Big Data: Data Wrangling Boot Camp
Big Data Vs

Chuck Cartledge, PhD

16 September 2016

Table of contents I

1 Introduction

2 Big Data's Vs

3 A laundry list of Vs

4 Q & A

5 Conclusion

6 References

Focusing on BD Vs

"What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions."

Lohr [15]

Image from [6].

Doug Laney, META Group

The origin of "Big Data" ideas and definitions.

- Started in the e-commerce Mergers and Acquisitions arena
- Used to explain why traditional Relational Database Management Systems (RDMS) wouldn't scale
- Intended audience was non-technical management

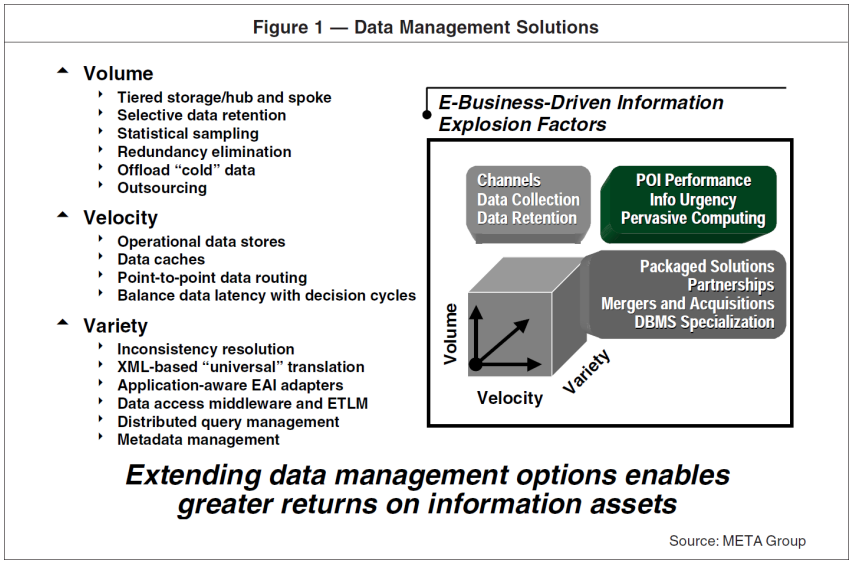


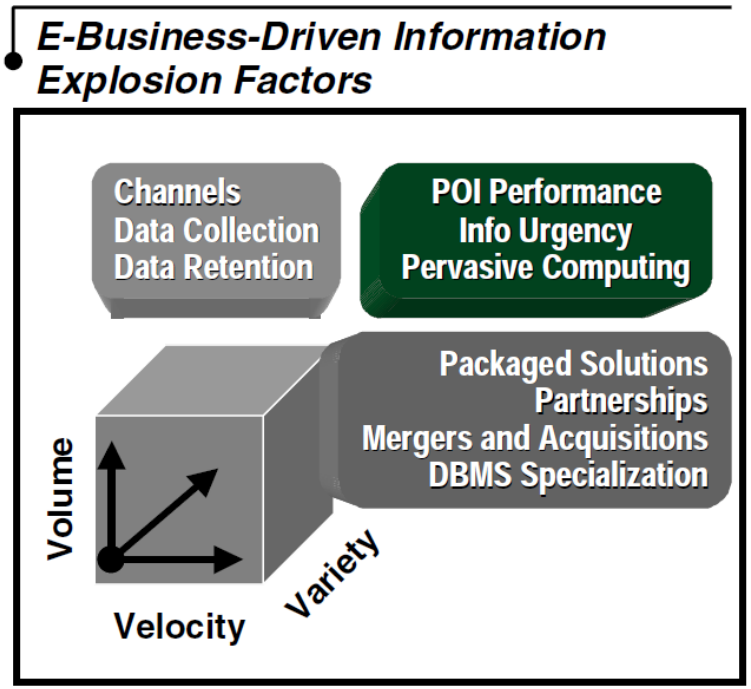
Image from [12].

Take away: traditional RDMS don't/won't scale and different approaches are needed.

Laney's original BD Vs

Figure 1 — Data Management Solutions

- **Volume**
  - Tiered storage/hub and spoke
  - Selective data retention
  - Statistical sampling
  - Redundancy elimination
  - Offload "cold" data
  - Outsourcing
- **Velocity**
  - Operational data stores
  - Data caches
  - Point-to-point data routing
  - Balance data latency with decision cycles
- **Variety**
  - Inconsistency resolution
  - XML-based "universal" translation
  - Application-aware EAI adapters
  - Data access middleware and ETLM
  - Distributed query management
  - Metadata management



Extending data management options enables greater returns on information assets

Source: META Group

Laney's Vs recapped

1. Velocity
   - Frequency of data generation/delivery
   - Think of data from a device, or sensor, robots, clicklogs
   - Real-time analysis is small (9%) [19].
   - Most Big Data analytics is batch
2. Variety
   - Data from a multitude of different sources.
   - Not all data is useful.
   - Data is lost during "normalization"
   - Hopefully not important data, when in doubt: keep it somehow
   - Gets away from relational databases
3. Volume
   - Store relational records?
   - Store transactional records?
   - How long to keep data available?
   - How to access data?
   - How to migrate data?

Volume — what does it mean for Big Data?

How much is there? And, how do we store it?

- Store relational records?
- Store transactional records?
- How long to keep data available?
- How to access data?
- How to migrate data?

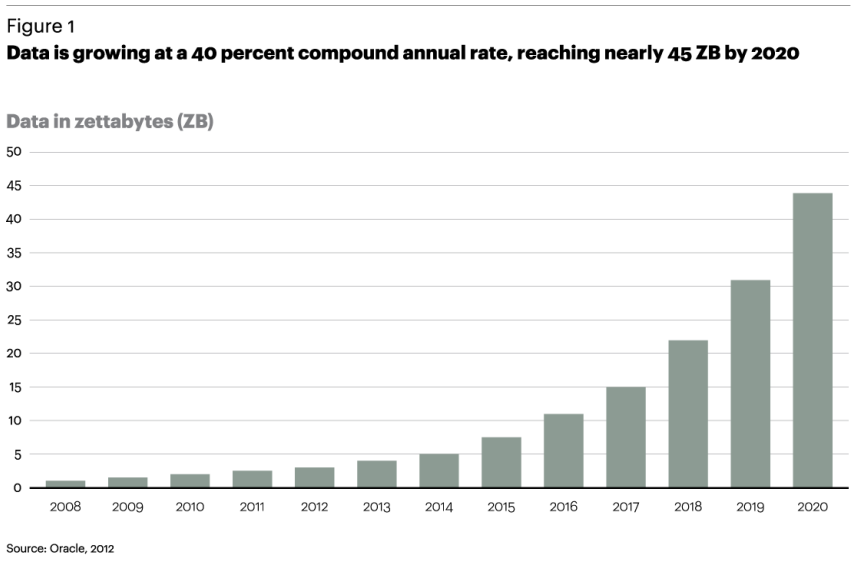
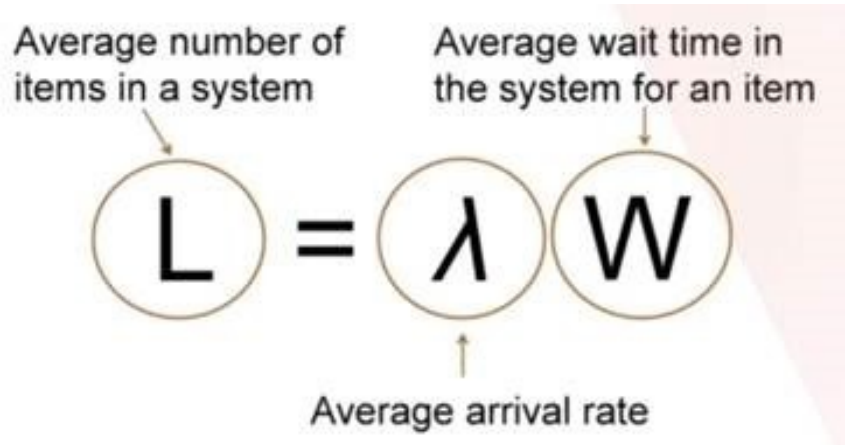


Image from [10].

See http://en.wikipedia.org/wiki/Metric_prefix for list of prefixes.

Velocity — what does it mean for Big Data?

- Frequency of data generation/delivery
- Think of data from a device, or sensor, robots, clicklogs
- Real-time analysis is small (9%) [19].
- Most Big Data analytics is batch



Known as "Little's Law" [13]

Take away: data is generated at a high speed, it must be analyzed before the next set of data is delivered.

Variety — what does it mean for Big Data?

Not all data is the same.

- Data from a multitude of different sources.
- Not all data is useful.
- Data is lost during "normalization"
- Hopefully not important data, when in doubt: keep it somehow
- Gets away from relational databases

The original Vs have been expanded

Lots more Vs.

1. Vagueness
2. Validity
3. Value
4. Variability
5. Variety
6. Velocity
7. Venue
8. Veracity
9. Viability
10. Vincularity
11. Virility
12. Viscosity
13. Visibility
14. Visible
15. Visualization
16. Vitality
17. Vocabulary
18. Volatility
19. Volume

We'll delve into these now.

Big Data as 3 Vs

Image from [20].

Big Data as 4 Vs

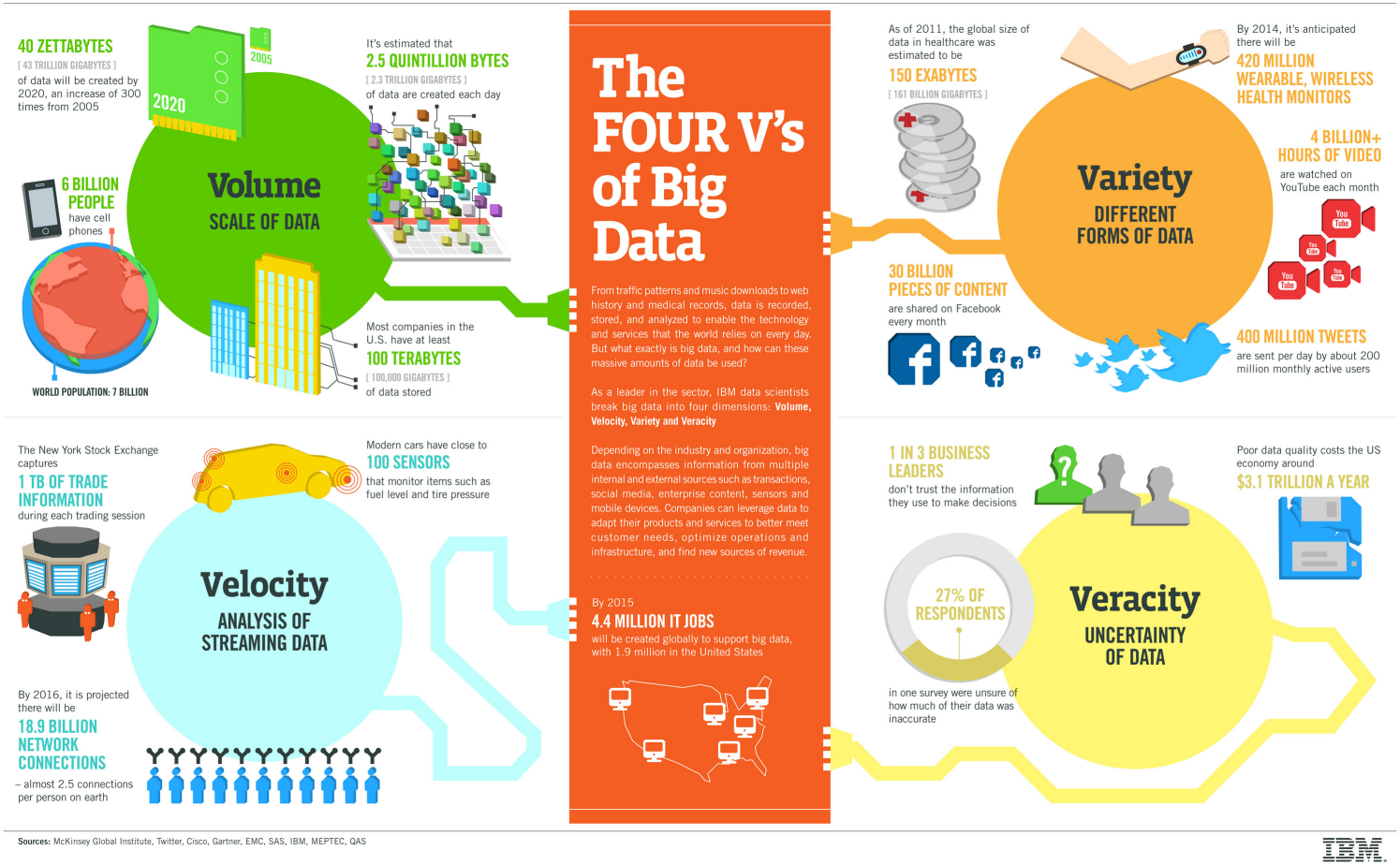


Image from [23].

Big Data as 5 Vs

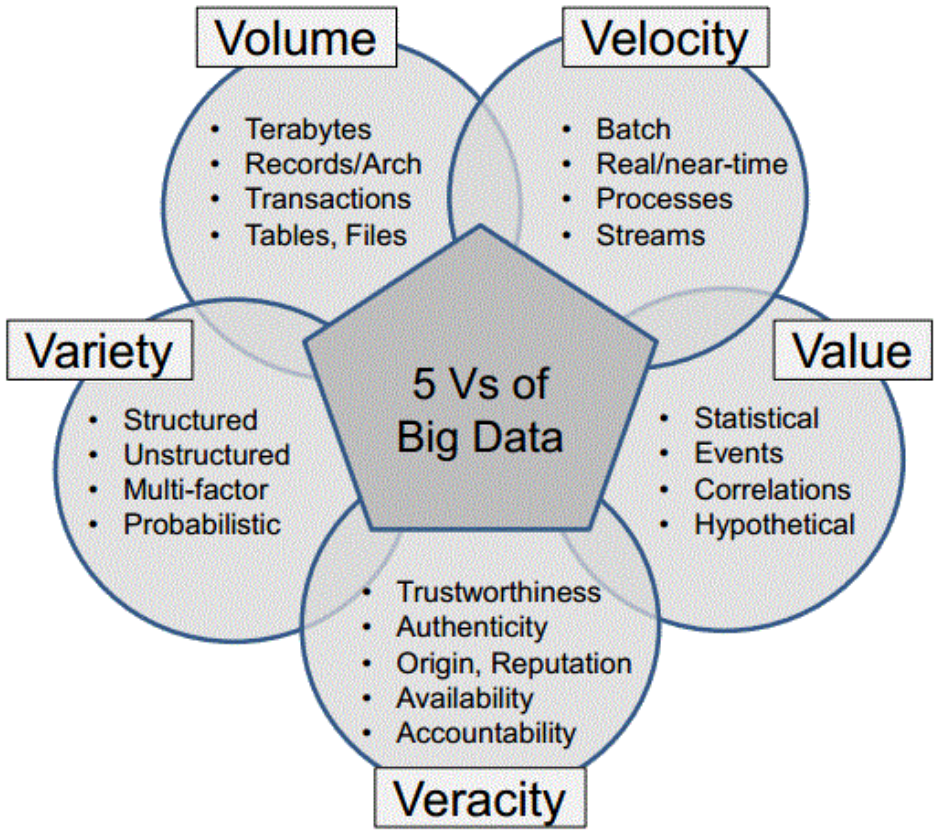


Image from [3].

Big Data as 6 Vs

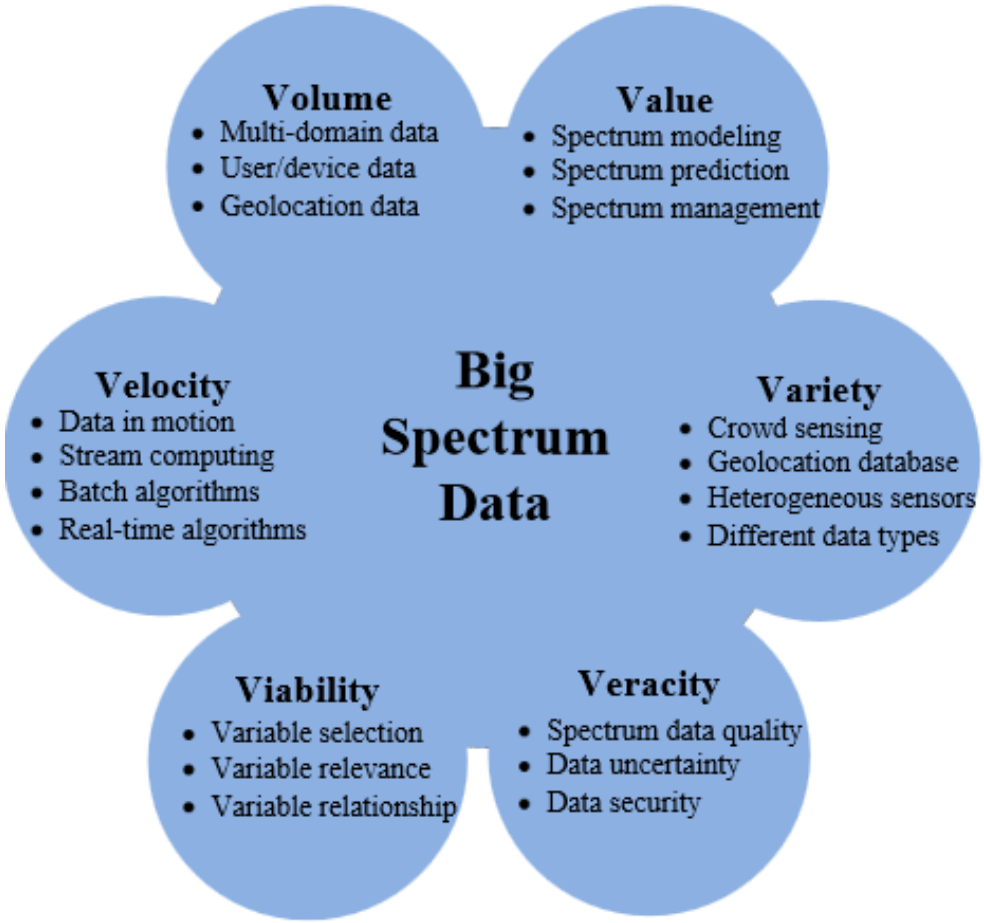


Image from [24].

Big Data as 7 Vs

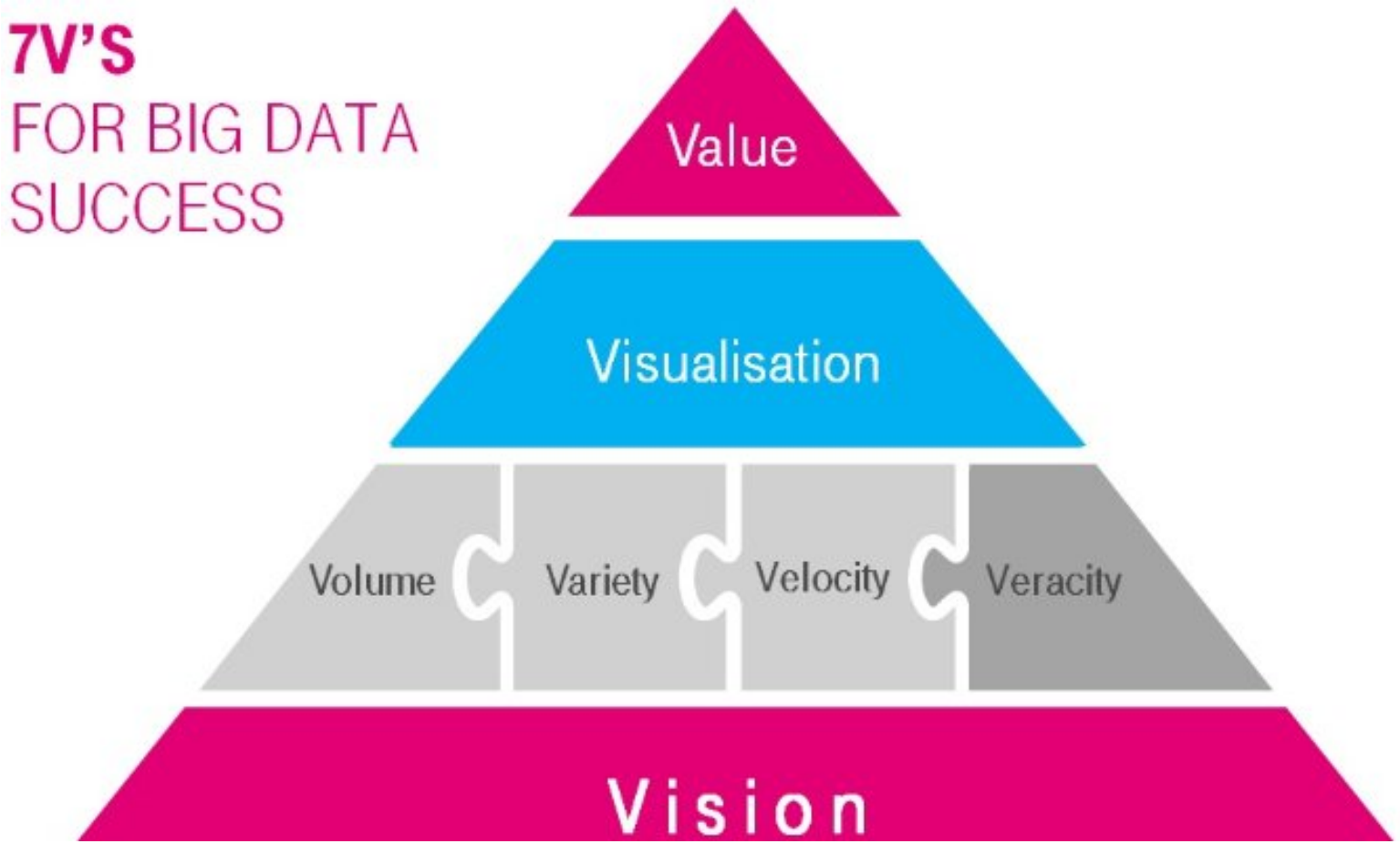


Image from [21].

Big Data as 8 Vs

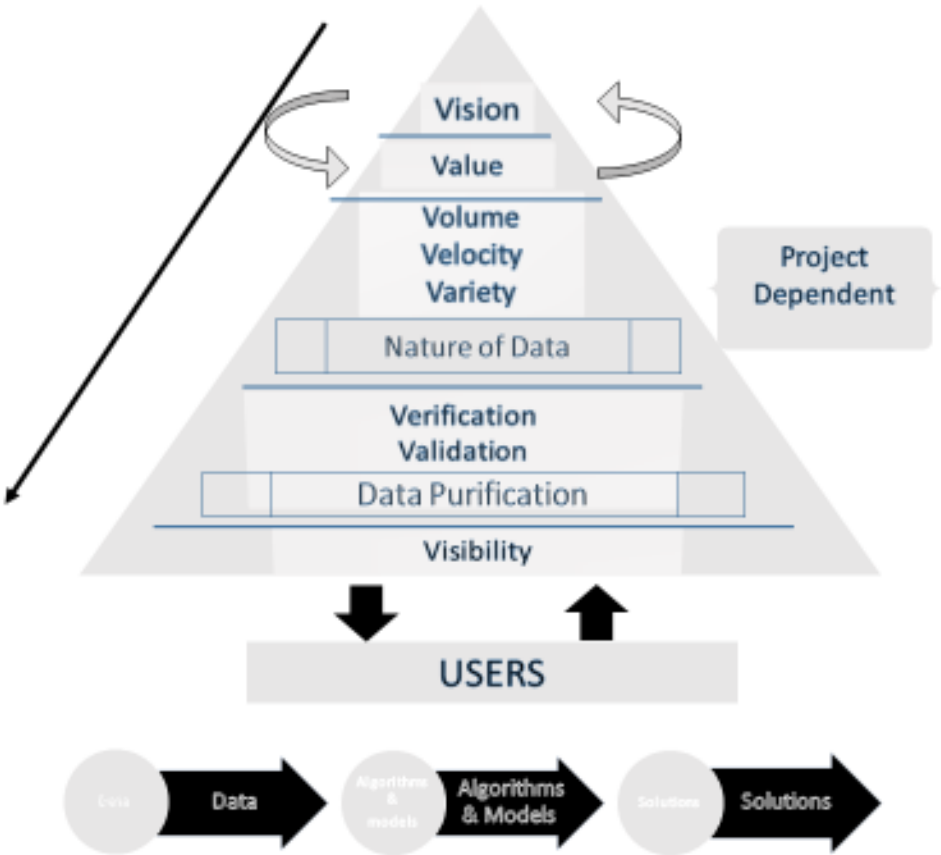


Image from [8].

Vs (part 1 of 7)

Num.	Year	V	Definition	Source
1	2001	Variety	. . . no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics.	[12, 16]
2	2001	Velocity	E-commerce has also increased point-of-interaction (POI) speed and, consequently, the pace data used to support interactions and generated by interactions.	[12]
3	2001	Volume	E-commerce channels increase the depth/breadth of data available about a transaction (or any point of interaction).	[12]

Vs (part 2 of 7)

Num.	Year	V	Definition	Source
4	2013	Validity	. . . is the data correct and accurate for the intended use.	[2, 14, 16, 17, 25]
5	2013	Value	How to determine the prescriptive value of data?	[2, 7, 14, 22, 25, 26, 11, 9, 4, 1]
6	2013	Variability	Many options or variable interpretations can confuse interpretation.	[2, 7, 16, 22, 26]

Vs (part 3 of 7)

Num.	Year	V	Definition	Source
7	2013	Veracity	. . . to the biases, noise and abnormality in data.	[2, 7, 14, 17, 25, 26, 18, 9, 4, 5, 1]
8	2013	Viability	. . . can the data be analyzed in a way that makes it decision-relevant?	[7, 16]
9	2013	Virility	. . . Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.	[26]

Vs (part 4 of 7)

Num.	Year	V	Definition	Source
10	2013	Viscosity	. . . used to describe the latency or lag time in the data relative to the event being described.	[26]
11	2013	Visibility	. . . the state of being able to see or be seen - is implied. [14, 25, 16]	
12	2013	Visualization	Making all that vast amount of data comprehensible in a manner that is easy to understand and read. With the right analyses and visualizations, raw data can be put to use otherwise raw data remains essentially useless.	[22]

Vs (part 5 of 7)

Num.	Year	V	Definition	Source
13	2013	Volatility	. . . how long is data valid and how long should it be stored.	[16, 17]
14	2014	Vagueness	. . . confusion over the meaning of big data (Is it Hadoop? Is it something that weve always had? Whats new about it? What are the tools? Which tools should I use? etc.)	[2]
15	2014	Venue	. . . distributed, heterogeneous data from multiple platforms, from different owners systems, with different access and formatting requirements, private vs. public cloud.	[2]

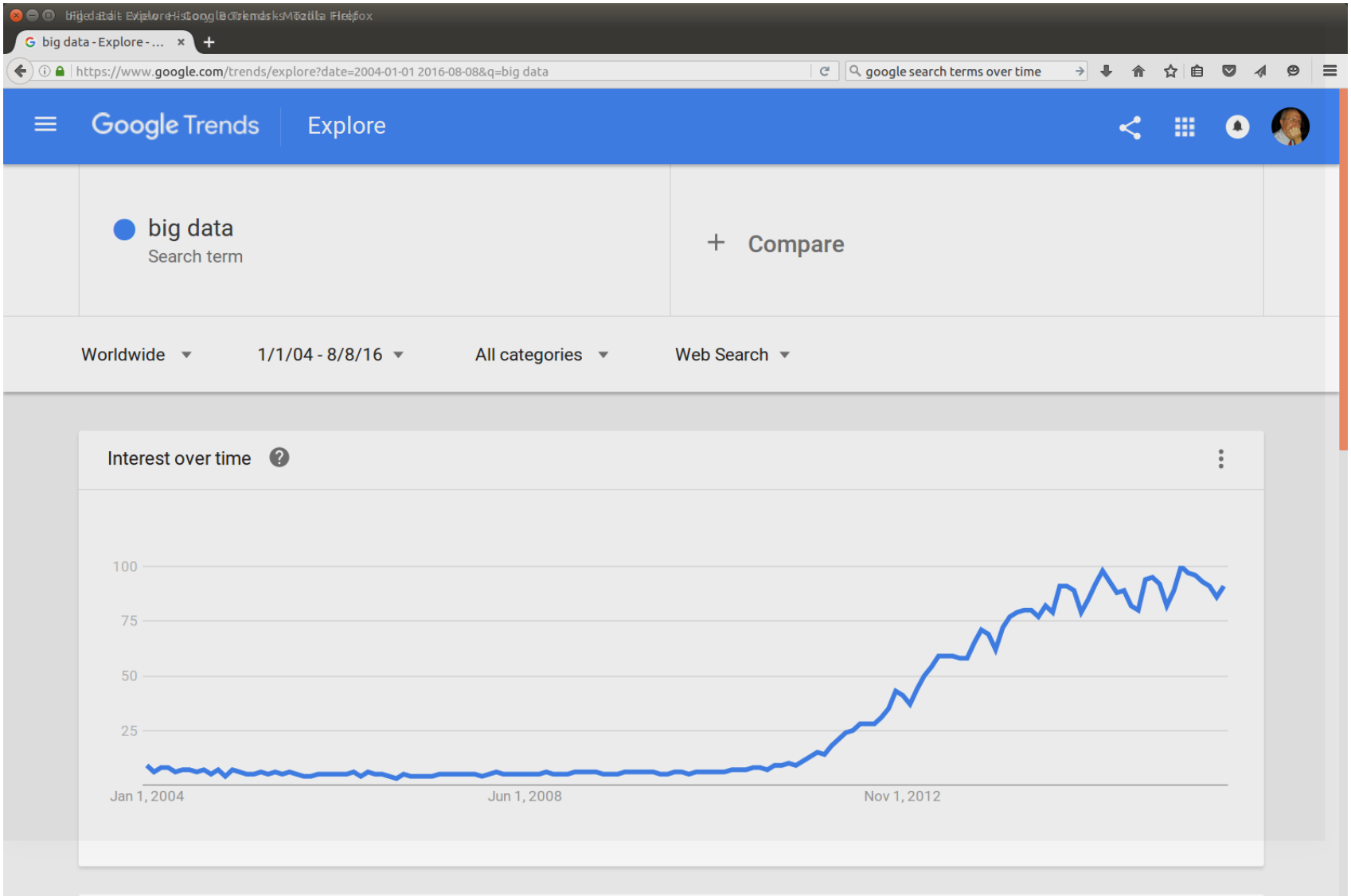
Vs (part 6 of 7)

Num.	Year	V	Definition	Source
16	2014	Vocabulary	. . . schema, data models, semantics, ontologies, taxonomies, and other content- and context-based metadata that describe the datas structure, syntax, content, and provenance.	[2]
17	2015	Vincularity	. . . it implies connectivity or linkage.	[16]
18	2015	Visible	We live in an increasingly visual world and the statistics of increase in the number of images and videos shared on the Internet is staggering.	[16]

Vs (part 7 of 7)

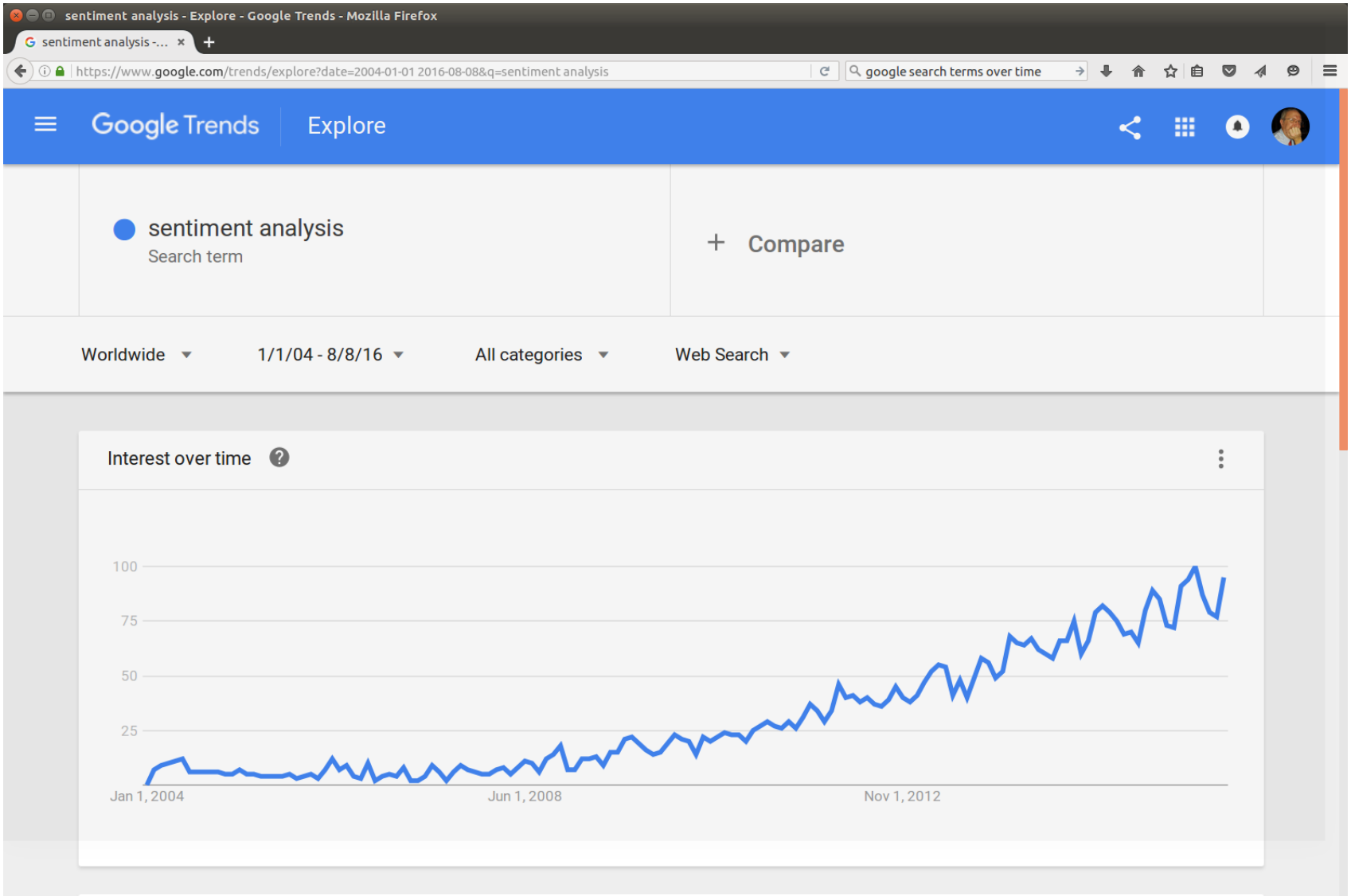
Num.	Year	V	Definition	Source
19	2015	Vitality	. . . criticality of the data is another concept that is crucial and is embedded in the concept of Value.	[16]

Big Data over time



Data source: Google Trends (www.google.com/trends).

Big Data over time



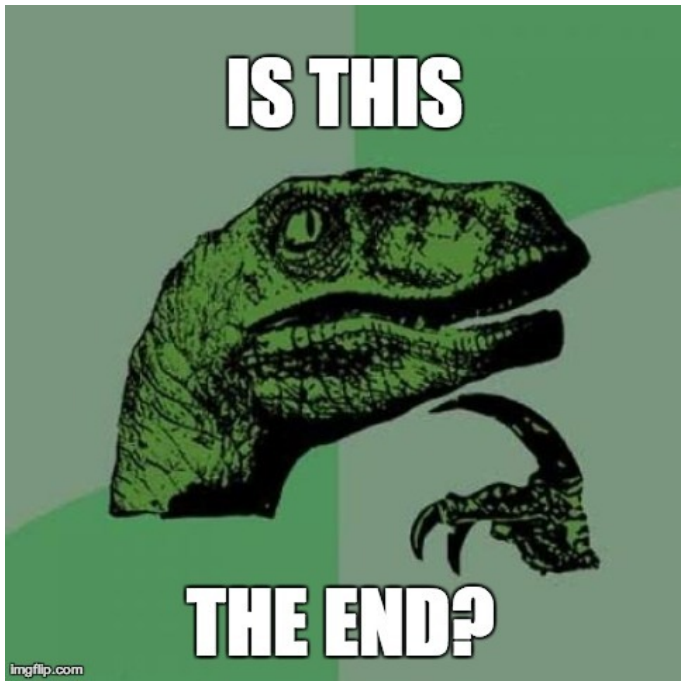
Data source: Google Trends (www.google.com/trends).

Q & A time.

"'The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,' said Deep Thought, with infinite majesty and calm."
Douglas Adams, The Hitchhiker's Guide to the Galaxy

What have we covered?

- Big Data Vs had a specific point of origin
- The list of Big Data continues to grow
- Big Data can be a very nebulous term



Next: Publicly available sources of Big Data.

References I

[1] Marcos D Assunção, Rodrigo N Calheiros, Silvia Bianchi, Marco AS Netto, and Rajkumar Buyya, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing **79** (2015), 3–15.

[2] Kirk Borne, Top 10 big data challenges - a serious look at 10 big data vs, `https://www.mapr.com/blog/top-10-big-data-challenges-%E2%80%93-serious-look-10-big-data-v%E2%80%99s`, 2014.

References II

[3] Patrick Cheesman, How big data can transform your understanding of your customers, http://www.patrickcheesman.com/how-big-data-can-transform-your-understanding-of-your-customers/, 2106.

[4] Yuri Demchenko, Paola Grosso, Cees De Laat, and Peter Membrey, Addressing big data issues in scientific data infrastructure, Collaboration Technologies and Systems (CTS), 2013 International Conference on, IEEE, 2013, pp. 48–55.

[5] Xin Luna Dong and Divesh Srivastava, Big data integration, Data Engineering (ICDE), 2013 IEEE 29th International Conference on, IEEE, 2013, pp. 1245–1248.

References III

[6] David Gewirtz, Volume, velocity, and variety: Understanding the three v's of big data, http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/, 2016.

[7] Seth Grimes, Big data: Avoid 'wanna v' confusion, http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?, 2013.

[8] Uma G Gupta and Mr Ashok Gupta, Vision: A missing key dimension in the 5v big data framework, International Business Research and Marketing **1** (2016).

References IV

[9] Pascal Hitzler and Krzysztof Janowicz, Linked data, big data, and the 4th paradigm., Semantic Web **4** (2013), no. 3, 233–235.

[10] Applied Innovations, Track website visitors, `http://www.appliedi.net/blog/track-website-visitors/`, 2010.

[11] Stephen Kaisler, Frank Armour, Juan Antonio Espinosa, and William Money, Big data: Issues and challenges moving forward, System Sciences (HICSS), 2013 46th Hawaii International Conference on, IEEE, 2013, pp. 995–1004.

[12] Doug Laney, 3d data management: Controlling data volume, velocity and variety, META Group Research Note **6** (2001).

References V

[13] John DC Little, A proof for the queuing formula: L= λ w, Operations Research **9** (1961), no. 3, 383–387.

[14] Rob Livingstone, The 7 vs of big data, `http://rob-livingstone.com/2013/06/big-data-or-black-hole/`, 2013.

[15] Steve Lohr, The age of big data, New York Times **11** (2012).

[16] Rajiv Maheshwari, 3 vs or 7 vs - whats the value of big data?, `https://www.linkedin.com/pulse/3-vs-7-whats-value-big-data-rajiv-maheshwari`, 2105.

References VI

[17] Kevin Normandeau, Beyond volume, variety and velocity is the issue of big data veracity, http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/, 2013.

[18] Wullianallur Raghupathi and Viju Raghupathi, Big data analytics in healthcare: promise and potential, Health Information Science and Systems **2** (2014), no. 1, 3.

[19] Philip Russom, Big data analytics, TDWI Best Practices Report, Fourth Quarter (2011).

[20] Diya Soubra, The 3vs that define big data, http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data, 2012.

References VII

[21] Vit Soupal, 7v's for successful big data project, https://www.linkedin.com/pulse/7vs-successful-big-data-project-vit-soupal, 2015.

[22] BI Staff, Why the 3vs are not sufficient to describe big data, https://datafloq.com/read/3vs-sufficient-describe-big-data/166, 2013.

[23] IBM Staff, The four v's of big data, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, 2016.

[24] InfoIvy Staff, How to use big data to predict utilization of a wireless network?, http://www.infoivy.com/2014/05/how-to-use-big-data-to-predict.html, 2014.

References VIII

[25] University of Technology Staff, The 7 vs of big data, http://mbitm.uts.edu.au/feed/7-vs-big-data, 2013.

[26] Bill Vorhies, How many vs in big data the characteristics that define big data, http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data/, 2013.