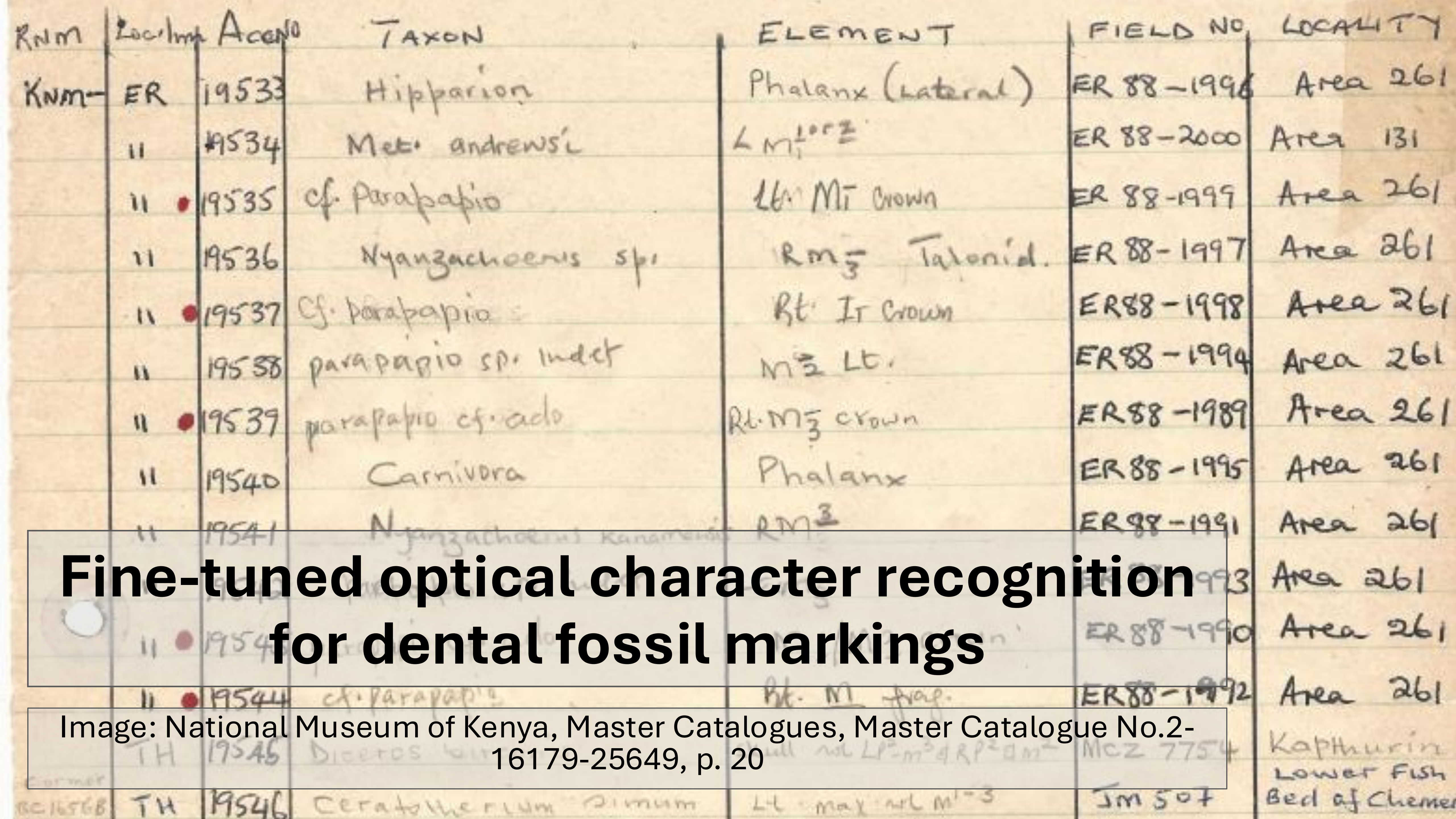
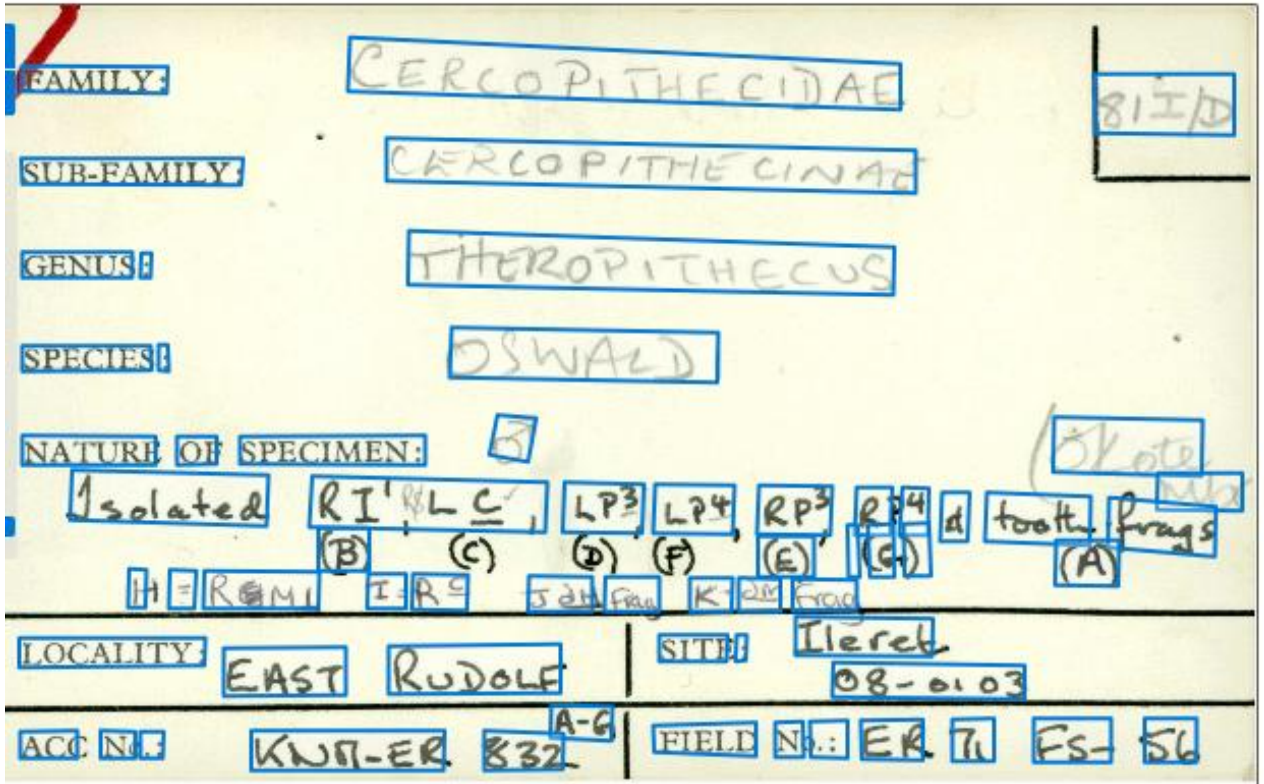


RNM	Loc/Inst	Acc No	TAXON	ELEMENT	FIELD NO	LOCALITY
KNM-	ER	19533	Hipparion	Phalanx (Lateral)	ER 88-1996	Area 261
	"	19534	Met. andrewsi	L $M_1^{1 or 2}$	ER 88-2000	Area 131
	"	19535	cf. Parapapio	Lt. M_1 crown	ER 88-1999	Area 261
	"	19536	Nyanzachoerus sp.	Rm_3 Talonid.	ER 88-1997	Area 261
	"	19537	Cf. parapapio	Rt. I_1 Crown	ER88-1998	Area 261
	"	19538	parapapio sp. indet	M_2 Lt.	ER88-1994	Area 261
	"	19539	parapapio cf. ado	Rt. M_3 crown	ER88-1989	Area 261
	"	19540	Carnivora	Phalanx	ER88-1995	Area 261
	"	19541	Nyanzachoerus kanamensis	RM^3	ER88-1991	Area 261
					[illegible]993	Area 261
	"	1954[illegible]			ER 88-1990	Area 261
	"	19544	cf. parapapio	Rt. M_1 frag.	ER88-1992	Area 261
	TH	19545	Diceros [illegible]	[illegible] LP^2-M^3 & RP^2-M^2	McZ 7754	Kapthurin
						Lower Fish
former BC16568	TH	19546	Ceratotherium simum	Lt. max. [illegible] M^{1-3}	JM 507	Bed of Chemer[illegible]

Fine-tuned optical character recognition for dental fossil markings

Image: National Museum of Kenya, Master Catalogues, Master Catalogue No.2-16179-25649, p. 20



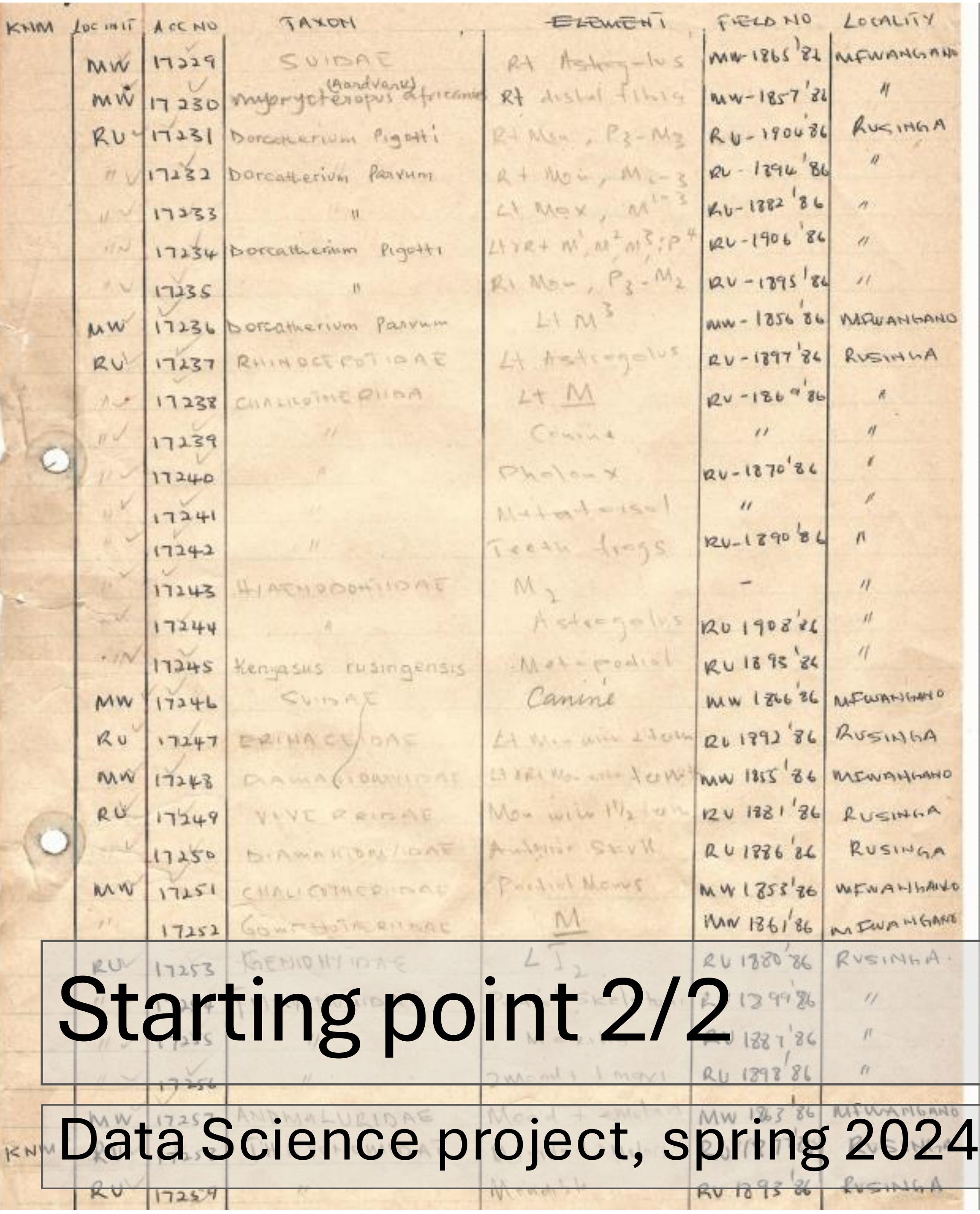
Starting point 1/2

Visualization & JSON sample from Azure Vision API output


```
{
  "text": "Isolated RI'RLC, LP3 LP4 RP3 RP 4",
  "boundingPolygon": [
    {
      "x": 255,
      "y": 1076
    },
    {
      "x": 2290,
      "y": 1115
    },
    {
      "x": 2287,
      "y": 1225
    },
    {
      "x": 253,
      "y": 1195
    }
  ],
  "words": [
    {
      "text": "Isolated",
      "boundingPolygon": [
        {
          "x": 277,
          "y": 1076
```

Starting point 2/2

Data Science project, spring 2024



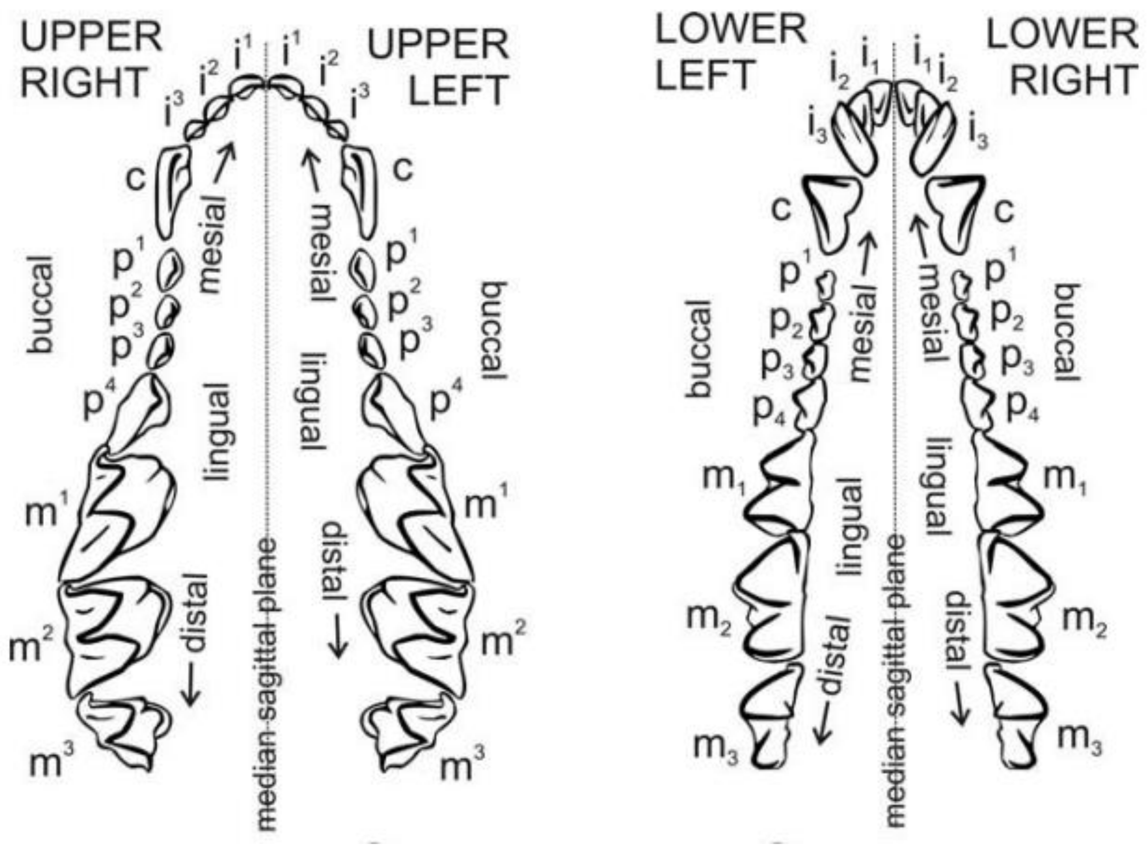
	KNM	ACC_NO	TAXON	ELEMENT	FIELD_NO	LOCALITY	TOOTH_RECORDS
2	MW	17229	Suidae	Rt Astragalus	MW- 1865 87	MFWANG AND	
3	MW	17236	Dorcatherium	L 1 M 3	MW- 1856 86	MEWANGANO	
4	RU	17253	GENIOHY IDAE	li	RU 1880 86	RUSINGA,	
5	11%	17255	GENIOHY IDAE	Maxilla	RU 181 86		
6	MW	17257	ANDMALURIDAE	Mond + 2 moles	MW 1863 86	MEWANGANO	
7	RU	17259			RV 1893 86	RUSINGA	
8	MW	7230	mayoryoteropus africanin	Rt distel fibia	MW - 1857	RUSINGA	
9	RUA	17231	Dorcatherium	R + Mou , P3 - Mz	RU-190486	RUSINGA	P3
10	RU	17237	RHINDET ROTIDAE	Lt Astragalus	RV- 1897 86	RUSINGA	
11	MW	17246	Suidae	Canine	MW 1866 86	MEWANGANO	
12	RU	17247	ERINACEIDAS	At Man with 2 tem	20 1892 86	RUSINGA	
13	MW	17248	DIAMA (TOMMEDAT	LA YRt Man with teamhat	MW 1855 86	MINANGANO	
14	RU	17249	Viverridae	Mon with 11/2	12V 1881 86	RUSINGA	
15	Mw	17251	CHALI COTHERHERE	Partial Marius	MW 1853 86	MFWAWGAIVO	
16	11	17254	IHRYO NOMIDAE	Partial skeleton	RU189986	MFWAWGAIVO	
17	RUI	17258	THEYONOMIDAE	Mexilet a motors	RU 18786	RUSINGA	
18	V	17260	THEYONOMIDAE		RU188386	RUSINGA	
19	. V	17261	THEYONOMIDAE			RUSINGA	
20	. V	17262	THEYONOMIDAE			RUSINGA	
21	. V	17232	Dorcatherium	R + Moi , Mi -	RU - 1894	RUSINGA	
22	. V	17233	Dorcatherium	LI MOX , MIT	KU-1882 86	RUSINGA	
23	. V	17234	Dorcaltherium Pigotti	: p	4 RU -1906 86	RUSINGA	
24	. V	17235	Dorcaltherium Pigotti	Rx mon P3 - M2	RV-1895	86 AND	(P3, M2)
25	. V	17238	CHALICOTHE RIIDA	Lt M	RV-186986	86 AND	
26	. V	17239		confine	RV-186986	86 AND	
27	. V	17240		photom	RV-1870 86	86 AND	
28	. V	17241		Metatarsal	RV-1870 86	86 AND	
29	. V	17242		Teeth frags	RU-1890 86	86 AND	
30	. V	17243	HATHODONTIDAE	M		86 AND	
31	. V	17244		Astragalus	20 190886	86 AND	
32	. V	17245	Kenyasus rusingensis	�� Met - podial	RU 18 93 86		
33	. V	17250	DIAMANTOMVIDAE	anterior SEVN	RU 1886 86	RUSINGA	
34	. V	17263	GOMEZ TO THE RIBAE	M	MIN 1861 86	MIWANGANY	
35	. V	17256	GOMEZ TO THE RIBAE	2 mond + 1 max!	RU 1898 86	MIWANGANY	
36	. V		seromus Pumilus	LM3	5		LM3
37	. V		seromus Pumilus	LM3	5	Bastabula	LM3
38	. V		Paraviacodus flynni	Rm orm +	LLJ129-84	KO 73	
39	. V		Heferocephalus 2p.		LLJ129-84 Knot	(2/73)	
40	. V		Democricetocon equator	RM -.	LLJ129-84 Knot	(2/73)	
41	. V		Aquatic monocots	i wl floral brack		Tugen Hills	
42	. V		Aquatic monocots	Unknown structure	(marked)	Tugen Hills	
43	. V		Plantae	fruit	-	Tugen Hills	
44	. V		, Fern	leaves		Tugen Hills	

Problem satement: find tooth records

Given:
- Catalogue scan images
- Azure AI outputs (generalist OCR* word-by-word readings & word bounding box pixel coordinates)
- Digitized catalogue tables (which words describe the elements)

Find:
Teeth mentioned in 'element' column
- Type (m, p, i or c)
- Index number (1-4)
- Upper / lower
- Left / right

(which tooth on the tooth row)



*Optical Character Recognition

High-level character recognition pipeline

Get words & bounding boxes under the 'Element' header

For each word under element header:

1. Classify each word: tooth or not tooth
   - Re.match('^[a-zA-Z]\d$|^[cC]$'), "letter followed by digit, letter 'c' or letter 'C'"
2. If not tooth: correct reading is the Azure output
3. If tooth: clean word using specialist models
4. custom model 1: word image to M,P, I or C (4-class image classification)
5. custom model 2: word image upper/lower jaw (2-class image classification)
6. custom model 3: index number (C -> 1 (no model), P -> 1,2,3,4, M -> 1,2,3, I -> 1,2,3)
7. Combine custom model results, eg. Upper third molar -> M3

Save output:

Teeth as tuple (eg. (p2, M1, M2) to tooth_records column

Element column value as words on row concatenated

Building a specialist model 1/5: the dataset

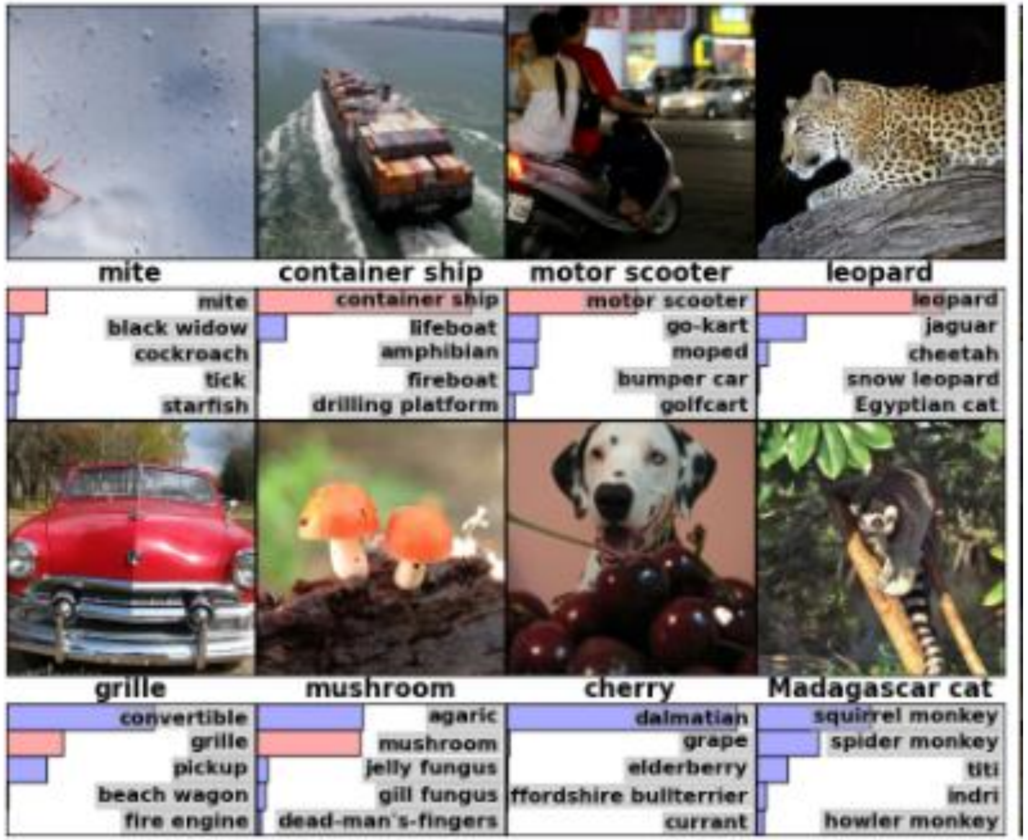
Example: image classification to M, P or I (custom model 1)

	tooth_type	image_i
0	M	0
3	M	3
4	P	4
5	M	5
7	M	7
...	...	...
43	I	43
44	I	44
45	I	45
46	I	46
47	I	47

Building a specialist model 2/5: base models

MNIST & ImageNet classifiers, example: AlexNet

```
AlexNet(
  (features): Sequential(
    (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU(inplace=True)
    (2): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU(inplace=True)
    (5): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU(inplace=True)
    (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (9): ReLU(inplace=True)
    (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU(inplace=True)
    (12): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (avgpool): AdaptiveAvgPool2d(output_size=(6, 6))
  (classifier): Sequential(
    (0): Dropout(p=0.5, inplace=False)
    (1): Linear(in_features=9216, out_features=4096, bias=True)
    (2): ReLU(inplace=True)
    (3): Dropout(p=0.5, inplace=False)
    (4): Linear(in_features=4096, out_features=4096, bias=True)
    (5): ReLU(inplace=True)
    (6): Linear(in_features=4096, out_features=1000, bias=True)
  )
)
```



Building a specialist model 3/5: previous transfer learning research work

Transfer learning (roughly): teaching a base model a new, related task

[1] P. Goel and A. Ganatra, “**A Pre-Trained CNN based framework for Handwritten Gujarati Digit Classification using Transfer Learning Approach**,” in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Jan. 2022, pp. 1655–1658. doi: 10.1109/ICSSIT53264.2022.9716483.

[2] M. Shopon, N. Mohammed, and M. A. Abedin, “**Bangla handwritten digit recognition using autoencoder and deep convolutional neural network**,” in *2016 International Workshop on Computational Intelligence (IWCI)*, Dec. 2016, pp. 64–68. doi: 10.1109/IWCI.2016.7860340.

[3] S. Chatterjee, R. Dutta, D. Ganguly, K. Chatterjee, and S. Roy,"**Bengali Handwritten Character Classification using Transfer Learning on Deep Convolutional Neural Network**". 2020. doi: 10.1007/978-3-030-44689-5_13.

[4] M. Akhlaghi and V. Ghods, “**Farsi handwritten phone number recognition using deep learning**,” *SN Appl. Sci.*, vol. 2, no. 3, p. 408, Feb. 2020, doi: 10.1007/s42452-020-2222-5.

[5] P. Goel and A. Ganatra, “**Handwritten Gujarati Numerals Classification Based on Deep Convolution Neural Networks Using Transfer Learning Scenarios**,” *IEEE Access*, vol. 11, pp. 20202–20215, 2023, doi: 10.1109/ACCESS.2023.3249787.

[6] A. Rasheed, N. Ali, B. Zafar, A. Shabbir, M. Sajid, and M. T. Mahmood, “**Handwritten Urdu Characters and Digits Recognition Using Transfer Learning and Augmentation With AlexNet**,” *IEEE Access*, vol. 10, pp. 102629–102645, 2022, doi: 10.1109/ACCESS.2022.3208959.

[7] K. Limbachiya, A. Sharma, P. Thakkar, and D. Adhyaru, “**Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks,**” *Sādhanā*, vol. 47, no. 2, p. 102, May 2022, doi: 10.1007/s12046-022-01864-9.

[8] N. Thuon, J. Du, and J. Zhang, “**Improving Isolated Glyph Classification Task for Palm Leaf Manuscripts**,” in *Frontiers in Handwriting Recognition*, U. Porwal, A. Fornés, and F. Shafait, Eds., Cham: Springer International Publishing, 2022, pp. 65–79. doi: 10.1007/978-3-031-21648-0_5.

[9] G. Zhao, W. Wang, X. Wang, X. Bao, H. Li, and M. Liu, “**Incremental Recognition of Multi-Style Tibetan Character Based on Transfer Learning,**” *IEEE Access*, vol. 12, pp. 44190–44206, 2024, doi: 10.1109/ACCESS.2024.3381039.

[10] A. F. Rizky, N. Yudistira, and E. Santoso, “**Text recognition on images using pre-trained CNN**,” Feb. 10, 2023, *arXiv*: arXiv:2302.05105. doi: 10.48550/arXiv.2302.05105.

[11] H. Zunair, N. Mohammed, and S. Momen, “**Unconventional Wisdom: A New Transfer Learning Approach Applied to Bengali Numeral Classification**,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sep. 2018, pp. 1–6. doi: 10.1109/ICBSLP.2018.8554435.

Building a specialist model 4/5: Training & evaluation

```
Epoch 26/50
60/60 - 0s - 825us/step - accuracy: 0.6250 - loss: 0.7820 - val_accuracy: 0.7
Epoch 27/50
60/60 - 0s - 7ms/step - accuracy: 0.6660 - loss: 0.7305 - val_accuracy: 0.818
...
Epoch 49/50
60/60 - 0s - 7ms/step - accuracy: 0.7537 - loss: 0.6275 - val_accuracy: 0.854
Epoch 50/50
60/60 - 0s - 844us/step - accuracy: 0.6250 - loss: 0.7502 - val_accuracy: 0.8
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

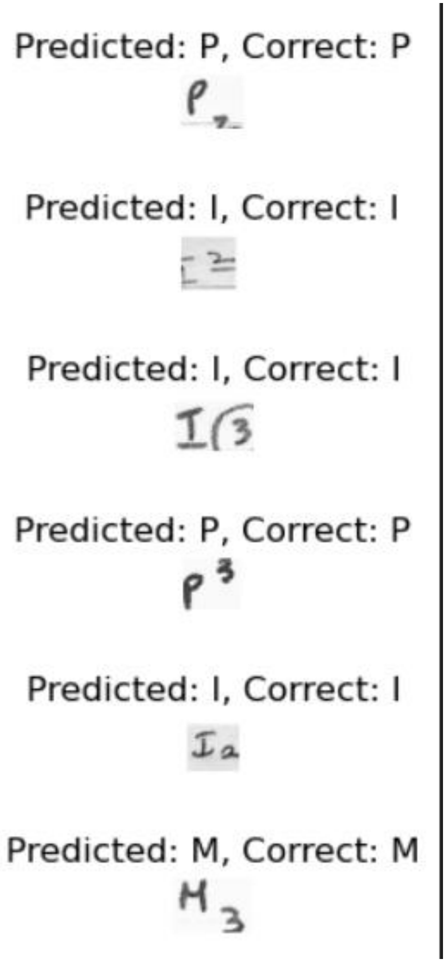
```python
loss, acc = model.evaluate(X_train, Y_train, verbose=1)
acc
```

```
16/16 ━━━━━━━━━━━━━━━━━━━━ 0s 17ms/step - accuracy: 0.9007 - loss: 0.3331

0.8911704421043396
```

+ Code + Mark

```python
test_loss, test_acc = model.evaluate(X_val, Y_val)
test_acc
```

```
2/2 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step - accuracy: 0.8614 - loss: 33.6445

0.8545454740524292
```



Building a specialist model 5/5: MLflow tracking

MPIC

Provide Feedback

Experiment ID: 2 Artifact Location: /home/riikoro/thesis/code/experiments/mlruns/2

> Description Edit

metrics.rmse < 1 and params.model = "tree" | Time created | State: Active | Sort: test accuracy | Columns

Table Chart Evaluation Experimental

	Metrics			Parameters			
Run Name	**Created**	**most frequent cl**	**test accuracy**	**training accurac**	**batch_size**	**data_v**	**layers_trained**	**num_epochs**

Run Name	Created	most frequent cl	test accuracy	training accurac	batch_size	data_v	layers_trained	num_epochs
illustrious-fowl-323	23 hours ago	0.33948339...	0.89090907...	0.90554416...	8	3	1	50
mercurial-jay-61	23 hours ago	0.33948339...	0.85454547...	0.89117044...	8	3	2	50
fortunate-skunk-897	4 days ago	0.33948339...	0.69090908...	0.67967146...	8	3	1	50
shivering-asp-677	4 days ago	0.33948339...	0.69090908...	0.72689938...	8	3	1	50
fun-loon-357	6 days ago	0.60576923...	0.66666668...	0.62032085...	6	3	-	150
masked-worm-324	4 days ago	0.58169934...	0.61290323...	0.80727273...	8	3	last 3	50
receptive-bat-707	4 days ago	0.58169934...	0.61290323...	0.61090910...	6	3	-	150
secretive-whale-618	5 days ago	0.58169934...	0.58064514...	0.61818182...	6	3	-	150
dapper-zebra-263	4 days ago	0.58169934...	0.54838711...	0.96727269...	8	3	last 3	150
abrasive-pug-616	4 days ago	0.58169934...	0.54838711...	0.72363638...	6	3	-	150
secretive-cod-762	4 days ago	0.58169934...	0.54838711...	0.72363638...	6	3	-	150
valuable-goat-156	5 days ago	0.58169934...	0.54838711...	0.62181818...	6	3	-	150
ambitious-mare-667	5 days ago	0.58169934...	0.51612901...	0.90909093...	6	3	-	150
carefree-hen-286	8 days ago	-	0.33333334...	0.3	-	3	-	-
agreeable-bear-49	8 days ago	-	0.33333334...	0.3	-	3	-	-

Summary

- Approach: start out with human-level classification (99,7%+) on simple tasks, generalize from there
- Premise: 97%+ accuracy on small (~1-5k samples) result set is better than 50-70% accuracy on large (~90K-300k samples) set
  - Manual annotation can only be skipped when accuracy is near-perfect
- Questions?
  - Question from me: have you encountered automated data digitization / cleaning for fossil catalogues in previous work?