SKILLS AND KNOWLEDGE

**Knowledge of:**
- Algorithms (e.g., machine learning, statistics)
- Analytic Thinking
- Best Practices
- Big Data Analytics
- Communication
- Concurrency
- Critical Thinking
- Data Modeling
- Data Practices (e.g., HIPAA, SOX)
- Data Security and Privacy
- Data Standards
- Data Structures
- Databases (e.g., SQL, NoSQL)
- Discrete Logic
- Distributed Systems
- Distributed Computing Methods
- Domain/Field Knowledge (i.e., deep & broad)
- Math
- Metadata Standards
- Numerical Methods
- Performance Metrics
- Programming
- Proper Use of Data (e.g., governance)
- Rapidly Evolving Technology
- Landscape
- Relational Algebra
- Research Methodology
- Resource Allocation
- Scientific Methodology
- Statistical Analysis
- Unstructured Data (e.g., images, text)
- Visualization

**BEHAVIORS**

**A successful big-data-enabled specialist is:**
- A choreographer
- A connector of domains/data
- A data lover
- A forecaster
- A mentor
- A multi-tasker
- A problem solver
- A risk taker
- A seeker of patterns
- A storyteller
- A strategic thinker
- Aninker
- Collaborative
- Curious
- Detail oriented
- Ethical
- Flexible
- Inclusive
- Logical
- Open-minded
- Organized
- Passionate
- Patient
- Respectful
- Self-directed
- Skeptical
- Socially aware
- Willing to question

**TRENDS/CONCERNS**

Accelerating data growth leads to fragmentation of ad hoc solutions.
Big data field evolving from individuals to discipline-specific to transparent, collaborative, and driven.

Demand for big data-enabled specialists is rapidly increasing, with supply of individuals with these skills lagging.

Difficulty in discovering poorly collected data.

Exponential growth of data.

Fragmentation of practices and tools exceeds the capacity of training programs and workforce professional development.

Growth of 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 needs.

Insufficient workforce to curate and clean data.

Insufficient workforce skilled in big data.

Lack of access to electric power to run data centers.

More complex statistical results/visualizations are increasingly present in media.

Need for ethical, safe harbor for data sharing.

Faster computing in developing nations creates new challenges.

Faster proliferation of diverse policies on governing data security.

Faster proliferation 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 rise 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.

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 localized problems.

Compute availability will be even the evening news.

Continuous stream in data but definition of the big data hype with a much greater focus on impact and ROI.

Data will be collected at even greater scales, yet software/tools/methods still lag behind.

Data and analytics 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 ensure reusability.

Using big data modeling and capture to change the mode of large tabular studies from local cases to global monitoring.

---

Panel

**Kirk Borne**

Professor of Astrophysics and Computational Science

George Mason University

Fairfax, Virginia

**Randy Buccarelli**

Programmer/Analyst

Scripps Institution of Oceanography

UC San Diego

La Jolla, California

**Tim Chadwick**

Principal Engineer

Dynamic Network Services, Inc.

Manchester, New Hampshire

**Benjamin Davidson**

Quantitative User Experience Researcher

Google

Boston, Massachusetts

**Lucy Dronning**

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 Mcweeney**

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

**Karthik Shah**

Principal Consultant

Strategy Solutions

Toronto, Canada

**Oceans of Data Institute**

**Ruth Knahr**

Director

**Profile Facilitators**

Joseph Ippolito

Joyce Malyn-Smith

---

Suggested Citation:

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