Please join the Research IT Reading Group for the following presentation:
The Berkeley Institute of Data Science program titled “Data Science Collective” provides opportunities for groups of student data scientists to work on real world problems. Campus Strategic Sourcing, a unit within our Supply Chain Management department, recently concluded a project aimed at better understanding the campus’s spend patterns. This project included a diverse set of 5 data scientists from campus and campus administration and resulted in some insights and areas of future study. More importantly, it provided opportunity for students to participate in the cost cutting activities within Berkeley and was a fun, engaging project.

Presenters:
- Andrew Clark, Strategic Sourcing Director
- Anthony Suen, Data Science Collective (DSC) coordinator and ISchool grad student
- Chris Holdgraf, DSC project lead and Neuroscience grad student

When: Thursday, May 21 from noon - 1pm
Where: 200C Warren Hall, 2195 Hearst St (see building access instructions on parent page).
Event format: The reading group is a brown bag lunch (bring your own) with a short ~20 min talk followed by ~40 min group discussion.

Please review the following in advance of the 5/21 meeting:
================================
Data Science Faire Poster provides a 1-page overview of the collaborative project between the Data Science Collective students and the UCB/UCSF Strategic Sourcing group.
Mid-semester video presentation for the UC Strategic Sourcing.
Final presentation slide deck.
Project code repository.

================================

Presenting:
Anthony Suen, Data Science Collective (DSC) coordinator @ BIDS, and ISchool grad student
[planned to attend/present but had to cancel: Andrew Clark, Strategic Sourcing Director]

Attending:
Aaron Culich, Research IT
Adam Fuchs, IST DB
Aron Roberts, Research IT
Camille Villa, Research IT
David Fulmer, Social Welfare
Greg Kurtzer, LBNL/Research IT
Indu Tandon, Human Resources
James McCarthy, SSL
John Lowe, Research IT
Mark Chiang, EDW
Oliver Heyer, ETS
Patrick Schmitz, Research IT
Ryan Lovett, Statistics
Scott Peterson, Library
Steve Masover, Research IT
Steven Carrier, Education
Susan Grand, DLab

Anthony Suen intro/presentation
================================
DSC is about bringing hands-on data science analysis for students to BIDS, and pairing that with pent up demand on the administrative side of UCB

~5 projects/semester; examples: uSeeData (viz for research), UC Strategic Sourcing (see reading materials for this meeting), Underclub (marketing analytics), DeStress (cognitive science, emotion derived from gigabytes of blog data), High Frequency Trading
Q & A

To the question of keeping administrative staff, faculty member/researcher involved in a project such as UC Strategic Sourcing, AS describes some steps:

=> designate a project leader, which takes weight off admin/faculty and gives a student project management experience
=> how discoveries in data are turned into appropriate actions or changes (in an administrative unit, for example) is not well worked out
=> tend to make project teams of ~4; needs to include project management aspect (lead/person) on the team

Q: How does DSC choose projects:

=> not a formal process/criteria established so far
=> need the data to be made available by the person proposing a project
=> helps for the person proposing the project to present, make a compelling case, present a cutting-edge element that will attract participation

Could have made the UC Strategic Sourcing project even better if someone who is expert in Industrial Relations research had been brought into the project early on to help set direction and identify new or interesting ways to approach the data analysis.

Q: How does a data analysis project start? When you're not even sure whether data will help solve any of several problems that may have been identified?

AS: Before we start cleaning data, we conduct a white board session exploring potential questions filtered against skills and interests of the team. Asking the questions is actually more difficult than cleaning/analyzing the data. There's a marketing/packaging aspect to attracting students, and setting that up might start with the project proposer meeting with AS or other DSC staff for an intake kind of session.

Q: How can potential partners prepare?

AS: Prepare to answer/provide:

=> What is core problem?
=> Why is it important?
=> Why can't you just do the project yourself (i.e., what are obstacles that require formation of a DSC team)?
=> How big is the data set? How clean/dirty?
=> Be able to speak to potential privacy or proprietary data concerns / restrictions
=> Can you meet for an intake session?
=> Can you present to potential team members?
=> Can you commit to regular check-ins with (and encouragement of) the team?

Some projects are ongoing, where people with an interest in a topic/subject area can join, broaden the project scope in terms of clients & data sets

Q: How are tools / toolkit chosen for the DSC's projects?

AS: If a client is tech savvy, s/he might drive that. So there's a certain project-driven aspect. In general and to the degree we can encourage: Python, IPython, R; github for a code repo.

Q: On project management training. What are you observing, what's needed?

AS: Can provide some guidance to fledgling PMs. How to manage team members that are not fitting in well or pulling their weight. How to be responsible for understanding both client needs and team needs/abilities. Regularity of meeting sessions. Delegating responsibilities. Identifying training needs (if a team member needs to come up to speed in using a technology.)

What we have definitely found: PM is what everybody lacks!

Q: Will DSC contribute to automation of data analysis or pull against that?

AS: Not convinced a computer is capable of asking the right questions. Hybrid will probably continue to win out.

Q: Anything from the projects that emerged as helping to define Data Science?

AS: Cleaning data, analysis, visualization, asking the right questions. All of these point to what Data Science is, but a broadly-accepted definition is not yet honed. One of the elements that is, perhaps, most interesting is how one takes results of an inquiry and turns it into something that impacts people or orgs based on the insights gained from the data.

Q: What's the range of scale of problem? What's the sweet spot for doing DSC projects in terms of resource requirements?

AS: A start-up project with about 1000 data points could be done using Excel. 80TB of data from stock market at the other end of the specrum. Sweet spot? Midsize projects, 300GB is probably about where "midsize" is now. "In the gigabytes range." Remains to be seen whether Berkeley will provide (or outside vendors) resources appropriate to the problems we'll take on.