Introduction to Big Data

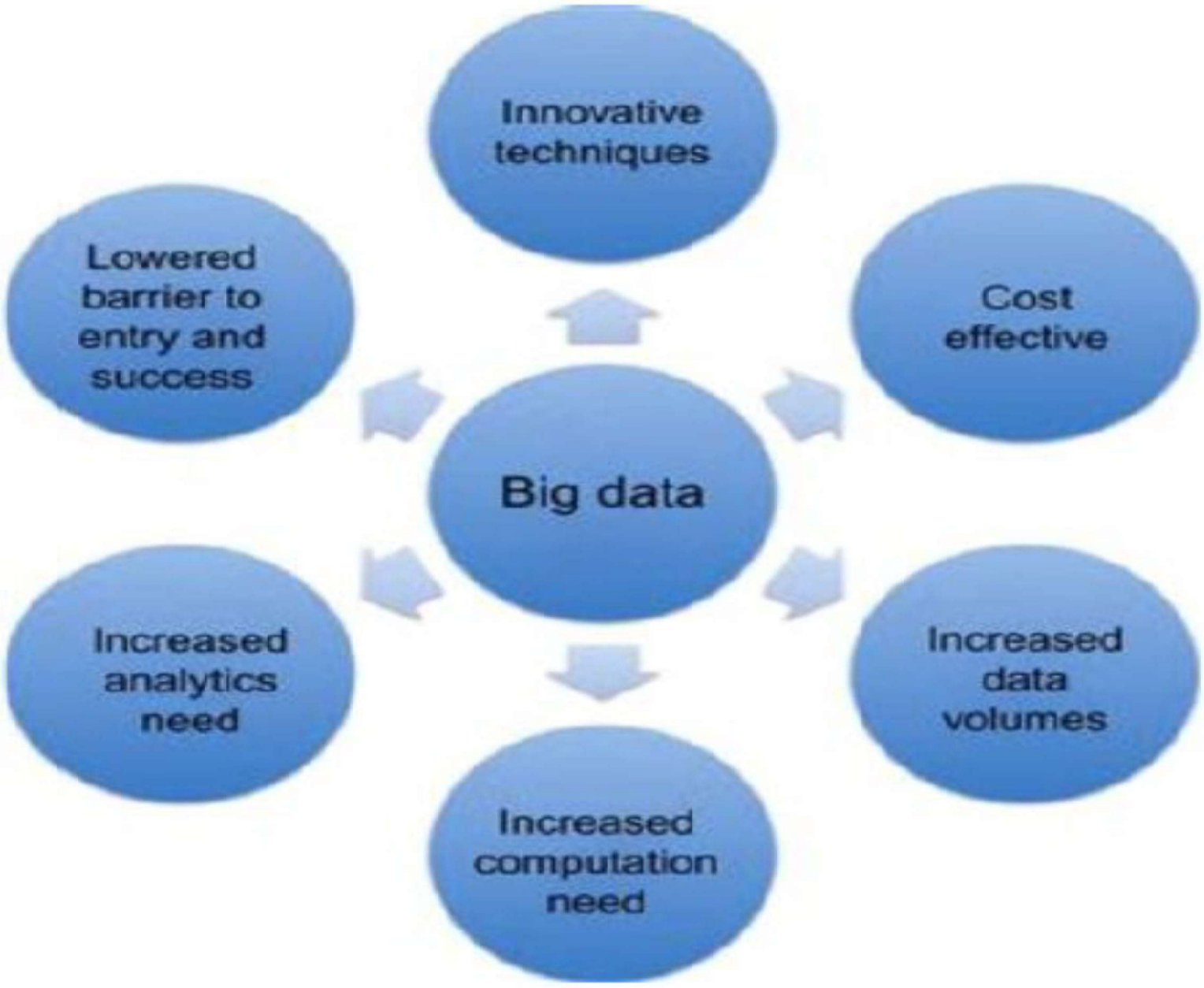
“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” – Gartner

BIG DATA is relentless. It is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor enabled instrumentation.

Definition and Characteristics of Big Data

- "Big data is high-**volume, high-velocity and high-variety information assets that** demand **cost-effective, innovative forms of information processing for enhanced insight and decision making." -- Gartner**
- which was derived from:
- "While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult.
- E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety.**
- much compile a variety of approaches to have at their disposal for dealing each."

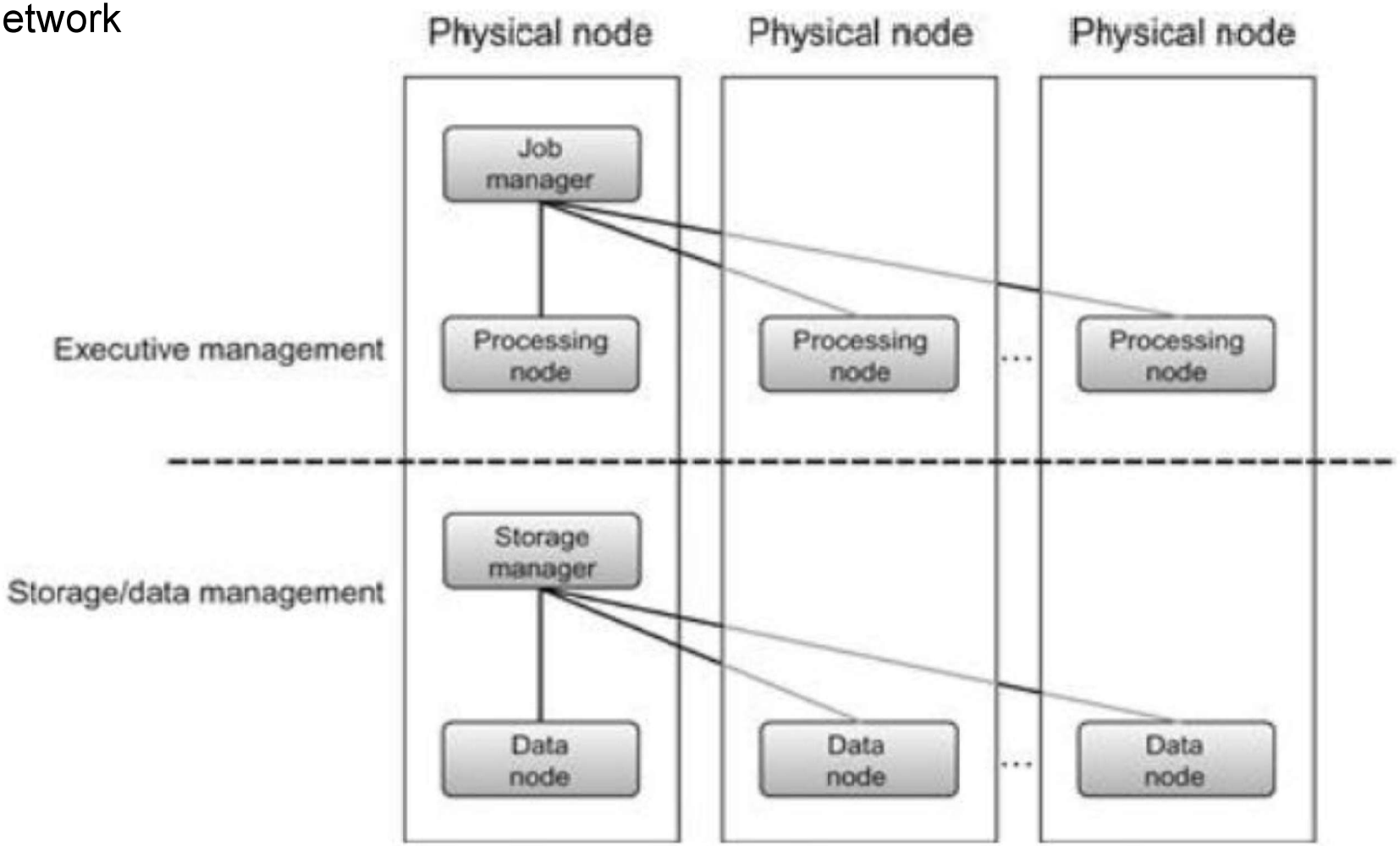
What made Big Data needed?



Key Computing Resources for Big Data

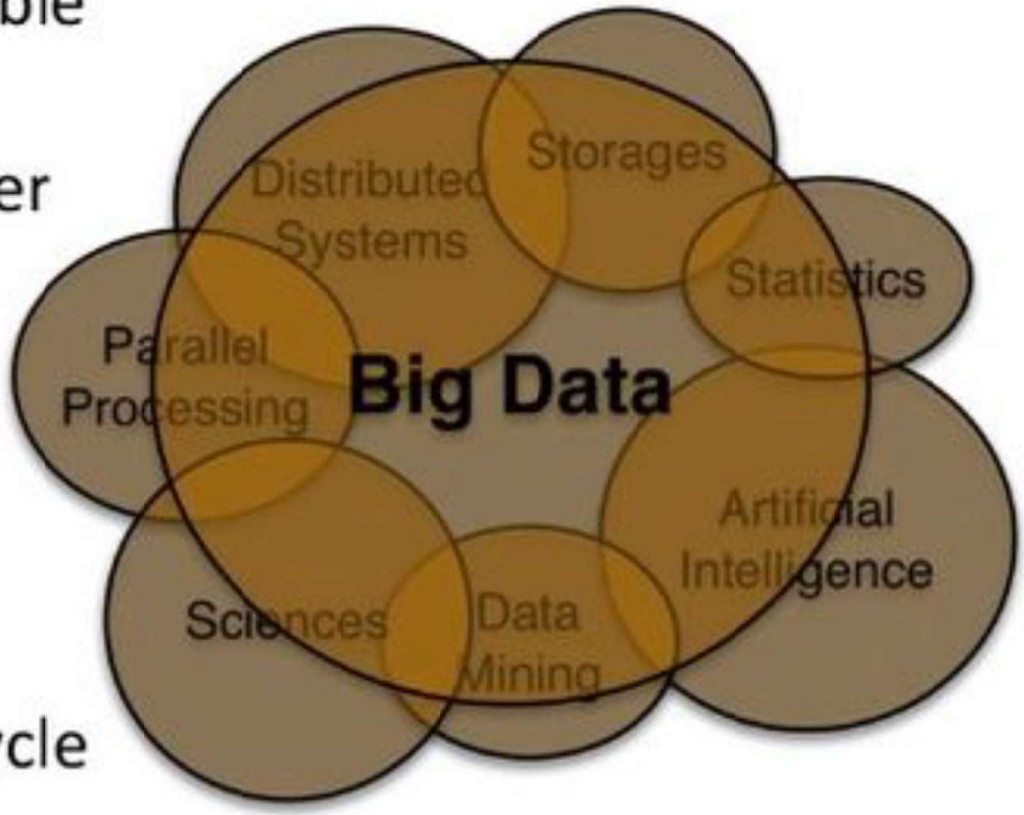
Processing capability: CPU, processor, or node.

- Memory
- Storage
- Network



What is Big data?

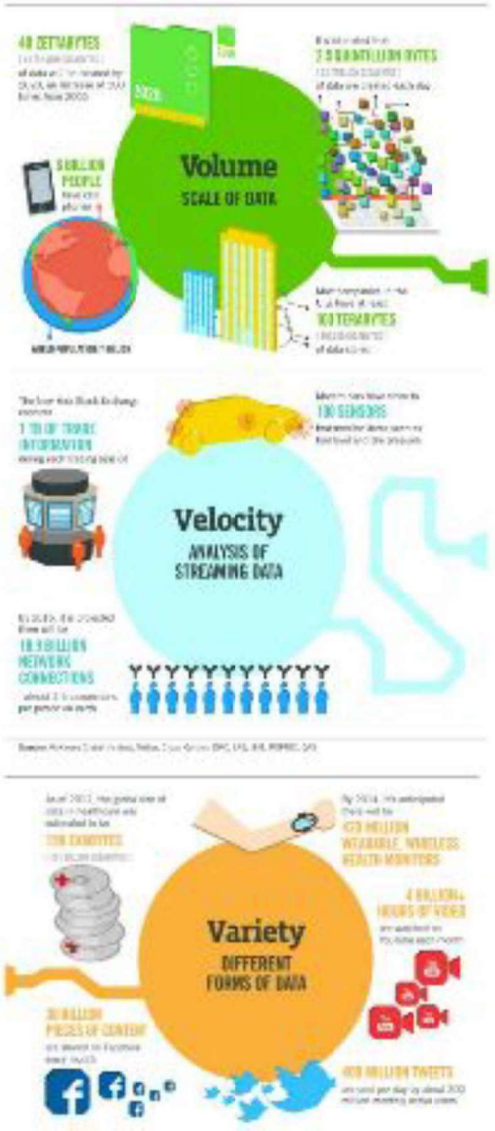
- There is lot of data available
  - E.g. Internet of things
- We have computing power
- We have technology
- Goal is same
  - To know
  - To Explain
  - To predict
- Challenge is the full lifecycle



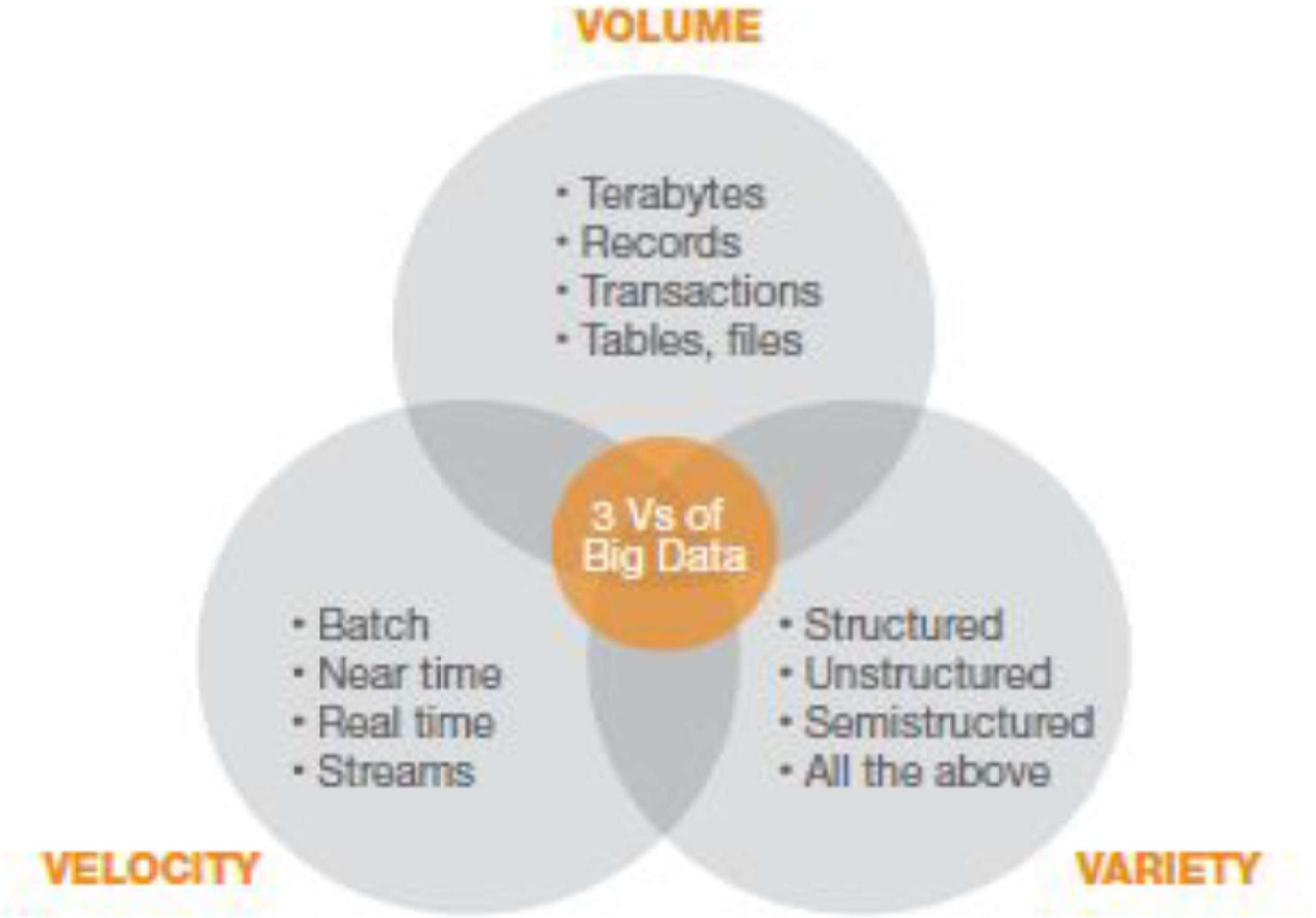
What is Big Data?

Basic attributes (Kitchin, 2014)

- High-volume
- High-velocity
- High-variety
- Exhaustivity (n=all)
- Fine resolution
- Relationality
- Flexibility



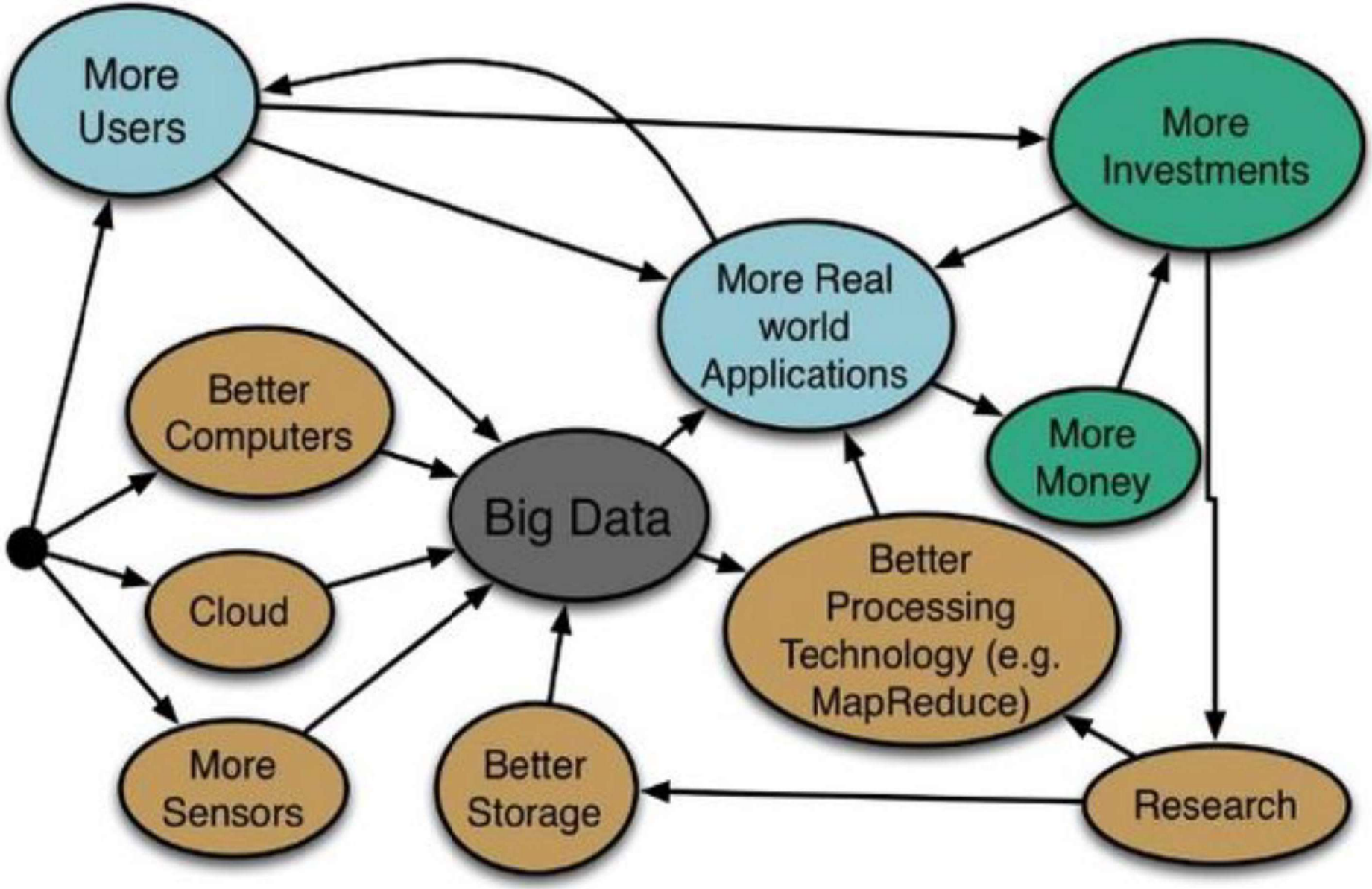
Defining Big Data Via the Three Vs



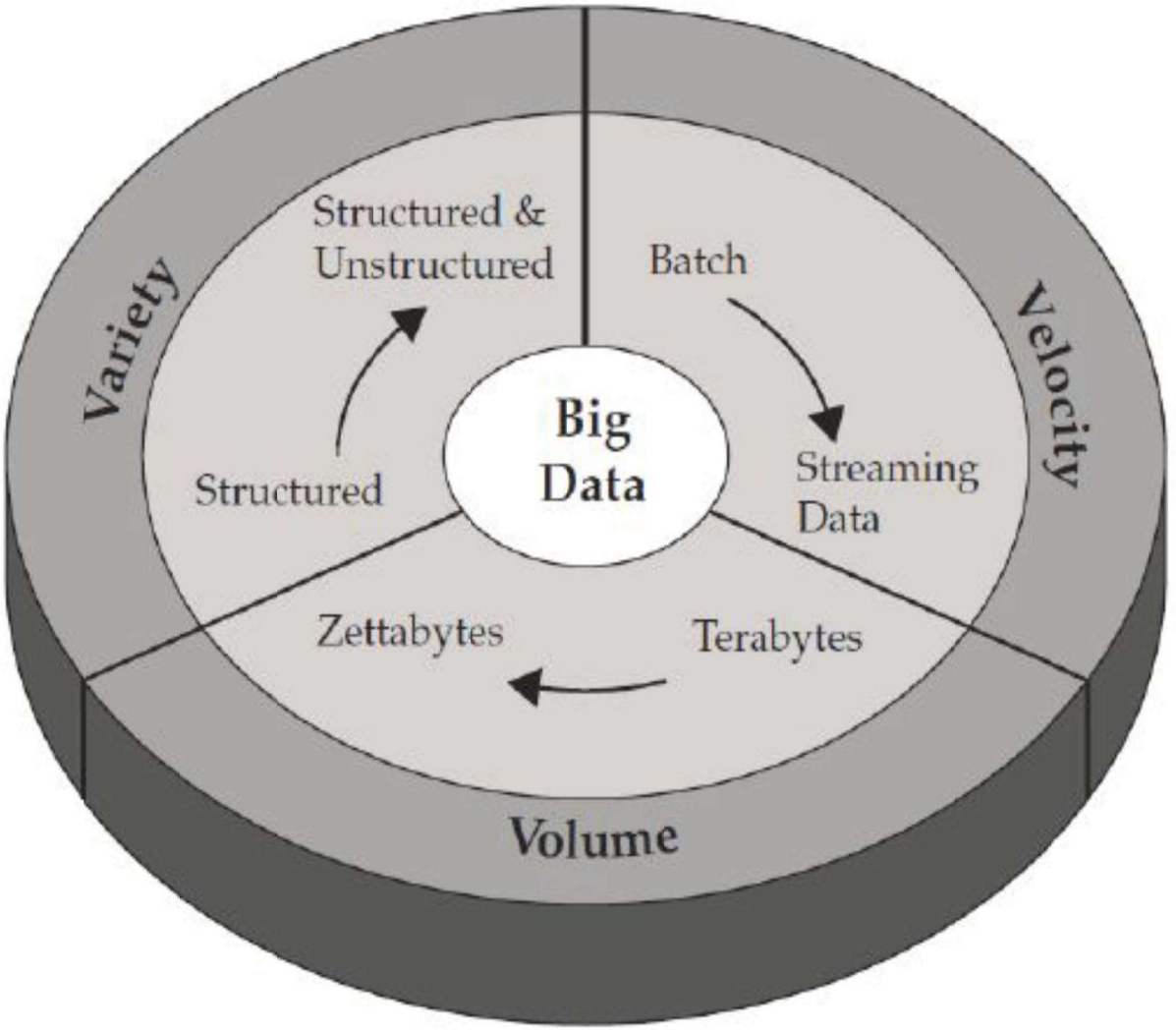
Big Data Analytics

	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

Drivers of Big Data

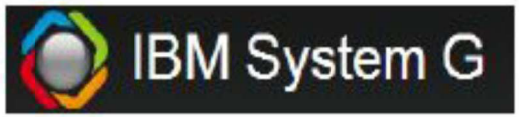


Characterization of Big-Data: volume, velocity, variety (V3)



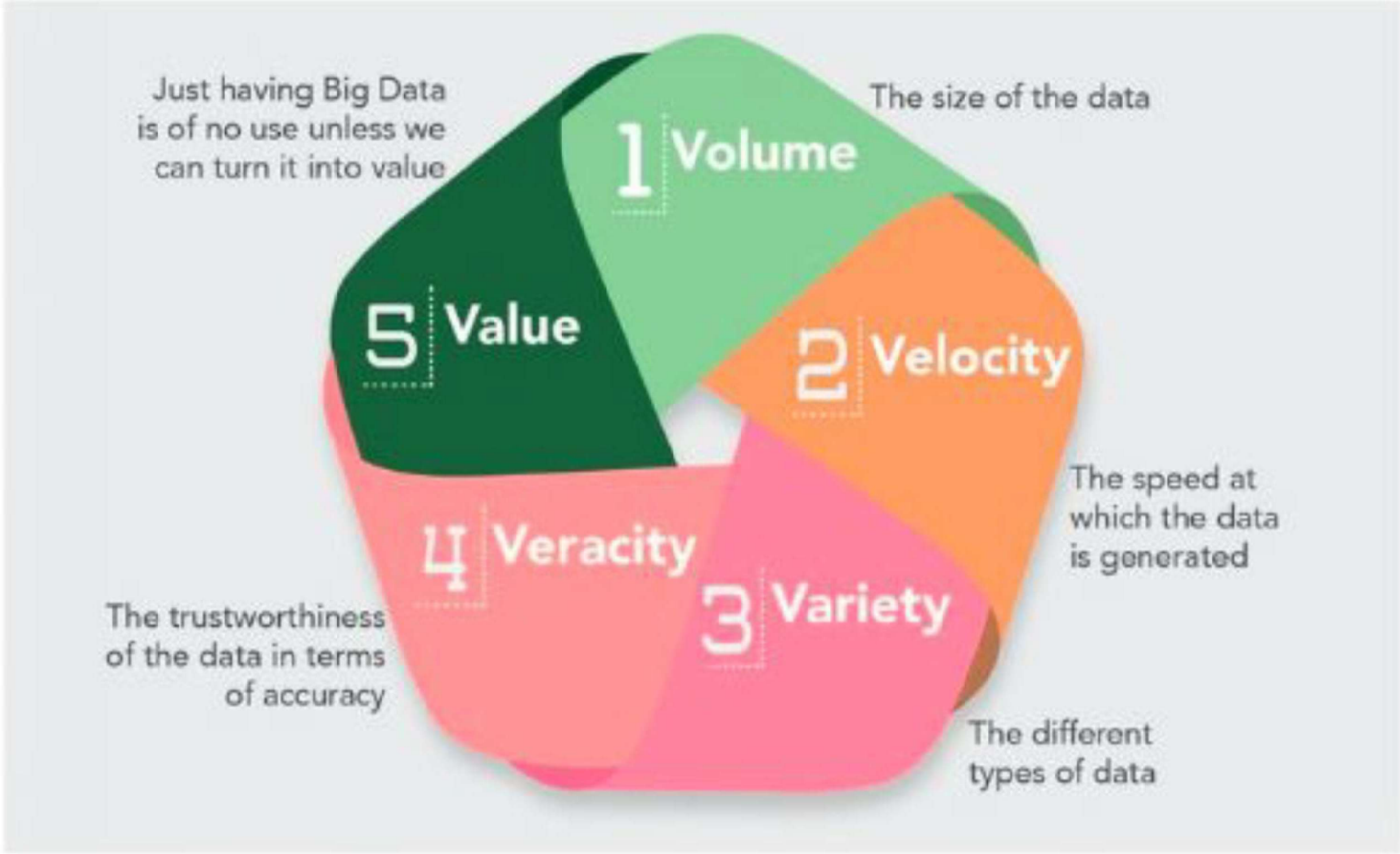
Big Data Examples -- Application Use Cases

1. System G for Expertise Location
2. System G for Recommendation
3. System G for Commerce
4. System G for Financial Analysis
5. System G for Social Media Monitoring
6. System G for Telco Customer Analysis
7. System G for Watson
8. System G for Data Exploration and Visualization
9. System G for Personalized Search
10. System G for Anomaly Detection (Espionage, Sabotage, etc.)
11. System G for Fraud Detection
12. System G for Cybersecurity
13. System G for Sensor Monitoring (Smarter another Planet)
14. System G for Cellular Network Monitoring
15. System G for Cloud Monitoring
16. System G for Code Life Cycle Management
17. System G for Traffic Navigation
18. System G for Image and Video Semantic Understanding
19. System G for Genomic Medicine
20. System G for Brain Network Analysis
21. System G for Data Curation
22. System G for Near Earth Object Analysis



Challenges of Conventional Systems

BigData Challenges & Characteristics

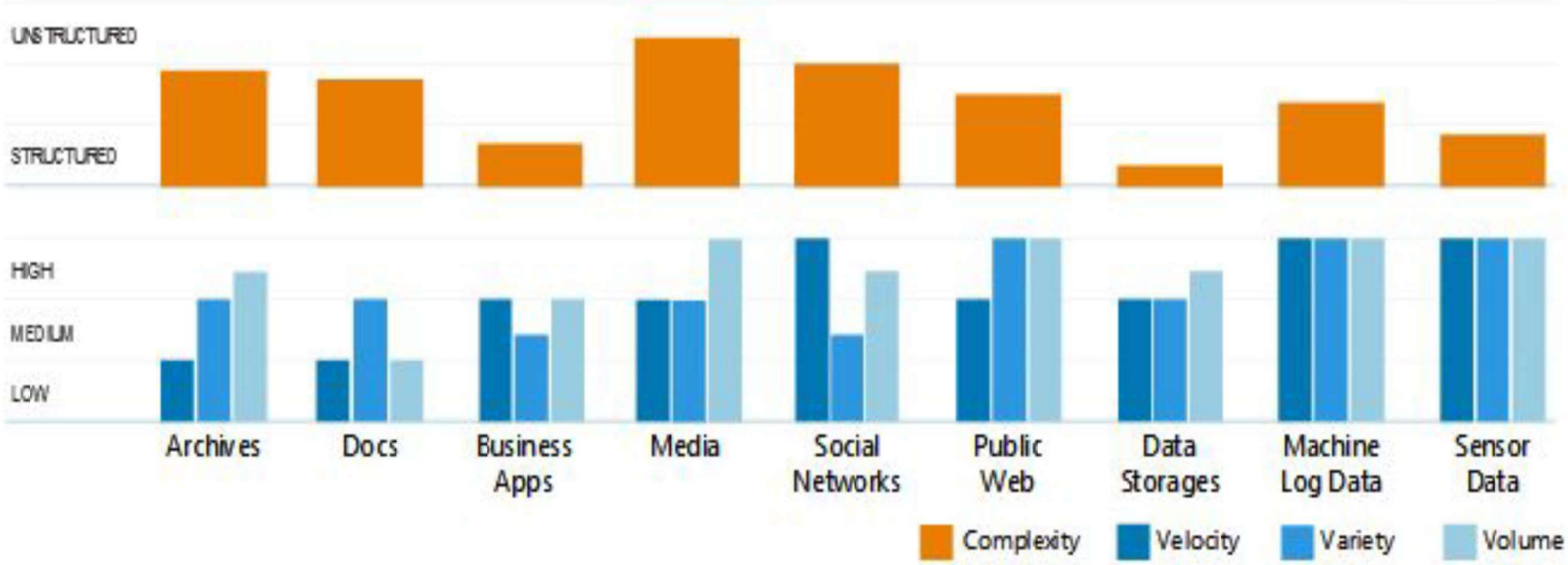


Ethical challenges

Potential consequences of Big data misuses and abuses (Bollier, 2010):

- imperil consumer freedom
- imperil civil security
- imperil personal privacy
- imperil of civil liberties

Big Data Challenges



Archives
Scanned documents, statements, medical records, e-mails etc..

Docs
XLS, PDF, CSV, HTML, JSON etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Media
Images, video, audio etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

Public Web
Wikipedia, news, weather, public finance etc

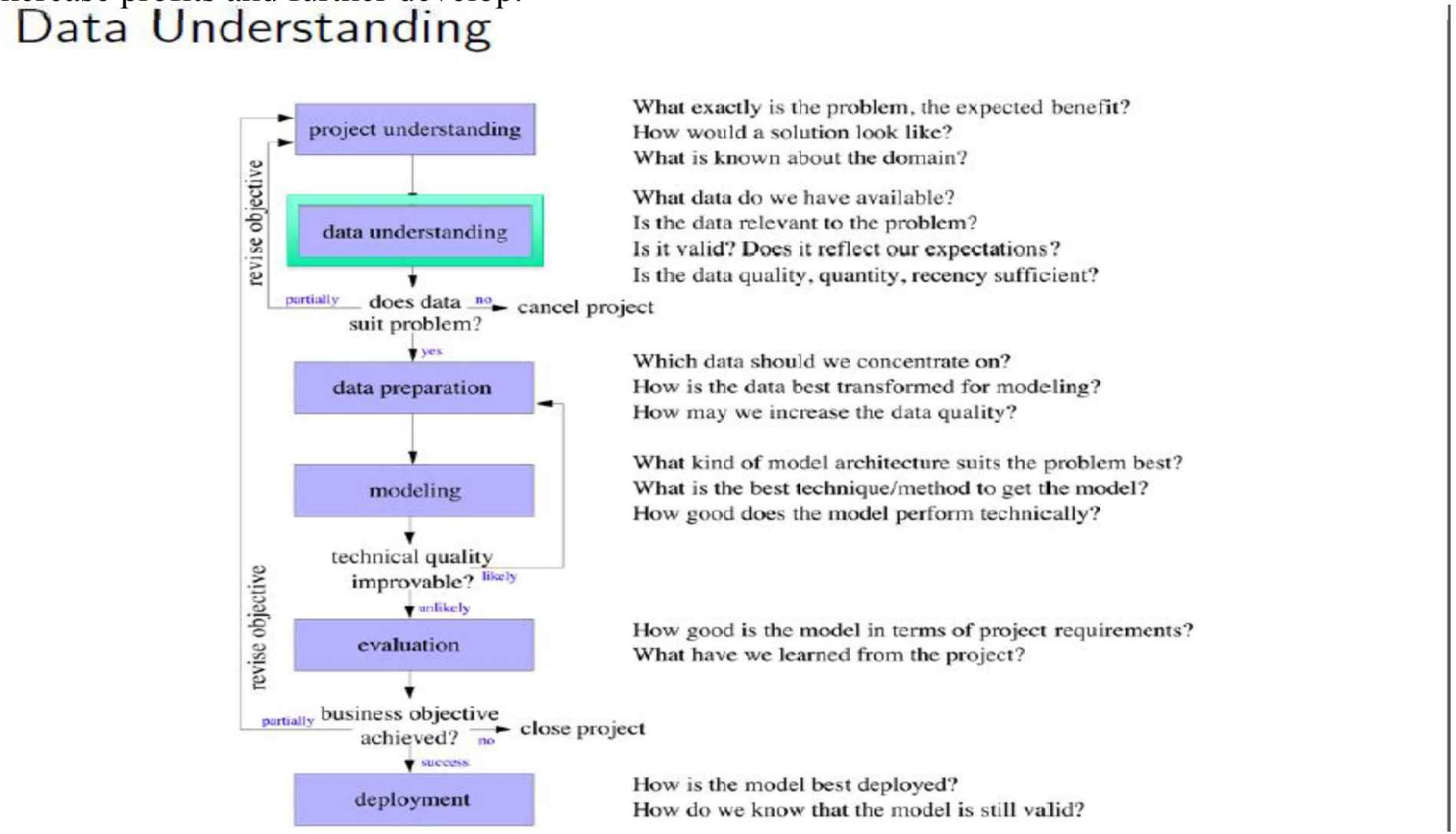
Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

Intelligent data analysis

Data analysis is the most powerful tool to bring into your business. Employing the powers of analysis can be comparable to finding gold in your reports, which allows your business to increase profits and further develop.



Types of attributes

categorical (nominal): finite domain

The values of a categorical attribute are often called classes or categories.

Examples: {female,male}, {ordered,sent,received}

ordinal: finite domain with a linear ordering on the domain.

Examples: {B.Sc.,M.Sc.,Ph.D.}

numerical: values are numbers.

discrete: categorical attribute or numerical attribute whose domain is a subset of the integer number.

continuous: numerical attribute with values in the real numbers or in an interval

Data quality

Low data quality makes it impossible to trust analysis results: "Garbage in, garbage out"

Accuracy: Closeness between the value in the data and the true value.

- Reason of low accuracy of numerical attributes: noisy measurements, limited precision, wrong measurements, transposition of digits (when entered manually).
- Reason of low accuracy of categorical attributes: erroneous entries, typos.

Data quality

Syntactic accuracy : Entry is not in the domain.
Examples: fmale in gender, text in numerical attributes, ...
Can be checked quite easy.

Semantic accuracy : Entry is in the domain but not correct.
Example: John Smith is female
Needs more information to be checked (e.g. "business rules").

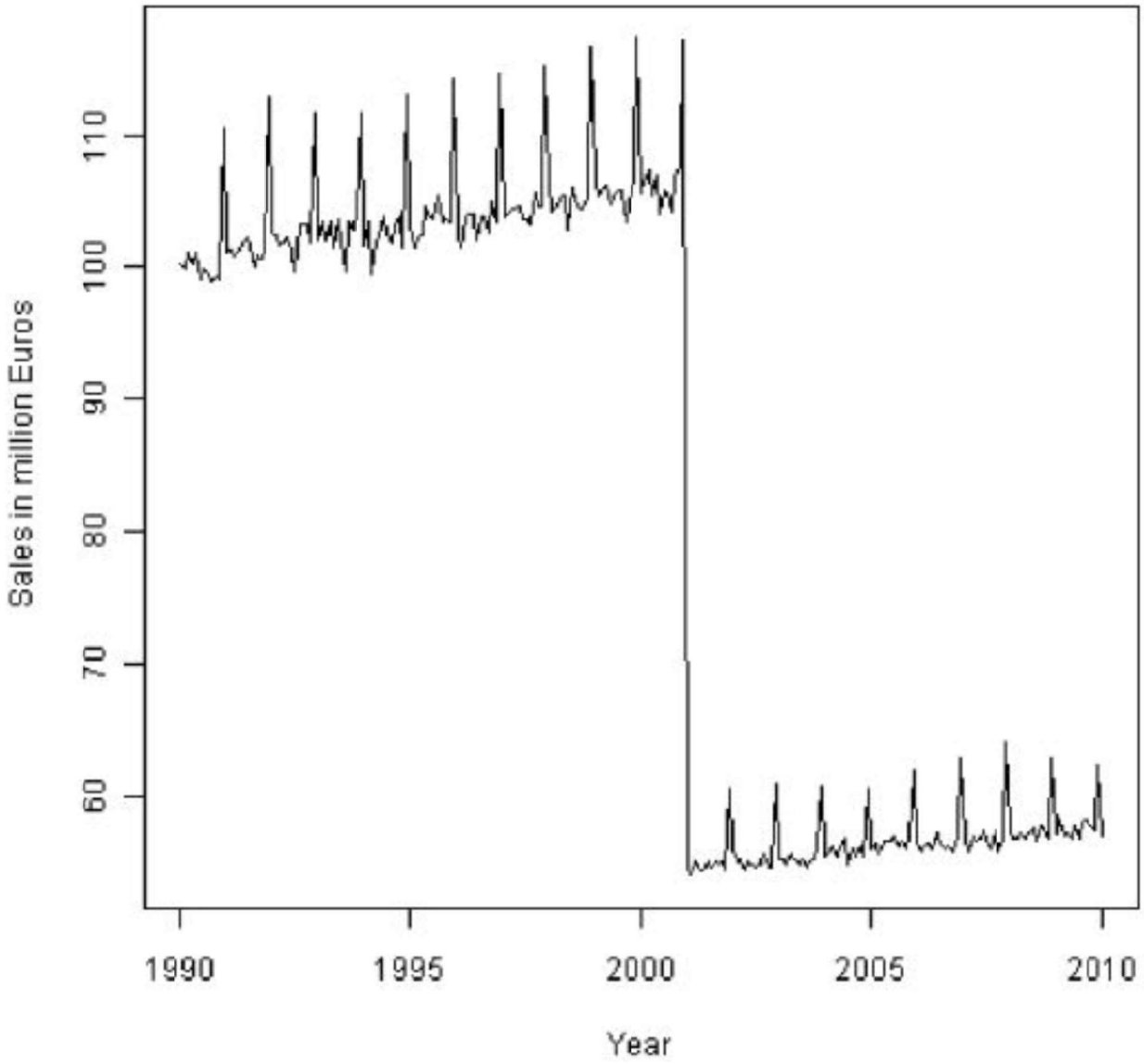
Completeness : is violated if an entry is not correct although it belongs to the domain of the attribute.
Example: Complete records are missing, the data is biased (A bank has rejected customers with low income.)

Unbalanced data: The data set might be biased extremely to one type of records.
Example: Defective goods are a very small fraction of all.

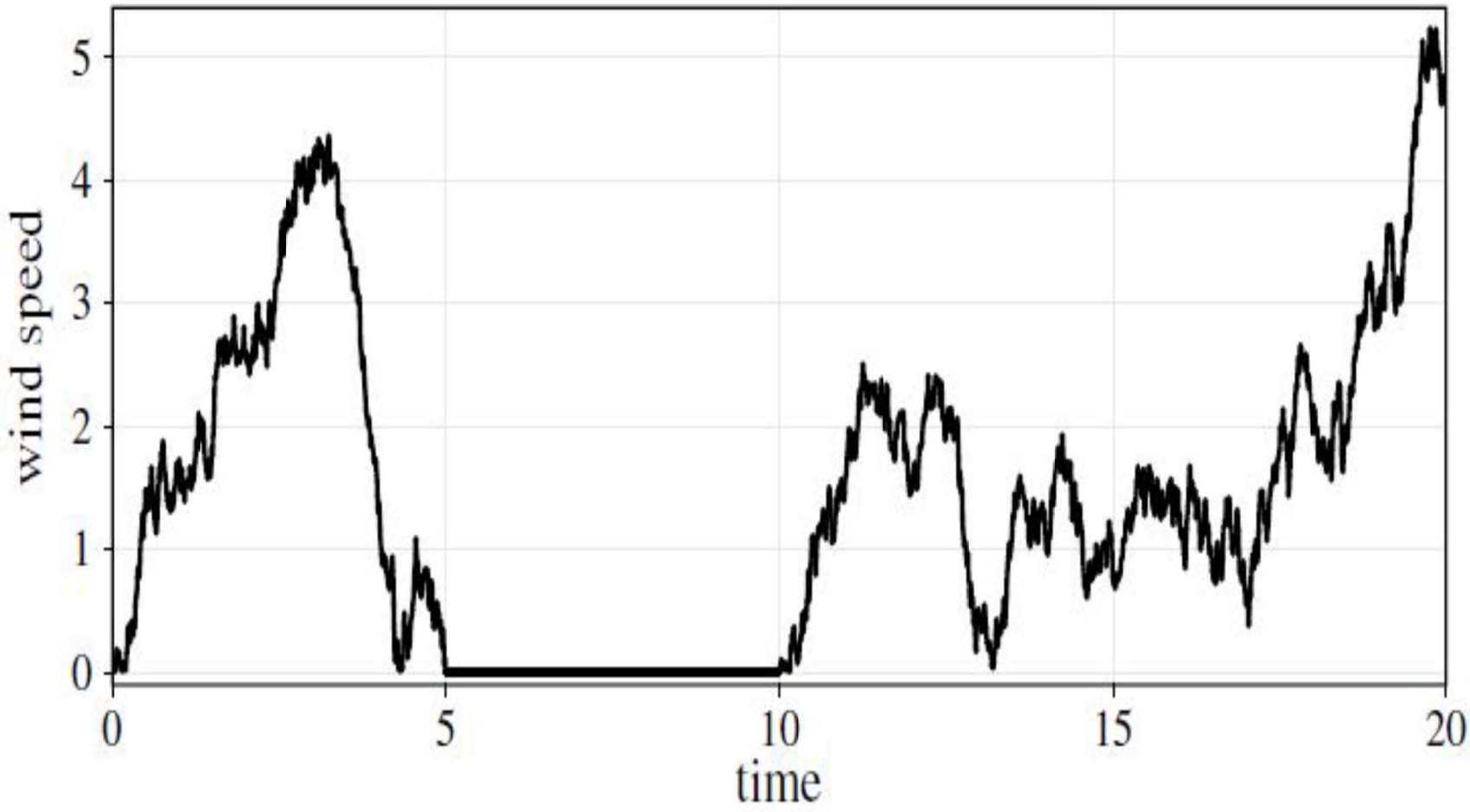
Timeliness: Is the available data up to date?

Data visualisation

Tukey: There is no excuse for failing to plot and look.



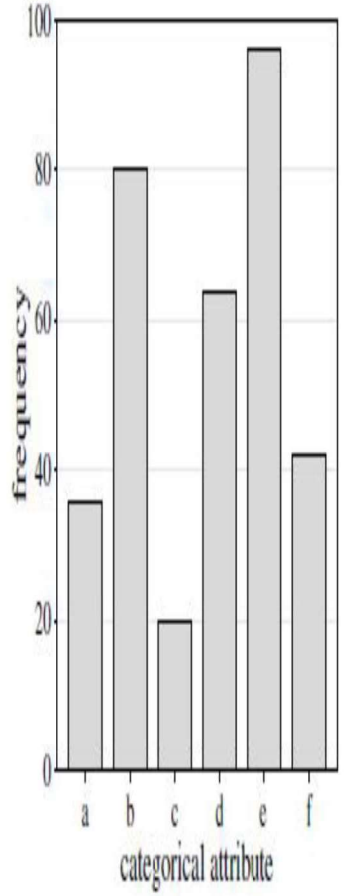
Hidden missing values



The zero values might come from a broken or blocked sensor and might be consider as missing values.

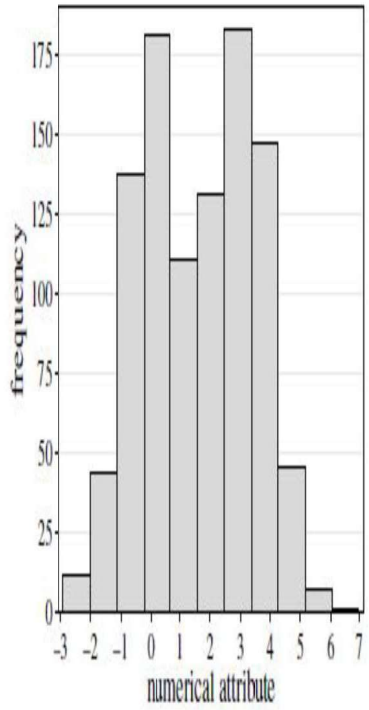
Bar charts

A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

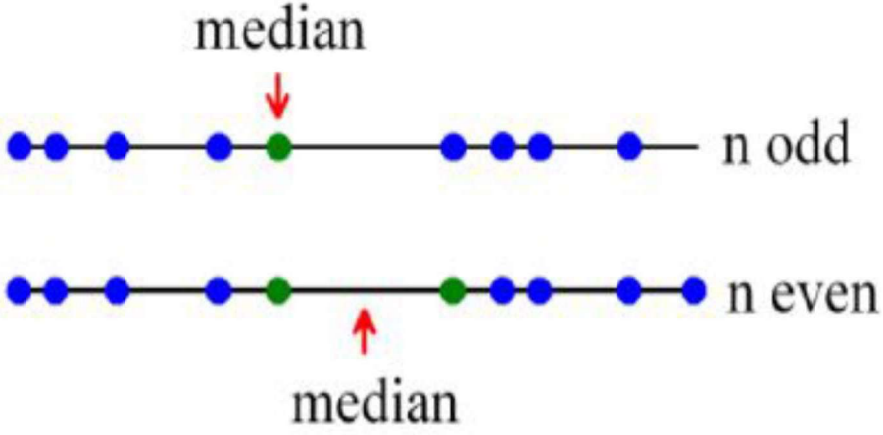


Histograms

A histogram shows the frequency distribution for a numerical attribute. The range of the numerical attribute is discretized into a fixed number of intervals (called bins), usually of equal length. For each interval the (absolute) frequency of values falling into it is indicated by the height of a bar.



Reminder: Median, quantiles, quartiles, interquartile range



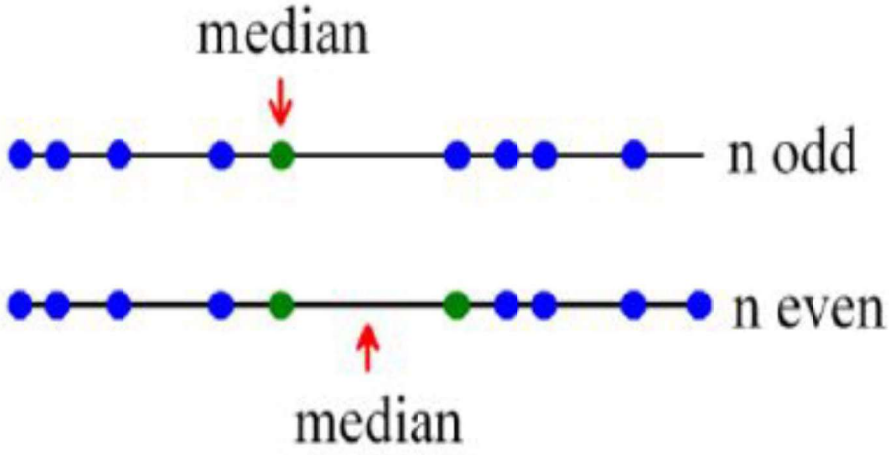
Median: The value in the middle (for the values given in increasing order).

q%-quantile ($0 < q < 100$): The value for which q% of the values are smaller and 100-q% are larger.
The median is the 50%-quantile.

Quartiles: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile).

Interquartile range (IQR): 3rd quantile - 1st quantile.

Reminder: Median, quantiles, quartiles, interquartile range



Median: The value in the middle (for the values given in increasing order).

q%-quantile ($0 < q < 100$): The value for which q% of the values are smaller and 100-q% are larger.
The median is the 50%-quantile.

Quartiles: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile).

Interquartile range (IQR): 3rd quantile - 1st quantile.

Example data set: Iris data

iris setosa

iris versicolor

iris virginica

- collected by E. Anderson in 1935
- contains measurements of four real-valued variables:
- sepal length, sepal widths, petal lengths and petal width of 150 iris flowers of types Iris Setosa, Iris Versicolor, Iris Virginica (50 each)
- The fifth attribute is the name of the flower type.

A checklist for data understanding

- Determine the quality of the data. (e.g. syntactic accuracy)
- Find outliers. (e.g. using visualization techniques)
- Detect and examine missing values. Possible hidden by default values.
- Discover new or confirm expected dependencies or correlations between attributes.
- Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
- Compare statistics with the expected behavior.
- Check the **distributions for each attribute**
  (unexpected properties like outliers, correct domains, correct medians)
- Check **correlations or dependencies** between pairs of attributes

Nature of Data: Categories of 'Big Data'

- Big data' could be found in three forms:
- Structured
- Unstructured
- Semi-structured

Structured Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

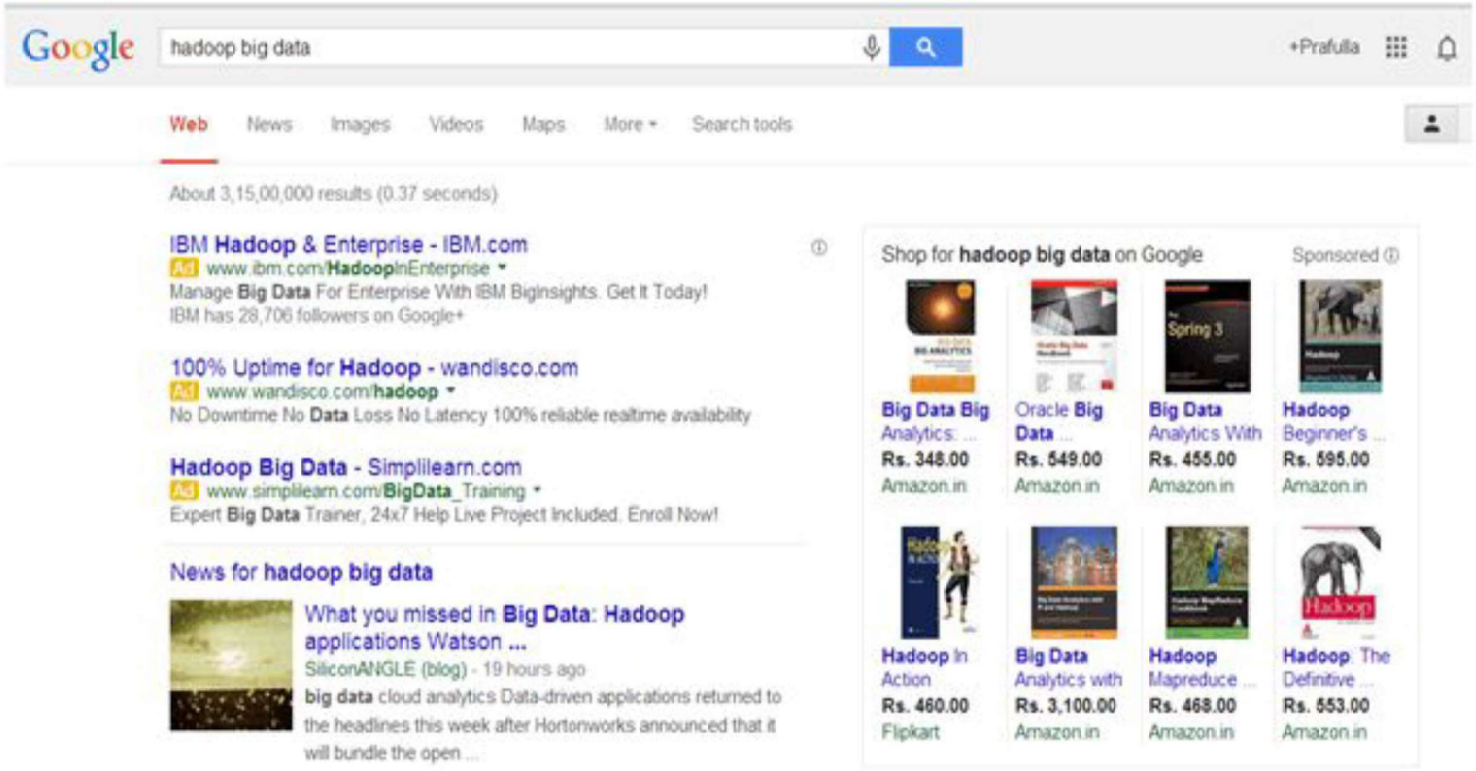
Examples Of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured Data

- Any data with unknown form or the structure is classified as unstructured data
- Examples Of Un-structured Data

Output returned by 'Google Search'



Semi-structured Data

- Semi-structured data can contain both the forms of data.
- We can see semi-structured data as a structured in form but it is actually not defined
- Example of semi-structured data is a data represented in XML file.

Examples Of Semi-structured Data

Personal data stored in a XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Nature of Data

Variety: Types of Data

- Structured data
  - Like tables with fixed attributes
  - Traditionally handled by relational databases
- Unstructured data
  - Usually generated by humans
  - E.g. natural language, voice, Wikipedia, Twitter posts
  - Must be processed into (semi-structured) data to gain value
- Semi-structured data
  - Has some structure in tags but it changes with documents
  - E.g. HTML, XML, JSON files, server logs

What is Big Data

- Use data from multiple sources and in multiple forms
- Involve unstructured and semi-structured data

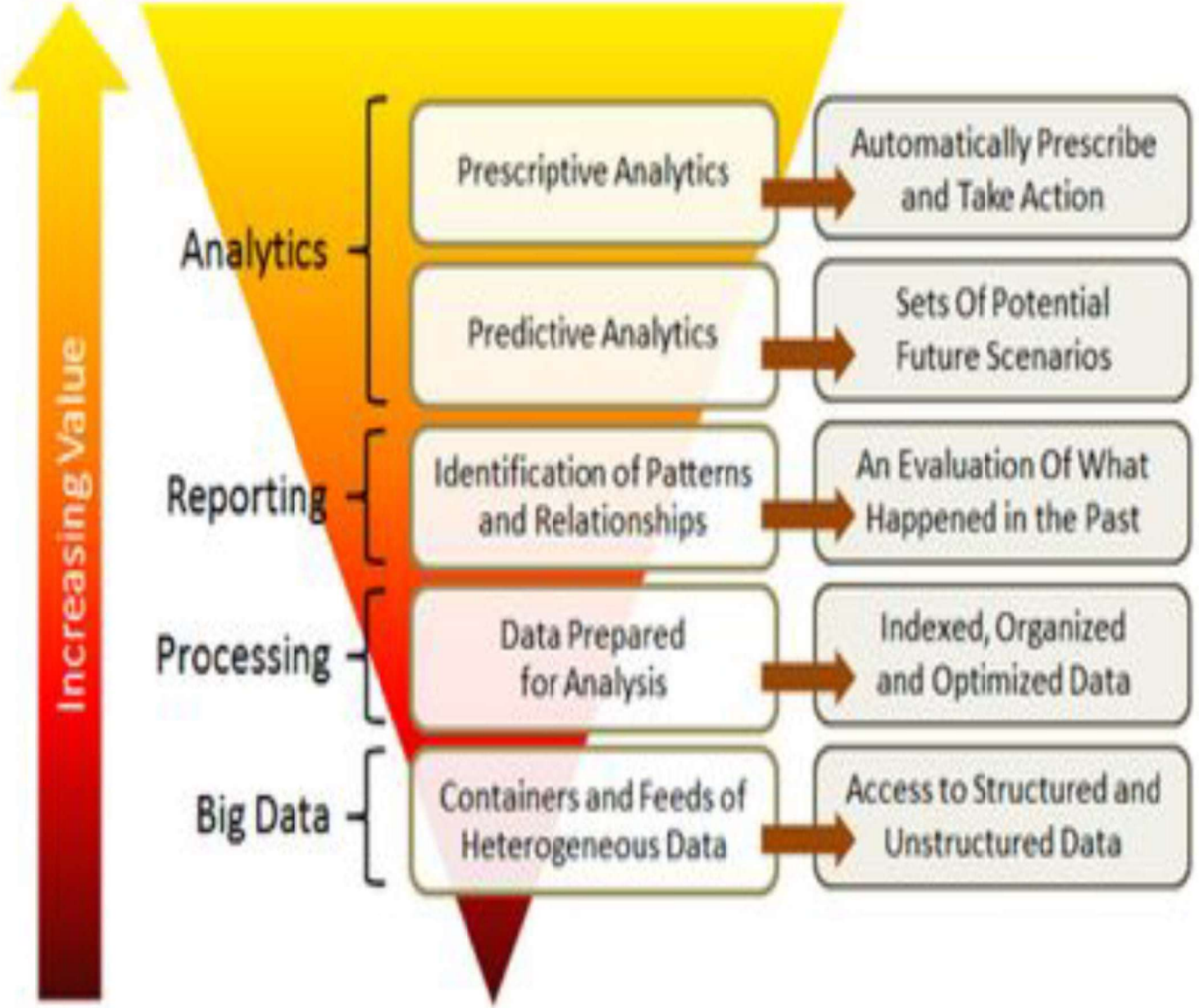
Types of Data Analytics and Value of Data

1. Descriptive analytics (Beschreiben)
   - "What happened?"
2. Diagnostic analytics
   - "Why did this happen, what went wrong?"
3. Predictive analytics (Vorhersagen)
   - "What will happen?"
4. Prescriptive analytics (Empfehlen)
   - "What should we do and why?"

The level of insight and value of data increases from step 1 to 4

The Value of Data (alternative view)

There are many visualizations of the processing and value chain [8]



Analytic Processes and Tools

- Train a classifier
- Preprocessing raw data
- Converting data into training data for classifier
- Converting classifiable data into vectors

Analytic Processes and Tools

- HDFS- Hadoop distributed file systems
- to enable the storage of large files, and does this by distributing the data among a pool of data nodes.
- The creation of a file in HDFS appears to be a single file, even though it blocks "chunks" of the file into pieces that are stored on individual data nodes.
- ZOOKEEPER- "Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services."
- HBASE- HBase is derived from Google's Bigtable and is a column-oriented data layout that, when layered on top of Hadoop, provides a fault-tolerant method for storing and manipulating large.
- HIVE- Hive is layered on top of the file system and execution framework for Hadoop and enables applications
- PIG- the Pig environment allows developers to create new user defined functions
- MAHOUT- Mahout is a project to provide a library of scalable implementations of machine learning algorithms on top of Map Reduce and Hadoop

Tools typically used in Big-Data scenarios

- NoSQL
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

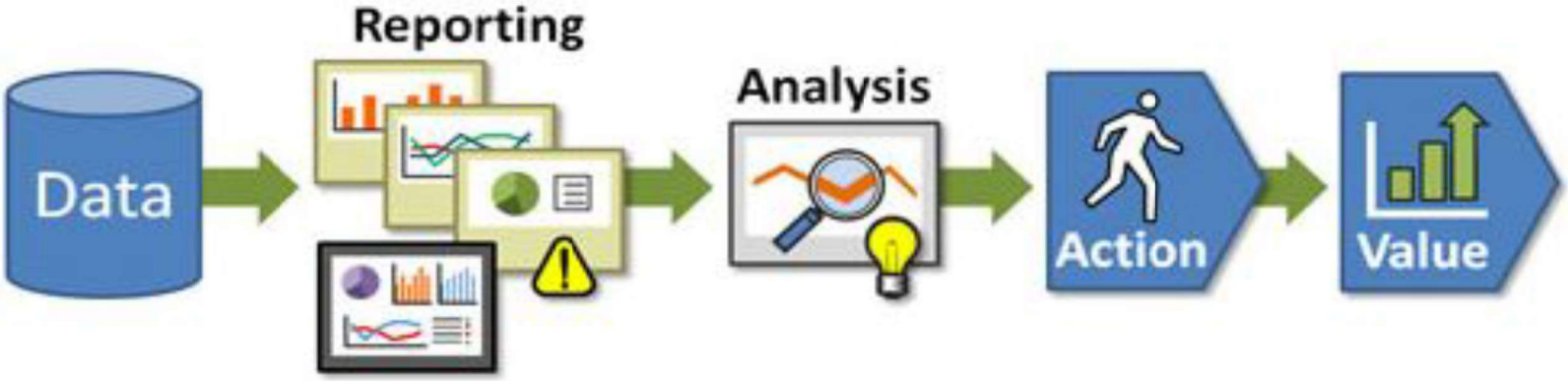
The project includes these modules:

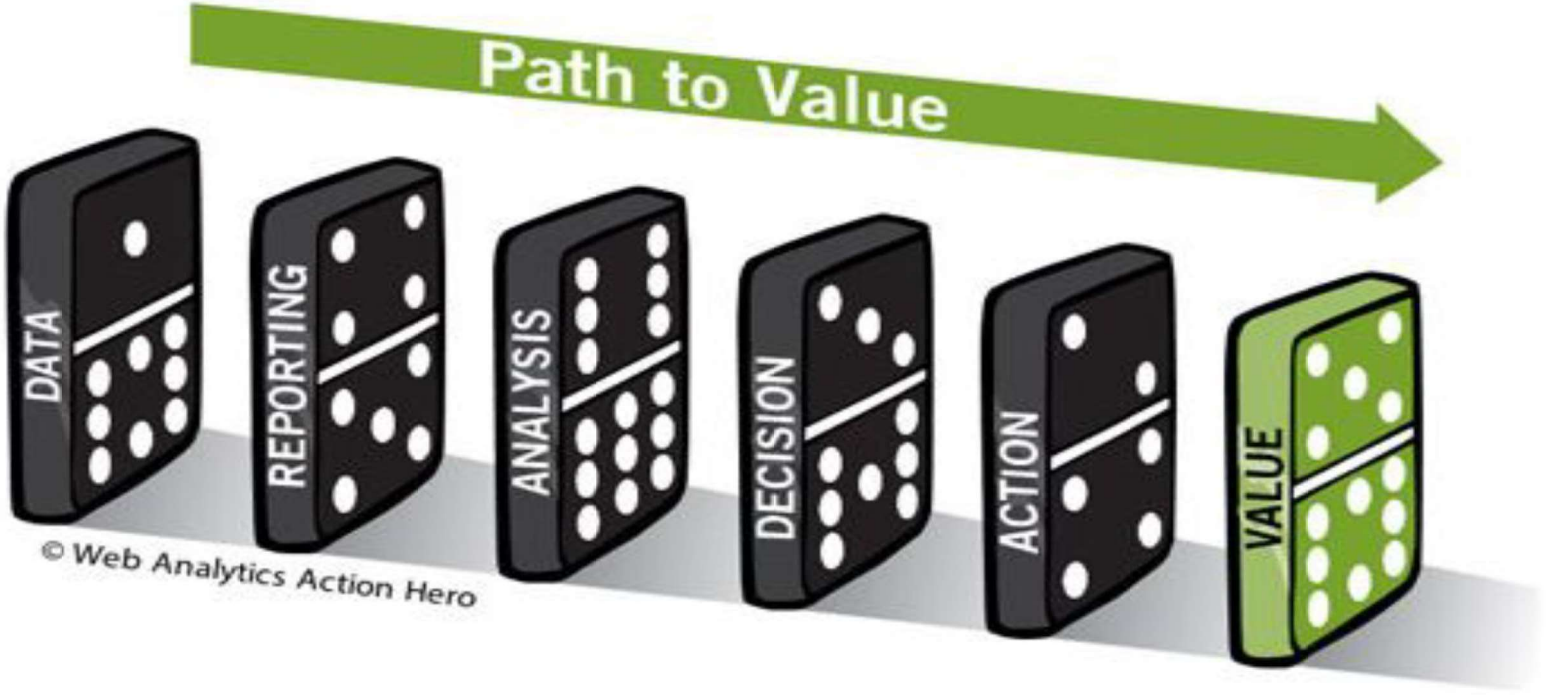
- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Analysis vs Reporting

- reporting translates data into information while analysis turns information into insights
- reporting should enable users to ask “What?” questions about the information, whereas analysis should answer to “Why”” and “What can we do about it?”
- 5 differences between reporting and analysis:
- 1. Purpose
- 2. Tasks
- 3. Outputs
- 4. Delivery
- 5. Value

	Purpose	Tasks	Outputs	Delivery	Value
Reporting	• Monitor and alert	• Build • Configure • Consolidate • Organize • Format • Summarize	• Canned reports • Dashboards • Alerts	• Accessed via tool • Scheduled for delivery	• Distills data into information for further analysis • Alerts company to exceptions in data
Analysis	• Interpret and recommend actions	• Question • Examine • Interpret • Compare • Confirm	• Ad hoc responses • Analysis presentations (findings + recommendations)	• Prepared and presented by analyst	• Provides deeper insights into business • Offers recommendations to drive action


Reporting
Data
Analysis
Action
Value


Path to Value
DATA
REPORTING
ANALYSIS
DECISION
ACTION
VALUE
© Web Analytics Action Hero

- **Purpose :** Before covering the differing roles of reporting and analysis, let's start with some high-level definitions of these two key areas of analytics.
- **Reporting:** The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.
- **Analysis:** The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.
- **Outputs :**three main types of reporting: canned reports, dashboards, and alerts.
- **Delivery: reporting** is more of a push model, where people can access reports through an analytics tool, Excel spreadsheet, widget, or have them scheduled for delivery into their mailbox, mobile device, FTP site, etc.
- analysis is all about human beings using their superior reasoning and analytical skills to extract key insights from the data and form actionable recommendations for their organizations

1. Purpose

- Reporting helps companies monitor their data even before digital technology boomed.
- Various organizations have been dependent on the information it brings to their business, as reporting extracts that and makes it easier to understand.
- Analysis interprets data at a deeper level. While reporting can link between cross-channels of data, provide comparison, and make understand information easier (think of a dashboard, charts, and graphs, which are reporting *tools* and not analysis reports), analysis interprets this information and provides recommendations on actions.

2. Tasks

- As reporting and analysis have a very fine line dividing them, sometimes it's easy to confuse tasks that have analysis labeled on top of them when all it does is reporting. Hence, ensure that your analytics team has a healthy balance doing both.
- Here's a great differentiator to keep in mind if what you're doing is reporting or analysis:
- Reporting includes building, configuring, consolidating, organizing, formatting, and summarizing. It's very similar to the abovementioned like turning data into charts, graphs, and linking data across multiple channels.
- Analysis consists of questioning, examining, interpreting, comparing, and confirming. With big data, predicting is possible as well.

3.Output

- Reporting and analysis have the push and pull effect from its users through their outputs. Reporting has a push approach, as it pushes information to users and outputs come in the forms of canned reports, dashboards, and alerts.
- Analysis has a pull approach, where a data analyst draws information to further probe and to answer business questions. Outputs from such can be in the form of ad hoc responses and analysis presentations. Analysis presentations are comprised of insights, recommended actions, and a forecast of its impact on the company—all in a language that's easy to understand at the level of the user who'll be reading and deciding on it.
- This is important for organizations to realize truly the value of data, such that a standard report is not similar to a meaningful analytics.

4.Delivery

- Considering that reporting involves repetitive tasks—often with truckloads of data, automation has been a lifesaver, especially now with big data. It's not surprising that the first thing outsourced are data entry services since outsourcing companies are perceived as data reporting experts.
- Analysis requires a more custom approach, with human minds doing superior reasoning and analytical thinking to extract insights, and technical skills to provide efficient steps towards accomplishing a specific goal.
- This is why data analysts and scientists are demanded these days, as organizations depend on them to come up with recommendations for leaders or business executives make decisions about their businesses.

5.VALUE

- This isn't about identifying which one brings more value, rather understanding that both are indispensable when looking at the big picture. It should help businesses grow, expand, move forward, and make more profit or increase their value.
- This Path to Value diagram illustrates how data converts into value by reporting and analysis such that it's not achievable without the other.
- **Data — Reporting — Analysis — Decision-making — Action — VALUE**
- Data alone is useless, and action without data is baseless. Both reporting and analysis are vital to bringing value to your data and operations.

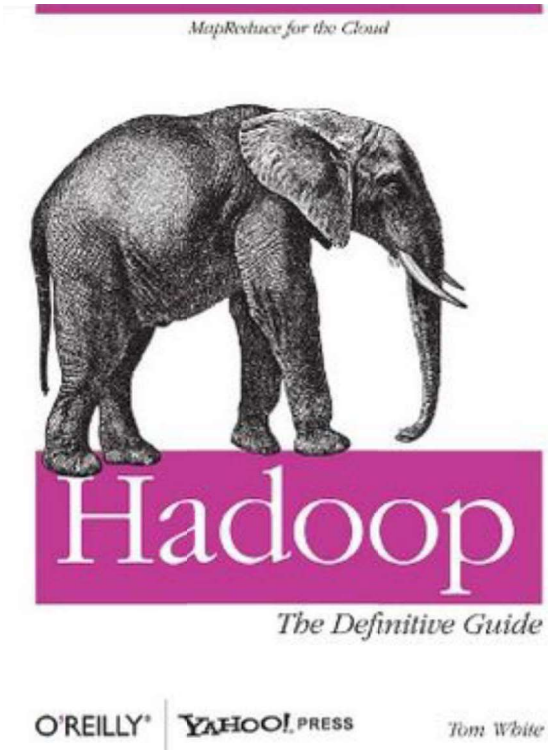
Reference

- https://blogs.adobe.com/digitalmarketing/analytics/reporting-vs-analysis-whats-the-difference/
- http://www.infinitdatum.com/blog/5-differences-between-reporting-and-analysis/

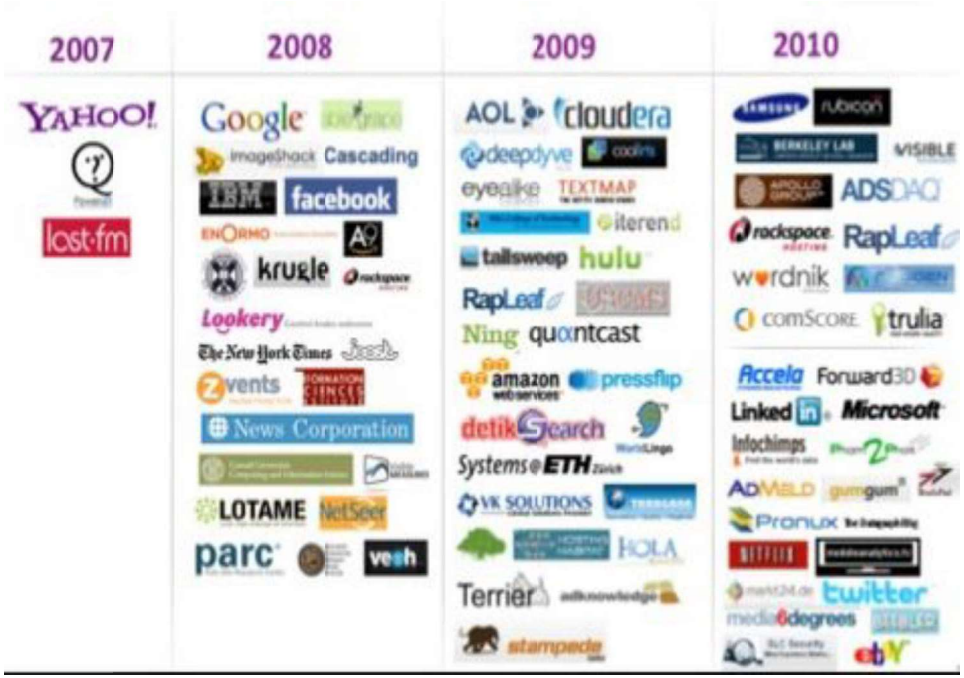
Modern Data Analytic Tools

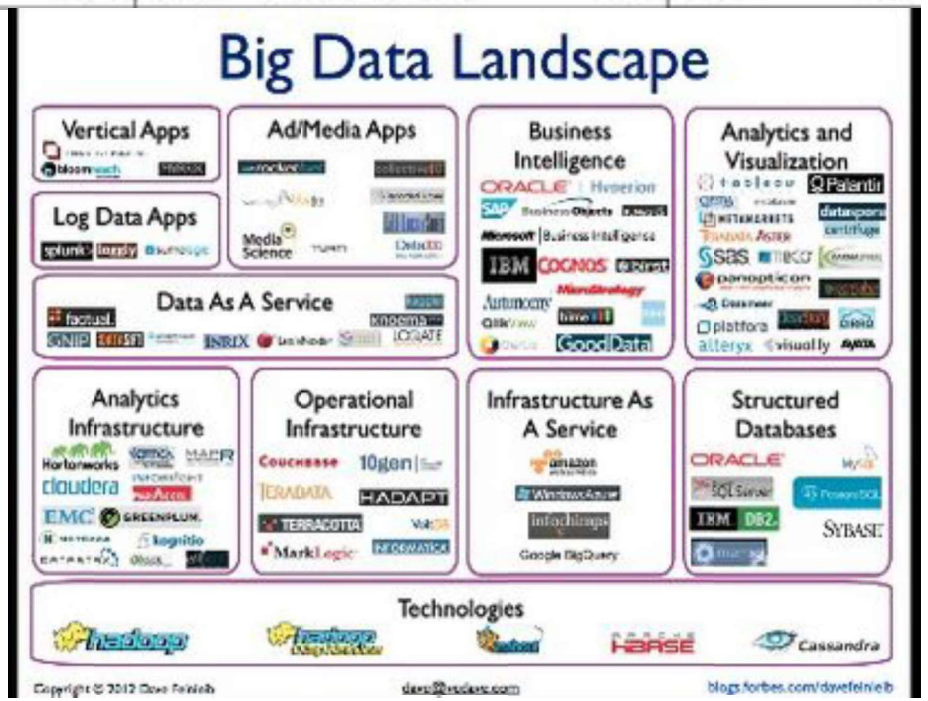
Tools typically used in Big-Data scenarios

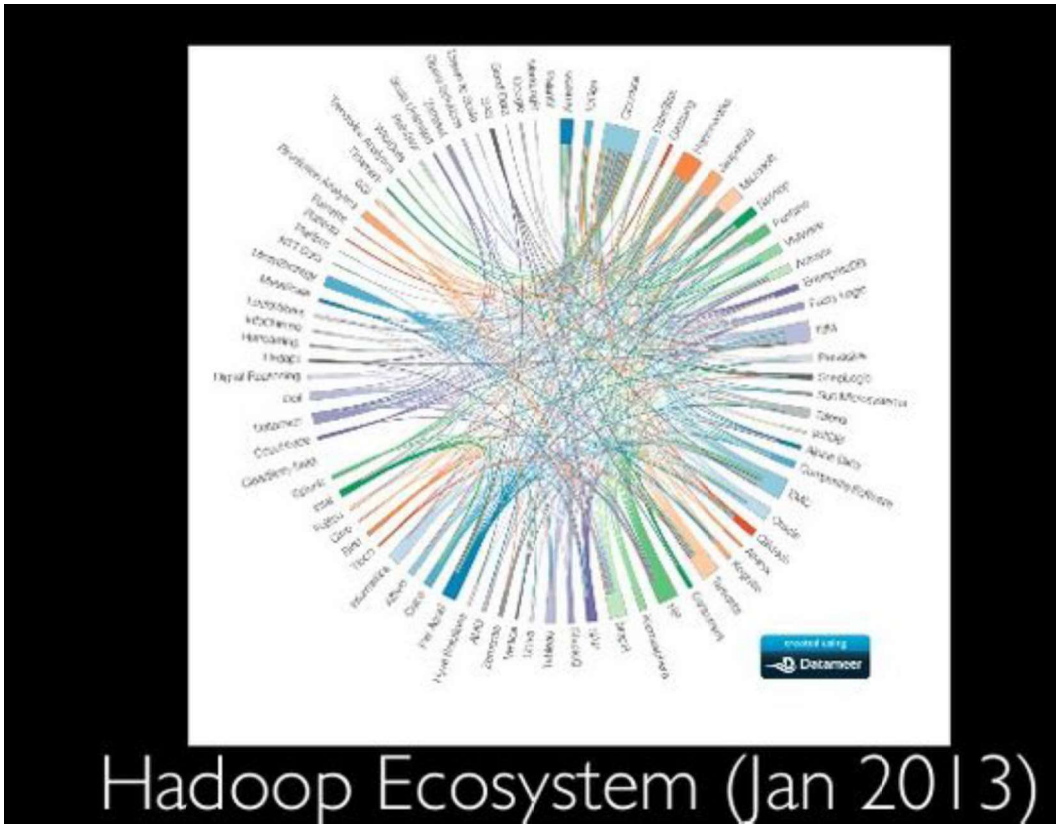
- NoSQL
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop


MapReduce for the Cloud
Hadoop
The Definitive Guide
O'REILLY
YAHOO! PRESS
Tom White

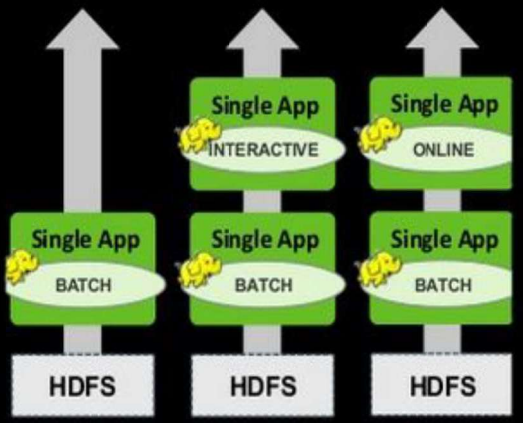

Hadoop
Hadoop Operations
MapReduce Design Patterns
Pig
HBase
Hive
Hadoop for Oracle Developers
Spring Data
Big Data for Chimps


2007
2008
2009
2010
YAHOO!
lost·fm
Google
facebook
Cascading
Lookery
The New York Times
News Corporation
LOTAME
parc
AOL
cloudera
deepdyve
eyealike
TEXTMAP
tailsweep
hulu
RapLeaf
Ning
quantcast
amazon web services
pressflip
detikSearch
Systems@ETH Zürich
VK SOLUTIONS
HOLA
Terrier
stampede
Accela
Forward3D
Linked in
Microsoft
Infochimps
AdMeld
gumgum
Pronux
twitter
media6degrees
ebY
rubicon
VISIBLE
ADSDAQ
rackspace
RapLeaf
wordnik
comScore
trulia

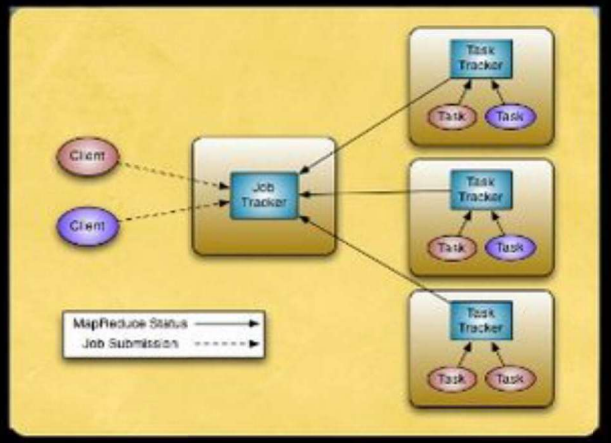

Big Data Landscape
Vertical Apps
Log Data Apps
Ad/Media Apps
Data As A Service
Business Intelligence
Analytics and Visualization
Analytics Infrastructure
Operational Infrastructure
Infrastructure As A Service
Structured Databases
Technologies
hadoop
Cassandra
HBASE
Google BigQuery
Copyright © 2012 Dave Feinleib
blogs.forbes.com/davefeinleib



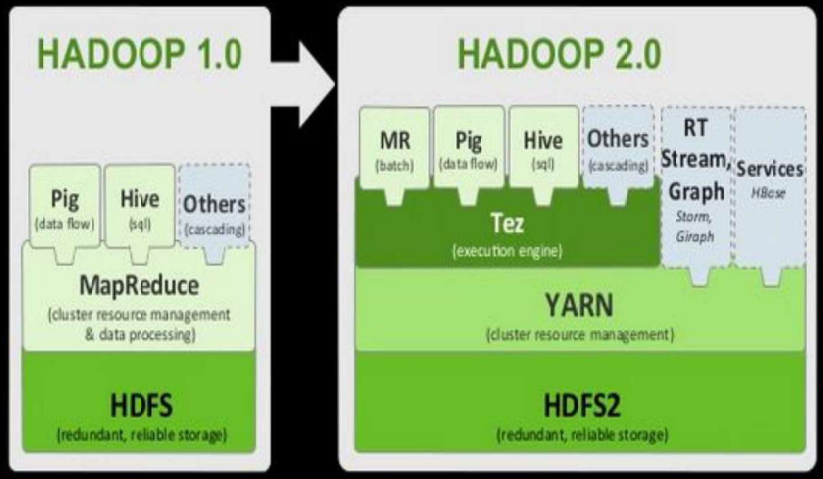
Hadoop Ecosystem (Jan 2013)



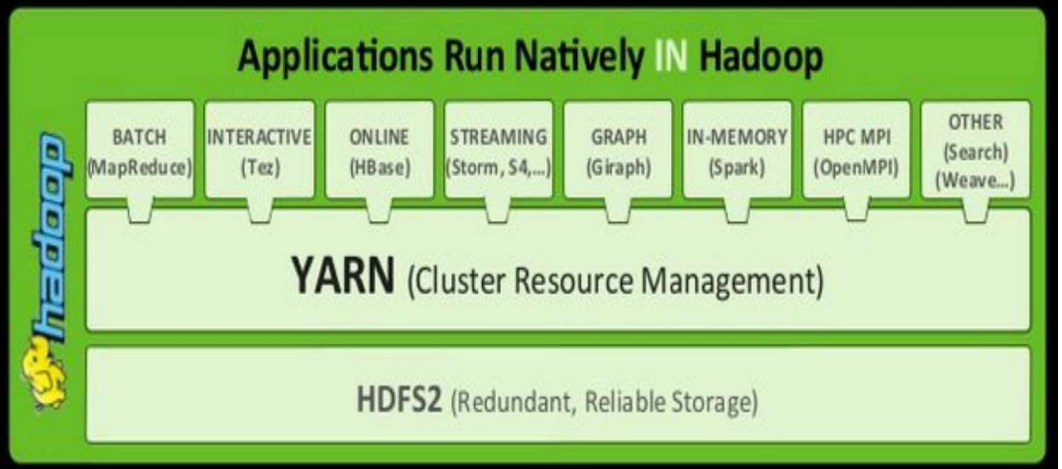
Hadoop 1.0



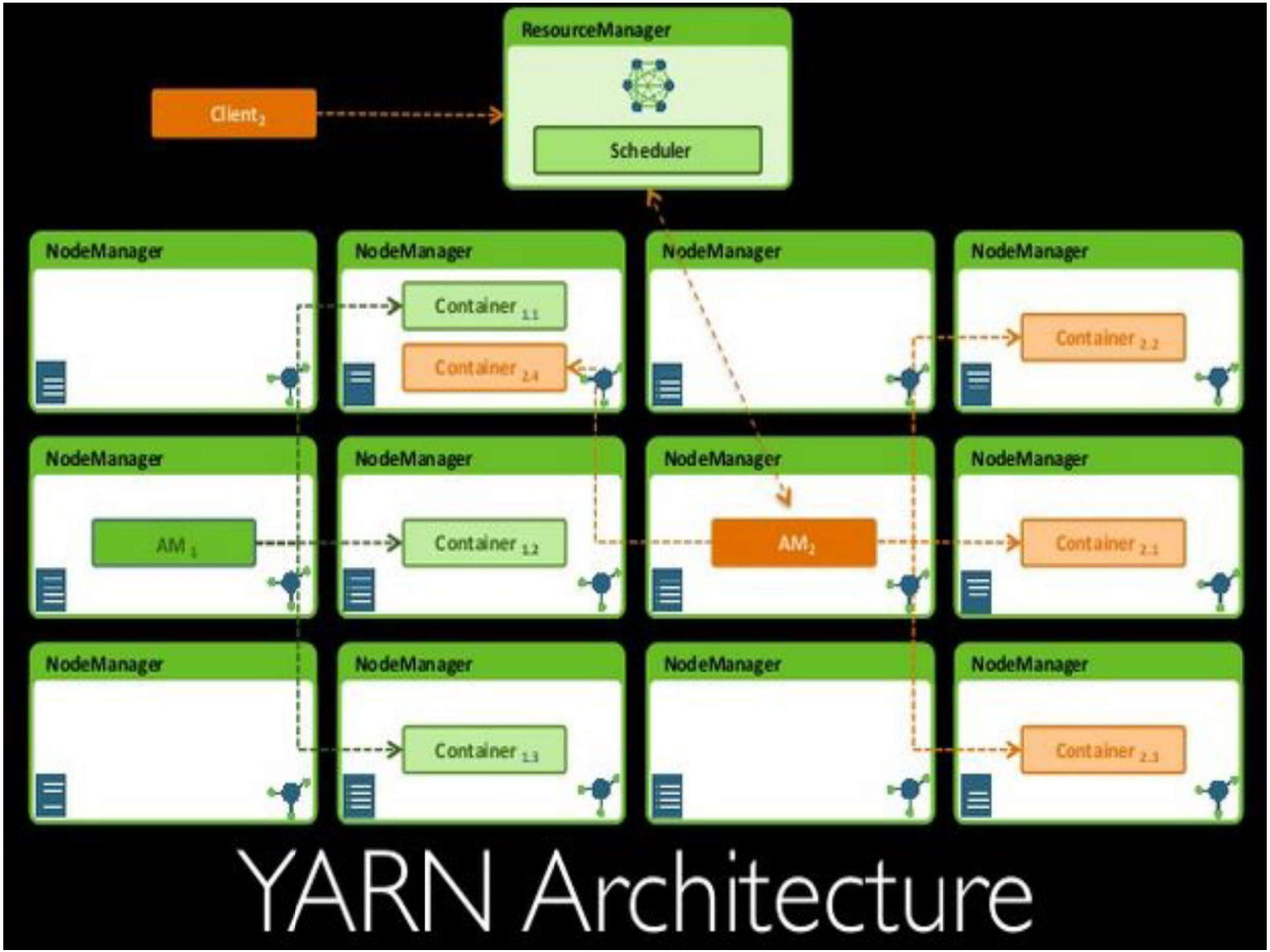
MapReduce 1.0



Hadoop 2.0



YARN Platform


ResourceManager
Client2
Scheduler
NodeManager
Container 1.1
Container 2.4
Container 2.2
AM 1
Container 1.2
AM2
Container 2.1
Container 1.3
Container 2.3
YARN Architecture

YARN

- Yet Another Resource Negotiator
- Resource Manager
- Node Managers
- Application Masters
  - Specific to paradigm, e.g. MR Application master (aka JobTracker)

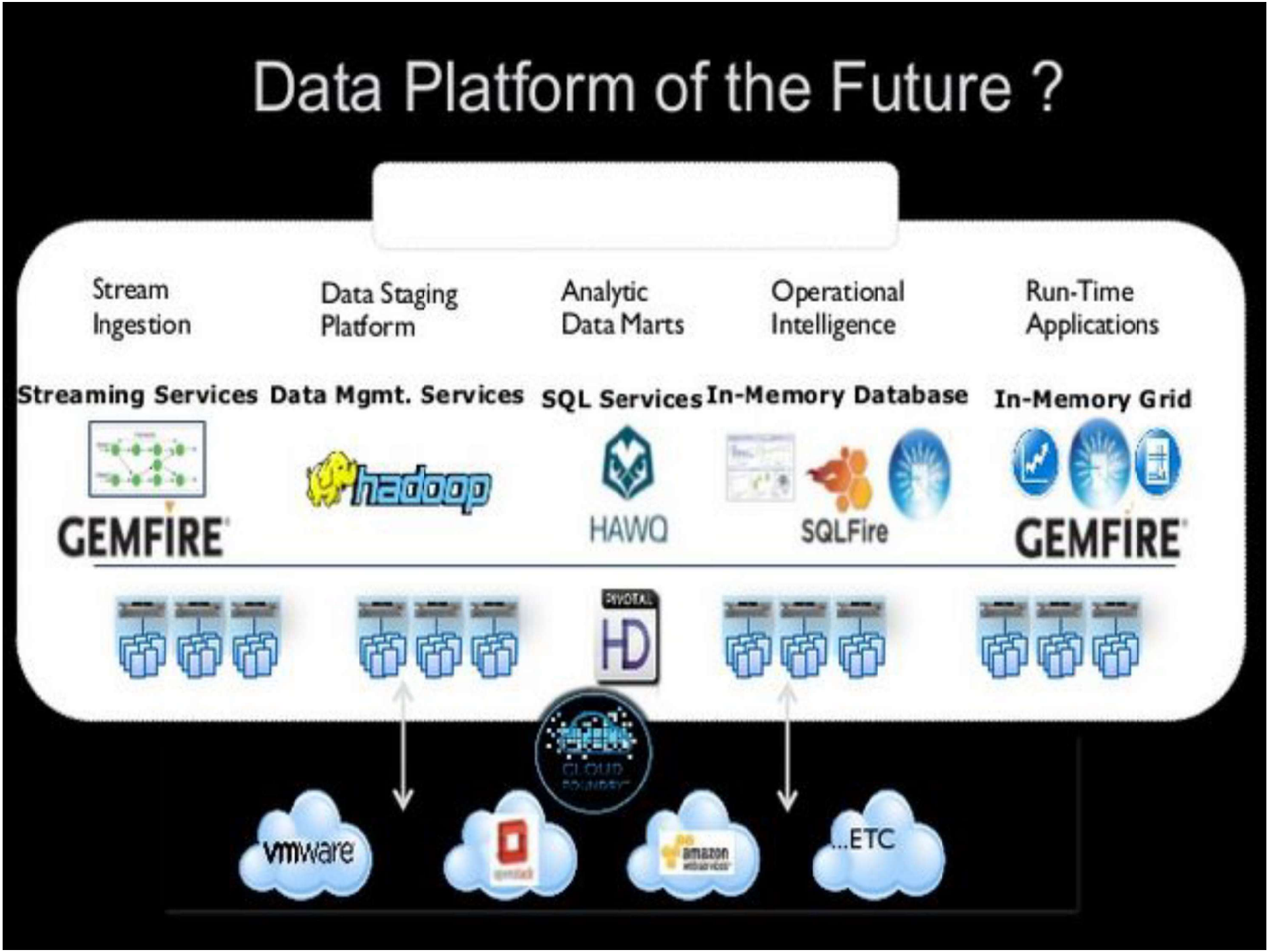
Beyond MapReduce

- Apache Giraph - BSP & Graph Processing
- Storm on Yarn - Streaming Computation
- HOYA - HBase on Yarn
- Hamster - MPI on Hadoop
- More to come ...

Hamster

- Hadoop and MPI on the same cluster
- OpenMPI Runtime on Hadoop YARN
- Hadoop Provides: Resource Scheduling, Process monitoring, Distributed File System
- Open MPI Provides: Process launching, Communication, I/O forwarding

Data Platform of the Future ?



- http://www.slideshare.net/GWOcon/great-wideopentalk

Statistical Concepts: Sampling Distributions

- The sampling distribution is a distribution of a sample statistic. While the concept of a distribution of a set of numbers is intuitive for most students.
- The sampling distribution is a distribution of a sample statistic. It is a model of a distribution of scores, like the population distribution, except that the scores are not raw scores, but statistics. It is a thought experiment; "what would the world be like if a person repeatedly took samples of size N from the population distribution and computed a particular statistic each time?" The resulting distribution of statistics is called the sampling distribution of that statistic.
- For example, suppose that a sample of size sixteen (N=16) is taken from some population. The mean of the sixteen numbers is computed. Next a new sample of sixteen is taken, and the mean is again computed. If this process were repeated an infinite number of times, the distribution of the now infinite number of sample means would be called the sampling distribution of the mean.
- Every statistic has a sampling distribution. For example, suppose that instead of the mean, medians were computed for each sample. The infinite number

Re-Sampling

- In statistics, **resampling** is any of a variety of methods for doing one of the following:
- Estimating the precision of sample statisticsEstimating the precision of sample statistics (mediansEstimating the precision of sample statistics (medians, variancesEstimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (**jackknifing**) or drawing randomly with replacement from a set of data points (**bootstrapping**)
- Exchanging labels on data points when performing significance tests (**permutation tests**, also called exact tests, randomization tests, or re-randomization tests)
- Validating models by using random subsets (bootstrapping, cross validation)
- Common resampling techniques include bootstrapping,

Statistical Inference

- **Statistical Inference, Model & Estimation**
- Recall, a **statistical inference** aims at learning characteristics of the population from a sample; the population characteristics are *parameters* and sample characteristics are *statistics*.
- A **statistical model** is a representation of a complex phenomena that generated the data.
- It has mathematical formulations that describe relationships between random variables and parameters.
- It makes assumptions about the random variables, and sometimes parameters.
- A general form: data = model + residuals
- Model should explain most of the variation in the data
- Residuals are a representation of a lack-of-fit, that is of the portion of the data unexplained by the model.

- **Estimation** represents ways of a process of learning and determining the population parameter based on the model fitted to the data.
- Point estimation and interval estimation, and hypothesis testing are three main ways of learning about the population parameter from the sample statistic.
- An **estimator** is particular example of a statistic, which becomes an **estimate** when the formula is replaced with actual observed sample values.
- **Point estimation** = a single value that estimates the parameter. Point estimates are single values calculated from the sample
- **Confidence Intervals** = gives a range of values for the parameter Interval estimates are intervals within which the parameter is expected to fall, with a certain degree of confidence.
- **Hypothesis tests** = tests for a specific value(s) of the parameter.
- In order to perform these inferential tasks, i.e., make inference about the unknown population parameter from the sample statistic, we need to know the likely values of the sample statistic. What would happen if we do sampling many times?
- We need the **sampling distribution** of the statistic It depends on the model assumptions about the population distribution, and/or on the sample size.
- **Standard error** refers to the standard deviation of a sampling distribution.

Prediction Error

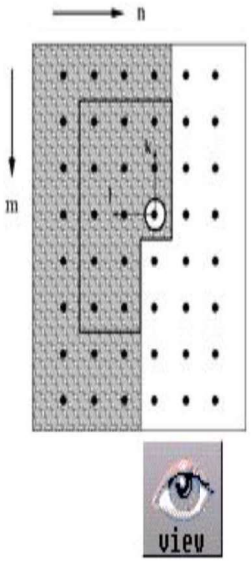
- Prediction error is a discontinuity attribute that removes the predictable image components and reveals the unpredictable.
- To use prediction error as a discontinuity attribute - the original goal and starting point of my project - one has to devise a prediction-error computation that predicts and removes the plane-wave volumes of sedimentary layers but that is incapable of predicting the discontinuities.
- Ref : http://sep.stanford.edu/public/docs/sep99/cohy_Fig/paper_html/node38.html
-

the prediction error is simply .

The region S_x defines which neighboring values contribute to the linear prediction. Causal predictions, that involve regions of the shape shown in Figure 35, lead to white noise driven output images. The region is calle *causal* since given a hypothetical scan from top to bottom and left to right all points of S_x lie to one-side of the predicted value $u(m,n)$.

pefDomain

Figure 35 Prediction-error domain. The causal domain ensures that the output of the prediction-error filter tends to be white noise.



To compute the prediction error of a given stationary image, we first find the prediction coefficients $a(k,l)$ that minimize the prediction error for all pixels of the input image. Once the prediction coefficients are known, convolution

computes the prediction error. In particular, a prediction-error filter potentially zeroes a random plane-wave, or a superposition of random plane waves, or a superposition of random constant-amplitude lines. I repres prediction-error computation as

where g is the input image, A is the prediction-error operator, and ϵ the residual prediction error.

To compute the prediction error of a nonstationary image, I, as usual, divide the image into stationary patches, compute the prediction error for each patch, and merge the patches containing the local prediction error single quilt. All result images of this section are smoothed along the vertical axis to suppress the prediction error's tendency to enhance high-frequency noise of the original unfiltered image.