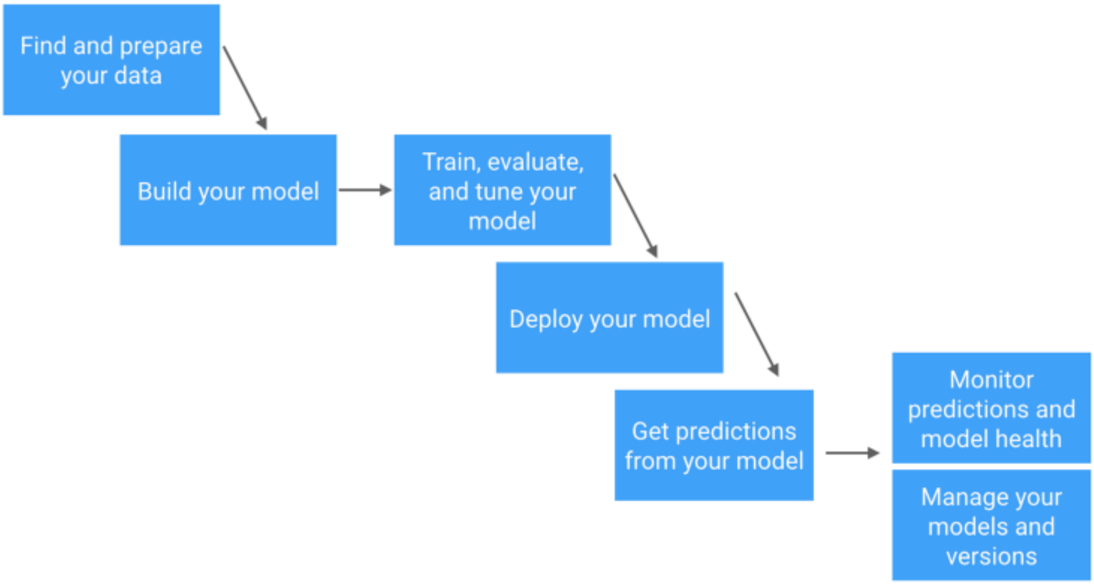
If you're a developer transitioning into data science, here are your best resources

1 April 2019


Find and prepare your data
Build your model
Train, evaluate, and tune your model
Deploy your model
Get predictions from your model
Monitor predictions and model health
Manage your models and versions

by Cecelia Shao

It seems like everyone wants to be a data scientist these days — from PhD students to data analysts to your old college roommate who keeps Linkedin messaging you to 'grab coffee'.

Perhaps you've had the same inkling that you should at least explore some data science positions and see what the hype is about. Maybe you've seen articles like Vicki Boykis' Data Science is different now that states:

> **What is becoming clear is that, in the late stage of the hype cycle, data science is asymptotically moving closer to engineering, and the skills that data scientists need moving forward are less**

> **visualization and statistics-based, and more in line with traditional computer science...:**

> Concepts like unit testing and continuous integration rapidly found its way into the jargon and the toolset commonly used by data scientist and numerical scientist working on ML engineering.

or tweets like Tim Hopper's:

What's not clear is how you can leverage your experience as a software engineer into a data science position. Some other questions you might have are:

What should I prioritize learning?

Are there best practices or tools that are different for data scientists?

Will my current skill set carry over to a data science role?

This article will provide a background on the data scientist role and why your background might be a good fit for data science, plus tangible stepwise actions that you, as a developer, can take to ramp up on data science.

> Want to see the latest data science roles? Subscribe to the biweekly ML Jobs Newsletter for new data science job openings in your inbox.

Data Scientist versus Data Engineer

First things first, we should distinguish between two complementary roles: Data Scientist versus Data Engineer. While both of these roles handle machine learning models, their interaction with these models as well as the the requirements and nature of the work for Data Scientists and Data Engineers vary widely.

> Note: The Data Engineer role that is specialized for machine learning can also manifest itself in job descriptions as 'Software Engineer, Machine Learning' or 'Machine Learning Engineers'

As part of a machine learning workflow, data scientist will perform the statistical analysis required to determine which machine learning approach to use then begin prototyping and building out those models.

Machine learning engineers will often collaborate with data scientists before and after this modeling process: (1) building data pipelines to feed data into these models and (2) design an engineering system that will serve these models to ensure continuous model health.

The diagram below is one way to view this continuum of skills:

There is a wealth of online resources on the difference between Data Scientists and Data Engineers — make sure to check out:

- Panoply: What is the difference between a data engineer and a data scientist?
- Springboard: Machine Learning Engineer vs Data Scientist
- O'Reilly: Data engineers vs. data scientists

As a disclaimer, this article primarily covers the Data Scientist role with some nod towards the Machine Learning Engineering side (especially relevant if you're looking at position in a smaller company where you might have to serve as both). If you're interested in seeing how you can transition to being a Data Engineer or Machine Learning Engineer, let us know in the comments below!

Your advantage as a developer

To everyone's detriment, classes around machine learning like 'Introduction to Data Science in Python' or Andrew Ng's Coursera course do *not* cover concepts and best practices from software engineering like unit testing, writing modular reusable code, CI/CD, or version control. Even some of the most advanced machine learning teams still do not use these practices for their machine learning code, leading to a disturbing trend…

Pete Warden described this trend as 'the Machine Learning Reproducibility Crisis':

> we're still back in the dark ages when it comes to tracking changes and rebuilding models from scratch. **It's so bad it sometimes feels like stepping back in time to when we coded without source control.**

While you may not see these 'software engineering' skills explicitly stated in data scientist job descriptions, having a good grasp of these skills as part of your background already will help 10x your work as a data scientist. Plus they'll come into use when it's time to answer those programming questions during your data science interview.

For some interesting perspective from the other side, check out Trey Causey's piece on 'Software development skills for data scientists' on skills that he recommends data scientists should learn to "write better code, interact better with software developers, and ultimately save you time and headaches".

Ramping up on data science

It's great that you have a good foundation with your software engineering background, but what's the next step towards becoming a data scientist? Josh Will's tongue-in-cheek tweet on the definition of a data scientist is surprisingly accurate:

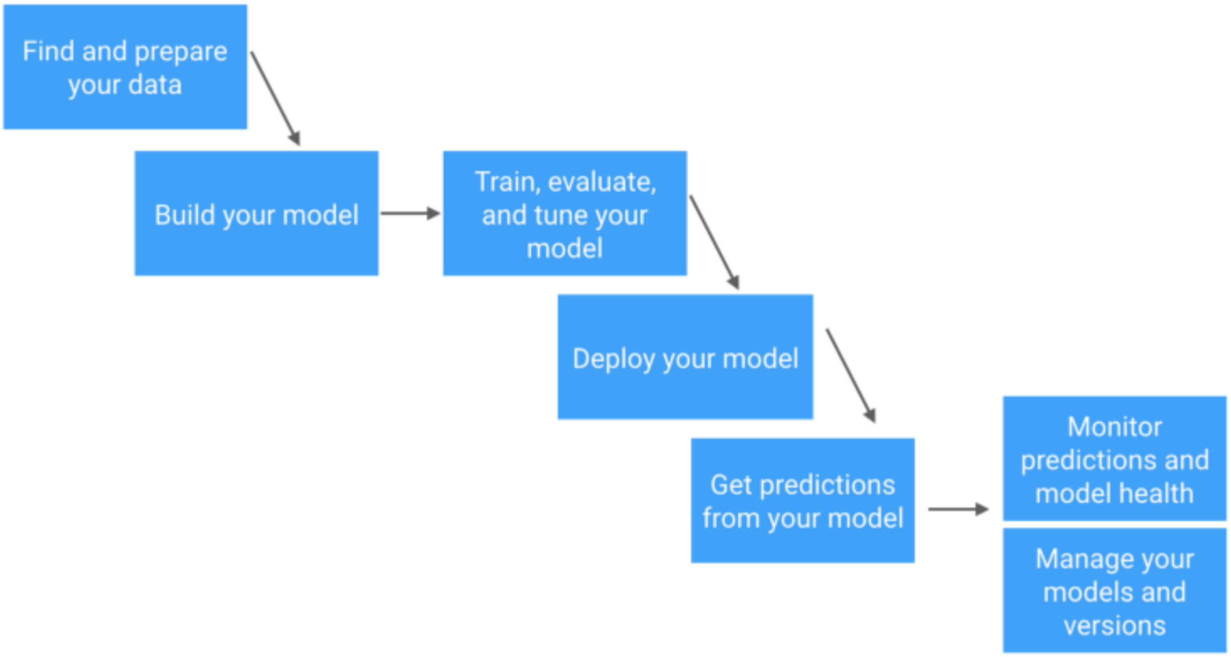
It hints at one of the topics you should catch up on if you're interested in pursuing a data scientist role or career: statistics. In this next section, we'll cover great resources for:

- **Building ML-specific knowledge**
- **Building industry knowledge**
- **Tools in the ML stack**
- **Skills and qualifications**

Building ML-specific knowledge

It's most effective to build a combination of theory-based knowledge around probability and statistics as well as applied skills in things like data wrangling or training models on GPUs/distributed compute.

One way to frame the knowledge you're gaining is to reference it against the machine learning workflow.


Find and prepare your data
Build your model
Train, evaluate, and tune your model
Deploy your model
Get predictions from your model
Monitor predictions and model health
Manage your models and versions

A simplified view of the machine learning workflow

> See this detailed workflow from Skymind AI

Here we list out some of the best resources you can find around machine learning. It would be impossible to have an exhaustive list and to save space (and reading time) we didn't mention very popular resources like Andrew Ng's Coursera course or Kaggle.

Courses:

- Fast.ai MOOC (free courses that teach very applied skills across Practical Deep Learning for Coders, Cutting Edge Deep Learning for Coders, Computational Linear Algebra, and Introduction to Machine Learning for Coders)
- Khan Academy
- 3Blue1Brown and mathematicalmonk youtube channel
- Udacity courses (including Preprocessing for Machine Learning in Python)
- Springboard AI/ML-specific track

Textbooks: **tried to find free PDFs online for most of these**

- Probabilistic Programming & Bayesian Methods for Hackers
- Probability and Random Processes
- Elements of Statistic Learning
- Linear Algebra Done Right
- Introduction to Linear Algebra
- Algorithm Design

Guides:

- Google Developers Machine Learning Guide
- Machine Learning Mastery Guides (for a good starting point, see this mini course on Python Machine Learning)
- Pyimagesearch (for computer vision)

Meetups: **primarily NYC-based ones**

- Papers We Love
- NYC Artificial Intelligence & Machine Learning
- DataCouncil.ai
- NY Artificial Intelligence

> For a cool starting point, check out Will Wolf's 'Open-Source Machine Learning Masters' on how you can structure your time across studying specific topics and working on projects to showcase expertise in a low-cost remote location.

Building industry-specific knowledge

If you have an inkling that you would like to be a specific industry like healthcare, financial services, consumer goods, retail, etc…, it is invaluable to catch up on the pain points and developments of that industry as it relates to data and machine learning.

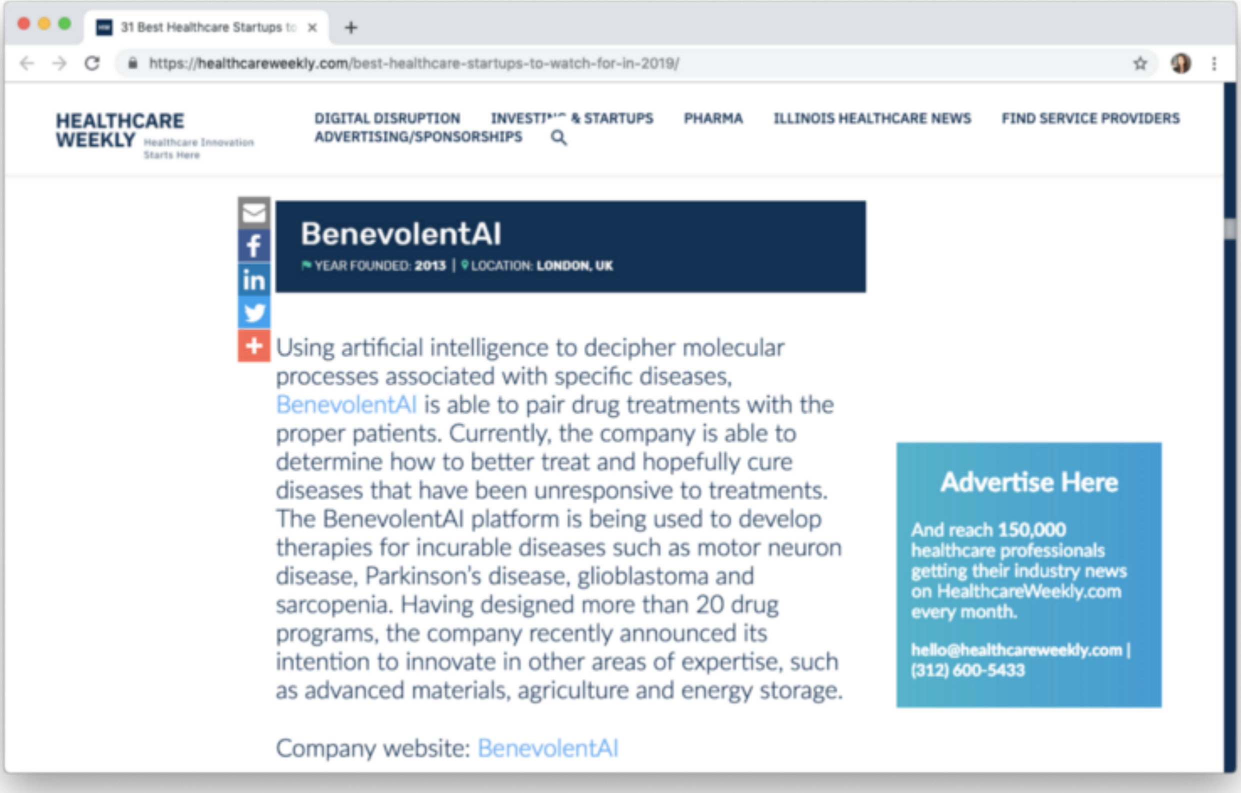
One pro tip = you can scan the websites of vertical-specific AI startups and see how they're positioning their value proposition and where machine learning comes into play. This will give you ideas for specific areas of machine learning to study and topics for projects to showcase your work.

We can walk through an example: let's say I'm interested in working in healthcare.

1. Through a quick google search for *"machine learning healthcare",* I found this list from Healthcareweekly.com on 'Best Healthcare Startups to Watch for in 2019'

> You can also do quick searches on Crunchbase or AngelList with "healthcare" as a keyword

2. Let's take one of the companies featured on the list, BenevolentAI, as an example.


31 Best Healthcare Startups to
https://healthcareweekly.com/best-healthcare-startups-to-watch-for-in-2019/
HEALTHCARE WEEKLY Healthcare Innovation Starts Here
DIGITAL DISRUPTION
INVESTI** & STARTUPS
PHARMA
ILLINOIS HEALTHCARE NEWS
FIND SERVICE PROVIDERS
ADVERTISING/SPONSORSHIPS

BenevolentAI

YEAR FOUNDED: 2013 | LOCATION: LONDON, UK

Using artificial intelligence to decipher molecular processes associated with specific diseases, BenevolentAI is able to pair drug treatments with the proper patients. Currently, the company is able to determine how to better treat and hopefully cure diseases that have been unresponsive to treatments. The BenevolentAI platform is being used to develop therapies for incurable diseases such as motor neuron disease, Parkinson's disease, glioblastoma and sarcopenia. Having designed more than 20 drug programs, the company recently announced its intention to innovate in other areas of expertise, such as advanced materials, agriculture and energy storage.

Company website: BenevolentAI

3. BenevolentAI's website states:

> We are an AI company with end-to-end capability from early drug discovery to late-stage clinical development. BenevolentAI combines the power of computational medicine and advanced AI with the principles of open systems and cloud computing to transform the way medicines are designed, developed, tested and brought to market.

> We built the Benevolent Platform to better understand disease and to design new, and improve existing treatments, from vast quantities of biomedical information. We believe our technology empowers scientists to develop medicines faster and more cost-efficiently.

> A new research paper is published every 30 seconds yet scientists currently only use a fraction of the knowledge available to understand the cause of disease and propose new treatments. Our platform ingests, 'reads' and contextualises vast quantities of information drawn from written documents, databases and experimental results. It

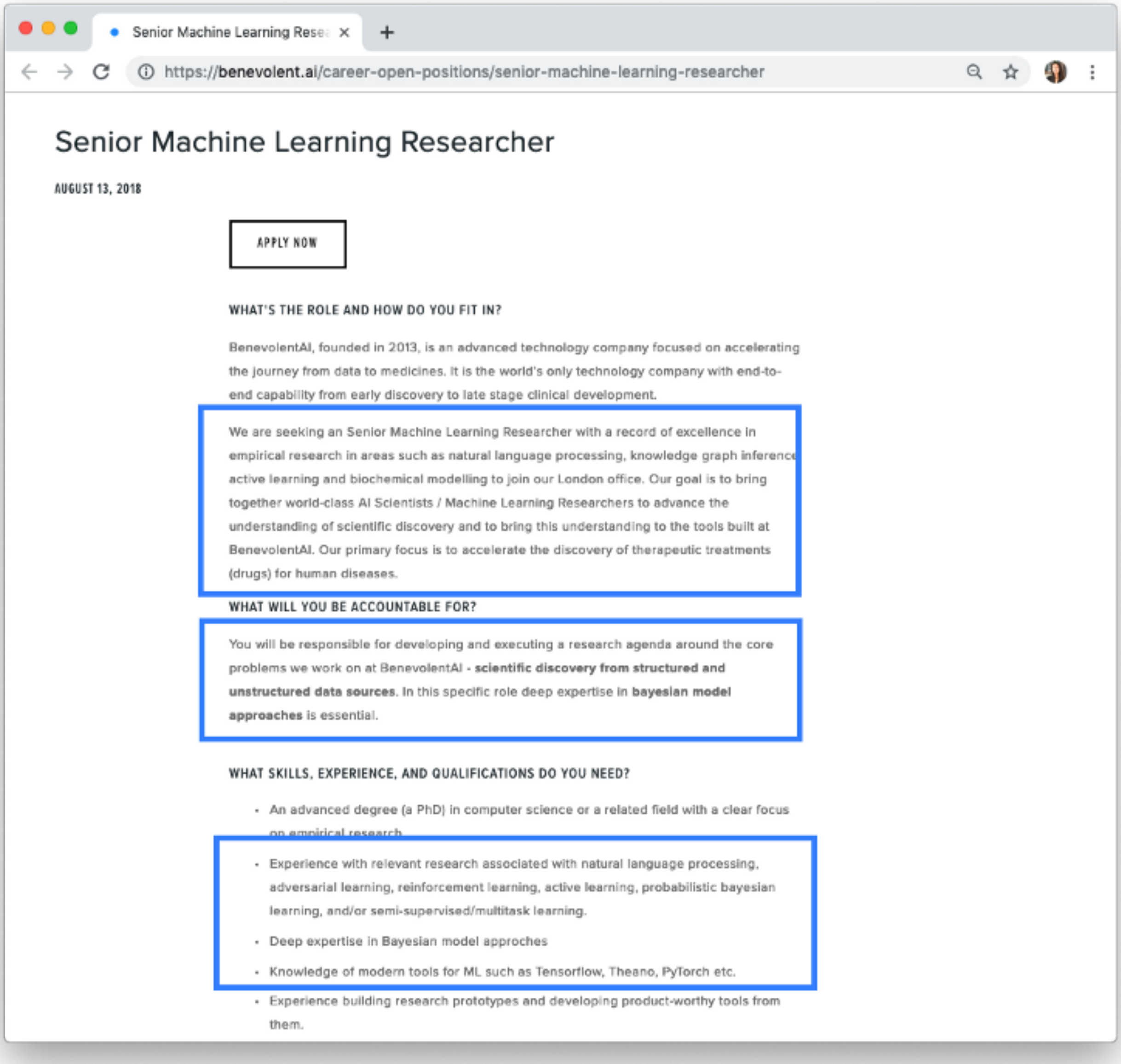
> is able to make infinitely more deductions and inferences across these disparate, complex data sources, identifying and creating relationships, trends and patterns, that would be impossible for a human being to make alone.

4. Immediately you can see that BenevolentAI is using natural language processing (NLP) and are probably working with some knowledge graphs if they're identifying relationships between diseases and treatment research

5. If you check BenevolentAI's career page, you can see that they're hiring for a Senior Machine Learning Researcher. This is a senior role, so it's not a perfect example, but take a look at the skills and qualifications they're asking for below:

Note:

- natural language processing, knowledge graph inference, active learning and biochemical modeling
- structured and unstructured data sources
- bayesian model approaches
- knowledge of modern tools for ML


Senior Machine Learning Rese
https://benevolent.ai/career-open-positions/senior-machine-learning-researcher
APPLY NOW

Senior Machine Learning Researcher

AUGUST 13, 2018

WHAT'S THE ROLE AND HOW DO YOU FIT IN?

BenevolentAI, founded in 2013, is an advanced technology company focused on accelerating the journey from data to medicines. It is the world's only technology company with end-to-end capability from early discovery to late stage clinical development.

We are seeking an Senior Machine Learning Researcher with a record of excellence in empirical research in areas such as natural language processing, knowledge graph inference active learning and biochemical modelling to join our London office. Our goal is to bring together world-class AI Scientists / Machine Learning Researchers to advance the understanding of scientific discovery and to bring this understanding to the tools built at BenevolentAI. Our primary focus is to accelerate the discovery of therapeutic treatments (drugs) for human diseases.

WHAT WILL YOU BE ACCOUNTABLE FOR?

You will be responsible for developing and executing a research agenda around the core problems we work on at BenevolentAI - **scientific discovery from structured and unstructured data sources**. In this specific role deep expertise in **bayesian model approaches** is essential.

WHAT SKILLS, EXPERIENCE, AND QUALIFICATIONS DO YOU NEED?

- An advanced degree (a PhD) in computer science or a related field with a clear focus on empirical research
- Experience with relevant research associated with natural language processing, adversarial learning, reinforcement learning, active learning, probabilistic bayesian learning, and/or semi-supervised/multitask learning.
- Deep expertise in Bayesian model approches
- Knowledge of modern tools for ML such as Tensorflow, Theano, PyTorch etc.
- Experience building research prototypes and developing product-worthy tools from them.

This should give you some steps for what to approach next:

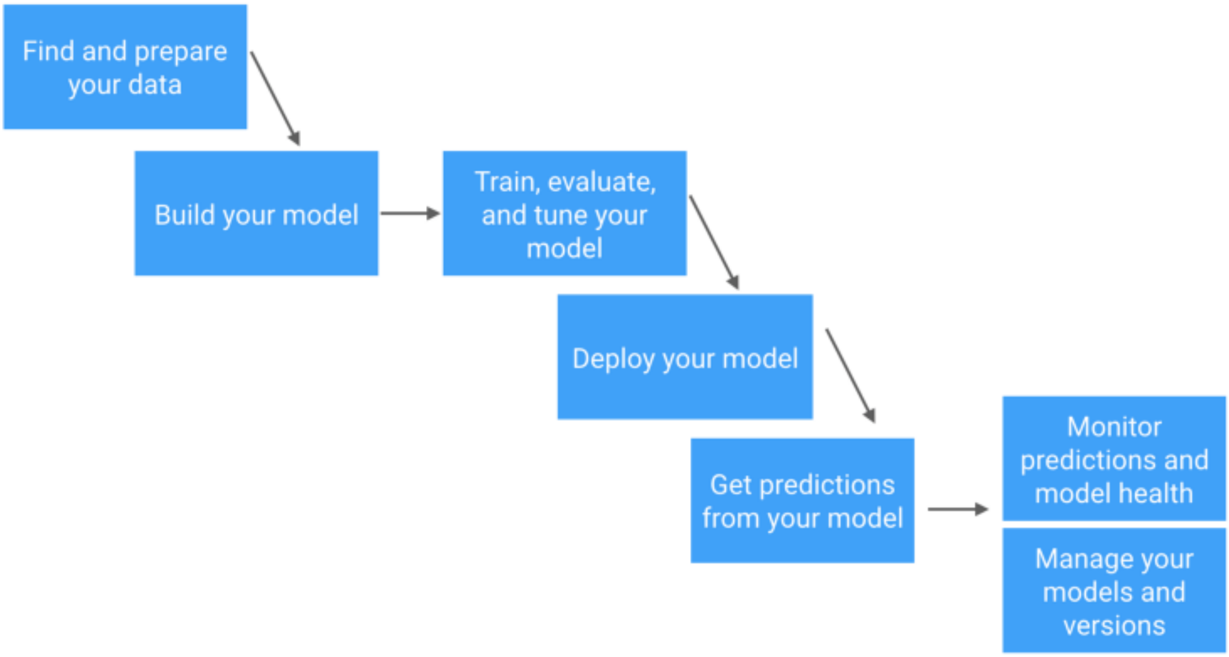
- working with structured data
- working with unstructured data
- classifying relationships in knowledge graphs (see a good resource here)
- learning bayesian probability and modeling approaches
- work on an NLP project (so text data)

We’re not recommending that you apply to the companies you find through your search, but rather see how they describe their customer’s pain points, their company's value propositions, and what kind of skills they list in their job descriptions to guide your research.

Tools in the ML stack

In the BenevolentAI Senior Machine Learning Researcher job description, they ask for *"knowledge of modern tools for ML, such as Tensorflow, PyTorch, etc…"*

Learning these modern tools for ML can seem daunting since the space is always changing. To break up the learning process into manageable pieces, remember to anchor your thinking around the machine learning workflow from above — *"What tool can help me with this part of the workflow?"* ?


Find and prepare your data
Build your model
Train, evaluate, and tune your model
Deploy your model
Get predictions from your model
Monitor predictions and model health
Manage your models and versions

To see which tools accompany each step of this machine learning workflow, check out Roger Huang's 'Introduction to the Machine Learning Stack' which covers tools like Docker, Comet.ml, and dask-ml.

Tactically speaking, Python and R are the most common programming languages data scientists use and you can will encounter add-on packages designed for data science applications, such as NumPy and SciPy, and matplotlib. These languages are interpreted, rather than compiled, leaving the data scientist free to focus on the problem rather than nuances of the language. It's worth investing time learning object-oriented programming to understand the implementation of data structures as classes.

To catch up on ML frameworks like Tensorflow, Keras, and PyTorch, make sure to go to their documentation and try implementing their tutorials end-to-end.

At the end of the day, you want to make sure that you're building out projects that showcase these modern tools for data collection and wrangling, machine learning experiment management, and modeling.

For some inspiration for your projects, check out Edouard Harris's piece on 'The cold start problem: how to build your machine learning portfolio'

Skills and qualifications

We left this section for last since it aggregates much of the information from the previous sections, but is specifically geared towards data science interview preparation. There are six main topics during a data scientist interview:

1. Coding
2. Product
3. SQL
4. A/B testing
5. Machine Learning
6. Probability (see a good definition vs. Statistics here)

You'll notice that one of these topics is not like the others (Product). For data science positions, communication about technical concepts and results as well as business metrics and impact is crucial.

> **Some useful aggregations of data science interview questions:**

> ?? https://github.com/kojino/120-Data-Science-Interview-Questions

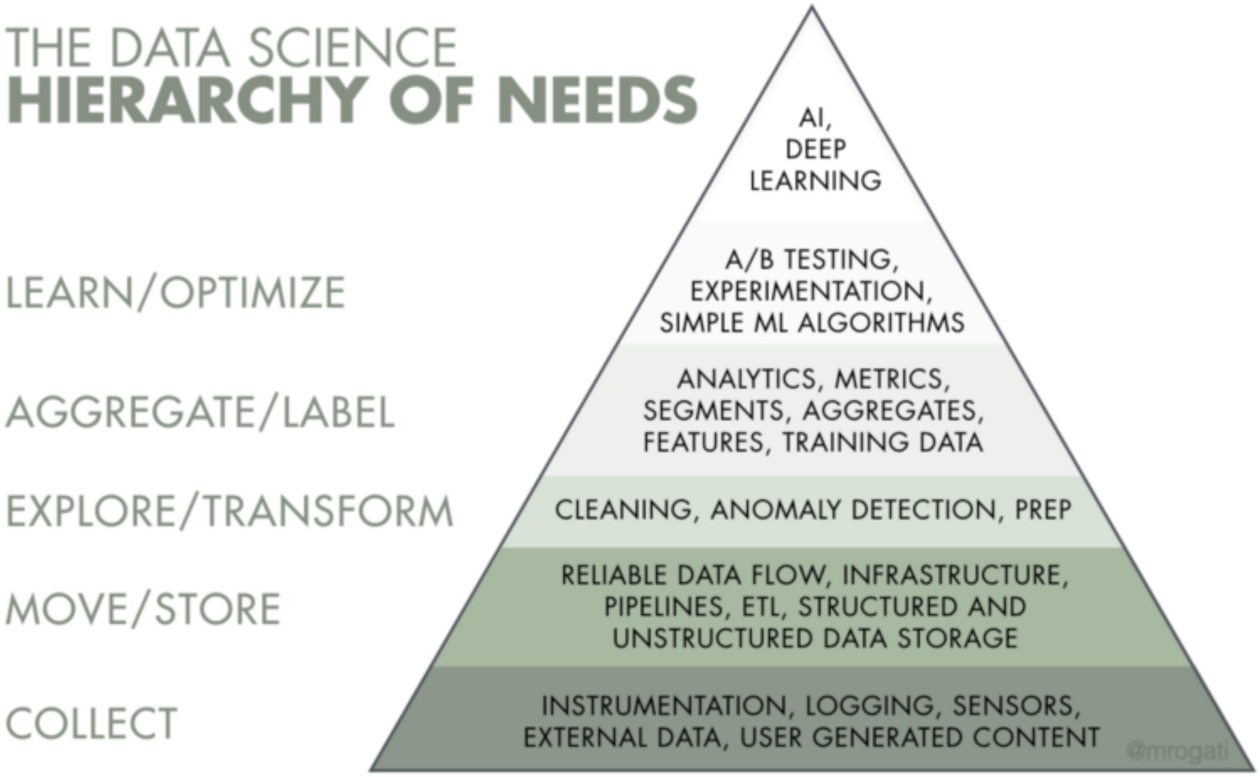
> ??https://github.com/iamtodor/data-science-interview-questions-and-answers

> ???? https://hookedondata.org/red-flags-in-data-science-interviews/

> ?? https://medium.com/@XiaohanZeng/i-interviewed-at-five-top-companies-in-silicon-valley-in-five-days-and-luckily-got-five-job-offers-25178cf74e0f

You'll notice that we included Hooked on Data's piece on 'Red Flags in Data Science Interviews' — as you interview for roles, you'll come across companies who are still building up their data infrastructure or may not have a solid understanding of how their data science team fits into the larger company value.

These companies may still be climbing up this hierarchy of needs below.


THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

The popular AI Hierarchy of Needs from Monica Rogati

For some expectation setting around data science interviews, I would recommend reading Tim Hopper's piece on 'Some Reflections on Being Turned Down for a Lot of Data Science Jobs'

Thanks for reading! We hope this guide helps you understand if data science is a career you should consider and how to begin that journey!

Want to see the latest data science roles? Subscribe to the biweekly ML Jobs Newsletter for new data science job openings in your inbox:

ML Jobs Newsletter - Revue
*Sign up to receive this biweekly curated list of data science job openings at the best companies in the industry. Roles…*www.getrevue.co

If this article was helpful, tweet it or share it.

Donate $5 and buy the world 250 hours of learning.