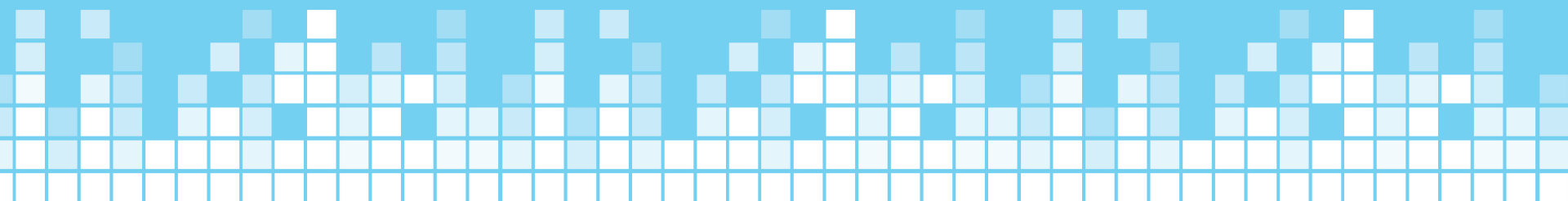




BIG DATA
SUMMIT

Big Data Case Studies

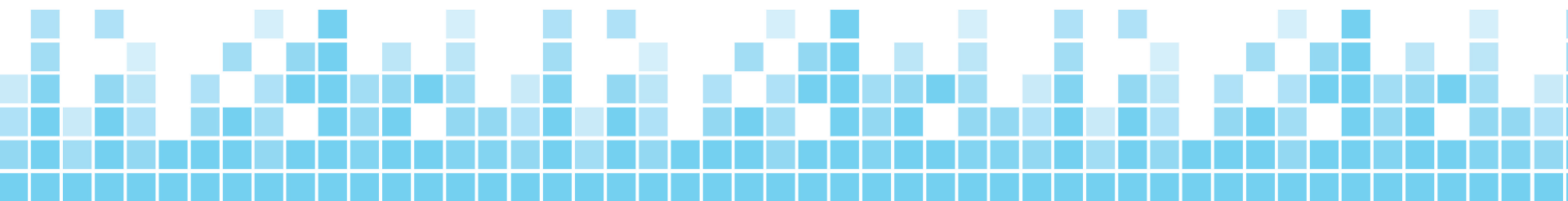
Sayan Bhattacharyya



The HathiTrust Research Center's Extracted Features Dataset:

An Opportunity for "Distant" Reading of
Millions of Books from the World's Great
Research Libraries

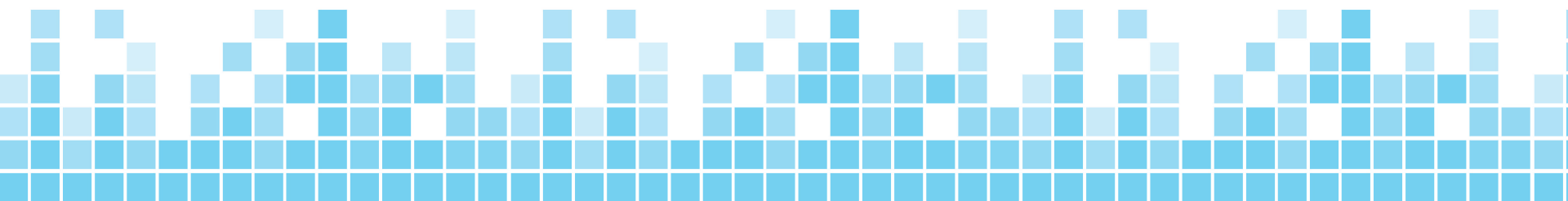
- Sayan Bhattacharyya (sayan@illinois.edu)
- Peter Organisciak (organis2@illinois.edu)
- J. Stephen Downie (Project PI)
Graduate School of Library and Information Science, UIUC, Urbana-Champaign
- Loretta Auvil and Boris Capitanu (Illinois Informatics Institute, UIUC) Ted Underwood (Department of English, UIUC, Urbana-Champaign)



WHAT

is

IT?



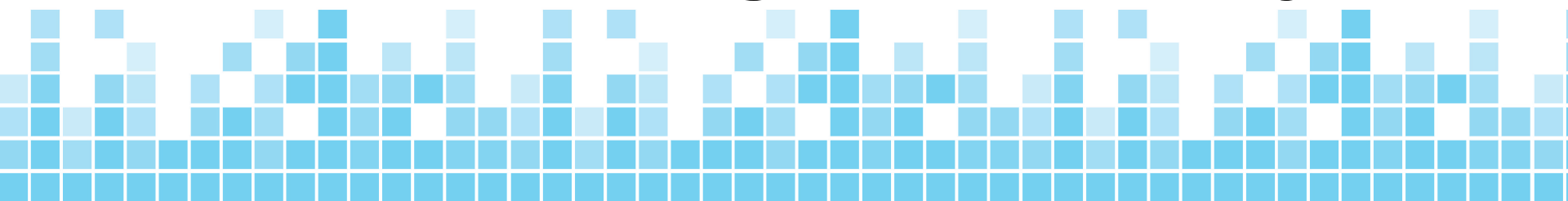
a dataset of
page-level
extracted features

for

scanned books

in the

HathiTrust Digital Library



Raw Text



Translation into features

(we drop you off here)



Algorithmic Use



had taken mental note of everything **that** was on **that** table. There were three plates laid, so **that** Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so **that** the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly. "This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something **that** for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck aggressively through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid *you* weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips **twitched** understandingly. She had expected Mrs. Rachel up; she had known **that** the

```
"that": {
  "DT": 1,
  "IN": 4,
  "WDT": 1
},
"the": {
  "DT": 5
},
"there": {
  "EX": 2
},
"thin": {
  "JJ": 1
},
"this": {
  "DT": 1
},
"though": {
  "IN": 1
},
"thought": {
  "VBD": 1
},
"three": {
  "CD": 1
},
"through": {
  "IN": 1
},
"to": {
  "TO": 2
},
"to-day": {
  "NN": 1
},
"twisted": {
  "VBN": 1
},
"twitched": {
  "VBD": 1
},
}
```

(Determiner,
Preposition or
subordinating
conjunction,
Wh-determiner)

(Verb, past
tense)

"Say, ye oppress by some fantastic woes,
Some jarring nerve that baffles your repose,
Who press the downy couch while slaves advance
With timid eye to read the distant glance;
Who with sad prayers the weary doctor tease
To name the nameless ever-new disease;
Who with mock patience dire complaints endure,
Which real pain and that alone can cure:
How would you bear in real pain to lie
Despised, neglected, left alone to die?
How would ye bear to draw your latest breath
Where all that's wretched paves the way to death?"

CRABBE.

It was a dark and stormy night; the rain fell in torrents—except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness. Through one of the obscurest quarters of London, and among haunts little loved by the gentlemen of the police, a man, evidently of the lowest orders, was wending his solitary way. He stopped twice or thrice at different shops and houses of a description correspondent with the appearance of the *quartier* in which they were situated,—and tended inquiry for some article or another which did not seem easily to be met with. All the answers he received were couched in the negative; and as he turned from each door he muttered to himself, in no very elegant phraseology, his disappointment and discontent. At length, at one house, the landlord, a sturdy butcher, after rendering the same reply the inquirer had hitherto received, added,—“But if *this* will do as vell, Dummie, it is quite at your service.” Pacing reflectively for a moment Dummie re-

had taken mental note of everything that was on that table. There were three plates laid, so that Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so that the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly. "This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something that for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck **agresively** through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid *you* weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips twitched understandingly. She had expected Mrs. Rachel up; she had known that the

sight of Matthew jaunting off so unaccountably would be too much for her neighbour's curiosity.

"Oh, no, I'm quite well although I had a bad headache yesterday," she said. "Matthew went to Bright River. We're getting a little boy from an orphan asylum in Nova Scotia and he's coming on the train to-night."

If Marilla had said that Matthew had gone to Bright River to meet a kangaroo from Australia Mrs. Rachel could not have been more astonished. She was actually stricken dumb for five seconds. It was unsupposable that Marilla was making fun of her, but Mrs. Rachel was almost forced to suppose it.

"Are you in earnest, Marilla?" she demanded when voice returned to her.

"Yes, of course," said Marilla, as if getting boys from orphan asylums in Nova Scotia were part of the usual spring work on any well-regulated **Avonlea farm** instead of being an unheard of innovation.

Mrs. Rachel felt that she had received a severe mental jolt. She thought in exclamation points. A boy! Marilla and Matthew Cuthbert of all people adopting a boy! From an orphan asylum! Well, the world was certainly turning upside down! She would be surprised at nothing after this! Nothing!

"What on earth put such a notion into your head?" she demanded disapprovingly.

This had been done without her advice being asked, and must perforce be disapproved.

"Well, we've been thinking about it for some time—all winter in fact," returned Marilla. "Mrs. Alexander Spencer was up here one day before

had taken mental note of everything that was on that table. There were three plates laid, so that Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so that the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly. "This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something that for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck aggressively through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid *you* weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips twitched understandingly. She had expected Mrs. Rachel up; she had known that the

sight of Matthew jaunting off so unaccountably would be too much for her neighbour's curiosity.

"Oh, no, I'm quite well although I had a bad headache yesterday," she said. "Matthew went to Bright River. We're getting a little boy from an orphan asylum in Nova Scotia and he's coming on the train to-night."

If Marilla had said that Matthew had gone to Bright River to meet a kangaroo from Australia Mrs. Rachel could not have been more astonished. She was actually stricken dumb for five seconds. It was unsupposable that Marilla was making fun of her, but Mrs. Rachel was almost forced to suppose it.

"Are you in earnest, Marilla?" she demanded when voice returned to her.

"Yes, of course," said Marilla, as if getting boys from orphan asylums in Nova Scotia were part of the usual spring work on any well-regulated Avonlea farm instead of being an unheard of innovation.

Mrs. Rachel felt that she had received a severe mental jolt. She thought in exclamation points. A boy! Marilla and Matthew Cuthbert of all people adopting a boy! From an orphan asylum! Well, the world was certainly turning upside down! She would be surprised at nothing after this! Nothing!

"What on earth put such a notion into your head?" she demanded disapprovingly.

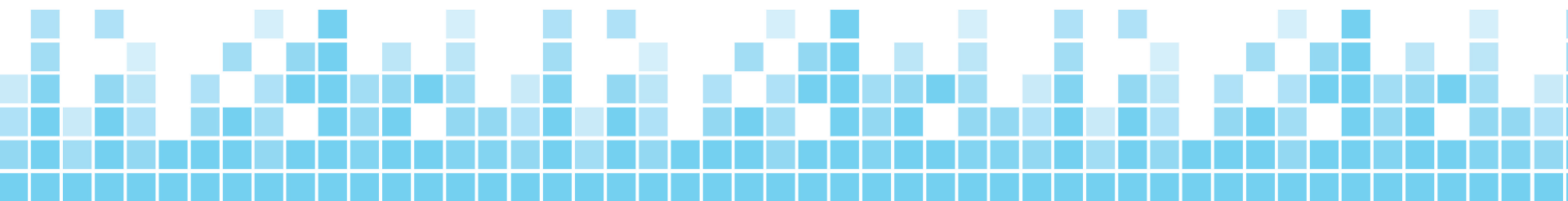
This had been done without her advice being asked, and must perforce be disapproved.

"Well, we've been thinking about it for some time—all winter in fact," returned Marilla. "Mrs. Alexander Spencer was up here one day before

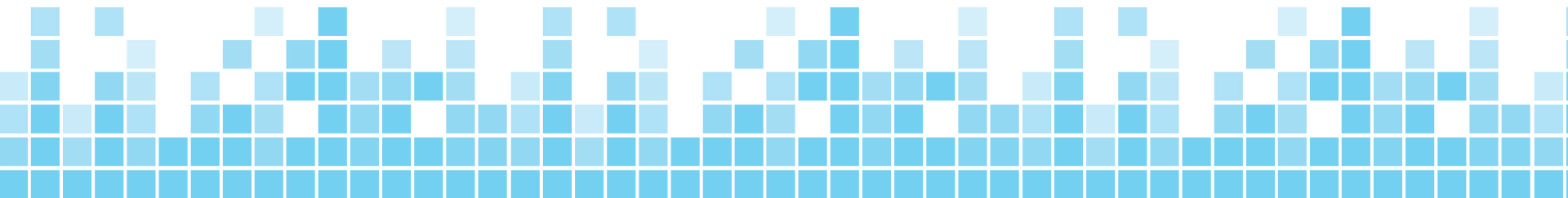
WHY

should you

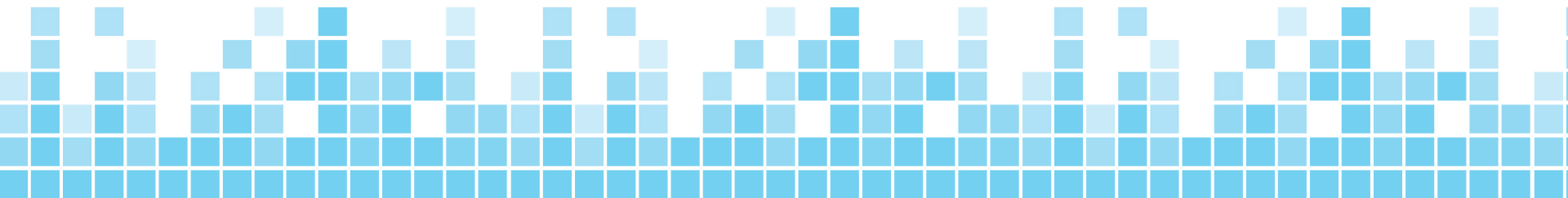
CARE?



- 1) It's huge
- 2) It's accessible
- 3) You can do cool stuff with it

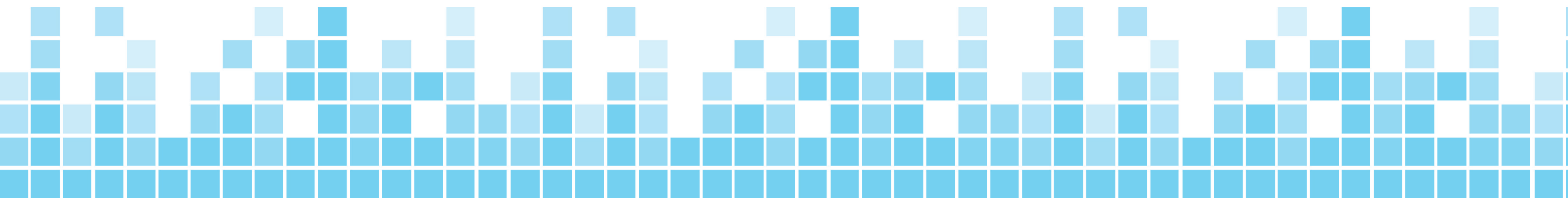


1) It's Huge



The HathiTrust Digital Library (HTDL)

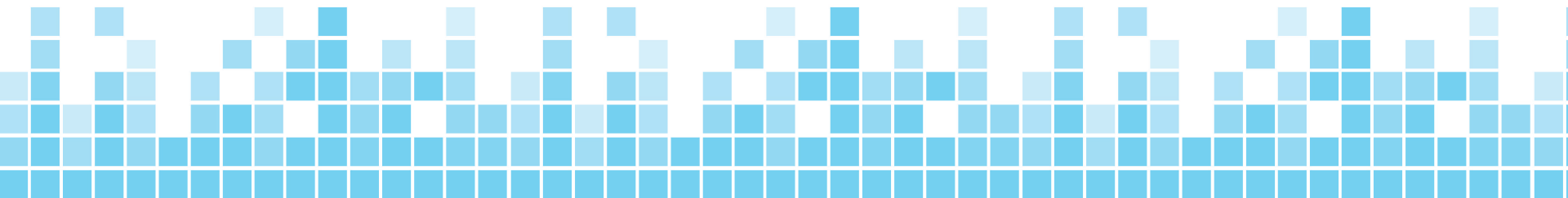
- Approximately 14 million books
 - from the world's great research libraries:
 - a large chunk of mankind's historical textual record of culture
- 1.8 billion pages
- 610 billion words
- Approximately 4.8 million of the 14 million books are in the public domain
 - Current applications are set up to work with these 4.8 million books





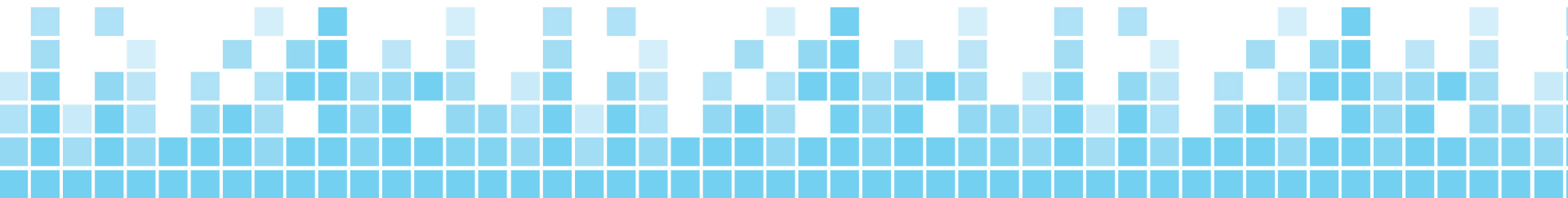
2) It's

Accessible

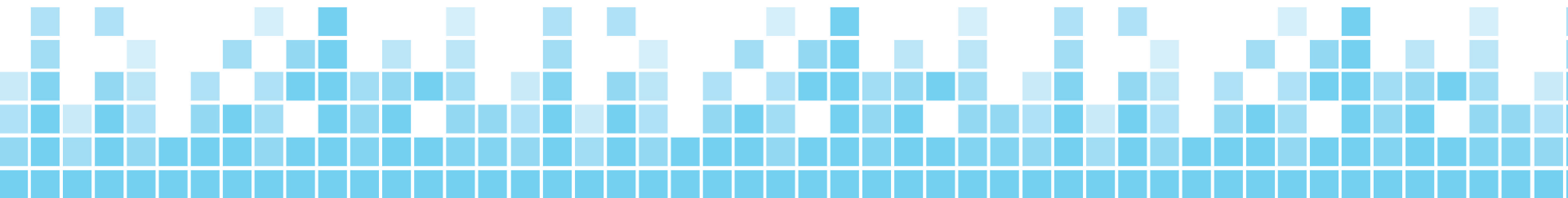


One of the largest
archives of pre-digital
human creation,
downloadable

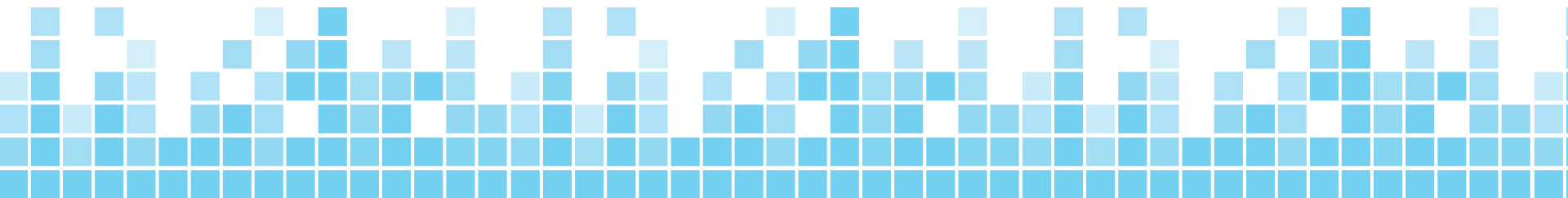
<https://sharc.hathitrust.org/features>



Can't share 10 million
in-copyright works,
but...



3) can do useful
stuff with it



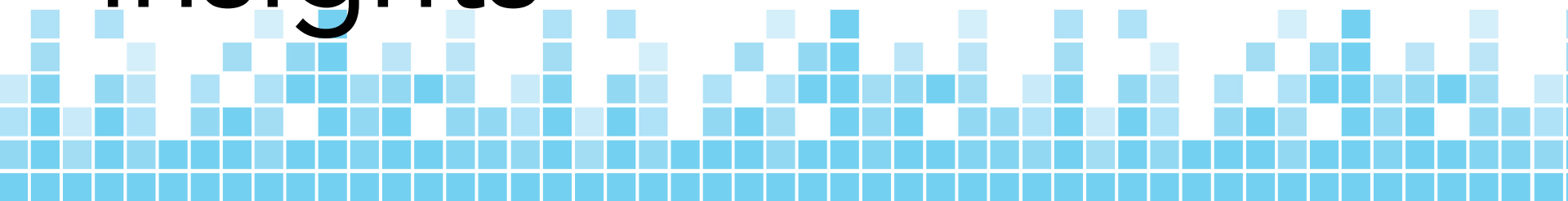
Large corpora allow for

–*historical*

–*cultural*

–*linguistic*

insights



Co-occurrence tables

by David Mimno

(Computer Science Dept., Cornell University)

Available for use at:

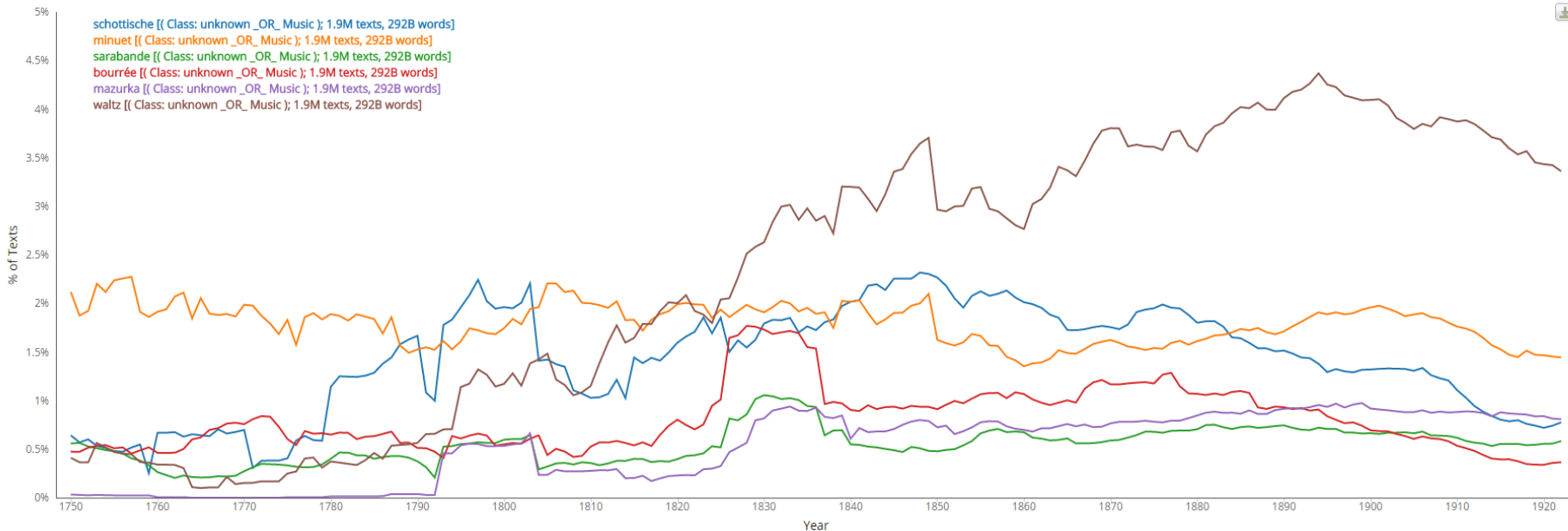
<http://mimno.infosci.cornell.edu/wordsim/nearest.html>

Explanatory article at:

<http://www.mimno.org/articles/wordsim/>



schottische	in	(Class: unknown OR Music)	-	+
minuet	in	(Class: unknown OR Music)	-	+
sarabande	in	(Class: unknown OR Music)	-	+
bourrée	in	(Class: unknown OR Music)	-	+
mazurka	in	(Class: unknown OR Music)	-	+
waltz	in	(Class: unknown OR Music)	-	+



Bookworm

<http://bookworm.htrc.illinois.edu>

Faceted visualization of trends over 4.8 million books

The HathiTrust+Bookworm project

Team Members:

– Current:

J. Stephen Downie, University of Illinois, Urbana-Champaign
Erez Lieberman Aiden, Baylor College of Medicine
Benjamin Schmidt, Northeastern University
Robert McDonald, Indiana University
Loretta Auvil, University of Illinois, Urbana-Champaign
Peter Organisciak, University of Illinois, Urbana-Champaign
Muhammad Shamim, Baylor College of Medicine
Sayan Bhattacharyya, University of Illinois, Urbana-Champaign
Leena Unnikrishnan, Indiana University

– Past:

Colleen Fallaw, University of Illinois, Urbana-Champaign
Matt Nicklay, Baylor College of Medicine

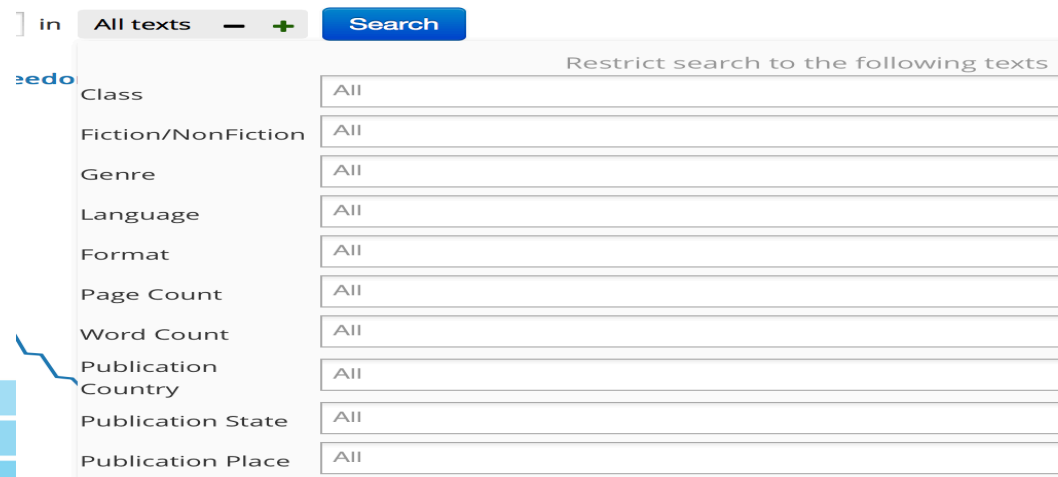


Funded by an NEH Implementation Grant (2014-2016)

Hooking up Extracted Unigrams with Bookworm: Advantages?

First advantage: Good metadata!

- HTDL has good and detailed metadata
 - metadata was meticulously created by librarians from contributing libraries
- allows for highly faceted queries:

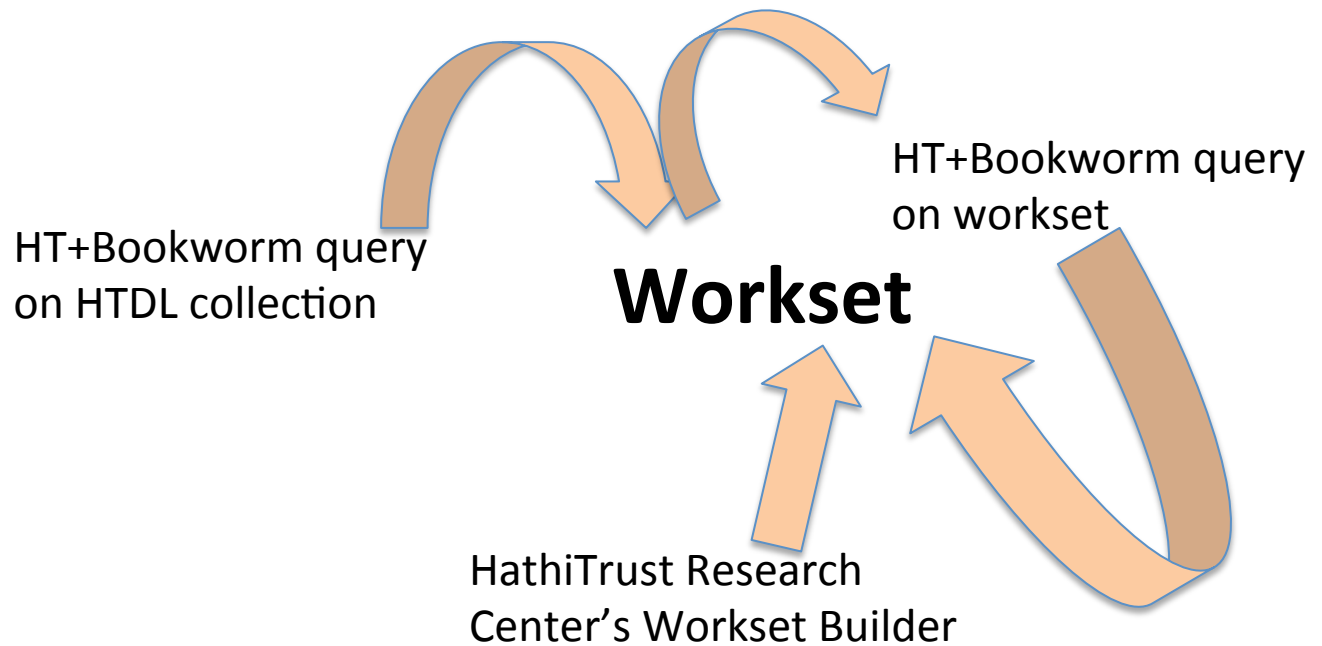


The screenshot shows a search interface with a search bar and a list of facets. The search bar contains the text "in All texts" and a "Search" button. Below the search bar is a table with the following facets and their current values:

Restrict search to the following texts	
Class	All
Fiction/NonFiction	All
Genre	All
Language	All
Format	All
Page Count	All
Word Count	All
Publication Country	All
Publication State	All
Publication Place	All

Hooking up HTDL with Bookworm: Advantages?

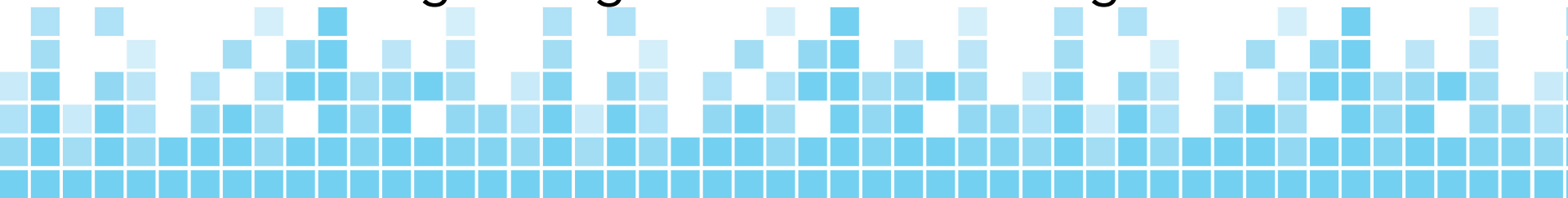
Second advantage: HTDL's *workset* functionality
(contd.)



**Workset creation and refinement
workflow**

How *scientific inquiry* meets *humanistic inquiry* in **culturomics** as performed by HT **+Bookworm**

- Scientific inquiry concerns:
 - *Generalization* across entities
 - Discovery of patterns *across* entities
- Humanistic inquiry concerns:
 - Close engagement with *specific* entities
 - Attending to singular instances among entities



DATASET

<https://sharc.hathitrust.org/features>

ACKNOWLEDGEMENTS

Boris Capitanu Ted Underwood

Loretta Auvil

Colleen Fallaw J. Stephen Downie

Benjamin Schmidt (Bookworm)

Special thanks to the National Center for
Supercomputing Applications (NCSA)

National Endowment for the Humanities (NEH)

Contact:

Peter Organisciak
organis2@illinois.edu

Sayan Bhattacharyya
sayan@illinois.edu

