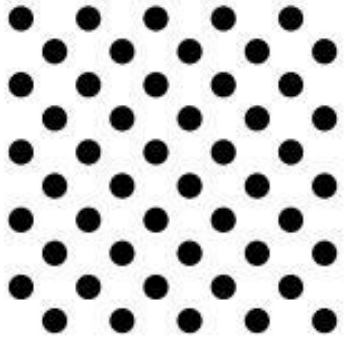
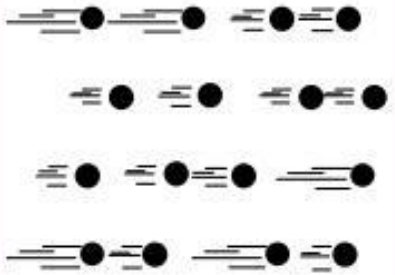
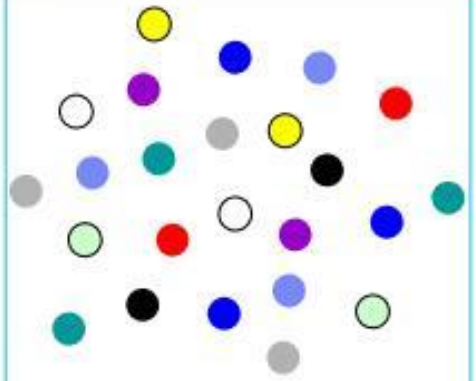
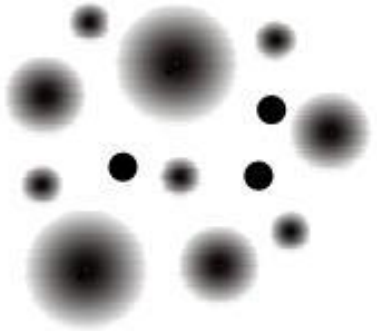
Module 1: Big Data

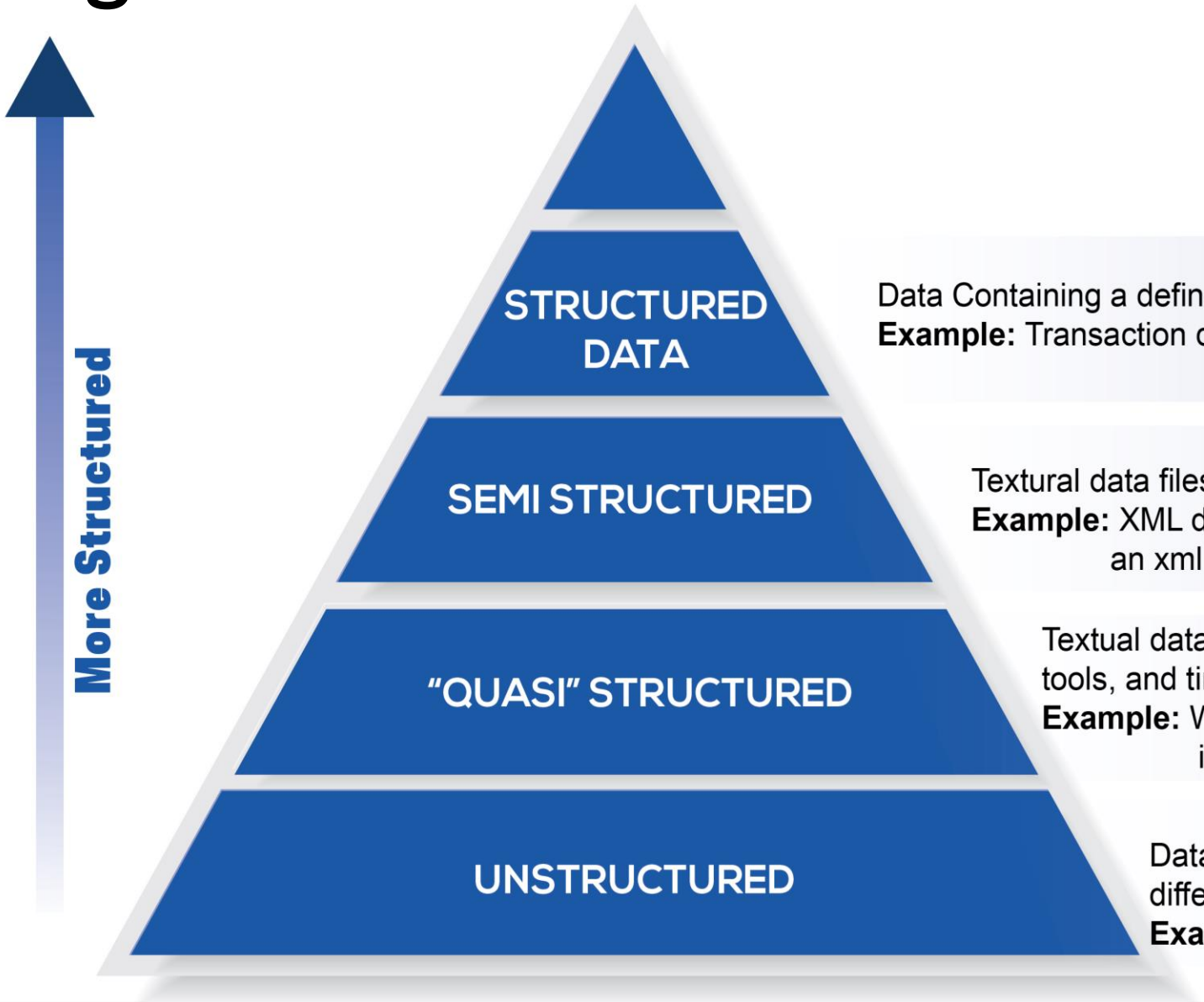
BeVera

“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…

Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data Characteristics: Data Structure



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

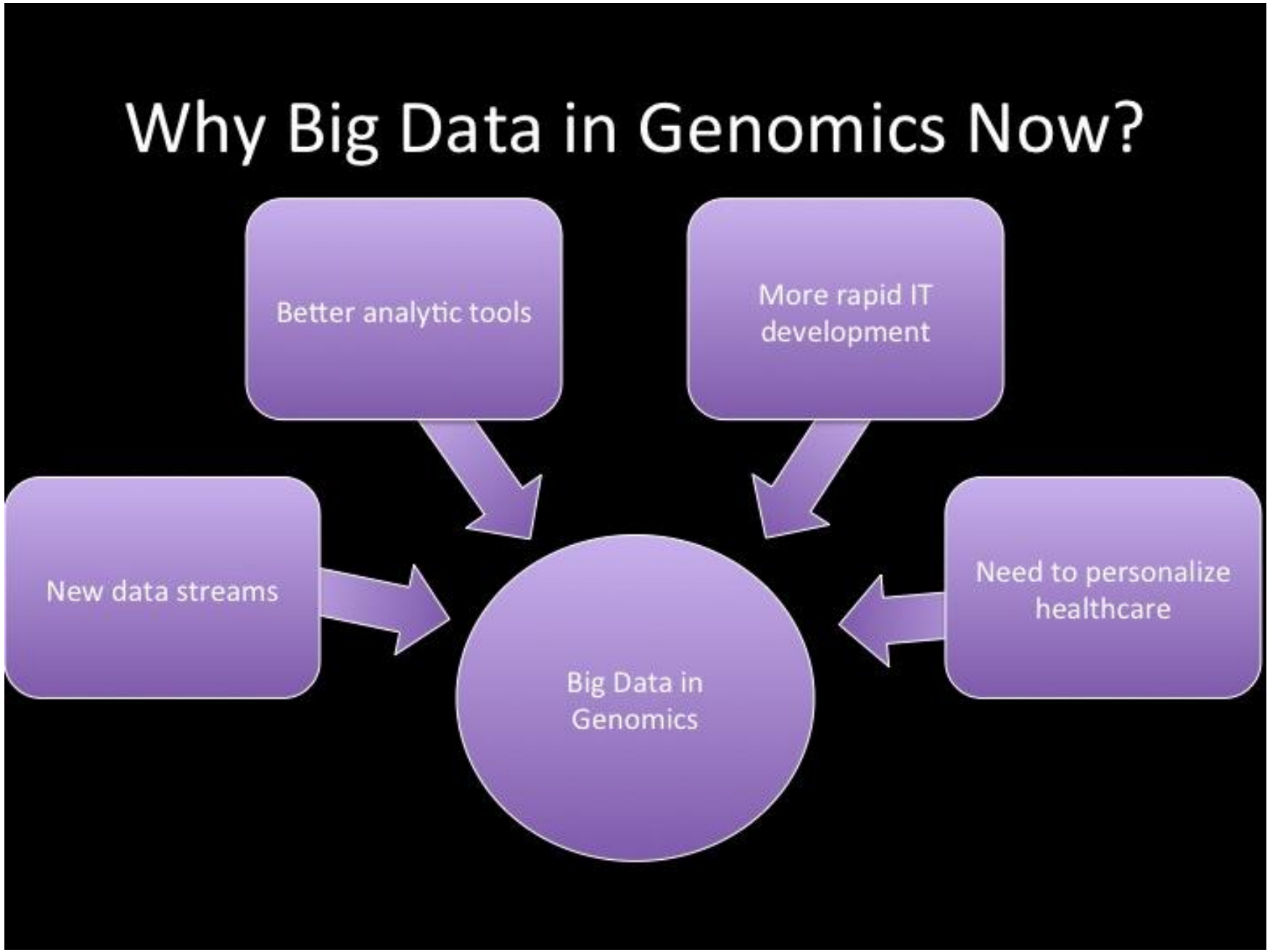
Textural data files with a discernable pattern, enabling parsing
Example: XML data files that are self describing and defined by an xml schema.

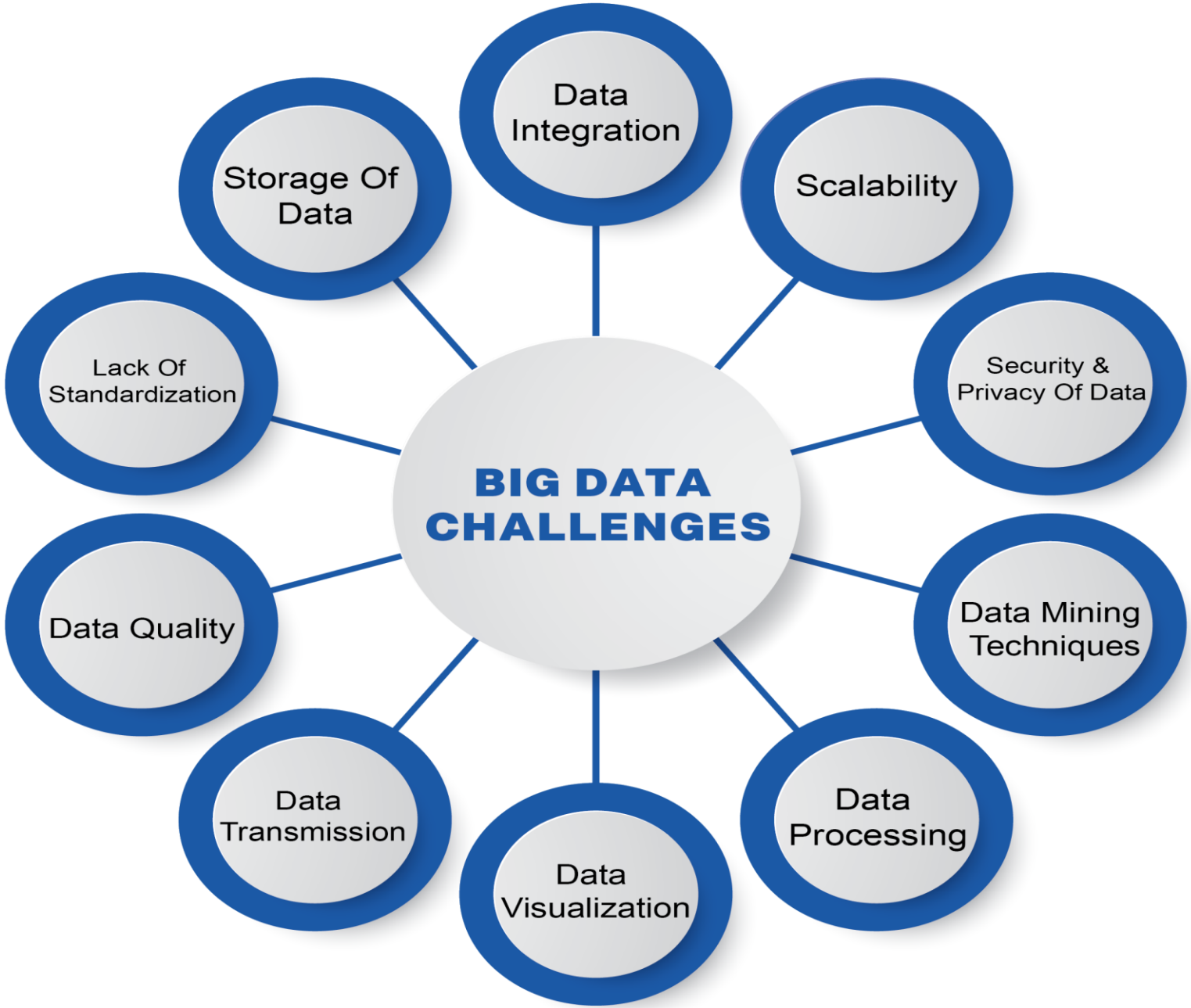
Textual data with erratic data formats, can be formatted with effort, tools, and time
Example: Web clickstream data that may contain some inconsistenscies in datavalues and formats..

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

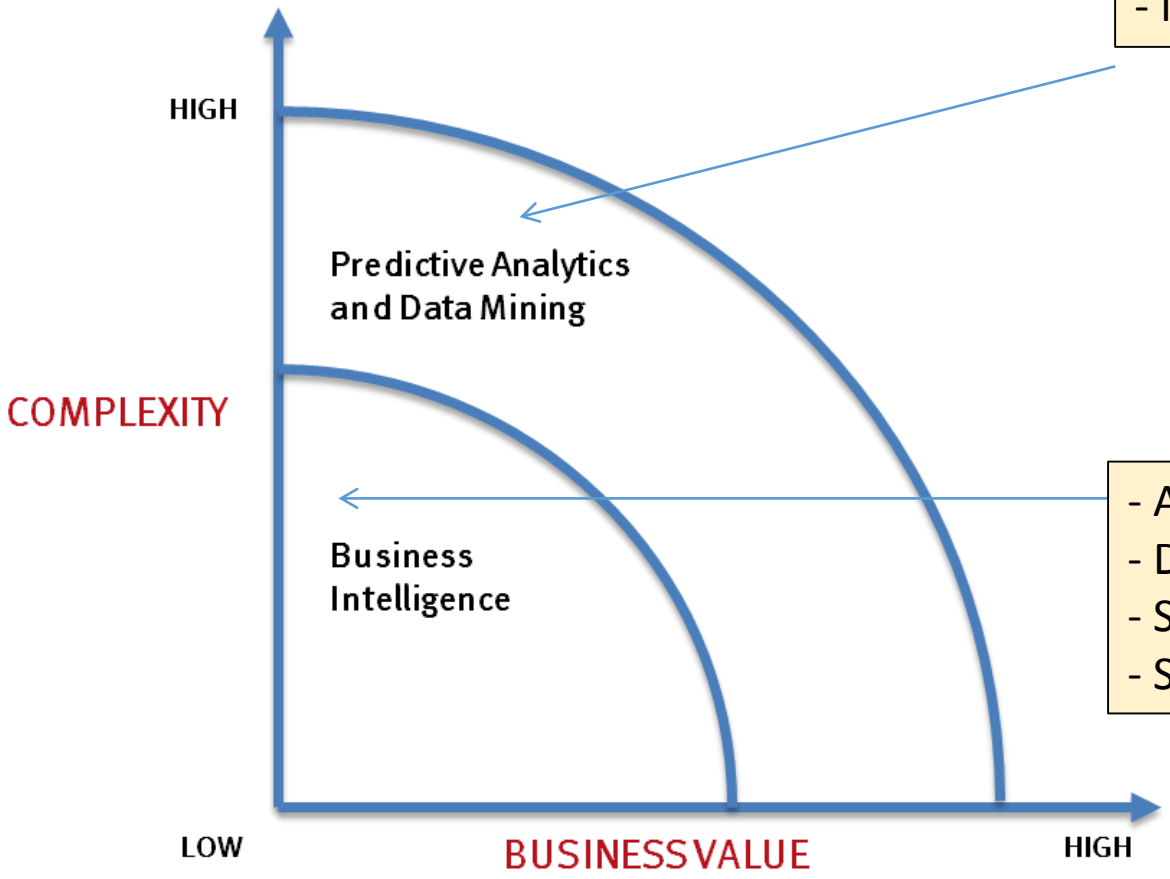

Why Big Data in Genomics Now?
Better analytic tools
More rapid IT development
New data streams
Need to personalize healthcare
Big Data in Genomics


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

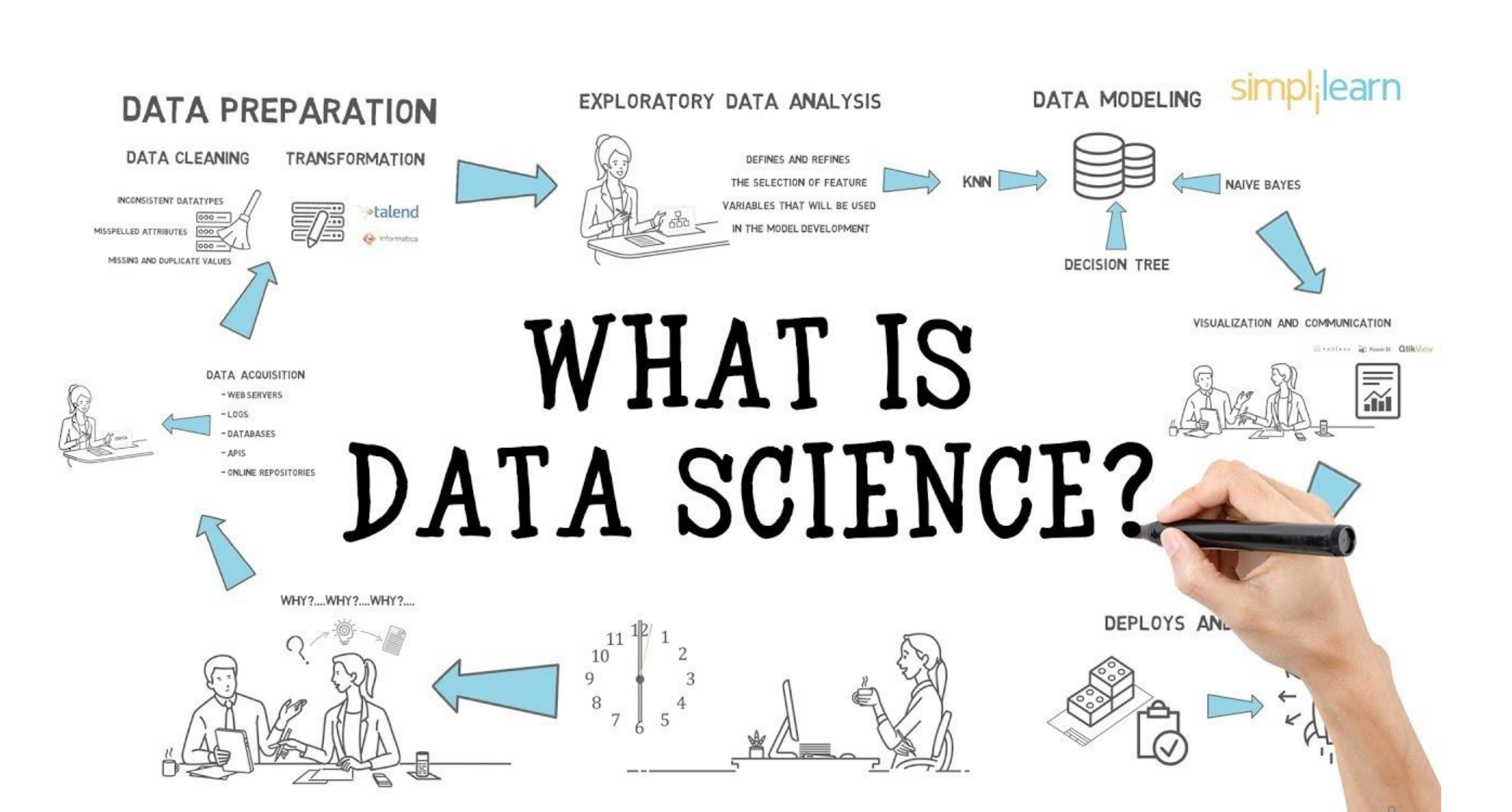
- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

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
DATA ACQUISITION
- WEB SERVERS
- LOGS
- DATABASES
- APIS
- ONLINE REPOSITORIES
WHY?....WHY?....WHY?....
DEPLOYS AN
WHAT IS
DATA SCIENCE?