# Habits

*Riaz Arbi*

*9 Feb 2018*

## A Data Scientific Approach to Equity Backtesting Research

**Masters in Data Science (STA????W) Research Proposal**

**University of Cape Town**

**Author: Riaz J Arbi**
**Supervisor: Associate Professor Tim Gebbie**

---

**Abstract**

Contemporary research into the cross-sectional variation in stock returns is fraught with replication challenges. This makes it difficult to validate important results and hampers the advance of knowledge in the field. This project addresses some of these challenges by fitting the backtesting research process to a data scientific workflow. By making extensive use of standard open source statistical programming languages the authors provide an environment in which total replication is trivial and common statistical errors are avoided by default.

**Keywords**

backtest, data workflow, overfitting, historical simulation, replication

---

## 1. Hypothesis

- There are major replicability issues surrounding academic research into the cross-sectional variation in stock returns. These issues can be largely mitigated by adhering to a data-scientific approach to data analysis.
- Robustness of results in the field of stock return research can be can be significantly improved by accounting for the risk of overfitting as the number of trials increases.

## 2. Literature Review

**note to supervisor** I'll flesh this out. Basically it'll be the argument we know and love about how the standards of reproducibility in science are not adhered to in the field and because of that we can't really admit much of it to the body of knowledge. I will rely heavily on the arguments made in the initial sections of Financial Charlasanism and Replicating Anomalies.

Three sections of the lit review will be -

1. A survey of several well-known backtests and the documentation of their workflow

- Fama (1973)
- Fama (1992)
- Daniel and Titman (1997)
- Hou (2017)

2. Reproducibility in science in general and the field in particular

- Setting the Default to Reproducible
- Why Most Published Research Findings Are False
- Star Wars: The Empirics Strike Back
- Evaluating trading strategies
- Replicating anomalies
- A manifesto for reproducible science

3. Statistical errors in backtests

- The probability of backtest overfitting
- Pseudo-mathematics
- Replicating anomalies

The lit review will be between 2 paragraphs and 1 page long.

## 3. Aims and Objectives

This project is firmly rooted in the meta of finance research. The objective is not to validate whether particular anomalies in the cross-sectional variation of stock returns exists. Rather, it is to outline and implement a system wherein researchers can investigate these questions in a statistically rigorous manner.

1. Survey of current academic backtest methods (see the github dissertation repository)

- A critique of the challenges around replicability because of lack of documentation.

- A critique of the challenges around validity because of poor statistical methods (IS/OOS).
- Discussion on how these challenges can be mitigated using standard data science tools.

2. Documentation of an working demonstration system that mitigates these challenges (see the github code repository).

- Source code along with README documentation of every phase of the project will be released to a public repository under and Apache 2.0 license.

3. An original replication case study which makes use of the demonstration system to replicate a widely cited academic paper in the field (see the github replication example).

## 4. Data Requirements Specification

All stages of the data collection and transformation will be transparently documented and source code will be made avaialble on a hosted repository.

All raw data wll be programmatically extracted from a Bloomberg Terminal using the Excel Add-In. Validation of this data will be conducted by randomly selecting ten financial reports from the universe and cross-checking the contents of those financials against the raw data.

All data cleaning and interpolation will be done using the dplyr and tidyr packages in the R statistical computing language.

## 5. Systems Requirements Specification

**Hardware Requirements**

```
- A computer running an x86 processor
- 50gb of hard drive space
- At least 8gb or RAM
```

**Software Requirements and Packages Used**

```
- Access to a Bloomberg Terminal with Excel installed (does not have ot be the development machine)
- A research machine running
    - Ubuntu 16.04 LTS
    - Jupyter Server
    - RStudio Server
    - Python 3.5
    - R 3.4.3
    - Nextcloud 12 for transferring data from the Bloomberg Terminal to the research machine
```

**Software Development Framework, configuration control and version control**

The primary objective of this dissertation is the creation of a script-based workflow that improves the equity backtesting research process. The complete codebase, as well as the dissertation and demonstration case will be made avaialble on a GitHub repository. The code will be released under an Apache 2.0 license.

The data will not be released due to data vendor licensing constraints. However, macro-enabled Excel workbooks that programmatically query the Bloomberg Excel Add-In for relevant data will be made avaialable in the public repository.

Version control of all project deliverables will be managed using the Git version control system. Commits will be pushed regularly to the publically avaialable GiHub repository to ensure timeous backups of the

## 6. Project Milestone Deliverables

| Date | Milestone | Code Reference Point | Status |
|---|---|---|---|
| October 2017 | Set up server with necessary dependencies | https://github.com/riazarbi/serversetup | Complete |
| November 2017 | Build Bloomberg Excel VBA Workbook to scrape data | https://github.com/riazarbi/backtest_workflow/tree/master/2_data_collection | Complete |

| Date | Milestone | Code Reference Point | Status |
|---|---|---|---|
| January 2018 | Scrape Bloomberg terminal for data | https://github.com/riazarbi/backtest_workflow/tree/master/2_data_collection | Complete |
| January 2018 | Merge raw files into single csv files | https://github.com/riazarbi/backtest_workflow/tree/master/3_data_cleaning | Complete |
| February 2018 | Clean, join and interpolate data | https://github.com/riazarbi/backtest_workflow/tree/master/3_data_cleaning | In progress |
| February 2018 | Transform raw data into Sqlite file | https://github.com/riazarbi/backtest_workflow/tree/master/3_data_cleaning | In progress |

| Date | Milestone | Code Reference Point | Status |
|------|-----------|----------------------|--------|
| March 2018 | Build out code to control for backtesting biases (look-ahead, survivorship etc) | https://github.com/riazarbi/backtest_workflow/tree/master/3_data_cleaning | Not Started |
| April 2018 | Perform a case study backtest | https://github.com/riazarbi/dissertation | Not Started |
| May 2018 | Debug, refactor, refine | . . . | Not Started |
| June 2018 | Add additional data sources: iNet, Datastream | | Not Started |
| July 2018 | Replicate case study backtest on alternative data and compare differential results | | Not Started |

| Date | Milestone | Code Reference Point | Status |
|---|---|---|---|
| August 2018 | Create second back-test and document workflow steps and benchmark timing | | Not Started |

## 7. References

Will fill out references from thelit review in proper fomat here.