Stylometry with R

Part 1. Textual Forces

Joanna Byszuk, Artjoms Šeļa and Maciej Eder

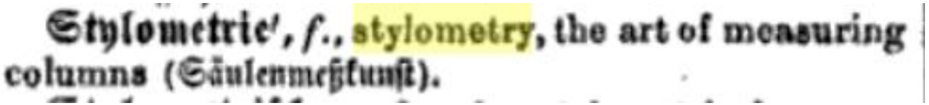
What is stylometry?

- A sub-field of computational text analysis that studies **differences** between texts
- Lutosławski 1897: method of “measuring stylistic affinities”

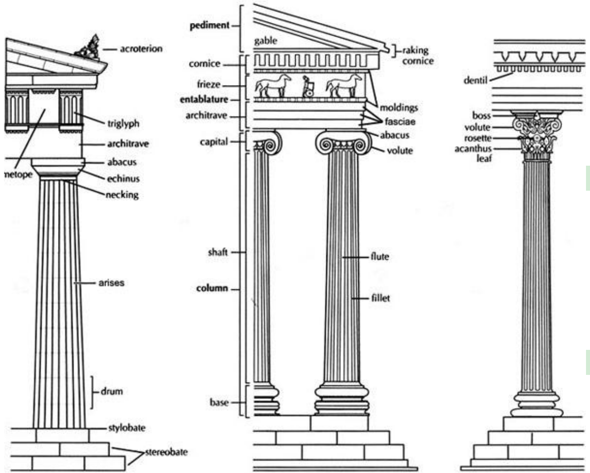
What is stylometry?

- A sub-field of computational text analysis that studies **differences** between texts
- Lutosławski 1897: method of “measuring stylistic affinities”

Don’t mix up with another “stylometry”: “the art of measuring columns”!



Stylometrie', *f.*, stylometry, the art of measuring columns (Säulenmeßkunſt).



A model of text

- Differences between texts can be expressed in a multitude of ways
- Central question is **how to represent** a text so it could be placed on a quantitative scale? I.e. how to **model** it?
- Short answer: all representations are ‘wrong’, but some are useful (or more useful than others)

A model of text

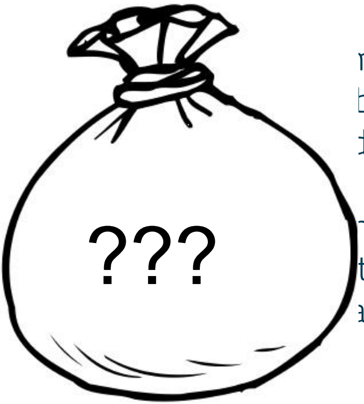
- Differences between texts can be expressed in a multitude of ways
- Central question is **how to represent** a text so it could be placed on a quantitative scale? I.e. how to **model** it?
- Short answer: all representations are ‘wrong’, but some are useful (or more useful than others)
  - Word frequencies?
  - Algorithmically inferred topics?
  - Part-of-Speech tags?
  - Networks of character connections?
  - Embeddings?
  - Sentiment scores?
  - ….

Silly things: bags of words

Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a “Penang lawyer.” Just under the head was a broad silver band nearly an inch across. “To James Mortimer, M.R.C.S., from his friends of the C.C.H.,” was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry – dignified, solid, and reassuring.

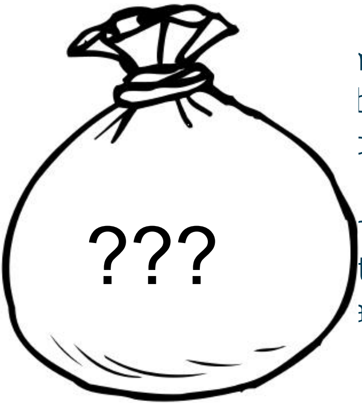
Silly things: bags of words

Mr. Sherlock Holmes, who was mornings, save upon those not infrequent
occasions when he was up all breakfast table. I stood upon the
hearth-rug and picked up the s d left behind him the night before. It was
a fine, thick piece of wood, bul which is known as a “Penang lawyer.”
Just under the head was a broa nch across. “To James Mortimer, M.R.C.S.,
from his friends of the C.C.H.,” th the date "1884." It was just such a
stick as the old-fashioned fam arry – dignified, solid, and reassuring.



Silly things: bags of words

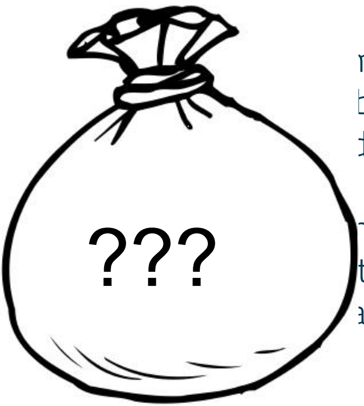
Mr. Sherlock Holmes, who was nornings, save upon those not infrequent
occasions when he was up all breakfast table. I stood upon the
hearth-rug and picked up the s d left behind him the night before. It was
a fine, thick piece of wood, bul which is known as a “Penang lawyer.”
Just under the head was a broa nch across. “To James Mortimer, M.R.C.S.,
from his friends of the C.C.H.,” th the date "1884." It was just such a
stick as the old-fashioned fam arry – dignified, solid, and reassuring.



the:	10
was:	7
a:	4
it:	3
of:	3
upon:	3
and:	2
as:	2
...	

Silly things: bags of words

Mr. Sherlock Holmes, who was nornings, save upon those not infrequent
occasions when he was up all breakfast table. I stood upon the
hearth-rug and picked up the s d left behind him the night before. It was
a fine, thick piece of wood, bul which is known as a "Penang lawyer."
Just under the head was a broa nch across. "To James Mortimer, M.R.C.S.,
from his friends of the C.C.H.," th the date "1884." It was just such a
stick as the old-fashioned fam arry – dignified, solid, and reassuring.

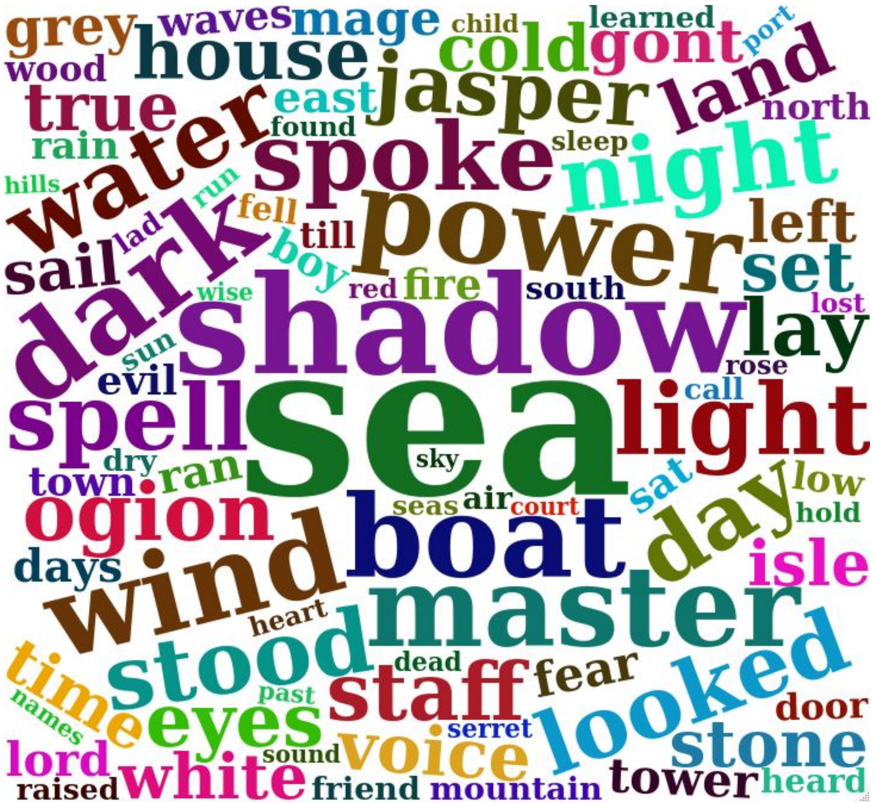


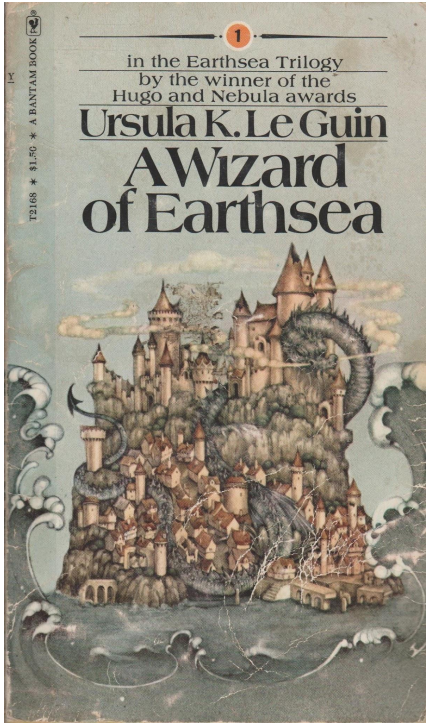
Doyle_Baskerville_p1: (10, 7, 4, 3, 3, 3, 2, 2,...)

Silly, but useful!

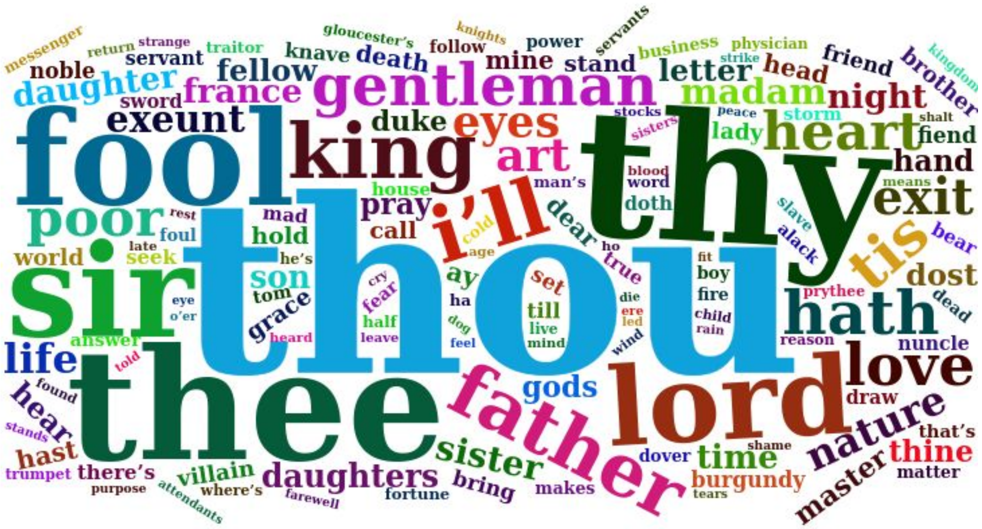
grey
waves
mage
child
learned
port
wood
house
cold
gont
land
true
east
jasper
north
found
rain
spoke
sleep
water
night
hills
run
fell
left
lad
till
power
sail
dark
boy
set
wise
red
fire
south
lost
shadow
lay
sun
rose
evil
call
spell
sea
light
dry
sky
town
ran
sat
low
seas
air
court
hold
ogion
boat
day
days
isle
wind
master
heart
dead
time
stood
fear
past
staff
looked
names
door
eyes
serret
lord
sound
voice
stone
white
raised
friend
mountain
tower
heard

Silly, but useful!

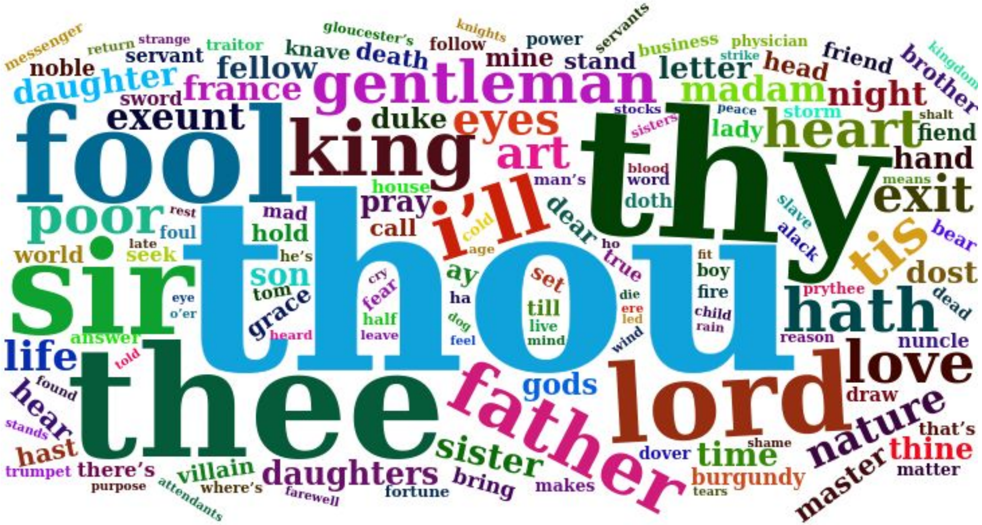


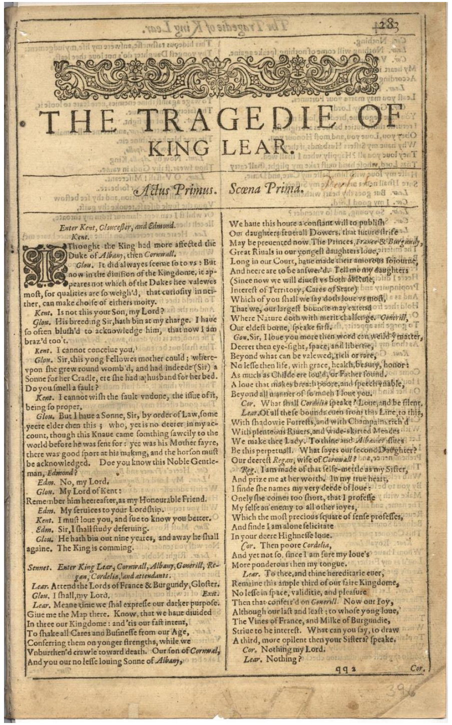


Silly, but useful!



Silly, but useful!





283

THE TRAGEDIE OF KING LEAR.

Actus Primus. Scœna Prima.

Enter Kent, Glouceſter, and Edmond.

Kent.

I Thought the King had more affected the Duke of *Albany*, then *Cornwall.*

Glou. It did alwayes ſeeme ſo to vs: But now in the diuiſion of the Kingdome, it appeares not which of the Dukes hee valewes moſt, for qualities are ſo weigh'd, that curioſity in neither, can make choiſe of eithers moity.

Kent. Is not this your Son, my Lord?

Glou. His breeding Sir, hath bin at my charge. I haue ſo often bluſh'd to acknowledge him, that now I am braz'd too't.

Kent. I cannot conceiue you.

Glou. Sir, this yong Fellowes mother could; whereupon ſhe grew round womb'd, and had indeede (Sir) a Sonne for her Cradle, ere ſhe had a husband for her bed. Do you ſmell a fault?

Kent. I cannot wiſh the fault vndone, the iſſue of it, being ſo proper.

Glou. But I haue a Sonne, Sir, by order of Law, ſome yeere elder then this; who, yet is no deerer in my account, though this Knaue came ſomthing ſawcily to the world before he was ſent for: yet was his Mother fayre, there was good ſport at his making, and the horſon muſt be acknowledged. Doe you know this Noble Gentleman, *Edmond*?

Edm. No, my Lord.

Glou. My Lord of Kent:
Remember him heereafter, as my Honourable Friend.

Edm. My ſeruices to your Lordſhip.

Kent. I muſt loue you, and ſue to know you better.

Edm. Sir, I ſhall ſtudy deſeruing.

Glou. He hath bin out nine yeares, and away he ſhall againe. The King is comming.

Sennet. Enter King Lear, Cornwall, Albany, Gonerill, Regan, Cordelia, and attendants.

Lear. Attend the Lords of France & Burgundy, Gloſter.

Glou. I ſhall, my Lord. *Exit.*

Lear. Meane time we ſhal expreſſe our darker purpoſe.
Giue me the Map there. Know, that we haue diuided
In three our Kingdome: and 'tis our faſt intent,
To ſhake all Cares and Buſineſſe from our Age,
Conferring them on yonger ſtrengths, while we
Vnburthen'd crawle toward death. Our ſon of *Cornwal*,
And you our no leſſe louing Sonne of *Albany*,
We haue this houre a conſtant will to publiſh
Our daughters ſeuerall Dowers, that future ſtrife
May be preuented now. The Princes, *France & Burgundy*,
Great Riuals in our yongeſt daughters loue,
Long in our Court, haue made their amorous ſoiourne,
And heere are to be anſwer'd. Tell me my daughters
(Since now we will diueſt vs both of Rule,
Intereſt of Territory, Cares of State)
Which of you ſhall we ſay doth loue vs moſt,
That we, our largeſt bountie may extend
Where Nature doth with merit challenge. *Gonerill*,
Our eldeſt borne, ſpeake firſt.

Gon. Sir, I loue you more then word can weild ÿ matter,
Deerer then eye-ſight, ſpace, and libertie,
Beyond what can be valewed, rich or rare,
No leſſe then life, with grace, health, beauty, honor:
As much as Childe ere lou'd, or Father found.
A loue that makes breath poore, and ſpeech vnable,
Beyond all manner of ſo much I loue you.

Cor. What ſhall *Cordelia* ſpeake? Loue, and be ſilent.

Lear. Of all theſe bounds euen from this Line, to this,
With ſhadowie Forreſts, and with Champains rich'd
With plenteous Riuers, and wide-skirted Meades
We make thee Lady. To thine and *Albanies* iſſues
Be this perpetuall. What ſayes our ſecond Daughter?
Our deereſt *Regan*, wife of *Cornwall*?

Reg. I am made of that ſelfe-mettle as my Siſter,
And prize me at her worth. In my true heart,
I finde ſhe names my very deede of loue:
Onely ſhe comes too ſhort, that I profeſſe
My ſelfe an enemy to all other ioyes,
Which the moſt precious ſquare of ſenſe profeſſes,
And finde I am alone felicitate
In your deere Highneſſe loue.

Cor. Then poore *Cordelia*,
And yet not ſo, ſince I am ſure my loue's
More ponderous then my tongue.

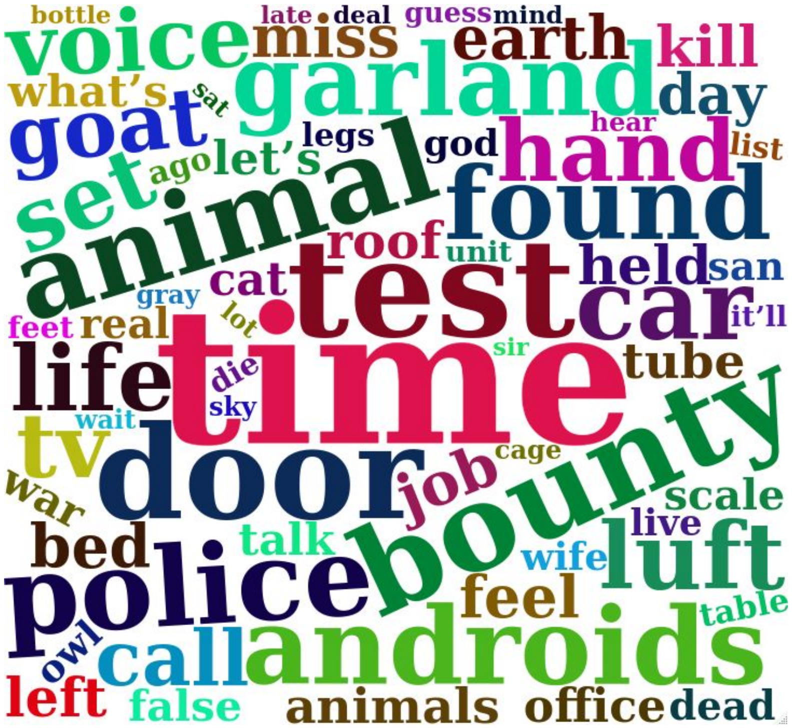
Lear. To thee, and thine hereditarie euer,
Remaine this ample third of our faire Kingdome,
No leſſe in ſpace, validitie, and pleaſure
Then that conferr'd on *Gonerill*. Now our Ioy,
Although our laſt and leaſt; to whoſe yong loue,
The Vines of France, and Milke of Burgundie,
Striue to be intereſt. What can you ſay, to draw
A third, more opilent then your Siſters? ſpeake.

Cor. Nothing my Lord.

Lear. Nothing?

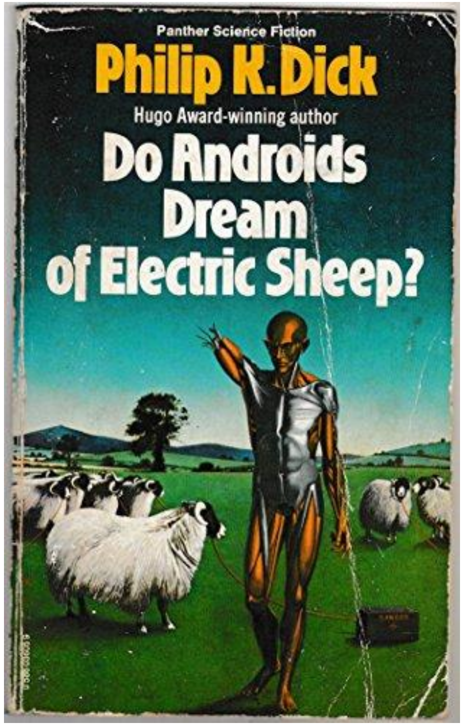
qq2 Cor.

Silly, but useful!



Silly, but useful!

bottle
late
deal
guess
mind
voice
miss
earth
kill
garland
what's
sat
day
goat
hear
legs
god
hand
list
set
ago
let's
animal
found
roof
unit
held
san
cat
gray
test
car
feet
real
lot
it'll
sir
life
time
tube
die
sky
wait
tv
cage
door
job
bounty
war
scale
live
bed
talk
wife
luft
police
feel
table
owl
call
androids
left
false
animals
office
dead


Panther Science Fiction
Philip K. Dick
Hugo Award-winning author
Do Androids Dream of Electric Sheep?

Proxies

- Word frequencies may serve as a proxy to **things we care about** in texts
- Word frequencies are the result of word choice -> word choice is a result of of **forces that organize texts**

Proxies

- Imagine LOTS of colorful dice of different shapes
- 1 text = 1 complex multidimensional die

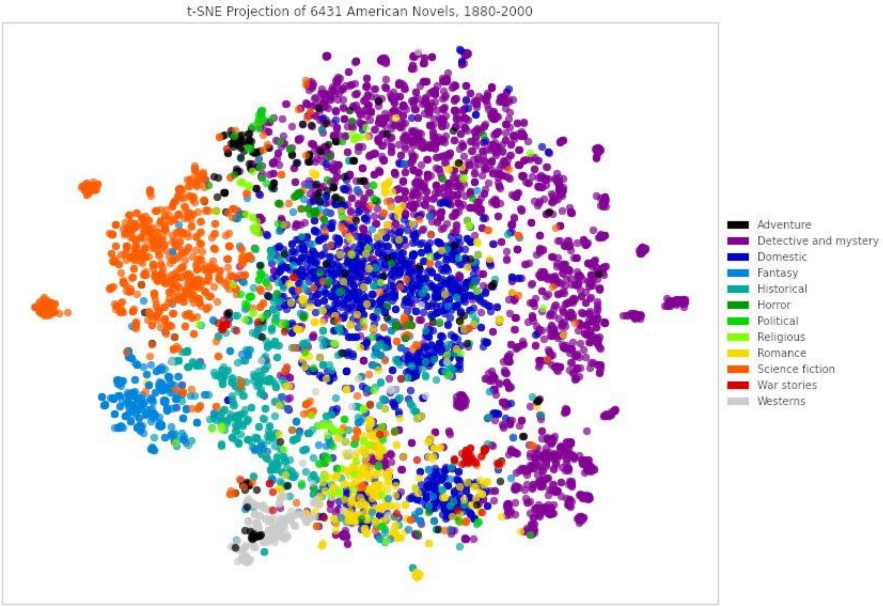
Let's toss them into some abstract space!

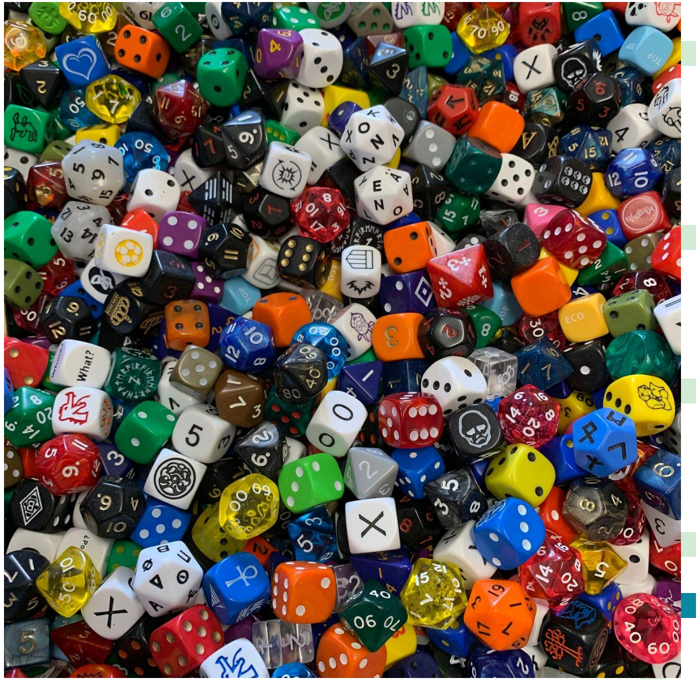
Proxies

Normal fair dice:
roll randomly all over the place, nice!

Text-dice: **OH NO…**

Text-dice rolling will be organized by multitude of forces



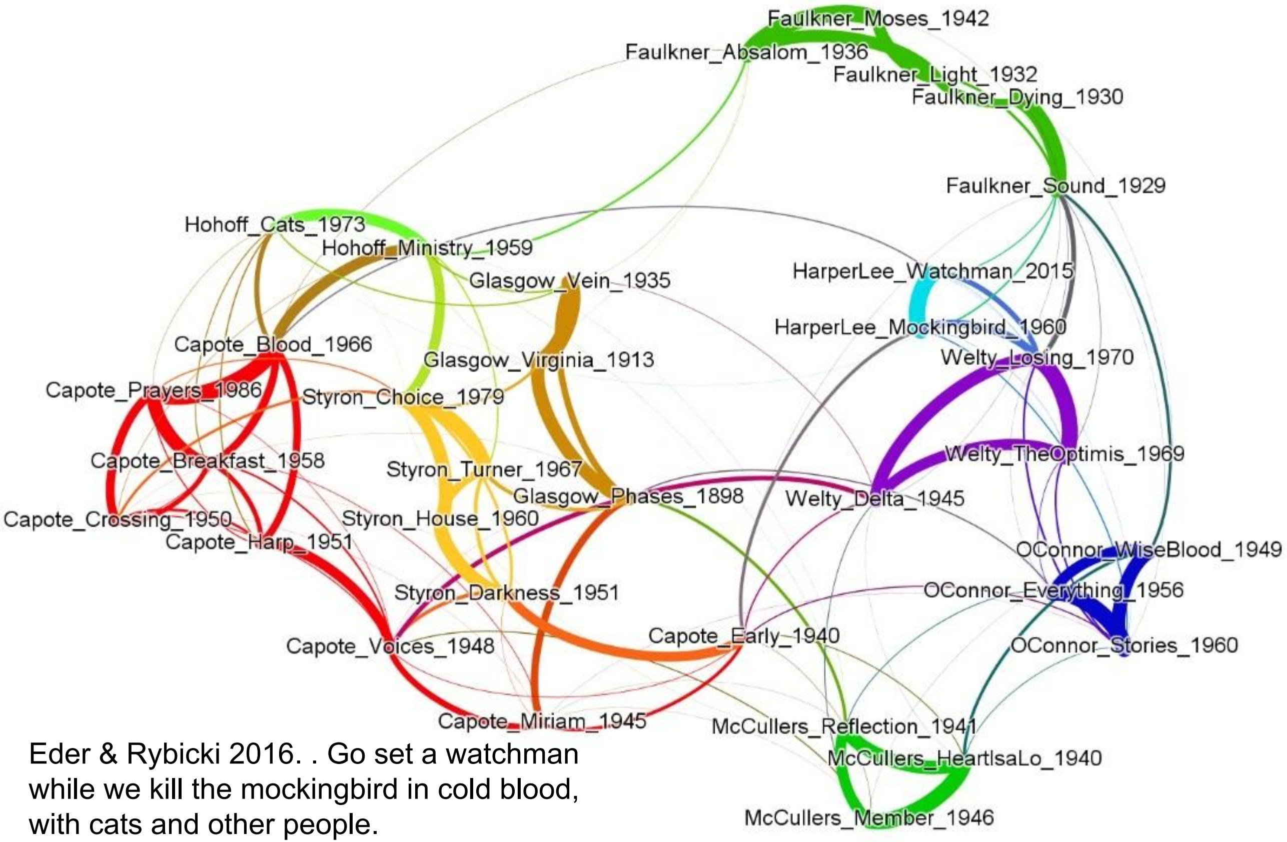


Curse & blessing of word frequencies

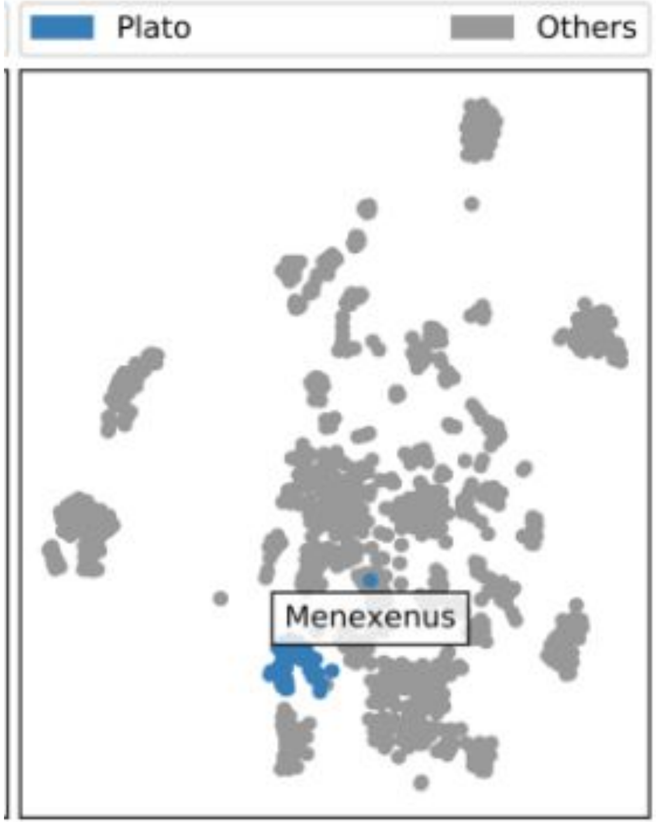
Forces that are “naturally” captured by word frequencies in a large corpus:

- **1. Authorship.** Two texts of the same author usually appear the closest to each other than to any outsider text
- **2. Modes of writing.** Fiction and nonfiction grew apart stylistically; Poetry books (esp. regularized verse) will always form a VERY exclusive party with themselves.
- **3. Genre.** Well-formed fiction genres (detective/mystery, sci-fi,
- **4. Chronology.** Global language change: Each generation of writers adopt slightly different version of language than the previous one (also: spelling conventions)
- **5. Gender.** Be aware of a historical gap between women and men writing (socially constructed)
- ...

Authorship



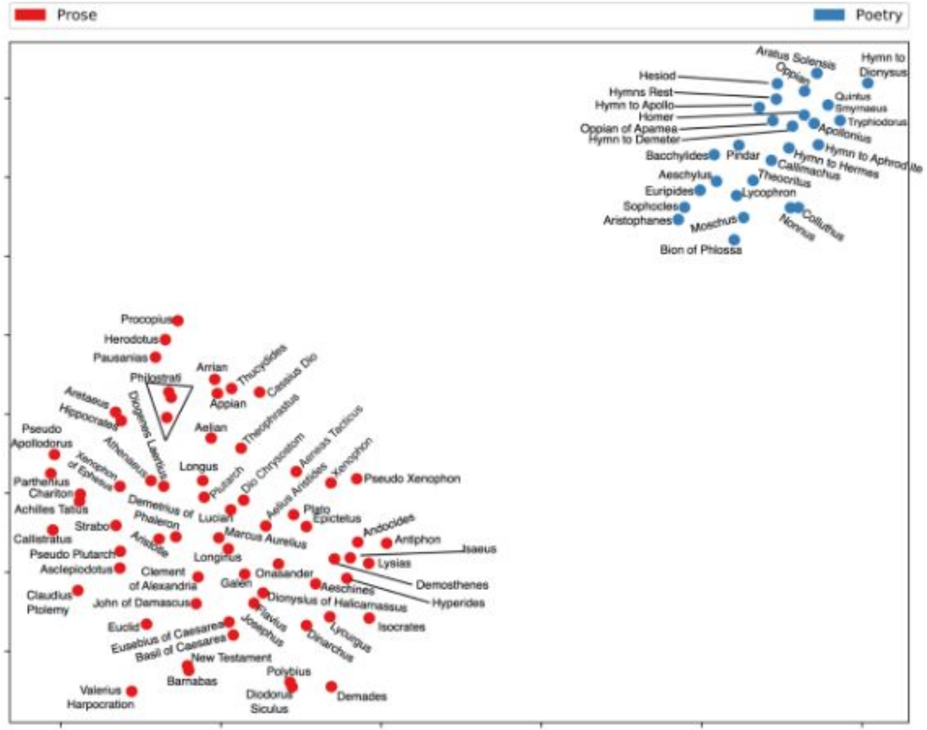
Eder & Rybicki 2016. . Go set a watchman while we kill the mockingbird in cold blood, with cats and other people.



- Plato’s texts with *Menexenus* as an outlier

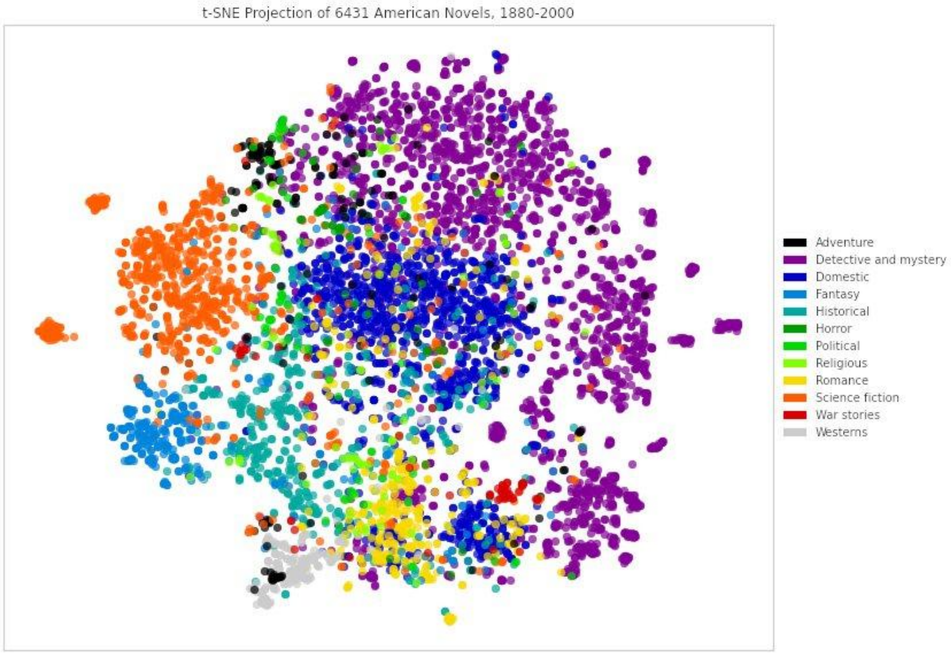
Storey & Mimno 2020: Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

Poetry parties with poetry

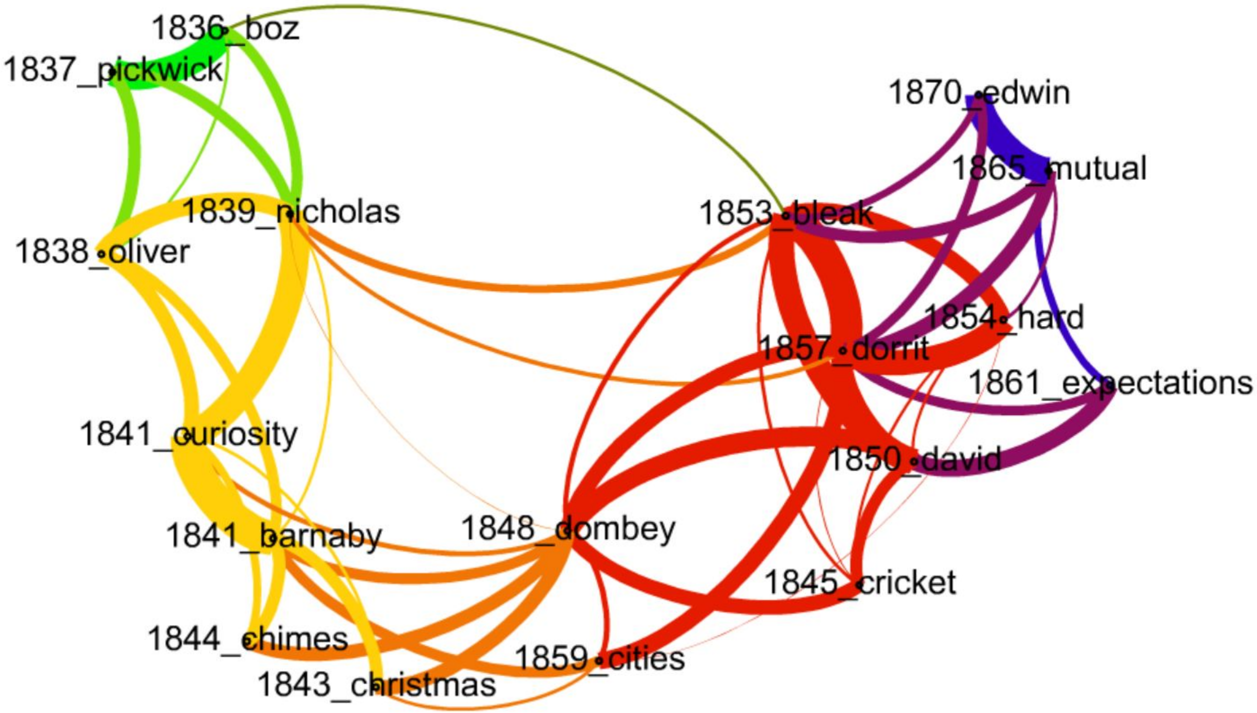


Storey & Mimno 2020: Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

Genres of fiction

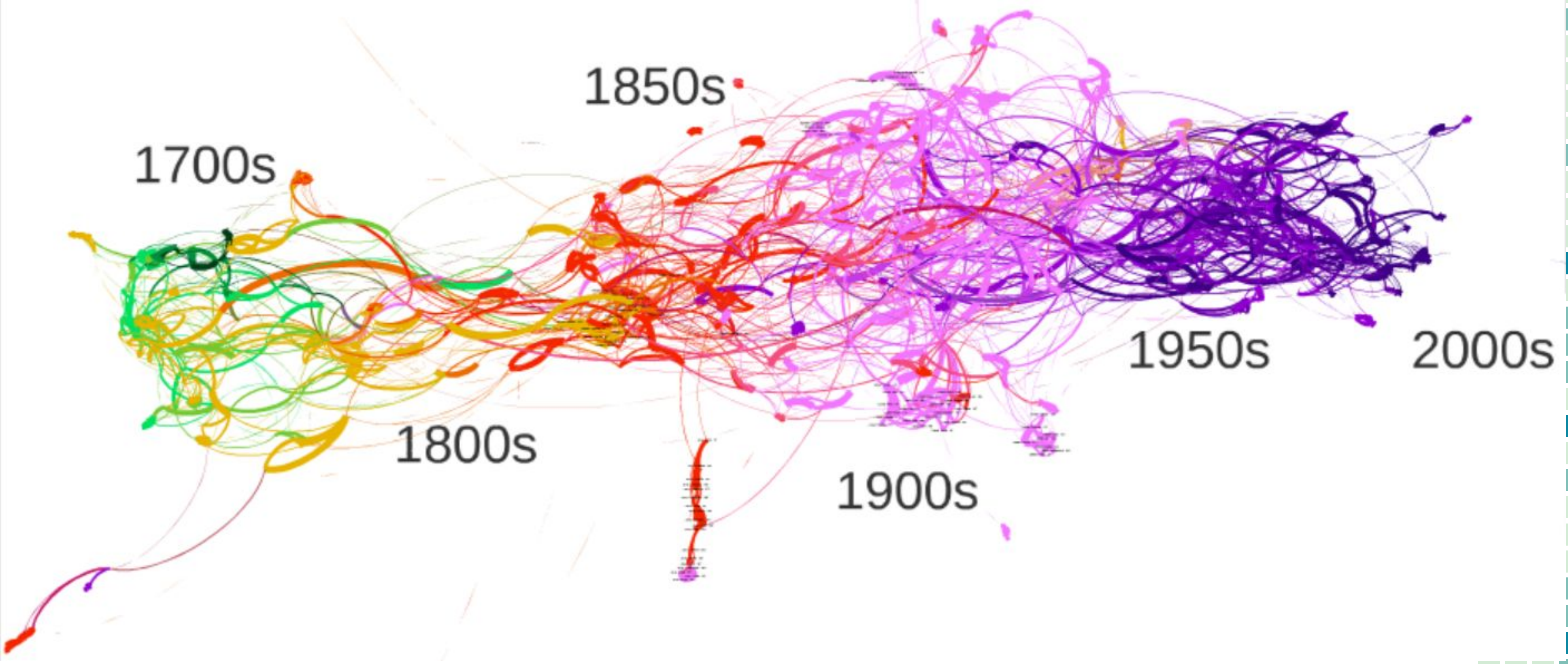


Jordan Pruett, *Twitter,* 2020
@pruett_jordan



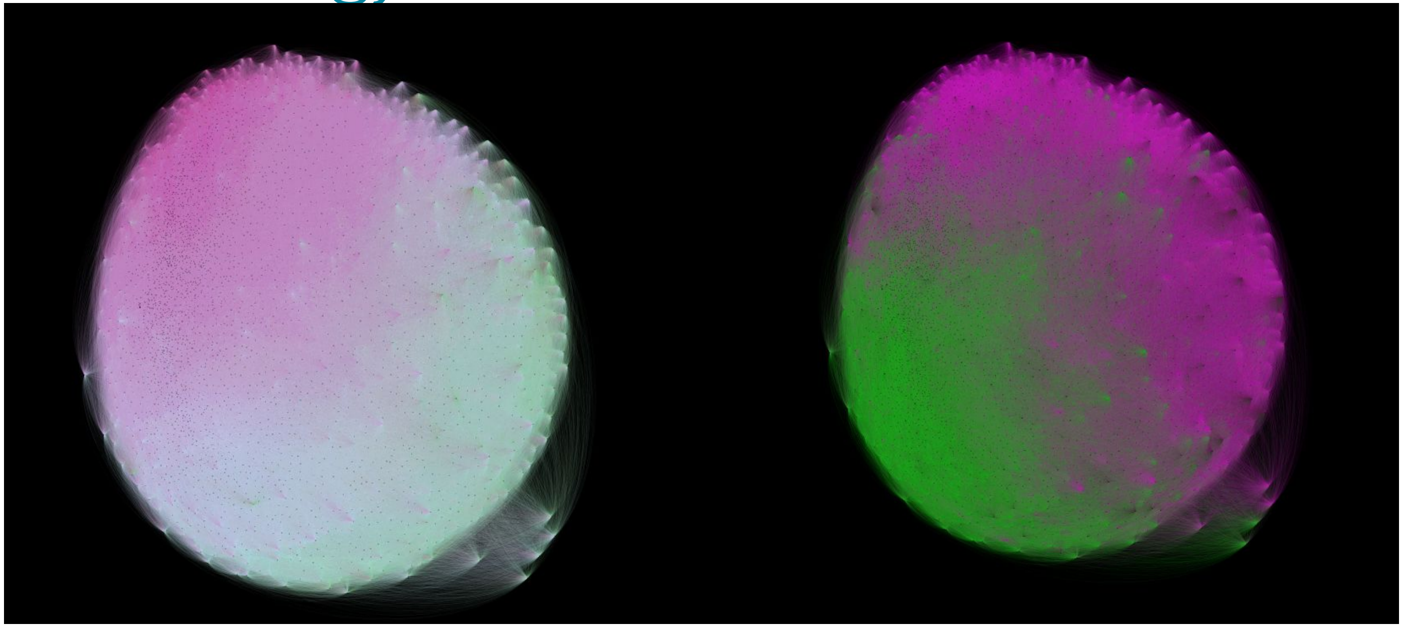
Rybicky 2016

Chronology: Global



Rybicky 2016

Chronology: Global

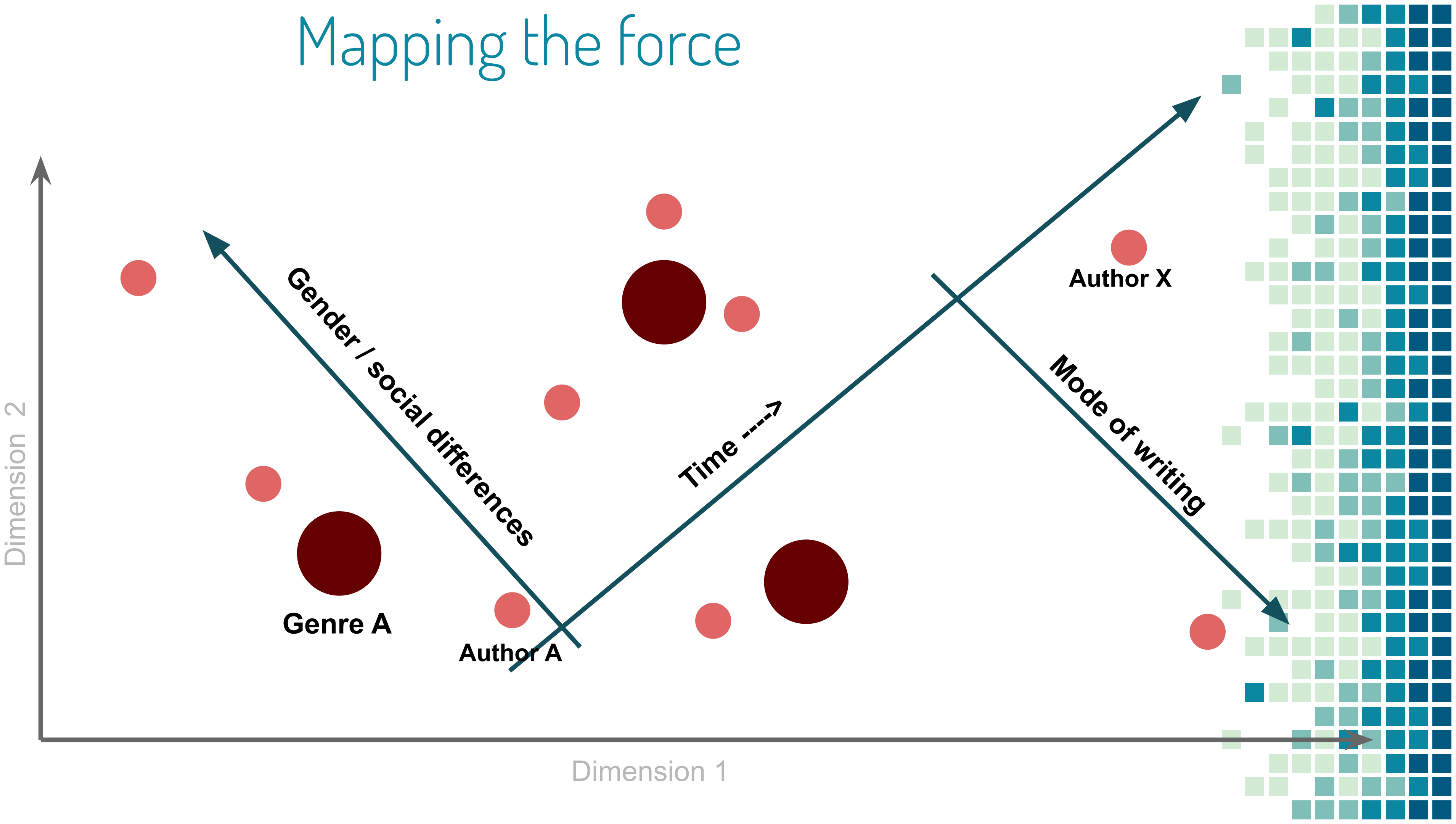


Time

Gender

Jockers 2013 *Macroanalysis*

Mapping the force



Thank you!