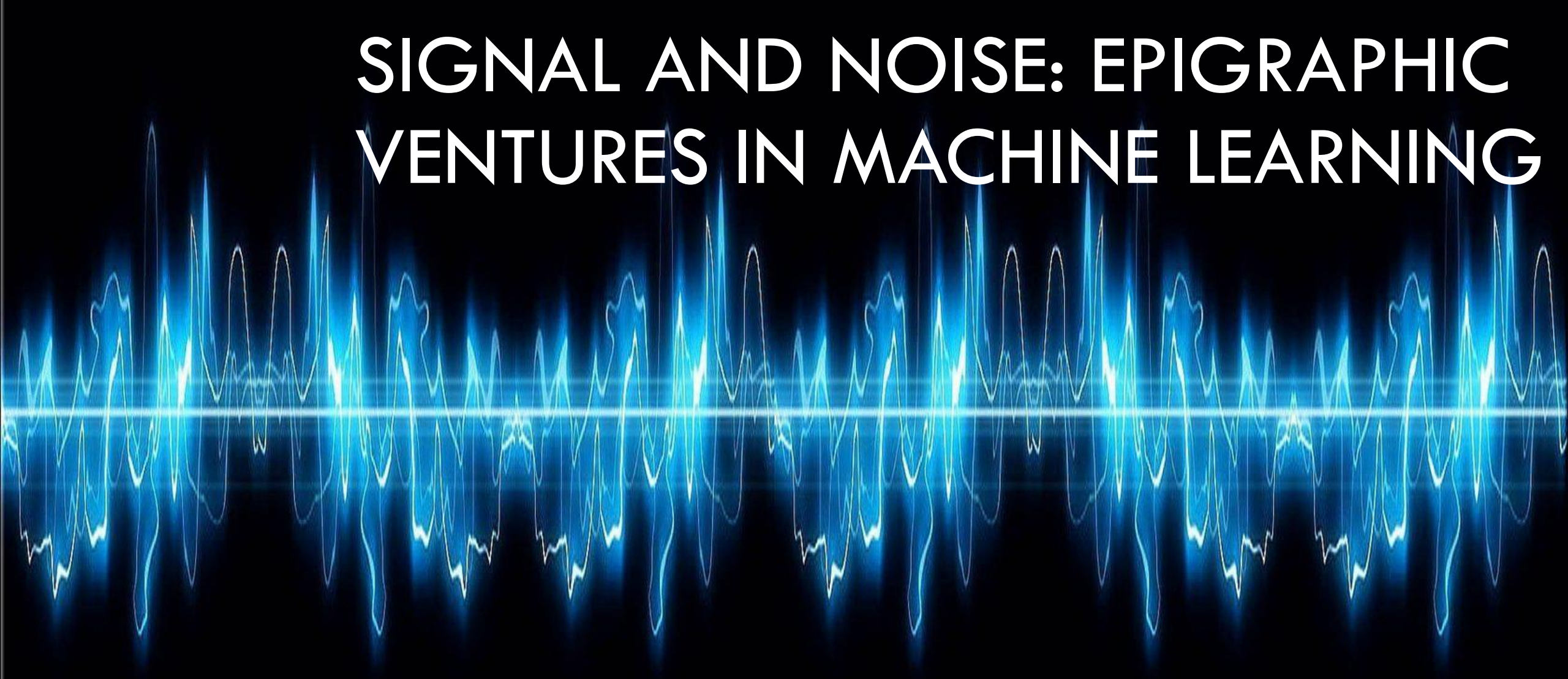
SIGNAL AND NOISE: EPIGRAPHIC VENTURES IN MACHINE LEARNING



DR. AARON HERSHKOWITZ

PROJECT MANAGER,

KRATEROS SQUEEZE DIGITIZATION PROJECT

ENCODE WORKSHOP

'AI AND ANCIENT WRITING CULTURES'

BOLOGNA, 23RD-27TH JANUARY 2023

ROADMAP

- Introduction and Apologia
- Overview of Epigraphic Machine Learning Projects
  - NLP and “Text Restoration”
  - Classification
  - Computer Vision
- Summary and Path Forward

LANGUAGE-BASED MACHINE LEARNING

ITHACA

- Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, Nando de Freitas
- https://ithaca.deepmind.com/
- https://github.com/deepmind/ithaca

ITHACA

- Text-focused, Greek language based
- Assesses 3 core problem areas:
  - “Text restoration”
  - Geographical location
  - Chronological placement

Without Restoration

Restoration of a damaged inscription, recording a decree from 485/4 BCE concerning the Acropolis of Athens (IG I^3 4B, CC BY-SA 3.0, WikiMedia).

With Restoration

Restoration of a damaged inscription, recording a decree from 485/4 BCE concerning the Acropolis of Athens (IG I^3 4B, CC BY-SA 3.0, WikiMedia).

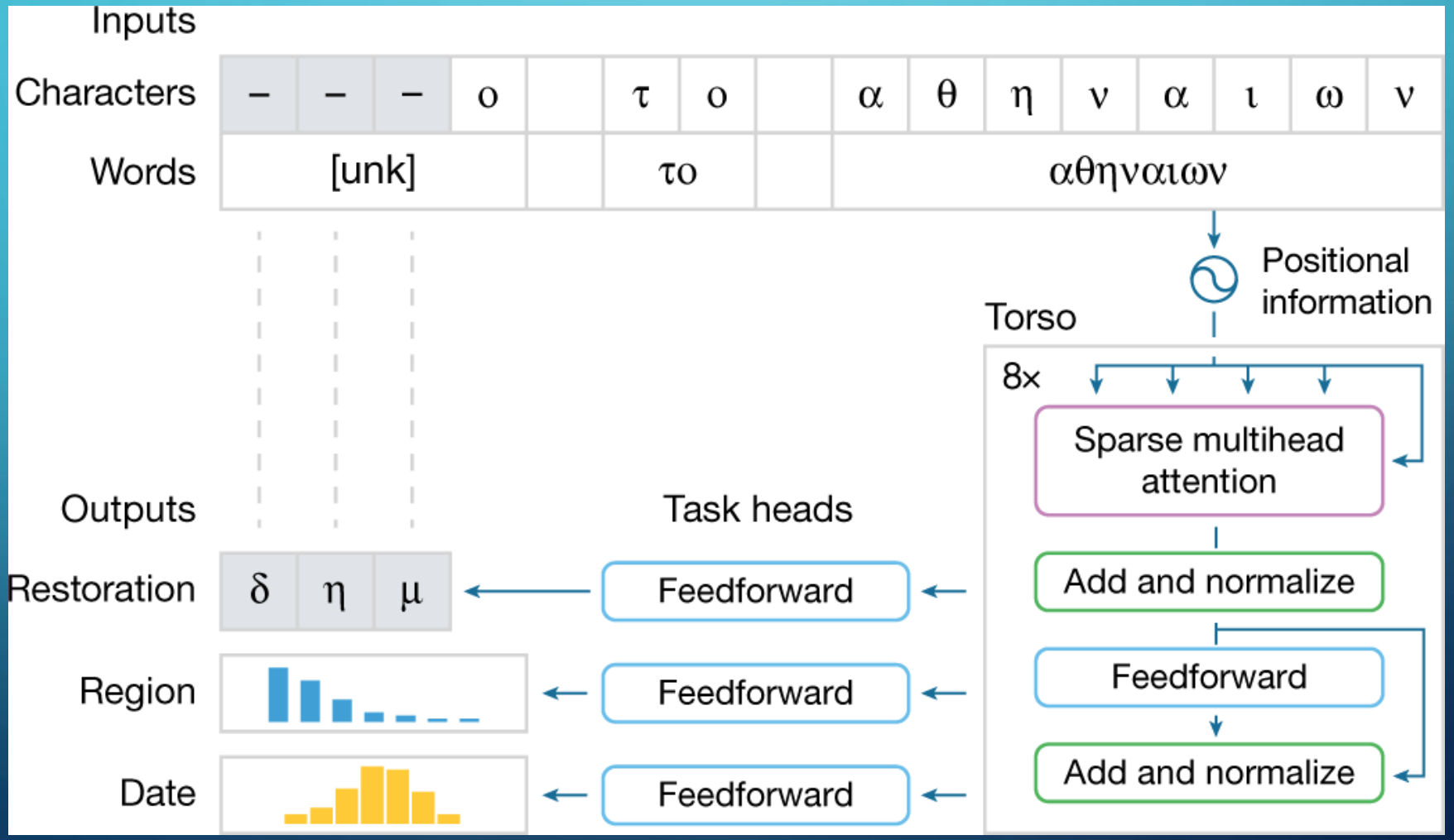
ITHACA

- Text-focused, Greek language based
- Assesses 3 core problem areas:
  - "Text restoration"
  - Geographical location
  - Chronological placement
- Important underlying dataset: I.PHI or PHI-ML (https://github.com/sommerschield/iphi)

ITHACA

- Uses deep neural network architecture (Deepmind)
  - Processes text input as character and word jointly
  - "Transformer" architecture directs focus and weights attention for task heads addressing three problem areas
- Builds on previous work in "Pythia" - sequence to sequence recurrent neural network for ancient-text restoration (https://wiki.digitalclassicist.org/Pythia)

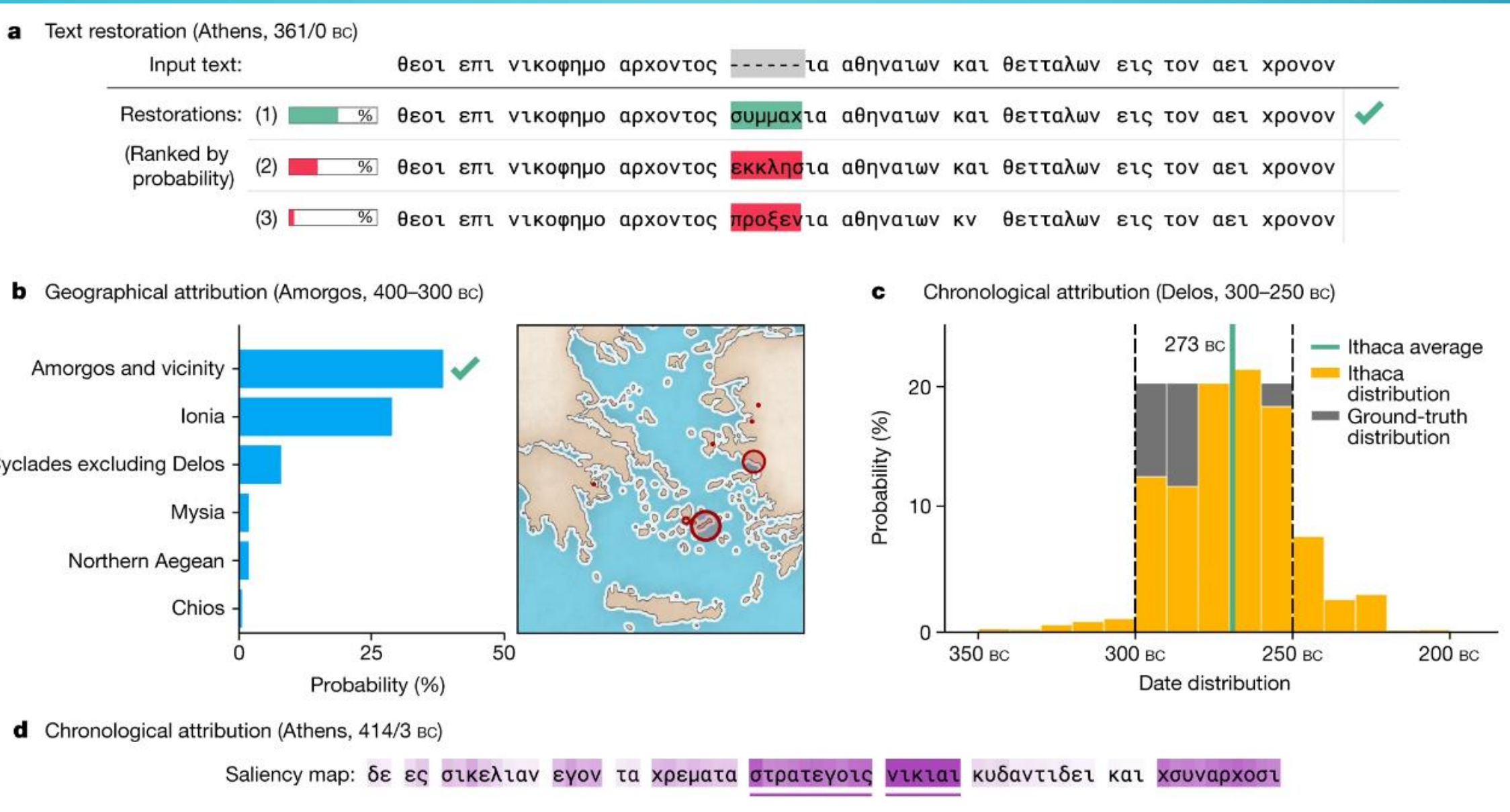
Ithaca's architecture processing the phrase 'δήμο το αθηναίων' ('the people of Athens')



ITHACA

- Uses deep neural network architecture (Deepmind)
  - Processes text input as character and word jointly
  - "Transformer" architecture directs focus and weights attention for task heads addressing three problem areas
- Builds on previous work in "Pythia" - sequence to sequence recurrent neural network for ancient-text restoration (https://wiki.digitalclassicist.org/Pythia)

ITHACA OUTPUT



ITHACA EXPERIMENTAL RESULTS

	Restoration			Region		Date
Method	**CER (%)**	**Top 1 (%)**	**Top 20 (%)**	**Top 1 (%)**	**Top 3 (%)**	**Years**
Ancient historian and 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

Evaluating methods for text restoration, geographical attribution (region) and chronological attribution (date) on I.PHI's test set of $n = 7{,}811$ inscriptions. For 'CER' and 'years', lower scores are better. For 'top 1', 'top 3' and 'top 20', higher scores are better. For each metric, the best performing method is in bold.

ITHACA

- Open access, underlying processes applicable to ancient texts broadly (papyrology, numismatics, codicology)
- Upside: designed as a tool, most impressive results in text restoration appear in human + machine context
- Downside: inability to provide clear and compelling rationale for conclusions (broad downside to ML)

CLASSIFICATION MACHINE LEARNING

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- Vojtěch Kaše, Petra Heřmánková, Adéla Sobotková
- Text-based, feature-focused, Latin language
- Goal: crosswalk inscriptions in Epigraphik Datenbank Clauss-Slaby (EDCS) to Epigraphic Database Heidelberg (EDH) categories
- https://github.com/sdam-au/LIRE_ETL

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- Latin Inscriptions of the Roman Empire dataset
- Aggregate of EDH and EDCS epigraphic datasets, focusing on inscriptions which are:
  - geolocated
  - within the borders of the Roman Empire in its highest extent
  - dated
  - in the dating interval intersecting the period from 50 BC to 350 AD

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- Latin Inscriptions of the Roman Empire dataset
- Dataset consists of 137,305 records and 110 attributes:
  - 49,916 inscriptions shared by the EDH and EDCS
  - 3,907 inscriptions recorded exclusively in EDH
  - 83,482 inscriptions originating solely from EDCS
- https://zenodo.org/record/5074774#.Y8clGnbMIuU

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- EDH inscription categories based on EAGLE Europeana Project standardized list of inscription types (22 unique categories)
- EDCS uses Latin labels and categories that do not overlap easily with EDH/EAGLE

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- Used *tfidf* vectorizer fed with 'bag-of-words' model created from 3 EDCS attributes:
  - inscription category and other metadata
  - predominant material or medium on which inscription is found
  - text of inscription (without "Leiden" markup)
- Continuous bigrams to capture formulaic language of inscriptions

CLASSIFYING LATIN INSCRIPTIONS OF THE ROMAN EMPIRE

- Supervised machine learning algorithms for document classification considered for preliminary testing:
  - Logistic Regression (LR)
  - Support-vector Machine (SVM)
  - Random Forests (RF)
  - Extremely Randomized Trees (ET)
- Implemented in Python 3 and Scikit-learn library

CLASSIFICATION MODEL SELECTION RESULTS

- Training set N = 4000
- C stands for inverse regularization strength
- n_estimators is number of estimators
- The F1(w) score is the harmonic mean of Precision (proportion of every observation predicted to be positive that is actually positive) and Recall (proportion of every positive observation that is truly positive)

classifier	C	n_estimators	avg. $F_1(w)$
LR	1		0.808297
LR	1000		0.831937
LR	10000		0.830611
SVM	1		0.310482
SVM	1000		0.760049
SVM	10000		0.828478
RF		10	0.815809
RF		100	0.825325
RF		1000	0.826737
ET		10	0.822891
ET		100	0.831064
ET		1000	0.830801

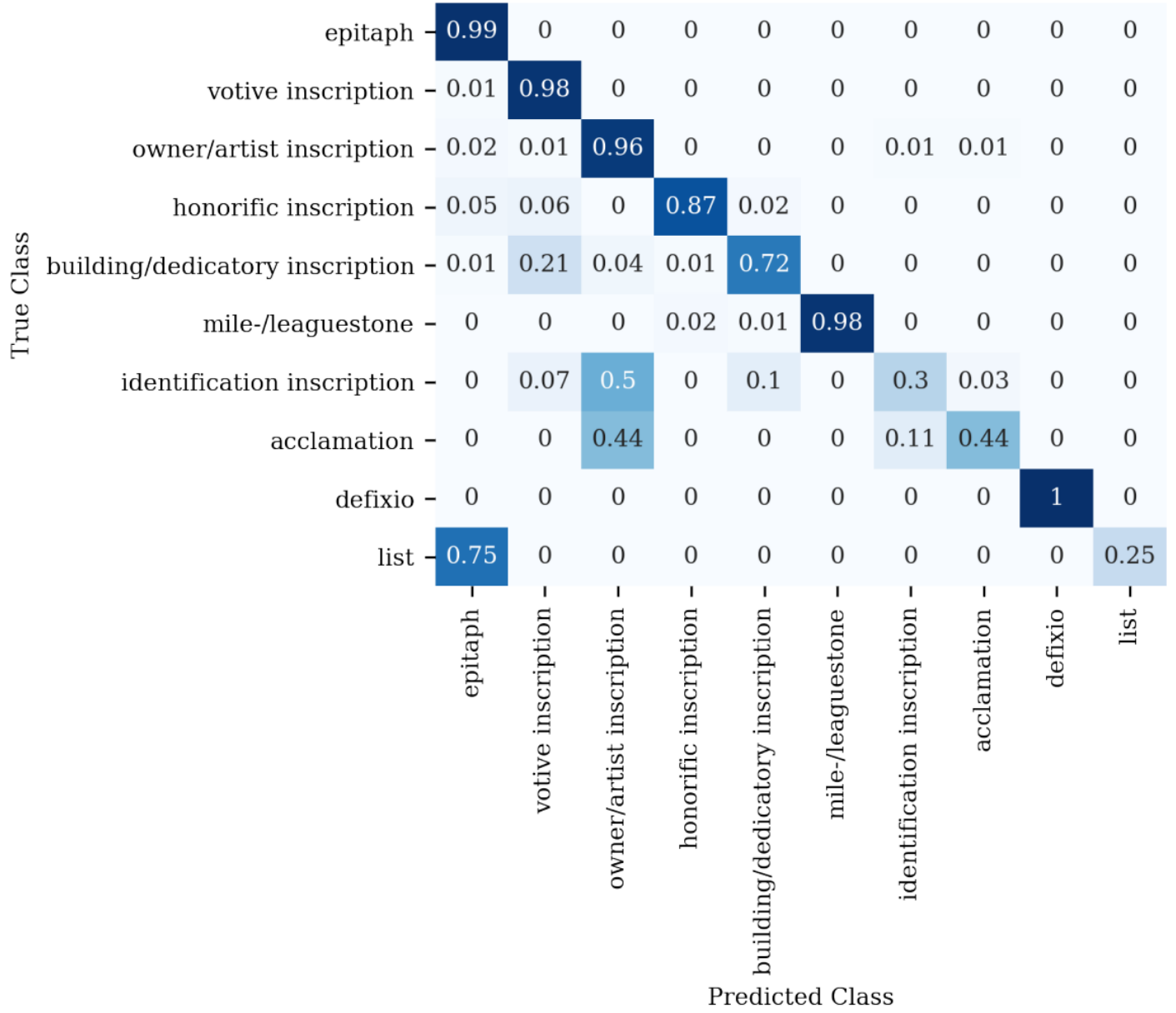
CLASSIFICATION MODEL TEST RESULTS

- Threshold is probability on 0-1 scale expressing level of certainty concerning predicted category

threshold ($\geq$)	proportion	N	$F_1(w)$	accuracy
0.40	0.96	4448	0.897102	0.905800
0.45	0.91	4225	0.923481	0.929467
0.50	0.89	4123	0.935463	0.940092
0.55	0.87	4027	0.945704	0.949839
0.60	0.85	3936	0.952090	0.955539
0.65	0.83	3853	0.957309	0.960550
0.70	0.81	3755	0.962801	0.965379
0.75	0.79	3653	0.965969	0.968519
0.80	0.76	3526	0.970297	0.972490
0.85	0.70	3253	0.978090	0.979096
0.90	0.67	3082	0.979371	0.980208
0.95	0.60	2762	0.981634	0.982259

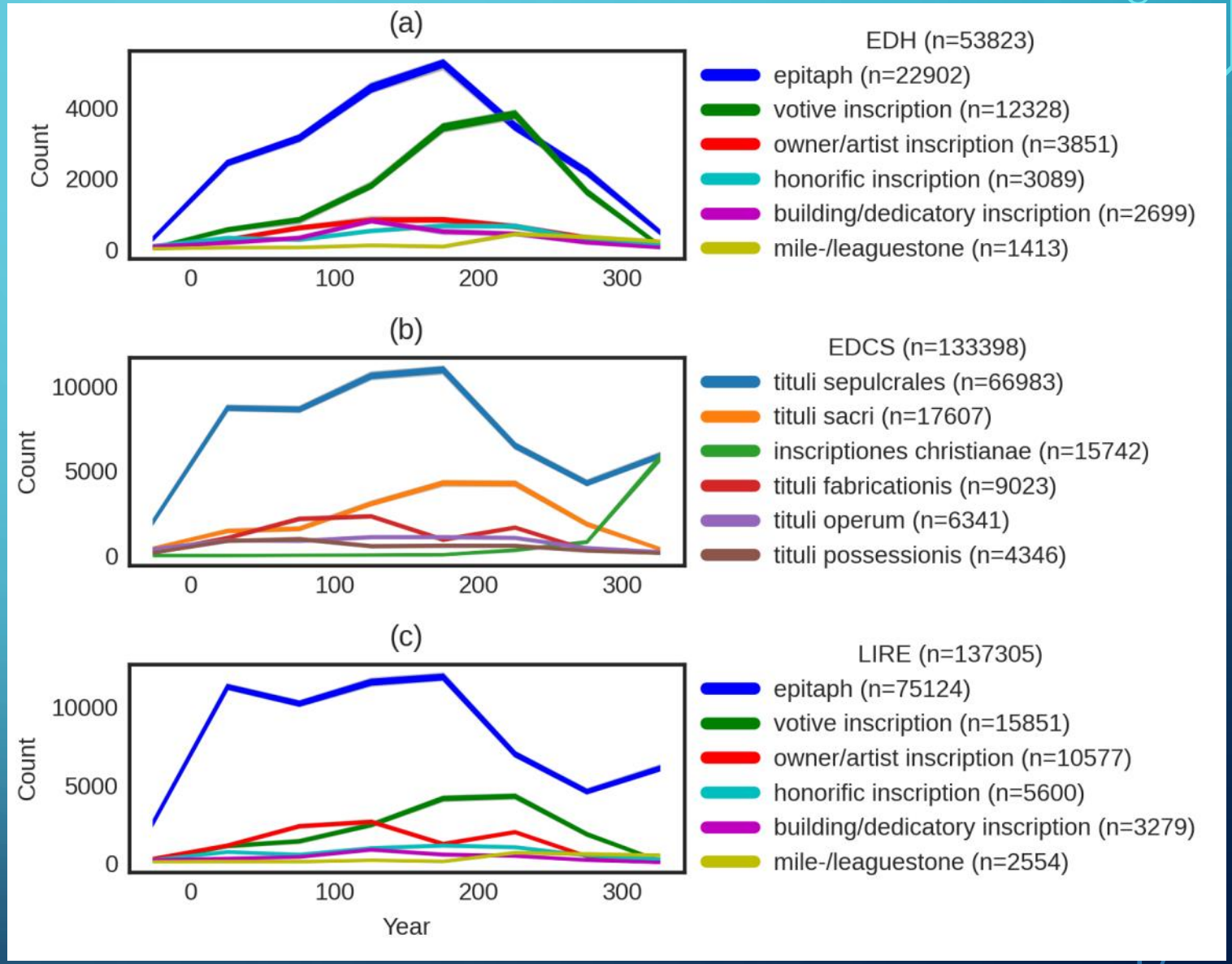
PRECISION TABLE FOR 10 MOST COMMON INSCRIPTION CATEGORIES

- Probability 0.6 and above only



TEMPORAL DISTRIBUTION OF 6 MOST COMMON INSCRIPTION TYPES

- (a) Inscription types as labeled by EDH, one label per inscription
- (b) Inscription types as labeled by EDCS, multiple labels per inscription allowed
- (c) Aggregate of manual EDH and automated EDCS classifications



COMPUTER VISION

AUTOMATED WRITER IDENTIFICATION

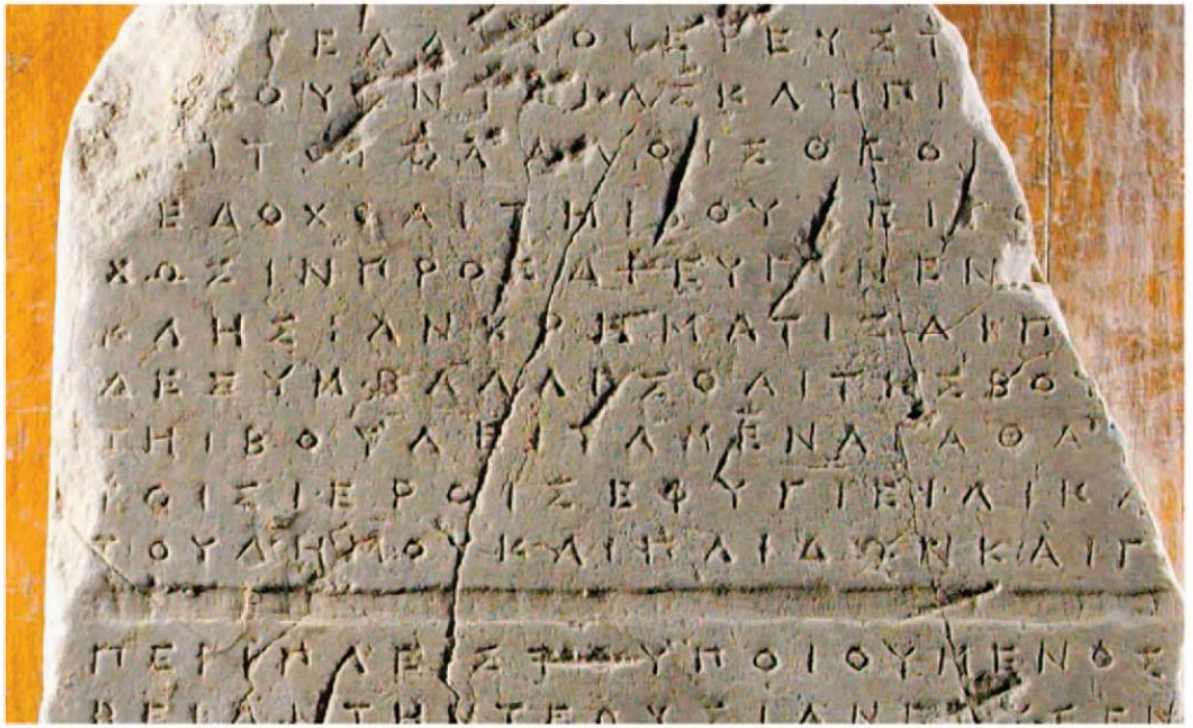
- Michail Panagopoulos, Constantin Papaodysseus, Panayiotis Rousopoulos, Dimitra Dafi, and Stephen Tracy
- Image-based, Greek language
- Attempt to automate S. Tracy's ability to match inscriptions with same stonecutter for dating purposes
- VERY EARLY – results published in 2007

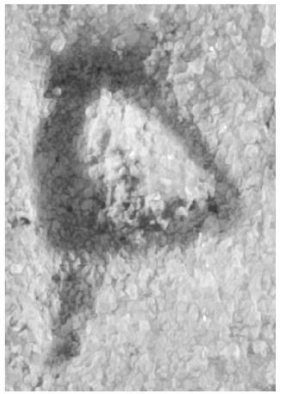
AUTOMATED WRITER IDENTIFICATION

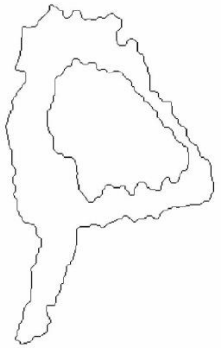
- Methodology:
  - Inscriptions photographed according to strict protocol
  - Images segmented and contours of letters extracted
  - Critical points detected for each letter
  - In each inscription, average contours to create “ideal prototype” for each letter
  - Compare inscription ideal prototype letters pairwise

AUTOMATED WRITER IDENTIFICATION

- Inscription carefully photographed
- Segmentation and contour detection





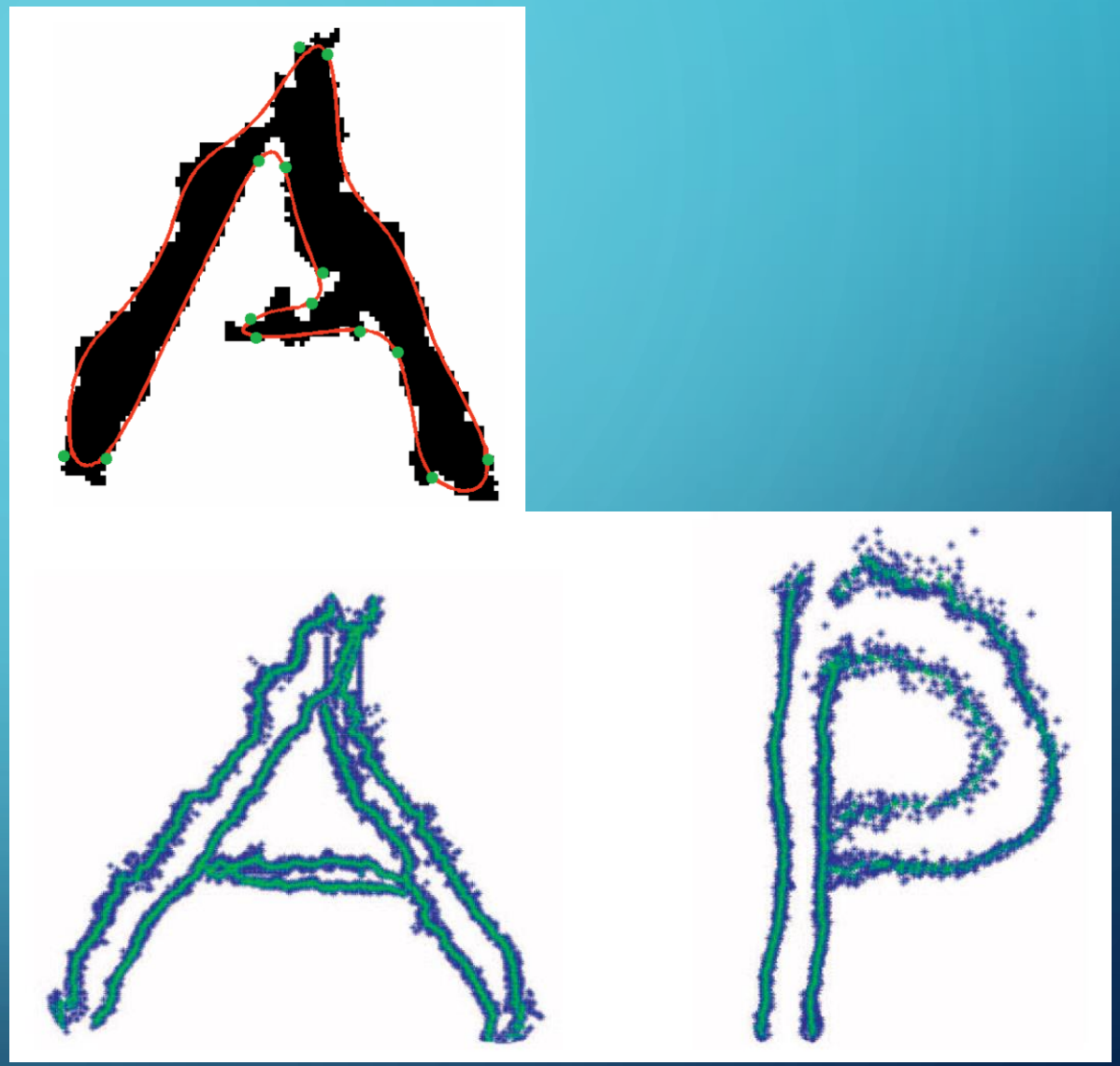


AUTOMATED WRITER IDENTIFICATION

- Methodology:
  - Inscriptions photographed according to strict protocol
  - Images segmented and contours of letters extracted
  - Critical points detected for each letter
  - In each inscription, average contours to create “ideal prototype” for each letter
  - Compare inscription ideal prototype letters pairwise

AUTOMATED WRITER IDENTIFICATION

- Critical point detection
- “Ideal prototype” creation

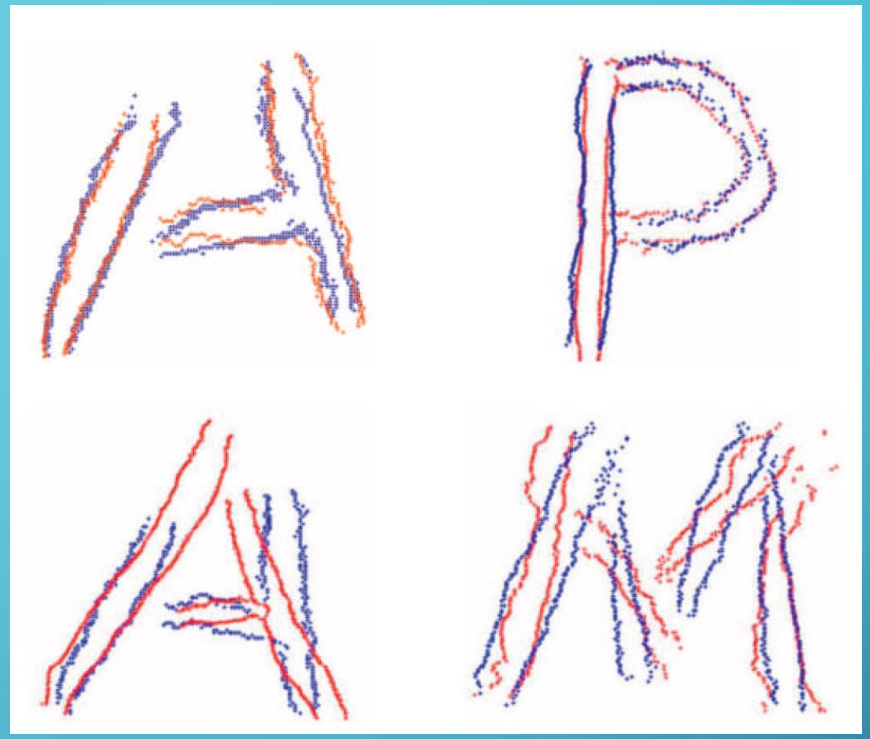


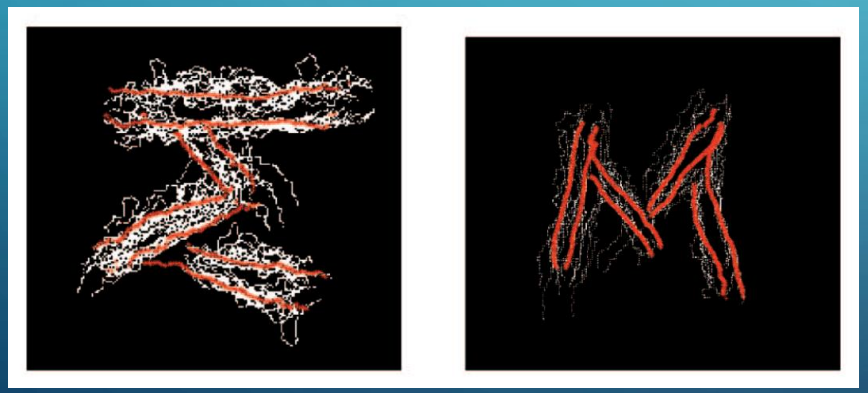
AUTOMATED WRITER IDENTIFICATION

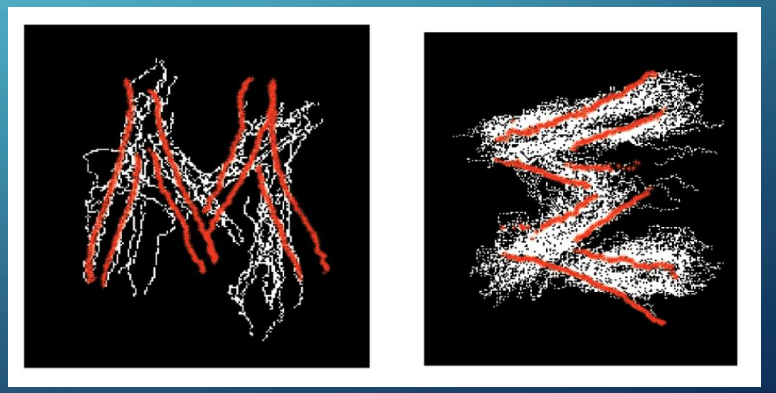
- Methodology:
  - Inscriptions photographed according to strict protocol
  - Images segmented and contours of letters extracted
  - Critical points detected for each letter
  - In each inscription, average contours to create “ideal prototype” for each letter
  - Compare inscription ideal prototype letters pairwise

AUTOMATED WRITER IDENTIFICATION

- Comparison of “ideal prototypes”
- Comparison of contour realizations to “ideal prototypes”







AUTOMATED WRITER IDENTIFICATION

- Writer identification task was successful in matching relatively small set of test inscriptions
- Funding was not received for further work
- Problem: lack of existing image set suited to segmentation

RECONSIDERING THE ROMAN WORKSHOP

- Charlotte Tupman, Dmitry Kangin, and Jacqueline Christmas
- Image-based, Latin language
- Interested in planning and execution of inscriptions

RECONSIDERING THE ROMAN WORKSHOP

- Like Ithaca/Pythia, aims to aid in reconstruction/dating, but via analysis of planning + layout of inscription
- Project separated into two phases:
  - Detection of image areas containing text (completed)
  - Detection of areas of individual characters within text (forthcoming)

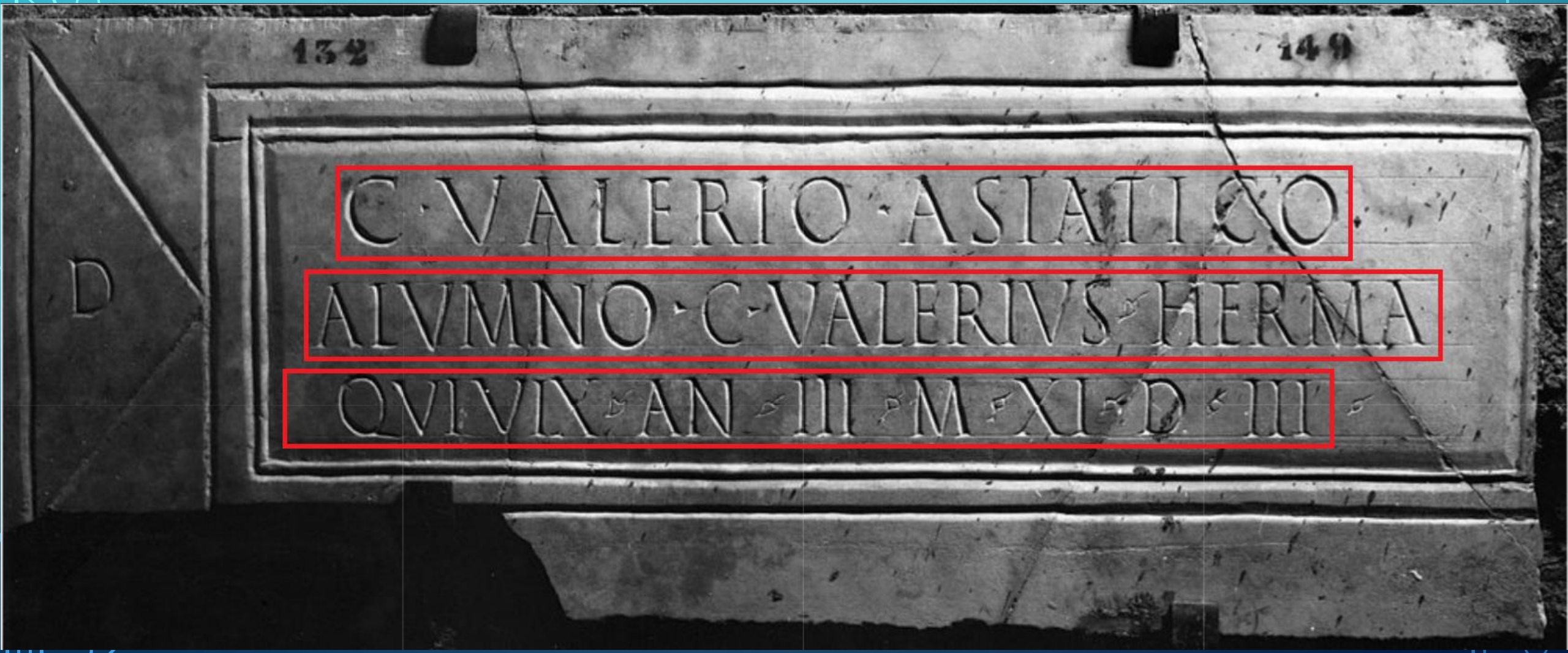
RECONSIDERING THE ROMAN WORKSHOP

- Selected EAST scene text detector algorithm:
  - Bounding box area for text
  - Robust performance on text detection
  - Fully differentiable model to enable downstream analysis of outputs
  - Easily reproducible model

RECONSIDERING THE ROMAN WORKSHOP

- Procedure to minimize manual labeling (no ground truth data):
  - Run algorithm to create bounding boxes
  - Select images with correct text identification and without false positives, then erase false negatives
  - Retrain the model on data set augmented by resulting ground truth

ROMAN WORKSHOP OUTPUT (E.G.) – EDH F006124 HD009818, AE 1987, 0112



RECONSIDERING THE ROMAN WORKSHOP

- Preliminary results:
  - Trained on 1,300 images, applied to 4,000 images
  - This number (10% of EDH images) determined by selecting inscriptions which could be feasibly analyzed by human eye to train the model
  - Enough success to proceed to character detection

RECONSIDERING THE ROMAN WORKSHOP

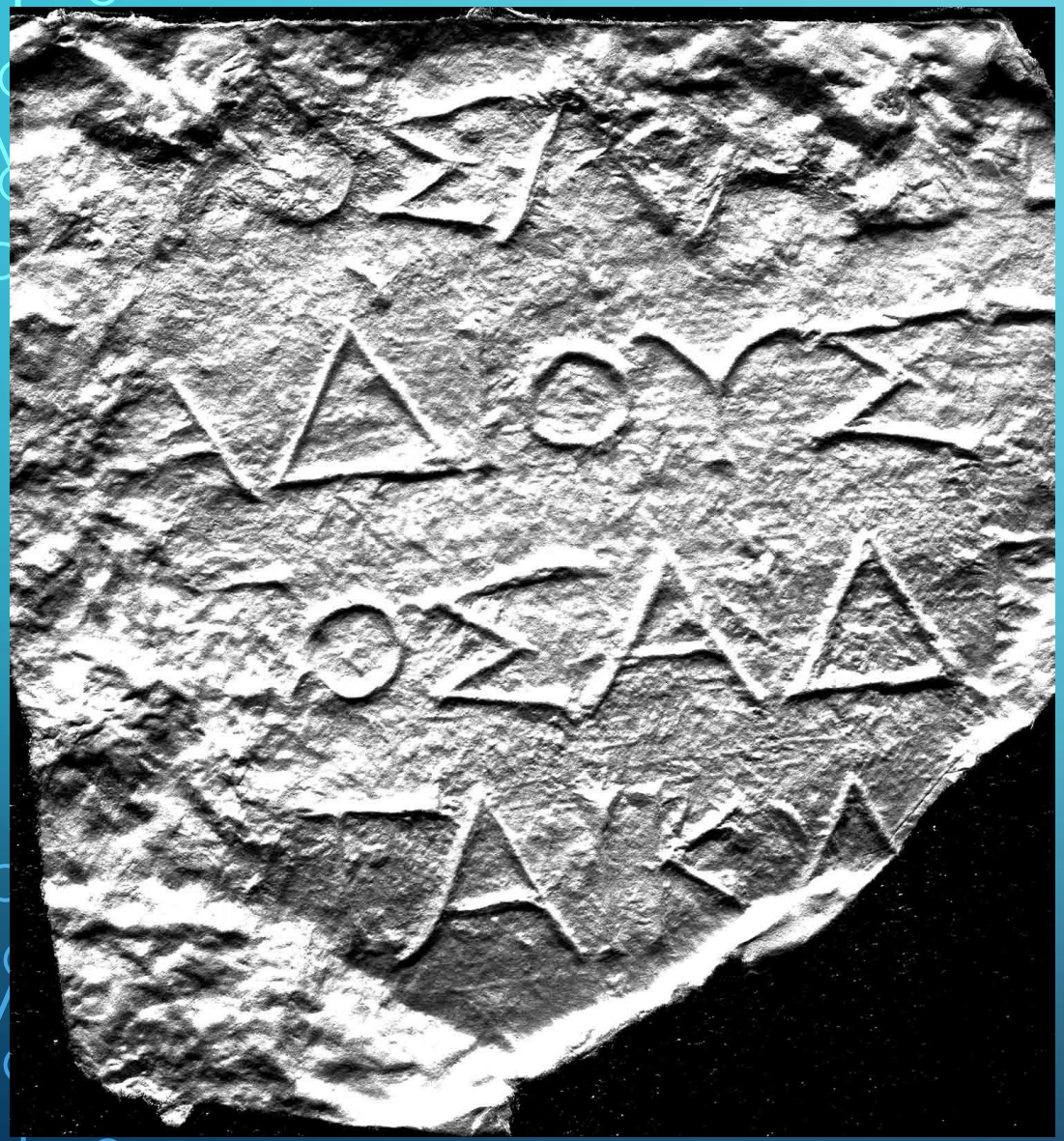
- Issues observed:
  - False positives in chisel marks and decoration
  - Overlapping bounding boxes
  - False negatives at beginning/ends of lines
  - False negatives caused by intense shade

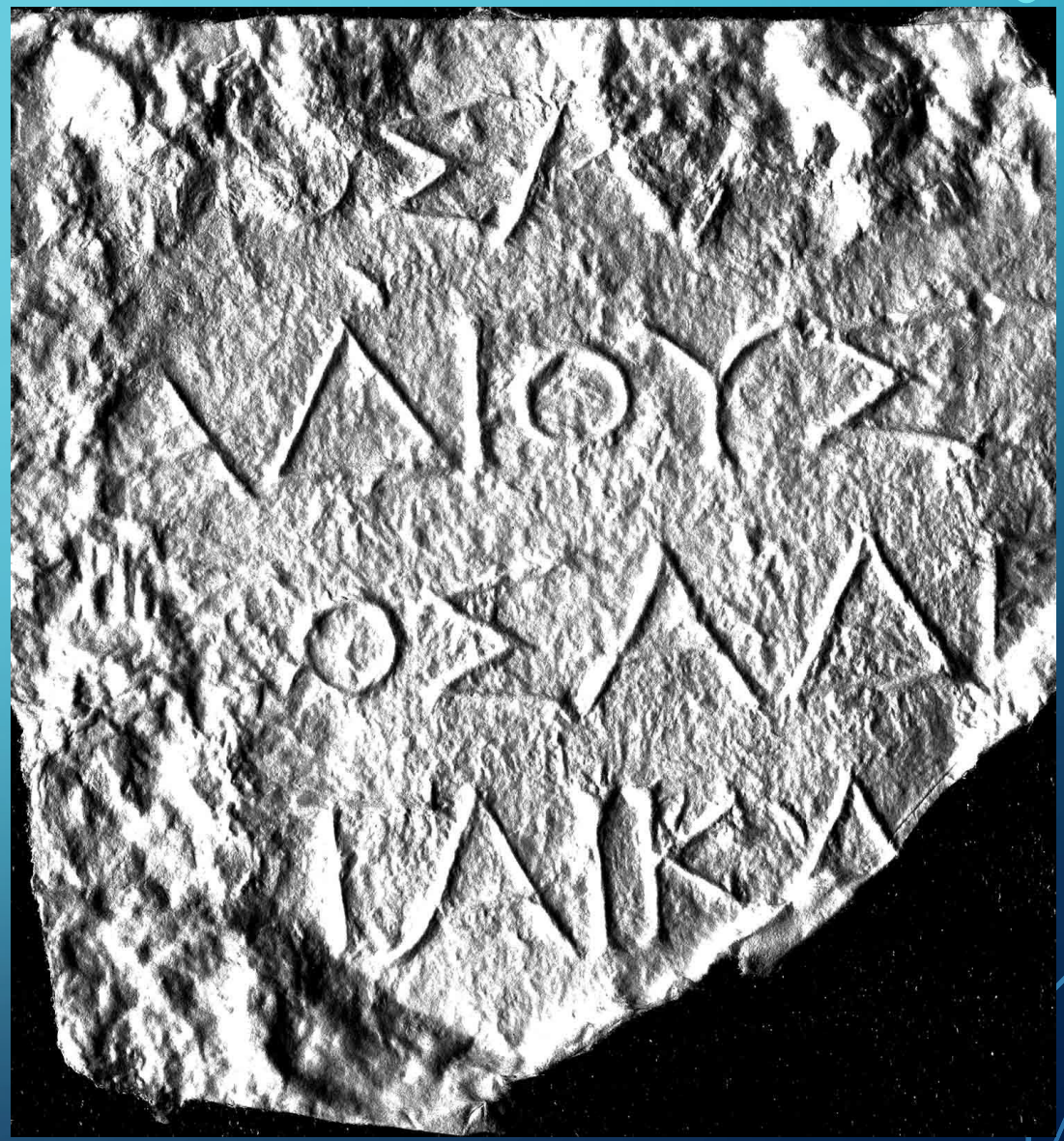
KRATEROS

- Angelos Chaniotis, Aaron Hershkowitz, Nicholas Howe, Stephen Tracy
- Primarily a digitization project aimed at epigraphic squeezes
  - 2D scanning of ~30,000 squeeze collection completed
  - 3D scanning (photogrammetry) under way
- S. Tracy is project advisor, interest in picking up earlier project
- Still seeking funding

THE KRATEROS DIGITIZATION PROCESS: 2D SCANNING

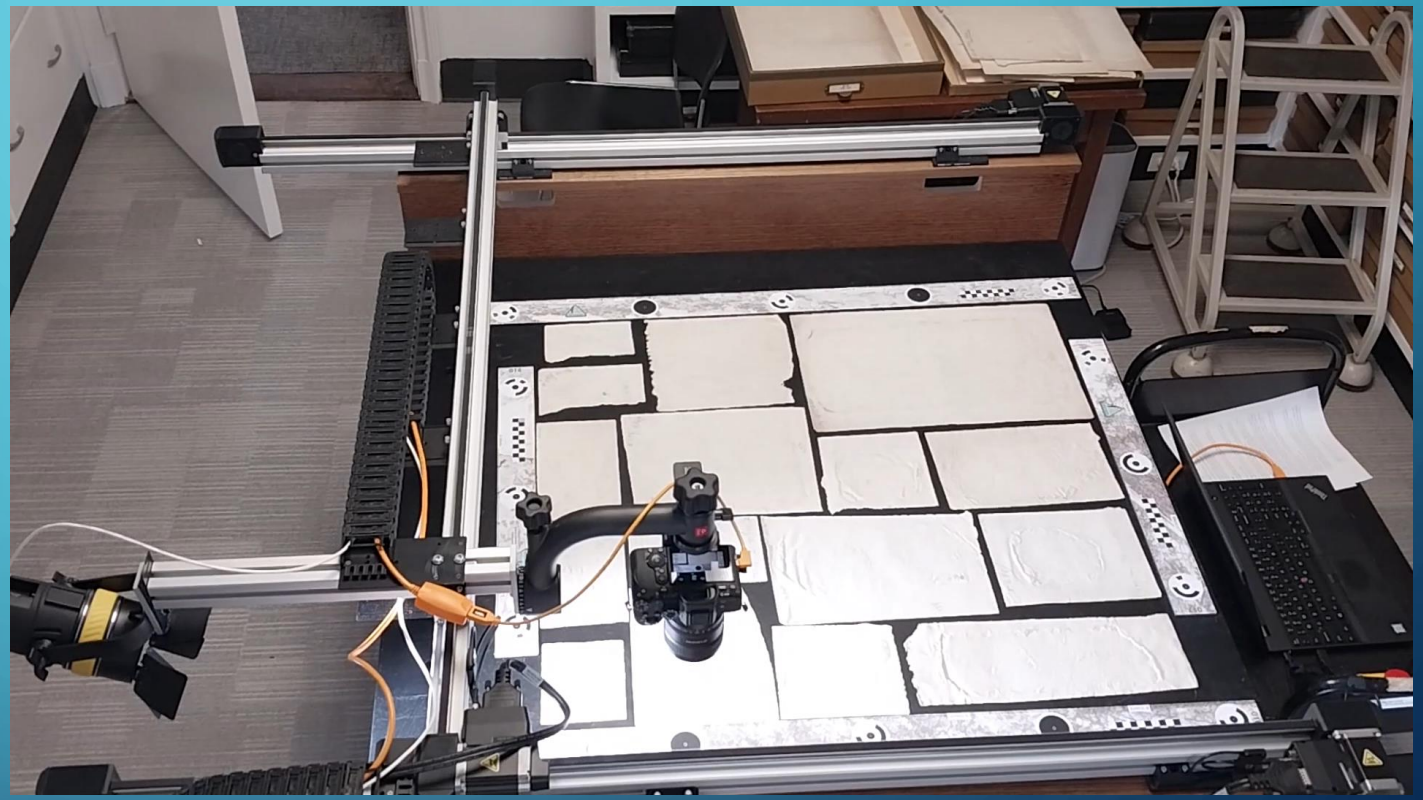
- WideTEK 25-600 flatbed scanner (18.5” x 25”)
- 3D-lighting, greyscale, 600dpi TIFFs
- Scan each squeeze twice: once upright, once rotated 90°

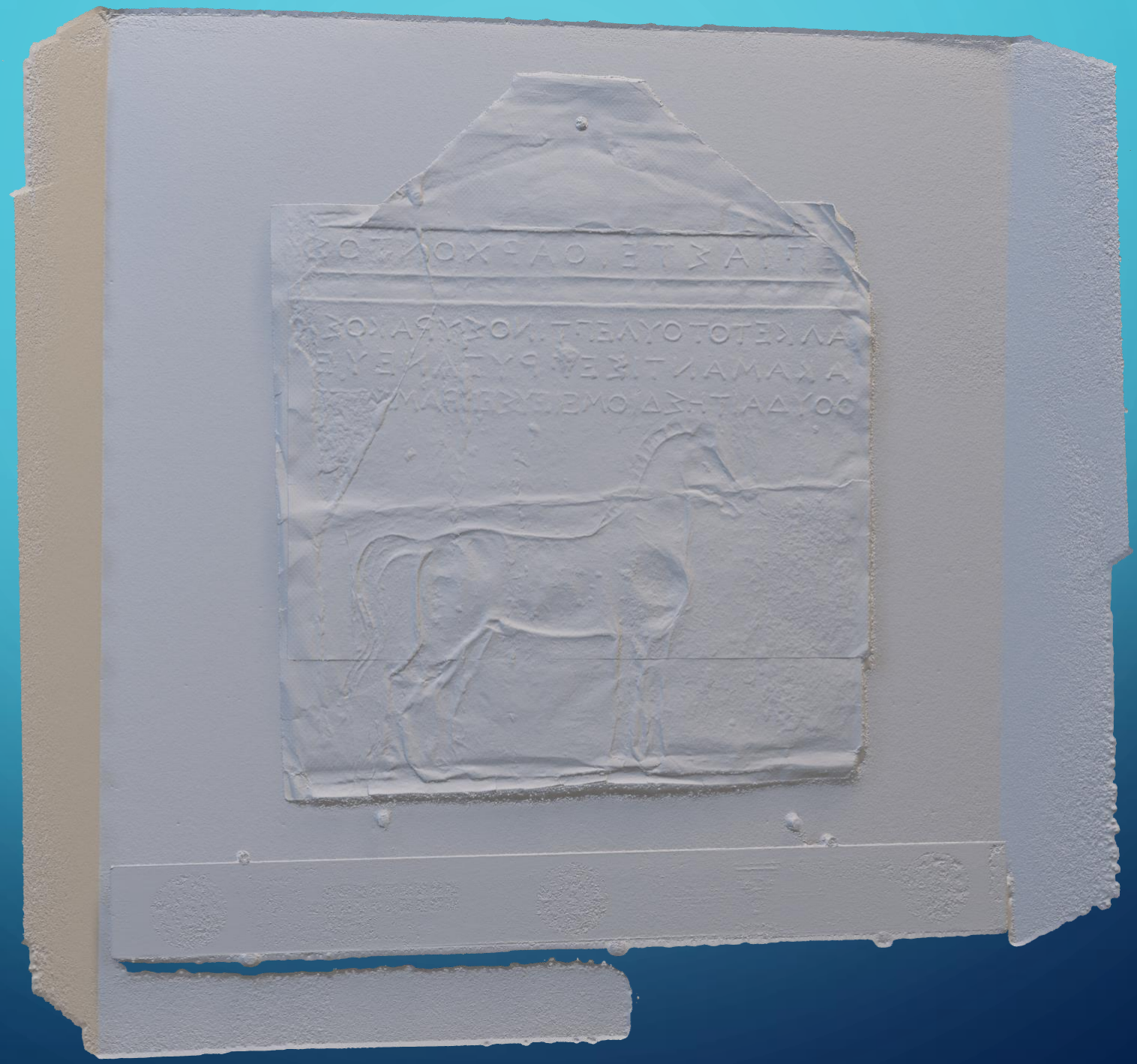




THE KRATEROS DIGITIZATION PROCESS: 3D SCANNING

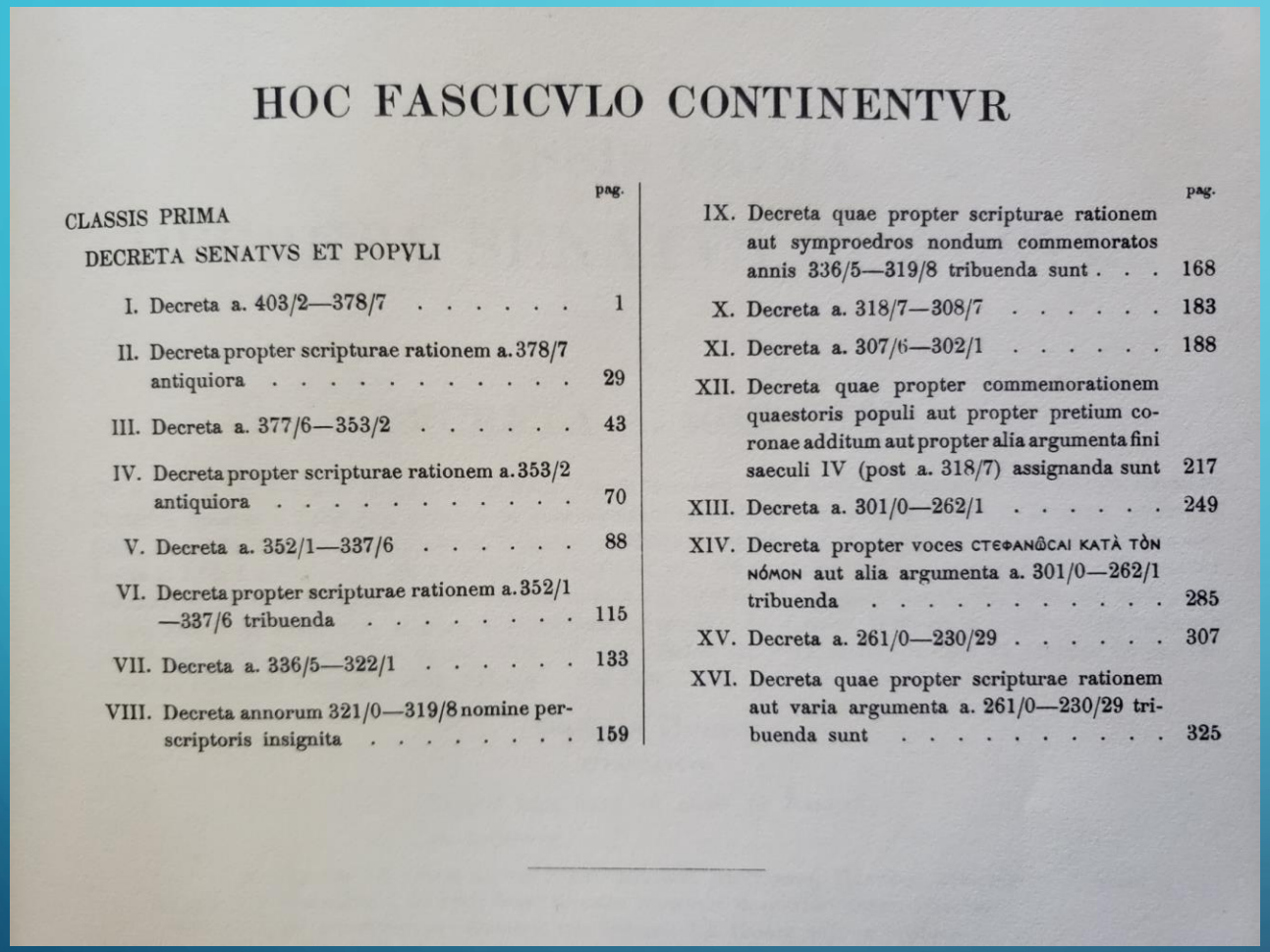
- Photogrammetry
- Automated XY Gantry on 1.2x1.2m table
- Multiple squeezes scanned at once, then “cut” out





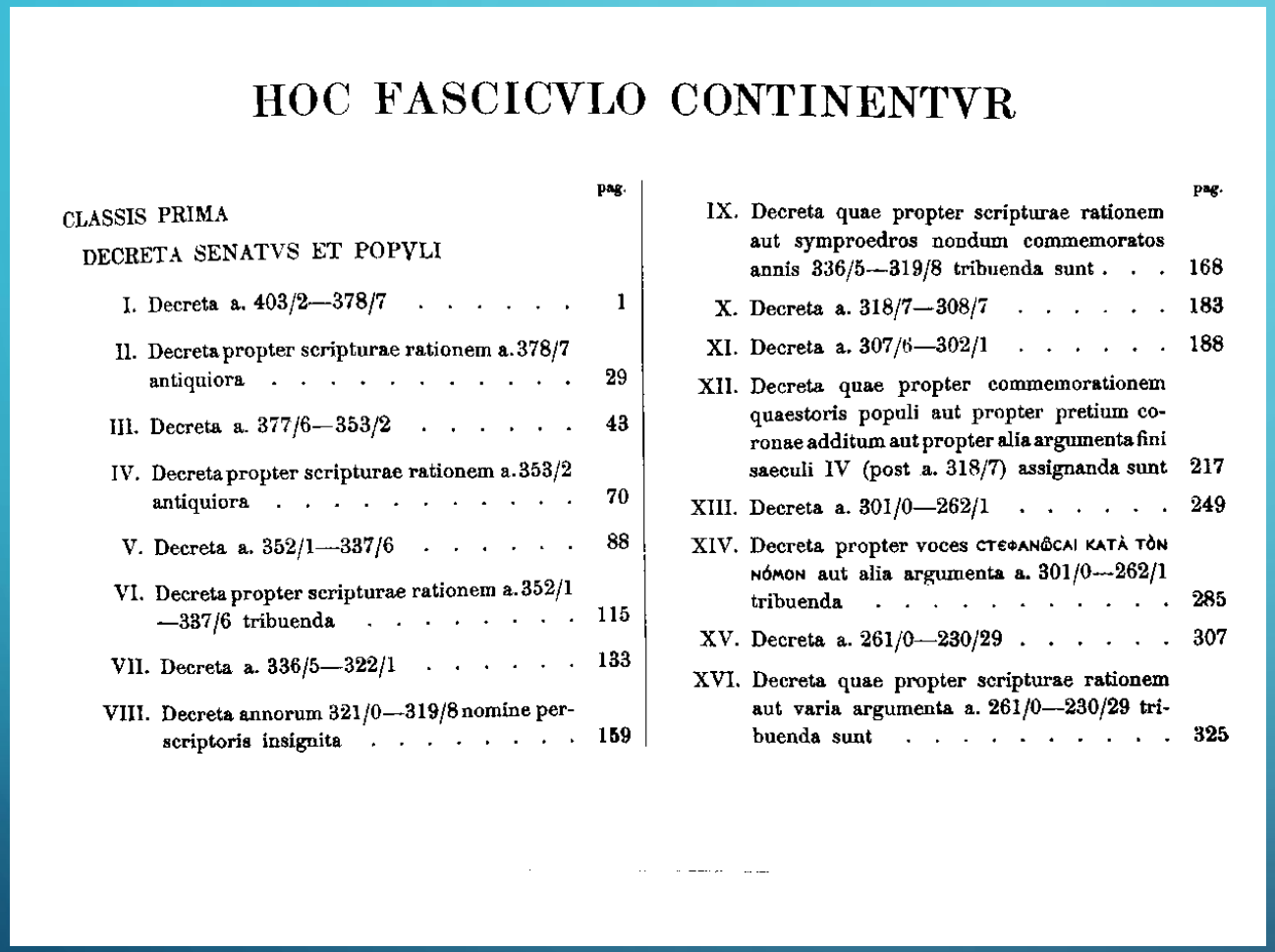
KRATEROS

- Same questions as Tracy et al. and Reconsidering the Roman Workshop
  - How to make use of existing photosets
  - Scene detection? Binarization?
- 3D mesh files present additional options:
  - “Gigamesh” (gigamesh.eu) for annotation and binarization
  - Cutting edge 3D scene detection



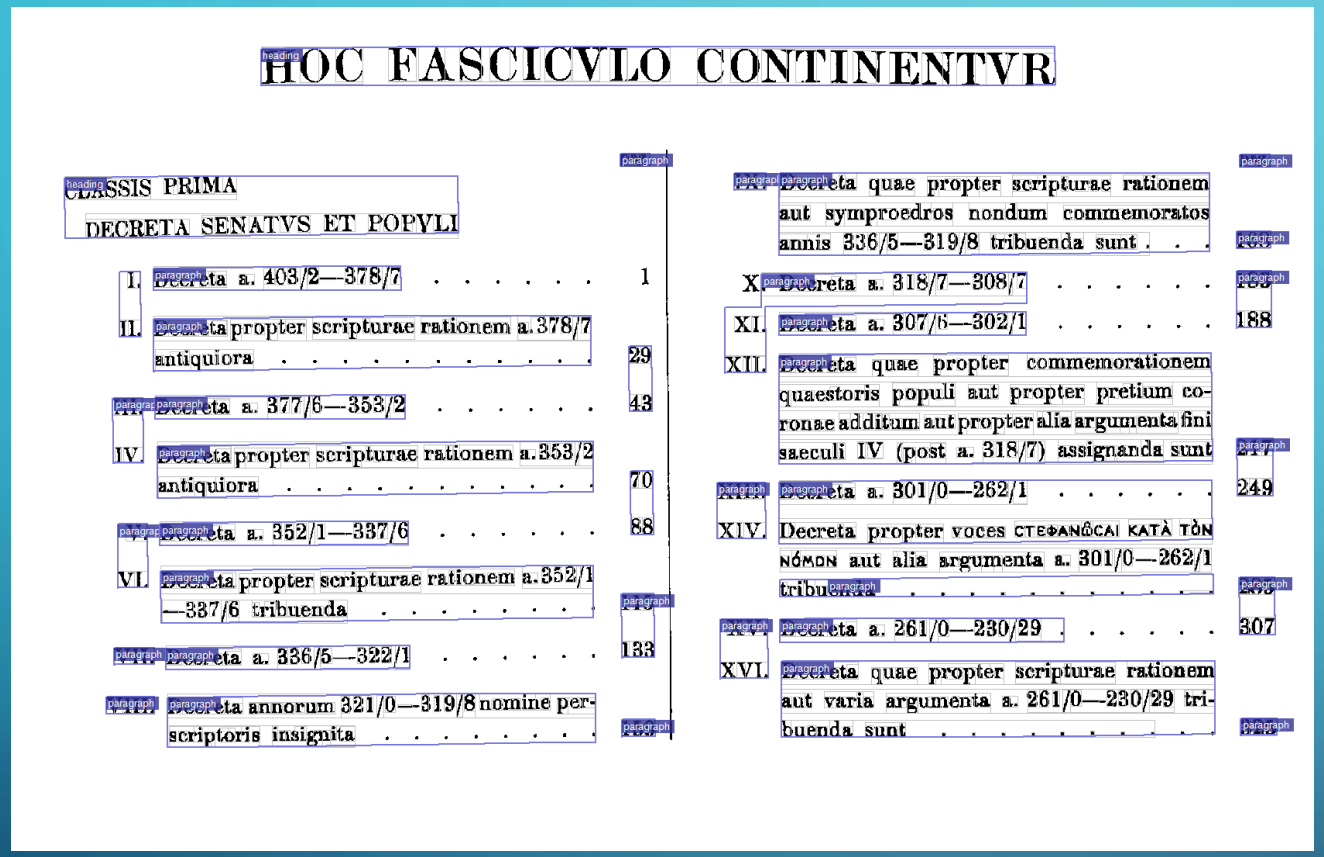
HOC FASCICVLO CONTINENTVR

PRINT OCR ("TESSERACT" ENGINE)



HOC FASCICVLO CONTINENTVR

PRINT OCR ("Tesseract" Engine)

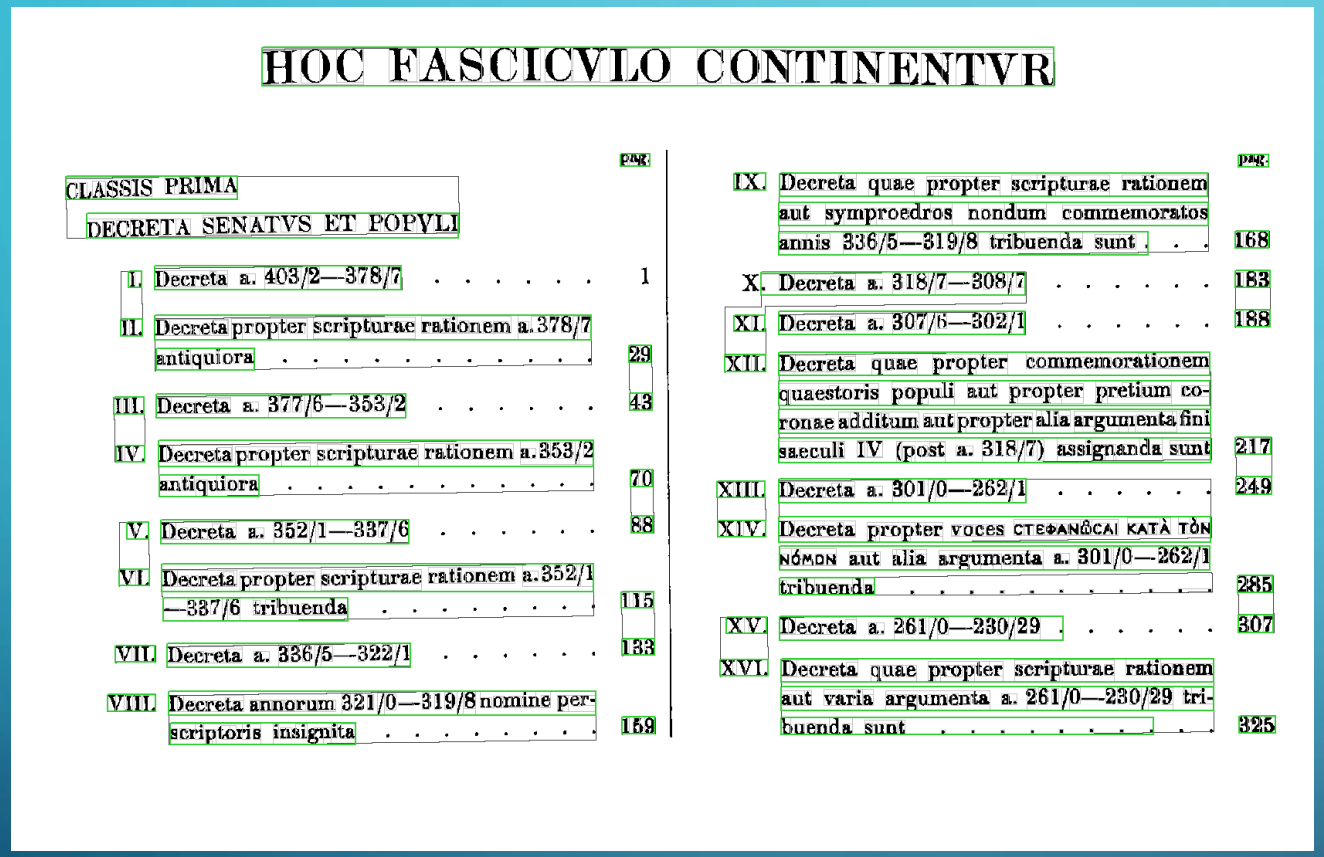


HOC FASCICVLO CONTINENTVR

CLASSIS PRIMA

DECRETA SENATVS ET POPVLI

PRINT OCR ("Tesseract" Engine)



HOC FASCICVLO CONTINENTVR

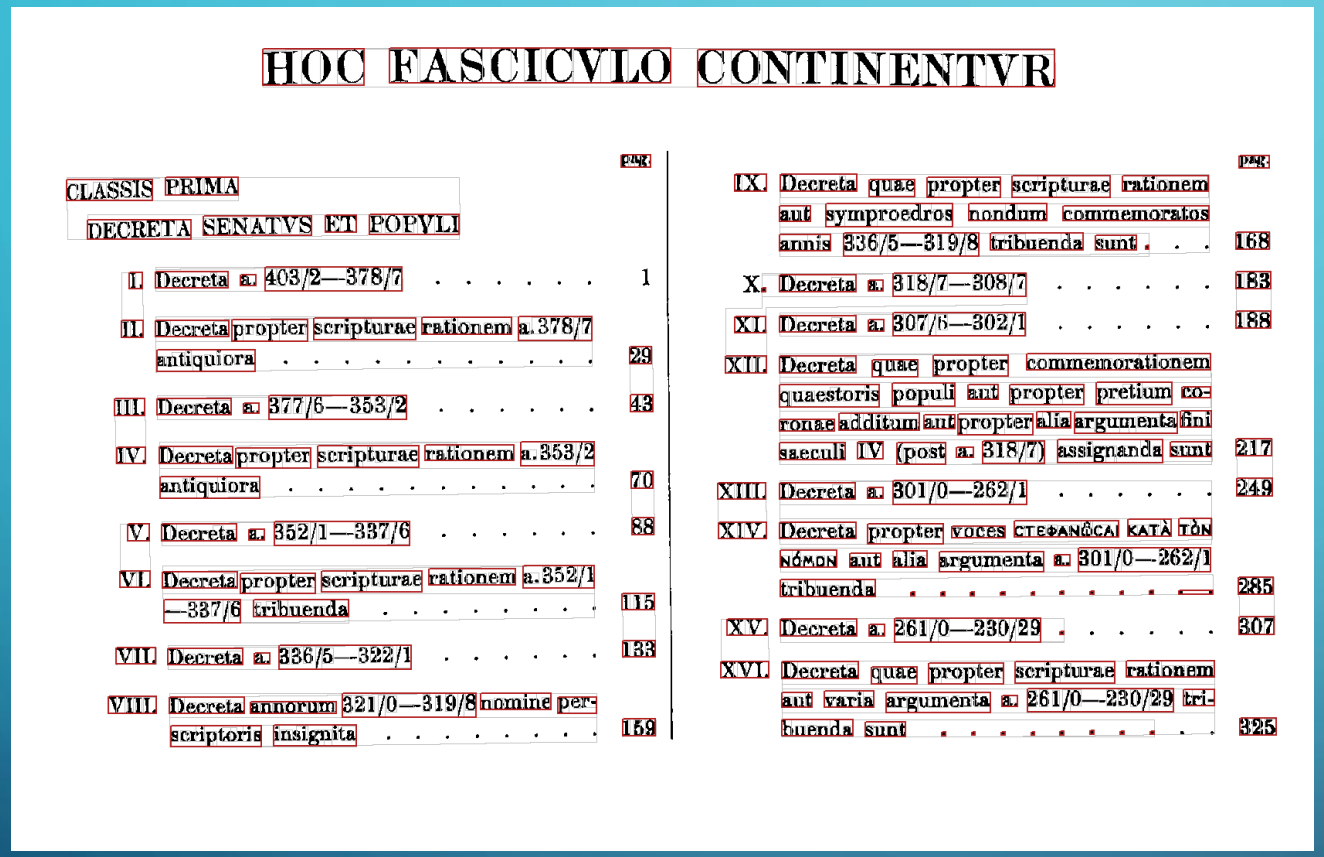
pag.

CLASSIS PRIMA

DECRETA SENATVS ET POPVLI

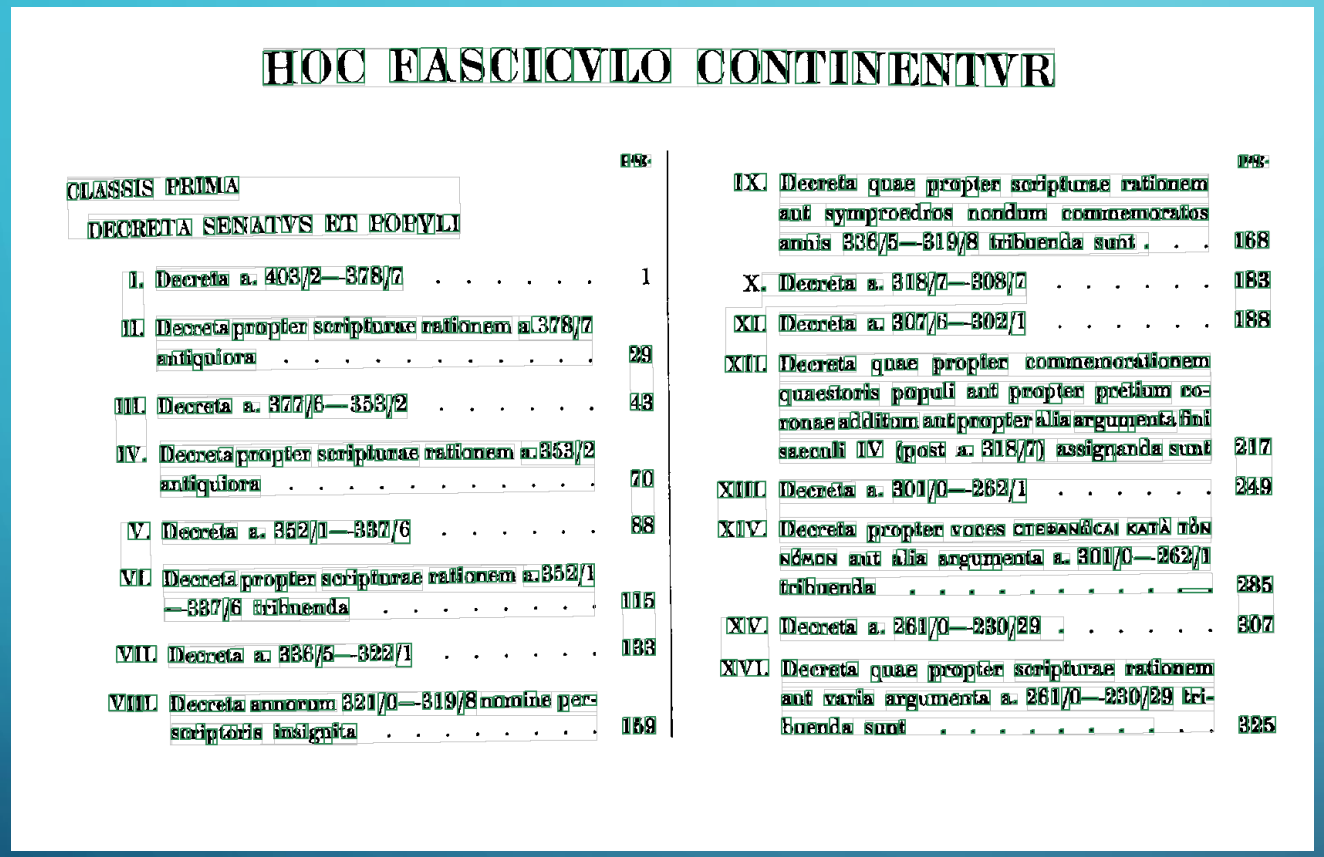
pag.

PRINT OCR ("Tesseract" Engine)



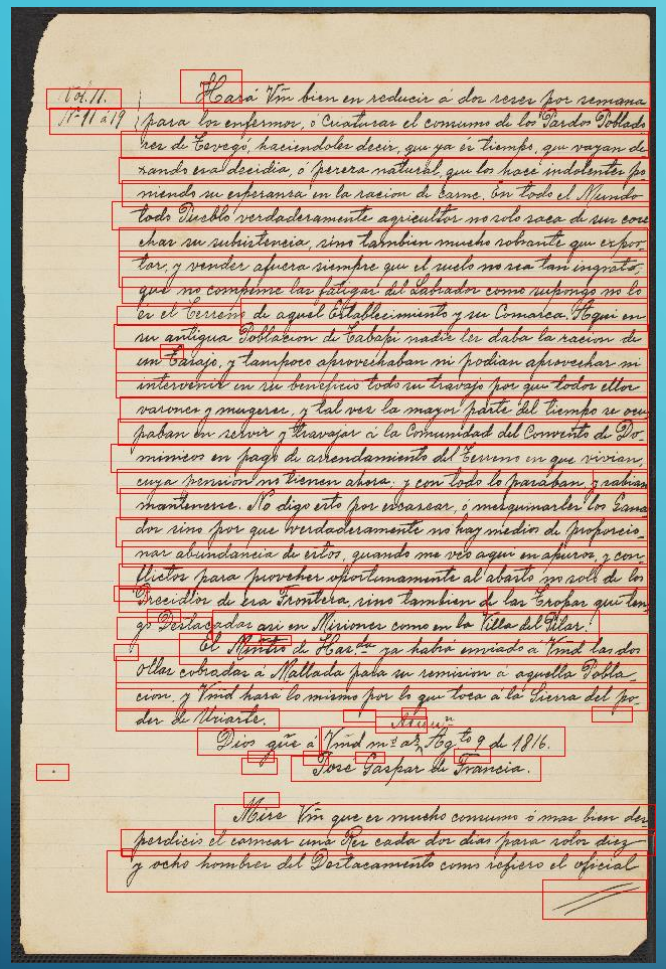
HOC FASCICVLO CONTINENTVR

PRINT OCR ("Tesseract" Engine)

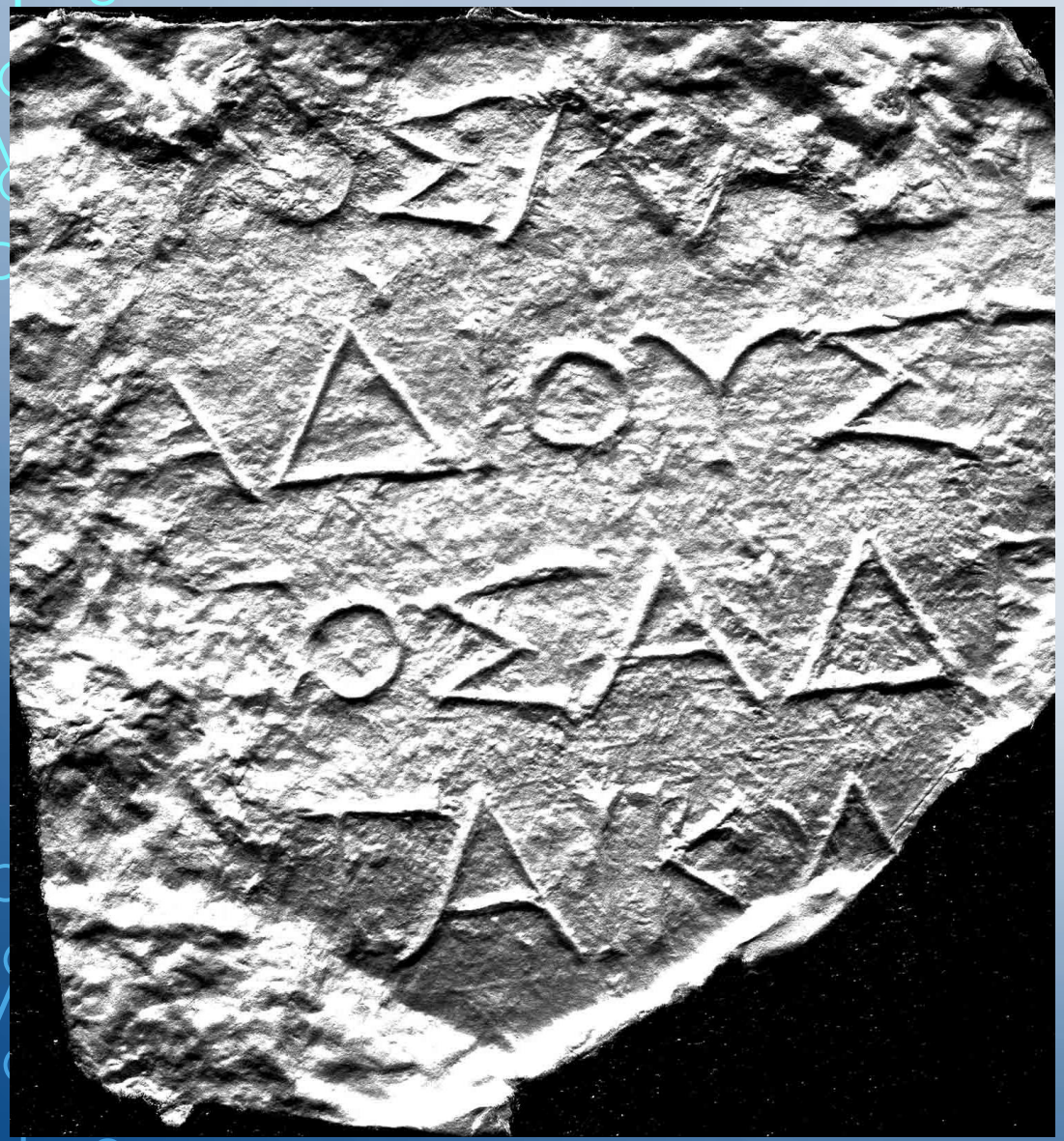


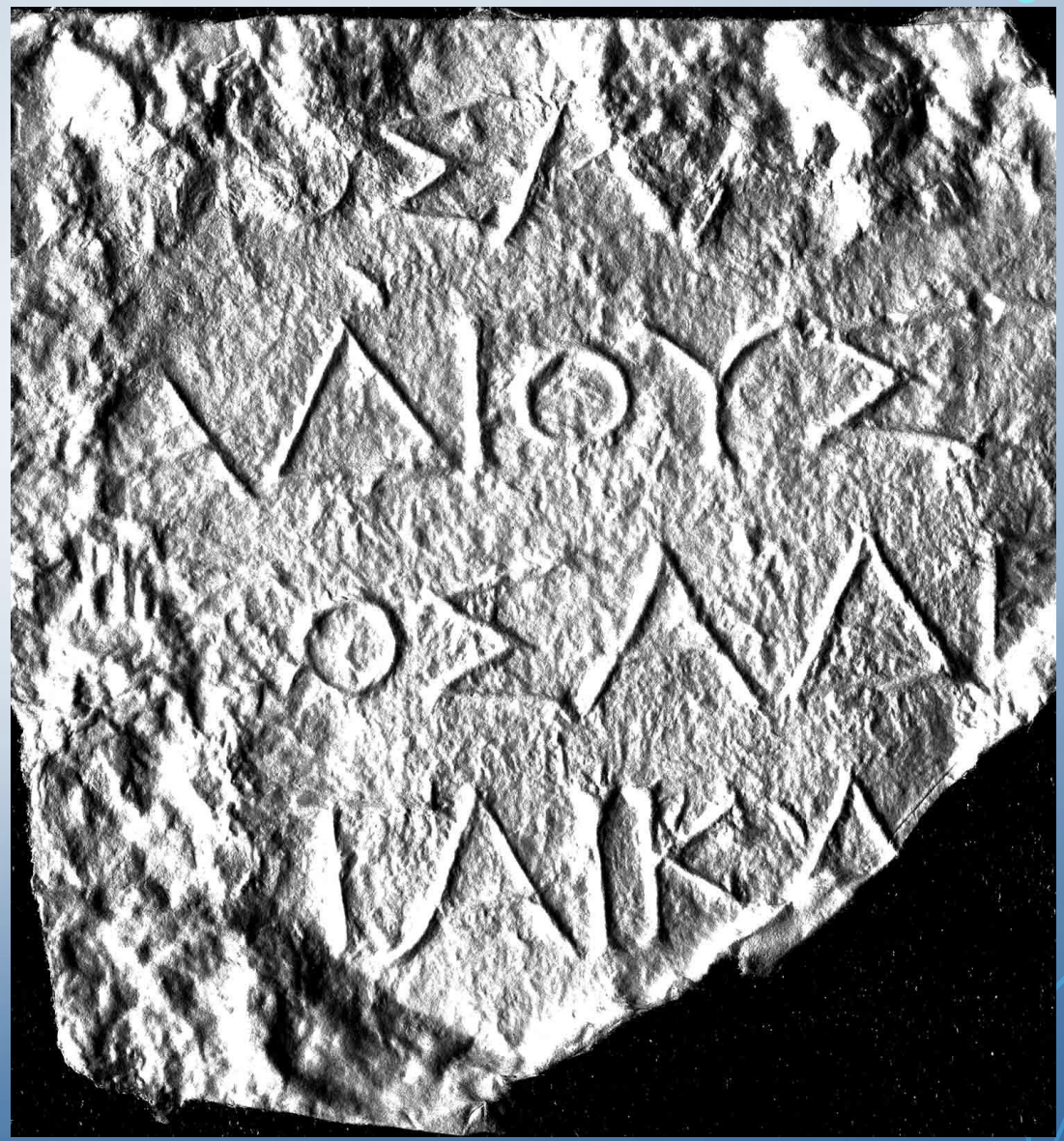
HOC FASCICVLO CONTINENTVR

PRINT OCR ("Tesseract" Engine)



MANUSCRIPT OCR
("KRAKEN" ENGINE)



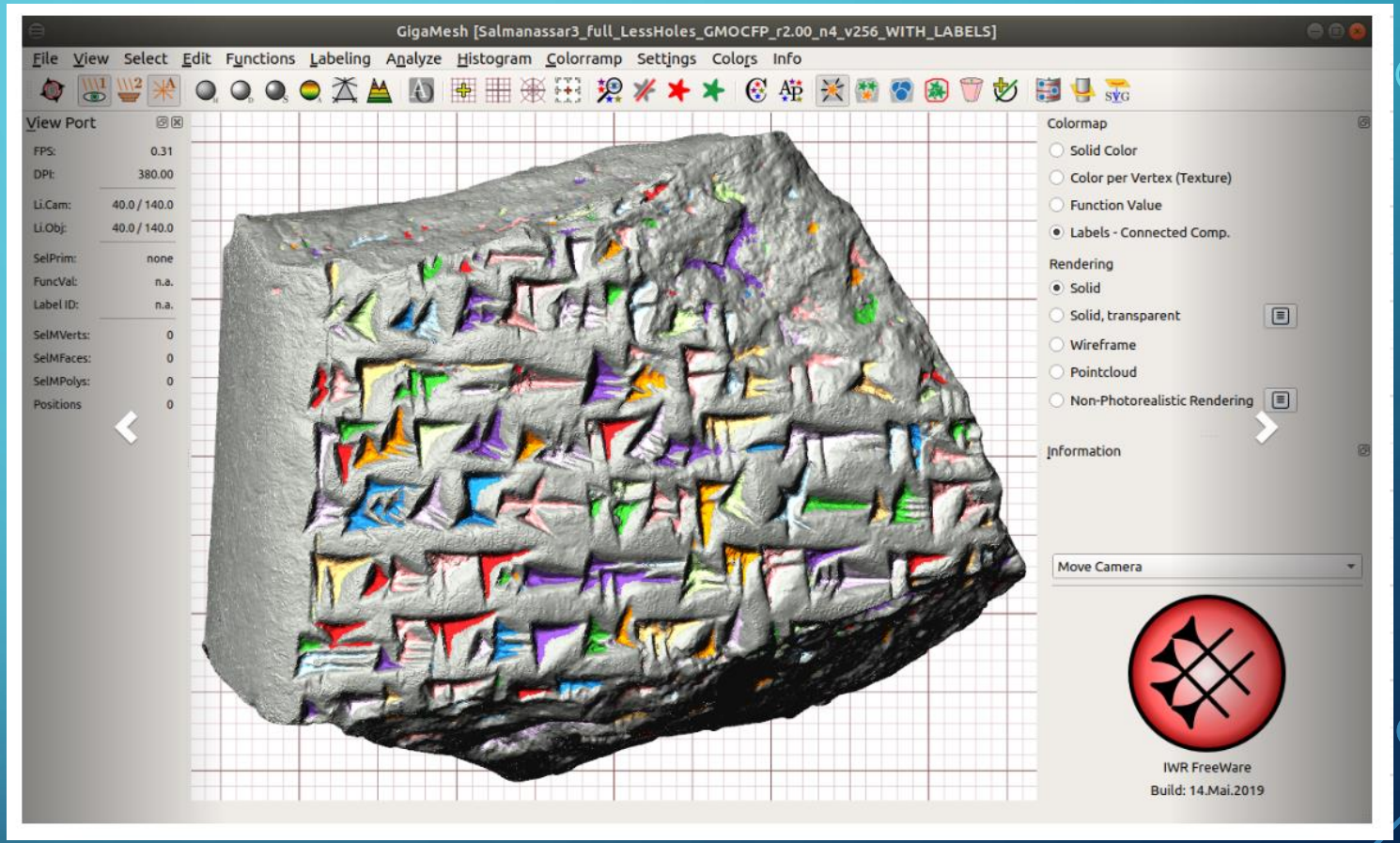


KRATEROS

- Same questions as Tracy et al. and Reconsidering the Roman Workshop
  - How to make use of existing photosets
  - Scene detection? Binarization?
- 3D mesh files present additional options:
  - “Gigamesh” (gigamesh.eu) for annotation and binarization
  - Cutting edge 3D scene detection

GIGAMESH

- Annotation and pairwise matching of alphabetic symbols
- Potential to use depth map for binarization



3D SCENE DETECTION

- Google Research Tensorflow3D
  - https://ai.googleblog.com/2021/02/3d-scene-understanding-with-tensorflow.html
  - https://github.com/google-research/google-research/tree/master/tf3d
  - Deep learning (like Ithaca)
  - Symbiotic with advances in self-driving vehicles (3D space parsing)

FINAL THOUGHTS

- Exciting potential in epigraphic text- and image-based machine learning
  - Huge data sets, but sometimes requiring substantial manipulation/cleaning
  - Possibility to significantly speed up (and improve) core epigraphic tasks
  - Discoveries in this area can be directly transferred into other areas
- Critical caveats
  - As always, funding is a limiting factor
  - Important to avoid framing ML as “superior alternative” to human expertise

THANK YOU!

The Krateros Project gratefully acknowledges the support of the National Endowment for the Humanities, the Charles and Lisa Simonyi Fund for Arts and Sciences, as well as support in memory of Fowler Merle-Smith.