Who is the big-data-enabled specialist?

Joe Ippolito, Senior Project Director
Oceans of Data Institute

Big Data & Analytics EdCon 2014
Las Vegas, Nevada
October 25, 2014
EDC creates learning opportunities for people around the world, empowering them to pursue healthier, more productive lives.
Why the Oceans of Data Institute?

• Workforce imperative
• Educational imperative
• Social imperative

“Decisions based on data and analytics will play an increasingly important role in business and society.” Davenport and Kim (Harvard Business School and KNDU Lab for Analytics Research), 2013

“Get familiar with Big Data now, or face a permanent pink slip” WSJ Apr 9, 2014
Addressing the challenge: Oceans of Data Institute

- Act as a hub for diverse stakeholders
- Develop and test curricula and tools
- Develop and test curricula and tools
- Build a research-based learning progression
- Convene
- Research

EDC OCEANS of DATA INSTITUTE
Developing an Occupational Profile

What are the skills, knowledge and behaviors of a “big data-enabled specialist”?

- Astrophysics
- Telecommunications
- Law Enforcement/Forensics
- Medical Informatics
- Utilities
- Climate Modeling
- Education
- Hazard Analysis
- Marketing
- Analytical Journalism
- Bioacoustics
- Journalism
DACUM

A process for Developing A Curriculum

A process for: (1) Job Analysis
   (2) Occupational Analysis
   (3) Process Analysis
   (4) Functional Analysis

Used by: Vocational-Technical Educators
         Business-Industry Trainers
         Government-Military Agencies
Principles behind our approach

• Engage expert workers
• A job is defined by tasks performed by expert workers
• To be performed well, work tasks require “enablers”
• Identify common cross-cutting work activities
• Identify all work activities linked to an occupational definition
Expert panel- August 14-15, 2014
Work session- August 14-15, 2015
Learning Occupation: The Big Data Enabled Specialist is an individual who wrangles and analyzes large and/or complex data sets to enable new capabilities including discovery, decision support and improved outcomes.

<table>
<thead>
<tr>
<th>Duties</th>
<th>Tasks</th>
</tr>
</thead>
</table>
| **1. Defines the Problem** | 1A. Identifies stakeholders needs  
1B. Determines stakeholders’ needs  
1C. Articulates the question  
1D. Aligns study to organizational goals and objectives  
1E. Translates question into a research plan  
1F. Designs the experiment  
1G. Develops deep domain knowledge of data source  
1H. Discerns whether Big Data is needed to solve problem  
1I. Identifies resources (e.g. experts, software)  
1J. Performs gap analysis  
1K. Assesses risk and bias involved in conducting study/project  
1L. Communicates cost/risk of study to stakeholders |
| **2. Wrangles Data** | 2A. Performs data exploration  
2B. Identifies data  
2C. Creates the data dictionary  
2D. Collects data  
2E. Assesses the extent methods to clean the data  
2F. Maps data across heterogeneous sources  
2G. Identifies outliers and anomalies  
2H. Cleans data  
2I. Transforms data  
2J. Synthesizes data  
2K. Defines new metrics/attributes based on accessible data  
2L. Performs data visualization |
| **3. Manages Data Resources** | 3A. Manages data life cycle  
3B. Conducts capacity planning of resources  
3C. Compiles with legal obligations  
3D. Applies ethical standards  
3E. Identifies tools that may be needed for purchase or modification  
3F. Protects data and results  
3G. Determines access to the data  
3H. Designs ETL workflow  
3I. Implements ETL workflow  
3J. Stores the data  
3K. Upserts data sources |
| **4. Develops Methods and Tools** | 4A. Researches current methods/models  
4B. Extends existing methods/models if possible  
4C. Selects tools/software/programming environment  
4D. Develops new methods/models  
4E. Runs simulations  
4F. Iterates correctness and scalability of methods/models  
4G. Validates methods/models with test cases  
4H. Disseminates methods/models for peer review  
4I. Documents methods/models |
| **5. Analyzes Data** | 5A. Develops analytic plan  
5B. Applies methods and tools  
5C. Conducts exploratory analysis (e.g. identifies anomalies, outliers, bias in sampling; visualizes)  
5D. Estimates results of the analysis (e.g. significance, effect, size)  
5E. Estimates precision and accuracy of answer  
5F. Determines level of confidence in results  
5G. Compares results with other findings  
5H. Answers the question (e.g. insights drawn from results)  
5I. Submits preliminary findings for peer review  
5J. Documents preliminary findings |
| **6. Communicates Findings** | 6A. Selects documentation media (e.g. dashboard, PowerPoint, e-mail)  
6B. Compiles report  
6C. Describes the problem, method and analysis  
6D. Identifies limitations (e.g. data use, data application methods)  
6E. Scopes data narrative based on time, depth and method  
6F. Prepares visualizations  
6G. Guides interpretation  
6H. Articulates conclusions  
6I. Contrasts alternative approaches and past results  
6J. Provides recommendations based on results  
6K. Tells “data story” to convey insight (e.g. talks to CEO) |
| **7. Engages in Professional Development** | 7A. Seeks out mentors  
7B. Stays current on emerging technologies, data, types and methods  
7C. Attends relevant Big Data conferences  
7D. Contributes new knowledge to the field  
7E. Maintains professional library  
7F. Participates in professional organizations  
7G. Mentors others  
7H. Engages in cross discipline training  
7I. Articulates value of Big Data activities to other departments/functions of the organization  
7J. Articulates evolving role of Big Data in supporting organizational goals |
Skills and Knowledge

Knowledge of:
- Algorithms (e.g. machine learning, statistics)
- Analytic Thinking
- Best Practices
- Big Data Analytics
- Communication
- Computer Error Tracking
- Critical Thinking
- Data Analysis
- Data Privacy
- Data Standards
- Data Structures
- DFS (e.g. HDFS, Littoral)
- Distributed Logic
- Distributed Systems
- Distributed Computing Methods
- Domain Field Knowledge (i.e. deep & broad)
- Math
- Metadata Standards
- Network (latency)
- Networking Protocols
- Numerical Methods
- Performance Metrics
- Programming
- Proper Use of Data (e.g. governance)
- Rapidly Evolving Technology
- Landscape
- Recreational Algebra
- Research Methodology
- Resource Allocation
- Scientific Method
- Statistics
- Unstructured Data (e.g. images, text)
- Visualization

Behaviors

A successful Big Data Enabled Specialist is...

- Detail
- Ethical
- Flexible
- Logical
- Open-minded
- Organized
- Patient
- Passionate
- Patient
- Respectful
- Self-directed
- Skeptical
- Socially
- Wiling
to

Skills in:
- Analytical Thinking
- Applying Statistical Methods
- Computational Thinking
- Computer Programming (e.g. R, Python)
- Critical Thinking
- Data Decoding (e.g. UTF, ASCII)
- Data Management
- Data Manipulation
- Data Security
- Database Administration
- Database Programming (e.g. DB, Query data tables)
- Internet Search Strategies
- Intra-company communications
- Keyboarding
- Machine Learning
- Manipulating data tables
- Parallel Programming (e.g. MPI, Hextap), MapReduce
- Problem Solving
- Project Management
- Relational Databases (e.g. Oracle, SQL)
- Research Methods
- Scanning Technical Literature
- Scripting
- Statistical Methods
- Synthetic Thinking (Big Picture)
- System Administration
- Time Management
- Troubleshooting
- Visualization Design
- Working with spreadsheets
- Writing

Equipment/tools/supplies

- Article Server/Search System (e.g. Google scholar, Web of Knowledge)
- Big Data Hardware (e.g. Clusters/ Servers, Networking {Infininode, Fiberlink}, Cloud {AWS, Azure, etc.})
- Collaborative Tools
- Compilers
- Cryptography libraries, programs, protocols
- Data Mining Tools
- Data Security Software and Appliance
- Data Warehouse (e.g. ETL Tools)
- Databases (e.g. SQL, NoSQL)
- Desktop Productivity (e.g. Word proc., Spreadsheet, Slide prog., e-mail)
- File systems (e.g. HDFS, GFS, Hadoop)
- Job Scheduler (e.g. HTCCondor, GridEngine)
- Knowledge Management Tools
- Knowledge Networks
- Map/Reduce (e.g. Hadoop, Spark, YARN, KEPLER)
- Operating Systems
- Personal Hardware (e.g. Desktop PC, Laptop, Smartphone, tablet)
- Programming Packages (e.g. Python, C#)
- Project Management Tools
- QA/ QC Tools
- Simulation Packages
- Slideshows (e.g. PPT, Keynote, Google Slide)
- Source Control Systems (e.g. SVN, Git)
- Statistics Packages (e.g. R, Matlab, SAS)
- Visualization and Analytics Software (e.g. D3, Tableau, Ayasdi)
- Workflow Tools (e.g. Proficiency Builder Edition)

Trends/Concerns

Accelerating data growth leads to fragmentation of ad hoc solutions
Big Data field evolving from individual disciplines to trans-disciplinary melting pot
Demand for Big Data Enabled Specialists is rapidly increasing, while supply of individuals with those skills is not
Difficulty in discovering poorly collected data
Fragmentation of practices and tools exceeds the capacity of training programs and workforce professional development
Growth in government involvement in organizational data practices
Increased need for real time analytics for streaming data
Increased risk to data security due to security breaches
Industry tools stand in contrast to workforce skill levels
More complex statistical results/visualizations are increasingly present in media
Need for ethical safe harbor for data sharing
Proclivity of computing in developing nations creates new challenges
Proclivity of diverse policies on governing data security
Proclivity of practices and internal tools exceeds the capacity of training programs and workforce professional development
Public interest in data literacy is growing
Public understanding of data remains low
Rapid drop in cost, along with rapid expansion in accountability and ubiquity of cloud-computing
Rapid obsolescence of technology and tools
The Big Data Enabled Specialist is transitioning from a technical role to a business driven role
The internet of things creates more data than existing capacity to process
The role of the Big Data Enabled Specialist is not well-defined in organizational culture

Five Years From Now ...

Client base will move to smaller organizations using larger data sets to solve more sophisticated problems
Computability will be on the evening news
Continuous increases in data but deflation of the Big Data hype with a much greater focus on impact and ROI
Data will be collected at even greater scales, yet software/ tool/methodologies still lag behind
Data and analysis will be provided more efficiently and transparently using new technologies and methods
Development of global data retention standards (e.g. safe harbor, templates)
Increase in data driven decision-making
Increase of data will increase solvability of crimes
Less hype and frenzy, and more productivity
Shift from documents/ PDF to interactive data methods and visualizations to insure reproducibility
Using Big Data modeling and capture to change the mode of global tectonic studies from local cases to global monitoring

Profile of a

Big Data Enabled Specialist

DRAFT

NOT FOR DISTRIBUTION

Initially Developed- August 15-16, 2014

Kirk Borne
Professor of Astrophysics and Computational Science
George Mason University
Fairfax, Virginia

Randy Bucciarelli
Programmer/ Analyst
Scripps Institution of Oceanography
UC San Diego
La Jolla, California

Tim Chadwick
Program Director
Jet Propulsion Laboratory
Pasadena, California

Benjamin Davison
Quantitative User Experience Researcher
Google
Boston, Massachusetts

Lucy Dotting
Associate Provost of Planning and Institutional Research
Columbia University
New York, New York

Ryan Kapaun
Law Enforcement Analyst
Eden Prairie Police Department
Eden Prairie, Minnesota

Juan Miguel Lavista Ferres
Principal Data Scientist
Bing Microsoft
Seattle, Washington

Shannon McWayne
Head of Division of Bioinformatics and Computational Biology
Oregon Health & Science University
Portland, Oregon

Jay Parker
Earth Scientist
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Steve Ross
Consultant on Data Quality Control
Corporate Editor
Broadband Communities Magazine
Revere, Massachusetts

Karli Shah
Principal Consultant
Strategic Solutions
Toronto, Canada

Joyce Malyn-Smith, Ed. D.
Boston, Massachusetts

Profile Facilitators

Joseph Ippolito
Cleveland, Ohio

Lucy Dotting
Associate Provost of Planning and Institutional Research
Columbia University
New York, New York

Ryan Kapaun
Law Enforcement Analyst
Eden Prairie Police Department
Eden Prairie, Minnesota

Juan Miguel Lavista Ferres
Principal Data Scientist
Bing Microsoft
Seattle, Washington

Shannon McWayne
Head of Division of Bioinformatics and Computational Biology
Oregon Health & Science University
Portland, Oregon

Jay Parker
Earth Scientist
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Steve Ross
Consultant on Data Quality Control
Corporate Editor
Broadband Communities Magazine
Revere, Massachusetts

Karli Shah
Principal Consultant
Strategic Solutions
Toronto, Canada

Profile Facilitators

Joseph Ippolito
Cleveland, Ohio

Joyce Malyn-Smith, Ed. D.
Boston, Massachusetts
Components of the Profile

- Learning Occupation
- Duties
- Tasks
- Enablers- Skills, knowledge, behaviors, tools/equipment
- Trends/ Future Concerns
The Big Data Enabled Specialist is an individual who wrangles and analyzes large and/or complex data sets to enable new capabilities including discovery, decision support and improved outcomes.
Major work responsibilities - Duties

- Defines the Problem
- Wrangles Data
- Manages Data Resources
- Develops Methods and Tools
- Analyzes Data
- Communicates Findings
- Engages in Professional Development
Constituent work activities- Tasks

• Total of 79

• Bear direct relation to duty

• Order does not signify priority
Enablers, Trends

- Skills
- Knowledge
- Behaviors
- Tools
- Future Trends/ Concerns
Validation of the Profile

• Importance of tasks
• Frequency of tasks
• Importance of knowledge, skills and behaviors
• General comments
Validation survey demographics

- 188 respondents / 100 Completed entire survey
- More than 15 industry sectors represented:
  - 35% IT
  - 11% STEM
  - 10% Law, Public Safety, Corrections
  - 9% Education and Training
  - 8% Health Science and Services
  - 6% Government & Public Administration
  - 5% Finance
Validation results

• Overall the profile has been validated

• 80%+ indicate essential, very important or important on all tasks and nearly all enablers

• Review and possible revisions to finalize document
Skills

• Analytical Thinking (96%)
• Critical Thinking (84%)
• Problem-solving (75%)
• Applying Statistical Methods (74%)
• Data Manipulation (70%)
Knowledge

- Analytic Thinking (89%)
- Algorithms (e.g., machine learning, statistics) (76%)
- Data Modeling (70%)
- Data Structures (7%)
- Best Practices (69%)
- Statistics (69%)
Behaviors

- A problem solver (89%)
- A lifelong learner (78%)
- Willing to question (78%)
- A seeker of patterns (67%)
- Open-minded (67%)
• Increased risk to data security due to security breaches (68%)
• The role of the big-data-enabled specialist is not well-defined in organizational culture (63%)
• Public understanding of data remains low (62%)
• Increased need for real-time analytics for streaming data (61%)
• Demand for big-data-enabled specialists is rapidly increasing, while supply of individuals with those skills is not (59%)
• Exponential growth of data (59%)
Five years from now…

• Increase in data-driven decision making (70%)
• Continuous increase in data but deflation of the big data hype with a much greater focus on impact and ROI (59%)
• Data and analysis will be provided more efficiently and transparently using new technologies and methods (57%)
Questions Driving Our Next Steps With The Profile

- How can we measure how well an individual performs the tasks of the big-data-enabled specialist? (Rubrics/Assessment Framework)

- Can the same process be used to develop profiles of other big data occupations? (e.g. big-data-enabled managers)

- What are the clusters of skills that comprise occupations under the umbrella of the big data enabled specialist? (middle skills development)

- What does a big data career pathway look like?
ODI: Using the Profile to improve educational programs

Working with workforce educators to:

- Design degree and/or certificate programs
- Align curriculum with local and regional industry needs
- Guide performance-based assessments
- As a framework for student portfolios
- Develop authentic real world projects
- Evaluate internships/ work-based learning activities
- Work with advisory committees to identify local and regional employers’ needs and interests
Thank you! Questions?

For more information, contact:
Ruth Krumhansl, Director
Oceans of Data Institute
oceansofdata@edc.org

And visit our website –
http://www.oceansofdata.org

Joe Ippolito
jippolito@edc.org