Geospatially Enabling Duke’s Enterprise Data Warehouse Tool

Howard Shang, Senior IT Director
Sohayla Pruitt, Senior Analyst
Duke Health Technology Solutions
Agenda

- Brief Overview of Duke’s Enterprise Data Warehouse Tool: *Duke Enterprise Data Unified Content Explorer (DEDUCE)*
- Geospatial Capabilities Being Implemented within DEDUCE
- Tools and Methodologies Utilized
• 3 Hospitals (in Durham and Raleigh, North Carolina)
• Dozens of Duke clinics and over 200 affiliated clinics
• 61,749 Inpatient admission—FY2010
• 1.8 million outpatient visits—FY2010
• 85,512 surgical procedures, including endoscopy—FY2010
• 2600+ employed & affiliated physicians—FY2010
Duke’s Enterprise Data Warehouse Environment

Duke’s Data Warehouse Group (DWG)
- 32 Full-time employees
- 1 Project Manager, 3 Resource Managers, 3 BI Developers, 4 Architects, 8 Data Analysts, 6 ETL Developers, 3 QA Analysts, 1 Application Developer, 3 Health Intelligence Analysts

Decision Support Repository (DSR)
- 13 major data sources
- 3 Terabytes of data
- Enterprise class Oracle 10g server
DEDUCE Overview

- Duke Enterprise Data Unified Content Explorer (DEDUCE) is a web-based query tool provided free of charge to Duke researchers and developed for cohort identification and data extraction.

- DEDUCE uses business intelligence to provide investigators the ability to filter millions of administrative and clinical records that are generated during patient care and integrated within Duke’s Enterprise Data Warehouse: the Decision Support Repository (DSR).

- DEDUCE has several tools that support different activities and ways to interact with data:
  - Guided Query: a wizard to create simple cohorts and receive pre-defined data exports.
  - Cohort Manager: a more advanced tool that allows sophisticated filtering for creation of complex cohorts and construction of custom data exports.
Common Uses of DEDUCE

- **Preparing for Research:**
  - How many patients meet the study criteria?
  - What are their gender and age?

- **Identifying potential participants with clinical research:**
  - Who are they?
  - Where have they been seen?
  - Who is their primary care provider?
  - When is their next appointment?

- **Extracting data to support clinical research analyses, quality improvement initiatives, and metrics reporting.**
DEDUCE Fact Sheet

Current Feature Set
- Project Centric
- Filter Wizard
- Cohort Manager
- Text Analytics
- Chart Review
- Smart Dimensions
- Internal Data Dictionary
- Asynchronous Query Execution
- Data structure agnostic
- MRN/Patient Key Upload
- Export Manager

Technology

Front End:
- IIS Web Server
- ASP.NET
- Active Directory for Authentication

Back End:
- Oracle 10g Database
- Specialized Data Structures
- Structured Data and Unstructured Text
- EIRB

Security and Patient Confidentiality

Roles-based authorization and account approvals
- Authentication through enterprise Active Directory accounts
- IRB protocol authorization process
- Distinct levels of access to de-identified data, limited data sets, and PHI
- Data usage agreements and terms of use
- Perpetual audit trails

HDWA 2012
Ottawa, Canada
With so many features...

- Role Based Authorization
- Guided Query
- Chart Review
- Text Analytics
- Cohort Manager
- DEDUCE
With so many features...

the addition of geospatial mapping and analytics is the next logical step
Geospatial Capabilities Being Implemented within DEDUCE

*a phased approach*:

**Phase I**

**Geospatially enable Duke’s Enterprise Data Warehouse**

- **How**: Standardize and Geocode patient addresses
- **Tools**: SAS Data Management Studio / USPS Knowledge Base / Tom-Tom Rooftop Geocoding Data Pack
- **When**: Complete as of August 2012!!!
Geospatial Capabilities Being Implemented within DEDUCE

*a phased approach:*

**Phase I**

Geospatially enable Duke’s Enterprise Data Warehouse

- **How:** Standardize and Geocode patient addresses
- **Tools:** SAS Data Management Studio / USPS Knowledge Base / Tom-Tom Rooftop Geocoding Data Pack
- **When:** Complete as of August 2012!!!

**Phase II**

Geospatially Enable DEDUCE with Mapping Dashboard

- **Capabilities:** Cohort visualization for different levels of PHI access: Dot Mapping, Density Heat-Maps, Block-Group Summaries.
- **Tools:** ESRI ArcServer, Javascript API, Dojo Toolkits, ASP.NET
- **When:** DEDUCE v4.3 release in November 2012!
Geospatial Capabilities Being Implemented within DEDUCE

*a phased approach:*

**Phase I**

*Geospatially enable Duke’s Enterprise Data Warehouse*

- **How:** Standardize and Geocode patient addresses
- **Tools:** SAS Data Management Studio / USPS Knowledge Base / Tom-Tom Rooftop Geocoding Data Pack
- **When:** Complete as of August 2012!!!

**Phase II**

*Geospatially Enable DEDUCE with Mapping Dashboard*

- **Capabilities:** Cohort visualization for different levels of PHI access: Dot Mapping, Density Heat-Maps, Block-Group Summaries.
- **Tools:** ESRI ArcServer, Javascript API, Dojo Toolkits, ASP.NET
- **When:** DEDUCE v4.3 release in November 2012!

**Phase III**

*Advanced Geospatial Queries and Analysis within DEDUCE Mapping Dashboard*

- **Capabilities:** Cohort queries by geospatial, demographic, and socioeconomic criteria and basic analytics to determine geo-statistical correlations.
- **Tools:** Built within existing front-end framework, but deploy more robust algorithms to pre-compute geospatial, demographic, and socioeconomic relationships.
- **When:** DEDUCE v5.3 release slated for Fall 2013.
Geospatial Capabilities Being Implemented within DEDUCE

*a phased approach*:

**Phase I**
- **Geospatially enable Duke’s Enterprise Data Warehouse**
  - **How**: Standardize and Geocode patient addresses
  - **Tools**: SAS Data Management Studio / USPS Knowledge Base / Tom-Tom Rooftop Geocoding Data Pack
  - **When**: Complete as of August 2012!!

**Phase II**
- **Geospatially Enable DEDUCE with Mapping Dashboard**
  - **Capabilities**: Cohort visualization for different levels of PHI access: Dot Mapping, Density Heat-Maps, Block-Group Summaries.
  - **Tools**: ESRI ArcServer, Javascript API, Dojo Toolkits, ASP.NET
  - **When**: DEDUCE v4.3 release in November 2012!

**Phase III**
- **Advanced Geospatial Queries and Analysis within DEDUCE Mapping Dashboard**
  - **Capabilities**: Cohort queries by geospatial, demographic, and socioeconomic criteria and basic analytics to determine geo-statistical correlations.
  - **Tools**: Built within existing front-end framework, but deploy more robust algorithms to pre-compute geospatial, demographic, and socioeconomic relationships.
  - **When**: DEDUCE v5.3 release slated for Fall 2013.

**Phase IV**
- **Geospatial Predictive Analytics within DEDUCE Mapping Dashboard**
  - **Capabilities**: Identify new patients or neighborhoods where a particular phenomenon has a high likelihood of future occurrence.
  - **Tools**: TBD
  - **When**: TBD
Phase I Goal:
Geospatially Enable Duke’s EDW (the DSR)

<table>
<thead>
<tr>
<th>Address as typed by someone</th>
<th>Standardized version</th>
</tr>
</thead>
<tbody>
<tr>
<td>121 Jones street</td>
<td>121 Jones St</td>
</tr>
<tr>
<td>Montgomery, AL</td>
<td>Montgomery AL 36104-4945</td>
</tr>
<tr>
<td>2993 Johnson</td>
<td>2993 Johnson Pl</td>
</tr>
<tr>
<td>Wantagh New York state 11793</td>
<td>Wantagh NY 11793-2836</td>
</tr>
<tr>
<td>2701 Phillips Ave., charlotte, N.C.</td>
<td>2701 Phillips Ave, Charlotte NC 28209-7029</td>
</tr>
<tr>
<td>616 Ivory</td>
<td>616 Ivory Rd SE</td>
</tr>
<tr>
<td>rio rancho NM, 87124</td>
<td>Rio Rancho NM 87124</td>
</tr>
</tbody>
</table>

Create USPS Verified and Standardized Records
Create Geocoded Records

HDWA 2012
Ottawa, Canada
An automated ETL job using SAS Data Management Studio, the USPS knowledge pack, and the Tom-Tom Rooftop +6 Geocoding Data Pack, ran initially on all address records, and runs nightly on new address records.

Requires NO manual interaction with the address data.

Time to complete:
- 5.3M records: 8hrs
- Nightly process: minutes

Phase I Results Summary:
Completed August 2012!
Phase I Results Summary:  
*Completed August 2012!*

- An automated ETL job using SAS Data Management Studio, the USPS knowledge pack, and the Tom-Tom Rooftop +6 Geocoding Data Pack, ran initially on all address records, and runs nightly on new address records. 
- Requires NO manual interaction with the address data. 
- Time to complete: 
  - 5.3M records: 8hrs
  - Nightly process: minutes

**ETL Job Analyzes Address Table within the DSR**

- Table contains 5.3 million records related to a Patient’s mailing and/or residential address. 
- Derived from multiple source systems. 
- Entire address is stored in a single text string. 
- Contains a vast number of non-address strings, misspellings, and incomplete address information.

**Methodology**
An automated ETL job using SAS Data Management Studio, the USPS knowledge pack, and the Tom-Tom Rooftop +6 Geocoding Data Pack, ran initially on all address records, and runs nightly on new address records. Requires NO manual interaction with the address data.

Time to complete:
- 5.3M records: 8hrs
- Nightly process: minutes

Table contains 5.3 million records related to a Patient’s mailing and/or residential address.
- Derived from multiple source systems.
- Entire address is stored in a single text string
- Contains a vast number of non-address strings, misspellings, and incomplete address information.

Returns USPS Verified and Standardized Address Fields
- 4.5 million records verified against USPS, parsed, and standardized into standard address fields.
- Useful new fields added and populated that are not collected from source systems (i.e. County, fully qualified 9-digit zip codes, etc.)
Phase I Results Summary:  
*Completed August 2012!*

**Methodology**
- An automated ETL job using SAS Data Management Studio, the USPS knowledge pack, and the Tom-Tom Rooftop +6 Geocoding Data Pack, ran initially on all address records, and runs nightly on new address records.
- Requires NO manual interaction with the address data.
- Time to complete:
  - 5.3M records: 8hrs
  - Nightly process: minutes

**ETL Job Analyzes Address Table within the DSR**
- Table contains 5.3 million records related to a Patient’s mailing and/or residential address.
- Derived from multiple source systems.
- Entire address is stored in a single text string
- Contains a vast number of non-address strings, misspellings, and incomplete address information.

**Returns USPS Verified and Standardized Address Fields**
- 4.5 million records verified against USPS, parsed, and standardized into standard address fields.
- Useful new fields added and populated that are not collected from source systems (i.e. County, fully qualified 9-digit zip codes, etc.)

**Geocodes Verified USPS Addresses to Rooftop Level of Accuracy**
- 3.7 million address records geocoded to rooftop level of accuracy.
- Census FIPS code fields added for county, tract, block group, block, MSAs, etc.
Phase II Goal: Geospatially Enable DEDUCE

- Spatially Enabled DSR
- Phase II Goal: Geospatially Enable DEDUCE
- Construct a cohort query in DEDUCE
- Visualize the results in the mapping dashboard
- Spatially Enabled DEDUCE
Next Release of DEDUCE—November 2012!

- Send cohorts from Cohort Manager to Mapping Dashboard for viewing
- View cohorts in three ways depending on PHI authorization: Dot Map, Density Surface (Heat Map), or Summarized by Census Block Group
- Modify layer visibility, order, and symbology
- Change base map (variety of options: satellite, canvas, hybrid, street map, etc.)
- Navigate map (zoom, pan, etc.)
- Export map to PDF
- Save mapping project

Feature Set

- ASP.NET
- ESRI ArcServer
- Javascript API
- Dojo Toolkits

Technology

- Same roles-based authorization and account approval
- Same IRB protocol authorization process
- Same Distinct levels of access to de-identified data, limited data sets, and PHI
- Same Data usage agreements and terms of use
- Same authentication through enterprise Active Directory accounts

Security and Patient Confidentiality

Phase II Capability Summary:
Phase III and IV Capability Summary:

*Slated for Fall 2013 Release*

Upon successful implementation of Phase III and IV, DEDUCE users will also be able to:
Upon successful implementation of Phase III and IV, DEDUCE users will also be able to:

- Further refine and visualize their cohort based on geospatial characteristics (i.e. only those patients who live near fast food restaurants, live in low income neighborhoods with substandard education)

Phase III and IV Capability Summary:  

*Slated for Fall 2013 Release*
Phase III and IV Capability Summary: 
*Slated for Fall 2013 Release*

Upon successful implementation of Phase III and IV, DEDUCE users will also be able to:

- Further refine and visualize their cohort based on geospatial characteristics (i.e. only those patients who live near fast food restaurants, live in low income neighborhoods with substandard education)
- Understand the relationship between their patient cohort and the hundreds of geospatial, demographic, and socioeconomic variables that are pre-calculated and housed in the DSR

![Graphs showing trends in distance to fast food, gyms, and roads for cohort and all patients.](image-url)
Upon successful implementation of Phase III and IV, DEDUCE users will also be able to:

- Further refine and visualize their cohort based on geospatial characteristics (i.e. only those patients who live near fast food restaurants, live in low income neighborhoods with substandard education)
- Understand the relationship between their patient cohort and the hundreds of geospatial, demographic, and socioeconomic variables that are pre-calculated and housed in the DSR
- Identify NEW areas or patients where the particular disease or phenomenon should be closely monitored from a prevention perspective

Phase III and IV Capability Summary:
Slated for Fall 2013 Release
Phase III and IV Methodology Summary:

*Slated for Fall 2013 Release*

Integration of Location-Based Queries, Geospatial Analysis, and Predictive Analytics into the DEDUCE Mapping Dashboard
Phase III and IV Methodology Summary:

*Slated for Fall 2013 Release*

Integration of Location-Based Queries, Geospatial Analysis, and Predictive Analytics into the DEDUCE Mapping Dashboard
Phase III and IV Methodology Summary:

Slated for Fall 2013 Release

Geospatial Data Collection

Pre-calculation of Geospatial Relationships per Address

100’s of Geospatial Distance and Value Variables computed, to include:
- Distance to Highways
- Distance to Restaurants
- Distance to Hospitals
- Distance to Fitness Facilities
- Distance to Health Food Stores
- Median Income of Block Group
- Male/Female Ratio of Block Group
- Predominant Race of Block Group
- Education Level of Block Group

Integration of Location-Based Queries, Geospatial Analysis, and Predictive Analytics into the DEDUCE Mapping Dashboard
Phase III and IV Methodology Summary:  
*Slated for Fall 2013 Release*

- Geospatial Data Collection
- Pre-calculation of Geospatial Relationships per Address
- Enabling of Geospatial Queries within Cohort Manager
- Integration of Location-Based Queries, Geospatial Analysis, and Predictive Analytics into the DEDUCE Mapping Dashboard
Phase III and IV Methodology Summary: 
*Slated for Fall 2013 Release*

- Geospatial Data Collection
- Pre-calculation of Geospatial Relationships per Address
- Enabling of Geospatial Queries within Cohort Manager
- Enabling the computation of geo-Statistical Relationships
- Integration of Location-Based Queries, Geospatial Analysis, and Predictive Analytics into the DEDUCE Mapping Dashboard
Benefits of Geospatial Integration:

- By integrating geospatial software and data into the DSR and DEDUCE, it will allow our community to understand more of what is happening, the contributing factors, and the ability to predict what might happen in the future within a geographic space.

- The geospatial data that gets integrated with the traditional health care data collected for a patient and their encounter will be very useful—as it will contribute a valuable set of data elements that would not be collected upon interaction with a patient (i.e. educational attainment, income, primary mode of transportation, distance from primary care clinics, distance from fitness facilities, etc.)

- Just as our DEDUCE users do not have to be expert data modelers and SQL programmers in order to access the data they need to conduct their research, they will similarly have access to valuable geospatial data and analysis without having to understand where to acquire the geospatial data, how to process it, and how to interact with geospatial software to get the information and analysis they need.

- This approach to research and management will transform how our organization will operate, allowing for:
  - Cost Savings and Increased Efficiency
  - Better Decision Making
  - Better Patient Care
  - More Efficient Self-Managed Care Among Patients
Contact Information

Howard Shang
Senior IT Director

• Duke Health Technology Solutions
• howard.shang@duke.edu

Sohayla Pruitt
Senior Analyst

• Duke Health Technology Solutions
• sohayla.pruitt@duke.edu
• Responsible for leading the design, development, and implementation of the geospatial infrastructure within Duke’s EDW.
• M.A. and B.A. in Geosciences, Specializing in Geographic Information Systems (GIS) and Remote Sensing (RS)
• 14 years experiences with Geospatial Analysis
• Managed and conducted GIS, RS, and Geospatial Predictive Analytics across multiple disciplines and geographies for Duke, GeoEye, RTI, the Department of Homeland Security, State and local law enforcement agencies, NASA-Goddard Space Flight Center, and the Smithsonian Environmental Research Center.