Big Data in Science & Problem Solving

Hunter Heaivilin

## Using Big Data in Problem Solving

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

testsuper235 and sub45



1. **“Static tag cloud visualization (word cloud) of key terms appearing in abstracts of Big Data-related papers.”** via 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>.

## What is big data?

As with many emerging concepts, a uniform definition is fleeting:

“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.”

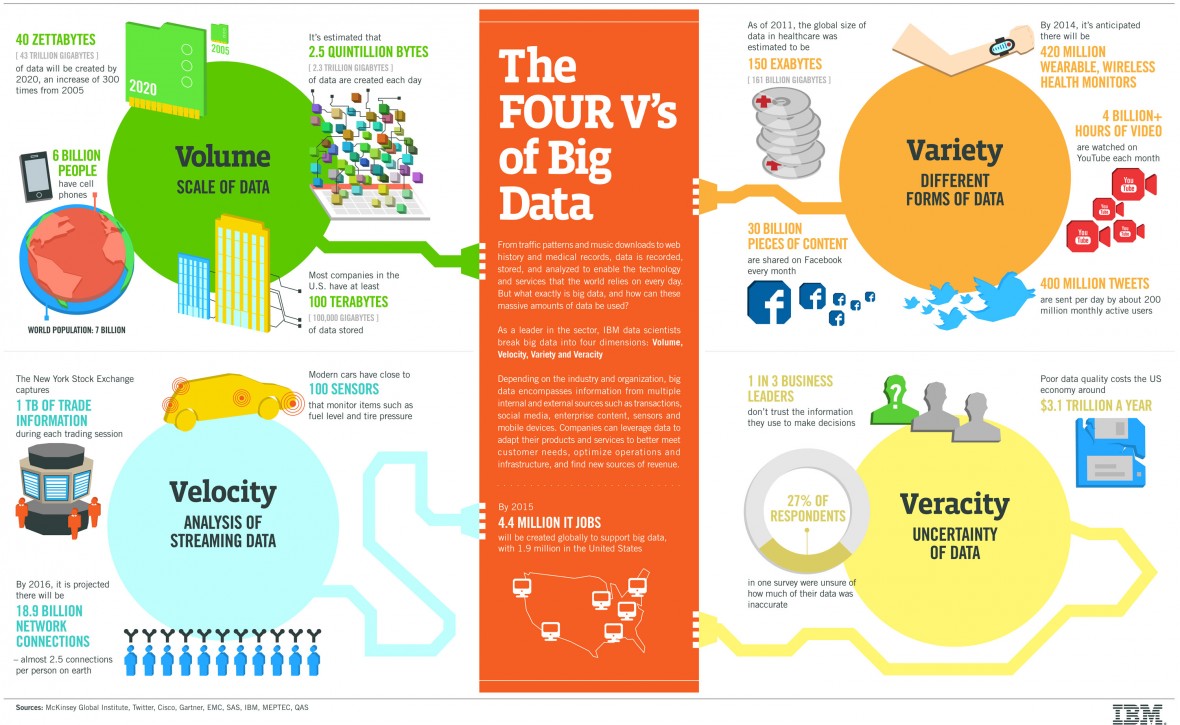
1. De Mauro, Andrea, Marco Greco, and Michele Grimaldi. “What Is Big Data? A Consensual Definition and a Review of Key Research Topics.” AIP Conference Proceedings 1644, no. 1 (February 9, 2015): 97–104. <https://doi.org/10.1063/1.4907823>.

“Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques includ- ing, but not limited to: NoSQL, MapReduce and machine learning.”

* 1. 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>.

However, most like to use some set of V’s [3,4] in the definition. Most often:

* Volume: the size of data generated
* Velocity: the speed of data creation
* Variety: the spread of types of data
* Veracity: the uncertainty, error, or other issues the data



IBM’s Take

### Why big data now?

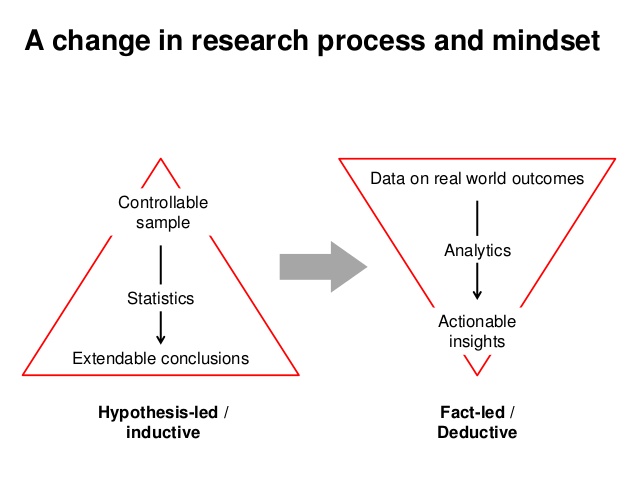
Machines generate data at this point, more than people

How is it different from small data?

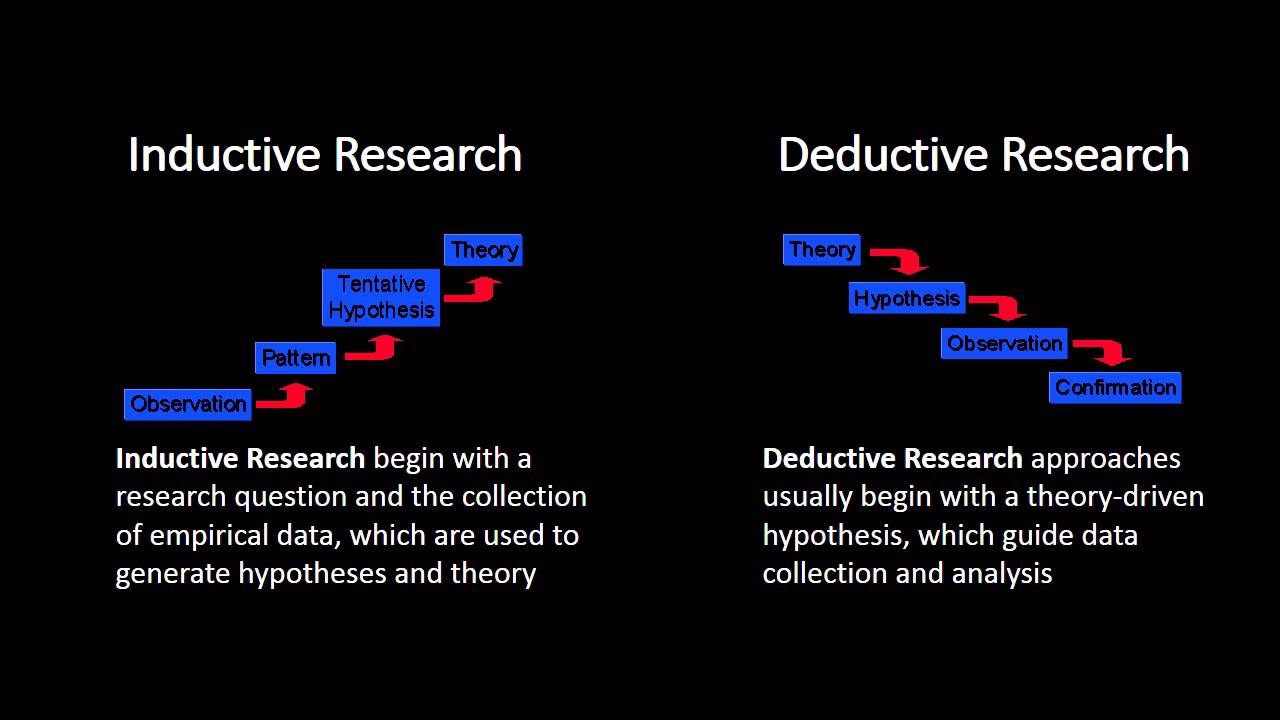
## Big data and problem solving

A decade ago big data was heralded by some as the [end of theory](https://www.wired.com/2008/06/pb-theory/) and [displacing the scientific method](find%20a%20link%20about%20this). Needless to say this was [not recieved well](find%20a%20paper%20about%20the/a%20response%20to%20Anderson) by all.

### Inductive vs deductive approaches



Different Strokes



different methods

**possibly connect to SOM issues in the past?**

* The Problem of Induction - Hume

The [SOC-Hub](link%20to%20mainpage) aims to use mass data entry and points to …

### Machine Learning

The volume of data and the technological requirements to process it have led to novel approaches to analysis like Map Reduce and Machine Learning. Machine learning can …

### Data-Driven Investigation

## Problems in Big Data solutions

#### Bias

There is not uniform access to the systems that generate big data, and thus there will almost inevitably be a slant in any analysis performed on the data. This is in part addressed by good understanding of how the data is generated and the steps involved in processing.

#### Privacy

#### Correlation over causation?

[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 generally accepted scientific practice is left by the wayside in preference of the power of correlation.

## References/Figures

This sentence will appear smaller.

bigger is it?

**adsfasd**

testadf

1. 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>.
2. 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>.

“The Four V’s of Big Data.” IBM Big Data & Analytics Hub. Accessed February 6, 2018. <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>. 3. Laney, Doug. “3D Data Management: Controlling Data Volume, Velocity and Variety.” META Group Research Note 6, no. 70 (2001).

4.40

**underlined text sample**

A [linked phrase](http://example.com/). At the bottom of the document:

Images

test out this thang