**Data Science interview tips**

* Why PhDs?
* Tell me about your research
* How do you problem solve?
* Taking feedback
* Communicating the big picture
* Where do you fit in?

**Highlight examples of things that they’re looking for**

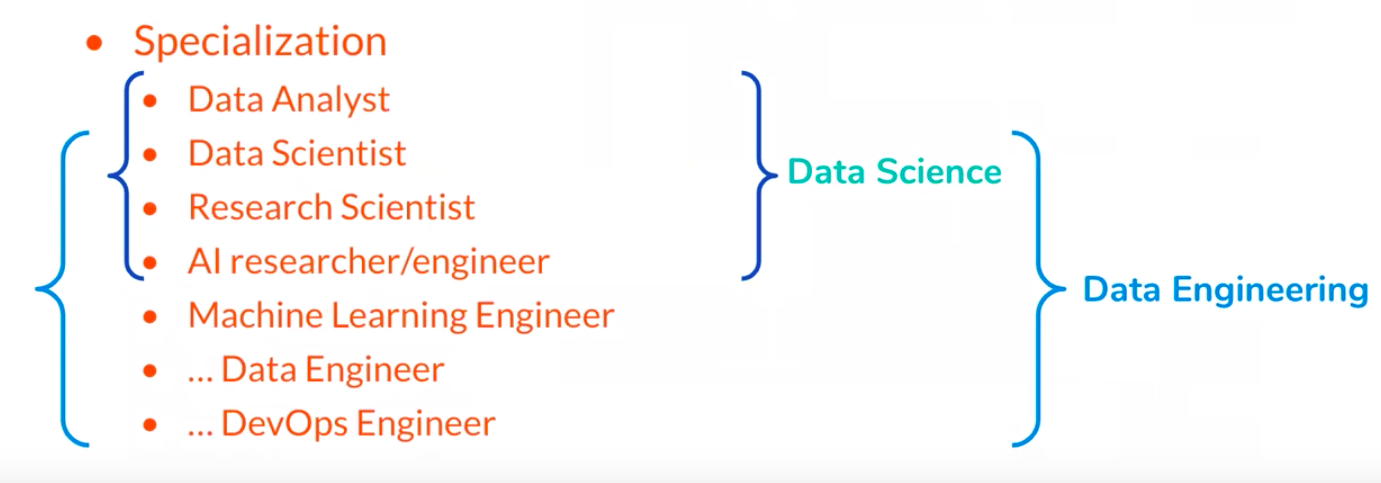
* **Make a table of the things they’re looking for and make columns for Fluidigm and columns for my project of when I’ve done that**
* **Sprinkle jokes in ☺**
* **Do the same with Alector**

**Data science skills (scale of 0 to 3, 0 being none at all)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Skill/quality** | **Experience (3 being most comfortable)** | **Desire to do more (3 being highest)** | **Comment** |
| Math and statistics |  |  |  |
| machine learning | 1 | 3 |  |
| statistical modeling | 1 | 3 |  |
| experimental design | 2 | 3 |  |
| bayesian inference | 1 | 3 | Did well with grade but had some difficulty grasping concepts |
| supervised learning | 1 | 3 |  |
| unsupervised learning | 2 | 3 |  |
| optimization | 0 | 2? |  |
| Programming and database |  |  |  |
| CS fundamentals | 1 | 3 |  |
| scripting (e.g. Python) | 2 | 3 |  |
| statistical computing packages (R) | 2 | 3 |  |
| SQL | 1 | 3 | Learn as needed |
| parallel databases | 0 | 0 |  |
| MapReduce | 0 | 0 |  |
| experience with AWS | 1 | 2 |  |
| Soft skills |  |  |  |
| passionate about business | 3 | 3 |  |
| curious about data | 3 | 3 | Motivation for me to move to bioinformatics |
| influence without authority | 3 | 3 |  |
| hacker mindset | 3 | 3 | Overcame limits of API |
| problem solver | 3 | 3 |  |
| strategic, proactive, collaborative | 3 | 3 |  |
| Communication |  |  |  |
| ability to engage with senior management | 2 | 3 | C1 HT |
| storytelling | 2 | 3 |  |
| translate data to decisions | 2 | 3 | C1 HT |
| visual art design | 2 | 3 | Name check Cairo |
| knowledge of visualization tools | 2 | 3 | CFTR |

# Job market

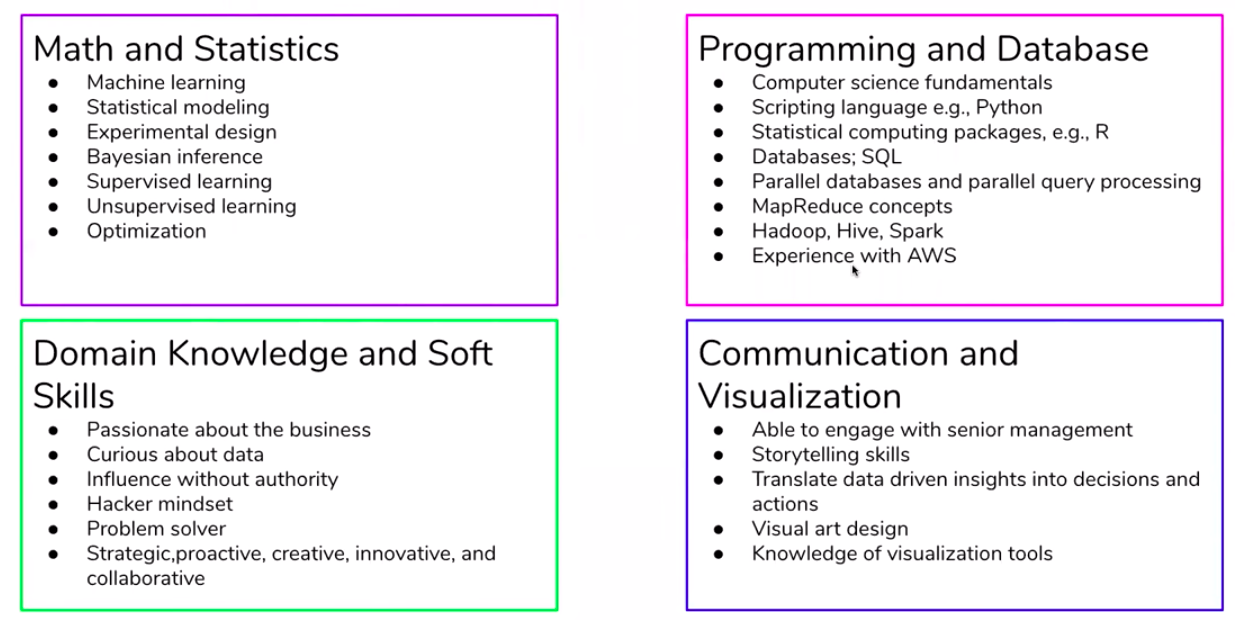
* + DS jobs are growing and the specialization of jobs are growing
  + Data engineering **uses some APIs to get big data**



* Who are academic scientists?
  + Plan/execute study takes months/years
  + Collect and clean data
  + Use programming and statistics/ML to discriminate signal and noise
  + Convey results to the scientific community
  + There’s a gap between what we get training for academia and what we want to show as a data scientist
* Who are data scientists?
  + **Plan and execute a study in week-long sprints (talk about how I did it over a period of months – show github)**
  + **Collect and clean data (API)**
  + **Use programming and statistics/ML to discriminate signal and noise**
  + Convey results to **team/company/investors (talking to internal team)**
  + **Make data-informed decisions that directly impact a product and business (e.g. data on whether they should launch a new product, C1 HT)**
  + **Test**

# Why PhDs?

* + Transferrable skills – don’t need all of this but emphasize more the communication/visualization and soft skills
    - E.g. adjusting storytelling skills depending on whether I’m talking to scientists or people
    - ID the skills I’ve done here and talk about:
      * what you’ve done but don’t like
      * what you’ve done and like
      * **what you haven’t done but want to learn**



* Highlight things where the job fits you
* Highlight things that are important across all job skills. **E.g. are you a problem solver? What’s an example? (maybe have a couple of stories)**
  + How did you assess it? Did you create a feasible timeline? Was it completed and what were the results that provided impact?
  + Outdated RNA-seq pipeline
    - Old reference, old tools – publications had shown that these tools were error-prone. It was a risk not only in data assessment but also in the perception of our company.
    - It was nice as a set pipeline but it was difficult to modify for ad hoc analysis.
    - I estimated that I would have a basic pipeline within a few weeks with continuing iterations.
    - New pipeline allowed us to see differences we expected between some parameters that we expected to see, but was not present before.
    - The pipeline also included metrics that couldn’t be studied but were of previous interest.
    - Re-wrote it, provided new metrics, provided visualizations that are automatically generated so that people can assess their experiment and be able to make the next decision whether it’s an experimental optimization or whether it’s closer to show to marketing or beta-test customers, etc.
  + SC Total RNA-seq
  + Education study
    - talk about how new legislation affecting start dates might be a really cool piece of data to add to the model
* **Identify what you want to learn. Know the breadth of some stuff, but then highlight what you want to specialize in. Be able to convey what you don’t know and then have something that you want to be fluent in.**
* For data science: focus on math and statistics box, especially ML
  + Improve my comfort in statistics and ML
  + Be better at some novel visualizations (d3?)
* For data engineering: focus on programming and database, have to know SQL
  + Large companies use tons of SQL

# Tell me about your research or tell me about your background

* + Behavioral part of interview
  + Will apply to insight or any job interview or talking to recruiter
* Describing your background effectively
  + Want it to be a very good 30 seconds; concise and clear
  + Want it to be specific to the job (e.g. provide data science relevant things to a data science job discussion)
    - If you talk about something, ground it into something that the average person can relate to.
    - Convey impact, why it’s interesting, and skills it highlights
  + Listen to directions and queues
  + Keep it conversational and interesting!
    - Remember a person is interviewing lots and lots of people – stand out!
    - **Use Toastmaster techniques in terms of vocal variety and body language**
  + Anything in my background is fair game – know it!
    - Turn this into an advantage
    - Focus on the projects that worked out the best or you know most about (e.g. talk about how TRseq has lead to sales); pick a “flashier” project
  + STAR(T)
    - Situation
    - Task (what task did you do; what did I specifically do?)
    - Action (what action did you take – similar to above)
    - Results
    - Takeaway (what’s the impact; saying it was in a good journal isn’t really that interesting – they what to know the DS impact)
  + Imagine stepping into a new place where you have never been before. A select group of neurons in your hippocampus are reacting to that experience. What is special about that subset of neurons that made it more likely to activate than other subsets? How did those cells get chosen? And what is the representation of that experience on a genetic level? That is what our study was about. We pioneered some techniques that allowed us to get at this question, being first to publish about a full transcriptome expression of activated single neurons. We found genes related to synaptic proteins and growth factors being expressed in these cells, suggesting that they were on their way to changing their connections. That project was my connection to my current role. In that project, I found myself asking some questions related to the data but my skillset was limited in addressing it. This started to coincide with the increasing popularity of data science. When I started in my current role, I made it a priority to get better at data science and bioinformatics.
  + In my industry role, C1 HT…
* You’ll be asked: Tell me about a time when you had a conflict with a team member? Or conflict with adviser?
  + The best explainer is data.

# How do you problem solve? (talking about my side project)

* + On paper a lot of people actually look similar, they use the same keywords, etc.
  + They want to know you solve problems
* Showing code effectively
  + Start at the top (high-level) – **use table of contents to your advantage**
    - Dig into most interesting aspects
  + Keep it conversational
  + Mind the time
  + Never go through line-by-line (very important!)
  + Basic script:
    - “Here’s a high-level problem I’m trying to solve”
    - “Here’s how I downloaded, cleaned data – considerations I had to make”
    - “Here’s the algorithm I looked into and why I chose it”
    - “Here are the most relevant features”
    - “Here are the final visualization that shows the main result”
  + Keep it conversational and keep it moving
    - The interviewer knows what you’re doing, e.g. splitting data, etc.
    - Touch on points that should be taken for granted (e.g. cleaning data, splitting, turning into categorical variables… don’t get stuck in the weeds)
* Morgan:
  + Don’t worry too much about having a clean story
  + Have clear motivation
  + Be able to explain rationale for analytical choices I made
  + Acknowledge the weaknesses and have ideas for what I’d do next if I had more time
  + Is data the limitation? What additional data would be helpful to get?
  + Is there a different technical approach I’d want to explore?
* If you’re not showing code:
* How you would address a new scenario: Amazon sells millions of products per second. They might want to launch a new coupon. You’re the DS, who is supposed to advise them on what to do. You have 1 week… 1 month.
  + Data? (What kind do they have?)
  + Approach? (Can you get features of how the coupon was presented to customers, labeled with those that have bought, to build a model that can predict which coupon presentation is effective? Or specific industry? Or what metric would be used)
  + Evaluation? (How do you know it’s right? What graphs and visualizations would you create? Maybe do some clustering analysis.)
  + Validation?
  + Note: Facebook runs experiments internally (on their own employees) but Amazon may not want to launch too many experiments. What do you want to do if they **don’t** want to experiment? (They may want to put you in an intentionally challenging situation and see how you’d approach it.) You can dig into data more. (Maybe use historical data that was a similar situation. If you had a week, or month, or 3 months – what would you do?)
  + They want to see that you can outline a plan and provide a framework for it. They want to see this:
    - You can structure a problem
    - Creatively approach it in a way that’s feasible and uses existing data if possible; if not, what new experiments would you do?
* Taking feedback
  + How do I take feedback and learn from it
  + Demonstrate that you can give things a thought but also show that you can learn, work on a team, and let other people contribute/ask questions
  + Demonstrate that you can **iterate constantly (note that the theme of iterating is an Insight culture thing)**
  + Show that you can work with the person
  + Ask them constructively for feedback
  + Check in with the person who is asking for feedback
  + Shows that you’re a communicator and team player
  + Get out of the solo/academic ownership mindset
* Communicating the big picture
  + Product, not science
  + Product focuses on the end user, data-driven answers (an end-user could be an executive or decision-maker)
  + Science focuses on the novel method, the interesting exploratory problem
  + Get an answer, figure it out, move through, then try to get a better answer.
* Where do you fit in?
  + What type of data scientist do you want to be?
  + Have you evaluated the field?
  + **Look at blogs, job postings, companies, industries…**
  + Consider work-life balance, mobility in the company, etc.
  + Whatever experience you’ll get, you’ll be good at it… so think carefully about what you want to get experience in – analyst, visualization, engineering, etc. – it puts you on a certain trajectory
  + Where do you fit – what do you consider important?
* Other:
* Why do Insight now?

# Demo notes

* Goal is to do this in 3-5 min. (aim for 3 minutes - 400 words; will be interrupted with questions)
* Sprinkle keywords in (bold/highlight)
* Make the intro/thesis and conclusion tight.

## Intro

* **I investigated features of high schools that promote low-income student success. The hope is to identify features that can be affected by policy changes to help low income students graduate and become college eligible at a higher rate.**
* I am interested in this question because during college and graduate school, I mentored low-income students and so it made me think about how different schools are successful in preparing their students for long-term success. I focused on CA public high schools.

## Body/data

### To obtain features of schools that could be meaningful for my target, I pulled in data from various sources.

* To obtain features of each school’s local community, I obtained **census data** for each zip code in the state of CA.
* To get detailed enrollment information of the schools, I pulled in information from the **CA public schools database**.
* The CA public schools database lacked some details and therefore I turned to a third data source, the **GreatSchools website**. If you’re not familiar with this site, it has information on academic performance, equity, and most importantly, it had data on low-income college eligibility.
* Here is where I use an **API** that I obtained from GreatSchools. The API allowed me to get the URL for each high school through an **XML** page. However, other than the overall rating, I found that the API was **limited** in providing the other details on the page. This includes these other ratings, and other features I thought could be **important and actionable for my target**, such as teacher salary and average student to teacher ratio. Therefore, I wrote **custom functions to scrape these kinds of details from the html.**
* One consideration I had to make was when different schools had the same. For example, there are three high schools in the state with the name “Golden Valley” and so I scraped school information on a district-by-district basis to make sure the right school was merged with the right census information.
* After cleaning and merging, I obtained 1,152 schools with 221 features.

## Modeling and feature identification

* After looking at the variability in the data for my target, which is what percentage of the high school’s low income students are college eligible, I saw that the data was largely normally distributed with a set of schools that are performing really well.
* I turned this into a logistic regression classification problem, splitting the data at one standard deviation over the mean. Data to the right of the dashed line, I’m calling the “good-exceptional” group of schools while below are the “poor-average” schools.

## Conclusion

* -if i had time... later start time... disproportionate
* this can help shape policy.... saying what's in the top features and what's not the top features.
* This is important because it helps point out what matters in policy versus not.
* -can't just pay teachers more
* -a school just has to be performing... bus them in
* -there's policy changes now... like start time
* -interesting to track ... if they have shift work
* some of the features are more actionable than others.
* this is a policy change that could potentially hurt... it was debated as to whether it would hurt low income students

## Potential questions

* Did you try other models?
* Yes but it seemed like the answers weren’t as clear cut with regression….
  + ?I turned it into a classification problem to make things simpler…

## Other demo notes

* Here is my table of contents which serves as an overview of what I did in this project.
  + Here is where I bring in the census data. I cleaned the data columns so that it was more human-readable and so that missing data could be interpreted appropriately in pandas. The categories in the census data include what industries do people work in, their primary mode of commute to work, whether they’re covered by health insurance, etc.
  + Here is where I bring in the school name which has zip codes and some demographic information.
  + Another limitation is that there are features that are not completely actionable… for example…
    - these are:
    - these aren’t
  + Here is where I create my model. I looked at my target and saw that it was largely normally distributed with some outliers on the upper end. Since I am most interested in schools that had the most successful low income students, I turned this into a classification problem by splitting the group at one standard deviation above the mean. The upper end I call the “exceptional” group which I distinguish from the “average-low” group.
  + I split, standardize, and do my model. I applied logistic regression with lasso regularization. I decided to use lasso since it can help do my feature selection for me, since there are over 200.
  + Here you can see my error metrics. These metrics…
  + What I learned is…
    - I have features but they’re not actionable.
    - What’s a little surprising are features that didn’t seem to matter.
    - Teacher salary, class ratios
  + DEMONSTRATE THIS: From a skills standpoint, I saw that I had to constantly iterate, getting to one result (or lack of a result) forced me to come up with a creative way of getting around it