Doodling with IRS 990s and NIH Exporter Data

Haskins Laboratories, a case study

David Braze

September 6, 2021

Executive Summary

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. In id erat non orci commodo lobortis. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Nunc rutrum turpis sed pede. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Nunc porta vulputate tellus. Sed bibendum. Donec at pede. Vivamus id enim.

Keywords: public data, non-profit organizations, Internal Revenue Service, IRS form 990, Propublica Nonprofit Explorer, National Institutes of Health, NIH Exporter, NIH grants, Haskins Laboratories

Contents

Preliminaries

This document is an exercise in working with data from two main sources. The first source is IRS form 990, "Return of Organization Exempt from Income Tax." The 990 is an informational form that non-profit organizations (NPO) must file with the Internal Revenue Service (IRS) each year (Internal Revenue Service 2020a). The completed forms are publicly available records as a matter of federal (law? policy?). The second source of information used here is details of individual grants from National Institutes of Health (NIH). The NIH Exporter gives access to information about all individual grants made by the NIH (National Institutes of Health n.d.). That information includes the direct and indirect moneys paid to grantees each year for each grant.

NIH grant data can be downloaded from the NIH Exporter web tool (National Institutes of Health n.d.). The data files are available for each fiscal year since 1985 in either CSV or XML format. A convenient place to get IRS 990s is from the ProPublica Nonprofit Explorer (Roberts et al. 2013). It is necessary to search for the particular NPOs you are interested in, and individually download form 990s for each tax year of interest. The form for a specific NPO/tax-year will be available either as a scanned PDF, if it was filed as a paper document, or as an XML file, if it was filed electronically. All NPOs are required to file electronically from 2019 forward.

In digging into what can be gleaned from these data sources, I will use the Haskins Laboratories as a case study. Why Haskins Labs? First, because there is both IRS data and NIH grant data available for the Labs. It is a 501(c)(3) non-profit organization, so is required to file an IRS 990 each year, and has been filing those documents electronically since 2014; all NPOs are required to file electronically starting in 2019. Electronic filing is significant for present purposes because it means that the 990 data is available in XML format, making it easy to access. Form 990s that have been filed as paper documents are only available as scanned PDF files, often not of very good quality. This makes data extraction error prone and tedious (but not impossible!). With regard to grant data, Haskins is a research organization and has received significant NIH research funding each year since at least 1970. So, there is a wealth of information about the grants that NIH has made to Haskins over the years available in the NIH ExPORTER database.

A second reason for using Haskins records in this exercise is more personal. I was a senior scientist on the Haskins research staff from 2002 until 2019. So, I have some curiosity about what nuggets of historical interest can be found in these public data sources. Certainly, there are other accounts of the lab's history which are both more detailed and more human than what can be gleaned from these federal databases (cf. Fowler and Shankweiler 2021, n.d.). As previously noted, information from documents filed with the IRS is only easily accessible from 2014 forward. As I write (September, 2021), the most recent IRS 990 available is for the 2019 tax year. NIH data is both more up to date and goes back further in time. Together, the IRS and NIH records might provide an interesting complement to other views of Haskins' history, but constructing an overview of Haskins history is not really my goal. I'm simply using my connection to the labs as one source of motivation for exploring these databases.

The real point of this exercise is simply to pick the lowest hanging fruit out of all that might be gleaned from these two sources. I am in no way trying for a comprehensive exploration of the data. The summaries (figures and tables) that I provide here are best thought of as rough and simple overviews of the source data. They are not tailored to any particular goal. One can easily imagine projects based on either of these data sources that involve in-depth examinations of how individual organizations change over time, or comparisons of groups of organizations, or both.

Summary of IRS i990s

Figure 1 shows annualized revenue, expenses, and net assets over the six year period where IRS data is available in the easily parsable XML format. Figure 2 shows those same values, with the addition of curves reflecting the decline in purchasing power of the dollar over time (dashed lines).

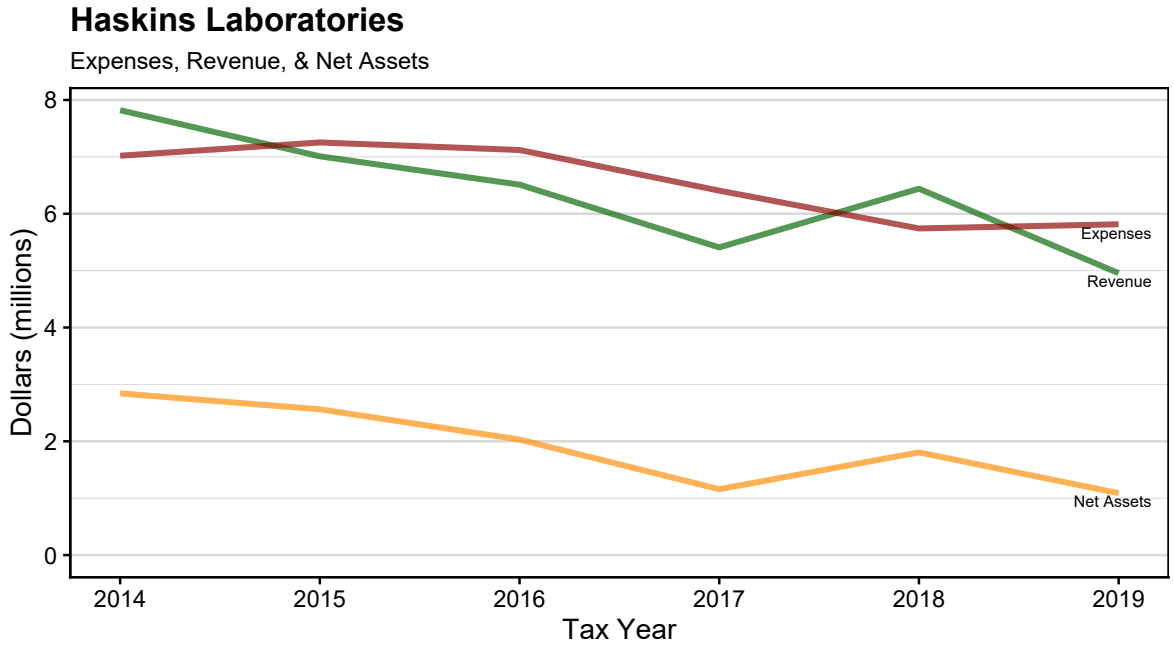

Figure 1: Data is taken from IRS form 990s. 'Revenue' is the total revenue line from the 990; 'Expenses is' total expenses; 'Net Assets' is total assets minus total liabilities.

In Figure 3, I've added a line to Figure 1 representing NIH grant money received by the labs. I've done this even though, as best I can determine, IRS annual data and NIH annual data use slightly different fiscal years. IRS 990 data corresponds to the calendar year; the IRS allows other options, but Haskins uses a calendar year for its filings. The NIH fiscal year runs from October through September and fiscal years are labeled according the year in which they end. Nonetheless, it is a useful comparison. Figure 3 also includes NIH data one year forward and and one year back from IRS data, just because. Keep in mind that the amounts represented by *NIH Total Costs* are already included in the *Revenue* curve. NIH Total Costs is the sum of direct + indirect funds for all NIH grants where Haskins is the primary grantee; subawards to Haskins, where the primary grant goes to another organization and Haskins is a sub-contractor, are not included in this quantity. It is clear that NIH Total Costs are a significant proportion of total revenue in each year, but the proportion varies considerably, and declines rather dramatically over the most recent 3 years.

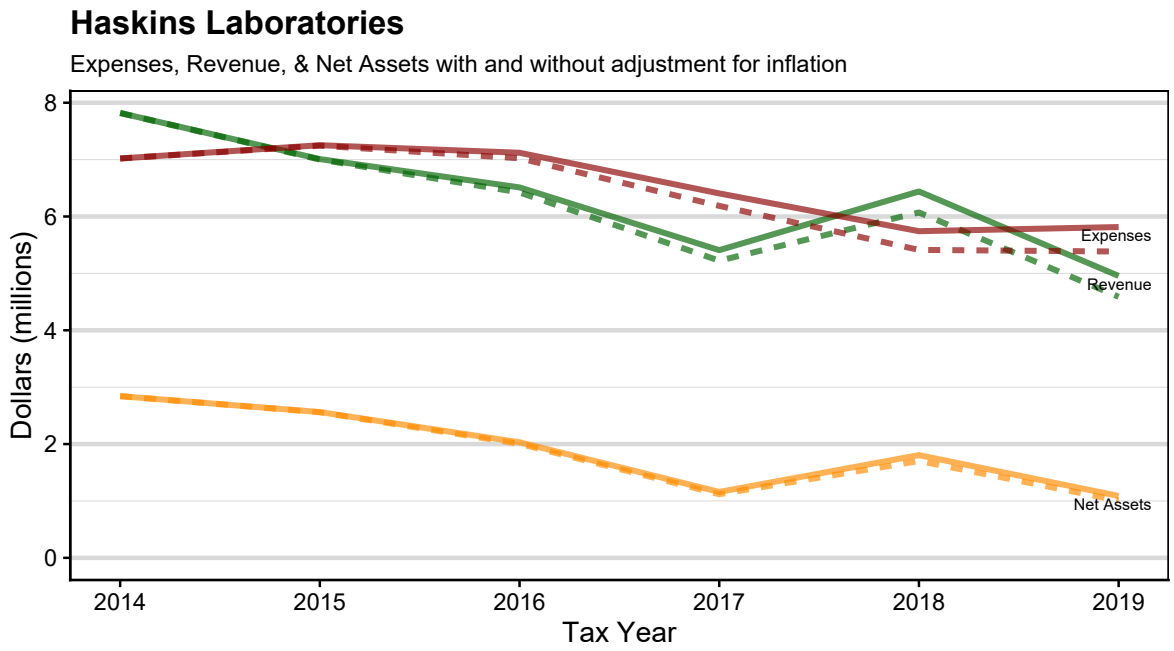

Figure 2: Data is taken from IRS form 990s. Solid lines reflect reported dollar values. Dashed lines are adjusted for inflation to equivalent 2014 purchasing power (using CPI data). 'Revenue' is the total revenue line from the 990; 'Expenses' is total expenses; 'Net Assets' is total assets minus total liabilities.

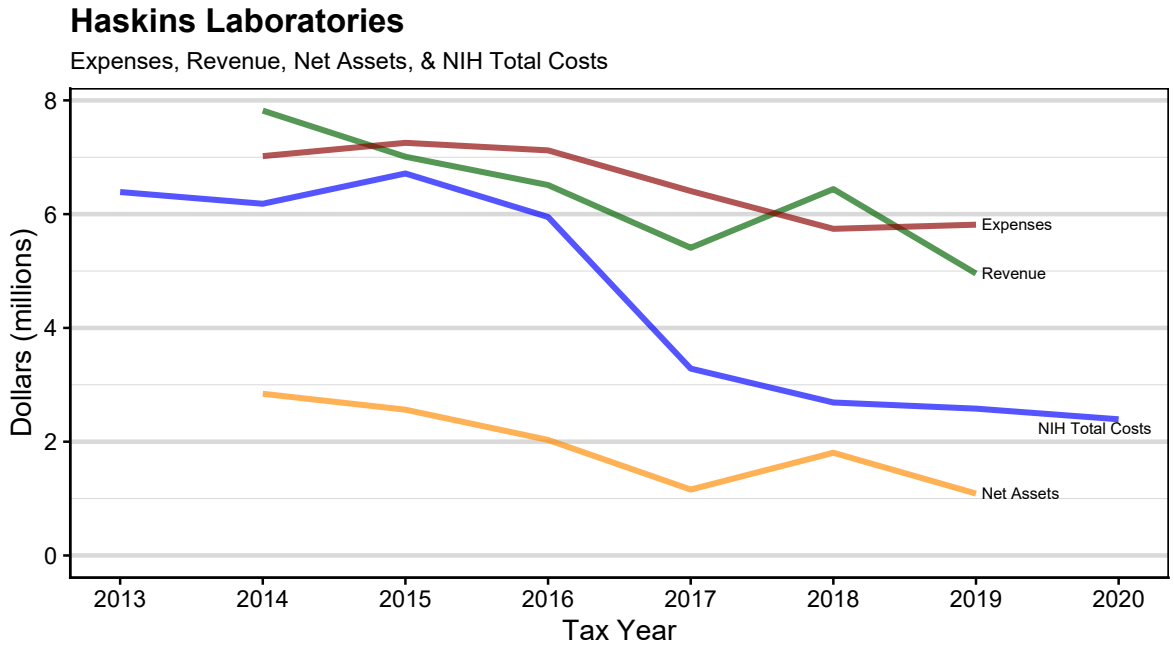

Figure 3: Revenue, expense, and asset figures are taken from IRS form 990s. NIH Total Costs comes from NIH Exporter. 'Revenue' is the total revenue line from the 990; 'Expenses is total expenses; 'Net Assets' is total assets minus total liabilities. 'NIH Total Costs' is direct + indirect costs for all NIH grants where Haskins is the primary grantee (subawards to HL are not included).

Haskins Revenue

IRS 990s identify revenue as belonging to any one of more than a dozen categories. Most of these are not used by Haskins. Figure 4 shows the breakdown of Haskins' revenue by category as reported in the labs' 990s. The maximum at each time-point in Figure 4 is equal to the *Revenue* quantities depicted in green in figures 1, 2, and 3.

I've also subdivided the IRS *government grants* category from the IRS 990 into two parts: NIH grant money that goes directly to Haskins as primary performance site, derived from NIH ExPORTER data, and 'other' grant money. NIH Total Cost in figure 4 is the same quantity shown in blue in Figure 3. The quantities shown in 4 as *Other Gov. Grants* is the difference between the *Revenue* and *NIH Total Costs* curves in Figure 3. Keep in mind the previously noted asynchrony in fiscal years for NIH vs. IRS sourced data.

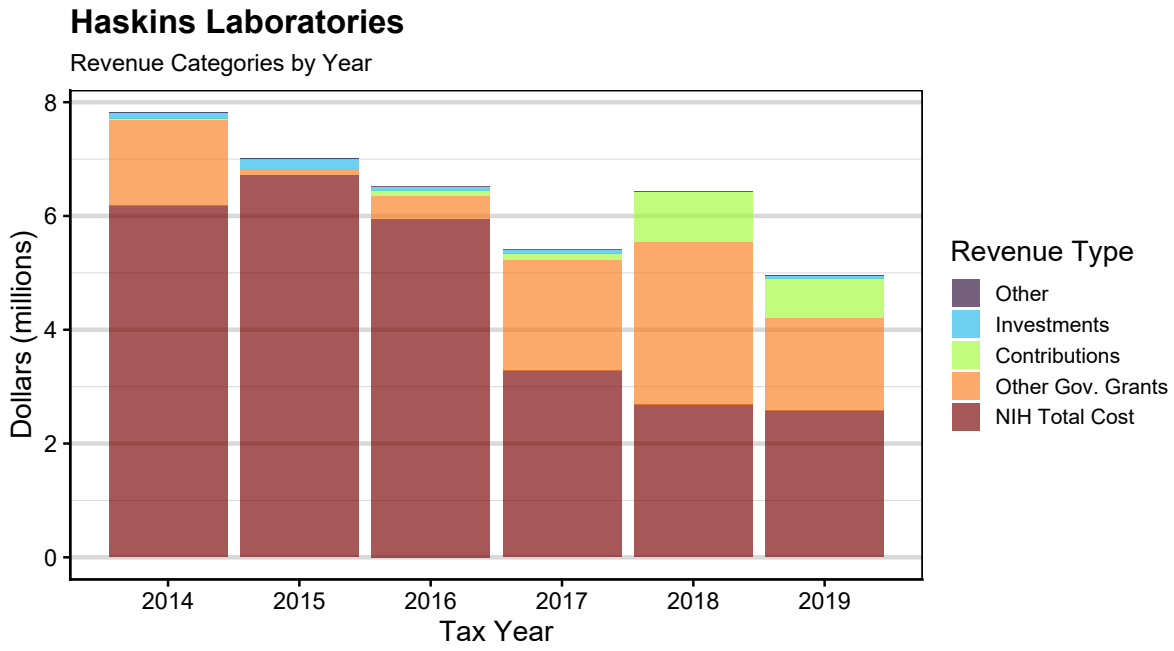

Figure 4: 'NIH Total Cost' is is direct + indirect funds for all NIH grants where Haskins is the primary grantee (subawards to HL are not included); 'Other Gov. Grants' is income from other government grants (including, I think, NIH subawards); other categories are self explanatory.

Haskins Expenses

Figure 5 shows the partitioning of Haskins Expenses by IRS defined category. Inspection of the figure makes clear that salaries are the single largest expense over the period under consideration. 'Other Expenses' is second largest. Grants from Haskins to others are a smaller but still non-trivial expense in each year. These are divided into two categories: Grants to Organizations, and 'Other Grants'. The first category seems fairly straightforward, but it is not clear to me what would fall into the second. Regardless, more information on Haskins grant-making activities can be found in Schedule I, for grants to US based organizations, and in Schedule F, for international grants (Internal Revenue Service 2020b, 2020c). Read on.

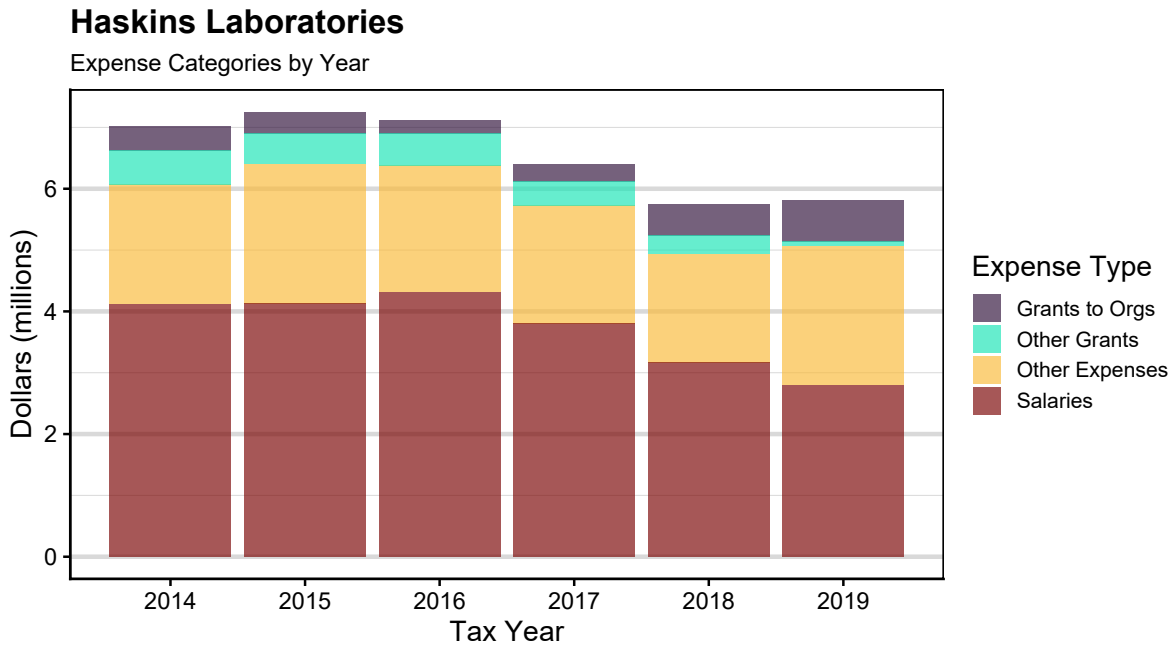

Figure 5: Four expense categories are in use by Haskins Laboratories over this six year period: Grants to Organizations, Other Grants, Other Expenses, and Salaries.

Grantees

US Grantees Information about Haskins grants (AKA subawards) to US-based organizations can be found in Schedule I (Internal Revenue Service 2020c). Figure 6 shows the yearly total of grants made by Haskins to US-based organizations each year, further broken down by recipient organization. Three of 10 grant recipients received money in all six years from 2014 through 2019 (CUNY, UCSF, Yale). Three received money in three of six years (NYU, UConn, USC). The remaining four organizations each got money from Haskins in just one of the six years.

International Grants Schedule F provides some information about grants going outside the United States (Internal Revenue Service 2020b). Sadly, the information provided is less detailed than that for US-based grantees. It does not include information about grantees as such, but only about the amount of each grant and the general geographic region where the grantee is located.

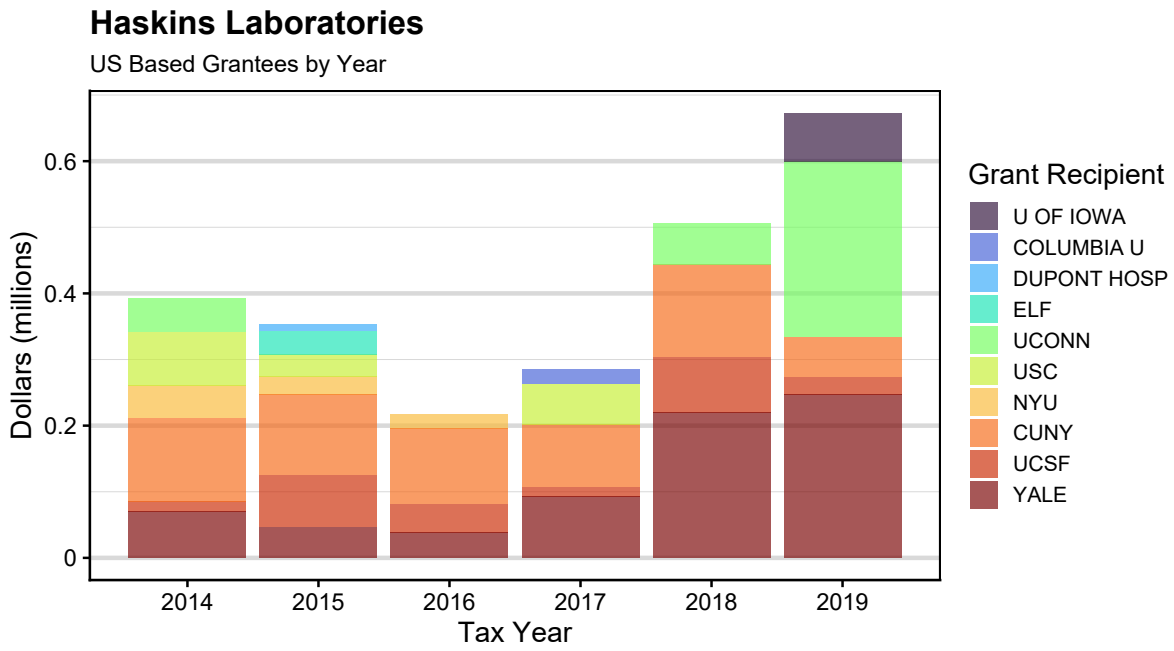

Figure 6: Yearly total of grants made by Haskins to US-based organizations, by recipient. Organization key: YALE = Yale University; UCSF = University of California at San Francisco; CUNY = Research Foundation of CUNY; NYU = New York University; USC = University of Southern California; UCONN = University of Connecticut; ELF = Endangered Language Fund; DUPONT HOSP = Dupont Hospital for Children; COLUMBIA U = Teachers College at Columbia University; U OF IOWA = University of Iowa.

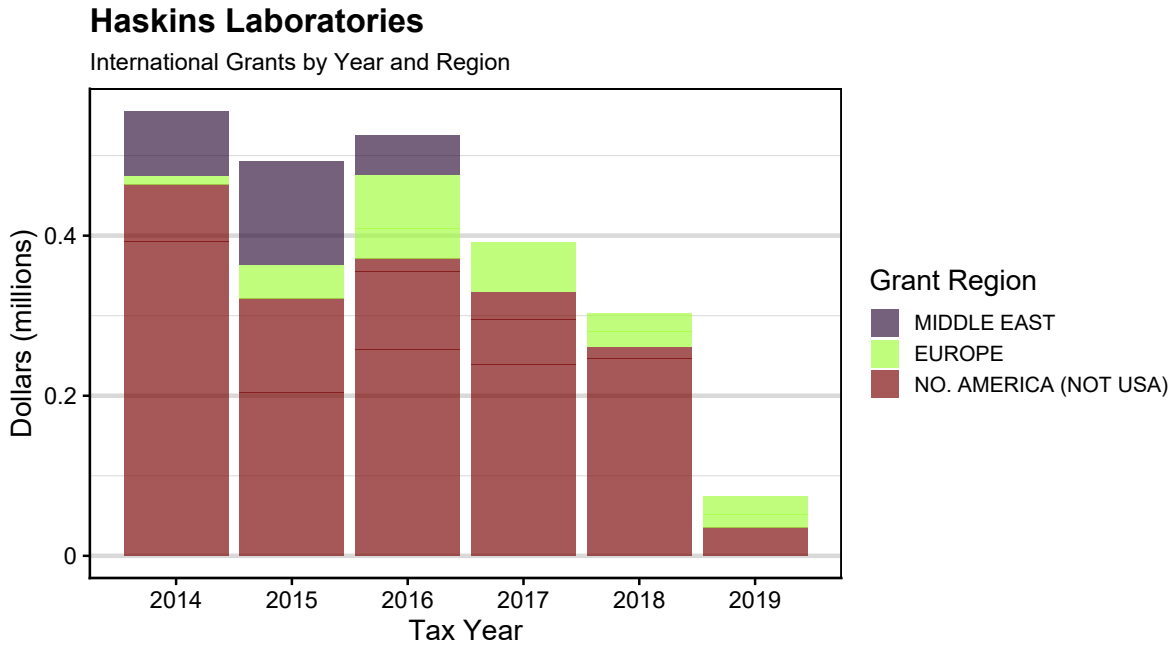

Figure 7: Yearly total of grants made by Haskins outside the USA, by region.

Directors, Officers, & Key Employees

Tables in this section are based on information from i990 Part VII, which identifies several categories of individuals having particularly significant roles within each NPO. These categories include: members of the board of directors, officers of the corporation, key employees, and highly compensated employees.

One of the challenges in dealing with named individuals in form 990s, of whatever category, is that when the same individual is mentioned across multiple years, their name is not always rendered consistently. One common issue is to do with middle names. Sometimes these are included in full, sometimes as initial only, sometimes they will be omitted altogether. Other variations do crop up as well. A similar problem exists for titles/roles. In the case either of names or of roles, if you want to make use of these as *data*, then there is a certain amount of work that must be done up-front to ensure that they are properly normalized.

Directors

Table 1 shows board members (Directors) over time and their roles on the board as listed on each year's IRS form 990. Note that, as is usual in non-profit organizations, the President/CEO of the organization is an *ex officio* member of the Board of Directors. So, Dr. Pugh is listed on both Table 1 and Table 2. P. Rubin and A. Abramson also appear in both tables. They each served terms as corporate officers and as members of the board, but their terms as directors and as officers did not overlap.

Corporate Officers

Table 2 shows Haskins Laboratories' corporate officers over time, and their roles, as listed on each year's IRS form 990. Note that Arthur Abramson passed away in December of 2017 (Whalen 2019), although he is still listed as an officer (Coporate Secretary) in subsequent years.

Office of the Connecticut Secretary of the State (HL is registered in CT) makes business filings publicly available on its website (e.g., https://service.ct.gov/business/s/onlinebusinesssearch?businessName=haskins%20laboratories). So with a little work, it will be possible to link and compare information from IRS 990s with information, including details about officers, from a business's filings with the SotS. Unfortunately, not all states are similarly transparent in making corporate filings so easily accessible.

Another possible source of linking information, one that may be more peculiar to Haskins Labs than it is of general utility, is the Open Payrolls website ("OpenPayrolls - the Largest Nationwide Salary Database" 2021). Open Payrolls aggregates payroll information for public employees at the local, state, and federal level. At least some of Haskins corporate officers serve in their roll at the labs as a side hustle, whereas their day jobs are as faculty members at public colleges or universities. So if an officer's Haskins salary, as reported on its IRS 990s, seems low, there may be more to see at Open Payrolls.

Highly Compensated Employees

In IRS terms, *highly compensated employees* (HCE) are those who are not officers of the corporation, but who received reportable compensation (i.e. regular salary, not including deferred or non-salary

Table 1: Board members over time and their roles on the board as listed on each year's form 990. CHAIR = chairperson; DIR = director; PCH = past chairperson; SEC = secretary of the board; TRE = treasurer.

	2014	2015	2016	2017	2018	2019
ARTHUR S ABRAMSON	SEC	–	–	–	–	–
CLAUDIA CARELLO	CHAIR	PCH	–	–	–	–
STEPHEN R ANDERSON	DIR	SEC	SEC	–	–	–
JOSEPH MOLDER	DIR	DIR	DIR	–	–	–
JEREMY TEITELBAUM	DIR	DIR	DIR	–	–	–
LEONARD KATZ	DIR	DIR	DIR	–	–	–
STEVEN M GIRVIN	DIR	DIR	DIR	DIR	–	–
DAVITA GLASBERG	–	–	–	DIR	DIR	–
WILLIAM H BAKER JR	DIR	DIR	DIR	DIR	DIR	DIR
DONALD SHANKWEILER	DIR	DIR	DIR	DIR	DIR	DIR
SHEILA E BLUMSTEIN	DIR	CHAIR	CHAIR	CHAIR	CHAIR	CHAIR
KENNETH R PUGH	DIR	DIR	DIR	DIR	DIR	DIR
MICHAEL ALMOND	DIR	TRE	TRE	TRE	TRE	TRE
SUSAN BRADY	DIR	DIR	DIR	DIR	DIR	DIR
LINDA C MAYES	–	DIR	DIR	DIR	DIR	DIR
OVID JL TZENG	–	DIR	DIR	DIR	DIR	DIR
ERNIE TEITELL	–	–	–	SEC	SEC	SEC
GERRY ALTMANN	–	–	–	DIR	DIR	DIR
FRANK KEIL	–	–	–	DIR	DIR	DIR
JULIE WASHINGTON	–	–	–	DIR	DIR	DIR
MARVIN CHUN	–	–	–	–	DIR	DIR
PHILIP RUBIN	–	–	–	–	DIR	DIR
JULI WADE	–	–	–	–	–	DIR

Table 2: Corporate officers (not officers of the board) over time. CEO = Chief Executive Officer; PRES = President; SEC = Corporate Secretary; VP = Vice President; VPF = VP of Finance & Administration; VPR = VP of Research; VPSO = VP of Scientific Operations. Note that Arthur Abramson passed away in December of 2017.

	2014	2015	2016	2017	2018	2019
PHILIP E RUBIN	CEO,VP	VP	VP	–	–	–
KENNETH R PUGH	PRES	PRES	PRES	PRES	PRES	PRES
DOUGLAS H WHALEN	VPR	VPR	VPR	VPR	VPR	VPR
JOSEPH P CARDONE	VPF	VPF	VPF	VPF	VPF	VPF
VINCENT L GRACCO	VPSO	VPSO	VPSO	VPSO	VPSO	VPSO
ARTHUR S ABRAMSON	–	SEC	SEC	SEC	SEC	SEC

compensation) of at least $100,000 in a given year. Note that several officers of the corporation also received salaries of more than $100,000. Table 3 lists non-officer HCEs.

TODO: Plot total salary reported for officers and HCEs by year as percentages of total salary paid by HL each year (cf. Figure 5).

Table 3: Highly compensated employees over time. A '1' indicates the person was listed as an HCE on that year's i990. Zeros (0) should not be interpreted as an indicator of employment status.

	2014	2015	2016	2017	2018	2019
SUSAN GALLI	1	1	1	0	0	0
EINAR MENCL	1	1	1	1	1	0
DAVID BRAZE	1	1	1	1	1	0
JULIE VAN DYKE	1	1	1	1	1	0
BETTY J DELISE	1	1	1	1	1	1
MARK TIEDE	0	0	0	0	1	1

Summary of NIH Grant Activity

This data set includes a total of 61 National Institutes of Health grants awarded to Haskins Laboratories due to the efforts of 38 PIs in the period spanning fiscal years 1990 to 2020. Note that NIH actually provides weekly data releases during the active fiscal year (i.e.,), but I have not included those "up to the minute" data here, mostly because it would be a little tedious to go that extra step and aggregate the weekly reports into a single partial data set for the current year (National Institutes of Health n.d.).

Lifespans and Types of Grants

Figure 8 shows the lifespans of grants to Haskins within the analyzed period, while also highlighting the type of grant in each case. For each grant, activity codes run from the first active fiscal year to the last, with a point (letter) at each intermediate active year. Grants are sorted by last year of activity so that grants with more recent funding are closer to the bottom.

There are some oddities here. Notice that for some grants(e.g., DC000183) there seem to be intermediate years without activity. I speculate that these are periods of unfunded extension. Also, for most grants the 'activity' code does not change, but for a few it does (e.g., DC000183, DC000121). I don't know what to make of that.

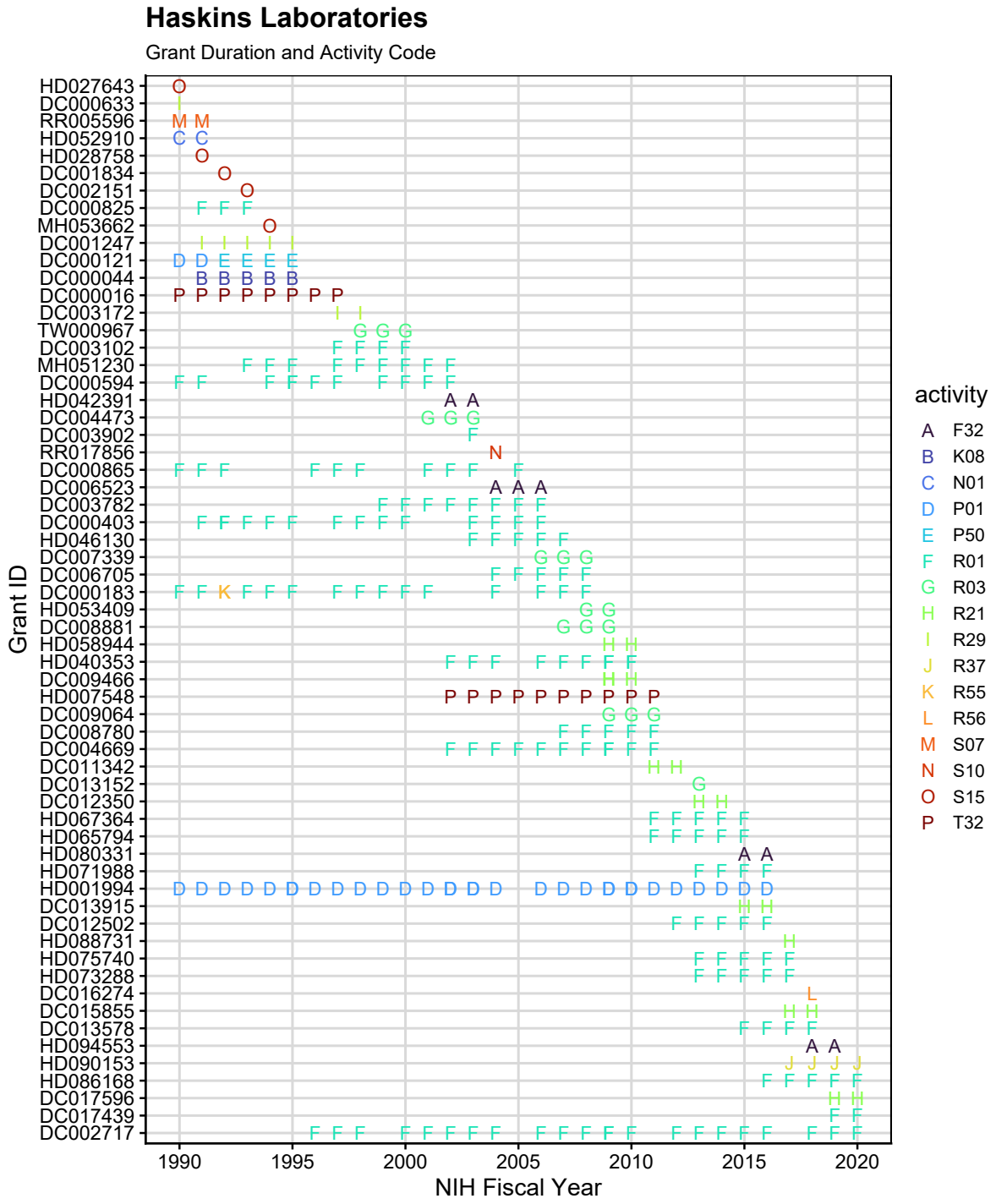

Figure 8: Lifespans of Grants. Grants are sorted by final year of activity within depicted span. I surmise that lacunae within a grant's span (missing activity codes) are likely due to 'no cost extensions.'

Primary Investigators

Years of Grant Activity per PI

Figure 9 shows the years in which each primary investigator has grant income. The figure shows grant activity at Haskins Laboratories only. Any grant activity at other organizations is not included here. Two instances of Co-PIs are indicated as composite individuals. Note years in which a PI has grant income only through unfunded extensions are not included included here. Grants are sorted by last year of activity so that grants with more recent funding are closer to the bottom.

As with IRS form 990 data, PI names are sometimes not consistently rendered in the NIH grants database. That is, when the same individual is mentioned across multiple grants/years, their name is not always rendered consistently. Middle names are a common source of problems. Sometimes these are included in full, sometimes as initial only, sometimes they will be omitted altogether. Married names are another common source of discrepancies. Unlike the IRS, the NIH has included a unique numerical ID for PIs, `pi_ids`, which *should* be consistent across name variations. If the goal is to combine data using names of individuals across IRS and NIH sourced data, there is even more work to be done. For example, names from IRS 990s are typically in first_name last_name order, but the NIH database lists PI names as last_name, first_name.

TODO: Another useful view would be to sort grants by **first** year of activity. This would highlight the number of grant starts in each fiscal year.

Total Number of Grants & Grant Income per PI

TODO: insert table of grant count and grant income by PI and year

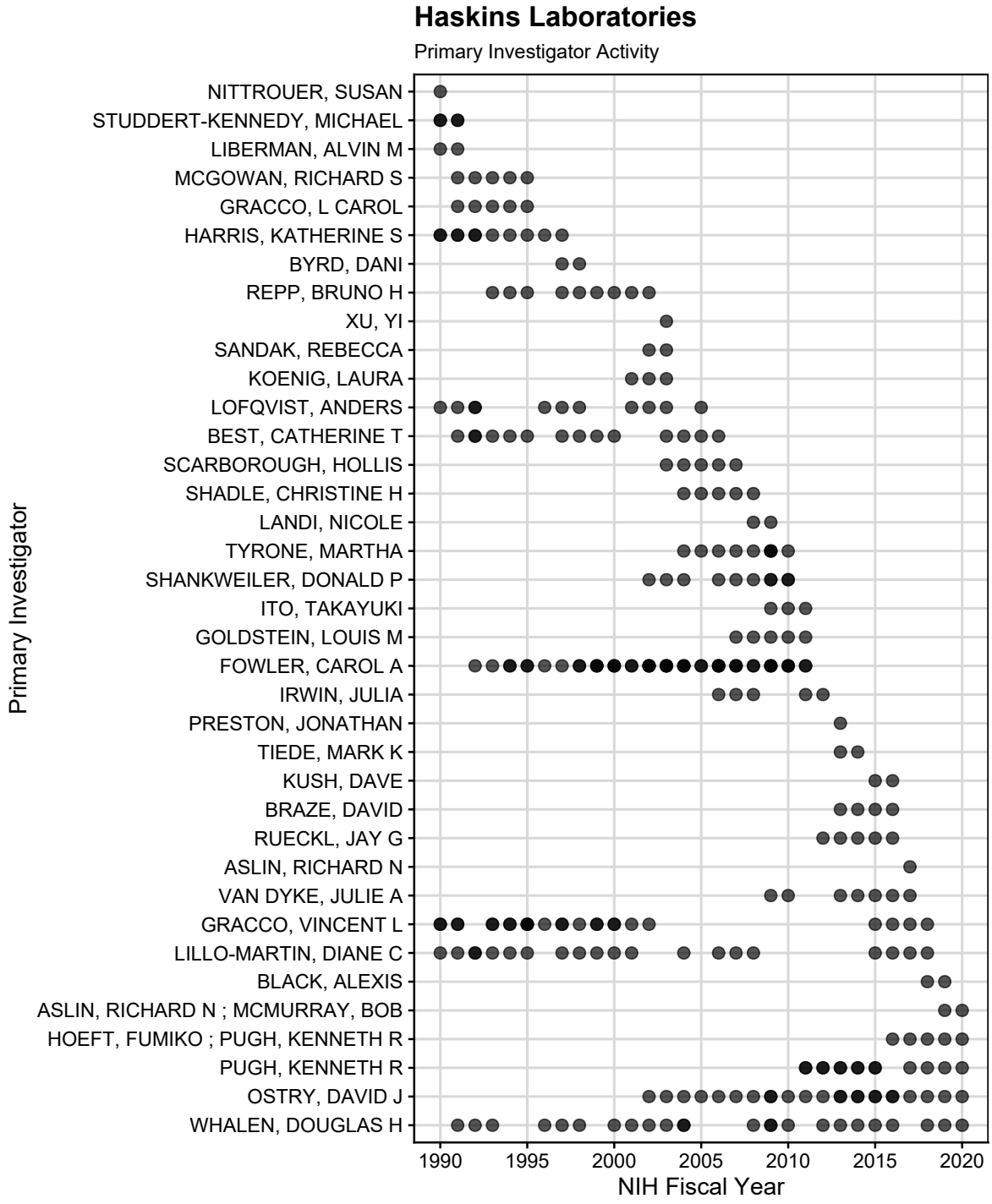

Figure 9: Years in which each primary investigator generated revenue to Haskins through one or more NIH grants are indicated with solid points. PIs are sorted by final year of activity within depicted time span. This chart indicates only NIH grant activity at Haskins; possible funding from sources is not included, nor is a PI's grant activity at other organizations. Co-PIs are shown as composite individuals. Note that a year in which a PI has grant income only through unfunded extensions will be blank (no point).

NIH Grant Income and Effective Indirect Rates

Figure 10 shows NIH Total Costs per year since 2000 to Haskins in blue. Total Costs are the sum of Direct Costs and Facilities and Administrative Costs (National Institutes of Health 2021). NIH Direct Costs, which are only provided as disaggregated from total costs since 2012, are shown in grey. The blue curve here corresponds to the blue curve in Figure 3, but extends backward in time.

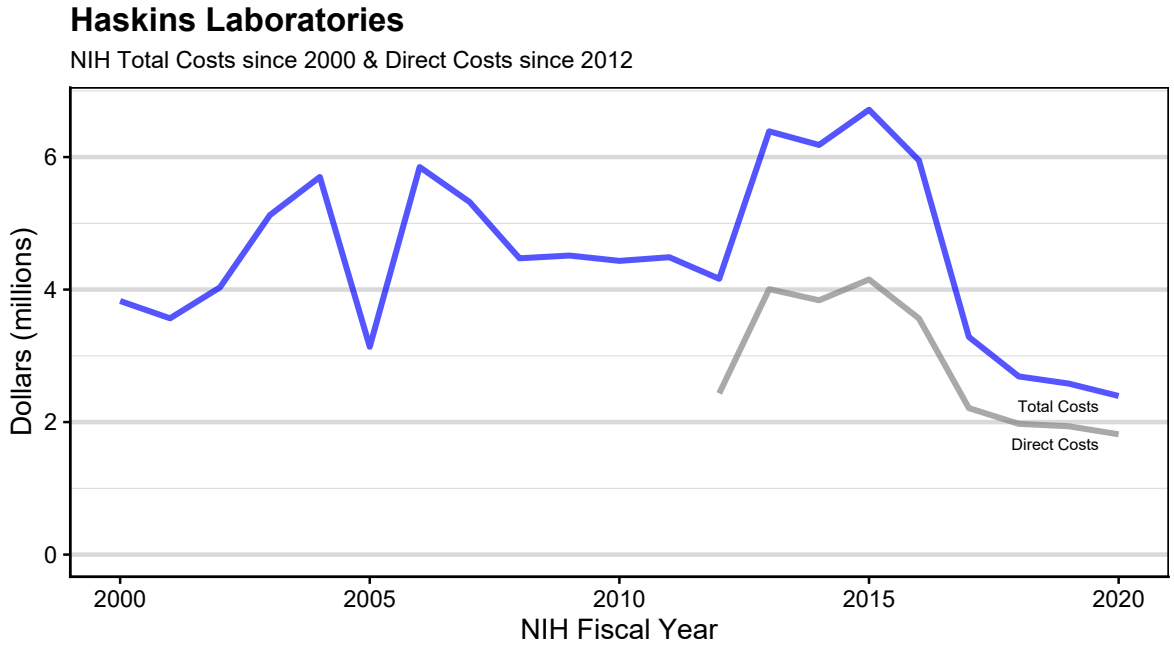

Figure 10: NIH Total Costs per year to Haskins in blue. Note that the NIH database only includes total cost numbers for grant activity from FY 2000 forward. NIH Direct Costs per year are in grey, which are only provided as disaggregated from total costs beginning in 2012.

TODO: Show direct costs by year as a stacked barchart with fill=activity_code.

TODO: Show direct costs by year as a stacked barchart with fill=pi_names.

TODO: Show effective F&A rate as a percentation of Total Costs. Although, grantee negotiated F&A rate is normally presented as a percentage of Direct Costs. Maybe show it both ways?

Appendix A: Analysis Software

All data summaries and analyses in this report were carried out using the *R* statistical environment, version 4.1.0 (R Core Team 2019). The report itself was produced using an Rmarkdown workflow. The following table lists the non-base R packages used in building the report. To see a full citation for a specific package, assuming you have both *R* and the particular package installed, call (e.g.) *citation(dplyr)* from the *R* prompt.

Table 4: R packages used in this report.

package	version	date
cowplot	1.1.1	2020-12-30
dplyr	1.0.7	2021-06-18
forcats	0.5.1	2021-01-27
fs	1.5.0	2020-07-31
ggplot2	3.3.5	2021-06-25
ggrepel	0.9.1	2021-01-15
here	1.0.1	2020-12-13
kableExtra	1.3.4	2021-02-20
knitr	1.33	2021-04-24
lubridate	1.7.10	2021-02-26
ParseIRS990	0.1.1	2021-07-28
purrr	0.3.4	2020-04-17
readr	2.0.0	2021-07-20
stringr	1.4.0	2019-02-10
tidyr	1.1.3	2021-03-03
xml2	1.3.2	2020-04-23

Appendix B: About the Author

David Braze, Ph.D., is a researcher and consultant with a background in linguistics and reading research. He has more than 25 years experience investigating the cognitive and social foundations of language, literacy, and educational achievement, including 17 years as a Senior Research Scientist at Haskins Laboratories. His research at Haskins, funded by the National Institutes of Health, emphasized the neurobiology of language and reading, and applications to education. Dr. Braze has broad experience consulting in the business, government, and non-profit sectors.

email: davebraze@gmail.com
website: davebraze.org
Google Scholar Page

References

Fowler, Carol A, and Donald Shankweiler. 2021. *Language and Life: Haskins Laboratories' First Half Century*. New Haven, CT: Haskins Press.

Fowler, Carol A., and Donald Shankweiler. n.d. "Haskins Laboratories Oral Histories and Transcriptions." Accessed September 13, 2021. https://haskinslabs.org/about-us/history/oral-histories-transcriptions.

Internal Revenue Service. 2020a. "Instructions for Form 990 Return of Organization Exempt from Income Tax." https://www.irs.gov/instructions/i990.

———. 2020b. "Instructions for Schedule F (Form 990)." https://www.irs.gov/instructions/i990sf.

———. 2020c. "Schedule I (Form 990) with Instructions." https://www.irs.gov/pub/irs-pdf/f990si.pdf.

National Institutes of Health. 2021. "NIH Grants Policy Statement: 7.3 Direct Costs and Facilities and Administrative Costs." https://grants.nih.gov/grants/policy/nihgps/HTML5/section_7/7.3_direct_costs_and_facilities_and_administrative_costs.htm.

———. n.d. "NIH ExPORTER." Accessed August 25, 2021. https://exporter.nih.gov/.

"OpenPayrolls - the Largest Nationwide Salary Database." 2021. https://openpayrolls.com/.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Roberts, Brandon, Ken Schwencke, Mike Tigas, Sisi Wei, Alec Glassford, and Andrea Suozzo. 2013. "Nonprofit Explorer." *ProPublica*. https://projects.propublica.org/nonprofits/.

Whalen, D. H. 2019. "Obituary: Arthur S. Abramson." *Journal of Phonetics* 72 (January): 83–84. https://doi.org/10.1016/j.wocn.2019.01.001.