

# Predicting the past with deep neural networks

*Yannis Assael\*, Thea Sommerschildt\*, Brendan Shillingford, Mahyar Bordbar,  
John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos,  
Jonathan Prag, Nando de Freitas*



Università  
Ca' Foscari  
Venezia



UNIVERSITY OF  
OXFORD



ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS



Google Cloud



Google Arts & Culture

ὄλβιος ὅστις τῆς ἱστορίας ἔσχε μάθησιν  
*Happy the man who has gained knowledge through history*

— Euripides (c. 480–406 BCE), Greek tragedian  
*Antiope*, uncertain fragment

# ancient history

## The setting

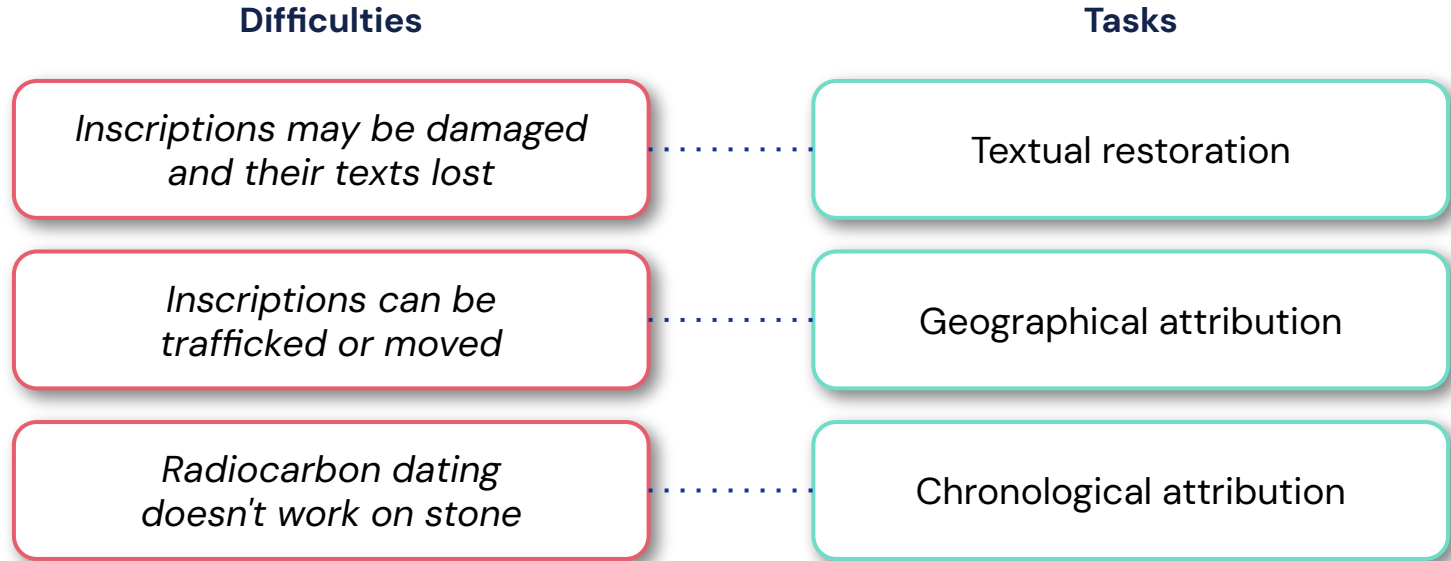
- Ancient History relies on disciplines such as Epigraphy, the study of inscribed texts, for evidence of the recorded past.
- These texts are known as "inscriptions".
- They offer firsthand evidence for the thought, language, society and history of ancient civilisations.
- Thousands of inscriptions have survived to our time. However, their study is far from straightforward.



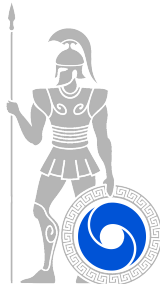
Damaged decree concerning the Acropolis of Athens (485/4 BCE).

IG I<sup>3</sup> 4B. (CC BY-SA 3.0, Wikimedia)

# the historian's workflow



***These are time-consuming and highly complex tasks;  
traditional methods are inefficient.***



# AI for ancient history



Restored decree concerning the Acropolis of Athens (485/4 BCE).

IG I<sup>3</sup> 4B. (CC BY-SA 3.0, Wikimedia)



dataset

# I.PHI dataset

- Based on the Packard Humanities Institute's database.
- Filter human annotations, render the text and metadata machine-actionable.
- First multi-task dataset for Epigraphy with 70k inscriptions.
- Consists of tuples of:
  - Corrupted text
  - Geographical metadata (*ancient region*)
  - Chronological metadata (*date range*)

<b>Split</b>	<b>Inscriptions</b>	<b>Characters</b>	<b>Words</b>
Training	63,014	19,559k	2,915k
Validation	7,783	2,503k	373k
Test	7,811	2,416k	360k

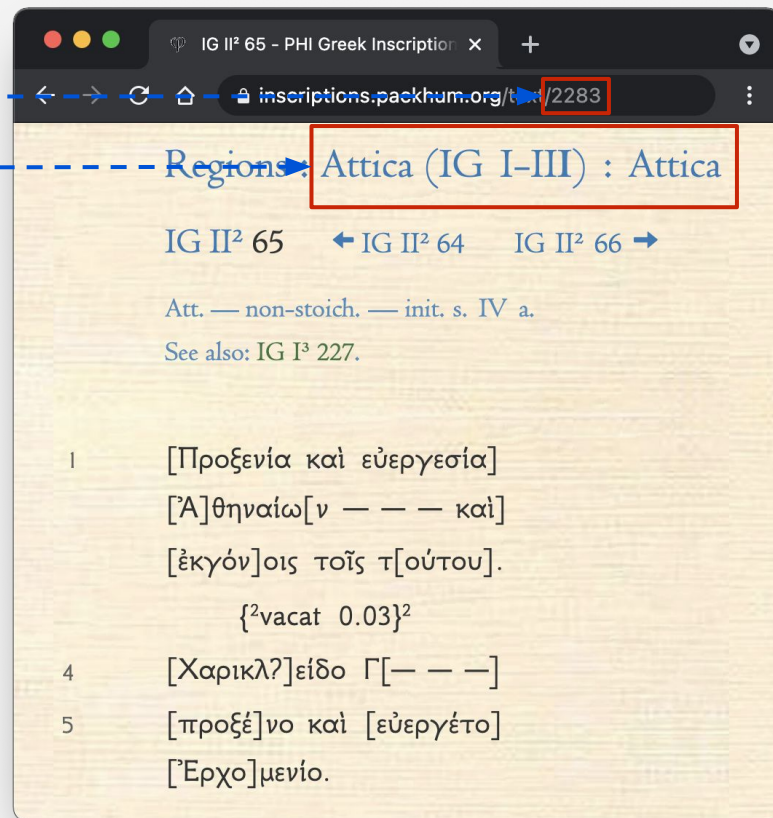
Statistics for the I.PHI corpus.

# data generation

Inscription ID

Geographical metadata

- 84 regions
- Based on IG



IG II<sup>2</sup> 65 — PHI Greek Inscription

inscriptions.packhum.org/text/2283

Region: Attica (IG I-III) : Attica

IG II<sup>2</sup> 65 ← IG II<sup>2</sup> 64 IG II<sup>2</sup> 66 →

Att. — non-stoich. — init. s. IV a.  
See also: IG I<sup>3</sup> 227.

1 [Προξενία καὶ εὐεργεσία]  
[Ἀ]θηναίων — — — καὶ  
[ἐκγόν]οις τοῖς τ[ούτου].  
{<sup>2</sup>vacat 0.03}<sup>2</sup>

4 [Χαρικλ?]εῖδο Γ[— — —]

5 [προξέ]νο καὶ [εὐεργέ]το  
[Ἔρχο]μενίο.

<https://inscriptions.packhum.org/text/2283>



# data generation


## Inscription ID

## Geographical metadata

- 84 regions

## Chronological metadata

- referring to historical eras & intervals in several languages
- lacking in standardized notation ("early", "first half", "1st half", "beginning", "beg.")
- fuzzy wording ("late 7th/6th c. BC", "ca. 100 a.?", "bef. 64 AD")



IG II² 65 - PHI Greek Inscription

inscriptions.packhum.org/text/2283

Region: Attica (IG I-III) : Attica

IG II² 65 ← IG II² 64 IG II² 66 →

Att. — non-stoic — init. s. IV a. ← 400-350 BCE

See also: IG I³ 227.

1 [Προξενία καὶ εὐεργεσία]  
[Ἀ]θηναίω[ν — — καὶ]  
[ἐκγόν]οις τοῖς τ[ούτου].  
{²vacat 0.03}²

4 [Χαρικλ?]εῖδο Γ[— — —]

5 [προξέ]νο καὶ [εὐεργέτο]  
[Ἔρχο]μενίο.

<https://inscriptions.packhum.org/text/2283>

# data generation

## Inscription ID

## Geographical metadata

- 84 regions

## Chronological metadata

- referring to historical eras & intervals in several languages
- lacking in standardized notation ("early", "first half", "1st half", "beginning", "beg.")
- fuzzy wording ("late 7th/6th c. BC", "ca. 100 a.?", "bef. 64 AD")

## Ancient Greek text

- noisy, non-standard human annotations / markup
- epigraphers annotate missing characters with "-"

IG II² 65 - PHI Greek Inscription

inscriptions.packhum.org/text/2283

Region: Attica (IG I-III) : Attica

IG II² 65 ← IG II² 64 IG II² 66 →

Att. — non-stoic — init. s. IV a. 400-350 BCE

See also: IG I³ 227.

1 [Προξενία καὶ εὐεργεσία]  
[Α]θηναίων — — — καὶ  
[ἐκγόν]οις τοῖς τ[οῦτου].  
{²vacat 0.03}²

4 [Χαρικλ?]εῖδο Γ[— — —]

5 [προξέ]νο καὶ [εὐεργέ]το  
[Ἔρχο]μενίο.

προξενια και ευεργεσια  
αθηναίων---και  
εκγονοις τοις  
τουτου. χαρικλ-  
ειδο γ---  
προξενο και  
ευεργετο  
ερχομενιο.

<https://inscriptions.packhum.org/text/2283>

# modern datasets

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

500,000,000,000 tokens =

400,000,000,000 words =

1,000,000,000 A4 pages =

60 km thick book



## modern datasets

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

500,000,000,000 tokens =

400,000,000,000 words =

1,000,000,000 A4 pages =

60 km thick book

## ancient datasets (PHI)

Split	Inscriptions	Words	Chars
Train	34,952	2,792k	16,300k
Valid	2,826	211k	1,230k
Test	2,949	223k	1,298k

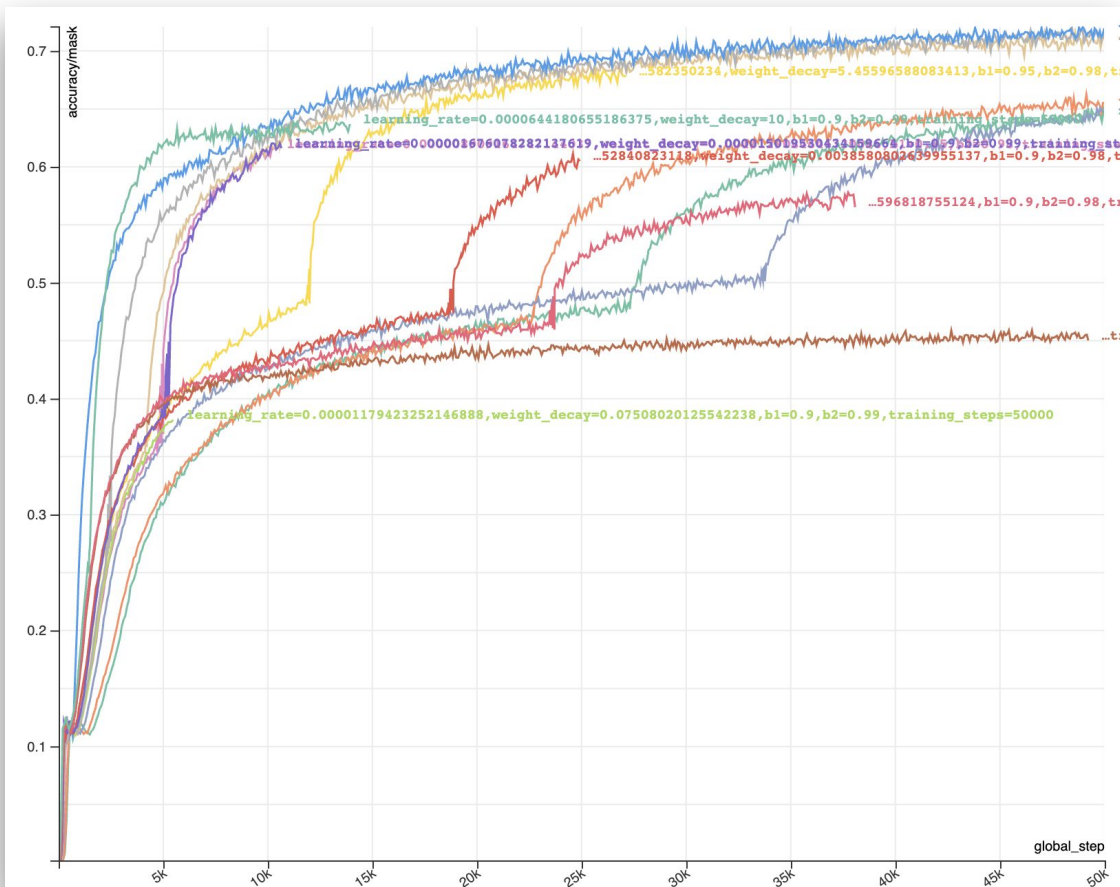
Table 1: Statistics for the PHI-ML corpus.

3,226,000 words =

8,065 A4 pages =

40 cm thick book

# experiments over a few weeks



Restoration  
accuracy of  
Ancient Greek  
and Latin texts  
(higher is better)



# data augmentation

**Original:** υπό το δήμο το αθηναίων και αναγραψάτω ο γραμματεύς

---

**Text clipping:** αθηναίων και αναγραψάτω

---

**Text masking:** υπ--το δή-----θ-να-ων--αι αν-γρ--ά-ω ο γρα---τεύς

---

**Word deletion:** υπό δήμο το και αναγραψάτω ο γραμματεύς

---

**Sentence swap:** αναγραψάτω ο γραμματεύς. υπό το δήμο το αθηναίων

---

**Label smoothing:**



ithaca

# how to do it?

## Collaboration

*“If you only build for yourself, you have the best intuition. But if you don't build for others too, you'll hit a ceiling.” - Prof. Charles Isbell*

## Explainable

Visual aids to guide intuition

## Synergy - not substitution

Show collaborative potential

## Accessibility

Enable wider access and facilitate further research (FAIR principles, Open Access)

## Scalability

How can we contribute at scale?

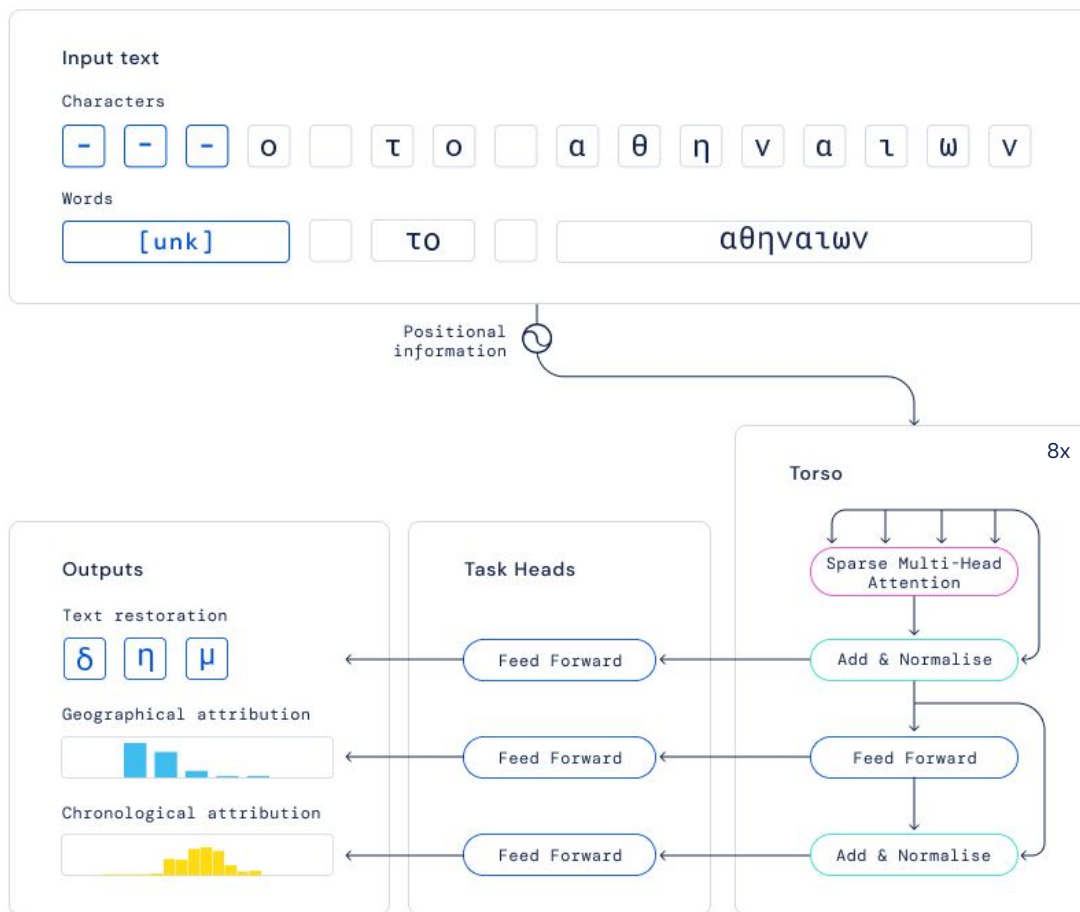


ITHACA





# Ithaca - a neural network model



# interpretable outputs

## Providing multiple restoration hypotheses

(IG II<sup>2</sup> 116, Athens 361/O BCE)

input text:	θεοι επι νικοφημο αρχοντος -----ια αθηναιων και θετταλων εις τον αει χρονον
restorations:	1.  % θεοι επι νικοφημο αρχοντος <b>συμμαχ</b> ια αθηναιων και θετταλων εις τον αει χρονον
(ranked by probability)	2.  % θεοι επι νικοφημο αρχοντος <b>εκκλησ</b> ια αθηναιων και θετταλων εις τον αει χρονον
	3.  % θεοι επι νικοφημο αρχοντος <b>προξεν</b> ια αθηναιων και θετταλων εις τον αει χρονον

Correct

## Inputs that affect Ithaca's restorations (saliency map)

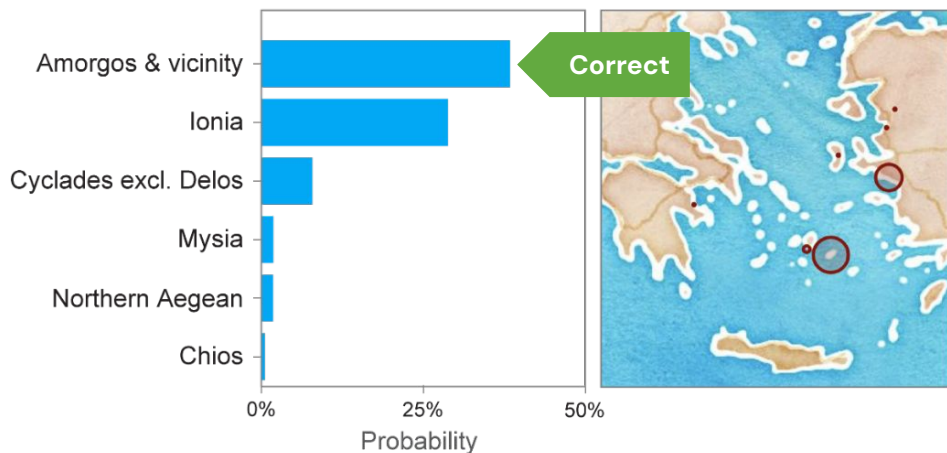
restored characters (per step)	"μ":	θεοι επι νικοφημο αρχοντος ---μ--ια αθηναιων και θετταλων εις τον αει χρονον
	"χ":	θεοι επι νικοφημο αρχοντος ---μ-χια αθηναιων και θετταλων εις τον αει χρονον
	"υ":	θεοι επι νικοφημο αρχοντος -υ-μ-χια αθηναιων και θετταλων εις τον αει χρονον
	"σ":	θεοι επι νικοφημο αρχοντος σμ-μ-χια αθηναιων και θετταλων εις τον αει χρονον
	"μ":	θεοι επι νικοφημο αρχοντος συμμ-χια αθηναιων και θετταλων εις τον αει χρονον
	"α":	θεοι επι νικοφημο αρχοντος συμμαχια αθηναιων και θετταλων εις τον αει χρονον

"Gods. In the archonship of Nikophemos. Alliance of the Athenians and Thessalians for all time."

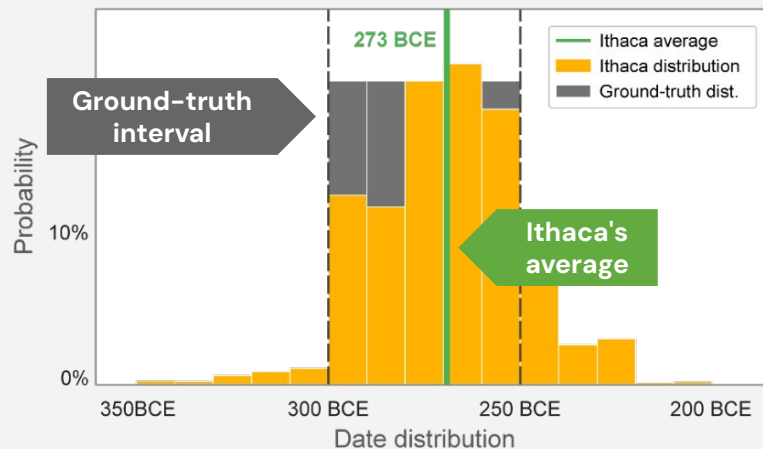


# interpretable outputs

**Geographical attribution**  
(IG XII 7, 2, Amorgos 400–300 BCE)



**Chronological attribution**  
(IG XI 4, 579, Delos 300–250 BCE)



# interpretable outputs

## Saliency maps for chronological attribution

(IG I<sup>3</sup> 371, Athens 414–413 BCE)

δε ες σικελιαν εγον τα χρεματα στρατεγοις νικιαι κυδαντιδει και χουναρχοσι

- Ithaca focuses on the personal name (Νικίας, "Nikias") and the Greek commanders' rank (στρατεγοῖς, "generals").
- Ithaca dates the inscription to 413 BCE, matching the exact range proposed by historians (414–413 BCE).



evaluation

# Experimental evaluation

## Restoration

- The pairing of Ithaca with an epigrapher results in a 3-fold improvement in Top-1 prediction.

	Restoration			Region		Date
Method	CER↓	Top-1↑	Top-20↑	Top-1↑	Top-3↑	Years↓
Ancient Historian & Ithaca	18.3%	71.7%				
Ithaca	26.3%	61.8%	78.3%	70.8%	82.1%	29.3
Pythia	47.0%	32.6%	53.9%			
Ancient Historian*	59.6%	25.3%				
Onomastics				21.2%	26.5%	144.4

Our model

Prior work

Humans



# Experimental evaluation

## Restoration

- The pairing of Ithaca with an epigrapher results in a 3-fold improvement in Top-1 prediction.

## Geographical attribution

- Ithaca predicts 84 regions with 82% Top-3 accuracy.

	Restoration			Region		Date
Method	CER↓	Top-1↑	Top-20↑	Top-1↑	Top-3↑	Years↓
Ancient Historian & Ithaca	18.3%	71.7%				
Ithaca	26.3%	61.8%	78.3%	70.8%	82.1%	29.3
Pythia	47.0%	32.6%	53.9%			
Ancient Historian*	59.6%	25.3%				
Onomastics				21.2%	26.5%	144.4

Our model

Prior work

Humans

# Experimental evaluation

## Restoration

- The pairing of Ithaca with an epigrapher results in a 3-fold improvement in Top-1 prediction.

## Geographical attribution

- Ithaca predicts 84 regions with 82% Top-3 accuracy.

## Chronological attribution

- Ithaca predicts dates within a 29.3 year average (3 year median).

	Restoration			Region		Date
Method	CER↓	Top-1↑	Top-20↑	Top-1↑	Top-3↑	Years↓
Ancient Historian & Ithaca	18.3%	71.7%				
Ithaca	26.3%	61.8%	78.3%	70.8%	82.1%	29.3
Pythia	47.0%	32.6%	53.9%			
Ancient Historian*	59.6%	25.3%				
Onomastics				21.2%	26.5%	144.4

Our model

Prior work

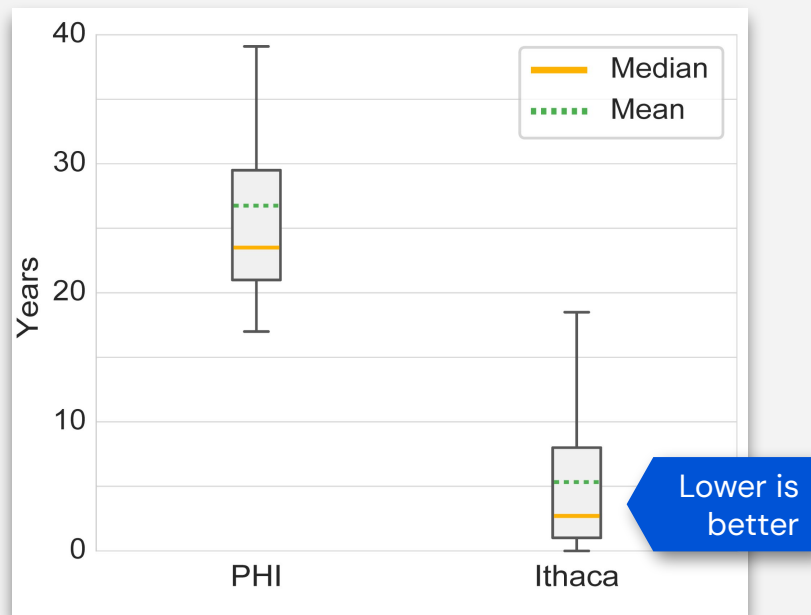
Humans





# Redating history

- Exploiting the unprecedented size of our dataset, Ithaca is more accurate than the PHI ground-truths (5 years vs 27 years distance).
- Predictions **independently align** with recent breakthroughs concerning the dating of influential political inscriptions.
- Ithaca is contributes to key methodological debates in Ancient History.



Ithaca's predictions vs PHI ground-truths compared to modern historical re-evaluations.

**intermission**

# cybersecurity Gemini evaluations

```
C/C++
diff --git a/arch/powerpc/kernel/traps.c b/arch/powerpc/kernel/traps.c
index d9f10f2fc372..5ed4c2ceb5ca 100644
--- a/arch/powerpc/kernel/traps.c
+++ b/arch/powerpc/kernel/traps.c
@@ -900,14 +900,13 @@ void kernel_fp_unavailable_exception(struct pt_regs *regs)

void altivec_unavailable_exception(struct pt_regs *regs)
{
-#if !defined(CONFIG_ALTIVEC)
    if (user_mode(regs)) {
        /* A user program has executed an altivec instruction,
           but this kernel doesn't support altivec. */
        _exception(SIGILL, regs, ILL_ILLOPC, regs->nip);
        return;
    }
-#endif
+
    printk(KERN_EMERG "Unrecoverable VMX/AltiVec Unavailable Exception "
           "%lx at %lx\n", regs->trap, regs->nip);
    die("Unrecoverable VMX/AltiVec Unavailable Exception", regs, SIGABRT);
```

non-security patch

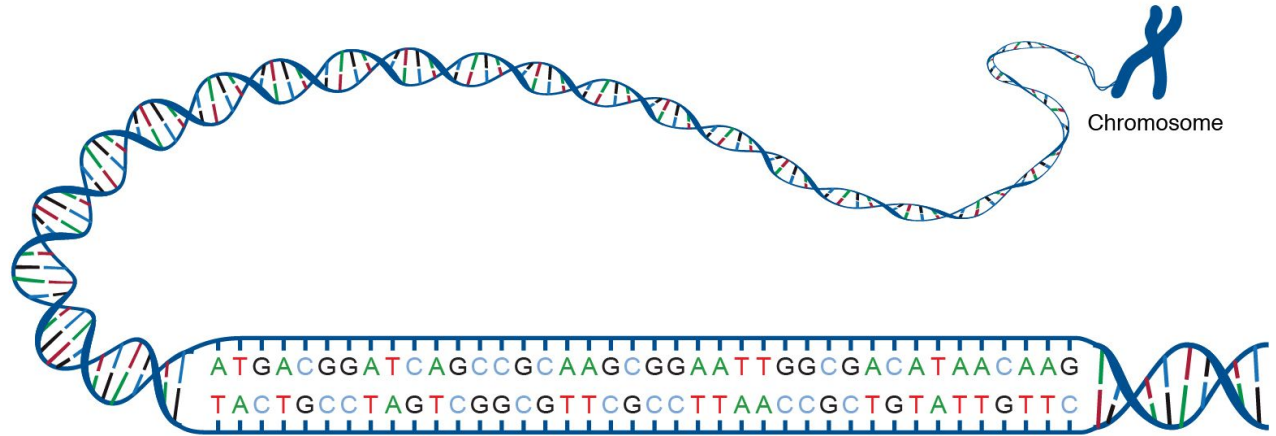
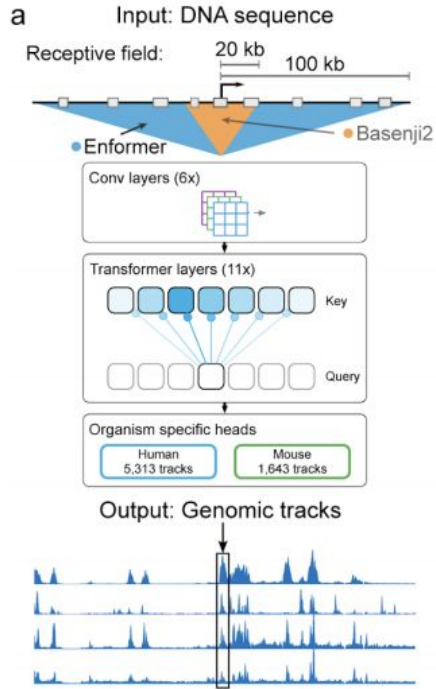
security patch

Task	Model	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	AUC
Wang et al. (2019) <i>Patch Classification</i>	Ultra 1.0	74.0 ± 2.0	75.5 ± 2.6	70.9 ± 4.1	73.1 ± 2.7	0.820 ± 0.018
	Pro 1.0	66.0 ± 2.8	81.9 ± 5.5	40.9 ± 4.2	54.5 ± 4.7	0.735 ± 0.030
SPI <i>Patch Classification</i>	Ultra 1.0	58.5 ± 2.2	58.9 ± 3.7	57.2 ± 4.0	57.9 ± 2.2	0.605 ± 0.023
	Pro 1.0	52.6 ± 3.0	52.3 ± 3.5	55.3 ± 4.7	53.7 ± 3.7	0.530 ± 0.031
DiverseVul <i>Function Classif.</i>	Ultra 1.0	53.8 ± 2.4	57.3 ± 6.9	30.8 ± 3.7	39.9 ± 4.0	0.581 ± 0.046
	Pro 1.0	51.6 ± 3.5	52.3 ± 7.1	42.3 ± 3.1	46.7 ± 4.3	0.533 ± 0.051

Table 9 | Cybersecurity: Vulnerability detection evaluation metrics.<sup>8</sup>



# DNA is a sequence of “words”



[Z Avsec, V Agarwal, D Visentin, J R Ledsam, A Grabska-Barwinska, K R Taylor, Y Assael, J Jumper, P Kohli, D R Kelley, 2021]



**conclusion**



# ithaca.deepmind.com is for all researchers

ITHACA

## Restoring and attributing ancient texts with deep neural networks

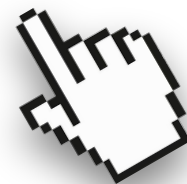
Xoipe! Welcome to Ithaca's interactive interface. Please follow the instructions below to begin restoring and attributing ancient Greek inscriptions. You will also find more information on the Ithaca project, links to the article and examples of Ithaca in action.



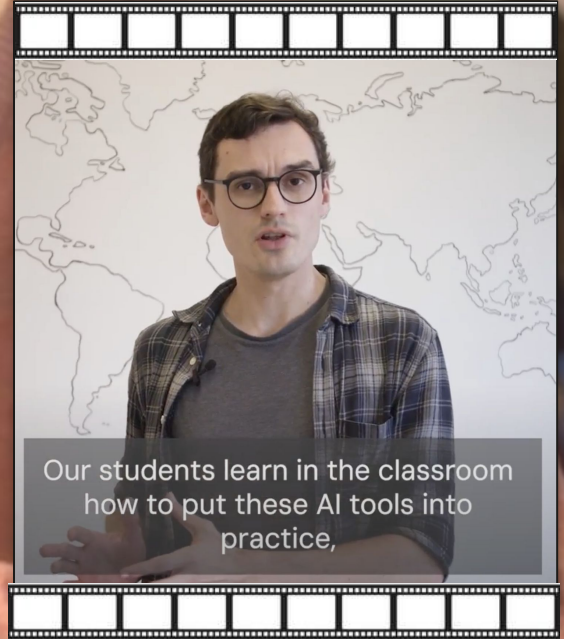
ithaca.deepmind.com

[ithaca.deepmind.com](https://ithaca.deepmind.com) is for all researchers

[ithaca.deepmind.com](https://ithaca.deepmind.com)



# ithaca's adoption



Our students learn in the classroom  
how to put these AI tools into  
practice,



# ithaca's adoption

- Schools using ithaca
  - More than 80 teachers in Europe are using Ithaca in the class
- Researchers using Ithaca
  - Hekatompedon inscription
  - More than 330 new jobs are submitted per week



**PART II: GET STARTED WITH ITHACA**

When we get started with the AI model Ithaca, we use a Notebook. That is a file where HTML, Markdown, Python code, images, text ... can be used interchangeably. More important: the code is ready for you!

So you don't have to immerse yourself in a course on "programming in Python" before you can get started with the AI model. However, it is important that you have:

- a computer with internet browser;
- a stable internet connection;
- a Google or Gmail account.

**STEP 1: INSTALL**

First, the AI model, associated libraries and packages must be downloaded and installed in the Notebook. You will have to perform this step every time you use this Notebook again. The reason is: after each session, the underlying system deletes all files and variables.

You launch the installation by pressing the **▶** button once.

**1) Installatie van Ithaca-onderdelen**

Installatie van het AI-model, startbestand ...

Klik eenmaal op de play-knop **▶** hier links.

Show code

**STEP 2: TEXT INPUT**

In the next step we need to enter our text fragment. When doing this, be sure to use the Greek keyboard and omit capital letters or punctuation. Place a question mark on the missing characters. When it is fully typed you can load it into the AI model by pressing the **▶** button.

Voer hieronder jouw tekst in.

tekst: ἰ ῶ ρ α ν α . α σ τ ῆ ρ α ς π ρ ο ῖ ῶ ρ α

Ingevoerd? Klik eenmaal op de play-knop

**STEP 3: OUTPUT**

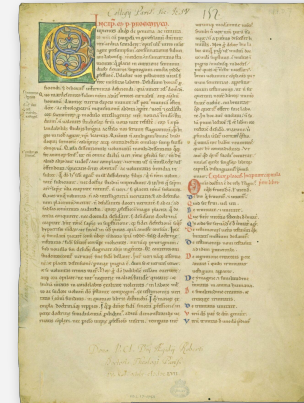
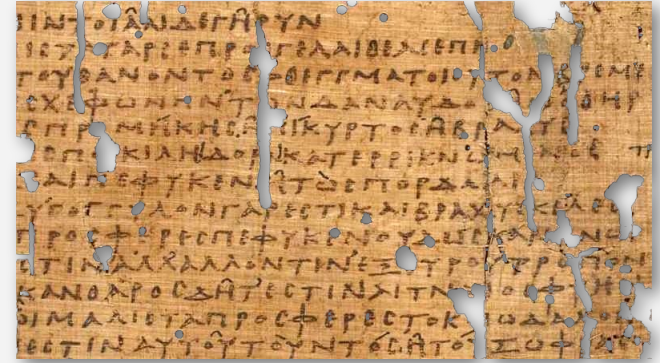
Once your text is entered, the Ithaca model will be able to calculate three things: 1) 20 hypotheses to recover the fragment, 2) situate the text in time, and 3) situate the text in space.

vii

οἴγω, and in our decree, ἀνοίγω appears in section 6); and δύνειν from δύ(ν)ω (to enter). We found this option by entering the clause in the **PYTHIA** model, which offers possible restorations of Greek text based on artificial intelligence (Assael, Sommerschild, Prag

# Ithaca for Humanities

- Ithaca's wider appeal:
  - All disciplines dealing with ancient texts (philology, papyrology, codicology, ...)
  - Any language (ancient or modern)
- Identification and study of newly discovered / forged artefacts.
- **The transformational impact of this work lies in delivering state-of-the-art research aids from the Sciences which extend the scope of the Humanities.**



# AI for Humanities

- Increasing number of publications per year.
- Map the interdisciplinary fields.
- Inspire future research.
- AI for the ancient world: focus on collaboration, linguistic diversity, decision-support, interpretability, human-in-the-loop.

**Joint efforts of specialists in both the Sciences and the Humanities is key to producing relevant, robust and cogent scholarship.**

## Machine Learning for Ancient Languages: A Survey

Thea Sommerschild\*  
Ca' Foscari University of Venice

Yannis Assael\*  
DeepMind

John Pavlopoulos\*  
Athens University of  
Economics and Business

Vanessa Stefanak  
DeepMind

Andrew Senior  
DeepMind

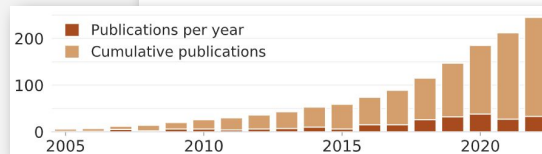
Chris Dyer  
DeepMind

John Bodel  
Brown University

Jonathan Prag  
University of Oxford

Ion Androutsopoulos  
Athens University of  
Economics and Business

Nando de Freitas  
DeepMind



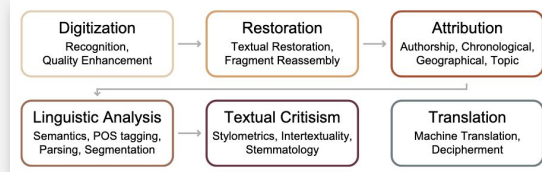
...ever, their study is text-based tasks, from finding the authorship of ancient texts, but in enabled analyses on a few microscopes and

...ent texts written in any language, script and medium, spanning over three and a half millennia of civilisations around the ancient world. To analyse the relevant literature, we introduce a taxonomy of tasks

...ent. This work offers out by the synergy

...active collaboration

...elling scholarship; work promotes and Machine Learning.



# Returning to Ithaca

- Our work hearkens back to Nature's interdisciplinary tradition of scientific communication between Science (Darwin, Einstein and Hawking) and Antiquity (Schliemann).
- **Research like Ithaca can unlock the cooperative potential between AI and historians,**
- **Delivering research aids that extend the scope of ancient history and the Humanities.**

NATURE

[Oct. 3, 1878

## THE ANCIENT CAPITAL OF ITHACA

IN a recent letter to the *Times* Dr. Schliemann describes his search for the ancient capital of the island of Ithaca. He began his researches in the valley called



A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE.

# ευχαριστώ!



assael.gr



yannisassael



iassael



iassael

