Big Data in Science & Problem Solving

Hunter Heaivilin

## Big Data from the Ground Up

This post will provide 1) an outline of the concept of big data, 2) how big data is changing the way we approach problem solving, 3) how the SOC-HUB fits as a big data project, and 4) an example of how big data is used for soil organic carbon assessment.

## What is big data?

“Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.”  
- [De Mauro, Greco, and Grimaldi, in What Is Big Data? A Consensual Definition and a Review of Key Research Topics (2015)](https://doi.org/10.1063/1.4907823)

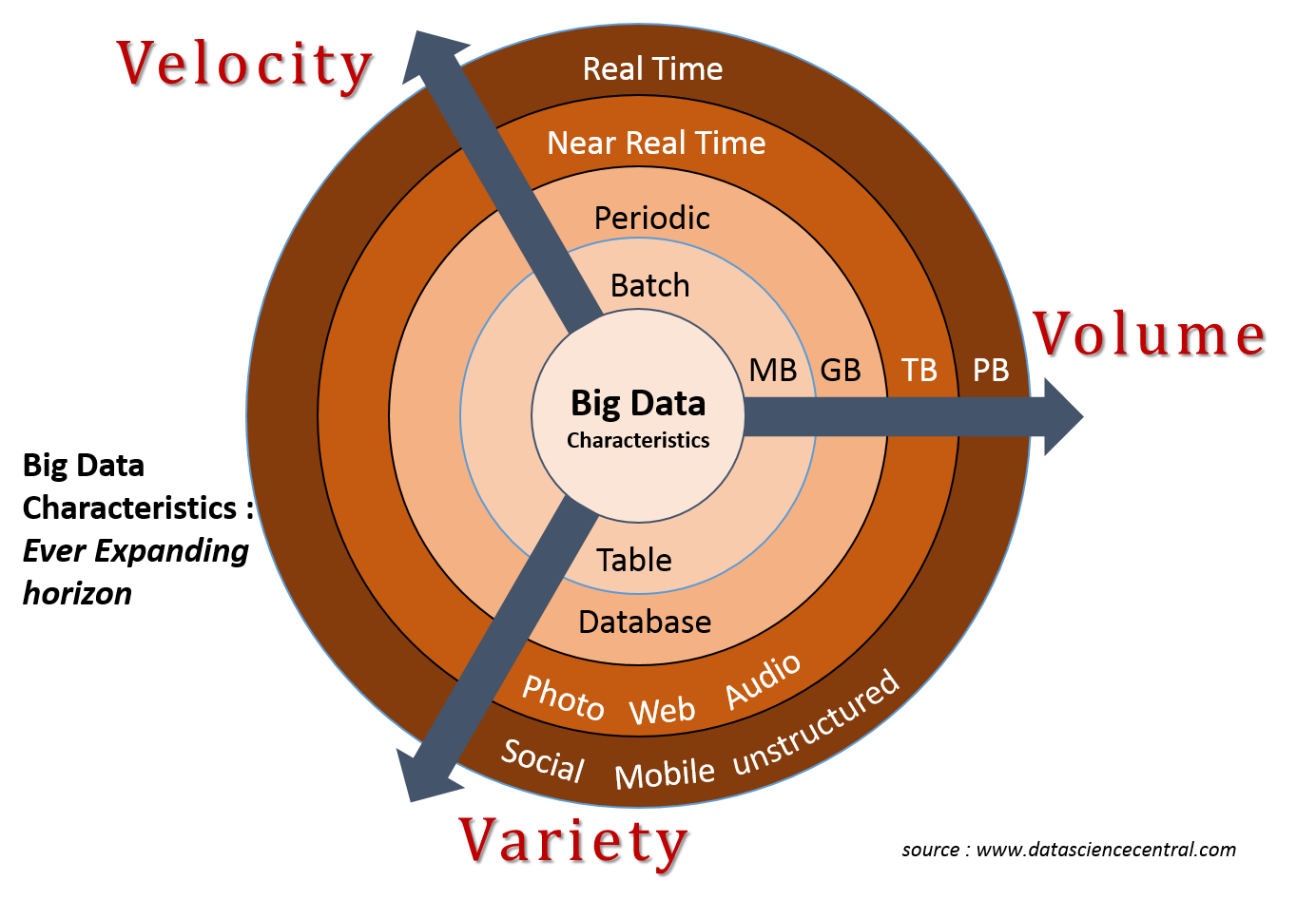


Growth in human population and the way we interface with technology have increased the amount of human-generated data produced. Machine-generated data is produced more and more as machines increasingly are able to operate independent of human intervention. These factors have dramatically increased the amount, speed and kinds of data created, while also requiring advances in how this data is processed. Together, the data and and the processing now compose the core of the concept of ‘big data’.

Unifying definitions of novel and developing ideas are hard to come by, but as [big data evolved conceptually over the past two decades (Diebold, 2012)](https://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf) theoretical overlaps have helped outline the contours of big data.

[Laney’s 2001 [pdf]](https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf) work, while not the first to frame an approach, identified ‘Volume’, ‘Velocity’ and ‘Variety’ as crucial factors of large format (originally termed “3D”) data management. The 3 V’s, as they became known, are now roughly defined as follows:

* **Volume**: the size of data used or generated
* **Velocity**: the speed of data creation, processing,
* **Variety**: the spread of types and structures of data



<https://hadoop.cesga.es/workshop/img/3v.png>

In the years since Laney’s work many have added more V’s to the list. For example, as early as 2013 [IBM added ‘Veracity’](https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/)(with an [infographic](http://www.ibmbigdatahub.com/infographic/four-vs-big-data) to prove it) to highlight the importance of uncertainty, error, or other issues the data. By [2016](https://web.archive.org/web/20160114004105/http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data), ‘Value’ was being included by IBM ([of course with another infographic](http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data)). As far as V’s go it seems any number greater than two is fine, be it [seven](https://www.impactradius.com/blog/7-vs-big-data/), [ten](https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx), or even [forty-two](https://www.elderresearch.com/company/blog/42-v-of-big-data).

[Ward and Barker’s 2013](http://arxiv.org/abs/1309.5821) survey of big data definitions yielded a process and technology focused summary, stating that “Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.”

[De Mauro et al’s 2015](https://doi.org/10.1063/1.4907823) meta-analysis of big data definitions, quoted at the beginning of this post, includes ‘Value’ as a fourth V as well as the technology and methods focus encompassed by Ward and Baker (2013). They define value as transforming information into actionable knowledge with economic impacts. While not all big data projects have an economic outcome focus IBM has, unsurprisingly, [jumped on board](http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters) with adding this V to their lists.

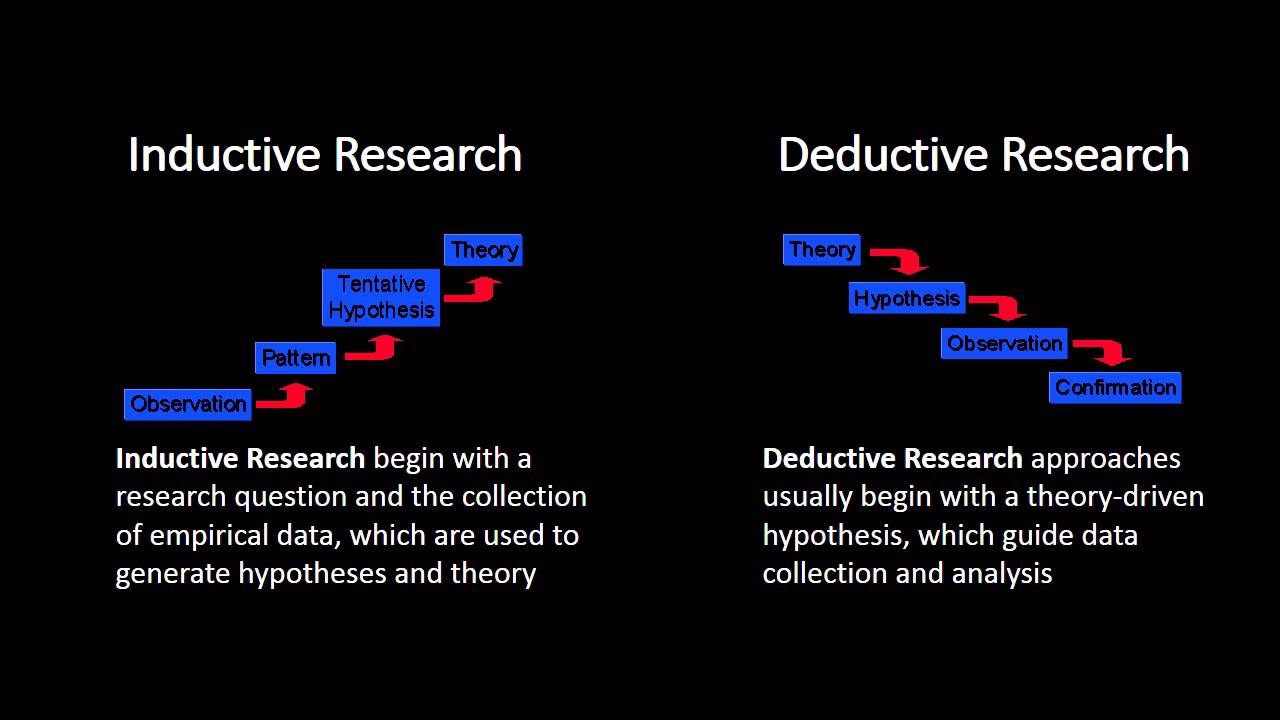
## Big data and Problem Solving

“All models are wrong, but some are useful”  
-[George Box, in Science and Statistics (1976)](https://doi.org/10.2307/2286841)

A decade ago big data was heralded by some as both the [end of theory (Anderson, 2008)](https://www.wired.com/2008/06/pb-theory/) and centuries of the hypothesis-driven scientific method. Since then neither theory nor hypothesis testing have fallen by the wayside, though [similiar discussion continues](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766450/). Part of these [debates](https://www.sciencenews.org/blog/context/why-big-data-bad-science) contrast inductive and deductive reasoning and research.

### Inductive vs Deductive Approaches

Not new concerns, nearly three centuries ago, Hume outlined what has become known as the Problem of Induction. In broad strokes, the problem is that while we constantly use previous experiences to make judgments about future experiences, inductive processes offer insufficient justification for causal conclusions. To wit, the past can help us to make a good guesses about the future (probability) but we can never be sure of exactly what it will hold (certainty).



Deductive research works to address this issue by starting with generally accepted theory and progressing through hypothesis development and experimentation to produce specific observations that can be confirmed with certainty.

By contrast, inductive research begins with observations to identify patterns from which tentative hypotheses and generalized theories can be developed.

The problem of induction highlights that these generalized theories are impossible to verify. Additionally, small observation sizes weaken the probability of generalized theory to have broader value. The disruptive potential of big data, as laid out by Anderson ([2008](https://www.wired.com/2008/06/pb-theory/)), is that when observation sizes become increasingly vast it calls into question if correlation, albeit massive, is sufficient as a scientific output.

To some degree this mimics the distinction between [confirmatory and exploratory research (Jaeger and Halliday, 1998)](http://www.jstor.org/stable/3893289). While deductive research aims to confirm a hypothesis through specific (and logically defensible?) observations, exploratory research parses observations to discern patterns and produce hypotheses to be tested.

[Google’s Flu rate prediction bungle](http://science.sciencemag.org/content/343/6176/1203) provides an interesting case of how bias, sampling, and other errors come easily into play when the gleam of powerful correlation trumps the use of generally accepted scientific practice.

### Data-Driven Science

For a long time the scientific method has dictated that study should follow a distinct process to be valid. While fidelity and replicability are important, this also led to questioning of the scientific validity of important work such as human genome mapping project since it was not hypothesis fueled (i.e., deductive or confirmatory) but more so an exploratory (i.e., inductive) data generating process ([Carroll and Goldstein, 2009](https://www.nature.com/articles/nmeth0409-237).

Now, however data-driven science has begun to lose its stigma; both in the sense of data-driven research that generates data of potential use (see ‘omics’ research review in [Carroll and Goldstein, 2009](https://www.nature.com/articles/nmeth0409-237)) as well as using big data exploration to drive further exploration or produce initial hypotheses that can be tested using a more deductive approach.

### Enter Machine Learning

“Machine learning is about predicting the future based on the past.”  
-[Hal Daume III, in A Course in Machine Learning (2017)](http://ciml.info/)

The multiple challenges of processing big data, from high dimensionality and massive sample size, require novel approaches to analysis ([Fan, Han, and Liu, 2014](https://arxiv.org/abs/1308.1479)). [Machine Learning](http://news.codecademy.com/what-is-machine-learning/) has become an increasing popular approach. Machine learning is a subset of artificial intelligence that uses pattern recognition to predict outcomes based on analysis of and training with input data. This involves giving computers (machines) the ability to improve on a task (learn) without being explicitly programmed along the way. The goal is to develop the machine capacity to generalize correctly beyond sample training data (input) to produce increasingly accurate predictions (outputs). Ideally, the input data should have little noise and much volume.

As mentioned above, for Ward et al., the choice processing techniques can be as useful in defining a big data project as the qualities of the data itself (e.g., size, structure, etc), as big data will often require novel or specific advanced processing approaches. Machine learning is an example of an advanced processing approach that can be used for big data analysis.

### So, are we there yet?

Is the [SOC-HUB](https://powellcenter-soilcarbon.github.io/SOC-Hub/) a big data project? In light of the information on the above, lets look at a few questions to help determine if we’re wading into big data.

**First, is the SOC data charaterized well by the definitions offered above, namely the 3 V’s?**

The **volume** of data is currently limited to information held in previously published papers or being produced by ongoing research.

The **velocity** of the data is the speed at which people (read: graduate students) will be be able to parse academic papers, enter the data into the proper template, and submit that data into the hub. As the userbase isn’t likely to reach that of a social media platform, SOC-HUB data velocity will be low.

The **variety** of data will be somewhat limited to the structure of the templates used. Meaning that while initial data, housed in research articles, may be unstructured and in various formats, the data entered into the SOC-HUB will be structured to meet the needs of later processing.

**Second, will data processecing use or require techniques that are specialized to address the complexities of big data?**

At current there are opportunities to use machine learning and the existing soil carbon datasets to expire the possibility of improved modeling. The SOC-Hub is intended in part to parse articles and structure their findings into a format that can be used by advanced data processing techniques.

### What might a soil carbon big data project look like?

In early 2018, [Flathers and Gessler](https://dl.sciencesocieties.org/publications/jeq/articles/0/0/jeq2017.08.0318) published a paper that “describes a framework of input data, processing code, and outputs designed around the concept of modeling SOC for the cereal grain producing region of the northwestern United States”. Briefly, their approach uses mutliple sources of soil organic carbon data along with other geospatial data (e.g., climate variables, topography, crop land cover, etc) to model and map SOC.

Their work outlines a process that brings together a **variety** of data (" produced by a diversity of organizations at different times and different spatial and temporal scales, using varied units of measurement,“) of pretty good **volume** (”∼180 GB“) that has a processing **velocity** that can vary as new data becomes avaialble or other input variables are explored. Another overview, very simply explained, can be [read here [pdf]](https://www.reacchpna.org/sites/default/files/tagged_docs/6d.3.pdf).

While the SOC-HUB is not currently a ‘big data’ project per se, the data that is being collected and processed is invaluable for future research that engages big data to help understand the drivers and dynamics of soil organic carbon.

#### References

* Anderson, Chris. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” Wired Magazine 16, no. 7 (2008): 16–07. <https://www.wired.com/2008/06/pb-theory/>.
* Daumé III, Hal. (2012). A course in machine learning. Self-published. <http://ciml.info/>
* De Mauro, Andrea, Marco Greco, and Michele Grimaldi. “What Is Big Data? A Consensual Definition and a Review of Key Research Topics.” AIP Conference Proceedings 1644, no. 1 (February 9, 2015): 97–104. <https://doi.org/10.1063/1.4907823>.
* Diebold, Francis X. “On the Origin(s) and Development of the Term ‘Big Data.’” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, September 21, 2012. <https://papers.ssrn.com/abstract=2152421>.
* Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314. <https://arxiv.org/abs/1308.1479>
* Flathers, E., & Gessler, P. E. (2018). Building an Open Science Framework to Model Soil Organic Carbon. Journal of Environmental Quality. <https://dl.sciencesocieties.org/publications/jeq/articles/0/0/jeq2017.08.0318>.
* Green-Lerman, Hilary. “Ask a Data Scientist: What’s Machine Learning?” Codecademy News, February 16, 2018. <http://news.codecademy.com/what-is-machine-learning/>.
* Hume, David. A Treatise of Human Nature, 2003. <https://www.gutenberg.org/ebooks/4705>.
* Jaeger, Robert G., and Tim R. Halliday. “On Confirmatory versus Exploratory Research.” Herpetologica 54 (1998): S64–66. <http://www.jstor.org/stable/3893289>.
* Laney, Doug. “3D Data Management: Controlling Data Volume, Velocity and Variety.” META Group Research Note 6, no. 70 (2001). <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> [pdf]
* Mazzocchi, F. (2015). Could Big Data be the end of theory in science?: A few remarks on the epistemology of data‐driven science. EMBO reports, 16(10), 1250-1255. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766450/>
* Ward, Jonathan Stuart, and Adam Barker. “Undefined By Data: A Survey of Big Data Definitions.” ArXiv:1309.5821 [Cs], September 20, 2013. <http://arxiv.org/abs/1309.5821>.