Data Driven Decision Making in Utah Government: Assessment for the Use of Big Data

Data Driven Decision Making Project

Volume 5: Process - Generating Value
September 2015

Submitted to:
Dave Fletcher
Chief Technology Officer
State of Utah
dfletcher@utah.gov

Point of Contact:
Frank Crichton
Project Manager
SAIC
frank.h.crichton@saic.com

Table of Contents

Executive Summary

Opportunities exist to utilize Utah's abundant data resources to improve the health and well-being of Utah's citizens and make services more cost-effective. This task paper offers two scenarios that demonstrate the value to be derived from using analytics. In addition, a general process model is presented that provides a roadmap for a broader implementation of data analytics across Utah government.

1 Introduction

This document is one in a series of five documents, describing an assessment of the state of the art in technology and skills for a cost-effective deployment of a Big Data solution. This Task 3 Technology Roadmap document describes the technology choices for standing up a big data environment, along with some indication of the relative pricing for the technology choices. The companion documents in this study are Task 1 Policy and Governance, Task 2 People Skills and Collaborations, Task 3 Technology Roadmap, and Task 4 Business Case.

2 Evidence-Based Management

Some years ago at the Gartner Symposium, an annual gathering of CIOs and senior technologists run by an independent technology assessment company, attendees were surprised to find the conference theme to be 'Balancing Cost, Risk and Growth'. Rather than a traditional technology focus, Gartner was subtly introducing the concept of evidence-based decision-making to many attendees who were used to relying on personal technology preferences. Attendees were told throughout the 5 days about the importance of analytics in managing IT as the economy began to rebound from the recession.

Figure 2-1 The Cycle of Data Science Facilitates Evidence-based Management

Today evidence-based decision-making is becoming widely accepted and States using these techniques to identify opportunities to improve services, undertake actions to get the most value out of limited resources and root out waste, fraud and abuse. To accomplish this, however, requires disciplined analysis of information to

produce actionable knowledge. The model in Figure 2-1, above, depicts the cycle whereby data is collected and curated, information is analyzed, and knowledge is acted upon based on evidences derived from the data.

2.1 Data Driven Organization

States like Utah have abundant data; the question is whether it is going to be used effectively. For years, the storage of data in most organizations has followed a pattern very reminiscent of the organization chart. All data is 'owned' by one organization for their purpose, use and benefit. Data is shared reluctantly, and then usually in response to a specific mandate or high visibility project. Furthermore, there is a tendency for "HiPPO"[1] decisions - decisions made by the "highest paid persons' opinion" rather than decisions based on evidence derived from the data.

To embrace evidence-based management, however, two behaviors must be developed: (1) data must be documented and shared, and (2) a core function of data stewardship and analysis (center of excellence or CoE) must be established to lead the way. When done properly, these two behaviors feed off of one another—data allows the CoE to conduct worthwhile analyses and the success of the CoE fosters a culture of data sharing and the accumulation of more data for analysis. Together, they pave the way for a snowball-effect whereby the identification of useful data is expanded, the value of data is recognized by the rank and file, and analytics becomes a norm.

That is a data-driven organization; it could be Utah.

2.2 Data Science - Agile Analytics

In developing the value proposition for data science, successful organizations such as the military and North Carolina Government Data Analysis Center (GDAC) point out that 'they didn't necessarily end up where they started to go.' In other words, it was intuitively obvious that analyzing data would be beneficial to determining the success and flaws of governmental operations. But, in the same way that the first smart phone was not developed to allow one to deposit checks from the dining room table (but does), nascent data scientists did not imagine the range of opportunities to improve the delivery and cost-effectiveness of government services.

The point here is that analytics is inherently agile, and becomes more agile with the passing of time. The analytics team gathers knowledge and experience, new data becomes available that can be added to the inventory, and sophisticated tools become both even more capable and simpler to use.

[1] HiPPO was a term coined by Avinash Kaushik and Ronny Kohavi in 2006, first blogged by Kaushik at http://www.kaushik.net/avinash/seven-steps-to-creating-a-data-driven-decision-making-culture/

3. General Process Model

Generating actionable knowledge for a data-driven organization consists of a number of steps in a data and analytics lifecycle. The data lifecycle is provided in Figure 2-2.

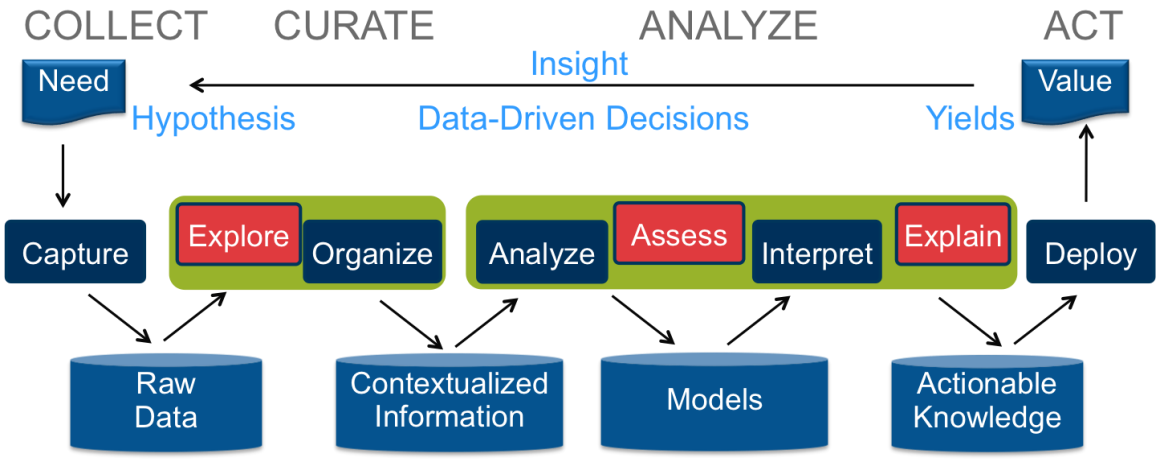

Figure 2-2: Data Science Lifecycle Model

Following this lifecycle provides the process model to address any hypothesis that is judged worthy for exploration.

A.1 Business Need

Develop a needs assessment to identify the most critical issues for the state that are in need of data-driven decisions. Determine the relevant stakeholders, and brainstorm on potential scenarios for better evidence-based decision-making.

- Produce analysis of value proposition based on tangible and intangible benefits
- Document the benefits of data sharing
- Publicize the availability of data for government, business, NGO, and public use
- Meet with program staff and identify critical issues
- Devise abatement strategies based on the anticipated results

Note that knowledge must be actionable in the end. Analytics that can't be used provides no value.

A.2 Technical Needs

- Create the data management framework necessary for the data lake

- Design and prepare the pilot virtual data warehouse for the study using available data
- From the business hypotheses prepare the technical approach to test the hypothesis
- Identify the desired data to pinpoint demographic, business, or environmental stressors
- Expand the collection of data through public and private partnerships
- Provide training on data submission, use, and analysis

3.1.2 Collect

The first step is to identify the data relevant to the scenario. This data would then be placed into Utah's data lake. Pursuing multi-modal solutions will benefit from a wide variety of data—structured and unstructured, point and non-point source, health and environment, local and national, etc. Depending upon the hypothesis, candidate data sources for any environmentally-related scenario could span:

- Conduct an inventory of additional available data including source
    - State of Utah environmental databases
        - Environmental monitoring and reporting beyond air quality
        - Emissions inventories
        - TRI and
        - Permit limits
        - Base geospatial data
        - Aerial photography
    - Health trends and statistics
        - Population growth
        - Illness data
        - Mortality Trends
        - Costs of access to care
    - Utah natural resource data
        - Weather data
        - Pollen counts
        - Flora and fauna observations
        - Park visitation
    - Economic data
        - Personal income data
        - Seasonal variations in employment
        - Business and employer income generation
        - Agriculture and pesticide data
        - Tourism contribution to economy
    - Other State data sources that have a relationship with/to environmental quality (DMV vehicle data, traffic flow surveys, natural gas sales, housing starts, etc.)

  - Other sources from Utah entities
    - Local governments and regional partnerships
    - Business and NGOs
    - Academia
  - National and western regional data sources that include Utah
    - EPA national and Region 8
    - NIH
    - DOE
    - DOI
    - USDA
    - Western Governors Conference

3.1.3 Curate

- Explore the data
  - Review the identified sets and associated meta data
  - Determine common data to facilitate linkage
  - Decide if data is most effective being hosted local or remote
  - Document APIs where data is remote
  - Determine if all or part must be re-formatted for ease of use
  -
- Organize into a data framework
  - Determine data fusion strategy
  - Develop hosting strategy – local or cloud
  - Prepare instructions for access
  - Decide upon funding strategy including charge-back/showback plan

3.1.4 Analyze

Perform initial analysis to determine feasibility given the data and the analytical method

- Review resulting data to identify inferences, possible trends, and correlations
- Refine analysis as desired and execute again to yield more targeted results
- Assess results and determine how best to present the information
- Overlay data onto geospatial platform and/or infographics package
- Prepare briefing on results and potential actions

3.1.5 Act

The increased availability of relevant data and metadata that is virtually integrated can 'arm' government, business, NGOs, and the public with information to identify and balance the highest priority actions, the most important geographic areas to target, and the actions deemed least intrusive to the Utah lifestyle.

- Discuss data with program staff and identify possible actions in response

- Deliver briefing to decision makers
- Receive approval for actions
- Conduct pilot actions
- Assess success from actions
- Develop broader action plan
- Execute action plan for set timeframe

3.1.6 Evaluate

For each analysis, a continual improvement cycle is undertaken to determine whether analysis was valid, needs adjustment to continue, or should be completed.

4. Generating Value

Data science and an analytical center of excellence can support Utah's move to evidence-based management. To do so, however, the results must deliver tangible and intangible value. As noted in Task 4, some states with analytical functions have identified better ways to deliver services, reduced the cost of government through efficiency, and targeted waste, fraud and abuse.

Two scenarios are presented below that offer potential value to Utahns. Each has a different value proposition as depicted in the following table:

Scenario	Value Proposition
Air Quality	Low cost air monitoring in residential neighborhoods is a cost-effective method to understand the actual concentrations of woodsmoke and target activities to reduce PM 2.5 exposure.
Skills Gap	Provides indicators of why Utahns do not complete college and suggest possible reasons that, with more study, could be actionable.

Following the scenarios, a summary table of three other use cases is presented.

4.1 Air Quality Targeting Scenario

This scenario focuses on reducing residential woodsmoke and its contribution to the PM 2.5 problems in the Salt Lake basin. By installing a network of low cost, neighborhood monitoring stations with real time data transmission, Utah DAQ can identify areas with the most persistent PM 2.5 challenges and conduct more effective source abatement activities including regulatory actions, enforcement actions, and citizen information provision.

The value proposition is that more location-specific and actionable data on particulate matter in Salt Lake neighborhoods could be collected, curated, analyzed and solutions proposed at or below the cost of an additional stationary monitoring station.

4.1.1 Understanding Need

A review of the literature on Utah's air quality challenges demonstrates support for multi-modal solutions and the value of public and private roles in pollution abatement. This scenario discusses how the expansion of monitoring and data collection on particulate matter can foster more sophisticated analysis to target the most cost effective and politically acceptable solutions.

A combination of geography, weather, economic activity and lifestyle combine to make air quality an important concern among government, business, and the public. It is widely recognized that Utahn's quality of life is negatively impacted through health impacts, economic costs, and reduced enjoyment of Utah's spectacular scenery and recreational opportunities.

Among the findings of a study by the Utah Foundation on Air Quality was that "woodsmoke was a greater source of particulate matter than previously thought." As much as 10% of the PM 2.5 originates from woodsmoke. [2]

The relevant facts are:

- Sizable numbers of homes in the Salt Lake area burn wood as a primary source of heat; many others burn wood for recreation and ambiance
- Woodstoves and fireplaces emit a significant amount of PM 2.5
    - A typical fireplace emits 3,373 times as much particulate matter as a gas furnace.
    - Heating a single home with a wood stove is equal in emissions to heating hundreds of homes with natural gas.
- Utah DAQ has 5 stationary monitoring stations in Salt Lake County serving its 808 square miles and 1.08 million residents.
- A stationary monitoring location costs approximately $20,000 per instrument plus facility preparation and housing. For example, a new Wyoming monitoring site cost $190,000.
- A study commissioned by the California Air Resources Board demonstrated that concentrations of PM from woodsmoke varied as much as 300% within neighborhoods in a small town.
- Utah has taken steps to communicate with residents about no-burn days but these efforts are not targeted and are not completely effective
- DAQ's enforcement efforts are limited to responding to citizen complaints

[2] Source: "The Air We Breathe: A Broad Analysis of Utah's Air Quality and Policy Solutions" Utah Foundation, Research Report Number 719, January 2014

resulting in only 16 violations in 2014.
- Alternatives to wood stoves include natural or propane gas burners that are vastly more efficient.

4.1.2 Technical Approach

Addressing the problem of wood-burning stoves will require obtaining sensor data that is both accurate, and provides proper coverage across critical areas. This would need to be matched with weather data that predicts inversion conditions and on-the-ground temperatures; utility data of the residences with wood-burning stoves versus other forms of heat sources; GIS registration of addresses to physical locations; additional point sources that could exacerbate the conditions and push into a "red" day situation. Bringing all this contextual information together for past woodsmoke emissions with weather prediction models for upcoming conditions would allow the prediction of days when residential wood-burning should be restricted.

4.1.3 Collect

From a data perspective, this effort focuses on achieving better identification of residential areas that experience higher exposure to PM 2.5. To accomplish this could require a wide variety of data both structured and unstructured. To *collect* data for this scenario, SAIC recommends:

- Conduct an inventory of relevant, available data:
    - State of Utah environmental databases
        - Environmental monitoring and reporting
        - Emissions inventories
        - Base geospatial data
        - Aerial photography
    - Utah natural resource data
        - Weather data
    - Other State data sources that have a relationship with/to environmental quality (DMV vehicle data, traffic flow surveys, natural gas availability, housing starts, etc.)
    - Utah's GIS database of addresses
    - Other sources from Utah entities
        - Local governments and regional partnerships
        - Business and NGOs
        - Academia

- Conduct a gap analysis to understand of data availability

- Develop a plan to fill gaps using locally sited, lower cost particulate monitors.
    - Identify locations
    - Develop cost justifications and deployment strategies for monitors

- Prepare data management plan for processing, hosting and presentation

- Procure and deploy monitors
- Initiate data collection in accordance with plan

4.1.4 Curate

- Explore the data
  - Review the identified sets and associated meta data
  - Determine common data to facilitate linkage
  - Decide if data is most effective being hosted in the data lake or having it remain in its remote location
  - Document APIs where data is remote
  - Determine if all or part must be re-formatted or interpolated to a new grid for ease of use

- Organize into a data framework
  - Prepare schema for logical data integration
  - Develop hosting strategy – local or cloud
  - Prepare instructions for access
  - Decide upon funding strategy including charge-back/showback plan

4.1.5 Analyze

- Formulate lines of inquiry for analysis
  - Identify neighborhoods and locations with high readings of PM 2.5
  - Predict emissions in neighborhoods of PM 2.5 based on temperature and winds indicative of an inversion condition
  - Determine residential addresses for the regions at risk
  - Target mailings based on target addresses regarding
    - Impact of burning wood for heat on air quality
    - Explanation of requirements during inversions
    - Offers and incentives to switch heat source to natural gas based on availability (utility maps)
  - Identify locations to most effectively position enforcement personnel during no-burn days
- Analyze multi-seasonal trends for the neighborhoods most often exhibiting "red" days
- Assess results and determine how best to present the information

Prepare briefing on results and potential actions

- Refine code as desired and execute again to yield more targeted results

4.1.6 Act

Based on the data, the following mitigations could be pursued:

- Overlay data onto geospatial platform and/or infographics package
- Discuss data with program staff and identify possible actions in response
- Coordinate information with legislature, neighborhood representatives, and law enforcement
- Send out mailings in the fall about the hazards of wood-burning during inversion days
- Create a GIS display of particulate counts as a contour plot overlay, or by neighborhoods as a web page or in a mobile app
- Based on predicted areas where PM 2.5 concentration is the greatest,
    - Increase information dissemination to those areas through mail, internet, smartphone applications, social media, and community/religious meetings
    - Pre-position Utah's limited number of enforcement officers to reduce illegal burning during 'red' days
    - Establish a citizens air watch program to identify those burning wood on days of poor air quality
- Develop longer-term strategies
    - Offer incentives to those areas where the impact of moving from wood to an alternative source would do the most good.
    - Refine other pollution abatement strategies based on the data collected on woodsmoke in neighborhoods

4.1.7 Evaluate

- Review monitoring data to determine if actions had desired affect
- Discuss in social media to allow public to comment on effectiveness
- Determine if further effort required to improve pollution abatement
    - Re-run analysis with different or more refined data
    - Adjust action plan based on lessons learned
    - Consider adjunct activities to increase program effectiveness
    - Relocate monitors that are ineffective
- Conduct validation of program results annually and decide on approach
    - Continue program as in prior year
    - Adjust program based on new data

4.2 The Skills Gap

One of the challenging issues facing Utah is the creation of an educated workforce. There is motivation for students to have higher education degrees. Studies have indicated the disparity in incomes based upon education. According to the Census Bureau, young adults between the ages of 25 and 34 with a college degree working year round, earn about 40 percent more than similar age adults with some college who have not completed a degree, and about two-thirds more than similar age adults

with just a high school diploma. [3] For the state, this results in lost revenue when students drop out and do not complete their degree. Given the lower unemployment rates for college graduates, this same study estimated that the lost revenue for students who began in 2002 but did not complete their degree as a loss of $17 million dollars, with a lost state income tax revenue of $870,000. Over time, this has a substantial negative impact on the Utah economy.

A second issue is the concern over the skills for the workforce. According to a recent study by Chegg[4], only 27% of college graduates are hired for a job related to their major. Furthermore 39% of hiring managers said recent college graduates they have interviewed in the past two years were completely or very prepared for a job in their field of study. In a study of Utah college graduates, one in every 4.5 freshmen at a Community College already had a bachelor's degree. This implies they graduated college, but could not find a job.

4.2.1 Understanding Need

The numbers on drop-out rates, while informative, are not actionable; they are trailing indicators. They are measuring an end result but do not indicate any way to affect that end result. They are merely estimating the magnitude of the problem. To be actionable, policy makers need to understand the cause and effect relationships that lead to individuals having either dropping out, or having to change career paths due to a lack of skills.

4.2.2 Technical Approach

To affect change in the skills gap there must be a determination of leading indicators that could provide information to decision-makers on areas where they can take action. This would take the form of determining for example the characteristic differences between graduates who find employment in their field, and those who do not; and determining the most prevalent reasons that students do not complete their degree work.

4.2.3 Collect

At the heart of any analytics is in fact the data that you collect. If the root cause was just in high school grade point averages, then those could be used to better steer students to colleges appropriate to their skills, or provide them with additional

[3] American Institutes for Research, http://www.air.org/sites/default/files/downloads/report/AIR_High_Cost_of_Low_Graduation_Aug2011_0.pdf

[4] "Bridge That Gap: Analyzing the Student Skill Index", https://www.insidehighered.com/sites/default/server_files/files/Bridge%20That%20Gap-v8.pdf

college preparatory classes in high school or in a community college prior to university admittance.

The best way to determine the dominant factors is to collect the data about all factors, and then in analysis determine which ones are correlated. From the most highly correlated factors you then determine which of those are causal. This would require combining education data at all levels, with workforce data, with community data, with socio-economic data, with family history, with early childhood interventions, etc.

4.2.4 Curate

This type of study cannot be performed on aggregate data. It could only be done with a "360 degree" view of a student. Only by integrating the data around the student's circumstances can a meaningful analysis be done.

At its core, root causes can only be determined if the data contains the factors that matter. As you do not know what those factors are, you have to figure out how to prepare the data so you can meaningfully use the information for each student. To compare across years, you can use ages for starting kindergarten, elementary, middle and high school, and college instead of dates. You can group students into the community they grew up in and not their address. You can take their grade point averages separately in core classes of English, history, math, and science. You can see if they had computer and typing and writing skills classes. You can see if they were active in sports or clubs. It is often the transformed combination of individual data fields that lets you capture the characteristics of the individual in a way that allows comparisons with other students.

4.2.5 Analyze

A range of methods can be used to determine the features that are highly correlated with dropping out of college. One simple analysis would be for example to use a decision tree, that takes the answer (such as dropped out or not) and looks at all the dimensions of the data (pre-college education, college major and minor, socio-economic upbringing, education of parents and siblings, employment of family members) and splits the population among those characteristics. It can be viewed as a series of if-then statements, until you can clearly identify the characteristics that separate graduates from drop-outs. Another possibility is to produce a linear-regression model that determines the weights across the contributory factors to determine the driving forces in the drop-out rate.

4.2.6 Act

Depending upon the highly correlated factors, they can be treated as leading-indicators. If there is a high correlation to inter-generational factors, then family assistance or early learning programs can be targeted. If it is driven by community

factors, then greater emphasis can be placed on community factors. Resources can be allocated at schools for at-risk students to provide one-on-one tutoring.

While the scenario discussed here is a simplistic look at a complex problem, until the full picture of information for students is analyzed and the dominant factors or leading indicators are determined, then no targeted action can be taken.

4.2.7 Evaluate

Purely measuring drop-out rates and evaluating students only for the current position in high school or college is insufficient. Once leading indicators are found, then the students in the targeted group for assistance can be tracked, and any changes from previous drop-out rates at each level can be analyzed. Performed on a year-by-year basis, it will be clear whether the mitigation is effective. If not, then either another mitigation strategy will need to be tried, or a different leading indicator will need to be targeted.

4.3 Additional Scenarios

Beyond air quality and educational skills, there are a myriad of other opportunities to use big data in Utah government. The following table presents three common use cases that governments have found can deliver tangible and intangible benefits. Example actions/approaches are presented for each process in the model as described in section 3.0, above.

Use Case	Processes	Value Proposition
Health care cost effectiveness for Medicaid patients	Collect • Medicaid patient health records • ER visit records • Birth records • Parent/child information • Geospatial base information • Provider name and address	• Receiving care for chronic disease improves quality of life and reduces cost of delivery • Identifying target pregnancies for education and follow up calls improves likelihood of receiving proper prenatal care. • Predict density of Medicaid patients so ensure availability of local providers
	Curate • Identify linking data for each file • Build integrated datasets of relevant data •	
	Analyze • Build patient 'profile' of relevant data for all of those covered • Identify those with chronic disease • Identify those with inconsistent appointments based on Medicaid billing • Find parents with Medicaid yet few/no	

	prenatal visits • Identify other relevant factors based on correlation	
	Act • Flag pregnancies based on previous lack of prenatal care for caseworker follow up • Target those not seeking care for chronic disease for follow up • Redirect those using ER to other, lower cost providers	
Fraud reduction in income and health care support program	Collect • Income tax records • AFDC records • Subsidized school lunch records • Real and personal property transactions • Medicaid recipients • Geospatial base data	• Reduce waste, fraud and abuse in public assistance programs. • Improve value of investment in public assistance • Build confidence in state government
	Curate • Identify linking data for each file • Extract relevant integrated datasets from each to make size of files more manageable to search •	
	Analyze • Identify those with indicators of income (reported income, large taxable purchases, multiple vehicle registrations, etc.) beyond the expected means of those on public assistance • Correlate to receipt of AFDC, subsidized lunches, Medicaid, etc.	
	Act • Investigate high correlations to determine if acceptance of public assistance is fraudulent	
Inter-generational poverty	Collect • Medicaid • Food Stamps • Birth Records • School Records • Justice Records • Taxes • Community Joblessness	• Using the data across all government, justice, and educational services, analyze the patterns for childhood circumstances

	• Community housing • Community transportation services Curate • Organize information relating to individuals • Organize information relating to households • Organize information relating to the surrounding community Analyze • Examine across households for their parent and child relationships, the common characteristics of the household the individual grew up in • Develop predictive models based on an individual's household and community as a child to determine the dominant factors for those who remain in the poverty cycle versus those who break the cycle Act • Begin designing services that can provide extra services for the children in those households to align with the factors most likely to break the cycle • Begin measuring activities and participation in offered services	determining the driving factors for when the children broke out of the poverty cycle, and for those that did not • Based on these correlations, prioritize services that align with those most likely factors. • While the confirmation of the selected factors as leading indicators of breaking the cycle won't be confirmed for many years, the participation in those services can be tracked and evaluated
Growth management	Collect • Topography • Zoning and land use • Housing starts with address • Rental unit occupancy rates • Current traffic corridor data • Business license applications with physical location • Air quality monitoring data • Utility service areas Curate • Identify linking data for each file • Extract relevant integrated datasets from each to reduce file size Analyze • Develop predictive models for growth based on where businesses are locating, existing and future transportation, land	• Enhance/expand current approaches to predicting service demands • Use integrated data to better understand and communicate impacts of growth to citizens • Facilitate more accurate, longer range planning

	suitable for development • Develop predictive models for growth impacts including traffic congestion, air quality emissions, utility demands, etc.	
	Act • Utilize predictive models to manage growth • Ensure that schools and utilities are sited and available for new residential areas • Identify where developers can contribute to extension of public services • Arrange zoning to encourage growth around desirable amenities and reduce need for driving	

5. Conclusion

The scenarios presented above offer examples of how Utah can harness the power of data analytics. Starting with one or two pilot efforts, Utah government can learn the ropes, test the waters, and demonstrate the value proposition. Whether these two scenarios or a different one, the general process model provides a road map for implementing an analysis function in the State.