Data Is a Sandwich
Lessons from the Computational Literary Field

Andrew Goldstone
Rutgers University, New Brunswick

University of Kansas
September 19, 2023

ABOUT HATHITRUST →

Preserving knowledge.
Empowering possibilities.

18 million and counting. At HathiTrust, we are stewards of the largest digitized collection of knowledge allowable by copyright law. Why? To empower scholarly research, create transparency, and inspire curiosity.

hathitrust.org.

living large

A change in how we look at *all* of literary history: canonical and non-canonical: together…. And that's really my hope, as I have said: to come up with a new sense of the literary field as a whole….

A larger literary history requires other skills: sampling; statistics; work with series, titles, concordances, incipits—and perhaps the "trees" that I discuss in this article.

Franco Moretti, "The Slaughterhouse of Literature," *Modern Language Quarterly* 61, no. 1 (March 2000): 208.

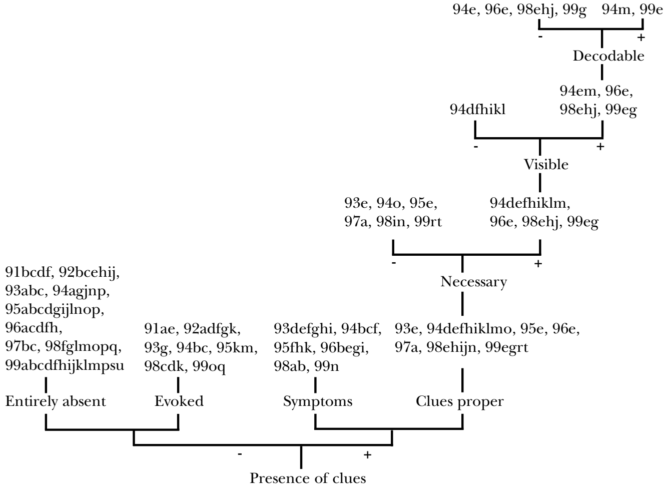
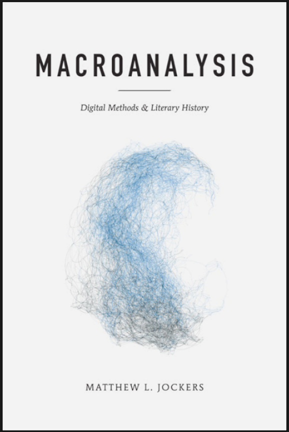

94e, 96e, 98ehj, 99g
94m, 99e
-
+
Decodable
94dfhikl
94em, 96e, 98ehj, 99eg
-
+
Visible
93e, 94o, 95e, 97a, 98in, 99rt
94defhiklm, 96e, 98ehj, 99eg
-
+
Necessary
91bcdf, 92bcehij, 93abc, 94agjnp, 95abcdgijlnop, 96acdfh, 97bc, 98fglmopq, 99abcdfhijklmpsu
91ae, 92adfgk, 93g, 94bc, 95km, 98cdk, 99oq
93defghi, 94bcf, 95fhk, 96begi, 98ab, 99n
93e, 94defhiklmo, 95e, 96e, 97a, 98ehijn, 99egrt
Entirely absent
Evoked
Symptoms
Clues proper
-
+
Presence of clues

Figure 2 Clues in the Strand magazine, 1891–99

Moretti, 219, figure 2.

The initial sample included the twelve *Adventures of Sherlock Holmes*, written in 1891 and 1892, and seven stories drawn from *The Rivals of Sherlock Holmes*, *Further Rivals of Sherlock Holmes*, and *Cosmopolitan Crimes*, all edited by Hugh Greene between 1970 and 1974.

Ibid., 214n8.



Through the study and processing of large amounts of literary data, the method calls our attention to general trends and missed patterns that we must explore in detail and account for with new theories.

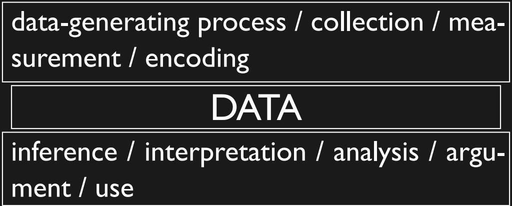
Matthew L. Jockers, *Macroanalysis: Digital Methods & Literary History* (Urbana: University of Illinois Press, 2013), 29.

We began with canon and archive as our objects of study, and with redundancy and type-token ratio as the means to investigate them; but then, the relationship between means and ends silently reversed itself: canon and archive moved to the periphery of our discussions, while redundancy and type-token ratio were increasingly occupying their center.

Mark Algee-Hewitt et al., “Canon/Archive. Large Scale Dynamics in the Literary Field” (Stanford Literary Lab, 2016), 12.

N. Saum, “Bánh mì thịt nướng,” Wikimedia Commons.

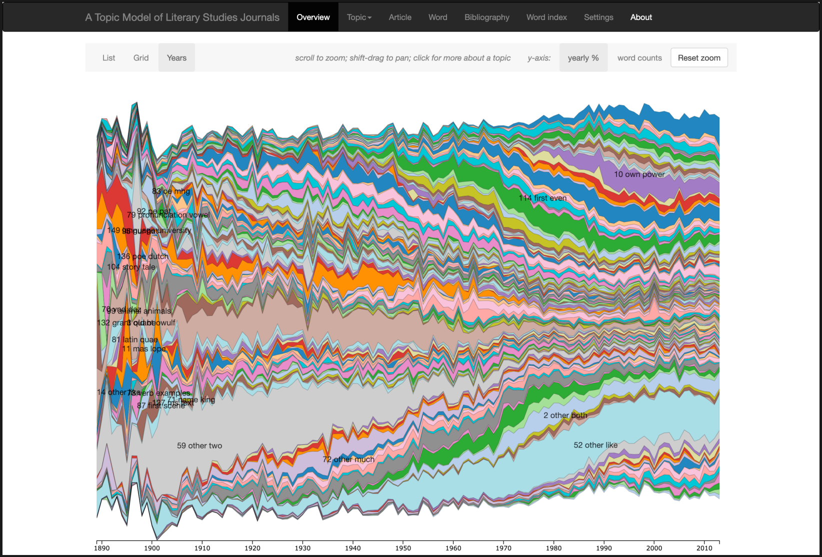
…data is a *metaphorical* sandwich



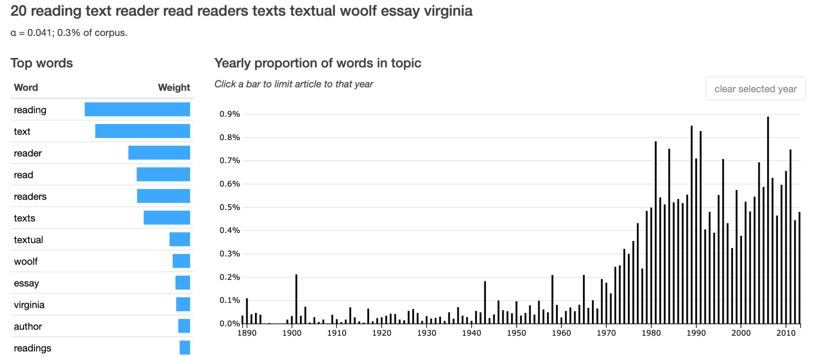
data-generating process / collection / measurement / encoding

DATA

inference / interpretation / analysis / argument / use



“Quiet Transformations of Literary Studies,” https://www.sas.rutgers.edu/virtual/ag978/quiet/#/model/yearly.



20 reading text reader read readers texts textual woolf essay virginia

α = 0.041; 0.3% of corpus.

Top words

Word	Weight
reading	
text	
reader	
read	
readers	
texts	
textual	
woolf	
essay	
virginia	
author	
readings	

Yearly proportion of words in topic

Click a bar to limit article to that year

Yearly proportion of corpus estimated to be drawn from topic *reading text reader read*, https://www.sas.rutgers.edu/virtual/ag978/quiet/#/topic/20.

author	*JML* subjects	author	*M/M* subjects
Joyce, James	8%	Eliot, T. S.	5.4%
Pound, Ezra	4.9%	Stein, Gertrude	3.8%
Yeats, William B.	3.9%	Joyce, James	3.5%
Conrad, Joseph	3.3%	Beckett, Samuel	3.4%
Beckett, Samuel	3.1%	Lewis, Wyndham	2.5%
Eliot, T. S.	2.9%	Woolf, Virginia	2.5%
Hemingway, Ernest	2.8%	Marinetti, Filippo	2.1%
Woolf, Virginia	2.8%	Pound, Ezra	2%
Kafka, Franz	2.6%	Kafka, Franz	1.4%
Lawrence, D.H.	2.3%	Kenner, Hugh	1.1%
Williams, W.C.	2.3%	Yeats, William B.	1.1%

MLAIB subject author headings from the *Journal of Modern Literature*, 1970–1990, and *Modernism/Modernity*, 1994–2014. Details: osf.io/frcys.

a big sandwich is still a sandwich

The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over.

Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014 (January 14, 2011): 176.

But, whenever you see something like this you should ask: is that all that data really doing anything? Could they have done the same research if the data could reach to the Moon and back only once? What if the data could only reach to the top of Mount Everest or the top of the Eiffel Tower?

Matthew Sagalnik, *Bit by Bit: Social Research in the Digital Age*, 2.3.1, bitbybitbook.com.

(crickets)

> Because what the hell do we care? We make plenty of money elsewhere. Do we really need this as a business? And if it's this difficult, at some point you say "screw it."

"Google insider" interviewed in 2016, qtd. in John B. Thompson, *Book Wars: The Digital Revolution in Publishing* (Cambridge: Polity, 2021), 139.

(crickets)

> Because what the hell do we care? We make plenty of money elsewhere. Do we really need this as a business? And if it's this difficult, at some point you say "screw it."

"Google insider" interviewed in 2016, qtd. in John B. Thompson, *Book Wars: The Digital Revolution in Publishing* (Cambridge: Polity, 2021), 139.

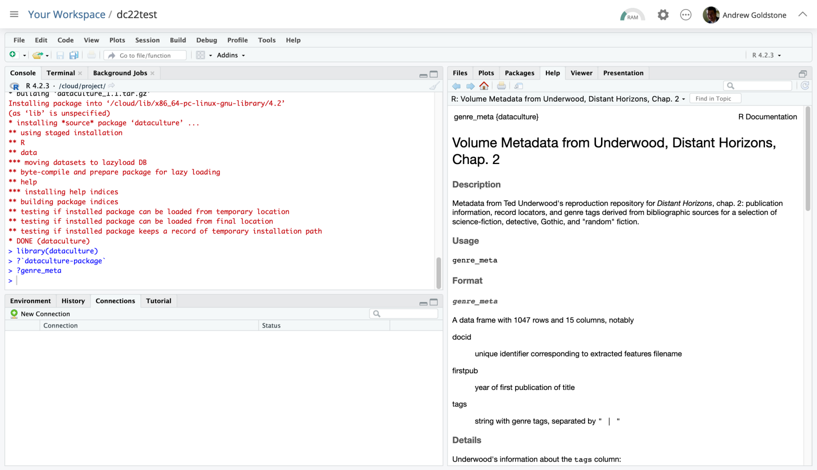
> In the beginning, there was Google Books…Fast forward to today: After more than a decade of evolution…

www.google.com/googlebooks/about/history.html

works cited

https://github.com/agoldst/dataculture

- Folgert Karsdorp, Mike Kestemont, and Allen Riddell, *Humanities Data Analysis: Case Studies with Python*.
- Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (University of Chicago Press, 2021).


Your Workspace / dc22test
RAM
Andrew Goldstone
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.3
Console Terminal Background Jobs
R 4.2.3 · /cloud/project/
Files Plots Packages Help Viewer Presentation
R: Volume Metadata from Underwood, Distant Horizons, Chap. 2
Find in Topic
Environment History Connections Tutorial
New Connection
Connection
Status

```
* building ‘dataculture_1.1.tar.gz’
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.2’
(as ‘lib’ is unspecified)
* installing *source* package ‘dataculture’ ...
** using staged installation
** R
** data
*** moving datasets to lazyload DB
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (dataculture)
> library(dataculture)
> ?`dataculture-package`
> ?genre_meta
>
```

genre_meta {dataculture} R Documentation

Volume Metadata from Underwood, Distant Horizons, Chap. 2

Description

Metadata from Ted Underwood's reproduction repository for *Distant Horizons*, chap. 2: publication information, record locators, and genre tags derived from bibliographic sources for a selection of science-fiction, detective, Gothic, and "random" fiction.

Usage

`genre_meta`

Format

`genre_meta`

A data frame with 1047 rows and 15 columns, notably

docid

unique identifier corresponding to extracted features filename

firstpub

year of first publication of title

tags

string with genre tags, separated by " | "

Details

Underwood's information about the `tags` column:

A session in RStudio on posit.cloud.

shave no yaks

```r
install.packages("tidyverse")
install.packages("remotes")
remotes::install_github("agoldst/dataculture")

library(tidyverse)
library(dataculture)
```

shave no yaks

Detective fiction and science fiction display a textual coherence…and hey sustain it over very long periods (160 or perhaps 200 years)….Instead of being more volatile than communities of reception, textual patterns turn out to be, if anything, more durable.

Underwood, 40.

```r
library(knitr) # for kable
genre_meta |> filter(recordid == "8886538") |>
    select(author, title, tags) |>
    kable()
```

author	title	tags
Wells, H. G.	The first men in the moon	teamblack | anatscifi | locscifi

Example metadata from Underwood's *Distant Horizons*, chap. 2 replication data.

```r
genre_meta |>
    filter(str_detect(tags, "scifi")) |>
    count(gender) |>
    kable()
```

gender	n
f	28
m	182
NA	4

Recorded genders of authors in Underwood’s SF corpus.

```r
genre_meta |>
    filter(str_detect(tags, "scifi")) |>
    mutate(tags=str_split(tags, " \\| ")) |>
    unnest(tags) |>
    filter(str_detect(tags, "scifi")) |>
    count(tags, gender) |>
    pivot_wider(names_from=gender, values_from=n) |>
    kable()
```

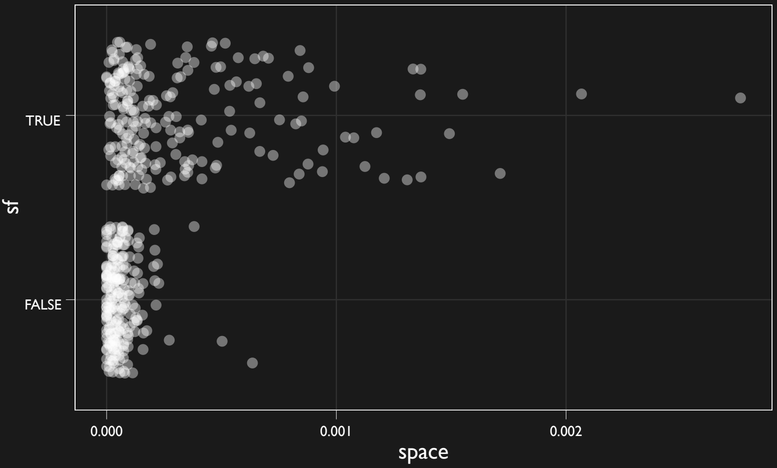
tags	f	m	NA
anatscifi	1	35	NA
chiscifi	18	137	4
femscifi	9	NA	NA
locscifi	2	19	NA

Recorded genders of authors in Underwood’s SF collections, by tag.

Works randomly selected from HathiTrust Digital Library, using fiction metadata developed in the NEH-funded project "Understanding Genre in a Collection of a Million Volumes." "Random selection" here means that the volumes were selected randomly but then approved or rejected by the author, to avoid stray volumes of nonfiction, classical poetry, juvenile works, etc.

Underwood, github.com/tedunderwood/horizon, qtd. in `help(dataculture::genre_meta)`.

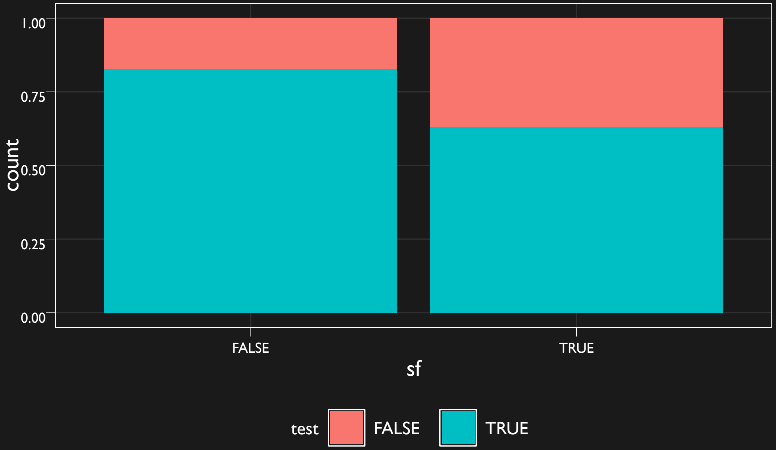
```r
n_sf <- genre_meta |>
    filter(str_detect(tags, "scifi")) |>
    nrow()
sf_set <- genre_meta |>
    select(docid, author, title, firstpub, tags) |>
    mutate(sf=str_detect(tags, "scifi"),
           random=str_detect(tags, "random") & !sf) |>
    filter(sf | random) |>
    group_by(sf) |>
    # randomly choose n_sf random volumes
    # (and n_sf SF volumes, but that's all of them)
    slice_sample(n=n_sf) |>
    ungroup()
space_test <- sf_set |>
    mutate(space=genre_features[docid, "space"])

ggplot(space_test) +
    geom_point(aes(x=space, y=sf), alpha=0.4,
               position="jitter", color="white")
```


TRUE
FALSE
sf
0.000
0.001
0.002
space

good enough for Federation work

```
space_test |>
    mutate(test=(space > 0.0001) == sf) |>
    ggplot(aes(x=sf, fill=test)) +
        geom_bar(position="fill")
```

good enough for Federation work



Classifying SF using the frequency of the word space with a cutoff at 0.0001 achieves an accuracy of 72.9%.

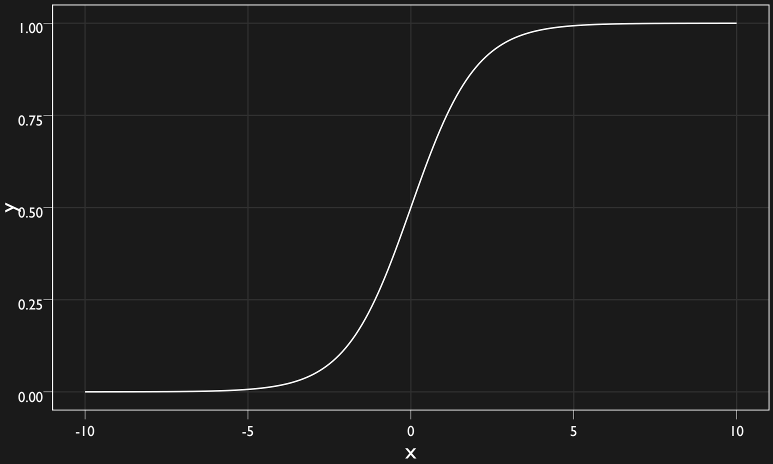
logistic regression in one slide

- convert texts to feature vectors $x_1, x_2, x_3, ...$
  - $x_1 =$ frequency of ”the” in the text
  - $x_2 =$ frequency of ”star” in the text
  - $x_3 =$ frequency of ”child” in the text
  - …about 4100 features (rare words ignored)
- for each text, record $y = 1$ if SF, $y = 0$ otherwise
- pretend every case is a (biased) coin flip
- bias of the coin assumed to depend systematically on x_i as:

$$P(y = 1|x_i) = \frac{1}{1 + \exp\left(-\left(b_0 + \sum_i b_i x_i\right)\right)}$$

- find best fit b_i using training data (“best”…)
- now you have an SF-detector: for any text $x_1, x_2, ...$
  - if $\hat{P}(x_1, x_2, ...) \geq 0.5$, guess it’s SF

```r
data_frame(x=seq(-10, 10, 0.1)) |>
    mutate(y=1 / (1 + exp(-x))) |>
    ggplot(aes(x, y)) + geom_line(color="white")
```


1.00
0.75
0.50
0.25
0.00
y
-10
-5
0
5
10
x

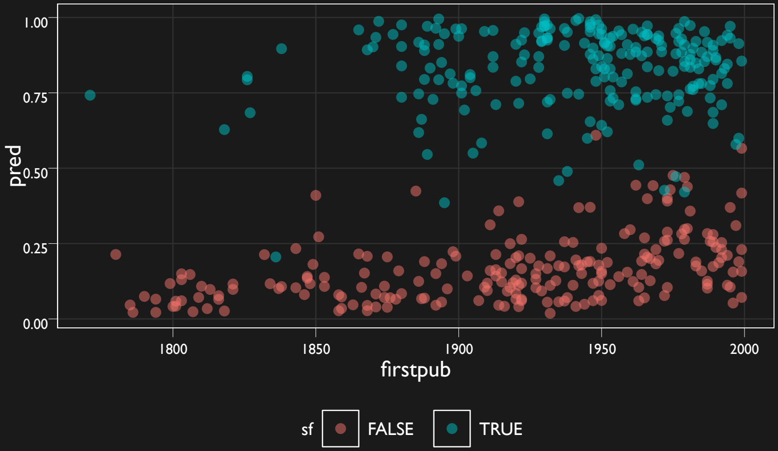
retrodicting SF

```
library(Matrix)
library(glmnet)
```

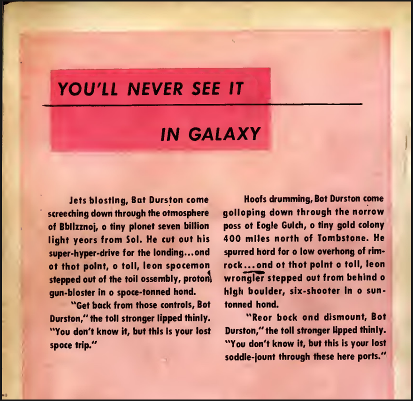
retrodicting SF

```
m4 <- cv.glmnet(x=genre_features[sf_set$docid, ], y=sf_set$sf,
                family="binomial", alpha=0, standardize=T,
                nfolds=10, type.measure="class")

sf_set |> mutate(
    pred=predict(m4, newx=genre_features[sf_set$docid, ],
                 type="response", s="lambda.min")) |>
    ggplot(aes(firstpub, pred, color=sf)) +
        geom_point(alpha=0.5)
```

retrodicting SF



L_2-regularized logistic regression used to predict “probability” of each volume being SF. The 10-fold CV estimate of classification accuracy is 0.91 (estimated s.d. 0.01; due to correlations between texts by the same author this is an underestimate).



YOU'LL NEVER SEE IT IN GALAXY

Jets blostlng, Bat Durston come screeching down through the otmosphere of Bbllzznoj, o tiny plonet seven billion light yeors from Sol. He cut out his super-hyper-drive for the londing...ond ot thot polnt, o toll, leon spocemon stepped out of the toil ossembly, proton gun-bloster in o spoce-tonned hond.

"Get back from those controls, Bot Durston," the toll stronger lipped thinly. "You don't know it, but thls ls your lost spoce trip."

Hoofs drumming, Bot Durston come golloping down through the norrow poss ot Eogle Gulch, o tiny gold colony 400 mlles north of Tombstone. He spurred hord for o low overhong of rimrock...ond ot thot polnt o toll, leon wrongler stepped out from behind o hlgh boulder, six-shooter ln o suntonned hond.

"Reor bock ond dismount, Bot Durston," the toll stronger lipped thinly. "You don't know it, but this is your lost soddle-jount through these here ports."

Galaxy 1, no. 1 (October 1950): back cover, Internet Archive.

sandwich preferences

- small samples with interesting variation
  - not “we scanned everything, good luck to you”
- rich metadata/detailed sourcing
  - footnotes, codebooks, originals…
- human-assigned high-level categories
  - if year of publication and author gender are your only covariates, go back
- meaningful arguments in view from the start
  - and a reflexive attention to the possibility of goal displacements through to the end