2018 Big Data Spain, Madrid

Building A Data Science Team From Scratch

Catherine Zhou

SENIOR DATA SCIENTIST + MANAGER

codecademy

Building A Data Science Team From Scratch

… cómo iniciar un equipo de data science cuando todos usan Excel

[how to start a data science team when everyone is still using Excel]

About Me_

twitter @catherinezh

#rstats

#rladies

#codecademy

- Based in New York City
- Data Scientist Manager @ Codecademy
  - Company's first data science hire
- 10+ years of statistical programming
  - 6 yrs of professional work
  - 4 yrs of technical hiring & interviewing

TOPIC

Lessons Learned from Both Sides of the Hiring Table

codecademy

Agenda

By the end of this talk you will be able to:

- **Hire the right type** of data professional
- Decide how to **structure your data team**
- Conduct a **data science technical screen**
- **Identify the right tools** for assessments
- When and how to **scale up your team**

codecademy

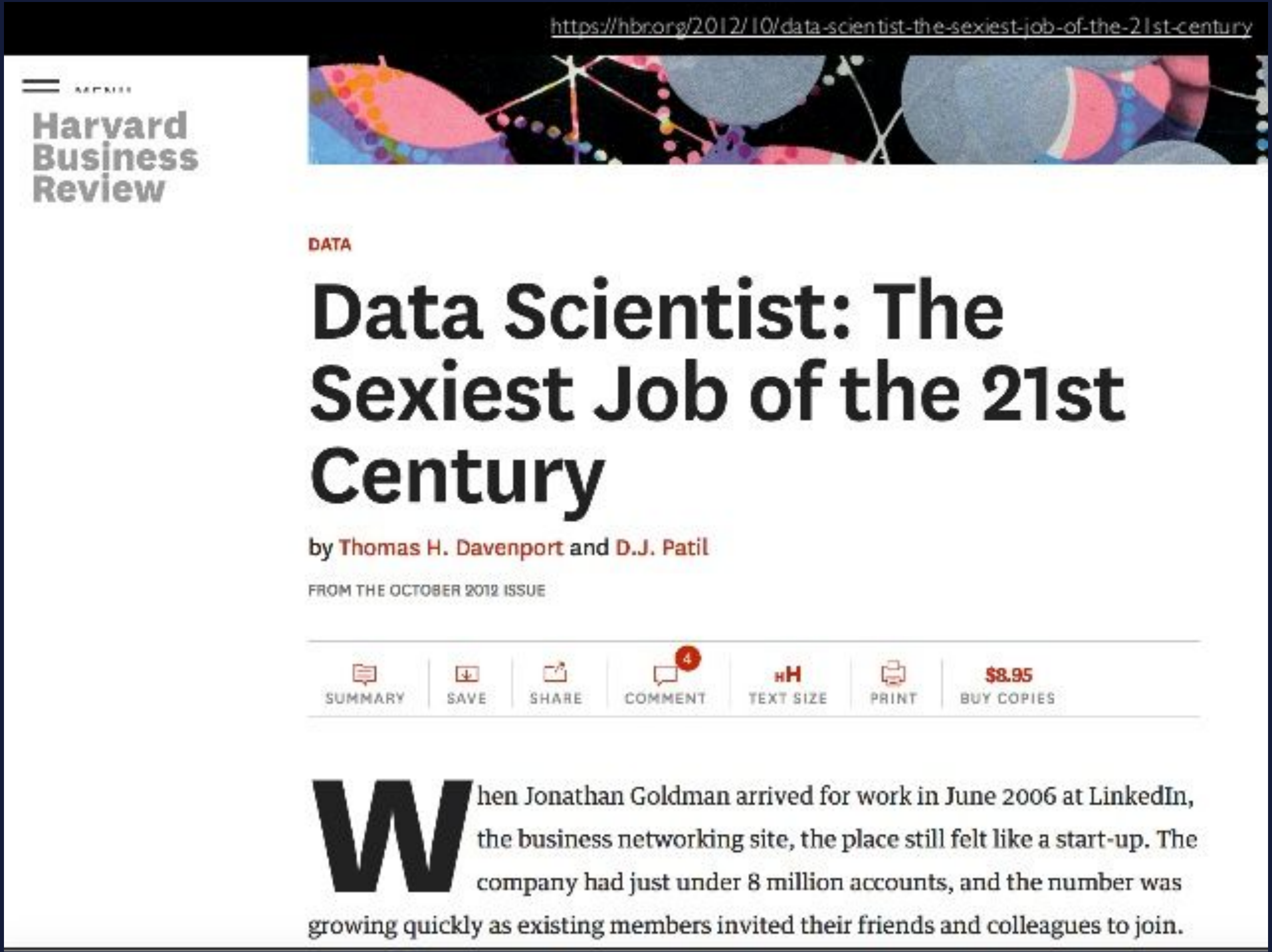
PART ONE

Data Science Hiring_

codecademy

LESSON ONE_

Everybody wants to hire data scientists,
but no one can agree on a job title and description.


https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
MENU
Harvard Business Review

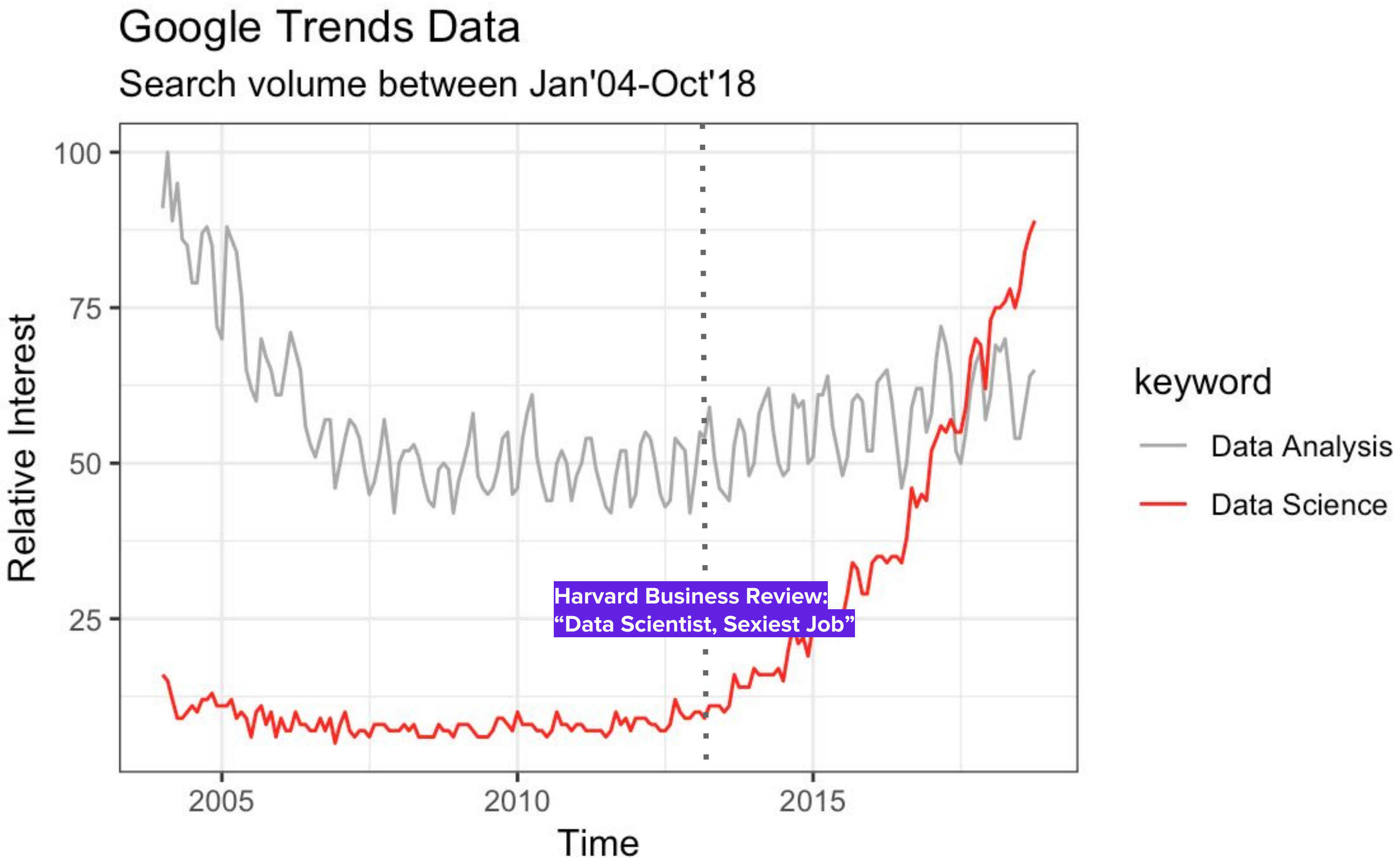
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.

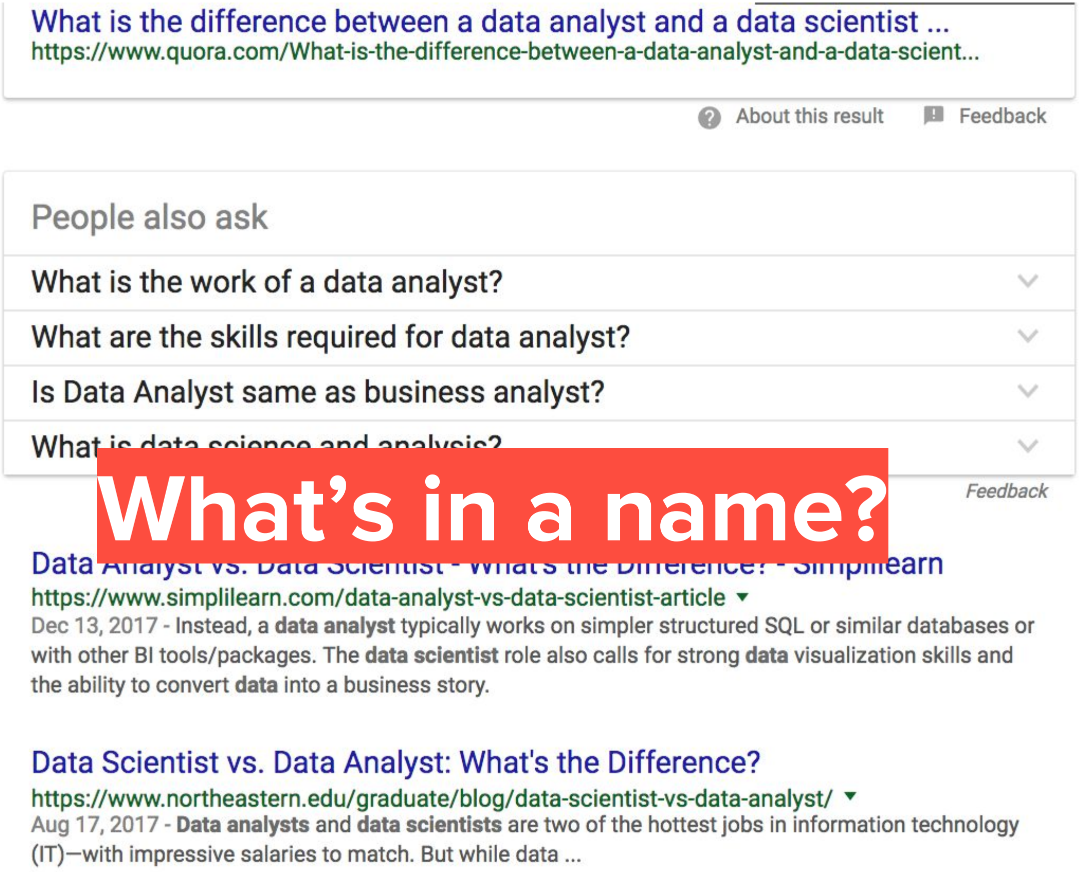


Google Trends Data

Search volume between Jan'04-Oct'18

Harvard Business Review:
“Data Scientist, Sexiest Job”

What’s in a name?



What is the difference between a data analyst and a data scientist ...
https://www.quora.com/What-is-the-difference-between-a-data-analyst-and-a-data-scient...

People also ask

What is the work of a data analyst?

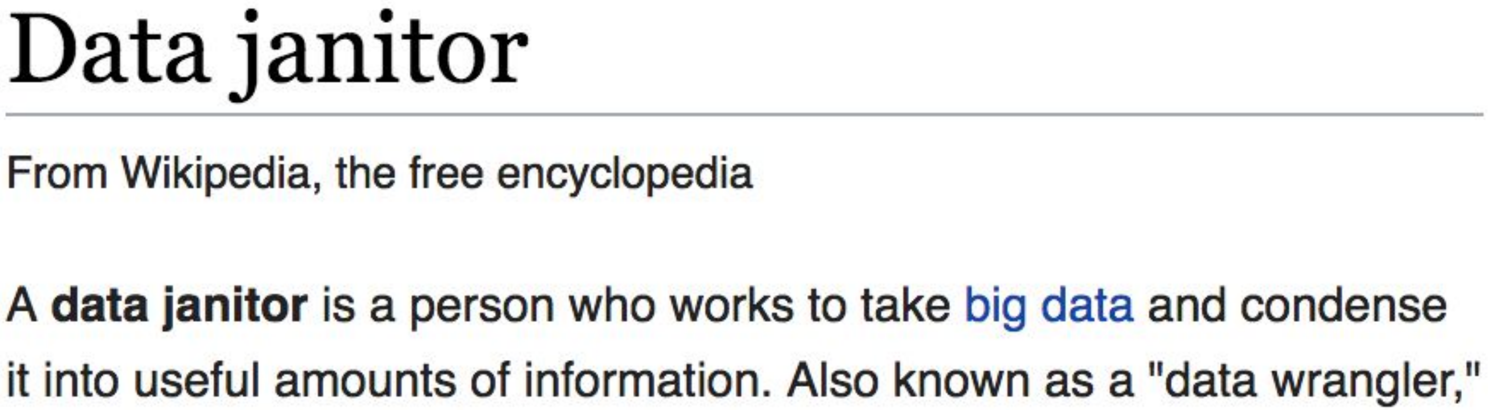
What are the skills required for data analyst?

Is Data Analyst same as business analyst?

What is data science and analysis?

Data Analyst vs. Data Scientist - What's the Difference? - Simplilearn
https://www.simplilearn.com/data-analyst-vs-data-scientist-article
Dec 13, 2017 - Instead, a **data analyst** typically works on simpler structured SQL or similar databases or with other BI tools/packages. The **data scientist** role also calls for strong **data** visualization skills and the ability to convert **data** into a business story.

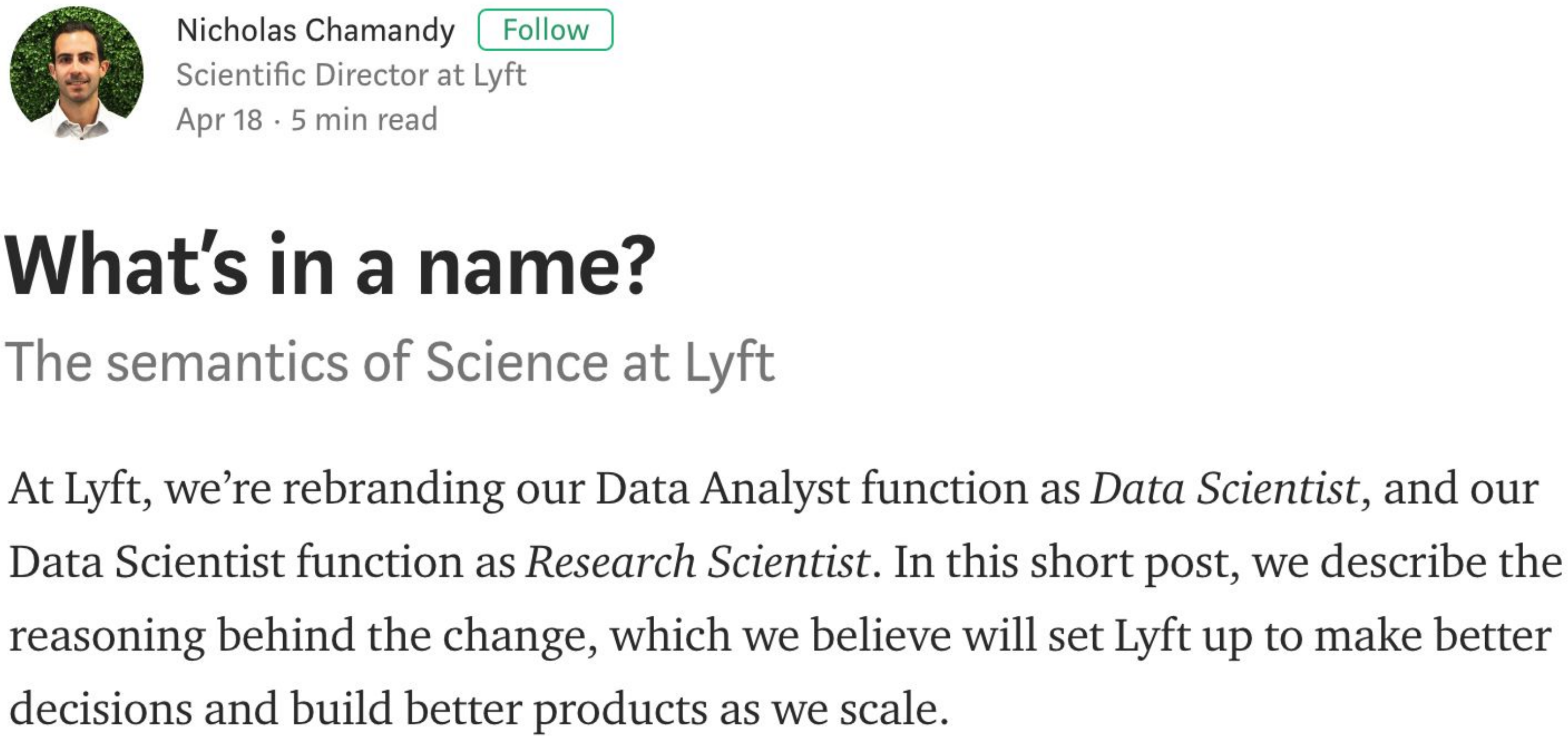
Data Scientist vs. Data Analyst: What's the Difference?
https://www.northeastern.edu/graduate/blog/data-scientist-vs-data-analyst/
Aug 17, 2017 - **Data analysts** and **data scientists** are two of the hottest jobs in information technology (IT)—with impressive salaries to match. But while data ...



Data janitor

From Wikipedia, the free encyclopedia

A **data janitor** is a person who works to take big data and condense it into useful amounts of information. Also known as a "data wrangler,"



Nicholas Chamandy

Scientific Director at Lyft

Apr 18 · 5 min read

What's in a name?

The semantics of Science at Lyft

At Lyft, we're rebranding our Data Analyst function as *Data Scientist*, and our Data Scientist function as *Research Scientist*. In this short post, we describe the reasoning behind the change, which we believe will set Lyft up to make better decisions and build better products as we scale.

LESSON ONE (CONT.)

Types of Data Scientists_

DATA SCIENTIST TYPE A

VS.

DATA SCIENTIST TYPE B

- *Michael Hochster, Data Science Director at Stitch Fix (Quora 2014)*

DATA SCIENTIST TYPE ANALYZE

VS.

DATA SCIENTIST TYPE B

- Michael Hochster, Data Science Director at Stitch Fix (Quora 2014)

codecademy

DATA SCIENTIST TYPE ANALYZE

Primarily concerned with making sense of data or working with it in a fairly static way. Very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.

VS.

DATA SCIENTIST TYPE BUILD

- *Michael Hochster, Data Science Director at Stitch Fix* (Quora 2014)

DATA SCIENTIST TYPE ANALYZE

Primarily concerned with making sense of data or working with it in a fairly static way. Very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.

VS.

DATA SCIENTIST TYPE BUILD

- *Michael Hochster, Data Science Director at Stitch Fix (Quora 2014)*

DATA SCIENTIST TYPE ANALYZE

Primarily concerned with making sense of data or working with it in a fairly static way. Very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.

VS.

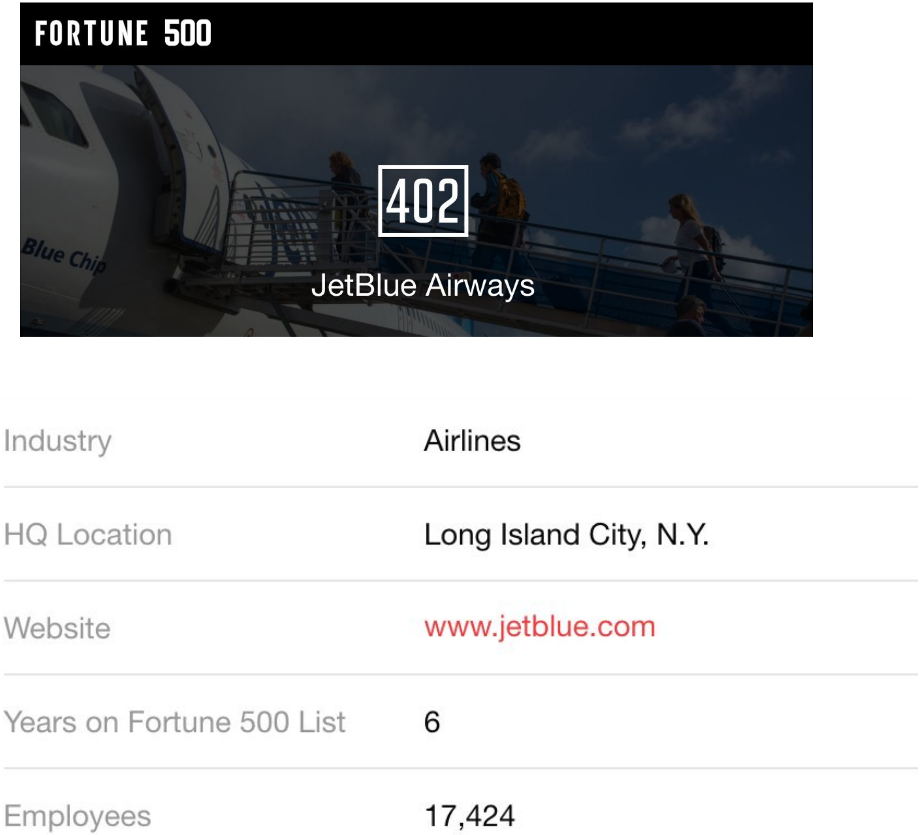
DATA SCIENTIST TYPE BUILD

Share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. Mainly interested in using data "in production." They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results).

- *Michael Hochster, Data Science Director at Stitch Fix (Quora 2014)*

LESSON TWO_

The stage your organization is in will guide the type of data scientist you should hire.

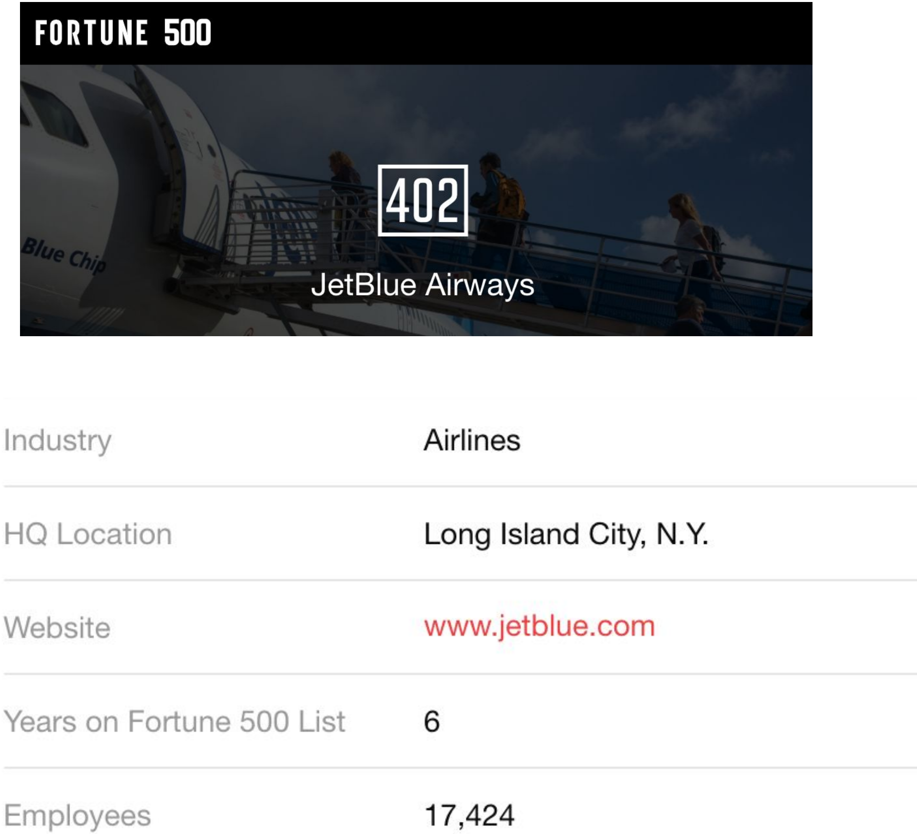


FORTUNE 500

402

JetBlue Airways

Industry	Airlines
HQ Location	Long Island City, N.Y.
Website	www.jetblue.com
Years on Fortune 500 List	6
Employees	17,424

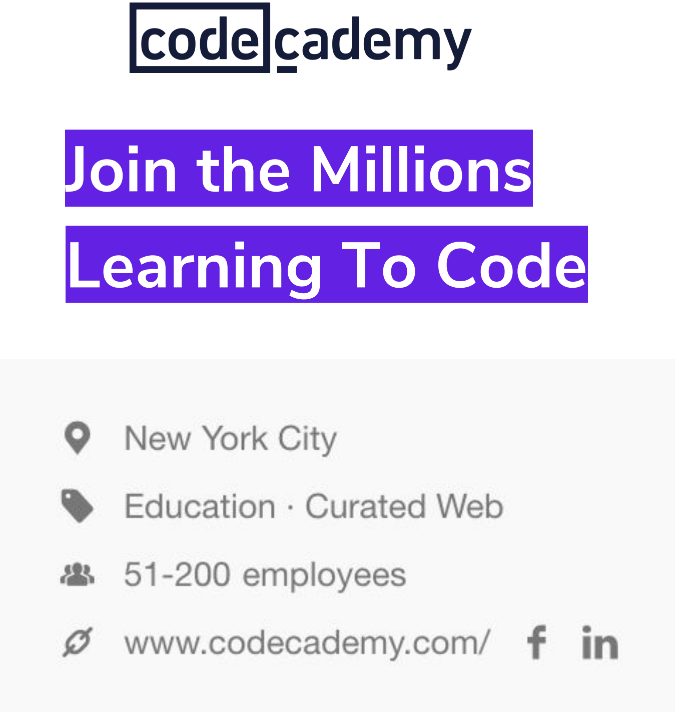


FORTUNE 500

402

JetBlue Airways

Industry	Airlines
HQ Location	Long Island City, N.Y.
Website	www.jetblue.com
Years on Fortune 500 List	6
Employees	17,424



codecademy

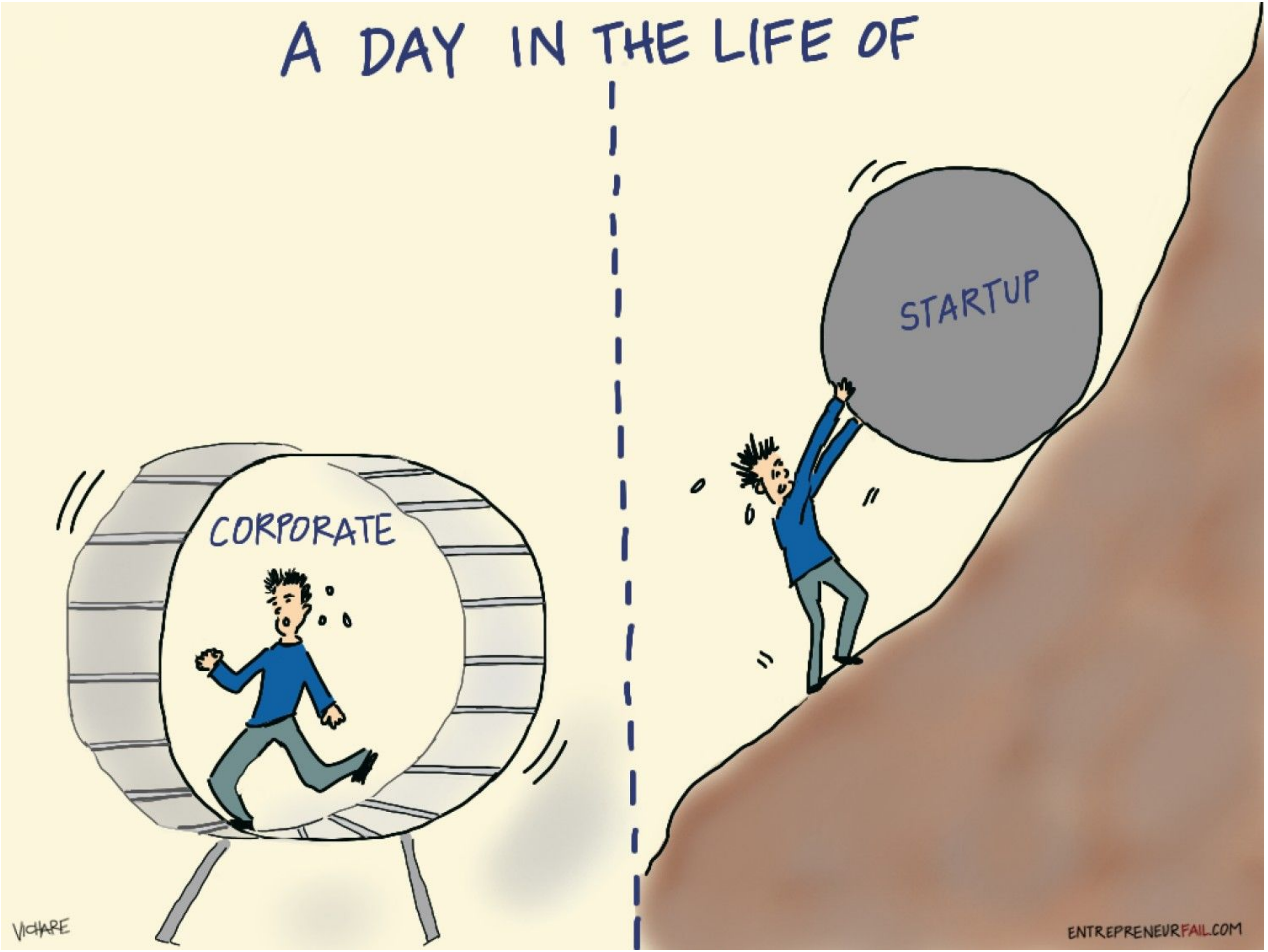
Join the Millions Learning To Code

New York City

Education · Curated Web

51-200 employees

www.codecademy.com/ f in


A DAY IN THE LIFE OF
CORPORATE
STARTUP
VICHARE
ENTREPRENEURFAIL.COM

WHAT STAGE ARE YOU IN?

Assess what type of Data Science work your organization needs.

DATA SCIENCE FOCUS AREAS

Early Stage:

- Data collection phase: database design, ETL
- Build foundation: biz intelligence, reporting

DATA SCIENCE FOCUS AREAS

Early Stage:

- Data collection phase: database design, ETL
- Build foundation: biz intelligence, reporting

Growth Stage:

- Define & measure key metrics, produce insights
- Run experiments on new and existing features

DATA SCIENCE FOCUS AREAS

Early Stage:

- Data collection phase: database design, ETL
- Build foundation: biz intelligence, reporting

Growth Stage:

- Define & measure key metrics, produce insights
- Run experiments on new and existing features

At Scale:

- Develop and deploy models into production.
- Automate analysis at scale

codecademy

WHAT IS THE STATE OF YOUR DATA WAREHOUSE?

Important note:

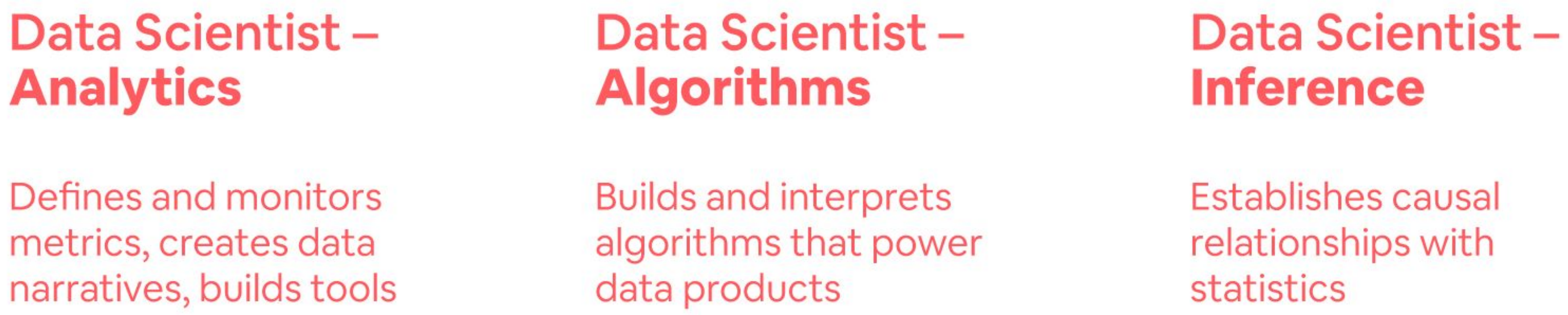
The age of the company is not inherently tied to the state the data warehouse is in.

Older companies often possess technical debt from ties to legacy software and old infrastructure.

LESSON THREE_

Start with DS generalists before specialists.
Strong teams come from diverse backgrounds.

DATA SCIENTIST TRACKS



Data Scientist – **Analytics**

Defines and monitors metrics, creates data narratives, builds tools

Data Scientist – **Algorithms**

Builds and interprets algorithms that power data products

Data Scientist – **Inference**

Establishes causal relationships with statistics

- *Elena Grewal, Head of Data Science at Airbnb (2018)*

PART TWO

Technical Interviews_

codecademy

LESSON FOUR_

Structure effective data science interviews to weed out the data science “experts”.

CODECADEMY HIRING PROCESS

1. Resume Review
2. Phone Interview
3. Pair Programming
4. Take-Home Project
5. On-Site Interview

Screening for:

Technical Skill & Business Sense_

Phone interview

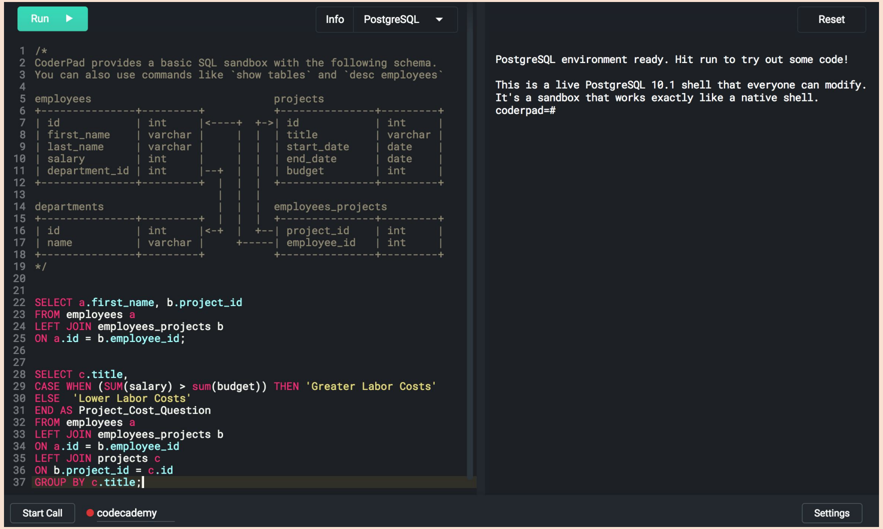
Question 1:

What's your favorite R or Python package, and why?

Question 2:

Tell me about an analytical project you worked on recently.

CODECADEMY HIRING PROCESS



```sql
/*
CoderPad provides a basic SQL sandbox with the following schema.
You can also use commands like `show tables` and `desc employees`

employees                             projects
+---------------+---------+           +---------------+---------+
| id            | int     |<----+  +->| id            | int     |
| first_name    | varchar |     |  |  | title         | varchar |
| last_name     | varchar |     |  |  | start_date    | date    |
| salary        | int     |     |  |  | end_date      | date    |
| department_id | int     |--+  |  |  | budget        | int     |
+---------------+---------+  |  |  |  +---------------+---------+
                             |  |  |
departments                  |  |  |  employees_projects
+---------------+---------+  |  |  |  +---------------+---------+
| id            | int     |<-+  |  +--| project_id    | int     |
| name          | varchar |     +-----| employee_id   | int     |
+---------------+---------+           +---------------+---------+
*/


SELECT a.first_name, b.project_id
FROM employees a
LEFT JOIN employees_projects b
ON a.id = b.employee_id;


SELECT c.title,
CASE WHEN (SUM(salary) > sum(budget)) THEN 'Greater Labor Costs'
ELSE  'Lower Labor Costs'
END AS Project_Cost_Question
FROM employees a
LEFT JOIN employees_projects b
ON a.id = b.employee_id
LEFT JOIN projects c
ON b.project_id = c.id
GROUP BY c.title;
```

```
PostgreSQL environment ready. Hit run to try out some code!

This is a live PostgreSQL 10.1 shell that everyone can modify.
It's a sandbox that works exactly like a native shell.
coderpad=#
```

PAIR PROGRAMMING

- Last 15 min of 45 min phone interview
- Test ability to write advanced SQL queries using CoderPad.
- Bonus points if SQL queries are optimized

TAKE-HOME PROJECT

- Provide toy dataset of company data
- Candidates have a week to send back a write-up
- Findings should include business conclusions
- Can include supporting visualizations and models
- Assess their code, creativity, statistical soundness

ON-SITE INTERVIEW

- Take-Home Project Presentation
  - Present findings to cross-functional panel
- In-Person Whiteboarding Challenge
  - Design a database schema
- 1-1 interviews with the rest of the team

LESSON FIVE_

Don’t limit your hiring pool by focusing on previous job titles. Screen applicants based on their skills.

FOCUS ON SKILLS, NOT JOB TITLES

Look beyond just the data scientist title. Many industries have been slow to adopt the data scientist title. Data scientists can also be called analysts, engineers, quants, specialists -- even data janitors, apparently! It's a relatively new field and titles are not standardized.

FOCUS ON SKILLS, NOT JOB TITLES

Job Seekers:

- **When reviewing job postings, don’t focus on the job title.**
- **Focus on the responsibilities listed in the job description.**
- **Search keywords for specific skills: SQL, R, Python, clustering, AB testing, regression, NLP.**

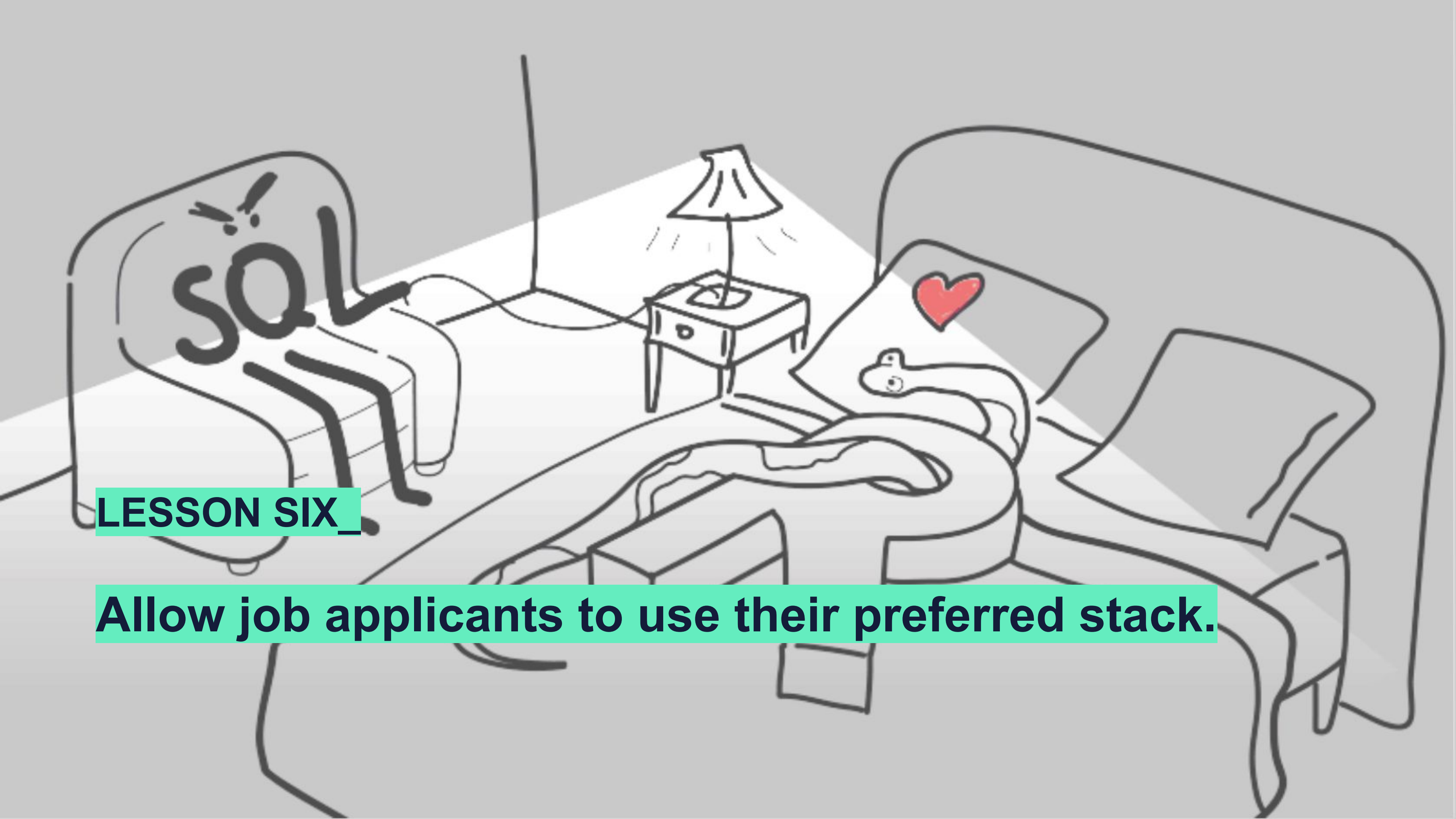
FOCUS ON SKILLS, NOT JOB TITLES

Job Seekers:

- When reviewing job postings, don’t focus on the job title.
- Focus on the responsibilities listed in the job description.
- Search keywords for specific skills: SQL, R, Python, clustering, AB testing, regression, NLP.

Interviewers:

- DS title hasn’t been around for long, and titles vary by industry.
- Over-focusing on previous job titles will limit your applicant pool.
- Write truthful job descriptions: list specific projects and tools.



LESSON SIX_

Allow job applicants to use their preferred stack.

WHAT LANGUAGE SHOULD I USE?

Job Seekers:

- Use the tools you’re most comfortable with to showcase your skills.
- Now is not the time to experiment with programs or methods you’re not familiar with.

WHAT LANGUAGE SHOULD I USE?

Job Seekers:

- Use the tools you’re most comfortable with to showcase your skills.
- Now is not the time to experiment with programs or methods you’re not familiar with.

Interviewers:

- Be flexible. Let candidates use their preferred tools for their assessment.
- Their preferred stack might even include non-programming tools, and sometimes that’s okay!

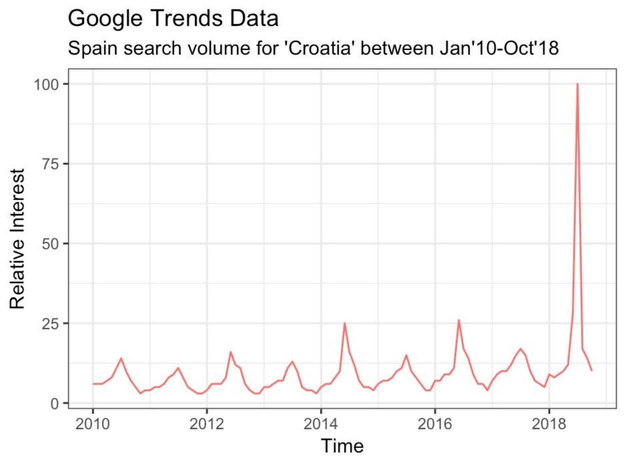
PART THREE

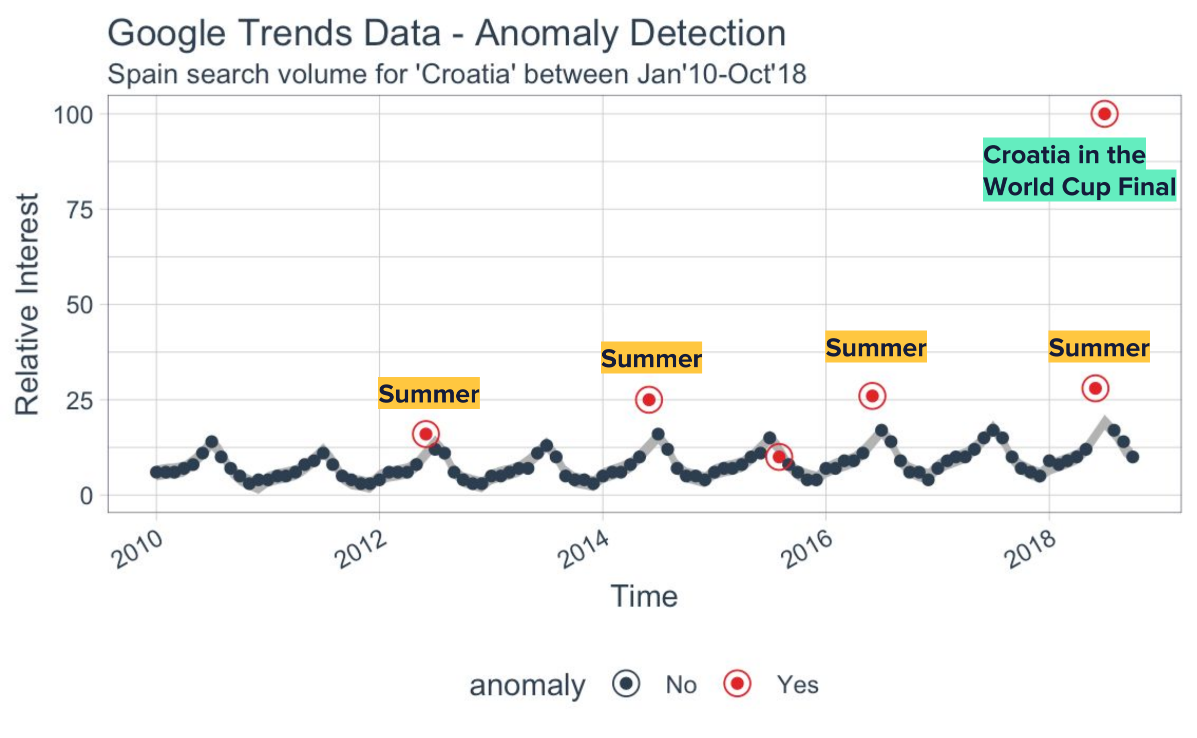
Final Lessons_

LESSON SEVEN_

Data science isn’t always the ‘sexiest job’.

EXPECTATION





Summer

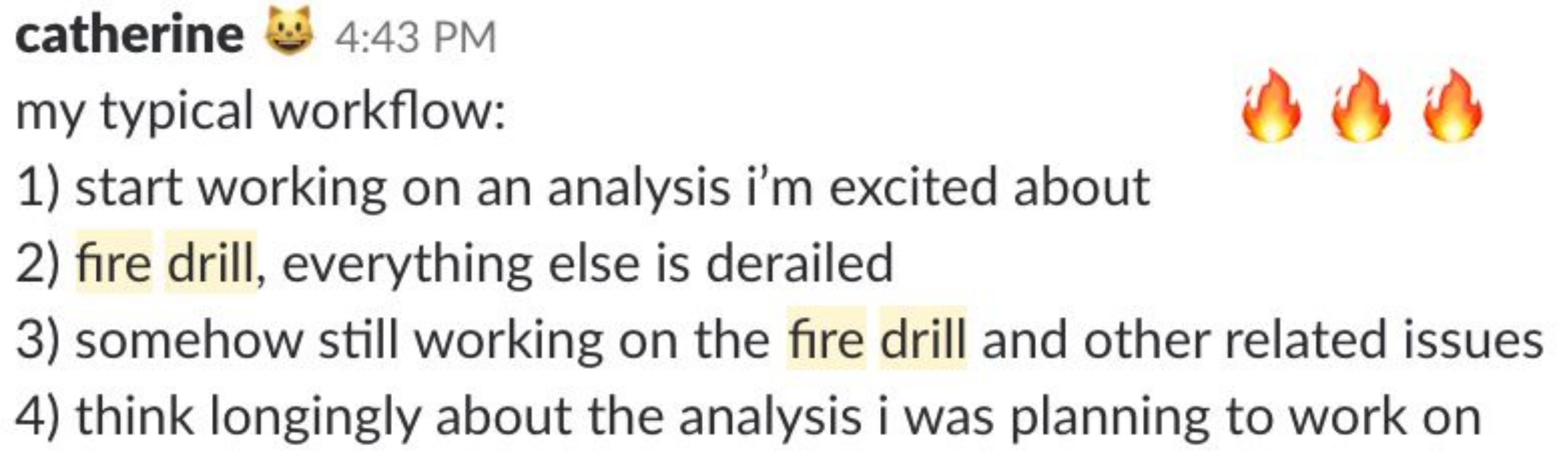
Summer

Summer

Summer

Croatia in the World Cup Final

REALITY



catherine 😺 4:43 PM

my typical workflow: 🔥🔥🔥

1) start working on an analysis i'm excited about

2) fire drill, everything else is derailed

3) somehow still working on the fire drill and other related issues

4) think longingly about the analysis i was planning to work on

LESSON EIGHT_

For job seekers and hiring managers, prioritize finding a good mutual fit.

codecademy

JOB SEEKERS

- Know what type of role you want:
  - Generalist vs Specialist
  - Early vs Late Stage Company
- What skills do you bring to the table?

JOB SEEKERS

- Know what type of role you want:
    - Generalist vs Specialist
    - Early vs Late Stage Company
- What skills do you bring to the table?

HIRING MANAGERS

- Assess candidates based on company needs for the year
- For skills that are “nice to have”, screen for potential
- Ensure that you offer job seekers mutual growth and satisfaction
- Keep your data scientists happy and interested!

Additional Resources

- Codecademy DS Reddit AMA
- Github: Slides For This Talk
- Airbnb: Data Science Tracks
- Doing Data Science At Twitter
- Lyft: Renaming Data Scientists
- Data Analyst vs Scientist

¡HASTA LA PRÓXIMA, MADRID!

twitter @catherinezh

github @cattystats

#rstats

#rladies

codecademy