Predictive Modeling Blueprint

Casey Quinn and Ahmed Alawami

March 29, 2016

Predictive Modeling Approach: Overview

1. **Define the Problem**
2. **Understand the Data: Data Types**
3. **Understand the Data: Missing Values**
4. **Model Tuning Metric**
5. **Model Evaluations and Tuning**
6. **Github Collaboration**
7. **General Collaboration**

Define the Problem

Model Complexity Constraints

- Offline, Time-averaged, "Instantaneous" Implementation
  - Computation Resources
  - Computation Time
  - Data Handling

Model Accuracy and Precision Requirements

- "Exact" vs "Rough" Predictions

Understand the Data: Data Types

Continuous

- Independent Observations?
  - Check for Autocorrelation
- Exact vs. Classification Ranges

Date/Time

- Numeric
- Factor

Classification

- Reduce factor lists as needed
- Variable combinations if many variables

Understand the Data: Missing Values

Find Them

- VIM: aggr function

Deal with Them

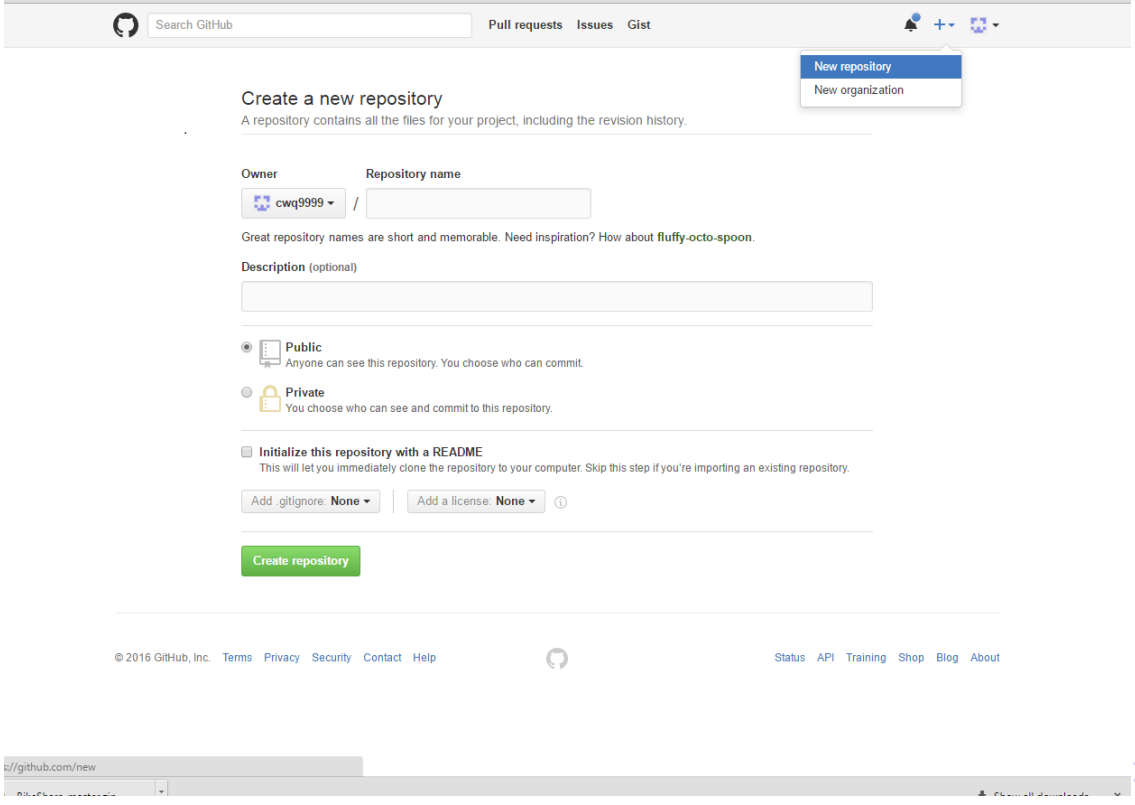
- **Removal**
  - Entire Variable
  - Individual Observations
- **Imputation**
  - VIM package
  - mice package
- **Set as unique classification**
- **Leave as is**
- **Handling test/prediction missing data**

Model Tuning Metric

Model Evaluations and Tuning

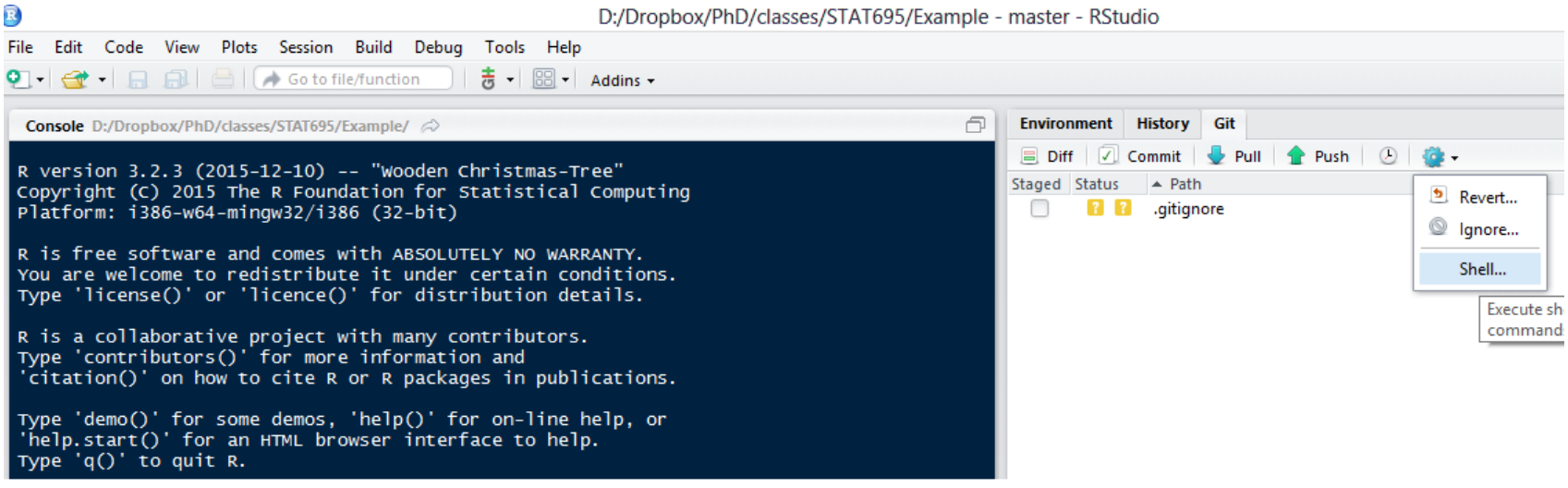
Github Collaboration

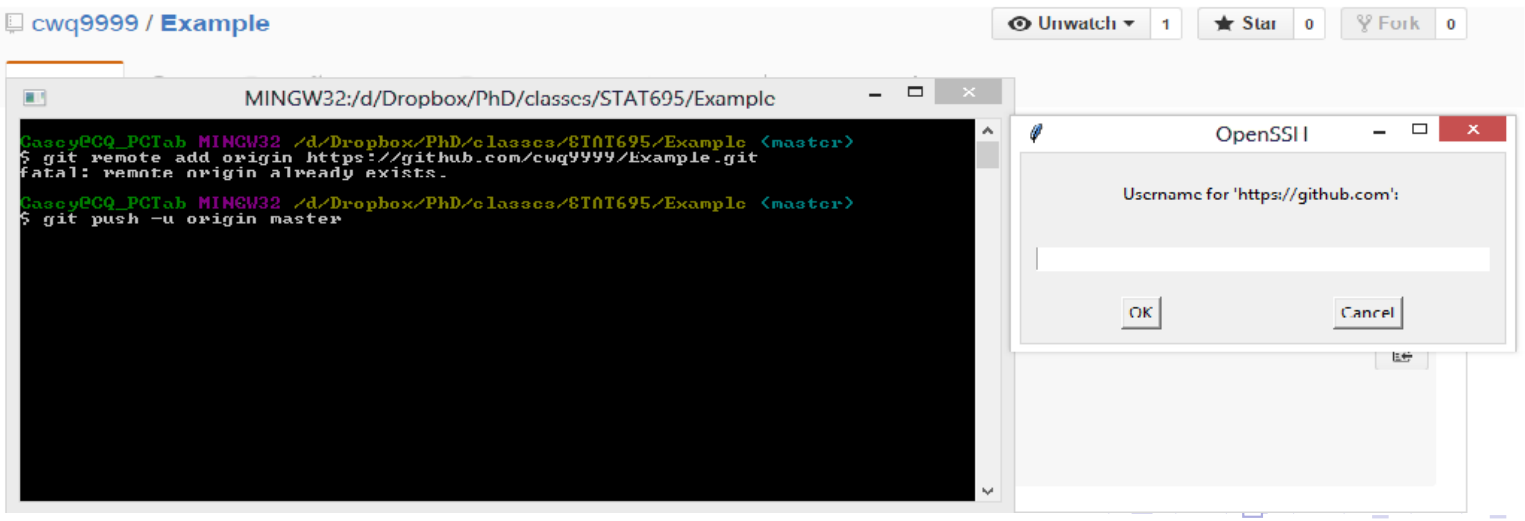
1. Chose a team leader.
2. Team leader creates a local folder with a descriptive name for the project.
3. Team leader creates a local RProject with the same name as the folder.
4. Team leader creates a Github repository with the same name as the RProject (Don't initialize with README).



Github Collaboration (cont.)

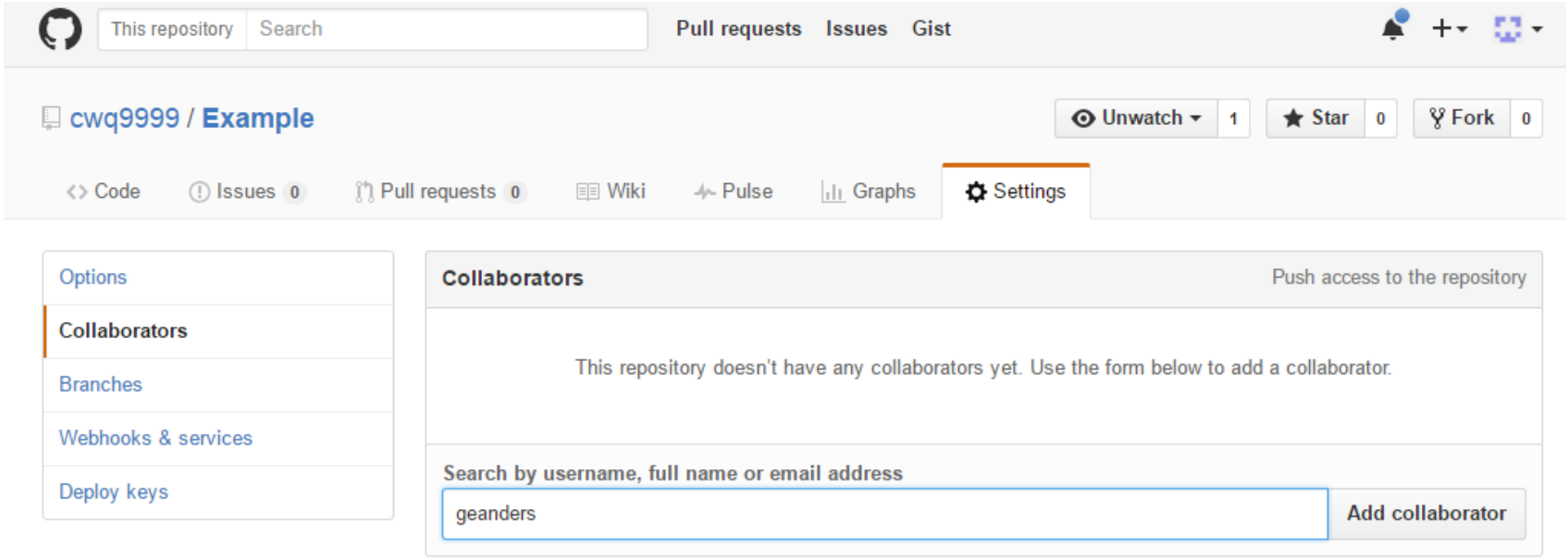
5. Push an existing repository from RStudio Shell.
   - git remote add origin https://github.com/cwq9999/Example.git
   - item git push -u origin master





Github Collaboration (cont.)

6. Add team members as collaborators.
   - Collaborators can use Pull/Push buttons under the Git tab.
   - No need to fork the repo or request pull requests.



General Collaboration: Team Efficiency (My two cents)

1. Select a team leader
2. Quick Individual Exploratory Data Analysis
3. Initial Exploratory Data Analysis Discussion(s)
   - Define Training/Test Datasets
   - Determine Outcome Requirements
   - Determine Metric(s)
4. Divide into Specialty Groups
   - General Linear Models
   - Trees
   - Support Vector Machines
   - etc.
5. Present and Compare Models
6. Select Model Type and Refine as a Group or Subgroups

General Collaboration: Require vs Library

General Collaboration: Relative vs Absolute Pathnames

General Collaboration: RProjects and Dropbox

General Collaboration: Operationg Systems