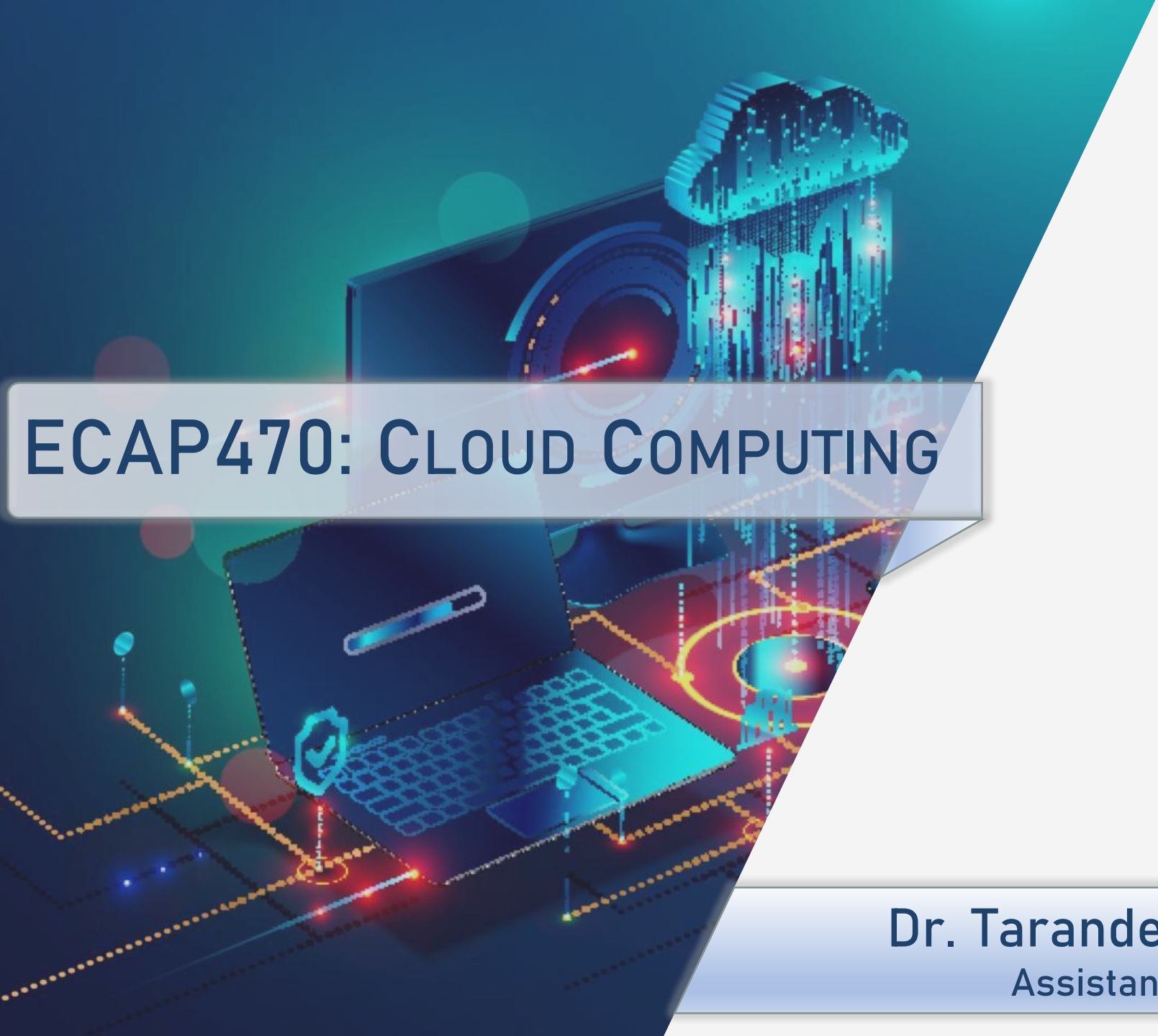


ECAP470: Cloud Computing

Dr. Tarandeep Kaur
Assistant Professor

Learning Outcomes

After this lecture, you will be able to,

- ✓ Learn about Big Data and its characteristics.
- ✓ Understand the different V's of Big Data.

Big Data

Collection of data sets so large & complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Big Data

Data whose scale, diversity, and complexity requires new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

Big Data

Correlations exist.

Facts & Figures

Walmart

Facebook

AT/T Database

Facts & Figures

Walmart

Facebook

AT/T Database

Facts & Figures

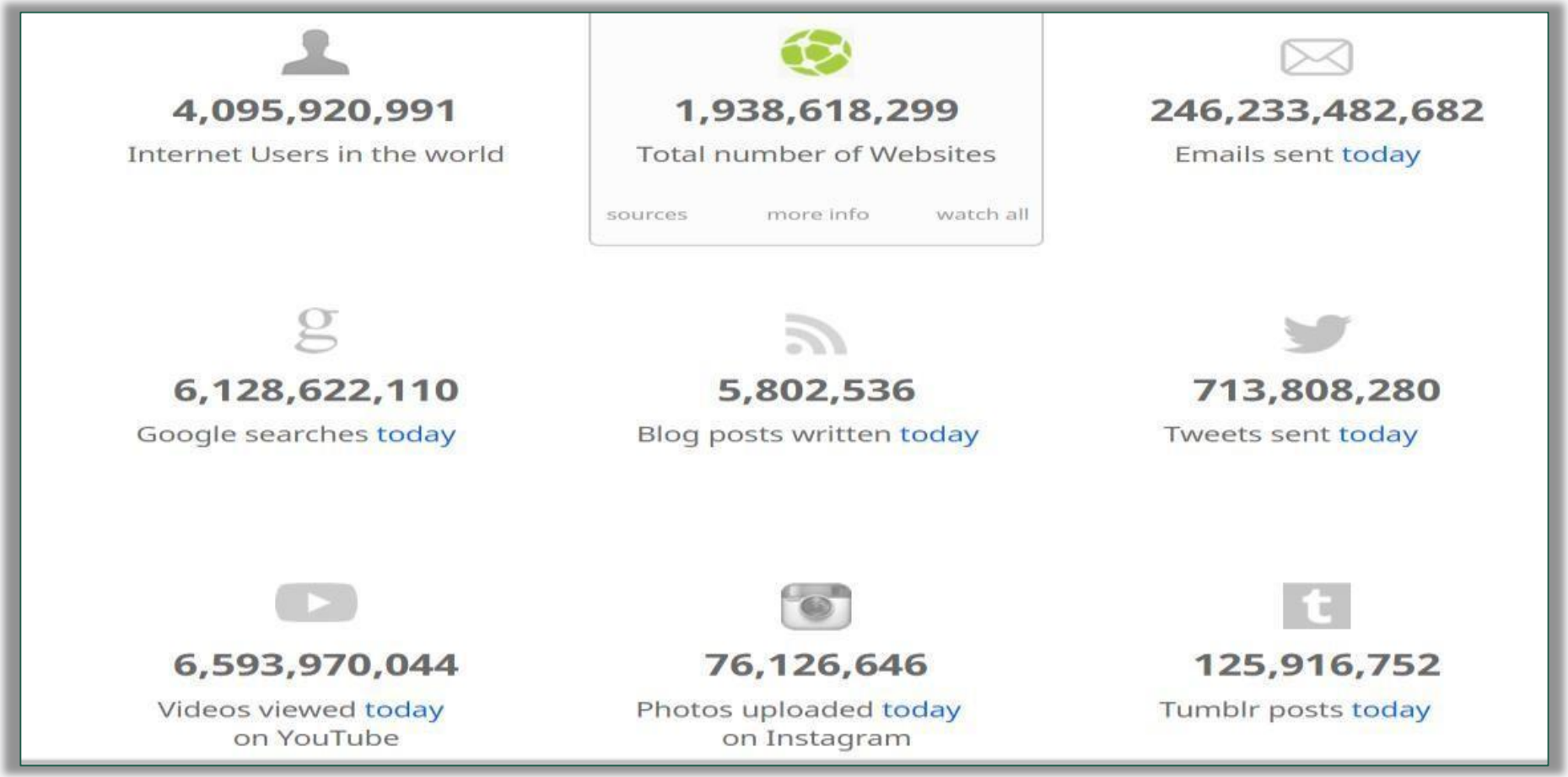
Walmart

Facebook

AT/T Database

From Where So Much Data Comes Up

Sources



From Where So Much Data Comes Up

Data on the Internet:

Internet Live Stats - Internet Usage & Social Media Statistics

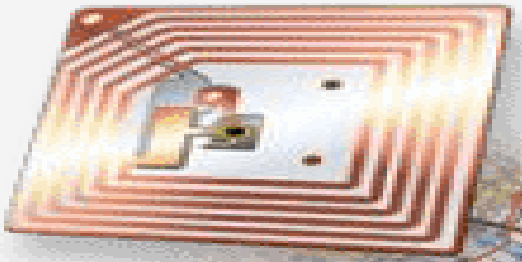
12+ TBs
of tweet data every day

? TBs of
data every day

25+ TBs **of log data every day**

From Where So Much Data Comes Up

30 billion RFID tags today
(1.3B in 2005)



4.6 billion camera phones world wide

100s of millions of GPS enabled Devices sold annually

76 million smart meters in 2009…
200M by 2014

2+ billion people on the Web by end 2011

Introduction to Big Data

Big Data Computing

From Where So Much Data Comes Up

- Progress & innovation is no longer hindered by the ability to collect data.
- Big Data facilitates the ability to manage, analyse, summarize, visualize & discover knowledge from the collected data in a timely manner & in a scalable fashion at a faster pace.

From Where So Much Data Comes Up

- Progress & innovation is no longer hindered by the ability to collect data.
- Big Data facilitates the ability to manage, analyse, summarize, visualize & discover knowledge from the collected data in a timely manner & in a scalable fashion at a faster pace.

From Where So Much Data Comes Up

Social media and networks
(all of us are generating data)

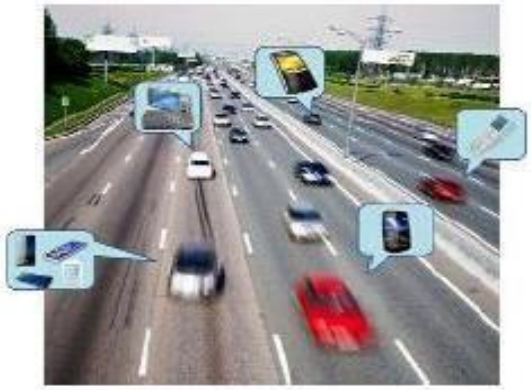
Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

An Example of Big Data at Work

Crowdsourcing



Computing

Sensing

Real time traffic info

Issues with Traditional Database Methods

- Traditional RDBMS queries
- Traditional tools
- Benefits of Big Data
- Other issues

Issues with Traditional Database Methods

Issues with Traditional Database Methods

Traditional RDBMS queries

Traditional tools

Benefits of Big Data

Other issues

Issues with Traditional Database Methods

- Traditional RDBMS queries
- Traditional tools
- Benefits of Big Data
- **Other issues**

Traditional Model Has Changed

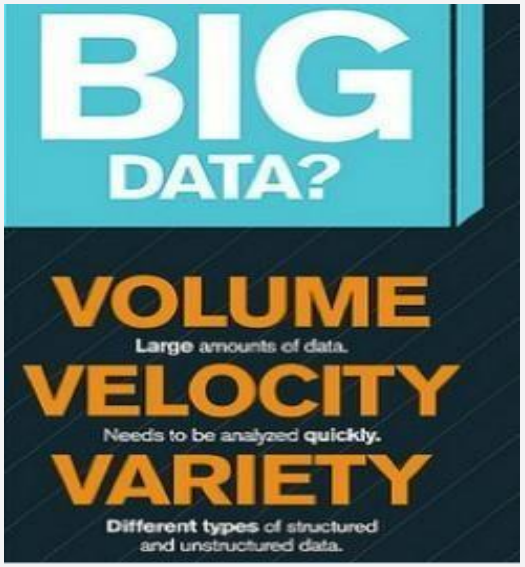
- Model of Generating/Consuming Data has Changed.
- Old Model

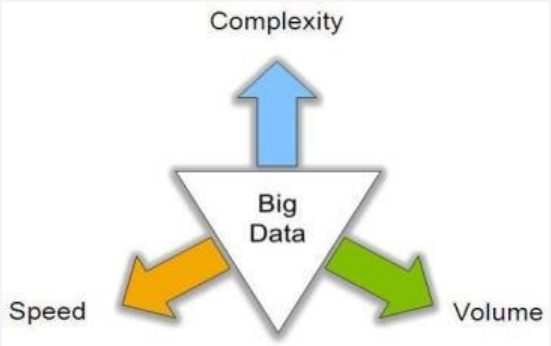
Traditional Model Has Changed

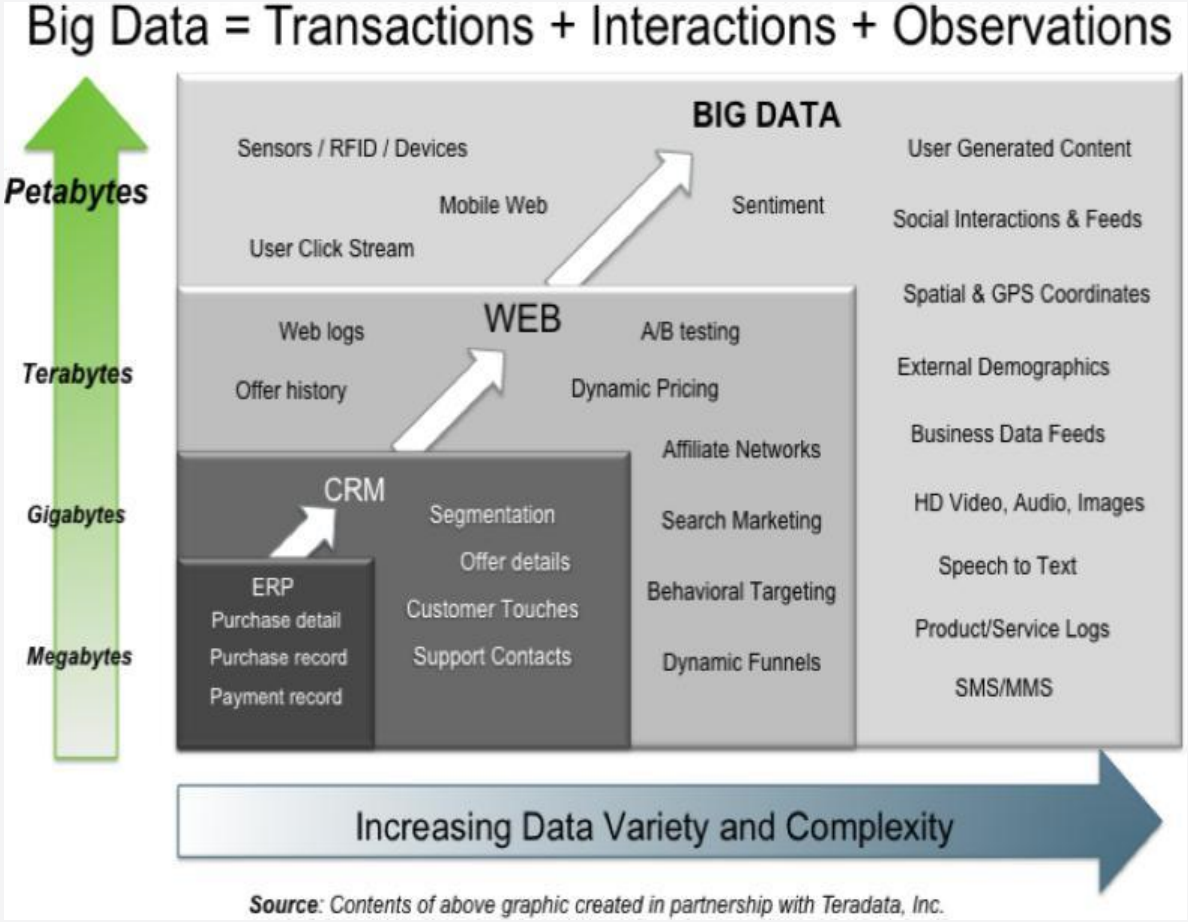
New Model

Big Data (3V's)

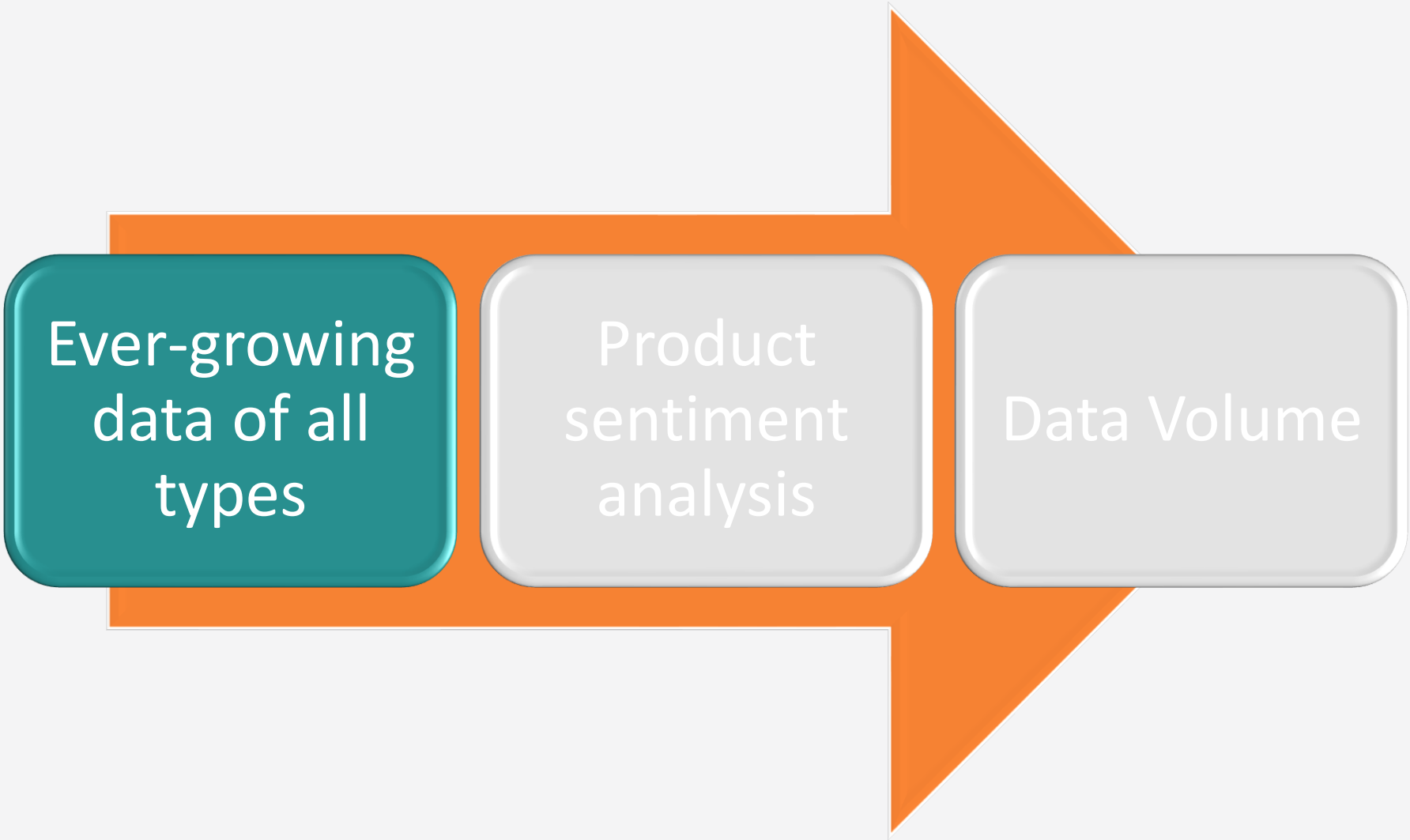
3V's: Volume, Velocity and Variety



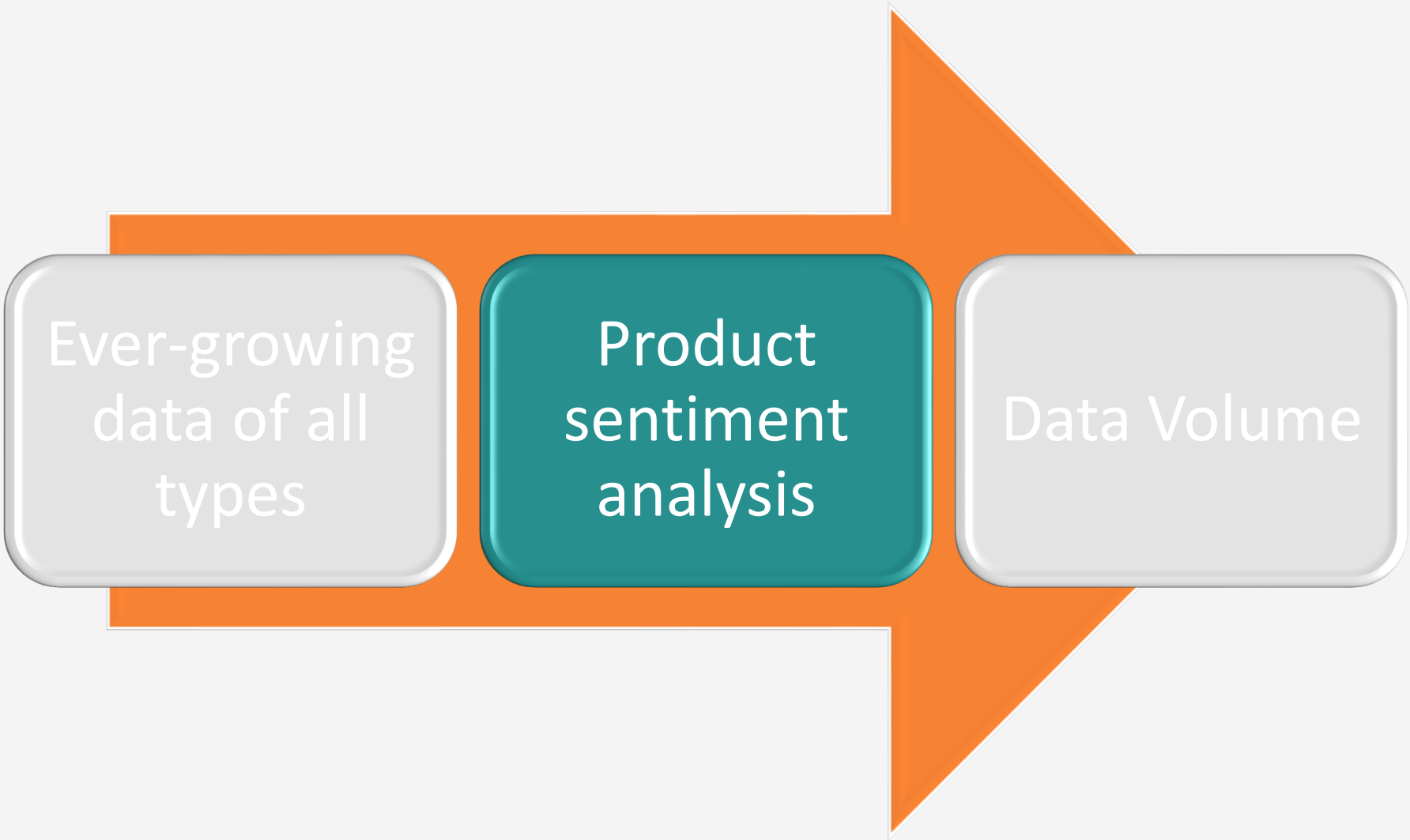




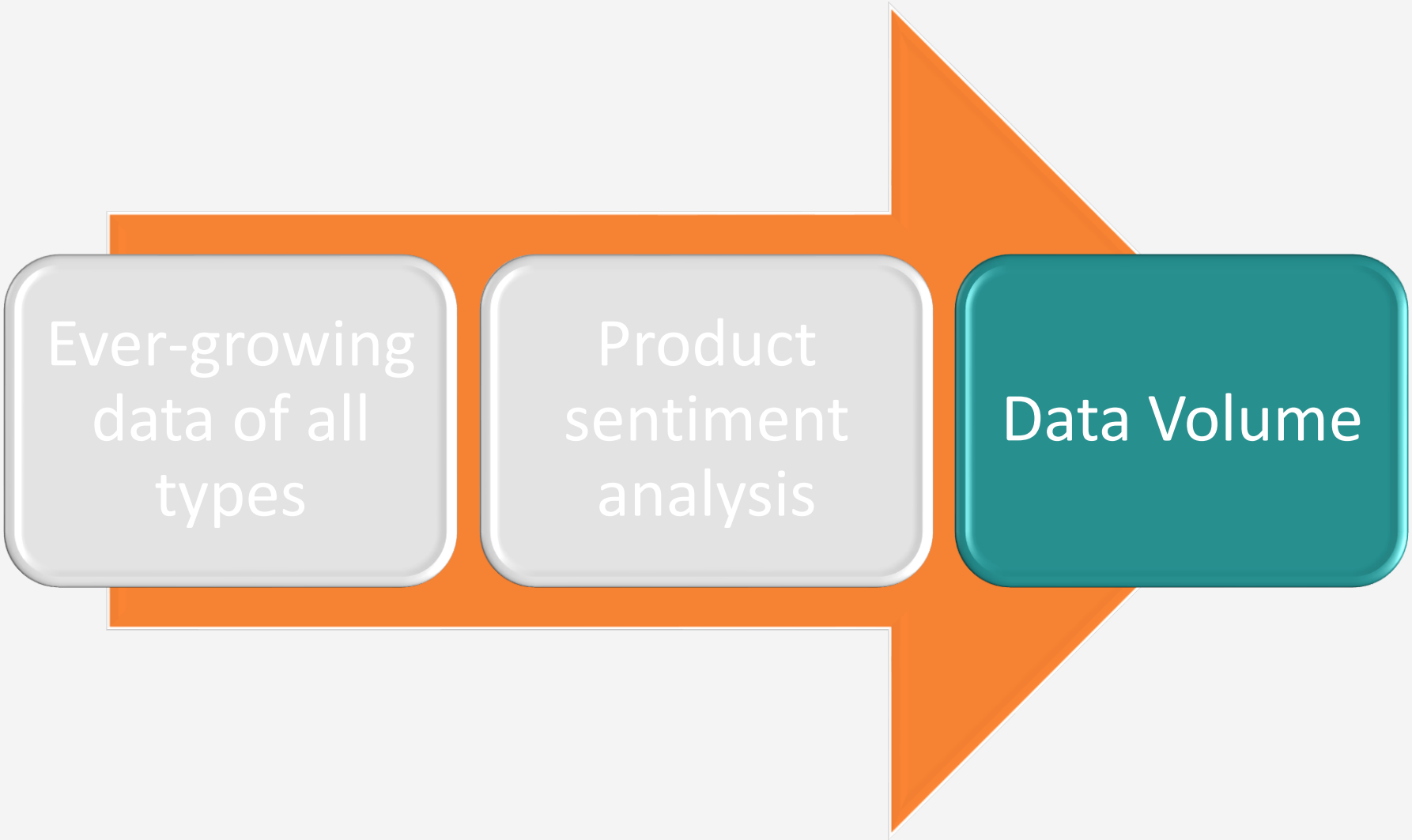
Volume (Scale)



Volume (Scale)

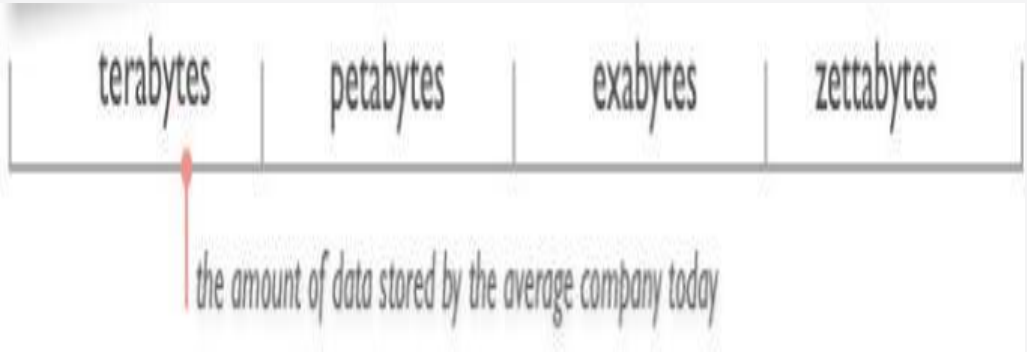


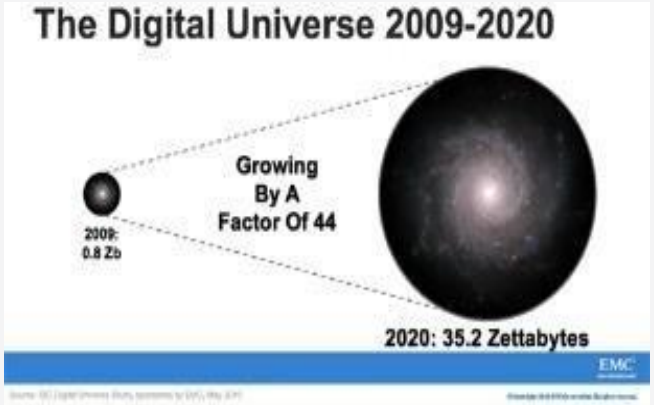
Volume (Scale)

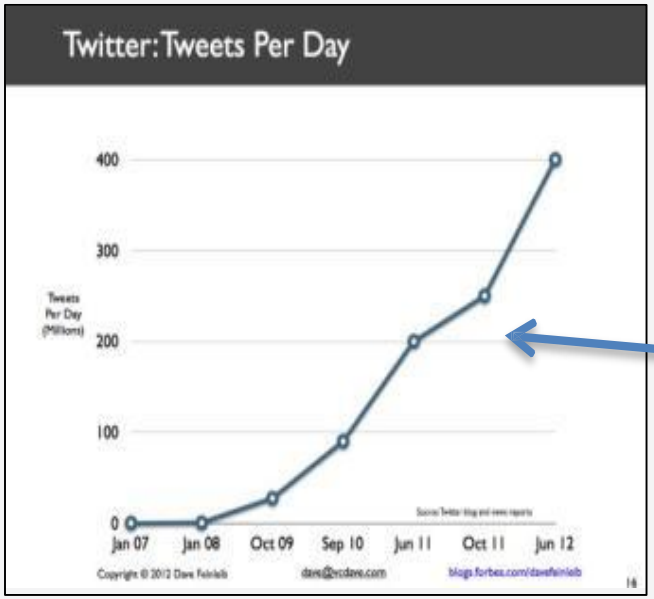


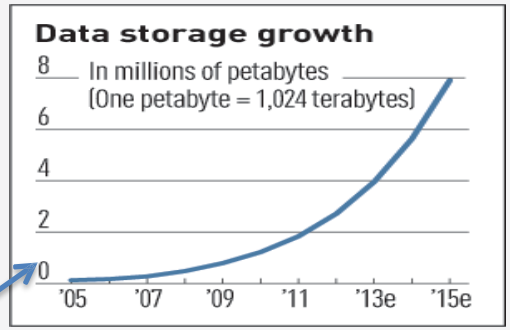
Volume (Scale)

Data Volume









Volume (Scale)

Earthscope

- World's largest science project.
- Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles amassing 67 terabytes of data.

Volume (Scale)

Earthscope

- Analyzes seismic slips in San Andreas fault, but also plume of magma underneath Yellowstone & much more.

Velocity (Speed)

- Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used.
- Data is begin generated fast & needs to be processed fast.

Velocity (Speed)

- Online Data Analytics
- Late decisions→→missing opportunities

Examples:

  - E-Promotions
  - Healthcare monitoring

Variety (Complexity)

Big data can handle any type of data–

- Structured Data: (example: tabular data)
- Unstructured Data: Text, sensor data, audio, video
- Semi Structured: Web data, log files

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

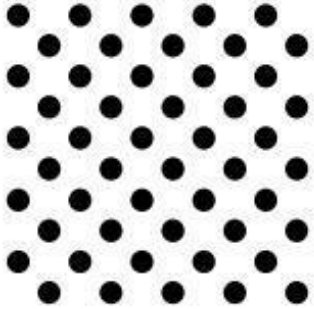
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML) Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- A single application can be generating/ collecting many types of data
- Big Public Data (online, weather, finance etc)

3 Big V's (+1)

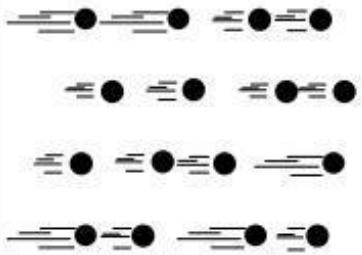
Volume



Data at Rest

Terabytes to exabytes of existing data to process

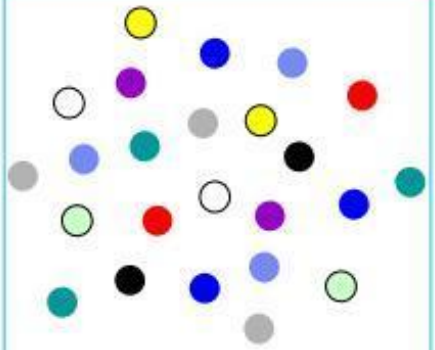
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

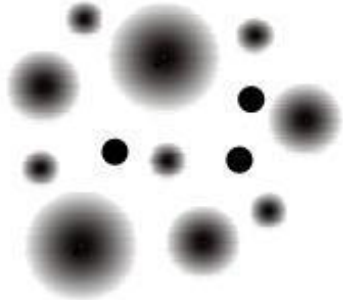
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

3 Big V's (+1)

Plus 1:

Value

3 Big V's (+1) (+ N more)

- Veracity
- Valence
- Validity
- Variability
- Viscosity & Volatility
- Viability
- Venue
- Vocabulary
- Vagueness ….

Valence

Refers to the connectedness of big data.

Validity

Accuracy and correctness of the data relative to a particular use.

Example: Gauging storm intensity.

Variability

- How meaning of the data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

Variability

- How meaning of the data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

Variability

- How meaning of the data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

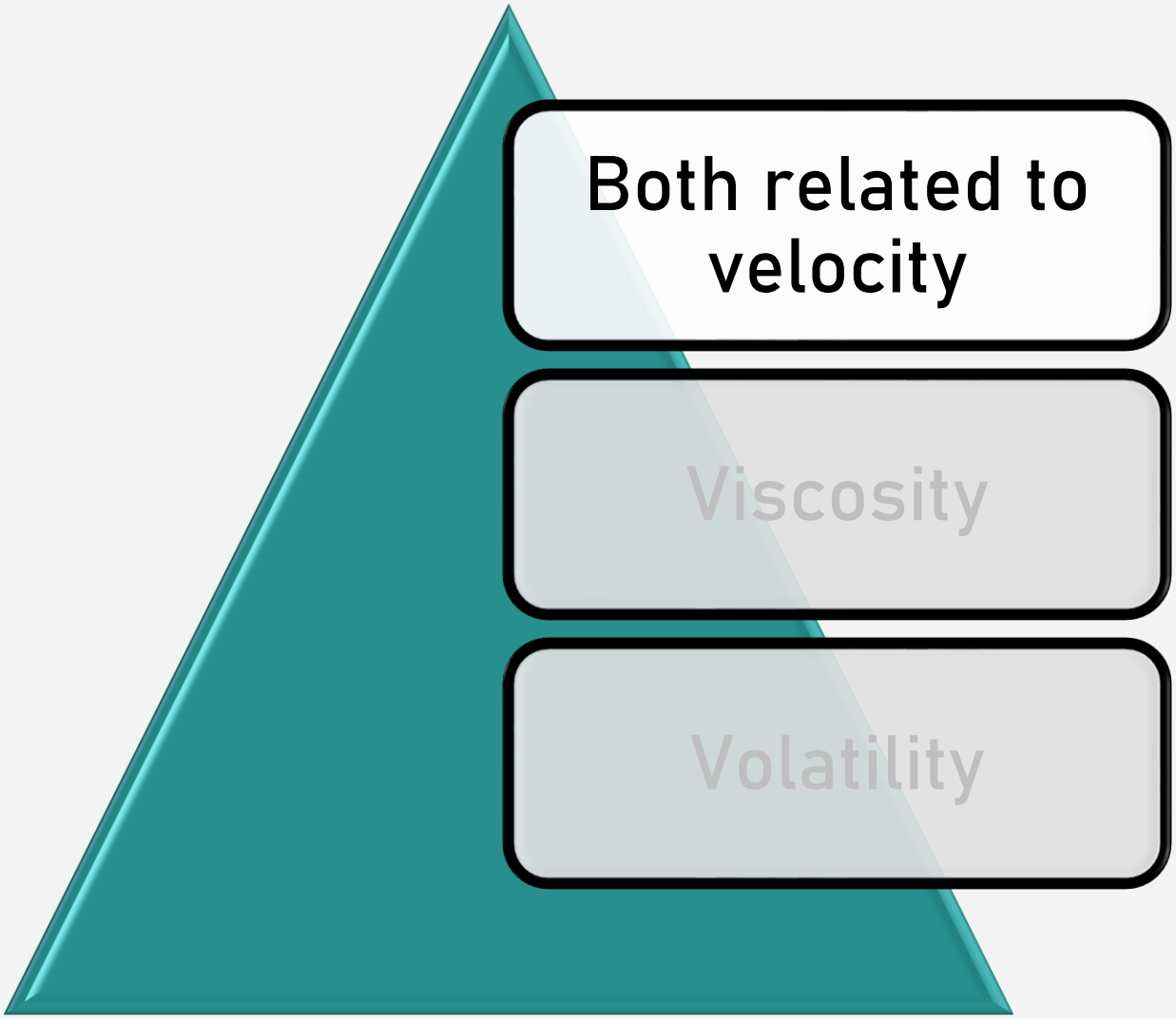
Variability

- How meaning of the data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

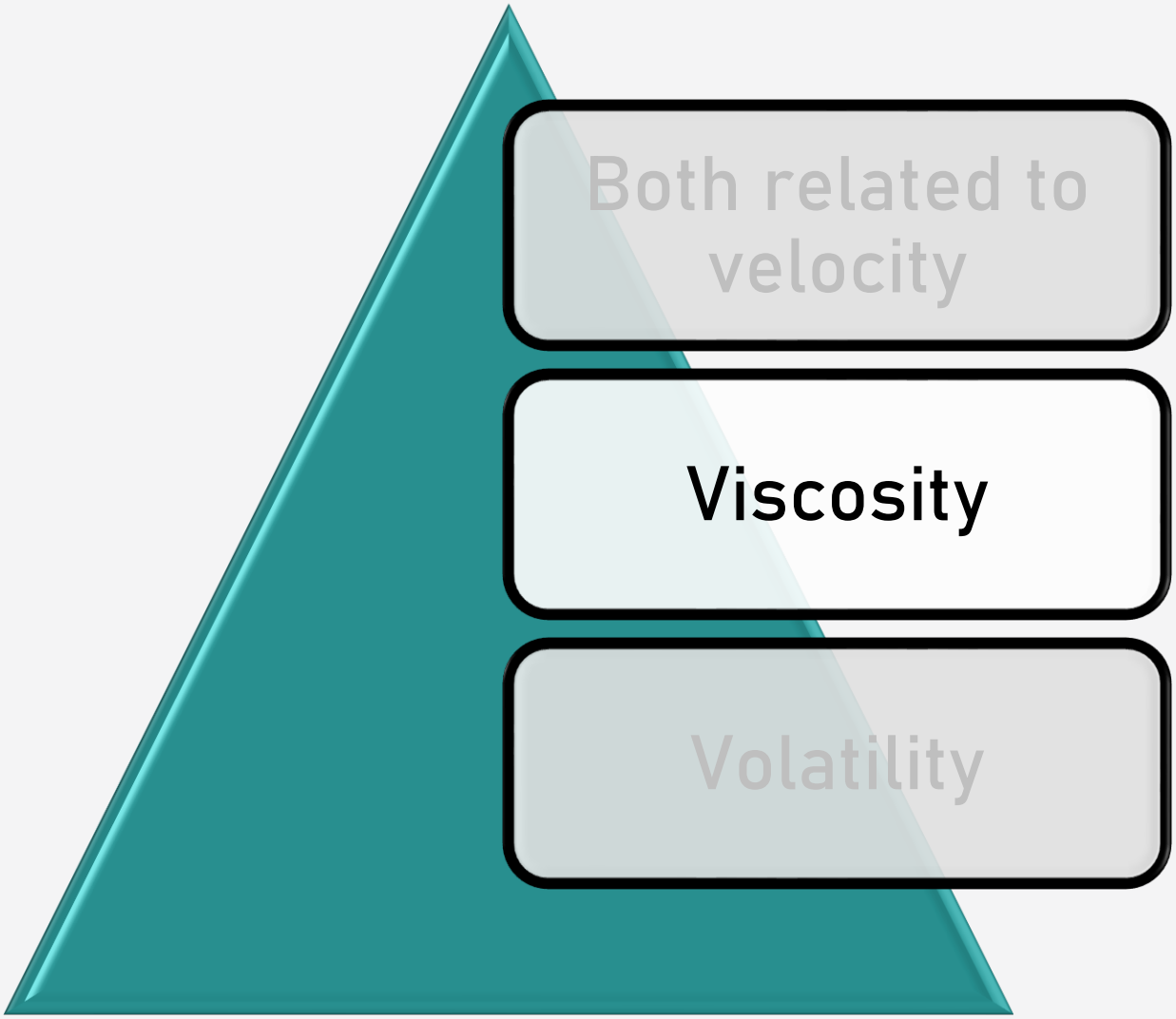
Variability

- How meaning of the data changes over time
- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

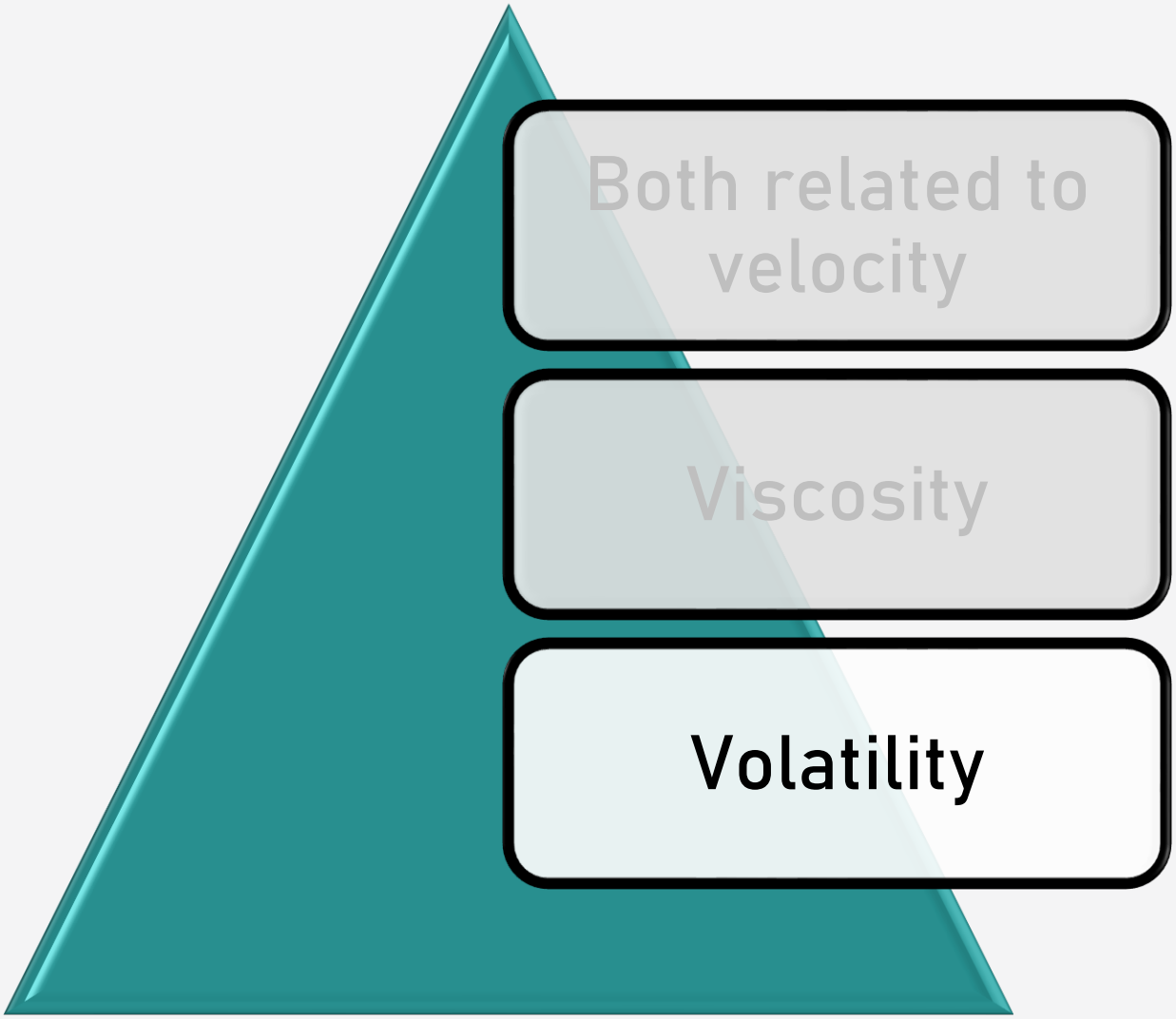
Viscosity & Volatility



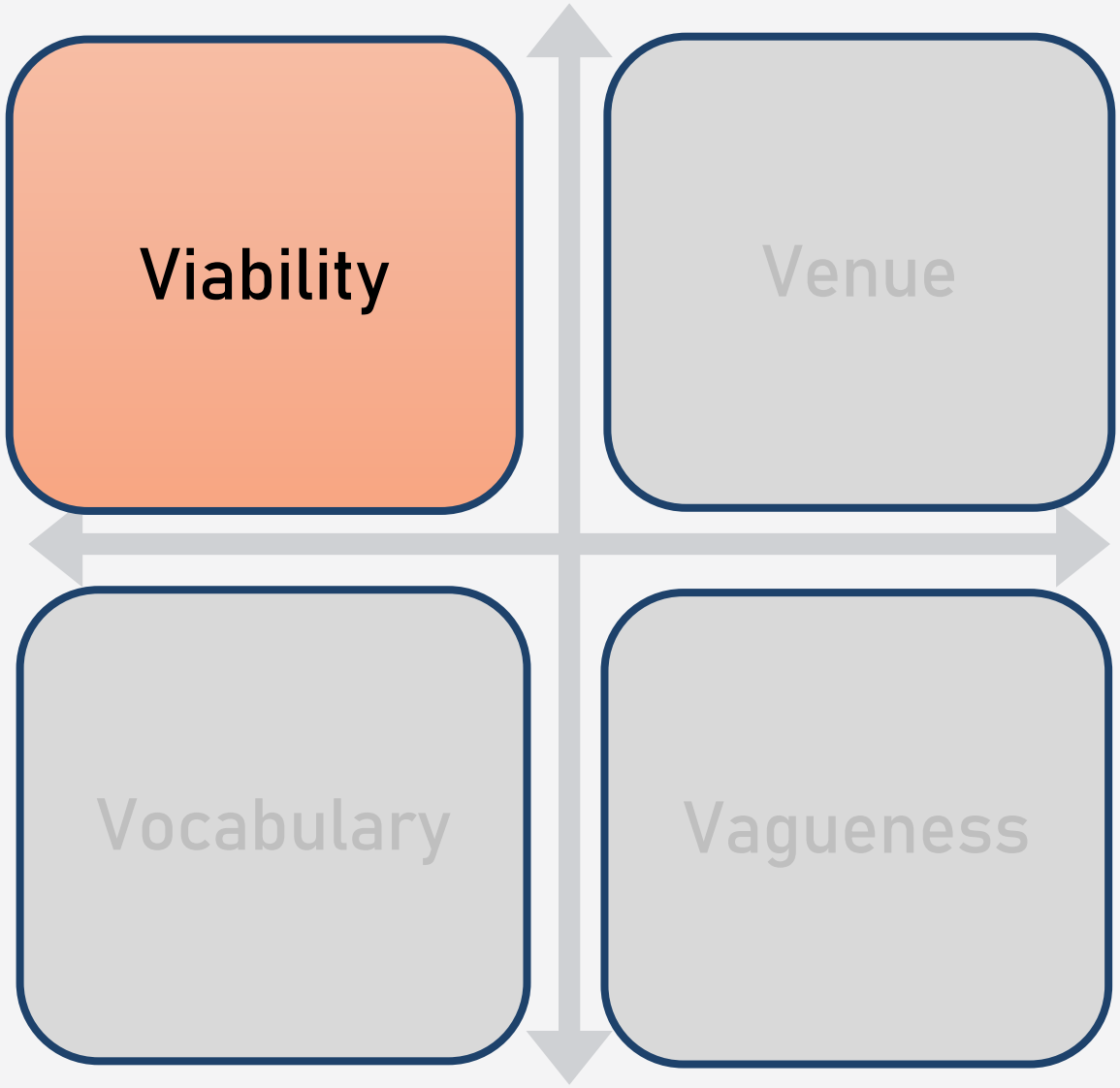
Viscosity & Volatility



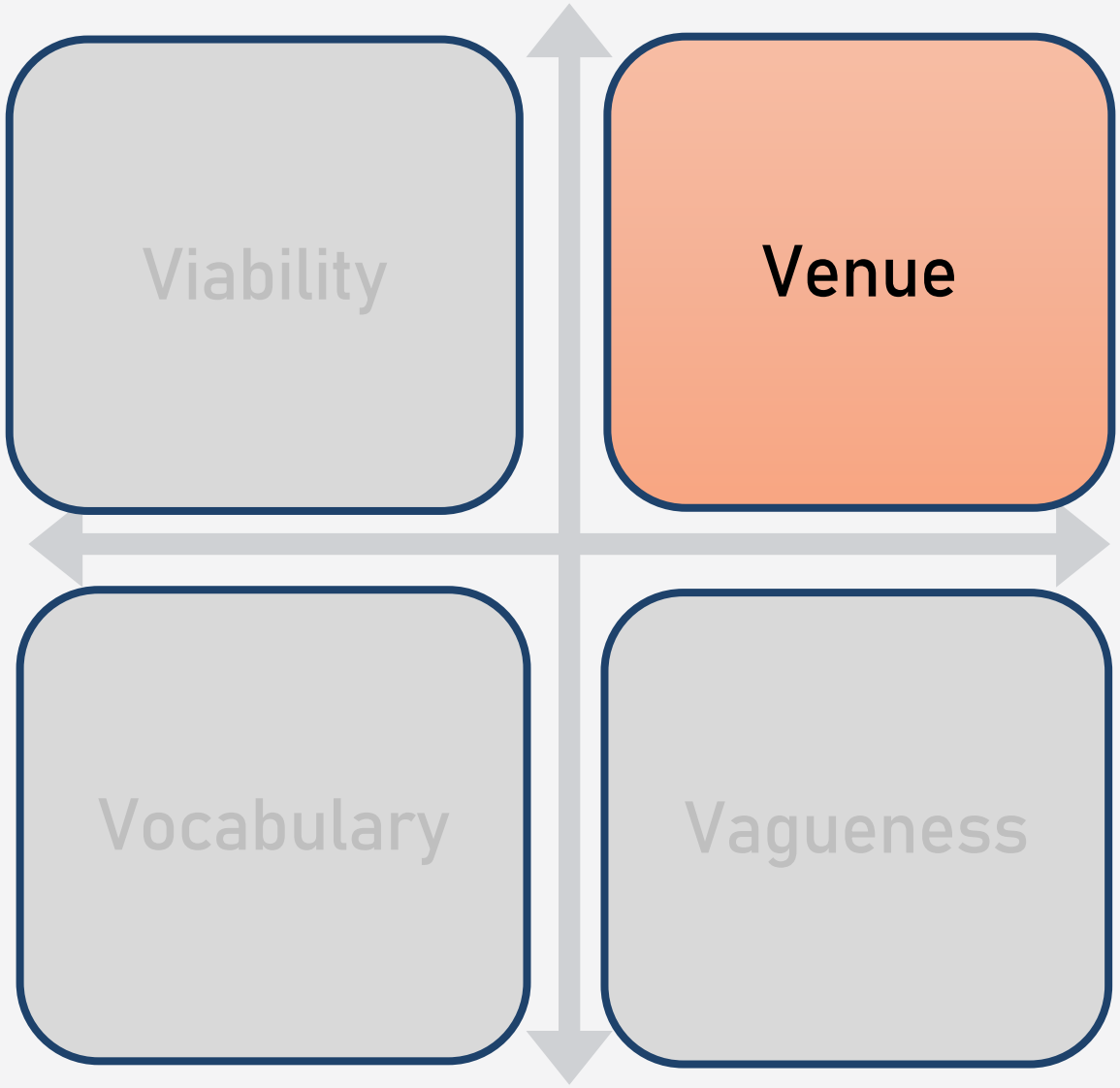
Viscosity & Volatility



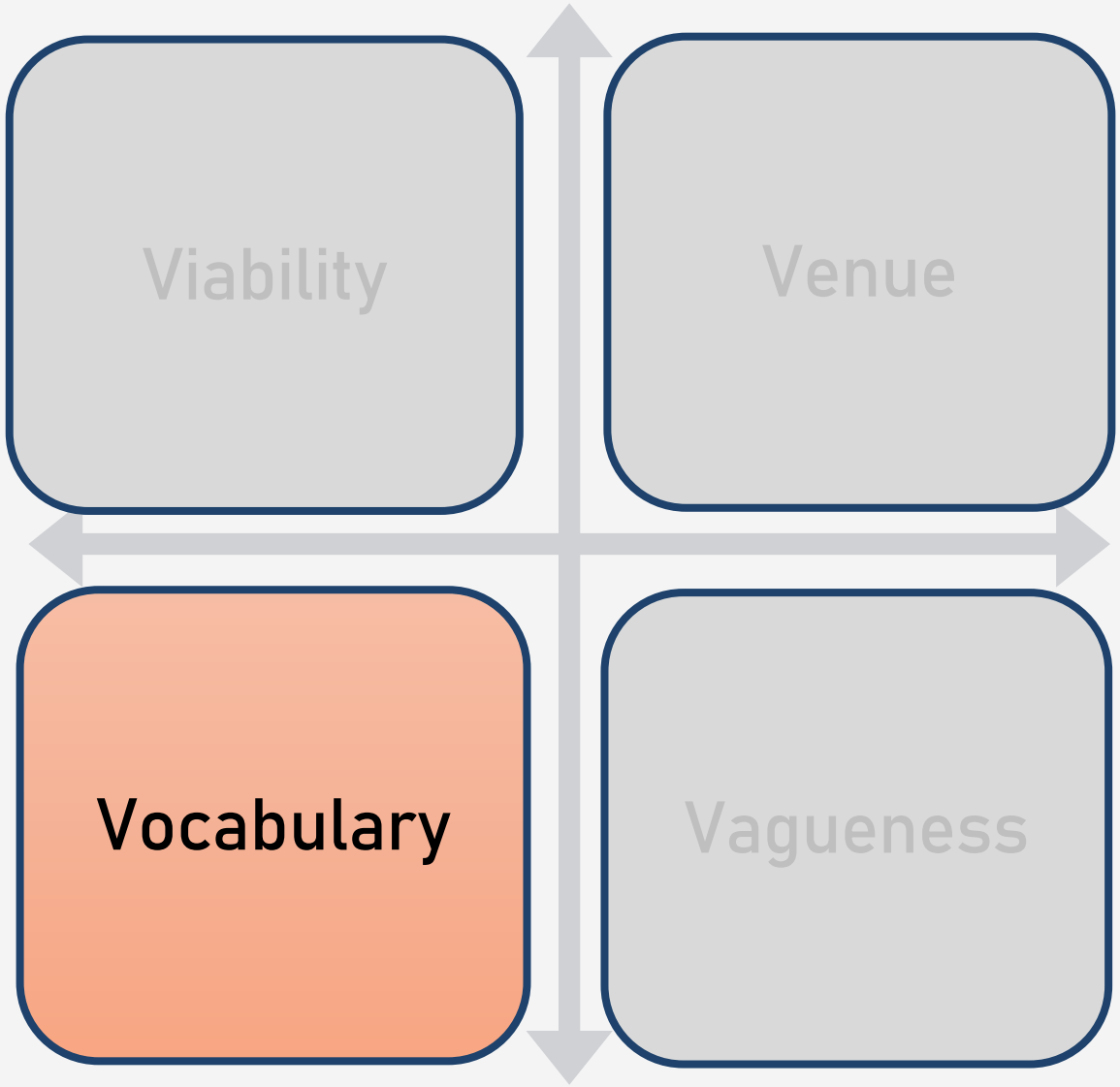
More V's



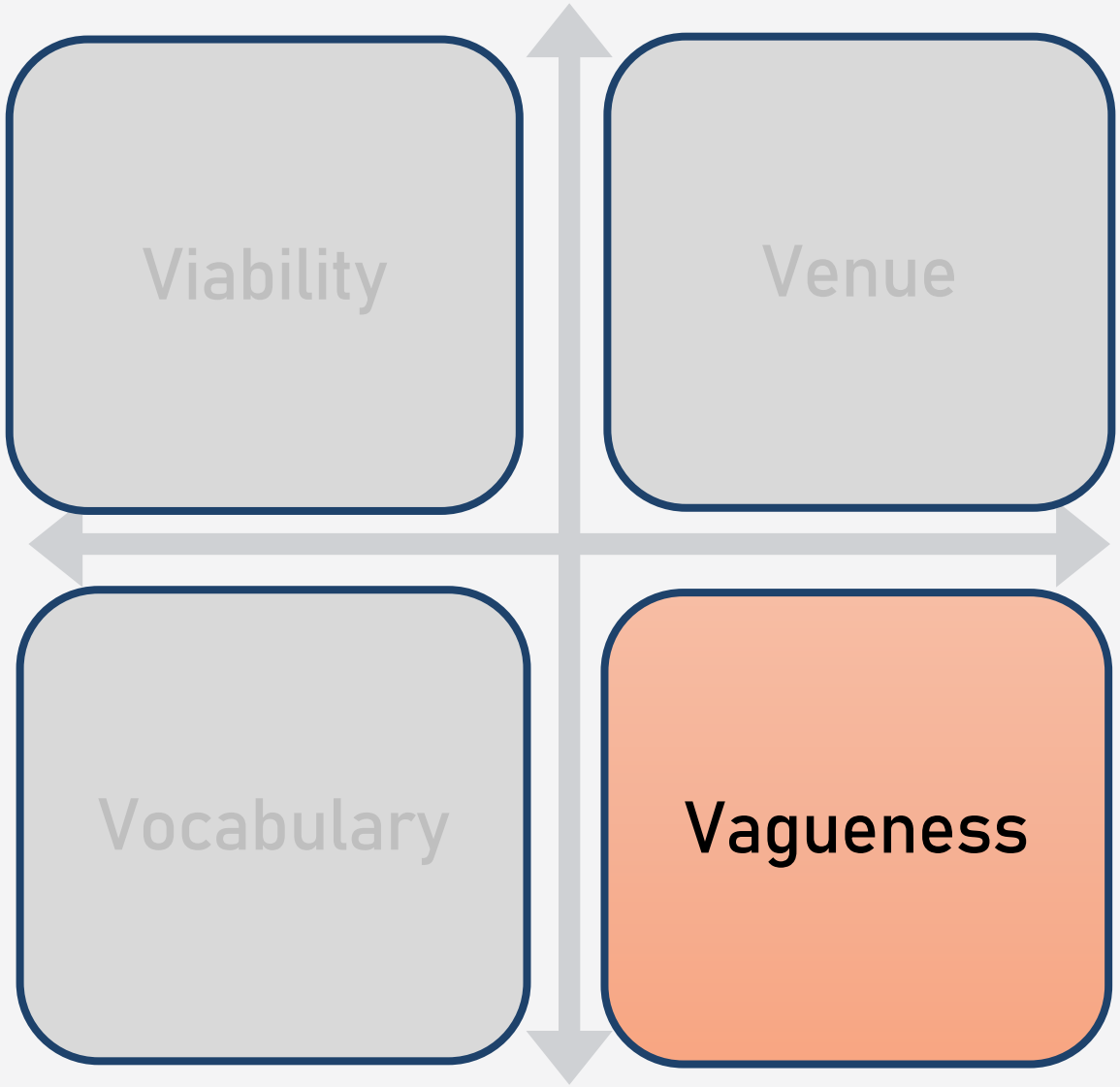
More V's



More V's



More V's

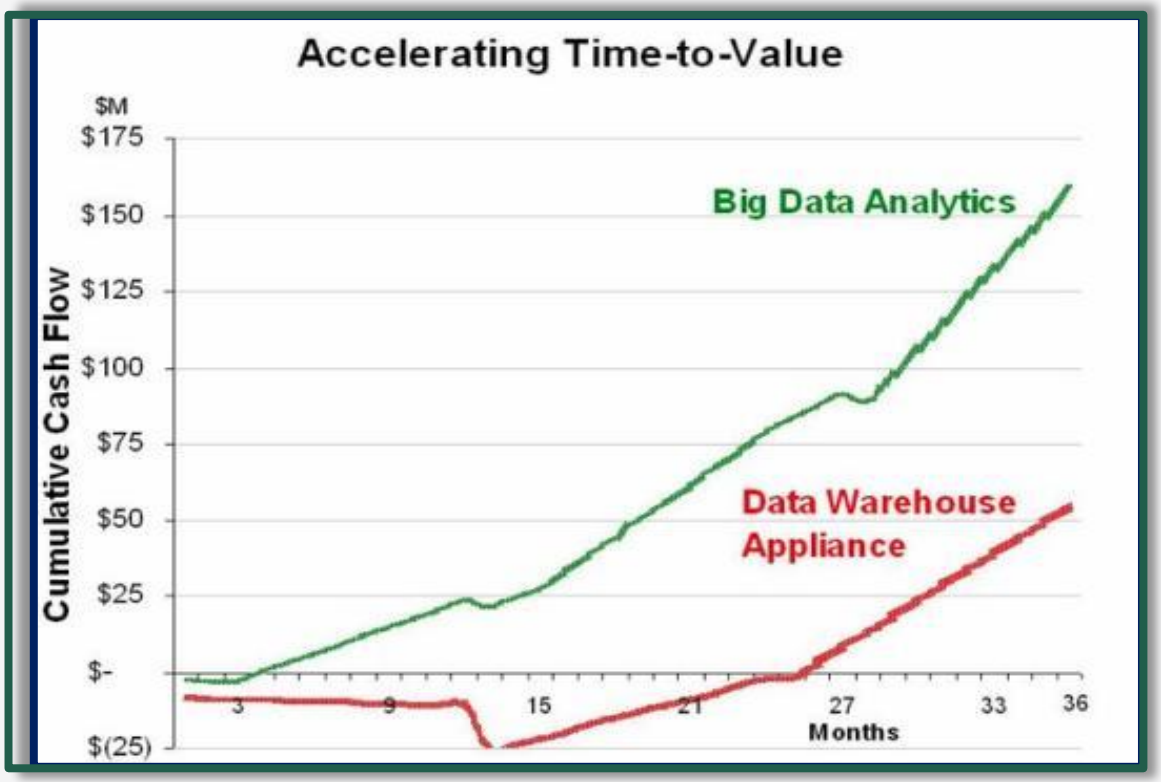


Big Data Analytics

- How to manage very large amounts of data and extract value and knowledge from them.
- Big data is more real-time in nature than traditional Data Warehouse (DW) applications.

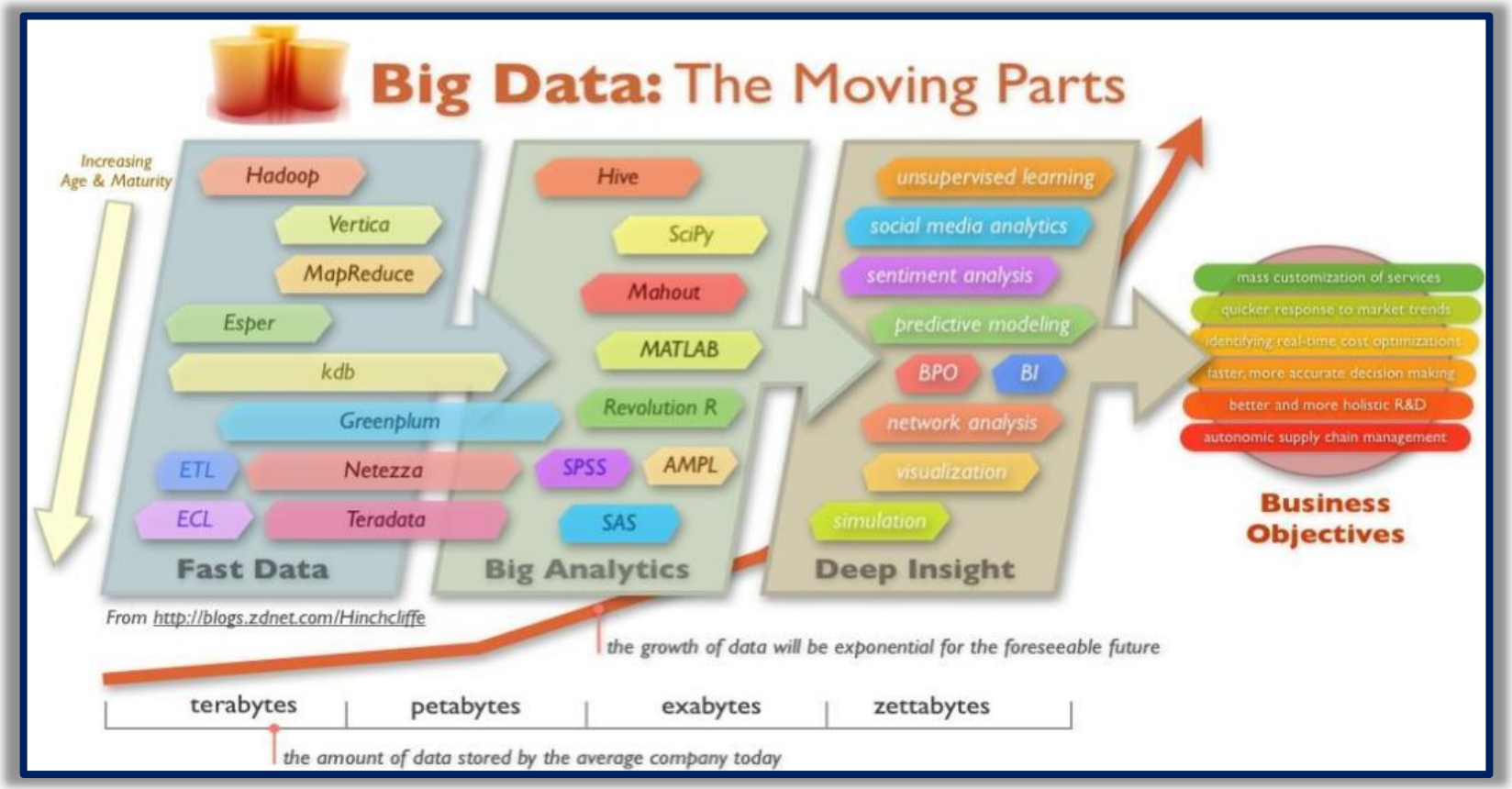
Big Data Analytics

Traditional DW architectures (Exadata, Teradata) are not well-suited for big data apps.



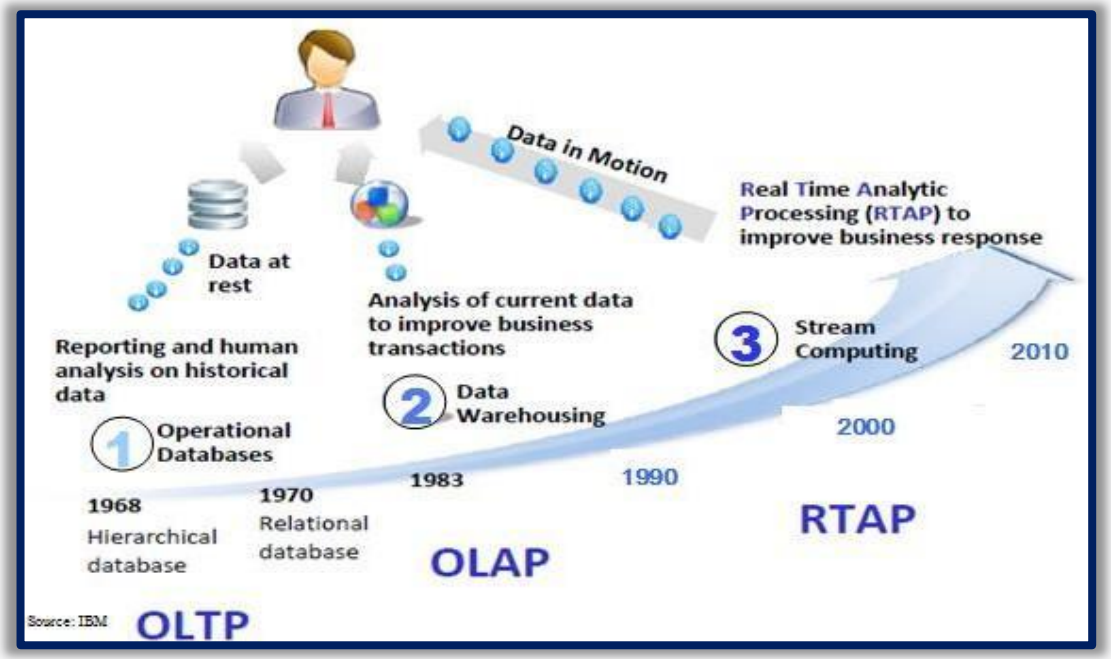
Big Data Analytics

Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps.



Harnessing Big Data

- OLAP: Online Analytical Processing (DBMSs)
- OLTP: Online Transaction Processing (Data Warehousing)
- RTAP: Real-Time Analytics Processing (Big Data Architecture & Technology)



Big Data Processing

Stage 1: Business case

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Stage 4: Data extraction

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Stage 4: Data extraction

Stage 5: Data aggregation

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Stage 4: Data extraction

Stage 5: Data aggregation

Stage 6: Data analysis

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Stage 4: Data extraction

Stage 5: Data aggregation

Stage 6: Data analysis

Stage 7: Visualization of data

Big Data Processing

Stage 1: Business case

Stage 2: Identification of data

Stage 3: Data filtering

Stage 4: Data extraction

Stage 5: Data aggregation

Stage 6: Data analysis

Stage 7: Visualization of data

Stage 8: Final analysis result

Real-Time Analytics/ Decision Requirement

Product Recommendations that are Relevant & Compelling

Influence Behavior

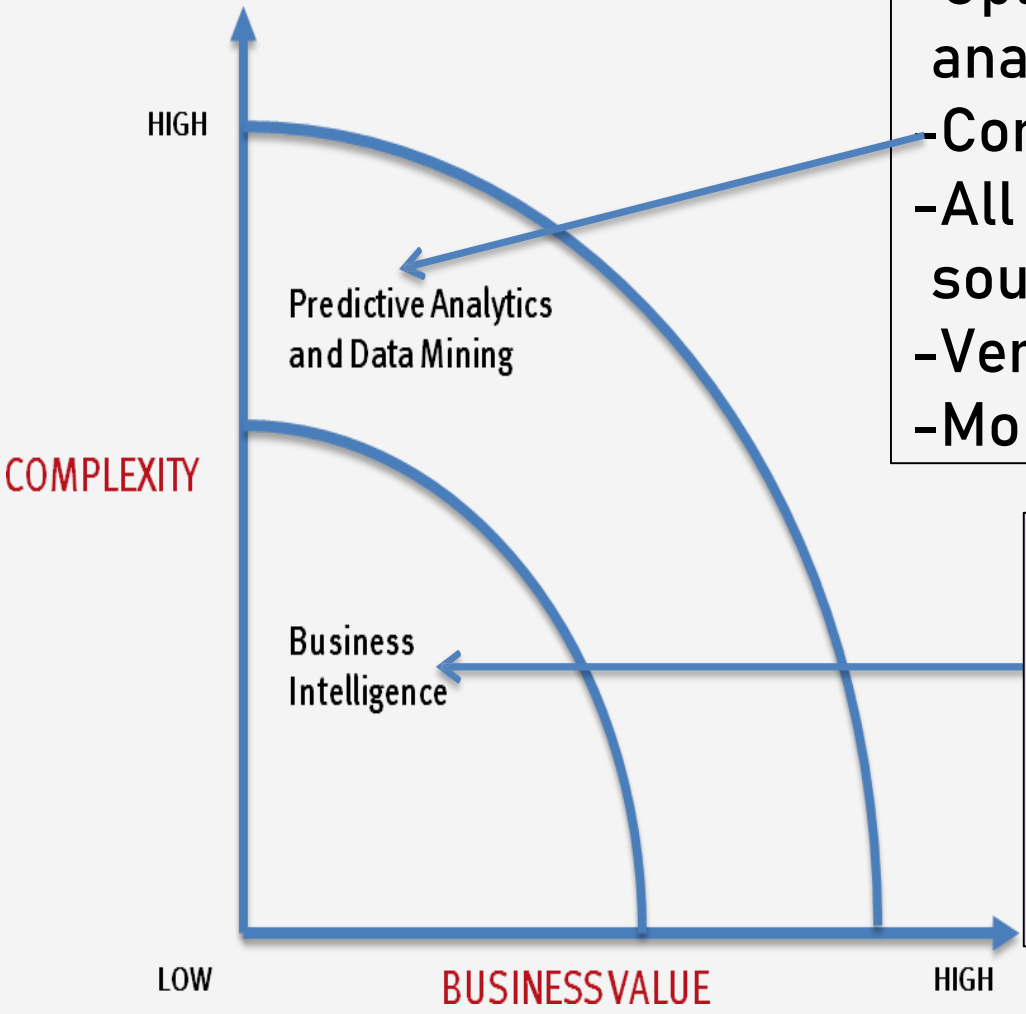
Learning why Customers Switch to competitors and their offers; in time to Counter

Improving the Marketing Effectiveness of a Promotion while it is still in Play

Preventing Fraud as it is Occurring & preventing more proactively

Friend Invitations to join a Game or Activity that expands business

What's Driving Big Data

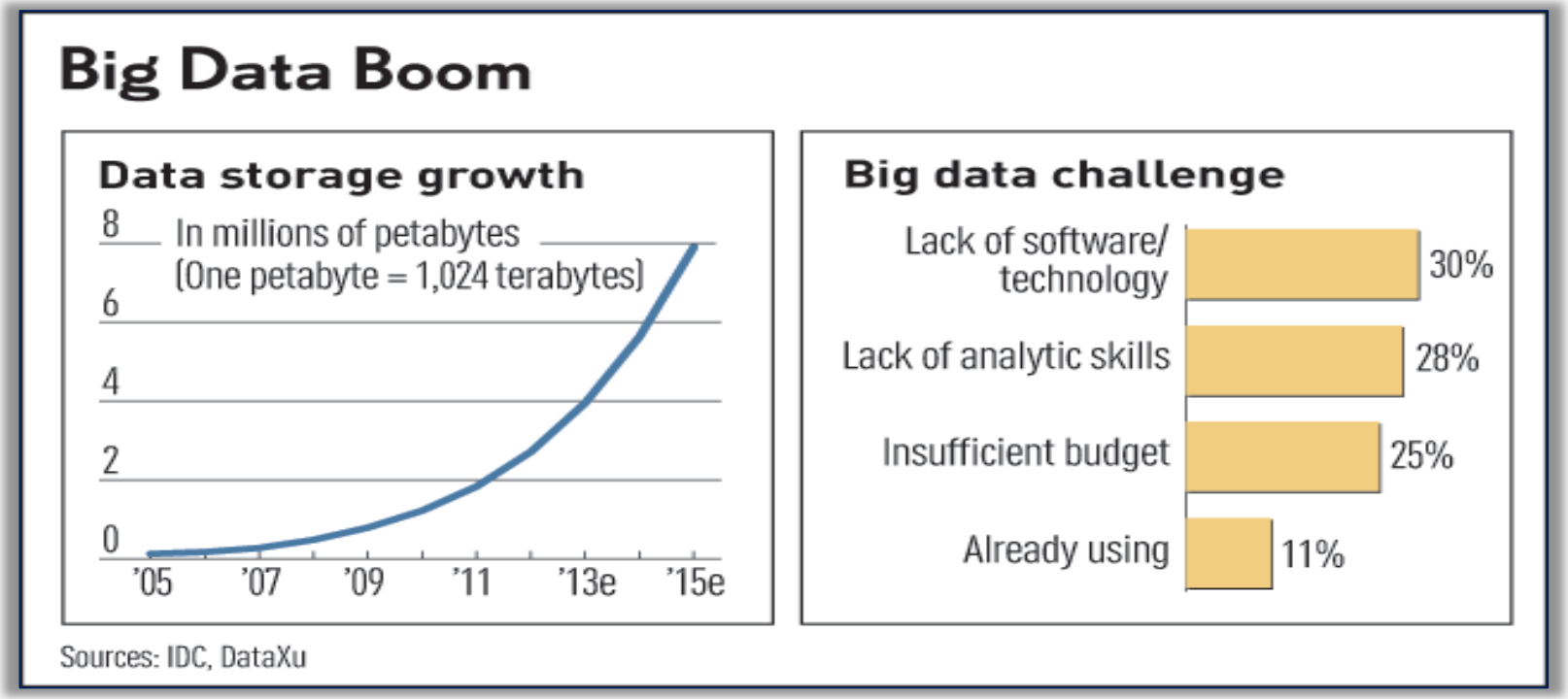


-Optimizations and predictive analytics
-Complex statistical analysis
-All types of data, and many sources
-Very large datasets
-More of a real-time

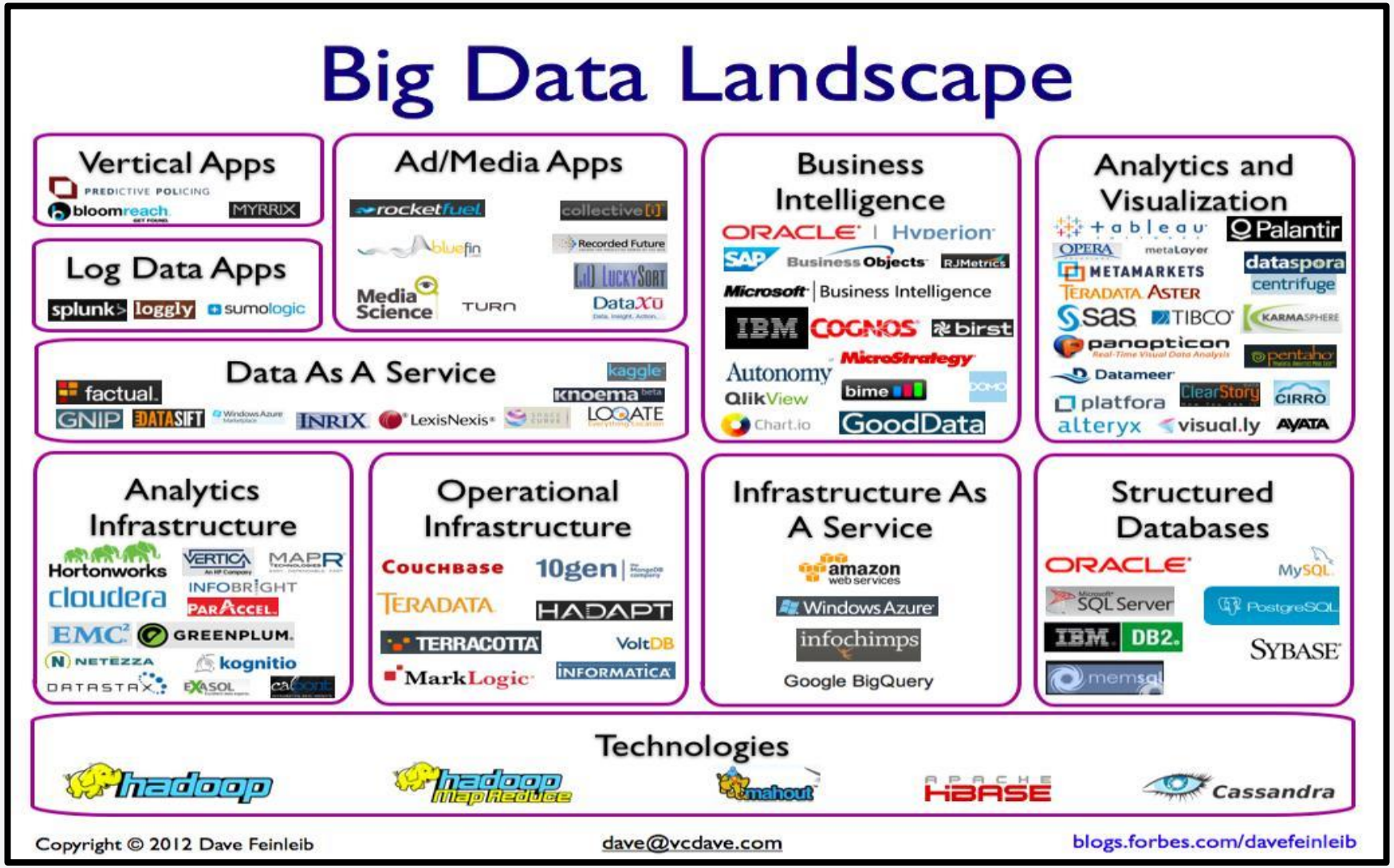
-Ad-hoc querying and reporting
-Data mining techniques
-Structured data, typical sources
-Small to mid-size datasets

Challenges in Handling Big Data

- Bottlenecks in technology
- Technical skills
- Other challenges



Big Data Landscape



That's all for now...