My Proposal

Update:  
If we want to stick strictly to things that are easier to measure,  I think there’s two distinct but related threads here:  Which jobs are dedicated to data activities versus which jobs have a stated data component.    The first question involves understanding how many kinds of jobs have “data” in the title and in which industries do they appear?  We can compare the current government listings for occupations from the BLS with some data from indeed or linkedin (if that is easy to pull.  I know LinkedIn data is challenging!).  Alternatively, if we might want to look at skill mismatch to see what skills or technologies are listed for certain job listings compared to skills or technologies that appear on resumes for people with that job title.  Skill mismatch and job polarization is a hot topic in labor econ right now.

If we want the time component, we can focus on just the government job descriptions and see how many jobs mention “data” work regardless of the title.  My hypothesis is that if we define a core set of words for technical work (data, database, analysis, etc.) and look for them in the job descriptions, we would find that in early years they only appear in IT related jobs but in later years they spread to other areas.

Before we choose one, why don’t you take a look at the historical OOH documents that would need to be parsed (They can be found in the FRASER archive from the St. Louis Fed.  Here’s an example from 1990: <https://fraser.stlouisfed.org/files/docs/publications/bls/bls_2350_1990.pdf>)  
As well as any API or scraping suggestions for Indeed and LinkedIn.  We don’t want to do ANYTHING that violates terms of service so if we need permission, we’ll ask.  
  
As much as I want to think of this like an economics research project where we formulate a question and then look for data to answer it, we need to approach it like a data science project:  data first.  Let’s see which data sources are easiest to obtain and work with and then formulate a question based on that given the possibilities we’ve already outlined.  
  
Does that make sense?

Dear Professor Sandra,

There were three things that you mentioned in your initial brainstorm for ideas.

The first boils down to using text mining to classify jobs and see if particular jobs use particular word tokens. Here, you mentioned that creating a measure of data literacy would be amazing. The second relates to how the need for data skills have changed over time based on features such as salary negotiations and data in the BLS occupational handbook. The third relates to how education for data things have evolved and the shift towards a data-oriented sphere in higher education.

They’re all about data. But I think there’s a common theme that encompasses all three, and it tries to answer a very simple question*: how are data scientists actually used in companies?* Are data scientists a valuable asset? Data science is 100% valuable, we can focus on the ways it is helpful and explore the landscape of both data-centered companies (companies with a data product) and companies leveraging data scientists for their operations, be it in healthcare or some other application area.

Let me illustrate the discrepancies that I want to “expose” with this research:

1. A job posting may mention all these methodologies that the applicant needs to know like Hadoop, Spark, and comprehensive experience in machine learning – but on the job, the actual toolset used is significantly more narrow and limited. I actually have personal experience in this. I had a data science internship early 2020. Their job description involved ML knowledge and use of a NLP library (spaCy) but in reality, most of the work they need was just data cleaning. It’s common knowledge that 80% of data science is data cleaning, but I still believe this trend exists in areas that isn’t data cleaning as well.
2. In practice, many companies do not need “data scientists”. Methods such as deep learning models or decision trees are often overkill for tasks that need just data visualization and more business intelligence related tasks. This is especially true for startups that has a non-data related product.
3. When you’re the lone wolf in a company with no real infrastructure set up, and the company’s programming and statistics knowledge is really lacking, it can be immensely difficult to set up these infrastructures because only you understand the technical content.
4. Most companies don’t have their data infrastructure together. Healthcare companies often operate through just Excel even when you may think that a hospital may already have proper database system in place.
5. Management might be looking for data insights that are aligned with their agenda instead of real insights that come from what the data actually says.

**Delivery**

To do a technical analysis on this subject, I think the best source of information are interviews with actual data science professionals. I can personally conduct them, and then record the transcripts of our conversation as a text document. Then, to measure the discrepancy between the job posting and the actual job, I can do a text analysis on the terms used and make a metric to measure a ratio of skills people actually need and skills people actually use in the job. Or I can find the average percentage of skills listed in the job description that are actually used in the job and gather a reasonable sample size on this finding.

From this intuition, we may even make a model that classifies the integrity of a job description. “Fraudulent” job postings in this case would be job postings or skills that are less likely to be actually implemented in a job.

We can come up with a list of all the use cases that are mentioned by interviewees and come up with a set of the text “terms” that are described by interviewees and compare them to the “terms” that are mentioned in the job description. The intersection and differences between these sets could be meaningful. For example, an employee might be doing a lot of data engineering related work or setting up cloud infrastructures in a job that emphasizes machine learning in their job description. We can find the likelihood of particular terms of being left out of the job description.

Of course, these are all ideas – the actual research will be informed by your experiences and input. But I think this is a good starting point and gives many branching points for where we could take this research! Thanks in advance, let me know your thoughts and feedback!