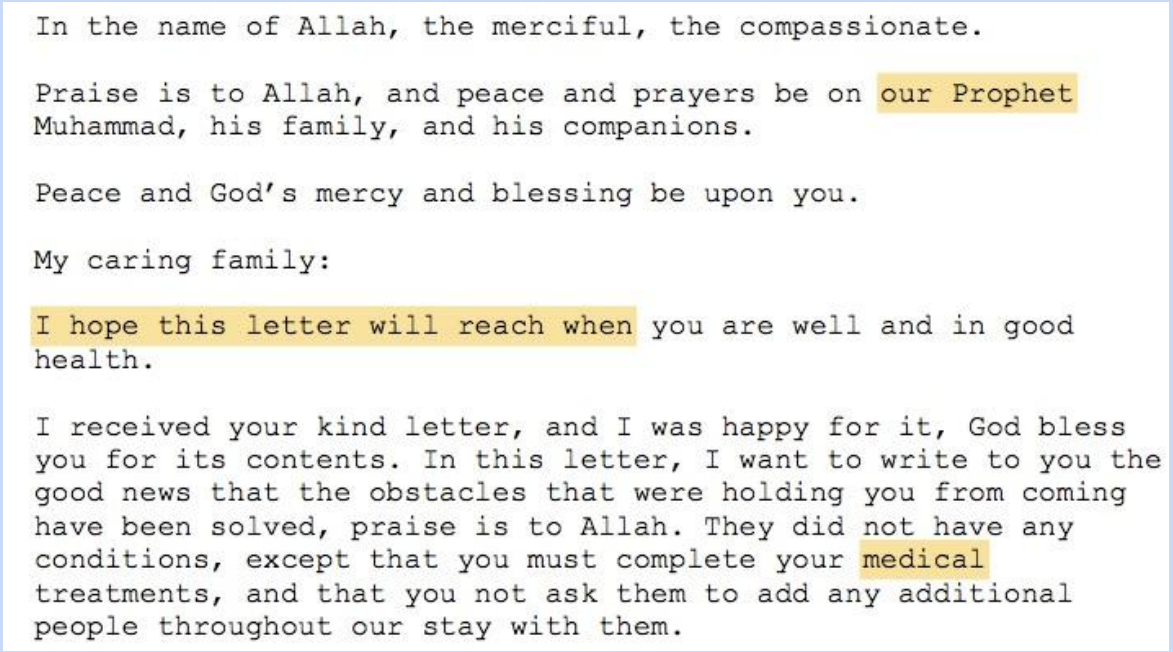
Writing is the outlining and shaping of letters to indicate audible words which, in turn, indicate what is in the soul. It comes second after oral expression. It is a noble craft, since it is one of the special qualities of man by which he distinguishes himself from the animals. Furthermore, it reveals what is in (people's) minds. It enables the intention (of a person) to be carried to distant places.... It enables (people) to become acquainted with science, learning, with the books of the ancients, and with the sciences and information written down by them. Because of all these useful aspects, writing is a

Architecture Design



Demonstrate Image Parser

- PDF Converter to text images
- Translate using language translation Watson's API
- Ground Truth:934 words
- Translated:714 words



In the name of Allah, the merciful, the compassionate.

Praise is to Allah, and peace and prayers be on our Prophet Muhammad, his family, and his companions.

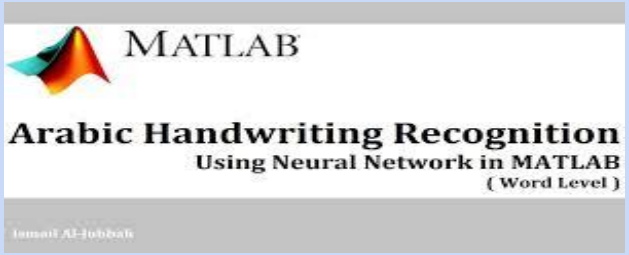
Peace and God's mercy and blessing be upon you.

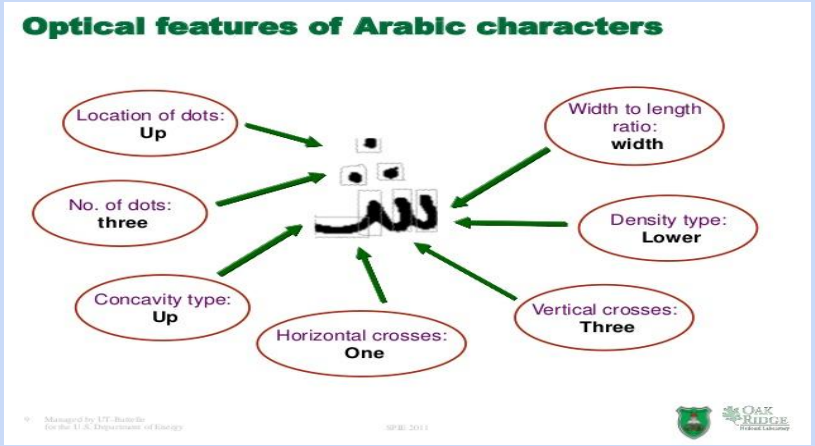
My caring family:

I hope this letter will reach when you are well and in good health.

I received your kind letter, and I was happy for it, God bless you for its contents. In this letter, I want to write to you the good news that the obstacles that were holding you from coming have been solved, praise is to Allah. They did not have any conditions, except that you must complete your medical treatments, and that you not ask them to add any additional people throughout our stay with them.

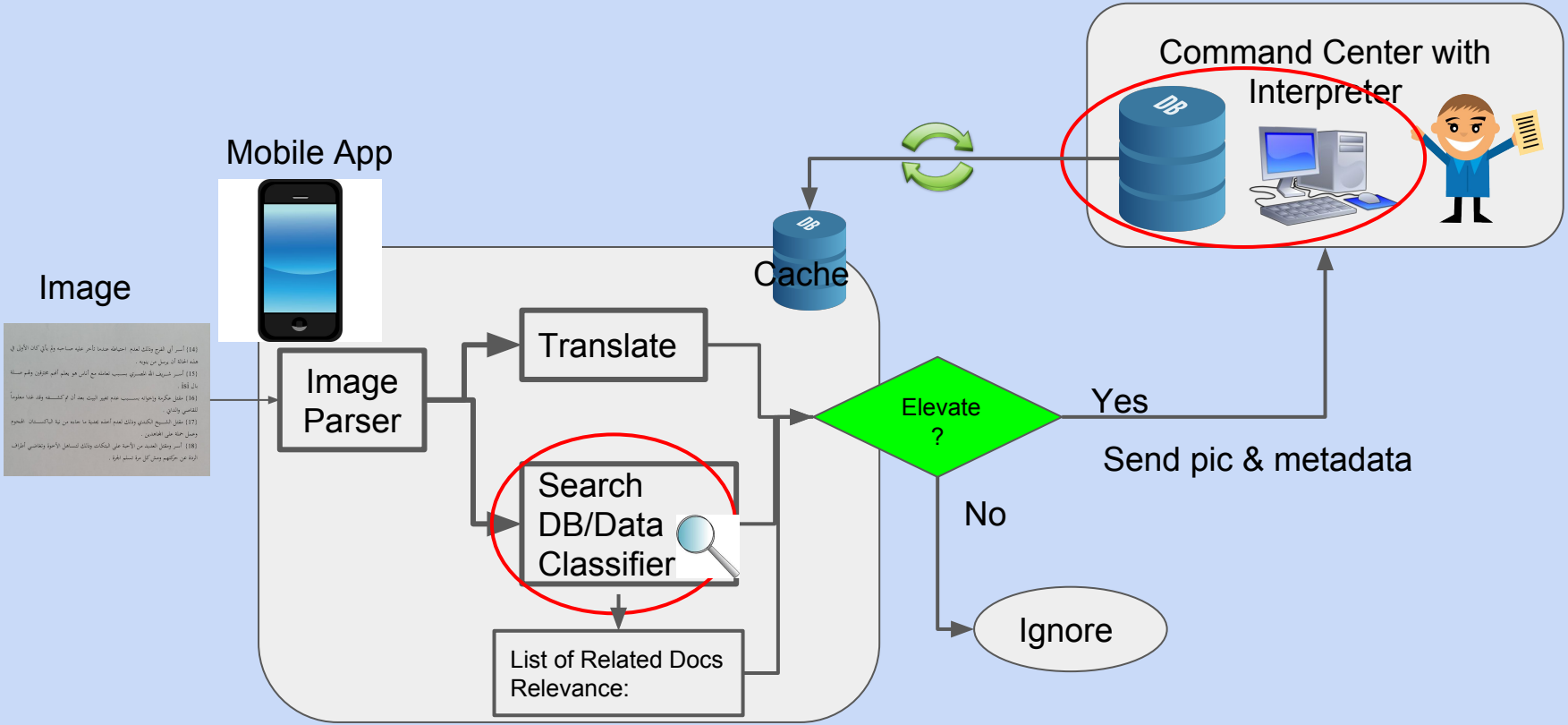
Reuse of Software, Services, Data...





- Google Translate API to create words for the Classifier

Architecture Design



Limitations and Assumption

Limitations:

- Time and memory constraints (mobile computer)
- Connectivity to central database
- Central intelligence about label of documents in other language

Assume:

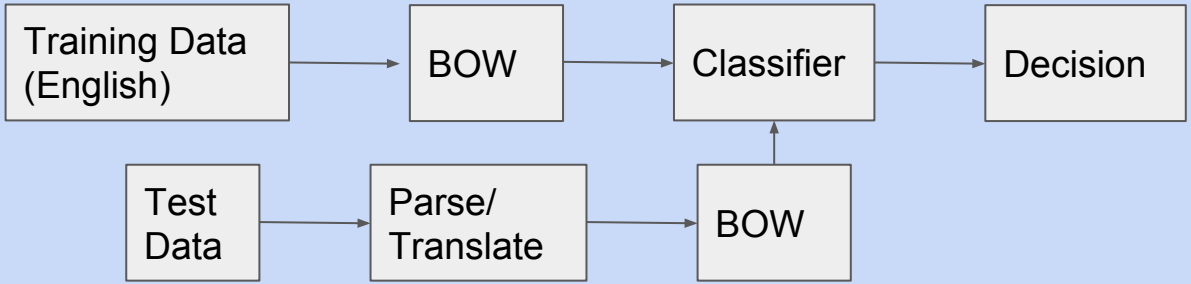
- Documents can be parsed into words
- Words can be mapped to their English word/words (Dictionary lookup)

Classifier Solution

Key Idea:

- Convert document to (BOW) histograms

- Leverage corpus of labeled documents in native language (Also just BOW)
  - Classifiers quickly tell what kind of document it is
  - Also classify as “Important for further detailed translation”



Benefits:

- Simple classifiers on BOW histograms are just linear feature vectors
  - Low memory, ultra fast decisions
- As more intel is gained, the decision rule(s) can be updated

Reuse of Software, Services, Data...

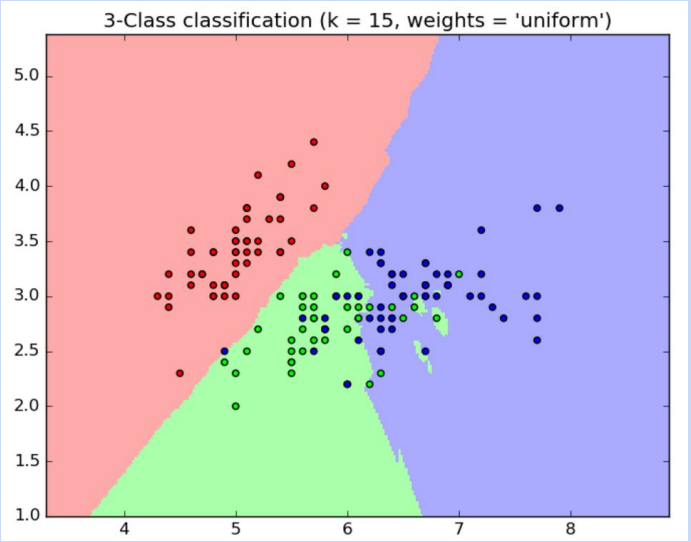
Sklearn learn library

For creating BOW:

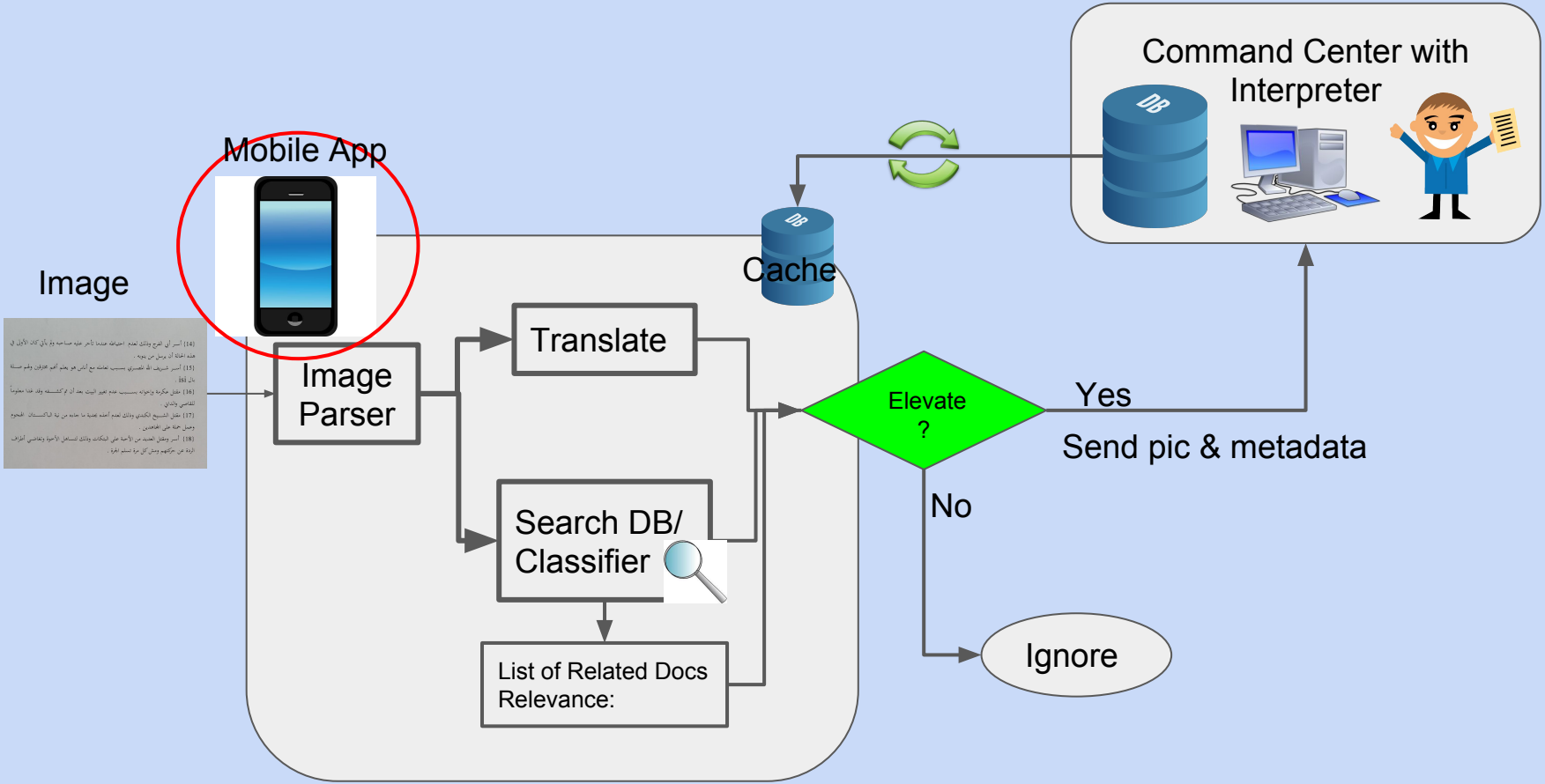
```
>>> from sklearn.feature_extraction.text import CountVectorizer
>>> vectorizer = CountVectorizer(min_df=1)
>>> vectorizer
      CountVectorizer(analyzer=...'word', binary=False, decode_error=...'strict',
        dtype=<... 'numpy.int64'>, encoding=...'utf-8', input=...'content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1, ...
```

For Classifier Training/Classification:

```
>>> from sklearn import neighbors # nearest neighbor
```



User Interface



Summary

- The challenge with translating Arabic has to do with image parsing.

- Presented an end-to-end solution with options for Image Parsing, Data Classification and decision making.
  - Google Search concept to find related documents
  - Neural Network techniques
  - Machine Learning techniques

- Acknowledgments:
  - Amazing Team
  - Thank you Labhack and Sponsors!

- upshot.strikingly.com

List of Resources

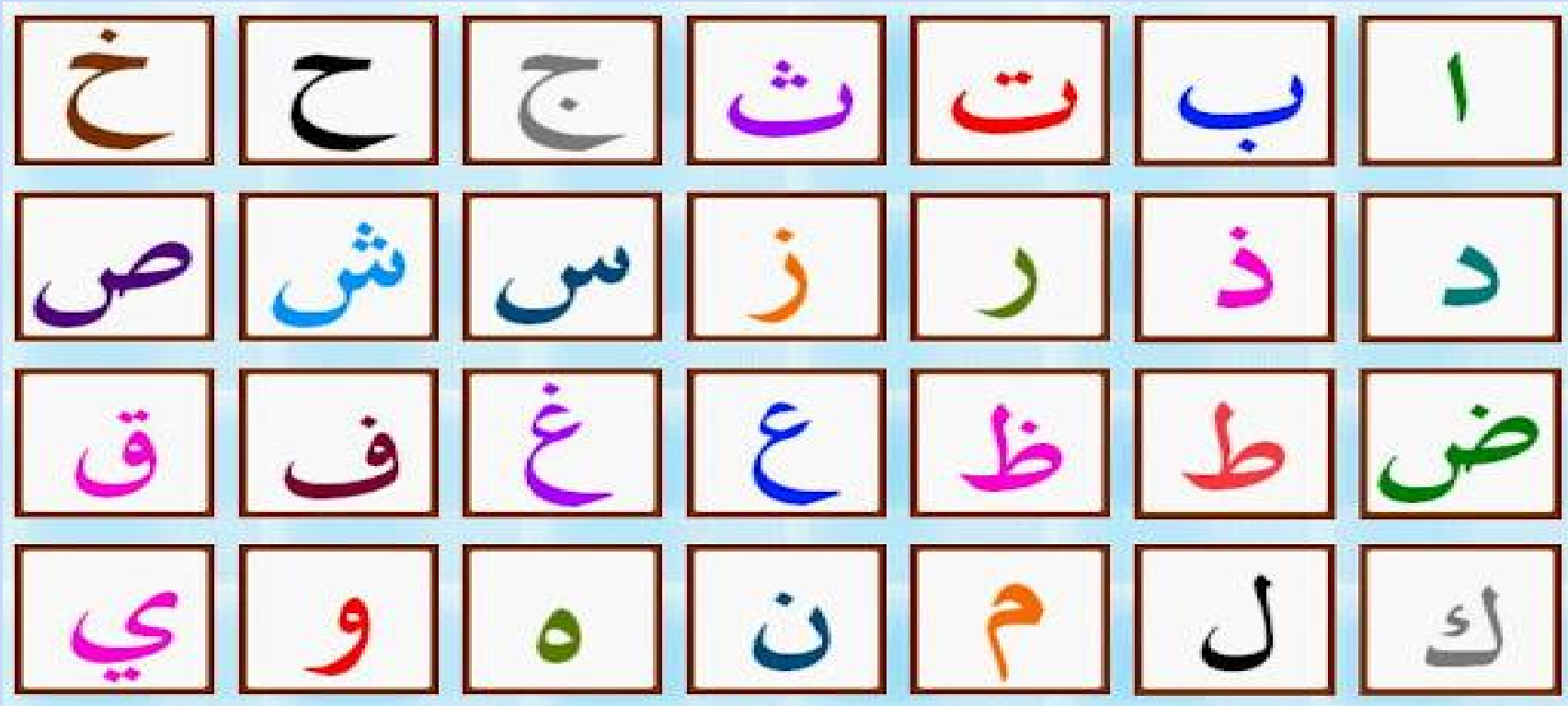
- http://arxiv.org/pdf/1206.1518.pdf - research paper for recognizing Arabic characters
- http://www.amara.org/en/videos/cI4VRKdprrGv/info/arabic-handwriting-recognition-using-neural-network-in-matlab-word-level-user-manual/ - video tutorial to use Matlab's Arabic Handwriting Recognition feature
- https://console.ng.bluemix.net/catalog/services/visual-recognition/ - specs for Watson's Visual Recognition service
-

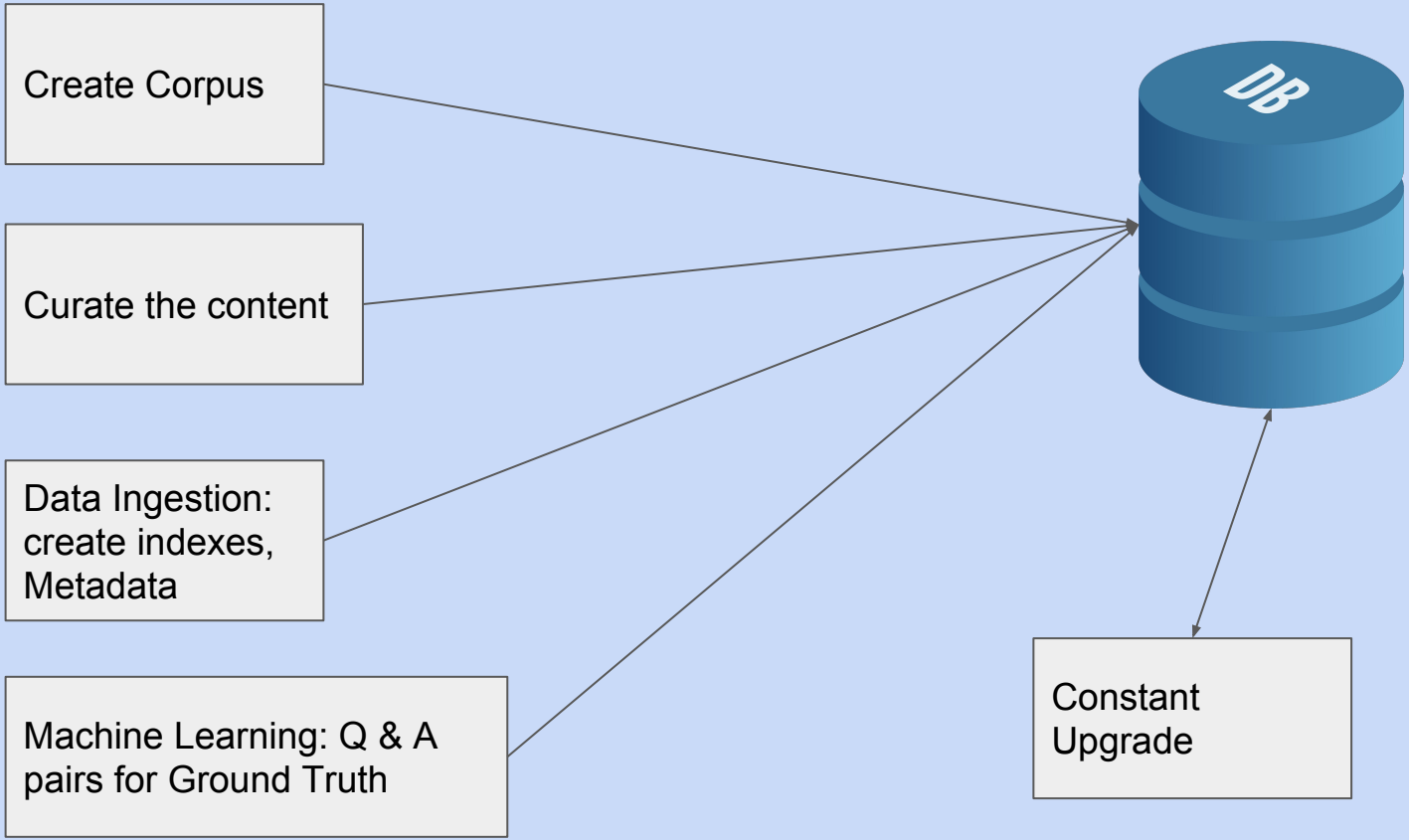
Backup Slides

Language Recognition

- There are thousands of languages spoken on Earth, but the top ten languages cover most speech.
- In field operations, English would be expected to be the default language in only a small percentage of situations.
- Arabic is spoken by

Simple Yet Complicated




Create Corpus
Curate the content
Data Ingestion: create indexes, Metadata
Machine Learning: Q & A pairs for Ground Truth
DB
Constant Upgrade

What Classifier?

- Classifier 1: Is it important for detailed translation by human?
- Classifier 2: What type of document is it like?/Which documents that I know are like it?

- You want:
  - Decision confidence
  - Fast lightweight decision
  - Small data decision update
    - **Nearest neighbor:** Distance implicitly tells you confidence about what type of document it is
    - **Relevance vector machine (RVM):** Gives you the category as well as explicit confidence score [0,1]
      Simple weight vector **(super small memory)**

Limitations

Hidden assumption:

- BOW for word -> word translation will give proper representation for classifying using our English corpus.

If not:

- Learn an intermediate mapping that re-weights other language BOW to native language BOW.
  - Simple least squares reweighting of histogram
  - Cross correlational analysis

Emails:

Craig.Boman@gmail.com

Joann.Luu@gmail.com

rmcchesney@threescale.com

sriveravi@gmail.com

cleverclue@gmail.com

d8tascientist@gmail.com

(one more email omitted)