# Springboard Capstone

## Guidelines

**Write a description for three capstone project ideas. Your ideas can be broad and high-level. The descriptions should address the problem and identify the data you’ll use to solve it. You do not need to talk about specific methods and techniques.**

How to Pick Your Capstone Projects

● Is it a real problem that someone cares about?

● Is there real data available?

● Is the data easy to acquire and clean?

*I want a real problem with real data available that is relatively easy to acquire and clean*

**First Capstone Project**: For your first capstone project:

a. Propose a topic that is straightforward in both approach and complexity.

b. (Optional) Choose a project that leverages any specific industry expertise you may already have from prior work experience.

c. The goal is to get a handle on a larger project without feeling overwhelmed, so select a project that doesn’t require web scraping, which is often more difficult than it appears.

Step 1. Initial Project Ideas

Datasets:

1. Quandl,
2. US Government Open Data,
3. UCI Machine Learning Repository,
4. Kaggle competitions,
5. Mode Analytics​, or ​
6. Google's public datasets directory​
7. Data is Plural

Use Google Dataset Search

Once you have a few ideas, narrow them down to the one idea that you’re most excited to work on

## Ideas

* Look at public datasets
* Check out forum for ideas
* Check out Springboard Youtube Channel
* Ask friends and family
* Do Google Search for interesting ideas
* Ask Lisa and Jason for capstone ideas?
* Ryan says do 311 data or churn model

### Public Datasets

#### Quandl:

#### US Government Open Data

* Topics:
  + Consumer
  + Education
  + Finance
  + Health
  + Public Safety

#### Mode Analytics

1. CrunchBase
   1. Data about start-ups and VCs
2. 538:
   1. economics,
   2. culture

#### Google Public Dataset Directory

1. Labor Productivity and Costs from the US Bureau of Labor Statistics
   1. Labor productivity through time
2. Economy size through time
3. OECD Factbook – Where will software engineers come from?

#### Data is Plural

* Two thousand billionaires
* What do you do with a PhD in science
* What kills us – CDC
* Business owners
* financial well-being

Google Dataset Search

#### Forum

Classic Topics

* Health: Predict if a patient is at risk of disease X; health factors of disease X
* Crime in city X: which factors contribute most to crime; likelihood of crime in certain areas; predicting crime in areas
* Predict property prices
* Predicting if a loan will be accepted; predicting loan defaults

#### Choose a dataset that is relatively clean and easy to use. As a rule of thumb, if you think that you’ll have to spend more than one week acquiring and cleaning your data, you may want to reconsider and find a cleaner dataset.

#### Two thousand billionaires

* What factors contribute most to becoming a billionaire? What are the commonalities among billionaires?
* What factors distinguish American billionaires from European ones? What factors distinguish software billionaires from other ones? Young billionaires from old ones? Exploring the dataset to find interesting features
* How has the composition of billionaires changed over time?
* What distinguished billionaires from mere-millionaires?
* 2,615 rows in the dataset: billionaires in 1996, 2001 and 2014
* 30 Columns: Year, Name, Rank, Citizenship, Countrycode, networthUSbillion, selfmade, typeofwealth, gender, age, industry, IndustryAggregates, region, north, politicalonnection, founder, generationofinheritance, sector, company, companytype, relationshiptocompany, foundingdate, gdpcurrentus, sourceofwealth, notes, notes2, source, source2, source3, source4

Dataset:

* Peterson Institute For International Economics: <https://piie.com/publications/working-papers/origins-superrich-billionaire-characteristics-database?ResearchID=2917>
* Forbes: <http://www.forbes.com/billionaires/list/>

Description:

* This dataset contains biographical information about every billionaire in the world for the years 1996, 2001, and 2014: gender, age, country of origin, industry, company type, source of wealth, and so on

Question/Problem to Analyze:

* What factors contribute most to becoming a billionaire? What are the commonalities among billionaires?
* What factors distinguish American billionaires from European ones? What factors distinguish software billionaires from other ones? Young billionaires from old ones?
* How has the composition of billionaires changed over time?
* What distinguished billionaires from mere-millionaires? (would require a dataset of millionaires, which would be vastly larger than the fewer than 2,000 billionaires in the world, no? )

Notes/Concerns:

* Already a complete working paper on this topic. Would not be doing original work. Would simply be reproducing the author’s work, no?
* This project looks like one that only requires Excel, specifically the use of Pivot Tables. It seems like a classic business intelligence problem, not a data science one. I know the line is often blurry, but still, shouldn’t there be number crunching of some sort to reveal patterns?

#### What do you do with a PhD in science?

Dataset:

* National Science Foundation’s Survey of Doctorate Recipients :
  + https://www.nsf.gov/statistics/srvydoctoratework/#sd&tools&micro&profiles&tabs-1
* Sponsored by the National Center for Science and Engineering Statistics and by the National Institutes of Health

Description:

* Biennial Survey conducted since 1973
* Provides Demographic, education, and Career History Information about individuals with a research doctoral degree in science, engineering, or health (SEH) from US institutions
* Data available from 1993
* Key variables:
  + Demographics (e.g., age, race, sex, ethnicity, citizenship)
  + Educational history
  + Employment Status
  + Field of degree
  + Occupation

Problems/Questions:

* Which doctorates earn the most money?
* Which doctorates stay in academia the most? Industry? Government? NGOs? Etc.
* Which doctorates live the longest?
* Which doctorates are healthiest?
* Which doctorates have the most children?
* Which doctorates have the most prior education?
* How has the demographics of doctorate earners changed with time?
* What are the main ways in which doctorate earners differ from non-doctorate earners? (Would need data about non-doctorate earners for this)

Notes/Concerns:

* Lots of numerical data here, so that looks good. Too much?
* Data rich
* Also have access to National survey of college graduates – compare to regular graduates?

#### What Kills Us

Dataset:

* CDC Underlying Cause of Death Database:
  + https://wonder.cdc.gov/ucd-icd10.html

Description:

* County level mortality statistics of U.S. residents, based on death-certificates
* 1999 to 2014
* Provides information about:
  + Geography
  + Demographics
  + Place of death
* Also have “compressed mortality” datasets going back to 1968

Problems/Questions:

* How does cause of death differ by location, demographics, etc.
* What diseases are most likely to occur for each state?

Notes/Concerns:

* Doesn’t let me export into an Excel file; have to query on their website, I think

#### Business owners

Dataset:

* U.S. Census Bureau’s Survey of Business Owners and Self-Employed Persons:
  + https://www.census.gov/programs-surveys/sbo/about.html

Description:

* “Provides the only comprehensive, regularly collected source of information on selected economic and demographic characteristics for businesses and business owners by gender, ethnicity, race, and veteran status.”
* Collected every 5 years since 1972
* all nonfarm businesses filing Internal Revenue Service tax forms
* Business ownership is defined as having 51 percent or more of the stock or equity in the business and is categorized by:
  + Gender: Male; female; or equally male/female
  + Ethnicity: Hispanic; equally Hispanic/non-Hispanic; non-Hispanic
  + Race: White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; some other race; minority; equally minority/nonminority; nonminority
  + Veteran status: Veteran; equally veteran/nonveteran; nonveteran
  + Publicly held and other firms not classifiable by gender, ethnicity, race, and veteran status
* Characteristics of Business Owners:
  + How Initially Acquired Business
  + Year Acquired Ownership of Business
  + Primary Function(s) in Business
  + Average Hours Per Week Spent Working
  + This Business Primary Source of Income
  + Prior Experience Owning a Business
  + Highest Level of Education Completed
  + Age of the Owner in 2012
  + Owner Born a Citizen of the United States
  + Service-Disabled and Other Veteran Characteristics

Questions/Problems:

* Has the education of business owners increased with time?
* Has the percentage of the economy made up by businesses grown or shrunk?
* Has the number of small businesses grown or shrunk?
* What are the main sectors that small business owners work in?
* How old are most entrepreneurs when they start their first business?
* How does the success of second businesses compare to first businesses?
* Are firms with more owners more successful than firms with fewer owners? Older vs younger?

Notes/Concerns:

* Lots of data in tables that I’m able to export; many, many tables

#### financial well-being

Dataset:

* Consumer Financial Protection Bureau’s National Financial Well-Being Survey:
  + https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

Description:

* 6,000 responses to the agency’s 10-question Financial Well-Being Scale
* additional demographic and financial information
* Includes information on:
  + Income and employment
  + Savings and safety nets
  + Past financial experiences
  + Financial behaviors, skills, and attitudes

Questions/Problems:

* What are the main factors contributing to financial well-being? Top 5
* What are best practices for financial well-being for someone just graduating university? Mid-life? Retiring?
* What financial attitudes are best for becoming wealthy?
* How does having children impact one’s financial attitudes, behaviors?
* Does education improve one’s financial well-being and attitude towards money
* What are the main ways men and women differ when it comes to spending habits?
* How does greater understanding of finance (stocks, mutual funds, volatility, inflation, compound interest, etc.) affect financial behaviour?

Notes/Concerns:

* Have stub code to read the survey data into Python, R, etc.
* Survey Data File is **all** numbers; little string or categorical data which could be a downside, I think; in other words, survey is small; on 2nd thought, might not be an issue because the numbers stand in for verbal responses
* Comes with a user’s guide, which is a plus
* Many, many columns, a big plus

## Kaggle

### Competitions: https://www.kaggle.com/competitions?search=tag:%27finance%27

Featured: these are public competitions with significant prize money meant to solve commercial problems. Prize winners grant the sponsor a non-exclusive license to their work, and will present their results via a detailed write-up and possible screencast (see below).

Playground: are public competitions set up to be fun, quirky and idea-driven, rather than to solve a specific business or research problem. These are for fun and not for any prize.

#### Playground

* DonorsChoose.Org Application Screening: https://www.kaggle.com/c/donorschoose-application-screening
  + Predict whether teacher’s proposals are accepted
  + The goal of the competition is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved, using the text of project descriptions as well as additional metadata about the project, teacher, and school.
  + Comes with a tutorial that provides an overview of the dataset as well as a quick-start guide to building my first model in TensorFlow and submitting my entry to Kaggle
  + Is a Machine Learning course
  + Competition ended 6 months ago
  + Competition sponsored by Donorschoose.org and Google Machine Learning Crash Course

#### Featured

* Home Credit Default Risk: <https://www.kaggle.com/c/home-credit-default-risk>
  + Finished 2 months ago, ran for 1 week with 70k of prize money
  + Predict clients’ repayment abilities
  + Sponsored by Home Credit International
* Santander Product Recommendation : <https://www.kaggle.com/c/santander-product-recommendation>
  + predict which products their existing customers will use in the next month based on their past behavior and that of similar customers
  + $60k of prize money
  + Ran for about 1 week in 2016

### Datasets:

* Credit Card Fraud Detection: https://www.kaggle.com/mlg-ulb/creditcardfraud/home
  + The datasets contains transactions made by credit cards in September 2013 by european cardholders.
  + we have 492 frauds out of 284,807 transactions (unbalanced dataset)
  + features not original: transformed via PCA to preserve confidentiality; only time, amount, and Class (fraud or not) are preserved
* Daily News for Stock Market Prediction: <https://www.kaggle.com/aaron7sun/stocknews/home>
  + Data is news data (from reddit) and stock data from DJIA
  + Binary classification task: DJIA rose or stayed the same
  + Guy said for his NLP and Deep Learning course, so maybe not good
* Lending Club Loan Data: <https://www.kaggle.com/wendykan/lending-club-loan-data/home>
  + Data from 2007-2015
  + Predict loan acceptance
* Default of Credit Card Clients Dataset: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home
  + Default Payments of Credit Card Clients in Taiwan from 2005
  + Payment default prediction
  + 25 variables (including default payment (1=yes, 0=no)
  + Some ideas for exploration:
    - How does the probability of default payment vary by categories of different demographic variables?
    - Which variables are the strongest predictors of default payment?
* US Consumer Finance Complaints: <https://www.kaggle.com/cfpb/us-consumer-finance-complaints>
  + Predict something about the company response from the complaint?