Big Data
- Introduction

Dr. Qing “Matt” Zhang
mzhang@itu.edu
ITU

Introduction

- About the instructor
- About you
  - Experience with Java?
  - Experience with Linux?
  - Experience with database?
  - Experience with Hadoop?
  - Expectation from this course?

Grading

Assignments	35%
Discussions	5%
Mid-term	30%
Final	30%

Textbook

Hadoop: The Definitive Guide

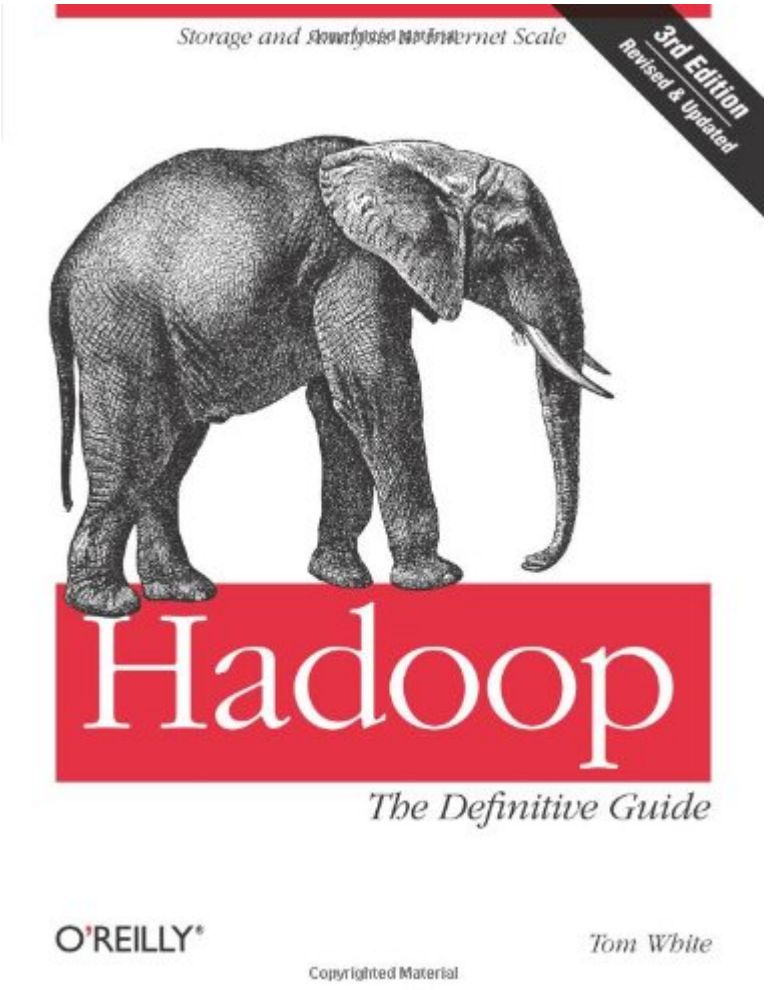
April 11, 2015

(4th edition)

by Tom White (Author)

ISBN-10: 1491901632

ISBN-13: 978-1491901632



Schedule

- Jan. 23/24
  - Big Data Overview
  - Hadoop Architecture
  - Hadoop Installation
  - Assignment
- Mar. 5/6
  - Hadoop Admin
  - mid-term
  - Hadoop Development

Schedule (cont'd)

- Apr. 23/24
  - Hadoop Development
  - Other NoSQL stack (Hive, Spark, etc …)
  - final

What is Big Data?

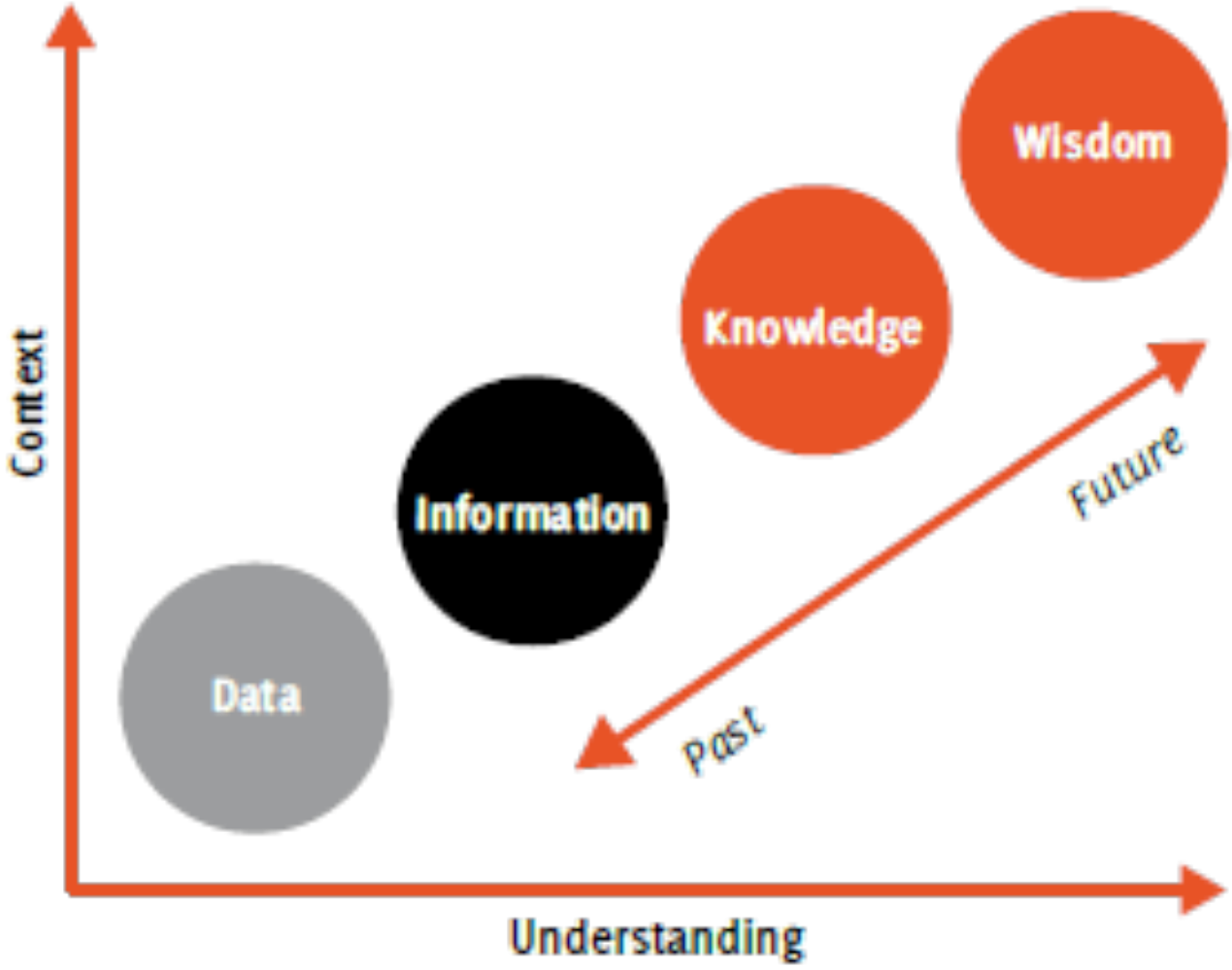
Data

- Data is generated from all businesses/ everyday life
  - Social media
  - Mobile devices
  - Financial transactions
  - Business processes
  - Sensor, IoT

Examples

- Every day
  - 864 million tweets
    - http://www.internetlivestats.com/twitter-statistics/
  - 4.5 billion Facebook “likes”
    - http://blog.wishpond.com/post/115675435109/40-up-to-date-facebook-facts-and-stats
  - One billion objects added to Amazon S3 storage

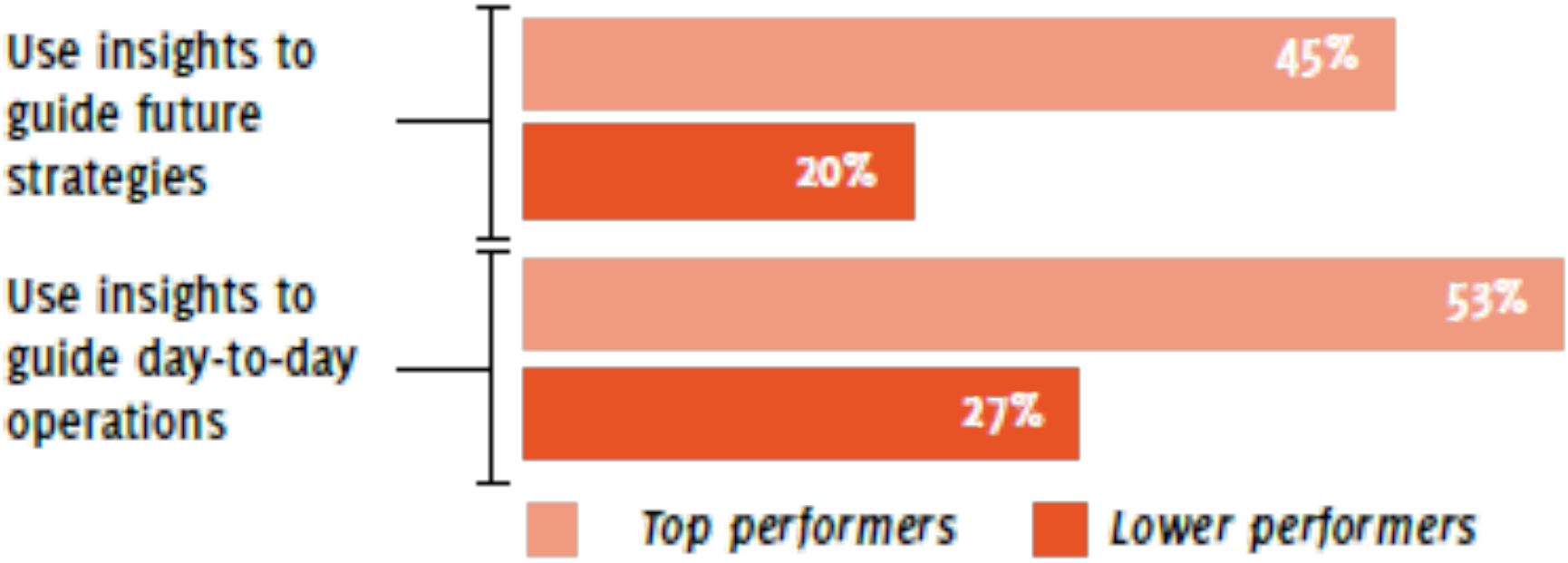
What is Data / Information?



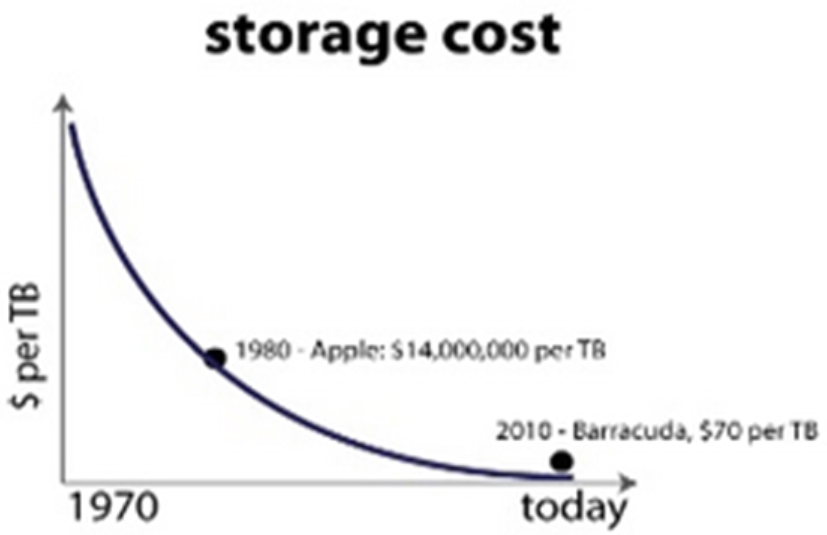
Data is Value

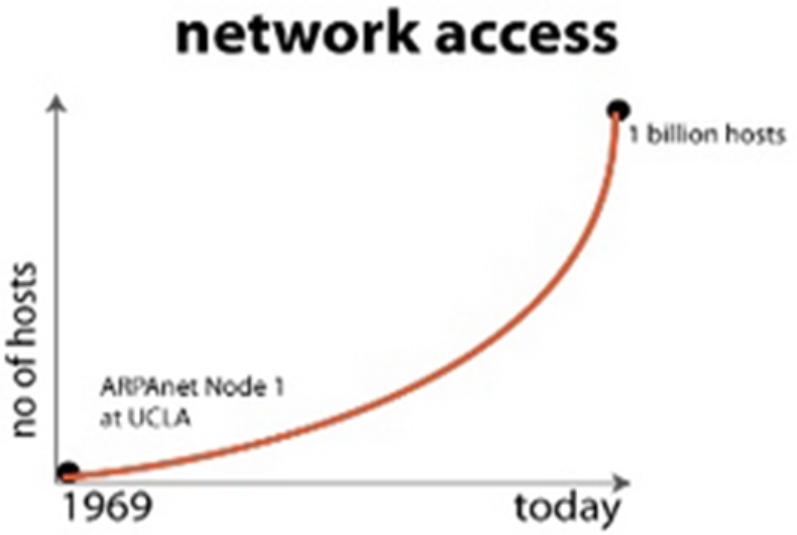
- Data has many valuable applications
  - Marketing analysis
  - Product recommendations
  - Demand forecasting
  - Fraud detection
- Must process it to extract the value

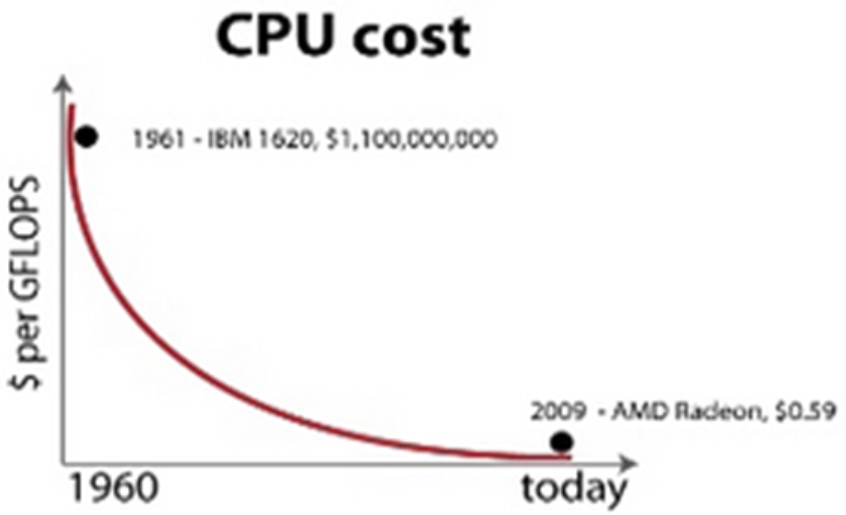
Data is Value

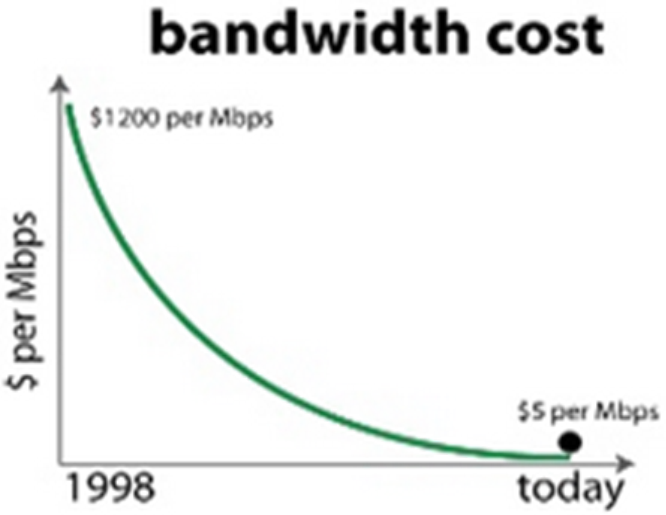


Exponential Growth of Data

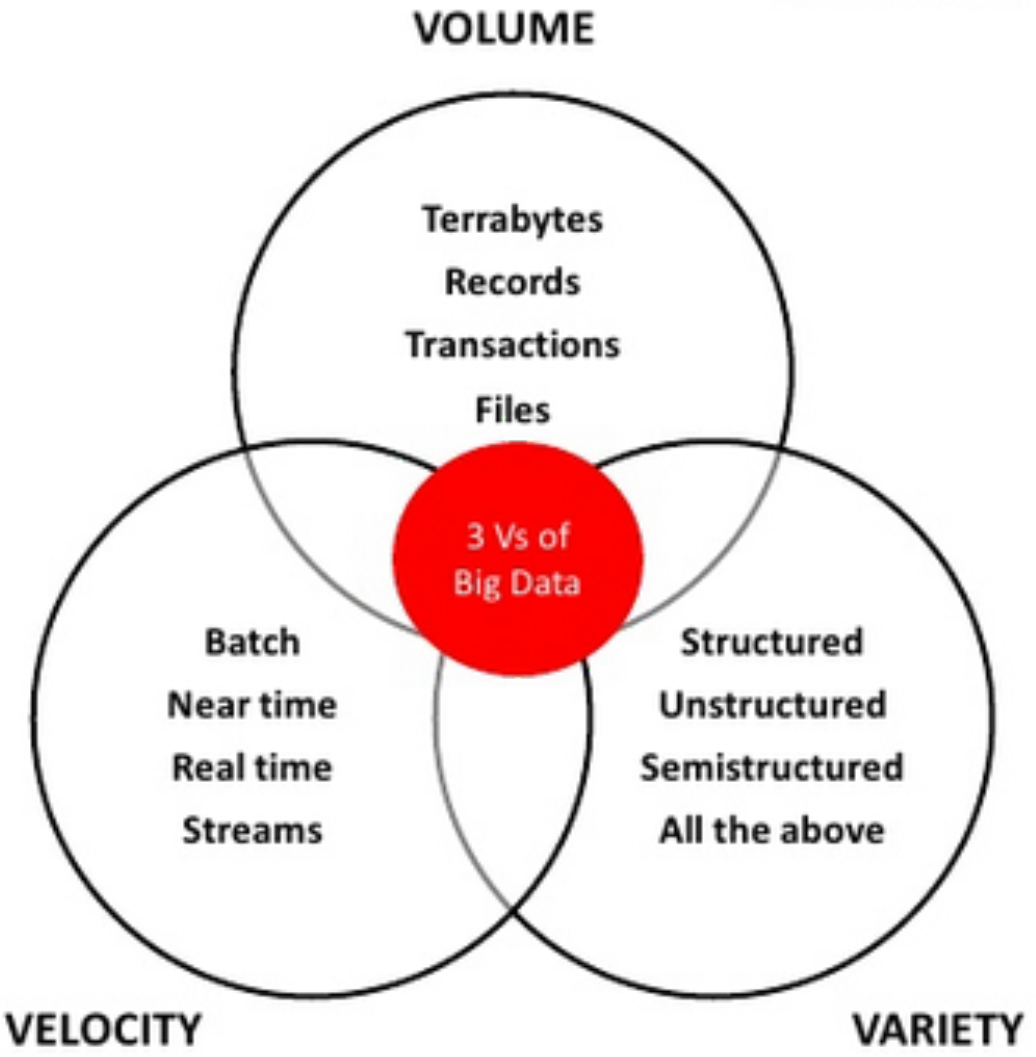








3 V's that Define Big Bata



Volume (Terabytes (1000 GB - 2^{40}) → Zettabytes (billion TB - 2^{70}))

Variety (Structured + Semi-structured + Unstructured)

Velocity (Batch + Streaming Data)

Structured Data

- Structured data:
  - Pre-defined schema imposed on the data
  - Highly structured
  - Usually stored in a relational database system
- Example
  - numbers: 20, 3.1415,. . .
  - dates: 01/01/2015
  - strings: ”Hello World” . . .
  - Roughly 20% of all data out there is structured

Semi-Structured Data

- Inconsistent structure
- Cannot be stored in rows and tables in a typical database.
- Information is often self-describing (label/ value pairs).
- Example
  - XML
  - logs
  - tweets
  - sensor feeds

CAP Theorem

- CAP Theorem (Eric Brewer, 2000)

  For highly scalable distributed system, you can only have two of the following

  - Consistency
    - all nodes sees the same data at the same time
  - Availability
    - every request receives a response
  - Partition tolerance
    - System continues to operate despite link/node failuer
- **Volume + Velocity** ➔ No Consistency
- **Implication**: Big data solutions must stop worrying about consistency if they want high availability

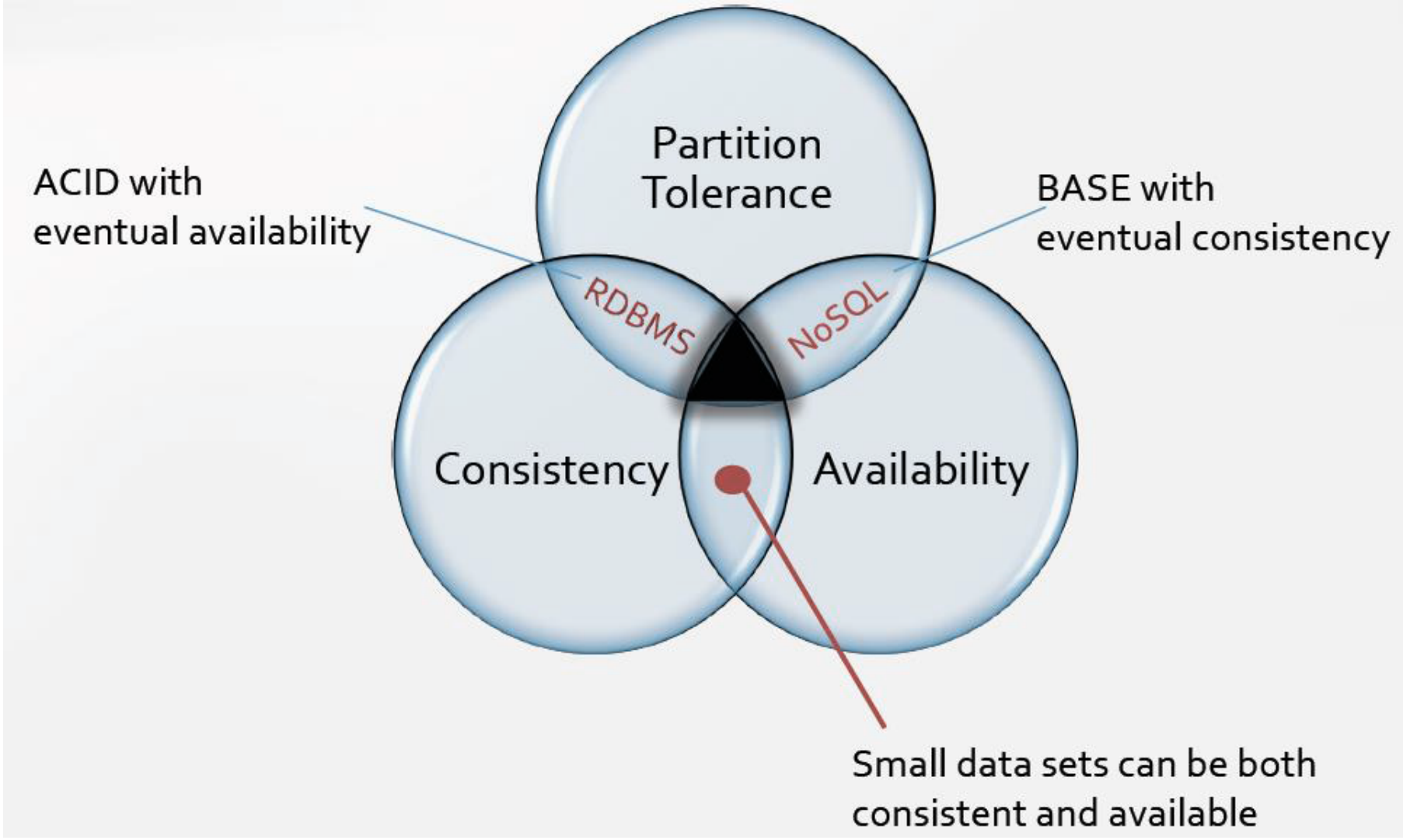
ACID and BASE

- RDBMS provide **ACID**
  - Atomicity: entire transaction fail/succeed
  - Consistency: valid state before/after transaction
  - Isolation: multiple transactions occurring at same time won't affect each other
  - Durability: once transaction committed, data persists

ACID and BASE(cont'd)

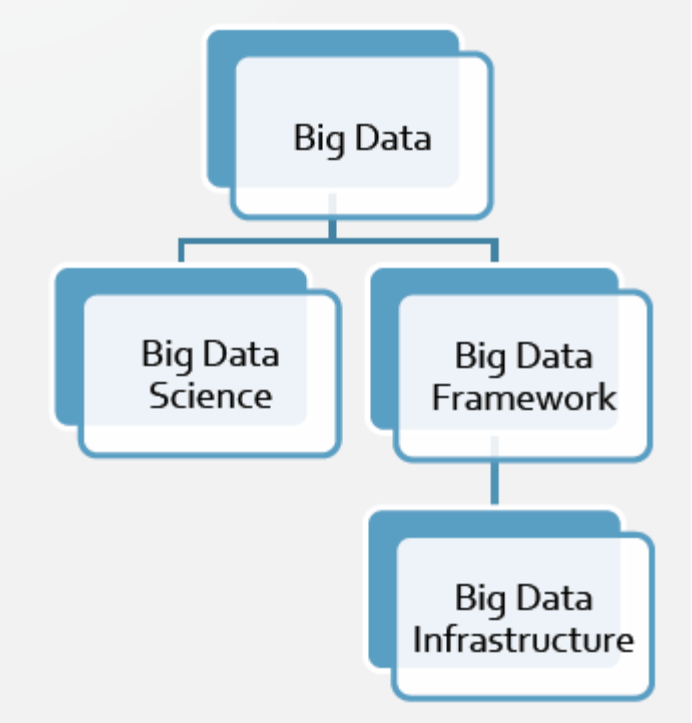
- NoSQL provides **BASE** (**B**asically **A**vailable, **S**oft state, **E**ventual consistency):
  - Basically Available: allow parts of the system to fail
  - Soft state: an object may have multiple simultaneous values
  - Eventual Consistency: consistency happen over time

ACID and BASE Virtualized



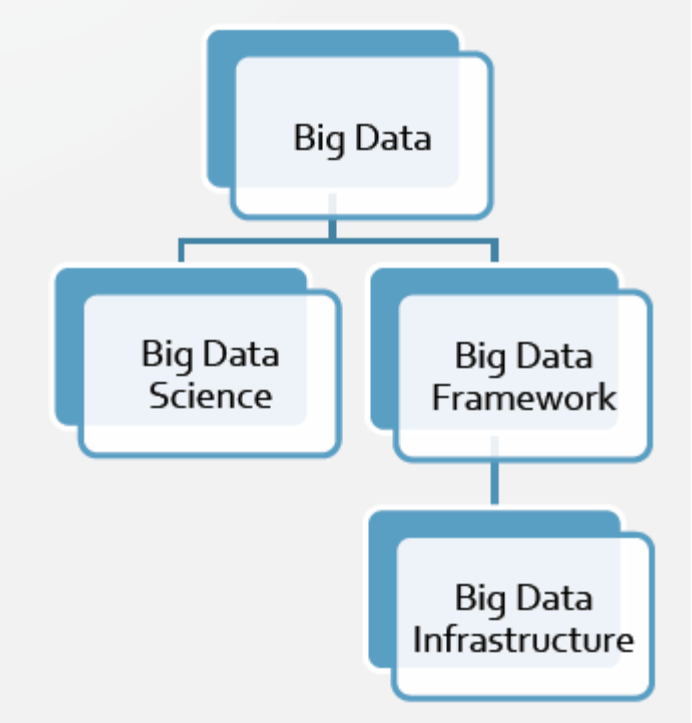
Big Data Science

- **Big Data:** The 3 Vs characterizing Big Data limits the ability to perform analysis using traditional RDBMS
- **Big Data Science**: the study of techniques covering acquisition, conditioning, and evaluation of Big Data; including information technology and math science



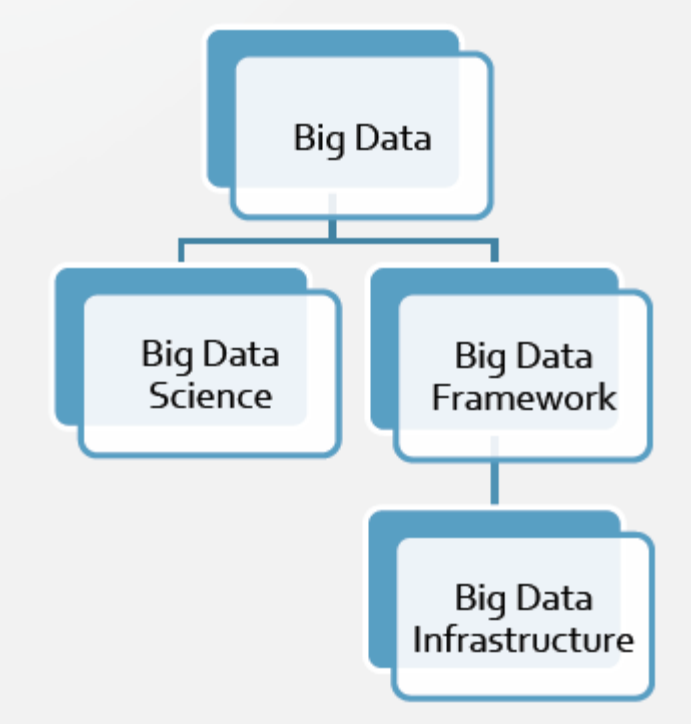
Big Data Frameworks

- **Big Data Frameworks**: software libraries and algorithms that enable distributed processing and analysis of big data across clusters of CPUs, or GPUs, etc



Big Data Infrastructure

- **Big Data Infrastructure**: instances of one/more Big Data Frameworks that include Management APIs and servers (physical or virtual) to solve specific Big Data problem or to serve as general purpose analysis and processing engine



Technologies Used with Big Data

- **Data Mining (DM):** is the process of extracting useful information from massive quantities of complex data
- **Machine Learning (ML):** is the study of systems that improve their performance with experience (typically learning from the data)

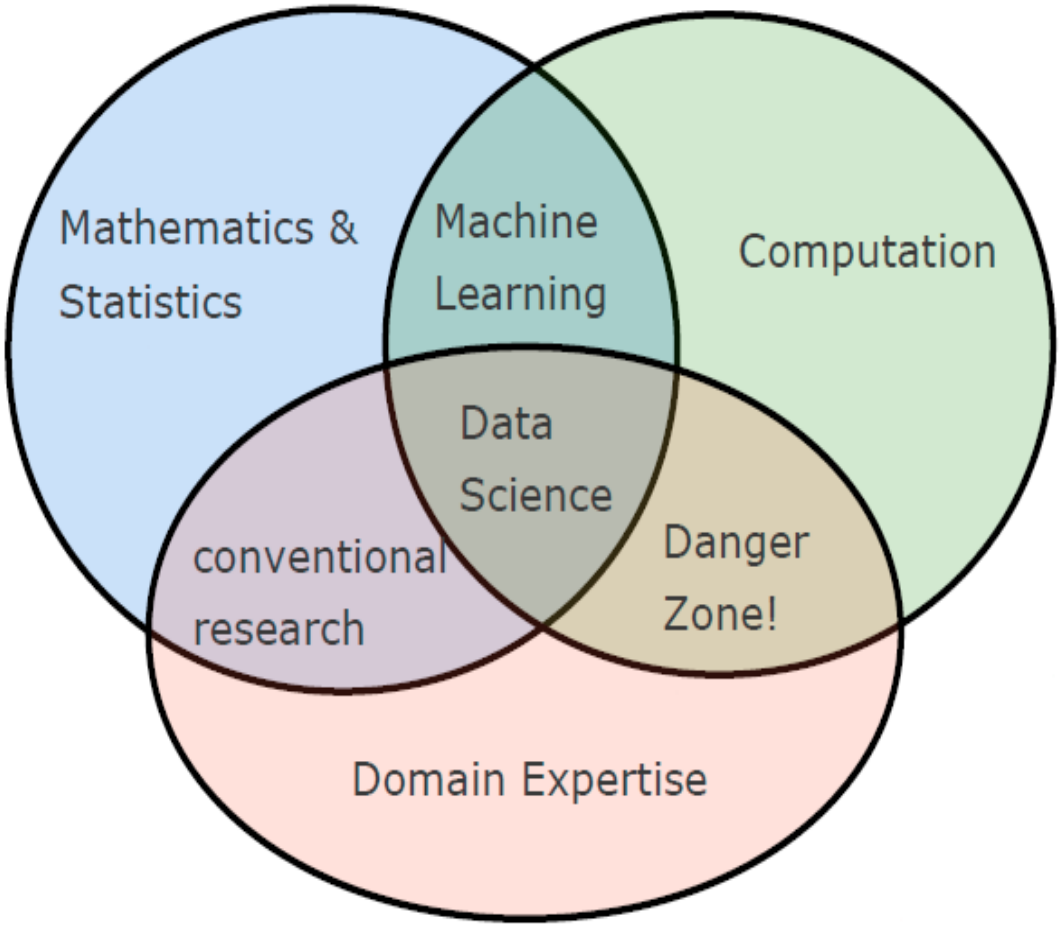
Technologies Used with Big Data

- **Artificial Intelligence (AI):** is the science of automating complex behavior such as learning, problem solving, and decision making
- Many of the technique used are statistical in nature, but are very different from classical statistics

Data Science

Automatically extracting knowledge from data:

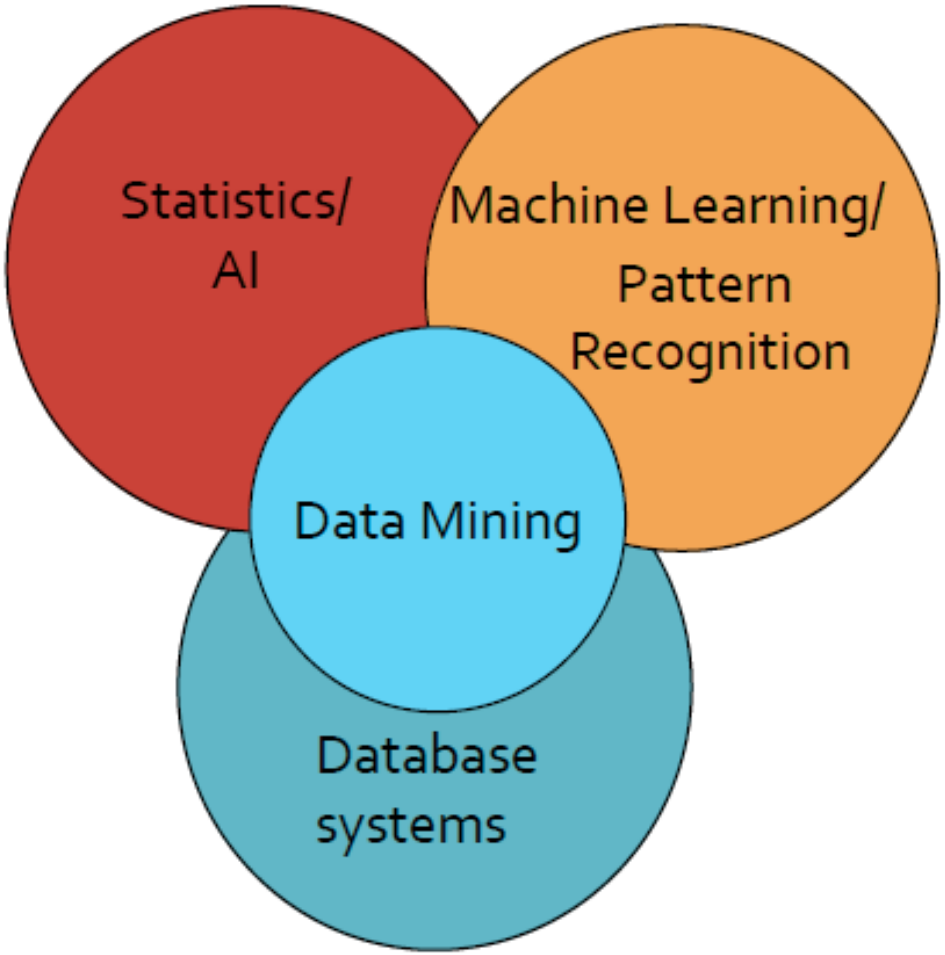
- Mathematics & Statistics
- Computer Science
- Domain Expertise



Data Mining

Overlap of machine learning, statistics, AI, databases but with more stress on

- Scalability
- Algorithms and architectures
- Automation for handling large data



Programming Platforms

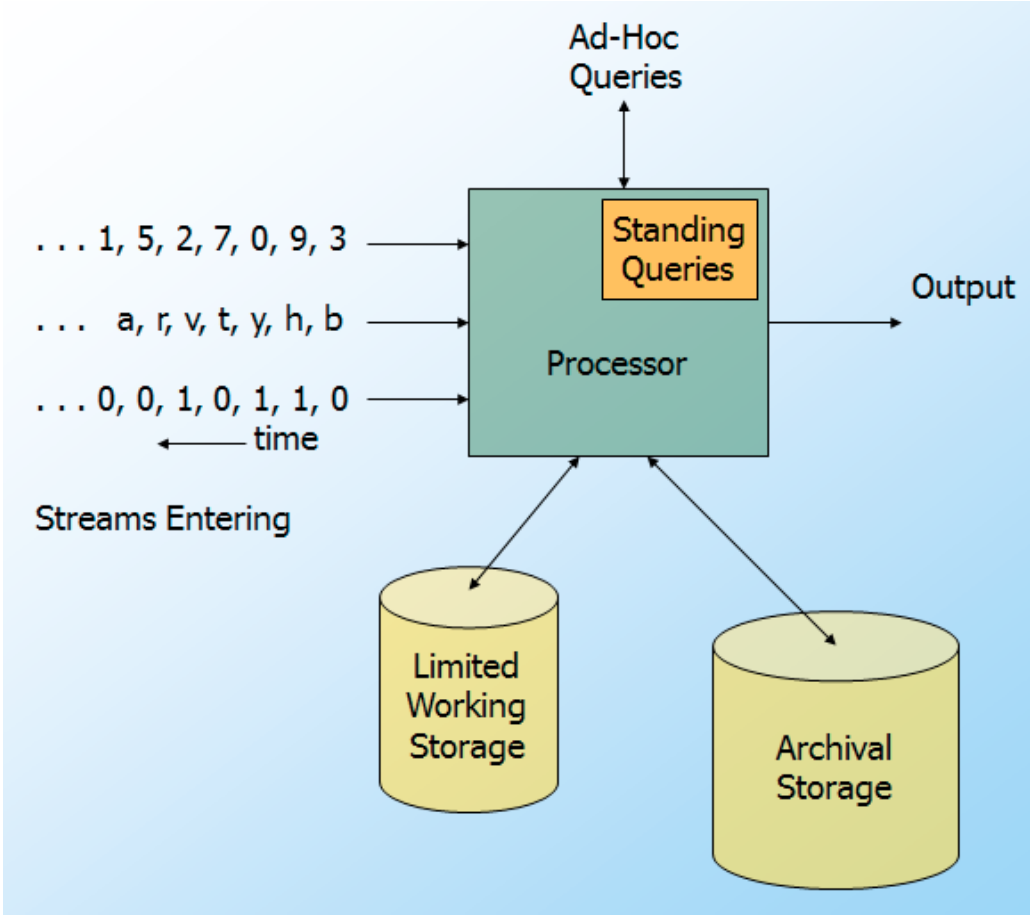
- **Popular Programming Abstractions:**
  - Hadoop MapReduce
  - Pipelines: Hadoop is not an island; requires orchestration between many diverse technologies
  - Graphs: Pregel
  - Iterations: many analysis tasks involve iterations (Vectorwise)
  - DISC: Data Intensive Scalable Computing

Big Data Processing

- Data size is large enough that cannot be processed in a single node
- What do we mean by “processed”?
  - **CRUD:** Create Read Update Destroy + potentially huge amount of actual processing
  - Big data examples come from machine generated domain, e.g., web crawling or tracking, real-time sensor data, logfiles, etc.
- So for Big Data, CRUD ➔ Crud or perhaps **CRAP** (Create Read Analytical Processing)

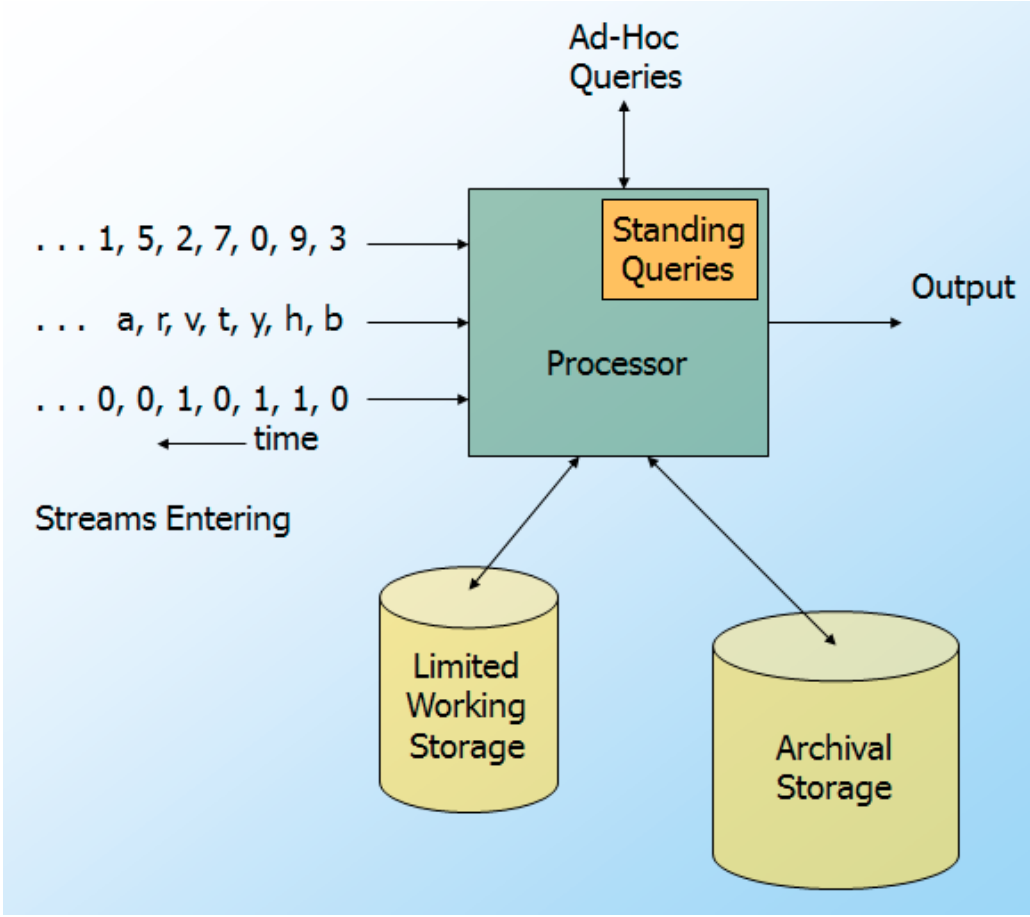
Data Streams

- Data Is Infinite/ Never-Ending
- In SQL, input is under the control of the programmer



Data Streams

- Stream Management is important if input rate is controlled externally
- The system cannot store the entire stream in time



Applications of Streams Processing

- Mining query streams
  - Example: What queries are more frequent today than yesterday?
- Mining Click Streams
  - Yahoo! Wants to know which of its pages are getting an unusual number of hits in the past hour?
- Sensors need monitoring, especially when there are many sensors of the same type, feeding into a central controller!

Applications of Streams Processing

- Telephone call records are summarized into customer bills
- IP packets can be monitored at a switch and used for optimal routing and to detect denial-of-service attacks!
- Sliding Windows: where queries are about a window of length (N); most recent received elements

Different Types of Data

- Data is high dimensional
- Data is a graph
- Data is infinite/never ending
- Data is labeled

Different Models of Computation

- MapReduce
- Streams and online algorithms
- In-memory processing

Data Analytics - Prediction

- Prediction (explaining specific attribute of the data in terms of other attributes):
- Classification (predict discrete value) and regression (estimate numeric value)
- Rule-based, case-based, and model-based learning

Data Analytics - Modeling

- Modeling (describing the relationships between many attributes and many entities):
  - Representation and heuristic search
  - Clustering (modeling group structure of data)
  - Bayesian networks (modeling probabilistic relationships)

Data Analytics - Detection

- Detection (identifying relevant patterns in massive/complex datasets):
  - Anomaly Detection (detecting outliers, novelties, etc.)
  - Pattern Detection (e.g., event surveillance, anomalous patterns)
  - Applications to biosurveillance, crime prevention, etc

Models vs. Analytic Processing

- To a DBMS person, data mining is an extreme form of analytic processing – queries that examine large amounts of data. Result is the answer of the query.
  - Given a billion numbers, a DBMS person would compute their average and standard deviation.
- To a statistician, data mining is the inference of models. Result is the parameters of the model.
  - Given a billion numbers, a statistician might fit the billion to the best Gaussian distribution and report the mean and standard deviation of the distribution

Example: Google's Infrastructure

- 200+ processing
- 200+ terabyte database
- 10^{10} total clock cycles
- 5¢ average advertising revenue

Example: Google's Infrastructure

- ~3 million processors in 200+ processing
- X86 processors, IDE disks, Ethernet communications.
  - Reliability is through redundancy and software management.
- **Partitioned Workload**:
  - Data: Web Pages, indices distributed across processors
  - Function: crawling, index generation, index search, documents retrieval, Ad placement

DISC

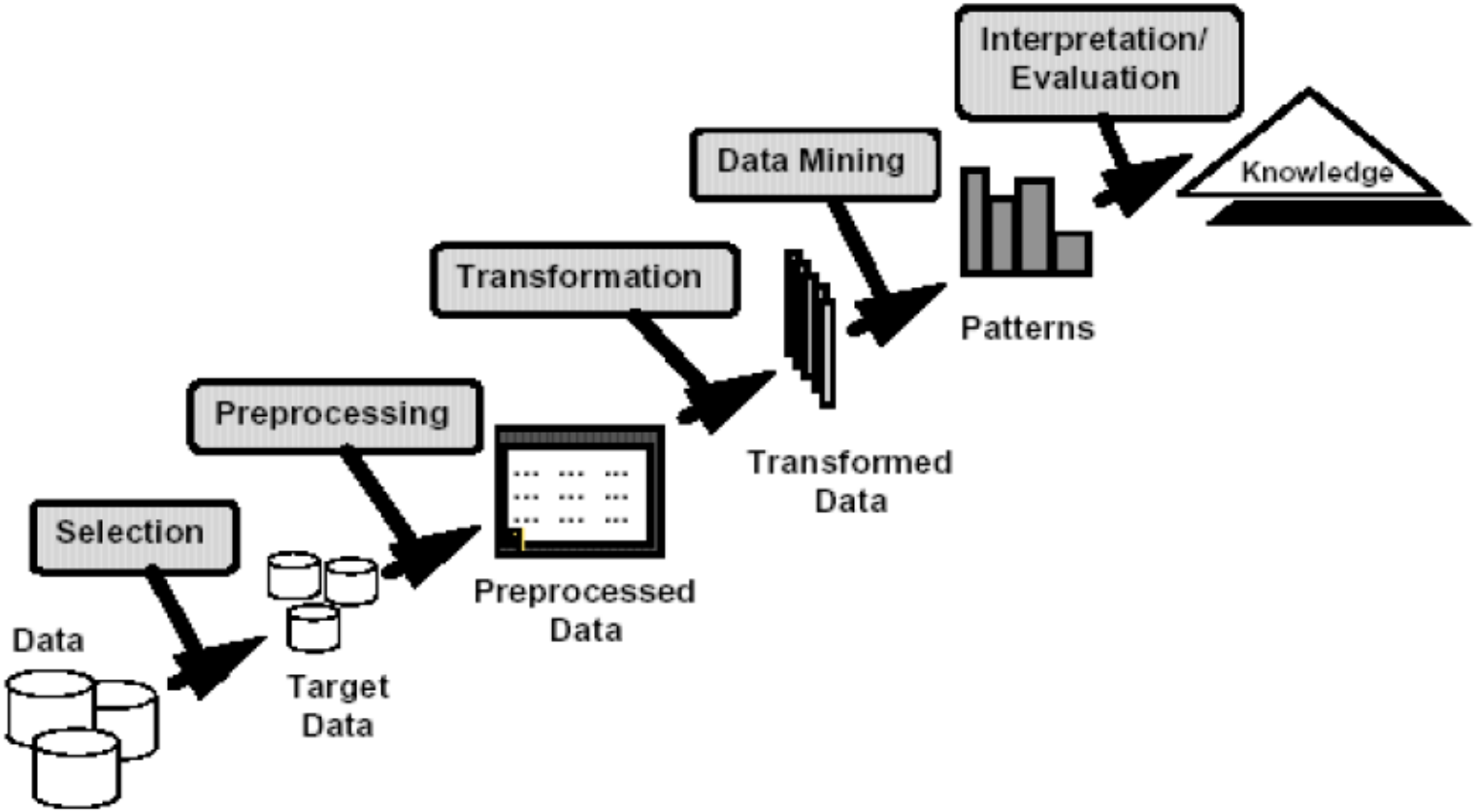
- **Data-Intensive Scalable Computer (DISC):**
  - Large-scale computer centered around data: collecting, maintaining, indexing, computing
  - Similar systems at Yahoo! and Microsoft
- **DISC Environment:**
  - **Architecture**:
    - Cloud computing
  - **Operating System**:
    - Hadoop
  - **Programming Model**:
    - MapReduce

DISC Challenge

- **Computation accesses 1 TB in 5 minutes**:
  - Data distributed over 100+ disks
  - Compute using 100+ processors
  - Connected by 1 Gbps Ethernet
- **System Requirements**
  - Lots of disks
  - Lots of processors
  - Located in close proximity
- **Interactive Access**
- **Robust Fault Tolerance**

Data Mining Role in Big Data

- Non-trivial discovery of implicit, previously unknown, and useful knowledge from massive data



Data Mining Functions in Big Data

- Association rule discovery
- Classification
- Clustering
- Recommendation systems: Collaborative filtering
- Link analysis and graph mining

Data Mining Tasks in Big Data

- Managing Web advertisements
- Descriptive Methods: find human-interpretable patterns that describe the data
- Predictive Methods: Use some variables to predict unknown or future values of other variable

Real-world Problems

- **Real-world Problems to be addressed:**
  - Recommendation systems
  - Association rules
  - Link analysis
  - Duplicate detection (Dedup)

Various Tools

- Linear algebra (e.g., Singular Value Decomposition (SVD))
- Optimization (e.g., stochastic gradient descent)
- Dynamic programming (e.g., frequent itemsets)
- Hashing (e.g., Locality Sensitive Hashing (LSH), Bloom filters)

Why Big Data Matters

Why Big Data Matters?

- **Big Data Enable new things!**
  - Google – 1st big success of Big Data
  - Social Network (Facebook, Twitter, LinkedIn, etc.)
- **Location analytics**
- **Healthcare**
  - Personalized medicine
- **Semantics and AI!**

Big Data Bubble

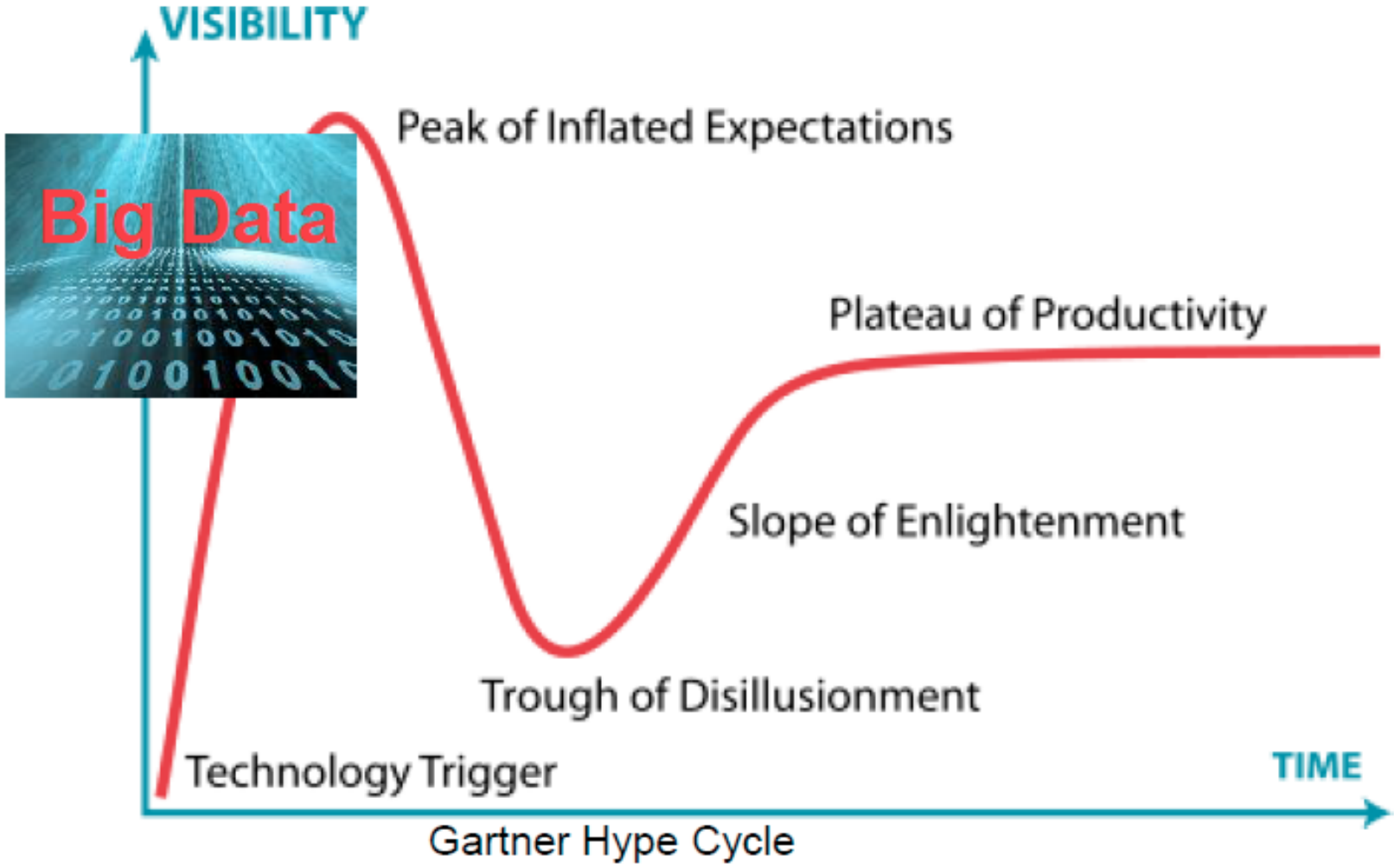
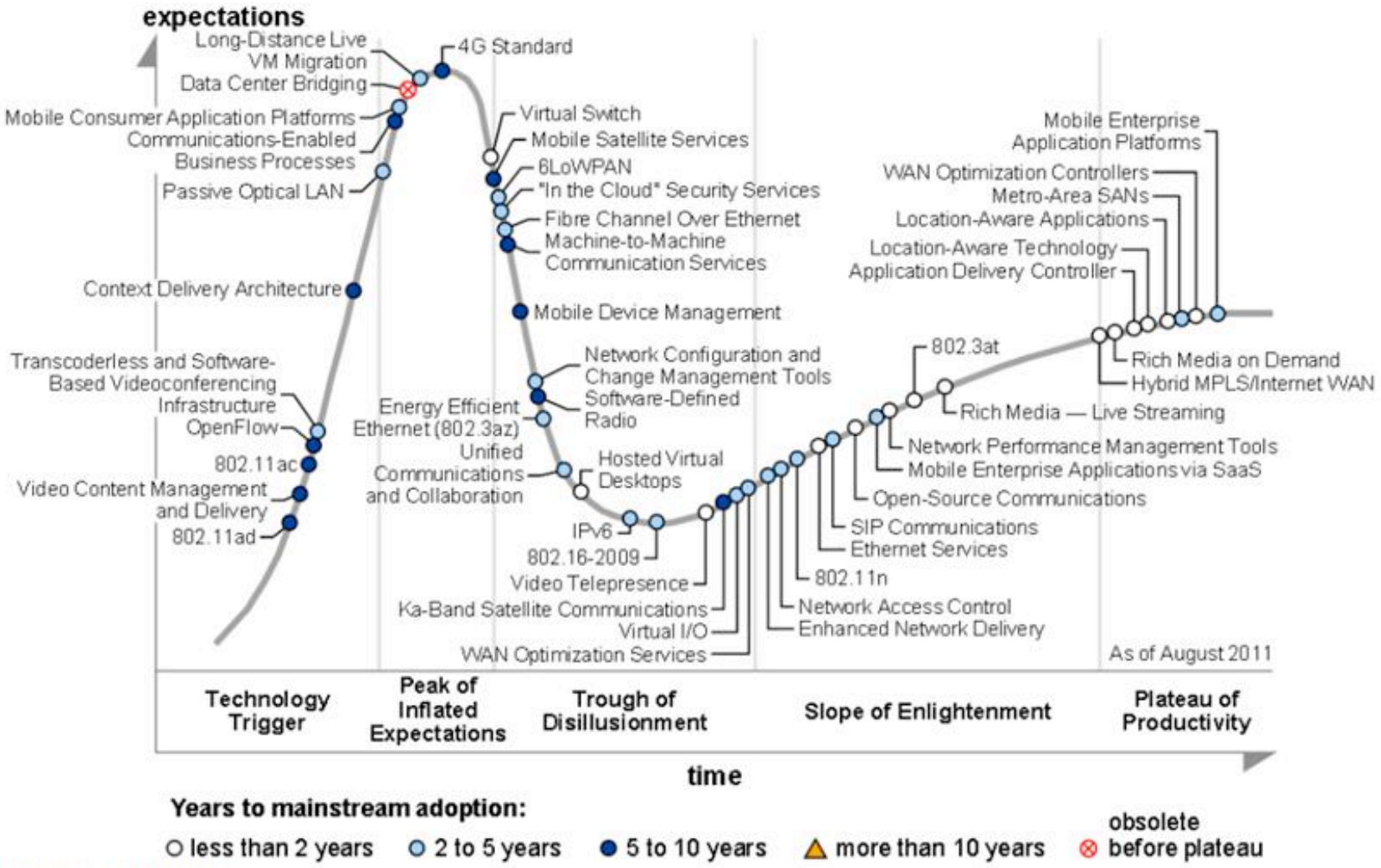
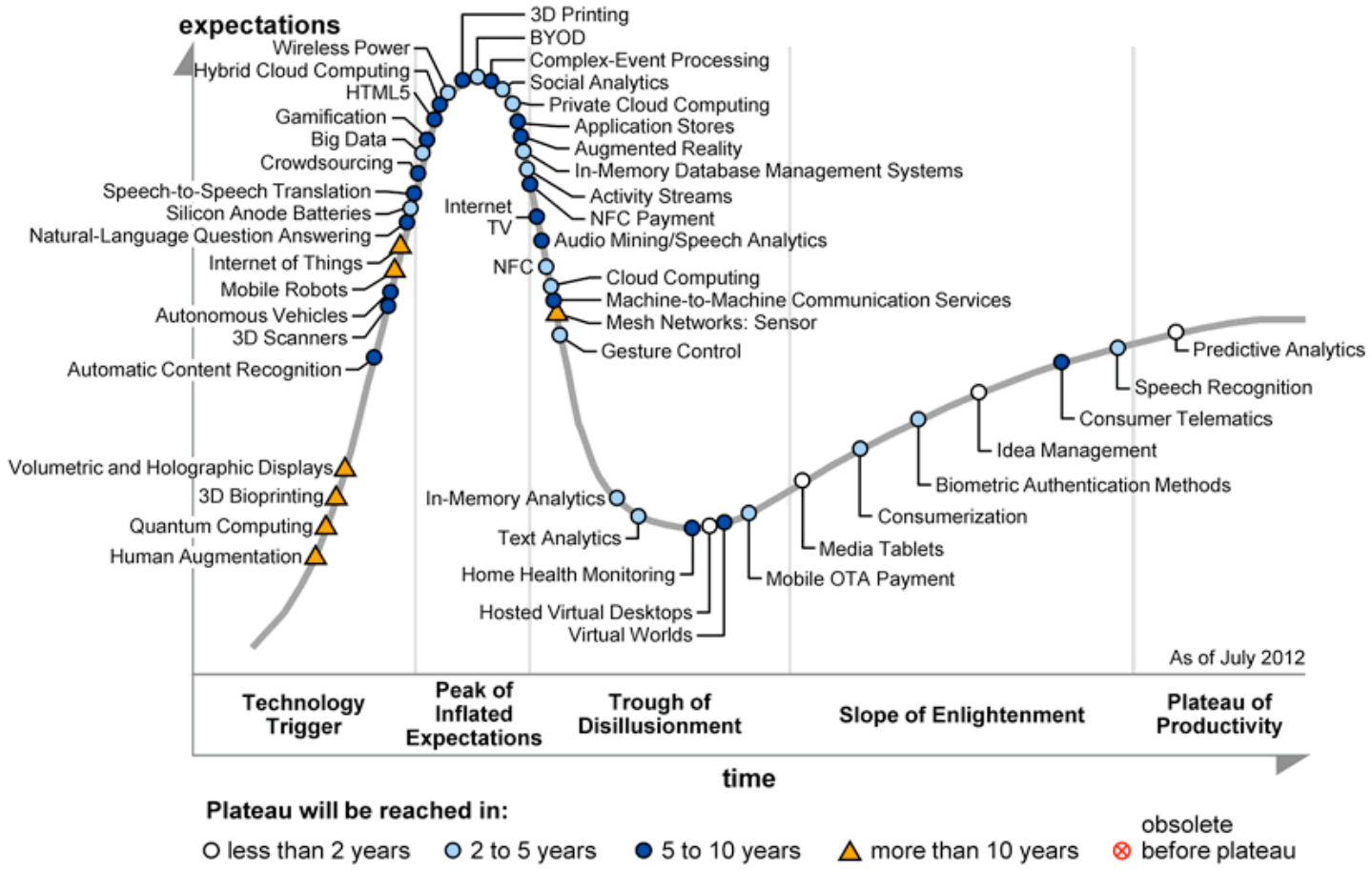
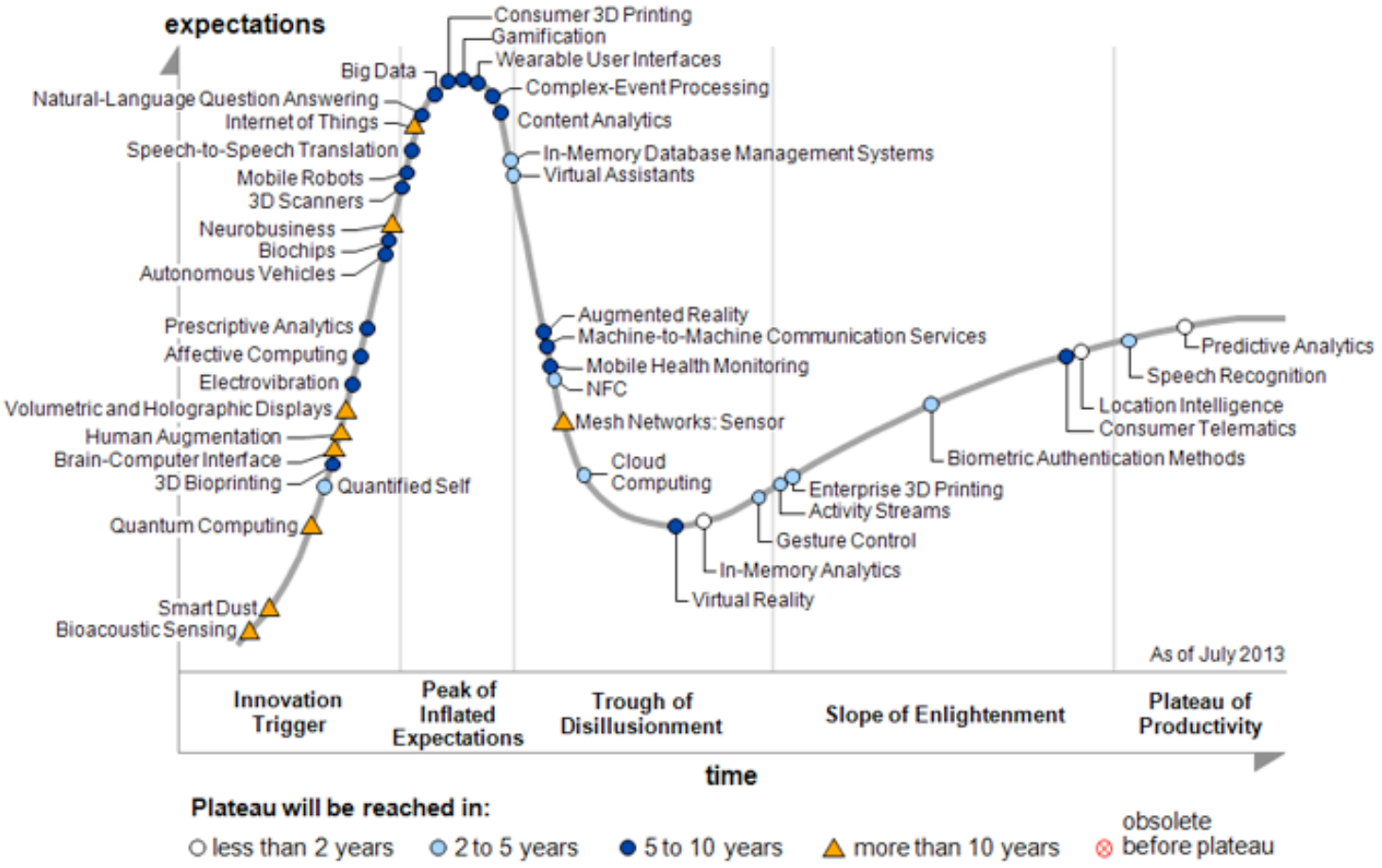


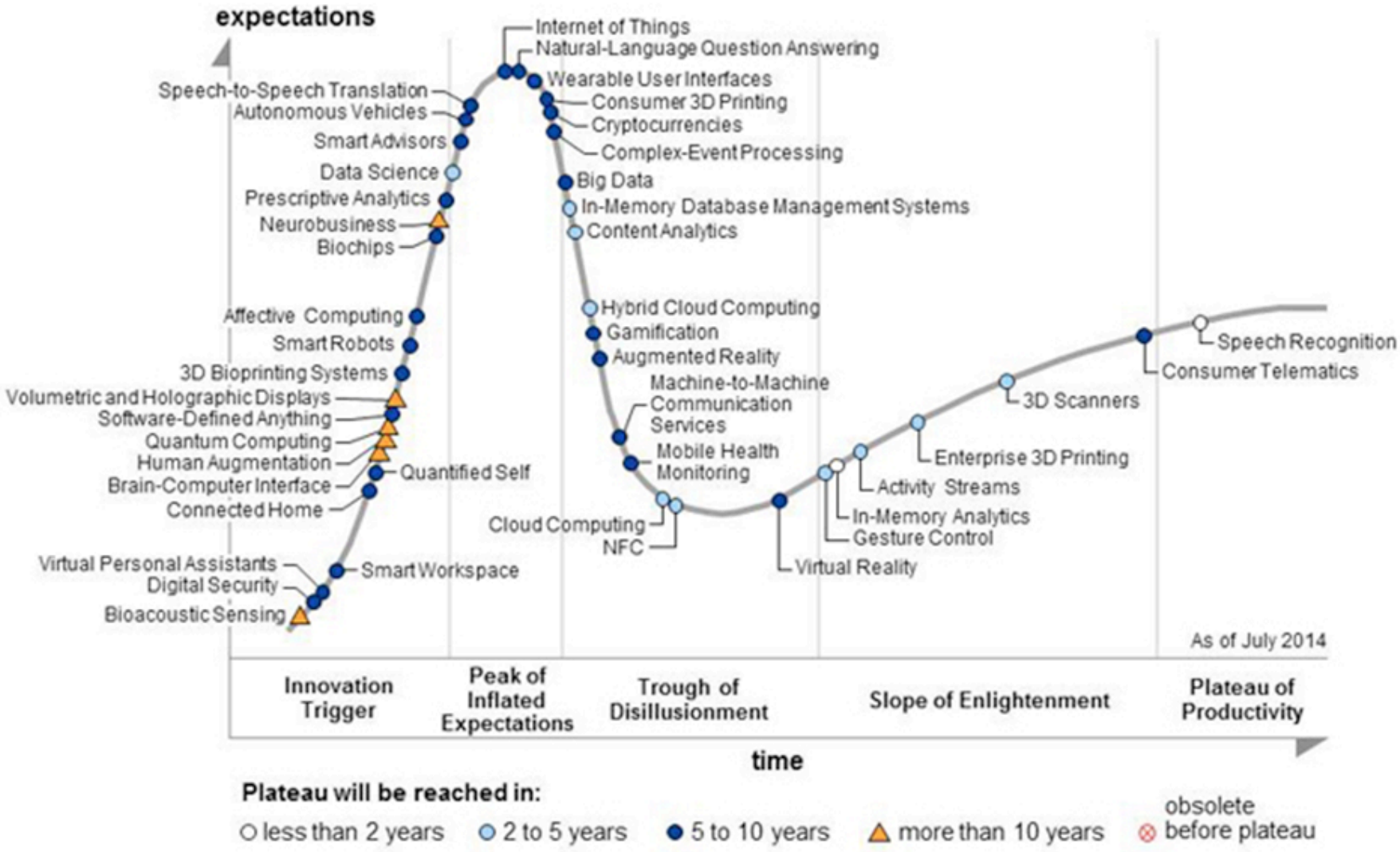
Figure 1. Hype Cycle for Networking and Communications, 2011



Source: Gartner (August 2011)


expectations
Wireless Power
Hybrid Cloud Computing
HTML5
Gamification
Big Data
Crowdsourcing
Speech-to-Speech Translation
Silicon Anode Batteries
Natural-Language Question Answering
Internet of Things
Mobile Robots
Autonomous Vehicles
3D Scanners
Automatic Content Recognition
Volumetric and Holographic Displays
3D Bioprinting
Quantum Computing
Human Augmentation
3D Printing
BYOD
Complex-Event Processing
Social Analytics
Private Cloud Computing
Application Stores
Augmented Reality
In-Memory Database Management Systems
Activity Streams
Internet TV
NFC Payment
Audio Mining/Speech Analytics
NFC
Cloud Computing
Machine-to-Machine Communication Services
Mesh Networks: Sensor
Gesture Control
In-Memory Analytics
Text Analytics
Home Health Monitoring
Hosted Virtual Desktops
Virtual Worlds
Media Tablets
Mobile OTA Payment
Consumerization
Biometric Authentication Methods
Idea Management
Consumer Telematics
Speech Recognition
Predictive Analytics
As of July 2012
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau


expectations
Consumer 3D Printing
Gamification
Wearable User Interfaces
Big Data
Complex-Event Processing
Natural-Language Question Answering
Content Analytics
Internet of Things
In-Memory Database Management Systems
Speech-to-Speech Translation
Virtual Assistants
Mobile Robots
3D Scanners
Neurobusiness
Biochips
Autonomous Vehicles
Augmented Reality
Prescriptive Analytics
Machine-to-Machine Communication Services
Affective Computing
Mobile Health Monitoring
Electrovibration
NFC
Volumetric and Holographic Displays
Mesh Networks: Sensor
Human Augmentation
Brain-Computer Interface
Cloud Computing
3D Bioprinting
Quantified Self
Enterprise 3D Printing
Biometric Authentication Methods
Activity Streams
Quantum Computing
Gesture Control
Predictive Analytics
Speech Recognition
In-Memory Analytics
Location Intelligence
Consumer Telematics
Virtual Reality
Smart Dust
Bioacoustic Sensing
As of July 2013
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau


expectations
Internet of Things
Natural-Language Question Answering
Speech-to-Speech Translation
Autonomous Vehicles
Wearable User Interfaces
Consumer 3D Printing
Cryptocurrencies
Smart Advisors
Complex-Event Processing
Data Science
Big Data
Prescriptive Analytics
In-Memory Database Management Systems
Neurobusiness
Content Analytics
Biochips
Affective Computing
Hybrid Cloud Computing
Smart Robots
Gamification
Augmented Reality
3D Bioprinting Systems
Machine-to-Machine Communication Services
Volumetric and Holographic Displays
Software-Defined Anything
Quantum Computing
Mobile Health Monitoring
Human Augmentation
Quantified Self
Brain-Computer Interface
Connected Home
Cloud Computing
NFC
Virtual Personal Assistants
Smart Workspace
Digital Security
Bioacoustic Sensing
Virtual Reality
Activity Streams
In-Memory Analytics
Gesture Control
Enterprise 3D Printing
3D Scanners
Speech Recognition
Consumer Telematics
As of July 2014
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau


expectations
Advanced Analytics With Self-Service Delivery
Autonomous Vehicles
Internet of Things
Speech-to-Speech Translation
Machine Learning
Wearables
Cryptocurrencies
Consumer 3D Printing
Natural-Language Question Answering
Smart Advisors
Micro Data Centers
Digital Dexterity
Software-Defined Security
Neurobusiness
Citizen Data Science
Biochips
IoT Platform
Connected Home
Affective Computing
Smart Robots
3D Bioprinting Systems for Organ Transplant
Volumetric Displays
Human Augmentation
Brain-Computer Interface
Quantum Computing
Hybrid Cloud Computing
Augmented Reality
Enterprise 3D Printing
Gesture Control
Virtual Reality
Autonomous Field Vehicles
Cryptocurrency Exchange
Bioacoustic Sensing
People-Literate Technology
Digital Security
Virtual Personal Assistants
Smart Dust
As of July 2015
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

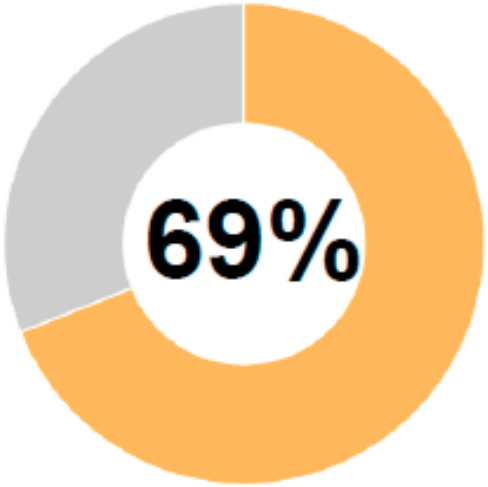
Why Big Data Matters?

- Provides groundbreaking opportunities for enterprise information Management and decision making
- The rate of information growth appears to be exceeding Moore's law
- The amount of data is exploding; companies are capturing and digitizing more information

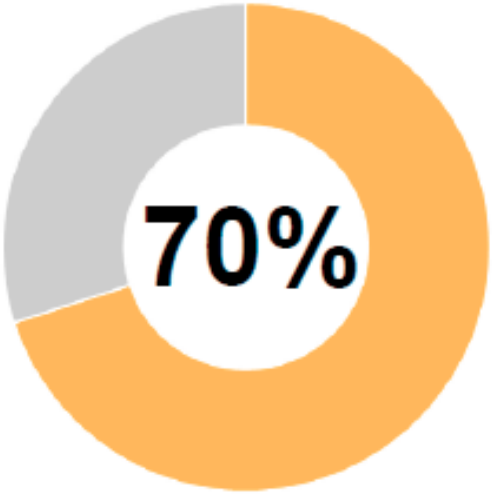
Why Big Data Matters?

- 35 zettabytes (billion TB) of data will be generated and consumed by the end of the decade
- Hadoop (HDFS, HBase, MapReduce) and Memcached are gaining a lot of momentum

Why Big Data Matters?

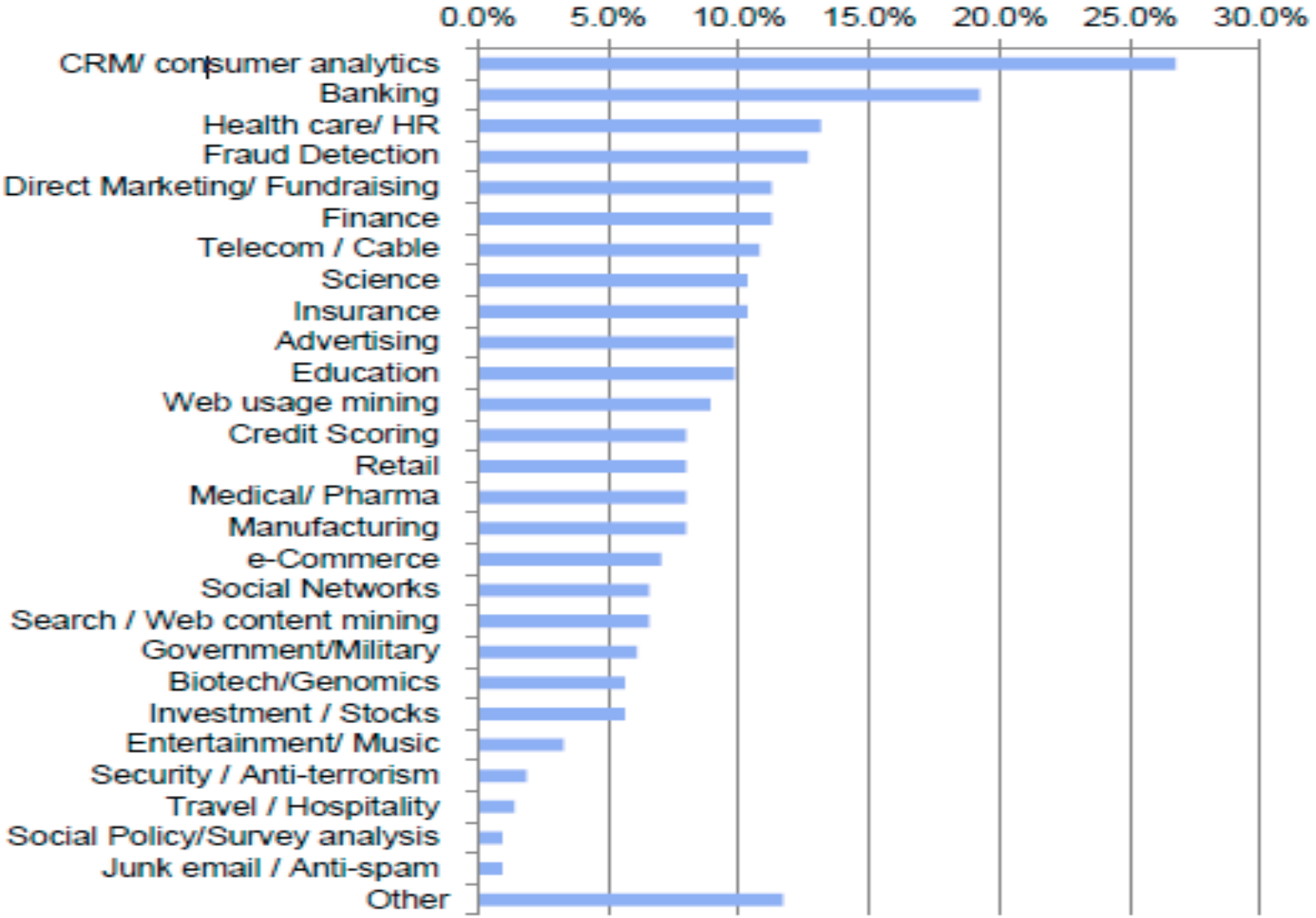


Extremely important for **competitive advantage**

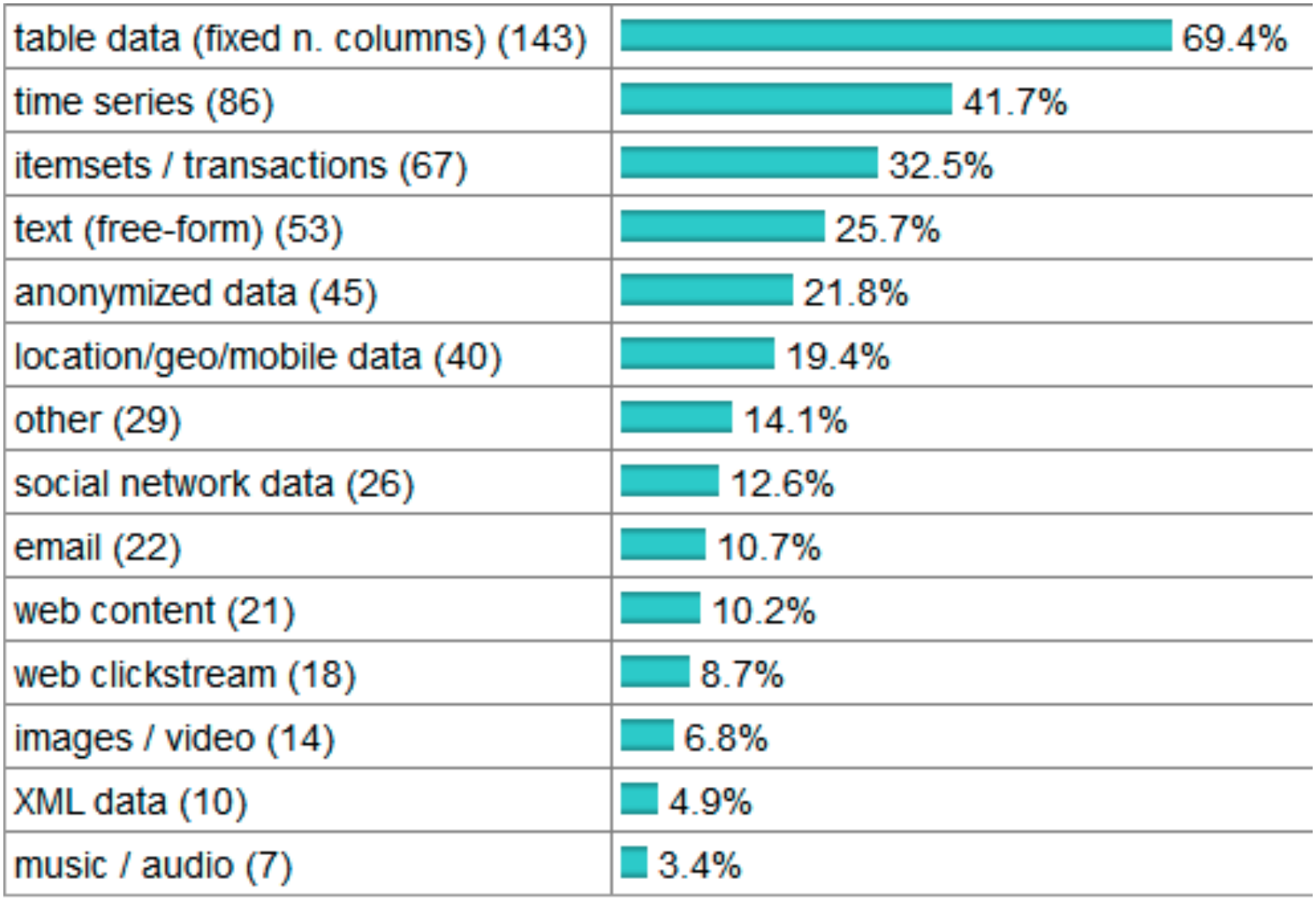


Helped **manage costs or improve operations**

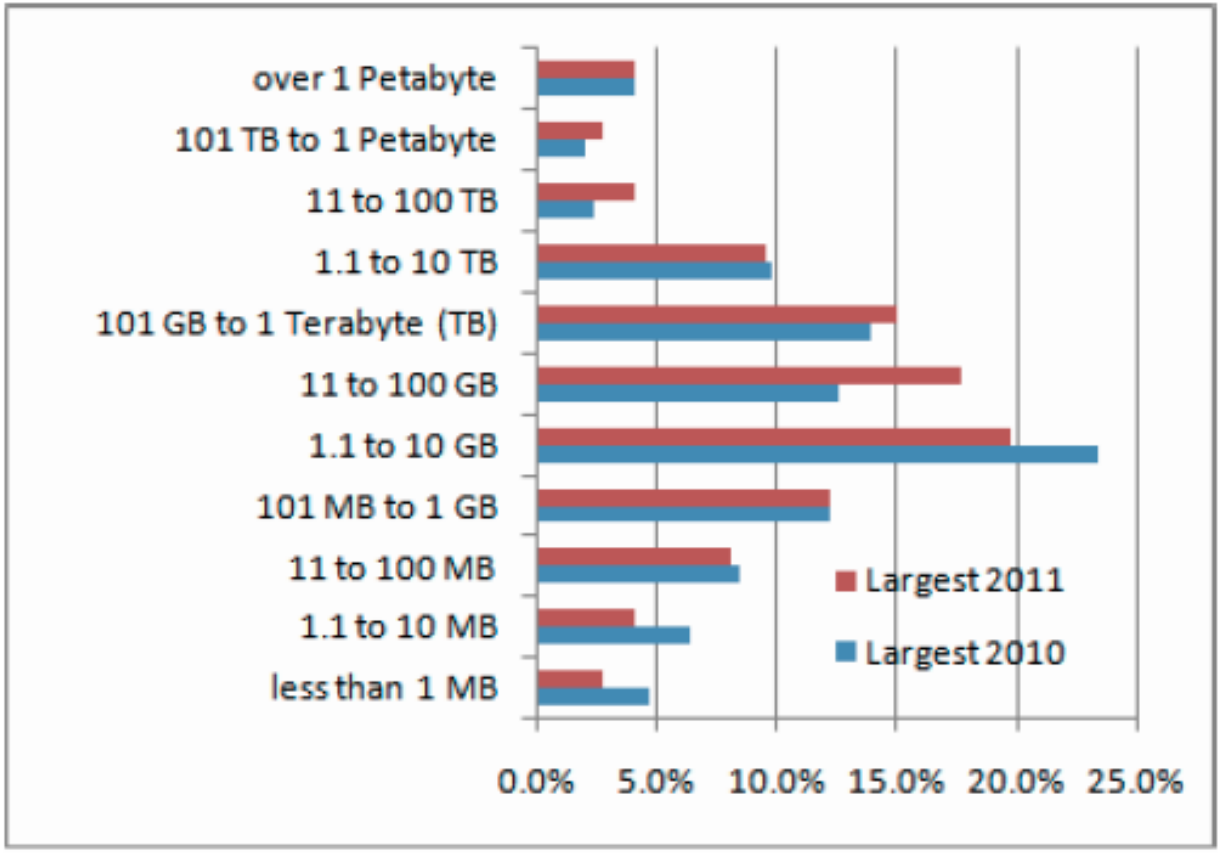
Where Data Mining and Analytics are applied?



Data Types Analyzed/Mined



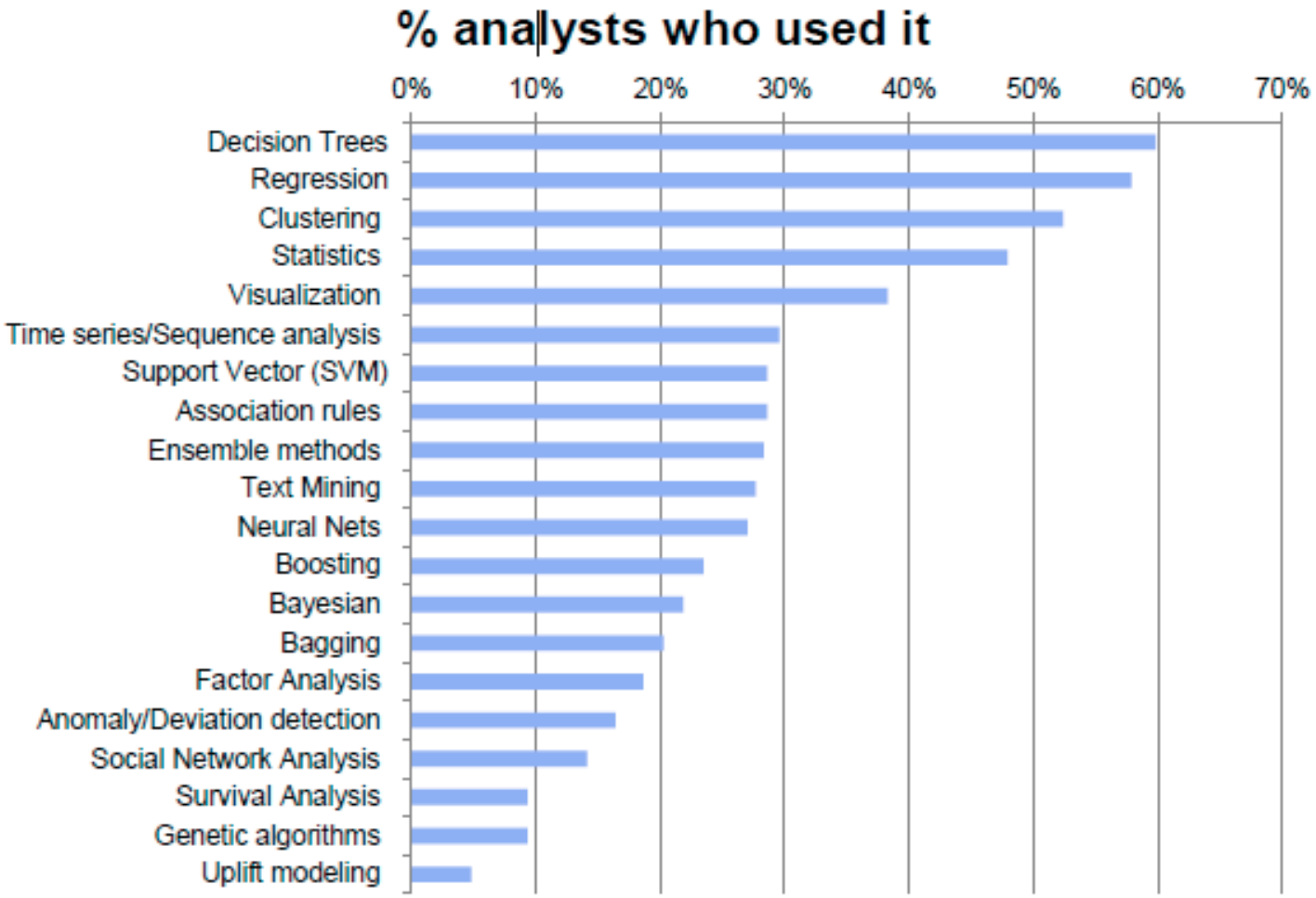
Largest Data Set Analyzed/Mined



2011 median dataset size ~10-20 GB, vs. 8-10 GB in 2010.

Increase in 10 GB to 1 PB range

Algorithms Used for Data Analysis/ Mining



Summary

Summary

- Big Data is best described in terms of 3Vs (Volume, Variety, Velocity)
- Big Data matters as it can give the enterprise an advantage in a competitive market