Chapter1

Fundamentals of Big Data Analytics

Basanta Joshi, PhD

Asst. Prof., Depart of Electronics and Computer Engineering

Program Coordinator, MSc in Information and Communication Engineering

Member, Laboratory for ICT Research and Development (LICT)

Member, Research Management Cell (RMC)

Institute of Engineering

basanta@ioe.edu.np

http://www.basantajoshi.com.np

https://scholar.google.com/citations?user=iocLiGcAAAAJ

https://www.researchgate.net/profile/Basanta_Joshi2

Presenter Profile - Dr. Bhuvan Unhelkar

(BE, MBA, PhD, **FACS, CBAP®, PSM**)

- Professor of IT, USF (Sarasota-Manatee Campus)
- Founder, MethodScience.com & PlatiFi.com
- Courses: Adv Prog. Design (UML), Agile PM, Data & Security Analytics, NoSQL, Sr. Project
- PhD from University of Technology, Sydney (UTS), 1997 "*Effect of Granularity of OO Design in Modelling an Enterprise and its application to Financial Risk Management*"; Guide: Prof. **Brian Henderson-Sellers**
- Author: **25** Books (AI, Mobile, Agile, Green ICT, and Big Data Strategies for Agile Business)
- Supervisor: **8** PhD Completions;
- **Fellow** of the Australian Computer Society; **IEEE Sr. member,** Life Member, Computer Society of India & BMA
- Hon. Prof. Amity Univ. (India), Western Sydney Univ. (Australia)
- Past President – Rotary club of Sarasota Sunrise; Rotary Club of St Ives, Paul Harris Fellow; AG); TiE;
- www.unhelkar.com & www.methodscience.com

- I will be teaching in Collaboration with Prof. Bhuvan


What is Big Data?

why is there such *buzz* around BIG DATA?

Isn’t *Big Data* just big hype?

… How big is BIG?

What is it that we are really talking about?

“Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.”

-- Forrester

"Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers."

-- Forrester

“Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”

-- Gartner

“Big Data in general is defined as high volume, velocity and variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.”

-- Gartner

“Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.”

-- O’Reilly

"Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it."

-- O'Reilly

“Big data is the data characterized by 3 attributes: volume, variety and velocity.”

-- IBM

RANDOM WORDS

"Big data is the data characterized by 3 attributes: volume, variety and velocity."

-- IBM

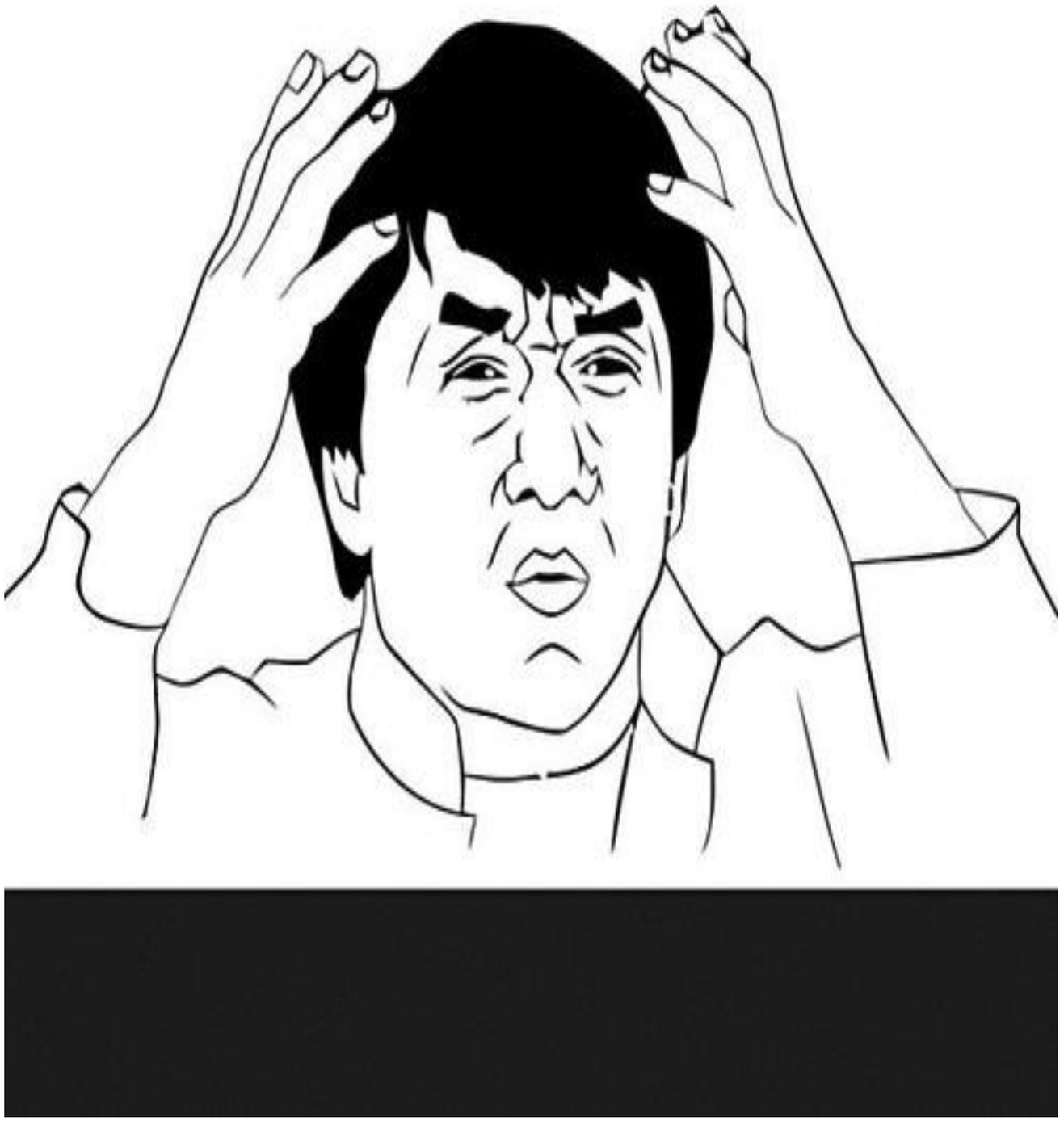
“Big data is the data characterized by 4 key attributes: volume, variety, velocity and

value.”

-- Oracle

"Big data is the data characterized by 4 key attributes: volume, variety, velocity and value."

-- Oracle



Let’s look at

Big Data

in a different way.

What was your first computer?


64
CTRL
SHIFT
RESTORE
RETURN
CLR HOME
INST DEL

What was its “Big Data” limit?

DevOps Borat

@DEVOPS_BORAT

Small Data is when is fit in RAM.
Big Data is when is crash because is not fit in RAM.

2/6/13, 8:22 AM

Let’s try again…

Byte : one grain of rice

Byte : one grain of rice

Kilobyte : cup of rice

Kilobyte

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Megabyte

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Gigabyte

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Terabyte

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Petabyte

Byte : one grain of rice

Kilobyte : cup of rice

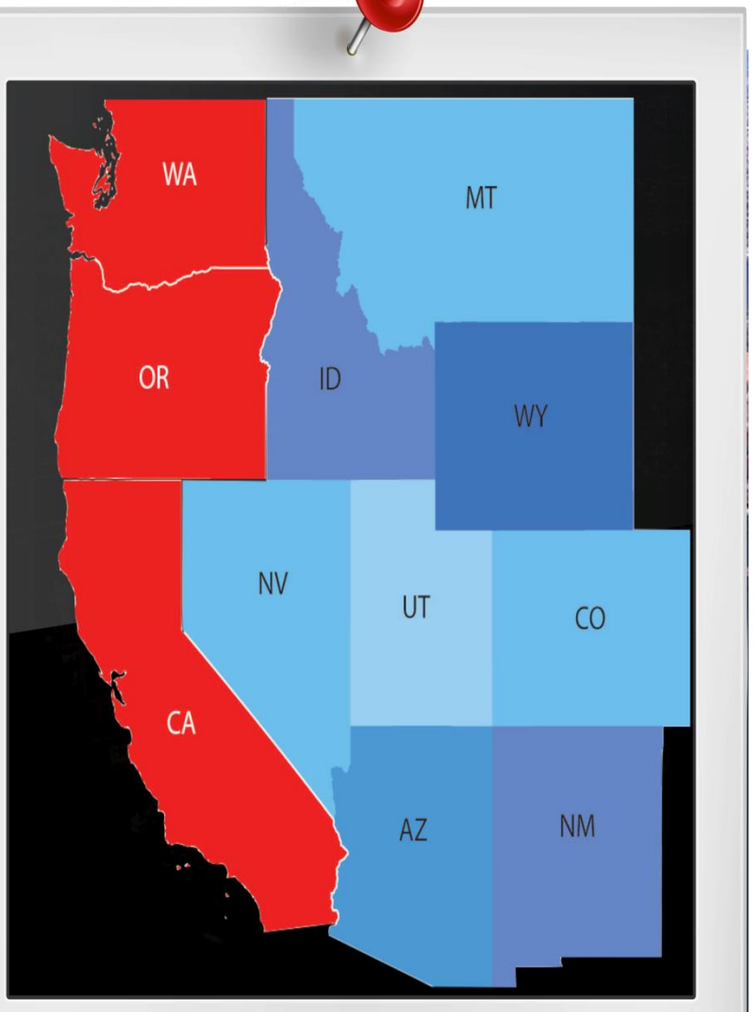
Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Blankets west coast states



Exabyte

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Zettabyte

Byte : one grain of rice
Kilobyte : cup of rice
Megabyte : 8 bags of rice
Gigabyte : 3 Semi trucks
Terabyte : 2 Container Ships
Petabyte : Blankets Manhattan
Exabyte : Blankets west coast states
Zettabyte : Fills the Pacific Ocean
Yottabyte : A EARTH SIZE RICE BALL!

Yottabyte

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Yottabyte : A EARTH SIZE RICE BALL!

Byte : one grain of rice

Kilobyte : cup of rice

Hobbyist

Megabyte : 8 bags of rice

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Yottabyte : A EARTH SIZE RICE BALL!

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Hobbyist

Desktop

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Yottabyte : A EARTH SIZE RICE BALL!

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

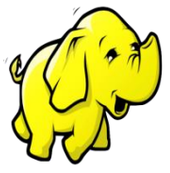
Petabyte : Blankets Manhattan

Yottabyte : A EARTH SIZE RICE BALL!

Hobbyist

Desktop

Internet



Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

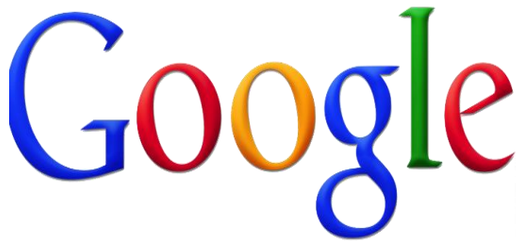
Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Pacific Ocean

Yottabyte : A EARTH SIZE RICE BALL!



Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

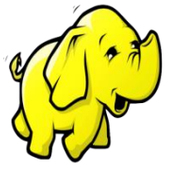
Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Hobbyist

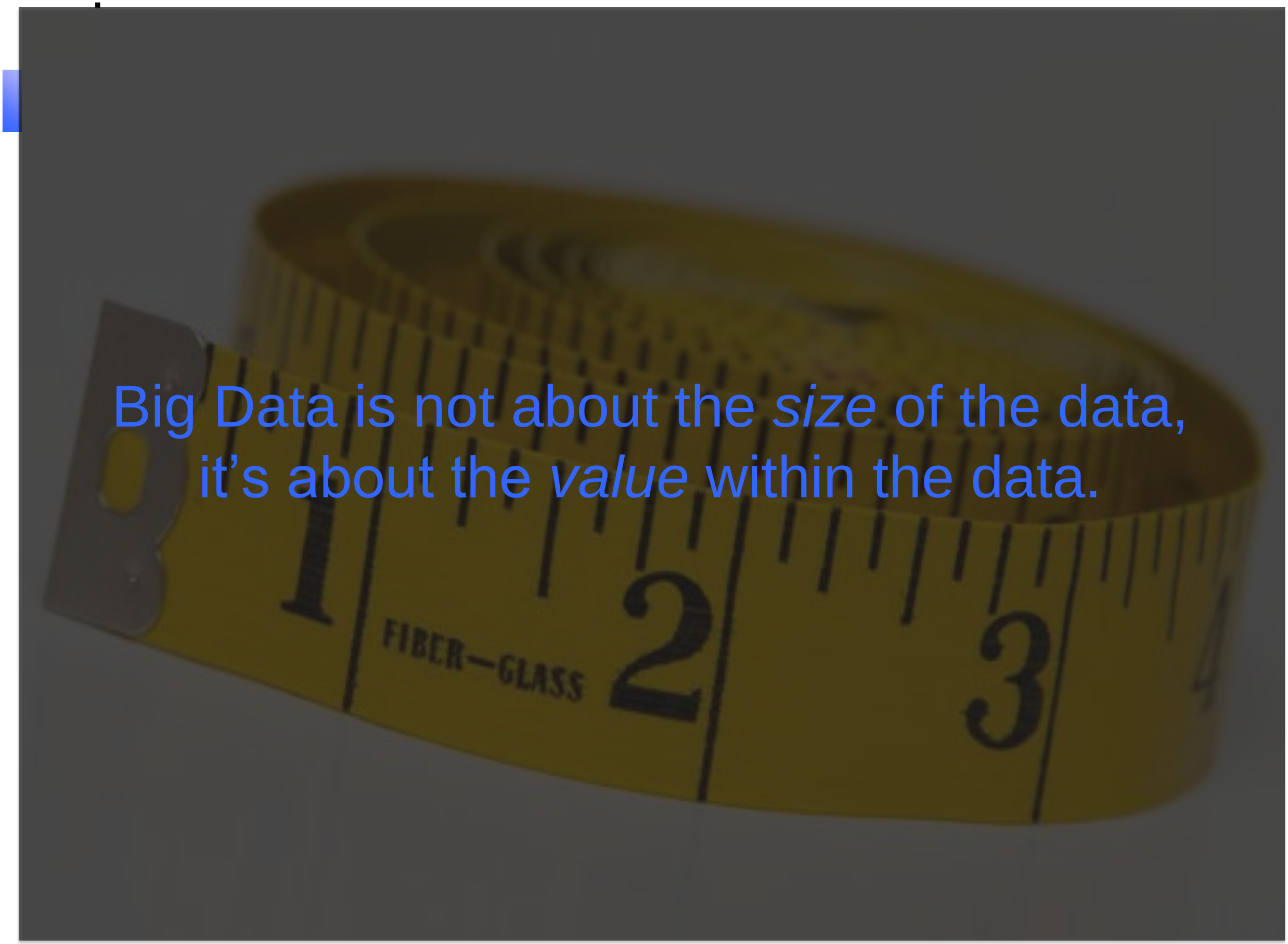
Desktop

Internet



Big Data

“Utah Governor Herbert on June 11, 2012 told the annual meeting of the National Governors' Association that the NSA's 1,500,000 square foot data center being built outside Salt Lake City will be the first facility to house a yottabyte of data.”



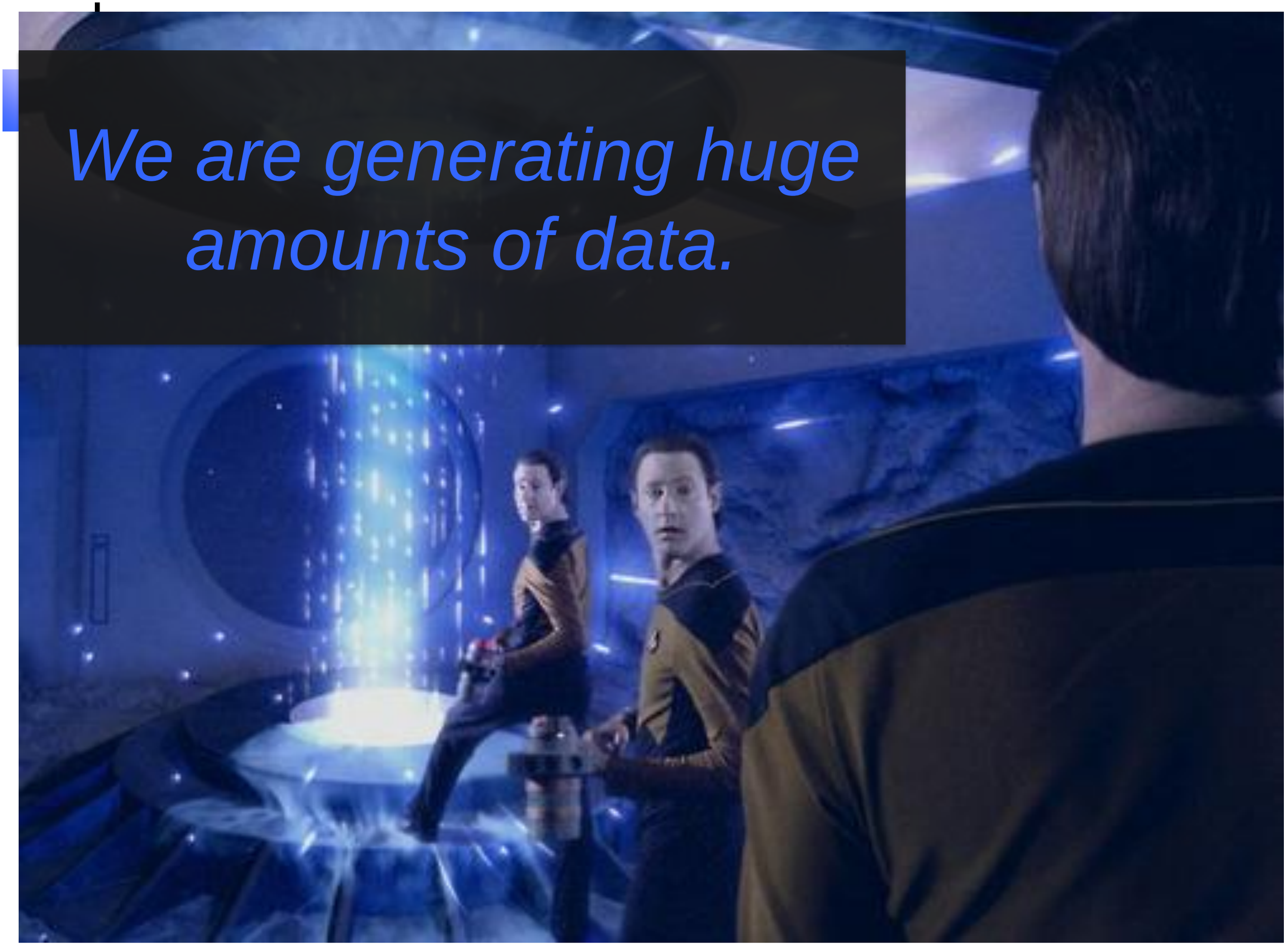
Big Data is not about the *size* of the data, it's about the *value* within the data.

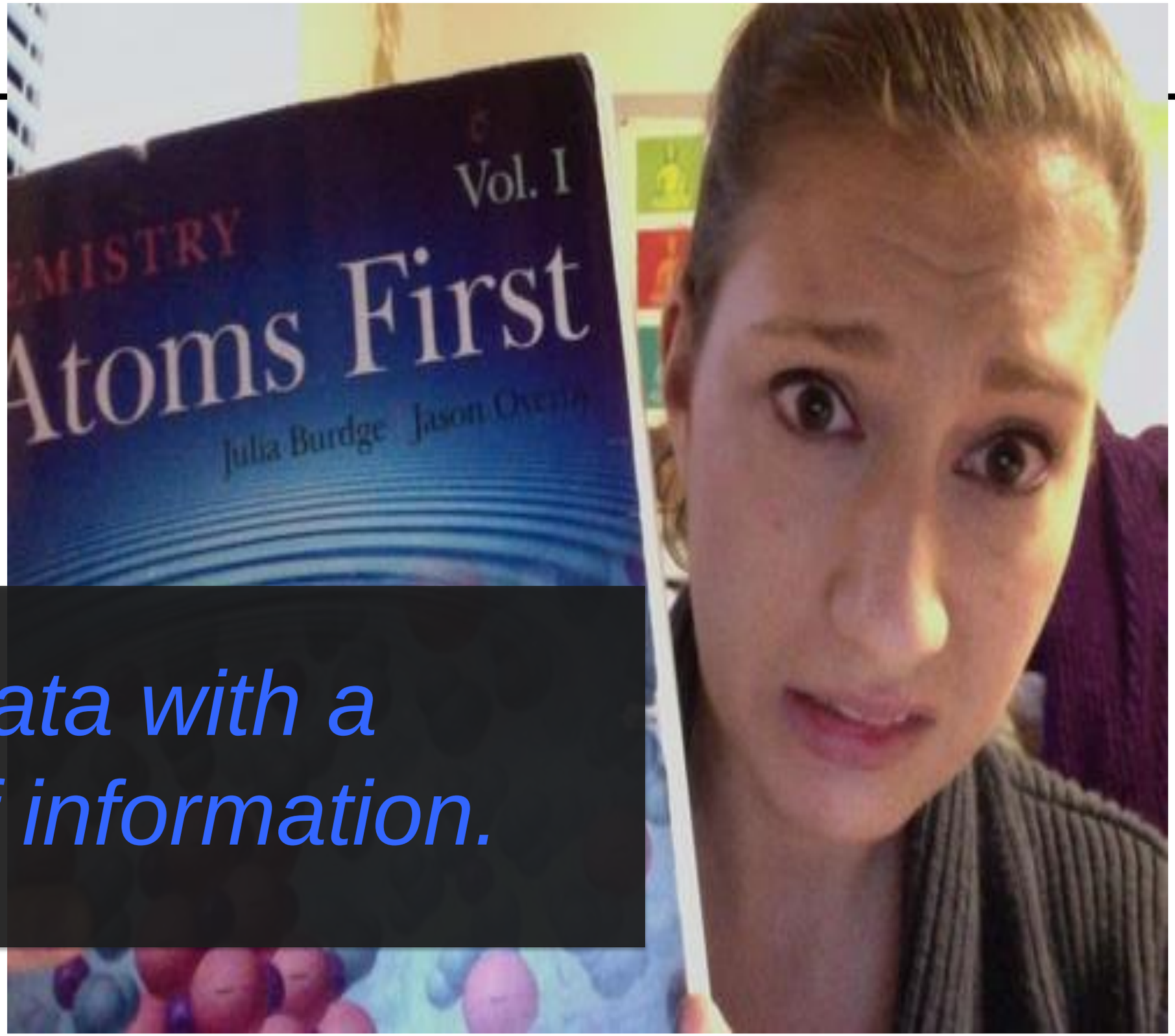

So, what is
value?

Market Intelligence?
Business Intelligence?
Secret Intelligence?

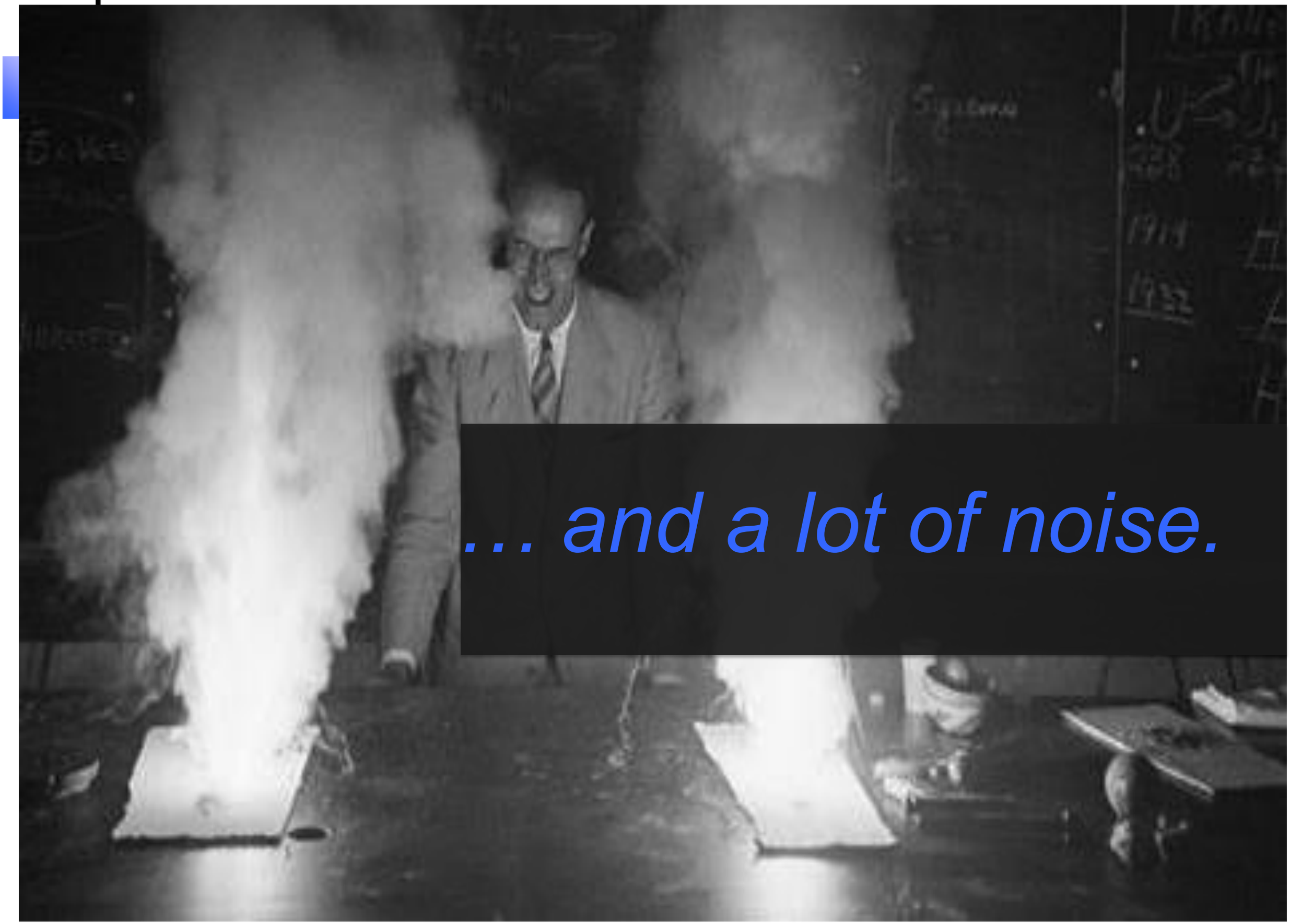

Our society is leaving
behind a digital footprint.


People are living online and we all are expressing our attitudes, likes and dislikes, our opinions and perspectives.


We are generating huge amounts of data.



Data with a lot of information.


… and a lot of noise.

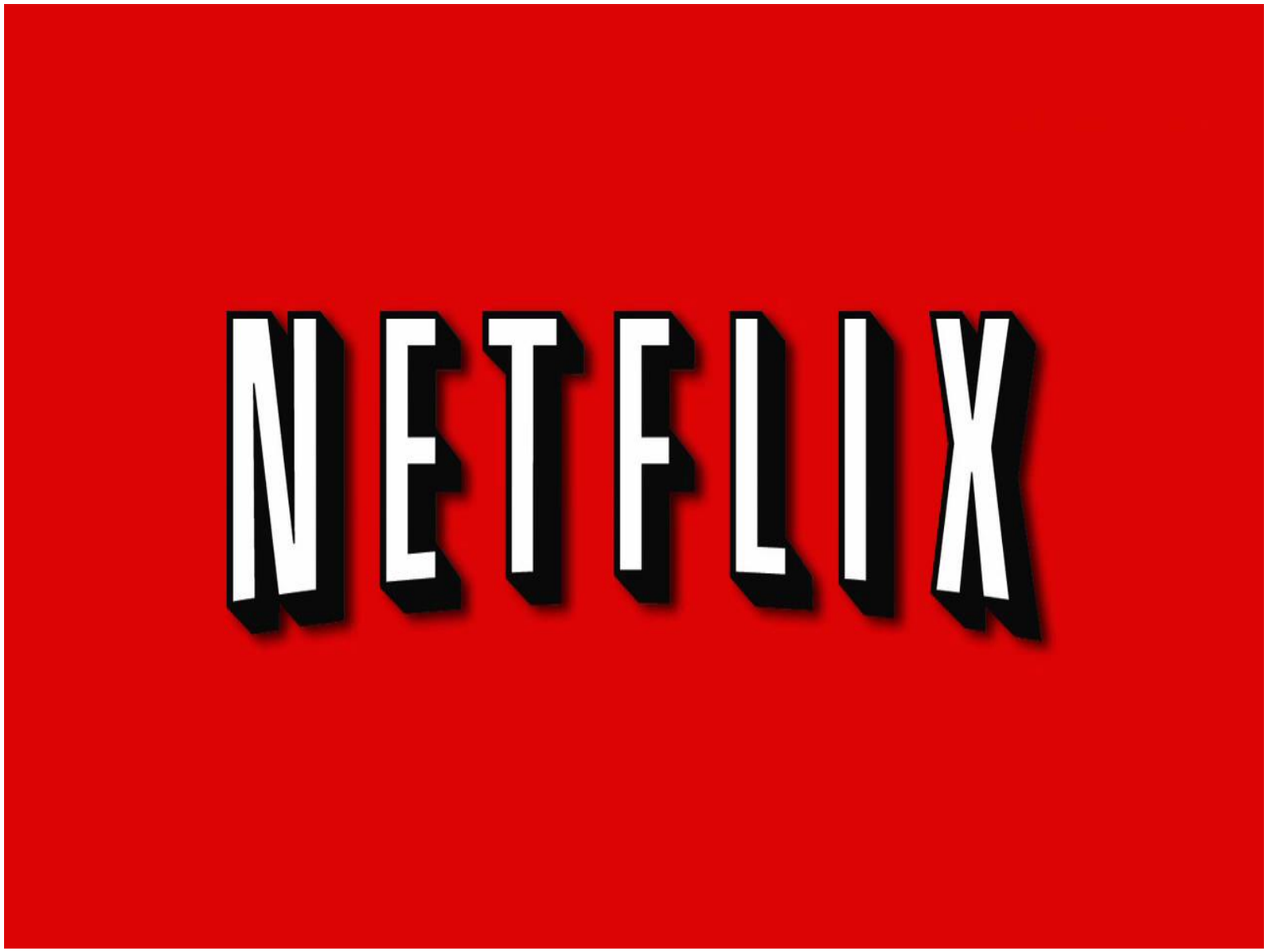

The ability to hear the signal
from the noise is the key…


to unlocking the human conversation
that is taking place around us.


Listening to this
conversation
can lead us to…

Unexpected Discoveries

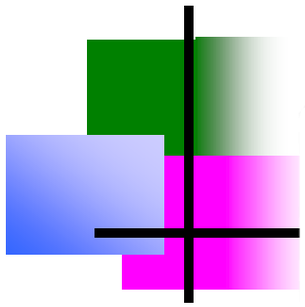
Case Study


NETFLIX


A NETFLIX ORIGINAL SERIES
HOUSE of CARDS

"House of Cards" is one of the first major test cases of this Big Data-driven creative strategy. For almost a year, Netflix executives have told us that their detailed knowledge of Netflix subscriber viewing preferences clinched their decision to license a remake of the popular and critically well regarded 1990 BBC miniseries. Netflix's data indicated that the same subscribers who loved the original BBC production also gobbled down movies starring Kevin Spacey or directed by David Fincher. Therefore, concluded Netflix executives, a remake of the BBC drama with Spacey and Fincher attached was a no-brainer, to the point that the company committed $100 million for two 13-episode seasons.

Did it work?



David Armano
@armano

Just started. So far. Awesome. #houseofcards #GetGlue getglue.com/tv_shows/house...

2/16/13, 7:02 PM

House of Cards

Ruthless Congressman Francis Underwood and his ambitious wife Claire will stop at nothing to ascend the ranks of power. This wicked political drama slithers through the back halls of greed...

The Studio Executive
@studioexec1

Watching #HouseofCards. Kevin Spacey's utterly brilliant as the slimy conniving politician with the quick wit. The man has such range.

2/24/13, 12:15 PM

Beau Willimon
@BeauWillimon

Wow- According to IMDB at least, #HouseofCards "Most Popular TV Show in the World" right now. THANK YOU FANS webpronews.com/house-of-cards... @netflix

2/15/13, 3:35 PM

12 RETWEETS **10** FAVORITES

The Atlantic
@TheAtlantic

The Real History Behind the Politics of #HouseOfCards theatln.tc/Ye557b

2/22/13, 5:00 AM

The Very Real History Behind the Crazy Politics of 'House of Cards'

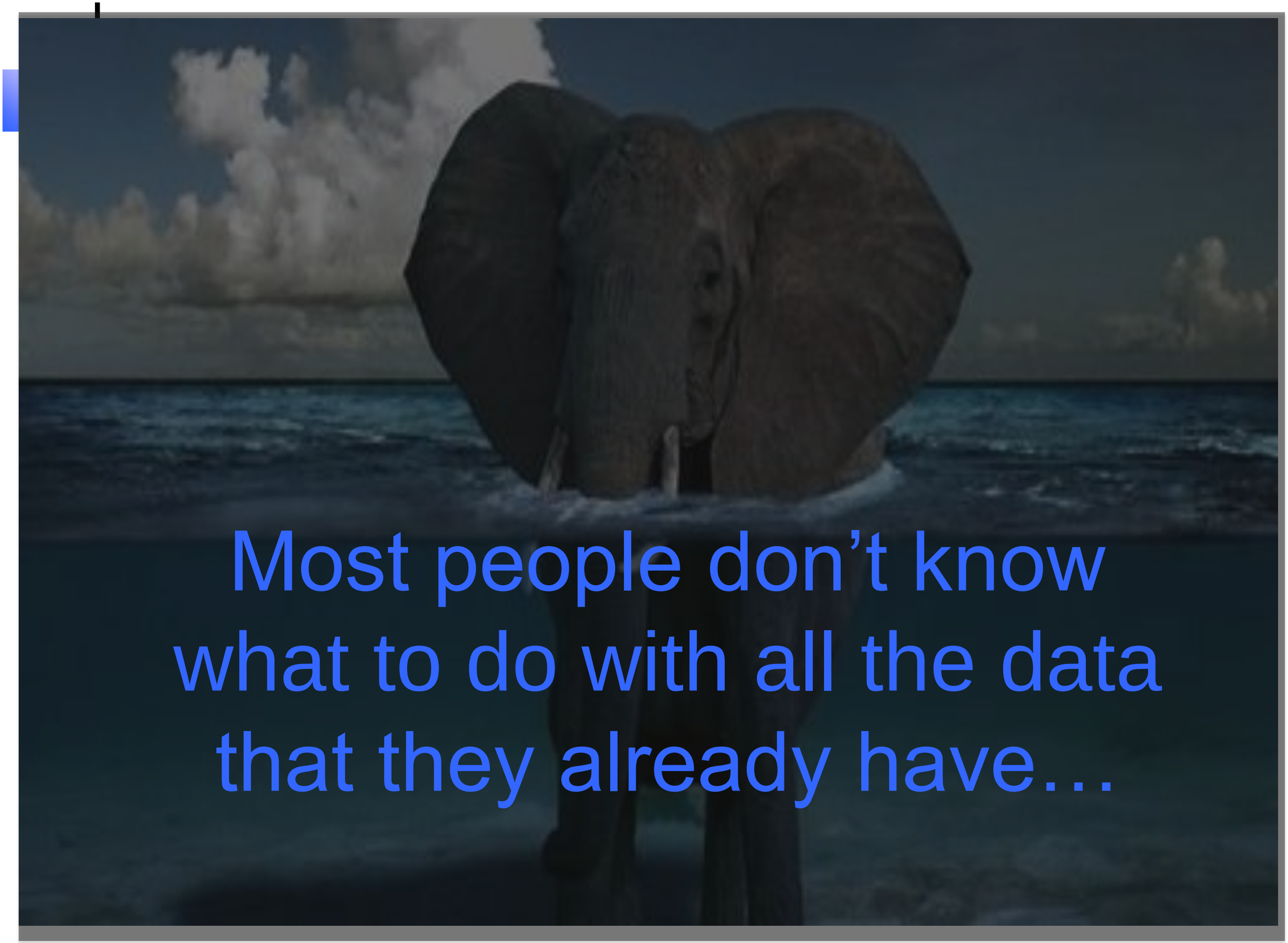
A few of the show's more outlandish moments are uncomfortably similar to real life.

AnikaChapin
@AnikaChapin

I've finished #HouseofCards and I don't know what to do with myself now. Maybe I'll start evilly manipulating people to fill the void.

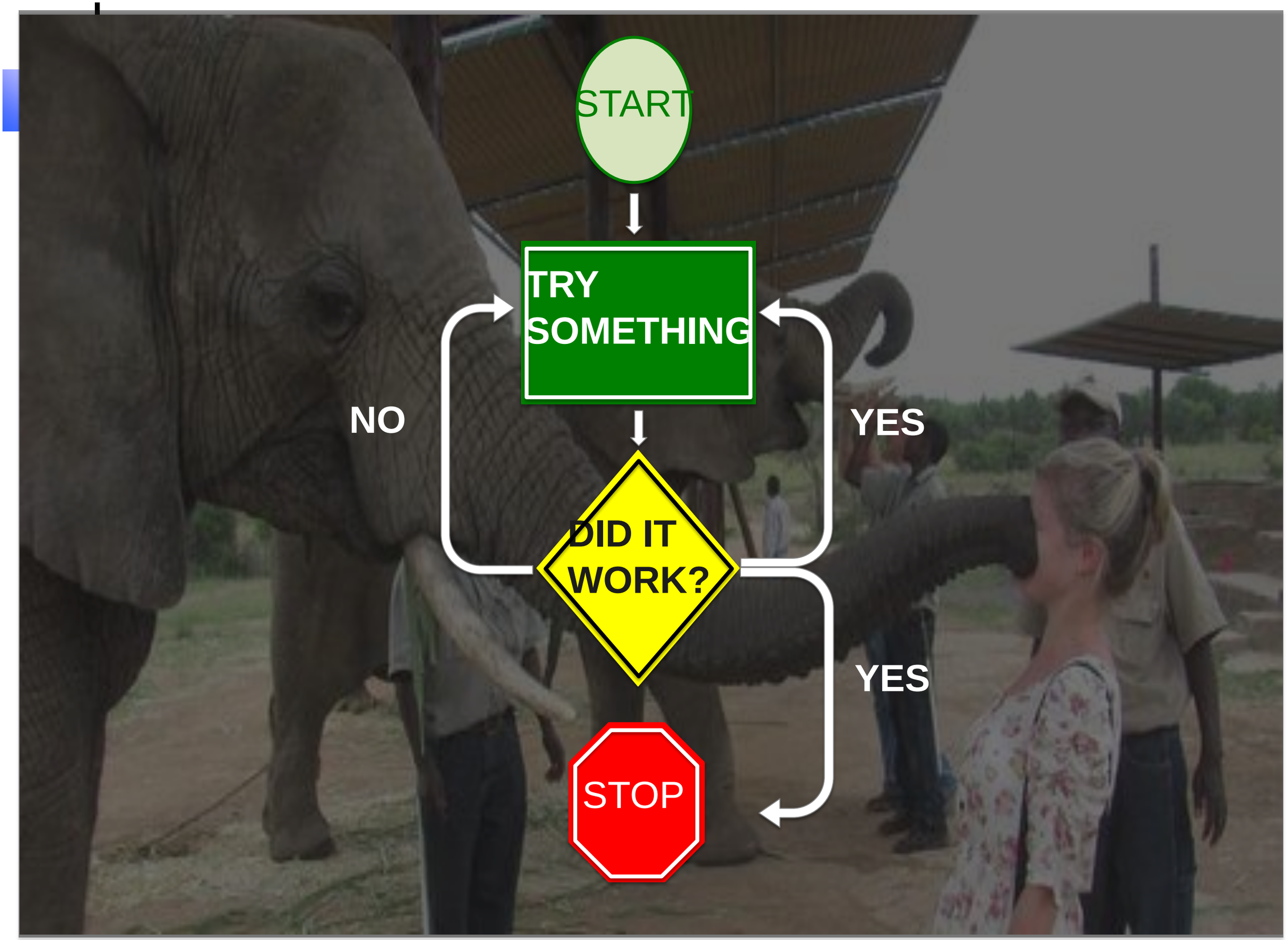
2/24/13, 6:51 AM


How do you
mine (and mind)
all this data?


Most people don’t know
what to do with all the data
that they already have…


Get Big
by starting
small

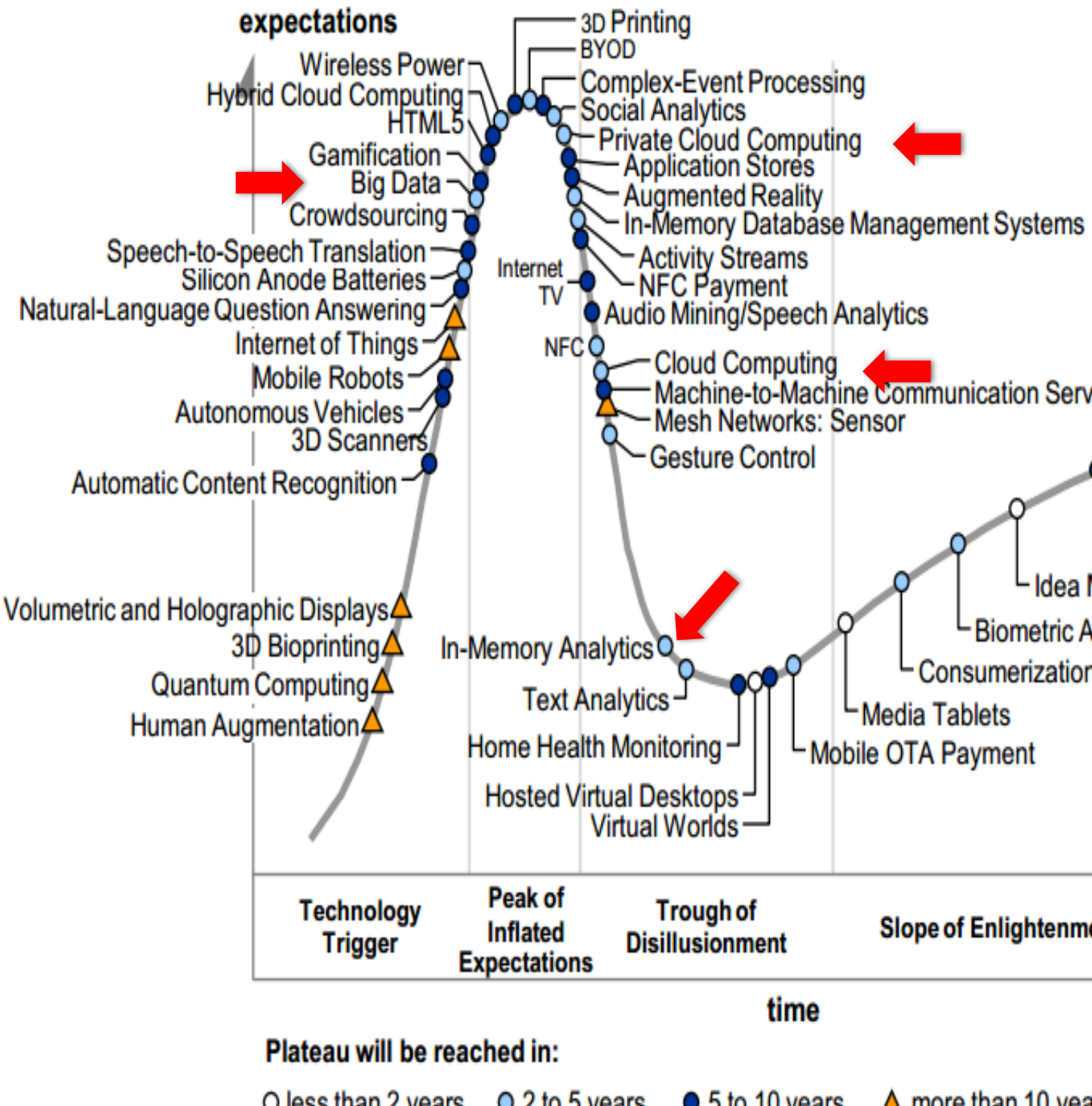

Focus on
Business Impact.


START
TRY SOMETHING
DID IT WORK?
NO
YES
YES
STOP


Big Data isn't *big*,
if you *know* how to
use it.

SMART DATA

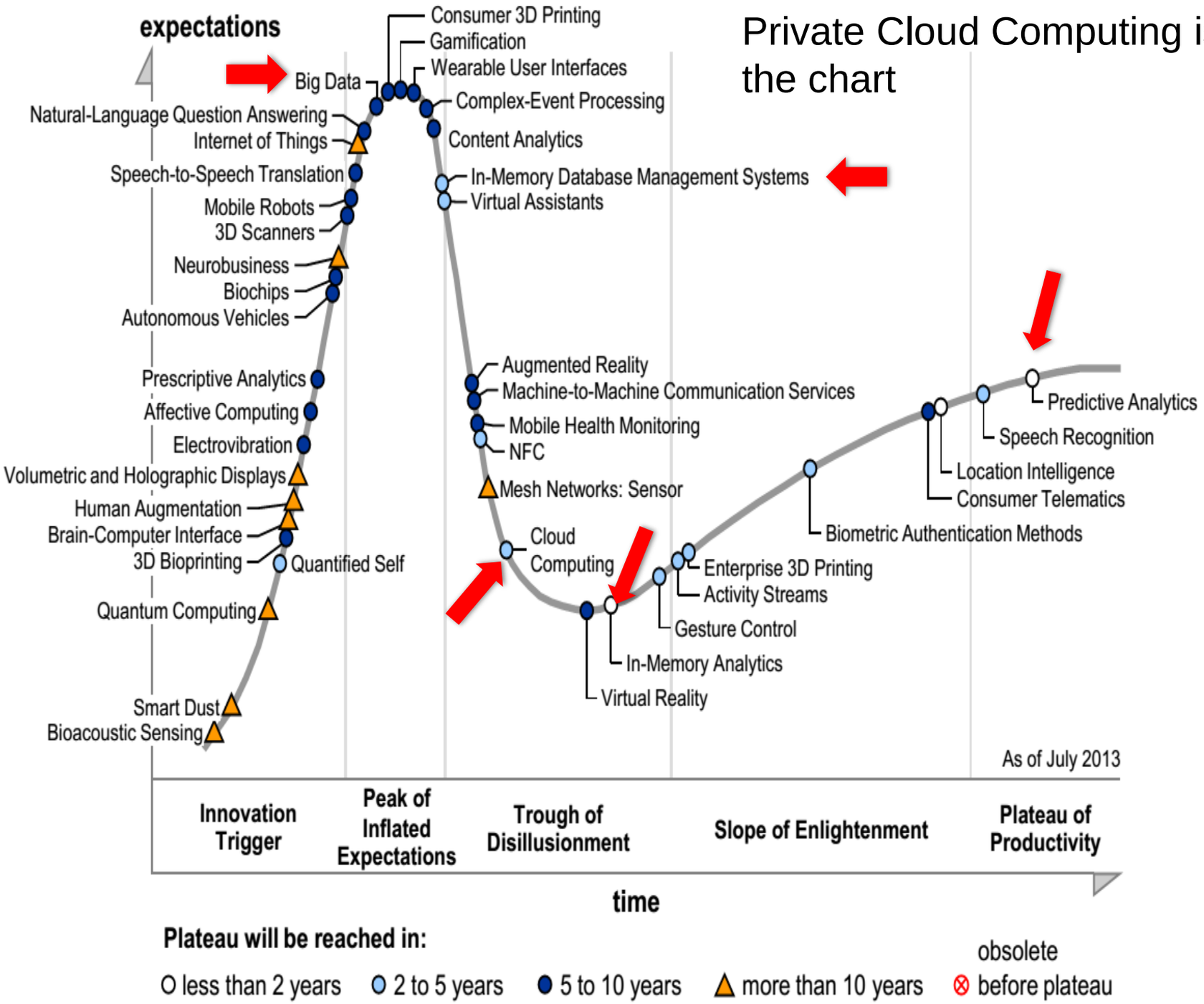
Emerging Technologies Hype Cycle 2012



48 technologies are listed in this year's hype cycle which is the highest in last ten years. Year 2008 was the lowest (27)
Gartner Says in 2012: *We are at an interesting moment — a time when the scenarios we've been talking about for a long time are almost becoming reality.*

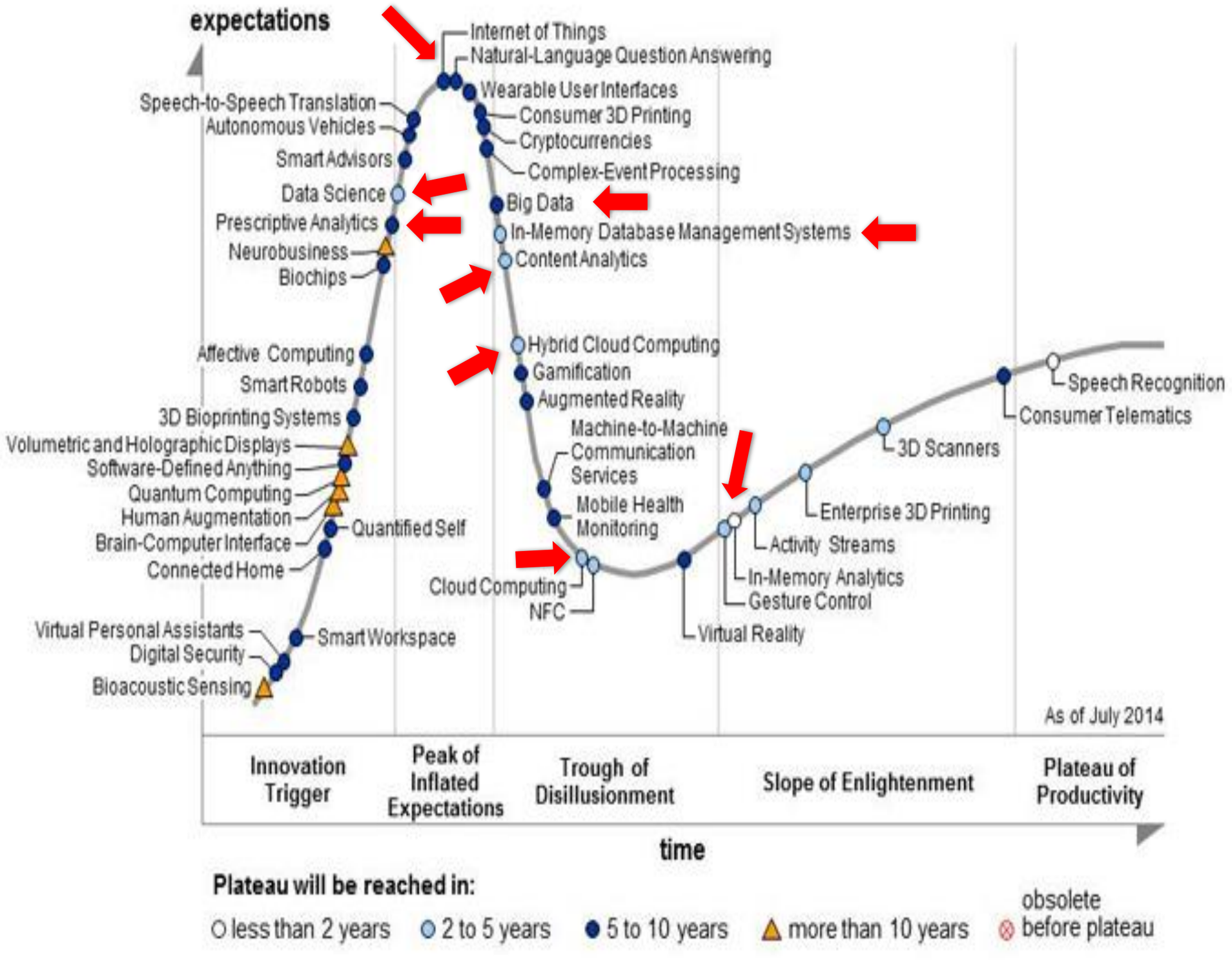
Emerging Technologies Hype Cycle 2013

Private Cloud Computing is off the chart

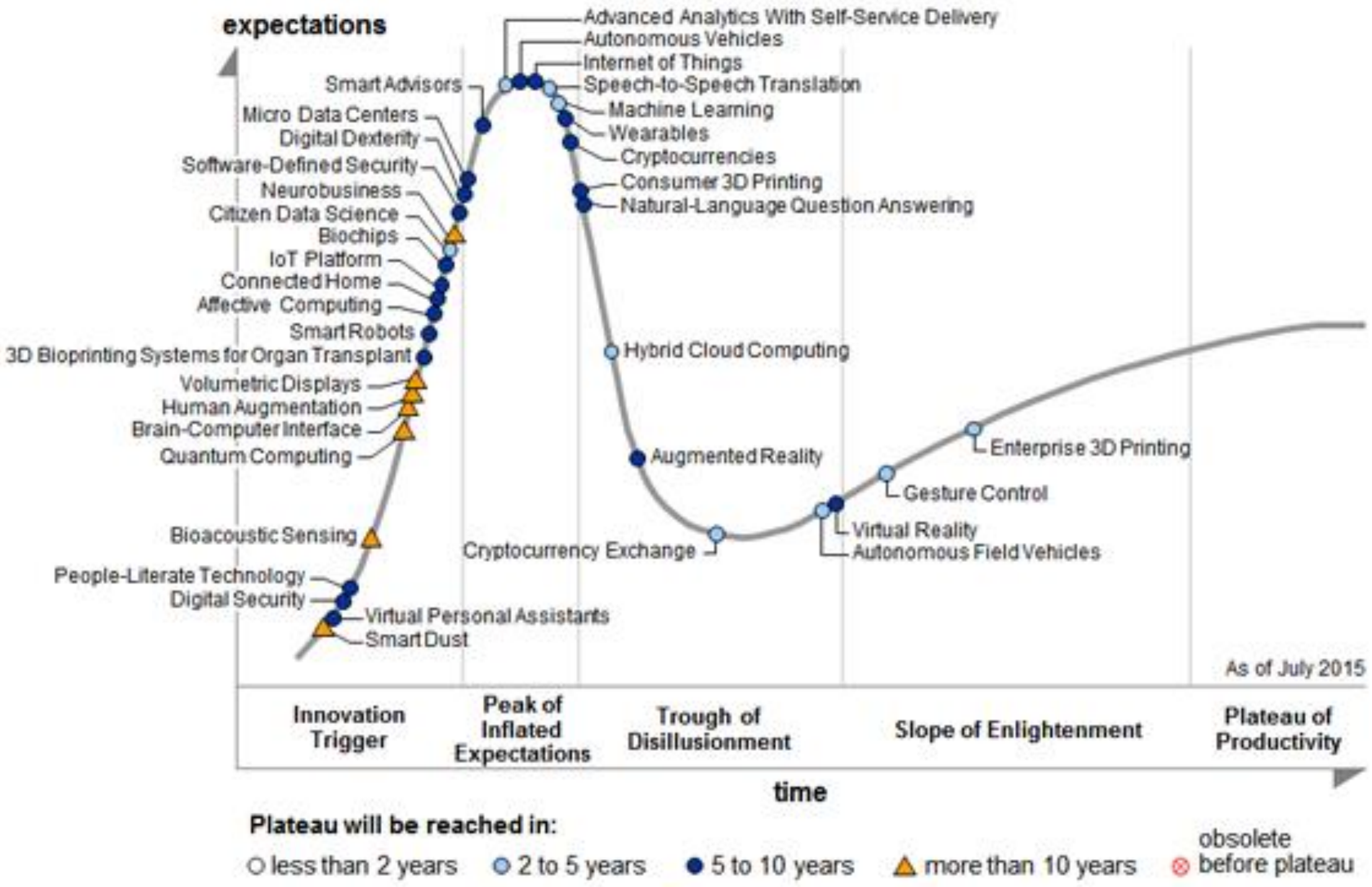


http://public.brighttalk.com/resource/core/19507/august_21_hype_cycle_fenn_lehong_29685.pdf

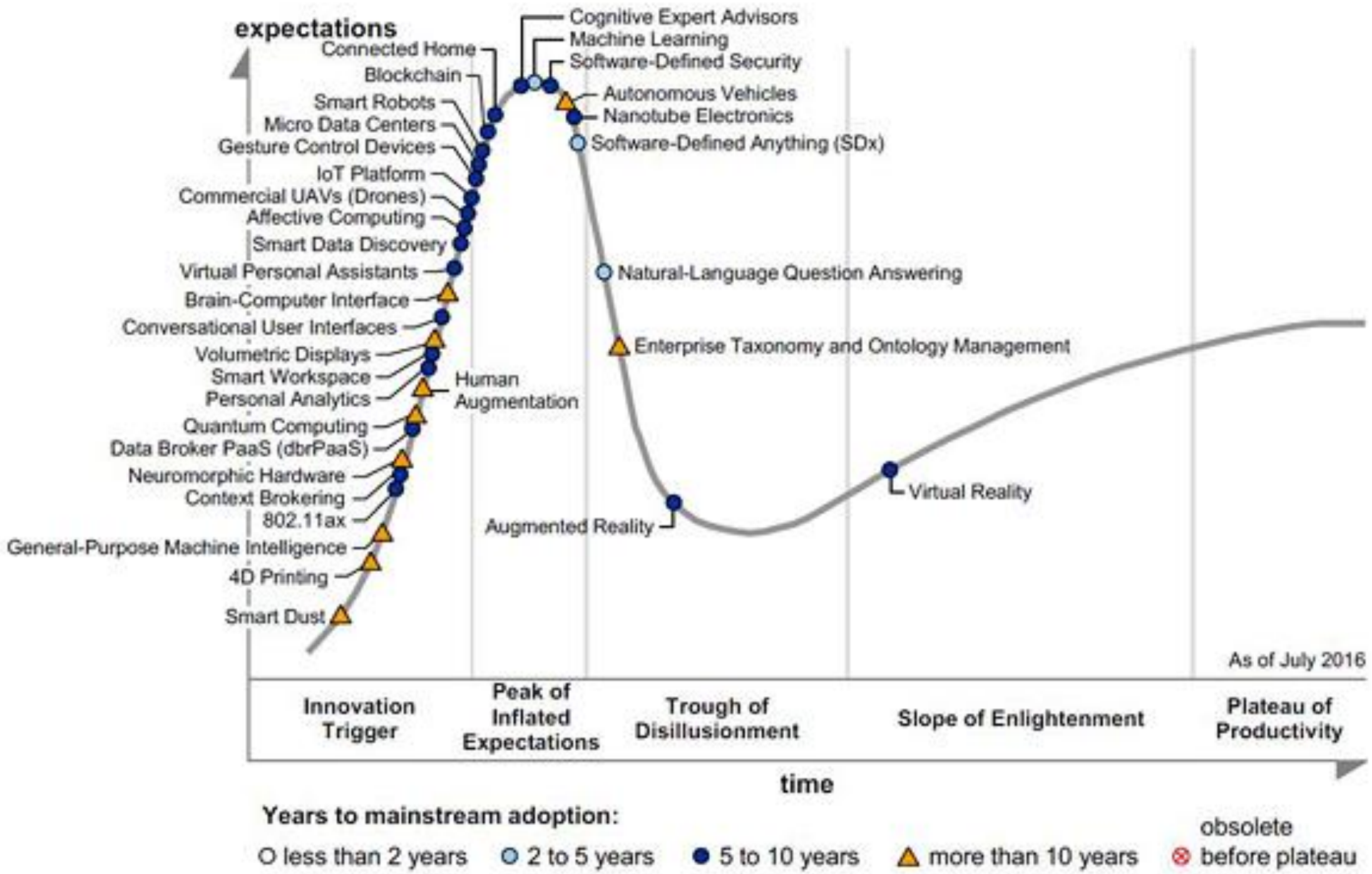
Emerging Technologies Hype Cycle 2014



Emerging Technologies Hype Cycle 2015



Emerging Technologies Hype Cycle 2016



What's Big Data?

No single definition; here is from Wikipedia:

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
  - The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Big data world then

Big data world then

"Big" is a Relative Term
(Context dependent)
IBM 5MB Hard Drive 1956

Big data world then

- The library at Alexandria was the center of data in the ancient world
- Created in 300 BC by Ptolemy II of Egypt
- People from around the known world would congregate and spend their entire lives studying

The ‘Data Lake’ of Antiquity

www.extentia.com

Big data world then

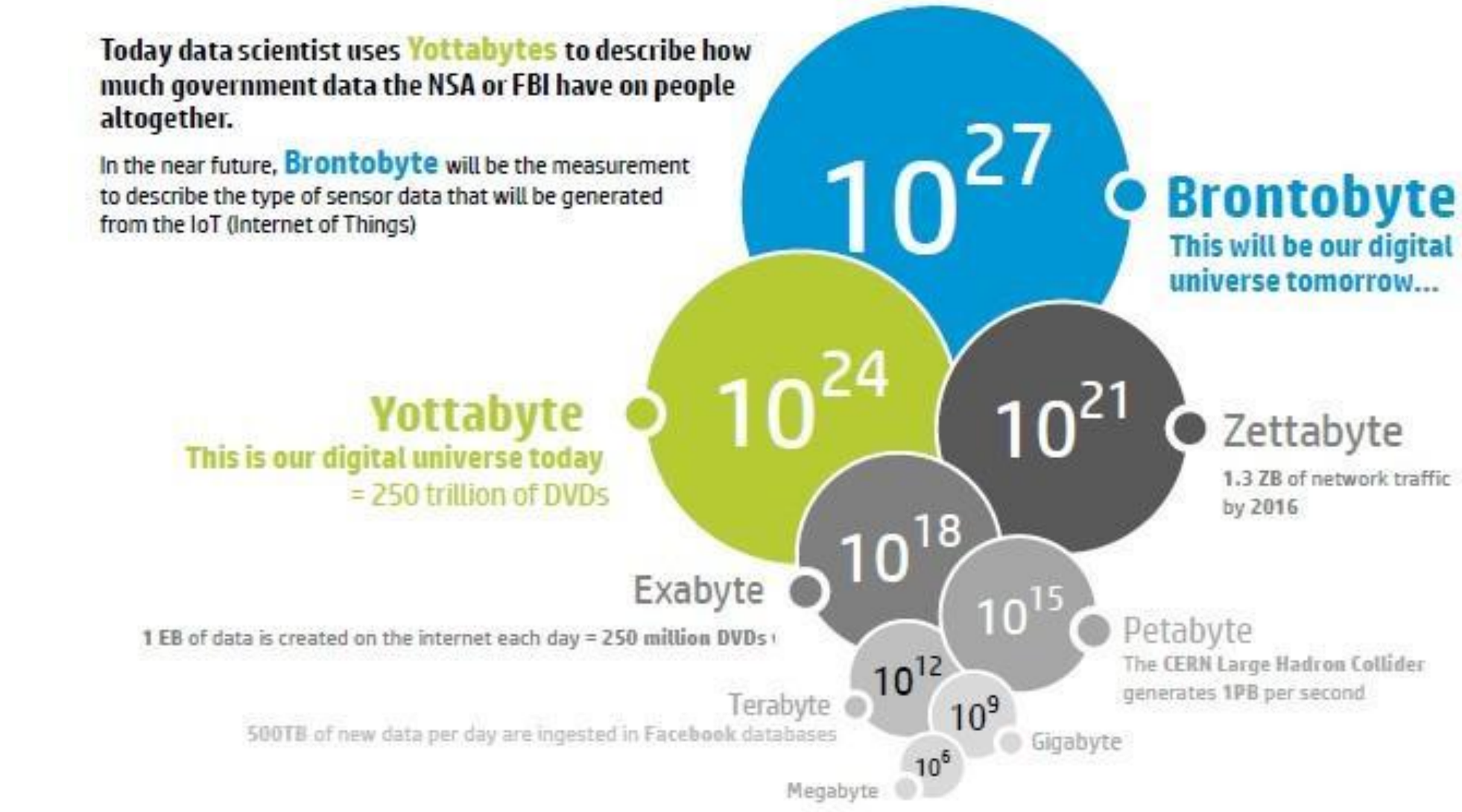
Two men operating a mainframe computer, circa 1960. It's amazing how today's smartphone holds so much more data than this huge 1960's relic. (Photo by Pictorial Parade/Archive Photos)

Big Data Analytics by Vikram Neerugatti

Big Data Statistics 2020

- **Every person will generate 1.7 megabytes in just a second.**
- **Facebook has gained around 2.7 billion active monthly users and generates 4 petabytes of data per day**
- **Facebook** stated that 3.14 billion people were using at least one of the company's core products (Facebook, WhatsApp, Instagram, or Messenger) each month.
- **YouTube currently counts 2 billion monthly active users and 500 hours of content are uploaded to the platform every minute s**
- **Amazon has 150 million mobile users**
- **Twitter users send more than 540,000 tweets every minute.**
- **Over 2.5 quintillion bytes of data is generated worldwide every day.**
- **The amount of global data sphere subject to data analysis will grow to 5.2 zettabytes by 2025.**
- **By 2021, insight-driven businesses are predicted to take $1.8 trillion annually from their less-informed peers.**
- **Data-driven organizations are 23 times more likely to acquire customers than their peers.**
- **Businesses are spending $187 billion on big data and analytics in 2019.**
- **91.6% of firms worldwide confirm an increased pace in investment in big data in 2019.**

Describing Data Size



https://twitter.com/paolopisani/

Big Data

WHY??

- "Retrieval of information".
- "Need of past history"
- "Science and research"
- "Simulation and modeling"
- "Forecasting"
- "Increased population"
- ".....Many more...."

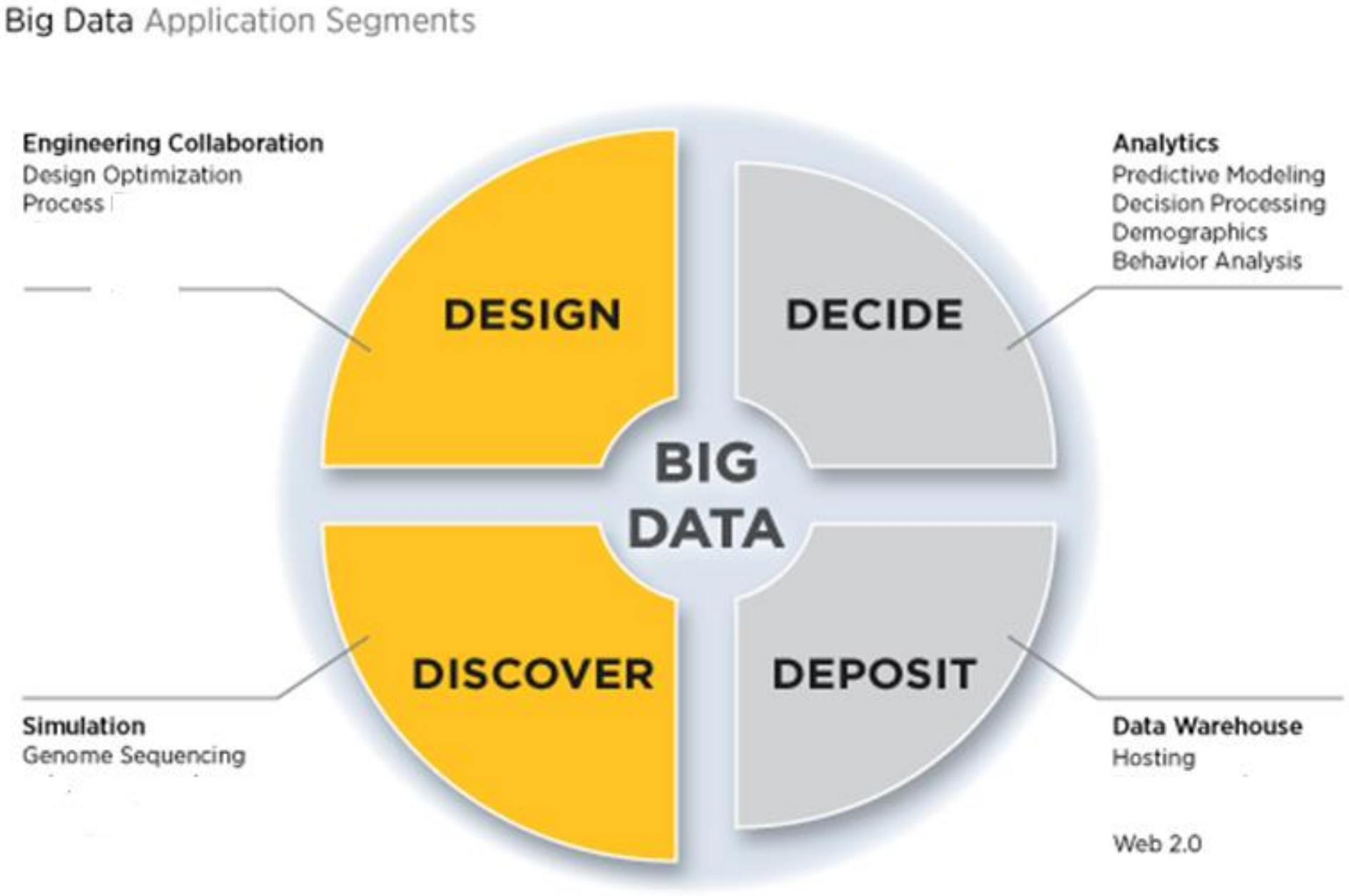
Big Data

- *Big data is a term applied to a new generation of software, applications, and system and storage architecture.*
- *It designed to provide business value from unstructured data.*
- *Big data sets require advanced tools, software, and systems to capture, store, manage, and analyze the data sets,*
- *All in a timeframe Big data preserves the intrinsic value of the data.*
- *Big data is now applied more broadly to cover commercial environments.*

Big data

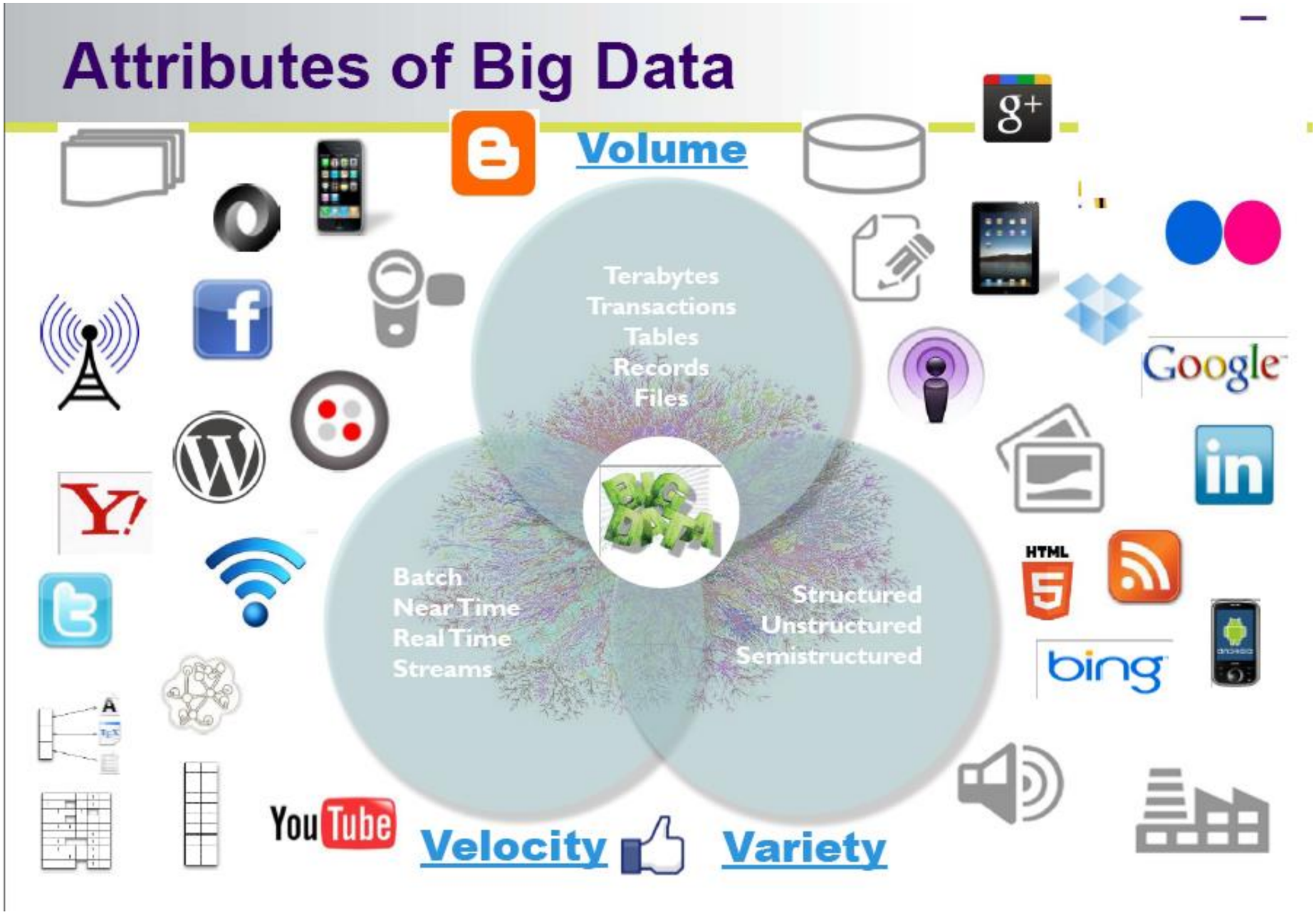
- *Four distinct applications segments comprise the big data market.*
- *each with varying levels of need for performance and scalability.*
- *The four big data segments are:*
- *1) Design (engineering collaboration)*
- *2) Discover (core simulation – supplanting physical experimentation)*
- *3) Decide (analytics).*
- *4) Deposit (Web 2.0 and data warehousing)*

Big Data


Big Data Application Segments
Engineering Collaboration
Design Optimization
Process
Analytics
Predictive Modeling
Decision Processing
Demographics
Behavior Analysis
DESIGN
DECIDE
BIG DATA
DISCOVER
DEPOSIT
Simulation
Genome Sequencing
Data Warehouse
Hosting
Web 2.0

Big Data
“Data Driven” Web 2.0 onwards.

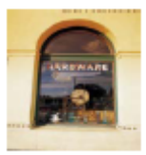
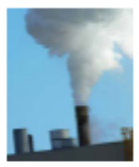
Big Data


Attributes of Big Data
Volume
Terabytes
Transactions
Tables
Records
Files
Batch
Near Time
Real Time
Streams
Structured
Unstructured
Semistructured
Velocity
Variety
Google
bing
HTML
YouTube

Big Data

Big Data

The Big Data Opportunity

Financial Services	Healthcare
	
Retail	**Web/Social/Mobile**
	
Manufacturing	**Government**
	

Big Data

Ten Common Big Data Problems

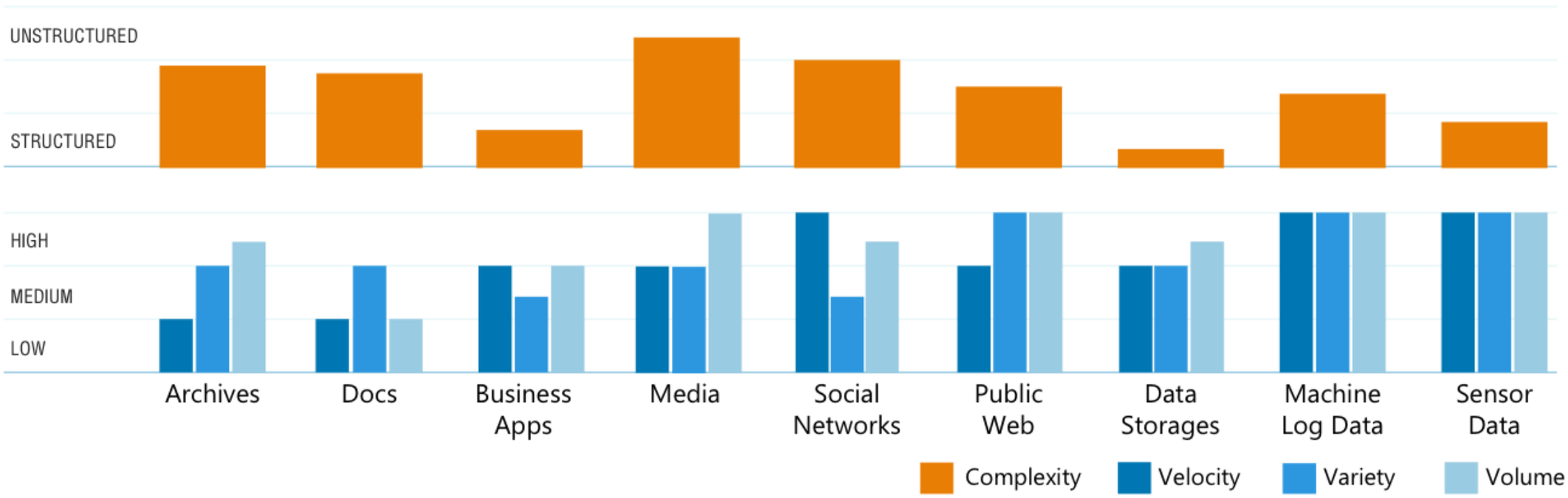
1. Modeling true risk
2. Customer churn analysis
3. Recommendation engine
4. Ad targeting
5. Transaction analysis
6. Analyzing network data to predict failure
7. Threat analysis
8. Trade surveillance
9. Search quality
10. Data "sandbox"

Big Data Challenges

Top 5 Big Data Challenges

1. Deciding what data is relevant
2. Cost of technology infrastructure
3. Lack of skills to analyze the data
4. Lack of skills to manage big data projects
5. Lack of business support

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.

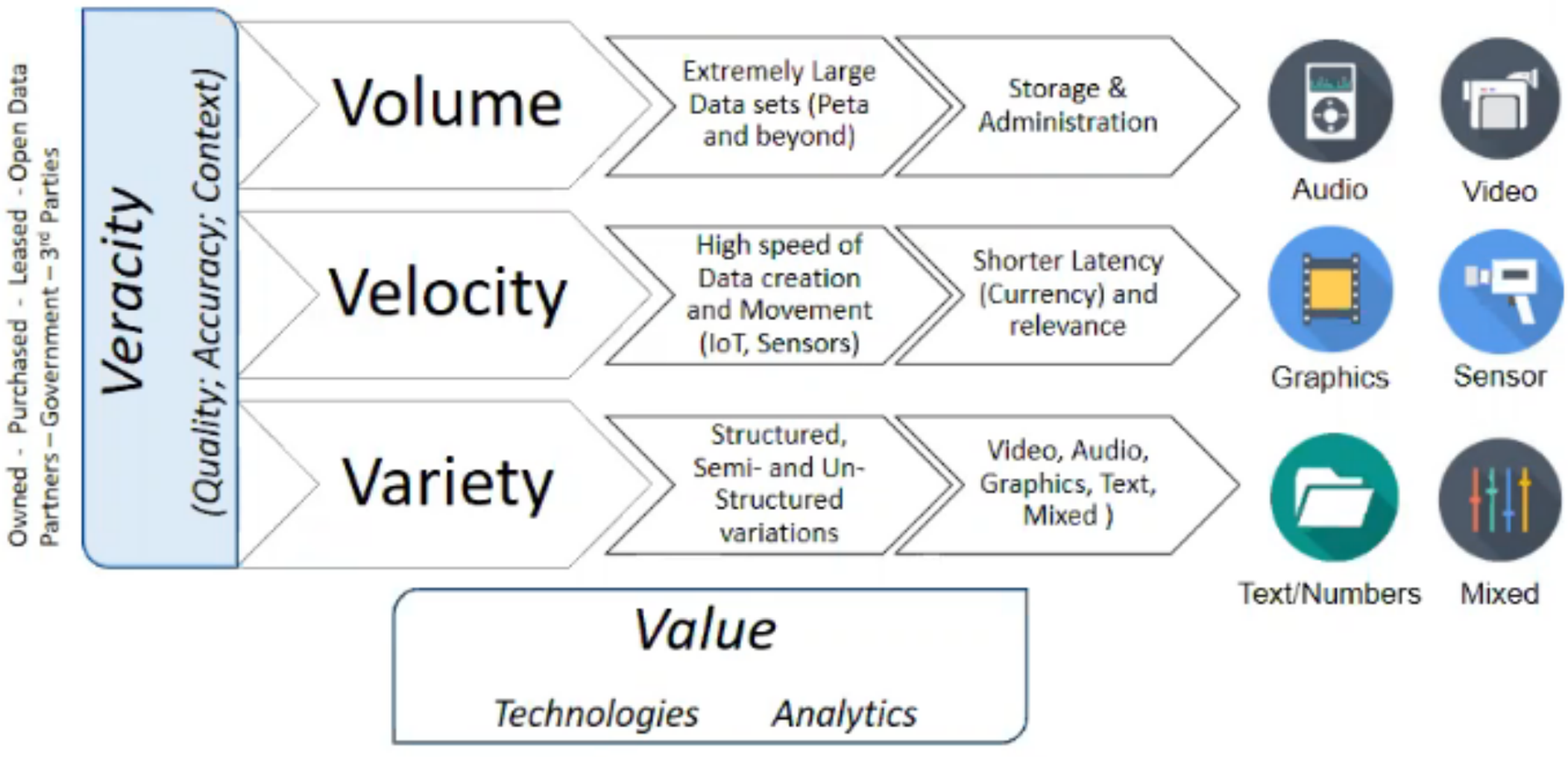
Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

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

Applications

V's of Big data

Figure 3.3 : Detailed Characteristics of Big Data's 3+1+1 Vs and the Types and Categories of data. (T fifth V for Value is the focus of BDFAB)



Each DATA characteristic impacts the way Big Data Strategies and corresponding Solutions are designed, developed and used

Sources of Big data

12+ TBs

of tweet data every day

? TBs

of data every day

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

30 billion

RFID tags today (1.3B in 2005)

76 million

smart meters in 2009… 200M by 2014

4.6 billion

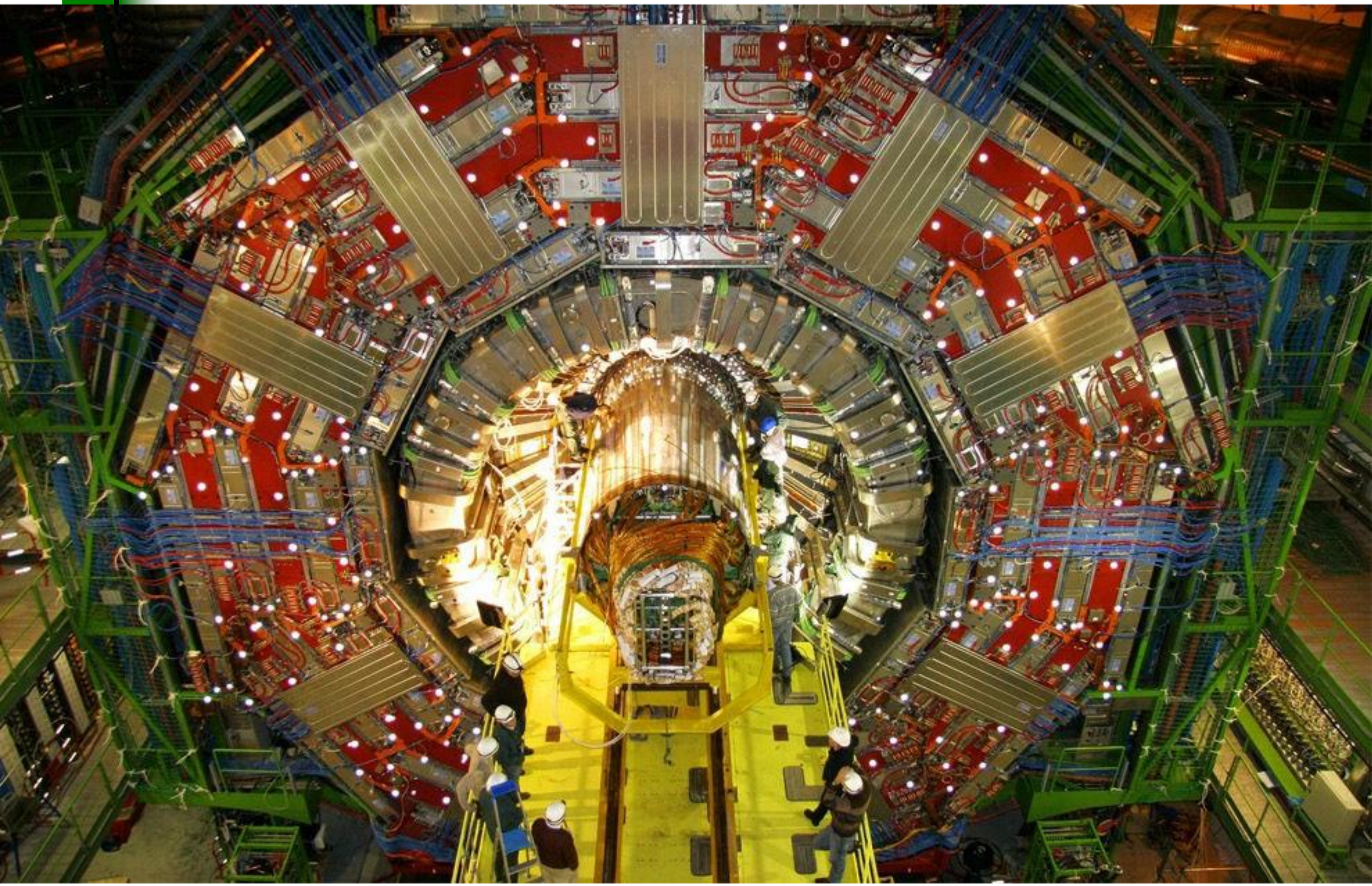
camera phones world wide

100s of millions of GPS enabled

2+ billion

devices sold annually

people on the Web by end 2011



CERN’s Large Hydron Collider (LHC)

The Earthscope

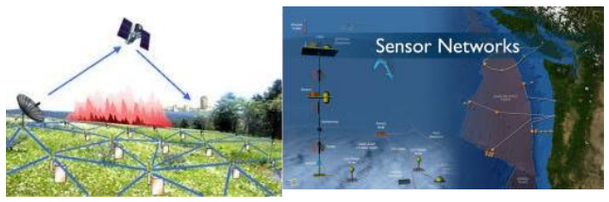
- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--ul)

Real-time/Fast Data

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

Scientific instruments
(collecting all sorts of data)

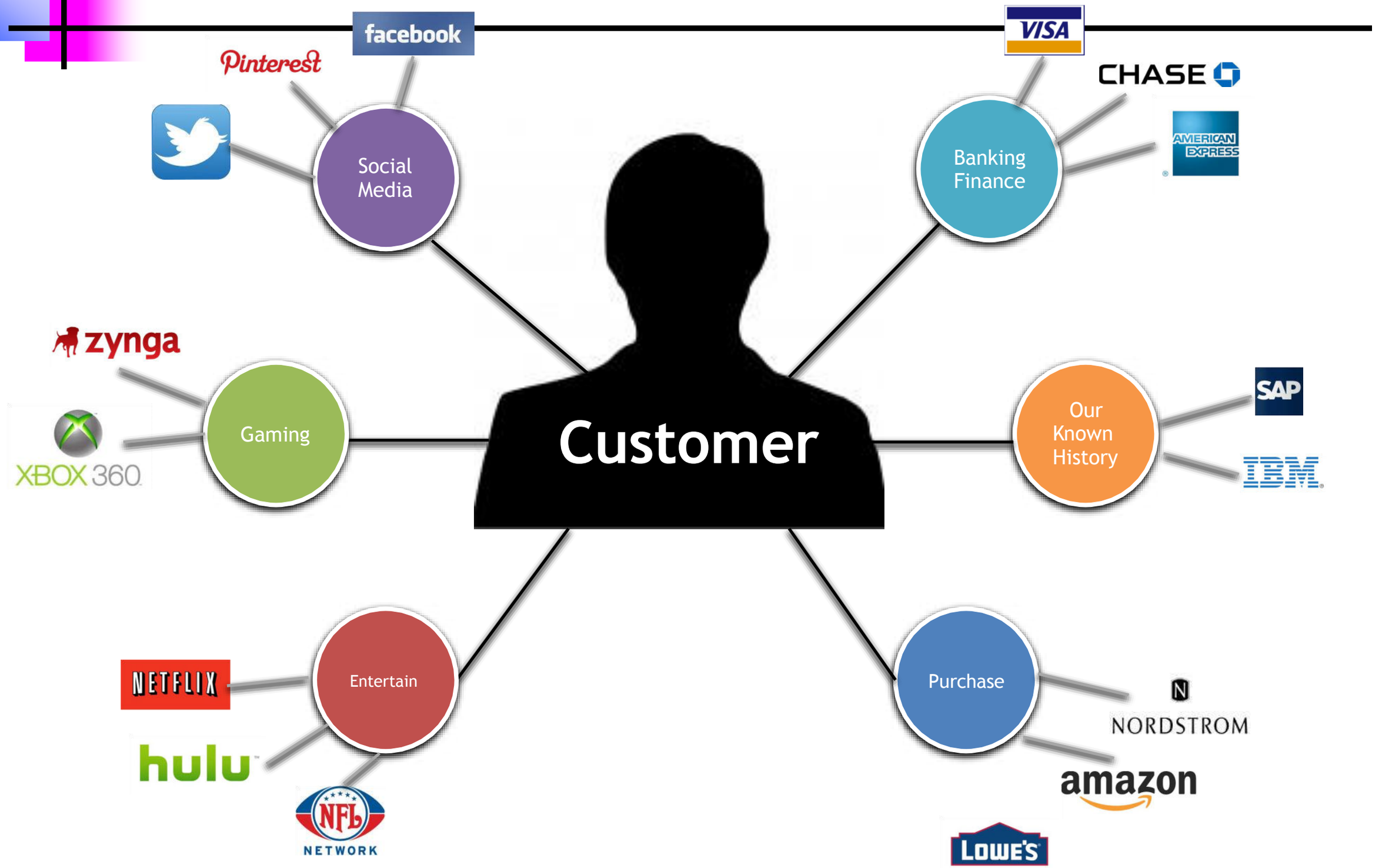
Mobile devices
(tracking all objects all the time



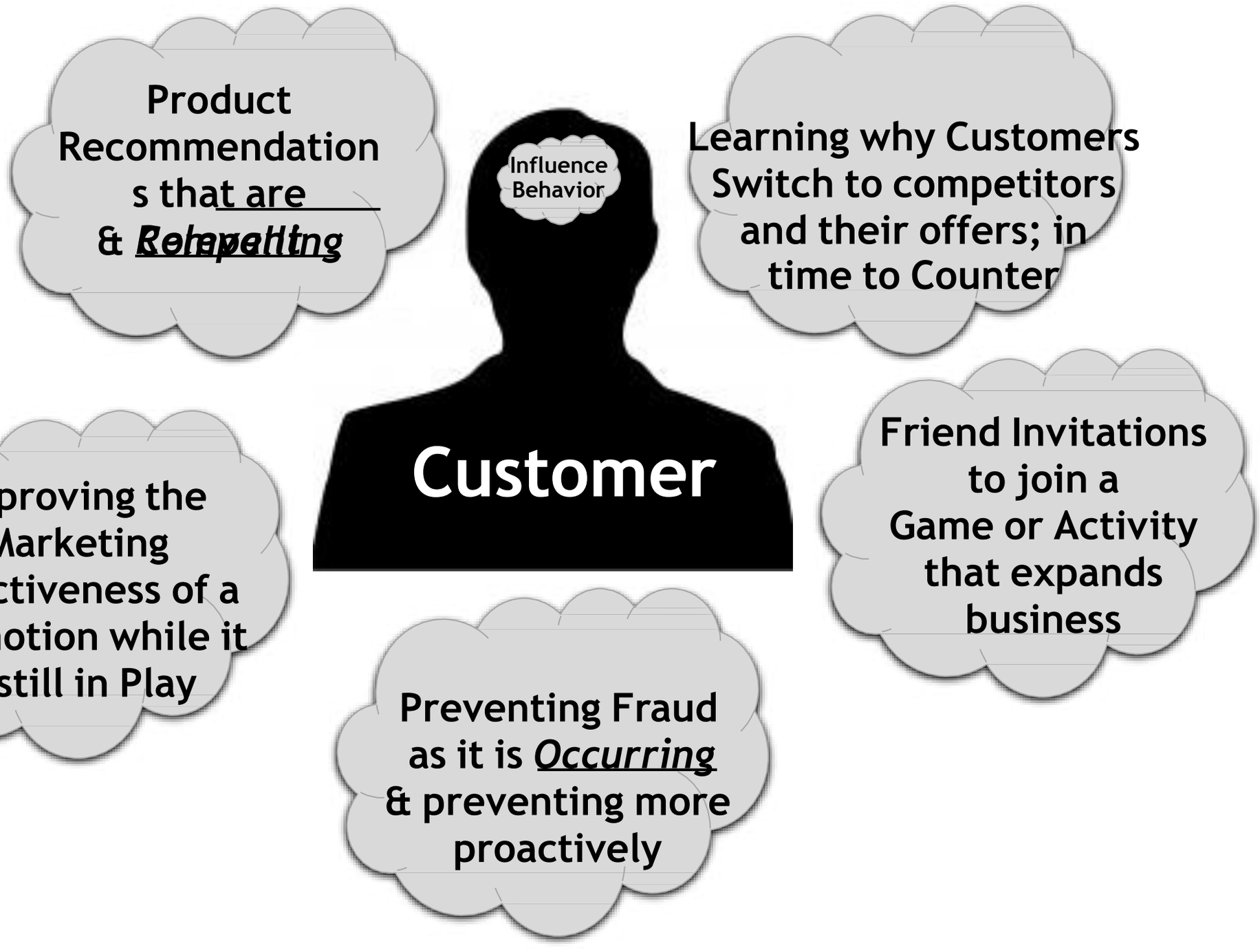
Sensor technology and networks
(measuring all kinds of data)

- 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

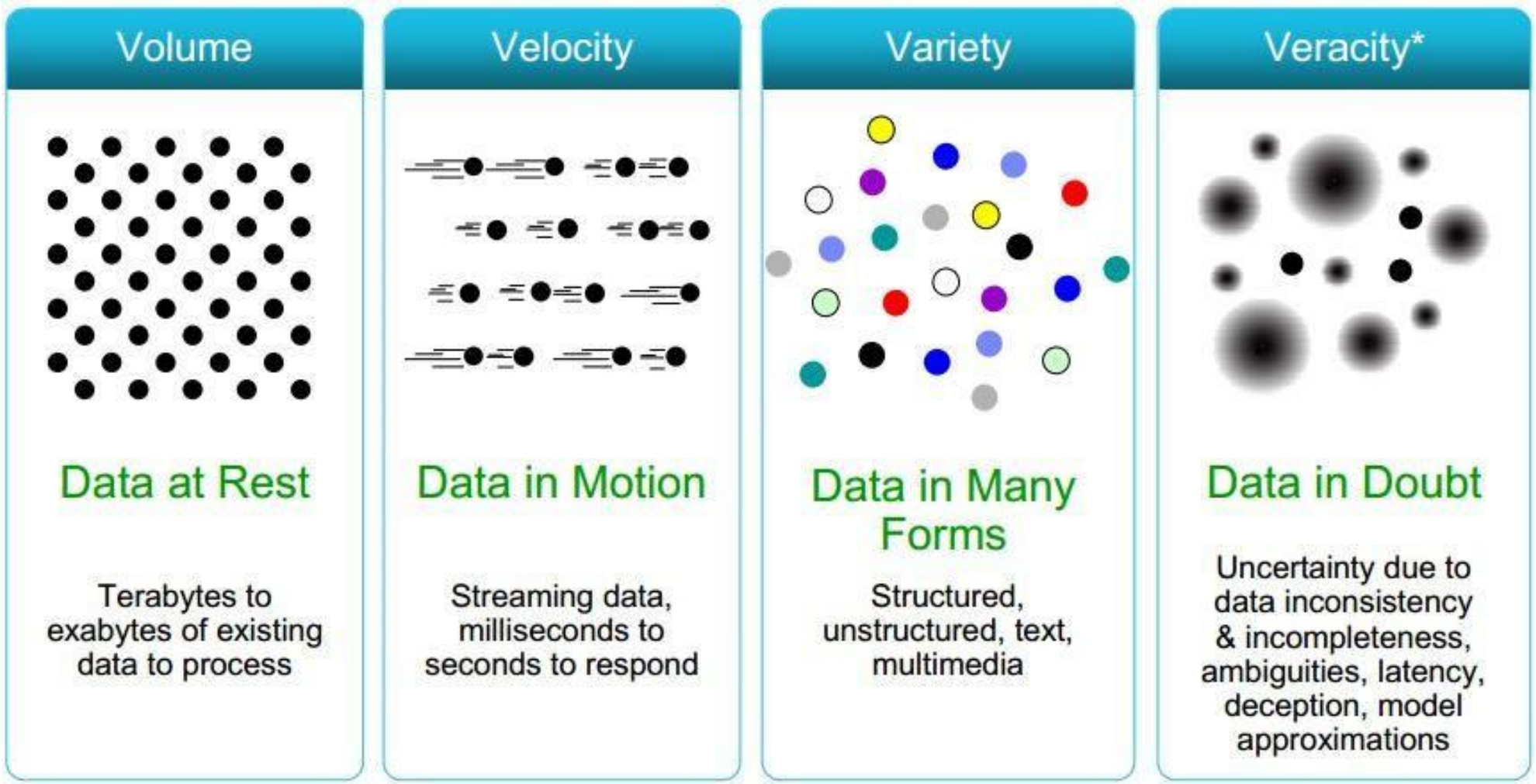
A Single View to the Customer



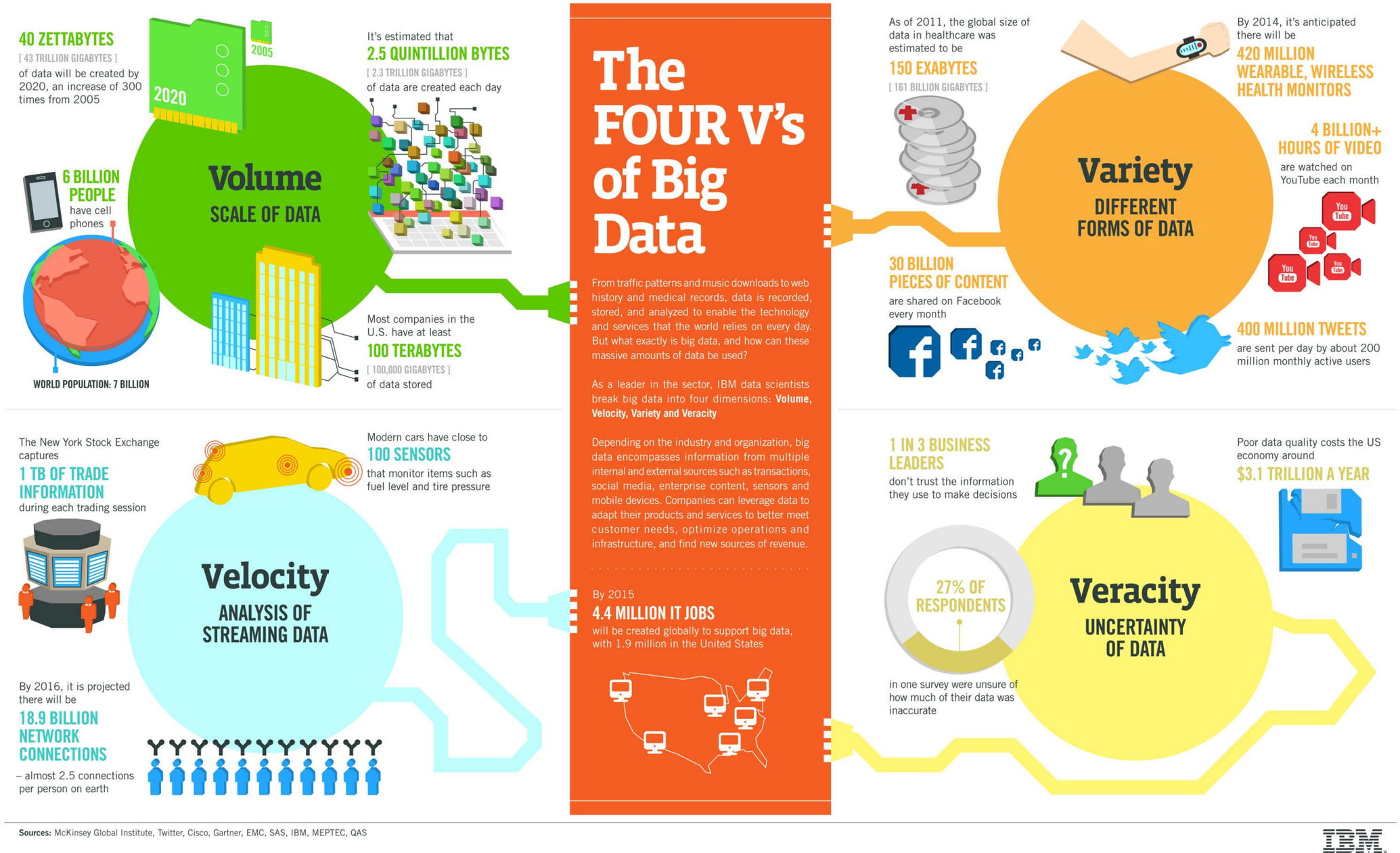
Real-Time Analytics/Decision Requirement



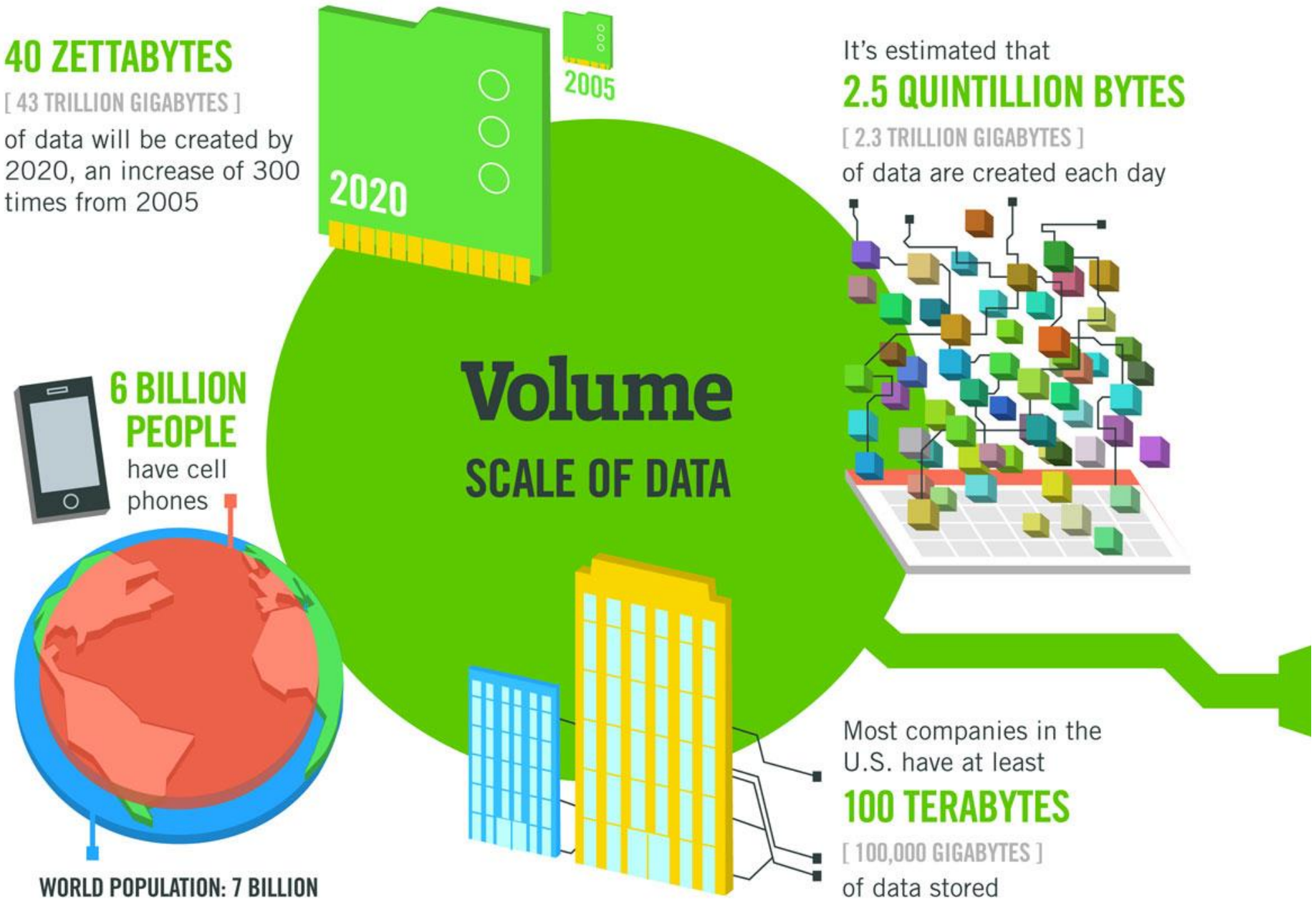
Some Make it 4V's



V's of Big data



V's of Big data



V's of Big data

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

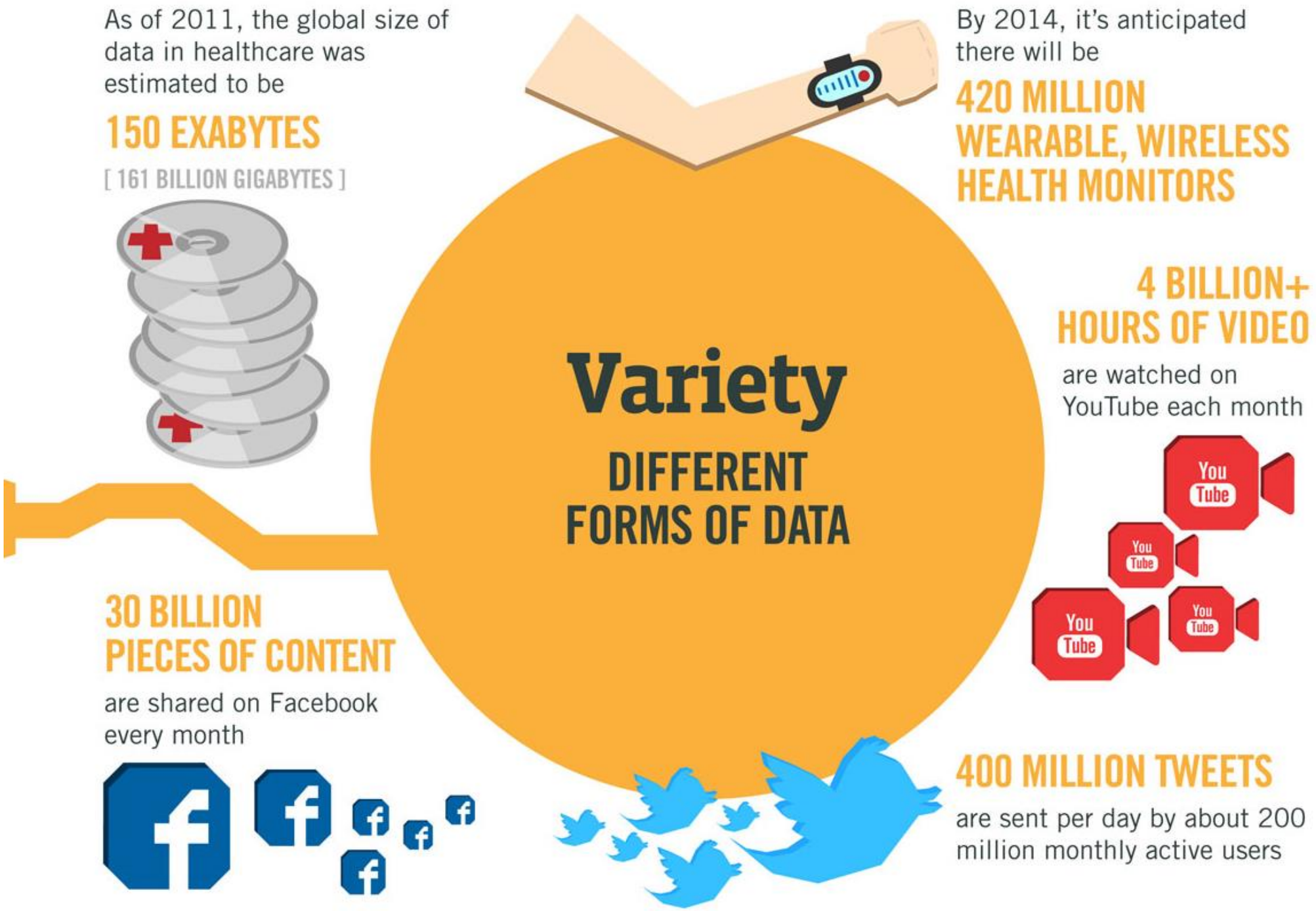
V's of Big data

420 million wearable,

wireless health monitors

4 billion hours of video are watched on YouTube each month

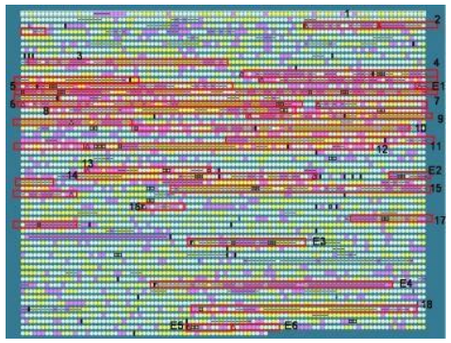
400 million tweets are sent per day 200 million mo

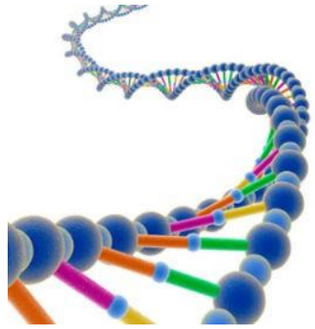


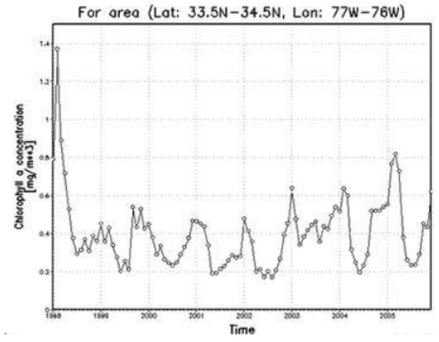
Variety (Complexity)

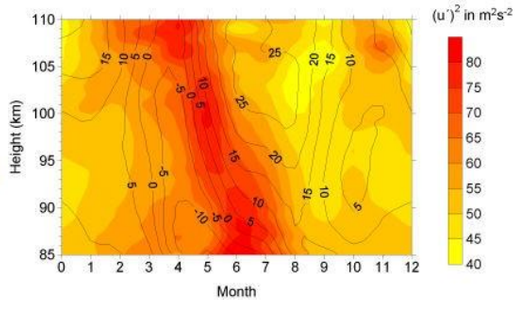
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data

  Big Public Data (online, weather, finance, etc)
-



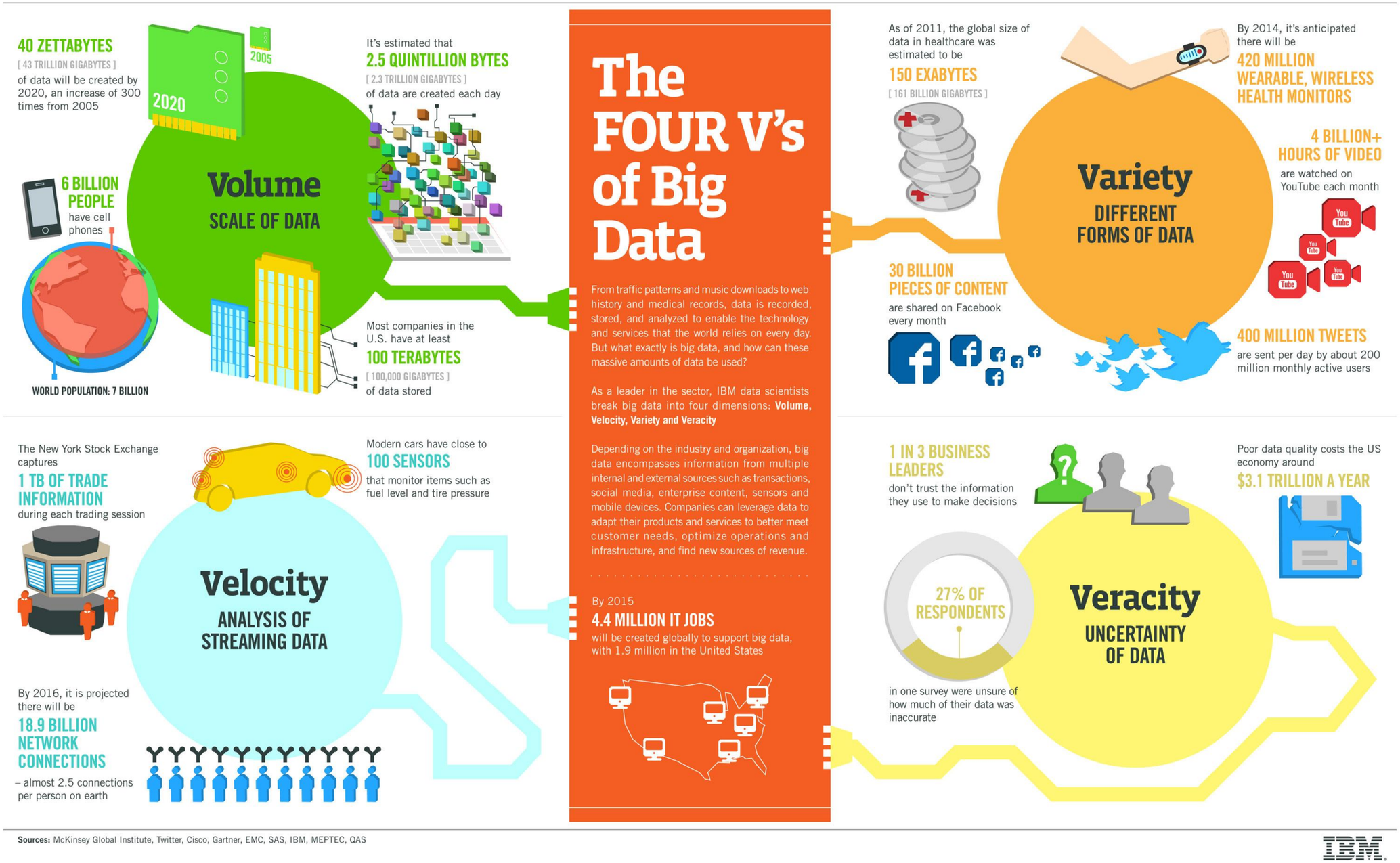




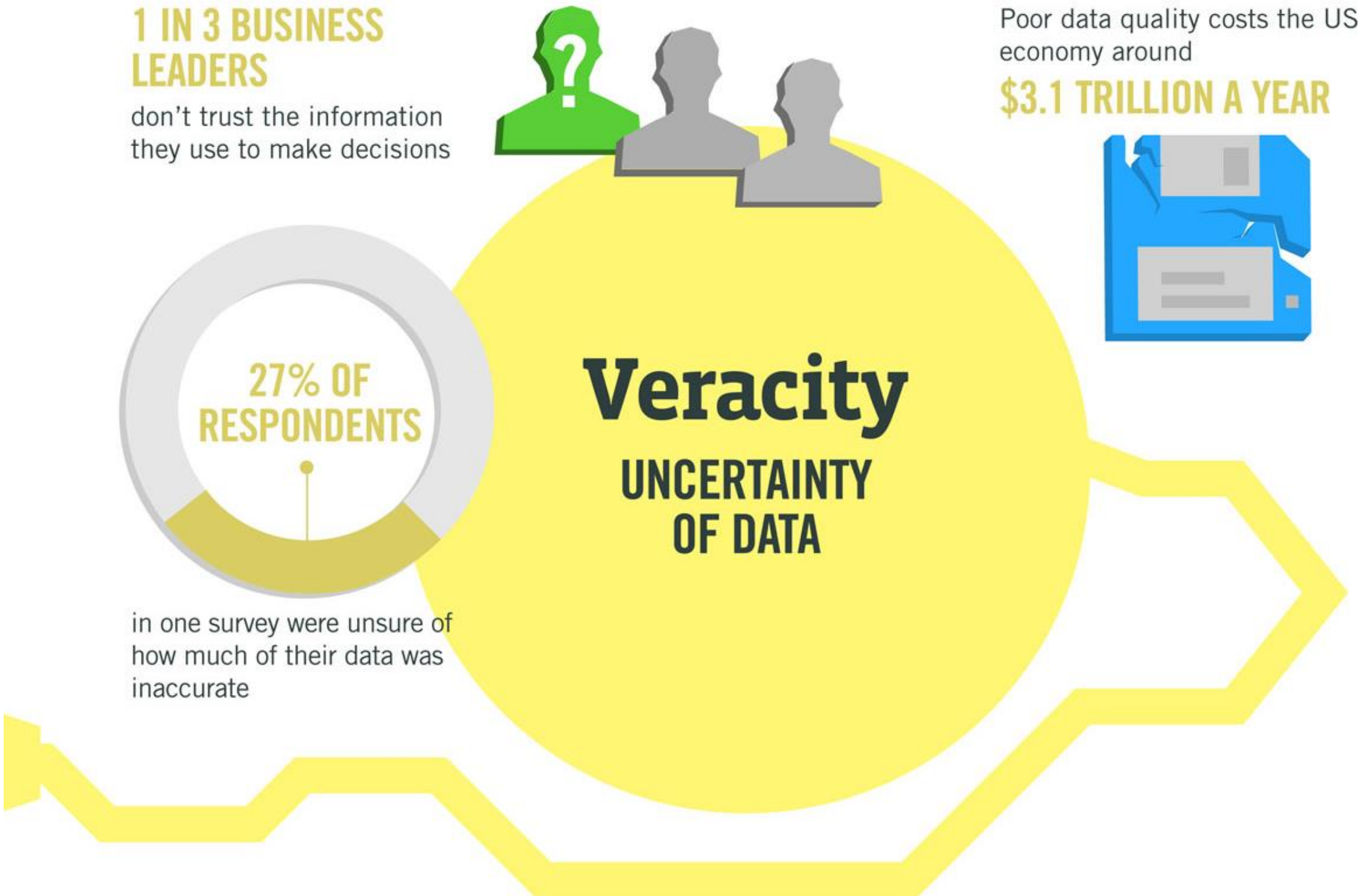


To extract knowledge轉 all these types of data need to linked together

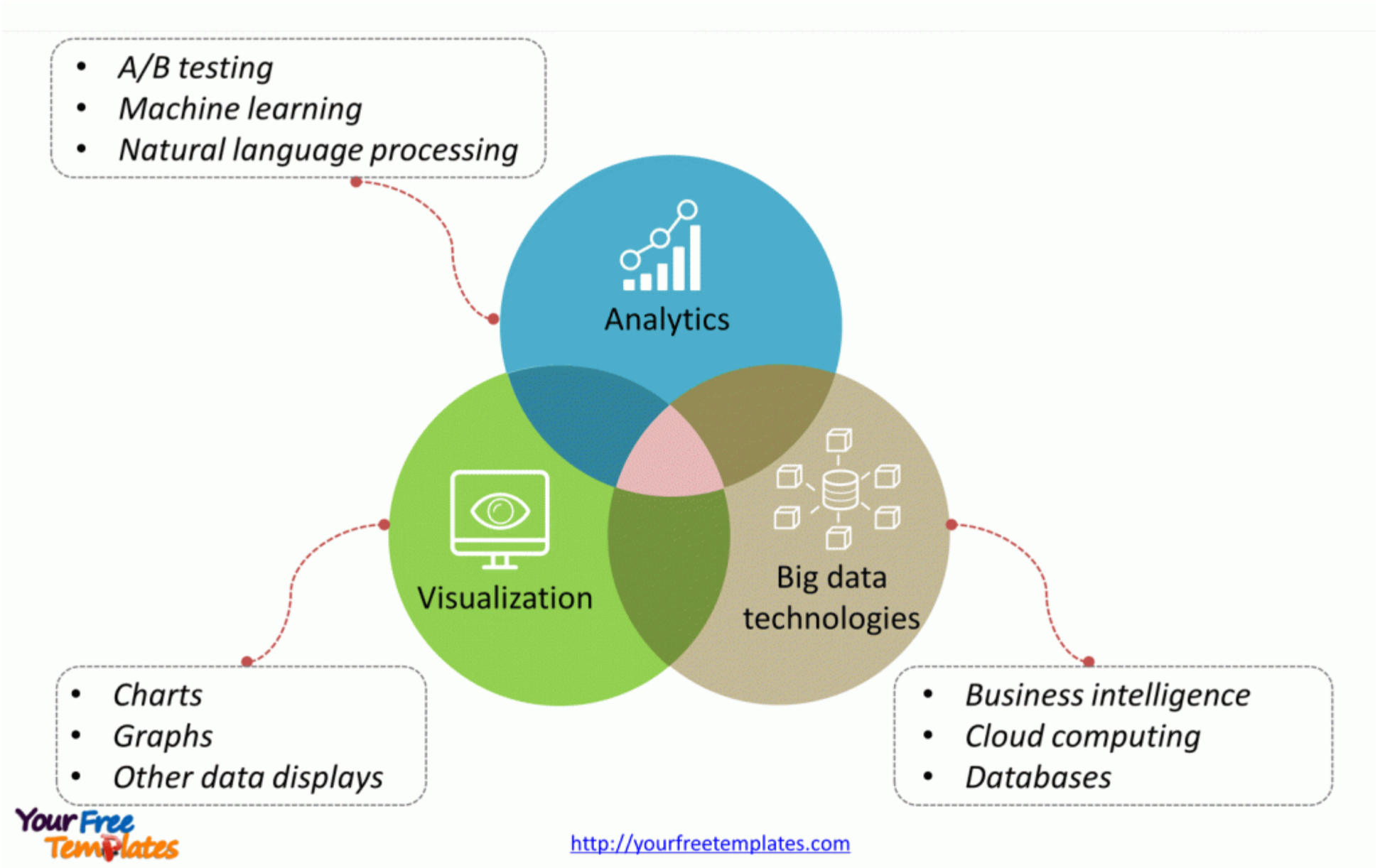
V's of Big data



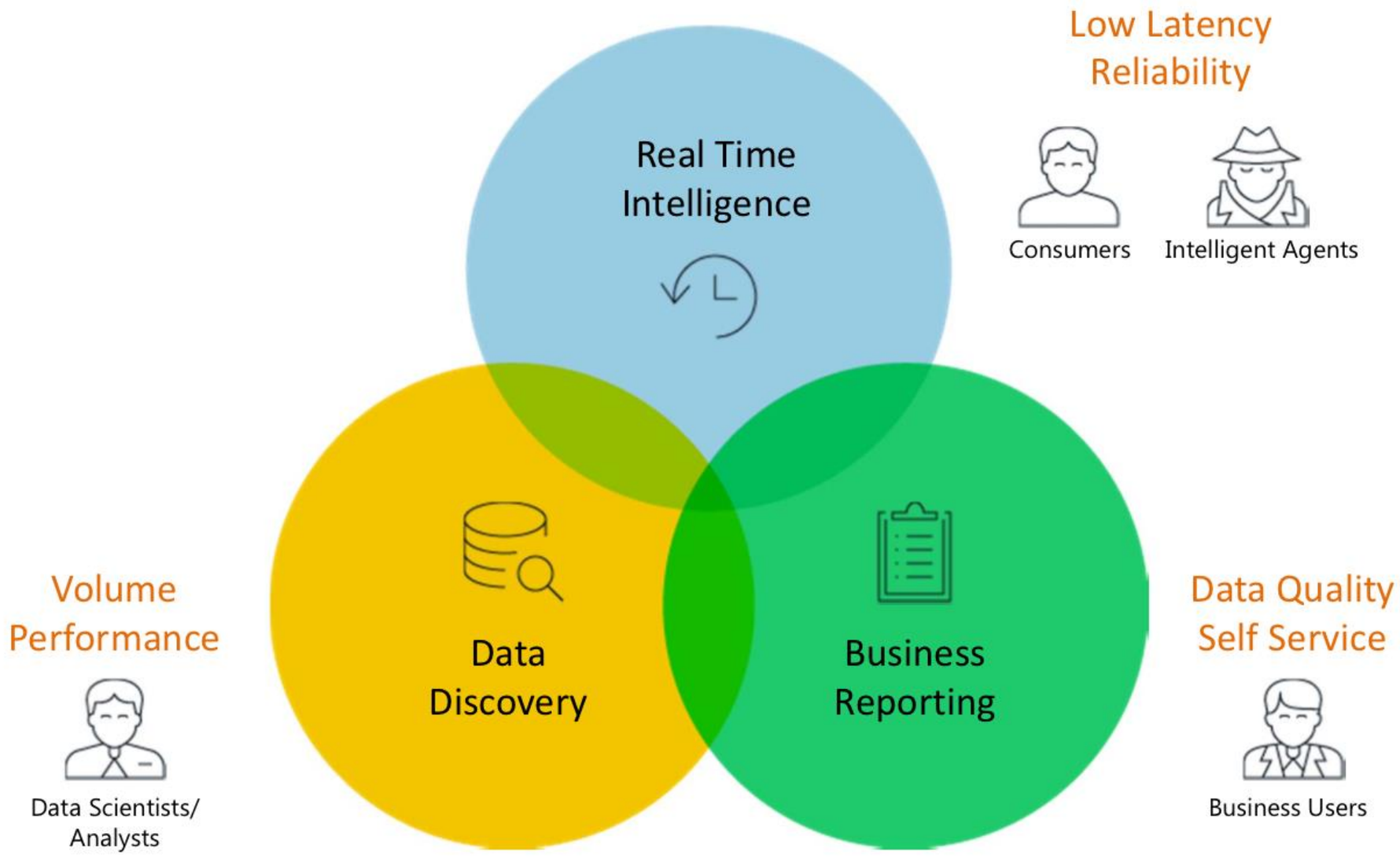
V’s of Big data



Technologies for Big data

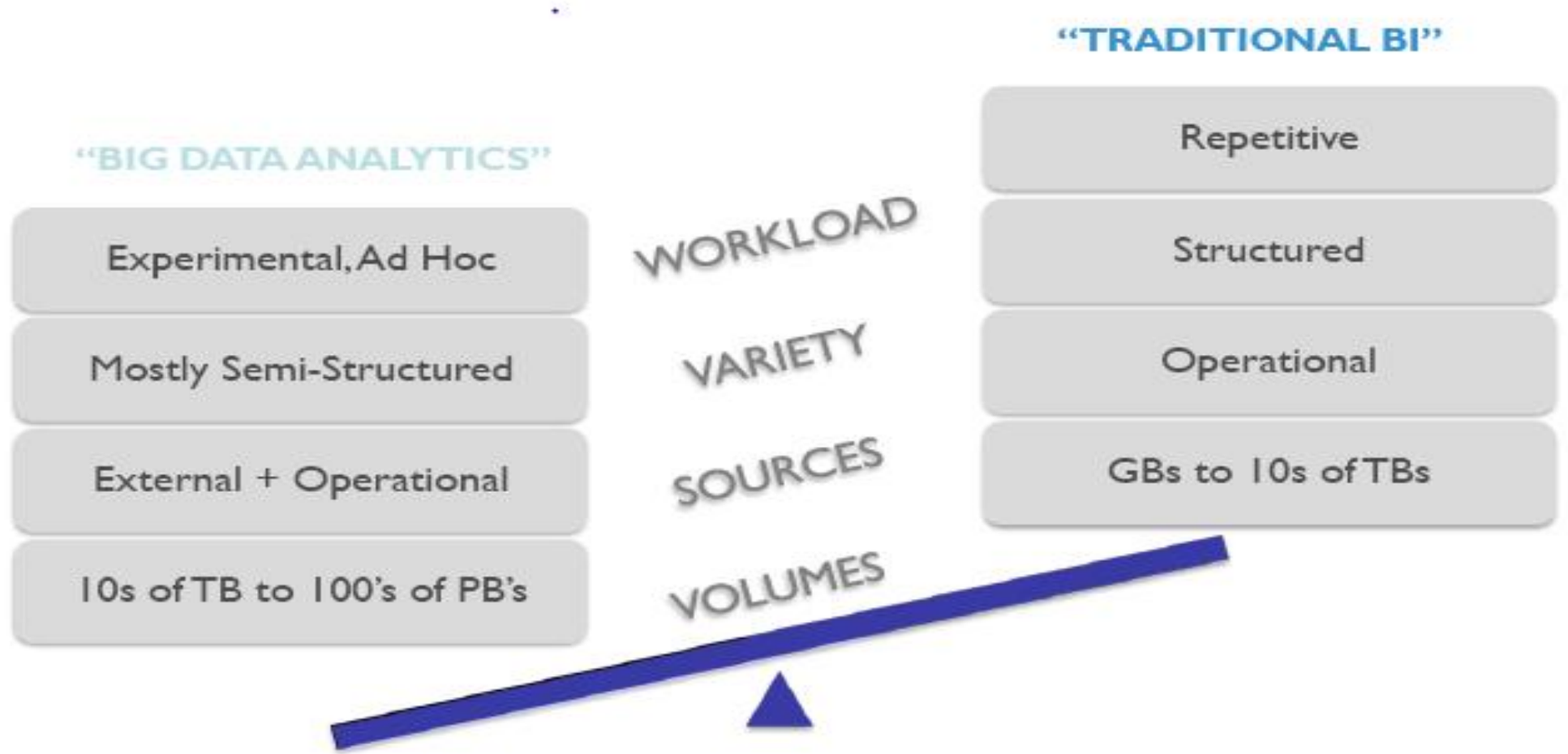


Big Data Analytics Use Cases



Big Data

Big Data Is Different than Business Intelligence



Big Data Analytics

- *Big data analytics is the process of examining large amounts of data of a variety of types.*
- *The primary goal of big data analytics is to help companies make better business decisions.*
- *analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs.*

Data Analytics

- Big data Consist of

  – uncovered hidden patterns.

  – Unknown correlations and other useful information.

  ➢Such information can provide business benefits.

  ➢more effective marketing and increased revenue.

Data Analytics

- Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines.
- such as predictive analysis and data mining.

- But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.

- Traditional data warehouses may not be able to handle the processing demands posed by big data.

Data Analytics

- *The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce.*

- *Known about these technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.*
- *big data analytics initiatives include*
  - *internal data analytics skills*
  - *high cost of hiring experienced analytics professionals,*
  - *challenges in integrating Hadoop systems and data warehouses*

Data Analytics

- *Big Analytics delivers competitive advantage in two ways compared to the traditional analytical model.*

- *First, Big Analytics describes the efficient use of a simple model applied to volumes of data that would be too large for the traditional analytical environment.*

- *Research suggests that a simple algorithm with a large volume of data is more accurate than a sophisticated algorithm with little data.*

Data Analytics

- *Big Analytics supporting the following objectives for working with Big Data Analytics:*

- *1. Avoid sampling / aggregation;*

- *2. Reduce data movement and replication;*

- *3. Bring the analytics as close as possible to the data.*

- *4. Optimize computation speed.*

Data Analytics

- *The term "analytics" refers to the use of information technology to harness statistics, algorithms and other tools of mathematics to improve decision-making.*
- *Guidance for analytics must recognize that processing of data may not be linear.*
- *May involve the use of data from a wide array of sources.*
- *Principles of fair information practices may be applicable at different points in analytic processing.*
- *Guidance must be sufficiently flexible to serve the dynamic nature of analytics and the richness of the data to which it is applied.*

The Power and Promise of Analytics

- *Big Data Analytics to Improve Network Security.*
- *Security professionals manage enterprise system risks by controlling access to systems, services and applications defending against external threats.*
- *protecting valuable data and assets from theft and loss.*
- *monitoring the network to quickly detect and recover from an attack.*
- *Big data analytics is particularly important to network monitoring, auditing and recovery.*
- *Business Intelligence uses big data and analytics for these purposes.*

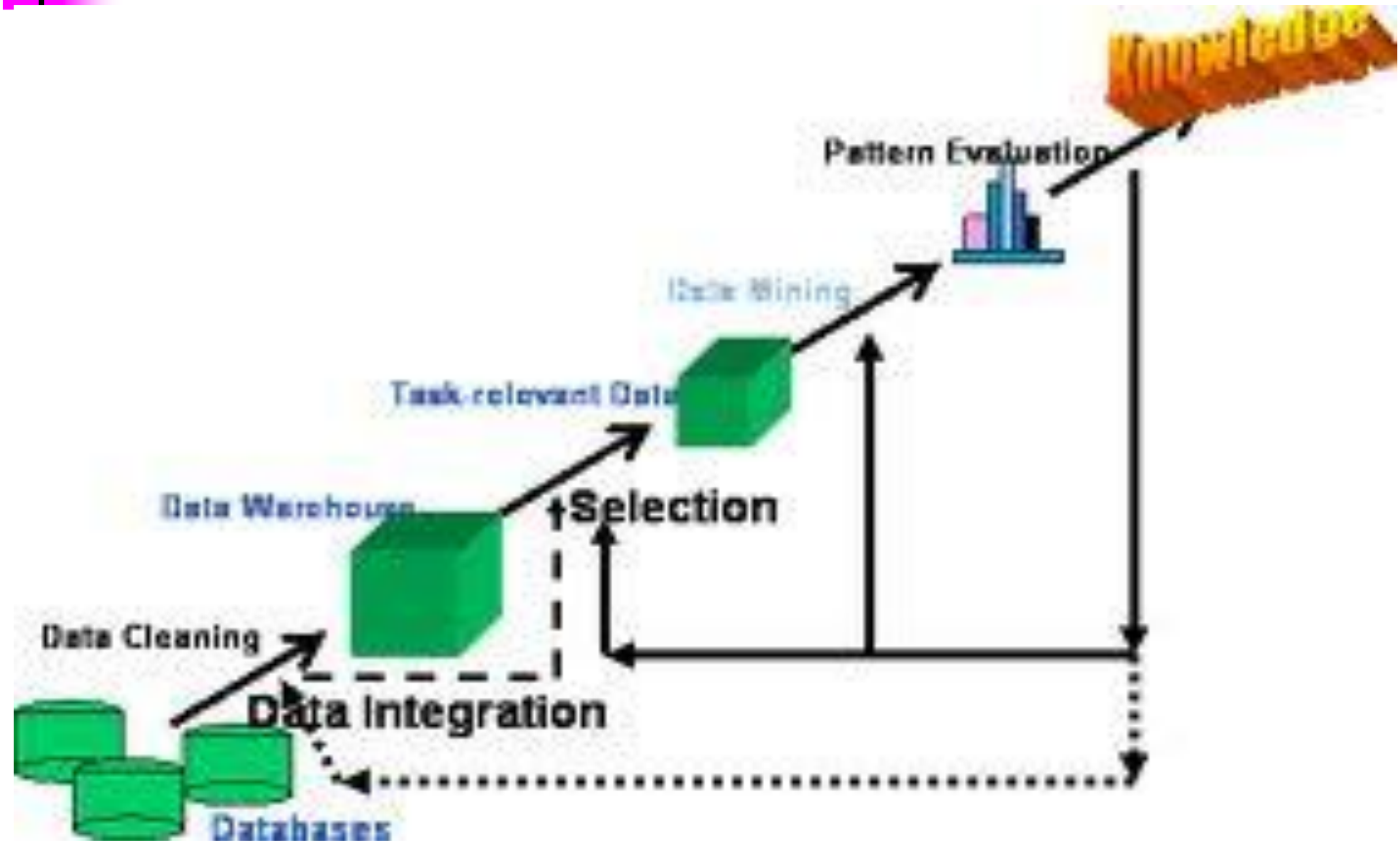
The Power and Promise of Analytics

- *Reducing Patient Readmission Rates (Medical data)*
- *big data to address patient care issues and to reduce hospital readmission rates.*
- *The focus on lack of follow-up with patients, medication management issues and insufficient coordination of care.*
- *Data is preprocessed to correct any errors and to format it for analysis.*

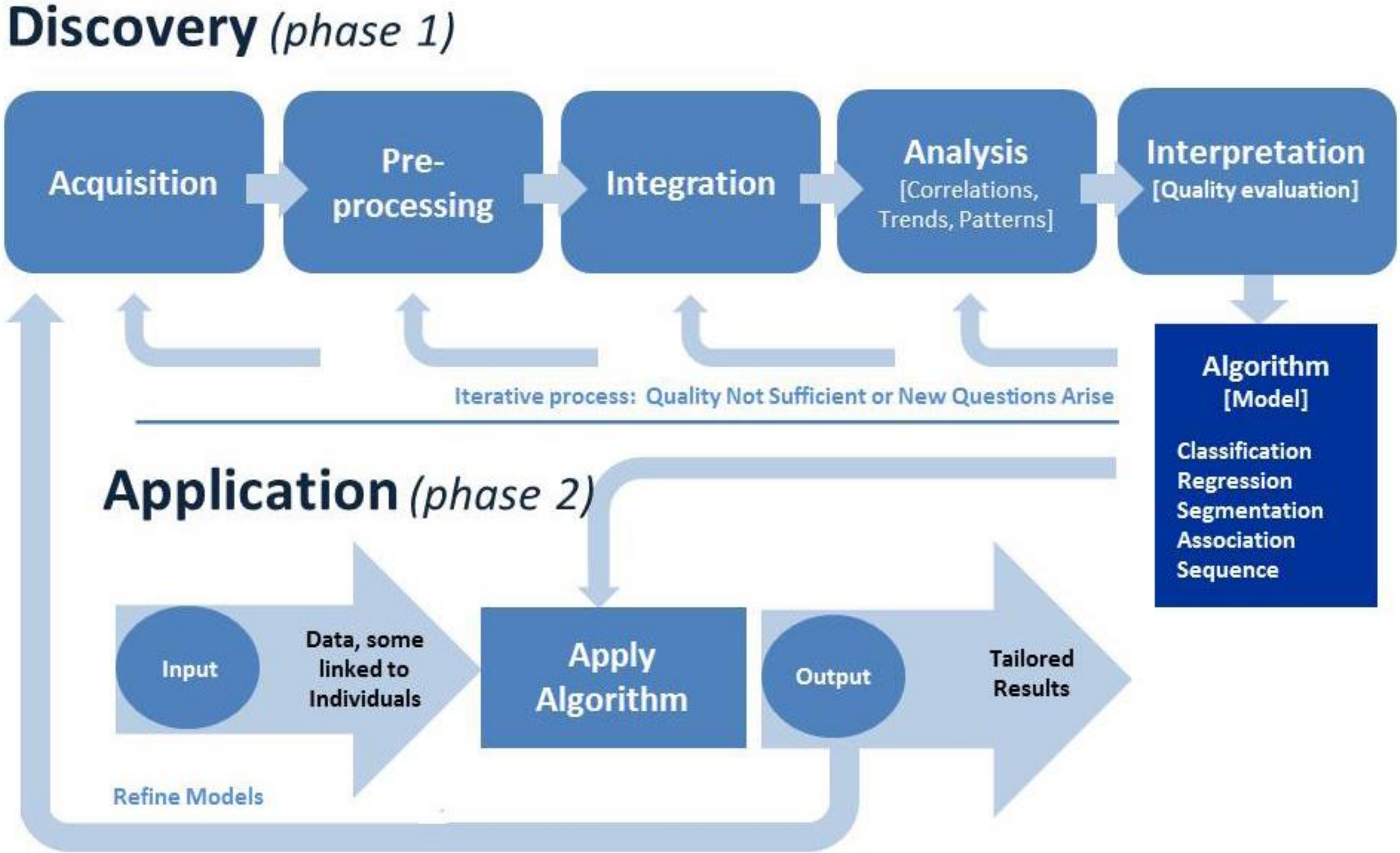
The Power and Promise of Analytics

- Analytics to Reduce the Student Dropout Rate (Educational Data)
- Analytics applied to education data can help schools and school systems better understand how students learn and succeed.
- Based on these insights, schools and school systems can take steps to enhance education environments and improve outcomes.
- Assisted by analytics, educators can use data to assess and when necessary re-organize classes, identify students who need additional feedback or attention.
- Direct resources to students who can benefit most from them.
- etc…..

Process of Data Analytics (KDD process)



The Process of Analytics



The Process of Analytics (Phase-1)

- This knowledge discovery phase involves
  – *gathering data to be analyzed.*

  – *pre-processing it into a* format that can be used.

  – consolidating (more certain) it for analysis, analyzing it to discover what it may reveal.

  – and interpreting it to understand the processes by which the data was analyzed and how conclusions were reached.

The Process of Analytics (Phase-1)

- *Acquisition –(process of getting something)*
- *Data acquisition involves collecting or acquiring data for analysis.*
- *Acquisition requires access to information and a mechanism for gathering it.*

The Process of Analytics (Phase-1)

- ***Pre-processing*** *–:*

- *Data is structured and entered into a consistent format that can be analyzed.*

- *Pre-processing is necessary if analytics is to yield trustworthy (**able to trusted**), useful results.*

- *places it in a standard format for analysis.*

The Process of Analytics (Phase-1)

- *Integration –:*
- *Integration involves consolidating data for analysis.*
  - *– Retrieving relevant data from various sources for analysis*
  - *– eliminating redundant data or clustering data to obtain a smaller representative sample.*
  - *– clean data into its data warehouse and further organizes it to make it readily useful for research.*
  - *– distillation into manageable samples.*

The Process of Analytics (Phase-1)

- *Analysis –; Knowledge discovery involves*

  – *searching for relationships between data items in a database, or exploring data in search of classifications or associations.*

  – *Analysis can yield descriptions (where data is mined to characterize properties) or predictions (where a model or set of models is identified that would yield predictions).*

  – *Analysis based on interpretation, organizations can determine whether and how to act on them.*

The Process of Analytics (Phase-1)

- *Interpretation –:*
  - *Analytic processes are reviewed by data scientists to understand results and how they were determined.*
  - *Interpretation involves retracing methods, understanding choices made throughout the process and critically examining the quality of the analysis.*
  - *It provides the foundation for decisions about whether analytic outcomes are trustworthy*

The Process of Analytics (Phase-1)

- *The product of the knowledge discovery phase is an algorithm. Algorithms can perform a variety of tasks:*
- *Classification algorithms categorize discrete variables (such as classifying an incoming email as spam).*
- *Regression algorithms calculate continuous variables (such as the value of a home based on its attributes and location).*
- *Segmentation algorithms divide data into groups or clusters of items that have similar properties (such as tumors found in medical images).*
- *Association algorithms find correlations between different attributes in a data set (such as the automatically suggested search terms in response to a query).*
- *Sequence analysis algorithms summarize frequent sequences in data (such as understanding a DNA sequence to assign function to genes and proteins by comparing it to other sequences).*

The Process of Analytics (Phase-2)

- *Application*
  - *Associations discovered amongst data in the knowledge phase of the analytic process are incorporated into an algorithm and applied.*

  - *for example, classify individuals according to certain criteria, and in doing so determine their suitability to engage in a particular activity.*

  - *In the application phase organizations reap (collect) the benefits of knowledge discovery.*

  - *Through application of derived algorithms, organizations make determinations upon which they can act.*

Goals for Analytics Guidance

- *Recognize and reflect the two-phased nature of analytic processes.*
  - *Traditional methods of data analysis usually involve identification of a question and analysis of data in search of answers to that question.*
  - *Use of advanced analytics with big data upends that approach by making it possible to find patterns in data through knowledge discovery. Rather than approach data with a predetermined question.*
  - *The results of this analysis may be unexpected.*
  - *Moreover, this research may suggest further questions for analysis or prompt exploration of data to identify additional insights, through iterative analytic processing.*

Goals for Analytics Guidance

- *Provide guidance for companies about how to establish that their use of data for knowledge discovery is a legitimate business purpose.*
  - allow for processing of data for a legitimate business purpose, but provide little guidance about how organizations establish legitimacy and demonstrate it to the appropriate oversight body.

  - Guidance for analytics would articulate the criteria against which legitimacy is evaluated and describe how organizations demonstrate to regulators or other appropriate authorities the steps they have taken to support it.

Goals for Analytics Guidance

- ***Emphasize the need to establish accountability through an internal privacy program that relies upon the identification and mitigation of the risks the use of data for analytics may raise for individuals.***

  – *how fair information practices are applied, it is important that organizations implement an internal privacy program that involves credible assessment of the risks data processing may raise.*

  – *Risk mitigation may involve de-identification and pseudo- nominisation of data, as well as other controls to prevent re-identification of the original data subject.*

Goals for Analytics Guidance

- *Take into account that analytics may be an iterative process using data from a variety of sources.*
  - analytics is not necessarily a linear process. Insights yielded by analytics may be identified as flawed or lacking, and data scientists may in response re-develop an algorithm or re-examine the appropriateness of the data for its intended purpose and prepare it for further analysis.

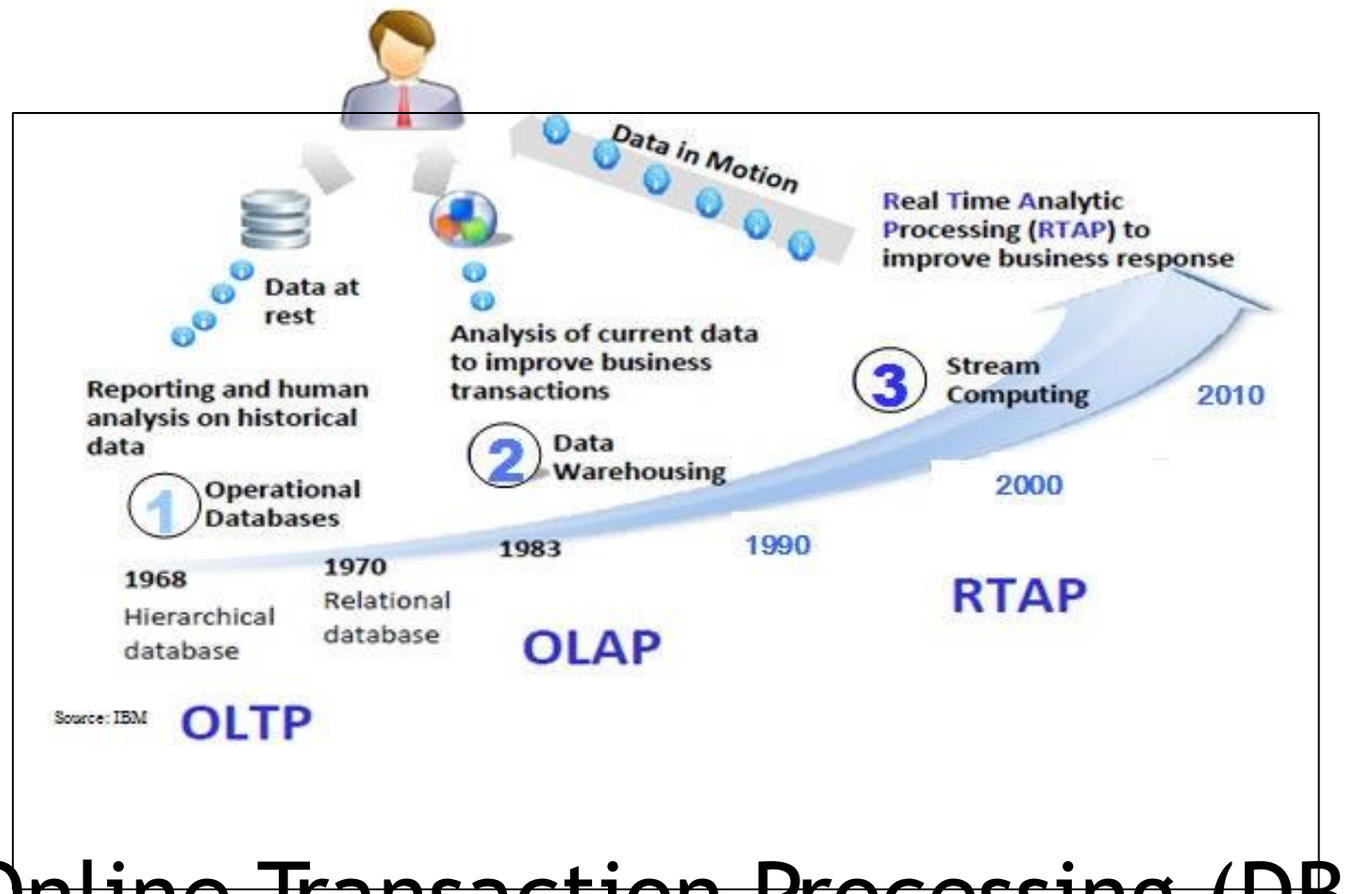
  - Knowledge discovery may reveal that data could provide additional insights, and researchers may choose to explore them further. Data used for analytics may come from an organization's own stores, but may also be derived from public records.

  - Data entered into the analytic process may also be the result of earlier processing.

Data Analytics

- *Conclusion*
  - *Analytics and big data hold growing potential to address longstanding issues in critical areas of business, science, social services, education and development. If this power is to be tapped responsibly, organizations need workable guidance that reflects the realities of how analytics and the big data environment work.*
  - *Such guidance must be grounded on the consensus of*
    - *international stakeholders.*
    - *data protection authorities and regulators.*
    - *business leaders.*
    - *academics and experts.*
    - *and civil society.*
- *Thoughtful, practical guidance can release and enhance the power of data to address societal questions in urgent need of answers.*
- *A trusted dialogue to arrive at that guidance will be challenging, but cannot wait.*

Harnessing Big Data

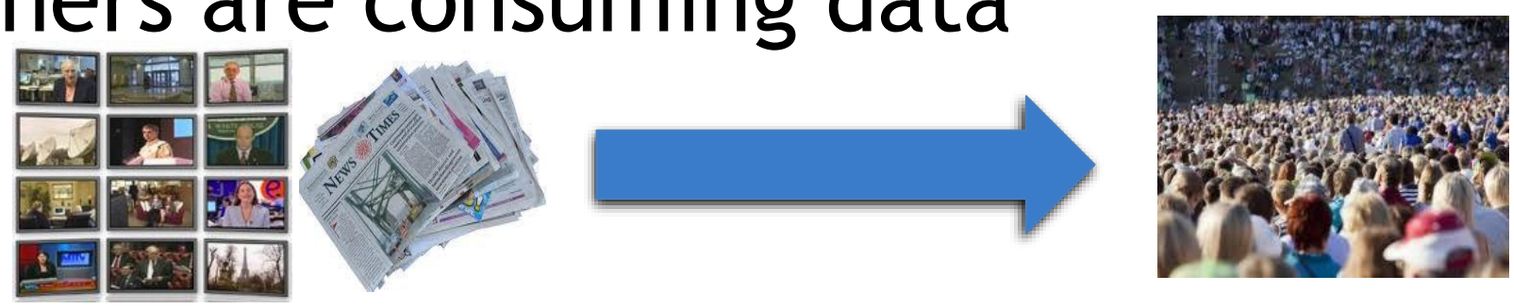


- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

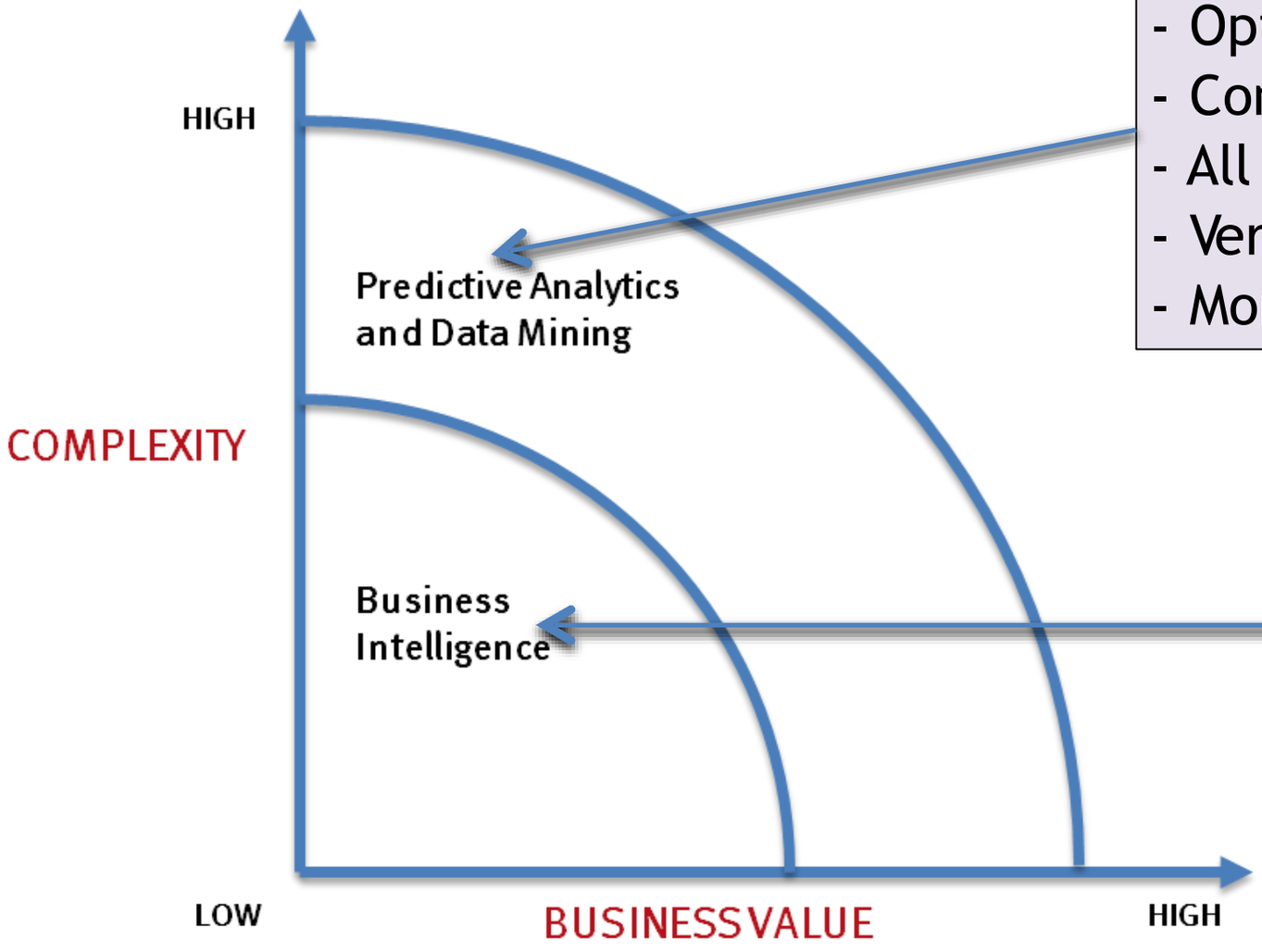
Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data

15

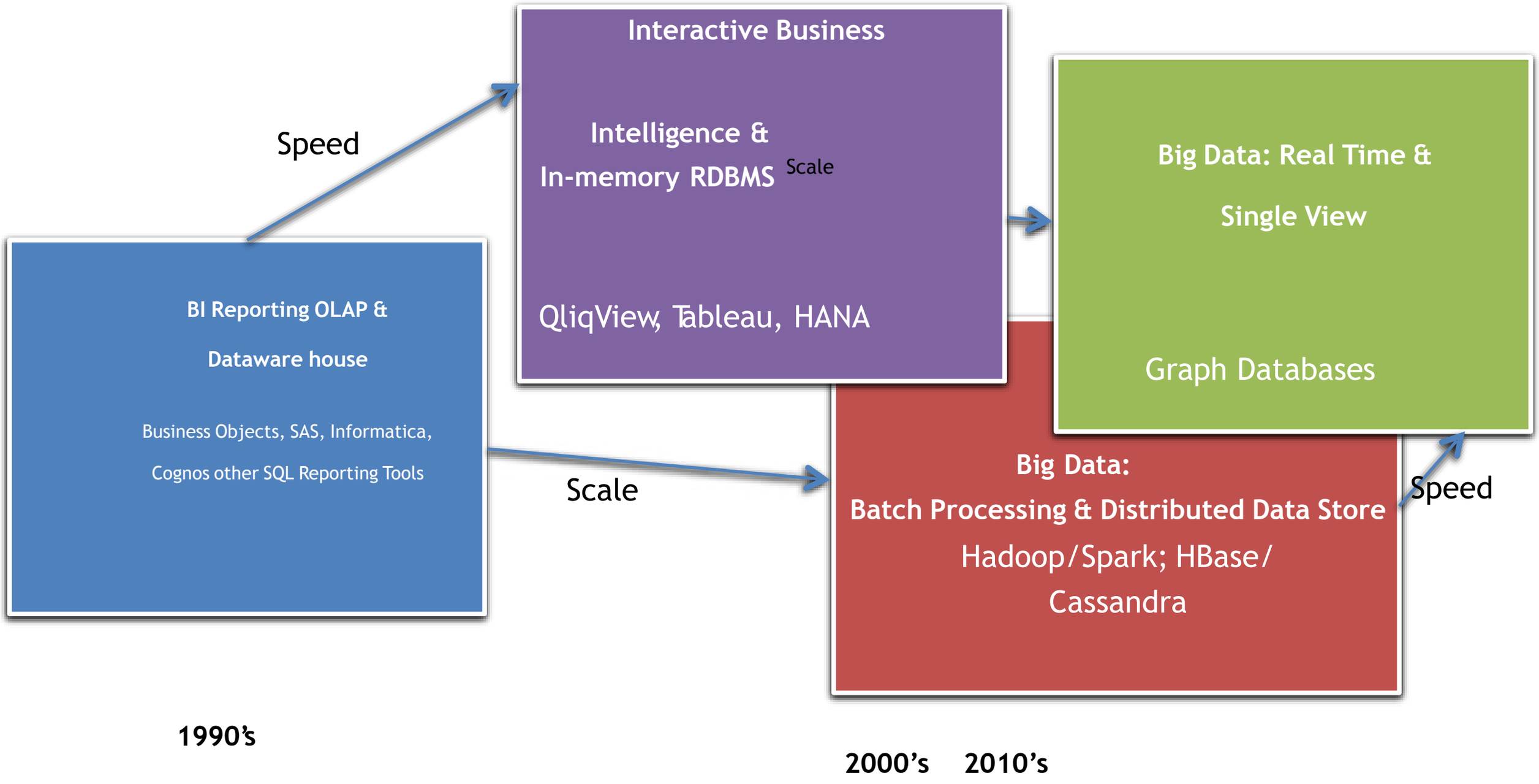
What's driving Big Data



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

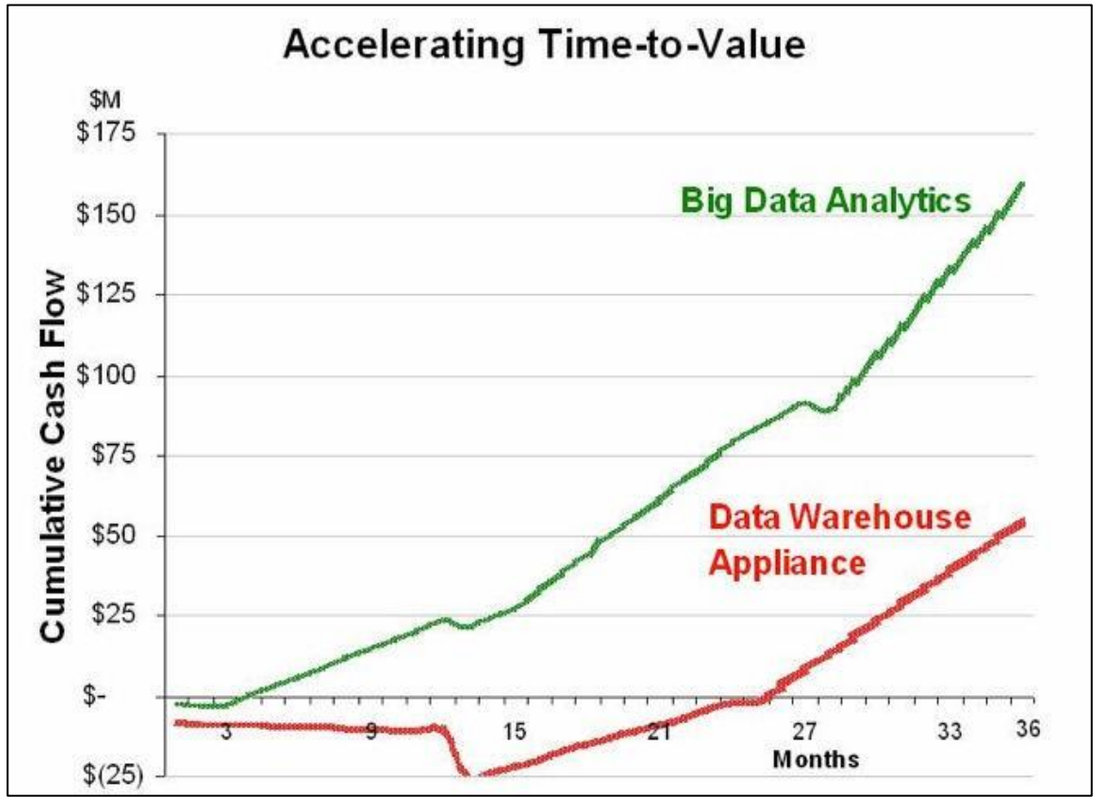
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

THE EVOLUTION OF BUSINESS INTELLIGENCE



Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- 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 Analytics

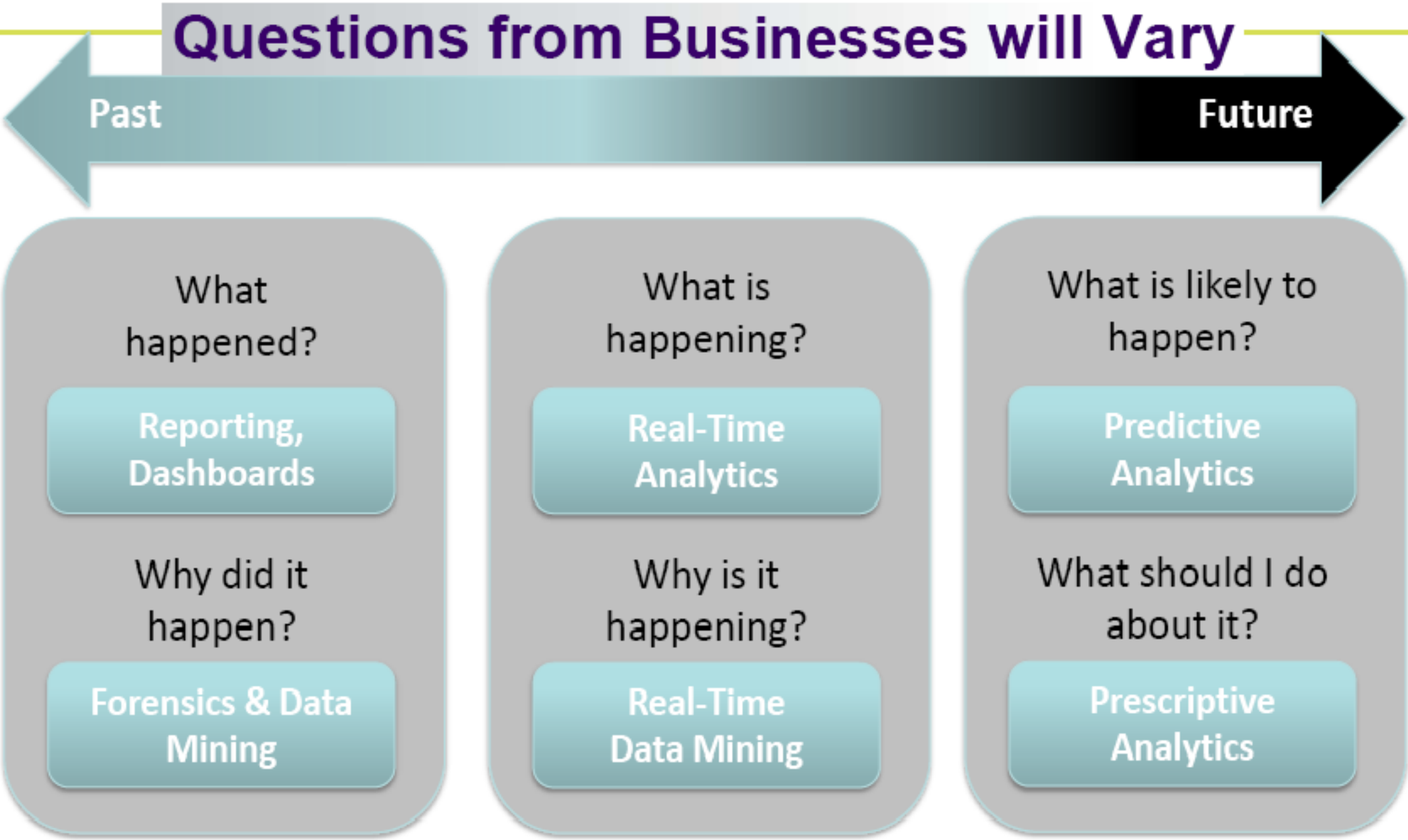
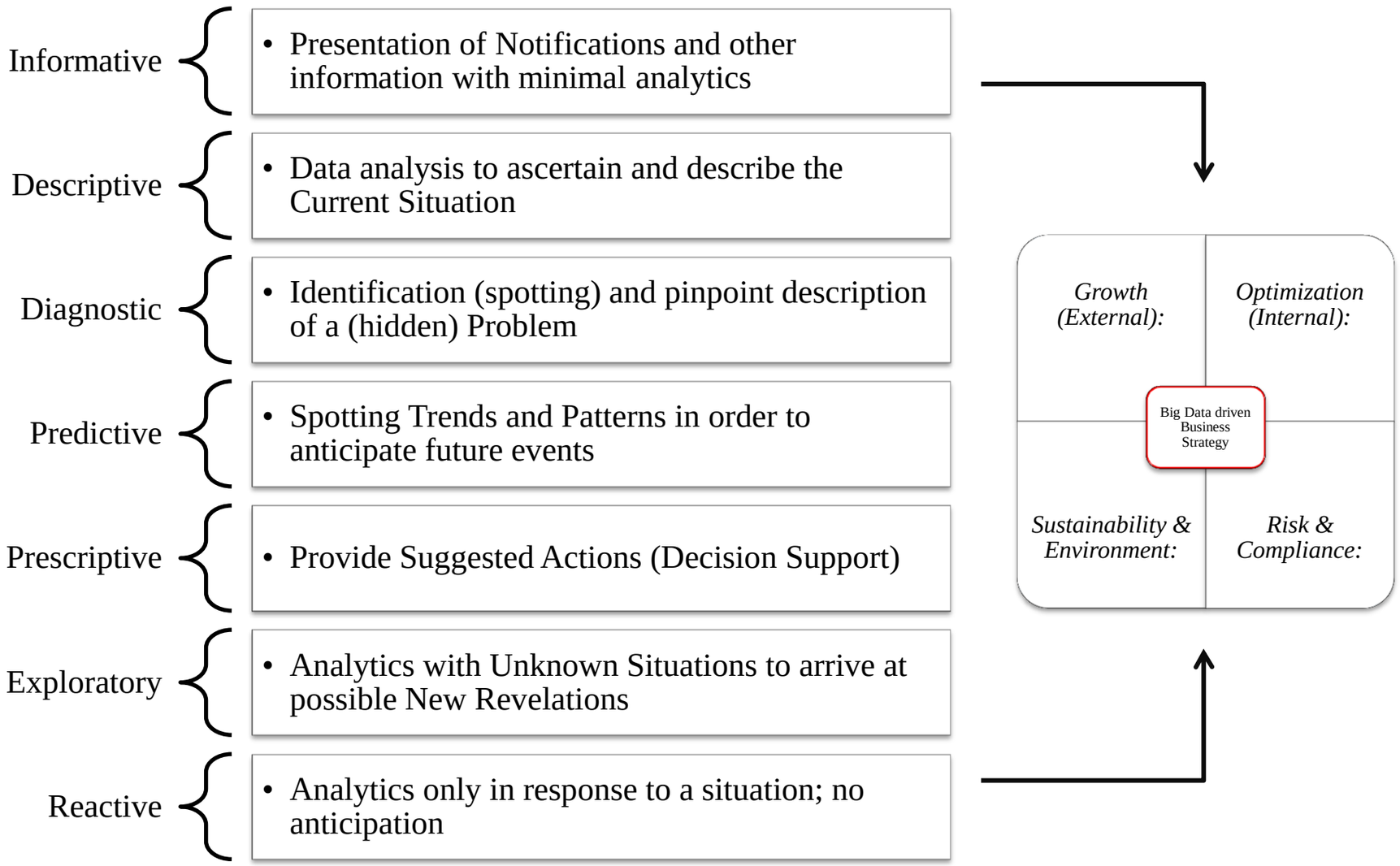


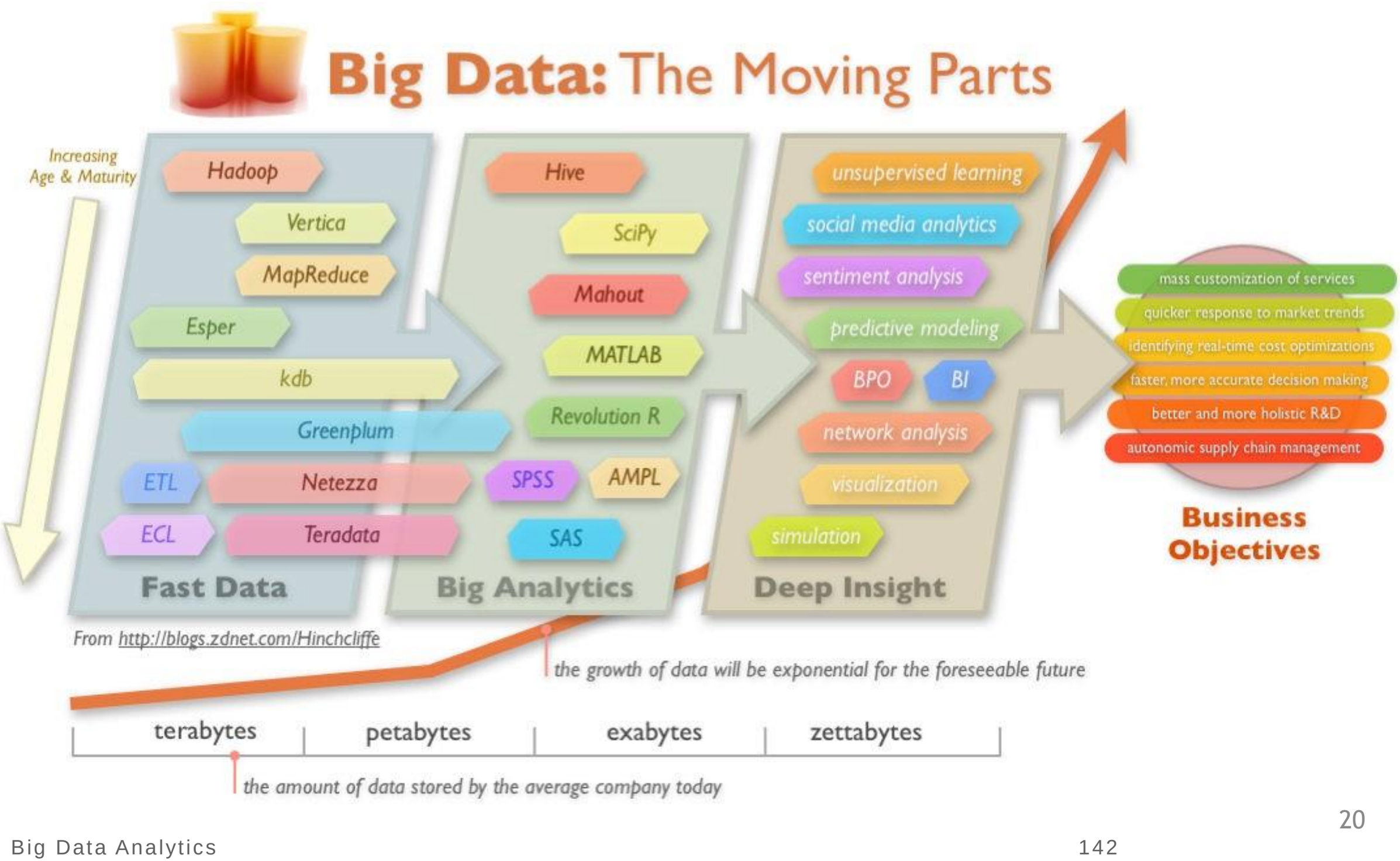
Figure 3.11: Various Analytics Categories provide Agile Business Values (and form basis of Business Strategies)



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

Big Data Technology



Big Data Landscape

Vertical Apps

PREDICTIVE POLICING

bloomreach

MEDIDATA

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel collective[i]

Media Science TURN DataXu

Business Intelligence

ORACLE | Hyperion

SAP Business Objects Cognos

Microsoft | Business Intelligence

IBM COGNOS birst

MicroStrategy

Autonomy bime

QlikView

ClearLab GoodData

Analytics and Visualization

+ableau Palantir

OPERA

METAMARKETS datameer centrifuge

TERADATA ASTER

SAS TIBCO

panopticon

Datameer

platfora GIRRO

alteryx visual.ly AYATA

Data As A Service

factual

INFOCHIMPS LOQATE

GNIP DATAMARKET INRIX LexisNexis

Analytics Infrastructure

Hortonworks VERTICA MAPR

cloudera INFOBRIGHT ParAccel

EMC GREENPLUM

NETEZZA kognitio

ORTSTRA SQL

Operational Infrastructure

Couchbase 10gen

TERADATA HADAPT

TERRACOTTA VoltDB

MarkLogic INFORMATICA

Infrastructure As A Service

amazon web services

Windows Azure

infochimps

Google BigQuery

Structured Databases

ORACLE MySQL

SQL Server PostgreSQL

IBM DB2. SYBASE

mongoDB

Technologies

hadoop hadoop MapReduce mahout HBASE Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Landscape

Infrastructure

NoSQL Databases
10gen, DataStax, basho, Couchbase, Cloudant, HYPERTABLE, Neo4j, SCYLLA, mapr graph

NewSQL Databases
MarkLogic, memsql, paradigm4, SQLFire, DRAMPSCALE, VoltDB, NUODB

Hadoop Related
cloudera, HADAPT, Hortonworks, MAPR, infochimps, HSTREAMING, VertiCloud, Zettaset, IBM, MORTAR, Microsoft, GREENPLUM, amazon, Qubole, aginity

MPP Databases
VERTICA, kognitio, ParAccel, GREENPLUM, TERADATA, NETEZZA, InfiniDB, SQL Server

Management / Monitoring
DATADOG, oceansync, StackIQ, OUTER THOUGHT, boundary

Storage
inktank, panasas, APPLIDATA, Compuverde, nimblestorage, Cleversafe

Cluster Services
LexisNexis, HPCC Systems, Acunu

Security
Stormpath, iMPERVA, TRACEVECTOR, codefortytwo software, DATAGUISE

Crowdsourcing
CROWD COMPUTING systems, CrowdFlower, amazon, mechanicalturk

Collection / Transport
aspera, nodeable

Analytics

Analytics Solutions
platfora, Palantir, PERVASIVE, Datameer, KARMASPHERE, DataMeer, CRYSTAL KNOWS, datasphere, PRECOG

Statistical Computing
SKYTREE, sense, ATHEWING, REVOLUTION ANALYTICS, MATLAB, sas, SPSS

Sentiment Analysis
GENERAL SENTIMENT, Owen, crimson hexagon

Location / People / Events
RapLeaf, Flipt op, Recorded Future, PlaceIQ, RADIUS

Real-Time
PerScreen, feedzai

Data Visualization
Quid, visual.ly, centrifuge, ACTUATE, Kitenga, metaLayer, Ayasdi, ClearStory, tableau, ISS, Quantum4D

Social Media
bitly, Tracx, bluefin, simplereach, Datamin r

Analytics Services
THINK BIG, McKinsey&Company, accenture, OPERA, Mu Sigma

Big Data Search
elasticsearch, Autonomy

IT Analytics
splunk, sumologic

Crowdsourced Analytics
DataKind, kaggle

SMB Analytics
sumAll, RJMetrics, custora

Applications

Ad Optimization
DataXu, m6d, MediaMath, Match, bluekai, rocketfuel, thetradedesk, TURN, 33across

Publisher Tools
VISUAL IQ, Yieldex, yieldbot, CONTENT PLANT

Marketing
CLICKFOX, RETENTION SCIENCE, Sailthru, LATTICE ENGINES, bloomreach, TellApart

Industry Applications
NEXT BIG THING, KNEWTON, finance, wonga, numberFire, MileSense, BILLGUARD, Climate Solutions, Bloomberg

Application Service Providers
CONTINUITY, collective[i]

Data Sources

Data Marketplaces
factual, DataMarket, Windows Azure Marketplace

Data Sources
premise, DATASIFT, knoema, GNIP, infochimps

Personal Data
BASIS, Withings, JAWBONE, RunKeeper, fitbit

Cross Infrastructure / Analytics
SAP, sas, IBM, Google, Microsoft, vmware, amazon, 1010data, ORACLE, METAMARKETS, TERADATA, NetApp

Open Source Projects

Framework
Hadoop, HDFS

Query / Data Flow
Hive, Pig

Data Access
riak, mongoDB, Cassandra, SciDB, HBASE, CouchDB, Sqoop

Coordination / Workflow
ZooKeeper, talend, Oozie

Real - Time
Storm

Statistical Tools
R, SciPy

Machine Learning
Mahout

Cloud Deployment

Matt Turck (@mattturck) and Shivon Zilis (@shivonz)

Big Data Landscape 2016 (Version 3.0)

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, NUODB, memsql, splice MACHINE, VOLTDB, MariaDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, An HP Company, NETEZZA, an IBM Company, Actian, kognitio, EXASOL, dremio

Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, Put potential to work:, MuleSoft, snapLogic, BedrockData, xplenty

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk>, DATADOG, Procana, DRIVEN, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

App Dev
apigee, CASK, Keen IO, Typesafe, DRIVEN

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, Bottlenose, interana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine DATA LABS, plotly, ARIMO, MODE, dataiku, tonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Qlik, looker, Roambi, SISENSE, ZOOMDATA, datorama, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, salesforce Wave Analytics, birst, GoodData, kyvos insights, platfora, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach

Real-Time
amazon web services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, H2O.ai, amazon web services, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, ViSENZE, PredictionIO, glowfi.sh

Speech & NLP
NarrativeScience, NUANCE, WolframAlpha, semanticmachines, Gridspace, ARRIA, api.ai, cortical.io, maluuba, MindMeld, iDIBON, yseop

Horizontal AI
IBM Watson, Cortana., sentient technologies, VIV, nervana SYSTEMS, vicarious, nara, Numenta, HyperScience, SI SCALED INFERENCE, Descartes Labs, clarifai, DEXTRO, Geometric Intelligence, MetaMind

Search
Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
OPERA SOLUTIONS, Mu Sigma DO THE MATH, EXL analytics, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumall, Airtable, retention SCIENCE, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, 6sense, QUANTIFIND, ACTIONIQ, fusemachines, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, ClickFox, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, eBrevia, PREMONITION

Ad Optimization
AppNexus, criteo, MediaMath, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

Security
CYLANCE, CounterTack, cybereason, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, siftscience, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
facebook, Clara, KASISTO, lumiata

Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
Affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, Payoff, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

Education / Learning
KNEWTON, Clever, declara, PANORAMA EDUCATION, knowre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics
amazon web services, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA An HP Company, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, Spark, MESOS, TEZ, Apache Kylin, Flink, CDAP

Query / Data Flow
SLAMDATA, HIVE, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, accumulo, mongoDB, CLOUDERA, cascading, SciDB, kafka, CouchDB, riak, OPENTSDB, nifi

Coordination
talend, OOZIE, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
ScalaLab, NumPy, SciPy

Machine Learning
mllib, mahout, Apache SINGA, Aerosolve, MADlib, Caffe, torch, CNTK, TensorFlow, learn, FeatureFu, WEKA, VELES, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practicefusion, fitbit, VALIDIC, Withings, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, StockTwits, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS TECHNOLOGY, Sense 360

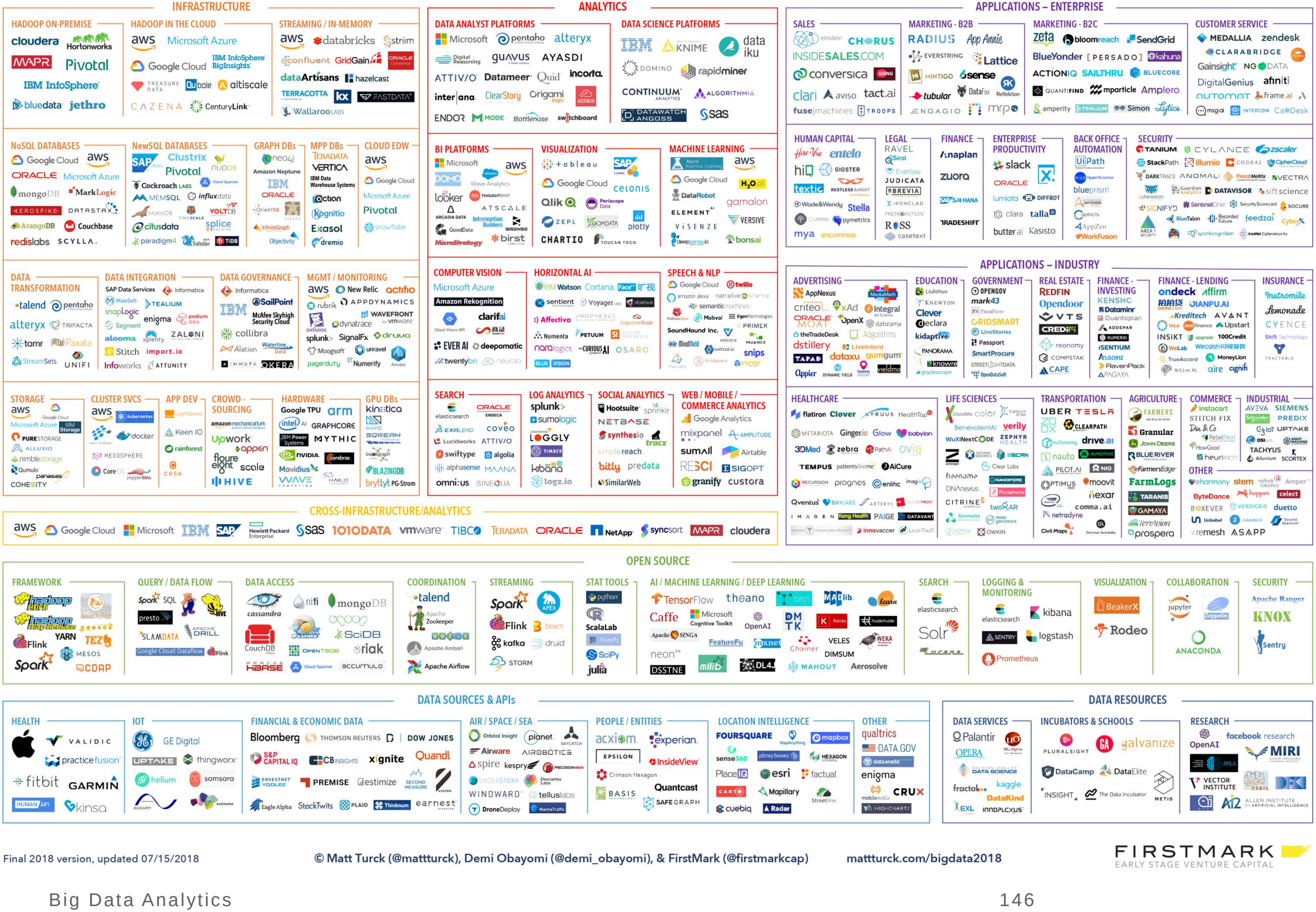
Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

Last Updated 3/23/2016

FIRSTMARK

BIG DATA & AI LANDSCAPE 2018



Final 2018 version, updated 07/15/2018

mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

DATA & AI LANDSCAPE 2019

July 16, 2019 - FINAL 2019 VERSION

mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

2020AI GERMAN STARTUP LANDSCAPE

247 AI STARTUPS

CONTRIBUTORS:

DIGITAL+ PARTNERS · EARLYBIRD · e CAPITAL ENTREPRENEURIAL PARTNERS · Google Cloud · High-Tech Gründerfonds · HV · LAKE STAR · NVIDIA · Si · UVC UNTERNEHMERTUM VENTURE CAPITAL

ENTERPRISE INTELLIGENCE

Audio Data

melodrive · Aaron.ai · PRECIRE · audEERING · audatic

Visual Data

Mobius Labs · NAVVIS · 3dvisionlabs · IDnow · DEEP VA · terraloupe · VISION IMPULSE · dyo deevio · EyeEm

Internal Data

cobrainer · keeeb · LANA · inspirient · KENDAXA GROUP

Market Data

scoutbee · pricefx · LYTICS · MAPEGY · Vencortex · NEUROFLASH

Sensor Data

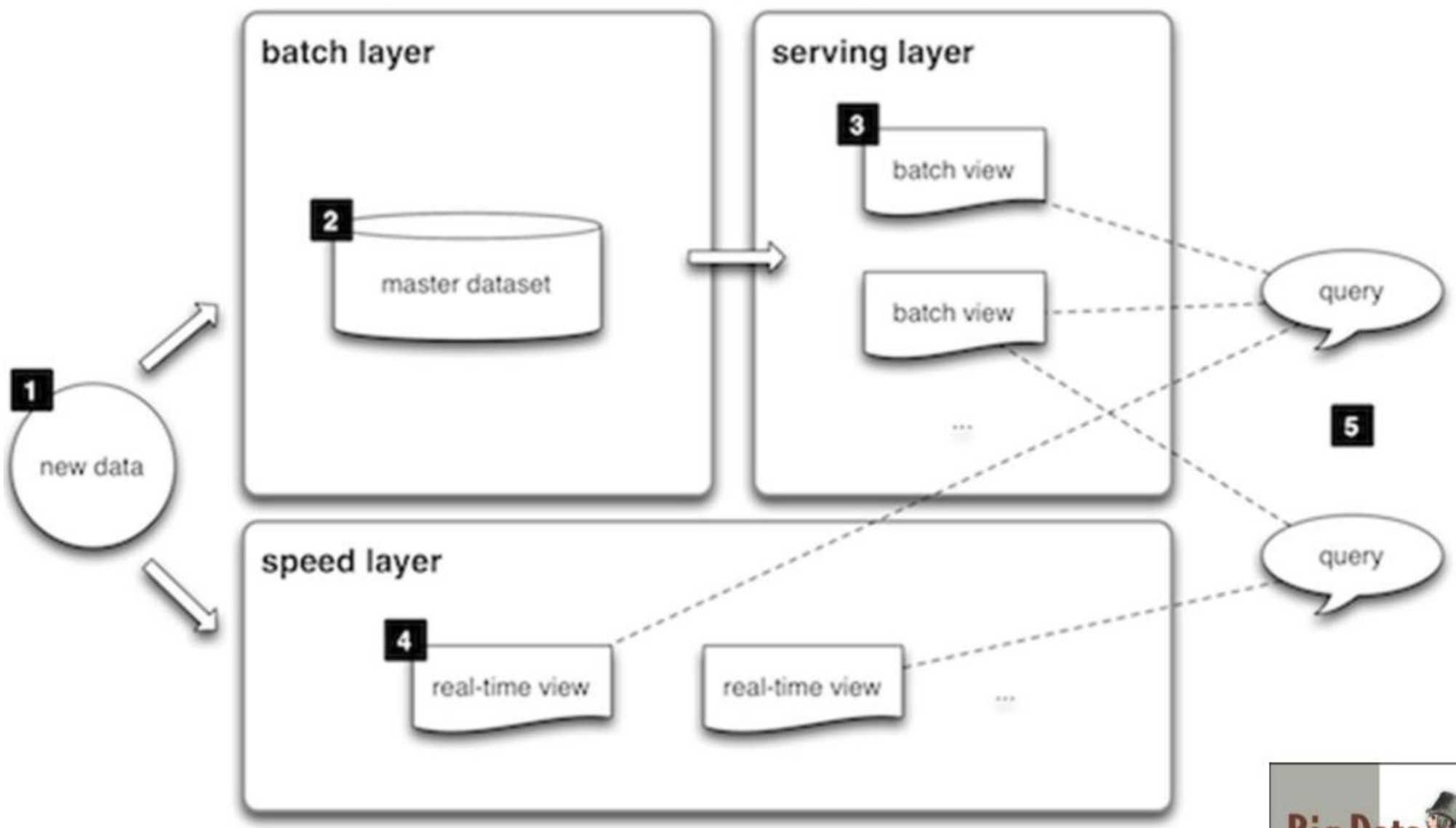
maiot · KONUX · MOTIONTAG · ZOLITRON TECHNOLOGY · TERAKI

Geo Data

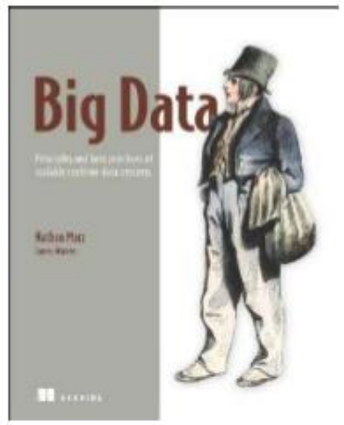
GEOSPIN · Eagle AI · Targomo

Text Data

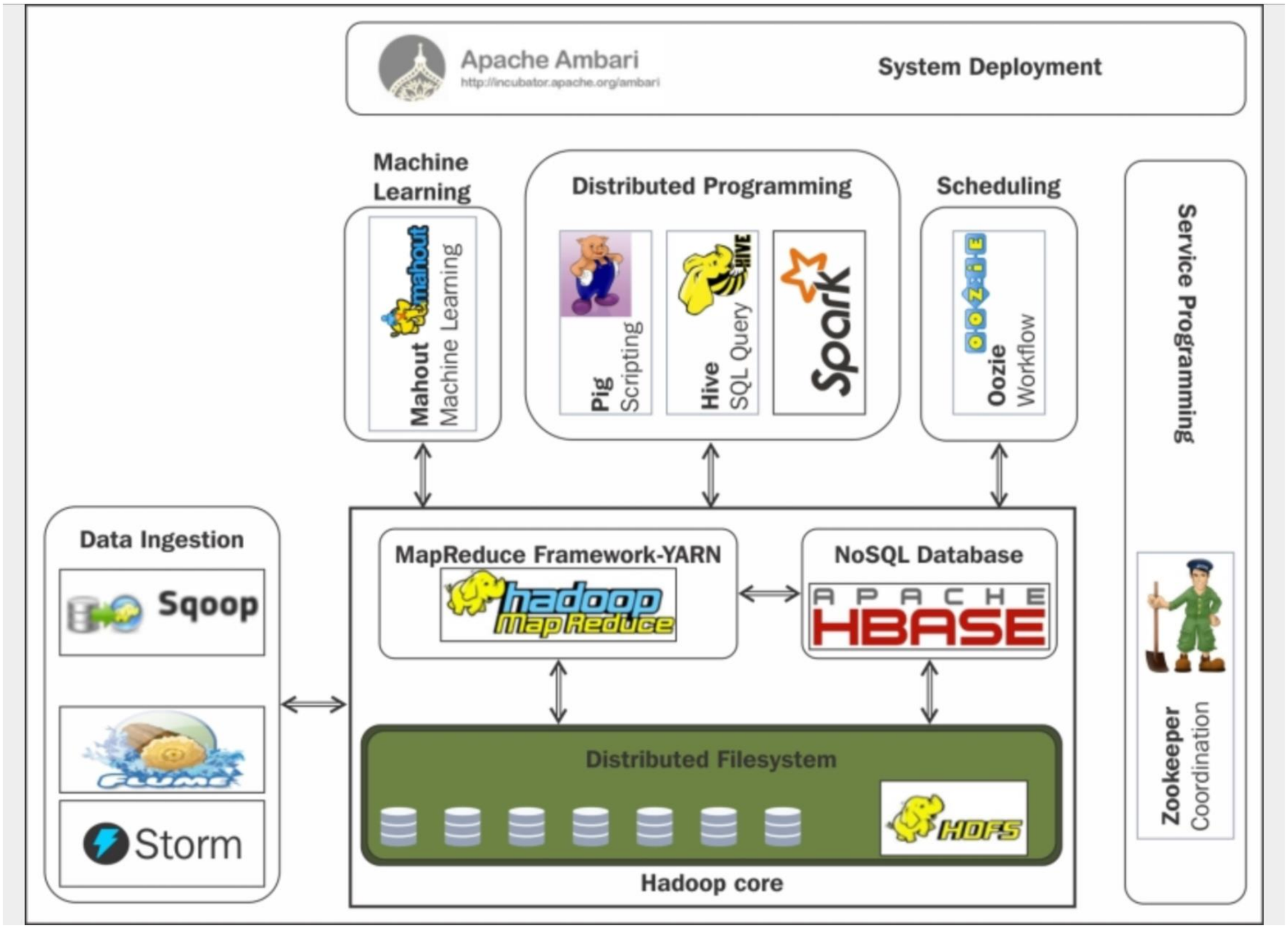
THINGS THINKING · Hypatos · lengoo · curiosity.ai · fileee · gini · LATERAL · Pyoneer · EVANA


batch layer
serving layer
3
batch view
2
master dataset
batch view
query
1
new data
...
5
query
speed layer
4
real-time view
real-time view
...

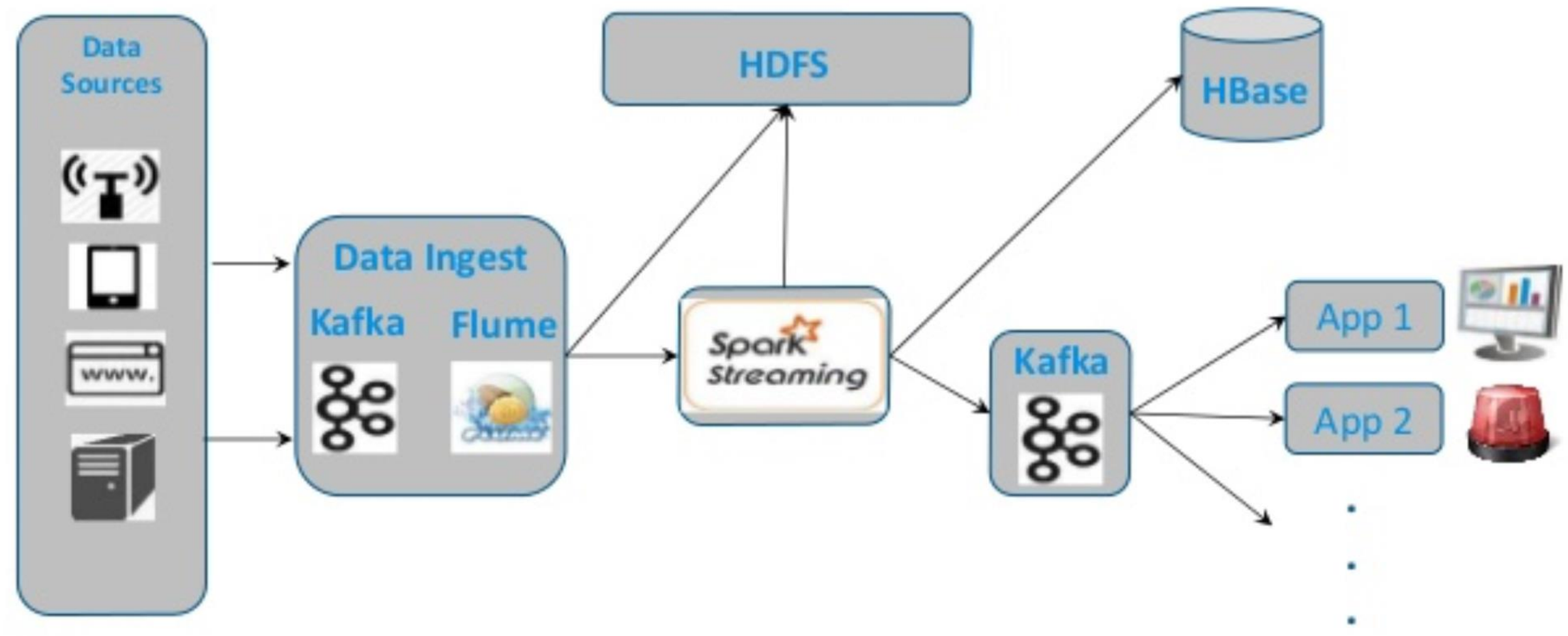
Source:


Big Data

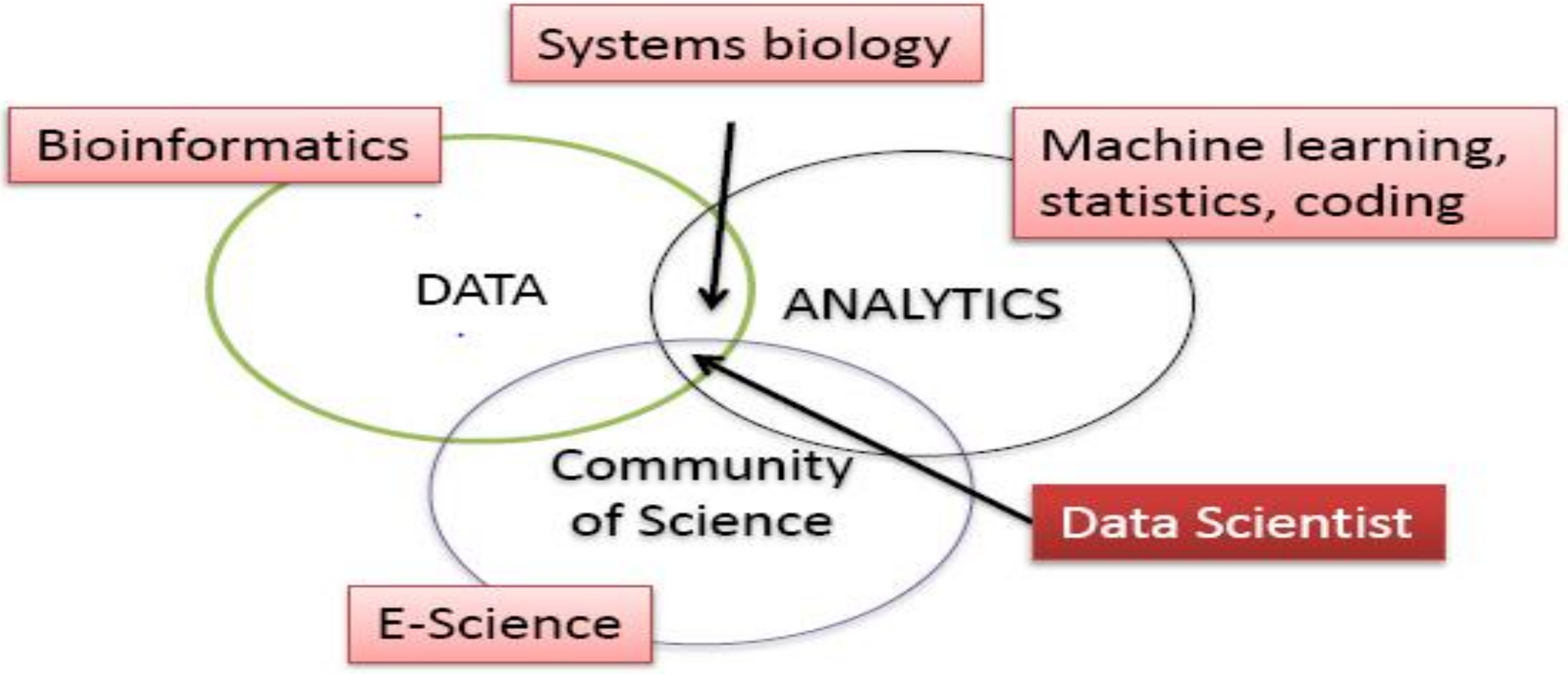
Hadoop Ecosystem



Stream Processing Architecture



Data Scientist



Data Scientist

- *Data scientist includes*
- *Data capture and Interpretation*
- *New analytical techniques*
- *Community of Science*
- *Perfect for group work*
- *Teaching strategies*

Data scientist

- *Data scientists require a wide range of skills:*
  - *Business domain expertise and strong analytical skills*
  - *Creativity and good communications.*
  - *Knowledgeable in statistics, machine learning and data visualization*
  - *Able to develop data analysis solutions using modeling/analysis methods and languages such as Map-Reduce, R, SAS, etc.*
  - *Adept at data engineering, including discovering and mashing/blending large amounts of data.*

Data scientist

- *Data scientists use an investigative computing platform*
  - *to bring un-modeled data.*
  - *multi-structured data, into an investigative data store for experimentation.*
  - *deal with unstructured, semi-structured and atructured data from various source.*

Data scientist

- *Data scientist helps broaden the business scope of investigative computing in three areas:*
- *New sources of data – supports access to multi-structured data.*
- *New and improved analysis techniques – enables sophisticated analytical processing of multi-structured data using techniques such as Map-Reduce and in-database analytic functions.*
- *Improved data management and performance – provides improved price/performance for processing multi-structured data using non-relational systems such as Hadoop, relational DBMSs, and integrated hardware/software.*

Role of Data scientist

- *Goal of data analytics is the role of data scientist*
  - *Recognize and reflect the two-phased nature of analytic processes.*
  - *Provide guidance for companies about how to establish that their use of data for knowledge discovery is a legitimate business purpose.*
  - *Emphasize the need to establish accountability through an internal privacy program that relies upon the identification and mitigation of the risks the use of data for analytics may raise for individuals.*
  - *Take into account that analytics may be an iterative process using data from a variety of sources.*

Current trend in Big data Analytics

- Iterative process (Discovery and Application)
- In general:
- Analyze the unstructured data (Data analytics)
- development of algorithm (Data analytics)
- Data Scrub (Data engineer)
- Present structured data (relationship, association)
- Data refinement (Data scientist)
- Process data using distributed engine. E.g. HDFS (S/W engineer) and write to No-SQL DB (Elasticsearch, Hbase, MangoDB, Cassandra, etc)
- Visual presentation in Application sw.
- QC verification.
- Client release.

Thank you !!!