# Modelling the Interpretation of Literary Allusion with Machine Learning Techniques

N. Coffee[1], J. Gawley[1], C.W. Forstall[1], W.J. Scheirer[2,3], D. Johnson[4], J.J. Corso[4], and B. Parks[2]

1. Dept. of Classics, State University of New York at Buffalo
2. Dept. of Computer Science, Univ. of Colorado, Colorado Springs
3. Depts. of Molecular & Cellular Biology, Computer Science, and Center for Brain Science, Harvard University
4. Dept. of Computer Science and Engineering, State University of New York at Buffalo

University at Buffalo
The State University of New York

UCCS University of Colorado Colorado Springs

HARVARD UNIVERSITY

TESSERAE

Work supported by NEH Start-Up Grant Award No. HD-51570-12

# What is the Tesserae Project?

```
Tessera (Latin): 1) a small square or block; 2)
a tablet bearing a password; 3) a token divided
between friends, so they or their descendants
can recognize one another when meeting again.
```

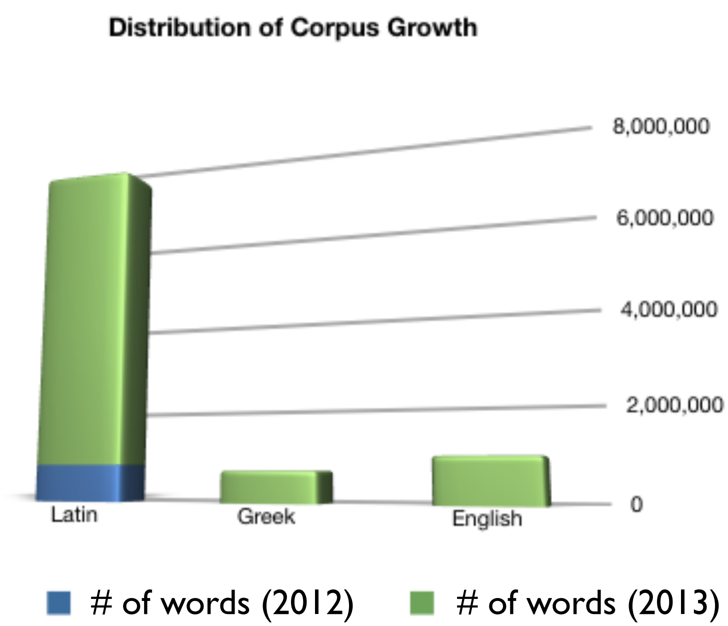
Tesserae is a freely available tool for detecting allusions in literary text.

http://tesserae.caset.buffalo.edu/

http://tesserae.caset.buffalo.edu/blog/

TESSERAE

# What's in Tesserae?




2012: Subset of canonical Latin poetry

2013: Ingestion of all of Perseus Latin and a subset canonical Greek texts

TESSERAE

# Tesserae Search

Parameters allow
for fine-grained search

| | |
|---|---|
| SOURCE: | Vergil - Aeneid |
| TARGET: | Lucan - Pharsalia - Book 1 |
| UNIT: | Phrase |
| FEATURE: | lemma |
| NUMBER OF STOP WORDS: | 10 |
| STOPLIST BASIS: | corpus |
| MAXIMUM DISTANCE: | 999 |
| DISTANCE METRIC: | frequency |
| DROP SCORES BELOW: | 0 |
| SCORING TEAM FILTER: | ON OFF |

Compare Texts

## Top Results

| BC | Target Phrase | Aeneid | Source Phrase |
|---|---|---|---|
| 1.359 | Si licet, exclamat, Romani maxime rector / Nominis et ius est, veras expromere voces; | 2.279 | Ultro flens ipse videbar / Compellare virum et maestas expromere voces: |
| 1.367 | Duc age per Scythiae populos, per inhospita Syrtis / Litora, per calidas Libyae sitientis arenas. | 4.41 | Hinc Gaetulae urbes, genus insuperabile bello, / et Numidae infreni cingunt et inhospita Syrtis; |
| 1.132 | totus popularibus auris / Impelli, plausuque sui gaudere theatri: | 6.816 | Quem iuxta sequitur iactantior Ancus, / nunc quoque iam nimium gaudens popularibus auris. |
| 1.38 | scelera ipsa nefasque / Hac mercede placent: | 7.317 | Hac gener atque socer coeant mercede suorum: |
| 1.237 | Constitit ut capto iussus deponere miles / Signa fore, stridor lituum clangorque tubarum / Non pia concinuit cum rauco classica cornu. | 11.192 | It caelo clamorque virum clangorque tubarum. |
| 1.237 | Constitit ut capto iussus deponere miles / Signa fore, stridor lituum clangorque tubarum / Non pia concinuit cum rauco classica cornu. | 2.313 | Exoritur clamorque virum clangorque tubarum. |
| 1.450 | Et vos barbaricos ritus moremque sinistrum / Sacrorum, Druidae, positis repetistis ab armis. | 12.836 | Morem ritusque sacrorum / adiciam faciamque omnis uno ore Latinos. |

TESSERAE

# How do we rank results?

| BC | Target Phrase | Aeneid | Source Phrase | Parallel Type | Tess Score | Commentators |
|---|---|---|---|---|---|---|
| 1.359 | Si licet, exclamat, Romani maxime rector / Nominis et ius est, veras expromere voces; | 2.279 | Ultro flens ipse videbar / Compellare virum et maestas expromere voces: | 4 | 9.721 | R |
| 1.367 | Duc age per Scythiae populos, per inhospita Syrtis / Litora, per calidas Libyae sitientis arenas. | 4.41 | Hinc Gaetulae urbes, genus insuperabile bello, / et Numidae infreni cingunt et inhospita Syrtis; | 4 | 9.343 | V,R |
| 1.132 | totus popularibus auris / Impelli, plausuque sui gaudere theatri: | 6.816 | Quem iuxta sequitur iactantior Ancus, / nunc quoque iam nimium gaudens popularibus auris. | 5 | 9.247 | V,R |
| 1.38 | scelera ipsa nefasque / Hac mercede placent: | 7.317 | Hac gener atque socer coeant mercede suorum: | 5 | 9.020 | TB,V,R |
| 1.237 | Constitit ut capto iussus deponere miles / Signa fore, stridor lituum clangorque tubarum / Non pia concinuit cum rauco classica cornu. | 11.192 | it caelo clamorque virum clangorque tubarum. | 4 | 8.883 | R |
| 1.237 | Constitit ut capto iussus deponere miles / Signa fore, stridor lituum clangorque tubarum / Non pia concinuit cum rauco classica cornu. | 2.313 | Exoritur clamorque virum clangorque tubarum. | 5 | 8.883 | R |
| 1.450 | Et vos barbaricos ritus moremque sinistrum / Sacrorum, Druidae, positis repetistis ab armis. | 12.836 | Morem ritusque sacrorum / adiciam faciamque omnis uno ore Latinos. | 3 | 8.838 | R |

$f(t)$ is the frequency of each matching term in the target phrase

$f(s)$ is the frequency of each matching term in the source phrase

$d_t$ is the distance in the target

$d_s$ is the distance in the source

$$score = \ln\left(\frac{\sum \frac{1}{f(t_i)} + \sum \frac{1}{f(s_i)}}{d_t + d_s}\right)$$

TESSERAE

The Lucan commentaries are Heitland and Haskins 1887, Thompson and Bruère 1968 (TB), Viansino 1995 (V), and Roche 2009 (R).

# Parallel Types

**5. High formal similarity in analogous content**

**4. Moderate formal similarity in analogous context; or High formal similarity in moderately analogous context.**

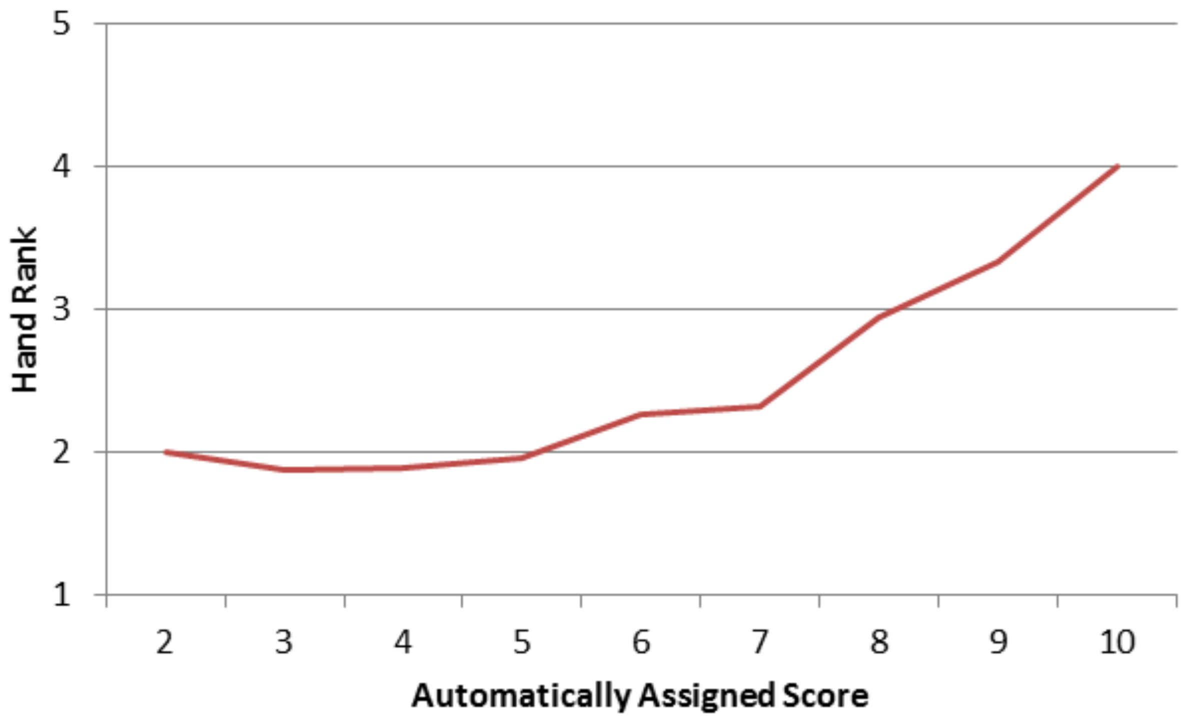
3. High / moderate formal similarity with very common phrase or words; or High / moderate formal similarity with no analogous context; or Moderate formal similarity with moderate / highly analogous context.

2. Very common words in very common phrase; or Words too distant to form a phrase.

1. Error in discovery algorithm, words should not have matched.

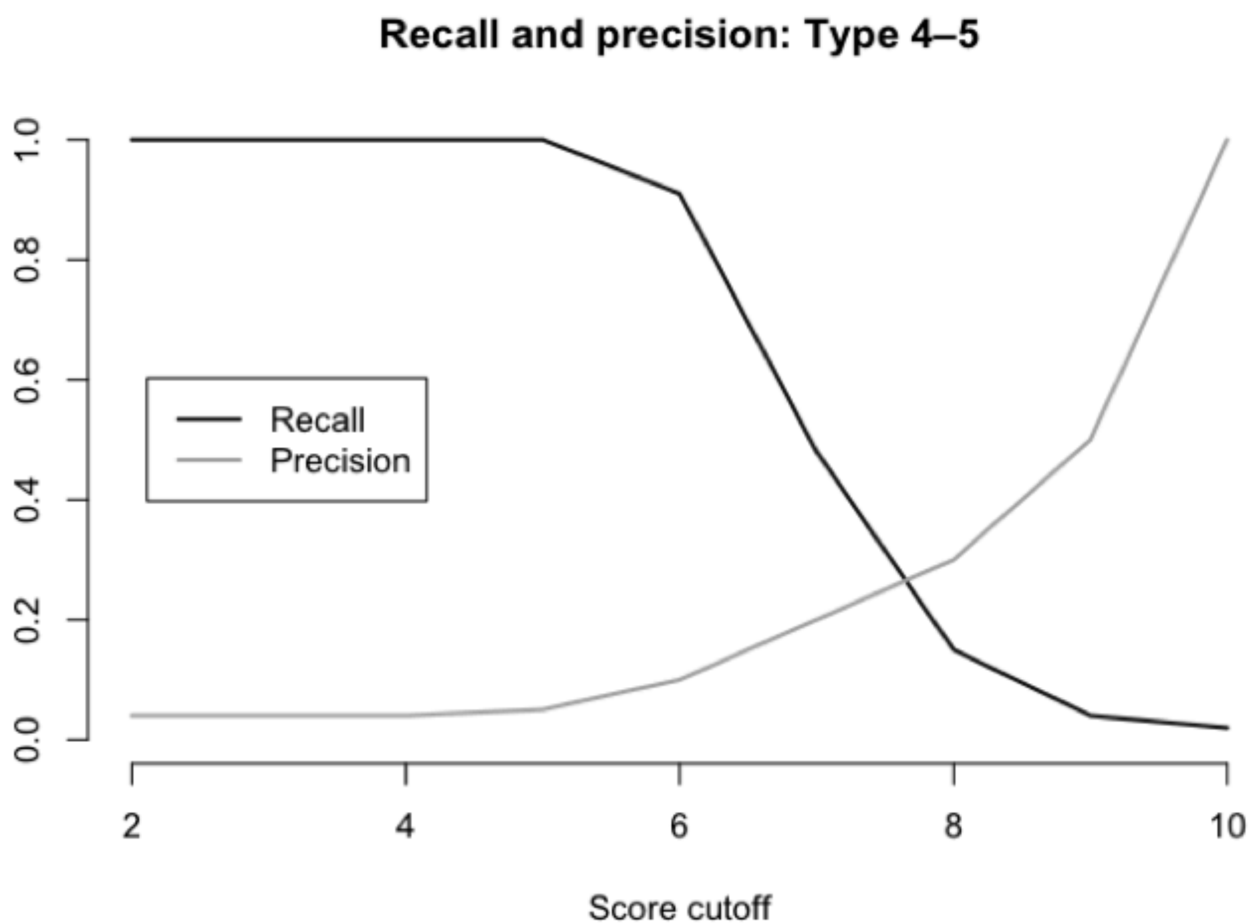
TESSERAE

# Average Hand Rank of Parallels per Automatic Score for Lucan / Vergil Benchmark Test



# String matching is good, but...

Tesserae Lucan / Vergil
Benchmark Results




TESSERAE

# Can we learn what allusion is to find new instances in a large corpus?

NY Times 11.23.2012
http://goo.gl/ROPdr

Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

FACEBOOK
TWITTER
GOOGLE+
SAVE

Machine Learning has the potential to be transformative for complex analysis tasks in literary study

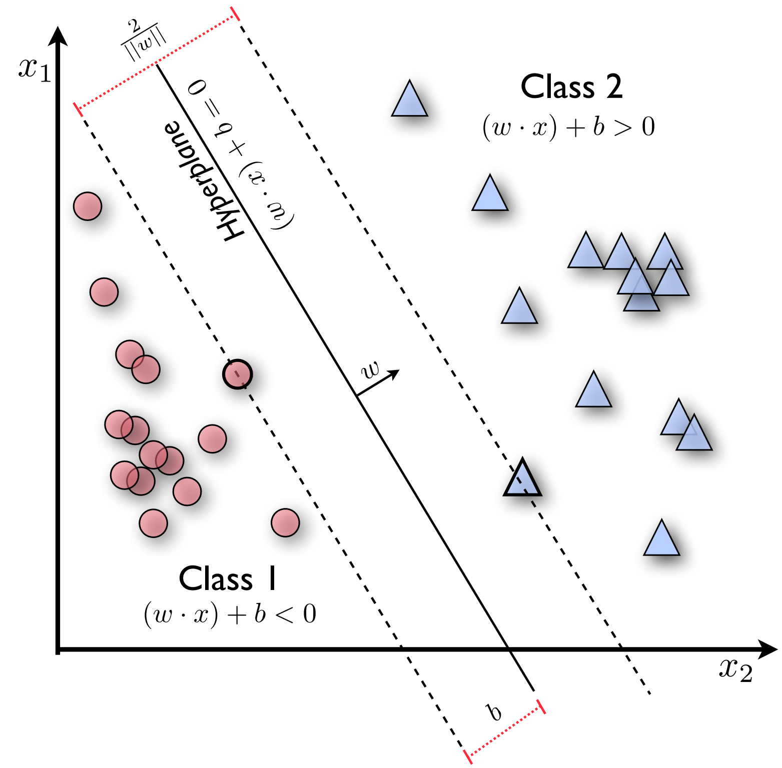
TESSERAE

# Machine Learning and DH

"...what we have today in terms of literary and textual material and computational power represents a moment of revolution in the way we study the literary record"

- Matt Jockers, *Macroanalysis*

- Familiar DH areas using ML
  - Distant Reading
  - Authorship Attribution
  - Stylometry
- Effective tools
  - [Mallett]
  - [R]




TESSERAE

# What are the limitations of what the DH community has been looking at?

- Straightforward classification: use algorithms as a “black box”
- Training with a small set of hand-tuned features
- Closed set evaluation

# Novel applications of machine learning beyond what we've all seen before...

- [Feature Learning]
- [Topic Modeling for Non-Lexical Matching]
- [Open Set Machine Learning]

# Learning Relevant Features

# Features that Express Allusion

- Bamman and Crane 2008[1]
  - token similarity, n-grams and syntactic structure
- Gawley et al. 2012[2]
  - word frequency, distance between words, matching inflected word forms
- This work: greatly expanded feature set
  - bi-gram frequency, frequency of individual words, character-level n-grams and edit distances

1. D. Bamman and G. Crane. The Logic and Discovery of Textual Allusion. LaTeCH, 2008.

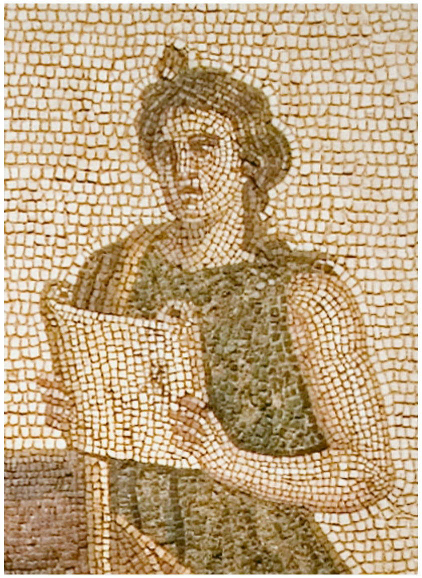
2. J. Gawley, C.W. Forstall, and N. Coffee. Evaluating the literary significance of text re-use in latin poetry. DHCS, 2012.

# Benchmark Data

Lucan, *Bellum Civile*, Book 1

Bust of the Roman poet Lucan, Córdoba, Spain (cc) CC-BY-3.0 Cruccone

Vergil, *Aeneid*



Virgil Mosaic Bardo Museum Tunis

- 3,400 pairs of sentences sharing at least one word
- Each pair was graded (1 - 5), establishing a "bronze set" of ground-truth data

TESSERAE

# Complete Feature Set

## 102 Features

Word Matches *BC*
Word Matches *AEN*
Word Matches Both
Stem Matches *BC*
Stem Matches *AEN*
Stem Matches Both
Unique Forms of Word Matches
Unique Forms of Stem Matches
Word Matches Doc. Specific Mean Freq. *BC*
Word Matches Doc. Specific Mean Freq. *AEN*
Word Matches Doc. Specific Mean Freq. Both
Word Matches Doc. Specific Min. Freq. *BC*
Word Matches Doc. Specific Min. Freq. *AEN*
Word Matches Doc. Specific Min. Freq. Both
Word Matches Doc. Specific Inv. Freq. *BC*
Word Matches Doc. Specific Inv. Freq. *AEN*
Word Matches Doc. Specific Inv. Freq. Both
Word Matches Corpus-wide Mean Freq. *BC*
Word Matches Corpus-wide Mean Freq. *AEN*
Word Matches Corpus-wide Mean Freq. Both

Word Matches Corpus-wide Min. Freq. *BC*
Word Matches Corpus-wide Min. Freq. *AEN*
Word Matches Corpus-wide Min. Freq. Both
Word Matches Corpus-wide Inv. Freq. *BC*
Word Matches Corpus-wide Inv. Freq. *AEN*
Word Matches Corpus-wide Inv. Freq. Both
Phrase Matches Doc. Specific Mean Freq. *BC*
Phrase Matches Doc. Specific Mean Freq. *AEN*
Phrase Matches Doc. Specific Mean Freq. Both
Phrase Matches Doc. Specific Min. Freq. *BC*
Phrase Matches Doc. Specific Min. Freq. *AEN*
Phrase Matches Doc. Specific Min. Freq. Both
Phrase Matches Doc. Specific Inv. Freq. *BC*
Phrase Matches Doc. Specific Inv. Freq. *AEN*
Phrase Matches Doc. Specific Inv. Freq. Both
Phrase Matches Corpus-wide Mean Freq. *BC*
Phrase Matches Corpus-wide Mean Freq. *AEN*
Phrase Matches Corpus-wide Mean Freq. Both
Phrase Matches Corpus-wide Min. Freq. *BC*
Phrase Matches Corpus-wide Min. Freq. *AEN*

Phrase Matches Corpus-wide Min. Freq. Both
Phrase Matches Corpus-wide Inv. Freq. *BC*
Phrase Matches Corpus-wide Inv. Freq. *AEN*
Phrase Matches Corpus-wide Inv. Freq. Both
Mean TF-IFD Word Matches in Phrases *BC*
Mean TF-IFD Word Matches in Phrases *AEN*
Mean TF-IFD Word Matches in Phrases Both
Cum. TF-IFD Word Matches in Phrases *BC*
Cum. TF-IFD Word Matches in Phrases *AEN*
Cum. TF-IFD Word Matches in Phrases Both
Max. TF-IFD Word Matches in Phrases *BC*
Max. TF-IFD Word Matches in Phrases *AEN*
Max. TF-IFD Word Matches in Phrases Both
Mean TF-IFD Word Matches in Text *BC*
Mean TF-IFD Word Matches in Text *AEN*
Mean TF-IFD Word Matches in Text Both
Cum. TF-IFD Word Matches in Text *BC*
Cum. TF-IFD Word Matches in Text *AEN*
Cum. TF-IFD Word Matches in Text Both
Max. TF-IFD Word Matches in Text *BC*

Max. TF-IFD Word Matches in Text *AEN*
Max. TF-IFD Word Matches in Text Both
Mean TF-IFD All Words in Phrases *BC*
Mean TF-IFD All Words in Phrases *AEN*
Mean TF-IFD All Words in Phrases Both
Cum. TF-IFD All Words in Phrases *BC*
Cum. TF-IFD All Words in Phrases *AEN*
Cum. TF-IFD All Words in Phrases Both
Max. TF-IFD All Words in Phrases *BC*
Max. TF-IFD All Words in Phrases *AEN*
Max. TF-IFD All Words in Phrases Both
Mean TF-IFD All Words in Text *BC*
Mean TF-IFD All Words in Text *AEN*
Mean TF-IFD All Words in Text Both
Cum. TF-IFD All Words in Text *BC*
Cum. TF-IFD All Words in Text *AEN*
Cum. TF-IFD All Words in Text Both
Max. TF-IFD All Words in Text *BC*
Max. TF-IFD All Words in Text *AEN*
Max. TF-IFD All Words in Text Both
Semantic Similarity

Dist. Between Furthest Matching Words *BC*
Dist. Between Furthest Matching Words *AEN*
Dist. Between Furthest Matching Words Both
Dist. Between Lowest-freq Words Doc. Specific *BC*
Dist. Between Lowest-freq Words Doc. Specific *AEN*
Dist. Between Lowest-freq Words Doc. Specific Both
Dist. Between Lowest-freq Words Corpus-wide *BC*
Dist. Between Lowest-freq Words Corpus-wide *AEN*
Dist. Between Lowest-freq Words Corpus-wide Both
Dist. Between Highest TF-IDF Words in Phrases *BC*
Dist. Between Highest TF-IDF Words in Phrases *AEN*
Dist. Between Highest TF-IDF Words in Phrases Both
Dist. Between Highest TF-IDF Words in Text *BC*
Dist. Between Highest TF-IDF Words in Text *AEN*
Dist. Between Highest TF-IDF Words in Text Both
Levenshtein Edit Distance
Character-level Uni-gram Count
Character-level Bi-gram Count
Character-level Tri-gram Count
Character-level Bi-gram Frequency
Character-level Tri-gram Frequency

TESSERAE

# Learning Relevant Features

**Objective**: learn relevant combinations of features in the presence of often incomplete data.

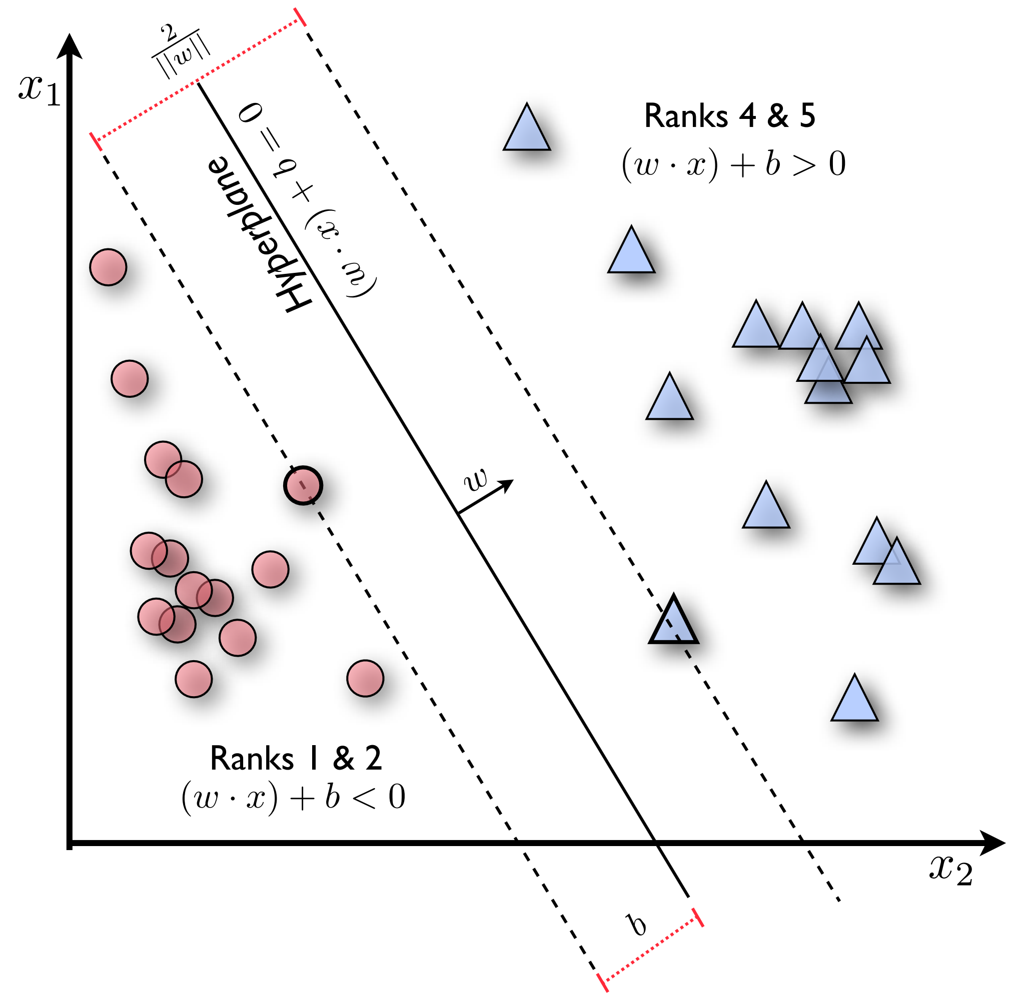
**Task 1**: find good separation between high-ranked parallels (ranks 4 & 5) and low-ranked parallels (ranks 1 & 2) for *Bellum Civile* and the *Aeneid*.

**Task 2**: find good separation between commentator parallels and non-commentator parallels.

# Why two different evaluation tasks?

- Neither task is ideal by itself
  - Rank 4/5 vs. 1/2 classification problem involves our own subjective hand-ranking
  - Commentator vs. non-commentator classification problem gives no weight to meaningful parallels that the commentators did not record

# Support Vector Machines




$w$ is the weight vector, which gives us some sense of relative feature importance

TESSERAE

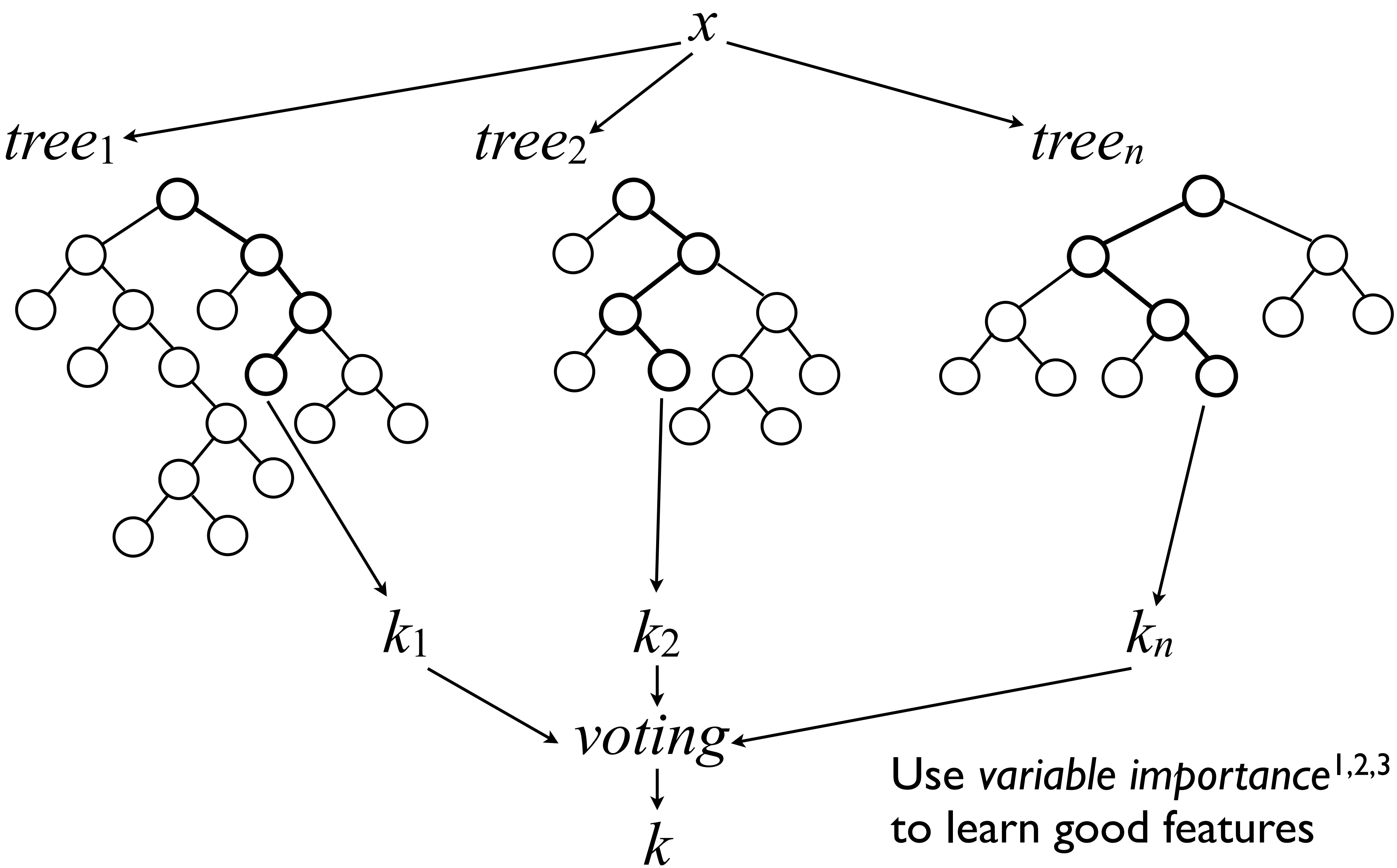
# Does SVM provide good separation?

- Rank 4/5 vs. 1/2 Classification Problem:

  **Area Under the Curve (AUC): 81.5%**

This suggests that *multiple* quantifiable patterns do exist across allusions, which can be captured algorithmically.

# Random Forest




Use *variable importance*[1,2,3]
to learn good features



1. L. Breiman, "Random Forests," Machine Learning 45(1), 2001.
2. T. Tobata, "Approaching Dickens' Style Through Random Forest," DH 2012.
3. C. Xiong, D. Johnson, R. Xu, and J. J. Corso, "Random forests for metric learning with implicit pairwise position dependence. ACM SIGKDD 2012.

# Does Random Forest provide good separation?

- Rank 4/5 vs. 1/2 Classification Problem:

  **Area Under the Curve (AUC) between: 82% - 83%**

- Incomplete data: not all dimensions are present for every data point
  - Use proximities to implicitly replace missing dimensions
  - Imputation and Marginalization

# Top 25 SVM Features:
# Rank 4/5 vs. 1/2 Classification Problem

Mean-TFIDF-Word-Matches-in-Phrases-AEN
Phrase-Matches-Doc-Specific-Mean-Freq-BC
Dist-Between-Highest-TFIDF-Words-in-Text-BC
Cum-TFIDF-Word-Matches-in-Phrases-AEN
Mean-TFIDF-Word-Matches-in-Text-Both
Dist-Between-Furthest-Matching-Words-AEN Levenshtein-Edit-Distance
Word-Matches-Corpus-Wide-Min-Freq-Both
Word-Matches-Doc-Specific-Min-Freq-BC
Cum-TFIDF-Word-Matches-in-Text-BC Phrase-Matches-Corpus-Wide-Mean-Freq-AEN
Word-Matches-Doc-Specific-Mean-Freq-BC
Max-TFIDF-Word-Matches-in-Phrases-BC
Mean-TFIDF-all-Words-in-Phrases-AEN Word-Matches-AEN
Word-Matches-Doc-Specific-Min-Freq-Both
Stem-Matches-BC Word-Matches-Corpus-Wide-Min-Freq-BC
Phrase-Matches-Doc-Specific-Min-Freq-AEN
Word-Matches-Corpus-Wide-Min-Freq-AEN
Dist-Between-Lowest-Freq-Words-Doc-Specific-AEN
Max-TFIDF-all-Words-in-Text-Both Phrase-Matches-Corpus-Wide-Inv-Freq-AEN
Unique-Forms-of-Word-Matches
Mean-TFIDF-all-Words-in-Text-BC

# Top 25 Random Forest Features: Rank 4/5 vs. 1/2 Classification Problem

Cum-TFIDF-all-Words-in-Text-BC
Max-TFIDF-all-Words-in-Text-BC
Cum-TFIDF-all-Words-in-Text-AEN
Character-Level-Trigram-Count
Mean-TFIDF-all-Words-in-Text-AEN
Character-Level-Unigram-Count
Phrase-Matches-Corpus-Wide-Mean-Freq-BC
Phrase-Matches-Corpus-Wide-Mean-Freq-Both
Word-Matches-Corpus-Wide-Min-Freq-BC
Phrase-Matches-Corpus-Wide-Inv-Freq-Both
Character-Level-Trigram-Frequency
Max-TFIDF-all-Words-in-Phrases-BC
Phrase-Matches-Doc-Specific-Mean-Freq-Both
Phrase-Matches-Doc-Specific-Mean-Freq-AEN
Mean-TFIDF-all-Words-in-Text-BC
Character-Level-Bigram-Frequency
Phrase-Matches-Corpus-Wide-Mean-Freq-AEN
Phrase-Matches-Corpus-Wide-Inv-Freq-AEN
Phrase-Matches-Doc-Specific-Inv-Freq-AEN
Cum-TFIDF-all-Words-in-Phrases-Both
Character-Level-Bigram-Count
Cum-TFIDF-all-Words-in-Text-Both
Cum-TFIDF-all-Words-in-Phrases-BC
Cum-TFIDF-all-Words-in-Phrases-AEN
Max-TFIDF-Word-Matches-in-Phrases-AEN

# Top 25 Random Forest Features: Commentator vs. Non-Commentator Classification Problem

Cum-TFIDF-all-Words-in-Text-AEN
Max-TFIDF-all-Words-in-Text-BC
Character-Level-Bigram-Count
Word-Matches-Corpus-Wide-Min-Freq-Both
Character-Level-Unigram-Count
Phrase-Matches-Corpus-Wide-Inv-Freq-Both
Phrase-Matches-Corpus-Wide-Inv-Freq-BC
Character-Level-Trigram-Count
Phrase-Matches-Doc-Specific-Mean-Freq-BC
Word-Matches-Corpus-Wide-Min-Freq-AEN
Phrase-Matches-Corpus-Wide-Mean-Freq-BC
Character-Level-Bigram-Frequency
Phrase-Matches-Corpus-Wide-Mean-Freq-Both
Phrase-Matches-Corpus-Wide-Mean-Freq-AEN
Character-Level-Trigram-Frequency
Cum-TFIDF-all-Words-in-Phrases-AEN
Phrase-Matches-Doc-Specific-Mean-Freq-Both
Phrase-Matches-Doc-Specific-Mean-Freq-AEN
Word-Matches-Corpus-Wide-Min-Freq-BC
Mean-TFIDF-all-Words-in-Text-BC
Unique-Forms-of-Word-Matches
Phrase-Matches-Doc-Specific-Inv-Freq-AEN
Mean-TFIDF-Word-Matches-in-Phrases-AEN
Cum-TFIDF-all-Words-in-Phrases-Both
Cum-TFIDF-all-Words-in-Phrases-BC

# Are any weightings correlated?

SVM and Random Forest
Rank 4/5 vs. 1/2 Classification Problem

**Mean-TFIDF-all-Words-in-Text-BC**

Phrase-Matches-Corpus-Wide-Inv-Freq-AEN

**Phrase-Matches-Corpus-Wide-Mean-Freq-AEN**

**Word-Matches-Corpus-Wide-Min-Freq-BC**

# Are any weightings correlated?

Random Forest
Rank 4/5 vs. 1/2 Classification Problem and
Commentator vs. Non-Commentator Classification Problem

Phrase-Matches-Doc-Specific-Inv-Freq-AEN
Phrase-Matches-Doc-Specific-Mean-Freq-AEN
Max-TFIDF-all-Words-in-Text-BC
Phrase-Matches-Corpus-Wide-Mean-Freq-Both
Character-Level-Unigram-Count
Phrase-Matches-Corpus-Wide-Mean-Freq-BC
Character-Level-Bigram-Count
Phrase-Matches-Corpus-Wide-Inv-Freq-Both
Cum-TFIDF-all-Words-in-Text-AEN
Character-Level-Trigram-Frequency
**Mean-TFIDF-all-Words-in-Text-BC**
Phrase-Matches-Doc-Specific-Mean-Freq-Both
Character-Level-Bigram-Frequency
Character-Level-Trigram-Count
Cum-TFIDF-all-Words-in-Phrases-BC
**Phrase-Matches-Corpus-Wide-Mean-Freq-AEN**
Cum-TFIDF-all-Words-in-Phrases-Both
Cum-TFIDF-all-Words-in-Phrases-AEN
**Word-Matches-Corpus-Wide-Min-Freq-BC**

TESSERAE

# Analysis

- Features universal to our benchmark experiment
  - **Phrase-Matches-Corpus-Wide-Mean-Freq-AEN**
  - **Mean-TFIDF-all-Words-in-Text-BC**
  - **Word-Matches-Corpus-Wide-Min-Freq-BC**
- Phrase level features are interesting: what makes an allusion extends beyond the matching words.
  - We can measure this in cases where there are no matching words
- Global features (corpus- and text-wide) are also a signature of a particular poet's style
  - In this case, Lucan

# Analysis

- Summary of features that are important for Vergil: Overall sense of word rarity
- Summary of features that are important for Lucan: Targeted rare words
- Are these particular features specific to the benchmark set???

# Topic Modeling for Non-Lexical Matching

TESSERAE

# Another type of allusion

- Macrobius (5th century)
  Recognized thematic similarity as a characteristic of allusion

## Example[1]

praeterea iam nec mutari pabula refert
quaesitaeque nocent artes, cessere magistri.

(Vergil *Georgics* 3.548-9)

Besides, it makes no difference now to change their feed,
healing arts do harm when applied, their masters withdraw in defeat.

nec requies erat ulla mali: defessa iacebant
corpora, mussabat tacito medicina timore.

(Lucretius *De Rerum Natura* 6.1178-9)

Nor did the evil know any respite: their bodies lay exhausted, physicians reduced to muttering in silent fear.

TESSERAE

1. Kaster, R., Ed. 2011. *Macrobius Saturnalia Books 6-7*. Loeb Classical Library. Cambridge, MA, Harvard University Press.

# Topic Modeling for Matching Allusions

- Objective: improve recall by finding additional parallels based on context

| Query: "Rubiconis aquas" | LSA Score |
|---|---|
| **2. post Cilicasne uagos et lassi Pontica regis proelia barbarico uix consummata ueneno ultima Pompeio dabitur prouincia Caesar** | 0.99977112 |
| 4. iam gelidas Caesar cursu superauerat Alpes ingentisque animo motus bellumque futurum ceperat ut uentum est parui Rubiconis ad undas | 0.99919581 |
| 3. non si tumido me gurgite Ganges summoueat stabit iam flumine Caesar in ullo post Rubiconis aquas | 0.89826238 |
| 5. sed non in Caesare tantum nomen erat nec fama ducis sed nescia uirtus stare loco solusque pudor non uincere bello | 0.023670167 |
| 1. Bella per Emathios plus quam ciuilia campos iusque datum sceleri canimus | 0.0 |
| 6. turba minor ritu sequitur succincta Gabino Vestalemque chorum ducit uittata sacerdos Troianam soli cui fas uidisse Mineruam | 0.0 |
| 7. certe populi quos despicit Arctos felices errore suo quos ille timorum maximus haut urguet leti metus | 0.0 |
| 8. quodque nefas nullis inpune apparuit extis ecce uidet capiti fibrarum increscere molem alterius capitis | 0.0 |
| 9. rupta quies populi stratisque excita iuuentus deripuit sacris adfixa penatibus arma quae pax longa dabat | 0.0 |

# Algorithmic Approach

- Latent Semantic Analysis (LSA) from the [Gensim] Package
- Query: 14 lines around target sentence
- Documents: 14 lines around target sentences throughout the entire reference corpus
- Features: bag-of-words representation, with the inflected form of each word replaced with the set of all possible stems
- Free parameter: number of topics

# A match to Roche's sensitivity[1] to thematic similarity without close verbal resemblance

***Civil War* 1.498 – 511**

qualis, cum turbidus Auster
reppulit a Libycis inmensum Syrtibus aequor
fractaque ueliferi sonuerunt pondera mali,
desilit in fluctus deserta puppe magister
nauitaque et nondum sparsa conpage carinae
naufragium sibi quisque facit, sic urbe relicta
in bellum fugitur. *nullum iam languidus aeuo*
*eualuit reuocare parens coniunxue maritum*
*fletibus, aut patrii, dubiae dum uota salutis*
*conciperent, tenuere lares*; nec limine quisquam
haesit et extremo tunc forsitan urbis amatae
plenus abit uisu: ruit inreuocabile uolgus.
o faciles dare summa deos eademque tueri
difficiles!

***Aeneid* 3.1-12**

postquam res Asiae Priamique euertere gentem
immeritam uisum superis, ceciditque superbum
Ilium et omnis humo fumat Neptunia Troia,
diuersa exsilia et desertas quaerere terras
auguriis agimur diuum, classemque sub ipsa
Antandro et Phrygiae molimur montibus Idae,
incerti quo fata ferant, ubi sistere detur,
contrahimusque uiros. uix prima inceperat aestas
et pater Anchises dare fatis uela iubebat,
litora cum patriae lacrimans portusque relinquo
et campos ubi Troia fuit. *feror exsul in altum*
*cum sociis natoque penatibus et magnis dis.*

TESSERAE

1. Roche, P., Ed. 2009. *Lucan: De bello civili. Book 1*. Oxford, Oxford University Press.

# Additional *Bellum Civile* 1 – *Aeneid* commentator parallels recovered (12)

| BC Line | AEN Line | Shared Context | Num. Topics | Rank |
|---|---|---|---|---|
| 1.60 | 1.291 | Divine destiny of Caesar; peace | 10 | 4 |
| 1.139 | 4.441 | The blowing wind; tree | 20 | 4 |
| 1.141 | 2.626 | The blowing wind; tree | 15 | 2 |
| 1.193 | 2.774 | An apparition | 20 | 28 |
| 1.193 | 3.47 | An apparition | 15 | 42 |
| 1.291 | 11.492 | Horses | 20 | 30 |
| 1.490 | 11.142 | Flight | 15 | 46 |
| 1.504 | 2.634 | Abandonment | 15 | 1* |
| 1.504 | 3.11 | Abandonment; Nautical Imagery | 15 | 1 |
| 1.673 | 2.199 | Omens; terror | 15 | 24 |
| 1.676 | 4.68 | Dido as Bacchant | 15 | 1 |
| 1.676 | 6.48 | Prophecy | 15 | 32 |
| 1.695 | 6.102 | Frenzied Discussion | 20 | 29 |

*** denotes a parallel also found by Tesserae Version 3 scoring.**

TESSERAE

# Available in Tesserae

http://tesserae.vast.uccs.edu/cgi-bin/lsa.pl

Back to Tesserae

Target: Lucan – Bellum Civile – Book 1

**LUCAN.BELLUM_CIVILE.PART.1**

Click to select a phrase (plus surrounding context).
Matches in vergil.aeneid.part.1 will be highlighted at right.

1.1 Bella per Emathios plus quam civilia campos
1.2 Iusque datum sceleri canimus, populumque potentem
1.3 In sua victrici conversum viscera dextra,
1.4 Cognatasque acies, et rupto foedere regni,
1.5 Certatum totis concussi viribus orbis
1.6 In commune nefas, infestisque obvia signis
1.7 Signa, pares aquilas, et pila minantia pilis.
1.8 Quis furor, o cives, quae tanta licentia ferri,
1.9 Gentibus invisis Latium praebere cruorem?
1.10 Cumque superba foret Babylon spolianda tropaeis
1.11 Ausoniis, umbraque erraret Crassus inulta,
1.12 Bella geri placuit nullos habitura triumphos?
1.13 Heu quantum terrae potuit pelagique parari
1.14 Hoc, quem civiles hauserunt, sanguine, dextrae,
1.15 Unde venit Titan, et nox ubi sidera condit,
1.16 Quaque dies medius flagrantibus aestuat horis,
1.17 Et qua bruma, rigens ac nescia vere remitti,
1.18 Adstringit Scythico glacialem frigore pontum!
1.19 Sub iuga iam Seres, iam barbarus isset Araxes,
1.20 Et gens si qua iacet nascenti conscia Nilo.

Back to Tesserae

Source: Vergil – Aeneid – Book 1

Number of Topics: 15

**VERGIL.AENEID.PART.1**

1.1 Arma virumque cano, Troiae qui primus ab oris
1.2 Italiam, fato profugus, Laviniaque venit
1.3 litora, multum ille et terris iactatus et alto
1.4 vi superum saevae memorem Iunonis ob iram;
1.5 multa quoque et bello passus, dum conderet urbem,
1.6 inferretque deos Latio, genus unde Latinum,
1.7 Albanique patres, atque altae moenia Romae.
1.8 Musa, mihi causas memora, quo numine laeso,
1.9 quidve dolens, regina deum tot volvere casus
1.10 insignem pietate virum, tot adire labores
1.11 impulerit. Tantaene animis caelestibus irae?
1.12 Urbs antiqua fuit, Tyrii tenuere coloni,
1.13 Karthago, Italiam contra Tiberinaque longe
1.14 ostia, dives opum studiisque asperrima belli;
1.15 quam Iuno fertur terris magis omnibus unam
1.16 posthabita coluisse Samo; hic illius arma,
1.17 hic currus fuit; hoc regnum dea gentibus esse,
1.18 si qua fata sinant, iam tum tenditque fovetque.
1.19 Progeniem sed enim Troiano a sanguine duci
1.20 audierat, Tyrias olim quae verteret arces;

TESSERAE

# Open Set Machine Learning

# How well are we really doing on classification tasks?

- Lots of good work in classification, but nearly all of it is in a closed set context, e.g.
  - Jockers et al. LLC 2008[1]
    - Book of Mormon
  - Jockers and Witten LLC 2010[2]
    - Federalist Papers
  - Eder 2010[3]
    - English novels, Polish Novels and Latin Prose
  - Eder and Rybicki 2013[4]
    - English, German, French, Italian, and Polish Novels

TESSERAE

1. M. Jockers, D. Witten, and C. Criddle, "Reassessing authorship in the 'Book of Mormon' using delta and nearest shrunken centroid classification," LLC 23(4): 465–91, 2008.
2. M. Jockers and D. Witten, "A comparative study of machine learning methods for authorship attribution," LLC 25(2), 2010.
3. M. Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem," DH 2010.
4. M. Eder and J. Rybicki, "Do Birds of a feather really flock together, or how to choose training samples for authorship attribution," LLC 28(2), 2013.

# Notable Exceptions

- Schaalje and Fields, LLC 2011[1]
- Koppel et al. English Studies 2012[2]
- Solutions reduce to thresholds over similarity scores...

Can we do better?

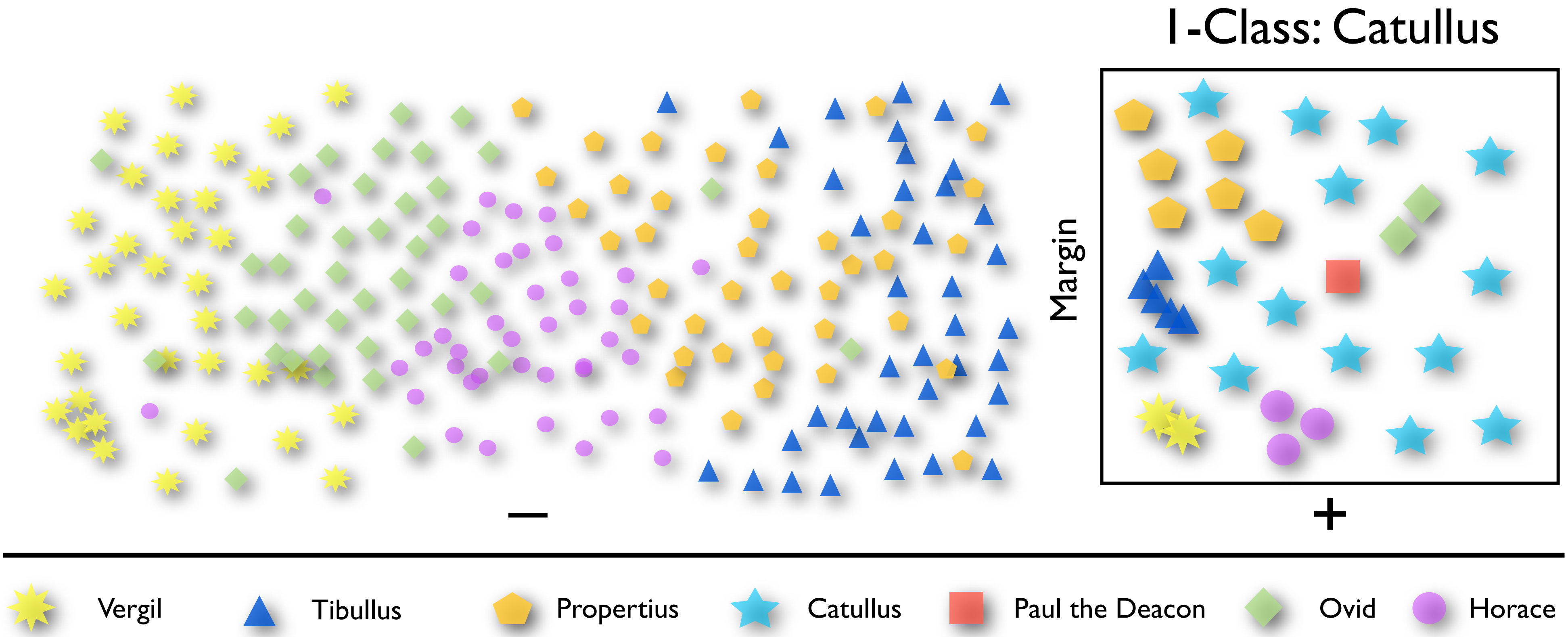
1. G. Schaalje and P Fields "Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes," LLC, 21(1), 2011.

2. M. Koppel, J. Schler, S. Argamon, and Y. Winter, "The Fundamental Problem of of Authorship Attribution," English Studies, 93(3), 2012.

TESSERAE

# Assessing Stylistic Similarity

Forstall et al. LLC 2011[1] - 1-class SVM




Bad density estimator for under-sampled positive training data -
great when the positive class is complete

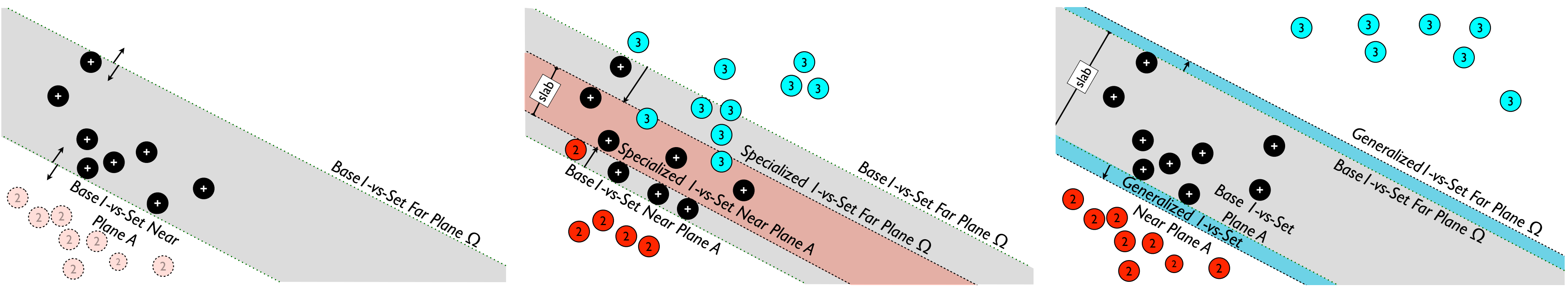
TESSERAE

1. C.W. Forstall, S. Jacobson, and W.J. Scheirer, "Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*," LLC, 26(3), 2011

# Open Set Machine Learning

## 1-vs-Set Machine[1]

Minimize risk of the unknown +
empirical risk over the training data




1. W. Scheirer, A. Rocha, A. Sapkota, and T. Boult, "Towards Open Set Recognition," IEEE T-PAMI, 36(3), 2013.

TESSERAE

# Tesserae Bibliography


N. Coffee, J.-P. Koenig, S. Poornima, C.W. Forstall, R. Ossewaarde and S.L. Jacobson, "The Tesserae Project: Intertextual Analysis of Latin Poetry," LLC, 28(2), 2013.

N. Coffee, J.-P. Koenig, S. Poornima, C.W. Forstall, R. Ossewaarde and S.L. Jacobson, "Intertextuality in the Digital Age," Transactions of the American Philological Association, 142(2), 2012.

C.W. Forstall, W.J. Scheirer and S.L. Jacobson, "Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*," LLC, 26(3), 2011.

C.W. Forstall and W.J. Scheirer, "Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound," Proc. of DHCS, 1(2), 2010.


Research Blog: http://tesserae.caset.buffalo.edu/blog/

TESSERAE