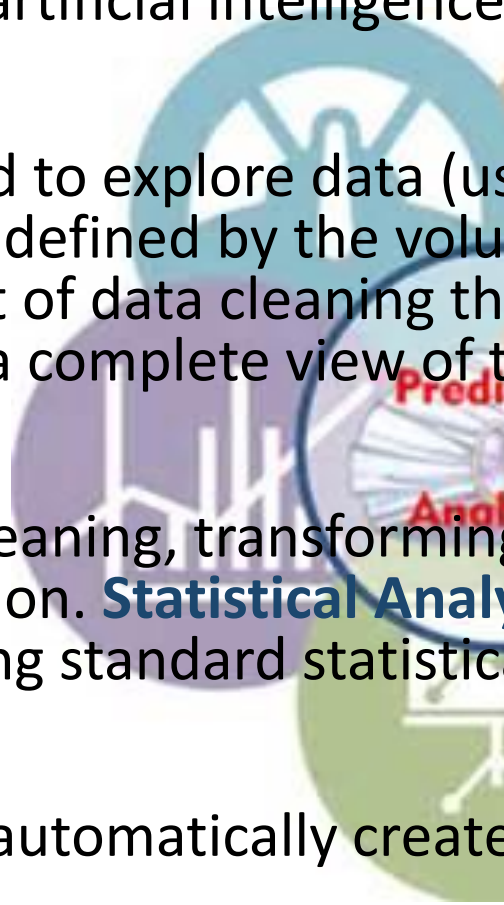
Module 3: Data Science

BeVera

Data Mining? Statistical Modeling? Predictive Analytics?

- **Predictive analytics** is the branch of advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses techniques from data mining, statistics, modeling, machine learning, and artificial intelligence.

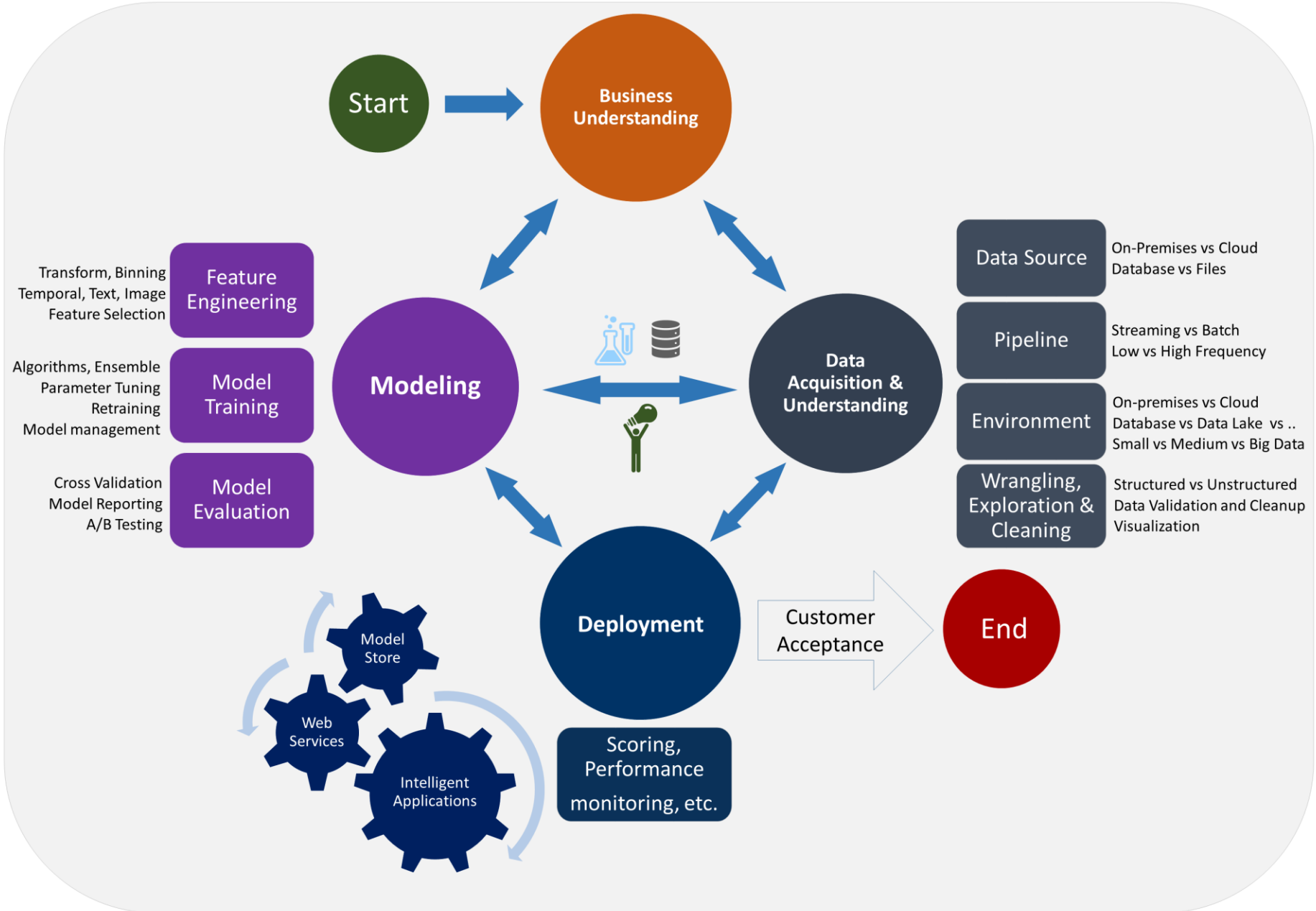
- **Data mining** is an analytics process designed to explore data (usually large amounts of data – also known as “big data”). Big data is defined by the volume, velocity, variety, variability and veracity. Data mining is a part of data cleaning that prepares data from multiple sources for analysis. This provides a complete view of the customer interactions.



- Data analysis is the process of inspecting, cleaning, transforming and modeling data with the objective of discovering useful information. **Statistical Analysis** enables to validate assumptions, hypothesis, and test them using standard statistical models.

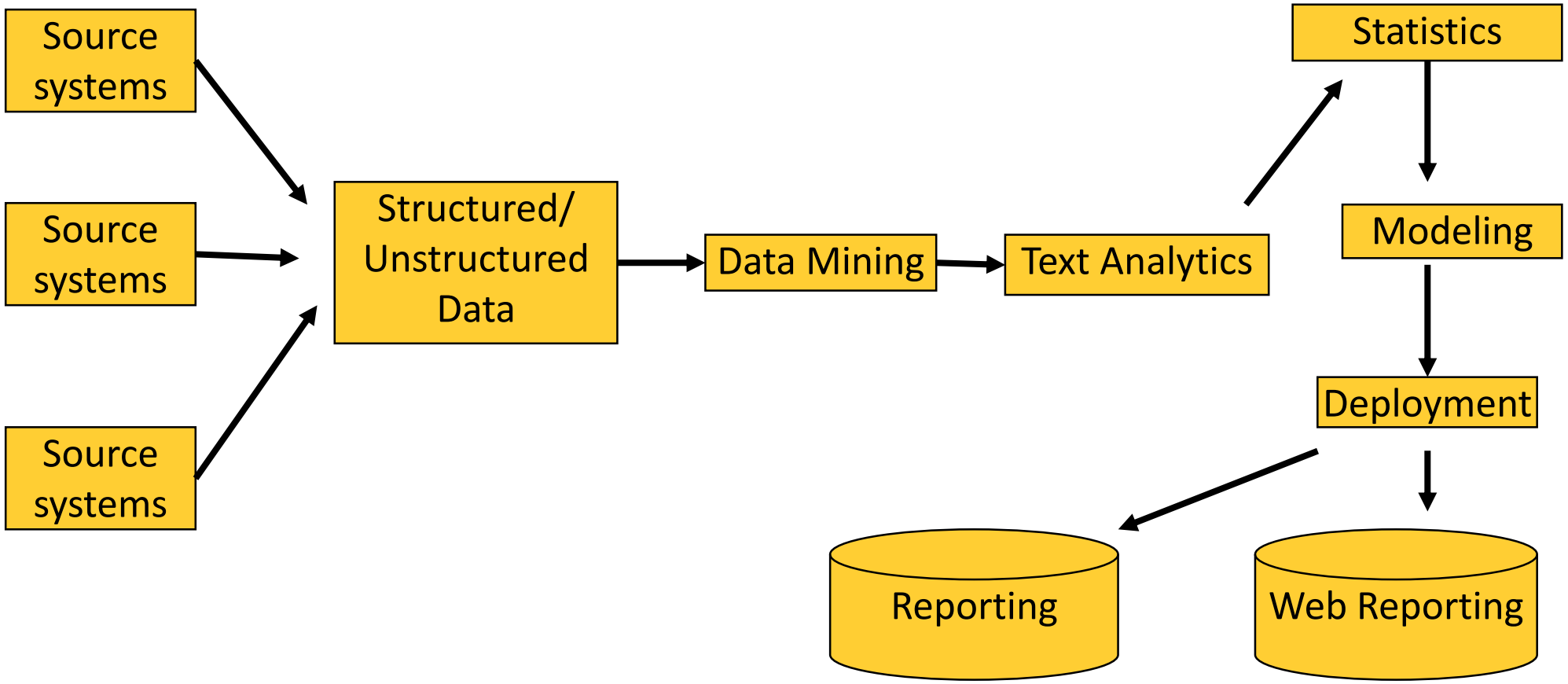
- **Predictive modeling** provides the ability to automatically create accurate predictive models about future.

Data Science Lifecycle



Data Mining? Statistical Modeling? Predictive Analytics?

Predictive Analytics



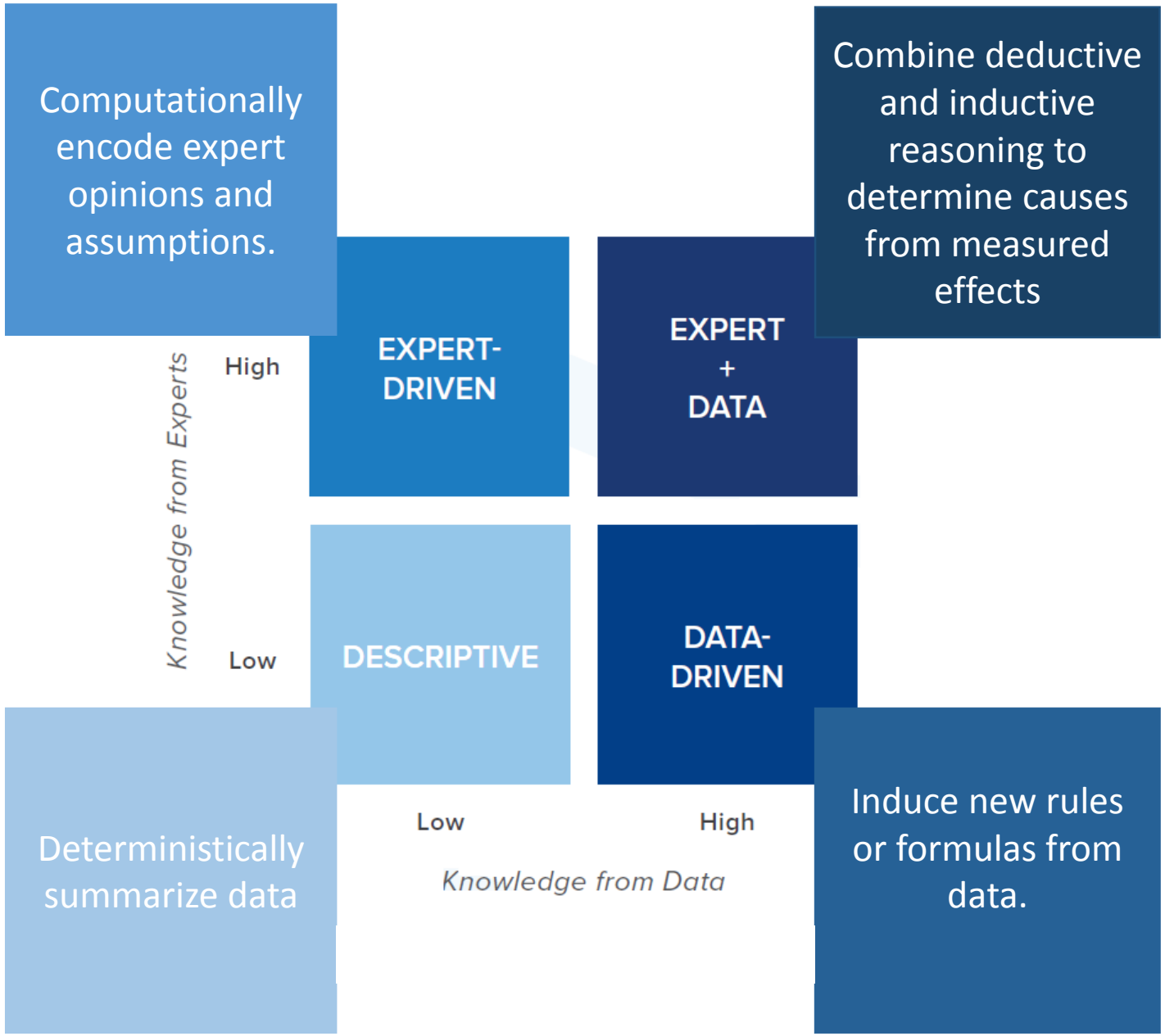
4 Levels of Knowledge Model

https://www.elderresearch.com/consulting-services/ten-levels-of-analytics

Sources of knowledge whether data or expert are independent. Combine data- and expert-driven approaches to maximize insight and value.

Expert-driven approaches is preferred if:

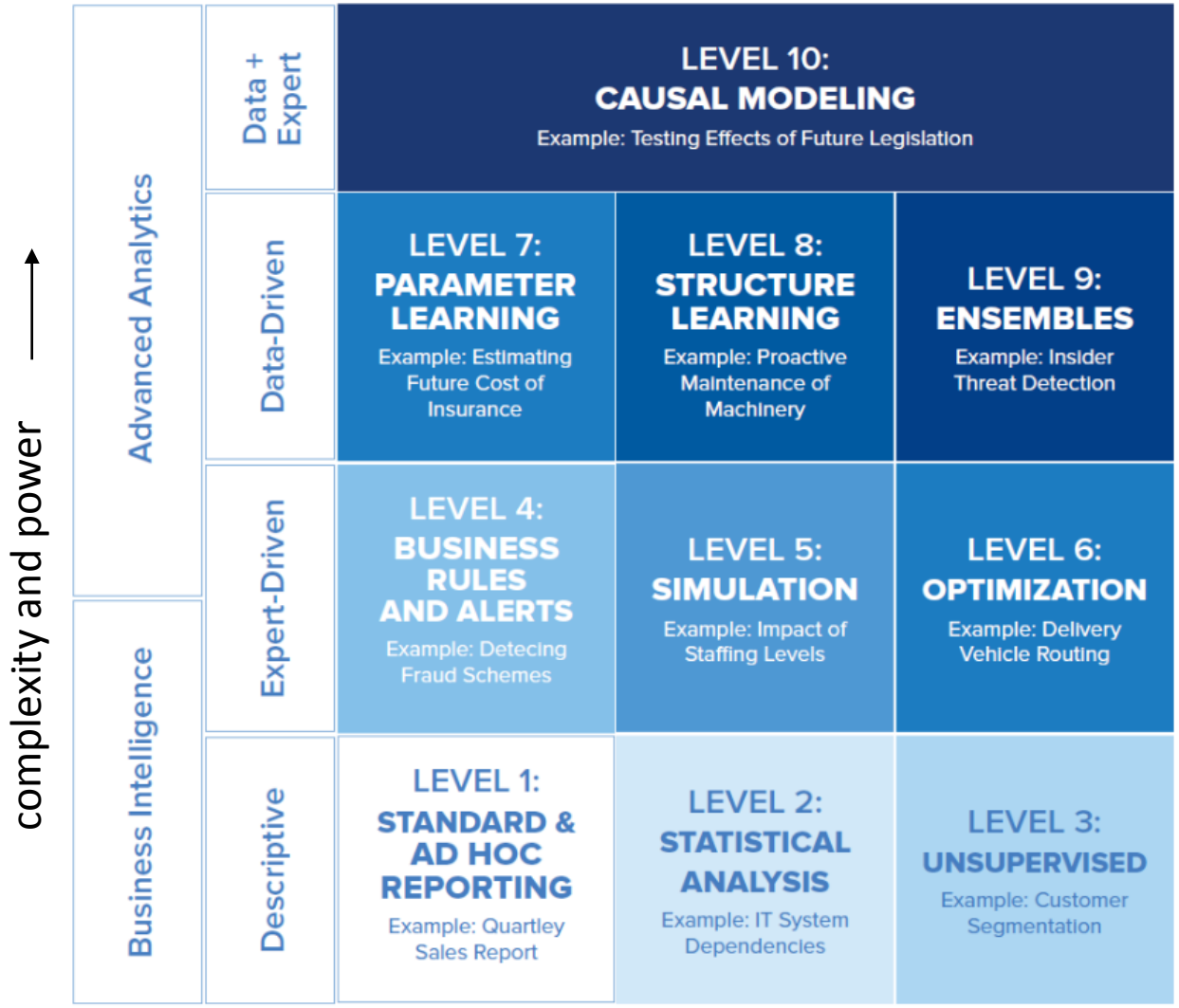
- an accepted premise
- the data is filtered
- poorly represents the full situation



Data –driven modeling:

- is inductive
- graded superior to expert-driven ones.
- allow unknown rules or relationships to be discovered from the data
- are less susceptible to the biases and misconceptions common to human reasoning.

10 Levels of Analytics



https://www.elderresearch.com/consulting-services/ten-levels-of-analytics

Analytic Techniques and their Disciplines

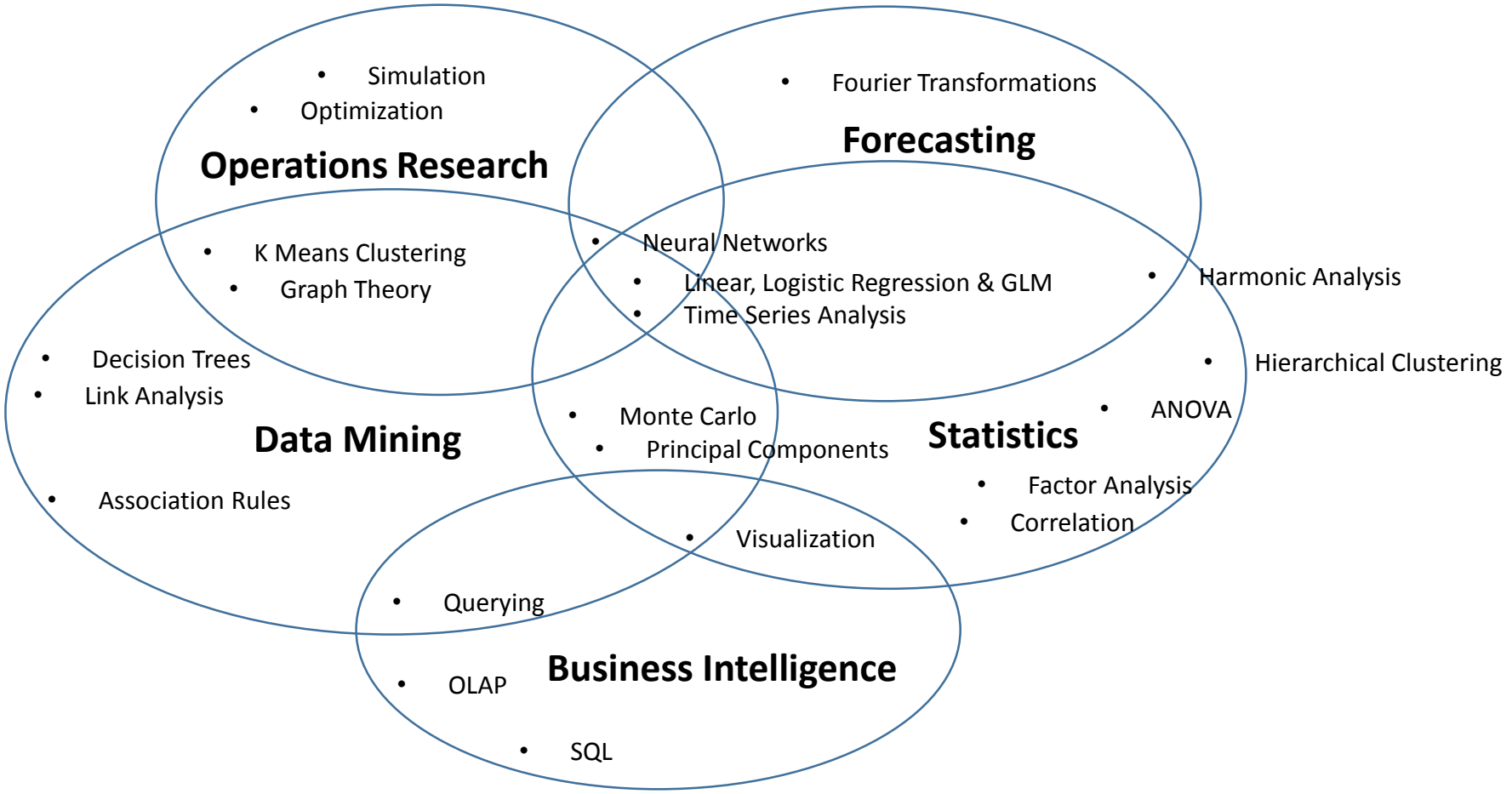
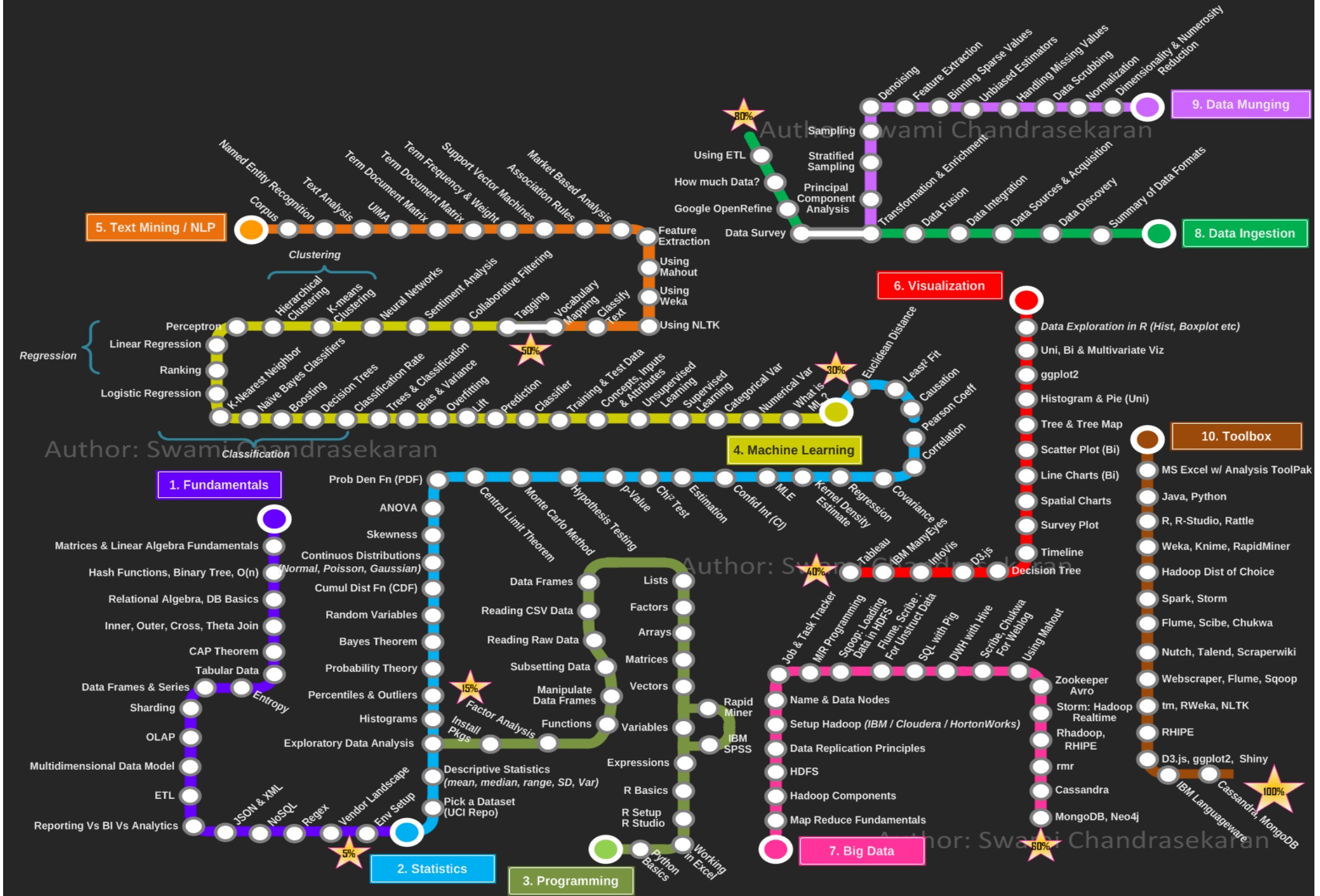


**Diagram developed by John Elder and Dustin Hux*


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Using Mahout
Using Weka
Feature Extraction
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
5. Text Mining / NLP
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
50%
80%
Using ETL
How much Data?
Google OpenRefine
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
9. Data Munging
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
8. Data Ingestion
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran