Quantitative Analysis of Text

genres are icebergs: with a visible portion floating above the water, and a much larger part hidden below, and extending to unknown depths.

Quantitative Analysis of text

“operationalize” a text

“concepts are transformed into a series of operations—which, in their turn, allow to measure all sorts of objects. Operationalizing means building a bridge from concepts to measurement, and then to the world”—Moretti

Quantitative Analysis of text

“operationalize” a text

Docuscope: Smart dictionary

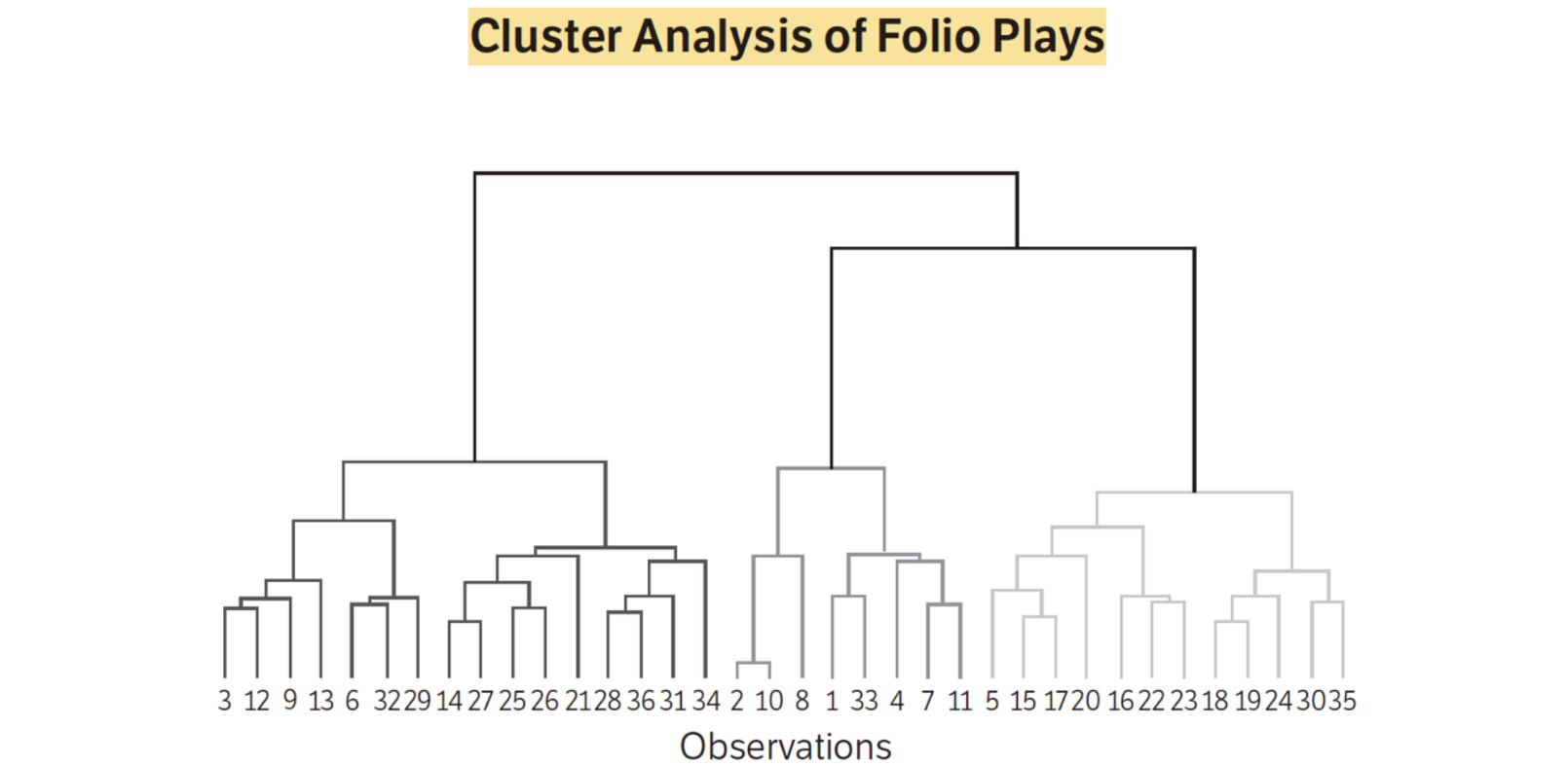
vs

MFW: Most frequent words

Quantitative Analysis of text

Unsupervised: Cluster Analysis

Quantitative Analysis of text



A Midsummer Night's Dream (3)
Twelfth Night (12)
Much Ado About Nothing (9)
Two Gentlemen (13)
Measure for Measure (6)
Othello (32)
Julius Caesar (29)

The Winter's Tale (14)
Cymbeline (27)
Antony and Cleopatra (25)
Coriolanus (26)
Henry VIII (21)
Hamlet (28)
Troilus and Cressida (36)
Macbeth (31)
Timon of Athens (34)

All's Well That Ends Well (2)
Taming of the Shrew (10)
Merry Wives of Windsor (8)
A Midsummer Night's Dream (1)
Romeo and Juliet (33)
Comedy of Errors (4)
Merchant of Venice (7)
The Tempest (11)

Love's Labours' Lost (5)
1 Henry IV (15)
2 Henry IV (17)
Henry V (20)
1 Henry VI (16)
King John (22)
Richard II (23)

2 Henry VI (18)
2 Henry VI (19)
Richard III (24)
King Lear (30)
Titus Andronicus (35)

Figure 1: Dendrogram illustrating clustering of Shakespeare plays rated on Docuscope's Language Action Types (LATs) produced in 2003. Clustering method: complete linkage, Euclidean distances. Notice the presence of comedies in the first and third columns, late plays and tragedies in the second, and histories in the fourth and fifth. "Incorrect classifications" such as *Othello* and *Love's Labours' Lost* are discussed on Witmore's blog, www.

Quantitative Analysis of text

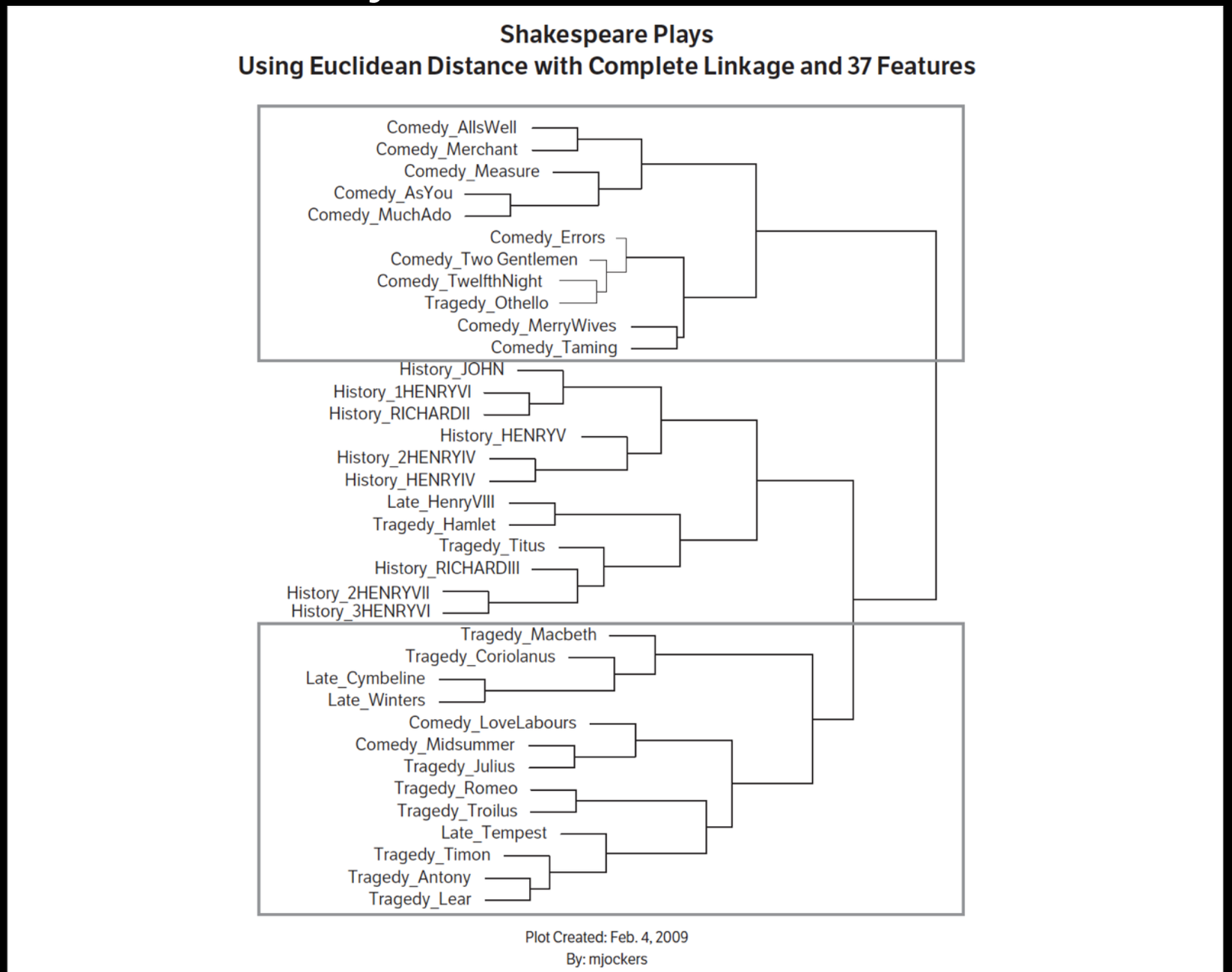


Figure 3.2: Dendrogram of Shakespeare First Folio plays using Most Frequent Words with major clusters highlighted. Here Jockers used the 37 features from the Shakespeare plays that had a mean relative frequency of greater than or equal to .03%. Note the similarity between this tree and Docuscope's diagram in fig. 1.1, with the close pairings of *Winter's Tale* and *Cymbeline*; *2 Henry VI* and *3 Henry VI*, and the proximity of *Coriolanus* to the *Cymbeline-Winter's Tale* pair.

As soon as school was over, we met again.

Quantitative Analysis of text

PCA: Principal components analysis

Quantitative Analysis of text

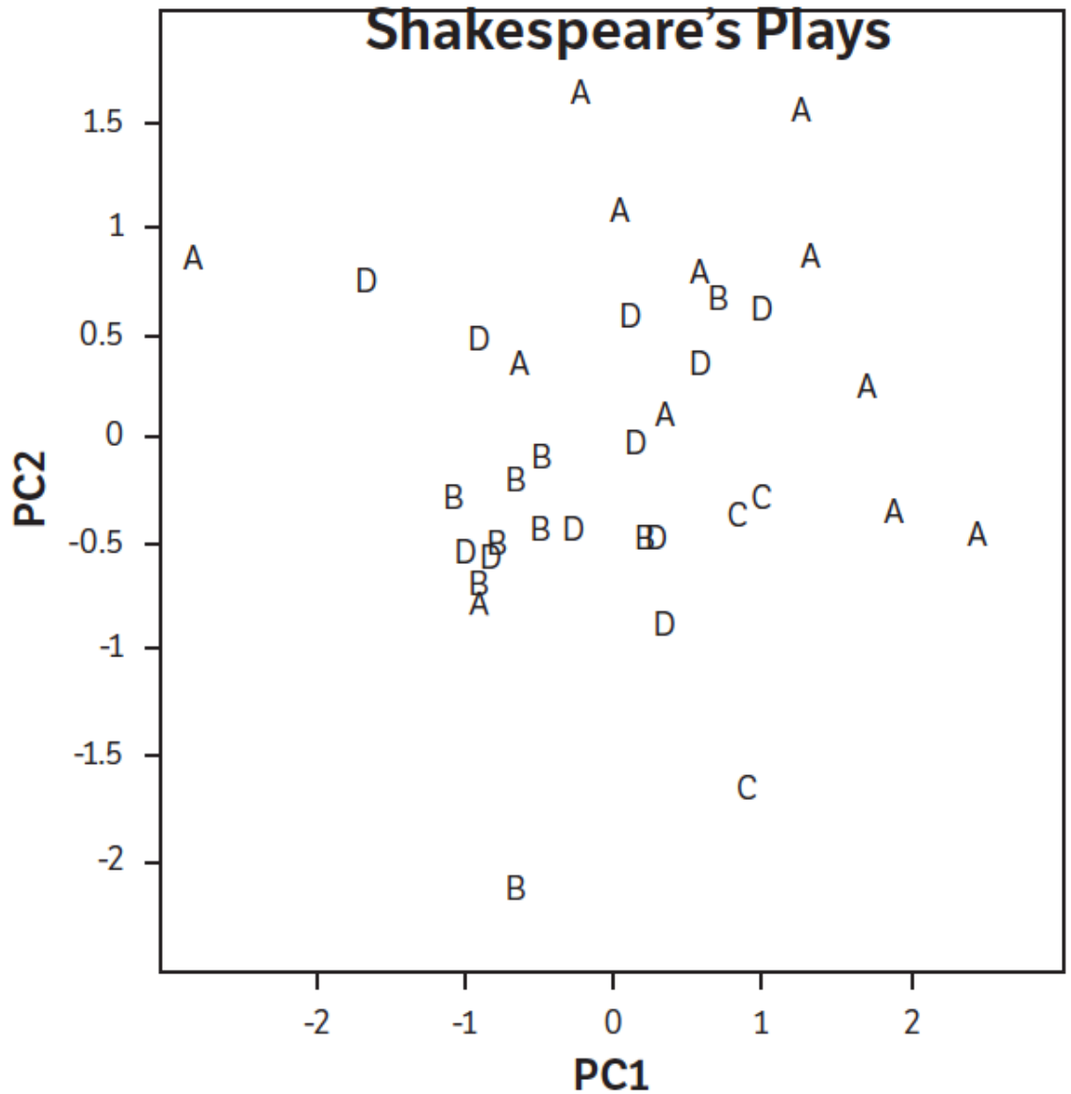


Figure 4.2: Scatterplot matrix in which Shakespeare's plays are rated on their first two principal components after having been counted by Docuscope and analyzed in terms of aggregates of LATs. PCA performed on the covariance matrix, unscaled data. Item key: A = comedy, B = History, C = Late Plays, D = Tragedies. Note how the two components place comedies in the upper right quadrant, histories in the lower left, and several late plays in the lower right (whereas tragedies, for some reason, are dispersed all over the field).

Quantitative Analysis of text

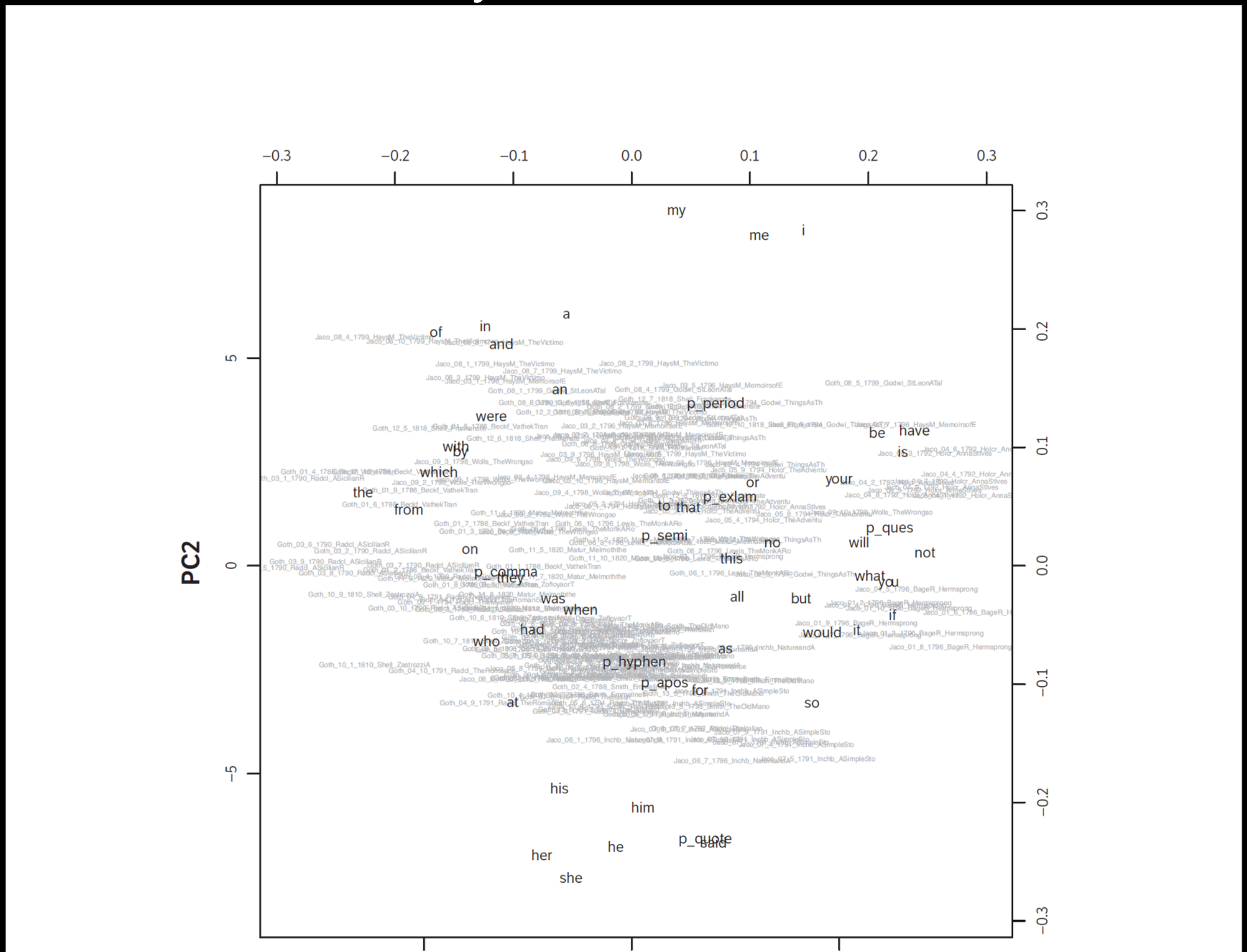
Category: genre

genres, like buildings, possess distinctive features at every possible scale of analysis: mortar, bricks, and architecture

Quantitative Analysis of text

Category: genre — 19th-century novel

Quantitative Analysis of text



Quantitative Analysis of text

Category: genre

Did we think they had produced new knowledge?

The answer, of course, was NO

Quantitative Analysis of text

Category: genre

Nothing in common in terms of units of analysis, everything in common in terms of results…

Quantitative Analysis of text

Category: genre, “the great unread”

One could give Docuscope and MFW thousands of texts of unknown generic affiliation, and see where they would fall in the gravitational field of be!er-known genres. One could envisage generation-by-generation maps of the literary universe, with galaxies, supernovae, black holes...

Quantitative Analysis of text

Category: genre

Roughly speaking, we found that
the gothic novel averages less talk
and more action than
the Jacobin.

Trying to understand how a computer
reads, reading into the results,

Rough speech is the explanatory
version of looking for constellations
the random stars

Quantitative Analysis of text

Category: genre breaks down

Separation of the genres, not genres themselves

Quantitative Analysis of text

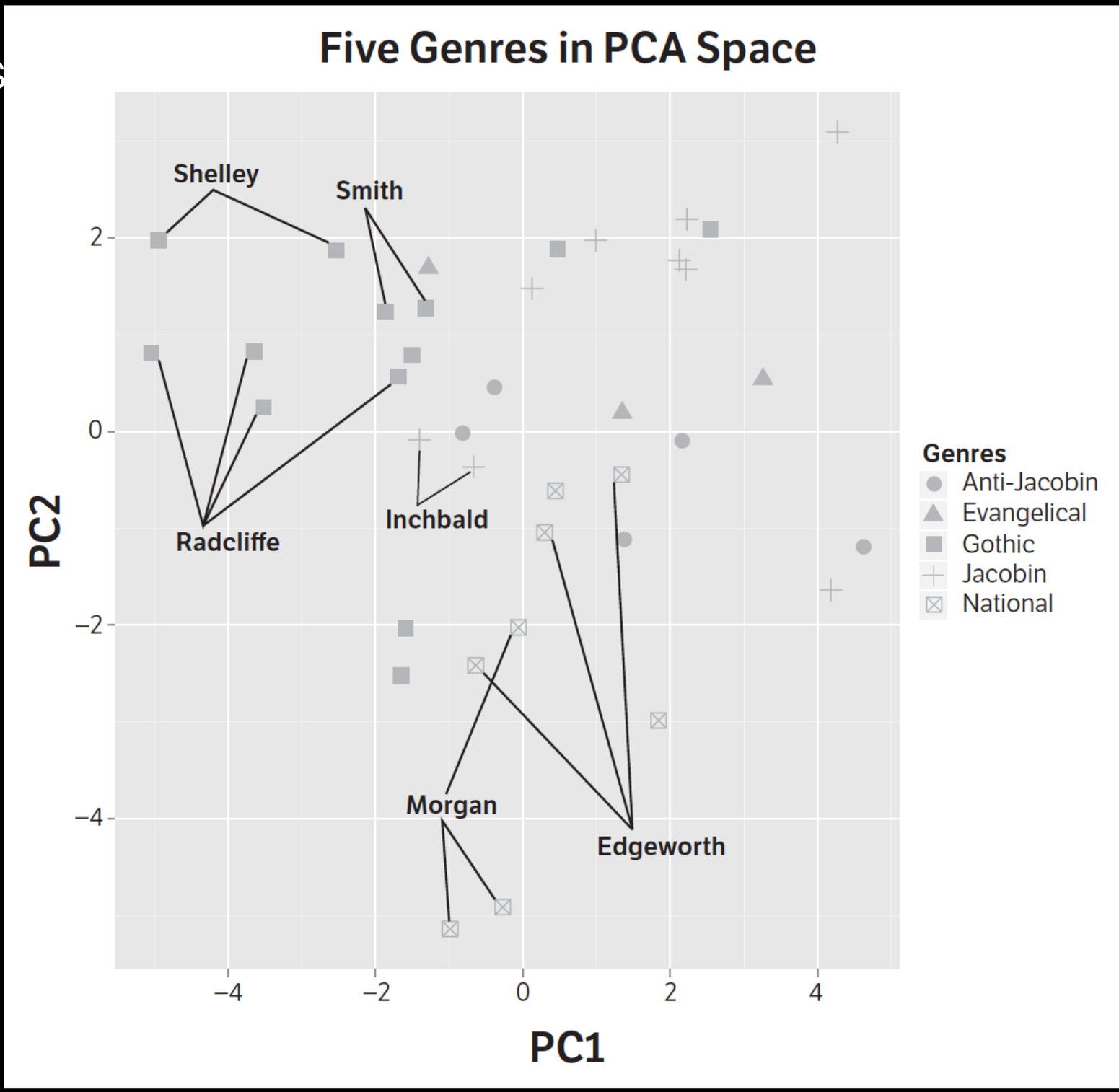
Category: genre breaks down

Why should authors be so much more recognizable than genres?

Probably, because Docuscope and MFW are very good at capturing something all writers do, whether theyknow it or not: using imperceptible linguistic pa!erns that provide an unmistakable stylistic “signature”

Quantitative Analysis of text

Category: genre breaks



Quantitative Analysis of text

Category: genre breaks down

Why did it do so “well” with
Shakespeare and so poorly with
19th-century novels?

Quantitative Analysis of text

Category: genre breaks down

Why did it do so “well” with Shakespeare and so poorly with 19th-century novels?

Also, authors exist, but do genres?

Oxford Shakespeare: whether a society should value imaginative reading

Quantitative Analysis of text

Question of domain: taxonomy

The question of what you were studying--genres of the single author, authors across time, genres across time

Quantitative Analysis of text

Domains: Chronology, Author, Genre

Earlobes, fingernails ... “involuntary signs”

There is a problem with earlobes and fingernails: good as they might be at identifying the author of a painting, they are worthless at explaining its **meaning**.

Quantitative Analysis of text

The experiment then turned into an **exploration**

There is something paradoxical in these traits that classify so well, and explain so little.

Assigning values/ quantities to language the only way to run computer read processes

Final projects: building a data set

The experiment then turned into an **exploration**

Final projects: building a data set

exploratory analysis

Measurements —> Meaning

1. Reading / Research: **choosing your subject/domain**

"Pericles," and the "Two Noble Kinsmen" should be thrust in, the last wrote by him and Fletcher, if he wrote not the greater part of that play he never wrote a line).

A Catalogue of all the Comedies, Histories, and Tragedies contained in this Booke.

THE TEMPEST.

Actus Primus. Scæna Prima.

A ship at sea

A tempestuous noise of Thunder and Lightning heard: Enter a Ship-master, and a Boteswaine.

Master.

Bote-swaine.

Bote. Heere Master: What cheere?

Mast. Good: Speake to th'Mariners: fall too't, yarely, or we run our selves a ground, bestirre, bestirre. *Exit.*

Enter Mariners.

Bote. Heigh my hearts, cheerely, cheerly my hearts: yare, yare: Take in the toppe-sale: Tend to th'Masters whistle: Blow till thou burst thy winde, if roome enough.

Enter Alonso, Sebastian, Anthonio, Ferdinando, Gonzalo, and others.

Alon. Good Bote-swaine have care: wher's the Master? Play the men.

Bote. I pray now keepe below.

Ant. Where is the Master, Boson?

Bote. Do you not heare him? you marre our labour, Keepe your Cabins: you do assist the storme.

Gonz. Nay good be patient.

Bote. When the Sea is: hence, what cares these roarers for the name of King? to Cabine; silence: trouble us not.

Gon. Good yet remember whom thou hast aboord.

Bote. None that I more love then my selfe. You are a Counsellor, if you can com nand these Elements to silence, and worke the peace of the present, wee will not hand a rope more, use your authoritie: If you cannot, give thankes you have liv'd so long, and make your selfe readie in your Cabine for the mischance of the houre, if it so hap. Cheerely good hearts: out of our way I say. *Exit.*

Gon. I have great comfort from this fellow: methinkes he hath no drowning marke upon him, his complexion is perfect Gallowes: stand fast good Fate to his hanging, make the rope of his destiny our Cable, for our owne doth little advantage: If he be not borne to bee hang'd, our case is miserable. *Exit.*

Enter Boteswaine.

Bote. Downe with the top-Mast: yare, lower, lower.

upon this howling: they are lowder then the weather, or our office: yet againe? What do you heere? Shall we give ore and drowne, have you a minde to sinke?

Sebast. A poxe o'your throat, you bawling, blasphemous incharitable Dog.

Bote. Worke you then.

Ant. Hang cur, hang, you whoreson insolent Noysemaker; we are lesse afraid to be drownde, then thou art.

Gonz. I'le warrant him for drowning, though the Ship were no stonger then a Nutt-shell, and as leaky as an unstanched wench.

Bote. Lay her a hold, a hold, set her two courses off to Sea againe, lay her off.

Enter Mariners wet.

Mari. All lost, to prayers, to prayers, all lost. *Exeunt.*

Bote. What must our mouths be cold?

Gon. The King, and Prince, at prayers, let's assist them, for our case is as theirs.

Sebast. I am out of patience.

Ant. We are meerely cheated of our lives by drunkards, This wide-chopt-rascall, would thou mightst lye drowning the washing of ten Tides.

Gonz. Hee'l be hang'd yet,
Though every drop of water sweare against it,
And gape at widst to glut him. *A confused noyse within.*
Mercy on us.
We split, we split, Farewell my wife, and children,
Farewell brother: we split, we split, we split.

Ant. Lets all sinke with' King *Exit.*

Seb. Let's take leave of him. *Exit.*

Gonz. Now would I give a thousand furlongs of Sea, for an Acre of barren ground: Long heath, Browne firrs, any thing; the wills above be done, but I would faine dye a dry death.

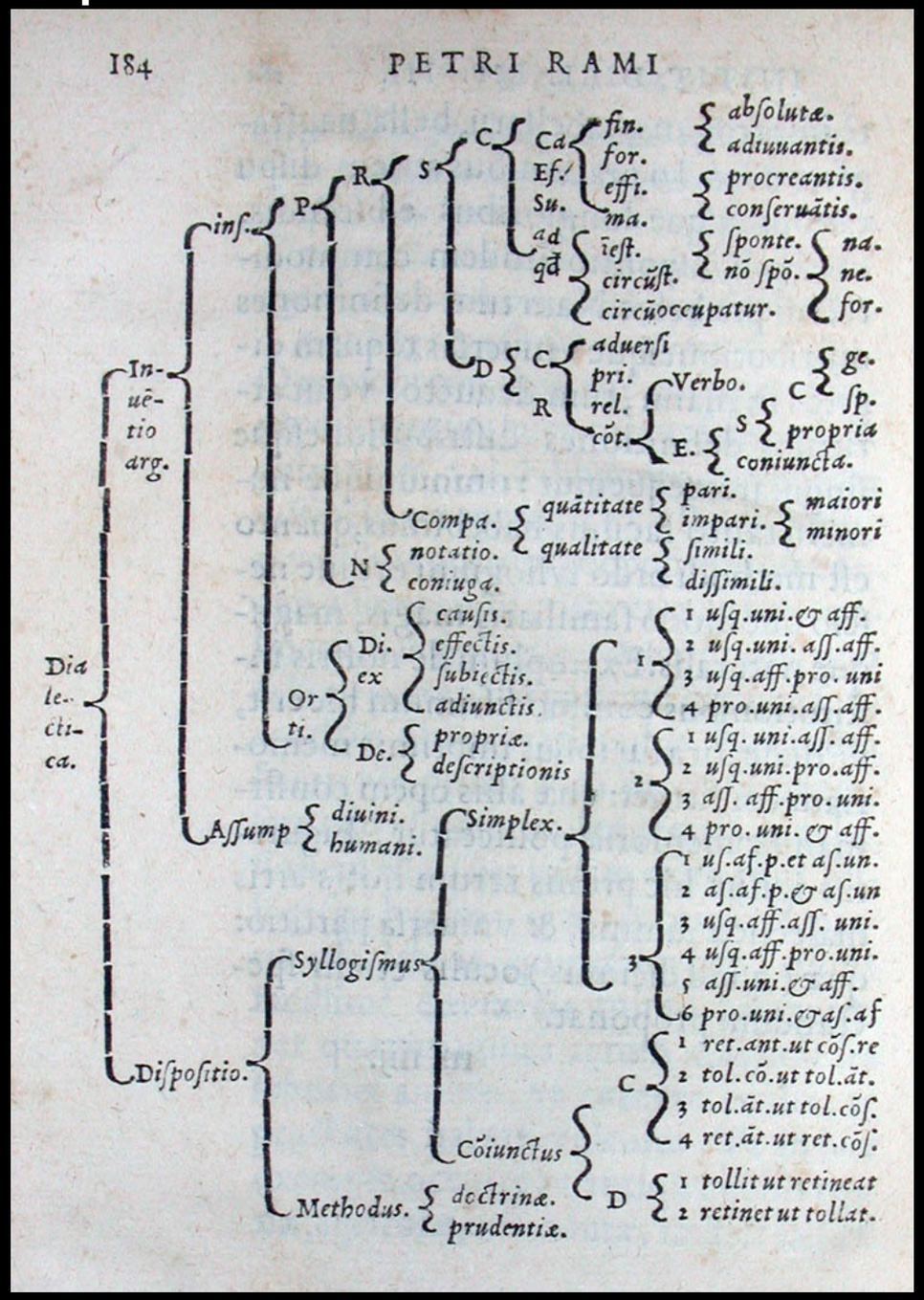
Scæna Secunda.

Enter Prospero and Miranda.

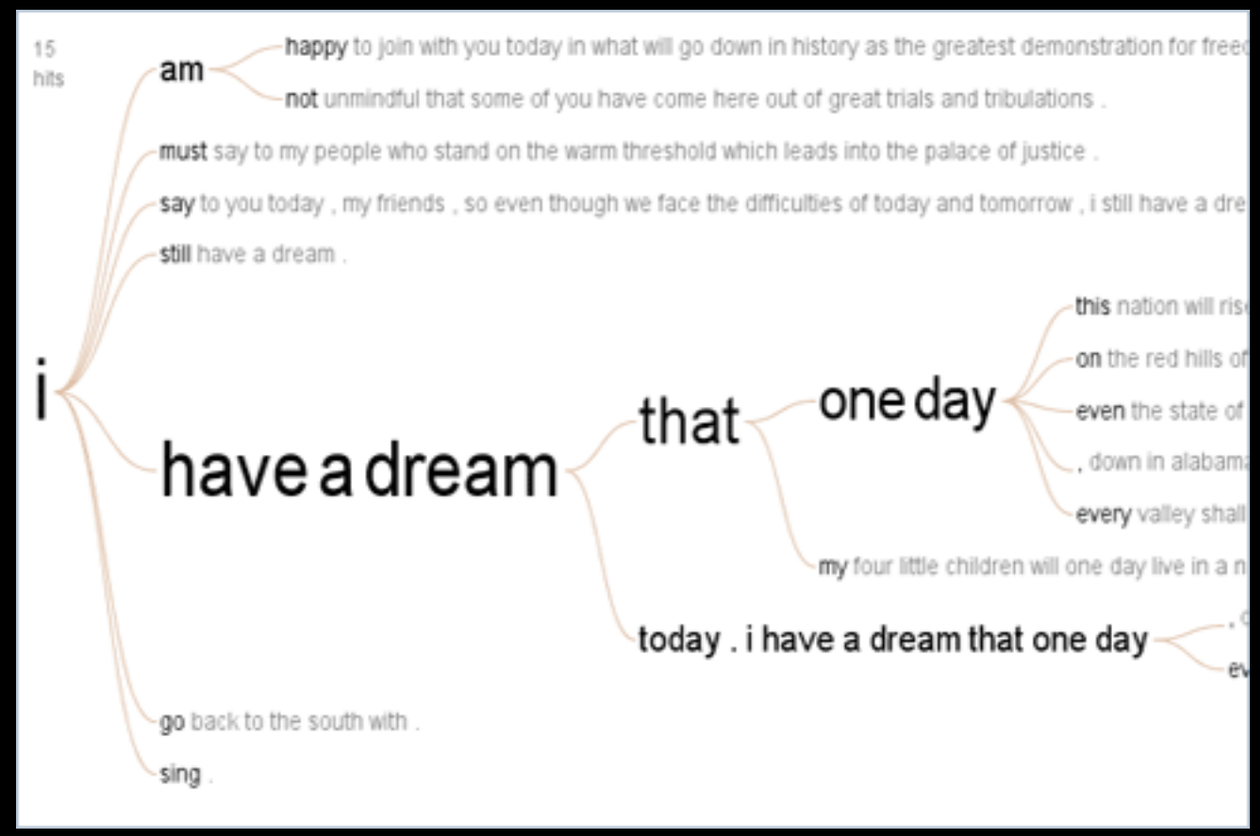
Mira. If by your Art (my deerest father) you have
Put the wilde waters in this Rore; alay them:
The skye it seemes would powre downe stinking pitch.
But that the Sea, mounting to th' welkins cheeke,

Measurements —> Meaning

2. Organizing / Taxonomies: **finding the data and imagining a shape for it**



Digitization



Left: Petrus Ramus, classification structure from *Dialectique,* 1555
Right: Fernanda Viegas & Martin Wattenberg, Word Tree 2007

Measurements —> Meaning

3. Building / Data Structures: **designing your data architecture**

```
mondial=# SELECT country, max(population)
mondial-#     FROM city
mondial-#     WHERE population IS NOT NULL
mondial-#     GROUP BY country
mondial-#     ORDER BY country
mondial-#     LIMIT 15;
 country |   max
---------+---------
 A       | 1761738
 AFG     | 2435400
 AG      |   22219
 AL      |  418495
 AND     |   22256
 ANG     | 2107648
 ARM     | 1066264
 AUS     | 4605992
 AZ      | 2150800
 B       |  507911
 BD      | 7423137
 BDS     |   88529
 BEN     |  665100
 BF      | 1475223
 BG      | 1270284
(15 rows)

mondial=#
```

Measurements —> Meaning

4. Searching / Aggregating

File Edit View Insert Cell Kernel Help Trusted Python 3

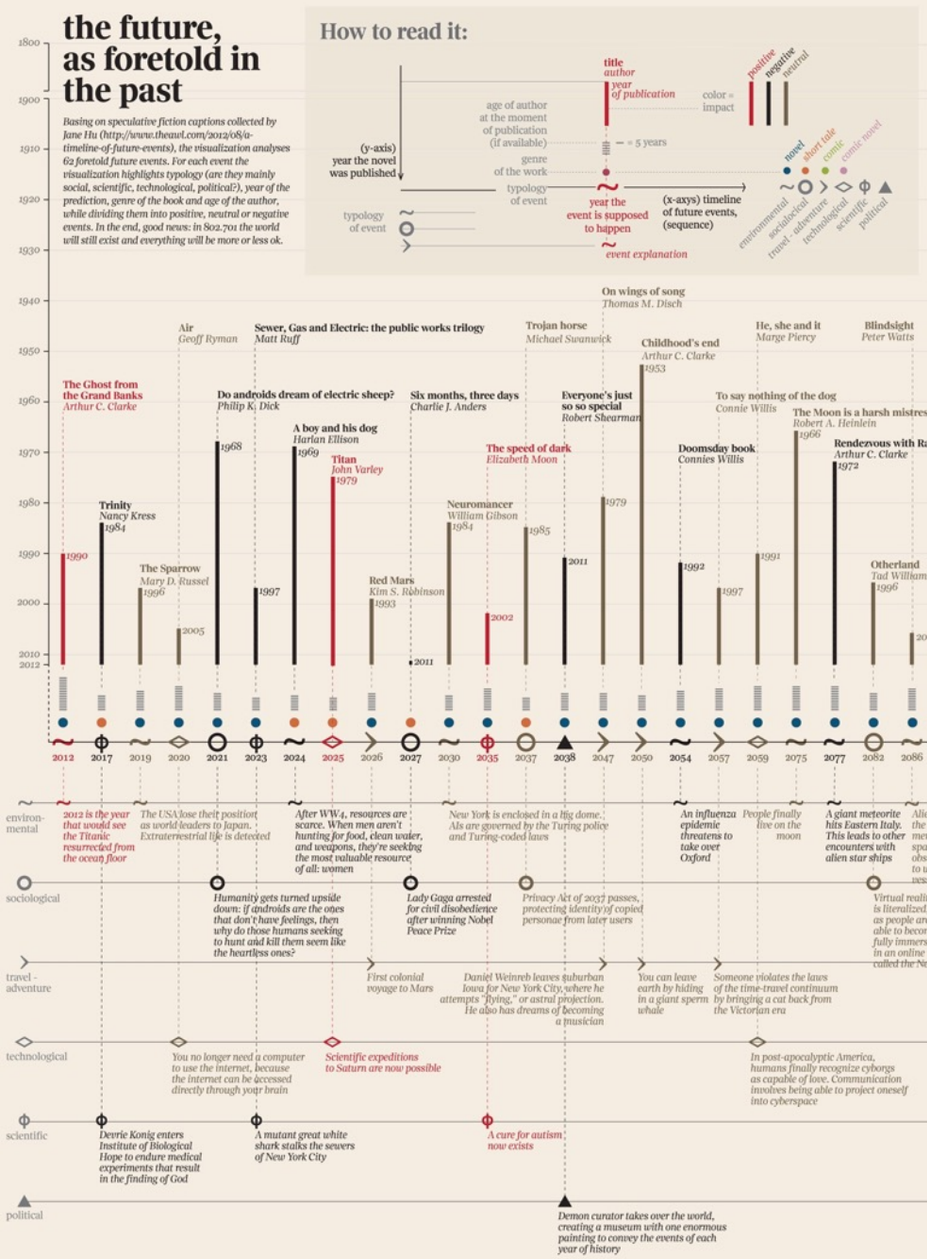
Code

In [5]:
```
urls = []
url_failed = []
for director in directors:
    search_d = director.lower().replace(" ","+")
    url_string = 'http://www.imdb.com/find?&q=' + search_d + '&s=all'
    search_name = director.lower()
    raw_html = urlopen(url_string).read()
    soup_doc = BeautifulSoup(raw_html, "html.parser")
    next_url = soup_doc.find(class_='result_text')
    if next_url is not None:
        good_url = next_url.a['href']
        cleaner_url = good_url.split("?ref")
        if cleaner_url[0].startswith('/name'):
            urls.append(cleaner_url[0])
        else:
            url_failed.append(search_name)
        print(cleaner_url)
```

Measurements —> Meaning

5. Displaying / Generating Knowledge

Design/Aesthetics
Interactivity/Screen Space

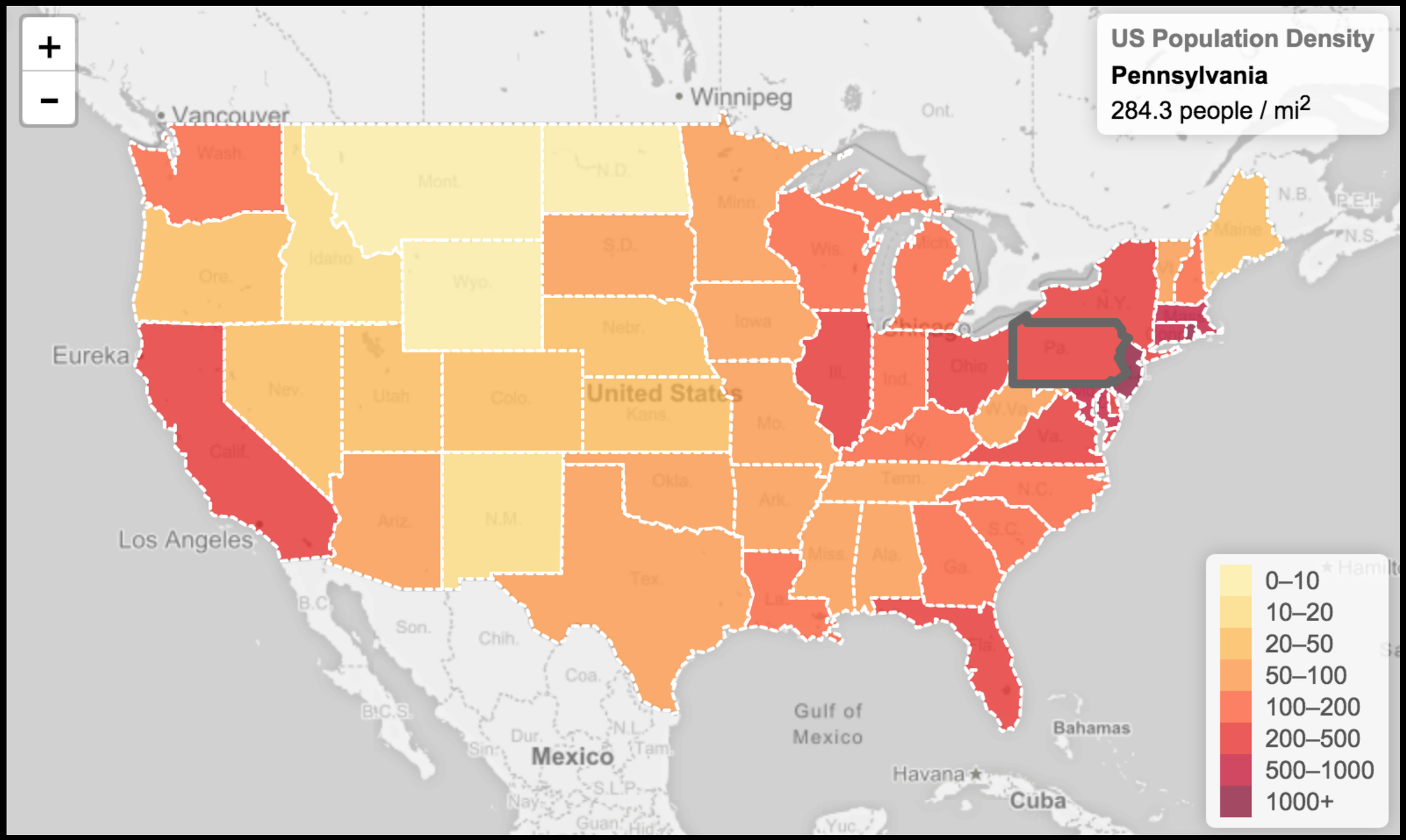


Giorgia Lupi

Measurements —> Meaning

Interactive map in leaflet.js

5. Displaying / Generating Knowledge



Final projects: building a data set

exploratory analysis > **map as visual presentation**

Maps are an artificial category that you must adhere to…

There may be a visual form more well suited to your subject, but we all agree the world exists…

Interrogating the form screen-based maps…

All screen/visual representations of data are a type of mapping…

Your focus will be building output for the map, not on the design

Final projects: building a data set

Choosing a subject

1. what is the subject of the project?
2. what is your main research question?
3. what is/are your data source(s)?
4. how will you transform the data into your own data set: scraping, regex, etc?
5. what will your architecture be?
6. how can this data set be geocoded?
7. on the map, what level of study would be displayed: Country, State, City, neighborhood, etc?
8. what information would be displayed when you click on/rollover a country (city, etc)?

Final projects: building a data set

Choosing a subject

1. what is the subject of the project?
2. what is your main research question?
3. what is/are your data source(s)?
4. how will you transform the data into your own data set: scraping, regex, etc?
5. **what will your architecture be?**
6. how can this data set be geocoded?
7. on the map, what level of study would be displayed: Country, State, City, neighborhood, etc?
8. what information would be displayed when you click on/rollover a country (city, etc)?

Final projects: building a data set

Levels of information the map can contain (properties)

Geometry--Points/Shapes: these are combinations of latitudes and longitudes that leaflet will translate to the screen for you.

Colors/size: these are the obvious ways to visually separate different points or shapes.

Groups: Groups allow you to have different layers of shapes/ points visible.

Headline/Lede: when you roll over a point or shape this is the most immediate information that you want to convey

Article: this is for deeper reading/analysis--on click you can get as much information as you want to output on the page

See sample simple maps and geojson output —note they are intentionally unimpressive