## Artificial Intelligence as an Archival Science

From Manuscript Collation with OCR Models to Cultural Circulation in Newspapers

David A. Smith<br>https://khoury.neu.edu/home/dasmith/

## The Uncanny Valla



Lorenzo Valla (Imagines philologorum) UMass Gallery of Philologists


Donation of Constantine (Santi Quattro Coronati, Rome)

# JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 

## INFERENCE IN AN AUTHORSHIP PROBLEM ${ }^{2,2}$

A comparative study of discrimination methods applied to the authorship of the disputed Federalist papers

Frederick Mostrller
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
David L. Wallace
University of Chicago
This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of The Federalist papers; and to propose routine methods for solving other authorahip problems.

Word counts are the variables used for discrimination, Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as $a n$, of, and upon, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like war, ezecutive, and legislature do not.


Lorenzo Valla (Imagines philologorum) UMass Gallery of Philologists




Erasmus De duplici copia verborum ac rerum... Basel, 1521.


Desiderius Erasmus (Hans Holbein, Nat'l Gallery, London)

Goethe: Marianne, have you ever tried translating Persian poetry?

Marianne von Willemer: Yes, Goethe, I have. It can be quite challenging to capture the beauty and depth of the original poetry in another language.

Goethe: That is true. The rhythms and word choice in Persian poetry are quite different from those in German.

Marianne von Willemer: Yes, and the concepts and themes often require careful consideration to translate accurately and beautifully.

Goethe: One approach I have found useful is to first understand the meaning and intent of the original poem, and then try to find German words and phrases that convey the same meaning and emotion.

Marianne von Willemer: That is a good approach. It is also important to consider the cultural context and historical background of the poem, as that can often provide valuable insight into the intended meaning and impact.

Goethe: Absolutely. And when in doubt, it can be helpful to consult with others who are more familiar with Persian literature an $\approx$ Try again

## Question: What is the name of the Russian mercenary group?

Answer: Wagner group.

## Question: Where is the Wagner group?

Answer: In Syria.
Question: Who is the leader of the Wagner group?
Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.


## Abundance

 and Loss

此


February 2012



Oral history with 83 year old female, Colorado (1978) starting at 1:53
Luton, Ruby (Speaker) I Kyle, Evelyn (Interviewer) I Nelson, Justina (Interviewer)


## The Roots of Big Data

- Big Government
- Big Business
- Big Science
- Social change: Living online
- Digitizing the past


Metropolitan Museum of Art
all Tunisian security forces were also placed on immediate alert. President Bourguiba subsequently informed a news correspondent that Tunisian police controls were particularly rigid in the area of the French base at Bizerte. Bourguiba is reported to fear that French units stationed either in Bizerte or in eastern Algeria may be encouraged to enter Tunisian territory. Publicly, however,

## SECRET $\square$

25 Apr 61 CENTRAL INTELLIGENCE BULLETIN Page 2

Digitized US newspapers


Digitized US newspapers


## Single-sheet US newspapers



## L. Putnam

## The transnational and

 the Text SearchableL. Klein

The Image of Absence: Archival Silence, Data
Visualization, and
James Hemings
S. Hartman

Venus in Two Acts

At the same time, the new topography of information has systematic blind spots. It opens shortcuts that enable ignorance as well as knowledge. Digital search offers release from place-based research practices that have been central to our discipline's epistemology and ethics alike.

But in spite of what Ed Folsom $(2007,1571)$ has extolled as the "epic transformation" of the archive, characterized not only by increased access to content, but also by the proliferation of paths through that content that are facilitated by the digital archive's underlying database structure, the issue of archival silence-or gaps in the archival record-remains difficult to address.

This double gesture can be described as straining against the limits of the archive to write a cultural history of the captive, and, at the same time, enacting the impossibility of representing the lives of the captives precisely through the process of narration.

The method guiding this writing practice is best described as critical fabulation.


# Modeling Human <br> Information <br> Processing 

## Machine Learning as an Archival Science



## This Talk



## Automatic Collation for Diversifying Corpora

## $\mathfrak{A l i d} d$ leburp

天n this ¥aper are publishey the ¥ublit orders, zesolutions, zaiws, wublit
prietor.
MIDDLEBURY VT
leasant and agreable employmen: for oung people at a time when little else can e done.-Albany Cultivator.

TESTIMONY AGAINST RADICAL. ISM.
In Document 109, of the House of Repesentatives, furnished us by Mr. Burke, ast printed, a Report of Mr. Ellsworth, Jommissioner of the Patent Office, we have latistics and observations, on the Agricularal and Manufacturing interests. We opy from this document the following im. ortant testimony :
ortant tes
Let us listen to the wisdom of those those opinions are recorded for our encour-gement-more especially since such opinons come from individuals who do not anction protection, except such as incidenally arises from the raising of a revenue:"To be independent for the comforts of fe, we must fabricate them ourselves. We ust now place the the manufacturer by the ide of the agriculturist. The grand inqui-

## MISCELLANEOUS.

THE RESTING PLACE.
-

EY CHARLES MACKAY.
Tell me, ye winged winds,
That round my path way roar,
Do ye not know some spot.
Where mortals weep no more ?
Some lone and pleasant dell, Some valley in the west,
Where free from toil and pain,
The weary soul may rest ?
The lond wind dwind led to a whisper low, And sigh'd for pity as it answered "No!"

Tell me thou mighty deep,
Where billows round may play
Know'st thou some favored spot,
Some island far away,
Where weary man may find
The bliss for which he sighs,
' Remember me to Lady ber me to Horatio! ren friends! Doctor, remem Tell him, I have made Hamilton and Horatio t repeated his remembran ton and Horatio, and to he said several times. G less agitated, and, at la ask questions about wh This led his mind to whom he sent and inqu iety, exclaiming aloud $h$ he was alive unless he s agitated at the captai mented his being unabl do what was to be done assurance given of the $c$ quarter-deck. At last, and he instantly grew 1 tened to his report abs fleet, directed him to an he should die, but obse half an hour longer.

Middlebury People's Press, 22 March 1843

THE RESTING PLACE.
XV CHAELES MACKAT.
Tell me, je winged windi, That round my palhay roar, Do ye not know some spoU Where mortals treep no more 1
Some lone and pleasant dell,
Some valley in the west,
Where free from toil and pain,
The weary soul maj rest 1
The lond wind dwindled to a nhisper low,
And tigh'd for pitir as it answcred "No!"
Tell me thou mighty deep,
Where billows roand may play
Know'st tlion some farored spot,
Some tland far away,
Whero weary man may 6nd
The bliis for nhich he sigbi,

THE RESTING PLACE.
gy charles mackay.
Tell me, ye winged winds, That round my path way roar,
Do ye not know some spot.
Where mortals weep no more ?
Some lone and pleasant dell, Some valley in the west,
Where free from toil and pain,
The weary soul may rest?
The lond wind dwind led to a whisper low, And sigh'd for pity as it answered "No!"

## THE INQUIRE.

Tell me, ye winged winds,
That round my pathway roar,
Do ye not know some spot Where mortals weep no more? ${ }^{\text {- }}$ Some lone and pleasant dell, Some valley in the west, Where, free from toil and pain, The weary soul may rest ? The loud wind dwindled to a whisper low, And sighed for pity as it whispred "No!"

## THE INQUIRY.

Tell me ye winding winds, That round my pathway r Do ye not know some spot Where mortals weep no me Some lone and pleasant dell, Some valley in the west,
Where, free from toil and pa
The weary soul may resi?
The loud wind dwindled to a whisp And sighed for pity, as it answered;

```
Settlements in America. 269
mines of Potofi, which winding amongt the openings of the Cordillera, difcharges itfelf at laft into the Paraguay; and this river is
```

\{
"begin": 74,
"text": "• openings of the Cordillera, difcharges itfelf\n", "witnesses": [
\{
"begin": 42, "id": "K023456.001_0278/manual",
"text": " openings of the Cordillera, difcharges itfelf"
\},
"begin": 72, "id": "cihm_32607_0294",
"text": "openings of the Cordillera, difcharges itfelf\n"
\},
"begin": 380, "id": "accountofeuropea01burk_0306",
"text": " openings of the\nCordillera, difcharges itfelf"
\},
"begin": 380, "id": "anaccounteuropea01burkiala_0302",
"text": " openings of the\nCordillera, difcharges itfelf"
\},
"begin": 381, "id": "cihm_32428_0314",
"text": " openings of the\nCordillera, difcharges itfelf" \},



\{

```
"begin": 349,
"text": "\sigma\tau\varepsilonı\lambda\varepsilon\nu \etȧ \varepsiloṅ\nu 'P\omegá\muпп ’Екı\imath\lambda\eta\sigma\varepsiloń\alpha lxa-\n",
"witnesses": [
    {
        "begin": 365, "id": "pg.Vol.-1.njp.32101077772331_ocr_036",
        "text": "\sigma\tau\varepsilon\\lambda\varepsilonv \grave{ 文 'Pú\mu\eta 'Екк\lambda\eta\sigmaí\alpha íк\alpha-\n"}
    },
        "begin": 368, "id": "pg.Vol.-1.nnc1.0022106693_ocr_024",
        "text": "\sigma\tau\varepsilon\\lambda\varepsilonv \etȧ \varepsiloǹv 'P\omegá\mun 'Екк\lambda\eta\sigmaí\alpha iк\alpha-\n"
    },
        "begin": 369, "id": "pg.Vol.-1.mdp.39015014764065_ocr_026",
        "text": "\sigma\tau\varepsilon\\lambda\varepsilon\nu \grave{ \varepsiloǹ\nu 'P\omegá\muпn 'Eкк\lambda\eta\sigmaí\alpha íк\alpha-\n"}
    },
        "begin": 370, "id": "pg.Vol.-1.hvd.32044073501405_ocr_022",
        "text": "\sigma\tau\varepsilonı\lambda\varepsilonv \etȧहेv 'Pú\mu\eta\ 'Eкк\lambda\etaoí\alpha iк\alpha-\n"
    },
        "begin": 378, "id": "pg.Vol.-1.hvd.32044054121090_ocr_028",
        "text": "\sigma\tauعı\lambda\varepsilon\nu \grave{ \̀\nu 'Pú\muпп 'Екıl\lambda\eta\sigma\xi\alpha lxa-\n"}
    }
```


## Duplicate Witnesses

| Corpus | Manually <br> transcribed | Manual + other |
| :--- | ---: | ---: |
|  | witnesses |  | Total witnesses


| Richmond |
| :--- |
| Dispatch |$\quad 2,184,000 \quad 949,000(43 \%) \quad 25,529,000$

EEBO/ECCO
TCP
E,579,000 5,480,000 (64\%) 529,000

## Characters



How many possibilities?

## Words



How many possibilities?

## Lines



How many possibilities?

## More Distant Supervision

- More context means...
- More naturally learning variant forms
- Learning context-dependent orthography
- Implicitly learning language models
- Spending less time on segmentation
- But with search algorithms taking longer


## Usual HTR Paradigm

Training and testing on different pages of the same manuscript



Keinen-Schoonbaert (2019). Results of the RASM2019 Competition on Recognition of Historical Arabic Scientific Manuscripts.


Hafiz, Divan
Bodleian Library MS Ouseley Add. 26


Hafiz, Divan
Philadelphia Free Lib. MS Lewis O 75

## Multiple Sequence Alignment

i find the states suffering from the effects of a`civil war resistance i lound the states buffering from the effects`of a civil war resistance
i found the states suffering from the effects`of a civil war resisauce \(t\) found tb states suffering from`th effects of a`ciru wsr iiesietanee i found the states suffering from the effects ofa civil war`resistance
i found the state suffering from the effects of`a civil war resistance i loutid the states sullering irani`the ef`fects of a civil war resistance

```
i f-in-d the -states suffering -from-`the -ef-fects -of -a-`civil w-ar -r-esi-stance
i loun-d the -states buffering -from-`the -ef-fects-`of -a -civil w-ar -r-esi-stance
i foun-d the -states suffering -from-`the -ef-fects-`of -a -civil w-ar -r-esi-s-auce
t foun-d t-b -states suffering -from-`th- -ef-fects -of -a-`ci-ru w-sr -iiesie-tanee
i foun-d the -states suffering -from-`the -ef-fects -of--a -civil w-ar-`r-esi-stance
i foun-d the -state- suffering -from-`the -ef-fects -of-`a -civil w-ar -r-esi-stance
i loutid the -states sullering -irani`the -ef`fects -of -a -civil w-ar -r-esi-stance
```




July 2, I 866


July 3, I 866


A linear stemma is an HMM s.t. each state is an LM



## Automatic Collation for Diversifying Corpora

Step \#1: User uploads manuscript images and digital edition of text

Step \#2: ACDC tool produces a very poor (dirty) OCR transcription of the manuscript image of Häfiz's Divān

Input \#1: Manuscript images of Heāfiz's Divān



- : used to indicate incorrectly transcribed character

Input \#2: Open-access digital texts of Haafiz's Divān



This line text and image are paired and become training data


This line text and image are not paired and are set aside

Step \#4: ACDC tool selects top line image-transcription pair for inclusion in training data because their degree of alignment meets determined threshold and rejects bottom line due to low alignment

Step \#3: ACDC tool aligns dirty OCR transcription of manuscript images of $\boldsymbol{H}$ äfiz's Divān produced in step \#2 with digital text (input \#2) from step \#1 by identifying matching patterns of characters in the two texts


Firūzābādī，al－Qāmūs al－muhițt

al－Jazūlī，Dalā＇il al－khayrāt
Staatsbibliothek zu Berlin，Glaser 133


## Automatic Collation



```
"id": "sbzb_glaser_33.pdf_page_42.xml",
"begin": 3148,
"text": "مش%\\\",
"witnesses": [
    {
        "ref": 0,
        "id": "leipzig_ms_gabelentz_60.pdf_page_49.xml",
    "begin": 1740,
    "alg1": "المترك\n",
    "alg2": "المشركين الذين لا صلح بيننا وبينهم\n",
    "matches": 79,
    "text": "المتركين لا صلح بينا وبنهم\n"
},
    "ref": 1,
    "id": "qamus_muhit.txt",
    "begin": 1023̄61,
    "alg1": "بينا وبينهم\المشركين ال\ين لا صلح ",
    "alg2": "المشركين الذين لا صلح بيننا وبينهم"\",
    "matches": 107,
    "text": "بينا وبينهم\\/\مشركين الذين لا صلح "
}
]
```


## Manuscript Families for Training

| author | title | language | MSS | transcribed lines |
| :---: | :---: | :---: | :---: | :---: |
| al-Jazūİ | Dalā'il al-khayrāt | Arabic | 28 | 1602 |
| Firūzābādi | al-Qāmūs al-muhiṭ | Arabic | 13 | 1254 |
| Heāfiz | Divān | Persian | 11 | 981 |
| Sa'di | Gulistān | Persian | 17 | 1865 |
| Taftāzānī | Sharh al-Aqā'id al-Nasafiya | Arabic | 12 | 1140 |
|  |  |  | 81 | 6842 |

## Test MSS: Disjoint from training

language
Arabic
Turkish
Persian
mixed
total




## Page Layout Analysis

| $\square$ body | $\square$ figure |  |
| :--- | :--- | :--- |
| $\square$ title | $\square$ caption |  |
| $\square$ pageNum | $\square$ note |  |
| $\square$ catchword | $\square$ | signature |



Cohn, Gustav: Die deutsche Frauenbewegung. Berlin, 1896.

| Bild: 0015 V | << vorherige Seite | nächste Seite >> | 1.0 |
| :---: | :---: | :---: | :---: |

## y) armort.

Diefe Edrift ift aut einem ©uffage ber „Deutiden Funb. [ idan*" bervorgegangen, ben idj als abgerumbeted શüdlein meiner Edjuefter gum Ceburtitage uiberreiden mollte. Vielleidft finbet et aufer ibr nodf einige anbere Sefer. Sm Gegeriage fu bem erten abbrud if barin mandeed neu georonet, emveitert, ers gängt, theils aut ber alteren unb neueren Qiteratur, theils ant ben Thatiaden ber Cegenmart, beren if zumal bon Gng* lant einigel nadd erneutem ©ิerienaufenthalte beimgebradjt habe.

Sd batte bie abfidit, ein altel englijded Motto vor. juicten . .

> They are slaves who dare not be In the right with two or threa* (Lowell, Stanzas on Freedom.)
aber ef [decint mir, bafg ef niidft mebr paft.
Inniverfitāt Mōttingen, Grbe §uni 1896.
©. ©oljn.
nächste Seite >>
田
<TEI>
<body>
<pb facs="\#f0015"/>
<div n="1">
<head>Vorwort.</head><lb/>
<milestone rendition="\#hr" unit="section"/>
<1b/>
<p>Diese Schrift ist aus einem Aufsatze der
\&\#x201E; Deutschen Rund-<1b/>
schau\&\#×201C; hervorgegangen, den ich als abgerundetes Büchlein meiner<lb/>
.
Vielleicht findet<lb/>
es außer ihr noch einige andere Leser. Jm Gegensatze zu
dem<1b/>
ersten Abdruck ist darin manches neu geordnet,
erweitert, er-<lb/>
gänzt, theils aus der älteren und neueren Literatur,
ganzt, theils
theils $<1 \mathrm{~b} />$
aus den Thatsachen der Gegenwart, deren ich zumal von Eng-<1b/>
land einiges nach erneutem Ferienaufenthalte heimgebracht habe. $</$ p><1b/>
<p>Jch hatte die Absicht, ein altes englisches Motto vor-<1b/>
zusetzen\&\#x2026;</p><1b/>
<cit>
<quote> <hi rendition="\#right">\&\#x201E; <hi
rendition="\#aq">They are slaves who dare not be<1b/>
In the right with two or three.</hi>\&\# x 201 C ; </hi>
</quote><lb/>
<bibl> <hi rendition="\#right">(<hi
rendition="\#aq">Lowel1, Stanzas on Freedom.</hi>)</hi> </bibl>

## </cit><1b/>

<p>aber es scheint mir, daß es mir nicht paßt. </p><1b/>
rendition="\#g">Göttingen</hi>, Ende Juni 1896.
</dateline><lb/>
<byline>
<hi rendition="\#right">G. Cohn.</hi>
</byline><lb/>

## Page to Align

## Mative val An Bien













 allue $=4$ Wanellomen nein，etit is he vione entr，etm olvite， m ＋iters
 an micnen betren niveter de


 4 thele suakept $m$ Nen $=$ Vahm then



 Konk in pid to kive




Molve melat lione $\quad$ me


 achat anty an








而浽
 nelate Mataneman ging satis ang
 Eanc







皆 1




Fine－grained FA


## Region Detection

- U-net DTA - U-net WWO - F-RCNN DTA - F-RCNN WWO



## Viral Texts

## 

B) Whe walton \& sons.
montpelier, thursday, january $20,1848$.
VOL. XLII, No. 10....-WHOLE NO. 2153.
twathmag state Journal.

Poctry.
thiméritan.






 , mionemion


 rowno








$\rightarrow 2$








## THE HUSKERS

## st d. at. Whitties.



 ar Mat






 Alt thering
Fravily the
And atelsit, 1 ,


Brcoen ibe the

shats



 $7 *$
 Numat mas mat,

 nel







## f)umorous.

## "That lant Julep!

A shony temperame itany.
Not masy years since, in ooe of our Missisippi fiver cities, the landloed of a poled hotrl, in spite of the good fere erved upon bis tahle toek an aversion to his food, and, coabsing himself so the faid exchisively, doank himaelf out of existence. He was a joily specimen of in poblican, ard had endesred Himself, in his latser days, to a oouple of young boarders, by frequesaly inviting them to share in what be called "/a comfortable carouse." These listle reunions, ns we beGore stated, were brought to a close by the prineipal roddenly dropping out of the set, sid being no lonter able to "shand up" to his share of the liguor.The deecened's wife being absent from The deecneers wife being absent froms the city, his friends comfortably pocked
him in lice for preservation until her rehims in ice for prescravios onfil
turs. The fectings of the hopefal youns consivialiats were sadly shocked at the idea of sojoursing is the same locuse wich the inamimate body of their regretied comparion. It was all well enoogh when he was able to stand the ligquor and driek his share of it, bet the thooght of his being peched away in such cold quarters made thees shiver with a strange appreheng to dissija
the room their oum loreed to example to have it verance.

## C19 newspapers were an all-purpose medium <br> perance. At lengsh the wife arrived. and the prullican was consigned to his

 long hotne; bat we regret to add that with his departure the desim for spother 'comfortable earouse" Legan to manifest$\qquad$ NEW AND DERIRABLE RAI.T, NAIL.S, GL.ABS, by she loesl or in smaller lots

CROCKIII, GLABS \& CHINA WARE, it city prices
900 prime Noth-West DUFPAL.O ROBES-cheap ensugh 5000 yds. rich PRINTB, GINGIIAMS \& FUILNITURES; 9000 yds, of benstiful styles of wersted

## roll wew PAPER HANCINGS,

 Montjelier, Nov. 25th, 1847.
## FOR SALE

 o- Fowhi. for makime Har ivon. 는


MORE GOODS ATTHEqBELPCANBNTOHE


## STORRS \& LANGDON,

ARE opening this week from Boston and New-Yook, and immense Stock of

## GOODS FOR THE WIN'TER THADE.

10 Hlses of very elseap SUGAR, with COPPEE, TEA \& SPICEs, in any OIL.s, PAINTA, FBH, FIUUIT, TOBACCO, SNUYP, \&e, at low prices; 40 cheats of that superior Y, H. TKA, at the same low prict-bod'т ponest IT I

Heasekeeping asd Mechanical HARD.WARE, of eew paterns

40 bales COTTMNS, TICKING, DRILLING, WICKING \& BATTING

CIOAK AND DRESS GOODS.
to With THIMMINGS TO MATCH EXACTLY. -D
The sbove and a matritade of other goods are offered at the lowest prices.
next,

Jrom aomesuc enemies bo our jree insubuuons.
Our object is briefly this: To perfect an orgamzation, modeled after that of the Constitution of the United States, and coextensive with the confederacy. Its object and principles, in all matters of nationsl concern, to be uniform and identical, whilst in all local matters, the component parts shall remain independent and sovereign within their respective limits.

The great result to be ob'ained-the only one which can secure a perfect guaranty as to our fu-ture-is Usios; permanent, enduring, fraternal Union! Allow me then to impress upon your minds and memories the touching sentiments of the Father of his country, in his farewell address : "The Uxiry of Government which constitutes you one people," says Washington, "is justly dear to you, for it is the main pillar in the edifice of your real independence, the support of your tranquility at home, of your peace abroad, your prosperity, even that liberty you so jusfly prize.

*     * It is of moment that you should properly estimate the immense value of your National Union, to your collective and individual happiness. You should cherish a cordial, habitual, and immo-

Cheay BE OF MUS

A COLLE: ZUREA

Beautifully III Designos, in I. subjecte:
No. 1. $\mathrm{Ma}_{\mathrm{A}}$
" 2. Bes
" 3. Det
" 4. Res
" 5. Mnı
" 6. Cos
Published Broadway, ( S tey The ab the country on
may29-1m

## Political Speeches



## News

## Sentimental Stories

## THE LAST FAREWELE.

## BOBERT EMMET AND HIS Love.

'Twas the evening off a lovely day-the last dny of the noble and ill-fated Emmet. A young ladystood at the gate and desired admittance in to the dungeon. She was closely vailed, and the keepar could not immagine who she was, nor why one of such proud bearing should be a suplient at the the prison door. However, he granted the boon-led her to the dungeon, opened the massive Iron door, then clesed it again, and the lovers were alone. He leaned against the prison wall with downeast head, and his arms were folded upon his breast. Gently she raised the vail fom her face, and Emmet torned to gaze upon all the earth contained for him-the girl whese sunny brow io the days of boybood had been his polar star-the maiden who had sometimes made him thinh 'the world was all sunshine.' The clanking of the heavy chains sounded like a death knei! to her ears, and she wept like a child. Emmet said but little, yet he preseed her warmly to his bosom, and their feelings held a

## Travel Narratives

## A Journey under Paris.

A correspondent of a Swedish journal furnishes an interesting account of a subterranean voyage made through one of the admirably constructed sewers of Paris. The boat which conveyed the party was reached by descending a flight of steps to the depth of about forty-five feet. The boat, $z$ flat-bottomed affair, was lighted by four lamps. The sewer is an archway, fifteen feet high, and of equal breadth, with a ditch of canal about ten feet wide, wherein all the dirt and filth of Paris is carried away. On the sides are sidewalks, which together are about four feet wide. The whole is built of beautiful sandstone, and is kept remarkably neat and clean. No stench or bad smell was perceptible. The denser portion of the filth is carried away through large drains beneath the sidewalks.
The sidewalks are excellent, and exhibit no signs of dampness, while the walls of the archway are kept whitewashec, and are at all times as white as the driven snow. The structure possesses the properties of an immense speaking tube, the workmen being able to converse at the distance of two miles from each other. The echo is very strong and lasting. The fabric is said to be built after a

## A HUSBAND'S CONFESSION.

ay tam smoons.
I never undertook but once to set át naught the authoity ofmy rife: You know her way, cool, quiet, but determined ns ever giew. Just aftry we were married and all was going nice and cozy she got me into the habit of doing all the churning, She never asked me to do it, you know, but then whe-why it was done just in this way. She finished brealfast rathor before the ore morning and slipped away from the table, whe filled the charn with cilust, and sat it jast where I couldn't he'p seeing what it yantedi So I look hoid regularly enough, nud churned till the butler cone. She did not thank me bat looked so nice and eveet about it, that I felt sell paid. Well, when the next chirpiog day crme along, she did the same thingi and I folloned suit and leiched the bulter. Again and again it way done just so; and I vas legularly in for it every there. Not a word said, you knon, of cuarse. Well, hy' and bye, this begain to be rather inssome, I wanted she should just ask me, but whe never did, and I couldn't eny anyihing about it to save my life, and so on we went. At last 1 made a resolve that I'would not churn another lime, unl ss sha asked me, Churning'day

## Jokes

## $\mathfrak{A l i d} d$ leburp

天n this ¥aper are publishey the ¥ublit orders, zesolutions, zaiws, wublit
prietor.
MIDDLEBURY VT
leasant and agreable employmen: for oung people at a time when little else can e done.-Albany Cultivator.

TESTIMONY AGAINST RADICAL. ISM.
In Document 109, of the House of Repesentatives, furnished us by Mr. Burke, ast printed, a Report of Mr. Ellsworth, Jommissioner of the Patent Office, we have latistics and observations, on the Agricularal and Manufacturing interests. We opy from this document the following im. ortant testimony :
ortant tes
Let us listen to the wisdom of those those opinions are recorded for our encour-gement-more especially since such opinons come from individuals who do not anction protection, except such as incidenally arises from the raising of a revenue:"To be independent for the comforts of fe, we must fabricate them ourselves. We ust now place the the manufacturer by the ide of the agriculturist. The grand inqui-

## MISCELLANEOUS.

THE RESTING PLACE.
-

EY CHARLES MACKAY.
Tell me, ye winged winds,
That round my path way roar,
Do ye not know some spot.
Where mortals weep no more ?
Some lone and pleasant dell, Some valley in the west,
Where free from toil and pain,
The weary soul may rest ?
The lond wind dwind led to a whisper low, And sigh'd for pity as it answered "No!"

Tell me thou mighty deep,
Where billows round may play
Know'st thou some favored spot,
Some island far away,
Where weary man may find
The bliss for which he sighs,
' Remember me to Lady ber me to Horatio! ren friends! Doctor, remem Tell him, I have made Hamilton and Horatio t repeated his remembran ton and Horatio, and to he said several times. G less agitated, and, at la ask questions about wh This led his mind to whom he sent and inqu iety, exclaiming aloud $h$ he was alive unless he s agitated at the captai mented his being unabl do what was to be done assurance given of the $c$ quarter-deck. At last, and he instantly grew 1 tened to his report abs fleet, directed him to an he should die, but obse half an hour longer.

Middlebury People's Press, 22 March 1843

THE RESTING PLACE.
ey charles mackay.

Tell me, ye winged winds,
That round my path way roar,
Do ye not know some spot.
Where mortals weep no more ?
Some lone and pleasant dell,
Some valley in the west,
Where free from toil and pain,
The weary soul may rest ?
The lond wind dwindled to a whisper low, And sigh'd for pity as it answered "No!"

## THE INQUIRY.

Tell me, ye winged winds,
That round my pathway roar,
Do ye not know some spot
Where mortals weep no more?
Some lone and pleasant dell,
Some valley in the west,
Where, free from toil and pain, The weary soul may rest ?
The loud wind dwindled to a whisper low, And sighed for pity as it whispered "No!"


"He touched his harp, and nations heard, entranced."

## THE INQUIRY.

Tell me ye winding winds, That round my pathway rear,
Do ye not know some spot
Where mortals weep no more-
Some lone and pleasant dell,
Some valley in the west,
Where, free from toil and pain,
The weary sonl may rest?
The loud wind dwindled to a whisper low
And sighed for pity, as it answered, "No!"

THE RESTING PLACE.
gy charles mackay.
Tell me, ye winged winds, That round my path way roar,
Do ye not know some spot.
Where mortals weep no more ?
Some lone and pleasant dell, Some valley in the west,
Where free from toil and pain,
The weary soul may rest?
The lond wind dwind led to a whisper low, And sigh'd for pity as it answered "No!"

## THE INQUIRE.

Tell me, ye winged winds,
That round my pathway roar,
Do ye not know some spot Where mortals weep no more? ${ }^{\text {- }}$ Some lone and pleasant dell, Some valley in the west, Where, free from toil and pain, The weary soul may rest ? The loud wind dwindled to a whisper low, And sighed for pity as it whispred "No!"

## THE INQUIRY.

Tell me ye winding winds, That round my pathway r Do ye not know some spot Where mortals weep no me Some lone and pleasant dell, Some valley in the west,
Where, free from toil and pa
The weary soul may resi?
The loud wind dwindled to a whisp And sighed for pity, as it answered;

## GRAND HAVEN, $\boldsymbol{M}$

> A PARODY.
> BY AN OLD BACHELOR.
> Tell me, ye winged winds, That round my pathway roar, Do ye not know some spot Where women fret no more? Some lone and pleasant dell; Some "holler" in the ground, Where babies never yell, And cradles are not found ? The loud wind blew the snow into my face, And snickered as it answered-" Nary place."

> Tell me, thou misty deep, Whose billows round me play,




Local, not global, alignment



```
sn83035487/1853-12-31/ed-1 (1853-12-31) Anti-slavery bugle. (New-Lisbon, Ohio) 1845-1861
The language of some household song, The perfume of some cherished flower Though gone from outward sense, belong To memory's after hour.
So with our words, or harsh, or kind,
Uttered, they ate not all forgot,
They leave their Influence on the mind,
Post on, but perish not ;
As they are spoken, so they fal
Upon the spirit spoken to,
Scorch it like drops of burning gall,
Or sooth like honey-dew.
So with our dcods, for good or ill,
They have thoir power, scarce understood
Then let us use our better will
To make them rife with good:
Like
```

harp_harp0011-4 (1855-09-01)
Harper's New Monthly Magazine Volume 0011 Issue 64 (September 1855)
The language of some household song,
The perfume of some cherished flower,
Though goue from outward sense, belong
To Memorys after-hour

So with our words: or harsh or kind
Uttered, they are not all forgot:
They have their influence on the mind,
Pass onbut periels noL
So with our deeds: for good or ill,
They have their power scarce understood;
Then let us use our better will,
To make them rife with good

## How Much Reprinting?



## How Much Reprinting?



## Where was the Reprinting?





Papers following the Lexington [Mo.] Weekly Caucasian in 1872 (Virality and Racial Terror)

## Cascade on a Graph



## Reprints by Editors' Affinities

- $>1$
$\square>=10$
$\square>=100$


Newspaper Affinity

## Cascades as Trees



## Tree Inference

| Data | Method | F1 |
| :--- | :--- | :--- |
| Separate cascades | DST Basic | 0.62 |
|  | DST Enhanced | 0.95 |
| Attach-first baseline | 0.93 |  |
| Merged cascades | DST Basic | 0.04 |
|  | DST Enhanced | 0.25 |
|  | Attach-first baseline | 0.04 |
|  | Sup. DST Basic | 0.16 |

## Network Inference





## Node Prediction



MemeTracker data from Wang et al., 2017

## Stay Tuned for the Book

$\leftarrow \rightarrow$ C © staging.manifoldapp.org/project/going-the-rounds



House action


## Affordable Care Act ("Obamacare")



## Citation Lineages (isnads)

حدثنا أبو داود قال: حدثنا هشام، عن قتادة، عن الحسن عن سمرة، أن النبي صلى اله عليه وسلم

Abu Dawud transmitted to us, saying, "Hisham transmitted to us, from Qatada, from al-Hasan, from Samura that the Prophet, may the peace and blessing of God be on him . . ."

Confidering the Magnificence of their Houfes, I fhould be apt to think that both the Materials and Workmanfhip were cheap. ${ }^{1}$ M. Lepidus's Houfe was the firt that had a Marble Door-cafe. ${ }^{m}$ Afterwards they had gilded ones, or rather plated with Gold: ${ }^{n}$ Then they began to cafe their Houfes with Marble: I think it was Mamurra, Cefar's Mafter-Carpenter in Gaul, that built the firtt of this kind. ${ }^{\circ}$ Afterwards they gilded their very Walls. ${ }^{\mathrm{P}}$ Within their


#### Abstract

g Muerádav $\dot{\varepsilon} \pi 7$ à $\dot{n} \mu i \sigma \varepsilon s$. $\quad$ M Mueciday $\pi \varepsilon v$ - Senecam Rhetorem. Primum Romx parietes тíxorla xj slaxooiar. Plutarch. in Mario. i M. Lepido \& Q. Catullo Coff, ut conflat inter ditigentiffimos authores, domus pulchrior non fuit Romæ quam Lepidi ipfius: at hercule intra annos triginta quinque, eadem centefimum Incum non obtinuit. Plin. lib. 36 cap. 15. k Publius Clodiusquem Milo occidit, centies $\&$ quadragtes octies domo emptâ habitavit. Plin. lib. 36. cap. 15. 1 Plin. lib. 36. cap. 6. m Hieronymus in Epilt. ad Marcellam, n Fabianus Papirius apud crufta Marmoris operuiffe totius domus fux in Cælio Monte Cornelius Nepos tradidit Mamurram Formiis natum, Equitem Romanum, Præfectum fabrûm C. Cæfaris in Gallia Plin. lib. 36. cap. 6. o Plin. lib. 33. cap. 3. Hieronymus in Epift. ad Gaudent. Petronius, ædificant auro. P Parietes Tyriis \& Hiacynthinis \& illis regiis velis, que vos operofe refolutatranffiguratis, pro pieturâ abutuntur. Tertullian. De Habitu Mulierum.


## Diachronic Syntactic Constraints

I will always remember my first visit to Boston.
?My first visit to Boston will always be remembered by me.
Their first visit to Boston will always be remembered by the French triplets.



## Diachronic Syntactic Constraints




## beach

## $\downarrow$

Text-to-image models: DALL•E, Parti, Stable Diffusion, ...


## High Imageability Words



MRC Psycholinguistics Database

## Low Imageability Words



MRC Psycholinguistics Database

## Correlation w/Human Judgements



## Artificial Intelligence as an Archival Science

# Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning 

Eun Seo Jo<br>Stanford University<br>eunseo@stanford.edu

Timnit Gebru<br>Google<br>tgebru@google.com


#### Abstract

A growing body of work shows that many problems in fairness, accountability, transparency, and ethics in machine learning systems are rooted in decisions surrounding the data collection and annotation process. In spite of its fundamental nature however, data collection remains an overlooked part of the machine learning (ML) pipeline. In this paper, we argue that a new specialization should be formed within ML that is focused on methodologies for data collection and annotation: efforts that require institutional frameworks and procedures. Specifically for sociocultural data, parallels can be drawn from archives and libraries. Archives are the longest standing communal effort to gather human information and archive scholars have already developed the language and procedures to






Shent = oustlither
(1ว) 1มเाt.
von
(1) GUTME.

 1819.$)$

$$
\begin{aligned}
& \text { كر آن شيرين پسر خونم بريزد } \\
& \text { دلا چین شير مـادر كن حالالش }
\end{aligned}
$$

O mein Herz! wenn dein Blut das süße Mädchen vergießet, Soll es wie Muttermilch dir gedeihen.

If that sweet one spill my blood, O heart it, like mother's milk, lawful hold.


O mein Herz! wenn dein Blut das süße Mädchen vergießet, Soll es wie Muttermilch dir gedeihen.

If that sweet one spill my blood, O heart it, like mother's milk, lawful hold.

## Clarivate Acquiring ProQuest for \$5.3B

By Lindsay McKenzie // May 18,2021

Research information company Clarivate yesterday announced plans to acquire information content and technology company ProQuest for $\$ 5.3$ billion. The deal is expected to close in the third quarter of this year.

Leaders at the two companies described their businesses as highly complementary during a call for investors. They said the acquisition will enable the organization to serve students, faculty members and staff members with data, resources and software at every stage of the academic research process.

Clarivate was formerly the intellectual property and science division of Thomson Reuters but was spun off into an independent company in 2016. The company is well-known for its annual Journal Citation Reports, as well as research discovery and workflow management tools such as Web of Science, EndNote and ScholarOne.


Dissertations

## American Periodicals Series

## THE LADIES HOME JOURNAL.



NOS P4E MNS ONS

 mesor innin mon


macrink.
 answan




WE KNOW YOU

 of 2ns. N.. varing
 Пn.

One IProp at a Time. I WtfC: says the late John Foster, ' is en. pe-; sie have it, bat as co.. lo.ing it; we have she u-e of it, hut a, cots. massing iL Supporte a man in Pcume Aim" rre ander the doom so stay there till his death; and there lo there ber hi, uve a dark of water. to which it is certain none can ever be added. lie knows, .o, that the quantity is wdt very great; he penetrate to ascertain ho. much, but it may be very little. lie has drawn fr.m it, by of . a good while already, and from it every day. Bult hor $r$ sld he feel each time os . *nd each time of thinking of itl Not as if he had a perennial spring to go to; non ' I have a reservoir, I may be at ease' Nol ' I had water yeaterday-I bare water to-day; but my had it, and my having it to-day. is the very ca.se that I shall to4 have it otn -,ne day that iv itig. Atd at the time I am Sn thin fatal !' So of our mortal, life! And yet, men are very -ed to admit the plain troth, that lire is a thing whith they are in ion other way than as ; and even in [hi. of po..n it every day lev a Ara,.z Ditrus.-- Chitterden, chief magistrate of Vermont, was of humble birth, and rove by the force of talent to his station. Yet while Glovernor of the green-mon..tain still continued u heep the same tavern, upon the steep bill side, that he kept for many years before. One evening, a drove op and accs ted him than: *( Chitteaden, a. chief of Vermon"t I render, yot due homage; but vs landlord Chittenden, IU thank to out heor.,:

## Recommendations for Cultural Institutions



ANDREW W.

NATIONAL ENDOWMENT FOR THE
Humanities

## Recommendations for Cultural Institutions

- Mellon/NEH report: Establish "service bureau" for OCR.
- Library of Congress report: Libraries should pool resources or contract with non-profit organizations to provide machinelearning support.
- OCLC report: Libraries should write specifications for "datascience service plans".
- OCR-D project: Develop digitization workflows for libraries, training for historical typefaces.


## Al for the Human Record

- Archive-scale transcription for printed, handwritten, spoken documents
- Inferring relational structure: quotes, cites, and document genealogies
- Modeling text/image interactions: figures, maps, and plans
- Search systems for iterative, interpretable, collaborative modeling
nothing but the the history of the lat if ever we did rties would be so $r$ the defeat of an he defeated party leserved calamity, ar in which we ever likely to continent more when it began, stance of a purely e to have even a for the purpose of ppression, and not retation of some it as independent n . If it be posnd America, it is of doing so. It what we want. ength is what we etween two great revent, or greatly as, hopeless, and vil war in characsies of national
it would seem good opportunity more a cordial least if we may the other side of
events, unless the accounts from many quarters as to General Schenck's instructions are utterly belied, the new American Ambassador will bring us quite reasonable, though not perhaps wholly admissible demands, - demands which we certainly ought to consider most gravely, and of which we should do well to yield frankly and freely all that we should ourselves feel called upon, in the same circumstances, to press. If we do so, General Schenck's mission may make England safer and stronger than she has ever been since the close of the Civil War in 1865, and will give her a reputation for moderation and candour as well.

ENGLISH PUBLIC OPINION ON THE WAR. Some of the philosophers should turn their attention from the subject of spectroscopic investigations and the invention of electrometers, galvanometers, hygremeters, and so forth, to the far more difficult problem of inventing a mode of measuring the intensity and diffusion of political wishes and convictions. No task at present is more difficult for a Statesman than this. There are, indeed, all sorts of shades of difference between the character of really prevalent and preponderant public opinions, of which no man, however acute, ever forms more than a purely conjectural impression, and of which, nevertheless, any respectably-accurate measure would be a matter of the highest political importance. For instance, there is at times a public opinion on one side of a question which is very widely diffused, but of very slight in-tensity,-which, in fact, amounts to nothing more than a wish in a particular direction without a will, and still more without any intention of submilting to a considerable sacrifice rather than not carry out the will into action. Again, there is such a thing as

## Thanks!



Ryan
Cordell


Sarah Savant


Rui
Dong


Giulia
Taurino


Alejandro Toselli


Abby Mullen


Shaobin Xu


Ryan Muther


Si
Wu

And funders at the Mellon, NSF, NEH, NIH, QNL, and ERC.

