Teaching & Learning with Geoscience Data

Kim Kastens

Work done at Education Development Center & Lamont-Doherty Earth Observatory of Columbia University

Presented at: Summit on Undergraduate Geoscience Education 8 January 2015, Austin, TX
Learning Science from Scientific Data: Why bother?

*Reason #1:* Students can grasp the evidence base that underlies the big ideas of science, rather than having to take these ideas on authority.

*Reason #2:* The world faces tough decisions and society is making some bad decisions. We want to raise up a generation who have the skills and disposition to make decisions based on evidence.

*Reason #3:* Those of our undergraduates who become scientists will need this skill set in their careers.
(A) Unstructured observation with human senses

(B) Student-collected small datasets

(C) Professionally collected large datasets, well-structured problems

(D) Professionally collected large datasets, ill-structured problems

Challenging transition

Consolidating skills

Challenging transition

Consolidating skills

Challenging transition
(A) Unstructured observation with human senses

(B) Student-collected small datasets

(C) Professionally collected large datasets, well-structured problems

(D) Professionally collected large datasets, ill-structured problems

How does this transition happen? And how can it be better supported?

Consolidating skills

Challenging transition

Consolidating skills

Consolidating skills

Thanks to Ruth Krumhansl, EDC & Margie Turrin, L-DEO & Sandra Swenson, John Jay College
Most pre-college data experiences have been with small, student-collected data sets
What is involved in this transition?

Student-collected data

Professionally-collected data

Day in the Life of the Hudson

Kim aboard Joides Resolution, Leg 107
Embodied, experiential grasp of the natural setting and data collection methods

(from School in the Forest powerpoint, http://www.blackrockforest.org/docs/about-the-forest/schoolintheforest/)

(from Using a Digital Library to Enhance Earth Science Education, Rajul Pandya, Holly Devaul, and Mary Marlino)
Dozens of data points

Air temperature at noon from Clement, 2002

Megabytes

Observed Precipit. Anomaly OND 2002
Shaded ONLY for “ABOVE-Normal” & “BELLOW-Normal”
[CAMS_OPI data, courtesy of NCEP/CPC]
Simple, transparent tools and techniques

Sophisticated tools & techniques

Air temperature at noon

Interpret one data set at a time

Multiple data sets with interactions; varying data types
What was the salinity at noon on April 16?
Common sense lines of reasoning. Multi-step chains of reasoning.

Spatial, temporal, statistical reasoning. (Wainwright, 2002)
- Student-collected data
- Embodied, experiential sense of circumstances
- Dozens to hundreds of data points
- Simple, transparent tools & techniques
- Interpret one data set at a time
- “Common sense” lines of reasoning
- Single step causal chains

- Professionally-collected data
- Sense of circumstances from metadata
- Megabytes
- Complex tools & techniques; black boxes
- Multiple data sets and their interactions
- Temporal, spatial, quantitative and other lines of reasoning
- Multi-step lines of reasoning

Day in the Life of the Hudson

http://www.umt.edu/urelations/rview/winter10/brikanarova.html
(1) Use pre-selected snippets of high insight:effort ratio data ("Data Puzzles").

(2) Nest a small student-collected data within a larger dataset.

(3) Ask students to commit to a prediction of what they will see before they start making data visualizations.

(4) Provide an array of candidate hypotheses.
**High insight: effort ratio data snippet**

**Procedure:**

1) Curriculum developer identifies a small snippet of authentic data that embodies an important and widely-taught scientific concept, and develops data visualization(s) that foreground the patterns or relationships emerging from that concept.

2) Students view data visualizations on screen or paper, and answer guiding questions about the system represented by the data (not just about how to decode the data).

3) Students experience a rewarding “Aha! moment” of recognition when they see the process they have previously studied conceptually manifest in real world data.

Decoding question: “What regions of the ocean had the warmest temperatures during April?"

Aha! Question: Propose a hypothesis that might explain why the average temperature in Bordeaux is higher than in Halifax, even though they are at the same latitude.

(from Krumhansl, 2014, EDC Earth Science.)
Procedure:

1) Students collect and interpret a local dataset.

2) (optional) Students from multiple schools combine similar datasets to aggregate a larger sample or span a larger area.

3) Students interpret larger professionally collected dataset(s) which encompass and expand beyond the circumstances of their self-collected dataset.

A Day in the Life of the Hudson:  http://www.ldeo.columbia.edu/edu/k12/snapshotday/
Combine with other school groups’ data to explore variation across space.

Combine with professionally collected data to explore changes through time.

Piermont Pier, NY. Salinity (psu)

(from Turrin, M., & Kastens, K. A. (2010). In Earth Science Puzzles: Making Meaning from Data and http://www.hrecos.org/)
Procedure:

1) Based on either a conceptual model, physical model or computational model, students predict what data from the system under consideration would look like under various conditions.

2) Students examine professionally collected data taken under a range of conditions, looking for the presence or absence of predicted patterns.

**Hypothesis Array**

**Procedure:**

1) Students are provided with text descriptions or sketches of several alternative working hypotheses (the “choice array”) that might depict a process or structure of the system under consideration.

2) Students explore a database of professionally collected data, seeking to assemble evidence in support of one of the hypotheses.

Analyzing and clearly articulating the strategies used by experts..... was not as valuable as providing a visual array of candidate answers.

Four ways to scaffold students’ transition from small, student-collected datasets to large, professionally-collected data bases

(1) Use pre-selected snippets of high insight:effort ratio data (“Data Puzzles”).
(2) Nest a small student-collected data within a larger dataset.
(3) Ask students to commit to a prediction of what they will see before they start making data visualizations.
(4) Provide an array of candidate hypotheses.
(A) Unstructured observation with human senses

(B) Student-collected small datasets

(C) Professionally collected large datasets, well-structured problems

(D) Professionally collected large datasets, ill-structured problems

Challenging transition

Consolidating skills

Learning curve

Consolidating skills

Consolidating skills

Thanks to Ruth Krumhansl & Joe Ippolito, EDC
A methodology for occupational analysis

Premise: experienced and respected practitioners can best define and describe their job or profession

Product:
- Definition of the job/career/profession
- Duties & Tasks
- Knowledge, Skills, Tools & Behaviors
Developing an Occupational Profile

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

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

EDC OCEANS of Data INSTITUTE
# Occupational Profile

## Duties

1. **Defines the Problem**
   - 1A. Identifies stakeholders
   - 1M. Negotiates plan, including deadlines and budgets
2. **Wrangles Data**
   - 2A. Performs data exploration
   - 2M. Writes software to automate tasks
3. **Manages Data Resources**
   - 3A. Manages data lifecycle
   - 3B. Conducts capacity planning of resources
4. **Develops Methods and Tools**
   - 4A. Researches current methods/models
   - 4B. Extends existing methods/models, if possible
5. **Analyzes Data**
   - 5A. Develops analysis plan
   - 5B. Applies methods and tools

## Tasks

<table>
<thead>
<tr>
<th>Duties</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A. Identifies stakeholders</td>
<td>1B. Determines stakeholders’ needs</td>
</tr>
<tr>
<td>1M. Negotiates plan, including deadlines and budgets</td>
<td>1N. Creates requirement document (sign-off)</td>
</tr>
<tr>
<td>2A. Performs data exploration</td>
<td>2B. Identifies data</td>
</tr>
<tr>
<td>2M. Writes software to automate tasks</td>
<td>2N. Documents the process</td>
</tr>
<tr>
<td>3A. Manages data lifecycle</td>
<td>3B. Conducts capacity planning of resources</td>
</tr>
<tr>
<td>3M. Determines access to the data</td>
<td>3C. Complies with legal obligations</td>
</tr>
<tr>
<td>4A. Researches current methods/models</td>
<td>4B. Extends existing methods/models, if possible</td>
</tr>
<tr>
<td>4M. Validates methods/models with test cases</td>
<td>4C. Selects tools/software programming environment</td>
</tr>
<tr>
<td>5A. Develops analysis plan</td>
<td>5B. Applies methods and tools</td>
</tr>
<tr>
<td>5M. Compares results with other findings</td>
<td>5C. Conducts exploratory analysis (e.g., identifies anomalies, outliers, bias in sampling; visualizes)</td>
</tr>
<tr>
<td>5N. Determines level of confidence in results</td>
<td>5E. Estimates precision and accuracy of answer</td>
</tr>
<tr>
<td>6A. Designs the experiment</td>
<td>6E. Translates study into a research plan</td>
</tr>
<tr>
<td>6M. Develops deep domain knowledge of data source</td>
<td>6G. Identifies outliers and anomalies</td>
</tr>
</tbody>
</table>

*Etc.*
How well is the current education system doing at preparing students for the tasks and duties of the big-data-enabled specialist?

- Disciplinary Core Ideas
- Cross-cutting Themes
- Practices of Science & Engineering
  #4: Analyze & interpret data
Comparison of ODI occupational profile tasks with NGSS Performance Expectations

<table>
<thead>
<tr>
<th>DUTIES</th>
<th>TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Analyzes Data</td>
<td>5A. Develops analysis plan</td>
</tr>
<tr>
<td></td>
<td>5B. Applies methods and tools</td>
</tr>
<tr>
<td></td>
<td>5C. Conducts exploratory analysis (e.g., identifies anomalies, outliers, bias in sampling; visualizes)</td>
</tr>
<tr>
<td></td>
<td>5D. Evaluates results of the analysis (e.g., significance, effect, size)</td>
</tr>
<tr>
<td></td>
<td>5E. Estimates precision and accuracy of answer</td>
</tr>
<tr>
<td></td>
<td>5F. Determines level of confidence in results</td>
</tr>
<tr>
<td></td>
<td>5G. Compares results with other findings</td>
</tr>
<tr>
<td></td>
<td>5H. Answers the question (e.g., insights drawn from results)</td>
</tr>
<tr>
<td></td>
<td>5I. Submits preliminary findings for peer review</td>
</tr>
<tr>
<td></td>
<td>5J. Documents preliminary findings</td>
</tr>
<tr>
<td>6. Communicates Findings</td>
<td>6A. Selects documentatio n media (e.g., dashboard, PowerPoint, e-mail)</td>
</tr>
<tr>
<td></td>
<td>6B. Compiles report</td>
</tr>
<tr>
<td></td>
<td>6C. Describes problem, method, and analysis</td>
</tr>
<tr>
<td></td>
<td>6D. Identifies limitations (e.g., data use, data application methods)</td>
</tr>
<tr>
<td></td>
<td>6E. Scopes data narrative based on time, depth, and method</td>
</tr>
<tr>
<td></td>
<td>6F. Prepares visualizations</td>
</tr>
<tr>
<td></td>
<td>6G. Guides interpretation</td>
</tr>
<tr>
<td></td>
<td>6H. Articulates conclusions</td>
</tr>
<tr>
<td></td>
<td>6I. Contrasts alternative approaches and past results</td>
</tr>
<tr>
<td></td>
<td>6J. Provides recommendations based on results</td>
</tr>
<tr>
<td></td>
<td>6K. Tells “data story” to convey insight (e.g., talks to CEO)</td>
</tr>
<tr>
<td>7. Engages in Professional Development</td>
<td>7A. Seeks out mentors</td>
</tr>
<tr>
<td></td>
<td>7B. Stays current on emerging technologies, data types, and methods</td>
</tr>
<tr>
<td></td>
<td>7C. Attends relevant big data conferences</td>
</tr>
<tr>
<td></td>
<td>7D. Contributes new knowledge to the field</td>
</tr>
<tr>
<td></td>
<td>7E. Maintains professional library</td>
</tr>
<tr>
<td></td>
<td>7F. Participates in professional organizations</td>
</tr>
<tr>
<td></td>
<td>7G. Mentors others</td>
</tr>
<tr>
<td></td>
<td>7H. Engages in cross-discipline training</td>
</tr>
<tr>
<td></td>
<td>7I. Articulates value of big data activities to other departments/ functions of organization</td>
</tr>
<tr>
<td></td>
<td>7J. Articulates evolving role of big data in supporting organizational goals</td>
</tr>
</tbody>
</table>

Abundant in NGSS | Potentially (implicitly) abundant in NGSS | Sparse in NGSS | Absent from NGSS
Occupational Profile tasks that are *well-represented* in NGSS

1. Defines the Problem  
   1B. Determines stakeholders’ needs  
   1C. Articulates the question  
   1E. Translates question into a research plan  
   1F. Designs the experiment  
   1G. Develops deep domain knowledge of data source

2. Wrangles Data  
   2D. Collects data

5. Analyzes Data  
   5A. Develops analysis plan  
   5B. Applies methods and tools  
   5D. Evaluates results of the analysis (e.g., significance, effect, size)  
   5H. Answers the question (e.g., insights drawn from results)
Occupational Profile tasks that are absent from NGSS

2. Wrangles Data
   2A. Performs data exploration
   2G. Identifies outliers and anomalies
   2N. Documents the process

3. Manages Data Resources
   3D. Applies ethical standards
   3F. Protects data and results

4. Develops Methods and Tools
   4F. Iterates correctness … of … models

5. Analyzes Data
   5F. Determines level of confidence in results

6. Communicates Findings
   6D. Identifies limitations (e.g., data use, data application methods)
Bottom line:

• It’s a long, complicated pathway to grow a populace that has the skills and disposition to use data as part of their tool-kit when confronted with a difficult question or problem.

• There are effective instructional templates to build on experience with small, student-collected datasets towards proficiency with large, complex, professionally-collected data.

• Big-data enabled professionals value a suite of skills around data quality, data safety, and data ethics that may be missing from today’s students.
For more information: