# Outline of the Visualization

**DECIPHERING INDUS VALLEY SCRIPT** 

## DECIPHERING THE INDUS VALLEY SCRIPT

This Project Is An Attempt To Decipher The Ancient Indus Valley Script By Comparing It To Modern Tamil (A Dravidian Language).



History And
Background

Data
Preprocessing

Statistical
Analysis

Machine
Learning

Github
Repository

Click on buttons to access each section

Can interact with the moving 3D background

#### HISTORY AND BACKGROUND

Go back to main section by clicking this button



This project is an attempt to decipher the ancient Indus Valley script by comparing it with modern Tamil (a Dravidian language).

Our strategy is to: Convert Tamil script from syllabic to logosyllabic by converting morphemes to signs. Perform a statistical analysis to compare the converted script with the Indus script.

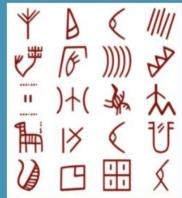
Here are notes on terminology:

Tamil - A Dravidian language

Syllabic script - A writing system whose characters represent syllables.

Logosyllabic script - A writing systems whose characters represent syllables, words and morphemes.

Morpheme - The smallest part of a word that changes its meaning. Ex- 'ing' in 'following'



The Indus Valley or Harappan civilization is India's oldest civilization spanning from 5500 BCE to 1300 BCE. However, their writing system is still undeciphered. The obstacles to deciphering the script have included:

- The exact uses of the artifacts are mostly unknown.
- The root language of the script is unknown.
- The texts are short (mean = 4.5 signs).
- Lack of bilingual inscription/tablet, E.g. Rosetta stone which recorded two Egyptian scripts and ancient Greek
- The complete corpus of the text was not widely available.



Since 2013, however, there have been breakthroughs in the field of population genetics that have given clues about the root language of the script. The 2013 paper `South Asia: Dravidian Linguistic History' authored by Professor Franklin C. Southworth and Dr David W. McAlpin reconstructed Proto Dravidian (one of the two main Indian language families) vocabulary and found similarities with Proto Elamite vocabulary (the language of the Zagros region). Through their archeological and linguistic research, they were able to conclude that the Harappan language is most likely **Proto Dravidian** (with influences from Proto Elamite). Furthermore, it was concluded that this languge has evolved into the major modern Dravidian languages (Tamil, Telugu, Kannada, etc).

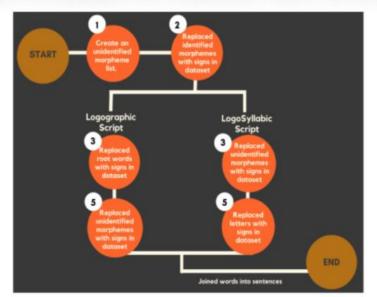
We are testing the hypothesis that Proto Dravidian is the language of the Harappans by comparing modern Tamil with the Harappan script.

First section explains History and Background of the project

#### DATA PREPROCESSING

Using the labelled Tamil database provided by the Institute of Formal and Applied Linguistics (UFAL) See Tagged Dataset. The dataset consists of Tamil news articles tagged by part of speech. Some of the morphemes including clitics and postpositions were alreedy separated. However, the dataset does not separate noun and verb morphemes. Tamil word morphology is exclusive to nouns and verbs.

We split the database into morphemes, clitics, and lemmas and formulated tables stored in our PostgresSQL server.



indes signs		Morphemelleparated text	۵	Morpheme feet	â	id text		Part of Speech fact	•	Sub-Pert of Speech &	Oi N		Tense test	٠	Person &	Number	Gender test	
	-0	(m.l.m.fum.fig)		affing .		-283	1	Verb		[167]	10	41	Present.		Int Person	Pioni-	Houselfic	
	-1	(a/ia/limig)		alling		-211	1	Verb		hall.	ja	15	Present		Ini Person	Pini.	[1/6]	
	2	Countries		Tower		231	1	Vers		(net	[8]	41	Just.		2nd Pinton	Sngriar	(tot)	
	1	pertain)		indext:		-213		Vers		[167]	Ja	41	Present.		3rd Person	Post	Henrythi	
	4	DESCRIPTION		mpault		-283	1	Verb		(n/)	İn	41	Fature		3rd Person	Plant.	Fenetho	
	5	DESCRIPTION		magn		213	1	Yers		(w)	þ	et	Tyrsoles	4	242 Person	PMI.	Honeric	
	6	Security		Convin		-2818	•	Vers		Inel	19	10	Just[		2nd Person.	Plant:	[tut]	
	7	(e.f.ertg)		Rep		-263	1	VWD		[84]	[N	41	Present		3id Fwaces	Flore	Hotorific	
	1	(m/fig)		aRp.		-221		Yes		[68]	[e	4	Present.		310 Person	Steple:	[set]	
	. 9	in in the		#89g		2014		Yes		(Not)	24	et .	hut.		totPorson	PMS	hull.	
	10	(m,i,m,i)(s)		ARID		-741	1	Ven		[mil]	ja.	41	Present		Int Person	Piet-	Hotstife	
	11	(m,1,m,7 <u>2</u> )		49 <sub>2</sub>		-201	ı	Veta		[nd]	Įn	41	(mil)		310 Person	Stepare:	tionattic	
	12	sa-fattige		Bay		-213	ì	Verb :		(net)	Į4	eş.	Present		3-d Person	PMP	Bull	
	13	(mim/s)		ARID		-231	ı	Verb		put.	15	41	2not		2nd Person	Plani	bull	
	14	(m.).m.(lg)		ARD		-211		yes		(nit)	[1	45	[rul]		Talt Person	Stopular	[14]	
	15	0-a-m		64		-214	1	Yero		Liefoli verbal perficie	Įis	et.	(tot)		(ret	[ref	(iut)	
	16	(8.7.8.1)		25		-211	,	ven		Levicar versal particle	24	et	Jwill.		Inuit	(had	[1//]	
	17	8-A-A		CC.		-333	2	Veto		Lexical vertal particle	10	25	hull.		(ref)	(ruit	[tot]	
	18	8-14-79		49		-214		SWID		Auxiliary verbal particle	ĵs	et .	Post		hell	[14]	[147]	
	19	(ALID/)		Bulb		-200	,	Vero		Dud .	1.0	cative	Note:		hut.	buit.	but.	

	Geretren		சென்ன	607		NEN-3SN		Cas=N P	er=3 Num	S   Gen=N			
	alge Ca	அருகே		pp									
	vef	ਲੀ		NEN-3SN-		Cas-N P	er=3 Num	-5 Gen=N					
	பெரும்	usinflia	பெரும்ப	LEST()		NEL-35N		Cas-L P	er=3 Num	-S Gen-N			
	<b>Edia</b>	கரீன		NEN-3SN-		Cas-N P	er-3 Num-	S Gen=N					
		பீல்டு		NEN-3SN-		Cas=N P	er:3 Num	S Genek					
	gefor	நவீனம்											
		விமான	ub da		NO35N-		Per-3 N	um=5 Gen	1-86				
	றிவைய	15040	5/6	நிலைய	nib		NND-35N-		Cas=D P	er-3 Num	-S  Gen=N		
	SHORT.	46		Tg									
13.	offente	dualla	N	MANI- 35H		Casable	Secret Man	Sigenati	18				

index d trigits		Sertence feet
111	0	சென்னை அருகே ஜீ பெரும்புதாரில் நீரின் பீல்டு ( நூரின் ) விமான நிலையத்துக்குக்கதுள் நிலம் யாறுக்கும் பாநிப்பு இல்லாத வகையில் எடுக்கப் படும் என்று முதல்வர் கருணாநிதி உறந்
2	- 1	இது தொடர்பாக , அவர் புதன்கிழமை வெளியிட்ட அறிக்கை
3	2	நாடு முழுவுகள் விமானப் போக்குவரத்தில் ஏற்பப்பு வநம் எனர்ச்சியைக் கரத்தில் கிளையும், புக்கிய நகரங்களில் உள்ள விமான நிலையுக்களை வீரியபடுத்தவடம், புதிகாக உலகத்
4	3	മളങ്ങൾ, പുകളികരി, ഗ്രമ്ബം, Pankaggir, Oraman ചൂടില വിലന്ത് ട്രിനുവല്ക്കേൽ വേലവിള്ള പുളില ട്രീലപ്പടൻ കരുപ്പടക്കാലവിട്ടിന്ന് ഉപ്പെട്ടും വള്ളിക്കുന്നു.
5	4	கர்நாடகத்தில் உம், ஆந்திரத்தில் உம் செரின் மேடு விமான நிலையுக்களை அமைந்த தமிழகத்தை முற்றிக் கொண்டு விட்டனர்.
6	5	ஆனால், இந்த வரிசையில் செள்ளை அருகே அமைக்கப்படகள்ள டுரின் பீல்டு வீமான நிலையத்துக்கு பாடுப்பு புற்படும் வகையில் அடுமுக போதுச் செயலாளர் ஜெயலலிதா , பா.ம.க. நீ
7	6	യള്ളിച ത്യര പുട്ടിപ്പെടെ ട്രെത്ത്യിയുള്ളപത്ത് ട്രോഗ്രേത്യൽ തുപയർ പോയ മടക്തിൽ തുപ്പാലപ്പെ ലൂറിൽവരെത്ത്വെ പുറ്റിർത്രയ അത്യോഗ കണ്ട്യൂന്ന് തന്നെ ഉത്തര്യവായിരുന്നു.
ē .	.7	இவறித்து அவர் வெளியிட்டின் அறிகளை
9	-	பண்டாட்டு அடையாளங்களைப் பாதகாக்க தோட்போருள் ஆய்வுத் தரை உறவாக்கப்பட்டு , தனிச் சுட்டங்கள் இயற்றப்பட்டு உள்ளன
10	9	ஆனால், இந்த அளம்படி உருவாவதற்கு முன்ப அப்பகுடுவில் வரழ்ந்தவர்கள் தான் பாதுகாத்து வந்தனர்.
11	10	கப்படிப் பாதுகாதது மக்களை, அவர்களது வாழ்வி, ங்கள்இலிருந்து அகற்றி, உனறாட்டு அகுடுகளதுக மற்றுதின்ற வகையில், மத்தியில் ஆளும் காக்கிரல்- இ மு.க. டைடனி அரசு புதி
12	11	இச்சட்டம் ஐஞ்து தமிழ் உள்ளிட்ட மாறிய மொழிகளில் உம் மற்றும் அந்தில் தமிழ் அளித்து நாளிதற்களில் உம் முழுப்பக்க அளவில் விளம்பரப்படுத்தி உள்ளது.
12	12	படுப் சட்டத்தின் பத , பாத்தாக்கப்பட்ட நின்னவுச் சின்னத்த 2000 அடி வரை ஏந்த கட்டுமானம் உய கட்ட அதுமதி இல்லை.

Log	osylla	bio	Tamil Sentences SQL Table Snippet
	index bigint		Setterios fait
-1		0	5000 106 5001 5002 2008 5003 3004 ( 5005 ) 5006 5007 2006 5001 2001 155 5008 5009 3003 2001 3006 5010 85 5011 3004 2008 5012 3005 5013 3006 98 5014 5015 5016 2021 47 .
2		1	100 100 100 100 100 100 100 100 100 100
3		2	5024 (1 5006 3009 5025 2008 5026 3010 5027 3006 5028 3004 2000 2025 5029 2008 5090 3011, 5031 5032 3012 2012 2008 107 5006 5007 3012 2012 2000 5033 3013 149, 5034 3007 5035 5006 5007 3012 2012 20
4		3	5041, 5042, 5044, 5000 9045-000 9005-0005-0005-0005-0005-0005-0005
5		4	5051 2021 2008 149, 9552 2021 2008 149 9503 9504 9505 9507 3612 2012 2009 5034 2006 5553 2021 2009 5054 9525 5050 9521 2089 5055 2022 2017 2034.
6		5	5056, 127 505/-3006/2008 5000-108 5006-2025-108-107 5003 5004 5008 5006/2006/2008-5008 5011/-3006/2008 508-5003 508-5003 508-5008 508-7-2008 5027-2024-2023-57
7		6	5087 5088 9054 3025 1008 2019 107 5009 5079 3025 3071 5072 2007 5078 5074 2012 2000 3009 5075 3027 5011 3004 2008 5017 3028 5079 2000 5077 3009 5078 5079 3006 157 5080 5081 5082 5088 5052 2021 47
		7	5884 5000 5002-3004-3002-107 5023
9		8	5085-3010-5086-3012-2012-2000-3009-5087-2025-5069-5070-2021-5088-5049-3017-100, 5089-3023-5071-3012-2012-5090-3009-100-5017-3030_
10		9	5856, 121 5091 5049-3031-2001 5092 141-5093-3004-2008 5094-3032-3021-2050-3024-2012 119 5087-2006 5095-2019-2017-2034
11		10	5096-3009 5087-2021 5072-2000, 5020-2012-2006 5087-2012-12 5096-2037, 5099-3010 5100-2012-154 5101-2035 5011-3006-2008, 5102-3006-2008 5103-3006 5104 - 5105 5106 5038 5024-3007 5107 5008 507
12		11	5071-27_51085109-202251105111-2012-2008-149485112-2021-2008-1495115 5114-2012-2008-1495115 5115-2008-5117-2023-2008-517-2023-2008-517-2023-2008-517-2023-2023-2023-2023-2023-2023-2023-202
13		12	5847 5071-2021-2007-144, \$087-3095-116-\$118-3022-\$119-2021-12-\$120 5121-150 115-\$122-149-\$123-\$125-3040

Second section explains the Data Processing required prior to Machine Learning and Statistical Analysis

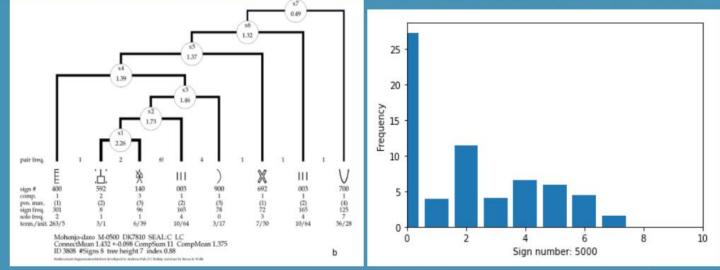
#### STATISTICAL ANALYSIS

Go back to main section by clicking this button



We are performing a statistical analysis in Python on both the Tamil and Indus scripts to evaluate metrics such as sign connectivity, positional frequency and polyvalency. Some methods we will use include multivariate segmentation method and frequency histograms.

We are using positional histograms to find possible initial, medial, and terminal signs on our converted Tamil dataset. We have identified key sign pairs and have calculated their connectivities. Using this, we create our segmentation trees. The higher the segmentation tree, the higher the connectivity between the sign pair. We can then compare this to Indus sign groups. The goal is to identify morphemes with similar distributions in both scripts.



C	reating	Co	nn	ect	ivity	Data	frame			
co	nnectivity_o nnectivity_o nnectivity_o	ff["co								
	Sign pairs	i	j	NPI	NPPI	NPT	NPPT	NPP	NIT	Connectivity
0	[5018', '5019']	5018	5019	0.2	0.000000	0.000000	0.000000	0.5	0.0	1.960000
1	[5019', '3007']	5019	3007	0.0	0.180446	0.000000	0.180446	0.5	1.0	2.639109
2	['3007', '5020']	3007	5020	0.2	0.000000	0.200000	0.000000	0.0	-1.0	0.420000
3	[5020', '5021']	5020	5021	0.0	0.000000	0.000000	0.000000	0.0	0.0	1.500000
4	[5021', '5022']	5021	5022	0.2	0.000000	0.000000	0.000000	0.0	0.0	1.460000
5	['5022', '2022']	5022	2022	0.0	0.056252	0.000000	0.056252	0.0	0.0	1.387495
6	['2022', '5023']	2022	5023	0.2	0.000000	0.134008	0.000000	0.0	-1.0	0.433198

#### MACHINE LEARNING

Go back to main section by clicking this button



We programmed a SVM algorithm to automatically perform morpheme extraction on Tamil. We referenced the paper 'Morpheme Extraction and Lemmatization for Tamil using Machine Learning' which used SVM to perform Tamil morpheme segmentation resulting in a high F score. Morpheme extraction is an essential part of morphological analysis. After preprocessing, the words in our dataset and SQL tables were used for our SVM. We obtained the form, lemma, part of speech identifier, and identified the morphemes within each word. After consulting the paper 'Morpheme Extraction and Lemmatization for Tamil using Machine Learning', we identified the current letter of the word, prefix, vowel/consonant, noun, and verb as vital features for the algorithm.

We programmed to manually fill the features\_df with the appropriate features:

- Identified if the current letter of the word was a vowel (1) or consonant (0)
- Noted if the word was a noun indicated by a 1 and a verb represented by a 1 in their columns
- We extracted the letters before the current letter to fill the prefix column
- Added programming to identify the morpheme boundaries in each word as our output label for the model

Before implementing the model, Label Encoding was used on the letters column and prefix columns to allow for numerical values only. Features selected were the index, vowel/consonant, noun, verb, letter\_label\_encoded, and prefix\_label\_encoded. A Standard Scaler was applied to avoid numerical overflow while running the model. The output label for our model was defined as morpheme boundary.

We used a 75% training and 25% testing data split with a rbf kernel with a gamma of 0.8 to achieve an accuracy of 76%.

Next Steps: Use the ML model on a bigger modern Tamil dataset and on an old Tamil dataset to perform morpheme segmentation to use extracted morphemes in statistical analyses

acc\_score = balanced\_accuracy\_score(y\_test, y\_pred)
(acc\_score)\*100
76.05529650972424

<pre># Create a DataFrame from the confusion matrix. cm_df = pd.DataFrame(     cm, index=["Actual 0", "Actual 1"], columns=["Predicted 0", "Predicted 1"] cm df</pre>	cm =	<pre>splay the confusion matrix confusion_matrix(y_test, y_pred)</pre>
cm, index-["Actual 0", "Actual 1"], columns-["Predicted 0", "Predicted 1"]		
cm df	1	in, index-[ needs v , needs i ], columns-[ recoletes v , recoletes i ]/
The state of the s	cm_d	
		Predicted 0 Predicted 1

# Print the im print(classific				, y_pred))			
	pre	rec	spe	f1	geo	iba	sup
0	0.82	0.81	0.71	0.81	0.76	0.58	9642
1	0.70	0.71	0.81	0.71	0.76	0.57	6067
avg / total	0.77	0.77	0.75	0.77	0.76	0.58	15709

	Predicted 0	Predicted 1
Actual 0	7820	1822
Actual 1	1759	4308

### Future Additions

#### Program the ability to:

- Allow the user to enter in a Tamil dataset
- Take the inputted data and segment it into morphemes
- Replace the morphemes with signs to convert to logosyllabic
- Perform statistical analysis on signs and return results
- Webpage will be published on Heroku