IBM

Smarter education for smarter planet

IBM Innovation Centre for Education

Welcome to:

Big Data, Knowledge Representation in Taxonomies and Ontologies & Application of Advanced Analytics to Cognitive Computing

Unit objectives

After completing this unit, you should be able to:

- Understand the concepts of dealing with human-generated data and big data
- Learn about 4 V's of bigdata and bigdata architecture with Hadoop ecosystem
- Gain knowledge on types of data and data services
- Understand taxonomies and ontologies, advanced analytics leads to cognitive computing
- Gain an insight into key capabilities in advanced analytics
- Learn about the relationship between statistics
- Understand the concepts of predictive analytics, text analytics, business value of text analytics, contents image analytics, and speech analytics

Association between cognitive computing and big data

- Big data and cognitive computing are highly distinct in their purpose or mission. The mission of big data is best understood as the next generation of the traditional IT function of storage and organization of machine-based enterprise information now extended to include different types of data handled in new ways. This includes the tools that tell us what is in these collections.

- Cognitive computing, on the other hand, seeks the meaning in the data. Cognitive computing is best understood as an innovation in methodology for the field of analytics. Cognitive computing wants to break through the constraints of analytics based on backward facing numerical calculations and static presentations of results for human review.

Dealing with human-generated data

Figure: Source of Human generated data

Source: https://images.app.goo.gl/XtLq4keifb9zsyHa6

Volume, variety, velocity, and veracity

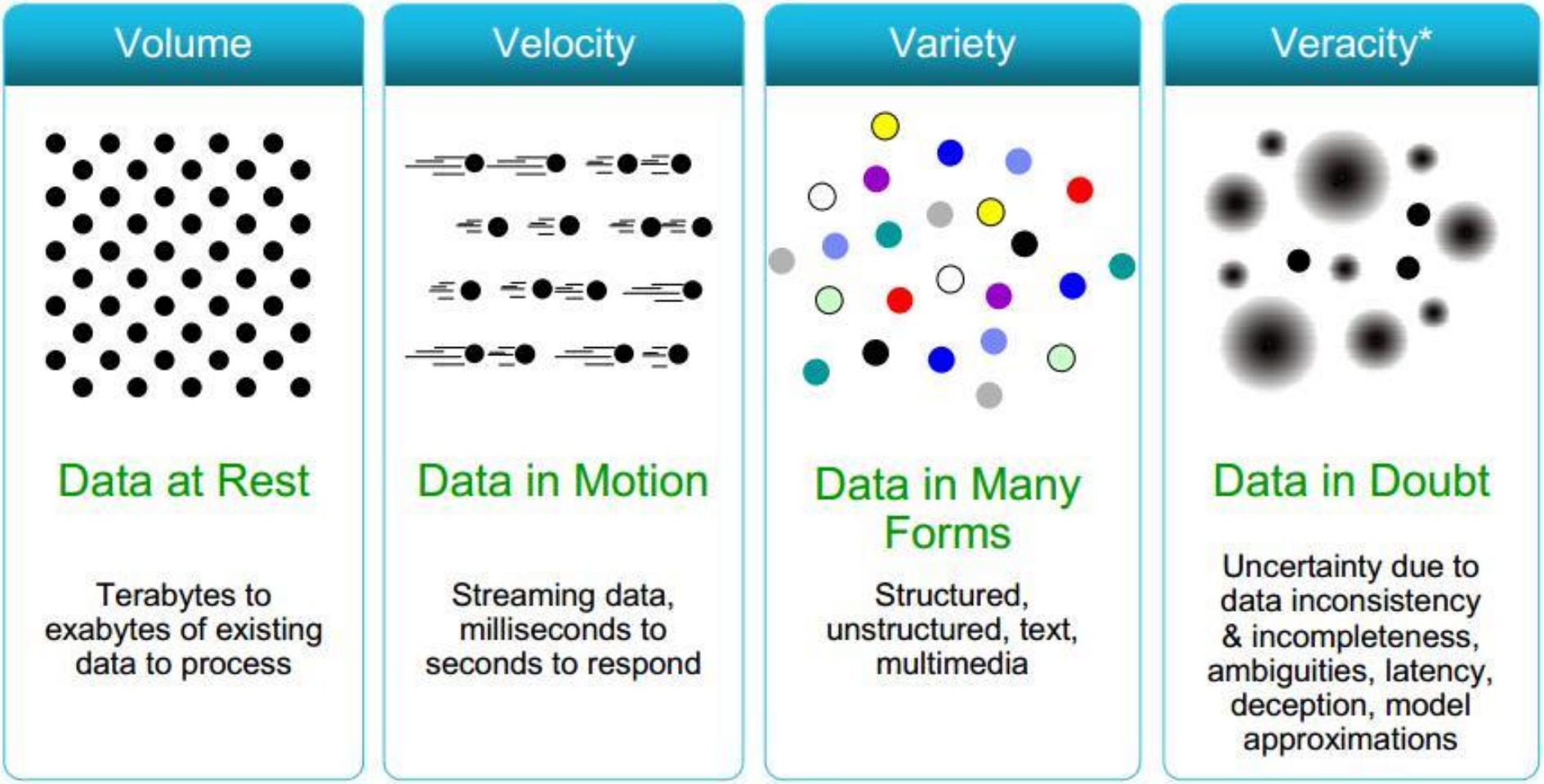


Figure: 4V's of BigData

Source: https://images.app.goo.gl/QDopqx5ahxPyMmNB6

Big data architecture

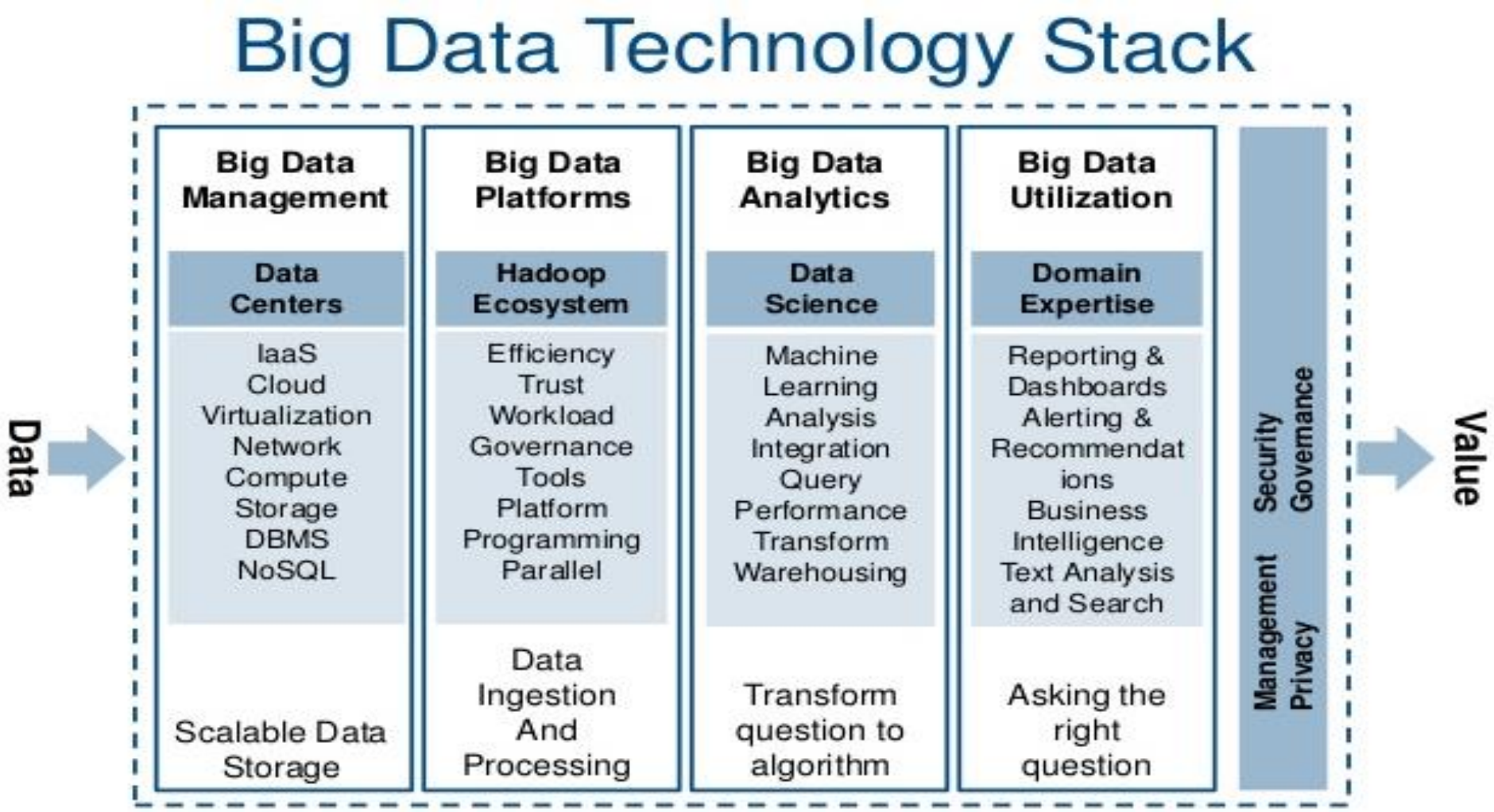


Figure: Big data technology stack

Source: https://images.app.goo.gl/zCAMMw9HmsDPfuaQ9

Structured and unstructured data functions

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Figure: Difference between Structure and unstructured Data functions

Source: https://images.app.goo.gl/dbQL9ZBDLkqQjFPS7

Data services and tools

- Data services and tools operations are:
  - To manage structured, non-structured data streams, a distributed file system was required. A global database is also a requirement for integrated data analysis from many sources.
  - Serialized systems are needed to enable both permanent data storage and remote procedure calls.
  - Coordination resources are important to develop a leveraged extremely scattered data framework.
  - Hadoop (one of the key techniques for big data entity) is used to collect, transform and load (ETL) for the loading and the transformation of structured and unstructured data.
  - Workflow resources methodology used to synchronize computing items over a broad data system.

Data warehouses analytical processing (OLTP/OLAP)

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Figure: OLTP and OLAP

Source: https://images.app.goo.gl/3RgfjxcbHZHNXJNj6

Hadoop (1 of 2)

- Open-source data storage and processing API
- Massively scalable, automatically parallelizable
  - Based on work from Google
    - GFS + MapReduce + BigTable
  - Current Distributions based on Open Source and Vendor Work
    - Apache Hadoop
    - Cloudera – CH4 w/ Impala
    - Hortonworks
    - MapR
    - AWS
    - Windows Azure HDInsight

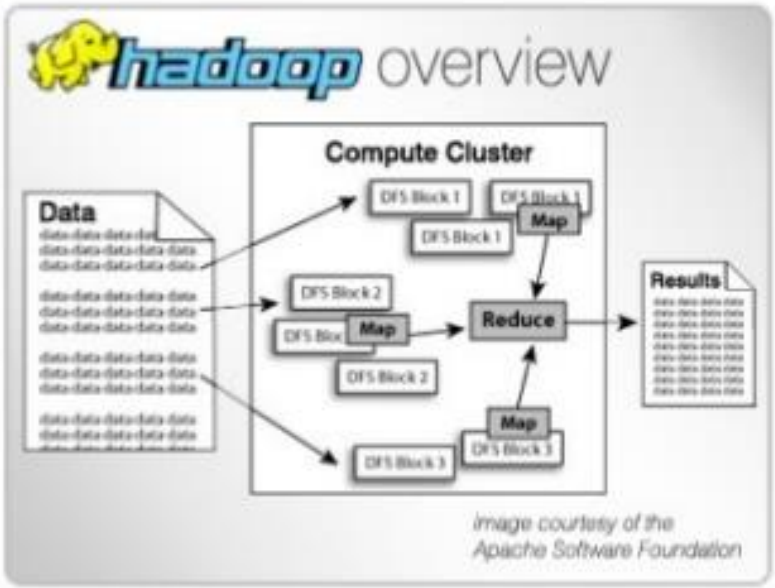


Figure: Hadoop Overview

Source: https://images.app.goo.gl/v4HqQw1jcoGY21T9A

Hadoop (2 of 4)

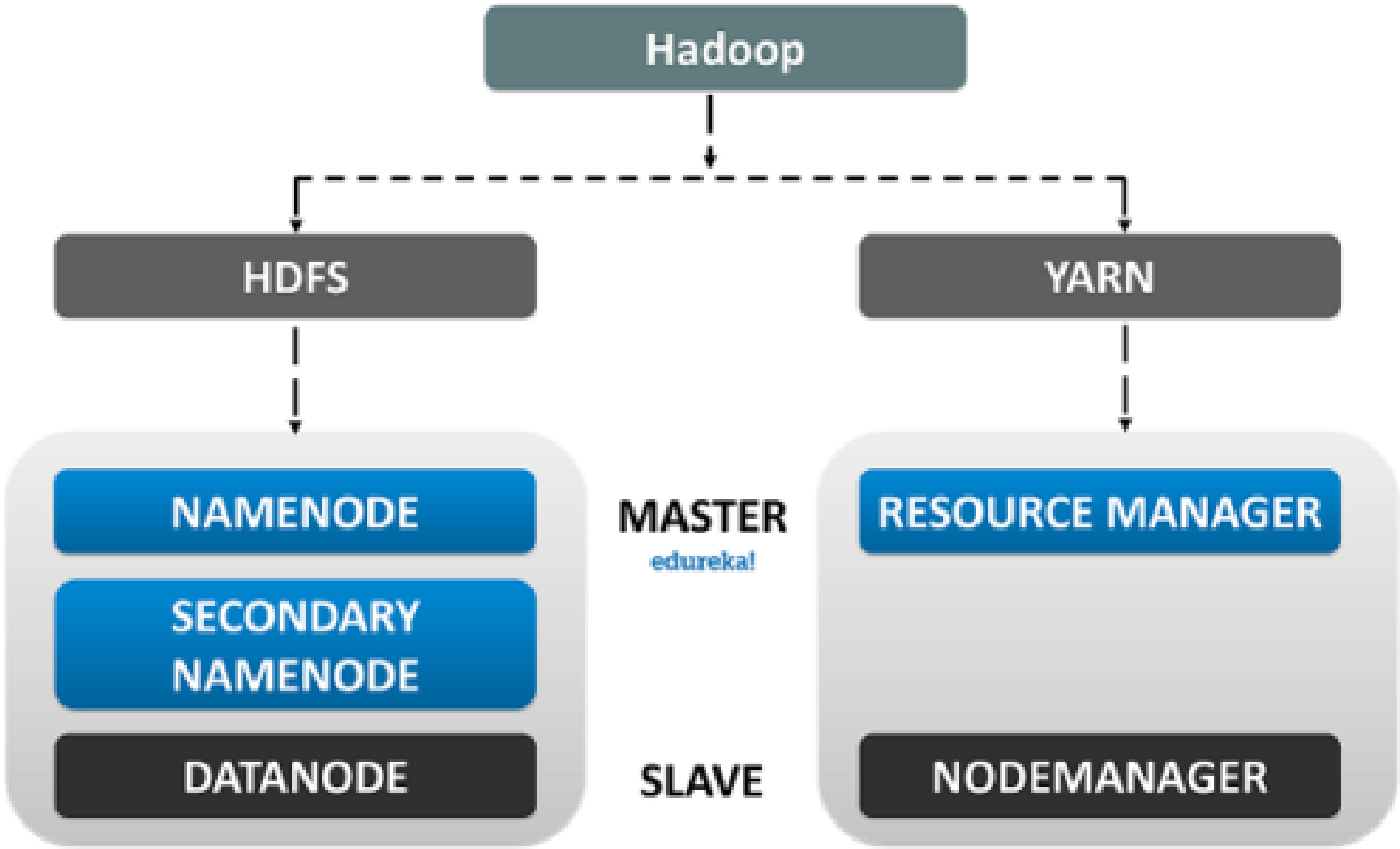


Figure: Hadoop components

Source: https://images.app.goo.gl/C5rMGyEjQ1BxoFAn9

Hadoop (3 of 4)

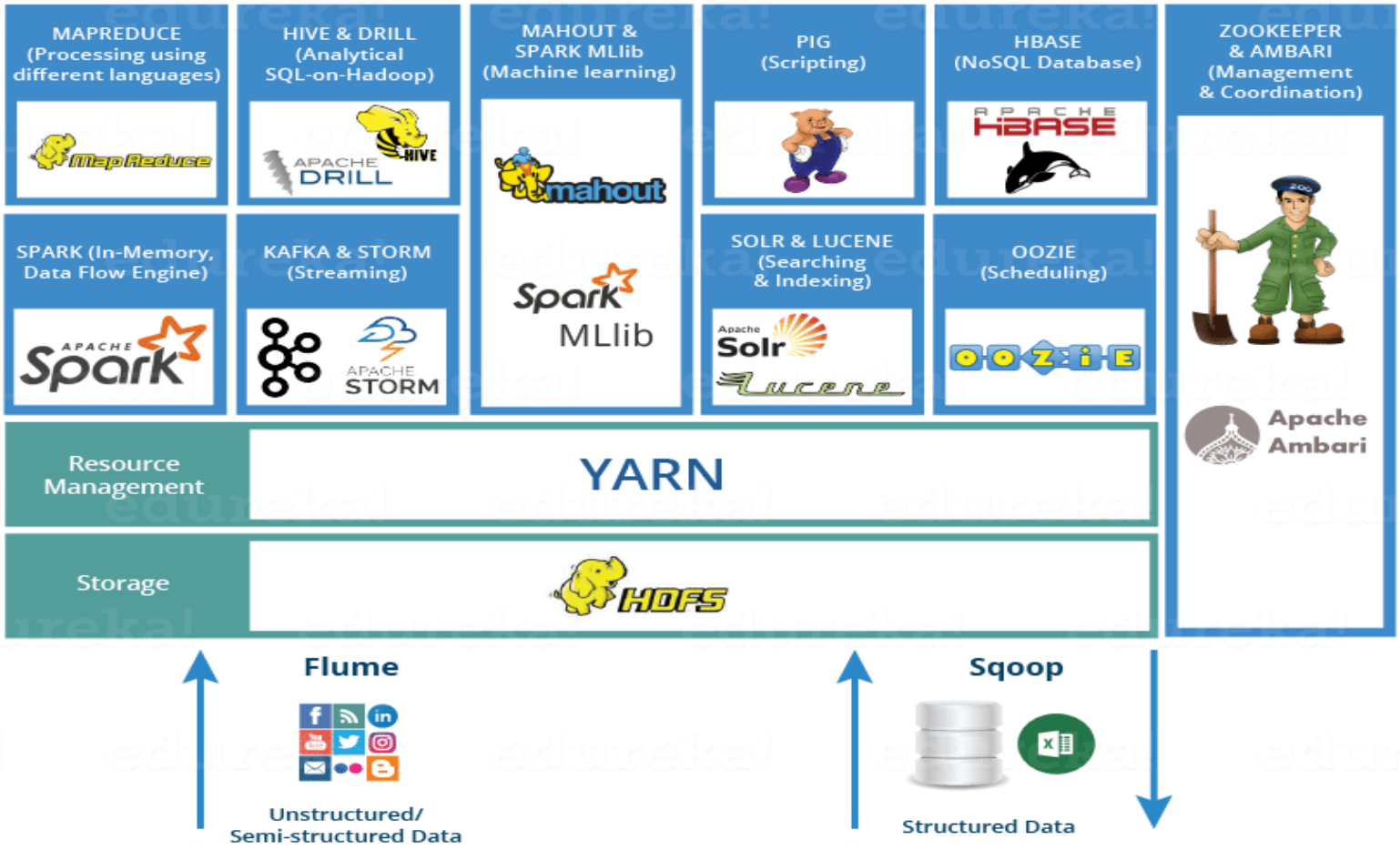


Figure: Hadoop ecosystem

Source: https://images.app.goo.gl/46i2iMrF3dAGr31Q6

Hadoop (2 of 2)

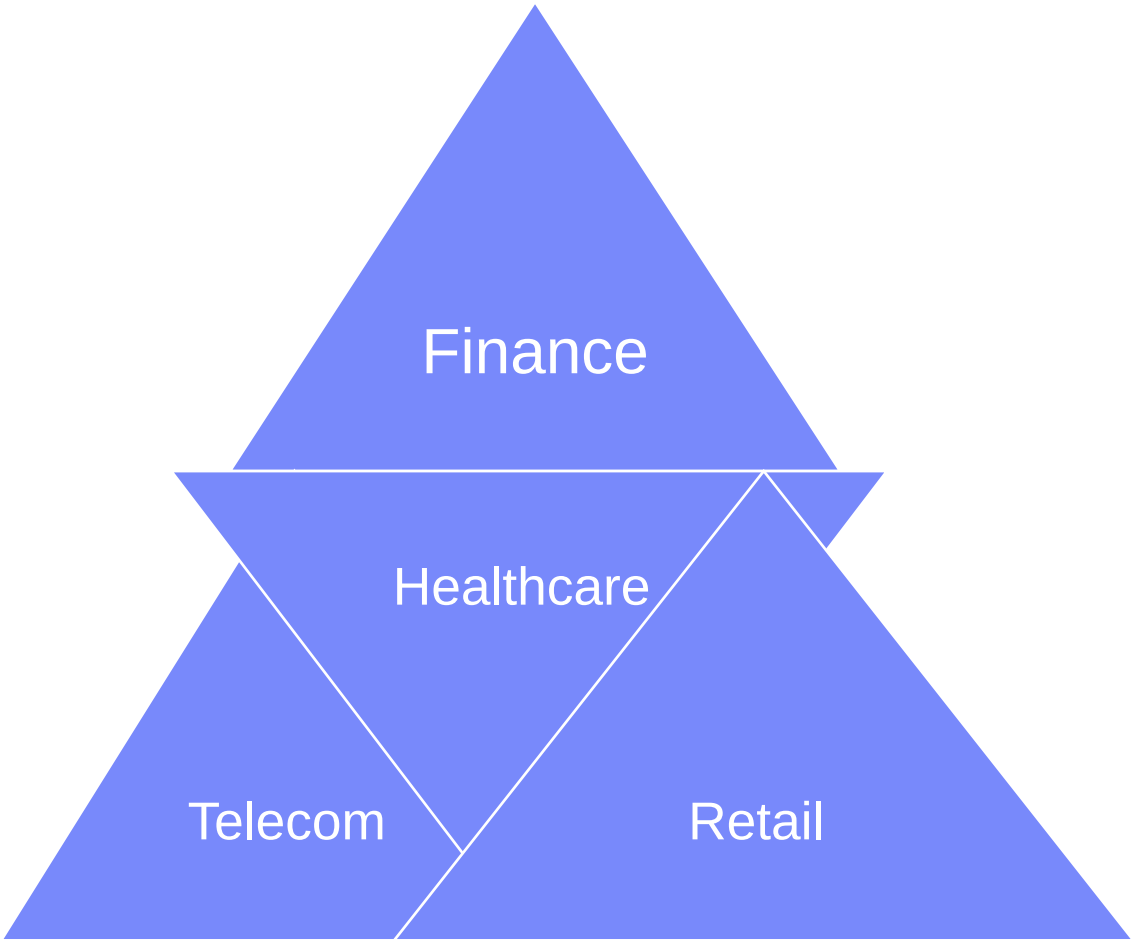


Figure: Use cases

Data in motion and streaming data

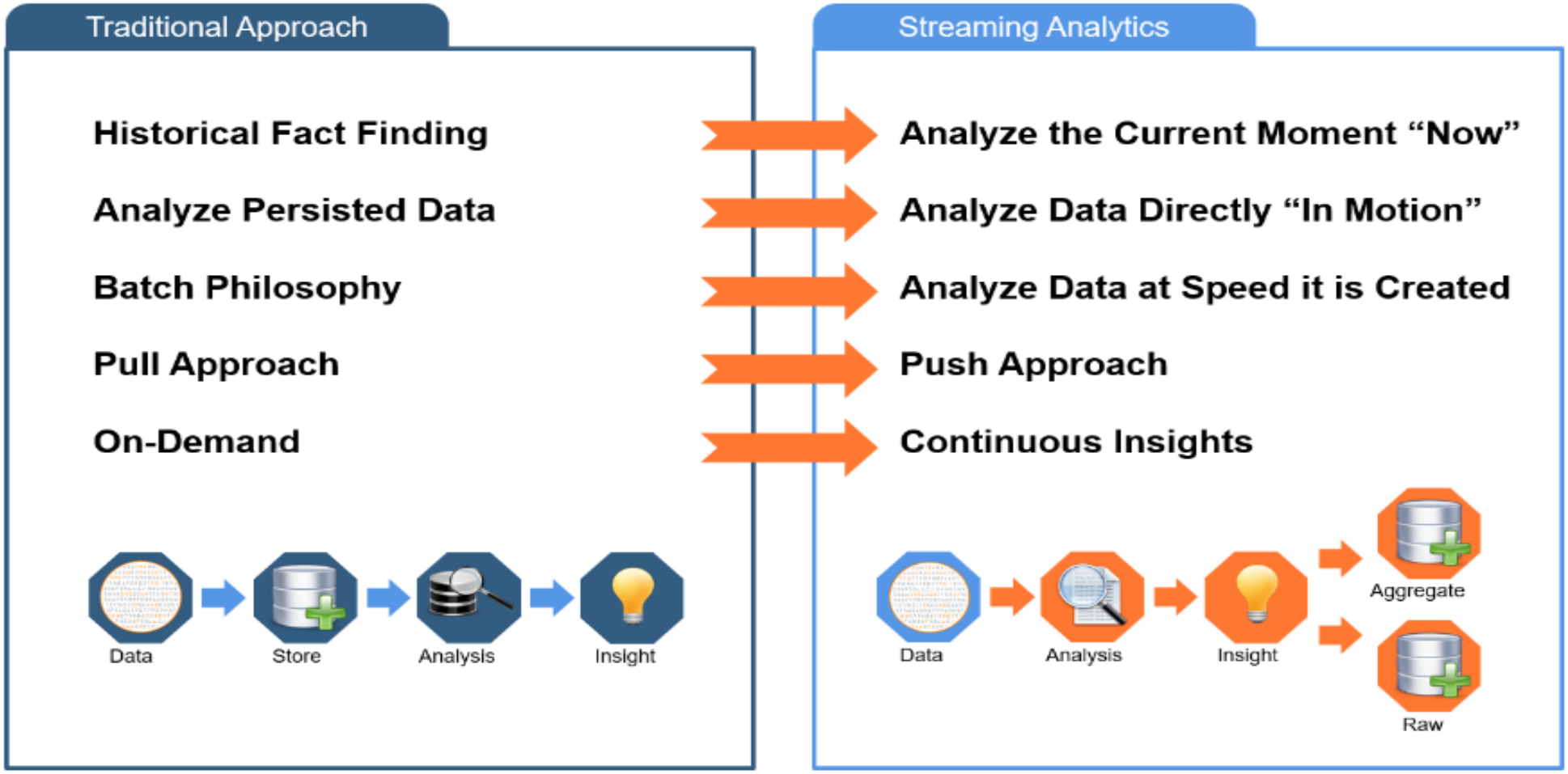


Figure: Data in motion and Streaming Data

Source: https://images.app.goo.gl/X6X6byfnfaWbs1zGA

Analyzing dark data

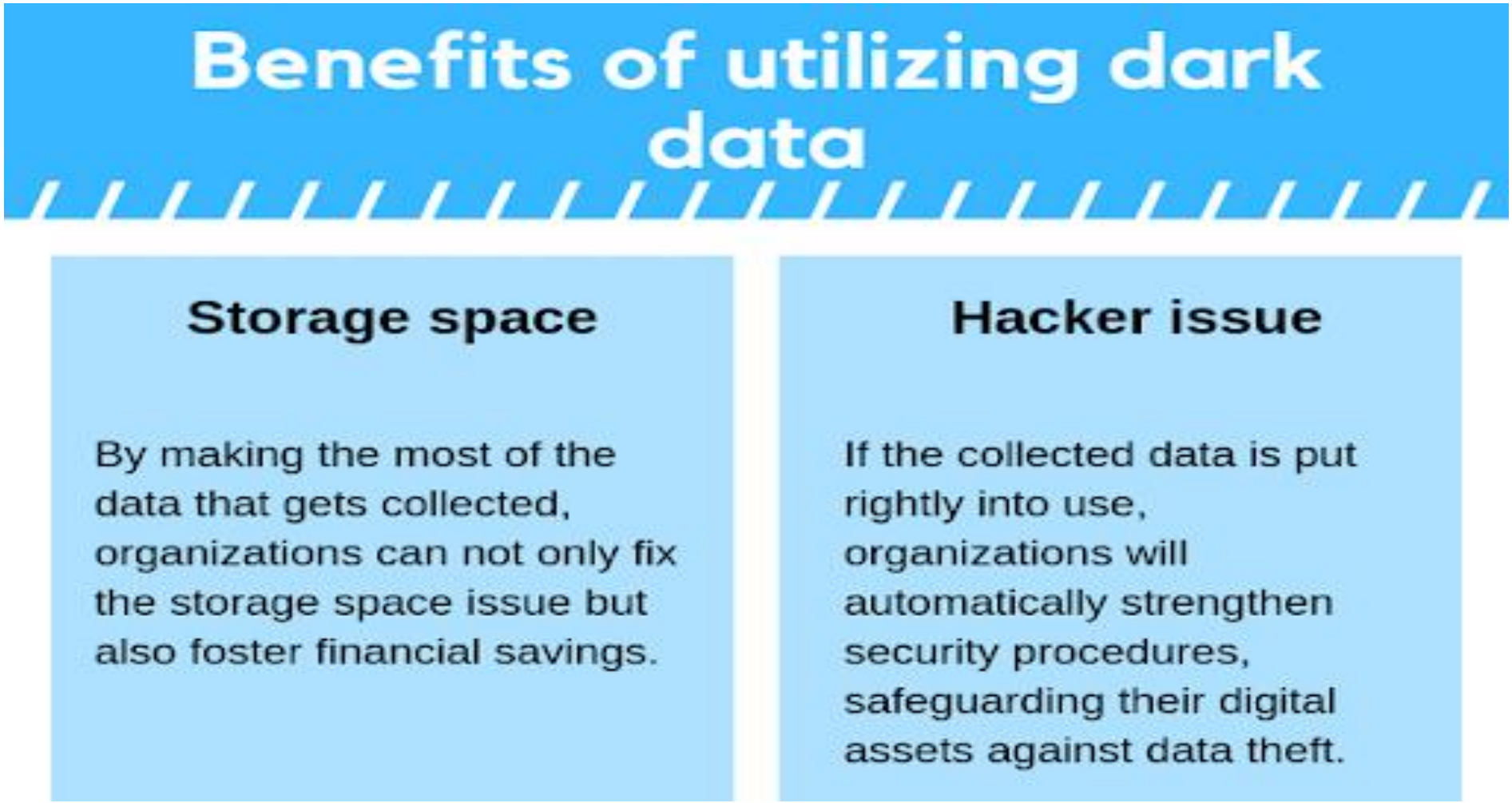


Figure: Dark Data Benefits

Source: https://images.app.goo.gl/RaaiHLucuW6teQ4t5

Big data integration into conventional data

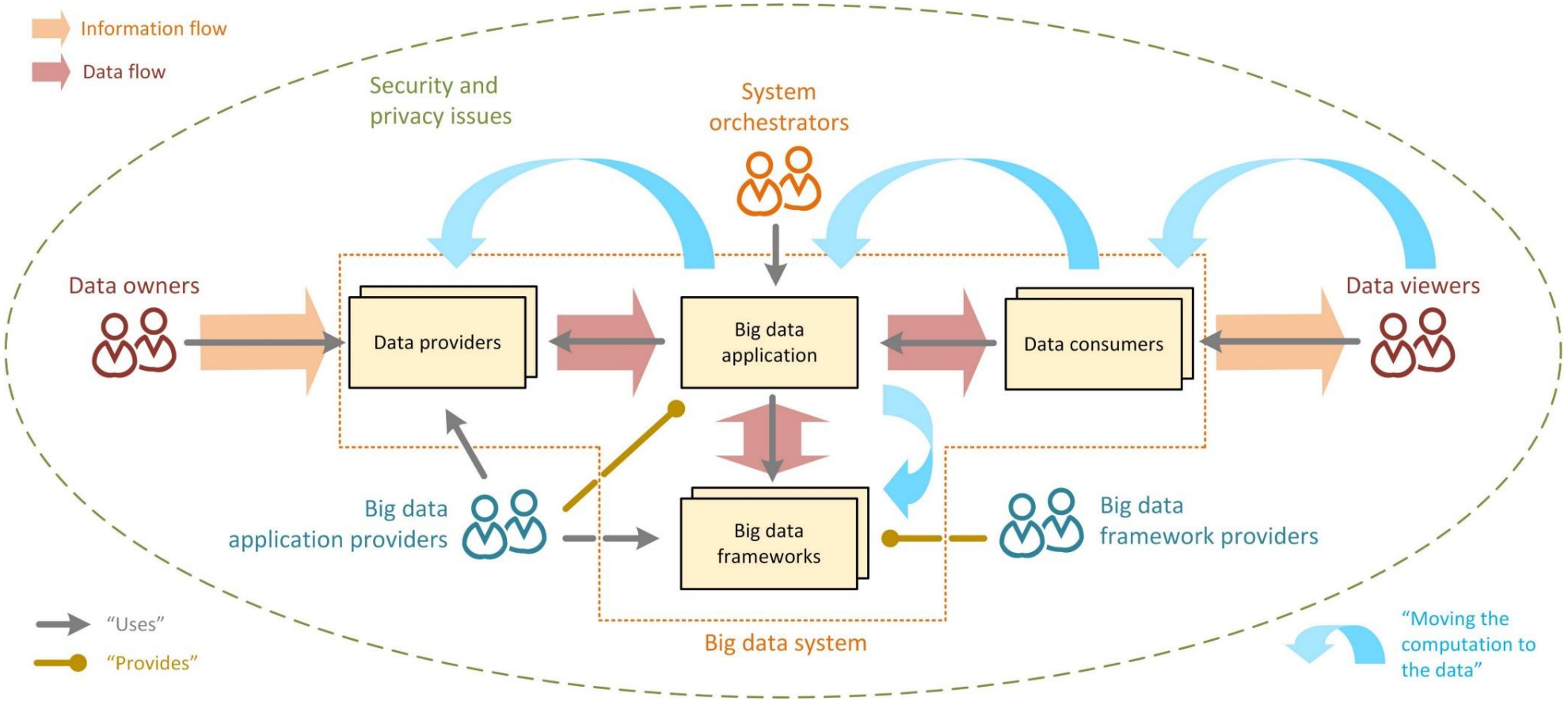


Figure: BigData Integration with existing dataset

Source: https://images.app.goo.gl/2Ug7F5ptP5FcHPCz7

Representing knowledge

- A machine sounds like an empty box unless it is encoded with some features or information. Therefore, to make it a valuable machine, it is required to put the necessary knowledge in it. So that it could understand it and is able to take the right decisions.

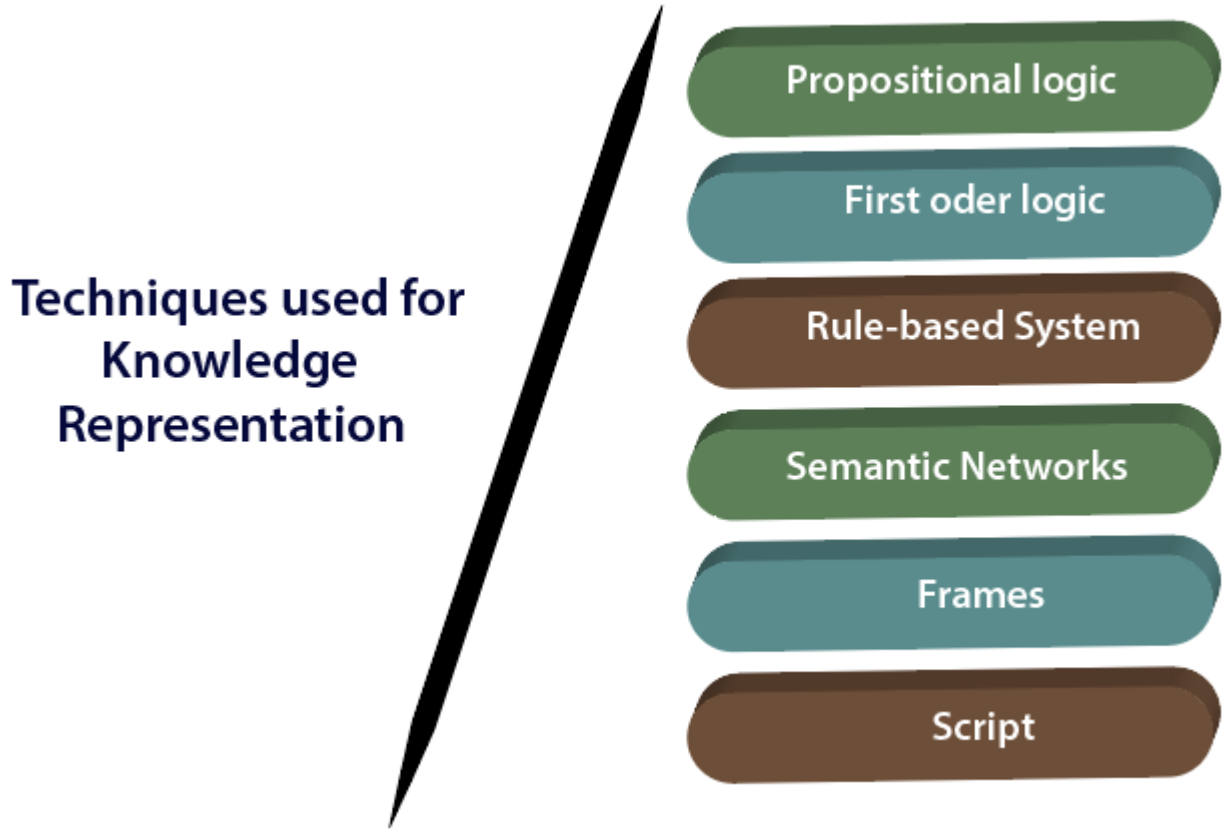


Figure: Knowledge Representation

Source: https://images.app.goo.gl/GMKLEoA2brJPDa986

Defining taxonomies and ontologies (1 of 2)

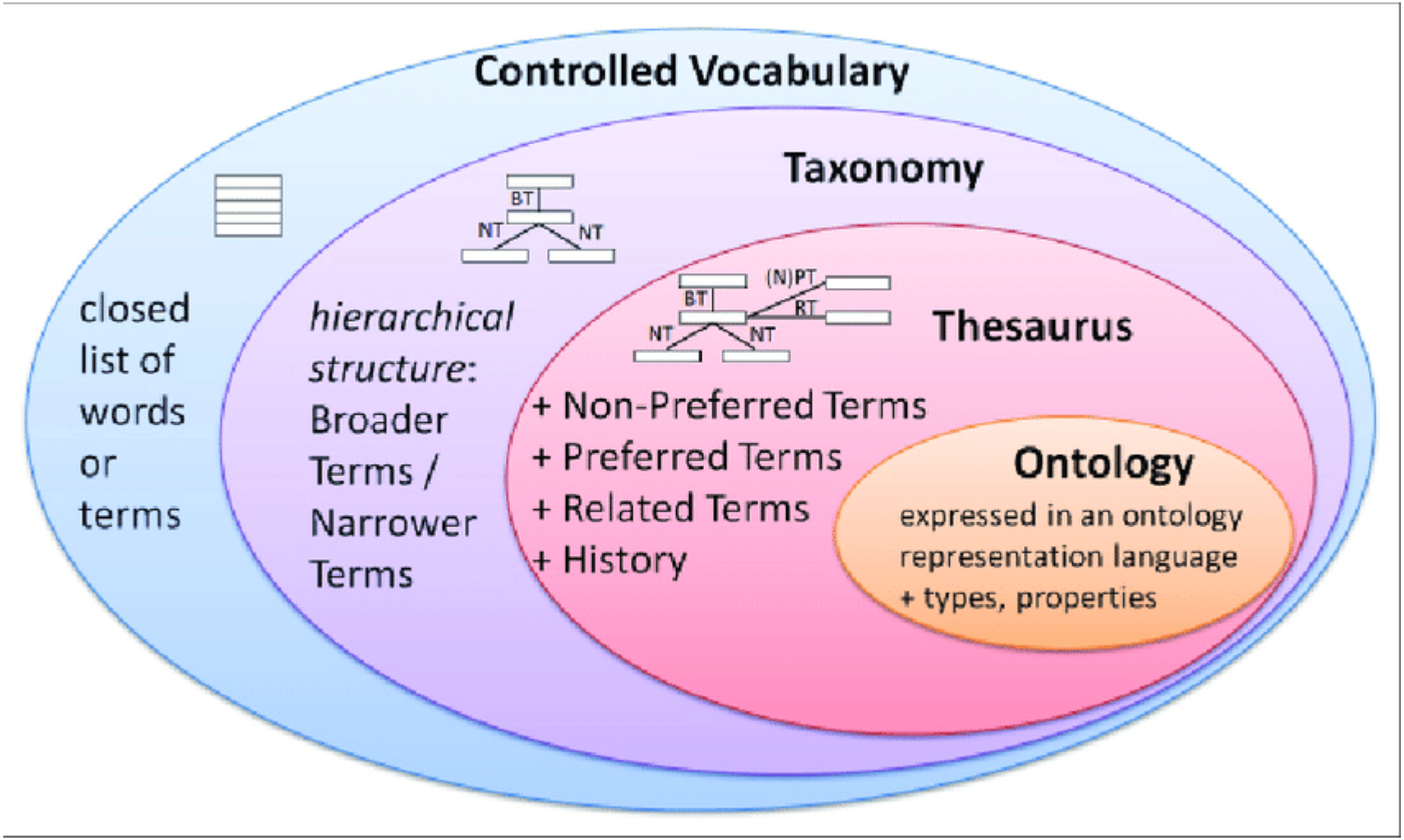


Figure: Taxonomies and Ontologies

Source: https://images.app.goo.gl/en7ArCHkmDgwrJEi8

Defining taxonomies and ontologies (2 of 5)

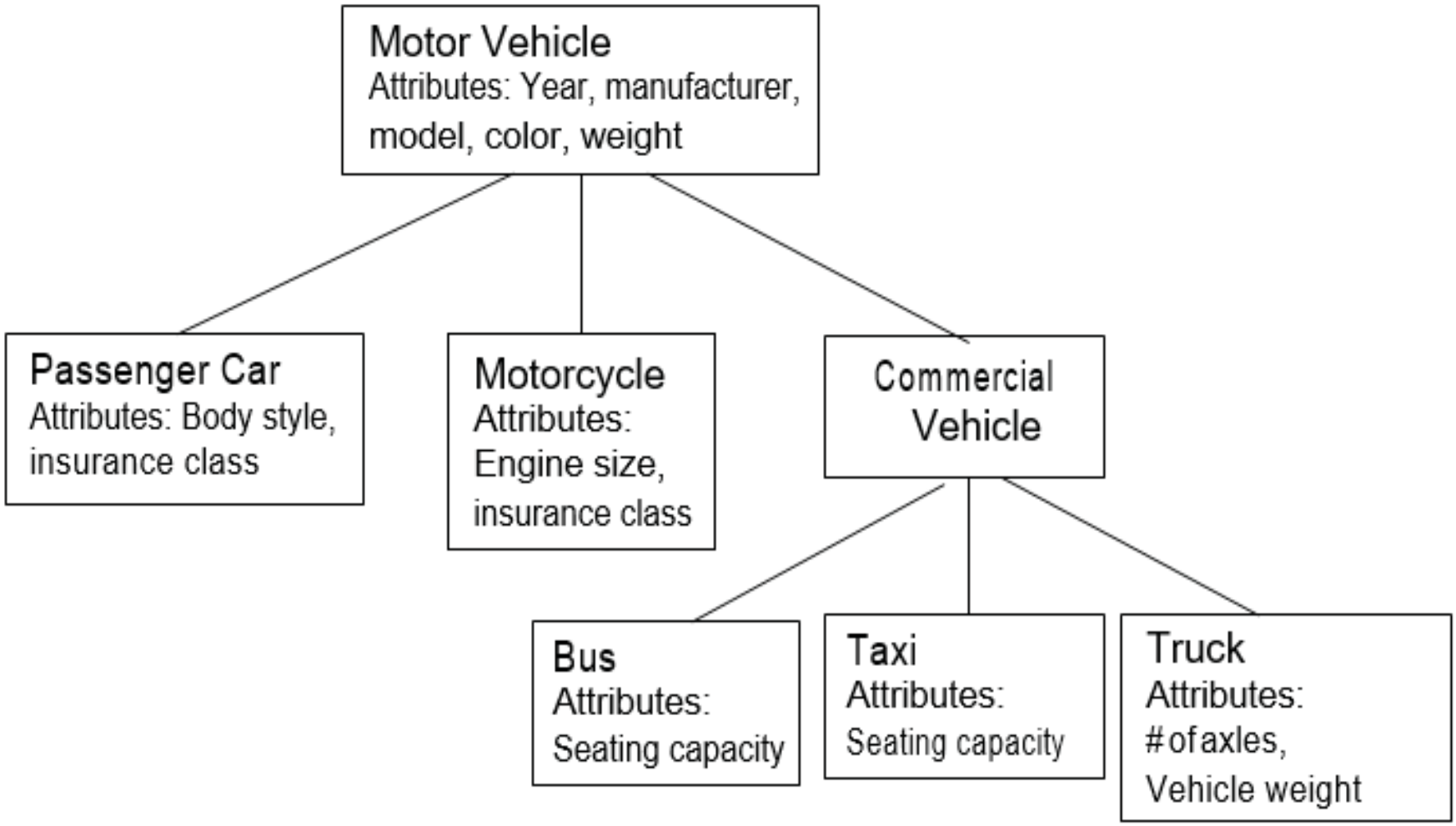


Figure: Motor vehicle types

Defining taxonomies and ontologies (2 of 2)

- Explanation of how information can be interpreted.

Target interpretation
black-control-open-cfile
black-control-semiopen-dfile
doubled-pawn-for-white
uncoordinated-pieces-for-white
weak-pawn-structure-for-white
Model's interpretation
black-control-semiopen-dfile
black-controls-black-squares
uncoordinated-pieces-for-white
weak-pawn-structure-for-white

Figure: Explanation of how information can be interpreted for chess game

Source: https://images.app.goo.gl/RKttjYS8aX4aWFmb6

Defining taxonomies and ontologies (4 of 5)

- Automotive diagnostics and repair.

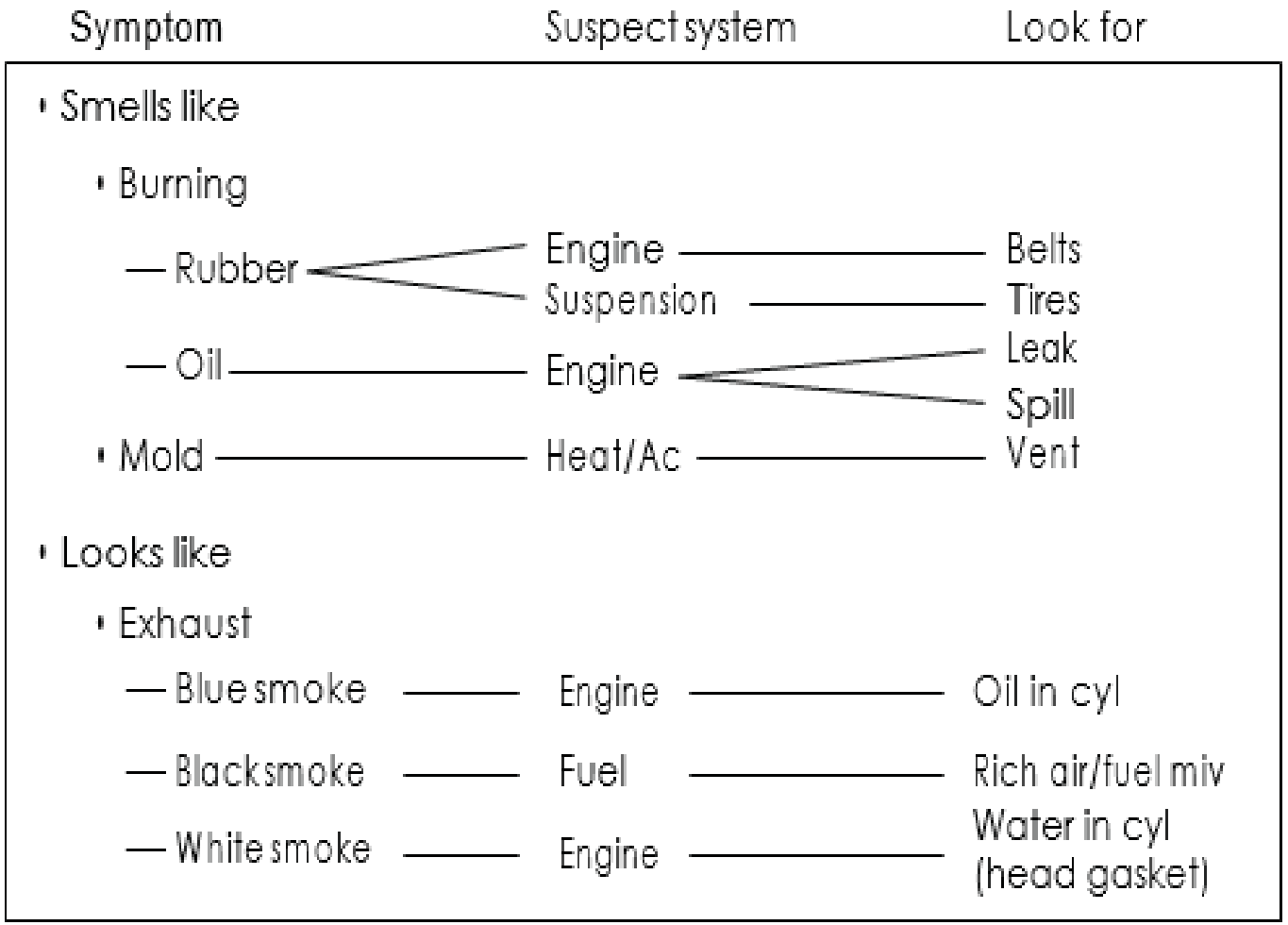


Figure: Automotive diagnostics and repair

Defining taxonomies and ontologies (5 of 5)

- Healthcare.

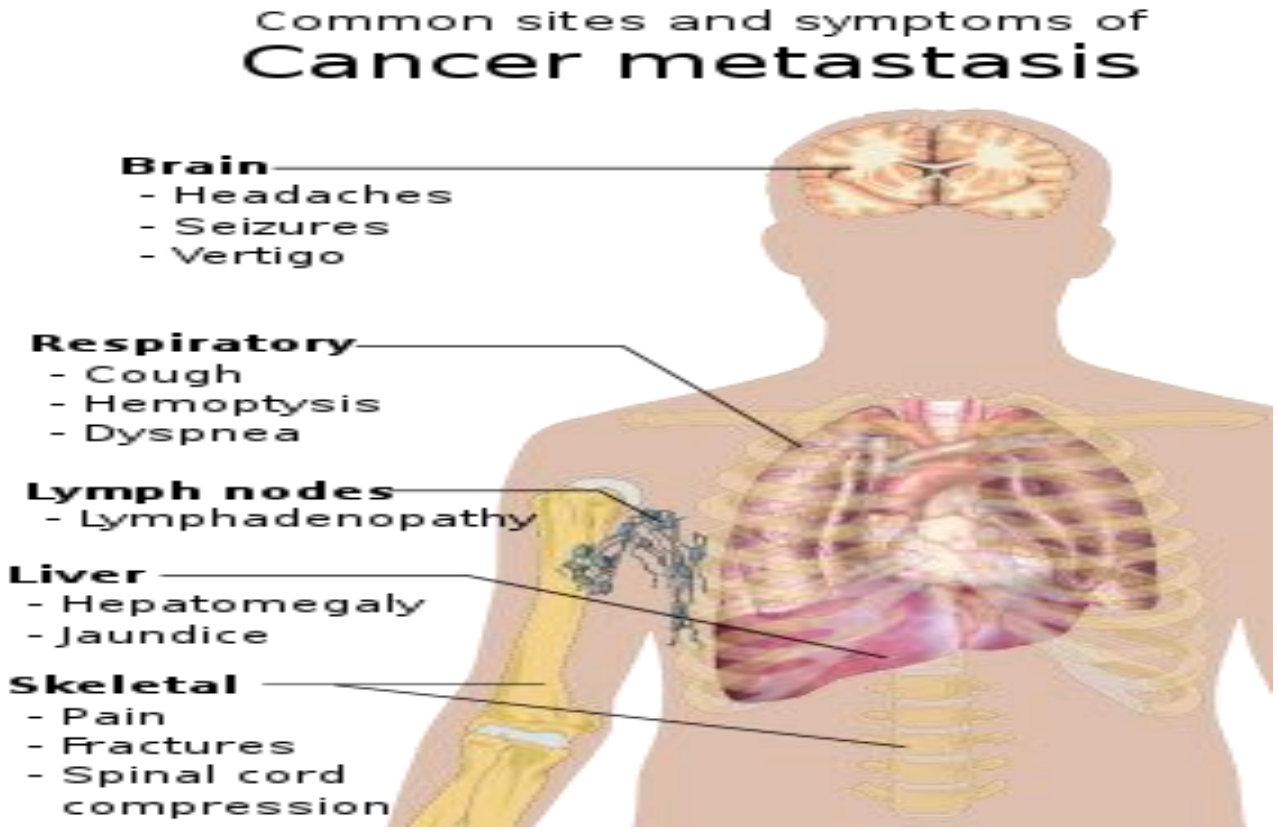


Figure: Common Types of Cancers

Source: https://images.app.goo.gl/jdypAD9DWbTirs7TA

Managing multiple views of knowledge

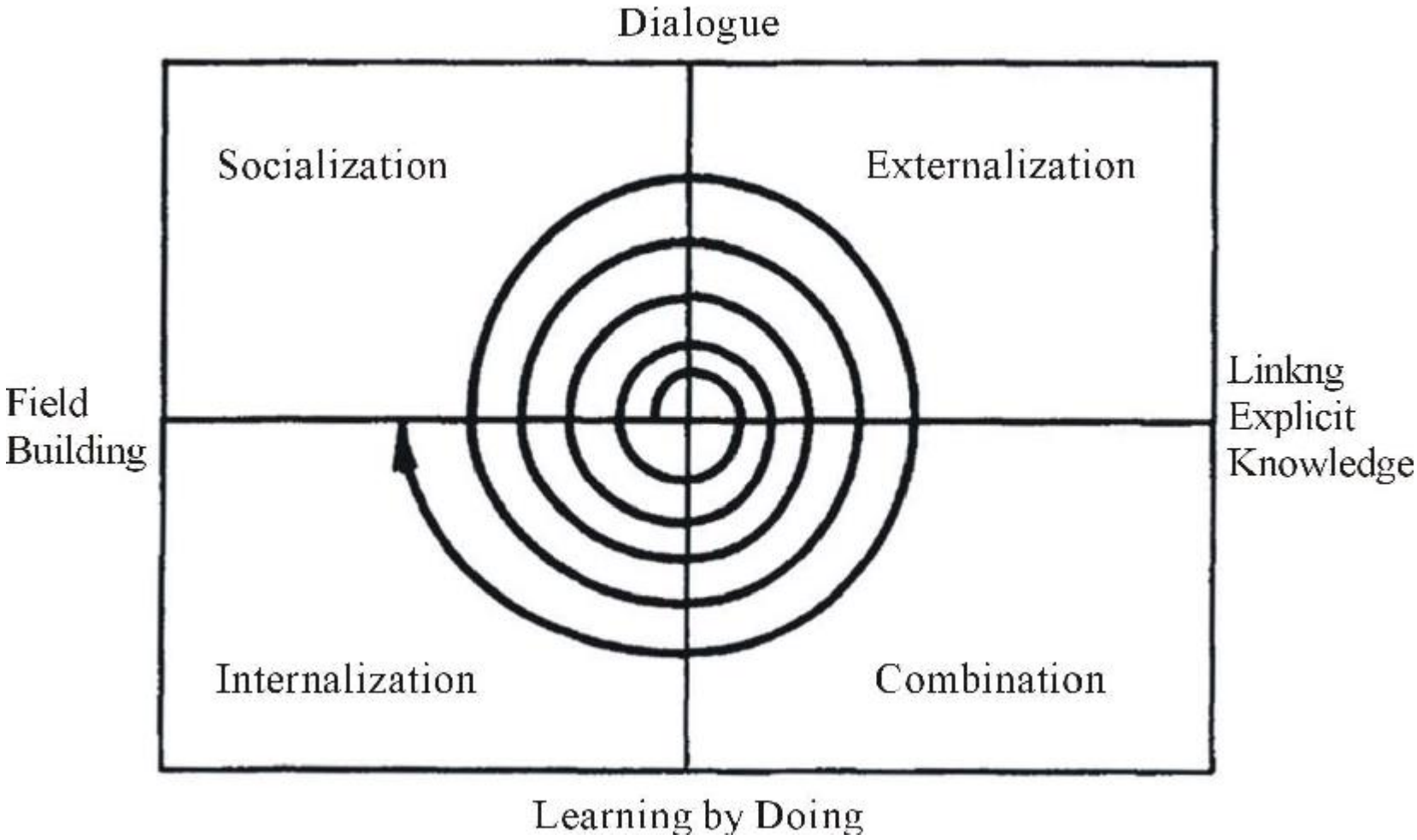


Figure: Knowledge Management

Source: https://images.app.goo.gl/evvDUH29ccUJD8S46

Self evaluation: Exercise 10

- To continue with the training, after learning the various steps involved in cognitive analytics and Advanced NLP operations, it is instructed to utilize the concepts of NLP and Big Data to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 10: NLP for cognitive Analytics – NLTK Package

Taxonomies

Taxonomies provide machines ordered representations. According to Bowles, a Taxonomy represents the formal structure of classes or types of objects within a domain. Bowles noted that taxonomies:

- Follow a hierarchic format and provides names for each object in relation to other objects.
- May also capture the membership properties of each object in relation to other objects.
- Have specific rules used to classify or categorize any object in a domain. These rules must be complete, consistent and unambiguous
- Apply rigor in specification, ensuring any newly discovered object must fit into one and only one category or object
- Inherits all the properties of the class above it but can also have additional properties.

- Finding a book or document in a library or locating a specific website in Google, requires a Taxonomy.

Ontology (1 of 2)

- Ontology as a subset of Taxonomy, but with more information about the behavior of the entities and the relationships between them.
- Ontology as a domain: “Including formal names, definitions and attributes of entities within a domain”.
- The W3C refers to an Ontology as a more complex and quite formal collection of terms.
- Ontologies factor the thinking about how a domain influences such elements as choices of maps and models, rules and representations, and required operations.

Ontology (2 of 2)

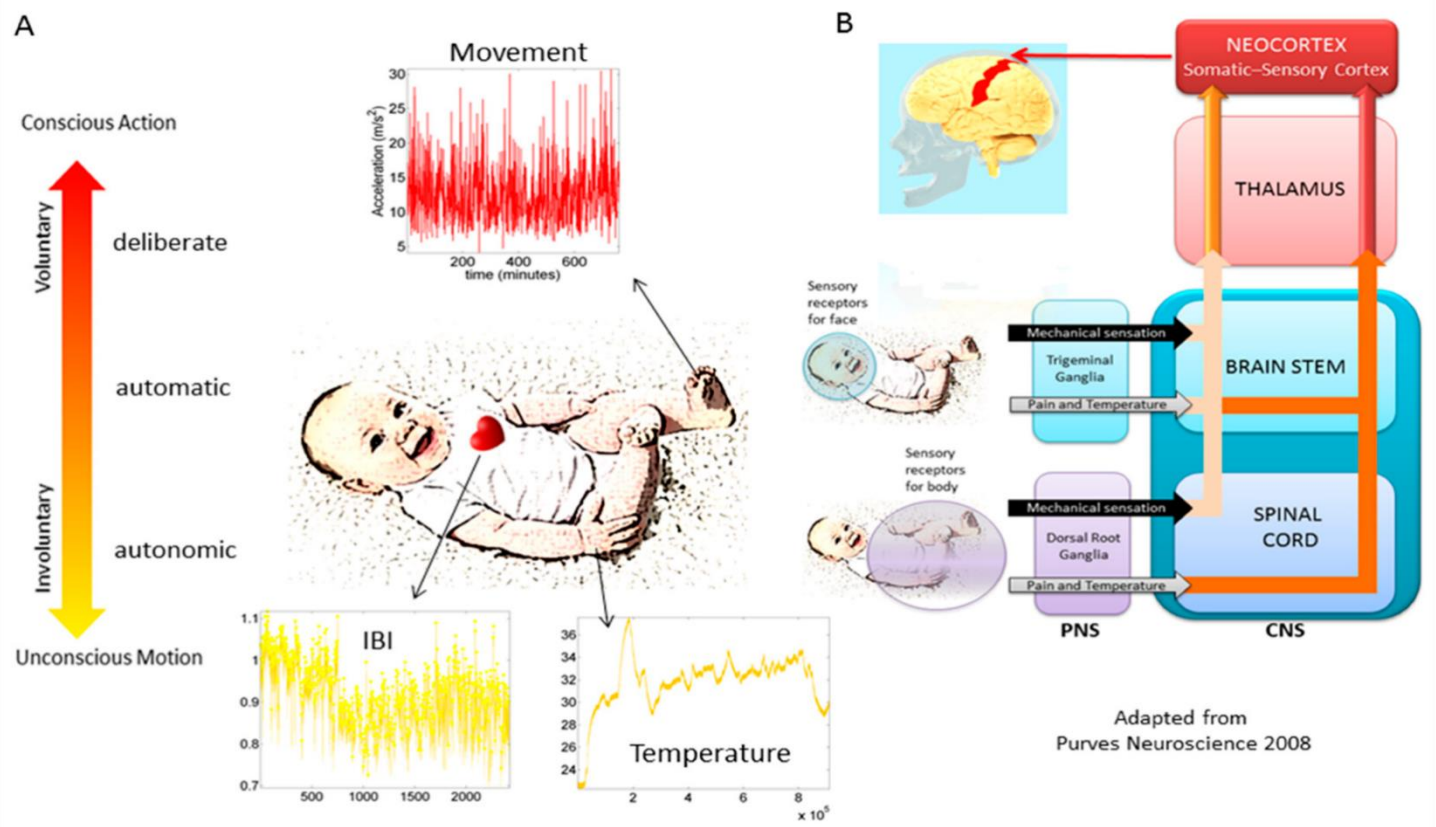


Figure: Taxonomies Progress—Autism in the Diagnostic and Statistical Manual of Mental Disorders

Source: https://images.app.goo.gl/awaAVAA5f79gGt646

Other methods of representing knowledge (1 of 2)

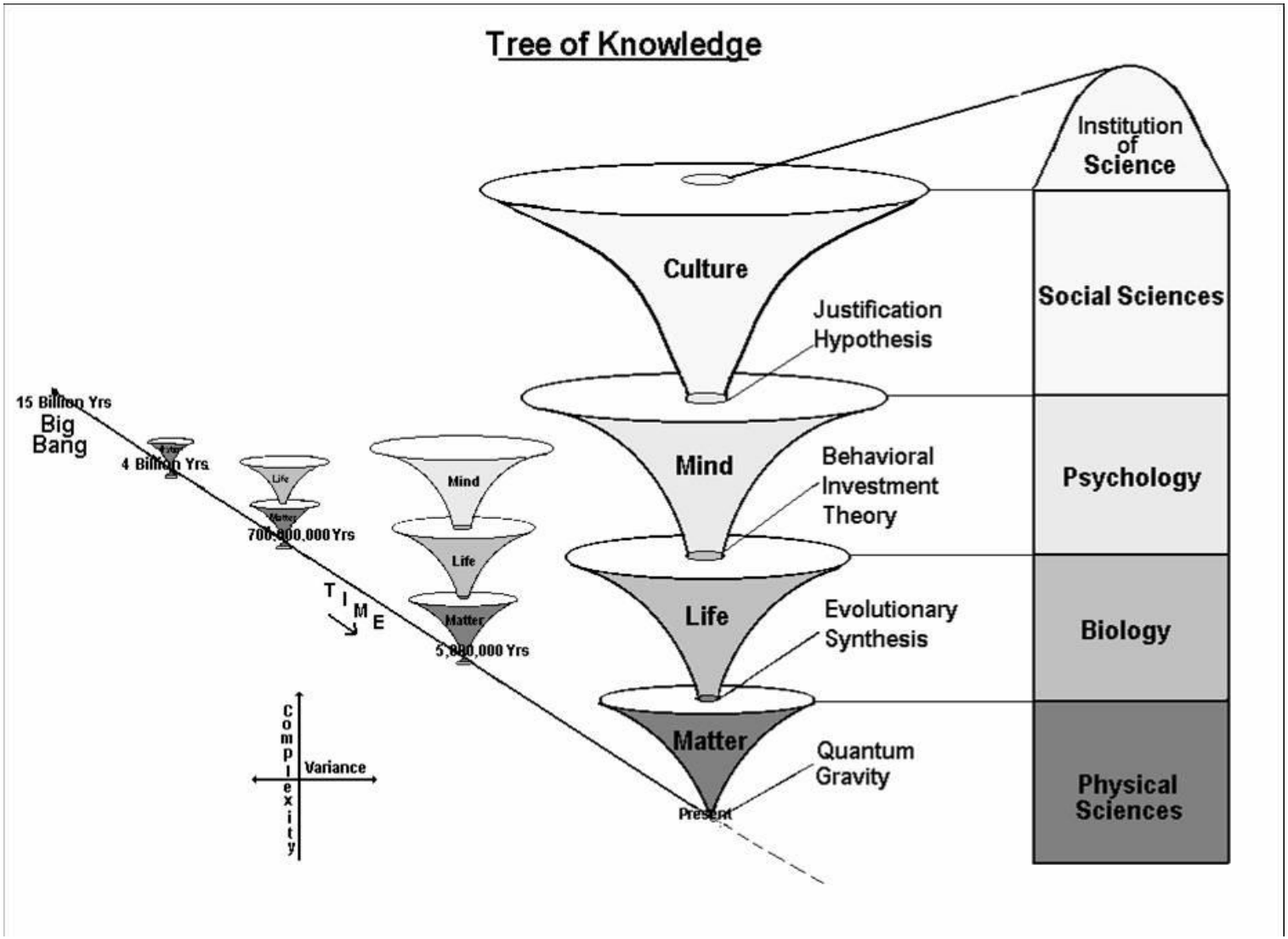


Figure: Simple Tree

Source: https://images.app.goo.gl/DEDycbEVWYAQUMWYA

Other methods of representing knowledge (2 of 2)

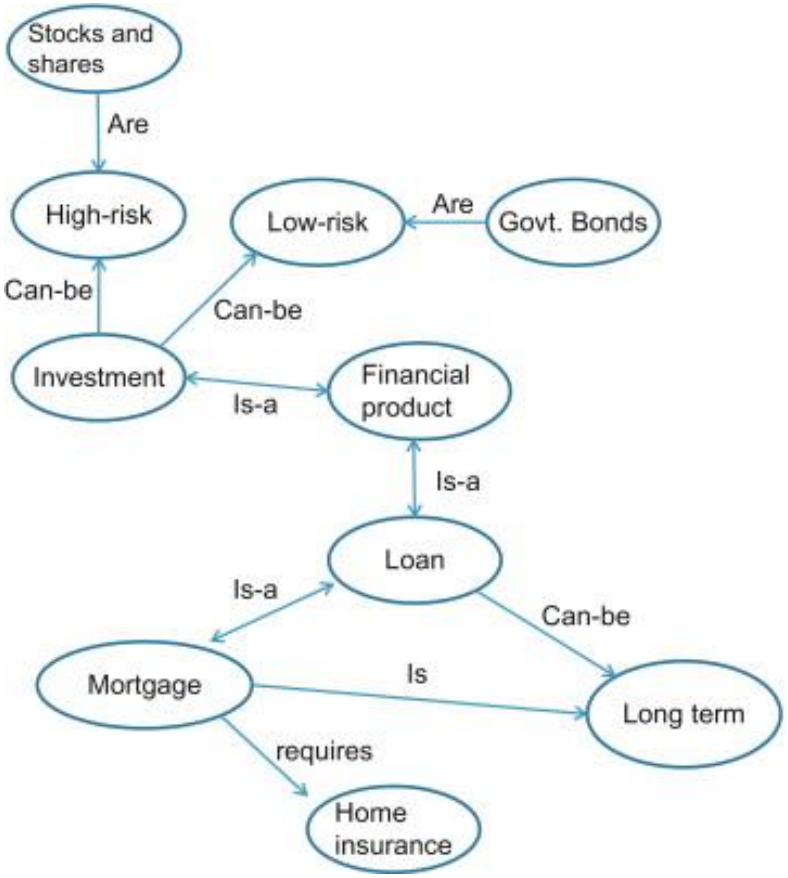


Figure: Semantic Web

Source: https://images.app.goo.gl/ghPXVjhpKsGLbctYA

Persistence and policy significance

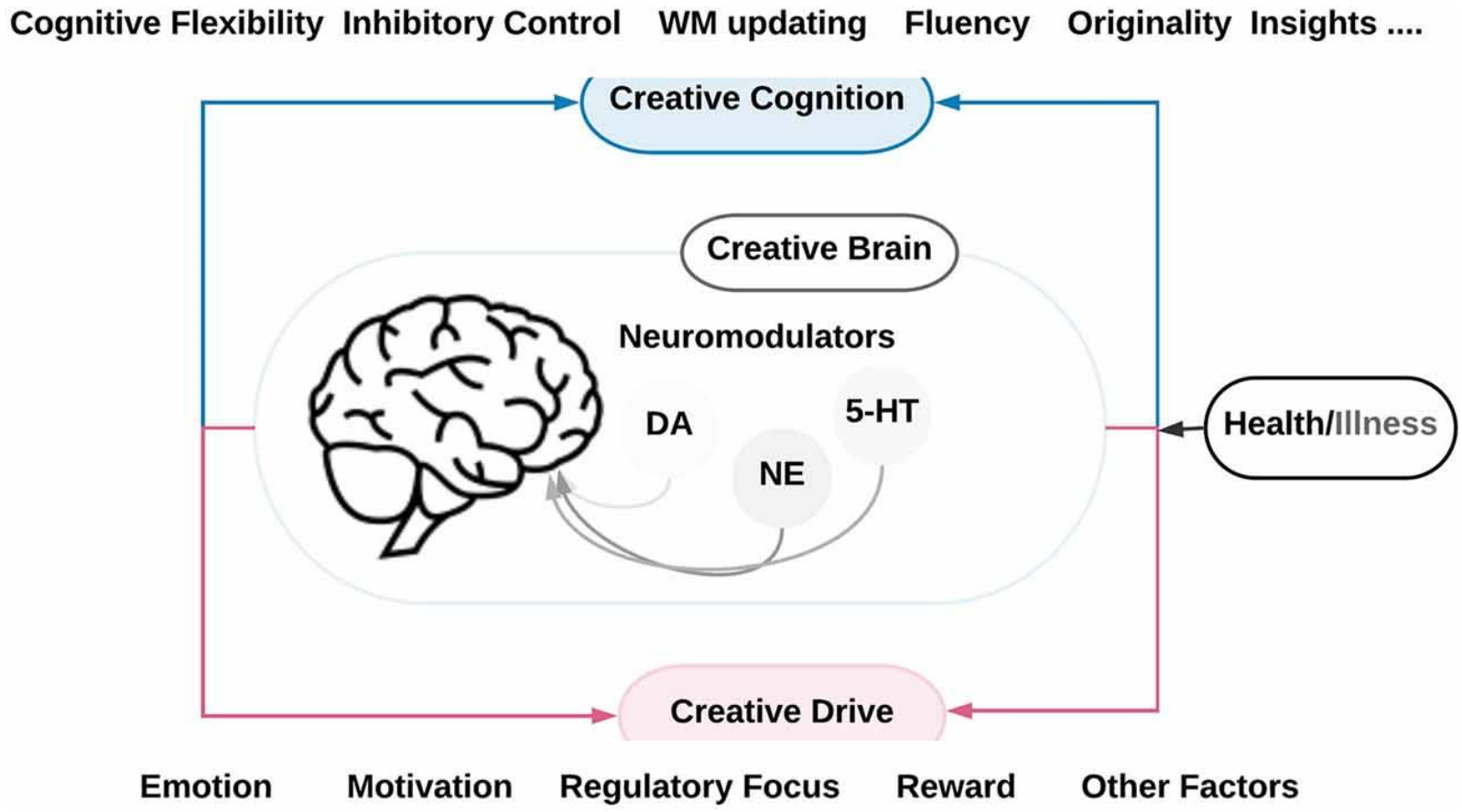
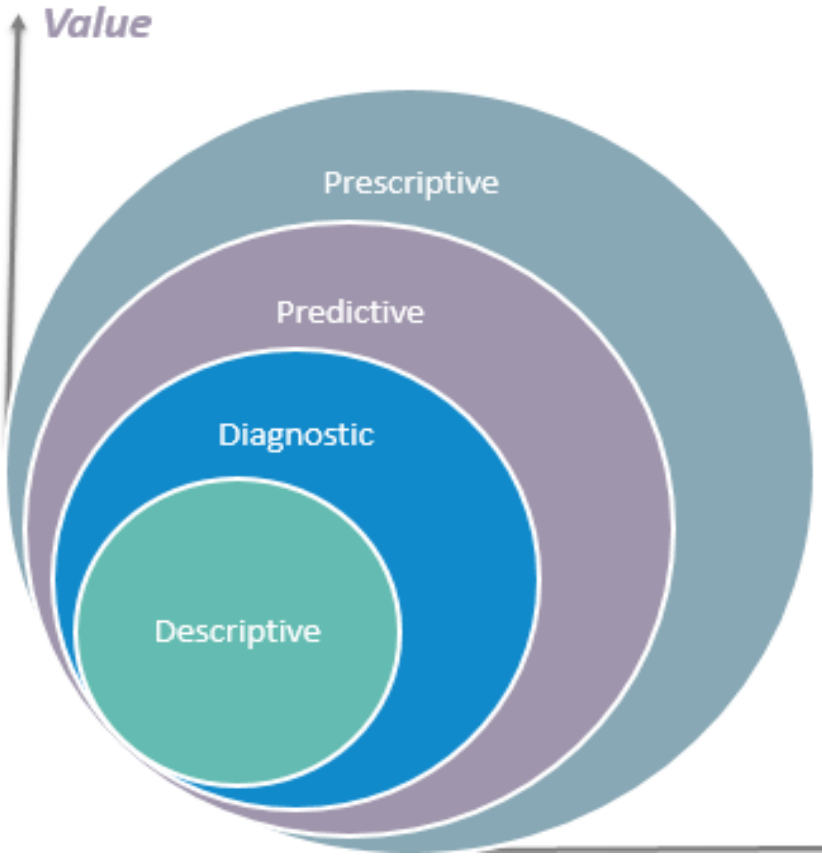


Figure: Persistence and Policy Significance for cognitive approach

Source: https://images.app.goo.gl/o92R43dqvKLTSxRg9

Advanced analysis is on a cognitive computer path (1 of 2)



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Figure: Types of analytics

Source: https://images.app.goo.gl/HAb8uLHFHhMZdKqx8

Advanced analysis is on a cognitive computer path (2 of 2)

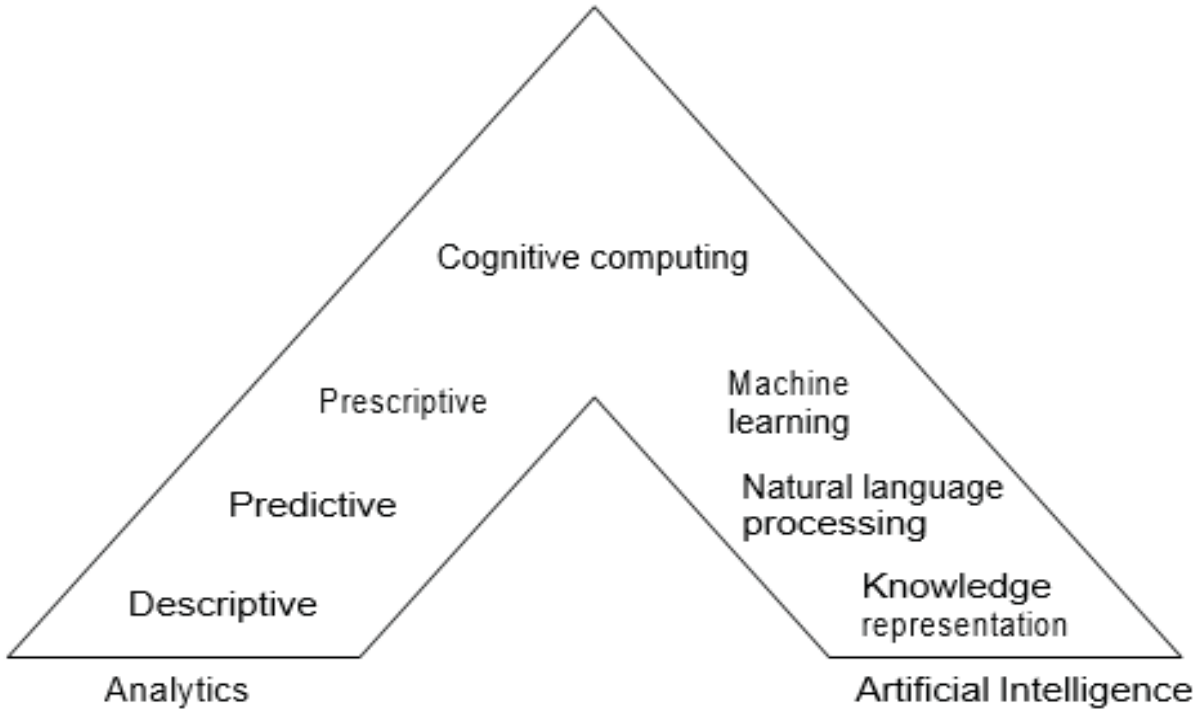


Figure: Converging technologies: analytics and artificial intelligence

Expert research primary capabilities

- The relationship between statistics, the mining of information and the learning process.

	MACHINE LEARNERS	STATISTICIANS
Network/Graphs vs. Models	Network/Graphs to train and test data	Models to create predictive power
Weights vs. Parameters	Weights used to maximize accuracy scoring and hand tuning	Parameters used to interpret real-world phenomena - stress on magnitude
Confidence Interval	There is no notion of uncertainty	Capturing the variability and uncertainty of parameters
Assumptions	No prior assumption (we learn from the data)	Explicit a-priori assumptions
Distribution	Unknown a priori	A-priori well-defined distribution
Fit	Best fit to learning models (generalization)	Fit to the distribution

Figure: Machine learners and Statisticians difference

Source: https://images.app.goo.gl/gecsxJkt2QruTM8JA

Self evaluation: Exercise 11

- To continue with the training, after learning the various steps involved in cognitive analytics and Time Series operations, it is instructed to utilize the concepts of Time Series and Big Data to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 11: Analyzing Time Series Data for cognitive processing

Usage of computational machine learning for analytics process

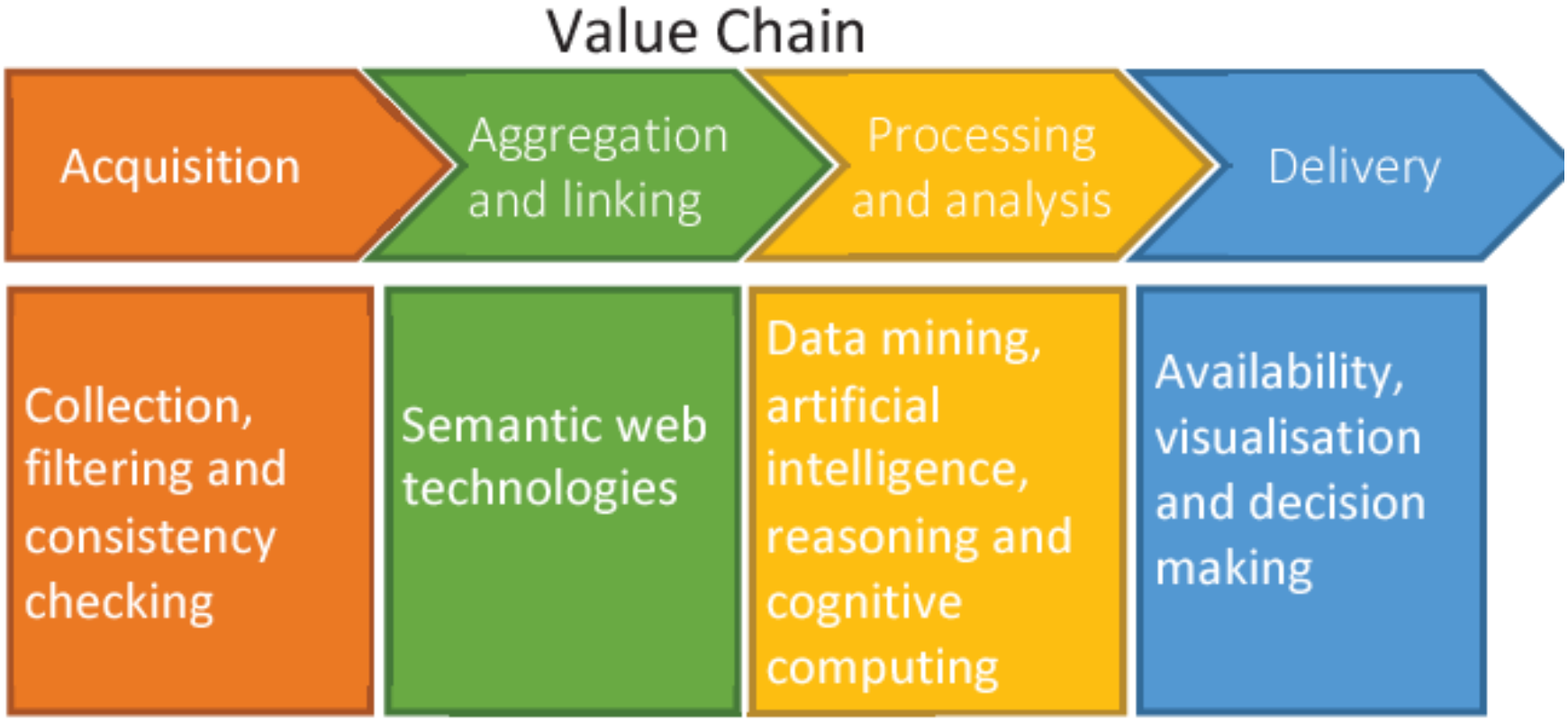


Figure: Value chain for ML Analytical process

Source: https://images.app.goo.gl/hTS61b8hiVTKC8Xu9

Models based on the kind of outputs from the algorithms (1 of 2)

- Classification.

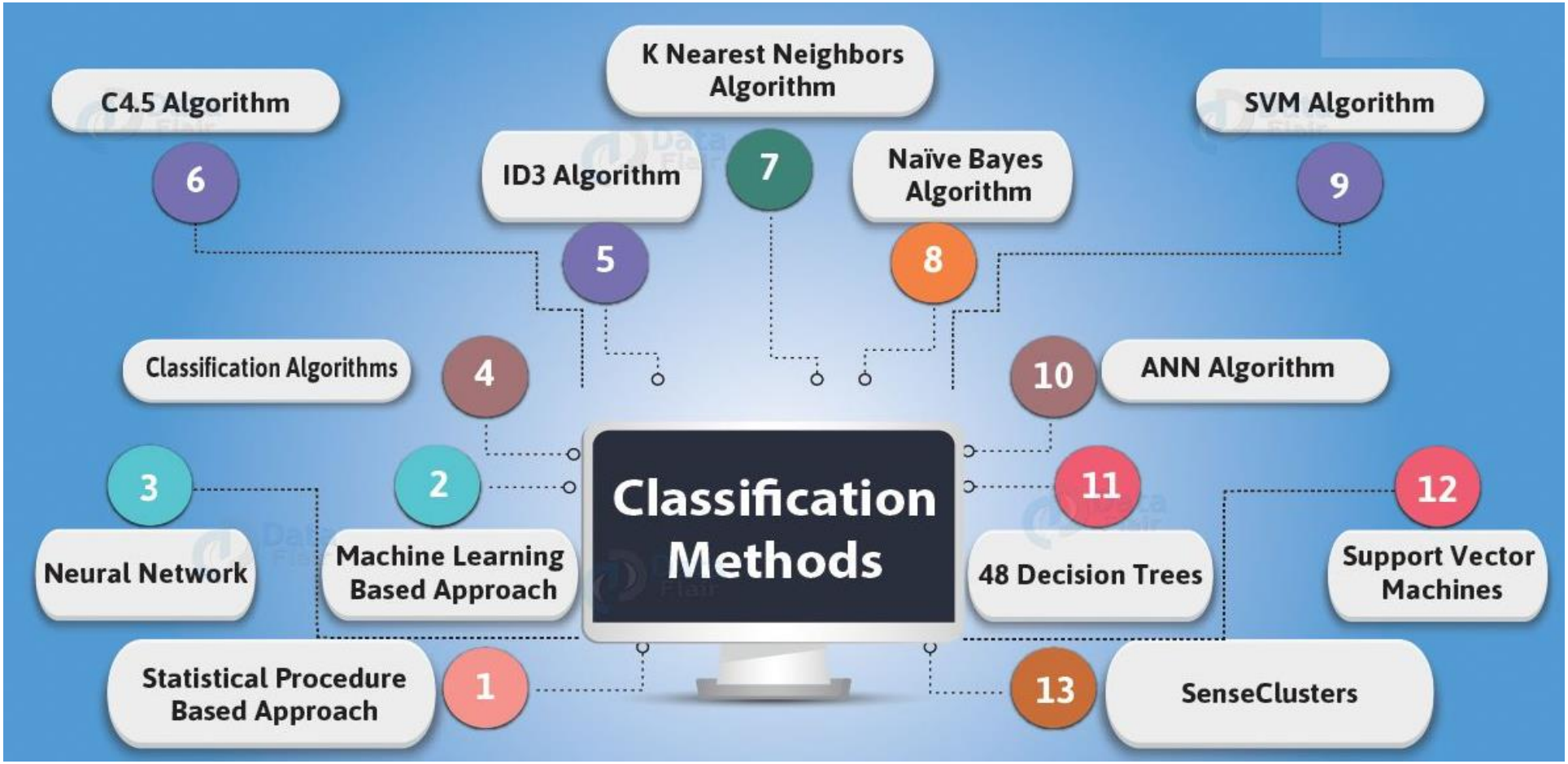


Figure: Classification algorithms

Source: https://images.app.goo.gl/pUwTkr7cBDSH4jAD9

Models based on the kind of outputs from the algorithms (2 of 3)

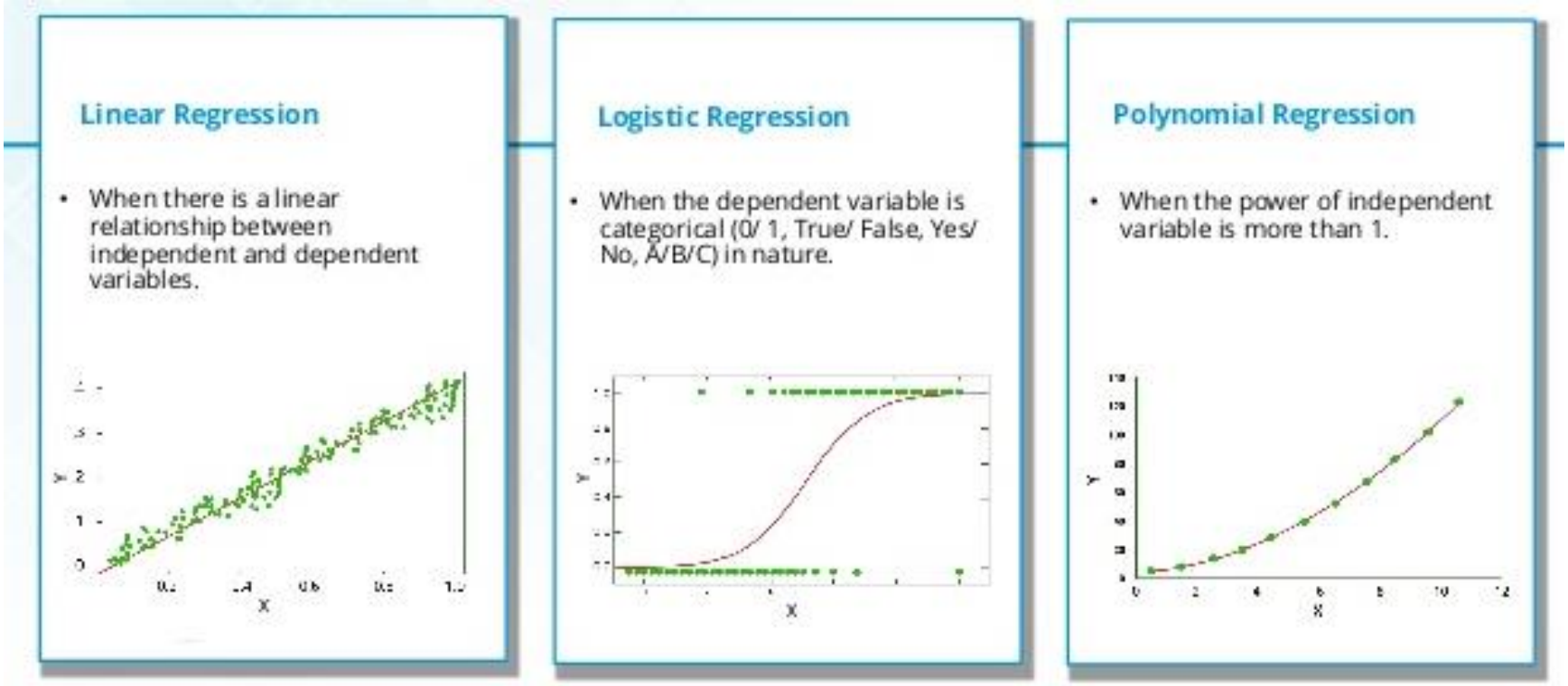


Figure: Regression algorithms

Source: https://images.app.goo.gl/iuijWm2rNSePtYca8

Models based on the kind of outputs from the algorithms (2 of 2)

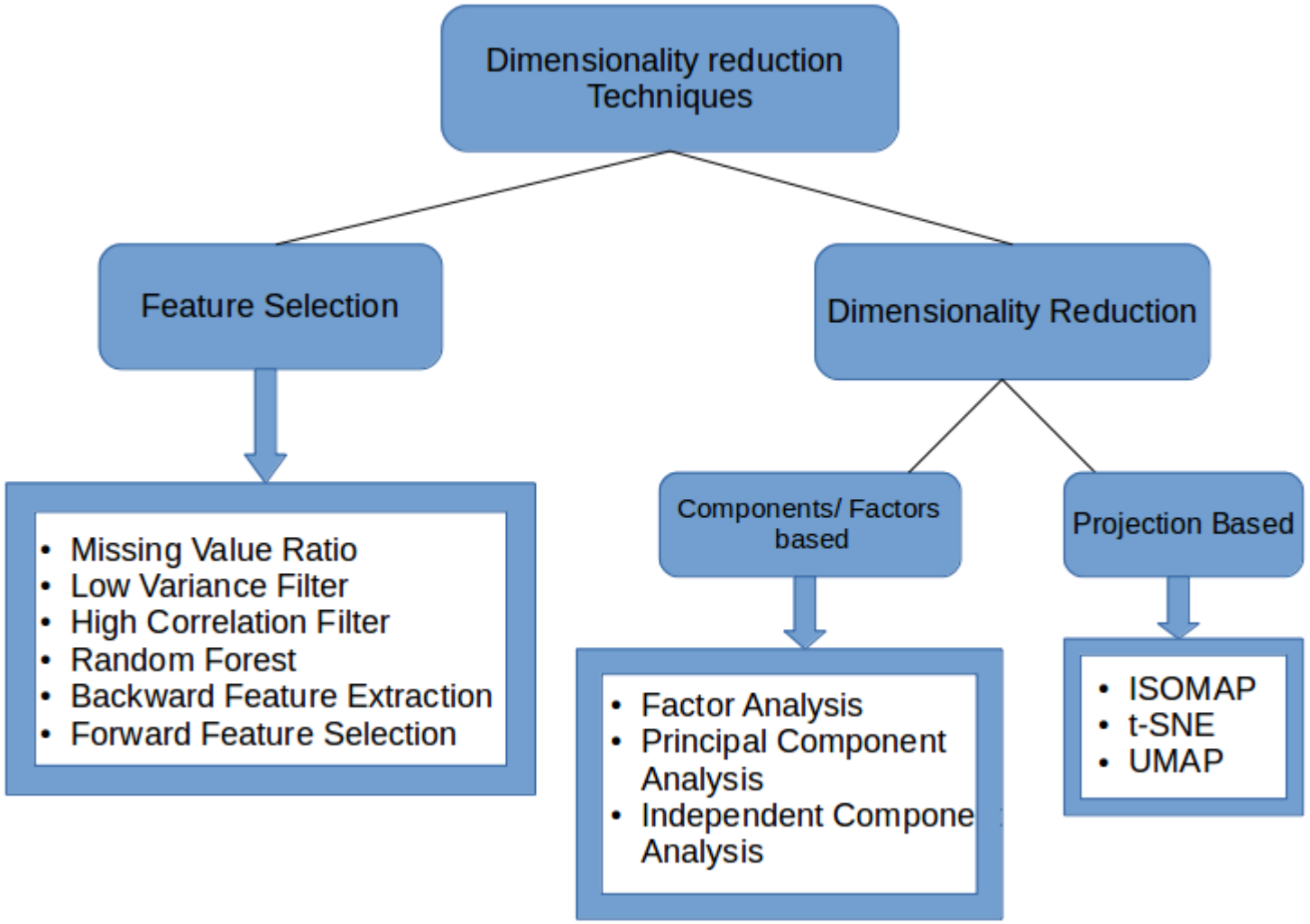


Figure: Dimensionality Reduction Techniques

Source: https://images.app.goo.gl/Hw3KW757vVBnREvS8

Self evaluation: Exercise 12

- To continue with the training, after learning the various steps involved in cognitive analytics and Speech analysis operations, it is instructed to utilize the concepts of ML Techniques and Speech recognition package to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 12: Cognitive Speech Recognition

Predictive analytics

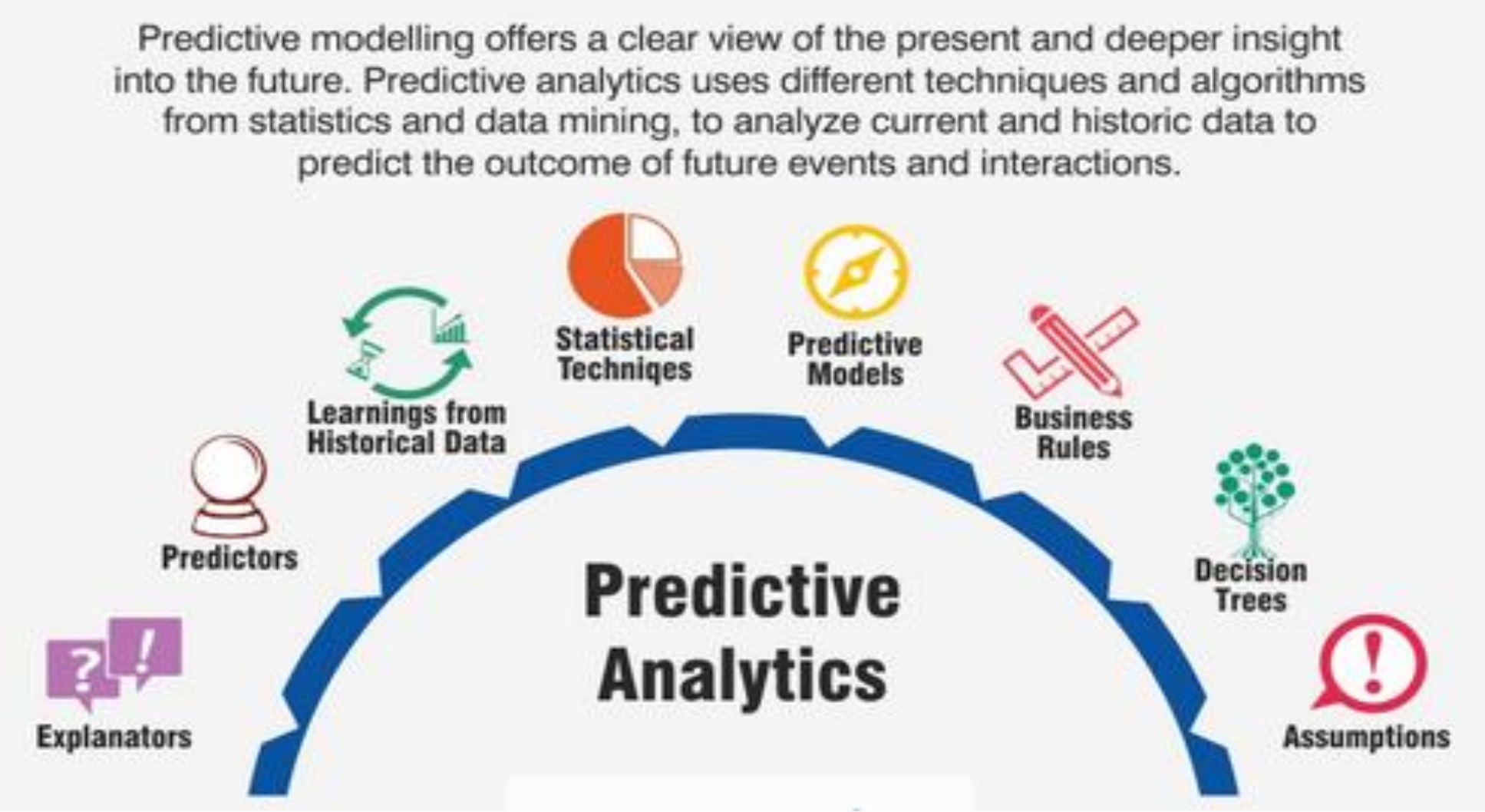


Figure: Predictive Analytics

Source: https://images.app.goo.gl/GWxor7JKRqtbtKeo8

Business value of predictive analytics

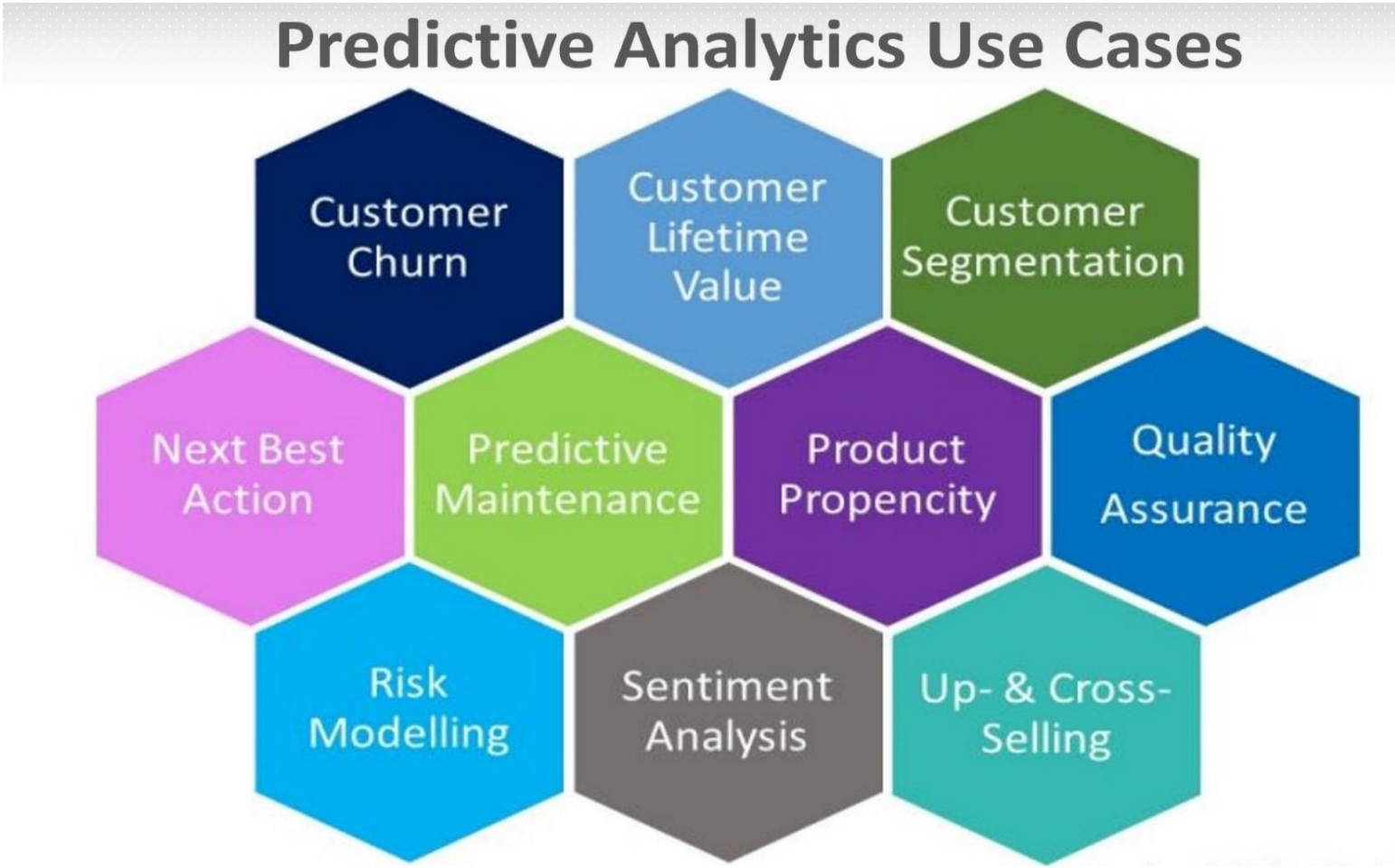


Figure: Predictive Analytics Use Cases

Source: https://images.app.goo.gl/Xm13ArsrgbFsUMRdA

Text mining and text analytics (1 of 2)

- Text Mining and Text Analytics solve the same problems but use different techniques and are complementary ways to automatically extract meaning from text.

- Text Analytics is developed within the field of computational linguistics. It has the ability to encode human understanding into a series of linguistic rules which are generated by humans are high in precision, but they do not automatically adapt and are usually fragile when tried in new situations.

- Text mining is a newer discipline arising out of the fields of statistics, data mining, and machine learning. Its strength is the ability to inductively create models from collections of historical data. Because statistical models are learned from training data they are adaptive and can identify "unknown unknowns", leading to the better recall. Still, they can be prone to missing something that would seem obvious to a human.

- Text analytics and text mining approaches have essentially equivalent performance. Text analytics requires an expert linguist to produce complex rule sets, whereas text mining requires the analyst to hand-label cases with outcomes or classes to create training data. Due to their different perspectives and strengths, combining text analytics with text mining often leads to better performance than either approach alone.

Figure: Text mining and text analytics

Text mining and text analytics (2 of 2)

- Information retrieval.
- Data preparation and cleaning.
- Segmentation.
- Tokenization.
- Stop-word numbers and punctuation removal.
- Stemming.
- Convert to lowercase.
- POS tagging.
- Create text corpus.
- Term-Document matrix

Significant variations in data mining and text processing

Level of text preprocessing needed

	Domain Specific / Noisy Texts	General / Well Written Texts
Lots of data	- Moderate pre-processing - Text enrichment could be helpful	- Light pre-processing - Text enrichment could be helpful, but not critical
Sparse data	- Heavy pre-processing - Text enrichment is important	- Moderate pre-processing - Text enrichment could be helpful

Figure: Text Preprocessing labels

Source: https://images.app.goo.gl/cpam8UkbT3u6SKrTA

Self evaluation: Exercise 13

- To continue with the training, after learning the various steps involved in cognitive analytics and Heuristic Search operations, it is instructed to utilize the concepts of ML Techniques and Heuristic Search package to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 13: Heuristic Search in AI and cognitive analysis

Sentiment analysis

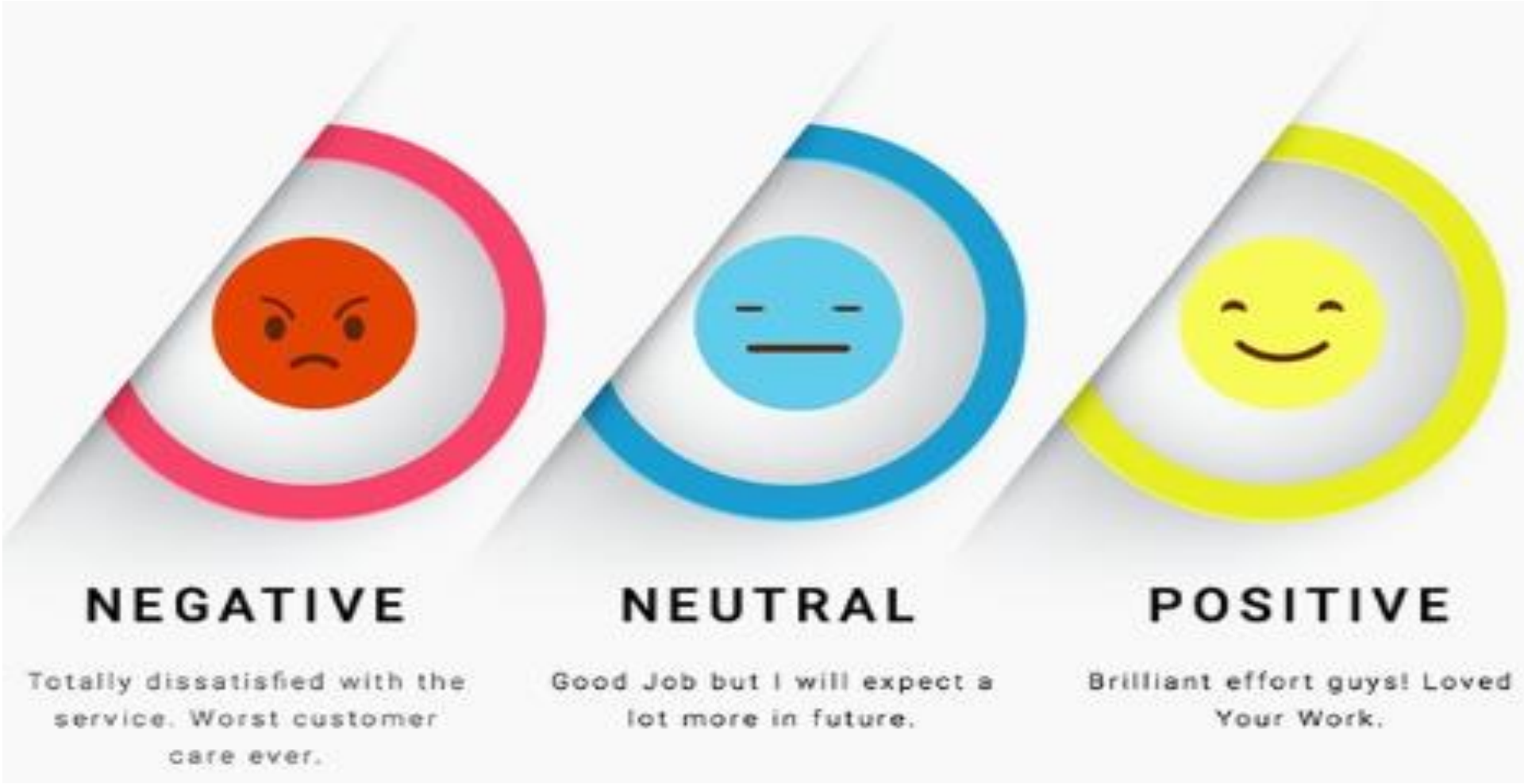


Figure: Sentimental Analysis

Source: https://images.app.goo.gl/4GoJA8gEj245eoBQ7

Text analytics business value

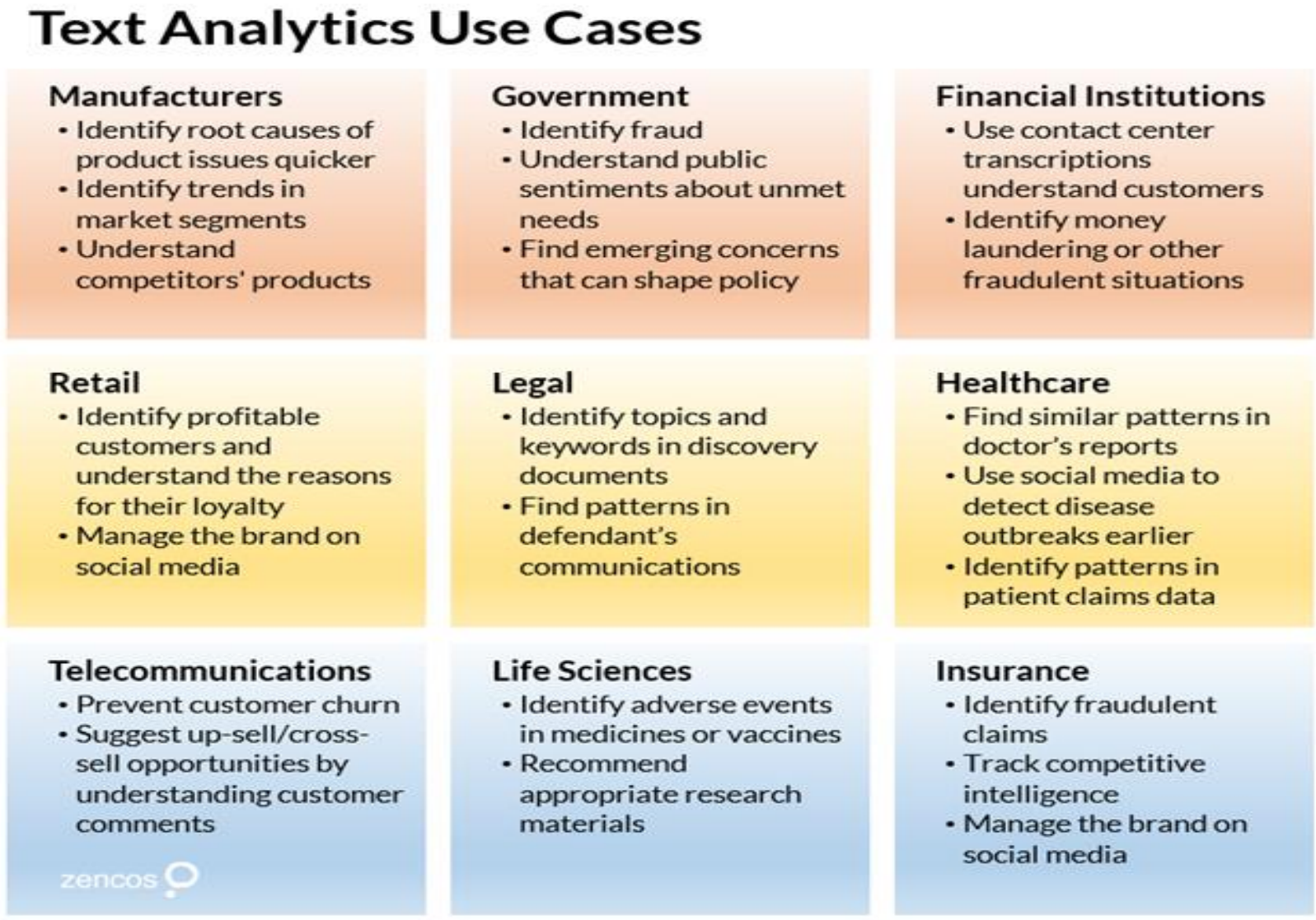


Figure: Text Analytics Use Cases

Source: https://images.app.goo.gl/iXxn3xph5uLSXrgf9

Image analytics

- Image analysis (also known as “computer vision” or image recognition) is the ability of computers to recognize attributes within an image.

- Image analytics can also identify faces within photos to determine sentiment, gender, age, and more. It can recognize multiple elements within a photo at the same time, including logos, faces, activities, objects, and scenes.

- There are some big advantages to looking at both text and images when analyzing social media data:
  - Images do not require translation making image analysis extremely useful in a global strategy.
  - Looking at a more complete data set enables businesses to more effectively incorporate social insights into decision-making.
  - Images can tell a completely different story than text mentions (Example: Text-based analysis of conversation around Disney’s Frozen, shows adults in their 30s. Image analysis of the same conversation shows the movie’s true audience, children).

Speech analytics

- Speech analytics is the process of analyzing recorded calls to gather customer information to improve communication and future interaction.

- The process is primarily used by customer contact centers to extract information buried in client interactions with an enterprise. Although speech analytics includes elements of automatic speech recognition, it is known for analyzing the topic being discussed, which is weighed against the emotional character of the speech and the amount and locations of speech versus non-speech during the interaction.

- Speech analytics in contact centers can be used to mine recorded customer interactions to surface the intelligence essential for building effective cost containment and customer service strategies.

- The technology can pinpoint cost drivers, trend analysis, identify strengths and weaknesses with processes and products, and help understand how the marketplace perceives offerings.

Using advanced analytics to create value

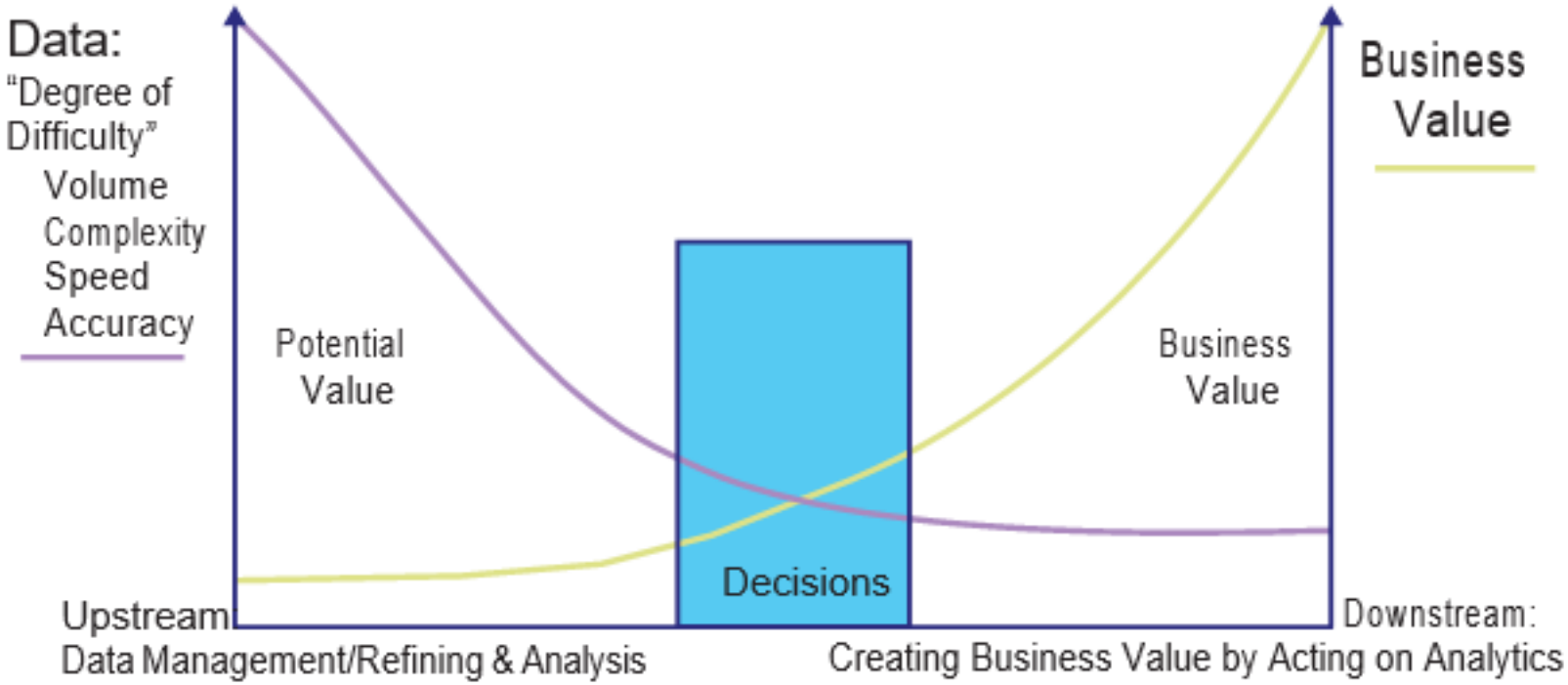


Figure: Business value

Source: “An Executive Guide to Analytics Infrastructure," January 2014 by STORM Insights, Inc.

Self evaluation: Exercise 14

- To continue with the training, after learning the various steps involved in cognitive analytics and Basic Gaming operations, it is instructed to utilize the concepts of ML Techniques , Gaming techniques and different Search package to perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 14: Gaming based on cognitive computing

Checkpoint (1 of 2)

Multiple choice questions:

1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____
   a) Improved data storage and information retrieval
   b) Improved extract, transform and load features for data integration
   c) Improved data warehousing functionality
   d) Improved security, workload management, and SQL support

2. Which of the following involves predicting a categorical response?
   a) Regression
   b) Summarization
   c) Clustering
   d) Classification

3. Which of the following method can be used to combine different classifiers?
   a) Model stacking
   b) Model combining
   c) Model structuring
   d) None of the above

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____.
   a) Improved data storage and information retrieval
   b) Improved extract, transform and load features for data integration
   c) Improved data warehousing functionality
   d) Improved security, workload management, and SQL support

2. Which of the following involves predicting a categorical response?
   a) Regression
   b) Summarization
   c) Clustering
   d) Classification

3. Which of the following method can be used to combine different classifiers?
   a) Model stacking
   b) Model combining
   c) Model structuring
   d) None of the above

Checkpoint (2 of 2)

Fill in the blanks:

1. _____ is simplest class of analytics.
2. Data that summarize all observations in a category are called _____ data.
3. Predicting with trees evaluate _____ within each group of data.
4. _____function is used for k-means clustering?

True or False:

1. Model based prediction considers relatively easy version for covariance matrix. True/False
2. Predictive analytics is same as forecasting. True/False
3. OpenCV is use for Image Processing. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Predictive** is simplest class of analytics.
2. Data that summarize all observations in a category are called **summarized** data.
3. Predicting with trees evaluate **homogeneity** within each group of data.
4. **k-means** function is used for k-means clustering?

True or False:

1. Model based prediction considers relatively easy version for covariance matrix. **False**
2. Predictive analytics is same as forecasting. **False**
3. OpenCV is use for image processing**. True**

Question bank

Two mark question:

1. What is big data analytics?
2. What is Hadoop ecosystem?
3. What is sentimental analysis?
4. What is text analytics?

Four mark question:

1. What is the difference between text analytics and text mining?
2. Explain types of analytics.
3. Explain the components of word cloud.
4. Explain dark data process.

Eight mark question:

1. Explain business values of analytics.
2. Explain the difference between ontology vs taxonomy.

Unit summary

Having completed this unit, you should be able to:

- Understand the concepts of dealing with human-generated data and big data
- Learn about 4 V's of bigdata and bigdata architecture with Hadoop ecosystem
- Gain knowledge on types of data and data services
- Understand taxonomies and ontologies, advanced analytics leads to cognitive computing
- Gain an insight into key capabilities in advanced analytics
- Learn about the relationship between statistics
- Understand the concepts of predictive analytics, text analytics, business value of text analytics, contents image analytics, and speech analytics