High Level Summary

Concepts formulated and worked in Volume 1 Definition — first half.

1. Identified Big Data in terms of its usefulness not actionable

Useful is more grounded and without legal connotations

1. Inserted the domain of linguistics

Big Data is particularly suited to “wide columns” or documents

1. Foregrounded the concept of public welfare to identify the scope of potential benefit and its inherent inclusiveness
2. Replaced lifecycle with life

Lifecycle as a concept does fit the nature or outcome of Big Data.

Is Big Data a series of cycles or one long unfolding variation.

Lifecycle refers to a view from the data analyst or scientist’s perspective. Data life refers to the nature of data itself.

1. Replaced “raw” data with “data in the state in which it was produced”.

This view has important consequences for governance, ownership and privacy.

1. Clarified the way Big Data transforms data.
2. Identified the Big Data phenomenon of inclusion.

Inherent in massive datasets.

1. Identified the phenomena of appearance and recognition as central to Big Data.

There are others.

Below is the comment template.

The Definitions Volume 1 with changes highlighted is also supplied.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Item #** | **Type** | **Page #** | **Line #** | **Section** | **Comment (with rationale)** | **Suggested Change** |
| 1 | g | iii | NA |  | Keywords: Big Data Definitions | insert |
| 2 | e | vii | 7 | Ex Sum | provide | enable |
| 3 | g |  | 7 | “ | well governed and useful | insert |
| 4 | g | vii | 9 | “ | linguistics | Insert after computer science, |
| 5 | g | vii | 9 | “ | useful knowledge from knowledge and from data including for the economy and public welfare. | Insert after “obtain” |
| 6 | e | 4 | 164 | 2 Big Data Definitions | useful | Insert after of |
| 7 | g | 4 | 180-182 | “ | In the evolution of knowledge looked at through the prism of data systems, there have been a number of times when the need for efficient, cost effective, well governed and useful data analysis has forced a change in existing technologies. |  |
| 8 |  | 4 | 193 | “ | revolution | Evolution (something of a minor point) |
| 9 | g | 4 | 195-196 | “ | including…usefulness and applicability for public welfare. | Insert after |
| 10 | g | 4 | 205 | Big Data Definitions | The value of each of these characteristics including their interactions with public welfare and economy influences the overall design of a Big Data system, resulting in different data system architectures or different data process orderings to achieve needed efficiencies and public welfare. |  |
| 11 | g | 5 | 210 | “ | and useful | Insert after efficient |
| 12 |  | 5 | 232 | “ | How much uncertainty is increased by these loosely coupled resources and processes including those that follow below needs to be explored and addressed as this new paradigm unfolds. | Insert after systems. On line 232 |
| 13 | g | 5-6 | 253-256 | “ | However understanding Big Data will require a broader viewpoint than a self-referencing viewpoint can contain. Not only does Big Data carry the characteristics of volume, variety, velocity and variability but it is *because* of these characteristics that a new inclusiveness has been discovered (emerged). | clarification |
| 14 | g | 6 | 267 | “ | and benefit | Insert after analysis |
| 14.1 | G | 6 | 270 | “ | inclusive | Insert after “of” |
| 15 | g | 6 | 295 |  | and most likely the concept of a data lifecycle itself. | Insert after processes on line 294 |
| 16 | g | 6 | 297 | “ | Replace life cycle with processing steps |  |
| 17 | g | 7 | 302-316 |  | Replace rich for raw | Raw is a hold-over from extractive industries metaphor and does not reflect the state in which it was produced. |
| 18 | g | 7 | 303-310 |  | In a volume use case, the data is often stored in the rich ~~raw~~ state in which it was produced—before being organized in some way such as (sometimes the organization is referred to as extract-load-transform). There are many consequences of the persistence of data stored in the rich state in which it was produced ~~in its~~ ~~raw~~ ~~rich state.~~ A couple of them are that one a schema or model for the data analysis is only applied when the data is retrieved for preparation and analysis and two the data provider can always be determined. This is an example or evidence that it is the ~~raw~~ data that remains in the state in which it was produced that is the richest for Big Data because many different schemas can be applied to the data without altering the richness the raw data contains. This Big Data concept is described as schema-on-read. | Replace existing text |
| 19 | g | 7 | 311-316 |  | **Schema-on-read** is the application of a data schema or model to transform the dataset to better align with the model through preparation steps, changes to dataset, such as transformations, cleansing or removing what one analyst considers unnecessary for his or her purposes, and integration at the time the data is read from the database. It is important to keep in mind that one analysts (context) or dirt is another analyst’s person’s data. | Replace existing text |
| 20 | g | 7 | 323-324 |  | and further evidence of how Big Data diverges from a traditional or unexamined data life cycle model or paradigm. | Add after architectures |
| 21 | g | 7 | 329 |  | The growing need for Interoperability in Big Data can be seen as evidence that in Big Data interoperability means inclusion and creates special problems for the Framework metaphor. It is not clear how the limits of inclusion are determined if at all in Big Data. | insert |
| 22 | g | 7 | 336 |  | and public welfare | Insert |
| 23 | g | 7 | 343 |  | By working out methods for communication among resources including uncertainty metrics, the same scaling is now available to data-intensive applications. | insert |
| 24 |  | 8 | 351 | Data Science Definitions | In many data science projects, the data stored in the form in which it was created, | Replace existing |
| 25 |  | 8 | 352 |  | ~~is browsed first~~ is thought about first (analyzed first) | replace |
| 26 | g | 8 | 356 |  | Public welfare | insert |
| 27 | g | 8 | 358 |  | The**data science paradigm** is the appearance or recognition of useful knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing. | Replace existing  The claim by the data scientist that he/she/they have extracted something valuable (actionable) is relying too much on a real world extractive industry metaphor and allows for the claim that he/she/they have discovered something the value of which belong to he/she/they. It is also arguable whether something was discovered or simply emerged by looking at it which is more an attribute of human perception and/or many other possible scenarios. Because of this doubt no credit or ownership can be taken exclusive of the source of the data. At best the value must be shared with the source, sources of the data. |
| 28 |  |  | 360 |  | Data science can be understood as the appreciative activities happening in the processing layer of the system architecture, within data at rest in the data layer, in order to realize an instance of the knowledge richness inherent in data stored in the manner in which it was received. | Replace existing |
| 29 |  |  | 363 |  | The life of **data** is its usefulness and the benefits that emerge. | Replace existing |
| 30 |  |  | 366 |  | **Analytics** is the appearance, synthesis and communication of knowledge from data. | Replace existing |
| 31 |  |  | 371 |  | life of data | Insert after “steps in the” |
| 32 |  |  | 373 |  | **Data science** is the appearance of actionable knowledge through the analytics of the life of data. | Replace |
| 33 |  |  | 378 |  | as well as linguistics | insert |
| 34 |  |  | 379 |  | ~~by employing deep expertise~~ | remove |
| 35 |  |  | 380 |  | Public welfare | insert |
| 36 |  |  | 383 |  | A **data scientist** is a practitioner who has sufficient knowledge and appreciation in the overlapping regimes of business needs, domain knowledge, analytical skills, software and systems engineering and public welfare to manage the life of data. | Replace |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |