

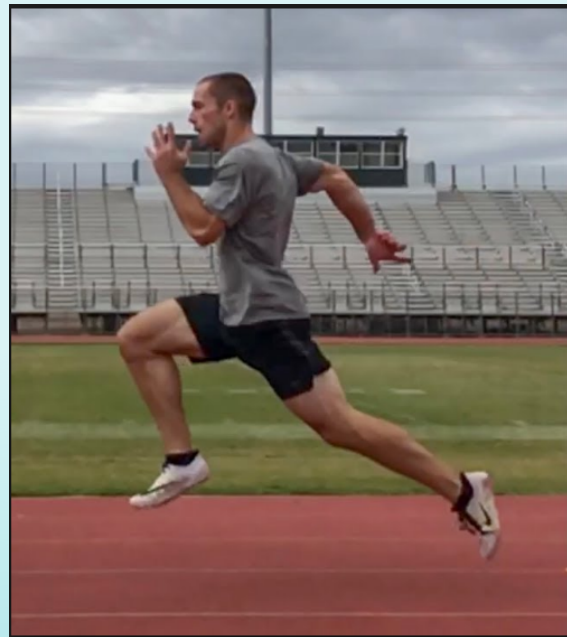


# Reproducibility Fast Track: Sprint, Leap of Faith, Hurdles, and Long Distance

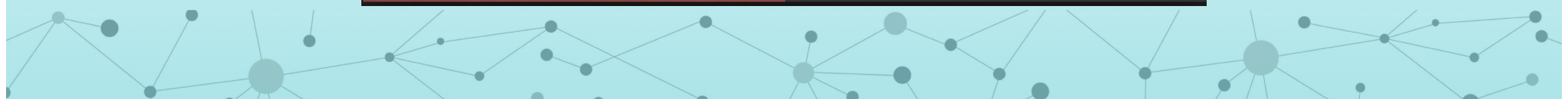
Janet Heslop  
CISER



# Preparation for the Sprint

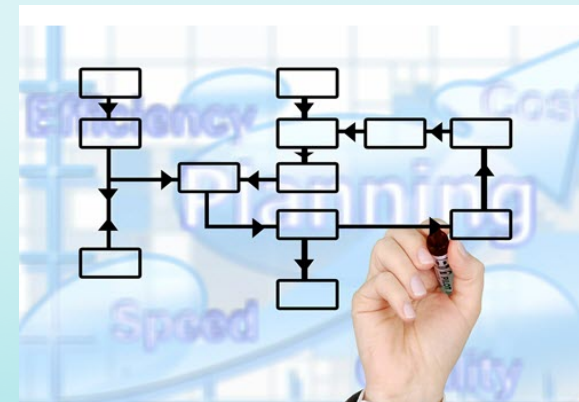


**SPEED  
WORKOUTS  
FOR  
SPRINTERS**



# R-squared Business Process

- Product and Services Description
- Scope
- Risks, Assumptions, Limitations, and Constraints
- Marketing Plan
- Benefits
- SWOT
- Executive Summary



## Warmup Exercise



- Are we clear about how R-squared will create value to the researcher?
- Can we convey three to six capabilities that R-squared uniquely provides?
- Do we know the “sweet spot”?
- Can everyone on the team clearly articulate R-squared?





# Product and Services Description

- What is the exact service we are providing?
  - CISER's Results Reproduction (R-squared) service **reproduces the computational results of a research study** prior to publication (or at check-points in the research), and make its **reproduction materials independently and easily understandable for reuse** and/or ready for publication submission.
- What does the service provide/solve/achieve?
  - CISER will **guide the researcher through the reproduction steps**, which requires access to the code, data, and supplementary documentation. **All variances from the original are reviewed with the researcher** with suggestions for correction. The revised materials from the researcher are again reviewed for full reproducibility. The **end result is a publication-ready package of reproduction materials** that has been verified to reproduce the study's results.





## Scope

- Value Creation
  - Who are we serving?
  - What are we offering?
- Value Delivery
  - How will we deliver service?
  - What are we able to achieve together?
- Value Capture
  - What are the benefits?



# Risks, Assumptions, Limitations, and Constraints





## Risks:

- As the service scales, quality control measures will need to be monitored closely
- Researcher did not keep good records on creation of final dataset
- Non-disclosure agreements need to be signed





## Assumptions:

- The data are clean and error free when provided to the R-squared team
- Ability to hire qualified and reliable staff to scale the program





## Limitations:

- Research provided is self-reported by client
- Use of known software applications only
- R-squared staff do not review methodology
- R-squared staff will need to be added to the restricted-data contracts





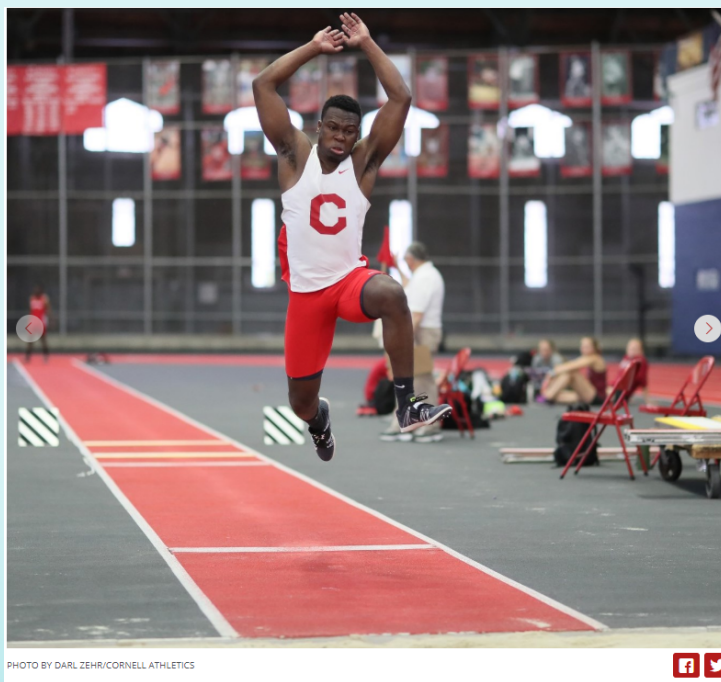


## Constraints:

- System environment, including OS, capacity and storage; client may have to provide access to their environment
- Turnaround time is dependent on timely researcher feedback



Leap.....



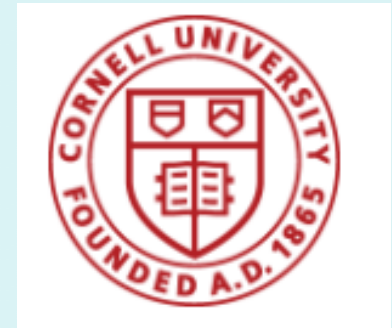
.....Of  
Faith





## Phase I – Internal to Cornell

- Features for Inclusion
- Market Plan Based on Market Positioning
- Opportunity Calculations
- Required Tools and Resources
- Budget Requirements and Forecasting





## Features for Inclusion

- Computationally reproducible scientific paper based on the figures self-reported by the researcher -- code, data, and documentation.
  - Thesis/dissertation prior to final electronic submission
  - Prior to initial journal submission (or any stage along the way)
  - Revise and resubmit scientific paper
  - Post-publication of scientific paper





## Define the Market to Create a Marketing Plan

- Market Research
  - Previous CISER Pilots
  - Market Postion
  - Opportunity Calculations



# Market Position

- Faculty
- Research staff
- Graduate students
- [Journals]
- [Publishers]





# Opportunity Calculations



- MPS, Research Master's, and Doctoral, Fall 2018: 4,822
- Conferred degrees, 2018: 1,625
- Faculty, Fall 2018: 1,875
  
- Original goal: 25% of conferred degrees and faculty publications

# Budget Requirements and Forecasting

- Based on the numbers presented....



- Today, we are currently not charging. 😊



# Opportunity Calculations -- Phase I - Actual

- FY18 R-squared projects: 5
- FY19 R-squared projects: 20
- FY20 R-squared projects: already some in the queue
  
- FY19 R-squared trainings, pitches, seminars: 18
- FY19 R-squared trainings, etc. attendance: 123



# Required Tools and Resources

- Advertising
  - Up-to-date web pages; Handouts;
  - Workshops; Seminars; Guest lectures
- Fresh compute server at the start of each project
- Archival storage
- Qualified staff
- Funding for all of the above



## Benefits

- “Researchers are taking reproducibility seriously and looking for ways to improve the credibility and transparency of their claims.” ([Ball, 2018](#))
- Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined. ([NIH,2003](#))
- Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. ([NSF, 2019](#))



Original

**S**

- Ease of entry into market with minimal startup costs
- Successful and proven experience
- Recognized need at the university level
- Addresses university research concerns
- DOI available at time of publication for inline referencing
- Meet funding agency requirements for data management
- CISER Data Archive is home for curating data

**W**

- Currently, only one staff member with knowledge to sustain service
- Service is dependent on researcher keeping good records of how they prepared their dataset(s)
- Researchers lack an understanding about the importance and benefits of reproducible work
- Research domain knowledge limited to social science
- Getting authority to access restricted-data
- Turnaround time based on researcher and journal deadlines

**O**

- Offering campus a much-needed service
- Increasing CISER visibility
- Increasing CISER income potential
- Compliance with Transparency and Openness Promotion (TOP) Guidelines at Level 3 and Recommendation of the Reproducibility Enhancement Principles

**T**

- Quality control of results, as the service scales
- Ability to hire qualified staff to keep up with capacity
- Balancing growing R-squared services and funding staffing
- Legal framework around licensing, copyright and citations





Actual

**S**

- Ease of entry into market with minimal startup costs
- Successful and proven experience
- Recognized need at the university level
- Addresses university research concerns
- DOI available at time of publication for inline referencing
- Meet funding agency requirements for data management
- CISER Data Archive is home for curating data

**W**

- ~~Currently, only one staff member with knowledge to sustain service~~
- ~~Service is dependent on researcher keeping good records of how they prepared their dataset(s)~~
- Researchers lack an understanding about the importance and benefits of reproducible work
- Research domain knowledge limited to social science
- Getting authority to access restricted-data
- Turnaround time based on researcher and journal deadlines

**O**

- Repeat clients
- Credentialed Certification
- Offering campus a much-needed service
- Increasing CISER visibility
- Increasing CISER income potential
- Compliance with Transparency and Openness Promotion (TOP) Guidelines at Level 3 and Recommendation of the Reproducibility Enhancement Principles

**T**

- Quality control of results, as the service scales
- ~~Ability to hire qualified staff to keep up with capacity~~
- Balancing growing R-squared services and funding staffing
- Legal framework around licensing, copyright and citations





# Executive Summary

Reproduction of results is a core requirement of the scientific method. [Cornell Institute for Social and Economic Research](#) (CISER) has developed a service, [Results Reproduction](#) (R-squared), to [computationally reproduce the results](#) of a research study. In addition to reproducing the results, CISER also documents the research to make it [easy and independently understandable for reuse and/or ready for publication submission](#). A pilot of the service has proven the need at Cornell University, with 100% of the publications needing revising, even if marginally. Marketing the service in the social science domain is underway, along with pilots in additional domains. Enthusiasm for participation in the R-squared service is encouraging, with [support at the Vice Provost and Deanship levels](#).



Hurdles  
and  
Hoops....



# Conquering Challenges Along the Way

- Build it and they will come
- Approvals to create a faculty list and email broadly
- Getting the word out
- Defining expectations
- Funding













# THANK YOU

Feedback welcome and encouraged!

**Janet Heslop**

[CISER@Cornell.edu](mailto:CISER@Cornell.edu)

