Plenty of data

Germán Reyes

This version: January 2020

This document contains a curated list of over 500 datasets organized by topic, compiled from various sources (mainly from Jeremy Singer-Vine's phenomenal Data is Plural newsletter). I stopped updating the database in 2020.

Some notes on other data sources:

- Data is plural is a weekly newsletter with interesting databases. Most of the datasets in this document are verbatim from this newsletter.

- NBER has a list of datasets grouped by categories.

- J-PAL North America has cataloged a number of key US data sets. The catalog documents procedures on how to access data based on information provided by the originating agencies.

- Google has a search engine for datasets.

- data.word is a social network for data-oriented people.

- Scientific data is a journal devoted to publishing papers that compile novel data.

- DevEconData has datasets for development economists.

- Amazon web services public datasets.

- Some researchers have a section on their websites where they publish the data that they collected and used in their papers (e.g., Nathan Nunn or Matthew Gentzkow).

- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are readily available to any researcher. This is a great source of data.

Table of Contents

1 Microdata (household surveys, census data, etc.)

SEDLAC: Harmonized databases of socio-economic and labor market statistics, covering more than 300 surveys from 17 countries in Latin American and the Caribbean. Basically, this database compiles all the household surveys from Latin American and makes them comparable across each other.

LABLAC: Same as SEDLAC, but with labor force surveys (which are usually collected at a higher frequency, i.e., quarterly). **Panel Data**: There is panel data (quarterly and yearly) for a few countries: Argentina, Brazil, Mexico, Peru.

Enterprise Surveys: The Enterprise Survey are firm-level survey of a representative sample of an economy's private sector. The surveys cover a broad range of business environment topics including access to finance, corruption, infrastructure, crime, competition, and performance measures.

IPUMS: Project dedicated to collecting and distributing census data from around the world.

Nielsen datasets: Researchers can integrate the consumer panel and retail scanner datasets to enable additional types of research. By integrating these two datasets, researchers can determine not only the items purchased by panelists, but also the availability, prices, and promotions associated with other products that were on the shelf at the same time. Although it varies by year, 45–70 percent of the stores in the consumer panel data can be matched to scanner data.

Consumer Expenditure Survey. The survey collects data on spending, income, and a handful of characteristics about U.S. consumers.

Family money. The Panel Study of Income Dynamics is "the longest running longitudinal household survey in the world," according to its University of Michigan overseers. The study, which began in 1968, has interviewed more than 70,000 people, including four generations of some families. You can access the data for free, but you first need to register for an account and agree to a set of guidelines.

Interned Japanese Americans. The Densho Digital Repository is an archive of oral histories, photographs, newspaper clippings, and other primary sources relating to the internment of Japanese Americans during World War II. Among the materials: several datasets listing people sent to the internment camps, based on official government records. The largest dataset contains more than 100,000 entries and includes details such as each internee's "relocation" site, arrival date, hometown, birth year, time spent in Japan, marital status, religion, educational degrees, occupation, and military service. The National Archives hosts the raw data, as well as its documentation.

German companies. The Open Knowledge Foundation Deutschland and OpenCorporates have partnered to make Germany's official business register available to download in bulk. The dataset contains basic information about more than 5 million German companies, and more than

4 million associated officers. Although the dataset's landing page is written in German, its documentation is available in English.

Luxembourg Income Study Database harmonizes datasets with income, wealth, employment and demographic information for many countries.

1.1 Opinion poll surveys

Latinobarómetro: Opinion poll survey conducted in 18 LAC countries since 1990's, interviewing about individual socioeconomic background, and preferences towards political and social issues (1200 individuals per country per year).

Latin American Public Opinion Project (LAPOP): The Americas Barometer is an opinion poll survey that covers 28 nations including all of North, Central, and South America, and the Caribbean (1200 individuals per country per year).

World Values Survey: Is a global research project that explores people's values and beliefs, how they change over time and what social and political impact they have. Includes many LAC countries.

General Social Survey: is a sociological survey created and regularly collected since 1972 by the National Opinion Research Center at the University of Chicago. It is funded by the National Science Foundation. The GSS collects information and keeps a historical record of the concerns, experiences, attitudes, and practices of residents of the United States.

Affluence and Influence: Economic Inequality and Political Power in America. Dataset of over 2,000 survey questions about specific proposed policy changes (1964-2006). The policy preferences explored in the dataset range from raising the minimum wage to restrictions on abortion to sending troops to Bosnia. The degree of support for each of these proposed changes expressed by individuals in different income groups was then related to the subsequent policy outcomes.

Pan-African surveys. Afrobarometer "is a pan-African, non-partisan research network that conducts public attitude surveys on democracy, governance, economic conditions, and related issues in more than 35 countries in Africa." You can download data from the first six rounds of surveys, conducted between 1999 and 2015. You can also read the detailed questionnaires and explore the results online. **Note:** To download the data, you'll need to create a (free) account on the website.

Half a century of opinions. The University of North Carolina's Louis Harris Data Centerserves as "the national depository for publicly available survey data collected by Louis Harris and Associates, Inc." The online depository contains more than 1,000 Harris polls, some from

as early 1958. In total, they include "160,000 questions asked of more than 1,200,000 respondents."

Pew Research has many free data sets, including their "Global Attitudes Project" archive.

1.2 USA restricted-use data available through the census bureau (link)

- **Economic data (link)**

Census of Auxiliary Establishments (AUX)

Census of Construction Industries (CCN)

Census of Finance, Insurance, and Real Estate (CFI)

Census of Island Areas – Puerto Rico (CIAPR)

Census of Manufactures (CMF)

Census of Mining (CMI)

Census of Retail Trade (CRT)

Census of Services (CSR)

Census of Transportation, Communications, and Utilities (CUT)

Census of Wholesale Trade (CWH)

Compustat–SSEL Bridge (CSB)

Integrated Longitudinal Business Database (ILBD)

Longitudinal Business Database (LBD)

Ownership Change Database (OCD)

Report of Organization Survey

Standard Statistical Establishment List/Business Register (SSEL)

Annual Survey of Manufactures (ASM)

Current Industrial Reports (CIR)

Manufacturing Energy Consumption Survey (MECS)

Medical Expenditure Panel Survey – Insurance Component (MEPS–IC)

National Employer Survey (NES)

Quarterly Survey of Plant Capacity Utilization (QPC) /

Survey of Plant Capacity Utilization (PCU)

Survey of Manufacturing Technology (SMT)

Survey of Pollution Abatement Costs and Expenditures (PACE)

Annual Capital Expenditures Survey (ACES)

Annual Retail Trade Survey (ARTS)

Annual Survey of Entrepreneurs (ASE)

Annual Wholesale Trade Survey (AWTS)

Business Expenditures Survey (BES)

Business Research & Development and Innovation Survey (BRDIS)

Exporter Database (EDB)

Kauffman Firm Survey (KFS)

Manufacturers' Shipments, Inventories, and Orders (M3)

Manufacturers' Unfilled Orders Survey (M3UFO)

Monthly Retail Trade Survey (MRTS)

Monthly Wholesale Trade Survey (MWTS)

Quarterly Financial Report (QFR)

Quarterly Services Survey (QSS)

Service Annual Survey (SAS)

Survey of Business Owners (SBO)

Survey of Industrial Research and Development (SIRD)

Commodity Flow Survey (CFS)

Foreign Trade Data – Exports (EXP)

Foreign Trade Data – Imports (IMP)

Longitudinal Firm Trade Transactions Database (LFTTD)

- **Demographic data (link)**

American Community Survey (ACS)

American Housing Survey (AHS)

Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement (March)

Current Population Survey (CPS) Computer and Internet Use Supplement

Current Population Survey (CPS) Fertility Supplement (June)

Current Population Survey (CPS) Food Security Supplement (December)

Current Population Survey (CPS) School Enrollment Supplement

Current Population Survey (CPS) Tobacco Use Supplement

Current Population Survey (CPS) Unbanked and Underbanked Supplement (January & June)

Current Population Survey (CPS) Volunteer Supplement (September)

Current Population Survey (CPS) Voting and Registration Supplement (November)

Decennial Census

Harmonized Decennial Census: IPUMS-format census data

National Crime Victimization Survey (NCVS)

National Crime Victimization Survey (NCVS) Identity Theft Supplement

National Crime Victimization Survey (NCVS) School Crime Supplement

National Crime Victimization Survey (NCVS) Police Public Contact Survey

National Crime Victimization Survey (NCVS) Supplemental Victimization Survey

National Immunization Survey Evaluation Study (NISES)

National Longitudinal Mortality Study (NLMS)

National Longitudinal Survey (NLS)

National Survey of College Graduates (NSCG)

Rental Housing Finance Survey (RHFS)

Survey of Income and Program Participation (SIPP) Panels

Longitudinal Employer-Household Dynamics Data (link)

Business Register Bridge (BRB)

Employer Characteristics Files (ECF)

Employment History Files (EHF)

Geocoded Address List (GAL)

Individual Characteristics Files (ICF)

Quarterly Workforce Indicators (QWI)

Unit-to-Worker (U2W)

Measurement Initiative's (IMI) UMETRICS Data. The inaugural IMI UMETRICS data release includes information on awards, wage payments from awards to university research employees, vendor purchases, subcontracts, and the unit performing the funded research for 19 universities, released in May 2017. University employee data are linked to internal Census Bureau data products, such as the Decennial Census, American Communities Survey, Longitudinal Employee-Employer Household Dynamics database (LEHD), integrated Longitudinal Business Database, and other demographic information. Vendors paid by research grants are linked to the Business Register, Longitudinal Business Database, and the LEHD, providing a comprehensive description of the businesses paid by research grants.

2 Macro data

md4stata: Macro Data 4 Stata addresses homogenize several commonly used macroeconomic datasets and converts them into Stata.

International Monetary Fund's International Financial Statistics (IFS). The main source for macro data.

The Conference Board's Total Economy Database includes data on GDP, Population, Employment, Hours worked and Labor productivity since 1950 for 123 countries.

OECD's Population and Labor Force Statistics includes labor data broken down by gender, age groups and sector.

Economic crisis: Dates for Banking Crises, Currency Crashes, Sovereign Domestic or External Default (or Restructuring), Inflation Crises, and Stock Market Crashes (Varieties).

Maddison Project: The Maddison Project Database provides information on comparative economic growth and income levels over the very long run. The 2018 version of this database covers 169 countries and the period up to 2016.

Three centuries of UK macroeconomic data. The Bank of England publishes a spreadsheet of historical economic data going back, in some cases, to the late 1600s. The country's GDP in 1700 was £11.7 billion in 2013 prices. That's about 1/157th the size of the UK's GDP in 2015. And in November 1694, monthly short-term interest rates were roughly 6%.

Global food prices. The UN World Food Programme's vulnerability analysis group collects and publishes food price data for more than 1,000 towns and cities in more than 70 countries. The dataset, which goes back more than a decade, covers basic staples, such as wheat, rice, milk, oil, and more. It's updated monthly and feeds into (among other things) the UNWFP's price-spike indicators.

Federal Reserve forecasts. Before each meeting of the Federal Open Market Committee, the Federal Reserve's research staff prepares a set of economic projections known as the Greenbook. Those forecasts are

kept secret for five years, and then released to the public. The Philadelphia Fed's archive of public Greenbook data dates back to 1966, and contains both PDFs and structured data files.

Industrial sector data. Aswath Damodaran — a professor of finance at the NYU's business school — maintains a trove of data on per-sector financials, including effective tax rates, return on equity, and working capital ratios by industry. For most datasets, Damodaran publishes both current and historical versions.

American manufacturing. The Census Bureau's Annual Survey of Manufacturers provides state-by-state and industry-by-industry statistics for America's manufacturing sector. Metrics include the number of employees, annual payroll, "value added," beginning-of-year inventory, and many more.

Some notes on how to do some growth accounting work (like computing labor's share) are available at Tim Kehoe's website.

2.1 Trade

The Economic Complexity Index: (ECI) and the Product Complexity Index (PCI) are, respectively, measures of the relative knowledge intensity of an economy or a product. ECI measures the knowledge intensity of an economy by considering the knowledge intensity of the products it exports. PCI measures the knowledge intensity of a product by considering the knowledge intensity of its exporters.

Detailed U.S. import/export data. The newly-free dataset covers more than 17,000 commodities, including a category for "magic tricks, practical joke articles; parts and accessories."

Global trade dynamics. The Atlas of Economic Complexity has collected decades of import/export data from the United Nations Comtrade database, and then applied "a unique method to clean the data to account for inconsistent reporting practices." You can download the raw data, learn more about the cleaning process in the FAQ, explore current and historical trade flows, and browse the Atlas's rankings of countries by "economic complexity."

Interstate commodity flows. The federally funded Freight Analysis Framework "integrates data from a variety of sources to create a comprehensive picture of freight movement among states and major metropolitan areas by all modes of transportation." For each year between 2012 and 2015, the database "provides estimates for tonnage (in thousand tons) and value (in million dollars) by regions of origin and destination, commodity type, and mode.

United Nation's COMTRADE has data on country to country trade by different levels of aggregation.

Christopher Magee's Bilateral Trade Dataset joint with information usually used for gravity equations.

2.2 Banking and Financial sector

Global Findex Database: provides in-depth data on how individuals save, borrow, make payments, and manage risks. It is the world's most comprehensive database on financial inclusion that consistently measures people's use of financial services across countries and over time

Global financial history. The Jordà-Schularick-Taylor Macrohistory Database claims to be "the most extensive long-run macro-financial dataset to date." It contains dozens of variables — GDP per capita, long-term interest rates, and the timing of systemic financial crises, for example — for 17 "advanced economies".

Historical credit ratings. The SEC requires Moody's, Standard & Poor's, and other "nationally recognized statistical rating organizations" to report their rating assignments and changes (e.g., upgrades, downgrades, withdrawals) going back to 2010. The agencies publish the reports as XBRL-formatted files, and update them monthly. But "because most researchers are unfamiliar with XBRL and cannot easily locate the history files, this valuable resource has seen limited use," according to the Center for Municipal Finance's RatingsHistory.info, which now provides the reports as easier-to-use CSVs.

Financial well-being. The Consumer Financial Protection Bureau's National Financial Well-Being Survey collected more than 6,000 responses to the agency's 10-question Financial Well-Being Scale, plus additional demographic and financial information. The survey results, which were collected in late 2016, come with a detailed methodology and data dictionary.

Financial consumer complaints. The Consumer Financial Protection Bureau's consumer complaint database can be searched online, accessed via an API, and downloaded in bulk. The 915,000+ complaints the Bureau has received have been categorized into 18 financial product groups (e.g., mortgages, debt collection, student loans, cryptocurrency) and more than 160 kinds of issues (e.g., billing disputes, communication tactics, privacy). The agency says they "don't verify all the facts alleged in these complaints," but that they "take steps to confirm a commercial relationship between the consumer and the company."

Historical Bitcoin prices. You can download daily and hourly data for the index and its components.

Insurance premiums and payouts. ProPublica and Consumer Reports published an analysis of car insurance costs in four states, finding that "some major insurers charge minority neighborhoods as much as 30 percent more than other areas with similar accident costs." The reporters also published a detailed methodology and dataset supporting their findings. The dataset contains company-by-company insurance premiums for a (hypothetical) college-educated, excellent-credit, accident-free 30-year-old woman in each of 6,261 ZIP codes in the four states — California, Texas, Missouri, and Illinois. The dataset also includes several years of average (per-car) insurance payouts for each ZIP code, which the reporters obtained from state insurance commissioners

IMF's Systemic Banking Crises dataset includes information on systemic banking, currency and sovereign debt crises for 1970-2010, including some information on policy responses.

World Bank's Financial Development and Structure Dataset of indicators of financial development and structure across countries and over time (since 1960).

Peer-to-peer loans. The Lending Club, which matches borrowers with investors, publishes a dataset of all loans issued through its platform since 2007. The dataset's many fields include each loan's amount, term, interest rate, grade, status, and purpose (as a category, and often also a fuller description), as well as the borrower's employer, home ownership status, and annual income. You can also download all declined loans, i.e., those "that did not meet Lending Club's credit underwriting policy."

Financial statements. The SEC's Office of Structured Disclosure publishes data extracted from corporations' public financial statements. That dataset contains the numbers listed in each company's primary financial statements — balance sheets, cash flows, et cetera. An even more detailed version of the dataset includes plain-text notes from the filings, plus numbers from a broader array of forms. Both datasets are updated quarterly and go back to 2009.

2.3 Prices

Gas prices at each pump in Argentina during 2017

Daily price indices: Daily price indices, monthly, and annual inflation rates for Argentina and the US. Monthly data with annual inflation rates for Argentina, Brazil, China, Germany, Japan, South Africa, UK, US, 3 US sectors, and global aggregates (including Eurozone).

3 Education

PISA: The Programme for International Student Assessment (PISA) is a worldwide study intended to evaluate educational systems by measuring 15-year-old school pupils' scholastic performance on mathematics, science, and reading. It was first performed in 2000 and then repeated every three years.

Barro-Lee: Provides educational attainment data for 146 countries in 5-year intervals from 1950 to 2010. It also provides information about the distribution of educational attainment of the adult population over age 15 and over age 25 by sex at seven levels of schooling.

SAT, ACT, and AP scores. The California Department of Education publishes aggregate scores on these high-school tests for each county, district, and school going back to the late 1990s.

School testing. The Department of Education's EDFacts data tracks public grade schools' participation and proficiency rates on standardized math and reading/language exams. The files provide data on all students who took the tests, broken down by race/ethnicity, sex, disability status, homelessness, and more.

STEM surveys. The IPUMS Higher Ed portal provides data from three "leading surveys for studying the science and engineering (STEM) workforce in the United States." The surveys currently cover 1993 through 2013 and include questions about educational choices, demographics, employment outcomes, and more. Requires a free account.

Educational attainment. Researchers at the Vienna-based Wittgenstein Centre for Demography and Global Human Capital have developed a dataset of historical and projected education levels for 171 countries. For five-year age groups in each country, the project estimates the percentage of people in each of several categories of educational attainment — no education, primary education, secondary education, post-secondary education, and a few gradations in between. The dataset is available to browse and download via the Wittgenstein Centre Data Explorer – look for "Educational Attainment Distribution" in the "indicators" dropdown.

The StudentLife Study. Back in 2013, four dozen Dartmouth College students agreed to let a custom smartphone app surveil them for the StudentLife Study. During the 10 weeks of the spring academic term, the app collected data on the students' physical activity, GPS coordinates, eating schedule, sleep habits, phone usage, and more. The study combined all that information with a slew of other data, including the students' class deadlines, academic performance, and their responses to surveys about stress, depression, personality, and sleep quality. The study's public (and anonymized) dataset clocks in at 53 gigabytes. Related: "Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities," a recently-released academic paper that uses the StudentLife dataset.

What do you do with a PhD in science? The National Science Foundation's Survey of Doctorate Recipients "is a longitudinal biennial survey conducted since 1973 that provides demographic and career history information about individuals with a research doctoral degree in a science, engineering, or health (SEH) field from a U.S. academic institution." You can download aggregated data and detailed survey responses going back to 1993.

Which public colleges are the best value? Compare the 100 top values in public colleges and universities, or create your own custom lists of individual colleges or schools from a particular state.

A decade-plus of Seattle library checkouts. Seattle Public Library released a dataset tracking the total number of checkouts for each title by year and month from April 2005 to December 2016 (so far). The dataset isn't limited to physical books; it also includes e-books,

magazines, CDs, DVDs, and more. Individual library checkouts. The Seattle Public Library publishes a dataset of every checkout of every physical item (e.g., paperback books and DVDs, but not e-books) since April 2005. It currently contains more than 90 million rows.

Digest of Education Statistics has a compilation of data on all levels of education.

Delta Cost Project analyses trends in money received and spent in higher education.

Foreign gifts to U.S. universities. The Department of Education requires U.S. universities to report all major gifts from (and contracts with) foreign entities. The agency's database of these gifts and contracts currently covers 2012 to mid-2018, and includes 18,000+ entries from more than 150 schools

3.1 Innovation

R&D spending. The UNESCO Institute for Statistics' data on national research and development budgets contains estimates of personnel and total spending by field, funding source, and more.

Crowdfunding. A Lithuania-based web-scraping company has been collecting data on Kickstarter projects and Indiegogo campaigns every month. The datasets include (among other things) each project's number of backers, amount pledged, and category.

Innovation nation. A team of researchers published HistPat, a database containing county-of-residence data for 2.8 million U.S. patents granted between 1836 and 1975. The database covers approximately 83% of all patents granted to U.S. residents during that time, according to the authors.

Nobel Prizes. The prestigious Scandinavian awards have an API. The official documentation explains it succinctly: "The data is free to use and contains information about who has been awarded the Nobel Prize, when, in what prize category and the motivation, as well as basic information about the Nobel Laureates such as birth data and the affiliation at the time of the award. The data is regularly updated as the information on Nobelprize.org is updated, including at the time of announcements of new Laureates."

Patents and trademarks. The U.S. Patent and Trademark Office publishes a huge amount of bulk data, including detailed XML files that contain information about millions of patent/trademark applications, assignments, trials, and appeals. The agency also publishes a collection of "research datasets", which distill those bulk XML files into easier-to-use tabular data.

Patent geography. Researchers at two Swiss universities have created a dataset of inventors' and applicants' locations listed in 18.8 million patents filed between 1980 and 2014. The locations, which span 46 countries, are specified both by their geographic coordinates as well as their administrative areas (e.g. city, state, country).

Millions of scientists, and their migrations. ORCID is a nonprofit organization that provides unique identifiers for researchers — mostly scientists so far — to make it easier to distinguish between them. It has issued more than 3 million IDs so far, and provides annual bulk downloads of all researchers' public profiles. In many cases, the researchers have supplied their education and employment histories. That enabled Science magazine to analyze the migrations of more than 110,000 researchers who've listed multiple countries in these public CVs.

Science grants. The National Science Foundation publishes data on all of the grants the agency has awarded since the 1970s (and some earlier ones, too). Each grant is represented as an XML file, which contains information about the project, the awardee, and the NSF division that awarded the grant.

4 Health

The kids are alright. Every two years since 1991, the CDC has conducted the Youth Risk Behavior Survey, which asks high school students questions about drug use, sex, eating habits, and more. The results are available at the national, state, and district level.

Obesity over time. An international network of researchers who study noncommunicable diseases estimates the annual prevalence of obesity and diabetes for approximately 200 countries and territories around the world. The data currently covers 1975–2014 and is based, on 2,000+ surveys, according to the group.

Measuring up. A group of public health researchers have estimated the average height of adults in 200 countries over the course of a century. Their calculations are based on a re-analysis of 1,472 previous studies, which collectively measured nearly 19 million participants. The resulting dataset contains annual height estimates for both men and women born each year between 1896 and 1996. During that time, South Korean women's average height increased by approximately 8 inches, the largest gain of any group. These days, the Netherlands boasts the tallest men, and Latvia the tallest women.

Health habits. The CDC calls its Behavioral Risk Factor Surveillance System "the largest continuously conducted health survey system in the world." Every year, the survey asks more than 400,000 American adults about a range of health-related topics, from tobacco to seatbelt use, from alcohol consumption to arthritis, from HIV testing to immunizations. Annual datasets from 1984–2015 are currently available.

What kills us. The CDC's Underlying Cause of Death database provides county-level mortality statistics based on death certificates of U.S. residents for each year from 1999 to 2014. The tool lets you group the data by geography, demographics, place of death (e.g., inpatient hospital, hospice, home, etc.), and other variables. In 2014, for example, about 40,000 residents died of pancreatic cancer — with the highest rates coming in America's most-rural counties (~15.6 deaths per 100,000 residents) and the lowest rates in country's most-urban

counties (~11.3 per 100,000). The CDC's "compressed mortality" datasets contain slightly less detail, but go all the way back to 1968.

Local health metrics. The CDC's 500 Cities Project provides "city and census tract-level data, obtained using small area estimation methods, for 27 chronic disease measures for the 500 largest American cities." The metrics range from cancer prevalence to binge drinking to dental health to undersleeping.

University of Michigan's Health and Retirement Study (HRS) is a longitudinal panel study that surveys a representative sample of more than 26,000 Americans over the age of 50 every two years.

Healthcare service in Africa. The African Economic Research Consortium, African Development Bank, and the World Bank have partnered to create the Service Delivery Indicators program — "a new Africa-wide initiative" that dispatches teams of surveyors "to gauge the quality of service delivery in basic health services" across the continent. The initiative's de-identified data contains results for nine countries so far, including assessments of facility infrastructure, worker absenteeism, and patient case simulations.

4.1 Drugs

Who's using what drugs and how often?

This data set tracks the price of marijuana from September 2nd, 2010 until about the present.

Drug-free school zones in Tennessee. As part of a recent investigation, reporters at Reason Magazine used public records law to obtain geospatial data on each of Tennessee's 8,544 drug-free zones. In addition the geographic boundaries, the shapefile also includes each zone's name and type (school, childcare, park, or library).

5 Urban economics

Redlining. The Mapping Inequality project has digitized more than 150 of the "security maps" produced by the Home Owners' Loan Corporation between 1935 and 1940. Together, the maps "offer a view of Depression-era America as developers, realtors, tax assessors, and surveyors saw it — a set of interlocking color-lines, racial groups, and environmental risks." To download the data for a given map, click on the cloud icon in the top-right corner.

Opportunity Insights. Chetty et. al put out publicly available datasets that can be used by other researchers and practitioners to support their own work. These datasets allow you to analyze social mobility and a variety of other outcomes from life expectancy to patent rates by neighborhood, college, parental income level, and racial background. You can search for datasets by geographic level (e.g., Census tracts), by topic (e.g., education), or by the title of the paper.

How high? The German Aerospace Center is publishing global elevation data derived from its TanDEM-X satellite mission. For five years, two satellites orbited Earth together in a formation that allowed their radars to "'see' the same land area, but from slightly different perspectives" and to calculate elevations based on those differences. Although the most detailed versions of the data are "subject to restrictions due to the potential for commercial exploitation, and thus requires a scientific proposal," the least detailed version (which still clocks in at more than 90 gigabytes) can be downloaded for free.

Bridges. The Federal Highway Administration's National Bridge Inventory contains detailed data on more than 600,000 "highway bridges" in the United States. The inventory goes back to 1992 and contains scores of fields, including the bridge's age, condition, design, and materials.

Most population numbers tell you where people live. But legions of Americans commute for work across city, county, and state lines. The Census Bureau's Commuter-Adjusted Daytime Population Data accounts for these daily migrations. Manhattan's population (non-tourist) population doubles from 1.5 million to 3 million, by far the largest influx by raw numbers.

American infrastructure. The Department of Homeland Security published more than 250 infrastructure-related datasets, which had previously been marked as "For Official Use Only." The release covers a wide range of topics, including datasets on educational facilities, hurricane evacuation routes, poultry slaughterhouses, and sports venues.

What kind of economy does your county have? The USDA Economic Research Service's County Typology Codes categorize each U.S. county based on (a) its dependence on certain industries and on (b) various socio-economic factors. For example, the data classifies 219 counties as "mining-dependent."

Nine years of homelessness estimates. Every January, at the behest of the U.S. Department of Housing and Urban Development, volunteers across the country attempt to count the homeless in their communities. The result: HUD's "point in time" estimates, which are currently available for 2007–2015. The most recent estimates found 564,708 homeless people nationwide, with 75,323 of that count (more than 13%) living in New York City.

Space imagery. You can browse NASA's Image and Video Library online; you can also access it via NASA's API. Through that interface, you can search by caption, keyword, location, photographer, year created, and other fields; in return, you get structured data on each media file. The library was launched two years ago, bringing together more than 140,000 images, videos, and audio files that had previously been spread across dozens of separate collections.

6,000 years of urbanization. Earlier this month, researchers published "the first spatially explicit dataset of urban settlements from 3700 BC to AD 2000," along with a detailed methodology. The dataset digitizes and geocodes population numbers originally tabulated by historian Tertius Chandler (Four Thousand Years of Urban Growth) and political scientist George Modelski (World Cities: -3,000 to 2,000). Though "far from comprehensive," the authors say that the dataset a "first step towards understanding the geographic distribution of urban populations throughout history."

NYC streets: the good, the bad, and the closed. New York City's Department of Transportation publishes a bunch of data, including its own assessments of each street segment's quality on a 1-to-10 scale. It also publishes spreadsheets of all construction-related street closures, by intersection and by block, updated daily.

Neighborhood boundaries. Zillow has created a dataset outlining the boundaries of more than 17,000 neighborhoods in the United States' largest cities, spanning 49 states (all but Wyoming) plus D.C. and Puerto Rico

Privately owned public spaces. In certain cities, private developers can earn zoning concessions by converting sections of their properties into plazas, atriums, mini-parks, and other open-to-the-public spaces. You can download datasets of these "privately owned public spaces" in San Francisco, Seattle, New York City, and — thanks to a recent collaboration between Guardian Cities and local community group — London.

Where we live and build. The European Commission's Global Human Settlement Layer combines satellite imagery and census data to measure three things: population, building density, and urban/rural classification. The resulting datasets are fairly detailed — they provide population estimates for every 250-meter square in the world, for example — and are available for 1975, 1990, 2000, and 2015.

Evictions. A team led by Princeton sociologist and *Evicted* author Matthew Desmond has compiled the United States' first-ever national-scale, publicly-available database of eviction metrics. Desmond's Eviction Lab has collected more than 80 million records from cities, counties, and states across the country, and used them to calculate the number of evictions and eviction filings in each place. (Short methodology here; longer methodology here.) You can download the aggregate data in bulk (after supplying your email address) and explore it through an interactive map.

Building footprints. Microsoft's Bing Maps team has published an open dataset describing the outlines of nearly 125 million buildings in the United States. To build the dataset, the team trained neural networks to detect buildings' footprints in satellite images.

Hourly pedestrians. Melbourne, Australia, has placed dozens of pedestrian-counting sensors across the city, and publishes a dataset of the hourly observations going back to 2009.

Cities' CO2 emissions. An international team of researchers has created a dataset of 343 cities' CO2 emissions. The researchers aggregated and standardized the emissions data — largely self-reported — from three sources: the Carbon Disclosure Project, the Bonn Center for Local Climate Action and Reporting, and a new project at Peking University. The dataset includes cities large and small, from Lagos and Shanghai to Kadıovacık, Turkey (pop. 216) and Brisbane, California (pop. ~4,700). In addition to emissions, the dataset also provides contextual information about the cities, such as average household sizes and gasoline prices.

Metro-area segregation "[W]hy are so many cities and metropolitan areas still split along racial lines? And what is the role of local government in reinforcing those divides? To answer those questions, *Governing* conducted a six-month investigation of black-white segregation in the small cities of downstate Illinois." As part of the investigation, the magazine calculated (and published) school and residential segregation metrics for hundreds of U.S. metropolitan areas, based on the latest Department of Education and Census Bureau data

5.1 Transportation

Chicago cab rides. Chicago's city government published data on more than 100 million local taxi rides taken in the city since 2013. (The city gathers the data through "periodic reporting by two major payment processors believed to cover most taxis in Chicago.") The dataset contains each ride's start/end times, pickup/dropoff location (based on Chicago's "community areas"), distance, cost, payment type, and taxi company.

Ride-hailing. Chicago has become the first city to publish detailed data from ride-hailing services, such as Uber and Lyft. Officials released three datasets — on (anonymized) drivers, vehicles, and trips. The driver and vehicle datasets cover early 2015 through December 2018. The trip dataset covers only November and December 2018; even so, it includes more than 17 million rides. For each ride, the records contain the rough pickup and dropoff location, duration, the approximate fare and tip, and more.

Global bike-sharing. The citybik.es API provides access to live data on every bike-sharing station in more than 400 cities around the world. It's free, and the underlying software is open-source. What data you

get per station depends on the city, but typically includes the number of empty slots, number of available bikes, and location information.

Tens of millions of parking tickets. I Quant NY author Ben Wellington recently discovered that New York City had been "ticketing legally parked cars for millions of dollars a year." To reach that finding, Wellington analyzed three years of parking tickets, amounting to more than 30 million summonses. NYC isn't alone in providing parking ticket data; Philadelphia, Toronto, Baltimore, Seattle, and others publish similar datasets.

How long does it take to get to the nearest city? A team led by researchers at the University of Oxford's Malaria Atlas Project have estimated the time it would take (as of 2015) to get from any square kilometer in the world to the nearest city of 50,000+ people. The analysis, which improves upon a similar effort from 15 years earlier, benefits from "the first-ever, global-scale synthesis of two leading roads datasets – Open Street Map (OSM) data and distance-to-roads data derived from the Google roads database."

Chicago traffic camera violations. The Windy City publishes two datasets on traffic violations. One tallies the daily number of speeding violations in each Children's Safety Zone; the other, red-light violations at each camera-surveilled intersection. Both go back to July 2014. The city also publishes a spreadsheet of city-towed vehicles. Related: The Chicago Tribune's long-running investigation into the city's traffic camera troubles.

130 million traffic stops. "Police pull over more than 50,000 drivers on a typical day, more than 20 million motorists every year. Yet the most common police interaction — the traffic stop — has not been tracked, at least not in any systematic way," according to the Stanford Open Policing Project. To that end, the group has been collecting and standardizing traffic-stop data from state police agencies across America. Its first data release, published Monday, contains 130 million records from 31 states. The records vary by agency, but the most-complete states include the date, time, location, reason, and outcome of each stop; the driver's race, gender, and age; whether a search was conducted; and whether the search found contraband.

German rail. State-owned Deutsche Bahn AG is Europe's largest railway company by revenue, serving 12 million train and bus passengers each day. It also happens to publish a bunch of open data, including datasets on its routes, stations, platforms, and cargo facilities.

Public transit, curated. As a way to "lower the barrier" for analyzing public transportation data, researchers at Finland's Aalto University have published "a curated collection of [now more than] 25 cities' public transport networks in multiple easy-to-use formats including network edge lists, temporal network event lists, SQLite databases, GeoJSON files, and the GTFS data format." On the project's website, you can browse, visualize, and download each city's data

Chicago eviction trends. The Chicago-focused Lawyers' Committee for Better Housing has built a database of evictions in the city from 2010 to 2017. It aggregates nearly 300,000 evictions to the ward, community

area, and Census tract level, and contains metrics on case types, outcomes, legal representation, and more. There's a user guide, bulk download, and methodology.

Uber Movement: Anonymized data from over 2 billion Uber trips.

5.2 Energy, and electricity

Powering America. Every year, the U.S. Energy Information Administration requires thousands of power plants to report detailed data on fuel consumption and electricity generation. The datasets stretch back more than three decades, to 1989.

Electricity prices. In May 2016, U.S. residential consumers paid an average of roughly 12.8 cents per kilowatt hour of electricity. The price was lowest in Louisiana (9.28 cents) and Washington state (9.54 cents), and highest in Hawaii (26.87 cents) and Connecticut (21.63 cents). These data-points, and more, are available through the Energy Information Administration's electric power reports, which are updated monthly.

The grid. The U.S. Energy Information Administration publishes near-real-time data on the Lower 48's electrical grid. The datasets include net electricity generation, flows in and out of the country's various "balancing authorities," regional demand, and forecasts of demand. You can explore the data online, access it through the EIA's API, or download it in bulk. **Helpful:** The EIA's guide to the data and "known issues".

Oil concessions. The OpenOil project aims to collect and standardizes data oil and gas development contracts around the world. So far, they've gathered at least some data from more than 60 countries. They've also published a map of oil concessions in the Middle East and Africa.

Global gas and oil infrastructure. The Department of Energy's National Energy Technology Laboratory has published what it says is the "first-ever database inventory of oil and natural gas infrastructure information from the top hydrocarbon-producing and consuming countries in the world." The database contains tons of geospatial information and "identifies more than 4.8 million individual features like wells, pipelines, and ports from more than 380 datasets in 194 countries. It includes information about the type, age, status, and owner/operator of infrastructure features."

Power outages. Utility companies are required to report major power outages and other "electric disturbance events" to the Department of Energy within a business day (or, depending on the type of event, sooner) of the incident. The federal agency then aggregates the reports annual summary datasets. For each event, the data includes the time it began and was resolved, the geographic areas it affected, the type of incident, and the estimated number of customers affected.

Home energy consumption. For many decades, the Department of Energy's Residential Energy Consumption Survey has been asking people

about their homes' energy-related characteristics (e.g., number of bedrooms and roofing materials) and energy-consuming appliances (e.g., television size and dishwasher use). Then, the agency cross-references those answers with billing data collected "directly from energy suppliers under a mandatory authority granted by Congress." The survey has been conducted 14 times since 1978; survey microdata is available for the eight most recent iterations.

The Reference Energy Disaggregation Data Set has about 500 GB of compressed data on home energy use.

6 Housing

Millions of home loans. Data released through the Home Mortgage Disclosure Act. The law requires large lenders to publish details about each of their loans. You can download the raw data from the FFIEC, or slightly user-friendlier versions from the CFPB.

Historical mortgages. With the help of volunteers, the New York Public Library is transcribing 6,000+ mortgage and bond ledgers from Emigrant Savings Bank, founded in 1850 and the oldest such bank in the city. You can search the transcribed records, or download the (very) raw data.

British property sales. The UK's Price Paid Data contains virtually all of the country's residential property sales, with only a few exceptions. (Sales forced under court order are excluded, for example.) Each row includes the sale price, address, property type, and more. The full, multi-gigabyte dataset covers all sales since 1995, but you can also download files for individual years or the most recent month, or just search the dataset online.

Historical San Francisco rents. To help understand San Francisco's soaring real estate prices, Eric Fischer transcribed decades of apartment and house listings in the San Francisco Chronicle. For each year from 1948 through 1979, Fischer jotted down every monthly rent advertised in the paper on the first Sunday in April. (Similar data for 1979 through 2001 is available from San Francisco's Housing Study DataBook.)

NYC property taxes and exemptions. Property tax data in New York City is technically available to the public, but the city makes it difficult to access. So a pair of civic hackers liberated the data. Now you can download 1.1 million rows of bulk data, which details each property's type, assessed value, taxes due, owner's name, and more.

International house prices since 1975. The International House Price Database combines and standardizes house price indices from 23 countries — mostly in Europe and North America, but also including South Africa, Australia, New Zealand, Japan, South Korea, and Israel.

The dataset, published by the Federal Reserve Bank of Dallas, is deeply documented and updated quarterly.

Building permits. The Census Bureau's Building Permits Survey collects data from thousands of municipalities every month. For each municipality, metro area, and state, the datasets provide the number of permits issued for new residential housing, number housing units authorized, and total estimated value of the new construction.

Many millions of mortgages. Freddie Mac — the government-sponsored, publicly traded company also known as the Federal Home Loan Mortgage Corporation — publishes data on 23 million single-family home mortgages it has originated or guaranteed since 1999. The dataset includes the loan amount and interest rate, the borrower's credit score, the property type (e.g., condo, co-op, manufactured housing), metro area, first payment month, whether the borrower is a first-time homebuyer, and lots more. Freddie Mac requests that you register before downloading the data, but you can also access the files directly.

Home price indices. The Federal Reserve Bank of St. Louis publishes S&P/Case-Shiller Home Price Index data, which measures changes in average home prices over time. The monthly-updated datasets — copyrighted, but free to download — are available at a national and metro-area level, and go back several decades.

House price indices, part two. Two weeks ago, DIP featured Case-Shiller's home price index data. There are, in fact, several other prominent (and downloadable) house price indices, including the Federal Housing Finance Agency's House Price Index, the National Association of Realtors' indices, and Zillow's Home Value Index. Helpful: This guide to various home price indices and how they're constructed, by Jed Kolko, formerly Trulia's chief economist.

Subsidized housing. Earlier this month, the Department of Housing and Urban Development released its "Picture of Subsidized Households" report for 2016. The dataset describes the living conditions, demographics, and finances of families receiving subsidies via the agency's various programs — including public housing, Section 8 vouchers, and several others. The figures are provided for the entire U.S., by state, metro area, housing agency, city, county, Census tract, and even by housing development. HUD provides a data dictionary explaining each field, as well as a tool to query the data without downloading the entire dataset.

Real estate inventories. The National Association of Realtors publishes monthly real estate inventory data "at the national level, the 500 largest metropolitan areas, the 1,000 largest counties, and over 15,000 zip codes." The data, based on the realtors' multiple listing services, goes back five years and "tracks key market metrics including list prices, days on market, and total active inventory."

6.1 Communications

Internet access. Through its Form 477 program, the Federal Communications Commission collects detailed data on broadband internet

access in the United States. One of the easiest ways to access county-level data is through the agency's Mapping Broadband Health in America project, which overlays internet access data and physical health indicators. The latest tabulations come from 2014. In more than a quarter of counties with at least 1,000 residents that year, broadband reached less than 50% of the population.

Global System for Mobile Association (GSMA). The data provide, for all years between 1998 and 2012, yearly geo-located information on mobile phone coverage aggregated across all operators.

Computer memory prices. John C. McCallum has collected the advertised prices of computer memory over time. In 1957, one byte of memory cost $392, or the equivalent of $411 million per megabyte; today, one metabyte costs about a third of a cent.

Broadband access and cost. The U.S. Federal Communications Commission publishes a ton of data on the "wireline" telecommunications industry, including several datasets about broadband internet access. Among them: the places where providers offer service, subscriptions per 1,000 households in each Census tract, and a survey of plans available in urban areas. You can also find a spreadsheet of payphones-by-state at the bottom of that landing page.

7 Labor & Private sector

Minimum wages. Researchers at the Washington Center for Equitable Growth have compiled a dataset of current and historical minimum wages in America. The federal and state minimum-wage data stretches back to May 1974 — when the federal minimum was $2.00 per hour, or roughly equivalent $9.76 per hour in today's dollars — while the data for cities and counties starts in January 2004.

Deaths on the job. Since 1992, the US Bureau of Labor Statistics' has collected data on work-related deaths through its Census of Fatal Occupational Injuries. The results are presented as various cross-tabulations — by industry, demographic, circumstances, and more

Maternity leave policies at hundreds of American companies. The 600+ entries in this searchable, sortable database range from 3M to Amazon to Zynga, and list both paid and unpaid leave. The database, run by the women-in-the-workplace website FairyGodBoss.com, culls from published policies and employee tips. An introductory blog post provides more information.

Who really controls UK companies? The British government began requiring companies to identify all the people who exert power over them. The resulting "People with Significant Control" database contains each person's name, country of residence, nationality, and "nature of control" — e.g., ownership of large numbers of shares, voting rights, or the ability to appoint/remove directors.

Two thousand billionaires. Researchers have compiled a multi-decade database of the super-rich. Building off the Forbes World's

Billionaires lists from 1996–2014, scholars at Peterson Institute for International Economics have added a couple dozen more variables about each billionaire — including whether they were self-made or inherited their wealth.

Business owners. The Census Bureau's Survey of Business Owners and Self-Employed Persons "provides the only comprehensive, regularly collected source of information on selected economic and demographic characteristics for businesses and business owners by gender, ethnicity, race, and veteran status." The most recent data comes from 2012. The survey has been conducted every five years since 1972, but data from before 1992 is "available only in printed form."

Secret offshore companies. On Monday, the International Consortium of Investigative Journalists released data on 210,000 companies, trusts, and funds named in the massive Panama Papers leak. The database is searchable online and downloadable as several CSV files. The dataset includes companies' officers, registered addresses, and middlemen. It supplements a pre-existing cache of of 105,000 companies named in ICIJ's 2013 "Offshore Leaks" investigation.

Two decades of workplace sexual harassment complaints. Anonymized dataset listing all 170,000+ sexual harassment claims submitted to the U.S. Equal Employment Opportunity Commission between October 1995 and September 2016. For each claim, the dataset indicates the date the complaint was filed, the complainant's gender, and the general category of employer. Additional fields — available for most claims, but not all — indicate the complainant's birthdate, race, and national origin, as well as the employer's industry and approximate number of workers.

Employment discrimination cases. "Thousands of people report workplace discrimination to the government each year. Employers are rarely held accountable," according to an investigation by the Center for Public Integrity. Reporters Maryam Jameel and Joe Yerardi "analyzed eight years of complaint data — through fiscal 2017 — from the [U.S. Equal Employment Opportunity Commission] as well as its state and local counterparts, reviewed hundreds of court cases and interviewed dozens of people who filed complaints." The data (on more than 3.7 million allegations and their outcomes) and code are available online

Gender pay gaps in Great Britain. The UK government has begun requiring all companies with at least 250 employees in Great Britain (i.e., England, Scotland, and Wales) to report the pay differences between their male and female workers. The reports include the percentage gaps in hourly earnings, differences in bonus pay, and the proportions of male and female employees in each pay quartile. You can search the data online and also download it as a CSV.

Sneaker factories. Nike, Inc.'s manufacturing map displays 618 factories and material suppliers that the company uses to manufacture its products (as of May 2018). You can export the entire dataset, or browse and filter the data online. For each of the factories, the information includes the factory's name, address, product type, number of workers, percentage of workers who are female, and more.

Retail stores and products. Best Buy's API and Walmart's API both let you search their products and stores. Both also require (free) registration to obtain an API key. In 2016, Best Buy also published bulk data describing its products and stores.

8 International relations and conflict

Every United Nations vote, 1946–2014. This repository contains voting data from each of the UN General Assembly's the first 69 sessions. One spreadsheet summarizes the topic and results of each voted-upon resolution. (The dataset also indicates whether the U.S. State Department identified the vote as "important" — such those condemning human rights violations in Syria and North Korea — in its annual Voting Practices in the United Nations report.) Another file contains each country's individual voting decisions.

UN General Debate speeches. Each September, the United Nations gathers for its annual General Assembly. Among the activities: the General Debate, a series of speeches delivered by the UN's nearly 200 member states. The statements provide "an invaluable and, largely untapped, source of information on governments' policy preferences across a wide range of issues over time," write a trio of researchers who, earlier this year, published the UN General Debate Corpus — a dataset containing the transcripts of 7,701 speeches from 1970 to 2016.

Two decades of UN Security Council debates. A group of researchers have collected, parsed, and added metadata to all UN Security Council debates from 1995 through 2017. The dataset includes more than 65,000 speeches (with information about each speaker), extracted from nearly 5,000 meeting transcripts.

Data on Armed Conflict. PRIO has been involved in the collection of a number of datasets on armed conflicts, including among others the UCDP/PRIO Armed Conflict Dataset for the period 1946 to the present.

Wars: Correlates of War seeks to facilitate the collection, dissemination, and use of accurate and reliable quantitative data in international relations.

Territorial disputes. The Issue Correlates of War project, which started in 1997 with a focus on territorial disputes, gathers "systematic data on contentious issues in world politics." In addition to its two centuries of territorial claims, the project has also catalogued disputes over rivers, maritime zones, and ethnic groups, and compiled supplementary datasets on colonial history, historical country names, and more

Armed Conflict and Location Event Dataset The Armed Conflict Location & Event Data Project (ACLED) is a disaggregated conflict analysis and crisis mapping project. ACLED is the highest quality, most widely used, realtime data and analysis source on political violence and protest in the developing world. Practitioners, researchers and governments depend

on ACLED for the latest reliable information on current conflict and disorder patterns.

Armed conflict. The Uppsala Conflict Data Program maintains several large, interconnected datasets describing decades of war, genocide, and other armed hostilities.

Subnational conflicts. University of Michigan–based researchers have created "a repository of micro-level, subnational event data on armed conflict and political violence around the world." The project, dubbed xSub, standardizes information from 21 data sources, and includes conflicts in 139 countries between 1942 and 2016. For each administrative boundary (e.g., country, province, district) and data source, xSub's data counts the number of violent incidents by year, month, week, or day. The numbers are also broken down by the sides involved, who initiated the conflict, and what types of force were used.

Rebel groups and natural resources. The Resources and Conflict Project's Rebel Contraband Dataset "measures if and how rebel groups earn income from the exploitation of natural resources or criminal activities." The dataset spans 1990–2015, covers more than 70 countries, and specifies dozens of types of resources — such as oil, cannabis, gold, tea, and timber.

Peace agreements. The PA-X Peace Agreements Database contains structured information about 1,500+ "formal, publicly-available documents" that address "conflict with a view to ending it." The database covers more than 140 peace processes between 1990 and 2015, and each agreement has been coded for more than 200 variables — for instance, whether the agreement contains provisions about religious groups.

Historical battles. Political scientist Jeffrey Arnold has converted the U.S. Army Concepts Analysis Agency (CAA) Database of Battles from a series of Lotus 1-2-3 worksheets into tidier, easier-to-use CSV files. The dataset includes details of 660 battles — associated with several dozen wars — between 1600 and the mid/late-1900s. The fields indicate each battle's "name, date, and location; the strengths and losses on each side; identification of the victor; temporal duration of the battle," and more.

Coups d'état. Two political science professors at the University of Kentucky are compiling a dataset of coup attempts. So far, the dataset covers both successful and unsuccessful attempts from 1950 to late 2015. During those 65+ years, coup plotters have been foiled about half the time, with 236 victories and 238 failures. According to the dataset, Bolivia's top leaders have faced 23 coup attempts, including 11 successful overthrows — more than any other country by either metric.`

North Korea negotiations and provocations. The Center for Strategic and International Studies' Beyond Parallel project publishes several databases related to North Korean international relations — including 200+ negotiations between the U.S. and DPRK since 1990, and several hundred military provocations since 1958.

Militarized disputes. The Militarized Interstate Dispute datasets provide details about more than 2,200 instances between 1816 and 2010 where a government "threatened, displayed, or used force against another" — including each dispute's timing, participants, death count, result, and more. A supplementary database tracks the disputes' locations. The datasets are part of the Correlates of War project, which was founded in 1963 and which strives for "the systematic accumulation of scientific knowledge about war."

Rebel groups. "The Foundations of Rebel Group Emergence (FORGE) Dataset examines the roots of rebellion by considering the characteristics and activities of the 'parent' organizations from which rebel groups emerged," plus details such as "the organization's 'birthdate' and founding location, initial goals, ideology, and ethnic/religious foundations." The new dataset, developed by the University of Arizona's Jessica Maves Braithwaite and the University of Maryland's Kathleen Gallagher Cunningham, contains 430 rebel groups active between 1946 and 2011.

9 Politics

9.1 US politics

Legislative linguistics. The Sunlight Foundation's Capitol Words project lets you explore the frequency of words and phrases in the Congressional Record since 1996. For example: "weapons of mass destruction", "war" vs. "peace", or "Obamacare".

The State of The State of The States. FiveThirtyEight has collected the text of all 50 state governors' 2019 annual addresses, and has analyzed the most common words and phrases used by Republican and Democratic governors.

Congressional junkets. The U.S. House of Representatives requires all staff to reveal all "gift travel" — i.e., "free" trips that the government didn't pay for. The Office of the Clerk compiles those filings into a database containing each trip's dates and sponsors.

State legislator ideologies. Political scientists Boris Shor and Nolan McCarty's have assigned ideology scores, on a conservative-to-liberal scale, to every US lawmaker in all 50 state legislatures. The most recent update, published May 2018, covers more than 22,000 legislators from 1993 through 2016. Shor and McCarty derived the numeric scores from a combination of legislative voting records and responses to Vote Smart's "Political Courage Test."

County-level and precinct-level results. OpenElections, a Knight Foundation–funded project, aims "to create the first free, comprehensive, standardized, linked set of election data for the United States." They've made progress but are looking for additional volunteers. In the meantime, you can download county-level presidential results from the National Atlas of the United States for 2004, 2008,

and 2012 — or all combined. And you can download precinct-level results from 2002 to 2012 from the Harvard Election Data Archive.

State-level results. Perhaps better known for its campaign-finance data, the Federal Election Commission also publishes official state-level results for presidential, House, and Senate elections going back to 1982. The results include all official candidates, and sometimes even write-ins (depending on the state).

Presidential newspaper endorsements. Noah Veltman has collected all presidential endorsements (and non-endorsements) of 100+ major newspapers from 1980 (Reagan vs. Carter) to 2016.

Congressional Research Service reports, in bulk. The website EveryCRSReport.com provides unprecedented public access to reports from the Congressional Research Service — essentially the national legislature's think-tank. The website, which launched by Demand Progress and Congressional Data Coalition, also lets you download metadata and text for each report.

Comparing election forecasts. State-level predictions from all nine forecasters in the analysis.

TrumpWorld. Big database of organizations and people connected to President Trump, his family, advisers, and Cabinet picks. connections between more than 1,500 organizations and people altogether.

Historical congressional results, historical boundaries. Through the Constituency-Level Elections Archive and other sources, you can get historical election results for the U.S. Congress. And through the work of Jeffrey B. Lewis et al., you can get data describing the historical boundaries of each congressional district. In a Scientific Data article, quantitative geographer Levi John Wolf presented a dataset that brings the two types of information together, so that all congressional election results from 1896 to 2014 are "explicitly linked to the geospatial data about the districts themselves."

Political crowd estimates. The Crowd Counting Consortium, launched earlier this year, is a volunteer effort to "[collect] publicly available data on political crowds reported in the United States, including marches, protests, strikes, demonstrations, riots, and other actions." The team publishes monthly spreadsheets that list each crowd's date, location, type, and cause (e.g., "Oppose removal of confederate statue"); high and low size estimates; the number of reported arrests and injuries; links to sources; and additional details.

European protests, 1980 to 1995. A team led by University of Kansas professor Ron Francisco has collected and codified data on protests, strikes, and other "coercive acts" in dozens of European countries during the late 20th century. There's a row for each day of each protest, and each row specifies the issue at stake, the organizers, their target, the type of action, and the location — as well as the number of protesters, arrests, injuries, and deaths.

Protests in autocracies. Political science professor Nils B. Weidmann and collaborators have taken tens of thousands of reports — published by the AP, AFP, and BBC Monitoring — of political protests in autocratic countries and have turned them into structured data. The resulting Mass Mobilization in Autocracies Database is available to download (free registration required), and comes with documentation and code examples. The database currently covers 2003–15, with data for 2016–17 in the works.

Women running for the U.S. House. As the basis for his recent study, "Is Running Enough? Reconsidering the Conventional Wisdom about Women Candidates" (paywalled, but a draft is freely available), PhD candidate Peter Bucchianeri compiled a dataset of female candidates in House primary elections from 1972 to 2010. The spreadsheet covers 1,242 candidacies, and includes each candidate's party, votes garnered in the primary and general elections, the seat's incumbency status, the district's demographics, and more.

Plum presidential appointments. Every four years, Congress publishes United States Government Policy and Supporting Positions, better known as the Plum Book. The 2016 version, which is available as both PDF and Excel files, identifies more than 8,000 executive and legislative branch jobs subject to "noncompetitive appointment." Those positions include 1,710 presidential appointments.

Voters' attitudes and choices, over time. The Democracy Fund Voter Study Group, "a research collaboration comprised of nearly two dozen analysts and scholars from across the political spectrum," has published the participant-level data from its 2016 VOTER survey. It's a "unique longitudinal data set" that represents the "political attitudes, values, and affinities" of 8,000 American adults who were interviewed first in December 2011, then again before and after the 2012 election, and again in December 2016.

State lawmakers' financial disclosures. For a recent investigation into state legislators' financial interests, the Center for Public Integrity "analyzed disclosure reports from 6,933 lawmakers holding office in 2015 from the 47 states that required them." You can search through the disclosures and download the data. For each of the 11,000+ disclosed interests, the dataset includes the lawmaker's state, legislative body, and district; the name and industry of the financial interest; and a link to the lawmaker's personal disclosure form.

Executive orders. The U.S. Office of the Federal Register publishes structured data on every presidential executive order since 1994. For each of the 886 entries, the dataset provides the order's title, the date it was signed, the president who signed it, and where to find it in the Federal Register.

State campaign finance laws. The nonpartisan Campaign Finance Institute has launched a database of current and historical state campaign finance laws. The information goes back to 1996 and describes each state's contribution limits, various kinds of prohibitions, disclosure rules, and more. You can download the full dataset or explore it online.

Political ads online. Google recently launched a database of political ads "that have appeared on Google and partner properties." The searchable and downloadable dataset indicates the organization that paid for each advertisement, approximately how much they spent, how long the ad ran, what demographics were used for targeting, and roughly how many people it reached. A few months ago, Facebook launched a similar initiative, but you need to logged in to view it and you can't download the data. You can, however, get Facebook political-advertising data from at least two sources: A repository of 267,000 ads scraped from Facebook's official archive by NYU researchers, and ProPublica's ongoing, detailed database of ads and targeting parameters gathered through their Political Ad Collector.

Congressional district demographics. The Census Bureau's My Congressional District tool lets you browse (and download) demographic, socioeconomic, and business data corresponding to each of the country's 435 congressional districts. Political scientist Ella Foster-Molina has compiled a historical dataset containing similar information for 1972 to 2014; it also contains details about each district's representatives — such as their personal characteristics, the committees they served on, and the number of bills they sponsored.

9.2 International Politics

Presidential speeches: The dataset is comprised of 953 speeches delivered during the period 1819-2016 in Spain plus ten Latin American countries: Argentina, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Mexico, Paraguay, Peru, Paraguay, and Venezuela. The selected speeches constitute the respective countries' closest equivalent of the United States' "State of the Union Address."

Political party data, linked. Party Facts is a "collaborative data collection" that links various political-party datasets together. The project has two main tables. One contains basic information about 4,100+ political parties in more than 200 countries, including each party's mother-tongue name and English translation, year founded, and Wikipedia page. The second table cross-references each party with its unique identifier in 26 external datasets,

Parties and parliaments. ParlGov, "a data infrastructure for political science," has collected detailed information on 1,500+ political parties, the results of 900+ elections, and the formation of 1,400+ parliamentary cabinets. The 37 countries it covers include every member of the European Union plus certain non-EU members of the OECD (such as Israel, Turkey, and Canada — but not the United States). The datasets are available in several formats, can be explored online, and come with extensive documentation

Autocracies. The Autocratic Ruling Parties Dataset bills itself as "the first comprehensive data set on the founding origins, modes of gaining and losing power, ruling tenures, and other characteristics of autocratic ruling parties." The dataset, created by political science professor Michael K. Miller, covers nearly 500 parties in more than 150 countries between 1940 and 2015.

Euro-bank speeches. European Central Bank has begun publishing a spreadsheet of all executive board members' speeches since the late 1990s. The dataset contains each speech's date, speaker(s), title, subtitle, and text; the ECB says it will be updated every two months.

Political conditions. "The Rulers, Elections, and Irregular Governance (REIGN) dataset describes political conditions in every country each and every month. These conditions include the tenures and personal characteristics of world leaders, the types of political institutions and political regimes in effect, election outcomes and election announcements, and irregular events like coups, coup attempts and other violent conflicts." The latest dataset covers 200 countries, from 1950 to the present, and includes dozens of variables for each monthly snapshot.

Observatorio de Elites Parlamentarias de América Latina (Elites): Collects the perceptions and attitudes of Latin American parliamentarians in 17 countries and now has more than 4,000 interviews.

Polity IV project: the most widely used resource for monitoring regime change and studying the effects of regime authority. The POLITY2 measure is available since the early 1800s, which enables us to extend the analysis far back in time, instead of restricting ourselves to more recent periods.

Democracy. Varieties of Democracy bills itself as "a new approach to conceptualizing and measuring democracy" — one that "reflects the complexity of the concept of democracy as a system of rule that goes beyond the simple presence of elections." The project scores countries annually on five high-level aspects of democracy, which are further broken down (by thousands of country-experts, based on a detailed codebook) into hundreds of more granular "indicators," such as how often the government publicly attacks the judiciary, the extent to which authorities respect religious freedom, and the proportion of journalists who are women. Version 9 of the dataset, released earlier this month, covers 1789 to 2018 and includes 202 countries

Foundation for Election Systems ElectionGuide is the most comprehensive and timely source of verified election information available online. It provides content on: National elections around the world; Subnational elections of high interest; Political parties and candidates; Referendum provisions; News on election-related laws and

developments around the world; Political institutions and electoral systems; Election results and voter turnout.

A century of UK general elections On Monday, the British government published a dataset of voting results, by party and parliamentary constituency, for every UK general election since 1918 — merging modern data with a handful of historical sources.

Global Dataset on Events, Location and Tone (GDELT) an open-access database that, through an automated coding of newswires, collects information on the occurrence and location of political events, including protests, worldwide.

(Almost) every politician. On everypolitician.org, you can search and download data on 70,000+ legislators (past and present) from 233 countries. (Among those missing: Cuba, Ethiopia, and Qatar.) The dataset includes each lawmaker's party affiliation, years served, gender, social media profiles, and more.

World leaders. The Archigos dataset provides historical data the leaders of nearly 200 countries between 1875 and 2015. The dataset — a collaboration between political scientists Hein Goemans, Kristian Skrede Gleditsch, and Giacomo Chiozza — includes basic demographic information, plus categorizations of how each leader came to power, how they lost it, and their post-office fate

Russian election results. In a recently-updated paper, three academics say they've found "convincing evidence of election fraud" in federal Russian elections since 2004. To support their analyses, the researchers have published the underlying data, which includes polling station data from seven Russian elections (as well as one Polish and one Spanish election, which showed no such signs of fraud).

Global elections. The Constituency-Level Elections Archive, based at the University of Michigan, collects and standardizes results from lower-house legislative elections around the world. The latest release covers 1,591 elections from 136 countries.

Protests and political violence in Africa and Asia. The Armed Conflict Location & Event Data Project (ACLED), records the locations, dates, actors, and outcomes of "all reported political violence and protest events in over 60 developing countries in Africa and Asia." The Africa datasets currently go back to 1997 and cover more than 50 countries. The Asia datasets currently only go back to 2015, but ACLED's website says it's planning to add data soon going back to 2010. Both of the datasets are extensively documented, as is the methodology

Political party manifestoes. The Manifesto Project has collected and coded more than 4,000 electoral manifestoes from more than 1,000 political parties in more than 50 countries between 1945 and 2015. For each manifesto, the project's dataset indicates whether the document expresses support for/against dozens of policies and attitudes, including "market regulation," a "national way of life", "environmental protection," and "anti-imperialism."

Global digital activism 1,180 coded cases of digital activism from 151 countries and dependent territories, r from 1982 through 2012.

Russian presidential voting. Software engineer Michael Penkov has scraped the official, polling station-level results for Russia's recent presidential election, and made the data available as a single JSON file. He's also published an introductory Python notebook, which explains the data structure and provides English translations for the Russian field names.

European electoral polling. PollOfPolls.eu aggregates political polls from 30 European countries. The Vienna-based initiative has, for instance, collected and standardized more than 1,000 individual polls on British parliament since 2014, and 60 on the Bavarian state elections. You can download each set of standardized data as either JSON or CSV.

Ethnonationalism. Christina Isabel Zuber and Edina Szöcsik's Ethnonationalism in Party Competition dataset compiles ratings for more than 200 political parties in 22 European countries. Experts rated the parties twice — first in 2011, and then again in 2017 — on a range of factors, such as the centrality of ethnonationalism to the parties' platforms, and their positions on territorial autonomy for minorities

9.3 Lobbying and political campaigns

Foreign influence campaigns on Twitter. Twitter released data on the public activity of "3,841 accounts affiliated with the [Internet Research Agency], originating in Russia, and 770 other accounts, potentially originating in Iran." Together, the datasets "include more than 10 million Tweets and more than 2 million images, GIFs, videos, and Periscope broadcasts."

Foreign lobbyists. The United States' Foreign Agents Registration Act requires lobbyists who represent foreign governments to file paperwork with the Department of Justice. The database has long been available to browse online. Feature: full-text search, an API, and bulk downloads.

European Union lobbying. The EU publishes a searchable database of people and organizations registered to lobby the European Parliament and the European Commission. The website LobbyFacts.eu takes that data and makes it available via an API. LobbyFacts also scrapes the European Commission's disclosed lobbying meetings, which you can download here (warning: 10-megabyte direct download).

You've probably heard of PolitiFact, the Tampa Bay Times project that fact-checks what politician say. What you might not know: PolitiFact has an API. You can use it to fetch detailed data the project's national and state-level editions.

Campaign ad purchases. The FCC requires broadcasters to keep records of "all requests for broadcast time made by or on behalf of a candidate for public office." With the help of volunteers, Political Ad Sleuth gathers those records and enters them into a searchable, downloadable database. Note: Due, in part, to the difficulty of transcribing the (non-standardized) records, the information in the database is incomplete.

More political ads. The Internet Archive's Political TV Ad Archive uses audio fingerprinting to identify the campaign ads playing in key primary states. You can search the database, watch the ads, and download the data. The data file contains information about each ad's sponsor, pro/con-ness, TV network, and time of airing.

Government lobbying. U.S. lobbyists must notify Congress within 45 days of being retained by new clients. Every quarter after that, they're required to file activity reports that detail the agencies they lobbied, the topics they covered, and the income they earned. Bulk downloads of both types of reports are available as XML files from the House (going back to 2004) and from the Senate (since 1999). Although they receive the same filings, each chamber "follows different data-cleaning, processing, and editing procedures before storing the data," according to this recent GAO report.

California elections and campaign finance. Since 2014, the California Civic Data Coalition has been working to improve access to CAL-ACCESS, "the jumbled, dirty and difficult government database that tracks campaign finance and lobbying activity in California politics." Their cleaned-up datasets are updated often and include formats suitable for beginners, "database junkies," and masochists. The organization released data files cataloging every state ballot measure and candidate for public office since 2000.

Local lobbying. Some cities — including San Francisco, Los Angeles, and Austin — provide downloadable databases of lobbyists who've officially registered to influence their administrations. Chicago has gone one step further, publishing data on lobbyists' compensation, expenditures, gifts, and more.

The 2012, 2016 Presidential Campaign Finance data

10 Government data

The Worldwide Governance Indicators: Reports aggregate and individual governance indicators for over 200 countries and territories over the period 1996–2015, for six dimensions of governance.

Socrata's software powers open-data portals around the world. But downloading large datasets — e.g., this 2.8-gigabyte dataset of NYC parking tickets — from Socrata-powered portals can feel, well, sluggish. One solution: OpenDataCache.com, a free website that provides faster-to-download versions of virtually every dataset from 50+ Socrata portals.

Wiretaps. The Administrative Office of the United States Courts posts its annual "wiretap reports", which provide details on the wiretaps that state and federal judges have authorized. The agency published its 2018 report; the supplementary data includes each wiretap's jurisdiction, authorizing judge, date of authorization, type of intercept, number of communications intercepted, total cost, and more

Local incarceration, 1970–2015. The Vera Institute of Justice's recently-expanded Incarceration Trends project combines data from a range of government reports — such as the *Census of Jails* and the *National Corrections Reporting Program* — into a single, longitudinal, well-documented dataset. For each county and year, the dataset tallies the number of people admitted to jails and prisons, the average daily incarcerated jail and prison population, and other related details. Many of the counts are also broken down by race, ethnicity, and sex

Parking tickets in Chicago. "For the first time, the city's database, which tracks more than 28 million parking and vehicle compliance tickets, is easily available to the public," according to ProPublica Illinois, which has published the two-gigabyte dataset in collaboration with WBEZ. The dataset, which covers January 2007 to mid-May 2018, "includes information on when, where, and by whom tickets were issued; de-identified license plates; vehicle make; registration zip code; the violation for which the vehicle was cited; the payment status and more."

Open Knowledge International has just published its latest survey of openly available government data. This year's audit includes 112 countries and territories, up from 97 last year. The survey scores each based on the availability of datasets in 13 key categories (e.g., "election results," "government spending," and "pollutant emissions") and links out to the available datasets.

The New Mexico city publishes dozens of regularly-updated, well-documented datasets. Among them: government employee earnings, the number of daily visitors to the city's swimming pools, real-time bus locations, the geography of police beats, and the city's complete vendor checkbook.

City of Chicago Employee Salaries. This file contains salaries for the City of Chicago.

City of Phoenix Employee Salaries. City Official's salaries for the City of Phoenix, Arizona.

White House Staff Salaries Dataset. Information on the salaries of staff at the White House

Bills, bills, bills. Congress has finally begun publishing official bulk data on the status of its bills — something open-government advocates had been requesting for more than a decade. The bulk downloads include an XML file for each piece of legislation, with indicators tracking (among other things) committee referrals and actions.

Public Policy. The Correlates of State Policy Project aims to become a "one-stop shop" for data related to public policy in America's 50 states. So far, the project is tracking 700+ aspects of each state's laws, budgets, demographics, and more. Among the policy variables: Can pharmacies dispense emergency contraception without a prescription? Does the state ban corporal punishment in schools? and Does the state have an endangered species act? Don't miss the codebook, which describes the data and sources in greater detail.

Government payrolls. Vast trove federal payroll data. Those records — provided by Office of Personnel Management through the Freedom of Information Act — cover more than 40 years and millions of employees. The dataset includes salaries, titles, job types, and demographic variables. In many-but-not-all cases (per OPM's data release policies), it also includes names.

Military spending. The Stockholm International Peace Research Institute's Military Expenditure Database is based on official reports, International Monetary Fund yearbooks, newspaper articles, and other sources. It covers most major countries since the 1950s and more than 100 countries since 1988. The dataset also quantifies military spending on a per-capita basis, as share of the country's GDP, and as a proportion of total government spending. Also: The Defense Manpower Data Center publishes spreadsheets detailing the number of active and reserve U.S. personnel stationed in each state, territory, and foreign country. Previously: SIPRI's database of international arms transfers

The federal checkbook. From Treasury.io: "Every day at 4pm, the United States Treasury publishes data tables summarizing the cash spending, deposits, and borrowing of the federal government." Those data tables "catalog all the money taken in that day from taxes, the programs, and how much debt the government took out." On Monday, for instance, the government spent $481 million on the Postal Service. One hitch: The Treasury's data tables are (subjectively) ugly and (objectively) spreadsheet-unfriendly. So Treasury.io — an open-source civic project — continuously converts the files into good ol' tabular data. You can download individual tables as CSVs, get the whole dataset as a big SQLite database, or query the API.

U.S. Treasury sanctions. Through its Office of Foreign Assets Control, the Treasury publishes several datasets that describe the people and companies subject to U.S. economic sanctions. The two main listings are the Specially Designated Nationals and Blocked Persons ("SDN") and the Consolidated Sanctions List. Those contain only currently-sanctioned entities, but the Treasury also publishes (semi-structured) documents describing historical additions and removals.

Death and taxes in the Garden State. The nonprofit organization Reclaim The Records recently obtained New Jersey's death index, and has made it available to search and download. The records include structured data for 1,275,833 deaths in the state between 2001 and 2017, plus digitized images of the death index for 1901-1903, 1920-1929, and 1949-2000. The structured data contains each person's name, date of birth, date of death, and death certificate number — plus, for the most recent records, the locations of birth and death.

Public pension plans. Boston College's Center for Retirement Research compiles detailed financial data on state and local public pension plans. The database covers fiscal years 2001–18 and includes 180 public pension plans, which together "account for 95 percent of state/local pension assets and members in the US."

The Book of the States. The Council of State Governments' annual Book of the States compiles 50-state reference tables on a range of topics, including elections, finances, courts, and more. It has been published since 1935, and the tables for the past decade-plus are available as spreadsheets.

10.1Taxes

Three centuries of taxation. For 220 countries between the 1750s and 2018, the Tax Introduction Dataset tracks "the year of the first permanent introduction at the national level of government of six major taxes, as well as on the top statutory tax rate for that year." The six taxes are those on personal income, corporate income, inheritance, and general sales, plus VATs and compulsory social security contributions.

Taxes filed. The IRS publishes a ton of tax statistics. One of the most interesting portions: data aggregated from individual income tax returns (i.e., Form 1040s), which the IRS provides at the state, county, and ZIP code level. Those datasets' 100+ fields include details that range from the basic (e.g., the number of tax filings and total income reported) to the more obscure (e.g., the number of returns that included "educator expenses" and the total amount of overpayments refunded).

Nonprofit IRS filings. The Internal Revenue Service released a huge dataset of nonprofits' annual Form 990 filings, which provide details on program expenses, salaries, and more. More than 60% of Form 990s are filed digitally, according to the IRS. Previously, those forms were only available as images; now the IRS is publishing them as analysis-friendly XML files. (You can also download the data in bulk from the Internet Archive, thanks to Carl Malamud, the public domain advocate who led the fight for 990s-as-XML.)

Hillary Clinton Income Taxes: Adjusted gross income and taxes owed by Hillary are included for each year from 2000-2015.

Global tax revenues. The Organisation for Economic Co-operation and Development (OECD) has launched a database "providing detailed and comparable tax revenue information for 80 countries around the world." The Global Revenue Statistics Database, "which will expand to cover more than 90 countries by the end of 2018," breaks tax revenues into dozens of categories and subcategories — such as sales taxes, taxes on capital gains, and taxes on exports.

11 Police, law and order

The Police Open Data Census, created by Code for America fellows in Indianapolis, is tracking "currently available open datasets about police interactions with citizens in the US," including officer-involved shootings, use of force, and citizen complaints. The census currently covers 36 police departments. Related: The NYPD says it will start tracking all officer use-of-force incidents — not just gunfire — next year, the New York Times reports.

The Citizens Police Data Project has collected more than 56,000 allegations of police misconduct. The data, covering 2002-2008 and 2011-2015, includes demographic information about the complainant and the officer, as well as the type and location of the incident. Click here to download the raw data. Related: The City of Chicago's wide-ranging data portal includes a spreadsheet of every reported crime in the city since 2001; you can explore neighborhood trends via the Chicago Tribune.

Deaths in police custody. At least 6,913 people died while in the custody of Texas police, jails, and prisons between 2005 and 2015, according to the newly-launched Texas Justice Initiative. The data, gathered through freedom-of-information requests, contains the age, sex, and race/ethnicity of each person who died, as well as the general cause of death and a more detailed summary.

The Department of Justice is authorized to investigate police departments that display a "pattern or practice" of civil rights violations. In April, the Marshall Project began publishing a spreadsheet of the DOJ investigations into local law enforcement. The dataset, which is updated regularly, indicates when each case began, when it ended, and what type of agreement (if any) was reached.

Order in the courts. CourtListener gathers and publishes bulk data the Supreme Court, all federal appeals courts, and hundreds of other jurisdictions. The files include opinions, audio from oral arguments, dockets, and citations. It also has an API.

Local justice data. The Sunlight Foundation's Hall of Justice brings together "nearly 10,000" criminal justice datasets and research documents from across the United States. You can search for topics and filter by geography, publisher, and accessibility (open, open-but-not-machine-readable, restricted access, etcetera.).

"The watch list Chicago police fought to keep secret.". The Chicago Sun-Times has obtained and published an August 2016 copy of the Chicago Police Department's "Strategic Subject List," a database that scores nearly 400,000 (unnamed) people on a scale from 10 to 500, based on an algorithm that attempts to estimate their risk of being involved in gun violence (either as a shooter or a victim). The database includes demographic, geographic, criminal history, and other information about the people it ranks.

Federal court cases. The U.S. Federal Judicial Center's "Integrated Data Base" contains a longitudinal record of all federal criminal, civil, and appellate court cases going back to the 1970s, as well as bankruptcy cases going back to late 2007. Each dataset contains dozens of detailed fields — including each case's jurisdiction, name, docket

number, relevant legal statutes, and more — accompanied by explanatory codebooks. You can download single-year snapshots and cumulative files, or interactively select specific slices of data to export.

Amendments in America. 11,000 proposed amendments to the United States Constitution from 1787-2014

Court decisions. The Caselaw Access Project aims "to make all published U.S. court decisions freely available to the public online, in a consistent format, digitized from the collection of the Harvard Law Library." Currently, the project provides an API for fetching data on more than 6 million cases published between 1658 and 2018 — though public access is limited to downloading 500 cases per day. You can also download bulk data for all cases in Illinois and Arkansas, but getting bulk data for other states currently requires a research agreement.

11.1Crime, and misconduct

NYC felonies. The historically opaque New York Police Department has finally started publishing incident-level felony data — something that cities such as Chicago and Boston have done for years. The dataset includes the date, time, and approximate location of each offense. It currently covers the first nine months of 2015 and will (apparently) be updated quarterly.

Mass shootings in America. ShootingTracker.com provides datasets listing all U.S. mass shootings — defined as "when four or more people are shot in an event, or related series of events" — since 2013. So far in 2015, mass shootings have killed 447 people and wounded an additional 1,292.

Fatal and nonfatal officer-involved shootings. For an investigation published Monday, Vice News spent "nine months collecting data on both fatal and nonfatal police shootings from the 50 largest local police departments in the United States." They've published raw and standardized data on every shooting, plus the code they used to analyze it.

Good Jobs First's Violation Tracker calls itself "the first national search engine on corporate misconduct." The new database currently contains nearly 100,000 penalties for environmental, health, and safety violations — sourced from 13 U.S. regulatory agencies — since 2010. Search results can be downloaded as CSV files, which contain a few additional fields. (Tip: Search for "*" to get all cases.) The largest single fine? The Department of Justice's $20.8 billion penalty this year against BP.

Texas has placed the last words of every inmate executed since 1984 online.

The National Registry of Exonerations contains "every known exoneration in the United States since 1989—cases in which a person was wrongly convicted of a crime and later cleared of all the charges based on new evidence of innocence." For each of the 1,702 cases, the registry

includes details about the exoneree, the crime, and the factors — such as new DNA evidence — that contributed to the exoneration.

Risky predictions. "There's software used across the country to predict future criminals. And it's biased against blacks," a ProPublica analysis has found. The investigation focused on risk assessments and recidivism in Broward County, Florida, and found that black defendants were more likely than white defendants to be mislabeled as "high risk." The reporters have published their methodology, code, and the underlying data — including two years of Broward County risk assessments — on GitHub.

State prison admissions, by county. Reporters at the New York Times have assembled a dataset counting the number of inmates each U.S. county sent to state prison in 2006, 2013, and 2014. The reporters derived the numbers from the Bureau of Justice Statistics' National Corrections Reporting Program, which only certain researchers can access.

Domestic radicalization. The Profiles of Individual Radicalization in the United States (PIRUS) database "contains deidentified individual-level information on the backgrounds, attributes, and radicalization processes of nearly 1,500 violent and non-violent extremists who adhere to far right, far left, Islamist, or single issue ideologies in the United States" — including the Klu Klux Klan, the Taliban, and the Animal Liberation Front, among others. The dataset covers 1948 through 2013 and was released earlier this year by a team at the University of Maryland.

Hate crimes in the United States. Since the 1990s, the FBI has collected data on hate crimes from local law enforcement agencies. On Monday, the bureau released data for 2015, reporting "5,850 criminal incidents and 6,885 related offenses, as being motivated by bias toward race, ethnicity, ancestry, religion, sexual orientation, disability, gender, and gender identity." Those numbers are based on reports from 14,997 participating agencies. On the FBI's website, you can view and download summary tables of the most recent data. You can also download incident-specific data for 1992 through 2014 from the National Archive of Criminal Justice Data. Unfortunately, as ProPublica noted, the FBI dataset is "deeply flawed"; more than 3,000 law enforcement agencies don't participate in the program.

Crime in cities. The Marshall Project has collected and analyzed four decades of FBI data "on the most serious violent crimes in 68 police jurisdictions." The FBI data covers 1975 through 2014; the reporters "also obtained data directly from 61 local agencies for 2015 — a period for which the FBI has not yet released its numbers."

The serial killer information center was created to provide students, researchers, and the media with accurate data on serial killers.

Easier-to-use crime data. Earlier this month, the FBI and 18F released the first iteration of their Crime Data Explorer, a website that simplifies access to the FBI's Uniform Crime Reporting program. You can download bulk data on individual incidents, state and national trends, hate crimes, arrests, assaults on officers, police employees,

human trafficking, and cargo theft. You can also access the data via an API. Caution: The FBI's data collection program is voluntary; not all law enforcement agencies participate. United States crime from 1960 to 2012.

School shootings. Over the past year, reporters at the Washington Post "attempted to identify every act of gunfire at a primary or secondary school during school hours since the Columbine High massacre on April 20, 1999." Using a range of sources, the reporters "reviewed more than 1,000 alleged incidents, but counted only those that happened on campuses immediately before, during or just after classes." The resulting database currently contains more than 200 incidents and can be downloaded as a CSV. For each shooting, the database includes details about the location, timing, circumstances, shooter, casualties, and the school's students.

Fatal and non-fatal gun crime. On Thursday, Sarah Ryley, Sean Campbell, and I published a deeply-reported investigation into U.S. cities' failure to solve shootings — a year-long collaboration between The Trace and BuzzFeed News. To reach our quantitative findings, we analyzed (and standardized) three major FBI datasets, internal data from 22 police departments, and a database of Baltimore victims and suspects. Data, code, and methodologies for the analyses are available on GitHub.

The death penalty. Law professor Brandon L. Garrett has led an effort to compile data on every death sentence in the U.S. since the early 1990s. Garrett's "End of its Rope" database currently includes more than 4,900 sentencings, and specifies each defendant's name, race, and gender; the state, county, and year of the sentence; whether it was a resentencing; and whether the defendant has been executed. You can download the data, browse it online, and explore it via an interactive map.

How do gender and mental illness affect crime? This data set was collected explicitly with that question in mind.

11.2Terrorism

Terrorism incidents. The Global Terrorism Database, run by a University of Maryland–based consortium, is an "open-source database" of more than 170,000 terrorist events. The database, which currently covers 1970 through 2016, is well-documented and includes information about about the attackers, locations, weapons, victims, and more.

Terrorist attack dates: The Chicago Project on Security and Terrorism (CPOST) maintains a searchable database on all suicide attacks from 1974 through June 2016. The database includes information about the location of attacks, the target type, the weapon used, and systematic information on the demographic and general biographical characteristics of suicide attackers.

Terrorism prosecutions. "The U.S. government has prosecuted 808 people for terrorism since the 9/11 attacks. Most of them never even got close

to committing an act of violence." The underlying data — available on GitHub — contains each defendant's name and demographic details, as well as each case's description, status, charges, charge date, conviction date (if convicted), jurisdiction, and more.

Historical terrorist groups. Joshua Tschantret, a political science Ph.D. candidate at the University of Iowa, has compiled a dataset of 260+ terrorist groups formed between 1860 and 1969. For the purposes of the dataset, "terrorist groups are operationally defined as politically-motivated non-state actors using *bombings* or *assassinations*," Tschantret writes in an introductory article (PDF). About one-third of the groups in the dataset operated in the US, Russia, or China; the rest are spread across dozens of other countries.

11.3Guns

The Arms Transfer Database tracks the international flow of major weapons — artillery, missiles, military aircraft, tanks, and the like. Maintained by the Stockholm International Peace Research Institute (SIPRI), the database contains documented sales since 1950 and is updated annually. SIPRI provides a download tool, which outputs rich-text files, but it's also possible to download the data as CSV.

Gun dealers use the FBI's National Instant Criminal Background Check System to determine whether someone is allowed to buy a firearm. There isn't a one-to-one correlation between these background checks and gun sales, but they're said to be the best available proxy. The FBI publishes a PDF tallying the monthly number of firearm checks for each state and type.

Gun origins. The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) helps trace guns — such as those recovered at crime scenes by law enforcement agencies — back to their original manufacturers, wholesale distributors, dealers, and purchasers. Each year, ATF publishes a range of datasets based on these gun traces. The datasets for 2016 provide state-by-state tallies of gun caliber, state of original purchase, possessors' age, associated crime, and more.

Stolen guns. Missing Pieces is "a yearlong investigation by The Trace and more than a dozen NBC TV stations [that has] identified more than 23,000 stolen firearms recovered by police between 2010 and 2016 — the vast majority connected with crimes." To support the investigation, the reporters obtained more than 800,000 records of stolen and recovered guns, which they've standardized into a single CSV file and supplemented with a data dictionary. The dataset "contains nearly complete stolen-gun records for the states of California and Florida, both of which have centralized collections of gun-theft data," as well as records from nearly 300 other agencies across the country.

State gun laws. A team of researchers at the Boston University School of Public Health has collected data on the presence/absence of 133 different types of firearm laws in each U.S. state, for each year between 1991 and 2016. The legal provisions are grouped into 14 categories, such as background checks, "Stand Your Ground" laws, and child access prevention.

12 Development economics

Farmers in Africa. Between 2002 and 2004, researchers surveyed more than 9,500 farming households in 11 African countries to better understand how climate change might affect agricultural practices. They published the detailed results and documentation in Scientific Data. The dataset includes responses to questions about plantings, harvests, yields, water sources, animal purchases, taxes paid, and much more.

Data on the "Progresa" conditional cash transfer. Oportunidades (English: Opportunities) (now rebranded as Prospera) is a government social assistance (welfare) program in Mexico founded in 2002, based on a previous program called Progresa, created in 1997. It is designed to target poverty by providing cash payments to families in exchange for regular school attendance, health clinic visits, and nutrition support. Oportunidades is credited with decreasing poverty and improving health and educational attainment in regions where it has been deployed.

Nepal, post-earthquake. In April 2015, the Ghorkha Earthquake killed more than 8,000 people in Nepal, and destroyed hundreds of thousands of homes. In early 2016, a team led by the not-for-profit Kathmandu Living Labs, in collaboration with Nepal's government, undertook "a massive household survey using mobile technology to assess building damage in the earthquake-affected districts." The responses to that survey are now available at the 2015 Nepal Earthquake Open Data Portal; you can explore the data online or download it in bulk. In all, the datasets include details on millions of individuals, plus information about each surveyed household and building.

Famine warnings. "Created by USAID in 1985 to help decision-makers plan for humanitarian crises," the Famine Early Warning Systems Network (FEWS NET) "provides evidence-based analysis on some 34 countries." As part of its work, FEWS NET publishes geospatial shapefiles that score each country's "most likely food security outcome" on standardized scale: Minimal, Stressed, Crisis, Emergency, and Famine.

Development projects and outcomes. Earlier this year, Johns Hopkins professor Dan Honig released the Project Performance Database, which tracks the outcome ratings of international development projects (typically conducted by auditors on a four- or six-point scale). "The PPD is, at present, the world's largest" such database and "contains over 14,000 unique projects from eight agencies," including the World Bank, the Asian Development Bank, and others.

Social assistance programs The Social Assistance, Politics and Institutions database, developed at an United Nations University research center, "provides a synthesis of longitudinal and harmonized comparable information on social assistance programmes in developing countries, covering the period 2000-2015." For each program, such as Brazil's "Bolsa Familia," the database describes its basic characteristics, budget and financing, and population coverage. [

12.1 Foreign aid

U.S. foreign assistance. USAID, the Peace Corps, the U.S. African Development Foundation, and other agencies report data on foreign assistance spending to ForeignAssistance.gov. The full dataset includes detailed information for each grant and contract — and comes with data dictionary.

Global foreign aid. AidData, an organization based at the College of William has compiled a dataset of more than 1.5 million foreign aid projects between 1947 and 2013. Together, the dataset accounts for more than $7 trillion in commitments from 96 donors such as the U.S. government, UNICEF, the Nordic Development Fund, and the World Bank. AidData also publishes geospatial datasets and a data user guide.

International aid for maternal and child health. Researchers at the World Health Organization have assembled a dataset of international aid — both from official government assistance and private grants — devoted to reproductive, maternal, newborn, and child health from 2003 to 2013. The dataset, which the researchers described in a recent academic article, draws on 2.1 million records, and is based largely on the OECD's Creditor Reporting System.

12.2Migration and refugees

International labor treaties. Bilateral labor agreements regulate the migration of workers between two countries, and the Bilateral Labor Agreements Dataset aims to catalog as many of these treaties as it can. So far the University of Chicago Law School professors and researchers running the initiative have identified 582 treaties signed between 1945 and 2015. "However, this list is almost certainly underinclusive," they write. "Many BLAs are not deposited in the major international treaty databases and they often do not receive much, if any, publicity."

Refugee arrivals along the Western Balkans route. The UN's refugee agency is keeping track of daily refugee movements through Greece, Macedonia, Serbia, and farther along into Europe.

The Department of State publishes demographic reports on refugee arrivals since 2002. The data includes country of origin, resettlement city and state, religion, age, gender, and more.

Three decades of immigration policies. The Immigration Policies in Comparison (IMPIC) project has quantified the immigration regulations of 33 OECD countries between 1980 and 2010. The project, led by political sociologist Marc Helbling, dives deeply into the regulations related to four policy areas: labor migration, family reunification, asylum/refugees, and "co-ethnics." You can find the dataset's detailed codebook and methodology in this PDF

Immigrants, internationally. The United Nations publishes estimates of the number of foreign-born residents living in every country. The figures cover 1990 to 2015, at five-year intervals. The Vatican (100% foreign-born) and the United Arab Emirates (88%) had the highest proportion of immigrant residents in 2015; the U.S. (46.6 million) boasted the largest total immigrant population. The dataset also includes estimates by age, sex, and country of origin.

Rohingya refugees. The Humanitarian Data Exchange has collated dozens of datasets related to the Rohingya refugee crisis. Among them: the geographic boundaries of Rohingya refugee settlements in Bangladesh, the numbers of refugees living in those settlements, and the infrastructure available there.

34,361 European migration deaths. The Amsterdam-based activist group UNITED for Intercultural Action has, since the early 1990s, been collecting information about the deaths of Europe's refugee-seekers. The organization's volunteers "update the data annually, spending six months at a time verifying reports, categorising deaths and entering them into the database," according to The Guardian's story about the endeavor and its findings. "When the project began, they received physical clippings from a network of groups around Europe. Now, the data is collected from email submissions and Google Alerts in a number of languages." The story features a PDF-listing of the deaths, including the date the migrants were found dead, names and countries of origin (where known), and the causes of death. The Italian civic-data organization OnData has converted the PDF to a spreadsheet

Who receives H1-B visas?

12.3Slavery

The transatlantic slave trade. Slate Magazine's "The Atlantic Slave Trade in Two Minutes" — recently named a multimedia finalist for the American Society of Magazine Editors' annual awards — tracks 20,528 transatlantic voyages over 315 years. The information comes via SlaveVoyages.org, which provides searchable, downloadable records of ships' and captains' names, regions where slaves were purchased and sent, and more.

U.S. slave populations, 1790–1860. For more than a century, the U.S. Census collected slave population figures. An assistant professor at George Mason University has aggregated that data, and mapped it.

New Orleans slave sales, 1856–1861. Dataset of 15,377 slave sales, culled from remarkably detailed official records. Data for each sale includes demographic information about the slaves, seller, and buyer; the price paid; payment method; and researcher notes.

13 Climate, geography, demography

Maps data: Map the administrative areas of all countries, at various levels.

Nightlights data: The files are cloud-free composites made using all the available archived DMSP-OLS smooth resolution data for calendar years. In cases where two satellites were collecting data - two composites were produced. The products are 30 arc second grids, spanning -180 to 180 degrees longitude and -65 to 75 degrees latitude. A guide to use night light data.

Shifting global borders. What did the world's political boundaries look like in 1945? The lines between Swedish counties in 1968? The U.S. states in 1865? Thenmap, an open-source API and mapping tool, answers these questions and more

Inter- and intra-national boundaries. The Database of Global Administrative Areas aims "to map the administrative areas of all countries, at all levels of sub-division." With 386,735 divisions and counting, "this is a never ending project, but we are happy to share what we have."

Natural disasters: EM-DAT contains essential core data on the occurrence and effects of over 22,000 mass disasters in the world from 1900 to the present day. The database is compiled from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies.

The PRIO-GRID dataset is a spatio-temporal grid structure constructed to aid the compilation, management and analysis of spatial data over time. It consists of quadratic grid cells that cover all terrestrial areas of the world.

Historical climate data. The National Centers for Environmental Information maintains more than 20 petabytes of data, it says. Among the most useful slices is the Global Historical Climatology Network's data, which aggregates reports on temperature, precipitation, wind, and more from tens of thousands of climate-monitoring stations around the world.

Four decades of U.S. air quality. The Environmental Protection Agency collects air quality samples from thousands of monitoring stations across the country. The resulting datasets, which go back to the 1980s, are available as daily files, annual files, and via an API. The monitored pollutants include ozone, carbon monoxide, sulfur dioxide, nitrogen dioxide, particulate matter, volatile organic compounds, and more.

Global rainfall. To create the most detailed measurements of global rainfall ever, researchers at UC Santa Barbara's Climate Hazards Group harmonize data from satellites and on-the-ground weather stations. The dataset, known as CHIRPS, stretches back more than 30 years and is freely available.

Hourly rainfall. Since 1997, the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) algorithm has used satellite imagery to estimate rainfall rates around the world. The system's hourly, daily, monthly, and annual estimates can now be explored online and downloaded.

Air quality. The World Health Organization released its latest update to the Global Urban Ambient Air Pollution Database, which now covers nearly 3,000 cities in 103 countries. For each city, the dataset includes annual average density of two key categories of particulates (PM2.5 and PM10), as well as details regarding the data collection.

According to the organization's own analysis, "98% of cities in low and middle income countries with more than 100,000 inhabitants do not meet WHO air quality guidelines.

The United States of Land. In 2011, agriculture occupied about 22% of all land in the contiguous U.S., according to the National Land Cover Database. The NLCD classifies every 30-meter-by-30-meter chunk of land into one of 16 categories, including "woody wetlands," "cultivated crops," and "developed" land, at different intensities. (Alaska's unique landscape has earned it a few additional categories, such as "dwarf scrub.") The database is presented as raster files, so you'll need some geospatial software to dig in.

Where clouds congregate, and when. Researchers have analyzed 15 years of satellite imagery to create a nearly-global dataset of seasonal cloud coverage. The data — available at a kilometer-square resolution — could help scientists monitor and predict changes in ecosystems.

Provincial populations. National population data is easy to find. But it's much harder to find reliable, standardized population figures for finer-grained geographies. To that end, the World Bank has launched a pilot of its Subnational Population Database, which calculates estimates for 75 countries' major provinces/states/regions.

Post-disaster aerial imagery. After major natural disasters, NOAA's National Geodetic Survey routinely collects detailed aerial photos of the affected areas. For each disaster — including Hurricane Harvey, Hurricane Irma, and a couple dozen others — you can download the full set of (georeferenced) images, by date and survey flight.

Flood maps. FEMA's Flood Map Service Center publishes geospatial files that detail the agency's flood risk assessments — both current and historical. The maps include flood zones, levee locations, "base flood elevations,"

Global flooding. The Dartmouth Flood Observatory's Global Archive of Large Flood Events contains data about 4,500+ floods, dating back to 1985. It's updated often, and is available in Excel, XML, HTML, and geospatial formats. The variables include each flood's location, timespan, severity, main cause, and estimated impact. The organization also publishes detailed maps of the "maximum observed flooding" for specific disasters, such as for Hurricane Harvey and for Hurricane Irma.

Local UV exposure. The National Cancer Institute has estimated ultraviolet radiation exposure estimates for every county in the continental United States. The estimates, based on a peer-reviewed methodology and 30 years of data from the National Solar Radiation Data Base, can also be explored using the institute's mapping tool. Luna County, New Mexico had the highest estimated UV exposure at 5,723 Watt-hours per square meter; Clallam County, Washington, was exposed to the least estimated UV radiation, at 3,012 Wh/m^2.

Tropical cyclones. Through its International Best Track Archive for Climate Stewardship project, the National Oceanic and Atmospheric Administration publishes what it calls "the most complete global set

of historical tropical cyclones available." For each tropical cyclone — a category that includes typhoons, hurricanes, tropical depressions, and more — the dataset includes its position, wind speed, central pressure, and classification at six-hour intervals. The dataset is updated annually and includes some historical cyclones from as early as 1842.

U.S. wildfire costs. Stanford University's Big Local News project has compiled data from 100,000+ daily situation reports (known as "SIT-209"s) filed by federal firefighting authorities, detailing their efforts to suppress large wildfires. The dataset covers 2014 to 2017, and includes 240+ variables from each report, including estimated costs, damaged/destroyed buildings, injuries, fatalities, and more

The European Center for Medium-Range Weather Forecasts has an impressive looking collection of weather data.

Data set of every recorded meteor impact on Earth from 2500 BCE to 2012.

EU Climate Change Mitigation Policies. This dataset contains a number of climate change mitigation policies and measures (PAM) implemented or planned by European countries to reduce greenhouse gas emissions.

The Western hemisphere. The GOES-16 satellite was launched into orbit in November 2016, and it's been collecting near-realtime images and data ever since. (GOES stands for "Geostationary Operational Environmental Satellite.") It collects data on 16 different spectral bands, and it can capture a full image of the Western Hemisphere every 15 minutes, plus "an image of the Continental U.S. every five minutes, and two smaller, more detailed images of areas where storm activity is present, every 60 seconds." You can browse the images and data online, and also download them as NetCDF files.

Natural disaster satellite imagery. DigitalGlobe's open data program publishes georeferenced satellite imagery from before and after major natural disasters. The archive currently includes a couple dozen events, including recent flooding in Kerala and California's Carr Wildfire and Mendocino Complex Fire.

14 Text as data

Twitter API data: For text mining purposes.

Google Ngram: The Google Ngram Viewer is an online search engine that charts the frequencies of any set of comma-delimited search strings using a yearly count of n -grams found in sources printed between 1500 and 2008 in Google's text corpora in English, Chinese (simplified), French, German.

Google trends: Google Trends is a public web that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages. Example:

Word associations. The Small World of Words project "is a large-scale scientific study that aims to build a mental dictionary or lexicon in the major languages of the world." The experiment has asked hundreds of thousands of participants to list their immediate associations with various words (such as "telephone," "journalist," and "yoga"). In all, the project has collected more than 15 million responses. You can download the data, examine the project's analysis pipeline, and explore the responses online.

Enron corpus. During the course of its Enron investigation, the Federal Energy Regulatory Commission obtained the emails of approximately 150 (mostly high-ranking) Enron staff. You can find versions of the dataset — cleaned, deduplicated, and restructured in various ways — hosted by Carnegie Mellon, UC Berkeley, and Duke Law.

Dark Net Market archives, 2011-2015. Researcher Gwern Branwen has assembled an archive of listings posted to "dark net markets". Silk Road is the best-known among the group, but the collection covers scores of other markets, including Amazon Dark and FreeBay. The materials gathered from each site are slightly different; many include product advertisements and seller profiles. Warning: Some of the archives contain pictures, which may include offensive or disturbing imagery. And it's probably wise to heed Gwern's caveats: The scrapes "are large, complicated, redundant, and highly error-prone. They cannot be taken at face-value."

Between 2008-2012, what was being downloaded from The Pirate Bay?

In 2011, the New York Public Library launched a crowdsourcing project to transcribe its massive collection of restaurant menus, dating back to the 1850s. So far, volunteers have transcribed more than 1.3 million dishes, their prices, and where on the menu each dish appeared. The library publishes a spreadsheet of all the data, and updates it twice a month.

568,454 reviews of "fine foods" on Amazon. In 2013, Stanford University researchers published a paper examining how people's tastes "change and evolve over time." They drew, in part, on a dataset containing 13 years of Amazon reviews of gourmet foods. (Note: Not all foods were intended for humans.). The dataset comes in a slightly unconventional format; here's a Python script to convert it to a TSV file.

Word-emotion associations. Computational linguists at Canada's National Research Council used Mechanical Turk to crowdsource the emotional associations of 14,182 words. For each word, participants were asked whether it was "positive" and/or "negative", and whether it was associated with any of eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The resulting Word-Emotion Association Lexicon was first published in 2010.

Five years of Facebook posts from 15 news sites. Data analyst Patrick Martinchek has published a dataset of all Facebook posts from "15 of

the top mainstream media sources" — a group that includes The New York Times, The Wall Street Journal, NPR, Fox News, and other familiar sources — from January 2012 through Nov. 8, 2016.

Fake news on Facebook. 1,000+ posts from hyperpartisan Facebook pages.

Millions of Amazon reviews. Julian McAuley, an assistant professor at UC San Diego, has collected a massive amount of user-generated data from Amazon.com, including 142.8 million reviews and 1.4 million answered Q&As. Stanford has 35 million Amazon reviews available for download.

The Gray Lady of 19th century Havana. The University of Miami Libraries has digitized 53,000+ pages of La Gaceta de La Habana, "the paper of record during the Spanish colonial occupation of Cuba in the nineteenth century." The digitized editions span 33 of the years between 1849 and 1897.

The geography of language on Twitter. Quartz published an addictive tool that lets you map word usage on Twitter, by U.S. county. It's based on an academic analysis of 890 million geocoded tweets uttered between October 2013 and November 2014

The most important entries on Wikipedia. Germany-based researcher Andreas Thalhammer has applied PageRank — the algorithm at the heart of Google's origin story — to the world of Wikipedia. The result: the DBpedia PageRank dataset, which estimates the importance of each page based on the other pages that link to it. You can download the data directly, or query it online.

The Wikimedia Foundation publishes hourly pageview counts for each of its articles. It's a tremendous amount of data — about 90 megabytes, compressed, per hour. Luckily, there's also a tool for browsing individual pages' daily traffic stats.

All of Wikipedia is freely available, DBpedia is an attempt to synthesize it into a more structured format. Just what kind of edits do people usually make on Wikipedia? You can figure it out with this data set.

There's been a fair bit of controversy around all the bureaucracy of Wikipedia. Who's the ideal Wikipedia administrator? Well, they're voted for, and the data is available for download.

Words kids learn. Wordbank is an "open database of children's vocabulary development." So far, the Stanford-hosted project has gathered data from more than 71,000 standardized and anonymized vocabulary questionnaires across 23 languages. You could spend hours exploring the data online, charting how quickly children learn individual words, how quickly the same word (e.g., "grandma," "abuela," "бабушка") is learned in different languages, and connections between words. You can download the data for each word or for each child's vocabulary.

All 7.3 million StackOverflow questions.

Yelp has a freely available subset of their data, including restaurant rankings and reviews.

You can download all the PDFs on Arxiv

Someone has scraped the top 2.5 million Reddit posts and then placed them on GitHub.

Three million grocery orders. Groceries-on-demand startup Instacart has released a dataset containing 3 million orders from 200,000 (anonymized) users. "For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order," the company's head of data science writes. "We also provide the week and hour of day the order was placed, and a relative measure of time between orders."

Presidential Debate Tweets: 2000 tweets immediately following the first Presidential Debate in September 2016

15 Networks

Zachary's karate club: social network of friendships between 34 members of a karate club at a US university in the 1970s.

Les Miserables: coappearance network of characters in the novel *Les Miserables*.

Word adjacencies: adjacency network of common adjectives and nouns in the novel *David Copperfield* by Charles Dickens.

American College football: network of American football games between Division IA colleges during regular season Fall 2000.

Dolphin social network: an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand.

Political blogs: A directed network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance.

Books about US politics: A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers.

Neural network: A directed, weighted network representing the neural network of C. Elegans. Data compiled by D. Watts and S. Strogatz and made available on the web here.

Power grid: An undirected, unweighted network representing the topology of the Western States Power Grid of the United States.

Condensed matter collaborations 1999: weighted network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive between Jan 1, 1995 and December 31, 1999.

Condensed matter collaborations 2003: updated network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive. This version includes all preprints posted between Jan 1, 1995 and June 30, 2003. The largest component of this network, which contains 27519 scientists, has been used by several authors as a test-bed for community-finding algorithms for large networks.

Condensed matter collaborations 2005: updated network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive.

Astrophysics collaborations: weighted network of coauthorships between scientists posting preprints on the Astrophysics E-Print Archive between Jan 1, 1995 and December 31, 1999.

High-energy theory collaborations: weighted network of coauthorships between scientists posting preprints on the High-Energy Theory E-Print Archive between Jan 1, 1995 and December 31, 1999.

Coauthorships in network science: coauthorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006. A figure depicting the largest component of this network can be found here.

Internet: a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project.

UCINet data sets: Social network data sets released with the UCINet software by Steve Borgatti *et al.*

Pajek data sets: Example data sets released with the Pajek software by Vladimir Batagelj and Andrej Mrvar.

Indiana University data sets: A set of very large data sets, including some non-network data sets, compiled by the School of Library and Information Science at Indiana University. Network data sets include the NBER data set of US patent citations and a data set of links between articles in the on-line encyclopedia Wikipedia.

Duncan Watts' data sets: Data compiled by Prof. Duncan Watts and collaborators at Columbia University, including data on the structure of the Western States Power Grid and the neural network of the worm *C. Elegans*.

Laszlo Barabasi's data sets: Data compiled by Prof. Albert-Laszlo Barabasi and collaborators at the University of Notre Dame, including web data and biochemical networks.

Alex Arenas's data sets: Data compiled by Prof. Alexandre Arenas and collaborators at Universidad Rovira i Virgili, including metabolic network data and the network from their study of the collaboration patterns of jazz musicians.

Stanford Large Network Dataset Collection: A substantial collection of data sets describing very large networks, including social networks, communications networks, and transportation networks.

Social networks: Online social networks, edges represent interactions between people.

Networks with ground-truth communities: Ground-truth network communities in social and information networks.

Communication networks: Email communication networks with edges representing communication.

Citation networks: Nodes represent papers, edges represent citations.

Collaboration networks: Nodes represent scientists, edges represent collaborations (co-authoring a paper).

Web graphs: Nodes represent webpages and edges are hyperlinks.

Amazon networks: Nodes represent products and edges link commonly co-purchased products.

Internet networks: Nodes represent computers and edges communication.

Road networks: Nodes represent intersections and edges roads connecting the intersections.

Autonomous systems: Graphs of the internet.

Signed networks: Networks with positive and negative edges (friend/foe, trust/distrust).

Location-based online social networks: Social networks with geographic check-ins.

Wikipedia networks, articles, and metadata: Talk, editing, voting, and article data from Wikipedia

Temporal networks: Networks where edges have timestamps.

Twitter and Memetracker: Memetracker phrases, links and 467 million Tweets.

Online communities: Data from online communities such as Reddit and Flickr.

Online reviews: Data from online review systems such as BeerAdvocate and Amazon.

Lada Adamic: network of hyperlinks between weblogs on US politics.

Arxiv's High Energy Physics paper citation network.

16 Behavioral economics

Pick a card, any card. When researchers asked 1,354 people to name or visualize a playing card, 1 in 6 of them first chose the Ace of Spades. Here's the data, which includes each participant's three card choices, age, and gender.

Psychometric tests. The Open Source Psychometrics Project "provides a collection of interactive personality tests with detailed results that can be taken for personal entertainment or to learn more about personality assessment." You can download results from more than 30 such tests, including the Big Five Personality Test, the Kentucky Inventory of Mindfulness Skills, and Bob Altemeyer's Right-wing Authoritarianism Scale.

Prime psychology. Contest: Try to choose the smallest prime number that nobody else will pick. Now he's posted the results — a CSV listing the number of contestants who chose each prime number.

Brain scans. The Open Access Series of Imaging Studies (OASIS) project is "aimed at making MRI data sets of the brain freely available to the scientific community," with the goal of "[facilitating] future discoveries in basic and clinical neuroscience." So far, the project has published two collections: a cross-sectional dataset of scans from 416 people, ages 18 to 96; and a longitudinal dataset, based on 150 people aged 60 to 96, each of whom were scanned at least two different times.

More Brain scans. The Stanford Center for Reproducible Neuroscience launched OpenNeuro, "a free and open platform for analyzing and sharing neuroimaging data." (It's the successor to the center's earlier initiative, OpenfMRI.) You can, for instance, download scans of brains that were watching a particular episode of The Twilight Zone.

Software development time estimates. Derek M. Jones analyzes software-engineering data. Recently, he convinced a small software company to release a dataset documenting its internal time estimates, spanning 10 years, 20 projects, and 10,000+ tasks. For each task, the dataset indicates the number of hours it was predicted to take, how long it actually took, the (anonymized) developers it was assigned to, and more

Confidence. The Confidence Database is aggregating data from behavioral studies that have asked participants' how confident they were in their own assessments. As of its launch earlier this month, the database contains 145 datasets, 8,700 participants, and 4 million individual observations.

Game show gambles. To study how people make decisions in risky situations, a team of academics analyzed contestants' choices in 100+ episodes of Deal or No Deal that aired in the Netherlands, Germany, and the US. Their dataset is available through ICPSR (registration required) and the Wayback Machine.

16.1 Genoeconomics

GET-Evidence has put up public genomes for download.

The 1000 Genomes project has made 260 *terabytes* of genome data downloadable.

17 Entertainment industry

Buenos Aires Marathon. Every single racer's result (including names!), 2006-2016.

Tens of millions of movie ratings. MovieLens.org is a free, noncommercial movie recommender — sort of like Netflix, minus the ability to watch movies. The service is run by a research lab at the University of Minnesota. The lab publishes several datasets of user ratings and movie info. The largest contains 22 million ratings. Among movies with at least 1,000 ratings.

Famous people on Wikipedia. A group of researchers introduced Pantheon 1.0, "a manually verified dataset of globally famous biographies." It starts with 11,341 Wikipedia biography pages in 25 languages, and adds birthplace, birthdate, gender, occupations, and page views. You can download the data or explore it online.

Movie chatter. The Cornell Movie-Dialogs Corpus contains 220,579 "conversational exchanges" between 9,035 characters in 617 movies.

NBA refereeing. Since March 2015, the National Basketball Association has issued post-game reports reviewing referees' calls during the final two minutes of neck-and-neck games. The NBA publishes those reports as PDFs; journalist Russell Goldenberg has been converting them to spreadsheet-friendly CSVs. Goldenberg is also analyzing and visualizing the data — updated daily — to show, for example, which players are benefitting most from incorrect and missed calls.

TV talk. The GDELT Project and the Internet Archive have collaborated to make the latter's Television News Archive more powerfully searchable. Their new tool, announced in December, lets you search across "more than 5.7 billion words from over 150 distinct stations spanning July 2009 to present" at a sentence-by-sentence level. The results are downloadable as CSV or JSON files.

A decade of TV news words. The TV-NGRAM project pulls 14 TV stations' data from the Television News Archive and calculates how often each word (and two-word combination) was said during each 30-minute window. Most of the stations' counts go back 9 or 10 years, and all are updated daily.

Powerlifting. OpenPowerlifting.org "aims to create a permanent, accurate, convenient, accessible, open archive of the world's powerlifting data. In support of this mission, all of the OpenPowerlifting data and code is available for download in useful formats." So far, that includes 400,000+ performances at 9,000+ competitions in dozens of countries.

Speaking roles in 2016's blockbusters. Researcher Amber Thomas has parsed the transcripts of last year's 10 highest grossing films. The resulting data files indicate each character's number of turns speaking, number of words spoken, and gender.

Story arcs. "The WikiPlots corpus is a collection of 112,936 story plots extracted from English language Wikipedia." The plots describe movies, books, plays, TV series, TV episodes, video games, and other stories — essentially, any *thing that has a Wikipedia article with the word "plot" in one of its subheadings.

Top 100 Rotten Tomatoes Movies: Movies with 40 or more critic reviews vie for their place in history at Rotten Tomatoes. Eligible movies are ranked based on their Adjusted Scores.

TED Talks Dataset: Master list of 2,600 Ted Talks and descriptions.

Top 500 Albums: Dataset of Rolling Stone's 500 greatest albums of all time.

17.1Media

5,000+ Brazilian news outlets. Atlas da Notícia is a Brazilian project that aims to collect data on all local and regional news outlets in the country. The project released its first batch of data, which identified 5,354 newspapers and online publications in a total of 1,125 municipalities. The raw dataset is currently only available in Portuguese, but the aggregate tables have been translated into English.

Venmo transactions. Dan Salmon, a grad student who specializes in information security, has published data on more than 7 million Venmo transactions, which he downloaded from the mobile payment platform's public API. "I am releasing this dataset," he writes, "in order to bring attention to Venmo users that all of this data is publicly available for anyone to grab without even an API key."

Media coverage. Media Cloud, a collaboration between MIT and Harvard-based researchers, describes itself as "an open-source platform for studying media ecosystems." The project lets you track topics and keywords across thousands of sources — including mainstream news publications in the U.S. and many other countries — at both a story and sentence level. You can access Media Cloud's data via its dashboard or its API

Chyrons. The TV News Archive's new "Third Eye" project is extracting chyrons — those placards of text at the bottom of news broadcasts, also known as "lower thirds" — from four major cable networks: BBC News, CNN, Fox News, and MSNBC. The resulting database contains every chyron

that Third Eye's optical character recognition (OCR) software has extracted since late August.

Historic newspapers. Chronicling America — a project run by the Library of Congress and the National Endowment for the Humanities — provides information about more than 150,000 historic newspapers and access to digitized pages from many of them. Its API lets you search the database and doesn't require registration; its bulk data includes text from more than 12 million pages.

A decade of New York Times front-page stories. UC Davis professor Amber E. Boydstun oversaw the compilation of a dataset of every front-page article in the New York Times from 1996 to 2006. Each of the 31,034 articles have been categorized by topic, according a detailed codebook, and given a short summary.

Journalists killed, imprisoned, and missing. The Committee to Protect Journalists maintains a database of journalists who've been killed for reasons related to their work. The database goes back to 1992 and contains more than 1,300 entries, with details about the journalists, the circumstances of their deaths, and whether perpetrators have been convicted. More recently, the organization has also begun publishing data on journalists who've been imprisoned or gone missing.

17.2Culture and Language

The World Atlas of Language Structures. This database compares the phonological, grammatical, and lexical properties of hundreds of languages. One dataset looks at languages' counting systems. (Many use the decimal system, but Yoruba uses the vigesimal system and Danish uses a hybrid.)

Every place name in the United States. Since 1890, the U.S. Board on Geographic Names has been cataloguing, standardizing, and promulgating official names for the places we hike, swim, work, and call home. Along the way, it began publishing Geographic Names Information System (GNIS), a searchable and downloadable database containing all of its domestic nomenclature.

Folktales. The Aarne-Thompson-Uther Classification of Folk Tales organizes (mostly Indo-European) folktales into groups and hierarchies. As Atlas Obscura's Cara Giaimo puts it, the ATU is "like the Dewey Decimal System, but with more ogres." The ATU doesn't publish any downloadable versions of its data, but researchers studying the "ancient roots" of such stories have built a data-matrix that denotes the presence/absence of the 275 ATU "tales of magic" across 50 Indo-European-speaking populations.

The Survey of Scottish Witchcraft. The University of Edinburgh hosts an incredibly detailed, and deeply documented database of more than 3,000 accused witches in Scotland. The mania reached its quantitative peak in 1662, when, according to the database, 402 people were accused of witchcraft.

Prisoners' tattoos. The Florida Department of Corrections' public database contains a table describing current and released inmates' tattoos. That data includes each tattoo's location (e.g., "right arm," "stomach," "face") and description ("cross," "tribal," and "skull" being the most common).

17.3 Religion

Sister, Sister. In the wake of the Second Vatican Council in the 1960s, Sister Marie Augusta Neal conducted an enormous opinion survey of Catholic "women religious." More than 130,000 sisters responded to the 649 multiple-choice-question survey — the results of which the University of Notre Dame recently cleaned up and made available online.

Religion in America. The 2010 Religious Congregations and Membership Study counts, for more than 200 religious groups, the number of congregations and adherents in each U.S. state and county. In total, the study reported more than 344,000 congregations and more than 150 million adherents — nearly half of the 2010 U.S. population.

18 Miscellaneous

Nuclear accidents. Researchers in Europe have published a database of 216 nuclear energy accidents — a compendium they say is "twice the size of the previous best data set." For each accident, the database contains the date, location, description, and four measurements of severity: its ratings on the International Nuclear Event Scale and on the Nuclear Accident Magnitude Scale, the number of fatalities, and total monetary cost. (The three most expensive: Chernobyl, Fukushima, and a 1995 accident at Japan's Monju Nuclear Power Plant, estimated to have caused $15.5 billion in damages.)

The Ku Klux Klan, 1915–1940. Scholars at Virginia Commonwealth University have identified and mapped the locations of 2,000 KKK branches active in the early 20th century. The dataset contains the city, state, earliest-known-date, and sources for each "klavern."

Sexualitics.org is on a mission: "to contribute to human sexuality understanding through a Big Data approach." The site has detailed metadata on 800,000 adult videos, including titles, descriptions, view counts, and tags. It powers Porngram, an only-kinda-safe-for-work charting tool.

Miscellany. The University of Florida's Larry Winner has collected hundreds of "miscellaneous" datasets, many from niche academic studies.