Big Data: Data Wrangling Boot Camp
Big Data Overview and Concepts

Chuck Cartledge, PhD

16 September 2016

Table of contents I

On the way to a working definition of BD.

"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 [9]

Image from [3].

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

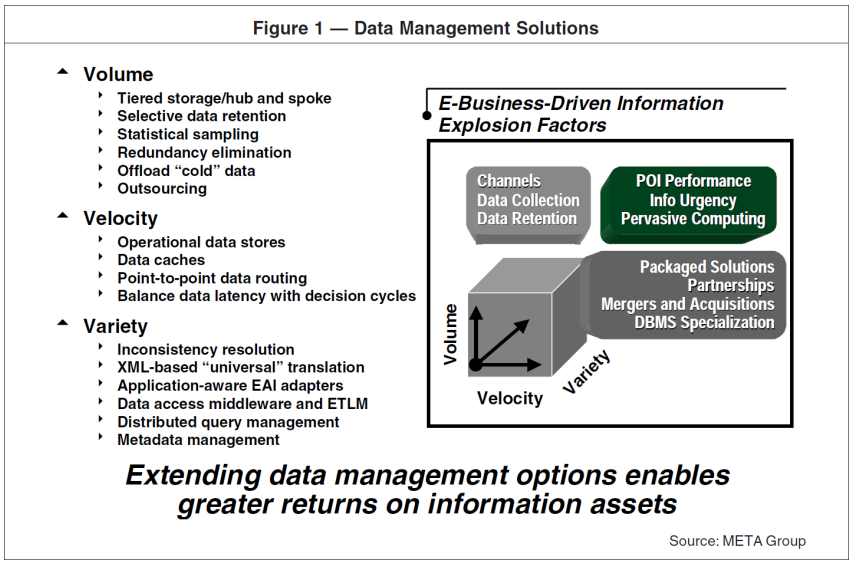


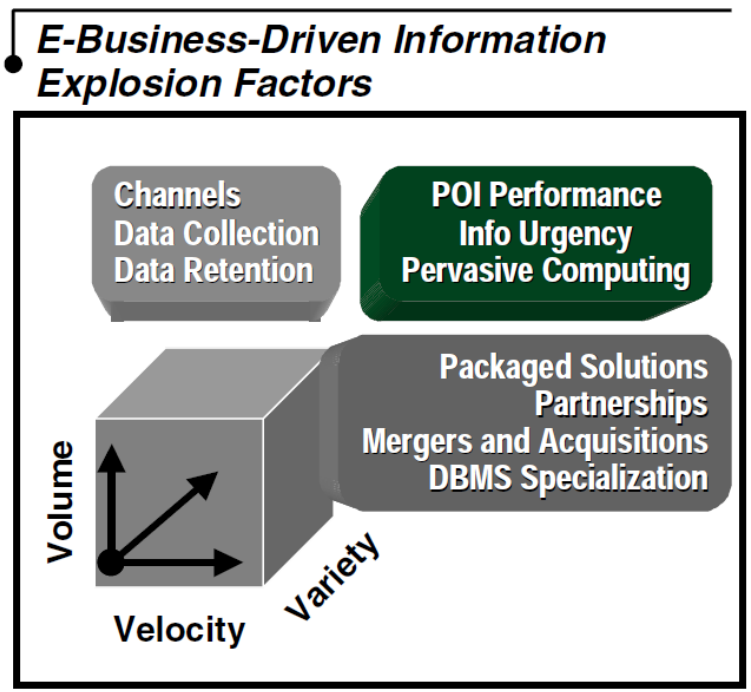
Image from [7].

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



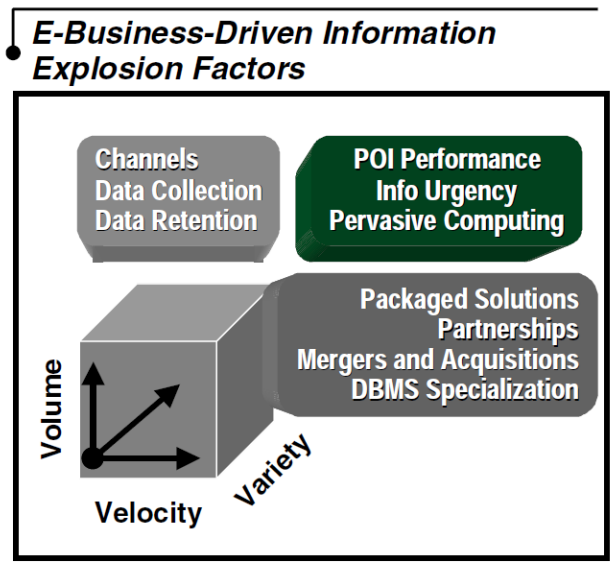
Extending data management options enables greater returns on information assets

Source: META Group

The 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

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?

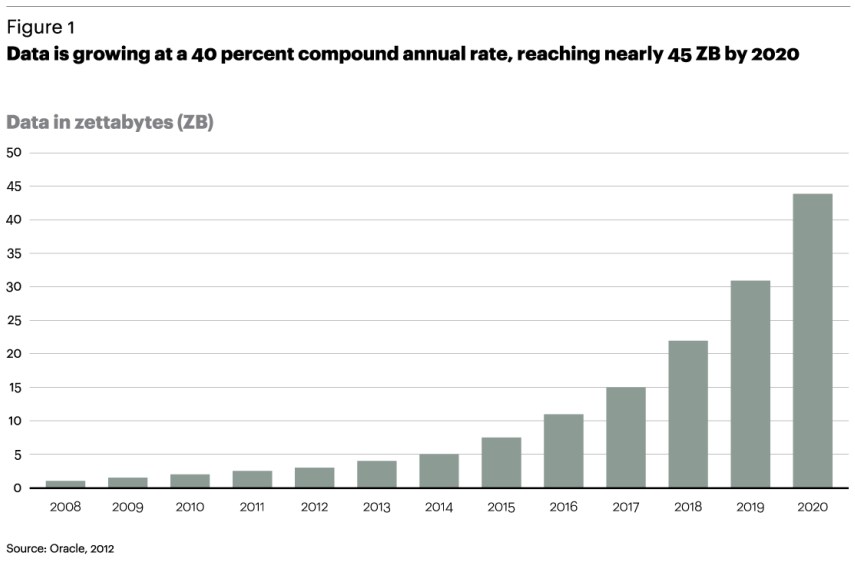
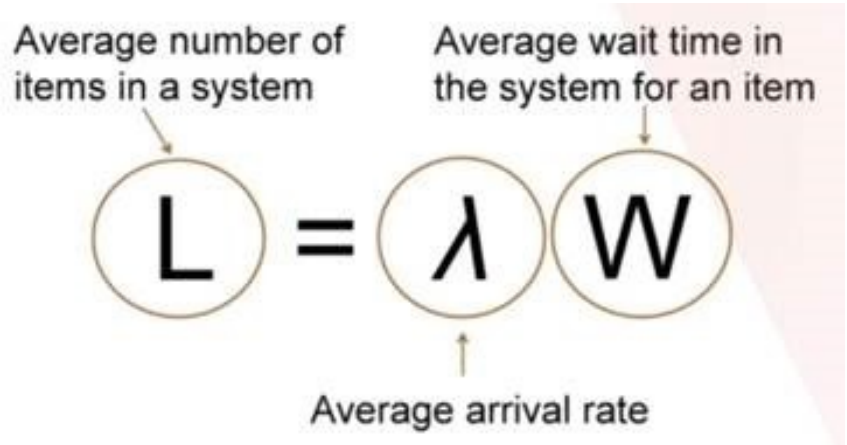


Image from [6].

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%) [10].
- Most Big Data analytics is batch



Known as "Little's Law" [8]

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

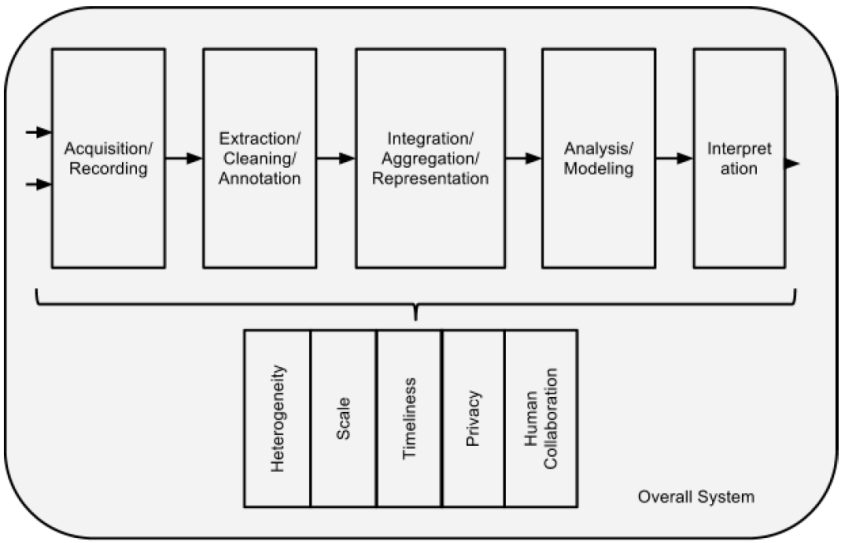
Lots more Vs.

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 talk about these later.

The Big Data challenges.

- Heterogeneity
- Scale
- Timeliness
- Complexity
- Privacy



The Big Data user changes the question[1].

Our friends the Vs

- Classic Vs (Variety, Velocity, Volume)
- Additional Vs

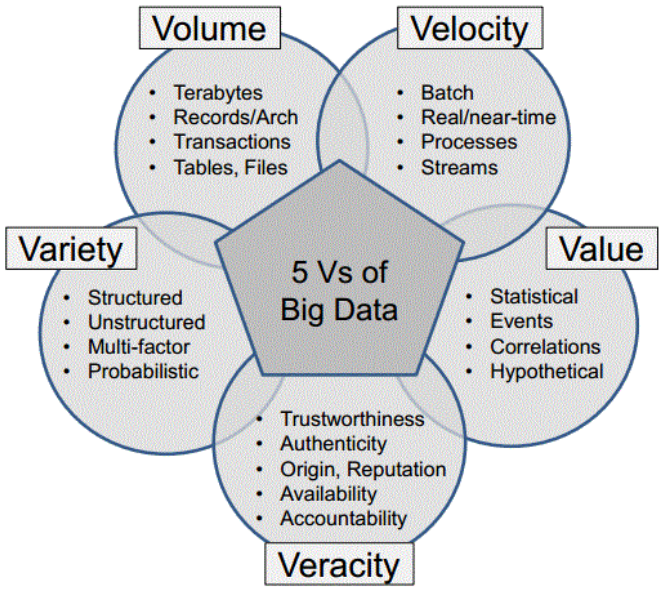


Image from [2].

The Vs tend to overlap.

Data sources

- Government:
  1. Medicare data
  2. NSA, DoD, NASA
- Private:
  1. Clickstream
  2. FICO
  3. Walmart
  4. Android devices
- Free:
  - Far too many to list. (See report.)

Image from [4].

Data characteristics

- Formatted/unformatted (even well-known numbers can be very different)
- Bits, bytes, tagged, free form
- Clean, messy
- Complete, fragmented

10000000
10000000 100000
<spaces> </spaces>
There are spaces.

We'll be looking at unformatted free form text.

Torrents of data

- Primary usage
- Secondary usage
- "Exhaust"
- Storage
  1. Accessibility
  2. Longevity
  3. Privacy

Image from [13].

Data can be intentional, or accidental, or by-products, but there is lots of it.

Big data players

- Visionaries – stand on the shoulders of giants and see new horizons
- Brokers – have seas and lakes of data at their disposal
- Scientists – dive into the seas and make the visions real

We will be performing a small part of the data scientist's labors.

A 50,000 foot view

What are the layers in this cake?

- User — the person (or thing) that want's something done
- Application — the program that does the work
- Operating system — arbitrates between multiple programs and limited resources
- Hardware — the silicone, copper, other tangibles that generate heat

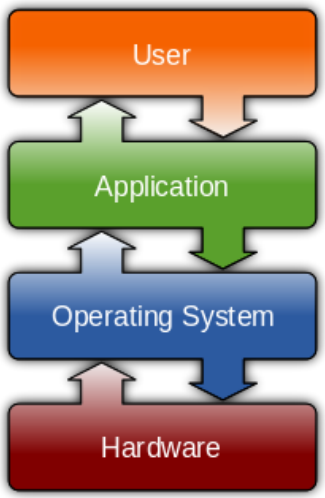


Image from [14].

Layering is a key concept.

Focusing on the OS

What does it do?

- Provides a user interface (maybe a Command Line Interface)
- Schedules access to the hardware
- Schedules the functions of the CPU

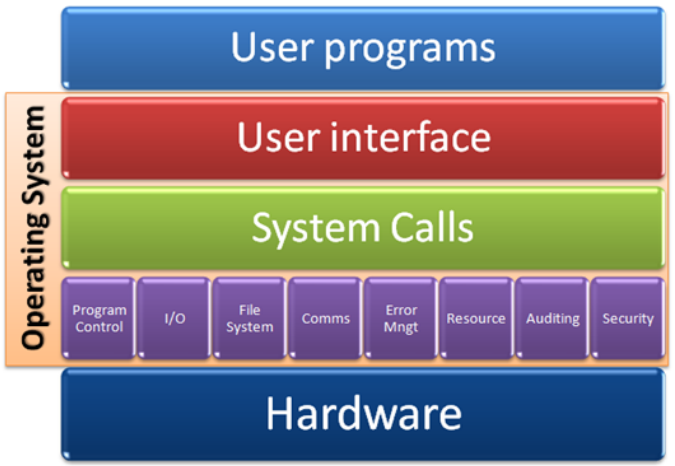
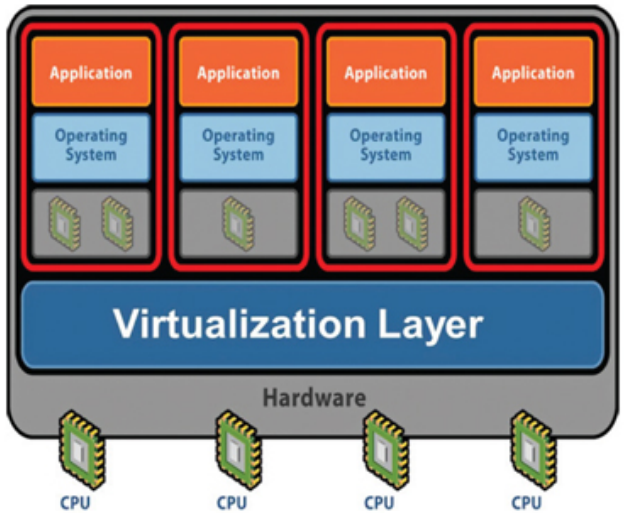


Image from [11].

An OS is a program (albeit, a large program). What if we could write a program that would run an OS as an application?

Tricking the upper layer

- Higher layers rely on lower layers for services
- Layers create interfaces
- Interfaces allow for hiding details



Virtualization software allows applications that previously ran on separate computers to run on one server machine.

Image from [5].

As long as the lower layer supplies all the services, the upper layer won't know where the services originated.

One hardware suite can run many OSs in virtual machines.

- Ultimately the hardware determines how many virtual machines can be run
- Faster CPU(s), more RAM, more network connections, more disks, . . . , more is better
- Fewer actual machines usually means lower power, lower cooling, cheaper upgrade path

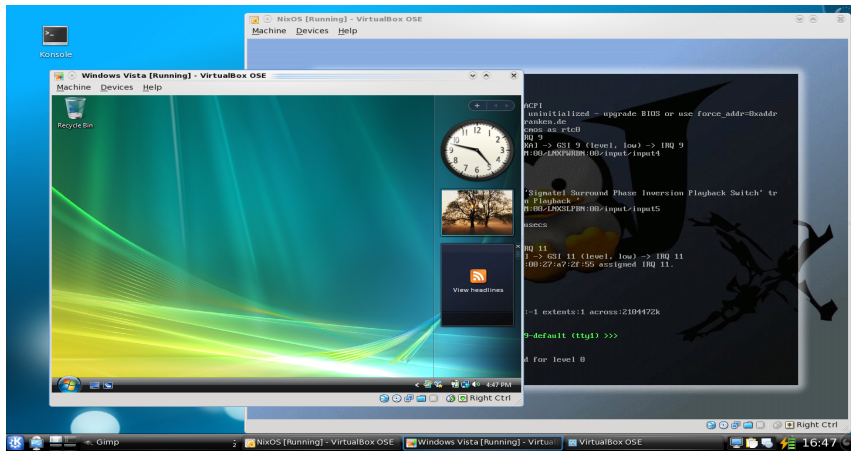


Image from [12].

With clever software, almost anything can be virtualized. Hadoop is clever software.

Anything that has to be fast.

- Underlying hardware suite is shared across all "machines"
- Mission critical applications

In summary.

- To use virtual machines, or
- To not use virtual machines.

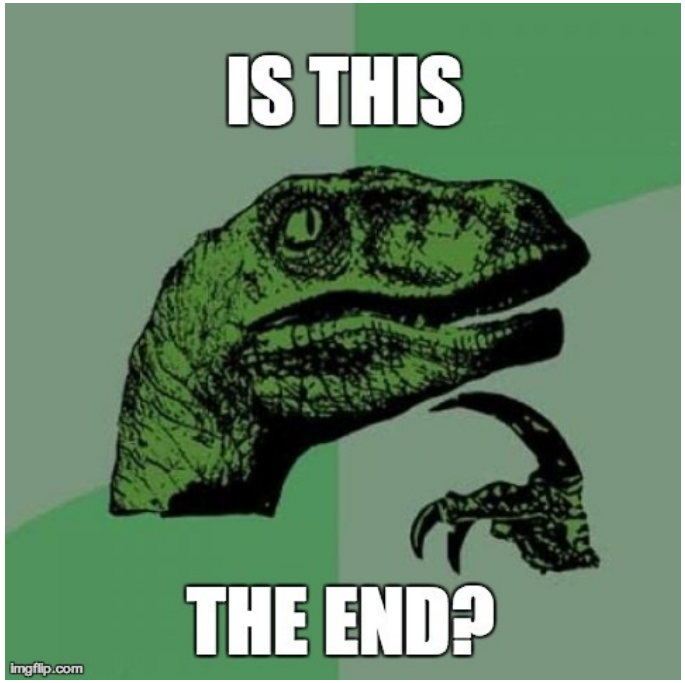
It depends on what is important. Many BD tools and techniques make use of virtualization.

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
- Big data has a list of challenges
- Big data can be very messy, and not neat and tidy
- Hinted at how BD tools and techniques use virtualization



Next: Understanding more about BD Vs.

References I

[1] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, and Michael Franklin, Challenges and opportunities with big data, Purde e-Pubs (2011).

[2] 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.

[3] 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.

References II

[4] Christian Hagen, KHalid Khan, Marco Ciobo, and Jason Miller, Big data and the creative destruction of today's business models, http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcg0eS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192, 2013.

[5] Paul Hodge, Virtualization 101: Understanding how to do more with less, https://www.isa.org/standards-and-publications/isa-publications/intech-magazine/2011/august/system-integration-virtualization-101-understanding-how-to-do-more-with-less/, 2011.

References III

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

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

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

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

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

References IV

[11] Willy-Peter Schaub, Unisa chatter operating system concepts: Part 2 system structures, http://blogs.msdn.com/b/willy-peter_schaub/archive/2010/01/07/unisa-chatter-operating-system-concepts-part-2-system-structures.aspx, 2010.

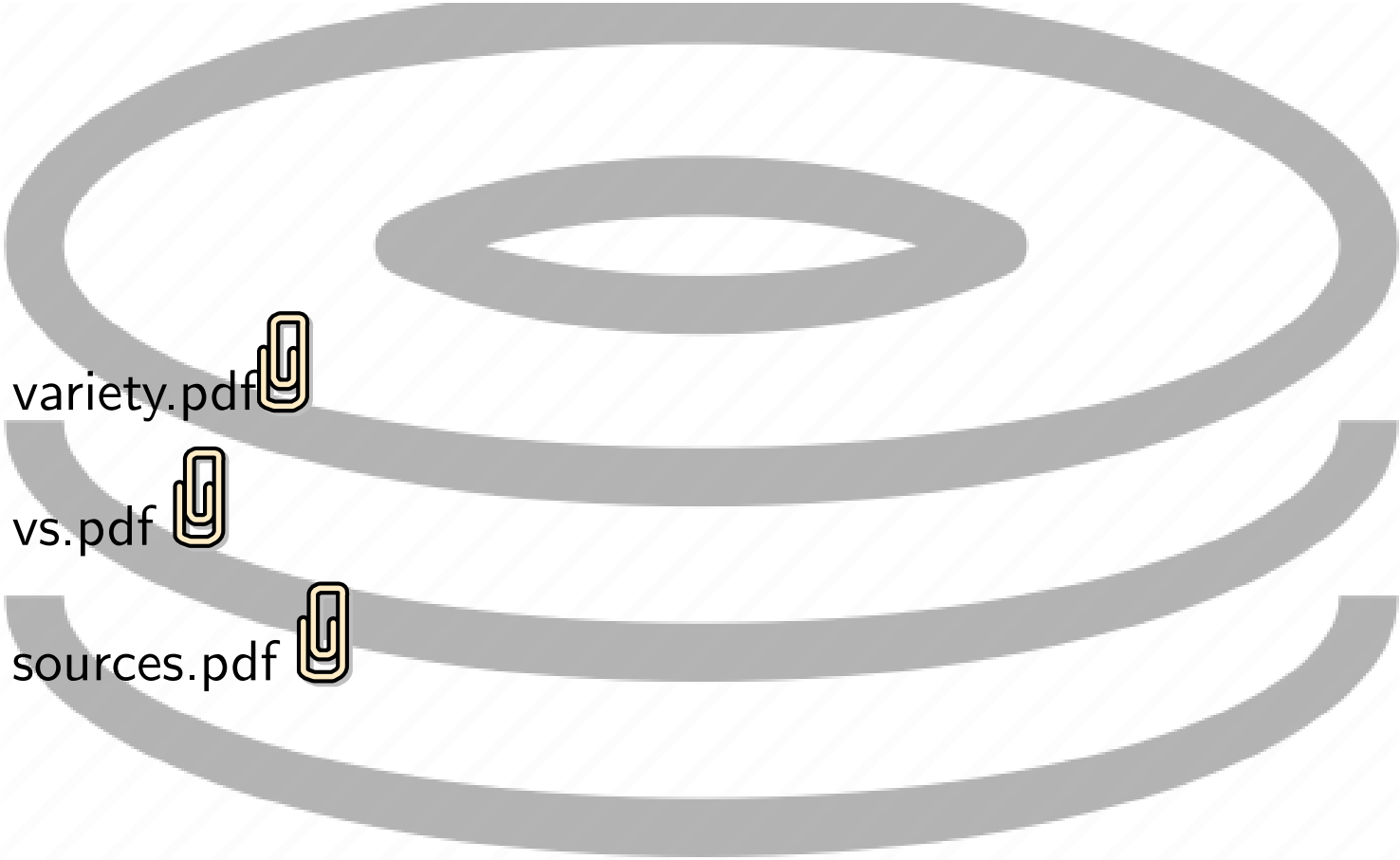
[12] NixOS Staff, Nixos screenshots, https://nixos.org/nixos/screenshots.html, 2016.

References V

[13] NYU Staff, Nyu launches initiative in data science and statistics to push advances in medicine, science, technology, and other fields, https://www.nyu.edu/about/news-publications/news/2013/02/19/nyu-launches-initiative-in-data-science-and-statistics-to-push-advances-in-medicine-science-technology-and-other-fields.html, 2013.

[14] Wikipedia, Software — wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Software, 2015.

Files of interest



1. variety.pdf
2. vs.pdf
3. sources.pdf