WWW.CIIA.XYZ

CII.IA®

Introduction to BigData

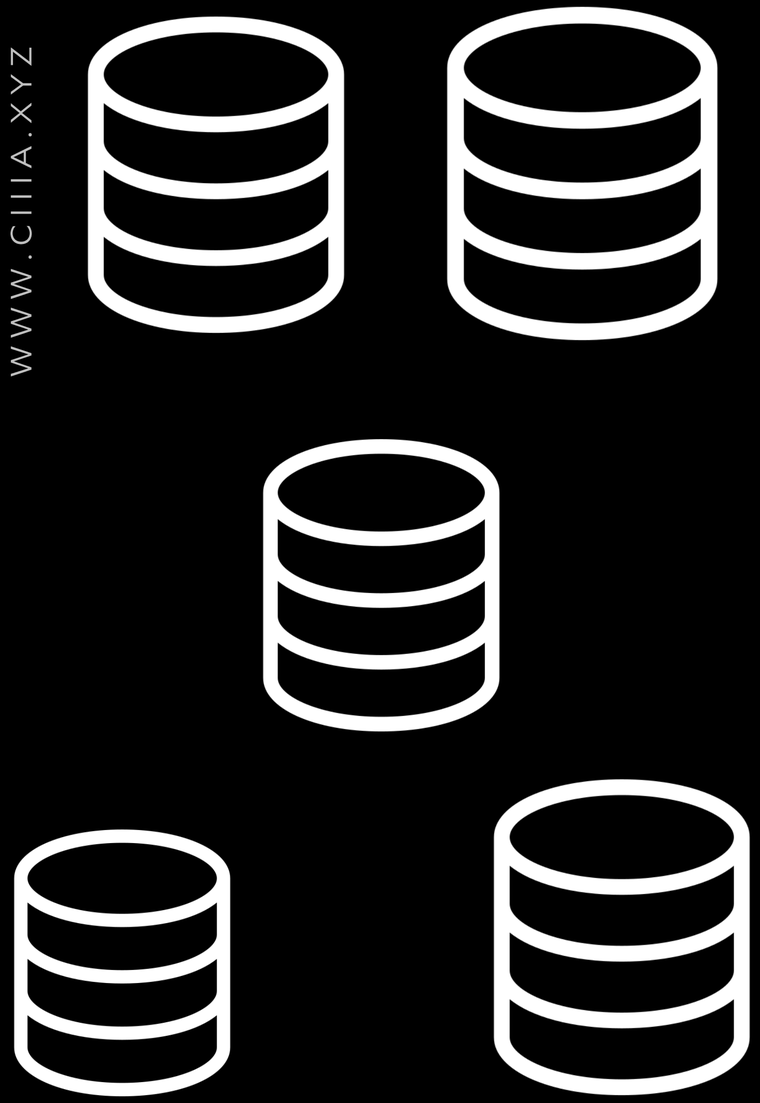
- Dra. MaryPaz Rico Fernández

WWW.CIIIA.XYZ

Big Data= ?

CII.IA®

WWW.CIIIA.XYZ



BigData= ?

- “Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” (McKinsey Global Institute)

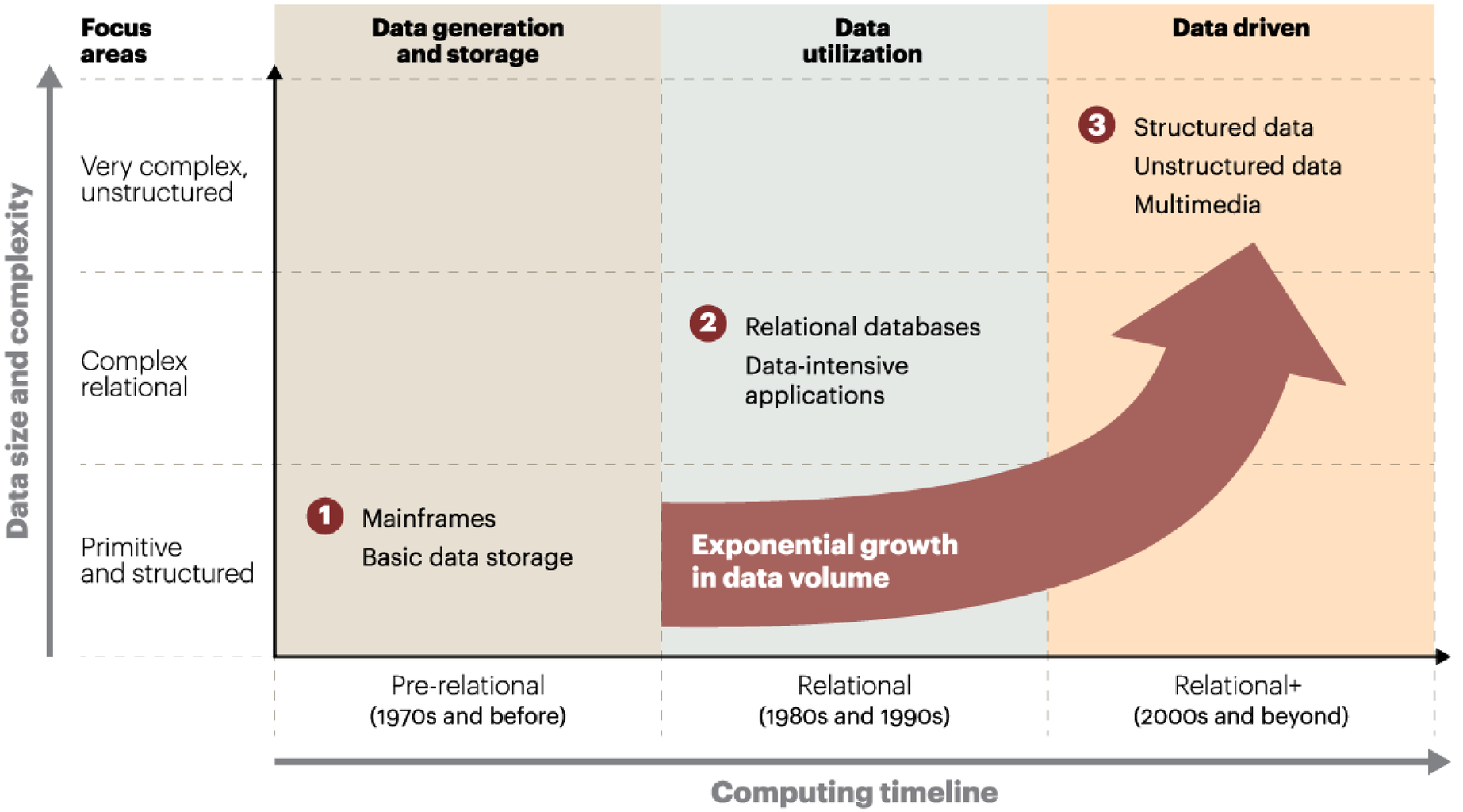
- “Big Data is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” (Wikipedia)

CII.IA®

WWW.CIIIA.XYZ

EVOLUTION OF BIG DATA

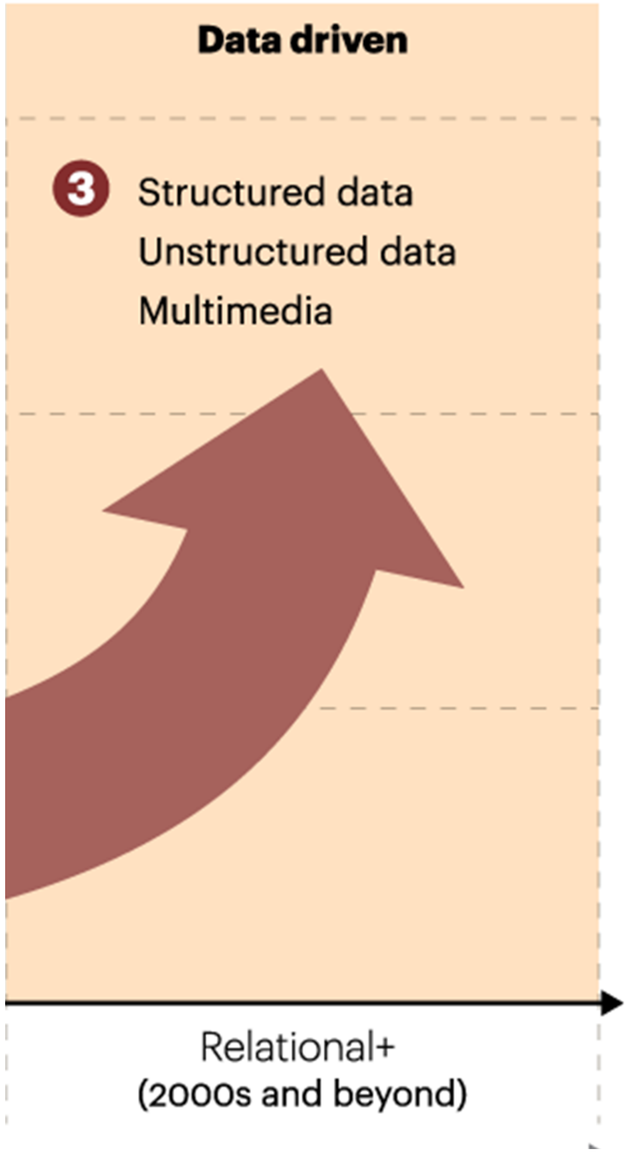
EVOLUTION OF BIG DATA



Source: A.T. Kearney analysis

HISTORICAL CONTEXT

- The AI Winter: before 2006 people thought deep neural networks cannot be trained.
- Theoretical breakthroughs in 2006 Learned how to train deep neural networks
- Cheap computational power – GPUs can run neural networks in parallel
- Big data – collecting massive data about the world – used for “schooling” ML



HISTORICAL CONTEXT

THE FIRST BIG DATA PROBLEM

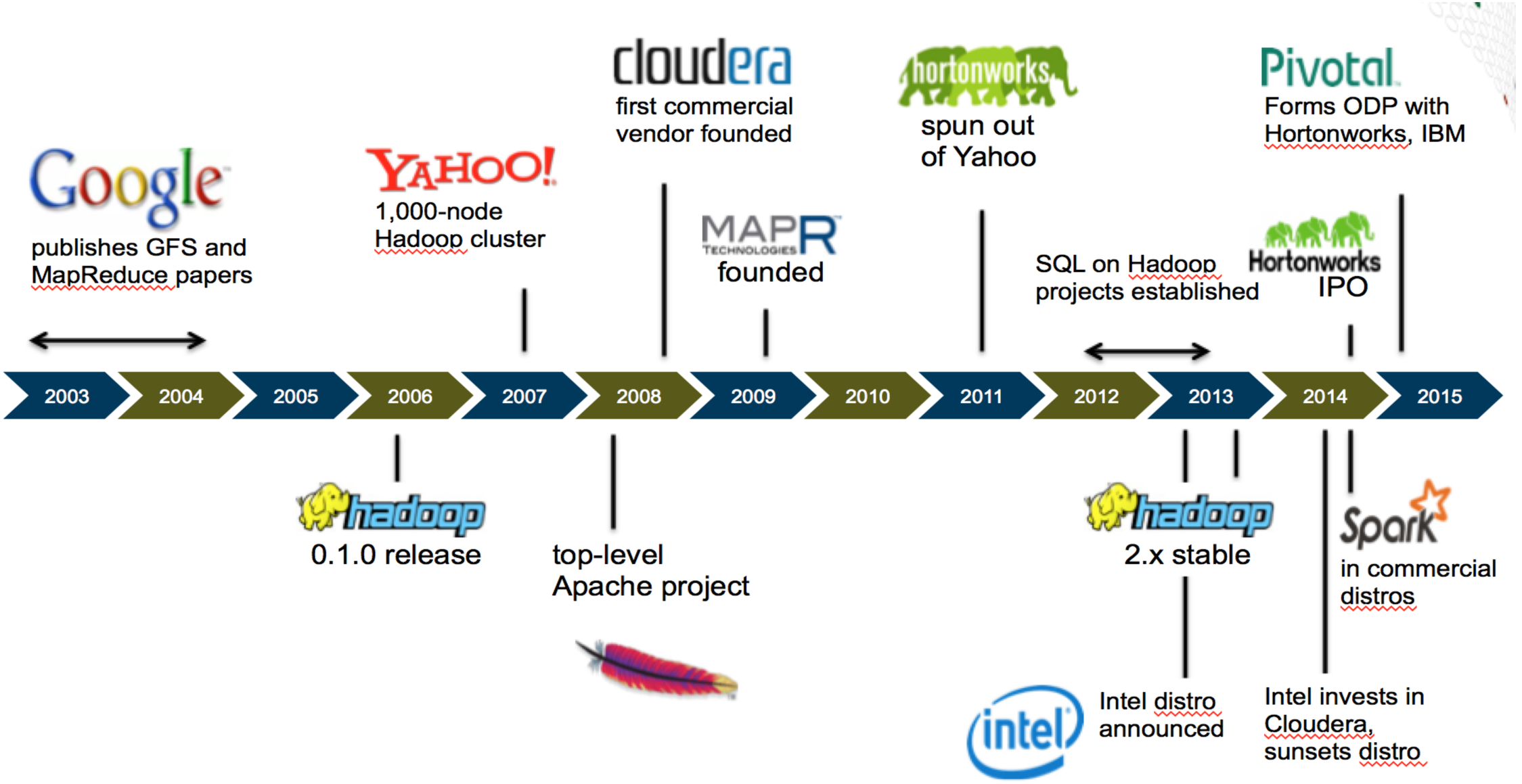
The Internet Search Engine:

In 2002, Google wanted to be able to crawl the web and index the content so that they could produce an internet search engine.

The standard method to organize and store data in 2002 was by means of relational database management systems (RDBMS) which were accessed in a language called SQL.

But almost all SQL and relational stores were not appropriate for internet search engine storage and retrieval because they were costly, not terribly scalable, not as tolerant to failure as required and not as performant as desired.

HISTORICAL CONTEXT



HISTORICAL CONTEXT

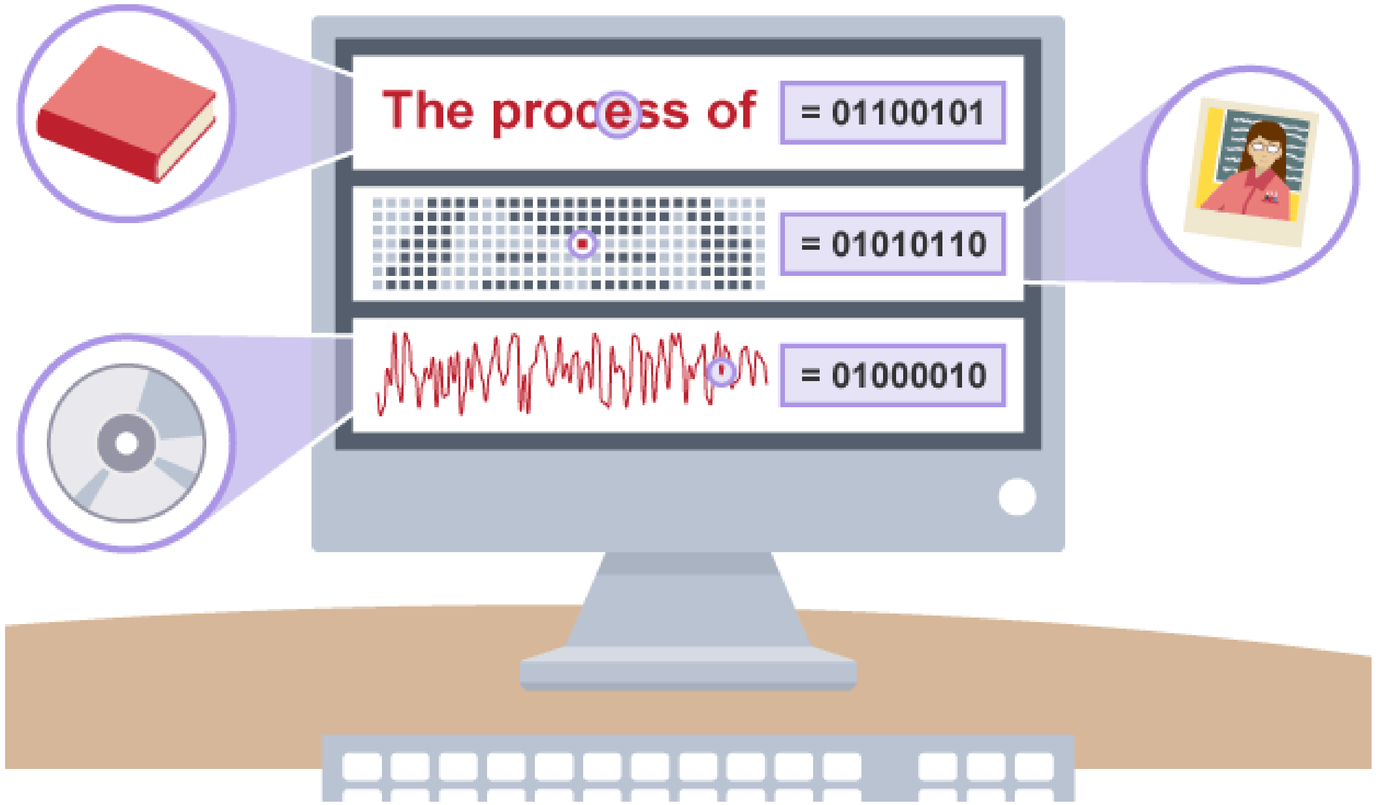
WWW.CIIA.XYZ

Johan Oskarsson, then a developer at Last.fm, reintroduced the term **NoSQL** in early **2009** when he organized an event to discuss **"open source distributed, non relational databases".** The name attempted to label the emergence of an increasing number of non-relational, distributed data stores, including open source clones of Google's Bigtable/MapReduce and Amazon's DynamoDB.[2]

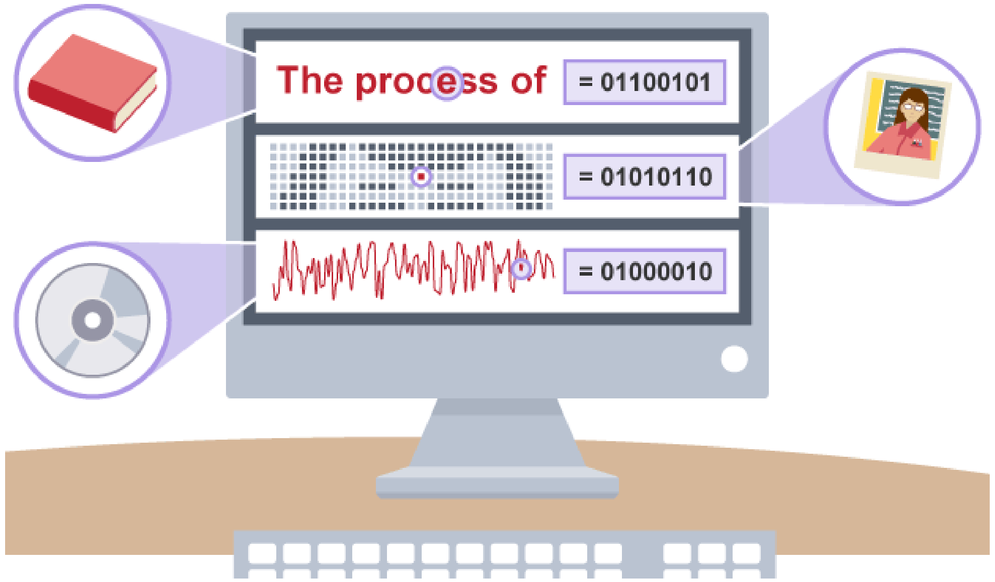
WWW.CIIIA.XYZ

BIG DATA

WHAT IS DATA?

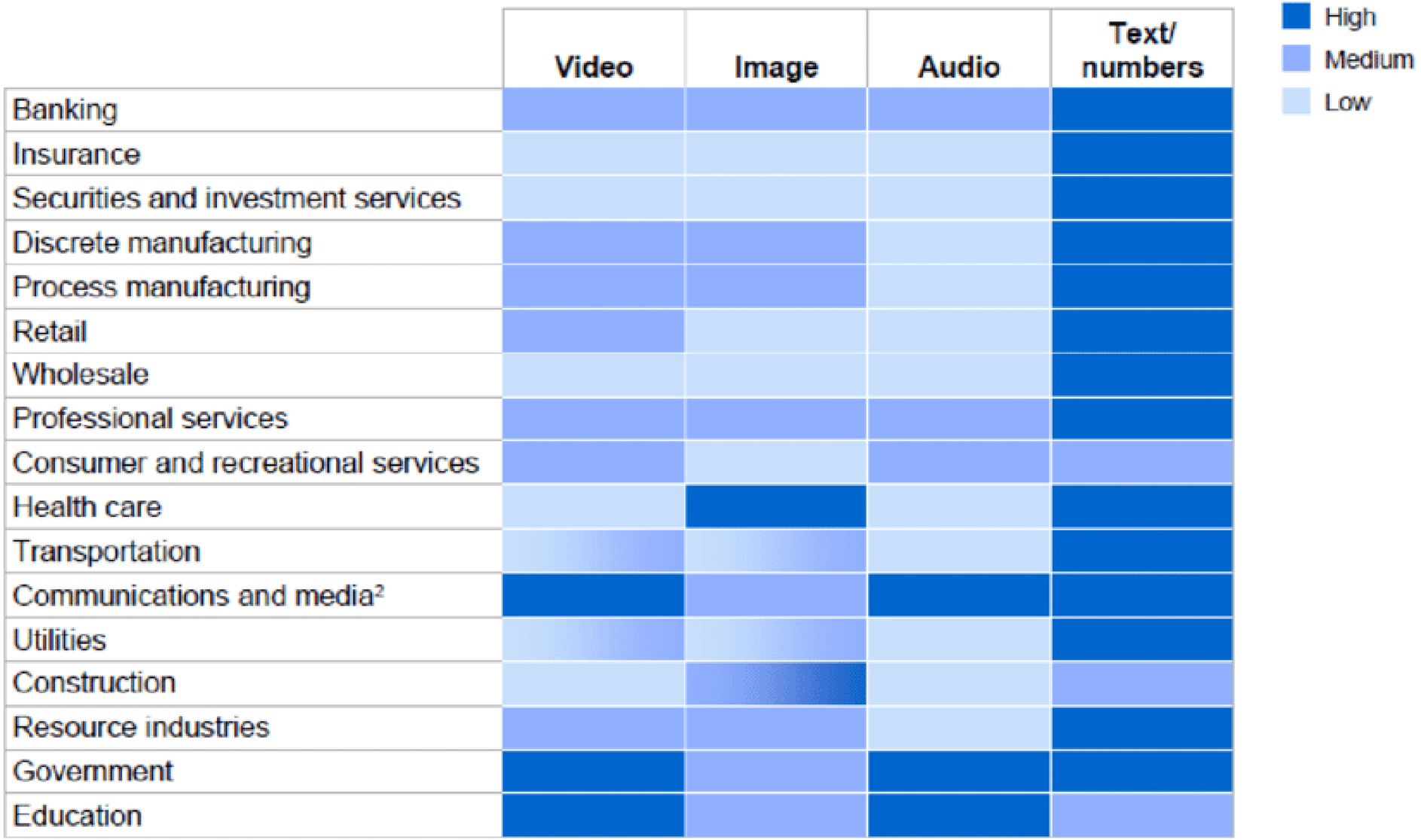


WHAT IS DATA?



Collection of different Variables and types from multiple sources.

TYPES OF DATA



1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

SOURCES OF DATA

Three types:

* Directed

* Automated

* Volunteered

SOURCES OF DATA

Three types:

- **Directed**

Organized and structured surveillance– personal or through technological lens, census, government forms, inspections

- **Automated**

Automated surveillance, smart electricity meters, electronic transportation tickets, passenger counting systems, car tolls, radar/LiDAR speedguns

- **Volunteered**

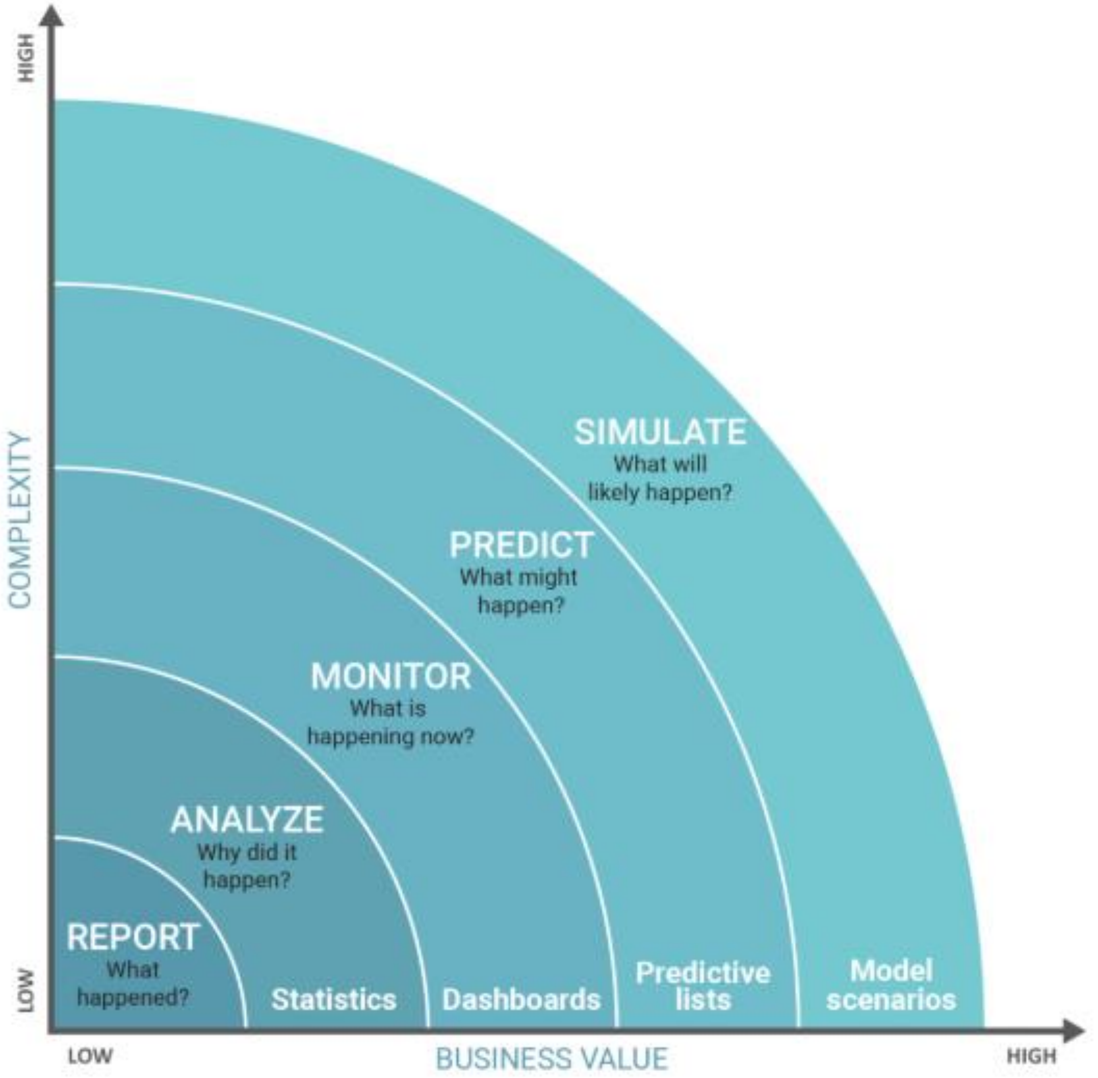
Smartphones/tablets, cameras, videos, GPSunits, medical devices

ANOTHER TYPES OF DATA

- Interaction data – web shopping, net banking..
- Scan data – machine-readable identification codes, chip card/smart card
- Sensor (sensed) data – continuous data streams from smart cities noise, temperature, light, CO2 ..

MAIN ACTIVITIES OF DATA

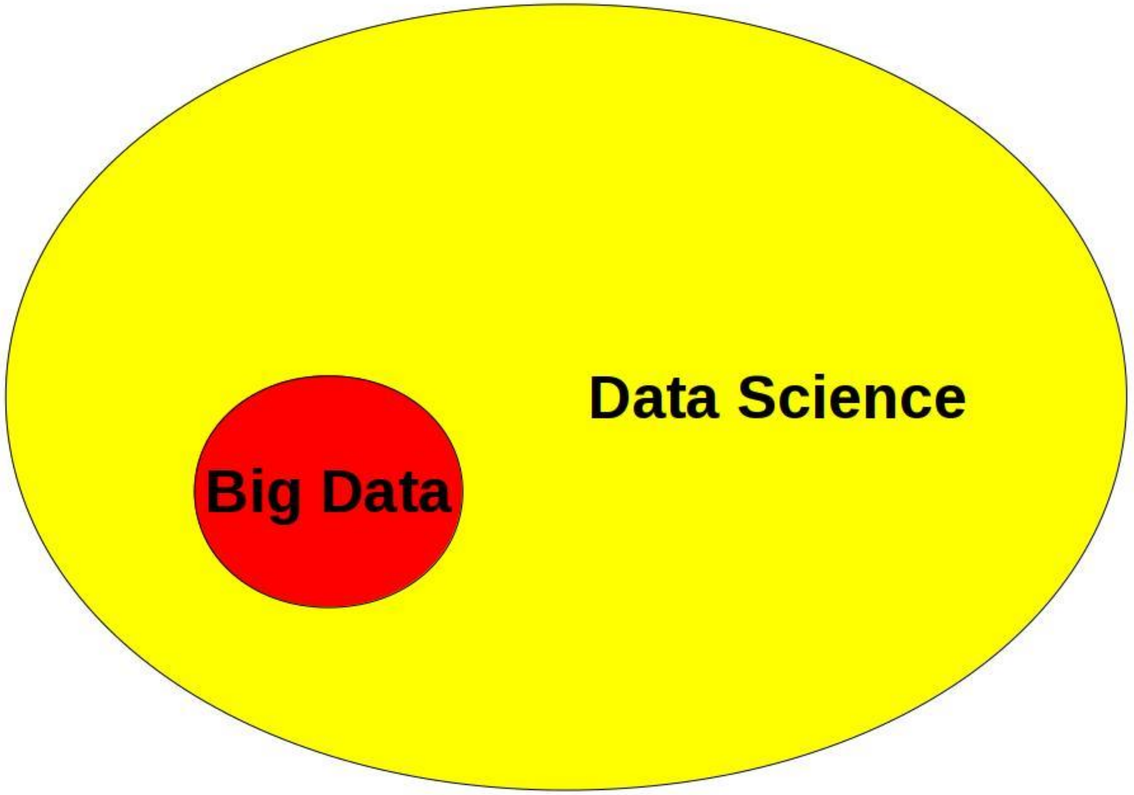
MAIN ACTIVITIES OF DATA



source: www.predictivelayer.com

BIG DATA UNDER THE DATA SCIENCE HOOD

Part of data science with efficient storage, extracting, processing, and analyzing information



CHARACTERISTICS OF BIG DATA: FIVE VS MODEL

Volume

Velocity

Variety

Value

Veracity

CHARACTERISTICS OF BIG DATA: FIVE VS MODEL

Volume

Large amounts of data generated every second (emails, twitter messages, videos, sensor data...)

Velocity

-The speed of data moving in and out data management systems (videos going viral...)

-“on-the-fly”

Variety

-Different data formats in terms of structured or unstructured (80%) data

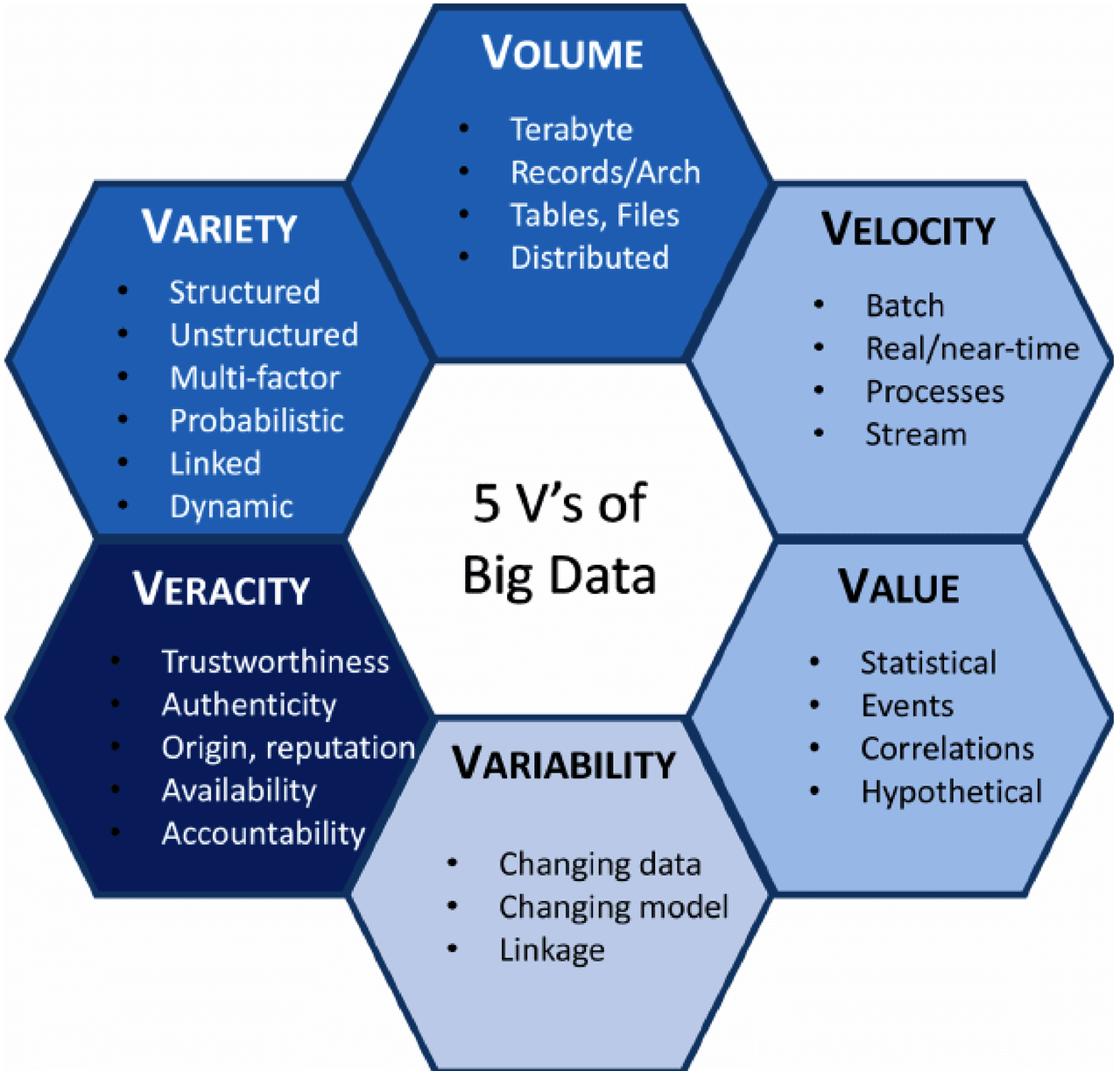
Value

-Insights we can reveal within the data

Veracity

-Trustworthiness of the data

CHARACTERISTICS OF BIG DATA: FIVE VS MODEL



BIG DATA ANALYTICS CHARACTERISTICS

- Value: produced when the analytics output is put into action
- Quality: well-formed data, Missing values and cleanliness
- Latency: time between measurement and availability

BIG DATA CHALLENGES

* Data Quality

* Data Discovery – Finding insights on Big Data

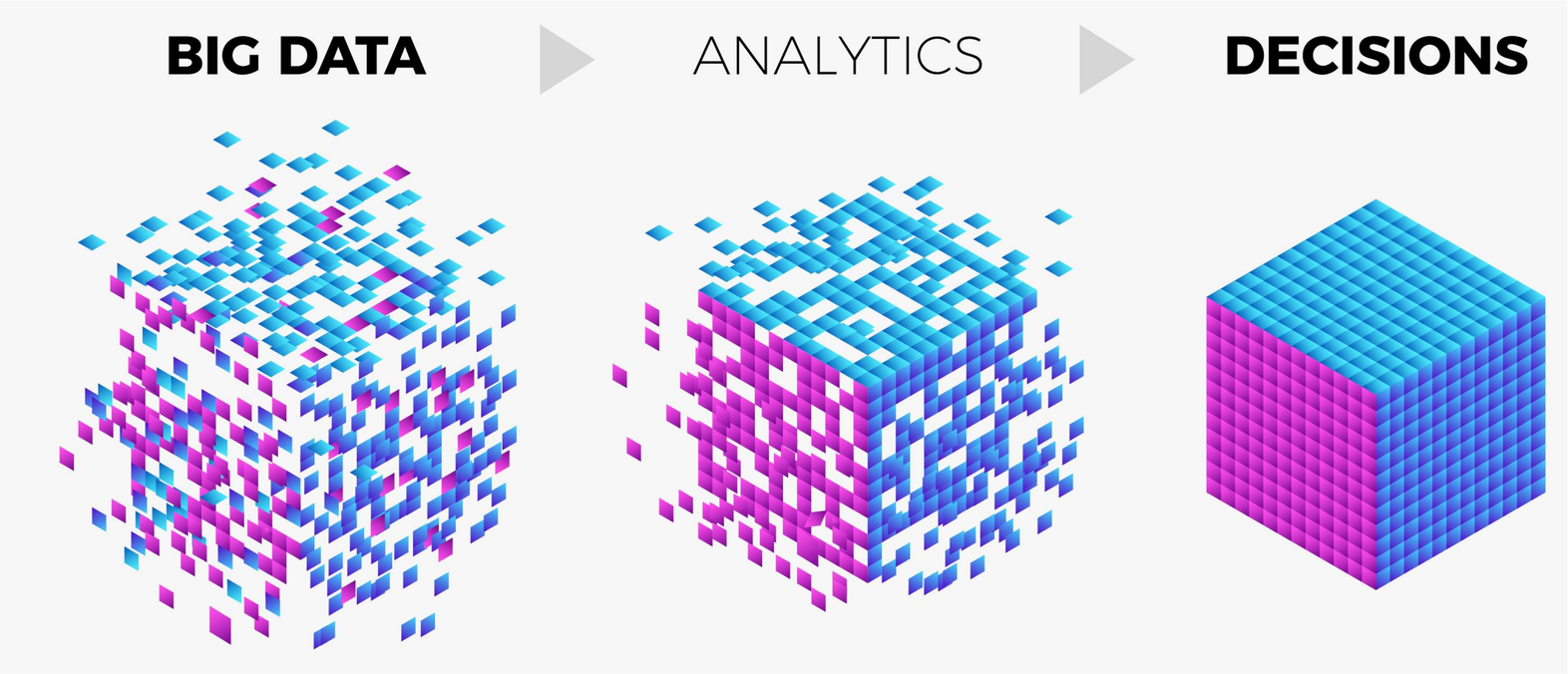
* Data Storage – complex problems of managing

* Data Analytics- advanced modelling and research

* Data Security

* Talent Acquisition – a sophisticated team of developers, data scientists and analysts along with Domain expertise

SKILLS REQUIRED FOR BIG DATA ANALYTICS



- Data processing (Analyst)
- Analyse and Modelling (Data Science and Statistics)
- Understand and Design (Data Engineering and Product)

SKILLS REQUIRED FOR BIG DATA ANALYTICS

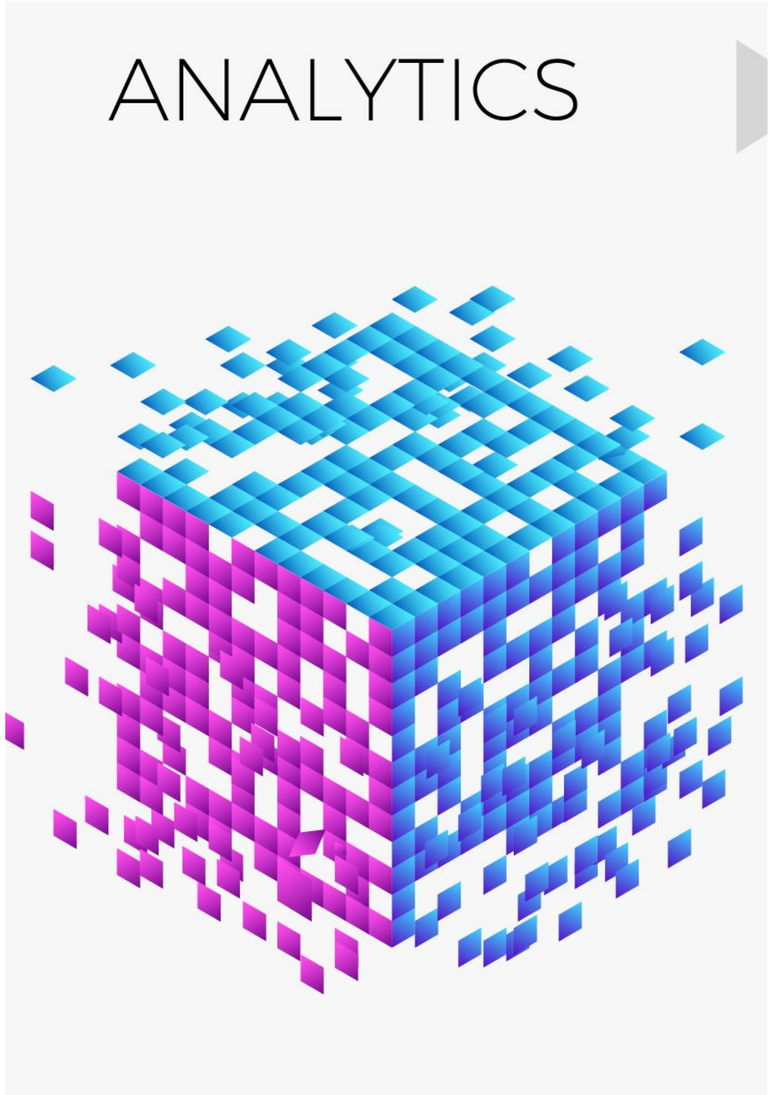
Data processing (Analyst)

- Large scale databases
- Software Engineering
- System/network Engineering

SKILLS REQUIRED FOR BIG DATA ANALYTICS

Analyze and modelling (Data Science and Statistics)

* Reasoning
* Knowledge Representation
* Multimedia Retrieval
* Modelling and Simulation
* Machine Learning
* Information Retrieval



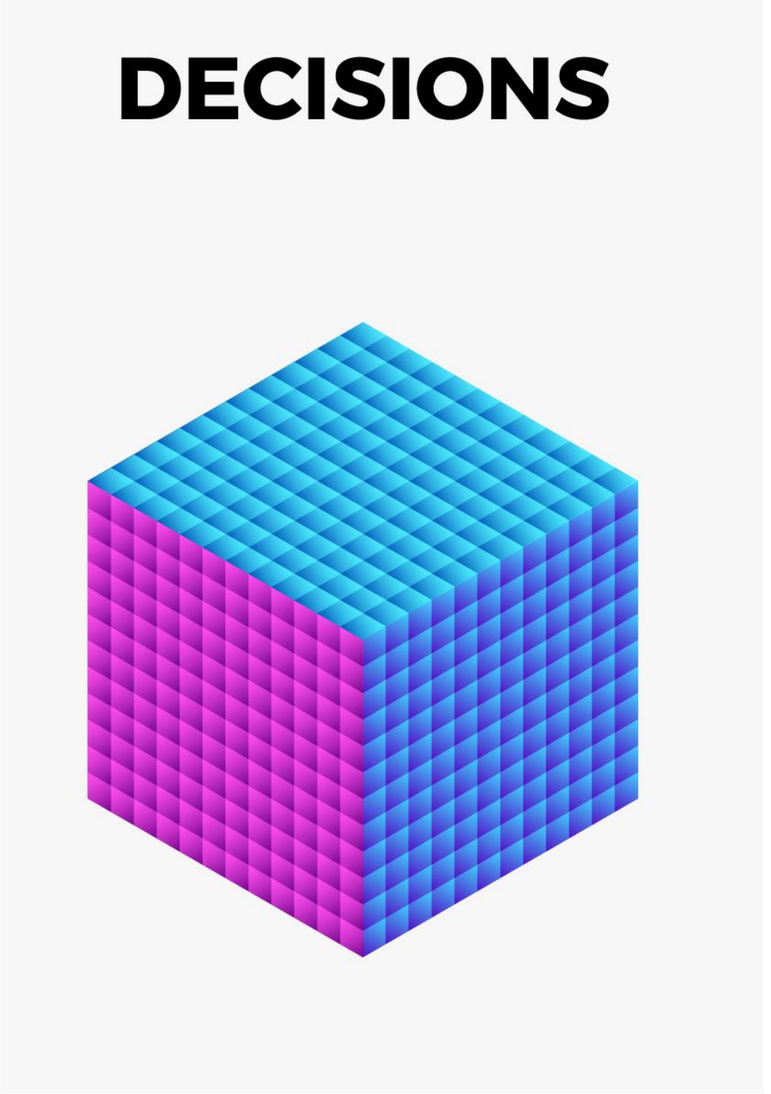
SKILLS REQUIRED FOR BIG DATA ANALYTICS

Understand and design (Data Engineering and Product)

Decision theory

Visual analytics

Perception Cognition



BIG DATA PLATFORM: SIX KEY IMPERATIVES

The Big Data platform manifesto: Imperatives and underlying technologies

* Data Discovery and Exploration

* Extreme Performance: Run Analytics Closer to the Data

* Manage and Analyze Unstructured Data

* Analyze Data in Real-Time

* A Rich Library of Analytical Functions and Tools

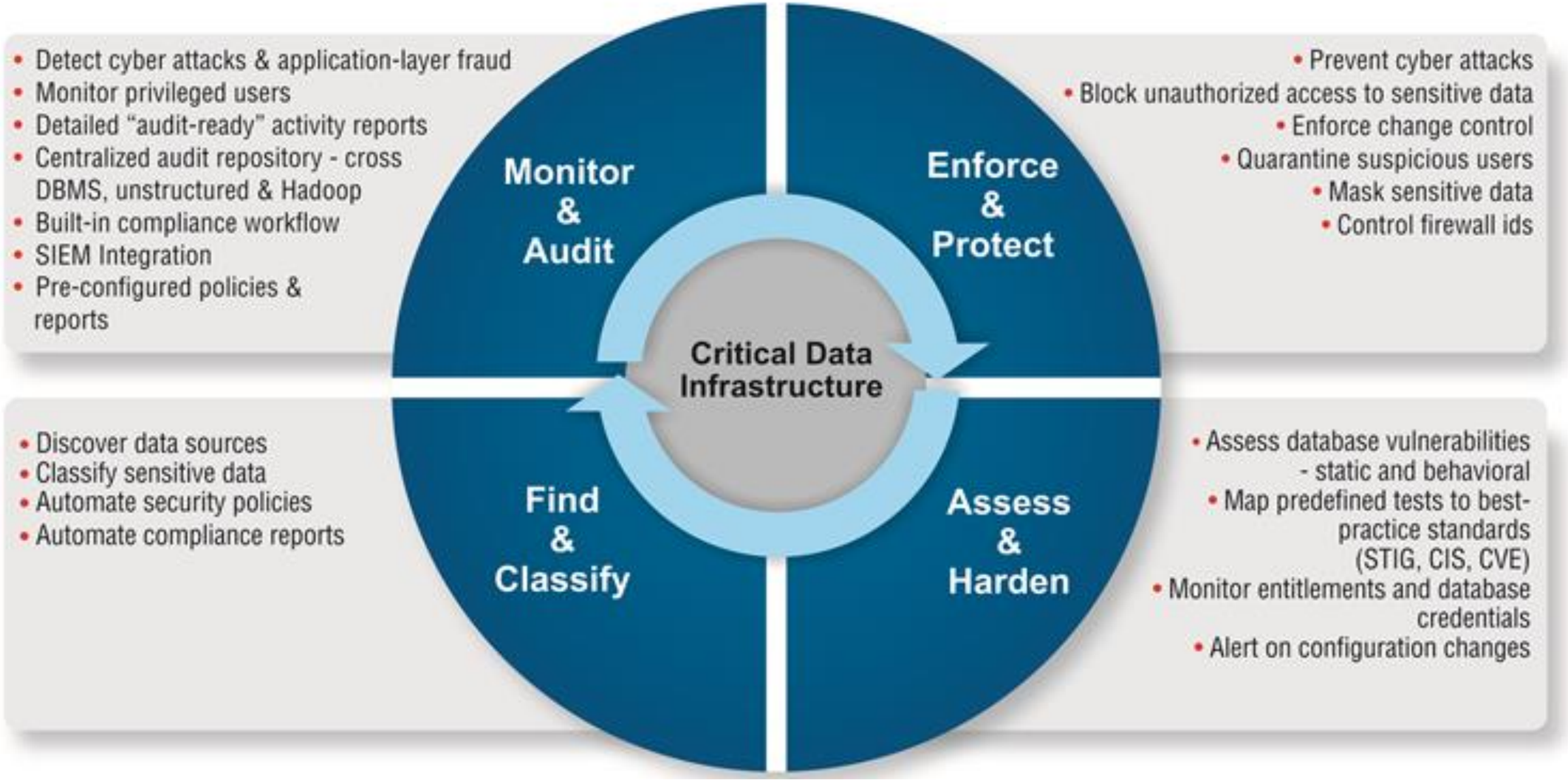
* Integrate and Govern All Data Sources

WHEN TO CONSIDER A BIG DATA SOLUTION

- Ingest data as quickly as possible and analyze
- Not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources
- For improved effectiveness of your algorithms or models: usage of all available Data

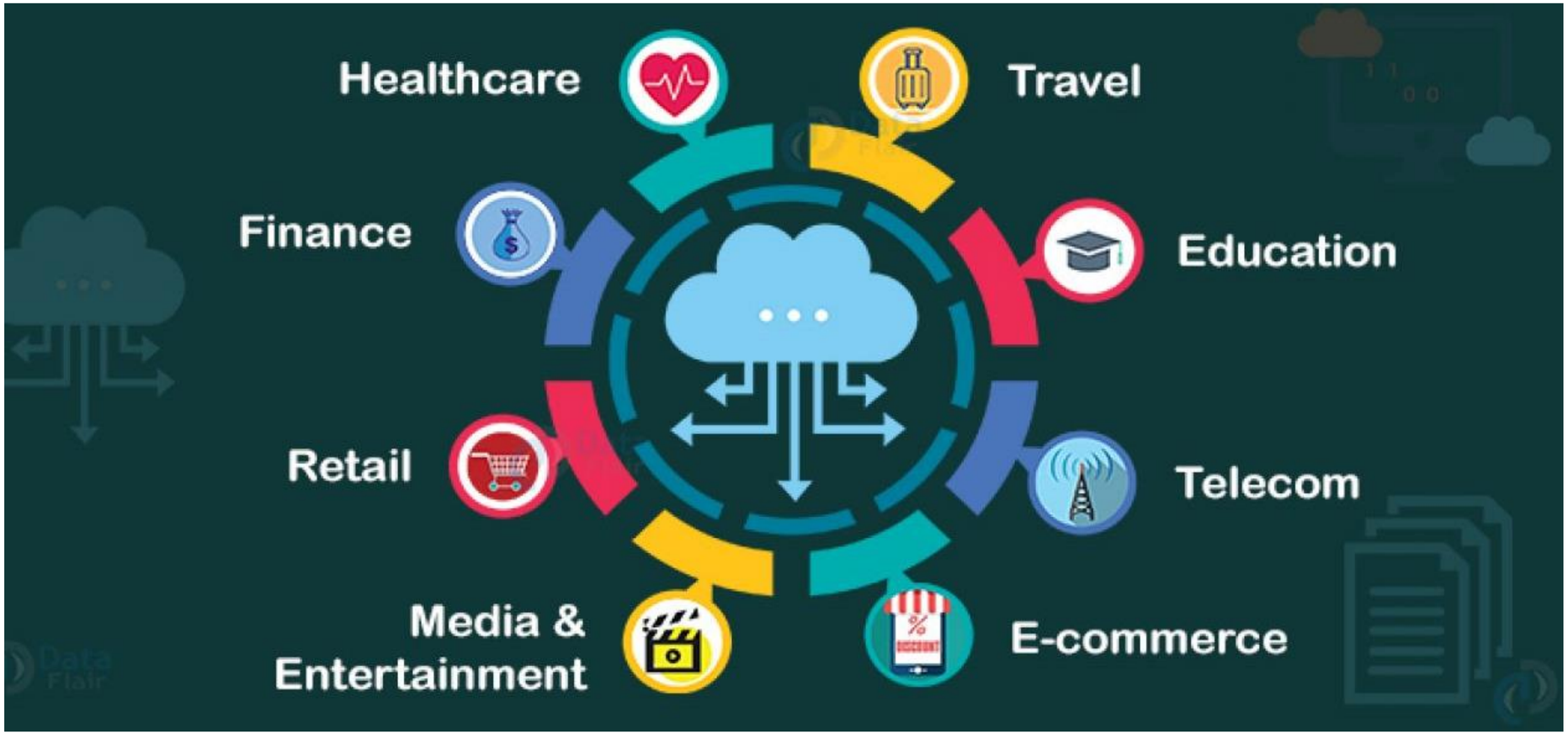
CHALLENGES

Address the full data security and compliance lifecycle



BIG DATA APPLICATIONS

BIG DATA APPLICATION



BIG DATA APPLICATIONS

Walmart

Save money. Live better.

WHEN TO CONSIDER A BIG DATA SOLUTION

Rank	Company Name	Revenues ($b)
1	 Wal-Mart Stores	469.2
2	 Exxon Mobil	449.9
3	 Chevron	233.9
4	 Phillips 66	169.6
5	 Berkshire Hathaway	162.5
6	 Apple	156.5
7	 General Motors	152.3
8	 General Electric	146.9

Some data:

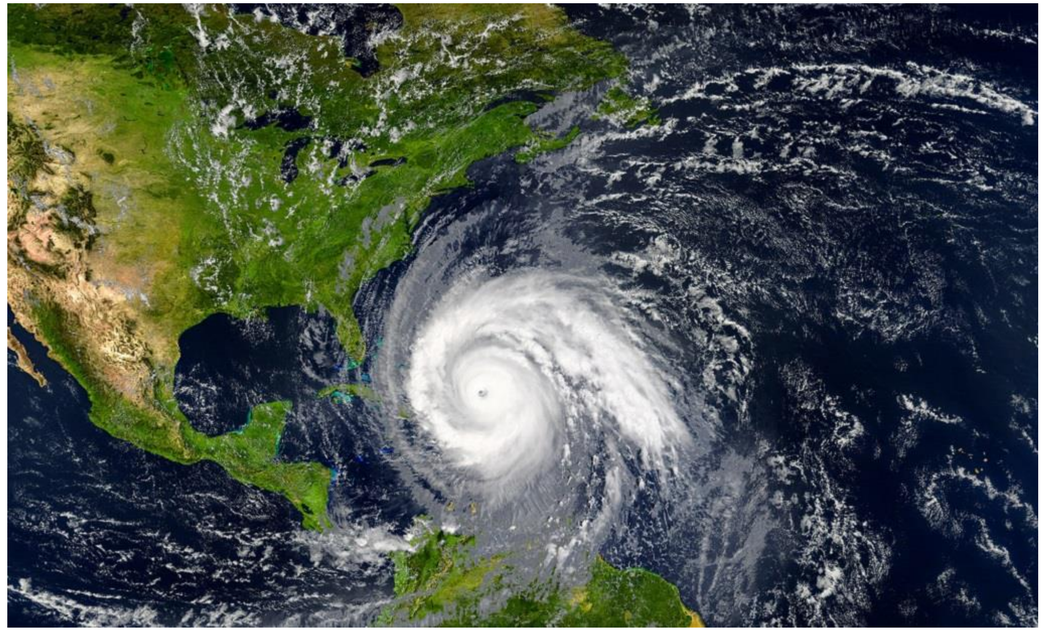
- 200 millons clients per week

- 10700 shops
- 27 countries
- 2 millons workeres
- 1,5 millons transactions per hour
- Many Terabytes of Information in real time
- Databases 3 Petabytes

BIG DATA APPLICATIONS

Walmart

Save money. Live better.



BIG DATA APPLICATIONS

FredCavazza.net

Social media ->90%of today's data

Fastest means of population feedback and provides cheap two-way information sharing

Personal nature of the data

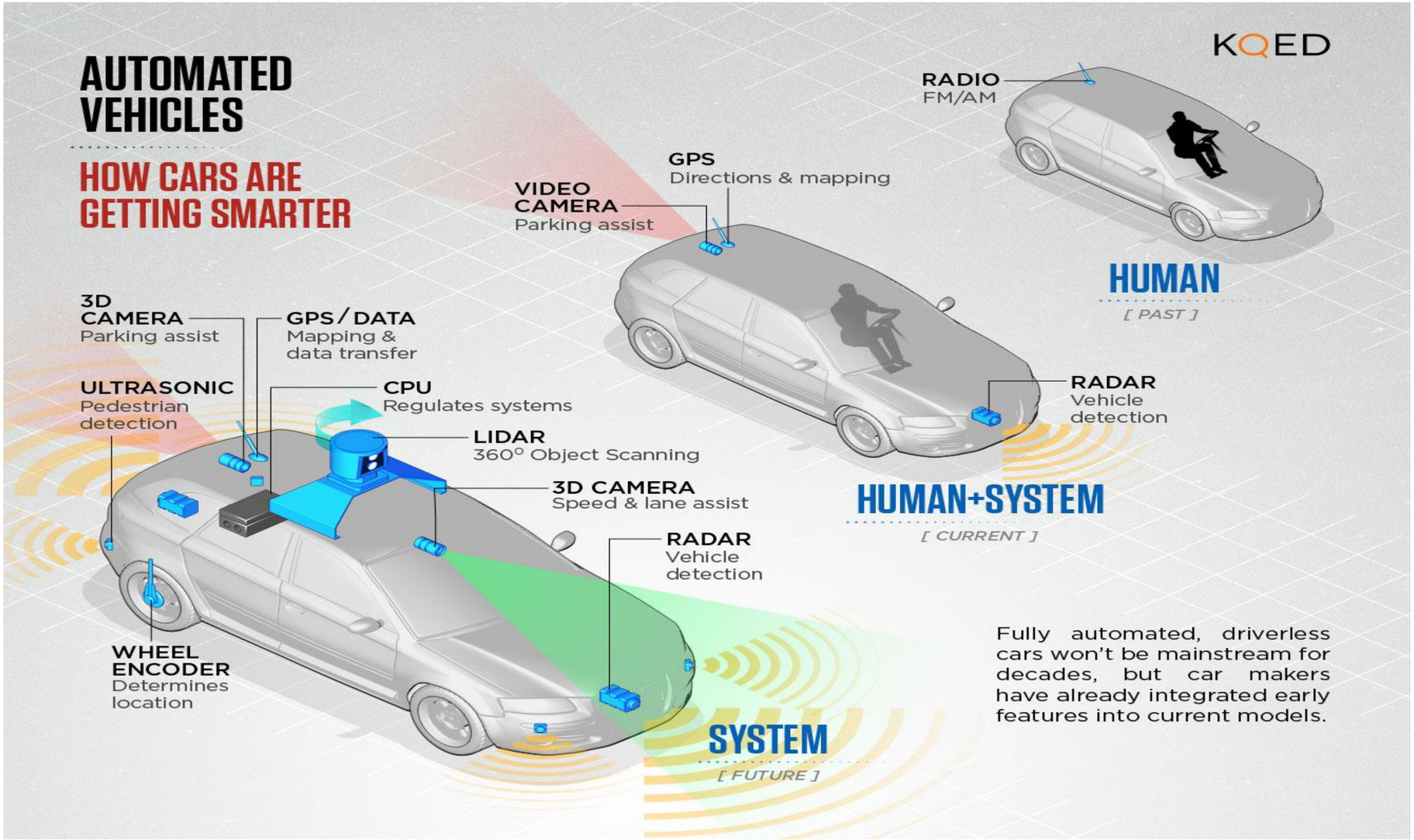
BIG DATA APPLICATIONS

FredCavazza.net

- More than 204 million email messages
- Over 2 million Google search queries
- 48 hours of new YouTube videos
- 684,000 bits of content shared on Facebook
- More than 100,000 tweets
- $272,000 spent on e-commerce

BIG DATA APPLICATIONS

•Development of the technologies for automated driving, telematics and mobility



BIG DATA APPLICATIONS

For Social Good

- GiveDirectly: Direct cash transfers to low-income families in Uganda and Kenya through mobile payments
- Using Data of kind of roofing, the IBM came up with algorithm to identify the needed household

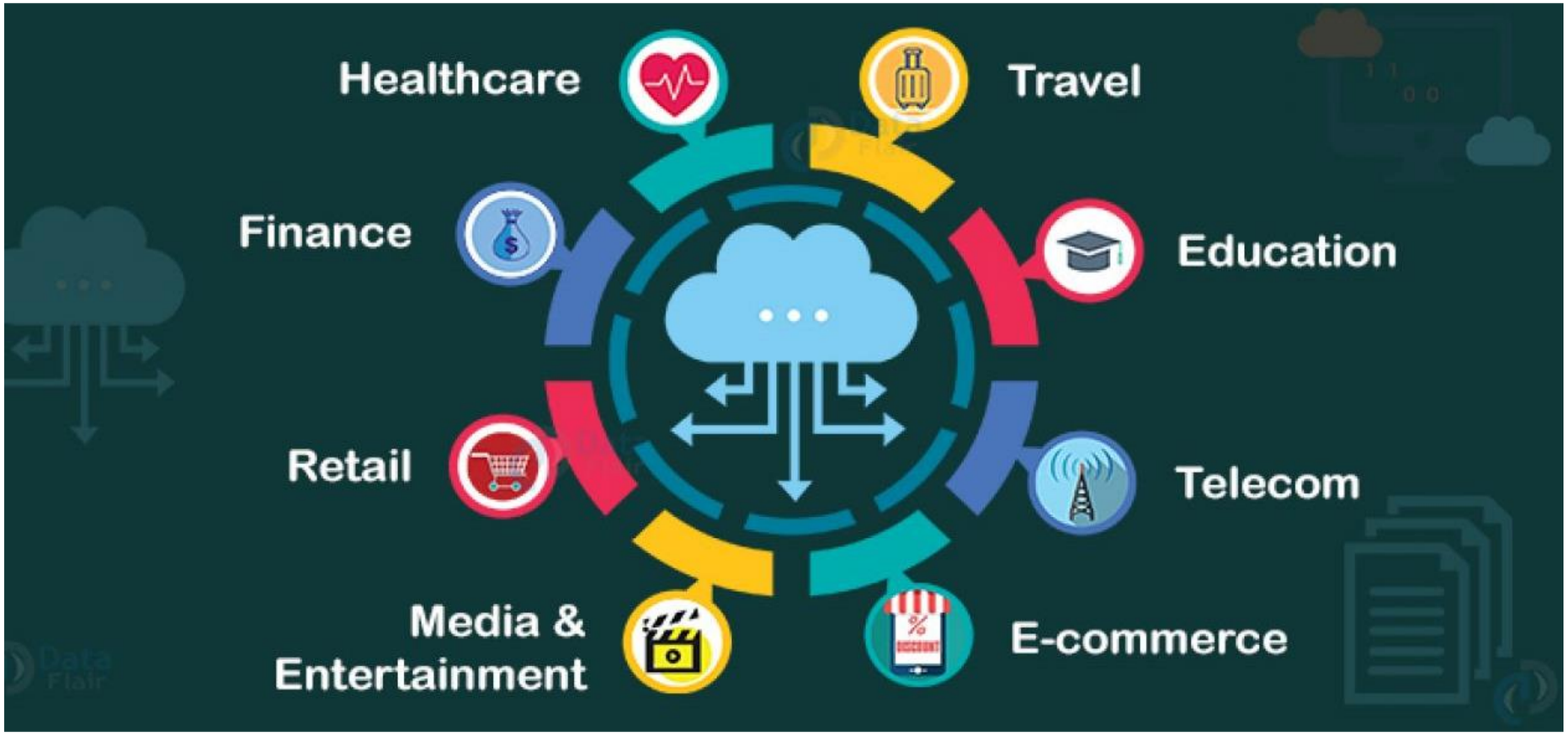
BIG DATA APPLICATIONS

For Social Good

Apollo Agriculture startup that’s revolutionizing Kenyan agriculture

Using machine learning, remote sensing, and mobile phones to deliver financing, farm products, and customised advice to smallholder farmers with radical efficiency and scalability

BIG DATA APPLICATION



WHAT COMPANIES USE BIG DATA

Multinational Companies in the Big Data / Data Analytics Space

WWW.CIIIA.XYZ

BIG DATA TECHNOLOGIES

BIG DATA TECHNOLOGIES

Processing and analysis of huge data sets is not feasible computationally

Special techniques and tools (e.g., software, algorithms, parallel programming, etc.)

Hadoop=

Hadoop is an Apache open source framework written in JAVA which allows distributed processing of large datasets across clusters of computers using simple programming models.

Hadoop is made up of several modules that are supported by a large ecosystem of technologies

Hadoop Main Components

- MapReduce, a distributed data processing model and execution environment that runs on large clusters of commodity machines
- Hadoop Distributed File System (HDFS), a distributed file system that runs on large clusters of commodity machines
- Hive, a distributed data warehouse
- Apache Mahout, for Machine Learning

Hadoop Ecosystem

Data Flair

oozie
(Work flow)

HCatalog
Table & schema Management

Pig (Scripting)

Hive (Sql Query)

mahout
(Machine Learning)

Drill (Interactive Analysis)

AVRO (JSON)

Thrift (Cross Language Service)

APACHE HBASE
HBASE (Columnar Store)

Sqoop
(Data Collection)

Flume
(Data Collection)

Zookeeper (Coordination)

Ambari

Apache Ambari (Management & Monitoring)

Mapreduce
(Data Processing)

Yarn
(Cluster Resource Management)

HDFS
(Hadoop Distributed File system)

Apache Spark

- Apache Spark is an open source framework for Big Data under Apache foundations
- Developed in the University of California, Berkeley's AMPLab by Matei Zaharia in 2009

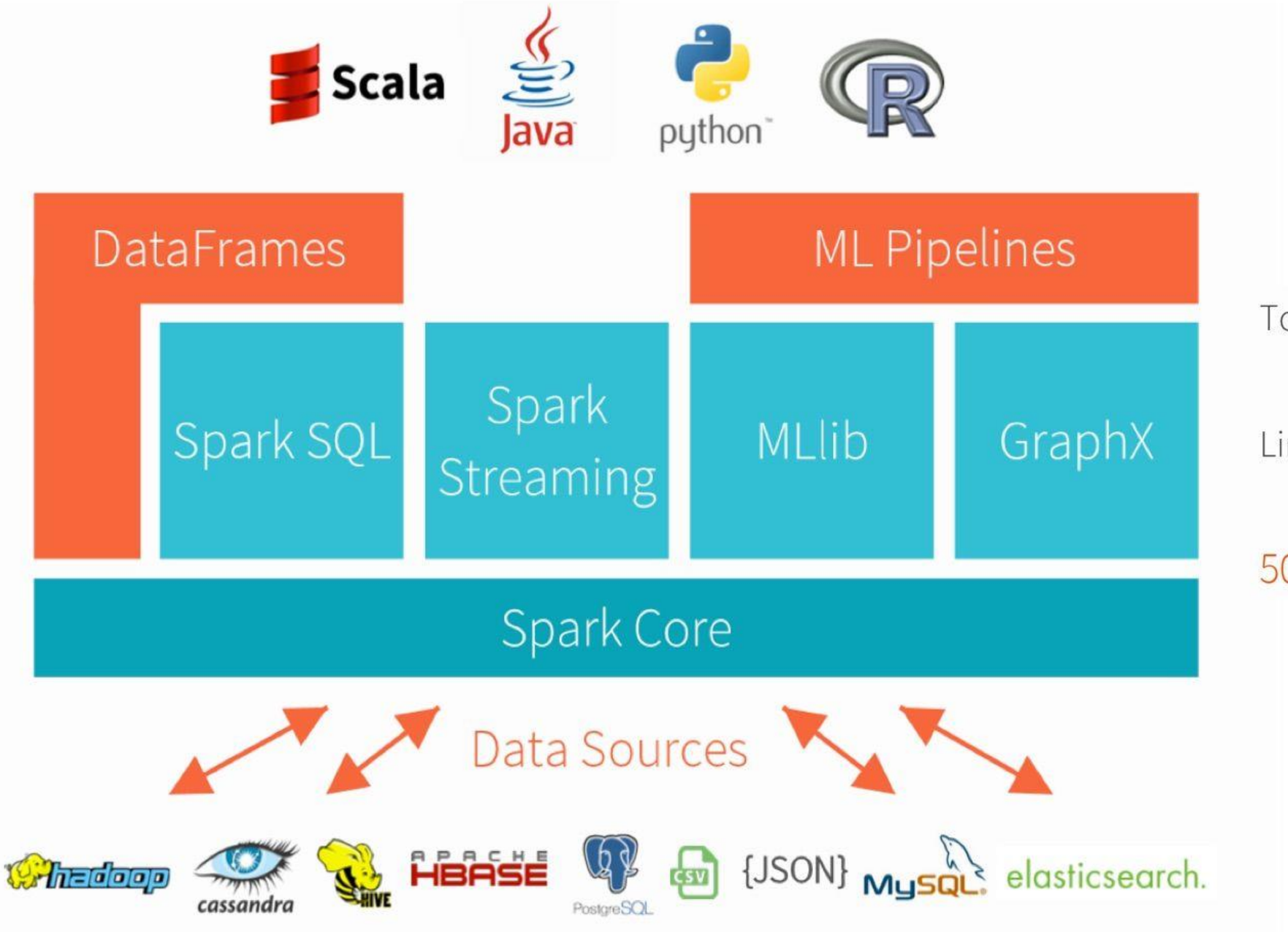
- Features like abstraction, ease of use, in-memory processing, streaming and batch processing, and Graph processing

- Spark provided the unified APIs in different languages like R, Python and Java

- The Spark ecosystem includes Spark Core, Spark SQL, Spark ML, Spark GraphX and Spark Streaming

Features of Apache Spark

- Ease of usage and well documented
- High-performance gains
- Advanced analytics
- Real-time data streaming
- Easy scaling and deployment



Total contributors: 150 → 500

Lines of code: 190K → 370K

500+ active production deployments

WWW.CIIIA.XYZ

BIG DATA & AI LANDSCAPE 2018

CIIIA®

V1 – Last updated 6/19/2018

 mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

WWW.CIIIA.XYZ

THANK YOU

#SHORTCUTTOTHEFUTURE

CII.IA®