Our next Research IT Reading Group topic will be: Analytics on UC strategic sourcing data with IPython and BCE

**When:** Thursday, October 23rd from noon - 1pm  
**Where:** 200C Warren Hall, 2195 Hearst St (see building access instructions below).  
**Event format:** The reading group is a brown bag lunch (bring your own) with a short ~20min talk followed by ~40min group discussion.

The presentation and discussion will be facilitated by Andrew Clark and Alexis Perez from UC Berkeley/UCSF Strategic Sourcing.

Andrew Writes writes: “Supply Chain Management - Strategic Sourcing reduces the time and money our campus spends on goods and services, so that our campus clients can use their time and money on the core mission of UC. This group of 8 uses data analytics, aggressive negotiation techniques, and partnerships with campus stakeholders to drive cost down and quality up. In the past three years, SCM-Strategic Sourcing has documented savings exceeding $28M for UC Berkeley (and about that for UCSF).

Four years ago, SCM-Strategic Sourcing was a fairly typical administrative unit in terms of data capability. They relied on IT as their data provider and used Excel as the sole analysis tool. In those dark ages, analysis was typically a 1-time ad-hoc effort without any hope of reproducibility, replication, or sharing between analysts. However, that all changed the day we were asked to analyze a data set larger than would comfortably fit in Excel. After a brief review of “R vs. Python” blogs and with the help of some early Coursera/Udacity classes, Andrew decided on Python as the team’s standard programming language and proceeded to upgrade the collective skillsets of the staff. Around the same time, SCM-Strategic Sourcing hired two new analysts and offered them only Python, Postgres and Bash for their analysis needs completely eschewing Excel. Fortunately, the ruse worked and the organization was transformed.

Today, Sourcing uses IPython Notebook, Pandas, Sci-Kit Learn and PostgreSQL to analyze UC’s Systemwide spend data, UC’s eProcurement catalog environments at 7 of campuses, and UC’s departmental spend patterns. The focus of the work shifted from “getting answers” to building “reproducible and auditable analytic pipelines” allowing the team to continually improve our analytic capabilities, reuse code and previous work, collaborate with analysts at both UCSF and Berkeley, and reproduce our analytic products for our stakeholders.

With a solid set of tools and the programming know how to be dangerous, SCM-Strategic Sourcing is working through the following challenges: They don’t have staff allocated for resource intensive ETL processes, their data is ever growing and “seemingly worthless” to other administrative groups, and we have an extreme impatience for high latency systems.

FLICTSecer-=--

Here’s the background material for review prior to our meeting:

Greg Macway (former Supply Chain Management - Strategic Sourcing Analyst) won a Berkeley campus Institutional Data Management and Governance award for most creative visualization for his BearBuy spend graph, which he created using Gephi.

Andrew G. Clark, UCSF Increases Consumer Value Through Optimal Vendor-Show Scheduling. Interfaces 41(4): 327-337 (2011) describes how a pre-merger UCSF Strategic Sourcing formulated a bipartite matching problem and solved it using binary integer programming to assign suppliers to supplier shows.

Alexis Perez’s SciPy 2014 presentation, Behind the Scenes of the University and Supplier Relationship, describes how Pandas and Python transformed a once tedious, time-consuming manual process into one that now takes only a few seconds to analyze supplier’s proposed price files and ensure the University is not paying more than contracted.

ATTENDING:

Aaron Culich, Joann Peterson UCSF/UCB, Mauricio Garzon, Chris Hoffman, Harrison Decker, Ron Sprouse, Bryan Hamlin, Aron Roberts, David Willson, David Greenbaum, John Lowe, Scott Peterson, Tim Dennis, Patrick Schmitz, Alexis Perez, Andrew Clark, Kerrie Hayes, Dav Clark, Steven Chan, Quinn Dombromski, Rick Jaffe (note-taker).

FRAMING QUESTIONS:
1. What benefits of scientific analytic tools would overcome the drawbacks of not using traditional business tools in your workflow?
2. How is “Scientific Analysis” different than “Business Analysis” and which tools are appropriate for each group?
3. Within higher-ed Administration, business units (like Supply Chain Management) are not proficient in data management, don’t know what ETL stands for, and generally don’t use data for strategic or tactical purposes beyond simple reporting. What changes that paradigm and how much money could be redirected to the core mission if our $2.1B enterprise became filled with data savants?

SLIDES:
UCSF SCM-BCE.pptx

NOTES:

Andrew asks: What if we had data savants in our administrations? How much money could we save if we streamlined our businesses?

Andrew presents slides

Who were we? Analytic journey over the past few years...

A-Team: a bit crazy, sometimes pretty, always strategic. Sometimes we have to be bad-asses, no way around it.

What we do?

• Structure and execute negotiations - prep campuses for successful negotiations.

• Run competitive bids and sign agreements

• Act as “bad guy” - to preserve timelines and chase savings.

• Own all items in BearBuy

• Destructive campaigns against things that don’t work

Org structure – see slide

2008 - using Excel; PeopleSoft at UCSF only. Dump from PeopleSoft, consume in Excel, two days of analysis. Exhausted. Didn’t use much of analysis. Couldn’t reproduce.

Fisher Scientific 800,000 items. Excel only accommodated 65,000.
2010 - Switched to VBA. Bad, but highly functional. Used into 2013. Standard inputs, outputs; standard process to transform the two. Difficult to learn.

2012 - Began experimenting with Python, R, SAS, BASH, PostgreSQL

2013 - Hired two analysts – Alexis was one, Sumanjit Mann, the other. (Sumanjit has since moved on.) Tricked them into using Python, BASH and PostgreSQL. Difficult at first.

BearBuy assessment due. Fast learning curve. Results surpassed expectations.

2014 - Postgres _ CitrisDB ; Python 2.7 via Anaconda Distro; Tableaux for reports; Git, Bash. All work done and saved in iPython notebook.

Recent prjects:

• BearBuy content analysis and cherry-picking contracts
  "What if we got the lowest price point on all BearBuy items?"

• Clustering UC departments by spend patterns

• Optimize ePro enablement decisions - more AP receipts v more POs v more catalog items. Broke gridlock between those stakeholders

Future:

• Hashing Procure to Pay (P2P) transactions and visualizing in D3.js. Develop fingerprints for all transactions. Identify metrics beyond satisfaction.

• Predicting the End of Blanket POs. Make bids in alignment with PO cycles (?)

John Lowe asks: What does a cluster of departments look like? What characteristics does it have?

Andrew opens iPython notebook: Data includes Spend Frequency (How often do you buy?) Narrowed down to five suppliers of interest. Avg monthly spend, PO, invoice features; spend dispersion among suppliers; sum of spend per supplier; median invoices per po (many items per po implies purchase from hosted catalog - good). Cluster 0 proved to be the point of focus

Bryan asks: In your workflow, how do you say 'this is the date range and the version of the data used in the notebook?'
Andrew describes how 30-minute conversation with Aaron helped him here: UCSF and UCB tech infrastructure completely different; BCE to the rescue. Analysis works across campus platforms. We've used Box to share data (csv file). Store iPython notebooks on Drive or Box, set the path to them. Versioning on tools has been a challenge – ie, different versions of Postgres.

Chris Hoffman asks: How does this relate to CalAnswers/EDW initiative? Is that the army version of the guerilla work you’re doing?

Andrew: We're doing exploratory stuff; begin with a premise. Enterprise wide effort must begin with better knowledge about business processes. (paraphrasing)

Patrick asks: RE: understanding patterns on campus - how to overlay taxonomy of products (e.g., all bioinformatics tools, money spent on them) - roll ups to aggregated data. How hard would it be to overlay a semantic model on the data?

Andrew: We still don't have our arms around that question. Suppliers generally represent commodity areas. UCOP has a data set / tool called Spend Radar (MySQL with CQlikview front-end) has some classification taxonomy. Fairly manual. New things go into unclassified bucket. We think of trying some machine learning techniques, but we don't have the data. Clustering methods wouldn't adhere to questions that humans generally ask.

We have new spend data every minute. No way to find all phones. We could find iPhone by text search, but that wouldn't bring up Samsung or dial phone.

John Lowe: There are product taxonomies.

Patrick: You get decent taxonomies for certain areas, but not generally across categories. Ontologies represent points of view – don’t work for all users.

Andrew - We’ve taken a people approach. IT purchasers talk to IT spenders. Help them. At UCSF, IT spend flows through one person. She requested us to manage those contracts, improve upon them. Since we now own them, we don't need to classify them.

Discussion of issues related to creating taxonomies, value of long tail, feature engineering as machine-learning equivalent to taxonomy engineering. Are we looking primarily for important or expensive items?

David asks: You said “we're an anomaly but we're having lots of fun.” What other groups would benefit by taking the approach you have?

Andrew: All of them. We start from the standpoint that we don't know, it's ok to fail, we're going to do things, measure the results, go on from there. (Methods include “behavioral economics” – lead buyers to better choices.)

David: Doing that method across an entire administration would be expensive.

Andrew: Yes. Hiring is difficult.
Aaron: Training – where did it come from? How much did it cost?

Alexis: On the job: DLab, StackOverflow, FreeNode webchat, piece by piece learning

Andrew: I gave Alexis and Sumanjit problems to solve.

David: Interested in admin staff internship model?

Animated conversation follows.

Andrew: Making an action and watching the results ripple across campus is super exciting.


--- END OF OFFICIAL TIME ---

Andrew: We have one metric: savings. We are aggressive. We’ve already surpassed this campus’s annual goal ($12M; We’re at $20M already and piling it on). Our analyst group is therefore, in some ways, R&D.

We have flexibility. We’re not tied to cues, we don’t issue POs. We focus on contracts. If you’re buying something that’s $100G or more, contact Joanne or John Arbolino.

Aaron: What sort of historical data exists? For example, “In the past ten years, who has bought a cluster? When will they expire?”

Patrick: We do that. Requires a lot of domain knowledge.

Andrew: VMWare deal: went on for years - trying to understand demand. For all the cool, fancy things we’ve done, our business process lets us down. Purchases go through all system, record shows “see quote.” Difficult to analyze.

John Lowe: You mentioned techniques for detecting power buyers. Can you detect people who fail to buy at the best price? Or buy from their brother-in-law?

Andrew: Haven’t found any malfeasance in our data yet.

Maurizio: Can the data be used to create a recommender matrix?

Andrew: That would come in with respect to shopping. That often happens outside of BearBuy.

Patrick: Can I find the power buyers for Chemistry?
David Willson, Andrew: The person making the buy in BearBuy is not often the person making the shopping decision.

Aaron: Sensitive data?

Andrew: Monthly click-through to certify "I will not put protected health information into the system." Doesn't guarantee that there's no PHI in the system. (There are tools to clean data of PHI.) The other type of sensitive data: suppliers’ prices.

Harrison: What were the difficulties in hiring? Job classification within UC system?

Andrew: Yes. We were lucky to get Alexis. From the job description, she wouldn't have known that the job was so analytical.

Alexis: I had worked at MIT, Stanford. I wanted to be at a similar institution.

Andrew: Our suppliers are beginning to pilfer our people. They pay more.

Discussion of skills needed - command line, Python, etc.

=== END OF NOTES ===