# What they don't teach you about data science

David Asboth

# Welcome

About me:

- data scientist + educator
- City & Data Bites alumnus
- half of the “half stack data science” podcast

About today:

- spicy hot takes about data science in the real world

# My path (for context)

- Software development (5 years)
- MSc Data Science (2016)
- Data scientist (4 years)
- Data science educator  (5 months)

# The plan

But please interrupt any time!

A series of hot takes to address:

1. Data science in the wild
2. What skills you **really** need
3. How to prepare for “the real world”

# Hot take 1:
# you won't all work at Google

# Why can't we all be Big Tech?

Because in "half-stack" world:

- Data often collected by accident
- Colleagues usually not techy
- Questions are ambiguous
- Success criteria undefined
- Interpretability matters

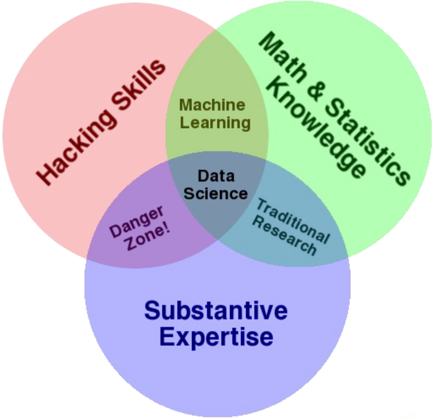
Takeaways:

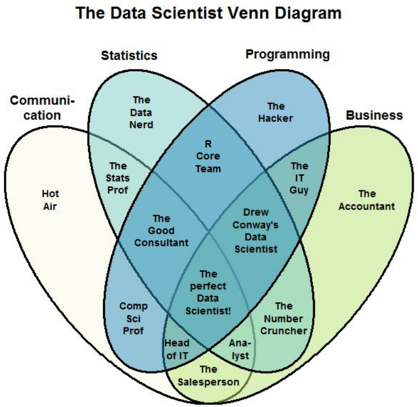
- Data scientists need to be pragmatic problem solvers, not PhD statisticians
- This is the majority of data science (despite what you may read online)

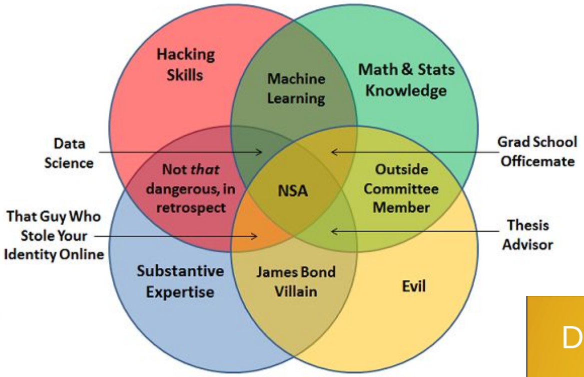
(Caveat: all my advice assumes this will be you)

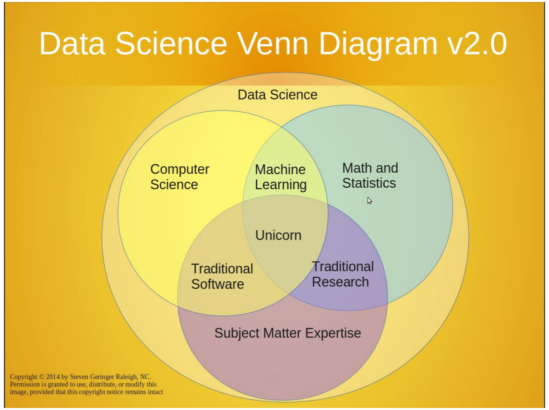
# Hot take 2:
# machine learning is overrated

# What does the world say you need to know?



# What does reality look like?

Reminder:

- Data often collected by accident
- Colleagues usually not techy
- Questions are ambiguous
- Success criteria undefined
- Interpretability matters

scikit-learn doesn’t have a function to solve any of these :-(

# Case study: vehicle sale predictor

- Model to predict probability that a car will sell at auction
- Logistic regression, 6 variables, surfaced via Tableau
- “Successful” from a machine learning point of view, but…
    - Turned out data was never available in time
    - And some pilot customers were unable to act on the recommendations
    - Failure was organisational/cultural not technical

The reasons this model wasn’t used were unrelated to the technical aspects!

# Why does data science fail?

From *[Data science fails (GitHub)]()*

Failures are categorised into:

- Organisational
- Intermediate
- Product Planning
- Product One-Off
- Product Ongoing

No mention that “we don’t have good enough machine learning models”

# Data science hats

**Statistician** - for avoiding unwarranted conclusions

**Scientist** - for thinking about hypotheses to test, not going in blind

**Customer -** to ensure you solve the right problem in the right way

**Developer** - to stay D.R.Y. and facilitate move from PoC to production

# So... what are you saying?

- Data scientists must be good data analysts first(?)
- The tech skills (programming, ML) are **important**
- But they're:
    - Just tools
    - Assumed
    - Only one part of a wider skillset

# What are these other skills you need to succeed?

According to Prof Roger Peng's *The Tentpoles of Data Science*

"Data Science is

1. the application of **design thinking** to data problems;
2. the creation and management of **workflows** for transforming and processing data;
3. the negotiation of **human relationships** to identify context, allocate resources, and characterize audiences for data analysis products;
4. the application of **statistical methods** to quantify evidence; and
5. the transformation of data analytic information into coherent **narratives and stories**"

We agree all of these are needed for success, but <u>most courses only have time to teach 2 & 4</u>

# Hot take 3:
# data “cleaning” is the most important thing you will do

# “Data cleaning IS analysis”

According to Randy Au: Data Cleaning IS Analysis, Not Grunt Work

- “80% is cleaning data” doesn’t mean 80% of time fixing date formats
- Understanding what’s behind the data takes a long time (requires an **analyst** skillset)
- It will be your most valuable contribution as a data scientist

# Case study: competitor data

- Automated download of competitors' public catalogues
  - Saved 30 minutes/day on a previously manual process
- Curated ("cleaned") and surfaced it via Tableau
- Actually looked into the data to find… INSIGHTS

Most of this project was what one might naively call data "cleaning"

# So... what are you saying?

Given that:

1. tech skills are important but not everything, and
2. a lot of value in data science is in the "cleaning"

What **should** you learn?

# Hot take 4:
# you don’t need SVMs

# Which topics are overtaught?

- Support vector machines (sorry)
- Deep learning (unless you're working with images)
- P-values (hopefully not!)

We need a toolbox to build things, but it doesn't have to be enormous to start with

# Which topics are undertaught?

- Exploring data & iterating **at speed**
  - Learn **more** pandas/tidyverse
- True "hacking skills" (e.g. web scraping)
- Working with categorical data
- Time series (!)
- Recommendation engines (maybe)

# Case study: recsys

- Built a recommendation engine to suggest:
  - "similar" cars for existing buyers
  - potential buyers for upcoming sales
  - potential sales for existing buyers outside of their regular auction house
- Recommendations refreshed daily, surfaced via Tableau
- Used **daily**

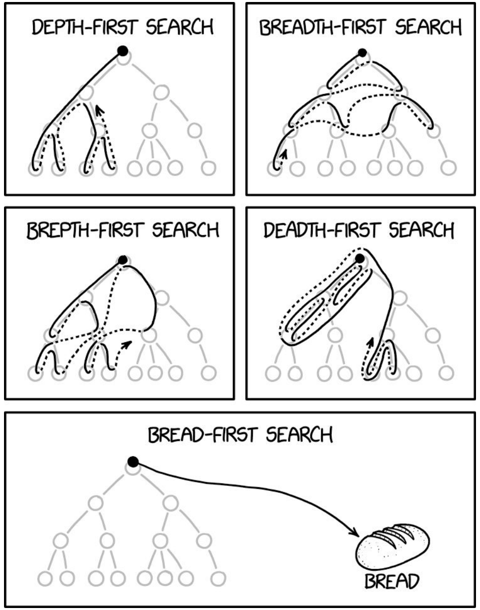
Groundwork was laid months ahead by attending a "buyer team meeting" and just listening

# So… what are you saying?

- Data science education **<u>isn't</u> broken**
- The tech skills are assumed in the job so they **must** be taught in a curriculum
- And yet there's a mismatch between classroom + reality
- This is a **<u>hard problem</u>**

# How to learn data science *(after formal education)*

# Breadth first




- Awareness of multiple topics is better than deep expertise in fewer
- Learning on the job is critical, don't assume you need to know it all up front
- "Jack of all trades"/ "T-shape" etc.

# Build things/ Do project work

- Take the internship
- Show how you solve problems
- Data science lifehack: build things no one asked for - **cultivate this mindset**

# Stay a 'beginner'

- Always ask the 'stupid' questions
- Be curious about how the industry you're in (not just 'data science') works
- Question the basics ("why do we measure X in the first place?")

Recommended reading: The joys of being an absolute beginner – for life

# In conclusion...

- Data science varies a lot, but most of it isn't FAANG
- Figure out what excites you about it + optimise
- It's all about problem solving, not the tools
- Practise, practise, practise

# Q&A

half stack
data science