ISMT S-117, Text Analytics and Natural Language Processing

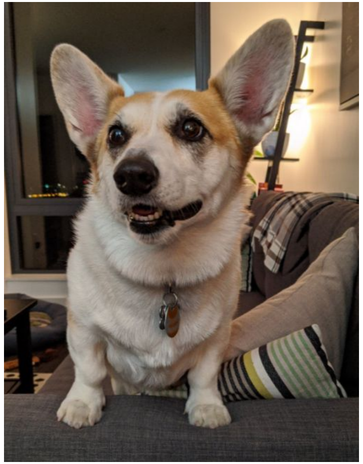
Research Design for Data Science Projects

Dr. Andrew Therriault
Northeastern University

JULY 20, 2020

Introducing myself...

- Education highlights:
  - NYU Politics (PhD, MA, and BA)
  - Vanderbilt Political Science (Post-Doc)
- Professional highlights:
  - Director of Data Science @ Democratic National Committee (2014-16)
  - Chief Data Officer @ City of Boston (2016-18)
  - Data Science Manager @ Facebook (2018-2019)
  - Data Science and Strategy Consultant (2019 - present)
- A few random personal highlights:
  - Living in East Boston with Amelia (→→→)
  - Go-to karaoke song: *Folsom Prison Blues* by Johnny Cash
  - Secret talent: breakfast burritos
  - Quarantine coping mechanism: night cheese.



How does data science provide value?

- Data scientists don't *do* anything themselves - instead, they help others to do their jobs better
- The value added by data science comes from providing information to customers / stakeholders, which enables better decision-making
    - **Strategic**: used to determine overall goals and direction
    - **Tactical**: used to accomplish specific objectives in implementing that strategy
- Data science is most valuable under resource constraints (broadly defined), because it is most often used to allocate resources efficiently and effectively
    - **Efficiency**: allocating resources where they are most needed
    - **Effectiveness**: making sure allocated resources have the greatest impact possible

Data science as an information process

- How do we use data?
  - **Collection**: finding out what's going on
  - **Presentation**: sharing what's going on
  - **Analysis**: finding patterns in what's going on
  - **Interpretation**: figuring out why this is going on
  - **Extrapolation**: predicting what's going to happen next
- The "science" in data science comes especially from the last two, but science is involved in all parts of the process

Data science as applied research

- Connecting data science to academic and applied research
  - Academic research produces knowledge primarily for its own sake, in order to develop and refine theories about how the world works
  - Applied research aims to solve specific problems, and is often based on theory but only rarely contributes back to theory
  - Data science is a generalization of applied *quantitative* research across fields, with a focus on using technology to answer questions quickly and at scale
- Quality research requires careful planning in order to ensure validity of the information delivered and to minimize risks of bias and error
  - **Validity**: is your data telling you what you think it is?
  - **Bias risk**: are the patterns in your data representative of patterns in the real world?
  - **Error risk**: do you have enough “signal” in your data to see through the “noise”?

Designing for data science

- Data science projects combine traditional research design with engineering, product development, and project management
- The design phase of a project requires you to gather information and make decisions about what you want to accomplish
- What a project design offers:
    - Clarifies the *purpose*, *scope*, and *deliverables* of your project
    - Identifies the requirements for success and highlights potential risks
    - Provides an opportunity to get concrete feedback before development starts
    - Gives the project team a shared vision and roadmap to work from
- A design for a data science project is *not* a formal framework or plan, but should provide most of the information needed to create one

What to include in a design

1. The context of the problem
2. A specific research question to answer
3. The source(s) of the data you’ll use to answer the question
4. What you’ll do with the data to get to the answer
5. Any other requirements or dependencies for the project
6. Any risks or limitations you foresee
7. How you’ll validate your results
8. What you’ll deliver to your end user
9. How the deliverables will be supported and maintained

1. The context of the problem

- What is the overall goal you're working toward?
    - This is broader than the data science problem (e.g., "selling more cars", "improving patient survival rates", "recommending songs people will like")
    - The goal should ideally suggest some way to measure success
- Who are your stakeholders / customers / end users (collectively, your "partners")?
    - These may be different (e.g., campaign managers vs. field directors vs. volunteers)
    - You should try to identify their particular wants / needs / challenges
- What do they want to know that they don't know right now?
- How would that help them succeed?

2. A specific research question to answer

- What decisions (strategic and/or tactical) would be better made if your partners had more information?
  - This should connect back to the wants / needs / challenges of your partners and what they don't already know
- What measurement(s), estimate(s), or prediction(s) would better inform those decisions?
  - Look back at the "Data science as an information process" slide - for example, an interpretation might be an estimate of a causal relationship (e.g., the effect of an advertisement on purchase rates), while an extrapolation might be a predicted sales forecast for the next year
  - Be as specific as possible about what information you want to produce and deliver
- Are you offering a causal answer, or a descriptive / predictive one?

3. The source(s) of the data you'll use

- What inputs do you need to answer this research question?
  - Two general types of data sources: **observational** and **experimental**
    - Observational data is produced by measuring the results of natural processes (without interference), while experimental data is the result of measuring a researcher-designed process
    - Observational data tends to offer greater *external validity*, while experimental data offers greater *internal validity*
- How will you acquire this data?
  - Think at this stage about access, cost, and how the data source is maintained (and whether that may change as a result of your use of the data)
- Are there are limitations or challenges in acquiring / using this data?
  - Think here about privacy, security, legal restrictions, reliability, format, etc.

4. What you’ll do with the data

- How will you transform the raw data sources into answers to your research question(s)?
  - Be as specific as possible, but leave room for trying alternatives (e.g., use “a binary classifier” rather than “a logistic regression” because there are many options)
  - Be sure to think about the data wrangling and merging required to get to a dataset that you can model or analyze
- What exactly does this approach to analysis / modeling your data tell you?
  - For example, a classification model doesn’t literally tell you what will happen in the future - it tries to approximate the data generating process in a way that allows for extrapolation to new data points
- If your question involves causality, how do you demonstrate it?

5. Any other requirements or dependencies

- What kind of infrastructure, software, services, etc., are required for the project, and are you relying on anyone else to support those?
- Do you need input or help from another person or organization in developing the project or answering critical questions?
- Do you need something external to happen (e.g., a software update, budget allocation, or new hire) in order to start / finish the project?
- Will putting your deliverables into practice require time, resources, etc., from your partners?

6. Any risks or limitations you foresee

- What could cause your project to not produce a useful result, or to produce one that has unacceptable levels of bias or error?
    - For example: if an underlying data source is biased, it's often impossible to eliminate bias in the analysis / modeling of that data
    - When collecting new data for a model, you often don't know ahead of time if the patterns are strong enough in a given sample to generate worthwhile predictions
- Are there limitations in how your results can or should be used which prevent it from being applied more broadly?
    - A result might be applicable in one context, but not very helpful in others (e.g., a model that's trained using one part of a population may not predict well for others)
    - There may be unavoidable levels of uncertainty or known edge cases in your results that require you to work around them

7. How you'll validate your results

- What are the internal metrics you can look at (both formal and informal) to check the accuracy and validity of your results?
    - Could be specific quantities (R-squared values, accuracy rate, etc.) or more general patterns (distribution of predictions across groups)
    - "Sanity checks" are often just as important: do the patterns you're seeing in your analyses / estimates / predictions align with your prior expectations (violation risk vs. inspection history)?
- What external data sources can you compare against?
    - Comparable or similar work done by others or in other contexts
    - Related external indicators that should be highly correlated (e.g., perceived covid risk vs. local infection rate)

8. What you’ll deliver to your end user

- What specific form(s) does your answer take?
  - Is it a data file? A dashboard? An application? A report?
  - How polished does it need to be?
  - You may need to deliver multiple variations to fit specific use cases
- What is the delivery mechanism and schedule?
  - Are you delivering a prototype / MVP and then iterating, or waiting for a final product?
- What kind of documentation do you need to produce?
  - Different kinds of end users may have different needs
  - Technical documentation and user documentation are distinct
- Do you need to provide any kind of training or hand-off?

9. Supporting and maintaining deliverables

- Does your project update its data and outputs automatically, or does it require someone to do it by hand?
- Do users need ongoing training and support?
- Which of the dependencies and requirements need to continuing being met going forward?
- Who fixes things if something breaks, and what happens to the users if it's not available?
- What's a realistic lifespan for what you're creating?

A few last thoughts...

- These questions and suggestions are a template, but not a formula, so adapt them as needed to fit your situation
- We haven't specifically discussed how to account for the time / effort required for a project, but it's important to keep that in mind when deciding on a realistic scope
- You may not have all the answers to these questions before starting, but this will help to identify those situations and work on them
- With experience, you may find that a different approach works for you, so think of it as a starting point to try and see how it fits for you

Questions?