

A decorative graphic on the left side of the slide. It consists of a blue parallelogram and a light green parallelogram, both tilted at an angle. The blue shape is in the foreground, and the green shape is partially behind it. They are set against a dark blue background with faint, lighter blue diagonal stripes.

DS Capstone Sprint 1



Goal

- Predict whether or not a given startup company will succeed
- Success = acquired/IPO or \$X fundraised and still operating
- According to various sources about 90% of startups fail
- Can this be done using data that's more readily accessible?



Method

What I'd like to do:

- Analyze company descriptions
- Involve logos
- Use news articles

What I'm actually doing for now:

- Analyzing more conventional data (amount fundraised, category list, date founded)



Potential Impact

- Investors could invest more efficiently
- New startups could position themselves to have a better chance of succeeding
- It's interesting
- I'd feel good about it



The Data

Concerns:

- Heavily skewed, most are still operating (as of 2015)
- Not trustworthy:
 - No documentation for data - source is probably Crunchbase
 - Harvard is a startup? Verizon?
 - Almost 4,000 companies procured funding before being founded
 - 9 companies raised a total of \$10 or less in funding



Next Steps

- Wrangle the categories column into more usable format
- Define success more clearly
- Get a status update on the companies using URLs
- Run some basic models
- Switch focus to working with text data/company descriptions