EDC OCEANS of DATA INSTITUTE

Pervasive and Persistent Understandings about Data

Kim Kastens, 26 January 2014

The Oceans of Data Institute has developed and is testing a hypothesized learning progression toward "data scientist," which involves passing through four domains: (A) children observe the real world with their human senses in an unstructured way, (B) students work with small datasets that they collected themselves, (C) students analyze and interpret large, professionally collected datasets in the context of well-structured problems, and (D) students or professionals analyze and interpret large, professionally collected datasets in the context of ill-structured problems (Figure 1).

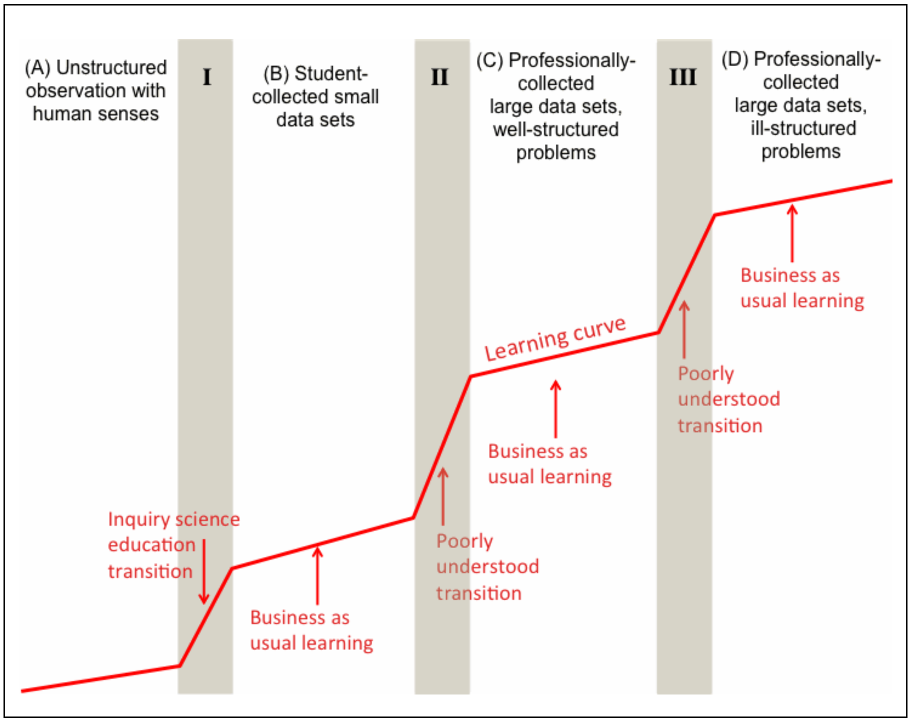
(A) Unstructured observation with human senses
I
(B) Student-collected small data sets
II
(C) Professionally-collected large data sets, well-structured problems
III
(D) Professionally-collected large data sets, ill-structured problems
Learning curve
Inquiry science education transition
Business as usual learning
Poorly understood transition
Business as usual learning
Poorly understood transition
Business as usual learning

Figure 1 sketches the broad outline of a sequence of stages that could culminate in an individual who can use large, professionally collected datasets to solve the sort of ill-structured, complex problems that abound in adult life.

Implicit in the declaration that this is a "learning progression" is the implication that there are skills and/or understandings that build across this entire trajectory.

Stating that the model of Figure 1 is a "learning progression" connotes that there are skills and/or understandings that persist and build across this entire trajectory, from unstructured observation with the human senses all the way through to large datasets and ill-structured problems. Articulating what these understandings might, how they manifest in each of the domains of Figure 1, and how to carry these understandings across each of the transitions of Figure 1, is part of the Ocean of Data Institute's strategic initiative on data in education. To give the reader a sense of our thinking, here are some examples:

- *Acquiring data from a complex world necessitates choices and trade-offs. You can't measure everything, all the time. As a consequence, every dataset is a subset or sampling of the referent system, leaving out as much or more than it includes (Table 1).* An individual who is advancing appropriately across the trajectory of Figure 1 is aware of the limitations of data, and this awareness persists whatever the topic and whatever the data type.

*Table 1: All datasets are incomplete**

	(A) Unstructured observation using human senses	*(B) Small, student-collected datasets*	*(C) Large, professionally collected datasets, well-structured problems*	*(D) Large, professionally collected datasets, ill-structured problems*
Earth Sciences	If we dig a hole in the sand at the beach to see what is in there, we never get to the bottom; there is always a deeper layer to sample.	A student dataset of phases of the moon has data gaps on cloudy nights.	In a professionally collected dataset of ocean salinity, the data is interpolated between research vessel sampling stations.	All available geological and geophysical data together give only an indirect and imprecise indication of the amount of petroleum in an oil reservoir.
Physical Sciences	Humans can see only a limited range of wavelengths.	In a dataset of acceleration of balls on ramps, only a certain number of ball sizes and masses are examined, and only a certain range of ramp heights.	In a professionally collected dataset of per capita electricity use, the data only go back to a certain date, leaving out the early years of the industrial revolution.	All available data together give only an approximate estimate of the risk of a nuclear power plant accident.
Life Sciences	When you look at a tree, you see only about half of it above ground; an equally important part is underground.	When students grow plants with varying amounts of sunlight and water, they typically measure them once a day; any variations in growth rate throughout the day/night cycle are not in the data.	Genome sequences are available for some species but not for others.	Tests of the efficacy and safety of a new drug are done on a sample population; other patients may respond differently

*The incompleteness of datasets as representations of the referent system can manifest in several ways:

- The data/observations cover only a limited spatial area or volume (e.g., the above-ground part of the tree, the finite depth of the hole)
- The data/observations are made at discrete locations, whereas the phenomenon of interest is continuous (e.g., the ocean salinity data).
- The data/observations cover only a limited interval of time (e.g., the energy use time series)
- The data/observations are made at discrete times, whereas the phenomenon of interest is continual (e.g., the growing plants).
- The data/observations cover only a limited set of the possible conditions (e.g., the height of the physics ramp).
- The data acquisition sensor or instrument (including human senses) is only sensitive or responsive over a certain range of the observed phenomenon (e.g., the wavelengths of light observed by human visual system).
- The data/observations sample only selected individuals or groups out of a larger population (e.g., the drug test).

- *Data can be used to make inferences about events of the past.* When an event or events (*sensu* Shipley, 2009) occurs it may leave a trace (Cleland, 2001, 2002), a physical manifestation in the arrangement of molecules. Human senses or instrumental sensors may be able to detect this trace, even long after the event. By looking at patterns in traces, and combining these observations with reasoning based on knowledge of the workings of the relevant system, humans can make inferences about the event. Reading from left to right across Figure 1: (A) from the observation of small round droppings, one can infer that a deer was present; (B) from analyzing the products of a student-initiated chemical reaction, one can infer that the reaction took place; (C) from the distribution of genes across a population, one can infer that the population passed through an evolutionary bottleneck; (D) by combining multiple lines of evidence, each fragmentary and incomplete, scientists were able to infer that a huge extraterrestrial impactor struck the Earth 65 million years ago.

- *Events leave traces in the real world, and by looking at the traces, humans can sometimes make inferences about the events. Some forms of scientists' data attempt to make permanent and visible traces that would otherwise be ephemeral or invisible.* For example: big rainfall events in a watershed dump a spike of fresh water into a river and can change its chemistry. For a tidal river like the lower Hudson, that means a decrease in salinity. If the scientist or science student personally experiences the big rainfall event and then is on site to sample the salinity change, then he/she can make a causal inference about the connection between then. But both the rainfall spike and the river chemistry change are ephemeral. A time series graph made with data from an environmental monitoring station can make a permanent record that scientists can contemplate and interpret at leisure (Figure 2), without needing to be on site while the event is happening.

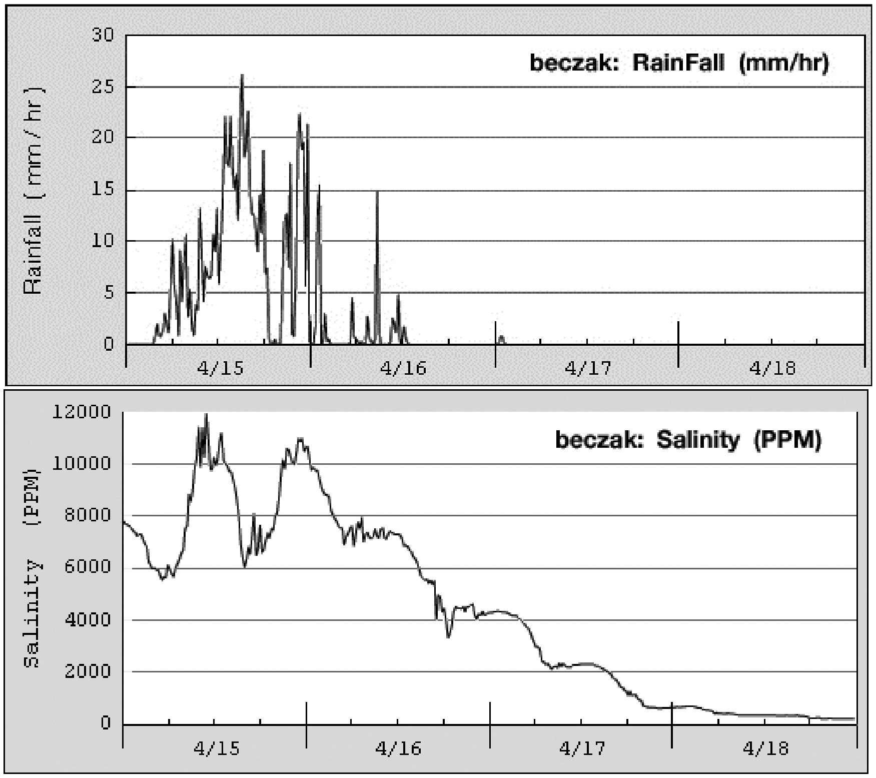

Figure 2: A spike of rainfall in the Hudson River watershed (upper) leads to a decrease in the salinity of the river over the next several days. The scientists' system of recording and representing these phenomena make permanent traces of these ephemeral events.

(From Kastens & Turrin, 2010.)

- *A single observable parameter can reflect multiple processes or influencers.* In working with real data from natural or human systems, is not generally the case that a single causal process or phenomena controls the parameter that is being measured or observed. It is often the case that several causes or influencers may contribute to the observable or measurable signal, sometimes on different time scales. For example, salinity measured in an estuary depends on the upstream/downstream position of the measurement, the phase of the tide, the amount of rainfall on the watershed in recent days to weeks, and the water usage patterns of the humans living in the watershed. Likewise, heart rate of a mammal depends on body size, fitness level, and extent of recent exertion.

- *Data and observations are useful for answering scientific or practical questions about the represented system and solving problems within the represented system.* Reading from left to right across Figure 1: (A) observations with the human senses can answer the question of whether an apple is ripe for picking; (B) student- collected data in a citizen science project can help solve the problem of which technique is most effective for preventing the regrowth of invasive plants; (C) professionally collected data can answer the question of whether global climate is changing; (D) census data are used in solving the problem of where to draw the boundaries of congressional districts. When faced with a tough problem or a challenging question, an individual who is advancing appropriately across Figure 1 is inclined to seek out and use relevant data.

Note that these meta-understandings carry across topics and across disciplines (from Earth science to life sciences and physical sciences), as well as across the trajectory of Figure 1. Students emerging from a data-infused K-12 science education should know these sorts of meta-understandings at an explicit, intellectual level, but they should also have internalized them as habits of mind to be applied automatically when contemplating a new dataset or new data type.

References Cited

Cleland, C. (2001). Historical science, experimental science, and the scientific method. *Geology, 29*, 987-990.

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science, 69*, 474-496.

Kastens, K. A., & Turrin, M. (2010). *Earth science puzzles: Making meaning from data*. Washington, DC: National Science Teachers Association.

Shipley, T. F. (2008). An invitation to an event. In T. F. Shipley & J. M. Zachs (Eds.), *Understanding events: From perception to action* (pp. 3-30). Oxford: Oxford University Press.