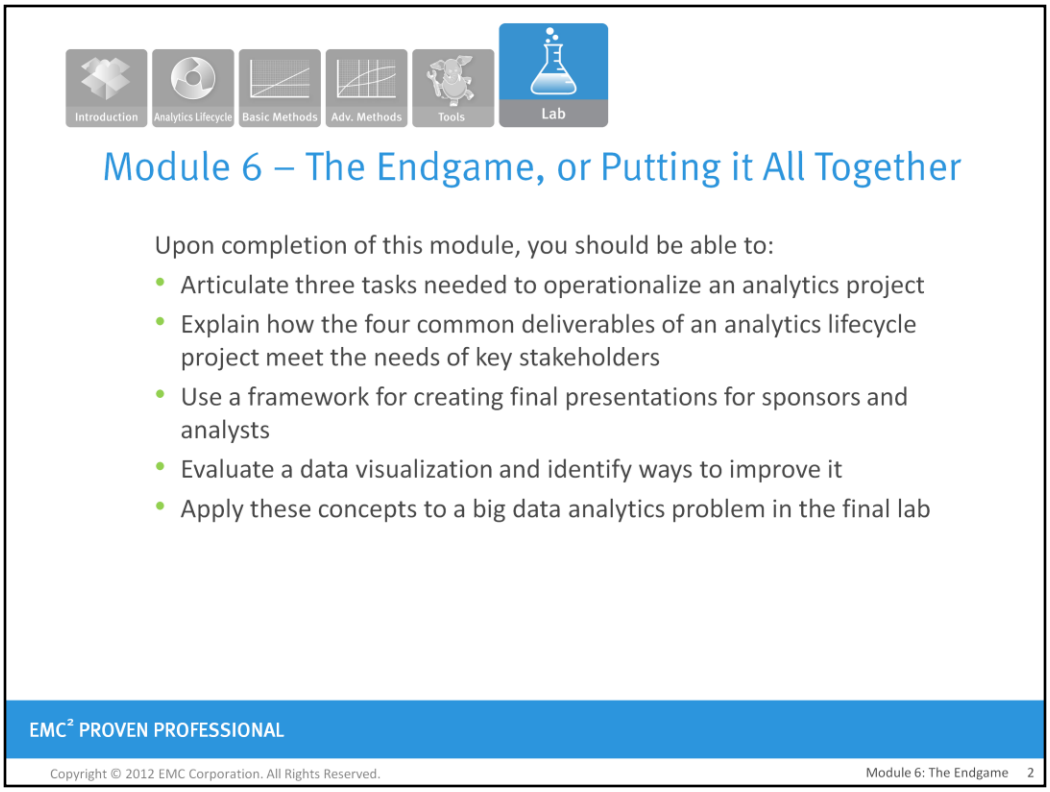
Lab

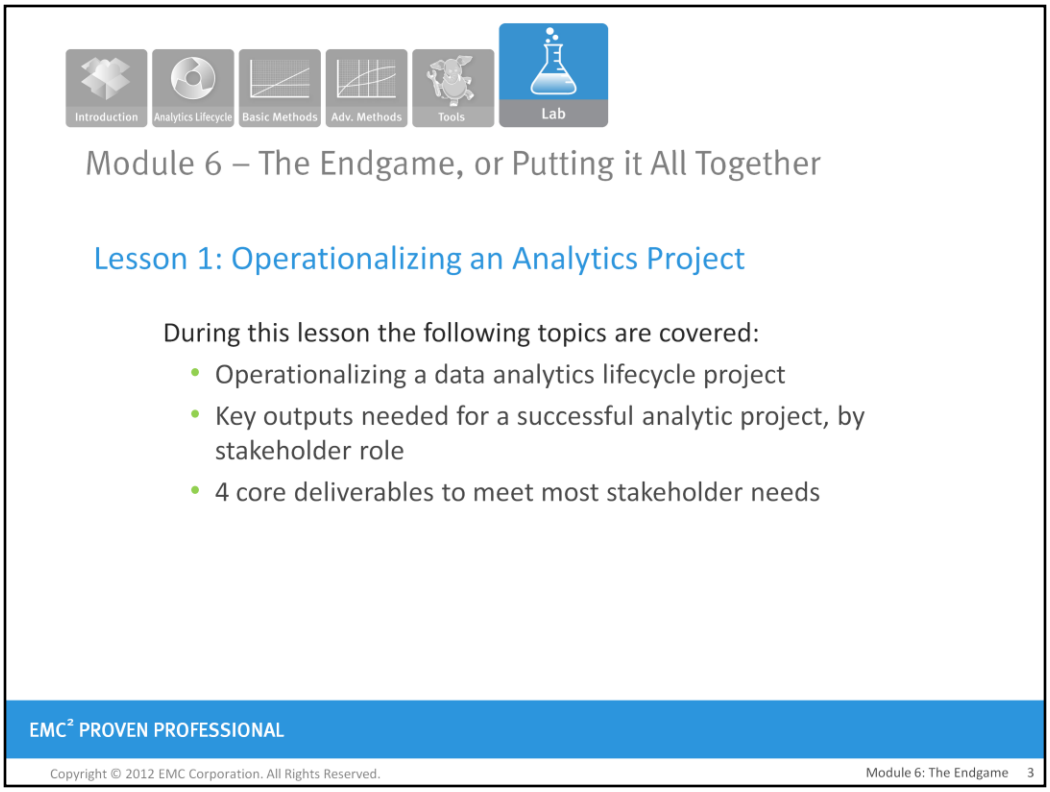
Module 6 – The Endgame, or Putting it All Together

EMC² PROVEN PROFESSIONAL

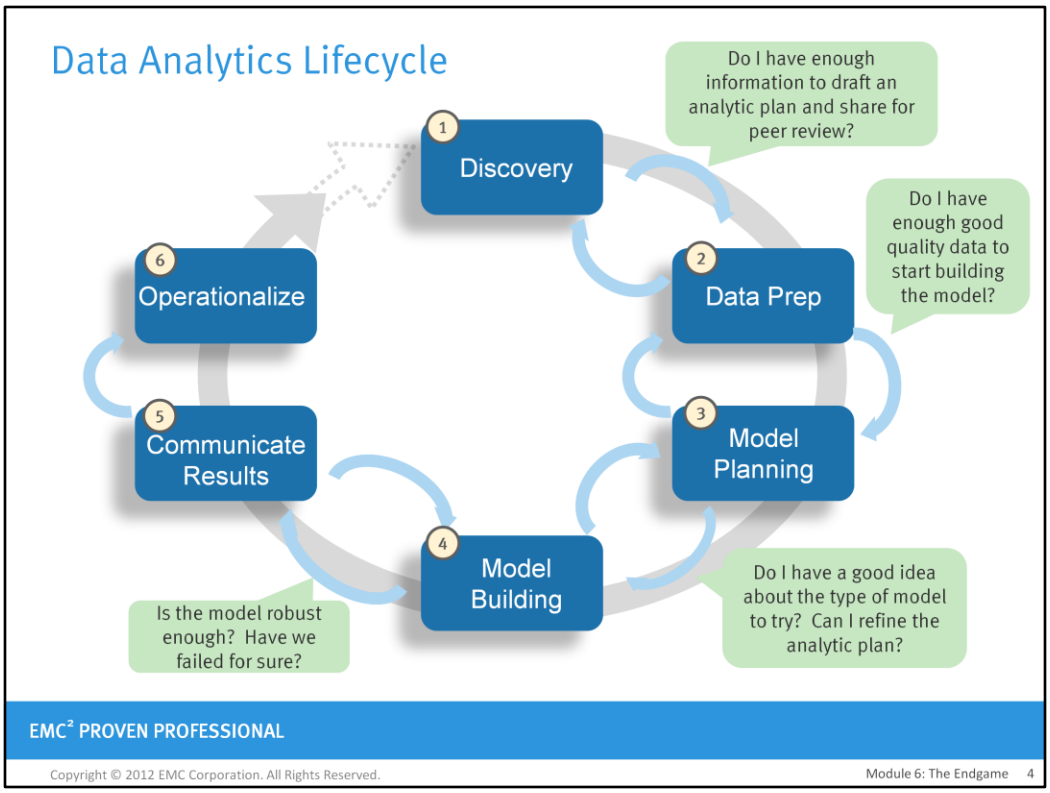


This module covers the following lessons:

- Operationalizing an analytics project
- Creating the final deliverables
- Data visualization techniques to support your final presentation
- Final lab on big data analytics

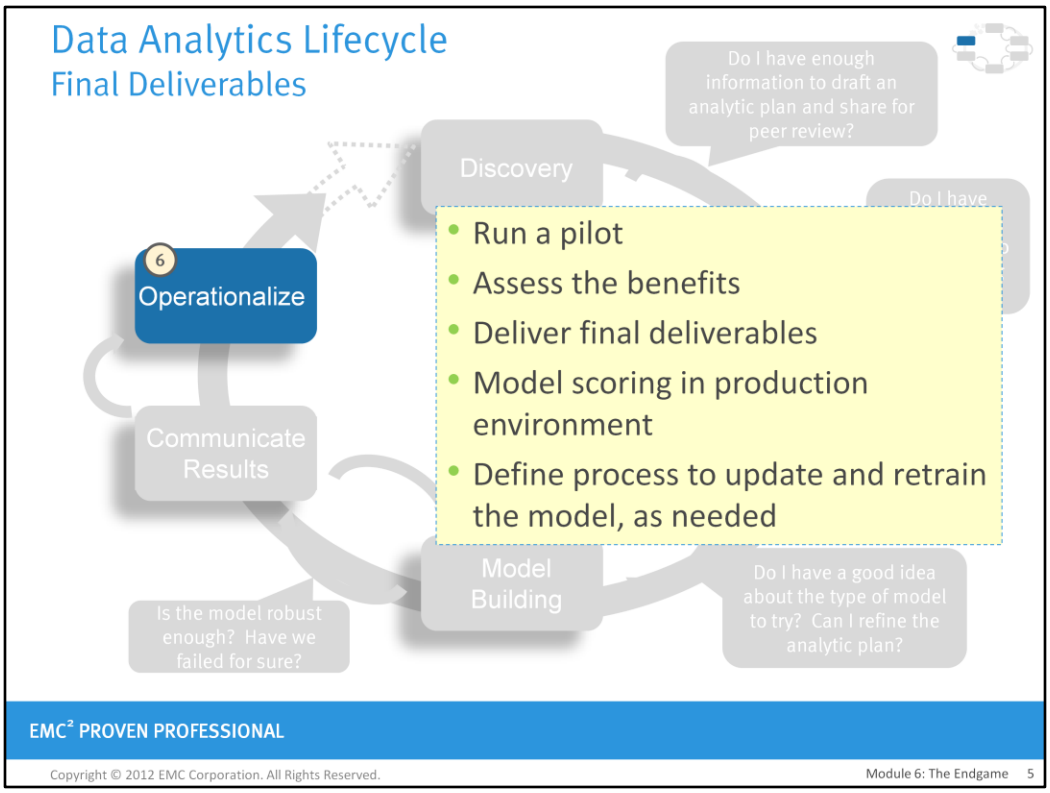


This lesson covers putting a data analytics lifecycle project into action, the outputs of a successful analytic project and the core deliverables.



This is a graphic portraying the data analytic lifecycle that we explored in Module 2. In this Endgame module, we will focus on the final phase of the cycle, "Operationalize". In this phase, the project team will deliver final reports, briefings, code, and technical documents. In addition, the conclusion of this phase includes running a pilot project, and implement your models in a production environment.

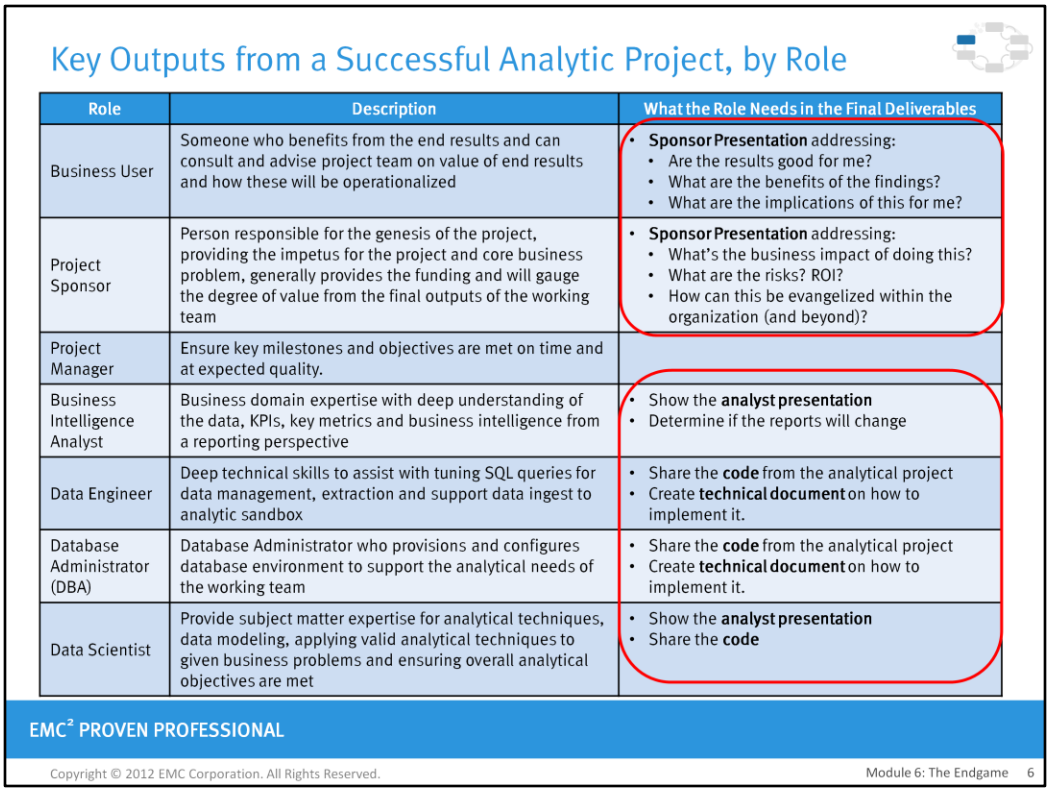
As stated in Module 2, you can perform a technically accurate analysis, but if you cannot translate the results into a language that speaks to the audience, people will not see the value and much of your time will have been wasted. For this reason, we will spend time in this End Game module to show you how to put together a clear narrative summary of the work and convey it to key stakeholders.



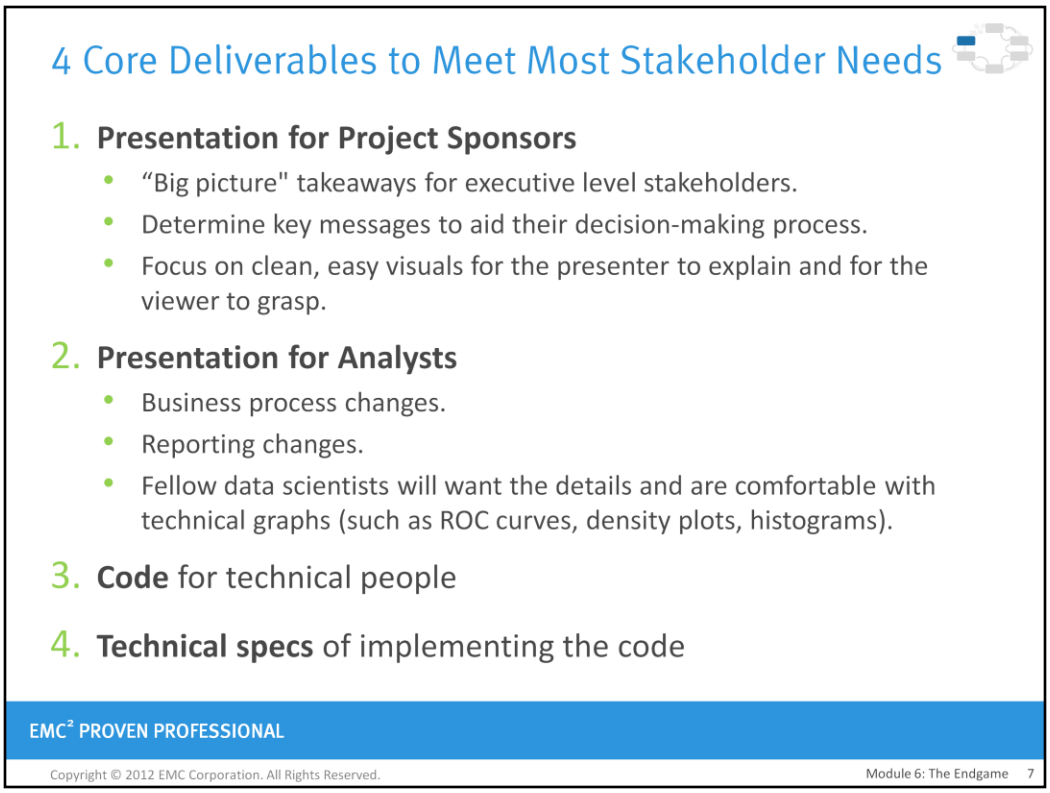
As mentioned in Module 2, the Data Analytic Lifecycle, the final phase of the lifecycle is focused on operationalizing the project. **In this phase, you will need to assess the benefits of the work that's been done, and setup a pilot so you can deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users.**

Your ability to quantify the benefits and share them in a compelling way to the stakeholders will determine if your work will move forward into a pilot phase and ultimately be run in a production environment. For this reason, it is critical that you identify the benefits and state them in a clear way in the final presentations. In the subsequent lesson, we will introduce you to a framework to share your work with the key stakeholders in a clear and concise way, to help illustrate the work that was done and share its potential value.

As you scope the effort of a pilot project for your model, **consider running the model in a product environment for a discrete set of single products, or a single line of business, which will test your model in a live setting**. This will allow you learn from the deployment, and make any needed adjustments before launching across the enterprise. Keep in mind that this phase can bring in a new set of team members – namely those engineers who are responsible for the production environment, who have a new set of issues and concerns. They want to ensure that running the model fits smoothly into the production environment and the model can be integrated into downstream processes. While executing the model in the production environment, **look to detect anomalies on inputs before they are fed to the model. Assess run times and gauge competition for resources with other processes in the production environment**.



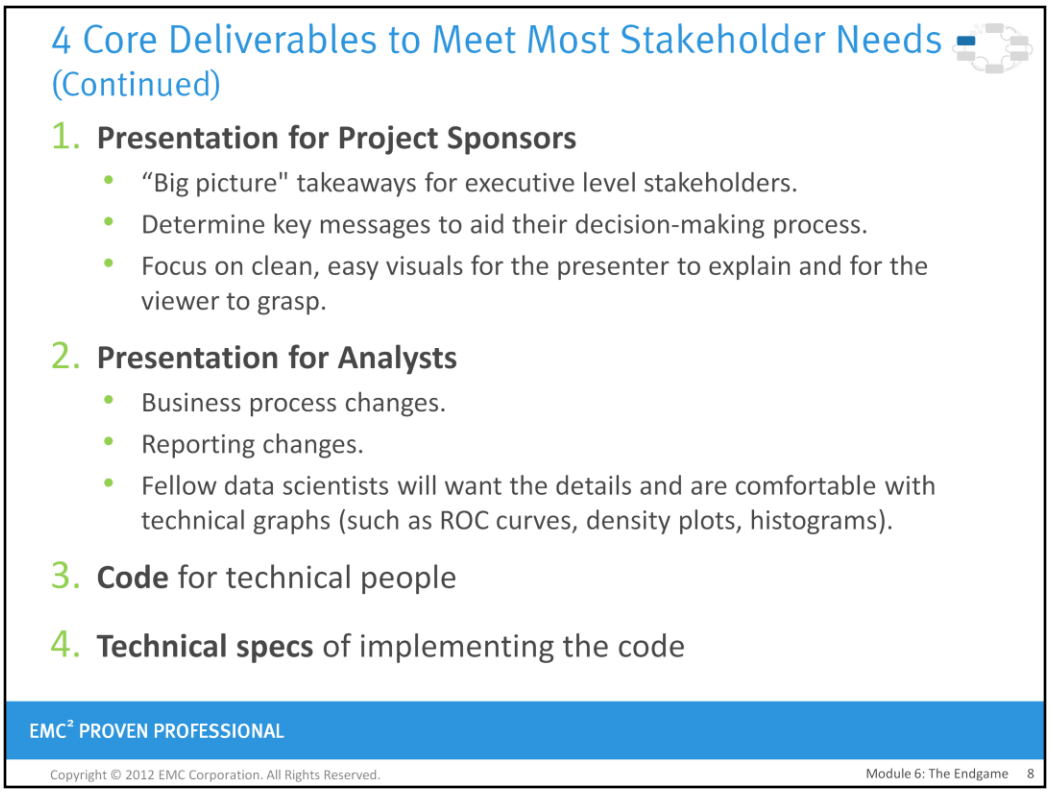
As you begin to frame the final deliverables, keep in mind the interests of each of the main stakeholders and be sure to frame your presentations in a way that will address their interests and concerns. Be prepared to discuss how your work will impact end users and others in the business.



We will provide further details on how to create the final deliverables in the next lesson. Here are a few general guidelines about preparing the results of the analysis for sharing with the key sponsors:

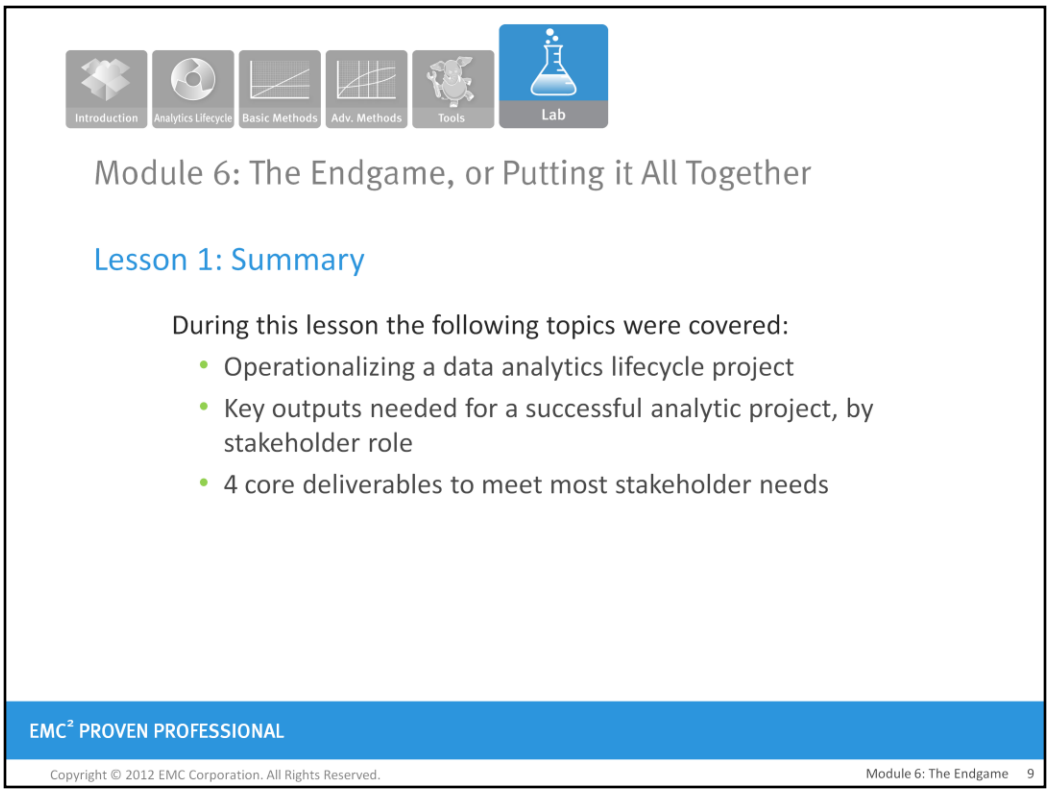
1) The more the audience is comprised of **executives**, the more succinct you will need to be. Most executive sponsors attend many briefings in the course of a day or a week. Ensure your presentation **gets to the point quickly and frames the results in terms of value to the sponsor's organization.** For instance, if you are working with a bank to analyze cases of credit card fraud, highlight the frequency of fraud, the number of cases in the last month or year, and how much of a cost or revenue impact there is to the bank (or focus on the reverse, how much more revenue they could gain if they address the fraud problem). This will demonstrate the business impact better than deep dives on the methodology. You will need to include supporting information about analytical methodology and data sources, but generally only as supporting detail or to ensure the audience has confidence in the approach you took to analyze the data.
2) If presenting to **other analysts**, focus more time on the **methodology and findings**. You can afford to be more expansive in describing the outcomes, methodology and the analytical experiment with a peer group, as they will be more interested in the techniques, especially if you developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems.
3) **Use imagery when possible**. People tend to remember mental pictures to demonstrate a point more than long lists of bullets.

<Continued>

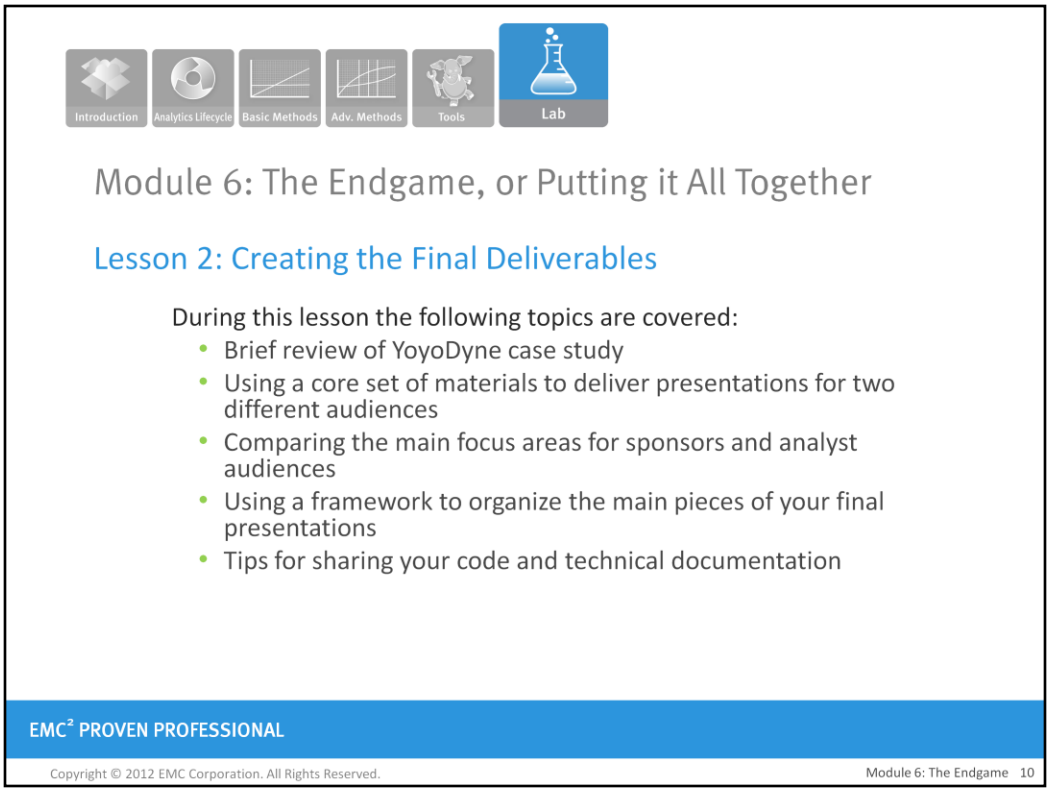


Additional references to learn more about best practices for giving presentations:

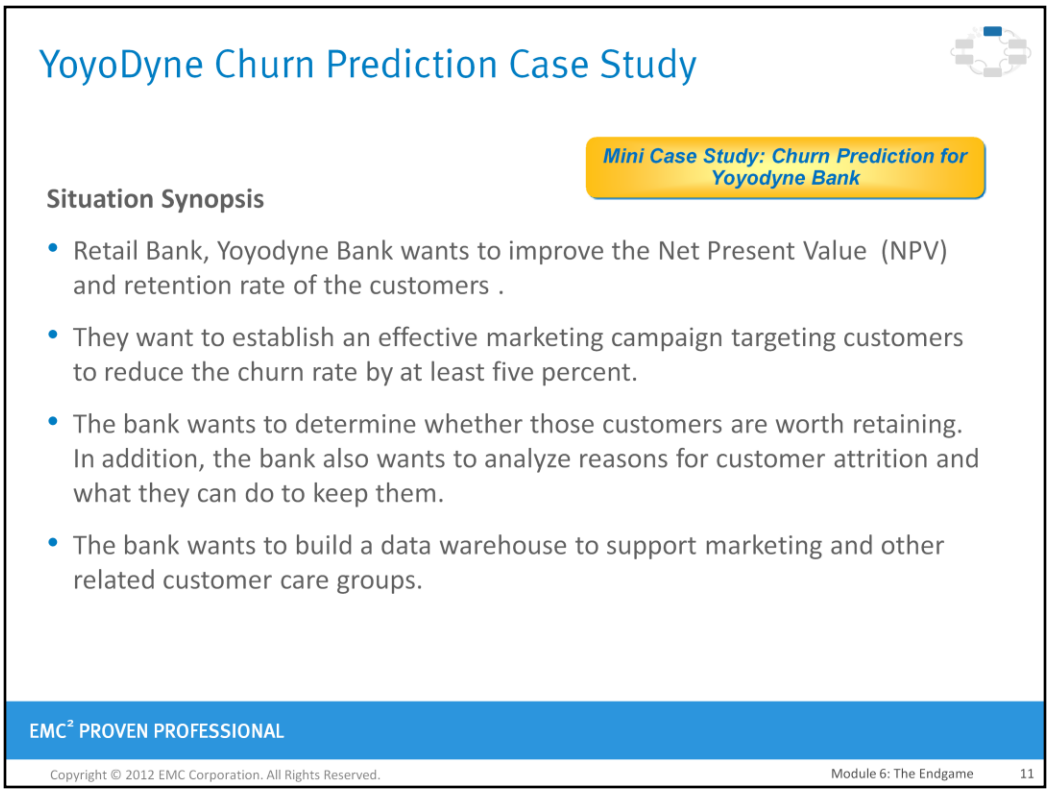
- Say It With Charts, by Gene Zelazny. Very simple reference book on how to select the right graphical approach for portraying data, and for ensuring your message is clearly conveyed in presentations.
- Pyramid Principle, by Barbara Minto. Minto pioneered the approach for constructing logical structures for presentations in threes. Three sections to the presentations, each with 3 main points. This will teach you how to weave a story out of the disparate pieces that emerge from your work.
- Presentation Zen, by Garr Reynolds. Teaches you how to convey ideas simply and clearly, use imagery in presentations, and shows many before and after versions of graphics and slides.



This lesson covered what is needed to put a data analytics lifecycle project into action, the outputs of a successful analytic project and the core deliverables.



This lesson covers presents and discusses a fictional case study.



Shown is a synopsis of the YoyoDyne Churn Prediction Case Study, which was introduced in Module 2. We will use this scenario throughout this lesson to show examples of how you would write a final narrative summary for the case study example.

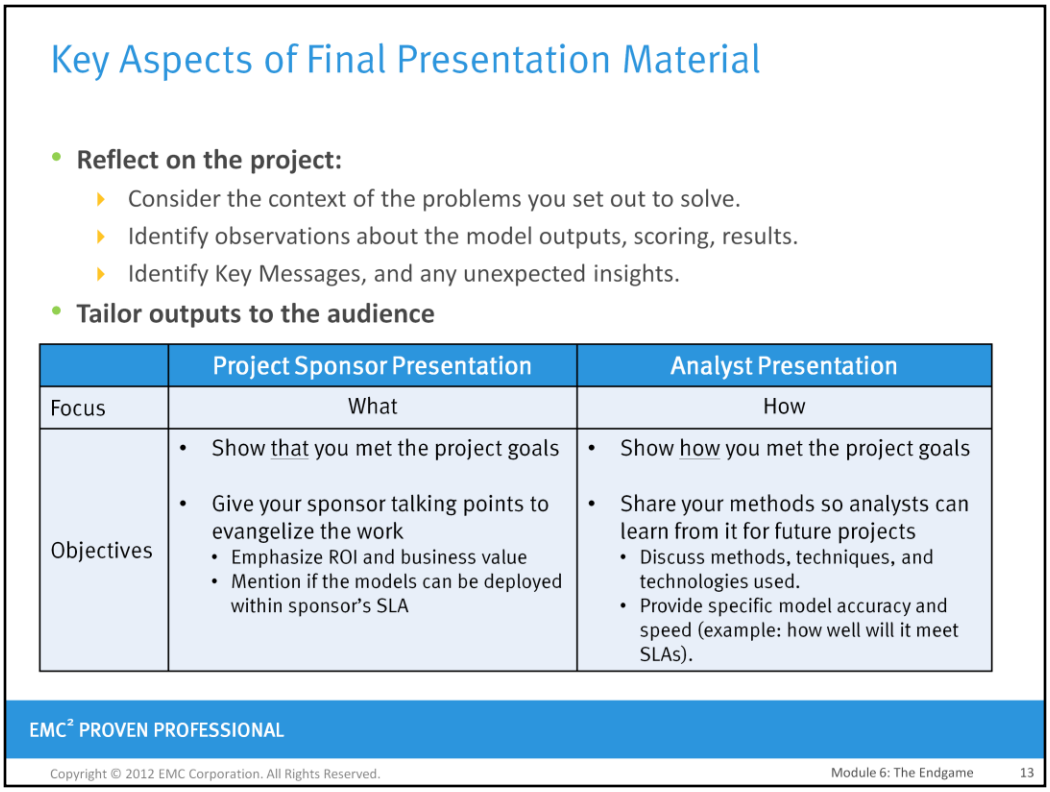
Use Analytic Plan to Guide Final Presentation

Mini Case Study: Churn Prediction for Retail Banking

Components of Analytic Plan	Retail Banking: Yoyodyne Bank
Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates
Data & Scope	5 months of customer account history
Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn
Result & Key Findings	Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save $3 million in lost customer revenue and avoid $1.5 million in new customer acquisition costs each year.

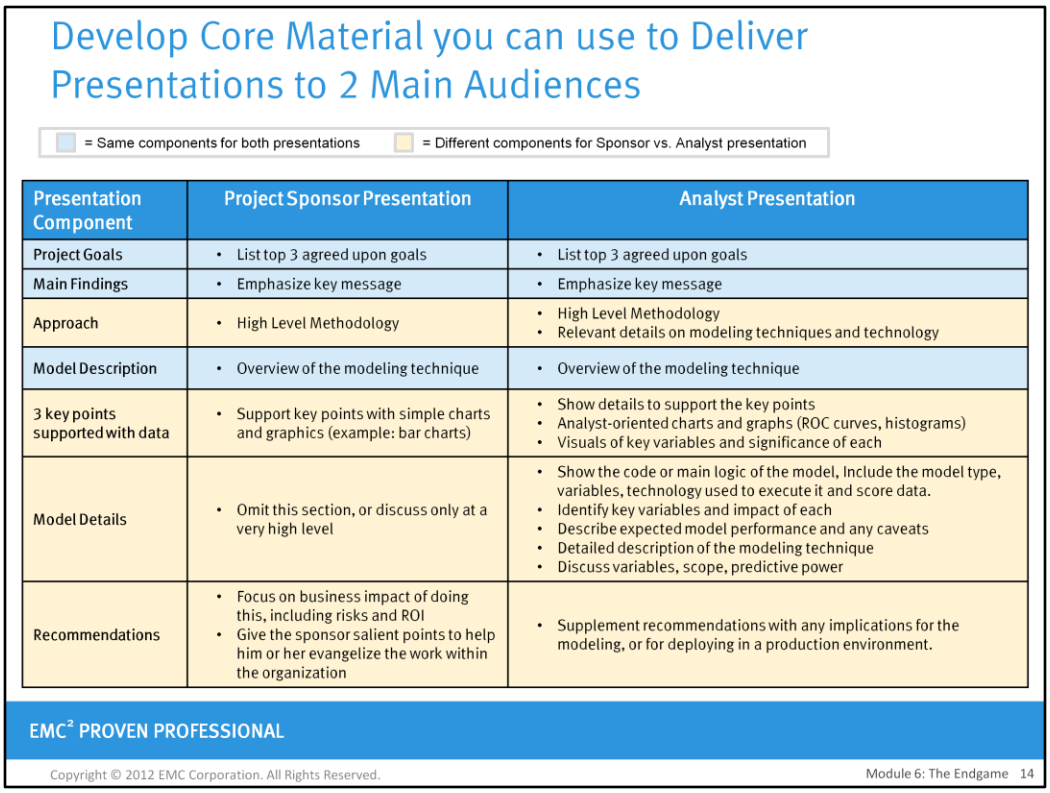
EMC² PROVEN PROFESSIONAL

This slide represents an outline of an analytic plan for the mini case study. In addition to guiding your model planning and methodology, the analytic plan can serve as a guide for the main points of the final presentation. Within the analytic plan are components that can be used as inputs for writing about the scope, underlying assumptions, modeling techniques, initial hypotheses, and key findings.

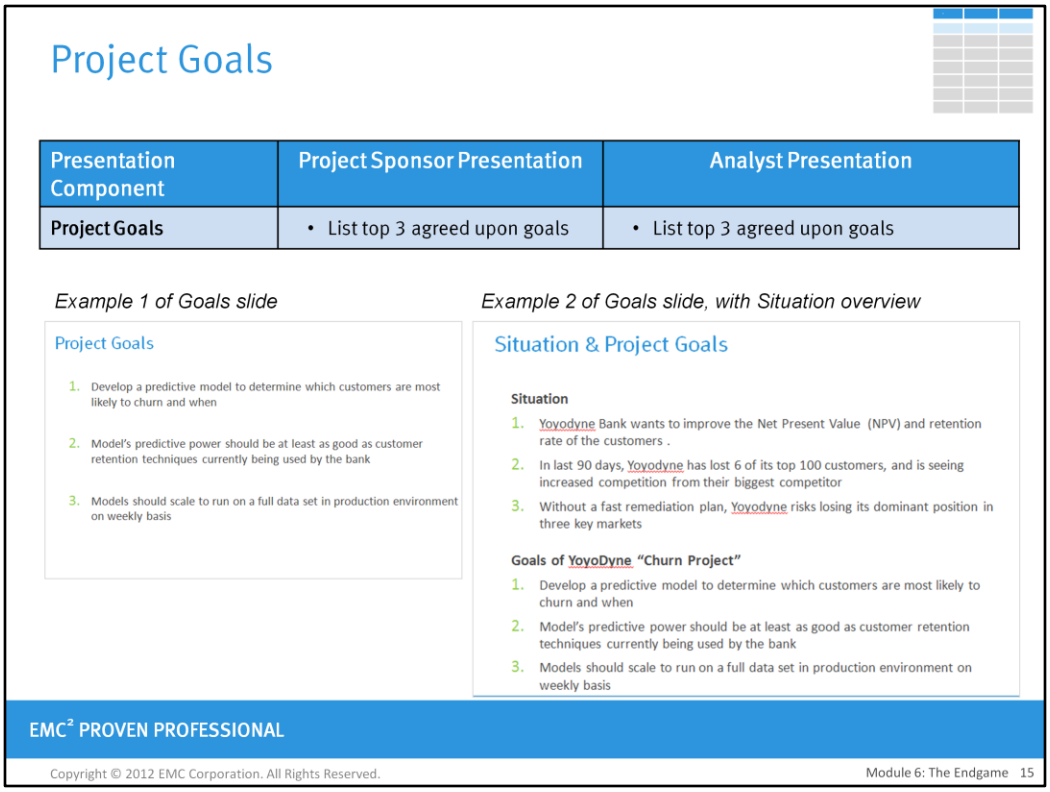


As shown, the **Sponsor presentation focuses on the "what**".....What did you do, what is the ROI and business value. **The Analyst deck focuses on the "how"**....how did you do it, how did you opt to solve the problem, etc.

Ideally, you should consider starting the development of the final presentation during the project, rather than at the very end. This approach will ensure that you always have a version of the presentation with working hypotheses to show stakeholders, in case you need to show a work in process (WIP) version on short notice. In fact, **many analysts write the executive summary at the outset of a project, then continually refine it over time** so that at the end of the project, pieces of the final presentation are already done. This approach also avoids forgetting key points or insights you may discover during the project, and reduces the amount of work you will need to do on the presentation at very end of the project.



Shown above are the main components of the final presentations for the project sponsor and an analyst audience. Notice that you can create a core set of material in these 7 areas, which can be used for the two presentation audiences. Three areas (Project Goals, Main Findings and Model Description), can be used as is for both presentations. Others areas need additional elaboration, such as the Approach. Other areas, such as the Key Points, require different levels of detail for the analysts and Data Scientists than for the project sponsor.

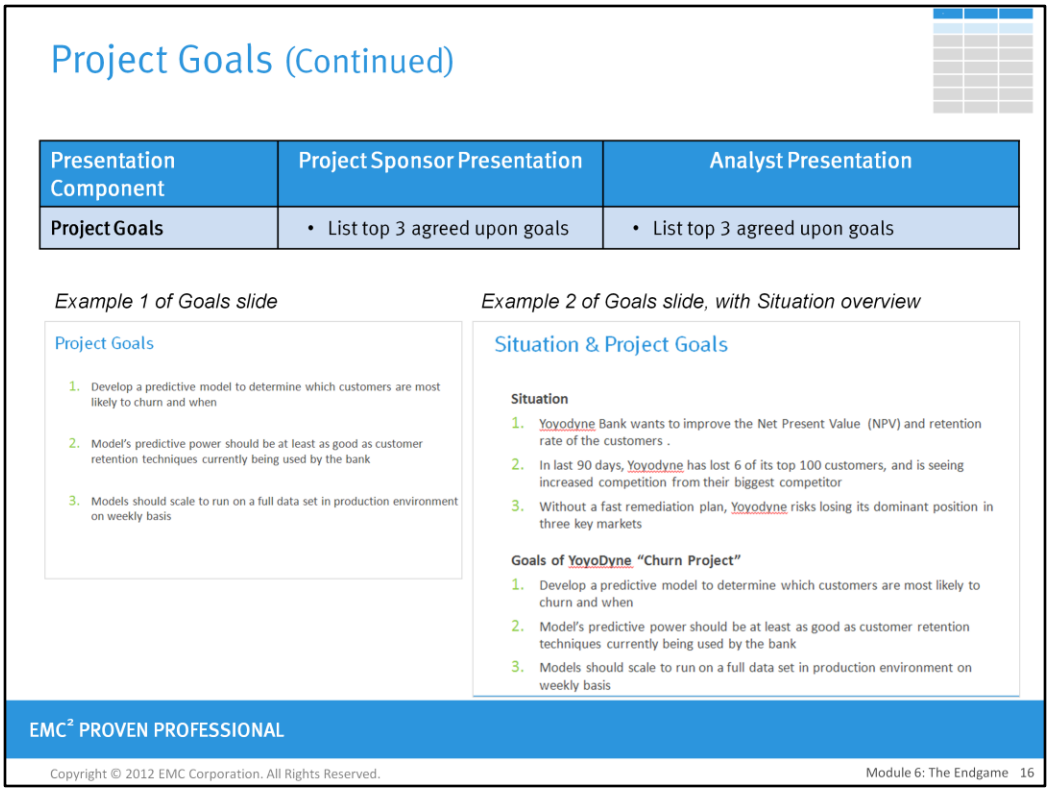


The Goals portion of the final presentation is generally the same, or very similar, for sponsors and for analysts. In each case, you will need to reiterate the goals of the project in order to lay the groundwork for the solution and recommendations that are shared later in the presentation. In addition, the Goals slide serves to ensure there is a shared understanding between the project team and the sponsors and confirm they are aligned in moving forward in the project. Generally, the goals are agreed on early in the project and it is good practice to write them down and share them to ensure the goals and objectives are clearly understood by both the project team and the sponsors.

The slide shows two examples of slides on Project Goals. **Example 1** shows three goals, describing the need to create a predictive model to anticipate customer churn and the expectation that the resulting model is at least as accurate as the current methods that YoyoDyne bank uses, and will be able to run in their production environment within the bank's SLA.

Example 2 shows a variation on the Project Goals slide. This shows a summary of the situation prior to listing the goals. Keep in mind that when delivering final presentations these deliverables get shared within organizations and the original context can be lost, especially if the original sponsor leaves or changes roles. For this reason, it is good practice to briefly recap the situation prior to showing the project goals. Adding a situation overview to the Goals slide does make it appear busier, so you will need to determine whether to split this into a separate slide or keep it together, depending on your audience and your style of delivering the final presentation.

<Continued>



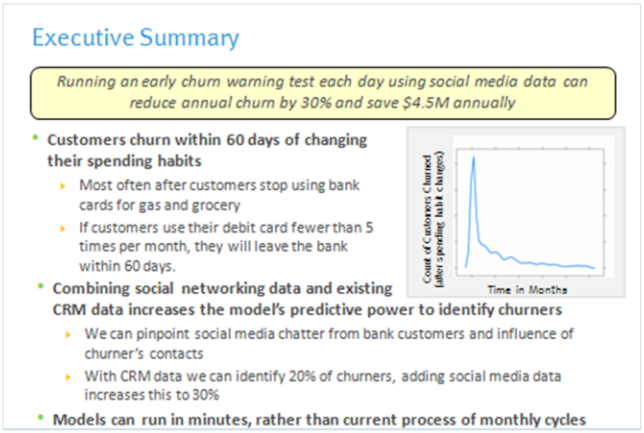
One method for writing the situational overview in a succinct way is to summarize it in three bullets as follows:

- **Situation:** Give a one-sentence overview of the situation that has led to the analytical project
- **Complication:** Give a one-sentence overview of the need for addressing this now. Something has triggered the organization to decide to take action at this time. For instance, they lost 100 customers in the last 2 weeks, they now have an executive mandate to address an issue. Or they have lost 5 points of market share to their biggest competitor in the last 3 months. Usually, this sentence represents the driver for why a particular project is being initiated at this point in time, rather than in some vague time in the future.
- **Implication:** One-sentence overview of the impact of the complication. For instance, if the bank doesn't address their customer attrition problem, they stand to lose their dominant market position in three key markets. Focus on the business impact to illustrate the urgency of doing the project.

Main Findings (Executive Summary)

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Main Findings (executive summary)	• Emphasize key message	• Emphasize key message

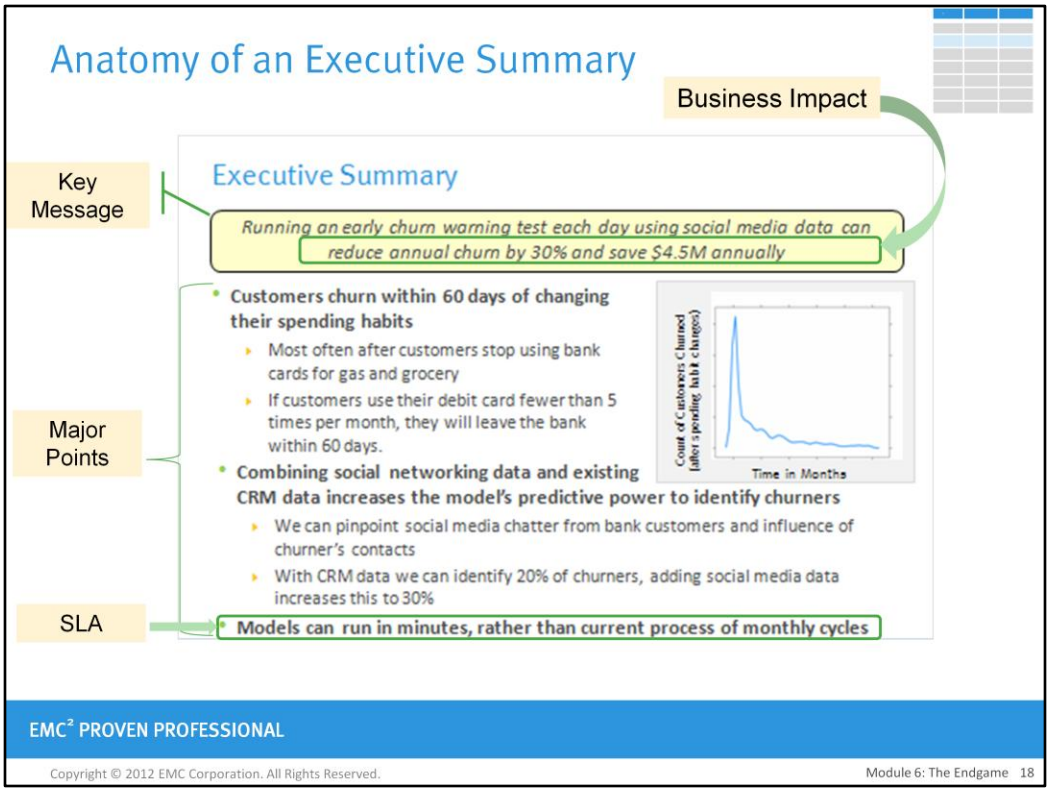
- Enable reader to grasp full synopsis in 1 slide
- Frame outcomes in terms of business value
- Generally same, or very similar, for both types of audiences



EMC² PROVEN PROFESSIONAL

Writing a solid Executive Summary to portray the main findings of a project is crucial. In many cases, it may be the only portion of the presentation hurried managers will read. For this reason, it is imperative to make the language clear, concise and complete. Someone consuming the executive summary should be able to grasp the full story of the project and the key insights in a single slide. In addition, this is an opportunity to provide key talking points for the executive sponsor to use to evangelize the project work with others in the customer's organization.

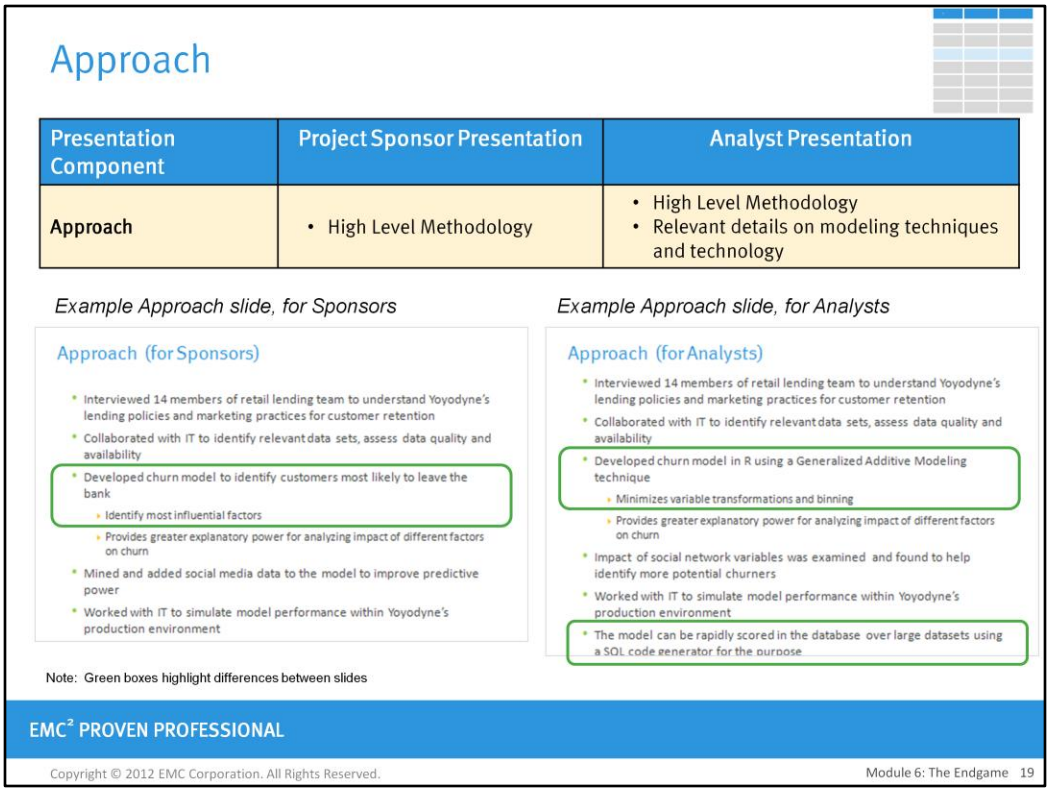
Be sure to frame the outcomes of the project in terms of business value, which is especially important if presentation is for the sponsor.



Shown above is an Executive Summary, with the main components called out in captions. Keep in mind this is not the only format for conveying the Executive Summary – this will vary based on the author's style, although many of the key components are common themes in Executive Summaries.

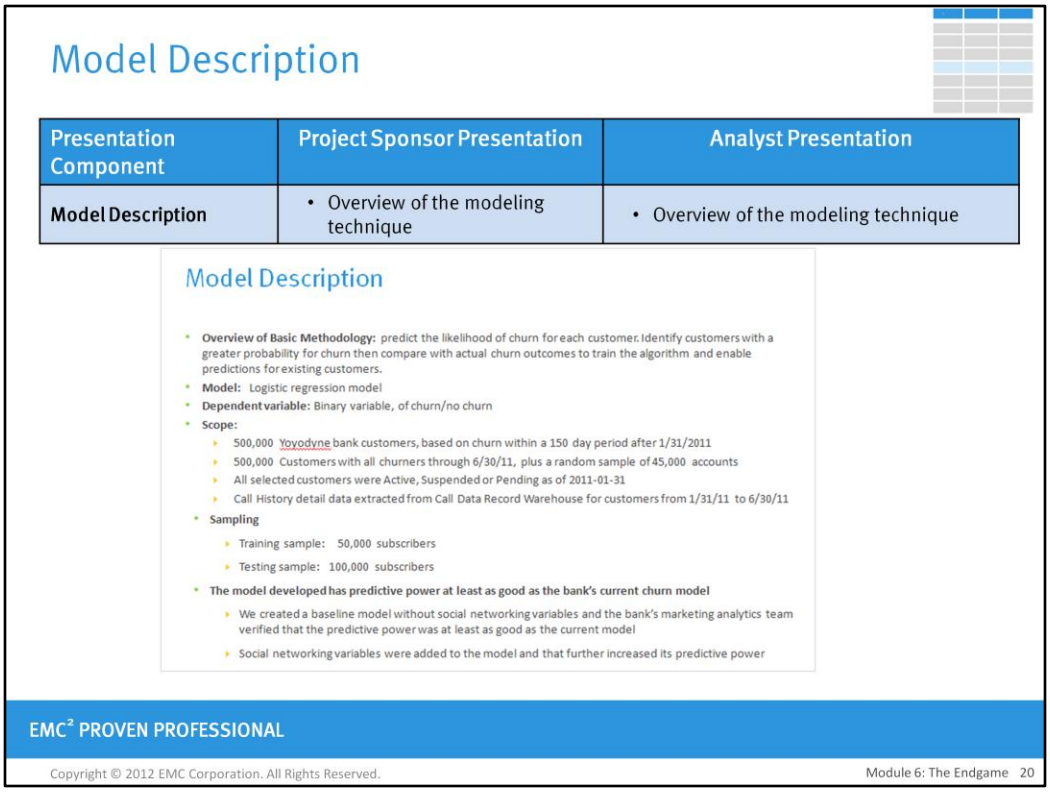
We recommend making your key message very clear and conspicuous at the forefront of the slide. This can be set apart with color as shown above, or you can use some other technique to draw attention to it. Your key message may become the talking point that executives or the project sponsor takes away from the project and uses to support your recommendation for a pilot project, so this needs to be succinct and compelling. **To make this message as strong as possible, measure the value of the work and quantify the cost savings, revenue, time savings, or other benefits to make the business impact concrete.**

Support the key message with 3 major points. You can have more than three major points, although going beyond three ideas makes it difficult for people to recall the main points, so you need to ensure that you keep the ideas clear and limit it to the few most impactful ideas you want the audience to take away from your work. If you list 10 key points, your message will become diluted and the audience may remember only 1 or 2. In addition, since is an analytical project, be sure to make one of your key points related to if, and how well, your work will meet the sponsor's SLA. Finally, although not required, it may be a good idea to support your main points with a visual, which will support the major points written at right. Visual imagery will serve to make a visceral connection with the reader, and help him or her retain the main message.



In the Approach portion of the presentation, you will need to **explain the methodology you pursued on the project**. This can include interviewing domain experts, the groups you collaborated with in the organization, and a few statements about the solution you developed. The objective of this slide is to ensure the audience is clear on the course of action you pursued, and understands it well enough to explain it to others in the organization. Also be sure to **include any additional comments related to your working assumptions as you did the work, this can be critical in defending why you pursued a specific course of action.**

When explaining the solution you developed, keep it at a high level for the project sponsors. If presenting to analysts or data scientists, add in additional detail about the type of model used, the technology and the actual performance of the model during your tests.
Finally, as part of the description on your approach, you may also want to mention constraints from systems, tools, or existing processes, and any implications for how these things may need to change due to this project.



Although the Model Description slide can be used for both audiences, the interests and objectives differ for each. For the sponsor, you will need to articulate the general methodology without getting into too much detail. You will need to convey the basic methodology followed in your work so the sponsor can communicate this to others within the organization. To do this, focus on explaining the general methodology you used in way that will enable your sponsor to convey it to others, and **provide talking points**. **Mention the scope of the data used to illustrate thoroughness and provide confidence that you used an approach that was an accurate portrayal of their problem and as free from bias as possible.** One of the key traits of a good Data Scientist is the ability to be skeptical of one's own work. This is an opportunity to view the work and the deliverable with a critical eye and consider how it will be received by the audience. Try to make sure it is an unbiased view of the project and the results.

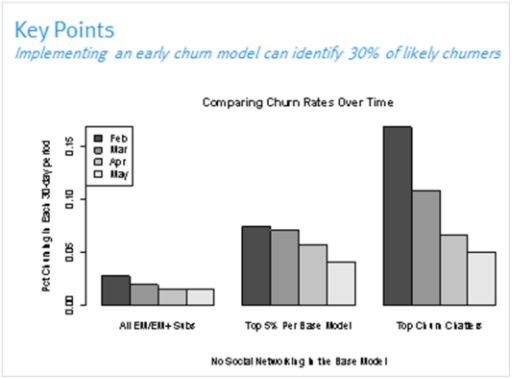
Assuming the model will meet the agreed upon SLAs, mention that the model will meet the SLAs based on performance of the model within the testing or staging environment.

Analysts will want to understand the details of the model, including the decisions you made in constructing the model, and the scope of the data extracts for testing and training. Be prepared to explain your thinking on this, as well as the speed of running the model within the test environment.

Key Points Supported With Data

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Key Points Supported With Data	• Support key points with simple charts and graphics (such as bar charts)	• Show details to support the key points • Analyst-oriented charts and graphs (ROC curves, histograms) • Visuals of key variables and significance of each

- Identify key points based on your insights and observations resulting from the data and model scoring results
- Illustrate your key points with charts and visualizations
- Use simpler charts for Sponsors



EMC² PROVEN PROFESSIONAL

Shown is an example slide showing some supporting detail regarding the rate of bank customers who would churn in various months. **When developing your key points, consider the insights that will drive the biggest business impact and are defensible with data**. For project sponsors, use simple charts such as bar charts, which illustrate data clearly and will allow the audience to understand the value of the insights. This is also where you will foreshadow some of your recommendations, and begin tying together your ideas to demonstrate what led to your recommendations and why. Creating clear, compelling slides to show your key points will make the recommendations more credible and more likely to be acted upon by the customer.

For analyst presentations, you will need more granular graphics. In this case, you may want to show a dot density chart or a histogram of a data distribution to support decisions you made in your modeling techniques. We will further discuss basic concepts of data visualizations in the next lesson.

Model Details

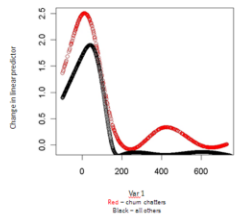
Presentation Component	Project Sponsor Presentation	Analyst Presentation
Model Details	• Omit this section, or discuss only at a very high level	• Show the code or main logic of the model, Include the model type, variables, technology used to execute it and score data. • Describe expected model performance and any caveats • Detailed description of the modeling technique, variables, scope, predictive power

Model Details

- Candidate variables: 22 from CRM, 154 from call history, and 12 social networking variables
- Through PCA and discussion with domain experts, we reduced ~190 variables to the 9 most predictive of customer churn
- General Additive Model (GAM) model built in R :

```
gam.wsn.by2 <- bam(volchurn.120.p ~
s(var1, bs="cs",by=c30,k=length(custom.knots))
+s(var2 , bs="cs",by=c30)
+s(var3, bs="cs",k=5)
+s(var4 ,bs="cs",k=5,by=c30)
+s(tvar5 ,bs="cs",k=5)
+var6
+var7
+ s(var8)
+ s(var9) ,
knots=list(var1=custom.knots),
data=train.df, family=binomial, weight=weight, gamma=1.4)
```

Var 1 has a larger and earlier impact on churn-chatters



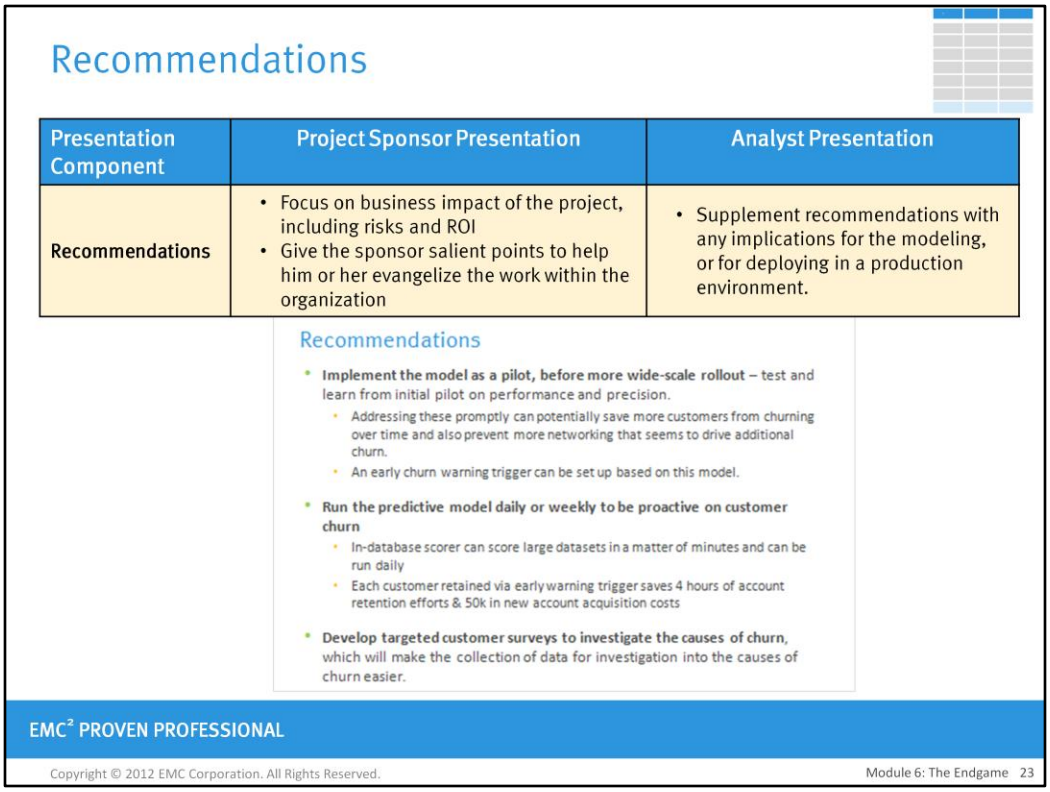
EMC² PROVEN PROFESSIONAL

Model details are needed to share additional information with people who have a more technical understanding than the sponsors, those who will need to implement the code, or colleagues on the analytical team.

In this section, **discuss the variables you used in the model and explain how or why you selected the ones you did.** In addition, you should share the actual code (or at least an excerpt) you developed to explain what was created and the basic mechanics of how it will operate. This will also serve to foster discussion related to any additional constraints or implications related to the main logic of the code. In addition, you can use this section to illustrate details of the key variables and the predictive power of the model, using analyst-oriented charts and graphs, such as histograms, dot density charts & ROC curves.

Discuss the speed with which the model can run in the test environment, the expected performance in a live, production environment, and the technology needed. This kind of discussion will address how well the model can meet the organization's SLA.

Finally, **include any additional caveats of the model and model performance**, such as systems or data the model will need to interact with, performance issues, or how to feed the outputs of the model into existing business processes. Describe the relationships of the main variables on your objectives (such as the effects of key variables on predicting churn, or the relationship of key variables to other variables). You may even consider making suggestions to improve the model, highlight any risks to introducing bias into the modeling technique, or describe certain segments of customers who may skew the overall predictive power of the methodology.

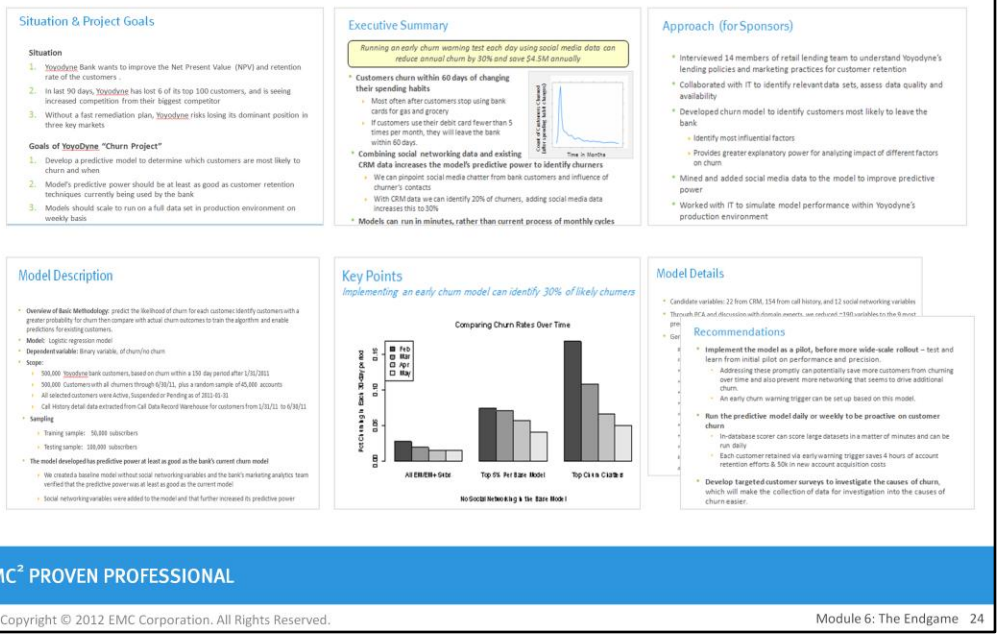


Create a set of recommendations that will include how to deploy the model from a business perspective within the organization and any other suggestions on the rollout of this logic, depending on your knowledge of the domain area from the discovery phase.

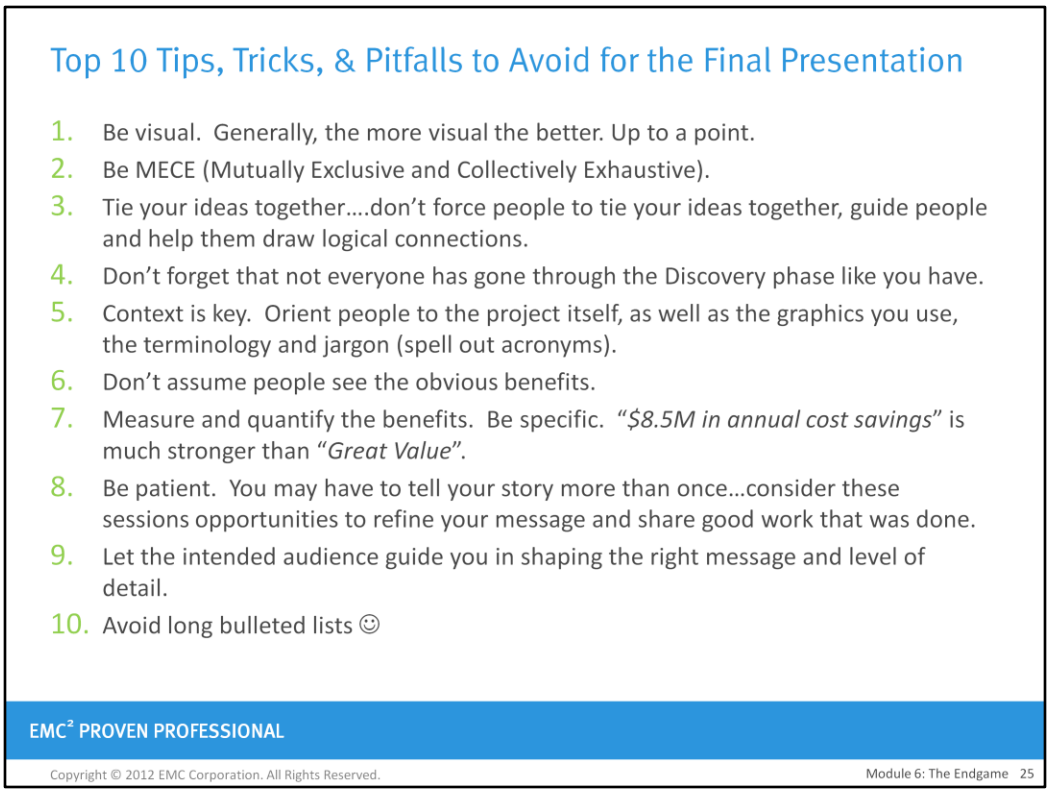
Attempt to measure the impact of the improvements and state how to leverage this within the recommendations. For instance, you might mention that every customer retained represents a time savings of 6 hrs for one of the bank's account managers or $50k in savings of new account acquisitions (due to marketing costs, sales, and system-related costs).

Focus on the actions you recommend to operationalize the work and the benefits the customer will receive as a result of implementing these recommendations.

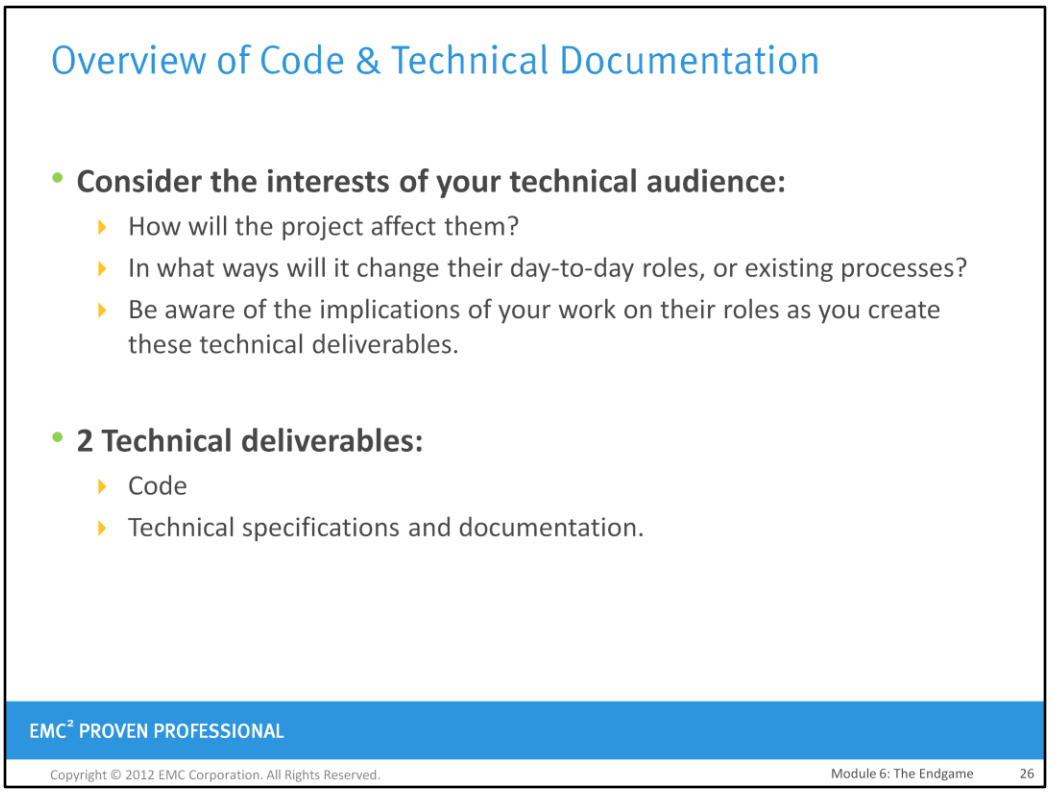
Quick Summary of Final Presentation Components



Shown is a summary of the main components of the final presentation.



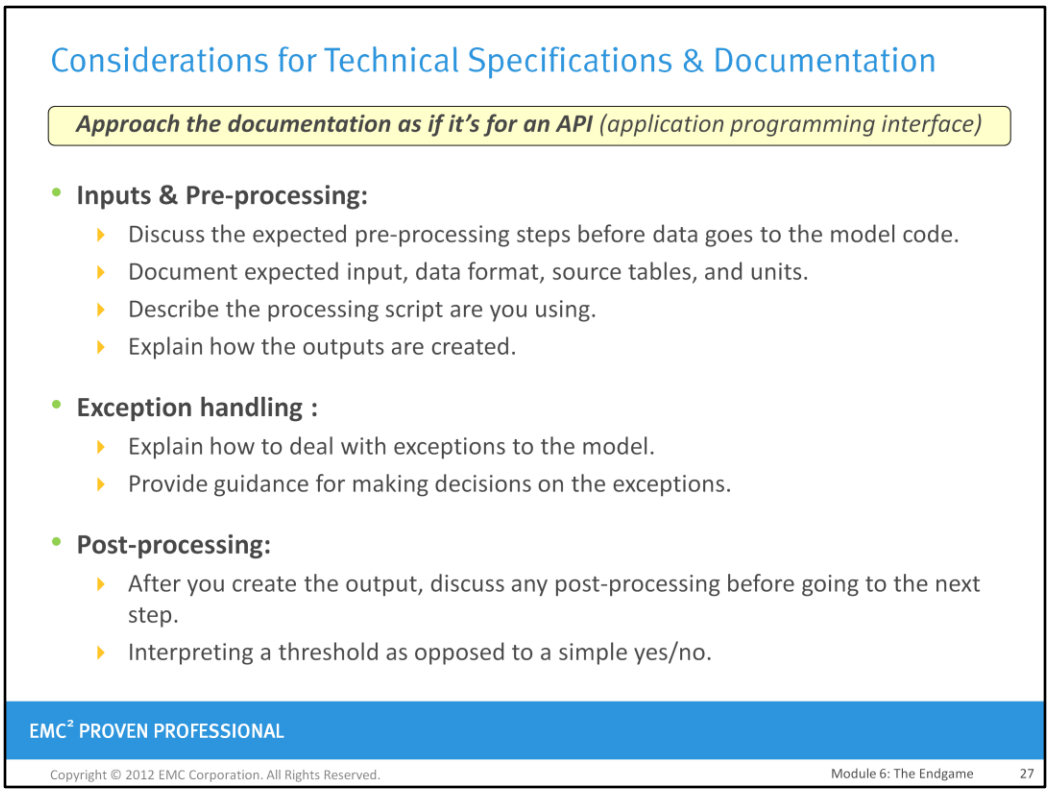
Here is a list of tips, tricks, and common pitfalls to avoid when creating presentations. One of the biggest mistakes authors make when creating presentations is forgetting to take time to set the context in presentations and in presenting the findings. When doing this, keep in mind that you always need to orient the viewer to the work you have created. This is true when writing your goals and situational overview for the project as a whole, and also for creating charts to give the right amount of context to orient the viewer to your main message. Context is key in conveying information in a way the audience will easily understand.



Now that you have learned a framework for authoring the final presentations, you also need to consider how to deliver the actual code you developed and the technical documentation that you need to support it. In doing this, consider how the project will affect the end users, and the technical people who will need to implement the code you developed.

Think through the implications of how your work will affect the recipients of the code, the kinds of questions they will have, and their interests. For instance, indicating that your model will need to perform real-time monitoring may require extensive changes to IT runtime environment, so you may need to consider a compromise of batch processes or nightly processes. In addition, you may need to get the technical team talking with the project sponsor to ensure the implementation and SLA will meet the business needs during the technical deployment.

Plan to address questions from IT related to how computationally expensive is will be to run the model in the production environment. You can provide a sense of this based on how well the model ran in the test scenarios and if there is additional fine-tuning that can be performed to optimize performance in the production environment.



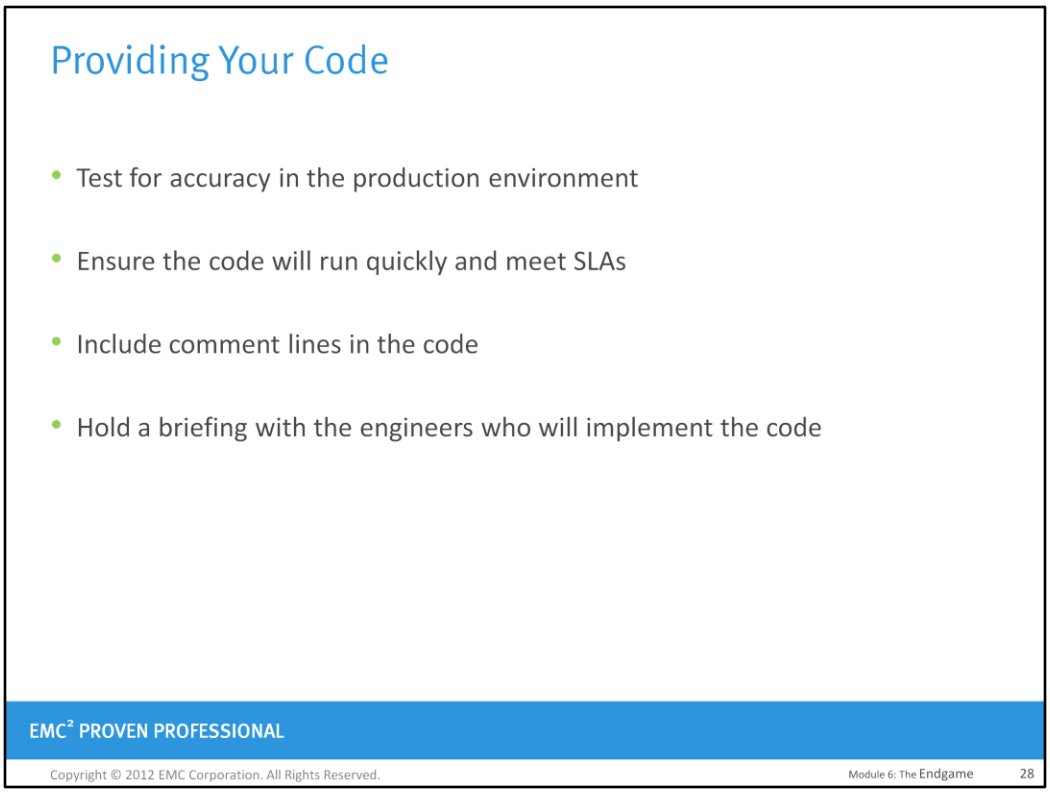
When considering the final deliverables, think as if you are delivering an API. To do this, you will **need to consider the inputs, outputs, and other system constraints to enable a technical person to implement it, even if they have not had a connection to the analytics project up to this point.**

Think about the documentation you create as a way to introduce the data your model needs, the logic it is using, and how other related systems will need to interact with it in a production environment for it to operate well.

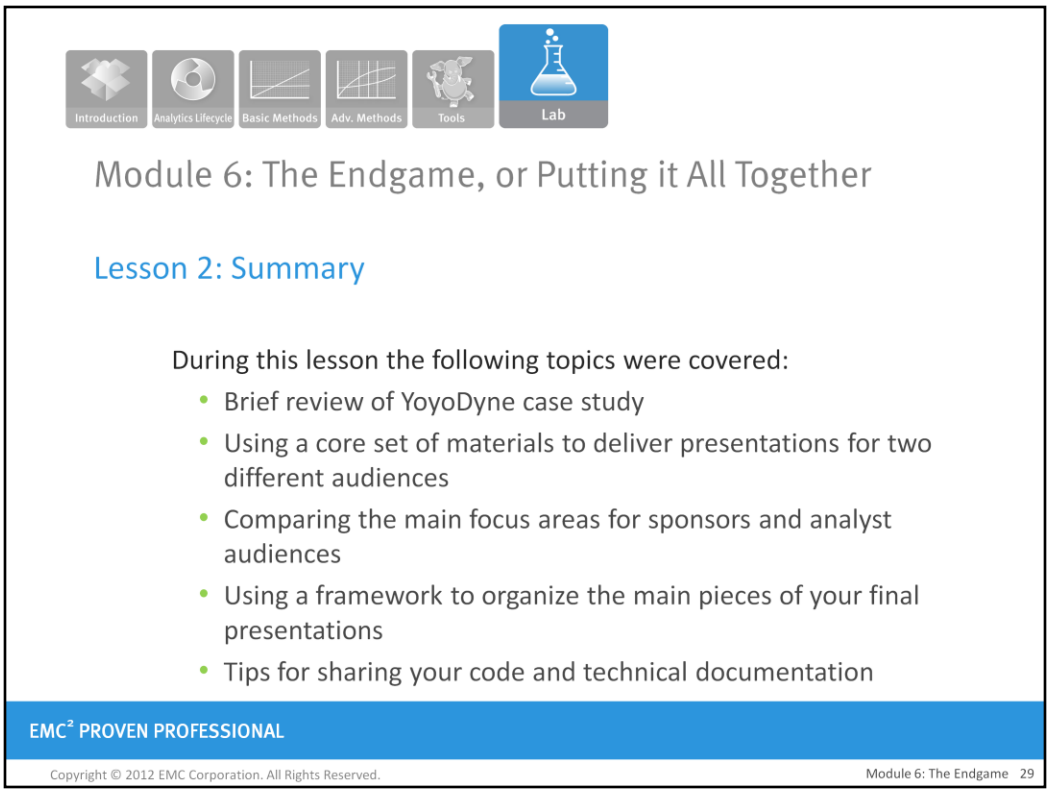
Consider writing specifications including specific details about the inputs the code will need, the data format and structures (such as do you need structured data, and does the expected data need to be numeric or string formats). Describe any transformations that need to be done on the input data before the code can use it, and if you created any scripting to perform these tasks. These kinds of details are important when other engineers need to modify your code, or point to a different dataset or table if and when their environment changes.

Regarding exception handling, consider how you handle data that is outside the expected data ranges of the model parameters, and how you will handle missing values, null values, zeros, NAs, or data that is in an unexpected format or type. Consider how you will treat these exceptions, or if there are implications on downstream processes that will need to be accounted for.

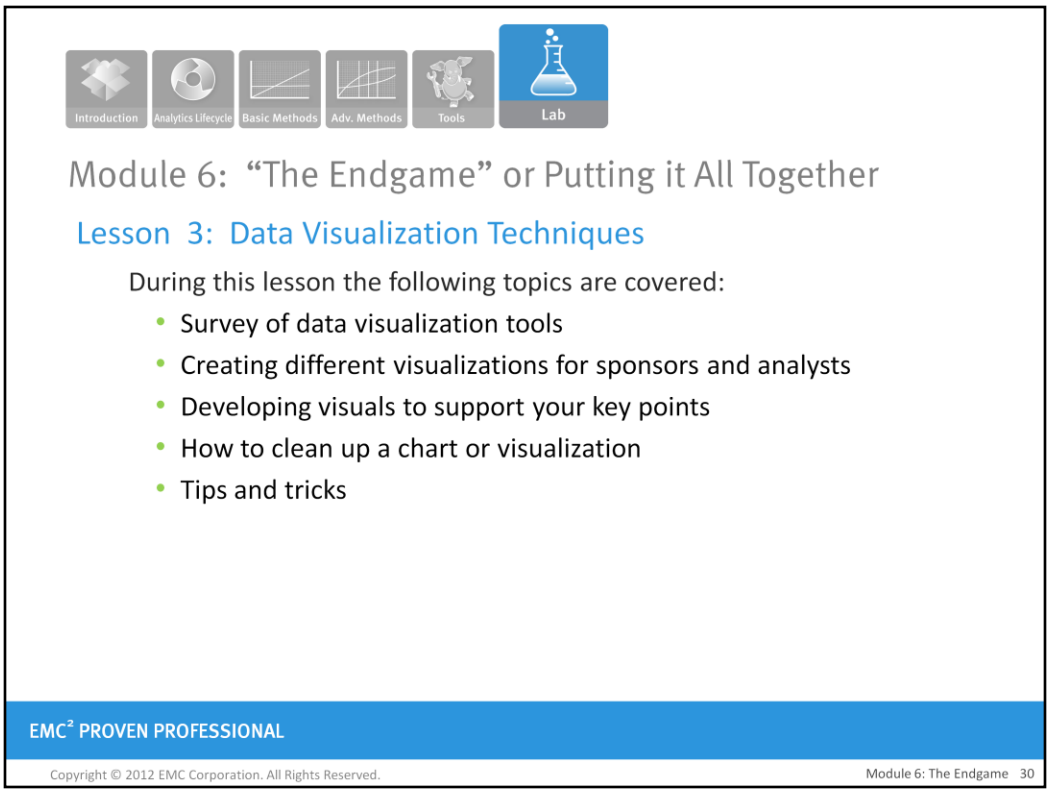
Regarding the model outputs, you will need to explain to what extent you post-process the output. For example, if you return a probability of churn, identify the scoring threshold to decide which customer accounts to flag as "in danger" or at risk of churn. In addition, you will need to make provisions for adjusting this threshold and training the algorithm, either in an automated learning fashion or with human intervention.



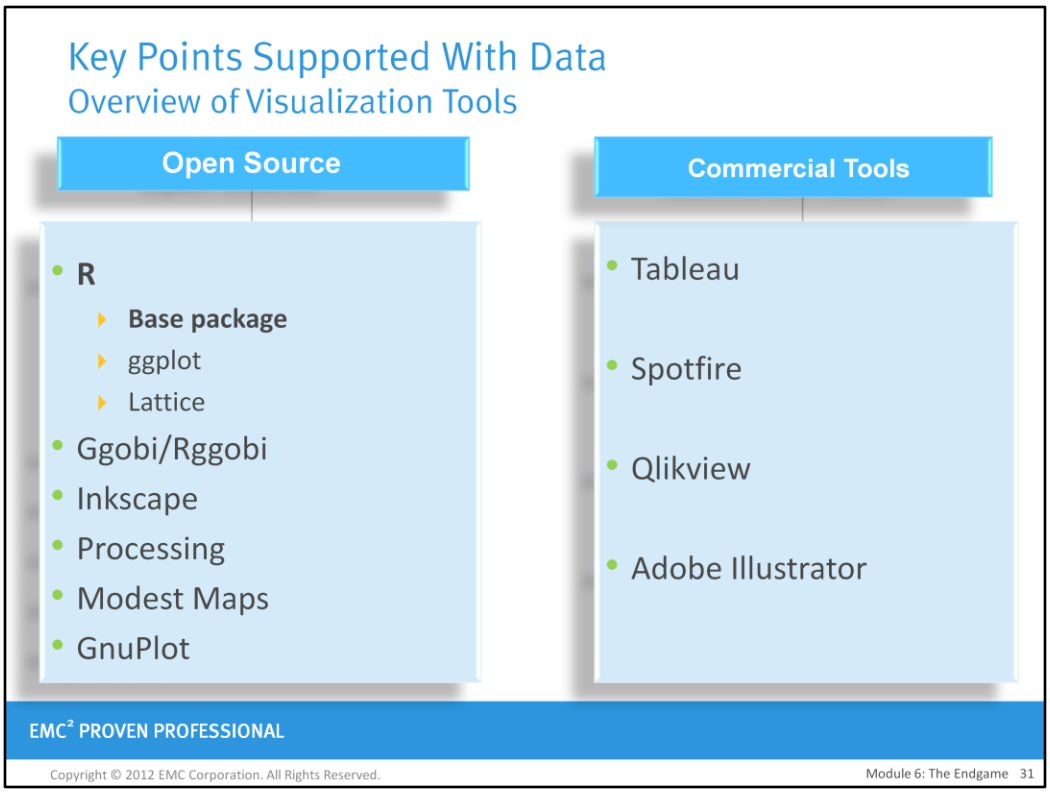
Although you will need to create technical documentation, many times engineers receive the code and try to use it without reading through all of the documentation. For this reason, it is important to **add extensive comments in the code itself**. This will serve to guide the people implementing the code on how to use it, what pieces of the logic are supposed to do, and guide them in stepping through the code and becoming familiar with it. If you can do a thorough job adding comments in the code, you will make it easier for someone else to maintain the code and help tune it in the runtime environment. In addition, this will help the engineers make edits to the code when their environment changes or they need to modify processes that may be providing inputs to the code or receiving its outputs.



This lesson covered a fictional case study.



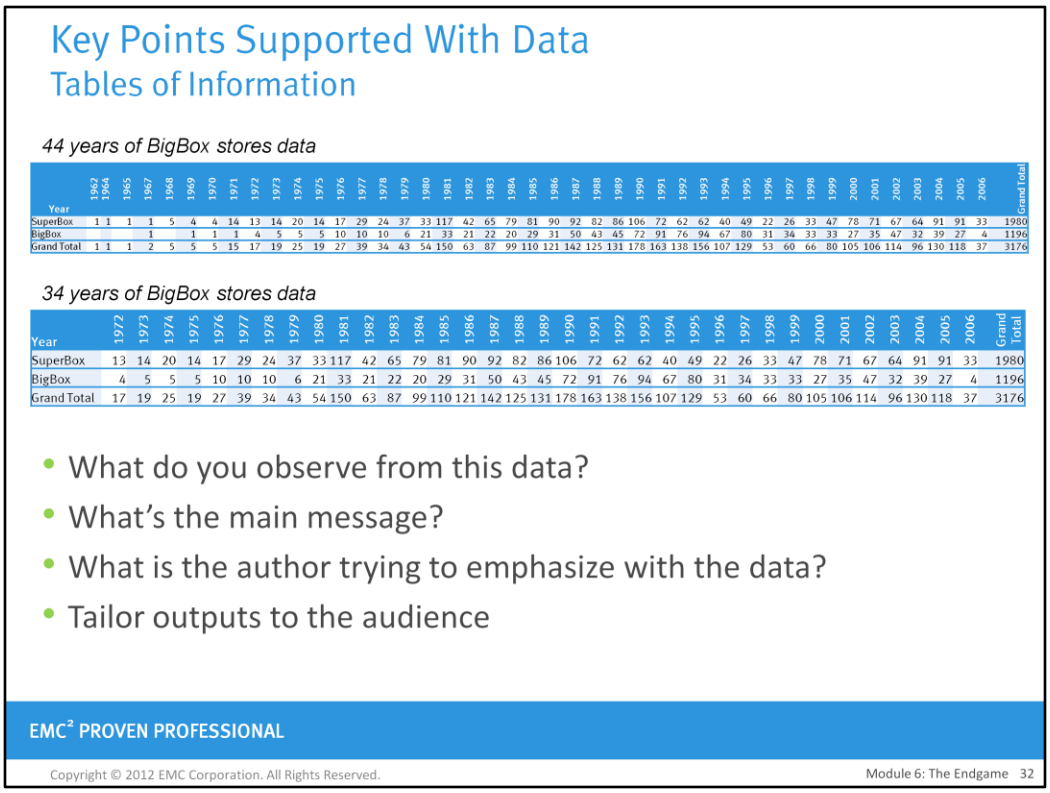
This lesson covers tools and recommendations for creating visuals for sponsors and analysts.



Many great visualization tools are on the market to help you in creating clear graphics for presentations and applications. Here is a listing of some of the more popular tools. As the volume and complexity of data has grown, users are becoming more reliant on using crisp visuals to illustrate key ideas and also to portray rich data in a digestible way.

In this course, we are using R as our main tool for data analysis and visualization. Over time, the open source community has developed many additional libraries to give you more options for portraying data visually. We are focusing on the base package of R in this course, although ggplot provide many more options for creating professional looking graphics, as does the Lattice library. For good examples of using open source visualization tools, you may refer to Nathan Yau's website, flowingdata.com or his book Visualize This, which provides additional methods for developing visualizations with many more open source tools.

Regarding the commercial tools listed above, **Tableau, Spotfire (by Tibco), and Qlikview function as data visualization tools, and also as business intelligence tools.** Due to the growth of data in the last few years, organizations for the first time are beginning to favor ease of use and visualization in business intelligence over more traditional BI tools and databases. These tools make visualization easy with good user interfaces. Adobe Illustrator is listed as some professionals will use this to enhance visualizations made in other tools. Inkspace is an open source tool used for similar use cases, with much of Illustrator's functionality.



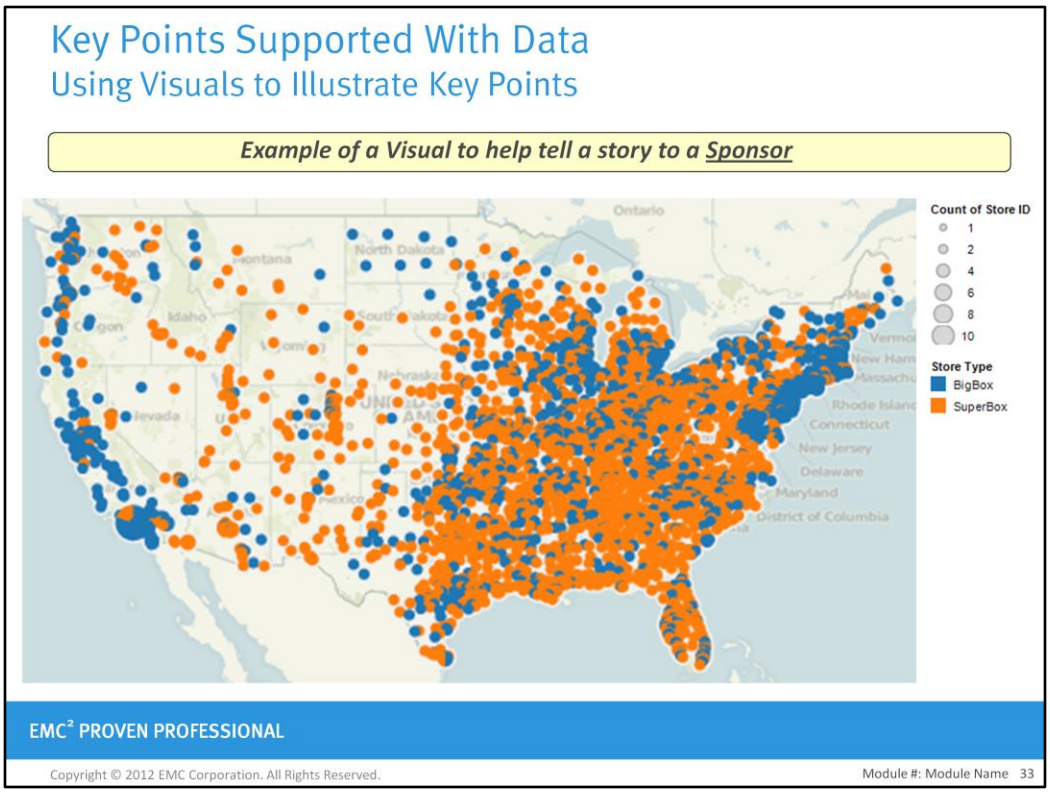
It is more difficult for people to observe the key insights when data is in tables than in charts. To underscore this point, in "Say it with Charts", Gene Zelazny mentions that **to highlight data create a visual out of it, such as a chart, graph or other data visualization.** The converse is also true. If for some reason you choose to downplay the data, leaving it in a table will draw less attention to it and make it more difficult for people to digest.

The way you choose to organize the visual in terms of the color scheme, labels and sequence of information will also influence how the viewer processes the information and what they believe is your key message from the chart.

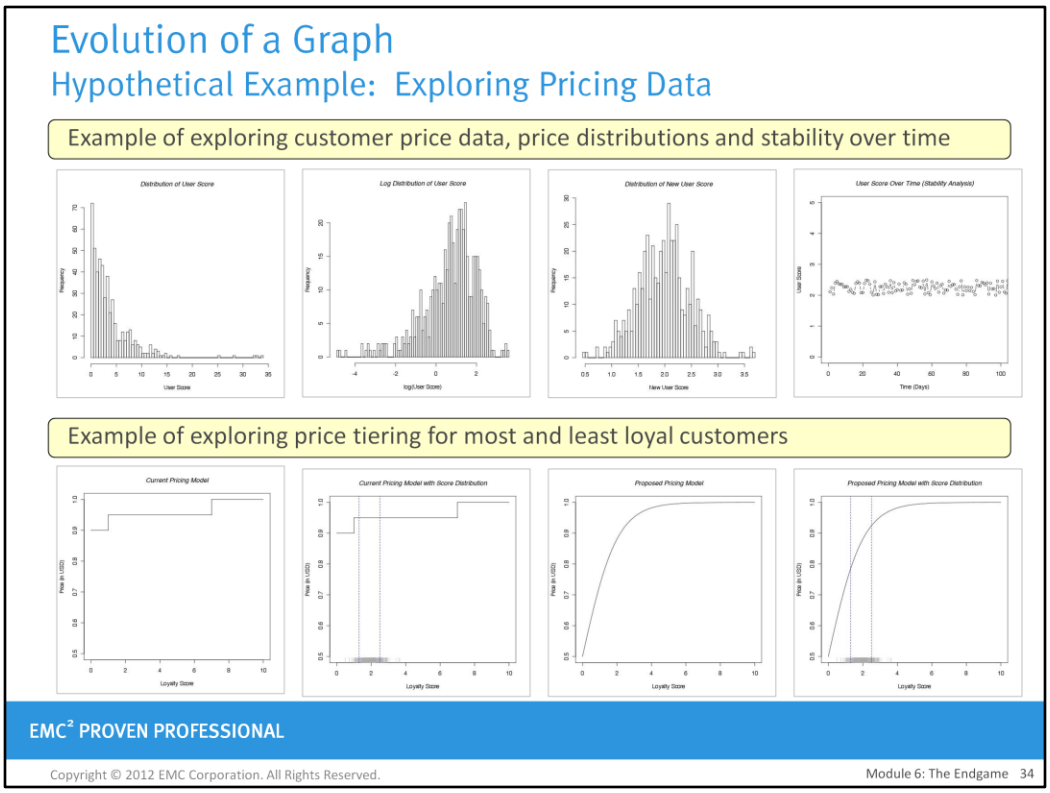
The table shows many data points, and given the layout of the information it is difficult to take away the key points at a glance. There are several observations in the data (if you look closely), such as …

1) BigBox experienced strong growth in the 1980s and 1990s
2) By the 1980s, BigBox began adding more SuperBox stores to its mix of chain stores
3) SuperBox outnumber regular stores nearly 2 to 1

Depending on the point you wish to make, take care to organize the information in a way that will intuitively enable the viewer to take away the same main point you want them to. Otherwise, they will guess at your main point and may take away something different than what you intended.

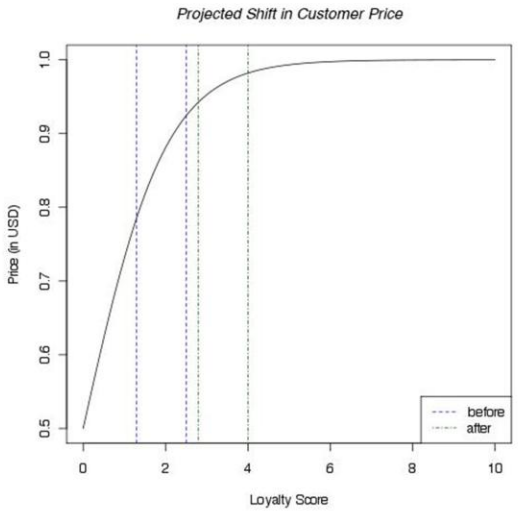


Above is a map of the U.S., showing the geographic location of BigBox stores. This is an example of a much more powerful way to depict data than in small tables, and this would be well suited to a sponsor audience. For a sponsor audience, you could also use simpler techniques, such as bar charts or line charts.



The graphics above portray a hypothetical example of some of the steps a Data Scientist may go through in analyzing customer pricing data. Data Scientists typically iterate and view the data many different ways, framing hypotheses, testing them and exploring the implications of a given model. In this case, we are looking at visual examples of pricing distributions, fluctuations in pricing, and exploring the differences in price tiers before and after implementing a new model to optimize price. The first row of visualizations depict distributions of the data, in raw and log form, as well as scatterplot of the prices over time to gauge the variability and consistency of the data. The second row shows price tiering and how this can change depending on the optimization methods. These visualizations illustrate how the data may look as the result of the model, and will help a data scientist understand the relationships within the data at a glance.

Evolution of a Graph, Analyst Example

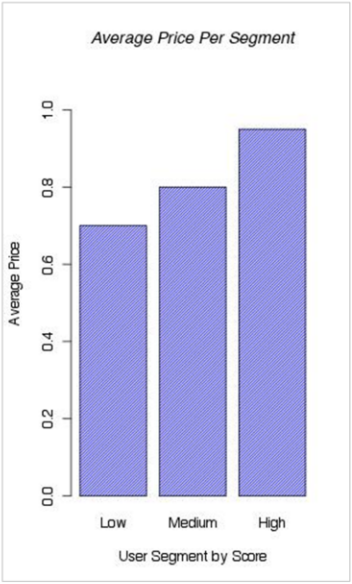


- *Implementing new price tiering approach increases the precision of price promotions by 23%*
- *Price optimization model explains 92% of customer behavior*
- *Model can be run in production environment on daily basis, if needed, to tailor changes to direct mail campaigns and web promotional offers*

EMC² PROVEN PROFESSIONAL

Above is an example of the output from the price optimization project scenario, showing how one may present this to an audience of other Data Scientists or data analysts. This shows a curvilinear relationship between price tiers and customer loyalty, when expressed as an index. Note that the comments at the right of the graph relate to the precision of the price targeting, the amount of variability in robustness of the model, and the expectations of model speed when run in a production environment.

Evolution of a Graph, Sponsor Example



- *Before the project, pricing promotions were offered to all customers equally*
- *With the new approach:*
  - *Highly loyal customers do not receive as many price promotions, since their loyalty is not strongly influenced by price*
  - *Customers with low loyalty are influenced by price, and we can now target them for this purpose better*
- *We project multiple cost savings with this approach*
  - *$2M in lost customers*
  - *$1.5M in new customer acquisition costs*
  - *$1M in reductions for pricing promotions*

EMC² PROVEN PROFESSIONAL

Above is an example of the output from the price optimization project scenario, showing how one may present this to an audience of project sponsors. This shows a simple bar chart to depict the average price per customer or user segment. This is a much simpler looking visual than the prior slide, and this one clearly shows that customers with lower loyalty scores tend to get lower prices, due to targeting from price promotions.

Note that the comments at the right of the graphic relate to explaining the impact of the model at a high level and the cost savings of implementing this approach to price optimization.

Key Points Supported With Data
Common Representation Methods

If you want to compare this kind of information….	…consider this kind of chart
Components	Pie chart
Item	Bar chart
Time Series	Line chart
Frequency	Line charts, histograms
Correlation	Scatterplot, side-by-side bar charts

EMC² PROVEN PROFESSIONAL

Shown are some basic chart types to guide you in considering that different types of charts are more suited to the situation depending on the data you have and the message you are attempting to portray.

The table is by no means exhaustive, it is illustrative to convey the most basic data representations, which can be combined, embellished and made more sophisticated depending on the situation and the audience. Consider the message you are trying to communicate, then choose an appropriate visual to support the point. Misusing charts tends to confuse the audience, so be sure to take into account the data type and message when choosing a chart.

Pie charts are designed to show the components, or parts relative to the whole set of things. It is also the most overused chart. If you are going to use a pie chart, use it when showing only 2-3 items in a chart and only for sponsor audiences. Bar charts and line charts are used much more often, and are very useful for showing comparisons and trends over time. For bar charts, horizontal bar charts allow you to fit the text labels better and provide more horizontal space to fit them next to a chart, even though many people tend to use vertical bar charts. Vertical bar charts tend to work well when the labels are small, such as when showing comparisons over time using years.

For frequency, histograms will show the distribution of data, and are useful for showing information to an analyst audience or to data scientists. The data distributions are typically one of the first steps in visualization data to prepare for the model planning. When doing correlation, scatterplots are useful to compare relationships among variables.

As with any presentation, consider the audience and their level of sophistication when selecting the chart to convey your message. These charts are simple examples, but can easily become more complex with additional data variables, combining charts together, or adding animation where appropriate.

For additional reference on data types and their related charts, you may want to look at the URL: http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

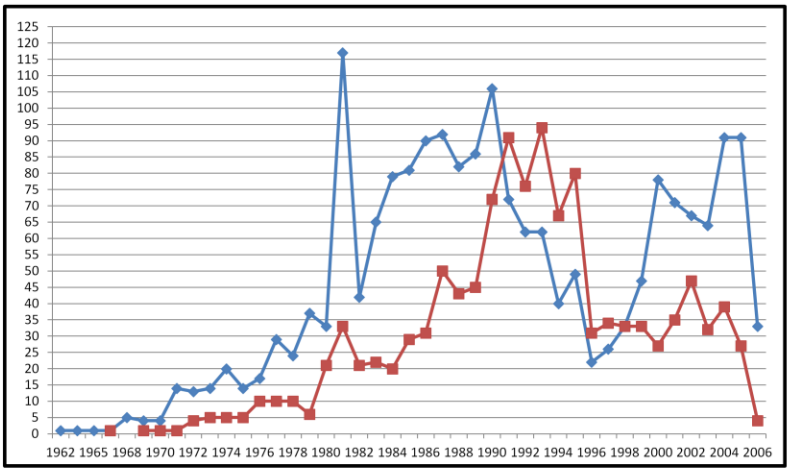
How to Clean Up a Graphic, Example 1

The Before Picture

- What are the main messages here? What is the author trying to emphasize?
- What's wrong with this picture?

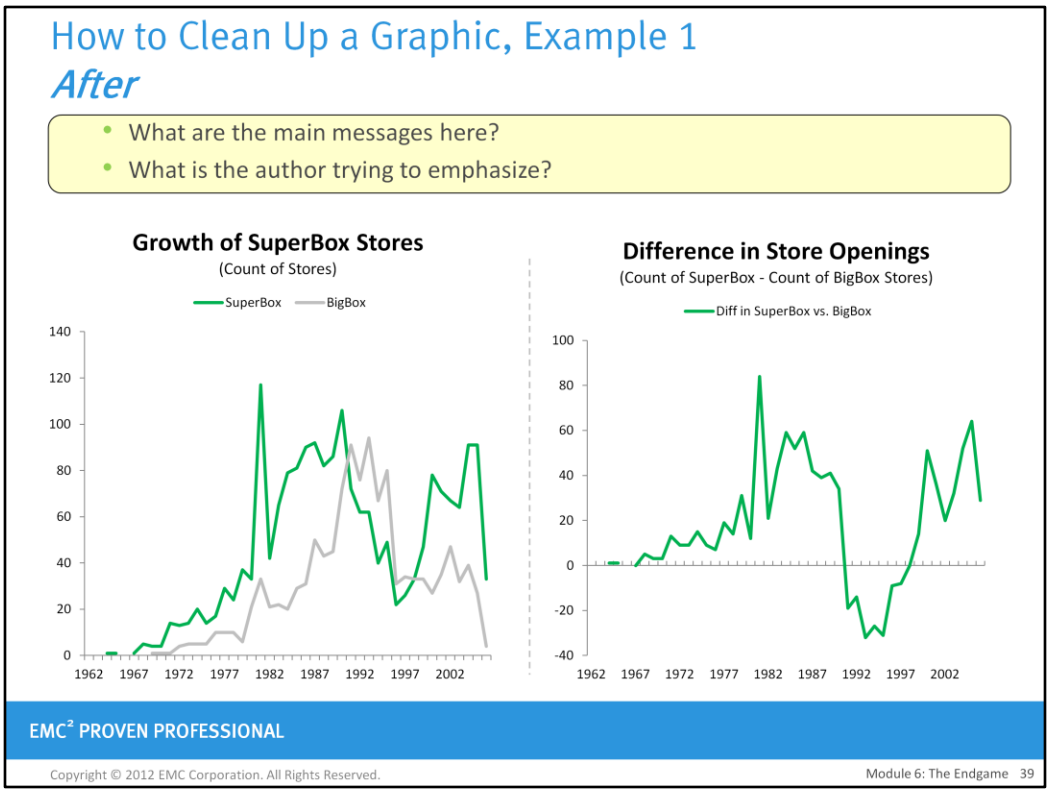
Chart junk

1. Horizontal Grid Lines
2. Chunky data points
3. Overuse of emphasis colors; lines & border
4. No context or labels
5. Crowded axis labels



EMC² PROVEN PROFESSIONAL

Shown is an example of a line chart comparing two trends over time. It's a busy looking chart and contains a lot of "chart junk", which distracts the viewer from the main message. Shown at the left of the chart are some of the chart junk this visual suffers from, which are easily addressed as shown in the next slide. Note that there is no clear message associated with the chart and no legend to provide context for what is shown.



These are two examples of cleaned up versions of the chart on the previous page. Note that the problems with chart junk have been addressed, there is a **clear label and title for each chart to reinforce the message, and color has been used in ways to highlight the point the author is trying to make**.

Note the amount of white space being used in each of the two charts shown. Removing grid lines, excessive axes, and the visual noise within the chart allows you to create very clear contrast between emphasis colors (the green line charts) and the standard colors (the light gray of the BigBox stores). When creating charts, it is best to do most of your main visuals in standard colors, light tones or color shades so that you can choose to add stronger emphasis colors to emphasize the main points and draw attention to the parts of the graphic that demonstrate your main points. In this case, we have made the trend of BigBox stores in light gray to fade into the background, but not disappear, while making the SuperBox stores trend in a bright green and make it prominent to support the message the author is making about the growth of the SuperBox stores.

An alternative is shown at right. If the main message is to show the difference in the growth of new stores, you could simplify the chart further and choose to graph only the difference between SuperBox stores compared to regular BigBox stores. **Two examples are shown to illustrate different ways to convey your message, depending on what it is you would like to show.**

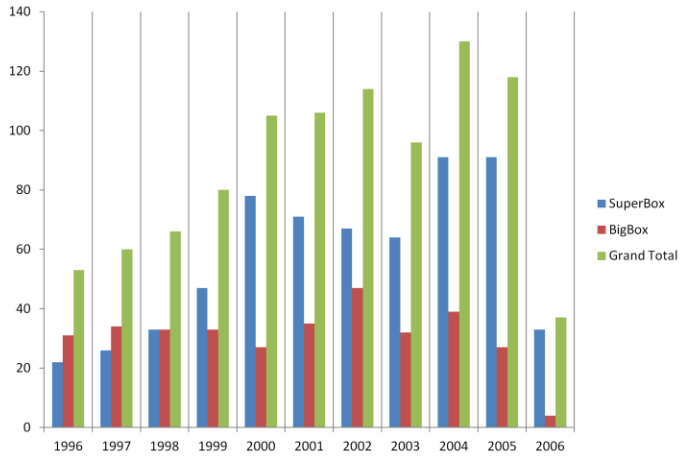
How to Clean Up a Graphic, Example 2

The Before Picture

- What are the main messages here? What is the author trying to emphasize?
- What's wrong with this picture?

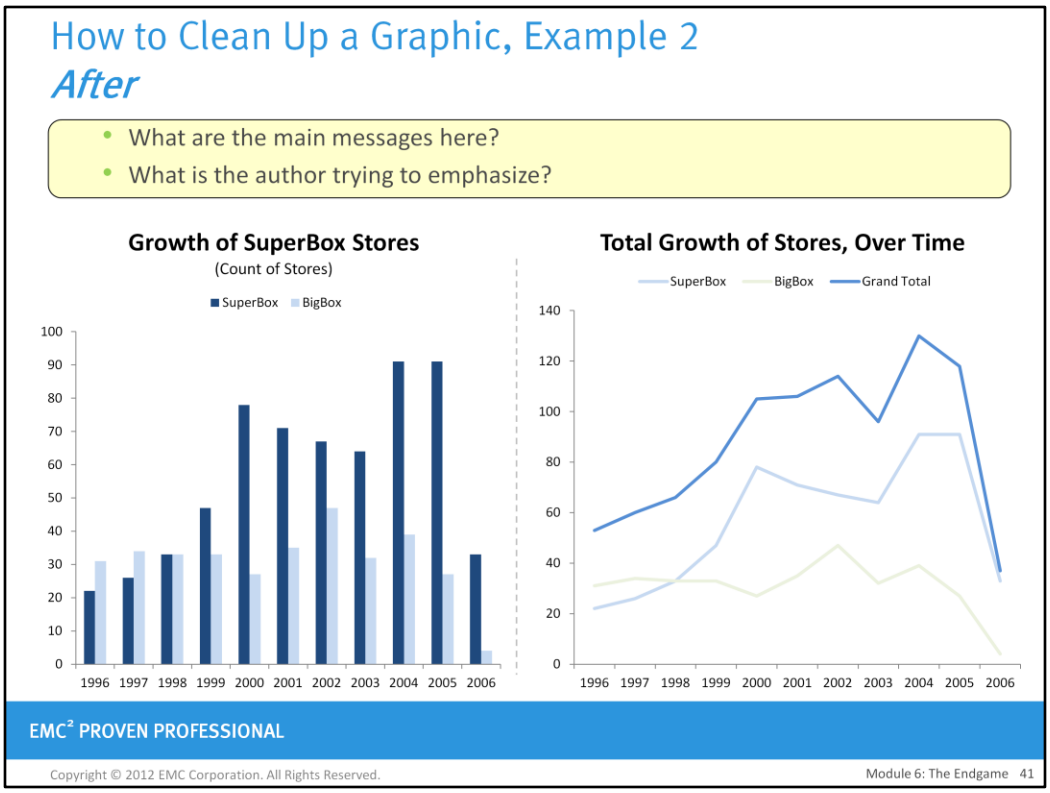
Chart junk

1. Vertical Grid Lines
2. Too much emphasis colors
3. No chart title
4. Legend at right restricts chart space
5. Labels are too small



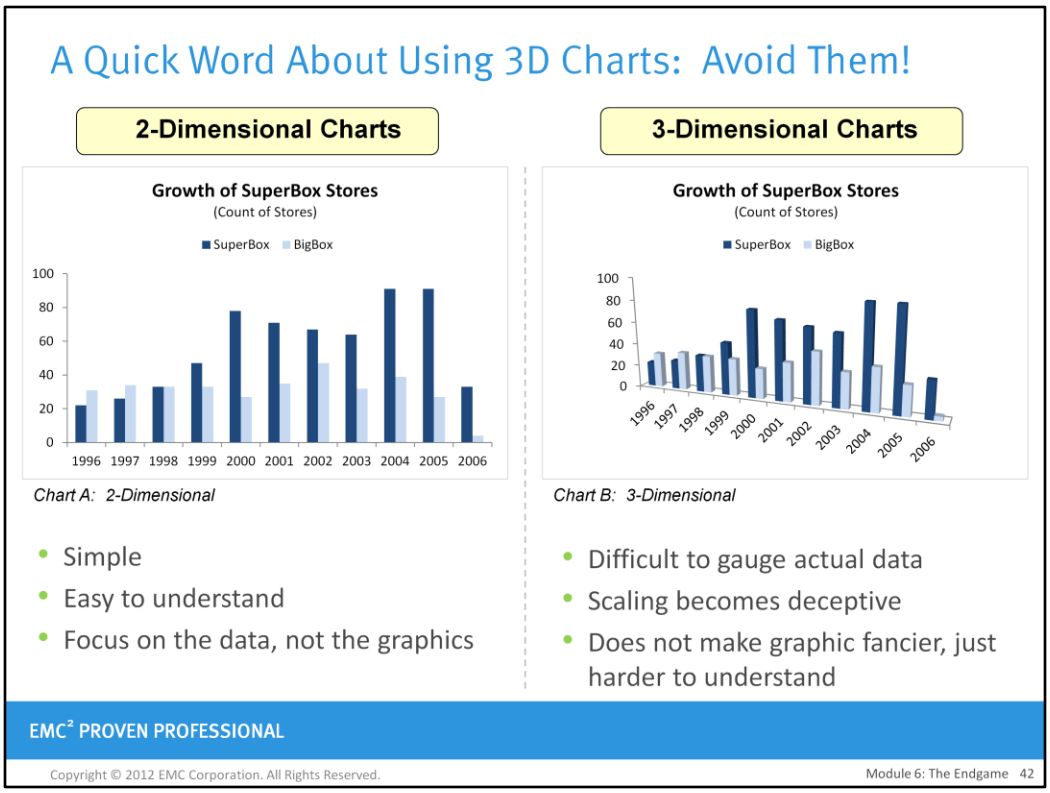
EMC² PROVEN PROFESSIONAL

Here is a sample graphic with typical problems related to chart junk, including misuse of color schemes and lack of context. Shown at left are the main problems with the graphic, with cleaned-up alternatives to this visual on the subsequent pages.

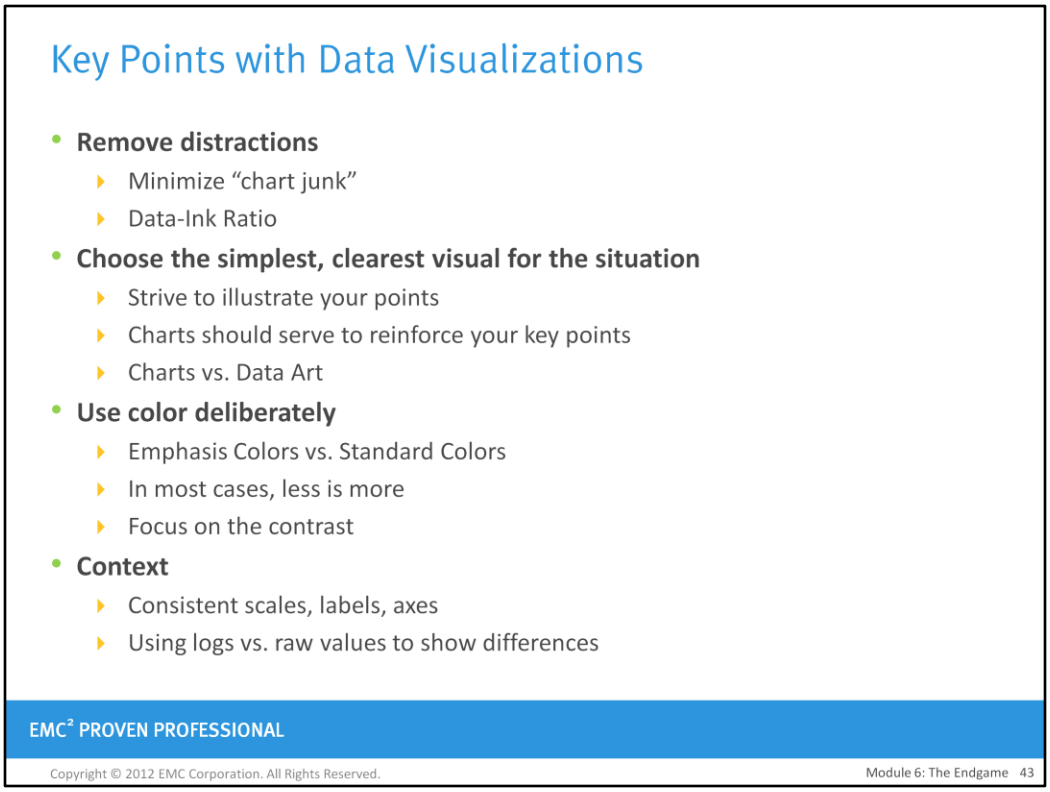


Shown are some simplified and cleaned up versions of the previous slide's graphic. These show two options for modifying the graphic, depending on the main point the presenter is trying to make.

The chart on the left of the slide shows strong, emphasis color (dark blue) representing the SuperBox stores, to support the chart title about the Growth of SuperBox Stores. If the presenter wanted instead to talk about the total growth of BigBox stores, a line chart (shown on the right) showing the trends over time would be a better choice. In both cases, we have removed the noise and distractions within the chart, have de-emphasized data we wish not to speak to, and made prominent data that will reinforce our key point as stated in the chart's title.



Shown is a side-by-side comparison of two charts. As mentioned, 3-dimensional charts often distort scales and axes, and impede viewer cognition. The charts on the left and right portray the same data, although when looking at Chart B it is more difficult to judge the actual height of the bars. In addition, the shadowing and shape of the chart cause most viewers to spend time looking at the perspective of the chart, rather than the height of the bars, which is the key message and purpose of this visual.

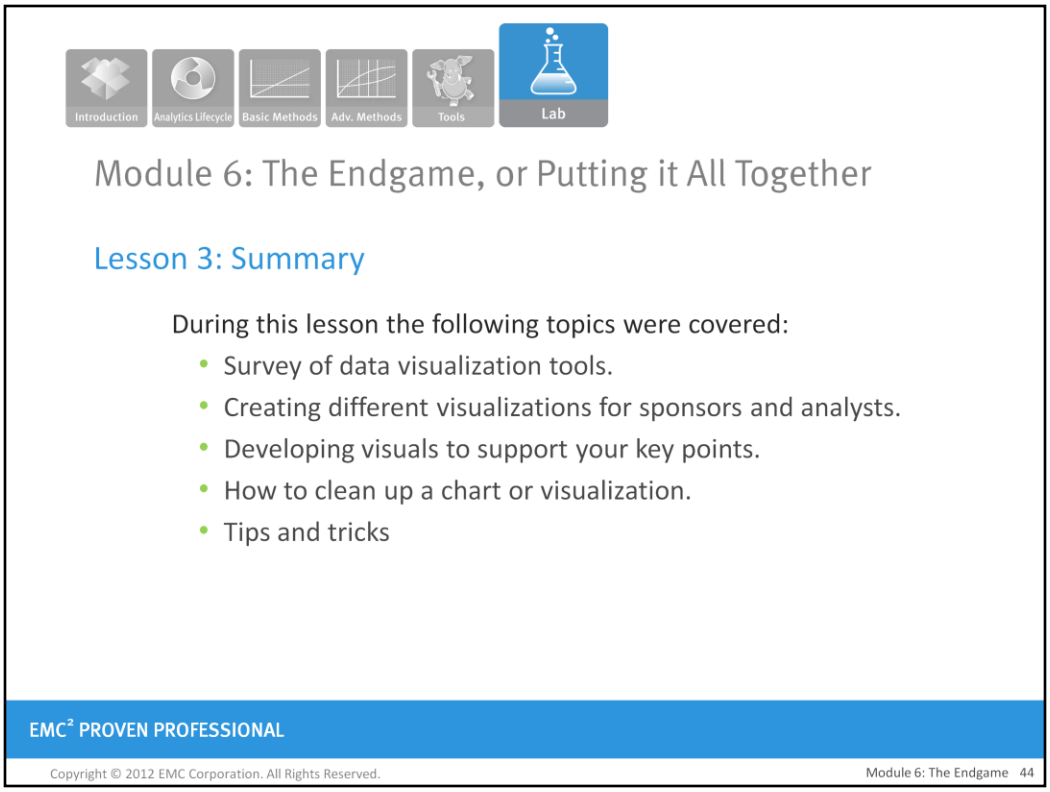


The points listed summarize many of the key ideas on the preceding examples. Following these few ideas about minimizing distracts in slides and visualizations, communicating clearly and simply, using color in a deliberate way and taking time to provide context will address most of the common problems in charts and slides. These few guidelines will enable you to have crisp, clear visuals that support your story, without needing to become a data artist.

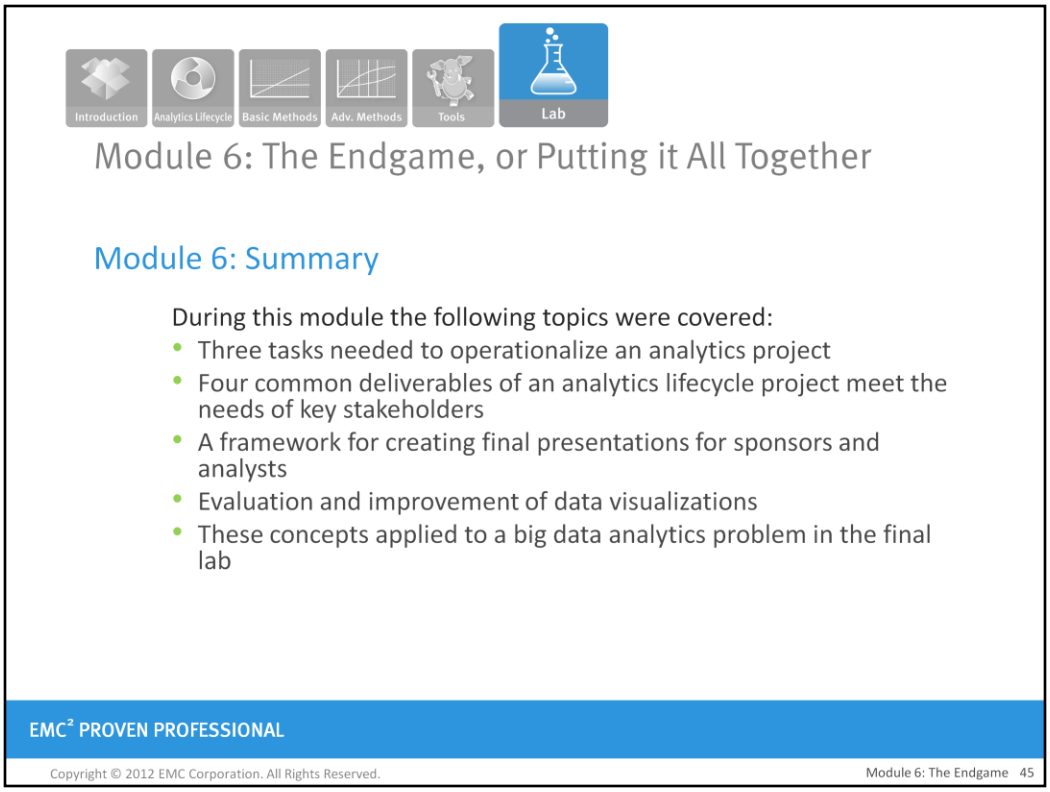
Similar to the idea of removing chart junk is being cognizant of the data-ink ratio. Data-ink refers to the actual portion of a graphic that is used to portray the data itself, while non-data ink represents for labels, edges, colors, and other decoration. The ratio = (data-ink)/(total ink used to print the graphic). In other words, the greater the ratio of data-ink in your visual, the more data rich it is and the fewer distractions it has. For more information and further examples, see http://www.infovis-wiki.net/index.php/Data-Ink_Ratio .

In most cases, **the best way to show your visual is using the simplest, clearest visual to illustrate your point. To do this, remove distractions and avoid unnecessary embellishment in the visual**. Keep in mind that you are trying to find the best, simplest method for transmitting your message, rather than data art which is about how data it can be represented in a creative way, and can be an end in and of itself.

Context is critical to orient the viewer of a visualization, as people have immediate reactions to imagery on a pre-cognitive level. To this end, make sure to use thoughtful usage of color, and orient the viewer with scales, legends, and axes. Using logarithms to normalize data is a useful way to fit data into a visualization, as we showed earlier in this lesson when showing a chart for analysts, showing account values distributed lognormally. This can be a very useful technique when you want to show a wide range of data, such as a broad range of income values or population sizes.



This lesson covered tools and recommendations for creating visuals for sponsors and analysts.



This module covered these topics.

Course Summary

Key points covered in this course:

- Participation as a data science team member on big data and other analytics projects by:
  - Deploying a structured lifecycle approach to data science and big data analytics projects
  - Reframing a business challenge as an analytics challenge
  - Applying analytic techniques and tools to analyze big data, create statistical models, and identify insights that can lead to actionable results
  - Selecting optimal visualization techniques to clearly communicate analytic insights to business sponsors and others
  - Using tools such as R and RStudio, MapReduce/Hadoop, in-database analytics, and window and MADlib functions
- How advanced analytics can be leveraged to create competitive advantage and how the Data Scientist role and skills differ from those of a traditional business intelligence analyst

EMC² PROVEN PROFESSIONAL

These are the key points covered in this course.

BIG DATA ANALYTICS & DATA SCIENCE ASSOCIATE COURSE

Introduction to Final Lab

EMC² PROVEN PROFESSIONAL

Final Lab Overview

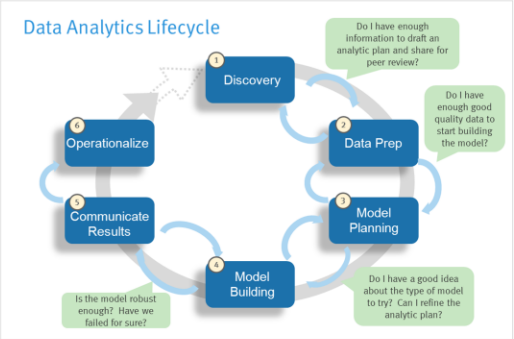
This lab allows students to apply what they have learned from the analytical methods and tools to a big data problem using the analytics lab environment.

- After completing the tasks in this lab you should able to:
  - Explore the big data set provided
  - Assess data quality, outliers and training sets
  - Use R and PSQL statements during your analysis of big data
  - Select the appropriate analytical methods for the problem and data set
  - Create a narrative summary of your findings, using the methods shared earlier in this module

EMC² PROVEN PROFESSIONAL

Introduction to Final Lab

- Each student will perform the end-to-end analytics lifecycle process in an accelerated way
  - From Discovery through creating final presentations
  - Students will have 6-8 hours to perform the lifecycle
  - Work on your own, or in teams



- Key Milestones
  - Frame the problem and develop an analytical plan
  - Checkpoint 1: Prepare data for analysis
  - Checkpoint 2: Select the variables, visualize the data
  - Checkpoint 3: Choose and execute the model, interpret results
  - Prepare and deliver presentation (either a Sponsor or an Analyst presentation)

EMC² PROVEN PROFESSIONAL

Introduction to Final Lab
Synopsis of the Problem

- **Overview**
  - A financial planning company, FPC, would like to expand the set of services they offer by creating an online site for loan advice.
  - Potential home loan borrowers can enter information about their personal finances and the kind of home loan they want, and the site will return the probability of getting such a loan, along with some general advice about how to increase their likelihood of success.
- **Goals for the Data Scientist on the Project**
  - Determine if it will more effective to develop different models or one model and why.
  - Determine if you need to ask borrowers for personal demographic information, or if you can build a robust enough model without this data
  - Determine how accurate the model is and provide suggestions for the kind of general advice FPC can put on their website

EMC² PROVEN PROFESSIONAL

Introduction to Final Lab
Data & Additional Considerations

- **Data**
  - Data for this lab is the housing loan database assembled by U.S. federal agencies to support the Home Mortgage Disclosure Act (HMDA)
  - This database of identifies the census tract location of almost every housing loan and housing loan application (tens of millions) made in the United States each year
  - The data provided for analysis in this lab is for the year 2010

- **Considerations**
  - Note your decision points to record why you chose do certain things with the data or select specific analytical methods for the problem set
  - Document your assumptions

EMC² PROVEN PROFESSIONAL

Final Presentation for FPC

Data Science & Big Data Analytics

Loan Underwriting Prediction Model

Project Sponsor Presentation

EMC² PROVEN PROFESSIONAL

Situation & Project Goals

Situation

1. A financial planning company, FPC, would like to expand the set of services they offer by creating an online site for loan advice.
2. FPC feels competitive pressure to expand their online services , as other companies are gaining more online revenue from prospective applicants
3. If FPC can create a successful pre-screening website, they can fend off competitive pressure and supplement their traditional lending business

Goals

1. Create a POC model to predict successful loan origination based on a limited set of loan and borrower characteristics, avoid using sensitive data if possible
2. Deploy model on FPC web site as a service for potential home loan borrowers, which can drive customers toward FPC for more focused, personal financial planning to achieve their life goals.
3. Develop an efficient model that can provide fast online responses and give an answer within a 45 second SLA

Project Goals

1. Create a POC model to predict successful loan origination based on a limited set of loan and borrower characteristics

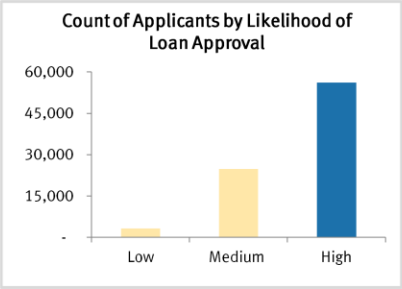
2. Deploy model on FPC web site as a service for potential home loan borrowers, which can drive customers toward FPC for more focused, personal financial planning to achieve their life goals.

3. Develop an efficient model that can provide fast online responses and give an answer within a 45 second SLA

Executive Summary

Loan advice website would drive new revenue for FPC and predict loan approvals with 68% accuracy

- **FPC domain experts identified 3 main growth areas for FPC resulting from a new prescreening website**
  - Close highly qualified applicants online, refer others for more personal loan advice
  - Partner with real estate brokers for referrals
  - Drive new revenue from advertising with partner real estate brokers



- **Reasonably accurate pre-screening predictions can be made with limited data**
  - Loan Amount, Applicant Income and Loan Type are most influential factors
  - Sensitive demographics can be omitted, these factors show negligible model improvement
- **Initial POC indicates model can run online within 45 second SLA**
  - Collaborated with IT to benchmark model performance in simulated production environment
  - Initial tests show acceptable speed, but further testing is suggested

EMC² PROVEN PROFESSIONAL

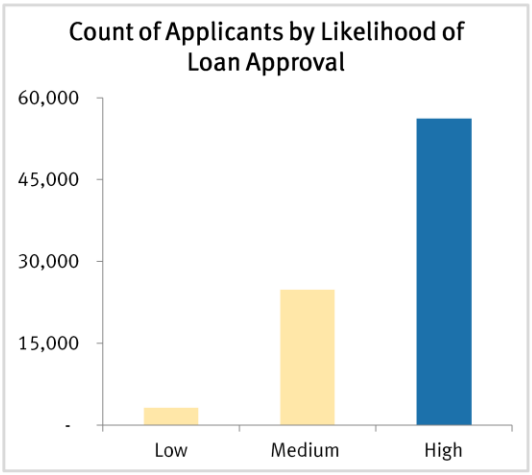
Approach

- Consulted with FPC's advisory staff to understand underwriting policy and identify factors that affect a home loan applicant's likelihood of approval
- Collaborated with IT to identify relevant data sets, assess data quality and availability
- Developed predictive model to identify loans most likely to be underwritten
  - Identify most influential factors
  - Provides greater explanatory power for analyzing influence of specific factors on loan origination
- Worked with IT to simulate model performance within FPC's production environment

Model Description

- **Overview of Basic Methodology**: predict the likelihood of loan approval for each applicant. Identify loans with a greater probability for successful underwriting then compare with actual loan originations to train the algorithm and enable predictions for future users.
- **Model:** Logistic regression model
- **Dependent variable:** Binary variable, of origination/no origination
- **Scope & Sampling**
  - 238,000 loan applications for Maryland from 2010
  - All selected loans were Approved or Denied
  - Training sample: 23,800 loans
  - Testing sample: 47,600 loans
- **The model developed has reasonable predictive power for the dataset provided**
  - Loan Type, Loan Amount and Applicant Income provide the most influence over the probability of approval
  - Model shows marginal improvement with sensitive demographic information, such as Applicant Age or Gender

Majority of FPC Have High Likelihood of Loan Approval



- Most of FPC's prospective customers fall in High likelihood of approval category
- Suggest tailoring offerings to transition high likelihood applicants to full application and close loans online
- Medium tier applicants are good candidates for loan counseling services

EMC² PROVEN PROFESSIONAL

Recommendations

- **Implement the model on FPC site as a pilot to limited set of site visitors –** test and learn from initial pilot on model outputs.
  - Tuning model predictions can encourage users to visit in-person advisors and minimize frustrations by unexpected negative predictions.
  - Outputs can be further enhanced to give users advice on how to increase likelihood of successful loan origination.
- **Embed links to filtered home listing based on user inputs**
  - *Example*: Show ~$300K homes to users who could afford loan based on income, regardless of desired loan amount
  - Charge partner real estate brokers for PPC (pay-per-click) traffic
- **Over time, refine model by analyzing whether more general or specific advice leads to user conversions**, encouraging users to visit with an in-person advisor for a personalized financial plan.

Final Presentation for FPC

Data Science & Big Data Analytics

Loan Underwriting Prediction Model

Analyst Presentation

EMC² PROVEN PROFESSIONAL

Situation & Project Goals

Situation

1. A financial planning company, FPC, would like to expand the set of services they offer by creating an online site for loan advice.
2. FPC feels competitive pressure to expand their online services , as other companies are gaining more online revenue from prospective applicants
3. If FPC can create a successful pre-screening website, they can fend off competitive pressure and supplement their traditional lending business

Goals

1. Create a POC model to predict successful loan origination based on a limited set of loan and borrower characteristics, avoid using sensitive data if possible
2. Deploy model on FPC web site as a service for potential home loan borrowers, which can drive customers toward FPC for more focused, personal financial planning to achieve their life goals.
3. Develop an efficient model that can provide fast online responses and give an answer within a 45 second SLA

Project Goals

1. Create a POC model to predict successful loan origination based on a limited set of loan and borrower characteristics

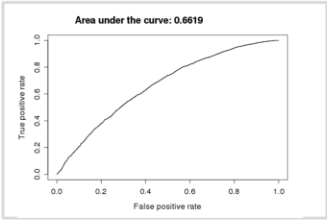
2. Deploy model on FPC web site as a service for potential home loan borrowers, which can drive customers toward FPC for more focused, personal financial planning to achieve their life goals.

3. Develop an efficient model that can provide fast online responses and give an answer within a 45 second SLA

Executive Summary

Loan advice website would drive new revenue for FPC & predict loan approvals with 66% accuracy

- **FPC domain experts identified 3 main growth areas resulting from new website**
  - Refer applicants to more personal loan advice by FPC
  - Partner with real estate brokers for referrals
  - Drive new revenue from advertising with partner real estate brokers



- **Reasonably accurate pre-screening predictions can be made with limited data**
  - Loan Amount, Applicant Income and Loan Type are most influential factors, and predict outcomes accurately 65-70% of the time
  - Sensitive demographic data can be omitted; factors such as Race & Gender show negligible model improvement in most cases
- **Initial POC indicates model can run online within 45 second SLA**
  - Collaborated with IT to benchmark model performance in simulated production environment
  - Initial tests show acceptable speed, but further testing is suggested

EMC² PROVEN PROFESSIONAL

Approach

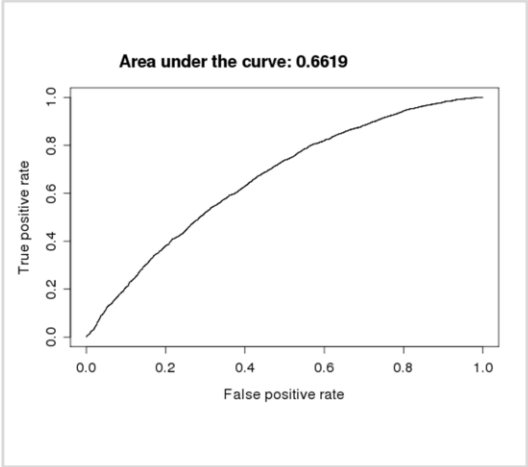
- Consulted with FPC's advisory staff to understand underwriting policy and identify factors that affect a home loan applicant's likelihood of approval
- Collaborated with IT to identify relevant data sets, assess data quality and availability
- Developed predictive model to identify loans most likely to be underwritten
  - Identify most influential factors
  - Provides greater explanatory power for analyzing influence of specific factors on loan origination
- Worked with IT to simulate model performance within FPC's production environment

Model Description

- **Overview of Basic Methodology**: predict the likelihood of loan approval for each applicant. Identify loans with a greater probability for successful underwriting then compare with actual loan originations to train the algorithm and enable predictions for future users.
- **Model:** Logistic regression model
- **Dependent variable:** Binary variable, of origination/no origination
- **Scope & Sampling**
  - 238,000 loan applications for Maryland from 2010
  - All selected loans were Approved or Denied
  - Training sample: 23,800 loans
  - Testing sample: 47,600 loans
- **The model developed has reasonable predictive power for the dataset provided**
  - Loan Type, Loan Amount and Applicant Income provide the most influence over the probability of approval
  - Model shows marginal improvement with sensitive demographic information, such as Applicant Age or Gender

EMC² PROVEN PROFESSIONAL

Model Predicts Accurate Outcome 66% Of The Time



- Model predicts accurate outcome 66% of the time
- Must determine if this is acceptable, or supplement with additional credit data
- AUC/ROC data points:
  - Null deviance: 14,006 on 20,824 degrees of freedom
  - Residual deviance: 13,258 on 20,699 degrees of freedom
  - Psuedo R2 = 1 – (deviance/null deviance), or 1 – (13258/ 14006) = 0.0534

EMC² PROVEN PROFESSIONAL

Above is an example of the output from the price optimization project scenario, showing how one may present this to an audience of other Data Scientists or data analysts. This shows a curvilinear relationship between price tiers and customer loyalty, when expressed as an index. Note that the comments at the right of the graph relate to the precision of the price targeting, the amount of variability in robustness of the model, and the expectations of model speed when run in a production environment.

Recommendations

- **Implement the model on FPC site as a pilot to limited set of site visitors –** test and learn from initial pilot on model outputs
  - Tuning model predictions can encourage users to visit in-person advisors and minimize frustrations by unexpected negative predictions.
  - Outputs can be further enhanced to give users advice on how to increase likelihood of successful loan origination.
- **Scope requirements for a successful pilot as a POC**
  - Identify a **line of business** or specific region to participate in the pilot
  - Identify **production environment constraints** with regard to network bandwidth, compute power, storage capacity
  - Identify **people to implement the model in product**, and monitor for retraining the model based on data inputs

Recommendations

- **Embed links to filtered home listing based on user inputs**
  - *Example*: Show ~$300K homes to users who could afford loan based on income, regardless of desired loan amount
  - Charge partner real estate brokers for PPC (pay-per-click) traffic
- **Over time, refine model by analyzing whether more general or specific advice leads to user conversions**, encouraging users to visit with an in-person advisor for a personalized financial plan.