Data Science and Analytics in Government: Putting the “I” back into I.T.

Data ➔ Information ➔ Knowledge ➔ Insights (Discovery & Decisioning)

Kirk Borne

George Mason University, Fairfax, VA • www.kirkborne.net • @KirkDBorne
Challenges in Digital Government

Demands on Data Analysts
Multiple web properties, stakeholders, and constituencies require agency analysts to extract insights across a variety of web user communities and portals (internal self-service, internal cybersecurity, agency “customers”, public/citizen “customers”).

Lack of True Automation
Lack of automated processes prevents analysts from achieving targeted end-user web content delivery and performing in-depth analyses of massive web analytics data logs.
Dividends earned via Web Data Analytics Applications in Digital Gov

- **Web behavior modeling & segmentation.**
- “**Personalized**” web portal experience and navigation aids (recommendations, contextual tabs) for different user communities.
- **More effective internal self-service & external citizen services.**
- **Fraud detection & prevention** (e.g., “do not pay” initiatives in social services agencies).
- **Threat detection & mitigation** (external & insider threat).
Extracting Information, Knowledge, Insights, and Data-to-Decisions (D2D) from Big Data is hard!

Data ➔ Information ➔ Knowledge ➔ Insights (Discovery & Decisioning)
Data Science and Big Data Analytics are putting the “Information” back into I.T.

- Analytics turns Big Data into something useful:
  Data → **Information** → Knowledge → Discovery + Decisions → **Insights & Innovation**!

- Data-driven processes boost innovation = Big ROI


Rationale for Collecting Big Data

• The end of demographics!
  - If we collect a complete set of data for all items (people, or anything else) within any domain of study, then we would have a “perfect” statistical model for that domain.

• X-informatics! = Data Science for any domain X
  - Anything we want to know about that domain X is specified and encoded within the big data collection.
  - *Examples*: Geoinformatics, Bioinformatics, …

• The goal of Big Data Science is to find those encodings, patterns, and knowledge that help us to understand and improve the world.
  - *More examples*: Urban informatics, Security informatics, Transportation informatics, Finance informatics, Climate informatics, Sustainability informatics, …
Big Data: What is it good for?

*The 3 D2D’s*

- ✔ Knowledge Discovery
  - *Data-to-Discovery (insights)*
- ✔ Data-driven Decision Support
  - *Data-to-Decisions (insights)*
- ✔ Big ROI (Return On Innovation) !!!
  - *Data-to-Dividends (in digital government)*
The **BIG** Big Data Challenge: *Big Data for Social Good*

- **Examples:**
  - Smart X (= utilities / grid, homes, cities, transportation, highways, …)
  - Precision X (= medicine, farming, weather forecasting, supply chain, …)
  - Personalized X (= healthcare, education, marketing, …)
  - Optimized X (= natural resource usage, pricing, shipping, scheduling, …)
  - Predictive X (= water usage / drought, climate change, resource loads, …)
  - Computational Sustainability (data-driven evidence-based decisions)
  - Quality control via comprehensive data collection & novelty detection
  - Asset Management and ERP (Enterprise Resource Planning)
  - Causal Analysis (financial fraud prediction & prevention, risk analysis, …)
  - Social Event Alerts / Early Warnings ("social listening" in social media)
  - Customer Analytics ("Voice of the Customer", sentiment analysis, …)
  - … and much more …
Love thy data
(Data Science commandment #4)


Truly show your data some love by giving it something useful to do!
True Data Love = **Amo** = the **All-important MIPS mOdel** for **Dynamic Data-Driven Application Systems (DDDAS)**

- **MIPS** =
  - Measurement – Inference – Prediction – Steering

- **This applies to any Network of Sensors:**
  - Web user interactions & actions (web analytics data), Cyber network usage logs, Social network sentiment, Machine logs (of any kind), Manufacturing sensors, Health & Epidemic monitoring systems, Financial transactions, National Security, Utilities and Energy, Remote Sensing, Tsunami warnings, Weather/Climate events, Astronomical sky events, ...

- **Machine Learning enables the “IP” part of MIPS:**
  - Autonomous (or semi-autonomous) Classification
  - Intelligent Data Understanding
  - Rule-based
  - Model-based
  - Neural Networks
  - Markov Models
  - Bayes Inference Engines

**Alert & Response systems:**
- LSST 10million events
- “Mars Rover” anywhere
- Automation of any data-driven operational system

[http://dddas.org](http://dddas.org)
3 Big Data D2D examples

- Knowledge Discovery
  - LSST (discovery)
- Data-driven Decision Support
  - Mars Rovers (decisions)
- Big ROI (Return On Innovation) !!!
  - Analytics Automation & Insights Generation
Astronomy Big Data Example

The LSST (Large Synoptic Survey Telescope)
LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

(mirror funded by private donors)
8.4-meter diameter primary mirror = 10 square degrees!

Hello!
LSST = Large Synoptic Survey Telescope
http://www.lsst.org/

(mirror funded by private donors)
8.4-meter diameter primary mirror = 10 square degrees!

Construction began August 2014 (funded by NSF and DOE)

Hello !
LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

- 8.4-meter diameter primary mirror = 10 square degrees!
- (mirror funded by private donors)
- 100-200 Petabyte image archive
- 20-40 Petabyte database catalog
**LSST Key Science Drivers: Mapping the Dynamic Universe**

- Complete inventory of the Solar System (Near-Earth Objects; killer asteroids???)
- Nature of Dark Energy (Cosmology; Supernovae at edge of the known Universe)
- Optical transients (10 million daily event notifications sent within 60 seconds)
- Digital Milky Way (Dark Matter; Locations and velocities of 20 billion stars!)

**LSST in time and space:**

- **When?** ~2022-2032
- **Where?** Cerro Pachon, Chile

![Architect’s design of LSST Observatory](image)
LSST Summary
http://www.lsst.org/

• 3-Gigapixel camera
• One 6-Gigabyte image every 20 seconds
• 30 Terabytes every night for 10 years
• Repeat images of the entire night sky every 3 nights: *Celestial Cinematography*
• 100-Petabyte final image data archive anticipated – *all data are public!!*
• 20-Petabyte final database catalog anticipated
• Real-Time Event Mining: ~10 million events per night, every night, for 10 years!
  – Follow-up observations required to classify these
  – Which ones should we follow up? …

... Decisions! Decisions! ( = D2D !)
3 Big Data D2D examples

- Knowledge Discovery
  - LSST (discovery)

✓ Data-driven Decision Support
  - Mars Rovers (decisions)

- Big ROI (Return On Innovation) !!!
  - Analytics Automation & Insights Generation
Mars Rover Example
Mars Rover: intelligent data-gatherer, mobile data mining agent, and autonomous science decision-support system

Rove around the surface of Mars and take samples of rocks (experimental technique: mass spectroscopy = data histogram)

Intelligent Data Operations in Action:

• Supervised Learning (search for rocks with known compositions)
• Unsupervised Learning (discover what types of rocks are present, without preconceived biases)
• Association Mining (find unusual associations)
• Clustering (find the set of unique classes of rocks)
• Classification (assign rocks to known classes)
• Deviation/Outlier Detection (one-of-kind; interesting?)
• On-board Intelligent Data Understanding & Decision Support Systems (Fuzzy Logic & Decision Trees & Cased-Based Reasoning)

= Science Goal Monitoring:
  – "stay here and do more"; or else "move on to another rock"
  – "send results to Earth immediately"; or "send results later"
Mars Rover = Decision Science Engine (enabling data-to-decisions = DTD)

- Decisions are based on data mined, prior experience, new knowledge, and the set of learned rules.
- Rover acts autonomously, without human intervention, in Deep Space environment.
- Actions are driven by mining actionable data from all sensors.

http://legacy.samsi.info/200506/astro/presentations/tut1loredo-7.pdf
The first mile is the hardest

“This red line indicates the change in this red line over a period of time.”

The last mile is the hardest
Knowledge Discovery for multi-source Data:
Heterogeneous data collections are the new normal

The “first mile” challenge: integrating multi-source data

Data Source #1: Images
Data Source #2: Time Series
Data Source #3: Catalogs
Data Source #4: Model Outputs

Information Extracted:
- Shapes & Patterns
- Distributions & Frequencies
- Science Metadata
- Model Parameters

Unsupervised Methods
- Association Mining
- Clustering
- Principal Components
- Link Analysis

KDD Tools

Supervised Methods
- Neural Networks
- Bayesian Networks
- Markov Models
- Support Vector Machines

Data to Information to Knowledge

New Knowledge on correlations, causal connections, and interdependencies between events, objects, processes within any application domain
Decision Support from Data: Early Warning System for Astronauts in Space (actionable intelligence!)

The “last mile” challenge: deriving Actionable Intelligence

CME = Coronal Mass Ejection
SEP = Solar Energetic Particle
3 Big Data D2D examples

- **Knowledge Discovery**
  - *LSST (discovery)*

- **Data-driven Decision Support**
  - *Mars Rovers (decisions)*

✓ **Big ROI (Return On Innovation) !!!**
  - *Analytics Automation & Insights Generation*  
    (= dividends from digital government initiatives)
Decision Analytics – based on massive amounts of information (Big Data – What is it good for? Decision Support and Innovation!)

From Devices......

... Intentions...

... Demographics...

... Location, weather, and other geographic attributes...
Decision Analytics – based on massive amounts of information (Big Data – What is it good for? Decision Support and Innovation!)

From Devices......

... Intentions...

The “first mile” challenge:
integrating data from multiple sources

... Demographics...

... Location, weather, and other geographic attributes...

The “last mile” challenge:
deriving Actionable Intelligence
The BIG Big Data Challenge: identifying, characterizing, & responding to millions of events in real-time streaming data

- **Astronomy example:**
  - Real-Time Event Mining: deciding which of the ~10 million events every night need follow-up observations (for maximum scientific return)

- **Web Analytics example:**
  - Web Behavior Modeling and Automated System Response (from online interactions & web browse patterns, personalization, user segmentation, advanced analytics discovery, …)

- **Many other examples:**
  - Health alerts (from EHRs and national health systems)
  - Tsunami alerts (from geo sensors everywhere)
  - Cybersecurity alerts (from network logs)
  - Social event alerts or early warnings (from social media)
  - Preventive Fraud alerts (from financial program application data)
Enter... Advanced Analytics Automation!

• **Web Data Analytics:**
  – Outlier / Anomaly / Novelty / Surprise detection
  – Clustering (= New Class discovery, Segmentation)
  – Correlation & Association discovery
  – Classification, Diagnosis, Prediction

• … for D2D:
  – Data-to-Discoveries
  – Data-to-Decisions
  – Data-to-Dividends
Automating Analytics as-as-Service (AaaS)

• Based on SYNTASA’s Marketing Analytics-as-a-Service™ (MAaaS)
• “Mars Rover in a box”
  – Your business rules determine the goals, decision points, alerts, and responses.
  – Moving beyond historical **hindsight** and **oversight** (Descriptive & Diagnostic Analytics) to new world of **insight** and **foresight** (Predictive & Prescriptive AaaS), eventually achieving **right sight** (Cognitive Analytics = the 360 view, enabling the right action, for the right web user, at the right place, at the right time).
• Mining multi-portal big data streams (across the agency’s departments)
• Personalization and Customization (“segment of one”)
• Decision Automation in a rich content (Big Data) environment

Based on Marketing Analytics-as-a-Service™ (MAaaS) from http://www.syntasa.com/
The SYNTASA™ Analytics Automation Stack for Web User Behavior Modeling

**Applications**
- Segmentation
- Personalization
- Targeting
- SEO

**Analytics**
Visualization, Statistics, Anomaly, Prediction, and recommendation

**Machine Learning**
Feature Selection, Modeling, Evaluation

**Data Services**
Ingestion, storage, processing, Marketing schema

**Data Input**
- Web
- Search
- CRM
- Social
- Mobile
- Local
- Partner
- 3rd Party

© SYNTASA 2014 and Confidential
Summary: Maximize Your Agency’s Insights (discovery), Information utility (decisions), and Innovation (dividends) through Data Science and Analytics Maturity

Right Sight = The right action at the right time in the right place = Cognitive Analytics

= the 360° view

What’s the Question?

Hindsight through Descriptive Analytics & Oversight through Diagnostic Analytics

Foresight through Predictive Analytics

Insight through Prescriptive Analytics

Right Sight = The right action at the right time in the right place = Cognitive Analytics

= the 360° view

What’s the Question?