In this lecture you will see that data is more than factual information. You will examine the importance of data analysis for business value and how data can address common business needs. And you will explore how to define a problem statement from those needs using a formalized approach. Then you will you will walk through developing an analytics rational statement from a problem statement using key metrics. You will explore how you can use these metrics to refine problem statements and how to refine these problem statements and analytical rationale statements so they align with analytics solutions. Finally,  you will you will examine the process for identifying data requirements and the needed data sources to solve your analytics problem statement using data analysis. You will see how to create data assumptions and constraints to guide your analysis along with defining a methodology with testing procedures to develop your problem framework and action plan.

By the end of this lesson you will have demonstrated the ability to:

* Explain the process in identifying and assessing business needs and issues (i.e. usability requirements).
* Create a business and analytics problem statement from a business need and/or issue.
* Define an initial set of business benefits for data analysis.
* Develop an analytics rationale statement based on a business problem statement.
* Define the process in identifying key metrics for solving business and analytics problem statements.
* Refine the problem statement to ensure alignment with the identified analytics solution.
* Identify the data requirements and related data sources to assist in solving the analytics problem statement.
* Create a set of data assumptions and constraints required for data analysis.
* Define the methodology and testing procedures required to develop the problem framework and action plan.

**Importance of Data**

[Importance of Data](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774451_1)

|  |
| --- |
| While listening, consider the following:   * What is data? * Why is data important? * How would you use and manage data? |

**What is Data?**

*When you hear data, what do you think of?*

Data is the electronic found in computers and the non-electronic information stored on paper and in people's heads. Data does not include just observable facts, but also the interpretation of observations that people remember and believe. For example, data includes the temperature observed on a thermometer saying 23 degrees Celsius. Data includes recording the temperature someone feels it is without using a thermometer. Data includes recording what someone thinks the temperature was yesterday. All are data – and it is these differences in data that makes data analysis complicated.

Data can be aggregated, analyzed, and used to make a profit, improve health, or influence public policy. And despite the volume and velocity of data, useful information is synthesized from data with the increasing technological capacity for collecting, storing, and analyzing the electronic versions of videos, pictures, sounds, and documents without being overwhelmed.

**Business Importance of Data Analysis**

There are different kinds of data. When you try to understand data, you are doing data analysis.

Data analysis includes the activities you do to understand and support the information needs of your organization and its stakeholders. This includes supporting your customers, employees, and business partners. Data analysis includes more than just support. Data analysis is also the examination, cleansing, transforming, and modeling of data. Data analysis helps you understand how the data fits together.

Data is the key to competitive advantage. This means data allows you to out-compete the competition. Organizations that have both reliable and high quality data about their customers, products, services, and operations make better decisions than those without reliable data. Data is considered the currency, the lifeblood, the oil for the information economy. You cannot conduct business transactions without data. This means that data has value and can be treated as asset that can be invested in to obtain future value.

Analyzing different kinds of data can help you achieve different kinds of business value.

**Category Data** This data allows you to communicate information about things. Data is used to classify and assign types to things. For example customers are classified according to market categories or business sectors; products are classified by colour, model, size, etc; and product orders are classified as open, delivered, and closed.

**Resource Data (or Reference Data)** This data allows you to conduct operational processes. Data is used to create basic profiles of resources that are used during operational processes. Profiles are for products, customers, suppliers, facilities, organizations, and accounts.

**Business Event Data (or Transactional Business Data)** This data allows you to conduct business transactions. Data created while operational processes are being executed. Examples of this data include customer orders, supply invoices, cash withdrawal, and business meetings.

**Detailed Transaction Data** This data also allows you to conduct business transactions using internet and sensors. This is data produced by point of sale systems (POS) in stores and online systems, internet interactions such as social media, sensors, machines, and personal devices. This is typically high volume data that is often referred to as Big Data.

Video

### Identifying Business Needs

[Identifying Business Needs](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774451_1)

To identify something essential or important such as a business need, you should consider the kind of addressing the sort of business needs well suited for data analysis. Business needs occur at all levels of the organization from the board room, to corporate governance, to the core processes that run the company. Despite the level of the business need, different kinds of data are well suited to address specific activities for improving the performance for your company. For further information on business needs analysis, refer to “[How to define business needs through business analysis](http://www.dummies.com/business/business-strategy/how-to-define-business-needs-through-business-analysis)”.

To read more, click [here](https://georgiancollege-my.sharepoint.com/:b:/g/personal/katie_tiwari_georgiancollege_ca/EdYTDofIjqtHkUVHaqwdcAIBILdv8jGuGEN2-AxOZE-vmQ?e=ca5I5x)

Whether public, private, or nonprofit, a business serves a market, executes a mission, and — presuming all goes well — fulfills the vision that the leaders have set for that business.

Throughout the course of operations, business leaders set goals and objectives for their enterprise, and they rally teams to work hard and deliver on them. These goals and objectives are business needs;they are the things the business must have or achieve to run, to be profitable, to serve effectively, and to deliver successfully on its mission.

Business needs defined at the highest level may include *capability needs* (statements about providing certain services, delivering a suite of products, assisting others in need, or ensuring the business’s own operational effectiveness) or *improvement needs* (suggestions meant to increase efficiencies or decrease costs, effort, or time-to-market).

Articulating and defining business needs is a part of the activity called *enterprise analysis* and includes identifying and understanding the business’s goals; articulating its strategic direction; and capturing any key concerns pertaining to the business’s successes, challenges, risks, or problems.

Successfully identifying business needs requires critical thinking, analysis, and insight. The business leaders of a company may not clearly tell you what they need, but they’ll probably suggest solutions they want, complain about capabilities they don’t have that’d be helpful, and talk a lot about opportunities they could go after if they only had the hottest new technologies. Therefore, you must do a lot of interpreting.

Digging into the source of leaders’ wishful thinking can give you information about their business objectives and targets. When you work toward identifying why they need those things, you identify the core activities or drivers of the business. A popular business analysis acronym is *IRACIS,* or “increase revenue, avoid costs, improve service.” Typically, the business needs are related to one of these targets.

To serve needs effectively, business analysts must articulate

* What objectives or goals are being served or attempted in specific business area(s)
* What results or outcomes are desired
* What issues or problems are getting in the way
* What solutions are being suggested or considered for implementation or adoption in order to get the business needs met

|  |  |  |
| --- | --- | --- |
| * **ID** | **Problem** | **Opportunity** |
| 1 | Manually tracking student registrations for classes takes too long. | Automate the registration process. |
| 2 | Manually tracking instructor availability and other business rules is too cumbersome. | Automate instructor availability and additional business rules. |
| 3 |  | Our competitors do all their scheduling manually, and we want to be first to market. |

You typically express business needs as broad statements characterizing strategic (and sometimes lofty) goals or as specific statements describing tactical objectives, such as what will be done by when. Consequently, effectively meeting business needs may require broad solutions, like a collection of organizations or operational stakeholders that participate in a variety of initiatives, subsequently designed to meet the objectives.

[Problem Statements](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774451_1)

|  |
| --- |
| Once you have identified the business need, your next step is to define your business problem statement. Your problem statement is the first step to handling the analysis and implementation of a new environment in a structured and procedural manner. In other words, when the problem is addressed, it results in a stable new environment that incorporates desired changes.  To read more, click [here](https://georgiancollege-my.sharepoint.com/:b:/g/personal/katie_tiwari_georgiancollege_ca/ESRohjYKxWxGp1swywgrWsUB_A2aZKv9UyZ_7mxpLTkHzg?e=eQLIYe)  **This brief video serves as a great reminder** that if you’re putting your effort towards solving a problem, it better be the right problem to solve given your situation.  Writing functional problems statements is the key to creating actionable solutions — but how do you go about writing this problem statements? You need to be SMART. Intrigued? **Watch this** [video](https://georgiancollege-my.sharepoint.com/:v:/g/personal/katie_tiwari_georgiancollege_ca/EamHUGRhJGRFmGvCzdg5ru8BhS9UrWFG2k_l6kgnKPo5bw?e=fJCFm0) to find out what SMART problem statements are all about. |

### Analytics Rationale Statement

[Analytics Rationale Statement](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774451_1)

Previously you covered how to develop problem statements using real world business needs and priorities. You will now walk through developing something called an analytical rationale statement.

**Analytics Rationale Statements**

The analytics rationale statement gives you an explanation of the issue or problem being addressed by your project using analytics and metrics. The rationale statement also gives you constraints for how your project is implemented. The rationale statement specifies the following:

• Brief analysis or summary of the problem or problems to be addressed by the project.

• Project goals

• Precise using quotes, examples, references, research data, and press articles.

• Relevant to the audience

• References to similar goals, problems and projects.

Often you will discover that nailing down your rationale statements requires you to understand the exact problem your project needs to address. This is because the problem really is a symptom or side effective of the real problem.

Example: Perhaps the supplier delivering the product late is just a symptom. It really is just a hiccup in the process. “Supplier scheduling results in product being delivered more than 5 days later” is the real problem.

By understanding the real problem, it becomes an integral part of the project implementation. To develop your rationale statement further, you focus on fine tuning your problem statements by specifying the relationships between effect, problem, and cause. To develop your rationale statement, you will alternate between your problem statements and rationale statement until you fully understand the problem and the scope of the project you need to address your problem. The secret to solving your problem is in its proper identification.

Once you understand your problem and what your rationale statement is, the next step is to determine the high level objectives of your project.

Example: “Reduce impact of scheduling on product delivery delays so products are delivered when needed.”

Now you have a strong idea of what your project is set out to do and how to go about doing it.

### Identifying Key Metrics

To further develop your rationale and problem statements, you will need to consider the metrics. Metrics not only allow you to better define and identify problems, they also give you a solid understanding of the kind of data your project needs to use to solve the problem.

**Identifying Key Metrics**

There are so many metrics you can use. Rather than reinventing the wheel, you can pick and choose the metrics organizations typically use for operational excellence. Operational excellence is an element of both organizational leadership and organizational intelligence. It focuses on meeting customer expectations, while stressing the successful application of principles, systems, and tools toward the sustainable improvement. By using these kind of metrics, you can ensure you problem and rational statements cover the full extent of the problem and the kind of project you need to solve it.

The process for selecting metrics are as follows:

**Step 1 – Select metrics based on problem statement**

This is where you read the problem statement and see which objectives and metrics you can use to make your problem statement even clearer. Which metrics can you collect that shows the project is addressing and solving the problem?

Example: “Customer receives product late.”

You review the objectives and metrics and find these objectives and metrics match:

Objective: Improve process responsiveness

Metric: Cycle time (from start of production until product completed) Objective: Delivery responsively to customers

Metric: Lead times, from order to delivery

**Step 2 – Adjust problem statement to be more concise**

Using the objectives and metrics you selected, rewrite the problem statement so it is clear and precise taking into consideration cause and effect.

Example: “Customer receives product late.” becomes “Improve process and delivery responsiveness with shortening [cycle time], [lead times], and [time for completion of product].”

**Step 3 – Update rationale statement and project goals**

Add further information to your rationale statement that reflects your updated problem statements. This can include project goals and constraints for implementation.

Example: “Improve process and delivery responsiveness with shortening [cycle time], [lead times], and [time for completion of product]” is summarized as “Improve process and delivery responsiveness with shortening time frames” in the rationale statement.

**Develop and sustain supplier relationships**

Suppliers are organizations that provide your company with products, services, and raw materials so your company can efficiently operate. Here are objectives and metrics you can use.

***Objective: Lower cost of ownership***

This is lowering the direct and indirect costs of a product or system. Metrics include:

• Activity-based cost of acquiring materials and services (includes cost of ordering, receiving, inspecting, storing, and coping with defects)

• Cost of purchasing is percentage of total purchase price

• Percent of purchases made electronically (Electronic Data Interchange (EDI) or Internet)

• Supplier ratings: quality, delivery, cost

***Objective: Achieve just-in-time supplier capability***

This is reducing flow times within a system and reducing response times of suppliers and customers. Metrics include:

• Lead time from order to receipt

• On-time delivery percentage

• Percent of late orders

• Percent of orders delivered directly to production processes by suppliers

***Objective: Develop high quality supplier capability***

This is increasing your supplier’s capability of controlling quality, delivery, quantity, and price. Metrics include:

• Part-per-million or percent of defects, incoming orders

• Percent of suppliers qualified to deliver without incoming inspection

• Percent of perfect orders received

***Objective: Use new ideas from suppliers***

This is implementing suggestions and recommendations your supplier may have for your company regarding the products, services, and raw materials they provide. Metrics include:

• Number of Innovations from suppliers

***Objective: Achieve supplier partnership***

This is used to manage the number of active supplier partnerships you company has. Metrics include:

• Number of suppliers providing services directly to customers

***Objective: Outsource mature, non-core products and services***

This is managing the aspects of your organization that is not part of your core business or processes. Metrics include:

• Number of outsourcing relationships

• Performance of outsourcing partners

**Produce products and services**

An organization is in business to produce products and services to customers to obtain a profit. Objectives and metrics include the following.

***Objective: Lower the cost of producing products / services***

This is increasing the profit obtained by selling a product. Metrics include:

• Activity-based cost of key operating processes

• Cost per unit of output (for organizations producing homogeneous outputs)

• Marketing, selling, distribution, and administrative expenses as percentage of total costs

***Objective: Continuously improve processes***

This is producing better products with less wasted material and effort. Metrics include:

• Number of processes with substantial improvements

• Number of inefficient or non-value-added processes eliminated

• Part-per-million defect rates

• Yield percentage

• Scrap and waste percentage

• Cost of inspection and testing

***Objective: Improve process responsiveness***

This is producing more products in less time. Metrics include:

• Cycle time (from start of production until product completed)

• Process time (time the product is actually being processed)

• Process efficiency (ratio of process time to cycle time)

***Objective: Improve fixed asset utilization***

This is increasing the effective use of your equipment and machines. Metrics include:

• Percent of capacity utilization

• Equipment reliability (percent of time available for production)

• Number and percent of breakdowns

• Flexibility (range of products /services that processes can produce and deliver)

***Objective: Improve working capital efficiency***

This is increasing the efficiency of managing your company’s inventory. Metrics include:

• Days' inventory, inventory turnover

• Days' sales in receivable

• Percent of stockouts

• Cash-to-cash cycle (days of accounts receivable plus days of inventory less days of payables)

Distribute products and services to customers

This is managing the logistics of getting your products to your customers. Objectives and metrics include the following.

***Objective: Lower cost-to-serve***

This is reducing the cost to serve your customers. Metrics include:

• ABC cost of storage and delivery to customers

• Percent of customers reached by a low cost to serve channels; switching customers for manual and telephone transactions to electronic ones

***Objective: Delivery responsively to customers***

This is shortening the delivery time of products coming to customers. Metrics include:

• Lead times, from order to delivery

• Time from completion of product / service until ready for use by customer

• On-time delivery percentage

***Objective: Enhance quality***

This is improving the quality of products based on customer feedback. Metrics include:

• Percent of items delivered with no defects

• Number frequency of customer complaints

**Risk management**

When businesses operate, they expose themselves to uncertainty. The management of this uncertainty is known as risk management. Objectives and metrics include the following.

***Objective: Manage financial risk /maintain High credit quality***

This is managing the uncertainty of finances during the operation of the company. Metrics include:

• Bad debt percentage

• Percent of uncollectible receivables

• Exposure or losses from interest rate, foreign exchange, or commodity price fluctuations

• Inventory obsolescence and spoilage

• Debt-to-equity ratio

• Interest coverage ratio

• Months of payroll held in cash

***Objective: Manage operating risk***

This is managing the risk of not being able to serve your customers. Metrics include:

• Order backlog

• Percent of capacity from existing and backlogged orders

***Objective: Manage technical risk***

This is managing the risk of being out marketed or out performed by your competition. Metrics include:

• Technology ranking of products and processes compared to competitors

Excellent selection

Certain objectives and metrics are the most key for operational excellence. These objectives and metrics can include the following.

***Objective: Lower the customer's cost; increase the customer's profit***

*•* Price, relative to competitors

• Customer's cost of ownership

• Customer's profitability from own company's products and services

***Objective: Deliver zero-defect products and services to customers***

• Part-per-million PPM or percent of defect rates experienced by customers

• Number and percent of customer complaints

• Number of incidents of warranty and field service repairs

***Objective: Deliver products on time***

• Percent on-time delivery

• Customer lead time (from order to delivery)

• Percent of perfect orders (defect free products and services delivered to the right location at the right time)

***Objective: Offer excellent selection***

• Index a product or service offerings measuring percent of customer needs covered

• Percent of stockouts

**Financial**

A successful strategy is the most important for the success of your company. Objectives and metrics can include the following.

***Objective: Becoming the industry cost leader***

Can your company; compete on price?

• Cost per unit, benchmarked against competitors

• Percent of annual reduction in cost per unit of output

• Percent of cost budget variance

• General, selling, administrative expenses per unit of output or per location

***Objective: Maximize use of existing assets***

Is your company moving product?

• Sales/asset ratio

• Inventory turnover ratio

• Free cash flow

• Inventory efficiency (NPV of new projects to total investment)

• Product and development pipeline to capacity available

• Percent of invoices paid on time

***Objective: Increase account share with existing customers***

Are you growing customers?

• Percent of growth in existing customers' businesses

***Objective: Increase revenue from new customers***

Are you gaining customers?

• Dollar revenue from acquiring new customers

Culture

High business performance doesn’t just come from the processes of your organization. It also comes from the culture, people, and how people work together. Objectives and metrics can include the following:

***Objective: Develop skills and quality management and process improvement***

This is developing skills, capabilities, and competence. Metrics include:

• Percent of employees trained in quality management techniques

• Number or percent of employees qualified at "black belt" SixSigma quality level. For more information on SixSigma, refer to https://www.sixsigma.com.

• Percent of employees with knowledge and training in activity-based management, just-in-time, theory of constraints

***Objective: Technology that facilitates process Improvement and customer satisfaction***

This is improving the tracking and communications of your organization. Metrics include:

• Percent of employees who obtain immediate feedback from operations

• Percent of customers who can track order status electronically

***Objective: Culture for continuous improvement***

This is aligning your companies culture towards continuous improvement. Metrics include:

• Employee survey on culture for continuous improvement and knowledge sharing

• Number of new process improvement ideas generated

• Percent of employee process improvement suggestions adopted

• Number of ideas for quality and process Improvement shared across multiple organizational units

• Performance improvement from employee suggestions and actions (cost savings, defect reduction, yield enhancement, process time reductions)

By focusing on objectives and metrics for operational excellence you can ensure your projects will address all the problems they are set out to solve.

### Confirming Completeness

By now you should have a good idea of the actions and activities your project will need to complete to solve the problems specified. To confirm your analytical rationale statement is complete, you will need to consider the kind of analysis and data you’ll need. By the knowing the sort of analyses needed, you’ll know if there is any further information missing from the analytics rationale statement.

You will use two kinds of modeling to complete your analysis –exploratory and descriptive. Exploratory modeling uses statistics, data mining, and predictive analytics to match model results with existing data. Descriptive modeling uses algorithms to define and refine relationships between variables within the data. Preliminary versions of both these models are good checklists for the confirming rationale statement is complete.

### Making the Final Check

When developing rationale statements there is always the danger of seeing only the trees and not the forest. To see the forest from the bird-eye view, you need to confirm that what you currently know about the project covers the needs of those the project is being done for. They need to make a decision.

For your project to be successful, you need to align both the problem statement and analytics rationale statement to the potential analytical solutions that the project will recommend. When your project completes, those the project is for will make multiple decisions on your project and its outcomes. For your project to be successful you must anticipate what these decisions are and the characteristics these decisions have. Characteristics include:

• Values, goals, and objectives relevant to the decision

• Nature of the decision

• Uncertainty of the decision

• Consequences once the decision is made

• Technology and tools used

• Elements of a good option or alternative

Some aspect of decision making is usually included in the analytics rationale statement.

### Data Requirements

[Data Requirements](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774451_1)

To solve your problem your project is set out to accomplish, you must perform data analysis. To do data analysis, the data must be set up a certain way. The specifying of this is called data requirements. Data requirements specify the structures, relationships, and data points. Data requirements can also be determined by decisions, laws, policies, and how data is consumed and used within the organization.

**Data Requirements**

There are different categories of data requirements. To ensure your data requirements are complete, you specify data requirements for each category. Here are a few of these categories.

**Definitions and Concepts**. These requirements specify the meaning and scope of terms used in requirements. Definitions are unique and unambiguous and may include aliases for the terms being defined.

* • Example: Prospect is a customer when prospect is registered with the corporate app. Customer is a prospect that has initiated the buying process.

**Movement.** These requirements specify how data flows from a source to destination. Sources and destinations can be within a single system, across multiple systems, and spanning external systems.

* • Example: Customer information is transmitted from ERP System to Visualization Tool. (ERP is short for Enterprise Resource Planning which is business planning software that integrates all functions of a company together.)

**Transformation.** These requirements specify how data is processed through specific processes, rules, or algorithms.

* • Example: Customer is associated to the group “future buyer of product Y” when customer purchases products X0, X1, and X2.

**Business Objectives.** These requirements specify which data to use and how it is used to support business objectives through prescriptive/predictive analytics to reduce risk and improve business operation quality, efficiency, communication/training effectiveness, and speed of change adoption.

* • Example: Customer information is submitted to Sales team to improve Sales team performance.

**Models.** These requirements specify the data needed and how the data is used in models for successful data analysis. These include requirements that specify what makes a good model, how data is related to each other, how the model can be validated, and how the model can be optimized and improved.

* • Example: Customer information is submitted to Classification algorithm to predict the number of future purchases.

**Visualizations and Reports**. These requirements specify how dashboards, reports, and visualization tools will consume and present the data to audiences. These requirements also specify which audiences will be using the visualizations and how they will use the visualizations during individual / group analysis and decision making.

* • Example: Customer information is displayed to manager using line chart.

**Datasets**. These requirements specify the dataset: the data structure, data included, data excluded and the valid values for the data. These requirements can also specify statistical validity of data, how the data can be interpreted and analyzed.

* • Example: Customer information contains name, products purchased, and amount paid.

**Data security.** These requirements specify the authorized personnel that have access to the data during collection, processing, and analysis. These requirements can also specify security strategies such as encryption and de-identification.

* • Example: Analyst accesses customer information during its collection.

**Data quality.** These requirements specify what is considered an acceptable level of quality for the data when it’s used for discovery, classification, profiling, and mapping. Data quality requirements answer the following questions:

• Where does information reside?

• What kinds of information are present based on standardized patterns?

• How is data populated and structured?

• Example: Customer address information is 100% accurate 75% of the time.

Once you determined your data requirements, you will know which data you will need to solve your problem statements.

### Data Sources

Once you identified and understand your data requirements, it is time to identify the data sources you need to solve your problem statements. There are different kinds of data sources you can use. You would select the best one that meets your data requirement.

**Data Sources**

There are times you will find your data requirements too restrictive, meaning there are no data sources you can find to meet them. In these situations, you would loosen your requirements so you can move options for the data sources you use.

For predictive and prescriptive analysis, you could use the data found internally to your organization. You can also use external data outside your organization from other sources. By using both internal and external data, you can gain strong insights from your data analysis. Here are a few of these external data sources you can use.

Open Data. This is data on everything from government, health, and science top popular games and data trends. A popular open data source is Kaggle. You can visit Kaggle here: https://www.kaggle.com/datasets

Government Open Data. This is data collected by governments on it policies, industries, and citizens. Examples of these include:

• USA – https://www.data.gov

• Canada – https://open.canada.ca/en/open-data

You can use this data to analyze market trends, talent pools, and compensation packages. There are also organizations that aggregate this government data to facilitate your collection and analysis of data. One such organization is the World Economic Forum. You can access it here: https://www.weforum.org/search?query=dataset

Vendor Communities. This is data analysis projects completed by experts using specific vendor tools. An example of this is a community for Tableau called Makeover Monday.

Remember, when solving your problem statements, you focus on your internal data first. Once you understand your organization’s internal data, you use the external data sources to refine your data analysis.

Example: You work for a retail chain of stores. You first analyze the internal information that your company has collected. You then go to an open data source and cross reference what you learned with traffic patterns data to gain further insights on your company’s customers.

### Data Assumptions and Constraints

After you define your data requirements and examine the data sources that meet those requirements, you would define the assumptions and constraints.

**Data Assumptions and Constraints**

After you define your data requirements and examine the data sources that meet those requirements, you would define the assumptions and constraints for the data in those data sources to help guide your data analysis. An assumption is something accepted as being true while a constraint is a limitation or restriction.

Example: Assumption: Customers purchase product Y after purchasing product X. Constraint: Types of customers for insurance lines include auto, fire, and marine.

When specifying your assumptions and constraints, you would consider exploratory, descriptive, and inferential analysis.

• Exploratory analysis: This analyzing datasets to determine its characteristics. Statistics and non-statistical methods can be used.

• Descriptive analysis. This uses statistics to summarize data and makes generalizations of a population in a dataset.

• Inferential analysis. This uses statistical tests to determine if observations in the data are from chance or from cause and effect.

Also, when specifying your assumptions and constraints, you need to consider these during data analysis:

**Cause or correlation.** When examining and analyzing historic data, you must rule out the possibility of coincided events with the real causes.

• Example: Customers buy both product A and product B. With deeper analysis you discover that customers don’t buy Product B because they bought Product A. They buy Product A when it’s raining and they buy Product B when it’s drizzling.

**Time.** When predicting outcomes, you must rule out effects from the passage of time.

• Example: Your customer watched a cowboy movie 5 years ago. However, they never watched one again since. Because of this, you would not recommend cowboy movies to them now.

**Interpretation.** You must anticipate which data will be examined and interpreted and ensure that it is interpreted correctly.

• Example: Customer information includes only information from 2018 for auto insurance products. Do not extrapolate this data for different years or insurance products.

**Measurement.** You must ensure measures are credible, reliable and unbiased.

• Example: When using a measurement such as [total number of customers], that measurement must include all customers. Sometimes customer measurements are calculated without considering all customers that bought all products. For instance, some customers that bought low end services may not be counted as a customer. This will make the [total number of customers] misleading and unreliable.

### Data Analysis Project Proposals

Now that you have your requirements, assumptions, and data you are ready to draft a proposal of your data analysis project.

**Data Analysis Project Proposals**

Now that you have your requirements, assumptions, and data you are ready to draft a proposal of your data analysis project. Your data analysis project’s main goal is to improve decision making by finding useful patterns and insights form the data. Often this involves programmatically examining large amounts of data from different perspectives that summarizes the data so useful patterns and relationships are discovered. From these patterns, predictions and forecasts can be determined. There are different kinds of analytics you can focus on during your project.

**Descriptive Analytics.**

“Let’s figure out what we can do, determine our options, and then next steps”

These analytics use historical data to understand and analyze past business performance. Business information is categorized and consolidated to suit the needs of your audience. These kinds of analytics are presented by executive management dashboards, middle level management key performance indicator scorecards, and operational level management charts. No assumptions are made during the analysis regarding what decisions need to be made or what actions need to be carried out.

**Predictive Analytics.**

“Let’s talk about our options then next steps”

These analytics applies statistics to identify patterns and relationships then uses that understanding to make predictions about future events. Assumptions are made during the analysis regarding what decisions need to be made and what actions need to be considered.

**Prescriptive Analytics.**

“Let’s talk about our next steps”

These analytics identify decisions to be made and to initiate appropriate action to improve business performance. Statistical optimization and simulations are used to determine the best solution among various choices.

Once you have determined which kind of analytics to focus on, your next step is to determine how you plan to analyze the data. Here are a few different approaches you can use.

**Classification and class probability estimation.** This is where you predict for each individual in a population which class or group this individual belongs to.

• Example: Among all the customers of the company which are likely to respond to a given offer? A customer can either belong to the “respond” class or “not respond” class; or a customer can belong 25% to the “respond” class and 75% to the “not respond” class.

**Regression or value estimation.** This is where you predict the value of a variable for each individual using historic data.

• Example: How much will mobile customer use the service? You would predict this by looking at other similar individuals in the population and their historic usage.

**Similarity matching.** This is where you identify similar individuals based on data known about them.

• Example: Which companies are similar to our company’s best business customers? Similarity matching is used for making product recommendations by identifying people with similar likes and recommending to you what they like.

**Clustering.** This groups individuals in a population together by the similarity but not driven by any specific purpose. In doing so, it is a faster approach.

• Example: Do our customers form natural groups or segments?

**Co-occurrence grouping.** This finds associations between entities based on transactions involving them.

Example: What items are commonly purchased together? Burgers and buns is a prime example.

**Profiling.** This characterizes the typical behaviour of an individual, group, or population.

• Example: What is the typical cell phone usage of this customer segment? Profiling cell phone usage may require a complex description of night and weekend airtime averages, roaming charges, text minutes, and so on. Behaviour is described over the whole population or at the level of an individual person.

**Link prediction.** This predicts connections between items by suggesting that a link should exist and estimate the strength of that link. Link prediction is used in social networking systems to recommend people you should connect to.

• Example: Since you and Karen share common friends maybe you'd like to be Karen's friend?

**Data reduction.** This takes large sets of data and replaces it with a smaller set of data which emphasizes what is considered important information in the larger data sets. The smaller dataset is easier to manage and process.

**Casual modeling.** This helps you understand and predict what events and actions influence others.

• Example: For advertisements to consumers you observed that the targeted customers purchased at a higher rate that those that were not targeted. Was this because of the advertisements or were these customers planning to buy anyway?

Once you determined the kind of analytics you are focusing on and your approach to data analysis, you are ready to draft a preliminary data analysis project proposal, also known as a statement of work that will become your action plan. You will notice the analytics project proposal has similar sections to a business case.

• **Executive Summary**. This section gives an overview of the whole proposal.

• **Analytics Rationale Statement.** This section is the analytics rationale statement. You covered in a previous lesson.

• **Problem**. This section contains information on your problem statements, metrics and success criteria. You covered this in a previous lesson.

• **Data Requirements**. This section is your list of data requirements written in a concise form.

• **Data.** This section is a list of your data sources and data sets used for the project. It also includes descriptions of what they are and where they are located.

• **Data Analysis Approach**. This section includes your methodology, algorithms, and software tools you will use for your data analysis.

• **Project Overview**. This section includes the scope and goals for your project and how you’re going to report your projects progress using which measures.

• **Project Plan**. This section contains information on what tasks the project will include and what will be delivered and when. This also includes who may be doing the work. Schedules can range from hours, to weeks, to months. Tasks can include:

• Collect information

• Analyze data

• Develop algorithm

• Develop model

• Test and optimize model

• Execute post-mortem

Procedures you need to complete for your dataset and model are extremely important for you to define ahead of time. It is a good idea to constantly assess your model using different subsets of your dataset. Here are the different kinds of procedures you will perform to define.

**Training procedures.** This is data you use to fit your model. Be cautious to not over fit your model.

**Validation procedures.** This is data you use to predict errors during selection.

**Testing procedures.** This is data you use to assess the error for the final model.

**Performance procedures.** This is data you use to performance tune your model’s algorithms so you run these procedures and additional data experiments within reasonable time frames.

When using different subsets of your dataset, you must ensure they are randomly divided. You must make your datasets representative of real world data. You also cannot make your datasets too similar during these procedures. In doing so, you will deliver a poor and ineffective model.

**Project Cost/Budget.** This section is an estimate of the costs for delivering the project.

For examples of the different kind of information that can be found in data analysis project proposals, refer to Columbia University’s Proposals located at <http://www.ee.columbia.edu/~cylin/course/bigdata/proposals.pdf>

### Summary

In this lecture you walked through what data is and that it is more than just factual electronic information. Data can also include the subjective interpretation of observations. You examined the importance of data analysis for business value and how it contributes to business communication, conducting operations, and completing business transactions. You have seen how to identify the common business needs and how to define a problem statement from those needs using the Definition stage of the Intervention Strategy Model (ISM).

Then you walked through developing an analytics rational statement from a problem statement. You refined the problem statement and analytics rationale statement using objectives and key metrics commonly used for operational excellence. You explored how you refine problem statements and analytical rationale statements by considering the decision making needs of people the project is for. By considering decisions and technology, you saw how to align the problem statement with analytics solutions.

Finally, you examined the different categories for data requirements that range from business outcomes, data sets, and visualizations. You saw how to use data requirements to select data sources from open data sources and how to define assumptions and constraints to support your data analysis. You walked through what is in an data analysis project proposal, and its sections including data requirements, approaches, project plans, and project costs.

**Software tools**. You also need to consider the software tools used to help facilitate your analysis.

• Example: Tableau will be the visualization tool of choice.

With keeping all of these in mind, specifying assumptions and constraints is a straightforward process. You just walk through each of your data requirements and if possible, expand on them by specifying your assumptions and constraints.

• Example: The data requirement is “customer information contains name, products purchased, and amount paid.” A potential assumption can be “customers can be up sold to Product B from Product A 60% of the time”.

Ensure your assumptions are brief and to the point.

During this lesson you will examine the process for calibration and how to evaluate data structures to best solve analytical problem statements. You will walk through how to select and use calibration model techniques and processes to increase accuracy and usefulness in data models.You will also see how to select the best software tools to identify relationships in your datasets. Finally, you will walk through how to assess dataset assumptions and constraints to develop effective data models. You will see how to identify selection criteria to create and evaluate data models and gauge your data model’s effectiveness with a data model selection scorecard.

By the end of this lecture you will have demonstrated the ability to:

* Assess dataset assumptions, limitations, and constraints in order to develop effective data models.
* Identify the selection criteria for data model creation and evaluation.
* Create a data model selection scorecard (i.e. evaluation and validation).
* Evaluate data model structures to help solve analytic problem statements.
* Assess and select appropriate software tools to successfully identify potential relationships in the data set(s).
* Select calibration model techniques and processes to increase accuracy and viability in data models.

### Validation and calibration with linear regression

Model calibration is the process of adjusting your model parameters to achieve a goodness of a fit for data that your model hasn’t experienced before. This video walks through an example of calibration, comparing predictions to observations. It then shows how rankings and adjusting the model makes better predictions.

While watching consider the following:

• Why should you validate your model?

• What is the difference between concordance and calibration?

• How do you calibrate your model?

[Calibration](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

Calibration is an approach used for fitting a statistical model to a dataset to better predict outcomes.

**Calibration**

Calibration is an approach used for fitting a statistical model to a dataset to better predict outcomes. When developing your statistical model, you constantly confirm that your model is accurate and precise. And you will use different datasets to ensure that your model is a good model. Calibration is straightforward. To perform calibration, you observe the difference of your model’s predictions to actual observations, outcomes and data then adjust your model’s parameters.

Example: When treating patients with a disease, you want your model to better predict your patient’s response to their treatment. After comparing your model’s predictions to the actual results, you adjust your parameters so your model has a better fit – that it is calibrated.

Calibration is a process that is repeated over and over until you have the results you want. Here are the basic steps for calibration.

**Step 1** – Preparation. Before starting your calibrations, you determine your objectives and measures for a successful calibration. You must define your acceptance criteria before you start. Because when you do start, you’ll be repeating the process over and over until you get the results you need.

**Step 2** – Dataset. You examine and confirm your dataset has the right characteristics that will result in your model being successfully calibrated. Is your dataset a good sample? Does your dataset reflect reality? Can you split your dataset to do better model validation and calibration?

**Step 3** – Calibrate. You then use a dataset and compare your model’s predictions to the actual results recorded in your dataset. You then adjust your model’s parameter values so your model predicts accurately what is recorded in your dataset. You repeat this step as often as you need to by using different datasets. Datasets should have these characteristics.

• Spatial variation (values differ at different spatial locations) to ensure model validity when using different datasets.

• Temporal variation (time variation) to adjust for data changing after collection.

• Heterogeneity (diverse) when average data is available but the data it’s based on is not available.

**Step 4** – Results. You review the results of your model’s predictions to determine you have completed your model’s calibration. You compare your results to your acceptance criteria to confirm you’re done.

### Assessing Data Structures

One of the key factors for your model’s successful calibration is the data structure in your dataset. Not having the right structure or including the right data in that structure will result in your model not being useful for making predictions.

**Assessing Data Structures**

One of the key factors for your model’s successful calibration is the data structure in your dataset. Not having the right structure or including the right data in that structure will result in your model not being useful for making predictions. Consider the following for a good data structure.

**Terminology**. Your data structure naming is intuitive and reflects the business terminology used in your company, your problem statements, and in your data analysis proposal.

* • Example: Customer, sales staff, and revenue.

**Intuitive**. Your data structure must be easy to follow when discussed and easy to understand when viewed visually.

* • Example: Hierarchies, trees, and networks.

**Efficient.** Your data structure is optimal for data analysis by both humans and machines.

* • Example: Fields in a dataset that are often used together are located right next to each other. This makes it easier for humans to associate these fields together in relationships. This makes it easy for machines to process because having fields right next to each other can make algorithms easier to program.

**Spatial and Temporal Variance.** Your data structure includes fields that support these kinds of variances.

* • Example: You look at the spatial variance of data points based on the coordinates for a map of the city. You notice that a lot of points are close together. There is a higher density of data points for a specify city block. You also look at temporal variance of data points where events occur at specific point of time of the month. You see a larger concentration of the data points on the first two days of the month.

**History and Events**. Your model’s main purpose is to predict outcomes. These outcomes are often events that repeat over time.

* • Example: Employee resignation, customer complaint, year 2018, and month April.

**Business.** Your data structure includes measures and ties to your business entities and organizations. These ties can be names of processes, rationale, policies, and organizations.

* • Example: Core process, finance, and production.

**Future Use**. Your data structure includes fields that your business will need to use to predict outcomes in the near and not so near future.

* • Example: Currently, purchases are best predicted by price when purchasing cell phones from your provider. Analysis shows that a shorter time of delivery is another factor. To be ready for the change in the organization’s business model, closest location is added to the data structure.

### Calibration Techniques

**Calibration Techniques**

Calibration is a straightforward yet very repetitive process. It involves quite a bit of trial and error. So always prepare and have a plan before you start calibrating your model. Also make notes of what you did, because there is always a strong likelihood you will need to start over and repeat your steps. Here are the main techniques for calibration. Depending on your model, you may find it easier to combine the techniques or use one technique over the other.

**Technique 1** – Calibration based on predicting outcome values. You compare your model’s predictions and calibrate using linear regression. Example include integers such as 1, 2, 3, 4, 5 and real numbers like 1.2, 3.2, 4.9.

**Technique 2** – Calibration based on predicting outcome categories. You compare your model’s predictions and calibrate using logistic regression. Examples of categories are 1, 2, 3, A, B, and DB. Each category is mutual exclusive from each other.

**Technique 3** – Calibration based on predicting outcome classifications. You compare your model’s predictions and calibrate using Bayes inference. Classes are similar to categories. The difference is that categories are mutual exclusive while classes can be mutual exclusive or be contained in other classes. Examples of classes are trees, spruce, oak, bushes, and strawberry bushes.

### Effective Software Tools

[Effective Software Tools](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

When performing your data analysis, it is very important that you use the best tool for the job. But the right tool highly depends on what you need to do and what your level of skill is with data analysis using the tool

**Effective Software Tools**

There are a lot of different software packages and programming languages out there. Here are a few of the features these software tools can have.

• Visualizations and data analysis such as budgeting, forecasting, multivariate, regression, time series, and text, and survival analysis

• Scorecards, reporting, dashboards, and ad hoc analysis

• Analytics such as predictive and web analytics

• Statistical analysis and process control

• File management, data management, pre-processing, merging and blending data

• Extension, customization, algorithms, and programming

• Model development, validations, calibrations, and anomaly detection

• Questionnaires, interviewing, and web surveys

• Simulations and experimentations

• Business intelligence

• Industry specific problems

• Beginner and advanced skill levels Small and large datasets

The best tool for the job is the one that does what you need it to do. Therefore, features are important. However, there are other criteria you need to consider for your software tool.

* • **Problem to be solved**. Some problems may require detailed programming while others just require visualizations and using features right out of the box.
* • **Level of expertise in statistics.** Your level of skill in statistics, may it be a beginner or an expert will determine the kind of tool you should use.
* • **Level of expertise in the tool**. You should use tools that you and your company use and know very well. You should not use tools that are unfamiliar.
* • **Prototype, dashboard or product**. Some tools are geared towards a certain set of features. Some tools are good at prototyping and test running models. Other tools are good as showing visualizations and dashboards. And some programming languages are really robust at building products to make data science and data analysis happen.
* • **Privacy of data**. Sometimes the data you’re analyzing is sensitive, so how it is stored and moved must be considered. Some tools do not provide the level of security with the data that you and your company needs.

As mentioned, there are software packages and programming languages. Here are the best software packages on the market for statistics.

* • Tableau –This helps you transform data into insights that make an impact. Read more about Tableau at https://www.capterra.com/p/149389/Tableau/
* • Domo –This provides you with a customizable platform for statistical analysis, visualizations, and collaboration between decision makers. Read more about Domo at https://www.capterra.com/p/161507/Domo/
* • Matlab–This is a programming environment for algorithm development, data analysis, and visualizations. Read more about Matlab at https://www.capterra.com/p/125813/MATLAB/
* • SPSS –This helps you find relationships in data to predict future events and what’s likely to happen next. Read more about SPSS at https://www.capterra.com/p/13990/SPSS/
* • QlikSense–This helps you create personalized and interactive visualizations, dashboards and reports. Read more about QlikSense at https://www.capterra.com/p/145530/QlikSense/
* • Statistix–This provides you with basic and advanced statistics for researchers. Read more about Statistix at https://www.capterra.com/p/125836/Statistix/
* • Minitab 18 –This helps you with quality improvement statistical education. Read more about Minitab 18 at https://www.capterra.com/p/109731/Minitab-17/
* • Stata –This helps you with data analysis, data management, and graphics. Read more about Stata at https://www.capterra.com/p/119880/Stata/
* • SAS Business Intelligence –This helps you discover data, do interactive reporting, apply analytics, and then monitor those analytics. Read more about SAS Business Intelligence at https://www.capterra.com/p/127/SAS-Business-Intelligence

To read about more software packages, go to https://www.capterra.com/statistical-analysis-software

Shifting gears, here are the best programming languages for data science on the market.

**R** – This is a free and powerful alternative. It is used by statisticians, data scientists, and also Wall Street traders. It allows you to manipulate data and create graphics with only a few lines of code. It is highly extensible with new packages and features constantly becoming available. R is often used as a prototyping language where the prototypes are handed to written in Java or Python. Read more about R at https://www.r-project.org.

**Python** – This has a sophisticated data mining capability and is practical for building products. It is intuitive and easier to learn than R and is growing in features and new packages for statistics. Read more about Python at https://www.python.org.

**Julia** – This fills in the gaps that other languages such as R, Python, Java, MatLab and SAS leave open. Juila is a high-level, fast, and expressive language. It’s faster than R. It’s more scalable than Python. It’s extremely easy to learn. The downside is that Julia is still a fairly new language. Read more about Julia at https://julialang.org.

**Java** – This is the language found deep in the skeletons of the biggest tech companies. It is the foundational language for data engineering infrastructures. Java doesn’t provide the same level of competency in statistics and visualizations as R and Python do. But, Java is ideal for building high performance products and systems. Read more about Java at https://www.java.com/en

**Scala** – This is a Java based language used for doing statistics, machine learning at large scales, and building high-level algorithms. Read more about Scala athttp://www.scala-lang.org.

### Model Selection in Multiple Regression

[Model Selection in Multiple Regression](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

Model selection in regression can involve removing unimportant variables. This video covers two strategies for model selection for auction data on eBay that involves adding and removing variables.

While watching consider the following:

• How do you identify a smaller model?

• When do you use backward selection?

• When do you use forward selection?

### Effective Data Models

[Effective Data Models](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

The data model is the structure of the data used by algorithms to make predictions.

**Effective Data Models**

The data model is the structure of the data used by algorithms to make predictions. Your data model has two views to it; a top down view and a bottom up view. The top down view is determined by your data requirements and shows how groups of fields are related to other groups of fields within your dataset. The bottom up view specifies the variables of your predictive model and shows the relationships between fields within the dataset. Using these two views, your data model ties your data requirements to your predictive model.

**Example:** At the high level your data model includes relationships between groups of fields. Customer information with a one to many relationship to products purchases. At the low level, your data model includes fields and relationships between fields. Confidence Interval, product cost, Customer Name purchases Product on Day, and etc.

You will put a lot of effort into your data model. To make it effective, you must consider what your data model is being developed for. You need your data model to run algorithms to predict outcomes and unknown values. These outcomes don't have to be just about the future. They can be about the past and present too. Examples include:

* • Past: Credit score estimating the likelihood that a potential customer will default become a write off.
* • Present: Spam filtering that the email you received is spam.
* • Future: Identifying bank accounts that are most likely to be defrauded.

The effectiveness of your data model relies on it being simple and directly addressing the specific decision making problems and data processing problems that it is set out to do. Do your best to answer these questions.

* • What is the business reason for the prediction? Planning? Decision making? Intervention?
* • How accurate do you require the predictions?
* • What is the cost of an inaccurate prediction?

To ensure your data model is effective, you have to ensure you are not specifying relationships between fields and elements that will result in overfitting your model. Overfitting is the result of you training data to the exact data at the expense of generalizing unseen or unknown data points. It is the equivalent of your model memorizing data points rather than being able to predict outcomes and values. This equates to bad predictions, defeating the purpose your data model is set out to accomplish.

* • As the Nobel Laureate Ronald Coase said "if you torture the data long enough, it will confess."
* • And your data will confess to anything.

There is no method or approach that prevents you from overfitting. Your model is always in danger of overfitting -- you just can't avoid it. You must rely on your knowledge, expertise, and experience to respond to and reduce overfitting. To reduce overfitting, consider the following while developing your data model.

* • Complex models are only necessary if the phenomenon being predicted is complex.
* • Visually show with diagramming what your data model requires your algorithms to predict based on the training and testing of different data sets
* • Specify cross-validation in your data model. Cross-validation is a systematic way of splitting up a single data set into multiple datasets, each used for different steps during your model's validation; different steps such as training and testing.

In the end, it's up to you, not approaches or algorithms, to ensure your data model is effective.

### Data Model Assumptions and Constraints

**Data Model Assumptions and Constraints**

Previously, you learned about assumptions and constraints for data requirements. Do not confuse these with the assumptions and constraints of data models. These are very different things for data models. For instance, assumptions put boundaries around your models to guide analysis. You have assumptions and constraints that are determined by your data requirements. You also have assumptions and constraints that are used to specify your data to be used for algorithms. Here are a few of the different kinds of assumptions.

• **Patterns**. These are assumptions specifying patterns found in the data that the data model relationships and structure need to take into consideration.Example: When a customer purchases burgers, there is a 80% chance of them purchasing buns.

• **Distributional**. These are assumptions specifying your data’s probability distribution of observations and random errors.Example: Customer purchases have a bell-shape distribution

• **Relationships**. These are assumptions specify your data’s statistical relationships between fields. Example: Price and purchase frequency have a linear relationship

• **Dependence.** These are assumptions specifying your data’s patterns in joint probability distributions of different observations or random errors. Example: Observations or errors for purchase frequency are statistically independent.

Constraints for data models are different too. Constraints, also known as hard constraints are conditions of an optimization problem that the solution needs to satisfy. There are several types of constraints.

• **Equality.** This is the relationship between two fields (variables) that are equal. Example: A = B

• **Inequality.** This is the relationship between two fields (variables) that are not equal. Example: A <> B

• **Integer.** This is specifying that a field (variable) must only have integer values. Example: A = 1, 2, 3, 4, 5 ...

• **Text List.** This is specifying that a field must only have specific text values. Example: A = Apple, Banana, Cucumber ...

Hard constraints always need to be satisfied. Soft constraints, on the other hand are non-mandatory constraints. These are usually constraints around preferences around planning. Example: Preferring solution A over B but there is not enough data to provide A, so you will settle for B to incorporate into your plans.

### Keep in mind that you will need to confirm your assumptions and constraints in later phases of your analysis. You will confirm your assumptions based on the availability of data. If the data is not available, it may make sense to loosen or remove assumptions. Also, you will confirm the validity of constraints while testing the dataset to see if the data makes sense.

### Data Model Evaluation and Selection

**Data Model Evaluation and Selection**

When creating and evaluating your data model, you rely on your data requirements as a starting point. You then use the assumptions and characteristics of your data, to help determine your full data model. Note that when you have your full data model, you must evaluate your data model. This means that your data model achieves its objectives at a high level as specified by the data requirements. Always confirm your data model meets requirements.

**Example:** Data requirements specify customer information. So your data model includes the fields customer name, location, and lifetime value. You realize that a preferred product is also customer information. So you add it to your data model. You feel your data model strongly satisfies data requirements.

After completing evaluation, you focus on making your data model simpler and specify improvements in your algorithm’s prediction accuracy. However, having one data model may not be adequate for your predictions. For instance, data models that specify linear regression may work in some cases while logistical regression or Bayes may work better in other cases. Sometime a combination may work even better. You will cover more about selection and the selection score card later in the lesson. Specify what you know and feel comfortable with. When doing data model selection, always keep in mind that "all models are wrong". This phrase was published in a 1976 paper in the Journal of American Statistical Association. The paper gives a warning of the tendency of models being over-elaborated and over-parameterized. To gain a better understanding of what assumptions/constraints you may need to specify for your data model review

### <http://statweb.stanford.edu/~jtaylo/courses/stats203/notes/selection.pdf>

### Evaluation Procedure For Your Data Model

### Evaluation Procedure For Your Data Model

### 

### Data Models Specifying Training, Validation, and Testing

**Data Models Specifying Training, Validation, and Testing**

Developing your data model (statistical model) was discussed in a previously in regards to calibration. Development involves three main activities – training, validation, and testing. To make sure your data model is effective for to complete these activities using algorithms, you must cautiously determine your best datasets and ensure your data model reflects them accurately. That is easier said than done. There is quite a bit of leg work you must do for your data sets.

Data quality. You need to ensure the quality of your data in your data set. And this should be checked as soon as possible.

* • Is the data correct?
* • Can incorrect data be corrected?
* • What data is missing?
* • Can missing data be added?
* • And most importantly, can the data be trusted?

To answer these questions, you can use counts, means, standard deviation, medians, frequency, and visuals with color schemes. Does your data model take into consider the quality of the data?

**Randomization.** You need to ensure the data in your data set are distributed within and across groups. You also need to ensure that in doing so, the data set is a representative sample. For instance, ripe apples can be red, green, and yellow. You wouldn't just include red apples in your data set. Does your data model consider randomization for specific fields?

**Characteristics.** You need to understand the characteristics of your data in your data set. The best way to do this is to look at your data through the following ways.

• Basic statistics of important fields and variables. This will give you a ballpark idea of how these variables behave.

• Scatter plots. This will give you a visual understanding of the relationships between fields and variables.

• Correlations and associations. This will give you a numerical understanding of the relationships between fields and variables.

• Cross-tabulations. This will give you a positional understanding of the relationships between fields and variables.

Are all relationships, constraints, and assumptions included in your data model?

**Events and Interventions.** Events occur and interventions are done. Both can drastically impact the data in your data set. You must ensure that event information, including from events caused by interventions are included in your data set. Events can drive and change data. Not including events can greatly reduce the prediction accuracy of your algorithms. For instance, when predicting customers to buy your product, customers can be influenced by various different marketing promotions. It's a good idea to include events created by these marketing promotions and use them to increase prediction accuracy. It is also a good idea to include customer's buys that were not determined by marketing promotions. This will allow you to factor out errors in predictions where the customer is going to buy despite the marketing promotion used. Consider the following when it involves events and interventions.

* • What actions and other events that occur before a specific event, during that specific event, and after that specific event.
* • If unknowns are needed to be known before the event.
* • Combinations of events occurring at the same time.
* • Event doesn't occur or doesn't occur in a way expected.
* • Unexpected events occurring.

Does your data model include fields that make it easy to analyze events?

**Your expertise.** You must make sure you have the expertise to analyze the data in the data set. You must know and be able to apply the needed subject matter expertise. Does your data model reflect the expertise of those that will use it?

Once you understand the data in your data set, you are ready to specify the data sets in your data model for training, validation, and testing. Make sure you specify the following for each activity at that you use a different data set for each activity.

* • Training. You specify a data set to fit the parameters of your model to the data.
* • Validation. You specify a data set to avoid overfitting and fine tune the parameters of your model. You specify tests for the accuracy of your model's predictions with known values.
* • Testing. You specify a data set to test the strength and accuracy of your model’s predictions of unknown values.

### • For each activity, you specify a visual and a probability distribution table that would show the results of your model's predictions. With those, you can adjust your data sets accordingly so your algorithm’s increases in prediction accuracy. Making your data model right and effective can be a time consuming task.

### Data Model Selection Scorecard

[Data Model Selection Scorecard](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

The data model selection scorecard lists criteria that determines the effectiveness of a data model.

**Data Model Selection Scorecard**

The data model selection scorecard lists criteria that determines the effectiveness of a data model. This scorecard allows you to compare the strengths and limits of each of your data models. Here are a few of the main criteria,

• AIC (Akaike Information Criterion) This is an estimate of the relative information lost when the model represents the process that generated the data. Read more about AIC at http://brianomeara.info/tutorials/aic/

• AICc (Second-order Akaike Information Criterion) This builds on AIC and takes into account sample size. Read more about AICc as http://brianomeara.info/tutorials/aic/

• BIC (Bayesian Information Criterion) This uses the likelihood function with AIC that gives a score on the estimate of model performance on a testing dataset. Read more about BIC at http://stanfordphd.com/BIC.html

• RIC (Risk Inflation, Threshold) This focuses on the adjustment to the risk when using AIC. Read more about RIC at https://projecteuclid.org/euclid.aos/1176325766

You can also use model evaluation and validation criteria for data model selection too. Here’s a few of them.

• Confidence Interval. This tells you how reliable a statistical estimate is. Wide confidence intervals means that your model is not effective.

• Confusion Matrix. This is used in clustering. When the concentration of observations is in the diagonal it means the higher the accuracy and predictive power of your clustering.

• Gain and Lift Chart. Lift is the ratio between the results obtained with and without the model.

• Kolmogorov-Smirnov Chart. This is a non-parametric statistical test used to compare two distributions and assess how close they are to each other.

• Chi Square. This is a parametric statistical test similar that aggregates observations in buckets.

• ROC curve. This is a graphical plot illustrating the performance of a binary classifier system as its discrimination threshold varies.

• Root Mean Square Error. Also know as RMSE is used to compute goodness of fit. It is the square root of the absolute value of the correlation coefficient between true values and predicted values.

• For further information, please read https://www.datasciencecentral.com/profiles/blogs/7-important-model-evaluation-error-metrics-everyone-should-know

### These different criteria will allow you systematically gauge the effectiveness of your data models. Always use what you understand and feel the most confident and comfortable with.

### Summary

[Summary](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774465_1)

During this lecture you examined the calibration process and the kind of datasets you need to consider using for calibrating your models. You walked through how to evaluate your data structures to best solve your problem statement using criteria such as terminology, intuitiveness, and having a connection to the business. You stepped through the different calibration techniques to increase accuracy and usefulness of you models. Such techniques include Bayes and regression to predict values, categories, and classes. You saw how to select the best software tools to identify relationships in your datasets by considering your level of expertise in statistics, the tool, and the complexity of the problem you need to solve.

Finally, you assessed dataset to specify data model assumptions and constraints to for probability distributions, combining variables, and when different dependent variables may rely on the same independent variables. You examined how to identify selection criteria to create evaluate, select, train, validate, and test your data models. You dove deep into the kinds of data sets you need to have to develop an effective data model. To do a better job with comparing models and selecting models, you walked through the common measures used on data model selection score card. A few of these measures are AIC, AICc, BIC, RIC, Confidence Intervals, Confusion Matrices, Gain and Lift Charts, Kolmogorov-Smirnov Charts, Chi Squares, ROC curves, and Root Mean Square Error.

# Week 7 - Exploratory Data Analysis

Top of Form

Bottom of Form

## Content

### Overview

[Overview](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)

|  |  |  |  |
| --- | --- | --- | --- |
| During this lecture you see how to explain the business drivers, insights, and metrics for analytics problem statements and how to identify and use problem solving approaches to solve analytical problems. You will walk through how to create data analytics scorecards to both document and communicate key findings.  Then you will see how to perform exploratory data analysis to identify the characteristics in data and determine data requirements to improve the validity of your dataset. You will also walk through approaches to identify patterns and correlations in data using statistical analysis.  By the end of this lecture you will have demonstrated the ability to:   * + Explain the key business drivers (i.e. insights, business performance metrics, etc.) for the analytics problem statements.   + Identify available problem solving approaches to help solve the analytics problem.   + Create a data analytics scorecard to document and communicate key findings.   + Perform exploratory data analysis to identify main characteristics within the data.   + Determine preliminary data manipulations requirements (i.e. harmonize, rescale and/or clean) to increase the validity in the dataset.   + Perform advanced statistical analysis to identify data patterns and correlations.  Balanced Scorecards [Balanced Scorecards](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Balanced scorecards are performance metrics used for strategic management to identify problems, track progress, and help determine when management should intervene.  While watching consider the following:   * Where did the balanced scorecard come from? * What perspectives are included in the balanced scorecard? * Why are you successful when you focus on three strategic objectives to influence change?  Business Drivers [Business Drivers](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  You use data to determine insights and findings that your organization can take advantage of. Insights and findings are the understanding of specific cause and effect chains within a given context. They can be information, an understanding, an introspection, or an observation.  **Business Drivers**  To better identify and understand findings, you should consider the business drivers in our organization. Business drivers are resources, processes, or conditions that that critical for the continued success and growth of our organization. Data, especially higher quality data lends itself well to certain business drivers. When you understand these drivers, you will better understand why the problem exists, allowing you to do an effective job of selecting business performance metrics, and crafting succinct problem statements so the problems are easier to analyze and solve.  In essence, business drivers are the key to your company’s competitive advantage – a more favorable business and market position than your competitors. Here are a few business drivers.   * • Strategic. High quality data helps you disseminate and use strategic information by enabling multiple uses for that information. * • Decision making. Better decisions come from more reliable and high quality data about your customers, products, services and operations. * • Regulatory compliance. High quality data helps you satisfy regulatory compliance by recording evidence. This evidence also allows you to better respond to evolving legislation. * • Business readiness. Data helps your company to be both aligned and prepared to evolve your products, services, and data to take advantage of business opportunities and emerging technologies. * • Business continuity. Data and its reliable management supports business continuity by reducing the risk of any disruptions your company may face. * • Reducing risk. Data helps anticipate and identify threats to your finances, reputation, improper data access, and legal ramifications regarding regulatory compliance. * • Reduce costs. High quality data reduces costs by preventing the use of out-of-date or incorrect data that would lead to the wrong decisions and courses of action. * • Improving processes. Data helps improve processes for time-to-market, regulatory compliance, organizational efficiency, data management, technology, compliance monitoring, and vendor management. * • Improving systems. Data helps you better manage data movement and identify and replace existing systems with new ones. * • Human capital. Having better data helps reduce training costs and lower the impact of staff turnover by providing the data’s context, history, and origins. * • Communications. Data helps your company communicate knowledge, progress, and actions timely and consistently. * • Search and Discovery. Having more and better data for searches and discovery improves customer and employee satisfaction. Search and discovery allows you to conduct one search across multiple types of content available across your organization. Content can include word documents, spreadsheets, project plans, bios, web pages, videos, and audio files. With just one search, you discover.   The more familiar you are with your company’s business drivers, the more understanding you will have on why the problem exists and how to best solve it. Solving Problems [Solving Problems](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  You can be provided with all the information, tools and approaches to help you with data analysis. However, they are only as good as your skills in problem-solving.  **Solving Problems**  With strong problem-solving you gain the most benefit from your knowledge and expertise by translating your company’s business drivers, metrics, and problem statements into findings. Here are the typical problem-solving approaches you'll use.   * • Research Solutions. You research and use existing known problems and their known solutions. Example: You what to know how a customer feels about your company’s product. You research the internet and uncover customer sentiment algorithms commonly used to solve your problem. You then reuse the solution. * • Simplify Problems. You reduce the complexity of your problem to the root problem so it's easier to solve. This may involve replacing aspects of the problem with a different problem that is much easier to solve. Example: You want to predict purchases customers will make at a supermarket chain. Rather than solving the problem by using a detailed analysis on your customer’s buying behaviour, you focus on a small number of products and what items the product is most likely sold with. * • Simplify Models. You reduce the complexity of your model to make it easier to solve your problem. Example: You see that your model uses variables that don’t significantly improve the accuracy of your model’s predictions. You decide to remove them because the extra processing power is just not worth it. * • Combine Models. A model may address only a part of the problem. Use multiple models to address the whole problem. Example: Your problem is to predict the likelihood of your customers buying both products A and B. That is difficult to solve. However, you have two other models. One model is good at predicting product A purchases and the other is good at predicting product B purchases. * • Brute Force. You test all possible solutions until you determine which is the right solution. Example: You prepare three different models. You methodically go through each model and combinations of models to determine what is the best model to use. You methodically go through repetitive steps that could last several hours to several days. “Success is 10% inspiration and 90% perspiration.” – Thomas Edison. * • Division. You break up your problem into multiple components and address each component individually. This is also known as "divide and conquer". Example: Your problem is to predict the likelihood of your customers buying both products A and B. That is difficult to solve. However, you can split up your problem into two problems. When will customers buy product A; and when will customers buy product B.   1. • Elimination. You eliminate aspects of the problem that make the problem unsolvable and focus only on the aspects of the problem that can be solved. • Call the customer -> Not useful for prediction   2. • Meet the customer -> 10% of meetings translate to sales   3. • Get the order -> Actual sale. Prediction not needed   4. • Customer pays -> Actual sale. Prediction not needed   Example: You want to predict how much product your customer will buy next year. This is extremely difficult to determine using behavioural analytics and trends analysis. It is much easier to look at the whole process of selling to a customer then eliminate what can’t be used to predict sales.   * • So you would eliminate all parts of the process except for “meet with the customer” and use it to predict the number of products your customer will buy.   When solving problems, keep it simple. Data Analytics Scorecard **Data Analytics Scorecards**  Before diving into scorecards, think a little bit about prediction and how prediction can even be possible. Is it possible to predict something in an uncertain future? In providing a loan to a home-buyer predictions can be made on borrowers. Such a prediction could look like this: borrowers will default 17% of the time with high risk loans while they will default 3% of the time with low risk loans. So predicting a least part of the future is very possible. And to predict the future, you use analytics.  Analytics is the discovery, interpretation, and communication of patterns found in the data. Your organization uses analytics to describe, predict, and improve business performance and business planning by uncovering findings. A subset of analytics is predictive analytics that analyzes current and historical facts to make predictions about the future, current, or past otherwise unknown events. Analytics in all their forms are used to help identify and assess risks and opportunities. And this guides current and future courses of action. Be cautious with your terminology. Metrics, key performance indicators, and critical success factors, though similar mean very different things.  **Metric.** This is just a measure.  Example: Time it takes to complete an activity.  **Key Performance Indicator.** This is a measure that quantifies management objectives used to measure strategic performance. They have targets.  Example: Number of new customers with a target of 10 customers to be acquired in 10 weeks.  **Critical Success Factor.** This is an element, not a measure that is necessary for the organization to be successful at achieving its mission.  Example: Employee engagement  Let's switch gears to scorecards. A scorecard, also known as a balanced scorecard, is a strategy performance management tool that uses key performance indicators to track execution and monitor the results of actions according to objectives.  To view an example of a Scorecard, please click here:  https://www.toolshero.com/wp-content/uploads/2017/10/balanced-scorecard-template-pdf-toolshero.pdf  • Scorecards focus on the strategic agenda of your organization  • Scorecards have a selection of a small number of data to monitor  • Scorecards are a mix of financial and non-financial data.  Scorecards differ from dashboards. Scorecards have a standard structure while dashboards provide "at-a-glance" view of key performance indicators relevant to objectives, business processes, or activities. Scorecards are for strategy performance while dashboards are for anything regarding progress.  To view an example of a Dashboard, please click here:  http://atlassian.wpengine.netdna-cdn.com/devtools/fisheye-source-crucible-code-review-dashboard.png  • Though dashboards do not have a standard, scorecards do. Scorecards consist of the following sections.   * • Perspectives. These perspectives include Financial, Customer, Internal Processes, and Learning & Growth. * • Objectives. These are verb-noun phrases specified in the strategy plan. Each objective is included under a perspective * • Measures. These are Key Performance Indicators * • Traffic light Indicators. These are red, yellow, or green symbols or colours providing a measure’s performance.   To develop a good scorecard, you focus on the objectives of each perspective in your scorecard. Here’s what you focus on for each of your scorecard’s perspectives. The details of this section may remind you of operation excellence which was covered in an earlier lesson.   * • Financial. You focus on high-level financial measures. "How do we look to shareholders?"   Examples: cash flow, sales growth, operating income, return on equity.   * • Customer. You focus on measures that answer "What is important to our customers and stakeholders?"   Examples: percent of sales from new products, on-time delivery, share of high tier customers’ purchases, ranking by high tier customers.   * • Internal business processes. You focus on measures that answer "What must our company excel at?"   Examples: cycle time, unit cost, yield, new product introductions.   * • Learning and growth. You focus on measures that answer "How can we continue to improve, create value and innovate?"   Examples: development time for new product generation, life cycle to product maturity, time-to-market versus competition.  To develop an even stronger scorecard, you would establish strategic themes that weave their way through each of the perspectives that tie together objectives that impact each other. When you hear organization’s use the term “alignment”, this is what they mean.   * • Example: Your organization wants to reinvent its offering and its image. It wants to move away from being perceived as being traditional and be seen as exciting, innovative, and an industry leader. You decide to work an “innovation” theme into your scorecard across all four perspectives. You adjust the already existing objectives as follows. * • Financial: Increase revenue from core products-   -> Increase revenue from innovations in core products   * • Customer: Introduce new software apps to meet demand    Introduce innovative software apps to meet demand   * • Internal Business Processes: Reduce costs in core processes    Reduce costs in core process by identifying innovative opportunities   * • Learning and Growth: Increase productivity of talent pool    Increase productivity and innovation competencies of talent pool   * • Now your objectives are aligned by the theme “innovation” using the words “innovation” and “innovative”  For further information on scorecards read Effective Performance Management with the Balanced Scorecard Technical Report and Measuring Performance: Seven Good Reasons to Use a Scorecard. **How to Present Scorecards**  [How to Present Scorecards](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Though scorecards have a standard structure, people interpret scorecards in many different ways. Primarily scorecards are for tracking and monitoring progress, but how people look at the sections and tie them together can vary from person to person. What is different for each person are the problems that they have. Whatever you come up to remedy the problem, this remedy must be accepted by your colleagues and your organization. This means you can't just be looking at problems, models, and datasets. You need to consider the behaviour and culture of those in your organization to tailor your findings and solutions. Here are a few things you should consider when providing any information about your findings.   * Use. How will the information you're providing be used? * Impact. Who will be impacted by what the information you've providing? * History. What happened when others provided that same information? * Politics. How will influencers react to your information? * Win-win. What information can you also include so most, if not all will benefit?   Never narrow your focus to just solving "the problem". You need to consider people first and ensure that you are really solving "their problem", not just the one specified in your analysis. Exploratory Data Analysis [Exploratory Data Analysis](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Exploratory data analysis is the first step where you examine data to answer questions and guide decisions when you don’t have all the information.  [Hide Course Menu](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1) Menu Management Options  * [Refresh](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1) * [Display Course Menu in a Window](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Course Menu:[22W Issues and Challenges in AI - 01 (AIDI1004-22W-20855)](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  * [Announcements](https://gc.blackboard.com/webapps/blackboard/content/launchLink.jsp?course_id=_331384_1&tool_id=_132_1&tool_type=TOOL&mode=view&mode=reset) * [Course Information](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6589022_1&mode=reset) * [Faculty Information](https://gc.blackboard.com/webapps/blackboard/content/launchLink.jsp?course_id=_331384_1&tool_id=_180_1&tool_type=TOOL&mode=view&mode=reset) * [Weekly Learning](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6589023_1&mode=reset) * [Assignments & Tests](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6589021_1&mode=reset) * [Tools](https://gc.blackboard.com/webapps/blackboard/content/launchLink.jsp?course_id=_331384_1&tool_id=_11_1&tool_type=TOOL&mode=view&mode=reset) * [My Grades](https://gc.blackboard.com/webapps/blackboard/content/launchLink.jsp?course_id=_331384_1&tool_id=_195_1&tool_type=TOOL&mode=view&mode=reset) * [Faculty Feedback Survey](https://gc.blackboard.com/webapps/blackboard/content/launchLink.jsp?course_id=_331384_1&tool_id=_3313_1&tool_type=TOOL&mode=view&mode=reset) * [Student Help & Support](https://help.blackboard.com/Learn/Student) * [Student Help & Support](https://help.blackboard.com/Learn/Student)  Week 7 - Exploratory Data Analysis Top of Form  Bottom of Form ContentOverview [Overview](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)   |  |  | | --- | --- | | During this lecture you see how to explain the business drivers, insights, and metrics for analytics problem statements and how to identify and use problem solving approaches to solve analytical problems. You will walk through how to create data analytics scorecards to both document and communicate key findings.  Then you will see how to perform exploratory data analysis to identify the characteristics in data and determine data requirements to improve the validity of your dataset. You will also walk through approaches to identify patterns and correlations in data using statistical analysis.  By the end of this lecture you will have demonstrated the ability to:   * + Explain the key business drivers (i.e. insights, business performance metrics, etc.) for the analytics problem statements.   + Identify available problem solving approaches to help solve the analytics problem.   + Create a data analytics scorecard to document and communicate key findings.   + Perform exploratory data analysis to identify main characteristics within the data.   + Determine preliminary data manipulations requirements (i.e. harmonize, rescale and/or clean) to increase the validity in the dataset.   + Perform advanced statistical analysis to identify data patterns and correlations.   Suggested Readings:   * + [Effective Performance Management with the Balanced Scorecard Technical Report](http://www.cimaglobal.com/Documents/ImportedDocuments/Tech_rept_Effective_Performance_Mgt_with_Balanced_Scd_July_2005.pdf)   + [Measuring Performance: Seven Good Reasons to Use a Scorecard](http://www.measurementinternational.com/pdfs/SevenReasonsforScorecard.pdf)   + [8+ Sample Scorecards](https://www.sampletemplates.com/business-templates/scorecard-template.html)   + [Principles and Procedures of Exploratory Data Analysis](http://brainimaging.waisman.wisc.edu/~perlman/papers/HumanFactors/10%20behrens%20principles%20and%20procedures%20of%20exploratory%201997.pdf) | Overview of Exploratory Data Analysis With Haberman Dataset – Towards AI —  The Best of Tech, Science, and Engineering  Photo on [Towards AI](https://towardsai.net/p/data-analysis/overview-of-exploratory-data-analysis-with-haberman-dataset) |  Balanced Scorecards [Balanced Scorecards](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Balanced scorecards are performance metrics used for strategic management to identify problems, track progress, and help determine when management should intervene.  While watching consider the following:   * + Where did the balanced scorecard come from?   + What perspectives are included in the balanced scorecard?   + Why are you successful when you focus on three strategic objectives to influence change?  Business Drivers [Business Drivers](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  You use data to determine insights and findings that your organization can take advantage of. Insights and findings are the understanding of specific cause and effect chains within a given context. They can be information, an understanding, an introspection, or an observation.  To read more, click [here](https://georgiancollege-my.sharepoint.com/:b:/g/personal/katie_tiwari_georgiancollege_ca/Ef_ulYxwPRtIp8jVgYJuhMYBgJPenuPFQcN05VQ1AEaU6A?e=y4fho3) Solving Problems [Solving Problems](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  You can be provided with all the information, tools and approaches to help you with data analysis. However, they are only as good as your skills in problem-solving.  To read more, click [here](https://georgiancollege-my.sharepoint.com/:b:/g/personal/katie_tiwari_georgiancollege_ca/EdLn6jM2q_RDva5H1aBgnEIBeISq_kxB8DGVQ2Jv4Z-S7w?e=f5NzUN) Data Analytics Scorecard [Data Analytics Scorecard](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  To read more, click [here](https://georgiancollege-my.sharepoint.com/:b:/g/personal/katie_tiwari_georgiancollege_ca/EZM9N2Etp99Dlvp3JMdgDysBrnY9w3V0vG624kmgjeOz_A?e=b66hzo) How to Present Scorecards [How to Present Scorecards](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Though scorecards have a standard structure, people interpret scorecards in many different ways. Primarily scorecards are for tracking and monitoring progress, but how people look at the sections and tie them together can vary from person to person. What is different for each person are the problems that they have. Whatever you come up to remedy the problem, this remedy must be accepted by your colleagues and your organization. This means you can't just be looking at problems, models, and datasets. You need to consider the behaviour and culture of those in your organization to tailor your findings and solutions. Here are a few things you should consider when providing any information about your findings.   * + Use. How will the information you're providing be used?   + Impact. Who will be impacted by what the information you've providing?   + History. What happened when others provided that same information?   + Politics. How will influencers react to your information?   + Win-win. What information can you also include so most, if not all will benefit?   Never narrow your focus to just solving "the problem". You need to consider people first and ensure that you are really solving "their problem", not just the one specified in your analysis. Exploratory Data Analysis [Exploratory Data Analysis](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Exploratory data analysis is the first step where you examine data to answer questions and guide decisions when you don’t have all the information.  While watching consider the following:   * + What is the main purpose of exploratory data analysis?   + What five things does exploratory data analysis help you do?   + What three things can data mining help you accomplish?  Exploratory Data Analysis [Exploratory Data Analysis](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Earlier you were exposed to exploratory data analysis, known as EDA for short. In this part, you will go deeper into what EDA is and its formal approaches.  **Exploratory Data Analysis**  EDA is used to analyze datasets to explore the dataset's main characteristics using visual methods complemented with statistical analysis. Data mining which is the automatic extraction of useful information and relationships from vast amounts of data is considered a subset of EDA. Often people think that EDA doesn't involve models or statistics, and that is true up to a point. It's not mandatory that you use models and statistical methods with EDA, but EDA is much more effective when you do. Many advocate that using visual graphs based on simple models and probability distributions are the best path for you to take to determine the best models to use. With them, you will uncover information on insights, variables, anomalies, assumptions, and structures.   * • Insights – Determine causes and relationships based on observations * • Variables – Determine most impactful variables * • Anomalies – Determine which data points are outliers * • Assumptions – Determine assumptions for statistical analysis and which tools to use * • Structure – Determine when further data collection and data manipulation is needed   You must graph and display the data in your dataset to better understand your data. There is no substitute for seeing the data from an eagle eye's view. You would do this by plotting the data on various graphs. You then would extract the patterns based on what you are seeing. The graphing you should use first in EDA are:  **Box Plot**  Box plots are a quick way to examine a variable and determine its shape and center is one view. Box plots answer these questions.   * • Is a variable significant? * • Does the location (shift in distribution) differ between subgroups? * • Does the variation (distributes around the mean) differ between subgroups? * • Are there any outliers? * • Are there data points that don’t follow the rules?  |  | | --- | | *Plotly (n.d.). BoxPlot Fig 9. Retrieved June 1, 2018 from https://plot.ly/static/img/literacy/boxplot/boxplotfig9.jpg* |  Statistical Analysis **Statistical Analysis**  Once you determine your plots graphs, your next step is to determine the characteristics in your data. Here are the main characteristics you can determine about your data by looking at graphs.   * • Distribution. This is the summary of frequency of individual values and ranges for a variable. * • Central Tendency. This is the estimate of the center of a distribution. Mean (average), median (middle value), and mode (most frequently occurring value). * • Dispersion. This is the scatterness of observations from the average.   Once you determined these characteristics, you can use them to determine the type of statistical model you can use, either using univariate or multivariate equations.   * • Univariate means depending on one variable. * • Multivariate means depending on more than one variable.   To know which statistical models to use, you would examine the data, asking the following questions. Is there a distribution? How many centers do your distribution or distributions have? You do a first pass at distributions by determining what kind of distributions for the variables you are seeing. Are you seeing any distributions, univariate distributions or multivariate distributions?  When you see any probability distributions, you use a parametric model.   * • When you have only two univariate distributions, you would use t-Test. For more information on the t-Test, go to http://docs.statwing.com/examples-and-definitions/t-test/ * • When you have two or more multivariate distributions, you would use ANOVA. For more information on ANOVA, go to https://explorable.com/anova * • When you have two or more univariate distributions, you would use MANOVA. For more information, go to http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova1.pdf   When you see no probability distributions, you use a non-parametric model.   * • When you have only two univariate equations, you would use Mann-Whitney U Tests. For further information, go to http://www.statisticssolutions.com/mann-whitney-u-test/ * • When you have two or more univariate equations, you would use Kruskal-Wallis Test. For further information, go to http://www.statisticssolutions.com/kruskal-wallis-test/ * • When you have two or more multivariate equations, you would use Median Test. For further information, go to http://davidmlane.com/hyperstat/viswanathan/Median\_Test.html   If it was decided that you use a non-parametric model, are the observed “counts” significantly different?  Determine if you see the counts as integers or as categories. In both cases you, would use the chi-squared test. For more information, go to https://explorable.com/chi-square-test  Is there a relationship between variables? Does one variable predict another? When you decided on using a parametric model.   * • When you have only two variables, you would use Pearson coefficient Test. For more information, go to https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php * • When the variables have a continuous relationship, you would use Linear regression For more information, go to http://www.statisticssolutions.com/what-is-linear-regression/ * • When the variables have a binary relationship, you would use Logistical regression. For further information, go to http://www.statisticssolutions.com/what-is-logistic-regression/   When you decided on using a non-parametric model, you would use a Rank correlation. For more information, go to https://www.medcalc.org/manual/rank\_correlation.php Once you have an idea of what your model could look like. You would create an ad hoc version of your model and use your dataset to plot it and your model’s predictions on a graph. You then compare your dataset plot graphs to your model’s prediction plot graphs. And compare the differences. Once you feel you better understand the data, you would train, validate, and calibrate your model so your model fits your data. But before you do that, you must examine your dataset to see if it needs to be manipulated so your dataset is more reliable. **Data Manipulation**  [Data Manipulation](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_331384_1&content_id=_6774480_1)  Data manipulation can be interpreted as the process of changing data in order to make it easier to read or be more organized.  **Data Manipulation**  All your visual graphs are driven by your data. Therefore, you must ensure your data is correct and right. When comparing your visuals together and looking for patterns, you'll have a good idea where the data points don't look right. To make the patterns more apparent, you will determine requirements for how to manipulate the data so the graphs are much easier to interpret. Here are your options.  **Cleaning**  This is where you examine the data and replace missing data and correct existing data through algorithms and manual manipulation. Here are a few things you should consider while cleaning your data.  **Missing Data.** Data is expected but not included in the dataset.  Examples: Blanks and missing records.  **Erroneous Values.** These are misspelled values, reversed digits, letters in a number field, or wrong values.  Examples: Ottowo instead of Ottawa; 321 instead of 123; and 22 instead of 40.  **Type Conversion.** Numbers are treated as categories and not integer values  Example: Values for number of products sold are treated as binary categories instead of range of integer values.  **Entity Resolution.** Having two different values for the same thing.  Example: Having two prices for the same product when there should be just one price.  **Valid Ranges.** Values for a variable fit within the constraints for that variable.  Examples: Income is above zero. An integer variable has values 1, 2, 3,... and not the value 1.5.  **Outliers.** Data points that don’t fit the pattern make sense.  **Harmonization**  This is improving the quality of your data using machine learning algorithms.  **Naïve Bayes Classifier Algorithm.** This allocates an element in a population to a specific category.  Example: Spam filter.  **K Means Clustering Algorithm.** This groups elements into clusters.  Example: Web search results on a keyword “Jaguar” comes up with Car, Mac OS, and Animal.  **Support Vector Machine Algorithm.** This categorizes data into classes by separating these classes by lines or hyperplanes. It doesn’t make strong assumptions and doesn’t overfit the data.  Example: Classifying product reviews.  **Apriori Algorithm.** This creates association rules between data. If A happens, then B will happen at a given probability.  Example: Predicting retail purchases based on what has already been purchased.  **Linear Regression Machine.** This shows the impact of changing variables.  Example: Predicting future demand of a product.  **Decision Trees.** This is a graphical representation of using branching to show all possible outcomes of a decision.  Example: Determine a restaurant recommendation.  **Random Forests Machine.** This uses a “bagging” approach create a bunch of decision trees and polling the results of each decision tree and providing the prediction the most common to all the decision trees.  Example: Determine a more accurate restaurant recommendation.  **Logistical Regression Machine.** This predicts the outcome of a categorical variable.  Example: Predict if there will be snowfall in Toronto tomorrow morning.  **Artificial Neural Networks.** This mimics the interconnections in the human brain to do complex cognitive activities.  Example: Recognizing faces. Netflix recommending videos to its customers.  **Nearest Neighbours.** This groups similar data points together. The more similar the data points, the closer these data points were.  Example: Recommender systems. For further information on these machine learning algorithms, go to https://www.dezyre.com/article/top-10-machine-learning-algorithms/202  **Rescaling**  You can rescale the data to improve its interpretability and make it easier to pull together a statistical model. You could use natural log transformations and inverse probability transformation to normalize distributions, square root transformation to address variances, and logarithmic transformation for linear trends.  For further information review How and why to harmonize, rescale and cleansing data located at https://stat.ethz.ch/education/semesters/as2015/analytics/Data\_Analytics\_2015-11-04.pdf  Always define your requirements and make sure they ensure your data is correct and right. With every dataset, run through the **10 Cs checklist.**   * • Completeness: all data fields are complete * • Correctness: data is accurate * • Consistency: data is consistent with field and concept definitions * • Currency: data is not obsolete * • Collaborative: data not an opinion but expert consensus * • Confidential: data is secure * • Clarity: data is legible and comprehensible * • Common Format: data is in an easily used format * • Convenient: data is conveniently accessible * • Cost-effective: cost of data collection and data use makes up for its costs of doing so  Play it safe by always anticipating problems with your data.Problems with Statistics During this lecture you learned about exploratory analysis, statistical analysis, and data manipulation. However, you will waste a lot of time and effort if you are not cautious with the techniques you use. Because statistics is at the center of exploratory data analysis and because statistics involves you interpreting the data, it is all too easy for you to go down the wrong road. Here are a few things you should consider with doing your statistical analysis.   * Incorrectly using statistical tests. * Not adjusting for multiple comparisons in the data. * Not defining and explaining statistical tests used. * Not doing sample size estimation. * Not using robust statistical techniques.   Make sure your conclusions involving your data doesn’t diminish the reliability, validity and the verifiability of your models and results. Summary During this lecture you reviewed business drivers such as strategy, business continuity, decision-making, regulatory, efficiencies, and reducing risks and costs. You explored the different problem solving approaches such as simplifying, combining, and dividing models and problems to better analyze and solve those problems. You saw what analytics is and the differences between metrics being just measures and key performance indicators being measures for strategic performance. You also saw the difference between scorecards and dashboards; though similar scorecards have a standardized structure while dashboards do not. You created a scorecard and explored what to consider when presenting your scorecard to an audience.  Then you saw how to perform exploratory data analysis using scatter plots, histograms, box plots, and probability plots. By examining these visuals you identify characteristics in the data such as distribution, central tendency, and dispersion. Based on that information you determine statistical analysis approaches to choose the kind of models you need and attempt to reproduce the patterns you visually see. And you walked through data manipulation and the approaches you can use so your data will be reliable and your statistical model viable. |