Astronomy & Physics to Data Science

Taka Tanaka

with thanks to Viviana Acquaviva, Mehmet Alpaslan, Duane Lee, Daisy Leung, Jeffrey Silverman

www.linkedin.com/in/takatanaka
Twitter: @astrobassball

Table of Contents

Why I made these slides

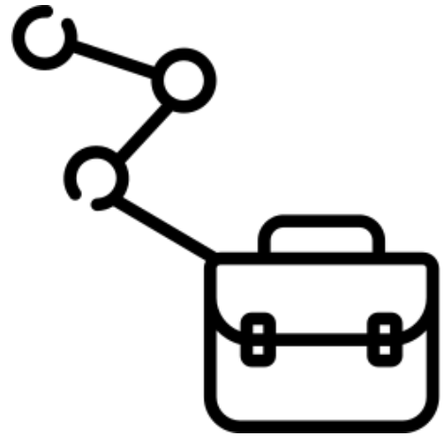
- I had a lot of great advice when I transitioned to industry data science. I'm trying to pay it forward.

- I've had conversations with over 100 academics curious about the transition. A lot of the same questions come up.

- My hope is that these slides could serve as a reference and starting point for future conversations.

- All opinions are from my own perspective—biases, privilege and all. Your mileage may vary.

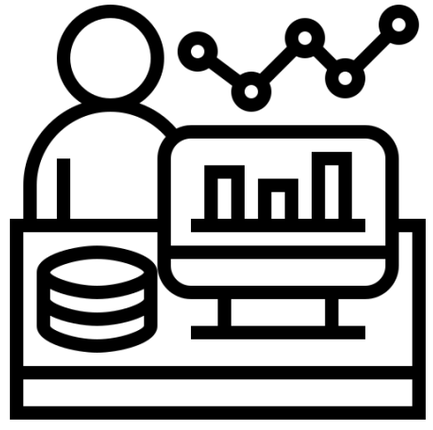
Context: Where am I coming from?



- PhD Astronomy, Columbia U. (2006 - 2011)
- Postdoctoral fellow in Munich, Germany (2011 - 2014)
- Postdoc & Research Faculty, joint appointment at two NY universities (2014-2016)
- Insight Data Science, NYC (2017)
  A bootcamp for PhDs to transition to industry data careers.
- Data scientist for a consulting firm (2017 - 2019
  - Promoted to manager in 2018
  - Worked with executives at large companies and startups in media and tech, planning and implementing new data science initiatives.
- Sr. Manager for a publicly traded health and wellness company (2019 - 2020)
  - Hired to start a new data team tasked with using advanced data methods to guide company strategy.
- Director of Data Science for a serialized fiction startup (2020 - 2021)
- Managing Director, Head of Data at a financial company specializing in residential real estate (2021 - present)
- Overall: >10,000 resumés reviewed for data positions; over 100 candidates interviewed; dozens of hiring committees.

About the job

What does a data scientist do?



- The “data scientist” title has come to encompass such a wide variety of possible roles and responsibilities, that it’s become a bit of a running joke.

- Depending on the job, the role may require a lot of software engineering, lots of dashboard building, or lots of consulting and storytelling. You may spend a lot of your time cleaning your data, pushing code, building models, or preparing presentations. You may be building something totally new, or working with legacy code.

- That being said, the *vast* majority of “data scientist” roles will require machine learning (ML), natural language processing (NLP), or both. Roles that require AI (deep learning / neural nets) are relatively rare and usually distinguished by the job title (e.g. “AI engineer”). You’ll be expected to know Python or R.

- Roles that don’t require ML/NLP may be called “data analyst” or similar—but not always! Non-ML roles might require SQL, dashboard skills, and some/no Python/R.

- Roles specializing in ML may also be called “machine learning engineer.” Roles specializing in dashboards and reports may be called “analytics engineer.”

Will I be challenged?

- In my experience, many of the core functions of the job (while the exact balance will vary, as it can in academia) are the same: abstract thinking, project planning, building and debugging code, exploring and researching solutions, visualizing data, sharing and storytelling, communicating with technical and nontechnical audiences….

- The main difference is the context of those tasks. Instead of, say, studying exoplanet atmospheres, you may be improving image processing in medicine, evaluating the user experience for streaming services, finding predictors for customer/user behavior, running experiments in apps etc.

- See the next slide, and the slide toward the end on “questions to ask about specific opportunities,” but I know very few data scientists (and those with comparable titles) that are bored with the profession. These positions tend to be greatly valued by companies and industries, and teams are usually comprised of clever, methodical people who enjoy using their minds.

What skills from PhD/ academia do you use?

- The ability to tackle ambiguous, complex, and/or open-ended problems.
    - The ability to hear about a problem in a meeting for the first time, and immediately start thinking about how you might approach it, anticipating potential time sinks and failure modes.
        - The ability to start hacking together some code and proofs of concept.
        - Making effective visuals and plots that summarizes a data story.
- Presentation, teaching, collaboration, project management—especially for roles with a heavy dose of liaising with non-technical stakeholders (clients, non-data teams, execs).
- Being able to field challenging questions from authority figures.

What are some example projects?

- Dashboards & automated reports
- Data pipelines (ETL processes) and other infrastructure
- Segmentation/clustering—e.g. identify common member profiles
- Predictions (prob. for user to pay/churn; revenue; inventory demand; click on ad)
- Optimization (inventory management; route planning; scheduling)
- Recommenders. “If you liked X, you might like…”
- Experimentation—statistical tests for in-app changes (A/B tests), ad delivery, recommendations.
- NLP—parse meaning, automatic text generation, identify topics, flag problematic content.
- Computer Vision—use image data to extract text or meaning
- Integrating above into an app or internal data processes

What are some common tools for the job?

- Cloud computing tools:
  AWS (S3, Redshift, Data Pipelines, EC2, ECS…)
  GCP (GCS, BigQuery, Colab…)

- Containers:
  Docker, Kubernetes

- Version control (git)

- Dashboarding/BI (Looker, Tableau, PowerBI)

- Coding: development environments, notebooks…

- Others: Data discovery, QA, documentation…

What’s the pay?

- It can really depend on the location and industry!

- Data Analyst roles earn a median salary of ~$70k, according to Glassdoor (2020)

- As of spring 2017, the median base salary (before bonuses) offered to Insight Data Science fellows in NYC was 120k. (Source: Insight program manager.)
  Note that these are candidates with PhDs, coming through a well known bootcamp.
  The same number was the median income for US “data scientist” positions as of late 2021.

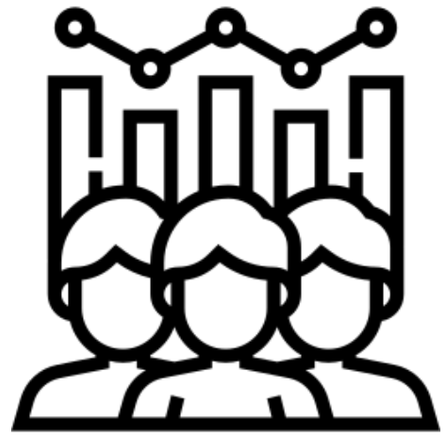
- As of fall 2019, the typical salary for senior/manager data science roles at tech companies in NYC was in the ballpark of 140-160k.

- Bonuses are typically around 10-20% of salary (sometimes much higher in banking). A part of your bonus may be tied to company or team performance. (Bonuses are typically taxed at a higher rate, around 40%.)

- Glassdoor is a decent resource to get ballpark ideas for salaries. It’s also a good place to gather anonymous feedback about specific companies from current and past employees.

- Talking to someone is also a good way to gauge whether you’re being fairly compensated.

How is the DS job market?



- My perception is that it's getting slightly harder to break in than it was 5 years ago. I think this is because of:
    - More people with a few years of experience.
    - Standardization of "the right skill set" for junior roles. Many recruiters and hiring managers like to see specific algorithms and programming tools listed in your resume.
    - The above goes hand-in-hand with the maturation of DS teams in industry. If a team works with specific tools, candidates who know (about) them have an advantage. (This is something you can prepare for, ahead of interviews.)
    - The proliferation of boot camps and BS/MS programs that specialize in data science.
    - COVID-19 note: while some industries have been affected more than others, tech and e-commerce companies continued to hire (some growing very aggressively) through the pandemic.

Titles (variable)

Tech / Most companies
Individual Contributor Track

- (Sr.) Principal
- (Sr.) Lead DS/DA/MLE…
- (Sr.) Data Scientist / Data Analytst/ Analytics Eng. / ML Eng…
- (Associate/Jr.) DS/DA/AE/ MLE…

Tech / Most companies
Management Track

- C-level/EVP
- SVP
- VP
- (Sr.) Director
- (Sr.) Manager

Wall Street / Banking

- C-level
- Managing Director
- SVP / Executive Director
- VP
- Associate
- Analyst

Getting hired

What's the hiring process like?

- Plan on the search taking months, not weeks. It's definitely not as easy as "PhD here, job please!"
- Typically, 3-4 weeks from application to last interview. Some roles may take months, and I've seen offers made in less than a week from the initial conversations.
    - Phone screen with a recruiter, typically to check for basic communication skills; understanding of the position; salary and other expectations.
    - A phone interview with a data scientist to screen for experience, basic ML knowledge, simple coding questions.
    - Data challenge or take-home test. They'll send a data set and some instructions, and you'll be expected to analyze a data set and/or build a ML/NLP model then explain your approach and findings. For some roles, you'll be asked to present to a non-technical audience.
    - On-site interview. Common to have 3-5 interviews over several hours to gauge technical skills, communication skills, cultural/company fit.
    - Offer and negotiation.
- Read up on the company, industry and role ahead of time, so that you can ask meaningful questions at each stage.

What's the hiring process like? (cont'd.)

- It's my belief that with preparation and practice, you can get to a state where (i) you're getting asked for a phone call on a decent fraction of your applications, and (ii) if you get that first phone call, you're consistently making it to the on-site interview stage.

    - If you're not getting to the phone screen, consider revising your resume. (This is the most common barrier I see for academics transitioning into industry.)

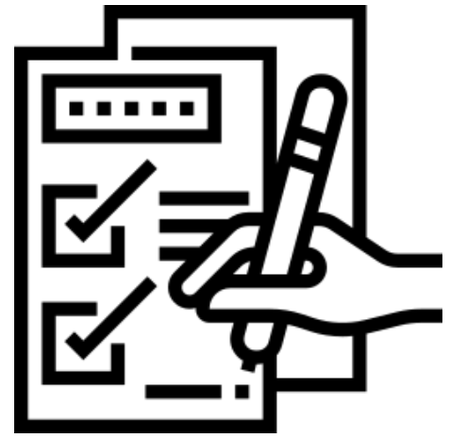
    - If you're not making it past the phone screen or the initial technical conversation, consider talking about it with people in industry to understand why not.

    - If you're not making it past the data challenge, again it's a great idea to walk through your submission with an industry data scientist.

    - For interviews, it's practice, practice, practice. The goal should be to get to the same standard as a good talk: presenting yourself professionally and speaking knowledgeably, while exhibiting enthusiasm and interest.

- But beyond that point, there's a lot of luck involved. Another candidate might be (or feel like) a better fit for them. They might decide that they want someone with more experience in their specific industry. I think of getting through the on-site interviews thinking "That went pretty well" as a "**shot on goal**," and that the way to get a job is to repeat that feeling until you get an offer.

How can I prepare?



- Polish your resume. Have people you know in industry look it over. Do NOT write it like a CV. Most people won't know the words "redshift" or "boson"—or even care. **Prominently** list the coding/ML tools you've used.
  - Example of what not to do: Moving-mesh GMHD simulations of accretion onto high-redshift AGNi, Jones et al., 2019, *MNRAS*
  - Example of how to present the above: Project lead for a computational study (C++, bash, git; cloud GPUs) that led to new phenomenological insights on supermassive black holes in the early Universe.

- Make a LinkedIn profile. Fill out the various sections, add a professional photograph, keep it up-to-date.

- Build significant depth in SQL, and either Python or R. There are more roles focusing on Python than R, but it varies.

- Build broad familiarity with common ML and NLP algorithms, (e.g. when you might use a random forest; what word2vec is) and focused experience with a few specific algorithms. (See study resources slide at end.)
  If you don't use these in your academic work, then do a self-contained side project that sits in a Jupyter notebook, a blog post, or a public GitHub repo. Pick something interesting to you (so that you have the domain knowledge and can follow through).

- Read up on relevant business fields for the roles you're interested in (e.g. digital marketing, tech blogs like Netflix, AirBnB).

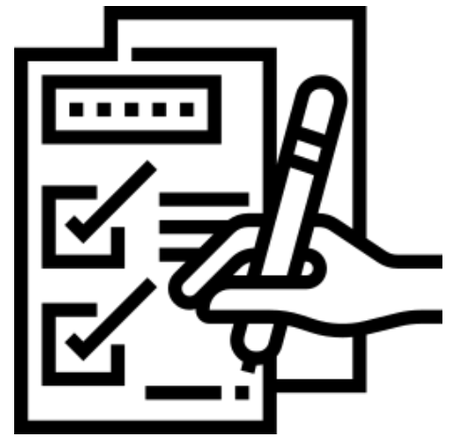
- Do mock/practice interviews.

- Make connections. **Referrals are probably the most powerful way to get an interview.** Ask people to have coffee with you, pick their brains (over video chat is okay!).

- Create a cover letter, and tailor it to every position you apply for. (Don't apply to jobs you don't want to do this for.)

- In screen + interview stages: research the company, and have a **list of questions at each interview stage**. These questions are an excellent way to demonstrate that you've thought deeply about the position and how you would fit at the company.

How can I prepare? (cont'd.)



- (*I'm assuming here that you have enough coding and stats to navigate, say, analyzing and replicating results in your academic field. I think that is sufficient for most data scientist and data analyst jobs.*
  Also see the "Some learning resources" slide at the end.)

- SQL:
  - Do you know your joins (e.g. left vs. inner)?
  - Can you aggregate (sum/average/count, group by)?
  - Can you use WHERE and HAVING clauses?
  - CASE WHEN?
  - Other: null values, window functions…

- ML:
  - scikit-learn basics: linear regression, random forest, feature importance, clustering, eval metrics…
  - "How would you deal with collinearity?"
  - "Can you describe gradient boosting, and the advantages it offers?"

- Tools & proper nouns checklist (not exhaustive, and you probably don't need to know everything):
  - AWS offerings: S3, Redshift, Sagemaker…
  - GCP offerings: GCS, BigQuery, Vertex…
  - MS Azure…
  - Snowflake
  - Containers (Docker, Kubernetes)
  - Dashboards (Tableau, PowerBI, Superset…)
  - ETL tools (Airflow, Spark…)

- Interviewing:
  - Have you done research into the company? Do you have thoughtful questions to ask about the work and the team?
  - Can you talk about your experience positively (no complaining)?
  - Can you spin how your experience is relevant to the role you're interviewing for, and how the skills you've accrued are transferrable?

The screening call

- Don’t “phone it in”! Employers typically talk to dozens of candidates at this stage and are looking for anything to narrow their list.

- Do your homework on the role and the company. Be able to speak more specifically than “Your company does X; I love X!” or “I want to work for a big/small company.”

- Articulate why you’re interested in <u>this</u> role (as opposed to any other role at another company).

- Be able to speak clearly, succinctly, and confidently about your expertise & past work. Especially for your research and thesis, have a <u>concise</u> “elevator pitch” that avoids the jargon of your academic subfield and uses terms a tech team will be familiar with. (Writing it down and rehearsing can help avoid rambling answers.)

- Don’t focus on why you think you’re qualified (they read your resumé, that’s why they want to talk to you); talk up why you’d be an awesome addition to their team. Instead of telling them you know tool X, tell them about how you impactfully used X.

- Ask questions about the role and the company (e.g. who does the position report to, size of team, examples of projects and business initiatives).

- Speak as a qualified and professional intellectual looking for the right job for you, not a displaced academic asking for a job.

A note on junior roles

- Landing an entry-level (or <2 years experience) job is a unique challenge due to the large number of applicants per opening who are qualified for the role on paper. Companies will be looking for reasons to shorten the list.
    - See previous slide on doing the homework on the company and the role.
    - They will focus on matching on the tools they use (telling them you're a "fast learner" isn't likely to do much, but specifics to back it up should), and fit for the role and "culture." If the job description mentions specific tools and skills, brush up on them before interviews. A candidate with a bachelor's degree who can convey familiarity with the relevant tools can be far more appealing than a PhD who can't.

- Being overqualified can be a disadvantage. Hiring managers can be weary of candidates who may outgrow the role quickly or use it as a stepping stone to a job somewhere else. You may need to work to convince them that you really want their job (instead of just any job) by citing specifics of why it's a good fit for you—e.g. mentorship, growth opportunity, affinity to company mission or culture.

I’m nervous that I can’t code well enough

- As mentioned in an earlier slide, DS roles can vary in how much they focus on software engineering and coding.

- Most roles don’t require you to write perfect code. Even if you have to google a bunch of stuff, if you code in a way that’s collaborative and sensible (e.g. descriptive variable names, comments, knowing when to write a function or class), that’s usually sufficient.

- For live technical interviews, treat it as a conversation. Explain what you’re thinking, let them know when you need a moment to think, and be honest when you’re stuck and need a hint. (Most interviewers I know would be fine with you saying “I know I can use such-and-such a function here, but I’d have to look up the exact syntax.”)

- You’ll find yourself getting more efficient with company data challenges over time, because many of the tasks are similar (explore data, build a quick model, validate, summarize your findings & recommendations).

Should I reach out to people to ask about specific opportunities?

- YES. Find someone on LinkedIn that's at that company, with any relevant connection whatsoever (a common acquaintance, same alma mater or former workplace, is a data scientist, etc.).

- This is because most companies offer referral bonuses (typically 5-10k) for candidates they end up hiring.

Is there a right time to apply?

- There’s no set “hiring season,” but this may vary from industry to industry. Jobs come and go quickly. So apply to a bunch as they come up.

- Starting near the end of the bonus cycle may impact bonus eligibility. (Bonuses may be paid at end of calendar year, end of fiscal year, or on 1-year anniversary of hire.)

Should I do a bootcamp?

- All things being equal, yes. Especially if you want help with current DS tools, jargon, and business thinking. It's also a great way to build your professional network. You will also get structured support around things like resume polishing, interview prep, and studying.

- It will constrain timing of job applications. (You're probably not going to be applying much during the bootcamp.)

- Bootcamps differ in what they offer and how they fit to individual needs. I was happy with Insight; I know others who weren't.

- My personal advice: don't pay upfront for a bootcamp or certification; make sure that it's free to you, or that the fees are contingent on finding a high-paying job.

Data science as a career

Some questions to consider about specific opportunities

- Are you motivated by the company's business? Are you morally comfortable helping the business?
- Is data science central to their business? (Or are they doing it because they feel like they have to?)
- What's the company culture like? How do they share and act on their values on topics like diversity & inclusion? Do they engage their community, and how?
- How good is the technical infrastructure? What does the rest of the tech org look like?
  e.g.: Do they have a clean data lake maintained by a data engineering team, or will you be hacking it together as you go?
  Are code base and ML models well documented?
- What will your career trajectory look like at this company? How will they support your growth?
- Are you comfortable with the people who'd be your teammates and manager? Are they supportive / do they give a damn?
- What will your first project be? Who do you go to for help when you don't know how to do something?
- What are their work hours actually like? Flexibility? How often will people expect you to respond to emails or Slack outside work hours? What's their PTO policy? Do people actually use their vacation days and unplug?
- Are company social functions important to you? How does the team/company socialize? Do people socialize outside work hours? How much alcohol is involved?
- What are the perks you like? (e.g. Foosball table, budget to attend conferences/classes?) How important are they to you?

How often and why do people change jobs in industry?

- Typical data science tenures are 2-3 years. Also not unusual for people to stay at a job for 5+ years, or move on after ~1 year.

- People tend to change jobs because they choose to.
  - Career/pay progression.
  - Looking for something new (different industry, different responsibilities, alignment with career goals, culture, work-life balance).
  - A job change may be planned, or you may be approached. (You will receive more and more communication from recruiters as you gain experience and seniority.)

Can I stay connected to my academic community?

- I think so. Some things I've done since moving to industry:
  - Was on a review paper with N>>1 authors.
  - Corresponded on scientific ideas.
  - Attended the AAS conference three times in four years, attended talks and posters. **There is a recurring "ATDS" splinter session** specifically meant to foster connections between academic and industry (ex-)astronomers.
  - I try to be a resource for academics who want to learn more about making the transition.

Some learning resources

- Andrew Ng's machine learning lectures
- towardsdatascience.com
- Chris Albon's machine learning flashcards
  (He has a standing offer for a 100% off coupon if you cannot afford the $12 price.)
- Books:
  - Cracking the Coding Interview
  - Statistical Learning (James, Witten, Hastie)
  - The Hundred-page Machine Learning Book
  - SQL for Data Scientists (Teate)
  - Build a Career in Data Science (Robinson, Nolis)

Feedback?

- Was this helpful?
- What would you change?
- I would love to hear from you!

www.linkedin.com/in/takatanaka
Twitter: @astrobassball