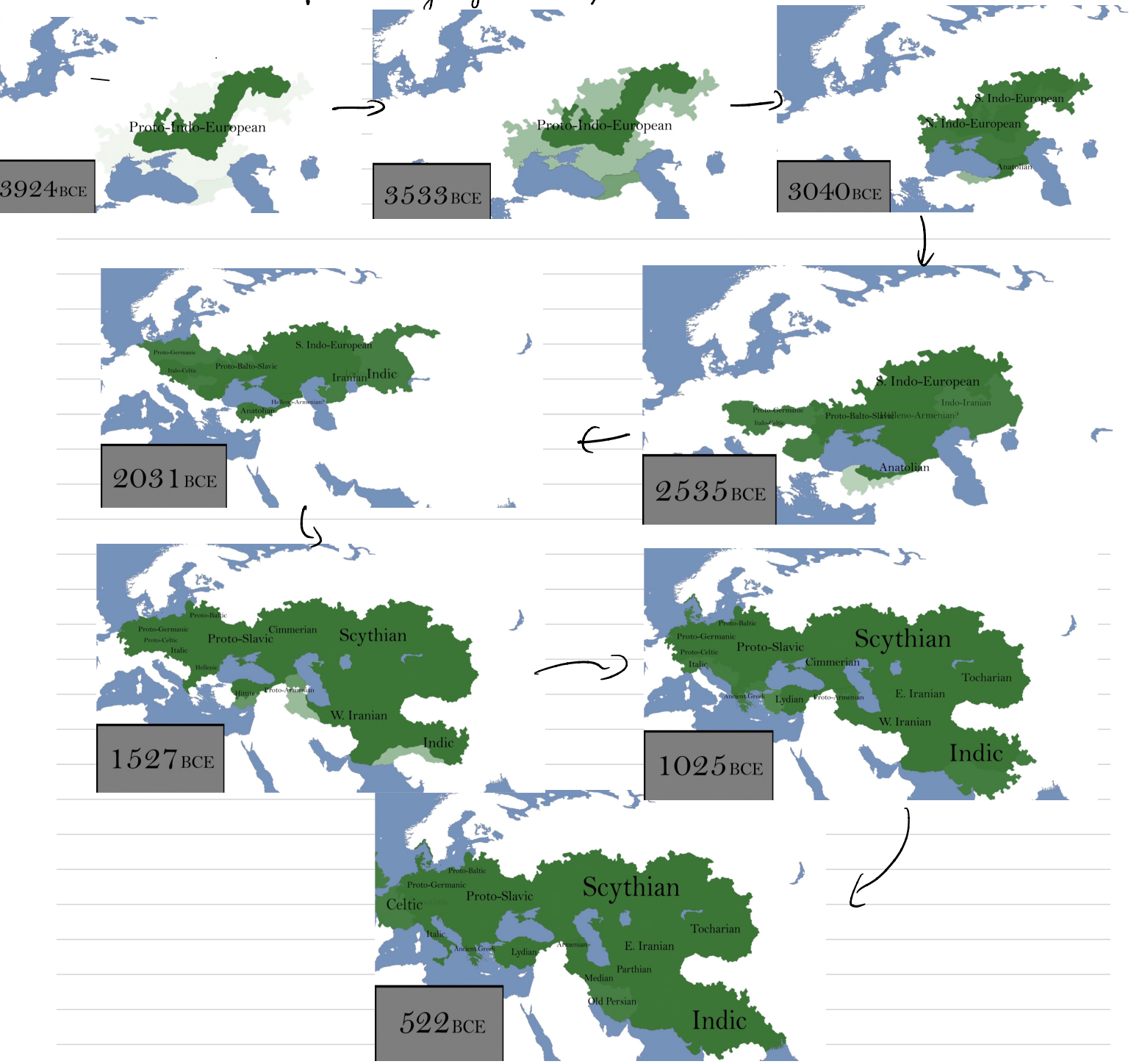
Indo European language Family Analysis

- by Madhunil.

Project Proposal.

All Indo European languages are connected to each other. These language are known by atleast 60% of the entire human population. & this family includes very popular languages such as English, Spanish & Hindi. & they cover a large portion in the Eurasia area. from iceland to North India. Hence studying the common roots of these languages is important. As the more we know about it, the more we know about this common culture. Which everyone has decided to call the Proto Indo European Culture speaking the common but hypothetical Proto Indo European language. (PIE)



PIE was a hypothetical language which had no evidence in written form but this was the answer given to the 235 years old question. which was, Is there a common origin to all Indo European

language? & Is there any explaination for uncanny similarities found between the languages as far apart as sanskrit & icelandic?

So our current research focuses on applying latest available natural language processing techniques to this 200 year old question.

Our prime objectives in this research would be to gain insights on two important things.

1) In which chronological order did the Indo European languages evolved.

2) Based upon our findings of the semantic & syntactic similarities, can the PIE language be reconstructed.

But finding answers to these questions is not a easy task. It requires a lot of research on history, linguistics & machine learning. Apart from the knowledge it will also require efforts on the coding & data collection, cleaning & modeling side. But this being my two year old idea. I have done some research on history, linguistics & data availability, which can provide a head start to us.

So in this project our aim would be to,

i) scrape required data from the internet. It includes text in the given list
[Rigved, Avesta, Iliad, Odyssey, Hesiod, Sappho, other relevant texts]

ii) First get English translated text corpora to find combined topics & also find particular topics for specific texts. for comparision and analysis.

iii) perform word embeddings on the entire text corpus.

iii) Get translated words for the common topics. & check for semantic & syntatic similarities using word embeddings

ii) Decide after above steps are done.

✱ Other data science problems.

There are a total of 10 books in the Rigved. It is said by the linguists that Book 1 & 10 are new books, & Books 2-8 are old books written by 7 Author families. As 2-8 books of the Rigved are known as author books. We can divide the text or hymns in those books into labelled data for each of the author books. Then we can divide that labelled data into training & test data. & perform task of "Author prediction".
This will prove that there is actually some empirical evidence that a machine learning model can actually identify author just based on the way in which the book is written.

For eg: In our 'Rigved' Analysis we have found that Book 1, 9, 10 have. "Raatri" word for "night" where as Book 2-8 have "Nukt" word for "night" which is more similar to its ancient form night, nox, Nychta etc.,
many such patterns can be found by the machice learning algorithm for us, which can help us answer many questions.

Hence 2-8 are also called kul-mandala i.e. family books kul being family.

✱ Data preprocessing steps done till now.

→ removed number from beginning

→ removed stop words.

→ made everything lowercased.

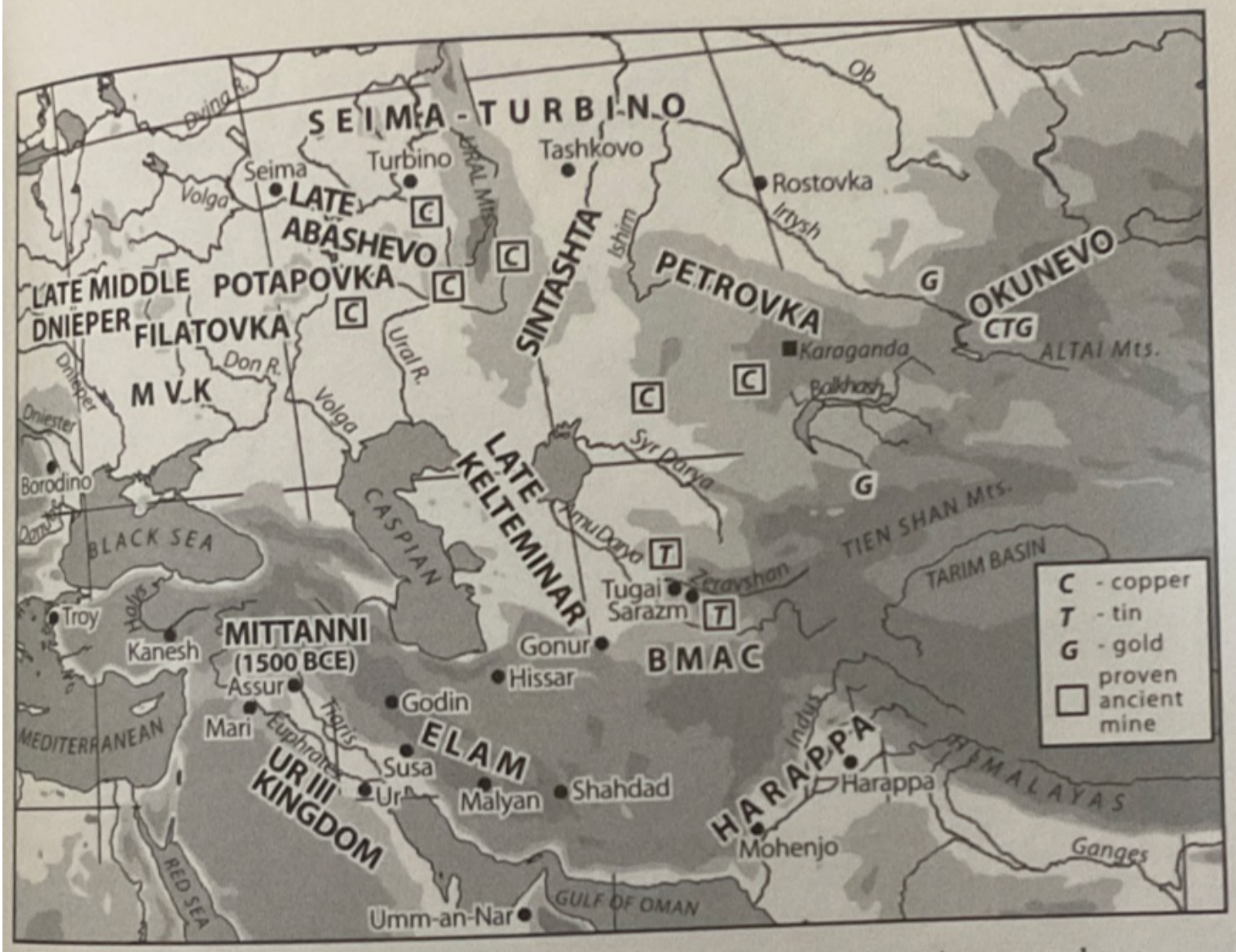
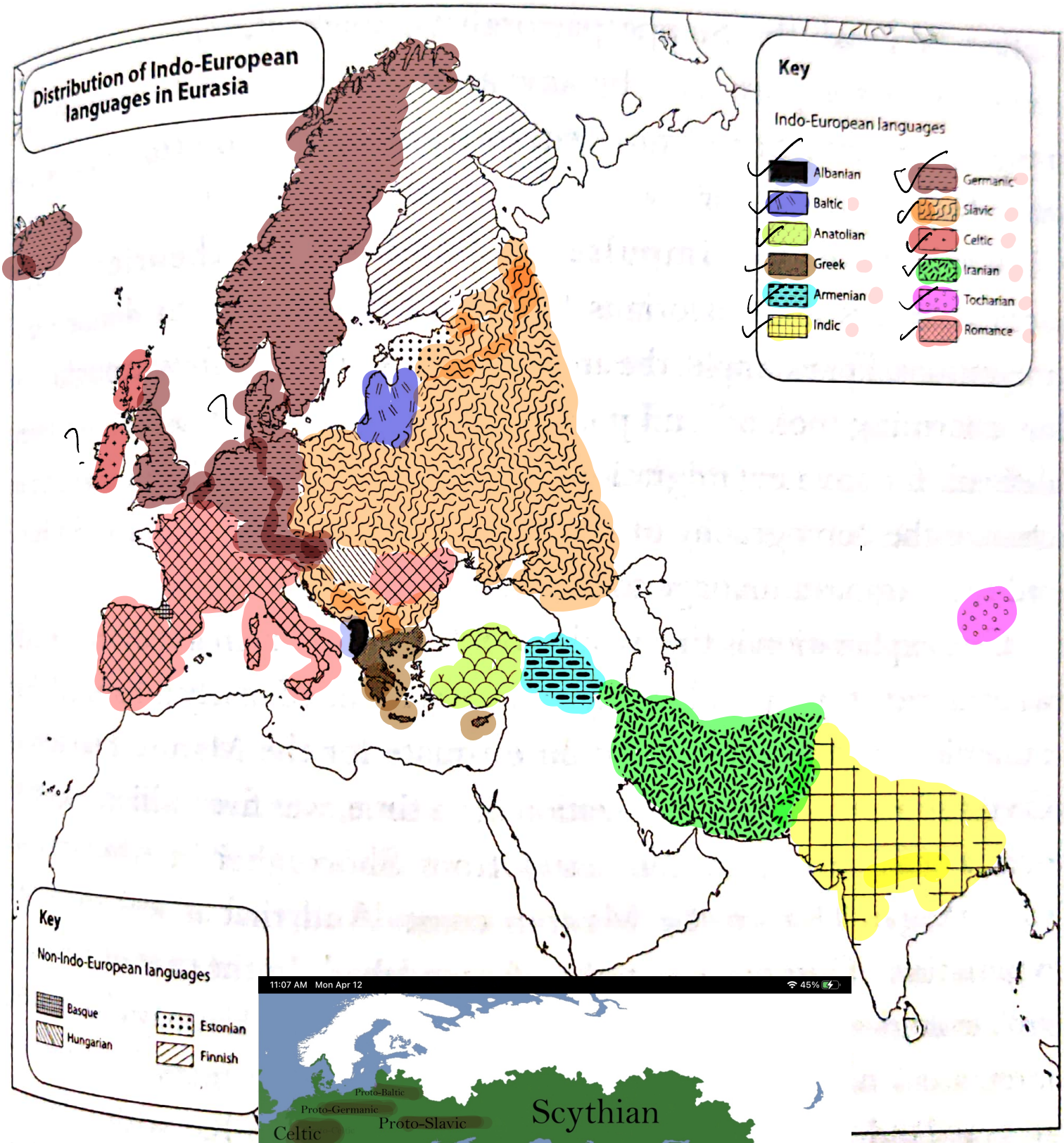
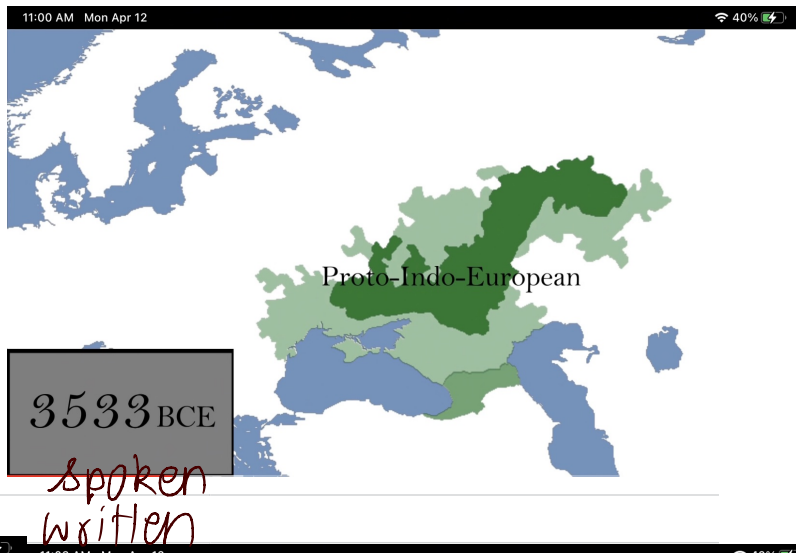


Figure 16.1 Cultures of the steppes and the Asian civilizations between about 2200 and 1800 BCE, with the locations of proven Bronze Age mines in the steppes and the Zeravshan valley.



Spoken
written



Spoken
written

Spoken
written

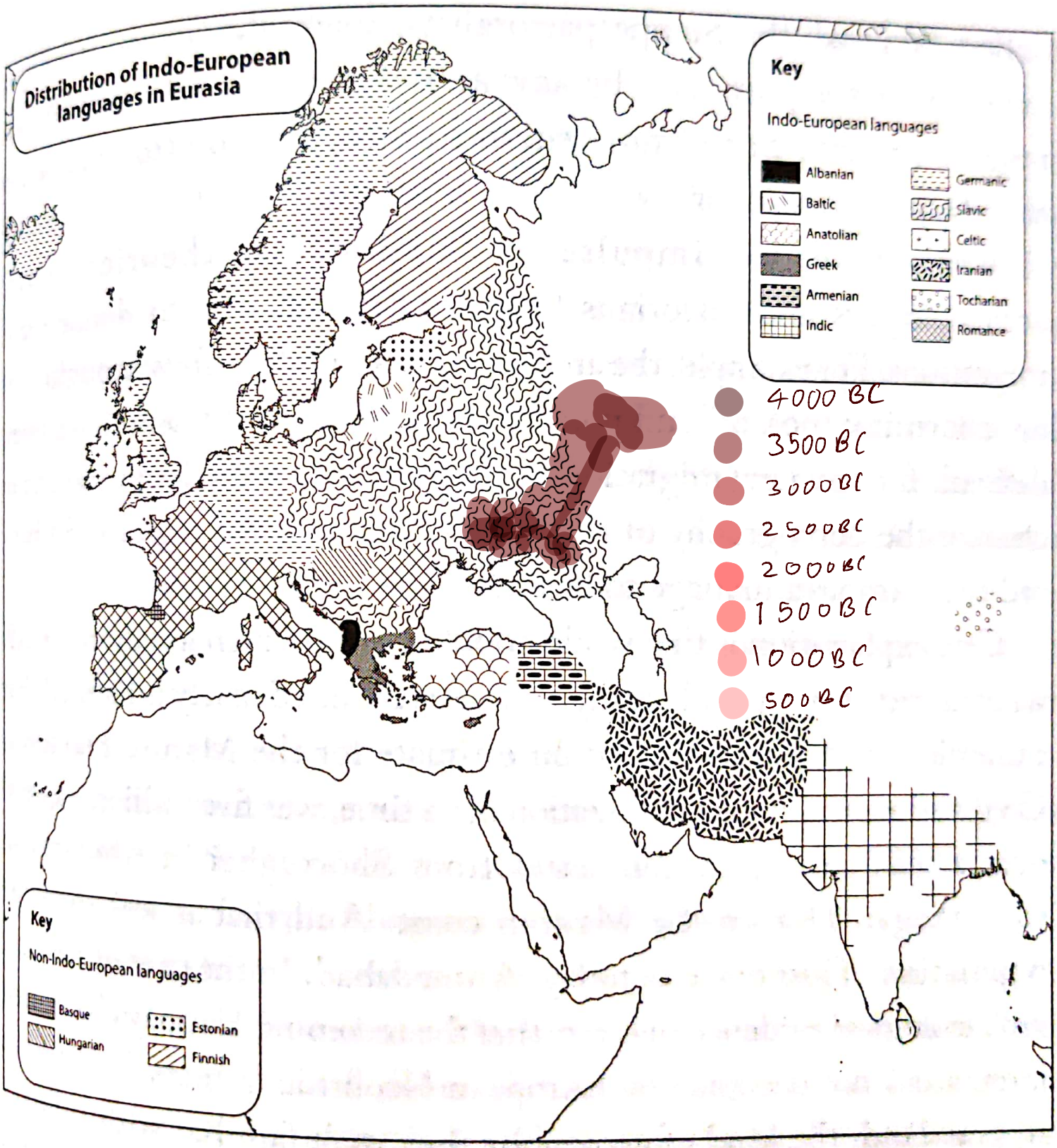
Spoken –
written

Spoken
written

Spoken
written – sanskrit ? did the writing system come to this region yet?

Spoken
written

Spoken ÷
written ÷



ST

Ancient and Modern Indo-European Languages from the 12 branches

- Sanskrit: **asmi, asi, asti.** — Indic — 1700BC(S) 1500 BC (W)
- Avestan: **ahmī, ahī, astī.** — Iranian — (S) — (W)
- Homeric Greek: **eimi, essi, esti.** — Greek — (S) — (W)
- Latin: **sum, es, est.** — Romance — (S) — (W)
- Gothic: **em, ert, est.** — germanic — (S) 300 AD (W)
- Hittite: **ēšmi, ēšši, ēšzi.** — ANATOLIA — 1680 BC – 1178 BC
- Old Irish: **am, at, is.** — Celtic — 70 – 900 AD
- Russian: **esmy, esi, esty.** — Slavic — (S) — (W)
- Lithuanian: **esmi, esi, esti.** — Baltic — (S) — 1500 AD (W)
- Albanian: **jam, je, ishtë.** — Albanian ✗ — (S) — (W)
- Armenian: **em, es, ê.** — Armenian ✗ — (S) — (W)
- Tocharian: **-am, -at, -aṣ.** — Tocharian ✗ — (S) — (W)

The inhabited world can be divided into twelve **major regions**.

Major Regions of the Inhabited World

Credit: Essential Humanities

Linked topics

writing → scripts →

trading → transactions → cities → civilization

COLOUR KEY TO THE WORLD HISTORY TIMELINE

Middle East
South Asia
East Asia
Europe (and colonial offshoots)
Sub-Saharan Africa
pre-colonial Americas
the Steppe

The Ancient World

ca. 3500 BC-500 AD

TIMELINE OF THE ANCIENT WORLD

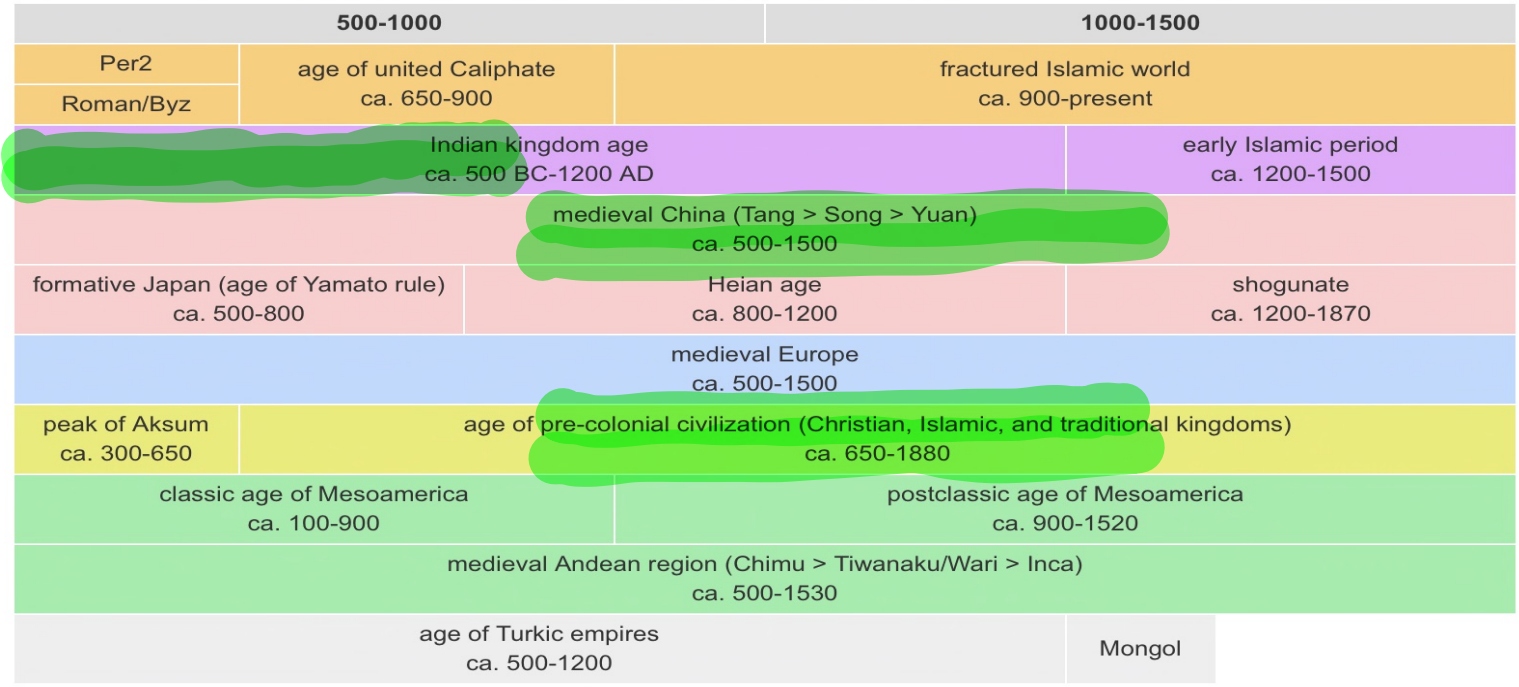
3500-3000 BC	3000-2000 BC	2000-1000 BC	1000 BC-0	0-500
Mesopotamian civilization ca. 3500-550 BC			Per1	inter-Persian / Per2
Egyptian civilization ca. 3000-550 BC				Ptolemaic / Roman/Byz
Indus civilization ca. 2500-1500 BC		Vedic age ca. 1500-500 BC	Indian kingdom age ca. 500 BC-1200 AD	
ancient China (Xia > Shang > Western Zhou > Han) ca. 2000 BC-500 AD				
		Aegean civilization ca. 2000-1200 BC	Greek age ca. 1200 BC-0	
			Roman Republic ca. 500 BC-0	Roman Empire ca. 0-500
		early Nubian civilization ca. 2000-1000 BC	Kush ca. 1000 BC-300 AD	Ak
		formative age of Mesoamerica ca. 1500 BC-100 AD		classic age ca. 100-900
			ancient Andean region (Chavin > Moche/Nazca) ca. 1000 BC-500 AD	
			ancient Steppe empires ca. 1000 BC-500 AD	

from 4500

almost stopped existing after 1900 BC due to climate change except few ultra rich trader communities

Per1	**First Persian Empire**	ca. 550-330 BC
inter-Persian	**inter-Persian period**	ca. 330 BC-200 AD
Per2	**Second Persian Empire**	ca. 200-650
Ptolemaic	**Ptolemaic Egypt**	ca. 330 BC-0
Roman/Byz	**Roman > Byzantine Egypt**	ca. 0-650
Ak	**peak of Aksum**	ca. 300-650

TIMELINE OF THE MEDIEVAL WORLD

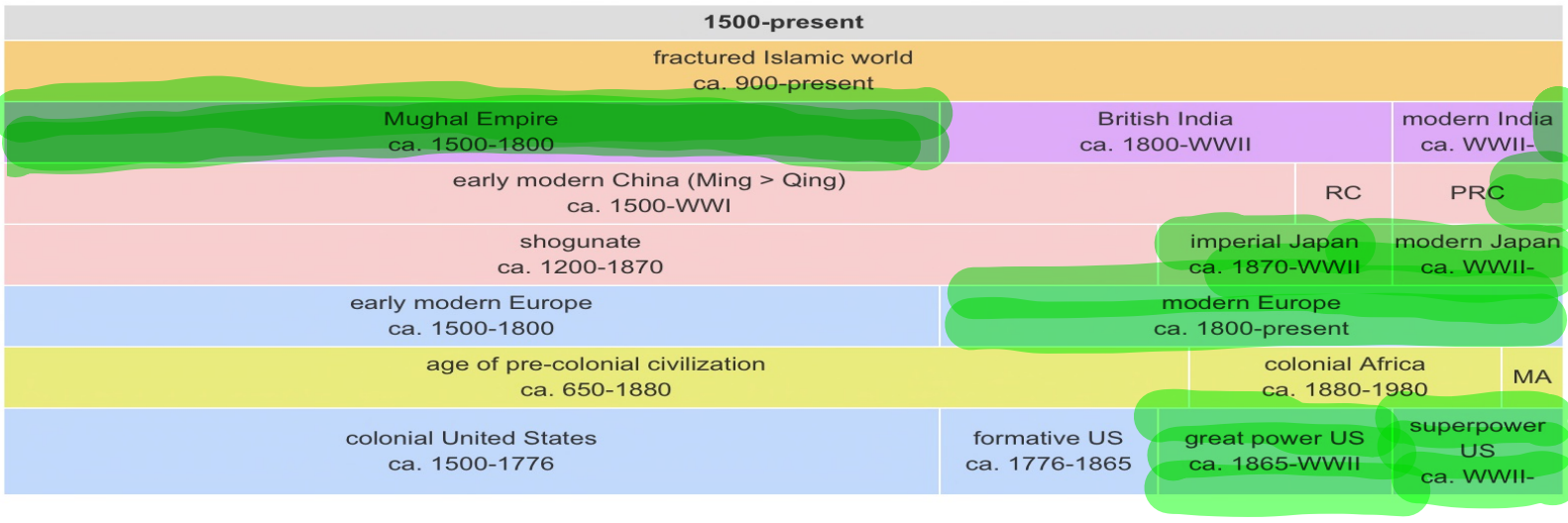


Per2	**Second Persian Empire**	ca. 200-650
Roman/Byz	**Roman > Byzantine Egypt**	ca. 0-650
Mongol	**Mongol Empire**	ca. 1200-1300

NOTES

the primary power of medieval Eastern Europe was the **Byzantine Empire**
the primary powers of medieval Western Europe were **France**, **England**, and the **Holy Roman Empire**

TIMELINE OF THE MODERN WORLD



RC	**Republic of China**	ca. interwar period
PRC	**People's Republic of China**	ca. WWII-present
MA	**modern Africa**	ca. 1980-present

NOTES

the primary powers of Reformation Europe (ca. 1500-1650) were **Spain**, **France**, and **Austria**
the primary powers of Europe from the Enlightenment to WWI (ca. 1650-WWI) were **France**, **Britain**, **Austria**, **Prussia** (later Germany), and **Russia**
the primary powers of Europe since WWI (ca. WWI-present) have been **France**, **Britain**, **Germany**, and **Russia**
the two superpowers of the Cold War (ca. 1945-91) were the **United States** and **USSR**
since the Cold War, the **United States** has reigned as the world's only superpower;

Golden ages in the past.

1st	Han dynasty	206 to 220	for	14 years
2nd	Golden age of India	300 to 600	for	300 years.
3rd	Tang dynasty	618 to 907	for	289 years
4th	Ming dynasty	618 to 1279	for	661 years
5th	Islamic Golden age	622 to 1258	for	636 years
6th	Renessance			
7th	America & western world			

The origins. [still a hypothesis]

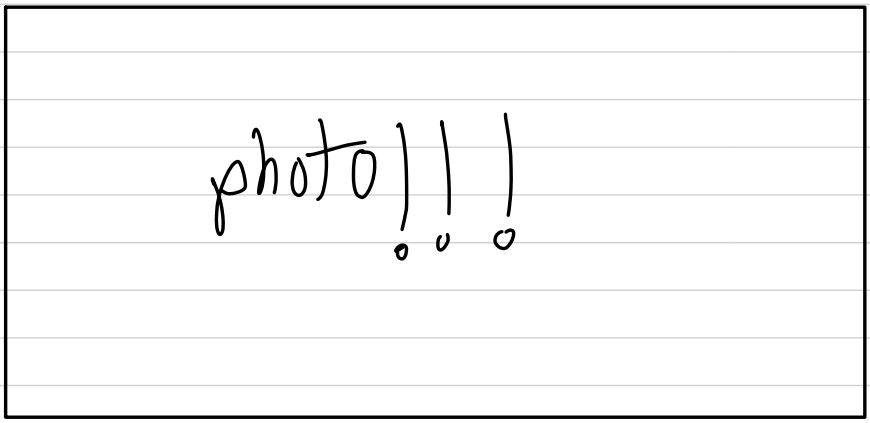
Sintashta culture. (2100 – 1800 BCE).

→ last rites exactly as mentioned in the Rigvedic family books.

→ could be the original aryavarta.

→ All current domesticated horses in the entire world share 98% of their ansestry with the sintashta horse.

→ had fortified settlements, which Indra apparently destroyed.



→ Old Indic which was used in sintash culture was also used Mittani people of North syria (1500 BC).

→ Common Indo Iranic tongue was spoken in the sintashta culture (proof?).

→ Rigvedic later books (1, 9, 10) include burial info of the sintashta elite. & also the horse sacrifice.

→ Rigvedic family books (2-8) include sacrificial feast info which used to take place to feed 100s during the funeral of sintashta elite.

→ dog sacrific also mentioned in the rigved but not performed in Sintashta. it was done in the Srubnaya culture.
↑
??. (volga steppes).

→ around 1500 BC Mittani dominance was seen near Hittites, both were indo-european tribes, but mittanis worshipper rigvedic gods Indra, Varunna, which were called during a treaty.

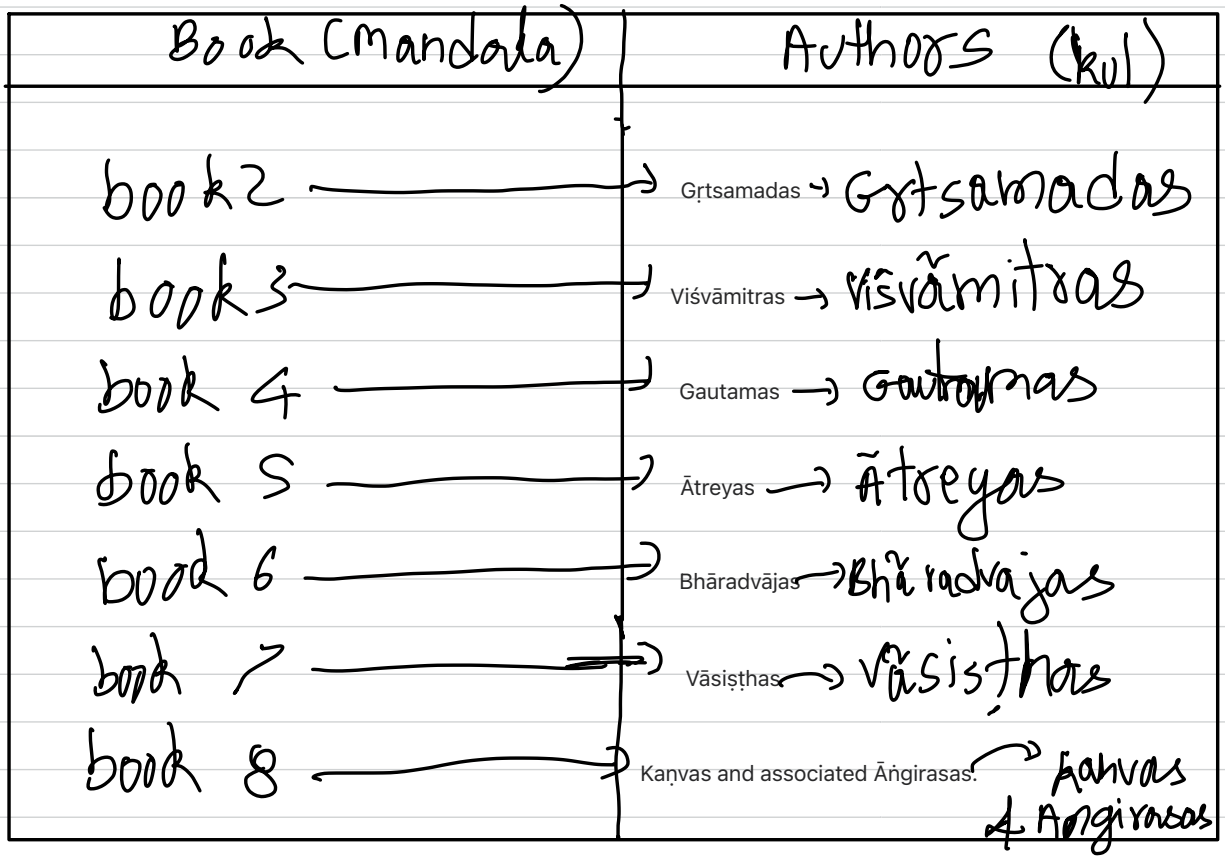
→ also around 1500 BC Zingust say the rigved was compiled. after which we see massive collaborations between the Bharat-Puru tribe & the rest of the nine Aryan & non Aryan tribes. this collaboration is what ultimately formed the Kuru state around 1300BC.

→ Some call this collaboration a massive success to save all tribal cultures, including those of the much advance Indus valley civilization. It is also compared by some to the treat of west phalia in Europe around 1500CE, which ultimately preserved many christian & European cultures in process bringing the renaissance & then leading to exponential scientific development.

→ But exact parallels can be drawn to this & the acient hindu vedic society, which also brought cultural renaissance to all the tribes linked to the Kuru state & in parallel promoting the upanishadic thought. which then lead to massive development in the Hindu society, ultimatey recognized as a golden age by the western world during around 300 AD. but like it was already a state since 1300BC & a civilization since 4000 BC.

✱ More obeservations to continue based on further research on the topic. hopefully making some advances with the use of Natural Laguage processing techniques.

✱ Treating RigVed as a Author Prediction problem.



Next steps to take

✱ label 90% of the data & start training a LLM. on the labelled data.

✱ Test the LLM on remaining 10% of unlabelled data.

✱ produce accuracy of the language model in 1) GLUE bench mark score
2) cross entropy / perplexity
3) BLUE score

4) Grammar & readability

5) must generate language in a readable format.

6) we should access factors such as language fluency, coherence, contextual understanding, factual accuracy & ability to generate relevant & meaningful responses.

Reconstructing PIE (π)

- Scrape all ancient books in latin script
- Combine all to make a single dataset
- train LLM on that "dataset"
- produce what will be PIE in latin script.

	ancient books available		date
Sanskrit	Rig ved		2100 - 1500 BCE
Avestan/Persi	Avesta		600 - 500 BCE
Greek	Illiad		700 - 600 BCE
Greek	Odessey		700 - 600 BCE
Latin