Lost Boy

The blog of @ldodds

Do data scientists spend 80% of their time cleaning data? Turns out, no?

Leigh Dodds Data January 31, 2020September 29, 2020 4 Minutes

It's hard to read an article about data science or really anything that involves creating something useful from data these days without tripping over this factoid, or some variant of it:

Data scientists spend 80% of their time cleaning data rather than creating insights.

Or

Data scientists only spend 20% of their time creating insights, the rest wrangling data.

It's frequently used to highlight the need to address a number of issues around data quality, standards, access. Or as a way to sell portals, dashboards and other analytic tools.

The thing is, I think it's a bullshit statistic.

Not because I don't think there aren't improvements to be made about how we access and share data. Far from it. My issue is more about how that statistic is framed and because its just endlessly parroted without any real insight.

What did the surveys say?

I've tried to dig out the underlying survey or source of that factoid, to see if there's more context. While the figure is widely referenced its rarely accompanied by a link to a survey or results.

Amusingly this IBM data science product marketing page (https://www.ibm.com/analytics/data-science) cites this 2018 HBR blog post (https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists) which cites this 2017 IBM blog (https://www.ibm.com/cloud/blog/ibm-data-catalog-data-scientists-productivity) which cites this 2016 Crowdflower survey (https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf). Why don't people link to original sources?

In terms of sources of data on how data scientists actually spend their time, I've found two ongoing surveys.

- The Crowdflower (now Figure Eight (https://www.figure-eight.com/)) data science survey. This has run annually since 2015. The 2018 is here (https://visit.figure-eight.com/rs/416-ZBE-142/images/Data-Scientist-Report.pdf), but you'll need to fill in a form to see the 2019 version in your inbox. (https://visit.figure-eight.com/WC-2019-StateofAIReport_RegLP2018GTM.html?source=Web&medium=FEResourceCenter)
- The Kaggle Data Science Survey (https://www.kaggle.com/c/kaggle-survey-2019), which has run since 2017. You can see some charts for the 2018 survey in this notebook (https://www.kaggle.com/paultimothymooney/2018-kaggle-machine-learning-data-science-survey)

So what do these surveys actually say?

- Crowdflower, 2015 (https://visit.figure-eight.com/rs/416-ZBE-142/images/Crowdflower_Data_Scientist_Survey2015.pdf): "*66.7% said cleaning and organizing data is one of their most time-consuming tasks*".
  - They didn't report estimates of time spent
- Crowdflower, 2016 (https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf): "*What data scientists spend the most time doing? Cleaning and organizing data: 60%, Collecting data sets; 19% …*".
  - Only 80% of time spent if you also lump in collecting data as well
- Crowdflower, 2017 (https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf): "*What activity takes up most of your time? 51% Collecting, labeling, cleaning and organizing data*"
  - Less than 80% and also now includes tasks like labelling of data
- Figure Eight, 2018 (https://visit.figure-eight.com/rs/416-ZBE-142/images/Data-Scientist-Report.pdf): Doesn't cover this question.
- Figure Eight, 2019: "*Nearly three quarters of technical respondents 73.5% spend 25% or more of their time managing, cleaning, and/or labeling data*"

    - That's pretty far from 80%!
- Kaggle, 2017 (https://www.kaggle.com/surveys/2017): Doesn't cover this question
- Kaggle, 2018 (https://www.kaggle.com/paultimothymooney/2018-kaggle-machine-learning-data-science-survey): "*During a typical data science project, what percent of your time is spent engaged in the following tasks? ~11% Gathering data, 15% Cleaning data…*"
    - Again, much less than 80%

Only the Crowdflower survey reports anything close to 80%, but you need to lump in actually collecting data as well.

Are there other sources? I've not spent too much time on it. But this 2015 bizreport article (http://www.bizreport.com/2015/07/report-data-scientists-spend-bulk-of-time-cleaning-up.html) mentions another survey which suggests "*between 50% and 90% of business intelligence (BI) workers' time is spend prepping data to be analyzed*".

And an August 2014 New York Times article (https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html) states that: "*Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data*". But doesn't link to the surveys, because newspapers hate links.

It's worth noting that "Data Scientist" as a job started to really become a thing around 2009 (https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/). Although the concept of data science is older (https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/). So there may not be much more to dig up. If you've seen some earlier surveys, then let me know.

Is it a useful statistic?

So looking at the figures, it looks to me that this is a bullshit statistic. Data scientists do a whole range of different types of task. If you arbitrary label some of these as analysis and others not, then you can make them add up to 80%.

But that's not the only reason why I think its a bullshit statistic.

Firstly there's the implication that cleaning and working with data is somehow not worth the time of a data scientist. It's "data janitor work (https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html)" work. And "It's a waste of their skills to be polishing the materials they rely on (http://don't particularly love munging and cleaning data either. It's a waste of their skills to be polishing the materials they rely on.)". Ugh.

Who, might I ask, is supposed to do this janitorial work?

I would argue that spending time working with data. To transform, explore and understand it better is absolutely what data scientists should be doing. This is the medium they are working in.

Understand the material better and you'll get better insights.

Secondly, I think data science use cases and workflows are a poor measure for how well data is published. Data science is frequently about doing bespoke analysis which means creating and labelling unique datasets. No matter how cleanly formatted or standardised a dataset its likely to need some work.

A sculptor has different needs than a bricklayer. They both use similar materials. And they both create things of lasting value and worth.

We could measure utility better using other assessments than time spent on bespoke work.

Thirdly, it's measuring the wrong thing. Actually, maybe some friction around the use of data is a good thing. Especially if it encourages you to spend more time understanding a dataset. Even more so if it in any way puts a break on dumb uses of machine-learning.

If we want the process of accessing, using and sharing data to be as frictionless as possible in a technical sense, then let's make sure that is offset by adding friction elsewhere. E.g. to add checkpoints for reviews of ethical impacts. No matter how highly paid a data scientist is, the impacts of poor use of data and AI can be much, much larger.

Don't tell me that data scientists are spending time too much time working with data and not enough time getting insights into production. Tell me that data scientists are increasingly spending 50% of their time considering the ethical and social impacts of their work.

Let's measure *that*.

Published by Leigh Dodds

Director of Delivery @ODIHQ. Chair of @BathHacked. Author of @datapatterns. Proud Dad. View all posts by Leigh Dodds