Introduction to Big Data:
Harnessing the Benefits of Data Powered Governance

Board of Directors Item 10 | January 24, 2020

Panel Members

Ray Major
Chief Analytics Officer, SANDAG

Professor Rajesh K. Gupta
Computer Science Director,
Halıcıoğlu Data Science Institute, UC San Diego

Andrell Bower
Chief Data Officer, City of San Diego

Christina Drummond
Senior Analyst, Ohio State University,
Moritz College of Law Program on Data and Governance

Sanjiv Nanda
Entrepreneur-in-Residence, SANDAG
Data Science: Connecting the dots in emerging science, society & living

Rajesh K. Gupta
Director

HALICIOĞLU DATA SCIENCE INSTITUTE
UC San Diego
Data have always been with us.

What has changed?

1. Ubiquity and scale of data that can be collected
1. Ubiquity and scale of data that can be collected
2. Development of sophisticated analytics tools
1. Ubiquity and scale of data that can be collected
2. Development of sophisticated analytics tools

Use of data in ways not previously imagined
New algorithms for locally sensitive hashing, approximate computing
Similar structures have similar effects.

Toxicology tested by matching against a massive DB of molecular structures and safety data.

Machine predictions that can reduce crime or amount of jailing.
New algorithms for locally sensitive hashing, approximate computing

Similar structures have similar effects

Toxicology tested by matching against a massive DB of molecular structures and safety data

Machine predictions that can reduce crime or amount of jailing

Data Science is about connecting the dots...

...to create new knowledge

...improve quality of life

...protect the environment around us
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner’s article “The Unreasonable Effectiveness of Mathematics in the Natural Sciences” examines why so much of physics can be neatly explained with simple mathematical formulas such as \( f = ma \) or \( e = mc^2 \). Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematical behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

**Learning from Text at Web Scale**
The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are

Data being used everywhere in daily lives

Detection

Segmentation

Lots of training data:
- Vision systems trained on millions of images
- Speech systems training on tens of thousands hours of annotated data
HDSI: A Convergence of Multiple Long-term Initiatives Launched in March 2018

*There is perhaps no other scientific or intellectual endeavor today that is more important than the advancement of data science.* — Peter Lee, Corporate Vice President, AI & Research, Microsoft

*Photo by Lan Yao/UC San Diego*
188 Faculty Drawn from All Areas of University

Understanding and Protecting the Planet
Enriching Human Life and Society
Exploring the Basis of Human Knowledge, Learning and Creativity
Understanding Cultures and Addressing Disparities in Society

HDSI

Environment Technology Health Microbiome Science Big Data

HALICIOĞLU DATA SCIENCE INSTITUTE

UC San Diego
Role of Data at SANDAG

Speaker: Ray Major
SANDAG is creating an environment where economic and social benefits are recognized and achieved using empirical information.

**Data Science** → **Regional Plan**

**Analytic Continuum**

- **Reporting**
- **Analysis**
- **Monitoring**
- **Forecasting**
- **Predictive**
- **Prescriptive**

- What happened
- Why did it happen
- What is happening now
- What may happen
- What possibly will happen
- Based in Real Time
- Based on History
- What actions should be
SANDAG's Evolution as a Data Driven Organization

- 1999: Intelligent Transportation System Architecture
- 2005: Regional Arterials Management System
- 2006: Integrated Corridor Management
- 2009: Work begins on Activity Based Model (ABM)
- 2010: INRIX Travel Time Data Incorporated in ABM
- 2012: Streetlight Origin-Destination Data Acquired for ABM
- 2013: Grant to Develop Solution for Analytics
- 2016: ATRI Data Commercial Vehicles
- 2017: AirSage Data for Military Base Trip Validation
- 2018: Regional Data Hub
- 2019: Data Governance (2018)
- 2020: Data used to drive Regional Plan

Where (BIG) Data Comes From

**Primary**
- Sensors
- Signals
- Cameras
- Public transit
- Traffic incidents
- Road closures
- Detectors

**Secondary**
- Navigation apps
- Weather
- Inrix/Streetlight
- Google
- Apple
- Amazon
- Uber/Lyft
- Scooters/bikes
Why Big Data is Essential in Making Policy Decisions for the Region

Anonymization/Privacy of the Economic, Demographic, and Transportation Data
Smart Cities & Open Data

Andrell Bower, Chief Data Officer
**Background**

**Open data** – an initiative launched in 2015

PUBLIC FACING...
Engaged with residents and increased transparency

... WHILE SHIFTING CULTURE
Data is a valuable asset that helps us meet strategic goals

---

**Inform policy**

We can decide the next category to add to Get it Done by conducting **content analysis** and **text mining** on reports submitted to the “Other” category.
Improve service

We used **911 call volumes** and **ambulance time on task** to calculate the **optimal number of ambulances** for each of our response zones.

Evaluate programs

A tool that calculates **parking meter utilization** by block using **daily transaction data** will help program managers pilot changes and make timely adjustments.
Thank you

Andrell Bower, abower@sandiego.gov

Visit our Open Data Portal

View our open-source data projects on Github
Central Ohio Case Studies in Civic Data Privacy & Governance

Christina Drummond, M.A. ISTP
Program Manager and Senior Analyst
Moritz College of Law, The Ohio State University
Drummond.120@osu.edu

Presentation to SANDAG Board of Directors
January 24, 2020

Efforts

Professional Networks

SMRT COLUMBUS

JOIN THE CIVIC DATA PRIVACY LEADERS NETWORK

MID OHIO REGIONAL MORPC PLANNING COMMISSION

METROLAB NETWORK
Regional Efforts

Regional Data Advisory Committee (RDAC)

- Public, private, nonprofit, academic data experts
- Regional Data Agenda provides strategic roadmap for regional data efforts
- Working groups progress Regional Data Agenda objectives

morpc.org/committees/regional-data-advisory-committee/
Contact Aaron Schill, Director, Data & Mapping. aschill@morpc.org
MORPC Regional Data Agenda

Specific action items support:

1. Fostering stakeholder collaboration
2. Data capacity building and education
3. “Good” data governance
4. Innovating via procurement and development
5. Model access, inclusivity, and equity

MORPC Regional Data Advisory Committee

- Data Policy Needs Survey & Toolkit
- Regional Information and Data Group
- Regional Municipal Fiber Strategy
- Sustainability Dashboard

Active Working Groups
Surveying and resourcing civic data governance needs

Data topics: e.g. cybersecurity, open data, privacy & ethics, contracting, sourcing and provisioning

Operations: Policy, procedures, staffing and training

Guiding data delivery & engagement

Personas span:
- Mid-sized City Planner
- Township Administrator
- Engaged Resident
- Elected Village Council Member
- Nonprofit Employee
- Consultancy Project Manager
- Civic Tech Enthusiast
To empower our residents to live their best lives through responsive, innovative and safe mobility solutions.

To demonstrate how an intelligent transportation system and equitable access to transportation can have positive impacts on every day challenges faced by cities.
TECHNICAL WORKING GROUP (TWG)

Data
- Identify use cases for data to solve community challenges & continue to make the city smarter

Technical
- Architecture & Design Recommendations
- Best Practices
- Tech Strategies
- Licensing / IP

Policy
- Data Management
- Data Privacy
- Policy Roadmap

SMART COLUMBUS
DATA PRIVACY AND MANAGEMENT PLANS

Data: Privacy, Management and Policies
Data Management Plan

- Governs data within Smart Columbus OS
- Addresses metadata management, data size, data acquisition, data access and data use.

https://smart.columbus.gov/programs/smart-city-demonstration

Data Inventory for each dataset:

- Description of Data
- Data Type
- Dataset
- Source of the Data
- Responsible Party
- Collection Approach
- Frequency
- Needed for Performance Management
- Period of Collection
- Users of the Data
- Value of the Data to the Users
- Does it Contain PII
- Relevant Data Standards
- Access Policies for the Data
- Where is the data located?
SMART COLUMBUS DATA PRIVACY PLAN

• High-level plan to protect privacy and secure data
  • Privacy controls
  • Security controls
  • De-identification
  • Ongoing governance

• A commitment to respect and to be a good steward of personal information

smart.columbus.gov/programs/smart-city-demonstration

HOW WERE THEY DEVELOPED?

• Guiding Light Documents
• Iterative Drafts with Special Reviews
• Diverse Drafting Team

smartcolumbusos.com/share-your-data
Types of Guiding Documents

Laws, including:
• ORC: Personal Information Systems, Public Records
• Federal: Privacy Act 1974
• International: GDPR

NIST Standards/Guides
• Security categorization standards, defining breach impacts (FIPS 199)
• Minimum security requirement specs, risk-based process control selection (FIPS 200)
• Information type to security category mapping (800-60v1)
• Protection of PII confidentiality (800-122)
• Security and privacy controls (800-53) *Control defining

Guiding documents continued…

Policy Reports/Playbooks
• Open Data Privacy Playbook (Green et al/Berkman)
• De-identification protocol for Open Data (IAPP)

Policy from other cities
• Open Data Release Toolkit – Data SF/San Francisco
• Open Data Risk Assessment - Seattle
• Data Privacy Plan – Connected Vehicle Pilot / Tampa
Iterative Plan Development Process

Drafting Dream Team

- Over two years
- 38 people from across 22 organizations
- Facilitated by Moritz Program on Data and Governance
- Diverse perspectives:
  - Public Sector IT
  - Lawyers
  - Information Privacy Professionals
  - Corporate Chief Privacy Officers
  - Unaffiliated national privacy experts
  - OSU students
**Data Privacy Plan Elements**

- Data Stewardship
- Applicable Law Compliance
- PII Definitions
- Administrative/Legal Safeguards
- Scope of plan related demonstration data
1. Notice and Consent
2. Data minimization
3. PII Use/Sharing
4. Data retention
5. Access, correction, deletion
6. Transparency
7. Accountability
8. De-identification
9. Data Curation, incl.
   Data Inventory spec.
   Benefit/Risk Analysis process
10. Privacy Testing
11. Data compartmentalization
12. IRB
13. Privacy and Security Board
15. Contractor / 3rd party compliance
16. Privacy Impact Assessments
Data Privacy Plan Elements

Specified Security Controls

1. Encryption
2. Physical controls
3. Access controls, with logs
4. Role/ID-based authorization
5. Penetration Testing
6. Secure software development
7. System monitoring
8. Data loss prevention

Specified Security Controls

9. Patching, Antivirus, Malware Checks
10. DPP training
11. Accountability and event review
   • Incident response plan spec.
Data Privacy Plan Elements

Privacy Impact Assessment

- PIA question prompts
  - Data description, use, retention
  - Internal and external sharing/disclosure
  - Data collection notice, consent
  - Data access, redress, correction
  - Project security controls

Additional Resources
Civic Data Privacy Leaders Network

- Peer-to-peer knowledge & resource sharing among local government officials

Future of Privacy Forum

- Resources and reports related to civic data privacy and governance

Contact: Kelsey Finch, Senior Counsel, Future of Privacy Forum. kfinch@fpf.org

MetroLab

- City/university partnerships
- Enabling transformation to data-driven decision-making
- Annual Summit

San Diego / UCSD members
Next OS: Clearinghouse for Transportation Data
Sanjiv Nanda
What is the Next OS?

Who...
- Planners and Policymakers
- Transportation Operators & Service Providers
- Residents and Visitors Goods and Freight

What...
- Informed and responsive planning and decision making
- Ability to balance supply and demand across the transportation system
- Seamless access to transportation information and services

How...
- A platform with public and private data and travel modeling and forecasting tools
- A suite of dashboards, online simulations, analytics and dynamic control tools
- Access to real-time information and a suite of trip planning tools

Next OS: Data Clearinghouse, Suite of Data Driven Tools and Applications

Freeway Service Patrol Incident Response
Next Steps

Regional Plan Board presentation topics (January – March 2020):

✓ Our Local Economy
✓ Big Data – Harnessing the Benefits of Data-Powered Governance
  - Overview of Air Quality, Vehicle Miles Traveled, and Greenhouse Gas Emissions Requirements
  - Primer on Modeling
  - Environmental Impact Reports

Spring 2020: Presentation of the Vision for the 2021 Regional Plan