USING DIMENSIONALITY REDUCTION AND TAG PARAMETER SPACES TO STUDY HISTORICAL CHANGE IN A LARGE DOCUMENT ARCHIVE

Tim Hitchcock (Professor of Digital History, University of Sussex)

William J Turkel (Professor of History, The University of Western Ontario)

OVERVIEW

- In this presentation we discuss one approach to studying historical change in a large document archive, *The Old Bailey Proceedings Online*.
- In addition to the texts themselves, we are working with two kinds of representation.
- The first is a set of XML tags that were added to the trial accounts when the digital archive was created. Since these tags were drawn from small finite sets, we can think of them as dimensions that can be used to categorize each trial in a tag parameter space.
- The second is a dimension reduction technique, Stable Random Projections (Schmidt 2018). Each SRP is a small sketch, or fingerprint, of a given trial. Each trial can be located in a space of SRPs.

THE OLD BAILEY PROCEEDINGS ONLINE

- The Old Bailey Proceedings Online (oldbaileyonline.org) is a fully searchable edition of 197,745 criminal trials held at London's central criminal court between 1674 and 1913.
- The trials range in length from eight words (the shortest) to over 155 thousand words (the longest). This raises challenges for comparing texts to one another.
- Comprising more than 127 million words, the *Proceedings* are much too large to be read in their entirety.
- In the past decade, we have applied various text mining and machine learning techniques to the digitized *Proceedings*, creating more than 1.78 million derivative files in the process.

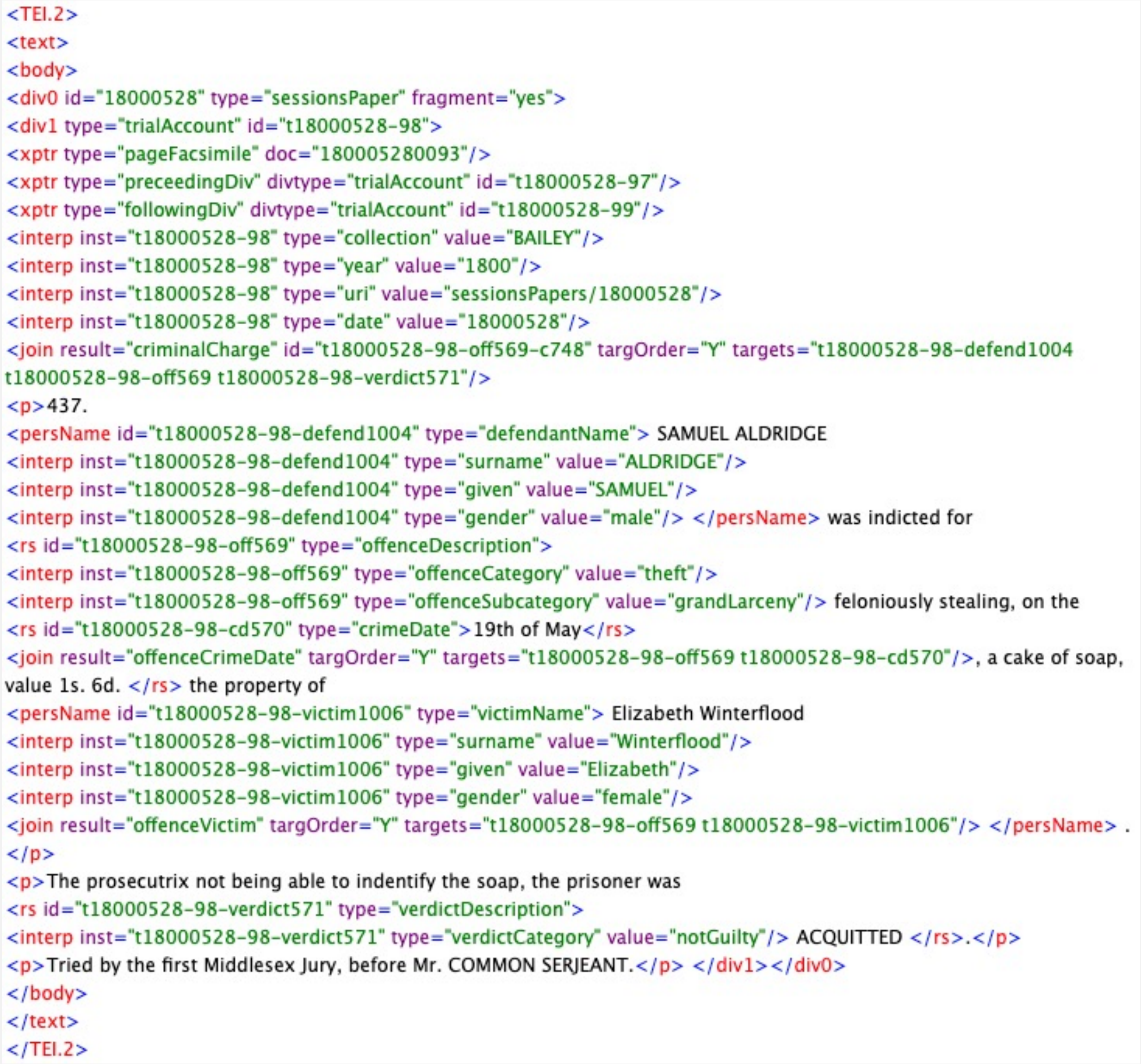
SAMPLE TRIAL SHOWING MARKUP

AS RAW TEXT:

437. SAMUEL ALDRIDGE was indicted for feloniously stealing, on the 19th of May , a cake of soap, value 1s. 6d. the property of Elizabeth Winterflood .

The prosecutrix not being able to indentify the soap, the prisoner was ACQUITTED .

Tried by the first Middlesex Jury, before Mr. COMMON SERJEANT.



```
<TEI.2>
<text>
<body>
<div0 id="18000528" type="sessionsPaper" fragment="yes">
<div1 type="trialAccount" id="t18000528-98">
<xptr type="pageFacsimile" doc="180005280093"/>
<xptr type="preceedingDiv" divtype="trialAccount" id="t18000528-97"/>
<xptr type="followingDiv" divtype="trialAccount" id="t18000528-99"/>
<interp inst="t18000528-98" type="collection" value="BAILEY"/>
<interp inst="t18000528-98" type="year" value="1800"/>
<interp inst="t18000528-98" type="uri" value="sessionsPapers/18000528"/>
<interp inst="t18000528-98" type="date" value="18000528"/>
<join result="criminalCharge" id="t18000528-98-off569-c748" targOrder="Y" targets="t18000528-98-defend1004
t18000528-98-off569 t18000528-98-verdict571"/>
<p>437.
<persName id="t18000528-98-defend1004" type="defendantName"> SAMUEL ALDRIDGE
<interp inst="t18000528-98-defend1004" type="surname" value="ALDRIDGE"/>
<interp inst="t18000528-98-defend1004" type="given" value="SAMUEL"/>
<interp inst="t18000528-98-defend1004" type="gender" value="male"/> </persName> was indicted for
<rs id="t18000528-98-off569" type="offenceDescription">
<interp inst="t18000528-98-off569" type="offenceCategory" value="theft"/>
<interp inst="t18000528-98-off569" type="offenceSubcategory" value="grandLarceny"/> feloniously stealing, on the
<rs id="t18000528-98-cd570" type="crimeDate">19th of May</rs>
<join result="offenceCrimeDate" targOrder="Y" targets="t18000528-98-off569 t18000528-98-cd570"/>, a cake of soap,
value 1s. 6d. </rs> the property of
<persName id="t18000528-98-victim1006" type="victimName"> Elizabeth Winterflood
<interp inst="t18000528-98-victim1006" type="surname" value="Winterflood"/>
<interp inst="t18000528-98-victim1006" type="given" value="Elizabeth"/>
<interp inst="t18000528-98-victim1006" type="gender" value="female"/>
<join result="offenceVictim" targOrder="Y" targets="t18000528-98-off569 t18000528-98-victim1006"/> </persName> .
</p>
<p>The prosecutrix not being able to indentify the soap, the prisoner was
<rs id="t18000528-98-verdict571" type="verdictDescription">
<interp inst="t18000528-98-verdict571" type="verdictCategory" value="notGuilty"/> ACQUITTED </rs>.</p>
<p>Tried by the first Middlesex Jury, before Mr. COMMON SERJEANT.</p> </div1></div0>
</body>
</text>
</TEI.2>
```

XML TAGS

- The trials have been marked up with XML tags reflecting defendant, victim, offence, verdict, punishment, and other categories.
- Each of these categories has been further subdivided into subcategories.
- So, for example, many of the trials in the *Proceedings* are thefts, and some of these thefts are burglary, some housebreaking, some highway robbery, etc.
- Since these tags were drawn from a small, closed vocabulary, their possible values define a parameter space: each trial can be thought of as a point in a multidimensional space.
- A pair of trials might be tagged *offcat-kill* (offence category is a killing); these are nearer to one another (along this one dimension at least) than they are to trials tagged *offcat-theft*.

TAG PARAMETER SPACE

IT IS POSSIBLE FOR A PAIR OF TRIALS TO SHARE THE SAME POINT IN TAG PARAMETER SPACE. THERE ARE 6,589 TRIALS THAT HAVE THE FOLLOWING SET OF TAGS:

- Offence category: theft
- Offence subcategory: grand larceny
- Defendant gender: male
- Victim gender: male
- Verdict: guilty
- Punishment: transport

This is one unique set of tags

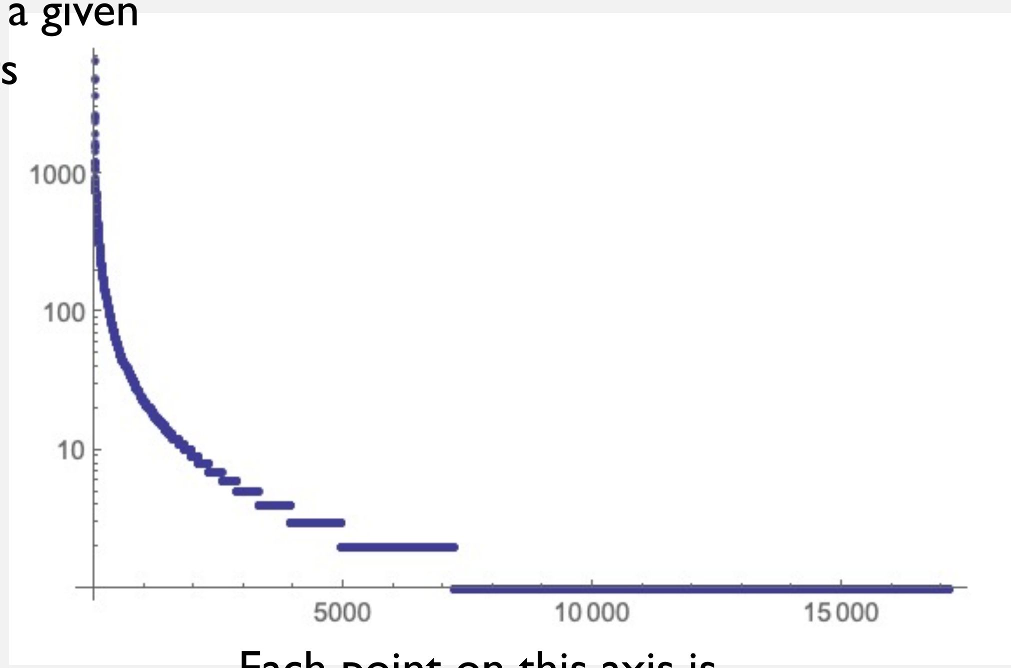
IT IS ALSO POSSIBLE FOR A TRIAL TO HAVE A UNIQUE COMBINATION OF TAGS, AND THIS IS THE CASE FOR THOUSANDS OF TRIALS. THERE IS EXACTLY ONE TRIAL WITH THIS SET OF TAGS:

- Offence category: violent theft
- Offence subcategory: highway robbery
- Defendant gender: indeterminate
- Victim gender: male
- Verdict: guilty
- Punishment: death

This is a different unique set of tags

A LONG-TAIL DISTRIBUTION

Number of trials matching a given set of tags



Each point on this axis is one unique set of tags

- There are 22 unique sets of tags that each represent more than one thousand trials. 48,653 trials in total are described by one of these 22 sets.
- 279 tag sets represent between 100 and 999 trials. 74,627 trials in total are described by one of these sets.
- 1636 tag sets represent between 10 and 99 trials. 44,858 trials in total are described by one of these sets.
- The remaining unique sets of tags represent between 1 and 9 trials. 29,607 trials in total are described by one of these sets.

STABLE RANDOM PROJECTIONS

- The Stable Random Projections (SRPs) of Ben Schmidt (2018) are a relatively new and powerful dimension reduction technique. Think of SRPs as sketches or fingerprints of documents.
- Each SRP represents a single trial (no matter how long or short) with a vector of 160 real numbers, and each of those numbers is a measurement along a particular dimension that encodes many aspects of the source text simultaneously.
- Besides being compact and very fast to compute, SRPs have the wonderful property of preserving important aspects of the text in a way that allows one to do useful work with them. Rather than comparing the texts directly, for example, you can compare their SRPs.

BAG OF DICTIONARY WORDS

Trial t18000528-98 as raw text:

437. SAMUEL ALDRIDGE was indicted for feloniously stealing, on the 19th of May , a cake of soap, value 1s. 6d. the property of Elizabeth Winterflood .

The prosecutrix not being able to indentify the soap, the prisoner was ACQUITTED .

Tried by the first Middlesex Jury, before Mr. COMMON SERJEANT.

Unordered collection of words that appear in the text and can also be found in English language dictionary:

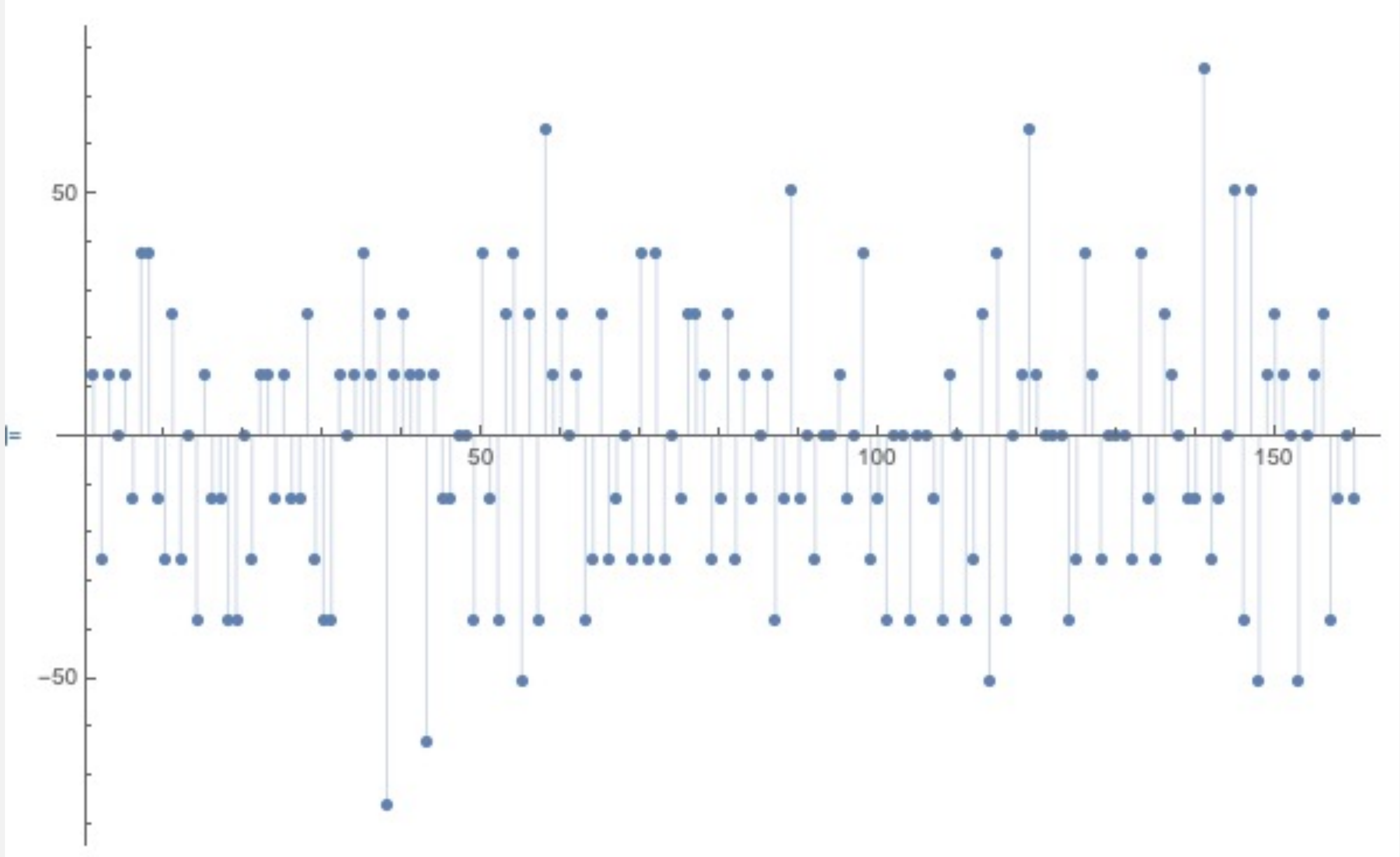
{"able", "acquitted", "cake", "common", "indicted", "jury", "prisoner", "property", "soap", "stealing", "tried", "value", "yes"}

Note that stopwords, proper names and misspelled words (e.g., "indentify") do not appear, and word order and frequency is lost.

HOW SRPS ARE CREATED

- The text of each trial is converted into a bag of dictionary words as shown on the previous slide
- The SRP is initialized to a vector of 160 zeros
- For each word in the bag of dictionary words…
  - Take its binary SHA-1 hash, which will be a vector of 160 zeros and ones. The same word will always return the same hash vector. Suppose the first word is 'bank'. The first eight bits of the SHA-1 hash for the word 'bank' are [1,0,1,1,1,1,0,1,…]
  - For each 1 in the SHA-1, add one to the SRP in the corresponding position. SRP is now [1,0,1,1,1,1,0,1,…]
  - For each 0 in the SHA-1, subtract one from the SRP in the corresponding position. SRP is now [1,-1,1,1,1,1,-1,1,…]
  - Go to the next word in the bag and repeat the process
- For each dimension of the SRP, the appearance of a given word in the text will reliably increase or decrease the score.
- "The net result of this process is that each dimension contains some information about the word counts for every word; the dimension is marginally higher if the bit for that dimension's SHA-1 hash is 1, and marginally lower if the bit is 0." (Schmidt p13)

SRP FOR A SINGLE TRIAL

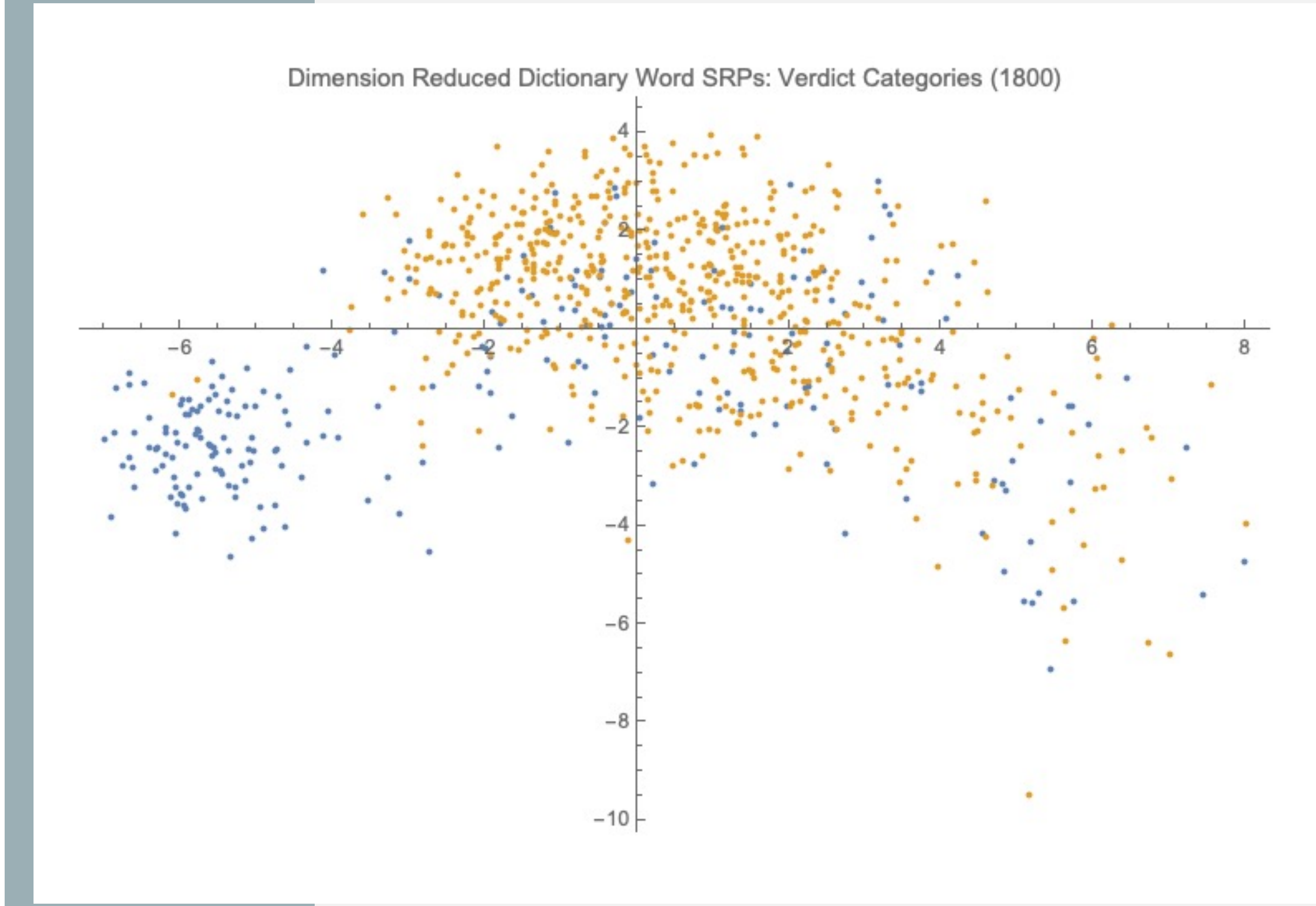


DIMENSION REDUCTION

In order to visualize SRPs, the 160 dimensions must be further reduced to two dimensions. There are several ways to do this. This dimension reduction step preserves the coherence of local clusters. "The x and y axes are arbitrary, but at both large and small scales the algorithm tries to position groups of similar documents near to each other. "

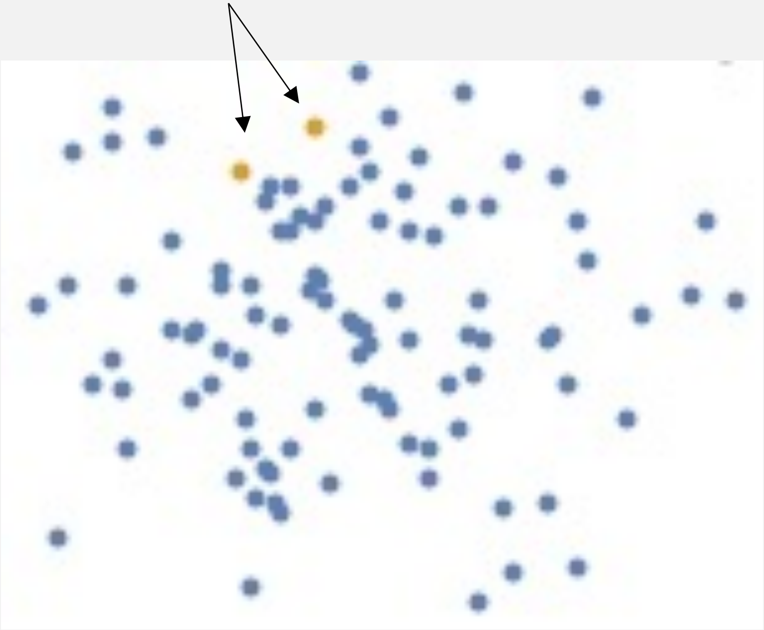
"If a cluster is coherent in the visualization, then it also exists in some sense in the higher-dimensional SRP space; relations that do not exist in the visualization may exist in the underlying data, or may be lost."
(quotes Schmidt p13, p21)

The visualization on the right shows SRPs for trials in the year 1800, not guilty verdict in blue and guilty verdict in orange.



ZOOMING IN TO EXPLORE OUTLIERS

If we zoom in to the previous figure, we see two guilty trials whose nearest neighbours in the SRP space are trials with not guilty verdicts



One of these two trials was of a pair of defendants, one who was found guilty and the other not guilty (t18001203-17). So, it is not surprising that this trial patterned with trials with a not guilty verdict.

The other of these two trials is discussed on the next slide.

We can use Cosine Distance on a pair of SRPs to assess how close the trials are to one another in the space. The trial with the lowest Cosine Distance from a trial of interest is its nearest neighbour.

NEAREST NEIGHBOURS

The other guilty verdict trial under consideration (t18000528-130) reads:

"469. WILLIAM MARKE was indicted for feloniously stealing, on the 27th of May , two loaves of sugar, value 13s. the property of William Jackson .

The prisoner, at the recommendation of his Counsel, pleaded GUILTY.

Judgment respited till next Sessions .

Tried by the first Middlesex Jury, before Mr. Baron CHAMBRE."

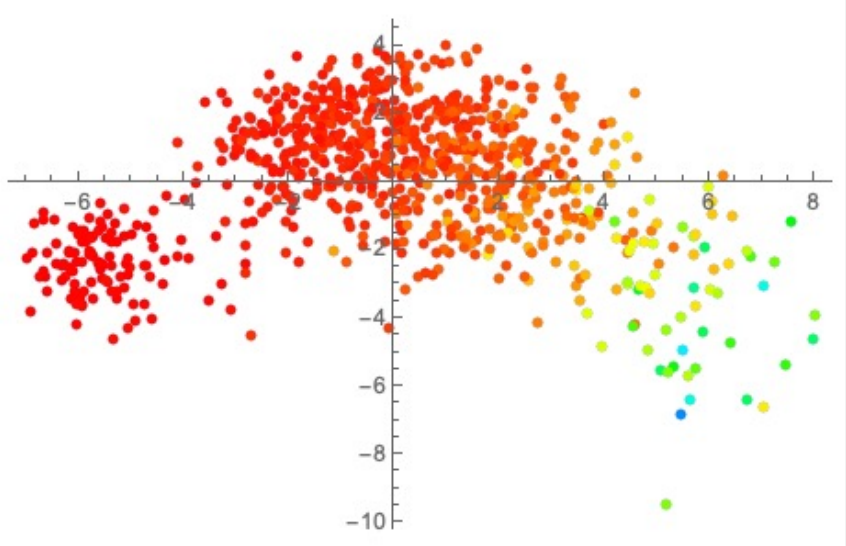
The three nearest trials to this one in SRP space for the year 1800 are t18000917-4, t18001029-48 and t18000528-98.

All are grand larceny trials that were tried by the first Middlesex Jury. In each of these trials, the prisoner was acquitted because of lack of evidence.

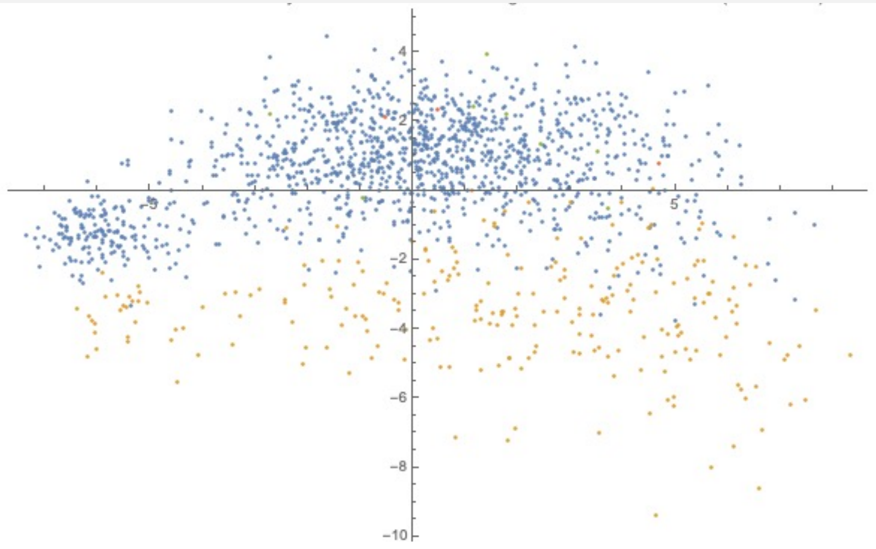
As we investigate outliers or anomalies like this, we generate new questions and avenues to explore. We don't know why the judgment was respited in this case, and William Marke's sentence outcome remains unknown. But we might investigate other similar cases where there was no punishment.

AUTOMATED CLUSTERING IN THE SPACE OF SRPS REVEALS PATTERNS MEANINGFUL TO HUMAN ANALYSTS

THE LENGTH OF A TRIAL IN WORDS (YEAR IS 1800, SHORTEST TRIALS ARE RED, THEN INCREASING LENGTHS ARE YELLOW, GREEN AND CYAN)



OFFENCE CATEGORIES (1780-1809) VIOLENT THEFT IN BLUE, KILLING IN ORANGE



CONCLUSION: USING THE TWO REPRESENTATIONS TOGETHER

- More generally, we are using SRPs in conjunction with the parameter space created by the XML tags to assess the representativeness of trials in particular periods of time
- This allows us to identify outliers and anomalies, as we showed above with the two guilty verdict trials that resembled ones with not guilty verdicts (that is to say, were nearest in SRP space to trials with not guilty verdicts)
- As Schmidt showed in his own examples, clusters in SRP space occur at a variety of scales, and can often be mapped onto classifications that are meaningful to human observers (e.g., represented by the XML tags)

ASK US ABOUT…

- We are happy to talk about anything presented here, of course, but there are also a lot of things we didn't have space to mention. Here are some other things we've been thinking about in conjunction with this work:
  - Changes in the digital humanities over the course of our research: new approaches to tagging, powerful new analytic methods, exponential growth of computing resources
  - The emergence of manslaughter
  - Punishments many and various
  - The rise of policing
  - Machine learning with SRP features

REFERENCES

- Schmidt, Benjamin. Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries. *Cultural Analytics* (4 Oct 2018) https://doi.org/10.22148/16.025