**Ben Klemens**

2031 11th Street NW
DC 20001
(213) 926-6336
ben@klemens.org

See also:
https://ben.klemens.org/
https://en.wikipedia.org/wiki/Ben_Klemens

*Employment*

---

※ Senior Statistician, Economist — 2015–present
U.S. Treasury Office of Tax Analysis

- Wrote papers on the intersection of intellectual property and tax policy.
- Provided extensive analysis of income and tax changes due to migration.
- Designed systems for projecting federal tax revenues from corporations, and for use in what-if scenarios for alternative tax policies.

※ Principal Researcher — 2009–2015
U.S. Census Bureau Statistical Research Division

- Led a group in design and implementation of *Tea*, a package used for certain aspects of American Community Survey and 2010 Census processing.
- Oversaw four employees, advising them on their research and helping them to implement Census projects.

※ Senior Statistician — 2006–2009
National Institute of Mental Health

- Found genetic segments statistically indicated to be implicated in bipolar disorder.

※ Nonresident Fellow 2003–2009
Brookings Institution

- Wrote a book, law review article, and several op-eds (Wall Street Journal, Washington Post, ...) on the state of intellectual property law.
- Was a technical advisor on development of the Brookings Weak State Index.
- Made radio and TV appearances (NPR, Voice of America, ...) discussing tech policy.
- Organized two conferences on intellectual property issues.

※ Executive Director, End Software Patents project 2007–2008
Free Software Foundation

- Wrote public information and made public appearances on the threats of the expanded scope of patent law and patent trolls.
- Organized a small staff of campaigners.
- Spoke with media about the campaign and patent policy.
- Wrote an amicus brief for a Federal Circuit case, with help from legal staff. FSF reused the amicus brief as a brief to the Supreme Court in *Bilski v Kappos*.

※ Contractor 2004–2005
World Bank

- Designed, implemented, and calibrated an agent-based model of labor migration through the Europe and Central Asia region.

※ Postdoctoral Fellow 2004–2005
Johns Hopkins University Department of Economics

※ Teaching Assistant 2000–2003
Caltech Humanities & Social Sciences Dept, incl. economics, law, psychology, anthropology

※ Research Assistant 2000
to Michael Chwe of the NYU Dept of Politics

※ Risk Control Analyst 1998–99
ABN AMRO, Inc

- Oversaw and reported on all of the desks of a regional brokerage firm, including OTC, Forex, pink sheet stocks, options trading, government bonds, and commodity futures
- Built a reporting system to gather information from all desks, calculate risk metrics, and report daily to management.

※ Research Assistant 1992–95
to Dali Yang of the U of Chicago Political Science Dept

*Education*

---

※ PhD in Social Science (Microeconomics and Game Theory) 2003
Caltech

※ Participation in the Empirical Implications of Theoretical Models Summer Institute 2002
Harvard University Center for Basic Research in the Social Sciences

※ MS in Social Science 2001
Caltech

※ NASD Series 7 certification 1998

※ BA in Economics 1996
University of Chicago. Included one year at London School of Economics studying mathematics and labor economics; dean's list.

*Government*

---

※ Policy Committee member 2020
Biden presidential campaign

※ Chair 2013–2015
Transportation Committee, DC ANC1B

*Technical papers*

---

※ An Analysis of U.S. Domestic Migration via Subset-stable Measures of Administrative Data June 2021
Journal of Computational Social Science

※ Valuing patents and trademarks in complex production chains March 2020
Journal of the Knowledge Economy

※ Intellectual Property Boxes and the Paradox of Price Discrimination May 2017
Center for Economic Priorities Working Paper

※ Unemployment Insurance and worker mobility, by Ryan Nunn, Laura Kawano and Ben Klemens February 2018
Urban/Brookings Tax Policy Center report

※ Estimating local poverty measures using satellite images: A pilot application to Central America, by Ben Klemens, Andrea Coppola, and Max Shron — June 2015
World Bank Policy Research Paper #7329

※ A Useful Algebraic System of Statistical Models — May 2013
Census Bureau Research Report Series (Statistics)

※ A Peer-based Model of Fat-tailed Outcomes — April 2013
Arxiv

※ Mutual Information as a Measure of Intercoder Agreement — October 2012
Journal of Official Statistics

※ Tea for Survey Processing — September 2012
United Nations Economic Commission for Europe Conference of European Statisticians

※ Finding Optimal Agent-based Models — September 2007
Center on Social and Economic Dynamics Working Paper #49

※ A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder — May 2007
*Molecular Psychiatry*(13:2), pp 197-207, by AE Baum, N Akula, M Cabanero, I Cardona, W Corona, *B Klemens*, TG Schulze, S Cichon, M Rietschel, MM Nathen, A Georgi, J Schumacher, M Schwarz, R Abou Jamra, S Hofels, P Propping, J Satagopan, NIMH Genetics Initiative Bipolar Disorder Consortium, SD Detera-Wadleigh, J Hardy, and FJ McMahon.

※ Social Influences and Smoking Behavior — February 2006
Brookings Report to the American Legacy Foundation

※ An Efficient Network Generation Method: Interpersonal Networks and the Distribution of Links — December 2005

※ Dissertation: Information Aggregation — May 2003
California Institute of Technology

*Books*

---

※ 21st Century C — September 2012
O'Reilly Media

※ Modeling with Data — October 2008
Princeton University Press

※ Math You Can't Use: patents, copyright, and software — November 2005
Brookings Institution Press

*Popular press articles*

| | |
|---|---|
| ※ Keeping science reproducible in a world of custom code and data<br>Ars Technica | November 2021 |
| ※ An Inclusive, Cyberpunk Future Is In the Cards, by Ben Klemens and Liz Landau<br>WIRED | February 2021 |
| ※ The Heisenberg Uncertainty Principle of Social Science Modeling<br>Scientific American | 7 July 2020 |
| ※ Software patents poised to make a comeback under new patent office rules<br>Ars Technica | 10 January 2019 |
| ※ The Beauty Contest: How Cities are Shaped by What We Think Others Think, with Erica C Barnett<br>Strong Towns | November 2018 |
| ※ 5 Q's for Census Statistics Expert Ben Klemens<br>Interview with The Center for Data Innovation | October 2014 |
| ※ Copyright Law and the Progress of Science and the Useful Arts by Alina Ng<br>Science and Public Policy | May 2012 |
| ※ U.S. expanding the law - domestic and foreign - to benefit corporations<br>San Francisco Chronicle, 17 February 2008 p E | February 2008 |
| ※ The Rise of the Information Processing patent<br>Boston University Journal of Science and Technology Law, 14:1, pp 1–37 | January 2008 |
| ※ U.S. Patent Imperialism Hurts American Interests<br>Washington Post, 25 August 2006 | August 2006 |
| ※ Net neutrality fosters competition between technologies<br>Op-ed distributed by Scripps-Howard News Service, 17 August 2006 | August 2006 |
| ※ The Supreme Court's Patent Trilogy: An Analysis<br>Brookings Institution | May 2006 |
| ※ The Gravity of the U.S. Patent Swindle<br>Wall Street Journal, 25 March 2006, p A9 | March 2006 |
| ※ New Legal Code<br>IEEE Spectrum, pp 60-62 | August 2005 |
| ※ Software Patents Don't Compute<br>IEEE Spectrum, pp 56-59 | July 2005 |

※ The Computer-shaped Hole in the Patent Reform Act — July 2005
Brookings Institution

※ Shadowing Bush — November 2004
Brookings Institution

※ Social Norms and Voter Turnout — January 2004
Brookings Institution

*Live events*

---

※ Workshop Co-organizer — July 2019
Agent-Based Models for Exploring Public Policy Planning, Lorentz Center at University of Leiden

※ A Simulation of Nonresponse and Imputation — August 2013
2013–14 Program on Computational Methods in Social Sciences (CMSS)

※ Designing a cross-paradigm modeling framework — May 2013
Neyman Seminar at the University of California, Berkeley, Department of Statistics, 8 May 2013; invited talk at Stanford University Department of Statistics 14 May 2013.

※ Appearance in *Patent Absurdity*, a documentary — 2010

※ Co-producer and speaker for Brookings Conference, "The Limits of Abstract Patents in an Intangible Economy" — January 2009

※ Speaker, NISS exploration workshop on agent-based modeling — November 2008

※ Interview by David Levine on Stanford Radio's Hearsay Culture — 10 August 2007

※ Panelist: Tech Policy Week — 5 May 2007

※ DNA Pooling for GWAS with Illumina Infinium Assays — October 2006
Poster by Amber E. Baum, Nirmala Akula, *Ben Klemens*, Imer Cardona, Winston Corona, Andrew Singleton, John Hardy, Sevilla Detera-Wadleigh, Francis J. McMahon

※ The Kojo Nnamdi Show: "The Legal Battle over Software" (National Public Radio discussion) — January 2006

※ Producer and moderator for "Software and Law: Is Regulation Fostering or Inhibiting Innovation?" December 2005

※ Taught graduate class on networks and information Feb-May 2005
Johns Hopkins Econ dept

*Major technical projects*

---

At the U.S. Treasury 2015–2020

※ Tax revenue

- PURPOSE: Revamp the system for determining who gets audited by the IRS. Given a tax return, what filers are most likely to have large adjustments? How can the system be designed to minimize algorithmic bias?
- DATA SET: U.S. individual tax returns, about 100 million per year.
- METHODS: Python's SciKit stack. Details are proprietary.
- OUTPUT: An online learning system to select returns for audit. Testing to begin in 2021.

※ Migration

- PURPOSE: Observe correlates to within-US migration. Holding all else constant, do middle-income households move more than higher- and lower-income? How do lifetime earnings for college graduates who move after school differ from those who don't?
- DATA SET: The U.S. formal economy, 1.7 billion observations.
- METHODS: Imported the data into the IRS Statistics of Income division's Hadoop server, reduced the data via SQL queries via Hive, via Spark, via Python. Due to time and resource limitations, did the final statistical calculations in C. Paper was in LaTeX using a macro language to insert the statistics into the paper. A POSIX shell script runs the cross-server pipeline from read-in to paper output.
- OUTPUT: white paper, at http://dx.doi.org/10.2139/ssrn.3197362

※ Corporate tax calculator

- Purpose: How much corporate tax revenue will the U.S. take in over the next decade? If a change to tax law is made, what is the effect on tax revenue, and who wins and loses?
- Data set: Sample of corporate tax returns, about 50,000 per year.
- Methods: The Federal Reserve, Treasury, and Congressional Budget Office generates a set of macroeconomic forecasts used across the US Government. I use these to evolve the balance sheets of corporations in the sample year-to-year. With another staff member, developed a system for tracking and projecting corporate assets over time. I developed a domain-specific language to encode tax forms, which compiles to plain C, and is used to calculate the final revenue figures. Graphical output via Bokeh via Python, with a few HTML+javascript pages. Data is stored in an SQLite database.
- Output: Non-public reports to White House offices. A first draft of the tax calculator portion, for individual tax returns instead of corporate, is available at https://b-k.github.io/1040.js/ and and https://github.com/b-k/py1040

At the World Bank 2015

※ Estimating local poverty measures using satellite images: A pilot application to Central America

- Purpose: Can satellite imagery be used to augment traditional surveys to improve poverty measurements?
- Data set: MODIS night lights data set, 250GB.
- Methods: I reduced the raw data to MySQL tables, joined to survey microdata, then ran linear regressions familiar to World Bank employees.
- Output: World Bank working paper: https://tinyurl.com/lights-and-surveys

At the US Census Bureau 2010–2015

※ Apophenia

- Purpose: Can we provide the familiar tools of data analysis such as data frames, simple regression functions, tools for Bayesian graphical models, in plain C?
- Methods: The system predates SciKit, Pandas, and the mature R ecosystem (circa 2010 R had severe data size limitations), but its data frames and model objects look and behave very much like those used in standard toolkits today. Included rudimentary R and Python front-ends.
- Output: See https://apophenia.info, or download the Debian package: https://packages.qa.debian.org/a/apophenia.html

※ Tea for survey processing

- Purpose: American Community Survey (the ACS, about two million per year), portions of the U.S. Census (the population of a handful of cities).
- Data set: Can we simplify the tools used to process surveys and the U.S. Census? How can we simplify the means of collating, editing, and imputing missing data?
- Methods: an R package, based partly on Apophenia.
- Output: Used for the ACS and small portions of the 2010 Census. Available at https://github.com/rodri363/tea

At the National Institute of Mental Health 2007

※ Bipolar disorder Genome-wide association study

- Purpose: What genetic markers are associated with bipolar disorder?
- Data set: 550,000 genetic markers for about 1,000 individuals.
- Methods: A team of psychologists interviewed subjects to class them into bipolar and not-bipolar groups, blood was drawn and gene-sequenced. I set up a server with a MySQL database and parsed the data. The statistical analysis was a sequence of 550,000 t-tests adjusted to the details of genetic data handling.
- Output: Journal article, at https://www.nature.com/articles/4002012