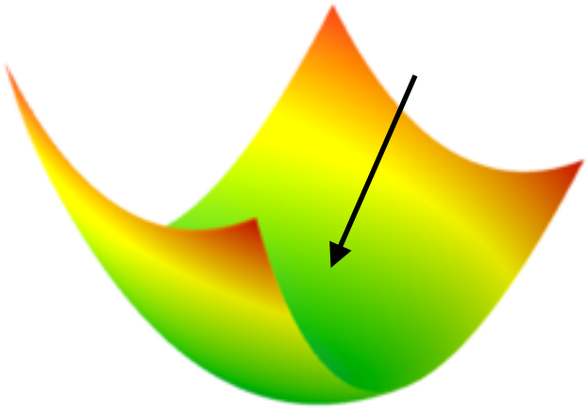
Machine Learning in the Trenches

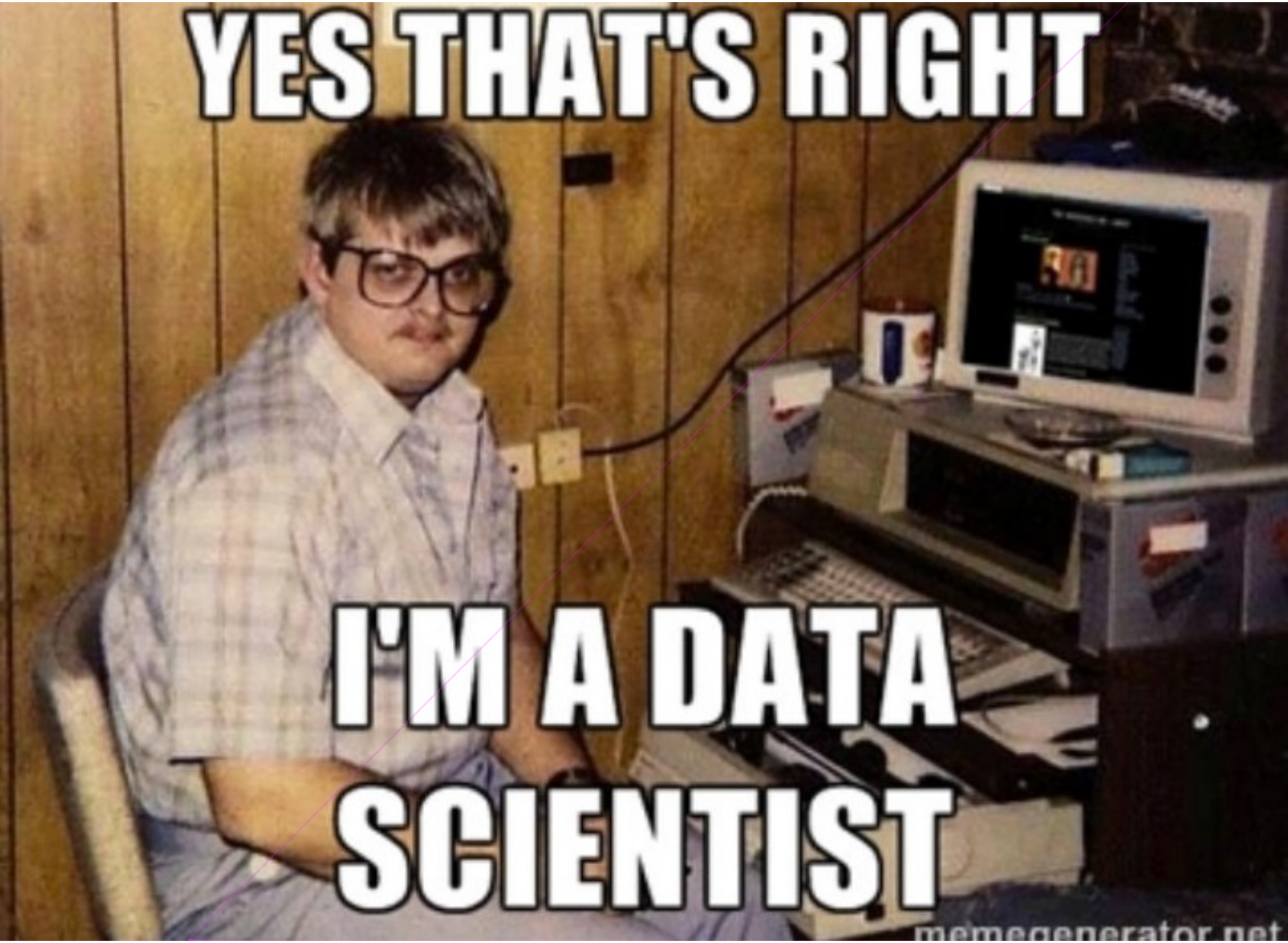
Will Landecker
@thewill

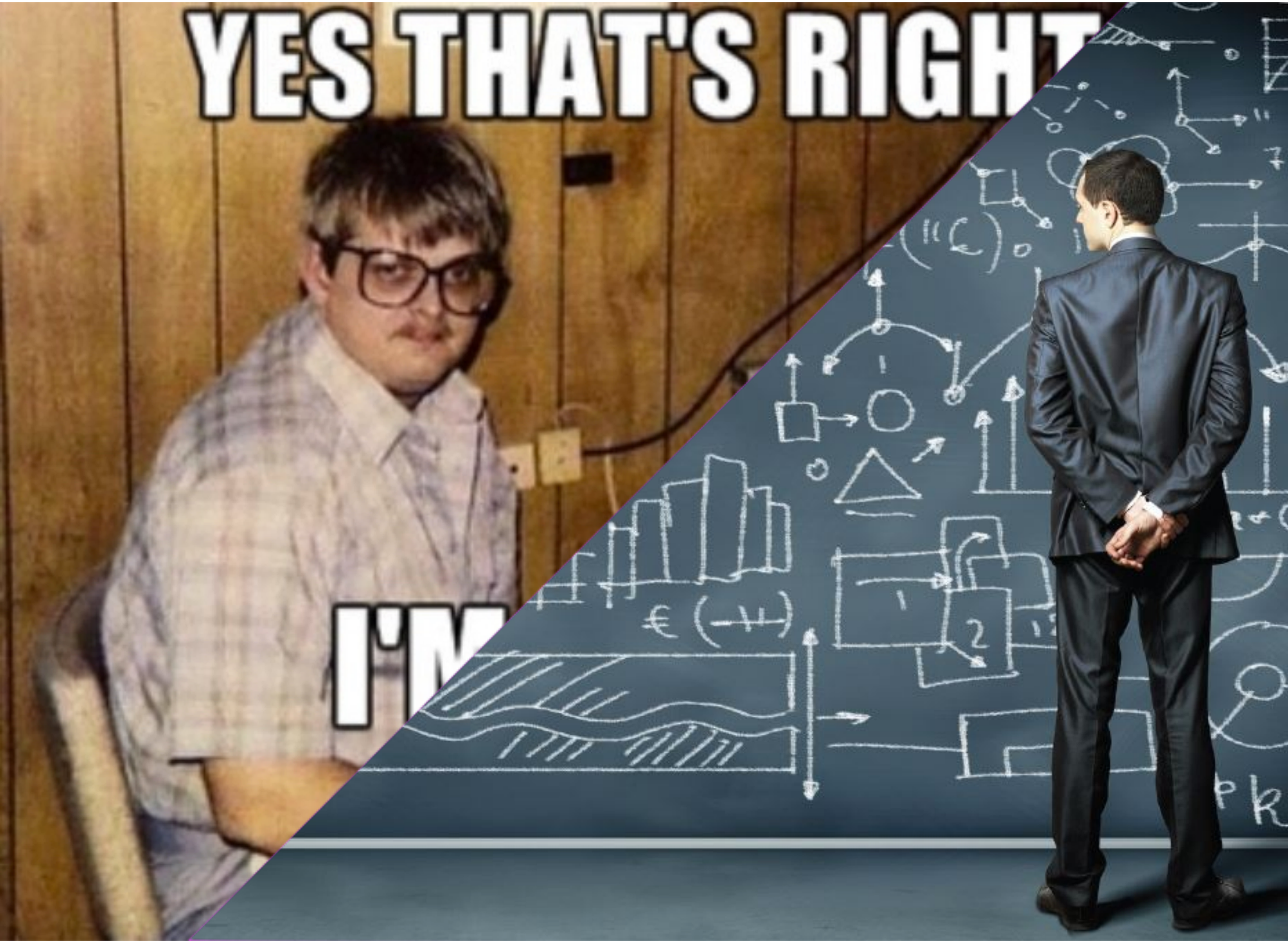


Who is this guy?

- Reed College (math)
- PhD Computer Science under Melanie Mitchell
- Data scientist @ twitter
- Data scientist @ lyft

data science?


YES THAT'S RIGHT
I'M A DATA SCIENTIST
memegenerator.net


YES THAT'S RIGHT
I'M

25 Best Jobs in America

33k 10k 422

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

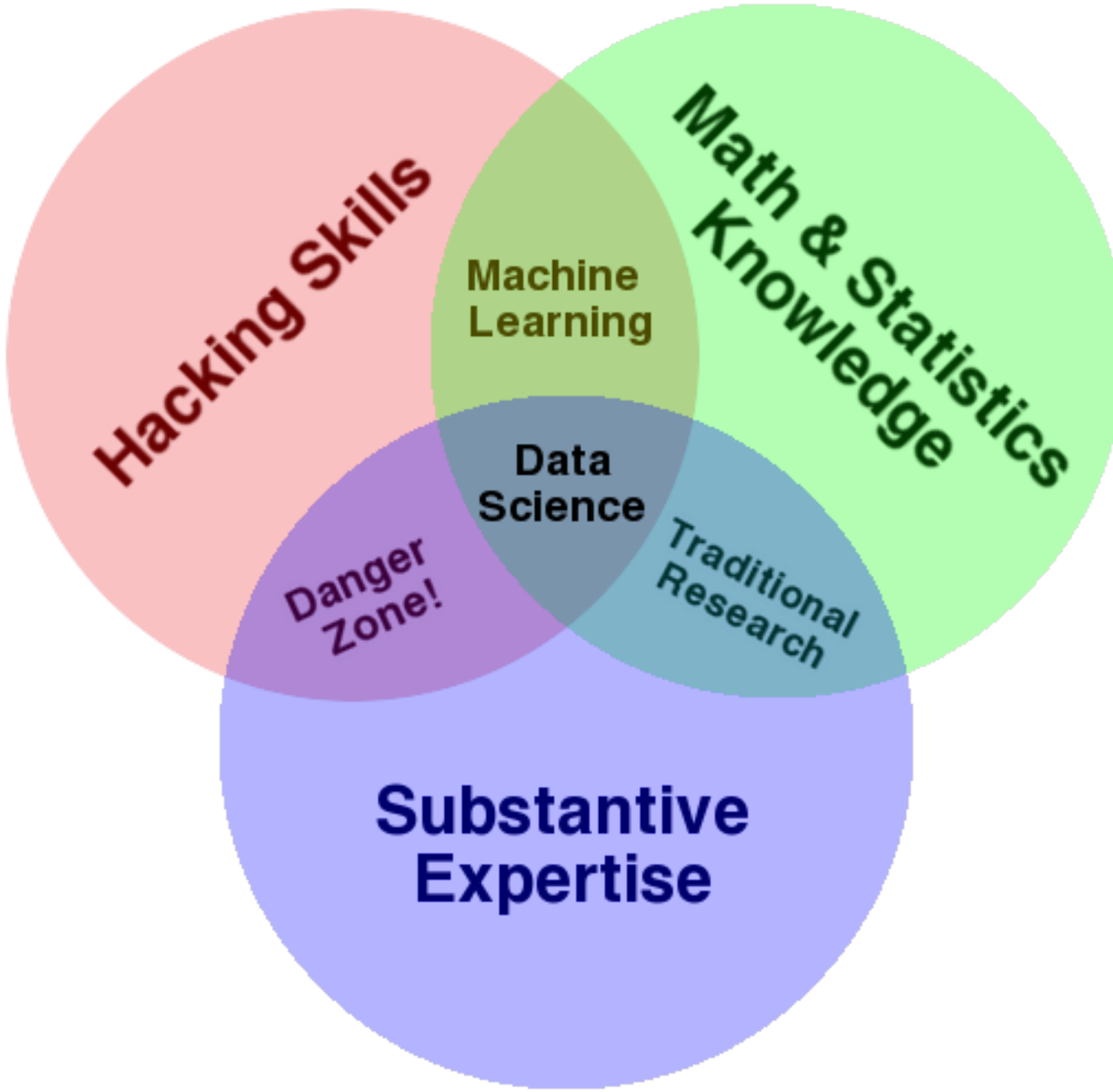
United States 2016

1

Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

(glassdoor)

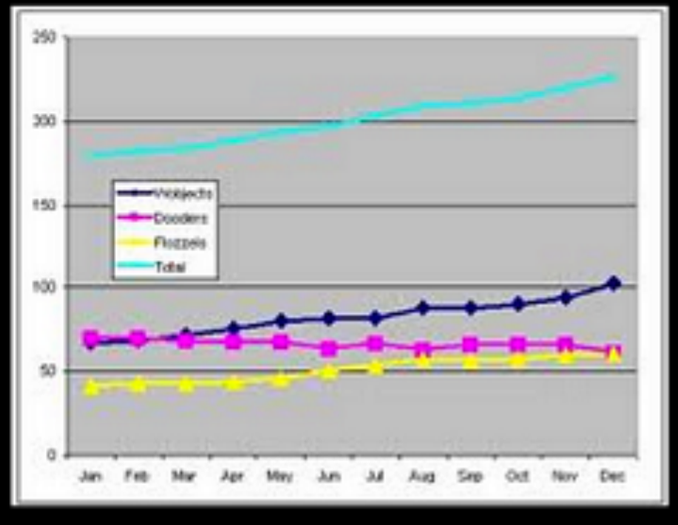

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Data Scientist

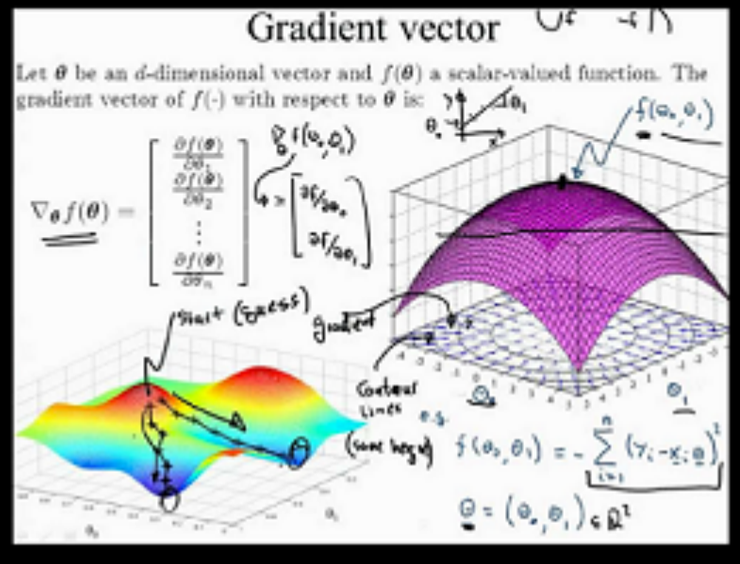
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

What I actually do

Data scientists get asked to do:

- Machine learning research
- Data visualization
- Analytics (counting things)
- Database administration
- Building ETL pipelines; data engineering
- Good old fashioned software engineering

Trey Causey
@treycausey

Following

Dirty secret: the majority of work most data scientists do isn't heavy duty ML, sophisticated algorithms, or cutting edge deep learning.

RETWEETS 21 LIKES 46

9:46 AM - 28 Feb 2016

Reply to @treycausey

Trey Causey @treycausey · Feb 28
On good days, I'm happy to figure out what a good denominator is.

4 28

View other replies

Rob Story @oceankidbilly · Feb 28
@treycausey the "I count things" joke is played out, but seriously there's a lot of counting things.

7

Tom Augspurger @TomAugspurger · Feb 28
@oceankidbilly @treycausey don't forget installing things.

3

What I do

- Machine learning research

  - Choosing / modifying / creating learning algorithms & classifiers

  - Feature engineering

- Finding sources of useful data

- Scientific communication

- Translating business goals into objectives functions I can optimize

- Implementation (sometimes)

What I do

$$\mathcal{L}(x, \lambda) = \frac{1}{2}\|x - x_0\|_2^2 + \lambda(\Phi x - y)$$

- How can machine learning / statistics solve this problem?
- What are the features? What are the labels?
- What imbalances & biases exist in that data?
- What will make for a successful product? What objective function can I use as a proxy?
- Given everything above, what's the best optimization algorithm for the job?
- What is the right operating point for this classifier?
- What are the shortcomings of this solution?
- How will it be implemented in production?

Portland State
@Portland_State

Follow

Free donuts and coffee and big smiles at @PSU_Alums' kick-off to 100-day countdown until commencement 🎓 #2016PDXGrad

RETWEETS 2 LIKES 20

10:53 PM - 4 Mar 2016

Portland State
@Portland_State

Follow

Free donuts and coffee and big smiles at @PSU_Alums' kick-off to 100-day countdown until commencement 🎓 #2016PDXGrad

RETWEETS 2 LIKES 20

10:53 PM - 4 Mar 2016

Joey Devries
@JosephDevries
Follow
Everyone go wish my brother @JackMorlenMusic A Happy 21st Birthday! Hope you had a wicked day brother
RETWEETS
126
LIKES
365
11:21 AM - 6 Mar 2016

How old is each user?

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

Industry ML vs. Academic ML

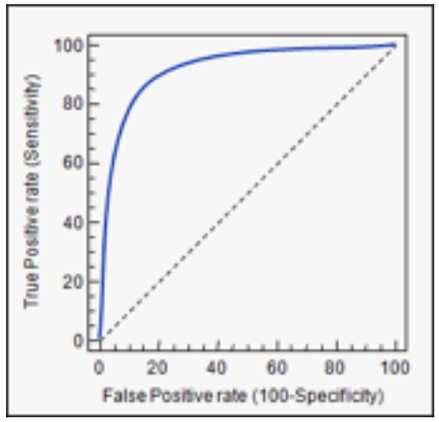
(for me)

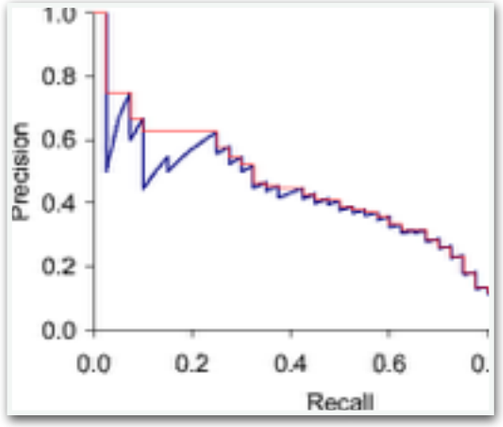
- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy



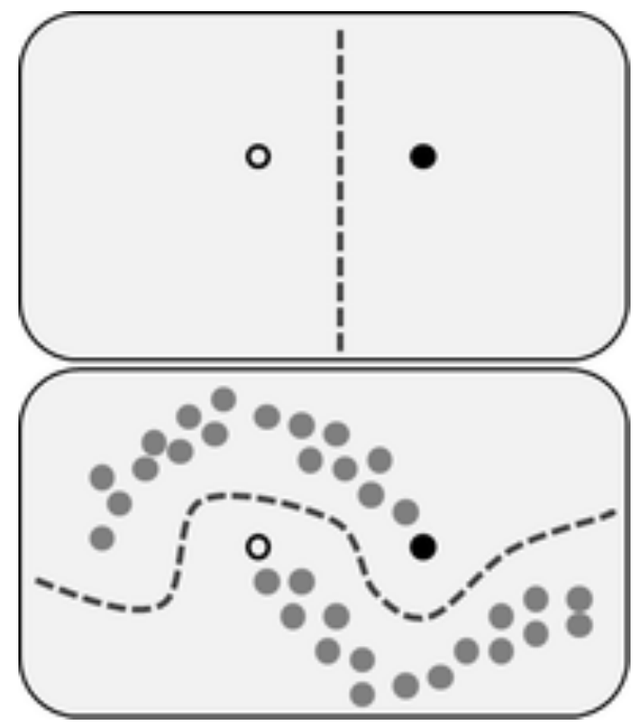


- AUC, F1, Precision @ X% Recall, etc
- How does the metric of success translate to the loss function & the optimization algorithm?
- Sampling, weighting, calibrating

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy



- Unsupervised feature engineering

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

We thank the reviewers for their comments. We are particularly pleased that Reviewer_1 and Reviewer_4 have identified the importance and merits of the submission.

We believe that Reviewer_2 misunderstood some of the terminology and technical aspects of the paper. Reviewer_2 takes issue with Line 343, which claims that DM has an advantage in the less sparse case. In Fig. 2, it is clear that DM outperforms the other algorithms when the sparse codes have more non zeros (higher …

- Advocate for the data
- Expertise != attacking others
  (teach others without being a jerk)

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

- IPython Notebooks
- Scalding / Spark / Hadoop
- Databases

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

- IPython Notebooks
- Scalding / Spark / Hadoop
- Databases

Industry ML vs. Academic ML

(for me)

- Finding features & labels
- Very unbalanced data
- Semi supervised
- Communication
- Tools
- Data Privacy

Resources:

Cosma Shalizi's textbook
http://stat.cmu.edu/~cshalizi/ADAfaEPoV

(and his blog)
bactra.org/weblog/

Trey Causey's blog
treycausey.com

Thanks for listening.
Questions?