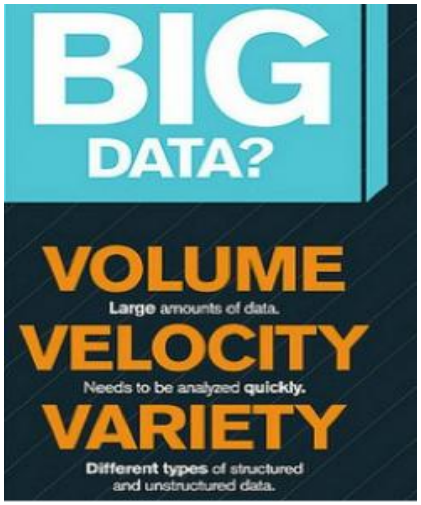
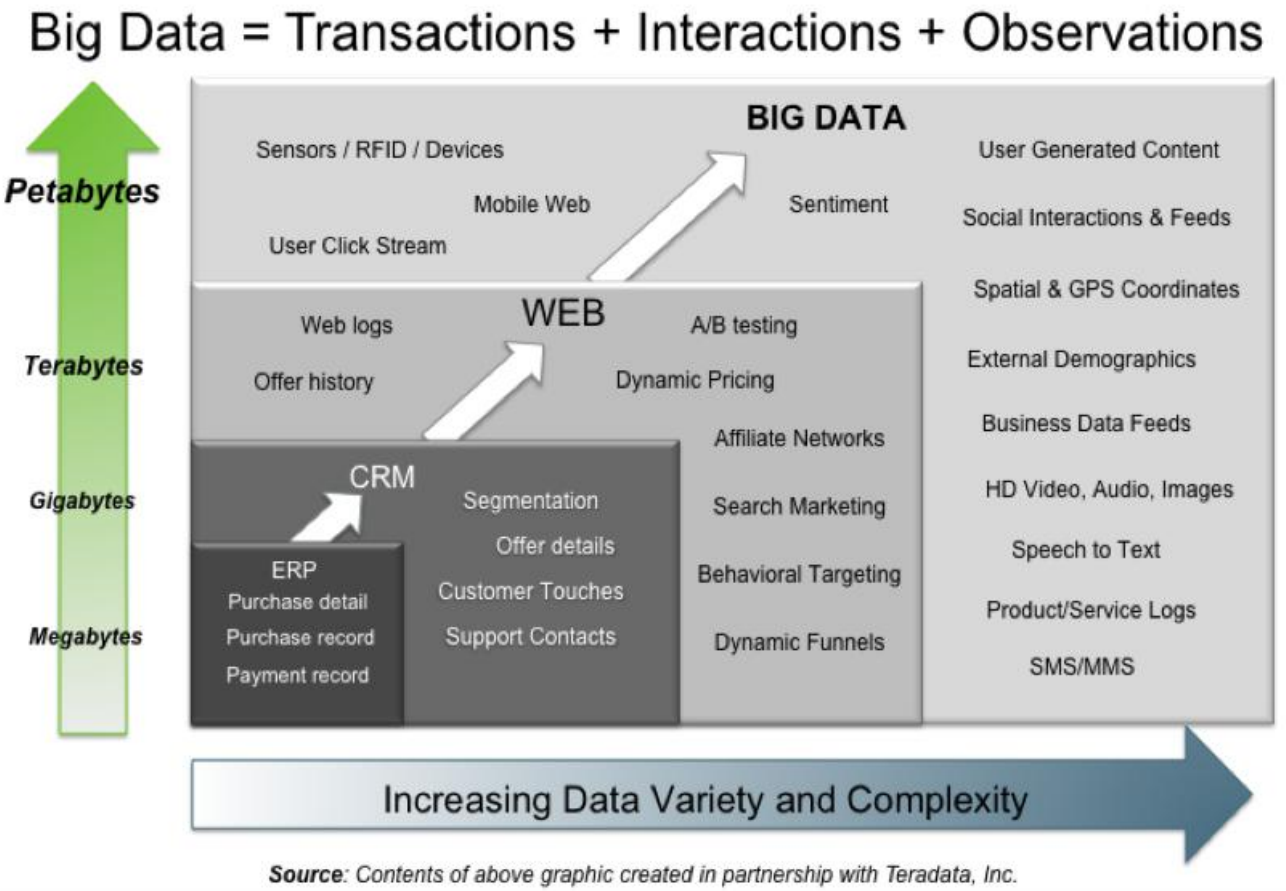
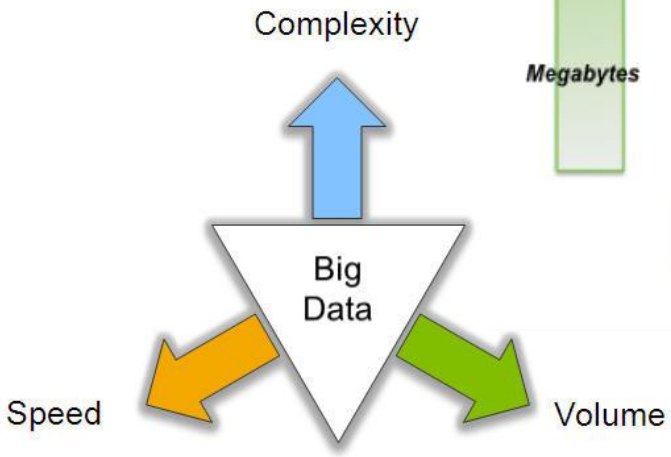
BIG DATA CHARACTERISTICS

Big Data: 3V's

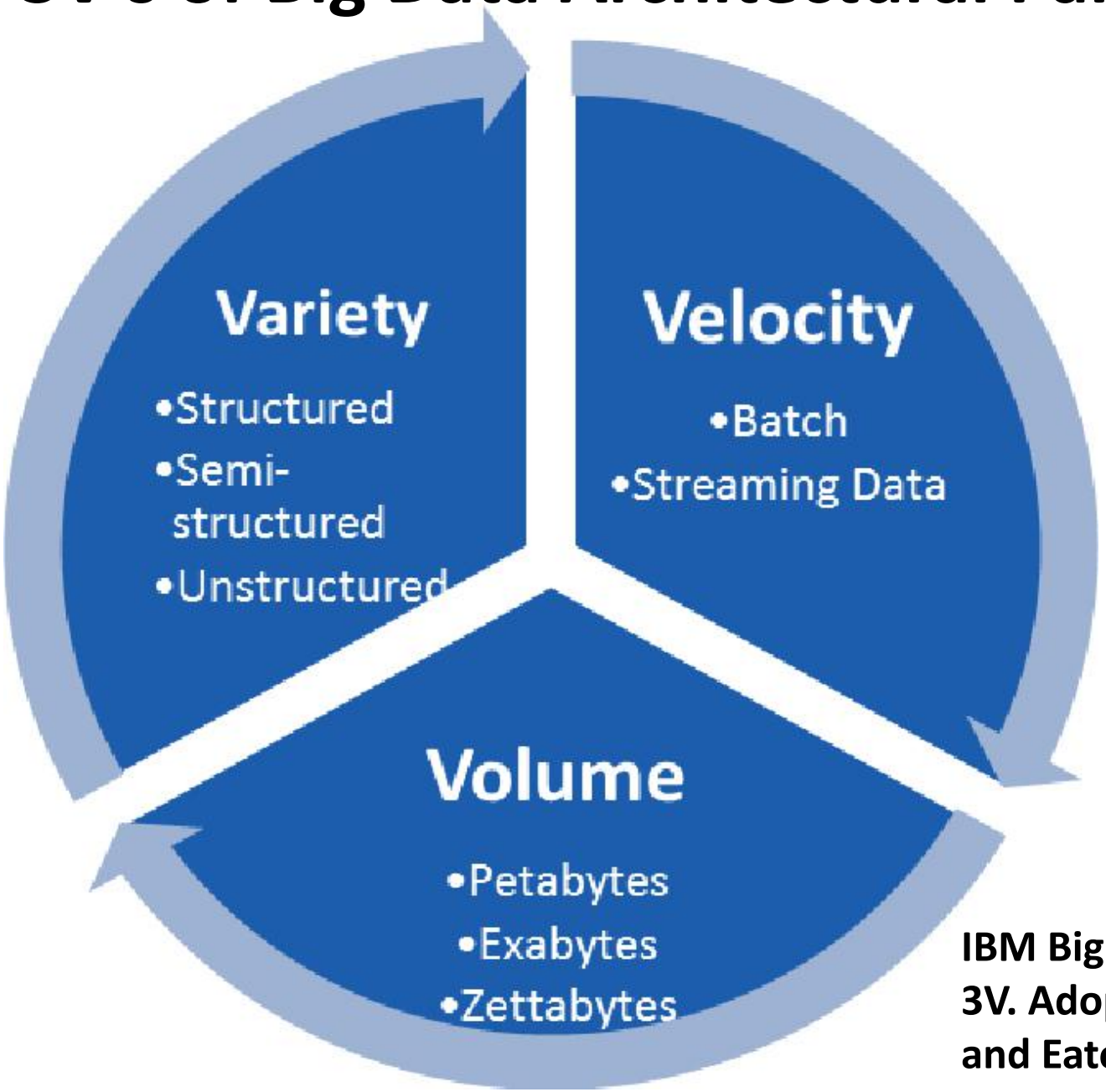




***Source**: Contents of above graphic created in partnership with Teradata, Inc.*



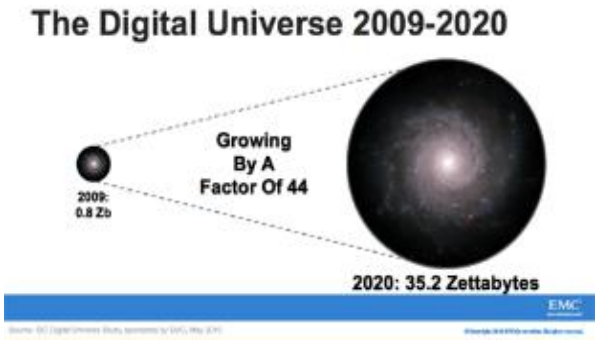
3V's of Big Data Architectural Paradigms

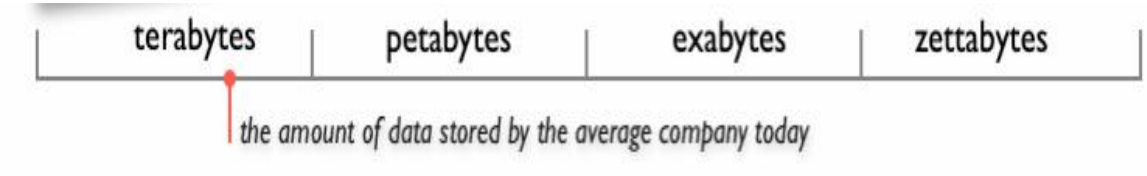


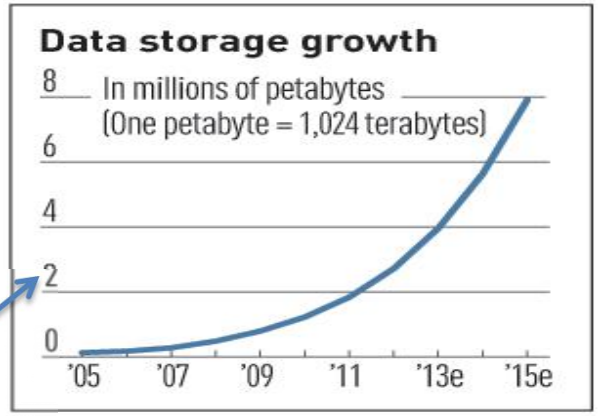
IBM Big Data characteristics – 3V. Adopted from (Zikopoulos and Eaton 2011)

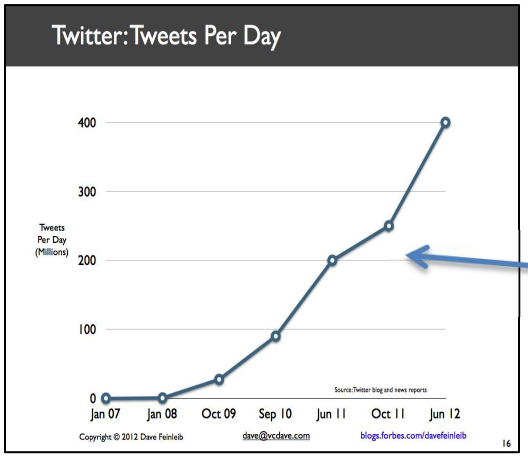
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

12+ TBs
of tweet data every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

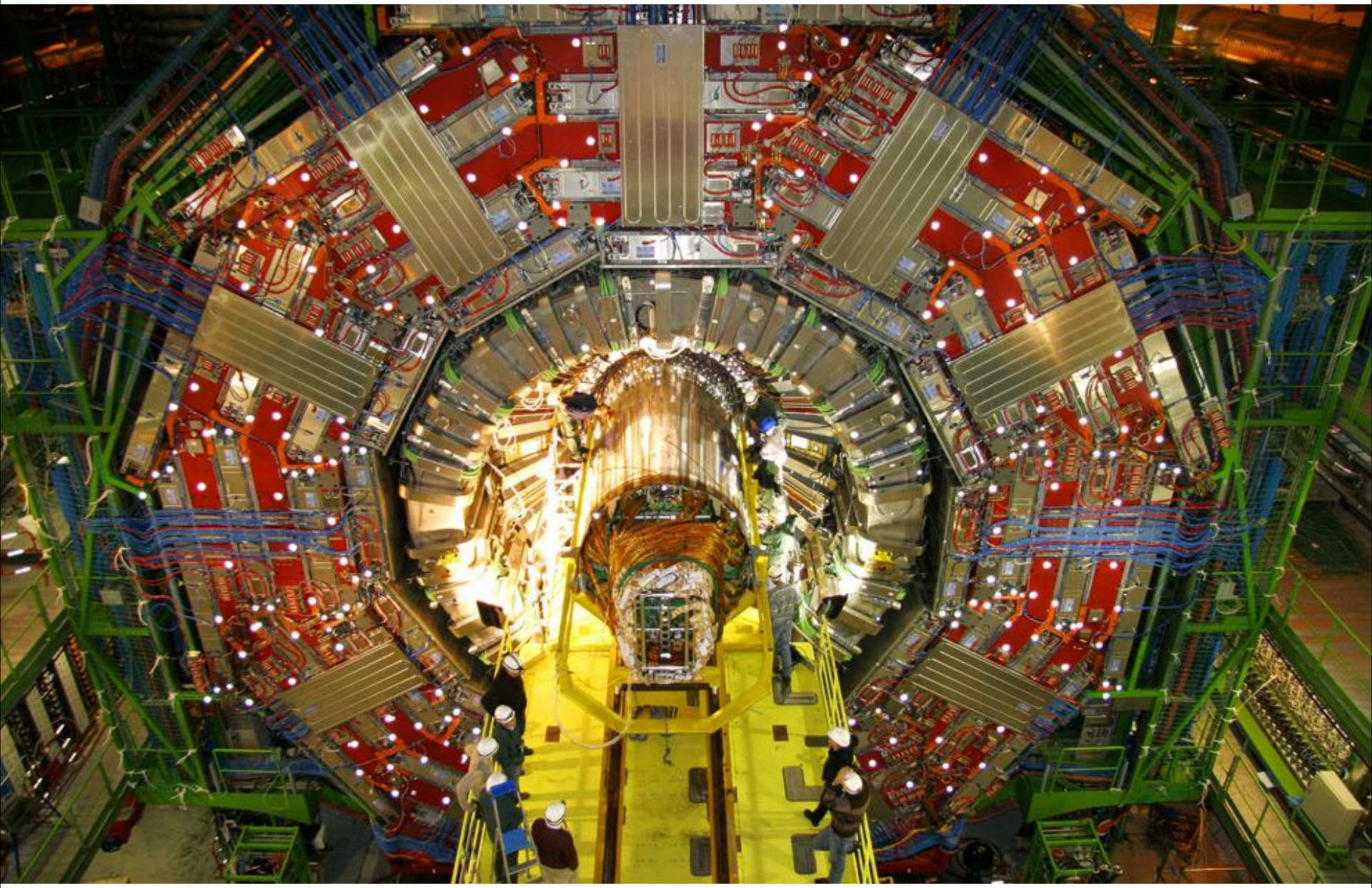
? TBs of data every day

100s of millions of GPS enabled devices sold annually

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

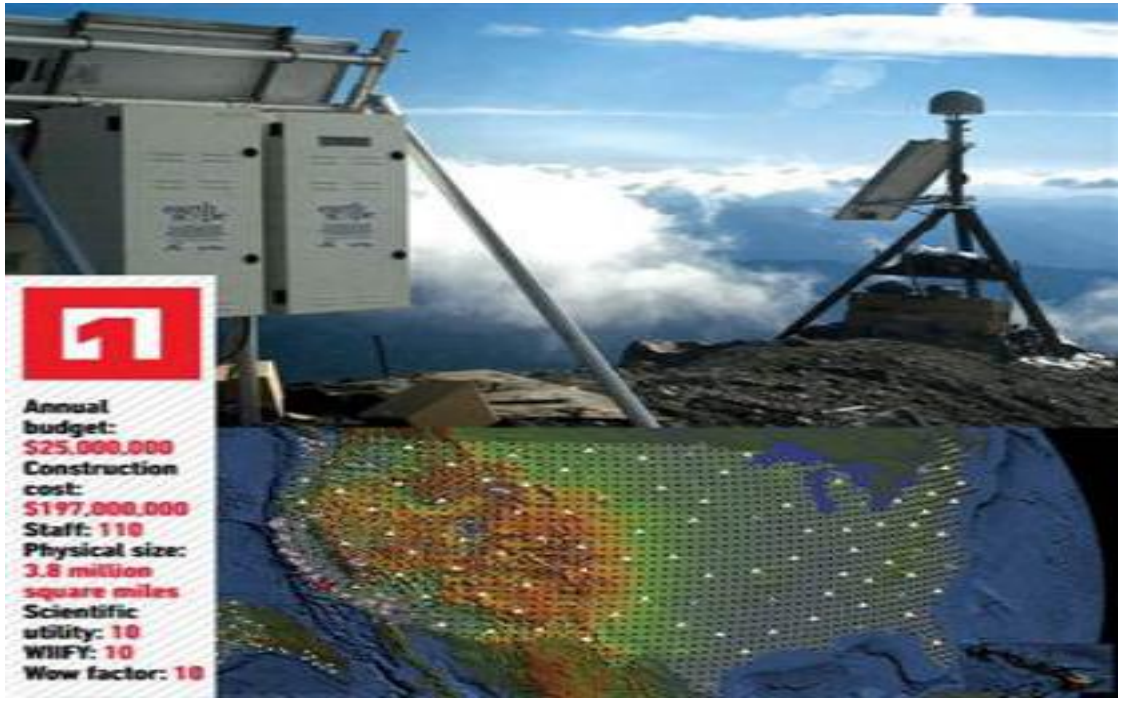
2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014



CERN’s Large Hydron Collider (LHC) generates 15 PB a year

The Earthscope

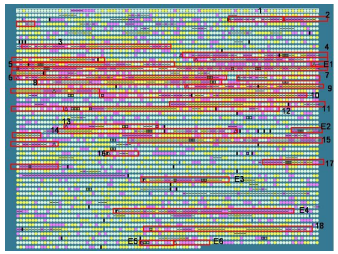


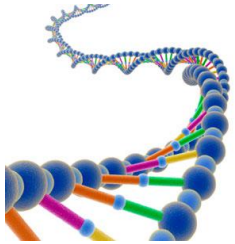
- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.

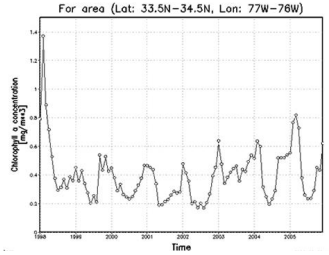
(http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)

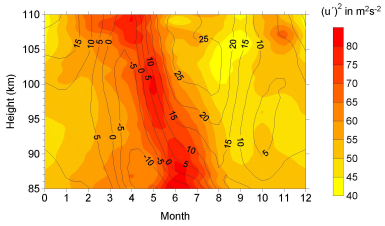
Variety (Complexity)

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



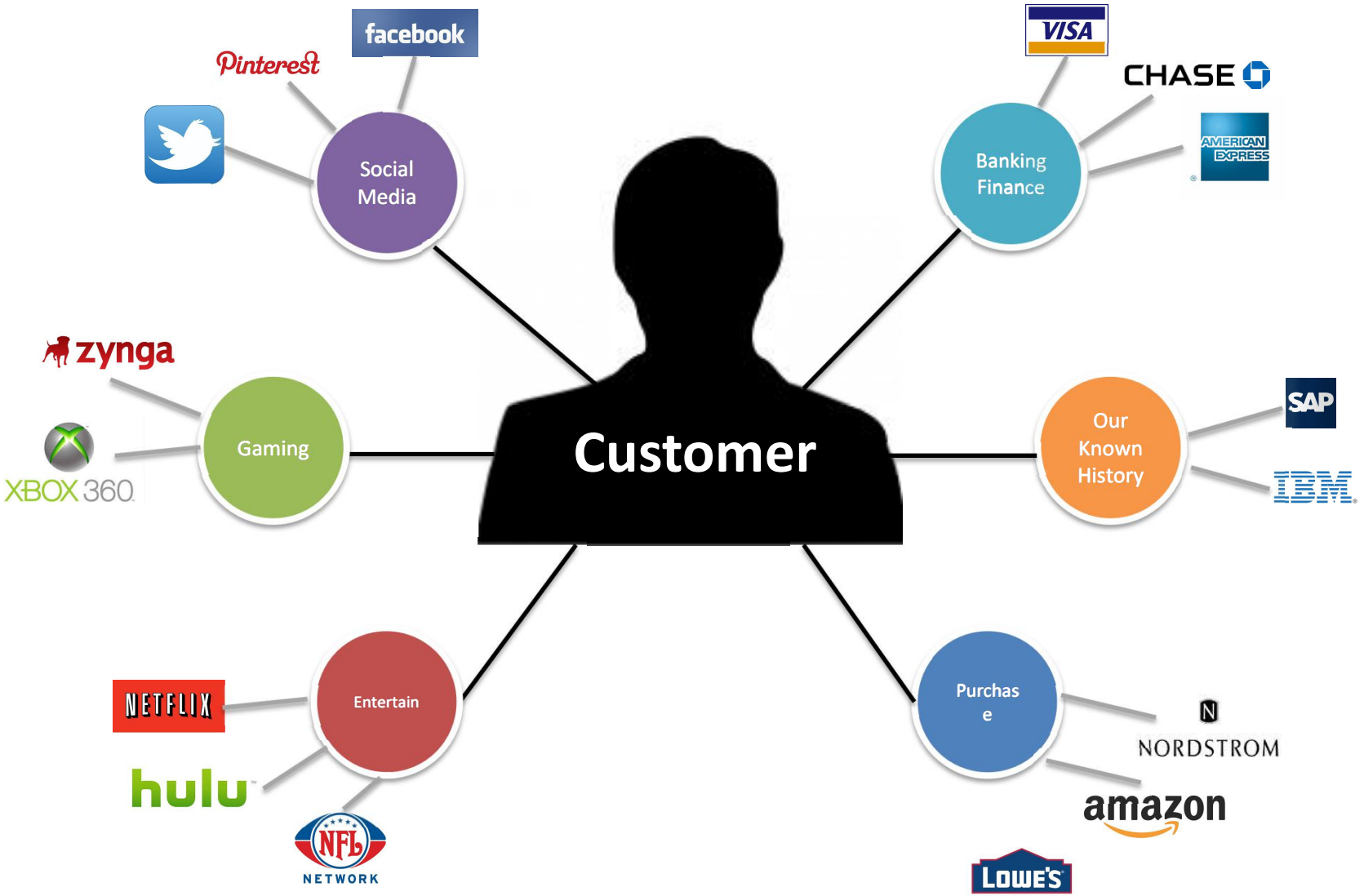






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

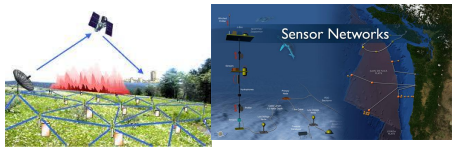
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

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

Scientific instruments
(collecting all sorts of data)

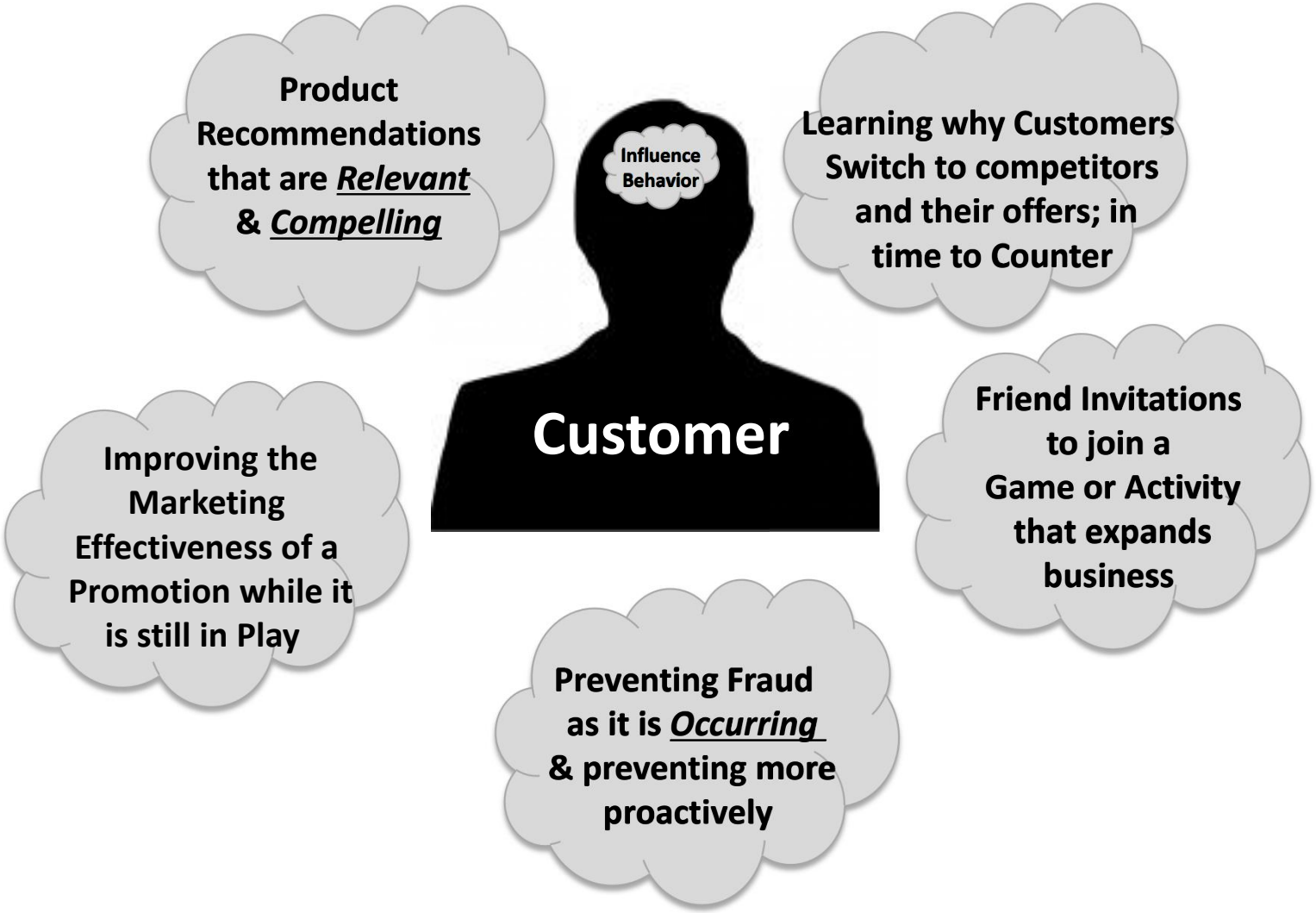
Mobile devices
(tracking all objects all the time)



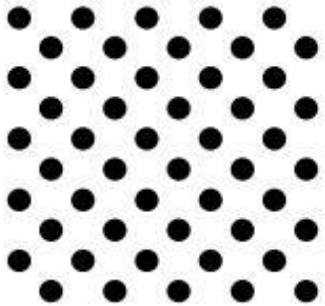
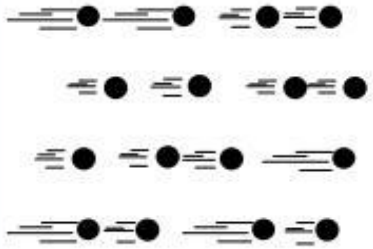
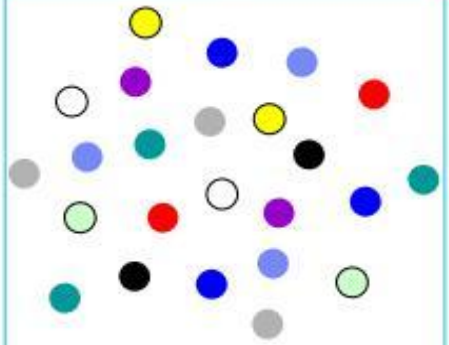
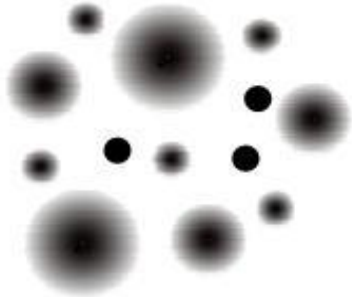
Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Real-Time Analytics/Decision Requirement



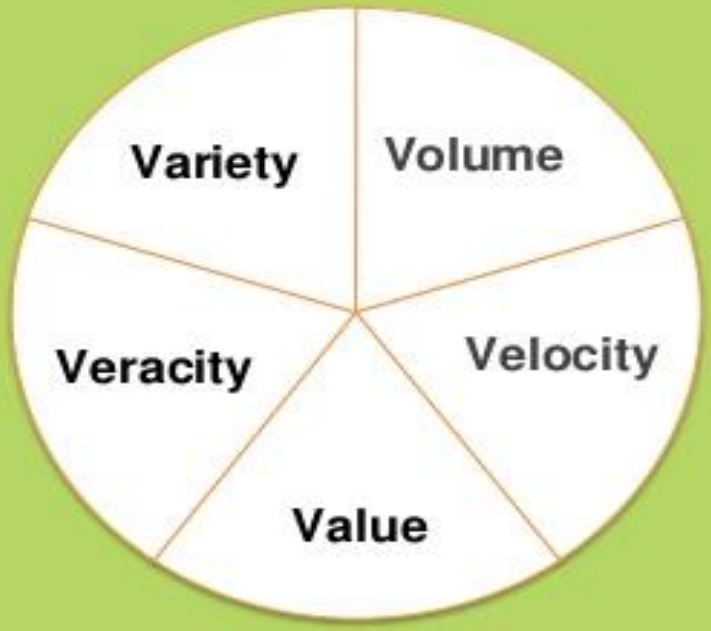
Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Some Make it 5 V's

To get a better understanding of what Big Data is, it is often described using

5 Vs:

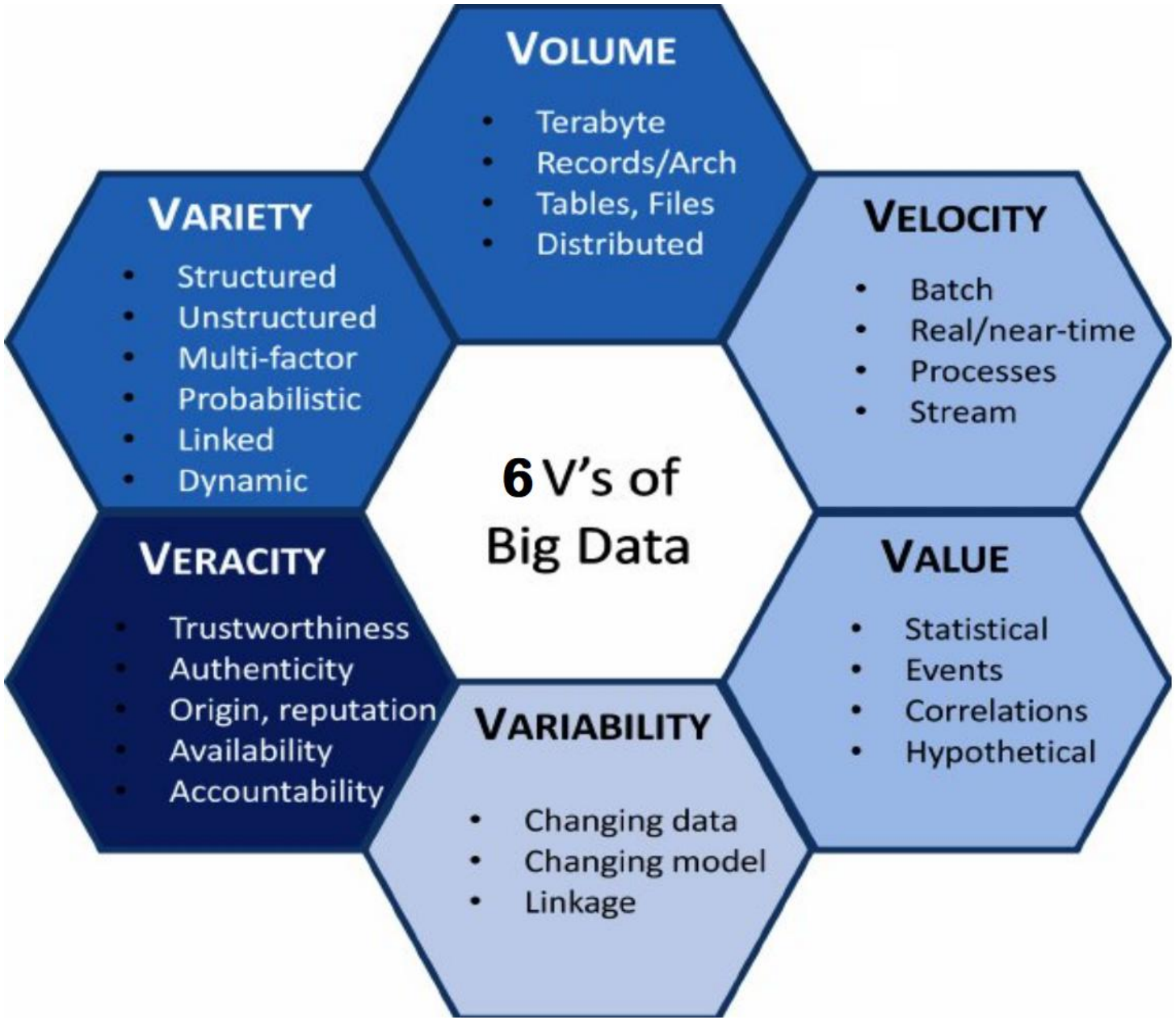


BROUGHT TO YOU BY THE BESTSELLING AUTHOR OF...

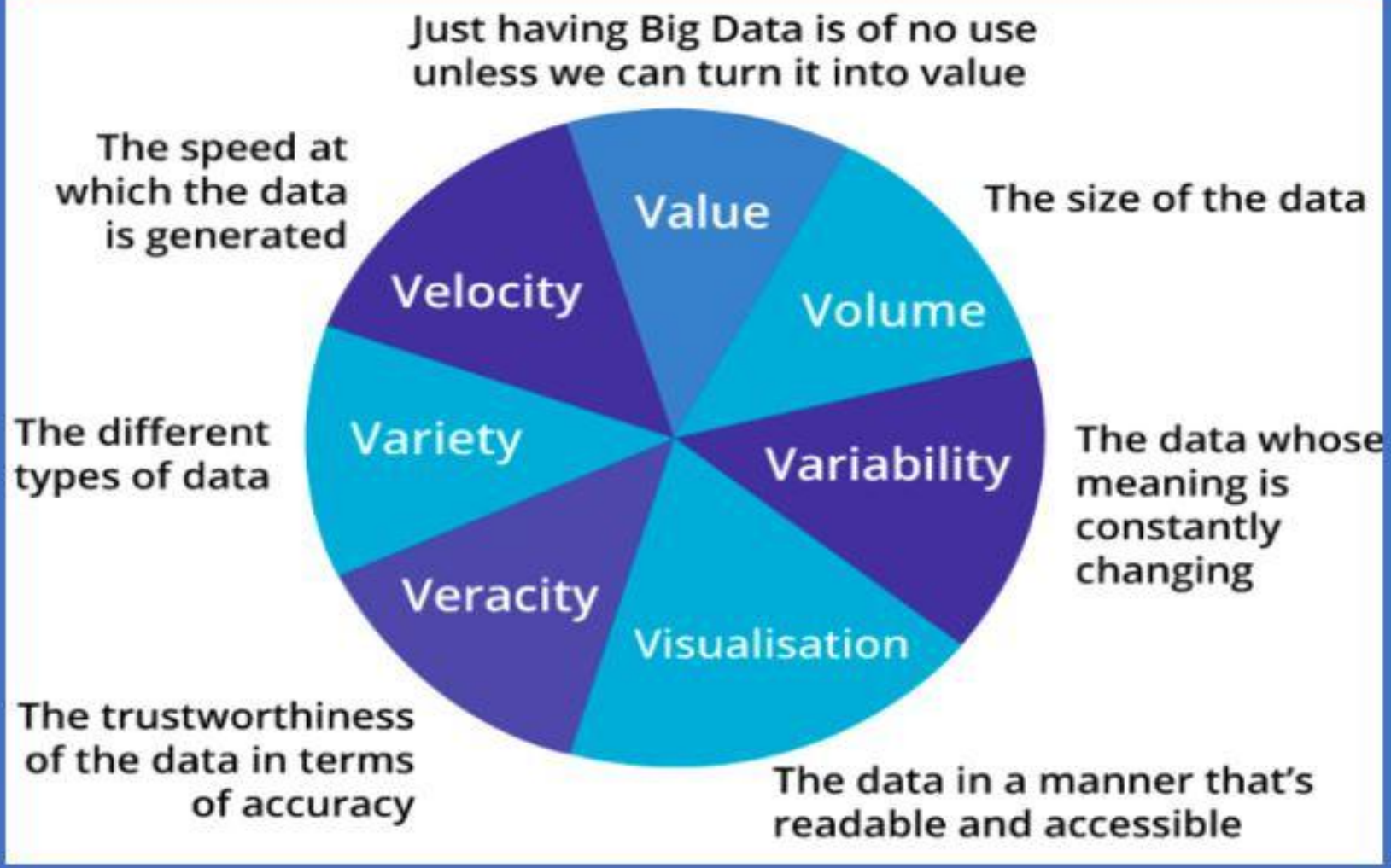
Value

- Value is defined as the usefulness of data for an enterprise.
- The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.
- Value is also dependent on how long data processing takes because analytics results have a shelf-life; for example, a 20 minute delayed stock quote has little to no value for making a trade compared to a quote that is 20 milliseconds old.
- Data that has high veracity and can be analyzed quickly has more value to business.

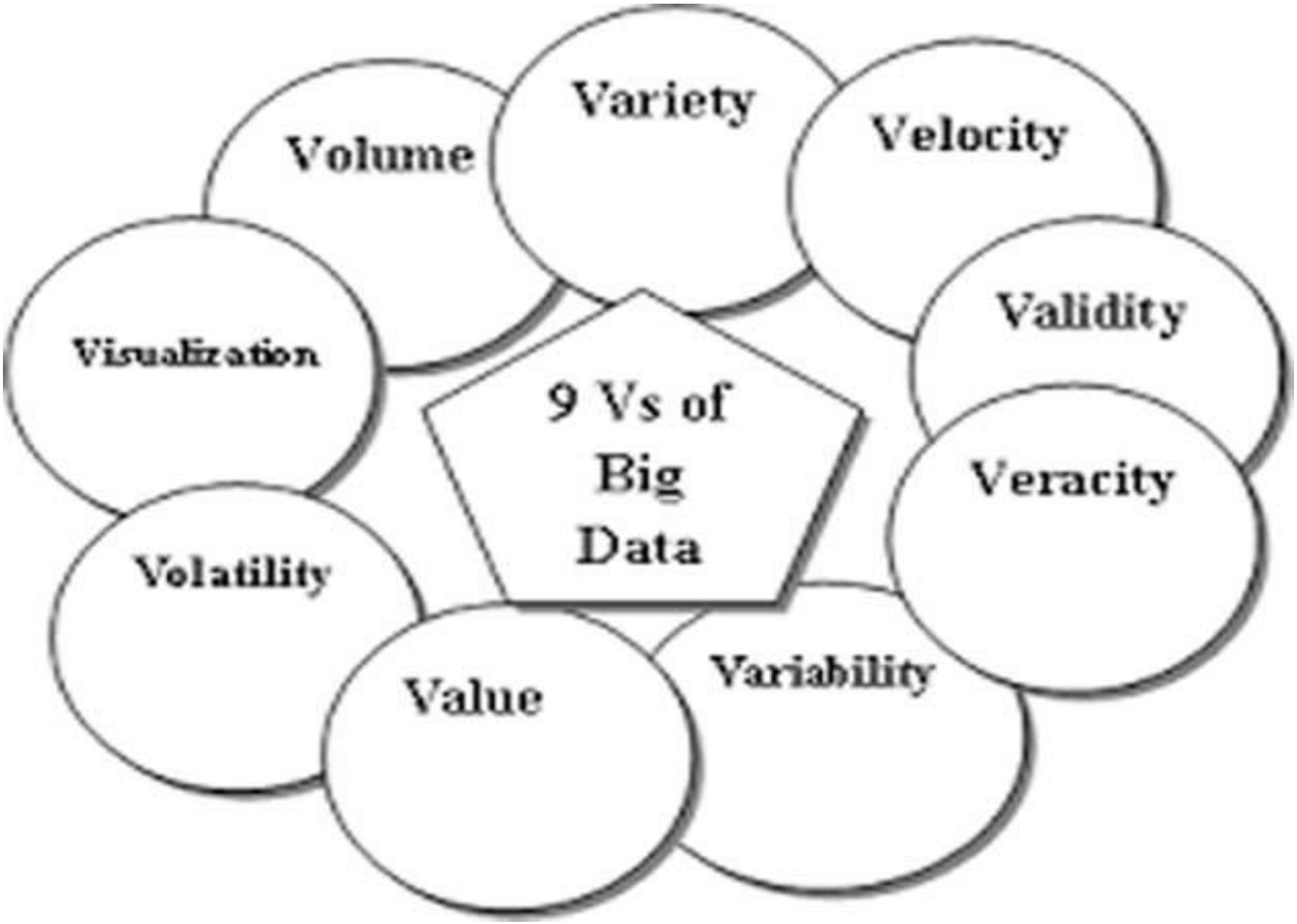
The 6 V's Big Data traits



The 7 Vs OF BIG DATA



The 9 V's Big Data traits



10 V's Big Data

1. Volume
2. Variety
3. Velocity
4. Veracity
5. Value
6. Variability
7. Visualization
8. Voloatility
9. Validity
10. Vulnerability

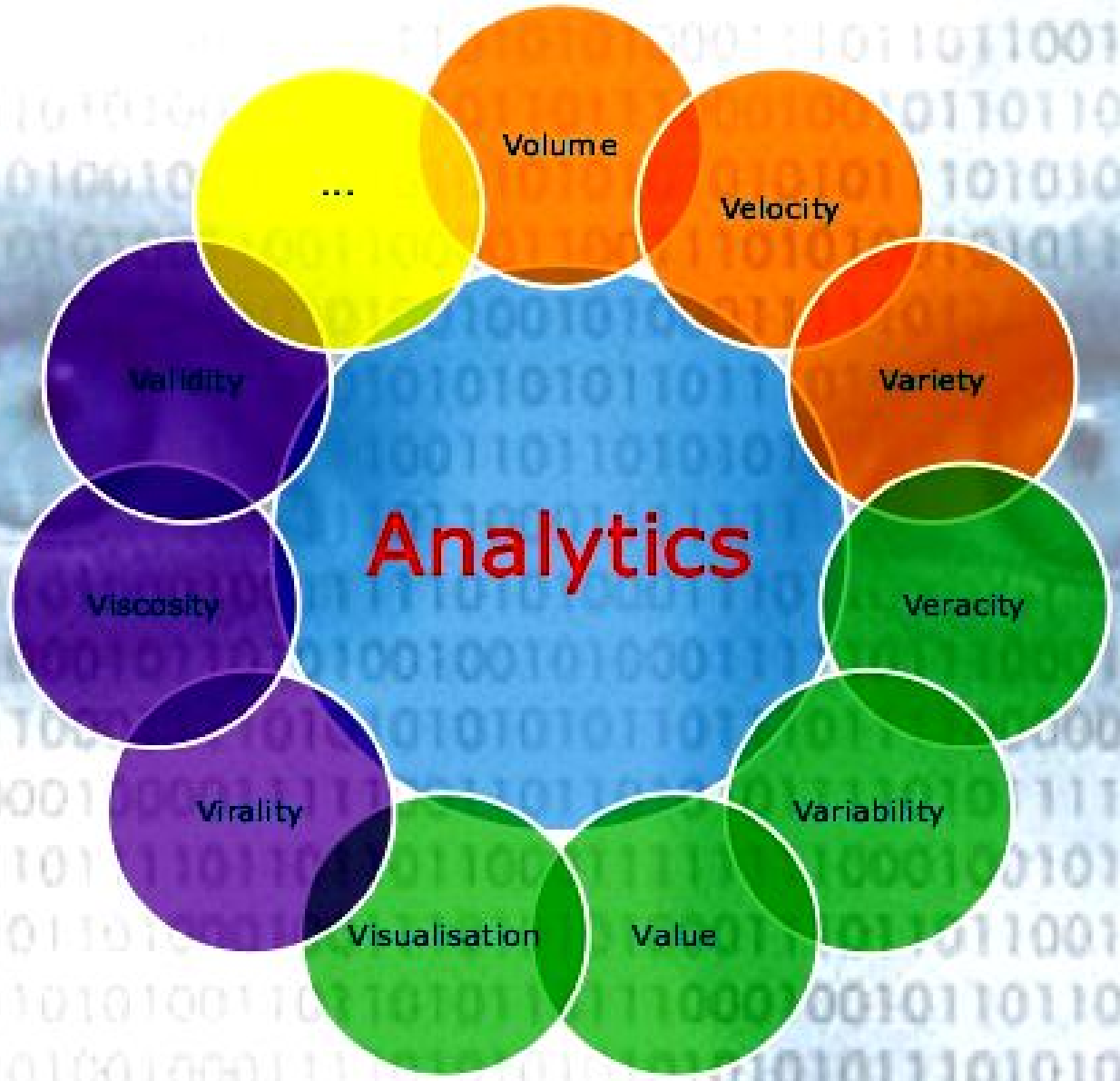
Volotility: How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?

Vulnerability: Big data brings new security concerns. After all, a data breach with big data is a big breach.

Big Data Analytics

- Big data is more **real-time in nature** than traditional DW applications
- Big data analytics reformed the ways to conduct business in many ways, such as it improves **decission making, business process management**, etc.
- Business analytics uses the data and different other techniques like information technology, features of statistics, quantitative methods and different models to provide results.
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps

Big Data: about V's and one A



Big Data Analytics

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

Types of Data Analytics

The main goal of big data analytics is to help organizations make smarter decisions for better business outcomes.

With data in hand, you can begin doing analytics.

- ***But where do you begin?***
- ***And which type of analytics is most appropriate for your big data environment?***

Looking at all the analytic options can be a daunting task. However, luckily these analytic options can be categorized at a high level into three distinct types.

- **Descriptive Analytics,**
- **Predictive Analytics,**
- **Prescriptive Analytics**

Descriptive Analytics - *(Insight into the past)*

- Descriptive Analytics, which use data aggregation and data mining to provide insight into the **past** and answer:
  - ***"What has happened in the business?"***
- Descriptive analysis or statistics does exactly what the name implies they "Describe", or summarize raw data and make it something that is interpretable by humans.
- The past refers to any point of time that an event has occurred, whether it is one minute ago, or one year ago.
- Descriptive analytics are useful because they allow us to learn from past behaviors, and understand how they might influence future outcomes.

Descriptive Analytics (cont..)

- The main objective of descriptive analytics is to find out the reasons behind precious success or failure in the past.
- The vast majority of the statistics we use fall into this category.
- Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers.

Predictive Analytics - *(Understanding the future)*

- Predictive Analytics, which use statistical models and forecasts techniques to understand the future and answer:
  - ***"What could happen?"***
- These analytics are about understanding the future.
- Predictive analytics provide estimates about the likelihood of a future outcome. It is important to remember that no statistical algorithm can "predict" the future with 100% certainty.
- Companies use these statistics to forecast what might happen in the future. This is because the foundation of predictive analytics is based on probabilities.
- These statistics try to take the data that you have, and fill in the missing data with best guesses.

Predictive Analytics (cont..)

Predictive analytics can be further categorized as –

- Predictive Modelling –What will happen next, if ?
- Root Cause Analysis-Why this actually happened?
- Data Mining- Identifying correlated data.
- Forecasting- What if the existing trends continue?
- Monte-Carlo Simulation – What could happen?
- Pattern Identification and Alerts –When should an action be invoked to correct a process.

Sentiment analysis is the most common kind of predictive analytics. The learning model takes input in the form of plain text and the output of the model is a sentiment score that helps determine whether the sentiment is positive, negative or neutral.

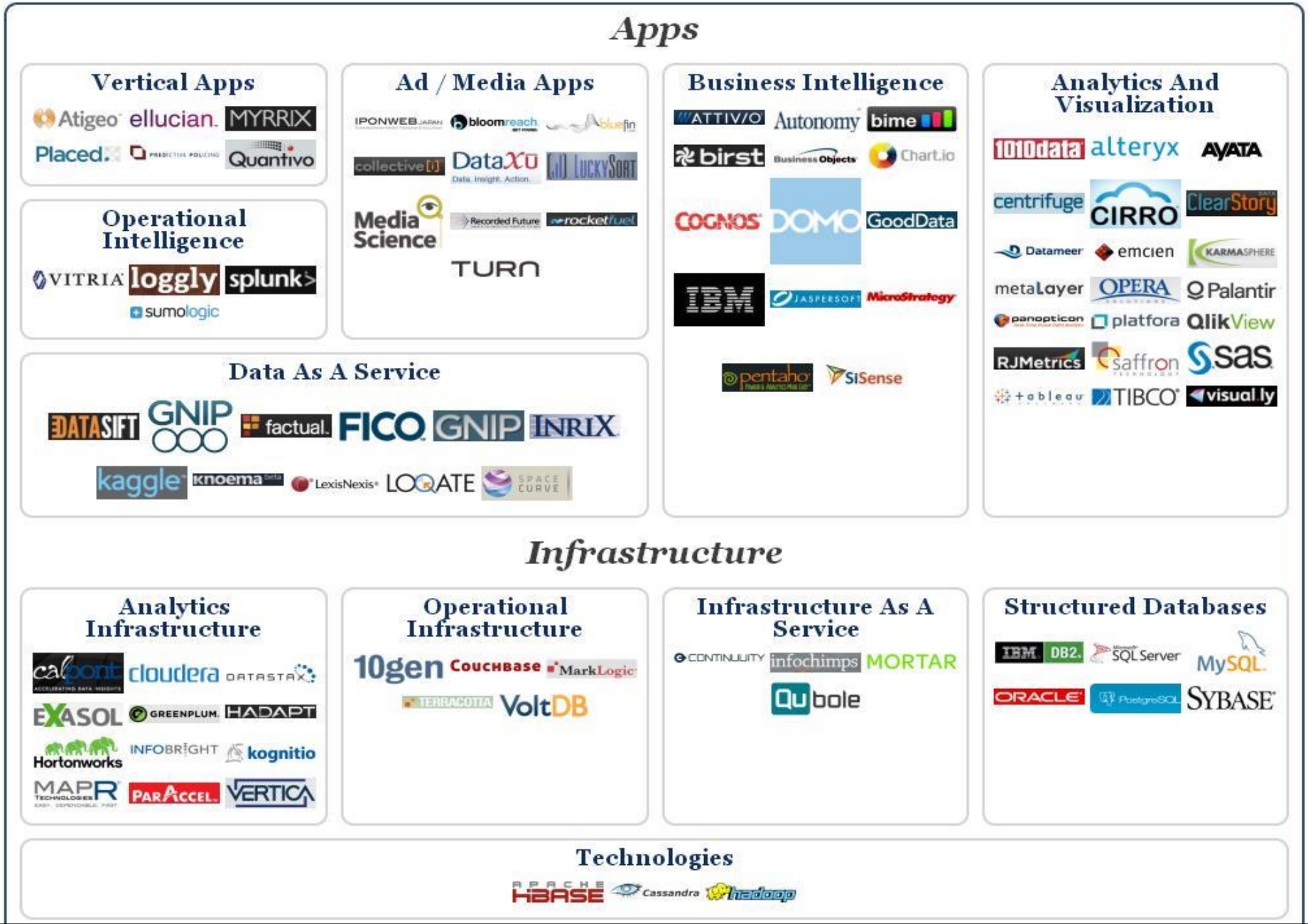
Prescriptive Analytics - *(Advise on possible outcomes)*

- Prescriptive Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer:
  - ***"What should we do?"***
- The relatively new field of prescriptive analytics allows users to "prescribe" a number of different possible actions to and guide them towards a solution. In a nut-shell, these analytics are all about providing advice.
- Prescriptive analytics is the next step of predictive analytics that adds the spice of manipulating the future.

Prescriptive Analytics (cont..)

- Prescriptive analytics is an advanced analytics concept based on,
  - Optimization that helps achieve the best outcomes.
  - Stochastic optimization that helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.
- Prescriptive analytics is a combination of data, mathematical models and various business rules. The data for prescriptive analytics can be both internal (within the organization) and external (like social media data).
- Prescriptive analytics can be used in healthcare to enhance drug development, finding the right patients for clinical trials, etc.

The Big Data Landscape



Big Data Technology