AUTOMATICALLY IDENTIFYING JOIN CANDIDATES IN THE CAIRO GENIZAH

Lior Wolf, Nachum Dershowitz and students
The School of Computer Science, Tel-Aviv University

Roni Shweka, Yaacov Choueka
The Friedberg Genizah Project, Jerusalem

Cairo Genizah

a collection containing ~250,000 fragments of mainly Jewish texts discovered in the late 19th century

discarded codices, scrolls, and documents, written mainly in the 10th to 15th centuries

spread out in tens of libraries and private collections worldwide

enormous impact on 20th century scholarship in a multitude of fields

The Friedberg Genizah Project

a philanthropically-funded project to digitally photograph and organize all Genizah fragments

cataloging the fragments

sharing all data on-line

Cambridge (CUL)

mosseri

Cambridge (Westminster)

London (BL)

Oxford

Manchester

Birmingham

Jerusalem (JNUL)

Tel-Aviv

Haifa

Ramat-Gan

New York (JTG)

Cincinnati (HUC)

Washington (Freer)

Toronto

Philadelphia (CAJS)

Paris (AIU)

Strasbourg (BN)

Vienna

Geneva

Moscow

Kiev

Budapest

St Petersburg (RNL)

Ann Arbor, Basel, Berlin, Copenhagen,
Dublin, Montreal, Munich, Princeton, Schoeyen, Turin

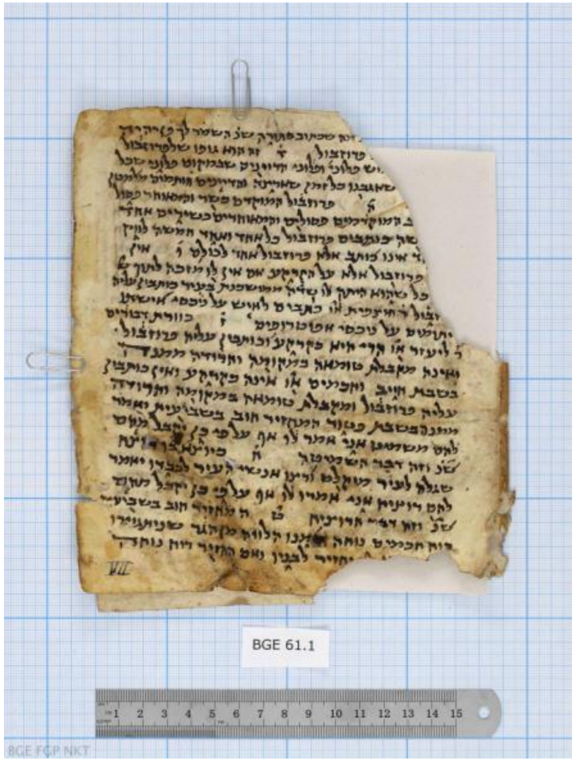
Basic notion: join

A **join** is a set of manuscript-fragments that are known to originate from the same original work.

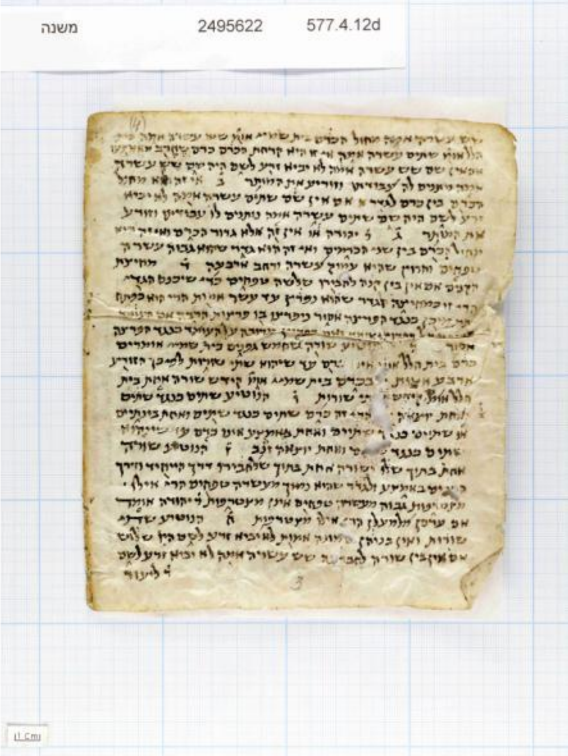
Known joins are documented in catalogs

Catalogs (very partial list)	Entries
Adler, Elkan Nathan Catalogue of Hebrew Manuscripts in the Collection of Elkan Nathan Adler., Cambridge, 1921	1026
Cowley, Arthur Ernest Photocopy of Unpublished Typescript Catalogue of the Hebrew Manuscripts in the Bodleian Library	1318
Gottstein, M.H. "Hebrew Fragments in the Mingana Collection," Journal of Jewish Studies V (1954), 1954	40
Halper, Benzion Descriptive catalogue of Genizah fragments in Philadelphia, Philadelphia, 1924	487
Lutzki, Morris Catalogue of Biblical Manuscripts in the Library of the Jewish Theological Seminary, Photocopy of Unpublished Typescript (New York: JTS)	927
Neubauer, Adolf, Cowley, Arthur Ernest Catalogue of the Hebrew Manuscripts in the Bodleian Library, Vol. II, Oxford, 1886-1906	2199
Reif, Stefan C. Hebrew manuscripts at Cambridge University Library, Cambridge, 1997	126
Schwab, Moise "Les Manuscrits du Consistoire Israelite de Paris Provenant de la Gueniza du Caire," REJ LXII (1911), pp. 107-119, 267-277; LXIII (1911), pp. 100-120, 276-296; LXIV (1912), pp. 118-141., ., 1911-1912	1896
Schwarz, A.Z. , Loewinger D.S. and Roth E. Die hebraieschen handschriften in Oesterreich (ausserhalb der Nationalbibliothek in Wien), New York	185
Wickersheimer, Ernest Catalogue général des manuscrits des bibliothèques publiques de France. Départements, Tome XLVII : Strasbourg, Paris, 1923	3
Worman, E.J. Hand-list of pieces in Glass of theTayler-Schechter Collection. Photocopy of Unpublished Handwriting	2291
Worrell, William Hoyt, Gottheill, Richard James Horatio	50

Example of our discoveries: Mishnah

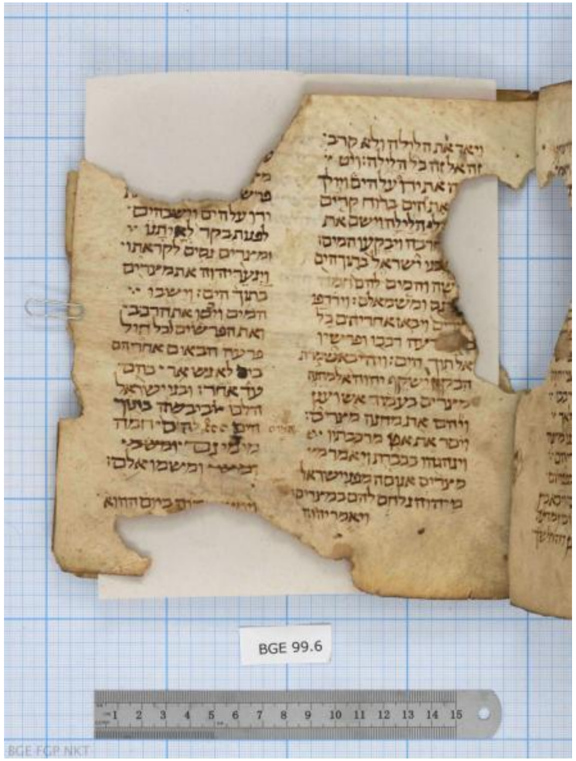


Geneva

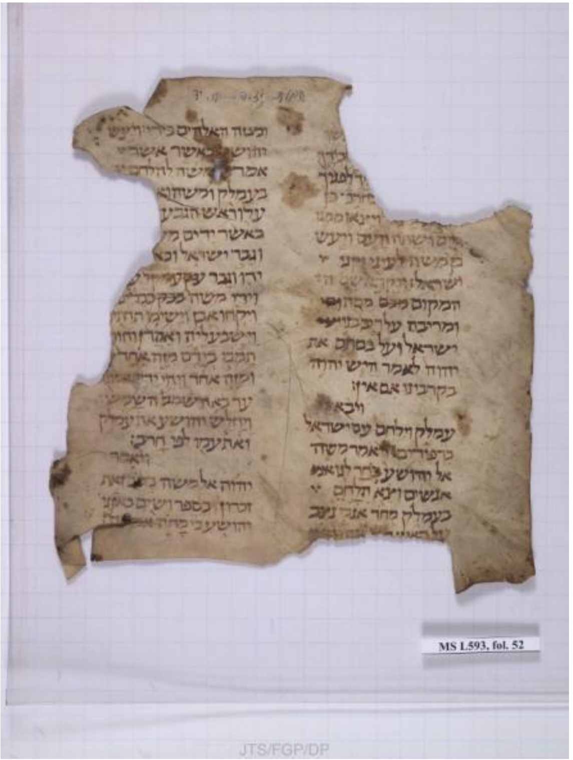


Jerusalem

Example of our discoveries: Bible (square letters)

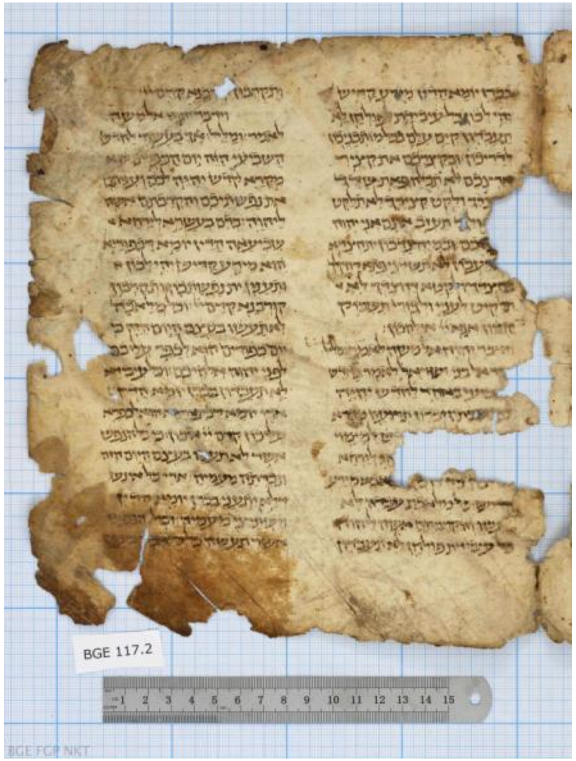


Geneva

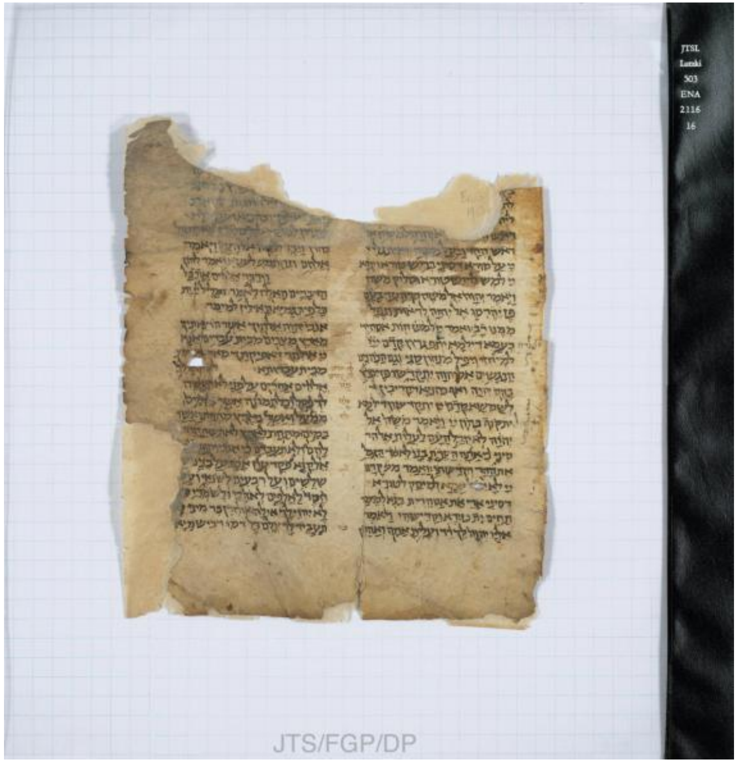


New york

Example of our discoveries: Bible (Aramic)

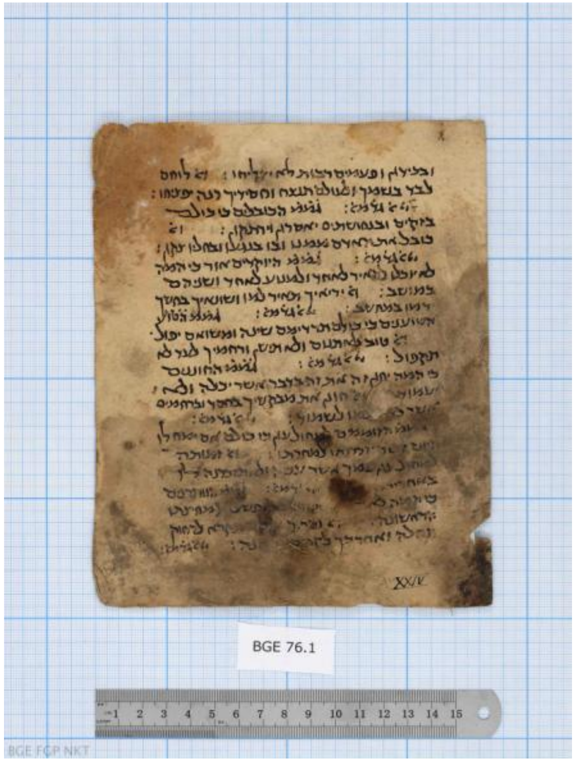


Geneva

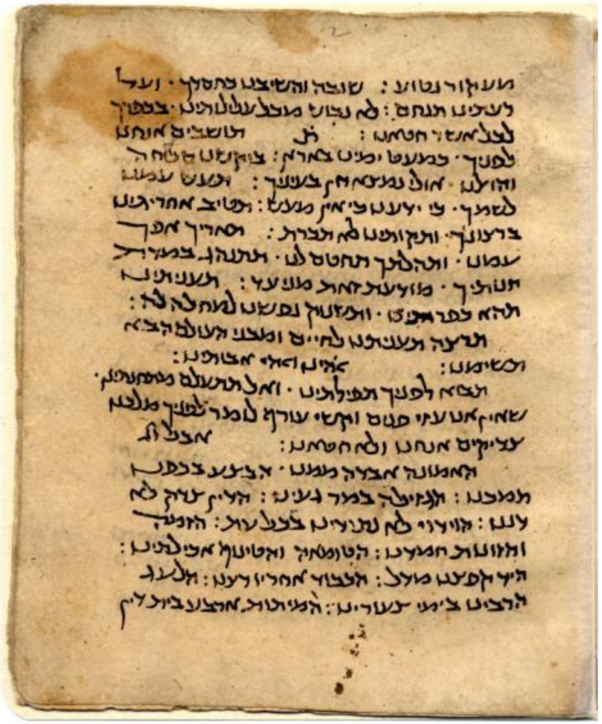


Vienna

Example of our discoveries: Liturgy

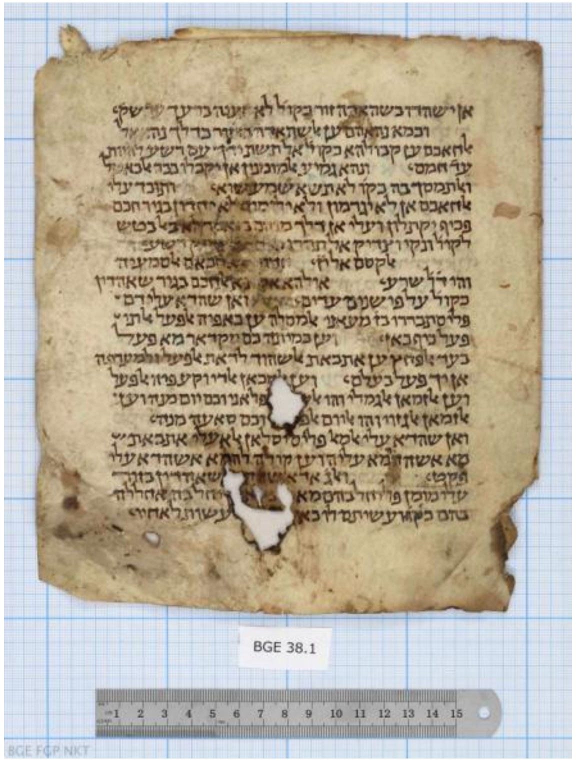


Geneva

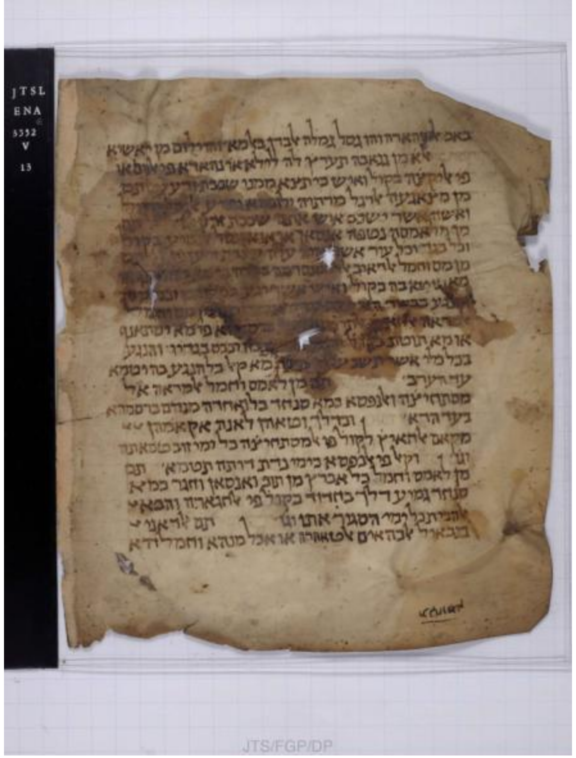


Pennsylvania

Example of our discoveries: Lost halakhic monograph of Rav Saadya Gaon (10 cent.) in judeo-arabic

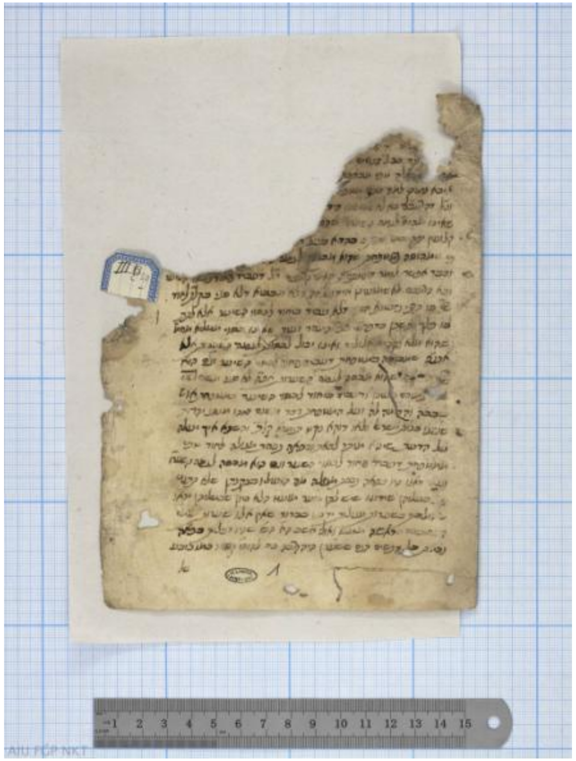

BGE 38.1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
BGE FGP NKT

Geneva

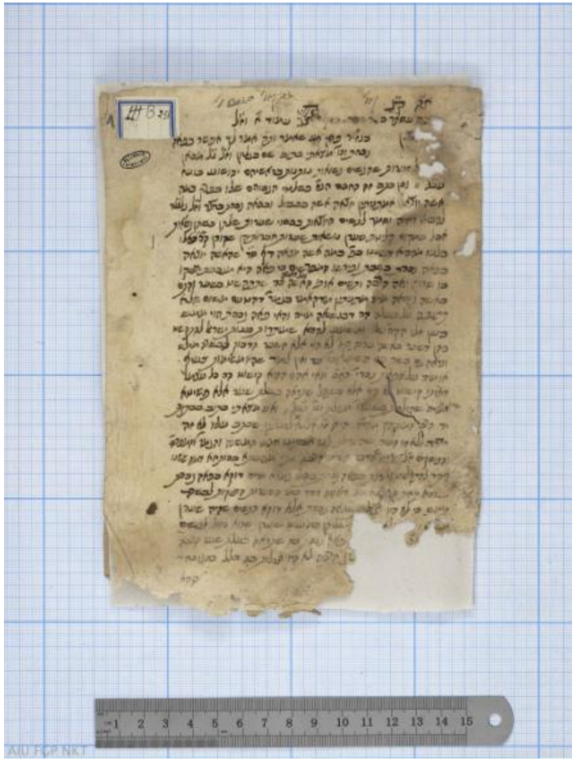

JTSL
ENA
5352
V
13
JTS/FGP/DP

New York

Example of our discoveries: halakhic responsa



Paris (shelf 0002134)



Paris (shelf 0001272)

Related work: writer identification

Much of the existing work is for Latin letters

Typical pipeline:

preprocessing -> segmentation -> letter-based matching

Another pipeline:

collect global statistics

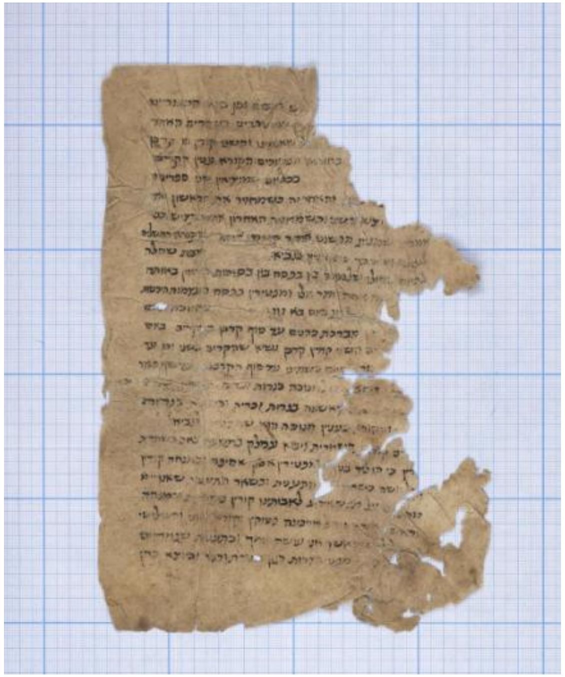
Hebrew letters are not connected

Here: focus on joins rather than on identifying a writer among a list of potential writers

Where to start?

Preprocessing the images

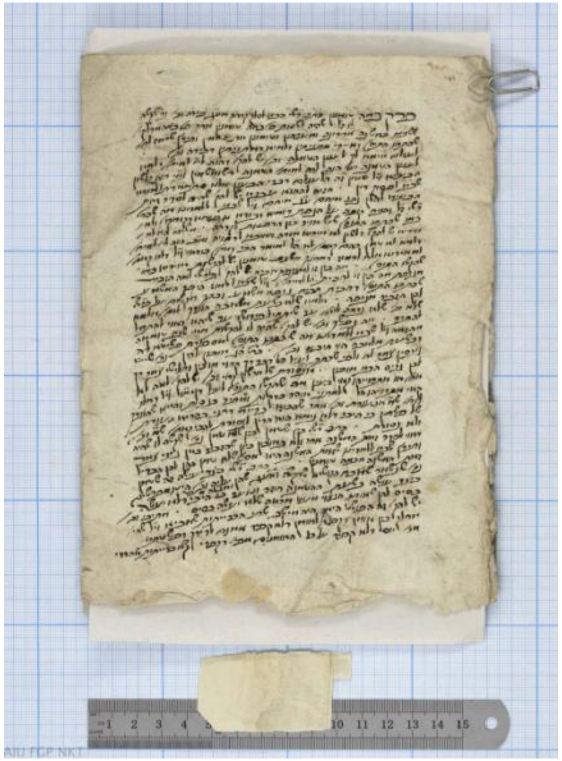
Foreground segmentation



Preprocessing the images

Foreground segmentation

Removal of rulers

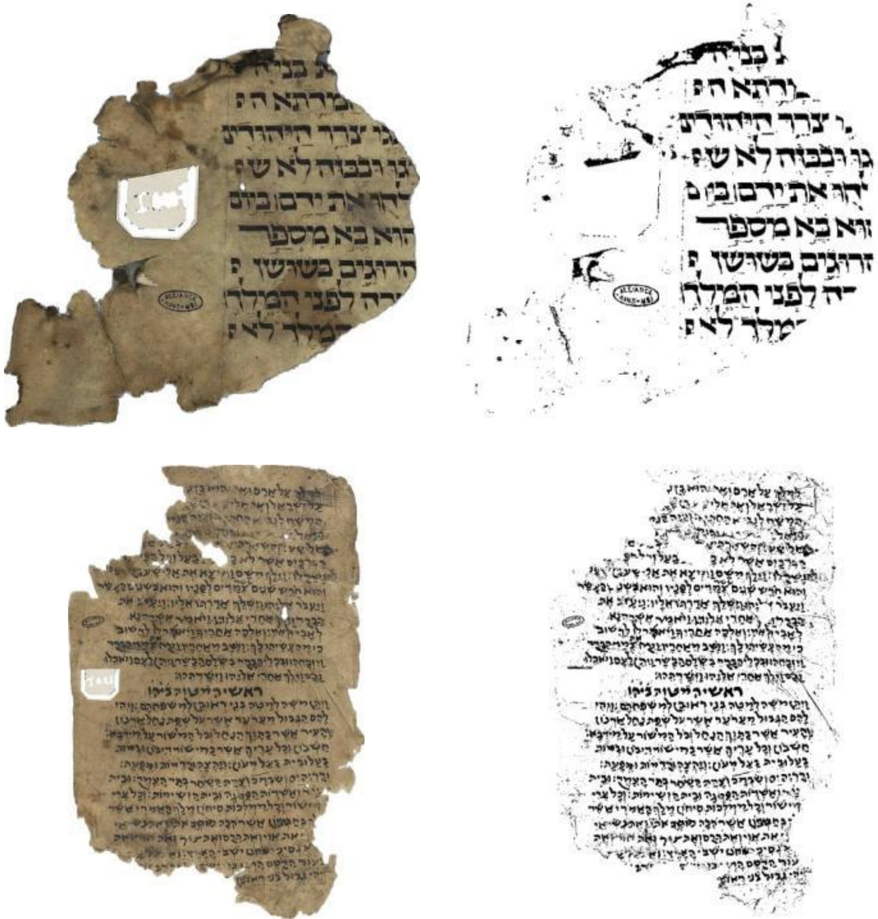


Preprocessing the images

Foreground segmentation

Removal of rulers

Binarization



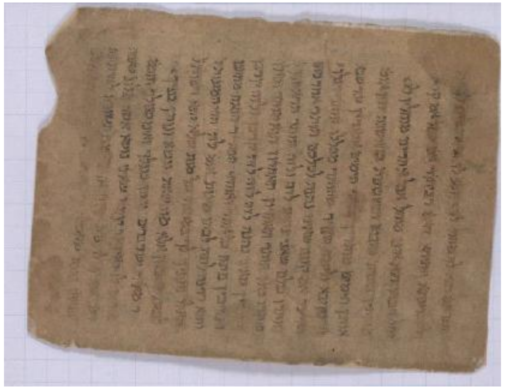
Preprocessing the images

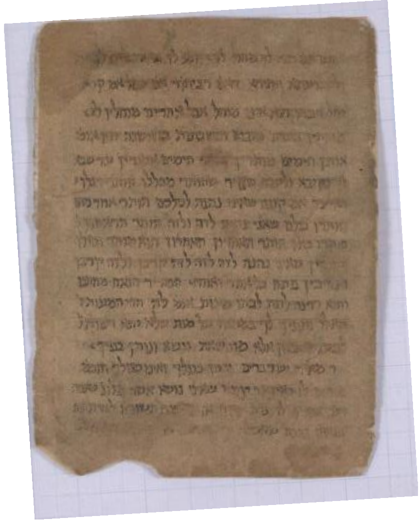
Foreground segmentation

Removal of rulers

Binarization

Auto-alignment





Preprocessing the images

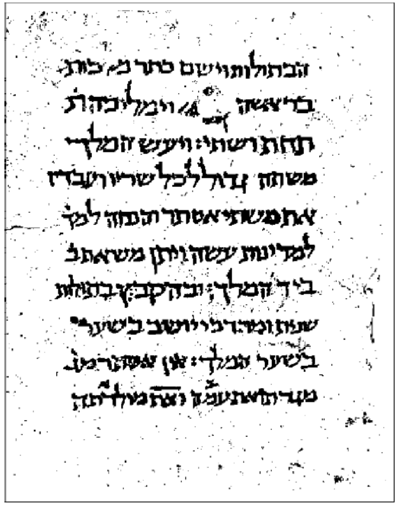
Foreground segmentation

Removal of rulers

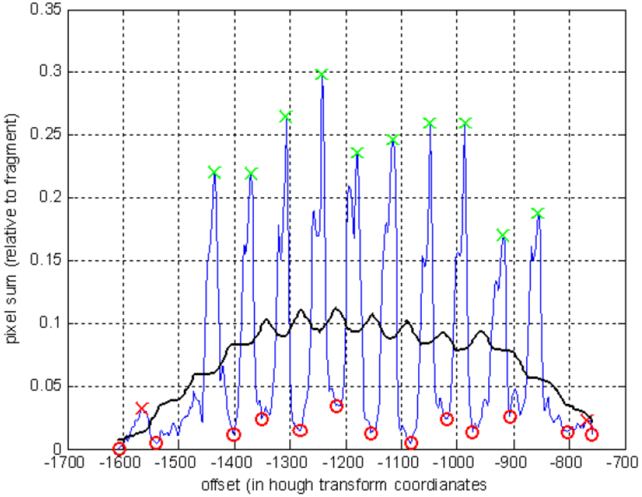
Binarization

Auto-alignment

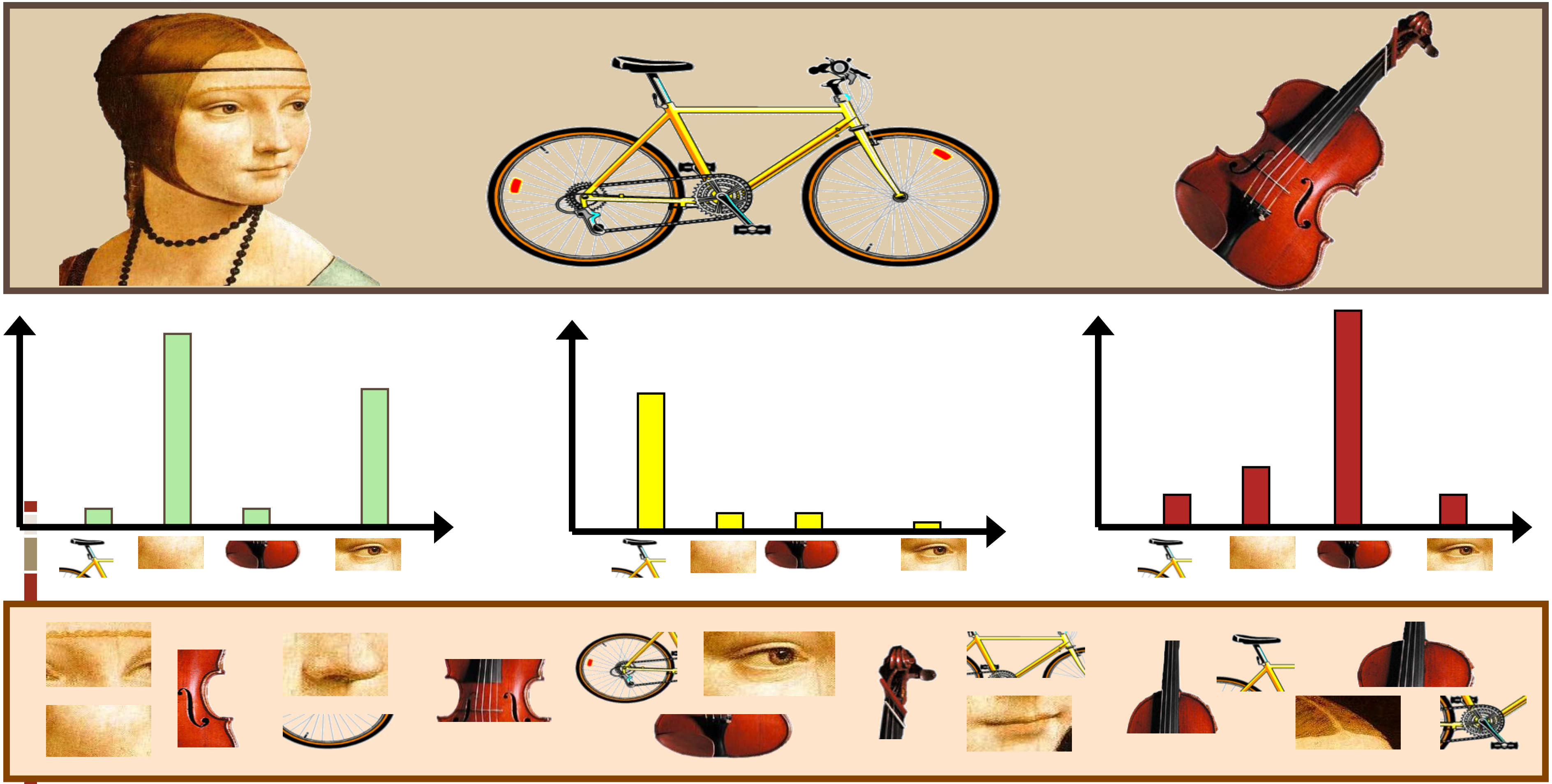
Taking physical measurements



הבתולות וישם כתר מלכות
בראשה וימליכה ת
תחת ושתי: ויעש המלך
משתה גדול לכל שריו ועבדיו
את משתה אסתר והנחה למ
למדינות עשה ויתן משאת
ביד המלך: ובהקבץ בתולות
שנית ומרדכי יושב בשער
בשער המלך: אין אסתר מג
מגדת את עמה ואת מולדתה



Bag of keypoints representation



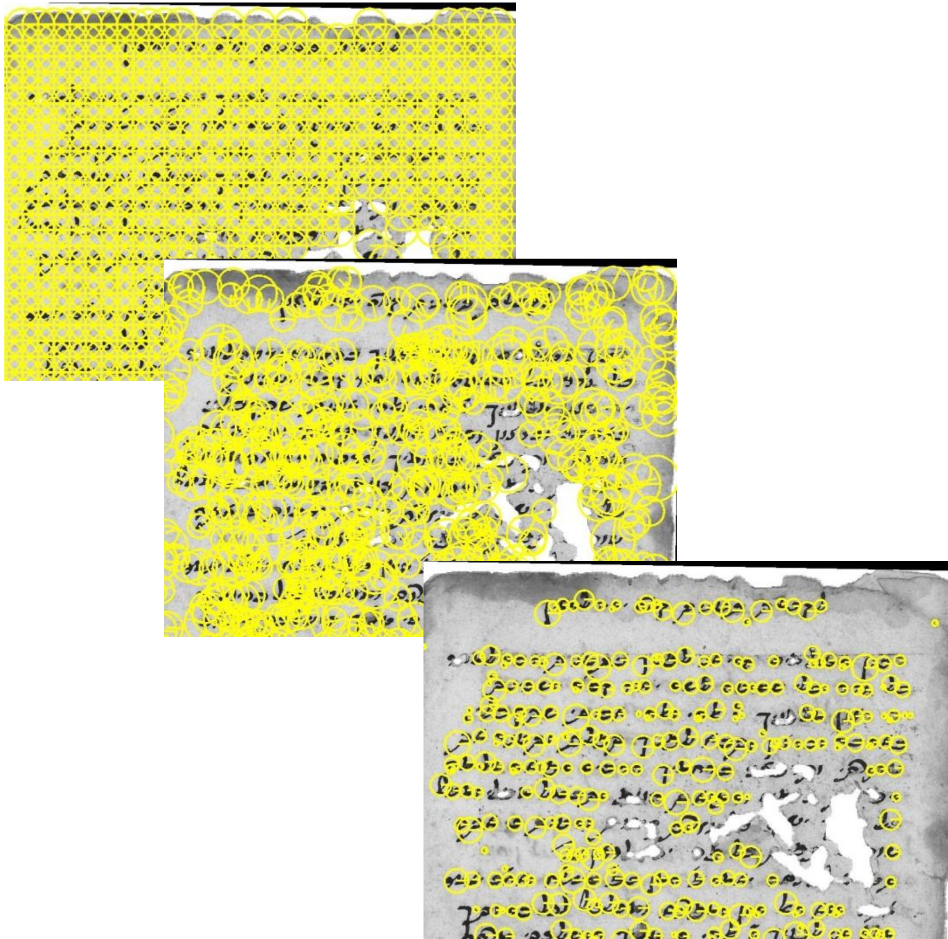
Slide images credit: Li Fei-Fei

Selecting keypoints

Grid points

Lowe's SIFT DoG detector

Proposed Connected components based



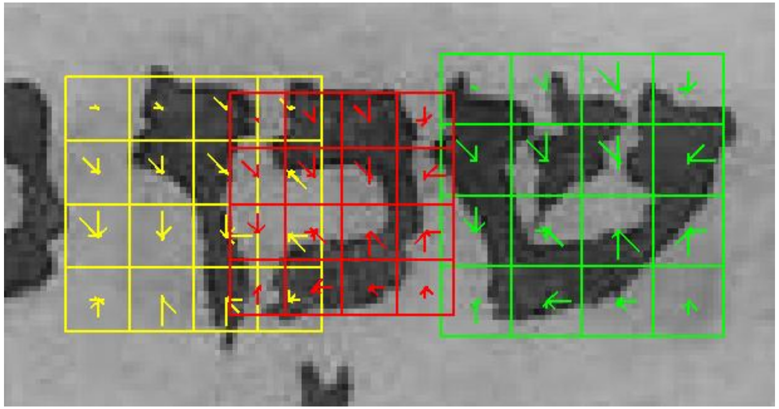
Local descriptors

SIFT

PCA-SIFT

Patch binary values fit to 32x32

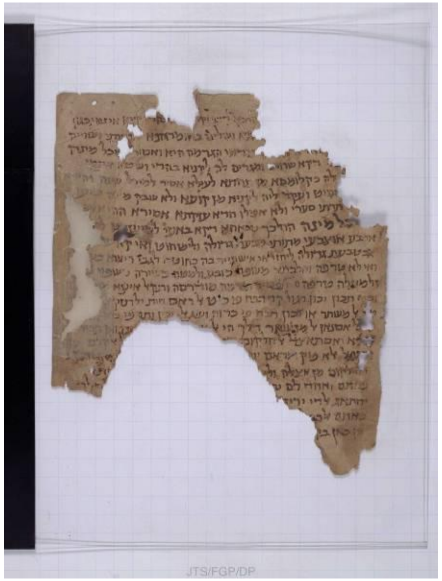
Patch Binary stretched to a height of 32 pixels



Dictionary learning

Used a separate set of 150 images, from which 20,000 keypoints were extracted, and clustered by using k-means

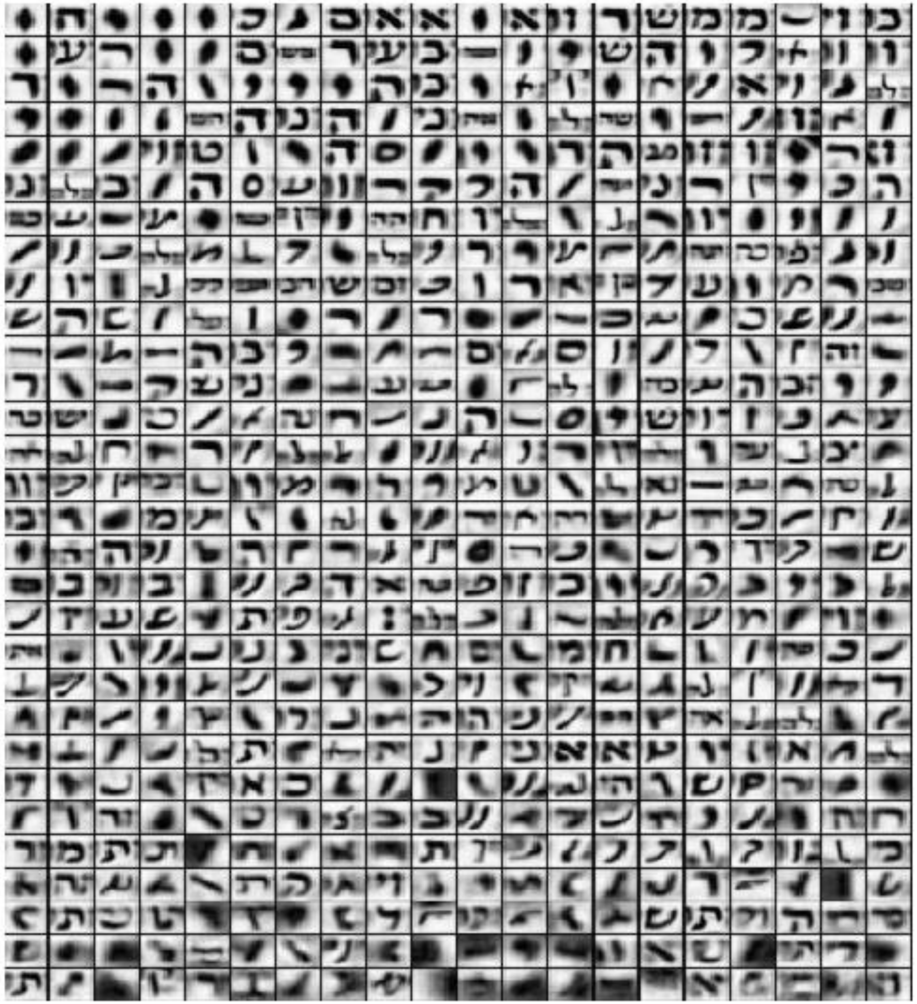
Performance is stable once k>400



Dictionary learning

Used a separate set of 150 images, from which 20,000 keypoints were extracted, and clustered by using k-means

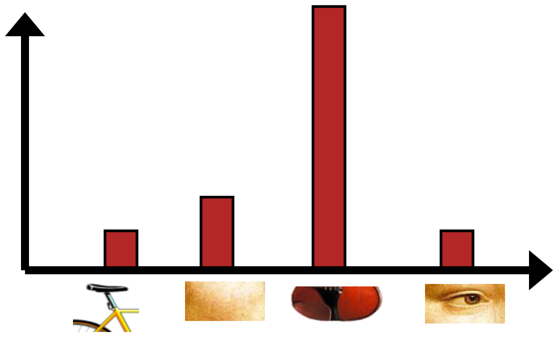
Performance is stable once k>400

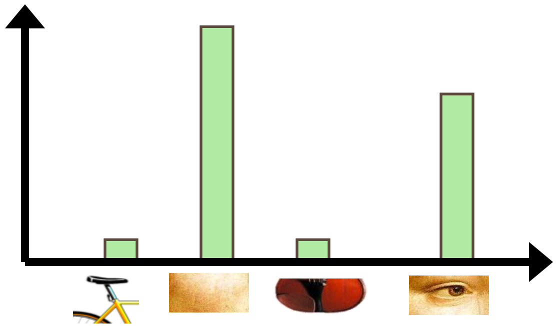


Vectorization

Histograms: counting the number of image keypoints of each type. Normalized by either L2 or L1

Minimal distance over all image keypoints to each dictionary word





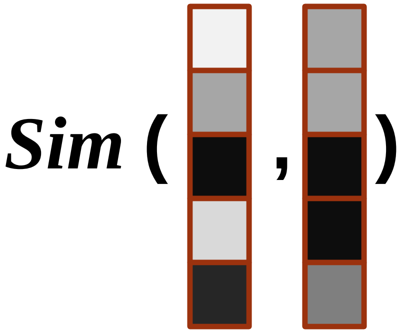
Similarity computation

Take 2 vectors (v_1,v_2) return a similarity value κ_{12}.

Ideally, there is a threshold such that

$\kappa_{12} > \theta$ ➔ join

$\kappa_{12} < \theta$ ➔ not a join



Methods used:

L2 distance of vectors

L1 distance of vectors

Hellinger norm

SVM of vector of absolute differences $\sum_i w_i \, | a_i - b_i |$

OneShot Similarity (ECCV*LFW*’08,ICCV’09)

Computing the “One-Shot” Similarity

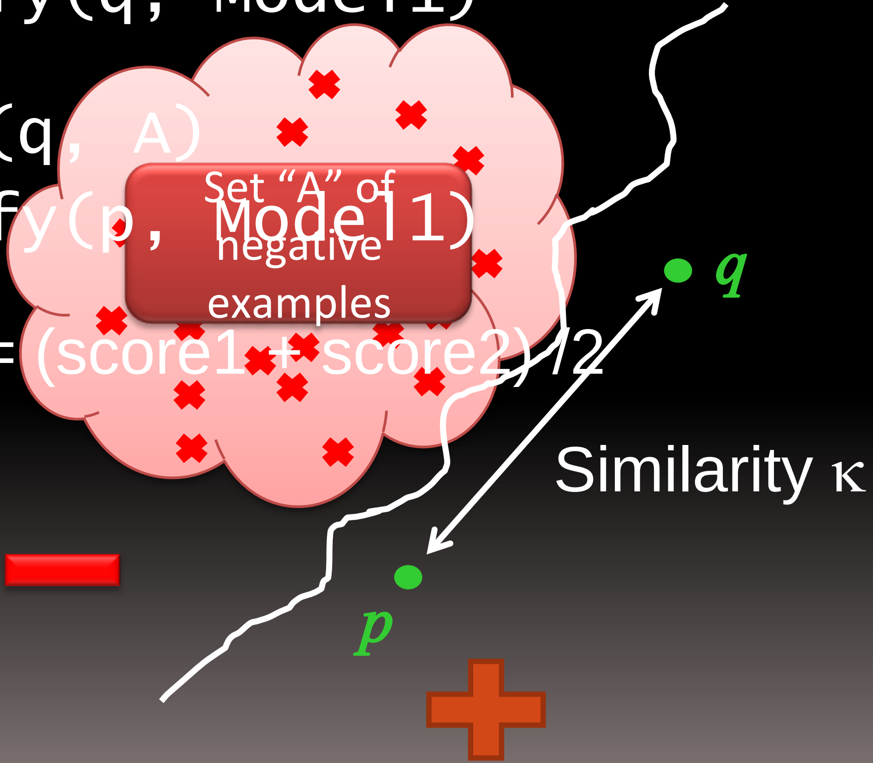
Step a: `Model1 = train (p, A)`

Step b: `Score1 = classify(q, Model1)`

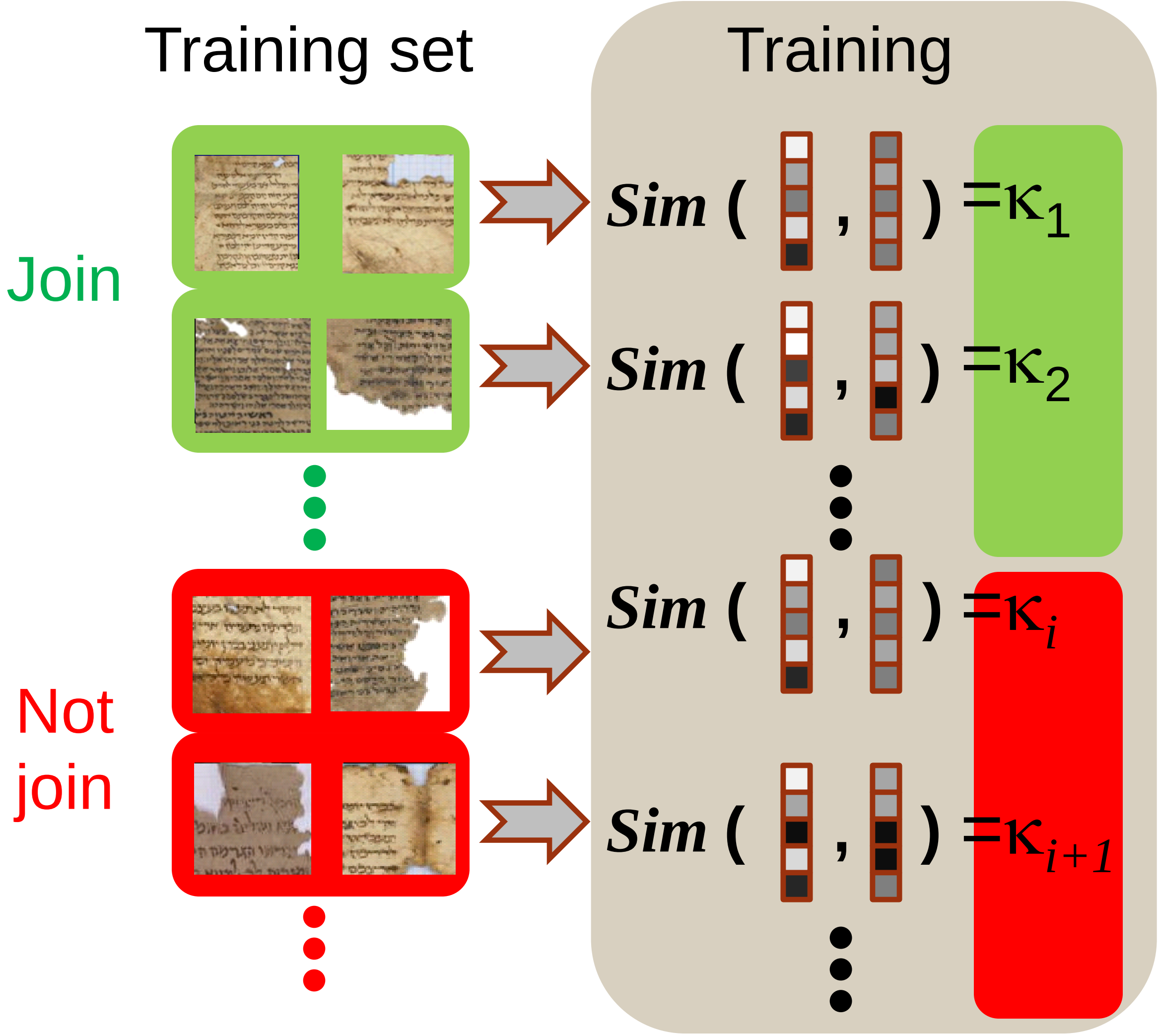
Step c: `Model2 = train (q, A)`

Step d: `Score2 = classify(p, Model1)`

`One-Shot-Sim` = (score1 + score2) /2



One similarity/vector/descriptor



Training set

Training

Join

$Sim(\ ,\ ) = \kappa_1$

$Sim(\ ,\ ) = \kappa_2$

Not join

$Sim(\ ,\ ) = \kappa_i$

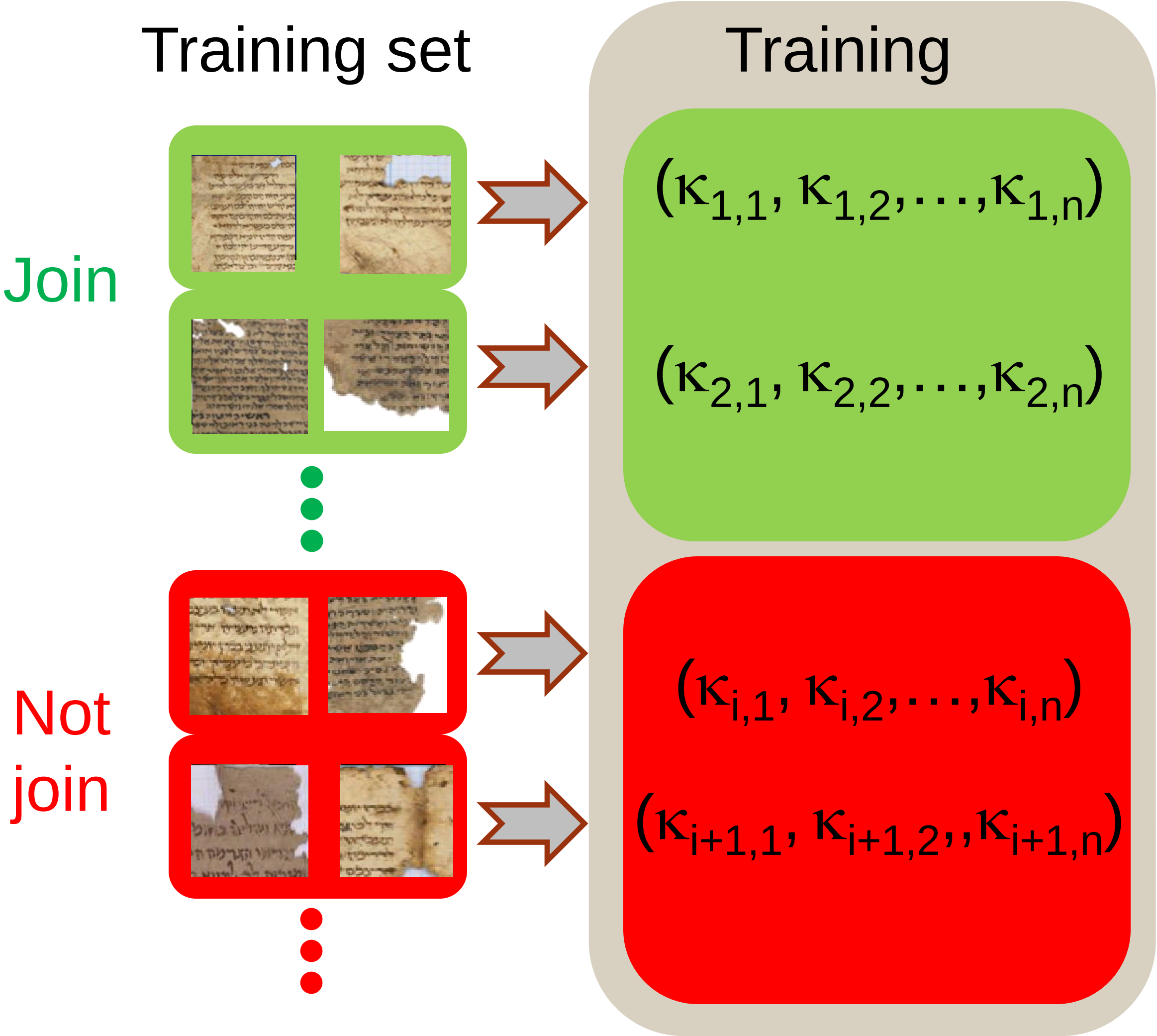
$Sim(\ ,\ ) = \kappa_{i+1}$

(a) Compute global descr. for images

(b) Measure descr. Similarity (e.g., L2 norm)

(c) Train classifier (e.g., SVM) to threshold join from not-join

Multi similarities/vectors/descriptors



Training set

Training

Join

$(\kappa_{1,1}, \kappa_{1,2},\dots,\kappa_{1,n})$

$(\kappa_{2,1}, \kappa_{2,2},\dots,\kappa_{2,n})$

Not join

$(\kappa_{i,1}, \kappa_{i,2},\dots,\kappa_{i,n})$

$(\kappa_{i+1,1}, \kappa_{i+1,2},,\kappa_{i+1,n})$

Train with a vector of similarity values

Combine with SVM

Our benchmark

Modeled after the Labeled Faces in the Wild benchmark

Actually two benchmarks:

View 1 – used to tune parameters

View 2 – used to test performance

~30000 leaves

<u>View 1:</u>

3 splits

each 1000 positives, 2000 negatives

<u>View 2:</u>

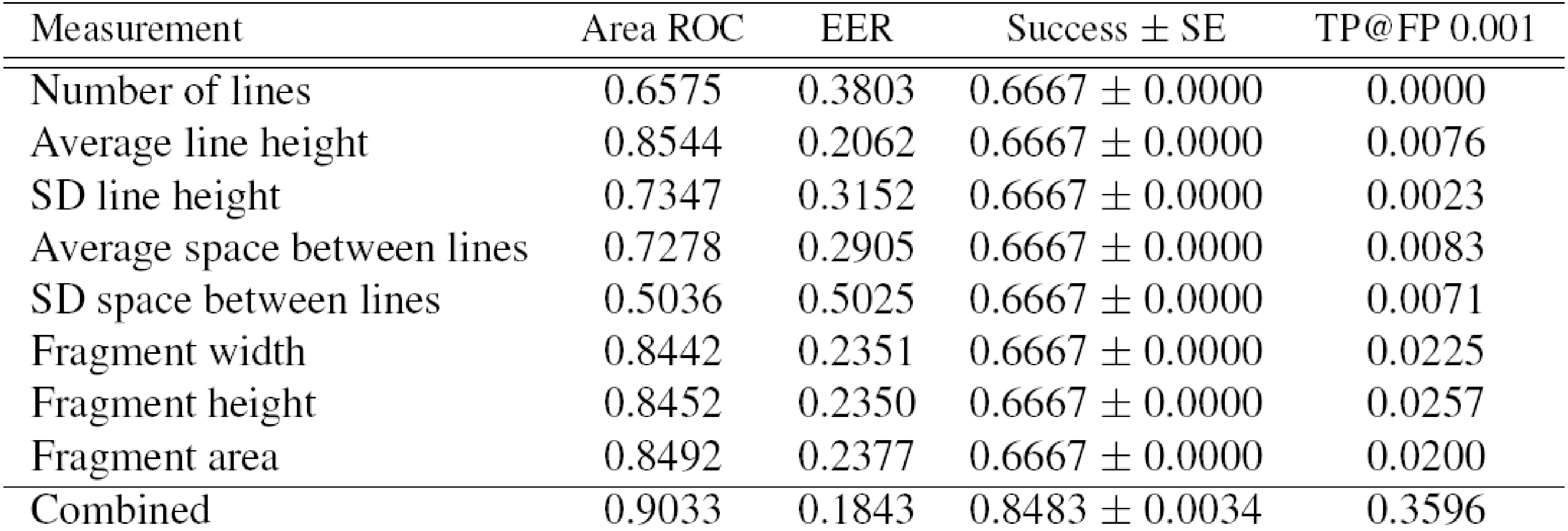
10 splits

each 1000 positives, 2000 negatives

joins are not shared between splits*

Results

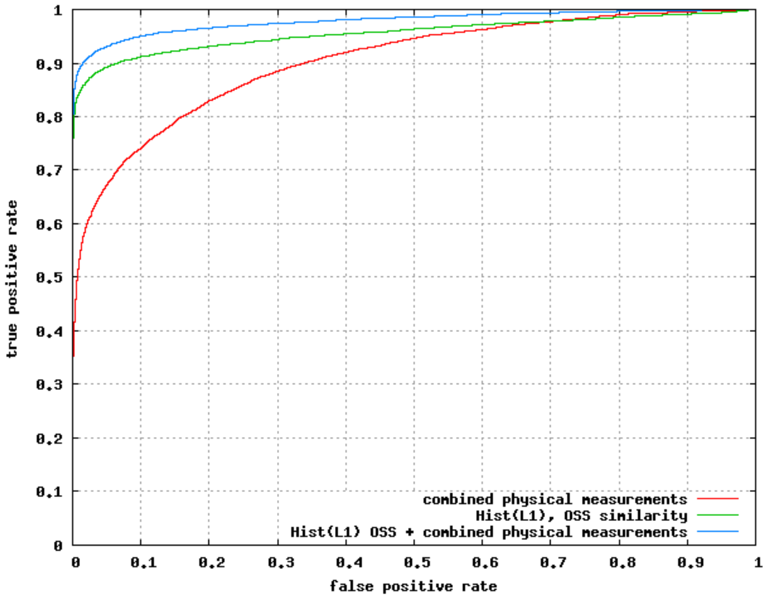
Physical measurement based identification

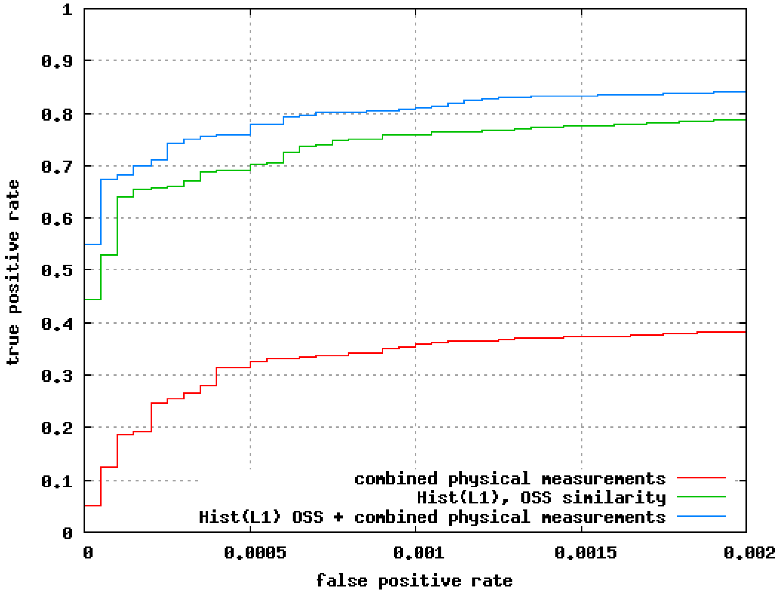


Measurement	Area ROC	EER	Success ± SE	TP@FP 0.001
Number of lines	0.6575	0.3803	0.6667 ± 0.0000	0.0000
Average line height	0.8544	0.2062	0.6667 ± 0.0000	0.0076
SD line height	0.7347	0.3152	0.6667 ± 0.0000	0.0023
Average space between lines	0.7278	0.2905	0.6667 ± 0.0000	0.0083
SD space between lines	0.5036	0.5025	0.6667 ± 0.0000	0.0071
Fragment width	0.8442	0.2351	0.6667 ± 0.0000	0.0225
Fragment height	0.8452	0.2350	0.6667 ± 0.0000	0.0257
Fragment area	0.8492	0.2377	0.6667 ± 0.0000	0.0200
Combined	0.9033	0.1843	0.8483 ± 0.0034	0.3596

Combining similarities

Combination	Area ROC	EER	Success rate± SE	TP@FP of 0.001
Physical Combined	0.9033	0.1843	0.8483 ± 0.0034	0.3596
OSS of Hist (L1)	0.9667	0.0918	0.9374 ± 0.0034	0.7600
OSS of Hist (L1) + Physical	0.9785	0.0627	0.9566 ± 0.0028	0.8116





Paleographic information

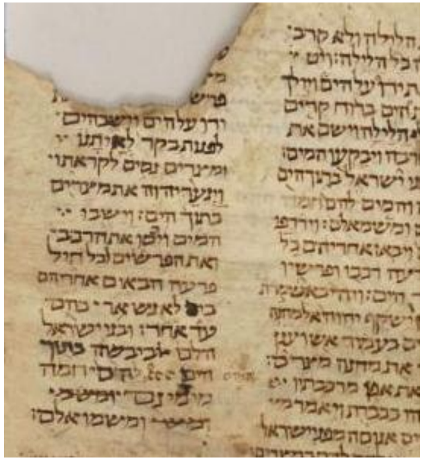
Script-style ("font")

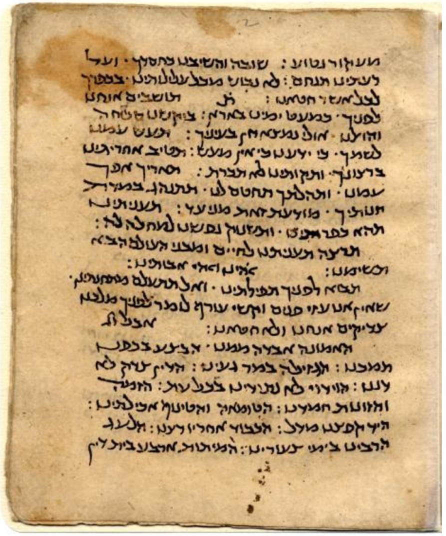
Square

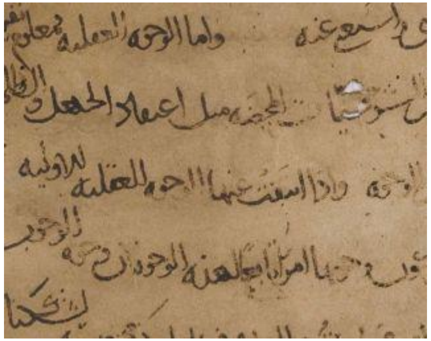
Semi-cursive

Cursive

Arabic







Paleographic information

Boost in performance for join finding

A join typically has only one script-type

Performance increase to 84.51% true positive rate, at a false-positive rate of 0.1%

To obtain this boost, the system does NOT need to decide which script type is used

Many novel joins found

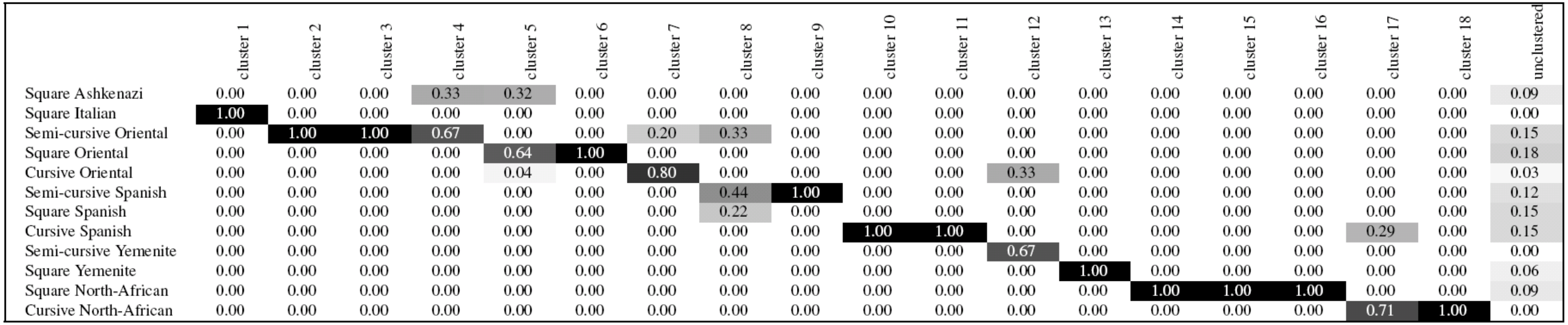
Hundreds of *new* joins found by passing candidate lists to Genizah researchers

Some joins are of great importance

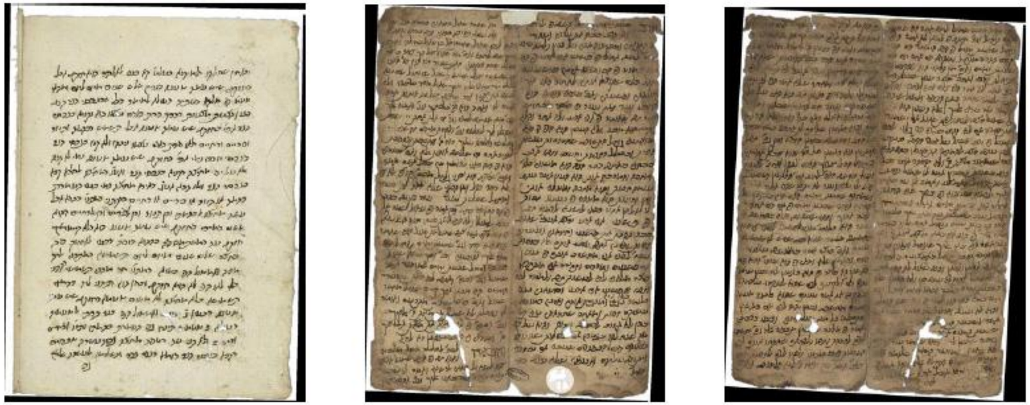
About 30% of looked at joins *between collections* are true joins

Range	% correct	After cleaning
1-2000	24.0%	44.8%
6000-9000	13.5%	18.0%

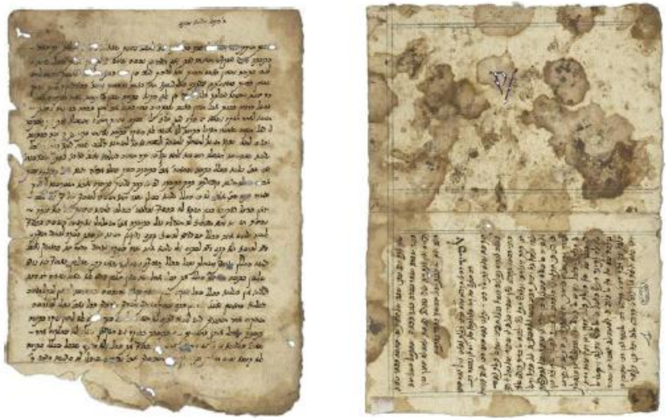
Paleographic information



	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	cluster 11	cluster 12	cluster 13	cluster 14	cluster 15	cluster 16	cluster 17	cluster 18	unclustered
Square Ashkenazi	0.00	0.00	0.00	0.33	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
Square Italian	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-cursive Oriental	0.00	1.00	1.00	0.67	0.00	0.00	0.20	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Square Oriental	0.00	0.00	0.00	0.00	0.64	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18
Cursive Oriental	0.00	0.00	0.00	0.00	0.04	0.00	0.80	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Semi-cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
Square Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.15
Semi-cursive Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Square Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.06
Square North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.09
Cursive North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	1.00	0.00



Cursive North-African



Cursive Spanish

Paleographic Classification

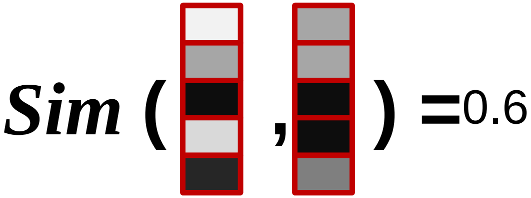
Given a fragment, provide best matching candidates within a corpus of gold standard paleography samples

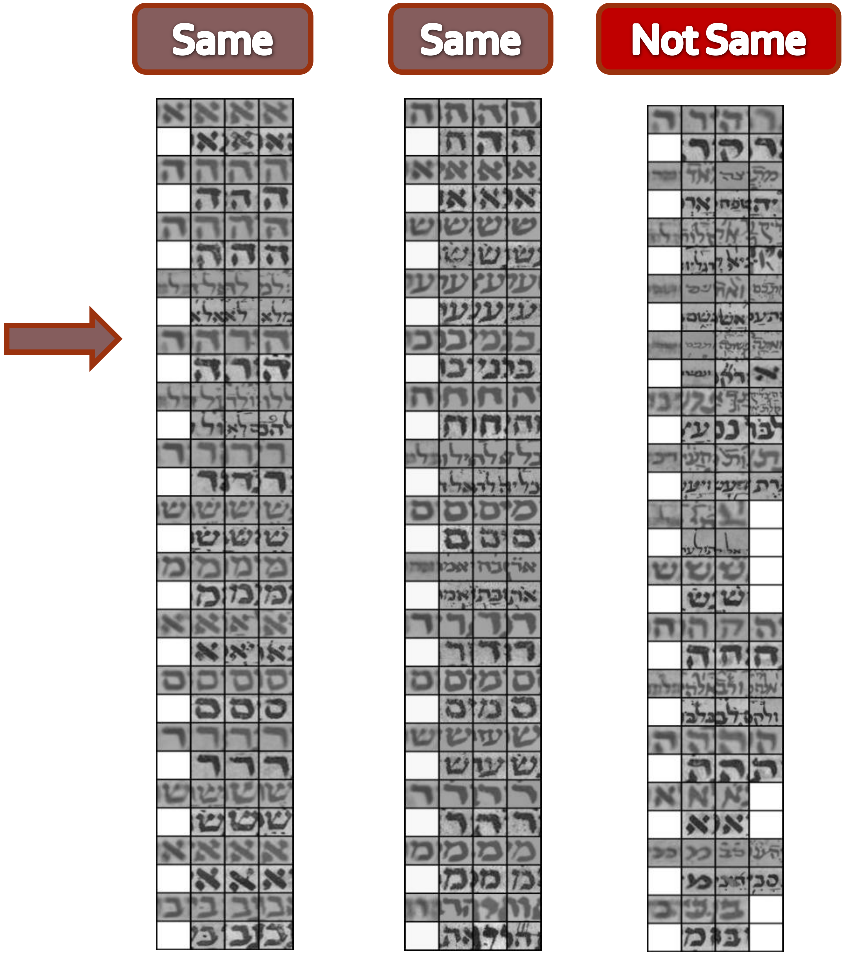
Classify

365 script types in the book
Example pages + Sample letters

אסופת כתבים עבריים מימי הביניים
כרך ב:
כתב ספרדי
מלאכי בית־אריה
עדנה אנגל
האקדמיה הלאומית הישראלית למדעים

Black box problem





Each chart compares 2 documents:

One block for each prototype

Sorted by contribution to the similarity score

Subtracted influence*

*Each time remaining descriptors are projected to the subspace perpendicular to the selected descriptors

Conclusions

1. Joins can be found automatically
2. Considerable value to the fields of medieval history and Jewish research
3. Help us improve

Ongoing work:

1. From pairs to manuscripts (ICCV2011)
2. Paleographic information (CPDAii ,ICIP2011)
3. Content