

Understanding Big Data





"6.3 Million Google Searches Take Place
Every Minute"



"241 million emails are sent per minute"

Activity	Amount of activity done in 60 seconds.
Emails sent	241 million
WhatsApp messages sent	41.6 million
Global hours spent online	25.1 million
Searches on Google	6.3 million
Facebook posts liked	4 million
Reels sent via DM on Instagram	694,000
X (Twitter) posts sent	360,000
Taylor Swift song streamed	69,400
Hours of content watched on Twitch	48,000
LinkedIn resumes submitted	6,060

%

%

%

%

“90% of the world's data has been generated in the last two years alone.

This relentless pace shows no signs of slowing down.”

NNNNN

"By 2025, an estimated 463 exabytes of data will be created each day globally.

That's like watching the entire Netflix catalog over 4 million times every single day!"



“There will be over 75 billion Internet of Things (IoT) devices connected world wide by 2025, constantly generating a massive stream of data.”

Big Brother meets Big Data, in an office near you

Forbes / Tech

MAY 27, 2015 @ 10:20 AM

34,550

How Big Data

Seems like A LOT of data!

Is this Big Data?

The Little Black Book of Billion

ove Public

E

ers to

Français

New York
Advertisi

Team will help lure marketers with tools to predict which articles will resonate with certain readers
better target advertising

Science Tools for

Data v

Make Better Business

According to Accenture Report



Interpretations of Big Data

- **Truly Massive Volume:** Datasets reaching into petabytes or exabytes, simply too large to store or open in traditional desktop software.
- **Relative to Capabilities:** Big Data for one might be Small Data for others.
- **Relative with Time:** What is large-scale today will likely seem small-scale in the near future!
- **Outgrowing Traditional Tools:** Dataset that is difficult to manage using traditional database systems.
- **Marketing Hype:** Sometimes "Big Data" is used as a buzzword to sell products or services.



Beyond Buzzwords ...

Big Data : Datasets that are too large, complex, and rapidly-changing to be effectively managed and analyzed using traditional data processing tools and methods.



Beyond Buzzwords ...

Big Data : Datasets that are too large, complex, and rapidly-changing to be effectively managed and analyzed using traditional data processing tools and methods.

When the volume starts to prevent us from doing what we need with the data.

Beyond Buzzwords ...

Big Data is about traditional tools reaching their breaking point.

Big Data : Datasets that are too large, complex, and rapidly-changing to be effectively managed and analyzed using traditional data processing tools and methods.

Beyond Buzzwords ...

Variety of sources – customer data, social media feeds, sensor data etc.

How fast new information is constantly being generated!

Data : Datasets that are too large, complex, and rapidly-changing to be effectively managed and analyzed using traditional data processing tools and methods.

**Big Data is about a fundamental
shift in how we need to approach
data**

– not just having more of it!

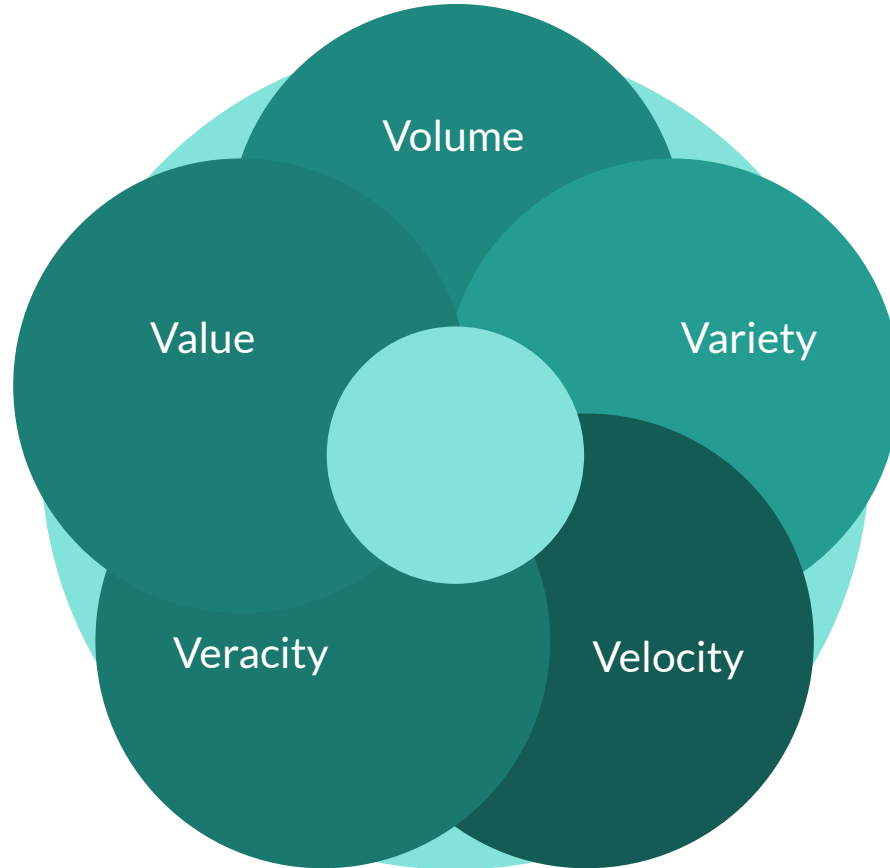
**Big Data is a
shift in how we
data**

Let's understand the defining
characteristics of Big Data.

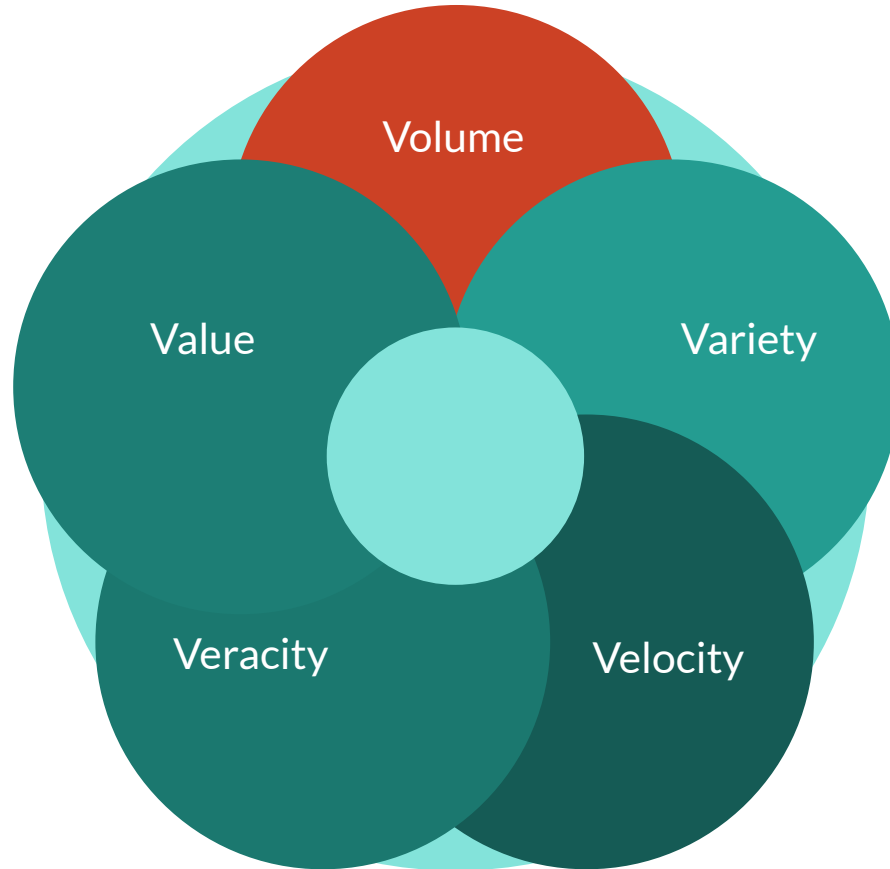
– not just having more of it!

V's of Big Data

Big Data V's



Big Data V's



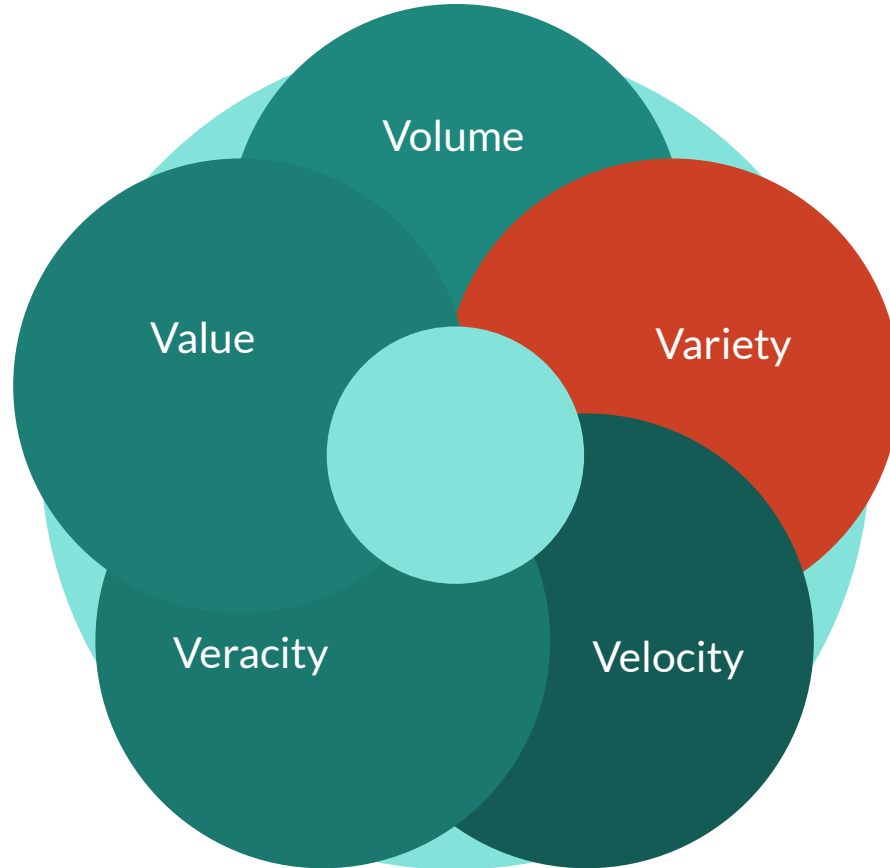


- Massive Datasets Beyond Traditional Capabilities
 - Numbers that are hard to even comprehend.
 - Increasing exponentially everyday.



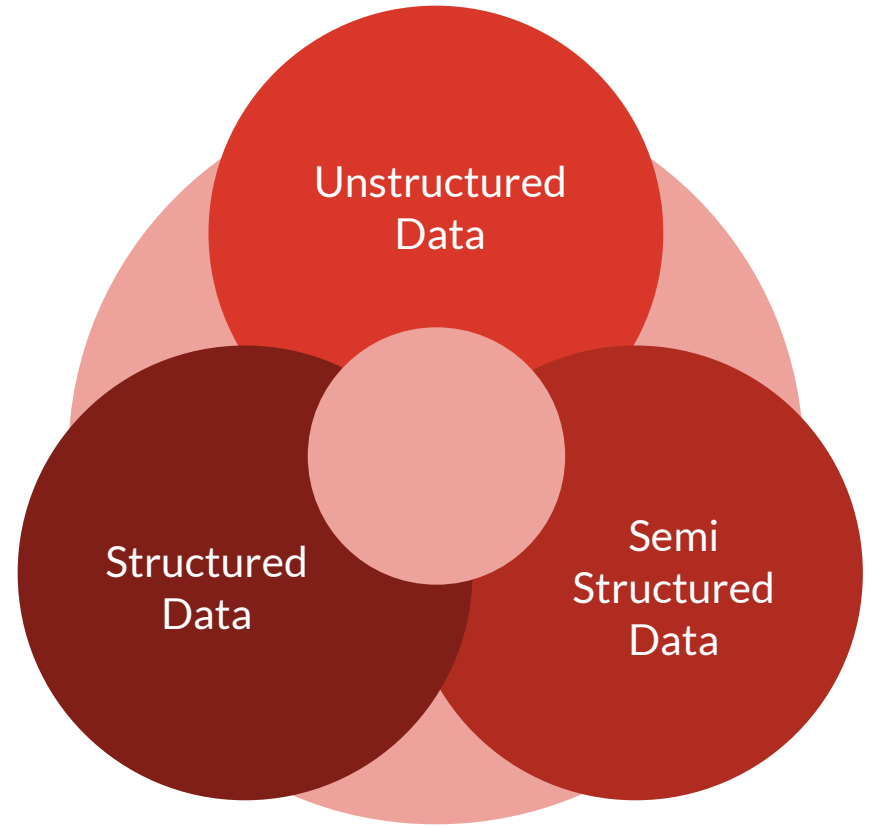
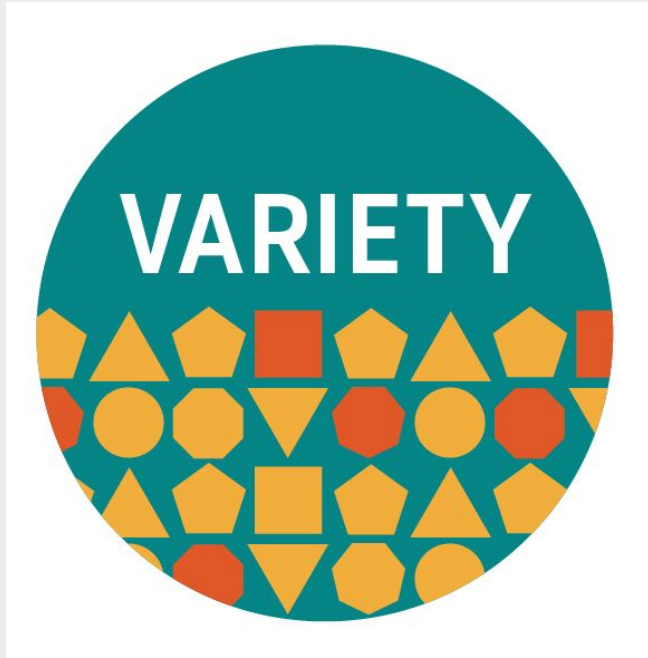
- Massive Datasets Beyond Traditional Capabilities
 - Numbers that are hard to even comprehend.
 - Increasing exponentially everyday.
- The Large Hadron Collider generates 1 petabyte of data per second.
 - 1 petabyte = 10^{15} bytes
 - 1 petabyte = 1,000 terabytes
 - 1 petabyte = 1 million gigabytes

Big Data V's

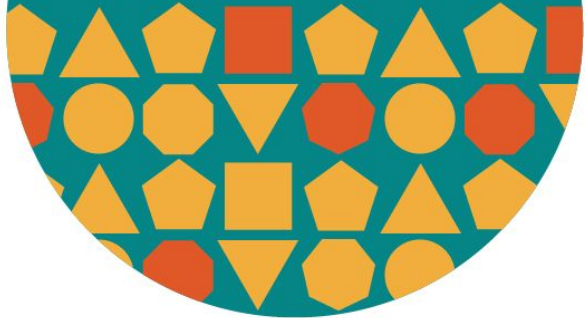




- Big Data isn't just numbers, dates & strings.
- Different forms that data can come in -
 - Text
 - Images
 - Voice ... etc.
- Handling this messiness requires new approaches.



VARIETY

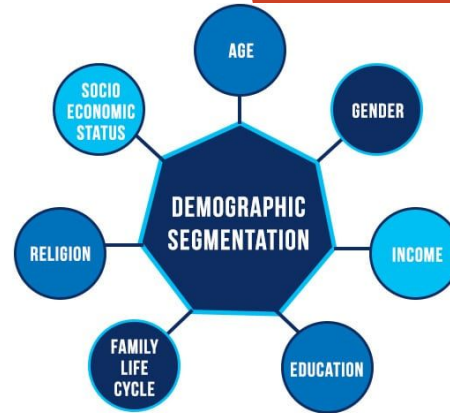


Structured Data

- Highly organized, easily searchable.
- Fits neatly into rows and columns.

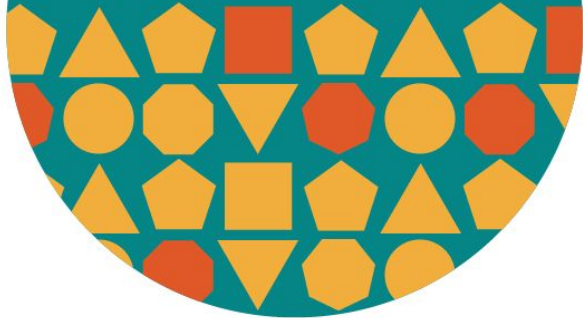
VARIETY

Structured Data



InvoiceNo	StockCode	InvoiceDate	CustomerID	Country
1001	A	1/1/16 13:40	12	X
1001	B	1/1/16 13:40	12	X
1001	C	1/1/16 13:40	12	X
1002	B	1/2/16 18:15	12	X
1002	C	1/2/16 18:15	12	X
1003	B	1/1/16 17:30	13	Y
1004	C	1/1/16 17:30	13	Y
1005	C	1/6/16 12:45	14	Z
1005	B	1/6/16 12:45	14	Z
1006	D	1/20/16 11:00	14	Z

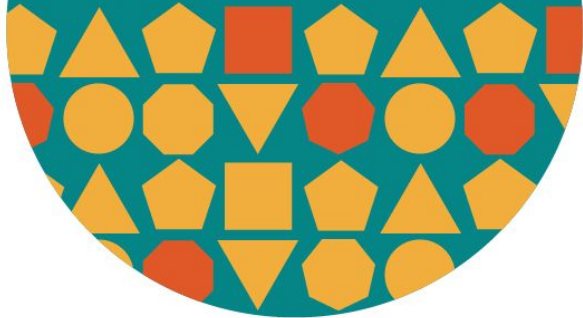
VARIETY



Structured Data

- Highly organized, easily searchable.
- Fits neatly into rows and columns.
- Example -
 - Demographic data: Age, income, zip code in a table
 - Transactional records: Sales figures, customer orders

VARIETY



Semi Structured Data

- Has some internal organization, but not rigid format like a table.

VARIETY

Semi Structured Data

- Has some internal organization, but not rigid format like a table.
- Example -
 - Email: Email body is less structured.

VARIETY



Unstructured Data



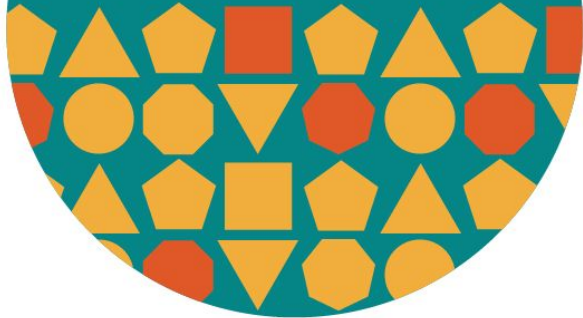
- No inherent structure.
- Requires significant processing to extract meaning.

VARIETY

Unstructured Data

- No inherent structure.
- Requires significant processing to extract meaning.

VARIETY



Unstructured Data

- No inherent structure.
- Requires significant processing to extract meaning.
- Example -
 - Social Media Posts: Raw text, emojis etc.
 - Images & Videos: Pixel data or a sequence of frames.

Summary



Characteristic	Structured	Semi-Structured	Unstructured
Organization	Highly organized, fits neatly into rows and columns	Internal tags or metadata provide some structure	No inherent organization
Examples	Transactional records, spreadsheets	Website logs, XML/JSON data, emails	Social media posts, images, customer reviews
Ease of Analysis	Easily analyzed with traditional tools	Requires specialized tools for parsing and analysis	Requires the most advanced techniques (often machine learning)
Insights Potential	Clear answers to well-defined questions	Can reveal patterns and trends	Holds potential for complex, nuanced insights

Summary

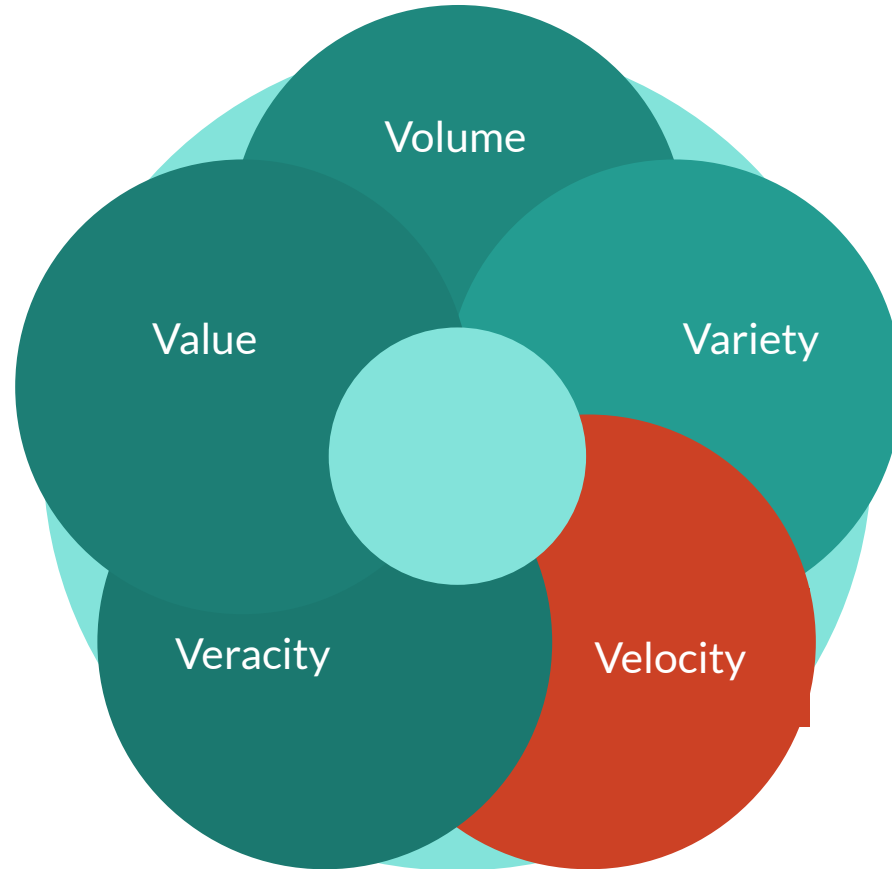
Clean

Messy

Very
Messy

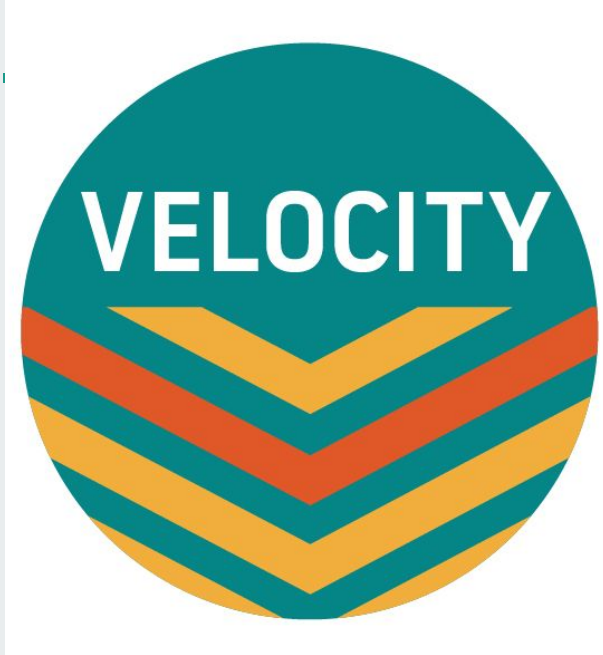
Characteristic	Structured	Semi-Structured	Unstructured
Organization	Highly organized, fits neatly into rows and columns	Internal tags or metadata provide some structure	No inherent organization
Examples	Transactional records, spreadsheets	Website logs, XML/JSON data, emails	Social media posts, images, customer reviews
Ease of Analysis	Easily analyzed with traditional tools	Requires specialized tools for parsing and analysis	Requires the most advanced techniques (often machine learning)
Insights Potential	Clear answers to well-defined questions	Can reveal patterns and trends	Holds potential for complex, nuanced insights

Big Data V's



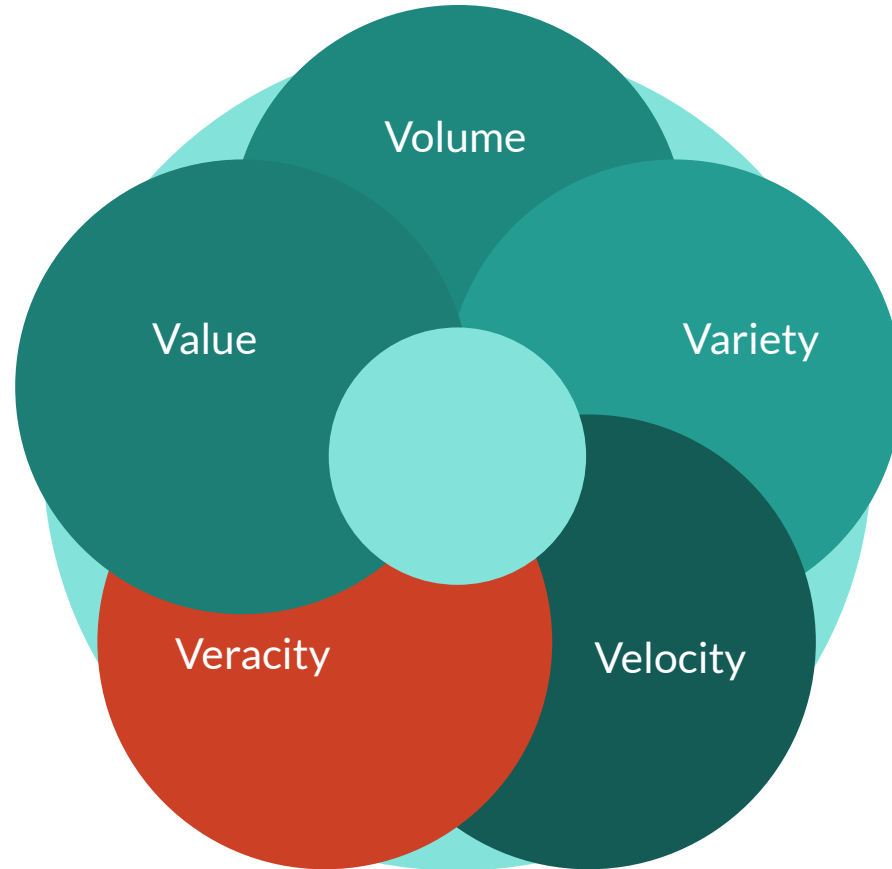


- How quickly data is generated.
- The speed at which it needs to be analyzed.



- How quickly data is generated.
- The speed at which it needs to be analyzed.
- Examples:
 - Stock Market: Prices fluctuate by the second, requiring algorithmic trading, not just end-of-day reports.
 - Social Sentiment: Monitoring social media in real-time for brand reputation or during major events.

Big Data V's



Veracity

Can We Trust the Data?





Veracity

Can We Trust the
Data?

- With so many sources, quality can vary wildly.
- Misleading data can lead to wrong insights.

Veracity
Can You
Data?

Recall



ality can

wrong

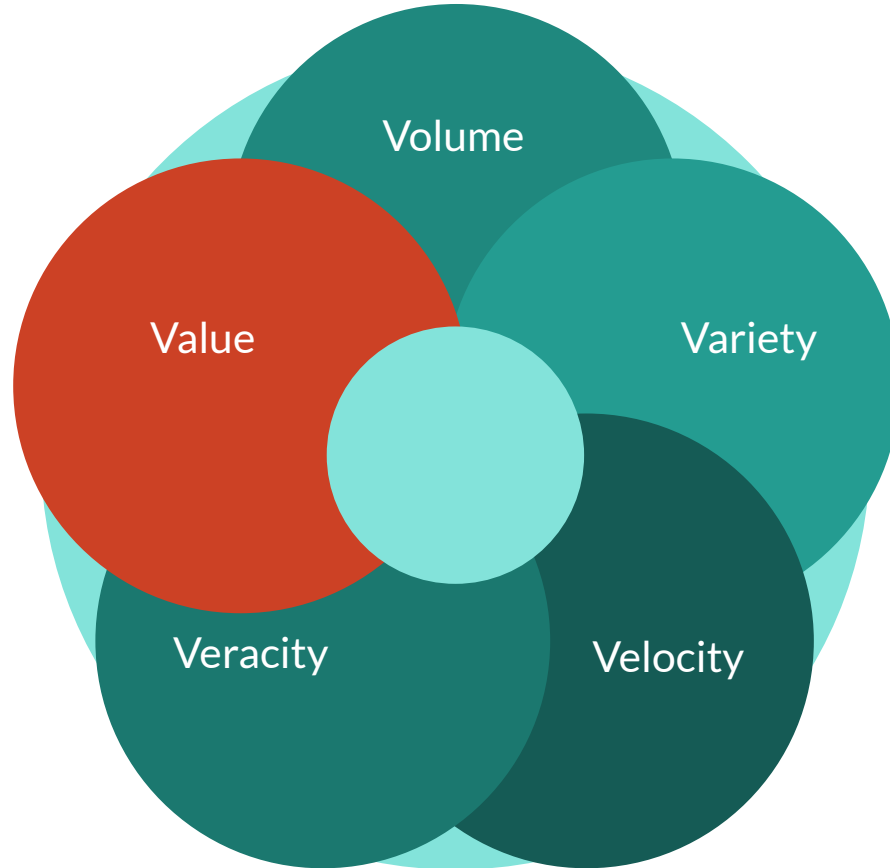


Veracity

Can We Trust the Data?

- With so many sources, quality can vary wildly.
- Misleading data can lead to wrong insights.
- Examples:
 - Inconsistent Sensor Data: Faulty sensors feeding errors into a system.
 - Fake Reviews or Social Bots: Manipulated online information distorting sentiment analysis.

Big Data V's



Value

Unlocking
Actionable
Insights





Value

Unlocking
Actionable
Insights


- Big Data only matters if it helps us do something better –
 - Smarter decisions
 - Breakthroughs
 - Increased efficiency



Value

Unlocking Actionable Insights

- Big Data only matters if it helps us do something better –
 - Smarter decisions
 - Breakthroughs
 - Increased efficiency
- Examples:
 - **Targeted Recommendations:** Personalized product suggestions leading to increased sales.
 - **Predictive healthcare:** Analyzing patient data to identify those at high risk and intervene early.



Value
Unloc
Action
Insigh

Takeaway

Focus on how data translates to better decisions, new products, or improved lives!

What did we learn?

Velocity == Missed Opportunities

Missed windows for action or slow decision making.

Veracity == Inaccurate Results

Insights can be wrong or misleading.

Volume == Storage Bottlenecks

Limit on what can be stored.

Variety == Complexity Overload

Analysis becomes difficult or impossible.

Value == Wasted Potential

Leaving potential gains on the table.

Velocity == Miss

Missed windows

Veracity == In

Insights can be v

Volume == Sto

Limit on what ca

Variety == Com

Analysis become

Value == Wast

Leaving potential

Big Data offers huge potential, BUT it
requires the right skills and tools to
harness.

Velocity == Miss

Missed windows

Veracity == In

Insights can be v

Volume == Sto

Limit on what ca

Variety == Co

Analysis become

Value == Wast

Leaving potential

Big Data offers huge potential, BUT it
requires the right skills and tools to
harness.

HADOOP!