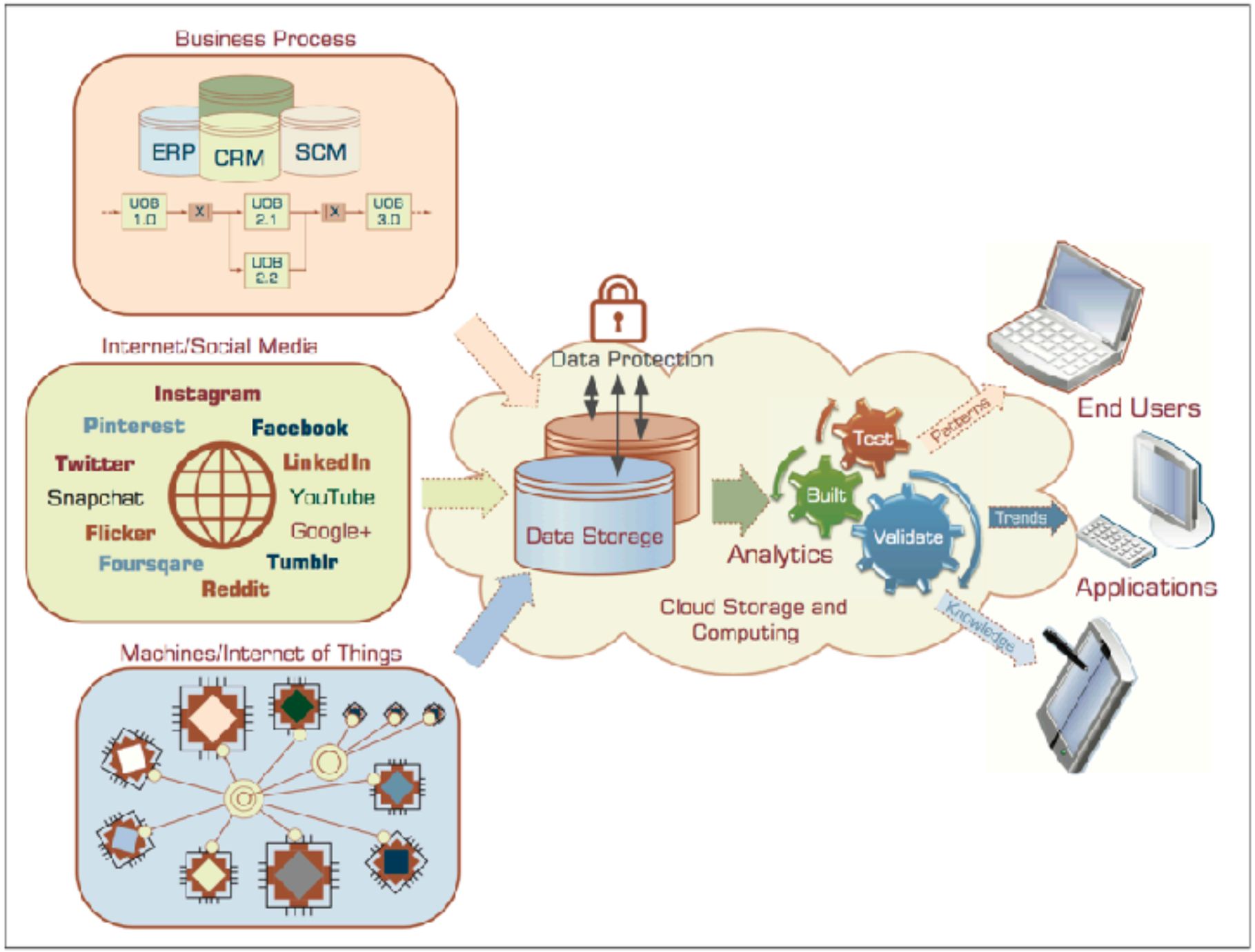
Business Data Analytics 2

Sadi Evren SEKER

Descriptive Analytics



Data Quality

- Data Source Reliability: Do we have the right confidence and belief in this data source?
- Data Content Accuracy: Do we have the right data for the job?
- Data Accessibility: Can we easily get to the data when we need to?
- Data Security and Privacy: Is data secured to only allow those people who have the authority and the need to access it and to prevent anyone else from reaching it
- Data Richness: Means that all the required data elements are included in the data set

Data Quality 2

- Data Consistency: means that the data are accurately collected and combined/ merged.

- Data Currency / Timeliness : means that the data should be up-to-date (or as recent/new as it needs to be) for a given analytics model

- Data Granularity: requires that the variables and data values be defined at the low- est (or as low as required) level of detail for the intended use of the data

- Data Validity: is the term used to describe a match/mismatch between the actual and expected data values of a given variable.

- Data Relevency: means that the variables in the data set are all relevant to the study being conducted.

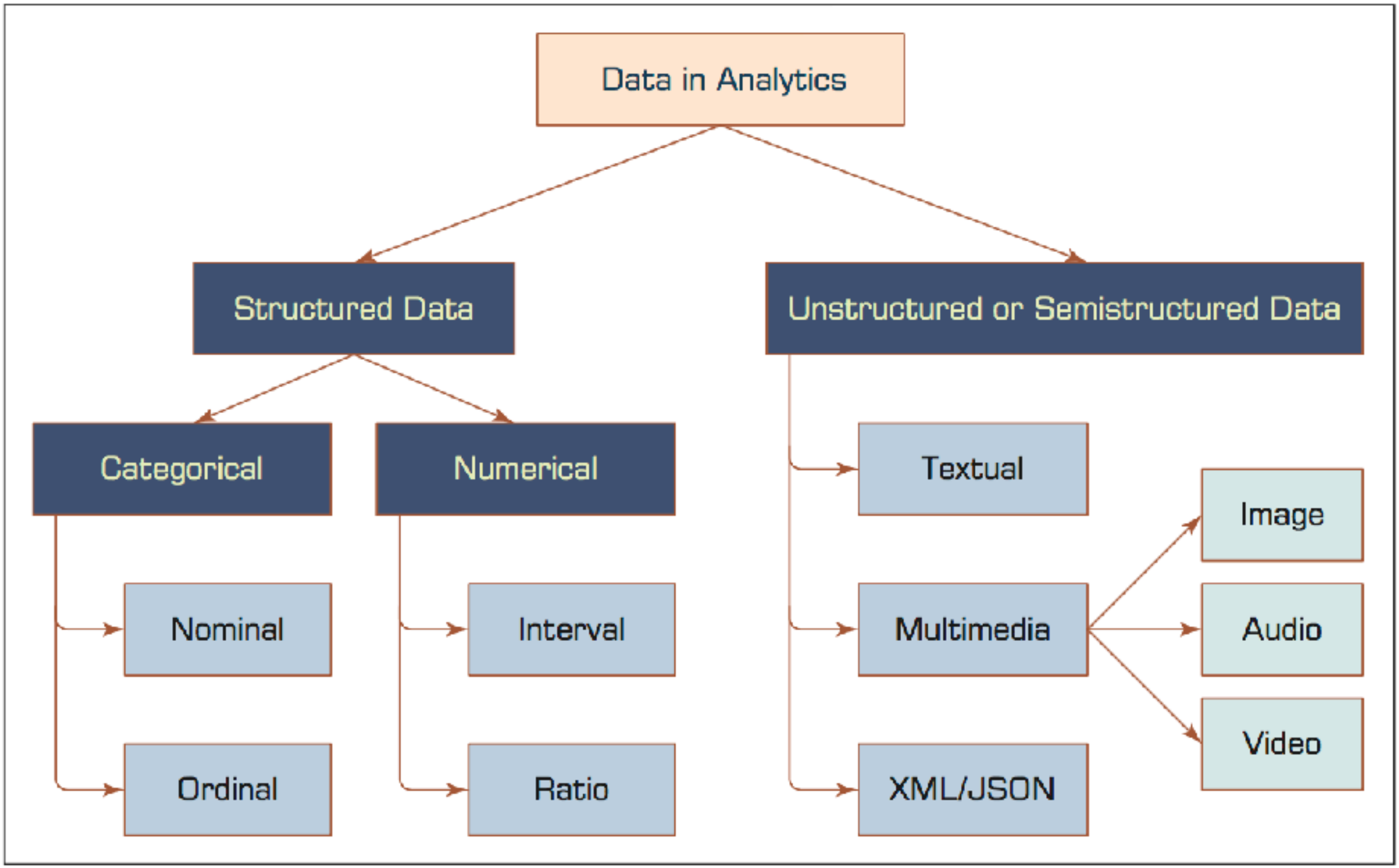
Questions

How do you describe the importance of data in analytics? Can we think of analytics without data?

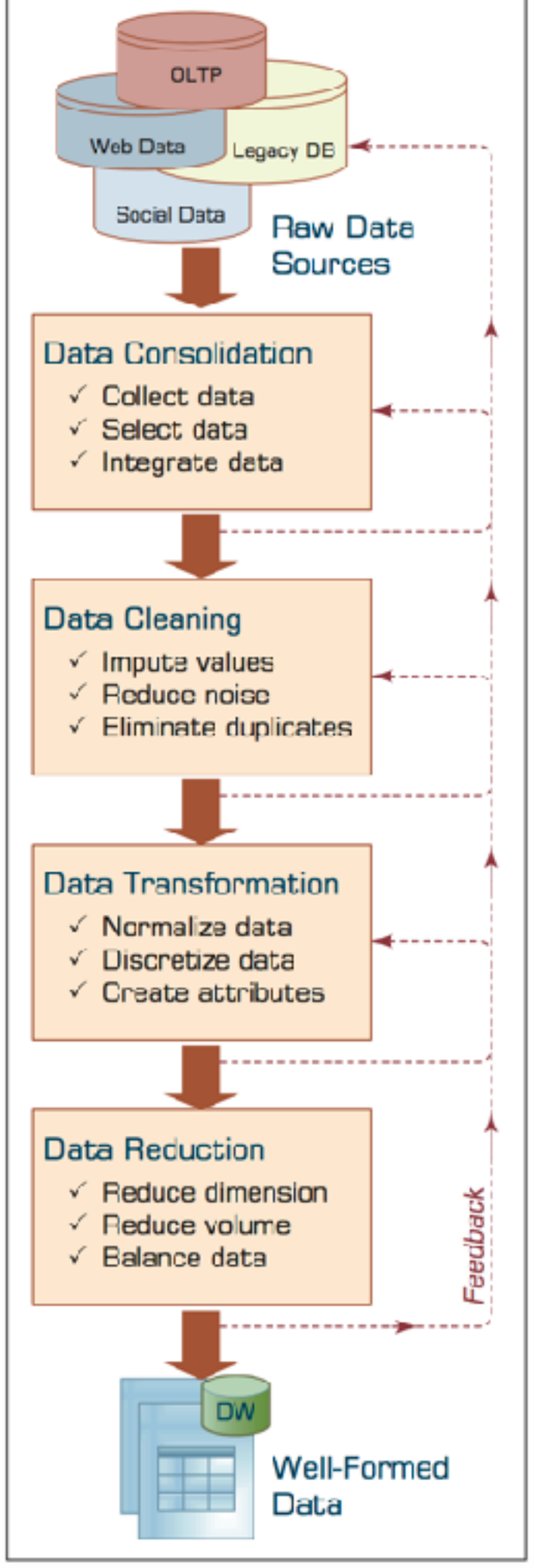
Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?

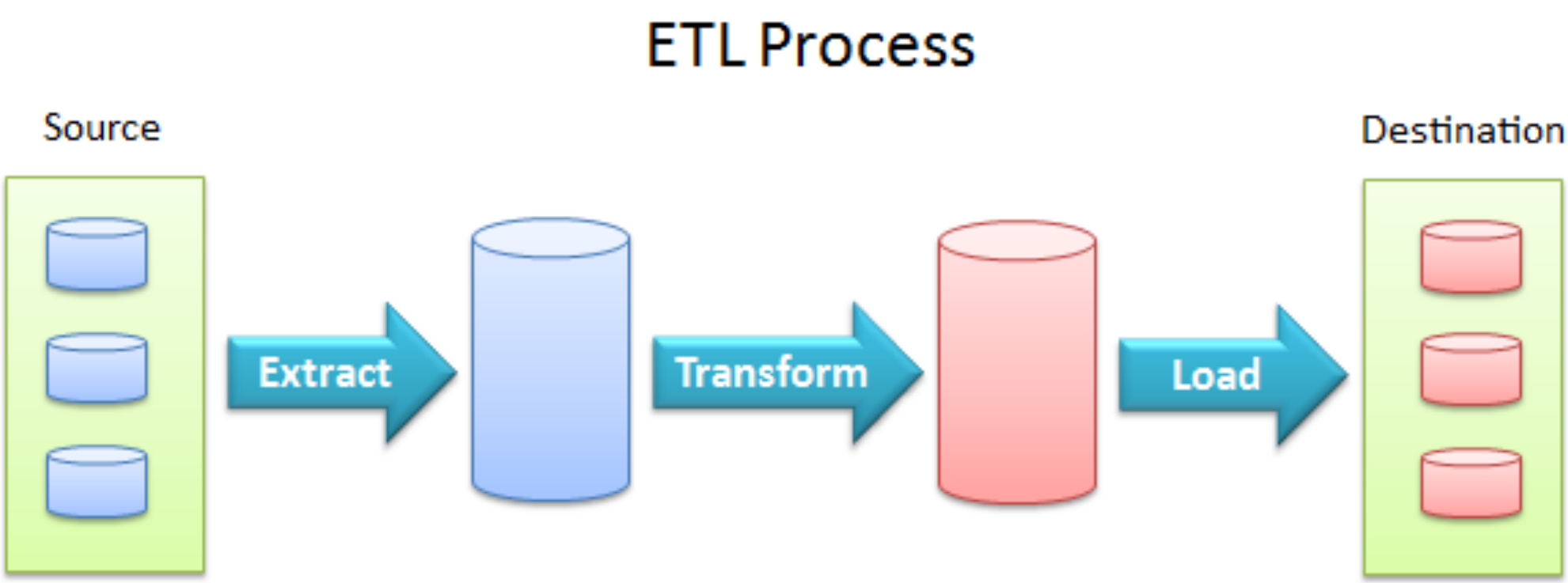
Where does the data for business analytics come from?
In your opinion, what are the top three data-related challenges for better analytics?

What are the most common metrics that make for analytics-ready data?


Data in Analytics
Structured Data
Unstructured or Semistructured Data
Categorical
Numerical
Textual
Image
Nominal
Interval
Multimedia
Audio
Video
Ordinal
Ratio
XML/JSON

PreProcessing

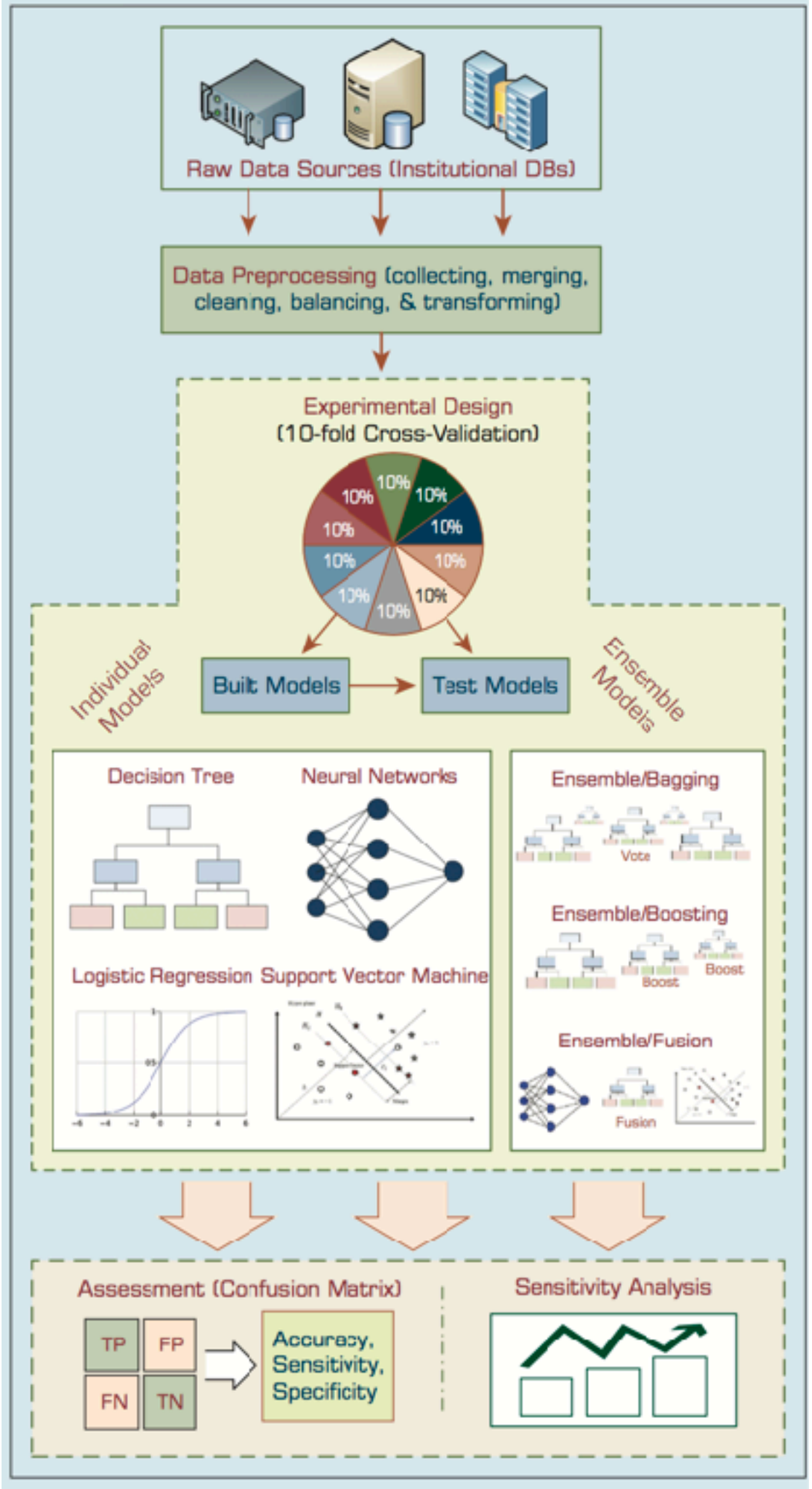



ETL Process
Source
Destination
Extract
Transform
Load

ETL

Extract **T**ransform **L**oad

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.


Raw Data Sources (Institutional DBs)
Data Preprocessing (collecting, merging, cleanng, balancing, & transforming)
Experimental Design
(10-fold Cross-Validation)
10%
10%
10%
10%
10%
10%
10%
10%
10%
10%
Individual Models
Built Models
Test Models
Ensemble Models
Decision Tree
Neural Networks
Logistic Regression
Support Vector Machine
Ensemble/Bagging
Vote
Ensemble/Boosting
Boost
Boost
Boost
Ensemble/Fusion
Fusion
Assessment (Confusion Matrix)
TP
FP
FN
TN
Accuracy,
Sensitivity,
Specificity
Sensitivity Analysis