Hiring Your First Data Scientist

Boston Data Festival 2016

Terran Melconian

terran@jobcase.com

https://www.jobcase.com/p/terran.melconian

twitter @terranmelconian

Motivation

I never want to hear this again:

“We hired a data scientist once…
and we couldn’t use any of the things he produced, so after six months we fired him.”

Role-Defining Questions

- What do you expect your projects to deliver?
- What do you have as input for your projects?
- How quickly do you need to have the results?

Deliverable: Money

- It is not time to hire a data scientist yet
- First develop a more specific plan for *how* you are going to make the money.
- Finding ways to turn technical projects into revenue is an executive or management function, not a data scientist function.
- Hire an experienced manager or start with a consultancy to identify and scope opportunities.

Deliverable: Reports

- Data Analyst, Business Intelligence
- Produces reports of varying polish, from Excel files for internal use to shiny reports deliverable to customers
- Translate data to business actions
  - e.g. manage advertising budgets and allocations to performance targets, target features to customer segments
- To get nice reports, plan to also spend on BI tooling
- No process automation unless provided by vendor

Deliverable: Data

- Data Scientist, Predictive Modeling
- Produces some kind of data file which you can then load into your operational systems
  - Product categories, cross-sell recommendations, user preferences, ad bidding, etc.
  - Latency is usually daily at best
- Uses modeling tools in R, Python, or commercial products
  - Probably already has a preferred/familiar toolset.
- Many uses require allocation of engineering resources to consume results

Deliverable: Final Code

- Machine Learning Engineer, Data Scientist
  - It is unfortunate that “data scientist” is used to describe these two distinct skill sets
- Specialized member of your engineering team.
- Writes algorithms which run in your live systems or product
  - e.g. search, live content recommendations, real time advertising bidding
- May or may not also do offline analyses with off-the-shelf tools

Deliverable: Publications

- Research Scientist
- Getting papers published is a specialized skill
- Look for prior publication history in the same field

Input: No Idea

X

- Don’t know what you have or what form it’s in
- Not time to hire a data scientist yet
- You need a Director of Business Intelligence or a consultant who has experiencing finding data assets

Input: To Be Purchased

- Examples:
  - Going into TV ads; need demographics for targeting
    - Buy additional data to merge to customer database
- Buying data is not a core data science skill – look for specific experience.
- Business consulting background may be useful

Input: Analytics Data

- What makes it analytic data?
  - You can get the data out quickly (minutes)
  - You can get it without worrying about operational impact
- Your existing systems might already meet these requirements
- This is the easiest place to be; congratulations.

Input: Unsuitable Data

- Must first transform and store your data in a form and technology which meets analytics requirements
- Generally a separate hire from a data scientist, in your engineering or operations department
- Data Warehousing, Data Engineering, Data Architecture

Small Data

- Fewer than 1,000 rows
  - You want a statistician
- 1,000 to 100,000 rows
  - Your project selection should be influenced by your data availability.
  - “Do we have enough data to answer this question?”

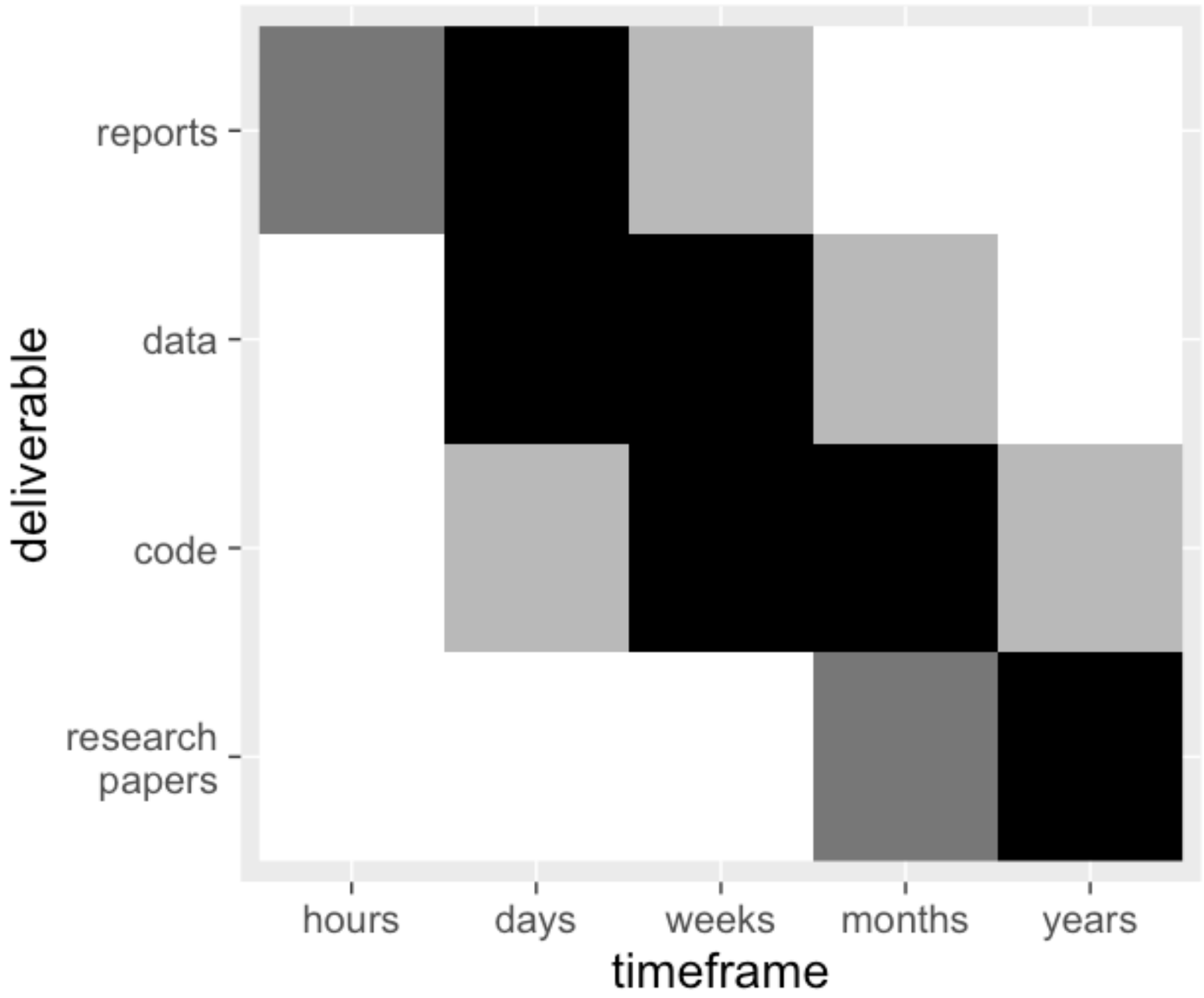
Medium Data

- up to ~2 GB: In memory
  - This is the easiest case and you will have the most options for tooling and algorithms.
- up to ~200 GB: On disk, one machine
  - You will have some constraints on the types of models and algorithms you can use.

Big Data

- More than fits on one machine: Big Data
  - Can you do your projects on a subset of your data?
  - Working with big data requires different algorithms and tools; look for this experience specifically
- For a long-running cluster with ongoing data flows, need a data warehouse operations team

Project Duration



Hiring: Use An Exercise

- Exercise protects you from your own misconceptions about requirements
- Therefore you must make it as realistic as possible, subject to constraints of hiring.
  - Type of deliverable
  - Type of input data
  - Size of input data
  - Adjust scope to make project achievable

Exercise: Data

- Take actual data from a real project. Scrub identifying information if necessary and leave as much as possible as-is.
- Match real format, e.g. database dump
- Is there labeled data (i.e. truth/what happened)?
- If data is large, give the candidate a pre-imaged cloud machine, or cluster of machines, to use.

Exercise: Instructions

- Framing of the question is very important
  - "Make a model to predict which prospects will convert" is very different from "recommend how to improve conversion".
  - Who is translating from business requirements to task definition?
- Tell them who their audience is
- If you are hiring someone to write code, they must deliver code and someone must read it

Exercise: Evaluation (1)

- Did they answer the question you posed, at the correct level of generality?
- Put the people who will actually be consuming their work product in the room, not just the hiring manager and peers.
  - Did they communicate in an appropriate way which inspired confidence in this audience?
  - Would this deliverable be accepted and used?

Exercise: Evaluation (2)

- All real projects involve feedback and iteration
  - Provide a realistic simulation of one feedback iteration as part of the exercise. Be clear on roles.
- Ask “suppose we roll this out and it doesn’t work; what do you do?”
  - All tools have limitations – does the candidate understand these limitations and how to address?

Summary

- Start from a clear understanding of what your data science project will deliver and how it will advance your business goals.
- Hire with an exercise which has realistic inputs and a realistic deliverable, evaluated by your actual users of the deliverable.
- If nobody can pass your exercise, consider splitting the role into two smaller roles.

Q&A

Terran Melconian

terran@jobcase.com terran@consistent.org

https://www.jobcase.com/p/terran.melconian

twitter @terranmelconian