GETTING PLUGGED INTO DATA SCIENCE

Caitlin Hudon | Data Scientist @ Web.com

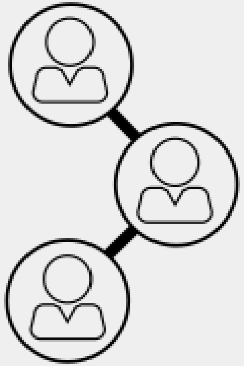
@beeonaposy

MY DEFINITION OF “PLUGGED IN”

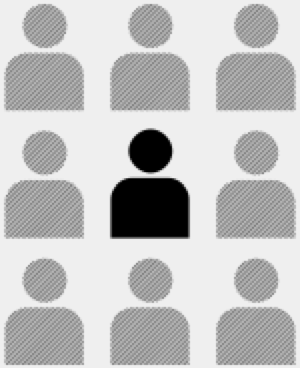
+ Being involved
+ Putting yourself out there
+ Sharing your experiences
+ Caring about the field beyond the work

WHY GET PLUGGED IN?

+ Helps (a lot) with getting jobs
+ Build a network of friends
+ Open up new resources for growth
+ Speaking invitations
+ Build your personal brand



BEING PLUGGED IN HELPS YOU **STAND OUT**



GETTING PLUGGED IN

STAYING PLUGGED IN

1. **Building your network**
2. Developing underrated skills
3. Helping employers find you

4. Beating imposter syndrome
5. Giving back to the community

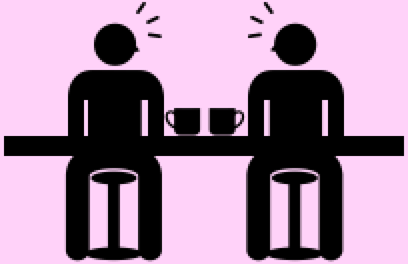
Source: giphy

NETWORKING IS AWKWARD

Source: giphy

THINK OF NETWORKING AS **FINDING YOUR PEOPLE**

AND **CONNECTING** WITH THEM



CONFERENCES

Source

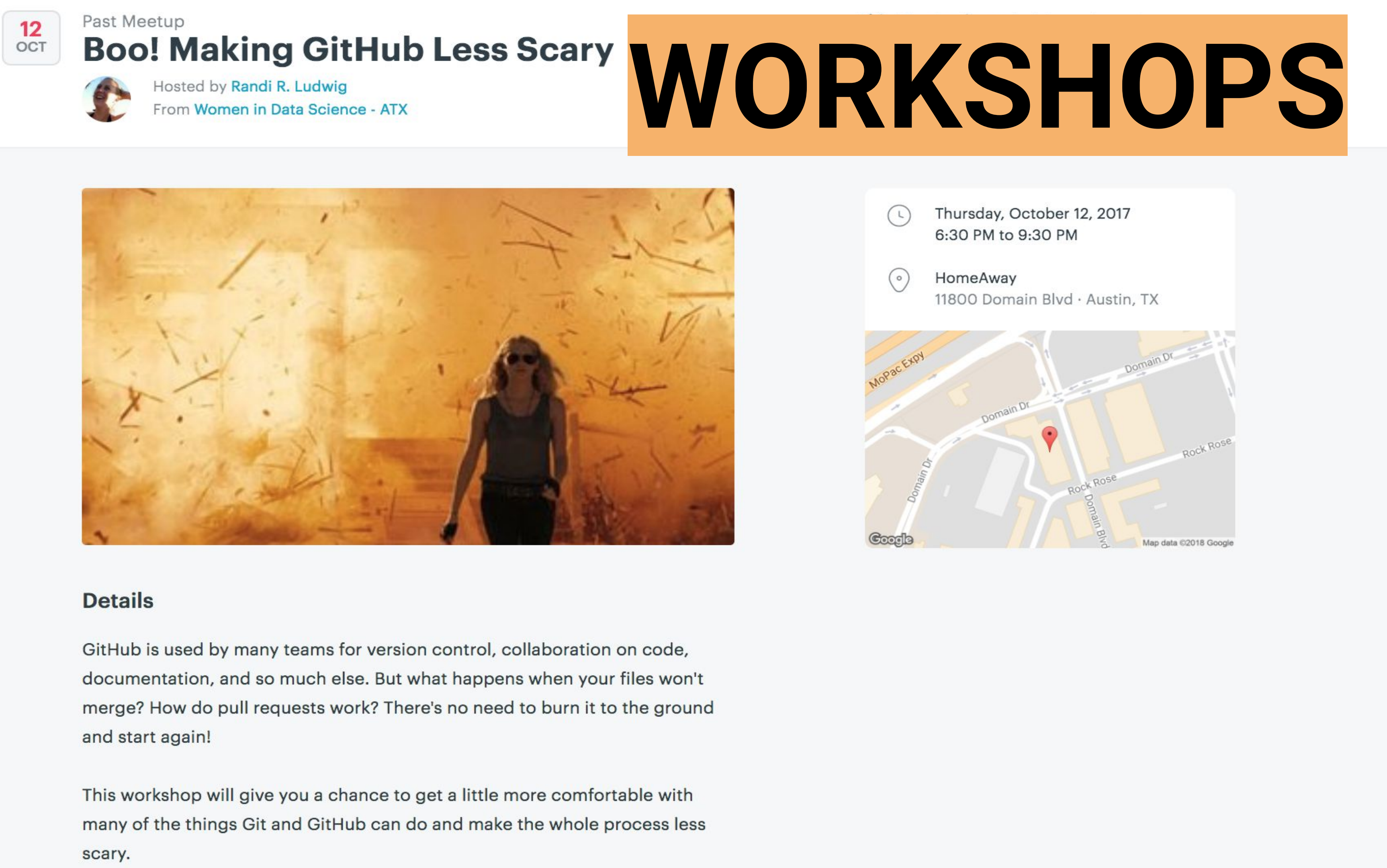
Meetup

JOIN GROUPS

(like R-Ladies!)

HAPPY HOURS

WORKSHOPS



12
OCT

Past Meetup

Boo! Making GitHub Less Scary

Hosted by Randi R. Ludwig
From Women in Data Science - ATX

Thursday, October 12, 2017
6:30 PM to 9:30 PM

HomeAway
11800 Domain Blvd · Austin, TX

Details

GitHub is used by many teams for version control, collaboration on code, documentation, and so much else. But what happens when your files won't merge? How do pull requests work? There's no need to burn it to the ground and start again!

This workshop will give you a chance to get a little more comfortable with many of the things Git and GitHub can do and make the whole process less scary.

TWITTER

#rstats

#pydata

#datascience

Source: giphy

HACKATHONS

Source: giphy

BOOK CLUBS

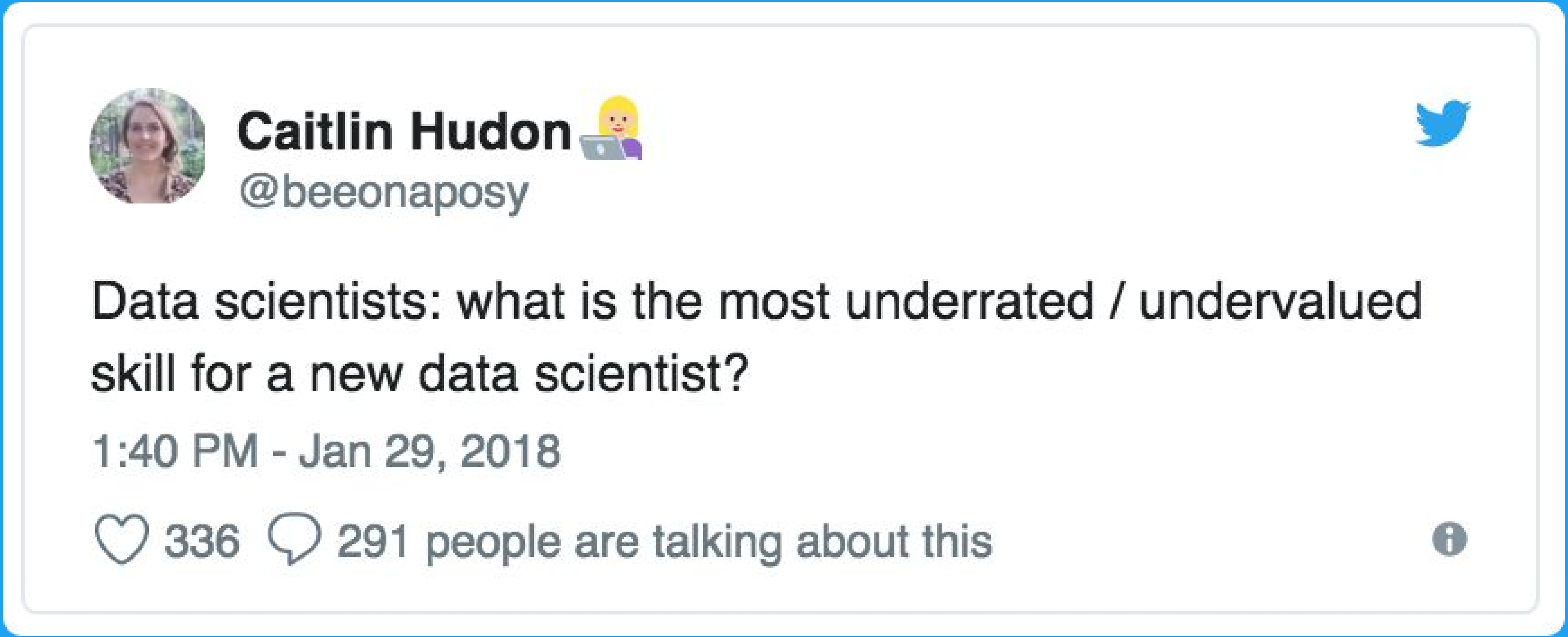
Regression Methods in Biostatistics
SAMPLING
Mathematical Statistics with Applications
James · Witten · Hastie · Tibshirani
An Introduction to Statistical Learning
Abraham Ledolter
Introduction to Regression Modeling
PEARSON
Introduction to Mathematical Statistics

	n	price	ABV
so fresh + so clean	5	$5-5.5	4.5-5.7
hop heads heaven	8	$4.5-6.5	5-9
malty madness	5	$5.25-5.5	4.9-8.3
the dark side	8	$5-7	5-11.1
wheat withdrawal	2	$5.25-5.5	4.8-5.2
other side	5	$5.5-7.5	5.8-9.2
cider city	2	$5.75-6	6.5-7.2
box wine	1	$5	

1. Building your network
2. **Developing underrated skills**
3. Helping employers find you

4. Beating imposter syndrome
5. Giving back to the community

UNDERRATED SKILLS



Caitlin Hudon 👩‍💻
@beeonaposy

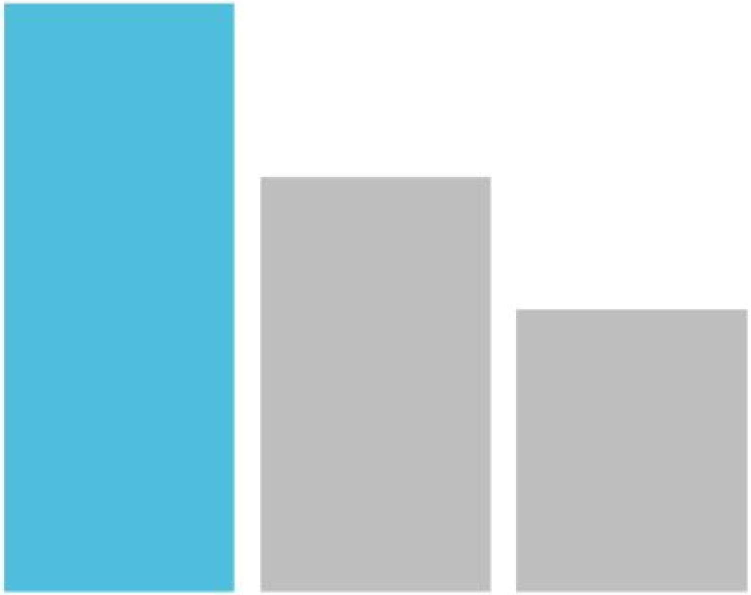
Data scientists: what is the most underrated / undervalued skill for a new data scientist?

1:40 PM - Jan 29, 2018

♡ 336 💬 291 people are talking about this

COMMUNICATION

Click rate for new feature emails to prospects is X% higher than previous feature email and Y% higher than non-feature email click rate.



Relevance: 🟢 Trustworthiness: 🟢 Repeatability: 🟡

Basic Queries

```
-- filter your columns
  SELECT col1, col2, col3, ... FROM table1
-- filter the rows
  WHERE col4 = 1 AND col5 = 2
-- aggregate the data
  GROUP by ...
-- limit aggregated data
  HAVING count(*) > 1
-- order of the results
  ORDER BY col2
```

Useful keywords for **SELECTS**:

DISTINCT - return unique results

BETWEEN a **AND** b - limit the range, the values can be numbers, text, or dates

LIKE - pattern search within the column text

IN (a, b, c) - check if the value is contained among given.

Data Modification

```
-- update specific data with the WHERE clause
  UPDATE table1 SET col1 = 1 WHERE col2 = 2
-- insert values manually
  INSERT INTO table1 (ID, FIRST_NAME, LAST_NAME)
     VALUES (1, 'Rebel', 'Labs');
-- or by using the results of a query
  INSERT INTO table1 (ID, FIRST_NAME, LAST_NAME)
     SELECT id, last_name, first_name FROM table2
```

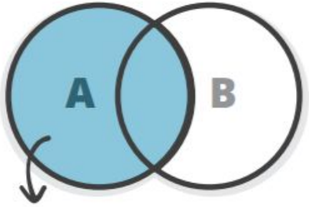
Views

A **VIEW** is a virtual table, which is a result of a query. They can be used to create virtual tables of complex queries.

```
CREATE VIEW view1 AS
SELECT col1, col2
FROM table1
WHERE ...
```

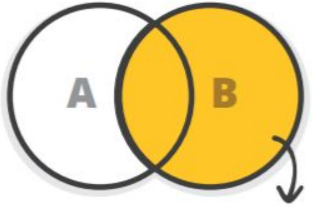
The Joy of JOINs

SQL



***LEFT OUTER JOIN** - all rows from table A, even if they do not exist in table B*

***INNER JOIN** - fetch the results that exist in both tables*



***RIGHT OUTER JOIN** - all rows from table B, even if they do not exist in table A*

Updates on JOINed Queries

You can use **JOIN**s in your **UPDATE**s

```
UPDATE t1 SET a = 1
FROM table1 t1 JOIN table2 t2 ON t1.id = t2.t1_id
WHERE t1.col1 = 0 AND t2.col2 IS NULL;
```

NB! Use database specific syntax, it might be faster!

Semi JOINs

You can use subqueries instead of **JOIN**s:

```
SELECT col1, col2 FROM table1 WHERE id IN
   (SELECT t1_id FROM table2 WHERE date >
      CURRENT_TIMESTAMP)
```

Indexes

If you query by a column, index it!

```
CREATE INDEX index1 ON table1 (col1)
```

Don't forget:

Avoid overlapping indexes

Avoid indexing on too many columns

Indexes can speed up **DELETE** and **UPDATE** operations

Useful Utility Functions

```
-- convert strings to dates:
  TO_DATE (Oracle, PostgreSQL), STR_TO_DATE (MySQL)
-- return the first non-NULL argument:
  COALESCE (col1, col2, "default value")
-- return current time:
  CURRENT_TIMESTAMP
-- compute set operations on two result sets
  SELECT col1, col2 FROM table1
  UNION / EXCEPT / INTERSECT
  SELECT col3, col4 FROM table2;
```

Union - returns data from both queries

Except - rows from the first query that are not present in the second query

Intersect - rows that are returned from both queries

Reporting

Use aggregation functions

COUNT - return the number of rows

SUM - cumulate the values

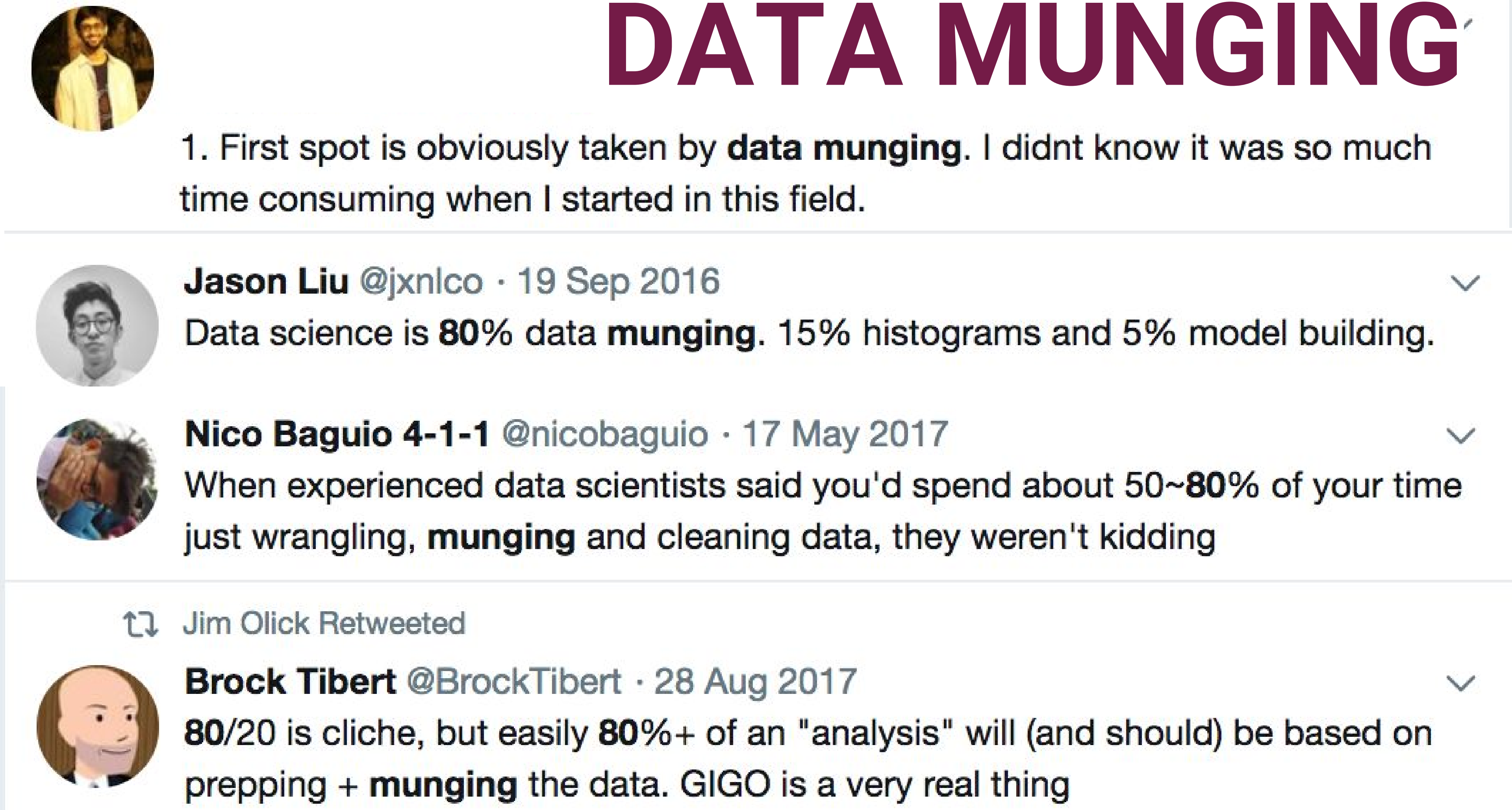
AVG - return the average for the group

MIN / MAX - smallest / largest value

BROUGHT TO YOU BY

XRebel

DATA MUNGING



1. First spot is obviously taken by **data munging**. I didnt know it was so much time consuming when I started in this field.

Jason Liu @jxnlco · 19 Sep 2016

Data science is **80%** data **munging**. 15% histograms and 5% model building.

Nico Baguio 4-1-1 @nicobaguio · 17 May 2017

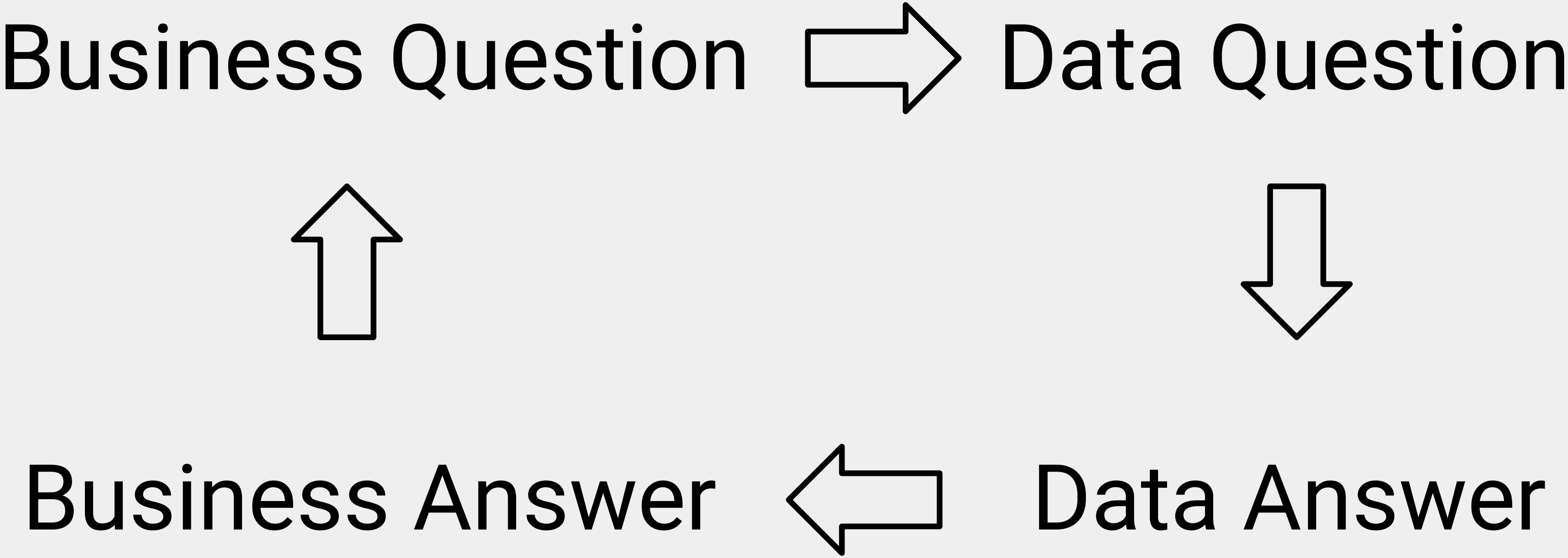
When experienced data scientists said you'd spend about 50~**80%** of your time just wrangling, **munging** and cleaning data, they weren't kidding

Jim Olick Retweeted

Brock Tibert @BrockTibert · 28 Aug 2017

80/20 is cliche, but easily **80%**+ of an "analysis" will (and should) be based on prepping + **munging** the data. GIGO is a very real thing

BUSINESS CONTEXT



Business Question

Data Question

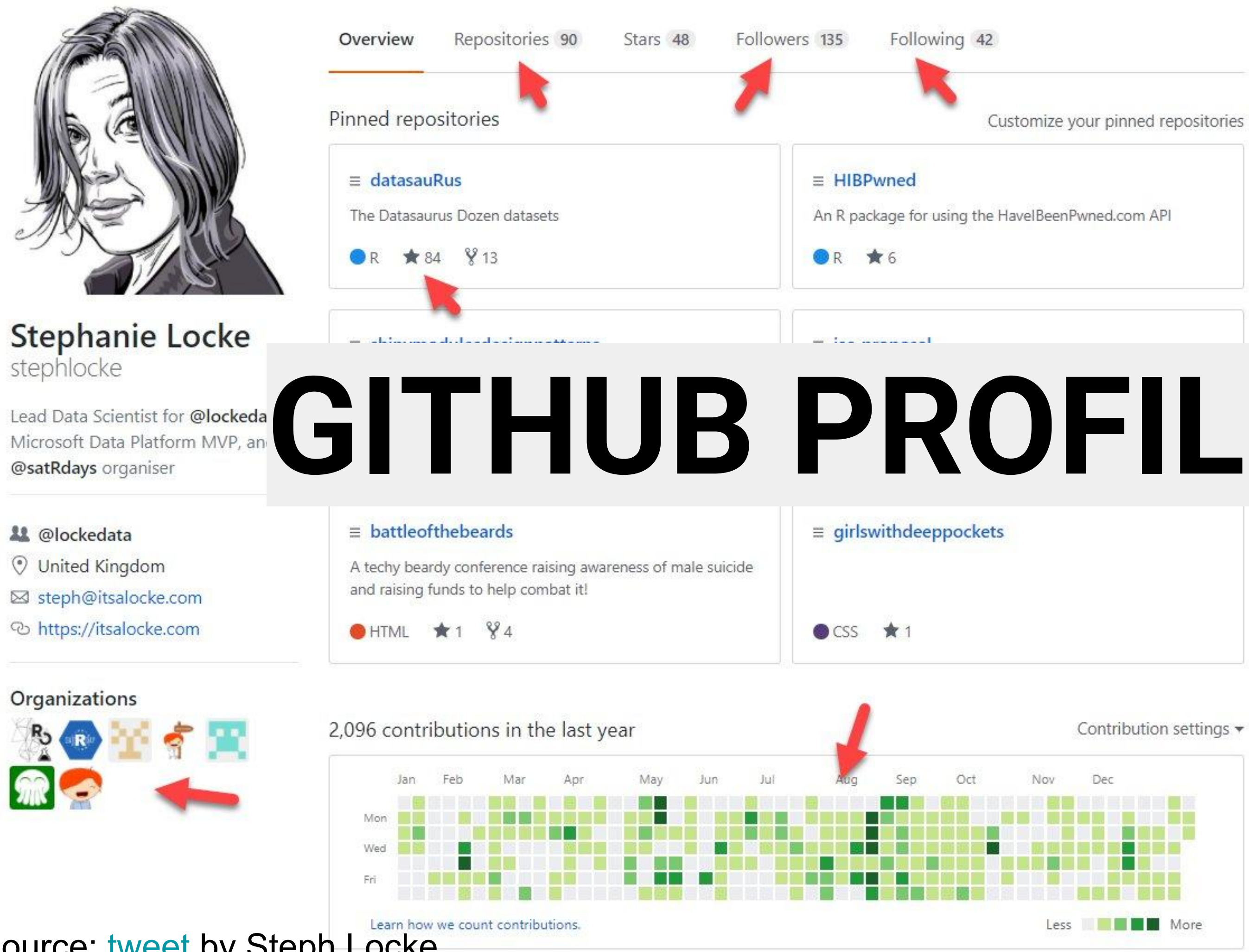
Data Answer

Business Answer

1. Building your network
2. Developing underrated skills
3. **Helping employers find you**

4. Beating imposter syndrome
5. Giving back to the community

GITHUB PROFILE



Overview Repositories 90 Stars 48 Followers 135 Following 42

Stephanie Locke
stephlocke

Lead Data Scientist for @lockeda
Microsoft Data Platform MVP, an
@satRdays organiser

@lockedata
United Kingdom
steph@itsalocke.com
https://itsalocke.com

Organizations

Pinned repositories

Customize your pinned repositories

datasauRus
The Datasaurus Dozen datasets
R ★ 84 13

HIBPwned
An R package for using the HaveIBeenPwned.com API
R ★ 6

battleofthebeards
A techy beardy conference raising awareness of male suicide and raising funds to help combat it!
HTML ★ 1 4

girlswithdeeppockets
CSS ★ 1

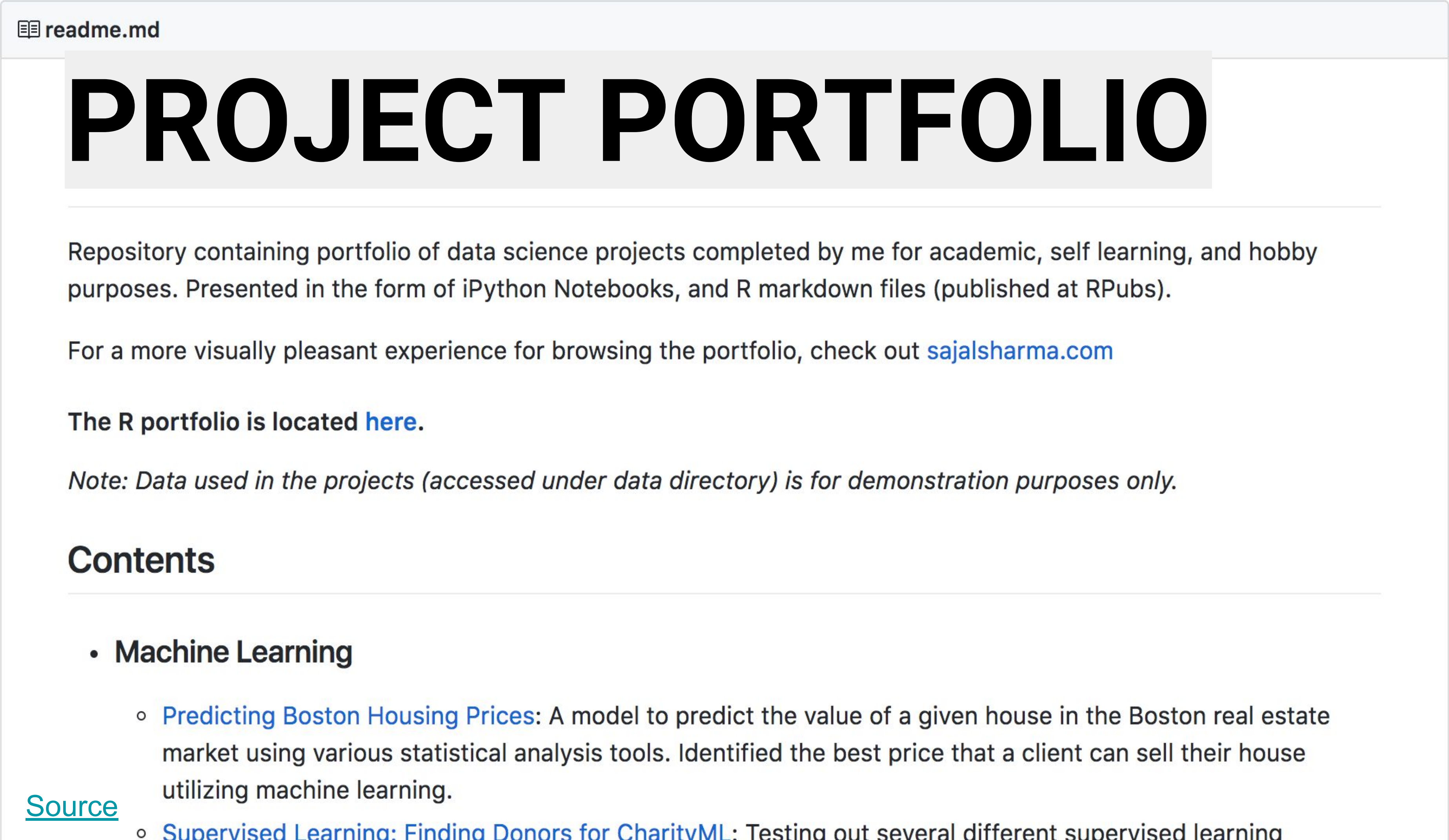
2,096 contributions in the last year

Contribution settings ▾

Learn how we count contributions.

Less 🟩🟩🟩🟩 More

Source: tweet by Steph Locke



readme.md

PROJECT PORTFOLIO

Repository containing portfolio of data science projects completed by me for academic, self learning, and hobby purposes. Presented in the form of iPython Notebooks, and R markdown files (published at RPubs).

For a more visually pleasant experience for browsing the portfolio, check out sajalsharma.com

The R portfolio is located here.

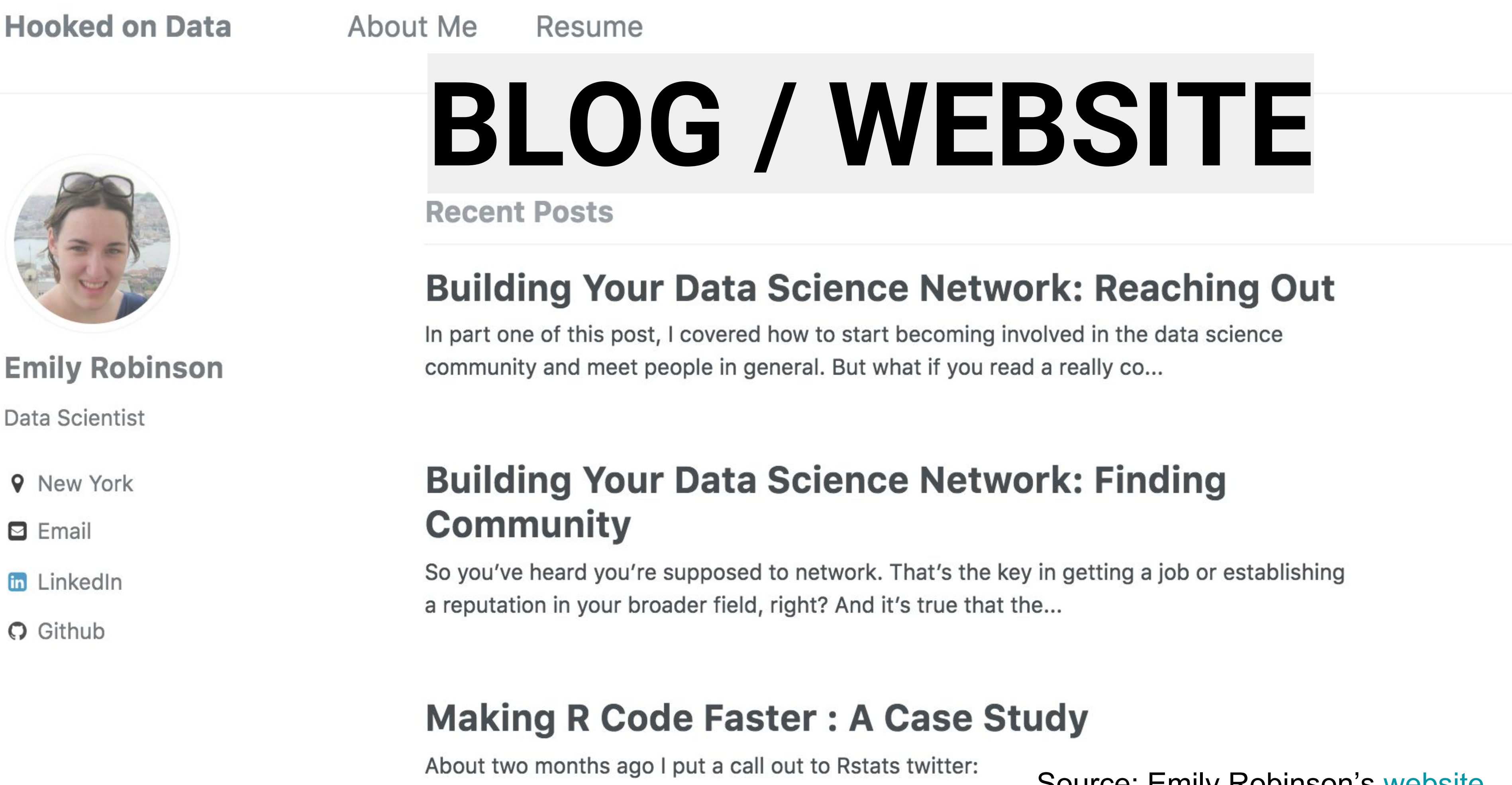
Note: Data used in the projects (accessed under data directory) is for demonstration purposes only.

Contents

- **Machine Learning**
    - Predicting Boston Housing Prices: A model to predict the value of a given house in the Boston real estate market using various statistical analysis tools. Identified the best price that a client can sell their house utilizing machine learning.
    - Supervised Learning: Finding Donors for CharityML: Testing out several different supervised learning

Source

BLOG / WEBSITE



Hooked on Data

About Me

Resume

Emily Robinson

Data Scientist

New York

Email

LinkedIn

Github

Recent Posts

Building Your Data Science Network: Reaching Out

In part one of this post, I covered how to start becoming involved in the data science community and meet people in general. But what if you read a really co...

Building Your Data Science Network: Finding Community

So you've heard you're supposed to network. That's the key in getting a job or establishing a reputation in your broader field, right? And it's true that the...

Making R Code Faster : A Case Study

About two months ago I put a call out to Rstats twitter:

Source: Emily Robinson's website

TWITTER

+ Live-tweet conferences
+ Find cool blog entries
+ Keep up with DS news
+ Build your network
+ Share your projects

LINKEDIN

I’m interviewing for your former position at Company A. I saw that you’re at Company B now. It looks like our backgrounds are similar (and we have similar interests -- I applied at Company B too).

I’d love to hear any advice you have about working in Austin’s data scene.

I’m glad you reached out to me. What job did you apply for at Company B? We have a Sr. Marketing Data Analyst job on my team that seems like a good fit for you. I’ll do my best to get you an interview for it.

1. Building your network
2. Developing underrated skills
3. Helping employers find you

4. **Beating imposter syndrome**
5. Giving back to the community

IMPOSTER SYNDROME

“A false and sometimes crippling belief that one's successes are the product of luck or fraud rather than skill”

DATA SCIENCE IS
A NEW FIELD

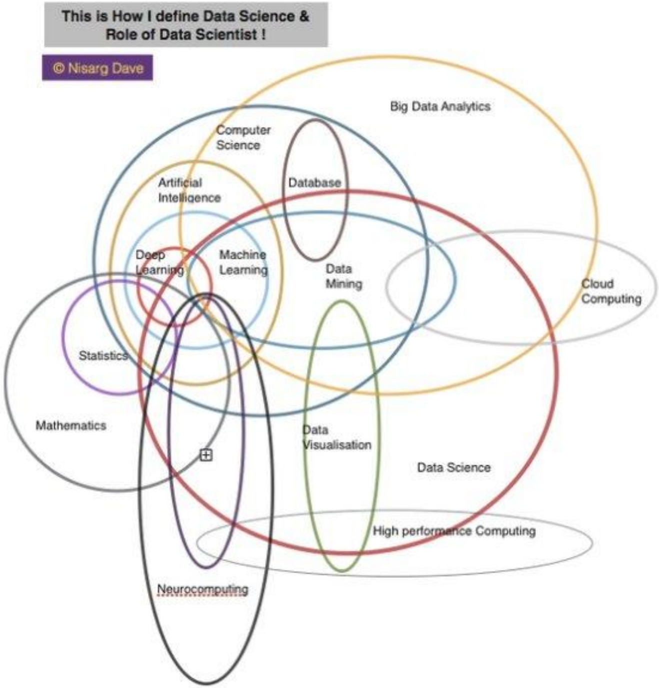
~~“Business analyst”~~

~~“Data analyst”~~

~~“Research scientist”~~

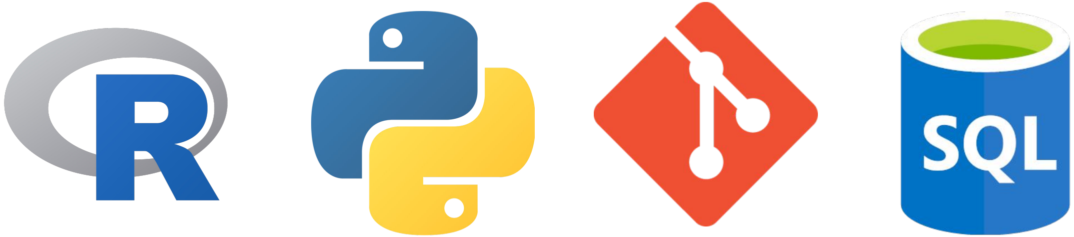
DATA SCIENCE IS
A COMBINATION OF OTHER FIELDS

- Data analysis
- Statistics
- Software engineering
- Machine learning
- Visualization
- Database administration
- Business acumen

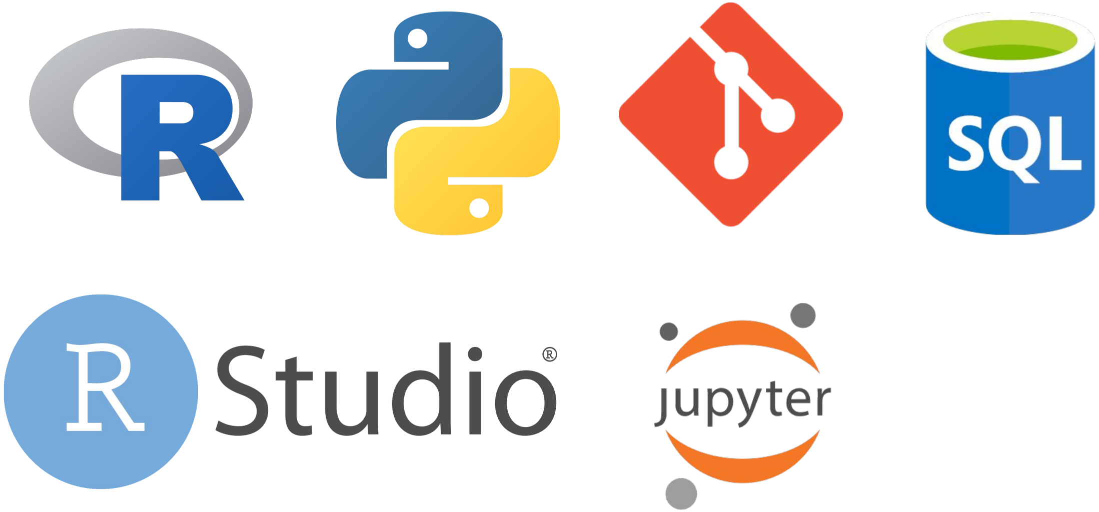


Source: Machine Learning 101 deck

DATA SCIENCE IS
CONSTANTLY EXPANDING



DATA SCIENCE IS **CONSTANTLY EXPANDING**



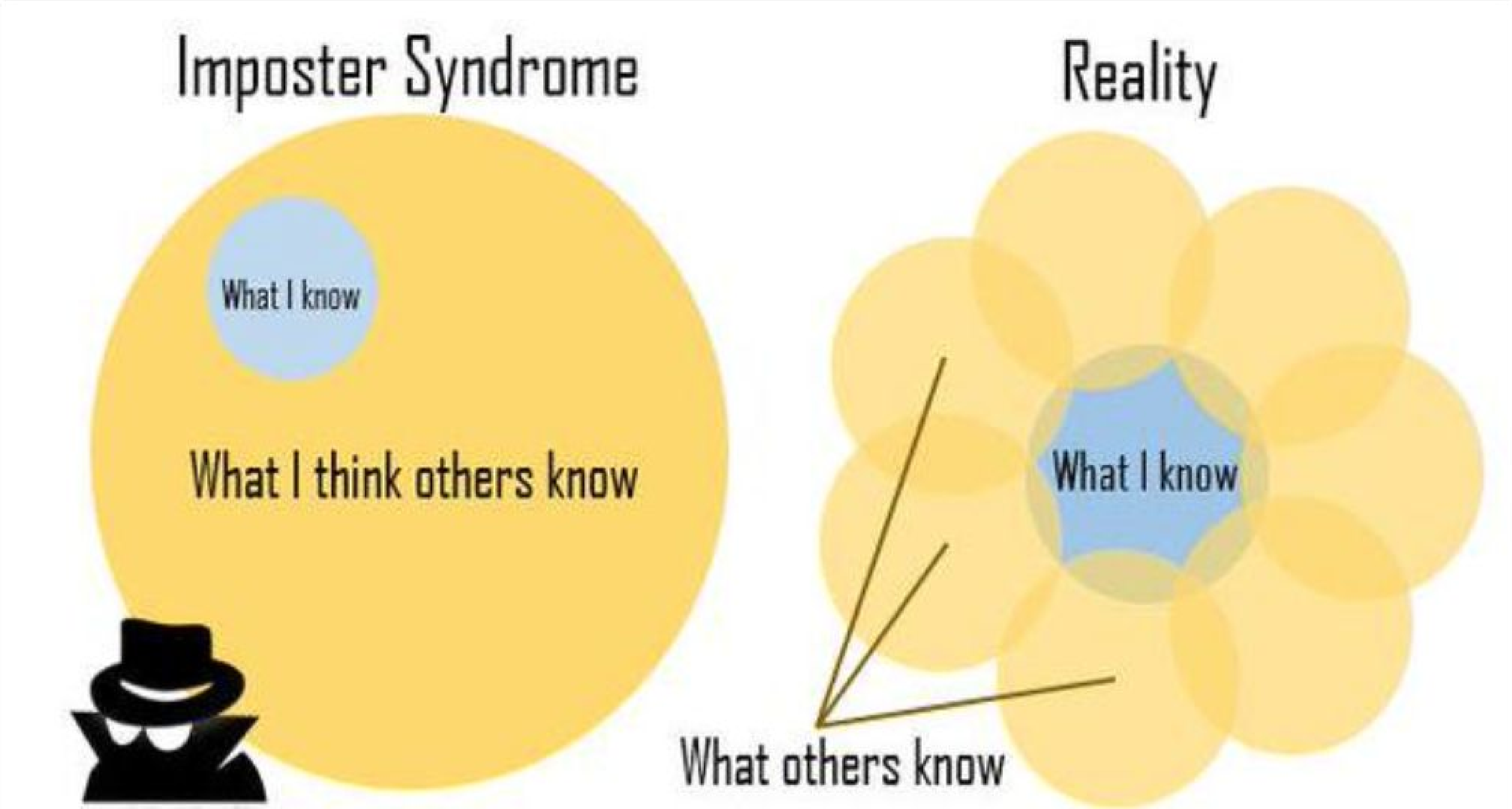
DATA SCIENCE IS **CONSTANTLY EXPANDING**

R
SQL
tidyverse
www.rstudio.com
R Studio®
jupyter
scikit
learn

DATA SCIENCE IS **CONSTANTLY EXPANDING**

DATA SCIENCE IS **CONSTANTLY EXPANDING**

R
SQL
tidyverse
www.rstudio.com
mongoDB
R Studio®
jupyter
scikit learn
Spark
hadoop



Source: Hugh Kearns’ tweet

MY APPROACH

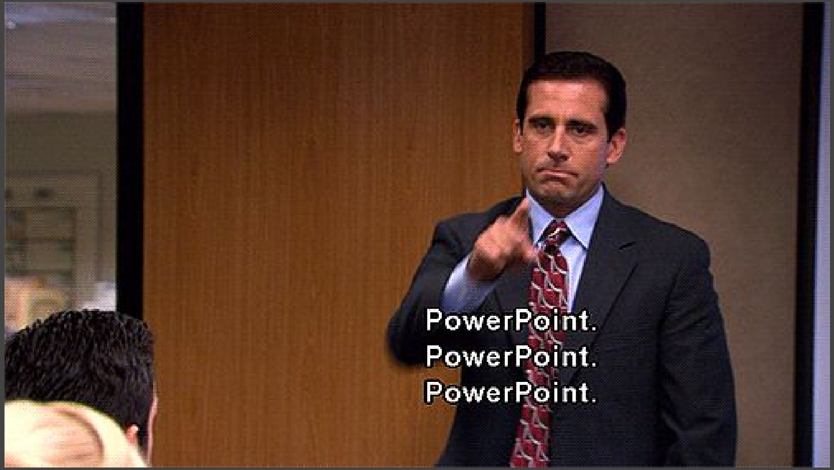
I will never be able to learn everything there is to know in data science — I will never know every algorithm, every technology, every cool package, or even every language — and that’s okay.

1. Building your network
2. Developing underrated skills
3. Helping employers find you

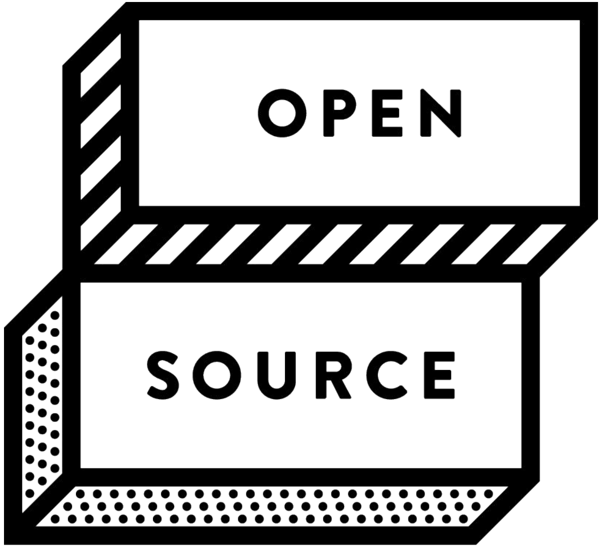
4. Beating imposter syndrome
5. **Giving back to the community**

GIVE A TALK

- Start local
- Lightning talks
- Beginner-level is great!



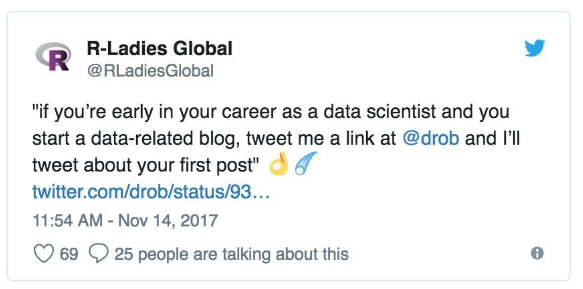
THINK OPEN SOURCE



- Help the next person
- Contribute to OS projects
- Share your work

KEEP BLOGGING

- + Audience: you, 2 weeks ago
- + Tracking your learning
  → #DSlearnings
- + Share what’s helped you

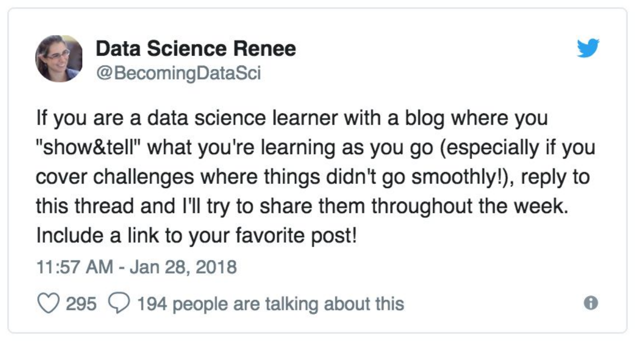


R-Ladies Global
@RLadiesGlobal

"if you’re early in your career as a data scientist and you start a data-related blog, tweet me a link at @drob and I’ll tweet about your first post" 👌☄️
twitter.com/drob/status/93…

11:54 AM - Nov 14, 2017

♡ 69 💬 25 people are talking about this



Data Science Renee
@BecomingDataSci

If you are a data science learner with a blog where you "show&tell" what you're learning as you go (especially if you cover challenges where things didn't go smoothly!), reply to this thread and I'll try to share them throughout the week. Include a link to your favorite post!

11:57 AM - Jan 28, 2018

♡ 295 💬 194 people are talking about this

RESOURCES: GETTING PLUGGED IN

- Building your data science network (finding community and reaching out), [Emily Robinson]
- Questions to ask in interviews [Julia Evans]
- Making peace with personal branding [Rachel Thomas]
- How to build your personal brand as a new web developer [Rick West]

RESOURCES: STAYING PLUGGED IN

- Learning at work [Julia Evans]
- Contributing to open source [Julia Evans]
- Advice to aspiring data scientists: start a blog [Dave Robinson]

THANK YOU!

Caitlin Hudon

@beeonaposy

caitlinhudon.com