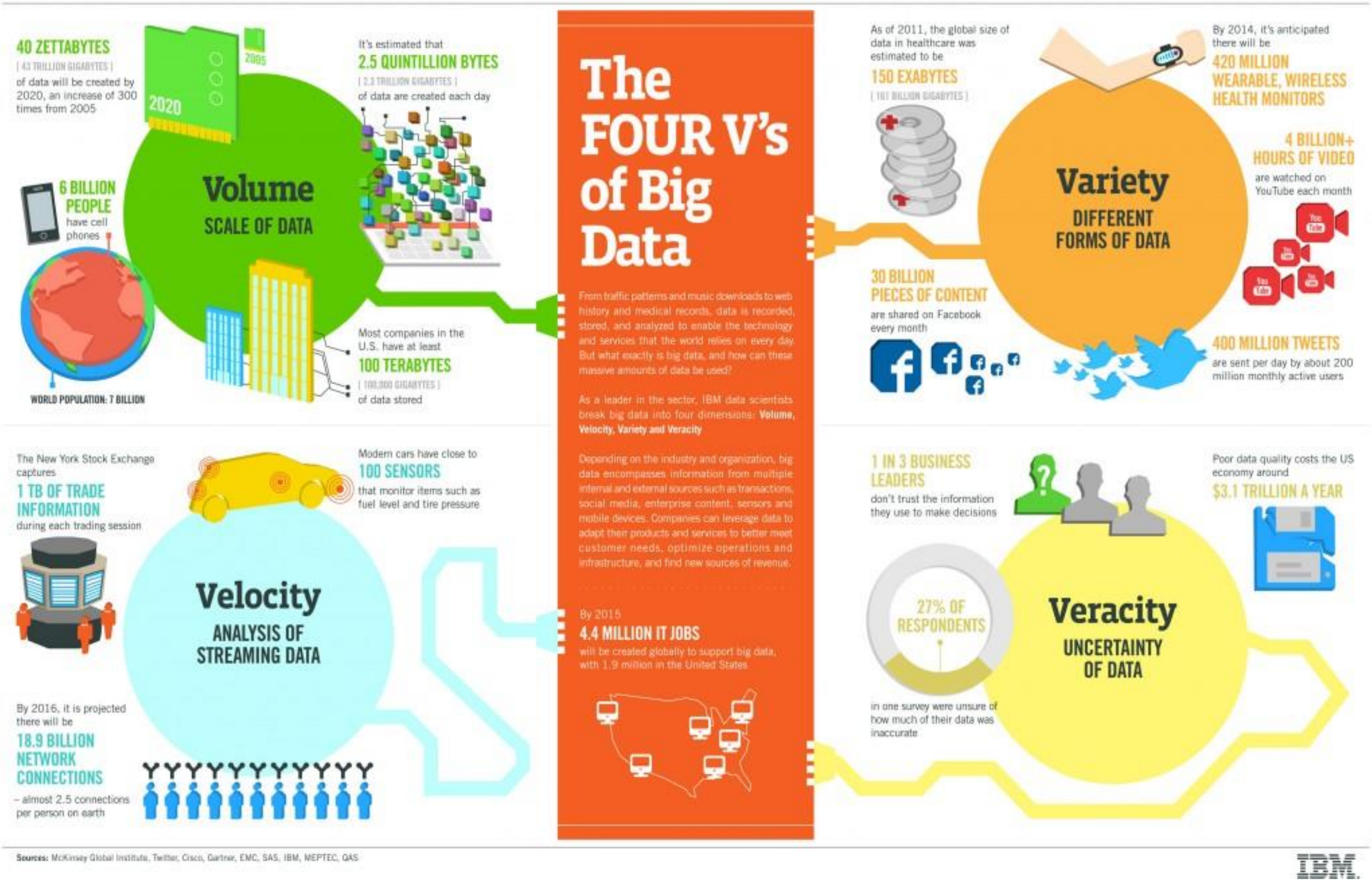
Module 1: Big Data

BeVera

“Big data is a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis.” (Lisa Arthur, CMO Network, 8/15/2013). “Big Data” is data whose scale, diversity, and complexity require new architecture, new tools, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity
ANALYSIS OF STREAMING 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

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

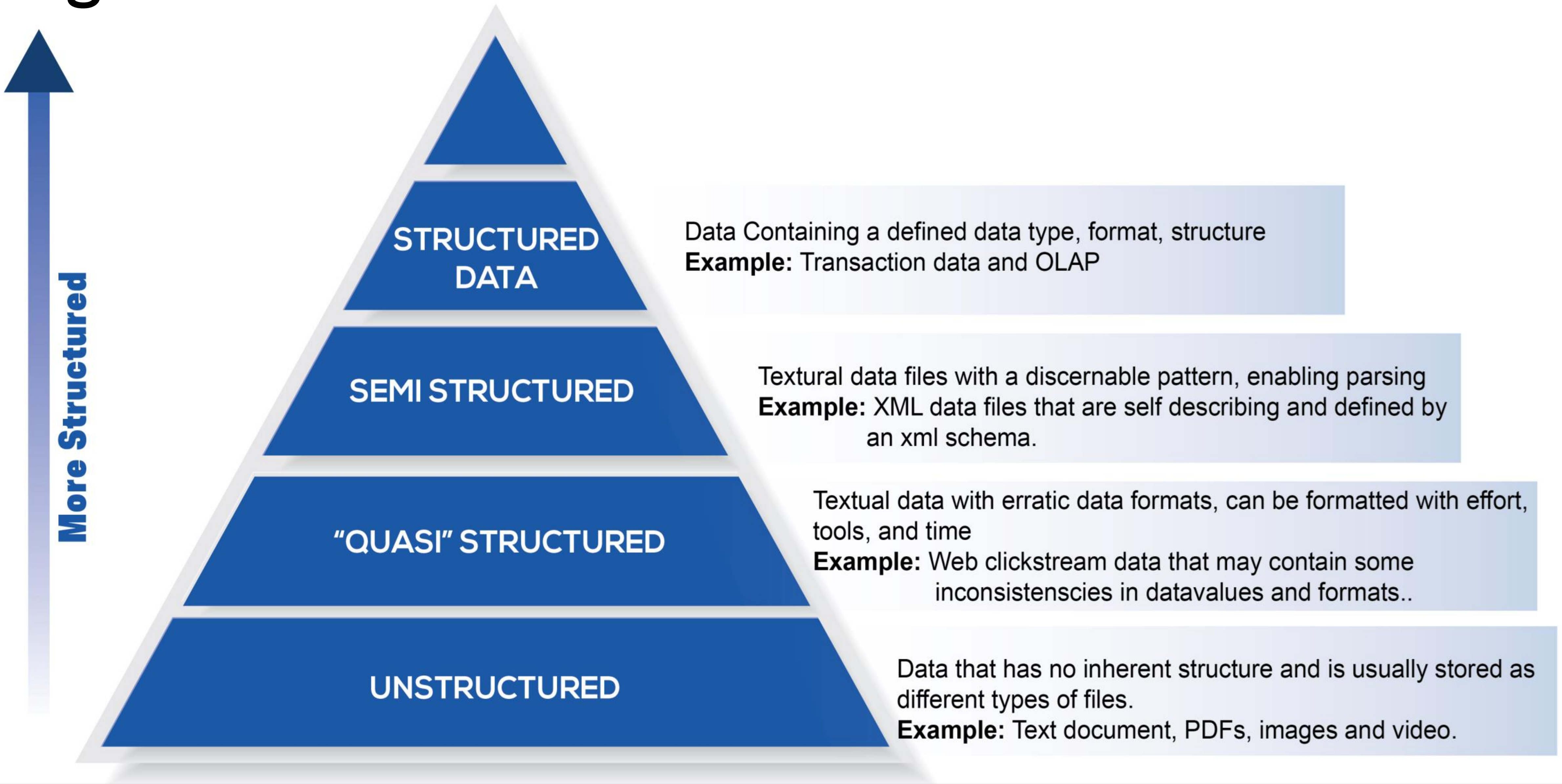
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM.

Big Data Characteristics: Data Structure



More Structured

STRUCTURED DATA

Data Containing a defined data type, format, structure
Example: Transaction data and OLAP

SEMI STRUCTURED

Textural data files with a discernable pattern, enabling parsing
Example: XML data files that are self describing and defined by an xml schema.

"QUASI" STRUCTURED

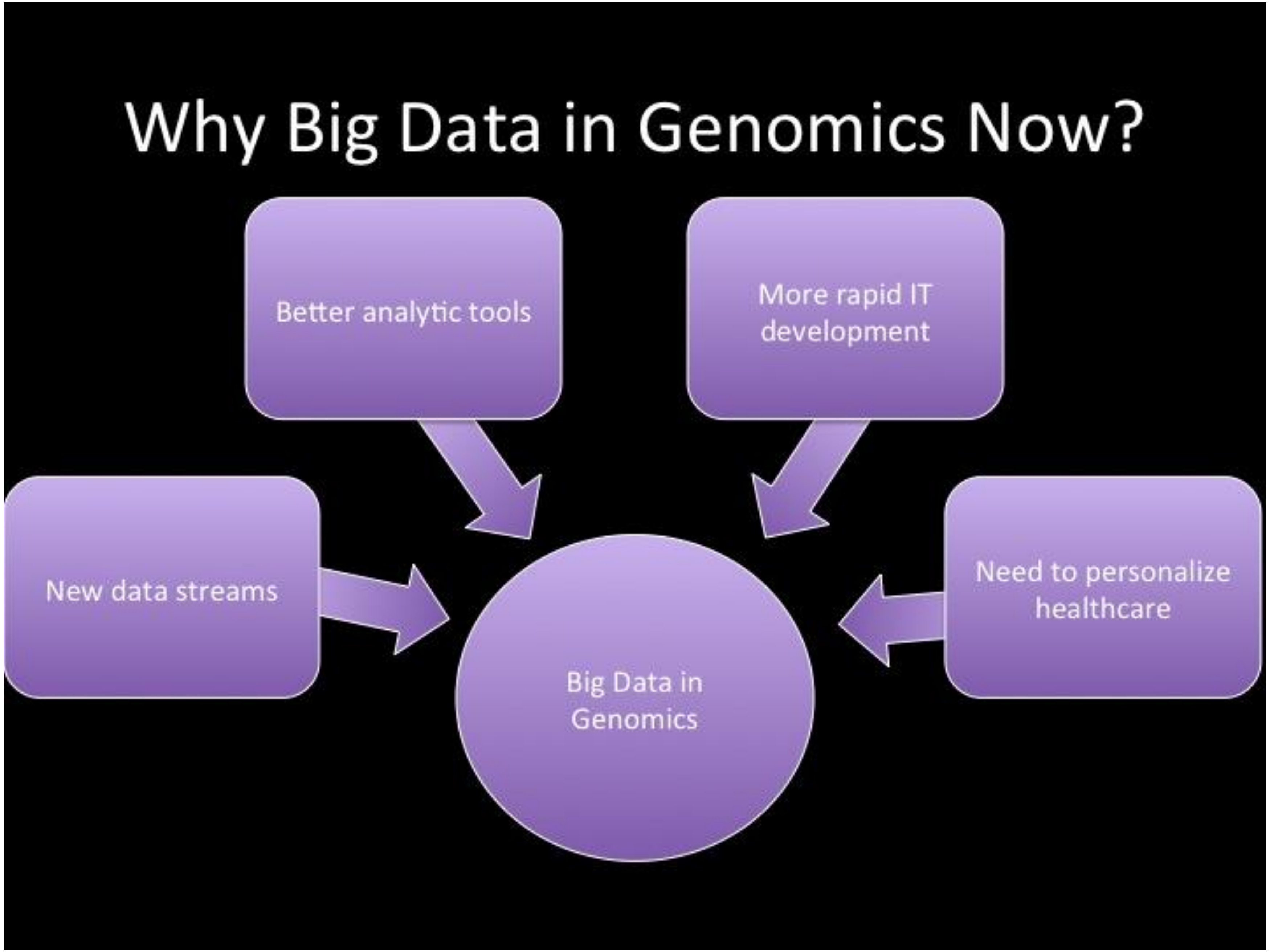
Textual data with erratic data formats, can be formatted with effort, tools, and time
Example: Web clickstream data that may contain some inconsistenscies in datavalues and formats..

UNSTRUCTURED

Data that has no inherent structure and is usually stored as different types of files.
Example: Text document, PDFs, images and video.

Traditional Data vs. Big Data

- Challenges
- Advantages



Why Big Data in Genomics Now?

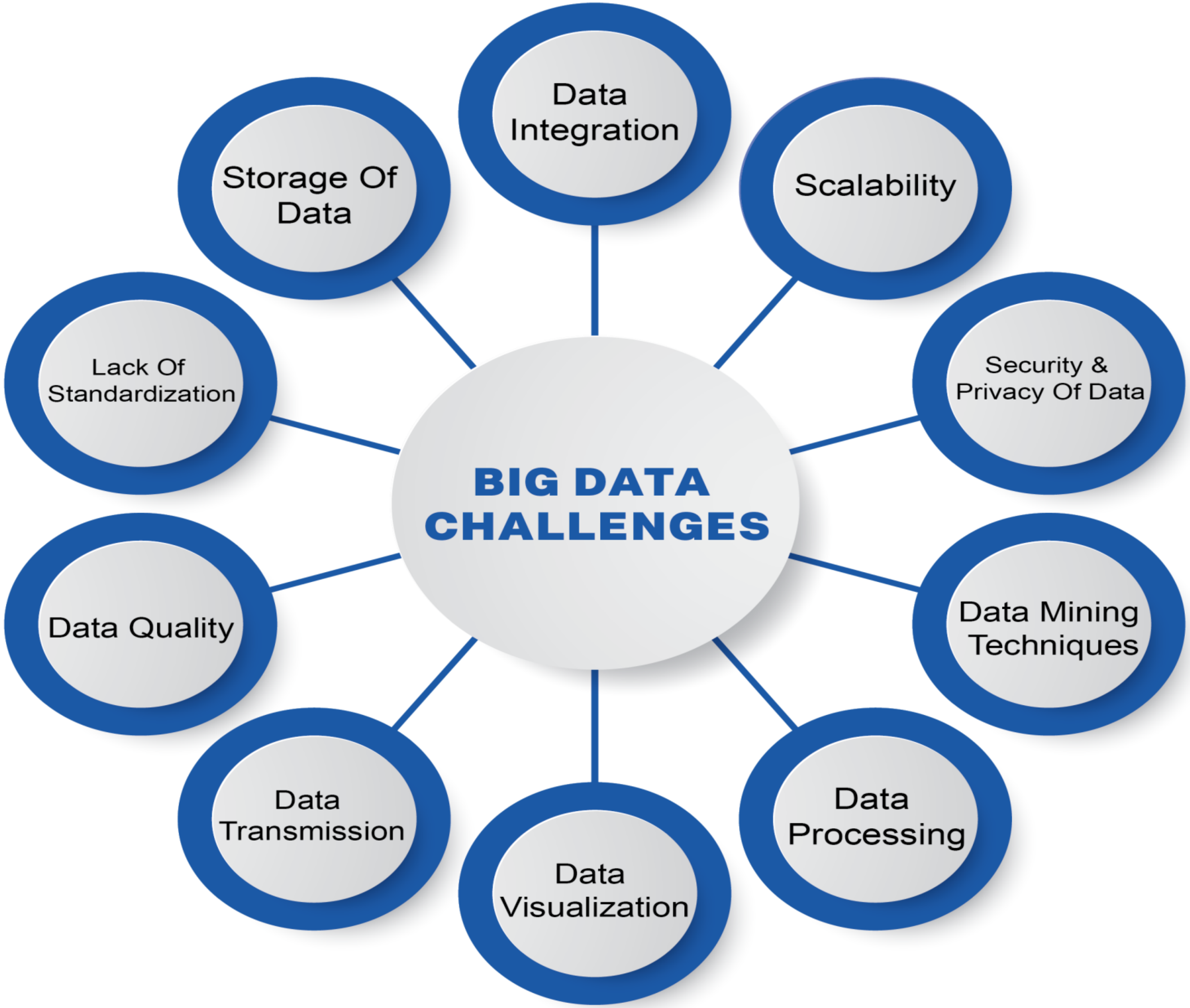
Better analytic tools

More rapid IT development

New data streams

Need to personalize healthcare

Big Data in Genomics



BIG DATA CHALLENGES

Data Integration

Scalability

Security & Privacy Of Data

Data Mining Techniques

Data Processing

Data Visualization

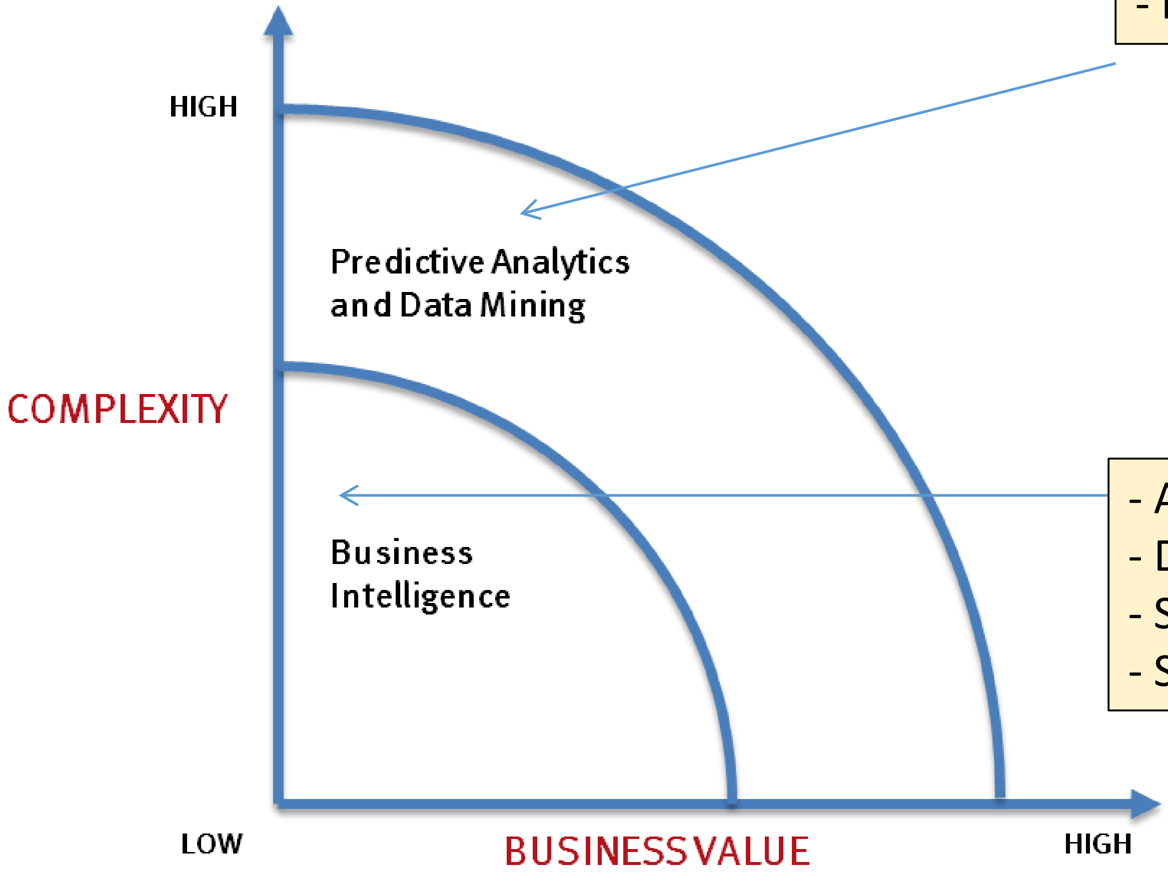
Data Transmission

Data Quality

Lack Of Standardization

Storage Of Data

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

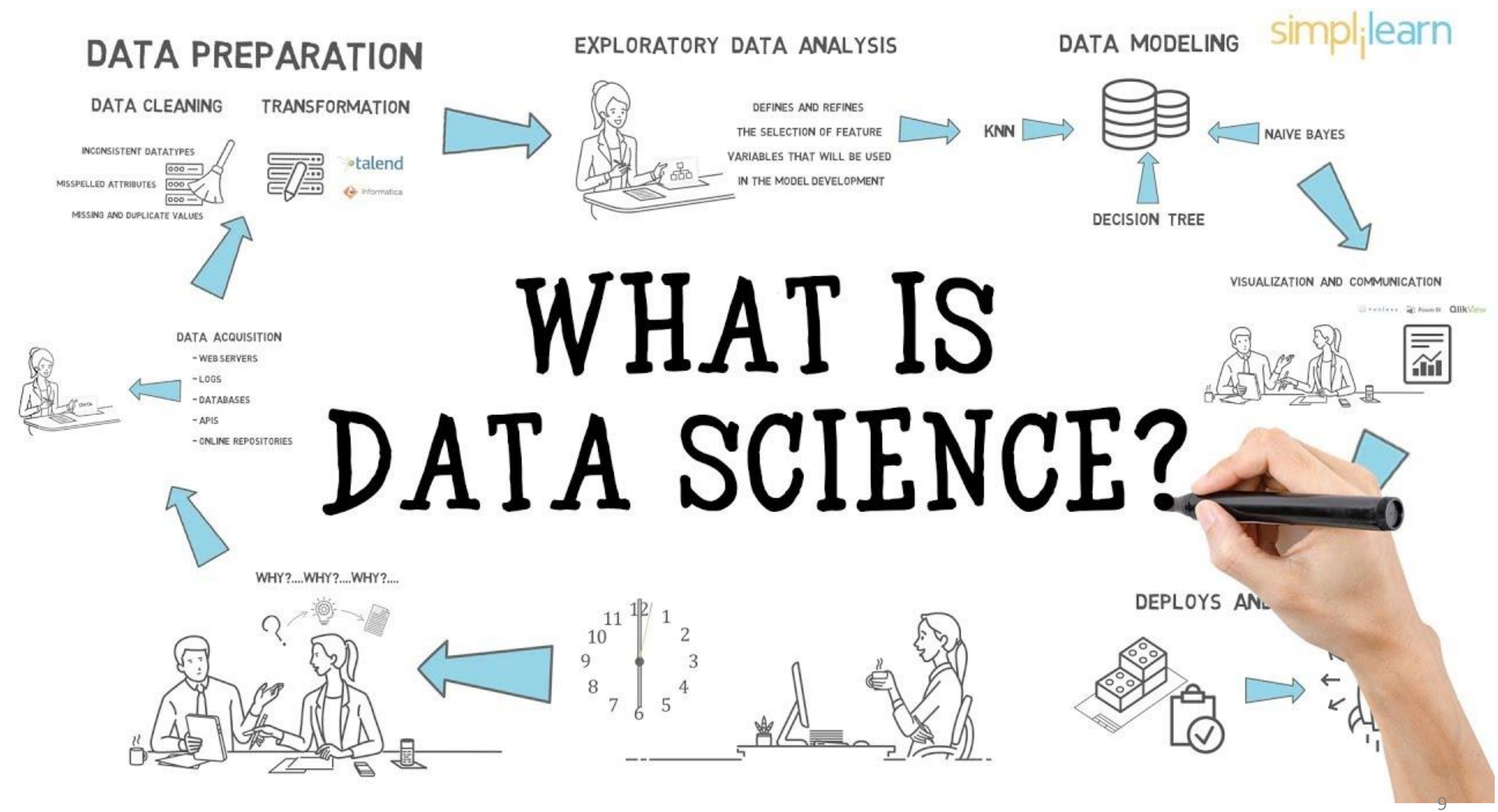
Common questions

- What if…
- What’s the optimal scenario for our business?
- What will happen next? What if these trends continue? Why is this happening?

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

Common questions

- What happened last quarter?
- How many did we sell?
- Where is the problem? In which situation?


DATA PREPARATION
DATA CLEANING
TRANSFORMATION
INCONSISTENT DATATYPES
MISSPELLED ATTRIBUTES
MISSING AND DUPLICATE VALUES
talend
informatica
EXPLORATORY DATA ANALYSIS
DEFINES AND REFINES
THE SELECTION OF FEATURE
VARIABLES THAT WILL BE USED
IN THE MODEL DEVELOPMENT
DATA MODELING
simplilearn
KNN
NAIVE BAYES
DECISION TREE
VISUALIZATION AND COMMUNICATION
Power BI
QlikView
WHAT IS
DATA SCIENCE?
DATA ACQUISITION
- WEB SERVERS
- LOGS
- DATABASES
- APIS
- ONLINE REPOSITORIES
WHY?....WHY?....WHY?....
12
1
2
3
4
5
6
7
8
9
10
11
DEPLOYS AN