



Module 3 - Session 1 - Data Analysis

Working effectively with data

CivicDataLab

2021/09/08 (updated: 2021-09-09)

Things we will **not** cover

1. This is not an introduction to Statistics
2. No, we'll not learn AI and ML
3. We will not learn about using the Microsoft Excel Analytics ToolPak addin
4. We will not use any additional analytical tools like Stata or SPSS to analyse data

What we will cover instead

1. In what ways can analysis of legal data help us
2. Common challenges with any data analysis project
3. Asking the right questions
4. Data Analysis use cases
 1. Court Performance
 2. Case level analytics
 3. Comparing entities - Creating Indices
 4. Text Based Analysis
 5. Criminal Justice Data Explorers - Commons Project
 6. FBI Crime Data Explorer
 7. Human Rights Data Analysis Group (HRDAG)
5. How can we analyse the data in an un-biased way
6. The methodology matrix - What to use when

Analysing legal data can help us in

1. Understanding how the judiciary functions
2. Creating more accountable institutions
3. Analysing the structural capacity of justice system across the country
4. Designing evidence-based judicial reforms
5. Designing better court/case management practices
6. Predict/Simulate the impact of laws on different stakeholders
7. Re-engineering judicial processes

and the list continues

But there are hurdles to cross

1. Availability of granular datasets
2. Bias in data collection and analysis
3. Beneficiaries not in alignment with the proposed outcomes
4. Setting up a multidisciplinary team that:
 1. Understands the context
 2. Has the right set of skills to analyse data
 3. Can publish with purpose
5. Ethical use of datasets
6. Sustainability
7. Periodic data updates

and the list continues as well ..

It starts with asking the right questions

If I had only one hour to save the world, I would spend fifty-five minutes defining the questions, and only five minutes finding the answers.

Albert Einstein

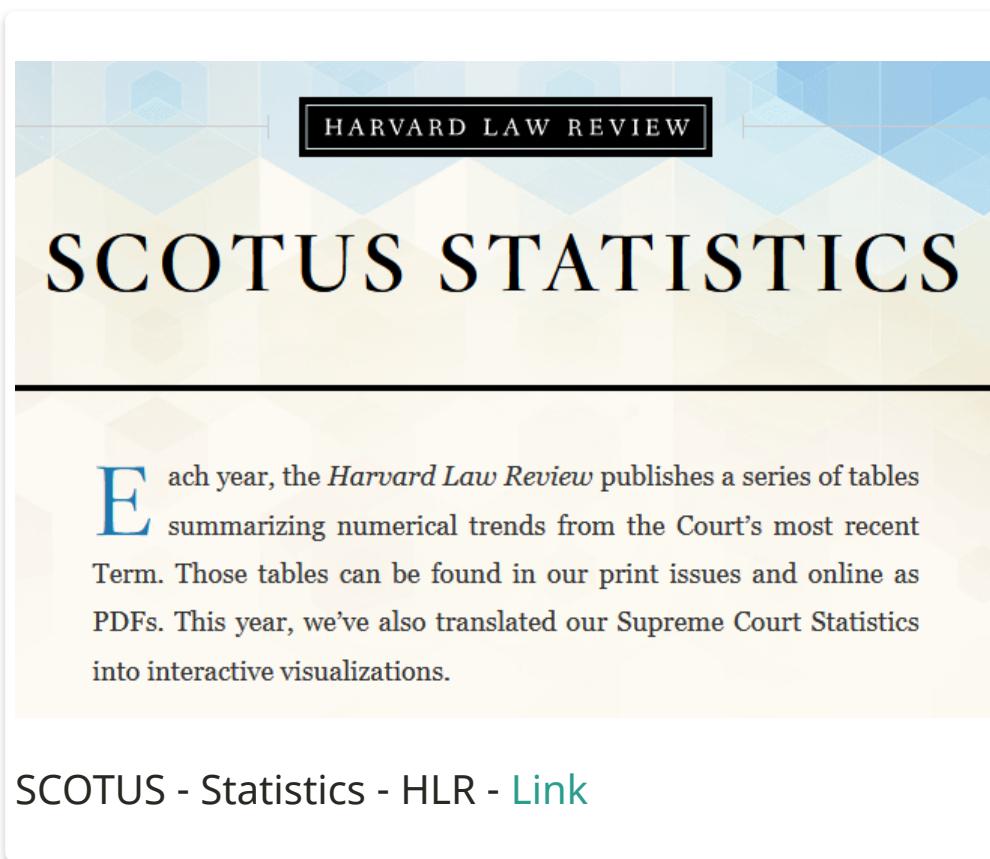
The 100 Questions Project - Initiatives

1. Migration
2. Gender
3. Governance

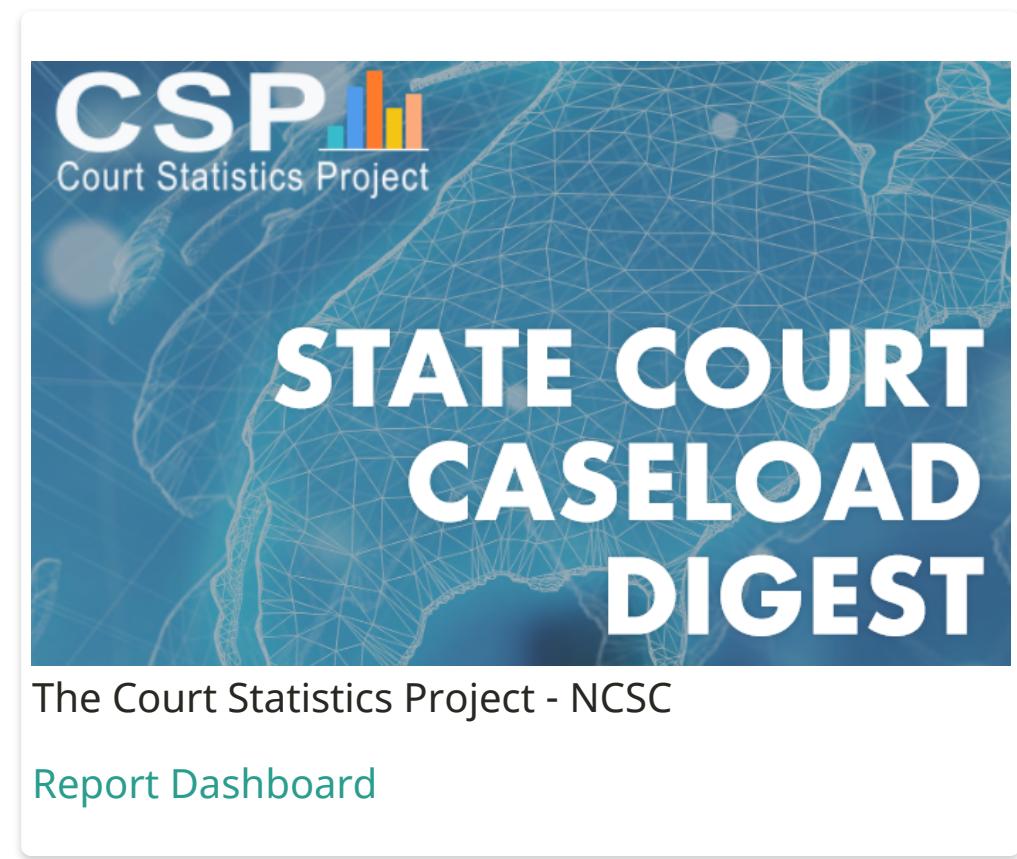
Explore other initiatives at <https://the100questions.org/>

Data Analysis - Use-Cases

Analysing Court Performance



A graphic for the Harvard Law Review's SCOTUS Statistics. It features a light blue background with a geometric pattern of overlapping triangles. In the top left corner, there is a black rectangular box containing the text "HARVARD LAW REVIEW" in white capital letters. Below this, the words "SCOTUS STATISTICS" are written in large, bold, black capital letters. A horizontal line separates this section from the text below. The main text area begins with the letter "E" in blue, followed by the sentence: "ach year, the *Harvard Law Review* publishes a series of tables summarizing numerical trends from the Court's most recent Term. Those tables can be found in our print issues and online as PDFs. This year, we've also translated our Supreme Court Statistics into interactive visualizations." At the bottom of the graphic, the text "SCOTUS - Statistics - HLR - [Link](#)" is displayed in black.



A graphic for the Court Statistics Project's State Court Caseload Digest. The background is a dark blue with a light blue wireframe map of the United States. In the top left corner, the letters "CSP" are written in large white capital letters, with three vertical bars of increasing height to its right. Below "CSP", the words "Court Statistics Project" are written in smaller white capital letters. In the center, the words "STATE COURT CASELOAD DIGEST" are written in large, bold, white capital letters. At the bottom of the graphic, the text "The Court Statistics Project - NCSC" and "Report Dashboard" are displayed in white.

Supreme Court - Statistics Pack



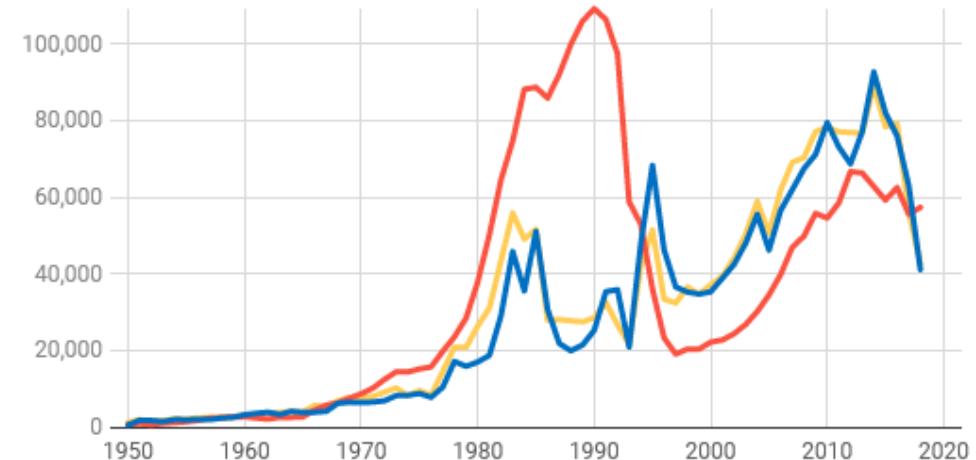
Opinions by Sitting.....	6
Time Between Oral Arg. And Opinion	7
Workload.....	8-9
Opinions Authored by Each Justice....	10
Total Opinion Authorship.....	11
Total Opinions Over Time.....	12
Majority Opinion Authorship.....	13
Majority Opinion Distribution.....	14
Frequency in the Majority.....	15
Strength of the Majority.....	16
Merits Cases by Vote Split.....	17
Unanimity.....	18-19
5-4 Cases.....	20-22
Justice Agreement.....	23-26
Oral Argument.....	27-30

United States (*SCOTUS blog*)

Pendency in the Supreme Court of India

1950-2018: no. of cases instituted, disposed, pending

— Institution — Disposal — Pending



Institution = new cases; Disposal = disposed cases; Pending = pending cases

India (*by CLPR*)

Other Court level analysis:

- A Quantitative Analysis of the Indian Supreme Court Workload - Nick Robinson
- An Empirical Assessment Of The Collegium's Impact On Composition Of The Indian Supreme Court

Analysing Case Laws

Table 2: Types of government related cases in our data

Sr.No.	Civil writ petitions	Govt. as Petitioner	Govt. as respondent	Total (% of government cases)
1.	Writ petitions	10,476	1,06,179	1,16,655 (84.69)
2.	Miscellaneous petitions	3,493	5,168	8,661 (6.28)
3.	Land Acquisition related cases	3,180	3,663	6,843 (4.9)
4.	Arbitration petitions and applications	221	2,837	3,058 (2.54)
5.	Civil suits	696	1,818	2,514 (1.8)
Total		18,066	1,19,668	1,37,734 (100)

Litigation in public contracts - NIPFP

The CaseLaw Access Project by Harvard Law School



CAP includes all official, book-published United States case law — every volume designated as an official report of decisions by a court within the United States.

Case Law Access Project - [Link](#)

Our data

360 years of United States caselaw

State and Federal Totals

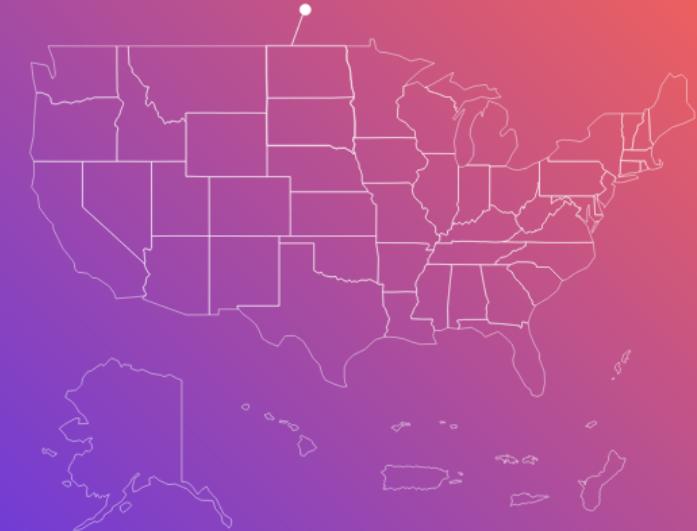
67,08,785
Unique cases

3,55,49,133
Pages scanned

Federal Totals

17,60,899
Unique cases

1,00,75,368
Pages scanned



The CaseLaw Access Project - Use Cases



Caselaw Visualization Blog
The Caselaw Visualization Blog by João Marinotti (HLS, 2020) shares data visualizations about U.S. caselaw using CAP data.

Creating a Case Recommendation System Using Gensim's Doc2Vec
Case Recommendation System by Minna Fingerhood.

Do Elected and Appointed Judges Write Opinions Differently?
Michael Nelson and Steven Morgan at Pennsylvania State University summarize their research using CAP data.

Tracking semantic change in US Court opinions and Congress speeches
Abdul Abdulrahim at the University of Oxford shares his experience using CAP data to see what language can tell us about ideological changes surrounding reproductive rights between the U.S. Supreme Court and Congress over time.

Telling Stories with CAP Data: The Prolific Mr. Cartwright
John Bowers presents "my experience working with the Caselaw Access Project's publicly available Illinois dataset as evidence for a more optimistic narrative – namely that applying quantitative techniques to corpuses primarily associated with the qualitative disciplines can help us to uncover and relate stories which might otherwise go unnoticed."

- **Telling Stories with CAP Data: The Prolific Mr. Cartwright**

Comparing entities - Creating Indices



IJR Ranking



	Overall rank (out of 18)		IJR 2020 pillar ranks (out of 18)				IJR 2020 scores (out of 10)						
	IJR 2020		IJR 2019		Police	Prisons	Judiciary	Legal aid	Overall	Police	Prisons	Judiciary	Legal aid
	Rank	Change	Rank	Change	13	4	5	1	5.77	4.62	5.45	6.40	6.90
Maharashtra	1	1	↔		13	4	5	1	5.77	4.62	5.45	6.40	6.90
Tamil Nadu	2	3	↑		5	6	1	11	5.73	5.40	5.28	7.22	5.22
Telangana	3	11	↑		10	2	6	6	5.64	4.89	5.69	6.14	5.93
Punjab	4	4	↔		12	13	2	3	5.41	4.72	4.20	6.78	6.35
Kerala	5	2	↓		14	5	3	7	5.36	3.89	5.45	6.68	5.84
Gujarat	6	8	↑		8	10	8	9	5.17	5.14	4.63	5.56	5.39
Chhattisgarh	7	10	↑		2	11	4	15	5.13	5.63	4.58	6.56	4.11
Jharkhand	8	16	↑		6	15	9	4	5.12	5.36	3.90	5.30	6.18
Haryana	9	5	↓		9	16	7	5	4.94	4.99	3.39	5.82	6.07
Rajasthan	10	14	↑		16	1	10	13	4.93	3.75	6.32	5.27	4.71
Odisha	11	7	↓		3	9	15	8	4.90	5.59	4.67	3.91	5.64
Andhra Pradesh	12	13	↑		4	7	14	14	4.81	5.43	5.25	4.28	4.37
Bihar	13	17	↑		11	3	18	2	4.65	4.73	5.67	2.66	6.57
Karnataka	14	6	↓		1	14	12	16	4.59	5.71	4.02	4.75	4.08
Uttarakhand	15	15	↔		7	18	13	10	4.48	5.30	3.14	4.61	5.25
Madhya Pradesh	16	9	↓		18	8	11	12	4.39	3.17	4.78	5.05	4.86
West Bengal	17	12	↓		17	12	16	17	3.89	3.75	4.58	3.69	3.63
Uttar Pradesh	18	18	↔		15	17	17	18	3.15	3.80	3.24	3.16	2.54

India Justice Report

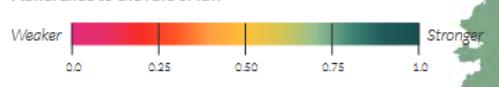
World Justice Project - Rule of Law Index



Rule of Law Around the World

The table below shows the overall scores and rankings of the WJP Rule of Law Index 2020 by country rank. Scores range from 0 to 1, with 1 indicating the strongest adherence to the rule of law.

Adherence to the rule of law



	Overall Score*	Global Rank
Denmark	0.90	1
Norway	0.89	2
Finland	0.87	3
Sweden	0.86	4
Netherlands	0.84	5
Germany	0.84	6
New Zealand	0.83	7
Austria	0.82	8
Canada	0.81	9
Estonia	0.81	10
Australia	0.80	11
Singapore	0.79	12
United Kingdom	0.79	13

	Overall Score*	Global Rank
Antigua and Barbuda	0.63	34
Namibia	0.63	35
St. Lucia	0.62	36
Rwanda	0.62	37
Mauritius	0.61	38
Croatia	0.61	39
Greece	0.61	40
The Bahamas	0.61	41
Georgia	0.60	42
Botswana	0.60	43
Grenada	0.59	44
South Africa	0.59	45
Dominica	0.58	46

Covering 128 countries and jurisdictions, the Index relies on national surveys of more than 130,000 households and 4,000 legal practitioners and experts to measure how the rule of law is experienced and perceived worldwide.

[World Justice Project - Report](#)

Text based analysis - Natural Language Processing



BLACKSTONE

Open source natural language processing for Legal Texts

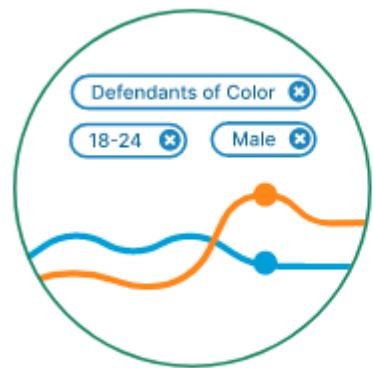
Blackstone is a spaCy model and library for processing long-form, unstructured legal text. Blackstone is an experimental research project from the Incorporated Council of Law Reporting for England and Wales' research lab, ICLR&D.

Open source natural language processing for Legal Texts

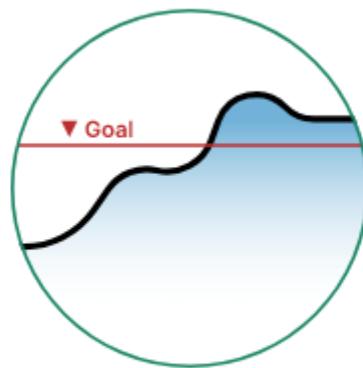
Criminal Justice Data - The Commons Project



Commons is a co-created space for the community, police, prosecutors, and courts to make criminal justice data transparent & shared goals public.



Share a goal, meet that goal



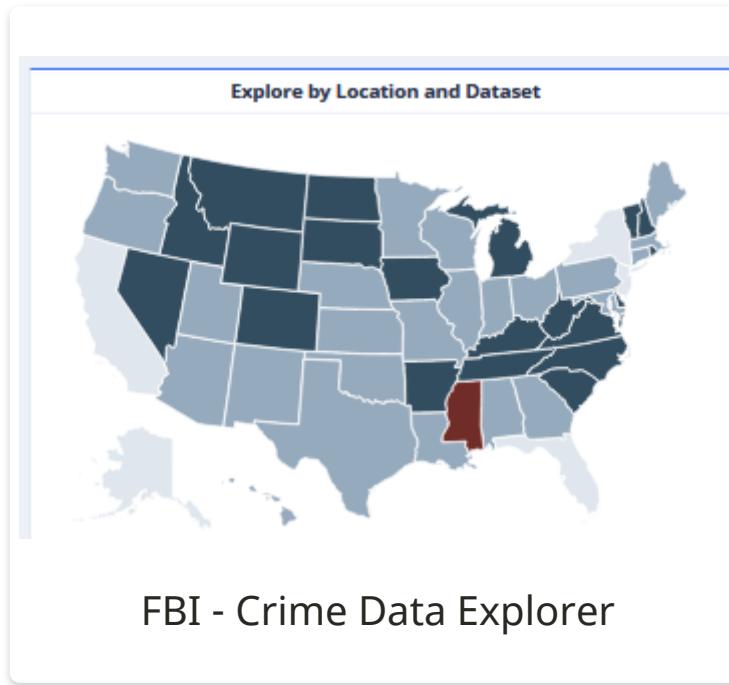
Look at trends month over month



Share findings directly with policymakers and media

Explore here - <https://measuresforjustice.org/commons/yoloda/>

Incident based Crime Data Explorer - FBI



[Explore here](#)

Mapping Police Violence - Alternate Datasets



Human Rights Data Analysis group (HRDAG)

The Human Rights Data Analysis Group is a non-profit, non-partisan organization that applies rigorous science to the analysis of human rights violations around the world. Their work has been used by truth commissions, international criminal tribunals, and non-governmental human rights organizations across countries.

Projects

- [Violent Deaths and Enforced Disappearances During the Counterinsurgency in Punjab](#)
- [Estimating the Human Toll in Syria](#)
- [How many people disappeared on 17-19 May 2009 in Sri Lanka?](#)
- [A dataset on Human Rights violations in Guatemala during the period 1960-1996](#)

Other Resources

- [Talks and Discussions](#)

Biases in data analysis

Confirmation bias is the tendency to seek out or interpret data to confirm beliefs you already hold. It does this to the exclusion of contrary evidence.

Biases in data analysis

Confirmation bias is the tendency to seek out or interpret data to confirm beliefs you already hold. It does this to the exclusion of contrary evidence.

Selection Bias is when the group chosen to be analyzed is not representative of the population you are trying to draw conclusions about.

Biases in data analysis

Confirmation bias is the tendency to seek out or interpret data to confirm beliefs you already hold. It does this to the exclusion of contrary evidence.

Selection Bias is when the group chosen to be analyzed is not representative of the population you are trying to draw conclusions about.

Survivorship bias is the tendency to draw conclusions based on things that have survived, some selection process, and to ignore things that did not survive.

A story from WW2



How to avoid bias

1. Better documentation of data collection practices, research questions, analysis methods, assumptions, stakeholders, funders, etc.
2. Conducting reproducible research experiments
3. Peer reviewed research
4. Regular audits
5. Correct framing of the research[1] questions[2]
6. Curating a representative sample - Exploratory data analysis will help in assessing if a sample is representative enough.

[1]What is a research question [2]How to frame your research questions equitably.

Data Analysis Methodologies

What to use when

It can be hard to know which methodology to use when designing a data project. The most important thing for the success of your selection (and equity) is to make sure that your chosen methodology matches the kinds of research questions you have.

Read more here

WE ALL COUNT
project for equity in data science

We All Count Methodology Matrix

	WHO	WHAT	WHERE	WHEN	WHY
3	Select an option in each column -->	-	-	-	-
4	Bayesian Latent Variables	A, B, E	A, B	A, B	A, B, C
5	Bayesian Net Analysis	A, B, C, D	A, B	A, B	A, B, C
6	Collaborative Yarning	A, B, C, D, E	A, B, C	A, B	A, B, C, E, F
7	Convergent Parallel Design	B, D, E	A, B	A	B, C
8	Difference in Difference	B, D, E	A	A	B
9	Directed Acyclic Graph Analysis (DAG)	A, B, C, D	A, B	A	C, E, F
0	Focus Groups	A, B, C, D, E	A, B, C	A, B	A, B, C, F
1	Key Informant Interviews	A, B, C, D, E	A, B, C	A, B	A, B, C, F
2	Latent Class Analysis	A, C	A, D	A	A, B, C

Methodology Matrix - [Link](#)

Resources

- [Statistics for lawyers](#)
- [OpenIntro - Statistics](#)
- A course by SDGAcademy on [Measuring Sustainable Development](#)
- [The Seductions of Quantification - Measuring Human Rights, Gender Violence, and Sex Trafficking](#)
- [The Summer Institutes in Computational Social Science - SICSS](#)
- [Bit By Bit - Social Research in the Digital Age](#)

Queries and Feedback