Finding Research Data

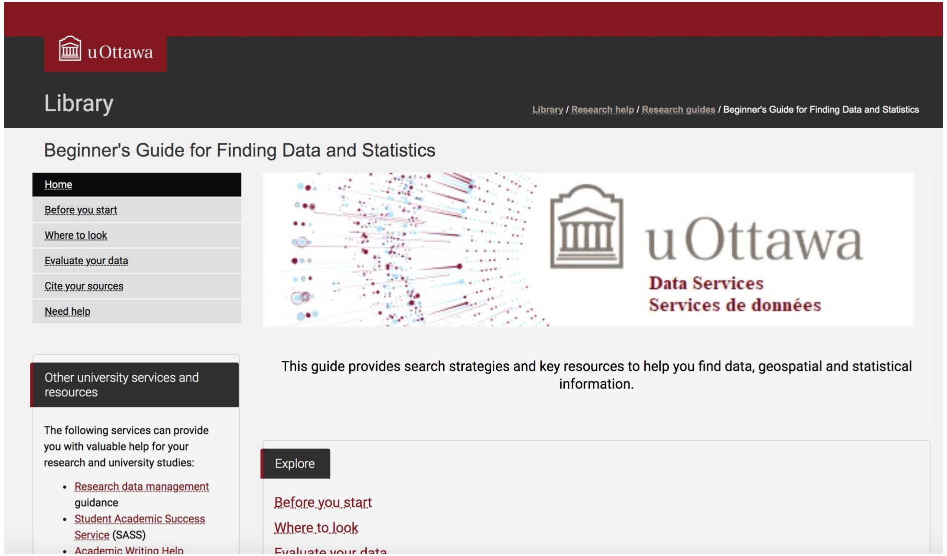
DH Toolbox: Fall 2020

Yoo Young Lee
@yooylee

About this workshop

- Interactive and hands-on workshop
- Based on research studies
- Technical components for a beginner
- Broad topics

About this workshop



uOttawa

Library

Library / Research help / Research guides / Beginner's Guide for Finding Data and Statistics

Beginner's Guide for Finding Data and Statistics

- Home
- Before you start
- Where to look
- Evaluate your data
- Cite your sources
- Need help

This guide provides search strategies and key resources to help you find data, geospatial and statistical information.

Other university services and resources

The following services can provide you with valuable help for your research and university studies:

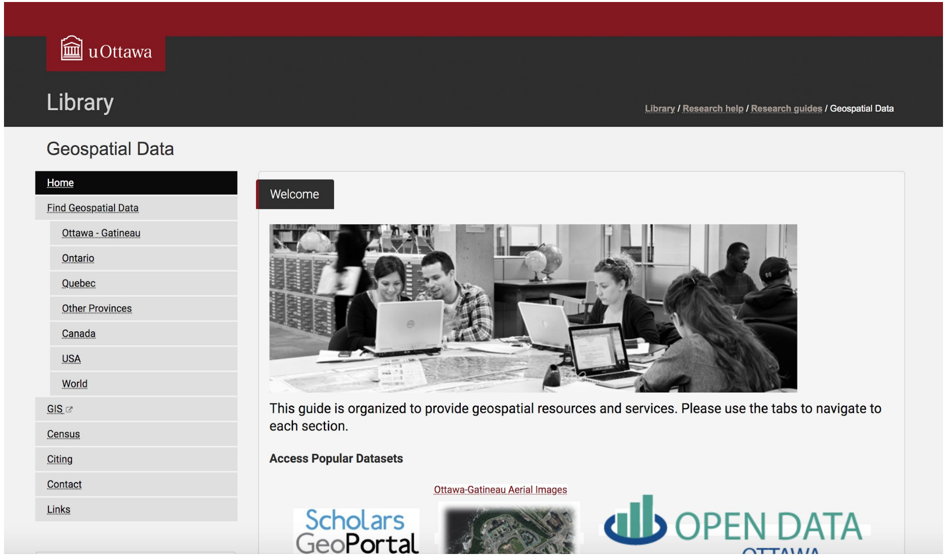
- Research data management guidance
- Student Academic Success Service (SASS)
- Academic Writing Help

Explore

Before you start

Where to look

Evaluate your data



uOttawa

Library

Library / Research help / Research guides / Geospatial Data

Geospatial Data

- Home
- Find Geospatial Data
  - Ottawa - Gatineau
  - Ontario
  - Quebec
  - Other Provinces
  - Canada
  - USA
  - World
- GIS
- Census
- Citing
- Contact
- Links

Welcome

This guide is organized to provide geospatial resources and services. Please use the tabs to navigate to each section.

Access Popular Datasets

Ottawa-Gatineau Aerial Images

Beginner’s Guide for Finding Data and Statistics

Geospatial Data

Data services team:
https://uottawa.libguides.com/intro-datastats-eng/help

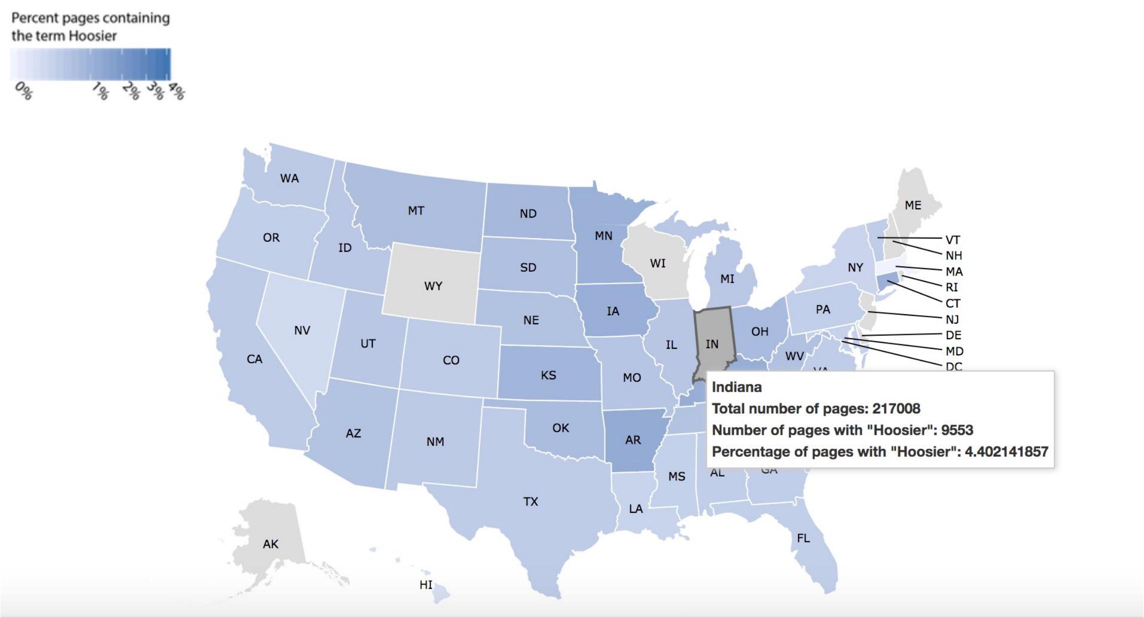
René Duplain: rene.duplain@uottawa.ca

About me and about you

First poll: What types of data have you used for your studies?

Second poll: Have you used APIs to access datasets for your research?

#1. Chronicling Hoosier

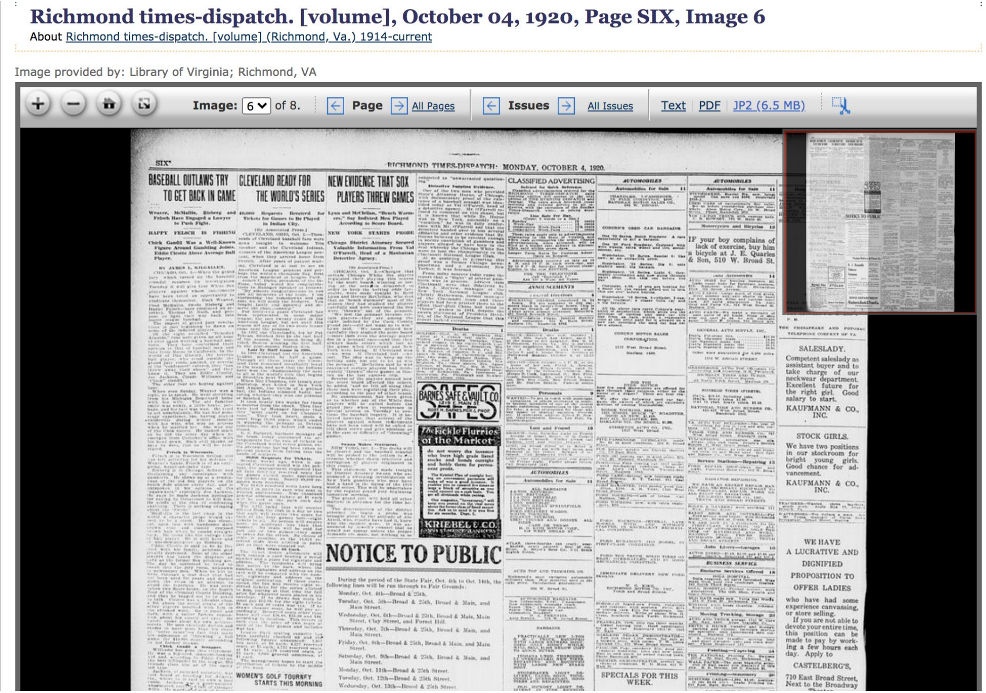


Indiana
Total number of pages: 217008
Number of pages with "Hoosier": 9553
Percentage of pages with "Hoosier": 4.402141857

Chronicling Hoosier (Palmer, Polley, & Pollock, 2016)

- Usage of the word hoosier through time and across geographies

Chronicling America



Richmond times-dispatch. [volume], October 04, 1920, Page SIX, Image 6

About Richmond times-dispatch. [volume] (Richmond, Va.) 1914-current

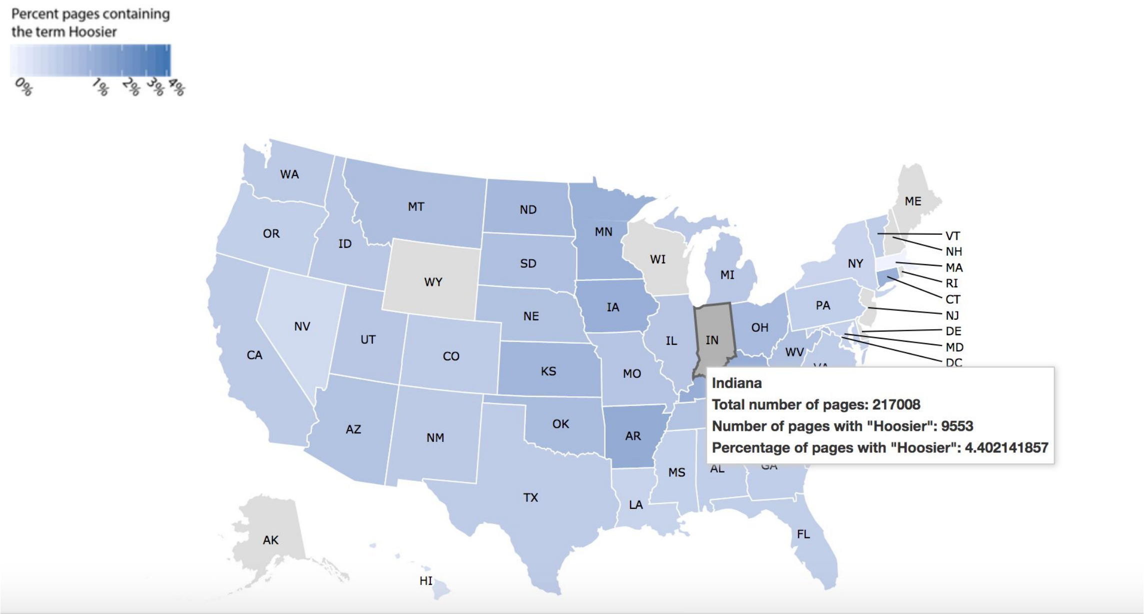
Image provided by: Library of Virginia; Richmond, VA

Image: 6 of 8.
Page
All Pages
Issues
All Issues
Text
PDF
JP2 (6.5 MB)
RICHMOND TIMES-DISPATCH: MONDAY, OCTOBER 4, 1920.
BASEBALL OUTLAWS TRY TO GET BACK IN GAME
CLEVELAND READY FOR THE WORLD'S SERIES
NEW EVIDENCE THAT SOX PLAYERS THREW GAMES
CLASSIFIED ADVERTISING
NOTICE TO PUBLIC

Chronicling America

- Access to information historic newspaper and select digitized newspaper pages
- API

#1. Chronicling Hoosier

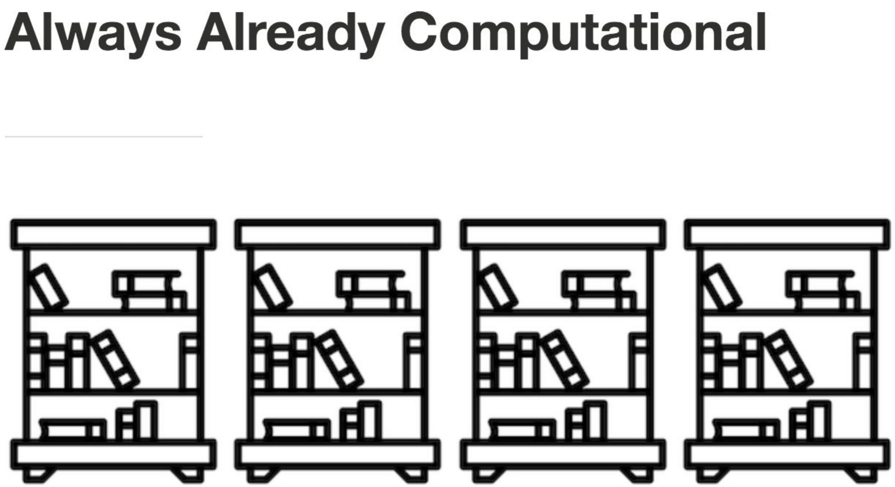


Indiana
Total number of pages: 217008
Number of pages with "Hoosier": 9553
Percentage of pages with "Hoosier": 4.402141857

Chronicling Hoosier
(Palmer, Polley, & Pollock, 2016)

- Usage of the word hoosier through time and across geographies

Collection as Data



Always Already Computational

Collections as Data

- National initiative to develop a strategic direction to creation of a collection as data framework and identification of methods for making computationally amenable library collections

API access

- API: Application Programming Interface
- Set of definitions and protocols for building and integrating application software.

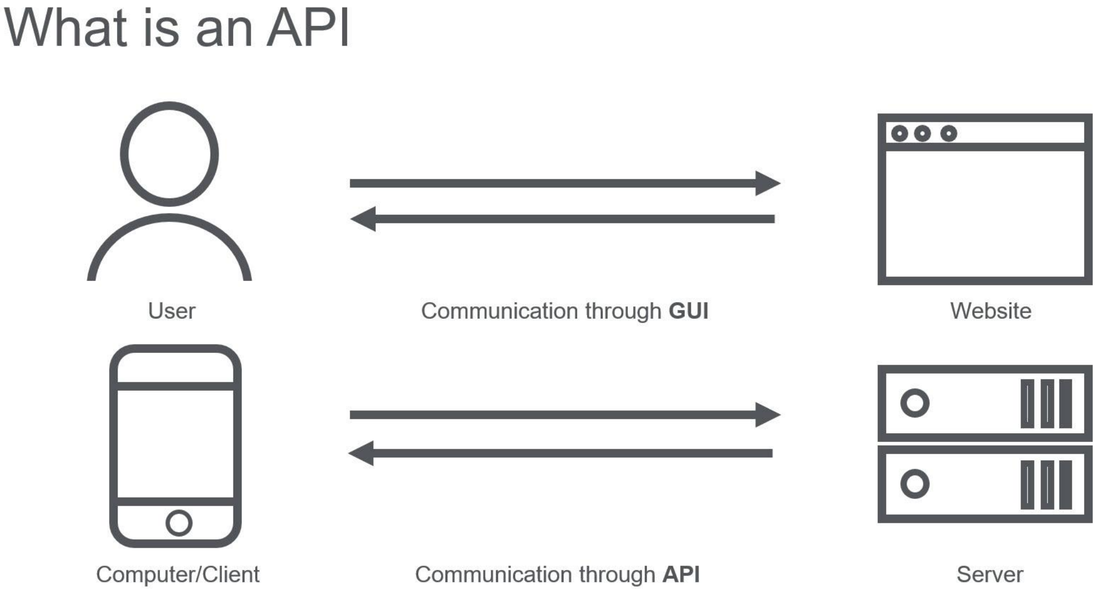


Photo Credit: Scopus, What is an API

API access to digital collections

- In Canada:
  - The University of British Columbia (UBC) Library's Open Collections and API documentation (API key is required)
  - Open Data from Library and Archives Canada (LAC) and API documentation (API key is not required)
    - Soldiers of the First World War - CEF
    - Maps, Plans and Charts of Canada
  - Open Parliament and API documentation

API access to digital collections

- In the States:
    - Metropolitan Museum of Art Collection and API Documentation
    - HathiTrust Bibliographic and API Documentation
    - HathiTrust Data and API Documentation
    - National Library of Medicine and API Documentation
    - Chronicling America and API Documentation
    - Digital Public Library of America (DPLA) and API Documentation
    - Library of Congress and API Documentation
    - Data.gov and API Documentation
    - New York Public Library Digital Collections and API Documentation

API access to digital collections

- Around the world:
    - OECD Data and API Documentation
    - Europeana and API Documentation
    - World Bank and API Documentation
    - UN Comtrade and API Documentation
    - eurostat and API Documentation
    - Internet Archive and API Documentation

API access to digital collections

- Databases:
  - arXiv and API Documentation
  - CORE and API Documentation
  - CrossRef and API Documentation
  - IEEE Xplore and API Documentation
  - PLoS and API Documentation
  - Science Direct and API Documentation
  - SCOPUS and API Documentation
  - Springer and API Documentation
  - Web of Science and API Documentation

Hands-on Activity: Chronicling America and OpenRefine

Third poll: Are you interested in learning more about API access to datasets?

Fourth poll: Have you used a large dataset of tweets for your research?

#2. The 42nd Canadian Federal Election on Twitter

An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter (Ruest and Milligan, 2016)

- Dataset: 4 million tweets during the 2015 Canadian Federal Election using the hashtag #elxn42
- Tool: Twarc developed by Ed Summers

Datasets on Dynamics of Coronavirus on Twitter

- Aguilar-Gallegos et al., 2020
- Datasets: 8,982,694 Twitter posts
- Tool: rtweet R package using Twitter REST API search
- Data range: January 21 to February 12, 2020
  - The data collection started when there was a boom in the use of hashtags related to the coronavirus outbreak in China. Those hashtags became trending topics very quickly. This way, the dataset covers the initial dynamics of the coronavirus outbreak on Twitter.

Documenting the Now

- DocNow is a tool and a community developed around supporting the ethical collection, use, and preservation of social media content.
- Twarc: command line tool and Python library for archiving Twitter JSON data.

Tools to collect Twitter Data

- Twarc (command line tools)
- Tweepy (Python)
- rtweet (R)
- Social Feed Manager (web applications)
- NVIVO (presentation on November 18, 2020)
- Netlytic
- TAGS for Google Sheets
- RapidMiner (presented in the DH Toolbox Series in 2019)
- Social Media Research Toolkit developed by Social Media Lab at Ted Rogers School of Management, Ryerson University

Hands-on Activity: Collect and visualize your tweets using twarc

Hands-on Activity: Collect and your tweets using TAGS

Large dataset of tweets

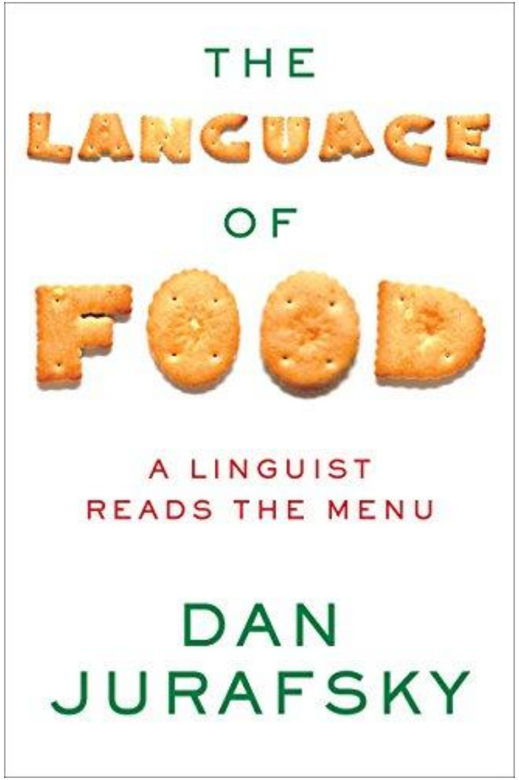
- DocNow Catalog
- Internet Archive
- Kaggle
- Zenodo
- Your institutional data repository including Dataverse

Computational Resources

- High performance computing (PHC) via Compute Canada
- Dr. Jarno van der Kolk, Senior Scientific Computing Specialist
  - Email: jvanderk@uottawa.ca
- Storage: https://it.uottawa.ca/professors/storage-solutions

Fifth poll: Do you trust datasets from internet?

#3. The Language of Food



THE LANGUAGE OF FOOD

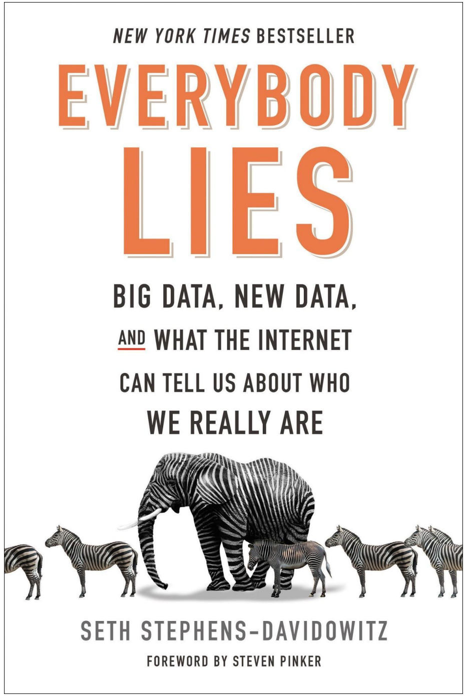
A LINGUIST READS THE MENU

DAN JURAFSKY

Narrative framing of consumer sentiment in online restaurant reviews (Jurafsky et al., 2014)

- Dataset: online reviews from 2006 to 2011 in seven cities from Yelp.com (Boston, Chicago, Los Angeles, New York, Philadelphia, San Francisco, and Washington D.C.)
  - 887,658 reviews from 6,548 restaurants

Everybody lies: big data, new data, and what the Internet can tell us about who we really are


NEW YORK TIMES BESTSELLER
EVERYBODY LIES
BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE
SETH STEPHENS-DAVIDOWITZ
FOREWORD BY STEVEN PINKER

Stephens-Davidowitz, 2017

- People’s search for information is, in itself, information.
- Rejected by five academic journals due to peer-reviewers doubt on the data source (INTERNET!)

Detecting influenza epidemics using search engine query data

Ginsberg et al. (2009)

- Health-seeking behavior in the form of online web search queries, submitted by millions of users around the world each day.
- A method of analyzing large numbers of Google search queries to track influenza-like illness in a population

Association of the COVID-19 pandemic with internet search volumes: A Google Trends analysis

Effenberger et al. (2020)

- Data from Google Trends
- Tool: ggplot2 package (R)
- Public interest in COVID-19 correlates with the number of newly reported COVID-19 cases, with highest interest observed on average 11.5 days before the peak of newly reported COVID-19 cases

Google Trends

- Google Trends in Canada
- Access to a largely unfiltered sample of actual search requests made to Google
- Anonymized, categorized, and aggregated
- Real-time data and non-realtime data

Amazon product reviews for recommender systems

- Amazon product data by McAuley, UCSD
- Data contains product reviews and metadata from Amazon, including 142.8 million reviews from May 1996 to July 2014, and there is the updated version (2018).
- Sentiment analysis

Sixth poll: How many women are represented in Wikipedia?

#4. Public Recognition of Female Scholars in Wikipedia

Female scholars need to achieve more for equal public recognition, Schellekens et al. (2019)

- Discrimination in public recognition of scientific achievement gauged by inclusion in Wikipedia at any level of success.
- Dataset: Google Scholars, Wikipedia, and genderalize.io

What successful people had in common

Everybody lies: big data, new data, and what the Internet can tell us about who we really are, Stephens-Davidowitz, 2017

- Datasets: Wikipedia articles
- Findings:
  - Geography played an outsized role - the person was born near a large college town
  - Growing up near immigrants played a significant role

How to download Wikipedia and Wikidata?

- Access a dump of all Wikipedia (English version and French version)
  - A dump refers to a periodic snapshot of a database
- Access a dump of all Wikidata
- Using Python, OpenRefine, R or other tools to parse data - check out DH Toolbox in 2019 on Web scraping for DH

#5. Archived web content

Me Too movement (#MeToo)

- Movement against sexual abuse and sexual harrassment
- #metoo Web Archives Collection by Schlesinger Library
- #MeToo and the Women's Rights Movement in China Web Archive by Ivy Plus Libraries Confederation

How to create a WARC file?

- Webrecorder
- Conifer
- Chrome extension: WARCreate
- Command line: $wget

Where to find a WARC file?

- Dalhousie University Repository
- Simon Fraser University Repository
- University of Toronto Repository
- University of Victoria Repository
- University of Winnipeg Repository

Tools to analyze archived web content

- Archives Unleashed Toolkit
- Archives Unleashed Cloud
- Archives Unleashed Notebooks
- pylibwarc
- Ukwa-gsheets-utils
- warc (R)

#6. Nineteenth-Century Newspaper Analytics

Illustrated Image Analytics, Fyfe, Ge, and Aguayo (2016-2017)

- Develop techniques in computer vision and image processing for large-scale interpretation of historical illustrations
- Data source: British newspapers including the Graphic, The Illustrated Police News, and the Penny Illustrated Paper among the many nineteenth-century periodicals

Visual analysis

Tools

- scikit-image: image processing in Python
- Image Processing Toolbox in MATLAB
- TensorFlow to identify objects in images and identify similar images

Where to find images?

- Library Digital Collections Repositories :)
- Wikicommons (Download tools)
- The MNIST database of handwritten digits
- Caltech 101 categories
- The Street View House Numbers (SVHN) Dataset
- Kaggle

#7. YouTube as source of information on 2019 novel coronavirus outbreak: a cross sectional study of English and Mandarin content

- Khatri et al. (2020)
- Datasets: YouTube videos
- In the current outbreak, YouTube viewership remains high for both English and Mandarin content.
- The medical content of these videos is suboptimal and needs to be improved.
- Misleading videos in Mandarin were more popular and had higher viewership than useful videos

YouTube Videos Dataset

- YouTube-8M Segments Dataset
- Trending YouTube Video Statistics (Kaggle)
    - Daily trending YouTube videos including for the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India
- Statistics and social network of YouTube videos (2007 and 2008)

What I can help you...

- Explore unconventional datasets for your research
- Identify tools that can be applied to your research
- Guide through how to use tools for your research

Email: yooyoung.lee@uottawa.ca

Q&A

Thank You!