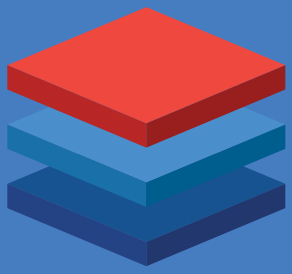
Andrew Fast, Ph.D.
John Elder, Ph.D.



THE TEN LEVELS OF ANALYTICS

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

About Elder Research

Elder Research is a recognized leader in the science, practice, and technology of advanced analytics. We have helped government agencies and Fortune Global 500® companies solve real-world problems across diverse industries. Our areas of expertise include data science, text mining, data visualization, scientific software engineering, and technical teaching. With experience in diverse projects and algorithms, advanced validation techniques, and innovative model combination methods (ensembles), Elder Research can maximize project success to ensure a continued return on analytics investment.

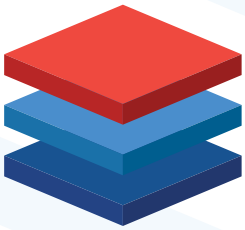
About the Authors

Dr. Andrew Fast, Chief Scientist at Elder Research, provides technical vision and direction to ensure that Elder Research remains at the forefront of analytics practice. Dr. Fast directs the research and development of new tools and algorithms for mining data, text, and networks and has published on an array of applications including detecting securities fraud using the social network among brokers, understanding the structure and behavior of criminal and violent groups, modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (featured on ESPN.com). Written with Dr. John Elder and four others, his Practical Text Mining book won the PROSE Award for top book in the field of Computing and Information Sciences in 2012.

Dr. Fast earned his MS and Ph.D. degrees in Computer Science from the University of Massachusetts Amherst, where he specialized in algorithms for causal data mining, and for analyzing complex relational data such as social networks.

Dr. John Elder, Founder and CEO of Elder Research, leads the largest and most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he's an adjunct professor. He's authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served 5 years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose "book of the year" awards.

Early versions of this work were published by the International Institute for Analytics (iianalytics.com)



THE TEN LEVELS OF ANALYTICS

1.0 INTRODUCTION

Every technical project involves some sort of analytics, ranging from simply reporting key facts, to predicting new events. Here, we define ten increasingly sophisticated levels of analytics so that teams can assess where they stand and to what they aspire. Along the way, we clarify definitions of three types of analytic inquiry and four categories of modeling technology. We illustrate these levels with examples using tabular data representations commonly found in spreadsheets and single database tables.

As our field's ability to collect and fuse data from different sources increases, advanced data types such as time series, spatial data, and graph data are moving into the analytic mainstream. In the second portion of this report, we extend the Levels to encompass these emerging data types, providing data complexity as second dimension for categorization alongside algorithmic sophistication.

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

2.0 ANALYTICS AS INQUIRY

Analytics is about asking questions from data and getting answers. It's what Aristotle and the ancient Greeks called "inquiry," and its two parts are the question being asked and the method of reasoning used to answer the question. The question, or type of inquiry, and the method of investigation, or analytic technique, have often been conflated. Let's clarify their distinctions.

2.1 CHARACTERIZING THE QUESTION

In his *Harvard Business Review* article "Analytics 3.0," Tom Davenport describes three types of analytic inquiry: Descriptive, Predictive and Prescriptive.[1] Similar categories have also appeared elsewhere.[2] Davenport summarizes that Descriptive Analytics report on the past, focusing on what has happened. Predictive Analytics uses past data to predict what might happen in the future. Prescriptive Analytics uses past data to understand what should happen, specifying optimal behaviors and actions for specific situations.

We agree with the definitions, but argue the three types are not parallel and separate categories; instead, Prescriptive describes how a model (of any type) is used. Most Predictive models are also Prescriptive, and some non-Predictive models are Prescriptive as well. All models have the potential to be Prescriptive, which we define as leading to an action without requiring human judgment.

To illustrate, let's look at examples that cover the four cases in Figure 1. Note that the dimensions are Predictive and Prescriptive. The only purely Descriptive analytics are in the non-Predictive and non-Prescriptive category.

ALL MODELS HAVE THE POTENTIAL TO BE PRESCRIPTIVE, WHICH WE DEFINE AS LEADING TO AN ACTION WITHOUT REQUIRING HUMAN JUDGMENT.

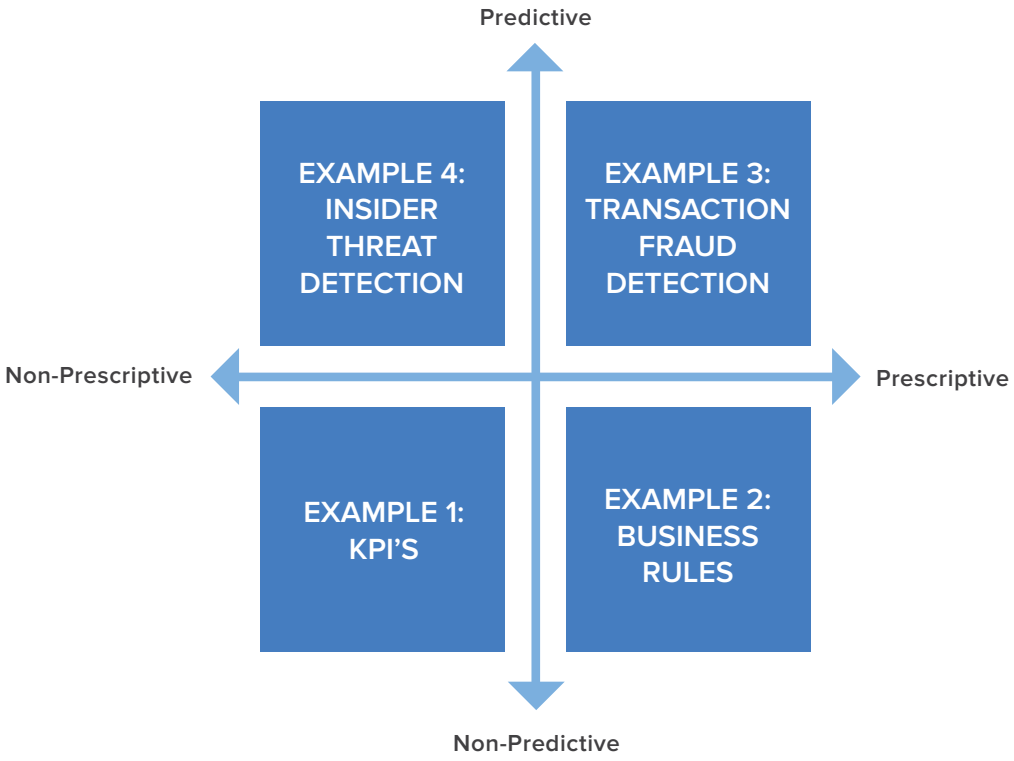


Figure 1: Four Analytic Problems

Example 1: KPIs (Key Performance Indicators): Reporting key metrics, such as the trend in year-over-year sales at each store, is a Descriptive task that can help experienced managers make useful decisions, especially when several KPIs are illustrated on a regularly updated dashboard. Nothing is being predicted or prescribed, but accurate and timely measurements can be very useful. As an engineering maxim goes, "you can't control what you don't measure."

Example 2: Business Rules: Subject Matter Experts (SMEs) may build automatic rules on top of key measurements that kick in when a threshold is met. For example, "If the average customer wait time exceeds 5 minutes" then "alert managers to start taking calls." Nothing is predicted, but automatic actions are initiated directly from key measurements.

Example 3: Transaction Fraud Detection: Fraud detection algorithms for credit card companies have to rapidly judge whether or not to allow a charge. They seek to head off fraud as soon as possible but must be calibrated to not overly disturb customers with inevitable false alarms. They are prescriptive as their predictive judgment immediately turns into action by denying or allowing a purchase. (If it is a false alarm, the customer is inconvenienced but can proceed to connect with humans who can overrule the algorithm).

Example 4: Insider Threat Detection: In large organizations, trusted insiders can intentionally or unintentionally compromise vital information. Their unending traceable activity is too vast and complex for investigators to track with any thoroughness, yet algorithms can continuously monitor staff behavior and predict a breach, or conditions that are ripe for a breach. In a "needle in a haystack" problem like fraud detection, the models can consider (in their limited way) every person, and rank order their risks to guide investigators. The predictions are not prescriptive though, as the judgment of what action, if any, to take – from counseling or reassignment, to investigation or arrest – must be made by experienced humans able to examine and consider the particular details of each case.

2.2 FOUR CATEGORIES OF MODELING TECHNOLOGY

We've described the types of questions answered by analytics; now let's break down the types of analytic inquiry available to answer those questions. Davenport makes the distinction between "Business Intelligence" (BI) and "Advanced Analytics." BI mainly employs descriptive techniques providing decision makers information about what has happened. "Advanced Analytics" uses algorithmic approaches drawn from fields such as data mining, machine learning, and operations research. This is a helpful distinction but does not distinguish between the types of knowledge required to make each type of analytics effective. We divide modeling technology into four categories based on the type of knowledge required for use:

1. **Descriptive –** deterministically summarize data.
2. **Expert-Driven –** computationally encode expert opinions and assumptions.
3. **Data-Driven –** induce new rules or formulas from data.
4. **Data+Expert –** combine deductive and inductive reasoning to determine causes from measured effects.

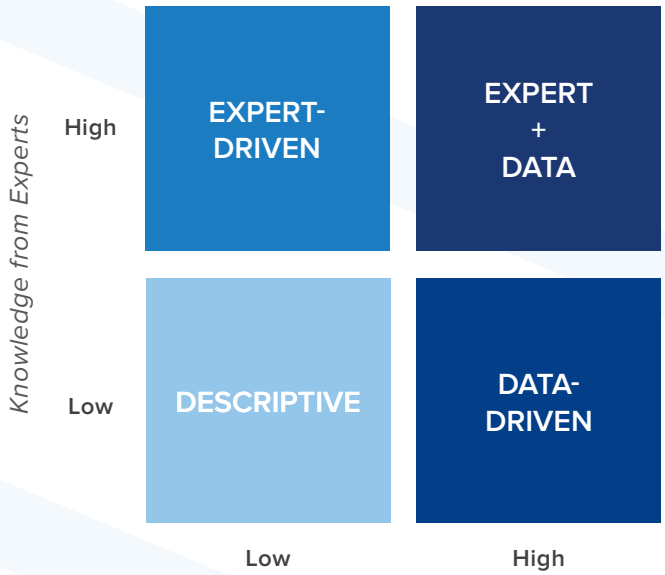


Figure 2: Data-driven vs. Expert-driven Knowledge

Expert-driven modeling is deductive – it reasons from theory to specific cases; data-driven modeling is inductive – it reasons from specific cases (data) to a theory (model). These sources of knowledge – data or expert – are independent; a modeling technology can rely on either or both, to varying degree, as shown in Figure 2.

Though there are potential pitfalls at all levels, we believe the accuracy and quality of the answer improves as you move up the levels. We grade data-driven inductive modeling approaches superior to expert-driven ones, as the inductive techniques allow unknown rules or relationships to be discovered from the data and are less susceptible to the biases and misconceptions common to human reasoning. On the other hand, expert-driven approaches are preferred if the data is filtered or poorly represents the full situation. Best, is to employ analytic approaches which combine both expert-driven and data-driven modeling.

3.0 TEN LEVELS OF ANALYTICS

There are further distinctions within each of the four categories of modeling technology that are useful for applying algorithms to specific problems. We have identified 10 increasingly complex levels of analytics. For each level, we provide an example question and list some of the major techniques employed. Very often, higher levels depend on techniques from lower ones; for example, data-driven analytics techniques often rely on optimization and simulation techniques when learning structure or parameters.

1. Standard and Ad Hoc Reporting
(queries and joins)

Example Question: How much did we sell last quarter by sales region?
Techniques: SQL, OLAP, Excel

2. Statistical Analysis

Example Question: Is the frequency of communication with the customer correlated with satisfaction?
Techniques: Means, standard deviations, correlations, principal component analysis (PCA)

3. Unsupervised Learning *(clustering)*

Example Question: What natural groupings or segments appear in our customer database?
Techniques: K-means, agglomerative models

4. Business Rules and Alerts

Example Question: When do we notify a manager that the system load is dangerously high?
Techniques: Thresholds, JRules, Drools

5. Simulation

Example Question: Would changing our onboarding procedure increase or decrease our average wait time?
Techniques: Monte Carlo simulation, stochastic modeling, agent-based modeling

6. Optimization

Example Question: What number of investigators should we put on each case to maximize expected return?
Techniques: Integer and linear programming

7. Parameter Learning

Example Question: How much additional cost do we expect this account to accrue in the next six months?
Techniques: Linear regression, logistic regression, Markov random fields, neural networks

8. Structure Learning *(automatic variable selection, model search)*

Example Question: How much additional cost do we expect this account to accrue in the next six months?
Techniques: Stepwise regression, decision trees, polynomial networks

9. Ensembles *(multiple models joined together)*

Example Question: How much additional cost do we expect this account to accrue in the next six months?
Techniques: Random forests, bagging, boosting, Bayesian model averaging, model blending

10. Causal Modeling *(a blend of expert assumptions and inductive learning to identify causal relationships from data)*

Example Question: How much of the reduction in fraud can be attributed to a change in investigative procedure?
Techniques: Causal Bayesian networks, Rubin's causal model, quasi-experimental designs

3.0 TEN LEVELS OF ANALYTICS (CONT)

Note that the three data-driven levels (7-9) all have the same example question. While there are infinite questions addressable by these powerful techniques, using the same example emphasize that the levels are fundamentally interchangeable in their goals, albeit not their sophistication.

Figure 3 summarizes the 10 levels, with complexity (and with it, power and danger) rising from the bottom to the top, as well as increasing from left to right. The upward dimension changes when moving from simple Descriptive analytic tasks to more Predictive and Prescriptive, and from BI to Advanced Analytics. The position from left to right reflects the intensity and complexity within a category.

Optimization is shown as the most advanced form of expert-driven technique, as domain knowledge is essential to creating a useful simulation or equation to optimize. But the search for parameter values is usually automated, so it can be considered a transitional form to the next level category that is data-driven.

Each level of the data-driven approaches increases complexity and power over the previous one. Parameter learning employs optimization to find the best parameter values for a fixed model structure. Structure learning performs an additional search over a large set of possible model structures. Ensembles combine multiple models having different strengths and weaknesses into a single model, which is typically more accurate and stable than any of its components. This combination of strengths makes ensemble methods the highest form of data-driven modeling.

Causal modeling draws from both data-driven and expert-driven techniques. It is like an automatic scientist using theory and data to refine a hypothesis. Expert-driven modeling depends on the expert knowing the cause, and data-driven modeling can reveal a possible cause, but only causal modeling can confirm a cause-and-effect relationship by combining both forms of knowledge to rule out alternatives.

ENSEMBLES COMBINE MULTIPLE MODELS HAVING DIFFERENT STRENGTHS AND WEAKNESSES INTO A SINGLE MODEL, WHICH IS TYPICALLY MORE ACCURATE AND STABLE THAN ANY OF ITS COMPONENTS.

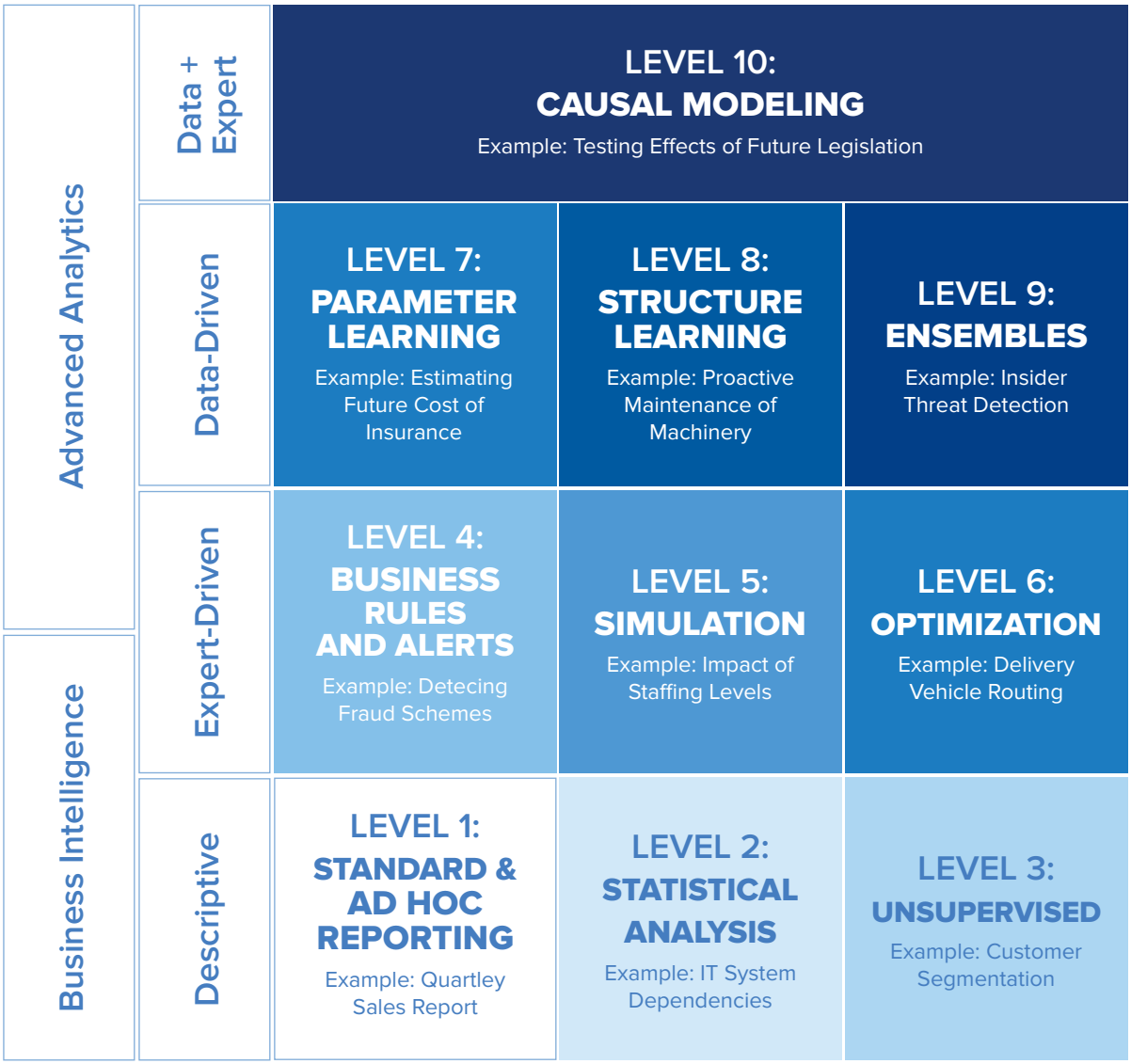


Figure 3: Ten Levels of Analytics

4.0 ADVANCED DATA TYPES

The past of predictive analytics has focused on algorithms, but its future will focus on data types. As organizations increase their reliance on analytics, the problems they face become harder, requiring more data of diverse data types for success. To solve these hard tasks, we must detect weak signals in the data by finding and extracting as much structure (or dependencies) as possible. Reducing complex data to a table will often obscure some of its value.

As our ability to collect and fuse data from different sources increases, advanced data types – with temporal, spatial, or link structure – are now moving into the analytic mainstream. We categorize techniques for handling data from text, sequences, time-series, space, and graphs. Though we use standard definitions of these data types, the lines between categories can be blurry as most data can be represented multiple ways. We will now briefly describe each data type and its major variations.

4.1 TEXT DATA

For most text analytics approaches to be applicable, a collection of text documents (a corpus) must be transformed into a numerical format. The most popular transformation, known as a "Bag of Words," creates a (huge) table where each row is a document, and each column is a word. The table entry notes the existence (or the total count) of a word in a document. Alternatively, text can be treated as a sequence where words are the focus and are viewed in the context of surrounding words (for example: defining a named entity as a string of proper nouns in a row; e.g., John Wayne).

Text data comes in two types: unstructured and semi-structured. Most free-form text, including letters, reports and web pages, are unstructured. Structured information enters from known context, such as an application form with text fields holding answers to particular questions. The added structure information provides tremendous value to analytics teams such as a way to add context with the "notes" fields in a record, or validate sentiment analysis by combing free form survey responses with numerical ratings.

4.2 SEQUENCE DATA

Sequence data can be ordered, but may not be time-based (such as gene sequences) or have a time stamp (such as the string of commands used in a software package). This data could be numeric or symbolic.

4.3 TEMPORAL DATA

Temporal data consists of measurements (readings/samples) with a time stamp. The samples can be taken at regular intervals (time-series data) or be based on events that occur at irregular intervals. Temporal data has much in common with sequence data and many techniques are applicable to both. A big factor with both is autocorrelation, where measurements at consecutive positions or time points are often very related. "Normal" table-based analytic techniques instead treat all cases as independent, so attention to this characteristic is vital for accurate models.

4.4 SPATIAL DATA

Spatial data consists of measurements indexed by x and y coordinates (x,y,z if 3-dimensional). Spatio-temporal data further adds a time dimension to this. Spatial data comes in three flavors, all characterized by spatial autocorrelation where nearby areas typically have very similar measurement values. The first is known as continuous spatial data or geostatistical data. It is analogous to time series in the spatial dimension, and consists of numerical measurements on a regular grid in space. A second focuses on spatial point processes where the location is a random variable, and not regularly-spaced. The third flavor of spatial data is on a lattice where regions of space border other regions (for example, counties in a state). This third type of spatial data is very closely related to graph data.

4.5 GRAPH DATA

In its simplest form, a graph consists of nodes (vertices) and edges (links, relations). This can be expanded to include types and/or attributes on both nodes and edges. Hypergraphs contain edges that connect more than two vertices. Graph data is equivalent to a relational database with multiple tables and foreign key relationships. It is very general, and many of the other advanced data types could be expressed as a graph. For example, sequence data could be represented by observations with foreign key relationships to the preceding and following events. For analytic productivity, we recommend using the tightest typing possible (e.g., sequence, instead of general graph) to take advantage of the structure of that type.

5.0 TECHNIQUES FOR ADVANCED DATA

There are two main ways to handle advanced data types: 1) reduce the complexity of the data so that a simpler technique can be used, or 2) create a custom technique tailored to the complex data type. The first method, which can be thought of as projecting the data down onto a workable space (with some loss of information), is the most common. With heterogeneous data, including graph, temporal, or text data, it is common to transform data into a table using simple aggregations such as COUNT, MIN, MAX, or AVERAGE. For example, the "Bag of Words" text transformation represents each document by a vector of counts for words appearing in the document. Surprisingly, these simple transformations are often effective in extracting useful information, perhaps because the new data sources are only recently being explored and they hold "low-hanging fruit."

New techniques for advanced data-types are often generalizations of techniques for tabular data and are motivated by capturing an element of the data that is missed by a simpler approach. For example, spatial regression adds a term to account for spatial correlation. If no such correlation is in evidence, the term drops out and the technique defaults to standard linear regression. Since advanced techniques can use a new signal in the data, they are preferable to reduction techniques that "lose" or "ignore" information, and they can extract more value out of analytics. However, with these advanced models, as with any more flexible modeling technique, there is always the danger of "over-fitting" to noise in the data leading to suboptimal performance. Care must be taken to ensure that each algorithm, no matter the sophistication, is properly evaluated.

With that as background, we assess the level of analytics currently being used with advanced data and suggest ways to move to higher levels. Moving higher, whether for tabular or advanced data, typically requires a tighter problem definition but produces improved results. We review the characteristics of each level and highlight techniques for different data types there, recognizing that many techniques can be applied to multiple types of data. Our list is illustrative and not exhaustive; there are simply too many specialty techniques to even be aware of them all.

Level 1: Standard and Ad Hoc Reporting

This level is dominated by counting cases after joining and selecting relevant data. The most common visualization of tabular data is histograms, which are applicable across all types of data. Specialized forms of histograms for advanced data include:

- **Table –** Histogram
- **Text –** Word Clouds
- **Spatial –** Chloropleth/Heat Maps
- **Graph –** Degree Distributions

Graph Link Analysis is a type of join and visualization for visual pathfinding in graph data. (Not to be confused with Text Link Analysis, which focuses on filling in factual templates of events).

Example: "Word clouds" are a popular way to quickly explore and explain textual data. Figure 4 is a word cloud created from customer responses to a satisfaction survey.

Figure 4: *A word cloud demonstrating customer satisfaction with an insurance company. Larger words are more frequent. (Orientation has no underlying meaning, and just improves presentation.)*

Level 2: Statistical Analysis

Statistical analysis applies mathematical metrics to data, including means and standard deviations, without requiring parameter fitting or model search. Many standard techniques such as correlations are applicable for every data type; however, there are specialized techniques that go further.

- Table – Correlation, Principle Components
- Text – Term Frequency – Inverse Document Frequency (TF-IDF), Singular Value Decomposition
- Sequence – Sequence Alignment
- Temporal – Autocorrelation, Trend Estimation
- Spatial – Edge Detection (Image)
- Graph – Centrality Measures, Social Network Analysis

Example: Modeling IT dependencies using network centrality measures. Customer-facing software applications rely on a collection of many different middleware applications, data sources, and other software. Each component can feed multiple applications. Centrality measures computed from the dependency graph can be used to highlight which components are most vital and therefore require special uptime requirements or maintenance handling.

Level 3: Unsupervised Learning

Unsupervised techniques are algorithmic in nature, involving search or parameter fitting, but do not require historical data to fit the parameters. This is often called clustering and many traditional clustering algorithms can be applied to advanced data types through a data transformation such as aggregation.

- Text – Topic Models
- Sequence – Sequence Analysis, Association Rules
- Graph – Community Detection, Graph Clustering

Example: Cluster analysis for major computer assisted drafting software. Software companies like to understand how their users are actually navigating through their software. Developers can then analyze those sequences to understand their classes of users and design modules and communication to better meet their needs. Similarly, crash reports containing a sequence of commands that led to the crash can reveal patterns leading to improvements.

Level 4: Business Rules and Alerts

Business rules capture expert judgment to define thresholds and interactions between variables. These rules are typically built or supported by the results of Statistical Analysis and Reporting. Very similar techniques are used across all data types, though adapted to fit the peculiarities of each. For example, an entire sub-area of text analytics has focused on processing text with business rules. Practitioners with a background in computational linguistics have developed a paradigm of automated text processing using linguistic rules, which are a natural extension of the grammatical rules underlying most languages. While many of the techniques are similar for other data types, the application of linguistic rules has been instrumental in the growth of text analytics.

Example: Contract Fraud prevention. Business rules and alerts are often used to signal patterns consistent with known fraud schemes, such as bid-rigging. For example, sometimes a large contract modification very soon after award is a sign of a *quid pro quo* scheme.

Level 5: Simulation

Similar to Business Rules and Alerts, many simulation techniques are applicable across all data types. These techniques include agent-based modeling, Monte Carlo methods, Cellular Automata, and Complex Systems modeling. Many simulation techniques were developed specifically to account for the richness of the real world and can be limited to work on simpler data types by producing simpler scenarios. Application to advanced data requires a skilled practitioner to frame the problem appropriately for the techniques.

Example: Simulation techniques can be used to project the impact on service responses of staffing changes for a large IT call center. One can run thousands of scenarios, based on historical information on trouble-ticket arrival rates, distributions for the times needed to handle different types of tickets, and the likelihoods that agents will pass tickets to others, to assess the impact of changes on quality and timeliness, and thereby make optimal decisions. Note that this example has both a temporal and network dimension.

Level 6: Optimization

Optimization techniques, including linear programming and quadratic programming, require one to specify constraints, resources, criteria of merit, goals, and the problem definition. Many of these techniques already account for temporal and spatial dimensions, as these were part of the original problems being solved. A number of dynamic programming techniques can be used to perform optimization such as Maximum Flow computation and Path Finding in a network. In special, albeit important, situations, such as all constraints and goals being linear, the optimization can be perfectly solved. But in general, even the very sophisticated global search algorithms provide a solution that is only probabilistically optimal, given realistic time constraints.

Example: Finding the most efficient method for distributing and routing vehicles to cover all of that day's deliveries in the most efficient way possible. This requires temporal and spatial data.

Level 7: Parameter Learning

Parameter learning uses historical data to fit a model that has been pre-selected by the user or by virtue of the method being used. For example, linear models assume there is a linear relationship between the output and the input features while the inputs themselves are selected by the analyst. There are many parameter-learning techniques across all data types, as parameter learning represents a balance of expert influence, ease of use, and data-driven reasoning.

- **Text –** Naïve Bayes Classifier, Conditional Random Fields
- **Sequence –** Hidden Markov Models, Conditional Random Fields
- **Temporal –** Forecasting (ARIMA, ARMA, etc.)
- **Spatial –** Markov Random Fields, Spatial Lag Models, Kriging
- **Graph –** Relational Markov Models, Statistical Relational Learning

Example: Entity Extraction from Text. Entity extraction models are used to identify proper nouns, such as person names, company names, or locations in text. These models typically treat text as a sequence and look for leading indicators that the words correspond to a named entity. For example, the words "Mr.", "Miss", "Ms." or "Mrs." typically precede a person name.

Level 8: Structure Learning

Structure learning automatically searches for the inputs and model structure; it puts a loop around traditional parameter learning approaches, hypothesizing different structures and/or inputs for each iteration. For example, stepwise regression adds and removes inputs to a traditional linear regression model, iteratively building the model until further changes stop helping. Also, decision trees are very popular for advanced data types and their construction occurs by continually redefining the problem, in particular, the data over which optimization is done.

That is, they recursively partition the data and select variables and thresholds appropriate for the subset of the cases on which the current partition is focused.

- **Spatial –** Exploratory Regression (step-wise spatial regression)
- **Graph –** Relational Dependency Networks, Statistical Relational Learning

Example: Forecasting which natural gas wells will need maintenance very soon. Features of wells and their conditions at a given time include weather metrics, spatial location, and streams of temperature and pressure information for each pipeline on the surface or underground. The spatial distribution of "frozen" wells can reveal that the relative productivity of a well is tied to its location. Some maintenance needs can be explained by surface conditions such as temperature or precipitation unrelated to underground conditions.

Level 9: Ensembles

The idea of combining multiple models into an ensemble[3] is not limited to tabular data and many of the modeling techniques in Level 7 can be run and combined into a single model for advanced data types. Ensemble model techniques such as random forests and gradient boosting can also be extended for advanced data types.

Example: Predicting upcoming liabilities for Auto Insurance. Auto accidents have a seasonal pattern, and weather plays a significant role, but other elements — holidays, school schedules, time of day, changes in traffic flow due to construction, relative amount of sun glare, etc. — also have an effect. Ensemble models can be used to model each of these components separately, then combined to perform better than each individually.

Level 10: Causal Modeling

Causal modeling is the most difficult level to extend to advanced data types because causal modeling requires reasoning about possible alternative causes, and richer data means more possible causes. For tabular data, techniques such as the Rubin Causal Model and Causal Bayesian Networks are proving useful. For sequence and temporal data, a type of weak causality known as Grainger Causality is often used. For sequence, temporal, spatial, and graph data, there are techniques for detecting quasi-experimental designs from data. Quasi-experimental designs are "natural experiments" where the data themselves provide some (but not all) of the structure of a full experiment.

Example: Determining the effect of new traffic laws. Since states adopt different safety regulations at different times, they can serve as ideal tests of the efficacy of those regulations. For example, the results from states that require the use of seat belts or limit the use of cell phones (or soon, allow autonomously-driven vehicles) can be compared to the statistics of surrounding states and their earlier selves, to determine the net effect of the new regulation.

6.0 CONCLUSION

The science of "inquiry" is composed of the question asked and the solution technique employed to answer it. We've categorized analytic questions in terms of their predictive and/or prescriptive nature. We also described four categories of modeling technologies, distinguishing the expert-driven from the data-driven. We then combined those perspectives to define 10 levels of analytics in terms of the questions addressed and techniques employed.

The levels apply to tabular data, but also to emerging and more challenging data sources, such as text, space, sequences, time-series, and graphs. One can often extract valuable results by projecting down the new data into a tabular format and employing conventional techniques. Still, the greatest returns come from customizing analysis algorithms to the inherent structure of the data source. We have briefly categorized ways this is already done, and we anticipate great advances from the furtherance of this approach in the near future.

This framework can be used to help teams understand where they are operating and what approaches are best for the analytic question and challenge they are facing. As one moves up the levels, the complexity and power of analytics increases. This makes possible greater accuracy when it's done right, but more danger if done wrong![4] Successful analytics teams know how to combine data- and expert-driven approaches to maximize insight and value.

Footnotes

1. Thomas Davenport, "Analytics 3.0," Harvard Business Review, December 2013. (https://hbr.org/2013/12/analytics-30)

2. http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279] Also see: http://www.predictiveanalyticsworld.com/patimes/prescriptive-versus-predictive-analytics-distinction-without-difference/

3. See for instance Seni & Elder (2010) Ensemble Methods in Data Mining: Improving accuracy through combining predictions, Morgan & Claypool; www.tinyurl.com/book2ERI

4. See "Top 10 Data Mining Mistakes," Chapter 20 in Handbook of Statistical Analysis & Data Mining Applications, by R. Nisbet, J. Elder and G. Miner, Academic Press (Elsevier), 2009. www.tinyurl.com/bookERI

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Friendship Landing Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673