Big Data Analytics: structured, semi-structured, and unstructured

DR. CHOUDUR LAKSHMINARAYAN
Department of Statistics and Data Sciences, TERADATA LABS, Advanced Algorithms R&D,

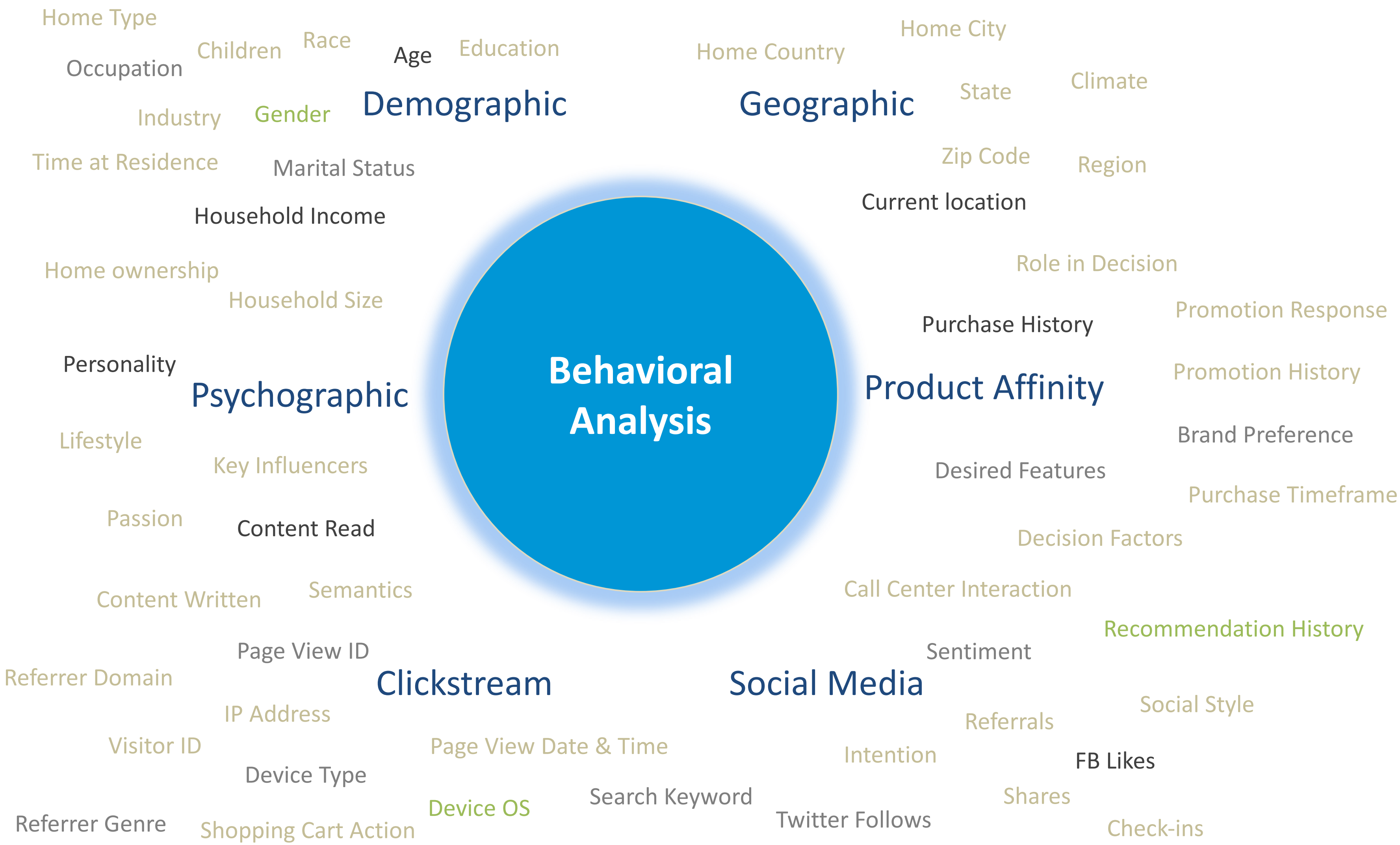
Span of Big Data Analytics:

Challenges in Data Management and Analytics

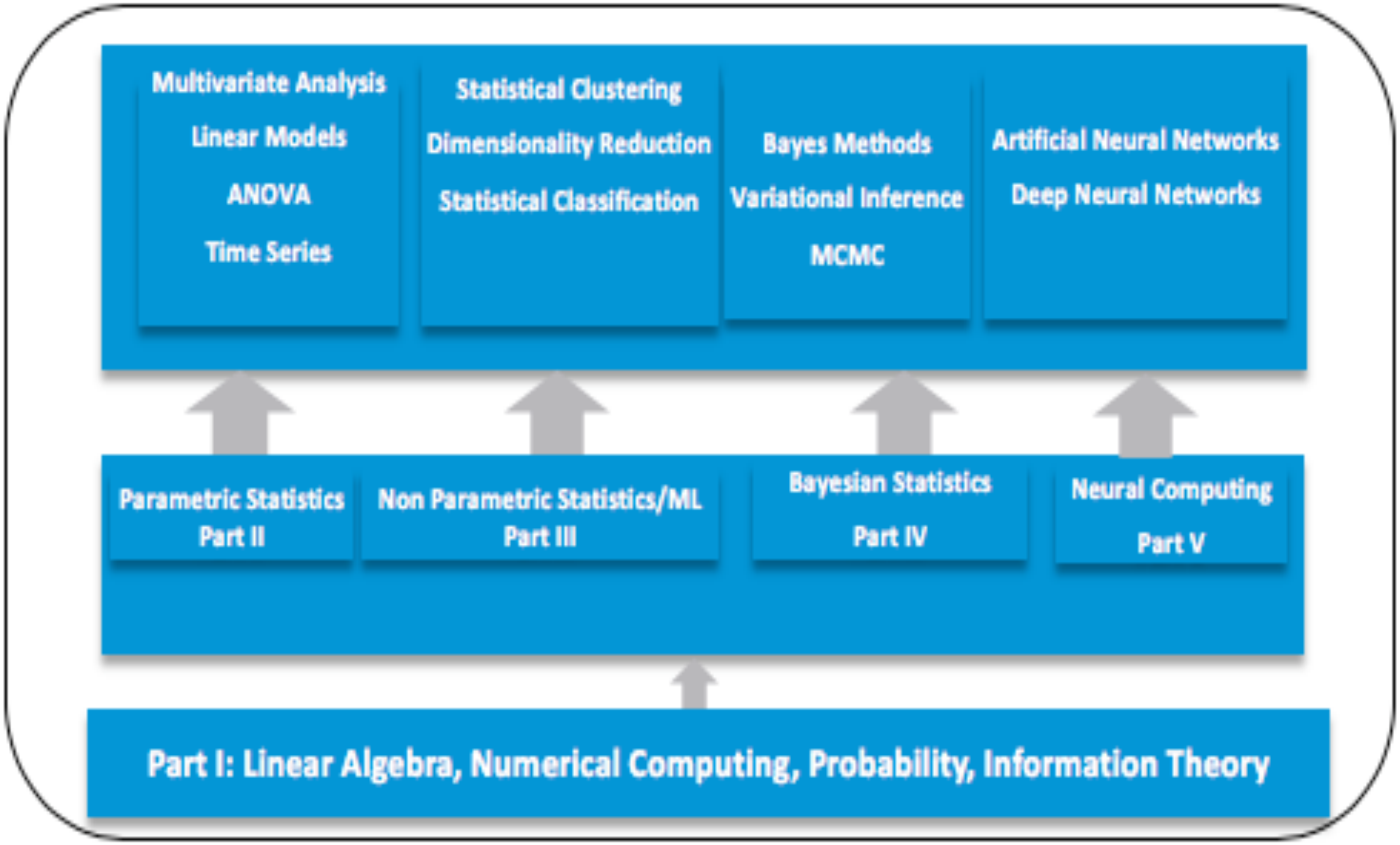
1. Dealing with highly distributed data sources
2. Tracking data provenance, from data generation through data preparation and Validating data
3. **Coping with sampling biases and heterogeneity**
4. Working with different data formats and structures
5. **Developing algorithms that exploit parallel and distributed architectures**
6. Ensuring data integrity
7. Ensuring data security
8. Enabling data discovery and integration
9. Enabling data sharing
10. Developing methods for visualizing massive data
11. **Developing scalable and incremental algorithms**
12. **Coping with the need for real-time analysis and decisio**n-making

Source: National Academy of Sciences, Frontiers in Massive Data Analysis

Typical Data

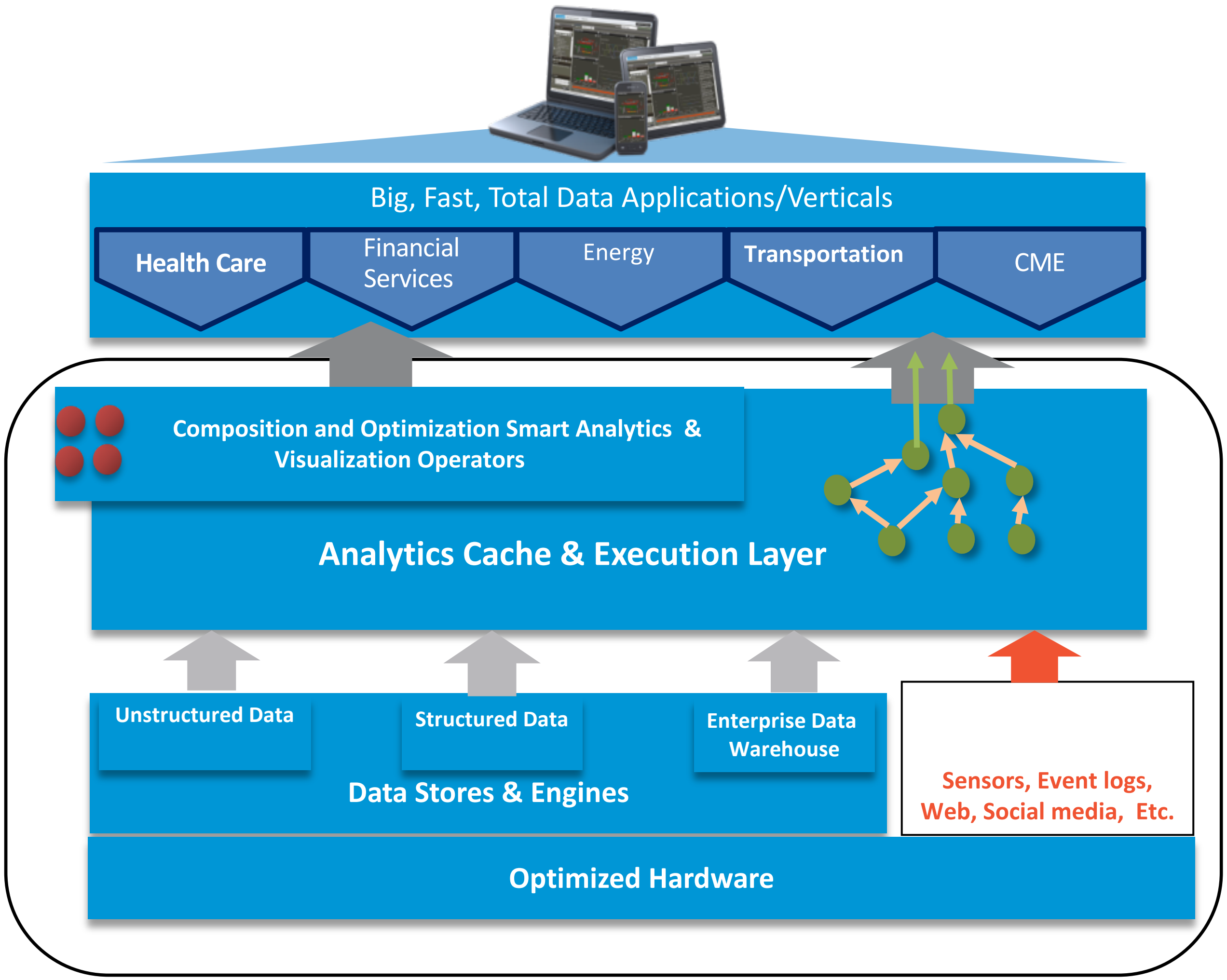


Elements of Data Science



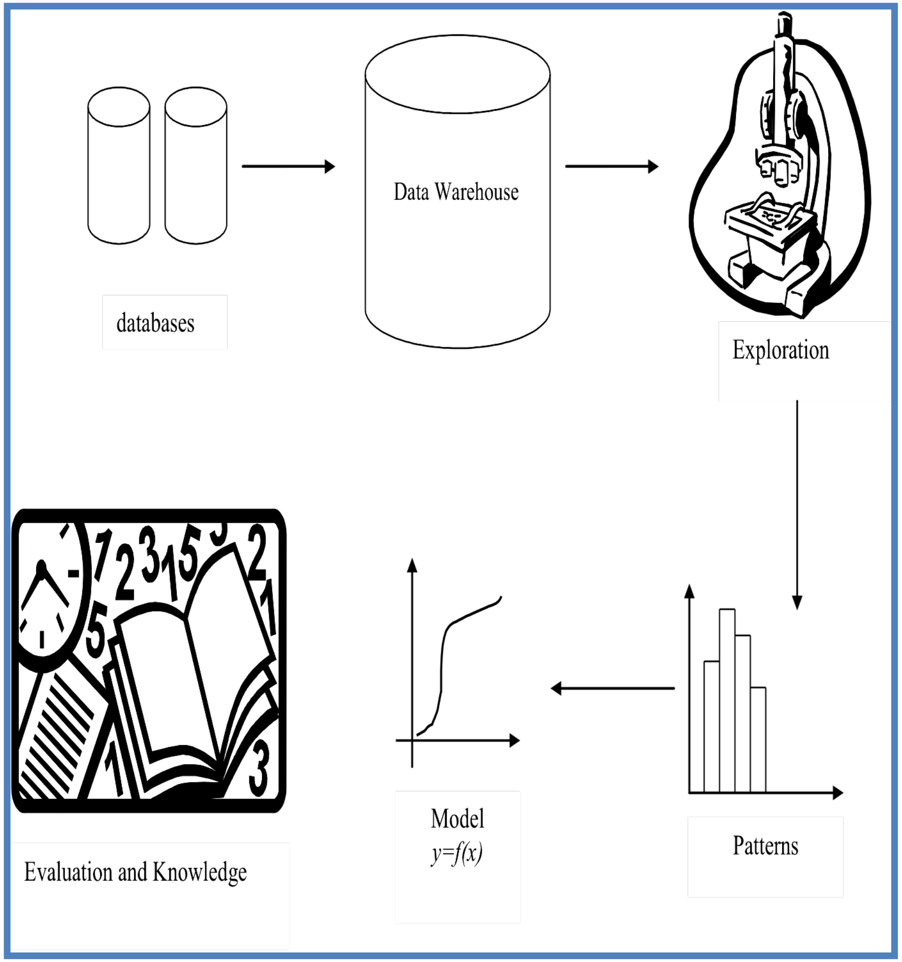
Data Science Life Cycle

The Future of Big Data Computing

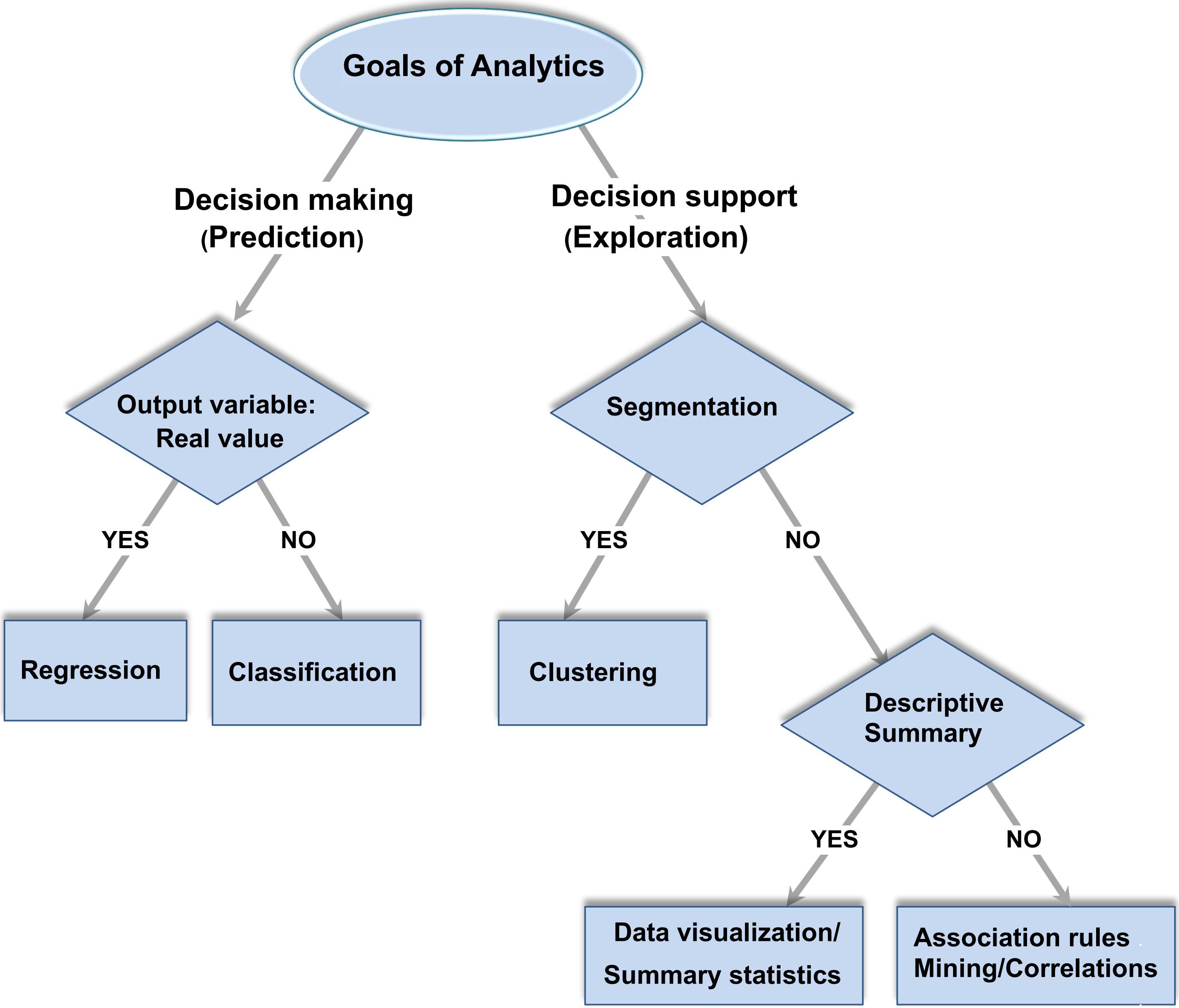


What is Massive Data Analytics?

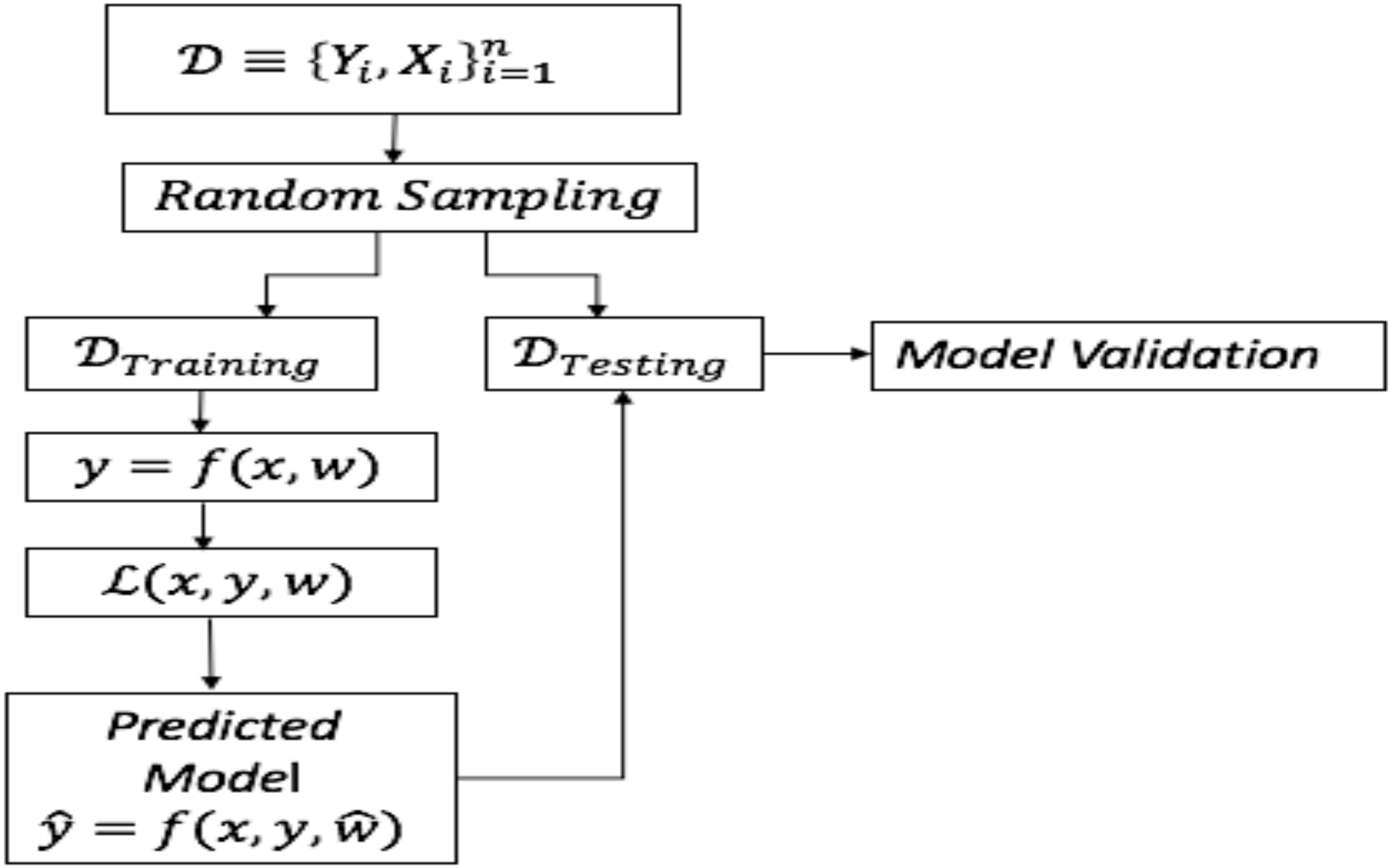
It is a process that involves a sequence of steps to uncover hidden nuggets of knowledge buried in a mountain of data



Mainstay of Analytics



Machine Learning at a Glance



Statistical Clustering

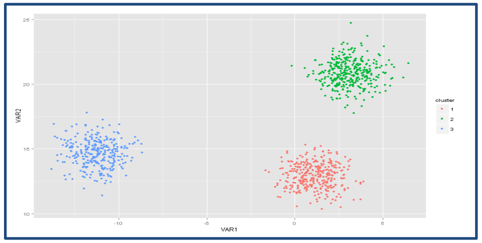
- **A cluster is a collection of data entities that are similar to one another within the same cluster and dissimilar to data in other clusters**
- **Clustering does not rely on pre-defined classes**
- **Clustering can be thought of as an ad hoc procedure (unsupervised learning)**
- **Cluster formation is based on a distance measure (or other metric) where observations are grouped into homogeneous groups**

Desirable Properties

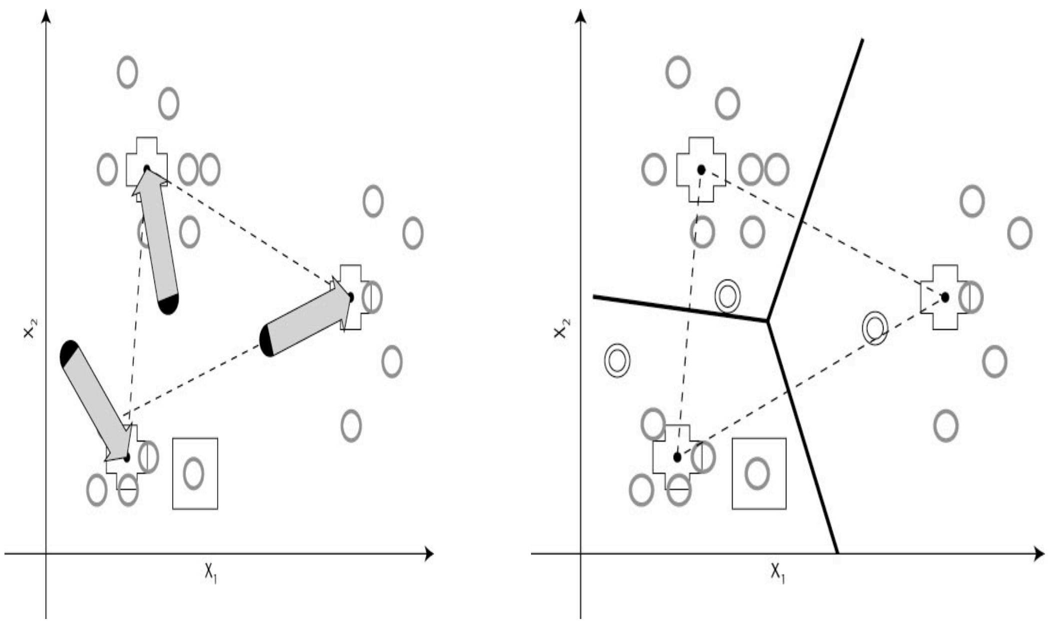
1. **Scalability**: Ability to deal with large data sets
2. **Attributes**: Ability to deal with binary, categorical, ordinal, numeric data
3. **Domain Knowledge**: Minimal requirement of area of application
4. **Noisy Data**: Ability to deal with outliers, and missing values
5. **Order of Records**: Insensitivity to order of records in the data set
6. **Dimensions**: Ability to deal with high dimensional data
7. **Homogeneity:** data within clusters are similar
8. **Heterogeneity:** Data between clusters are distinctly different
9. **Stability:** Cluster(s) formation is insensitive to changes in the data sets
10. **Actionable** : The clusters should be meaningful and meets approval of marketers

Contextualized Data → Clustering Algorithm → Clusters/Groupings

➢ Thus, we see clustering means groupings of data or dividing a large data set into smaller data sets of some similarity.



K-Means Algorithm



Steps involved in *K*-Means algorithm

1. Randomly select *k* data points to be the seeds from the data set
    1. They can be the first *k* records
    2. If the data has some order, choose widely spread records
2. Assign each subsequent record to the closest seed until all records in data set are assigned.
3. Then compute centroids as the average value of each attribute/dimension of all the observations in the cluster.
4. Using centroids as the seeds repeat step 2 onwards until cluster boundaries stop changing.

- The centroids are calculated from the points that are assigned to each cluster
- We use the *k*-means subroutine with (Hartigan-Wong) option
- At each iteration all cluster assignments are re-evaluated. Cluster boundaries are perpendicular bisectors of lines joining centroids
- We compared performance with the FASTCLUS procedure in SAS® for evaluation which is *k*-means (Hartigan-Wong) option

The assignment criterion is:

$$C(i) = \underset{i \leq k \leq K}{\operatorname{argmin}} \|\vec{x}_i - \vec{m}_k\|$$

Stopping Rules for Finding number of Clusters

Create a Raw Data file Consisting of Original Customer Data

Project and Transform the Data into an Attribute Space(Contextual Space)

Apply *K*-means iteratively to determine appropriate number of clusters for Upstream Segmentation
- **We use the SNR (Signal to Noise) Metric with a Stopping Rule Criterion to determine the number of clusters**
- **And exclude the Noise Clusters**

- **At each Iteration, *K*-Means (*k*=1,2,3,...) produces within cluster (Homogeneity) and Between cluster (Heterogeneity) Sum of Squares.**
- **The Within(SSW)+Between (SSB) is the Total Sum of Squares(SST)**
- **We calculate ϕ =1-SSW/SST**
- **We call ϕ the SNR**
- **We measure the Change in Δϕi = ϕi-ϕi-1from Iteration "*i*-1" to Iteration "*i*."**
- **When Δϕi < Median(Δϕi), The Corresponding Cluster Solution is Accepted**

Merge the Results from K-Means with the Original Data by Visitor ID
- **Interpret the Clusters By Customer Attributes to Generate Meaningful/Actionable Customer Segments**

- The **Akaike information criterion (**AIC) is commonly used for model selection
- Its performance is **not satisfactory** in practical applications: generates too many clusters
- (**AIC**) is a measure of relative quality of a *k*-means solution.
- That is, given a collection of clustering solutions, (*k*=1,2,3,…), AIC evaluates the quality of each *k*-means result, relative to the results for other values of *k*.
- So, AIC purportedly provides a means for selecting optimal *k*.
- In our application of AIC produced too many clusters that were not actionable to marketers

Analyzing Throughput of CVICU and NICU

- Problem Statement

- The quantitative analysis of identifying sources of variation and bottlenecks in clinical care in an ICU is known as complex care analytics (CCA)
- Patient level analysis may reveal actionable insights
- An approach to CCA requires studying patient data as a series of events/interventions across ailment categories in an ICU
  - To identify resource utilizations by cohorts of patients
  - Identify top *"k"* ailments by admission rates and analyze variations in medical practice by cost and outcomes

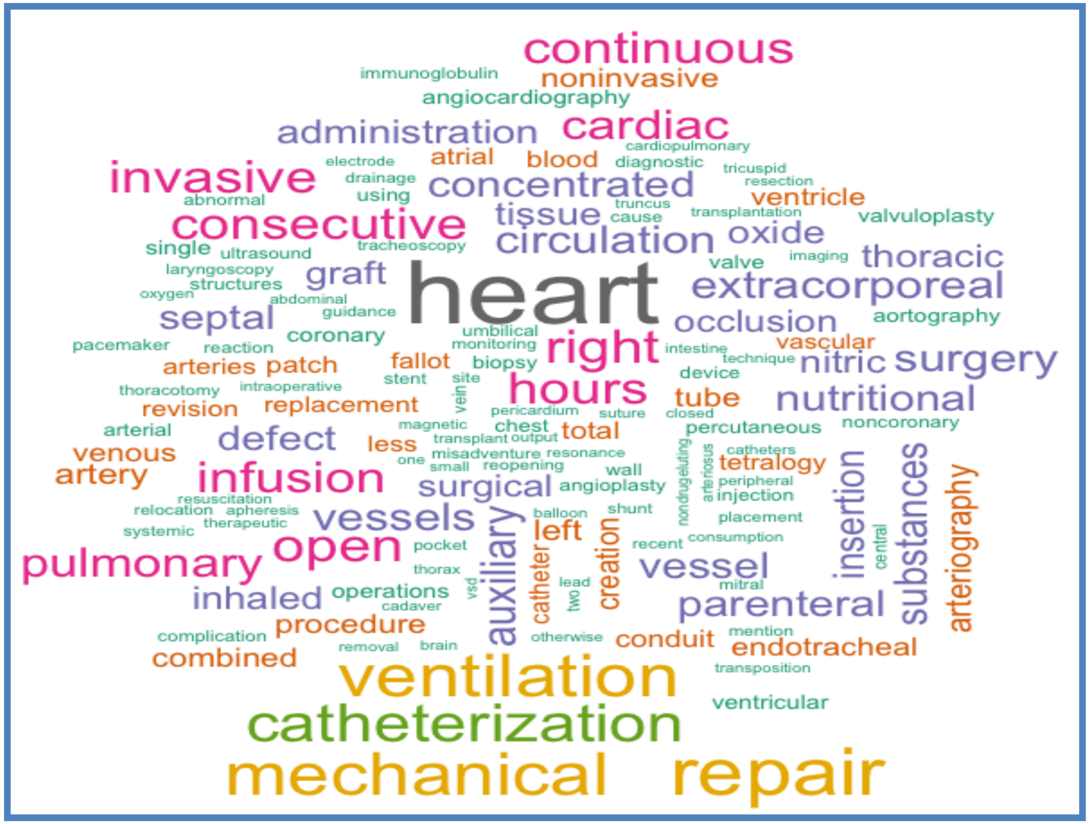
1st Case Study: Segmenting Patients in Intensive Care Units

Data and Methodology

- Patient level data from CVICU, NICU, PICU, Other Units
- Focus of Analysis: **CVICU**
- Analytical tools
    1. Exploratory Data Analysis
    2. Statistical Clustering
    3. Time series analysis
- Interpretation of Results by Domain experts

➢ **Cluster Analysis** revealed patient groups in relation to "resource consumption" and types "types of interventions"

Cluster 1 Interpretation

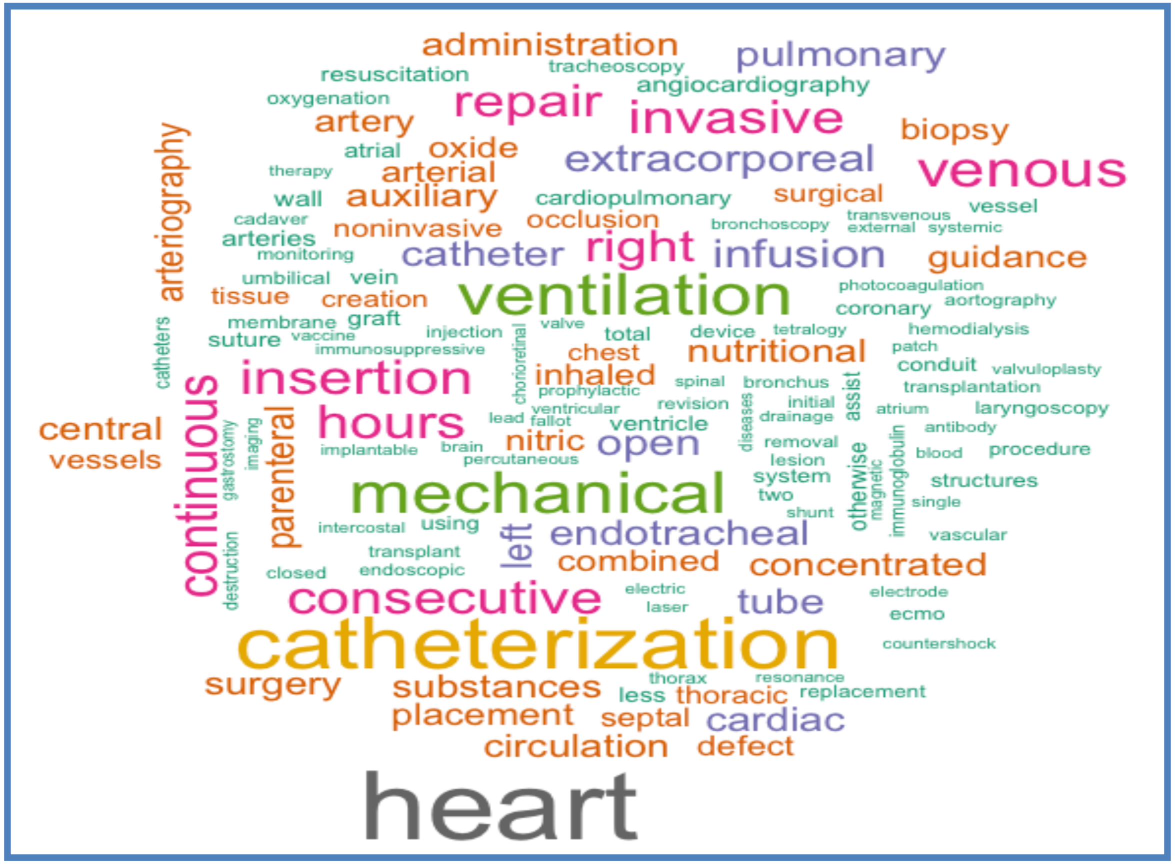


- Patients requiring **Surgical Operations** in CVICU
- Resource Intensive group

- Average Resource Utilization: Hospital resources consumed based on counts

Length of Stay	Radiology	UltraSound	MRI&CT	Blood Bank	Respiratory	Ventilator	Diagnostic Echo	Microbiology	Distinct Pharma	Medications	Distinct Laboratory
12.32	17.86	0.57	0.19	27.43	64.70	0.43	4.03	11.35	39.35	309.73	19.35

Cluster 2 Interpretation



- Patients in CVICU requiring Supportive and Diagnostic Care
- High Resource Consumption Group

- Average Resource Utilization: Hospital resources consumed based on counts

Length of Stay	Radiology	UltraSound	MRI&CT	Blood Bank	Respiratory	Ventilator	Diagnostic Echo	Microbiology	Distinct Pharma	Medications	Distinct Laboratory
27.50	39.37	4.00	0.32	73.79	119.84	1.68	8.89	35.21	63.47	757.11	37.00

Cluster 3 Interpretation

- Very Sick Children in CVICU
  - Birth Defects,
  - Pre-mature Births
- High Resource Consumption Group

- Average Resource Utilization: Hospital resources consumed based on counts

Length of Stay	Radiology	UltraSound	MRI&CT	Blood Bank	Respiratory	Ventilator	Diagnostic Echo	Microbiology	Distinct Pharma	Medications	Distinct Laboratory
6.85	11.45	0.29	0.03	16.74	35.23	0.58	1.65	4.48	30.77	157.32	14.23