

# DSC10 Final Lecture

Colin Jemmott, March 2019

# Part I

DSC 10 and Data Science

# So now what?

aka: What the hell? Coursera said I can be a data scientist in one week for \$29.95. Some dude donated millions of dollars and I worked my ass off for ten weeks. So why don't I know data science yet?

# What this class covered

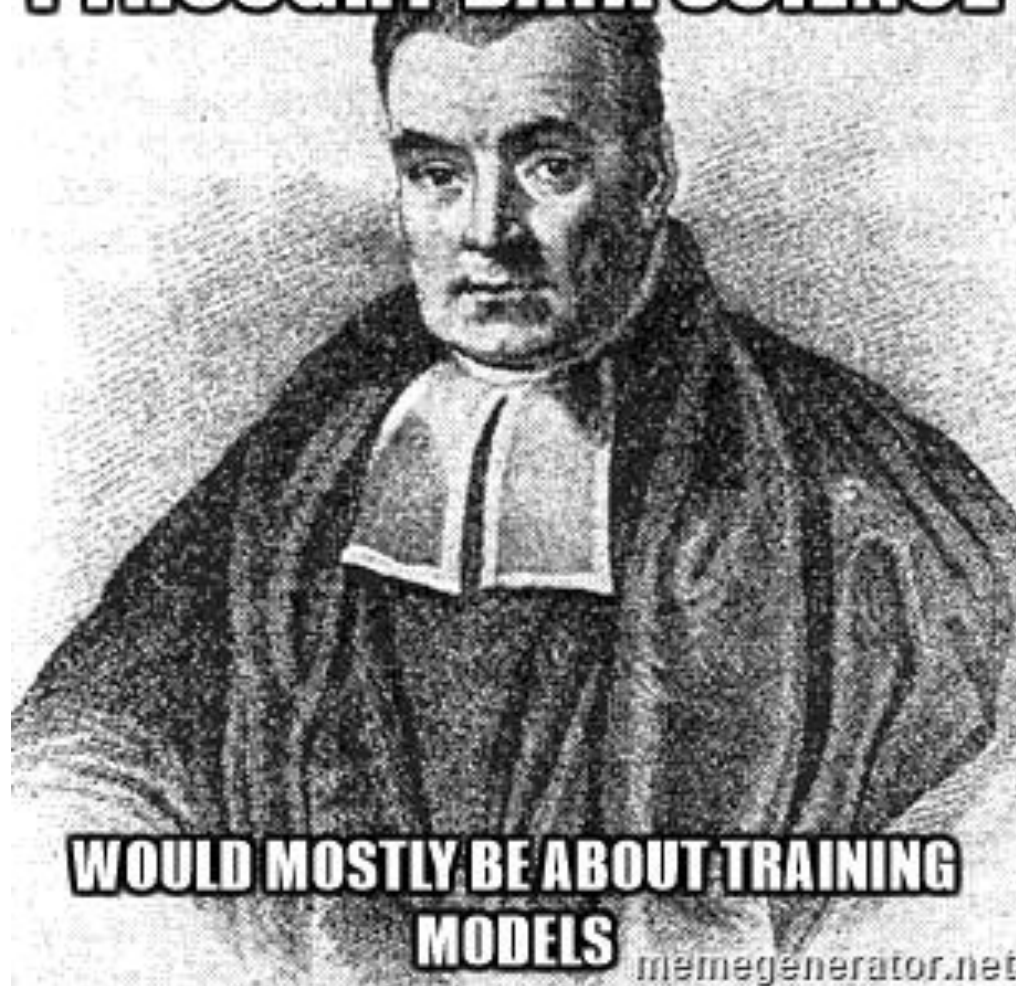
DSC10 is roughly equal thirds:

1. Python
2. Data
3. Inference

# What this class did *not* cover

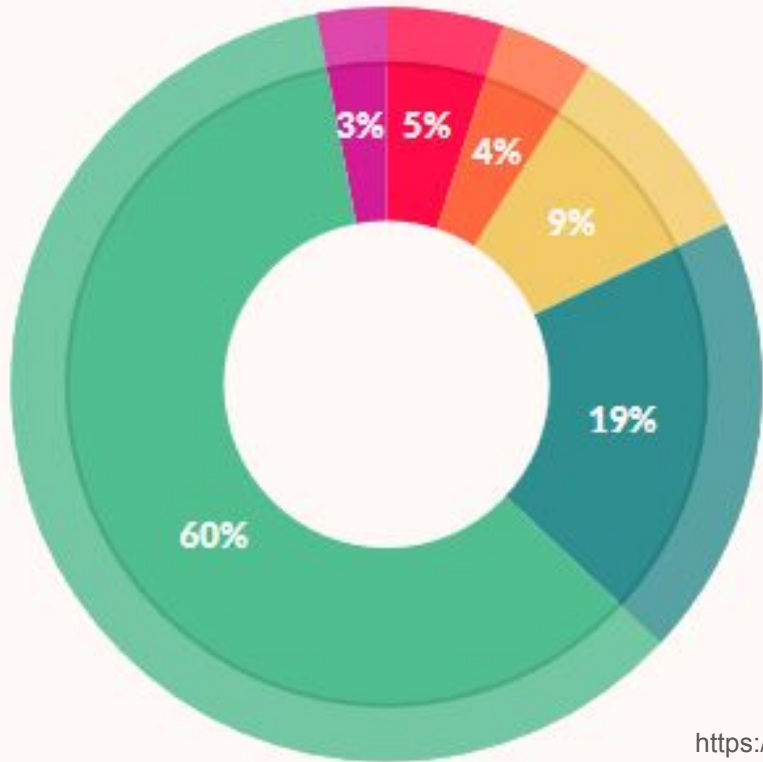
1. How to do data science

**I THOUGHT DATA SCIENCE**



**WOULD MOSTLY BE ABOUT TRAINING  
MODELS**

# What data scientists do all day



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

# The Data Science Process

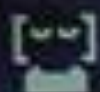
1. Identify the question
  - What is the goal? What is the scope?
2. Prepare the data
  - Find, source, clean. Check completeness, fix anomalies, perform QA
3. Analyze the data
  - Build models, data mine, run text analysis, etc.
4. Visualize the data
  - Complex results -> easy-to-digest visuals
5. Present your findings
  - Document, UI/UX, productize.



# Ethics

The Five Cs:

1. Consent
2. Clarity
3. Consistency and Trust
4. Control and Transparency
5. Consequences



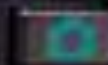
Did someone blink?



OK Exit

Nikon

SCENE

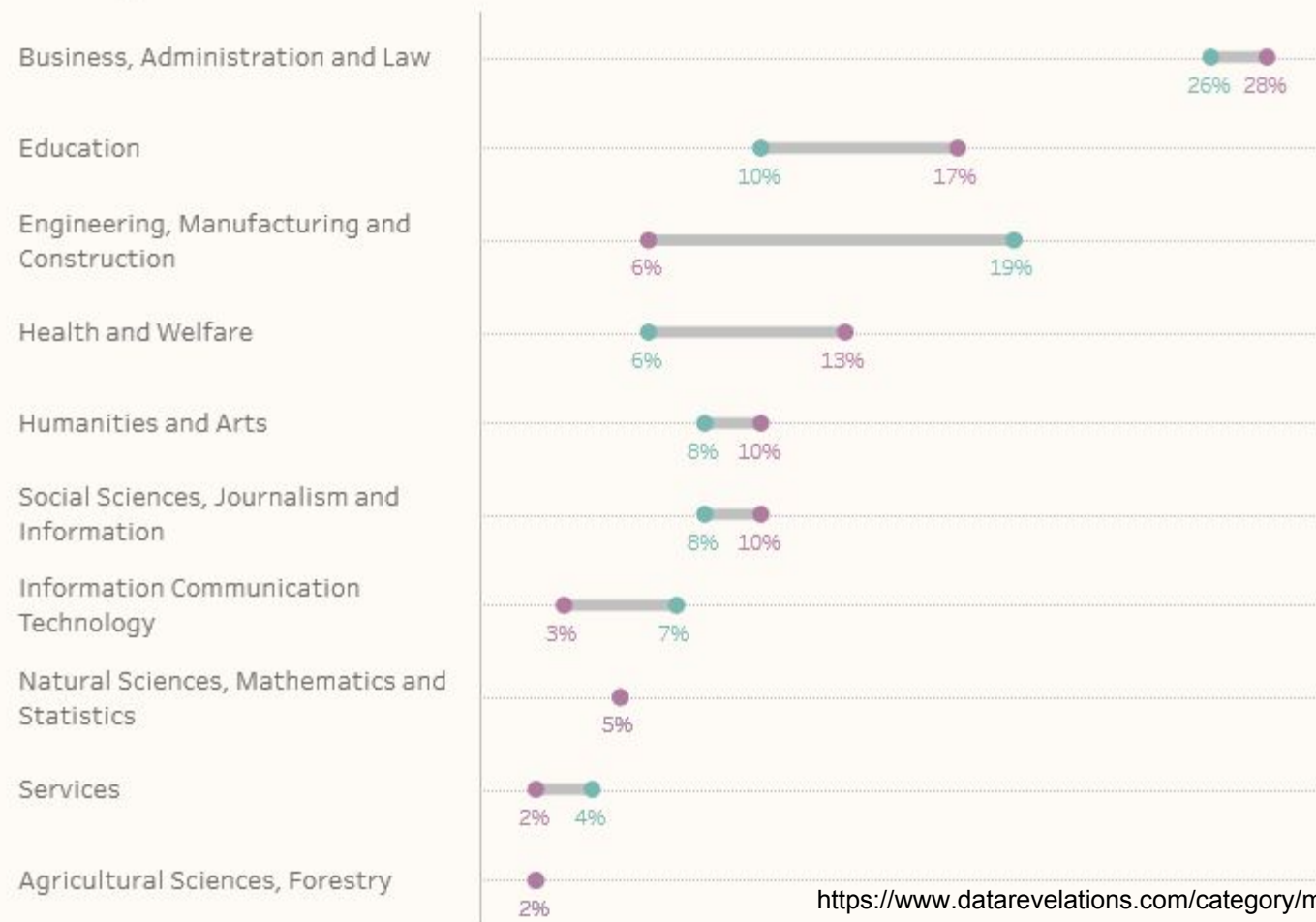


# Visualization



# Fields of Study

Women | Men





<https://haveibeenpwned.com/>

<https://www.google.com/maps/timeline>

# Part II

My Principles of Data Science

# My Principles of Data Science

1. Honesty
2. Curiosity
3. Impact
4. Humility

# Honesty with Collaborators

- Robustly and openly review other's work
- Have your work reviewed by them
- Exercise independent judgement
- Feedback should be carefully considered



# Honesty with Yourself

We all have limits and make mistakes. Understanding them is important.

There is no shame in acknowledging ignorance, but shipping algorithms or code you don't understand can be disastrous.

Remember to take ownership of both your failures and successes.

# Honesty With Users

- If a user is mislead, they may make important decisions based on falsehoods
- To prevent this:
  - Never ship bad results
  - Acknowledge and quickly fix mistakes
  - Check with users to see if they actually understand

# Curiosity About Data

Good analysis depends on understanding your data.

- Not just what it looks like, but how it reflects what happened in the real world.

When something seems off, have the tenacity to really dig in.

Good understanding of data beats good algorithms every time.

# Curiosity: Using Science

Controlled experiments are how we know what works and how to do better.

- Form a hypothesis
- Design a test
- Measure results

New data should be able to change your mind, even about strongly held beliefs.

# Curiosity: Self-Improvement

Your technical skills are the main thing holding you back.

You have:

- Data
- Resources
- Support

Your ability to shop good code is the limiting factor.

Continuous learning is the way to overcome that.

# Impact: How to Tell

*If people don't change their behavior,  
you didn't make an impact.*

# Impact: Get Unstuck

If you find yourself working really hard to make marginal improvements:

- Step back
- Embrace creativity
- Ask you are solving the right problem
- Seek radically different approaches
- Ask for help

# Impact: Follow

A secret of data science:

we generally make novel applications of existing methods  
instead of inventing totally new methods.



# Humility: Being Wrong is OK

Despite our best efforts, most of us are wrong most of the time.

- Find ways to validate your work. Or, more importantly, invalidate it.
- Look for edge cases
- Don't be fooled by randomness
- "Analysis is free of errors" isn't the same as "the results of the analysis are true"

# Humility: Easy > Hard

Science is hard.

If you start by solving easy problems with simple methods:

- You can iterate much more quickly
- You are less likely to make mistakes

Complex solutions drive up research cost, but also development, architecture, compute and maintenance

# Humility: Dead Ends

*Most of your research will fail.*

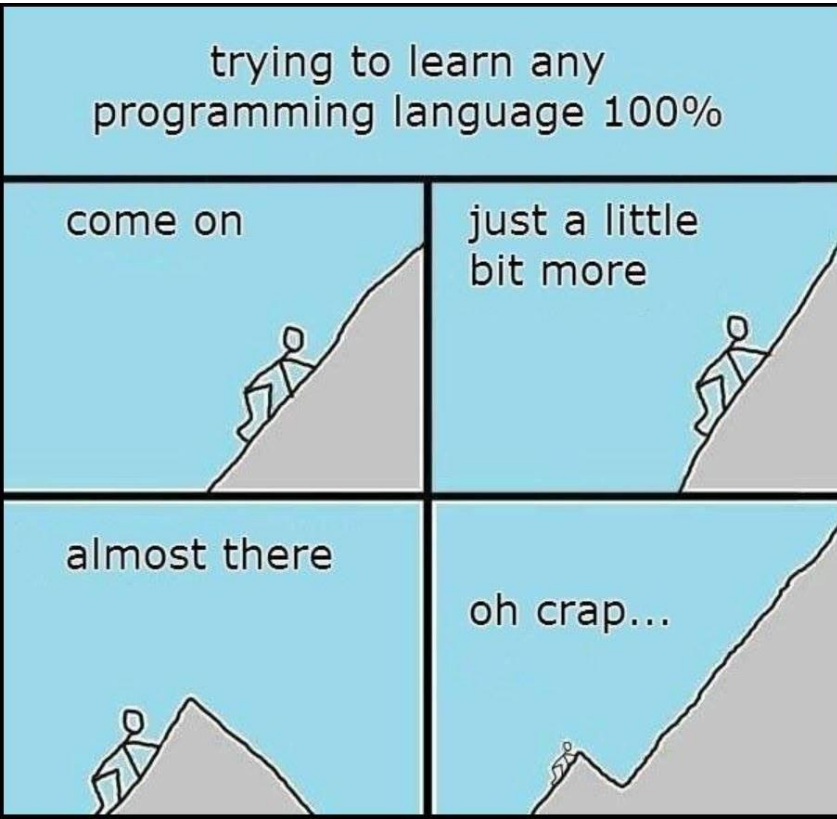
- Identify dead ends quickly
- Accept it with grace
- Document it
- Try again

# Part III

The Future

# The near future: DSC20

## Foundations of Programming and Data Structures



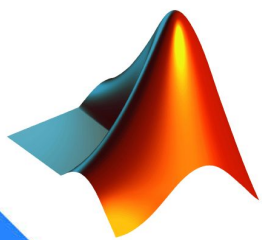
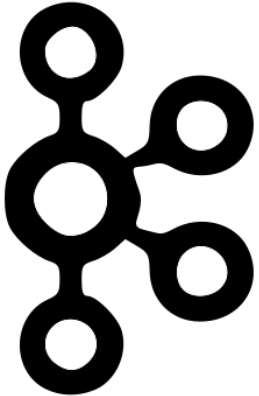
Python programming tutorials, resources and exercises:

- Software Carpentry
- Code Academy
- CodingBat

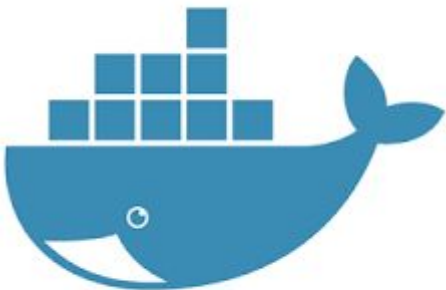
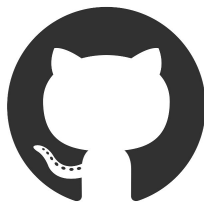
# The Future After UCSD

Get a job!

1. Skills
2. Projects
3. Portfolio
4. Interview



Skills



ArcGIS



# ~~Skills~~ Concepts

- Querying: SQL
- Web Visualization: d3.js
- Cloud: AWS
- Exploration: Tableau
- Collaborate: github
- Machine Learning: sklearn
- Anything weird: google for the python package

Recommendations are Colin's perception of the best / easiest / most fun tool in 2019, but these things change fast, and he is wrong almost all the time.

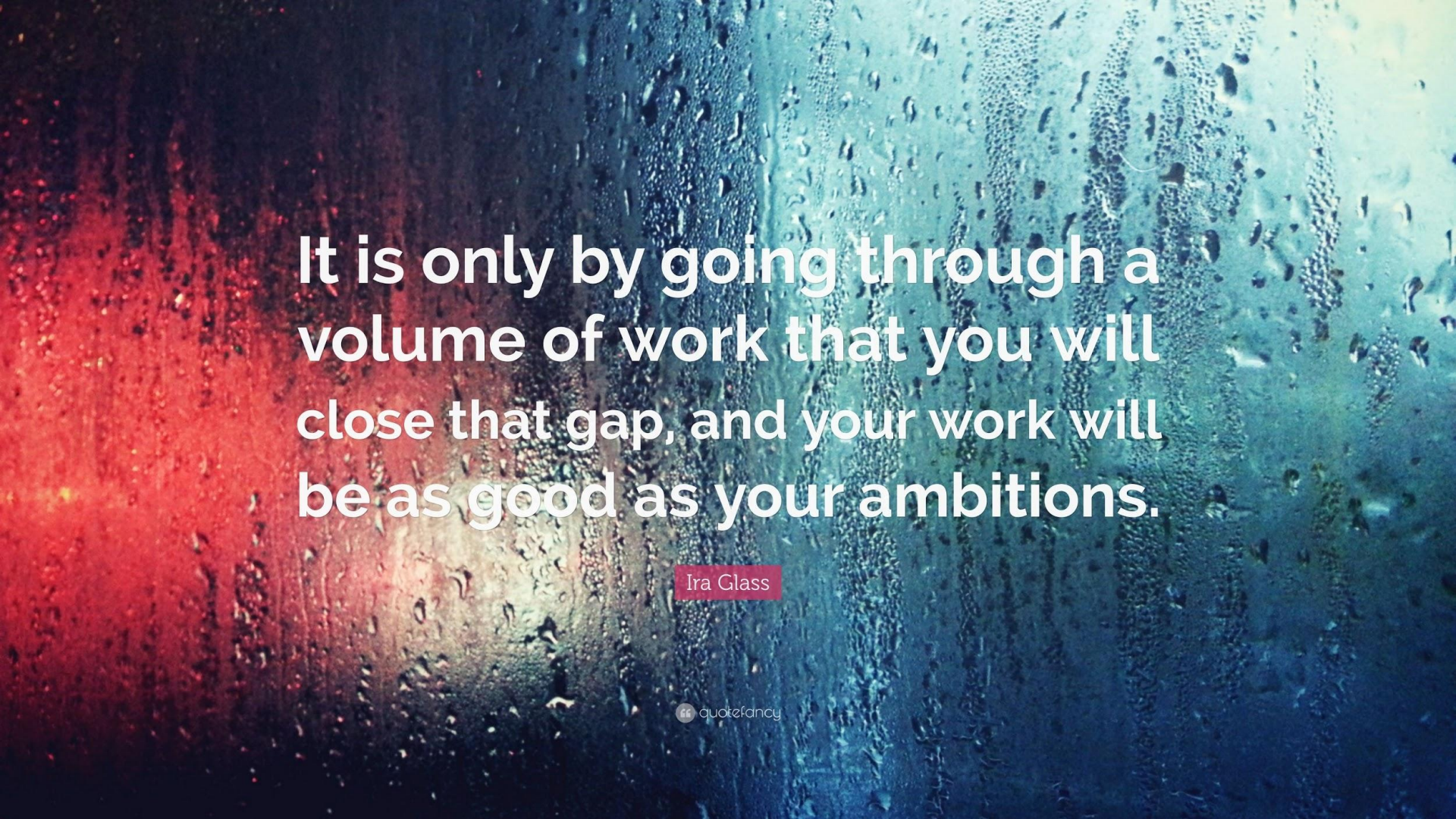


# How to get started

1. Don't wait to get started.
2. Steal
3. Build what you want to see in the world

## The Secret:

Do good work and share it with people



It is only by going through a  
volume of work that you will  
close that gap, and your work will  
be as good as your ambitions.

Ira Glass

# You need a portfolio

A resume isn't enough. Show off your projects.

There are a ton of options, but some I use:

- Github
- Professional social (Twitter / Medium / LinkedIn)
- Professional personal site

# Scott Cole

personal webpage

Home

Blog

Burritos of San Diego

CV

Data projects

I'm a PhD student studying [neuroscience](#) at [UC San Diego](#) and working in the [Voytek lab](#) studying brain rhythms. Specifically, I am working on a new analysis framework to extract more information from neural signals by parametrizing the waveform shapes of the brain rhythms. You can learn more about this in our recent [review](#) and [methods paper](#).

Otherwise, my main passion is in uncovering trends in data. Though I hope I'll be able to apply these skills to improving public good in the future, you can [see a list of my projects here](#) ranging from burritos to text mining to police misconduct ([CV](#), [resume](#)).

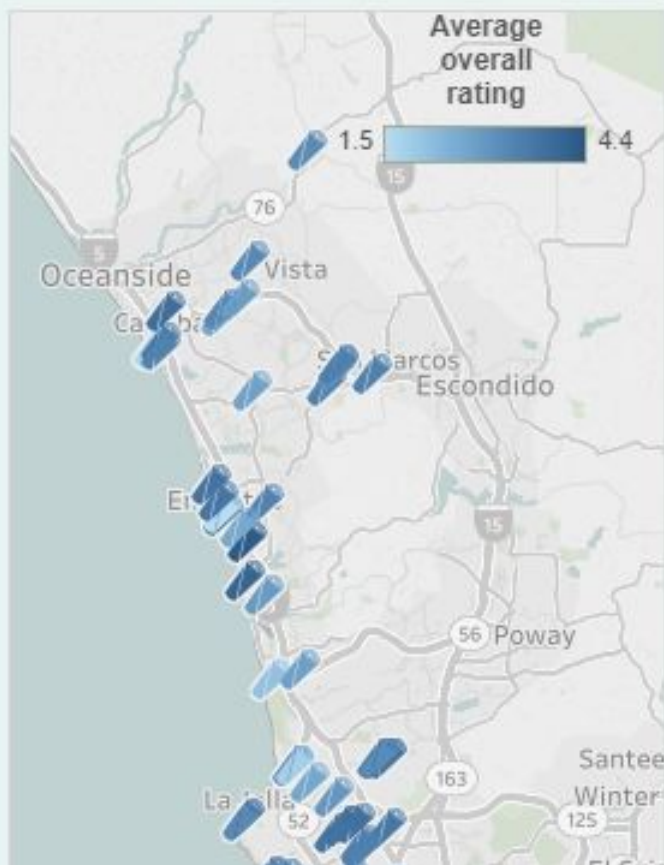




# San Diego Burrito Ratings



350 burritos were rated at 77 taco shops across San Diego using the 10 official burrito dimensions (for more information, go to <https://srcole.github.io/100burritos>). Use the sliders on the right to filter the map and locate burritos based on your standards.



Click a burrito on the map to see the details of that restaurant below.

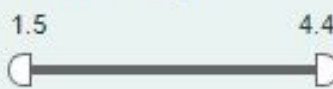
## Restaurant information

## Neighborhood

(All)

This visualization was based on a template by Will Griffiths, @wilsooon

## Overall rating



## Tortilla quality



## Yelp rating



## Temperature



## Cost



## Meat quality



## # burritos rated



## Nonmeat quality



## Volume (liters)



## Meat-to-filling ratio



# Stretch Goal

Contribute to an open source project.

*“If I was interviewing someone and found out they had a pull request accepted for code we use, I would hire them on the spot.”*

- Ben Taylor, Chief AI Officer at ZIFF

# Prepare for Interviews

## General

- Be ready for coding interviews
  - Cracking the Coding Interview
- Read about how to interview
  - But tech companies can be, ummm, *different*

## Company specific

- Do research about the company
- Do research about how they interview

85,526 views | Mar 1, 2019, 07:45am

# Radical Change Is Coming To Data Science Jobs



**Nate Oostendorp** Forbes Councils

**Forbes Technology Council** CommunityVoice ⓘ



# Radical Change Is Coming To Data Science Jobs

A calculator was once a person. Webmaster was once a hot career.

[A]dvancements in hardware and software took specialized skills and put them into the hands of generalists. While specialist jobs were lost, the democratization of these technologies unleashed waves of innovation, commerce and job creation.

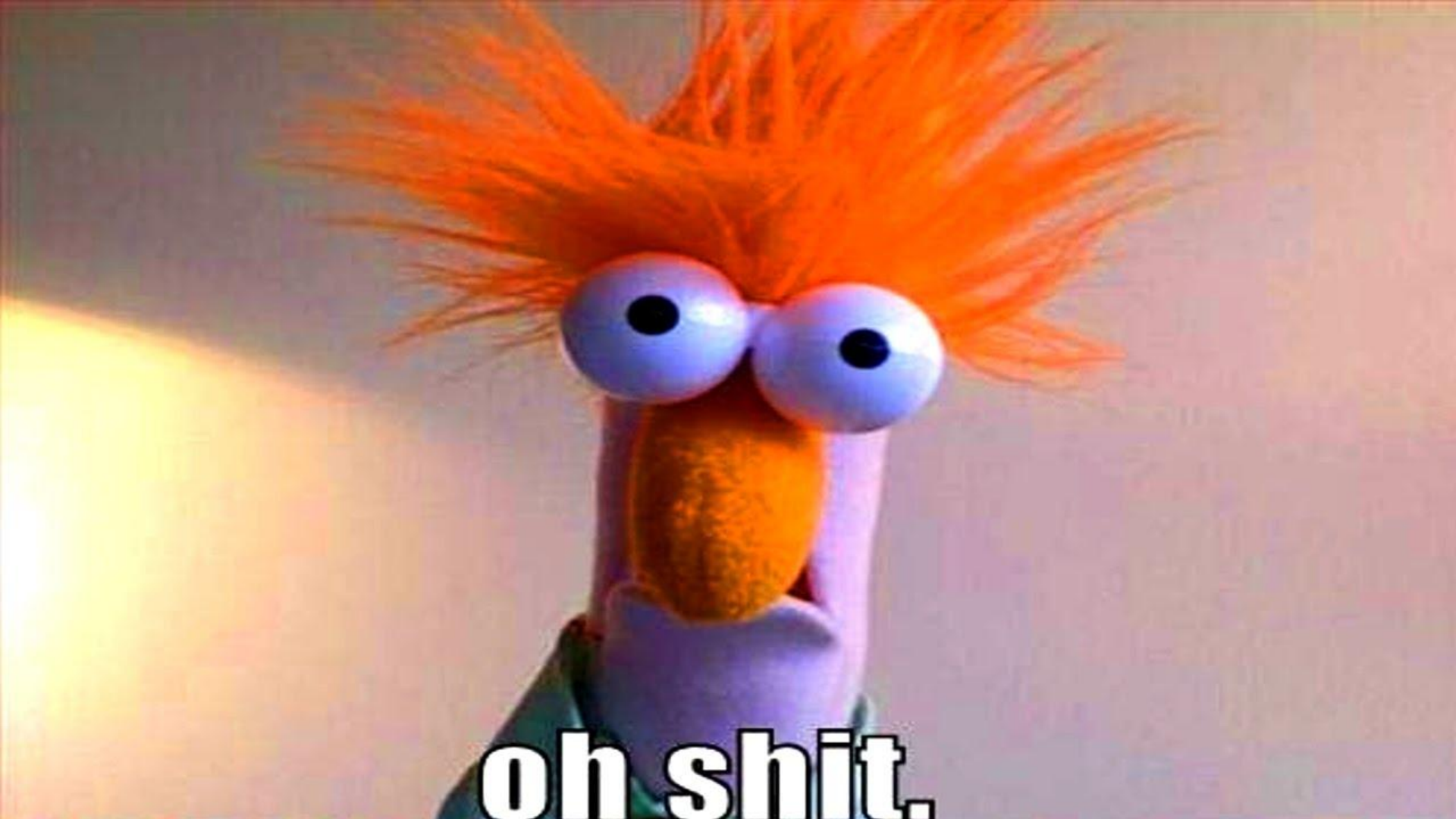
Similarly, I believe the job of data scientist as we know it today will be barely recognizable in five to 10 years....While their data science skills will be a strong career asset, a surprisingly small proportion of them will likely to be working as straight data scientists.

# Radical Change Is Coming To Data Science Jobs

When I studied computer science back in the way-back-when, compiler design was a required course. ... It was common to write pieces of commercial applications in machine language for faster performance.

Over the past few decades, successive layers of software functions have been abstracted into higher-level development tools.

Data science is quickly following the same progression.



**oh shit.**

# The Good News:

Drawing conclusions from data is always a good career.

Four paths (none of which is “data scientist”):

- Industry Specialist
  - Bring data science to new industries
- Analytics & Data Visualization
  - Bring data science to less technical people
- Data Engineering
  - Remember what all those data scientists are spending their time on?
- Go Deep (get a PhD)
  - Push the boundaries

# Stay in Contact

[https://twitter.com/colin\\_jemmott](https://twitter.com/colin_jemmott)

<https://www.linkedin.com/in/cjemmott/>

[cjemmott@ucsd.edu](mailto:cjemmott@ucsd.edu)

[cjemmott@seismic.com](mailto:cjemmott@seismic.com)

<https://www.cjemmott.com/>