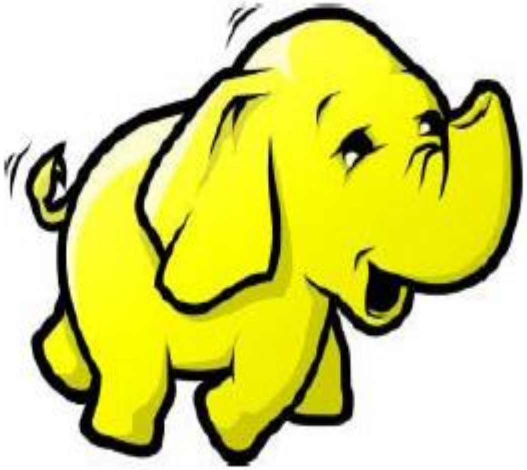
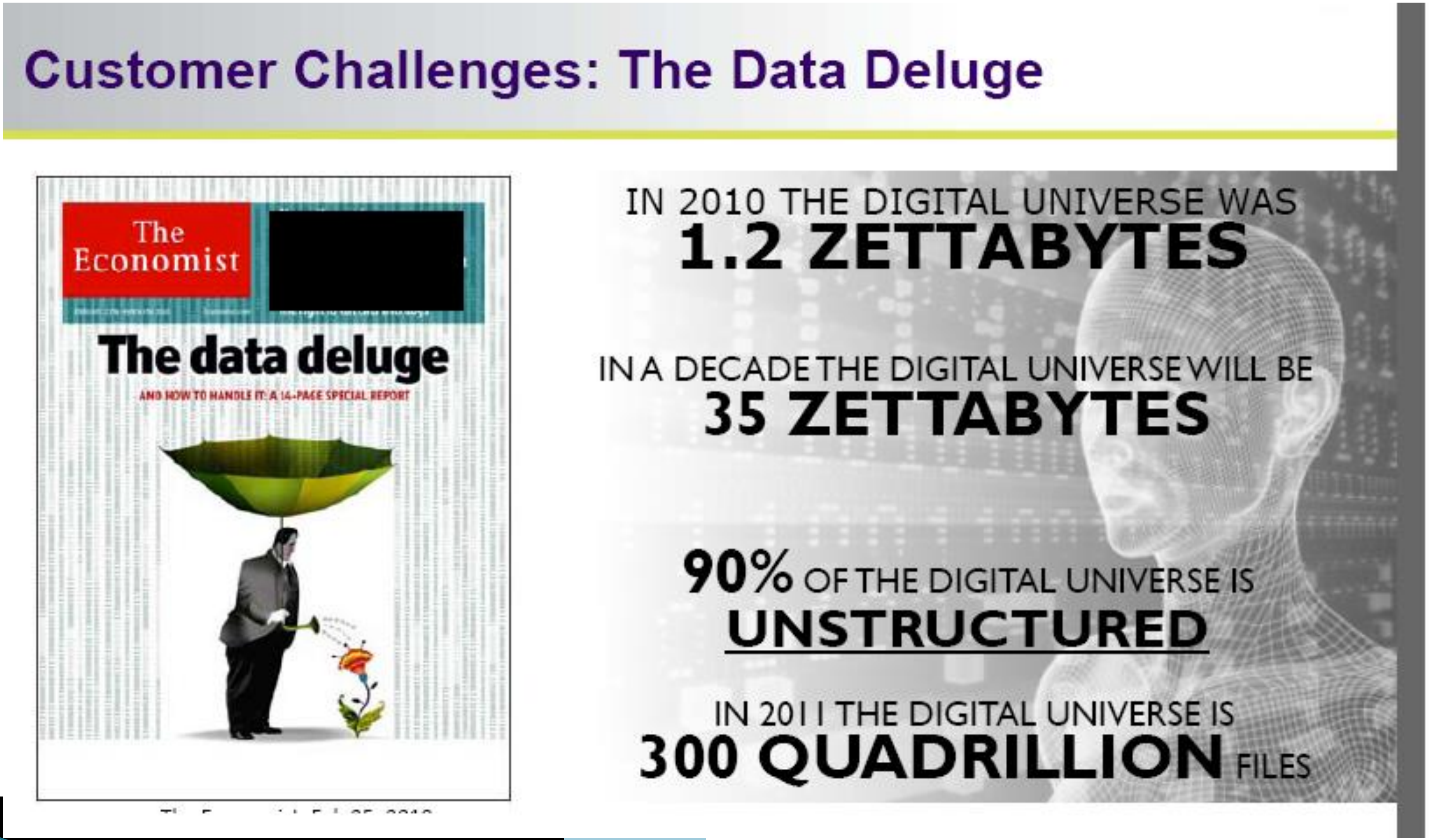
Big Data

BIG DATA

Data Challenges



Big Data



Customer Challenges: The Data Deluge

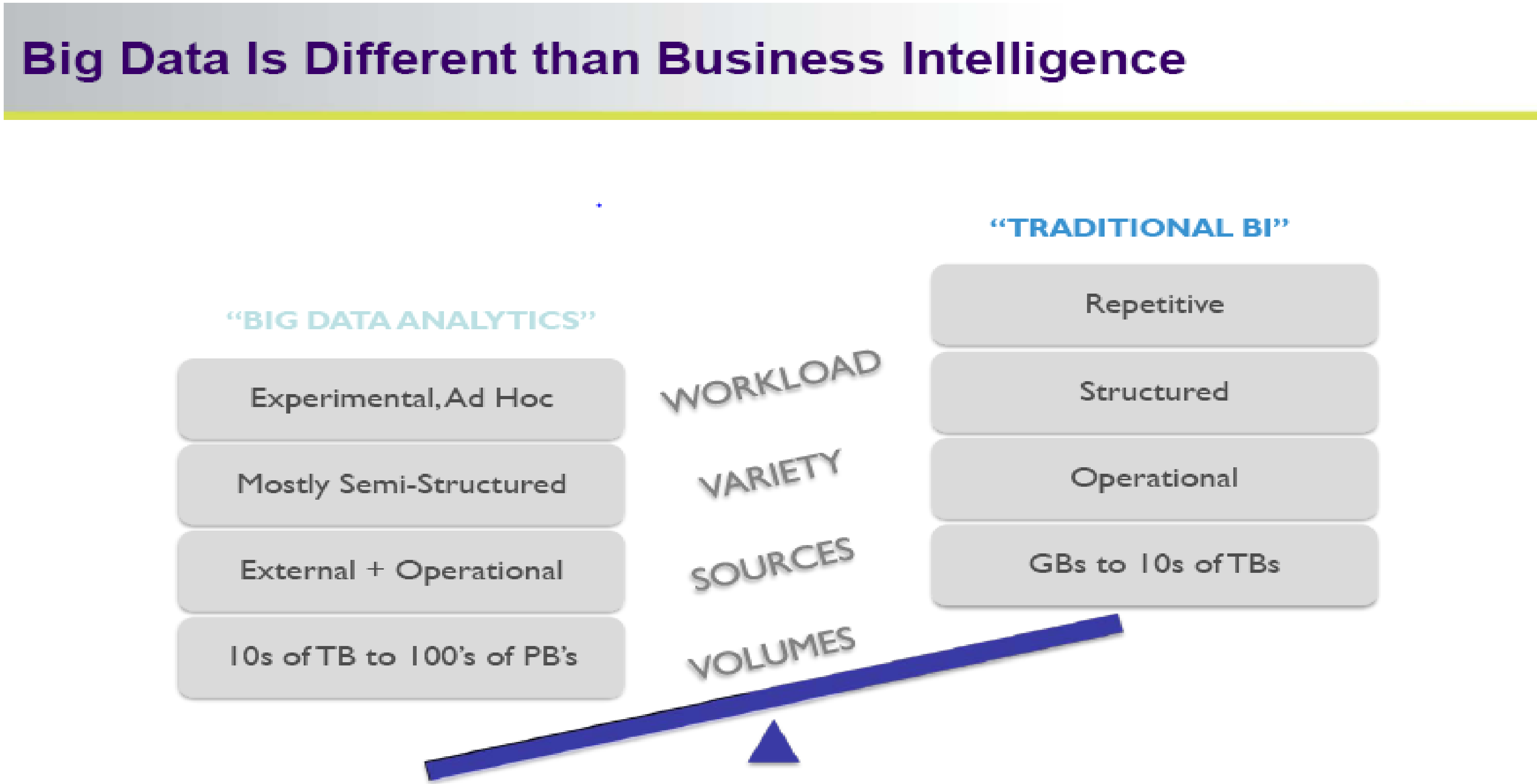
IN 2010 THE DIGITAL UNIVERSE WAS
1.2 ZETTABYTES

IN A DECADE THE DIGITAL UNIVERSE WILL BE
35 ZETTABYTES

90% OF THE DIGITAL UNIVERSE IS
UNSTRUCTURED

IN 2011 THE DIGITAL UNIVERSE IS
300 QUADRILLION FILES

Big Data



Big Data Is Different than Business Intelligence

"BIG DATA ANALYTICS"

Experimental, Ad Hoc

Mostly Semi-Structured

External + Operational

10s of TB to 100's of PB's

WORKLOAD

VARIETY

SOURCES

VOLUMES

"TRADITIONAL BI"

Repetitive

Structured

Operational

GBs to 10s of TBs

Big Data

WHY??

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

Big Data

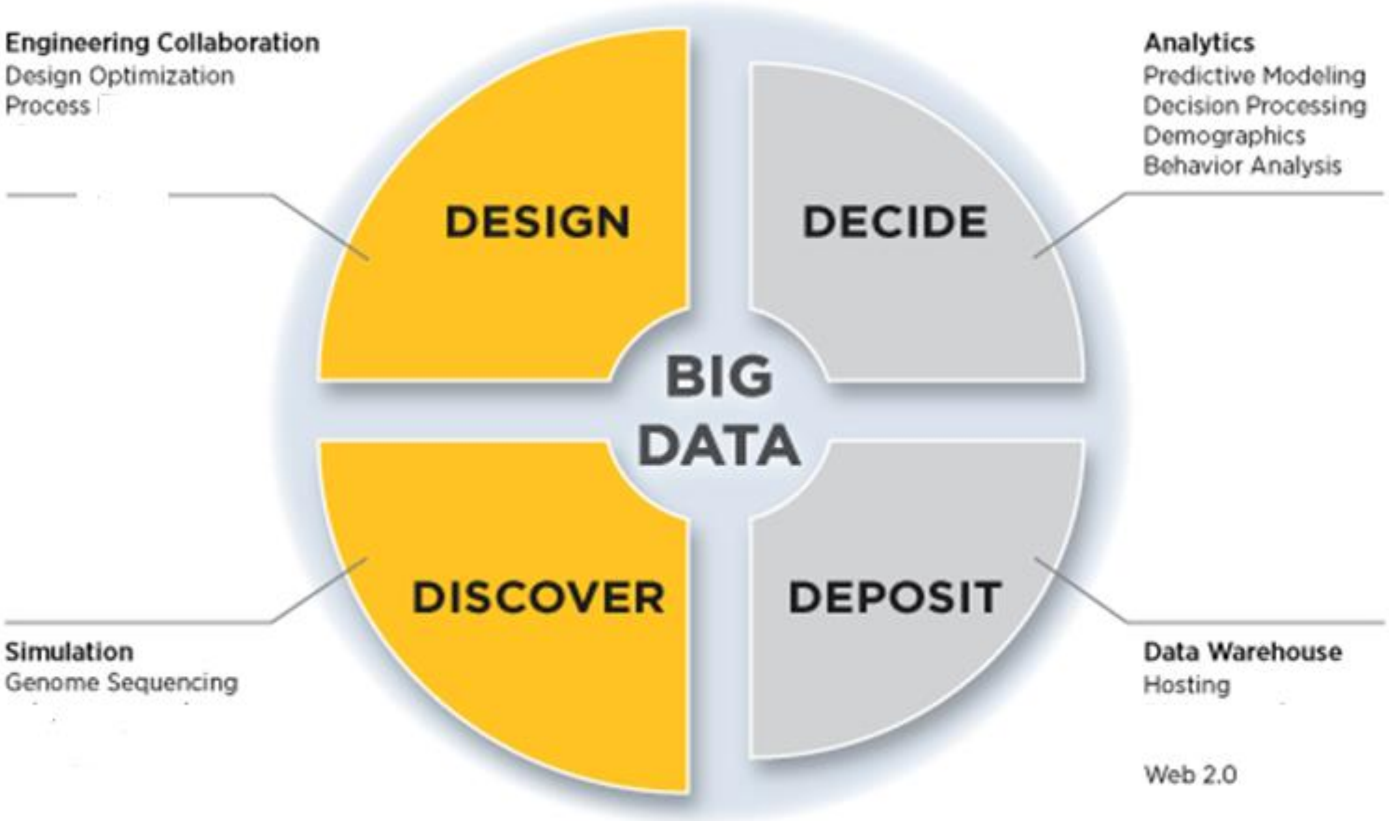
- *Big data is a term applied to a* ***new generation of software, applications, and system and storage architecture.***
- *It designed to provide* ***business value from unstructured data****.*
- *Big data sets* ***require advanced tools****, software, and systems to capture, store, manage, and analyze 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

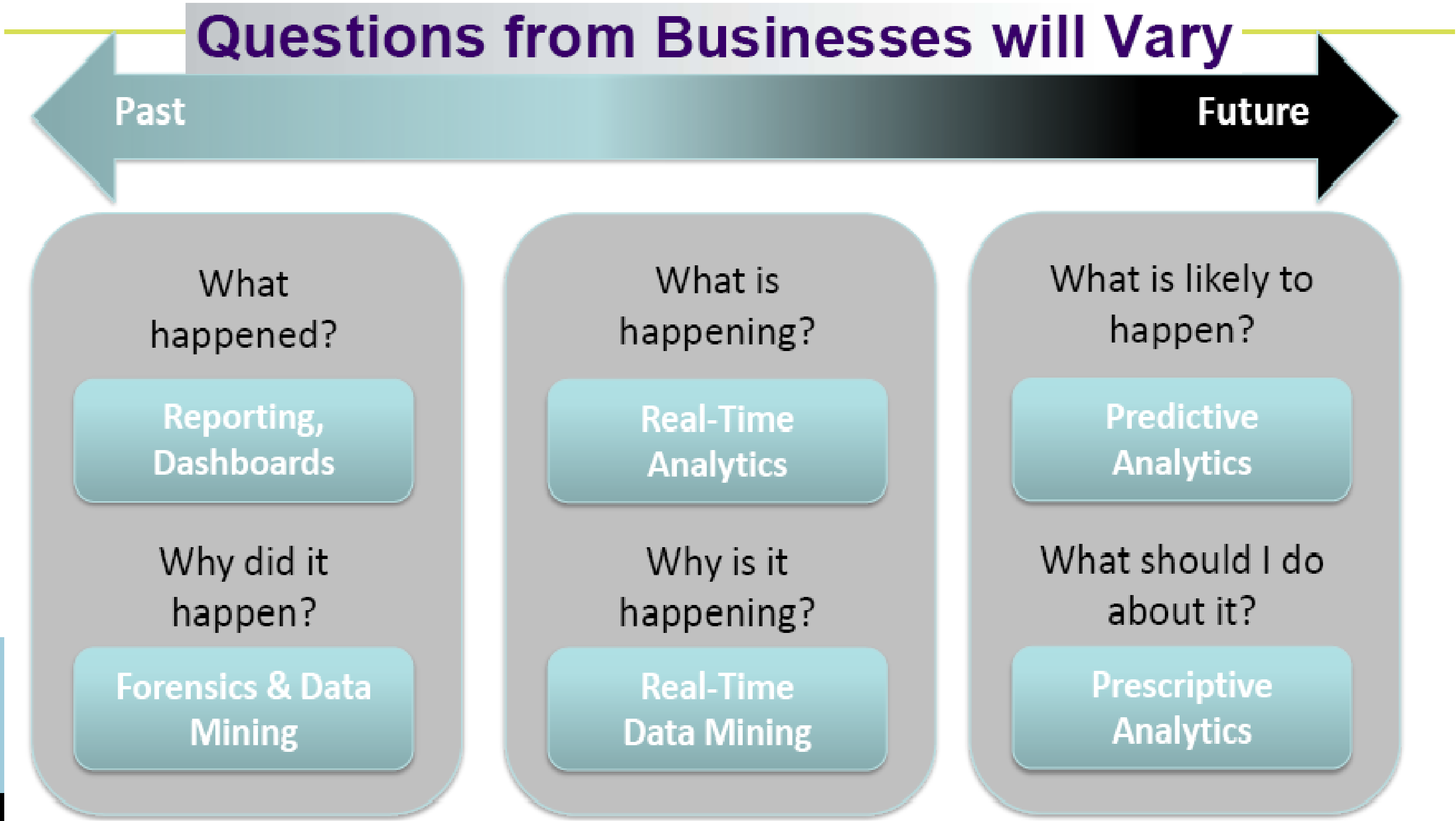
Analytics
Predictive Modeling
Decision Processing
Demographics
Behavior Analysis

Simulation
Genome Sequencing

Data Warehouse
Hosting

Web 2.0

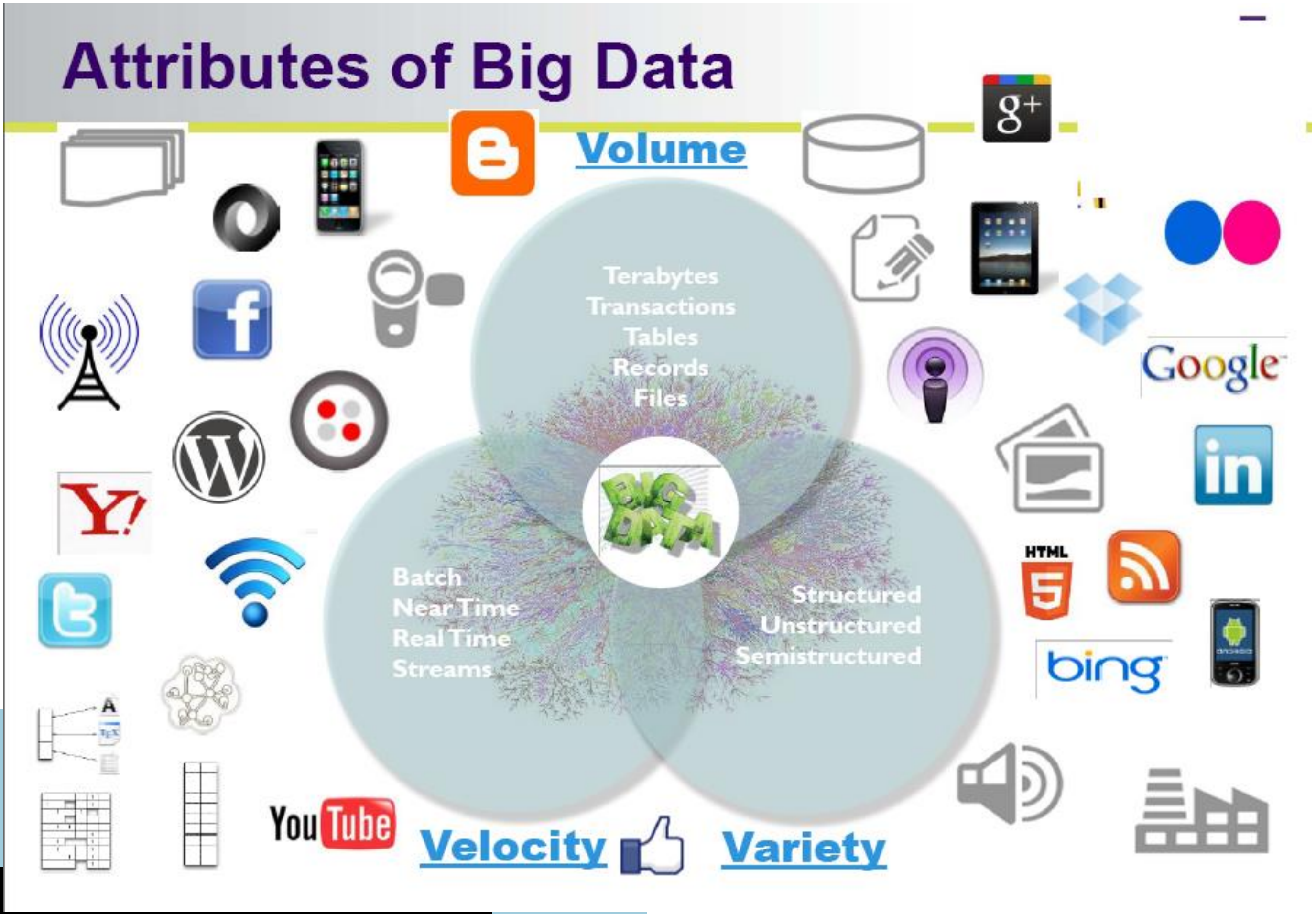
Big Data


Questions from Businesses will Vary
Past
Future
What happened?
Reporting, Dashboards
Why did it happen?
Forensics & Data Mining
What is happening?
Real-Time Analytics
Why is it happening?
Real-Time Data Mining
What is likely to happen?
Predictive Analytics
What should I do about it?
Prescriptive Analytics

Big Data
“Data Driven” Web 2.0 onwards.

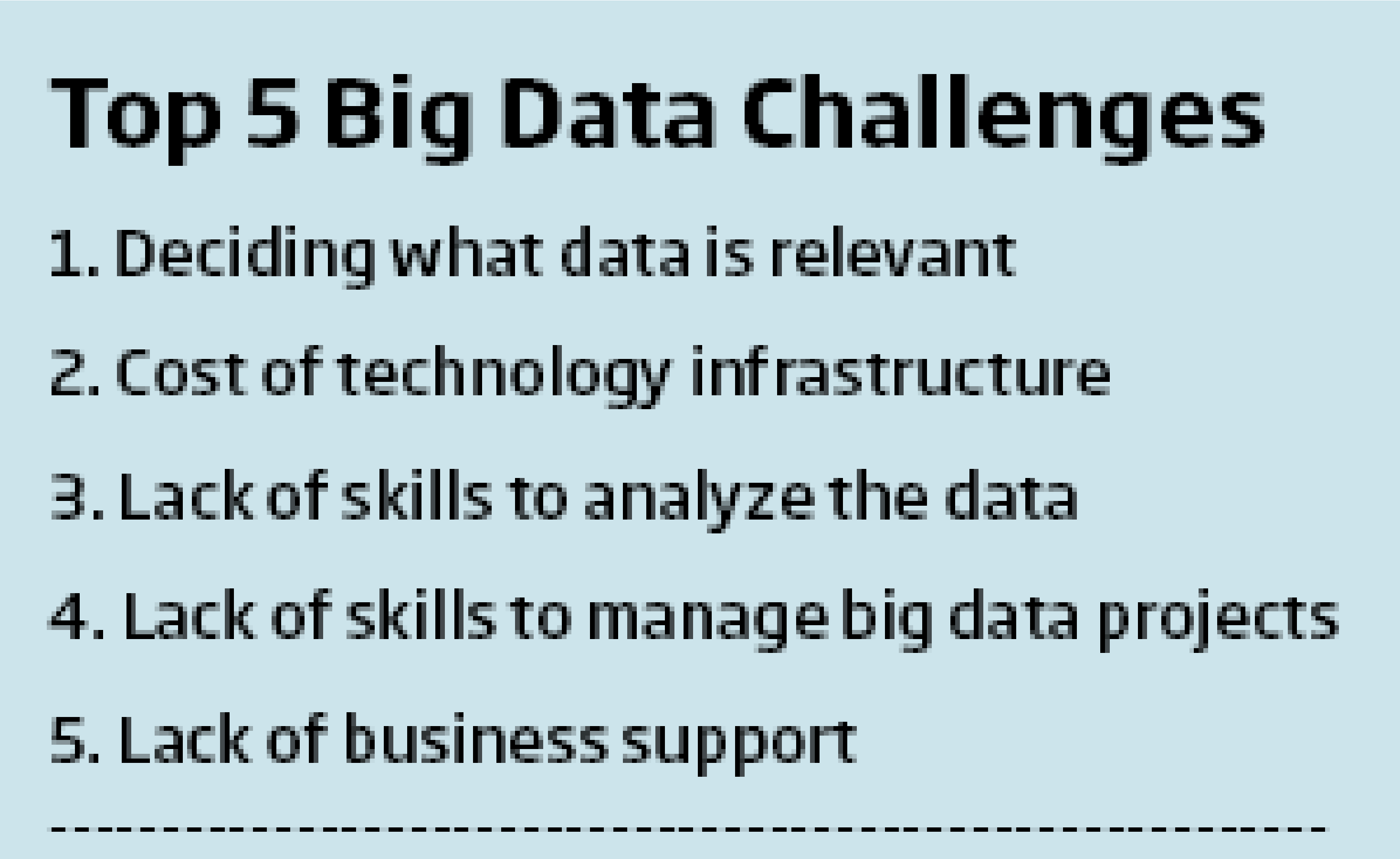
facebook
Google
amazon.com
twitter
Y!
Linked in
Bai du 百度

Big Data



Attributes of Big Data

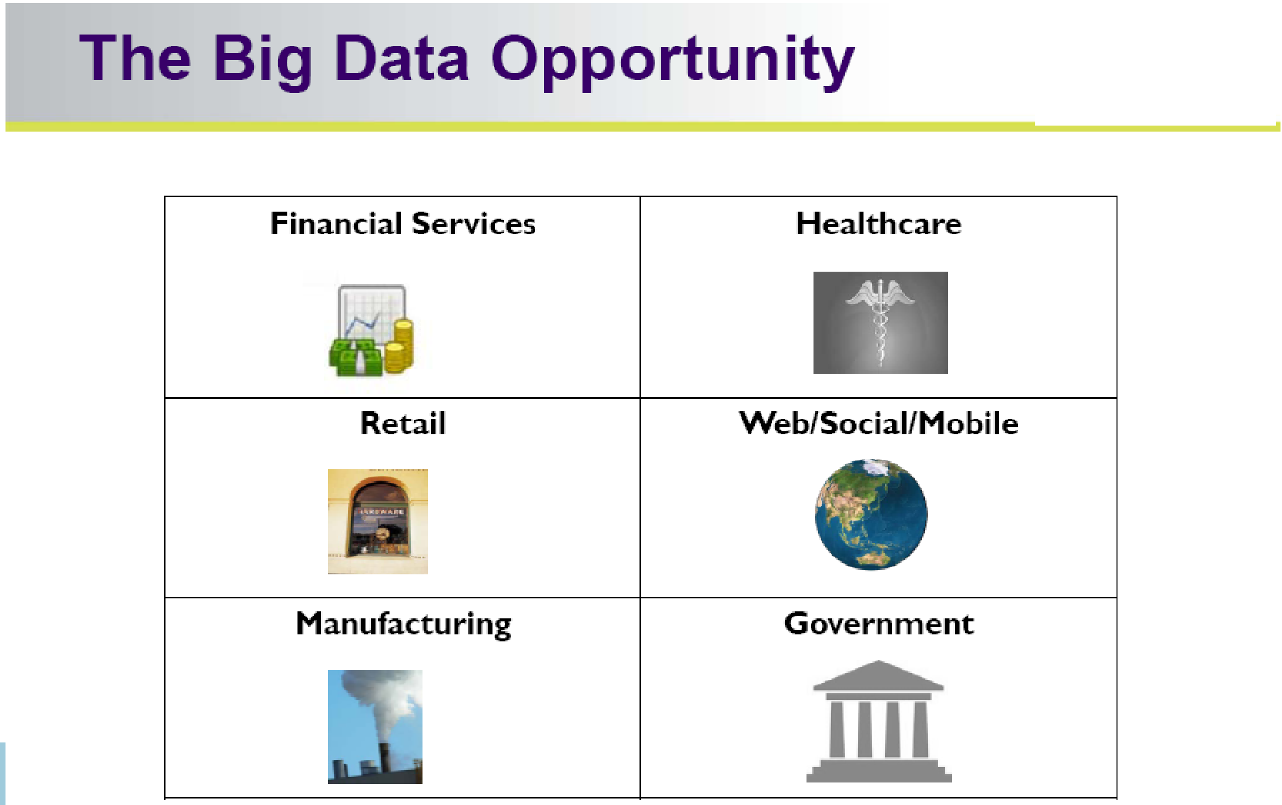
Big Data Challenges



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



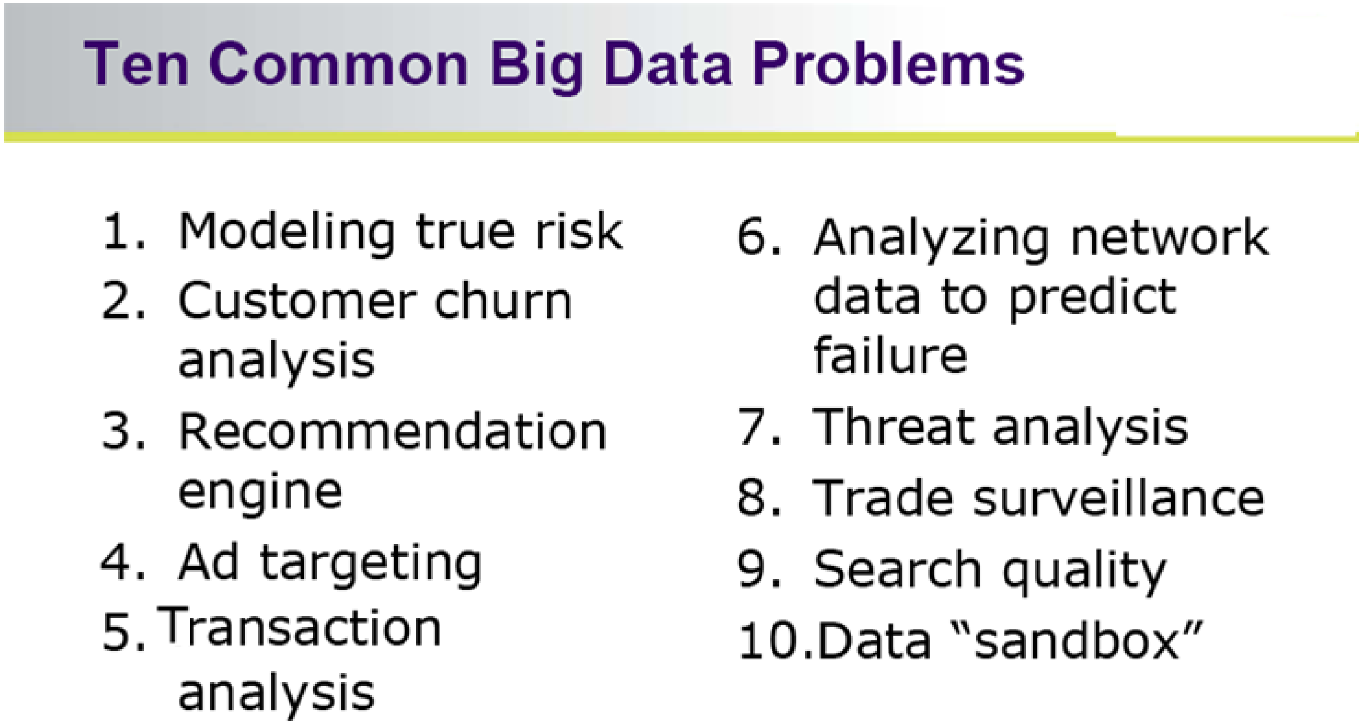
The Big Data Opportunity

Financial Services	Healthcare
Retail	Web/Social/Mobile
Manufacturing	Government

Big Data

Enterprise + Big Data = Big Opportunity

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”

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.*

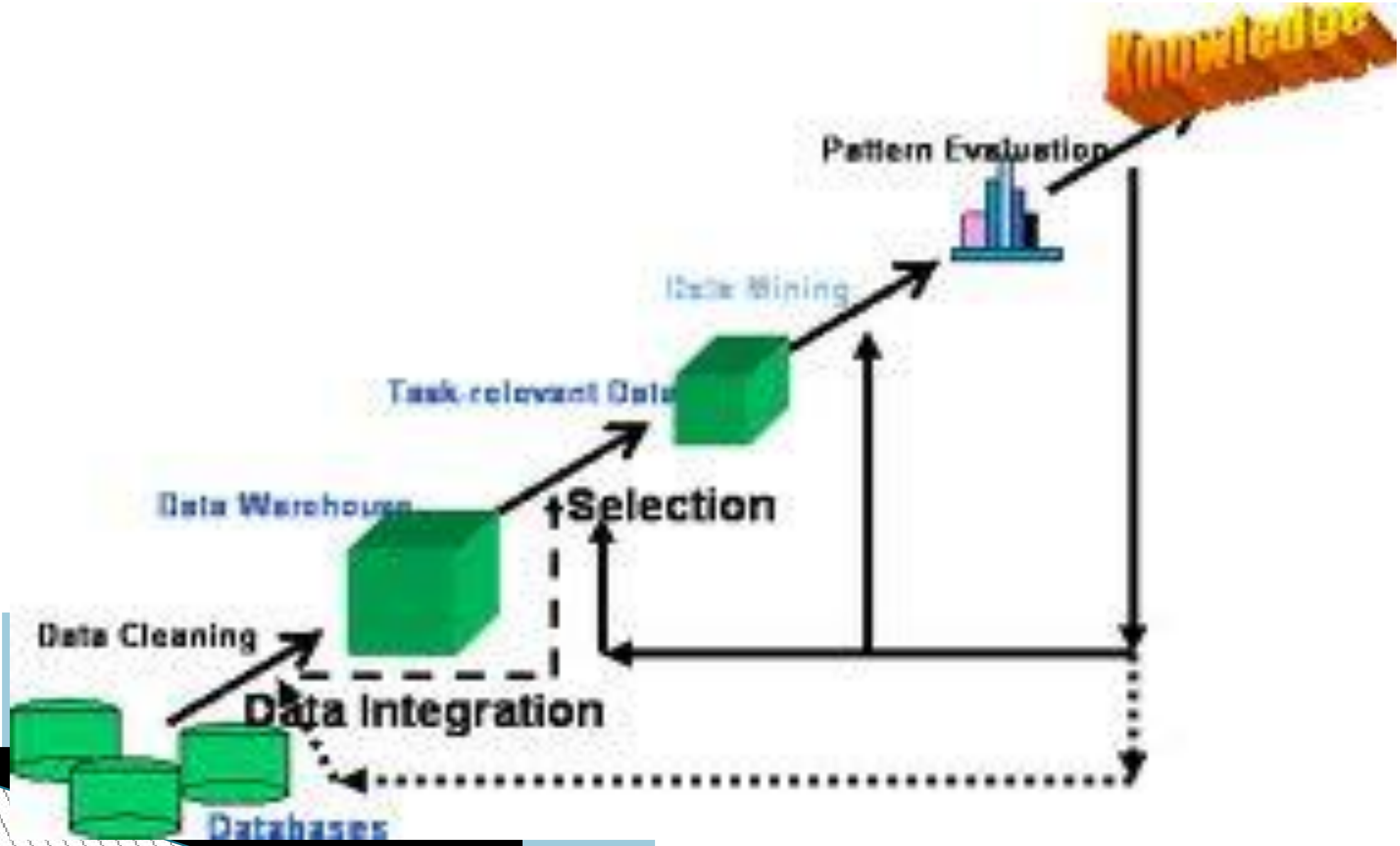
- *The focus on lack of follow-up with patients, medication management issues and insufficient coordination of care.*

- *Data is preprocessed to correct any errors and to format it for analysis.*

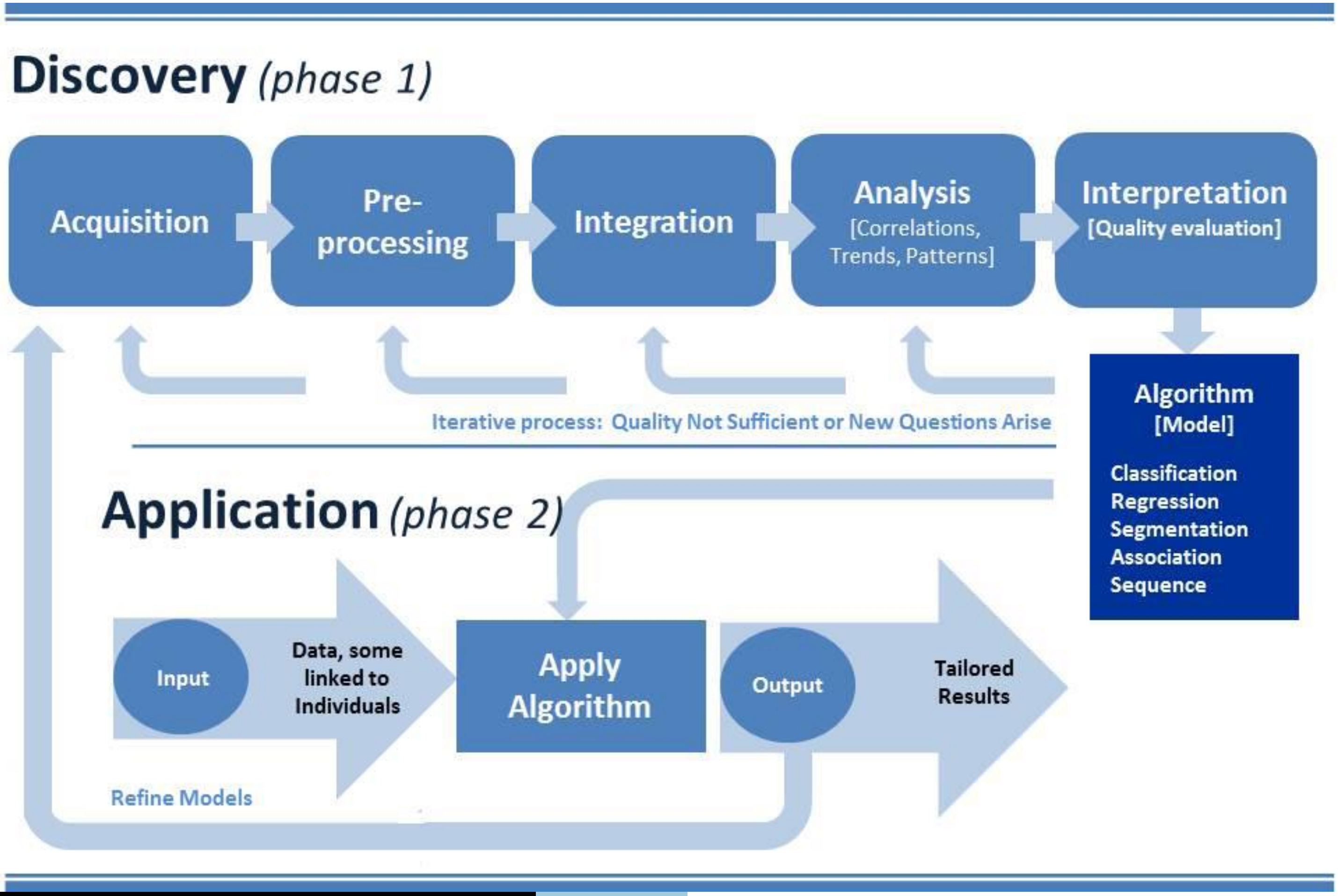
The Power and Promise of Analytics

- ***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.*
- [redacted]

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.

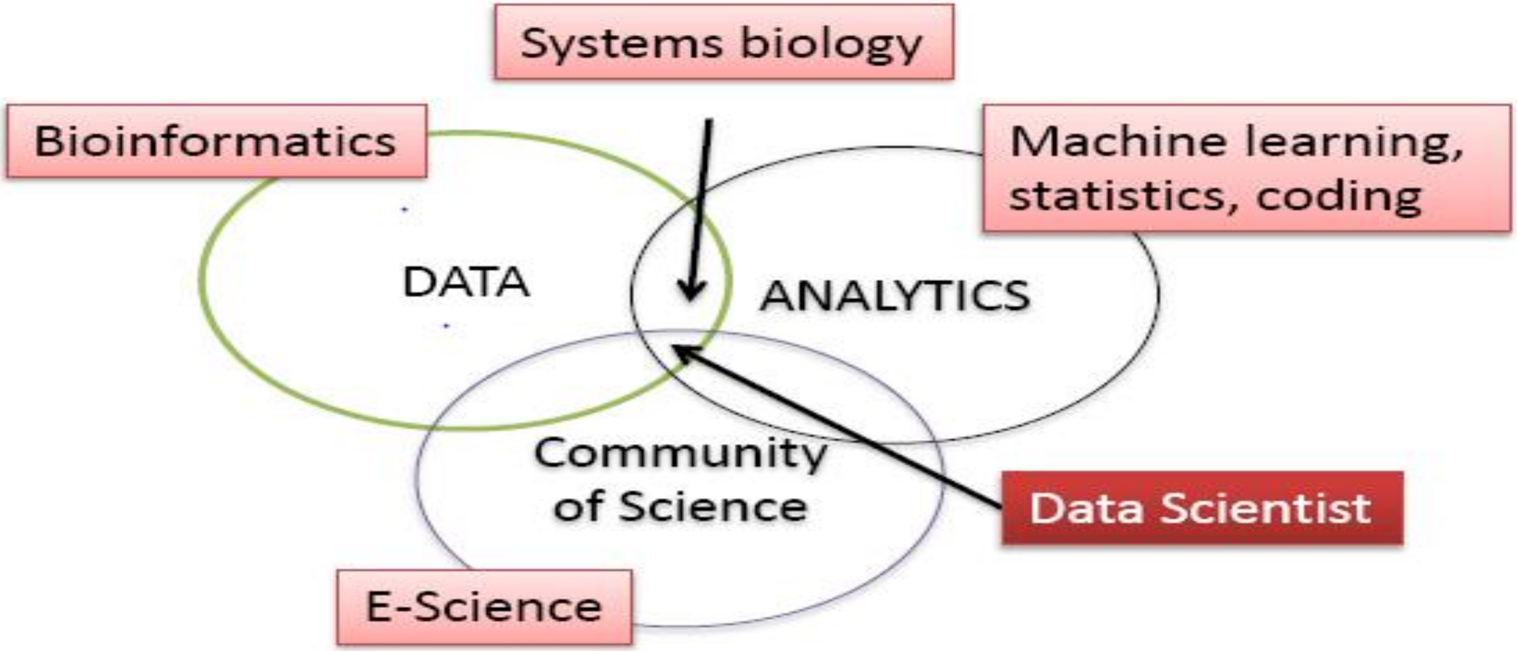
  - ◦ Knowledge discovery may reveal that data could provide additional insights, and researchers may choose to explore them further. Data used for analytics may come from an organization's own stores, but may also be derived from public records.

  - ◦ 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.*

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 structured/semi-structured/unstructured data (Data analytics)
- development of algorithm (Data analytics)

- Data refinement (Data scientist)
- Algorithm implementation 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.
- QA verification.
- Releaes.

Q&A/Feedback?