Workflow for Model Building

Joe

Session Objective

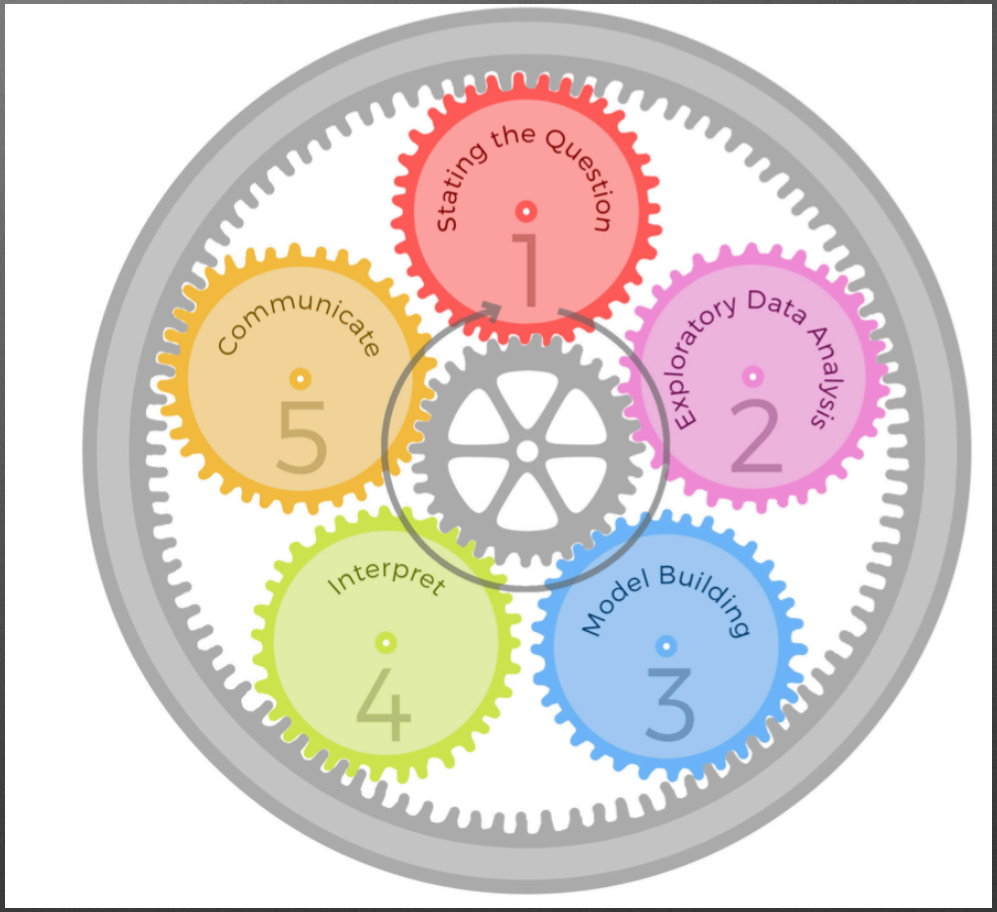
- Describe the 5 skills to oversee successful model synthesis
- List consequences of doing data science mindlessly

Data Science Workflow

Epicycles of Analysis

Data science is not a linear process.

Be careful about decisions, but also note that perfect is the enemy of good.



DS Expectation vs Reality

EXPECTATION	WORST CASE REALITY	REQUIRED SKILLS
Manager understands what is possible	Manager believes you are a wizard	1. Formulating Intelligent Questions
Data for this problem exists and is well documented	Data was collected by an intern two summers ago	2. Data Collection, Cleaning, & Exploration
Projects have well defined life cycles and metrics	You iterate infinitely on the same problem because there are no measures of effectiveness	3. Define MOEs and MOPs
You are a member of a skeptical community	You may be a lone data scientist, or work on isolated projects	4. Communicate Uncertainty 5. Treat all models with scrutiny

Forming Intelligent Questions

Ask the Right Question

Abraham Wald

https://www.fastcodesign.com/1671172/how-a-story-from-world-war-ii-shapes-facebook-today

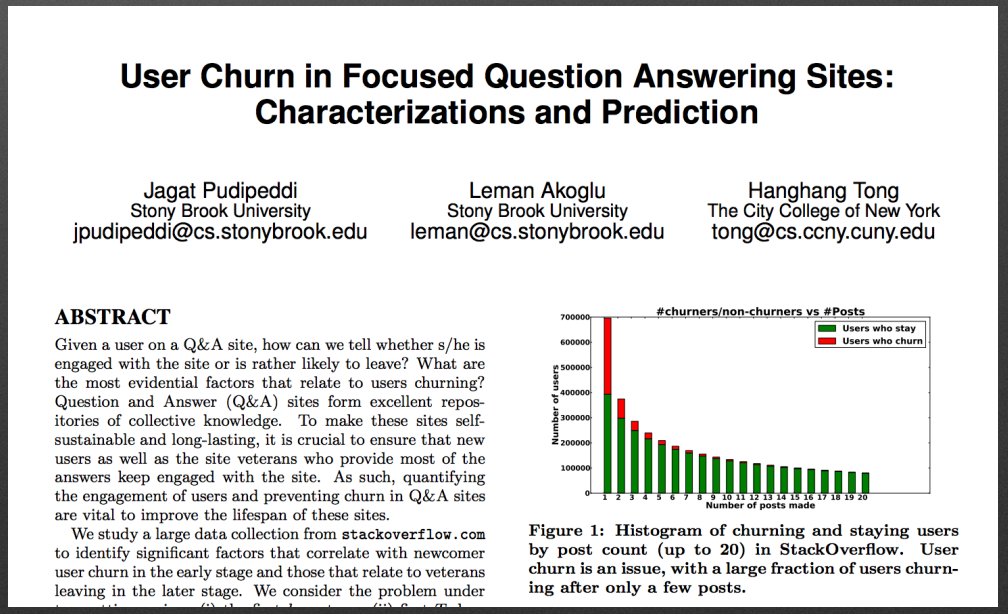
Do Your Homework

Is the problem old?

Yes - read a book

No - Google scholar

Google scholar is also a cool place to get inspiration for capstones ;-)



User Churn in Focused Question Answering Sites: Characterizations and Prediction

Jagat Pudipeddi
Stony Brook University
jpudipeddi@cs.stonybrook.edu

Leman Akoglu
Stony Brook University
leman@cs.stonybrook.edu

Hanghang Tong
The City College of New York
tong@cs.ccny.cuny.edu

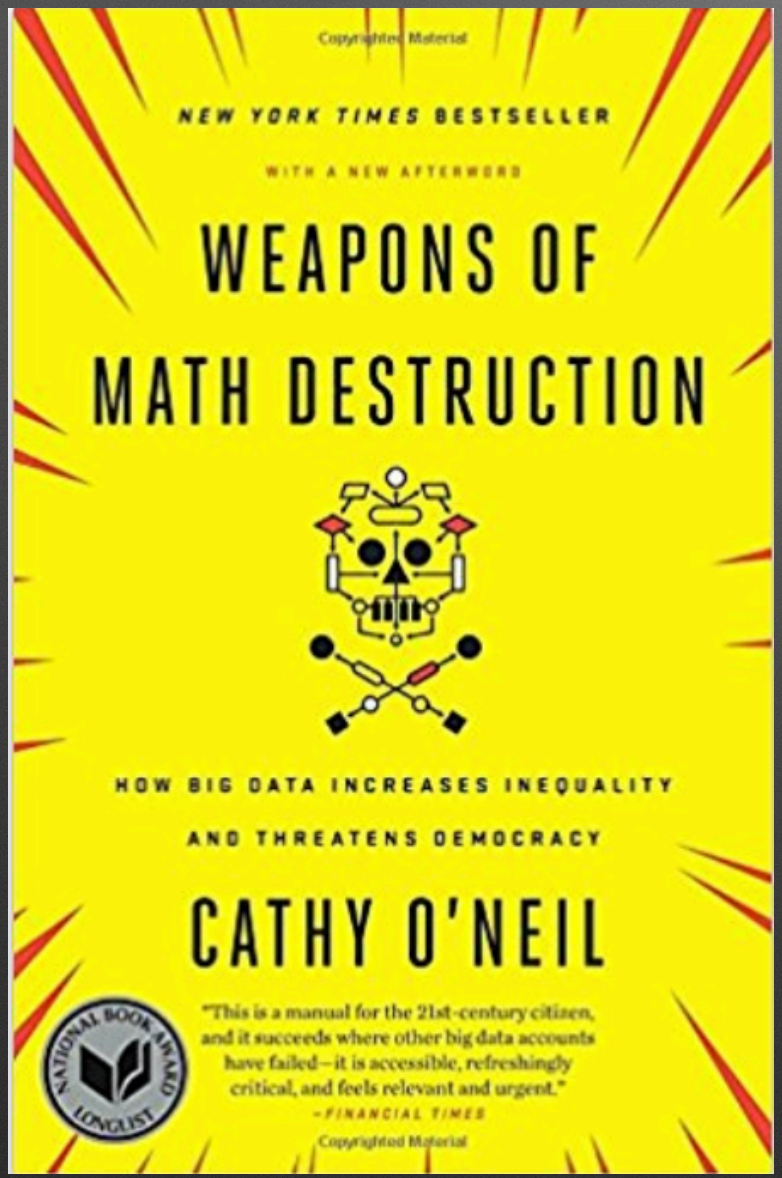
ABSTRACT

Given a user on a Q&A site, how can we tell whether s/he is engaged with the site or is rather likely to leave? What are the most evidential factors that relate to users churning? Question and Answer (Q&A) sites form excellent repositories of collective knowledge. To make these sites self-sustainable and long-lasting, it is crucial to ensure that new users as well as the site veterans who provide most of the answers keep engaged with the site. As such, quantifying the engagement of users and preventing churn in Q&A sites are vital to improve the lifespan of these sites.

We study a large data collection from `stackoverflow.com` to identify significant factors that correlate with newcomer user churn in the early stage and those that relate to veterans leaving in the later stage. We consider the problem under

Figure 1: Histogram of churning and staying users by post count (up to 20) in StackOverflow. User churn is an issue, with a large fraction of users churning after only a few posts.

Ask the Tough Questions



- Not all relevant factors exists in data
- Mindless models can:
  - Anonymize systematic racism
  - Bias against disabled (in violation of the law)
  - Violate civil liberties
  - Kill people

Data Wrangling

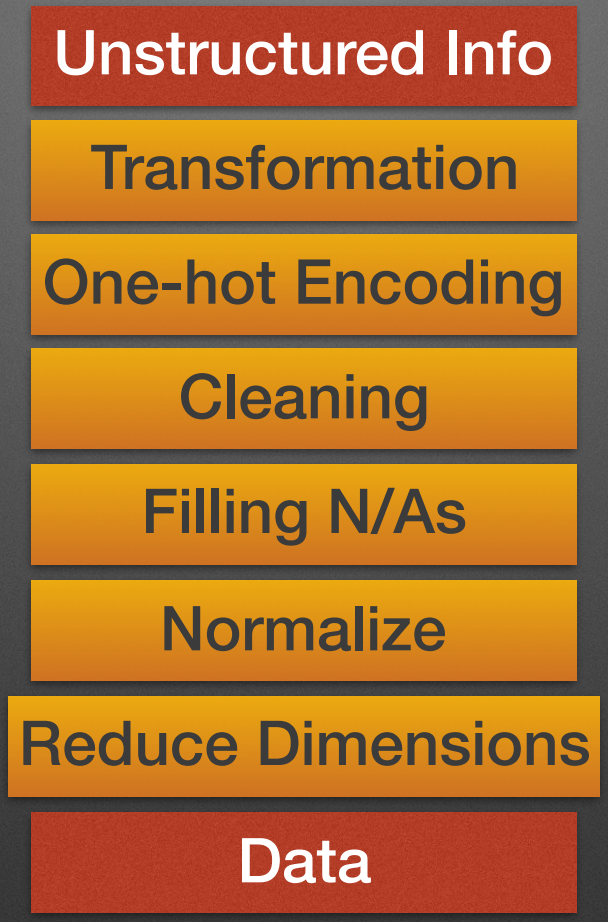
The Secret Lives of Scientists

Data:

- Well formatted
- Typically some mixture of numeric and categorical variables
- Models can read data
- Below is one tweet from the twitter api….
- Data wrangling is the process of turning unorganized information into data

{"created_at":"Thu Sep 21 16:06:00 +0000 2017","id":910897913097801729,"id_str":"910897913097801729","text":"I wish babe didn't have to work the night of Cojo concert because we would've def been there \ud83d\ude29","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":852773260177489922,"id_str":"852773260177489922","name":"Allie","screen_name":"AllieeAmise_","location":null,"url":null,"description":"KZS \u2665\ufe0f","translator_type":"none","protected":false,"verified":false,"followers_count":113,"friends_count":134,"listed_count":0,"favourites_count":178,"statuses_count":461,"created_at":"Fri Apr 14 06:39:23 +0000 2017","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/909215640594128897\/8xlikW6z_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/909215640594128897\/8xlikW6z_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/852773260177489922\/1505931694","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"e0060cda70f5f341","url":"https:\/\/api.twitter.com\/1.1\/geo\/id\/e0060cda70f5f341.json","place_type":"admin","name":"Texas","full_name":"Texas, USA","country_code":"US","country":"United States","bounding_box":{"type":"Polygon","coordinates":[[[-106.645646,25.837092],[-106.645646,36.500695],[-93.508131,36.500695],[-93.508131,25.837092]]]},"attributes":{}},"contributors":null,"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1506009960804"}

Data Wrangling Path



Transforming Data

- ML models are driven by mathematics, meaning that everything needs a mathematical representation
- Many data cleaning steps are specific to a particular technique
  - Text Tokenization, Stemming, etc for NLP
  - Scale normalization for clustering

Steps for EDA

1. Formulate Questions - Have an idea of what you are looking for.
2. Look at the top and bottom of your data -> head & tail will let you know that data was collected in a consistent fashion. If it's date ordered, it'll give you a range
3. Calculate Summary Statistics -> For very large data, use a subsample
4. Make plots of distributions
5. Form an opinion about how well a particular model should work.

MOEs & MOPs

When is a Model Sufficiently 'Good'

- Often times, defining the metrics that are used to measure the 'goodness' a model is the most crucial aspect of an analysis
- MOE - Measure of Effectiveness. How well does the model do the thing it is constructed to do. Some times this is quite straightforward, but for many problems this is not the case
- MOP - Measure of Performance. What does the model require to run well. Do you need a 300 core spark cluster to run it? Do you need 50,000 labeled training examples to build it? Do you need a quad GPU to run it in production?
- The time for building MOEs and MOPs are before your initial model is ever built!!!!!

Communicate Uncertainty

Flavors of Uncertainty

Fit Uncertainty

- Parameter Uncertainty - error from model parameters (i.e. parameter fit error)
- Parametric Variability - difference between features on which the data was built, and features that the set in question have (e.g. a model built on Texas data would not apply to Oregon)

Model Uncertainty

- Structural Uncertainty - fitting a round peg in a square hole
- Algorythmic Uncertainty - the limit of accuracy for a particular model

Sample Uncertainty

- Experimental Uncertainty - a particular sample will vary from the universal truth
- Interpolation Uncertainty - future events change the true value of measurements

Quantifying Uncertainty Takes Effort

My Ph.D. ->

THE STUDY OF THE Z BOSON TRANSVERSE MOMENTUM SPECTRUM RECORDED BY THE COMPACT MUON SOLENOID FROM 2010 LARGE HADRON COLLIDER DATA

TOTAL LENGTH	SPECIFIC TO THE PT MEASUREMENT	PAGES DEVOTED TO MEASURING PT	PAGES DEVOTED TO MEASURING ERROR
181	35	1	34

"The first principle is that you must not fool yourself – and you are the easiest person to fool." R. Feynman

Treat Models with Scrutiny

- 99% accurate models on predictive analytics are *extremely* rare for all but the most mundane tasks
- Have an idea of how accurate your model should be. More often than not, significant improvements in accuracy are not good signs…
- If unexpectedly good results are achieved, the epicycle needs to be revisited so that these results can be explained

Session Objective

- Annunciate the basic steps in the data science workflow
- List consequences of doing data science mindlessly