

# Capstone Check-In 1

Guidelines and Recommendations



---

# Capstone Check-In - Part 1

## What?

- Presenting three potential topics and problems
- Describing your:
  - Goals & criteria for success
  - Potential audience(s)
- **IDEALLY**, identifying 1-2 potential datasets/data sources

## When?

- Wednesday, April 29, after lunch

## How?

- Lightning Talks! A 3-5 minute presentation that covers 3 potential topics, including potential sources of data, goals, metrics and audience.

# {Project Idea}

- Data will be collected from:
  - Source1
  - Source2
- My MVP is: a model and something Else.
- My stretch goals include:
  - Goal1
  - Goal2
- My observations will be \_\_\_\_\_ and my target will be \_\_\_\_\_.

I will use   {what}   data  
to build a   {type}   model  
that predicts   {target}   values  
in order to   {value prop}  .

\_\_\_\_\_

# Additional Notes

- Some potential roadblock
- Something I want to research more is \_\_\_\_\_.
- I'm not sure if I can even accomplish \_\_\_\_\_.
- If anyone has recommendations on how to find \_\_\_\_\_, please let me know!

# EXAMPLE: Hit Streak Predictor

- Data will be collected from ESPN API and Some Stats Website
- My MVP is an daily scraper and the main classification model.
- My stretch goal is an automated pipeline that emails me every morning with the top 5 predictions for the day.
- My observations will be batters, representing a single matchup. My target will be binary, whether or not they got a hit that day.

I will use batter performance data to build a binary classification model that predicts whether or not a batter will get a hit in order to try and win the MLB “Beat the Streak” competition.

---

# Additional Notes

- Data collection/wrangling will be an issue due to the abundance of data. Each observation will need to be a single day for the batter so I will need to reformat a lot of the information I will have.
- I need to research more expert analysis to see what the important features might be.
- My stretch goal will be difficult, I will need help on automating the process and running it each day at a specific time.
- If anyone has recommendations on how to send emails with python, please let me know!

## EXAMPLE: Produce Image Classifier

- Data will be collected from the flickr API
- My MVP is a NN that beats baseline accuracy for broad produce categories.
- Stretch goal: a species specific model that is deployable on the phone.
- My observations will be single images and the target will be the fruit label.

I will use produce images to  
build a multi-class classification  
model that predicts produce  
type in order to improve a  
frustrating part of the checkout  
process..

---

# Additional Notes

- Image data is “fun” to work with
- We haven’t learned NN yet so you will have to be comfortable implementing topics on the fly.
- Image data is heavy.



# Resources

- <https://github.com/BrianLKane/capstone>
- <https://github.com/irinhwng/Image-Classification-of-Fruits-and-Vegetables>
- <https://gallery.generalassemb.ly/DSI?metro=>
- <https://toolbox.google.com/datasetsearch>
- [Places to Get Interesting Datasets](#) (from the Resources repo)