ISED RDU Data Integration Pilot Project Lessons Learned

PPX Symposium

May 2018

AKA Death by Bureaucracy

About the Project

Players

- Two universities
- One private sector company
- One non-profit volunteer
- ISED RDU
- ISED CIO

June 2017 to January 2018

Data Successfully Acquired

- CMC Microsystems client records
- IRAP
- Canadian Intellectual Property Office patent and industrial designs
- OSB and Corporations Canada open data
- ACOA, SADI, NSERC, SHRC, CIHI, and more…

In five months….
We identified and 'wrangled' multiple datasets, security cleared seven contractors thoroughly, set two contracts in play, cleaned 11 data sets and tried to use more, held 37 conference calls, e-mailed the CIO 139 times and continuously reached out across the department in search of higher quality datasets that might lend insight.

How the bureaucratic ride began…

One area of contracting was fine with our initial project model

Somewhere deep in corporate, however, someone else was not

We tried to make use of shared data, but there were no takers so we moved to Open Data

But it turned out that it was not so open after all. Complex XML formats were being released to non-profits for "free"

At the outset, we were told security was fine with Enhanced Reliability Clearance for all contractors

Somewhere deep in corporate, however, someone disagreed

As we course corrected and sought solutions, our colleagues in the CIO literally rescued the project on more than one occasion

But even in our CIO, skills, experience and expertise were still developing and all of this was not on the business plan

We had never articulated a business requirement

Service Blueprint: Death by Bureaucracy

Indicates project change and resource cost

	Physical Evidence	Revised project plan	Contracts	Security Forms (1)	Security Forms (2)	Security Forms (3)	Tokens and network accounts	Data	E-mails, phone calls, etc.	Work-around solutions
Contractor view	Decision to undertake project	Security protocols newly articulated	Contracts signed	Security Clearance Paperwork Complete	Additional Security Clearance Paperwork	Security Clearance Complete	Remote Access Accounts Established	Data cleaning and uploading	IT Issues, outages	Software requested
Line of interaction On Stage Contact	Meet with contracting & admin		Data wrangling		Knowledge gap identified		Begin project meetings	Data preparatory work	Help with data cleaning sought out	Help with software acquisition sought out
	Project plan laid out	New IT solution sought out		New IT solution found	Contingency plans made					
Line of visibility Back Stage Contact	Contracting authority advises	Contracting authority rejects	Assessment of existing prototype for new application				Remote access established		Standards identified	
	Contracting authority reviews									
Line of internal interaction Support processes (But for us, rescue processes)			CIO advises and makes recommendation	Prototype system resurrection	System maintenance and support		System rescue		Software installation	
				Resources reassigned	Loan project more powerful computers		New software assessment		Data conversion XML to CSV	

Lesson Learned #1:

Contracting is different for data science

- **The Rules as Previously Unstated:**

Data could not be sent to contractor firms without Enhanced Facility Screening – a PSPC process that requires several months to ensure servers in other organizations meet GoC requirements. This is not widely known or understood by contractors, especially the smaller, younger business we were dealing with…The professors had never heard of it either.

- **Collaborative Contracting:**

While a Gantt chart was shared, a project charter, sharable via the contracting process would have smoothed project management by keeping roles and responsibilities as well as timelines and inter-dependencies out in the open.

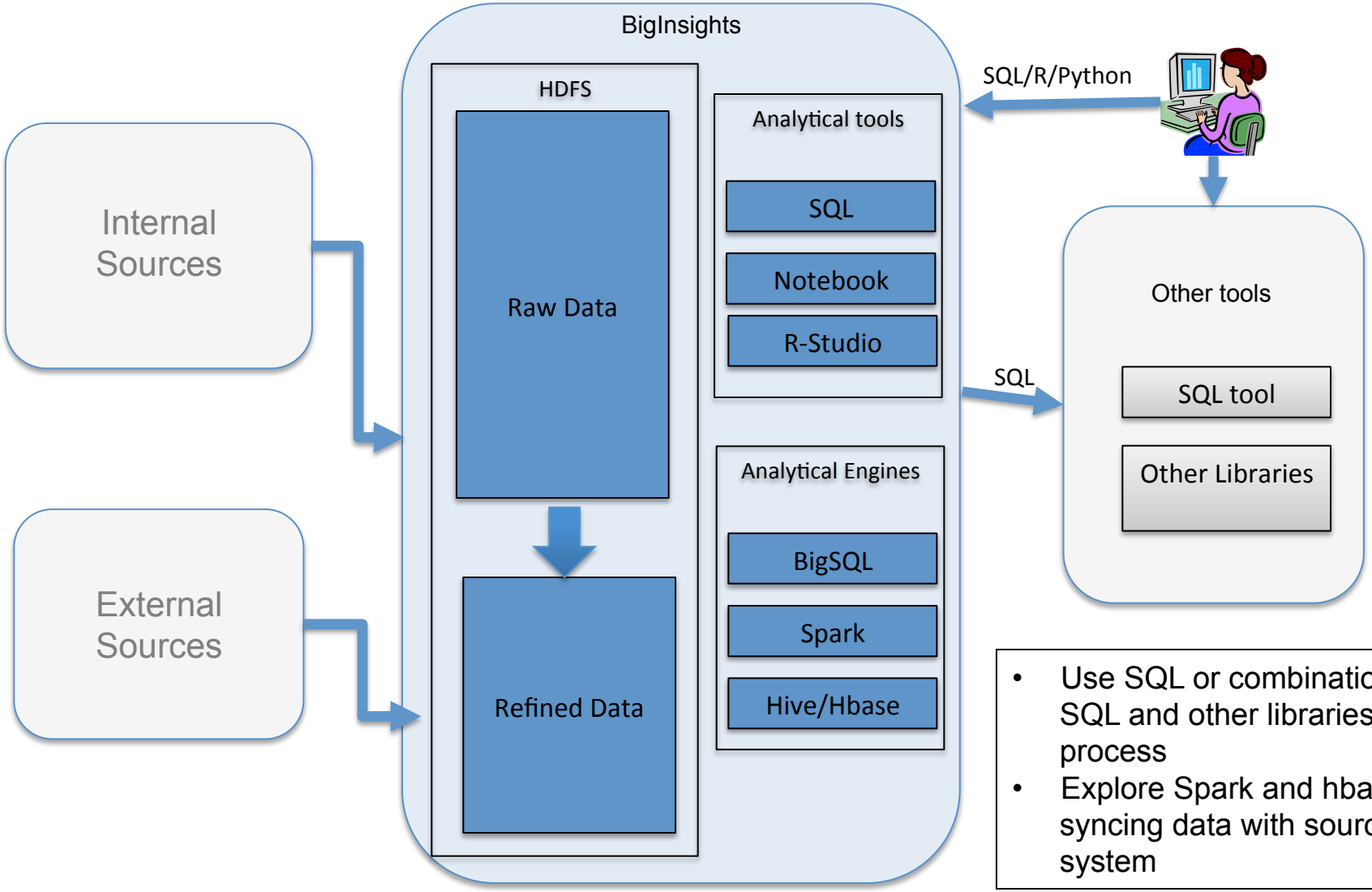
- **A Need for Greater Engagement:**

Several weeks into the contracts, we discovered later that we could have gotten academia for free…they were after access to data and longer term partnerships.

Insight:

We need to build stronger relationships with academia and private sector partners to facilitate data access and analysis and ensure they have secure servers as well as security cleared faculty and students to help us better understand our data. Universities and colleges both actively seek out data projects for their students.

ISED CIO: A Recycled Prototype to the Rescue



- Use SQL or combination of SQL and other libraries for ETL process
- Explore Spark and hbase for syncing data with source system

ISED CIO: A Recycled Prototype to the Rescue

Rescuing the project

Resource Re-allocation

Once You Offer Free Services…

Lesson Learned #2:

At ISED, data scientists do not have the tools they need to do their jobs

- **We Needed our CIO to Save the Day:**

ISED re-used an aging prototype Hadoop environment. While this allowed work to proceed, the environment was already out of date.

- **We Needed our CIO to Save the Day for Free:**

Being a prototype, both CIO and contractor hours spent managing the system and dealing with technical issues overwhelmed the project on several occasions. Remote access proved cumbersome for this work.

- **Our Tech Still Needed Improvement:**

There were challenges with respect to remote access ranging from technological limitations to quality assurance and collaboration (e.g. Skype is not our norm in government, but is the norm elsewhere).

- **And So Did Our Business Processes:**

New requests for software substantially slowed the process down. Contractors complained that things that took days or weeks would have taken minutes or seconds in their own companies.

Insight:
The tools we needed changed over the course of the project. The department needs a list of what is available quickly and at what cost. Computing power was also lacking. Where possible, we upgraded workstations, but there weren't enough on hand to go around…and this took more time…

Lesson Learned #3:
Data Scientists and the CIO need to work collaboratively without barriers

- **Responsiveness Matters:**

The capacity of the horizontal project team to move and adapt was key. Future initiatives will be more successful if they are built with joint teams that do not need to go through a help desk or queue to move forward.

- **The Future is Already Here:**

Both CS and EC groups can code, which can generate efficiencies in collaboration. In fact, our coop student worked in tandem with a CIO resource to decode some open data, posted in a complex XML format.

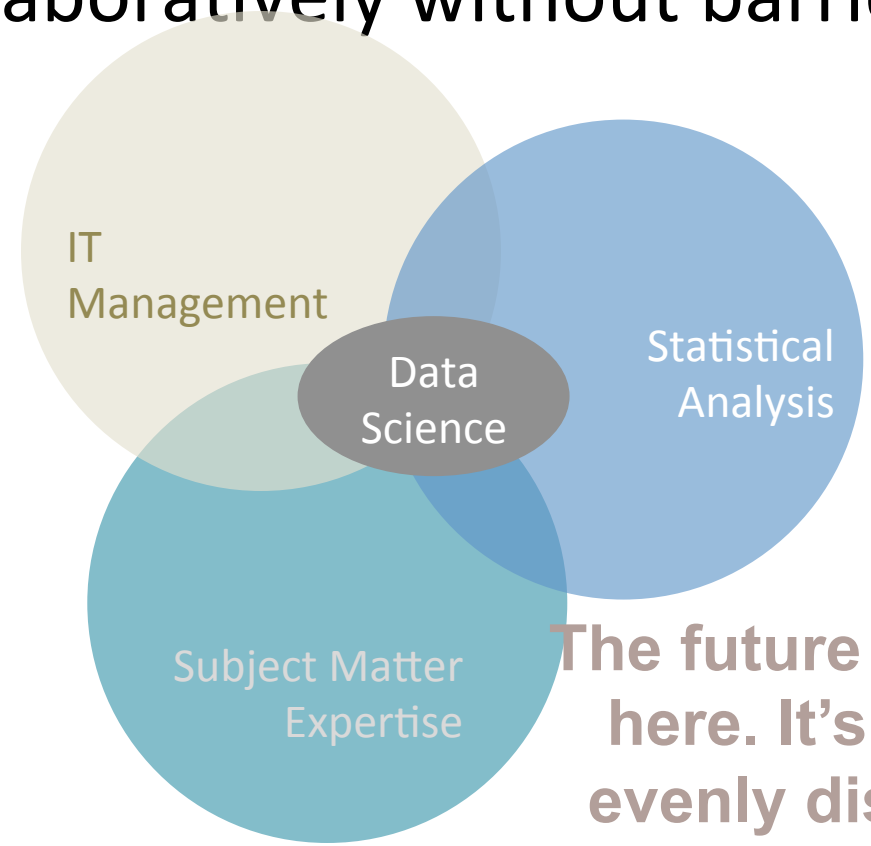
- **And Yet Our Software and Permissions are Not:**

Data science graduates use different tools than what ISED is accustomed to using: Linux, Python, Hive, Power BI etc. They do not like working in Windows. Contractors all stated that administrator-level rights on computers would have helped significantly.

Insight:
The project would have stalled without high levels of collaboration between the RDU, CIO and data scientists working on the project.
The CIO is not currently configured to provide this type of support on an ongoing basis to business units across the department.

Lesson Learned #3:

Data Scientists and the CIO need to work collaboratively without barriers



The future is already here. It's just not evenly distributed yet.

-William Gibson

Insight:
Data scientists are advanced IT clients and require close relationships with the CIO to do their work effectively.
There is also a convergence of skill sets in data science. True experts straddle the CS and EC groups.

Lesson Learned #4:

Data Scientists do not have the data they need to do their jobs

- **We asked for tax data …**
- **We asked for BR data…**
- **We asked for other data from programs and portfolio agencies…**
- **We went to Open Government and asked for data…**

- None of this data was included in our study. We were met with silence, negative legal interpretations and blank stares.

- We were also unable to use the project descriptions in the datasets we received because they were to scarcely populated. Unstructured data, such as what we store as PSFs, offer a rich dataset and should be explored going forward. **We have already had some success in the RDU.**

Insight:
The Government of Canada's open microdata is not necessarily accessible. Complex XML schema used by several ISED programs makes the use of the data prohibitive for those without advanced skill sets. This could be easily remedied.

Lesson Learned #4:

Data scientists do not have the data they need to do their jobs

Ultimately, we lost one of our universities because we just couldn't get the data in…and honestly, by that point, we weren't sure we could get them through security in time if we did.

Lesson Learned #5:

We need to evolve as data stewards

- **We Need to Better Understand Our Data:**

There are datasets in government still untapped that could be used as a basis for this work

- **We Need Standards:**

Without a measure of uniformity across our administrative datasets, they are of un-necessarily limited value

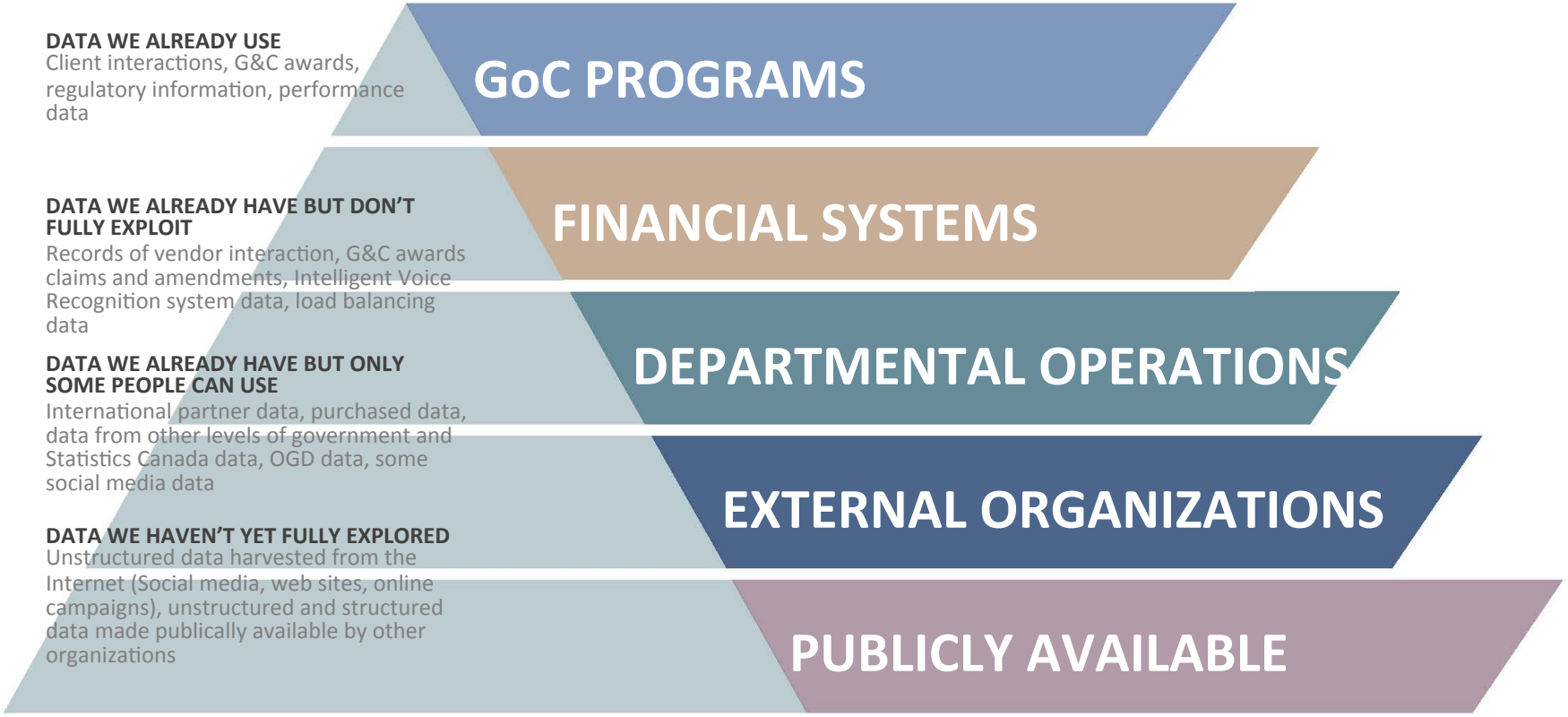
- **Then We Will Need to Manage Our Data**

Quality assurance, standardized concordance, data dictionaries and data storage formats really do all matter!!

Insight:
Cleaning data and loading data into the system was over 80% of the project work. Higher measures of accuracy will minimize the time and resources required for this work in the future.

Lesson Learned #4:

Data scientists do not have the data they need to do their jobs



Lesson Learned #6:

We Need to Shift Paradigms to Be Successful

- **Open Government Has Not Yet Reached its Full Potential:**

It was not possible to simply contact Open Government and ask for whatever micro data they had. It was found via hours of slogging though the online database of aggregate information.

- **Nous Sommes Différent!**

Not all Canadians (or software!) are equipped to deal with data that drifts between both our official languages. We need to consider enablers and accessibility with respect to open data. Accents, for example, are handled differently in different applications.

- **Convergence is Both a Challenge and a Solution:**

Subject matter experts need to be engaged in all phases – and not just data owners but programs and policy areas as well.

- **We Only Need a Rallying Point to Launch:**

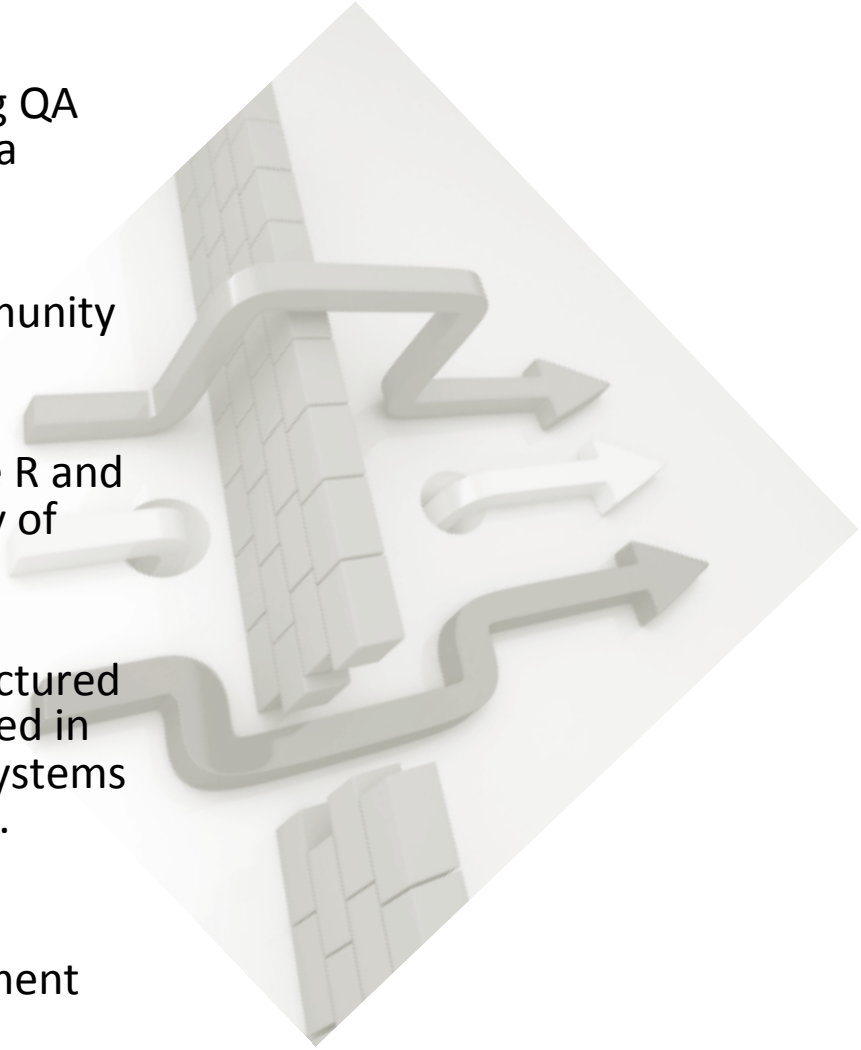
As data scientists across the organization have discovered the prototype Hadoop environment, they are experimenting and learning more about it and the free software that it leverages.

Insight:
Sharing across government departments with respect to experiences, skills and assets such as coding and data will be critical to success going forward.

…It is also time to start sharing data. Can we not be open by default?

Recommendations (AKA The Dream)

- Establish a team that does data asset R&D, sets standards, supports business lines in developing QA strategies and holds concordance tables for data integration
- Review how the CIO and the Data Science community interact and consider new collaboration models
- Formalize and provide additional support to the R and Python Users Group / Data Scientist Community of Practice
- Create an ISED data platform that liberates structured and unstructured internal data that could be used in analysis that is separate from other operating systems to allow for advanced analytical techniques (e.g. machine learning)
- Better package sharable data for Open Government and provide a data science-friendly portal



Some stuff to think about

We don’t typically assess the performance and efficiency of corporate functions

And yet they are critical to our projects being successful

Some stuff to think about

We have not even really identified all of our data

There are teras of unstructured data in our CIO

PSFs hold project descriptions, progress reports hold clues to project success we have not yet explored

Some stuff to think about

‘I thought we had the data and it was kind of disappointing. Now I know where you were all hiding it and it is actually pretty exciting. We can do a lot of interesting things with this, if your senior management is willing to invest.’

- Dr. Peter Taillon

Thank you

Contact Information