AI for Document Understanding

Final Presentation – 18.02.2019

TUM-DI-LAB Team

Wassim

Studies:
Master Mathematics

Alican

Studies:
Master Mathematical Finance and Actuarial Science

Cingis

Studies:
Master Mathematics in Data Science

Oliver

Studies:
Master Mathematics

Hey, my name is John.

I work at a big accounting consulting company.

Today is a bad day.

Investigate an audit of a big firm.

Done in 1995. No digital trace.

Letters.

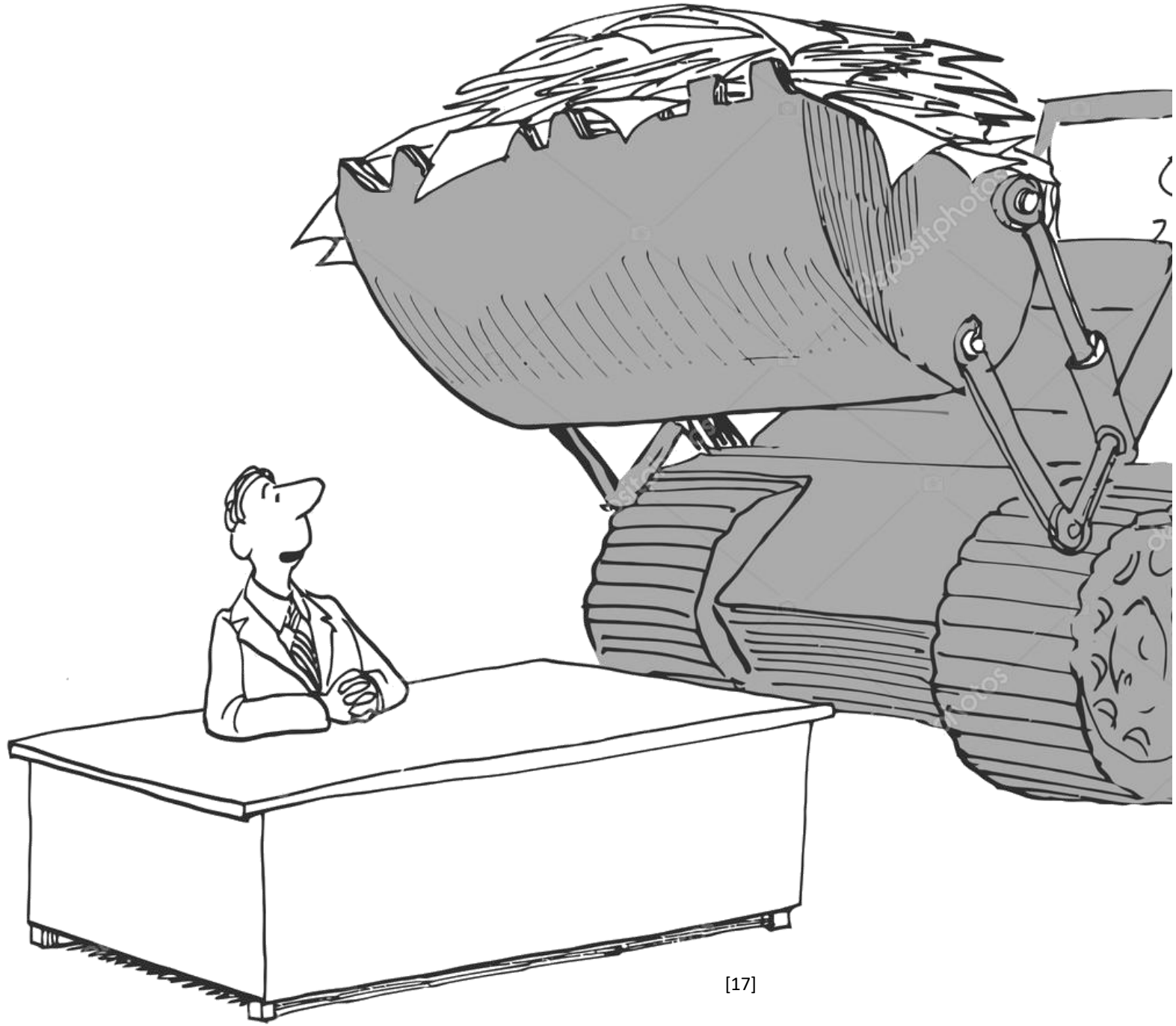
Receipts.

Invoices.

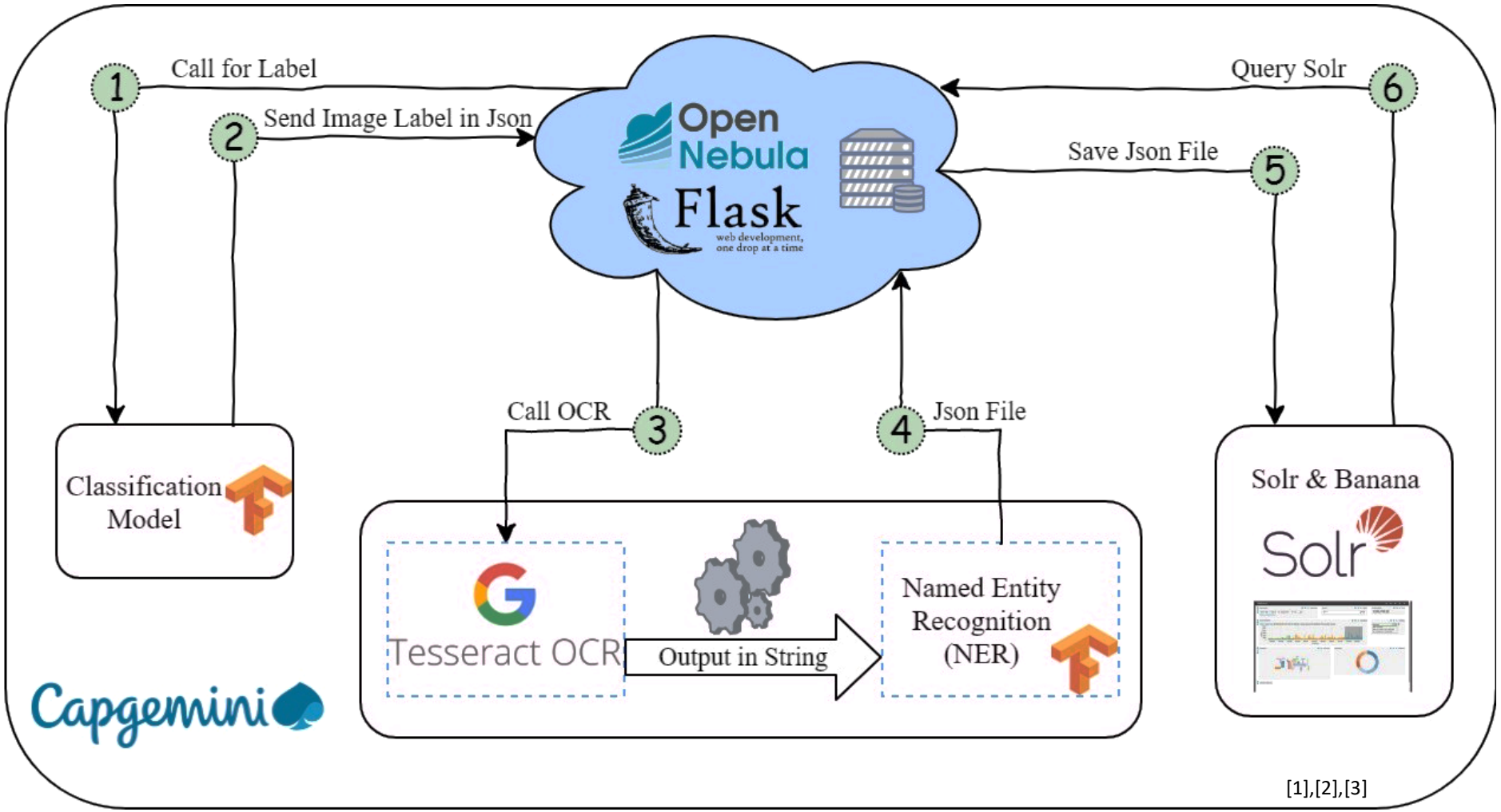
Emails.

Thousands of them.

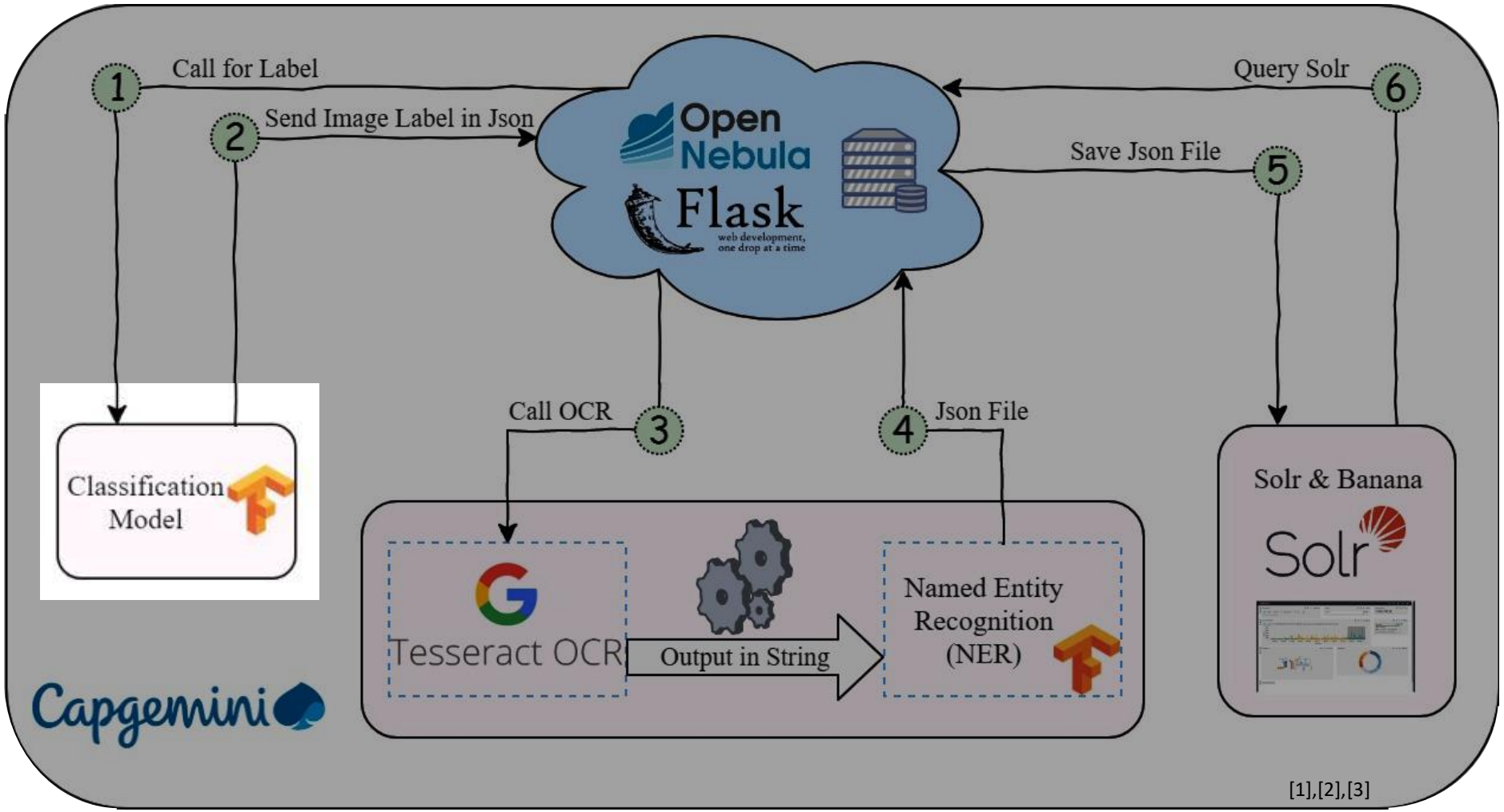
How it feels like?



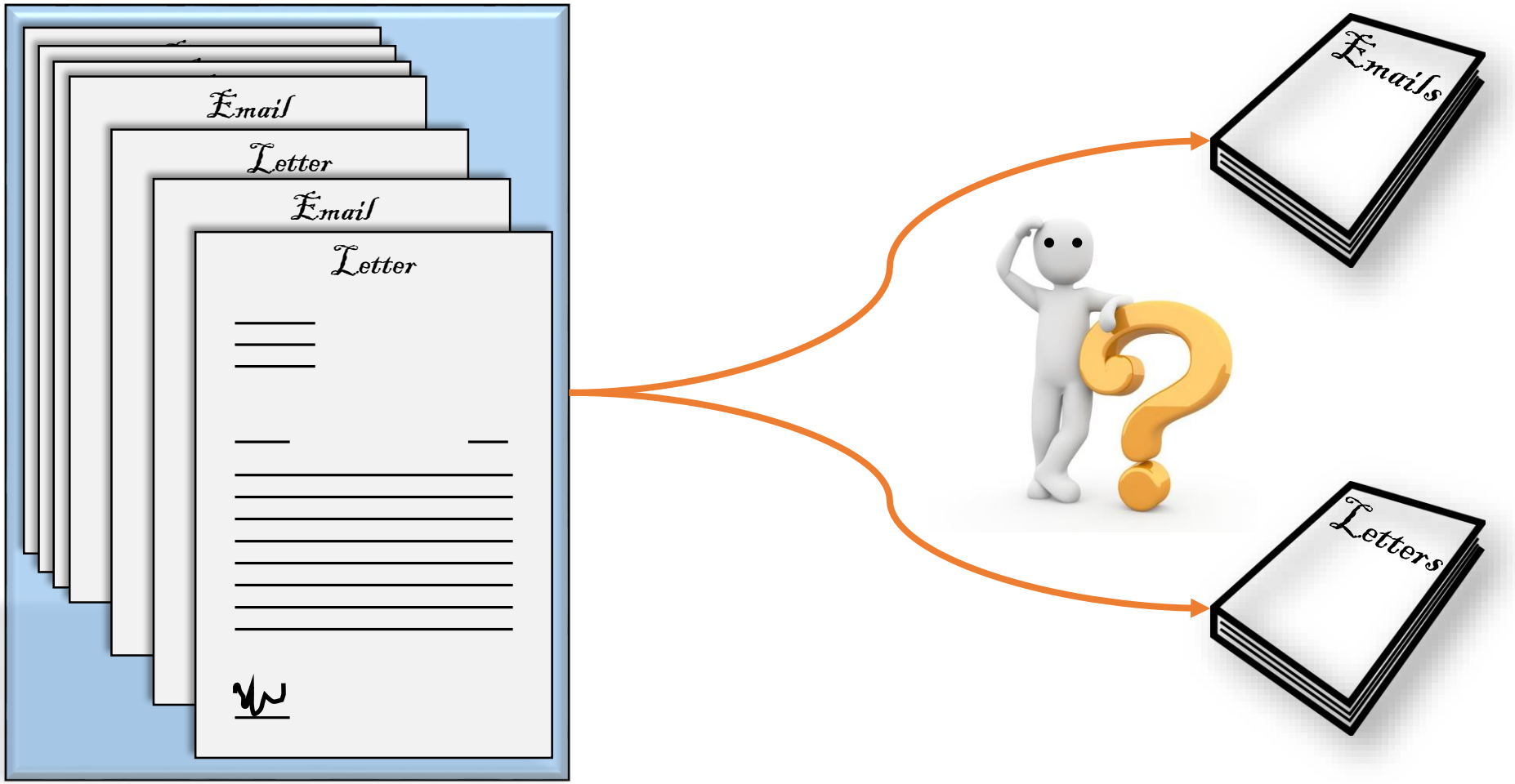
[17]


1
Call for Label
2
Send Image Label in Json
Open Nebula
Flask
web development, one drop at a time
Query Solr
6
Save Json File
5
Call OCR
3
4
Json File
Classification Model
Tesseract OCR
Output in String
Named Entity Recognition (NER)
Solr & Banana
Solr
Capgemini

[1],[2],[3]

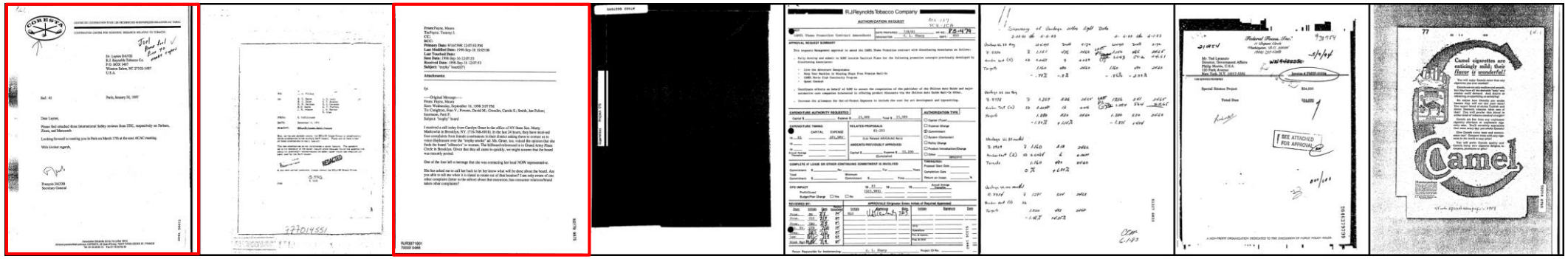

1
Call for Label
2
Send Image Label in Json
Open Nebula
Flask
web development, one drop at a time
Query Solr
6
Save Json File
5
Classification Model
Call OCR
3
4
Json File
Solr & Banana
Solr
Tesseract OCR
Output in String
Named Entity Recognition (NER)
Capgemini
[1],[2],[3]

Task



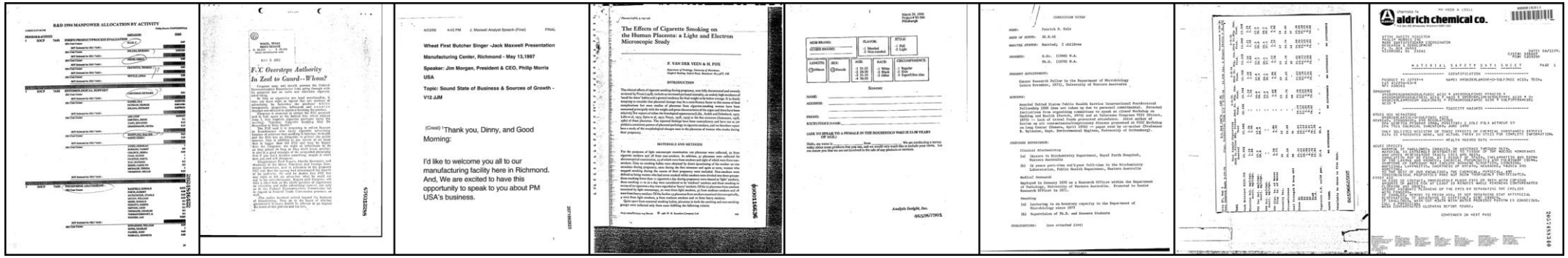
Dataset

- RVL-CDIP dataset



[14]

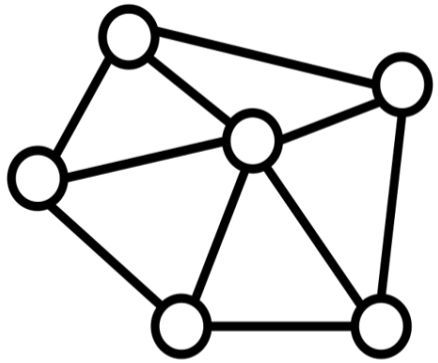
Letter | Memo | Email | Filefolder | Form | Handwritten | Invoice | Advertisement



[15]

Budget | News Article | Presentation | Scientific Publication | Questionnaire | Resume | Scientific Report | Specification

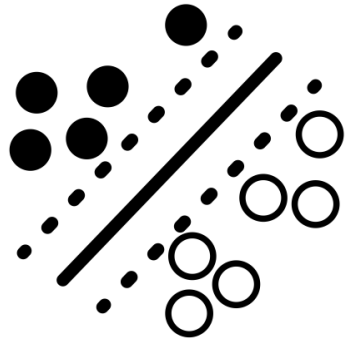
First Steps



Graph-Based [11]

+ Unconventional approach

- Complicated preprocessing
- Hard to implement



Content-Based [12]

+ Very good accuracy

- Complicated preprocessing

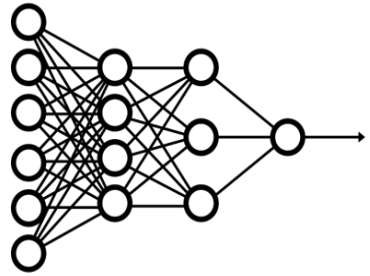
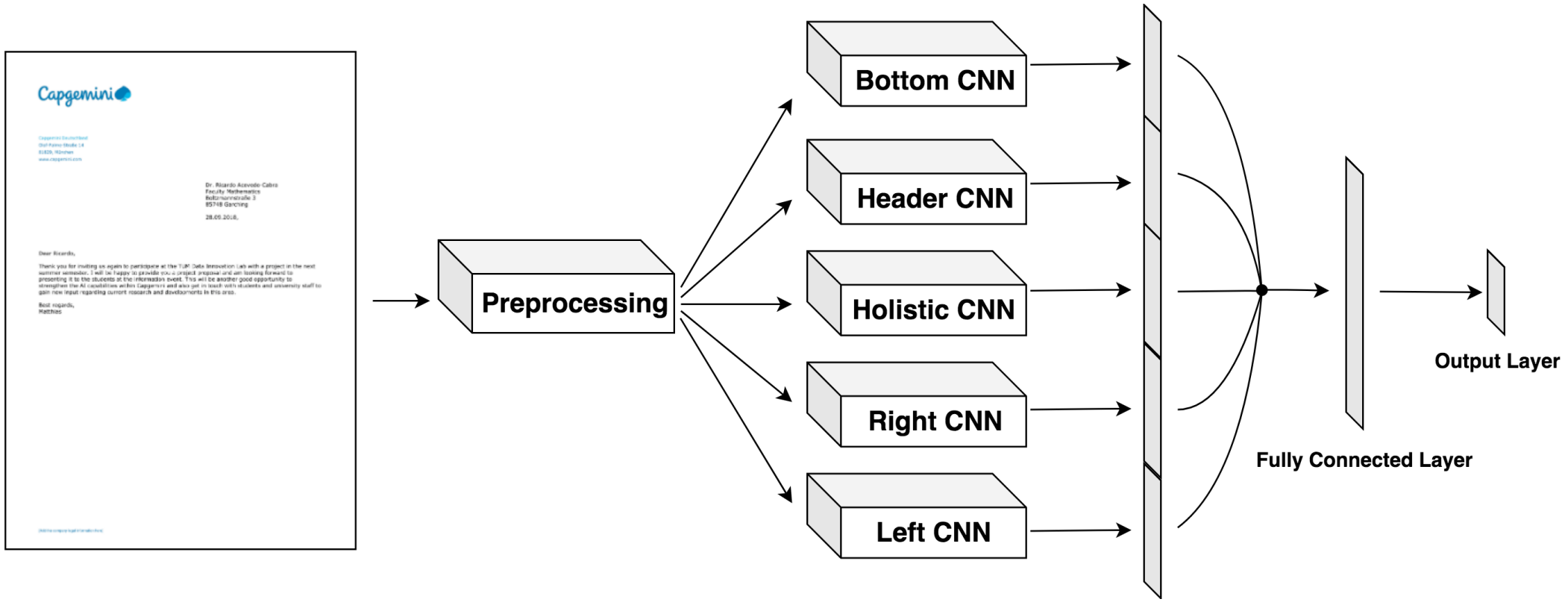


Image-Based [13]

+ State-of-the-art
+ Simple preprocessing
+ Matches our project goals

- Training time

Classification Model



- Validation accuracy: 0.98
- Test accuracy: 0.98

Preprocessing

Bottom

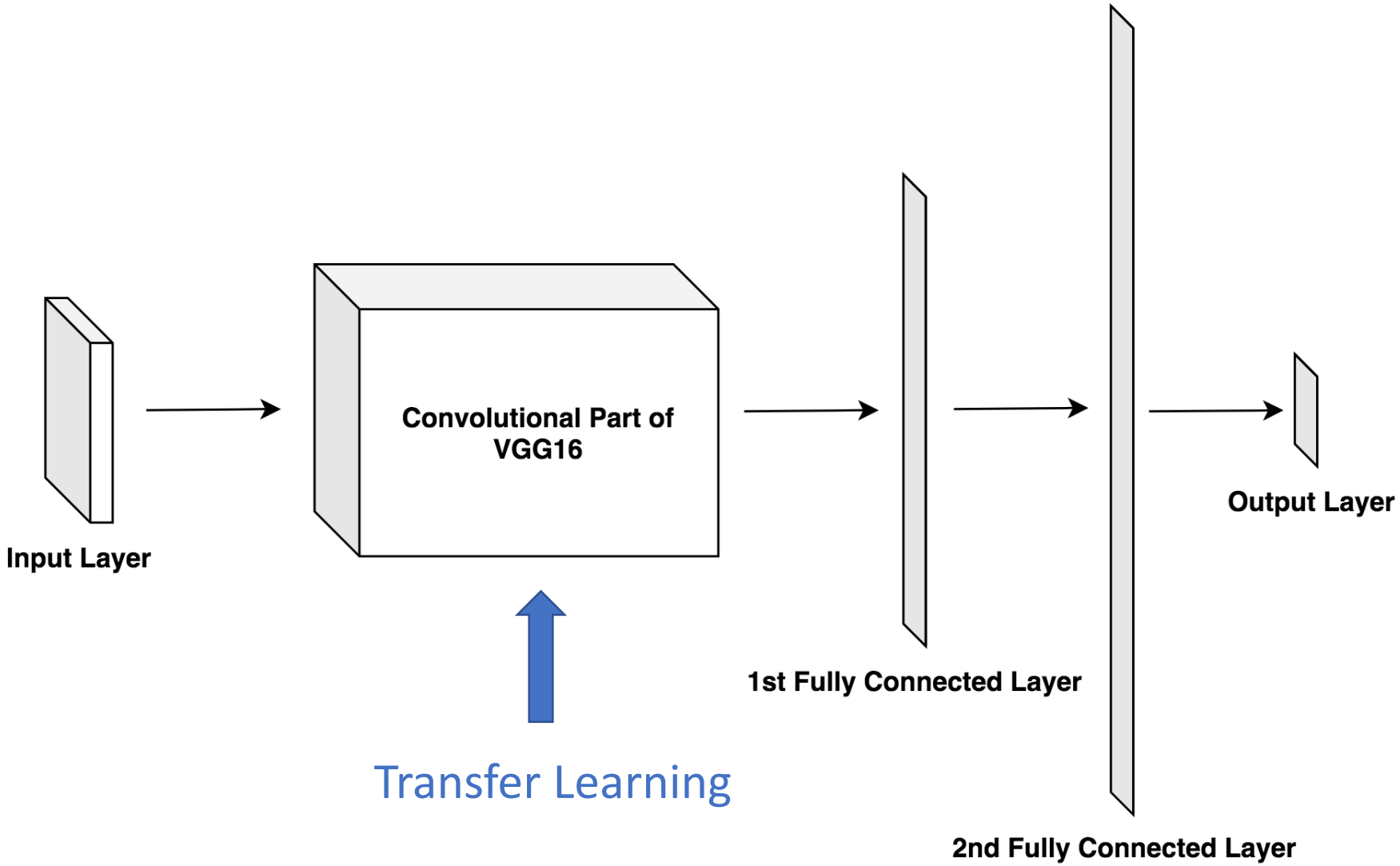
Header

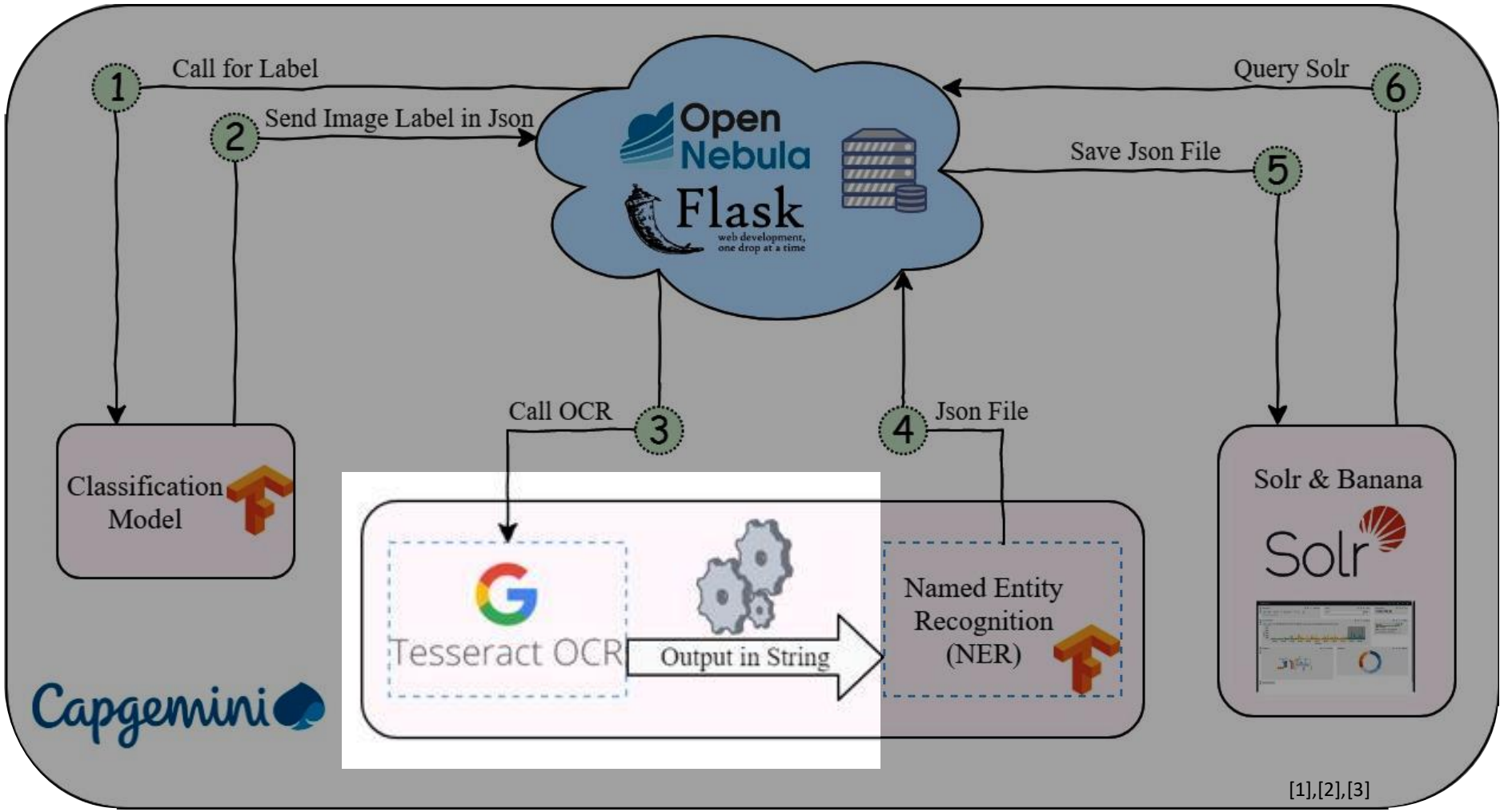
Holistic

Left

Right

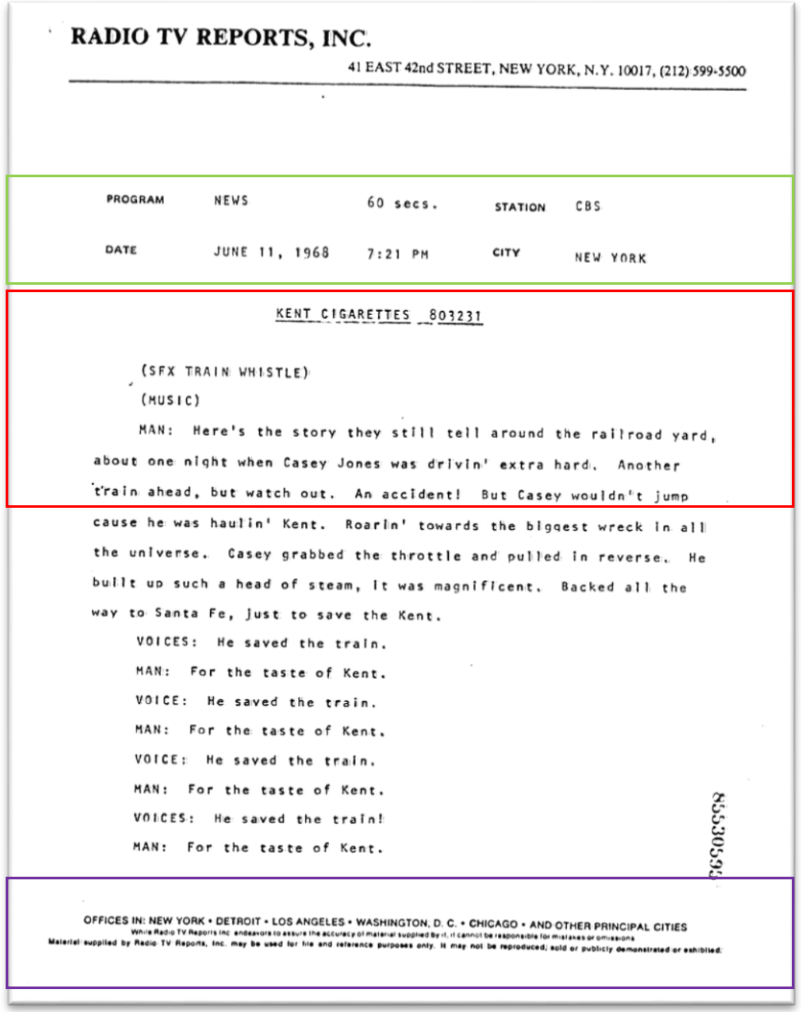
Region-Based & Holistic CNNs




1
Call for Label
2
Send Image Label in Json
Open Nebula
Flask
web development, one drop at a time
Query Solr
6
Save Json File
5
Call OCR
3
4
Json File
Classification Model
Tesseract OCR
Output in String
Named Entity Recognition (NER)
Solr & Banana
Solr
Capgemini

[1],[2],[3]

OCR Illustration



RADIO TV REPORTS, INC.

41 EAST 42nd STREET, NEW YORK, N.Y. 10017, (212) 599-5500

PROGRAM	NEWS	60 secs.	STATION	CBS	
DATE	JUNE 11, 1968	7:21 PM	CITY	NEW YORK	

KENT CIGARETTES 803231

(SFX TRAIN WHISTLE)
(MUSIC)
MAN: Here's the story they still tell around the railroad yard, about one night when Casey Jones was drivin' extra hard. Another train ahead, but watch out. An accident! But Casey wouldn't jump cause he was haulin' Kent. Roarin' towards the biggest wreck in all the universe. Casey grabbed the throttle and pulled in reverse. He built up such a head of steam, it was magnificent. Backed all the way to Santa Fe, just to save the Kent.
VOICES: He saved the train.
MAN: For the taste of Kent.
VOICE: He saved the train.
MAN: For the taste of Kent.
VOICE: He saved the train.
MAN: For the taste of Kent.
VOICES: He saved the train!
MAN: For the taste of Kent.

8553059

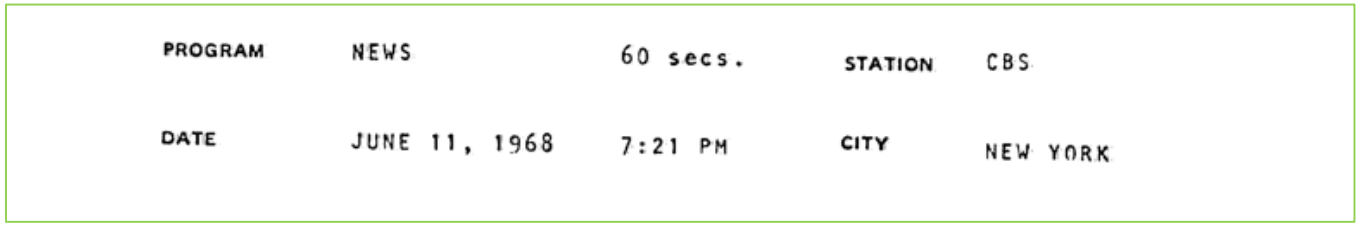
PROGRAM	NEWS	60 secs.	STATION	CBS	
DATE	JUNE 11, 1968	7:21 PM	CITY	NEW YORK	

KENT CIGARETTES 803231

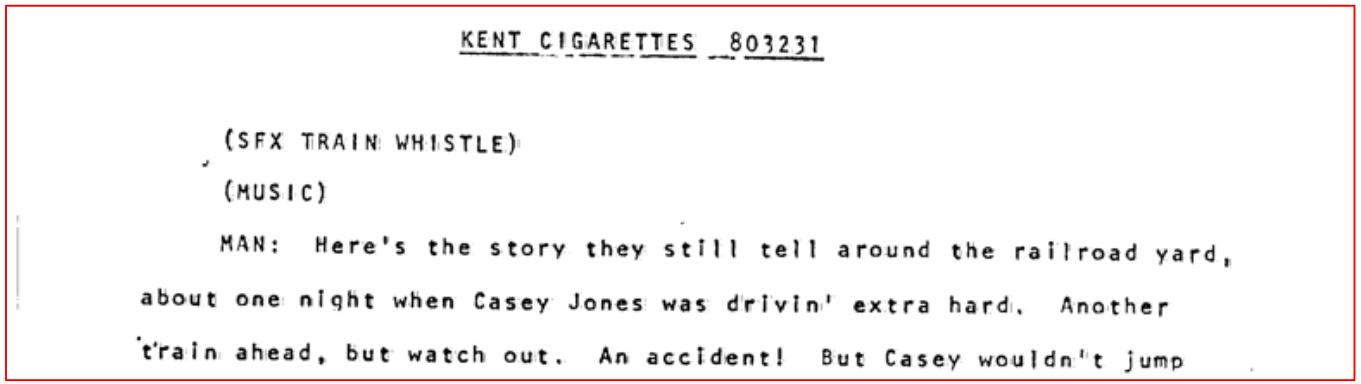
(SFX TRAIN WHISTLE)
(MUSIC)
MAN: Here's the story they still tell around the railroad yard, about one night when Casey Jones was drivin' extra hard. Another train ahead, but watch out. An accident! But Casey wouldn't jump

OCR Illustration

PROGRAM NEWS 60 secs. STATION CBS
DATE JUNE 11, 1968 7:21 PM CITY NEW YORK

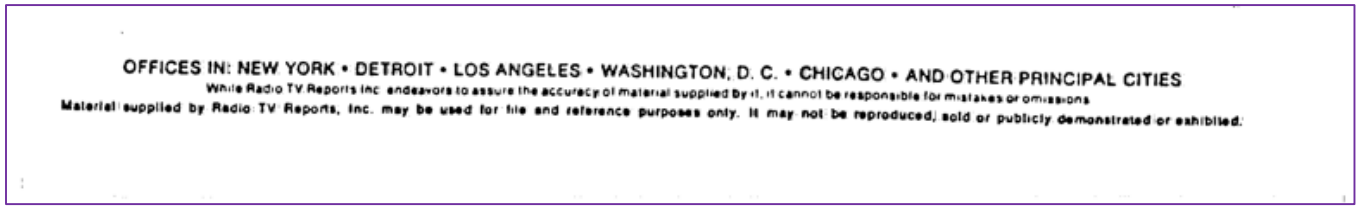

PROGRAM NEWS 60 secs. STATION CBS
DATE JUNE 11, 1968 7:21 PM CITY NEW YORK

KENT CIGARETTES 803231
(srx TRAIN WHISTLE)
(MUSIC)
MAN: Here's the story they still tell around the railroad yard, about one nlqht when Casey Jones was drlvln' extra hard. Another train ahead, but watch out. An accident! But Casey wouldn't jump

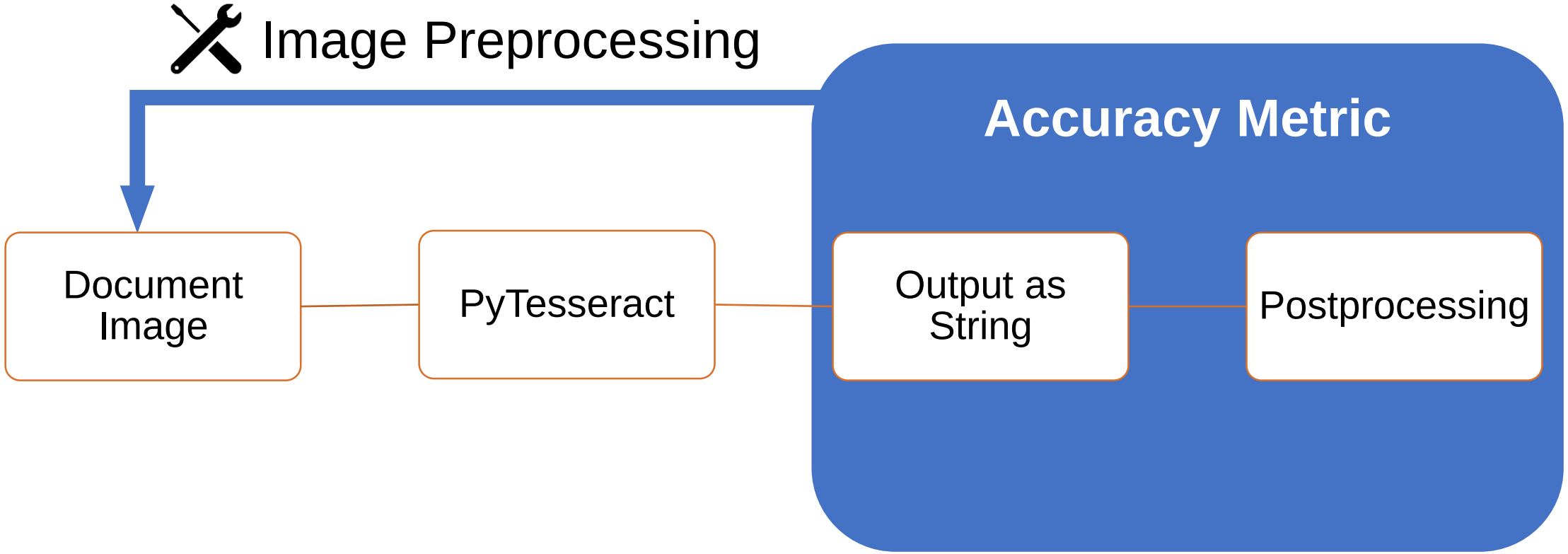

KENT CIGARETTES 803231

(SFX TRAIN WHISTLE)
(MUSIC)
MAN: Here's the story they still tell around the railroad yard, about one night when Casey Jones was drivin' extra hard. Another train ahead, but watch out. An accident! But Casey wouldn't jump

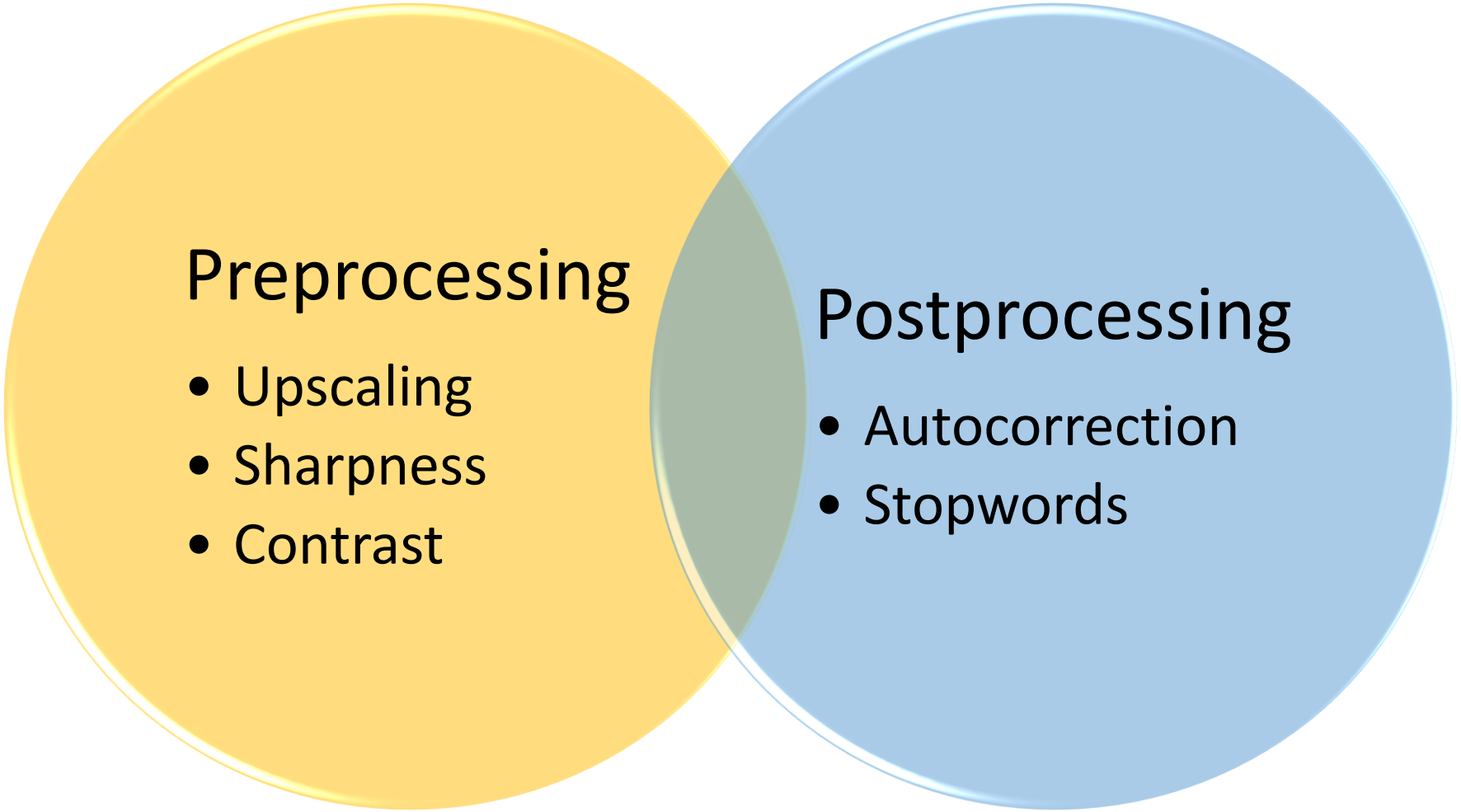
OFFICES IN: NEW YORK.DETROIT.LOS ANGELES.WASNINGTON D C.CHICAGO AND OTHER PRINCIPAL CITIES
While Padro TV Reoriis imc endeevors to ...


OFFICES IN: NEW YORK • DETROIT • LOS ANGELES • WASHINGTON, D. C. • CHICAGO • AND OTHER PRINCIPAL CITIES
While Radio TV Reports Inc endeavors to assure the accuracy of material supplied by it, it cannot be responsible for mistakes or omissions
Material supplied by Radio TV Reports, Inc. may be used for file and reference purposes only. It may not be reproduced, sold or publicly demonstrated or exhibited.

OCR Pipeline

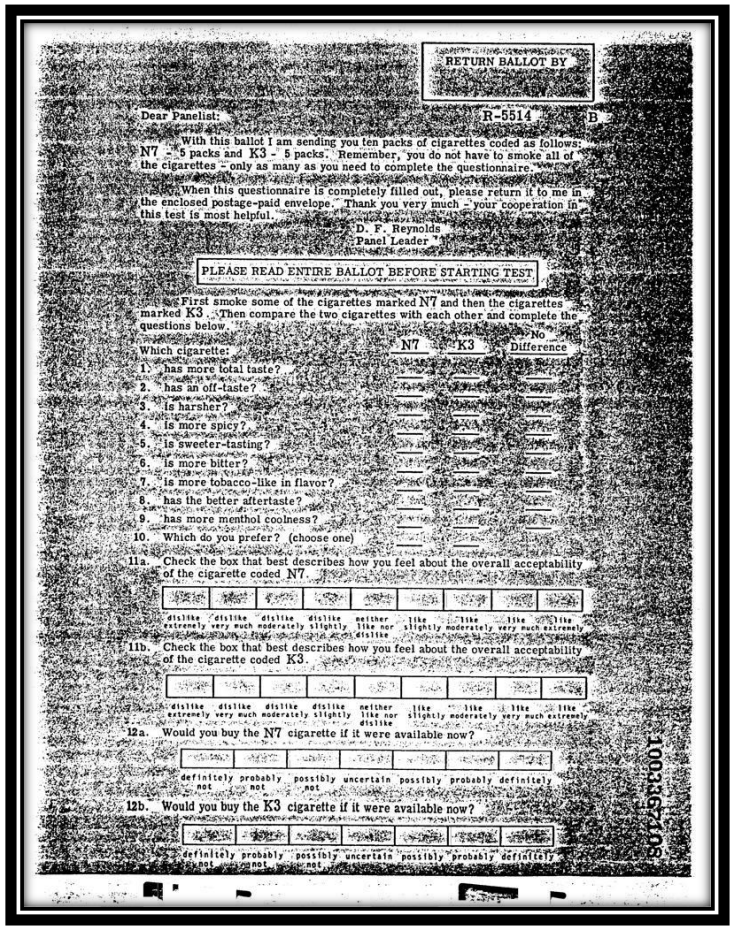


Preprocessing & Postprocessing



Usefulness of Preprocessing

Factors trained on



RETURN BALLOT BY

Dear Panelist: R-5514 B

With this ballot I am sending you ten packs of cigarettes coded as follows: N7 - 5 packs and K3 - 5 packs. Remember, you do not have to smoke all of the cigarettes - only as many as you need to complete the questionnaire.

When this questionnaire is completely filled out, please return it to me in the enclosed postage-paid envelope. Thank you very much - your cooperation in this test is most helpful.

D. F. Reynolds
Panel Leader

PLEASE READ ENTIRE BALLOT BEFORE STARTING TEST

First smoke some of the cigarettes marked N7 and then the cigarettes marked K3. Then compare the two cigarettes with each other and complete the questions below.

Which cigarette:	N7	K3	No Difference
1. has more total taste?			
2. has an off-taste?			
3. is harsher?			
4. is more spicy?			
5. is sweeter-tasting?			
6. is more bitter?			
7. is more tobacco-like in flavor?			
8. has the better aftertaste?			
9. has more menthol coolness?			
10. Which do you prefer? (choose one)			

11a. Check the box that best describes how you feel about the overall acceptability of the cigarette coded N7.

dislike extremely | dislike very much | dislike moderately | dislike slightly | neither like nor dislike | like slightly | like moderately | like very much | like extremely

11b. Check the box that best describes how you feel about the overall acceptability of the cigarette coded K3.

dislike extremely | dislike very much | dislike moderately | dislike slightly | neither like nor dislike | like slightly | like moderately | like very much | like extremely

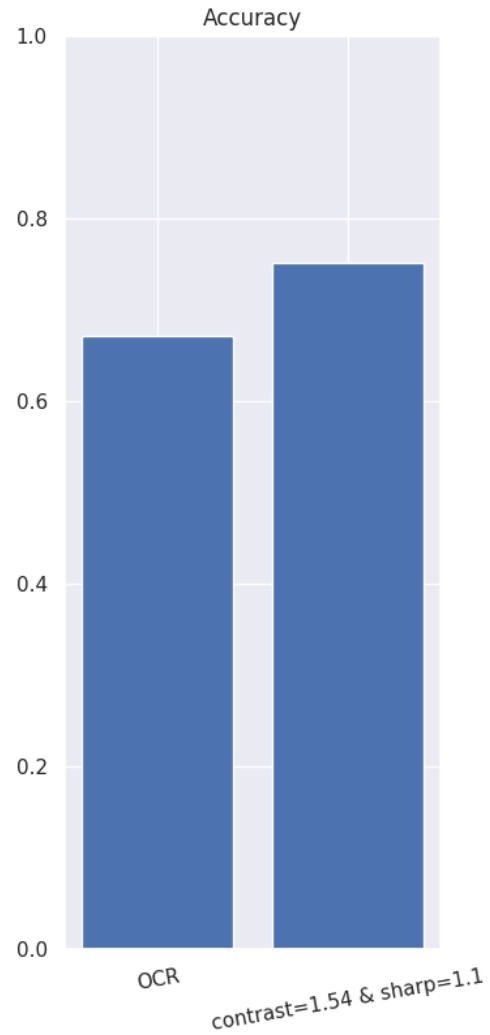
12a. Would you buy the N7 cigarette if it were available now?

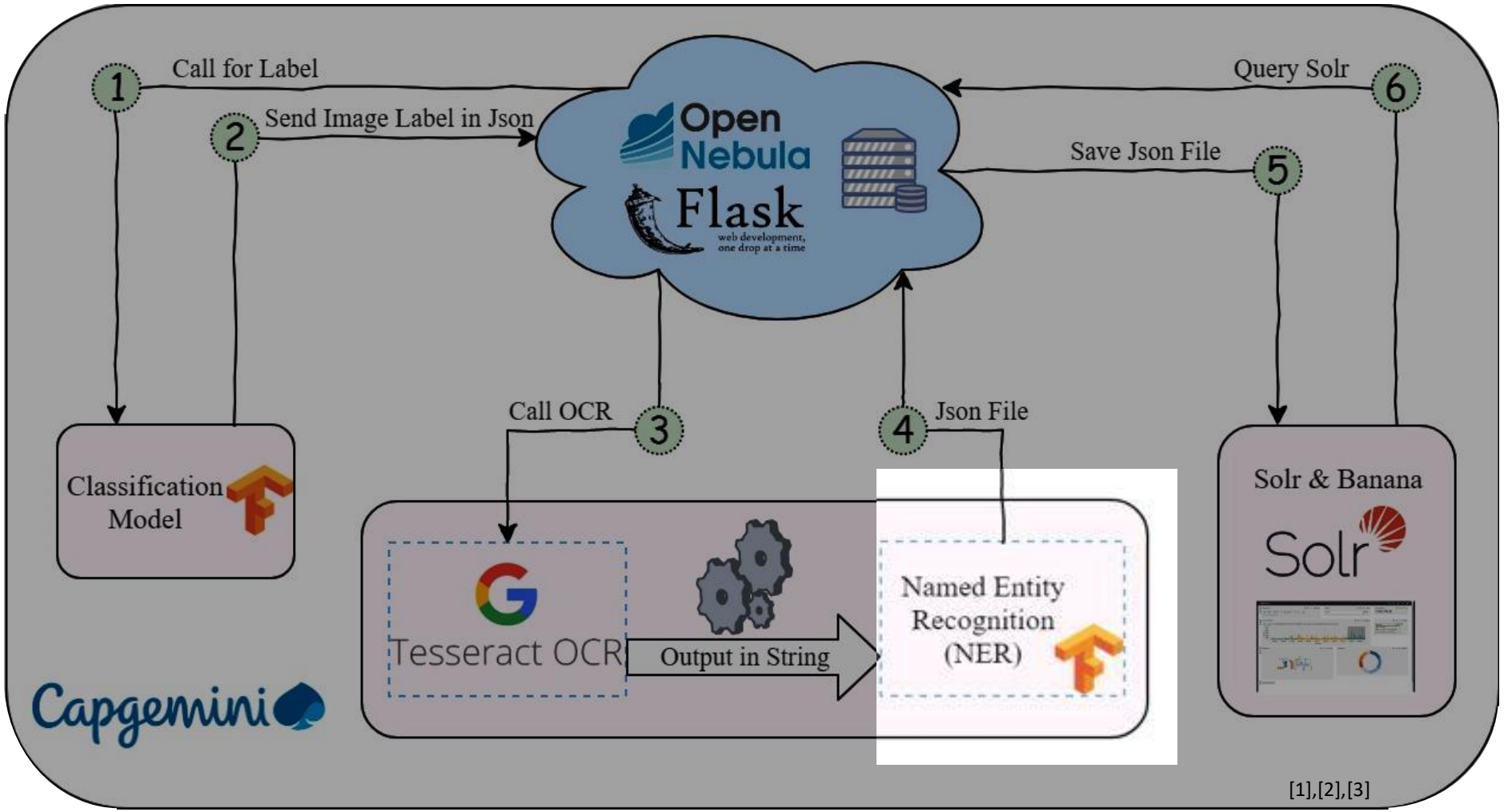
definitely not | probably not | possibly not | uncertain | possibly | probably | definitely

12b. Would you buy the K3 cigarette if it were available now?

definitely not | probably not | possibly not | uncertain | possibly | probably | definitely

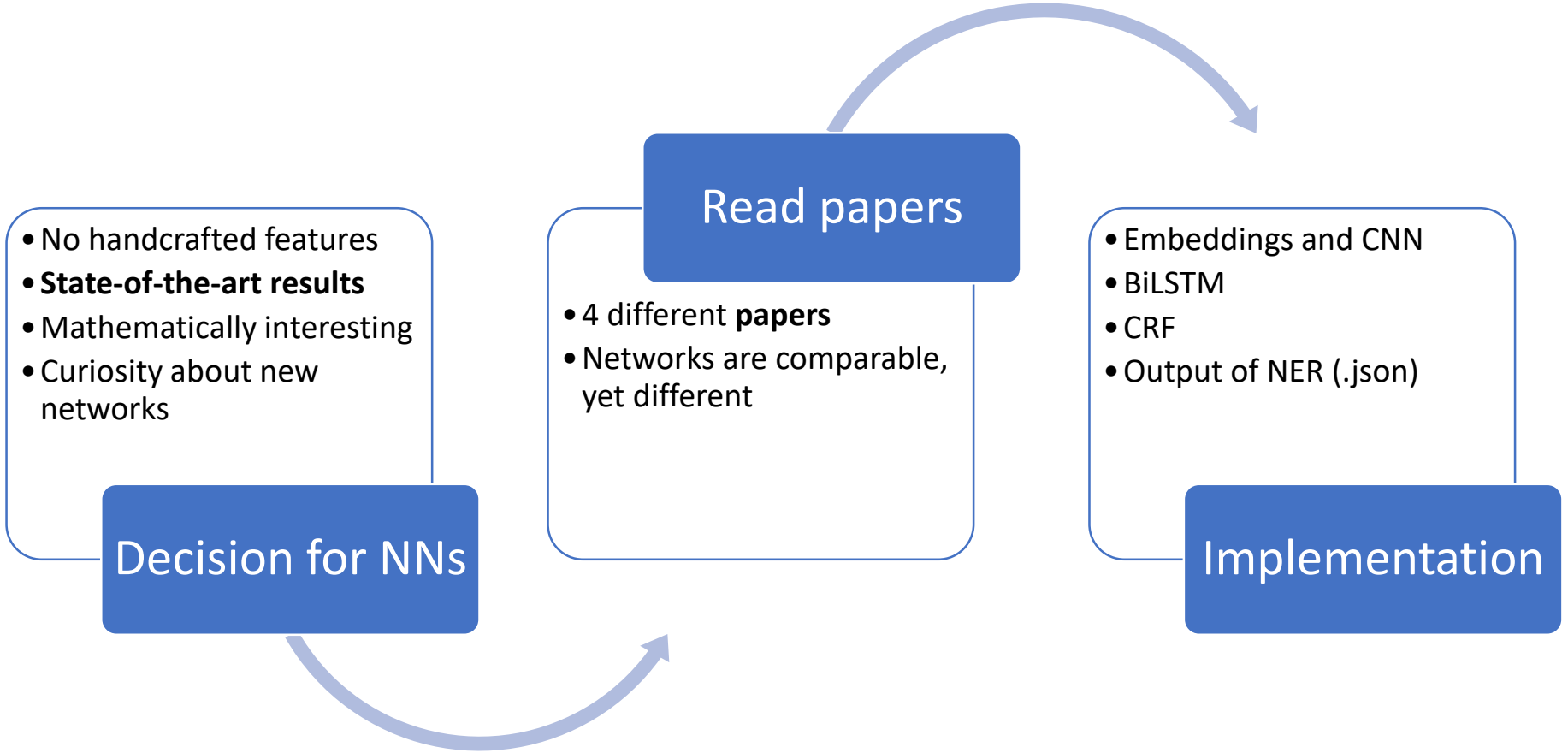
100-3-5-7-10-S




1
Call for Label
2
Send Image Label in Json
Open Nebula
Flask
web development, one drop at a time
Query Solr
6
Save Json File
5
Call OCR
3
4
Json File
Classification Model
Tesseract OCR
Output in String
Named Entity Recognition (NER)
Solr & Banana
Solr
Capgemini

[1],[2],[3]

Research – NER Algorithms



Named Entity Recognition

Named Entities: Organization Location Person

Named Entity Recognition

Named Entities: Organization Location Person

Facebook Inc. is endowing a new institute, led by Christoph Luetge , devoted to the ethics of artificial intelligence at the Technical University of Munich , in Germany .

[10]

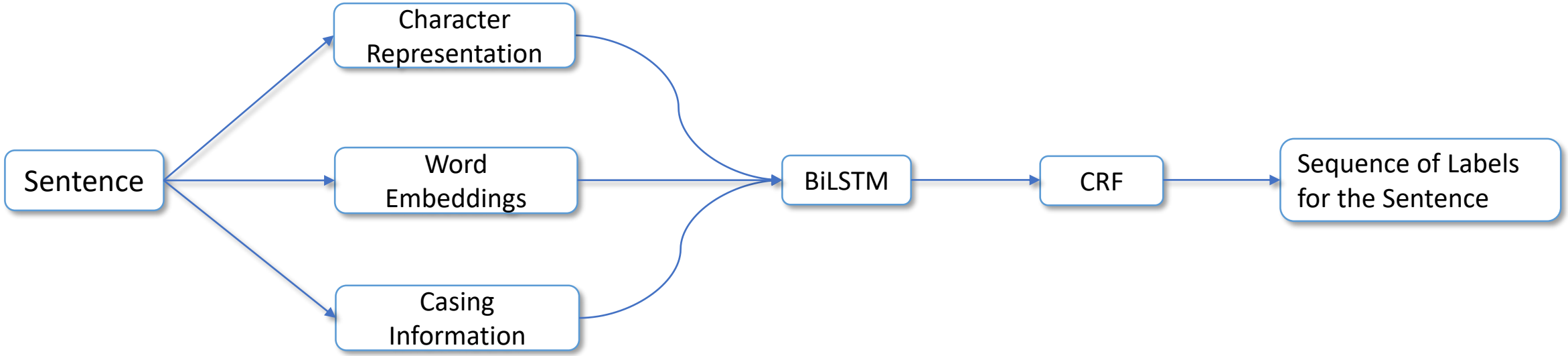
Named Entity Recognition

Named Entities: Organization Location Person

Facebook Inc. is endowing a new institute, led by Christoph Luetge, devoted to the ethics of artificial intelligence at the Technical University of Munich, in Germany.

[10]

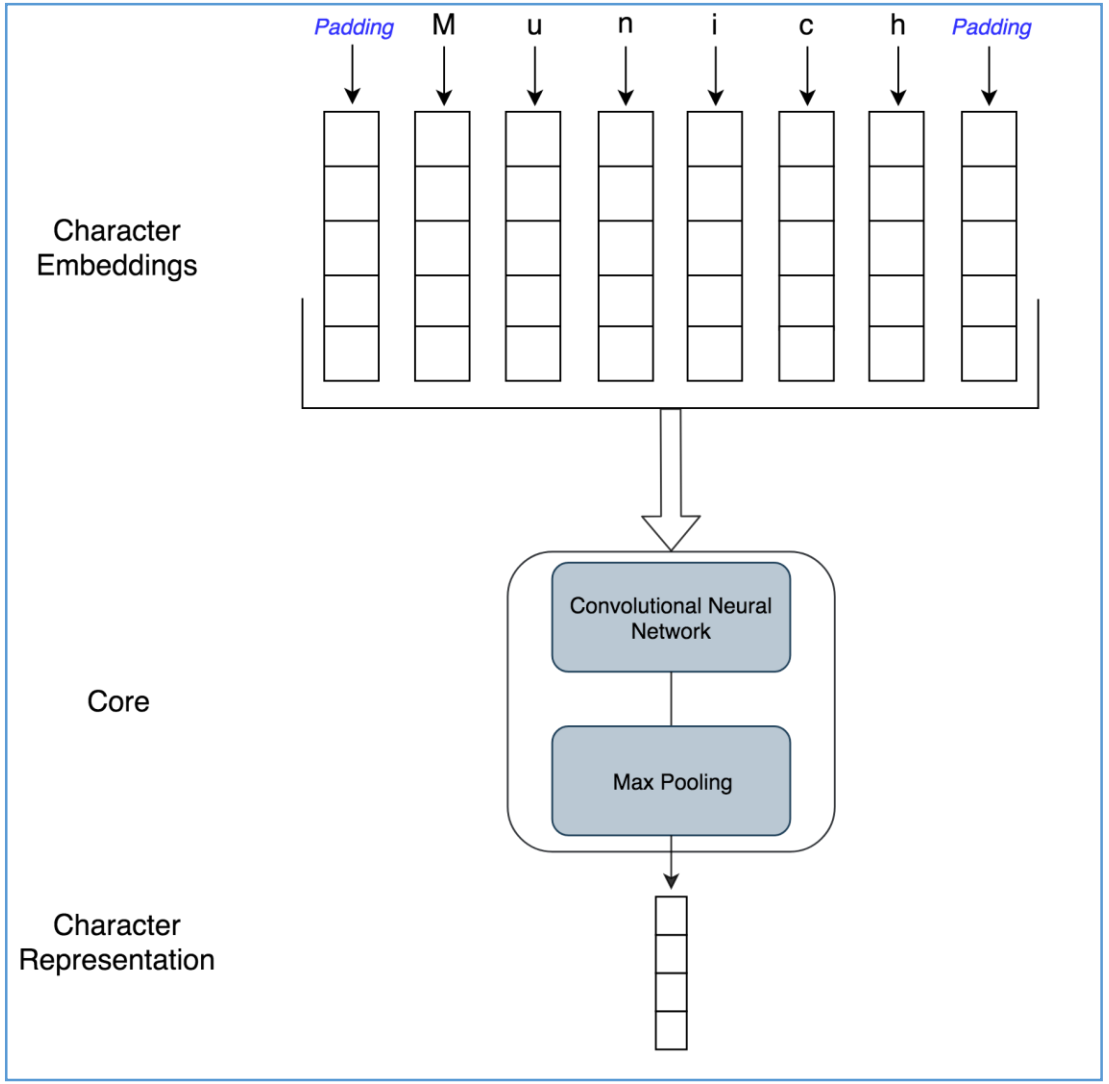
BiLSTM-CNN-CRF: Complete Architecture



- Validation accuracy: 94.2%
- Test accuracy: 90.8%

CNN-Extracted Features

CNN: Transform character-level information into *character-representation*

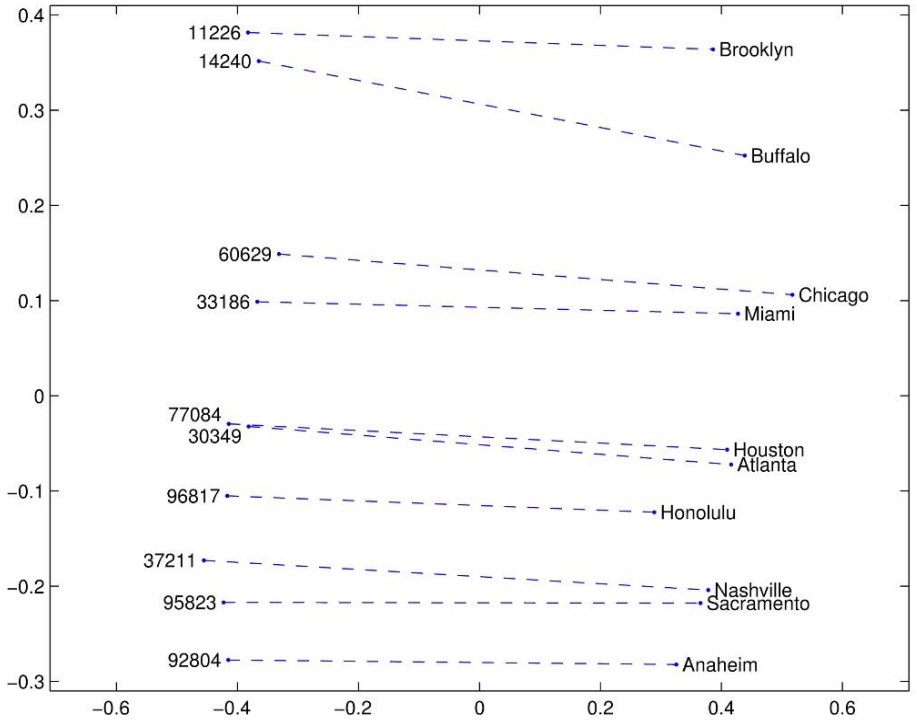


Word-Embeddings using GloVe

Word Embeddings:

- *Pretrained* word embeddings with *GloVe* 100 dimensions.

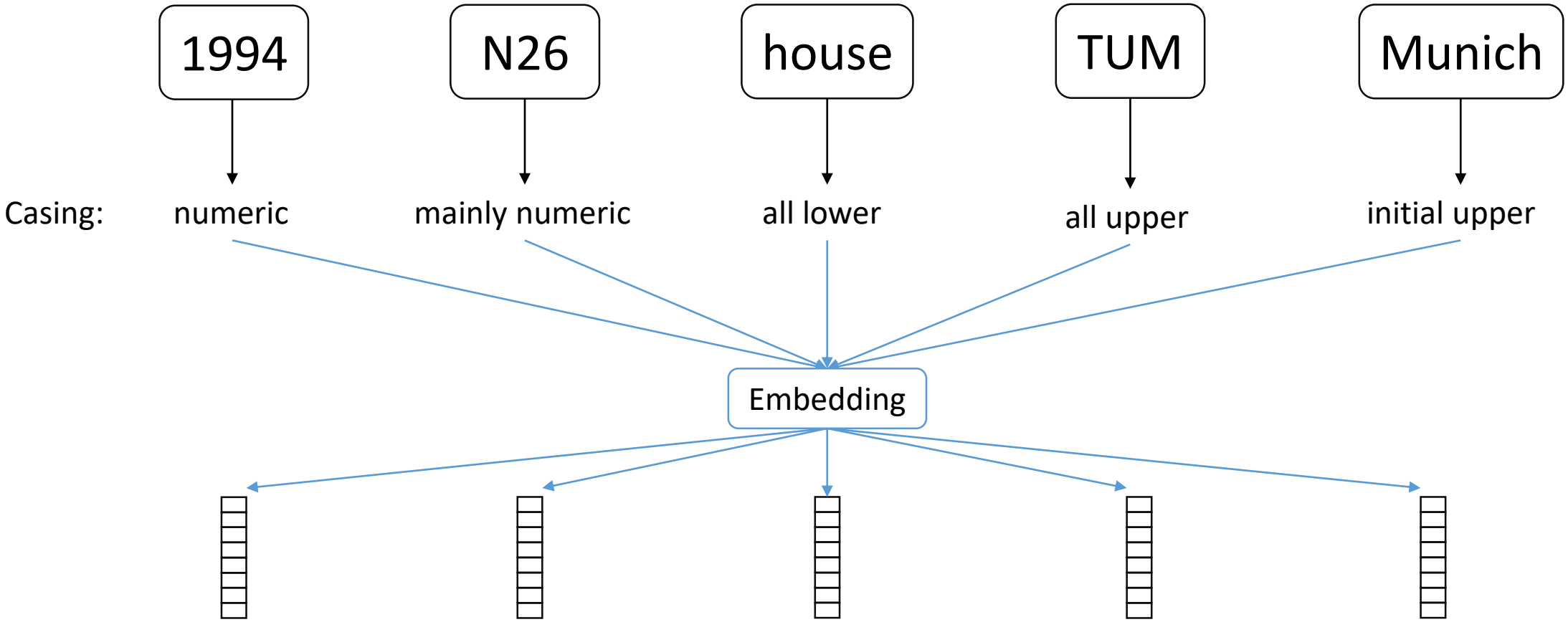
GloVe: Projection into 2D



[16]

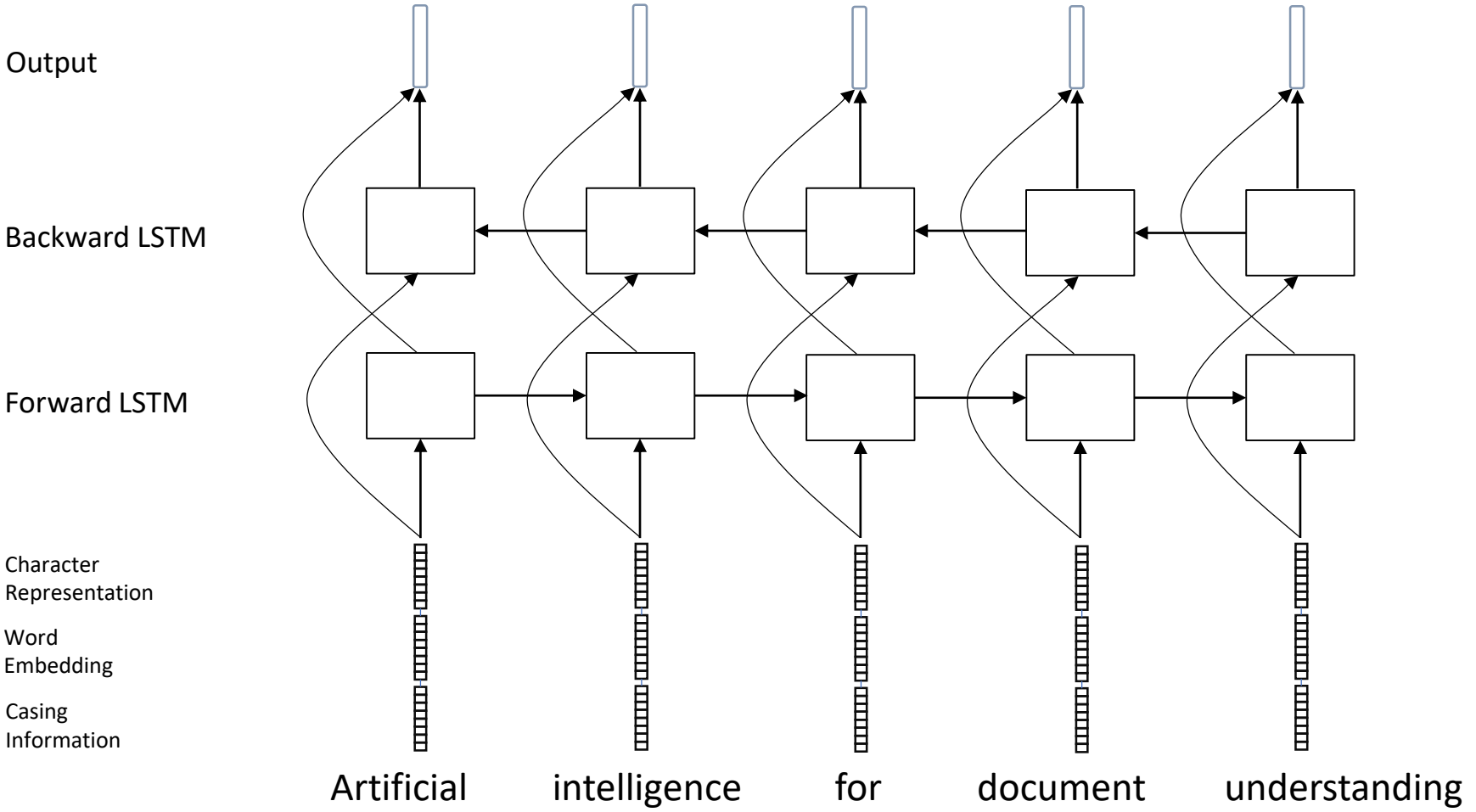
Casing Information

The casing of a word contains important information



Bi-Directional Long Short-Term Memory

Capture context on both sides of a word of a sentence



Conditional Random Field

Which sequence of labels is most likely for the sentence?

[…]	Technical	University	of	Munich,	in	Germany.

Conditional Random Field

Which sequence of labels is most likely for the sentence?

[…] Technical University of Munich, in Germany.

Not likely:

[…] Technical University (Organization) of (Person) Munich (Location), in Germany (Location) .

Conditional Random Field

Which sequence of labels is most likely for the sentence?

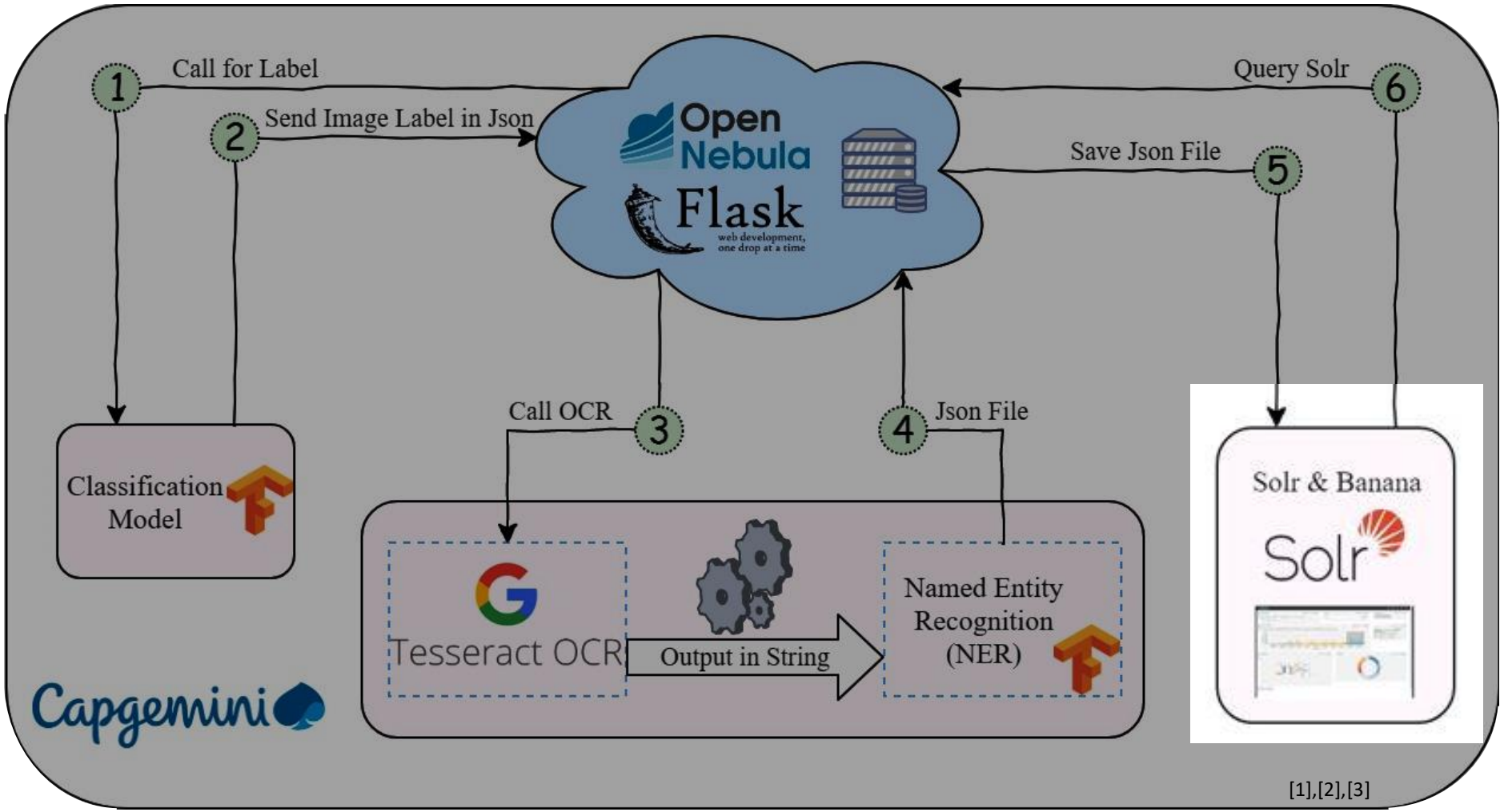
[…] Technical University of Munich, in Germany.

Not likely:

[…] Technical University (Organization) of (Person) Munich (Location), in Germany (Location) .

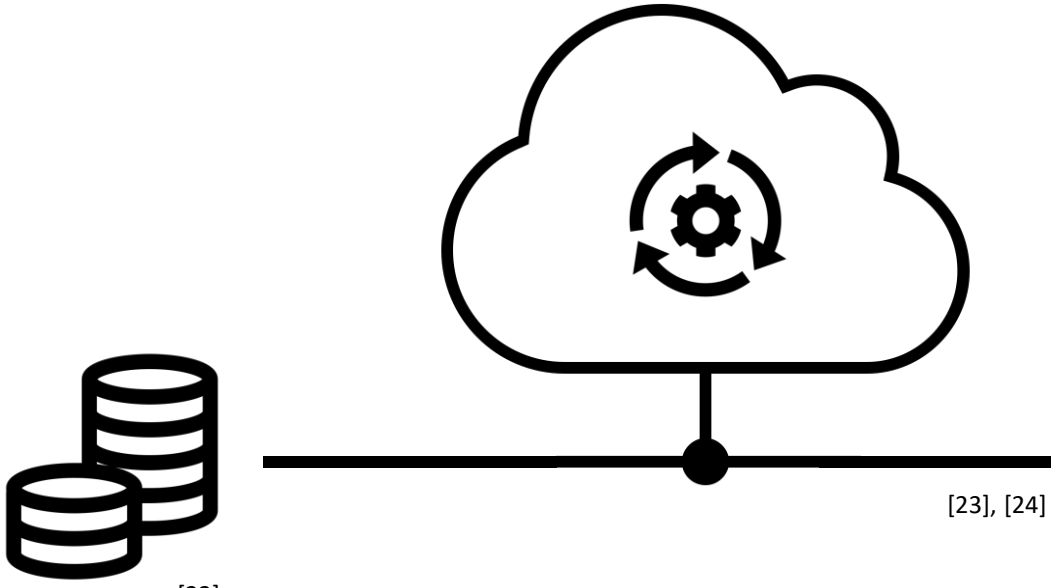
More likely:

[…] Technical University of Munich (Organization), in Germany (Location) .


1
Call for Label
2
Send Image Label in Json
Open Nebula
Flask
web development, one drop at a time
Query Solr
6
Save Json File
5
Call OCR
3
4
Json File
Classification Model
Tesseract OCR
Output in String
Named Entity Recognition (NER)
Solr & Banana
Solr
Capgemini

[1],[2],[3]

Search Module



[22]

[23], [24]

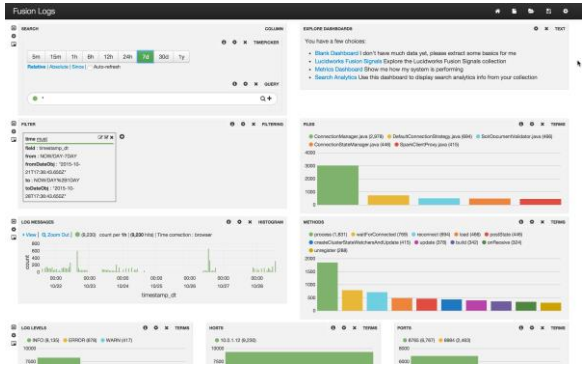
Solr

[3]

[21]

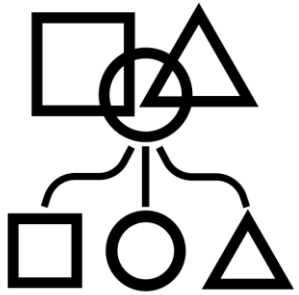
Banana Dashboard

[19]



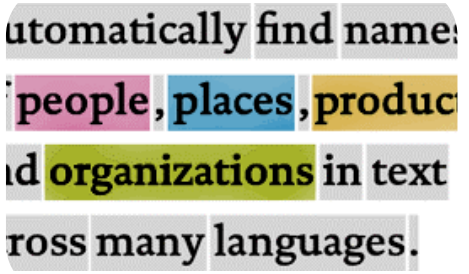
[20]

Goals and Outcomes



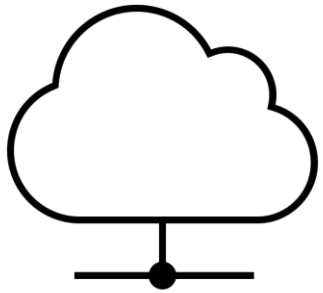
Classification [4]

Optical Character Recognition [5]

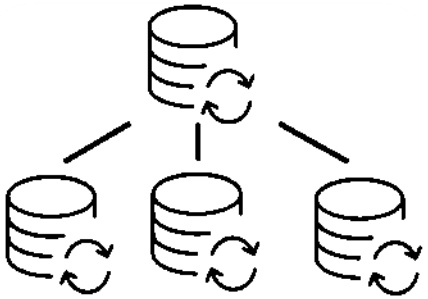


Named Entity Recognition [6]

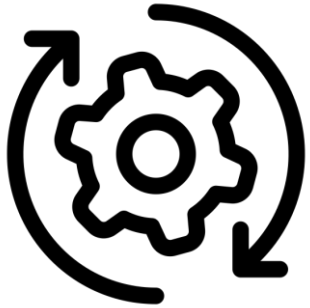
Search Module



OpenNebula [7]

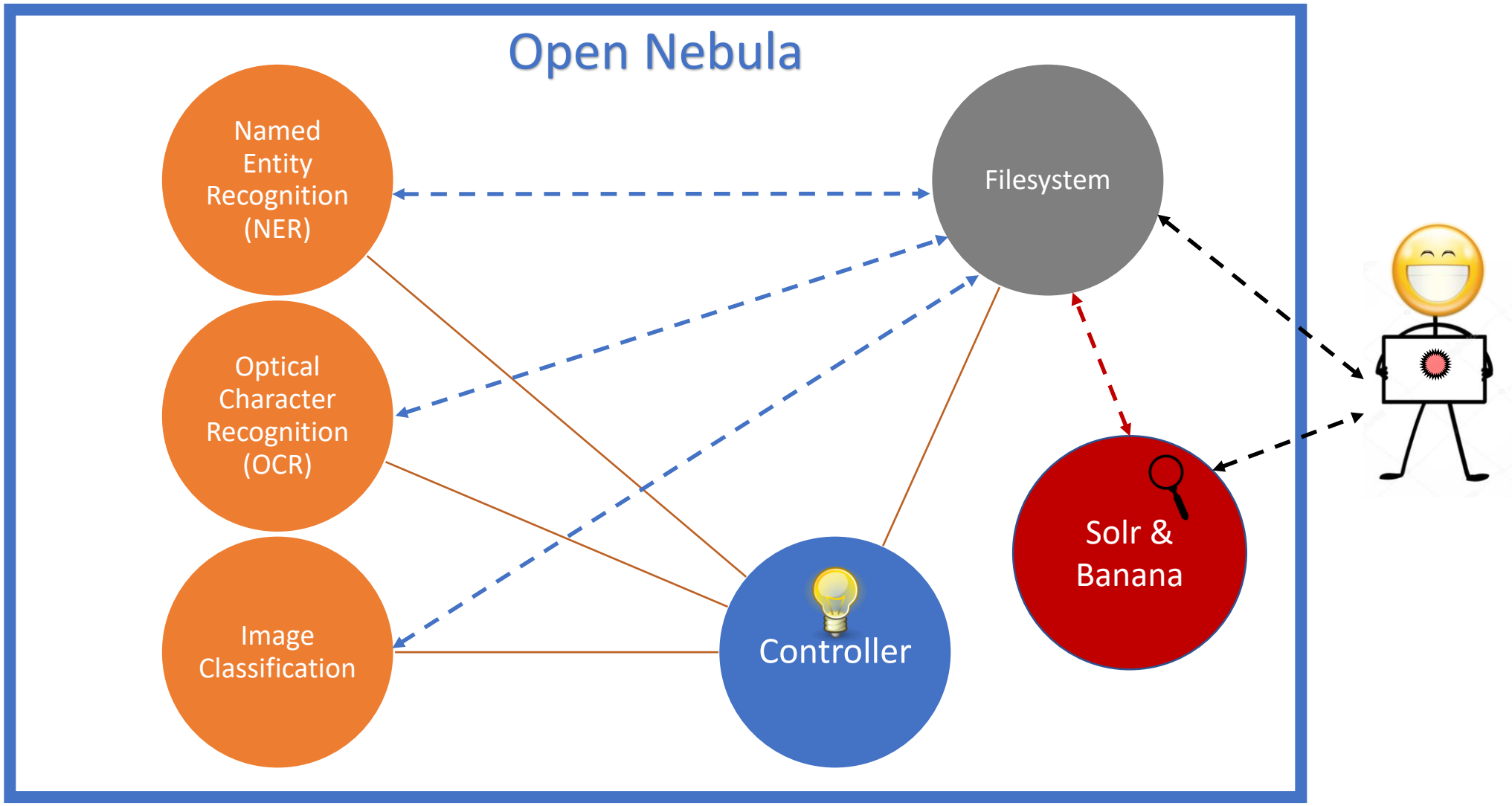


Flask [8]



Continuous Integration – Jenkins [9]

Cloud Structure (Flask)



Thank you for your Attention

References

- Logo OpenNebula, https://opennebula.org/referencing/, 2019-02-08 [1]
- Logo Flask, https://de.wikipedia.org/wiki/Datei:Flask_logo.svg, 2019-02-08 [2]
- Logo Solr, http://lucene.apache.org/solr/logos-and-assets.html, 2019-02-08 [3]
- Classification by cre.ativo mustard from the Noun Project [4]
- Content by Jyoti Vyas from the Noun Project [5]
- Logo NER, https://wordlift.io/blog/en/entity/named-entity-recognition/, 2019-02-13 [6]
- Logo OpenNebula, https://www.v3.co.uk/v3-uk/news/2356529/microsoft-gets-more-open-with-cross-platform-packerio-and-opennebula-cloud-tools, 2019-02-13 [7]
- Logo Flask, https://engineering.bitnami.com/articles/deploy-a-production-ready-mariadb-cluster-on-kubernetes-with-bitnami-and-helm.html, 2019-02-14 [8]
- process by Gregor Cresnar from the Noun Project [9]
- Facebook Endows AI Ethics Institute at German University TUM, Jeremy Kahn, https://www.bloomberg.com, 2019-01-20 [10]
- Graph, https://thenounproject.com/mb.icons/collection/network/?i=1775952, 2019-02-14 [11]
- SVM-SVD, https://thenounproject.com/search/?q=svm&i=1503831, 2019-02-14 [12]
- Network, https://thenounproject.com/smodgekar/collection/technology/?i=1714861, 2019-02-14 [13]
- RVL-CDIP sample1, http://www.cs.cmu.edu/~aharley/rvl-cdip/images/sample1.png, 2019-02-16 [14]
- RVL-CDIP sample2, http://www.cs.cmu.edu/~aharley/rvl-cdip/images/sample2.png, 2019-02-16 [15]
- GloVe, https://nlp.stanford.edu/projects/glove/, 2019-02-16 [16]
- Paperwork, https://de.depositphotos.com/62875115/stock-illustration-vector-paperwork-mood.html, 2019-02-17 [17]
- Banana, https://www.alamy.com/banana-logo-template-vector-icon-illustration-design-image159202591.html, 2019-02-17 [19]
- Dashboard, https://doc.lucidworks.com/fusion/2.0/Dashboards.html, 2019-02-17 [20]
- Magnifier, Search by Mello from the Noun Project, 2019-02-17 [21]
- Data, database by Aiden Icons from the Noun Project, 2019-02-17 [22]
- Cloud-Structure, cloud storage by un·delivered from the Noun Project, 2019-02-17 [23]
- Cloud, implementation by Tomas Knopp from the Noun Project, 20189-02-17 [24]