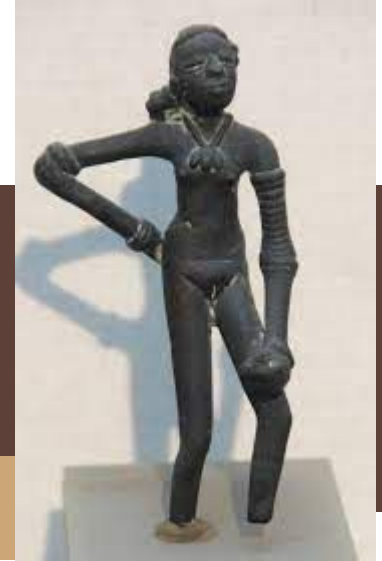


# Indus Valley Girls

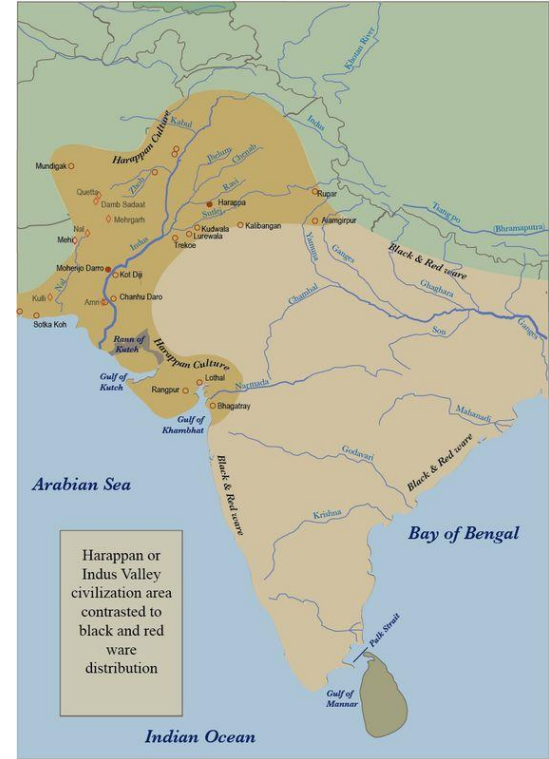
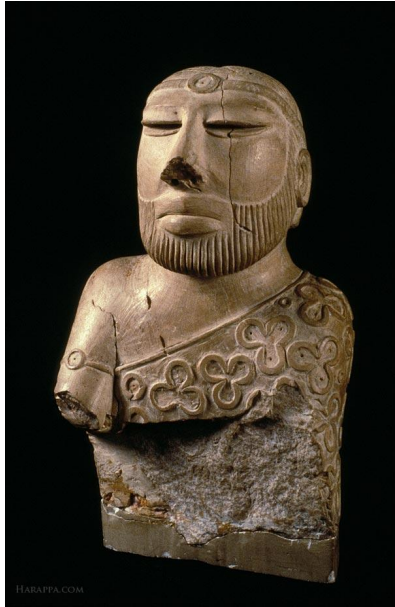
Deciphering the Indus Valley Script

Keerthana Jayakumar

Sonia Sharma



# The Indus Valley/Harappan Civilization



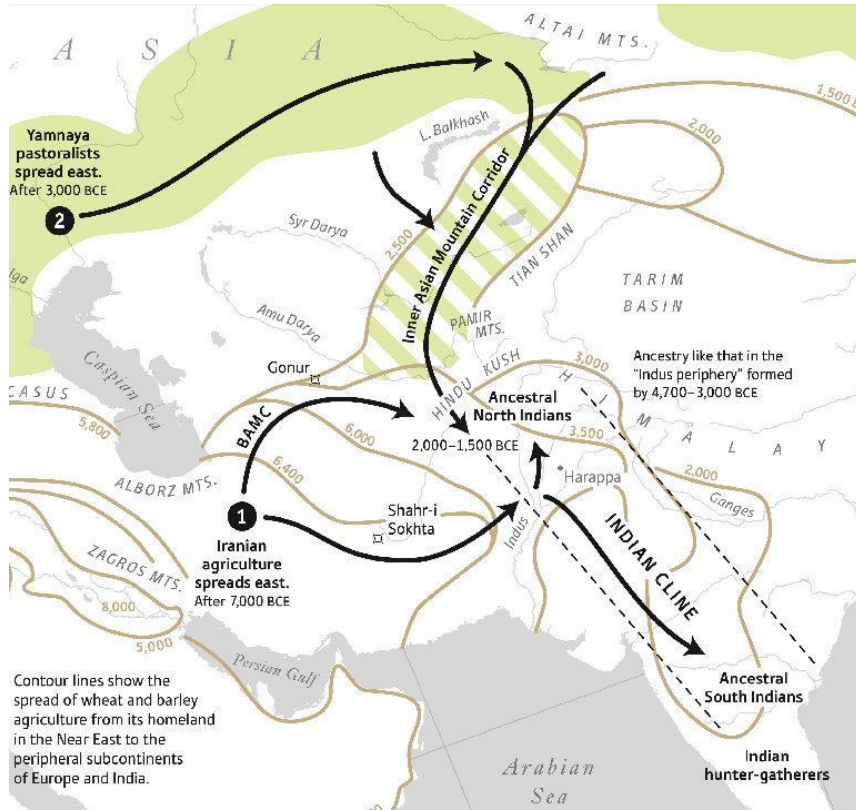
Oldest civilization in the Indian subcontinent

# Why has Indus Valley Script not been Deciphered

- Indus texts are very short.
- Corpus treated as one unit
- Lack of context and usage information
- Lack of bilingual inscriptions
- Debate over root of Indus language



# Clues to Root Language



- Prior to 2013, we didn't know who the Harappans were.
- First Indians and Iranian Agriculturalists were ancestors of the Harappans.
- Spoke Proto Dravidian with influences from Proto Elamite

**STRONG EVIDENCE FOR PROTO-DRAVIDIAN AS THE ROOT LANGUAGE OF THE INDUS SCRIPT**



# Purpose

To decipher ancient Indus Valley (Harappan) script by comparing it statistically with modern Tamil (Dravidian language)



பல்வேறு சமயங்களை உள்ளடக்கிய இந்த நாட்டில் யாருக்கும் பிறரது மதத்தை இழிவுபடுத்தும் அதிகாரம் வழங்கப்படவில்லை.

இந்து சமய மக்களும் இந்த நாட்டிற்கு பல தியாகங்களைப் புரிந்துள்ளனர்.

இந்நிலையில், தேசிய முன்னணியின் தோழமைக் கட்சியான அனைத்துக்கும் தாஜுடின் களங்கத்தை ஏற்படுத்தியுள்ளார் என்பதை மறவாதீர்.

தாஜுடின் அப்துல் ரஹ்மான் நாடாளுமன்றத்தில்

# Data Preprocessing

- Indus script is **Logosyllabic** (signs can be syllables or morphemes ex- 'cat-s' 'play-ing')
- Converted Tamil from Syllabic to Logosyllabic by converting grammatical **morphemes** (smallest unit of a word that changes its meaning) to signs.

## Logosyllabic Script

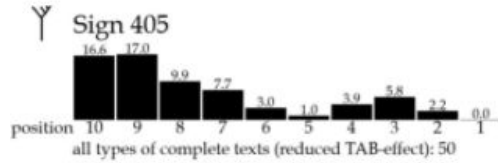
| 001-002 | -003-004   | -005                | -006                     | - 007                                 | -008          | -009 | -010 |
|---------|------------|---------------------|--------------------------|---------------------------------------|---------------|------|------|
| go      | accomplish | word-joining letter | negation<br>(impersonal) | nominalizer<br><i>he/she who does</i> | plural marker | to   | for  |

|                                 |           |           |   |           |                         |    |       |   |   |
|---------------------------------|-----------|-----------|---|-----------|-------------------------|----|-------|---|---|
| Tagged POS<br>Tamil<br>Database | பா.ம.க.   | பா.ம.க.   | N | NEN-3SN-- | Cas=N Per=3 Num=S Gen=N | 23 | Atr   | _ | _ |
|                                 | நிறுவனர்  | நிறுவனர்  | N | NNN-3SH-- | Cas=N Per=3 Num=S Gen=H | 24 | Atr   | _ | _ |
|                                 | ராமதாஸ்   | ராமதாஸ்   | N | NEN-3SH-- | Cas=N Per=3 Num=S Gen=H | 21 | Atr_M | _ | _ |
|                                 | ஆகியோர்   | ஆகியோர்   | N | NNN-3PA-- | Cas=N Per=3 Num=P Gen=A | 28 | Sb    | _ | _ |
|                                 | போராட்ட   | போராட்டம் | J | JJ-----   | _                       | 27 | Atr   | _ | _ |
|                                 | அறிவிப்பை | அறிவிப்பு | N | NNA-3SN-- | Cas=A Per=3 Num=S Gen=N | 28 | Obj   | _ | _ |
|                                 | வெளியிட்ட | வெளியிடு  | V | Vt-T---AA | Ten=T Voi=A Neg=A       | 0  | Pred  | _ | _ |

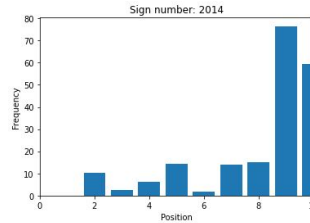
**Pandas** for preprocessing + **PostgreSQL instance on Amazon AWS** for database storage

# Statistical Analysis

## 1. Compared Positional Histograms.

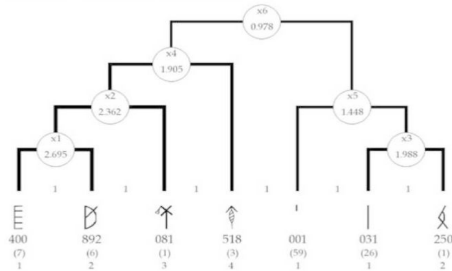


**Indus Script**



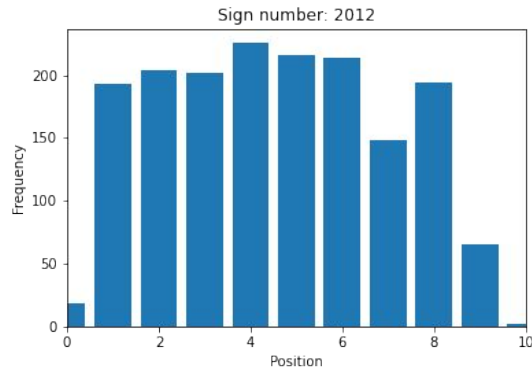
**Tamil**

## 2. Calculated connectivity between sign groups using multivariate segmentation analysis and compared it to Indus sign groups.

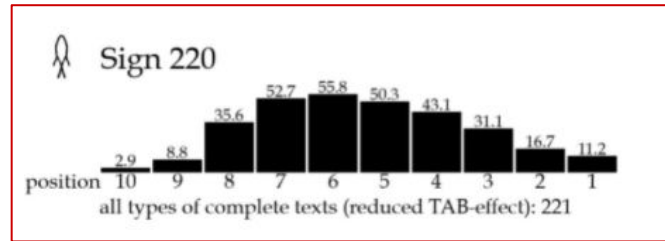


# Results - Possible Plural Marker

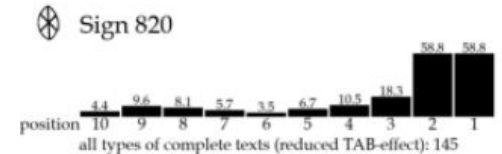
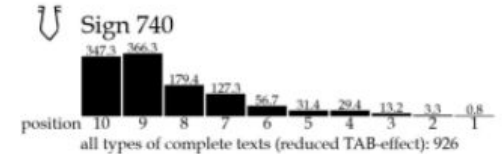
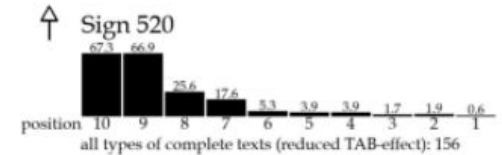
- The most frequent sign in the Tamil script is the plural marker 'கள்' (kal) or sign '2012'
- On analysing the most frequent Indus signs, one sign had a very similar positional distribution



Tamil



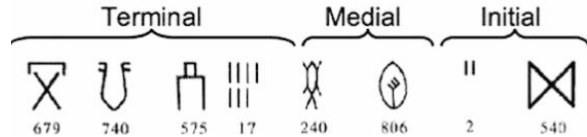
Indus Script
















# Results - Fish Signs and Sign 220

- Fish Signs are often associated with numerals - they were classified as units of measurement (ex- kg) (Bonta 1995)
- Fish signs mostly occur in medial context.



- Sign 220 is unique.** Can be found in initial, medial as well as terminal positions.
- Previously attributed to polyvalence
- Possibility that range is due to its function as plural marker

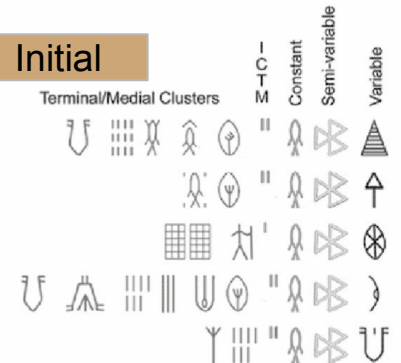
|  | Frequency | Sign Pairs             |
|--|-----------|------------------------|
| 2012 கள்'  | 551       | 147 different pairings |
| 220  | 475       | 52 different pairings  |

|  |  |  |  |  |
|--|--|--|--|--|
| <br>220<br>435 | <br>240<br>331 | <br>231<br>82  | <br>235<br>231 | <br>233<br>182 |
| <br>226<br>36 | <br>241<br>12 | <br>232<br>10 | <br>236<br>19 | <br>234<br>8  |

## Medial

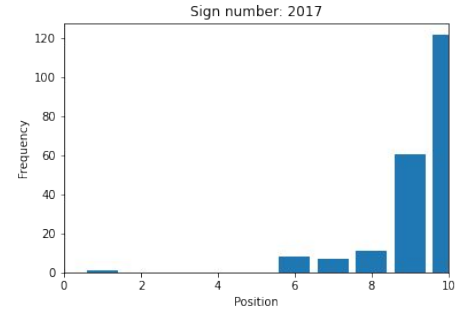
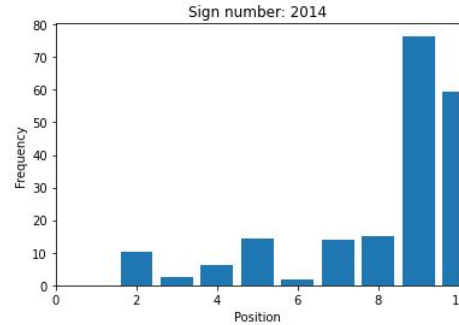
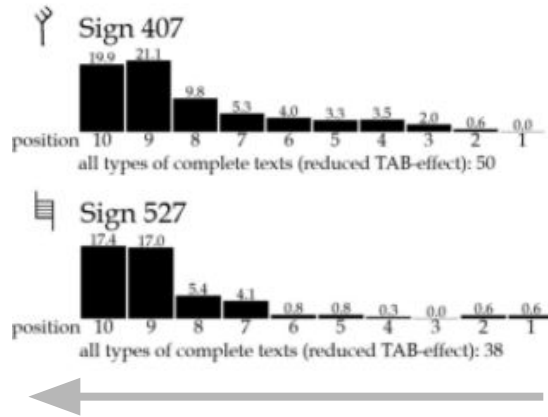


## Initial



# Results - Verb Identification

- Identified verb candidates
- Tamil has a verb terminal syntax
- Can demonstrate that the Indus script is verb terminal (Wells 2007)



- Next steps - Predict verb morphemes by comparing calculated connectivities

# Machine Learning

- Created a SVM algorithm to automatically perform morpheme extraction on modern Tamil
- Extracted features: index, current letter, vowel/consonant, noun/verb, current prefix. Output label: morpheme boundary

## Segmented Morphemes

|      |            |                     |                          |                                |               |      |     |
|------|------------|---------------------|--------------------------|--------------------------------|---------------|------|-----|
| pōka | muṭi       | y                   | āta                      | var                            | kaḷ           | ukku | āka |
| go   | accomplish | word-joining letter | negation<br>(impersonal) | nominalizer<br>he/she who does | plural marker | to   | for |

```
acc_score = balanced_accuracy_score(y_test, y_pred)  
(acc_score)*100
```

76.05529650972424

# Next Steps

- Continue comparison of two scripts.
- Use SVM machine learning algorithm to convert Simpler sentences to LogoSyllabic and run analysis.
- Convert Old (Sangam Tamil, Tamil Brahmi Script) to LogoSyllabic and run analysis.
- Developing a deep neural network to automatically compare similar positional histograms and connectivities
- Create feature in Visualization to:
  - Run the SVM algorithm on a Tamil csv file/text to segment morphemes
  - Automatically convert the segmented script to logosyllabic
  - Automatically run analysis and show results