**Data deluge** /deljuːdʒ/

* We create 2.5 exabytes [1018] of data every day, with 90% of the data in the world created in the last two years alone...
* Every hour, Wal-Mart **handles** 1 million transactions, feeding a database of 2.5 petabytes [1015], 170 times the data in the Library of Congress.
* The entire collection of the junk mail delivered by the U.S. Postal Service in one year is equal to 5 petabytes, while Google p…………es that amount of data in just one hour.
* 5 exabytes of data would c……….n all words ever spoken by human beings on earth
* The first gigabyte disk drives appeared in 1982, w………ing more than 100 kilograms; now terabyte drives are **consumer equipment**, and a 32 GB microSD card weighs about half a gram.
* Whether you look at bits per gram, bits per dollar, or **raw capacity**, storage availability has grown faster than CPU speed
* The **bulk** of information created nowadays can be called “**data exhaust**,” in other words, “the digitally **trackable** or storable actions, choices, and preferences that people g……………. as they **go about** their daily lives.

**Four Vs definition of big data**

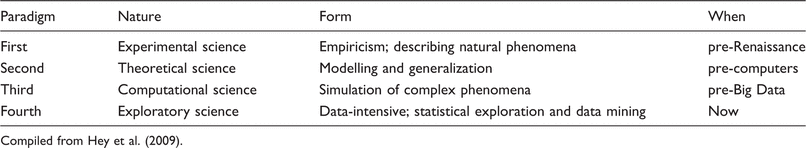
* **V**……………… (the amount of data): data v…………….s continue to increase …. an unprecedented **rate**. It’s still difficult to define “how big”: everyone can agree that whatever is considered “high v……………” today, will be even higher tomorrow;
* **V**……………. (different types of data and data sources): multiple data types, including structured, semi-structured and unstructured data. With the explosion of sensors, smart devices and social media technologies, data is being generated in **countless** forms, including text, web data, tweets, sensor data, audio, video, click streams, log files, and more;
* **V**……………..(data in motion): the speed at which data is created, processed and analyzed continues to accelerate. Higher v……………. is due to both the real-time nature of data creation, and the need to incorporate streaming data into business processes. Today, data is generated at a rate that is impossible for traditional systems to store and analyze.
* **V**……………..(data uncertainty): the level of **reliability** associated with certain types of data. High data quality is an important big data challenge, but even the best data **cleansing** /klenz/ methods cannot r…………….. the unpredictability of some data, like the weather or a customer’s buying decisions. It is necessary to **acknowledge** and plan for uncertainty.

**Data taxonomy** (UNECE[[1]](#footnote-1))

* **Human-sourced information**: subjective r……….d of human experiences, previously in books and works of art, and later in photos, audio, video. Now **digitized** and electronically stored everywhere from tweets to movies. (social networks growth);
* **Process-mediated data**: business processes data highly structured and include transactions, reference tables relationships. (RDBMS);
* **Machine-generated data**: the o…………t of sensors and machines, computer logs. Well-structured nature is **amenable to** computer processing, but its size and speed is often beyond traditional approaches (Internet of Things). As sensors **proliferate** and data volumes grow, this is becoming an **increasingly** important component of the data processed

**A paradigm shift in the production of knowledge**

* a paradigm **refers to** an accepted way of investigating the world and synthesizing knowledge common to a majority of researchers in a discipline at any given moment in time
* paradigm shifts occur because the dominant **mode** of science cannot **account for** particular phenomena or answer key questions and demands the formulation of new ideas



* **top-down paradigm**
  + plan what data you want to collect
  + try to provide explanations behind phenomena
  + r………… on probability theory to support your results
* **bottom-up paradigm**
  + **explore** available data to find information value
  + trust big data: you don't need to understand or explain what is going on
  + **avoid** data aggregation
* Thanks to big data, "there is now a better way. Petabytes allow us to say: ‘Correlation is enough.’ We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing **cluster**s the world has ever seen and let statistical algorithms find p……………s where science cannot … Correlation **replaces** causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." (Anderson, 2008)
* “Scientists no longer have to make **educated guesses**, construct hypotheses and models, and test them with data-based experiments and examples. Instead, they can **mine** the complete set of data **for** p……………s that **reveal** effects, producing scientific conclusions without further experimentation." (Prensky, 2009)

**A critique of “the new empiricism” based on inductive science**

* Although big data **seek to** be exhaustive, **capture** a whole domain and provide full resolution, it is both a representation and a sample and it is subject to **sampling bias**;
* Big data do not **arise from** nowhere. New analytics may present the illusion of automatically discovering **insights** without asking questions, but the algorithms used most certainly did **arise** and were tested scientifically for validity and veracity;
* Inanimate data can never speak for themselves. Making sense of data is always **framed** – data are examined through a particular **lens** that influences how they are interpreted. Correlations between variables within a data set can be **random** in nature and have no or little **causal** association, and interpreting them as such can produce serious ecological fallacies[[2]](#footnote-2)
* The idea that data can speak for themselves suggests that anyone with a reasonable understanding of statistics should be able to interpret them without context or **domain-specific** knowledge. However, without **subject matter experts** available to **articulate** problems in advance, you often get poor results.  Data need to be interpreted in context and through **domain-specific expertise**

**sources**: Rob Kitchin, *Big Data, new epistemologies and paradigm shifts* (2014); ESTP Training Course: *Introduction to Big Data and its Tools*

1. United Nations Economic Commission for Europe [↑](#footnote-ref-1)
2. **If a study shows that people who wear glasses have above average intelligence, assuming that everyone who wears glasses is intelligent is an example of ecological fallacy.** The fallacy occurs when group data is used to draw conclusions about individuals. [↑](#footnote-ref-2)