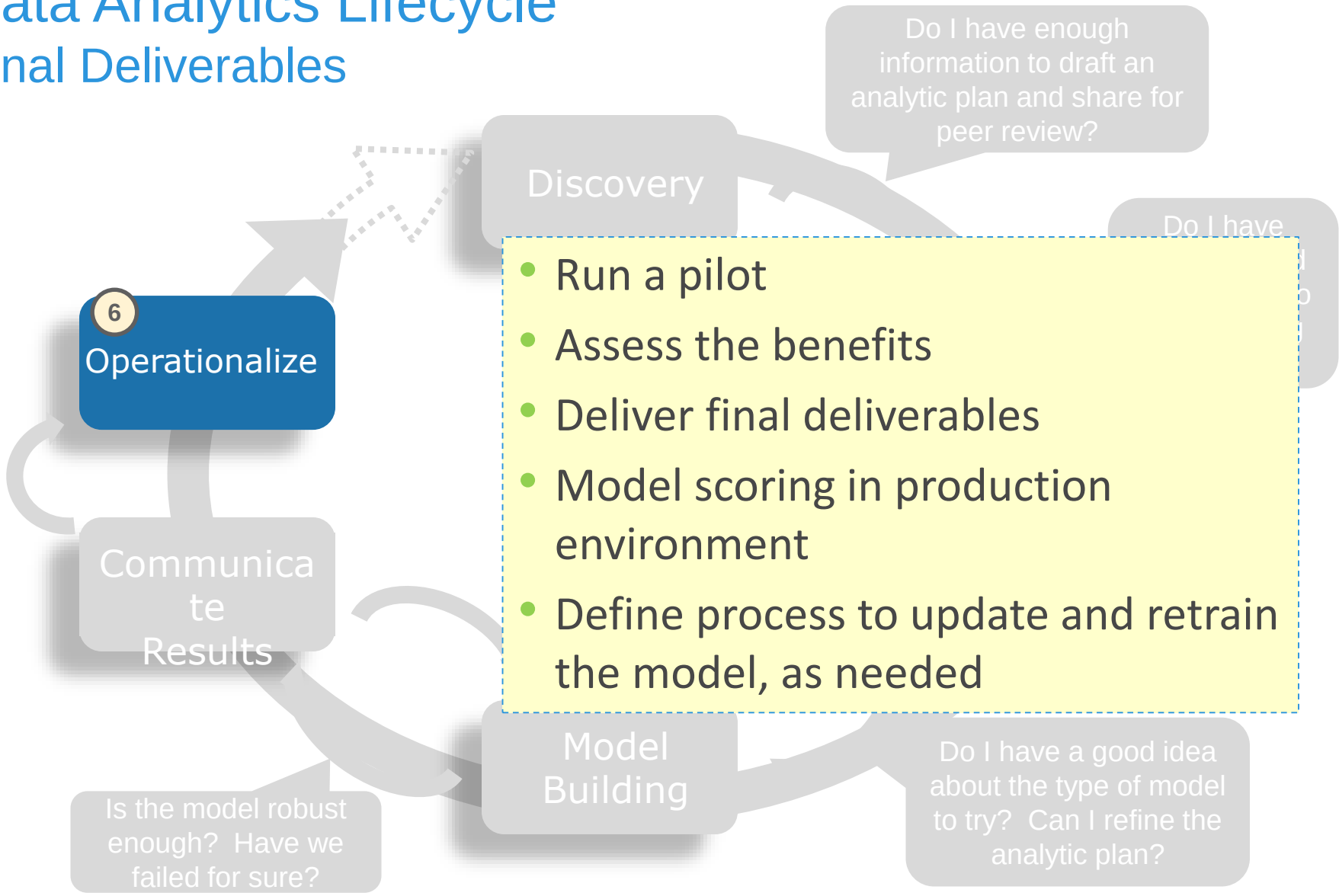
Module 6
Putting It All Together

Data Analytics Lifecycle
Final Deliverables



- Run a pilot
- Assess the benefits
- Deliver final deliverables
- Model scoring in production environment
- Define process to update and retrain the model, as needed

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Outputs from a Successful Analytic Project, by Role

Role	Description	What the Role Needs in the Final Deliverables

Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	• **Sponsor Presentation** addressing:
		  • Are the results good for me?
		  • What are the benefits of the findings?
		  • What are the implications of this for me?

Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	• **Sponsor Presentation** addressing:
		  • What's the business impact of doing this?
		  • What are the risks? ROI?
		  • How can this be go within the organization (and beyond)?

Project Manager	Ensure key milestones and objectives are met on time and at expected quality.	
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective	• Show the **analyst presentation** • Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met	• Show the **analyst presentation** • Share the **code**

Copyright © 2012 EMC Corporation. All Rights Reserved.

Core Deliverables to Meet Most Stakeholder Needs

1. **Presentation for Project Sponsors**
   - "Big picture" takeaways for executive level stakeholders.
   - Determine key messages to aid their decision-making process.
   - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
2. **Presentation for Analysts**
   - Business process changes.
   - Reporting changes.
   - Fellow data scientists will want the details and are comfortable with technical graphs (such as Receiver Operational Characteristic (ROC) curves, density plots, histograms).
3. **Code for technical people**
4. **Technical specs** of implementing the code

Copyright © 2012 EMC Corporation. All Rights Reserved.

Creating the Final Deliverables
ABC Churn Prediction Case Study

Situation Synopsis

- Retail Bank, ABC Bank wants to improve the Net Present Value (NPV) and retention rate of the customers .
- They want to establish an effective marketing campaign targeting customers to reduce the churn rate by at least five percent.
- The bank wants to determine whether those customers are worth retaining. In addition, the bank also wants to analyze reasons for customer attrition and what they can do to keep them.
- The bank wants to build a data warehouse to support marketing and other related customer care groups.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Use Analytic Plan to Guide Final Presentation

Components of Analytic Plan	Retail Banking: ABC 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.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Aspects of Final Presentation Material

- **Reflect on the project:**
  - Consider the context of the problems you set out to solve.
  - Identify observations about the model outputs, scoring, results.
  - Identify Key Messages, and any unexpected insights.
- **Tailor outputs to the audience**

	Project Sponsor Presentation	Analyst Presentation
Focus	What	How

Objectives	• Show that you met the project goals
	• Give your sponsor talking points to do the work
	  • Emphasize ROI (**Return on investment**) and business value
	  • Mention if the models can be deployed within sponsor's SLA (**service-level agreement**)
		• Show how you met the project goals
		• Share your methods so analysts can learn from it for future projects
		  • Discuss methods, techniques, and technologies used.
		  • Provide specific model accuracy and speed (example: how well will it meet SLAs).

Copyright © 2012 EMC Corporation. All Rights Reserved.

Develop Core Material Presentations to 2 Main Audiences

= Same components for both presentations = Different components for Sponsor vs. Analyst presentation

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project Goals	• List top 3 agreed upon goals	• List top 3 agreed upon goals
Main Findings	• Emphasize key message	• Emphasize key message
Approach	• High Level Methodology	• High Level Methodology • Relevant details on modeling techniques and technology
Model Description	• Overview of the modeling technique	• Overview of the modeling technique
3 key points supported with data	• Support key points with simple charts and graphics (example: 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
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. • Identify key variables and impact of each • Describe expected model performance and any caveats • Detailed description of the modeling technique • Discuss variables, scope, predictive power
Recommendations	• Focus on business impact of doing this, including risks and ROI • Give the sponsor salient points to do the work within the organization	• Supplement recommendations with any implications for the modeling, or for deploying in a production environment.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Main Findings (Executive Summary)

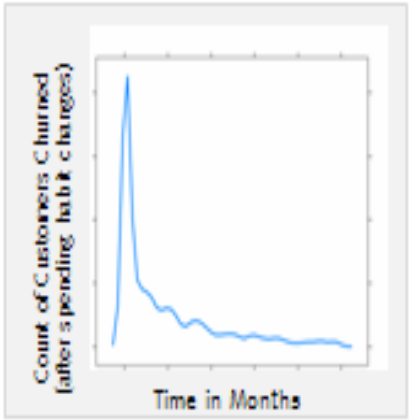
- Enable reader to grasp full synopsis in 1 slide
- Frame outcomes in terms of business value
- Generally same, or very similar, for both types of audiences.

* *CRM: Customer Relationship Management*

Executive Summary

Running an early churn warning test each day using social media data can reduce annual churn by 30% and save $4.5M annually

- **Customers churn within 60 days of changing their spending habits**
  - Most often after customers stop using bank cards for gas and grocery
  - If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.



- **Combining social networking data and existing CRM data increases the model's predictive power to identify churners**
  - We can pinpoint social media chatter from bank customers and influence of churner's contacts
  - With CRM data we can identify 20% of churners, adding social media data increases this to 30%
- **Models can run in minutes, rather than current process of monthly cycles**

Copyright © 2012 EMC Corporation. All Rights Reserved.

Anatomy of an Executive Summary

Key Message

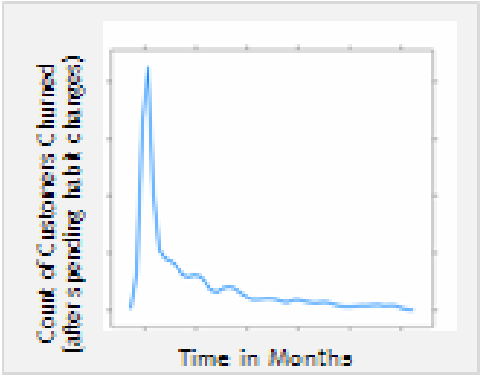
Business Impact

Executive Summary

Running an early churn warning test each day using social media data can reduce annual churn by 30% and save $4.5M annually

Major Points

- **Customers churn within 60 days of changing their spending habits**
  - Most often after customers stop using bank cards for gas and grocery
  - If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.



- **Combining social networking data and existing CRM data increases the model's predictive power to identify churners**
  - We can pinpoint social media chatter from bank customers and influence of churner's contacts
  - With CRM data we can identify 20% of churners, adding social media data increases this to 30%

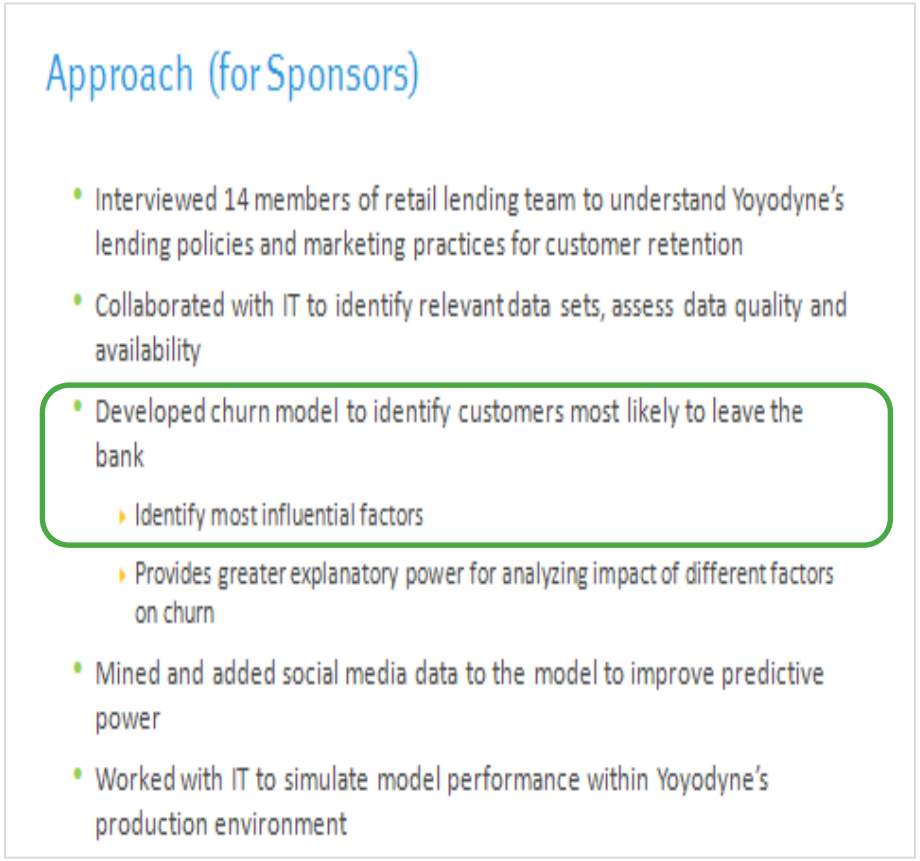
SLA

- **Models can run in minutes, rather than current process of monthly cycles**

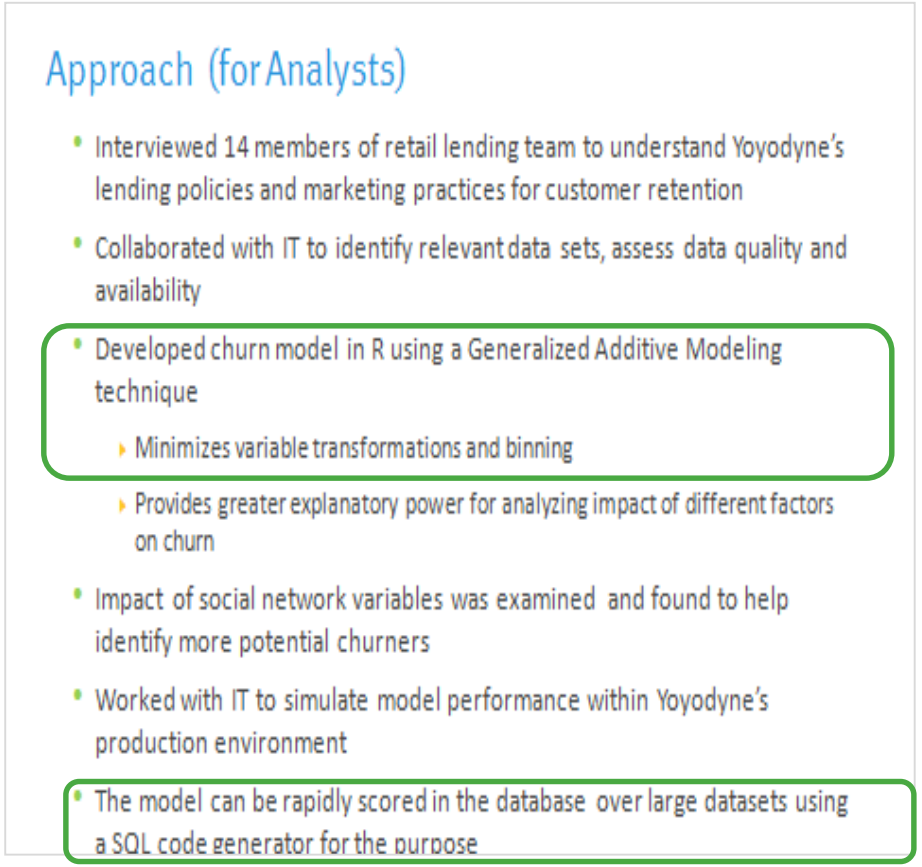
Copyright © 2012 EMC Corporation. All Rights Reserved.

Approach

Example Approach slide, for Sponsors



Example Approach slide, for Analysts



Note: Green boxes highlight differences between slides

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Points Supported With Data

- 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

Copyright © 2012 EMC Corporation. All Rights Reserved.

Recommendations

Recommendations

- **Implement the model as a pilot, before more wide-scale rollout** – test and learn from initial pilot on performance and precision.
  - Addressing these promptly can potentially save more customers from churning over time and also prevent more networking that seems to drive additional churn.
  - An early churn warning trigger can be set up based on this model.
- **Run the predictive model daily or weekly to be proactive on customer churn**
  - In-database scorer can score large datasets in a matter of minutes and can be run daily
  - Each customer retained via early warning trigger saves 4 hours of account retention efforts & 50k in new account acquisition costs
- **Develop targeted customer surveys to investigate the causes of churn**, which will make the collection of data for investigation into the causes of churn easier.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Tips, Tricks, & Pitfalls to Avoid for the Final Presentation

1. Be visual. Generally, the more visual the better. Up to a point.
2. Be MECE (Mutually Exclusive and Collectively Exhaustive).
3. Tie your ideas together….don't force people to tie your ideas together, guide people and help them draw logical connections.
4. Don't forget that not everyone has gone through the Discovery phase like you have.
5. Context is key. Orient people to the project itself, as well as the graphics you use, the terminology and jargon (spell out acronyms).
6. Don't assume people see the obvious benefits.
7. Measure and quantify the benefits. Be specific. "*$8.5M in annual cost savings*" is much stronger than "*Great Value*".
8. Be patient. You may have to tell your story more than once…consider these sessions opportunities to refine your message and share good work that was done.
9. Let the intended audience guide you in shaping the right message and level of detail.
10. Avoid long bulleted lists

Copyright © 2012 EMC Corporation. All Rights Reserved.

Overview of Code & Technical Documentation

- **Consider the interests of your technical audience:**
  - How will the project affect them?
  - In what ways will it change their day-to-day roles, or existing processes?
  - Be aware of the implications of your work on their roles as you create these technical deliverables.

- **2 Technical deliverables:**
  - Code
  - Technical specifications and documentation.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Considerations for Technical Specifications & Documentation

Approach the documentation as if it's for an API *(application programming interface)*

- **Inputs & Pre-processing:**
  - Discuss the expected pre-processing steps before data goes to the model code.
  - Document expected input, data format, source tables, and units.
  - Describe the processing script are you using.
  - Explain how the outputs are created.
- **Exception handling :**
  - Explain how to deal with exceptions to the model.
  - Provide guidance for making decisions on the exceptions.
- **Post-processing:**
  - After you create the output, discuss any post-processing before going to the next step.
  - Interpreting a threshold as opposed to a simple yes/no.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Providing Your Code

- Test for accuracy in the production environment
- Ensure the code will run quickly and meet SLAs
- Include comment lines in the code
- Hold a briefing with the engineers who will implement the code

Copyright © 2012 EMC Corporation. All Rights Reserved.

Data Visualization Techniques - Key Points Supported With Data Overview of Visualization Tools

Open Source

- **R**
  - **Base package**
  - ggplot
  - Lattice
- Ggobi/Rggobi
- Inkscape
- Processing
- Modest Maps
- GnuPlot

Commercial Tools

- Tableau
- Spotfire
- Qlikview
- Adobe Illustrator

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Points Supported With Data
Tables of Information. Note: BigBox and SuperBox are the names of the companies stores

44 years of BigBox stores data

Year	1962 1964	1965	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Grand Total
SuperBox	1 1	1	1	5	4	4	14	13	14	20	14	17	29	24	37	33	117	42	65	79	81	90	92	82	86	106	72	62	62	40	49	22	26	33	47	78	71	67	64	91	91	33	1980
BigBox			1		1	1	1	4	5	5	5	10	10	10	6	21	33	21	22	20	29	31	50	43	45	72	91	76	94	67	80	31	34	33	33	27	35	47	32	39	27	4	1196
Grand Total	1 1	1	2	5	5	5	15	17	19	25	19	27	39	34	43	54	150	63	87	99	110	121	142	125	131	178	163	138	156	107	129	53	60	66	80	105	106	114	96	130	118	37	3176

34 years of BigBox stores data

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Grand Total
SuperBox	13	14	20	14	17	29	24	37	33	117	42	65	79	81	90	92	82	86	106	72	62	62	40	49	22	26	33	47	78	71	67	64	91	91	33	1980
BigBox	4	5	5	5	10	10	10	6	21	33	21	22	20	29	31	50	43	45	72	91	76	94	67	80	31	34	33	33	27	35	47	32	39	27	4	1196
Grand Total	17	19	25	19	27	39	34	43	54	150	63	87	99	110	121	142	125	131	178	163	138	156	107	129	53	60	66	80	105	106	114	96	130	118	37	3176

- What do you observe from this data?
- What's the main message?
- What is the author trying to emphasize with the data?

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Points Supported With Data Tables of Information

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.

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Points Supported With Data Tables of Information

There are several observations in the data (if you look closely), such as …

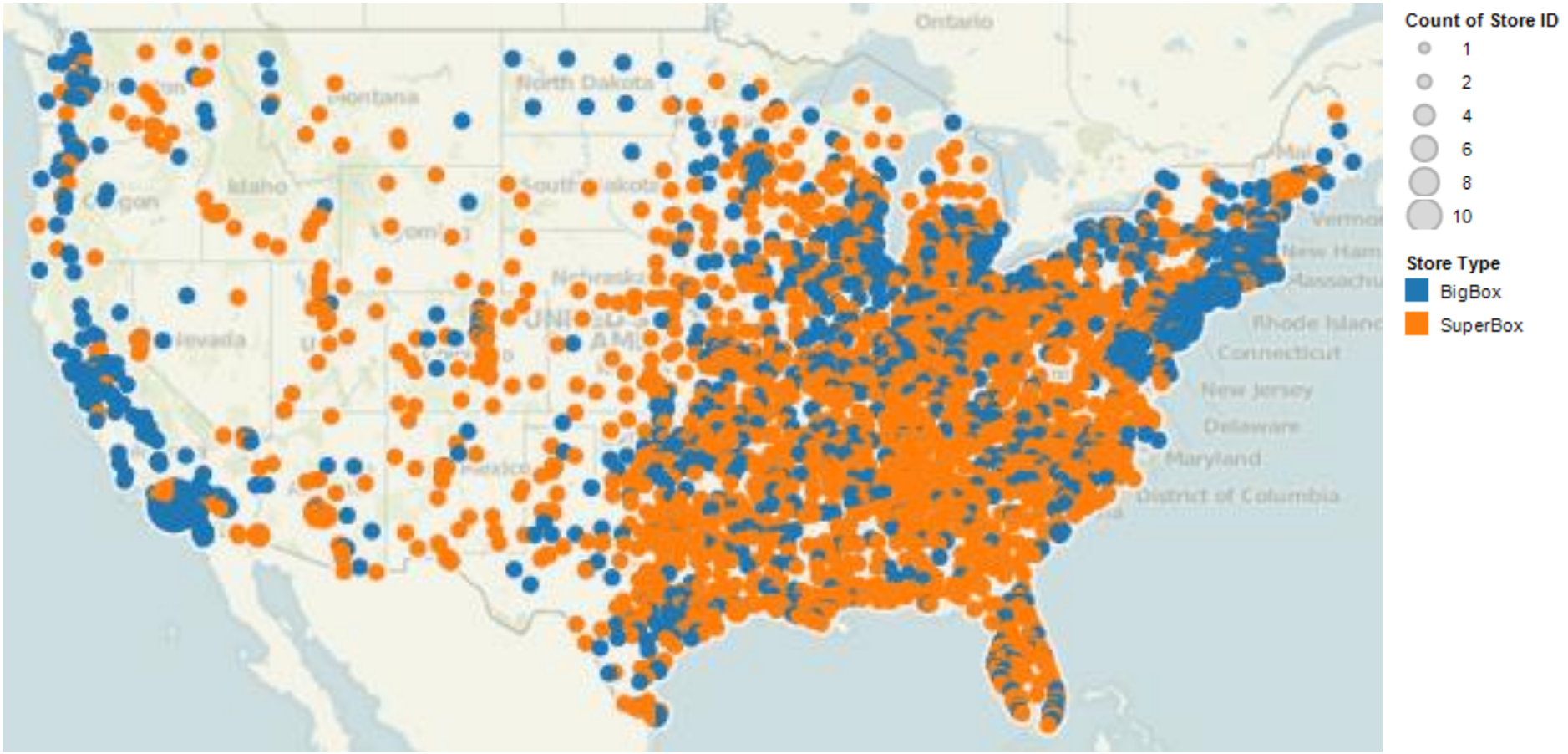
1) BigBox experienced strong growth in the 1980s and 1990s
2) By the 1980s, BigBox began adding more SuperBox stores to its mix of chain stores
3) SuperBox outnumber regular stores nearly 2 to 1

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.

Copyright © 2012 EMC Corporation. All Rights Reserved.

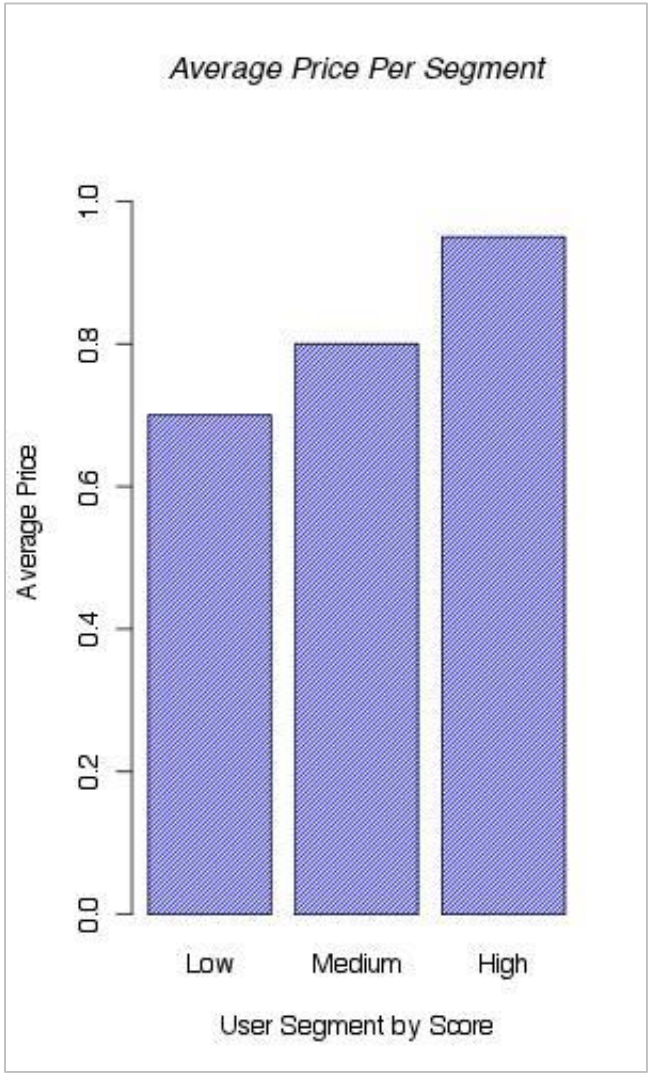
Key Points Supported With Data
Using Visuals to Illustrate Key Points

Example of a Visual to help tell a story to a <u>Sponsor</u>



Copyright © 2012 EMC Corporation. All Rights Reserved.

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*

Copyright © 2012 EMC Corporation. All Rights Reserved.

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

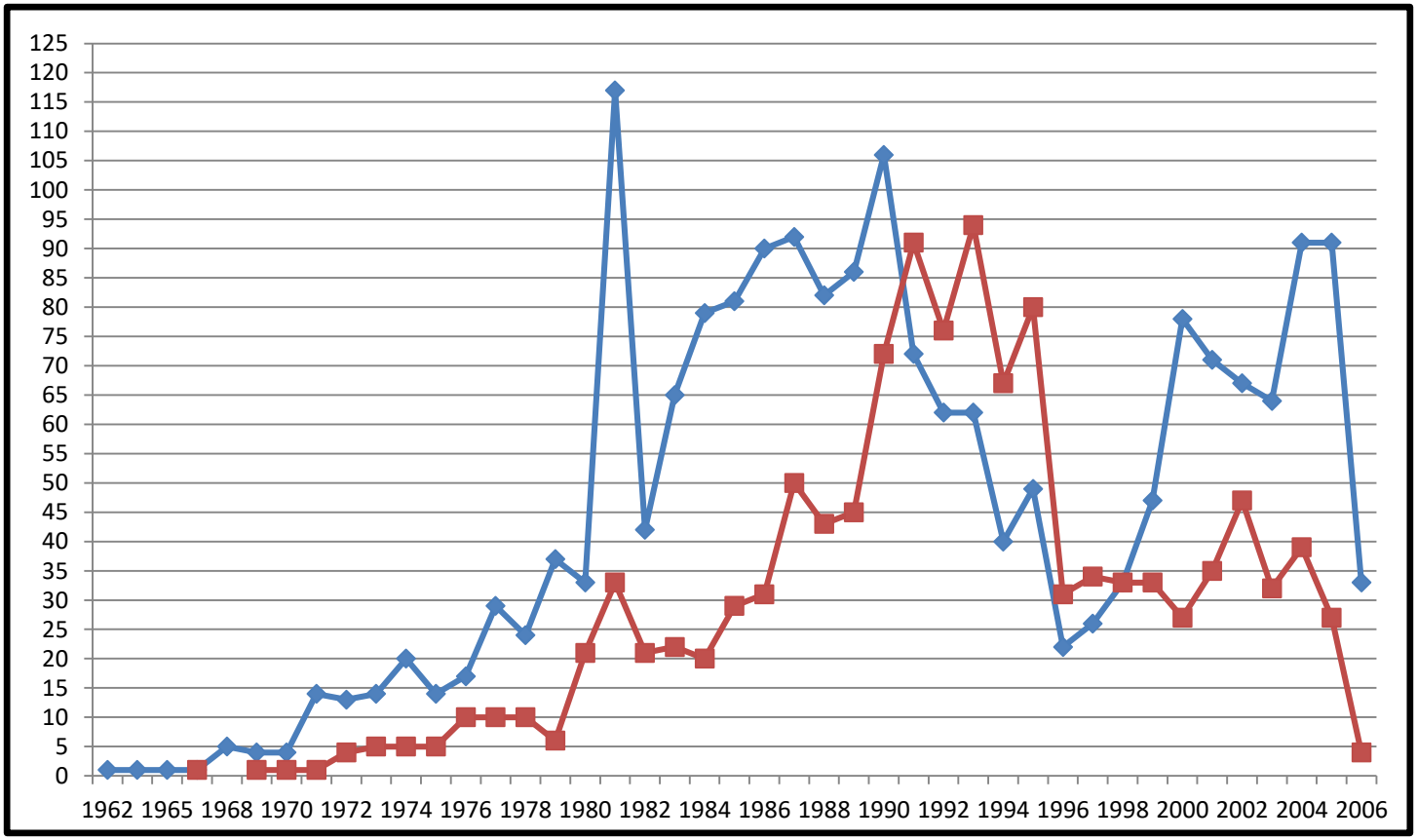
Copyright © 2012 EMC Corporation. All Rights Reserved.

How to Clean Up a Graphic - Example 1 Before Cleaning

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

Chart junk

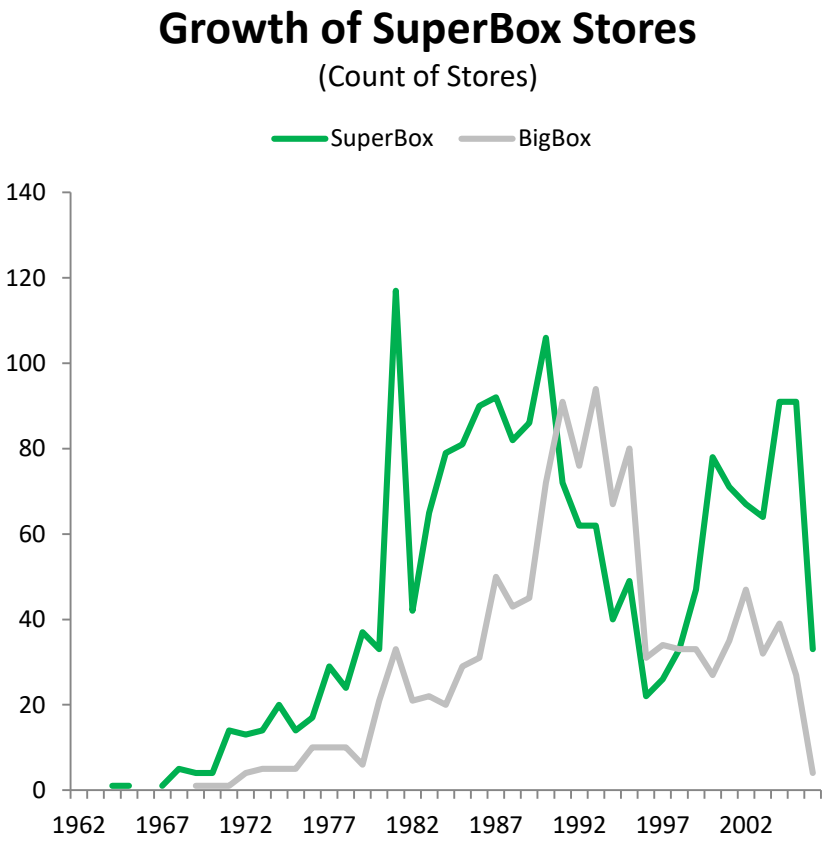
1. Horizontal Grid Lines
2. Chunky data points
3. Overuse of emphasis colors; lines & border
4. No context or labels
5. Crowded axis labels

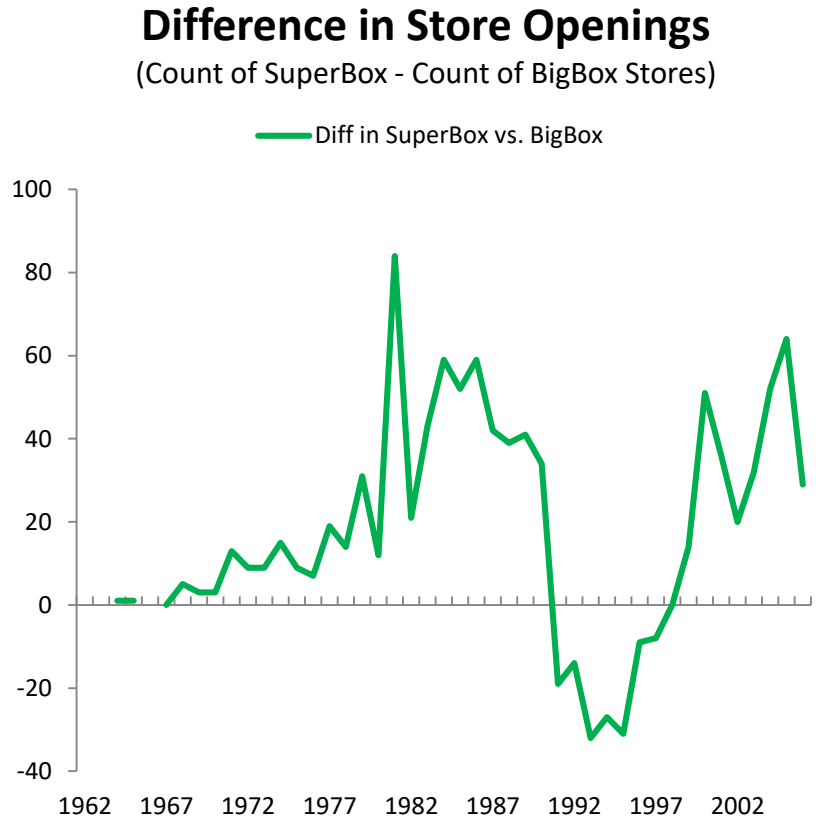


Copyright © 2012 EMC Corporation. All Rights Reserved.

How to Clean Up a Graphic - Example 1 After Cleaning

- What are the main messages here?
- What is the author trying to emphasize?





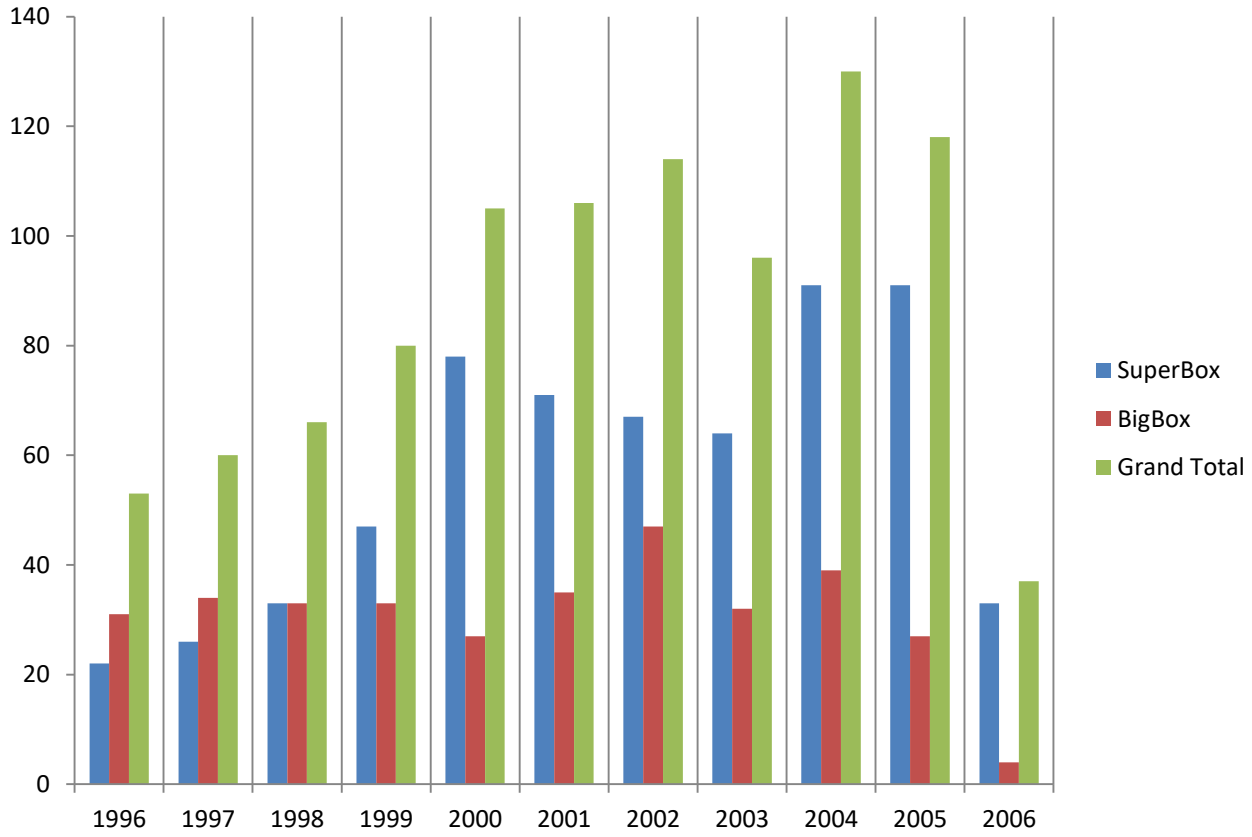
Copyright © 2012 EMC Corporation. All Rights Reserved.

How to Clean Up a Graphic - Example 2 Before Cleaning

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

Chart junk

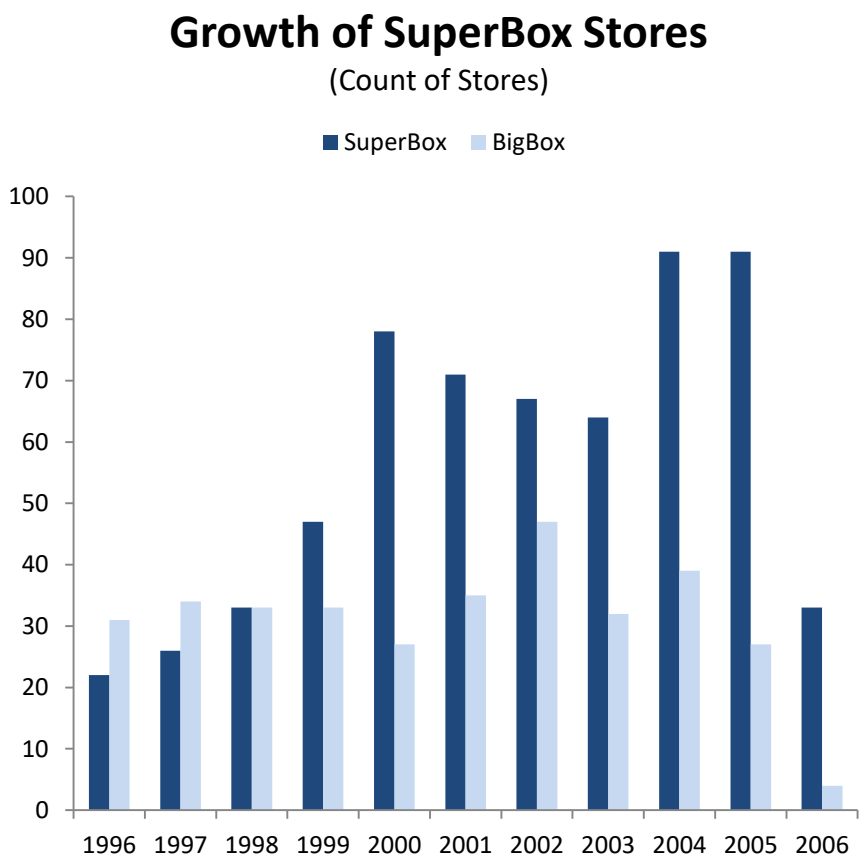
1. Vertical Grid Lines
2. Too much emphasis colors
3. No chart title
4. Legend at right restricts chart space
5. Labels are too small

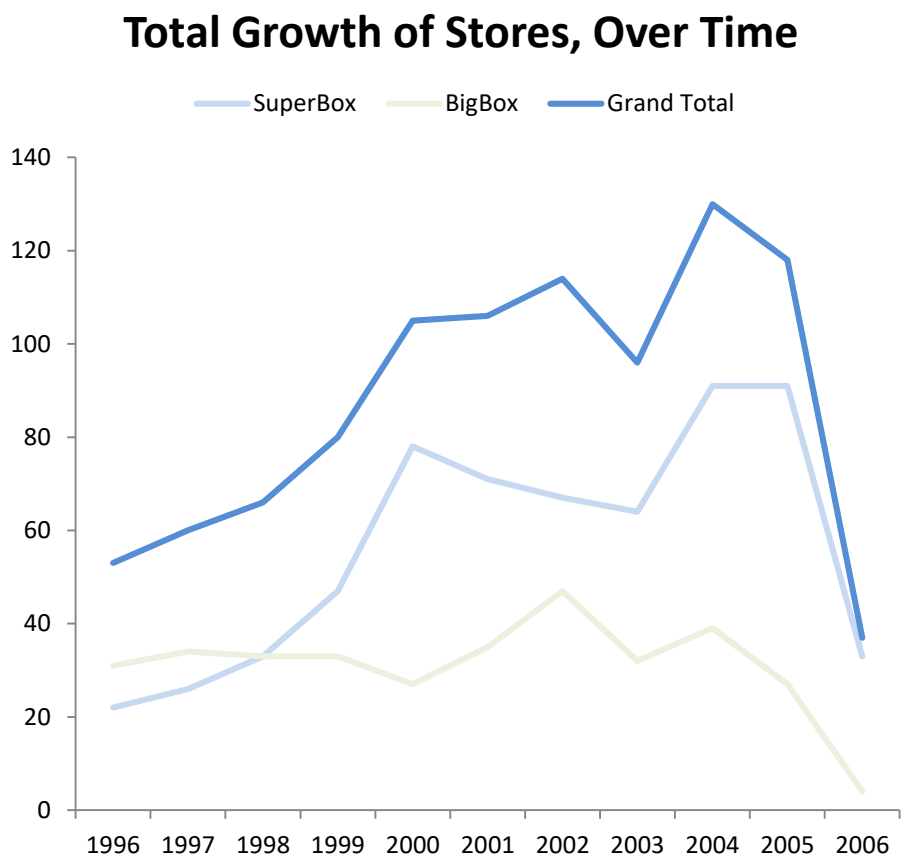


Copyright © 2012 EMC Corporation. All Rights Reserved.

How to Clean Up a Graphic - Example 2 After Cleaning

- What are the main messages here?
- What is the author trying to emphasize?





Copyright © 2012 EMC Corporation. All Rights Reserved.

Using 3D Charts: Avoid Them!

2-Dimensional Charts

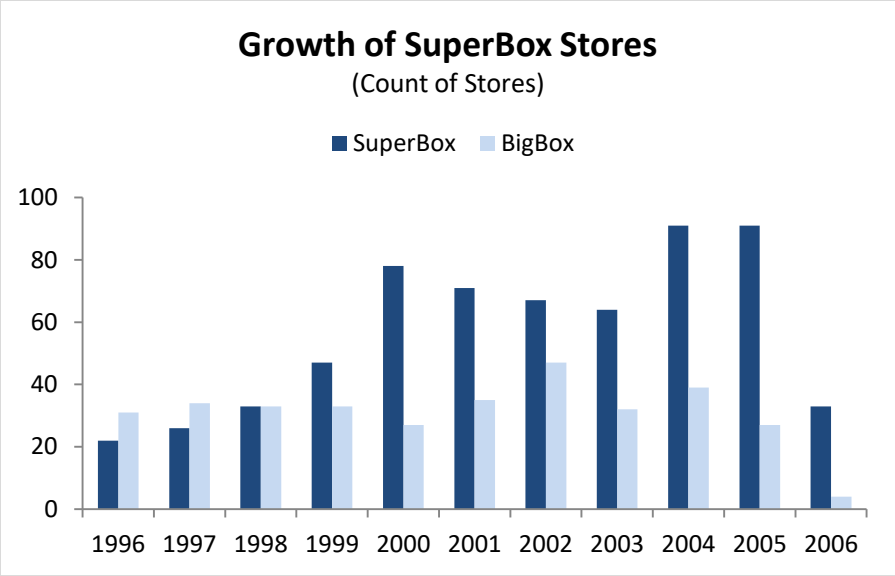


Chart A: 2-Dimensional

- Simple
- Easy to understand
- Focus on the data, not the graphics

3-Dimensional Charts

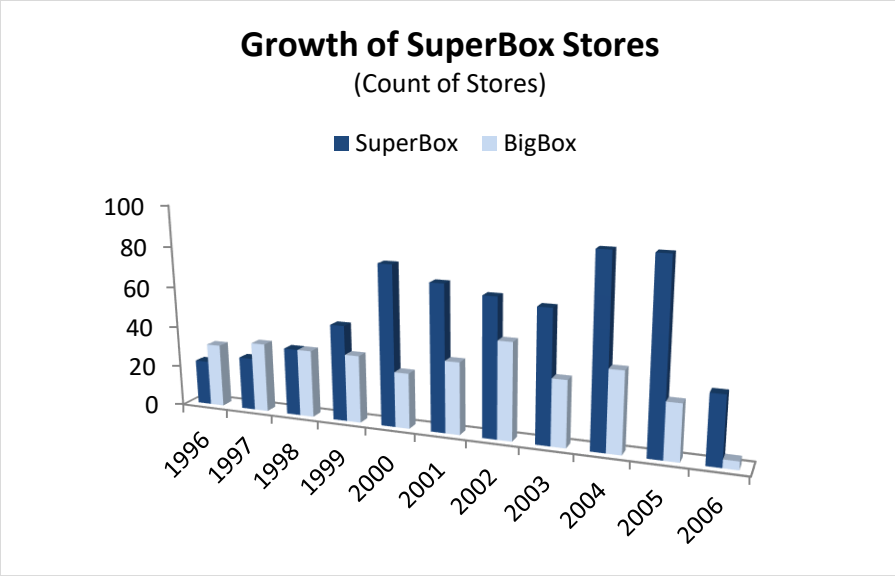


Chart B: 3-Dimensional

- Difficult to gauge actual data
- Scaling becomes deceptive
- Does not make graphic fancier, just harder to understand

Copyright © 2012 EMC Corporation. All Rights Reserved.

Key Points with Data Visualizations

- **Remove distractions**
  - Minimize “chart junk”
  - Data-Ink Ratio
- **Choose the simplest, clearest visual for the situation**
  - Strive to illustrate your points
  - Charts should serve to reinforce your key points
  - Charts vs. Data Art
- **Use color deliberately**
  - Emphasis Colors vs. Standard Colors
  - In most cases, less is more
  - Focus on the contrast
- **Context**
  - Consistent scales, labels, axes
  - Using logs vs. raw values to show differences

Copyright © 2012 EMC Corporation. All Rights Reserved.