UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

Opening the Data of the U.S. Court Systems

COURT LISTENER

Brian W. Carver

February 14, 2013

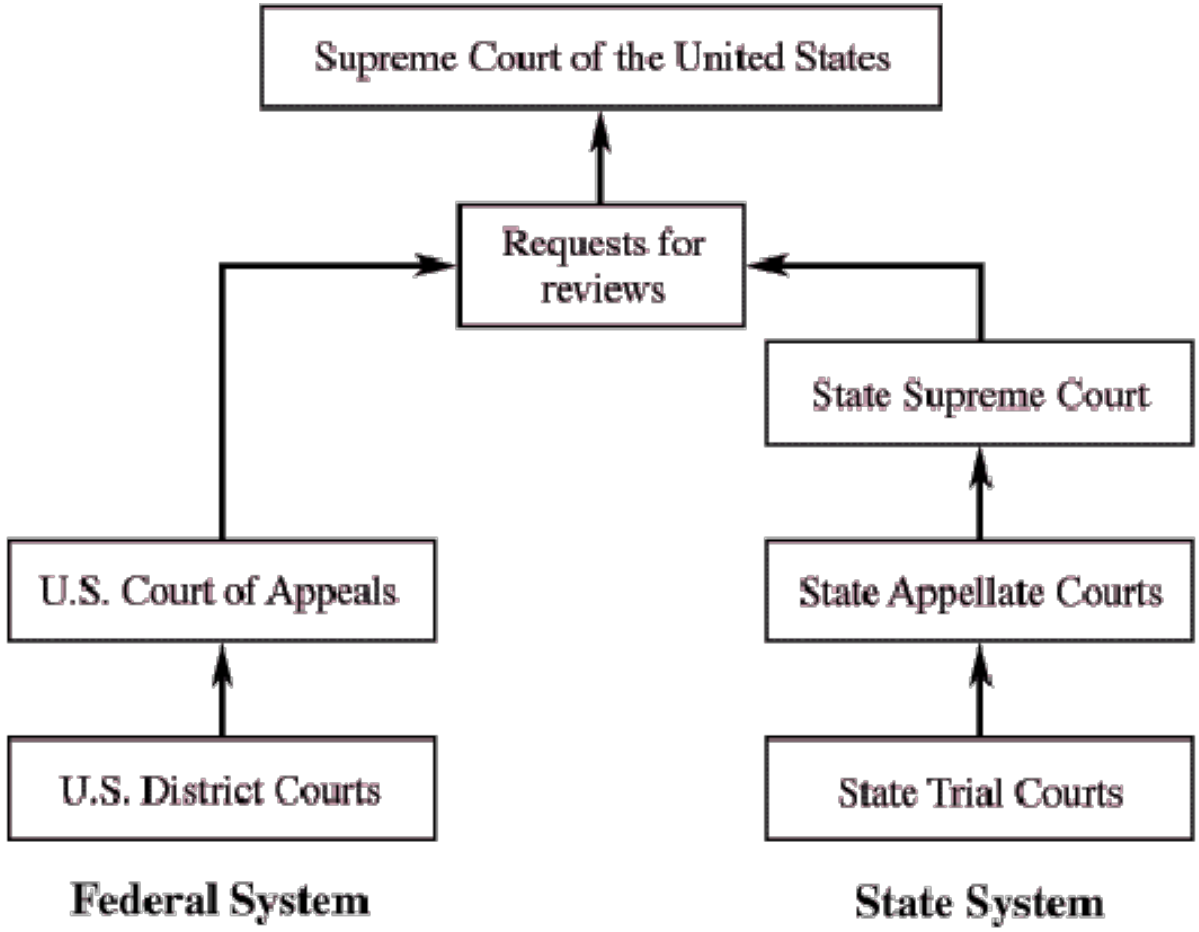
http://courtlistener.com

http://bitbucket.org/mlissner/

Outline

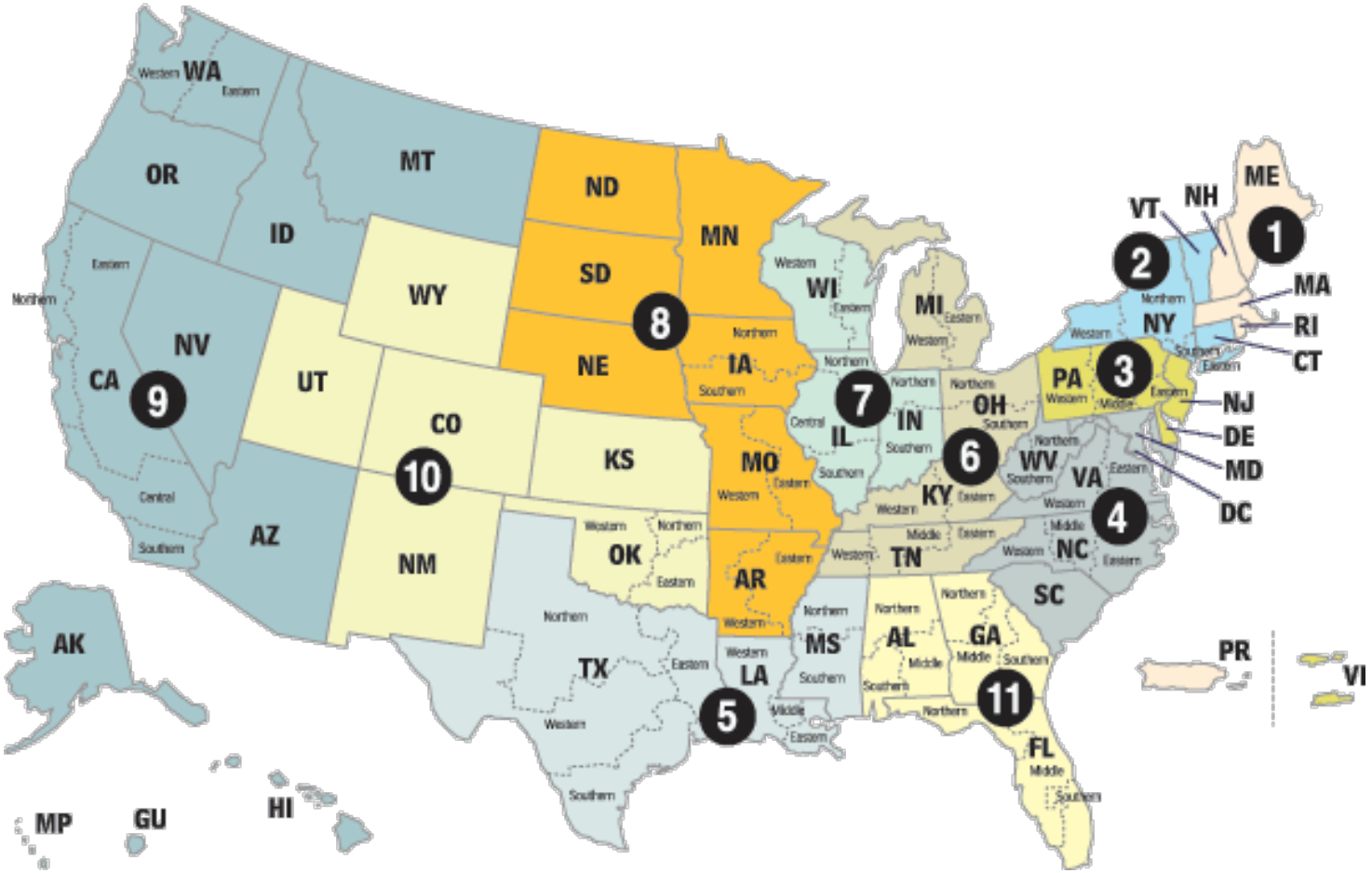
- Background
  - The U.S. Court Systems
  - My original problem
- Michael Lissner's Final Project (2010)
- Rowyn McDonald and Karen Rustad (2012)
- Current Projects:
  - Juriscraper
  - CITRIS Research
- Future Projects, Endgame, and Your Questions

Background: U.S. Court Systems



UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

Background: U.S. Court Systems



UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

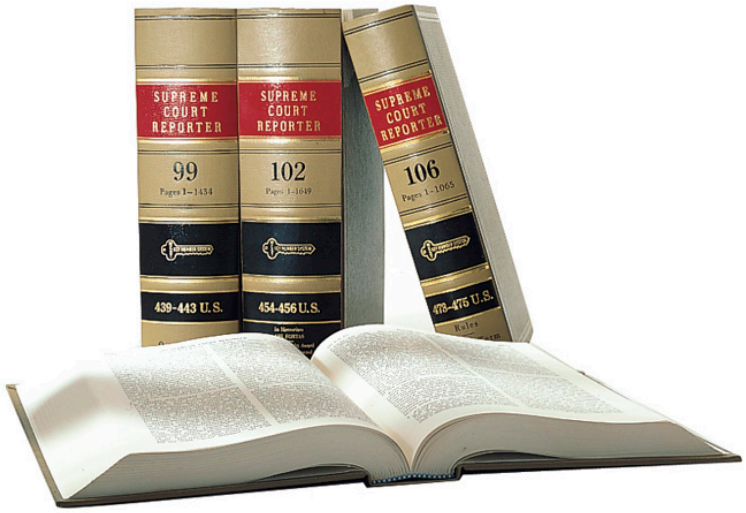
Background: U.S. Court Systems

Background: U.S. Court Systems

- Nearly all courts have websites.
- Nearly all courts put case-related resources on their websites, such as:
  - court documents (including briefs or motions filed in the case, deposition or hearing transcripts, and judicial opinions), and
  - a handful now also publish audio or video of oral arguments.
- The most common resource courts put online (and arguably the most important) are the judicial opinions. These are the orders and decisions of the courts. They *are* the law. I also think of them as data.

Background: U.S. Court Systems

For a couple hundred years, the primary means of distributing judicial opinions has been for the courts to provide them to private publishers (e.g., West) who print bound volumes known as "Reporters."



Background: My Original Problem

- I like to keep track of new opinions that are published in areas that I'm interested in.
- I found that the services that provide this sort of daily alerting were:
  - Not timely
  - Both incomplete and over-inclusive
  - Not sufficiently customizable
  - Extremely expensive
- Academic researchers aren't the only ones with this need. Journalists and practicing attorneys need this.

Michael Lissner's Final Project (2010)

- Michael created CourtListener to solve this daily alerting problem for the 13 federal circuit courts and the Supreme Court of the United States.
- By himself.
- CourtListener delivers email alerts of new opinions:
  - Every weekday at 5:30 p.m. PST (or weekly/monthly by request).
  - Like Google Alerts, it works using user-created search queries.
  - It permits queries using complex Boolean operations.
  - It is free (as in freedom and as in beer).

Michael Lissner's Final Project (2010)

- CourtListener is built from free software:
  - Ubuntu
  - Apache
  - PostgreSQL (formerly we used MySQL)
  - Python
  - Django
  - Solr (formerly we used Sphinx)
  - Celery & RabbitMQ
  - Tesseract
  - Mercurial

Michael Lissner's Final Project (2010)

- CourtListener's Document Data Model:
  - A unique ID we generate for citation purposes
  - Source: court website, resource.org, manual input.
  - SHA1
  - Date filed by court
  - Court
  - URL where retrieved
  - File location in our filesystem
  - Plain text and HTML of the doc, and HTML + citations
  - Cases cited
  - Precedential status
  - Block indexing?
  - Original block date
  - OCR?

Michael Lissner's Final Project (2010)

- CourtListener's Citation Data Model:
  - Our local URL : united-states-v-roy-gray
  - Full name of case : United States v. Roy Gray
  - Docket Number : 94-1298
  - Westlaw citation : 63 F.3d 57
  - LexisNexis citation :
  - Neutral citation :

Michael Lissner's Final Project (2010)

- CourtListener and Persistent Identifiers
  - The West Citation is *the* means of referring to a case, but
    1) it doesn't exist until West publishes the case in hard copy,
    2) West provides no API for learning of these citations,
    3) West doesn't publish *every* case, and
    4) sometimes they aren't even unique: (very short cases).
  - The Docket Number
    1) exists at the time of publication,
    2) but is the same for all documents in a given case,
    3) is often re-used by the SAME court (civil, criminal, bankr.), &
    4) is re-used by other courts.
  - Neutral Citations exist at the time of publication and are unique, but are only used by a handful of jurisdictions.

Michael Lissner's Final Project (2010)

- CourtListener's Distinguishing Features
  - Best daily alert service for federal appellate courts at any price.
  - Entirely free software; can be continued if I'm hit by a bus.
  - Provides bulk downloads to enable research or duplication.
  - Is the only free law site in the U.S. providing a citator.

Rowyn McDonald & Karen Rustad: 2012

- Building a Free, Open Source, Legal Citator
  - CourtListener's documents were largely plain text documents with no hyperlinks.
  - This project identified the citations within the documents, determined if the cited document was in the CourtListener database, and then created HTML versions of all the documents containing inter-connecting hyperlinks.
  - A 'citator' creates a citation index. Given a reference of a legal decision, a citator allows the researcher to find newer documents which cite the original document and thus to reconstruct the judicial history of cases and statutes.

Rowyn McDonald & Karen Rustad: 2012

- Identifying citations is easy. Right?
  - Ingle v. Landis Tool Co. (C.C.A.) 272 F. 464
  - Ingle v. Landis Tool Co., 272 Fed. 464 (3d Cir. 1921)
  - Ingle v. Landis Tool Co., 272 F. 464 (3d Cir. 1921)
  - Ingle v. Landis Tool Co., 272 Fed. 464 (3rd Circ., 1921)
  - Ingle v. Landis Tool Co., 272 Fed. 464 (C. C. A. 3d, 1921)
  - Ingle v. Landis Tool Co., 272 _____ LINE BREAK _____ F. 464 (3d Cir. 1921)

Rowyn McDonald & Karen Rustad: 2012

- They tokenized the text, used the Python Natural Language Toolkit (NLTK), and identified the basic well-behaved citations: 499 U.S. 340
- Look for a parenthetical to the right: (1991).
- Look for a 'v.' or 'In re' to the left:
  … v. Rural Telephone Service Co., 499 U.S. 340 (1991)
- But there are 'v. United States' and 'v. Holder'
- Take the next word to the left of the 'v.'
- Search for the citation in the CL database, using reverse matching to select among several candidates.

Rowyn McDonald & Karen Rustad: 2012

- Over 750,000 documents, they created 4.2 million citation links, performing 10.9 million queries.
- They matched over 80% of the citations found (more for Supreme Court cases, because our corpus is complete for that court)
- This enabled answers to new questions:
  - What U.S. court case is the most cited in the CL database?
  - Can we use the citation information to improve our relevance search results (a la PageRank?)
  - What would it take to create "Depth of treatment" scores?

Current Projects

- Juriscraper
  - We have spun off our court scrapers into a separate project called Juriscraper.
  - We found there were many people that had an interest in collecting court opinions but did not need the entire CourtListener machinery.
  - CourtListener is AGPLv3 while Juriscraper is BSD licensed.
  - You can help! We need scrapers for the 50 states.
  - Due to the modular design of our scraping system, the code required to scrape a given page will often be less than 30 lines.

Current Projects

- CITRIS Research
  - Proactive Legal Information Filtering and Retrieval (plifr)
  - Working with Yi Zhang of UC Santa Cruz Computer Science and her Ph.D. students.
  - Modifying the CourtListener interface to track documents that searchers visit or repeatedly visit and using that information to proactively recommend other documents, while showing the user what features of that document caused it to be recommended. (AKA torturing Berkeley law students).

Future Projects?

- Data mining the CourtListener corpus:
  - Clean up our database to include 'authoring judge' of each opinion and then train for authorial style, using results to identify the authors of 'per curiam' opinions.
  - Create an algorithm that can win Fantasy SCOTUS.
  - What can the cases cited by the parties and the amici in Supreme Court briefs, or the courts below, tell us about the cases that will be cited in the resulting Supreme Court opinion(s)? (Can this help us predict case outcomes?)
  - What does Justice Scalia's "word cloud" look like? How does it compare to Justice Kagan's?

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

Brian's Summary / Endgame

- Having all the opinions that were issued TODAY in any court that is part of the U.S. judicial system (and being able to consistently retrieve those on a daily basis) is only a few months of steady work away.
- Having all the opinions EVER ISSUED from any court that is part of the U.S. judicial system is a very big challenge, but we want to find a way to get there.
- We'd like to improve the metadata in our database so that it would enable even more interesting academic research on the bulk data downloads.

Your Questions

- What is the purpose of the case fingerprint?
- Why do some cases not have a downloadable version?
- When you crawl the web to collect court cases how do you conclude whether a document is the same document you already have?
- What type of text mining has been completed with the data dumps. Any interested projects to link us to?
- Are there plans to make the search function more user friendly by incorporating NLP elements? "Aetna v. Jeppesen"

Your Questions

- What is the more common use case for those searching? Finding ongoing cases or finding sources for older cases? [Noticed the default sort is Date as opposed to Relevance] Was sorting by Date/Recency by default a design choice?
- The categories in the left are referring the courts. Did you evaluate a different category system, like privacy, free speech, or others?

More Questions?

http://courtlistener.com

http://bitbucket.org/mlissner/