National Workshop on
Recent Trends in Big Data Analytics
February 27, 2016

BIG Data - > Big Decisions

By M.Saravanan

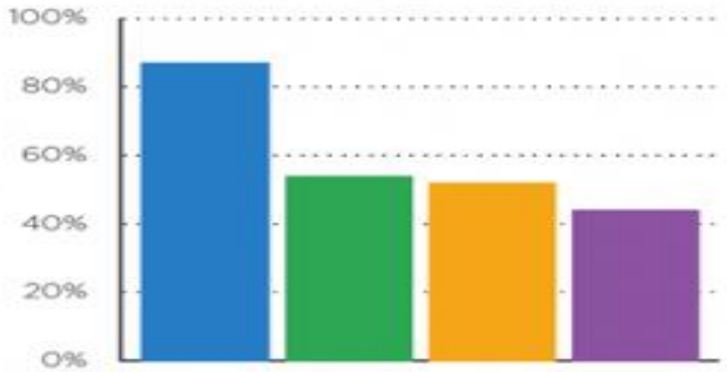
Big Data, Big Decisions

Capturing data is only the first step. Making the most of it takes the right capabilities, processes and strategy.

Are you drowning in data without a life vest? As gigabytes give way to terabytes and petabytes, business leaders have access to more raw performance data than ever before. Many organizations have so much data that they don't know how to harness it for business value.

Here's how building data management capabilities can help your organization turn big data into better business decisions.

BRIDGING THE GAP



87% of global senior managers and executives agree that better decision-making would improve their organization's financial performance...

...And **54% say** that a better ability to analyze data would improve decision-making.

Yet **52% of C-level** execs and **44% of other senior execs** have dismissed data because they didn't understand it.

From Big Data to Better Decisions

Data Flood

- 5x Companies are more likely to exceed expectations when they use analytical tools than their competitors
- 80% of the data collected by companies is wasted
-  of CEOs have the information that they need /want in their reports

Potential For Companies

- 10% increase in usability
- $ 2.01 billion increase annually
- 44 Trillion gigabytes of data created annually by 2020
- 6% Most profitable than their competitors
- 5% More productive compare to their competitors

Future of Business Intelligence

Wearable tech will increase from 14 billion to 32 billion by 2020

A greater dependence on Machine Learning

Increase in the need for Data Scientist

BIG DATA CULTURE CLASH: 10 KEY ISSUES & ACTIONS

1. Markets and customer behaviors change too fast for legacy decision-making to keep pace.

2. Insights from big data will separate business winners from losers.

3. "Build Data-Driven Culture" was voted the number one priority of *The Wall Street Journal's* CIO Network.

4. Collect data for *all* aspects of your business and build analytics to answer critical business questions.

5. Hire a Chief Data Officer to drive big data infrastructure agility *and* culture.

6. Hire (or train) multi-skilled generalists — people who "get" the tech *and* the business unit mission.

7. Embed or align technology analysts with the business units they support.

8. Analyze *all* your data — most companies look at only up to 15%.

9. Look for long-term trends; avoid tunnel vision.

10. CIOs can champion the value of data-driven decisions by:
    - Creating plain-English, customer-centric examples
    - Instilling confidence that data is sound
    - Guarding against "confirmation bias"
    - Exploring *all* potential outcomes of an analysis

Cover Story Investment Guide 2016 Forbes on Big Data: 20 Mind-Boggling Facts

Big data is not a fad. We are just at the beginning of a revolution that will touch every business and every life on this planet. Loads of people are still treating the concept of big data as something they can choose to ignore — when actually, they're about to be run over by the steamroller that is big data.

1.The data volumes are exploding, more data has been created in the past two years than in the entire previous history of the human race.

2.Data is growing faster than ever before and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet.

3.By then, our accumulated digital universe of data will grow from 4.4 zettabyets today to around 44 zettabytes, or 44 trillion gigabytes.

4.Every second we create new data. For example, we perform 40,000 search queries every second (on Google alone), which makes it 3.5 searches per day and 1.2 trillion searches per year.

5.In Aug 2015, over 1 billion people used Facebook in a single day.

6.Facebook users send on average 31.25 million messages and view 2.77 million videos every minute.

7.We are seeing a massive growth in video and photo data, where every minute up to 300 hours of video are uploaded to YouTube alone.

8.In 2015, a staggering 1 trillion photos will be taken and billions of them will be shared online. By 2017, nearly 80% of photos will be taken on smart phones.

9.This year, over 1.4 billion smart phones will be shipped – all packed with sensors capable of collecting all kinds of data, not to mention the data the users create themselves.

10.By 2020, we will have over 6.1 billion smartphone users globally (overtaking basic fixed phone subscriptions).

11.Within five years there will be over 50 billion smart connected devices in the world, all developed to collect, analyze and share data.

12.By 2020, at least a third of all data will pass through the cloud (a network of servers connected over the Internet).

13.Distributed computing (performing computing tasks using a network of computers in the cloud) is very real. Google GOOGL +2.17% uses it every day to involve about 1,000 computers in answering a single search query, which takes no more than 0.2 seconds to complete.

14.The Hadoop (open source software for distributed computing) market is forecast to grow at a compound annual growth rate 58% surpassing $1 billion by 2020.

15.Estimates suggest that by better integrating big data, healthcare could save as much as $300 billion a year — that's equal to reducing costs by $1000 a year for every man, woman, and child.

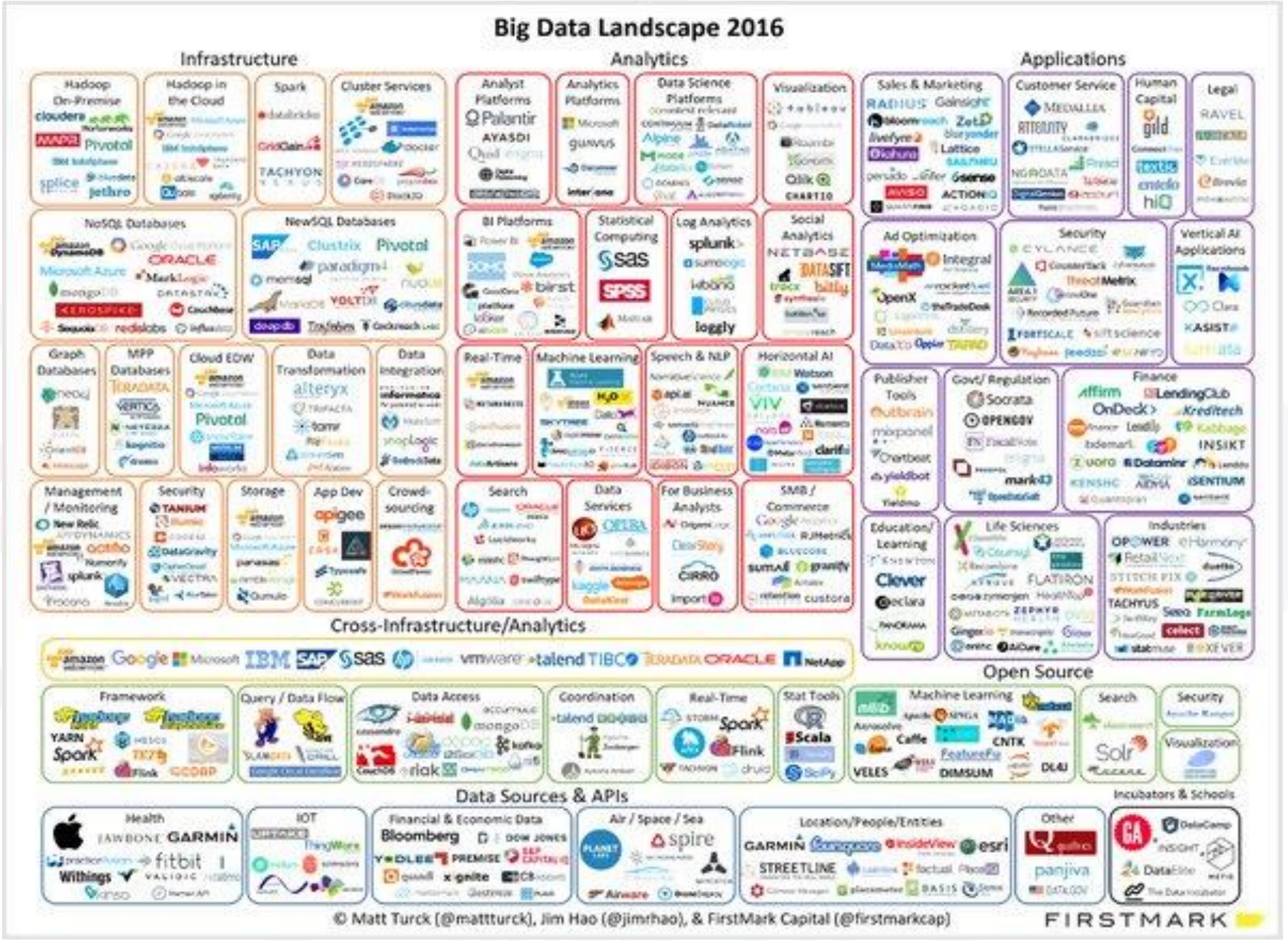
16.The White House has already invested more than $200 million in big data projects.

17.For a typical Fortune 1000 company, just a 10% increase in data accessibility will result in more than $65 million additional net income.

18.Retailers who leverage the full power of big data could increase their operating margins by as much as 60%.

19.73% of organizations have already invested or plan to invest in big data by 2016

20. **Interesting fact**: At the moment less than 0.5% of all data is ever analyzed and used, just imagine the potential here.



Big Decisions

Big Decisions Will Supplant Big Data

as the key initiative in business, government, and society

Big Individual Decisions

Big Company Decisions

Big Societal Decisions

Big Idea 2015, IBM Post

Big Individual *Decisions*

- More individual data than every before.
- Data about how we are driving through in-car sensors.
- Data about our health from smart watches and other wearable computers.
- Data about our cognitive abilities from daily tests on our smartphones.
- Data about our finances from multiple apps.
- Data about the traffic patterns on our commute.
- More of these data sources will add reasoning to help us make decisions and take action that improve our health, improve our safety, or improve our financial choices.
- We are even evolving the weather forecast to become more personalized to your life outdoors -- with forecasts for traveling, boating, surfing, running, meal planning, hair care, and 100s of other use cases.
- We have learned that turning data into decisions helps people make the most out of weather data

Big Company Decisions

- Companies have been investing heavily in Big Data.
- They are tracking every Stock Keeping Unit, every purchase, every ad impression, every referral, every hotel stay, product efficacy, and so on.
- Companies have been correlating the behavior of their customers with economic data, pricing data, weather data, news cycles, and every potential factor that might drive a customer decision.
- In a few leading cases, companies are getting more savvy about using the data to drive better and more precise decisions.
- We see better decisions to serve the right content to each customer from savvy recommendation engines of leading e-commerce and publishing companies.
- We see better decisions from retail ad campaigns, synced with sophisticated supply chain management, which vary by geography based on local weather and economic conditions.
- We see better investment and risk decisions in some technical trading models.
- Most companies are not yet using machine learning, not are they automated decision-making with better algorithms. Instead, they are feeding more data into the same human processes. The decisions are constrained by the human capacity to process the data, so only a small part of the Big Data is actually used.
- Recently, we should see more companies rethink their decision processes to make Bigger and Better Decisions using Big Data.

Big Societal *Decisions*

- Government officials and large non-profit leaders need to make decisions that affect society as a whole, in policy areas ranging from sustainable energy to education and healthcare access to deployment of national and international security efforts.
- To help make these decisions, governments have also long been the biggest collectors of data -- satellite observations of the Earth. geological surveys of the planer, economic and employment data, census data, transportation and traffic patterns, and millions of other topics.
- This data is compiled, released in summary reports (often weeks and months later), and used to help lawmakers and policymakers with Big Decisions.
- Recently, there have been efforts in the USA and other key governments to make the data accessible more rapidly and in more raw and granular fashion.
- This allows the data to be used by tax-paying citizens and businesses to improve their own decisions.
- It also helps create entire new business models, like our weather business, or Google and Microsoft's map business, or Bloomberg's economic indices.
- When we free Big Data, process it through faster and lower cost computers, and turn it into Big Decisions, we create an economic engine with more precision, more efficiency, and more time left for innovation.

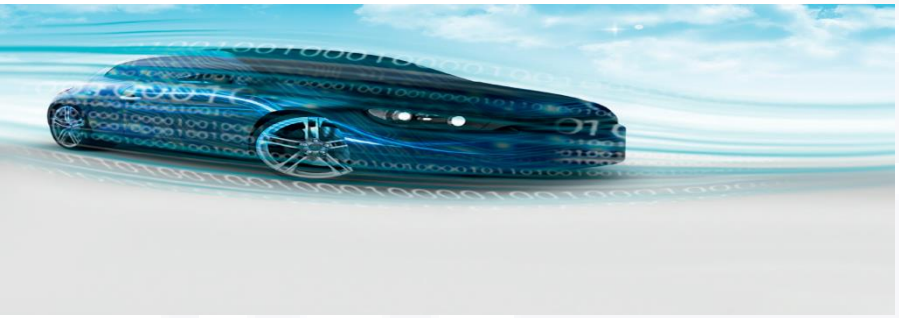
ECONOMIC INDICATORS - BLOOMBERG MARKET CONCEPTS

Discover the regiment upon which economic indicators are published and analyzed.

Identify how investors use economic indicators to gauge the health of the economy

Explain the qualities of good economic indicators.

Explore how economic indicators can be used to spot inflection points.



CONNECTED CAR: TRAFFIC AND DIAGNOSTICS
CONNECTING VEHICLES – CONNECTING CUSTOMERS
BIG DATA IN THE AUTOMOTIVE INDUSTRY.

Able to provide a wealth of information that would be invaluable to drivers, repair shops and automakers alike. To gain access to this data – and help the car talk – more and more vehicles are being fitted with sensors and connectivity solutions.

80 percent of all autos sold in 2016 will be connected.

That would equate to approximately 210 million talking cars cruising round our streets.

Compared to 45 million autos in 2011, that is a projected annual growth rate of over 36 percent.

Connected cars could provide a steady stream of data on vehicle movements, condition, wear and tear of parts, and ambient conditions.

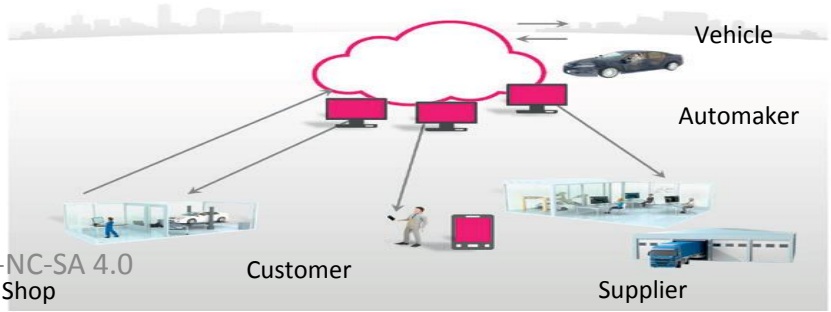
Extracting meaning from this mass of mixed data is no easy task. The challenge is transmitting the information, analyzing it and redistributing it to the relevant recipients – all at high speed

BIG DATA. BIG OPPORTUNITY

- To survive and thrive in today's ferocious markets, you need to have your finger on the pulse.
- Making the most of your mass data – to understand your customers, to identify market opportunities and anticipate relevant developments.
- To get to the top, and stay at the top, you must effectively manage the four key attributes of big data:
- ***Volume: the quantity of data to be captured continues to grow exponentially.***
- ***Velocity: bits and bytes have to be processed at high speed.***
- ***Variety: data comes in many formats, from diverse sources.***
- ***Value: data needs to be converted into meaningful insights.***
- To leverage these four Vs to your full advantage, you need a fresh approach: an end-to-end solution that aggregates, analyzes and visualizes mass data – quickly, simply and reliably

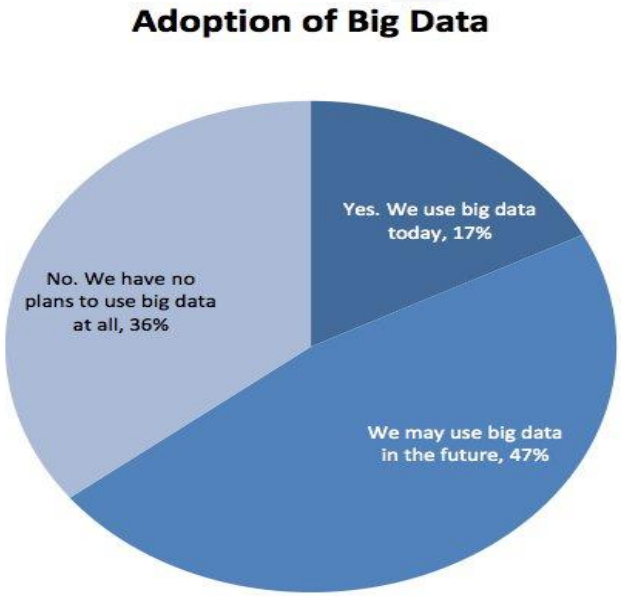
THE THREE KEY BENEFITS.
1. Improved car development
2. Enhanced planning of service and repair intervals due to early fault recognition
3. Increased customer satisfaction and retention

Real time Information Sharing

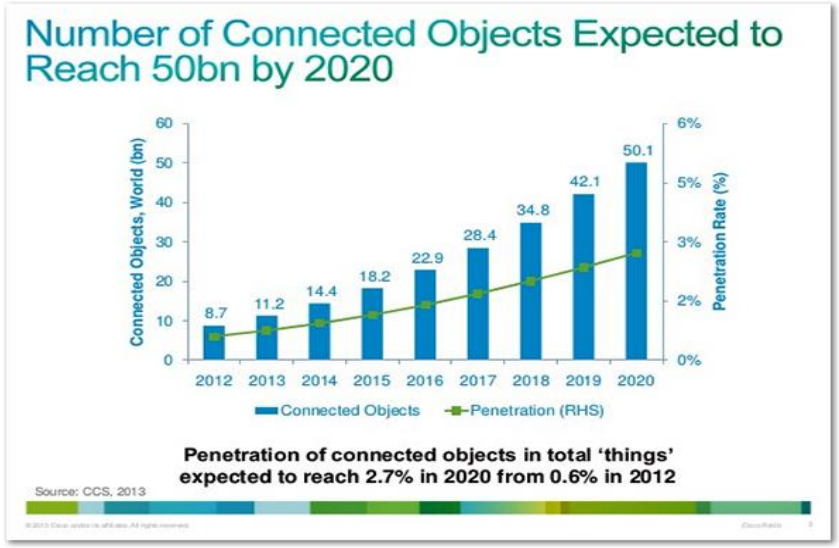


Connected devices

The global unmanned aerial vehicle (UAV) market is expected to show double-digit growth over the next five years, with government agencies and private sector companies in agriculture, energy, retail, utilities, mining, construction, real estate, news media, and other fields seeking ways to use drones in their operations and in their data collection and analytics.



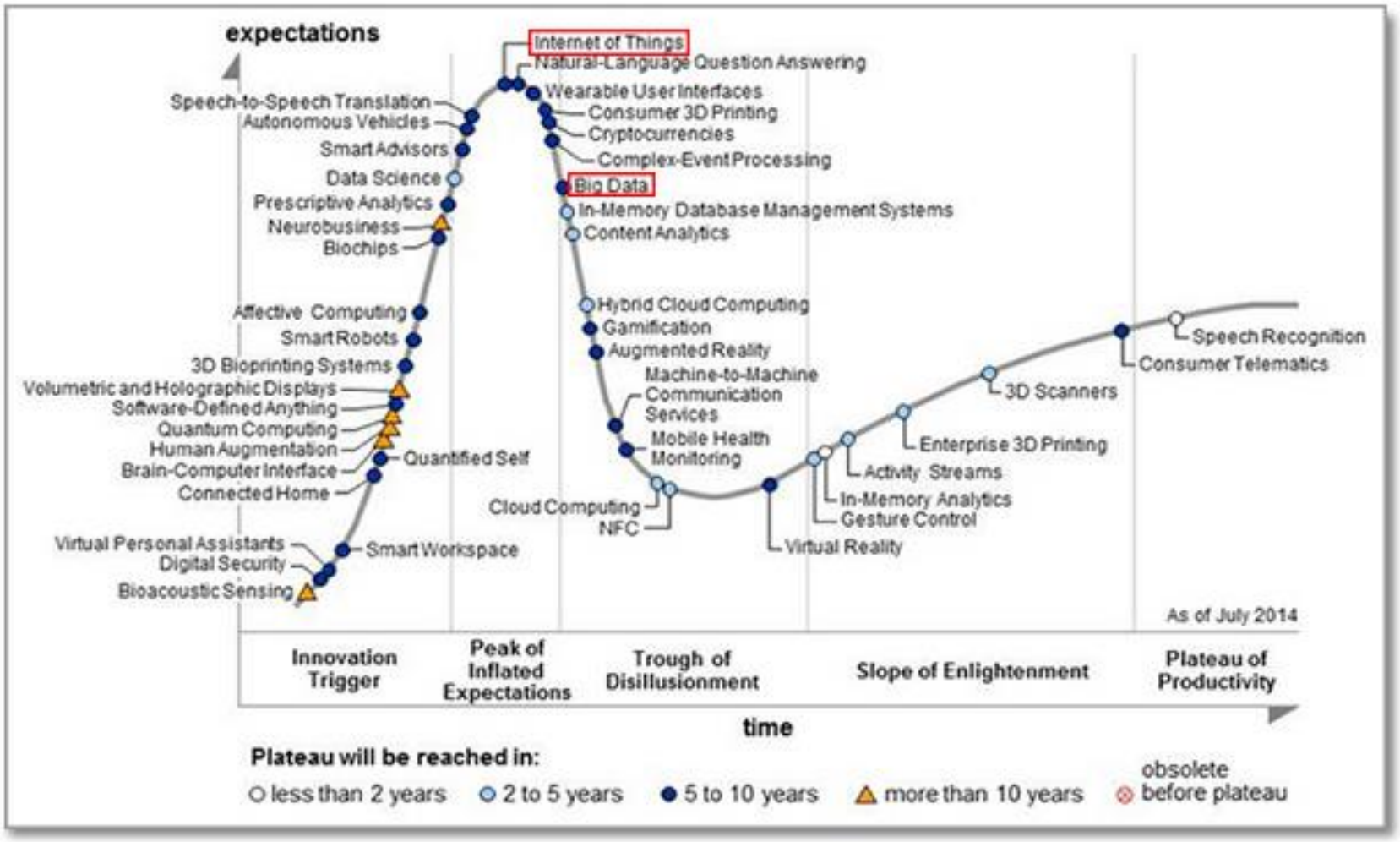
Drones collecting big data



Avoiding Big Data Headaches: Moving Hadoop to Production

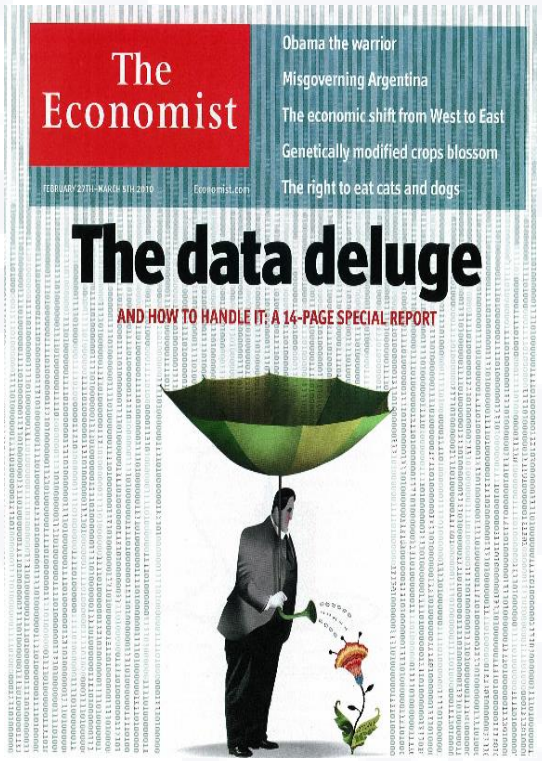
The internet of things and big data: Unlocking the power

The IoT will massively increase the amount of data available for analysis by all manner of organizations



Gartner's most recent Hype Cycle for Emerging Technologies:

What is big data?
Big Revolution?



5 Billions of mobile devices accessing to the broadband in 2010. 50 Billions are expected to 2020

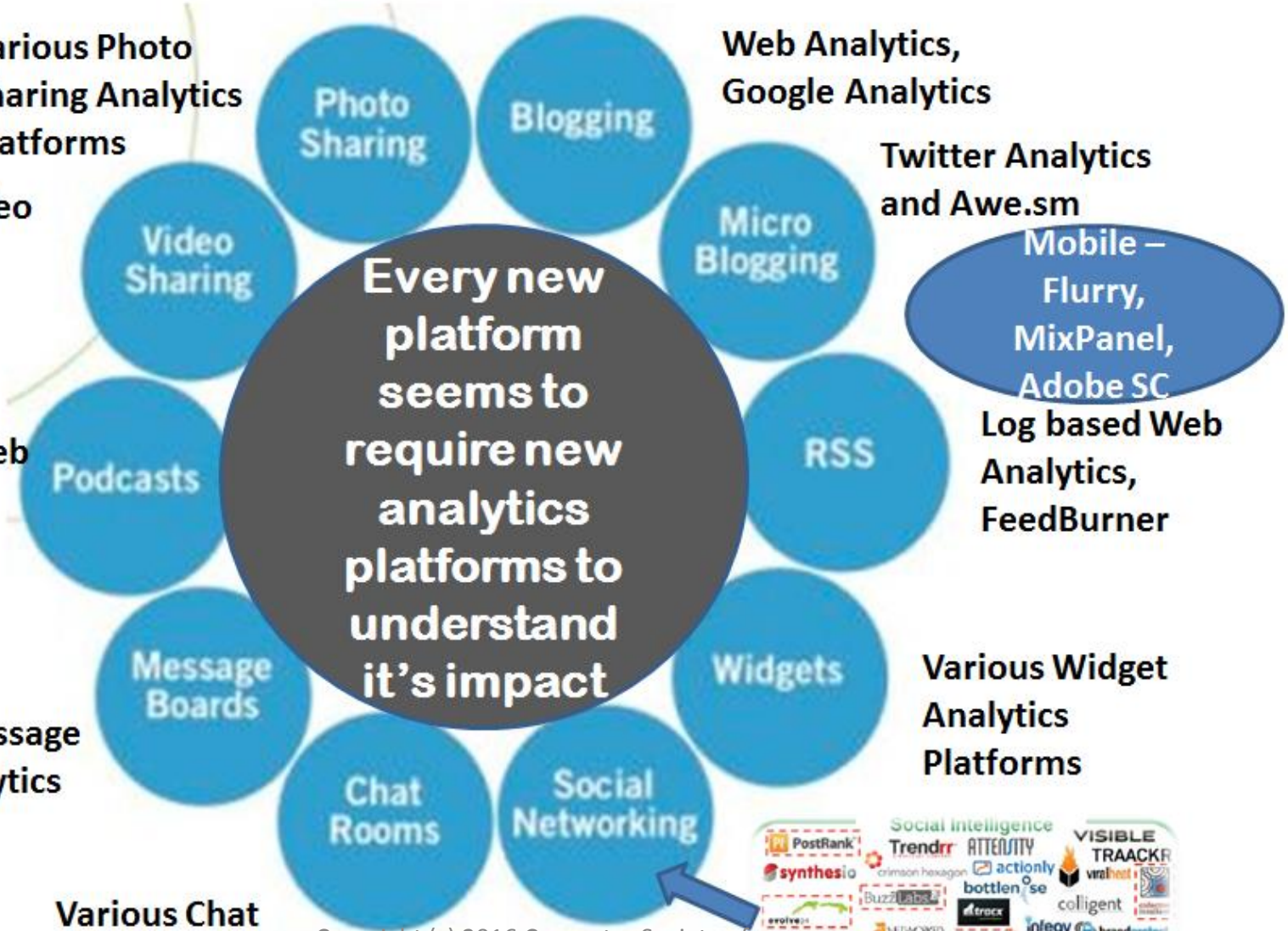
15 out of 17 sectors in US generate more digital info per company than US congress library

Smart meters and utilities measure installed in each US home produce more than 400 terabytes of new data every day

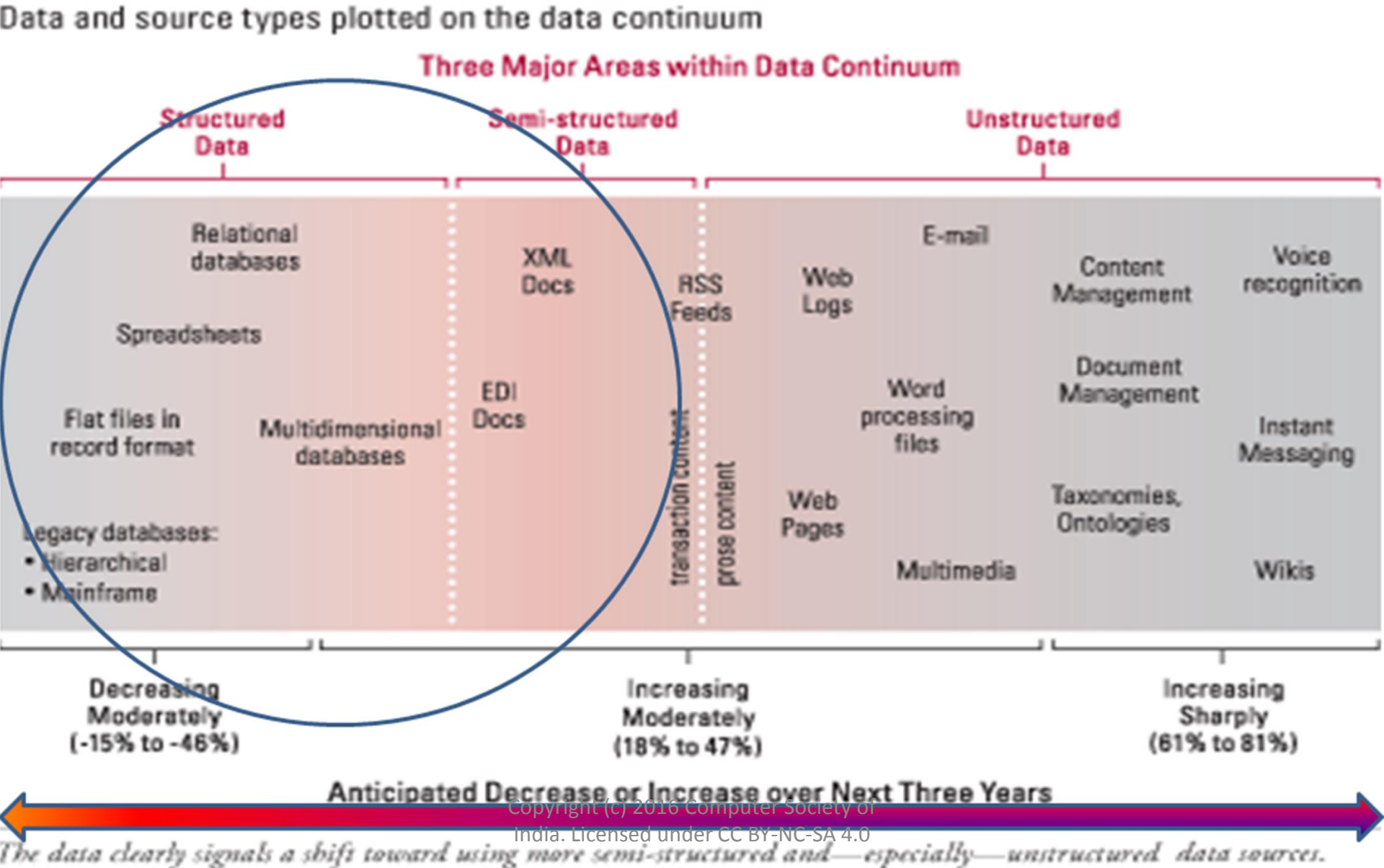
30 Billions of shared contents in Facebook every month. It is estimated to store at least 100 petabytes of pictures and videos alone.

Each second of high definition video generates more than 2000 times as many bytes as storing a page of text

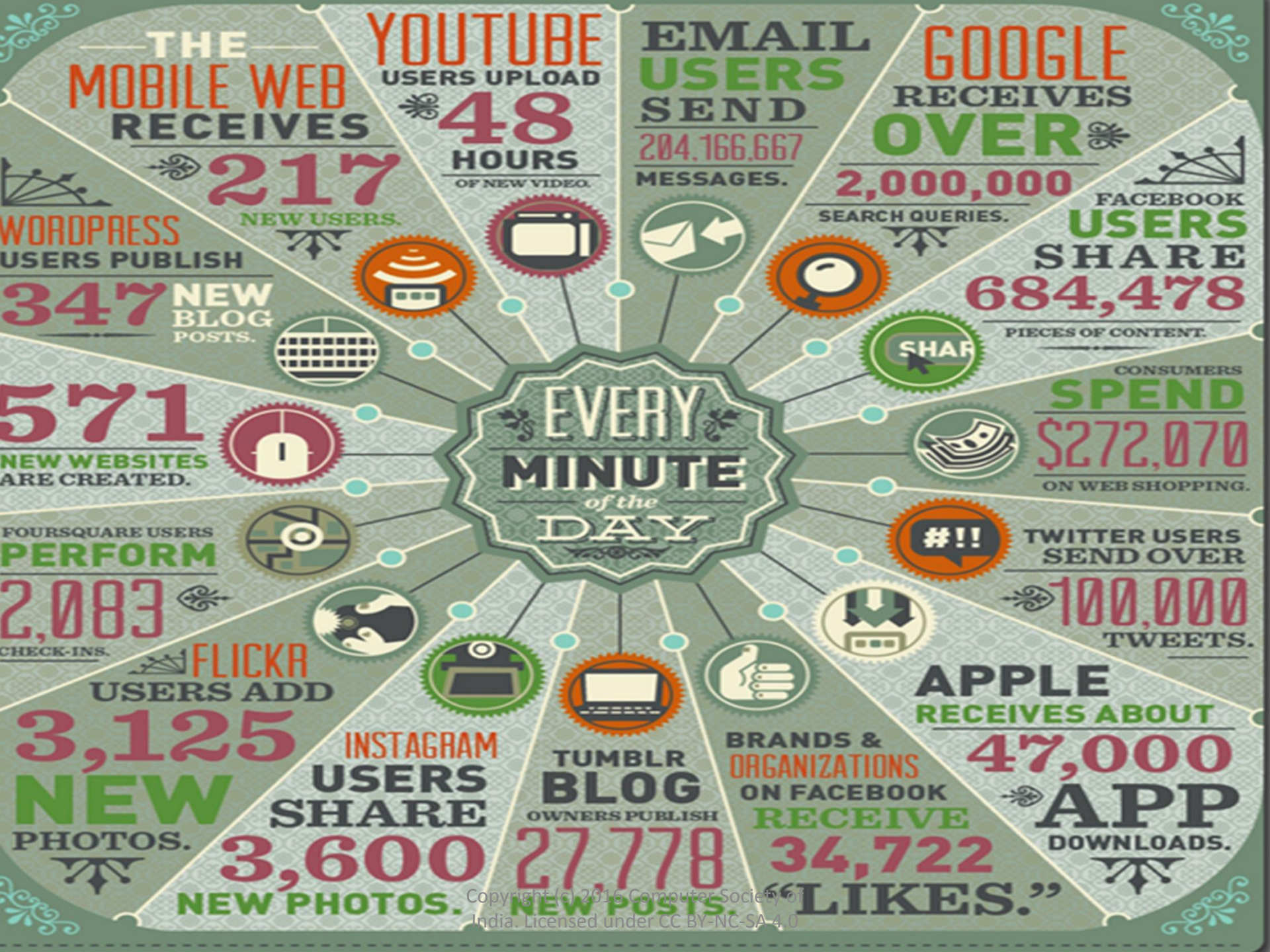
Current Situation – More types of data to mash up, so much of it *unstructured*



The Challenge: How much Structured Data can we capture from the Big Data *Continuum by using intelligent sensors?*



The data clearly signals a shift toward using more semi-structured and—especially—unstructured data sources.



Social Media Analytics

Key topics Include:

- **Analyzing social media marketing opportunities for the arts**
- **Exploring the behavioral and psychological factors that drive social media**
- **Understanding how to form and optimize social networks for the arts**
- **Designing effective communication strategies of social networks**
- **Implementing measurement strategies to track and measure the ROI of social media**
- **Discovering new social media technologies and learning how and when to employ them**

8 Big Trends in Big Data Analytics

1.Big data analytics in the cloud

2. Hadoop: The new enterprise data operating system

3. Big data lakes

4. More predictive analytics

5. SQL on Hadoop: Faster, better

6. More, better NoSQL

7. Deep learning

8. In-memory analytics

Cost Effective Over Time Hadoop Vs. Data Warehouse

Big Data – What Does It Really Cost?

Use cases

- **Example 1: Build a Data Warehouse**
- Objective: Build an enterprise data warehouse for a large financial institution
- Data volume: 500 TB
- Business requirements:
  - Large number of data sources, users, complex queries, analyses and analytic applications
  - Data integration and integrity
  - Reusability and agility to accommodate rapidly changing business requirements and long data life

- **Example 2: Build a Data Refinery**
- Objective: Refine the sensor output of large industrial diesel engines
- Data volume: 500 TB
- Business requirements:
  - Rapid, intensive processing of a small number of closely-related data sets
  - Analysis reads the entire dataset
  - Life of the raw data is relatively short
  - Small group of experts collaborate on analysis

Cost Comparison – A 5-Year Summary

Example 1: Data Warehouse

Cost	DataWarehouse Platform	Hadoop
System Cost Initial acquisition	$44.6 $10.8	$1.4 $0.2
Upgrades	$16.4	$0.3
Maintenance/support	$15.9	$0.2
Power/space/cooling	%1.5	$0.6
Administration	$7.7	$8.5
Application Development	$16.5	$36.0
ETL	$18.4	---
Complex queries	$88.7	$475.0
Analysis	$88.7	$219.0
Total cost	$265.0 million	$740.0 million

Example 2: Data Refinery

Cost	DataWarehouse Platform	Hadoop
System Cost Initial acquisition	$22.7 $5.5	$1.4 $0.2
Upgrades	$8.4	$0.3
Maintenance/support	$8.2	$0.2
Power/space/cooling	%0.6	$0.7
Administration	$0.8	$0.8
Application Development	$6.6	$7.2
ETL		
Complex queries		
Analysis		
Total cost	$30.0 million	$9.3 million

Winner

- **Data Warehouse**
- The data warehouse platform ($265 million) is far more cost-effective than a Hadoop solution ($740 million).
- Choosing the data warehouse platform in this case lowers the overall cost by a factor of 2.8.
- Further analysis shows that you will get essentially the same result for a data warehouse ranging in size from 50 TB to 2 PB.
- The development of complex queries and analytics are the dominant cost factors in the example.
- Of the $44 million estimated for EDW system cost, $10.8 million is the initial acquisition cost – about 4% of the TCOD.
- While it is common to focus on the first major outlay in the project—i.e., the acquisition of a platform—the total cost of the project is far more important, and other factors greatly outweigh all the system costs combined.

- **Hadoop**

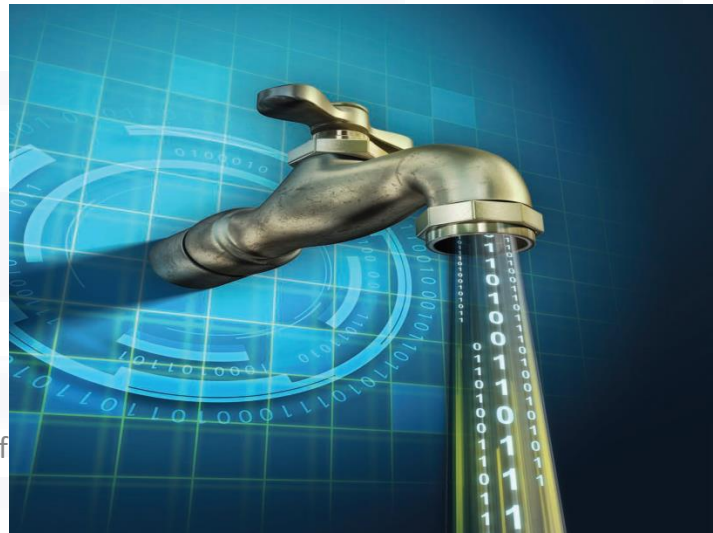
- Hadoop ($9.5 million) is a far more cost-effective solution than a data warehouse appliance ($30 million).
- The system cost for the data warehouse appliance is the dominant factor in this case.
- The system cost and its breakdown in the table above, where just $5.5 million of the $22.7 million system cost for the data warehouse appliance is incurred in the first year.

Present Industry Trends

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Big Data – A Future

- Challenge- The next frontier for innovation, competition, and productivity
- Evolution - large pools of data that can be captured, communicated, aggregated, stored, and analyzed—is now part of every sector and function of the global economy
- Deluge - a growing torrent



Big Data Every Where!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network and Media
  - APP






6 C's in Industry Scenario

- **Connection** (sensor and networks)
- **Cloud** (computing and data on demand)
- **Cyber** (model & memory)
- **Content/context** (meaning and correlation)
- **Community** (sharing & collaboration)
- **Customization** (personalization and value).

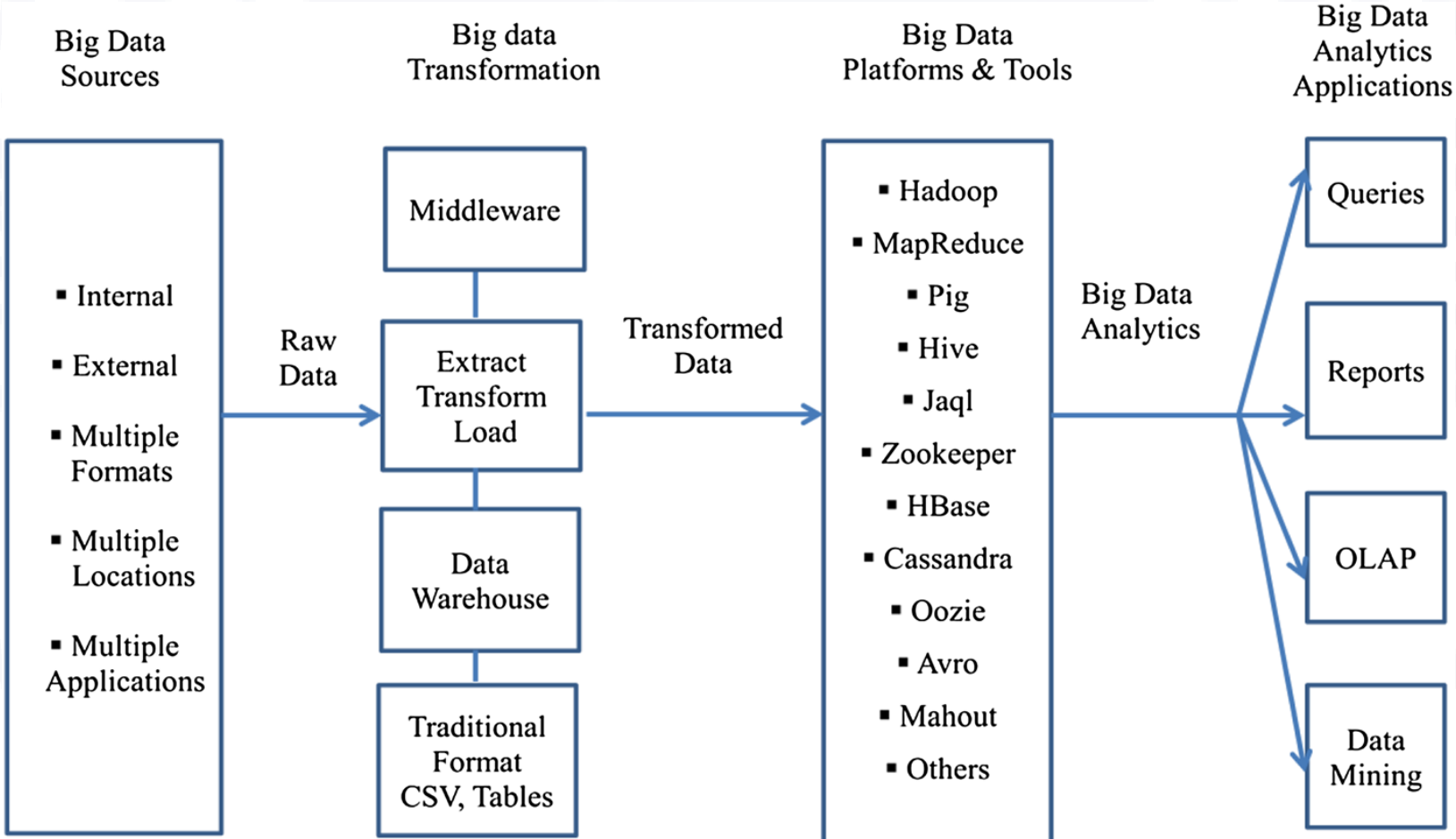
Great Values of BIG DATA

- Creating transparency
- Enabling experimentation to discover needs, expose variability, and improve performance
- Segmenting populations to customize actions
- Replacing/supporting human decision making with automated algorithms
- Innovating new business models, products, and services
- Industry Demonstration - increase productivity, quality and flexibility

BIG DATA Analytics

- ***Big data analytics*** is the process of examining ***big data*** to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions
- If big data is a haystack, analytics is how you find the needle.
- If it's a huge wave, analytics is a surfboard.
- If it's noise, analytics lets you hear the signal

Conceptual Architecture of Big Data Analytics



Smart Mega Cities – Requires full set of Solutions

Street Light Management

Public Safety

Smart travel

Management Control

Education

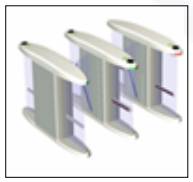
Fleet Management

VoD

Video Conferencing

HealthCare

KIOSK



Access Control

Parking Control

CCTV Monitoring

Waste Management

Facilities Control

Power Control

Light Control

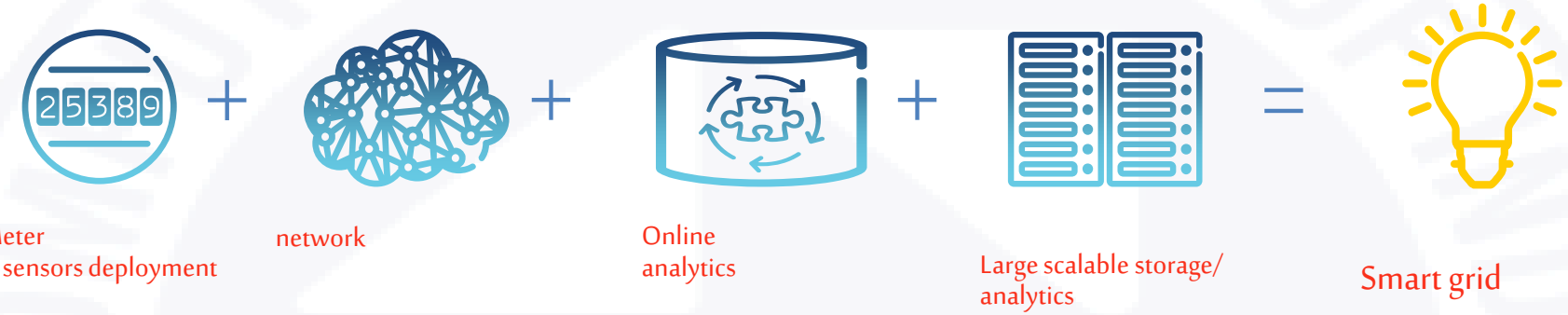
Big Data Sources and Types

- Every day 2.5 Petabytes of data are generated from new and traditional sources including climate sensors, social media sites, digital pictures and videos, purchase transaction records, cellphone GPS signals, and more.
- Big data is the combination of any type of data – structured and unstructured- such as text, sensor data, audio, video, clickstreams, log files and more

How much data?

- Google processes 40 PB a day
- Wayback Machine has 5 PB + 150 TB/month
- Facebook has 5.5 PB of user data + 35 TB/day
- eBay has 9.5 PB of user data + 90 TB/day
- CERN's Large Hydron Collider (LHC) generates 15 PB a year

Big data analysis and utilities



Enterprise analytics

- Consumer Profitability Management
- Financial and Operational Performance Management
- Market evolution

Grid operation analytics

- Asset management (Early fault detection)
- Manage & Predict Usage. Automatic grid accommodation
- Crisis analytics (Outage management)

Consumer analytics

- Behavioral analytics
- Dynamic pricing models
- Building energy management
- Social Media integration
- Energy Theft

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Big Data Components and Sources

- 4 V's

1. Velocity : Speed of data in and out
2. Volume : Increasing amount of data
3. Variety : Range of data types and sources
4. Veracity:

- Structured and Unstructured Data

1. RDBMS –based
2. Real-time data feeds and Social Media sources

Usage : Science and Research, Government and Corporations

Big Data Security and Privacy

- Secure data, collect it and then aggregate to evaluate
- ❖ Collection - Obtain visibility of data
- ❖ Integration – Understand the context
- ❖ Analytics – Utilize the intelligence

- Can the company trust its sources of Big Data?
- What information is the company collecting without exposing the enterprise to legal and regulatory battles?
- How will the company protect its sources, processes and decisions from theft and corruption?
- What policies are in place to ensure that employees keep stakeholder information confidential during and after employment?
- What actions are company taking that creates trends that can be exploited by its rivals?

Study by ISACA

7 Big Data Techniques – Business Value

1. Association rule learning
2. Classification tree analysis
3. Genetic algorithms
4. Machine learning
5. Regression analysis
6. Sentiment analysis
7. Social network analysis

Data Mining and Tools

- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Collaborative Filter [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

- automatically assign documents to categories
- categorize organisms into groupings
- develop profiles of students who take online courses

Decision Trees

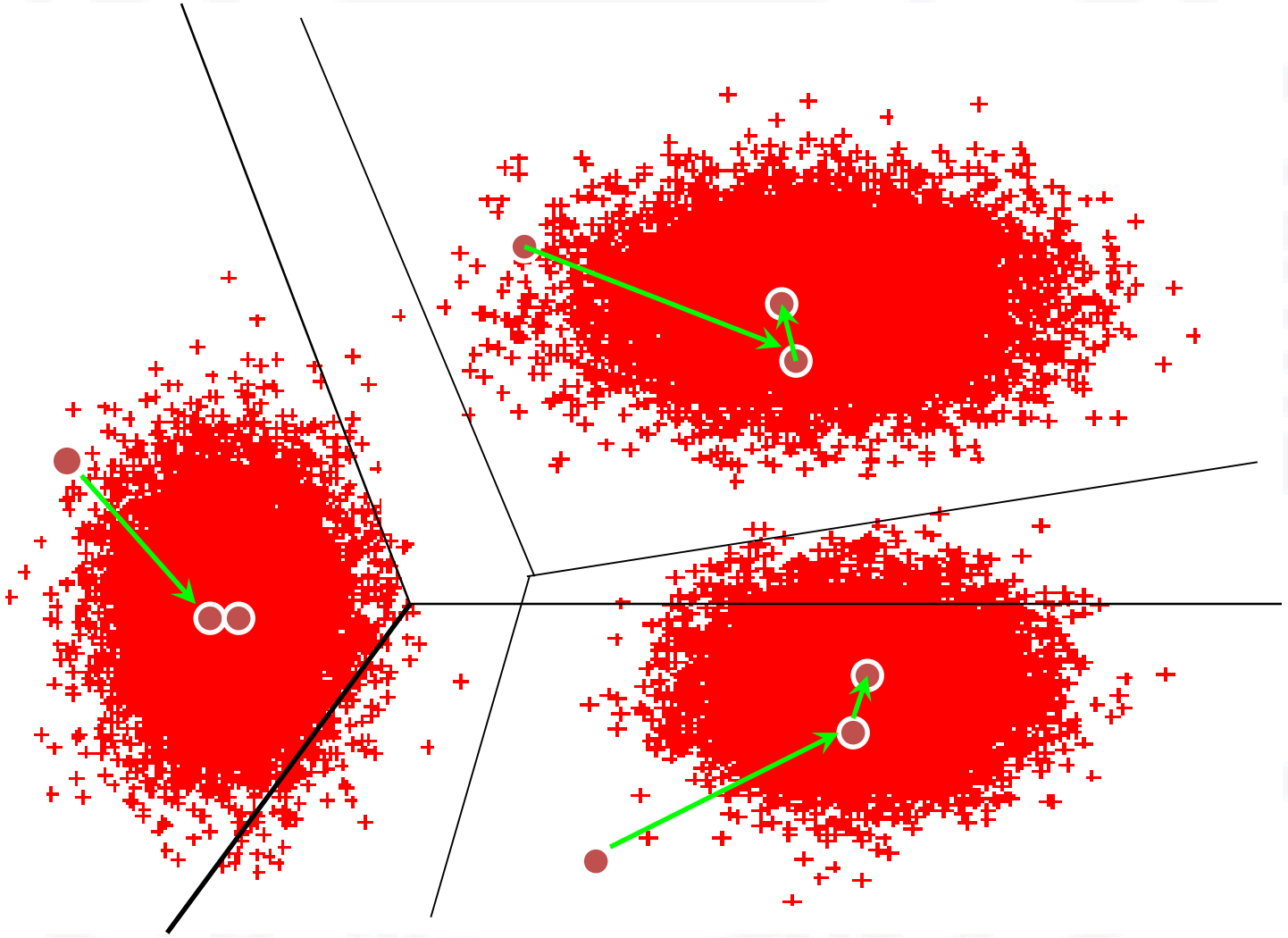
Example:

- Conducted survey to see what customers were interested in new model car
- Want to select customers for advertising campaign

sale	custId	car	age	city	newCar
	c1	taurus	27	sf	yes
	c2	van	35	la	yes
	c3	van	40	sf	yes
	c4	taurus	22	sf	yes
	c5	merc	50	la	no
	c6	taurus	25	la	no

training set

K-Means Clustering



Association Rule Mining

sales records:

transaction id	customer id	products bought
tran1	cust33	p2, p5, p8
tran2	cust45	p5, p8, p11
tran3	cust12	p1, p9
tran4	cust40	p5, p8, p11
tran5	cust12	p2, p9
tran6	cust12	p9

market-basket data

- Trend: Products p5, p8 often bough together
- Trend: Customer 12 likes product p9

- place products in better proximity to each other in order to increase sales
- extract information about visitors to websites from web server logs
- analyze biological data to uncover new relationships
- monitor system logs to detect intruders and malicious activity
- identify if people who buy milk and butter are more likely to buy breads

Collaborative Filtering

- Goal: predict what movies/books/… a person may be interested in, on the basis of
  - Past preferences of the person
  - Other people with similar past preferences
  - The preferences of such people for a new movie/book/…
- One approach based on repeated clustering
  - Cluster people on the basis of preferences for movies
  - Then cluster movies on the basis of being liked by the same clusters of people
  - Again cluster people based on their preferences for (the newly created clusters of) movies
  - Repeat above till equilibrium
- Above problem is an instance of **collaborative filtering**, where users collaborate in the task of filtering information to find information of interest

Other Types of Mining

- **Text mining**: application of data mining to textual documents
  - cluster Web pages to find related pages
  - cluster pages a user has visited to organize their visit history
  - classify Web pages automatically into a Web directory
- **Graph Mining**:
  - Deal with graph data

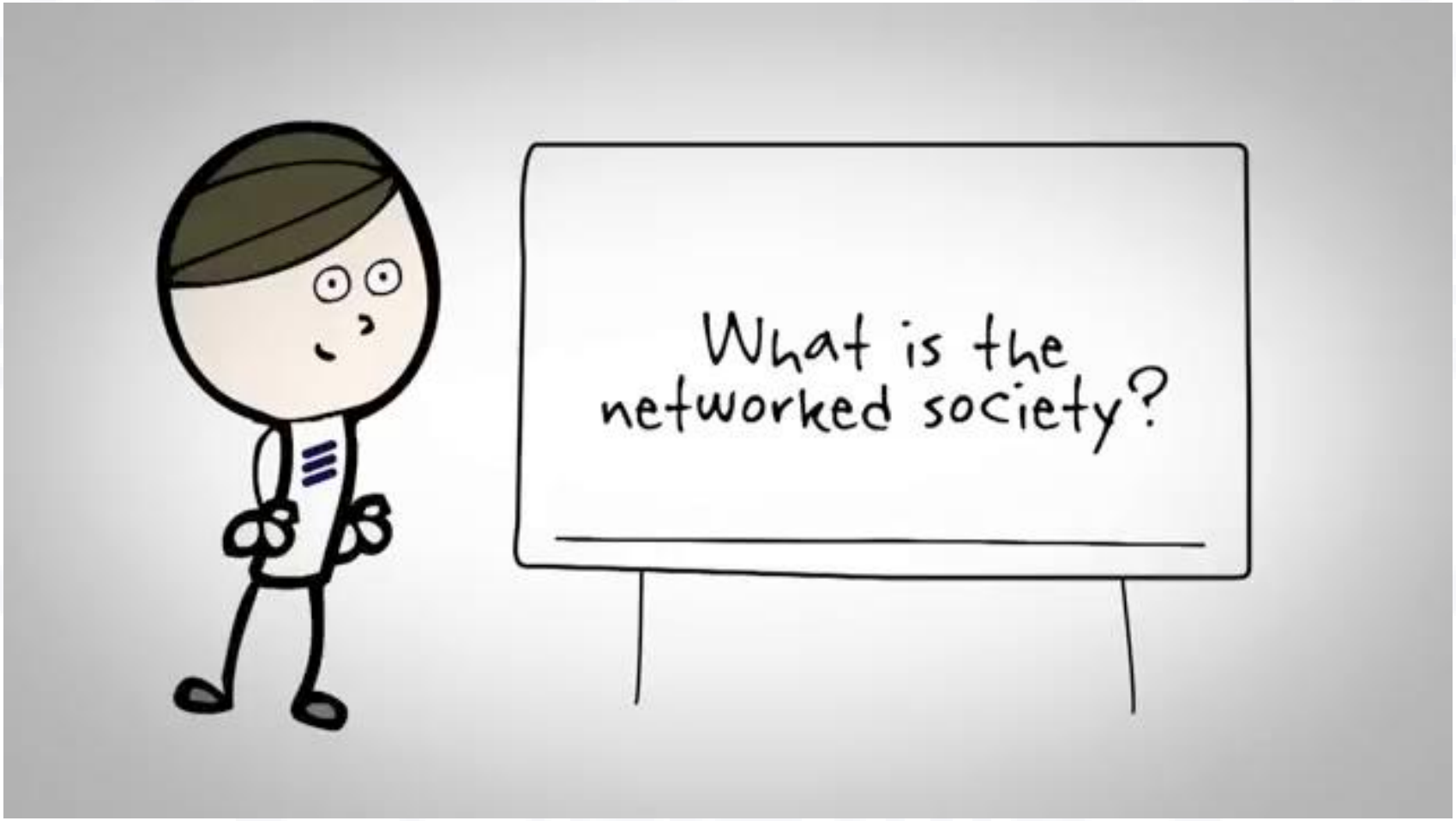
Data Streams

- What are Data Streams?
  - Continuous streams
  - Huge, Fast, and Changing
- Why Data Streams?
  - The arriving speed of streams and the huge amount of data are beyond our capability to store them.
  - “Real-time” processing
- Window Models
  - Landscape window (Entire Data Stream)
  - Sliding Window
  - Damped Window
- Mining Data Stream

Twitter can be the new crime buster

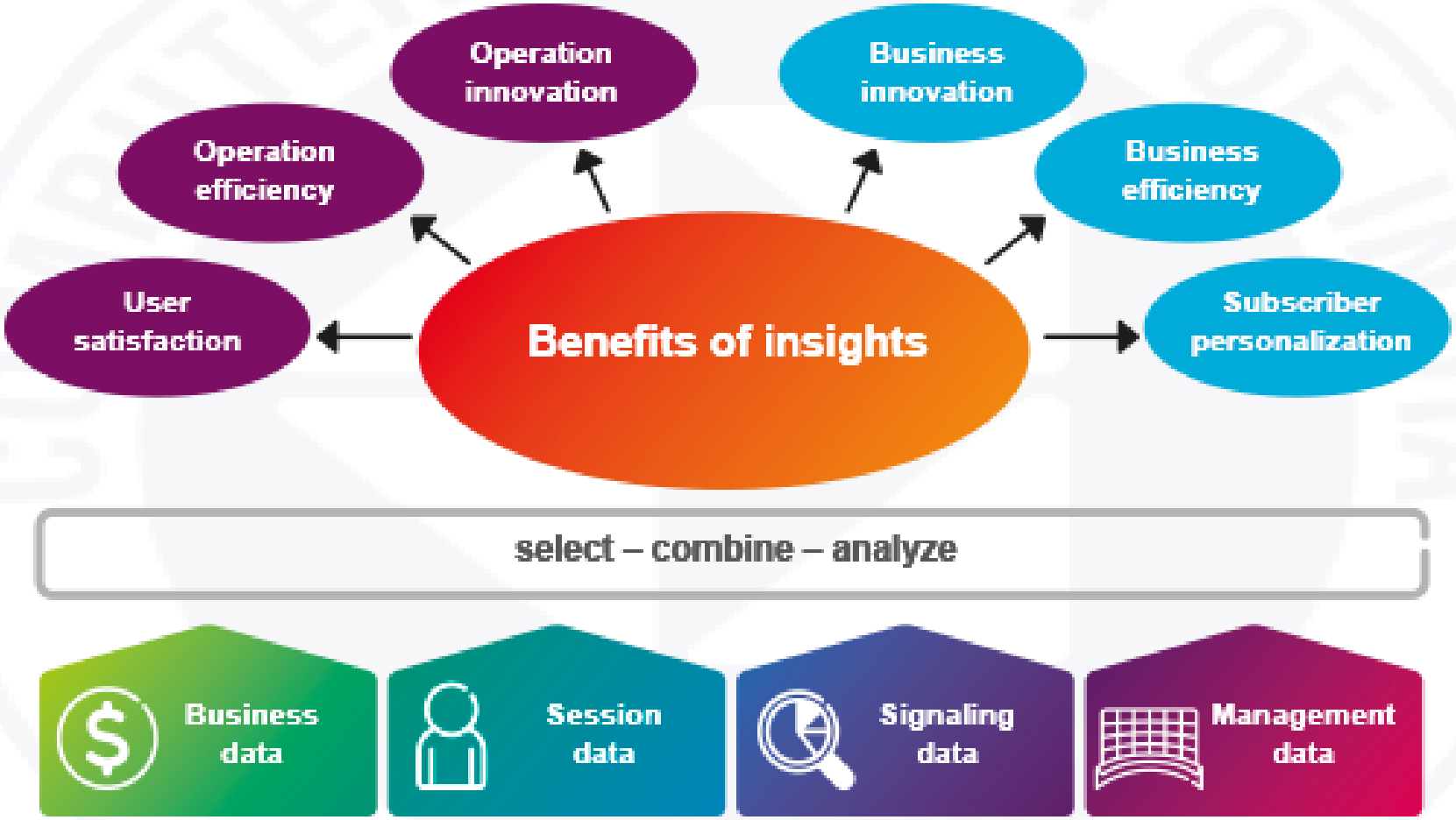
- Hidden in the Twittersphere are nuggets of information that could prove useful to crime fighters — even before a crime has been committed.
- A research paper published in the scientific journal Decision Support Systems few months before said the analysis of geo-tagged tweets can be useful in predicting 19 to 25 kinds of crimes, especially for offences such as stalking, thefts and certain kinds of assault.
- In this study, they analyzed tweets from the city of Chicago tagged to certain neighborhood's — measured by individual square kilometers — and the city's crime database . They were then able to make useful predictions about areas where certain crimes were likely to occur — something which could be helpful in deployment of police resources.
- "This approach allows the analyst to identify areas with historically high crime concentrations ," said the study. "Future crimes often occur in the vicinity of past crimes, making hot-spot maps a valuable crime prediction tool.

NETWORKED SOCIETY



BIG DATA IMPACT ON TELCO

BIG DATA ANALYTICS – TELECOMMUNICATIONS VIEW



Smarter networks, improved user experience, data monetization and churn prevention.